Multi-provider Service and Transport Architectures

Stefano Secci

POLITECNICO DI MILANO
Dipartimento di Elettronica e Informazione
DOTTORATO DI RICERCA IN INGEGNERIA DELL'INFORMAZIONE

MULTI-PROVIDER SERVICE AND TRANSPORT ARCHITECTURES

Doctoral Dissertation of:
Stefano Secci

Advisors:
Prof. Achille Pattavina
Prof. Jean-Louis Rougier

Tutor:
Prof. Antonio Capone

Supervisor of the Doctoral Program:
Prof. Patrizio Colaneri

2009 - XXII

Thèse

Présentée pour obtenir le grade de docteur
de Télécom ParisTech
Spécialité: Informatique et Réseaux

Stefano SECCI

Architectures de service et transport multi-opérateurs

Soutenue le 9 decembre 2009 devant le jury composé de:

Rapporteurs	Olivier BONAVENTURE	Université Catholique de Louvain, Belgique
	Bijan JABBARI	George Mason University, Etats-Unis
	Xavier MASIP BRUIN	Universitat Politècnica de Catalunya, Espagne
Examinateurs	Gérard MEMMI	Télécom ParisTech, France
	Michal PIÓRO	Politech. Warszawska, Pologne
		Lunds Universitet, Sweden
Présidente	Brunilde SANSÒ	École Polytechnique de Montréal, Canada
Directeurs de thèse	Achille PATTAVINA	Politecnico di Milano, Italie
	Jean-Louis ROUGIER	Télécom ParisTech, France

TELECOM PARISTECH
Département d'Informatique et Réseaux
Doctorat en Informatique et Réseaux

POLITECNICO DI MILANO
Dipartimento di Elettronica e Informazione
Dottorato in Ingegneria dell'Informazione

MULTI-PROVIDER SERVICE AND TRANSPORT ARCHITECTURES

Author: Stefano SECCI

Defended the 9 December 2009 in front of the committee composed of:

Referees: Prof. Olivier BONAVENTURE (Université Catholique de Louvain, Belgium)
Prof. Bijan JABBARI (George Mason University, USA)
Prof. Xavier MASIP BRUIN (Universitat Politècnica de Catalunya, Spain)

Examiners: Prof. Gérard MEMMI (Télécom ParisTech, France)
Prof. Michal PIÓRO (Politech. Warszawska, Poland; Lunds Universitet, Sweden)
President: Prof. Brunilde SANSÒ (École Polytechnique de Montréal, Canada)

Advisors: Prof. Achille PATTAVINA (Politecnico di Milano, Italy)
Prof. Jean-Louis ROUGIER (Télécom ParisTech, France)

Acknowledgements

I am sincerily grateful to my advisors, Professors Achille Pattavina and Jean-Louis Rougier for their unique and often complementary experience and support, suggestions and corrections, professionality and moral integrity, an advisory every Ph.D. candidate shall profit from. Professor Brunilde Sansò, as my de-facto third advisor, equally deserves my sincere admiration and respect. Following their example I could shape in the right direction my working attitude and scientific approach; even when their scope of expertise did not go that far, they could indicate me the right researchers to collaborate with.

I am very lucky to have collaborated with excellent researchers; in organization alphabetical order, Richard Douville, Ramon Casellas, Jean-Louis Le Roux, Mariusz Mycek and Michal Pióro and Artur Tomaszewski, Guido Maier and Federico Malucelli, Fioravante Patrone, Alberto Ceselli, and Massimo Tornatore. They all were unique colleagues that brought, and thankfully are still bringing directly with active collaborations or indirectly with their excellent researches, precious bricks to the consistency of my research work of the last four years.

I am grateful to Olivier Bonaventure, Bijan Jabbari and Xavier Masip Bruin for accepting to review the dissertation, and for helping me in improving it with their internationally recognised expertise.

A warm acknowledge goes to the students that had the opportunity to work on my projects, Aruna Prem Bianzino, Nabil Bachir Djarallah, Eric Elena, Michael Huaiyuan Ma and Matteo Marinoni, and to all the colleagues from the research teams I have co-worked in since 2004, from the Broadband Lab. of Ecole Polytechnique de Montréal, the Networking team of Politecnico di Milano, and the Network, Mobility and Security team of Telecom ParisTech.

A large number of friends, researchers and mates from various experiences, projects and conference meetings, persons coming from every corner of the world, enriched my cultural and professional experience, and probably also fuzzyly contributed to some of the bizarre ideas of this dissertation. The same equally stands for all the students I teached network protocols and architectures at Télécom ParisTech and EFREI schools in France, inevitably arising questions that let you reflect, with a learner standpoint, on subjects you though were a master in.

Last but surely not least, I deeply thank my family that resigned to accept my worldwide mobility anyway, and Anca, excellent companion and tremendous woman, I will always keep adoring you!

“Non sunt nova veteribus substituenda,
sed perpetuo iungenda fœdere"

Let us not substitute new things for old ones,
but add them piecemeal with awareness. (Anonymous)

“Nolite (...), neque mittatis margaritas vestras ante porcos (...)"

Do not throw your pearls before swines. (Notable)

Abstract

The dissertation presents various technical solutions to improve the level of collaboration among providers in support of inter-provider network services. The scientific contribution embraces different networking research facets, from IP routing to G-MPLS provisioning and network design optimization, applying concepts from graph theory, game theory and operations research.

By an in-depth analyzis of recent Internet routing traces, we show that the current inter-domain (connection-less) routing suffers from a lack of coordination that produces inefficiencies and frequent route deviations. With respect to this issue, relying on concepts of non-cooperative game theory, we propose coordination strategies to improve the current BGP routing across peering settlements, while preserving the providers' independence and respective interests. We show that their implementation can avoid congestion on peering links, reduce significantly the routing cost and successfully control the route deviations. The mathematical model can be extended to support a new form of peering agreement extended to multiple providers, but its adoption may appear too weak with respect to alternative solutions able to guarantee end-to-end cross-provider Quality of Service (QoS).

The support of strict end-to-end QoS constraints for added-value services imposes, indeed, a higher level of collaboration on the multi-provider agreement. It is required to reserve resources for own services in other providers' networks. These requirements bring towards a new interconnection model, the "provider alliance", as a cooperative framework that providers shall deploy to allow dynamic connection-oriented service routing and provisioning. We define the functional architecture of a service plane managing service-related data within the provider alliance, together with the instantiation and activation of multi-provider tunnel and circuit services. We highlight the required protocol

extensions for the distributed (router-level) path computation and the dynamic resource reservation, which have been implemented and validated in a testbed. We define, moreover, specific AS-level routing algorithms that scale with the proposed model, supporting pre-computation and directional transit metrics. Finally, we show how providers shall cooperate also to statically reserve link resources, in an optimal and distributed fashion, modelling the economical incentives and the strategic position of each provider in such a cooperation with the application of concepts from cooperative game theory (precisely, the Shapley Value concept).

In the second part of the dissertation, we tackle more physical issues related to the provisioning of tunnels and circuits across Internet eXchange Point (IXP) infrastructures. We present a novel very-high-capacity optical transport architecture, called the Petaweb, as a possible next generation IXP solution and, more generally, as a possible very-high-capacity transport architecture. It consists in a regular direct interconnection scheme of electronic access nodes via optical switches disconnected from each other. This structure can allow a simple inter-provider G-MPLS signalling, can drastically simplify traffic engineering operations, and can facilitate modular upgrades of network elements, at the expense of potentially higher installation costs.

We formulate the design dimensioning problem of the Petaweb composite-star topology, which is NP-Hard, and propose a scalable and efficient heuristic approach. Moreover, we propose a quasi-regular structure for the same transport architecture, less costly and slightly more complex (requiring wavelength conversion), for which we also formulate the design problem and propose an efficient heuristic. We argue by simulations that the physical dimensioning of classical multi-hop optical networks under additive path metric minimisation (such as the delay) would produce a solution that tends toward a quasi-regular Petaweb structure. To conclude, we analyze how practically a network planner decision-maker shall trade-off – when discriminating among many Petaweb solution alternatives – the various performance criteria with the level of reliability, survivability and availability.

Riassunto in lingua italiana

La tesi presenta diverse soluzioni tecniche atte a migliorare il livello di collaborazione tra operatori Internet nell'offerta di servizi di rete inter-operatore. Il contributo scientifico abbraccia diversi aspetti della ricerca del settore, coprendo problematiche di instradamento IP, di fornitura di servizi G-MPLS e di dimensionamento di rete, applicando concetti di teoria dei grafi, teoria dei giochi e ricerca operativa.

Attraverso l'analisi approfondita di recenti mappe di instradamento Internet, mostriamo che gli attuali schemi di instradamento (in modalità non connessa) soffrono di una mancanza di coordinamento che produce inefficienze e frequenti deviazioni dal miglior cammino scelto. Riguardo a questa problematica, affidandoci a concetti di teoria dei giochi non-cooperativi, proponiamo l'implementazione di strategie di coordinamento per migliorare l'attuale instradamento BGP sulle interconnessioni di *peering*, pur rispettando l'indipendenza e gli interessi rispettivi degli operatori. Risultati sperimentali mostrano che in tal modo si possono evitare congestioni sulle interconnessioni di *peering*, che si può ridurre significativamente il costo di instradamento e che si possono controllare con successo le deviazioni. Il modello matematico può essere esteso per supportare una nuova forma di accordo di peering esteso a più di due operatori, ma gli incentivi alla sua adozione potrebbero rivelarsi troppo deboli rispetto a soluzioni alternative in grado di garantire la qualità di servizio inter-operatore tra nodi terminali.

Il supporto di stringenti vincoli di qualità di servizio tra nodi terminali impone, infatti, un più alto livello di collaborazione all'accordo di interconnessione multi-operatore. Occorre poter prenotare risorse per i propri servizi nelle reti altrui. Tali requisiti portano alla definizione di un nuovo modello di interconnessione, la "alleanza multi-operatore", come struttura di cooperazione che gli operatori dovrebbero realizzare per permettere l'instradamento e la fornitura dinamici di servizi di rete in modalità connessa. Definiamo l'architettura funzionale di un piano di servizio multi-operatore per la gestione di dati specifici ai servizi all'interno dell'alleanza, così come per il supporto dell'instanziazione

e dell'attivazione di servizi (tunnel e circuiti) multi-operatore. Mettiamo in evidenza le necessarie estensioni protocollari per il calcolo distribuito dei cammini (a livello *router*) e per la prenotazione dinamica delle risorse, che sono state implementate e validate tramite banco di prova. Proponiamo inoltre degli algoritmi scalabili e specifici per l'instradamento delle connessioni al piano di servizio (a livello *AS*), con supporto di pre-computazione e di metriche di transito direzionali. Infine, mostriamo come gli operatori dovrebbero cooperare nella prenotazione statica delle risorse, in modo ottimale e distribuito, modelizzando gli incentivi economici e la posizione strategica di ognuno in tale cooperazione tramite l'applicazione di concetti di teoria dei giochi cooperativi (precisamente, il valore di Shapley).

Nella seconda parte della tesi, affrontiamo problematiche più fisiche relative alla prenotazione delle risorse per tunnel e circuiti attraverso infrastrutture di interscambio Internet (*IXP*). Presentiamo una nuova architettura di trasporto ottico ad altissima capacità, chiamata "Petaweb", come possibile soluzione di IXP di nuova generazione e, più in generale, come possibile architettura di rete di trasporto ad altissima capacità. La rete Petaweb consiste in uno schema regolare d'interconnessione diretta dei nodi di accesso attraverso commutatori ottici non interconnessi tra loro. Tale architettura può permettere una semplice segnalazione G-MPLS attraverso le frontiere d'operatore, può semplificare drasticamente le operazioni di ingegneria del traffico, e può facilitare l'aggiornamento modulare degli elementi di rete, a scapito di costi di installazione potenzialmente più importanti.

Definiamo il problema di progetto della topologia Petaweb a sovrapposizione di stelle, che è NP-completo, e proponiamo un approccio euristico che si dimostra scalabile ed efficiente. Inoltre, proponiamo un'alternativa struttura quasi-regolare, meno costosa e leggermente più complessa (richiedente conversioni di lunghezza d'onda), per cui pure definiamo il problema di progetto e proponiamo un efficiente algoritmo euristico. Argomentiamo tramite simulazioni che il dimensionamento fisico di classiche reti ottiche multi-salto con minimizzazione di metriche di percorso additivi (come il ritardo) produrrebbe una soluzione che tende verso una struttura di tipo Petaweb quasi-regolare. In conclusione, analizziamo come il decisore pianificatore di rete potrebbe - nella discriminazione tra possibili soluzioni alternative di tipo Petaweb - bilanciare i diversi fattori di prestazione e costo con il livello di affidabilità, scalabilità e disponibilità della soluzione offerta.

Résumé long en langue française

La thèse présente différentes solutions techniques capables de supporter différents niveaux de collaboration entre opérateurs Internet pour l'offre de services de réseau inter-opérateur. La contribution scientifique regroupe différents aspects de la recherche dans le domaine, en couvrant des problématiques de routage IP, d'approvisionnement de services de réseau G-MPLS et de dimensionnement de réseau, ceci en appliquant concepts de théorie des graphes, théorie des jeux et recherche opérationnelle.

Dans ce résumé long nous présentons avec un discret niveau de détail le contenu et les résultats de la thèse; les modèles et formulations mathématiques, les démonstrations, les justifications réelles, le détail sur les expérimentations et les références à l'état de l'art sont présentés dans le corps de la thèse.

Amélioration du routage IP dans le cœur d'Internet

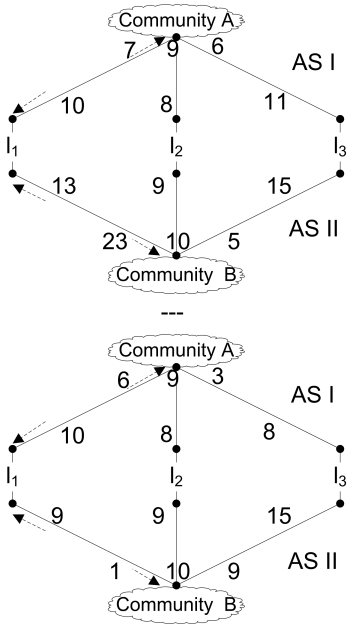
A travers l'analyse détaillée de récents plans de routage Internet, nous montrons au début de la thèse que les actuels schémas de routage (en mode non connecté) souffrent d'une absence de coordination qui produit inefficiences et fréquentes déviations du meilleur chemin choisi. Cela se manifeste au niveau du routage inter-domaine - réalisé par le protocole "BGP (Border Gateway Protocol)"- par le choix de la chaine d'opérateurs, mais aussi au niveau du routage interne - guidé par les protocoles "IGP (Interior Gateway Protocols)"- par le choix du nœud de sortie du réseau.

A long terme, ces phénomènes peuvent être très problématiques car ils peuvent sérieusement affecter la qualité de service offerte pour les services Internet en mode non connecté. Par rapport à cette problématique, en s'appuyant sur les concepts de théorie des jeux non-coopératifs, nous proposons l'implémentation de stratégies de coordination pour améliorer le routage BGP actuel au niveau des interconnexions de *peering*, tout en respectant l'indépendance et les intérêts respectifs des opérateurs. En fait, comme le

montre l'analyse des déviations, les interconnections de peering sont en train de devenir le vrai point d'étranglement dans l'architecture Internet; l'accord étant presque toujours sans paiements latéraux, deux opérateurs en peering ne sont pas économiquement et opérationnellement motivés à suivre les préférences de routage de leur voisin, ce qui produit de fréquentes congestions, et donc des déviations et des pertes de paquets.

Nous proposons alors différentes stratégies de coordination qui suggèrent de passer de simples choix égoïstes dans le routage inter-opérateurs à des choix plus collaboratifs qui trouvent leur justification rationnelle dans les équilibres de routage dits "équilibres de Nash". Des résultats expérimentaux - faisant référence à une émulation de l'interconnexion entre les réseaux de recherche européen et nord-américain - montrent que, ainsi, on peut éviter les congestions sur les liens d'interconnexion de *peering* entre opérateurs, qu'on peut réduire significativement le coût de routage et que les déviations peuvent être contrôlées avec succès.

Un exemple de jeu non-coopératif de coordination est représenté dans la figure suivante. Nous montrons deux configurations avec à coté les respectifs jeux en forme stratégique. Dans le premier, nous avons un seul équilibre de Nash (le profil en gras) qui est aussi le seul profile efficient dans le sens de Pareto, et donc représente une bonne solution optimale de routage. Dans le deuxième cas, il y a quatre équilibres de Nash (en gras), dont un seul Pareto-supérieur aux autres (en cursif) qui n'est cependant pas Pareto-efficient. Il y a aussi un profil Pareto-efficient (l_1, l_3) qui n'appartient pas à l'ensemble de Nash. La nécessité de définir des stratégies de coordination pour le choix des bons équilibres de Nash est donc évidente.



I\II	l_1	l_2	l_3
l_1	(17,36)^6	(19,32)^2	(16,38)^8
l_2	(15,23)^4	(**17,19**)^0	(14,25)^6
l_3	(18,18)^7	(20,14)^3	(17,20)^9

I\II	l_1	l_2	l_3
l_1	(16,10)^2	(19,10)^2	(13,16)^8
l_2	(**14,19**)^0	(**17,19**)^0	(11,25)^6
l_3	(***14,18***)^0	(**17,18**)^0	(11,24)^6

Figure 1: Exemple de jeu de coordination avec trois liens.

La structure du jeu de coordination est représentée dans la figure suivante. Le jeu est une composition de trois jeux: un jeu G_s de pure coordination ou "égoïste"défini par les coûts de routage interne des nœuds de sortie vers les liens de peering; un jeu G_d de pure externalité défini par les coûts de routage interne des liens de peering vers les nœuds de sortie; un jeu G_c de congestion construit suivant une fonction de congestion standard utilisant les débits des flots passant d'une frontière à l'autre. Le jeu résultant est un jeu cardinal avec potentiel, c'est-à-dire un jeu qui peut être défini complètement avec une seule fonction de coût pour les deux joueurs.

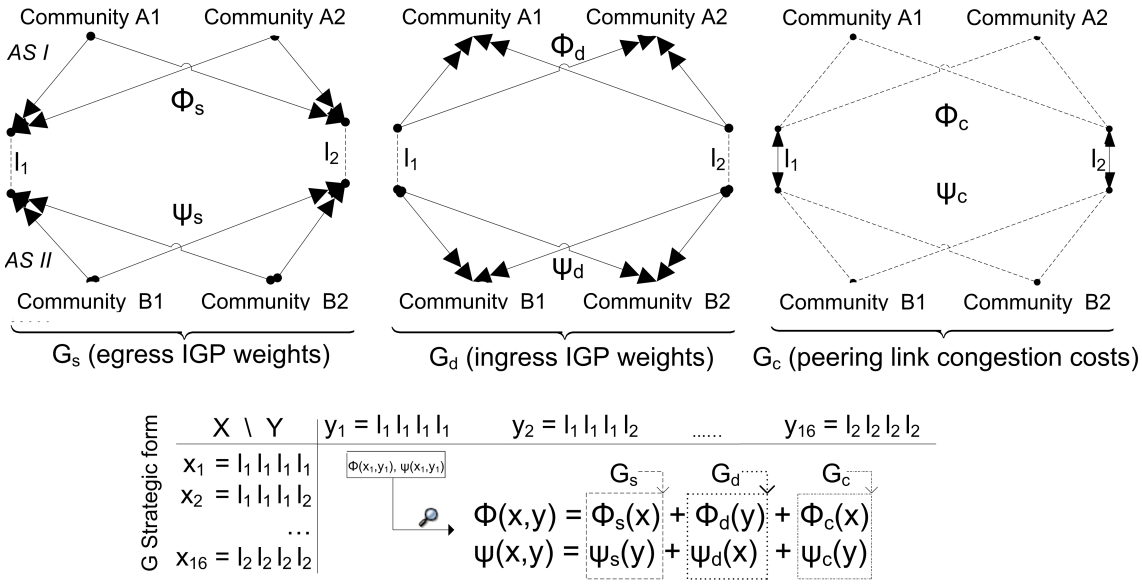


Figure 2: Exemple de composition d'un jeu de coordination avec 2 liens.

Choisir plus d'un équilibre de Nash comme solution de routage inter-domaine implique un routage multi-chemin à travers différents liens en même temps. Dans la figure suivante nous montrons, par exemple, les résultats des quatre stratégies de coordination proposées (nommées NEMP, Pareto-frontier, Unself-Jump et Pareto-Jump) comparées entre elles et avec le protocole BGP dans sa modalité multi-chemin ("BGP multipath") en terme d'utilisation maximale des liens de peering. On peut remarquer que l'utilisation peut être gardée avec sureté loin du seuil de congestion de 100%.

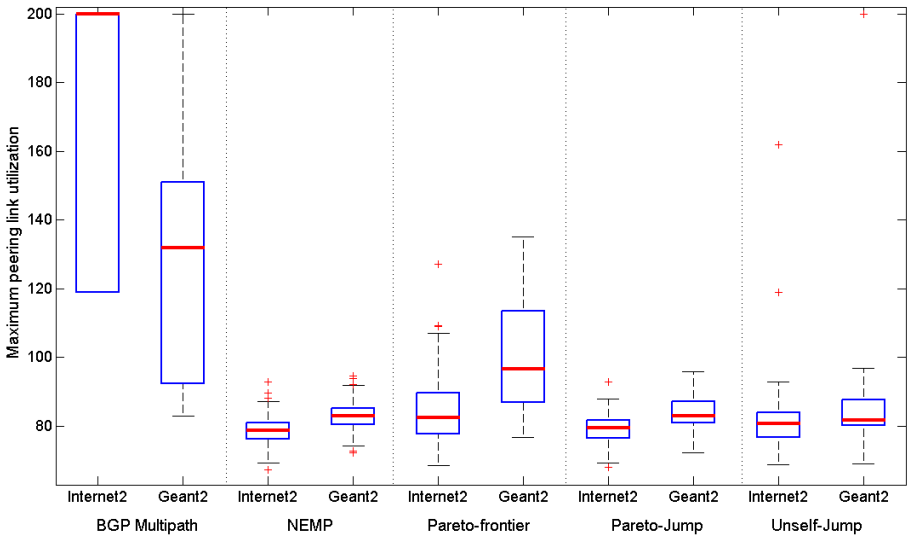


Figure 3: Utilisation maximale des liens de peering (représentation des quartiles par boxplot).

Un point fort de notre proposition est qu'une telle sélection des équilibres de Nash n'est pas lourde du point de vue calculatoire grâce à la propriété du jeu de routage d'être un jeu avec potentiel, qui implique que les minima de la fonction potentiel correspondent aux équilibres (et nous démontrons que l'inverse est aussi vrai pour notre jeu). Enfin, du point de vue de l'implémentation, notre amélioration pourrait être assez facilement déployée aujourd'hui en réutilisant un attribut du protocole BGP - le "MED (Multi-Exit Discriminator)"- qui est normalement désactivé sur les liens de peering, notamment à cause du problème de coordination susmentionné.

Le modèle mathématique présenté peut être étendu pour supporter une nouvelle forme d'accord de peering étendu à plus de deux opérateurs, et nous détaillons telle extension dans la thèse par souci de cohérence. Toutefois, nous discuterons les motivations économiques et pratiques à l'adoption du modèle de coordination étendu, en mettant en évidence qu'elles pourraient se révéler trop faibles par rapport à des solutions alternatives capables de garantir en mode strict (connecté) la qualité de service inter-opérateurs entre nœuds terminaux.

Proposition d'une nouvelle architecture pour l'extension des services de réseau à valeur ajoutée au-delà de la frontière des opérateurs Internet

Le support de fortes contraintes de qualité de service entre nœuds terminaux impose, en fait, un plus haut niveau de collaboration au niveau des accords d'interconnexion multi-

opérateurs. Il faut pouvoir réserver des ressources pour les services dans les réseaux d'autrui, et il faut qu'une telle opération soit justement et stratégiquement motivée.

Ces requis nous amènent à la définition d'un nouveau modèle d'interconnexion, "l'alliance multi-opérateurs", comme structure de coopération que les opérateurs devraient réaliser pour permettre le routage et la fourniture dynamiques de services de réseau en mode connecté. L'idée est de concevoir une architecture technologique et stratégique qui permettrait de remplir le vide (l'absence) de collaboration qui aujourd'hui rend impossible l'offre de services à valeur ajoutée au delà de la frontière des opérateurs. Au cours des dernières années, ce vide a été partiellement rempli par le développement d'applications aux couches supérieures (dites "overlay") qui, efficaces pour certains services - tel que notamment la distribution de contenu multimédia - ne suffisent pas pour des services interactifs en temps réel nécessitant de strictes garanties de qualité de service, et d'une ingénierie de trafic inter-opérateur.

Nous définissons l'architecture fonctionnelle d'un nouveau plan de service pour la gestion, à l'intérieur de l'alliance, de données spécifiques à de nouveaux services de réseau inter-opérateur, ainsi que pour leur instanciation et activation. Technologiquement, l'alliance multi-opérateurs s'appuie sur ce plan de service commun gérant les plans sous-jacents de gestion et de réseau de chaque opérateur. Ces deux derniers plans sont basés sur l'architecture protocolaire "MPLS-TE (Multi-Protocol Label Switching with Traffic Engineering extensions)"- récemment étendue en normalisation pour un fonctionnement au-delà de la frontière d'un seul opérateur (ces extensions portent le nom d'"inter-AS (Autonomous System) MPLS-TE", ou bien "inter-AS GMPLS"dans sa généralisation pour les couches basses) - et sur un réseau de serveurs de calcul de chemin appelé "PCE (Path Computation Element) architecture"qui a été également standardisé pendant cette thèse par l'"IETF (Internet Engineering Task Force)".

Dans ce contexte technologique, nous mettons en évidence les extensions protocolaires nécessaires non définies par les organismes de normalisation. Ces extensions concernent le calcul distribué du chemin (au niveau routeur) et la réservation dynamique des ressources à l'intérieur d'une alliance d'opérateurs, pour les message de signalisation et calcul entre les réseaux des différents opérateurs au sein de l'alliance. Ces extension concernent le protocole "RSVP-TE (Resource Reservation Protocol with Traffic Engineering extensions") et le protocole "PCEP (PCE communication Protocol"). Les extensions de ce protocole ont été implémentées et validées grâce à une plate-forme d'essai en collaboration avec un producteur français d'équipement de réseau et un centre de recherche espagnol.

Dans le cadre de l'alliance d'opérateurs, un nouveau problème de routage apparaît au plan de service. Le problème de routage se réfère au problème de composition d'un service de bout-en-bout en utilisant un répertoire d'éléments de service partagé par les opérateurs au sein de l'alliance. Un exemple de composition est montré dans la figure suivante où quatre éléments de service sont sélectionnés, satisfaisant certaines contraintes de qualité de service, dans la formation d'un service de bout-en-bout qui sera, dans les phases successives, configuré au niveau réseau.

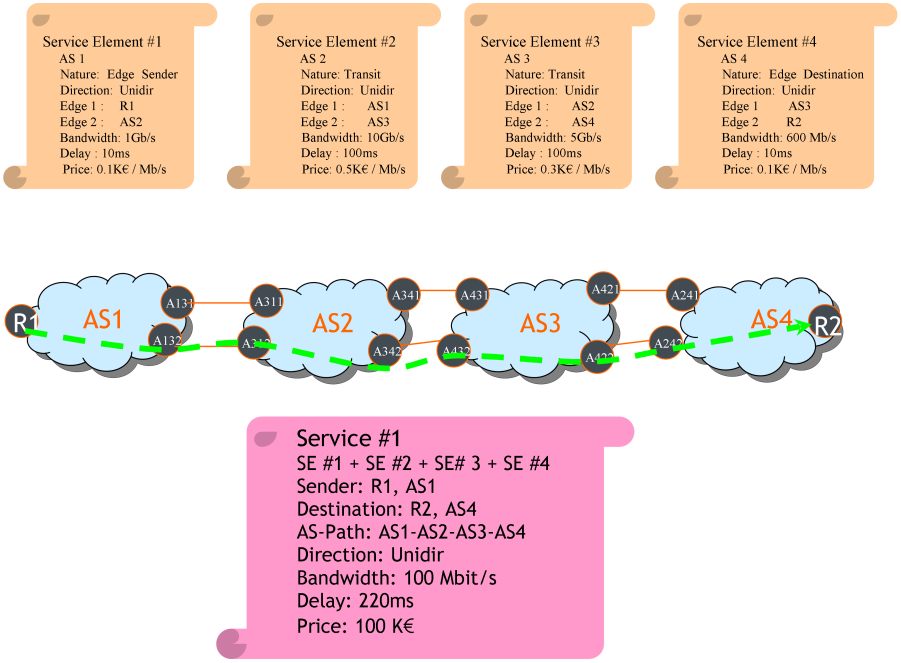


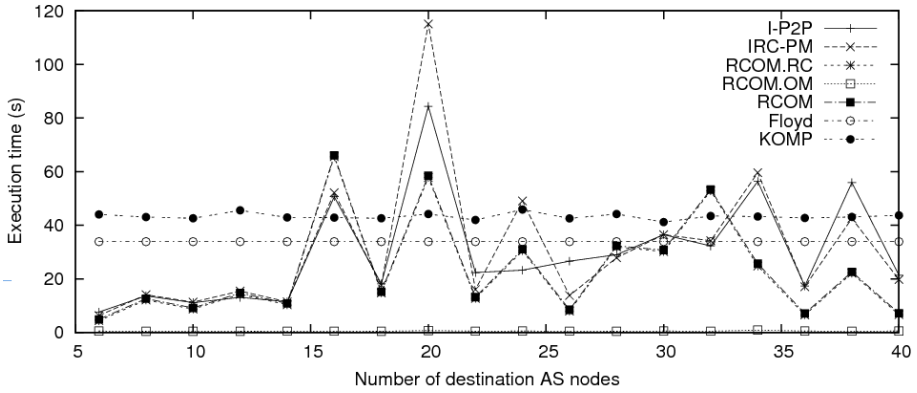
Figure 4: Routage au niveau de service comme composition d'éléments de service

Nous définissons clairement les requis du protocole de routage au plan de service dans le cadre d'une alliance d'opérateurs, et révisons l'état de l'art sur l'argument montrant que les meilleurs algorithmes dans la littérature ne satisfont pas adéquatement de tels requis. Un algorithme de routage au plan de service doit notamment:

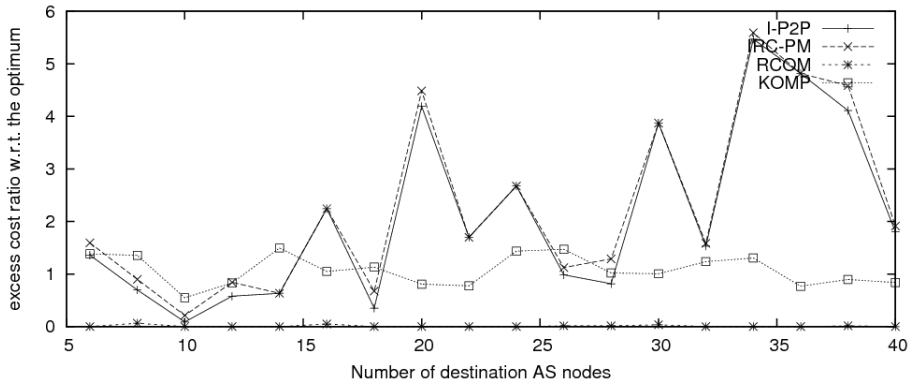
1. permettre l'application de politiques locales de routage (comme l'on fait avec le protocole BGP, par exemple, dans le mode non connecté);
2. intégrer, en passant à l'échelle, des métriques directionnelles qui caractérisent de façon abstraite le niveau de qualité de service garanti dans différentes directions à travers un réseau d'opérateur;
3. supporter la possibilité de pré-calculer une partie de la tâche - qui semble être une tache lourde même pour des topologies moyennes - en exploitant ainsi la présence des serveurs de calcul PCE;

4. intégrer à la fois le cas de calcul de routes point-à-point et le cas de routes point-à-multipoint;
5. intégrer la contrainte de diversité de chemins.

En ce sens, nous proposons un algorithme passant à l'échelle et qui est spécifique pour le routage des connections au plan de service (au niveau AS), avec support notamment des requis de pré-calcul et de métriques de transit directionnelles. Dans la figure suivante nous reportons une comparaison de performance entre notre algorithme nommé RCOM et d'autres algorithmes à l'état de l'art. En terme de complexité, le temps d'exécution nous montre que, grâce au calcul préliminaire possible d'une partie de la tâche utilisant l'algorithme nommé "Floyd"indiqué aussi dans la figure, RCOM est bien plus rapide que les autres algorithmes (KOMP, IRC-PM, I-P2P). En terme d'optimalité, RCOM peut exceptionnellement s'approcher de l'optimum.



(a) Temps d'exécution



(b) Gaps d'optimalité pour le coût de l'arbre de routage multi-chemin

Figure 5: Résultats pour une topologie de 300 nœuds

L'architecture définie pour l'alliance d'opérateurs se base donc sur une phase préliminaire de sélection des éléments de services de différents opérateurs, qui du point de vue mathématique représente un problème particulier de calcul de plus court chemin

sous contraintes multiples. Une fois les éléments sélectionnés, ils sont instanciés dans des phases successives, puis activés et enfin utilisés pour la transmission finale. Il est évident qu'en stade de l'instanciation des éléments de service il y a un risque que les ressources précédemment annoncées avec l'élément ne soient pas disponibles au moment de la requête. Pour rendre cette phase d'instanciation raisonnablement fluide, il est donc nécessaire (comme l'on fait pour les réseaux intra-opérateur dans le mode connecté) de procéder avec une réservation statique des ressources pour garantir un certain niveau de disponibilité pour les éléments de service annoncés aux autres opérateurs (autrement, en cas de refus successifs, la réputation de l'opérateur n'activant pas ses éléments serait compromise).

Dans ce sens, nous avons adapté une méthodologie définie à l'état de l'art pour l'optimisation des niveaux de réservation des liens relatifs aux connexions inter-opérateurs. Nous montrons comment les opérateurs pourraient coopérer efficacement à la réservation statique des ressources inter-opérateurs, d'une façon optimale et distribuée, en modélisant les motivations économiques et la position stratégique de chacun dans une telle coopération en appliquant des concepts de théorie des jeux coopératifs. Plus précisément, la valeur de Shapley des jeux de coalition représente une solution qui s'adapte correctement à notre contexte grâce à ses propriétés et à son habilité à modéliser les tensions entre joueurs et entre sous-coalitions. Nous proposons un modèle mathématique pour son adoption dans le cadre de l'alliance d'opérateurs. Pour justifier une réservation statique des ressources inter-opérateurs, la valeur de Shapley peut être utilisée pour partager coûts et gains relatifs à la nouvelle classe de service offerte. Son usage permet de peser correctement la contribution de chaque opérateur aux services des autres opérateurs. Par exemple, dans la figure suivante, nous montrons une distribution suivant la valeur de Shapley pour un réseau de sept opérateurs. Nous la comparons à la distribution relative au modèle actuellement utilisé pour les services en mode non connecté, c'est-à-dire tel que le gain soit collecté seulement par la source du trafic suivant un fonction de tarification approximativement proportionnelle au débit du trafic offert. Le diagramme montre que la répartition de la valeur peut amener à des grosses variations dans les imputations à chaque opérateur, ces variations représentant le montant nécessaire à motiver la collaboration entre opérateurs au sein de l'alliance.

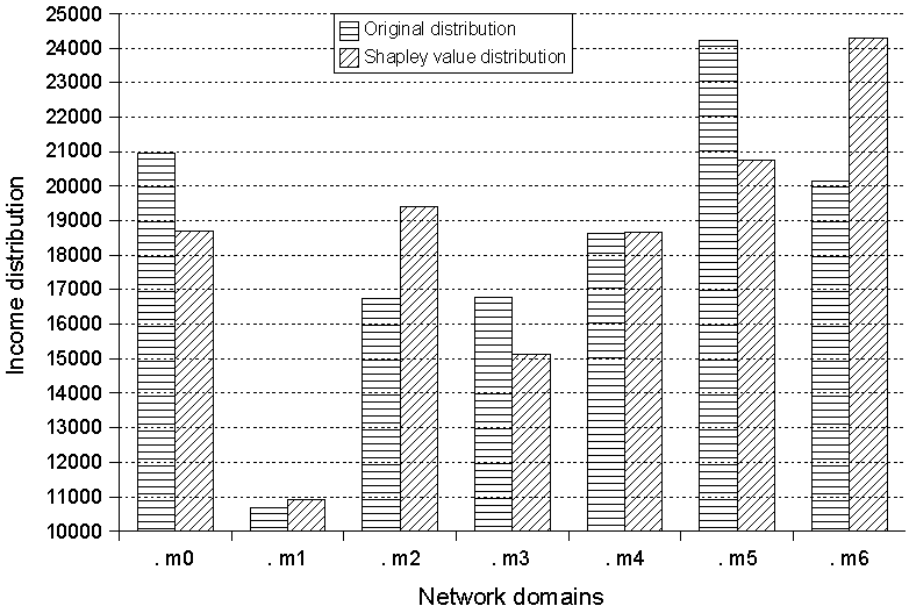


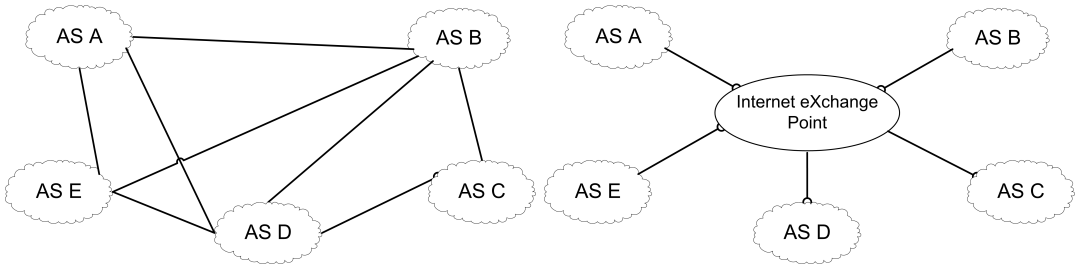
Figure 6: Comparaison de deux méthodes de distribution de la valeur à l'intérieur d'une alliance de sept opérateurs

Conception et design d'une nouvelle architecture de transport optique pour les points d'échange Internet

En perspective, nous nous attendons donc à ce que la demande de services interactifs à valeur ajoutée et, plus généralement, de services de réseau avec des contraintes strictes sur la qualité de service augmente dans les années qui viennent, pour répondre à de nouvelles applications, et aussi pour faire face à une probable baisse de la qualité du service Internet en son mode non connectée actuel..

Dans la première partie de la thèse nous montrons comment et pourquoi l'offre de services de réseau inter-opérateurs devrait passer à travers la formation d'une alliance d'opérateurs pour garantir un accord contraignant sur le respect de la relation collaborative entre opérateurs. Nous proposons une architecture adéquate de service, des algorithmes de routage adaptés et des mécanismes de motivation à la coopération.

L'établissement de tunnels et de circuits - utilisant par exemple la technologie inter-AS (G-)MPLS - à travers la frontière des opérateurs assume que la signalisation et la réservation des ressources puissent se faire à travers les infrastructures physiques d'interconnexion entre opérateurs, autrement appelées point d'échange Internet ou "IXP (Internet eXchange Point)". Comme le montre la figure suivante, un point d'échange Internet est normalement utilisé pour interconnecter plusieurs opérateurs en même temps, souvent pour établir des accords de peering vers un grand nombre de destinations, car l'alternative en maillage complet serait bien plus chère et éventuellement non praticable.



(a) Maillage complet entre opérateurs (b) Point d'échange Internet

Figure 7: Types d'interconnexion physique entre opérateurs Internet.

La transmission du trafic dans la large majorité - pour ne pas dire la presque totalité - des points d'échange Internet est aujourd'hui basée sur une simple commutation Ethernet. Ce choix de design est justifié par le faible coût des commutateurs Ethernet, par leur capacité croissante de réception et de transmission, et par le fait qu'aujourd'hui l'échange dynamique de trafic Internet se fait au niveau IP.

Dans un contexte futur où les opérateurs auront aussi le besoin d'établir dynamiquement des connexions en mode connecté à travers leurs frontières, par exemple selon les modalités présentées dans le cadre d'une alliance d'opérateurs, une infrastructure de point d'échange adéquate à la signalisation de tunnels et de circuits sera nécessaire, notamment dans le domaine optique.

Nous présentons une nouvelle architecture de transport optique à très haute capacité, appelée "Petaweb", comme une solution possible de point d'échange Internet de nouvelle génération et, plus généralement, comme possible architecture de réseau de transport à très haute capacité. Le réseau de transport Petaweb consiste en un schéma régulier d'interconnexions directes des nœuds d'accès passant par des commutateurs optiques qui ne sont pas interconnectés entre eux. Dans le cas du point d'échange, les nœuds d'accès représenteraient des routeurs de frontière d'opérateur, et les commutateurs seraient installés dans les locaux du point d'échange. Une telle structure, représentée dans la figure suivante, peut permettre une simple signalisation G-MPLS à travers les frontières d'opérateur, peut simplifier drastiquement les opérations d'ingénierie de trafic, et peut faciliter une mise à jour modulaire des éléments de réseau, aux frais de coûts d'installation potentiellement plus importants.

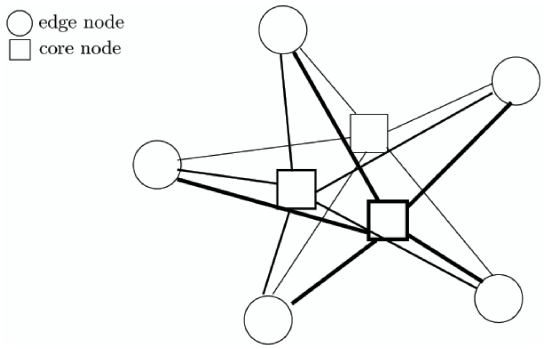
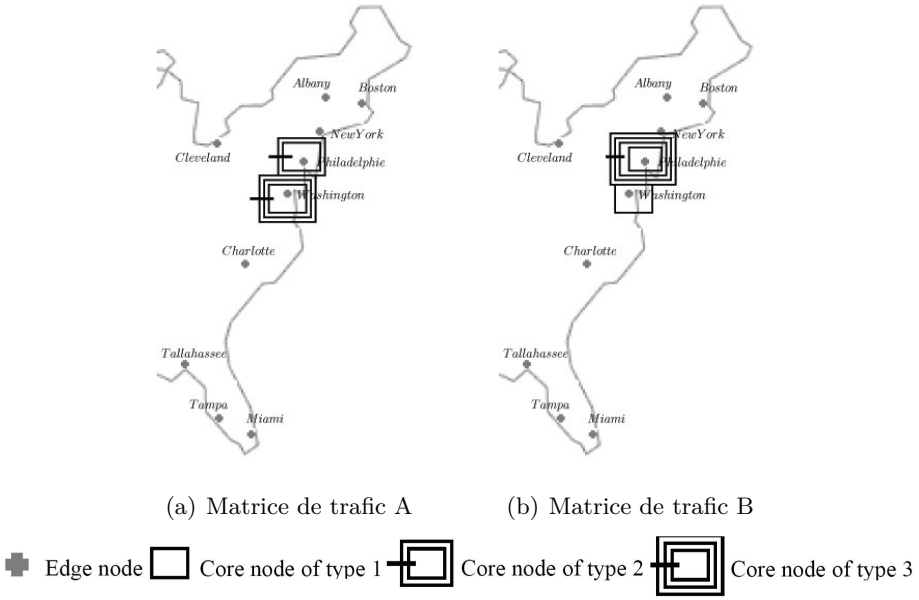


Figure 8: Architecture de transport de type Petaweb

L'architecture Petaweb a été originellement proposée par des chercheurs de Nortel Networks. Nous nous sommes intéressés - pour la première fois à l'état de l'art - au problème de dimensionnement de l'architecture Petaweb pour mieux étudier ses performances et mettre en avant des éventuelles amélioration à l'architecture initiale. Nous définissons le problème de dimensionnement de Petaweb et démontrons qu'il est NP-complet. Plus précisément, on démontre qu'il est une réduction du problème de placement d'équipements avec contraintes de capacité.

Nous énonçons la formulation mathématique d'optimisation linéaire à variables entières, que nous avons pu résoudre facilement pour des réseaux de taille moyenne. La figure suivante montre, par exemple, le résultat du placement des commutateurs optiques pour deux réseaux différents (avec deux matrices de trafic, A et B) avec 10 nœuds d'accès.

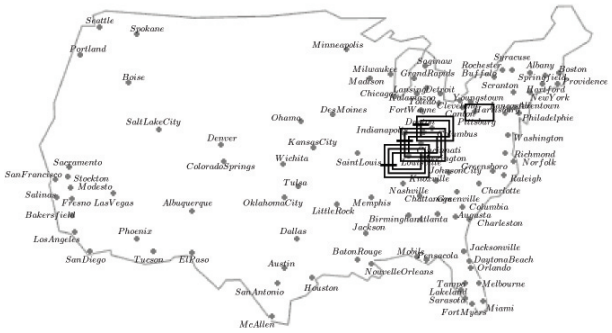


(a) Matrice de trafic A (b) Matrice de trafic B

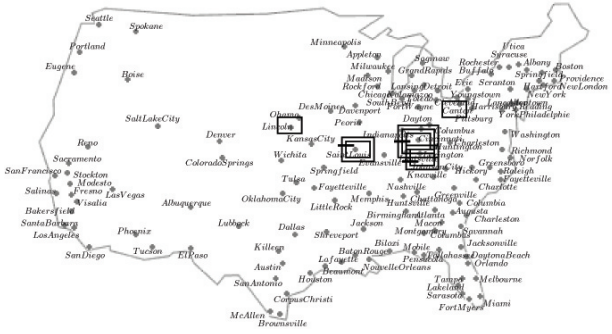
Figure 9: Réseaux de 10 nœuds dimensionnés par approche optimale (CPLEX)

Même si réalisable pour des petites et moyennes instances, l'approche optimale se révèle excessivement complexe et irréalisable pour des topologies plus grandes. Nous proposons une approche heuristique efficace et passant à l'échelle; l'heuristique est un algorithme de couplage entre les différents éléments du réseau (nœuds d'accès, nœuds de commutation, chemins optiques), répété plusieurs fois jusqu'à convergence. Les résultats montrent des seuils d'optimalité très satisfaisants.

Le modèle de coût de dimensionnement intègre des coûts d'installation et un coût de délai de propagation. L'intégration de coûts d'installation est une pratique habituelle dans le dimensionnement de réseaux de transport, tandis que le coût de délai de propagation est une nouveauté. Nous l'avons introduit dans notre contexte car, dans le cadre d'offre de services à valeur ajouté avec de strictes contraintes de qualité de service, la contrainte sur le délai deviendrait bien plus importante, surtout pour de grands réseaux. Par exemple, dans la figure suivante, nous montrons le résultat de dimensionnement pour des grands réseaux, avec un grand coût de délai de propagation. L'effet que cela peut avoir sur la topologie est donc évident, en créant des centre d'attraction près du barycentre pésé géographique du réseau.



(a) 100 nœuds d'accès

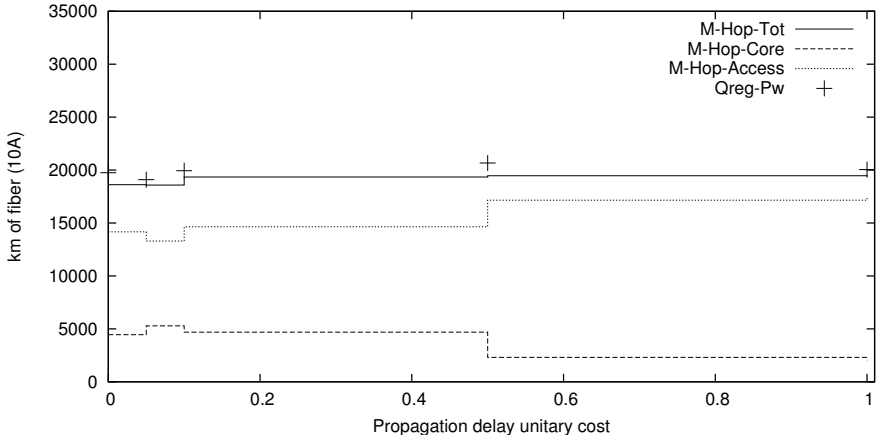


(b) 136 nœuds d'accès

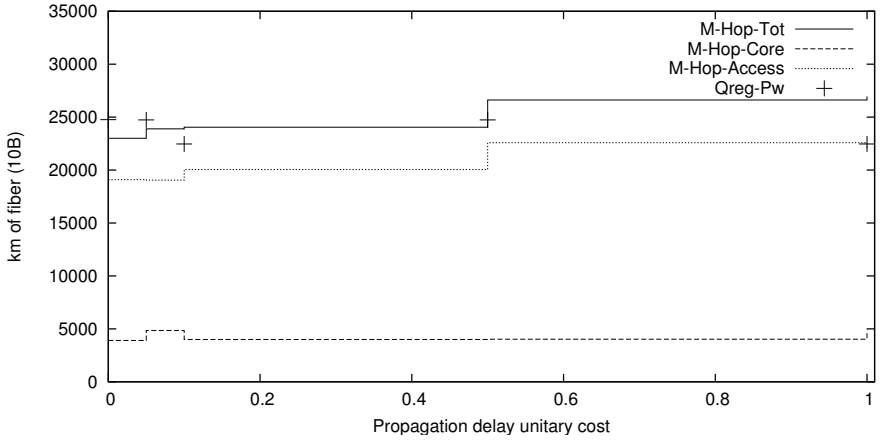
Figure 10: Réseaux de 100 et 136 nœuds dimensionnés par approche optimale avec un grand coût de délai de propagation. Matrice de trafic B.

Les résultats montrent aussi que le coût de dimensionnement total est très important, et que l'utilisation finale des liens peut être assez souvent très faible, avec une moyenne de moins de 20%. Pour éviter un tel gaspillage de ressources, nous proposons deux améliorations du modèle de réseau. Premièrement, l'allocation des ressources physiques dans les réseaux optiques peut désormais être faite en divisant dans le temps le canal de communication offert pas chaque longueur d'onde; on parle dans ces cas de "TDM (Time Division Multiplexing) over WDM (Wavelength Division Multiplexing)". Avec du TDM/WDM, des ressources physiques coûteuses peuvent être mieux utilisées; du point de vue algorithmique, le problème de dimensionnement se divise ainsi en deux sous-problèmes, un sous-problème d'allocation des ressources aux routes et aux fibres (nommé "RFA (Route and Fiber Allocation)") et un problème d'assignation des longueurs d'onde et des intervalles temporels (nommé "WTA (Wavelength and Time-slot Assignment)"). La deuxième amélioration consiste à concevoir une structure alternative quasi-régulière, moins coûteuse et légèrement plus complexe de la structure Petaweb régulière originale. La structure définie est "quasi"-régulière car la régularité de la structure originale de Petaweb est préservée et reste atteignable par simple mise-à-jour modulaire de la structure quasi-régulière.

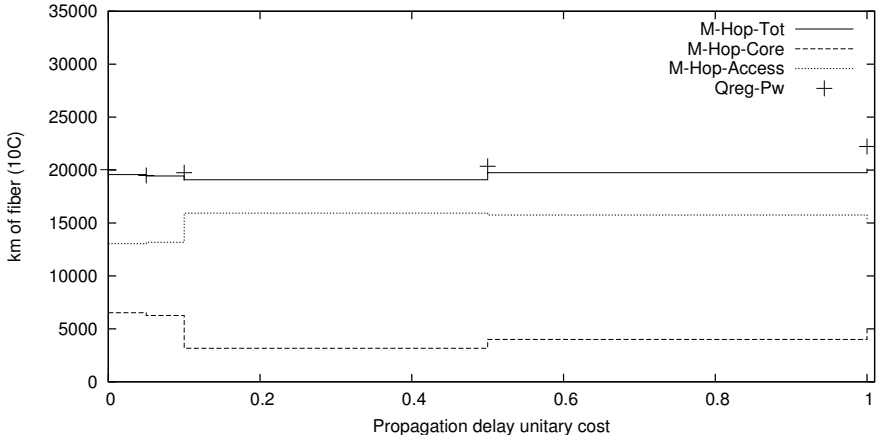
La structure quasi-régulière nécessite - à la différence de la structure régulière - de convertir la longueur d'onde dans les nœuds de commutation. Nous soutenons par simulations que le dimensionnement physique des réseaux optiques multi-saut avec minimisation de métriques de chemin additives (comme le délai) produirait une solution qui tend vers une structure de type Petaweb quasi-régulière. Comme le montre la figure suivante - pour trois cas de matrices de trafic avec différentes densités et distributions - en augmentant le coût du délai de propagation les kilomètres de fibres de cœur entre commutateurs optiques diminue, faisant tendre donc une structure classique multi-saut vers une structure mono-saut quasi-régulière.



(a) 10A



(b) 10B

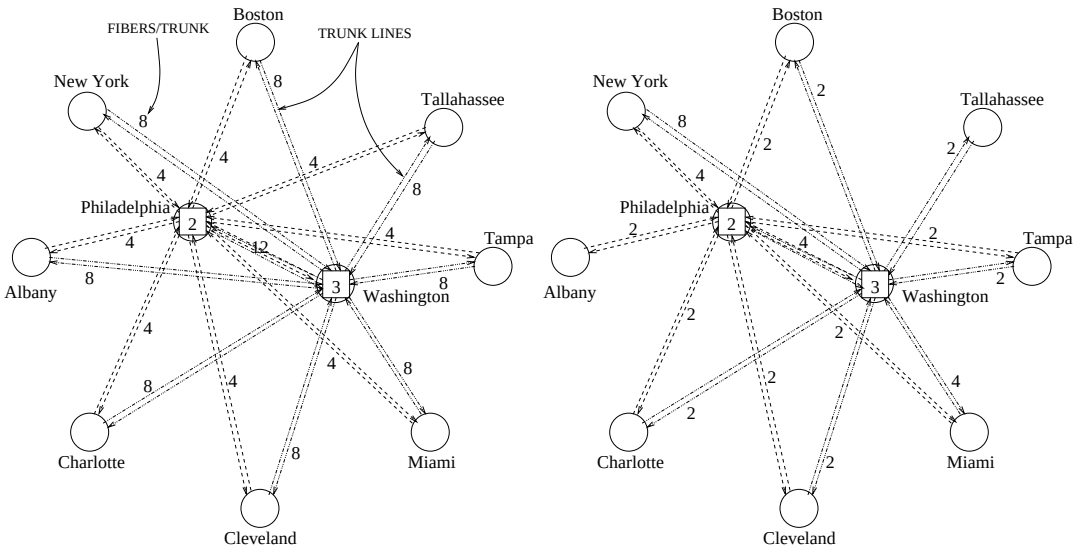


(c) 10C

Figure 11: Comparaison entre la structure multi-saut et la structure Petaweb quasi-régulière en terme de kilomètres de fibre.

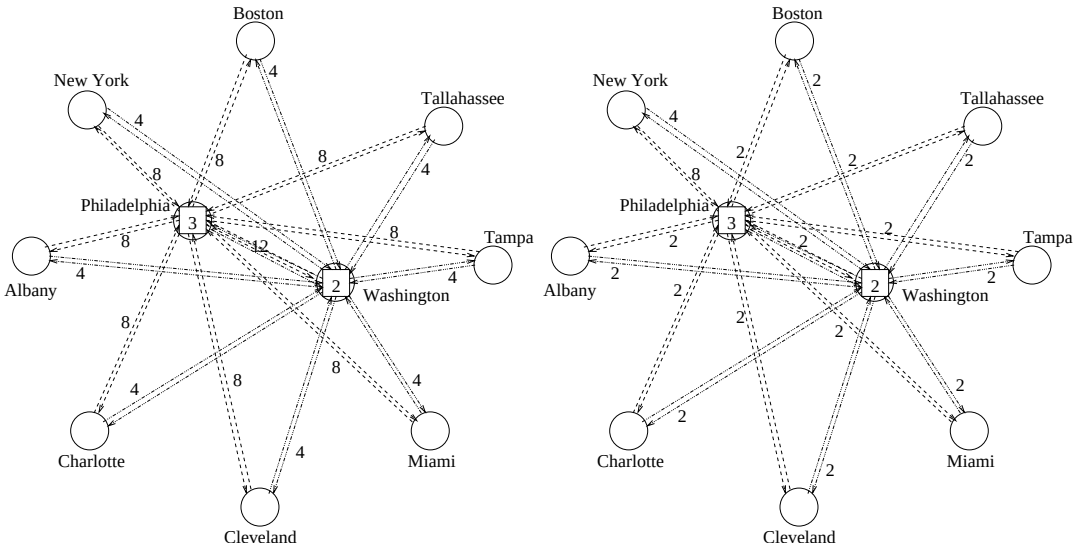
Nous définissons également le problème de dimensionnement de la structure quasi-régulière. Le problème de dimensionnement pour le cas quasi-régulier est plus complexe que pour le cas régulier. Le besoin d'une heuristique s'impose donc aussi pour des

topologies moyennes. Une simple heuristique consiste à enlever les éléments non-utilisé d'une topologie régulière optimisée. Dans les figures suivantes nous comparons le résultat d'allocation des ressources obtenu pour la structure régulière et celle quasi-régulière obtenue avec l'heuristique susmentionnée.



(a) 10A régulière (b) 10A quasi-regulière

Figure 12: Résultat de l'allocation des ressources pour le cas 10A

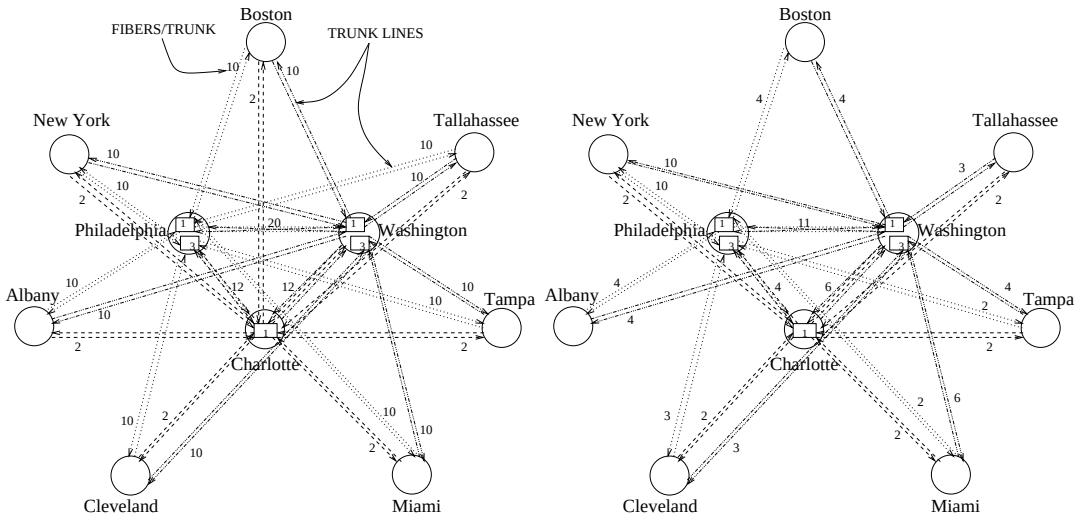


(a) 10B régulière (b) 10B quasi-regulière

Figure 13: Résultat de l'allocation des ressources pour le cas 10B

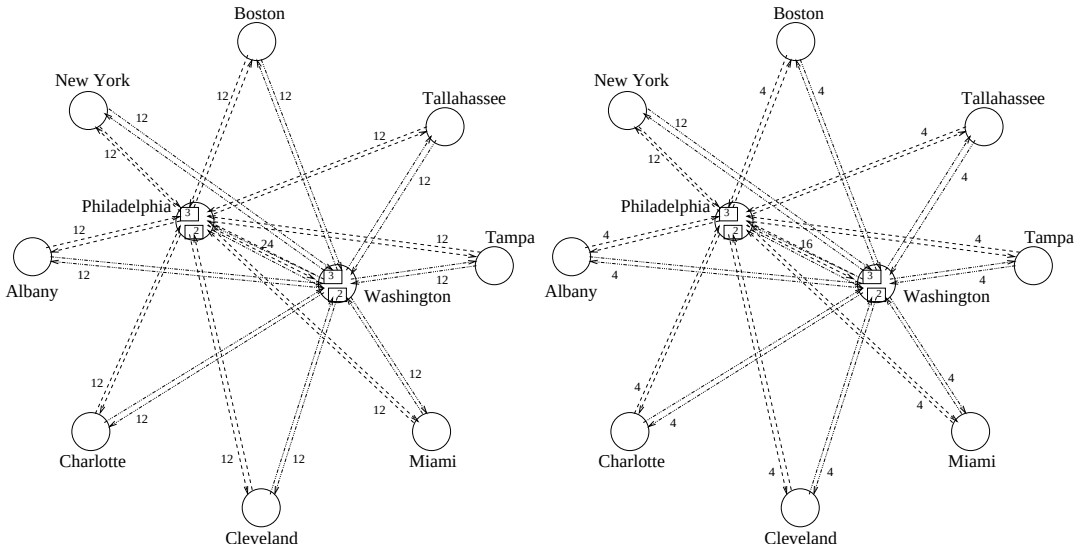
Les résultats montrent trois aspects importants avec l'adoption d'une structure quasi-régulière: l'utilisation des ressources peut pratiquement doubler, le coût total de dimensionnement pour être divisé par deux (ou plus), et la structure quasi-régulière peut

paraître peu fiable avec des nœuds isolés en cas de panne d'un seul lien. Le deuxième aspect est positif, mais le désavantage est évidemment le plus bas niveau de fiabilité de la solution. Nous avons donc intégré au modèle des contraintes de protection dédiée du chemin optique - nommées "DPP (Dedicated Path Protection"), ce qui peut garantir à chaque nœud d'accès au moins deux interconnexions vers le cœur du réseau même avec la structure quasi-régulière, comme le montre les figures suivantes.



(a) 10A regulière (b) 10A quasi-regulière

Figure 14: Résultat de l'allocation des ressources pour le cas 10A avec protection



(a) 10B regulière (b) 10B quasi-regulière

Figure 15: Résultat de l'allocation des ressources pour le cas 10B avec protection

La considération de contraintes de protection ne peut évidemment qu'augmenter le coût total du réseau par rapport à la solution sans protection. Nous avons défini un

meilleur modèle de design, et une meilleure heuristique correspondante, pour le dimensionnement direct de la structure quasi-régulière (sans passer par une régulière optimale), ce qui montre que le coût total de Petaweb à structure quasi-régulière peut être diminué d'environ 30%. La figure suivante montre le résultat de cette amélioration sur la topologie du réseau.

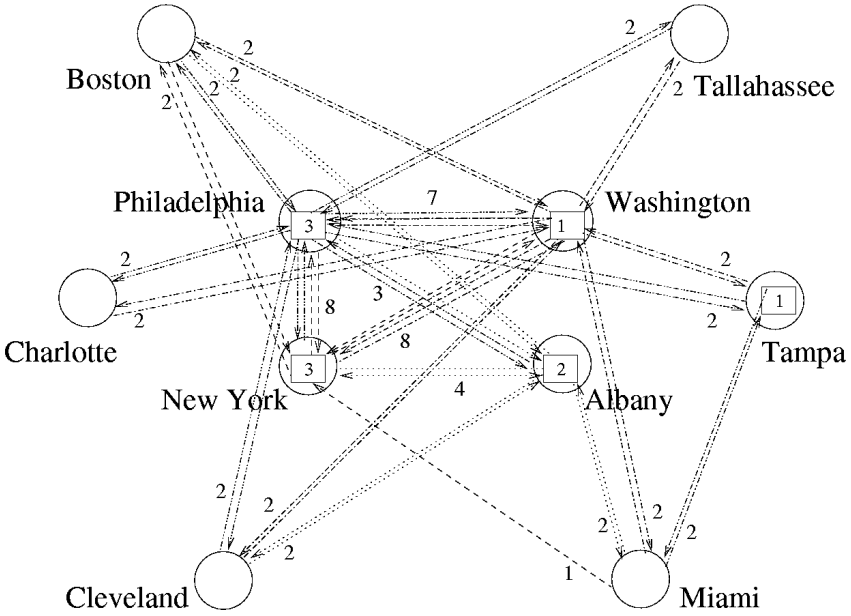


Figure 16: Résultat de l'allocation optimale des ressources pour le cas 10A en structure quasi-régulière avec protection

A la différence du problème d'allocation des ressources, le problème d'assignation des longueurs d'onde et des intervalles temporels est plus simple et peut être résolu avec un algorithme polynômial. Une fois les fibres allouées, une assignation sans blocage est toujours possible. L'adoption des contraintes de protection ne change pas la nature d'un algorithme généraliste que nous avons défini (des méthodes avancés d'assignation tenant compte, par exemple, du voisinage nécessaire entre time-slots relatifs à une même connexion, restent envisageables). La figure suivante montre un exemple d'affectation de longueurs d'onde et d'intervalles temporels à des chemins optiques entre un sous-ensemble de nœuds pour le cas 10A.

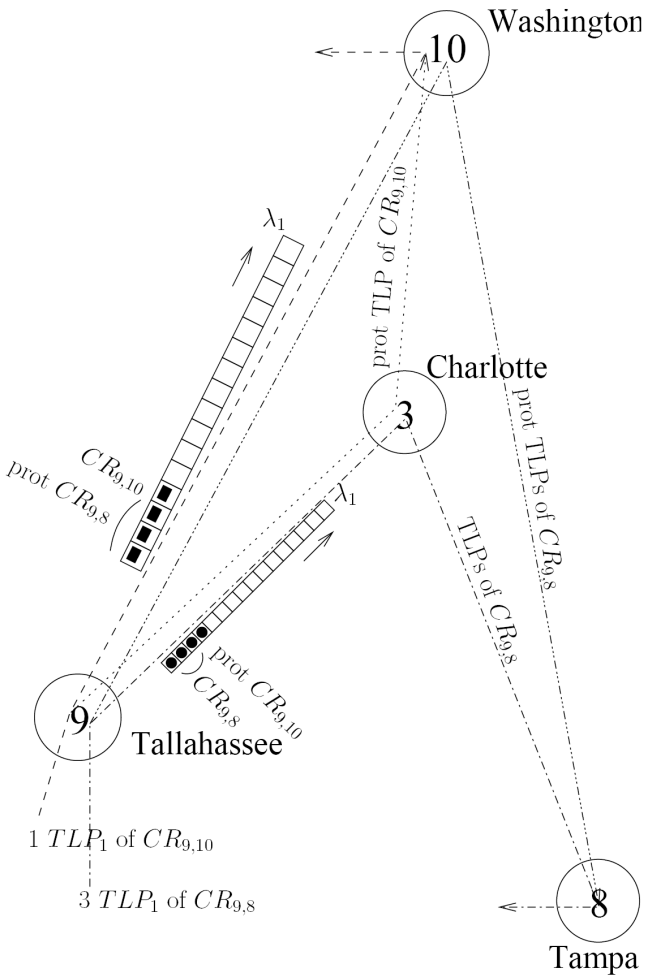


Figure 17: Exemple de résultat d'assignation des ressources pour le cas 10A.

Nous avons donc défini, pour la première fois, le problème de dimensionnement d'un réseau de transport optique de type Petaweb, proposé des approches optimales et heuristiques pour son dimensionnement, et mis en évidence et résolu certains points faibles de l'architecture. Afin de mieux caractériser la qualité de la solution offerte - donc d'une structure Petaweb quasi-régulière directement dimensionnée - nous avons étudié dans les détails ses propriétés de fiabilité, disponibilité et de capacité de rétablissement. Nous avons donc étudié le problème de mise-à-jour de l'architecture, et étudié sa réponse fonctionnelle dans différents cas de panne. Le tableau suivant montre, par exemple, les taux de rétablissement garantis pour les chemins optiques dans différentes configurations de réseau et pour différents cas de pannes; en indice on indique le taux de restauration correspondant dans le cas de fragmentation ultérieure optimale du chemin optique dans le domaine temporel et en fréquence avec retransmission partielle et pas entière. Globalement, les résultats confirment le compromis entre structure régulière et structure quasi-régulière et mettent en évidence des meilleures performances pour des

grandes topologies ayant, évidemment, plus de ressources non utilisées à disposition.

Modèle	*10A*		*10B*		*34A*		*34B*	
cas de panne	reg	q-reg	reg	q-reg	reg	q-reg	reg	q-reg
C.1: double lien								
C.1.1: entrée+sortie	$.13_{.40}$	$.7_{.19}$	$.8_{.33}$	$.3_{.15}$	$.57_{.74}$	$.28_{.59}$	$.44_{.68}$	$.25_{.52}$
C.1.2: autrement	$.25_{.49}$	$.21_{.31}$	$.22_{.41}$	$.18_{.34}$	$.65_{.81}$	$.39_{.72}$	$.66_{.9}$	$.35_{.71}$
C.2: double commutateur	$.71_{.21}$	$.46_{.15}$	$.63_{.16}$	$.34_{.08}$	$.96_{.23}$	$.74_{.13}$	$.93_{.11}$	$.51_{.14}$
C.3: double plan de commut.	$.88_{.52}$	$.53_{.24}$	$.69_{.31}$	$.39_{.11}$	$1_{.0}$	$.75_{.82}$	$.96_{.74}$	$.77_{.62}$
C.4: double site de commut.	$.05_{.10}$	$.04_{.03}$	$.00_{.00}$	$.00_{.00}$	$.31_{.34}$	$.18_{.20}$	$.25_{.11}$	$.8_{.31}$

Table 1: Niveaux de rétablissement pour différents cas de panne de plusieurs équipements de réseau

Finalement, un grand nombre de paramètres caractérisent différentes structures de type Petaweb. Dans la partie finale de la thèse, nous analysons comment le décideur planificateur de réseau pourrait - dans la discrimination entre possibles solutions alternatives de type Petaweb - balancer les différents facteurs de performance et de coût avec le niveau de fiabilité, de passage à l'échelle et de disponibilité de la solution offerte. Le problème de décision est un problème de décision multi-objectifs. Nous avons identifié dans les plans interactifs de décision, proposé par A. Lotov et al., la technique adaptée pour ce type de prise de décision. Dans la figure suivante, par exemple, nous montrons une procédure de décision basée sur la visualisation des différentes frontières de Pareto. En utilisant les logiciels appropriés, il est possible de faire varier les différents paramètres, d'identifier la configuration désirée et de choisir interactivement le point dans le plan de décision représenté qui répond le plus raisonnablement aux besoins du décideur.

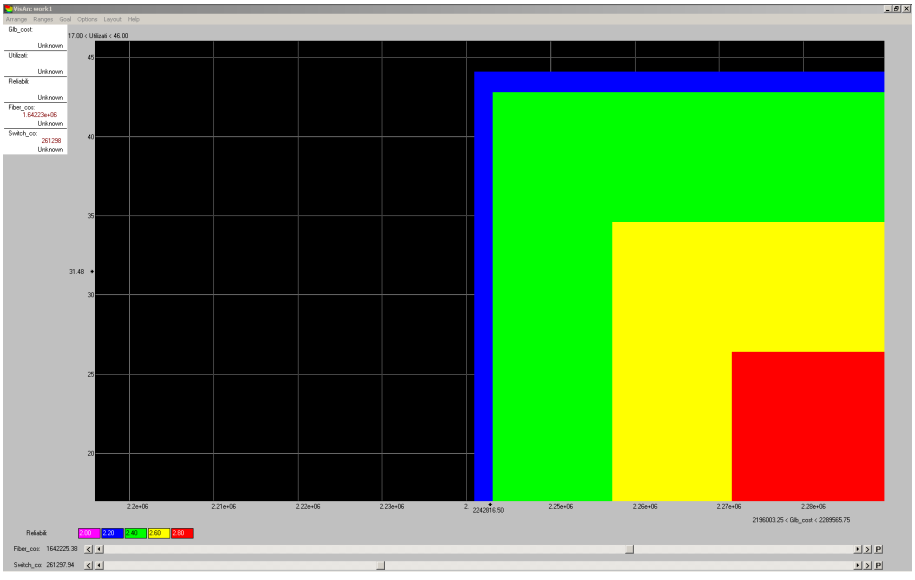


Figure 18: Plan de décision coloré avec cinq critères de décisions.

Conclusion

Dans cette thèse nous proposons différentes solutions de collaboration entre opérateurs pour l'amélioration d'Internet et pour l'offre de nouveaux services futurs. Les solutions se différencient selon un niveau croissant de collaboration, en passant de légers mécanismes de coordination pour le routage IP/BGP, par la définition d'une architecture de coopération pour des services de réseau inter-opérateurs à valeur ajoutée, jusqu'à la définition d'une nouvelle architecture de transport pour les points d'échange Internet.

Le niveau croissant de collaboration ne devrait pas être interprété comme un niveau croissant d'utopie, et donc d'impossibilité d'implémentation de la solution. A chaque niveau de collaboration, nous nous sommes inspirés d' outils adéquats pour la modélisation de stratégies complexes d'interaction entre opérateurs. Plus précisément, dans le protocole de routage coordonné - présenté en détail dans le chapitre 3 - l'égoïsme et les intérêts des opérateurs sont modélisés comme des requis impératifs à travers la théorie des jeux non-coopératifs. Les solutions données par le concept d'équilibre de Nash peuvent donner une solution bien lointaine de l'optimum bilatéral: le prix à payer est le prix de satisfaire les requis d'indépendance. Or, dans le cadre de l'architecture d'alliance d'opérateurs - décrite en détail dans les chapitres 4 et 5 - nous émulons de quelque façon la pratique courante dans les marchés financiers, avec encore une certaine compétition entre opérateurs partageant une plateforme technique commune qui permet la collaboration. De plus, le mécanisme de distribution de la valeur au sein de l'alliance - décrite en détail dans le chapitre 6 - suppose un encore plus grand niveau de coopération, tout en faisant l'hypothèse que certains opérateurs, qui bénéficieraient ponctuellement du status quo (par rapport au modèle de tarification), acceptent d'adopter un nouveau modèle pesant correctement la contribution de chacun aux services des autres opérateurs. Cette différence représente, encore, le prix qui pragmatiquement devrait être pris en considération pour modéliser l'indépendance et la tendance naturelle à faire sécession ou négocier que tout opérateur aura toujours dans un contexte d'interaction stratégique. Enfin, avec un point de vue d'encore plus longue période, l'architecture d'interconnexion Petaweb - objet des chapitres 7-10 - pourrait être implémentée comme une infrastructure physique commune de point d'échange Internet, légèrement plus coûteuse que les alternatives courantes, mais que les opérateurs devraient évaluer pour ses simplifications dans le routage et dans l'ingénierie de trafic inter-opérateurs.

Contents

List of Figures

List of Tables

Abbreviations

A2ASP: Any to Any Shortest Path
AAA: Authentication, Authorization, Accounting
ACT: Activation
ANR ACTRICE: Agence Nationale pour la Recherche - Approche Combinée de Technologies Réseaux Inter-domaine sous Contraintes Economiques
ADRENALINE: All-optical Dynamic REliable Network hAndLINg IP/Ethernet Gigabit traffic with QoS
AS: Autonomous System
ASA: AS Selection Agent
ASBR: AS Border Router
ASON: Automatically-Switched Optical Network
ATM: Asynchronous Transfer Mode
BGP: Border Gateway Protocol; IBGP/EBGP: Internal/External BGP
BRPC: Backward Recursive Path Computation
BXC: waveBand CrossConnector
CALC: Calculation
CN: Core Node
CN-r: Core Node of type r
CEPH: Convex Edgeworth-Pareto Hull
ClubMED: Coordinated Multi-Exit Discriminator
COMP: Composition
CPLP: Capacitated Plant Location Problem
CR: Connection Request
CTTC: Centre Tecnològic de Telecomunicacions de Catalunya
DiffServ: Differentiated Services
DPP: Dedicated Path Protection
DRAGON: Dynamic Resource Allocation in GMPLS Optical Networks
EGP: Exterior Gateway Protocol
EN: Edge Node
EON: European Optical Network; EONc: EON core
ERO: Explicit Routing Object
FDDI: Fiber Distributed Data Interface
FEC: Forwarding Equivalent Class
FOIRL: Fiber-Optic Inter-Repeater Link
FXC: Fiber CrossConnector
GMCP: Generalised Minimum Clique Problem

G-MPLS: Generalised Multi Protocol Label Switching
HD: High Definition
ICMP: Internet Control Message Protocol
IDM: Interactive Decision Map
IETF: Internet Engineering Task Force
IGMP: Internet Group Management Protocol
IGP: Interior Gateway Protocol
IGP-WO: IGP Weight Opimisation
ILP: Integer Linear Programming
INST: Instantiation
IntServ: Integrated Services
IP: Internet Protocol
I-P2P: Iterative Point to Point Selection
IRC-PM: Irrespective Routes Computation with Post Merging
ISO/OSI: International Organization for Standardization/Open Systems Interconnection
ITU(-T): International Telecommunication Union (Telecommunication Sector)
IXP: Internet eXchange Point
K-M: Kalai and Smorodinsky
LDP: Label Distribution Protocol
LAN: Local Area Network
LP: Linear Programming, or LightPath
LSP: Label Switched Path
LSR: Label Switching Router
MAINT: Maintenance
MG-OXC: Multi-Granularity OXC
MIP: Mixed Integer Programming
MPLS: Multi Protocol Label Switching
NA: Not Available
NMS: Network Management System
NSFNET: National Science Foundation Network
OSNR: Optical Signal to Noise Ratio
OSPF: Open Shortest Path First
OTDSS: Optical time Division Space Switch
OTN: Optical Transport Network
OTU: Optical Transport Unit
OXC: Optical Cross-Connect
P2MP: Point to MultiPoint
P2P: Point to Point
PCC: Path Computation Client
PCE: Path Computation Element
PCEP: Path Computation Element communication Protocol
PDP: Policy Decision Point, or Petaweb Design Problem
PEMP: Peering Equilibrium MultiPath
P-LP: Protected LightPath
PM: Policy Manager
PoP: Point of Presence
PTLP: Protection Ts-LigthPath
QoE: Quality of Experience

QoS: Quality of Service
RCOM: Route Collection and Optimal Matching
RECS: Route Enumeration and Clique Selection
RGM: Reasonable Goals Method
RFA: Route and Fiber Allocation
RR: Route Reflector
RSVP(-TE): Resource ReserVation Protocol (with TE extensions)
RWA: Route and Wavelength Assignment
SA: Service Agent
SC: Switched Connection
SDH: Syncronus Digital Hierarchy
SEA: Service Element Agent
SID: Service IDentifier
SIG: Service Signaling
SLS: Service Level Specifications
SONET: Synchronous Optical Network
SNMP: Simple Network Management Protocol
SOAP: Simple Object Access Protocol
SPC: Shortest Path Cost
SPLP: Simple Plant Location Problem
T1/T2/T3: Tier 1/2/3
TCP: Transmission Control Protocol
TDM: Time Division Multiplexing
TE: Traffic Engineering
TED: Traffic Engineering Database
TLP: Time-slotted LightPath, Ts-LightPath
TLP-h: ts-lightpath of class h
TWSR: Time Wavelength Space Router
VLAN: Virtual Local Area Network
VLSR: Virtual Label Switched Router
VoD: Video on Demand
WBS: WaveBand Switching
WDM: Wavelength Division Multiplexing
W-LP: Working LightPath
w.r.t.: with respect to
WTA: Wavelength and Time-slot Assignment
WTLP: Working Ts-LightPath
WXC: Wavelengths CrossConnector
XML: Extensible Markup Language

Chapter 1

Introduction

The boundaries of the Internet are nowadays blurred.

For common users the Internet may simply correspond to the terminal they use to surf or to the services provided to them (e.g., Skype, Youtube, etc). For some expert networkers and network engineers, the Internet may be defined, instead, as a physical interconnection of Autonomous Systems (ASs); for others, as a transport layer network, for some developers as a flat application layer network, etc. The common user's standpoint is undoubtedly the most relevant one that shall be considered by developers and engineers when defining new solutions to improve the Internet user experience. What really matters is the so called Quality of Experience (QoE) [85] for the user, however dependent from the Quality of Service (QoS)-enabler technology deployed at the lower layers.

Nowadays, providers can easily control the QoE by means of comprehensive traffic engineering mechanisms within their networks' boundaries. Routing protocol architectures such as Multi-Protocol Label Switching with Traffic Engineering (MPLS-TE) capabilities, Generalized MPLS (G-MPLS), allow to explicitly route connections over packet-based or circuit-based networks. This allows to reduce congestions and eventually ensure QoS for specific flows, e.g., for Video Streaming, High Definition (HD) Video-on-Demand (VoD), and other added-value services that are nowadays widely deployed in networks belonging to the same provider.

In the multi-provider scope, instead, no dynamic TE mechanisms is currently operated across provider boundaries. In order to offer some added-value services across the Internet, a current solution is to overlay QoS routing functionalities at the application layer with proprietary solutions (e.g., for real time services such as with Skype, or content delivery networks such as with Youtube). The notion of end-to-end service with QoS is therefore being defined at the application level, but even if quite efficient sometimes these mechanisms just do not work so well and can just guarantee a 'soft QoS', or a transient QoE. One may say that such a QoE provisioning is being done in a quasi-parasitic form with respect to the network, not disposing the carrier providers of adapted mechanisms of collaborative management. Current Internet overlay provisioning solutions, even if effective for some services, are limited by the lack of underlying QoS and TE, by the occurrence of unforeseen congestions and route deviations at the border between carrier providers (lack of efficient inter-provider TE), and by the (current) impossibility to

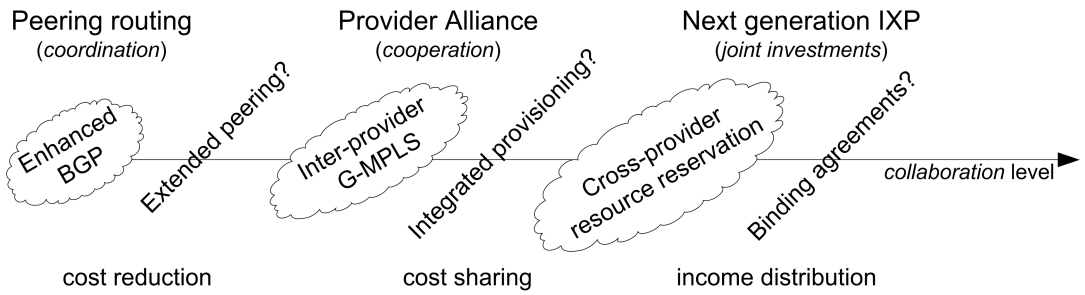


Figure 1.1: Dissertation rationales – a picture

communicate with underlay technologies.

Multi-provider collaborative technologies are unfamiliar to operators, nowadays. One of the main reasons lies in the lack of cooperation due to the structure of the Internet, where independent Autonomous Systems (ASs) interact with purely selfish routing. The aim of this dissertation (see Fig. 1.1) is the exploration of the different levels at which the collaboration among providers can take place, while modeling their independence and respective interests. We propose differently binding solutions with an increasing level of collaboration, passing from 'light' approaches to improve the current best-effort inter-domain IP routing (and generally decrease costs), toward the definition of 'strict' multi-provider connection-oriented architectures (whose costs are to be shared), and of cross-provider static resource planning frameworks (with common income distribution), finishing with the conception of joint interconnection transport infrastructures (relying on joint investments).

Structure of the dissertation

In the first part of the dissertation, Chapter 2-7, we focus on inter-provider routing issues, while in the second part, Chapter 8-10, we concentrate on physical transport issues. The third part, Appendix A-C, contains additional complementary information.

Chapter 2 is an introductory chapter that analyzes the current practice in Internet routing, arising the motivations for improvements. We highlight some pitfalls in the current Internet backbone. In particular, we emphasize the issue of route deviations and oscillations in IP routing, which we could detect using PlanetLab radar traces. The analysis focuses on the characterization of these events on top-tier provider interconnections to assess their importance for Internet QoS and reliability. We discovered that a non-negligible part of the Internet routes deflects quite frequently, and that others periodically oscillate.

In order to prevent from such issues, in **Chapter 3** we work toward the definition of 'light' multi-provider collaboration approaches. First, we review the state of the art in collaborative inter-domain traffic engineering. Then, we propose a *coordination* framework for providers interconnected via peering agreements (i.e., free transit of traffic between the respective customers only). The modeling of the bilateral free-transit routing decision passes through a particular usage of the Multi-Exit Discriminator (MED)

attribute of the Border Gateway Protocol (BGP) that we elect as transport medium of coordination routing and congestion information. The MED signaling is modeled using concepts from non-cooperative game theory, which can allow defining a technical framework, upon the strategic interaction among ASs, that produces rational routing solutions. A potential game can be built and strategically efficient strategies can be selected at a provably low complexity. We show on a realistic scenario that, besides successfully controlling the number of BGP route deviations, we can also avoid peering link congestions and significantly decrease the bilateral routing cost.

Whenever there are strict end-to-end QoS requirements, the collaboration among providers shall pass through specific interconnection agreements, thus instead of simple coordination we need a form of *cooperation*. Multi-provider binding agreements in fact define a form of "Provider Alliance", wherein costs and possibly revenues, related to added-value inter-provider network services, shall be shared. In **Chapter 4**, we propose a technical Provider Alliance framework for inter-provider MPLS/G-MPLS service management. In this connection-oriented framework, the cooperation intervenes at different layers. At the network plane, some protocols need to be extended - in fact, the Resource reSerVation Protocol with Traffic Engineering (RSVP-TE) and the Path Computation Element Protocol (PCEP) - to allow the automated provisioning of tunnels and circuits. At the management plane, inter-provider requests are filtered via policy managers and the service status is maintained. At a novel 'service plane', providers exchange service-related data, instantiate and activate new services. We define the required network elements at each plane, the inter-layer communications and the functional blocks of the provider alliance architecture.

