Design Optimization of the Petaweb Architecture

Anne Reinert, Brunilde Sansò, *Member, IEEE*, and Stefano Secci, *Student Member, IEEE*

***Abstract*—This paper explores the design modeling issues of the Petaweb, an optical network architecture that provides fully meshed connectivity between electronic edge nodes. The Petaweb is simple to manage, simplifies key networking functions such as routing and addressing and can offer a total capacity of several Petabits per second. From the topology standpoint, it is an unusual structure as the backbone nodes are totally disconnected whereas the edge nodes are all attainable in one-hop. The network design problem leads to a very hard combinatorial problem. We propose a model and a heuristic approach that is based on repeated matchings. Computational results concerning the modeling issues will be presented and thoroughly discussed.**

***Index Terms*—Capacitated location problem, composite-star network, dimensioning, matching, Petaweb, topological design.**

I. Introduction

The Petaweb is a new network structure that offers a total capacity of several petabits per second (10^{15} b/s) that was proposed for a next generation Internet [1], [2] [3]. 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} b/s). The structure provides fully meshed connectivity with direct optical paths between electronic edge nodes. It is composed of several optical cross-connectors (OXCs), 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 also be seen as a superposition of star structures as shown in Fig. 1. 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 fewer communication layers and simpler protocols than what we are used in the current Internet, thus greatly increasing network efficiency and communication speed. The proposed architecture also provides a high level of network reliability.

The large improvement in network efficiency of the Petaweb architecture comes at the expense of a significant increase in fiber costs as all of the edge nodes have to be connected to all of the core nodes. Another possible drawback of the proposed structure is that its topology is such that the upgrade of the network has to be carefully crafted. In [3], the Petaweb architecture was formally compared with an optical multihop 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 single-hop connectivity.

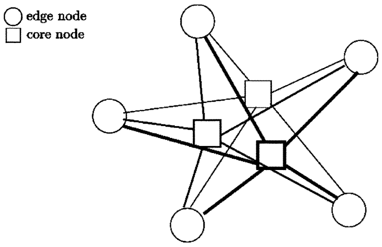


Fig. 1. Petaweb architecture: a composite-star structure.

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} b/s) [4], [5] .

From the telecommunication 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 of how to optimize separately or jointly those two different levels. In [6], a thorough review of all of the types of design problems and algorithmic resolutions can be found. The Petaweb, on the other hand, presents a different structure: all of the edge nodes are connected through a backbone switch and yet the backbone switches are disconnected among themselves.

In mathematical terms, the Petaweb design remains a location problem since we must decide where to place the core nodes. It presents similarities with the Capacitated Facility Location Problem [7] and, in particular, with the Single Source Facility Location Problem (SSFLP) [8], [9]. Nevertheless, the capacity and physical constraints that are present in the design make it a problem much more difficult to solve.

The objective of this paper is to formally define the Petaweb Design Problem and propose a mathematical formulation and an efficient resolution approach.

This paper is divided as follows. In Section II, we present the mathematical formulation for the Petaweb Design Problem, discuss its modeling details, and evaluate its computational complexity. In Section III, a heuristic approach to allow us to solve

Manuscript received August 08, 2005; revised August 29, 2006 and June 08, 2007; approved by IEEE/ACM Transactions on Networking Editor F. Neri. First published June 24, 2008; current version published February 19, 2009.

A. Reinert is with Capgemini, 31036 Toulouse Cedex 1, France (e-mail: anne.reinert@polymtl.ca).

B. Sansò is with the Department of Electrical Engineering, École Polytechnique de Montréal, Montreal, QC, Canada (e-mail: brunilde.sanso@polymtl.ca).

S. Secci is with the Department of Computer Science and Networks, TELECOM ParisTech, 75014 Paris, France, and with the Department of Electronics and Information, Politecnico di Milano, Italy (e-mail: secci@enst.fr).

Digital Object Identifier 10.1109/TNET.2008.925092

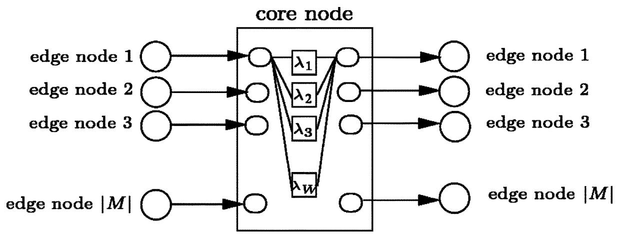


Fig. 2. Connection between the edge nodes and a core node.

large instances of the problem will be presented. The results of the approach will be later compared in Section IV with the solutions obtained with a general MILP solver with two different sets of traffic matrices. In that section, a sensitivity and a scalability study of the heuristic are also carried out. Concluding remarks and suggestions for further work are presented in Section V. Details on the heuristic implementation are provided in Appendix A.

II. Petaweb Design Problem

A. Switching System

The Petaweb is based on WDM technology. The fiber is composed of a fixed number of channels, with each channel corresponding to one wavelength. When the fiber enters a core node, it is demultiplexed in its channels, and each channel is connected to its associated switching plane. As depicted in Fig. 2, 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. Note that, to ease the figure interpretation, only the channels from and to edge node 1 are pictured. 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, 32, 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.

It is worth mentioning that, given the regularity of the core node architecture (same number of wavelengths per fiber and the same number of fibers per link), no wavelength conversion is required, and no wavelength continuity constraint needs to be applied [3]. However, if in a dimensioned network some links appear to be underused, the network planner may decide to reduce the number of fibers per link (if a core node has many switching planes) arbitrarily, while keeping the nonblocking system and requiring wavelength converters at some space switches. Furthermore, whenever it would appear from the given traffic patterns that by grooming traffic of different requests on the same wavelength important capacity savings can be obtained, the switching plane structure can be adapted to support time division multiplexing (TDM). In such a case, edge nodes would split the traffic over different time-slots, and the switching plane should be capable of aligning and multiplexing time slots: a proper switching plane architecture has been recently evaluated in [10].

B. General Description and Notation

The Petaweb Design Problem (PDP) consists of 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 useful notation.