At the service plane, a peculiar routing problem arises. At the AS-level, the network graph is weighted with directional metrics and is potentially very wide and dense so that the possibility of pre-computation - profiting from the PCE architecture - shall be considered so as to reduce the time complexity of common QoS routing algorithms. In **Chapter 5**, we first analyze the arising AS-level routing requirement, and then propose an ad-hoc routing algorithm and evaluate its performance for point-to-point and point-to-multipoint connections with respect to the state of the art. It presents an $O(n^3)$ on-line time complexity instead than more than $O(n^4)$ with classical algorithms, and better optimality trade-offs. We also discuss how at this level path diversity shall be considered, to deliver protection paths and to optimize the inter-layer communications and call acceptance.

Under the assumption of higher binding collaboration among providers, in **Chapter 6** we propose a distributed framework for static multi-provider link-reservation optimization and related income distribution. The solution is an extension of a work in the state of the art that proposes to solve the global multi-domain link reservation optimization at local domains via Lagrangean decomposition; we adapt it to make it feasible in a multi-provider strategic context. The proposition is to use the Shapley value - concept from cooperative game theory - as a power index in the distribution of the Provider Alliance income. We show that the result is fairer as with this distribution each provider's transit contribution is correctly weighted with the traffic volume it injects into the alliance.

Starting from **Chapter 7**, we deal with physical transport issues in providers' net-

works. In particular, in the light of simplifying the provisioning of MPLS/G-MPLS Label Switched Paths (LSPs) across providers' boundaries, and more generally of simplifying traffic engineering and upgrade operations in provider networks, we study a novel high-capacity optical transport network nicknamed "the Petaweb". The Petaweb infrastructure is peculiar in that it offers a direct optical path between electronic edge nodes via modular core nodes that are disconnected from each other. It may be implemented as high capacity connection-oriented multi-site Internet eXchange Point (IXP) infrastructure as well as a future-generation core network solution.

In **Chapter 8** we define the Petaweb dimensioning optimization problem. Moreover, we propose a cost-effective quasi-regular Petaweb structure and show how, under delay-constrained dimensioning optimizations, multi-hop core networks would tend toward a composite-star Petaweb-like structure. For both regular and quasi-regular structures, we formulate the optimization problem and propose heuristic dimensioning approaches, which demonstrate to be efficient in terms of time complexity and optimality also for large networks. We present the corresponding numerical results in the separate **Chapter 9** for the sake of presentation.

Finally, assuming the standpoint of a network planner decision-maker, in **Chapter 10** we discuss the trade-offs that shall be considered for the adoption of the Petaweb solution. In particular, we first analyze the reliability, availability and survivability properties of the infrastructure, present an upgrade procedure consistent with the proposed structures, and then propose to solve the multi-criteria decision-making problem to discriminate among all the possible Petaweb alternatives via the usage of Interactive Decision Maps (IDM).

Chapter 11 concludes the dissertation and contains suggestions for further work.

Appendix A presents principles of game theory, some of which are recalled across the dissertation. **Appendix B** integrates Chapters 8-10 with some additional details. **Appendix C** reports a study on the evaluation of wavebanding schemes in multi-hop optical network dimensioning.

Part I

Multi-Provider Service Architectures

Chapter 2

Current Practice in Internet Routing

In this chapter we discuss the current practice in Internet routing. First, we give a big picture of the Internet interconnection policies introducing some current stability issues. Then, we analyze Internet routing measurements trying to experimentally assess the relevance of these issues. We characterize route deviations and oscillations detected across top-tier backbone interconnections. The analysis highlights that they represent a major issue for the Internet reliability[1,2].

2.1 The Internet digital ecosystem

The current Internet interconnection infrastructure can be seen as a particular ecosystem in which providers interact via reciprocal interconnection policies to exchange Internet traffic data. With 'provider' in the following we refer to an economically independent Internet actor that provides Internet connectivity to a number of customers. A provider may in turn be customer of other providers to improve its Internet connectivity.

A provider may functionally operate via a number of 'Autonomous System (AS)' networks, where each AS can refer to a specific customer type or to a geographical location, e.g.. Each AS can, in turn, be composed of several 'domains', each representing, e.g., different regional parts of the same AS network, or parts of the network using specific routing technologies or equipment[3].

2.1.1 Current bilateral interconnection agreements

As the updated data of [193] show, since 2009 there are more than 30000 ASs. The most of these ASs are "stub"ASs, i.e., belonging to organizations or local providers at the borders of the Internet that do not offer transit to other ASs. Stub ASs are those generating and attracting a big fraction of the Internet traffic since a large part of Internet users is connected by them. Nowadays, the Internet routing architecture is based on the Border Gateway Protocol (BGP) [161], used by ASs to exchange routes toward

[1]The contents presented in this chapter are also presented in [31].

[2]The work presented in this chapter and the next one has been conducted in the framework of the Euro-NF INCAS (INter Carrier Alliance Strategies) research activity, and the Institut Telecom (Networks of the Future Lab) I-GATE (Internet - Game theoretic analysis of Traffic Engineering methods) project.

[3]In the literature, inter-AS routing is usually referred to as 'inter-domain' routing, adopting a generic meaning of 'domain'. In this dissertation, we often prefer to refer to it as 'inter-AS' or 'inter-provider' to precisely point out the scope of a routing policy. Moreover, the 'provider' and 'carrier' terms are often used interchangeably in the sequel.

destination networks (or IP prefix). In its last versions the BGP allows to implement *bilateral* interconnection agreements among ASs.

The most current inter-AS relationship is the customer/provider one that is normally settled with a "transit agreement". Moreover, an uncommercial open relationship can be settled with a "sibling agreement", in which friend ASs or ASs belonging to a same provider agree to exchange all the routes, which results in free open transit across the sibling AS's network for all Internet flows. A third common relationship, normally settled among providers with similar characteristics, relies on "peering agreements". In a peering agreement, two ASs agree for free reciprocal transit only for their respective customers' networks, because in this way they can obtain mutual economic and performance benefit (e.g., downgrading transit settlements and improving latency, respectively).

In Internetworking lingo, a provider is called "Tier-1"if it has only peering agreements and is never customer of another provider, "Tier-2"if a top-customer of Tier-1s, and "Tier-3"if it is a regional provider with a few peerings (rather with other Tier-3s) and giving access mainly to stub ASs and direct customers.

2.1.2 The BGP decision process and policy routing

The interconnection policies are eventually implemented via BGP. It is worth briefly reminding how the route selection is performed with BGP [161]. When multiple paths to a destination network are available, a cascade of criteria is employed to compare them. The first is the "local preference"through which local policies with neighbor ASs, mainly guided by economic issues, can be applied: e.g., a peering link (i.e., free transit) is preferred to a transit link (transit fees). The subsequent criteria incorporate purely operational network issues to select the best route:
(i) the route with a smaller AS hop count;
(ii) if the routes are received by the same neighbor AS, the route with a smaller MED (Multi-Exit Discriminator);
(iii) the route via the closer egress point ("hot-potato"rule), using as distance metric the egress IGP path cost;
(iv) the more recent route;
(v) the AS path learned by the router with the smaller IP ("tie-breaking"rule).

Considering these criteria, BGP selects the best route. This best route is eventually advertised to its peers, if not filtered by local policies. Indeed, the best route selection can be influenced by setting ingress and egress filters; in this way, one can modify the attributes of exchanged BGP routes (as, e.g., the local preference or the AS hop count), and also block some routes, on a per-prefix and per-neighbor basis. The result of such a policy routing are inter-AS routes that are often asymmetric and that cross a single peering settlement [122].

The Multi-Exit Discriminator (MED)

The MED is a metric that a downstreaming AS can attach to BGP route advertisements toward a potential upstreaming AS, to suggest an entry point when many exist. In this way, the upstreaming AS can prefer an entry point in the downstreaming AS toward advertised networks. By default, the MED is set to the corresponding intra-AS IGP path cost (from the downstream border router to the egress router). On transit links, subject to provider/customer agreements, the provider should always follow "MED-icated"routes suggesting preferred entry points because the customers pay for. This is not the case for peering settlements, and this can be considered as the main reason why the MED

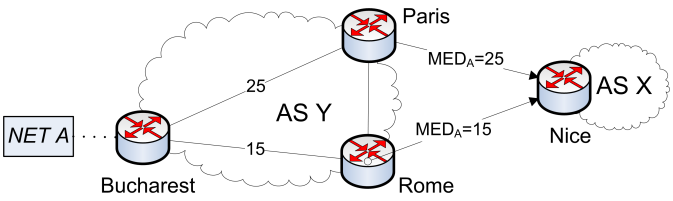


Figure 2.1: Multi-Exit Discriminator signalling example.

is often disabled between peers [165]. In Fig. 2.1, e.g., the upstream AS X selects a route for the network 'NET A'. It has two route alternatives through AS A: by the Paris router or the Rome router. MEDs are attached to the routes announced by AS A's Paris and Rome routers. If accepting MEDs, the AS X router will then select the route with smaller MED, hence the route passing by the Rome router. The default MED value is normally set equal to the IGP cost of the corresponding intra-AS path.

BGP Multipath

If the MEDs and/or the IGP path costs are equal, to avoid tie-breaking the load may be balanced on the equivalent routes. At the time being, such multipath extensions for BGP have not found consensus at the IETF, and for this reason there is no standard specification. However, some suggestions are indicated in [182]. As of our knowledge, the only implemented method carriers can use for multipath inter-AS routing is the "BGP Multipath"mode that some router vendors now provide (e.g., Juniper [137] and Cisco [138]), with some little variations on the routing decision [96]. Therefore, BGP Multipath allows adding multiple paths to the same destination in the routing table. This does not affect the best path selection: a router still designates a single best path and advertises it to its neighbors. More precisely, BGP Multipath can be used when more than one IBGP (Internal BGP) router have equivalent routes to a destination through many border routers, or when all of the candidates routes are learned via EBGP (External BGP). As stated in [182], other cases, with a combination of routes learned from IBGP and EBGP peers, should be avoided, as they may lead to routing loops for instance.

2.1.3 Routing coordination issues

On transit links, the provider is supposed to meet all the customer requirements in terms of link capacity upgrade or inter-AS routing preferences since the customer pays for. Differently, peering agreements do not rely on financial payoffs, and an AS is not motivated to follow the peer's preferences.

Top-tier carriers often 'peer' and rely on a large number of interconnection links in different locations or Points of Presence (PoP). Some AS neighbor's links may need to be upgraded to prevent from congestions, or at least a provider would like that the neighbor selects the upstream link following its preferences on the entry point. Both these requests may be acceptable if they are subject to payments, or otherwise if correctly balanced at the two sides; in particular, the second (the usage of downstream ingress routing preferences) could be implemented by using the MED attribute. However, on peering settlements they are not performed because not motivated by the free relationship.

As a matter of fact, the Internet interconnection topology is getting denser and wider, which increases the global Internet path diversity. A larger path diversity is very beneficial in that it can allow a better path selection and possibly future forms

of multipath routing at the inter-AS scope [63]. However, if not managed opportunely, in a 'best path' Internet a large path diversity risks to undermine the Internet route stability, especially across peering links where there is lack of coordination on the routing choice. This can dangerously weaken the foundations of peering agreements, which are becoming the real critical point and bottleneck of the Internet infrastructure. We are recently assisting, indeed, to an increasing occurrence of de-peerings (see, e.g., [93], at least between top-tier providers [62]), and to mutations of transit agreements into peering agreements and viceversa [88].

2.1.4 Coupling between internal and external routing

A main open issue in inter-AS routing is the coupling between Interior Gateway Protocol (IGP) routing and BGP routing. When deciding on the best route, a BGP router quite often uses the hot potato and least MED rules that depend on the least IGP path cost. A BGP route decision results in an AS path and in the corresponding egress IGP path to reach the AS border. Since in the last decade traffic engineering methods have been defined upon usage of IGP link weights, their reconfiguration can cause deviations in both IGP and BGP paths. Especially on top-tier large networks, these reconfigurations can be the result of scheduled IGP Weight optimization (IGP-WO) operations. It is worth mentioning that, besides this cause, BGP route deviations[4] can also be due to intra-AS topology changes (e.g., transient link or node failures).

BGP route deviations can be troublesome mainly because they can cause unexpected link congestions worsening the Quality of Service (QoS) in IP networks [89]. More generally, high levels of routing instability can lead to packet loss, increased network latency and time to convergence [125]. Several methods have been recently defined to anticipate, react or model this coupling and thus to prevent from deviations. In particular, in [94] the authors present an IGP-WO heuristic that assigns robust weights against possible deviations. Or, IGP graph extension tricks can be used to include inter-AS links in IGP routing [91]. In [102], the authors mathematically reformulate the egress routing problem with more expressive and efficient rules.

Persistent BGP route deviations may also be due to the joint use of the MED BGP attribute and BGP route reflectors [164]. However, ad-hoc methods have been defined to prevent from these oscillations, e.g., see [92], and moreover these events are not likely to happen among informed top-tier providers. Anyway, some routes with too frequent deviations ('flaps') can be ignored in BGP if Flap Damping Controls are enabled [162].

Therefore, BGP route deviations can cause AS path or intra-AS path changes within an unchanged AS path. In the sequel, we focus on the detection of these two types of BGP route deviations across top-tier AS interconnections. We study such deviations in order to better point out those aspects that can be improved in current Internet routing.

2.2 Detection of BGP route deviations

In the following, we aim to assess the importance of route deviations across top-tier interconnections, those that may rely on peering agreements, as an index of the dangerous lack of coordination in the Internet backbone.

[4]With the expression "BGP route deviation", we mean a change in the BGP route as a result of the BGP decision process, and not a deflection from a best chosen path inside an AS network, which is instead commonly referred to as "BGP route deflection"(see [65]).

2.2.1 Analysis methodology

Many techniques have been defined in the literature to active measure the Internet topology. In [124] there is a thorough state of the art on measurement techniques (up to 2007). The detection of route deviations needs a history of routing maps. An Internet routing map is a collection of paths from a set of monitors to a set of destination hosts. A history of routing maps can be stored sampling sequentially source-destination routes during an observation period.

In [123], the authors present a measurement framework that allows building a history of Internet routing maps. The framework stores traceroute-like samples, toward some thousands of destination hosts, collected from a few dozens of PlanetLab monitors [187]. Recorded 2008 data is now publicly available to the research community in [188]. For each PlanetLab monitor and sampling instant (round), a *tracetree* is stored as a file; a tracetree is a compact route tree from a source to many destinations that avoids some anomalies and useless ICMP signaling (replication of common paths among several destinations). We employed this data to build a history of Internet routing maps. As already mentioned, we focus on the detection of deviations across those top-tier AS interconnection (ideally peering interconnections) that are likely to suffer the most from these events. In order to select top-tier interconnections, we monitored all possible frontiers between the top-50 carrier providers in the caida ranking [194]; this can be done grouping the ASs belonging to the same provider. The total number of monitored frontiers is around 7300 (sibling frontiers were not considered).

We isolated the Radar data obtained from some monitors of different Planetlab sites. Each monitor has a random destination set of a few thousands of online IP hosts. We then extracted for each destination host the router-level and AS-level (or AS path) routes from the corresponding sources at each round. Then, we kept those crossing top-tier frontiers. If a route crosses more than one of such frontiers, we associate the route with the more ranked frontier so as to concentrate on the higher interconnection likely to represent peering settlements (indeed, it is known that each path in the Internet can cross at most one peering link [122]).

2.2.2 Intra-AS path deviations

In the sequel, we focus in those situations in which a BGP route deviation did cause an intra-AS path change, within a stable AS path. We only consider deviations occurring while the AS path remains stable during a chosen observation interval; in other words, for each stable AS path we study if there are intra-AS IP/router-level route deviations.

To detect intra-AS path deviations, for each IP-level route sample within a stable AS path, we isolate those crossing a top-50 frontier and deviating within one of the two ASs (again, if there are more than one top-50 frontier, we keep the more ranked one; if a deviating AS does not belong to a top-50 frontier, the deviation is not counted). Two kind of intra-AS deviations can be experienced:

- *internal deviation*: change of an intra-AS path with unchanged AS Border Router (ASBR). Such deviations can be caused by both BGP route deviations (different AS-level route, but same egress ASBR) and IGP route deviations or load balancing (same BGP route and egress ASBR); in the results there may be thus a bias that can not be eliminated with such measurements.

- *ASBR deviation*: change of an intra-AS path with at least one different ASBR. When the ingress ASBR changes, the deviation is due to a change of the routing

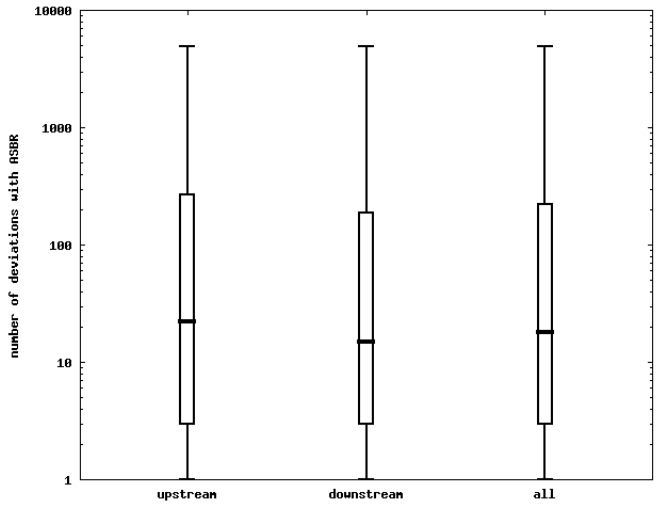


Figure 2.2: Boxplot statistics of the number of detected ASBR deviations

decision of the upstreaming AS; when the egress ASBR does, it is due to local decisions. Besides to the hot potato and least MED rules, such deviations may also be due to the usage of BGP Multipath (see Sect. 2.1.2). If the deviating ASBR is the ingress one, we count the deviation as 'upstream ASBR', otherwise as 'downstream ASBR'.

Because of the excessive risk of bias due to intra-AS IGP load balancing, we do not report internal deviation results (not enough pertinent). We focus instead on ASBR deflections that are less biased by IGP load balancing (in fact, behind the same ASBR multiple interfaces, with different IP addresses, may be detected).

Fig. 2.2 reports the boxplot statistics (minimum, first quartile, median, third quartile, maximum) for the number of ASBR deviations and for both 'upstream' and 'downstream' types, and for the both together. An AS path lifetime of 1h has been considered. We can observe that the 'upstream' deviations are more numerous than the 'downstream' ones. Moreover, the number of deviations is less than 30 for 50% of the deviating routes (i.e., the median is always minor than 30), and less than 120 for 75%.

Behind the fact that upstream deviations are slightly more numerous than downstream ones, one reason worth discussing may reside in the usage of the MED attribute of BGP across the top-tier interconnection. Across a given monitored frontier, if the MED signaling is enabled, MED-icated routes would be sent by the downstream AS to the upstreaming one to suggest an entry point for the upstream flow. In fact, a higher instability in the upstreaming AS may be a symptom of a frequent MED reconfiguration by the downstreaming AS. However, it is worth mentioning that a route change in the upstream AS could often imply an ASBR change in the downstream AS too.

Furthermore, some deviating routes may once suffer from an internal or ASBR deviation, and once an AS path deviation. This sort of deviation is likely to be related to the hot potato rule of BGP that compare all the routes (in fact, their IGP path cost) with no respect to their (downstreaming) AS neighbors. Such a deviation is not likely to be related to the least MED rule (MED=IGP transit path cost of the neighbor) since this last considers, instead, only the routes for the same (downstreaming) AS neighbor.

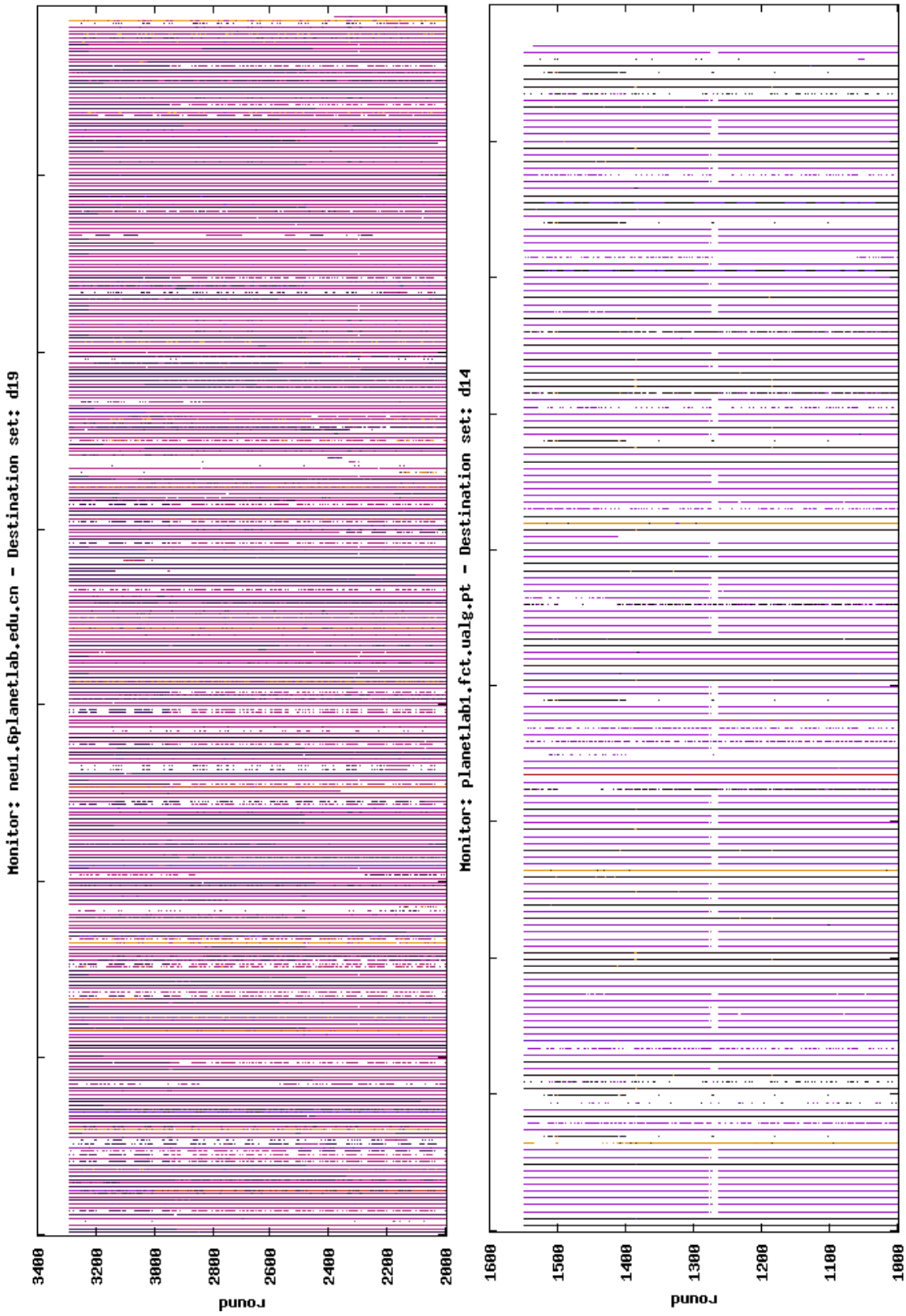


Figure 2.3: Route deviations from two PlanetLab monitors

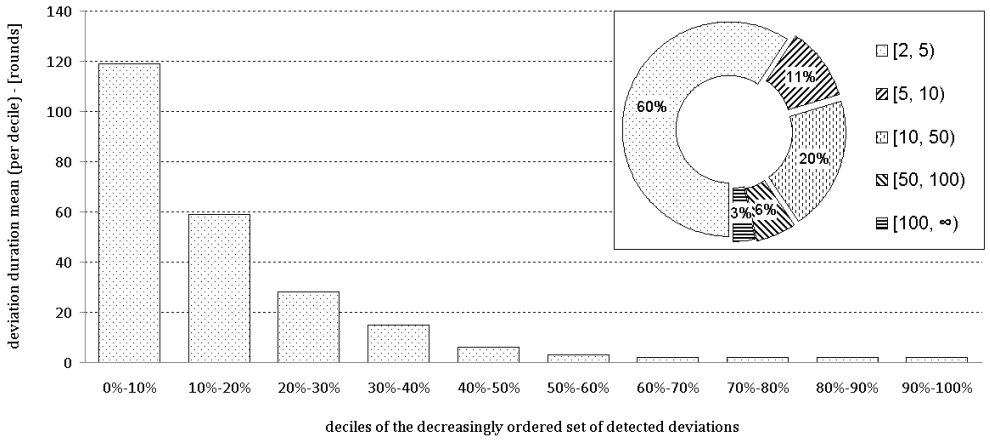


Figure 2.4: deviation duration decile distribution

2.2.3 AS path deviations and oscillations

In order to better characterize these phenomena, we monitored the deviations inducing an AS path change. Only the deviations in which a single AS in the AS path changes are considered, from an AS path to another one with the same length. In this way, we can better target those deviations likely due to IGP path cost variations rather than those due to the least AS path length rule of BGP (being the least AS path length priority to the least MED and hot potato rules). Focusing on such a 1-hop AS-level deviation we can also target those situations in which an upstreaming AS discriminates between two downstreaming AS both containing the destination host in their destination cone.

With a rapid glance, for a dozen of different PlanetLab monitor traces considered for our analysis, from 3% to 10% of the AS paths deviate, and from 1% to 3% oscillates. The whole observation period changes with the monitor and ranges from 1000 to roughly 3000 rounds; rounds are delayed of roughly 10 minutes, a tracetree can take up to 5 minutes to be stored, thus the observation period is very approximately of 10-30 days. In the following, we further characterize the detected AS path deviations and oscillations.

Deviation characterization

Fig. 2.3 gives a graphical representation of the AS path deviations detected from two sample PlanetLab monitors. The vertical axis is the time in the unit of round. At each round, a route toward each destination is recorded. In the horizontal axis we have different destination host identifiers (sequentially assigned). For each host, we have a vertical line composed of a sequence of points; a change of color corresponds to an AS path deviation. Each colored point represents the crossing of one of the top-50 frontiers, while a white point represents a crossing of unmonitored frontiers. At a glance, we can observe that the deviations can vary from quite random to more regular ones, and that they affect a small yet non negligible part of the destination hosts.

Looking deeper into this data, Fig. 2.4 represents the duration distribution of the deviations, with the average length for each decile. We also report the duration distribution across different duration intervals. We focus now only on AS path deviations within the top-50 frontier area only, i.e., only when an AS path crossing a top-50 frontier deviates toward another top-50 frontier. We can assess that:

- for the 10% longer deviations, the mean duration is 120 rounds (roughly 24h); they

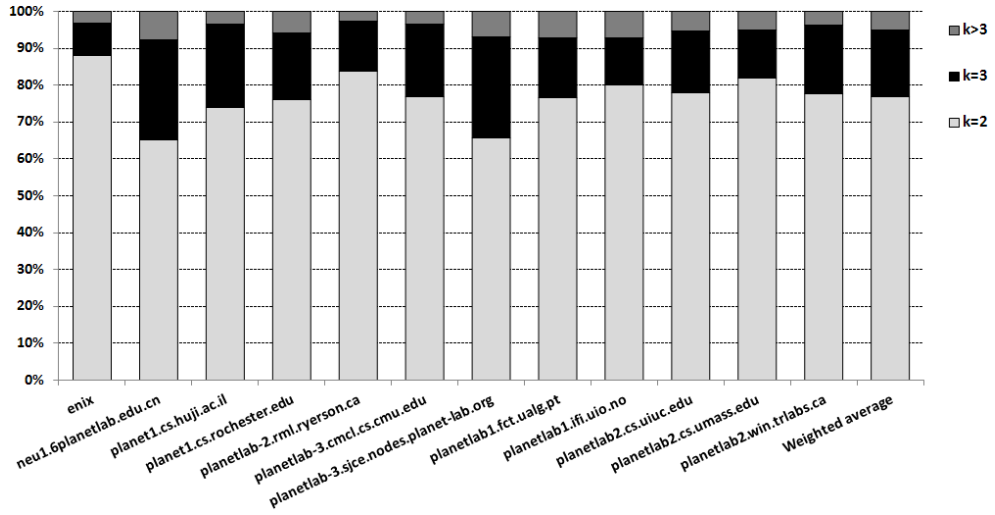


Figure 2.5: Distribution of the number of deviations per Planetlab source

represent, however, less than 3% of all the deviations;

- 59% of the deviations lasts less than 5 rounds (roughly 1h);
- 1/3 of the deviations lasts between 1h and 10h, or 1/3 more than 2h.

Therefore, the large majority of the AS path deviations manifests with a daily occurrence, which is probably linked to IGP path cost variations. Those deviations with longer duration are probably due to topology changes and are more likely to happen only once during the observation period, hence their long duration. Isolating them for some monitors in Fig. 2.5, we indicate the percentage and the number of destinations whose route deviates k times, with $k = 2$, $k = 3$ and $k > 3$ (the last column is the weighted percentage average). We can observe that there is a non negligible part of the destinations whose AS path deviates quite frequently, roughly 23% more than two times, and roughly 5% more than 3 times.

Oscillation characterization

When we observe three consecutive, not necessarily adjacent, deviations across the same frontier, we count an oscillation (or oscillating route) if the inter-deviation duration is equal with the approximation of a given error.

Fig. 2.6 reports the boxplot statistics of the number of oscillations observed per oscillating route (the last box considers all destinations). An error of two rounds is there considered. We have thus a median of 10 oscillations per oscillating route, with a first quartile of 4 and a third quartile of 40. To further characterize these oscillations, we looked into the period of each oscillating route. In Fig. 2.7, we represent a sort of periodogram in the time domain, where the vertical axis represents the number of routes that deviate with the period indicated in the horizontal axis; two lines corresponding to errors of 2 and 3 rounds are drawn. We can distinguish clear picks at durations spaced of roughly 5h, namely at 5, 10, 15, 21, 30 and 35 hours. This may reflect the different periods in daily traffic variations and related weight reoptimizations.

Nevertheless, the higher picks are within 2h, with more than 20 oscillating routes per period length. The reasons for these oscillations can be many. Some are likely to

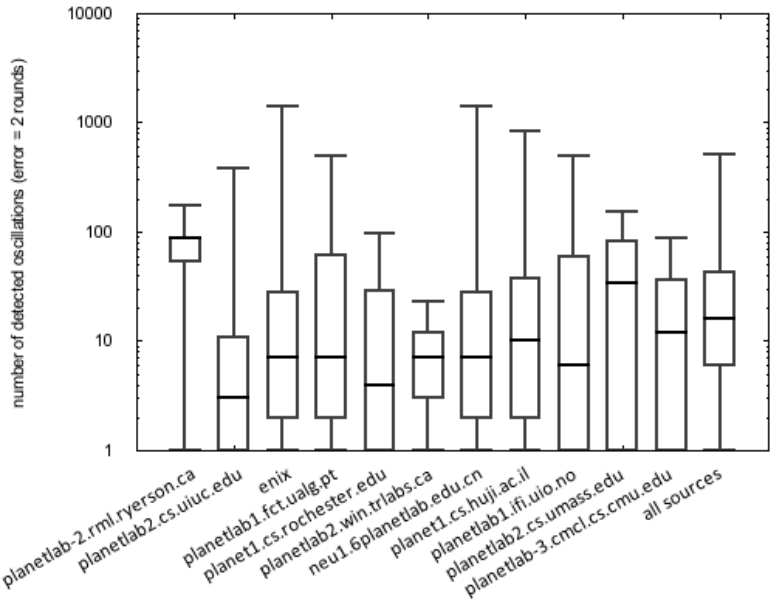


Figure 2.6: Per-destination statistics on the number of detected oscillations for 12 PlanetLab sources

reflect topology changes due to network link outages, or router off periods. A part is probably linked to critical situations in which the egress IGP path costs, toward different equivalent neighbor ASs, are very close, and a little variation of some IGP weights can trigger a deviation. Another part can be due to BGP Multipath routing (see Sect. 2.1.2). Finally, some of these oscillations may also be due to an incorrect usage of the MED attribute [164]. There is no definitive method to assign an oscillation to a cause. The important general assertion is that we have long oscillation periods probably due to IGP-WO operations, and short periods probably due also to particular BGP modes (e.g., multipath, route reflectors) or frequent IP topology changes.

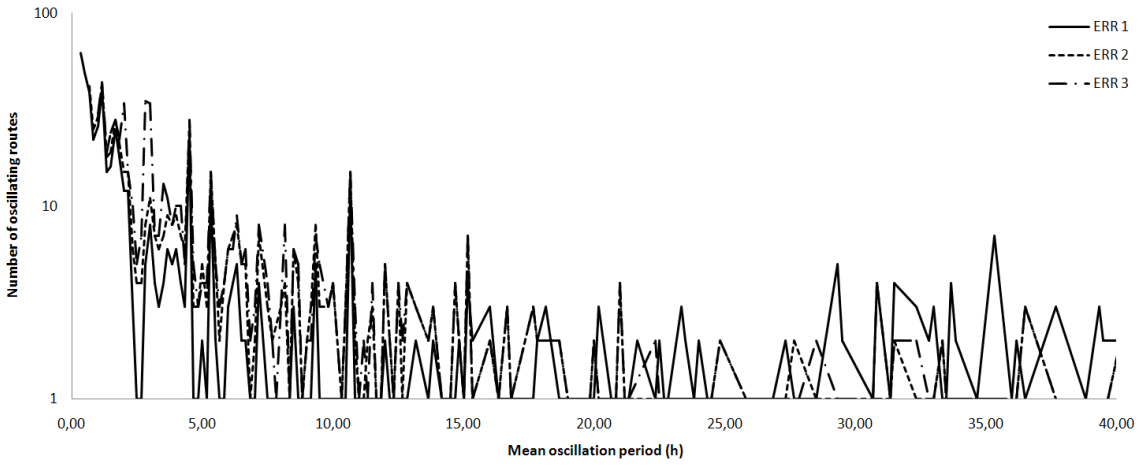


Figure 2.7: Time-domain oscillation periodogram of AS path deviations

2.3 Further work

It is worth remarking that our BGP route deviation characterization is slightly biased by BGP Multipath load balancing operations currently implemented at some carrier providers. This bias should be, however, not relevant since only a few vendors offer it as an option. Other bias, for the ASBR deviations only, could be introduced by IGP load balancing toward the same border router, and by a configuration in which two EBGP peer routers that are interconnected with multiple links and balance the load on them (see, e.g., chapter 6 in [95] for a technical insight): different IP addresses for the same router may be displayed by successive tracetrees. Indeed, the IP address that is returned with ICMP Echo Reply messages is the IP address of the router interface that is used to reply (not the interface from where the packet was received [69]).

Therefore, it is a fact that forms of BGP load balancing have introduced some bias in our characterization. We believe it is negligible for the 2008 tracetree samples, but one can not be sure on its impact. A further work [27] consists in re-sampling an Internet routing history on PlanetLab using the Paris traceroute software [68], instead than classical traceroutes or tracetrees. In fact, Paris traceroute allows detecting load balanced routes by controlling the IP packet header fields (used by hashing functions for load balancing). Preliminary results have confirmed that, indeed, BGP multipath seems practically not used nowadays. In any case, with both traceroute versions, there is still the issue that the measured path is unstable during the measurement period, so that successive probes may follow different paths [69].

Finally, some studies report that the routes towards popular destinations are stable [66], and that those AS paths along which the largest fraction of the traffic are stable too [67]. Those data are referred to 2002 and 2004 traces (respectively), periods in which the AS number was less than a half than today, and in which multi-homing and IP traffic engineering were probably much more unusual than today. In the light of recent technology advances, those analysis might need to be updated. To characterize the Internet dynamics in this sense, it would be interesting to resample Internet routing histories (with Paris traceroute) classifying the destination sets with respect to some AS ranking criterion.

2.4 Summary

In this chapter, we described the current practice in inter-carrier interconnection policy and routing. By an extensive analysis of a 2008 Internet routing history, we quantified and characterized BGP route deviations. We focused on those deviations that could be due to IGP/BGP routing interaction, and that happen across top-tier AS interconnections. This choice is guided by two main reasons. First, since top-ranked ASs dispose of a higher path diversity, across top-tier interconnections the risk of deviations due to IGP path cost minimization (hot potato or least MED BGP rules) is higher. Second, top-tier borders are likely to rely on (or to have been or to become) peering settlements (i.e., two ASs agree in free-transit between their customers' networks only); BGP routes across peering links risk to be instable because generally one AS is not bound to follow the preferences of the peer, which produces a lack of routing coordination (e.g., no MED signaling, or rare capacity upgrade), hence the risk of sudden congestions and IGP reconfigurations is higher.

To summarize, from the quantitative analysis of the routing history, we can conclude that an important ratio of backbone Internet routes deviate, and that still a non negligible part oscillate. Even if the real causes of these instabilities can not be classified without doubts, we can point out that for a large part of these events the main reason is the coupling between IGP and BGP routing that is not managed opportunely. It relies on the usage of IGP metrics and corresponding transit path IGP costs, which can be modified by IGP-WO operations and can then induce changes in the chosen BGP route. IGP path costs are used by two BGP rules (hot potato and least MED) that are independent of each other and that shall ideally be managed together to control the occurrence of BGP route deviations. These were the first motivations for the coordinated routing framework presented in the next chapter, in which we propose 'light' collaboration strategies for peering AS carriers upon a coordinated usage of the MED signaling.

Chapter 3

Coordinated Routing Framework for Peering Interconnections

We showed how current Internet routing suffers from route instability. In this chapter, we further investigate the issue and propose a robust routing framework that not only improves the route stability, but also prevents from inter-provider link congestion and strategically reduces the bilateral routing cost, which are in fact different aspects of the same issue. We rely on non-cooperative game theory concepts, whose principles are resumed in Appendix A. In the beginning of the chapter, we moreover review the state of the art on collaborative inter-domain traffic engineering[1].

3.1 Introduction

Frequent route deviations represent the major reason why nowadays peering links are becoming the main bottleneck of the Internet. Controlling congestion in such an environment is difficult, particularly due to the lack of coordination between providers, which use independent and 'selfish' routing policies.

In this chapter, we are interested in identifying possible 'light' coordination strategies that would allow carriers to better control their peering links, while preserving their independence and respective interests. We propose a robust multi-path routing coordination framework that relies on the Multi-Exit Discriminator (MED) attribute of BGP as signaling medium. Our scheme relies on a game theoretic modeling, with a non-cooperative potential game - called ClubMED (Coordinated MED) game - considering both routing and congestion costs. Within this framework, Peering Equilibrium MultiPath (PEMP) coordination policies can be implemented by means of selecting Pareto-superior Nash equilibria at each carrier - solutions which can be selected with low computational efforts by minimizing the game potential function. The selection of multiple equilibrium routing solutions for a same flow implies load balancing for some inter-provider flows across multiple links. We thus compare different PEMP policies to BGP Multipath schemes by emulating a realistic peering scenario. Our results show that the bilateral routing cost can be decreased by roughly 10%, the stability of routes can be significantly improved, and the congestion can be practically avoided on the peering links.

[1]The contents presented in this chapter are also presented in [2], [7], [9], [20], [21], [25] and [28].

3.2 Inter-provider routing issues

We aim at tackling the following major issues in inter-provider routing and peeering settlements.

3.2.1 BGP and selfish routing

The BGP decision process is summarized in Sect. 2.1.4. For peering settlements, two ASs have usually many links in several locations and can thus dispose of many routes to the same network through the same AS. By default, these routes have equal local preferences and AS hop count. Hence, the best route is chosen with respect to either the smaller MED or (if the MED is disabled) the smaller IGP path cost. The decision is taken in such cases minimizing the routing cost of a single peer: either the upstreaming AS's IGP path cost (hot-potato), or the downstreaming AS's weight (smaller MED). The challenge is thus the definition of methods that consider both the routing costs when taking the peering routing decision.

3.2.2 BGP route deviation

The peering routing decision with BGP thus relies on IGP routing costs. Nowadays, the interaction between IGP routing and inter-AS routing represents a major issue because IGP weights are optimized and reconfigured automatically. To react to non-transient network events, a carrier may re-optimize the IGP weights, inducing changes in the BGP routing decision, so that congestions might appear where not expected. This is the main cause of BGP route deviations such as those reported in the previous chapter. The challenge is thus the definition of methods to control the coupling between inter-AS and intra-AS routing, as the authors in [89] conclude after studying these interactions.

3.2.3 Peering link congestion

It should also be noted that the incentives for increasing the capacities of peering links are not straightforward. Indeed, peering agreements do not rely on any payment, as opposed to transit agreement. Controlling the load on the peering links is thus essential. However, this is difficult, as it requires setting very complex routing policies [136]. Furthermore, the current inability to estimate possible IGP weight variations, and thus to foresee the associated inter-AS route deviations they might cause, prevents carriers from controlling the inter-AS link congestion precisely. Whenever available, Multipath BGP is expected to reduce congestion, by better distributing the load over the different available routes (through the different peering links) with the same IGP costs. However, the choice of routes on which to distribute the load is based on internal costs, which might lead to inefficient traffic distribution for the peer's network. The challenge is thus the definition of scalable peering link control methods, with some collaboration among providers.

3.3 Related work on collaborative multi-domain traffic engineering and QoS

During the last years, a new research branch has been shaped around collaborative inter-domain TE methods. The goal is the definition of TE mechanisms not simply crossing multiple providers, but also relying on new forms of information exchange among providers to produce better multi-lateral (global) solutions. We do not intend covering, in

this related work review, novel egress/ingress traffic engineering methods, BGP tweaking and unilateral traffic engineering mechanisms, but instead those propositions about cross-provider collaboration in Internet routing and resource provisioning.

3.3.1 Objectives

An important goal is very well explained in [63]: the improvement of the global routing system with purely local control. It is claimed that in the short/mid term the Internet IP traffic – related to critical applications such as real-time interactive services – shall be better engineered with solutions that does not require global collaboration. The reasons for the absence of collaboration between ASs are argued in [70]. Firstly, the characteristics of the BGP policy management are rich enough to construct intricate intra-AS routing policies, but are poor when it comes to inter-AS collaboration. Moreover, the fact that ASes are not willing to disclose the details about their internal configuration and policies limits the possible degree of collaboration. In a nutshell, the independence of each Internet providers shall be sustained by every new framework that aspire being scalable and economically feasible.

The serious drawback of the lack of collaboration – in an increasingly denser, wider and less robust IP-based Internet – is the absence of routing convergence, which manifests with frequent deviations, oscillations and routing cycles, at some extent characterized in Chapter 2. Some important guidelines for collaborative methods for Internet routing can be arisen from [64] and [70]:
(i) the influence of neighboring ASs shall be limited, i.e., the coupling between internal and external routing shall be better controlled;
(ii) the routing system should have a limited reaction to minor routing changes, i.e., a form of cross-provider robust routing is needed;
(iii) any novel routing framework shall focus not all the Internet traffic but only popular or critical destination cones;
(iv) the BGP shall not be entirely substituted and any new solution shall be compliant with the principles of the BGP decision process;
(v) Internet-wide multipath routing is needed in a form that is not solely based on local unilateral decisions – in [136] unilateral multipath routing methods are reviewed and open multipath challenges are explored.

At the time being, the improvement objectives target critical applications such as real-time interactive services, currently relying on best-effort and connection-less connectivity. The guidelines above are conceived mainly for an incremental technology upgrade of the current Internet architecture, and only vaguely for future cross-provider differentiated connection-oriented solutions.

The point (iv) above derives mainly from scalability considerations. A replacement of BGP, without an incremental modification of its architecture, is not considered as feasible by networking researchers sensible to a practical implementation in the short/mid term. One shall not forget, indeed, that BGP is run on millions of routers nowadays. For the long-run, other solutions might be desirable; this is the reason why point (iv) is strongly debated nowadays, especially when possible complementary technologies are studied for differentiated connection-oriented services that require, as discussed in the sequel, a higher level of collaboration.

A few propositions of collaborative inter-AS traffic engineering methods can be counted in the literature. ‘Collaboration’ is sometimes intended as coordination and sometimes as cooperation, without an apparent agreed difference between the terms ‘coordination’ and ‘cooperation’. In our view, which follows multi-agent decision theory and

game theory lingo, a cooperation mechanism generally differs from a coordination one in that cooperation implies tools that are commonly managed or barely jointly 'bought, installed and maintained' and whose benefits are shared somehow; in short, cooperation implies binding agreements among collaborating agents. Instead coordination implies own tools implemented to increase first own benefits, then the global solution, by interacting with other actors; coordination does not imply binding agreements, and does not seek the global optimum, but an unilateral improvement, as the first objective. In the literature, instead, both the terms coordination and cooperation seem mixed and generally linked to 'negotiation' or information sharing mechanisms.

3.3.2 Coordination approaches

The idea of inter-domain route negotiation was given by [72]; the negotiation is in this case per-flow, where a flow is characterized by a number of routing options, including lifetime and bit-rate. The negotiation is reached with a flat propose/accept-reject negotiation process between neighboring ASs [73].

In [74], the inter-AS coordination is sought with bilateral route negotiations mechanisms among AS not necessarily neighbors. The outcome of a pull-based route retrieval negotiation is the establishment of cross-AS Internet flows routed using IP-in-IP encapsulation. A similar idea is developed in [75], where a 'virtual peering' among non adjacent domains, with IP-in-IP encapsulation, is proposed as a new form of Internet routing collaboration, other than classical adjacent peerings. The outcome of the virtual peering is the selection of bilaterally chosen path (across intermediate ASs) that may also encompass a form of inter-peer multipath for IP load-balancing.

3.3.3 Cooperation approaches

In another work [76], the collaboration goal corresponds to the maximization of the 'social' or global routing optimum. In order to achieve the goal, the decision process is demanded to novel 'smart routing managers', on the top of BGP yet able to configure BGP routers. The routing optimum relates to the choice of inter-AS paths that minimize routing cost and that are, moreover, QoS-feasible.

Obviously, the temptation is high to instantiate the QoS in the Internet rather than simply engineer the route selection process. How to guarantee QoS to Internet connections is an important open question. The QoS improvement may be reached by implementing load balancing and inter-AS multipath routing (see [136] and [75]), or by reserving cross-provider QoS. In the state of the art, many works have been dedicated to the BGP enhancement in order to support cross-provider QoS. Namely, the outcome of some European projects was mainly the definition of an enhanced BGP decision process, adding new QoS-dependent BGP rules [79] or managing QoS-class planes across the Internet [80] [81]. Finally, other works propose extensions to BGP for inter-AS lightpath provisioning, called Optical BGP (OBGP) [183].

Other studies show that BGP is not good for selecting path with QoS or traffic engineering constraints [84]. This is still true for OBGP, but refinements in this sense have been proposed as, e.g., [77]. In fact, another direction to provide 'strict' Internet network QoS is to rely on connection-oriented protocols such as MPLS, opportunely extended to the inter-AS scope to cope with distributed path computation and signaling (see the inter-AS MPLS and the Path Computation Element architectures and the related cooperative algorithms and protocols [169] - [181]). However, these technologies are not alternative to BGP but complementary in that the related potential services are not the

same managed with BGP (best-effort), and hence shall not be compared or inter-work. Their application (together with a review of related work) will be considered starting from Chapter 4 when treating multi-provider connection-oriented services.

Therefore, the approaches proposed in [72] - [75] consist in coordinated negotiation mechanisms. One may refer instead to the approaches [76] - [81] as cooperation mechanisms since their informed path selection relies on more information sharing and on commonly agreed (strictly binding) new routing policies.

3.3.4 Coordination or cooperation for IP services?

Game theory – see Appendix A – appears as the adapted mathematical framework to model complex and rational multi-agent interactions, so as to obtain solutions that are provably fair and stable.

In [86], e.g., the cooperative game-theoretic Shapley value concept is proposed as a rule to impute costs and profits related to Internet connections (we will se in Chapter 6 how their approach can be generalized for connection-oriented services). Another application of cooperative game theory is given in [82], where two domains bargain for a common utility sharing, yet the common utility computation being agreed by both parties. The negotiation is in this case modeled with concepts of game theory, more precisely as a 2-player bargaining problem, which indicates how binding cooperation among independent agents shall be mathematically modeled and solved (in fact, minimizing the Nash product of the unilateral utilities).

Nevertheless, for connection-less IP services, one may argue that cooperation is too much and coordination can be enough since, as in the cited work above, cooperative solutions pass through complex novel architectures that would finally present too serious management issues for finally 'just' best-effort services: interesting economical trade-offs would not be at the rendez-vous. There is a clear need to model the strategic interaction among independent providers with novel and 'light' tools.

Coordinated routing policies shall possibly be modeled with non-cooperative game theory, in which the utility of the agents are considered separately and not jointly in the computation of the strategy decision, without any binding condition on the coordination. In the rest of this chapter, we explore a possible non-cooperative game theoretic coordination solution to improve the current inter-domain routing, meeting - at least partially - the guidelines (i)-(v) indicated above.

And for differentiated services?

When the focus is not on the enhancement of current best-effort IP services, but on the definition of new solutions for added-value inter-AS services, i.e., for Internet services requiring a differentiated management (for QoS or TE) across multiple AS networks, the approach to adopt shall be different. In fact, added-value inter-AS services would be likely to represent a new source of revenues for the operators. Such network services could be needed to interconnect application providers or servers with strict QoS and reliability guarantees, rather than to interconnect common Internet users. The management of multi-provider differentiated services would thus require novel service architectures whose installation and maintenance costs could be motivated by their novel revenues. In this sense, novel *cooperative* network architectures can be economically feasible for differentiated connection-oriented services.

Approaches trying to adapt protocols defined for best-effort services (such as, e.g., OBGP [183]), even if technically feasible, could be applied only on a restricted scope

as, e.g., between multiple network domains of a same provider or friend providers. In our opinion, their extension at the large scale, between independent and selfish provider, would not be economically feasible and scalable.

At the state of the art, the technology to configure inter-AS connection-oriented services is ready (as summarized in Sect. 3.3.3). Some missing blocks are, however, needed. These aspects will be discussed in detail in the next chapters.

3.4 The ClubMED framework

We present the ClubMED (Coordinated MED) framework, in which coordinated MED signaling defines a non-cooperative peering game that allows peer ASs to coordinate towards rational, efficient and stable multipath routing solutions.

3.4.1 The ClubMED peering game

We propose to re-use the MED attribute of BGP (see Sect. 2.1.2) as the means to exchange loose routing costs and peering link congestion costs between peer networks, in the light that a coordinated MED signaling can help carriers to better collaborate in the load sharing decisions. Our scheme relies on a game theoretic modeling of the load sharing problem. Each peer is represented as a rational player that can take benefit by routing accordingly to a cost game built upon routing and congestion costs. The basic idea is to take the peering routing decision following efficient equilibrium strategy profiles of the game - in its one-shot form or repeated form - thus allowing higher collaboration.

We introduce the game on a simple example, depicted in Fig. 3.1, with two peers, AS I and AS II. Let us first define a *destination cone* as a set of customers' destination prefixes. On Fig. 3.1, Community A and Community B represent two critical destination cones that may deserve careful peer routing, e.g., because they produce high bit-rate flow aggregates. The inter-cone flows are supposed to be equivalent, for instance with respect to their bandwidth, so that their path cost can be fairly compared and their routing coordinated. We also assume that these cones represent networks that belong to direct customer ASs or stub ASs, which would often ensure that their entry point in a peer network is unique. This condition would reinforce the equivalence condition of the two flows, but is not, however, a strict requirement.

We propose that the two ASs coordinate the choice of the egress peering link for each outgoing flow, from Community A to Community B and vice-versa. A "ClubMED peering game"is built at R_a and R_b routers, called *ClubMED nodes*, using the egress IGP path cost, the ingress IGP path cost, the same costs for the peer announced via the MED, and endogenously-set peering link congestion costs. At ClubMED nodes, efficient equilibria can be selected (accordingly to the different policies presented in the next section) so as to decide the egress route(s) for each inter-community flow.

In order to take broader decisions, many pairs of inter-cone flows shall be considered in a same ClubMED game. In this way, the equivalence condition (e.g., on the bandwidth) can be extended to all the pairs together, not necessarily related to a same couple of ClubMED nodes. Therefore, the final ClubMED game derives from the superposition of many inter-community flows (e.g., in Fig. 3.2 we have 4 pairs and 8 flows). With multiple pairs of cones, carriers shall control the congestion on inter-peer links. The more egress flows are routed on a peering link, the more loaded the link and the congestion risk, and the higher the routing cost. Hence, we aim at weighting the inter-carrier links with congestion costs when congestion may arise due to the inter-peer flow routing. This could

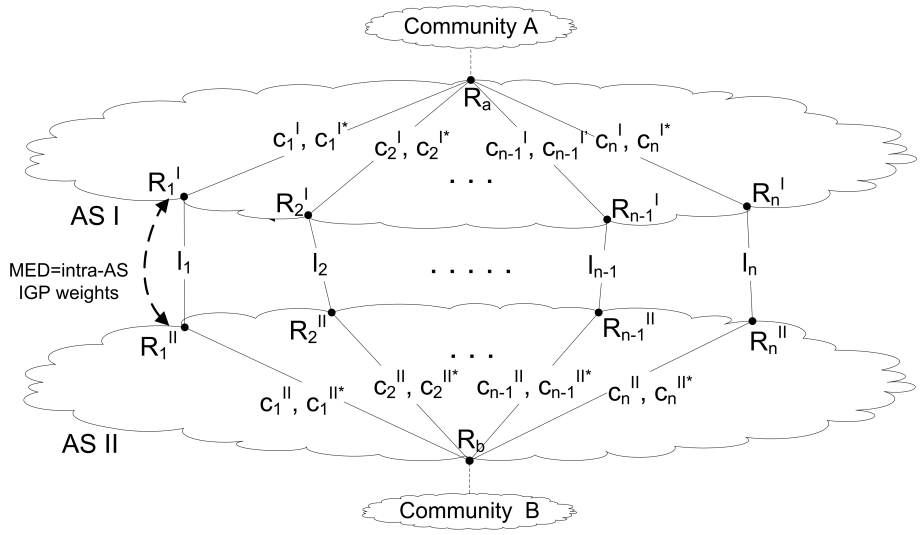


Figure 3.1: Single-pair ClubMED interaction example.

be alternatively done by modeling the inter-peer link in IGP-WO operations (e.g. [91]), but this would violate, however, the requirement of decoupling intra-AS from inter-AS routing [89].

Notations

The ClubMED game can be described as $G = G_s + G_d + G_c$, sum of a selfish game, a dummy game and a congestion game, respectively, as depicted in Fig. 3.2. Let X and Y be the set of strategies available to AS I and AS II (respectively): each strategy indicates the peering link where to route each inter-community flow. Let $(\phi(x,y), \psi(x,y))$ be the strategy cost vector for the strategy profile (x,y), $x \in X,\ y \in Y$. In Fig. 3.2, e.g., we have 4 pairs (A1↔B1,A1↔B2,A2↔B1,A2↔B2) and 2 links (l_1,l_2), and X and Y become $\{l_1l_1l_1l_1,\ l_1l_1l_1l_2,\ ...,\ l_2l_2l_2l_2\}$. For m pairs and n links, the game is the repeated permutation of m single-pair n-link games, thus with $|X|=|Y|=n^m$. G_s considers egress IGP weights only, modeling a sort of extended hot-potato rule (i.e., extended to many destination together for a same decision). G_d considers ingress IGP weights only, impacted by the other peer's routing decision (not taken into account in the legacy BGP decision process). G_c considers peering link congestion costs as explained hereafter.

Let c^I_{ji} and c^{II}_{ji} be the egress IGP weight from the j^{th} ClubMED node of AS I and AS II to the i^{th} peering link l_i, $i \in E$, $|E|=n$. Let ${c^I_{ij}}^*$ and c^{II*}_{ij} be the corresponding ingress weights, from the i^{th} link to the j^{th} ClubMED node.

$G_s = (X, Y; f_s, g_s)$, is a purely endogenous game, where $f_s,\ g_s : X \times Y \to \mathbf{N}$ are the cost functions for AS I and AS II, respectively. In particular, $f_s(x,y) = \phi_s(x)$, where $\phi_s : X \to \mathbf{N}$, and $g_s(x,y) = \psi_s(y)$, where $\psi_s : Y \to \mathbf{N}$. For the topology in Fig. 3.2, e.g., consider the profile $(\widetilde{x}, \widetilde{y})$ with $\widetilde{x} = l_1l_2l_1l_1$ and $\widetilde{y} = l_1l_1l_1l_2$; we have:
$f_s(\widetilde{x}, \widetilde{y}) = \phi_s(\widetilde{x}) = c^I_{11} + c^I_{12} + 2c^I_{21}$
$g_s(\widetilde{x}, \widetilde{y}) = \psi_s(\widetilde{y}) = 2c^{II}_{11} + c^{II}_{21} + c^{II}_{22}$.

$G_d = (X, Y; f_d, g_d)$, is a game of pure externality, where $f_d,\ g_d : X \times Y \to \mathbf{N}$, $f_d(x,y) = \phi_d(y)$ and $\phi_d : Y \to \mathbf{N}$, $g_d(x,y) = \psi_d(x)$ and $\psi_d : X \to \mathbf{N}$. For the above

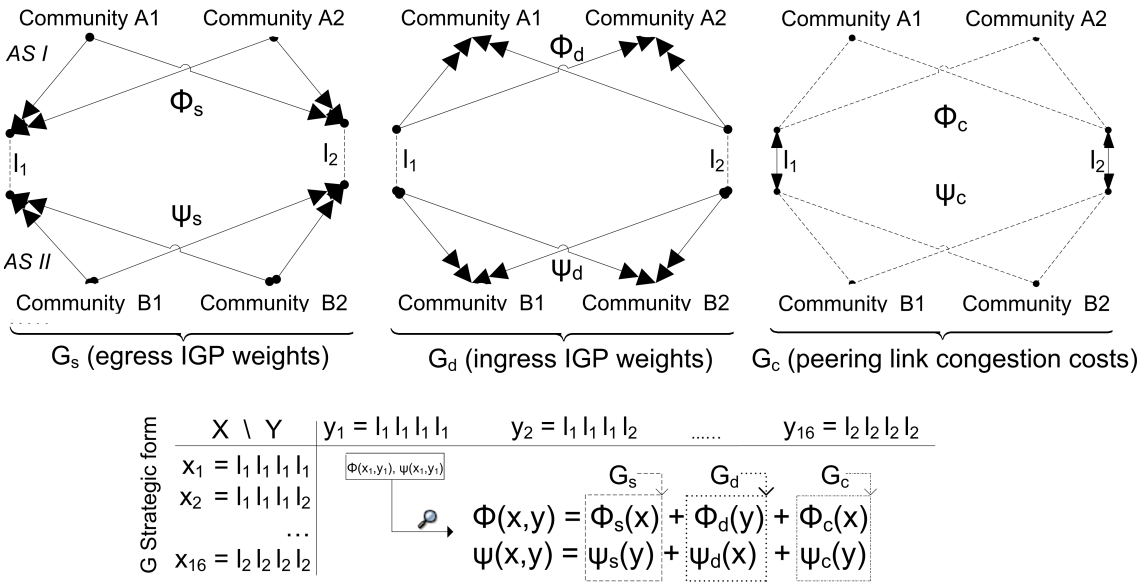


Figure 3.2: Multi-pair 2-link ClubMED game composition example.

example:
$f_d(\widetilde{x},\widetilde{y}) = \phi_d(\widetilde{y}) = 2c_{11}^{I\,*} + c_{12}^{I\,*} + c_{22}^{I\,*}$
$g_d(\widetilde{x},\widetilde{y}) = \psi_d(\widetilde{x}) = 2c_{11}^{II*} + c_{12}^{II*} + c_{21}^{II*}$.

$G_c = (X, Y; f_c, g_c)$ is an endogenous game too, where $f_c,\ g_c : X \times Y \to \mathbf{N}$. $f_c(x,y) = \phi_c(x)$ and $g_c(x,y) = \psi_c(y)$. In order to build the congestion game, the flow bit-rates have to be known. Let H be the set of inter-peer flow pairs, ρ_h the outgoing flow bitrate of the pair $h \in H$, and C_i the egress available capacity of l_i. With multipath, ρ_h can be partioned, and ρ_h^i is the fraction routed towards l_i. G_c should not count when $\sum_{h\in H} \rho_h \ll \min_{i\in E}\{C_i\}$, otherwise it would affect the G equilibrium selection. The congestion cost function should be monotone increasing with the number of flows routed on a link [103]; one can use (idem for $\psi_c(y)$):

$$\phi_c(x) = \sum_{i\in E|l_i\in x} \left\lceil K_i \frac{1}{C_i - \sum_{h\in H} \rho_h^i} \right\rceil \tag{3.1}$$

If $C_i < \sum_{h\in H} \rho_h^i$, $K_i = \infty$. Otherwise, K_i are constants to be scaled to make the cost comparable to IGP costs, e.g., such that it is 1 when the idle capacity is maximum, i.e., $K_i = C_i$.

Peering Nash equilibrium

$G_s + G_c$ is a cardinal potential game [45], i.e., the incentive to change players' strategy can be expressed in one potential function, and the difference in individual costs by an individual strategy move has the same value as the potential difference. G_d can be seen as a potential game too, but with null potential. Hence, the G potential $P : X \times Y \to \mathbf{N}$ depends on G_s and G_c only. As property of potential games [45], the P minimum corresponds to a Nash equilibrium and always exists (see Sect. A.3.5). The inverse is not necessarily true, but the following theorem proves it for G, thanks to the endogenous nature of G_s and G_c.

Proposition 3.4.1. *A ClubMED Nash equilibrium corresponds to the strategy profile with minimum potential.*

Proof. If (x^*, y^*) is an equilibrium, $P(x^*, y^*) \leq P(x, y^*)$, $\forall x \in X$. But: $P(x^*, y^*) = \phi_s(x^*) - \phi_s(x_0)$ and $P(x, y^*) = \phi_s(x) - \phi_s(x_0)$, $\forall x \in X$. Thus $P(x^*, y^*) \leq P(x, y^*)$, $\forall x \in X$, is equivalent to $\phi_s(x^*) - \phi_s(x_0) \leq \phi_s(x) - \phi_s(x_0)$, $\forall x \in X$, that is $\phi_s(x^*) \leq \phi_s(x)$, $\forall x \in X$. Hence x^* is a minimum for ϕ_s. Idem for y^*. So $P(x^*, y^*) = 0$, that is a minimum of P. □

The ClubMED peering Nash equilibrium is thus guided by the egress IGP weights and the congestion costs, and may not be unique when their sum is equal over different strategies. Moreover, the opportunity of using the minimization of the potential function to catch all the peering Nash equilibria represents a key advantage. It decreases the Nash equilibrium computation complexity, which would have been very high for instances with many links and pairs. When there are multiple equilibria (which happens in fact quite often), G_d can help in avoiding tie-breaking routing by the selection of an efficient equilibrium in the Pareto-sense.

Pareto-efficiency

A strategy profile p is *Pareto-superior* to another profile p' if a player's cost can be decreased from p to p' without increasing the other players' costs. The *Pareto-frontier* contains the *Pareto-efficient* profiles, i.e. those not Pareto-inferior to any other. In the ClubMED game, ingress costs affect the Pareto-efficiency (because of the G_d pure externality). In particular, given many Nash equilibria, the Pareto-superiority strictly depends on G_d. Fig. 3.3, e.g., depicts two cases with 3 links and their strategic forms (G_c is not considered). The exponent indicates the corresponding potential value. Egress costs are close to the egress points, and ingress costs are close to the communities. For the upper case, there is a single equilibrium, (l_2, l_2). For the lower one, there are four equilibria, and (l_3, l_1) is the single Pareto-superior one; however, it is not Pareto-efficient, but Pareto-inferior to (l_1,l_3), which is not an equilibrium because AS I will always prefer l_2 or l_3 to l_1 ($11 < 13$). This is due to the external effect of G_d. Indeed, it is possible that, after an iterated reduction of strategies, G assumes the form of a Prisoner-dilemma game, in which equilibria are Pareto-inferior to other profiles.

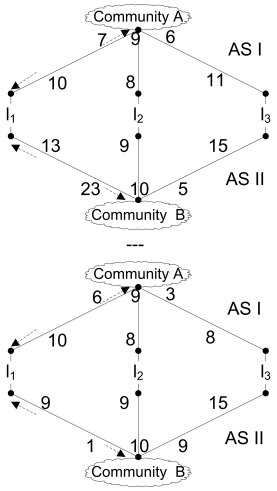
Note 1: To explicate P in calculus, we use a form in which we set to 0 the minimum of ϕ_s and ψ_s, i.e., $P_s(x_0, y_0) = 0$ where: $\phi_s(x_0) \leq \phi_s(x)\ \forall x \in X$, and $\psi_s(y_0) \leq \psi_s(y)\ \forall y \in Y$.

Note 2: In the simple example of Fig. 3.3, all the Nash equilibria have a null potential value, but this is not the case in general.

3.4.2 Modeling of IGP-WO operations

Nowadays, IGP weights are frequently optimized and these operations are often scheduled and automated. In this sense, we should assume that the ClubMED costs are subject to changes when the ingress/egress path changes. In the following we explain how, in the ClubMED framework, the coupling among IGP and BGP routing can be modeled to anticipate route deviations. We aim at selecting a robust peering equilibrium with an approach that is vaguely related to the idea presented in [105] to stabilize intra-AS routing with respect to traffic pattern variations.

At a given ClubMED node i of AS I, let $\delta_s^{i,j,I}$ and δ_s^{j,i,I^*} be the (i, j) path cost variations in the egress and ingress directions, respectively, when passing from the current routing to the routing profile $s \in X$ (idem $\delta_s^{i,j,II}$ and δ_s^{j,i,II^*} for AS II). δ variations could



I\II	l_1	l_2	l_3
l_1	(17,36)^6	(19,32)^2	(16,38)^8
l_2	(15,23)^4	(**17,19**)^0	(14,25)^6
l_3	(18,18)^7	(20,14)^3	(17,20)^9

I\II	l_1	l_2	l_3
l_1	(16,10)^2	(19,10)^2	(13,16)^8
l_2	(**14,19**)^0	(**17,19**)^0	(11,25)^6
l_3	(***14,18***)^0	(**17,18**)^0	(11,24)^6

Figure 3.3: 3-link examples.

be used to extend the G Nash set and Pareto-frontier. However, the δ should not be announced via the MED to avoid a large overhead and excessive insight in a carrier's operations. Each peer can just announce a directional path cost error. Let ϵ^I and ϵ^{II} be these egress cost errors for AS I and AS II, respectively. Being aware that IGP weights may significantly increase, an optimistic min-max computation can be:

$$\epsilon^I = \min_{(i,j)} \left\{ \max_{s \in X} \left\{ \delta_s^{i,j,I} \right\} / c_{i,j}^I \right\} \tag{3.2}$$

Similarly for ϵ^{II}, ϵ^{I^*} and ϵ^{II^*}. The ϵ cost errors represent good trade-offs between network information hiding and coordination requirement: not announcing per-link errors avoid revealing the δ variations; announcing directed errors (ingress and egress) allows reflecting the fact that upstream and downstream availability is likely to be unbalanced because of the bottleneck asymmetry in inter-AS links.

The ϵ errors induce a larger number of equilibria for the multipath routing solution. The game can be easily extended to take into account these error margins. They define *potential thresholds* under which a profile becomes an equilibrium. More precisely, the minimum potential strategies are found, then the other profiles that have a potential within the minimum plus the threshold (T_P) are considered as equilibria too. Each potential difference ΔP from (x_1, y_1) to (x_2, y_2) can be increased by $a_I(x_1, x_2) + a_{II}(y_1, y_2)$, where $a_I(x_1, x_2) = \epsilon^I(\phi_s(x_1) + \phi_s(x_2))$ and $a_{II}(y_1, y_2) = \epsilon^{II}(\psi_s(y_1) + \psi_s(y_2))$. An optimistic threshold can be:

$$T_P = \min_{x_1, x_2 \in X} \{a(x_1, x_2)\} + \min_{y_1, y_2 \in Y} \{a(y_1, y_2)\} \tag{3.3}$$

Indicating with $P(x_0, y_0)$ the potential minimum, all strategy profiles (x, y) such that $P(x, y) \leq P(x_0, y_0) + T_P$ will be considered as equilibria. This operation can also escape selfish (endogenous) solutions mainly guided by $G_s + G_c$, introducing Pareto-superior profiles in the Nash set.

3.5 Peering Equilibrium MultiPath (PEMP)

Within the ClubMED framework, peers would route accordingly to an equilibrium because it grants a rational stability to the routing decision. The Nash set and the Pareto-frontier may be quite broad, especially considering the IGP path cost errors. This leads to

different possible Peering Equilibrium MultiPath (PEMP) load balancing policies (upon these sets of profiles), which are presented below.

3.5.1 Implicit coordination

Assuming thus that ClubMED remains a fully non-cooperative framework, its implicit solution policy to which to coordinate without any signaling message is: *play the equilibria of the Nash set, and only the Pareto-superior ones if there is at least one.* Hence, it is feasible to natively implement a Nash Equilibrium MultiPath (NEMP) routing policy. When in the Nash set no Pareto-superior equilibria exist (as already mentioned, this can happen), NEMP is performed over all the equilibria. E.g, in the bottom of Fig. 3.3 AS I may balance the load on l_2 and l_3, being aware that AS II may balance its load on l_1 and l_2.

3.5.2 Repeated coordination

Given that the the G Pareto-frontier may not contain equilibria, in a repeated ClubMED context, an explicit coordination policy is: *play the profiles of the Pareto-frontier.* The ClubMED game would be repeated an indefinite number of times, indeed. From "folk-theorem"-like results [40], this policy is an equilibrium of the repeated game and grants a maximum gain for the players in the long-run. Nevertheless, the unilateral trust for such a strategy could decrease whether in a short period of analysis the gains reveal to be unbalanced and in favor of a single peer. The reciprocal trust among peers can thus affect the reliability of such a Pareto coordination.

Unselfish-Jump

Another policy is conceivable to guarantee a state of balance in gains in the short term, and thus helping to keep a high level of reciprocal trust. After shrinking the Nash set with respect to the Pareto-efficiency, for each equilibrium the ASs might agree to make both a further step towards the best available strategy profile (x^j, y^j) such that:

$$\psi(x^j, y^j) - \psi(x_0, y_0) + \phi(x^j, y^j) - \phi(x_0, y_0) < 0 \tag{3.4}$$

where (x_0,y_0) is the starting equilibrium. One AS may unselfishly sacrifice for a better bilateral solution: the loss that one may have moving from the selected equilibrium is compensated by the improvement upon the other AS. This policy makes sense only if the other AS is compensated with a bigger improvement, and returns the favor in the future.

Pareto-Jump

Instead, with the addition of the constraint:

$$\psi(x^j, y^j) - \psi(x_0, y_0) \leq 0 \ \wedge \ \phi(x^j, y^j) - \phi(x_0, y_0) \leq 0 \tag{3.5}$$

we select a Pareto-superior profile (not necessarily in the Pareto-frontier), without unselfish sacrifices. If at least one (x^j, y^j) is found we obtain a new profile set that is to be shrinked with respect to the Pareto-superiority for the final solution.

In the bottom example of Fig. 3.3, e.g., we would jump from the Pareto-superior Nash equilibrium (l_3, l_1) to the Pareto-superior profile (l_1, l_3). We would not have this

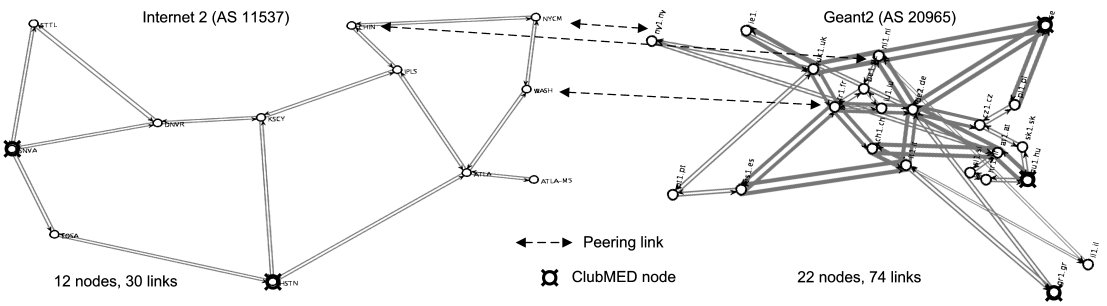


Figure 3.4: Internet2 - Geant2 peering scenario with 3 peering links.

jump for the Unselfish-Jump policy, that would prefer instead (l_1, l_1) with a global gain of 6 instead of "just"3 with (l_1, l_3).

Finally, note the last two policies (Unselfish-Jump and Pareto-Jump) are not binding: the implicit threat to change to one of the more selfish choices is enough.

Note 3: With the Jump policies, we assume that MEDs from different ASs are normalized to the same IGP weight scale in order to be comparable.

Note 4: We have a decreasing level of collaboration (thus trust), starting with a high level for 'Pareto-frontier', lower for 'Unselfish-Jump', still lower for 'Pareto-Jump' and basic coordination with 'NEMP'.

3.6 Performance evaluation

We evaluated the performance of the three PEMP routing policies with realistic simulations. We created a virtual interconnection scenario among the Geant2 and the Internet2 ASs, depicted in Fig. 3.4, emulating their existing peering with $n = 3$ cross-atlantic links. We considered $m = 6$ pairs of inter-cone flows among the routers depicted with crossed circles. The TOTEM toolbox [106] was used to run a IGP-WO heuristic, with a maximum IGP weight of 50 for both ASs. We used 252 successive traffic samples, oversampling the datasets from [107] for Geant2 and from [192] for Internet2 on a 8h basis (to cover all the day times). The original link capacity was scaled by 10 to create an intra-AS congestion risk. The inter-cone routing generates additional volume for the traffic matrices; we used a random inter-cone traffic matrix such that flows are balanced with 200 Mb/s per direction, which corresponds to 2/3 of the total available peering capacity. To evaluate the effectiveness of the congestion game we considered peering links with 100 Mb/s available per direction.

We compare the PEMP routing policies ('NEMP', 'Pareto-Frontier', 'Pareto-Jump', 'Unself-Jump') to the 'BGP Multipath' solution without and with ('...+MED') classical MED signaling enabled at both sides, and to a 'Full BGP Multipath' solution in which all the peering links (i.e., the available routes) are used for the multipath solution.

3.6.1 Routing cost

Fig. 3.5 reports the IGP routing costs statistics in BoxPlot format (minimum; box with lower quartile, median, upper quartile; maximum; outliers). We show four solutions: Full BGP Multipath; BGP Multipath as described in Sect. 2.1.2; the first PEMP policy, NEMP, without and with the congestion game G_c. For each method, we display the

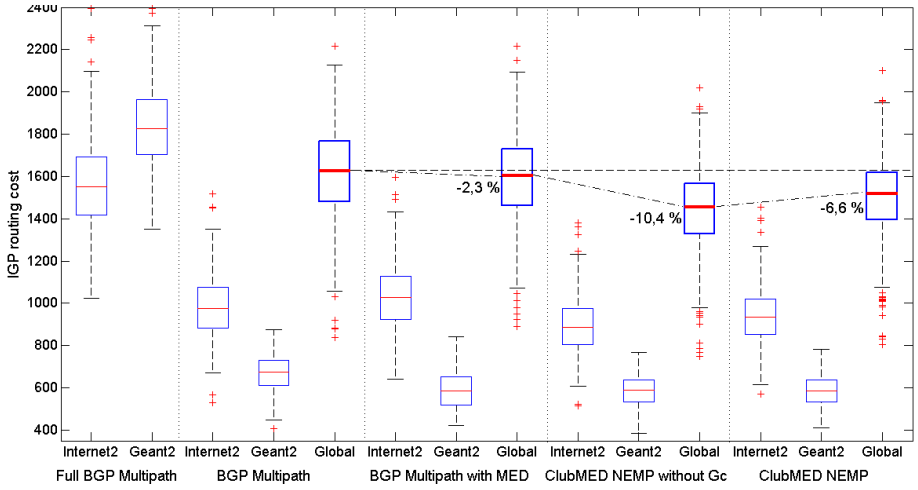


Figure 3.5: IGP routing cost Boxplot statistics: NEMP vs BGP Multipath.

Internet2, the Geant2 and the global IGP routing costs. We considered two ClubMED solutions, with and without the congestion game G_c (for the first two figures only).

The full BGP multipath solution obviously guarantees an even load on all the peering links. However, its routing cost almost doubles compared with normal BGP multipath, which balances the load only on equal cost paths (egress IGPs or MEDs). Curiously, the simple usage of the MED would decrease by 2% the cost of the BGP case without MED. This is probably due to the fact that for the most utilized network (Internet2) the ingress paths are more loaded than the egress one (hence, with higher ingress IGP weights), which leads to a lower global IGP cost. Another reason may be that with the MED enabled the chance of doing Equal Cost Multi-Path (ECMP) is higher: not only on equal IGP path cost routes, but also on equal MED routes. The ClubMED solution, instead, outperforms BGP with a median cost 10% lower without G_c, and 6,6% in its complete form.

Fig. 3.6 compares the four PEMP policies. With respect to NEMP, the Pareto policies give statistically very close results. This may sound disappointing: one may expect more from the Pareto-frontier and the Pareto-Jump policies. By analyzing the results in detail, we verified that the reason for this poor performance is that the Pareto-frontier often contains strategy profiles with the least cost for one peer and very high cost for the other peer. Such strategy profiles are not marked as Pareto-inferior because of the single peer's least cost and thus belong to the Pareto-frontier. Such situations are likely to be frequent since an uncongested intra-AS link may produce a IGP weight much lower than the others thus affecting the G profile cost components. And this risk is augmented in the Pareto-Jump policies since the new selected profiles can 'just' be Pareto-superior: they do not necessarily belong to the Pareto-frontier. However, for the Pareto-jump policy the median, the minimum and the upper and lower quartiles outperform the NEMP result; in fact, the starting Nash set for its Pareto-improvement is the NEMP one (see Sect. 3.5.2). Moreover, the Unselfish-Jump one is expected to outperform or equalise the Pareto-Jump strategy with respect to the routing cost since, without (3.5), it can be see as its relaxation. Indeed, as reported in Fig. 3.6, the Unselfish-Jump gives a median cost roughly 3% inferior to the NEMP cost and 1% inferior to the NEMP cost.

Fig. 3.7 further compares the Pareto-frontier, the Pareto-Jump and the Unselfish-Jump routing policies in terms of fairness in routing cost in the long-run. The horizontal axis is the round, i.e. a repetition of the ClubMED game with a new traffic

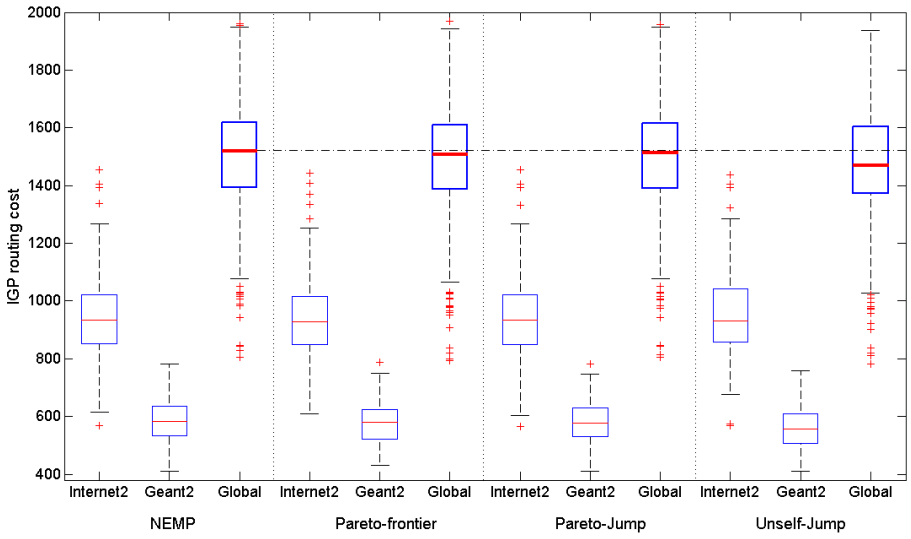


Figure 3.6: IGP routing cost Boxplot statistics: PEMP strategies.

matrix. The vertical axis displays the cumulative number of times in which a peer obtained a percentage gain with respect to the NEMP solution (e.g. for the first peer, $(\phi_{NEMP}(x,y) - \phi_{PEMP-str}(x,y))/\phi_{NEMP}(x,y))$ bigger than that of the other peer. In this way we can assess how fair is the solution of the repeated game in the long-run. Even if Geant2 is the final winner always, as expected the Pareto-frontier policy reveals to be the most fair one. This can be measured by the difference between the Internet2 and the Geant2 lines: the lowest with Pareto-frontier, the highest with Unselfish-Jump.

3.6.2 Route deviations

Fig. 3.8 reports the statistics of routing changes with respect to the previous round (with an upper bound equal to the total number of flows). The PEMP policies behave significantly better than BGP Multipath: they have a median of around 3 route deviations against 5, and the upper quartile and the maximum much lower. Interestingly, among the PEMP policies, the Pareto-frontier one statistically behaves better than the other policies for all the criteria but for the minimum. The reason may be that the Pareto-superiority condition applied on a very large set of candidate profiles (in fact, $n^{2m} = 531441$), offers a finer selection than the approximate potential threshold one. Finally, the Jump policies present a lower route stability with respect to all the statistical criteria. This is probably due to the fact that the jump from the Nash set is done without considering the cost errors.

As mentioned above, the original link capacity of both networks was scaled by 10 to create an intra-AS congestion risk. It is interesting to observe how the ClubMED framework can improve the route stability under 'normal conditions', in which an operator's network is largely overdimensioned.

In Fig. 3.9, we report the route deviations' number dynamics, together with the corresponding Boxplot statistics, obtained rerunning the simulations with original intra-AS link capacities, comparing the BGP solution to the ClubMED NEMP solution. The median of route deviations with NEMP falls to 0. The reason for this very good performance relates to the IGP-WO algorithm used to set the IGP weights. The IGP-WO cost function (such as the one implemented in TOTEM) assigns weights as function of the

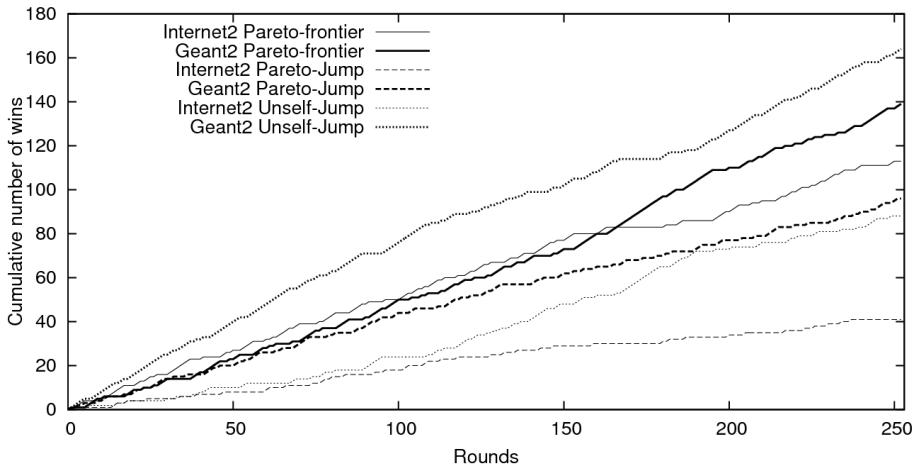


Figure 3.7: Dynamics of the cumulative number of wins.

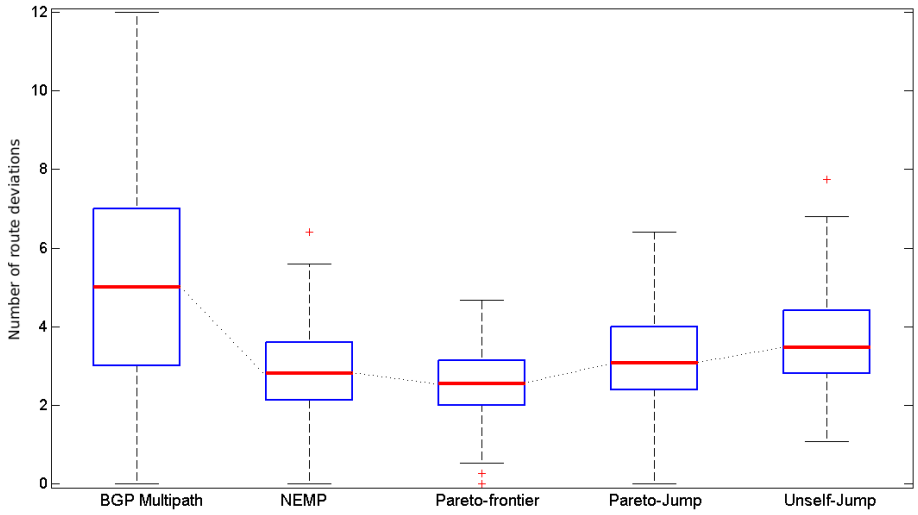


Figure 3.8: Number of route deviations.

expected load, so that with loads below 50% the variation in weight assignment is very low while it increases more than exponentially as the load approaches 100%. Therefore, we verified that ClubMED works even better with high available networks whose link IGP weight variations are contained.

3.6.3 Peering link congestion

Fig. 3.10 reports the Boxplot statistics maximum link utilization as seen by each peer, with the five above-mentioned methods. The PEMP policies except the Pareto-frontier one never caused congestion on peering links (utilization above 100%). The enabling of the Multipath mode in BGP does not have a significant effect on the peering link congestion. With ClubMED, instead, the multipath routing choice is carefully guided toward efficient solutions. The NEMP, Pareto-Jump and Unselfish-Jump policies show the median, the upper and lower quartiles always above 85%, remembering that with full BGP Multipath one would have the best $200/300 = 66,7\%$ utilization. The Pareto-frontier policy does not guarantee, however, a congestion-free solution, with a median close to 100% utilization. The reason for this behavior is the same as mentioned above:

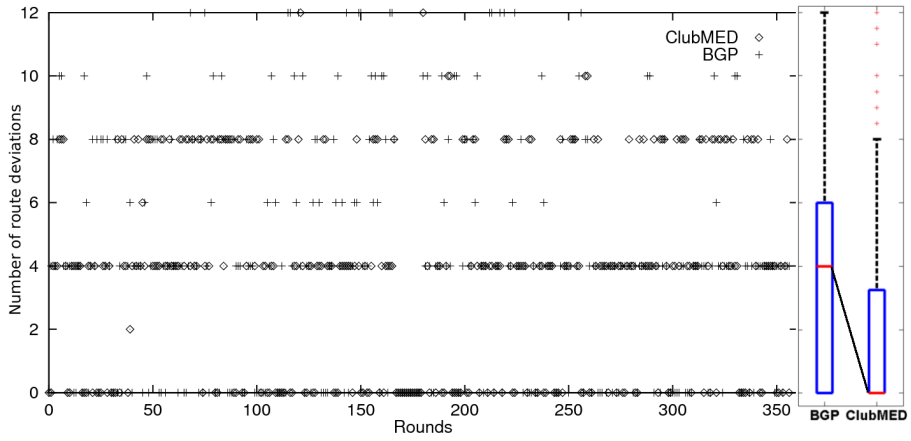


Figure 3.9: Number of route deviations with original link capacities.

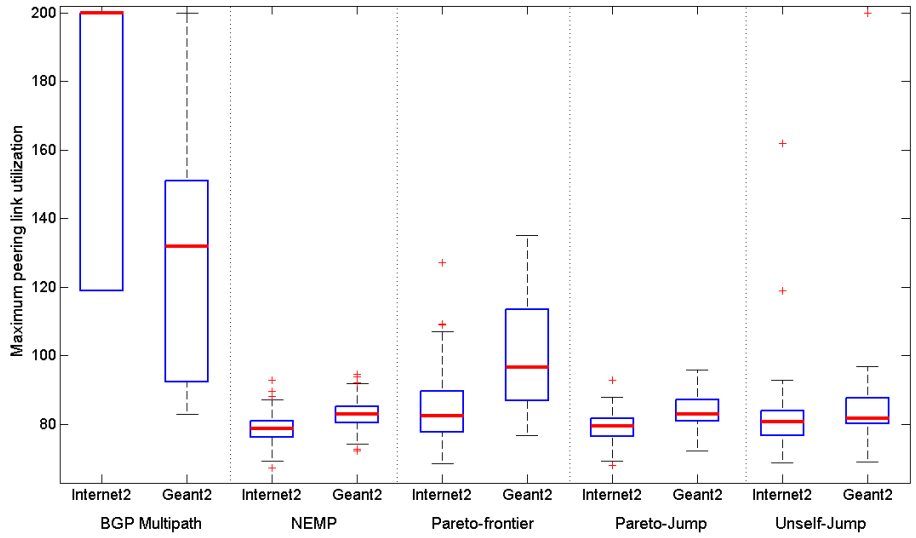


Figure 3.10: Maximum peering link utilisation boxplot statistics.

the Pareto-superiority condition may introduce highly asymmetric cost profiles in the multipath routing solution.

3.6.4 Time complexity

Fig. 3.11 reports the execution time for the PEMP policies. As expected the Pareto-frontier computation is excessively complex, with a $O(n^{2m})$ time complexity. The other policies have, instead, a polynomial complexity since they asymptotically depend on the minimization of a (mono-dimension) potential function to populate the Nash set. In fact, the other policies have an average computation time below 2 seconds (however, rare peaks of a few more seconds appear, probably due to the cases with very large Nash set, as can be seen cross-checking with Fig. 3.12). Hence, only the NEMP, Pareto-Jump and Unselfish-Jump policies shall be considered for a practical implementation (we have however introduced the Pareto-frontier case for a thorough comparison). Their execution times are acceptable in so much as the routing policies are computed after each IGP-WO, which can take much more time.

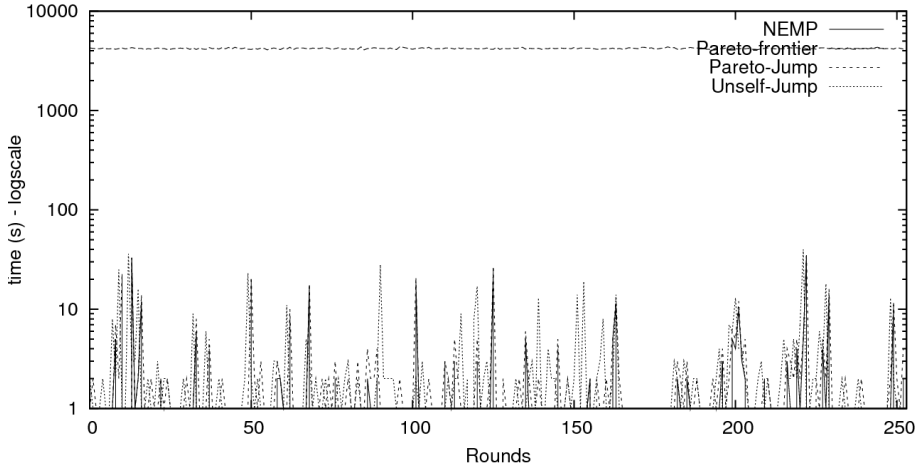


Figure 3.11: PEMP strategies execution time.

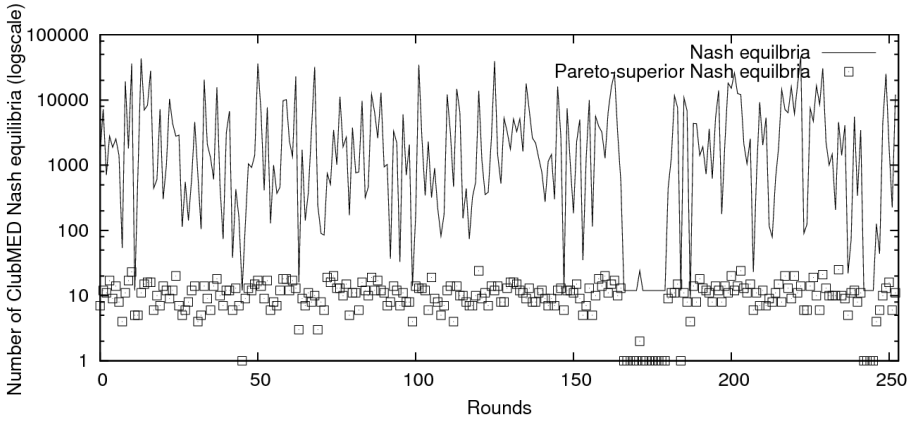


Figure 3.12: Nash set dynamics.

3.6.5 Nash equilibrium dynamics

Fig. 3.12 reports the number of ClubMED Nash equilibria and those Pareto-superior in a log-scale for all the rounds. The Pareto-superiority condition picks a few efficient Nash equilibria over broad sets, whose dimension varies significantly in time. This reveals a high sensitivity to the routing costs, probably due to the endogenous effect of G_c with high congestion costs; in fact, G_c cost components are not taken into account by the IGP ϵ errors.

3.7 Implementation aspects

The proposed framework does not require new protocol definitions or invasive extensions of existing ones. As partially already mentioned, there are important assumptions, possible intra-AS routing issues and ClubMED operations aspects that we discuss hereafter.

3.7.1 Technical assumptions

An assumption is that at each border a network management system is present to estimate traffic matrix, run IGP-WO and update IGP weights, as it happens nowadays for

large commercial Internet carriers. Nevertheless, from an algorithmic standpoint, the management operations or the IGP-WO algorithm behind IGP weight reconfiguration remain arbitrary and unilateral choices. The algorithmic requirement is that the ASs exchange IGP path cost variation information via the ϵ errors.

Moreover, the decision on the destination cones to include in the ClubMED communities should rely on an initial setting agreement between the peers. An initial setting agreement should also contain the scaling rules for the IGP weights (needed, however, only for the repeated coordination PEMP policies), which is particularly important especially for large providers that want to apply ClubMED with a large number of peers.

3.7.2 Routing and signaling

There are some routing and signaling aspects that relate to:

1. at the peering nodes:
   (i) the coding of multiple sub-attributes (ingress and egress IGP path costs, cost errors) into the MED attribute for the networks belonging to the destination cones (i.e., the prefix belonging to the ClubMED destination communities) – the new MED sub-attributes shall pass opaquely across intra-AS routers;
   (ii) the usage of the MED may be adapted on a per-community identifier fashion rather than on a per-prefix fashion, so as to aggregate the MED information; the community identifiers can in fact pass the AS frontier (i.e., no community strip operations on the prefix belonging to the ClubMED destination cones).

2. at the ClubMED nodes:
   (i) the modification of the BGP decision process at the "least MED"stage to select the multipath PEMP solution;
   (ii) the collection of the inter-peer flow bit-rate information for the congestion game (we assume that some metrology infrastructure, e.g., Netflow, is available).

3. with an IBGP AS core, there is no guarantee that at least one MED-icated route for each peering link will be visible at the ClubMED nodes, and (viceversa) that at least one route per ClubMED node will be visible at the peering nodes; let us call both kinds of routes 'ClubMED routes'. This can happen in some corner cases, in particular when some internal router compared ClubMED routes and announced only the best (with shortest IGP path cost) one. It is worth remarking, however, that the same issues would be present with BGP Multipath, and that in our simulations these route limitation cases were not considered (which actually yielded better than real solutions for BGP Multipath).

   Nonetheless, to deal with such corner cases, a BGP-friendly approach would be to limit the strategy set of a player to the 'visible' peering links at each ClubMED node; however, in the absence of specific signaling among each peer's ClubMED nodes announcing which peering links each peer considers in the strategy set, we would have a game with incomplete information in which the strategy sets considered by the peers are not completely known. The ClubMED game with incomplete information, even if respectful of BGP, may no longer be as effective as with complete information since a probability distribution shall be used by each peer over the different types of players (number and type of strategies) it could experience (see [40]).

4. in the case of configuration of BGP Route Reflectors (RRs) the visibility issue described above could be even more important. Moreover, ClubMED nodes should not behave as normal RR clients for the networks belonging to the ClubMED destination cones.

Let us further discuss the implementation aspects 3 and 4, pointing out the correlated signaling issues that should be tackled to avoid the incompleteness of the game information and to deal with RRs:
(i) only the ClubMED nodes play the game, not intermediate intra-AS routers (those in between ClubMED nodes and peering routers);
(ii) the ClubMED nodes should learn all the different peering routes in order to play the ClubMED game (avoid having only best paths);
(iii) intermediate nodes should forward packets to the proper egress router (without playing the game).

With respect to (i), there is no scalability issue in that only a few AS border routers are likely to be elected ClubMED nodes even in large networks.

With respect to (ii) and (iii), ClubMED nodes could just have configured an IBGP direct session with the peering routers. With a BGP-free-core configuration, i.e., direct BGP sessions only among AS Border Routers and an MPLS-managed AS core, the game could be played in its complete form. If RRs are configured, since their normal setting contrasts with (i) and (ii), they should announce the several routes with the same AS path to ClubMED nodes, at least for the routes whose prefix belongs to the ClubMED destination cones.

3.7.3 ClubMED execution policy

It is worth stressing that the ClubMED game does not require an execution of an IGP-WO for the computation of each PEMP solution. The G_c components do not depend on IGP metrics and can be updated when a peering link fail or when the inter-community flow bitrate (μ_h) change.

There is no need to compute the PEMP solution after each IGP-WO or after each inter-community flow bitrate variation. An appropriate execution policy, to be defined in a further work, should be able to assess the opportunity to re-run the PEMP computation at each side with respect to IGP weights and inter-community flow bitrate variations.

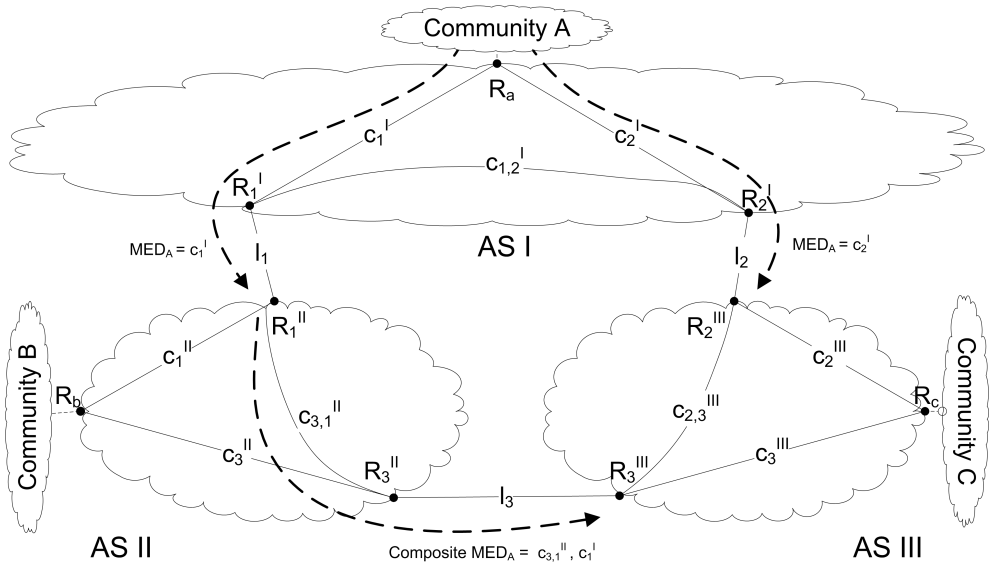


Figure 3.13: Extended peering scenario with 3 peers (for simplicity only AS I MED signaling and simple bidirectional costs are depicted).

3.8 Conception of Internet Extended Peering

Within the ClubMED framework it is thus possible to efficiently control the route deviations by fine-tuning the routing strategy. The major practical benefits from the implementation of the PEMP policies would be the trust-reinforcement of an existing peering agreement and the improvement of the provided QoS (hence QoE) related to the lack of congestions and frequent deviations. Let us try to see what happens if we generalize the 2-player game to a n-player game, to then discuss the practical incentives for such a novel model.

The extension of the ClubMED framework to more than two players could allow the definition of a sort of "Extended Peering"in which the border one provider has with the other neighbors (peers or candidate peers) is modeled as a single border. Please note that this differs from many sibling settlements (see Sect. 2.1.1) In an extended peering, only the peers' client traffic would be routed across the peering borders. In order to treat multi-peer borders as a single equivalent peering border, (in the extended framework) transit costs at each peer - from each neighbor to every other neighbor - shall be considered in the game modeling.

Referring to the 3-peer scenario in Fig. 3.13, e.g., AS I announces the destination prefixes of community A to AS II with the MED set to the intra-AS I routing cost c_1^I. AS II in turn announces the same prefixes to AS III. In such announcements, a composite MED is to be coded including the individual routing costs that the selection of the link l_3 by AS III would cause to the ASs in the extended peering chain, thus in this case to AS II ($c_{3,1}^{II}$) and to AS I (c_1^I)[2]; instead, the routing cost toward local communities, e.g. c_3^{II} for AS II, is sent via a normally MED-icated announcement. AS III disposes of two routes toward the community A, one through the direct link l_2 with AS I, one crossing AS II. Being aware of the costs that its routing decision causes to the other peers (given by the composite MEDs), the router R_c of AS III decides consistently with

[2]In the composite MED, the reference to the selected peering links over the inter-peer path is lost. Only the routing costs impacted to the peers matter.

the extended peering game strategy profiles. R_c decides toward what peering link to route the aggregate C→A+B flow, aware of the routing costs it implies for AS II (transit cost $c^{II}_{3,1}$, from l_3 to l_1 for C→A, and c^{II}_3 for C→B) and for AS I (c^I_1).

In the general case, many peering links can connect two peers. Moreover, many ASs can transit traffic toward the same destination community, and the AS chain lengths within the extended peering vary. While inter-community routing is distributed at the edge routers (e.g., R_a, R_b and R_c) following the extended peering game (thus bypassing BGP), transit routing decisions are, instead, taken at the peering routers (e.g., R^{II}_1 and R^{II}_3) following the normal BGP routing policy for the ingress peering flows (without specific route filtering). The peer routing costs, which depend on the peering router's decisions, are to be coded in the composite MED sent to the neighbors. For those MEDs that are composite, the smaller MED rule shall be applied to the sum of all the MED parts. Finally, it is possible that many "MED-icated"routes from different ASs have the same AS hop count. In such a case, MEDs from different ASs shall be normalized over a same IGP weight scale.

3.8.1 The extended peering game

The extended peering game is a straightforward extension of the 2-player game:

- the number of strategies increases due to the enlarged interconnection,
- G_s and G_c maintain the same structure,
- G_d includes also the exogenous transit costs toward the external destination communities; a transit cost is simply summed to the ingress cost for the internal destination community.

In Fig. 3.14 there is a so-built extended peering game example with 3 carriers, with just one link connecting two peers, and without G_c; the corresponding strategic form is in Table 3.1. The decision of routing on a link impact an egress cost for the deciding AS, an ingress cost and a transit cost for the next AS and an ingress cost for the last AS; and three routing decisions must be taken at community edge routers, one for A→B+C flows by AS I, one for B→A+C flows by AS II and one for C→A+B flows by AS III. In Table 3.1 there is the strategic form of the game; each cell corresponds to a strategy profile and indicates between brackets the routing cost for AS I, AS II and AS III, in the listed order; each peer strategy corresponds to a possible link where to route its own egress flow, thus e.g. l_1 and l_2 for AS I, l_1 and l_3 for AS II and l_2 and l_3 for AS III. We have a Nash equilibrium in (l_2,l_3,l_3). Similarly than in 2-player games, it is possible to have profiles Pareto-superior to the equilibrium(a), such as (l_1,l_1,l_3) that grants a lower cost for all ASs.

Under the assumption that the IGP costs of all the carriers involved in the extended peering are normalized to the same scale, the PEMP polices should be used to coordinate the extended peeering routing strategy. Given that the extended peering game is a straightforward extension of the 2-player one, we expect similar results and benefits for this framework (no simulation results will be provided herein).

3.8.2 Incentives for an extended peering

The incentives for implementing an extended peering coordination framework are not straightforward. The alternative closer solution would be a full mesh of classical bilateral

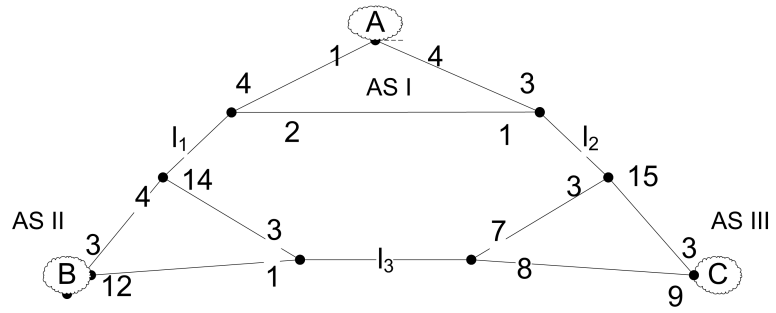


Figure 3.14: Extended peering game example.

III

l_2	I \ II	l_1	l_3
	l_1	(12,13,27)	(14,10,36)
	l_2	(11,19,28)	(13,16,37)

l_3	I \ II	l_1	l_3
	l_1	(7,36,20)	(9,33,29)
	l_2	(6,42,21)	(**8,39,30**)

Table 3.1: Strategic form of Fig. 3.14 example.

peering agreements. With respect to the best case with a full mesh of ClubMED-based bilateral peering agreements, an extended peering framework would have the following key advantages:

- *Extended balance*: it may be easier to agree on a peering among many carriers rather than among only two carriers since, e.g., the traffic balance condition may be reached more easily by considering a larger set of flows.
- *Higher Internet reliability*: congestions or outages at one peering border or de-peerings can be surrounded by an automatic rerouting of the traffic elsewhere within the extended peering settlement without losing visibility toward a piece of the Internet;
- *Larger path diversity*: the resulting increased path diversity can further improve the efficiency of the peering routing solution (with respect to routing costs, congestions and deviations).

Nevertheless, these incentives may be too weak because not appealing enough especially for those top-tier providers for which the existing peering settlements are well balanced, for which the reliability is not a relevant issue (with a number of peerings with high cone overlapping), and which would see the extended peering management too cumbersome already with a discrete number of communities.

The 'killer incentive' for a generic form of extended peering might be additional revenues related to novel added-value inter-provider services. The framework may, indeed, be used to differentiate the treatment of added-value services overlapping best-effort routing. However, such services may also require a connection-oriented network technology, with strict QoS requirements, which can be provided only if hard collaboration

schemes among providers are implemented. Instead “Extended Peering”with pro-active routing coordination, it would be probably more appropriate to refer to a “Provider Alliance”with more cooperation, within which an ad-hoc inter-provider routing architecture, totally decoupled from BGP, is provided. This more collaborative path is discussed and treated in the next chapters.

3.9 Summary

In this chapter we tackled the coordination issue for current BGP/IP routing. After reviewing related work, we modeled the routing on peering links as a non-cooperative game with the aim of allowing carriers to fine-select routes for critical flows by following efficient equilibrium multipath solutions. We presented the mathematical model of the game, composed of a selfish game, a dummy game and a congestion game. The first considers egress IGP costs, the second ingress IGP costs and the latter peering link congestion costs. The game components can be easily adapted to consider IGP cost variations due to IGP-WO re-optimizations.

We proposed a low-computational way to compute the Nash equilibria of the game. We presented four possible Peering Equilibrium MultiPath (PEMP) routing policies under which carriers can coordinate. The first one balances the load on the Pareto-superior Nash equilibria of the one-shot game. The second one balances the load on the Pareto-frontier, that is the equilibrium of the repeated game. Finally, the latter two policies improve the first strategy moving from the starting Nash set toward Pareto-superior and unselfish routing profiles (respectively).

We simulated these different PEMP policies with a realistic emulation of the peering between the North-American and European research networks. Results are compared with existing BGP Multipath, the current inter-AS multipath IP routing protocol. The results show that PEMP outperforms BGP Multipath in terms of routing cost, route stability and peering link congestion. In particular, the route stability is significantly improved and the peering link congestions can be practically avoided. Some differences exist between the four PEMP policies. Namely, the Pareto-frontier policy strategy is extremely complex and shall not be implemented. The others present some trade-offs but all represent promising solutions to perform an efficient and rational routing across peering links. In particular, the Unselfish-Jump policy represents the best trade-off between peering trust insurance, routing cost, congestion control, routing stability and execution time.

Finally, we showed that the extension of the framework to an arbitrary number of provider-players can be done straightforwardly by modeling additional transit routing cost in the dummy game. Such an extension might allow the definition of extended peering models that could increase the Internet reliability at the expense of some additional complexity at the border routers. Nevertheless, the incentives for the extended peering framework may be too weak whether strict inter-provider QoS guarantees are required.

Chapter 4

Towards Provider Alliances for Connection-Oriented Services

We concentrate on the automated provisioning of inter-provider services based on MPLS-TE/G-MPLS technology. For such services, simple 'light' coordination schemes, such as those defined in the previous chapter, would not allow meeting 'hard' QoS requirements needed by inter-provider tunnels or circuits. A higher level of collaboration is needed within an ad-hoc cooperative multi-provider service architecture solution object of this chapter. We conceive a provider alliance architecture where Traffic Engineering (TE) connections are established between the members of the alliance. We define the architecture for the automatic provisioning of inter-AS connection-oriented services, based on the introduction of a multi-provider service plane coupled with the PCE-based architecture. We report also some details of a testbed implementation[1][2].

4.1 Introduction

Providers usually deploy TE technologies such as MPLS-TE [163] to offer value added services. Nevertheless, these technologies are restricted to intra-AS networks, narrowing the scope of potential service offers. There is a clear demand nowadays to extend guaranteed service offers beyond provider boundaries. In order to support inter-provider services, it is needed to rely on inter-AS QoS mechanisms. The Internet Engineering Task Force (IETF) has defined an extension to the MPLS-TE technology, called inter-AS MPLS-TE, which allows establishing inter-AS, explicitly routed, TE connections with stringent QoS and availability constraints. In this chapter, we propose to enrich the current inter-AS MPLS-TE/G-MPLS technology in order to enable automatic provisioning of inter-AS TE services.

In the following section, we describe the inter-AS TE building blocks involved in our solution and point out their current limitations to achieve our goal. We then propose to introduce a service plane coupled with the PCE-based architecture show how it can allow negotiation of providers' service elements via successive phases of selection, instantiation

[1]The contents presented in this chapter are also presented in [5] and [10].

[2]The work presented in this chapter and the next one has been conducted in the framework of the Euro-FGI TEIDE (Traffic Engineering Inter-Domain Extension) research activity, the ANR ACTRICE (Agence Nationale pour la Recherche - Approche Combinée de Technologies Réseaux Inter-domaine sous Contraintes Economiques) project, and the CELTIC/EUREKA TIGER2 (Together IP, GMPLS and Ethernet Reconsidered - Phase 2) project

and activation. Finally, we define the necessary extensions to existing network protocols and management elements and we illustrate how they can be applied to offer automatic provisioning of inter-AS TE services.

4.2 Inter-AS MPLS-TE/G-MPLS

Within provider boundaries, the MPLS and G-MPLS technologies (possibly integrated with ASON, Automatically-Switched Optical Network [184], modules) allow establishing connections, called Label-Switched Paths (LSPs), based on any transport network solution (circuit or packet based). The TE extensions add the possibility to route a LSP explicitly, taking TE constraints into account: for instance verifying resource availability, switching capability and end-to-end or subpath protection possibility. As already mentioned, the IETF has extended the MPLS-TE technology to support the configuration of inter-AS LSPs, meeting the requirements in [169].

4.2.1 Inter-AS Path Signaling

A signaling protocol called RSVP-TE is used to establish LSPs. As explained in [174], an inter-AS LSP can be signaled in three ways:

- *LSP Nesting*: In this mode a local high level intra-AS LSP is used between AS border routers to transport many inter-AS LSPs sharing a common intra-AS subpath. For purely MPLS backbones, this corresponds to encapsulating an inter-AS tunnel into an intra-AS tunnel. For optical networks, this corresponds to grooming incoming inter-AS MPLS or G-MPLS LSPs into lower level intra-AS LSPs with coarser optical switching capabilities (fiber, waveband or wavelength).

- *Contiguous LSP*: In this mode a single end-to-end LSP is signaled across the ASs. The head-end router is connected to the tail-end router via a single signaling session. This means that single session and LSP identifiers are used across the inter-AS path. Hence the re-configuration of such an LSP is controlled by the head-end router, and intermediate ASs should not modify their local subpath.

- *LSP Stitching*: In this mode, intra-AS LSPs are signaled and then stitched at the boundaries to form a single inter-AS LSP perceived in the data-plane. From the control-plane standpoint the local intra-AS LSPs are signaled separately and every LSP has different source and destination (ingress and egress AS border routers). This signaling method would particularly be applied to the case in which some switching capabilities are not compatible with the nesting method. For instance a lambda-LSP cannot be nested in another lambda-LSP.

4.2.2 Inter-AS Path Computation

A LSP is to be signaled over a pre-computed path. A head-end router has full topology visibility within its AS, and can hence compute alone an end-to-end intra-AS path, but cannot compute alone an end-to-end inter-AS path. Two methods can be adopted for the inter-AS path computation:

- a per-AS path computation method, in which the source or ingress router determines the next AS and the ingress router in this next AS, and computes the corresponding subpath. Then the path computation is moved to the ingress router

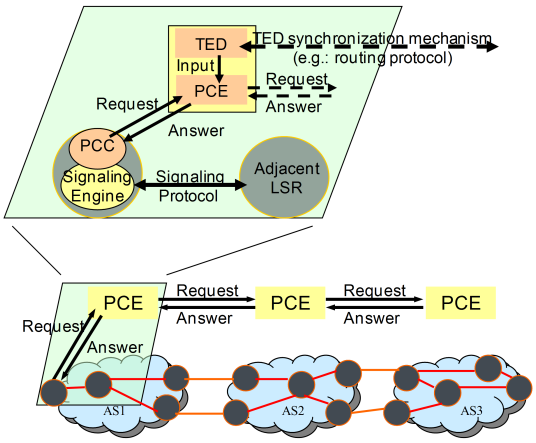


Figure 4.1: Communications relating to PCE

of the next AS, and so on up to the tail-end router. This simple method does not allow computing a shortest inter-AS path and may lead to several crankbacks that may affect the stability of the control plane;

- an inter-AS path computation method that takes as input the AS chain (i.e the succession of ASs to be used), and relies on computation servers present in each AS, called Path Computation Elements (PCEs), to collaboratively compute an inter-AS shortest path along the given AS chain.

4.3 Path Computation Element Architecture

The PCE architecture [170] consists in delocalizing the path computation from the head-end router to a PCE that computes paths on behalf of the head-end router. PCEs can collaborate together in order to compute constrained inter-AS paths, without being required to share any TE information with each other, thus solving the topology visibility issue. PCEs can be particularly useful when end-to-end constraints for protection or path diversity have to be kept into account. As depicted in Fig. 4.1, at least one PCE is required per AS. A PCE serves requests sent by Path Computation Clients (PCC), e.g., routers or switches, using information of a local TE Database (TED). A PCE can query the PCEs of other ASs to perform this computation, acting in turn as PCC. A PCE Communication Protocol (PCEP) has been defined to relay these requests and answer messages [177]. PCCs can dynamically discover external PCEs through extensions of existing routing protocols, meeting the requirements in [171]. As already mentioned, the PCE-based inter-AS path computation can be performed once the AS chain for the destination is known. Efficient distributed algorithms are needed for the end-to-end inter-AS path computation taking into account a pre-computed AS chain.

The procedure called “Backward Recursive Path Computation (BRPC)”is the one that seems best meeting the requirements of both operators and suppliers in terms of complexity and network information hiding [178]. It consists in computing recursively, at each PCE of the AS chain, an inverse tree of constrained shortest paths with one branch for each ingress border router (and towards the destination), starting from the destination AS. Each path might be a loose path containing only the tail router, the border routers, and the cost of the corresponding shortest path. The tree is sent back to the previous AS, which does the same, and so forth back to the source AS.

4.4 Contributions

As we have mentioned, the IETF developed solutions for inter-AS LSP set-up. However, we can outline the following open issues:

- for the PCE-based architecture, the standardization does not indicate how the input AS chain is calculated;
- the setup of an inter-AS LSP is subject to strong business, security and confidentiality aspects. Hence the setup of an inter-AS LSP must only be done between trusted entities and it requires a preliminary service instantiation and activation, in order to ensure billing and to manage routing and signaling requests at AS boundaries. Such procedures are beyond the scope of the IETF and are still not defined.
- inter-AS service provisioning steps shall be automated to deal with the complexity of cross-AS service management.

As already reminded in the previous chapters, for connection-less (best-effort) services the AS chain selection is performed via BGP, with a cascade of criteria to compare alternative paths. Through the first criterion (local preference) local policies mainly guided by economical issues can be applied: e.g., a peering link is preferred to a transit link. The subsequent criteria incorporate, instead, purely operational network issues. Hence gathering one path per destination AS from BGP is not advisable because paths are chosen regardless to QoS, availability and reliability constraints, and because the BGP decision process is too simple to model the complex AS relationships that will emerge with the generalization of 'extended peerings' (including valued added services) or 'provider alliances'. However, the economical distinction of paths granted by the BGP local preference criterion should not be lost, but, on the contrary, enhanced.

Within the ACTRICE project we have studied how the deployment of a multi-provider service plane can support an automatic provisioning of multi-AS LSPs, under a favorable business model. This service plane is to be adopted by an alliance of providers willing to collaborate for the delivery of multi-AS TE services, and willing to decrease the overwhelming operational efforts nowadays related to such a service (a chain of bi-lateral ad-hoc agreements). Within the provider alliance, imprecise TE information is to be transmitted by means of service elements through which each provider advertises its transit capabilities and policies. Further on, we focus on the service establishment within a provider alliance. It consists in a first phase of selection and validation of service elements to compose the AS chain, and in a second phase of service configuration (computation and signaling). In the following, we characterize the required inter-AS service plane and detail the inter-AS LSP provisioning steps.

Related work

Apart the IETF activities already resumed, previous relevant works on inter-domain MPLS-TE have been achieved, particularly in the context of AGAVE, DRAGON, EuQoS and MESCAL projects.

Within the NSF Dragon project [83] an experimental PCE-based framework for multi-domain provisioning of TE paths has been implemented. A distributed control plane across heterogeneous networks, with different switching technologies and granularities, has been tested, including mechanisms for authentication, authorization, accounting

(AAA), and scheduling. This work seems particularly useful for grid networks where economical constraints are absent and QoS constraints are often limited to availability and survivability.

The IST Mescal and Agave projects mainly recommend a provider-centric approach for connection-less services based on a cascaded model [81]: in this framework an AS can discover transit QoS capabilities only of adjacent ASs, and only towards specific destination networks. This limits the path diversity. Within these projects, and similarly in the EuQoS project, extensions to BGP have been proposed to advertise TE information [80] [79]. Other studies propose the combination of distributed overlay architectures and BGP extensions (e.g., [78]). Such approaches require changing BGP, which is problematic, given the number of existing routers deployed currently. The exchanging of QoS information on BGP is also questionable even if the proposers claim that it scales [80]. Moreover, these choices do not meet the route diversity requirement [113]: assuming a cascaded model, proposing extensions to BGP, not enough TE inter-AS routes per destination can be taken into account.

4.5 Notion of Inter-AS TE Service

As explained before, we can benefit from the mechanisms defined at the IETF for the configuration of inter-AS LSPs. However, this should rely on a preliminary agreement between providers on a common service plane, and should require the application of important TE and security policies at the providers' boundaries. We tackle these fundamental service and policy aspects, out of scope of the IETF. An "inter-AS TE service"' is composed of one or more inter-AS LSPs between a Head Router in the source provider and a Tail Router in the destination provider, crossing a chain of transit providers. An inter-AS LSP can be unidirectional for content distribution, or bidirectional for interactive services. We characterize an Inter-AS TE Service by the following parameters:

- address of the Head and Tail Routers;
- source and destination AS Numbers;
- AS chain;
- direction: unidirectional or bidirectional;
- bandwidth;
- Service Level Specifications (SLSs), containing performance parameters and a cost;
- protection level: unprotected, global protection, local protection;
- re-optimization: enabled or disabled.

An inter-AS TE service is the result of a composition of service elements offered by each operator. We introduce three service element categories:

- the "Edge Sender", which assures the routing from the head router of the sender AS toward an ingress router of a neighboring AS;
- the "Edge Destination", which assures the routing from an ingress router of the destination AS toward the tail router;

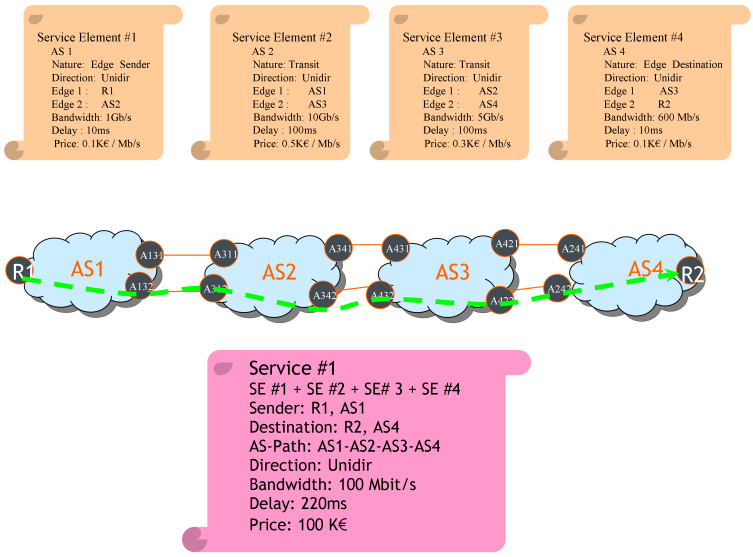


Figure 4.2: Service Elements

- the "Transit", which assures the routing from an ingress router of the AS towards an ingress router of the next AS.

Fig. 4.2 illustrates an example of inter-AS TE service, unidirectional and unprotected, between the R1 router of AS1 and the R2 router of AS4 across AS2 and AS3. It is the composition of four service elements, two transit, one sender and one destination. Each element indicates explicit edges as incoming and outgoing border routers.

4.5.1 Service Elements

To allow the composition of an end-to-end service, every operator advertises its service elements to the other members of the alliance. It is worth mentioning that the IP Sphere Forum is specifying a framework that, among other features, can allow reliable multi-provider advertisement of service data via (SOAP/XML-based) web-services [186]. We characterize a service element by the following parameters:

- local AS Number;
- nature of the service: Sender/Destination/Transit;
- direction: unidirectional or bidirectional;
- ingress and egress edges;
- upper bounds of performance parameters;
- maximum bandwidth that can be reserved for a given session;
- protection level (unprotected, global, local);
- transit price, per bandwidth and/or duration unit.

An edge can be identified explicitly by the address of a router or a group of routers, or implicitly by the neighbor's AS Number. Obviously, in the case of Edge Sender and

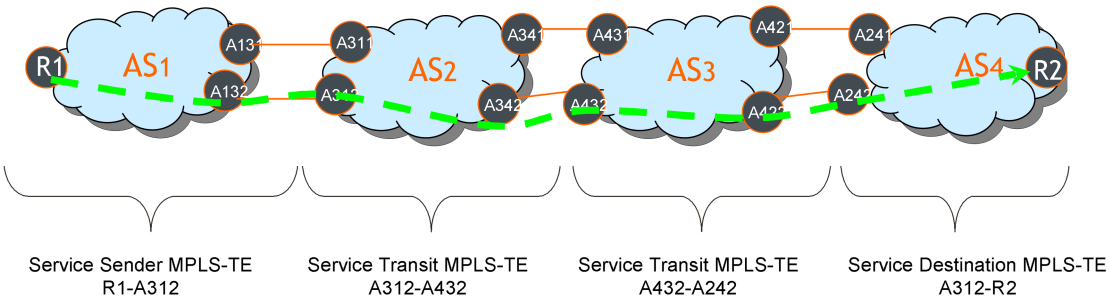


Figure 4.3: Service Elements Composition

Edge Destination service elements, one of the two edges would represent the head and the tail router (or group of routers), respectively. In Fig. 4.3, a set of service elements that contribute to the composition of an end-to-end network service is displayed. In this example, each element indicates implicit edges as incoming and outgoing ASs. The choice of these four service elements is guided by the compliance with the request parameters and the minimization of the service cost. Coherently, the AS chain is built as list of the corresponding ASs. Within the provider alliance, business relationships are defined by the ASs' policy in advertising service elements and in admitting requests. Existing transit or peering agreements for best effort IP routing may intervene in the alliance formation, or in the service element definition, or may be simply disregarded for inter-AS TE services.

It is worth remarking that the described framework is not restricted to MPLS-TE networks, but it encompasses ASON/G-MPLS-based networks. As previously mentioned, LSP nesting and stitching methods can locally rely on G-MPLS LSPs. The service plane signaling is agnostic on the used switching granularities, and the service element's parameters are not restricted for MPLS tunnels.

4.5.2 Requirements

We can now outline that the set-up of an inter-AS TE service requires the following subsequent steps.

1. The discovery of the service elements offered by each provider.
2. The composition of the service elements to form an end-to-end inter-AS TE service, i.e., the computation of the cheapest constrained AS chain. This computation should take care of:
   - the fact that transit service elements contain directional policies, from an ingress edge toward an egress edge. The AS graph is weighted thus by 'directional' metrics and not simple metrics;
   - the presence of computation servers, the PCEs, that could support pre-computation allowing to decrease the online time complexity (i.e., related to the task executed once the request is received).
3. The instantiation of the composed service. This should include:
   - a Connection Admission Control at the service plane to verify the availability of the service elements;
   - a confirmation of the SLS.

4. The activation of the service. This should include, in the following order:
   - the configuration of filters on policy managers of each AS to validate inter-AS PCEP and RSVP-TE messages in function of the instantiated service;
   - the configuration of the LSP on the head router.
5. The inter-AS path computation along the composed AS chain, via the PCE-based architecture. Inter-AS PCEP messages have to be filtered with:
   - the translation of some TE parameters (priority, class of service);
   - the filtering of topological information to assure the confidentiality;
   - the rejection if not compliant with the instantiated service (SLS, etc.).
6. The LSP signaling via RSVP-TE. Inter-AS RSVP-TE messages have to be filtered similarly to PCEP messages as indicated above.
7. The service maintenance. This should include:
   - accounting and measurement of the end-to-end and local performances;
   - fault detection, localization and reporting.

4.6 Architecture for the Automatic Provisioning of Inter-AS MPLS Services

In order to reduce the operational costs and minimize service set-up and restoration times, all the steps enumerated above need to be automated. The IETF has defined mechanisms for the steps 5 and 6 above. Nonetheless, extensions are needed to support the transport of a service identifier in RSVP-TE and PCEP messages, so as to apply the filters. The steps 1-4 and 7 need the introduction of a new plane, a service plane, which allows service information exchange among providers. This plane may for example rely on an adaptation of the IP Sphere Forum technical framework.

4.6.1 Functional Elements

The Fig. 4.4 illustrates the elements needed to automate the set-up of an inter-AS TE service. We can distinguish three layers. The network layer encompasses the ASs with their core and border routers (ASBR), and their PCEs. The network layer is guided by the management layer, where a Network Management Server (NMS) and a Policy Manager (PM) are needed as explained hereafter.

The management layer is in turn guided by the novel layer that we introduce, the inter-AS service layer. This layer is composed of per-AS agents called Service Element Agents (SEA), of end-to-end agents called Service Agents (SA), and of agents responsible of the service element composition called AS Selection Agents (ASA). At the network layer, the ASBRs are linked by an IP/MPLS link and interact via the RSVP-TE protocol. A router communicates with its local PCE via the PCEP protocol. The PCEs communicate via the PCEP protocol too, for the end-to-end path computation.

Between the network and the management layers, the ASBR communicates with its NMS via a network management protocol (e.g., SNMP, XMLConf, Telnet CLI) to set-up LSPs and to rise information about LSP status. An ASBR similarly communicates with its PM to perform the admission control of inter-AS RSVP-TE and PCEP messages

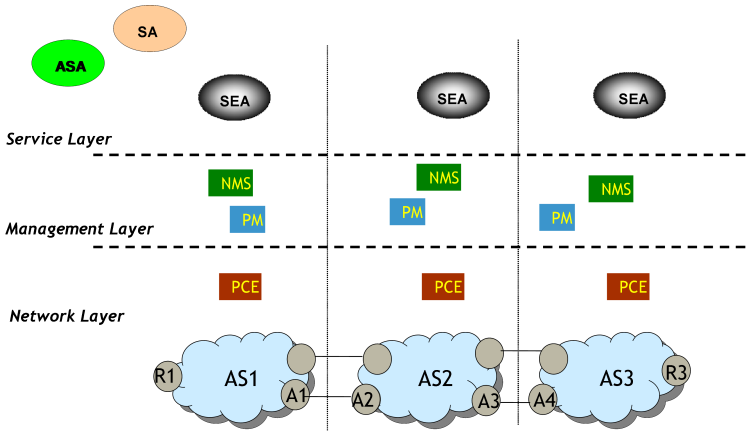


Figure 4.4: Inter-AS Multi-Layer Service Architecture

(this may rely, e.g., on a policy protocol such as COPS, Diameter or a yet to be defined SOAP/XML solution).

Between the service and the management layers, the NMS receives inter-AS LSP setup, change or deletion commands from the SEA and rises LSP status information to the SEA. The PM also communicates with its SEA to acquire filtering policies corresponding to the services activated at the service layer. The policies are to be indexed by a service identifier. PM/SEA and NMS/SEA communications may be based on a SOAP/XML protocol.

It is worth noting that there is no inter-AS communication at the management layer. At the service layer, the SA is responsible of the construction of the end-to-end service. It receives inter-AS LSP setup, change and deletion commands from an administrator. It queries the ASA to calculate an AS chain. Then, it communicates with the SEAs along the AS chain to instantiate and activate the service elements. SA/SEA and SA/ASA communications may be based on SOAP/XML too.

4.6.2 Functional Steps

We distinguish thus seven functional steps corresponding to the different life-cycle phases of an inter-AS TE service, at the service plane (Fig. 4.5) and at the management and network planes (Fig. 4.6).

1. *Service Discovery (DISC)*: This step consists in acquiring the inventory of all the service elements offered by the providers of the alliance.

2. *Composition of Service Elements (COMP)*: This step consists in determining the LSP's AS chain (Fig. 4.5a). The administrator triggers the composition, via the SA, at the ASA where constrained shortest path algorithms should be implemented. The ASA answers with one or many AS paths. Many diverse AS paths may be selected in order to increase the success of having at least one accepted, or to meet path diversity requests. Indeed, a selected AS path can fail during the instantiation, the activation, the path computation or the signaling phases. The COMP step follows the ideas of service element composition described in [134].

3. *Service Instantiation (INST)*: This action aims at verifying the availability of the service elements composing the AS chain, and at agreeing upon the final SLS and

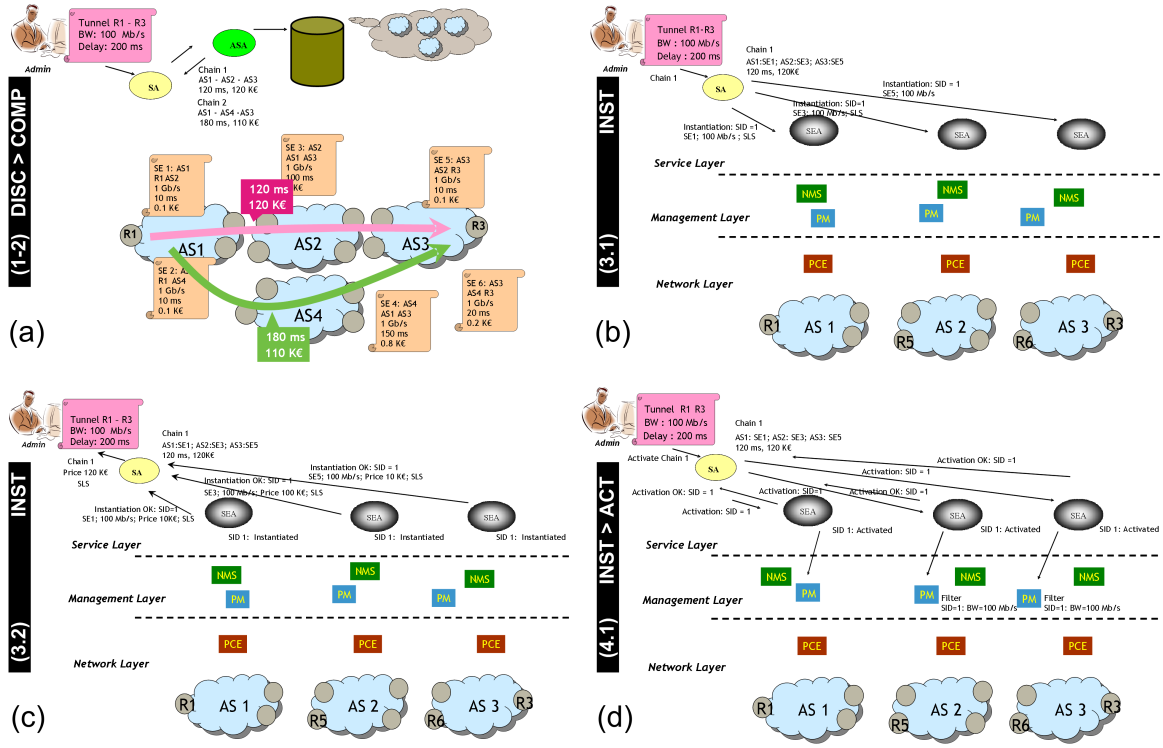


Figure 4.5: Discovery, Instantiation and Activation at the Service Layer

cost (Fig. 4.5b-c). An Instantiation message is generated at the source SA and is sent to the SEAs of the involved ASs. The request contains a Service IDentifier (SID) which is produced to identify the service during all its life-cycle. Then, in the case of availability the SEAs send back an Instantiation OK message with the current SLSs (potentially changed), and the current price otherwise it sends an Instantiation NOK message. If an element is not available or if the SLS is not acceptable, the SA can test another AS chain.

4. *Service Activation (ACT)*: This step consists in triggering the service establishment (Fig. 4.5d, Fig. 4.6e). As a first action, the SA sends to all the SEAs an Activation message with the SID. These SEAs send to the PMs a filtering policy associated to the SID, useful to filter inter-AS PCEP and RSVP-TE messages. If no error occurs, each SEA sends back an Activation OK message, otherwise an Activation NOK message. If all the responses are positive, the SA sends an Activation message to the source SEA, which then commands the LSP configuration to the local NMS with all the request details. This SEA sends to the SA an Activation OK message if no error occurs. Then, the NMS configures the inter-AS LSP on the head router, passing the SID and the AS chain in addition to base TE parameters.

5. *Path Calculation (CALC)*: This step consists in computing the inter-AS path via the PCE-based architecture (Fig. 4.6f-g). Acting as PCC, the head router sends a 'PCReq' message to its local PCE. This message is propagated along the PCEs of the AS chain up to the destination PCE, where the BRPC procedure can start. During the computation, a 'PCRep' message is propagated backwards towards the source PCE. A confidentiality key should be associated to this local information so as to retrieve the full intra-AS path during the signaling phase [179]. The

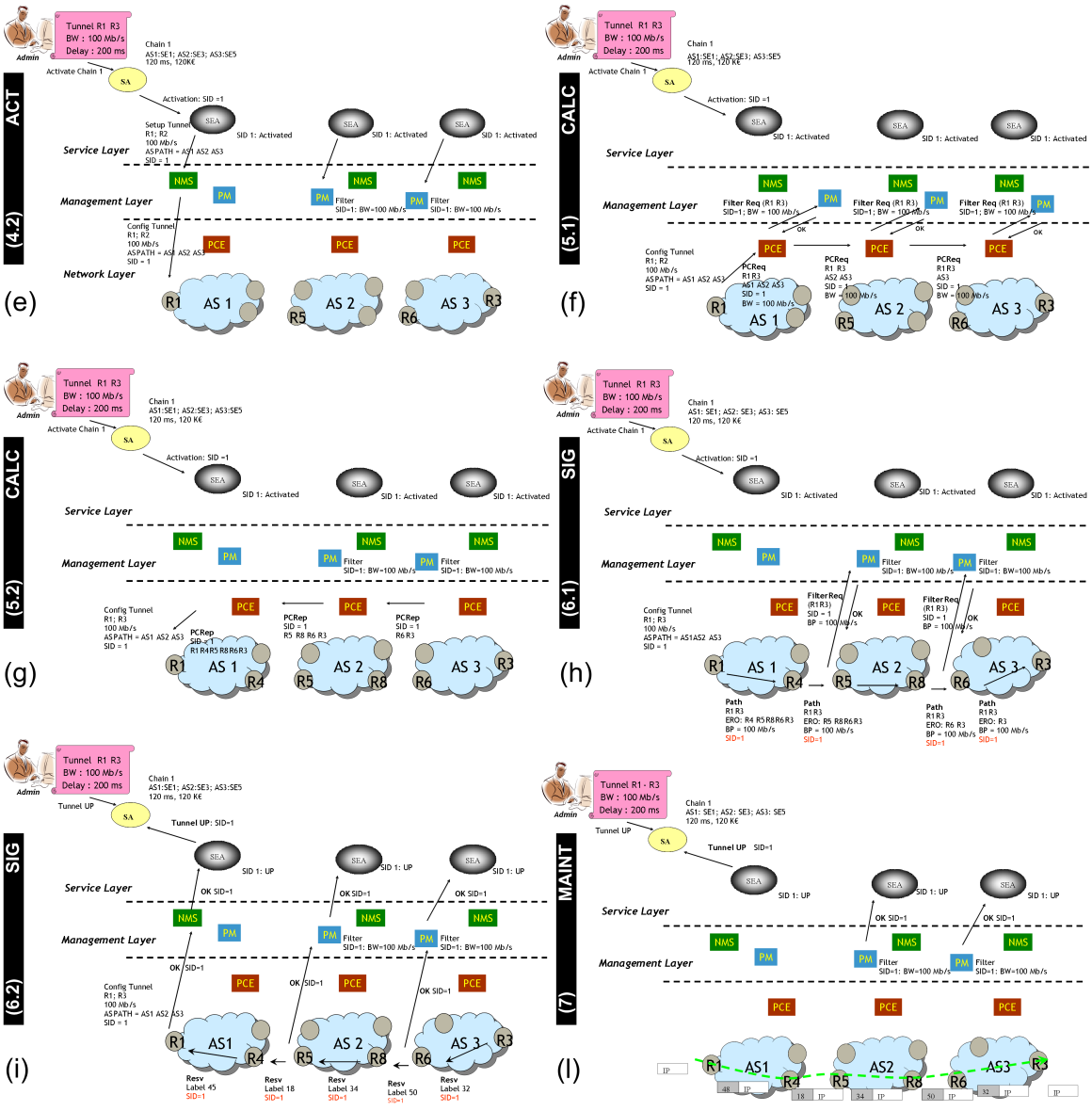


Figure 4.6: Computation and signaling at the management and network layers

novelty we introduce is in filtering the PCEP messages: by default an inter-AS PCEP message is to be rejected by a PCE for obvious reasons of security and confidentiality; it can be accepted only if it transports a SID corresponding to a service that has passed a preliminary activation at the service layer. In [180] a PCEP extension has been proposed to include a SID object. When a PCE receives a PCEP message, it transmits the request to its PM using a Filter Req message. The PM performs the following operations:

- it extracts the SID and the request's parameters;
- it looks for a filter indexed by the SID; if there is not, a PCErr message is sent back to the source PCC;
- if a filter is found, its application entails the deletion of certain objects, the change of others (e.g., the priority or the DiffServ class) or the rejection of

the request if some parameters do not comply with the activated service.

6. *Service Signaling (SIG)*: This step consists in the final signaling of the inter-AS LSP (Fig. 4.6h-i). When the source PCE computes the final path to be employed, it sends a 'PCRep' message to the source PCC containing an end-to-end path towards the destination router, as depicted in Fig. 4.6. This should be a loose path containing, for example, only the border routers to cross. Then the signaling can proceed as explained before, using as ERO object this loose inter-AS path, resolved locally via the confidentiality key. The novelty we introduce is in filtering the inter-AS RSVP-TE messages. The RSVP-TE 'Path' and 'Resv' messages should be extended to transport the SID [174]. Employing the SID, the ingress router of each AS queries the local PM to perform the needed filtering operations according to the instantiated service. When an ASBR receives a Resv message, it sends an OK message to the PM with the SID. The PM then decrements the bandwidth allocated to the service (in Fig 4.6i the LSP uses all the negotiated bandwidth, so there is no remaining bandwidth for the service).

7. *Service Maintenance (MAINT)*: Once the inter-AS LSP has been established (Fig. 4.6l), the events that can happen are the failure or the closing of the LSP. In case of failure, re-routing or re-provisioning operations should be executed. If a specific protection strategy was chosen at the corresponding service element, it should be implemented. Whether the failure happens on intra-AS links or routers, the recovery should not involve the service plane. If a failure on intra-AS equipment cannot be recovered, or if a failure happening on inter-AS links cannot be recovered by rerouting the LSP on an alternative path between the two involved ASs, a Status NOK message is sent to the service plane, and then the source SA is notified and should proceed with a new service request.

4.6.3 Dealing with Collateral Behaviors

It is evident that if a provider is not able to guarantee the SLS, or block the resources without paying further on, it should be penalized. Moreover, if a provider perturbs with unidentified inter-AS PCEP and RSVP-TE messages, or advertises service elements that show to be unavailable the most of the time, it should be penalized.

Furthermore, it is worth noting that collateral business behaviors may appear within the alliance, which can still be correct from an alliance standpoint. First, it is still possible to hide private bilateral agreements. Second, it is possible to prune or weight competitors' service elements. Third, an AS may avoid instantiating elements by blocking requests involving a specific AS. These behaviors may be detected locally, but this may not be sufficient. In order to automatically isolate them, some transaction statistics might be shared at the service plane (but also outside it in a parallel control architecture) by the providers (or also by a subcoalition of providers). This data shall contain the level at which the behavior is detected: at the service layer during the composition or instantiation, at the management layer during the message filtering, or at the network layer. By sharing this data, a SA can instruct its ASA with information on how to prune and weight some service elements. These transaction statistics may, for instance, be collected by a shared service broker whose data is populated by the local SA whether they have to report a proved incorrect behavior.

The rationale is to freely allow indirect regulation schemes among providers, in which the reputation or business reliability of a provider within the alliance is informally defined

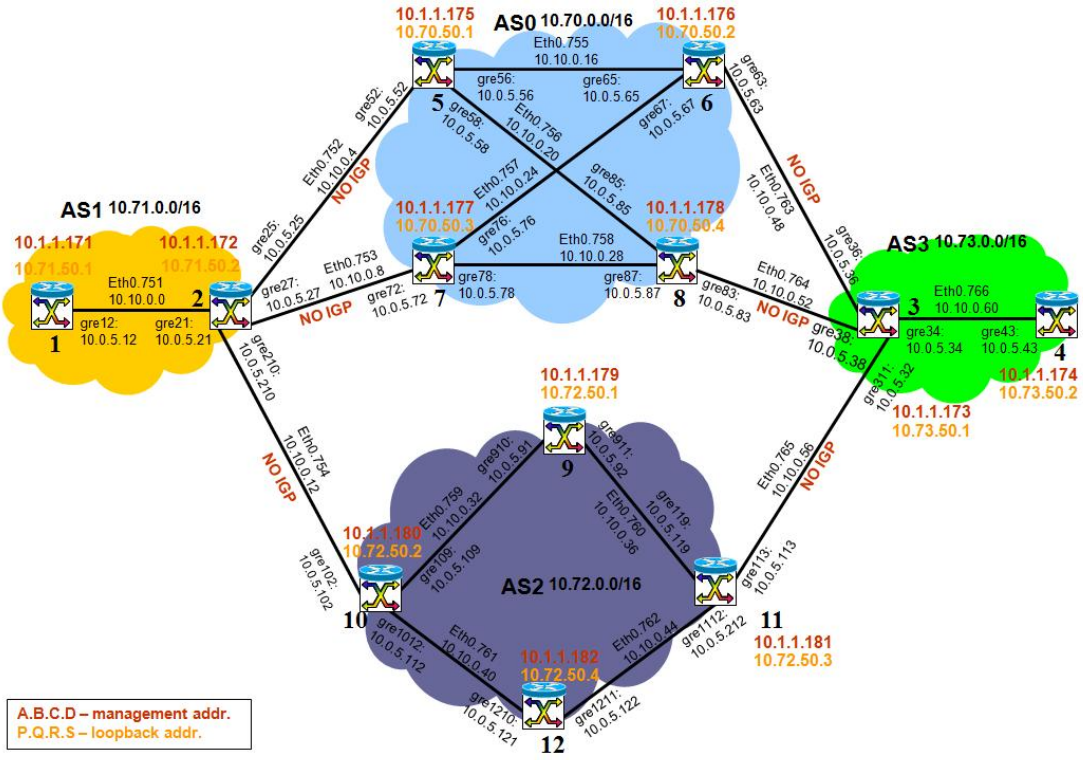


Figure 4.7: The reference topology

by its past behavior (guaranteed SLS, signaling perturbation, excessive policy blocking). With respect to an excessive blocking of service elements, it may not easy to detect it (a-posteriori data mining techniques upon transaction histories may be conceived). In general, an alliance member that isolate other providers' service elements for its clients' requests (second point above) could act to meet clients' requests and would not behave against the alliance cohesion. On the other hand, an alliance member not instantiating its service elements for external LSP requests (third point above) may be considered as a provider with an anti-alliance behavior. Therefore, hidden and/or bilateral routing policies within the provider alliance are still possible, but there is a alliance cohesiveness tension to care of when rejecting service element requests.

4.7 Testbed Implementation

We worked at a testbed implementation of the CALC and SIG blocks of the proposed architecture. Hence, we focus on the network plane and its interaction with the management and the service planes. We emulated the topology in Fig. 4.7. The testbed has been built upon the existing CTTC ADRENALINE+ (All-optical Dynamic REliable Network hAndLINg IP/Ethernet Gigabit traffic) testbed, presented in [158] and [159]. On this existing testbed, there was no EGP (Exterior Gateway Protocol) implementation available since it has originally been conceived for intra AS uses only. Hence, in order to implement a multiple AS scenario and inter-AS links, we decided to insert static routing information on the ASBR in order to guarantee the inter-AS reachability. The main implementation issues have been:

- The extension of the starting RSVP-TE implementation for the multi-AS environment;
- The inclusion of the SID object into RSVP-TE, and its management;
- The extension of the existing implementation of BRPC from the intra-AS inter-area to the inter-AS scope;
- The inclusion of the SID object into PCEP, and its management;
- The PCEP - RSVP-TE interworking aspects;
- The implementation of a separate policy management module to implement the policy manager functionalities.

4.7.1 RSVP-TE extension

Our first focus has been on issues related to inter-AS path signaling itself. As a first step, we thus deployed a scenario using per-AS path computation instead of the cooperative PCE-based approach. With per-AS path computation, we need to cope with the inter-AS visibility issue - that is, any node can only count on the local knowledge of its routing AS, given by the OSPF (Open Shortest Path First)-TE protocol. The initial Explicit Routing Object (ERO) used by the inter-AS RSVP-TE Path message only contains the strict list of unnumbered interface ID subobjects (nodeID, interfaceID) down to proper egress ASBR and then the final destination nodeID as a 'loose' subobject. Once the RSVP-TE message reaches the ASBR of the AS, the ERO (containing at this point only the destination as 'loose') has to be expanded once in order to include the next hop as 'strict' (nodeID of the ingress ASBR of the next AS). At the ingress ASBR of the following AS, the ERO has to be expanded again to include the strict list of subobjects down to the proper egress ASBR of the actual AS (or to the final destination in the case of destination AS), and so on up to the destination. This iterative procedure is depicted in Fig. 4.8 for a TE-LSP signaling with RSVP-TE from node N1 to node N4. Every node receiving an RSVP-TE PATH message erases itself from the contained ERO, and then looks for the next strict hop (if present), or expands the ERO with the above explained procedure, in the case of an ASBR. This issue was not considered on the existing testbed implementation of RSVP-TE since it was designed for multi-area, intra-AS scenarios. Indeed, in that case it was possible to rely on the area border routers belonging to multiple areas and thus having visibility on the topology of each area belonging to; such nodes are thus able to expand the ERO of the received packets that have to cross an area boundary.

RSVP-TE extension with the Service ID object

It is needed that RSVP-TE transports the SID object in order to apply policy management during the LSP establishment. The SID allows discriminating between authorized inter-AS RSVP-TE Path messages and unauthorized attempts, guaranteeing that only the first ones are processed, while the others are rejected. The SID values are local to each AS, and are first exchanged during the aforementioned service instantiation phase. Hence, a multi-AS service is identified by many local SIDs, one per AS[3]. The implemented SID computation rule is a function of the LSP Identifier (LSPID) and the destination node; then, the SID values are to be coherently checked at the reception.

[3]We previously depicted a single SID at the service layer for the sake of clarity.

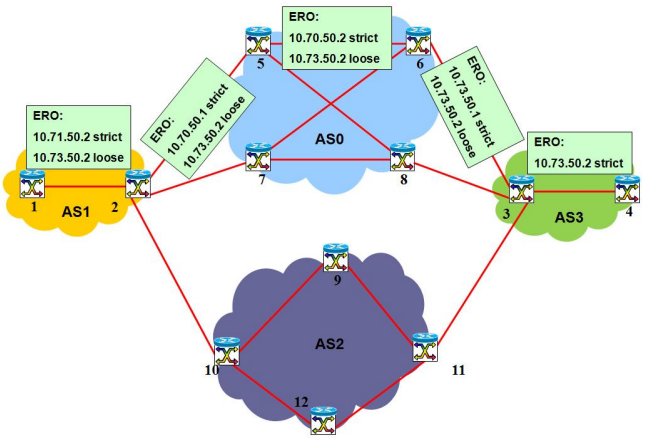


Figure 4.8: Example of ERO expansion for a path from N1 to N4

4.7.2 BRPC algorithm extension to the inter-AS scope

The BRPC path computation procedure was implemented in the ADRENALINE testbed. However, it was designed for the multi area, intra-AS scenario case only, with a PCE in each area. In this environment, border routers belong to multiple areas and have interfaces on each one while, in the multiple AS environment, the ASBRs are part of a single AS with the inter-AS links not belonging to any AS, neither announced by any IGP instance.

In order to overcome this issue, when an upstream PCE receives a tree of paths routed to the destination from the following downstream PCE, every branch starting at a given downstream ASBR is mapped to the proper inter-AS link and then to corresponding egress ASBR of the upstream AS. To do so, we use the aforementioned static information regarding the inter-AS links in order to extend and prune the BRPC tree branches in a first step. In this way, the tree of paths can be treated by the current PCE exactly as if we were in a multi-area context, reusing the same implementation of the BRPC algorithm. This approach supports the presence of multiple parallel links between AS pairs: if an incoming path can be mapped to more than one of the ASBRs in the upstream AS an extended path to each of the ASBRs including its (nodeID, interfaceID) is created. All the created paths are then added to the solution set, which is sorted using the total path metric so only the best is kept, while the others are pruned consistently with the BRPC algorithm. Fig. 4.9 depicts an example of the described procedure for a path computation from node N1 to node N4 with an AS chain equal to (AS1, AS0, AS3).

4.7.3 PCEP extension with the SID object

According to the proposed architecture, and in order to be able to apply policy management during the path computation, every PCEP PCReq message has to carry a SID object and every PCE has to check it in order to validate/reject a received path computation request: it allows discriminating between authorized PCReq messages (to be processed) and unauthorized attempts (to be discarded). The SID computation is performed similarly than for the inter-AS RSVP-TE extension. The PCEP implementation has been extended to include the object with proprietary Ctype objects. Fig. 4.10 shows the SID object as seen in a capture of a PCEP PCReq by the PCE1 for the PCE2, concerning a path computation request for a path going from node N1 to node N4 (the

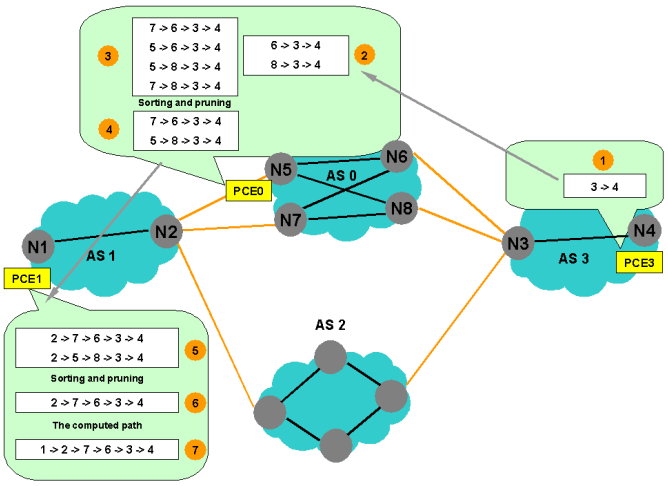


Figure 4.9: Received ERO mapping procedure

capture has been obtained using a version of the Wireshark tool properly extended in order to be able to correctly decode the new object). As highlighted by the zoom, the SID object contains the value assumed by the SID for the particular request.

4.7.4 PCEP and RSVP-TE interworking

When a node requests an LSP, it queries the local PCE for the ERO to be used in the RSVP-TE path signaling procedure. The ERO is computed by the PCEs with the above-mentioned procedure and returned to the requesting node. This allows forcing policies on the followed path differently from the case of simple OSPF-TE use. Moreover, this allows a global path computation on a specific AS chain given by the service plane. Hence, now the optimal 'strict' ERO (i.e., the complete list of sets [nodeID, interfaceID] representing the crossed nodes in the computed constrained path from the source toward the destination) is returned by the PCE and used by RSVP-TE to reserve a path and establish the LSP.

4.7.5 Policy management module

Meeting the policy requirements previously outlined (similar requirements are now formalized in [176]), the policy management functionalities have to be implemented in a module, named Policy Decision Point (PDP). This module, running on an external node, has to reply to the following types of request:

- SID request by PCC: a PCC requesting path computation asks to the PDP for the SID to be used in the request.
- SID and TE parameters check request: a PCE required for a path computation asks to the PDP if the parameters included in the request are coherent with the negotiated ones. This includes a control on the LSP ID[4]; it must belong to a negotiated interval, i.e., the TE parameters (in our testbed we used just the bandwidth) must correspond to the negotiated ones for the LSP ID; moreover, the SID must follow the right computation rules.

[4]in our testbed implementation, LSP ID and Tunnel ID take the same value, even if normally can take different values.

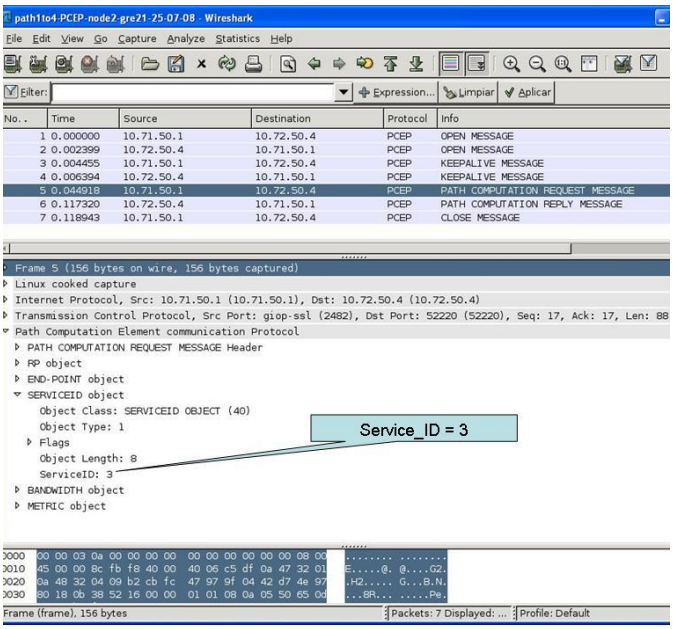


Figure 4.10: Wireshark capture of a PCEP packet with the SID object

- SID request for RSVP-TE PATH message: a node sending a RSVP-TE PATH message asks to the PDP for the SID to be used in the PATH message.
- SID check for RSVP-TE PATH message: a node receiving a RSVP-TE PATH message asks to the PDP if the SID used in the PATH message is consistent with the negotiated one. This check includes the control on the LSPID already mentioned above.

An ad-hoc protocol has been designed to handle the communications between the PDP and the different Policy Enforcement Points (PEP).

4.8 Summary

Nowadays, the MPLS-TE technology is largely deployed within providers' boundaries to support real-time and interactive services. The extension of these services at the multi-AS scope requires supporting inter-AS QoS guarantees between providers. The IETF has worked on extending existing protocols and architectures required to set-up inter-AS connections. These extensions are referred as inter-AS MPLS-TE/G-MPLS technology. However, some missing blocks are needed in order to automate inter-AS services. In this chapter, we first outlined these missing blocks and in particular the importance of a service plane. In the context of a provider alliance where TE connections are established between the members of the alliance, we defined the notion of inter-AS MPLS-TE service as a composition of service elements. We then proposed a comprehensive architecture based on three planes: service, management and network planes. We outlined the roles of each plane and showed how they can interact, showing how the inter-AS provisioning can be automated.

We performed a testbed experimentation of the required control plane extensions. We extended an existing testbed to the inter-AS scope, integrating the service architecture-related data (AS chain, "Service IDentifier"object and inter-AS service TE metrics) and the required protocol extensions (RSVP-TE and PCEP extensions to carry the SID object). The main achievements are the integration, the testing and the validation of the Service IDentifier object in the PCEP and RSVP-TE, and the related policy management.

Chapter 5

AS-level Routing in Provider Alliances

In this chapter we develop the service element composition functional step (COMP) of the provider alliance architecture defined in the previous chapter. The composition of service elements represents a peculiar routing problem at the AS-level that we define and for which we propose ad-hoc scalable algorithms[1].

5.1 AS-level routing requirements

The provider alliance architecture arises specific routing requirements for COMP QoS algorithms (obviously, 'QoS' implies multiple constraint support):

1. *Policy routing*: the source AS shall be able to apply local policies to influence the inter-AS route local selection, while having the highest possible visibility on inter-provider AS-level routes (i.e., it needs to have all the service elements);

2. *Directional metrics*: a shortest path algorithm with multiple constraints should deal with a graph weighted with service elements' parameters, which are directional in the sense that they are applied to a 3-tuple ingress node - transit AS - egress node, where the nodes can represent either ASs or neighbor ASs' ASBRs or group of ASBRs;

3. *Pre-computation*: the presence of computation servers, the PCEs, may allow reducing the online time complexity of the routing algorithm by pre-computing a part of the job;

4. *Multipoint routing*: so as to cope with any class of inter-provider service, the AS-level routing algorithm should encompass both point-to-point and point-to-multipoint inter-AS routes;

5. *Route diversity*: for each request, the source AS shall select a set of possible routes with a certain degree of diversity because of at least two reasons:

    - to decrease the request blocking probability by sequentially testing feasible routes that do not share critical common paths, so as to avoid inter-plane COMP $\rightarrow$ INST $\rightarrow$ COMP signaling loops;

[1] The contents presented in this chapter are also presented in [1], [11], [13], and [14].

- to offer diverse paths for service requests with a certain degree of reliability so as to provide path protection mechanisms.

Let us discuss the pertinence of each requirement (but the last, discussed further on) and the algorithmic implications.

5.1.1 Policy routing

The first requirement implies that the routing decision is not distributed. This guides toward a source-based algorithm, executed at the source AS that disposes of the required routing information (in particular, all the available path, hence a distance-vector like algorithm would reduce the available routing information). Based on this information, following the considerations given in Sect. 4.6.3, an AS might apply local policies, potentially hidden to the other ASs, and influence the routing decision by pruning the graph. The importance of the first requirement brought to the definition of a functional policy architecture in the recent RFC 5394 [176] (related to the PCE architecture), which states: "Network operators require a certain level of flexibility to shape the TE path computation process, so that the process can be aligned with their business and operational needs. Many aspects of the path computation may be governed by policies.". The idea is to let providers maintain a level of arbitrarity in the routing choice similar yet broader than that granted by the local preference in BGP routing.

5.1.2 Directional metrics

Within the service architecture described in the previous chapter, via the Service Elements each AS announces different transit costs and capabilities as function of both the entry and the exit ASs or ASBRs. Upon arrival of a request, an ASA employs this information to compose the service elements.

The second requirement specifies that the adopted routing algorithms should deal efficiently with directional metrics. There are two possible ways to meet this requirement, either one executes classical constrained shortest path heuristics on the pruned graph, or one designs a *search algorithm* to explore the graph following the metric directions. In the first case, in order to deal with directional metrics, the original graph should be extended, as depicted in the example of Fig.5.1: each AS node is to be exploded in a number of virtual nodes equal to the number of neighbors it is connected to. Then, directional metrics are to be applied to simple arcs connecting these new virtual nodes, while null metrics are to be applied to arcs connecting virtual nodes related to different originating nodes.

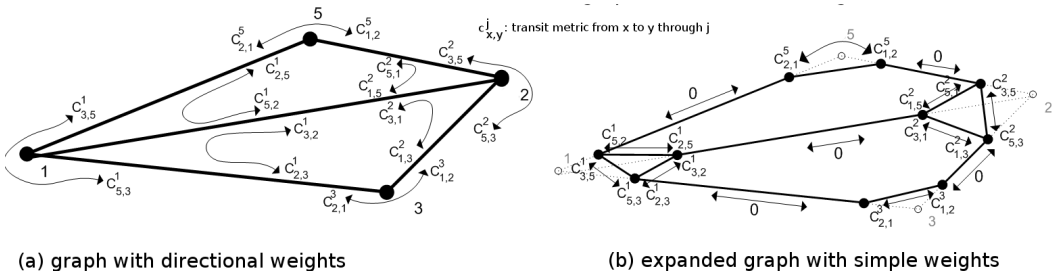


Figure 5.1: Example of graph extension required with classical QoS algorithms

The AS graph having a scale-free nature (i.e., a few nodes attract most of the arcs), those few connected ASs that still occupy a key position in the graph would find in

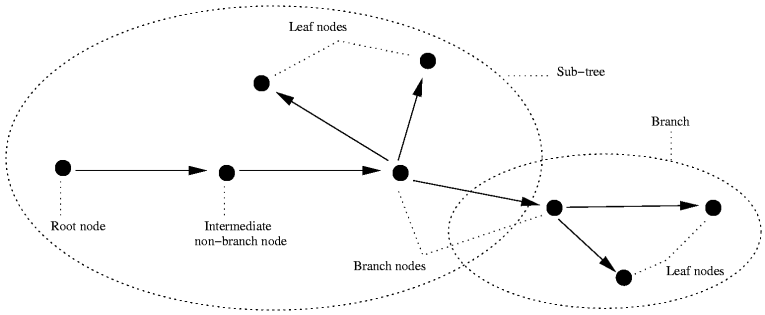


Figure 5.2: Point-to-multipoint tree

directional policies the most proper means to attract connections. We empirically discovered - extracting adjacency information from public BGP routing tables - that in the current AS graph with n ASs, an optimistic approximation for the average degree of an AS-node can be $\sqrt[3]{n}$ (still more optimistic for those hub top-tier ASs that would be likely to participate to a provider alliance). This suggests that the aforementioned extension requires approximately $n\sqrt[3]{n}$ new nodes and arcs, which implies that classical QoS routing heuristics with at least $O(n^3)$ time complexity on normal graphs would pass $O(n^4)$ on an AS graph with directional metrics.

5.1.3 Pre-computation for QoS routing

Common QoS routing algorithms minimize generic link costs while being subject to several constraints. Such algorithms are generally heuristic in that their solution is suboptimal, since the problem is NP-hard. As the number of constraints (additive, multiplicative, diagonal, etc.) used to guarantee certain performance to QoS paths (delay, jitter, bandwidth, protection, etc.) is expected to increase (normally more than 2), pre-computation schemes for QoS routing are highly desirable to reduce the online computational complexity, i.e., the post-request algorithmic complexity [135].

The idea is to let some routing tasks to be performed in advance so as to promptly provide a satisfactory path upon request. In practice, for source routing, one can devise to design an algorithm with a pre-computable initialization procedure independent of the QoS parameters of a path request. In our provider alliance architecture, we dispose of route computation resources potentially available at the PCEs that may be queried for such local pre-computation tasks.

5.1.4 Multipoint routing

Besides backhauling and inter-provider VoIP gateway interconnection, important services requiring inter-provider connection-oriented services are HD video content distribution, e.g., for VoD, Video streaming, telemedicine, teleconference. Such services require point-to-multipoint (P2MP) connections from one source to many destinations. Recent works carried within IETF have extended the MPLS-TE architecture in order to offer point-to-multipoint tunnels (i.e., P2MP TE-LSPs) [168], and have assessed the applicability of the PCE architecture for P2MP TE-LSPs [181].

An agreed taxonomy is needed to identify the elements of a P2MP path, called hereafter *AS tree* (Fig.5.2):

- Root/leaf node: source/destination node of a P2MP data transmission.

- Branch node: node that performs data replication.
- Intermediate node: non-branch and non-root node.
- Bud node: a leaf-and-branch node.

In a P2MP tree, a set of nodes can be classified as:

- P2MP sub-tree: part of a tree such that the root or an intermediate node is connected to a subset of leaves;
- P2MP branch: part of a sub-tree such that a single branch is connected to a subset of leaves.

5.2 Related work and our contribution

In the following, we discuss possible AS tree selection schemes that partially meet the above requirements, highlighting the open path for performance improvement.

5.2.1 Extensions of point-to-point algorithms

In the literature, a number of heuristics are based on the extension of point-to-point algorithms, the most of which can be classified under the following two classes.

Irrespective Routes Computation with Post Merging (IRC-PM)

A simple method relies on the following steps:

- compute the shortest route subject to all constraints for each leaf AS;
- join the subroutes of the routes sharing directional arcs.

We refer to this algorithm with the acronym IRC-PM. The resulting AS tree has sparse branches in sub-optimal positions. It is important to remark that resources (e.g., bandwidth) are shared on common links. Hence it is better to adopt algorithms allowing to reduce the tree cost by encouraging arc sharing.

Iterative Point-to-Point selection (I-P2P)

Breaking the P2MP problem into multiple P2P route selections, inter-AS routes tend to share (directional) arcs:

- compute the shortest inter-AS route subject to all constraints from the root AS to a first leaf AS;
- assign null cost to all directional arcs taken by the first route and compute the inter-AS route to the second leaf;
- repeat the process for every leaf AS.

We refer to this algorithm with the acronym I-P2P. An advantage of this approach is that it still does not require the knowledge of all leaf ASs during the tree computation, while being more sensitive to link sharing than IRC-PM. However, the solution (and its optimality) strongly depends on the order in which routes to leaf nodes have been computed.

5.2.2 Steiner tree

To avoid the dependency on leaf ordering, it is needed to compute the optimal tree that spans all the destinations at once, i.e., the so called Steiner tree [48]. This optimization problem is known to be NP-hard, and is more complex when taking into account additive constraints. The problem not being tractable for large instances, heuristics are needed. Heuristics for the Steiner problem have been studied extensively. A comparison of some of the main heuristics can be found in [109]; the two most promising source-based heuristics are in the sequel considered for the sake of comparison. The first one by Zhu et al. consists in an I-P2P variant, where a constrained version of Bellman-Ford algorithm is used iteratively [110]. The second is the Kompella's centralised algorithm [111], that is:

- compute the all pair constrained shortest paths and build the closure graph of shortest paths from the root to the leaves;
- find the constrained spanning tree of the closure graph;
- expand the spanning tree avoiding possible loops.

The overall time complexity of the Kompella's algorithm is $O(n^3D)$, where D is the integer value of the delay bound. For graphs with directional metrics, the time complexity of this heuristic after the graph expansion would thus become $O(n^4D)$.

5.2.3 Improvement motivations

Looking for a multipoint routing algorithm (requirement 4), supporting policy routing and thus source-based (requirement 1), and which can handle multiple QoS constraints, we fall into the class of source-based multipoint (or multicast) QoS routing algorithms. QoS routing requires the support of multiple metrics to bound the final path solution, some of which are 'multiplicative' (e.g., bandwidth, class of service, etc) and can be easily considered in source-based algorithms by pruning the graph, and some of which are additive (e.g., secondary costs, delay, jitters, hop count, etc). This was an intensive research topic of the 90s, which brought to many possible solutions very well summarized and compared in [109]. The authors clearly point out that the Kompella's [111] and the Zhu's algorithms [110] can be considered as the two source-based multipoint QoS routing algorithms that offer the better performance, especially with respect to time complexity, optimality and QoS constraint multiplicity aspects. Both the algorithms may be adapted to solve our AS-level routing problem and will be considered in the sequel for performance comparison.

We can arise, however, the following problems: such algorithms do not scale with directional metrics (requirement 2) and would thus both assume a time complexity bigger than $O(n^4)$, do not support pre-computation (requirement 3) and can not thus reduce the online time complexity [135], and do not seem to be effectively adaptable to support diversity constraints (requirement 5). In the following, we devise a novel ad-hoc routing algorithm that, instead, better meets the AS-level routing requirements. Its definition passes through the adoption of ideas of first-search approaches, namely the constrained k-shortest path A*prune [115] algorithm, and the usage of a pre-computable subalgorithm, i.e., the any-to-any unconstrained shortest path Floyd's algorithm [116].

5.3 The RCOM AS tree routing algorithm

To solve our specific routing problem we devise an ad-hoc heuristic called Route Collection and Optimal Matching (RCOM), composed of two steps:

1. Route collection: some feasible point-to-point routes towards each leaf AS node are collected
2. Optimal matching: the optimal matching of collected routes is reached minimizing the tree cost.

Differently than IRC-PM, RCOM retains a subset of feasible routes instead of only one route per destination. With respect to I-P2P, RCOM should be more flexible in branch and bud nodes placement, since it can reach a wider set of solutions. Last but not least, in Sect.5.3.3 we show that the more time consuming tasks can be pre-computed before the request arrivals (and independently of these requests).