M edge node set;

N set of potential core node locations;

T set of edge node pairs, with the origins different from the destinations, that is, $T \subset M \times M$;

V set of core node types;

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

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

C_{channel} channel capacity (in Gb/s);

W number of wavelengths per fiber;

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

K_r total capacity of a core node of type r, $r \in V$, (in Gb/s), $K_r = s_r \times W \cdot |M| \cdot C_{\text{channel}}$;

f_r cost of one core node of type r, $r \in V$;

P cost of one port in a core node;

γ scale factor for the cost of the ports;

F reference fiber cost per length unit;

$\phi(W)$ discrete function that scales F as a function of the number of wavelengths;

β cost representing the propagation delay per length and traffic unit;

Q_p traffic of an origin/destination pair p, $p \in T$, (in Gb/s);

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

d_{ip} 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 eth core node of type r located at i is opened and 0 otherwise;

$x_{\mathrm{ire,p}}$ binary variable equal to 1 if traffic Q_p is switched by the eth core node of type r located at site i and 0 otherwise.

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.

C. 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.

1) Cost of the Core Node: The cost of the core nodes 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 given core 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 given by $2|M|Ws_r$; 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 then given by $2|M|Ws_r\gamma^{(s_r-1)}P$. Factor γ is lower than 1 so that the cost per port decreases with the type of core node. For instance, if $\gamma = 0.95$ the cost of the ports of type 1 ($s_r = 1$) will be $2|M|WP$. On the other hand, the cost of the ports of type 2 ($s_r = 2$) will be $0.95 \times (2|M|WP)$ which implies an economy of 5%.

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

3) Propagation Delay Cost: The propagation delay cost term aims at choosing the edge core type and location so that the *pondered* propagation delay is minimized. The pondered term was used 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 traveled 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: $\sum_{i\in N}\sum_{r\in V}\sum_{e\in E}\sum_{p\in T}\beta d_{ip}Q_p x_{ire,p}$.

Thus, the objective function of the problem is

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

D. Constraints

1) Unicity of the Core Node Connection:

$$\sum_{i\in N}\sum_{r\in V}\sum_{e\in E} x_{ire,p} = 1 \quad \forall\, p \in T. \quad (2)$$

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

2) Linking Constraints:

$$x_{\mathrm{ire,p}} \le y_{\mathrm{ire}} \quad \forall\, i \in N,\ \forall\, r \in V,\ \forall\, e \in E,\ \forall\, p \in T. \quad (3)$$

This specifies that the traffic can be routed through the eth core node of type r located at site i only if this core node is active.

3) Core Node Capacity Constraints:

$$\sum_{p\in T} Q_p x_{\mathrm{ire,p}} \le K_r y_{\mathrm{ire}} \quad \forall\, i \in N,\ \forall\, r \in V,\ \forall\, e \in E. \quad (4)$$

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

4) Edge Node Capacity Constraints:

$$C_{\mathrm{channel}} \times W \times \sum_{i\in N}\sum_{r\in V}\sum_{e\in E} s_r y_{\mathrm{ire}} \le C_j \quad \forall\, j \in M. \quad (5)$$

This 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

$$\left\lfloor \frac{1\ \text{Tb/s}}{160\ \text{Gb/s}}) \right\rfloor = 5\ \text{fibers}$$

(with each fiber having 16 wavelengths of 10Gb/s) per direction. This can correspond, for instance, to one core node of type 1 and one of type 3, or five of type 1, etc.

5) Link Capacity Constraints:

$$\begin{aligned}\sum_{p\in T\ origin\ j} Q_p x_{\mathrm{ire,p}} &\le C_{\mathrm{channel}} \times W \times s_r y_{\mathrm{ire}}, \\ &\forall\, j \in M, \quad \forall\, i \in N, \\ &\forall\, r \in V, \quad \forall\, e \in E \quad (6)\end{aligned}$$

$$\begin{aligned}\sum_{p\in T\ destination\ k} Q_p x_{\mathrm{ire,p}} &\le C_{\mathrm{channel}} \times W \times s_r y_{ire}, \\ &\forall\, k \in M, \quad \forall\, i \in N, \\ &\forall\, r \in V, \quad \forall\, e \in E. \quad (7)\end{aligned}$$

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

6) Binary Constraints:

$$\begin{aligned} y_{\mathrm{ire}} &\in \{0,1\} \quad \forall\, i \in N,\ \forall\, r \in V,\ \forall\, e \in E, \\ x_{\mathrm{ire,p}} &\in \{0,1\} \quad \forall\, i \in N,\ \forall\, r \in V,\ \forall\, e \in E,\ \forall\, p \in T. \quad (8)\end{aligned}$$

E. The Mathematical Model

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

$$\min (1) \quad \text{subject to} : (2),\ (5),\ (6),\ (7),\ \text{and}\ (8).$$

Note that constraints (6) and (7) imply (4) and (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 $|T| + |M| + 2|N||V||E||M|$ ($\sim |T| + |N||M|$). Supposing that $N \equiv M$, and being $|T| \approx |M|^2$, the complexity of the PDP depends on a number of variables $\sim |M|^3$ and on a number of contraints $\sim |M|^2$.

As previously stated in the Introduction, the PDP has some similarities with the SSFLP that is known to be NP-hard. In the SSFLP, we have a set of customers that must be served by a single facility and there is a cost associated with 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 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.

III. Resolution Approach

Here, we present a heuristic method based on a repeated matching heuristic to be able to solve large instances of the problem. We first provide some key definitions used for problem reformulation before introducing the heuristic and discussing complexity issues.

A. Reformulation of the PDW

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), $i \in N, r \in V, e \in E$. i indicates the site of the core node, r is the type of the core node, and e is the identifier of the core node of type r at site i with which we are dealing. 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 $((i,r,e),D_k)$.

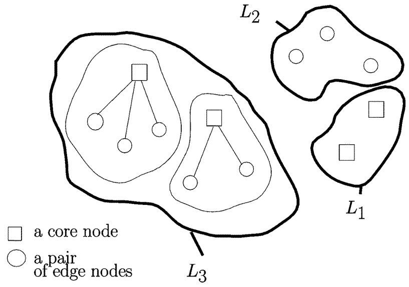


Fig. 3. Sets L_1, L_2, and L_3 associated with a packing Π.

A kit $((i,r,e),D_k)$ is said to be feasible if 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. Let us define a *packing* as 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: $((i_1,r_1,e_1),D_1),((i_2,r_2,e_2),D_2) \in \Pi, \Leftrightarrow (i_1,r_1,e_1) \neq (i_2,r_2,e_2)$, and $D_1 \cap D_2 = \emptyset$.

Given a packing Π, let us define L_1, L_2, and L_3. 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 $J_k = \bigcup_{(i,r,e,D_k) \in \Pi} p \in D_k$. Finally, 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. For example, in Fig. 3, $n_1 = 2$, $n_2 = 3$, and $n_3 = 2 \ldots$ Fig. 3 shows a packing Π whose cost can be determined as the sum of all of the terms of objective function (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 cost of the packing is reduced at each iteration, details will be given in Section III-B and in Appendix A.

B. Matching Problem

The classical matching problem can be described as follows. Let A be a set of q elements $h_1, h_2, \ldots, h_q$. A matching over A

is so that 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 and zero otherwise.

The matching problem consists in finding the matching over A that minimizes the total cost of matched pairs

$$\min \quad \sum_{i=1}^{q}\sum_{j=1}^{q} c_{ij} z_{ij} \tag{9}$$

$$\text{s.t.} \quad \sum_{j=1}^{q} z_{ij} = 1, \quad i = 1,\dots,q \tag{10}$$

$$\sum_{i=1}^{q} z_{ij} = 1, \quad j = 1,\dots,q \tag{11}$$

$$z_{ij} = z_{ji}, \quad i,j = 1,\dots,q \tag{12}$$

$$z_{ij} \in \{0,1\}, \quad i,j = 1,\dots,q. \tag{13}$$

Equations (10) and (11) ensure that each element is exactly matched with another one. Equation (12) ensures that, if h_i is matched with h_j, then h_j is matched with h_i. Equation (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, the elements of L_2 and the elements of 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 of the sets L_1, L_2 and L_3. 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 of L_1, L_2, or L_3.

The costs c_{ij} are stored in a matrix C. The dimension of cost matrix C is $(n_1 + n_2 + n_3) \times (n_1 + n_2 + n_3)$. Note that this dimension changes at each iteration.

C is a symmetric matrix composed of nine submatrices. 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 the elements of L_j as

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

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 Appendix A.

Once the cost matrix is calculated, the matching problem (9)–(13) is solved heuristically. The resolution is not easy because of the symmetry constraint (12). We have implemented the algorithm of Forbes [11] that is based on the method of Engquist [12]. The starting point for Forbes' algorithm is the solution vector of the matching problem without the symmetry constraint (12). Such a starting solution is obtained with the algorithm of Jonker and Volgenant [13] that was chosen for its speed performance. The output of the Forbes' algorithm is a symmetric solution vector that indicates the matchings to be performed between the heuristic elements.

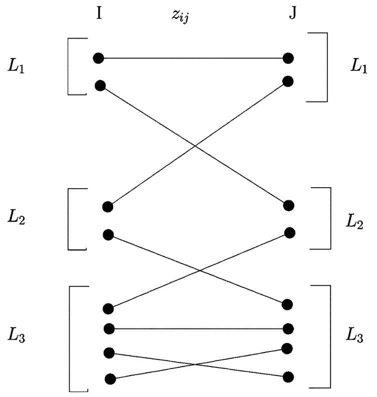


Fig. 4. Solution for the matching problem.

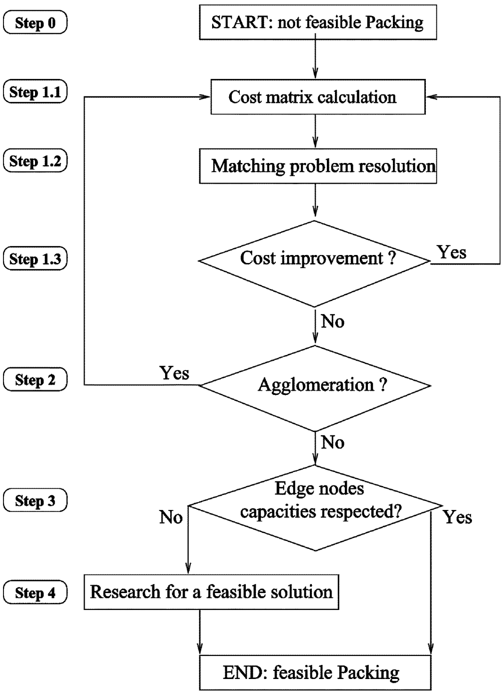


Fig. 5. Chart of the repeated matching heuristic for the Petaweb design. This figure was inspired by the work of Rönnqvist [8].

Fig. 4 illustrates a possible solution of the matching problem. The solution of the matching problem 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.

C. Repeated Matching Heuristic for the Petaweb Design

A global chart of the heuristic is given in Fig. 5.

- **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**: At each iteration, the cost matrix C is calculated for every block (Appendix A).
- **Step 1.2**: Then, the problem of finding the least costly matchings between the elements 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.
- **Step 1.3**: When the cost of the packing cannot be reduced any more, i.e., when the matching results do not produce cost improvement for the current packing, then proceed to **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 constraint 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.

D. Complexity

The complexity of the whole heuristic depends on its different subalgorithms and phases. The calculation of the cost matrix is straightforward except for two blocks of the matrix (see blocks 5 and 6 in the Appendix) 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 has a $O(n^2)$ complexity, where $n = n_1 + n_2 + n_3$. The Forbes' algorithm looks for a symmetric matching vector starting from the Volgenant's asymmetric solution vector; the algorithm creates a branch-and-bound tree whose dimensions increase during the research of a symmetric solution. However, in order to avoid excessive searches, we controlled the dimensions of the tree: when the search goes above a higher fixed bound without finding a solution, a nonoptimal solution with a forced symmetry is given back. Thus, the complexity of the matching resolution phase is kept under control by introducing suboptimal solutions. Not bad, since we deal with a heuristic that solves a succession of matching problems. The higher bound for the search tree was fixed to 1000 tree children.

IV. Computational Results

The proposed heuristic was tested using two networks, composed respectively 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, which is calculated using a gravity model based on urban populations and distances between cities. The urban populations were found in [14]. 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 and 10.692 Tb/s. The values for matrix B were 2.167 and 10.050 Tb/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 [15] and a formula to assess the distance between two points on a sphere [16] was used. The calculated distances were later compared and validated with a few air distances estimated at the University of Minnesota [17].