Algorithm 5.3.1: ROUTE COLLECTION(G)

procedure POP(c, d, h, π)
– $\overline{f}$: per-destination vector with counters of found routes so far
– a, d_a, c_a : next directional arc, delay and cost of a
– M : multicast group (set of leaf nodes)
if $h = H$
then
  if $\exists!$ leaf $d \mid c + SPC(\pi[h], d) < \upsilon(d)$
    then add π to ζ_{cand}
  if $\pi[h] \in M$
    then
      if $c < \upsilon(\pi[h])$
        then
          add π to ζ_{sel}
          $f(\pi[h]) \leftarrow f(\pi[h]) + 1$
          if $f(\pi[h]) \geq F$
            then update $\upsilon(\pi[h])$
else
  for $i \leftarrow 1$ **to** N
    do
      if i adjacent to $\pi[h]$, and $i \notin \pi$
        then
          $\pi[h+1] \Leftarrow i$
          $a \leftarrow (\pi[h-1], \pi[h], \pi[h+1])$
          if $h = 0$
            then POP($c, d, h+1, \pi$)
            else if $d + d_a < D$
            then POP($c + c_a, d + d_a, h+1, \pi$)

main
$H \leftarrow 1$, $\overline{\upsilon} \leftarrow \overline{\infty}$, $\zeta_{cand} \leftarrow \{\pi_0 = (root)\}$
while $\zeta_{cand} \neq \emptyset$ **or** $H < H_m$
do
  extract a subroute π from ζ_{cand}
  POP($cost(\pi), delay(\pi), H-1, \pi$)
  $H \leftarrow H + 1$

5.3.1 Route Collection

To collect the per-destination routes set, we devise an ad-hoc breadth-first-search algorithm with limited depth. It starts at the root, moves to unvisited neighbors, collects

the routes if a destination is attained, and so on, until no longer routes can be collected. It stops at a given number of hops or during the search by pruning branches depending on additive metric and cost bounds.

This approach was inspired by the A*prune algorithm [115], proposed to solve the constrained k-shortest paths problem. Our approach differs from it in that:
(i) since the final objective is the selection of the optimal tree, further pruning (besides that on the additive metrics) depending on the route cost is performed, giving priority to the least hop routes;
(ii) given that there is no need to sort the candidate routes (as A*prune does), the number k of shortest routes is not fixed and all the experienced (feasible) routes are collected (i.e., we do not need a best-first-search approach).

Collection algorithm

Let $\overline{v}$ be a threshold cost vector with one entry per destination. Each entry is a threshold re-calculated for each new route collection. The starting values are infinite. Then, an entry is initialized when at least F routes have been collected for that destination; F has to be chosen conveniently (we use $F = \sqrt[3]{n}$ in our simulations). Each threshold is calculated as the average cost of a subset of the F routes. In order to avoid taking into account the routes with a too high cost with respect to the others, for the threshold computation, within the first F routes we consider only those with a variance on the average cost less than the average of this variance. In this way, the threshold has a decreasing trend, with a starting value not excessively high. The least hop routes are, thus, privileged because the cost bound is higher in the first hops. Favoring routes of a few hops is a suitable approach for our specific problem, since long routes crossing several ASs risk to have a small number of arcs joint with the previously selected ones, and tend to have very high costs. In this way we try to cut a lot of branches that would have been considered by general purpose solvers for an exhaustive optimization.

Definition 5.3.1. A *projected cost* of a subroute is the sum of the current subroute cost and the cost of the shortest path from the tail of the subroute towards the leaf node.

It requires a pre-computation of the shortest paths costs from all intermediate nodes towards the leaves. The pseudo-code is given in Alg. 5.3.1. The search starts looking for feasible routes at 2 hops, then 3, and so on. At every iteration, the search looks only at those routes with equal hop number H, up to a given bound H_m. At every iteration, the subroutes in the set ζ_{cand} are the starting point of the search. At every call of POP(), c and d are the cumulative cost and delay of the route handled by the current route vector π with h hops number. When visiting the root neighbors ($h = 0$), π has only the root, and the delay is not verified. Then, the function recursively visits every neighbor of the subroute tail node, updating π, and evaluating the route feasibility on the cumulative delay. At the H^{th} hop, the route is collected in the set ζ_{sel} if a leaf is visited, if its cost is less than the threshold, and if the delay bound is respected; it is also added to ζ_{cand} for further expanding and possible selection in the next hop only if, for at least one destination, its projected cost is equal to or less than the threshold.

5.3.2 Route matching

The routes in ζ_{sel} define a subgraph built as superimposition of their directional arcs. The optimal tree is thus the minimal composition of directional arcs linking the root to the leaves within this subgraph, solvable via Integer Linear Programming (ILP) with a

low complexity given the limited size of ζ_{sel} and given that there is no need to check the additive constraints any longer. Indeed, forcing each destination to be crossed by at least one route, we assure that the leaves are reached and the constraints are satisfied.

5.3.3 Complexity and pre-computation

The RCOM complexity is thus dominated by the collection algorithm. The majority of the time is spent in computing the (unconstrained) shortest path costs, which are needed to determine the projected costs, in the collection algorithm. We propose to pre-compute them, prior to any request, and after any topological and costs update. This can stand when costs and topology are expected to change much less frequently than the requests arrival, and this hypothesis would apply to the presented multi-provider architecture. Hence, prior to any request (characterized by root, leaves, and end-to-end constraints) a simplified version of the Floyd's algorithm [116] can be used in order to pre-compute the cost of the shortest paths (SPC matrix in Alg.5.3.1) from any node to any node (A2ASP). Floyd's algorithm takes $O(n^4)$ time to compute (see the reasons in Sect. 5.2.2). The subsequent breadth-first search would have, without pruning, a time complexity of $O(n^{\frac{1}{3}H_m})$ for the worst case, approximating the base (branching factor) to $\sqrt[3]{n}$. Because of pruning, it is more efficient than that.

To improve the execution time, A2ASPs computation should be pre-computed, prior to any request, and triggered by topology and costs update. In this way the post-request worst case complexity of the collection becomes $\mathrm{O}(n^{\frac{1}{3}H_m})$.

For the sake of comparison, the centralized heuristics proposed so far for constrained multicast routing, as those in [109], do not have a sub-algorithm independent of the constraint values. For example, the Kompella's algorithm computes *constrained* A2ASPs to build the closure graph with a complexity proportional to the delay bound (see Paragraph 5.2.2). Or, the Zhu's algorithm [110] uses as starting point a least-delay spanning tree. Both Kompella's and Zhu's algorithms have an overall complexity equal to the post-request complexity, which is, for a graph with directional metrics, bigger than $\mathrm{O}(n^4)$ [109].

Note 1: It is worth mentioning that given the breath-first-search nature of the collection algorithm and the additive constraint transparency of the routes matching, an extension of the RCOM approach to multiple additive constraints would scale with the number of constraints (besides the delay).

Note 2: A restriction to a single destination for P2P paths is straightforward and slightly decreases the RCOM complexity: the matching task is trivialized to the choice of the shortest route among the collected ones.

5.4 Performance evaluation I

We compare the described algorithms in terms of optimality and execution time, and analyze the characteristics of the selected AS trees. We chose to use realistic topologies: we dumped the AS whois database containing interconnection data available at [193]. As stated before, our architecture is not meant to be used at Internet-wide scale (even the PCE-based one is not meant to be) but on a set of ASs collaborating to a common service plane. We use Internet topology estimations in order to be as realistic as possible. Two topologies are considered. The first one is built as following: among all the ASs, only those with at least 7 adjacencies are kept (in this way, we select those AS carriers potentially interested in inter-domain tunnel provisioning); then, only those ASs with

more than 2 adjacencies with the other ASs are kept for the final topology. The final topology, called ATL7, has 643 AS-nodes. The second topology, TOP300, is similarly built with the 300 most connected nodes of ATL7.

Capacities and costs

For capacity and cost assignment, we classify as Tier-3 (T3) an AS with a number of interconnections less than the average, Tier-1 (T1) one with a number of interconnections with non-T3 ASs over the average, and Tier-2 (T2) the remaining ones. This deviates from the conventional terminology (see Sect. 2.1.1) which does not apply to our framework since we overtake the BGP-policy-based peering and customer-provider relationships. Moreover, we prefer a degree-based instead of a betweenness-based ranking because this last could not apply, since shortest intra-alliance routes are potentially computed dynamically for each new request.

Considering a T3 not able to offer as much connectivity as T2s and T1s do, and the same for T2s with respect to T1s, we assign capacities to inter-AS links normally with different averages and deviations as indicated in [13].

Moreover, since the bottleneck is not at the intra-AS but inter-AS links, and since lower transit costs come with a higher availability, we approximate the directional transit cost equal to $K\frac{log[\beta min(C_{i,k},C_{k,j})]}{\beta min(C_{i,k},C_{k,j})}$, $K = 10^5$, for a directional arc (i,k,j) with links capacities of $C_{i,k}$ and $C_{k,j}$. We chose this function so that it decreases more than linearly as function of the product between the requested bandwidth and the minimal inter-AS capacity: the more available the transit capacity is, the less expensive the service element is; the more bandwidth is sold, the lower the per-bandwidth cost is. We halve the cost when the transit involves two AS of the same provider, and set it to zero when the threes of them do, so as to try to be more realistic (the per-provider AS grouping is a publicly available information).

Delay bounds

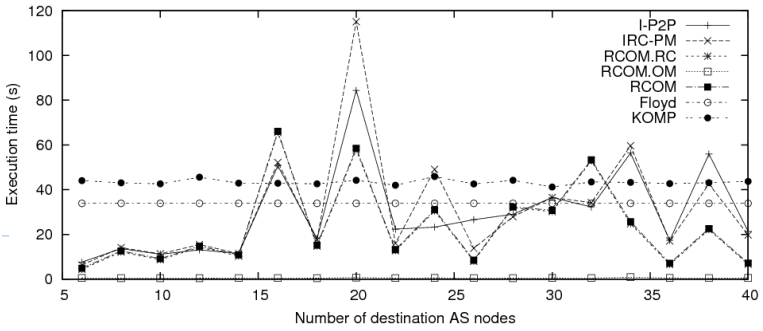
The significant factor affecting the end-to-end delay is the propagation delay [98]. According to the whois tags, we assign ASs to a country. Since carriers can operate in more continents, we calculate the directional transit delay bounds independently of the geographical position of the transit nodes, but as a function of the position of edge nodes, following a normal distribution with averages and deviations chosen on the basis of experimental round trip times (see [13]).

5.4.1 Algorithmic performance

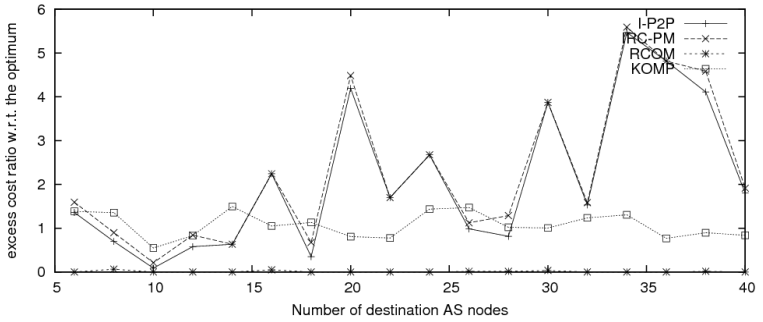
We test the algorithms for different destinations group sizes. Root and leaves are generated randomly. The delay bound is set to 1.5 s and the bandwidth to 6 Mb/s. Simulations run over a 3.4 GHz CPU, with 1 MB cache.

Execution time

Figures 5.3 and 5.4a display the execution times for the TOP300 and ATL7 topology cases as function of the destination group size. For ATL7, $H_m = 8$ (sufficient for this topology); for TOP300, $H_m = 5$ (also sufficient because of the smaller diameter). The optimal solution case is not plotted: it grows more than exponentially with the number of nodes. For RCOM we display: the total time ('RCOM'), the times of the collection



(a) TOP300: Execution times



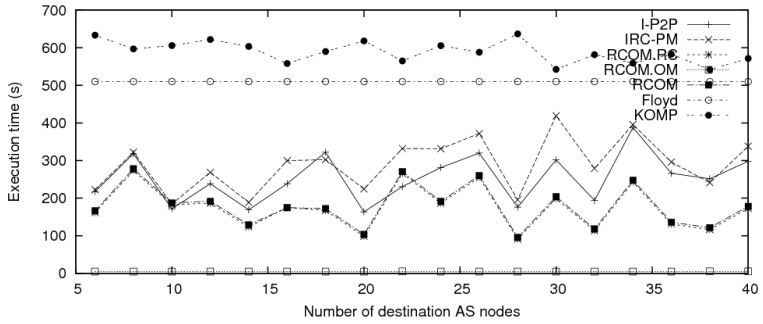
(b) TOP300: Tree cost gaps w.r.t. the optimum

Figure 5.3: Results for TOP300 topology

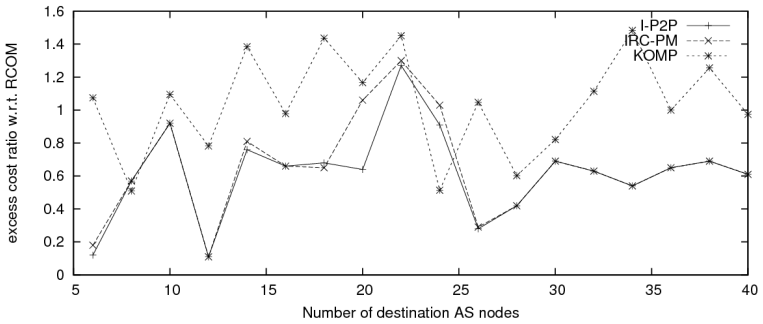
('.RC') and matching ('.OM') procedures. The cases of 'IRC-PM', 'I-P2P' and Kompella's ('KOMP') algorithms are also plotted. The time of the A2ASP computation ('Floyd') is separated since we assume that it can be pre-computed; in fact, it is constant since it is independent of the request parameters. We can assess that:
(i) The complexity of the RCOM route matching part is as more negligible as the topology grows;
(ii) As expected, KOMP is lower bounded by Floyd since it implements a constrained version of Floyd;
(iii) Including the A2ASP computation, RCOM has an execution time comparable to that of KOMP; without, it has almost always the lowest time;
(iv) I-P2P and IRC-PM have a close behavior, and both seem to scale worse than the other algorithms with the destination group and topology sizes;
(v) Larger instances (with more AS nodes) do not worsen the RCOM and KOMP complexity.

Optimality

Fig.5.3b displays the excess cost ratio (i.e., $1 \rightarrow 100\%$) with respect to the optimal solution for TOP300. For ATL7 this could not be computed, but Fig.5.4b displays the excess cost with respect to RCOM for ATL7. We can assess that:
(i) For the TOP300 topology, RCOM yielded an optimality gap largely under the 10%;
(ii) KOMP has always at least 50% excess cost with respect to RCOM;
(iii) I-P2P and IRC-PM give similar solutions.



(a) ATL7: Execution times



(b) ATL7: Tree cost gaps w.r.t. RCOM

Figure 5.4: Results for ATL7 topology

5.4.2 Solution characterisation

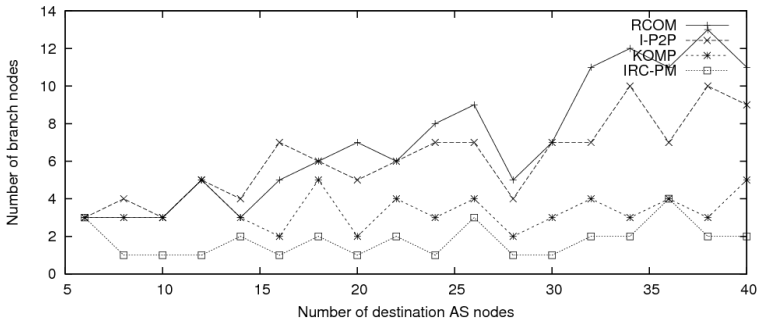
Node type

Fig.5.5 displays the number of branch, bud and intermediate nodes. The ATL7 results are considered. We can assess that:
(i) for RCOM and I-P2P the number of branch nodes increases with the number of ASs;
(ii) the number of branch nodes is lower bounded by KOMP and IRC-PM;
(iii) interestingly KOMP often gives more bud nodes than the other algorithms;
(iv) on the contrary, RCOM often has more branch nodes and less bud nodes than KOMP;
(v) in terms of intermediate nodes, RCOM represents a good trade-off between I-P2P and KOMP.

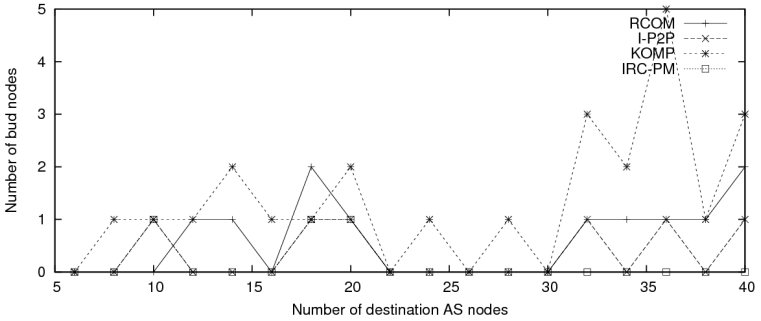
(ii) and (iii) may be explained as follows. While RCOM has an unconstrained A2ASP pre-computation for projecting costs during the constrained exploration and pragmatically discarding routes, KOMP has a constrained A2ASP computation for producing a closure graph where the minimum spanning tree is computed. The KOMP algorithm seems falling easily in local minima corresponding to longer routes. The possibility of branching at leaves is thus higher; indeed, the closure graph is not sensitive to the real hop number.

Tree slimness

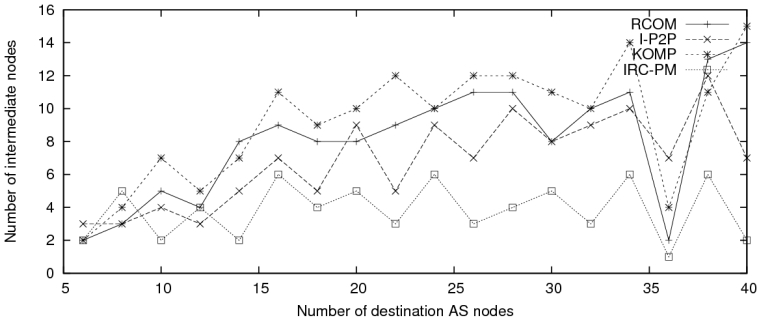
Let the utility of a directional arc be the number of destinations it allows to serve minus one. Let the tree slimness be defined as the ratio between the sum of all these utilities and the number of directional arcs the tree it is composed of. The slimness expresses how much the selected tree is exploited, or how much the selected tree has directional



(a) Number of branch nodes



(b) Number of bud nodes



(c) Number of intermediate nodes

Figure 5.5: Node characterization of the solution tree

arcs that are very used to reach several destinations. This is not intended as an overall evaluation parameter of a tree; however, one can deduce that the less optimal a tree is, the smaller its slimness is expected to be. We are motivated in analyzing this parameter because in multi-layer network, a major application of these algorithms, a computation in one layer can be followed by computations in other lower layers along the routes chosen in the upper layer. Hence, the slimmer the tree is, the simpler the under-layer path computation (and maybe signaling) might be in the case of multi-layer networks.

Fig.5.6 displays the slimness of solution trees obtained for the ATL7 graph.

We can assess that:
(i) RCOM offers the best slimness, i.e., the better utility of the tree;
(ii) KOMP offers the worst slimness;
(iii) I-P2P and IRC-PM behave better than KOMP but worst than RCOM.

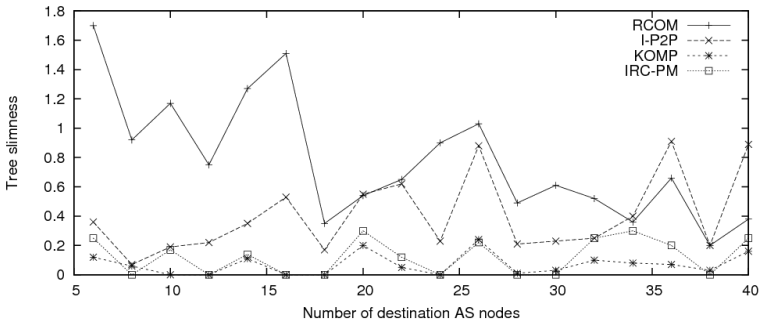


Figure 5.6: Solution tree slimness as function of M

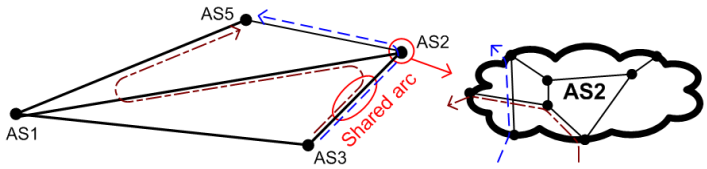


Figure 5.7: Example of two diverse inter-AS routes.

5.5 Route diversity in AS-level routing

We now deal with the last requirement in Sect. 5.1, the route diversity. As previously mentioned, a set of route alternatives (P2P AS paths or P2MP AS trees) should be selected to offer enough diversity for a successful route selection, or to set-up disjoint tunnels for protection purposes. For the first case, the route alternatives should be computed and tested one after the other to avoid signaling loops between the COMP and INST steps. For the latter case, it is possible to compute disjoint AS paths or AS trees sequentially with RCOM, by collecting in the RC step only those paths or trees that are disjoint with the first one. However, this can lead to blocking if the second path cannot be placed. Such issues can be more readily avoided if the set of route alternatives are computed in parallel [181]. In the following, we first concentrate on the P2P diverse route selection problem, the extension to P2MP routes (AS trees) being straightforward.

Definition 5.5.1. A *directional arc* denotes a succession of two inter-AS logical arcs linking three AS-nodes.

Definition 5.5.2. *Diverse routes* do not share any directional arc.

As depicted in Fig.5.7, forcing a directional disjointness, two route alternatives may concern the same AS-node, but involve different intra-AS directions and different inter-AS links. Note that only one route may be instantiated because of intra-AS resource availability. Indeed, different inter-AS directions can have a different intra-AS resource availability. We believe that such an AS-level disjointness constraints is more pertinent than other forms such as end-to-end disjointness. End-to-end disjointness at the AS-level would be, indeed, very hard to achieve. When two ASs are connected with a single inter-AS link, the end-to-end disjointness may not be guaranteed: this would be the case for the most part of the AS-node pairs in the Internet graph given its scale-free nature. In fact, the directional disjointness constraint allows exploiting the scale-free nature of the AS graph, which presents a few AS hubs interconnecting a lot of ASs.

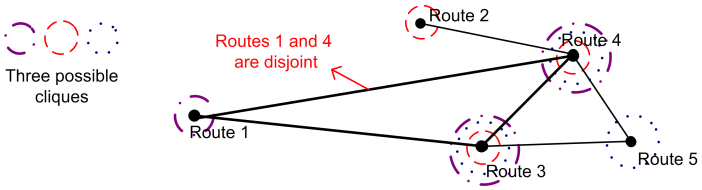


Figure 5.8: Three possible cliques of 3 diverse routes in a 4-route graph

5.5.1 Diverse AS-level routing problem

The diverse routing problem consists in selecting the less costly set of diverse routes satisfying a given connection request. The set of feasible routes ζ_{sel} can be collected with the Route Collection algorithm presented in Sect. 5.3.1. Then, a given number of diverse routes is kept, that is, a clique of diverse route has to be selected within ζ_{sel}.

Optimal clique selection

The next step consists in extracting the least cost clique of a diverse (collected) routes. Every route-element of ζ_{sel} has a cost and can be included in the final clique. We can see every route as a vertex, so that the least cost clique of vertices is the solution. In Fig.5.8, e.g., we have a 5-route graph from which only 3 cliques of 3 vertices can be extracted. This problem is linked to the Generalised Minimum Clique Problem (GMCP), with a fixed clique size. The routes of ζ_{sel} are considered as vertices, which are connected only if diverse.

The optimal clique selection sub-problem can be solved by ILP. The GMPC considers weighted vertices and links, and is NP-hard [49]. In our case only vertices have cost, and it becomes a node-weighted MCP, which is still not polynomial, but less complex and treatable for a few hundreds of routes. Let f_i be a binary variable equal to 1 if $i \in \zeta_{sel}$ is a clique member. The formulation is:

$$\min \sum_{i \in \zeta_{sel}} f_i c_i \tag{5.1}$$

$$s.t. \sum_{i \in \zeta_{sel}} f_i = a \tag{5.2}$$

$$(a-1) f_i - \sum_{j \in (\zeta_{sel} - \{i\})} f_j s_{i,j} \leq 0 \quad \forall i \in \zeta_{sel} \tag{5.3}$$

$$f_i \in \{0,1\} \quad \forall i \in \zeta_{sel} \tag{5.4}$$

The objective (5.1) is the minimisation of the clique cost. (5.2) sets a routes for the clique. (5.3) forces the clique membership. (5.4) sets the f binarity.

5.5.2 About route diversity for multipoint paths

As already mentioned, the extension of the AS-level routing algorithm to deal with the selection of several diverse AS trees is straightforward and not included. Two AS trees shall be considered as diverse if they do not share any directional arc. Please consider that, however, such a disjointness constraint may be too strict especially for small AS graphs with a few hubs. In such cases it might make more sense to consider as diverse the AS trees that do not share branch nodes, which may also decrease the computational complexity of the optimal matching step.

	$< 5\%$	$<50\%$	$< 100\%$
$a = 2$	80%	93%	99%
$a = 4$	75%	80%	99%
$a = 16$	69%	77%	96%

Table 5.1: RECS optimality evaluation.

5.6 Performance evaluation II

In order to test the diverse AS-level routing algorithm - nicknamed RECS (Route Enumeration and Clique Selection) in the following - on very large instances, this time the AS graph is built considering among all the ASs only those with at least 4 adjacencies (instead of 7 - then again only those with more than 2 adjacencies within the selection are kept in). The final graph has now 1716 ASs.

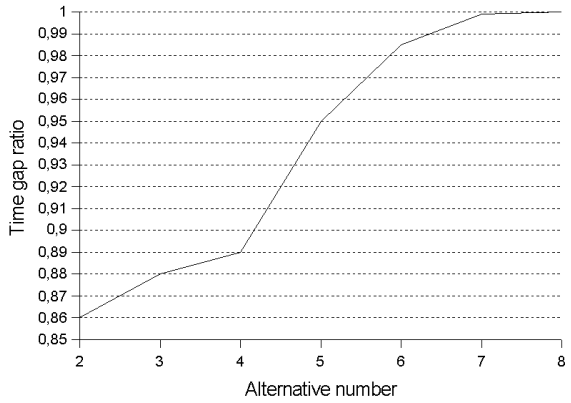
5.6.1 Algorithmic performance

Time complexity

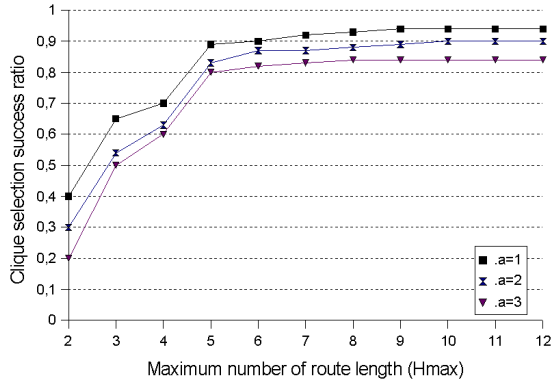
Fig. 5.9a displays the average execution time gap ratio between the proposed approach for diverse route selection (RECS) and the optimal result that could be obtained by ILP (CPLEX). The ratio is simply computed as $1 - t_{RECS}/t_{ILP}$, where t_{RECS} and t_{ILP} are the execution times under the two approaches; the results are displayed as function of a. 50 successful simulations are considered. Fig. 5.9a displays two curves: the dotted one considers the A2ASP computation time in t_{RECS}, while the continuous one does not. Indeed, the ASAs should compute the A2ASPs off-line prior to inter-AS route requests. The higher the number of alternatives is, the harder the optimal approach: the RECS approach scales with the number of alternatives. Indeed, given that the number of collected routes remains always under 1000, the clique selection requires only a few solution searches. Obviously with the A2ASP computation time we have just a shift.

Optimality

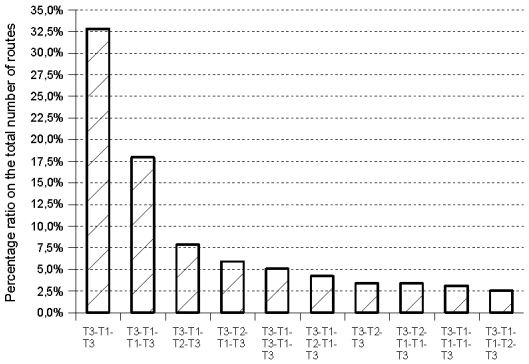
We compare the average deviation (of the selected clique cost) of RECS and the optimal approach. Each entry of Table 5.6.1 indicates how many of the performed simulations per case produced a solution with an optimality gap within 5%, 15% or 100%. Three cases are considered for 2, 4 and 16 route alternatives in the clique, with 50 simulations per case. For each case we show how often (in percentage) RECS solutions had an optimality gap that falls in the three intervals. We can assess that:
(i) RECS can give a solution less with an optimality gap within 5% more than once every two times;
(ii) it can guarantee a solution with an optimality gap within 100% for practically all the requests;
(iii) better optimality gaps are obtained with a small number of route alternatives, even if large numbers of alternatives still allow reaching an optimality gap within 50% the most of the times.



(a) Time gap ratio ILP - RECS



(b) Success ratio performance



(c) Hierarchical routes ratio

Figure 5.9: Simulation results

5.6.2 Solution characterisation

Connection admission

Fig. 5.9b reports the success ratio in selecting a clique for sizes $a = 1, 2, 3$, as function of the upper hop bound, and with 50 new simulations per case. We can assess that:
(i) the most part of ASs is attainable within 5 hops;
(ii) the exploration of the graph for more than 8 hops is not useful;
(iii) even for single-element degenerate cliques, a 100% success ratio was never reached because the bandwidth and the delay constraints limit the number of collected routes.

AS hierarchy

Fig.5.9c reports the 10 most selected hierarchical routes, for 100 new successful simulations with a hop bound of 8. We can assess that:
(i) all the routes have T3s as source and destination ASs;
(ii) more than 80% of routes count less than 5 hops;
(iii) a significant part has only T1s transit nodes, while the others use at least one T1.

Thus, less than 0.1% of ASs (the T1s) attract the most of the traffic. Such results prove that assuming, as we did, a carrier hierarchy where top-tier ASs dispose of more resources and can apply lower prices, the economically feasible routes are attracted by top-tier ASs. This does not preclude, however, a lower-tier AS to attract more routes if it can tune transit rates efficaciously.

5.7 Summary

In this chapter, we proposed heuristics for the AS-level source-based routing and diverse routing problems, as possible algorithms to by implemented in the COMP step of the provider alliance architecture. We highlighted the peculiar AS-level routing requirements and position our contribution with respect to the state of the art.

We have showed that with our heuristics, pre-computation of some tasks can be performed, which drastically speeds up subsequent routing computations at tunnel request arrivals. All in all, by means of extensive simulations, we argued that:
(i) exploiting pre-computation, our approaches are faster than the well-known algorithms;
(ii) multiple additive constraints do not affect the asymptotic time complexity;
(iii) they often reach the optimality and have a gap always largely under 10%;
(iv) they produce efficient trees with respect to under-layer computation issues;
(v) AS-level diversity constraints can be included in the routing algorithm, and their consideration does not decrease optimality and computational performances.

Chapter 6

A Cooperative Framework for Cross-Provider Resource Reservation

The cooperation in the provider alliance architecture is limited to the joint management of service elements and to extended computation and signaling schemes for the inter-provider TE service configuration. The service element availability shall be assured by a form of static cross-provider resource reservation so as to guarantee them a seamless instantiation. Indeed, having inter-provider resources reserved statically, then dynamically assigned to LSPs, would prevent from excessive crankbacks during the INST phase of the provider alliance architecture.

Within the same alliance, providers can still compete, and the outcome of the competition can be the formation of sub-coalitions (derived from reciprocal resource reservation levels), which strategically could mine the alliance inner trust. In order to guarantee stability to the alliance, these strategic situations shall be managed properly and fairly. The cooperative game theory (whose principles are resumed in Annex A) offers, by modeling inter-coalition binding interactions, strong mathematical solutions that can provably be fair and possess other desirable properties. In this chapter, we refine a multi-domain distributed resource reservation framework at the state of the art. In order to make it economically feasible in a strategic multi-provider scope (rather than in a flat multi-domain one), we adapt it by considering the Shapley value power index as a means to distribute the inter-provider service income[1][2].

6.1 Related work and motivations

In [126] and [127] a generic multi-domain routing problem is defined and possible decomposition methods are discussed. It consists in the optimization of bandwidth reservation levels on inter-domain links for per-destination traffic flows, possibly identified by traffic classes.

In [128] it is shown how to decompose the problem with respect to individual domains using sub-gradient optimization based on Lagrangean relaxation, and it is demonstrated how to solve an inter-domain routing optimization problem using a distributed process based on sub-gradient optimization combined with recovering of near-optimal bandwidth

[1] The contents presented in this chapter are also presented in [8], [19], [22], and [23].

[2] This work has been conducted in the framework of the Euro-NF INCAS research activity, the Institut Telecom (Networks of the Future Lab) I-GATE project and the CELTIC/EUREKA TIGER2 project.

reservation levels. For further details please refer to the original papers by Tomaszewski, Mycek and Pióro [126] - [130].

Besides the unique technical framework allowing to decompose the master optimization problem, the useful outcome for our model is the optimization result, i.e., the reservation levels for inter-domain links, on a per-flow basis, where flows can be aggregated at the AS-level. For the sake of clarity, in Sect. 6.1.2 we then report the mathematical notations re-adopted for our model.

6.1.1 Rationales

We are thus considering an interconnection scenario in which multiple domains interact to optimize link reservation levels, to improve their *multi-domain* routing, escaping a solution guided by unilateral and selfish choices toward a more effective global solution with cross-provider resource reservation. If the interconnection principle is undoubtedly attractive, and the proposed approach shall be suitable, the incentives are still not obvious in a *multi-provider* scope. In this chapter, we investigate this aspect and propose a game-theoretic income distribution scheme based on the Shapley-value concept [39] that shall motivate the adoption of the proposed resource reservation framework in the multi-provider scope.

Adopting the distributed multi-domain optimization approach of Tomaszewski et al., the global routing solution is likely to improve with respect to multi-domain throughput and load distribution efficiency. Nevertheless, it is likely that, in the corresponding optimal global solution, the computed reservation levels arise disparities among domains. Still acceptable when the involved domains belong to a same provider network, such routing disparities would decrease the reciprocal trust among individual providers. Eventually, the alliance agreement may not be settled for the lack of fair incentive schemes.

The framework of Tomaszewski et al. may be considered as a sort of extended peering agreement from which providers obtain mutual benefit without side payments. However, for such frameworks the agreement is binding (cross-provider resources are reserved in fact) and thus shall rely on side payments since the multi-provider optimization can arise disparities: in order to reserve bandwidth for external connections for which no direct earning is obtained, a provider may need a form of economical incentive. It is indeed possible that, by reserving bandwidth for external connections, a provider grants earnings to its 'peers' bigger than the earnings related to its own services. Instead of 'extended peering agreement' it is in fact more appropriate to refer still to the provider alliance concept, in which the collaboration include a cooperative framework with cross-provider QoS and availability guarantees.

It is thus needed to define a fair scheme for multi-provider cooperation in a way that it is not solely based on the generated traffic or absorbed traffic (i.e., Content or Eyeball behaviors, see [87]) but also accordingly to its *alliance transit contribution*, i.e., that takes into account how much a provider supports the services of the other providers allocating its network's resources. The originally conceived scheme of Tomaszewski et al. relies on the current Internet business model, with transit agreements settled between domains, wherein a provider income is supposed to derive from the Internet connections provided to its direct clients only. The idea is to conceive cooperative schemes that reward more pertinently the transit contribution each provider grants to the other providers.

As argued in [86] and [87], the Shapley value concept from *cooperative games* can be used to impute profits and costs considering the importance of each AS in the interconnected 'coalition' composed of ASs routing 'common' inter-AS flows. In this way, it is proven that ASs have the incentive to better route yielding to a common inter-domain

routing cost lower than with the current practice, besides than interconnection cost savings. We show that we can use the Shapley value within the framework of Tomaszewski et al., so as to take into account the strategic interaction among provider, by distributing accordingly the related income.

6.1.2 Adopted mathematical notations

The considered network model consists of a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the set of nodes $\mathcal{V}$ and the set of directed links $\mathcal{E}$ ($\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$). For a set of nodes $\mathcal{U} \subseteq \mathcal{V}$ we define the set $\delta^+(\mathcal{U})$ of links outgoing from set $\mathcal{U}$, and the set $\delta^-(\mathcal{U})$ of links incoming to set $\mathcal{U}$. More precisely, $\delta^+(\mathcal{U}) = \{e \in \mathcal{E} : a(e) \in \mathcal{U} \wedge b(e) \notin \mathcal{U}\}$ and $\delta^-(\mathcal{U}) = \{e \in \mathcal{E} : b(e) \in \mathcal{U} \wedge a(e) \notin \mathcal{U}\}$, where $a(e)$ and $b(e)$ denote the originating and terminating node, respectively, of link $e \in \mathcal{E}$. Besides, we shall write $\delta^{\pm}(v)$ instead of $\delta^{\pm}(\{v\})$, i.e., when $\mathcal{U} = \{v\}$ is a singleton.

$\mathcal{M}$ is the set of domains. Each node $v \in \mathcal{V}$ belongs to exactly one domain denoted by $\mathcal{A}(v)$. Hence, set $\mathcal{V}$ is partitioned into subsets $\mathcal{V}^m = \{v \in \mathcal{V} : \mathcal{A}(v) = m\}, m \in \mathcal{M}$. For each $m \in \mathcal{M}$, $\mathcal{E}^m = \{e \in \mathcal{E} : a(e), b(e) \in \mathcal{V}^m\}$ is the set of *intra-domain* links between the nodes in the same domain m. The set of all intra-domain links is denoted by $\mathcal{E}_{\mathcal{I}} = \bigcup_{m \in \mathcal{M}} \mathcal{E}^m$. Further, the set of all *inter-domain* links is denoted by $\mathcal{E}_{\mathcal{O}}$, where $\mathcal{E}_{\mathcal{O}} = \{e \in \mathcal{E} : \mathcal{A}(a(e)) \neq \mathcal{A}(b(e))\} = \bigcup_{m \in \mathcal{M}} \delta^+(\mathcal{V}^m) = \bigcup_{m \in \mathcal{M}} \delta^-(\mathcal{V}^m)$. Clearly, the set of intra-domain links is disjoint with the set of inter-domain links. Finally, the capacity of link $e \in \mathcal{E}$ is denoted by c_e.

Set $\mathcal{D}$ represents traffic demands between pairs of nodes. The originating and terminating nodes of $d \in \mathcal{D}$ are denoted by $s(d)$ and $t(d)$, respectively, and h_d is the traffic volume of d (in the same bandwidth unit as the link capacity). $\mathcal{D}(s,t) = \{d \in \mathcal{D} : s(d) = s \wedge t(d) = t\}$ denotes the set of all demands from node $s \in \mathcal{V}$ to $t \in \mathcal{V}$. In the sequel, z_d will denote the variable specifying the percentage of volume h_d actually handled in the network, i.e., $z_d h_d$ is the carried traffic of demand d. The set of all demands originating in domain m is denoted as $\mathcal{D}^m = \{d \in \mathcal{D} : s(d) \in \mathcal{V}^m\}$. The sets $\mathcal{D}^m = \{d \in \mathcal{D} : s(d) \in \mathcal{V}^m\}$, $m \in \mathcal{M}$, define a partition of $\mathcal{D}$.

Let x_{et} denote a variable specifying the amount of aggregated bandwidth reserved on intra-domain link $e \in \mathcal{E}_{\mathcal{I}}$ for the traffic destined for (a remote) node $t \in \mathcal{V}$. Then, for each inter-domain link $e \in \mathcal{E}_{\mathcal{O}}$ we introduce two flow variables: x^+_{et} and x^-_{et} denote (respectively) the amount of bandwidth reserved for traffic carried on e and destined for t that is reserved by domain $\mathcal{A}(a(e))$ (respectively, $\mathcal{A}(b(e))$) at which link e originates (respectively, terminates). Then for each domain $m \in \mathcal{M}$ we introduce the following flow vectors:

- $\boldsymbol{z}^m = (z_d : d \in \mathcal{D}^m)$, $\boldsymbol{x}^m = (x_{et} : e \in \mathcal{E}^m, t \in \mathcal{V})$
- $\boldsymbol{x}^{m+} = (x^+_{et} : e \in \delta^+(\mathcal{V}^m), t \in \mathcal{V})$, $\boldsymbol{x}^{m-} = (x^-_{et} : e \in \delta^-(\mathcal{V}^m), t \in \mathcal{V})$
- $\boldsymbol{X}^m = (\boldsymbol{z}^m, \boldsymbol{x}^m, \boldsymbol{x}^{m+}, \boldsymbol{x}^{m-})$.

The basic conditions that have to be fulfilled in each domain $m \in \mathcal{M}$ are flow conservation constraints and capacity constraints:

$$\sum_{e \in \delta^+(v) \cap \mathcal{E}^m} x_{et} + \sum_{e \in \delta^+(v) \setminus \mathcal{E}^m} x^+_{et} - \sum_{e \in \delta^-(v) \cap \mathcal{E}^m} x_{et} - \sum_{e \in \delta^-(v) \setminus \mathcal{E}^m} x^-_{et} = \sum_{d \in \mathcal{D}(v,t)} z_d h_d, \quad t \in \mathcal{V}, v \in \mathcal{V}^m \setminus \{t\} \tag{6.1a}$$

$$\sum_{t\in\mathcal{V}} x_{et} \leq c_e, \qquad e \in \mathcal{E}^m \tag{6.1b}$$

$$\sum_{t\in\mathcal{V}} x^+_{et} \leq c_e, \qquad e \in \delta^+(\mathcal{V}^m) \tag{6.1c}$$

$$\sum_{t\in\mathcal{V}} x^-_{et} \leq c_e, \qquad e \in \delta^-(\mathcal{V}^m). \tag{6.1d}$$

Let $\boldsymbol{\mathcal{X}}^m$ ($m \in \mathcal{M}$) denote the set of all vectors $\boldsymbol{X}^m$ satisfying constraints (6.1) and, possibly, certain extra domain-specific conditions. Such extra constraints can for example be implied by requirements for the weight-based shortest-path intra-domain routing (see chapter 7 in [131]) or by QoS-type conditions such as $z_d \geq 1, d \in \mathcal{D}^m$. The routing optimization problem can now be stated as follows:

$$\max \; F(\boldsymbol{z}) = \sum_{m\in\mathcal{M}} \sum_{d\in\mathcal{D}^m} z_d h_d \tag{6.2a}$$

$$\text{s.t.} \quad \boldsymbol{X}^m \in \boldsymbol{\mathcal{X}}^m, \qquad m \in \mathcal{M} \tag{6.2b}$$

$$x^+_{et} \leq x^-_{et}, \qquad e \in \mathcal{E}_{\mathcal{O}},\ t \in \mathcal{V}. \tag{6.2c}$$

The utility of such a formulation is the possibility to use Lagrangean decomposition via the dual problem, which can allow for distributed optimization of the master problem with a form of coordination among the local problems at each domain via successive (sub-)iterations.

6.2 A Shapley value perspective

The Shapley value concept is a game-theoretic solution for value imputation problems that offers interesting properties recalled below [39]. For this reason, it has been applied to very diverse fields [46]. In game theory, interacting agents are modeled as players that take decisions rationally considering the utility functions of all the players. In cooperative games, since some players may contribute more than others for the collaboration, the value imputation problem consists in how to distribute a global value (or revenue) among the players. How important is each player to the coalition, and what payoff can be reasonably expected, are questions to which cooperative coalitional game theory answers with many theoretical concepts. Among these concepts, the Shapley value considers the strategic weight (importance) of each player in the alliance to share the alliance value.

The Shapley value is calculated as follows:

- consider all the possible permutations of the players (e.g., if we have three players 1, 2, 3 the permutations are 123, 132, 213, 231, 312, 321);
- for each permutation and each player, calculate the marginal contribution that the player grants if he joins the coalition formed by the predecessor players (e.g., for the permutation 312, the contribution of 2 is $\mu(123) - \mu(13)$, that of 1 is $\mu(13) - \mu(3)$, that of 3 is $\mu(3) - \mu(\emptyset) = \mu(3)$);
- for each player, calculate its average marginal contribution.

The Shapley value is thus equal to zero for null players, which do not offer any marginal contribution to a coalition in any case, and equal to the single-player payoff for dummy players, which are indifferent in staying in the coalition or not. In our multi-provider framework, dummy players are those that reserve resources for external inter-provider connections but do not obtain the same from the other providers, while null player are those that not even reserve resources.

Mathematical formulation

The Shapley value can be used to assign the payoff of a player as function of his marginal contribution to the coalition. Given that the marginal contribution that a player brings to a coalition (i.e., the alliance income related to its connection services), varies as function of the players that already form the coalition, it is essential considering the order in which the player enters the coalition (or would enter if a coalition evaluates the opportunity of joining the new player).

Mathematically, we use the formulation of a coalitional game. We start with a function $\mu : \mathcal{P}(\mathcal{M}) \to \mathbb{R}$, that goes from subsets of players (partition set of $\mathcal{M}$) to reals, called the "worth function", with the properties:
(i) $\mu(\emptyset) = 0$;
(ii) $\mu(S \cup T) \geq \mu(S) + \mu(T)$, $\forall S, T \subseteq \mathcal{M} \mid S \cap T = \emptyset$.
The computation of μ will be explicated in the next section. The interpretation of the function μ is as follows: if S is a coalition of players which agree to cooperate, then $\mu(S)$ describes the total expected gain from this cooperation, independent of what the actors outside of S do. ii) is the "super-additivity"condition, hypothesis of classical cooperative game theory, which expresses the fact that collaboration can only help, and never hurts. A Shapley value imputation ω_i can thus be calculated for each player $i \in \mathcal{M}$ as function of μ:

$$\omega_i(\mu) = \sum_{S \subseteq \mathcal{M} \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} (\mu(S \cup \{i\}) - \mu(S)) \tag{6.3}$$

where the sum extends over all subsets S of $\mathcal{M}$ not containing player i. The formula can be justified if one imagines the coalition being formed one player at a time, with each player demanding its contribution $\mu(S \setminus \{i\}) - \mu(S)$ as a fair compensation, and then averaging over the possible permutations in which the coalition can form.

Properties

The Shapley value satisfies desirable properties of individual fairness, efficiency, symmetry, additivity and null player modelling (for a detailed characterization see [46]). In fact, it is the only payoff vector - defined on the class of all superadditive games - that satisfies these five properties. Namely, under a Shapley value distribution, in our framework every provider gets at least as much as it would have got without any collaboration, and two strategically equivalent providers obtain the same value. Moreover, the Shapley value distribution supports anonymity. That is, the labelling of the players doesn't play a role in the assignment of their payoffs, i.e., if i and j are two players, and μ^1 is the worth function that acts just like μ^2 except that the roles of i and j have been exchanged, then $\omega_i(\mu^1) = \omega_j(\mu^2)$. Finally, the Shapley value is the single imputation rule that supports marginality, i.e., which uses only the marginal contributions of a player as argument [46].

6.2.1 Coalitional game characterization

The computation of Shapley values for every domain $m \in \mathcal{M}$ of the coalition requires a procedure for evaluation of worth function $\mu(\mathcal{S})$ of arbitrary sub-coalition $\mathcal{S} \subseteq \mathcal{M}$. Let us consider a simple example of a multi-provider network (connectivity graph presented in Fig. 6.1) where node $m \in \mathcal{V} = \{m_1, m_2, m_3, m_4\}$ represents a single domain of the original network and, edge $e \in \mathcal{E} = \{e_1, e_2, e_3, e_4, e_5\}$ represents an aggregate of all the directed links between the pair of domains. Assume that there is a single demand d such that $s(d) = m_1$ and $t(d) = m_3$ of nominal volume $h_d = 1$. Let the optimal routing solution,

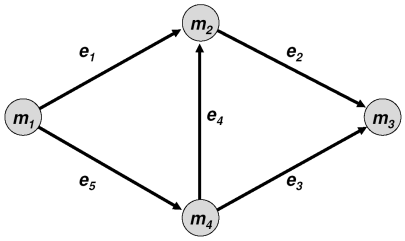


Figure 6.1: Connectivity graph of an exemplary multi-domain network

solution of (6.2), define that $z_d = 1$ and particular reservation levels are: $x_{e_1 m_3} = 0.5, x_{e_2 m_3} = 0.75, x_{e_3 m_3} = 0.25, x_{e_4 m_3} = 0.25, x_{e_5 m_3} = 0.5$. A worth function $\mu(\mathcal{S})$ for the sub-coalition $\mathcal{S} \subseteq \mathcal{M}$ is defined as the value of the objective function (6.2a) that could be achieved when only nodes $m \in \mathcal{S}$ and links $e \in \mathcal{E}_\mathcal{S}$ ($\mathcal{E}_\mathcal{S} = \{e \in \mathcal{E} : a(e) \in \mathcal{S}, b(e) \in \mathcal{S}\}$) would be active.

Therefore, $\mu(\mathcal{S})$ is computed upon the optimal routing solution for the grand coalition, so independently of which sub-coalition is active, the corresponding reservation levels are fixed. This can be considered as heuristic, since the worth value of a sub-coalition is not computed with respect to the optimal reservation levels that would be obtained from a restriction of (6.2) to $S \subseteq \mathcal{M}$. Nevertheless, it is more pragmatic since a provider has no final choice to enter or to leave the grand coalition, which is already imposed by business agreements that take into account not only direct profits but also other issues – e.g., the possibility for extending its customer's base. One can observe that the value of the worth function $\mu(\mathcal{S})$ is equal to zero for every sub-coalition $\mathcal{S} \subseteq \mathcal{M}$ such that does not contain both the source and the destination of demand d (i.e., domains m_1 and m_3) together with at least one from two transit domains (either m_2 or m_4). Hence, the only profitable sub-coalitions are $\mathcal{S}_1 = \{m_1, m_2, m_3, m_4\}, \mathcal{S}_2 = \{m_1, m_2, m_3\}$ and $\mathcal{S}_3 = \{m_1, m_4, m_3\}$ with respective worth functions $\mu(\mathcal{S}_1) = 1$, $\mu(\mathcal{S}_2) = .5$ and $\mu(\mathcal{S}_3) = .25$. Please note, that for coalition $\mathcal{S}_2$, as transit domain m_2 does not receive any flow on its incoming link e_4, it can not send to destination domain m_3 any more than .5 units of traffic that it receives directly from domain m_1.

The Shapley values are then computed using (6.3). The intermediate and the final results of this process are presented in Table 6.1. The first column of the table contains all the possible permutations of domains of the coalition $\mathcal{M}$, the next four columns contain marginal contributions of domains m_1, m_2, m_3 and m_4 respectively. The last row of the table contains the final Shapley values for every domain $m \in \mathcal{M}$ of the coalition.

6.2.2 Worth function computation

The Shapley value computation is complex. This is due to additional intrinsic complexity related to the structure of the optimal routing solution: a flow directed to a particular destination domain t is usually aggregated and has many source domains s, and its sub-flow paths are a-priori unknown.

Let $d_t(v), v \in \mathcal{M}, t \in \mathcal{M}$ denote the total traffic volume generated within v and directed to t. Let $\phi_t(\mathcal{S}, v), \mathcal{S} \subseteq \mathcal{M}, v \in \mathcal{S}, t \in \mathcal{S}$ denote the traffic volume that domain v has to direct to domain t when sub-coalition $\mathcal{S}$ is active. Let $\mathcal{E}_\mathcal{S}$ denote set of links active for sub-coalition $\mathcal{S}$ – i.e., $e \in \mathcal{E}$ such that $a(e) \in \mathcal{S}$ and $b(e) \in \mathcal{S}$. At last, let $\varphi_t(\mathcal{S}, e), \mathcal{S} \subseteq \mathcal{M}, e \in \mathcal{E}_\mathcal{S}, t \in \mathcal{S}$ (we refer to it as volume of link e) denote the volume of traffic to domain t carried on link e when sub-coalition $\mathcal{S}$ is active. The following

permutations	m_1	m_2	m_3	m_4	total
$m_1m_2m_3m_4$	0	0	.5	.5	
$m_1m_2m_4m_3$	0	0	1	0	
$m_1m_4m_2m_3$	0	0	1	0	
$m_1m_4m_3m_2$	0	.75	.25	0	
$m_1m_3m_4m_2$	0	.75	0	.25	
$m_1m_3m_2m_4$	0	.5	0	.5	
$m_2m_3m_4m_1$	1	0	0	0	
$m_2m_3m_1m_4$	.5	0	0	.5	
$m_2m_1m_3m_4$	0	0	.5	.5	
$m_2m_1m_4m_3$	0	0	1	0	
$m_2m_4m_1m_3$	0	0	1	0	
$m_2m_4m_3m_1$	1	0	0	0	
$m_3m_4m_1m_2$	.25	.75	0	0	
$m_3m_4m_2m_1$	1	0	0	0	
$m_3m_2m_4m_1$	1	0	0	0	
$m_3m_2m_1m_4$	.5	0	0	.5	
$m_3m_1m_2m_4$	0	.5	0	.5	
$m_3m_1m_4m_2$	0	.75	0	.25	
$m_4m_1m_2m_3$	0	0	1	0	
$m_4m_1m_3m_2$	0	.75	.25	0	
$m_4m_3m_1m_2$	.25	.75	0	0	
$m_4m_3m_2m_1$	1	0	0	0	
$m_4m_2m_3m_1$	1	0	0	0	
$m_4m_2m_1m_3$	0	0	1	0	
Shapley values	.3125	.229167	.3125	.14833	1

Table 6.1: Intermediate and the final Shapley values

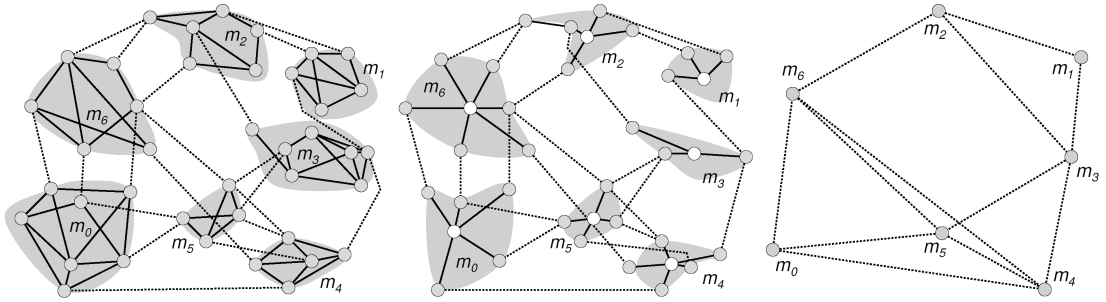
properties hold:

$$x_{et} \geq \varphi_t(\mathcal{S}, e) \qquad \mathcal{S} \subseteq \mathcal{M}, t \in \mathcal{S}, e \in \mathcal{E}_{\mathcal{S}}, \tag{6.4a}$$

$$\phi_t(\mathcal{S}, v) = d_t(v) + \sum_{e \in \delta^-(v) \cap \mathcal{E}_{\mathcal{S}}} \varphi_t(\mathcal{S}, e) \qquad \mathcal{S} \subseteq \mathcal{M}, v \in \mathcal{S}, t \in \mathcal{S}, \tag{6.4b}$$

$$\phi_t(\mathcal{S}, v) \geq \sum_{e \in \delta^+(v) \cap \mathcal{E}_{\mathcal{S}}} \varphi_t(\mathcal{S}, e) \qquad \mathcal{S} \subseteq \mathcal{M}, v \in \mathcal{S}, t \in \mathcal{S}, \tag{6.4c}$$

Property (6.4a) states that reservation levels get from the optimal routing solution define the upper bounds for link volumes. Property (6.4b) simply denotes that volume of domain v is equal to the volume of traffic that it generates together with the sum of volumes of its active incoming links. Finally, property (6.4c) denotes that sum of volumes of active links outgoing from domain v cannot exceed volume of that domain. Distribution of volume of domain v to its outgoing links is trivial if this domain volume exceeds the sum of reservation levels over active outgoing links – i.e., $\phi_t(\mathcal{S}, v) \geq \sum_{e \in \delta^+(v) \cap \mathcal{E}_{\mathcal{S}}} x_{et}, \mathcal{S} \subseteq \mathcal{M}, v \in \mathcal{S}, t \in \mathcal{S}$ – as in such a case, every active outgoing links gets volume equal to its reservation ($\varphi_t(\mathcal{S}, e) = x_{et}$). In the opposite case, where there is a surplus of reservation to use and due to the optimal routing solution got from the distributed optimization does not specify paths for particular demands (or sub-flows), it seems reasonable and simple to assume a fair weighted distribution of volume of domain v to its outgoing links – i.e., $\varphi_t(\mathcal{S}, e) = x_{et}/(\sum_{f \in \delta^+(v) \cap \mathcal{E}_{\mathcal{S}}} x_{ft})\phi_t(\mathcal{S}, v) \quad \mathcal{S} \subseteq \mathcal{M}, t \in \mathcal{S}, e \in \mathcal{E}_{\mathcal{S}}, v \in \mathcal{S}$.



(a) The original network (b) After *star aggregation* (c) Domain connectivity

Figure 6.2: Network topology abstraction schemes

Algorithm 6.2.1: WORTHFUNCTIONCOMPONENT$(\mathcal{S}, t)$

procedure DOMAINVOLUME$(\mathcal{S}, m)$
if not $domainReady(m)$

$$\textbf{then} \begin{cases} \phi_t(\mathcal{S}, m) \leftarrow 0 \\ \textbf{for each } e \in \delta^-(m) \cap \mathcal{E}_{\mathcal{S}} \\ \quad \textbf{do } \phi_t(\mathcal{S}, m) \leftarrow \phi_t(\mathcal{S}, m) + \text{LINKVOLUME}(\mathcal{S}, e, t) \\ domainReady(m) \leftarrow true \end{cases}$$

return $(\phi_t(\mathcal{S}, m))$

procedure LINKVOLUME$(\mathcal{S}, e, t)$
if $a(e) \in \mathcal{S}$
then return $(\phi_t(\mathcal{S}, v) x_{et} / (\sum_{f \in \delta^+(a(e)) \cap \mathcal{E}_{\mathcal{S}}} x_{ft}))$
else return (0)

main
for each $m \in \mathcal{M}$
do $domainReady(m) \leftarrow$ **false**
return (DOMAINVOLUME$(\mathcal{S}, t)$)

Let $\mu(S, t)$ denote a component of worth function of sub-coalition $\mathcal{S} \subseteq \mathcal{M}$ for traffic to destination domain t. To compute value of that component one may use Alg. 6.2.1. The algorithm assumes that a flow of traffic to domain t, which is induced by values of reservations taken form the optimal routing solution, forms an acyclic graph (in fact, the optimal routing solution does not always induce acyclic flows still, they could be easily made acyclic by simple preprocessing). Alg. 6.2.1 takes advantage of the observation that $\mu(S, t) = \phi_t(\mathcal{S}, t)$ – i.e., that value of worth function component, for particular sub-coalition $\mathcal{S}$ and destination domain t, is equal to the volume of domain t.

Finally, the worth function of the sub-coalition $\mathcal{S} \subseteq \mathcal{M}$ can be computed as

$$\mu(S) = \sum_{t \in S} \mu(S, t) \tag{6.5}$$

Then, (6.3) computes the Shapley value imputation for each provider.

$a(e)$	$b(e)$	m_1	m_6	m_0	m_5	m_4	m_3	m_2
m_0	m_4	155	0	0	0	3987	174	0
m_0	m_5	132	0	0	2694	0	1052	0
m_0	m_6	563	11174	0	0	0	0	1031
m_1	m_2	0	724	809	0	0	0	6384
m_1	m_3	0	0	0	495	625	1627	0
m_2	m_1	7669	0	0	0	0	0	0
m_2	m_3	0	0	0	0	0	2616	0
m_2	m_6	0	7092	1822	542	651	0	0
m_3	m_1	5596	0	0	0	0	0	0
m_3	m_2	0	246	0	0	0	0	5168
m_3	m_4	0	0	202	0	4992	0	0
m_3	m_5	0	794	1070	2274	0	0	0
m_4	m_0	0	0	2158	0	0	0	0
m_4	m_3	772	0	0	0	0	1333	495
m_4	m_5	0	0	0	9418	0	0	0
m_4	m_6	0	4862	0	0	0	0	214
m_5	m_0	0	0	7756	0	0	0	0
m_5	m_3	624	0	0	0	0	7257	393
m_5	m_4	0	0	0	0	7870	0	0
m_5	m_6	0	3413	0	0	0	0	212
m_6	m_0	0	0	8903	0	0	0	0
m_6	m_2	1280	0	0	0	0	724	9921
m_6	m_4	0	0	0	0	2467	0	0
m_6	m_5	0	0	0	1550	0	254	0

Table 6.2: Reservation levels in the domain connectivity graph

6.3 Numerical results

We tested the Shapley value based distribution algorithm for a multi-provider network consisting of seven domains. The original network topology of the network is presented in Fig. 6.2a, where thick lines represent intra-domain links and thin lines represent inter-domain links; independently of the type of the link, a single line represents a pair of oppositely directed unidirectional links of equal capacity). The considered traffic matrix is random.

To reduce the complexity of the original network, a *star-aggregation* of intra-domain networks was applied (according to the methodology described in [129]). For the resulting aggregated network (Fig. 6.2b) the distributed optimization process was run (it terminated in about one hundred of iterations). Then, we did the second-stage aggregation to receive a domain connectivity graph of the original network (Fig. 6.2c). Table 6.2 shows reservation levels (from the optimal routing solution) projected onto links of the domain connectivity graph. First two columns of the table identify the starting and the terminating node of a link, and the remaining columns depict the amount of bandwidth that is reserved on that link for traffic directed to particular destination.

Let us consider flow to the arbitrarily chosen destination domain – e.g., flow to domain m_3. That flow is depicted in Fig. 6.3, where number beside a link denote the amount of bandwidth reserved on that link for traffic to domain m_3. Considering these reservation levels, one can easily compute the amount of traffic to domain m_3 that each domain injects into the network, terminates or transits (results for such computations are presented in Fig. 6.3). The Table 6.3 shows how the two considered distributions divide the income related to flow to domain m_3 between particular domains (the original distribution refers to the implicit distribution rule assumed in [127] and [128], where

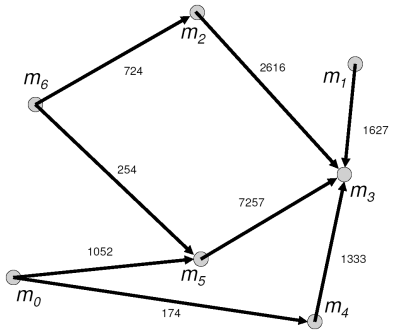


Figure 6.3: Reservations toward m_3

provider	injects	ends	transits
m_0	1225	0	0
m_1	1619	0	0
m_2	1902	0	773
m_3	0	12833	0
m_4	1091	0	174
m_5	5948	0	1325
m_6	1045	0	0

Table 6.3: Flow m_3 components

	Original distribution				Shapley distribution			
	i	t	tr	$\sum$	i	t	tr	$\sum$
m_0	1225	0	0	1225	408	0	0	408
m_1	1619	0	0	1619	809	0	0	809
m_2	1902	0	0	1902	x	0	x	1188
m_3	0	0	0	0	0	6010	0	6010
m_4	1091	0	0	1091	x	0	x	602
m_5	5948	0	0	5948	x	0	x	3408
m_6	1045	0	0	1045	323	0	0	323

Table 6.4: Flow m_3 related income distribution

a domain was awarded only for traffic that it injects into the network – there were no income components related to transiting nor terminating traffic). There are four columns for each distribution – i denotes income component related to traffic a domain injects into the network, t income component related to traffic terminated within a domain and tr income component related to traffic transited by a domain. Finally, column $\sum$ denotes the total income that is attributed to a domain by particular distribution.

Observing the Tables 6.3 and 6.4 one can easily conclude that the original distribution is unfair – as there are significant unpaid volumes of traffic terminated by domain m_3 and transited through domains m_2, m_4 and m_5. The second part of Table 6.4 shows that the proposed Shapley value distribution schemes offers significantly fairer results, as domains are awarded for every type of their contribution in the total income (the 'x's mean that the Shapley value attributed to transit domain cannot be easily divided into components related to injecting and transiting of traffic). Namely, the Shapley scheme assigns to m_3 the biggest share while the original scheme would assign a null income: without m_3 12833 units of traffic (c.f. the total ingress traffic at m_3 in Fig. 6.3) could not be provided, so the corresponding revenue is distributed fairly also to m_3 recognizing to it an income share of 6010. Or, m_0 and m_6 not reserving bandwidth for any external connection, receive roughly one third of the original share.

In this study case (restricted to a single destination, m_3), we can appreciate the application of the concept initially proposed by the authors of [87]. They claim that nowadays the Internet is characterized by "Content providers"(e.g., Youtube) that delivers traffic to "Eyeball providers"(e.g., Polish Telecom) that connect large communities of customers. Since the Content providers get revenue by selling services to customers of Eyeball providers using the network of "Carrier providers"(e.g., Opentransit), they propose to share the corresponding income among all the providers in the delivery chain. In our restricted study case, m_0, m_1 and m_6 can be seen as Content providers simply injecting traffic toward m_3, the Eyeball provider, crossing m_2, m_4 and m_5 that act as

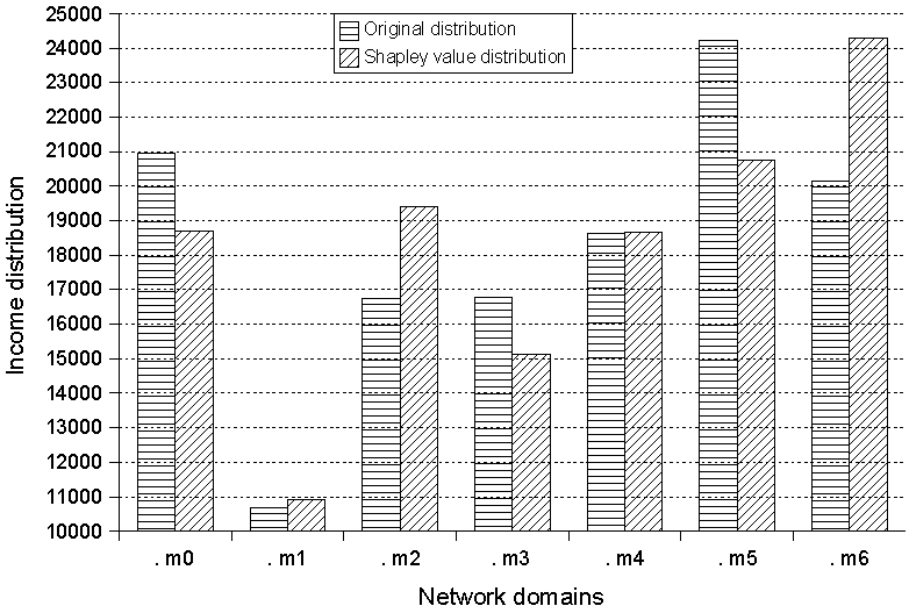


Figure 6.4: Income distribution schemes comparison

Carrier providers as well as Content providers.

In the general case, by contemplating a mixed Content/Eyeball/Carrier behavior for each domain, our framework somehow adopts and extends the concept proposed in [87], by coupling a routing decomposition optimization framework that deals with multiple connections, with a fair income distribution policy. For the general case, in Fig. 6.4 we compare the original distribution to the final Shapley values, which are computed summing all the contributions due to *all the flows* and *all the destinations*. We can better appreciate the global effect of the proposed distribution scheme. For each domain, the result is a fair weighting of the traffic injected, terminated and transited, following the Shapley imputation rule (6.3).

Those domains better interconnected, i.e., with more neighbors and more intra-domain availability (c.f. Fig. 6.2), are able to transit traffic (c.f. the reservation levels in Table 6.2) and get a higher share. This is the case for example of m_6 and m_2 that increase their share of 29% and 16%, respectively. Those domains that inject a lot of traffic but still offer an adequate transit for alliance connections, e.g., m_4, maintain a similar share. Instead, for those whose injected traffic volume is not sufficiently compensated with transit contribution, e.g., m_5, m_3 and m_0, the share decreases (of roughly 10%).

6.4 Summary

In this chapter we have presented a cooperative cross-provider resource reservation framework. We discussed under which circumstances this might result economically feasible. In order to support the adoption of the framework, we proposed a fair income distribution scheme relying on the Shapley value concept from cooperative game theory, showing how the complex issue of computing the Shapley values using decomposition result parameters can be solved heuristically.

By comparison with the original implicit income distribution policy, we show the benefits of the adoption of the Shapley value distribution scheme. Those providers that attract large volumes of traffic can receive an income for such a contribution. Those that do not balance their injected traffic volume with bandwidth reserved for external connection transit, see their income share decreased. Those that do not offer transit at all are fairly penalized. Our approach is a further step (after a few others such as [87]) toward the definition of feasible cooperative routing frameworks and acceptable business models for the future Internet.

As a further work we aim to refine the optimization decomposition method so as to allow a pro-active integration of the Shapley values. The idea is to control the amount of traffic volume a provider is allowed to inject within the alliance. It might be desirable to allow rewarding a provider's transit contribution directly with intra-alliance traffic injection ability by bounding the inter-provider throughput. A direct integration of the Shapley value in a cross-provider reservation level computation procedure may reveal to be too combinatorial. In this sense, a further work should address the challenge of finding acceptable approximation methods for the Shapley value computation.

Part II

Transport Architectures

Chapter 7

Physical Interconnection Issues in Provider Networks

In this chapter, we study provider interconnection issues with a physical transport standpoint, focusing on the physical issue of signaling circuits (namely, G-MPLS LSP) across provider network borders. In particular, being the Internet carriers quite often interconnected using Internet eXchange Points (IXP) infrastructures instead of direct bilateral interconnections, we are interested in novel lightpath-friendly IXP transport architectures. We review current solutions and indicate the requirements for a more effective IXP transport architecture, presenting a promising solution object of the last chapters of the dissertation.

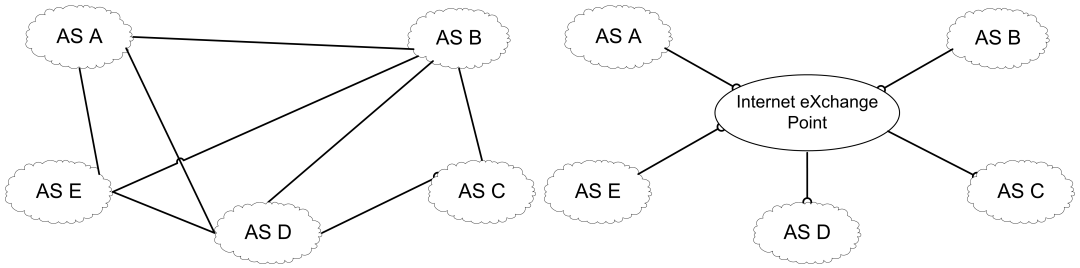
7.1 Internet infrastructure and inter-AS G-MPLS

In the dissertation so far, we have explored novel routing frameworks and architectures to support inter-provider services. Starting from Chapter 4, we have dealt with multiple facets related to the provisioning of inter-provider MPLS and G-MPLS LSP tunnels and circuits. In particular, in Sect. 4.2 we presented how an LSP can be functionally signaled across several provider borders. In order to practically deploy the multi-provider architecture, an implicit assumption is thus that the providers in a carrier alliance are interconnected either directly with one or several point-to-point links, either through a Internet eXchange Point (IXP) infrastructure that supports G-MPLS signaling.

Nevertheless, a direct interconnection between different ASs may be infeasible for a number of practical reason (mainly, economical and geographical). On the other hand, nowadays most of IXP infrastructures are not able to strictly reserve physical transport resources between two providers, whether they consists in simple MPLS LSP tunnels or in more critical wavelength-based ASON/G-MPLS LSP circuits. In the following, we review common IXP infrastructures, highlighting related issues for inter-AS G-MPLS and investigating a novel transport architecture as a candidate next generation IXP infrastructure.

7.1.1 Internet eXchange Points

An Internet eXchange Point (IXP) is a peering interconnection infrastructure among several providers that represents a practical alternative solution (for providers) to many bilateral point-to-point links. Physically an IXP is a single facility to which providers



(a) Mesh of bilateral interconnections (b) Internet eXchange Point

Figure 7.1: Peering interconnection types.

trench high-capacity optical fiber links. Nowadays, there are roughly 500 IXPs all around the world, each IXP processing on average 40 Gb/s of traffic (see [195] for an updated status). Providers connect to an IXP in order to dispose of a physical interconnection infrastructure with one or more 'peers' with which they have peering agreements (the traffic passing through an IXP is typically not billed), the traffic routing still being ruled by BGP routing policies .

The key advantage in interconnecting to an IXP is the possibility of peering with a potentially very large number of ASs using a unique physical connection. As already discussed in previous chapters, peering agreements can offer many advantages in terms of transmission speed, Internet connection reliability and cost (avoiding to settle new transit agreements or upgrading existing ones).

Strategically, one can distinguish between "public IXP"and "private IXP"types. The first is purely commercial: provider subscribers normally pay a flat rate function of the consumed bandwidth. Public IXPs are normally regional facilities that interconnect regional or national providers, which commonly manage their IXP facility, hence called 'private peering'.

Topologically, an IXP is thus a star interconnection between several providers, as depicted in Fig. 7.1b. The traffic to and from all the peers is thus aggregated in one (or a few) interconnection links. This is a scalable alternative to a mesh of private peering interconnection between peering providers, as depicted in Fig. 7.1a.

7.1.2 Physical IXP infrastructure solutions

We synthetically review some relevant propositions for IXP infrastructures.

LAN-based IXP

Physically, a typical IXP consists in a sort of Local Area Network (LAN) (e.g., see Fig. 7.1), with one or more network switches to which each of the participating ISPs interconnect. Some sources state that over 90% of IXPs use nowadays Ethernet (or Giga-Ethernet) switches. For some commercial ISPs, for example, the billing is a function of the used Ethernet ports. In the 90's, prior to the existence of high-capacity switches, IXPs typically utilised FOIRL (Fibre-Optic Inter-Repeater Link, the original standard for Ethernet over fibre) hubs or FDDI (fiber Distributed Data Interface, a token-ring like protocol) rings. As Ethernet became available in mid 90's, the most of IXPs migrated to Ethernet-LAN solutions.

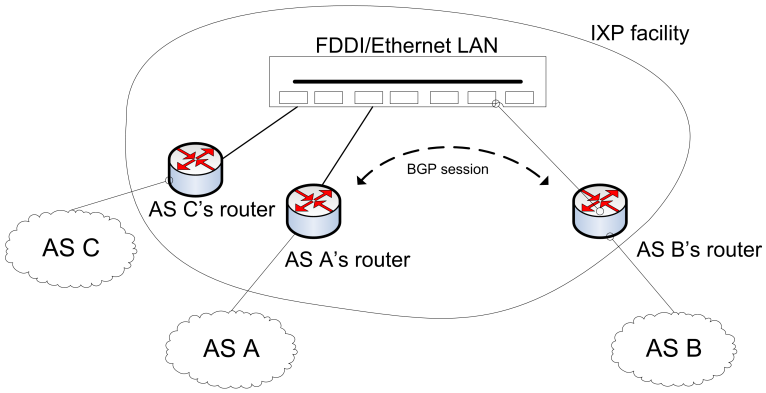


Figure 7.2: LAN-based typical IXP infrastructure

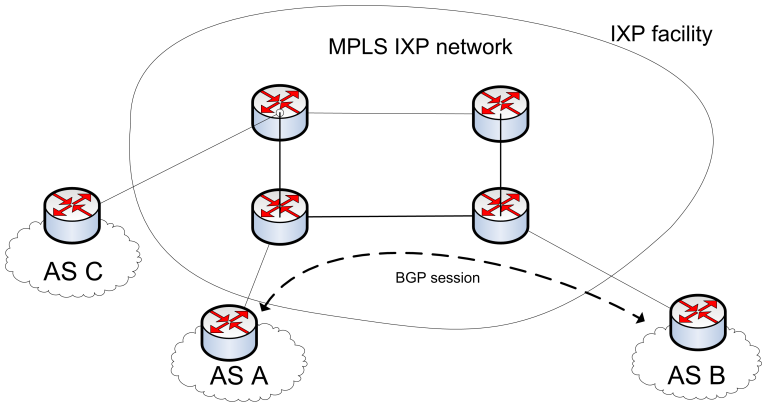


Figure 7.3: MPLS-based IXP infrastructure

Circuit switching IXP

Although its simplicity, a LAN-based solution presents security, management and scalability issues and drawbacks related to the shared bus [133]. Namely, the switching speed can not be guaranteed; each participating provider has to locate a router close to or into the IXP facility because of physical cable restrictions; bandwidth and latency performance – e.g., for inter-AS LSPs – can not be guaranteed. To overcome these issues, ATM switches could be used (and they were indeed briefly used at a few IXPs in the 1990s), but finally Ethernet has prevailed for its *simplicity*, low operational cost, low overhead and high scalability. Moreover, the recent introduction of Virtual LAN (VLAN) mechanisms in Ethernet equipment allowed to partially address the above-mentioned issues.

As a matter of fact, ATM has been pushed out from the core networks, where it has being replaced by (or integrated to) MPLS and ASON/G-MPLS. Eventually, the requirements for optical LSP-switching enabled IXP infrastructure are going to arise in the next future, especially when inter-AS G-MPLS service provisioning will need to be automated.

A step in this direction is the MPLS-based IXP infrastructure proposed by the authors in [133], depicted in Fig. 7.3. They propose to deploy an MPLS core in the IXP's facility so as to allow inter-AS MPLS, and validate the solution in a testbed. Obviously, as the authors point out, an MPLS-IXP network should not hold any routing information exchanged between ASs. One may wonder why, indeed, MPLS capabilities shall

be enabled in the IXP given that no specific routing functionalities is required to an IXP but only an interconnection functionality. Moreover, an intra-IXP network topology may require traffic engineering procedures within the IXP, which would increase the operational management and operational costs (to be shared among providers or to be imputed on them).

7.2 Future-generation IXP requirements

Therefore, there is no real need to have MPLS routers in the IXP facility, while still remains the need to support inter-AS G-MPLS. Moreover, both providers and IXPs would surely prefer not installing new routers in the IXP facility because of cost and space reasons, respectively. The natural solution is thus to leave edge G-MPLS routers outside the IXP facility and let IXP offering a simple and scalable physical (optical) interconnection solution.

We can arise the following requirements for next generation IXP infrastructures:

1. *inter-AS G-MPLS compliancy*: two customer providers shall be able to signal and reserve inter-AS MPLS or G-MPLS LSPs across an IXP;

2. *very high speed interconnectability*

3. *IXP facility space*: providers shall not deploy their routers or switches inside the IXP facility;

4. *IXP scalability and upgradeability*: in order to support a large number of customers and increasing volumes of traffic, the IXP infrastructure shall be modular and easily upgradeable;

5. *low OPEX*: as IXP customer providers either share the IXP costs (private IXPs) or support them with their subscription (public commercial IXPs), the IXP infrastructure shall ease traffic engineering and routing;

6. *IXP reliability and survivability*: the IXP solution shall offer an interconnection that is reliable and survivable against fiber, cable, core node and IXP site failures[1].

7.3 The Petaweb architecture solution

In [97] the authors propose to adapt the "Petaweb architecture"– originally proposed by Nortel Networks [117] – as physical infrastructure beyond a new model of IXP. Their suggestion is interesting in that such a solution could meet the above-mentioned requirements. In the following, we present the Petaweb architecture that, besides its applicability as a future-generation IXP infrastructure, also represents an interesting transport architecture for multi-hub Internet carriers.

[1]The last requirement deserves some attention. As already argued, the Internet reliability depends on the peering settlements, hence physically also on the IXP reliability. It is well know that in 2008, e.g., a failure in the Milan IXP would have isolated the large part of Italian servers. Or, in 2008 the Amsterdam IXP suffered from a power outage in one of its facilities, which lead to a drop of roughly 200 Gb/s of traffic on 500 Gb/s [204]. Therefore, a particular care on the reliability of transport architecture solution shall be taken.

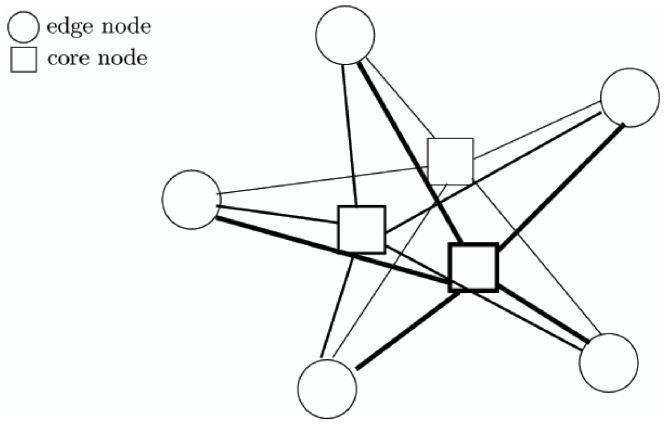


Figure 7.4: The Petaweb architecture: a composite-star structure

7.3.1 The Petaweb architecture

The Petaweb is a next generation transport network structure that offers a total capacity of several petabits per second (10^{15} bit/s) that was proposed in [117] [118]. The term Petaweb was coined because the architecture can deal with thousands of nodes each requesting an external capacity of terabits per seconds (10^{12} bit/s). The structure provides fully meshed connectivity with direct optical paths between some electronic edge nodes. It is composed of several OXCs (Optical Cross-Connectors), also named core nodes, that commute the traffic exchanged by the edge nodes. One particular feature is that each optical core node is connected to all edge nodes. Another peculiar characteristic is that the core nodes are not connected among themselves, making it a complete architectural breakthrough.

The Petaweb can be seen as a superposition of star structures as shown in Fig. 7.4. The great advantage of such a structure is the important simplification of key network functionalities such as routing, addressing and scheduling that is provided by the one-hop connection architecture. The term one-hop refers to having just one intermediate physical node between any pair of edge nodes. Such a simplification leads to greatly increasing network efficiency and communication speed. This comes at the expense of a significant increase in fiber costs as all the edge nodes have to be connected to all the core nodes. Moreover, its upgrade has to be carefully crafted. In [153], the a practical implementation of Petaweb architecture was compared with an optical multi-hop network and it was found that although the Petaweb requires a higher fiber length, it needs much fewer ports and no wavelength conversion thanks to the one-hop connectivity.

In order to study an implementation of the Petaweb infrastructure as a possible IXP solution, the authors in [97] practically propose a single-site restriction of the Petaweb integrated with a specific traffic engineering framework. That is, a Petaweb-like IXP solution in which all the core nodes are installed in the same switching site (the IXP facility), and some intelligence is located in route computation servers within the IXP. Nevertheless, with respect to the latter aspect, given the standardization activities on the distributed PCE architecture conducted in the last years, and provided the integration of the PCE architecture with a multi-provider service plane that support inter-provider routing as we proposed in previous chapters, the need to dispose of such an IXP intelligence does not seem to be practically justified. The idea of defining a Petaweb-like IXP solution – possibly not restricted to a single site – is, instead, a very interesting research challenge in rupture with the current practice and worth being explored.

7.3.2 Design aspects

The architecture includes core nodes of different sizes, and several fibers can connect an edge node to a core node. In order to construct a Petaweb, it is necessary to efficiently tackle the network design problem. That is, to find the location and the type of core nodes that will be placed in the network in order to satisfy the demand between edge nodes, while minimizing costs and respecting the architectural constraints. This is particularly important given that the Petaweb may be one of the largest networks ever designed and has been even proposed as a building block for the YottaWeb, a mega-network with aggregated capacities in the order of yottabits per second (10^{24} bits/s) [139].

From the design standpoint, the Petaweb design problem is unique since telecommunication networks are typically composed of a backbone and an access network and the design consists in how to optimize separately or jointly those two different levels. In [56] a thorough review of all the types of design problems and algorithmic resolutions can be found. The Petaweb, on the other hand, presents a different structure: all the edge nodes are connected through a backbone switch and yet the backbone switches are disconnected among themselves.

7.4 Summary

As an important step that shall be tackled in the definition of a complete multi-provider framework, in this chapter we presented the interconnection issue of providers' network. We introduced some requirements for future Internet eXchange point infrastructures. In particular, a viable solution is represented by the so-called Petaweb architecture.

The Petaweb design problem has been tackled, for the first time, in the master thesis [140] and [29]. Since then, we worked at the definition of optimal and heuristic design dimensioning approaches, considering different switching solutions, proposing a protection strategy, novel Petaweb-like structures, and assessing the reliability and the survivability of the studied solutions. This aspects are discussed in details in the last three chapters of the dissertation.

Chapter 8

Design Optimization of the Petaweb Architecture

In this chapter we define the Petaweb Design Problem and propose planning approaches under different design choices. In particular, we present how TDM/WDM switching and path protection can be implemented, and discuss when a quasi-regular structure shall be preferred. For the sake of seamlessness, the numerical results are presented in parallel with the next chapter and summarized across this chapter[1][2].

8.1 The Petaweb Design Problem

The Petaweb is based on the WDM (Wavelength Division Multiplexing) technology. Each fiber is composed of a fixed number of channels, each channel corresponding to one wavelength. At the ingress of each core node, each fiber is demultiplexed and each channel is connected to its associated switching plane. As depicted in Fig. 8.1, in a switching plane of such a core node there are W space switches each of which commutes channels of the same wavelength. The channels that are sent to the same destination edge node are multiplexed to the same link (to ease the interpretation, only the channels from and to edge node 1 are depicted).