The following default values were used: $W = 16$; $C_{\text{channel}} =$ 10 Gbit/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); $C_j = 1000$ Gbit/s for 10-node networks, $C_j = 2000$ Gbit/s for the 34-node network with traffic matrix A, and $C_j = 2800$ Gbit/s for the 34-node network with traffic matrix B. $\phi(W) = W$ is a discrete function used to scale the reference fiber cost F. F is assumed to be the cost of a single-wavelength fiber. When F is multiplied by $\phi(W) = W$, the resulting fiber cost is considered to be proportional to the number of wavelengths.

A. Results With Default Parameters

The first set of tests was run to solve the problem using the default parameters and using two resolution approaches: CPLEX and the proposed heuristic. The results are presented in Table I for the 10-node network and in Table II for the 34-node network.

TABLE I
RESULTS OBTAINED FOR THE 10-NODE NETWORKS

Network	Traffic A heuristic	Traffic A CPLEX	Traffic B heuristic	Traffic B CPLEX
Objective (/F)	2289564	2280980	2153868	2152920
Core node cost	11.4%	11.2%	11.9%	12.1%
Fiber cost	77.1%	77.8%	83.3%	83.8%
Delay cost	11.5%	11%	4.8%	4.1%
Time (s)	6	23650	11	232
Optimum gap	0.38%	-	0.04%	-

TABLE II
RESULTS OBTAINED FOR THE 34-NODE NETWORKS

Network	Traffic A heuristic	Traffic A CPLEX	Traffic B heuristic	Traffic B CPLEX
Objective (/F)	31940857	31837547	44757016	42406000
Core node cost	5.5%	5.3%	5.4%	5.3%
Fiber cost	82%	81.7%	82.2%	81.6%
Delay cost	12.6%	13%	12.5%	13.1%
Time (s)	217	579998	322	1614383
Optimum gap	0.32%	-	5.5%	-

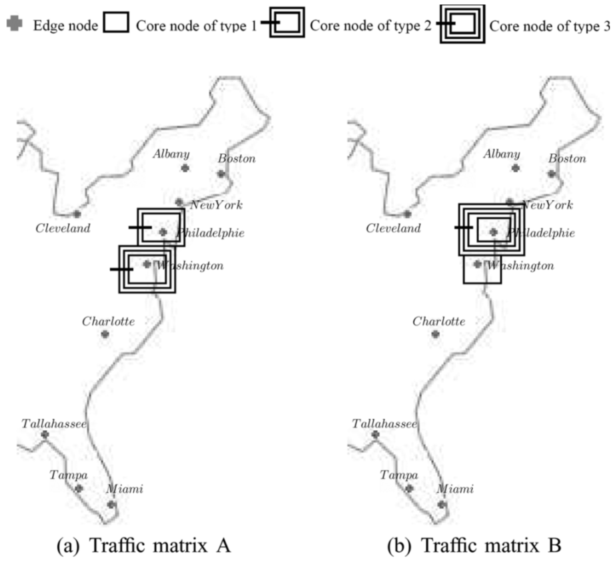
The gap in the last line is the discrepancy in percentage between the total network cost found by the heuristic and the total network cost found by CPLEX for the mathematical model. The costs have all been normalized to F. The actual solutions obtained for all instances treated are presented in Figs. 6–9.

In terms of computational complexity, we can see that these are extremely hard problems. In fact, in some of the instances, it took CPLEX up to 18 days to reach the best solution, and that was for a network of 34 nodes. 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.

Regarding 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 of the costs considered, for both the 10-node network and the 34-node network. There is, however, a slight difference between the cases with the A and B matrices' runs 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 the difference is, in the case of the matrix A, being absorbed by the delay cost. So, for this small network, the sparseness or fullness 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.

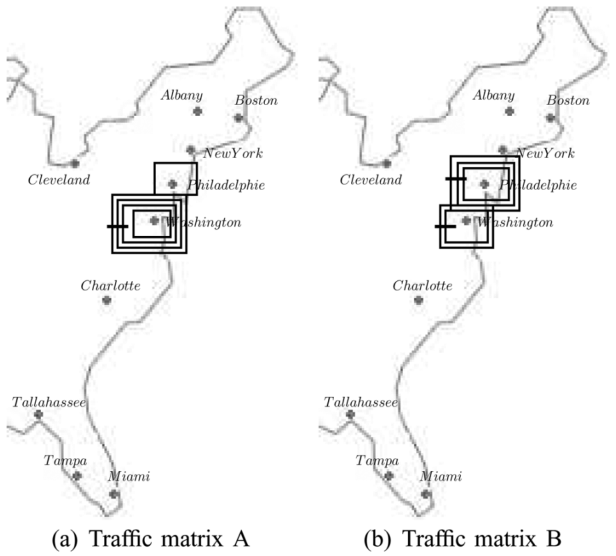
B. Sensitivity Studies

1) 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. 10 that the influence of the terms in the solution is quite different. When the delay costs are omitted, all of the switches are set at the center of mass of the map. On the other hand, when the fiber cost is set to zero but we keep a term to account for the delay, all of the switches are spread, with larger switches on the east part of the country where the higher origin–destination demand is concentrated.



(a) Traffic matrix A (b) Traffic matrix B

Fig. 6. 10-node networks with default parameters (CPLEX).

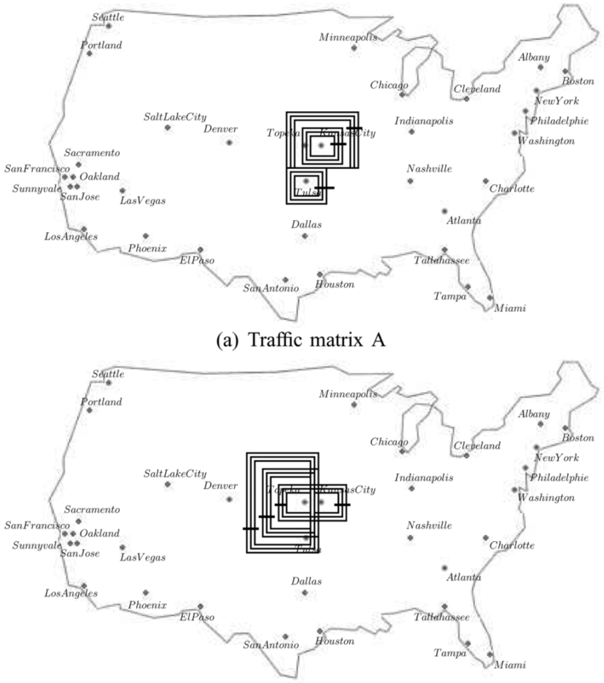


(a) Traffic matrix A (b) Traffic matrix B

Fig. 7. 10-node networks with default parameters (heuristic).

2) Propagation Delay Variation: Given the important influence of the delay term in the objective function, some sensitivity tests were made for the 34-node network with respect to the propagation delay cost. The weight parameter β for the propagation delay cost was progressively increased. The results for the traffic matrix A are presented in Table III and in Fig. 11.

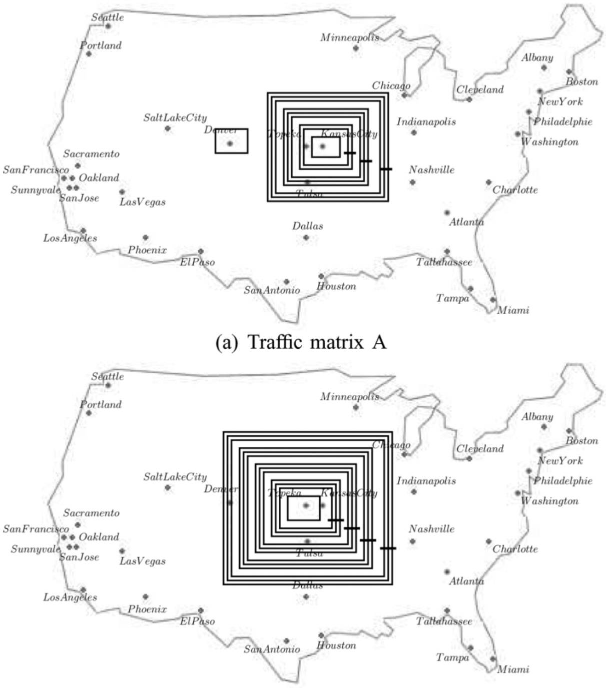
The importance of the term can be assessed from the results. Clearly, when the coefficient β increases, the active core nodes are increasingly more spread in the country. Thus, the added delay costs can be seen as a "natural" survivability term that prevents the location of all the resources in the same place. When we study Table III we can see that, as expected, the total cost



(a) Traffic matrix A

(b) Traffic matrix B

Fig. 8. 34-node networks with default parameters (CPLEX).

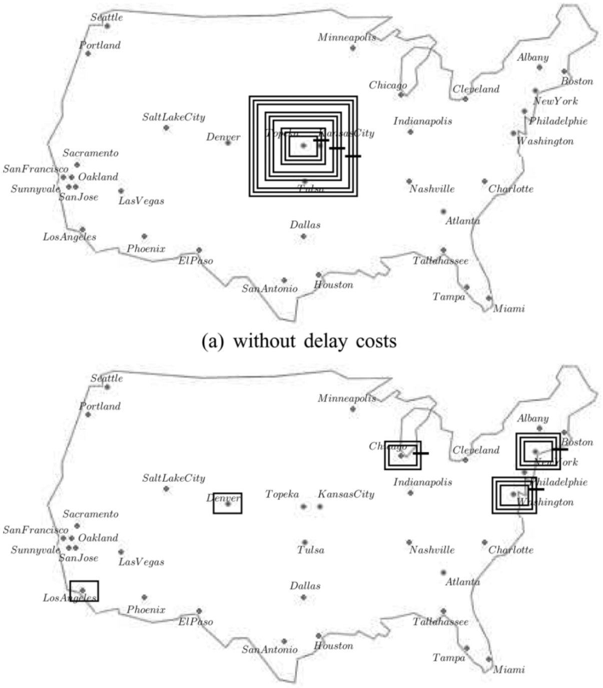


(a) Traffic matrix A

(b) Traffic matrix B

Fig. 9. 34-node networks with default parameters (heuristic).

of the network increases when β increases. Also expected is the proportion of the delay cost in the total cost. We cannotice 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.



(a) without delay costs

(b) without fiber costs

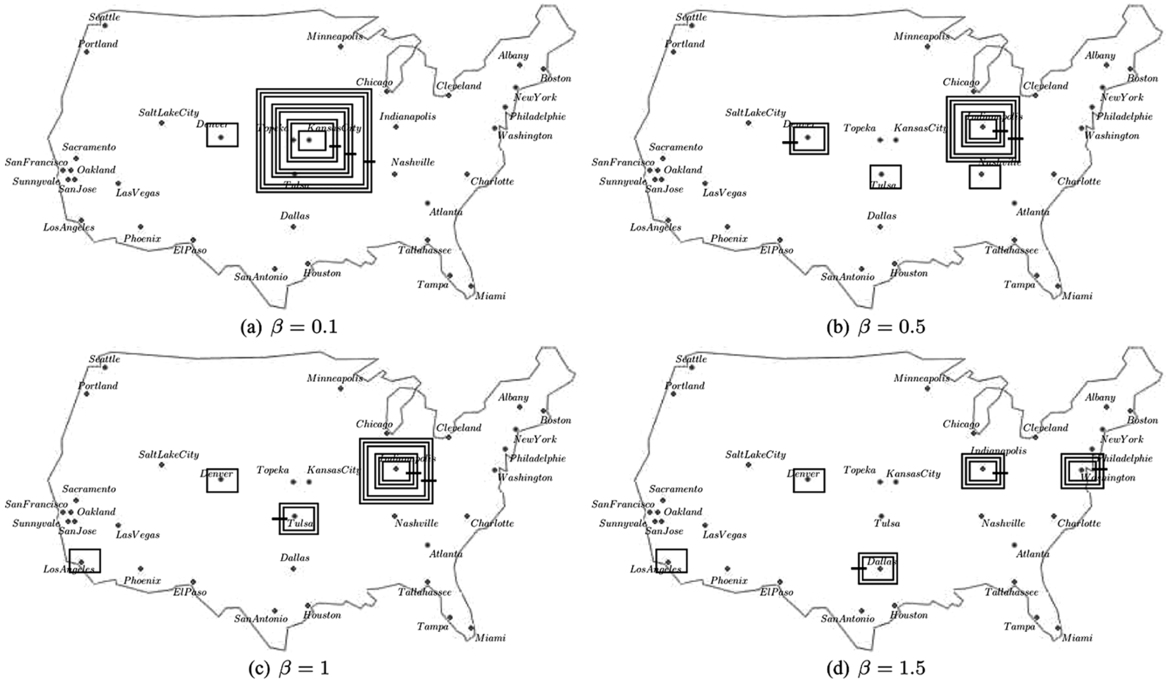
Fig. 10. 34-node network with default parameters, matrix A (heuristic).