The architecture includes core nodes of different sizes. For bigger core nodes, the number of space switches can be a multiple of the number of wavelengths. For example, with W=16 channels per fiber, a core node can have 16 or 32, or 48, or 64 space switches. Thus, we classify each core node by its type r, which represents the size of the core node. A core node of type r has s_r switching planes, each composed of W space switches. Note that several fibers can connect an edge node to a core node, since there is one connection to each switching plane; from now on, we call 'link' the set of fibers connecting an edge to a core node. Moreover, given the regularity of the core node architecture (same number of wavelengths per fiber, and of fibers per link), no wavelength conversion is required, and no wavelength continuity constraint needs to be applied.

[1] The contents presented in the next two chapters are also presented in [4], [12], [15] and [17].

[2] The work presented in this and the next two chapters has been conducted in the framework of the NSERC strategic grant nb. STPG 246/59, and the Euro-FGI, Euro-NF, e-Photon/One+ and BONE networks of excellence

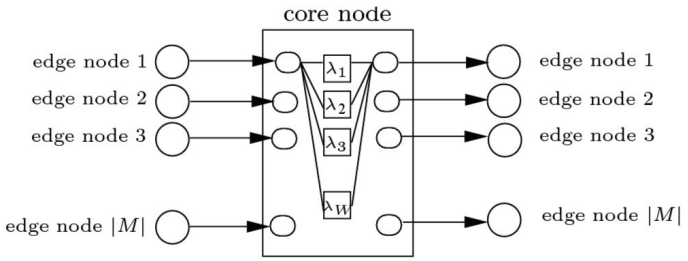


Figure 8.1: Connection between the edge nodes and a core node

8.1.1 General description and notation

The Petaweb Design Problem (PDP) consists in determining both the number and the optimal location of the core nodes given a general traffic matrix, and respecting a series of capacity and physical constraints so that a cost function is minimized. In other words, we want to know which core nodes should be opened, of which type they are and through which core node each traffic connection should be switched. From now on, we say that *a core node is open* at a site if that node specimen has to be installed in the site.

We assume that the location of edge nodes, the matrix of traffic between the edge nodes and the potential locations for the core nodes are given. Moreover, since two edge nodes generate two connection requests, one per direction, we do not assume any type of symmetry in the traffic routing, i.e., the two connection requests can be switched by different core nodes. Furthermore, it is also assumed that the potential locations for the core nodes are the sites of the edge nodes.

Let us introduce some notations.

N: the set of edge node sites, potential core node locations.

T: the set of edge node pairs, with the origins different from the destinations, that is, $T \subset N \times N$; this corresponds to the set of connection requests with one connection request per pair.

$s(p), t(p)$: the source and destination node of p, respectively.

V: the set of core node types.

s_r: the number of switching planes for core node of type r, $r \in V$

E: the set of the core node specimens of the same type that can be opened at one site, $E \subset \mathbf{N}^3$. $e \in E$ identifies an individual core node.

C_{ch}: the wavelength channel capacity (in Gb/s).

W: the number of wavelengths per fiber.

C_j: the capacity of edge node j, $j \in M$, (in Gb/s).

f_r: the cost of one core node of type r, $r \in V$.

P: the cost of one port in a core node.

[3]Note that as the E set is finite, the maximum number of core nodes that can be opened at a site is limited. Moreover, when core and edge nodes are in the same site, the distance between them is negligible (null), and the interconnection costless.

γ: the scale factor for the cost of the ports.

F: the reference fiber cost per length unit.

$\phi(W)$: scaling function of F as a function of the number of wavelengths.

β: the cost representing the propagation delay, per length and traffic unit.

Q_p: the traffic of a origin/destination pair p, $p \in T$, (in Gb/s).

δ_{ij}: the distance between the site i, $i \in N$, and the edge node j, $j \in M$.

d_{ip}: the sum of the distance between the origin edge node of the pair p. and the site i, and of the distance between the site i and the destination edge node of the pair p. If j and k are the origin and the destination on node pair p, then $d_{ip} = \delta_{ij} + \delta_{ik}$.

y_{ire}: binary variable equal to 1 if the e^{th} core node of type r located at i is opened and 0 otherwise.

$x_{ire,p}$: binary variable equal to 1 if traffic Q_p is switched by the e^{th} core node of type r located at i and 0 otherwise.

Cost function

We propose to integrate three different types of cost terms into the cost function: the cost of the core nodes, the cost of the fiber and a propagation delay cost. The last is added to provide flexibility to the network design model by avoiding choosing locations that imply too much propagation delay. The trade-offs between those terms will be part of the study.

Cost of the core node: This term is composed of a fixed cost f_r that depends on the type of node, and of a variable cost that depends on the ports. The cost of the ports in a node of type 1 ($s_r = 1$) is given by P times the number of ports. The number of ports in a core node of type r is then given by $2\,|N|\,W\,s_r$; the factor 2 comes from the fact that there must be entry and exit ports. The cost of the ports in a core node of type r is given by $2\,|N|\,W\,s_r\,\gamma^{(s_r-1)}\,P$. Factor γ is lower than 1 so that the cost per port decreases with the type of core node. If $\gamma = 0.95$, e.g., the cost of the ports of type 1 will be $2\,|N|\,W\,P$. On the other hand, the cost of the ports of type 2 will be $0.95 \times (2\,|N|\,W\,P)$ which implies an economy of 5%.

Cost of the fiber: This term is given by the expression:
$\sum_{i\in N} \sum_{r\in V} \sum_{e\in E} 2\,\phi(W)\,F\,s_r\,(\sum_{j\in N}\,\delta_{ij}) y_{ire}$.
Note that $\phi(W)\,F$ provides us with a unitary cost per length of fiber, $\phi(W)$ being a function that may depend on the manufacturer.

Propagation delay cost: This term aims at choosing the edge core type and location so that the *pondered* propagation delay is minimized. A pondered term allows to penalize long connections between origin destination edge node pairs that share high levels of traffic. The term is given by the product of the total distance travelled by a signal of a particular origin destination p by the total demand Q_p weighted by a factor β that is used to vary the importance of the propagation delay in the objective function, that is: $\sum_{i\in N} \sum_{r\in V} \sum_{e\in E} \sum_{p\in T} \beta d_{ip} Q_p x_{ire,p}$.

It is worth mentioning that this type of cost model is unusual for the literature in the area. More generally, the physical dimensioning[4] of optical networks has not been

[4]With 'dimensioning' we mean the operation consisting in selecting the location of new network equipments and sites

a hot research topic in the last decade. The reason is probably that physical transport networks are considered as built incrementally from existing facilities. Our approach assumes, instead, a from-scratch installation of a novel transport infrastructure, possibly interconnecting existing networks. The Petaweb cost model we propose can be, however, applied also to the design dimensioning of existing optical networks with some physical constraints, such as the node location or the node architecture. With respect to this issue, in Appendix C we show how a close cost model can be applied in the design dimensioning of optical networks with WaveBand switching features.

Another aspect worth discussing is the choice of limiting the design problem to facility location and link installation issues. More physical photonic processing aspects such as 3R regeneration costs are not considered. In fact, one may consider that such costs can be with some approximation spread in per-distance (F) or per-node cost (f_r), especially when the long-haul fiber links are, more or less, at close lengths (say, with a length difference a few dozens of km). When this can not stand, we think the cost model can be easily extended (at the expense, however, of additional complexity).

The final objective cost function of the Petaweb design problem is:

$$\begin{aligned} G(y_{ire}, x_{ire,p}) &= \sum_{i\in N}\sum_{r\in V}\sum_{e\in E}(2\,|N|\,W\,s_r\,\gamma^{(s_r-1)}\,P + f_r)\,y_{ire} \\ &+ \sum_{i\in N}\sum_{r\in V}\sum_{e\in E} 2\,\phi(W)\,F\,s_r\,(\sum_{j\in M}\delta_{ij})y_{ire} \\ &+ \sum_{i\in N}\sum_{r\in V}\sum_{e\in E}\sum_{p\in T}\beta d_{ip}Q_p x_{ire,p} \end{aligned} \tag{8.1}$$

The following constraints should hold.

Unicity of the core node connection

$$\sum_{i\in N}\sum_{r\in V}\sum_{e\in E} x_{ire,p} = 1, \quad \forall p \in T \tag{8.2}$$

It indicates that the total traffic exchanged by a pair of edge nodes must be routed through a single core node.

Linking constraints

$$x_{ire,p} \le y_{ire}, \quad \forall i \in N, \forall r \in V, \forall e \in E, \forall p \in T \tag{8.3}$$

It specifies that the traffic can be routed through the e^{th} core node of type r located at site i only if this core node is active.

Core node capacity constraints

$$\sum_{p\in T} Q_p x_{ire,p} \le s_r W|N|C_{ch}y_{ire}, \quad \forall i \in N, \forall r \in V, \forall e \in E \tag{8.4}$$

It states that the capacity of each core node must be respected.

Edge node capacity constraints

$$C_{ch} \times W \times \sum_{i \in N} \sum_{r \in V} \sum_{e \in E} s_r \, y_{ire} \leq C_j, \quad \forall j \in N \tag{8.5}$$

It guarantees that the capacity of the edge nodes is respected, i.e., it ensures that the transmission capacity of an edge node is equal or bigger than the switching capacity of all the network, which is directly proportional to the number of opened switching planes ($\sum_{ire} s_r y_{ire}$). Practically, it is a bound on the number of fibers through which edge nodes are linked to the network core. This necessarily would restrict in the optimization the choice of core nodes to be connected to. For instance, an edge node with capacity = 1 Tb/s can be at most connected to the network with $\left\lfloor \frac{1Tb/s}{160Gb/s}) \right\rfloor = 5$ fibers (with each fiber having 16 wavelengths of 10Gb/s) per direction. This can correspond, for instance, to 1 core node of type 1 and one of type 3, or 5 of type 1, etc.

Link capacity constraints

$$\sum_{p \in T}^{s(p)=j} Q_p x_{ire,p} \leq C_{ch} \times W \times s_r \, y_{ire}, \quad \forall j, i \in N, \forall r \in V, \forall e \in E \tag{8.6}$$

$$\sum_{p \in T}^{t(p)=k} Q_p x_{ire,p} \leq C_{ch} \times W \times s_r \, y_{ire}, \quad \forall k, i \in N, \forall r \in V, \forall e \in E \tag{8.7}$$

These ensure that the total link capacity is respected for all the links between each origin edge node and core node or each core node and edge node, respectively.

Binary constraints

$$y_{ire}, x_{ire,p} \ \in \{0, 1\} \tag{8.8}$$

8.1.2 The mathematical model

Now that we have defined all the variables, cost functions and constraints of the model we define the PDP as the following problem:

min (8.1) subject to: (8.2), (8.5), (8.6), (8.7), and (8.8).

Note that constraints (8.6) and (8.7) imply (8.4) and (8.3) which, therefore, were omitted from the final formulation.

This problem presents $|N||V||E|$ binary variables ($\sim |N|$) for the location of the core nodes and $|N||V||E||T|$ binary variables ($\sim |N||T|$) for the edge traffic switching through specific core nodes, for the worst case. The number of constraints of the problem is given by $|T| + |N| + 2|N||V||E||N|$ ($\sim |T| + |N||N|$). Being $|T| \approx |N|^2$, the complexity of the PDP depends on a number of variables $\sim |N|^3$ and on a number of contraints $\sim |N|^2$.

In mathematical terms, the PDP reminds a location problem since we must decide where to place the core nodes. It presents similarities with the Capacitated Facility Location Problem [57] and, in particular, with the Single Source Facility Location Problem (SSFLP) [58, 59]. Nevertheless, the capacity and physical constraints that are present in the design make it a problem much more difficult to solve. The SSFLP is known to be NP-hard: a set of customers must be served by a single facility, there is a cost associated to opening a facility in a particular location and a transportation cost from the facility

to the customer; each customer has a particular demand and each facility has a limited capacity. The problem is to find where to locate the facilities to minimize the cost of the network.

Proposition 8.1.1. *The Petaweb Design Problem is NP-hard.*

Proof. The SSFLP reduces to an instance of the PDP. To show the reduction, let us assume that in the PDP we create two edge nodes for each customer of the SSFLP and that both are in the same location. Those pairs of edge nodes that represent a customer will have a demand among themselves equal to the customer demand from a facility, all the demands between other edge nodes will be set to zero. The demand between edge nodes that has to be entered in the PDP is set equal to each customer demand from a facility in the SSFLP. The cost of the link between the potential core node location and each edge node in the PDP is set to half the cost between the potential facility location and the customer of the SSFLP. To account for the single type of facility, only one type of core node will be considered in the PDP. Also, the cost of installing a core node is equal to the cost of opening a facility. The capacity constraint of the core node in the PDP is set to the capacity of the facility in the SSFLP. Thus, the solution of this instance of the PDP will provide us with the solution of the SSFLP and the proof is completed. □

8.2 The resolution approach

We present a design method based on a repeated matching heuristic able to solve large instances. We first reformulate the problem before introducing the heuristic and discussing complexity issues.

8.2.1 Reformulation of the design problem

Let an edge node pair be designated by the letter p, $p \in T$. Let us remember that $p_1 = (i, j)$ is different from $p_2 = (j, i)$, i...e., between two edge nodes we have two edge node pairs, representing two different connection requests. A subset k of edge node pairs is designated by D_k so that $D_k \subset T$. For example, with three edge nodes, we could have : $T = \{(1,2),(1,3),(2,1),(2,3),(3,1),(3,2)\}$, $D_1 = \{(1,2),(1,3),(2,3)\}$ and $D_2 = \{(1,3)\}$. A core node is designated by the triplet (i, r, e).

Definition 8.2.1. A *kit* is composed of a core node (i, r, e), $i \in N, r \in V, e \in E$, and a subset D_k of edge node pairs.

A kit implies that the edge node pairs of D_k are assigned to the core node (i, r, e), i.e., each edge node pair of D_k commutes its traffic through the core node (i, r, e). In other words, in a kit D_k represents the set of all edge node pairs that are assigned to core node (i, r, e) for a given network configuration. The core node (i, r, e) and its assigned edge node pairs D_k will be denoted by $((i, r, e), D_k)$.

Definition 8.2.2. A *feasible kit* $((i, r, e), D_k)$ is such that the capacity constraints of the links between each origin edge node of D_k and the core node (i, r, e), and the capacity constraints of the links between the core node (i, r, e) and each destination edge node of D_k are satisfied.

Definition 8.2.3. A *packing* is a union of feasible kits.

Let $((i_1, r_1, e_1), D_1)$ and $((i_2, r_2, e_2), D_2)$ be two feasible kits. $((i_1, r_1, e_1), D_1)$ is composed of the core node (i_1, r_1, e_1) and the edge node pairs of D_1. $((i_2, r_2, e_2), D_2)$ is composed of the core node (i_2, r_2, e_2) and the edge node pairs of D_2.
These two kits form a packing Π if the following is true :
$((i_1, r_1, e_1), D_1), ((i_2, r_2, e_2), D_2) \in \Pi \Leftrightarrow (i_1, r_1, e_1) \neq (i_2, r_2, e_2) \wedge D_1 \cap D_2 = \emptyset$.

Given a packing Π: L_1 is the set of core nodes that are not active, i. e. that do not commute traffic, $L_1 = \{(i, r, e) \mid \forall D_k \subset T, ((i, r, e), D_k) \notin \Pi\}$; L_2 is the set of edge node pairs that are not assigned to a core node,$L_2 = \bigcup_{p \notin J_k} p \in T$ with $J_k = \bigcup_{(i,r,e,D_k) \in \Pi} p \in D_k$; L_3 is the set of active core nodes with their associated edge node pairs, i. e. the set of feasible kits, $L_3 = \Pi$.

Let us assume that L_1 has n_1 elements, L_2 has n_2 elements, and L_3 has n_3 elements. In Fig. 8.2, e.g., $n_1 = 2$, $n_2 = 3$ and $n_3 = 2$... Fig. 8.2 shows a packing Π whose cost can be determined as the sum of all the terms of objective function (8.1) applied *only* to the kits of L_3 plus a penalty cost for the unassigned pairs in L_2, $\mathcal{M} * n_2$, where $\mathcal{M}$ is a very large number.

In a repeated matching approach, we want to match elements of L_1, L_2 and L_3 so as to generate new sets L'_1, L'_2 and L'_3 that have a lower total cost.

The matching problem

The classical matching problem can be described as follows: let A be a set of q elements $h_1, h_2, ..., h_q$. Each $h_i \in A$ can be matched with only one $h_j \in A$. An element can be matched with itself, which means that it remains unmatched. Let c_{ij} be the cost of matching h_i with h_j. We have $c_{ij} = c_{ji}$. We introduce the binary variable z_{ij} that is equal to 1 if h_i is matched with h_j. The matching problem consists in finding the matching over A that minimizes the total cost:

$$\min \sum_{i=1}^{q} \sum_{j=1}^{q} c_{ij} z_{ij} \tag{8.9}$$

$$s.t. \quad \sum_{j=1}^{q} z_{ij} = 1, \quad i = 1, ..., q \tag{8.10}$$

$$\sum_{i=1}^{q} z_{ij} = 1, \quad j = 1, ..., q \tag{8.11}$$

$$z_{ij} = z_{ji}, \quad i, j = 1, ..., q \tag{8.12}$$

$$z_{ij} \in \{0, 1\} \tag{8.13}$$

(8.10) and (8.11) ensure that each element is exactly matched with another one. (8.12) ensures that if h_i is matched with h_j, then h_j is matched with h_i. (8.13) indicates that variable z_{ij} is binary.

In our heuristic, one matching problem is solved at each iteration between the elements of L_1, L_2 and L_3. At each iteration, the number of elements to be matched is $n_1 + n_2 + n_3$, where n_1, n_2 and n_3 are the current cardinalities. For each matching problem, the costs c_{ij} have to be evaluated. The cost c_{ij} is the cost of the resulting packing after having matched element h_i of L_1, L_2 or L_3 with element h_j.

The costs c_{ij} are stored in a matrix C, whose dimension $(n_1+n_2+n_3) \times (n_1+n_2+n_3)$ changes at each iteration. C is symmetric and composed of nine sub-matrix. Given the symmetry, only six blocks have to be considered. The notation $[L_i - L_j]$ is used to indicate the matching between the elements of L_i and L_j.

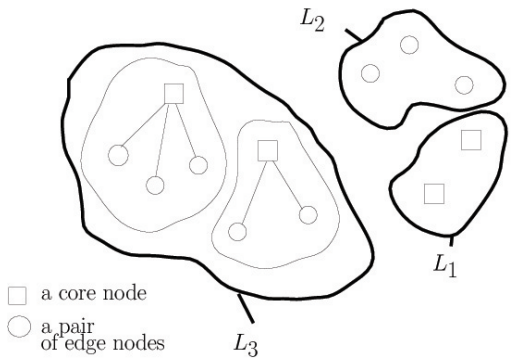


Figure 8.2: The sets L_1, L_2 et L_3 associated with a packing Π

$$C = \begin{pmatrix} \text{[L1-L1]} & \text{[-]} & \text{[-]} \\ \text{[L2-L1][L2-L2]} & \text{[-]} & \\ \text{[L3-L1][L3-L2][L3-L3]} & & \end{pmatrix} = \begin{pmatrix} \text{[1][-][-]} \\ \text{[2][3][-]} \\ \text{[4][5][6]} \end{pmatrix}$$

To avoid a matching between two elements, the matching cost is set to infinity (very high value in practice). This happens when capacity constraints on links or core nodes would not be respected, and when the matching involve the same element for blocks 1, 3 and 6. Furthermore, a matching between two elements can produce several results. In such a case, the result with minimal cost is chosen. We develop the matching costs for each block in the Appendix B.

Once the cost matrix is calculated, the matching problem (8.9)-(8.13) is solved heuristically. The resolution is not easy because of the symmetry constraint (8.12). We have implemented the algorithm of Forbes [54] that is based on the method of Engquist [53]. The starting point for Forbes' algorithm is the solution vector of the matching problem without the symmetry constraint (8.12); such a starting solution is obtained with the algorithm of Jonker and Volgenant [55] that was chosen for its speed performance. The output is a symmetric solution vector that indicates the matchings to be performed between the heuristic elements.

The solution is then analyzed. Some matchings result in new elements in L'_1, L'_2 and L'_3 whereas other elements disappear. For example, the matching between an inactive core node (i, r, e) of L_1 and an unassigned edge node pair p of L_2 results in the new element $((i, r, e), D = \{p\})$ of L_3.

8.2.2 A repeated matching heuristic

A global chart of the heuristic is given in Fig. 8.3.

- **Step 0** The algorithm starts with a feasible packing. We choose a packing where no core node is opened and no edge node pair is assigned: $L_1 = \{$all potential core nodes$\}$, $L_2 = \{$all origin/destination edge node pairs exchanging traffic$\}$, $L_3 = \emptyset$.
- **Step 1**: A series of feasible packings with decreasing cost is formed.
  - **Step 1.1**: The cost matrix C is calculated for every block.
  - **Step 1.2**: Then, the problem of finding the less costly matchings between the element of C is solved. If those matchings improve the packing cost, a new packing can be built by applying the matchings to the current packing.

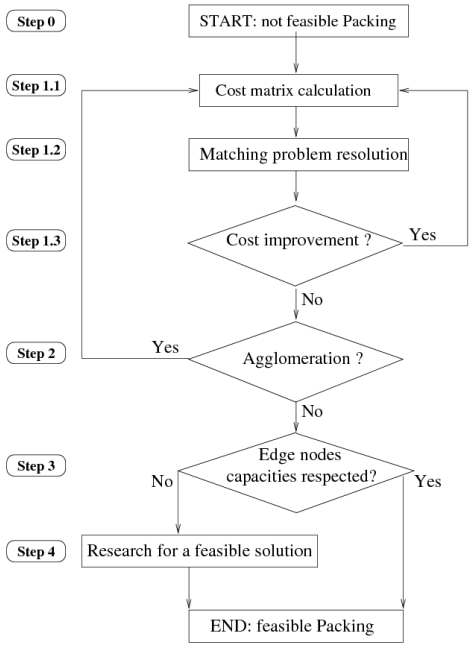


Figure 8.3: Chart of the repeated matching heuristic for the Petaweb design.

– **Step 1.3**: When the cost of the packing can not be reduced any more, i.e., when the matching results do not produce cost improvement for the current packing, then Step 2.

Step 2 the heuristic checks if the active core nodes can be agglomerated so as to take into account the scale economy in the core node cost. Given that $s_1 = 1, s_2 = 2$ and $s_3 = 4$ a core node of type 2 opened at a site presents the same capacity but it is less expensive than two core nodes of type 1. The same can be said for one type 3 compared with two type 2 core nodes. We underline that the heuristic could not do these agglomerations while building packings with lower cost. If at least one agglomeration is possible, a new packing is generated and the iterations are re-started. Such a process is repeated until no progress can be done.

- **Step 3**: Finally, one constraint must yet be verified: the edge node capacity constraint. This constraint has been omitted by now in order to allow multiple little kits to be built at the beginning of the algorithm and then be agglomerated.

- **Step 4**: Knowing the active core nodes in the current best solution, we verify if (8.5) is respected. If so, the heuristic stops, otherwise it searches for a feasible solution in restricting the number of active core nodes, as follows.

  If one edge node capacity is exceeded by one fiber, a core node of type 1 or the equivalent capacity must be closed in the network. Step by step, at each site, the equivalent of a core node of type 1 is closed and the optimal assignment of all edge node pairs to the core nodes remaining active is calculated. This assignment must verify the capacity of each core node still active and the link capacity between each edge node and each active core node. The optimal assignment is solved by ILP (CPLEX). Whenever the equivalent of a core node of type 1 is closed at one site, the total cost of the network with optimal assignment of the pairs is calculated. Finally, we choose the solution with the lowest total network cost.

If one edge node capacity is exceeded by two fibers, a core node of type 2 or the equivalent capacity must be closed in the network. Each combination is tried to close the equivalent of a core node of type 2 in the network.

If one edge node capacity is exceeded by more than two fibers, we randomly choose the core nodes that will be reduced in capacity or entirely closed.

Complexity

The complexity of the whole heuristic depends on its different sub-algorithms and phases. The calculation of the cost matrix is straightforward except for two blocks of the matrix (see block 5 and 6 in Appendix I) where a polynomial swapping problem depends on the number of connections in the network.

The resolution of the matching problem operates on the cost matrix through the Forbes' and the Volgenant's algorithms. In the worst case, the first has a $O(n^3)$ complexity while the second one $O(n^2)$, where $n = n_1 + n_2 + n_3$. The Forbes' algorithm looks for a symmetric matching vector starting from the Volgenant's a-symmetric solution vector; the algorithm creates a branch-and-bound tree whose dimensions increase during the research of a symmetric solution. But, in order to avoid excessive researches, we controlled the dimensions of the tree: when they pass a higher fixed bound without finding a solution, a non-optimal solution with a forced symmetry is given back. Thus, the complexity of the matching resolution phase is kept under control by introducing sub-optimal solutions. Not bad, since we deal with a heuristic that resolves a succession of matching problems. The higher bound for the research tree was fixed to 1000 tree children.

Numerical results

In Sect. 9.1 we present and discuss the results obtained from numerical simulations.

Summarizing, we show that the designed heuristic proved to be very efficient and scalable. For those network sizes for which we could find a lower bound, the heuristic presented an optimum gap of 5.5%. As expected, the solution time increases exponentially with the network size but some traffic matrix cases are more difficult to solve than others. Moreover, we show that the fiber accounts for up to 80% of the total costs. This is not surprising given that one of the shortcomings of the proposed architecture is precisely the large number of fiber connections that have to be established between the edge nodes. However, when changing the fiber cost function so that it is less dependent on the number of wavelengths per fiber, we found that the percentage of fiber costs could go down to 46%.

8.3 A quasi-regular Petaweb structure

In the dimensioned network only a portion of the transport capacity is to be reasonably assigned. Many optical fibers or links may be totally unused for the following reason: the traffic matrix contain a few peaks of traffic between two sites, and a lot of medium-low values; the peaks will induce high utilization of those optical links connected to the connection request's sites, while the other optical links used for low-rate CRs are to be under-used. Indeed, the activation of a switching plane in the network requires installing one fiber for every edge node.

In order to cope with this inefficiencies, a quasi-regular structure can be extracted from the optimal regular dimensioned network. The quasi-regular structure as attainable with future re-installations of the disabled fibers. After this removal, the final composite-star topology becomes irregular, or, better, quasi-regular. The physical connection between an Edge Node and a Core Node (or EN and CN in the sequel) may become partial, but sufficient for the reserved traffic.

The cost of the quasi-regular structure is expected to be significantly lower than the regular structure cost because the unused fibers are disabled in conjunction to the corresponding CN ports. An additional physical hypothesis should, however, be assumed with quasi-regular topologies: the switching planes of a same core node with several switching planes should be able to communicate to each other in order to multiplex lightpaths on the same fiber (if required).

8.4 TDM/WDM switching implementation

The link capacity constraints (8.6) and (8.7) implicitly assume that a form of TDM over WDM is performed. Connection requests from the same origin or toward the same destination nodes can be groomed in the same optical channel as far as their traffic volume sum does not exceed the link capacity. However, as the traffic matrix may present a profile with a long tail of low-rate connection requests and a few peaks (as in our considered datasets), the assumed end-to-end TDM/WDM grooming may reveal to be inefficient. In the following, we study how TDM/WDM switching enabled in core nodes can grant benefits to both network cost and resource utilization.

In [153] Blouin et al. assumed the use of TDM in the Petaweb to realize sub-channels within a lambda-channel. They proposed the use of electronic core nodes. In order to model TDM switching, we assume the switch architecture with time-slot interchanging functionalities that has been proposed by Huang et al. [154]. They introduced an all-optical TDM Wavelength Space Routers (TWSR) (e.g., resumed in [155]) where the time-slot switching is implemented using Optical Time-Division Space Switches (OTDSSs); the alignment of the time-slots can be done by the schemes of input synchronizer described in [148]. An OTDSS can reconfigure itself with the granularity of a time-slot and multiplexes in a time-slot basis. No buffering operations are required and a local access unit guides the alignment of the incoming time-slots. The TWSR is composed of a number of OTDSS equal to the number of used wavelengths and each OTDSS manages the time-slots of the same wavelength; therefore this kind of switch can easily replace the wavelength-driven space switch presented in Sect. 8.1, and we will thus assume the TWSR as the switching plane unit. In what follows, as suggested in [154], we call a time-slotted lightpath, a *ts-lightpath* (TLP).

As it emerges from specification of G.709 of the ITU-T [185], the use of TDM in WDM networks is useful for two main reasons: fractionating the lambda-channel into more sub-channels improves the network capacity utilization, the requested resources are smaller and the network cost is more competitive; many transport levels can be used, e.g., ITU-T G.709 defines, for the OTN interfaces, three different traffic levels. Therefore, we have to choose the number of TLP classes and the transport capacity for every class. In [156] the authors face this problem in the design of a WDM network with static traffic load and TDM channel partitioning. Referring to the Optical Transport Unit (OTU) hierarchy specifications, they chose three transport classes that correspond to the three OTU-rates (2.5 Gb/s, 10 Gb/s and 40 Gb/s). This choice is justified with the fact that, for the moment, operators use transmission systems with fixed transmission rates.

8.4.1 Time-slotted lightpath hierarchy

We introduce here a three-level hierarchy for a ts-lightpath inspired from the OTU hierarchy. However, we will not limit our choice of bit-rate values to the OTU rates. Let Z_h denote the transport capacity of a TLP of class h (TLP-h). A CR is best served using the minimum number of TLPs. This means that the traffic volume of a CR is rounded up to a value that is the sum of the transport capacity of several TLPs that can even be of different classes; the number and the classes of these TLPs is such that the rounding up value is as close as possible to the traffic of the CR. Let us assume that the rules to set the transport capacity Z_h are:

- $Z_1 = \frac{1}{2^n} C_{ch}$, $n \in \mathbf{N}$, represents the transport capacity of a time-slot; it is a fraction, multiple of 2, of the transport capacity of a wavelength (C_{ch});
- $Z_2 = C_{ch}$, that is the transport capacity of a wavelength
- $Z_3 = W\, C_{ch}$, that is the transport capacity of a fiber

The above capacities have been chosen so that switching operations, already simplified by the Petaweb architecture, are further simplified. An optical switch can commute a TLP-1 in a time-slot basis, and a TLP-2 switching the whole wavelength without any time-slot alignment; to switch a TLP-3 one may simply interconnect the incoming fiber with an outgoing fiber without any demultiplexing/multiplexing. The TLPs are then switched and transported independently and the original data flow is then recomposed at the destination EN.

The association to a TLP of a propagation delay cost directly proportional to its bit-rate, besides that to traveled path length, is an important novelty for the design of backbones with QoS) guarantee: the minimization of the global network cost will give priority to high bit-rate classes in getting the bandwidth over short paths. Indeed, high bit-rate lightpaths may accommodate traffic, belonging to video streaming and voice-over-ip services, which need the lowest possible propagation delay. An operator offering VoD services, e.g., and having a few video pump station in its geographical network, will present high bit-rate connection between the PoPs of the pump stations and the PoPs of the clients. Similarly, the gateway to public switched telephone networks may be located not in all the PoPs and the calls would be aggregated in high bit-rate flows. Therefore, with our model we can guarantee to high bit-rate CRs the priority for the attainment of short paths.

Constraints

The Petaweb design has to preserve the characteristics of network components and to simplify routing operations. With TDM/WDM switching, we can distinguish between Capacity and Coherency Constraints.

For edge nodes and optical links, the *Capacity Constraints* are modular, i.e., can be allocated and incremented only through discrete quantities: the optical link capacity can be increased by a multiple of the capacity of W lambda-channels at a time and must be verified for both directions; the capacity of an EN depends on the number of optical fibers connected to it.

Furthermore, we have to consider additional constraints to satisfy basic communication system requirements in terms of delay and buffering operations; we call these constraints *Coherency Constraints*:

1. All the TLPs of a CR should be transported on the same optical trunk line;

2. All the time-slots of a TLP should be transported on the same optical link;

3. All the TLPs of a CR should be transported contiguously in the time and in the frequency domains.

The 1^{st} Coherency Constraint assures that the traffic between two ENs is switched in the same site. Without this constraint one may loose too much time at the destination EN because of out-of-order buffering operations; two TLPs of the same connection may be switched in different sites cumulating different propagation delays. The 2^{nd} Coherency Constraint imposes that the time-slots used by a TLP must be switched in the same core node. The 3^{rd} has been introduced for three main reasons: to ease multiplexing/demultiplexing operations; to lighten buffering operations at destination ENs; to relate lost data in case of damage of a single switching plane to the minimum possible number of CRs.

8.4.2 Refinements to the design problem

The optimization problem consists in finding the best composite-star physical topology for the given set of TLPs, respecting the network model and the peculiar composite-star architecture. A pre-processing phase produces the optimal set of TLPs for the assigned traffic matrix; the resulting set drives the dimensioning of the physical topology and the assignment of its resources to the TLPs.

Dealing with the optimization of WDM networks with TDM channel partitioning upon a pre-assigned physical topology and with static CRs, the authors in [156] call their design problem RFWTA (Route, Fiber, Wavelength and Time-slot Assignment), keeping as the objective the minimization of the total number of fibers, which is the only variable in the physical topology. In our design problem we also need to determine an efficient routing and assignment of wavelengths and time-slots for a set of pre-assigned CRs. The Petaweb network design with TDM/WDM shall be seen as a joint dimensioning and assignment problem. We divide it into two sub-problems: Route and fiber Allocation (RFA), which treats the allocation of the resources guaranteeing an efficient routing, and Wavelength and Time-slot Assignment (WTA), which concerns the assignment of the allocated resources.

RFA algorithm

The task of the RFA problem is to find the optimal location of network components in order to efficiently switch all the TLPs of the virtual topology; every TLP has to be assigned to its switching CN so that its route and optical links are decided. The complete ILP formulation for the RFA problem resolution is reported in Appendix B.2. This formulation presents some fundamental differences from the previously presented one in order to manage the TDM/WDM switching and classes; the first two terms of the objective remain unchanged but the third determines the propagation delay cost considering the discrete capacity Z_h of a ts-lightpath. The resulting complexity is high, but not prohibiting for the assigned instances; we could control the number of TLPs, and thus the number of variables and constraints, by grooming end-to-end TLP-1s and TLP-2s of the same CR in sub-classes consisting of, respectively, 4 time-slots and 4 wavelengths.

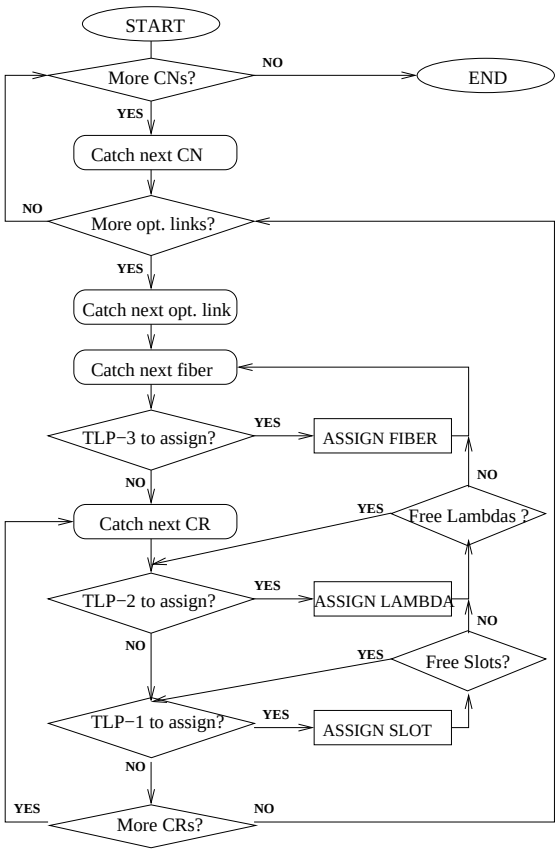


Figure 8.4: Flow chart for WTA resolution algorithm

WTA algorithm

The result of the RFA problem is the optimal location of the CNs to switch all the TLPs at minimum cost. The results contain, thus, the set of TLPs for every CN and, consequently, the assignment of every TLP to two optical links, one between the origin edge node and the CN, and one between the CN and the destination edge node. The task of the WTA problem is to assign to every TLP a subset of the wavelengths and the time-slots of its optical links. Let us remember that a TLP-3 needs one fiber, a TLP-2 one wavelength and a TLP-1 one time-slot. The only constraint to respect is the 3^{rd} Coherency Constraint.

The WTA flow chart is showed in Fig. 8.4. The input are the sets of enabled CNs, CN-ENs optical links for every CN, fibers for every optical link, and TLPs assigned to every optical link (grouped according to their class and CR). The algorithm starts considering one optical link at a time and assigns time-slots, wavelengths and fibers to the TLPs. As it can be noticed from the flow chart, the assignment of whole fibers to TLP-3s is done independently of their CR.

TDM/WDM for the quasi-regular structure

For the quasi-regular structure, we pass to the WTA algorithm only the fibers really exploitable, the others are ignored. For every optical link, we determine the number of fibers needed by the assigned TLPs, and if it is inferior to s_r the superfluous fibers are no longer considered. This operation does not require an undifferentiated grooming of

TLPs of different optical links over the same fibers. Obviously, the extraction of the quasi-regular topology does not require changes to the WTA, because one just changes the fiber number for the optical links.

Numerical results

In Sect. 9.2 we present and discuss the results obtained from numerical simulations.

Summarizing, the results show that there is only a small difference in cost, with a slightly higher core node cost and lower delay cost with the **A** traffic profiles, and conversely for the **B** profiles. The fiber cost ratio remains very high, that is roughly around 80%. Passing from the regular to the downgraded quasi-regular structure the network cost gets more than halved, and the fiber ratio passes to roughly 60%.

We also reported the average network utilization in the network, always under 20% for the regular Petaweb structure, and around 50% for the quasi-regular structure. It is worth remarking that while in IP/MPLS network optimization under-utilized links are desirable, in optical network dimensioning a high average link utilization is, instead, desirable because it grants that the investment produced an optimized transport facility. In this sense, the quasi-regular structure allows large investment savings with a still acceptable resource utilization.

Nevertheless, with a quasi-regular topology there may be edge nodes that are connected through only one trunk line to the transport network (e.g., see Fig. 9.10). A failure to the trunk line would totally disconnect the edge node. Another reliability issue may appear whether all the CNs are installed in a few sites, which is quite possible also with regular structures for small networks. An outage in the switching site might interrupt large volumes of traffic. It is thus needed to conceive protection strategy for the Petaweb architecture.

8.5 Protection strategies

Figures 9.10 and 9.11 report optimized 10-node Petaweb topologies. The regular topology has all the optical links enabled, and the optical links connected to a CN-r are composed of s_r fibers. In this solution, the CNs are located at Philadelphia and Washington, but both sites have also ENs. The quasi-regular topology, contemplates the deactivation of those unused fibers in the optimized regular topology; in this way, the network cost is reduced more than 55% and the network utilization more than duplicates. The quasi-regular topology can be thus determined heuristically by removing unused fibers and ports from an optimal regular Petaweb topology, without any change in the lightpath routes and switching schemes.

The drawback is that such a topology is not reliable. The Tallahassee and Albany edge nodes in Fig. 9.10b, e.g., are connected to the core through only one trunk line. In the case of failure of one of these trunk lines and if the network operator wants to adopt a restoration strategy, those ENs would remain totally isolated from the network and their outgoing and incoming traffic could not be restored at all. Moreover, consider the case where all the CNs are located in the same switching site (likely for small networks): all the ENs would be connected to the transport network through only one trunk line. It is thus necessary to introduce a protection strategy, which for every working lightpath provides a link-disjoint protection lightpath. Finally, it is worth noting that even if a switching plane failure affects only the lightpaths switched there, a failure of a whole CN or switching site, e.g., caused by facility flooding or power failure, might be a disaster.

8.5.1 Dedicated path protection

The willing is to design a reliable Petaweb transport network offering restoration functions for its lightpaths and nodes directly in the physical layer. Restoration techniques already exist in the electronic layers (IP, TCP, ATM, SDH), but, even if effective, fo high bit-rate they require signaling procedures that slow down the restoration [149]. We need to limit the restoration time because the Petaweb trunk lines may transport optical flows at a Tb/s rate, belonging to various CRs.

We excluded link protection techniques because with the Petaweb architecture the installation of backup trunk lines would require a replica of all the CNs in other switching sites at an enormous cost. It is required a path protection strategy that does not alter its qualities in terms of simplicity or its switching schemes, and that protects in the case of single trunk line failures. The protection functions are performed at the ENs and we shall avoid excessive signaling operations involving the CNs. Because of the high working rate of transmission and switching equipment, any ms of time elapsed for restoration may imply Tb/s of data loss. In the event of a trunk line failure, a long signaling phase to establish the protection paths for all the affected TLPs would be required; and it would involve not only the CNs and the EN connected by the faulty line, but even CNs of other switching sites candidate for the TLPs routing. Moreover, the ENs affected by the failure should have knowledge of the actual lightpath topology and routing schemes in order to provide for physical reconfigurations of CNs and compute in some way the optimal protection path. Thus, this signaling phase would be reasonably long, and also hazardous because the CN reconfiguration may not be successful.

We chose a Dedicated Path Protection (DPP) strategy: the protected signal is sent over two separate allocated paths, then the receiver selects one among them. For every working ts-lightpath (wTLP) of our network model we have to allocate a protection ts-lightpath (pTLP). In the 1+1 DPP case the signal is split at the origin over two disjoint paths, and thus there would not be a signaling phase. In the case of 1:1 DPP the pTLP is sent on the protection path only when the failure occurs, and it makes sense to enable a shorter path for wTLPs, and a longer path to pTLPs to be used in the case of failure along the working one. For this reason the optimization problem should give priority to wTLPs in the contention for short paths. A shared path protection would for sure guarantee a less expensive network requiring fewer resources, but the signaling would be important and would involve CNs and ENs introducing a significant restoration delay.

In the case of one trunk line failure all the wTLPs must be recovered from the allocated pTLPs. The DPP strategy requires the use of a *protection constraint*: every pTLP must be multiplexed on trunk lines different than those of the corresponding wTLP; in the Petaweb architecture this means that a pTLP must be switched in a different network site than that of its wTLP.

The application of path protection, with the above site-disjointness constraint, guarantees even *node protection* and *switching site protection*: if a core node fails, the TLPs it routed will propagate the failure to the destination edge nodes, which will recover the traffic from the corresponding protection paths; similarly, if a whole switching site is damaged and disconnected in the case of disasters, all the traffic its core nodes switched will be recovered and routed elsewhere. Therefore, even if the DPP constraint is supposed to increase the network cost and the equipment to be deployed, it offers in this specific one-hop optical architecture not only path protection, but also node and site protection. The application of node and site protection constraints in classical mesh networks create irregularities that generate or worsen the bottleneck in some spare links. For the Petaweb, this can not happen.

8.5.2 Refinements to the design problem

The WTA algorithm is transparent to the adoption of a DPP protection strategy. Indeed, working and protection ts-lightpaths are already allocated to disjoint optical links by the RFA optimization. The refined ILP formulation is in Appendix B.3, where a relaxed formulation with lower complexity is also proposed. Namely, the objective function scales the propagation delay cost of pTLPs in order to assign them longer paths than the corresponding wTLPs path. This not only allows the assignment of the best paths to the wTLPs, but also the minimization of the fiber distance for wTLPs and thus their outage risk.

Numerical results

In Sect. 9.3 we present and discuss the results obtained from numerical simulations.

Summarizing, the DPP constraints has a large impact on network cost, higher with the quasi-regular structure as expected. This is due to a larger number of CNs and switching sites that is forced. In the regular case, the network utilization increases, while for the quasi-regular case it decreases. This is probably due to the heuristic way used to retrieve the quasi-regular structure.

8.6 Optimal quasi-regular Petaweb design

After the optimal dimensioning of a regular Petaweb, a quasi-regular topology can thus be retrieved by simply dropping unused fibers and ports. However, it is important to highlight that this approach does not guarantee the optimality of the obtained quasi-regular topology. In the sequel, we tackle the problem of directly optimizing a quasi-regular Petaweb and propose a three-step heuristic.

8.6.1 ILP formulation

We complete the notation already introduced as follows:

$w(p)$: the bandwidth consumption of p in number of time-slots.

$\delta < 1$: the scaling factor for pTLPs delay costs.

$C_r = (Z_2/Z_1)Ws_r$: capacity of a core node of type r.

a_p^i, $\widetilde{a_p^i}$: binary variable equal to 1 if CR p is switched at site i; $\widetilde{a_p^i}$ relats to the corresponding protection flow.

y_{ir}: integer variable equal to the number of core nodes of type r to install at site i.

$f_{p(ire)}$ is the number of timeslots uploaded by the source node for CR p. to the e^{th} CN-r at site i (also equal to the number of downloaded timeslots).

$l^u_{j(ire)}$ (resp. $l^d_{(ire)j}$) is the number of fibers connecting EN j and CN (ire) (resp. CN (ire) and EN j).

All the capacities are measured in time slots; these terms are computed by dividing each value by Z_1 and rounding down to the nearest integer. In a similar way, each bandwidth consumption value $w(p)$ is measured in time slots, dividing by Z_1 and rounding up to the

nearest integer. The exact optimization of the quasi-regular topology can be pursued by solving the following ILP:

$$\begin{aligned}
&\min G(y,l,a,\widetilde{a}) = \sum_{i\in N}\sum_{r\in V} f_r y_{ir} + \\
+ &\sum_{j\in N}\sum_{i\in N}\sum_{r,e} (\phi(W)\Delta_{ij}F + W\gamma^{s_r-1}P)(l^u_{j(ire)} + l^d_{(ire)j}) \\
+ &\sum_{i\in N}\sum_{p\in T} \beta d_{ip} w(p) a^i_p + \sum_{i\in N}\sum_{p\in T} \delta\beta d_{ip} w(p) \widetilde{a^i_p} \qquad (8.14)
\end{aligned}$$

$$s.t. \sum_{i\in N} a^i_q = 1, \quad \forall p \in T \qquad (8.15)$$

$$\sum_{i\in N} \widetilde{a^i_p} = 1, \quad \forall p \in T \qquad (8.16)$$

$$a^i_p + \widetilde{a^i_p} \le 1, \quad \forall p \in T, \forall i \in N \qquad (8.17)$$

$$\sum_{re} f_{p(ire)} = w(q)(a^i_p + \widetilde{a^i_p}), \quad \forall p \in T, \forall i \in N \qquad (8.18)$$

$$\sum_{p}^{s(p)=j} f_{p(ire)} \le C_{ch} W l^u_{j(ire)}, \quad \forall j \in N, \forall (ire) \qquad (8.19)$$

$$\sum_{p}^{t(p)=j} f_{p(ire)} \le C_{ch} W l^d_{(ire)j}, \quad \forall j \in N, \forall (ire) \qquad (8.20)$$

$$\sum_{j\in N}\sum_{r,e} l^u_{j(ire)} \le \sum_{r\in V} s_r y_{ir}, \quad \forall i \in N \qquad (8.21)$$

$$\sum_{j\in N}\sum_{r,e} l^d_{(ire)j} \le \sum_{r\in V} s_r y_{ir}, \quad \forall i \in N \qquad (8.22)$$

$$\sum_{i\in N}\sum_{r\in V} C_r y_{ir} \le \max C_j \qquad (8.23)$$

$$y_{ir} \in Z_+, a^i_p, \widetilde{a^i_p} \in \{0,1\}, l_{ij} \in Z_+, f_{p(ire)} \in Z_+ \qquad (8.24)$$

The objective (8.14) is composed of the fixed cost of enabled CNs, the cost of single unidirectional fibers going to and coming from core nodes, and the propagation delay cost of working and protection TLPs. Constraints (8.15)-(8.17), (8.23) and (8.24) have the same meaning as (B.24)-(B.26), (B.29) and (B.30) of the previous formulation. Constraints (8.18), together with integrality conditions on the a^i_q and $\tilde{a}^i_q$ variables, ensure that the whole traffic of each connection request is switched in the same site. Constraints (8.19) and (8.20) model the capacity of installed fibers. (8.21) and (8.22) ensure that the number of fibers entering and exiting each site is coherent with the number of available switching planes.

8.6.2 A three-step heuristic

The formulation (8.14)-(8.24) is hard to optimize for a general purpose solver. Therefore, we devised a special-purpose heuristic algorithm.

First step We compute the optimal regular network by solving model (B.23)–(B.30). We indicate by $T^w(i) \subseteq T$ the set of connection requests having wTLPs assigned to a core node in site i in an optimal regular network. Similarly, we indicate by $T^p(i) \subseteq T$ the set of connection requests having pTLPs assigned to a core node in site i. Then, as in the seminal work of Cooper [50] on location problems, we proceed to an iterative two-step local search.

Second step Each site hosting core nodes is optimized independently. This requires, for each site i, fixing all the variables a_p^i and $\tilde{a}_p^i$ as follows:

$$a_q^i = \begin{cases} 1 & \text{if } p \in T^w(i) \\ 0 & \text{otherwise} \end{cases} \quad \tilde{a}_p^i = \begin{cases} 1 & \text{if } p \in T^p(i) \\ 0 & \text{otherwise} \end{cases}$$

and to solve Single-Site restrictions of the previous optimization Problem (SSOP):

$$\begin{aligned} \text{SSOP}(i) : \min s(y,l) = &\sum_{r \in V} f_r y_{ir} + \\ &+ \sum_{j \in N} \sum_{r,e} (\phi(W)\Delta_{ij} F + W\gamma^{s_r - 1} P) l^u_{j(ire)} \\ &+ \sum_{j \in N} \sum_{r,e} (\phi(W)\Delta_{ij} F + W\gamma^{s_r - 1} P) l^d_{(ire)j} \end{aligned}$$

subject to (8.18)-(8.24).

Third step Each switch installed in the second step is shifted to the location that minimizes the fiber and delay costs. The best location i^* of each core node during the second step is found by solving a set of (Weighted) Median Problems (WMP), one for each core node e of type r at site i:

$$\text{WMP}(ire) : i^* = \text{argmin}_{k \in J} \sum_{j \in J} (l^u_{j(ire)} \cdot d_{jk} + l^d_{(ire)j} \cdot d_{kj})$$

each WMP can easily be solved by inspection. Whenever a core node (ire) is moved during this step ($i^* \neq i$), the wTLPs of certain connection requests may be removed from i

$$T^w(i) := T^w(i) \setminus \{p \in T^w(i) | f_{p(ire)} > 0\}$$

and assigned to i^*

$$T^w(i^*) := T^w(i^*) \cup \{p \in T^w(i) | f_{p(ire)} > 0\}.$$

An analogous update is performed for the pTLPs; however, in order to prevent from using the same switching site for both the wTLPs and pTLPs of certain connection requests, when $p \in T^p(i)$ and $p \in T^w(i^*)$ the shift of the pTLPs is forbidden. In a similar way, if $p \in T^w(i)$ and $p \in T^p(i)$ the real and protection paths are swapped.

$$\begin{aligned} T^w(i) := T^w(i) \setminus \{p\}; T^p(i) := T^p(i) \cup \{p\} \\ T^p(i^*) := T^p(i^*) \setminus \{p\}; T^w(i^*) := T^w(i^*) \cup \{p\} \end{aligned}$$

Therefore, even connection requests with a small fraction of traffic switched by (ire) are shifted to i^*. Even if in a single iteration the solution cost may increase, this strategy helps the algorithm to escape local minima. Due to constraint (8.23), this shifting may even yield infeasible subproblems.

The second and third steps are iterated, until no more changes to the solution are made, or the algorithm encounters a previously visited solution.

Numerical results

In Sect. 9.4 we present and discuss the results obtained from numerical simulations.

Summarising, the experimental results showed that the direct quasi-regular Petaweb optimization can improve the network cost by roughly 70% with respect to the regular structure, and by roughly 30% with respect to a quasi-regular structure heuristically determined by removing unused equipment from the optimal regular structure. Moreover, the proposed heuristic for the quasi-regular optimization showed quite good computational results even for large networks.

8.7 Comparison with multi-hop core networks

An extensive comparison between the quasi-regular Petaweb and classical multi-hop irregular mesh architectures is needed, to finally assess the convenience of the Petaweb architecture in that which appears its most efficient structure. Previous studies of Blouin et al. [153] compared the regular Petaweb with multi-hop architectures, in the case of changing demands. They concluded that roughly 17% more fiber-km is needed for the regular Petaweb, while requiring roughly 66% less ports. However, the meaning of that comparison concerns the regular structure and is limited by the fact that the authors forced a five-stage structure for the multi-hop network core, limiting thus the degrees of freedom. Our objective is, instead, to compare the *quasi-regular* Petaweb to multi-hop irregular mesh structures, in terms of network cost, network utilization and fiber length. We leave all the degrees of freedom to the design dimensioning for the multi-hop case.

8.7.1 Switching Systems

In order to fairly compare the quasi-regular Petaweb to irregular mesh structures we need to use comparable switching systems, described in the following.

Petaweb's Core Node

Previously we studied a hierarchy of switching core nodes, with a cost decreasing with the size, was considered in the modeling. To perform a fair comparison we shall not adopt this hierarchy. For the same reason, we now assume that the fibers have the same number of wavelengths, W, and thus that each fiber requires W ports to be connected to a switching plane.

In the quasi-regular structure, it is possible to relax the constraint imposing each edge node to be connected to each core node with a single fiber. Unused fibers and ports can be removed to reduce the physical cost without modifying the switching scheme. In this case, wavelength conversion is still not necessary. However, whether at a given switching site there are many core nodes, with a quasi-regular structure we may further disable fibers by multiplexing/demultiplexing wavelength-channels switched by different core nodes, by allowing thus wavelength conversion. Indeed, the relaxation of the wavelength continuity constraint allows wavelength-channels coming from different edge nodes, but with the same destination, to be multiplexed into the same fiber (idem for the ingress stages).

Multi-hop network's Core Node

In this case a core node can have different sizes and a site can have a single core node. As illustrated in Fig. 8.5b, the switching plane has a number of incoming and outgoing fibers

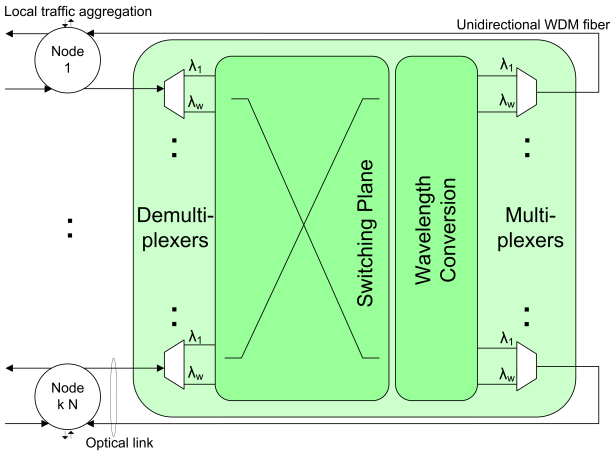


Figure 8.5: Multi-hop Core Node Structure. N: number of edge nodes

multiple of the number of edge nodes, and offers full wavelength conversion. For example, in a 10-node network a switching plane of size k can house $10 \cdot k$ ingress/egress fibers, independently of the origin/destination node, which can be an edge node or another core node. fiber links and ports can be disabled if unused. fiber links can be *access fiber links* connecting core nodes to edge nodes (and vice-versa), and *core fiber links* interconnecting two core nodes. Certainly, the number of core links can be bigger than the number of edge links since this might facilitates path finding. Also for this reason, we let the design procedure dimension the number of edge and core links at each core node.

8.7.2 Multi-hop Network Design Dimensioning

We here present the mathematical formulation used to model the multi-hop network dimensioning problem. The following notations are added:

$w(p)$: bandwidth consumption of p in number of wavelengths.

C: fixed cost of a core node switching unit.

We introduce the following variables, helped by Fig.8.6:

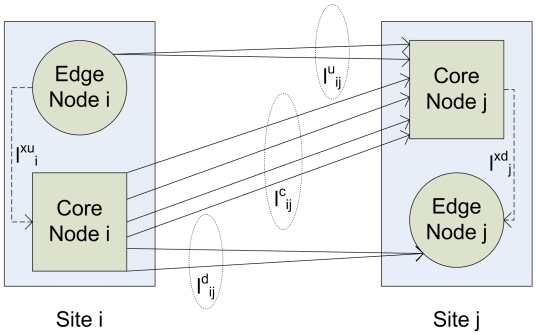


Figure 8.6: Counting locally added and dropped wavelength channels, and edge-to-edge, edge-to-core and core-to-core fiber links with l variables.

l^c_{ij}: number of fibers connecting the core node in site $i \in N$ to the core node in site $j \in N$;

l_{ij}^u (resp. l_{ij}^d): number of fibers connecting the edge node in site i to the core node in site j (resp. viceversa);

l_i^{xu} (resp. l_i^{xd}): number of fibers needed for local edge-to-core (resp. core-to-edge) interconnections;

a_{ij}^p: binary variable indicating if the lightpath of the connection request $p \in T$ is switched over the link (i, j);

x_p^u (resp. x_p^d): number of wavelength-channels for p added from the source edge node to the local core node, if any, bypassing a direct edge-to-core fiber (respectively, locally dropped for the core node, if any, to the destination edge node, bypassing a direct core-to-edge fiber);

The design dimensioning objective is thus:

$$\begin{aligned}\min G(y,l,a) &= \sum_{i\in N}\sum_{j\in N} \phi(W)\, F\, \Delta_{ij}\, (l_{ij}^c + l_{ij}^u + l_{ij}^d) \\ &+ \sum_{i\in N}\sum_{j\in N} WP(2l_{ij}^c + l_{ij}^u + l_{ij}^d) + \sum_{i\in N} WP(l_i^{xd} + l_i^{xu}) + \sum_{i\in N} C\, y_i \\ &+ \sum_{p\in T}\sum_{i\in N}\sum_{j\in N} \beta w(p) \Delta_{ij} a_{ij}^p \qquad (8.25)\end{aligned}$$

$$s.t. \quad \sum_{j\in N} a_{ij}^p - \sum_{j\in N} a_{ji}^p = \begin{cases} 1 & \text{if } i = s(q) \quad \forall i \in N \\ -1 & \text{if } i = d(q) \quad \forall p \in T \\ 0 & \text{otherwise} \end{cases} \qquad (8.26)$$

$$\sum_{p\in T}^{s(p)\neq i, t(p)\neq j} w(p) a_{ij}^p + w(p)\left[x_{p|s(p)=i}^u + x_{p|t(p)=j}^d\right] \leq W\, l_{ij}^c \quad \forall (i,j) \in N \times N \qquad (8.27)$$

$$\sum_{p\in T}^{s(p)=i} w(p) a_{ij}^p - w(p) x_{p|s(p)=i}^u \leq W\, l_{ij}^u, \quad \forall (i,j) \in N \times N \qquad (8.28)$$

$$\sum_{p\in T}^{t(p)=j} w(p) a_{ij}^p - w(p) x_{p|t(p)=j}^d \leq W\, l_{ij}^d, \quad \forall (i,j) \in N \times N \qquad (8.29)$$

$$\sum_{p\in T}^{t(p)=i} w(p) x_p^d \leq W l_i^{xd}, \quad \forall i \in N \qquad (8.30)$$

$$\sum_{p\in T}^{s(p)=i} w(p) x_p^u \leq W l_i^{xu}, \quad \forall i \in N \qquad (8.31)$$

$$l_j^{xu} + \sum_{i\in N} (l_{ij}^c + l_{ij}^u) \leq |N| y_j, \quad \forall j \in N \qquad (8.32)$$

$$l_i^{xd} + \sum_{j\in N} (l_{ij}^c + l_{ij}^d) \leq |N| y_i, \quad \forall i \in N \qquad (8.33)$$

$$a_{ij}^{q}, x_{q}^{d}, x_{q}^{u} \in \{0,1\}; \quad y_i, l_{ij}^{c}, l_{ij}^{d}, l_{ij}^{u}, l_{i}^{xu}, l_{i}^{xd} \in Z_+ \tag{8.34}$$

(8.25) expresses the minimization of the total network cost due to propagation delays, fibers and switching ports. The port cost for core-to-core fibers is double than that for edge-to-core and core-to-edge fibers: a core-to-core fiber link would be preferred to a new edge-to-core or core-to-edge link to route a demand if it disposes of enough capacity and if the cost of W additional ports is minor than the cost of a edge-to-core or core-to-edge fiber. Moreover, we add the port cost for local interconnection. (8.26) is the traffic conservation constraint, imposing that the flow leaving node i is balanced by the entering flow, except for the source (destination) node; (8.27) dimension the core-to-core fiber links; (8.28) and (8.29) dimension the edge-to-core and the core-to-edge fibres; (8.30) and (8.31) dimension the local edge-to-core and core-to-edge fiber interconnections; (8.32) and (8.33) enforce the maximum size of the core nodes; (8.34) imposes the binary constraint for a and x variables, and the integer constraint for l and y variables.

Numerical results

In Sect. 9.5 we present and discuss the results obtained from numerical simulations.

Summarizing, we showed that the amount of core fibers in multi-hop core networks decreases whether a larger impact is given to a propagation delay virtual cost in the design dimensioning objective. This suggests that as the propagation delay cost increases, the multi-hop structure tends to assume a composite-star configuration, which has no core fibers indeed. Moreover, the quasi-regular composite-star architecture presented at most 20% higher network cost because of its more stringent switching constraints. The simplification in traffic engineering operations that the composite-star core architecture offers may convince the decision-maker that such a small difference in network cost is not an issue.

8.8 Summary

Recent works on optical transport networks architectures have investigated a novel high capacity physical architecture with a composite-star topology structure. Such an architecture was originally proposed in [117] and nicknamed "the Petaweb"since it might serve a global traffic volume in the order of the petabit per second (10^{15} bit/s). Its structure provides fully meshed connectivity with direct optical paths between electronic edge nodes. It is composed of several core nodes that commute the traffic exchanged by the edge nodes without wavelength conversion. A particular feature is that each core node is connected to all edge nodes. Another peculiar feature is that the core nodes are not connected among themselves, making it a complete architectural breakthrough. This architecture might require a higher fiber distance value; however, the cost savings in operational engineering can be significant.

We tackled for the first time the design dimensioning problem of the Petaweb architecture. We demonstrated that it is NP-hard, by reduction to a facility location problem, and proposed optimal and suboptimal resolution methods. We analyzed the physical performance of this architecture, underlining that a very low resource utilization may characterize this architecture in the case of traffic matrix with a few peaks and a lot of low-rate connections. We showed that by removing unused equipments (fibers, ports) from the optimal regular architecture, the utilization can pass from 20% to 50%, roughly. As a consequence, the network cost gets almost halved. The resulting "quasi-regular composite-star"structure can still guarantee to reach a regular structure by simple addition of elements, at the expense of some wavelength conversions at the core nodes. However, the quasi-regular structure appeared not to be reliable for small networks because some edge nodes may become isolated. To overcome these aspects we identified as the suitable protection strategy the dedication path protection. This allows designing a reliable core network at, however, an almost double cost. Still, the quasi-regular structure obtained in this way is determined heuristically. We further propose a method to directly optimize a quasi-regular composite-star architecture, instead of determining it by downgrading an optimal regular one. The results show that the quasi-regular network cost can be further reduced by 30%.

In order to assess the convenience in adopting a quasi-regular Petaweb solution, we finally compared it with multi-hop mesh networks. We found that as additive lightpath costs are minimized (such as the propagation delay cost), the multi-hop structure tends to assume a composite-star configuration, which has no core fibers indeed. Moreover, the quasi-regular composite-star architecture presented quite higher network cost because of its more stringent switching constraints.

Chapter 9

Petaweb Design: Numerical Results

9.1 Regular Petaweb

The proposed heuristic was tested using two networks, respectively composed of 10 and 34 edge nodes. The locations of the edge nodes are specific cities of the United States. Two traffic matrices were used:

- Matrix A, which is a sparse matrix that was provided by the industry (Nortel Networks),
- Matrix B, that is calculated using a gravity model based on urban populations and distances between cities. The urban populations were found in [201]. Note that this matrix does not include any zeros, except on its diagonal.

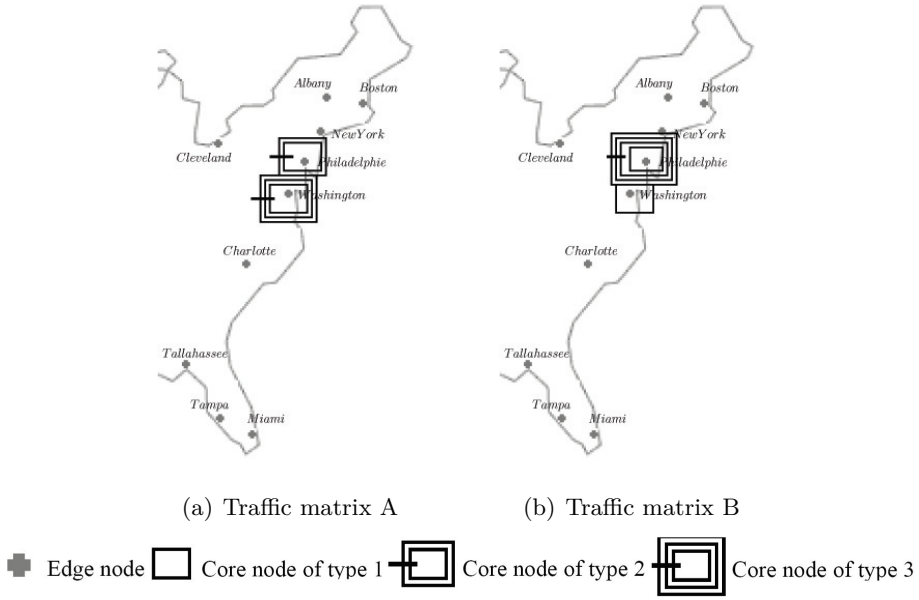
For the 10 and the 34 node networks, the total amount of traffic requested for all origin destinations of matrix A were respectively 2.1612 Tbit/s and 10.692 Tbit/s. The values, for matrix B were 2.167 Tbit/s and 10.050 Tbit/s.

The distance matrix between edge nodes was calculated as follows. To work with realistic distances, geographical coordinates were first found in an American national atlas [202] and a formula to assess the distance between two points on a sphere [199] was used. The calculated distances were later compared and validated with a few air distances estimated at the University of Minnesota [200].

The following default values were used: $W = 16$; $C_{ch} = 10$ Gb/s; $v = 3$ (number of types of core nodes); $e = 3$ (maximal number of core nodes of one type at one site), except for the 34-node network with traffic matrix B when $e = 4$ for core nodes of type 3; $s_1 = 1$, $s_2 = 2$, $s_3 = 4$; $\gamma = 0.95$; $P/F = 150$, $f_1/F = 20$, $f_2/F = 50$, $f_3/F = 100$, $\beta/F = 0.1$ (the unitary costs are furnished normalized to F); $C_j = 1000$ Gb/s for 10-node networks, $C_j = 2000$ Gb/s for the 34-node network with traffic matrix A, $C_j = 2800$ Gb/s for the 34-node network with traffic matrix B. Then, $\phi(W) = W$ as discrete function used to scale the reference fiber cost F. F is assumed to be the cost of a single-wavelength fibre. When F is multiplied by $\phi(W) = W$ the unitary cost of a fiber is considered proportional to the number of wavelengths.

Results with default parameters

We solved the problem using the default parameters and using two resolution approaches: CPLEX and the proposed heuristic. The results are presented in Table 9.1 for the 10-node network and in Table 9.2 for the 34-node network. The gap in the last line is the



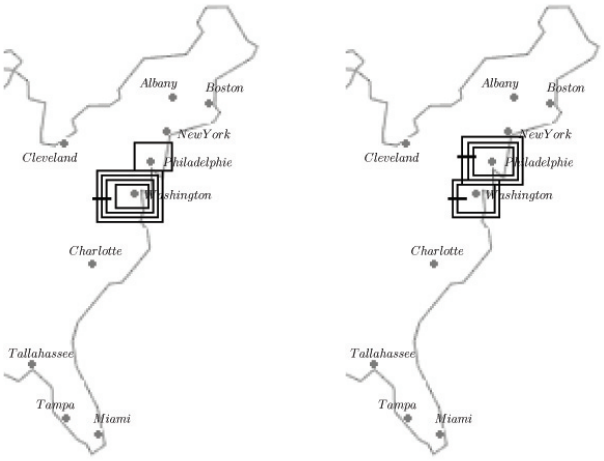
(a) Traffic matrix A (b) Traffic matrix B

Figure 9.1: 10-node networks with default parameters (CPLEX)

discrepancy in percentage between the total network cost found by the heuristic and the total network cost found by CPLEX for the mathematical model. All the costs have been normalized to F. The actual solutions obtained for all treated instances are presented in Figs. 9.1, 9.2, 9.3 and 9.4.

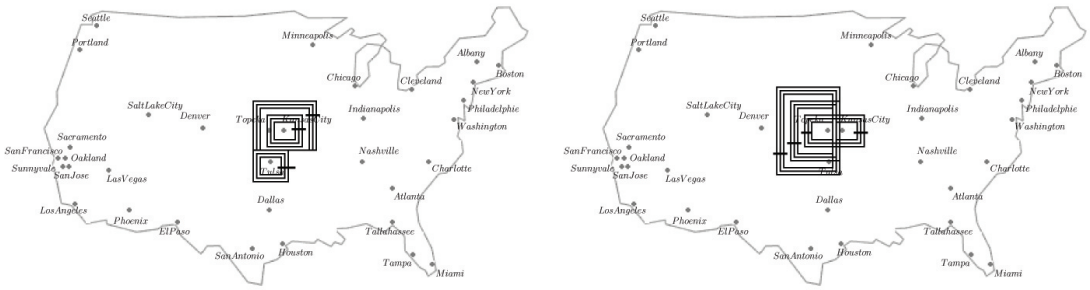
In terms of computational complexity we can see that these are extremely hard problems. In fact, for some instances, it took CPLEX up to 18 days to reach the best solution. These results underline the importance of creating an efficient heuristic approach. From the optimization standpoint it can be seen that the heuristic presents very good results, showing an optimum gap well below 1% in most of the instances and of 5.5% in the case of the 34 node example with a dense matrix. On the other hand, the resolution time is drastically reduced with the use of the heuristic going from days or hours to just seconds.

About the objective costs of the obtained solutions, the vast majority of the cost is allocated, as expected, to the fiber term, which amounts for roughly 80% of all the considered costs, for both the 10-node network and the 34-node network. There is, however, a slight difference between the cases with the A and the B matrices run for the 10-node network. In fact, whereas for the A matrix the percentage of the fiber costs are around 77%, for the B matrix it goes up to 83%. We see also that that difference is, in the case of the matrix A, being absorbed by the delay cost. So, for this small network, the density of the traffic matrix seems to have an impact on how the costs are allocated. The other interesting observation is that when we compare the 34-node network cases with the smaller instances, we see that the cost distribution is not affected by the traffic matrix. On the other hand we see that the percentage of the cost that goes to the core nodes is lowered from 12% to 5%: with 34-node networks we have less core nodes, but of higher types, and, thus, the switching planes are less expensive.



(a) Traffic matrix A (b) Traffic matrix B

Figure 9.2: 10-node networks with default parameters (heuristic)



(a) Traffic matrix A (b) Traffic matrix B

Figure 9.3: 34-node networks with default parameters (CPLEX)

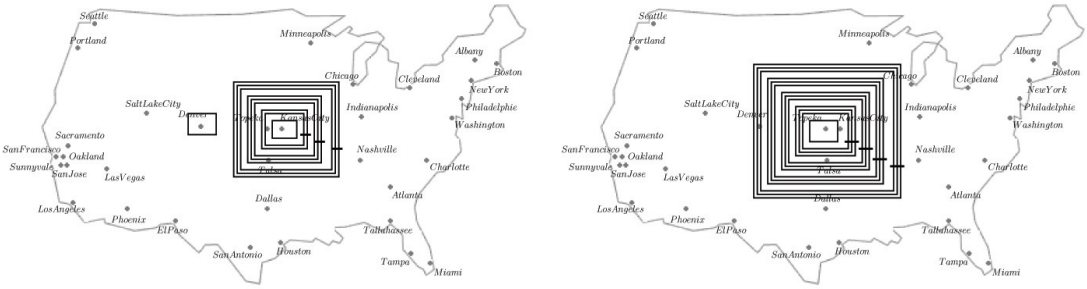
9.1.1 Sensitivity studies

Influence of delay versus fiber costs

To see the influence of the delay cost versus the fiber cost, we ran a test with traffic matrix A. In the first case the delay costs were omitted whereas in the second case the fiber cost was set to zero. It can be appreciated from Fig. 9.5 that the influence of the terms in the solution is quite different. When the delay cost is omitted, all the switches are set at the centre of mass of the map. On the other hand, when the fiber cost is set to

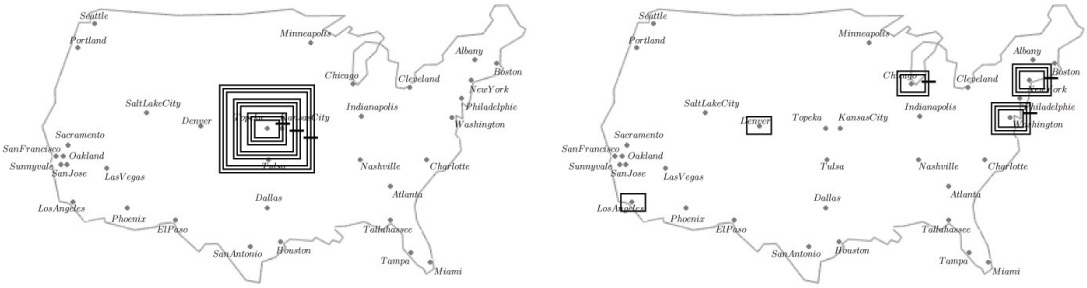
Network	Traffic A heuristic	Traffic A CPLEX	Traffic B heuristic	Traffic B CPLEX
Objective	2289564	2280980	2153868	2152920
fiber	77.1%	77.8%	83.3%	83.8%
CN	11.4%	11.2%	11.9%	12.1%
delay	11.5%	11%	4.8%	4.1%
Time (s)	6	23650	11	232
Optimum gap	0.38%	-	0.04%	-

Table 9.1: Results obtained for the 10-node networks



(a) Traffic matrix A (b) Traffic matrix B

Figure 9.4: 34-node networks with default parameters (heuristic)



(a) without delay costs (b) without fiber costs

Figure 9.5: 34-node network with default parameters, matrix A (heuristic)

zero but we keep a term to account for the delay, all the switches are spread, with larger switches on the east part of the country where the higher origin-destination demand is concentrated.

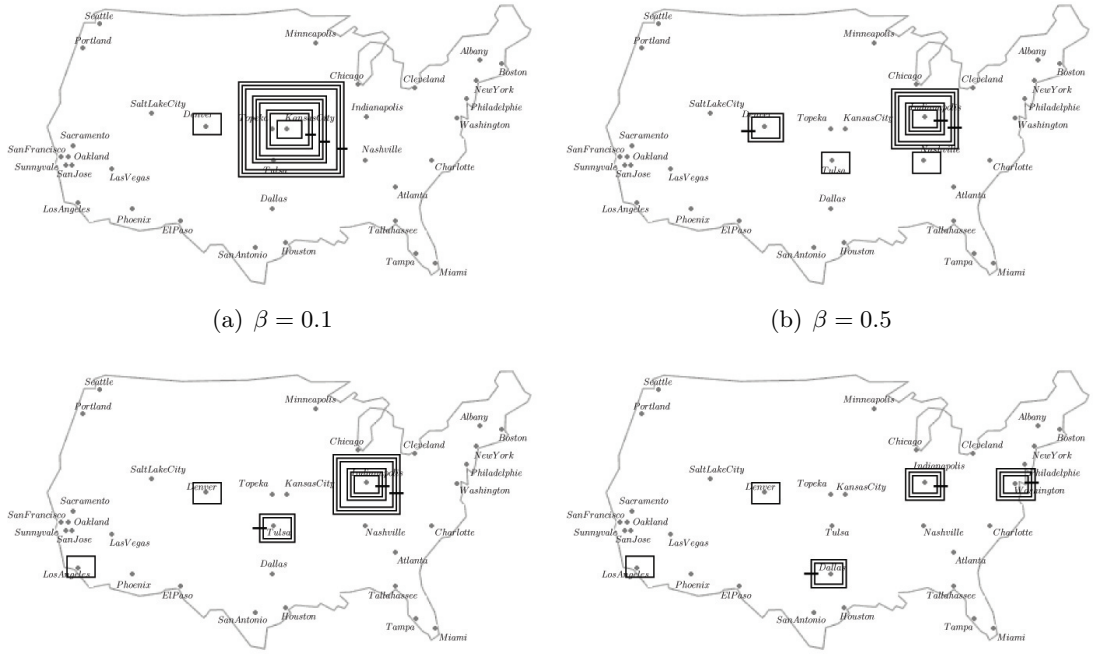
Propagation delay variation

Given the important influence of the delay term in the objective, some sensitivity tests were made for the 34-node network with respect to the propagation delay cost. The parameter β was progressively increased. The results for the traffic matrix A are presented in Table 9.3 and in Fig. 9.6.

The importance of the term can be assessed from the results. Clearly, when β increases, the active CNs are increasingly more spread in the country. Thus, the added delay costs can be seen as a 'natural' survivability term that prevents the location of all the resources in the same place. In Table 9.1.1 we can see that, as expected, the total

Network	Traffic A heuristic	Traffic A CPLEX	Traffic B heuristic	Traffic B CPLEX
Objective	31940857	31837547	44757016	42406000
fiber	82%	81.7%	82.2%	81.6%
CN	5.5%	5.3%	5.4%	5.3%
delay	12.6%	13%	12.5%	13.1%
Time (s)	217	579998	322	1614383
Optimum gap	0.32%	-	5.5%	-

Table 9.2: Results obtained for the 34-node networks



(a) $\beta = 0.1$ (b) $\beta = 0.5$

(c) $\beta = 1$ (d) $\beta = 1.5$

Figure 9.6: 34-node network with several delay weights, matrix A (heuristic)

β value	0.1	0.5	1	1.5
Objective	31940857	46864904	61244206	74650622
fibre	82%	60.3%	46.6%	41.7%
CN	5.5%	3.7%	2.9%	2.4%
delay	12.5%	36.0%	50.5%	55.9%

Table 9.3: Influence of the propagation delay cost for 34A (heuristic)

cost of the network increases when β increases. Also expected is the proportion of the delay cost in the total cost. We can notice that the percentage of the core node cost and of the fiber cost in the total cost decreases. In fact, the number and the type of the active core nodes are constant when β increases, which lowers the percentage of the cost of the core nodes. There is also the clear trade-off between the fiber and the delay cost that is underlined by these tests. The more the delay cost increases, the lower is the percentage of the fiber costs.

In Table 9.4 we report, for the 10- and 34-node cases, another type of test to assess how the variation of β influences the average lightpath length, and thus the propagation delay. The average is taken over all the origin-destination pairs of edge nodes in the examples. In the table we indicated as pedix of the average length the standard deviation to provide a measure of how much the average length represents the connections length. It can be seen from the table that when the weight of the propagation delay cost is increased, the length of the transmission path between an origin and a destination node is reduced.

With these results in mind, let us assume that it were possible to establish a direct link (0-hop) between every pair of edge nodes leading to a full-mesh network with a link lengths equivalent to the air distances between cities. Such a topology would be the fastest one from the standpoint of the connection speed, that is, it would be the topology

Case	$\beta = 0.1$	$\beta = 0.5$	$\beta = 1$	$\beta = 1.5$
10A	$1682_{\sigma=1884}$	$1635_{\sigma=1868}$	$1276_{\sigma=1453}$	$1266_{\sigma=1887}$
10B	$1802_{\sigma=1284}$	$1794_{\sigma=1277}$	$1794_{\sigma=1276}$	$1794_{\sigma=1276}$
34A	$3396_{\sigma=4024}$	$2862_{\sigma=3391}$	$2549_{\sigma=3132}$	$2411_{\sigma=3144}$
34B	$3289_{\sigma=1670}$	$3111_{\sigma=1652}$	$3184_{\sigma=1865}$	$3060_{\sigma=1691}$

Table 9.4: Average length of an origin-destination connection [km] as function of β. The pedix is the standard deviation.

	$\beta = 0.1$	$\beta = 0.5$	$\beta = 1$	$\beta = 1.5$
10A	$1309_{\sigma=1652}$	$1244_{\sigma=1634}$	$761_{\sigma=1201}$	$735_{\sigma=1191}$
10B	$416_{\sigma=1691}$	$415_{\sigma=1694}$	$415_{\sigma=1694}$	$415_{\sigma=1694}$
34A	$3279_{\sigma=3911}$	$2823_{\sigma=3346}$	$2372_{\sigma=2971}$	$2262_{\sigma=2882}$
34B	$4875_{\sigma=2521}$	$3849_{\sigma=1923}$	$2804_{\sigma=1854}$	$1038_{\sigma=2495}$

Table 9.5: Weighted average length of an origin-destination connection [km] as function of β. The pedix is the standard deviation.

that would provide the lowest propagation delay. Now we want to assess how far is the Petaweb design from that full-mesh topology. For this, we evaluate the average length of a connection for each of the Petaweb cases considered and define the *overhead* as the percentage length increment with respect to the corresponding full-meshed case length value. In Fig. 9.1.1 we report such an overhead as a function of β.

It can be observed from the figure that with the default value for β the overhead is under 100% for the 10-node cases, and close to 200% and 500% for the 34 node cases. Thus, we can see that the average propagation delay overhead increases with the network geographical extension and dimension. When β is incremented to 0.5, 1 and 1.5 a very significant overhead reduction is experienced in all the cases and it tends towards a 0% increase asymptote. For $\beta = 1$ the overhead of all but one case has been at least halved. A particular exception, however, seems to be the 10 B case (10 nodes and a full matrix demand). The phenomena could be explained by the fact that the core nodes for this case maintain the same location for $\beta \in \{0.5, 1, 1.5\}$ and, thus, the lightpaths follow the same routes.

We also considered the weighted average lightpath length and overhead where the weights are proportional to the origin-destination demand. The results are displayed in Table 9.5. It can be seen that the weighted overhead decreases for all the instances, but that there is a more marked tendency for the 34B case, which is the larger network with a dense traffic matrix. This is precisely the case where the influence of the origin-destination demand is the greatest, therefore it is not surprising that it is the one for which the weighted delay term has the more impact.

These results indicate that the danger of a bigger propagation delay supposed in [117] can be controlled during the planning of the Petaweb structure.

Variation of core node costs

The fixed unitary node cost f_r and the unitary cost per port P was first changed in the range $[-60\%, +60\%]$ of their default values. The tests produced no significant results: the route assignment for connection did not change significantly (the propagation delay was almost constant), and the number of switching planes and their location remained almost the same (i.e., the fiber cost was almost constant). The conclusion is that, within such a range of variation of the unitary costs the solution is not affected. We then

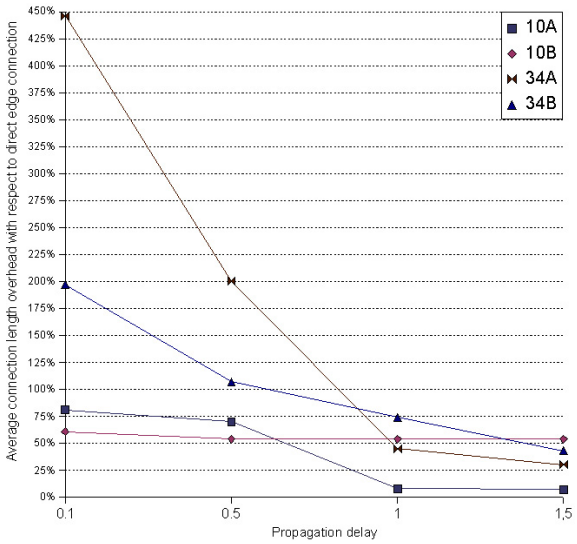


Figure 9.7: Length average overhead as function of β w.r.t. a full mesh network.

increased by 100%, 200% and 300 % the cost of the cores to see if such a major increase would lead to significant variations. The results can be seen in Table 9.13 where the total costs are provided, followed by the cost distribution in percentage. We can see how the costs distribution changes for all the considered cases. In all the instances, an increment in the core node is reflected in an increase of the core node cost percentage and a decrease of the fiber cost. However, it can be assessed that the percentage of the delay cost does not vary a lot. This means that the absolute value of the propagation delay increases when the node cost increases.

Therefore, we can conclude that whereas reasonable changes in the core node cost do not have an impact in the design, an important increment leads toward a design with higher propagation delays.

Sensitivity to fiber costs

Concerning fiber costs, the default data were obtained considering $\phi(W) = W$, i.e., assuming that the global fiber cost is proportional to the number of wavelengths. We want now assess the influence of this term in the final solution. We varied $\phi(W)$, considering an exponential dependence $\phi(W) = W\gamma^{\sqrt{W}}$, a logarithmic dependence $\phi(W) = \sqrt{W}ln(W)$, and a radical dependence $\phi(W) = \sqrt{W}$. With these types of $\phi(W)$ functions, the incremental cost from $W = 20$ to 40 is bigger than the incremental cost from $W = 80$ to 100, for example. The results are displayed in Table 9.6. We reported both the absolute values and the percentage values for the detailed costs, and, for the objective, we reported the percentage decrease with respect to the default case in bold. Since $W = 16$, the actual values of ϕ were 16, 13, 11 and 4. Thus, each of the non-default cases yield a fiber cost reduction of 18%, 29% and the 75%, respectively.