TABLE III
INFLUENCE OF THE PROPAGATION DELAY COST FOR 34A (HEURISTIC)

β value	0.1	0.5	1	1.5
Objective (/F)	31940857	46864904	61244206	74650622
Core node cost	5.5%	3.7%	2.9%	2.4%
Fiber cost	82%	60.3%	46.6%	41.7%
Delay cost	12.5%	36.0%	50.5%	55.9%

In Table IV, we report, for the 10- and 34-node cases, another type of test to assess how the variation of β influences the average length of a connection, and thus the propagation delay. The connection length is the length (in km) of a lightpath being established between an origin and a destination edge node. 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 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 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. 12, 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



(a) $\beta = 0.1$ (b) $\beta = 0.5$

(c) $\beta = 1$ (d) $\beta = 1.5$

Fig. 11. 34-node network, traffic matrix A, for several propagation delay weights (heuristic).

TABLE IV
AVERAGE LENGTH OF AN ORIGIN–DESTINATION CONNECTION [km] AS A FUNCTION OF β. THE PEDIX IS THE STANDARD DEVIATION

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 V
WEIGHTED AVERAGE LENGTH OF AN ORIGIN–DESTINATION CONNECTION [km] AS A 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	$4875_{\sigma=2521}$	$3849_{\sigma=1923}$	$2804_{\sigma=1854}$	$1038_{\sigma=2495}$

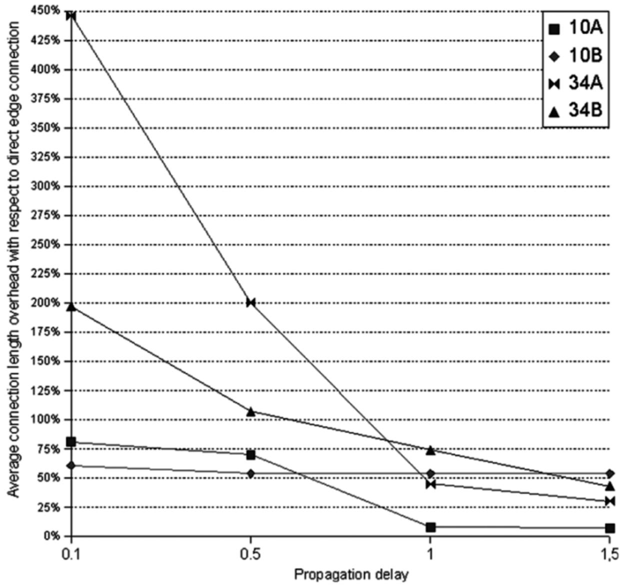


Fig. 12. Connection length average overhead as function of β with respect to a fully meshed network.

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 of the cases under study, 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 V. It can be seen that the weighted overhead decreases for all of 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 more impact.

The results of these tests make us conclude that the danger of a bigger propagation delay supposed in [1] can be controlled during the planning of the Petaweb structure.

3) Variation of Core Node Costs: The fixed unitary node cost f_r and the unitary cost per port P was first varied in the range $[-60\%, +60\%]$ of their default values. The tests produced no significant results: the route assignment per 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 increased by 100%, 200%, and 300% the cost of the cores to see if such a

TABLE VI
EFFECTS OF THE VARIATION OF CORE NODE COST. CN = CORE NODE

	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%	$2.6{\cdot}10^6$	21%	68%	9%	$2.5{\cdot}10^6$	22%	74%	4%	$3.9{\cdot}10^7$	11%	80%	9%	$5.5{\cdot}10^7$	11%	80%	9%
200%	$2.9{\cdot}10^6$	29%	62%	9%	$2.7{\cdot}10^6$	30%	67%	3%	$3.9{\cdot}10^7$	16%	75%	9%	$5.1{\cdot}10^7$	15%	75%	10%
300%	$3.1{\cdot}10^6$	35%	57%	8%	$3.0{\cdot}10^6$	36%	61%	3%	$4.1{\cdot}10^7$	20%	71%	9%	$5.3{\cdot}10^7$	19%	71%	10%

TABLE VII
RESULTS OBTAINED FOR DIFFERENT FIBER COSTS (HEURISTIC). $W = 16$. RESULTS FOR THE DEFAULT $\phi(w)$ ARE IN BOLD

10-node network, matrix A				
	$\phi(W) = W$	$\phi(W) = W\gamma^{\sqrt{W}}$	$\phi(W) = \sqrt{W}ln(W)$	$\phi(W) = \sqrt{W}$
Total cost	**2289564**	1983146 (-13.4%)	1774067 (-22.6%)	968542 (-67.7%)
Core node cost	**261011 (11.4%)**	278540 (14.05%)	278540 (15.70%)	278540 (28.76%)
Fiber cost	**1765254 (77.1%)**	1453612 (73.30%)	1257717 (70.89%)	451205 (46.59%)
Delay cost	**263299 (11.5%)**	250994 (12.65%)	237810 (13.51%)	238797 (24.66%)
10-node network, matrix B				
Total cost	**2153868**	1863349 (-14%)	1640993 (-23.8%)	828107 (-61.5%)
Core node cost	**256310 (11.4%)**	273750 (14.69%)	273750 (16.7%)	273750 (33.1%)
Fiber cost	**17694172 (83.65%)**	1493923 (80.17%)	1271497 (77.5%)	458610 (55.36%)
Delay cost	**103385 (4.61%)**	95675 (5.14%)	95746 (5.8%)	95746 (11.54%)
34-node network, matrix A				
Total cost	**31940857**	31289432 (-2%)	27509399 (-13.8%)	13378622 (-41.8%)
Core node cost	**1756747 (5.5%)**	2203530 (7.1%)	2203530 (8.01%)	2203530 (16.5%)
Fiber cost	**26191502 (82%)**	25518282 (81.5%)	21659389 (78.3%)	8073687 (60.3%)
Delay cost	**4024547 (12.5%)**	3567619 (11.4%)	3646479 (13.26%)	3101404 (23.2%)
34-node network, matrix B				
Total cost	**44757016**	43480527 (-2.8%)	3743923 (-8.3%)	17592830(-60%)
Core node cost	**2416878 (5.4%)**	2954390 (6.8%)	2893924 (7.7%)	2807480 (15.9%)
Fiber cost	**35790267 (82.1%)**	35572707 (81.8%)	29554327 (78.9%)	10709106 (60.9%)
Delay cost	**5594627 (12.5%)**	4953429 (11.4%)	4989672 (13.4%)	4076243 (23.2%)

major increase would lead to significant variations. The results can be seen in Table VI 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, important increment leads toward a design with higher propagation delays.

4) Sensitivity to Fiber Costs: Concerning fiber costs, the default data were obtained considering $\phi(W) = W$, that is, assuming that the global fiber cost is proportional to the number of wavelengths. In this part of the sensitivity analysis we wanted to assess the influence of this term in the final solution. For this, we varied $\phi(W)$. We considered 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 $W = 40$ is bigger than the incremental cost from $W = 80$ to $W = 100$, for example. The results are displayed in Table VII. 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 nondefault cases yield a fiber cost reduction of 18%, 29%, and 75%, respectively.

Also note 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 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 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.

C. Scalability of the Heuristic Approach

The heuristic has given very good results for 10- and 34-edge-node networks. Here, we increase the size of the networks to be treated to test the scalability of the heuristic method. The reader should be aware, however, that the very good optimality gaps that were obtained for 10- and 34-node networks may or may not be kept for larger networks, as CPLEX could not find a solution for larger instances.

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

TABLE VIII
RESULTS FOR SCALABLE NETWORKS WITH $\beta = 1$ (HEURISTIC)

Network:	40B	50B	60B	70B	80B
Total cost	$6.7 \cdot 10^7$	$6.6 \cdot 10^7$	$7.1 \cdot 10^7$	$7.2 \cdot 10^7$	$8.3 \cdot 10^7$
CN cost	3.9%	4.6%	4.6%	5.0%	5.5%
Fiber cost	60.0%	59.3%	62.8%	64.7%	70.5%
Delay cost	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	$8.9 \cdot 10^7$	$9 \cdot 10^7$	$1 \cdot 10^8$	$1 \cdot 10^8$	$1 \cdot 10^8$
CN cost	5.3%	5.2%	5.4%	5.4%	5.8%
Fiber cost	65.4%	66.4%	67.4%	69.3%	70.6%
Delay cost	29.3%	28.4%	27.2%	25.3%	23.6%
Iterations	28	27	27	30	38
Time (s)	90369	71619	555416	155500	505115

of one type that could be opened at one site was 4 and the maximum edge node capacity was $C_j = 3000$.

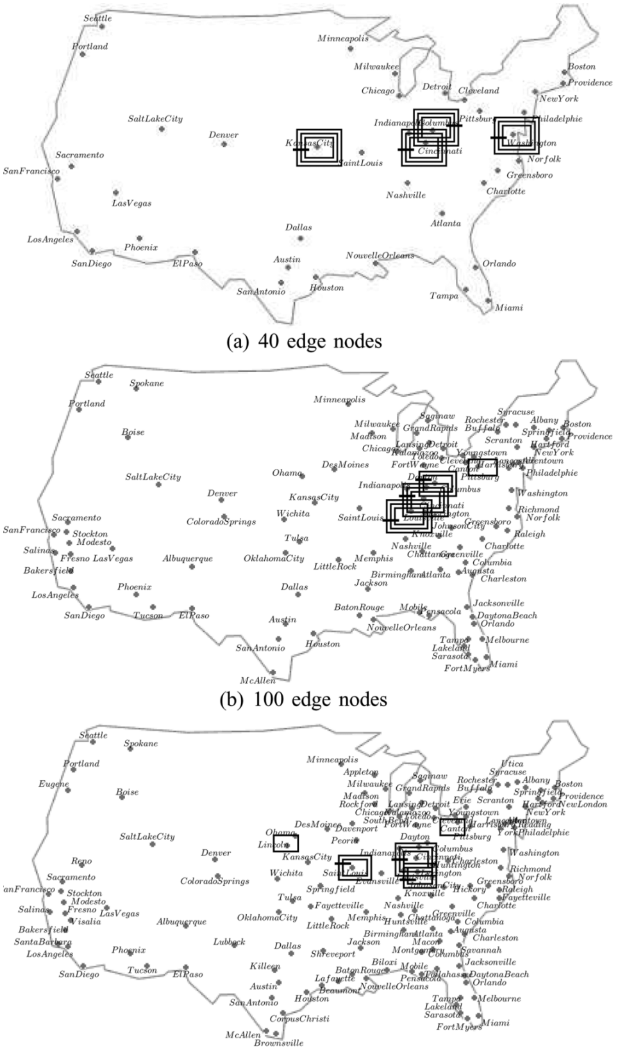
The results given by the heuristic for 40 to 136 edge nodes are given in Table VIII. As expected, the total network cost increases when the network size is growing. The proportions of the different costs in the total cost are kept almost constant. The fiber cost predominates with a percentage of 60% to 70% of the total cost. The delay cost comes next with a percentage of 25% to 35% of the total cost. The core node cost is the lowest with a percentage of 4% to 5.5% of the total cost. Three cases are illustrated in Fig. 13.

As expected, the resolution time increases with the network size. Nevertheless, it appears that some cases are quite more difficult to solve. For example, the 90-edge-node network calculation needs more than one week. The difficulty is located in the matching problem resolution. Other tests were triggered to better characterize the solution time. Fig. 14 illustrates the solution time diagram for 10- to 130-edge-node networks with all default parameters.

V. CONCLUSION AND FURTHER WORK

The Petaweb is a unique architecture that can yield important benefits to large-scale highly capacitated networks. From the networking standpoint it presents a completely different topology from the traditional access and backbone design concepts. In this paper, we have reviewed the architecture and formally defined the Petaweb design problem as a hard combinatorial problem that presents some similarities with a facility location problem. A mathematical formulation for general purpose MILP solver and a specialized and efficient heuristic method have been proposed.