It is worth noting that, for 10-node networks, the total cost reduction is close to the value of the fiber cost reduction that the specific $\phi(W)$ produces, thus implying a direct impact of the fiber cost on the total cost. Interestingly, this is not the case for the 34-node examples, in particular for the 34B case that presents reduction of the order of 2.8%, 8.3% and 60% in the total cost. The other interesting observation is that for all the three non-default experiments the core node costs increase a little bit when compared

Cost	$\phi(W) = W$	$\phi(W) = W\gamma^{\sqrt{W}}$	$\phi(W) = \sqrt{W}ln(W)$	$\phi(W) = \sqrt{W}$
10-node network, matrix A				
Total	**2289564**	1983146 (-13.4%)	1774067 (-22.6%)	968542 (-67.7%)
CN	**261011 (11.4%)**	278540 (14.05%)	278540 (15.70%)	278540 (28.76%)
fiber	**1765254 (77.1%)**	1453612 (73.30%)	1257717 (70.89%)	451205 (46.59%)
delay	**263299 (11.5%)**	250994 (12.65%)	237810 (13.51%)	238797 (24.66%)
10-node network, matrix B				
Total	**2153868**	1863349 (-14%)	1640993 (-23.8%)	828107 (-61.5%)
CN	**256310 (11.4%)**	273750 (14.69%)	273750 (16.7%)	273750 (33.1%)
fiber	**17694172 (83.65%)**	1493923 (80.17%)	1271497 (77.5%)	458610 (55.36%)
delay	**103385 (4.61%)**	95675 (5.14%)	95746 (5.8%)	95746 (11.54%)
34-node network, matrix A				
Total	**31940857**	31289432 (-2%)	27509399 (-13.8%)	13378622 (-41.8%)
CN	**1756747 (5.5%)**	2203530 (7.1%)	2203530 (8.01%)	2203530 (16.5%)
fiber	**26191502 (82%)**	25518282 (81.5%)	21659389 (78.3%)	8073687 (60.3%)
delay	**4024547 (12.5%)**	3567619 (11.4%)	3646479 (13.26%)	3101404 (23.2%)
34-node network, matrix B				
Total	**44757016**	43480527 (-2.8%)	3743923 (-8.3%)	17592830(-60%)
CN	**2416878 (5.4%)**	2954390 (6.8%)	2893924 (7.7%)	2807480 (15.9%)
fiber	**35790267 (82.1%)**	35572707 (81.8%)	29554327 (78.9%)	10709106 (60.9%)
delay	**5594627 (12.5%)**	4953429 (11.4%)	4989672 (13.4%)	4076243 (23.2%)

Table 9.6: Results obtained for different fiber costs (heuristic). $W = 16$. In bold the results for the default $\phi(W)$.

to the default, but then stay almost constant. Also, when we evaluate the non-default cases with the default, we see that there is an initial decrease on the delay cost and that when ϕ is lowered, it decreases even more or stays roughly the same. The delay cost decrease and the core node increase can be explained with the fact that the core nodes are driven to be located better near edge nodes because of less expensive fibers.

As a conclusion, there seems to be a clear impact on the fiber cost function and a net difference between the case where the fiber costs are proportional to the number of wavelength and the case they are not.

9.1.2 Scalability of the heuristic approach

The heuristic has given very good results for 10- and 34-nodes networks. In this subsection, we increase the size of the networks to be treated to test its scalability.

Some tests were made adding at each step some cities of the United States according to their decreasing population importance. For each test, a full traffic matrix was elaborated using the gravity model (matrix B). The sum of the total exchanged traffic was the same for all cases. The values of the parameters were the default values. The parameter representing the propagation delay was increased to $\beta = 1$. The maximum core nodes of one type that could be opened at one site was 4 and the maximum edge node capacity was $C_j = 3000$ Gb/s.

The results given by the heuristic for 40 to 136 edge nodes are given in Table 9.7. The total network cost increases when the network size is growing. The proportions of the different costs in the total cost are kept constant. The fiber cost predominates with a percentage of 60% to 70%, the delay cost comes next with a percentage of 25% to 35%, and the core node cost is the lowest with a percentage of 4% to 5.5%. Three cases are illustrated in Fig. 9.8.

Network:	40B	50B	60B	70B	80B
Total cost	$6.7 \cdot 10^7$	$6.6 \cdot 10^7$	$7.1 \cdot 10^7$	$7.2 \cdot 10^7$	$8.3 \cdot 10^7$
fiber	60.0%	59.3%	62.8%	64.7%	70.5%
CN	3.9%	4.6%	4.6%	5.0%	5.5%
delay	36.1%	36.2%	32.7%	30.3%	24.0%
Iterations	22	19	26	33	23
Time (s)	981	2109	6226	13497	13016
Network	100B	110B	120B	130B	136B
Total cost	$8.9 \cdot 10^7$	$9 \cdot 10^7$	$1 \cdot 10^8$	$1 \cdot 10^8$	$1 \cdot 10^8$
fiber	65.4%	66.4%	67.4%	69.3%	70.6%
CN	5.3%	5.2%	5.4%	5.4%	5.8%
delay	29.3%	28.4%	27.2%	25.3%	23.6%
Iterations	28	27	27	30	38
Time (s)	90369	71619	555416	155500	505115

Table 9.7: Results for scalable networks with $\beta = 1$ (heuristic)

Network	10A	10B	34A	34B
Objective	2281804	2155353	31995440	42596082
fiber	77.8%	83.27%	82.46%	81.27%
CN	11.22%	11.88%	5.44%	5.26%
delay	10.98%	4.86%	12.1%	13.47%
Time (s)	169	100	43452	1685
μ_R	17.91%	15.15%	16.83%	12.39%

Table 9.8: RFA solution

9.2 TDM/WDM and quasi-regular structure

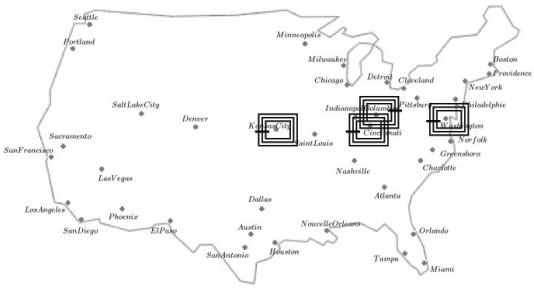
In this section, we show the results obtained implementing the resolution algorithms for the Petaweb design problem with TDM/WDM (see Sect. 8.4).

Given $C_{ch} = 10$ Gb/s and $W = 16$, $Z_2 = 10$ Gb/s and $Z_3 = 160$ Gb/s. Having to choose a value for Z_1, we verified experimentally that an appropriate value guaranteeing an acceptable rounding up value for our traffic models is Z_1=0.625 Gb/s. Furthermore, being Z_1 a fraction multiple of 2 of the capacity of a lambda-channel of 10 Gb/s, the OTU rates 2.5 Gb/s, 10 Gb/s and 40 Gb/s [185] are reachable through a composition of TLP-1 and TLP-2. Other parameter values not yet explicated are: $L_1 = L_2 = 12$, $L_3 = 20$, and $E_1 = E_2 = 1$ to avoid the generation of equivalent solutions (remembering that $K_r < K_{r-1}$), and $E_3 = 3$.

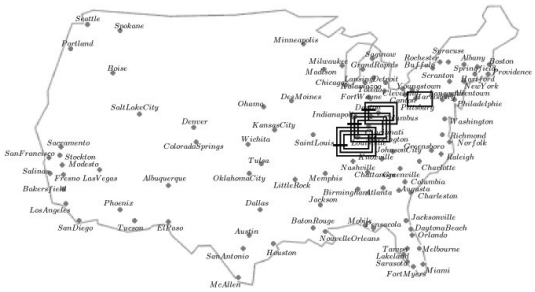
9.2.1 RFA results

Table 9.2 reports the final results for the 10A, 10B, 34A and 34B cases, Table 9.2.1 shows the changes obtained extracting a quasi-regular topology from the solution; the objectives are normalized to F. Figs. 9.10 and 9.11 illustrate the optimized regular and quasi-regular topologies for 10-node architectures. Fig. 9.12 displays the CNs geographical distribution for the optimized 34-node networks. The solutions for 10-node networks have an affordable execution time. For 34-node networks we contain the execution time setting the CPLEX upper cut-off value exploiting the objective values obtained through the heuristic for the regular Petaweb opportunely adapted to TDM/WDM.

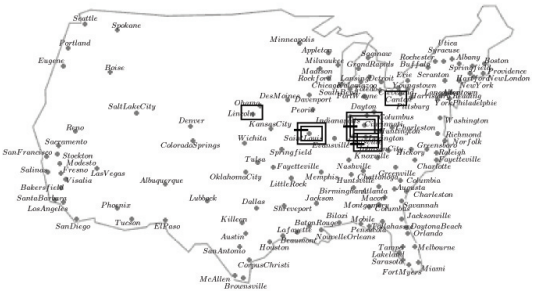
The results with the A matrixes provide better values of the network utilization coefficient (μ_R). The explanation is that the B traffic matrixes are dense and present



(a) 40 edge nodes



(b) 100 edge nodes



(c) 136 edge nodes

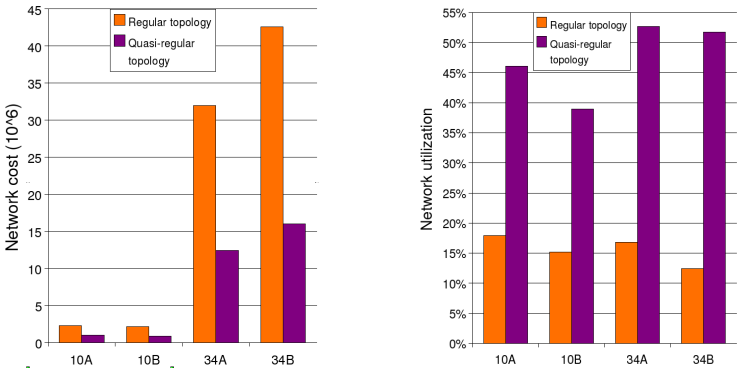
Figure 9.8: Scalable networks with $\beta = 1$ (heuristic). B traffic matrix.

Case	10A	10B	34A	34B
Total cost	982492	840006	12406718	15976542
fiber	63.13%	73.65%	63.81%	60.41%
CN	11.36%	13.88%	4.97%	3.67%
delay	25.51%	12.46%	31.21%	35.91%
μ_R	48.36%	38.97%	52.69%	51.73%

Table 9.9: RFA solutions changes using a quasi-regular topology

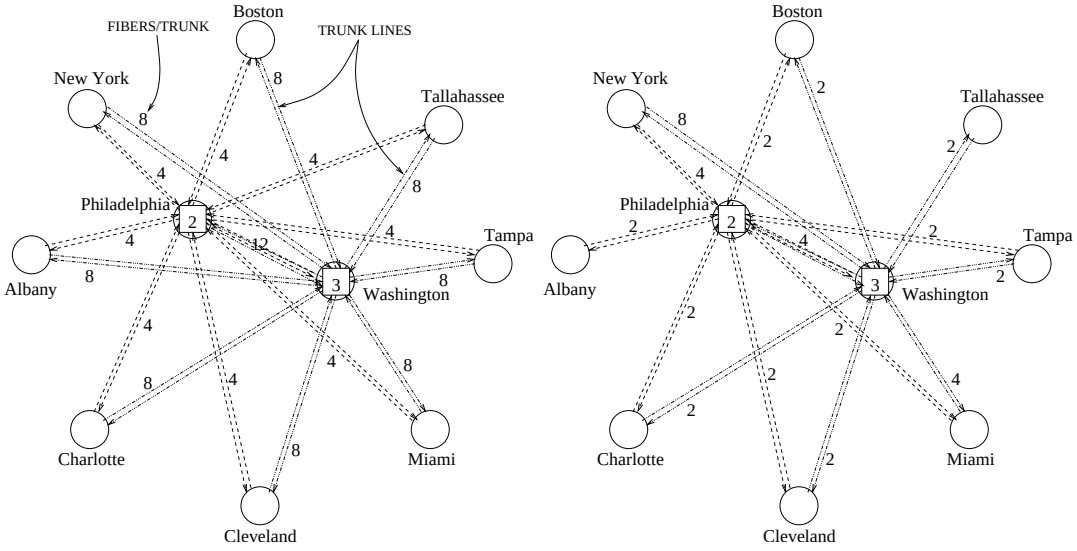
many CRs with low traffic demands; thus, the links used by these CRs are under-used. What are the changes if we can use the quasi-regular topology? As it can be assessed by Table 9.2.1, the network cost is dramatically reduced, more than 50% (Fig. 9.9a), and the network utilisation is doubled (Fig. 9.9b). This is due to the fact that the regular topology demands the allocation of too many unused fibers. Blouin et al. [153] estimated the quantity of fiber requested by a regular Petaweb as roughly the 17% more than a multi-hop architecture; now, we reduced considerably the quantity of fiber to install; for example, the km of installed fibers with the quasi-regular topology reduced of 65% for 10A, and of 72% for 34B.

In Fig. 9.10 and Fig. 9.11 for every connection between the EN and the switching site we indicate upon it the number of fibers to install. We now analyze in detail the regular topology for the 10A model (Fig. 9.10a): we have two enabled core nodes, a CN-2 in Philadelphia and a CN-3 in Washington. The optical trunk lines between every switching site and every EN are composed of two optical links in opposite directions. An optical link is composed of s_r fibers; thus, the ones connected to the CN-2 in Philadelphia are composed of 2 fibers, and the ones connected to the CN-3 in Washington are composed of 4 fibers. In the quasi-regular topology (Fig. 9.10b) the unused fibers were disabled and not considered in the solution. Entire trunk lines have been disabled because their optical



(a) Network cost (b) Network utilization

Figure 9.9: Comparison between quasi-regular and regular structures



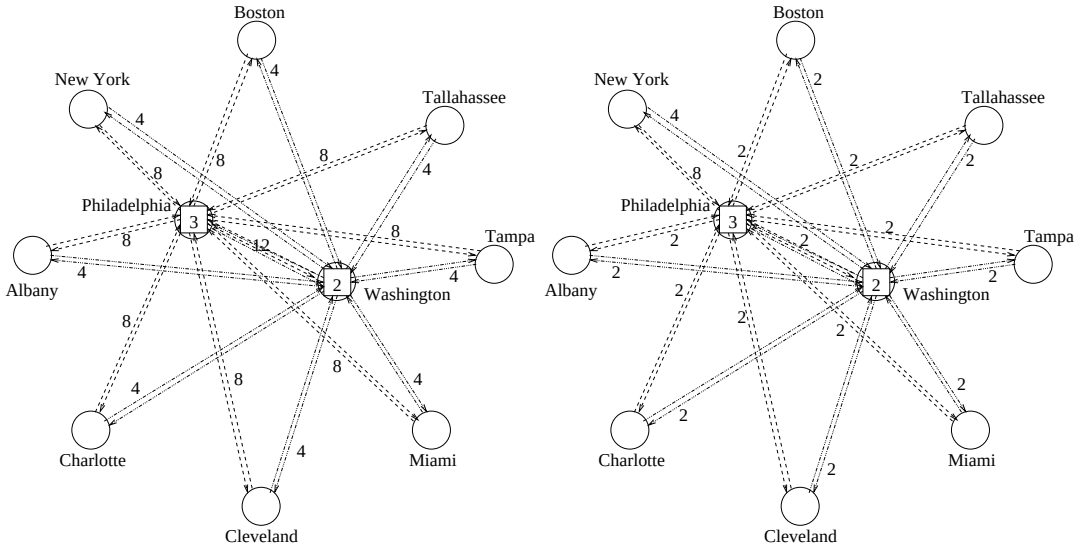
(a) 10A regular (b) 10A quasi-regular

Figure 9.10: Route and Fiber Allocation solution for 10A case

links were totally unused: the Washington-Albany one and the Philadelphia-Tallahassee one. Furthermore, the number of fibers per trunk line has generally decreased: the optical links connected to the CN-2 and the CN-3 are now composed of only one and two fibers (instead of 2 and 4), except the one with Miami, and the one with New York; the EN in New York asks more than 800 Gb/s to be switched, more than any other EN, and forces the opening of high type CNs.

For the model 10B (Fig. 9.11), on the contrary, one can observe that there is not any optical trunk line disabled. Why? Because the traffic matrix of 10B is dense, every EN is fully connected with the others and there are more TLPs to be switched. One has to pay attention to the optical trunk line Philadelphia-Washington that is not a direct connection between the CNs, but a trunk composed of optical links connecting the two ENs with the two CNs.

Therefore the cost reduction concerns unused fibers cost and disabled ports cost; consequently the cost allocation changes. In Fig. 9.13 we compare the cost allocation assuming the two topologies for 10-node cases. Adopting a quasi-regular topology, the weight of fibers cost decreases more than 10 percentage points and, thus, the weight



(a) 10B regular (b) 10B quasi-regular

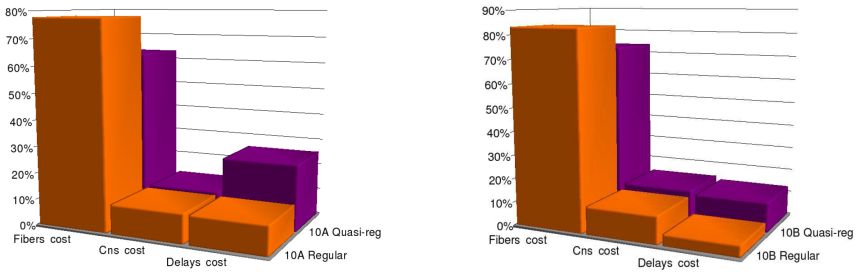
Figure 9.11: Route and fiber Allocation solution for 10B case

(a) 34A (b) 34B

Figure 9.12: Core nodes geographical distribution

of delays and CN costs increases; the delays cost assumes a weight of more than 10 percentage points than in the case with the regular topology where the fibers cost is over-estimated; the CN cost increases even if its absolute value decreases because the cost reduction due to fibers is more important than that due to ports. In the case of 34-node networks it is evident that the CN cost becomes unimportant at the expense of the delay propagation cost. And this is even more evident with a quasi-regular topology. The presence of high order CNs allows to assign the TLPs to a few trunk lines and to decrease to total CNs number.

As a consequence of the deletion of unused fibers and links, with the quasi-regular topology the network utilization increases very significantly and this indicates that, on average, the network fibers are used almost at the 50% of their transport capacity. This is a good result, considering that, thanks to the time-sharing we have better exploited the lambda-channels. Nevertheless, one can observe a problem in the solutions with a quasi-regular topology: there are ENs that are connected through only one trunk line to the transport network (Fig. 9.10). The same problem may appear in an optimized regular topology where all the CNs are installed in the same site.



(a) 10A (b) 10B

Figure 9.13: Cost allocation for regular and quasi-regular topologies

9.2.2 WTA results

The WTA task is solved with linear complexity. In a 10-node network with two enabled CNs we have a total of 40 optical links. We report a part of the 10A WTA solution concerning only two optical links.

We have to assign time-slots and wavelengths of every optical link to the TLPs whose traffic has been allocated on that link. Here we consider the optical link from Tallahassee to Washington and the one from Washington to Tampa, following the route of the outgoing TLPs of the EN in Tallahassee. Tallahassee and Tampa are connected to the CN-3 in Washington through only one optical link per direction. The connection requests are $CR_{9,8}$=1.6 Gb/s and $CR_{9,10}$=0.2 Gb/s; 3 TLP-1s serve $CR_{9,8}$, and 1 TLP-1 serves $CR_{9,10}$. The RFA solution indicates that the optical link Tallahassee-Washington has to transport only those TLP-1s (then the TLP-1 of $CR_{9,10}$ is directly dropped in Washington), while the optical link Washington-Tampa serve even TLPs of others CRs. Fig. 9.14 illustrates the assignment of time-slots and wavelengths.

The TLP-1s of $CR_{9,8}$ and $CR_{9,10}$ have been assigned to the first four time-slots on λ_1 of fiber 1 on the optical link Tallahassee-Washington. The other 3 fibers, 15 wavelengths and 12 time-slots of that optical link remain unused. Then the TLP-1s of $CR_{9,8}$ are routed on the optical link between Washington and the destination EN in Tampa, where they occupy the third, the forth and the fifth time-slots on λ_2 of the fiber 1. On this optical link three fibers and one wavelength remain unused. The other wavelengths and time-slots are assigned to the TLPs of other CRs terminated in Tampa.

What about the unused fibers? If one considers a regular topology the Tallahassee-Washington trunk line would have an utilization of 0.39% and the Washington-Tampa one of 23.44%; otherwise with a quasi-regular topology one would have an utilization of, respectively, 1.56% and 93.75%.

Fig. 9.14 illustrates how ts-lightpaths of different classes are assigned to different medium. While the outgoing TLPs of Tallahassee are of the lower class, and are assigned to a time-slot each, others TLPs of class 2 are routed on the used fiber between Washington and Tampa. The benefits of using many hierarchically ordered classes of service is evident, because we can use efficiently the link capacity thanks to the TDM over WDM; in the link Washington-Tampa we could set up 45 independently manageable ts-lightpaths on a fiber with only 16 wavelengths.

9.3 Dedicated path protection

In this section we discuss the results obtained implementing the resolution algorithms for the Petaweb design problem with TDM/WDM and DPP strategy (see Sect. 8.5).

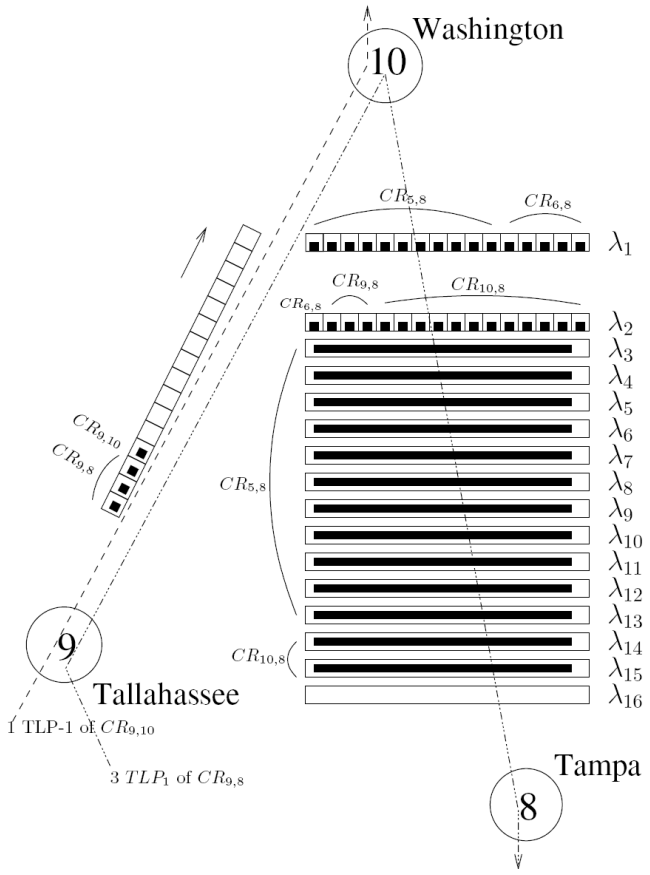


Figure 9.14: WTA for two optical links of the 10A solution

Because of the larger amount of traffic requested by working and protection lightpaths, we increased E_3 to 4, and the ENs capacity constraints C_j: $C_j = 2000$ Gb/s, for 34A $C_j = 4200$ Gb/s, for 34B $C_j = 4800$ Gb/s.

9.3.1 RFA results

Table 9.10 shows the results of the resource allocation problem (B.16)-(B.22). The left part refers to the regular topology while the right one refers to the quasi-regular topology. In the tables, μ_R indicates the network utilization, that is, the ratio between the transport capacity allocated for the TLPs and the global allocated capacity, in accordance with the definition given in [99]. Fig. 9.18 displays the CNs geographical distribution for the optimized 34-node networks.

Figs. 9.16 and 9.17 illustrate the optimized regular and quasi-regular topologies for the 10-node networks. As already mentioned, core nodes are not directly connected, but core nodes and edge nodes can be collocated in the same switching site, so that the direct links between Washington and Philadelphia in the figures, e.g., represent edge-to-core links. Table 9.10 show the results with path protection.

The results with the A matrices provide better values in terms of network utilization. The explanation is again that the B traffic matrices are dense and present many CRs with low traffic demands; thus, the links used by these CRs are under-used. With the quasi-regular topology, the network cost is still dramatically reduced, around 50%, and the network utilization has more than doubled.

	Regular structure				*Quasi-Regular structure*			
Network	*10A*	*10B*	*34A*	*34B*	*10A*	*10B*	*34A*	*34B*
Objective	4260644	4340130	61419238	77837599	2196002	1829473	27086832	36407731
fiber	77.58%	82.25%	82.27%	80.95%	65.85%	71.68%	67.13%	66.33%
CN	11.11%	11.80%	5.35%	5.24%	12.21%	14.28%	4.82%	4.15%
delay	11.31%	5.95%	12.37%	13.81%	21.95%	14.04%	28.06%	29.52%
Time	2768s	2330s	59.59h	62.1h	(same as regular)			
μ_R	23.19%	19.19%	17.38%	13.67%	46.39%	43.18%	46.94%	39.35%

Table 9.10: RFA solution with DPP

	Regular structure				*Quasi-Regular structure*			
Network	*10A*	*10B*	*34A*	*34B*	*10A*	*10B*	*34A*	*34B*
Objective	87%	101%	92%	83%	123 %	118 %	118 %	128%
fiber	86%	99%	91%	82%	133%	111%	129%	150%
CN	85%	100%	89%	82%	140%	124%	111%	157%
delay	92%	146%	96%	87%	92%	145%	96%	87%
μ_R	29%	27%	3.2%	10.3%	-4%	10%	-11%	-24%

Table 9.11: Differences in the RFA solution with and without DPP (in % with respect to the case without DPP)

Given that the cost reductions are due to unused fibers and disabled ports, the cost allocation changes as well with the quasi-regular topology. The fiber cost weight decreases more than 10 percentage points and, thus, the weight of delay and CN costs increases; the delay cost assumes a weight of more than 10 percentage points than in the case of the regular topology where the fiber cost is over-estimated; the CN cost increases even if its absolute value decreases because the cost reduction due to fibers is more important than that due to port cost. In the case of 34-node networks it is evident that the CN cost becomes unimportant at the expense of the delay propagation cost. And this is even more pronounced with a quasi-regular topology. The presence of high order CNs allows assigning the TLPs to a few trunk lines and to decrease the total number of CNs.

Once again, as it happened in the case without path protection (Fig. 9.10), the EN in New York fully uses the connected CNs because it has high traffic CRs. The 10A quasi-regular topology presents the trunk lines Charlotte-Boston, Charlotte-Albany and Philadelphia-Tallahassee disabled, but in this case the network remains survivable. And the quasi-regular topology for 10B is fully meshed with a big number of disabled fibers. As a consequence, as expected, with the quasi-regular topology the network utilization increases very significantly, indicating that, on average, the network fibers are used almost at the 50% of their transport capacity. This is a good result considering that, thanks to the time-sharing, we have better exploited the lambda-channels; the idle capacity is available for further network extensions, such as resource re-provisioning or low-level traffic provisioning.

9.3.2 WTA results

We analyze, for simplicity, only the optical links exiting the EN in Tallahassee for the 10A quasi-regular solution. As of Fig. 9.16, using the quasi-regular topology the trunk line Tallahassee-Washington went from 5 fibers per direction to 1 fiber from Tallahassee

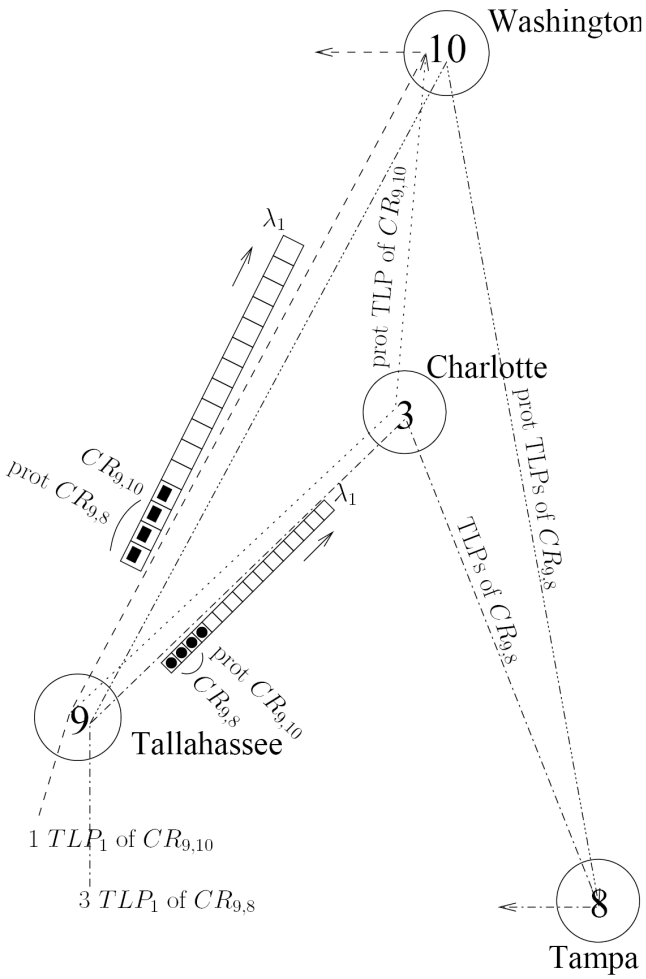


Figure 9.15: Routing and assignment in a study case of 10A solution

to the CN-1 in Washington and 2 fibers from Washington to Tallahassee. Thus the 4-fibers optical link between the EN in Tallahassee and the CN-3 in Washington has been disabled in the quasi-regular topology. Then, Tallahassee is connected to the CN-1 in Charlotte through one fiber per direction.

The EN in Tallahassee has only two outgoing CRs, one of 1.64 Gb/s with Tampa ($CR_{9,8}$), and one of 0.2 Gb/s with Washington ($CR_{9,10}$): the first is accommodated using three TLP-1s and the correspondent pTLPs; the second is served by one TLP-1 and its pTLP. The RFA results indicates that on the fiber going from Tallahassee to Washington one must transport the wTLP of $CR_{9,10}$ and the pTLPs of $CR_{9,8}$, and that on the fiber going from Tallahassee to Charlotte one must transport the wTLPs of $CR_{9,8}$ and the pTLP of $CR_{9,10}$. Fig. 9.15 shows the assignment and the route for these wTLPs and their correspondent pTLPs (only on the outgoing fibers of Tallahassee). The effect of choosing $\delta < 1$ can be appreciated by the fact that the path chosen for the 9-8 wTLPs is the shortest one, a total of 858+1132 = 1999 Km, while the path for the pTLPs is 1574+1799 = 3663 Km. On both the fibers quitting the edge node in Tallahassee we have yet 15 wavelengths and 12 time-slots available. This is possible because the EN in Tallahassee requests resources for only two CRs with low traffic demand.

Changes with respect to the case without path protection

Table 9.3.1 reports the difference in the results with or without DPP. The table presents the increase (in % of the case without DPP) of the objective function, the different costs and the utilization factor. Please note that in Table 9.10the % for the costs represents the weights of a particular cost (i.e., fiber cost) with respect to the total objective function whereas in Table 9.3.1 it represents the percentage increase in cost when DPP is used; likewise, whereas in Table 9.10 the last row represents the utilization, in Table 9.3.1 it represents the increase of the utilization. We can make the following observations:

- The costs increases due to the introduction of DPP are higher for the quasi-regular structure. This could be expected given that such a topology was chosen for cost reduction to begin with;
- The increase in cost for the regular topology is less than double (except for the 10B case) despite the fact that protection was added to all the connections;
- The fiber cost weight is roughly unchanged for regular networks, while it has increased for quasi-regular networks. However, once again, the increase is not that large;
- The global CNs size has approximately doubled, which can be verified in the result tables that shows an almost double absolute CN cost for the DPP case. Indeed, the CNs disposition is very similar than before as it can be stated that normally the CNs of the case without DPP are to be re-enabled in a dual site to switch the pTLPs/wTLPs of their TLPs;
- In the regular case, the network utilization has increased, the links are thus better exploited than before. On the other hand, for the quasi-regular case, the network utilization decreases for nearly all instances (except 10B).

Nevertheless, the execution time has increased, but it is still reasonable. Adding the protection constraint to the heuristic of the regular Petaweb (opportunely adapted for TDM/WDM and DPP), we could set up good CPLEX cut-off values.

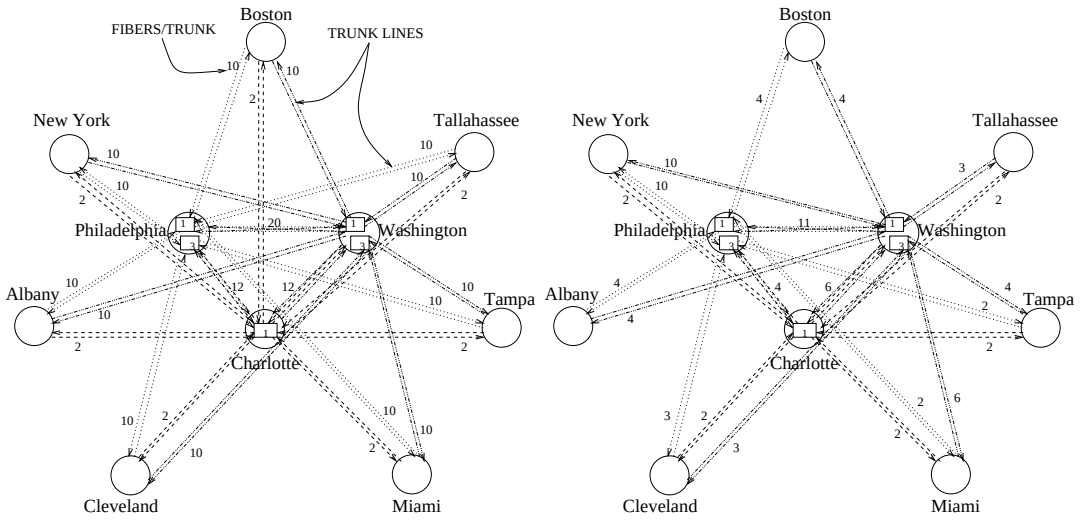
9.4 Optimal quasi-regular Petaweb

In this section we show the results obtained by implementing the resolution algorithms for the direct quasi-regular Petaweb optimization (see Sect. 8.6).

Quasi-regular topology optimization

In Table 9.4 we compare the results obtained by removing unused ports and fibers from the optimal regular Petaweb obtained in the previous experiment (removal), by using CPLEX to optimize formulation (8.14)–(8.24), and by running the three-step heuristic. In order to speed-up the computation, in the heuristic we limited CPLEX to explore at most 100 nodes of the branch-and-bound tree and to perform at most 100 cutting planes iterations at each node. Furthermore, we emphasized the generation of useful cuts by fine-tuning CPLEX internal parameters (their complete setting is available on request).

We report the results in Table 9.4, which consists of three horizontal blocks, one for each solution method. For each method we report the solution value, the computing time required to obtain the solution for each instance (whose id is indicated in the first



(a) 10A regular (b) 10A quasi-regular

Figure 9.16: RFA solution for 10A model with dedicated path protection

column). To complete the analysis, we report also the lower bounds on the optimal solution values obtained by CPLEX, and for the 3-step heuristic we indicate the total number of iterations performed (Tot. iter.) and the iteration in which the best solution was found (Best iter.).

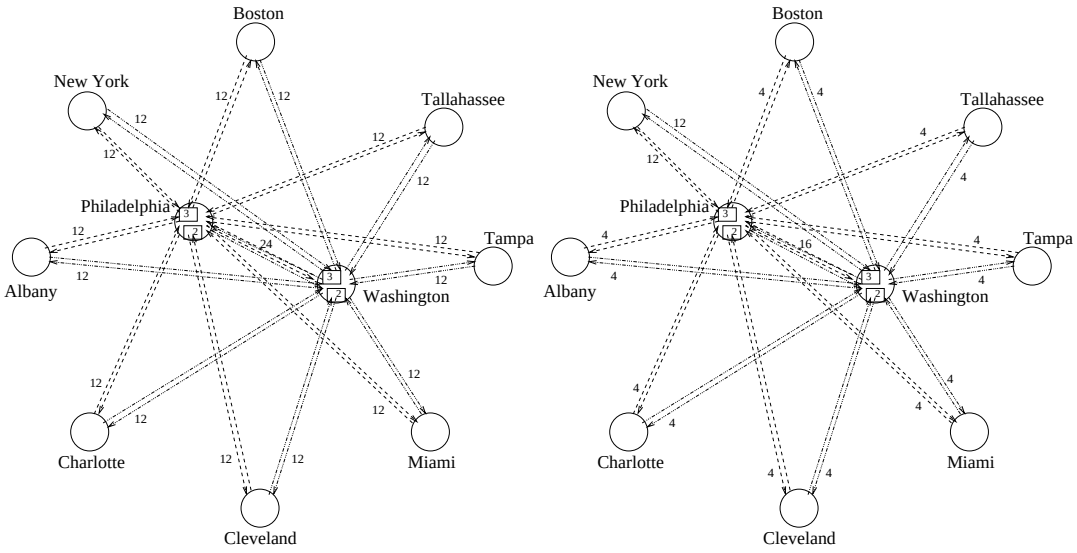
While on the smaller networks CPLEX was able to provide good solutions within two days, it was unable to optimize the larger instances: we aborted the execution after two and four days of computation for respectively instance 34A and instance 34B, and no feasible integer solution was found. The heuristic showed good computational results: on the smaller instances the quality of its solutions is the same as those of CPLEX, but the computing time is significantly smaller; moreover the heuristic substantially improved the best known solutions on larger instances with still affordable computing resources.

Both CPLEX and the heuristic methods showed that directly optimizing the quasi-regular topology actually yields substantial savings in terms of network cost.

A-posteriori analysis of the solutions

Fig. 9.19 show the CN geographical distribution; it can be noticed that the direct optimization of the quasi-regular topology allows a fairer distribution of core nodes. In fact, they are no longer concentrated in the weighted baricenter of the network. This may also offer an additional protection against large-scale disasters.

The quasi-regular topology obtained for instance 10A by downgrading the optimal regular is plotted in Fig. 9.16b, while the quasi-regular topology optimized through (8.14)–(8.24) is plotted in Fig. 9.20. By directly optimizing a quasi-regular topology, a core node is not forced to be connected to every site. Then, it becomes appealing to install core nodes in the regions of the network having isolated peaks of traffic. Some core nodes in Fig. 9.20, e.g., are installed far from the high traffic axes New York - Washington, New York - Philadelphia and Washington - Philadelphia. By looking at Fig. 9.16b, several core nodes were concentrated in these sites, and these axes required 10 or more fibers each. In the direct optimal solutions a big fraction of the traffic is switched to sites closer to some boundary CRs, like Tampa and Albany. In fact, we observed a significant reduction of fiber costs.



(a) 10B regular (b) 10B quasi-regular

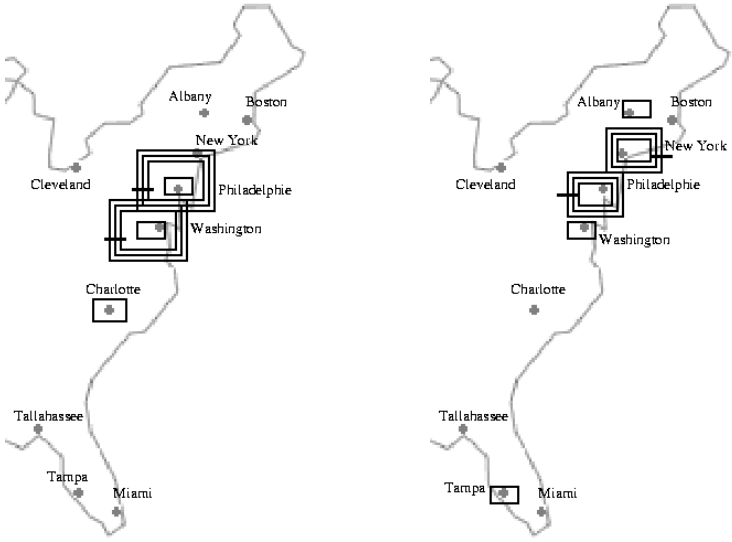
Figure 9.17: RFA solution for 10B model with dedicated path protection

(a) 34A (b) 34B

Figure 9.18: Core nodes geographical distribution for 34-node networks

Fig. 9.21a shows the fiber, delay and core node contributions to the solution cost for the 10A instance. We compared the optimal regular topology, a quasi-regular topology obtained by removal, and the optimal quasi-regular topology obtained with CPLEX. The removal of unused equipment from an optimal regular network implies a significant gain in fiber cost, 10% in cost ratio and 56% in absolute value. By directly optimizing the quasi-regular we obtain a solution in which the fiber cost decreased by 40% (from 1446067 to 863045), the delay cost increased by 7% (from 482022 to 516670), and the core cost decreased by 38% (from 268132 to 165115); with the number of switching planes unchanged at 11 (see Fig. 9.19), the 38% CN cost reduction is due to a lower number of installed CN ports. The impact of the core node costs in this solution remains almost the same, the delay costs are slightly higher, and the fiber costs dropped by 10%. Therefore, we observed that the direct optimization of a quasi-regular topology, besides lowering the fiber needed, may improve the CN configuration in terms of activated ports, and may refine lightpaths routes at the price of slightly higher propagation delays.

Finally, we stress how the direct design of a quasi-regular topology yields a network which is 65-75% cheaper than the optimal regular one (see Fig. 9.21b). The largest improvements can be observed on networks with high traffic loads (B scenarios): the



(a) Quasi-regular by removal (b) Optimal quasi-regular

Figure 9.19: Geographical distribution of core nodes (10A)

direct optimization methods allowed us to obtain a quasi-regular Petaweb 23-45% less expensive than a quasi-regular structure obtained by simple equipment removal.

9.5 Comparison with multi-hop structures

We finally assess the trade-offs between the quasi-regular Petaweb and classical multi-hop partially meshed solutions. Besides the **A** and **B** traffic profiles, we now also treat a **C** type that does not consider the inversely proportional dependence on the square distance, still considering the population product proportionality. **B** and **C** matrices are both opportunely scaled for having a global traffic volume comparable to that of the A matrix, i.e., roughly 0.8 Tbit/s. Each connection request now simply consists in a discrete number of wavelength-channels. In Fig. 9.22 we compare the profiles of the three traffic matrices for 10 node networks. Given that the considered instances are of 10 nodes, the switching plane unit is composed of 10 ingress/egress fibers (see Sect.8.7.1). The delay unitary cost is tuned on four different values: 0, 0.05, 0.1, 0.5 and 1 times the unitary per km fiber cost.

Cost allocation

In Table 9.13 we display the cost allocation solutions, for both the architectures, as function of the delay unitary cost (indicated by β/F). As expected the quasi-regular Petaweb requires more fiber link resources than the multi-hop structure, at most 7% more. This is detected by a higher cost allocation for core nodes in the multi-hop architecture, and by a higher delay cost for the quasi-regular Petaweb. Also expected is the fact that the composite-star architecture has a higher network cost. Indeed, it may be considered as a not optimal special case of the multi-hop structure with strong constraints on the switching system to keep the regularity as a target. These constraints force one switching plane per ingress fiber from a given edge node even if it is the single one connected, while in the multi-hop switching system an equal-in-size switching plane

Instance	Value	Time (s)	
Quasi-regular by removal			
10A	2196002	37	
10B	1829473	5	
34A	27086832	50834	
34B	36407731	3961	
Optimal quasi-regular			Lower bound
10A	1456683	>2 days	1211505
10B	1038975	>2 days	1031920
34A	NA	>2 days	10399274
34B	NA	>4 days	14962700
Quasi-regular by heuristic			Tot. iter. / Best iter.
10A	1544830	242	8 / 8
10B	1034150	176	3 / 2
34A	21541800	29194	11 / 11
34B	21656300	240063	31 / 27

Table 9.12: Results comparison for quasi-regular Petaweb design

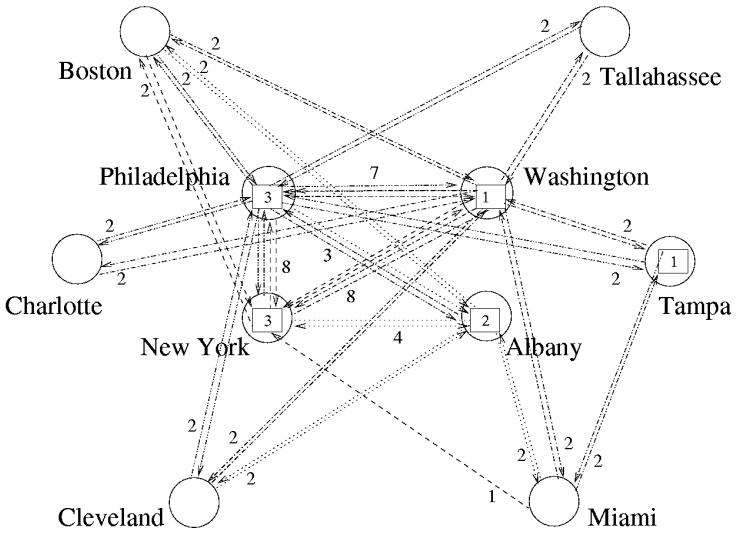
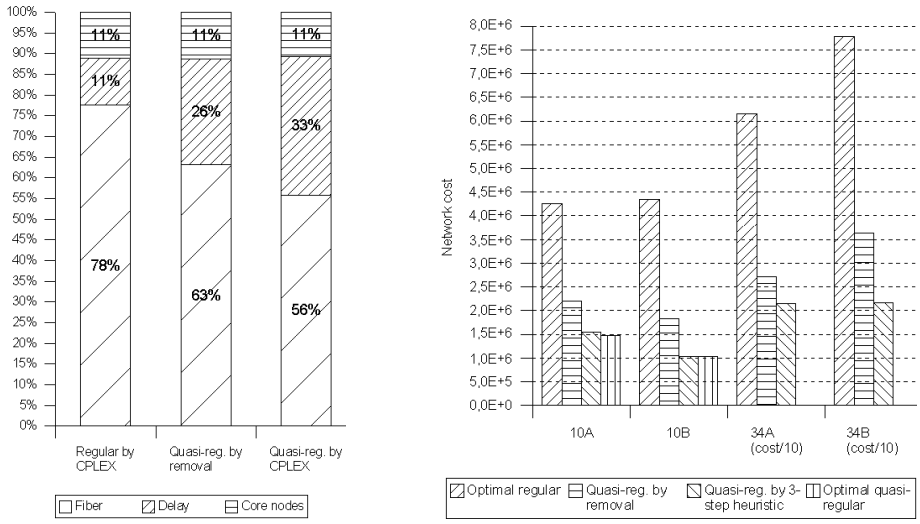


Figure 9.20: Optimal quasi-regular Petaweb topology (10A), with path protection

can house more than one fiber per edge node. However, the difference in cost is not exceedingly high: the quasi-regular Petaweb costs at most 20% more than the multi-hop structure, and in some cases the two solutions are very close. This suggests that multi-hop architectures may tend toward a quasi-regular composite-star structure when one lets all the degrees of freedom to the design dimensioning procedure.

Fiber length

In order to assess the relevance of the propagation delay cost in the design dimensioning, we analyze how it affects the amount of fiber, in km, needed for both the architectures. In Fig. 9.23 we indicate with a '+'point the quasi-regular Petaweb case, with a continuous step the multi-hop case, and with dotted steps the core part and the access part for the multi-hop case. The high fiber cost allocation for the quasi-regular Petaweb indicated in Table 9.13 is reflected by a larger amount of fiber, indeed. The multi-hop architecture allows an efficient fiber distribution and the overall amount slightly increases when the



(a) Cost distribution (10A) (b) Objective comparison

Figure 9.21: Cost distribution and objective comparison

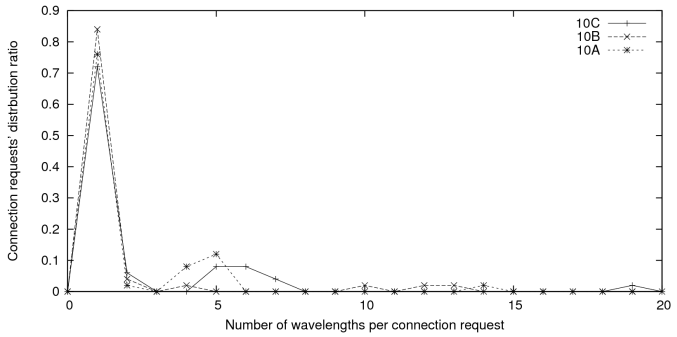


Figure 9.22: Profiles of the traffic matrices

propagation delay unitary cost is increased. Moreover, as the propagation delay is increased much more fiber is allocated to the access than to the core: the km of core fiber decreases, making the multi-hop network similar to the quasi-regular composite-star network, which has no core fibers indeed. If we look at this behavior for the three traffic profiles (10A, 10B, 10C), we can notice that the amount of core fibers is comparable for the three cases, while the amount of access fibers differs. In particular, the 10B profile has a very relevant impact on the access fibersŠ amount, since its matrix is dense and has a lot of single-wavelength connections (Fig. 9.22).

Fiber link utilization

Fig. 9.24 displays, for all the considered cases, the fiber resources utilization as of the classical definition given in [120]. We can affirm that for multi-hop networks:

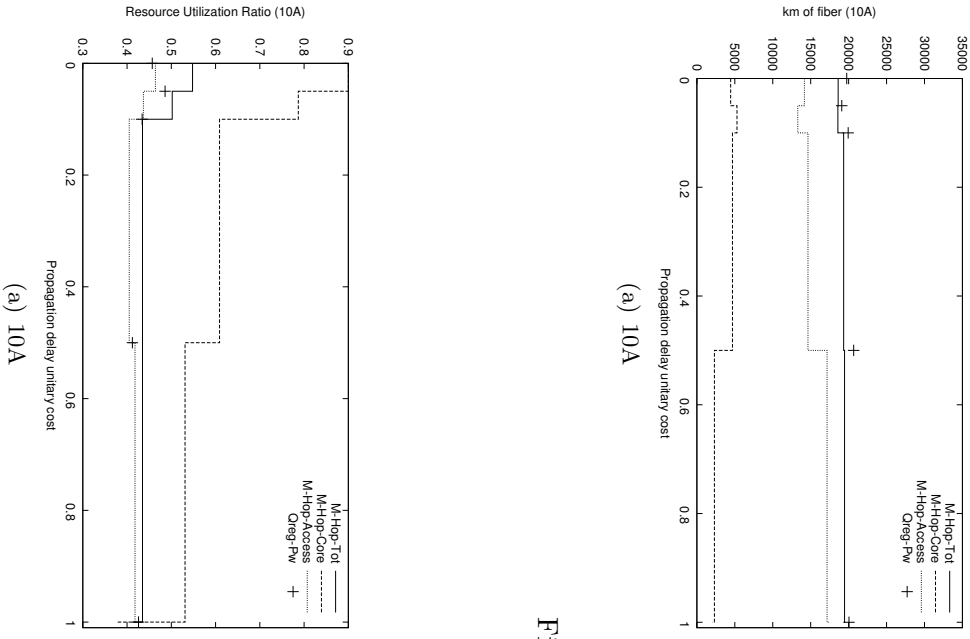
- the access fibers tend to be under-used.
- the resources utilization is worsened by higher weights of the propagation delay cost.

	10A				10B				34A				34B			
CN cost increase	Total cost	CN cost	fiber cost	delay cost	Total cost	CN cost	fiber cost	delay cost	Total cost	CN cost	fiber cost	delay cost	Total cost	CN cost	fiber cost	delay cost
100%	2.6	21%	68%	9%	2.5	22%	74%	4%	39	11%	80%	9%	55	11%	80%	9%
200%	2.9	29%	62%	9%	2.7	30%	67%	3%	39	16%	75%	9%	51	15%	75%	10%
300%	3.1	35%	57%	8%	3.0	36%	61%	3%	41	20%	71%	9%	53	19%	71%	10%

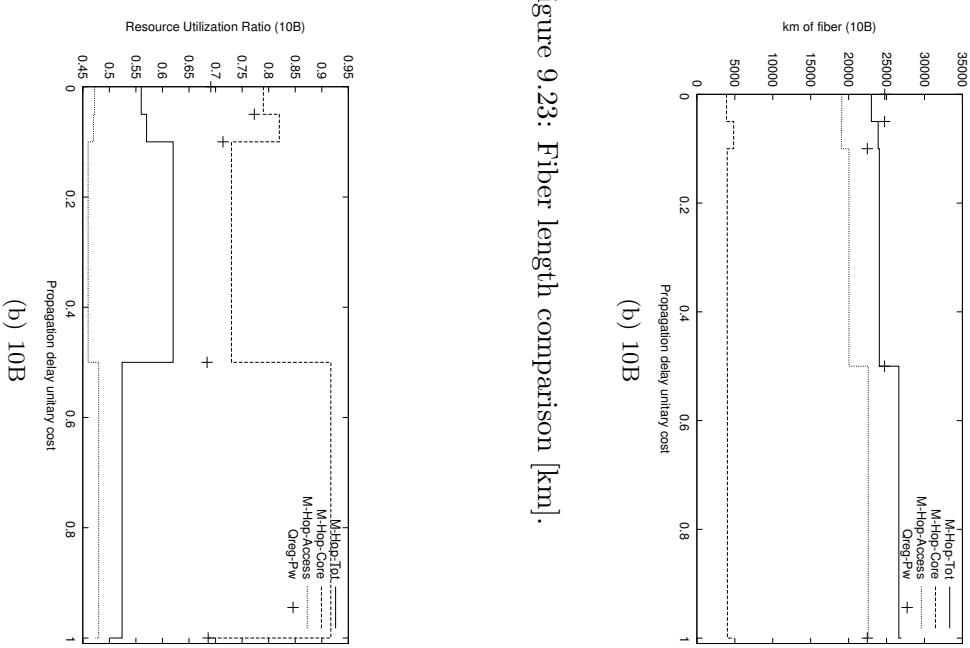
Table 9.13: Effects of the variation of core node cost. CN = Core Node. Total cost in millions ($\cdot 10^6$)

Quasi-Regular Composite Star Architecture															
	$\beta/F = 0$			$\beta/F = 0.05$			$\beta/F = 0.1$			$\beta/F = 0.5$			$\beta/F = 1$		
Case	10A	10B	10C	10A	10B	10C	10A	10B	10C	10A	10B	10C	10A	10B	10C
Objective	376	538	454	392	587	399	399	691	465	458	753	578	630	1019	645
fiber cost	86.2%	77.4%	86.4%	84.7%	76.7%	84.1%	82.6%	73.0%	82.4%	72.3%	63.7%	70.2%	61.9%	53.9%	60.9%
CN cost	13.8%	22.6%	13.6%	13.3%	21.0%	13.6%	13.7%	22.5%	13.2%	12.1%	18.6%	11.9%	10.7%	16.2%	9.4%
Delay cost	0.0%	0.0%	0.0%	2.0%	2.3%	2.3%	3.7%	4.5%	4.4%	25.7%	17.7%	17.9%	27.4%	30.0%	29.7%
Multi-Hop Irregular Mesh Architecture															
	$\beta/F = 0$			$\beta/F = 0.05$			$\beta/F = 0.1$			$\beta/F = 0.5$			$\beta/F = 1$		
Case	10A	10B	10C	10A	10B	10C	10A	10B	10C	10A	10B	10C	10A	10B	10C
Objective	372	530	390	382	571	397	398	605	397	457	690	480	523	801	572
fiber cost	80.0%	73.2%	80.3%	77.7%	72.3%	78.4%	77.6%	69.4%	76.9%	68.1%	61.7%	65.9%	59.2%	53.8%	56.2%
CN cost	20.0%	26.8%	19.7%	20.1%	24.8%	19.4%	18.7%	27.6%	18.8%	15.8%	23.0%	16.5%	13.8%	20.7%	14.7%
Delay cost	0.0%	0.0%	0.0%	2.1%	2.9%	2.3%	3.7%	3.0%	4.3%	16.1%	15.3%	17.6%	27.0%	25.5%	29.1%

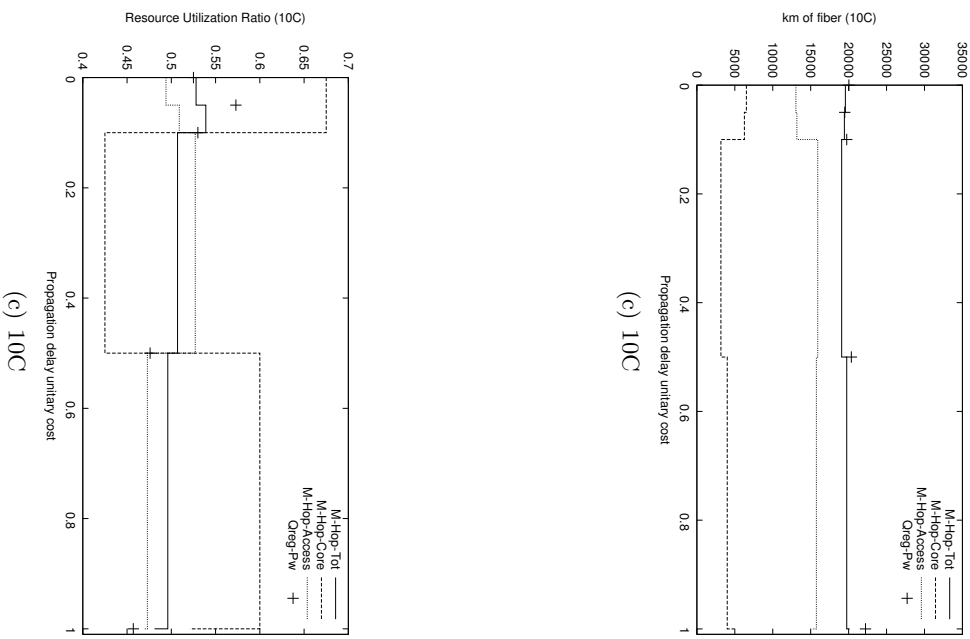
Table 9.14: Cost allocations as function of the propagation delay unitary cost and of the traffic profile. The objectives are in thousands; CN = Core Node.



(a) 10A



(b) 10B



(c) 10C

Figure 9.23: Fiber length comparison [km].

(a) 10A

(b) 10B

(c) 10C

Figure 9.24: Resource utilization ratio comparison.

Chapter 10

Performance Trade-offs for the Petaweb Planning

In this chapter we analyze performance trade-offs of planned Petaweb networks. We arise different criteria under which the various architectural solutions shall be evaluated. In particular, we stress the performance robustness of the Petaweb solutions with respect to reliability, survivability and availability aspects. Finally, given the quite large number of criteria that the decision maker (network planner) might consider when taking a decision on the final Petaweb structure to adopt, we present the related multi-criteria network planning decision problem and a possible way to solve it by means of Interactive Decision Maps[1].

10.1 Robustness Performance Evaluation

Network robustness refers to several measures related to the reliability, survivability and availability of the designed network. Reliability is meant as the capability of a system to durably grant the service, i.e., the capability to reduce end-user perceived failures; survivability represents the capability to rapidly recover a service from a failure that interrupted it; availability signifies the capability to fully access the service when it is required. We analyze how the Petaweb behaves with respect to these aspects with both regular and quasi-regular structures, and dedicated path protection. In our case, a failure can be a trunk line cut, a switching plane damage, a core node disconnection or a switching site disconnection.

10.1.1 Reliability and Seamlessness

When the network can recover seamlessly, i.e., automatically and immediately, from a network element failure, it can be considered reliable because the failure is not perceived by the end users beyond the edge nodes.

Looking at Figs. 9.16 and 9.17 one can notice that the optimized quasi-regular network is more reliable with DPP. Every EN gets connected to at least two switching sites, and every connection is split into trunk-disjoint wTLPs and pTLPs, so that the network is reliable against trunk line failure, core node or switching site disconnection. In the case of failure of one of the two trunk lines where a wTLP passes, e.g., the destination

[1]The contents presented in this chapter are also presented in [3], [16] and [18].

EN can recover the traffic of the wTLP from the pTLP. Moreover, a configuration with all the core nodes installed in the same site is no longer allowed. With DPP, there are at least two different switching sites, and if a single site is totally disconnected, or if a single core node is damaged, all the traffic can be recovered by core nodes in the other(s) enabled switching site(s). Whether a single switching plane of a core node fails, only the TLPs switched by that plane would be affected. Indeed, even the largest TLP class, TLP-3, requires a single switching plane. Whether an entire switching site gets disconnected, or the wTLP or the pTLP of each connection request is also switched in another site, and thus the service is not interrupted.

There exists a trade-off between the level of reliability and the geographical range of the network to be planned. The more the network sites are distant, the higher the trunk line failure probability. The more a trunk line failure is probable, the less reliable the network can be in the case of multiple failures.

Since the Petaweb is expected to offer longer lightpaths than classical mesh architecture solutions, this issue acts as shortcoming for the Petaweb topology, as it will tend to increase link failure probability. All in all, we can state that the wider the network, the higher the link failure probability. Even if concurrent double failure on disjoint trunk lines - possibly causing an unprotectable outage in a DPP-planned Petaweb - would still remain a very rare event, its occurrence may affect very large volumes. It is thus interesting to assess how much traffic may be reprovisioned in a planned network for such critical failure cases.

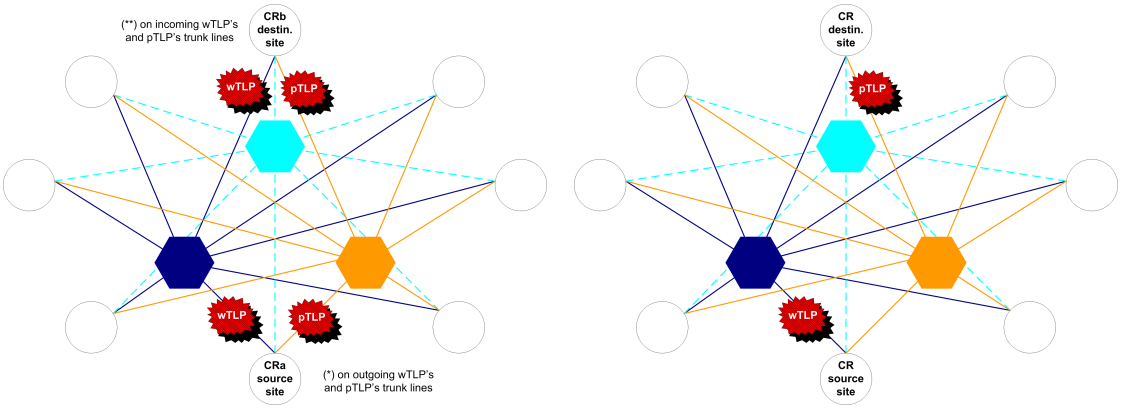
10.1.2 Survivability and Reprovisioning

The designed network is thus reliable with respect to single failures of fiber links, switching planes, core nodes or switching sites. It even stays reliable in the case of multiple fiber link, switching plane or core node failures if, respectively, the fiber links are part of the same optical trunk line or the switching planes or the core nodes operate in the same switching site. For such cases in a DPP Petaweb, the reprovisioning time (also measurable as Mean Time To Repair, MTTR [101]) is null. Otherwise, one may have service interruption if no other resources are available on alternative paths. In such a case, the interval between the downtime and the next uptime can be an index of the survivability degree of the network.

To recover from multiple failures blocking both the wTLP and the pTLP of a connection request, reprovisioning functions should be implemented at a control plane. Reprovisioning is possible if alternative resources can be instantiated. In such a case, the MTTR is expected to be under the expected MTTR for multi-hop networks, since the resource reservation over two optical links can be performed faster than for lightpaths with more than two links and one core switch.

Hence, if alternative resources are available, the source edge node should be able to select and instantiate them and reprovide the TLP. To evaluate the survivability of the Petaweb networks with DPP we monitor, for each wTLP-pTLP pair, the network ability to perform TLP Reprovisioning in the following blocking cases:

- C.1: double trunk line failure, at wTLP's and pTLP's trunk lines.
  - C.1.1: both the trunk are egress lines (Fig. 10.1a*) - from the source EN toward the switching CN - or ingress lines (Fig. 10.1a**) - from the switching CN toward the destination EN.
  - C.1.2: one trunk is an ingress line, the other an egress line (Fig. 10.1b).



(a) C.1.1 both egress (*) or both ingress (**)

(b) C.1.2 one egress and one ingress

Figure 10.1: C.1: double trunk line failure

- C.2: double core node failure, at both the wTLP's and pTLP's CNs.

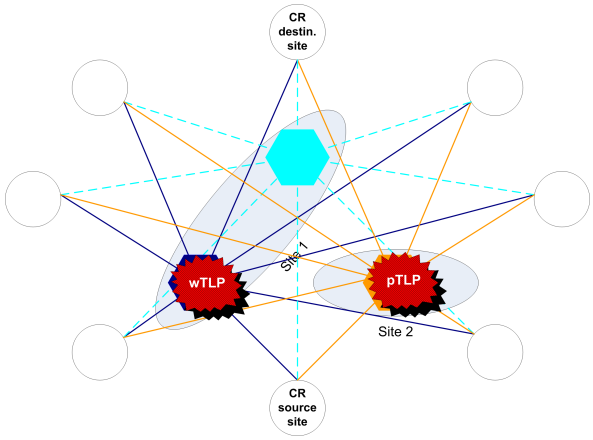


Figure 10.2: C.2: double core node failure (one switching site is here represented as composed of two core nodes, while the others of a single one).

- C.3: double switching plane failure (Fig. 10.3), at the wTLP's switching plane and at the pTLP's switching plane.
- C.4: double switching site failure (Fig. 10.4), at the wTLP's switching site and at the pTLP's switching site.

In what follows, we analyze the results with DPP for all the topology cases and for both regular and quasi-regular structures. For each network case (10A, 10B, 34A, 34B with regular or quasi-regular structure) we look after the chance of reprovisioning a single working TLP when both the corresponding wTLP and pTLP have been interrupted because of multiple equipment failures. We consider the C.1-C.4 multiple failure cases given above.

Table 10.1 indicates for each network case the ratio of TLPs that could be reprovisioned, having all the wTLP-pTLP pairs been considered under each multiple failure case. For those TLPs that could not be reprovisioned, the index indicates the ratio of traffic volume that might be reprovisioned fragmenting the TLP in several lower class

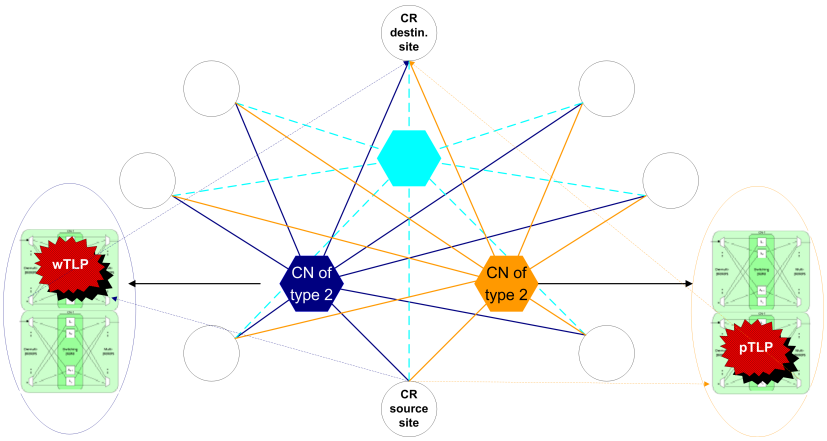


Figure 10.3: C.3: double switching plane failure (the core nodes where the failed switching planes are located are as being of type 2).

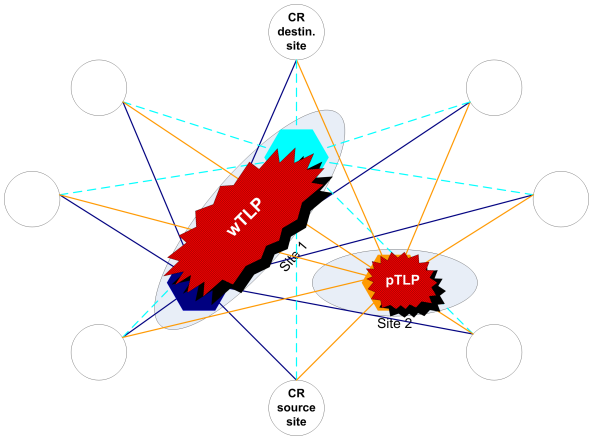


Figure 10.4: C.4: double switching site failure (the switching site is represented as composed of two core nodes, while the others of a single one).

TLPs. These two transversal parameters offer a good insight on the survivability of the network solution. As of Table 10.1, we can assess that:

- Large networks seem to perform better than smaller ones, as all the 10-node instances provided worse robustness results than the equivalent 34-node ones. This can be explained by the fact that larger switching site diversity and larger resource availability guarantee better TLPs reprovisioning, and this is the case for larger networks.
- Better robustness results are generally obtained when the traffic matrix is sparse (for the A cases). This can also be explained by larger resource availability produced by the extra sites that are to be considered in the case of the networks designed with the sparser matrix. In Figures 9.16, 9.17 and 9.18, e.g., we have, respectively, 3, 2, 7 (Fig. 9.18)a) and 8 (Fig. 9.18)b) switching sites; indeed, the more switching sites, the larger path diversity and resource availability there will be in case of failures.
- As expected, quasi-regular structures suffer much more than regular ones from equipment failure and site disconnection. Indeed, a larger switching and transport

resource availability guarantees better TLPs reprovisioning. Therefore the decision on the adoption of a quasi-regular or a regular structure might be pondered by a statistical analysis on the failure probability of the different network elements and on the site disconnection probability.

Particular observations

- with double trunk line failures (C.1) we have different survivability features for the two subcases:
  - with failures on both ingress or both egress trunk lines (C.1.1), a rapid glance to Figures 9.16-9.17 would suggest that the network should have a very low, close to zero reprovisioning ratio for 10-node networks because most of the edge nodes are connected to the backbone through only two trunk lines: when both trunk lines fail these edge nodes would get disconnected. However, Table 10.1 reports that the reprovisioning ratio is not close to zero, but between 3% and 14% for 10-node networks, and between 25% and 58% for 34-node networks. Indeed, after a closer look to Figs. 9.16-9.17 one may note that those edge nodes connected to the backbone with more than two trunk lines (e.g., Philadelphia and Washington sites for 10B), while being collocated with some core nodes, are those likely to be source or sink of most of the traffic.
  - when the two failed links are one ingress and one egress at different edge nodes (C.1.2), the reprovisioning performance significantly increases. The TLPs reprovisioning ratio goes over 60% for 34-node regular networks and over 15% for 10-node regular networks, for instance. For those TLPs that could not be totally reprovisioned, the resource reprovisioning ratio may be, however, satisfactory especially for large networks. Seemingly the TLPs that can be easily reprovisioned are those with a low rate.

- with double core node failure (C.2), a majority of the TLPs can be reprovisioned, but those TLPs that can not be reprovisioned are likely to be those with the highest bit rates. Indeed, we can notice that the index - representing the fraction of traffic of the failed connection that could be reprovisioned - are always values around 20% or less, i.e., only roughly 20% or less of the traffic of those TLPs that could not be totally reprovisioned might be reprovisioned (through lower class TLPs). This seems to be due to the fact that TLPs can be easily reprovisioned and switched by another core node, possibly co-located with the failed one, only if their rate does not exceed the idle capacity on the corresponding fibre links (probably equal to a small fraction of the link capacity).

- in the case of double switching plane failure (C.3), TLP reprovisioning is possible with very good statistics, reaching 100% success with regular 34A networks. This confirm the expectations of [117] about the high switching core reliability of the Petaweb core architecture.

- in the case of double switching site disconnection (C.4), even if the event presents very low probability of occurrence, 10-node networks are almost totally blocked, and 34-node networks can get seriously damaged.

  Indeed, the designed 10-node networks dispose of only 2 (10B) and 3 (10A) switching sites, while 34-node networks only 4 (34B) and 5 (34A) switching sites. For the

MODEL	*10A*		*10B*		*34A*		*34B*	
failure case	reg	q-reg	reg	q-reg	reg	q-reg	reg	q-reg
C.1: double trunk line								
C.1.1: both ingress/egress	$.13_{.40}$	$.7_{.19}$	$.8_{.33}$	$.3_{.15}$	$.57_{.74}$	$.28_{.59}$	$.44_{.68}$	$.25_{.52}$
C.1.2: otherwise	$.25_{.49}$	$.21_{.31}$	$.22_{.41}$	$.18_{.34}$	$.65_{.81}$	$.39_{.72}$	$.66_{.9}$	$.35_{.71}$
C.2: double core node	$.71_{.21}$	$.46_{.15}$	$.63_{.16}$	$.34_{.08}$	$.96_{.23}$	$.74_{.13}$	$.93_{.11}$	$.51_{.14}$
C.3: double switching plane	$.88_{.52}$	$.53_{.24}$	$.69_{.31}$	$.39_{.11}$	$1_{.0}$	$.75_{.82}$	$.96_{.74}$	$.77_{.62}$
C.4: double switching site	$.05_{.10}$	$.04_{.03}$	$.00_{.00}$	$.00_{.00}$	$.31_{.34}$	$.18_{.20}$	$.25_{.11}$	$.8_{.31}$

Table 10.1: Reprovisioning capabilities under multiple network equipment failures

10B case, double switching site disconnection blocks all switching simply because we have only two switching sites in the dimensioned network. For the 1OA case, only one can survive. For the 34A and 34B cases, only a few switching planes would get overloaded and would create congestion at the edges. However, whether the network planner disposes of certain statistics about specific failure site, additional site diversity or site avoidance constraints might be easily added to the design model in order to avoid installing large switching equipment in dangerous sites.

10.2 Availability and Petaweb Upgrade

The availability is an important criterion to evaluate the performance of a communication network, and in particular of WDM networks [141]; it represents the network ability to provide services. This depends on the availability of enough spare capacity and switching resources to serve new connection requests. Future connection requests might stem from a CR extension or from the addition of new ENs or network sites. We have found that optimized survivable Petaweb networks still present a significant amount of idle capacity, roughly 50% of the capacity resources remain available to accommodate further bandwidth requests.

A connection request extension would operationally consist in one or more new TLPs to be provided. The new TLPs' provisioning may or may not be feasible with respect to the available resources and with respect to DPP or delay constraints. Local equipment addition may be required and the upgrade problem may become complex. In the following, we treat such a problem, analyzing both the cases of quasi-regular and regular Petaweb network upgrade

Related work

An increase in traffic volume imposes changes in the network configuration; there are two ways to face such an increase: by updating or by upgrading the network.

Updating a transport network means configuring new circuits to further exploit the available equipment and resources; in that case, a reconfiguration of the network virtual topology may be useful to free more resources via re-optimization and hence postponing network upgrades [150]. For example, in [151] the authors tackled the problem of accommodating an expansion of the original traffic matrix for a pre-optimized WDM mesh network with the restriction that no more physical equipment should be added to

the existing infrastructure, and that only the existing idle capacity could be exploited without touching active lightpaths.

Upgrading a network means resizing its infrastructure and, optionally, reconfiguring its routes. An upgrade may require removal and/or addition of new equipment to satisfy a set of new end-to-end requests. In [152] the upgrade design problem for WDM mesh networks is solved through a methodology that exploits the idle capacity of an optimized network adding more resources if the idle ones are not enough. They do not consider reconfiguring the original connections.

In the following, we focus on the upgrade of the Petaweb architecture without reconfiguration; in an edge-controlled transport network, such as the Petaweb, the reconfiguration of the lightpaths routing would imply high data flows interruption for a significant gap of time. Moreover, the re-optimization of active lightpaths becomes no more an essential operation for this composite-star architecture because all the idle capacity is directly exploitable, differently than with meshed WDM networks. Another feature not explicitly taken into account in our model is the equipment removal. Even though that might be considered in some networks [121], it is not a real option in nationwide optical transport networks. In any case, this is a feature that can be easily incorporated into the model.

10.2.1 Petaweb upgrade

The proposed upgrade model for the Petaweb not only considers the addition of new core node and fiber equipment, but also the exploitation of the idle capacity that is present in the initial architecture. In fact, we found that optimized Petaweb networks still present a significant amount of idle capacity, which remains then available to accommodate subsequent bandwidth requests.

When the traffic volumes or the number of the connection requests between the existing edge nodes increase, they need new TLPs for which a route through a core node has to be decided. Thus, some switching sites that had few or low bit-rate Connection Requests, may now be updated with more core nodes, optical fibers and links. New switching sites may also be opened. We also consider that new edge nodes can be added to the network, which could imply that the opening of new switching sites is even more likely.

As previously stated, equipment removal is not considered in our update model. Therefore, the *upgrade cost* only includes the cost of the *new equipments* (i.e., fibres, core nodes and ports) and a propagation delay cost of the *new TLPs*. This latest term is added to make sure that the new TLPs are not routed on paths that will produce too long propagation delays. The logic of the upgrade model is to keep the initial Petaweb topology, whether it is regular or quasi-regular. Moreover, it is assumed that the existing optimized network was designed with a survivable strategy (DPP model) that is kept after the upgrade. Since we assume that existing core nodes and fibers cannot be removed, and that the number of new TLPs are likely to be fewer than the existing ones, the complexity of the upgrade problem is reasonably lower than that of the initial planning problem.

A comprehensive ILP formulation for the Petaweb upgrade for both the regular and the quasi-regular structure is presented in Sect. B.4.

10.2.2 Numerical results

The initial network status is defined by the 10-node networks dimensioned with TDM/WDM and DPP. We consider two scenarios: simple traffic increase and traffic increase with edge node additions.

Traffic increase

In this study case the traffic of every existing Connection Request is increased by 200%. The left side of Table 10.2.2 reports the upgrade results obtained solving the formulation (B.31)-(B.38) for the 10A and the 10B pre-planned network with regular and quasi-regular topologies.

The results show that the network utilization μ_R, defined as the ratio between the used and the available capacities, increases for both topologies as illustrated in Fig.10.5. Such an increase is more important for the regular topology. This is due to the equipment already installed that allow the new TLPs to be routed more efficiently than with the quasi-regular topology. This behavior seems to be confirmed by the average path length (weighted on the traffic unit), indicated by ν in Table 10.2.2; it is slightly bigger with quasi-regular topologies. This seems to indicate that, if a network operator foresees regular upgrades and want to route its TLPs in the most effective way, the cost to pay is the initial regularity. For the 10A model, e.g., the added traffic amount exploits mainly the idle capacity without enabling lots of new switching planes; indeed, the network utilization (μ_R) went from 23.19% and 46.39% to 31.6% and 54.1%, and the weight of the fiber cost felt by 5-7 percentage points (as depicted in Fig.10.6). In this case one can notice how the upgrade cost is bigger for a quasi-regular topology than for a regular topology; in the first case one has to install fibers that, instead, with a correspondent regular topology may have already been installed.

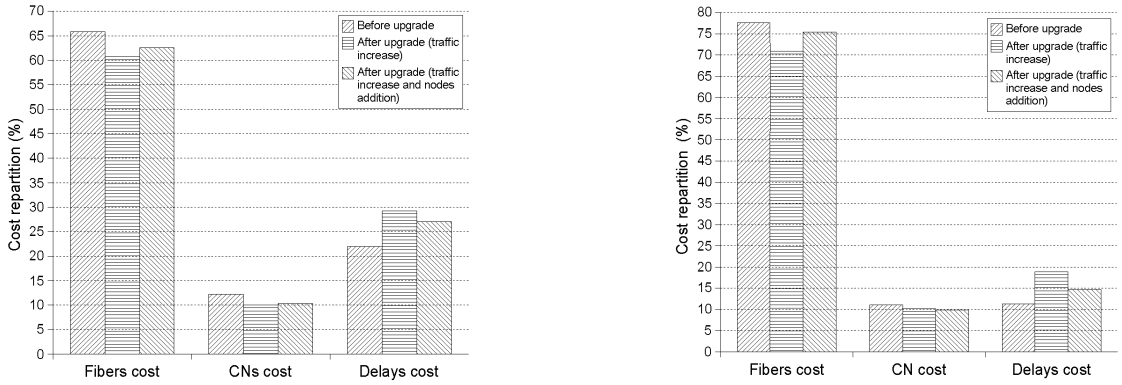
In Table 10.2.2 the cost distribution concerning only upgrade costs and the one concerning the whole network equipments (those installed before the upgrade together with those installed after) are portrayed. For comparison purposes the global cost distribution before and after the upgrade for the 10A case are illustrated in Fig. 10.6. We can see that the most remarkable effect of the new TLPs and subsequent network upgrade is an increase of the weight of the cost due to propagation delays and, thus, a decrease of the fiber cost and of the CN cost weights. The upgrade cost fraction due to new fibers and CNs is minor if compared to the one related to the whole network; on the contrary, the upgrade fraction due to the delay of the new TLPs is significantly bigger than the one related to the whole network. This confirms that the upgrade tends to exploit the existing resources rather than requiring new ones. And the difference is more evident for the upgrade of a quasi-regular topology, because the existing fibers are better exploited, and, even if new fibers are placed, the overall fiber cost weight still decreases.

Traffic increase and edge nodes addition

In this study case we increased by 200% the existing CRs, and we added 4 ENs to the existing ones. The right side of Table 10.2.2 displays the upgrade results obtained for 10A and 10B with, respectively, regular and quasi-regular topologies. The resulting networks are composed of 14 ENs. As expected, the addition of new ENs causes a large upgrade cost because of the new trunk lines that are needed to connect the new ENs to at least two CNs (because of the DPP constraint). Fig.10.6 shows how the equipment cost weights are still smaller than the correspondent value for the pre-existing network,

(a) regular topology (b) quasi-regular topology

Figure 10.5: Network utilisation before and after the 10A upgrade.



(a) regular topology (b) quasi-regular topology

Figure 10.6: Cost distribution before and after the 10A upgrade

but, in the case of the fiber cost, slightly higher than the value for the case with only traffic increase. Fig.10.5 reflects that under EN addition the overall network utilization may decrease, as it happens for the 10B case; indeed, the new installed trunk lines are under-used with respect to the old ones that were better exploited.

Observing the upgrade cost in the two cases we notice that it is lower for an existing quasi-regular topology. When new edge nodes are added to the network, they can be integrated installing new equipment, mainly new optical links. And with quasi-regular topologies these new optical links are composed only of the essential number of fibers, and nothing more. We can conclude that the upgrade with ENs addition is more convenient if one adopts a quasi-regular topology; the cost gain is significant and the network operator may prefer to start with a quasi-regular topology and to upgrade it only in case of new ENs addition. Until new ENs have to be added, it may be possible to accommodate increases of traffic only exploiting the present idle capacity, without additional physical equipments, i.e., through updates (see section 10.2); this update method may be a subject for further work (similar to the method used in [151] for mesh networks).

	traffic increase				traffic increase and nodes addition			
	regular topology		quasi-reg. top.		regular topology		quasi-reg. top.	
Model	10A	10B	10A	10B	10A	10B	10A	10B
cost	856983	975431	1304076	717565	4817336	3471536	2984283	1670231
upgrade cost distribution								
fibre	35.7%	74.9%	52.3%	69.6%	73.1%	32.2%	60.2%	75.0%
CN	5.7%	10.7%	6.3%	11.2%	8.8%	2.7%	8.9%	9.7%
delay	58.6%	14.4%	41.4%	19.2%	18.1%	65.1%	30.9%	15.3%
global cost distribution								
fibre	70.9%	80.9%	60.8%	71.1%	75.4%	82.7%	62.6%	73.3%
CN	10.2%	11.6%	10.0%	13.4%	9.9%	10.7%	10.3%	12.1%
delay	18.9%	7.4%	29.2%	15.5%	14.7%	6.5%	27.0%	14.6%
μ_R	31.6%	22.4%	54.1%	45.6%	26.9%	17.5%	51.7%	38.3%
ν	2924	894	2953	911	1719	1151	1717	1138
time (s)	1.3	23.9	1.2	36.2	1269	128	1178	86

Table 10.2: Upgrade solutions

	traffic increase				traffic increase and nodes addition			
	regular topology		quasi-reg. top.		regular topology		quasi-reg. top.	
Model	10A	10B	10A	10B	10A	10B	10A	10B
Cost	1874996	999035	1369553	1029584	6894998	3551366	3278257	1791541
gap	+118%	+2%	+5%	+43%	+43%	+2%	+9%	+7%
upgrade cost distribution								
fibre	69.2%	75.6%	43.2%	66.7%	76.5%	83.9%	60.8%	69.8%
CN	10.1%	11.3%	5.8%	12.7%	9.6%	9.4%	8.7%	10.7%
delay	20.7%	13.1%	51.0%	20.6%	13.9%	6.7%	30.5%	19.5%
global cost distribution								
fibre	75.3%	81.0%	63.7%	69.9%	77.1%	83.0%	62.8%	70.9%
CN	10.8%	11.7%	10.9%	13.7%	10.2%	10.7%	10.1%	12.5%
delay	13.8%	7.3%	25.4%	16.5%	12.7%	6.3%	27.1%	16.6%
μ_R	24%	21.5%	49.1%	47%	20.5%	16.7%	50.4%	42.6%
ν	2375	871	2375	871	1798	1074	1798	1074
gap	-23%	-2.6%	-24%	-4.5%	+4.6%	-7%	+4.7%	-5.9%

Table 10.3: Greedy upgrade solutions

Comparison with a greedy upgrade

In this section we comment on the results obtained applying an upgrade method based on a straightforward greedy strategy when compared with the results obtained with the method proposed in this paper. The greedy upgrade can be described as follows: when a new TLP is created, it is switched in the closest switching site with core nodes already installed. Then the two trunk lines supposed to route the TLP may be opportunely resized and new core nodes may be installed at that site. Note that the edge node capacity constraint (B.35) may not be respected for regular topologies. In such a case, the edge node should be replaced.

We analyze the results for the two previous study cases. The behavior of the greedy method is the same for an existing regular or quasi-regular architecture. In either case, the resulting network is suboptimal as can be seen in Table 10.2.2 where the gaps with respect to the optimal solution given by the upgrade are depicted. It can be seen that the greedy update may yield a solution costing twice as much as a solution produced by the optimized procedure. Interestingly, the worst differences are produced with the 10A matrices. For the 10B cases, the upgrade cost is not too large compared with the previous values; only one new switching plane was required. But, along with the 10A cases, we can see the worst values of fiber cost and network utilization: the route for TLPs was not carefully chosen. In terms of average path length, the greedy method gives better values than the optimal method; this can be seen by the gaps with respect

to the optimal solutions that are negative in almost all the instances. Clearly, plugging new connections to the closer CNs produces an improvement in overall path length, but this is often a more expensive choice with respect to the cost model. The upgrade cost distribution has a behavior very close to that of the global cost distribution: the greedy method can not profit efficiently of the available resources.

The proposed upgrade model underlines the importance of conducting a cost-effective upgrade when compared with the common practical paradigm 'plug where it is closer', often used in the industry. Such greedy upgrade provisioning method applied to the Petaweb brings a lower network utilization at significantly higher cost.

10.3 Final network planning decision: multiple criteria and Interactive Decision Maps

Analyzing the previous results, we might conclude that a regular topology is advisable if the network operator has a good initial budget and if frequent upgrades are foreseen; a quasi-regular topology is the best choice in case of low budget and rare upgrades, especially when the upgrade contemplates edge nodes addition.

Nevertheless, we are contemplating a scenario in which the network planner's decision shall be taken with respect to possible several other criteria. This kind of decision problem relates to multi-criteria decision-making where the decision maker is a human being with personal conscious preferences on the criterion trade-offs.

Given the quite large number of criteria (around cost, performance, quality of service and reliability) that one might consider when taking a network planning decision, a single-objective optimal choice is not pragmatically advisable. Differently, the decision problem shall consider the multiple criteria, falling in the class of Multiple Objective Optimization problems. In this context, also because eventually such planning decisions are always taken by human beings, and never by machines, the Interactive Decision Map (IDM) approach proposed by A. Lotov et al. is of interest [52]. In particular, the Reasonable Goals Method (RGM) technique with IDM seems a particularly expressive multi-criteria method to select alternatives from finite lists. RGM/IDM may represent a possible solution for network planning operations. After a quick review of the Reasonable Goals Method, we characterize our study case and present the simulation tests we performed using the Visual Market/2 (VM/2) software.

10.3.1 The Reasonable Goals Method

The RGM method is introduced for IDM in the book [52]. It relies on a representation of decision alternatives as points in a decision map, and on approximating and exploring the corresponding convex hull (i.e., the frontier of the envelope of alternatives). In particular, the approximation results in the so called Convex Edgeworth-Pareto Hull (CEPH), that includes the Pareto-frontier and all dominated alternatives. This operation corresponds to the first line step of the chart in Fig. 10.7. Fig. 10.8a depicts the CEPH of an example bi-objective decision problem of a worth that is evaluated with respect to its price and age. After, the variety of variants is considered by a computer (C) to process the CEPH. Then, the CEPH is displayed via IDM so that the Decision Maker (DM) can dispose of a visual representation of the interesting alternatives and their trade-offs. This interaction result in a choice from the DM of a *reasonable goal*, e.g., Fig. 10.8b.

The identified reasonable goal is thus a combination of the criteria not necessarily corresponding to an existing alternative, that is, an identified but potentially infeasible

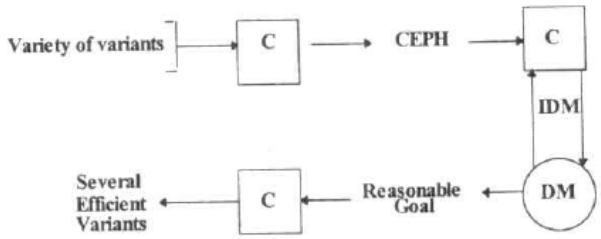
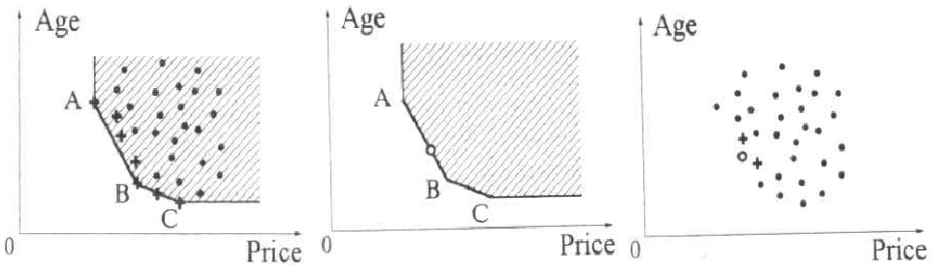


Figure 10.7: The steps of the RGM/IDM technique. C: computer processing. DM: Decision Maker. Source: [52].



(a) Convex Edgeworth-Pareto Hull (CEPH)

(b) The identified goal is denoted as an empty point

(c) The points close to the goal are selected

Figure 10.8: Identification of the reasonable goal (with minimization). Source: [52].

goal to which the DM aspire. The representation the CEPH alternatives via IDM induces a conscious identification of the reasonable goal by the DM. The next step is a selection of several feasible points close the identified goal (Fig. 10.8c), hence those that reflect the subjective preferences of the DM.

A choice has to be taken among the selected feasible alternatives using the identified goal. The RGM/IDM procedure is the following; Fig. 10.9 is a representative example, in which both the criteria shall be maximized, the filled circle are original feasible points close to the reasonable point, i.e., the square. First, the original points are projected toward the bound imposed by the reasonable point. Modified points – e.g., empty circles 1', 2', 4' and 5' in Fig. 10.9 – are then considered instead of original points. Those original points strictly dominated by the reasonable goal are not modified – e.g., point 3. Among the modified points and those possibly not strictly dominated by the goal, only the Pareto-superior one are kept, i.e., those that are not dominated by other modified points – e.g., 2', 3 and 4'. Finally, the original points corresponding to the Pareto-superior modified points are kept – e.g., 2, 3 and 4. The DM then receives the efficient variants.

10.3.2 Decision criteria and planning alternatives

The possible criteria to evaluate the solution variants can be:

- Global network cost
- Core node, fiber and port costs
- Link utilization
- Lightpath delay, and total fiber kilometers

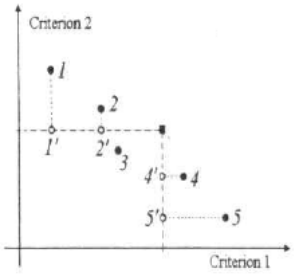


Figure 10.9: RGM selection procedure (example, maximization). Source: [52].

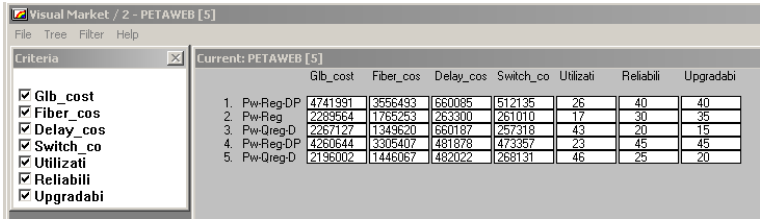


Figure 10.10: Petaweb trade-off VM/2 dataset.

- Reliability and upgradeability level

The different Petaweb structure alternatives are listed in the following:

- Petaweb regular (PwRg)
- Petaweb regular with TDM/WDM (PwRgTDM)
- Petaweb regular with DPP (PwRegDPP)
- Petaweb regular with DPP and TDM/WDM (PwRegDPPTDM)
- Petaweb quasi-regular (PwQg)
- Petaweb quasi-regular with TDM/WDM (PwQgTDM)
- Petaweb quasi-regular with DPP (PwQgDPP)
- Petaweb quasi-regular with DPP and TDM/WDM (PwQgDPPTDM)

10.3.3 RGM/IDM simulation with Visual Market/2

The previous simulations allowed us to fill with multiple criteria the variants of a RGM/IDM dataset, available in [198], reported in Fig. 10.10. Given the limited number of criteria of the shareware version of VM/2 [198], we did not simulate for more than 5 criteria, excluding the total fiber km, the lightpath delay and the port cost criteria (even because they can be inferred from the others). For the sake of clarity, we also excluded the cases PwQg, PwRgTDM, PwQgTDM.

In Fig. 10.11 the CEPH of the decision problem is displayed. The two axes index those that can be considered as the most important criteria, i.e., the resource utilization (to be maximized in network planning) and the global network cost (to be minimized). The reliability can be considered as the third criteria for importance, and given its low granularity it is displayed with superposed slices of different colors. The latter two

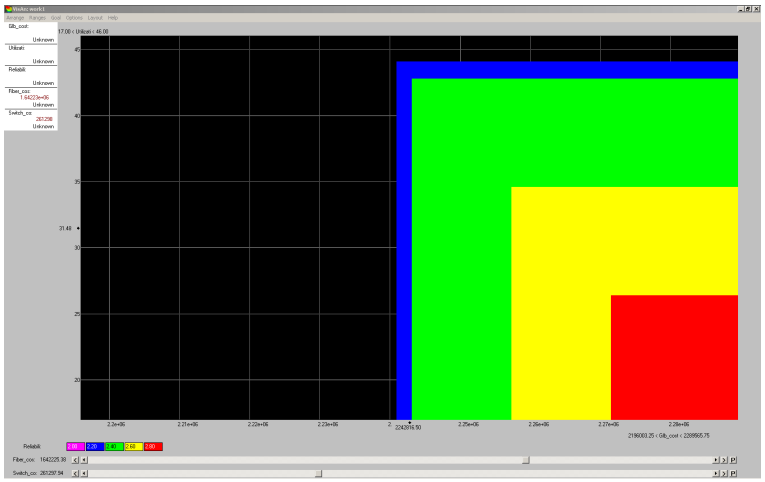
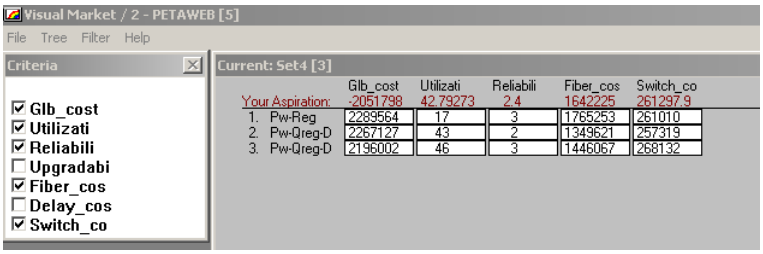


Figure 10.11: Color decision map for the five criteria. Reliability as third criterion, fiber cost as fourth criterion.

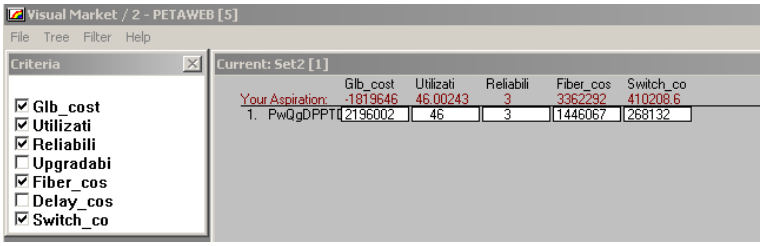
criteria can be explored with the two scroll bars in the bottom. In this expressive way it is thus possible to explore *visually* all the variant set. It is intuitive to notice that, e.g., an increase in the reliability passes through an increase of the global cost and a decrease of the network utilization. What the IDM gives especially to the DM are the trade-offs between all the criteria and, via the VM/2 software, all the possible trade-offs can be explored.

The DM then chooses a reasonable goal and a list of several selected alternatives is finally considered and suggested. In Fig. 10.12a we have for example the result of the application of RGM/IDM. The reasonable goal ('Your aspiration') is not feasible. The result is an ordered list of variants close to the goal, which corresponds to only those close variants whose modified points are not Pareto-inferior to each other are finally listed to the decision maker. Hence, for this example, the decision maker might finally opt for the classical regular Petaweb variant. Or, in Fig. 10.12b, with a different goal, we have that the list of variant is restricted to a single one, the quasi-regular Petaweb with DPP and TDM one.

In Fig. 10.13, we dispose of another point of view on the same decision problem. This time, the third criterion is the upgradeability. Which, generally speaking, is higher for regular structure, higher if TDM is not employed, lower if the DPP strategy is used. The fourth criterion is the delay cost instead of the fiber cost, which have approximately a linear relation. We verified a higher complexity in the computation of the CEPH for this configuration of criteria. Indeed, as displayed in Fig. 10.13, the CEPH is less straightforward than before. The CEPH can be explored so as to appreciate the effect of the criteria; we captured a video, available at [197], which displays the dynamics of the CEPH tuning the latter two criteria.



(a)



(b)

Figure 10.12: List of alternatives that are in line with the goal.

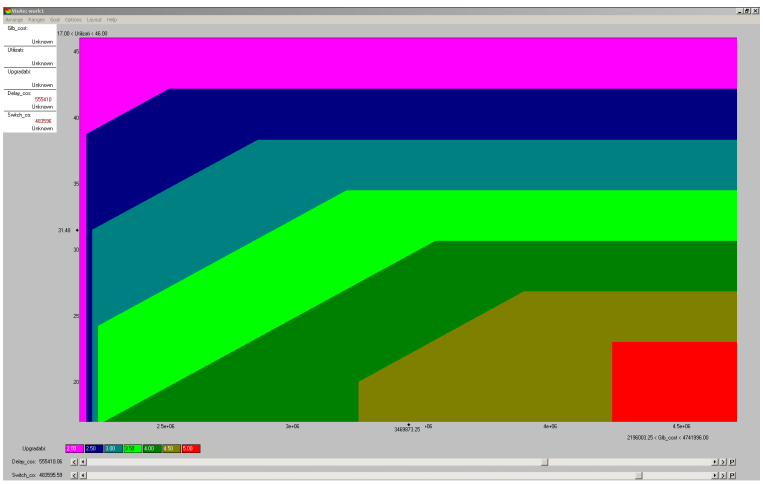


Figure 10.13: Color decision map for the five criteria. Upgradeability as third criterion, delay cost as fourth criterion.

Chapter 11

Conclusion

The dissertation embraces various solutions accounting for collaboration among providers in current and next-generation Internet service delivery and management.

The increasing level of collaboration shall not be interpreted as an increasing level of utopia, and thus of impossibility to implement the solution. At each stage of collaboration we took advantage of 'tools' that adequately modeled a strategic interaction between providers. Namely, in the framework presented in Chapter 3, the selfishness and independent interests of the Autonomous Systems are modeled as an imperative requirement; the coordinated routing solution relies on the concept of game equilibrium that can give a solution quite far from the bilateral optimum: the paid price is the price of meeting the requirement. The provider alliance functional architecture of Chapters 4 and 5 emulates somehow the current practice in bids or financial markets, with still a competition among providers within the common technical platform that allows the inter-working. Moreover, the income distribution scheme of Chapter 6 supposes a quite high level of collaboration and trust among providers, assuming that some providers that would benefit from the status quo would finally agree in changing the business model and in being less rewarded because of their limited transit contribution. This gap representing the pragmatic price needed to correctly take into account the independency and 'natural' tendency to secede or bargain each provider will always have in a strategic interaction framework. With a longer term standpoint, the interconnection Petaweb architecture object of Chapters 7-10 may be implemented as a joint IXP physical infrastructure, slightly more expensive than the current alternatives, which providers might however prefer to implement in order to have simplified and efficient cross-provider traffic engineering.

Each proposition brings novelties to the corresponding research area. The BGP coordination enhancements of Chapter 3 open an interesting path toward game-theoretic yet simple management of critical flows across peering settlements. Chapters 4 and 5 define a novel complete routing and provisioning architecture for multi-provider connection-oriented services that readily integrates recent advances in standardization bodies. Chapter 6 also opens a new path toward the definition of economically feasible business models for the future Internet. Chapters 7-10 report the first planning solutions for the Petaweb design at the state of the art. Chapter 10 brings, moreover, the original proposition of using interactive decision maps for operational network planning.

The complexity of the subject needed to be treated with the different mathemat-

ical tools presented and applied for this dissertation. Graph theory, cooperative and non-cooperative game theory, operations research and multi-objective decision theory are nowadays essential yet conceptually heavy bricks in the formation of networking researchers, and hopefully in the operations of future telecommunication engineers and experts. Our will is that the reader of this thesis would find, in at least one of the followed directions, interesting and useful ideas in the same time, ideally with some ideas to complete and improve these works.

Many open issues can come to mind. For example, the extended peering coordination framework even if defined and discussed, may need additional refinements in order to be considered as really interesting in a multi-carrier interaction scenario. Or, the provider alliance concept implicitly assumes that it is or will act in a competing multi-alliance framework. The alliance formation process is a problem that shall thus be formalized to arise the alliance join and secession conditions, and the inter-provider and inter-alliance interconnection policies. We initiated some works in this direction, falling in game theory concepts that are still not agreed and under discussion in the related community. Finally, cooperative cross-provider static resource reservation mechanisms shall be defined upon an pro-active online algorithmic usage of cooperative game theory concepts, instead than simply adopt them as income distribution schemes; we are currently working toward convergent and efficient solutions in this sense.

An attentive reader may be puzzled about how the solutions claimed in this thesis would relate with the network neutrality debate. As already mentioned, the coordinated routing framework may not be used for best-effort traffic but for implementing a sort of highway for selected traffic in the Internet. Similarly, the provider alliance framework may be pointed as against the network neutrality, even if it is conceived for novel inter-provider services and not necessarily Internet services in the common current sense. We believe that this aspect represents the very critical issue in the definition of a fair and economically feasible network neutrality policies. There is the need to regulate the matter in a way that the enabling of connection-oriented inter-provider network services is allowed. Moreover, the offering of a level play field to application providers (willing to access the inter-provider QoS) by network providers (possibly acting also as application providers) shall be fairly regulated, and so the QoS/TE connectivity between the end-user and the application providers connected by the same network provider. An interesting approach proposing a level play field regulated by application-provider/network-provider open interfaces is presented, for instance, in [160]. This is an interdisciplinary research subject that is likely to arise once the technology will be ready and the demand for Inter-provider connection-oriented services will explode.

This thesis has been structured trying to focus the attention on the part different type of readers would be interested in, introducing the contents synthetically without unnecessary large introductions, organizing two different research fields and methodologies in two parts, guiding the reader with brief introductions and summaries around each subject, and leaving some complementary parts in a large appendix. Finally, other less relevant aspects are not included, but are however presented in the related publications.

Last words are needed to further acknowledge the researchers, and their respective teams, whose external collaboration allowed to pursue the high value inter-disciplinariety

of the dissertation. In order of content exposition,
Prof. Fioravante Patrone from the University of Genoa for his inspiring suggestions on the game-theoretic aspects of the solutions presented in Chapters 3 and 6;
Dr. Ramon Casellas, Ing. Richard Douville and Ing. Jean-Louis Le Roux from Centre Tecnològic de Telecomunicacions de Catalunya, Alcatel-Lucent Bell Labs and Orange France Telecom for their industrial experience and pragmatic point of view about the propositions of Chapters 4 and 5;
Prof. Alberto Ceselli from the University of Milan and Prof. Federico Malucelli from Politecnico di Milano for their reliable review and contribution on quasi-regular Petaweb network design optimization presented in Chapters 8 and 9;
Prof. Michal Pióro and Mariusz Mycek from Warsaw University of Technology and Prof. Brunilde Sansò from École Polytechnique de Montréal for their precious and rare expertise in both operations research and telecommunications, for Chapter 6 and Chapters 8 - 10, respectively;
Prof. Alexander Lotov from the Russian Academy of Science for his useful course on multi-objective decision theory that allowed the decision-making analysis of Chapter 10.

Bibliography

Related publications

International Journals with peer review

[1] S. Secci, J.-L. Rougier, A. Pattavina, "AS-level source routing for multi-provider connection-oriented services", to appear in *Computer Networks*, 2011.

[2] S. Secci, J.-L. Rougier, A. Pattavina, F. Patrone, G. Maier, "Multi-Exit Discriminator Game for BGP routing coordination", to appear in *Telecommunication Systems*, 2011.

[3] S. Secci, B. Sansò, "Survivability and Reliability of a Composite-Star Transport Network with Disconnected Core Switches", *Telecommunication Systems*, Vol. 46, No. 1, Jan. 2011.

[4] A. Reinert, B. Sansò, S. Secci, "Design Optimization of the Petaweb Architecture", *IEEE/ACM Transactions on Networking*, Vol. 17, No. 1, pp: 332-345, Feb. 2009.

[5] R. Douville, J.-L. Le Roux, J.-L. Rougier, S. Secci, "A Service Plane over the PCE Architecture for Automatic Multi-Domain Connection-Oriented Services", *IEEE Communications Magazine*, Vol. 46, No. 6, June 2008.

[6] S. Secci, M. Tornatore, A. Pattavina, "Optimal Design for Survivable Backbones with End-to-End and Subpath Wavebanding", *OSA Journal of Optical Networking*, Vol. 6, No. 1, pp: 1-12, Dec. 2006.

International conferences with peer review

[7] S. Secci, J.-L. Rougier, A. Pattavina, F. Patrone, G. Maier, "PEMP: Peering Equilibrium MultiPath routing", in *Proc. of 2009 IEEE Global Communications Conference (GLOBECOM 2009)*, 30 Nov. - 4 Dec. 2009, Honolulu, USA.

[8] M. Mycek, S. Secci, M. Pióro, J.-L. Rougier, A. Tomaszewski, A. Pattavina, "Cooperative Multi-Provider Routing Optimization and Income Distribution", in *Proc. of 2009 7th Int. Workshop on the Design of Reliable Communication Networks (DRCN 2009)*, 25-28 Oct. 2009, Washington, USA.

[9] S. Secci, J.-L. Rougier, A. Pattavina, F. Patrone, G. Maier, "ClubMED: Coordinated Multi-Exit Discriminator Strategies for Peering Carriers", in *Proc. of 2009 5th Euro-NGI Conference on Next Generation Internet Networks (NGI 2009)*, Aveiro, Portugal, 1-3 July 2009. **Best Paper Award**.

[10] A.P. Bianzino, J.-L. Rougier, S. Secci, R. Casellas, R. Martinez, R. Munoz, N. Djarallah, R. Douville, H. Pouyllau, "Testbed Implementation of Control Plane Extensions for Inter-Carrier GMPLS LSP provisioning", in *Proc. of 2009 5th Int. Conference on Testbeds and Research Infrastructures for the Development of Networks and Communities (TRIDENTCOM 2009)*, Washington, USA, 6-8 April 2009.

[11] S. Secci, J.-L. Rougier, A. Pattavina, "AS Tree Selection for Inter-Domain Multipoint MPLS Tunnels", in *Proc. of 2008 IEEE International Conference on Communications (ICC 2008)*, Beijing, China, 19-23 May 2008.

[12] S. Secci, J.-L. Rougier, A. Pattavina, "Comparison of Quasi-Regular Composite-Star and Multi-Hop Structures for Core Networks", in *Proc. of 2008 IEEE International Conference on High Performance Switching and Routing (HPSR 2008)*, Shanghai, China, 15-17 May 2008.

[13] S. Secci, J.-L. Rougier, A. Pattavina, "On the Selection of Optimal Diverse AS-Paths for Inter-Domain IP/(G)MPLS Tunnel Provisioning", in *Proc. of IEEE 4th International Telecommunication Networking WorkShop on QoS in Multiservice IP Networks (IT-NEWS 2008 - QoS-IP 2008)*, Venezia, Italy, 13-15 Feb. 2008.

[14] S. Secci, J.-L. Rougier, A. Pattavina, "Constrained Steiner Problem with Directional Metrics", in *Proc. of EuroFGI Workshop on IP QoS and Traffic Control 2007*, Lisbone, Portugal, 7-9 Dec. 2007.

[15] S. Secci, A. Ceselli, F. Malucelli, A. Pattavina, B. Sansò, "Direct Optimal Design of a Quasi-Regular Composite-Star Core Network", in *Proc. of 2007 6th International Workshop on the Design of Reliable Communication Networks (DRCN 2007)*, La Rochelle, France, 07-10 Oct. 2007.

[16] S. Secci, B. Sanso, "Upgrade of a Composite-Star Optical Network", in *Proc. of IEEE/IFIP NTMS 2007*, 2-4 May 2007, appeared as book chapter in *New Technology Mobility and Security*, pp. 281-296, Springer Netherlands, 2007.

[17] S. Secci, B. Sansò, "Design and Dimensioning of a Novel Composite-star WDM Network with TDM Channel Partitioning", in *Proc. of 2006 IEEE 3rd International Conference on Broadband Communications, Networks and Systems (BROADNETS 2006)*, San José, CA, USA, 1-5 Oct. 2006.

[18] S. Secci, B. Sansò, "Optimization of a Dedicated Path Protected PetaWeb Architecture", in *Proc. of 2006 IFIP/INFORMS Networking and Electronic Commerce Research Conference (NAEC 2006)*, Riva del Garda, Italie, 19-22 Oct. 2006.

Other workshops and national conferences

[19] S. Secci, J.-L. Rougier, M. Mycek, M. Pioro, "A Shapley Value Scheme to Incent Provider Collaboration in the Future Internet", in *Proc. of 2010 24th European Conference on Operational Research (EURO XXIV)*, Lisbon, Portugal, 11-14 July, 2010.

[20] S. Secci, F. Patrone, "Potential Games for Internet Routing Coordination", in *Proc. of 2010 Spain - Italy - Netherlands 6th Meeting on Game Theory (SING 6)*, Palermo, Italy, 7-9 July, 2010.

[21] S. Secci, J.-L. Rougier, A. Pattavina, F. Patrone, G. Maier, "Peering Games for Critical Internet Flows", in *Proc. of 2009 Euro-NF Fifth International Workshop on Traffic Management and Traffic Engineering for the Future Internet*, 7-8 Dec. 2009, Paris, France.

[22] S. Secci, J.-L. Rougier, A. Pattavina, M. Mycek, M. Pióro, A. Tomaszewski, "Connection-oriented Service Management in Provider Alliances: a Shapley Value Perspective", in *Proc. of 2009 Euro-NF Fifth International Workshop on Traffic Management and Traffic Engineering for the Future Internet*, 7-8 Dec. 2009, Paris, France.

[23] M. Mycek, S. Secci, M. Pióro, J.-L. Rougier, A. Pattavina, "A Shapley value-based Incentive Scheme for Cooperative Multi-Provider Traffic Engineering", in *Proc. of 2009 16th Polish Teletraffic Symposium (PTS 2009)*, Lódz, Poland, 24-25 September, 2009.

[24] S. Secci, J.-L. Rougier, A. Pattavina, "Routage inter-domaine en mode connecté", in *Proc. of RESCOM 2008*, Saint-Jean-Cap-Ferrat, France, Jun. 2008.

Submitted

[25] S. Secci, J.-L. Rougier, A. Pattavina, F. Patrone, G. Maier, "Peering Equilibrium MultiPath Routing: a game theory framework for Internet peering settlements", submitted.

[26] E. Elena, J.-L. Rougier, S. Secci, "Characterisation of AS-level Path Deviations and Multipath in Internet Routing", submitted.

Thesis

[27] S. Secci, "Game Theory for Internetworking", minor Ph.D. thesis (due to the minor research activity of the doctoral program at Politecnico di Milano), Oct. 2008. Advisor: Fioravante Patrone. Published at the library of the Dep. of Electronics and Information, Politecnico di Milano, Nr: 2008.21.

[28] S. Secci, "Design and optimisation of a novel composite-star TDM/WDM network architecture : the Petaweb", Master thesis, Oct. 2005. Advisors: Brunilde Sansò, Achille Pattavina. Published at the Central Library of Politecnico di Milano, 2005. Inv.: 0TN900009185, Coll. T.D.L. 12427.

[29] S. Secci, "Multi-Provider Service and Transport Architectures", Ph.D dissertation. Advisors: Jean-Louis Rougier, Achille Pattavina. Télécom ParisTech and Politecnico di Milano.

Research reports

[30] S. Secci, J.-L. Rougier, A. Pattavina, G. Maier, M. Marinoni, E. Elena, "Detection of BGP route deflections across top-tier interconnections", ENST res. report nb 9287-2009. `http://www.tsi.enst.fr/publications/enst/techreport-2009-9287.pdf`

[31] S. Secci, J.-L. Rougier, A. Pattavina, F. Patrone, G. Maier, "Internet Extended Peering: a non-cooperative game theoretic framework", Telecom ParisTech research report, `http://www.tsi.enst.fr/publications/enst/techreport-2009-9286.pdf`

[32] S. Secci, "Broadband for All via Industrial Cluster in Fast Developing European Regions", D.2.4 deliverable for the INTERREG III.A project ACInD ("Adriatic Cooperation for Industrial Development"), Jan. 2008.

[33] S. Secci, B. Sansò, "A Survivable Composite-Star Optical Transport Network with Disconnected Core Switches", *Les Cahiers du GERAD*, G-2007-28, April 2007.

[34] S. Secci, B. Sansò, "Upgrade of a Composite-Star Optical Network", *Les Cahiers du GERAD*, G-2007-15, Mars 2007.

[35] S. Secci, J.-L. Rougier, "A constrained Internet graph for interdomain circuit transit", ENST research report, SR:IGR-06, 2006.

[36] S. Secci, B. Sansò, "Design and Dimensioning of a Novel Composite-Star WDM Network with TDM Channel Partitioning", *Les Cahiers du GERAD*, G-2006-35, May 2006.

[37] A. Reinert, B. Sansò, S. Secci, "Design Optimization of the Petaweb Architecture", *Les Cahiers du GERAD*, G-2006-86, Dec. 2006.

References

Game theory and operations research

[38] A.E. Roth, *The Shapley value, essays in honor of Lloyd S. Shapley*, Cambridge Univ. Press (1988).

[39] R.B. Myerson, *Game Theory: Analysis of Conflict*, Harvard Univ. Press, 1991

[40] F. Patrone, *Decisori (razionali) interagenti*, Pisa University Press (2006).

[41] M. Osborne, *An Introduction to Game Theory*, Oxford University Press, 2004.

[42] M. Osborne, A. Rubinstein, *A course in Game Theory*, MIT Press, Cambridge (MA, USA), 1994.

[43] R.-J. Aumann, M. Maschler, "Game theoretic analysis of a bankruptcy problem from the Talmud", *Journal of Economic Theory*, Vol. 36, No. 2 (1985).

[44] D. Monderer, L.S. Shapley, "Potential Games", *Games and Economic Behavior*, Vol. 14, No. 1 (1996).

[45] S. Moretti, F. Patrone, "Transversality of the Shapley value", *TOP*, Vol. 16, No. 1 (2008).

[46] F. Patrone, "Giochi con Potenziale", *Lettera Matematica Pristem*, Vol. 69, newsletter "Il decimo appuntamento con la Teoria dei Giochi"(2008).

[47] S. Chopra1, M.R. Rao, "The Steiner tree problem I: Formulations, compositions and extension of facets", *Journal of Mathematical Programming*, Vol. 64 (1994).

[48] A. Koster, S.P.M. Van Hoesel, A.W.J. Kolen, "The Partial Constraint Satisfaction Problem: Facets and Lifting Theorems", *Operations Research Letters*, Vol. 23 (1998).

[49] L. Cooper, "The transportation-location problem", *Operations Research*, Vol. 20 (1972).

[50] N. Menakerman, R. Rom, "Bin Packing with Item Fragmentation", in *Proc. of the* 7^{th} *Int. Workshop on Algorithms and Data Structures* (2001)

[51] A. V. Lotov, V. A. Bushenkov, G. K. Kamenev, *Interactive Decision Maps*, Kluwer Academic Pusblisher (2004).

[52] M. Engquist, "A successive shortest path algorithm for the assignment problem", *INFOR*, Vol. 20 (1982).

[53] M.A. Forbes, J.N. Holt, P.J. Kilby, A.M. Watts, "A matching algorithm with application to bus operations", *Australian Journal of Combinatorics*, Vol. 4 (1991).

[54] R. Jonker, A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems", *Computing*, Vol. 38 (1986).

[55] J.G. Klincewicz, "Hub Location in backbone/tributary Network Design: a Review", *Location Science*, Vol. 6 (1998).

[56] M.L. Balinski, "On Finding Integer Solutions to Linear Programs", *Mathematica* (1964).

[57] M. Rönnqvist, S. Tralgantalerngsak, J. Holt, "A Repeated Matching Heuristic for the Single-source Capacitated Facility Location Problem", *European Journal of Operational Research*, Vol. 116 (1999).

[58] M. Rönnqvist, K. Holmberg, D. Yuan, "An Exact Algorithm for the Capacitated Facility Location Problems with Single Sourcing", *European Journal of Operational Research*, Vol. 113 (1999).

[59] R. Jonker, A. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems", *Computing*, Vol. 38 (1986).

[60] B. Sansò, S. Soriano, *Telecommunication Network Planning*, Centre for Research on Transportation, 25^{th} Anniversary Series.

Networking

[61] M.A. Brown, C. Hepner, A.C. Popescu, "Internet Captivity and the De-peering Menace", in *Proc. of NANOG 45* (2009).

[62] L. Gao, J. Rexford, "Stable Internet routing without global coordination", *IEEE/ACM Transactions on Networking*, Vol. 9, No. 6 (2001).

[63] N. Feamster, J. Borkenhagen, J. Rexford, "Guidelines for interdomain traffic engineering", *ACM SIGCOMM Computer Communications Review*, Vol. 33, No. 5.

[64] T.G. Griffin, G. Wilfong, "On the correctness of IBGP configuration", in *Proc. of SIGCOMM 2002*.

[65] J. Rexford, J. Wang, Z. Xiao, Y. Zhang, "BGP routing stability of popular destinations", in *Proc. of IMW 2002*.

[66] S. Uhlig, V. Magnin, O. Bonaventure, C. Rapier, L. Deri, "Implications of the topological properties of Internet traffic on traffic engineering", in *Proc. of 19th ACM Symposium on Applied Computing* (2004).

[67] F. Viger, B. Augustin, X. Cuvellier, C. Magnien, M. Latapy, T. Friedman, R. Teixeira, "Detection, understanding, and prevention of traceroute measurement artifacts", *Computer Networks*, Vol. 52, No. 5 (2008).

[68] M. Crovella, B. Krishnamurthy, *Internet Measurement*, Wiley (2006).

[69] M. Yannuzzi, X. Masip-Bruin, O. Bonaventure, "Open issues in interdomain routing: a survey", *IEEE Network*, Vol. 19, No. 6 (2005).

[70] B. Quoitin, S. Uhlig, C. Pelsser, L. Swinnen, O. Bonaventure, "Interdomain traffic engineering with BGP", *IEEE Communications Magazine*, Vol. 41, No. 5 (2003).

[71] R. Mahajan, D. Wetherall, T. Anderson, "Towards coordinated interdomain traffic engineering", in *Proc. of HotNets-III 2007*.

[72] R. Mahajan, D. Wetherall, T. Anderson, "Negotiation-based routing between neighboring ISPs", in *Proc. of the 2nd conference on Symposium on Networked Systems Design & Implementation* (2005).

[73] W. Xu, J. Rexford, "MIRO: Multi-path interdomain routing", in *Proc. of ACM SIGCOMM 2006*.

[74] B. Quoitin, O. Bonaventure, "A Cooperative Approach to Interdomain Traffic Engineering", in *Proc. of NGI 2005*.

[75] A. Fonte, E. Monteiro, M. Yannuzzi, X. Masip-Bruin, J. Domingo-Pascual, "A Framework for Cooperative Interdomain QoS Routing", *Springer Series: IFIP*, Vol. 196 (2006).

[76] M. Yannuzzi, X. Masip-Bruin, G. Fabrego, S. Sanchez-Lopez, A. Sprintson, A. Orda, "Toward a new route control model for multidomain optical networks", *IEEE Communications Magazine*, Vol. 46, No. 6 (2008).

[77] M. Yannuzzi et al., "A proposal for inter-domain QoS Routing based on distributed overlay entities and QBGP", in *Proc. of 2004 WQoSR*.

[78] X. Masip-Bruin et al., "The EuQoS system: a solution for QoS routing in heterogeneous networks", *IEEE Communications Magazine*, Vol. 45, No. 2 (2007).

[79] D. Griffin et al., "Interdomain routing through QoS-class planes", *IEEE Communications Magazine*, Vol. 45, No. 2 (2007).

[80] M.P. Howarth et al., "Provisioning for inter-domain quality of service: the MESCAL approach", *IEEE Communications Magazine*, Vol. 43, No. 6 (2005).

[81] G. Shrimali, A. Akella, A. Mutapcic, "Cooperative Inter-Domain Traffic Engineering Using Nash Bargaining and Decomposition", in *Proc. of INFOCOM 2007*.

[82] T. Lehman, J. Sobieski, B. Jabbari, "DRAGON: a framework for service provisioning in heterogeneous grid networks", *IEEE Communications Magazine*, Vol. 44, No. 3(2003).

[83] C. Pelsser, O. Bonaventure, "Path selection techniques to establish constrained interdomain MPLS LSPs", in *Proc. of Networking 2006*.

[84] A. Van Moorsel, "Metrics for the internet age: Quality of experience and quality of business", in *Proc of* 5^th *International Workshop on Performability Modeling of Computer and Communication Systems* (2001).

[85] R.T.B. Ma, D. Chiu, J. C.S. Lui, V. Misra, D. Rubenstein, "Internet Economics: The use of Shapley value for ISP settlement", in *Proc. of 2007 ACM Conference on Emerging Network EXperiment and Technology (CoNEXT 2007)*.

[86] R. Ma, D.M. Chiu, J.C.S. Lui, V. Misra, D. Rubenstein, "Interconnecting Eyeballs to Content: A Shapley Value Perspective on ISP Peering and Settlement", in *Proc. of ACM SIGCOMM 2008*.

[87] P. Faratin, D. Clark, P. Gilmore, S. Bauer, A. Berger, W. Lehr, "Complexity of Internet Interconnections: Technology, Incentives and Implications for Policy", in *Proc. of TPRC 2007*.

[88] R. Teixeira, A. Shaikh, T.G. Griffin, J. Rexford, "Impact of Hot-Potato Routing Changes in IP Networks", *IEEE/ACM Trans. on Networking*, Vol. 16, No. 6 (2008).

[89] B. Huffaker et al., "Distance Metrics in the Internet", in *Proc. of 2002 International Telecommunications Symposium (ITS 2002)*.

[90] S. Balon, G. Leduc, "Combined Intra- and inter-domain traffic engineering using hot-potato aware link weights optimization", Arxiv preprint arXiv:0803.2824, 2008.

[91] T.G. Griffin, G. Wilfong, "Analysis of the MED oscillation problem with BGP", in *Proc. of ICNP 2002*.

[92] M. Brown, E. Zmijewski, A. Popescu, "Peering Wars: Lessons Learned from the Cogent-Telia De-peering", NANOG research report, http://www.nanog.org.

[93] S. Agarwal, A. Nucci, S. Bhattacharyya, "Controlling Hot Potatoes in Intradomain Traffic Engineering", SPRINT RR04-ATL-070677, 2004.

[94] S. Jalabi, D. McPherson, *Internet routing architectures*, Cisco Press, 2^nd edition.

[95] S. Uhlig, O. Bonaventure, "Designing BGP-based outbound traffic engineering techniques for stub ASes", *ACM SIGCOMM Computer Communication Review*, Vol. 34, No. 5(2004).

[96] P. He, G.V. Bochmann, "OSN-IX: A Novel Internet eXchange (IX) Architecture based on Overlaid-Star Networks", in *Proc of. 2008 Next Generation Internet Networks (NGI 2008)*.

[97] B. Choi, S. Moon, Z. Zhang, K. Papagiannaki, C. Diot, "Analysis of Point-To-Point Packet Delay in an Operational Network", in *Proc. of INFOCOM 2004*.

[98] F. Yegenoglu, E. Sherk, "Network characterization using constraint-based definitions of capacity, utilization, and efficiency", *IEEE Communications Magazine* Vol. 43, No. 9 (2005).

[99] R. Feuerstein, "Interconnecting the Cyberinfrastructure", in *Proc. of Cyber-infrastructure workshop* (2005).

[100] W.D. Grover, *Mesh-Based Survivable Networks*, Prentice Hall (2004).

[101] R. Teixeira, T. Griffin, M.G.C. Resende, J. Rexford, "TIE Breaking: Tunable Interdomain Egress Selection", in *Proc. of CoNEXT 2005*.

[102] F. Larroca, J.-L. Rougier, "Routing Games for Traffic Engineering", in *Proc. of IEEE ICC 2009*.

[103] Y. Wang, I. Avramopoulos, J. Rexford, "Design for configurability: re-thinking interdomain routing policies from the ground up", *Journal on Selected Areas in Communications*, Vol. 27, No. 3 (2009).

[104] P. Casas, L. Fillatre, S. Vaton, "Multi hour robust routing and fast load change detection for traffic engineering", in *Proc. of IEEE ICC 2008.*

[105] J. Lepropre, S. Balon, G. Leduc, "Totem: a toolbox for traffic engineering methods", in *Proc. of INFOCOM 2006.*

[106] S. Uhlig, B. Quoitin, S. Balon, J. Lepropre, "Providing public intradomain traffic matrices to the research community", *Computer Communication Review*, Vol. 36, No. 1 (2006).

[107] A. Farrel, I. Bryskin, *GMPLS Architecture and Applications*, Morgan Kaufmann (2006).

[108] H.F. Salama, D.S. Reeves, Y. Viniotis, "Evaluation of multicast routing algorithms for real-time communication on high-speed networks", *IEEE J. on Sel. Areas in Communications*, Vol. 15, No. 3 (1997).

[109] Q. Zhu, M. Parsa, J.J. Garcia-Luna-Aceves, "A source-based algorithm for delay-constrained minimum-cost multicasting", in *Proc. of INFOCOM 1995.*

[110] V.P. Kompella, J.C. Pasquale, G.C. Polyzos, "Multicast routing for multimedia communication", *IEEE/ACM Transactions on Networking*, Vol. 1, No. 3 (1993).

[111] A. Sprintson, M. Yannuzzi, A. Orda, X. Masip-Bruin, "Reliable Routing with QoS Guarantees for Multi-Domain IP/MPLS Networks", in *Proc. of INFOCOM 2007.*

[112] M. Yannuzzi, X. Masip-Bruin, S. Sanchez, J. Domingo-Pascual, A. Orda, A. Sprintson, "On the challenges of establishing disjoint QoS IP/MPLS paths across multiple domains", *IEEE Communications Magazine*, Vol. 44, No. 12 (2006).

[113] D. Di Sorte, G. Reali, "Minimum price inter-domain routing algorithm", *IEEE Communications Letters* Vol. 6, No. 4 (2002).

[114] G Liu, KG Ramakrishnan, "A*Prune: an algorithm for finding K shortest paths subject to multiple constraints", in *Proc. of INFOCOM 2001.*

[115] R. W. Floyd, "Algorithm 97: Shortest Path", *Communications of the ACM*, Vol. 5, No. 6 (1962).

[116] R. Vickers, M. Beshai, "Petaweb architecture", presented at *Networks 2000 - Toward Natural Network:* 9^{th} *International Telecommunication Network Planning Symposium*, Toronto (2000).

[117] F. J. Blouin, S. Yazid, B. Bou-Diab., "Emulation of a vast adaptative network", presented at *Networks 2000 - Toward Natural Network:* 9^{th} *International Telecommunication Network Planning Symposium*, Toronto (2000).

[118] J. P. G. Sterbenz, L. Chapin, R. Krishnan, "Routing issues in interconnecting IP Networks with the Petaweb", presented at *Networks 2000 - Toward Natural Network:* 9^{th} *International Telecommunication Network Planning Symposium*, Toronto (2000).

[119] F. Yegenoglu, E. Sherk, "Network characterization using constraint-based definitions of capacity, utilization, and efficiency", *Communications Magazine*, Vol. 43, No. 9 (2005).

[120] S. Chamberland, B. Sansò, "Update of two-level networks with modular switches", in *Proc. of ITC 1999.*

[121] L. Gao, "On inferring autonomous system relationships in the Internet", *IEEE/ACM Trans. On Networking*, Vol. 9, No. 6 (2001)

[122] M. Latapy, C. Magnien, F. Ouédraogo, "A Radar for the Internet", in *Proc. of ADN 2008.*

[123] B. Donnet, T. Friedman, "Internet topology discovery: a survey", *IEEE Communications Surveys and Tutorials*, Vol. 9, No. 4 (2007).

[124] C. Lobavitz, G.R. Malan, F. Jahanian, "Origins of Internet routing instability", in *Proc. IEEE INFOCOM 2009*

[125] M. Pióro, A. Tomaszewski et al., "A Subgradient Optimization Approach to Inter-domain Routing in IP/MPLS Networks", in *Proc. of Networking 2007*.

[126] A. Tomaszewski, M. Pióro, M. Mycek, "A Distributed Scheme for Inter-Domain Routing Optimization", in *Proc. of DRCN 2007*.

[127] M. Pióro, A. Tomaszewski, M. Mycek, "Distributed Inter-Domain Link Capacity Optimization for Inter-Domain IP/MPLS Routing", in *Proc. of Globecom 2007*.

[128] M. Pióro, A. Tomaszewski, M. Mycek, "A distributed scheme for optimization of inter-domain routing between collaborating domains", *Annales des Télécommunications; Special Issue on Inter-Domain Routing and QoS over Heterogeneous Networks* (2008).

[129] M. Pióro, A. Tomaszewski, M. Mycek, "A Scheme for Cooperatirve Optimization of Flows on Inter-Domain Links", in *Proc. of Polish-German Teletraffic Symposium* (2008).

[130] M. Pióro, D. Medhi, *Routing, Flow and Capacity Design in Communication and Computer Networks*, Morgan Kaufman (2004).

[131] N. Feamster, J. Borkenhagen, J. Rexford, "Guidelines for Interdomain Traffic Engineering", *ACM SIGCOM Computer Communications Review*, Vol. 33, No. 5 (2003).

[132] I. Nakagawa, H. Esaki, K. Nagami, "A design of a next generation IX using MPLS technology", in *Proc. of SAINT 2002*.

[133] J. Xiao, R. Boutaba, "QoS-aware service composition and adaptation in autonomic communication", *IEEE Journal on Selected Areas in Communications*, Vol. 23(2005).

[134] A. Orda, A. Sprintson, "Precomputation schemes for QoS routing", *IEEE/ACM Transactions on Networking*, Vol. 11, No. 4 (2003).

[135] Jiayue He, J. Rexford, "Towards Internet-wide multipath routing", in *IEEE Network magazine* Vol. 22, No. 2 (2008).

[136] "Configuring BGP to Select Multiple BGP Paths", JUNOS Documentation.

[137] "BGP Best Path Selection Algorithm", Cisco Documentation.

[138] J. Dégila, B. Sansò, "Topological design optimisation of a yottabit-per-second lattice network", *Journal on Selected Areas in Communications*, Vol. 22, No. 9 (2004).

[139] A. Reinert, *Conception et optimisation d'un réseau optique de prochaine génération: le Petaweb*, Master thesis, Ecole Polytechnique de Montréal (2004).

[140] D. Arci, G. Maier, D. Petecchi, A. Pattavina, M. Tornatore, "Availability models for protection techniques in WDM networks", in *Proc. of DRCN 2003*.

[141] X. Cao, V. Anand, C. Qiao, "Waveband switching in optical networks", *IEEE Communications Magazine* Vol. 41, No. 4 (2003).

[142] P.-H. Ho, H.T. Mouftah, J. Wu, "A scalable design of multigranularity optical cross-connects for the next-generation optical Internet", *Journal on Selected Areas on Communications*, Vol. 21, No. 7 (2003).

[143] A. Kolarov, T. Wang, B. Sengupta, M. Cvijetic, "Impact of waveband switching on dimensioning multi-granular hybrid optical networks", in *Proc. of ONDM 2005*,

[144] M. Lee, J. Yu, Y. Kim, C. Kang, J. Park, "Design of hierarchical crossconnect WDM networks employing a two-stage multiplexing scheme of waveband and wavelength", *Journal on Selected Areas on Communications*, Vol. 20, No. 1 (2002)

[145] Y. Shun, O. Canhui, B. Mukherjee, "Design of hybrid optical networks with waveband and electrical TDM switching", in *Proc. of Globecomm 2003*.

[146] G. Maier, A. Pattavina, M. Tornatore, "WDM network optimization by ILP models based on flows aggregation", *IEEE/ACM Transactions on Networking*, Vol. 15, No. 3 (2007).

[147] P. Gambini et al., "Transparent optical packet switching: Network architecture and demonstrators in the KEOPS project", *Journal on Selected Areas in Communications*, Vol. 16, No. 7 (1998).

[148] G. Maier, A. Pattavina, S. De Patre, M. Martinelli, "Optical network survivability: protection techniques in the WDM layer", *Photonic Network Communications*, Vol. 4, No. 3/4 (2002).

[149] Lu Shen, Xi Yang, B. Ramamurthy, "A load-balancing shared-protection-path reconfiguration approach in WDM wavelength-routed networks", in *Proc. of OFC 2004*.

[150] L. Barbato, G. Maier, A. Pattavina, "Maximum traffic scaling in WDM networks optimized for an initial static load", in *Proc. of ONDM 2003*.

[151] Y. Jintae, I. Yamashita, S. Seikai, K. Kitayama, "Upgrade design of survivable wavelength-routed networks for increase of traffic loads", in *Proc. of ONDM 2005*.

[152] F. J. Blouin, A. W. Lee, A. J. Lee, M. Beshai, "A comparison of two optical-core networks", *OSA Journal of Optical Networking*, Vol. 1, No. 1 (2001).

[153] N. F. Huang, G. H. Liaw, C. P. Wang, "A novel all-optical transport network with time-shared wavelength channels", *Journal on Selected Areas in Communications*, Vol. 18, No. 10 (2000).

[154] M. Sivakumar, S. Subramannian, "A performance evaluation of time-switching in TDM wavelength routing networks", in *Proc. of BROADNETS 2004*,

[155] A. Pattavina, M. Tornatore, A. De Fazio, G. Maier, M. Martinelli, "Static WDM network planning with TDM channel partitioning", in *Proceedings of Networking 2004*.

[156] Y.C. Huei, P.H. Keng, "Framework for shared time-slot TDM wavelength optical WDM networks", *OSA Journal of Optical Networking*, Vol. 5, No. 7 (2006).

[157] R. Martinez, R. Munoz, M. Requena, J. Sorribes, J. Comellas, G. Junyent, "ADRENALINE Testbed: architecture and implementation of GMPLS-based network resource manager and routing controller", in *Proc. of TRIDENTCOM 2006*.

[158] R. Munoz, C. Pinart, R. Martinez, J. Sorribes, G. Junyent, A. Amrani, "The ADRENALINE Testbed: Integrating GMPLS, XML and SNMP in transparent DWDM networks", *IEEE Communications Magazine*, Vol. 43, No. 8 (2005).

[159] S. Jordan, "Implications of Internet Architecture upon Net Neutrality", *ACM Transactions on Internet Technology*, Vol. 9, No. 2 (2009).

Standardization

[160] Y. Rekhter, T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, Mar. 1995.

[161] C. Villamizar et al., "BGP Route Flap Damping", RFC 2439, Nov. 1998.

[162] J. Malcolm et al., "Requirements for Traffic Engineering Over MPLS", RFC 2702, Sept. 1999.

[163] D. McPherson, V. Gill, D. Walton, A. Retana, "BGP persistent route oscillation condition", RFC 3345, Aug. 2002.

[164] D. McPherson, V. Gill, "BGP MED considerations", RFC 4451, Mar. 2006.

[165] L. Berger, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, Jan. 2003.

[166] D. Papadimitriou, J. Drake, J. Ash, A. Farrel, L. Ong, "Requirements for Generalized MPLS (GMPLS) Signaling Usage and Extensions for Automatically Switched Optical Network (ASON)", RFC 4139, July 2005.

[167] S. Yasukawa, "Signaling Requirements for Point-to-Multipoint Traffic Engineered MPLS LSPs", RFC4461, Apr. 2006.

[168] R. Zhang, J.-P. Vasseur, "MPLS Inter-Autonomous System (AS) Traffic Engineering (TE) Requirements", RFC 4216, Nov. 2005.

[169] A. Farrel, A. Vasseur, J. Ash, "A Path Computation Element (PCE) based architecture", RFC 4655, Aug. 2006.

[170] J.-L. Le Roux, "Requirements for Path Computation Element (PCE) Discovery", RFC 4674, Oct. 2006.

[171] A. Farrel, J.-P. Vasseur, A. Ayyangar, "A Framework for Inter-Domain Multiprotocol Label Switching Traffic Engineering", RFC 4726, Nov. 2006.

[172] D. Meyer, L. Zhang, K. Fall, "Report from the IAB workshop on routing and adressing", RFC 4984, Aug. 2007.

[173] A. Farrel, A. Ayyangar, J.-P. Vasseur, "Inter-Domain MPLS and GMPLS Traffic Engineering - Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 5151, Feb, 2008.

[174] J.-P. Vasseur, A. Ayyangar, R. Zhang, "A per-domain PC method for establishing Inter-domain TE LSPs", RFC 5152, Feb. 2008.

[175] I. Bryskin, D. Papadimitriou, L. Berger, J. Ash, "Policy-Enabled Path Computation Framework", RFC 5394, Dec. 2008.

[176] J.-P. Vasseur, J.-L. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, Mar. 2009.

[177] J.-P. Vasseur, R. Zhang, N. Bitar, J.-L. Le Roux, "A Backward Recursive PCE-based Computation (BRPC) procedure to compute optimal inter-domain Traffic Engineering Label Switched Paths", RFC 5441, Apr. 2009.

[178] R. Bradford, J.-P. Vasseur, A. Farrel, "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Key-Based Mechanism", RFC 5520, Apr. 2009.

[179] J.-L. Le Roux, R. Jacob, R. Douville, "Carrying a Contract Identifier in the PCE communication Protocol (PCEP)", draft-leroux-pce-contract-id-01, Mar. 2007.

[180] S. Yasukawa, A. Farrel, "Applicability of the Path Computation Element (PCE) to Point-to-Multipoint (P2MP) Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) Traffic Engineering (TE)", RFC 5671, Oct. 2009.

[181] A. Lange, "Issues in Revising BGP-4", draft-ietf-idr-bgp-issues-01, June 2003.

[182] M. Blanchet, F. Parent, B. St-Arnaud, "Optical BGP (OBGP): Inter-AS Lightpath Provisioning", ietf-draft-parent-obgp-01, March 2001.

[183] "Architecture for the automatically switched optical network (ASON)", ITU-T G.8080/Y.1304, Feb. 2003.

[184] "Interfaces for the Optical Transport Networks (OTN)", ITU-T G.709/Y.1331, Mar. 2003.

[185] IP Sphere Framework Technical Specification, www.ipsphereforum.org

WWW online resources

[186] PlanetLab website. http://www.planet-lab.org

[187] LIP6 complex networks website, radar traces. http://data.complexnetworks.fr/Radar

[188] Vito Fragnelli, "Game theory and its application to telecommunication. Ph.D. course slides", http://antlab.elet.polimi.it/PUB/Fragnelli_slides.pdf, (2007).

[189] Paul Walker, "A Chronology of Game Theory", http://www.econ.canterbury.ac.nz/personal_pages/paul_walker/gt/hist.htm.

[190] *Wikipedia, the Free Encyclopedia*, http://en.wikipedia.org/wiki/ Metropolitan_cities_of_Europe, and *United States Census 2000 Population and Housing*, http://www.census.gov/main/www/cen2000.html

[191] By courtesy of Y. Zhang. Internet2/Abilene topology and traffic dataset. http://www.cs.utexas.edu/Ÿyzhang/research/AbileneTM.

[192] The CIDR report, http://www.cidr-report.org.

[193] CAIDA ranking (website), http://as-rank.caida.org

[194] PCH, Internet eXchange Directory, https://prefix.pch.net/applications/ixpdir

[195] Visual Market/2 Petaweb example dataset: http://perso.enst.fr/secci/petaweb.vmt.

[196] Visual Market/2 Petaweb example video: http://perso.enst.fr/secci/petaweb-idm.avi.

[197] Visual Market/2 software evaluation: http://www.ccas.ru/mmes/mmed

[198] Geographical coordinates: http://www.fes.uwaterloo.ca/crs/geog165/gcoords.htm.

[199] Cartographic Laboratory Department of Geography, University of Minnesota, 1989, "Air Distances", Airmaps Inc.

[200] US Department of Commerce, Economics and Statistics Administration, "United States Census 2000 Population and Housing", http://www.census.gov/main/www/cen2000.html.

[201] "The National Atlas of the United States of America", United States Department of the interior geological Survey, Washington, DC (1970).

[202] AMS-IX website. http://www.ams-ix.net.

[203] IP Networking Lab website. http://inl.info.ucl.ac.be/blogs/08-10-28-power-failure-ams-ix-one-largest-ix-europe.

Part III

Appendix

Appendix A

Principles of game theory

This appendix is due to the minor research activity carried out within the Ph.D. program at the Politecnico di Milano, under the advisory of Fioravante Patrone, Professor of Game Theory at the Università di Genova[1].

A.1 Introduction

Game theory is a mathematical theory, a discipline, that even a reader with scarce mathematical basis can adopt to analyze every day life problems: simple classical games, but also social behaviors, financial trends or telecommunication systems. In this appendix, we depict principles of game theory, voluntarily avoiding mathematical formalisms and historical facts for the sake of seamlessness. Our first references are, for an Italian reader, the book from Fioravante Patrone [41] and, for an English reader, the book from Osborne [42].

Game theory can intervene in all those situations in which more than one decision-makers have to plan their optimal strategy to solve a problem. Hence, it does not intervene in those situations with a single decision-maker, ground for operations research.

In the following, before diving into the ways in which a player's strategy is to be built and modeled, we highlight principles, concepts and assumptions of game theory, starting from the characterization of the decision-maker.

A.2 The decision-maker, i.e., the player

The reader may think that life is a sentence to forced works, full of duties and difficulties. Others may, instead, think that life is a game, and that a person in charge of a decision, the decision-maker, is a player, as the game theory predicates. Not whatever player: an intelligent and rational player[2].

A.2.1 Intelligence

The starting hypothesis of game theory is that the player is infinitely intelligent, in the sense that he has infinite ability of deduction, computation and analysis. This is

[1] The contexts presented in this appendix are also presented in [28].

[2] From now on, we will refer to the decision-maker or to the player with the third male person for simplicity.

a strong hypothesis especially for wide games, such as the chess, since just the simple enumeration of all the possible strategies may take a lot of resources. However, the vast majority of the problems that are appealing for the game theory are simple, and thereby this hypothesis is practically not binding.

A.2.2 Rationality

The meaning of rationality in game theory may not correspond to the common meaning for the reader. A classical assumption of game theory is to consider a player as rational, i.e., the player is supposed to be able to choose among several alternatives and to take decisions consistently with the preferences the player has on (the consequences of) those alternatives.

Furthermore, the preferences of a rational player are transitive. To explain this with an example rather than mathematically, consider a player that prefers apples to pears, and pears to bananas. If he is rational, he does not prefers bananas to apples. Otherwise, interacting with rational players he could be swiped. Indeed, another player with an apple, a pear and a banana may give you an apple as a gift; then, he may propose to exchange your apple with the banana for a few Euros since you prefer bananas to apples; then, he may propose you to exchange your banana with the pear since you prefers pears to bananas; then, the apple for the pear, and so on so forth he can impoverish you to infinity.

A.3 Multi-agent decision problem

In a game, an intelligent and rational player would choose a move so as to maximize the expected utility that the choice could return.

Let us consider an example. In Table A.1 we have a table describing a reduced version of the following roulette game: “to play 100 euro on the number 13 (strategy P), or not to play (strategy NP); if 13 is selected by the roulette (36 possible numbers), the player earns 3500 otherwise looses his 100 units". Each cell of the table indicates the expected gain for the player as function of the state of nature, where for this case the state of nature is the number selected by the roulette.

action | state of nature	0	1	...	13	...	36
P	-100	-100	...	3500	...	-100
NP	0	0	...	0	...	0

Table A.1: The roulette game

Let $u : \mathbf{R} \to \mathbf{R}$ be the utility function associating to a monetary gain a perceived utility. In order to take a decision, the decision-maker needs to know the probability of occurrence for each state of nature. In this way he can estimate the expected utility of each possible move. A rational decision-maker for which such an utility function and such calculations have sense, i.e., for which - under risk conditions - the preference relation over a finite set of states can be written as an expected utility, is said to be a Von Neumann-Morgenstern decision-maker; this is the affection we adopt for the decision-maker in the following.

Risk

If the probability distribution on the network states is exogenously given, the decision problem is said to be *under risk conditions*. For instance, let us consider the game described by the Table A.1. If each state of nature is equally probable, the player can estimate an expected utility for the strategy P equal to $\frac{1}{37}u(3500) - \frac{36}{37}u(100)$, for the strategy NP equal to $u(0)$, and a rational player will choose P only if the first value is bigger than the latter.

Uncertainty

If the state of nature probability distribution is not externally given, the player has to devise it. The player assigns endogenously the probability of each state starting by his preferences on the problem data. In this case, the decision problem is said to be *under uncertainty conditions* (and the player in this case is usually called a Bayesian decision-maker).

Interactivity

If in the place of the state of nature events there is a list of possible choices of another player, we have two players and thus an interactive decision. In such a game, the interaction is not intended in the sense that one player can directly influence the decision of the other player, but in the sense that the game result depends on the decisions of both the players.

Even if practically an interactive game may seem very close, if not equivalent, to a decision problem under uncertainty, in this case the player has to consider that on the other side there is not the luck, but another rational player as himself with different strategies and preferences on the results. Surely, the states of nature can encompass aspects due to others' decisions. However, under uncertainty there is no awareness on the possible interaction with other agents. Hence, the procedures that assign probabilities to state of nature and to other player's decisions are different, because of the different meaning the external states assume.

A.3.1 Dominance

A classical assumption is that all the parameters of the game are common knowledge among the players[3], i.e., each player knows the other player's possible strategies, the preferences he has on the alternative strategies, and this fact is also known, it is known that this fact is known, etc. In a strategic interaction context, the decision problem is usually not simple except the case in which the players dispose of strongly dominant strategies. A few definitions are needed to characterize the three different dominance levels. We use in the following a different terminology than the classical one, adopting the one suggested in [41].

Weak dominance: A strategy weakly dominates another alternative strategy if the payoff it grants is greater than or equal to the alternative payoff whatever the strategy of the other player will be.

Strict dominance: A strategy strictly dominates another alternative strategy if it weakly dominates the other strategies, and if for at least one strategy of the other player the payoff it grants is strictly greater than the alternative payoff.

[3]In the following, for simplicity we will refer to a two player game, when not differently mentioned.

Strong dominance: A strategy strongly dominates another alternative strategy for a player if the payoff it grants is strictly greater than the alternative payoff whatever the strategy of the other player will be.

Therefore, if both players have a strongly dominant strategy, the rational solution of the game is the pair of the two dominant strategies; rational in the sense that if the two players are rational, their moves should stop in this solution point.

Consider for example the game whose "strategic form" is represented in Table A.2. In this case, each cell indicates the payoffs for player I and player II respectively, within brackets and separated by a comma. The player I has three possible strategies (T, M and B), and the player II has twos (L and R). Each cell contains the two payoffs for the corresponding pair of strategies of player I and player II. At a first sight one can notice that M is strongly dominated by both T and B, and thus I will never choose M. Under the condition that II knows the payoffs of I and that I is rational, II assumes that I will not play M, and since R is strongly dominated by L, II chooses L. Under the condition that I knows the payoffs of II, that I knows that II is rational and that II knows that I knows, I plays B. Hence by iterated reduction of strategies each player can determinate that the solution pair of strategies will be (B,L). Nevertheless, when there are no strongly dominating strategies, the iterated reduction of strategies has no bite.

I | II	L	R
T	(2,1)	(3,0)
M	(1,0)	(1,1)
B	(3,1)	(2,0)

Table A.2: A game with strongly dominating strategies.

A.3.2 Equilibrium strategies

An equilibrium of strategies is a pair of strategies such that, assuming the player II strategy given, player I's strategy gives a payoff greater than or equal to the payoffs of his other alternative strategies, and vice-versa[4]. This definition of strategic equilibrium is due to Nash, and it is commonly referred to as "Nash equilibrium". It has sense only for the class of games to which we have focused till now, that is the class of *non-cooperative games* in which there can not be *binding agreements* among players. A Nash equilibrium is thus the solution to which rational players of a non-cooperative game shall implicitly tend.

A Nash equilibrium satisfies all the players, because it naturally represents an acceptable point for rational players. This is due to the fact that both the players would not have interest in deviating from a Nash equilibrium. Indeed, if a couple of strategy does not correspond to a Nash equilibrium, at least for one player it would be better to change to a strategy that gives him a greater payoff (this is true because the players are rational).

[4]Please note that this differs from the weak dominance that stands for all the possible strategies of the other player, and thus not constrained to a single strategy of the other player.

Rationality and efficiency

Nevertheless, an issue may arise: even if the players are rational, an equilibrium may not be efficient.

Consider for example the game in Table A.3, commonly named “the prisoner’s dilemma". In this example, the payoffs may indicate the years spent out of jail, for two prisoners, different whether a prisoner confesses (C) or does not confesses (NC) a crime he has committed with the other prisoner. If both confess, each prisoner gets out from jail 1 year before; if both do not confess, both get out 2 years before; otherwise, who does confess gets out 3 years before and the other one sticks.

I | II	NC	C
NC	(2,2)	(0,3)
C	(3,0)	(1,1)

Table A.3: The prisoner’s dilemma.

The dilemma is the following: the Nash equilibrium apparently is (C,C), even if the efficient pair of strategies for both would be (NC, NC). The strategies for the Nash equilibrium (C,C) are strongly dominating strategies. (NC, NC) is not an equilibrium because each player would prefer to change strategy to earn 1 year more whether he assumes that the other player would choose NC, and viceversa. Hence this would not happen, being both players rational and intelligent: the assumption that the other player would choose NC is thus wrong and not rational.

The conclusion is that even if the players are rational and intelligent, this is not sufficient for getting an efficient result. Even if they could communicate somehow to agree on (NC,NC), a rational player should not hope the other will do what they agreed upon.

If cooperation is needed, it would have sense only if the players can subscribe binding agreements so that, for example, each prisoner can be reassured somehow that the other will not confess to reach the efficient solution (NC,NC). In this case we would have a *cooperative game*. For a non cooperative game, one may thus have inefficient Nash equilibrium solutions.

Multiple equilibria and coordination

Another issue that may arise in games is the Nash equilibrium non-uniqueness.

Consider the example game in Table A.4. There are two equilibria: (T,L) and (B,R). There is here room for an implicit coordination between players and both would rationally choose (T,L) since both have the same preference on that result.

I | II	L	R
T	(2,2)	(0,0)
B	(0,0)	(1,1)

Table A.4: A coordination game.

Consider now the game in Table A.5. There are two equilibria: (T,L) and (B,R), and the players have the same preferences for both the equilibria. This game is dramatic for

I | II	L	R
T	(1,1)	(0,0)
B	(0,0)	(1,1)

Table A.5: A pure coordination game.

both the players because a 1 bit coordination message would be enough to successfully play for (T,L) or (B,R). Hence, such a game is classified as a *pure coordination game*.

Consider finally the game in Table A.6. Also for this game we have the two equilibria (T,L) and (B,R), but the players have adverse preferences for them. For this reason, such a game is commonly named "the battle of the sexes": suppose that player I and II are a man and a woman, the man prefers to go the stadium (S) rather than to the theatre (T), while the woman prefers the inverse (T and S, respectively), and both have a null preference in going alone to the theatre or to the stadium and a not null preference, however, if they go in the not preferred place with the partner. Keeping such preferences, even with coordination between players one can not say what rational players would decide. A form of coordination would be useful only in that it could at least allow avoiding results with null payoffs.

I | II	S	T
S	(2,1)	(0,0)
T	(0,0)	(1,2)

Table A.6: The battle of the sexes.

An implicit coordination approach among players is to choose a Nash equilibrium that is efficient in the Pareto sense. The Pareto-efficiency is a formal criterion to determine the efficiency of a profile in a non-cooperative game. A strategy profile s is *Pareto-superior* to another profile s' if a player's cost for s is minor than for s', while the other players' costs for s are not bigger than for s'. In such a case s' is *Pareto-inferior* to s. A strategy profile is *Pareto-efficient* if it is not Pareto-inferior to any other strategy profile. The set of Pareto-efficient strategy profiles is the *Pareto-frontier* of the game. E.g., the (T,L) profile in Table A.4 is Pareto-efficient and Nash equilibrium. Nevertheless, generally a Pareto-efficient Nash equilibrium may not exist, or many may do. E.g, the game in Table A.3, has the Pareto-efficient profile (C,C), which is not however the Nash equilibrium.

Absence of equilibrium and mixed strategies

Finally, there exist games for which there are no equilibria at all.

Consider for example the game in Table A.7, the classical "matching pennies" (that falls into the class of *zero sum games* for which the sum of the payoffs is null). It is easy to verify that there are no equilibria.

I | II	Heads	Tails
Heads	(-1,1)	(1,-1)
Tails	(1,-1)	(-1,1)

Table A.7: Matching pennies game.

This may disappoint the reader, who may loose the trust in a theory that can not give solutions for a so simple game. However, there is a chance to overcome the non-existence of the Nash equilibrium. The key brick to help the decision in such situations is represented by the *mixed strategies*.

In mixed strategies, the player no longer chooses a single alternative, but chooses a probability distribution on the alternatives. Somehow the player can rely on a random process that implements his decision defined by the probability distribution. In mixed strategies, the best strategy consists in the best configuration of the player's probability distribution.

In non-cooperative games, the two players adopt *independent* random processes, and the probability distribution on the pair of strategies is given by discrete multiplication of both the processes[5]. For example, by associating to the game in the strategic form in Table A.7 the probability distribution in Table A.8, we have the *mixed extension* of the matching pennies game (by reflection the start game can be referred to as a *game with pure strategies)*. The player I chooses the first strategy with probability p, the second with probability $1-p$; similarly on his two alternative strategies, the player II does with probability q.

I | II	Heads: q	Tails: $1-q$
Heads: p	pq	$p(1-q)$
Tails: $1-p$	$(1-p)q$	$(1-p)(1-q)$

Table A.8: Probability distribution on strategies.

We can thus calculate the expected payoff for I, equal to $u_I(p,q) = pq(-1) + p(1-q)(1) + (1-p)q(1) + (1-p)(1-q)(-1) = (2-4q)p + 2q - 1$. Given this expected payoff, what is the best configuration of p for player I? When $(2-4q) > 0$, i.e. $q < 1/2$, the best reply is $p = 1$ in order to maximize $u_I(p,q)$. When $(2-4q) < 0$, i.e. $q > 1/2$, the best reply is instead $p = 0$. When $(2-4q) = 0$, i.e. $q = 1/2$, the best reply is any p. This reply defines the full line in Fig. A.1. For player II, the expected payoff is equal to $u_{II}(p,q) = (4p-2)q + 1 - 2p$. Similar considerations bring to the dashed line in Fig. A.1.

The Nash equilibrium in mixed strategies for the matching pennies game is given by the intersection between the two lines, thus when $p = q = 1/2$: both players play Heads with probability 0.5 and Even with probability 0.5 (which is not surprising for this game).

It is worth at this point clarifying the role of the Nash equilibrium with pure strategies and with mixed strategies:

- With pure strategies a (finite) game may not have Nash equilibrium, while with mixed strategy there always exists at least one Nash equilibrium.
- All the Nash equilibria with pure strategies appear with mixed strategies, in the sense that a "pure" Nash equilibrium appears with strategies' probabilities equal to 1 with mixed strategies.
- If a Nash equilibrium with pure strategies is obtained for strongly dominating strategies, with mixed strategies no new equilibrium is introduced, otherwise at least one new equilibrium is introduced (this is the case, for example, of the "battle of the sexes", Table A.6).

[5]This is thus similar to but clearly different than what mentioned in Sect. A.3

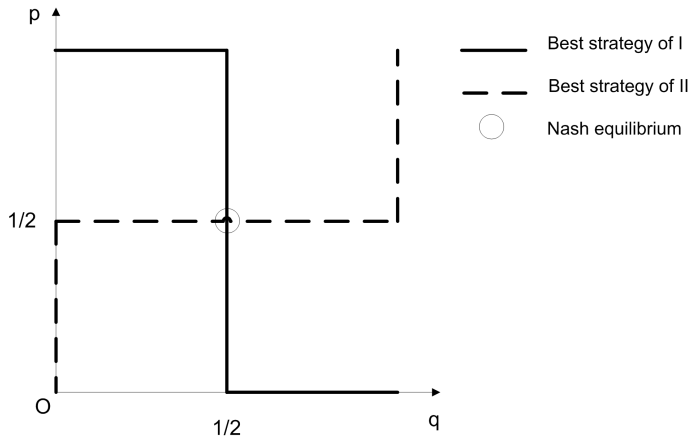


Figure A.1: Nash equilibrium in mixed strategies - Matching pennies game

I | II.1	NB	B
E	(1, 2)	(-1, 0)
NE	(0, 3)	(0, 2)

I | II.2	NB	B
E	(1, 2)	(-1, 3)
NE	(0, 3)	(0, 5)

Table A.9: Example of game with incomplete information. II.1: High building costs on the left side. II.2: low building costs on the right side.

- With both pure and mixed strategies, a Nash equilibrium is not necessarily a Pareto-efficient profile, and a Pareto-efficient profile is not necessarily a Nash equilibrium.

A.3.3 Incomplete information

Classically, information in a game is assumed to be common knowledge and the game is said to be with “complete information". However, a real player would typically not dispose of exhaustive information about relevant parameters of his decision problem, would not be able to choose the best alternative among the available alternatives, and would surely have limited computation and deduction abilities. It is not practically possible to use the finest accuracy in the computation of the preferences, for example. Or, the limited memory capacity narrows the deduction ability, or an equal satisfaction may be associated to $x\%$ near but not equal consequences. Under all possible effort, loosing much time computing strategies and payoffs, analyzing the problem and all possible alternatives, the player would approach the ideal solution strategy. But normally there is a limit to this tension to reach the ideal solution because of material bounds.

Relaxing the classical assumption that all the relevant parameters are known by the players, and keeping the classical assumptions of infinite rationality and intelligence, we have a game with “incomplete information". In such a case, the relevant parameters that a player does not know are typically the preferences/utilities, the number and the type of the other players. Two types of player can differ for example on the kind of available strategies.

In Table A.9 we depict a simplified game with incomplete information. In this example, the incompleteness relies on the fact that player I does not know the type of player II among two possible types (II.1, II.2). In particular, player I may be a company that has to decide if entering (E) on a market or not (NE), and player II may be the

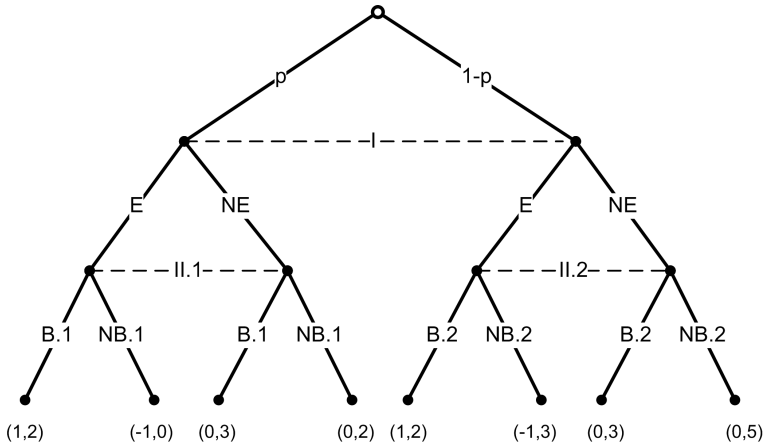


Figure A.2: Extended form of the Table A.9 game example

incumbent company on that market that has to decide if building (B) a new plant or not (NB). From the player I standpoint, he does not know if the incumbent company would have high building costs (left side in Table A.9), i.e. it is of type II.1, or if it would have low building costs (right side in Table A.9), i.e. it is of type II.2. How can player I solve such an issue: has he to consider II.1 or II.2 or both when defining its strategy?

In Fig. A.2 we have the "extended form"[6] of the game: each edge at the stage two represents a possible action of player I, and at the stage three we have actions of the player II. The game terminates at one of the bottom points where the utilities are indicated. By modeling the choice between II.1 and II.2 as a process that selects II.1 with probability p and II.2 with probability $1-p$, one can divide the game into the two possible subgame branches as depicted at the first stage of the figure. Between stage one and stage two we draw a horizontal dashed line to indicate the fact that when choosing between E and NE player I does not dispose of the information about the selection of the luck process. Also, the dashed line between the second and the third stage indicates that when choosing among B and NB the player II, II.1 or II.2 respectively, does not know the move of player I. The dashed lines are thereby useful to represent the *state information set with incomplete information.*

Player I can thus now easily find his optimal strategy, helped by the fact that the subgames have a single Nash equilibrium: (E, NB) for the game with high building costs, (NE,B) for the game with low building costs, thus to enter when the incumbent company would not build a new plant because of the high costs, not to enter when the costs are low and the incumbent company builds. Player I will thus enter only if the expected payoff for this strategy is better than otherwise, i.e. if $p\cdot(1)+(1-p)\cdot(-1) > p\cdot(0)+(1-p)\cdot(0)$, thus if $p > 1/2$.

Generalizing, one may assume that many pairs of player types (I.i,II.j) manifest for both the players, and that the probability that such a pair manifests is known by the single player. The single player is called "Bayesian player". The solution of such an incomplete information game is called "Bayesian equilibrium" or "Nash-Bayesian equilibrium". The probabilities that a pair of players manifests, or that a single player manifests, can be exogenously given or endogenously calculated by a player. The probability distributions that each player assigns to the types of the other players is called

[6]The extended form of a game is a form other than the strategic form or the characteristic, to represent a game.

I | II.1	L	R
T	$(5, 2)^4$	$(-1, -2)^0$
B	$(-5, -4)^{-6}$	$(1, 4)^2$

Figure A.3: Example of potential game.

belief. Based in his beliefs, player I/II is able to calculate the probability of having a type j/i of player II/I and can consequently select the optimal strategy with a straightforward generalization of the procedure described in the previous simple example.

A.3.4 Equilibrium selection issue

Ideally, non cooperative (interactive) games would have an equilibrium of strategies, being this equilibrium characterized by a pair of strategies (for two players games). However, a game could not have equilibria at all. Or, as usually it happens, a game could have more than one equilibrium. And when this happens, what is the “right" solution?

Game theory can not answer conclusively this last question. What game theory gives is the set of possible strategy that can be adopted to augment to (expected) player's payoff. When several equilibria are present, or when none exists, it is possible to solve the decision problem using mixed strategies. But also with mixed strategies one may have several, possibly new, equilibria. Among multiple equilibria, players may implicitly coordinate to play a Pareto-efficient one, but a Pareto-efficient equilibrium may not exist or many Pareto-efficient equilibria may exist. Finally, we showed how, in incomplete information games, by opportunely modeling the incompleteness of the information one can still find one or more solution strategies.

Several other methods have been defined to select one single equilibrium among several equilibria. Notably, if the dynamics of the game is known, i.e, if the dependences between players' strategies are known, an extended branching form with the strategy dependences can be drawn. On the extended form, one can restrict its focus on the equilibria that can be really experienced (selecting among the so-called *subgame* equilibria). In this way he can skip artificial equilibria that refer to strategies that could not be practically experienced. Nevertheless, a game may not have complex dynamics, there may not be complete information awareness, or the knowledge of the game dynamics may not be acquired.

A.3.5 Potential games

A class of games, the potential games, deserve some attention, especially because it seems having many application in transport network problems (an italian reader can refer to [47]; [45] is the seminal work).

A game possedes a ‘potential’ if the difference in passing from a strategy to another can be expressed with a single (potential) function for all the players. E.g., consider the example in Table A.3. The potential values are indicated as exponents of the payoff vectors. The difference from T to B for I is equal the potential difference, and so does the difference from L to R for II. Hence the strategic form of the game may be simply reduced to the potential function.

The simpler analogy with physics is that we can substitute two functions (a vector field) with one (a scalar one), besides being defined with the discard of a constant. A

very useful property of potential games is that, whether the strategy sets for all the players are finite, the maxima (or minima for cost games) correspond to Nash equilibria (please note that the inverse is generally not true), and hence that a Nash equilibrium always exists.

A.4 Cooperative games

What if the two players can communicate to agree on the strategy profile satisfying both of them, being assured that the agreement will be respected? In such a case we have a cooperative game. In a cooperative game, groups of players enforce a cooperative behavior. For 2 players, the cooperative game can be modeled as a bargaining problem as that we describe in the following.

A.4.1 Bargaining problem

A bargaining problem instance may be, for example, the problem of agreeing in the partition of a monetary amount M among two players[7]. There will be a value for each player that represents the status quo, the point corresponding to no cooperation. Let us indicate with d_1 and d_2 the utility for player 1 and 2 in this point, such that obviously $d_1 + d_2 \leq M$. The $d = (d_1, d_2)$ point represents thus the disagreement point. All point (x_1, x_2) such that $x_1 \geq d_1$ and $x_2 \geq d_2$ and $x_1 + x_2 \leq M$ is a possible solution, and all such points define a closed, convex, bounded and non-empty solution set where a reasonable solution should be found. Let F be the generic bargaining set (with upper boundaries not simply defined by $d_1 + d_2 \leq M$), as represented in Fig. A.4. (F, d) is a bargaining problem.

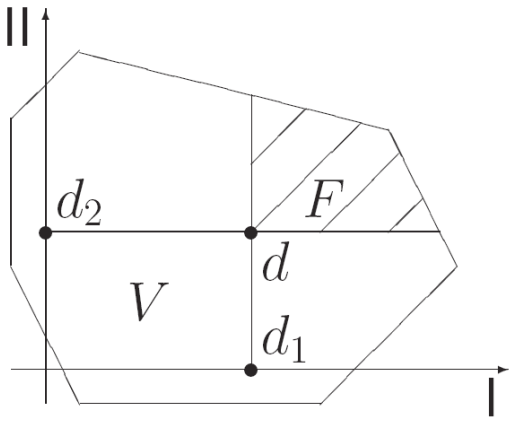


Figure A.4: Bargaining set.

How to find a reasonable solution to this problem, what is the point in F to agree upon? Two valuable answers are worth being mentioned here, the one from Nash and the one from Kalai and Smorodinsky. Before, some basic axioms are to be introduced:

1. Symmetry.
   If the solution set F is symmetric with respect to the bisector of the axes having d=(d_1, d_2) as centre and $d_1 = d_2$, then the solution is a point of the bisector (and

[7] In this example, we are assuming that players have an utility "linear with money" (otherwise stated, that they are risk neutral)

thus the same payoff for the two players). This means that the two players, equally rational and intelligent, have the same ability in bargaining.

2. Strong efficiency.
   The solution does not encompass cases in which a part of the value (M) is not distributed to any player.

   For example, in Fig. A.5 the red line highlights possible efficient solutions, while the adjacent horizontal and vertical lines indicate solution moves that do not allow the other player to improve and to distribute all the payoff amount (thus inefficient).

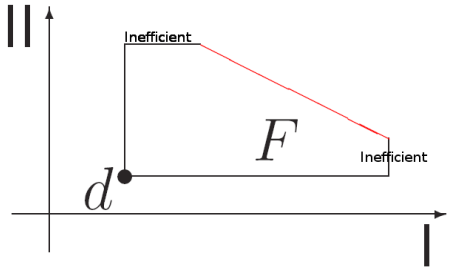


Figure A.5: Strongly Efficient solution subset

3. Scale covariance.
   The bargaining is not dependent on the scale of the utility function (like if a contract signed in Euro is converted in Dollars). Any coordinate transformation of (F,d) into (F',d') coherently transforms a solution in the origin space into a solution in the new space.

4. Independence from Irrelevant Alternatives (IIA).
   The solution point x^* of (F,d) remains the solution of (G,d) if $G \subset F$ and $x^* \in G$. The solutions in $F - G$ represent irrelevant solution alternatives.

5. Individual Monotonicity.
   Given $G \subseteq F$, shrinking of the original space F, if the maximum payoff point for player I in the shrinked space (G,d) is equal to the maximum payoff point in (F,d) (i.e., if it does not change after the space shrinking), then the new expected solution point in the shrinked space gives to player II a payoff less or equal to the old one; and viceversa.

Nash product

Under the above-mentioned 1-4 axioms, Nash proved that the optimal solution of the bargaining problem is given by the point that maximizes the product of the utility of the two players, where with utility here we intend the difference between the final payoff and the payoff given by the disagreement point. Hence:

$$N_s = argmax\{(x_1 - d_1)(x_2 - d_2)|(x_1, x_2) \in F\} \tag{A.1}$$

This may sound really strange, especially if the reader did not pay sufficient attention to the above axioms. Apart axioms 1-3 which are quite straightforward and reasonable, in particular for games with more than 2 players axiom 4 is debatable, instead. Consider this example. Among three candidates for an election, A, B and C, A wins. Afterwards,

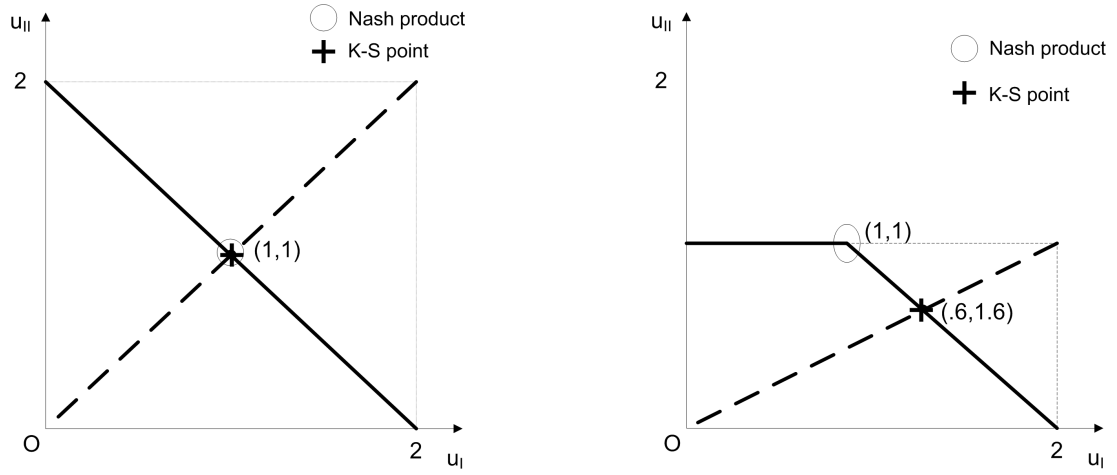


Figure A.6: Comparison example between Nash product and K-S solutions

one discovers that C was not eligible because of some matters. If one applies the IIA axiom, one should not repeat the elections, because C is to be considered as an irrelevant alternative given that A has already won on B. However, one could debate that given that the electoral program of C is very close to that of B, if the election had only two candidates, A and B, B could have won on A. Instead, with only two players, say A and B, the IIA axiom seems justified.

Kalai and Smorodinsky's solution

Therefore, even if the Nash product is the solution approach for games under 1-4 axioms, maybe more acceptable solutions could be reached if the axiom 4 were refined with axiom 5, as Kalai and Smorodinsky did.

The Individual Monotonicity introduces the following reasonable novelty with respect to the IIA: cutting possible alternative solution points of a player would reduce the chances that this player has to get a good solution and thus these points are not irrelevant. So, for the election game, the elections should be repeated.

Under 1-3 and 5 axioms, Kalai and Smorodinsky (K-S) proved that the solution is the intersection between the straight line from the disagreement point to an *utopia point* corresponding to the maximum desirable payoffs for the two players (usually outside the solution space), and the external higher border of the solution space. E.g., in Fig. A.6a we have that because of symmetry and efficiency, the solution is given by (1,1), which is also the Nash bargaining equilibrium. In Fig. A.6b, we suppressed a part of the solution set, in fact reducing convenient alternatives for player I. The Nash product is unchanged, thanks to the IIA axiom. The K-M solution is instead changed into (.6,1.6). Indeed the utopia point is not (2,2) any longer, but (2,1); the utility of player I decreased at the advantage of player II.

A.4.2 Coalitional n/player games

We have briefly depicted the most primitive and simple cooperative game, the bargaining problem, and possible solution approaches. What happens when players are more than

two? With more than two players, it is no longer a “simple" issue on how to bargain: groups of players, or coalitions, may ally in order to improve their utility.

We have a coalitional cooperative game when players can agree on a common strategy to play (assured that this agreement will be respected). Players can ally to improve their payoff. In the following, we focus on the case for which the players may share the total payoff of a coalition [Transferable Utility(TU)-games], special case of the more general case - we do not treat in this report - in which binding agreements indicate how to spread the payoff among players inside a coalition [Non Transferable Utility(NTU)-games].

A coalitional game can be fully represented via the “characteristic form", which indicates the set of players (N) and an utility function ν that associates a coalitional payoff to each possible coalition (S) of players ($\nu(S),\ \forall S \subseteq N$), including single player coalitions and the grand coalition. For example, let's consider a game with the three players A, B and C: $N = \{A, B, C\}$. Let $\nu(A) = 4$, $\nu(B) = 3$, $\nu(C) = 2$ indicate the utilities, payoffs, for single players. If the players collaborate and form some coalitions to increase their utility, let $\nu(AB) = 2$, $\nu(AC) = 2$, $\nu(BC) = 6$ and $\nu(ABC) = 1$ indicate the utilities for the possible coalitions. Each coalition is detrimental but $\{B,\ C\}$: B and C would ally to increase together their payoffs. A (N, v) coalitional game is thus fully characterized by the possible coalitions and the utility associated to each coalition.

In the previous example, clearly the best solution is not to run all together to form a single grand coalition, since single player or two players coalitions can get more than the grand coalition. The grand coalition would perhaps form and be stable if the utility function is *superadditive*, i.e., if for all possible pairs of disjoint coalitions, a mother coalition embracing all the players of sub-coalitions gives at least the same utility than the sum of the two sub-coalitions' utilities. In other words, a game is superadditive if allying has always a positive or null effect. For example, for $N = \{A, B, C\}$, the game is superadditive if $\nu(AC) \geq \nu(A) + \nu(C)$, $\nu(AB) \geq \nu(A) + \nu(B)$, $\nu(BC) \geq \nu(B) + \nu(C)$ and $\nu(ABC) \geq \nu(A) + \nu(B) + \nu(C)$. In general, if $\nu(S) \geq \sum_{i \in S} v(i),\ \forall S \subseteq N$.

Once a coalitional solution is determined, how the utility is spread among the players? The main assumption in cooperative game theory is that under superadditiviy the grand coalition N will form since this would be the best solution. The challenge is then to allocate the utility $\nu(N)$ among the players in the same coalition in a fair way. The solution is a final payoff allocation vector $\overline{x} = (x_i)_{i \in N}$, called *imputation*, that represents the payoff allocation for each player that satisfies the following conditions:

- Feasibility.
  The sum of the allocations to the players is at most equal to the grand coalition utility (i.e., $\sum_{i \in N} x_i \leq \nu(N), \forall N \subseteq N$).

- Individual rationality.
  Each player gets at least what he could get alone (i.e., $x_i \geq \nu(i) \forall i \in N$).

- Collective Rationality.
  The sum of player payoffs is at least equal to the amount of the grand coalition (i.e., $\sum_{i \in N} x_i \geq \nu(N)$). If the feasibility is also respected, the sum of the payoff allocations is thus equal to the utility of the grand coalition (i.e., $\sum_{i \in N} x_i = \nu(N)$).

A.4.3 Core

By extending the collective rationality constraint to many *coalitional rationality* constraints, that is, by imposing that the sum of the payoffs of the players of a coalition S is at least equal to the utility associated to that coalition (i.e., $\sum_{i \in S} x_i \geq \nu(S), \forall S \subseteq N$),

we obtain a set of possible solution imputations, called *core*, for a coalitional cooperative game.

What is the meaning of the core? The core includes the imputations under which no coalition has a value greater than the sum of its members' payoffs. Therefore, no group of players has incentive to secede from the grand coalition since he can not receive a larger payoff. The core is useful to select which solutions should not be chosen if one desires a stable grand coalition: the imputations not belonging to the core. Hence, if the core is empty, the grand coalition would not be a stable solution; indeed, in such a case, excessively strong intermediate coalitions offer attractive alternative solutions to the players. For example, let us consider the characteristic form:
$N = \{A, B, C\}$. $\nu(A) = 4$, $\nu(B) = 3$, $\nu(C) = 2$, $\nu(AB) = 8$, $\nu(AC) = 6$, $\nu(BC) = 5$, $\nu(ABC) = 12$.
The core contains the possible imputations for which the grand coalition ABC is advantageous. The imputation (5,3,3) does not belong to the core because $5 + 3 + 3 < \nu(ABC)$, even if $5 + 3 \geq \nu(AB)$, $5 + 3 \geq \nu(AC)$ and $3 + 3 \geq \nu(BC)$. The imputation (5,4,4) does, instead, belong to the core.

Example 1: the majority game

Three players A, B and C, can obtain together a certain value, say 1, that has to be shared among them. To share 1, they vote and the majority determines to who the value will be assigned. So, two players can agree and guide the results with their majority. That is: $N = \{A, B, C\}$; $\nu(A) = \nu(B) = \nu(C) = 0$; $\nu(AB) = \nu(AC) = \nu(BC) = \nu(ABC) = 1$.

It is a superadditive game. In order to calculate the core of the game we explicate the condition: $\sum_{i \in S} x_i \leq \nu(S), \forall S \subset N$, that is:

$$x_A + x_B \leq \nu(AB) = 1 \ , \quad x_B + x_C \leq \nu(BC) = 1 \ , \quad x_A + x_C \leq \nu(AC) = 1 \tag{A.2}$$

and summing all members we obtain:

$$2(x_A + x_B + x_C) \leq 3 \tag{A.3}$$

that is never satisfied, given that $x_A + x_B + x_C = 1$ always. Hence, the core is empty, there is no allocation in the core for the majority game.

A.4.4 Shapley Value

Since some actors may contribute more than others, the question arises on how to distribute fairly the gains among the actors. How important is each actor to the coalition, and what payoff can be reasonably expected? The players' worth is not directly considered in the core definition. The Shapley value is another type of payoff vector, which is calculated as follows:

- consider all the possible permutations of the players (e.g., if we have three players {1, 2, 3} the permutations are 123, 132, 213, 231, 312, 321);
- for each permutation and each player, calculate the marginal contribution that the player grants if he joins the coalition formed by the predecessor players (e.g., for the permutation 312, the contribution of 2 is $\nu(123) - \nu(13)$, that of 1 is $\nu(13) - \nu(3)$, that of 3 is $\nu(3) - \nu(\emptyset) = \nu(3)$);
- for each player, calculate the average of its marginal contributions on all the permutations.

	1958			1973		
Countries	Weight	%	Shapley	Weight	%	Shapley
France	4	23.53	0.233	10	17.24	0.179
Germany	4	23.53	0.233	10	17.24	0.179
Italy	4	23.53	0.233	10	17.24	0.179
Belgium	2	11.76	0.150	5	8.62	0.081
The Netherlands	2	11.76	0.150	5	8.62	0.081
Luxemburg	1	5.88	0.000	2	3.45	0.010
United Kingdom	-	-	-	10	17.24	0.179
Denmark	-	-	-	3	5.17	0.057
Ireland	-	-	-	3	5.17	0.057
Total	17	100.00	1.000	58	100.00	1.000

Figure A.7: Application of the Shapley values. EU Council 1958-1973. Source: [189].

The Shapley value is thus equal to zero for useless players, which do not offer any marginal contribution to a coalition in any case, and equal to the single-player payoff for dummy players, which are indifferent in staying in the coalition or not. The Shapley value can thus be used to assign the payoff of a player as function of his marginal contribution to the coalition. And given that the marginal contribution that a player brings to a coalition varies as function of the players that already form the coalition, it is essential considering the order in which the player enters the coalition (or would enter if a coalition evaluates the opportunity of joining the new player).

Fig. A.7 illustrates the results in applying the Shapley value calculus to evaluate the weight assigned to the Council EU countries. In 1958, the majority quota was 12 on 17 ($\approx 70\%$) while in 1973 it was 41 on 58 ($\approx 70\%$). The Shapley value points out a "worm" in the vote weight assigned to Luxemburg in 1958: a weight of 1 had no worth in 1958. The vote of Luxemburg was not significant for having the majority. In 1973, new members entered the EU, and the weight had to be reassigned: every member became significant.

For superadditive games, the Shapley value is an imputation, but not always belongs to the core, even if the core is non-empty. But if the game is convex, i.e., if the incentives for joining a coalition increase as the coalition grows (i.e., if $\nu(S \cup \{i\}) - \nu(S) \leq \nu(T \cup \{i\}) - \nu(T),\ \forall S \subseteq T \subseteq N \backslash \{i\}\ \forall i \in N$), the Shapley value belongs to the core.

A.4.5 Kernel

Another concept is introduced to select under a differently attractive criterion possible solutions to coalitional games. The "Kernel" contains all imputations such that no player has bargaining power over one another.

First, given an imputation $\overline{x}$, one calculates the maximum surplus $s_{i,j}(\overline{x})$ as the maximal amount Player i can gain without the cooperation of Player j by withdrawing from the grand coalition N, under payoff vector $\overline{x}$, to form another coalition S (i.e., $s_{ij}(\overline{x}) = \max\{\nu(S) - \sum_{k \in S} x_k : S \subseteq N \setminus \{j\}, S \ni i\}$). Hence, the maximum surplus is a way to measure one player's bargaining power over another. Intuitively, Player i has more bargaining power than Player j with respect to imputation x if the maximum surplus he has on j is bigger than the maximum surplus j has on him [i.e., if $s^{\nu}_{ij}(\overline{x}) > s^{\nu}_{ji}(\overline{x})$].

Then, the Kernel contains all imputations where no player has this bargaining power

over another, i.e., the imputations such that when Player i has more bargaining power than Player j, Player j is immune to Player i's threats [hence if $x_j \leq \nu(j)$] because he can obtain this payoff or a better payoff on his own:

$$[s^{\nu}_{ij}(\overline{x}) - s^{\nu}_{ji}(\overline{x})][x_j - \nu(j)] \leq 0 \ \wedge \ [s^{\nu}_{ji}(\overline{x}) - s^{\nu}_{ij}(\overline{x})][x_i - \nu(i)] \leq 0, \ \forall i, j \in N \qquad (A.4)$$

A.4.6 Nucleolus

Finally, another concept is worth being mentioned here: the “Nucleolus". The basic idea is that it is necessary to improve the situation of the player that is worst off.

The Nucleolus is computed on the imputation set as follows.

- calculate the coalitional surplus for each coalition S as the difference between the coalition worth $\nu(S)$ and the received payoff $\sum_{i \in S} x_i$;
- order the coalitions with respect to the coalitional surplus, decreasingly, and fill a coalitional surplus vector $\overline{\theta}(\overline{x})$ following this order;
- the Nucleolus is the imputation $\overline{x}_n$ such that $\overline{\theta}(\overline{x}_n)$ is the lexicographic minimum of all possible $\overline{\theta}(\overline{x})$ (given $\overline{a}$ and $\overline{b}$, $\overline{a}$ precedes lexicographically $\overline{b}$ if $\exists i \geq 1 | a_j = b_j \ \forall j < i \wedge a_i < b_i$; e.g., $(1, 2, 3, 4)$ precedes $(1, 2, 4, 4)$).

A game has a unique nucleolus and the nucleolus is always in the kernel. If the core is non-empty, the nucleolus belongs to the core and can be used for selecting a core element.

Visibly, the computation of the nucleolus is not trivial. One should execute the steps above for all imputations. Otherwise, it can be modeled as a mixed linear programming problem. The objective is to minimize the maximal coalitional surplus α such that:

$$\nu(S) - \sum_{i \in S} x_i \leq \alpha, \ \forall S \subset N \qquad (A.5)$$

And the grand coalition has always null surplus:

$$\sum_{i \in N} x_i = \nu(N) \qquad (A.6)$$

Solving this LP, many optimum solutions might be obtained for the same minimum α_0. If this is not the case, the solution vector $\overline{x}$ is the nucleolus, and α_0 is the maximal coalitional surplus. Otherwise, one extracts the set of binding coalitions S_0 and iterates the optimization changing (A.5) with:

$$\nu(S) + \sum_{i \in S} x_i = \alpha_0, \ \forall S \in S_0 \qquad (A.7)$$

when a coalition belongs to S_0. S_0 is to be built as set of coalitions that are likely to oscillate the least, typically one takes the least surplus set of coalitions. After at most n iterations the solution is unique and it is the nucleolus of the game.

Example 2: the bankruptcy problem

We report a fact listed in the chronology in [190]:

The Babylonian Talmud is the compilation of ancient law and tradition set down during the first five centuries A.D. which serves as the basis of Jewish religious, criminal

and civil law. One problem discussed in the Talmud is the so called marriage contract problem: a man has three wives whose marriage contracts specify that in the case of this death they receive 100, 200 and 300 respectively. The Talmud gives apparently contradictory recommendations. Where the man dies leaving an estate of only 100, the Talmud recommends equal division. However, if the estate is worth 300 it recommends proportional division (50,100,150), while for an estate of 200, its recommendation of (50,75,75) is a complete mystery.

The rule seems the following: when the estate is at most equal to the minimum claim, divide it equally; when it is at least equal to the maximum claim, divide it proportionally. When it is in between, a mysterious method is applied. In 1985, a study from Aumann and Maschler [44] clarifies the mysterious method, e.g., when the estate is 200, finding its justification in the nucleolus, as for the cases with an estate of 100 and 300. They model the bankruptcy problem as a cooperative coalitional game.

The characteristic form of the corresponding coalitional game is defined taking the worth of a coalition S to be what it can get without going to court, i.e., by accepting either nothing, or what is left in the estate E after each member i of the complementary coalition $N\backslash S$ is paid his complete claim. Thus $\nu(s) = \max\{0, E - claim(N\backslash S)\}$. By focusing on the mysterious case in which the amount left $E = 200$, let 1, 2 and 3 be the wives that claim 100, 200 and 300, respectively; hence, $\nu(1) = \nu(2) = \nu(2) = 0$ because 23, 13 and 12's claims are bigger than 200 and thus absorb all the estate, $\nu(23) = 100$ which is the amount left by 1 once she takes her claim of 100, $\nu(12) = 0$ and $\nu(13) = 0$.

Letting x_i be the final payoff allocation for wife i, the optimization problem is:

$$\min \alpha \tag{A.8}$$
$$s.t. \quad x_1 + x_2 + x_3 = 200 \tag{A.9}$$
$$x_1 + x_2 + \alpha \geq 0 \tag{A.10}$$
$$x_1 + x_3 + \alpha \geq 0 \tag{A.11}$$
$$x_2 + x_3 + \alpha \geq 100 \tag{A.12}$$
$$x_1 + \alpha \geq 0 \tag{A.13}$$
$$x_2 + \alpha \geq 0 \tag{A.14}$$
$$x_3 + \alpha \geq 0 \tag{A.15}$$

The optimum is reached for $\alpha = -50$, but there are many possible allocations

$$50 \leq x_i \leq 150, \ \forall i \in \{1, 2, 3\} \tag{A.16}$$

Taking for example (50, 50, 100), an imputation satisfying (A.9)-(A.15), and $S_0 = \{(1), (2, 3)\}$, the next optimization is:

$$\min \alpha \tag{A.17}$$
$$s.t. \quad x_1 + x_2 + x_3 = 200 \tag{A.18}$$
$$x_1 + x_2 + \alpha \geq 0 \tag{A.19}$$
$$x_1 + x_3 + \alpha \geq 0 \tag{A.20}$$
$$x_2 + x_3 = 150 \tag{A.21}$$
$$x_1 = 50 \tag{A.22}$$
$$x_2 + \alpha \geq 0 \tag{A.23}$$
$$x_3 + \alpha \geq 0 \tag{A.24}$$

The optimum corresponds to $\alpha = -75$ and this time the solution is unique and equal to (50, 75, 75), the solution suggested by the Talmud.

Also for the cases $E = 100$ and $E = 300$, applying the same methodology the solutions correspond to the nucleolus.

A.5 Summary

This appendix depicted some principles of game theory, starting with classical assumptions moving through non-cooperative games and cooperative games. We voluntarily skipped historical facts and mathematical formalisms. Historical references to the discipline evolution would add precious elements for the complete knowledge of the discipline (for a chronology see [190]), and mathematical formalisms would complete given statements and definitions, clarifying and enforcing their meaning. Ideally, after the reading of this appendix one might discover the usefulness of game theory for a problem one has in mind, maybe with a no clear idea on how to apply the presented principles, but for this one should continue reading the report to understand specific applications, and possibly referring to reference books.

We did not mention many other important aspects of game theory (too specific for being considered as principles). First, repeated game modeling was omitted. A repeated game consists in some number of repetitions of a base stage game, which is usually one of the well studied 2 player games; it captures the idea that a player has to take into account the impact of his current action on the future actions of other players. Second, the mechanism design problem was not discussed; the problem is to correctly choose the game form to meet the expectations of the payers. For example, what kind of auction model should be chosen for UMTS or Wi-Max frequency auctions? Third, sequential games were not completely discussed, and the use of the extended form not technically explored. Fourth, evolutionary games were not treated. This class of games analyzes the dynamics of the interaction between players with limited rationality and intelligence. Finally, NTU games - generalization of the coalitional games (TU games) as we treated them in this chapter - were not considered: the complexity level of their mechanisms may not be useful for telecommunications.

Appendix B

Petaweb design: modeling details

B.1 Petaweb regular design heuristic details

We develop the matching costs for each block of the cost matrix (see Sect. 8.2.1).

Block 1

Matching two inactive core nodes. Let (i_1, r_1, e_1) be the s^{th} node of L_1, (i_2, r_2, e_2) the t^{th} node of L_1. The matching cost is $c_{s,t} = \infty$ if $s \neq t$ or 0 if $s = t$.

Block 2

Matching an unassigned edge node pair with an inactive core node. Let p be the s^{th} pair of L_2 with origin/destination (j,k) and (i, r, e) be the t^{th} core node of L_1. The matching is allowed if the link capacity between the origin j of the pair p and the core node (i, r, e) on the one hand, and the link capacity between the core node (i, r, e) and the destination k of the pair p on the other hand, are respected: $Q_p \leq C_{ch}\, W\, s_r$. If the capacity constraints are verified, the matching results in a new element $((i, r, e), D = \{p\})$ of L_3 whose cost is the sum of the cost of the core node plus the cost of the fiber between the core node (i, r, e) and all edge nodes in the network and the cost of the propagation delay of the pair p traffic via the core node (i, r, e): $\beta d_{ip} Q_p$. The matching cost for the block 2 is finally:

$$c_{n_1+s,t} = \begin{cases} 2\,|N|\,W\,s_r\,\gamma^{(s_r-1)}\,P + f_r + 2\,\phi(W)\,F\,s_r\,\sum_{j\in N}\delta_{ij} + \beta d_{ip} Q_p \\ \qquad \text{if } Q_p \leq C_{ch}\,W\,s_r, \\ \infty \quad \text{otherwise} \end{cases}$$

Block 3

Matching two unassigned edge node pairs. If the two pairs are different, the matching is impossible and the cost is set to infinity. If a pair is matched with itself, it remains unmatched. The cost is twice the cost of one unassigned pair because each matching cost must appear twice in the objective function. Let p_1 be the s^{th} unassigned edge node pair of L_2 and let p_2 be the t^{th} unassigned edge node pair of L_2. The matching cost for the block 3 is: $c_{n_1+s,n_1+t} = \begin{cases} \infty & \text{if } s \neq t, \\ 2\mathcal{M} & \text{if } s = t \end{cases}$

Block 6

Matching two kits of L_3. Let $((i_1, r_1, e_1), D_1)$ be the s^{th} kit of L_3 and $((i_2, r_2, e_2), D_2)$ be the t^{th} kit of L_3.

If $s = t$, the element is matched with itself. The matching cost is twice the cost of one element (as above). The self-matching cost is then:
$2\,(2|N|W\,s_{r_1}\gamma^{(s_{r_1}-1)}\,P + f_{r_1} + 2\,\phi(W)\,F\sum_{j\in N}\delta_{i_1 j} + \beta\sum_{p\in D_1} d_{i_1 p}\,Q_p)$

If $s \neq t$, three cases must be considered:

Case I: All edge node pairs of D_1 and D_2 are assigned to (i_1, r_1, e_1).

This case is possible if the link capacity between each origin edge node of D_1 and D_2 and the core node (i_1, r_1, e_1) on the one hand, and the link capacity between the core node (i_1, r_1, e_1) and each destination edge node of D_1 and D_2 on the other hand, are respected. The matching cost for this case is then:

$$v_I = \begin{cases} 2|N|W\,s_{r_1}\gamma^{(s_{r_1}-1)}\,P + f_{r_1} + 2\phi(W)Fs_{r_1}\sum_{j\in N}\delta_{i_1 j} + \beta\sum_{p\in(D_1\cup D_2)} d_{i_1 p}Q_p \\ \quad \text{if} \quad \sum_{p\in(D_1\cup D_2)\in Orig_j} Q_p \leq C_{ch}\,W\,s_{r_1}, \quad \forall\, origin\, j \in (D_1 \cup D_2) \\ \quad \text{and} \quad \sum_{p\in(D_1\cup D_2)\in Dest_k} Q_p \leq C_{ch}\,W\,s_{r_1}, \quad \forall\, destination\, k \in (D_1 \cup D_2) \\ \\ \infty \quad \text{otherwise} \end{cases}$$

Case II: All edge node pairs of D_1 and D_2 are assigned to (i_2, r_2, e_2).

As the previous case, reversing the roles of the core nodes. The cost is:

$$v_{II} = \begin{cases} 2|N|W\,s_{r_2}\gamma^{(s_{r_2}-1)}\,P + f_{r_2} + 2\phi(W)Fs_{r_2}\sum_{j\in N}\delta_{i_2 j} + \beta\sum_{p\in(D_1\cup D_2)} d_{i_2 p}Q_p \\ \quad \text{if} \quad \sum_{p\in(D_1\cup D_2)\in Orig_j} Q_p \leq C_{ch}\,W\,s_{r_2}, \quad \forall\, origin\, j \in (D_1 \cup D_2) \\ \quad \text{and} \quad \sum_{p\in(D_1\cup D_2)\in Dest_k} Q_p \leq C_{ch}Ws_{r_2}, \quad \forall\, destination\, k \in (D_1 \cup D_2) \\ \\ \infty \quad \text{otherwise} \end{cases}$$

Case III: The core nodes (i_1, r_1, e_1) and (i_2, r_2, e_2) are both active.

This is a difficult case because the core nodes may exchange some edge node pairs. We then need to find the optimal assignment of the pairs to the two core nodes. Let us define w_p as a binary variable so that $w_p = 1$ if the pair $p \in D_1$ swaps its current core node for (i_2, r_2, e_2) and $w_p = 0$ otherwise. Also let z_p be binary variable so that $z_p = 1$ if the pair $p \in D_2$ swaps its current core node for core node (i_1, r_1, e_1) and $z_p = 0$ otherwise. The swapping problem can be formulated as:

$$v = \min \sum_{p\in D_1} g_p w_p + \sum_{p\in D_2} h_p z_p \tag{B.1}$$

$$s.t. \sum_{p\in D_1}^{s(p)=j} -Q_p w_p + \sum_{p\in D_2\in Orig_j} Q_p z_p \leq \epsilon_{wj}, \quad \forall\, origin\, j \in (D_1 \cup D_2) \tag{B.2}$$

$$\sum_{p\in D_1}^{s(p)=j} Q_p w_p - \sum_{p\in D_2\in Orig_j} Q_p z_p \leq \epsilon_{zj}, \quad \forall\, origin\, j \in (D_1 \cup D_2) \tag{B.3}$$

$$\sum_{p\in D_1}^{t(p)=k} -Q_p w_p + \sum_{p\in D_2\in Dest_k} Q_p z_p \leq \eta_{wj}, \quad \forall\, destination\, k \in (D_1 \cup D_2) \tag{B.4}$$

$$\sum_{p\in D_1}^{t(p)=k} Q_p w_p - \sum_{p\in D_2\in Dest_k} Q_p z_p \leq \eta_{zj}, \quad \forall\, destination\, k \in (D_1 \cup D_2) \tag{B.5}$$

$$w_p, z_p \in \{0,1\} \tag{B.6}$$

g_p and h_p are the marginal costs if the edge node pair p exchanges its core node. ϵ_{wj} and ϵ_{zj} are the surplus capacities of the links between the origin edge node j and (i_1, r_1, e_1) and (i_2, r_2, e_2), respectively. idem η_{wk} and η_{zk} of the links between (i_1, r_1, e_1) and (i_2, r_2, e_2), respectively, and the destination edge node k.

$$\begin{aligned}
g_p &= \beta d_{i_2 p} Q_p - \beta d_{i_1 p} Q_p, \quad \forall p \in D_1 \\
h_p &= \beta d_{i_1 p} Q_p - \beta d_{i_2 p} Q_p, \quad \forall p \in D_2 \\
\epsilon_{wj} &= C_{ch}\, W\, s_{r_1} - \sum_{p\in D_1 \in Orig_j} Q_p, \quad \forall\, origin\, j \in (D_1 \cup D_2) \\
\epsilon_{zj} &= C_{ch}\, W\, s_{r_2} - \sum_{p\in D_2 \in Orig_j} Q_p, \quad \forall\, origin\, j \in (D_1 \cup D_2) \\
\eta_{wk} &= C_{ch}\, W\, s_{r_1} - \sum_{p\in D_1 \in Dest_k} Q_p, \quad \forall\, destination\, k \in (D_1 \cup D_2) \\
\eta_{zk} &= C_{ch}\, W\, s_{r_2} - \sum_{p\in D_2 \in Dest_k} Q_p, \quad \forall\, destination\, k (D_1 \cup D_2)
\end{aligned}$$

The objective (B.1) is to minimize the cost of the packing. (B.2) and (B.3) are surplus capacity constraints for the links between each origin edge node j and the core nodes (i_1, r_1, e_1) and (i_2, r_2, e_2) respectively; idem (B.4) and (B.5) are for the links between (i_1, r_1, e_1) and (i_2, r_2, e_2) and each destination edge node k. (B.6) indicates the variables w_p and z_p are binary. The matching cost is finally:

$$\begin{aligned}
v_{III} &= 2|N|W s_{r_1} \gamma^{(s_{r_1}-1)} P + f_{r_1} + 2\phi(W) F s_{r_1} \sum_{j\in N} \delta_{i_1 j} \\
&+ \beta \sum_{p\in D_1} d_{i_1 p} Q_p + 2|N|W\, s_{r_2} \gamma^{(s_{r_2}-1)}\, P + f_{r_2} \\
&+ 2\,\phi(W)\, F\, s_{r_2} \sum_{j\in N} \delta_{i_2 j} + \beta \sum_{p\in D_2} d_{i_2 p} Q_p + v
\end{aligned} \tag{B.7}$$

Among the three cases whenever $s \neq t$, we choose the solution with minimal cost: $\min\{v_I, v_{II}, v_{III}\}$. At last, the matching cost for the block 6 is:

$$c_{n_1+n_2+s, n_1+n_2+t} = \begin{cases} 2(2|N|W\, s_{r_1} \gamma^{(s_{r_1}-1)}\, P + f_{r_1} + 2\,\phi(W)\, F s_{r_1} \sum_{j\in N} \delta_{i_1 j} + \\ +\beta \sum_{p\in D_1} d_{i_1 p}\, Q_p) \quad \text{if } t = s \\ \min\{v_I, v_{II}, v_{III}\} \quad \text{otherwise} \end{cases}$$

where v_I, v_{II} and v_{III} are given by (B.1), (B.1) and (B.7), respectively.

Block 4

Matching a kit of L_3 with an inactive core node of L_1. This is a particular case of block 6. Let $((i_1, r_1, e_1), D_1)$ be the s^{th} kit of L_3 and (i, r, e) be the t^{th} core node of L_1. The inactive node (i, r, e) can be seen as an active one with no assigned pair: $(i, r, e) = ((i_2, r_2, e_2), \emptyset) \in L_3$. The cost is: $c_{n_1+n_2+s,t} = \min\{v_I, v_{II}, v_{III}\}$ where v_I, v_{II} and v_{III} are given by (B.1), (B.1) and (B.7).

Block 5

Matching a kit of L_3 with an unassigned pair of L_2. Let $((i_1, r_1, e_1), D_1)$ be the s^{th} kit of L_3 and q be the t^{th} pair of L_2 with origin/destination (j,k). Two cases must be considered:

Case I: The unassigned edge node pair can be assigned to (i_1, r_1, e_1).

Then D_1 becomes $D_1 \cup \{q\}$. This is possible if: the link capacity between the origin edge node j and (i_1, r_1, e_1) on the one hand, and the link capacity between (i_1, r_1, e_1) and the destination edge node k on the other hand, are respected. The matching cost for this case is (now $q \in D_1$):

$$c_{n_1+n_2+s,n_1+t} = \begin{cases} 2|N|W\, s_{r_1}\gamma^{(s_{r_1}-1)}\, P + f_{r_1} + 2\,\phi(W)Fs_{r_1} \sum_{j\in N} \delta_{ij} + \beta \sum_{p\in D_1} d_{ip}Q_p \\ \quad \text{if} \quad \sum_{p\in D_1 \in Orig_j} Q_p \le C_{ch} W s_{r_1}, \quad \forall\, origin\, j \in D_1 \\ \quad \text{and} \quad \sum_{p\in D_1 \in Dest_k} Q_p \le C_{ch} W s_{r_1}, \quad \forall\, destination\, k \in D_1 \end{cases}$$

Case II: The unassigned edge node pair can not be assigned to (i_1, r_1, e_1).

If one capacity constraint is not respected, one pair or more have to be removed from the kit. A problem of pair exchange is then solved as for the block 6. The pair q is inserted in D_1. D_2 becomes an empty set. $D_1 = D_1 \cup \{q\}$ and $D_2 = \emptyset$.

We solve (B.1)-(B.6) without considering (B.3) and (B.5) where now w_p is defined as being equal to 1 if the pair $p \in D_1$ is detached from (i_1, r_1, e_1) and becomes an unassigned pair of L_2, and 0, otherwise. Also, $z_p = 0$, $g_p = \mathcal{M} - \beta\, d_{i1p}\, Q_p, \forall p \in D_1$ $h_p = 0$. ϵ_{wj} and η_{wk} can be negative. The set $\overline{D_1} \subset D_1$ corresponds to the edge node pairs assigned to (i_1, r_1, e_1) in the exchange problem solution. Let $\overline{n_1}$ be the number of elements in $\overline{D_1}$. The matching cost for this case is then:

$$c_{n_1+n_2+s,n_1+t} = 2|N|Ws_{r_1}\gamma^{(s_{r_1}-1)}P + f_{r_1} + 2\phi(W)Fs_{r_1} \sum_{j\in N} \delta_{i_1 j} + \beta \sum_{p\in \overline{D_1}} d_{i_1 p}Q_p + (n_1 - \overline{n_1})\mathcal{M} \tag{B.8}$$

At last, the matching cost for the block 5 is:

$$c_{n_1+n_2+s,n_1+t} = \begin{cases} 2|N|W\, s_{r_1}\gamma^{(s_{r_1}-1)}\, P + f_{r_1} + 2\phi(W)Fs_{r_1} \sum_{j\in N} \delta_{i_1 j} + \beta \sum_{p\in D_1} d_{i_1 p}Q_p \\ \quad \text{if} \quad \sum_{p\in D_1 \in Orig_j} Q_p \le C_{ch}\, W\, s_{r_1}, \quad \forall origin\, j \in D_1 \\ \quad \text{and} \quad \sum_{p\in D \in Dest_k} Q_p \le C_{ch}\, W\, s_{r_1}, \quad \forall destination\, k \in D_1 \\ \\ 2|N|W\, s_{r_1}\gamma^{(s_{r_1}-1)}\, P + f_{r_1} + 2\phi(W)Fs_{r_1} \sum_{j\in N} \delta_{i_1 j} + \beta \sum_{p\in \overline{D_1}} d_{i_1 p}Q_p + \\ +(n_1 - \overline{n_1})\mathcal{M} \quad \text{otherwise} \end{cases}$$

B.2 Regular Petaweb design with TDM/WDM

Let us complete the notation introduced in Chapter 8 with:

$F_{i,r}$: Total fiber installation cost due to the placement of a CN-r at site i

O_j/D_k: a subset of T with a fixed origin/destination site j

E_r: number of CN-r specimens that can be enabled in a site

H: set of TLPs classes

m_h: maximal number of TLP-h specimens for a CR, $h \in H$

(p, h, l): triple representing a TLP specimen, $p \in T, h \in H, 1 \leq l \leq m_h$

x^{ire}_{phl}: indicates if l^{th} TLP-h specimen of CR p exists and is switched by (i, r, e)

The ILP formulation for the RFA problem is:

$$\min G(\overline{y}, \overline{x}) \quad = \quad \sum_{(i,r,e)} (K_r + F_{i,r})\, y_{ire} + \sum_{(i,r,e)} \sum_{(p,h,l)} \beta\, d_{ip}\, Z_h\, x^{ire}_{phl} \tag{B.9}$$

$$s.t. \quad \sum_{r \in V} \sum_{e=1}^{E_r} x^{ire}_{ph_1l_1} = \sum_{r \in V} \sum_{e=1}^{E_r} x^{ire}_{ph_2l_2} \tag{B.10}$$

$\forall i \in M,\ \forall p \in T,\ \forall h_1, h_2 \in H,\ \forall l_1 | \, 1 \leq l_1 \leq m_{h_1},\ \ \forall l_2 | \, 1 \leq l_2 \leq m_{h_2},\ (h_1, l_1) \neq (h_2, l_2)$

$$\sum_{(i,r,e)} x^{ire}_{phl} \ = \ 1, \quad \forall (p, h, l) \tag{B.11}$$

$$\sum_{(i,r,e)} C_{ch} W s_r\, y_{ire} \leq C_j, \quad \forall j \in N \tag{B.12}$$

$$\sum_{(p \in O_j, h, l)} Z_h\, x^{ire}_{phl} \leq C_{ch} W s_r\, y_{ire}, \quad \forall j \in N \forall (i, r, e) \tag{B.13}$$

$$\sum_{(p \in D_k, h, l)} Z_h\, x^{ire}_{phl} \leq C_{ch} W s_r\, y_{ire}, \quad \forall k \in N \forall (i, r, e) \tag{B.14}$$

$$y_{ire} \in \{0, 1\}\,,\ x^{ire}_{phl} \in \{0, 1\} \tag{B.15}$$

The objective (B.9) is to minimise the total network cost, sum of the costs of all the enabled core nodes, of the optical fibers between the edge nodes and the enabled core nodes, and of the propagation delays cumulated by the TLPs;
(B.10) expresses the 1^{st} Coherency Constraint: two TLPs associated to the same CR must be switched by CNs in the same site and, thus, transported on the same optical trunk line. Given a switching site and a CR, for every couple (TLP_1,TLP_2) of the CR, with $TLP_1 \neq TLP_2$, the number of CNs in that site switching TLP_1 must be equal to the number of CNs switching TLP_2;
(B.11) expresses the 2^{nd} Coherency Constraint: the traffic transported through a TLP, from an origin EN to a destination EN, must be entirely switched in the same CN and, thus, transported in the same optical links. For every TLP, the number of CNs switching that TLP must be equal to 1;
(B.12) imposes the respect of the capacity constraint for every EN;
(B.13), (B.14) impose the respect of the capacity constraint for the optical links between every origin EN and every CN, and between every destination EN and every CN;
(B.15) indicates that x^{ire}_{phl} and y_{ire} are binary variables.

B.3 Regular Petaweb design with DPP

We present two different yet equivalent ILP formulation for the RFA problem resolution with dedicated path protection (and TDM/WDM), the first more straightforward than the second but more complex.

Refined formulation with DPP

Let us complete the notation introduced in Chapter 8 and above with:

wTLP/pTLP: working/protection TLP

K_r: global cost for a core node of type r

m_h: maximum number of wTLPs of type h for a single pair

$(p, h, l + m_h)$: triple identifying uniquely the pTLP of the wTLP (p, h, l)

δ: scaling factor for the pTLPs' propagation delay cost, $0 \leq \delta \leq 1$

Ω_w: set of all the wTLPs, $p \in T$, $h \in H$ and $0 < l \leq m_h$

Ω_p: set of all the pTLPs, $m_h < l \leq 2m_h$ – Ω: set of all the TLPs, $0 < l \leq 2m_h$

The ILP formulation for the RFA problem becomes:

$$\begin{aligned} \min G(\overline{y}, \overline{x}) \;&=\; \sum_{(i,r,e)} (K_r + F_{i,r})\, y_{ire} \\ &+\; \sum_{(i,r,e)} \sum_{(\Omega_w} \beta\, d_{ip}\, Z_h\, x^{ire}_{phl} + \sum_{(i,r,e)} \sum_{(\Omega_p} \delta\, \beta\, d_{ip}\, Z_h\, x^{ire}_{phl} \end{aligned} \tag{B.16}$$

$$s.t. \sum_{r \in V} \sum_{e=1}^{E_r} x^{ire}_{phl} + \sum_{r \in V} \sum_{e=1}^{E_r} x^{ire}_{phl_p} \leq 1, \quad \forall i \in M, \forall (p, h, l) \in \Omega_w, l_p = l + m_h \tag{B.17}$$

$$\sum_{(i,r,e)} x^{ire}_{phl} \;=\; 1, \quad \forall (p, h, l) \in \Omega \tag{B.18}$$

$$\sum_{(i,r,e)} C_{ch}\, W\, s_r\, y_{ire} \;\leq\; C_j, \quad \forall j \in N \tag{B.19}$$

$$\sum_{(p \in O_j, h, l) \in \Omega} Z_h\, x^{ire}_{phl} \;\leq\; C_{ch}\, W\, s_r\, y_{ire}, \quad \forall j \in N, \forall (i, r, e) \tag{B.20}$$

$$\sum_{(p \in D_k, h, l) \in \Omega} Z_h\, x^{ire}_{phl} \;\leq\; C_{ch}\, W\, s_r\, y_{ire}, \quad \forall k \in N, \forall (i, r, e) \tag{B.21}$$

$$x^{ire}_{phl} \in \{0, 1\}, y_{ire} \in \{0, 1\} \tag{B.22}$$

The objective (B.16) is to minimize the total network cost. The network cost is linearly computed as sum of the costs of all the enabled core nodes, of the optical fibers between the edge nodes and the enabled core nodes, and of the propagation delays cumulated by the TLPs; the second and the third term are the cost of the propagation delays for, respectively, the wTLPs and the pTLPs. The 1^{st} Integrity Constraint (Sect. 8.4.1) is implicitly partially considered by the propagation delay minimization that drives the TLPs of a same CR to follow the same route.

(B.17) ensures that the path protection constraint is respected.. The pTLP and the wTLP for the same connection request can not be routed on the same trunk line, i.e., can not be switched on the same switching site. Given a switching site and a wTLP, the number of enabled CNs switching that wTLP and the corresponding pTLP in that site must be lower or equal to 1.

(B.18) expresses the 2^{nd} Integrity Constraint: the traffic transported through a TLP, from an origin EN to a destination EN, must be entirely switched in the same CN and, thus, transported in the same optical links. For every TLP, the number of CNs switching that TLP must be equal to 1.

(B.19) ensures that the capacity constraint is respected at every EN.

(B.20) and (B.21) ensure that the capacity constraint on the optical links between every origin EN and every CN, and between every destination EN and every CN, is respected. (B.22) sets the binarity constraint on the variables.

The ILP complexity is hard, but not prohibitive for the assigned instances. By grooming end-to-end TLP-1s and TLP-2s of the same CR in virtual sub-classes consisting of, respectively, 4 time-slots and 4 wavelengths, we could control the number of TLPs, and thus the number of variables and constraints.

Note: fiber cost function $F_{i,r}$

We excluded in the simulations the cases of leased fiber lines in the design dimensioning. For the purposes of performance evaluation, a particular form of the cost function has been used. Since the installation requires s_r fibers per direction and per optical link, we set $F_{i,r} = \sum_j c(i,j,r) = 2\,\phi(W)\,F\,s_r \sum_j \Delta_{ij}$, where Δ_{ij} is the distance between sites i and j, and F is the cost of a single-wavelength fibre, which is then scaled by a discrete function $\phi(W)$ that depends on the number of wavelengths. We adopted $\phi(W) = W$ considering, thus, that the cost of a fiber is proportional to the number of wavelengths.

It is worth noting that the chosen form of $F_{i,r}$ is a theoretical modelling choice, since in the reality the fiber cost is not merely proportional to a unitary cost. As previously mentioned, there are trenching costs, amplifiers, regenerators, cross connect charges, etc.. However, when the geographical distances between network sites are very high, and of the same order of magnitude (as for our performance study cases, often hundreds of 100 km), it is acceptable to approximate the end-to-end fiber cost with $F_{i,r}$. In fact, point costs as for amplifiers, regenerators, and edge costs as for fiber trenching and cross connect, can be considered imputing an additional (quite small) constant cost contribution to a length-dependent unitary cost – mathematically, when the distance between sites present a small standard deviation from the average.

B.3.1 Relaxed formulation

The addition of the protection constraint (B.17) may lead to a too combinatorial problem for large instances. We propose in the following an alternative approach that, by relaxing some modeling features, offers a lower computational complexity. We start by designing the least expensive network which can satisfy the CRs; then, when the CNs are located and the CRs are assigned to the switching sites, we decide how to split each CR into the minimum number of TLPs.

Let us complete the notation introduced in Chapter 8 and above with:

$w(q)$: the bandwidth consumption of CR q.

$\delta < 1$: the scaling factor for pTLPs delay costs.

d_{iq}: the distance from the origin of CR q to site i plus the distance from site i to the destination of CR q.

$C_r = (Z_2/Z_1)W s_r$: capacity of a core node of type r.

a_q^i,$\widetilde{a_q^i}$: binary variable equal to 1 if CR q is switched at site i; $\widetilde{a_q^i}$ relates to the corresponding protection flow.

y_{ir}: integer variable equal to the number of core nodes of type r to install at site i.

All the capacities are measured in time slots; these terms are computed by dividing each value by Z_1 and rounding down to the nearest integer. In a similar way, each bandwidth consumption value $w(q)$ is measured in time slots, dividing by Z_1 and rounding up to the nearest integer.

The problem of finding the optimal regular Petaweb can be stated as follows:

$$\min G(y,a,\widetilde{a}) = \sum_{i\in N}\sum_{r\in V}(K_r + F_{ir})y_{ir} +$$
$$+ \sum_{i\in N}\sum_{q\in Q}\beta d_{iq}w(q)Z_1 a_q^i + \sum_{i\in N}\sum_{q\in Q}\delta\beta d_{iq}w(q)Z_1\widetilde{a_q^i} \tag{B.23}$$

$$s.t. \sum_{i\in N} a_q^i = 1, \quad \forall q \in Q \tag{B.24}$$

$$\sum_{i\in N}\widetilde{a_q^i} = 1, \quad \forall q \in Q \tag{B.25}$$

$$a_q^i + \widetilde{a_q^i} \le 1, \quad \forall q \in Q,\ \forall i \in N \tag{B.26}$$

$$\sum_{q|\text{head of } q=j} w(q)(a_q^i + \widetilde{a_q^i}) \le \sum_{r\in V} C_r y_{ir}, \quad \forall i \in N, \forall j \in N \tag{B.27}$$

$$\sum_{q|\text{tail of } q=j} w(q)(a_q^i + \widetilde{a_q^i}) \le \sum_{r\in V} C_r y_{ir}, \quad \forall i \in N, \forall j \in N \tag{B.28}$$

$$\sum_{i\in N}\sum_{r\in V} C_r y_{ir} \le \max C_j \tag{B.29}$$

$$y_{ir} \in Z_+, a_q^i, \widetilde{a_q^i} \in \{0,1\} \tag{B.30}$$

The objective (B.23) is to minimize the network cost; this is composed of the core node cost, the fiber cost, and a term accounting for the propagation delay of wTLPs and pTLPs, respectively. Constraints (B.24) and (B.25) impose a unique switching and protection site on each CR q, while constraints (B.26) ensure that working and protection paths of each CR are disjoint. Constraints (B.27) and (B.28) limit the number of optical links ending to and starting from a given switching site to the number of installed ports. Inequality (B.29) imposes a maximum capacity on each edge node, while conditions (B.30) restrict the variables to take integer and binary values.

B.3.2 Comparison

In Table B.1 we report the CPU time needed for optimizing the regular topology with the first formulation (column 'First model') and with the relaxed model (column 'Relaxed model'), for each instance (whose id is indicated in the first entry of each row). As expected, the computational effort required with the new procedure drastically decreases thanks to the TLPs post-selection.

Table B.1: Results comparison for the regular Petaweb design

Instance	First model exec. time (s)	Relaxed model exec. time (s)
10A	2768	37
10B	2330	5
34A	214524	50834
34B	223560	3961

It is worth remarking that this new method requires that once the optimal design is obtained, the problem of splitting the traffic of each CR into TLPs has to be solved, and these TLPs have to be assigned to time-slots, wavelengths and fibers. The optimal solution may require to split more TLP-3s and TLP-2s than with the first formulation in order to fully exploit the available capacity. From the computational point of view, this involves solving a set of *bin packing problems with item fragmentation* [51]. However, this proved not to be an issue. In fact, the size of the instances for these subproblems is typically very small, and their optimization can easily be tackled by general purpose solvers.

B.4 Petaweb upgrade problem formulation

In the following we present the ILP formulation for the Petaweb upgrade. The logic of the upgrade model is to use the same type of objectives and constraints of Sect. B.3 but forcing the design to keep the existing equipment, and altering the available capacity on each single link so that the media already in use is not considered to route the new traffic.

The existing network is identified by all the enabled core nodes, the set of TLPs they commute, and by the number of fibers per optical link. From these, the used and the available transport capacity can be extracted and considered in the capacity constraints. Regarding the objective, it is worth noting that the optimization will be carried aiming at the minimization of the current total equipment costs. Thus, to assess the cost of the update, the cost of the equipment already installed will be subtracted.

The set of new TLPs (p, h, l) identifies the additional traffic volume, and the set M comprehends the pre-existing ENs sites and the new ones, if any. Thus, the solution is an optimized network with a regular or a quasi-regular topology, it indicates where the new TLPs must be routed and the equipment that have to be installed to satisfy the additional traffic.

Let us complete the notation introduced in Chapter 8 and above with:

χ: set of core nodes of the existing optimized network

Q_j^{ire}: pre-used capacity from site j to the existing CN (i, r, e) if $(i, r, e) \in \chi$

Q_{ire}^k: pre-used capacity from the existing CN (i, r, e) to site k if $(i, r, e) \in \chi$

The mathematical formulation of the problem is the following.

$$\begin{aligned} \min G(\overline{y}, \overline{x}) \quad &= \quad \sum_{(i,r,e)} \left(K_r + F_{i,r}\right) y_{ire} \\ &+ \quad \sum_{(i,r,e)} \sum_{(p,h,l)\in\Omega_w} \beta d_{ip} Z_h x_{phl}^{ire} + \sum_{(i,r,e)} \sum_{(p,h,l)\in\Omega_p} \delta\beta d_{ip} Z_h x_{phl}^{ire} \end{aligned} \tag{B.31}$$

$$s.t. \;\; y_{ire} = 1, \quad \forall (i,r,e) \in \chi \tag{B.32}$$

$$\sum_{r \in V} \sum_{e=1}^{E_r} x_{phl}^{ire} + \sum_{r \in V} \sum_{e=1}^{E_r} x_{phl_p}^{ire} \leq 1 \;\; \forall i \in N, \forall (p,h,l) \in \Omega_w, l_p = l + m_h \tag{B.33}$$

$$\sum_{(i,r,e)} x_{phl}^{ire} \; = \; 1, \quad \forall (p,h,l) \in \Omega \tag{B.34}$$

$$\sum_{(i,r,e)} C_{ch} \, W \, s_r \, y_{ire} \; \leq \; C_j, \quad \forall j \in N \tag{B.35}$$

$$\sum_{(p \in O_j, h, l) \in \Omega} Z_h x_{phl}^{ire} \leq \left(C_{ch} W s_r - Q_j^{ire} \right) y_{ire}, \quad \forall j \in N, \forall (i,r,e) \tag{B.36}$$

$$\sum_{(p \in D_k, h, l) \in \Omega} Z_h x_{phl}^{ire} \leq \left(C_{ch} W s_r - Q_{ire}^{k} \right) y_{ire}, \quad \forall k \in N, \forall (i,r,e) \tag{B.37}$$

$$x_{phl}^{ire}, y_{ire} \in \{0,1\} \tag{B.38}$$

The objective (B.31) includes the cost of switches and fiber plus two cost terms to account for propagation delay: one for the pTLPs and one for the wTLPs. Note that the two terms are weighted differently to avoid that the pTLP and its corresponding wTLP contend for the same shortest path. (B.32) imposes the enabling of the existing core nodes; (B.33) is the protection constraint; (B.34) insures that a TLP must be switched only by one CN; (B.35) enforces EN capacity constraint; (B.36) and (B.37) impose the capacity constraints on the idle capacity for the optical links going from every core node and every edge node, and vice versa, subtracting the already occupied transport capacity; (B.38) defines the binary domain of the variables.

As it was already mentioned, the upgrade cost is obtained subtracting from the final objective value the equivalent cost of the pre-existing network. Also mentioned was the fact that the upgrade aims at a regular topology. Then if the initial topology was quasi-regular and the planner intends the update to keep a quasi-regular structure, the quasi-regular topology can be extracted from the regular one.

To extract the quasi-regular topology one proceeds taking into account every optical link in the optimized regular network, looking for how much of its fibers would be used by the TLPs routed there, and disabling those fibers that would not be used at all. So, a whole optical link may be disabled in the quasi-regular topology, and, also, a whole trunk line may be disabled (e.g., see Fig.9.10). Moreover, even the ports associated to the disabled fibers are not considered in the quasi-regular architecture. Hence the cost reduction concerns the cost of unused fibers and ports. Note that the TLPs remain associated to the same core node than in the regular topology and that the routes are not affected by the disabling of fibers and ports. Also note that the upgraded quasi-regular solution is not the optimal one since it is still obtained by downgrading the optimal regular one..

Appendix C

Evaluation of Waveband grouping in WDM network dimensioning

This appendix studies the performance of different WaveBand Switching (WBS) schemes[1]. In particular, we study the problem of designing a minimal cost WBS backbone serving traffic volumes of the order of the terabit-per-second, where the objective is to dimension both the switching core nodes and the physical links. In order to comply with the strict availability requirements of most of today applications, we allocate resources with a dedicated path protection strategy. The design objective is the minimization of the network cost, composed of fiber cost, port costs and lightpath propagation delay cost[2].

C.1 Related work

For WBS networks, the most studied switching node architecture is the single-layer multi-granularity OXC node (MG-OXC) described in [142]. This node architecture guarantees a better signal quality and a smaller ports number with respect to other node architectures. The authors in [142] delineate the WBS benefits and compare the architecture of multi-layer and single-layer MG-OXC showing the ports allocation in the two cases.

In [143] authors analyze how the traffic increase can be accommodated over an existing network having only OXC nodes, without considering fiber dimensioning and thus with a blocking switching node. Given the considered single-layer MG-OXC structure, the bandwidth unit to be switched can be only a wavelength, a waveband or a fiber, and the existing wavelength cross-connect part can not be over-dimensioned. In [144] authors assume a hierarchical optical node able to perform waveband and wavelength switching: only two kinds of ports are taken into account, wavebands switch and wavelength switch ports, and no fiber cross-connection is considered. In [145] authors propose an optimization method assuming a multi-layer MG-OXC performing fiber, waveband and wavelength conversion. Originally, they consider wavebands that group lambda-channels with the same destination; this represents a strict constraint for wavebanding application.

[1] The contents and results presented in this appendix are also presented in [6].

[2] Usually in mesh backbones design this last factor is not considered, or at most the number of hops is minimized. This may result in an inaccurate choice since a little hops number does not imply a short travelled distance; moreover longer paths have larger signal attenuation, connection and splitting losses and generally a worse OSNR value.

In [146] authors propose an hybrid switching architecture employing both all-optical and electrical fabrics for performing, respectively, all-optical waveband switching and TDM switching. That creates a very large design instance that they could solve only for small networks with a low number of wavelengths.

Summarizing, the following grouping strategies can be employed [142]:

- *End-to-end grouping*: with same source-destination pair [143, 144, 146];
- *One-end grouping*: lightpaths with same source or destination [145];
- *Subpath grouping*: grouping of lightpaths with a common subpath.

The application of subpath grouping seems the best choice, even if the complexity for medium networks may be very quite. On the other hand, end-to-end and one-end grouping do not allow to fully exploit the wavebanding capabilities.

C.2 Traffic and network model

The physical topology is characterized by a set of nodes identifying the sites where switching equipment is installed, and by a set of unidirectional interconnection arcs between those nodes. The number of fibers per arc is not pre-defined and has to be dimensioned. We call a set of fibers in the same direction and on the same arc an optical link. The fibers are equipped with $W = 16$ wavelengths, each with a channel capacity of $C_{ch} = 10$ Gb/s.

The traffic matrix consists of static Connection Requests (CRs) between two nodes whose bit-rate represents an aggregated flow, coming, e.g., from SDH rings at the concentration level. The CR traffic is determined through a gravitational model [4]: the traffic between two sites is directly proportional to the product of the population values of the metropolitan areas [191], and inversely proportional to the square of the distance between the sites; the resulting value is opportunely scaled to obtain a global volume in the order of the Tb/s as described in Sect. C.5. Thus, the traffic matrix is symmetric and the virtual topology is fully meshed.

We assume three main bit-rate classes for lightpaths; a CR is mapped over one ore more lightpaths of different bit-rate classes in order to obtain the best rounding-up value. So, the gap between the required transport capacity and the allocated capacity will be as little as possible. Let us indicate with Z_h the bit-rate of a lightpath (LP) of class h, $h \in H = \{1, 2, 3\}$; the number of lightpaths of class h (LP-h) per CR is the minimum possible to accommodate the CR traffic. We assume $Z_1 = C_{ch} = 10$ Gb/s for a LP-1, which has the transport capacity of a wavelength, $Z_2 = 4Z_1 = 40$ Gb/s for a LP-2, which has the transport capacity of a waveband grouping $R = 4$ wavelengths, and $Z_3 = WC_{ch} = 160$ Gb/s for a LP-3, which have the transport capacity of a fiber (also called from now on *multi-waveband*). The three main classes use thus as transport unit, respectively, the wavelength, the waveband and the fiber. Considering $CR_1 = 195$ Gb/s and $CR_2 = 30$ Gb/s, e.g., then CR_1 would be mapped onto 1 LP-3 and 1 LP-2 (instead of 1 LP-3 and 4 LP-1), and CR_2 onto 3 LP-1 (instead than 1 LP-2). The mapping of the traffic matrix over the bit-rate classes creates the set of LPs to be routed. The classes correspond with the bit-rates of SDH and OTN interfaces [185].

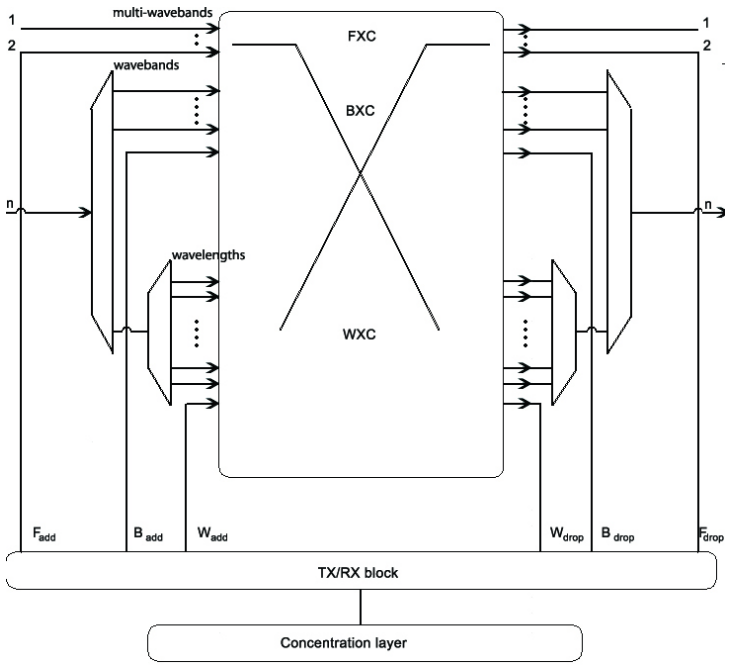


Figure C.1: Single-layer MG-OXC

C.3 Node architecture

We adopt the single-layer MG-OXC described in [142] and illustrated in Fig. C.1; it has one common optical switching fabric, which includes three logical parts: fiber Cross-Connector (FXC), waveBand CrossConnector (BXC) and Wavelengths CrossConnector (WXC); we will suppose full cross-connection features and, thus, no continuity constraint. At the input interface demultiplexing operations are performed and local fibers, wavebands or wavelengths are added, while at the output interface multiplexing and dropping are performed.

The incoming fibers that bypass the local site can be directly switched through FXC, without any wavelength and waveband demultiplexing; each of these bypassed fibers requires two ports, one at the input and one at the output stage. Also locally dropped fibers require two ports each, one port at input stage and one drop-port at output stage. The remaining incoming fibers transport one or more wavelengths or wavebands to be either bypassed or dropped at local site. These fibers are firstly demultiplexed in their wavebands. Then, the bypassed wavebands (those grouping wavelengths that are dropped at local site) can be directly switched by BXC and require two ports each. Also locally dropped wavebands require two ports each, one at input stage and a drop-port at output stage.

The remaining wavebands transport one or more wavelengths to be either bypassed or dropped at local site. Again, both the wavelengths bypassed through the WXC and the wavelengths dropped at the WXC output require two ports each. The same mechanism applies for locally added fibres, wavebands and wavelengths: one add-port at input stage and one port at output stage.

It is worth noting that wavelength continuity constraint strongly limits the benefits of wavebanding: as a matter of fact, in most of the previous works on wavebanding this constraint has not been included in the model. Indeed, the wavelength continuity constraint would implicitly imply even a waveband continuity constraint, decreasing the

chance for efficient subpath wavebanding. The application of these constraints would produce an over-dimensioned backbone with many unused wavelengths and wavebands, low network utilization and large amount of idle capacity. Moreover, the application of wavelength continuity constraint would produce more complex optimization problems as the authors pointed out in [147].

C.3.1 End-to-end grouping example

Let us consider a simple example to clarify the node architecture and to show how we can perform port allocation by end-to-end wavebanding. Consider the case that at a given site we have six fibers of sixteen wavelengths entering the site, and six exiting it. If we would use basic OXCs, every one of the input/output fibers would be demultiplexed/multiplexed in/from all its wavelengths. In this case, the OXC contains $6 \times 16 \times 2 = 192$ ports. Suppose now that the six entering fibers transport 4 LP-3, 7 LP-2 and 4 LP-1, among which 1 LP-3, 1 LP-2 and 1 LP-1 must be dropped, and 1 LP-3, 1 LP-2 and 1 LP-1 must be added. Thus, both the incoming and the outgoing fibers are fully used; the 4 LP-3 occupy one fiber each, and the 7 LP-2 and the 4 LP-1 are transported through 2 fibers. The required input ports for the FXC are, thus, $4+1=5$. Two fibers are demultiplexed into wavebands; between the 8 demultiplexed wavebands, 7 (the LP-2) are switched through a BXC with $7 + 1$ ports (1 port for the locally added LP-2). The eighth waveband is demultiplexed in four wavelengths, which are switched through WXC with $4 + 1$ ports (1 port for locally added LP-1). At the output interface the number of ports is the same than at the input interface; we have thus $4+7+4=15$ ports for bypass LPs and $1 + 1 + 1 = 3$ ports for the dropped LPs. Globally, we need a total of 36 ports, that is exactly twice as much the number of bypassed, dropped or added LPs, while with a OXC we would need 192 ports.

Addition of subpath grouping

Let us now consider the case that the six fibers transport 1 LP-3, 19 LP-2 and 4 LP-1 (dropped LPs remain 1 LP-3, 1 LP-2 and 1 LP-1). With a MG-OXC we would need $2 \times (1 + 19 + 4 + 1 + 1 + 1) = 54$ ports instead of the 36. Similarly, if we have 4 LP-3, 5 LP-2 and 12 LP-1 we need $2 \times (4+5+12+1+1+1) = 42$ ports instead of 36. A possible method to further reduce the ports number is to switch, as a single entity, groups of LPs that bypass a switching node even if they do not belong to the same end-to-end path. The objective is to group, where possible, those LPs of type 1 and 2 that bypass a node into, respectively, wavebands and *multi-wavebands*: e.g., 4 LP-1 can be grouped into one waveband and similarly 4 LP-2 can be grouped into one multi-waveband. This new functionality allows to apply the so called *subpath grouping*; it requires to identify for every hop all the LPs that bypass it and could form the *local wavebands* and *local multi-wavebands*. Moreover, consider the case in which at a given site there are 2 LP-2 that have not been grouped in multi-waveband since no associable LP-2 were available, and 2 wavebands created grouping LP-1. These entities can further be grouped in what we call an *heterogeneous local multi-waveband*.

C.3.2 Switching hierarchy

An agreed taxonomy is needed to identify the multi-granular entities that are switched in a MG-OXC. We employ the following classification (see Fig.C.2):

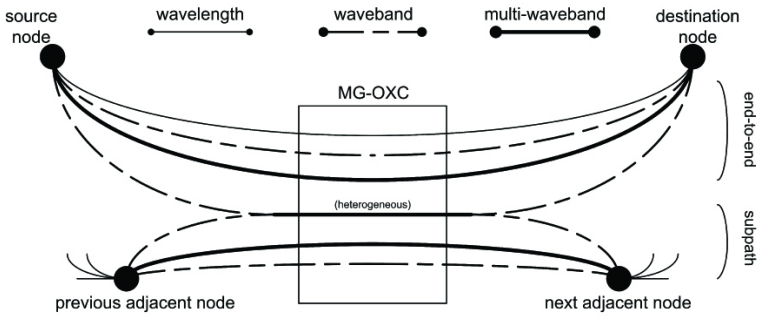


Figure C.2: Multi-granular channel entities

- **End-to-end wavelength**: a wavelength used by an end-to-end lightpath of the first bit-rate class (LP-1).
- **Waveband**: a set of (let us say R) wavelengths grouped as
  - *End-to-end waveband*: end-to-end LP of the second class (LP-2);
  - *Local waveband*: set of R end-to-end LPs (LP-1) bypassing a site.
- **Multi-waveband**: set of R wavebands grouped as
  - *End-to-end multi-waveband*: end-to-end LP of the third class (LP-3);
  - *Local multi-waveband*: group of R end-to-end wavebands bypassing a switching site;
  - *Heterogeneous local multi-waveband*: group of end-to-end wavebands *and* local wavebands.

In the following, we will model and analyze the following four cases:

OXC: basic OXC without any wavebanding;

MG-OXCs: MG-OXC with end-to-end wavebanding and multi-wavebanding;

MG-OXCh: local wavebanding and multi-wavebanding added to MG-OXCs;

MG-OXCc: heterogeneous local multi-wavebanding added to MG-OXCh.

C.4 Design dimensioning optimization

The purpose of our ILP model is to opportunely allocate resources (fibers, ports) and assign them to lightpaths in order to guarantee fast communications (i.e., with low propagation delay); also, it returns the optimal lightpaths routes and the wavebands and multi-wavebands that have to be created locally and end-to-end.

In other works on WBS, the usual design objective contains only the global ports number. We exploit a more refined cost function that captures the global network cost as a combination of distinct cost contributes: the cost of the fibers to be install and the cost of ports at switching nodes (real costs), and the cost due to propagation delays (virtual cost). The switch fixed cost is not considered because it has been considered negligible with respect to port cost. The propagation delay unitary cost of a LP is directly proportional to its bit-rate and to the travelled distance. Note that the propagation delay

cost is a virtual cost and it does not represent the real propagation delay; it is employed to allow the assignment of short routes to lightpaths, giving the priority to lightpaths with high bit-rates: consider that a crucial issue to be solved for selling VoIP and Video services with high QoS is to guarantee the lower possible delay in the transmission systems; such services should require LPs of high classes in our model, and these must have priority in getting the bandwidth over shorter paths.

We adopt a dedicated path protection strategy, as in Sect. 8.5 and B.3. In case of 1:1 dedicated path protection it makes sense to enable a shorter path for working lightpaths (wLPs), and a longer path to protection lightpaths (pLPs) to be used in case of failure along the working one.

In the following we discuss a set of ILP formulations to solve the design problem with wavebanding according to network model, constraints and node architecture proposed. The input data are the set of LPs to be routed and the pre-assigned physical topology. The outcome of the optimization is the number of fibers per link to install, the number of ports needed on each switching node and the assignment for each LP to the unidirectional optical links to be traversed. Let us complete the notation introduced in Chap. 8 and B.3 with:

$G(N, \epsilon_P)$: topology oriented graph; N is the set of nodes, ϵ_P the set of arcs

$(i,j) \in \epsilon_P$: arc, or optical link, between the node i and the node j

R: number of wavelengths/wavebands forming a waveband/multi-waveband; we set $W = R^2$ to have $Z_h = R\, Z_{h-1}$, $h < 3$. We will use $R = 4$.

$x_{i,j}^{p,h,l}$: indicates if the LP (p,h,l) passes over the link (i,j)

$l_{i,j}$: number of fibers to install on the link (i,j)

$$\begin{aligned} \min G(\overline{x}) \;=\; & \sum_{\Omega_w} \sum_{(i,j)} \beta\, Z_h\, d_{i,j}\, x_{i,j}^{p,h,l} + \sum_{\Omega_p} \sum_{(i,j)} \sigma\, \beta\, Z_h\, d_{i,j}\, x_{i,j}^{p,h,l} \\ & + \sum_{(i,j)} W\, F\, d_{i,j}\, l_{i,j} \;+\; P(\overline{x}) \end{aligned} \tag{C.1}$$

$$s.t. \quad \sum_{j \in N} x_{i,j}^{p,h,l} - \sum_{j \in N} x_{j,i}^{p,h,l} = \begin{cases} 1 & \text{if } i = s(p) \\ -1 & \text{if } i = t(p) \\ 0 & \text{otherwise} \end{cases} \qquad \forall i \in N,\, \forall (p,h,l) \in \Omega \tag{C.2}$$

$$x_{i,j}^{p,h,l} \;+\; x_{i,j}^{p,h,l+m_h} \;\leq\; 1, \quad \forall (i,j), \forall (p,h,l) \in \Omega_w \tag{C.3}$$

$$\sum_{\Omega} Z_h\, x_{i,j}^{p,h,l} \;\leq\; W\, C_{ch}\, l_{i,j}, \quad \forall (i,j) \tag{C.4}$$

$$x_{i,j}^{p,h,l} \;\in\; \{0,1\}; \quad l_{i,j} \in \mathbf{N} \tag{C.5}$$

(C.1) expresses the minimization of the total network cost due to propagation delays, fibers and switching ports ($P(\overline{x})$ varies for the different cases). The propagation delay cost associated to the pLPs is scaled by σ to avoid competition for the best path between a wLP and its pLP. (C.2) is the traffic conservation constraint, imposing that the flow leaving node i is balanced by the entering flow, except for the source (destination) node. (C.3) imposes the protection constraint, which requires that a pLP can not be routed

on a same link where the correspondent wLP is routed. (C.4) imposes the capacity constraint for every link of the network: the global bit-rate of the LPs traversing a link (i, j) must be minor to the capacity offered by the $l_{i,j}$ fibers to allocate on that link. (C.5) imposes the binary constraint for x, and the integer constraint for l.

$P(x)$ changes according to the scenarios of Sect. C.3.2 as in the following. Note that in case of locally added or dropped LPs, an additional add/drop port is not needed, it is substituted by an ingress/egress port.

OXC

If we assume a switching node composed of OXCs with no wavebanding functionalities, then every fiber has to be demultiplexed in all its tributary wavelength signals and thus requires $2W$ ports of unitary cost P to be installed; thus:

$$P(\overline{x}) = P_1(\overline{x}) = \sum_{(i,j)} 2\,P\,W\,l_{i,j} \tag{C.6}$$

MG-OXCs

If we include end-to-end waveband switching, the number of required ports at a switching node is equal to the double of the number of LPs traversing it, that can be end-to-end wavelengths, wavebands and multi-waveband channels; so:

$$P(\overline{x}) = P_2(\overline{x}) = \sum_{(i,j)} \sum_{(p,h,l)_{wp}} 2\,P\,x_{i,j}^{p,h,l} \tag{C.7}$$

MG-OXCh

Considering local wavebands and multi-wavebands in addition to end-to-end wavebands and multi-wavebands, the number of ports per fibers is obtained from the number of ports needed in the MG-OXCs case minus the number of ports saved thanks to local wavebands and multi-wavebands. The global port cost is calculated through the following notations and constraints:

$(i, j, k) \in N \times N \times N$, $(i, j) \in \epsilon_P$, $(j, k) \in \epsilon_P$, $k \neq i$, is a 1-hop arc;

$f_{i,j,k}^{p,h,l}$: set to 1 if the LP (p, h, l) passes over the 1-hop arc (i, j, k), to 0 otherwise;

$s_{i,j,k}^{h}$: number of 1-hop wavebands (h=1) or multi-wavebands (h=2) over the arc (i, j, k).

$$2\,f_{i,j,k}^{p,h,l} \;\leq\; x_{i,j}^{p,h,l} \;+\; x_{j,k}^{p,h,l}, \quad \forall (i,j,k), \forall (p,h,l)_{wp,\,h\in\{1,2\}} \tag{C.8}$$

$$R\,s_{i,j,k}^{h} \;\leq \sum_{(p,h,l)_{wp}} f_{i,j,k}^{p,h,l}, \quad \forall (i,j,k),\, \forall h \in \{1,2\} \subset H \tag{C.9}$$

$$f_{i,j,k}^{p,h,l},\; s_{i,j,k}^{h} \in \mathbf{N} \tag{C.10}$$

$$P(\overline{x}) = P_3(\overline{x}) = P_2(\overline{x}) - 2\,R\,P \sum_{(i,j,k)} \sum_{h\in\{1,2\}} s_{i,j,k}^{h} \tag{C.11}$$

MG-OXCc

We introduce the heterogeneous local multi-waveband, able to group end-to-end wavebands (not already grouped in multi-wavebands) with existing local wavebands. The global port cost is calculated as of the following:

$m_{i,j,k}$: number of heterogeneous local multi-wavebands over the 1-hop arc (i,j,k);

$g^h_{i,j,k}$: final number of local wavebands ($h=1$) and multi-wavebands ($h=2$) over (i,j,k).

$$R\, m_{i,j,k} \leq \sum_{(p,h,l)_{wp}}^{h=2} f^{p,h,l}_{i,j,k} - R\, s^2_{i,j,k} + s^1_{i,j,k}, \quad \forall (i,j,k) \tag{C.12}$$

$$g^1_{i,j,k} = s^1_{i,j,k}, \qquad g^2_{i,j,k} = s^2_{i,j,k} + m_{i,j,k}, \quad \forall (i,j,k) \tag{C.13}$$

$$g^h_{i,j,k},\ m_{i,j,k} \in \mathbf{N} \tag{C.14}$$

$$P(\overline{x}) = P_4(\overline{x}) = P_2(\overline{x}) - 2\, R\, P \sum_{(i,j,k)} \sum_{h \in \{1,2\}} g^h_{i,j,k} \tag{C.15}$$

In the OXC and MG-OXCs cases, the variable number is equal to $|N|^2(a|N|+1)$ and the constraint number equal to $|N|^2(1+\frac{3}{2}a)$, where a is the average number of LPs per node and $|N|$ the number of nodes, and supposing the a full meshed physical topology (worst case). The introduction of the subpath grouping increases the complexity by a factor N: in particular, the MG-OXCh case introduces other $|N|^3(2+Na)$ additional constraints and variables, and MG-OXCc needs adds further $3|N|^3$ constraints and variables.

C.5 Numerical Results

We consider the three case-study networks depicted in Fig.C.3: the EON, the NSFNET and a 6-nodes backbone extracted from the EON core that we will call EONc; the NSFNET has 14 nodes and 44 unidirectional arcs, the EON is a more interconnected backbone with 19 nodes and 78 arcs, and EONc has 6 nodes and 18 arcs. We analyze the results considering these three networks loaded by three different traffic volumes; the global traffic load is set equal to the global transport capacity of the considered backbones, if equipped with 1, 2 or 3 fibers per unidirectional arc. Then, the total traffic has been distributed among all the node couples in the network according the gravitational model introduced in Section C.2.

Cost values are expressed in unit of fiber cost F [4]: $P/F = 150$, $\beta/F = 0.1$. We set $\sigma = 0.9 < 1$. The optimization problem can be solved quickly for low traffic loads especially in the cases without subpath grouping; indeed, solutions for MG-OXCh and

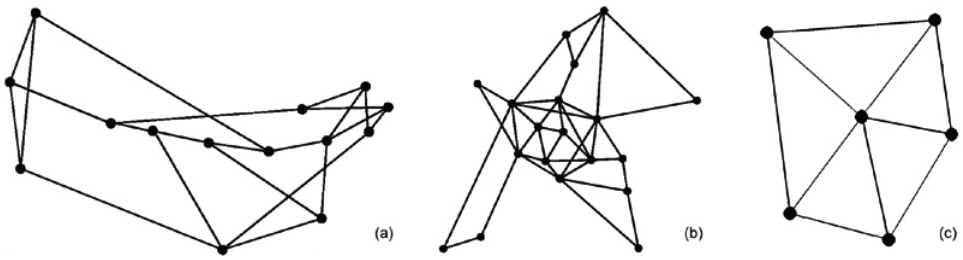


Figure C.3: Topology types: (a) NSFNET (b) EON (c) EONc

	NSF			EON			EONc		
Load case	1	2	3	1	2	3	1	2	3
OXC	10656	19257	28021	8945	14970	20970	1192	2282	3281
MG-OXCs	10039	17840	25741	8062	12985	17975	986	1782	2533
MG-OXCh	9781	17479	25350	NA	NA	NA	969	1772	2527
MG-OXCc	NA	NA	NA	NA	NA	NA	967	1767	2515

Table C.1: Number of allocated ports under different traffic volumes and cases

OXC	7904	14240	20992	11456	20256	28896	1984	3968	5760
MG-OXCs	3806	4762	5416	5370	6902	8234	544	636	772
MG-OXCh	3197	2796	4454	NA	NA	NA	492	558	728
MG-OXCc	NA	NA	NA	NA	NA	NA	476	551	724

Table C.2: Objectives (in thousands) under different traffic volumes and cases

MG-OXCc are more difficult to obtain. For example, the number of variables and constraints goes from 31196 and 25532 for the NSFNET-1/OXC case, to 139960 and 239324 for the NSFNET-3/MG-OXCc case (NSFNET-x stands for x-fibers per arc global traffic load).

We fixed a time limit of 36 hours to achieve the solution: when the solver does not reach the optimal solution within this limit, then, if at least a feasible integer solution reasonably close to the optimal solution is available, we report it, otherwise we report that the solution is not available (NA). This happens to NSFNET and EON in the most complex subpath wavebanding cases. We have, however, enough results over the EONc network to estimate the effectiveness of WBS in all the WBS scenarios.

Tables C.5 and C.2 contain respectively, the objective values of network cost, and the ports number for the three networks under the different traffic loads. The first table is useful to observe how the total network cost is affected by the application of wavebanding; the second table is a better indicator of how the different waveband grouping methods determine the ports to enable in the network.

As expected, WBS capabilities at the switching nodes allow for an increasing reduction of the network cost as well as of the ports number. In particular, the most significant gain is already achieved by means of end-to-end wavebanding (MG-OXCs case): grouping connections owing to the same source-destination pair results in a wide advantage especially for high network loads when, according to our traffic model, LP-2 and LP-3 are dominant with respect to LP-1.

WBS performance can be further improved grooming also connections that own to different end-to-end connections. Tables C.5 and C.2 show that the effect of subpath grouping can be still significant, especially for lower traffic loads: the results show that the MG-OXCh case induces a small decrease in network cost, yet a more significant gain on ports, especially for lower traffic loads. As a matter of fact, the subpath grouping tends to perform better when the network has more residual capacity and many connections with low-granularity are available and can be grouped: this two situations are more likely to emerge for low traffic loads. Moreover, dense topologies may benefit more from subpath grouping: in full-meshed networks we suppose that we can perform better subpath wavebanding than in ring topologies, for example. In [142] authors observed that the introduction of subpath grouping may induce longer lightpaths routes, because the joint enforcement of capacity constraints and minimization of WBS ports could lead to re-routing of connections over not-effective paths. In our work, this trade-off is solved according to the minimization of a different network cost.

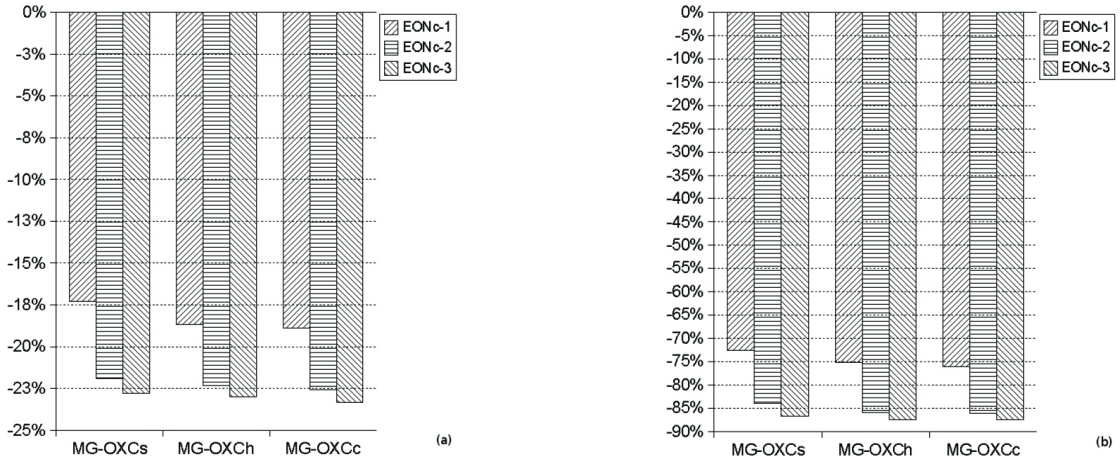


Figure C.4: (a) Objectives and (b) ports percentage reduction for EONc topology

Finally, in the MG-OXCs case we added the heterogeneous local multi-wavebanding functionality which grants a further, yet very small additional reduction in ports number and network cost; remarkable savings by heterogeneous local grouping may be achieved only in networks with few LPs, mainly of lower bit-rates classes. In MG-OXCs case, the design problem becomes very challenging and we could solve only for the EONc topology. Since network design and management becomes more difficult to tackle with, each network operator should decide according to its planning requirements if heterogeneous wavebanding is a viable choice to be implemented in its network.

Comparing the results on the three network topologies, we can observe that, in percentage, the global cost and port number reduction is smaller in NSFNET and EON than in EONc; this difference is related to the larger arcs length in NSFNET and EON: longer arcs imply larger fiber costs, while the port cost is not affected by the geographical dimension. As a consequence, wider networks tend to have higher fiber cost with respect to port cost in our model. The application of WBS reduces the cost amount due to ports, modifies the global propagation delay cost because lightpaths are deflected from the shortest path to increase WBS, while the fiber costs is not (at least directly) reduced by WBS.

In Figs. C.4 we draw the percentage savings of the various forms of WBS with respect to OXC case, for the different traffic loads. in the EONC case. In Fig. C.4a we report the global cost percentage decrease: from OXC to MG-OXCs, the objective reduces rapidly, till 22% with the EONc topology. This reduction increases for larger values of traffic load: the LP-1 number is stable in the three traffic scenarios and so the number of formed wavebands; on the contrary, the LP-2 number increases significantly as the number of formed multi-wavebands. In Fig. C.4b we report the percentage reduction of ports number. Considering the case with lower traffic, MG-OXCs reduces the ports number already over the 70% and this reduction keeps increasing till over 75% for MG-OXCh and MG-OXCc cases. For high traffic loads, the saving is even more consistent reaching the 88%.

In Fig. C.5 we illustrate how the global network cost is distributed among its components (fibers, ports, and propagation delays). Without WBS, the port cost represents about 26% of the global cost. The introduction of end-to-end wavebanding is able to reduce the port cost share to 4,57%: this share shows a very small additional decreases with subpath wavebanding. Note that the share of cost due propagation delay increases

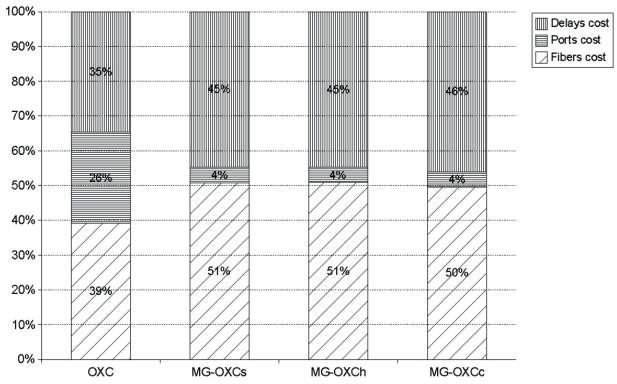


Figure C.5: Cost distribution for EONc-3 in different grouping cases

passing from MG-OXCs to MG-OXCh: this confirms that subpath grouping may imply longer paths for the grouped LPs.

C.6 Summary

We analyzed and quantified, for the first time, the effect of joint end-to end and subpath wavebanding in optical networks, carrying on a cost-effective design of wide-area case-study networks. We proposed and applied ILP formulations for a case-by-case analysis and showed the benefits of WBS technology, when switching operates at wavelength, waveband and fiber level. The best results on both network cost and ports number are achieved by means of subpath wavebanding; in particular we showed that: the application of end-to-end wavebanding by itself already leads to very good results, not far from the results achievable by subpath grouping, especially for high load; secondarily subpath grouping with heterogeneous grouping do not introduce relevant savings with respect to the case without heterogeneous grouping.

Globally we showed that WBS backbone can be less expensive up to 22% and can save more than 50% of ports in comparison with classical OXC-based networks. Further work is needed to develop heuristics for the design of WBS networks with subpath grouping, in order to apply this techniques to dense networks.

POLITECNICO DI MILANO
Dipartimento di Elettronica e Informazione
Piazza Leonardo da Vinci, 32, 20133 - Milano