In the design, we included equipment costs such as core and fiber costs and delay-related costs to allow for greater flexibility in the planning process. We used two different set of traffic matrices and two different network sizes to carry out the tests. We found that, with the default parameters, 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%.



(a) 40 edge nodes

(b) 100 edge nodes

(c) 136 edge nodes

Fig. 13. Scalable networks with $\beta = 1$ (heuristic). B traffic matrix.

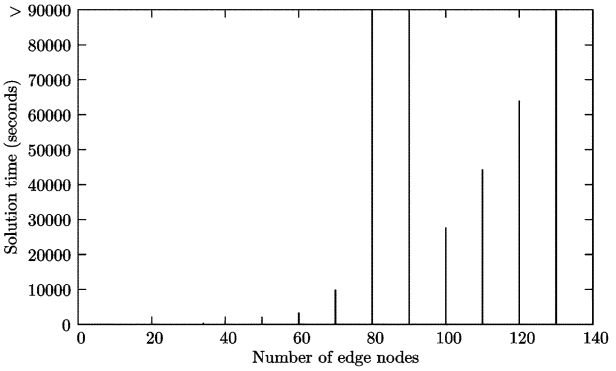


Fig. 14. Solution time for scalable networks (heuristic). B traffic matrix.

From the modeling standpoint we added a term to account for propagation delay when doing the Petaweb design. This term produced a topology where the connections were more direct, thus improving network efficiency, while acting as natural “reliability” enhancer by avoiding too much concentration of core nodes in the same sites.

The heuristic designed 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 cases are more difficult to solve than others. The resolution difficulty lays on the resolution of the matching problem, which can provide an avenue for future research to improve the resolution method.

It is worth mentioning that, even though the Petaweb architecture was conceived within the framework of the optical layer, the topological and design concepts explored in this paper could be applied to the higher layers. For instance, a composite star architecture could be the topology of the IP/(G)MPLS layer where the edge nodes would be the entry Label Switched Router (LSR), the core nodes the core LSR, and the edge-to-edge connections could be carried out through wavelength-switched paths.

As further work, we are currently investigating the issues of update and reliability of the Petaweb structure. In fact, one of the drawbacks of the architecture could be its scalability, given that the composite star topology should have to be maintained as networks grow in sizes. Two avenues are open: carefully plan the upgrade so that the topology is respected or allow a degree of irregularity in the network. In such a case, though, reliability could become a major issue.

Appendix
Matching Costs

Here, we develop the matching costs for each block of the symmetric cost matrix presented in Section III-B.

Block 1: **Matching two inactive core nodes.** Let (i_1, r_1, e_1) be the sth core node of L_1 and (i_2, r_2, e_2) be the tth core 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 sth pair of L_2 with origin/destination (j, k) and let (i, r, e) be the tth 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: $Q_p \leq C_{\text{channel}} W s_r$.

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

$$c_{n_1+s,t} = \begin{cases} 2|M| W s_r \gamma^{(s_r - 1)} P + f_r \\ +2\phi(W) F s_r \sum\limits_{j \in M} \delta_{ij} + \beta d_{ip} Q_p, \\ \quad \text{if} Q_p \leq C_{\text{channel}} W s_r \\ \infty, \qquad \text{otherwise.} \end{cases}$$

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 sth unassigned edge node pair of L_2 and let p_2 be the tth 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 sth kit of L_3 and let $((i_2, r_2, e_2), D_2)$ be the tth kit of L_3.

If $s = t$, the element is matched with itself. The matching cost is twice the cost of one element as explained above. We remind the reader that the cost of the kit $((i_1, r_1, e_1), D_1)$ is composed of the cost of the core node, the cost of the fiber between the core node (i_1, r_1, e_1) and all edge nodes and the cost of the propagation delay of the D_1 traffic pairs via the core node (i_1, r_1, e_1). The self-matching cost is then

$$2(2|M| W s_{r_1} \gamma^{(s_{r_1} - 1)} P + f_{r_1} \\ \qquad +2\phi(W) F \sum_{j \in M} \delta_{i_1 j} + \beta \sum_{p \in D_1} d_{i_1 p} Q_p).$$

If $s \neq t$, three cases must be considered:

Case 1) All edge node pairs of D_1 and D_2 are assigned to the core node (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

$$v_I = \begin{cases} 2|M| W s_{r_1} \gamma^{(s_{r_1} - 1)} P + f_{r_1} \\ +2\phi(W) F s_{r_1} \sum\limits_{j \in M} \delta_{i_1 j} \\ +\beta \sum\limits_{p \in (D_1 \cup D_2)} d_{i_1 p} Q_p, \\ \quad \text{if} \sum\limits_{p \in (D_1 \cup D_2) \in Orig_j} Q_p \leq C_{\text{channel}} W s_{r_1}, \\ \quad \forall\, origin\ j \in (D_1 \cup D_2) \text{and} \\ \quad \sum\limits_{p \in (D_1 \cup D_2) \in Dest_k} Q_p \leq C_{\text{channel}} W s_{r_1}, \\ \quad \forall\, destination\ k \in (D_1 \cup D_2) \\ \infty, \qquad \text{otherwise.} \end{cases} \tag{14}$$

Case 2) All edge node pairs of D_1 and D_2 are assigned to the core node (i_2, r_2, e_2).

This case is the same as the one before if we reverse the roles of the core nodes. The matching cost is then

$$v_{II} = \begin{cases} 2|M| W s_{r_2} \gamma^{(s_{r_2} - 1)} P + f_{r_2} + 2\phi(W) F s_{r_2} \sum\limits_{j \in M} \delta_{i_2 j} \\ +\beta \sum\limits_{p \in (D_1 \cup D_2)} d_{i_2 p} Q_p \\ \quad \text{if} \sum\limits_{p \in (D_1 \cup D_2) \in Orig_j} Q_p \leq C_{\text{channel}} W s_{r_2}, \\ \quad \forall\, origin\ j \in (D_1 \cup D_2) \\ \quad \text{and} \sum\limits_{p \in (D_1 \cup D_2) \in Dest_k} Q_p \leq C_{\text{channel}} W s_{r_2}, \\ \quad \forall\, destination\ k \in (D_1 \cup D_2) \\ \infty, \qquad \text{otherwise.} \end{cases} \tag{15}$$

Case 3) 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. A mathematical formulation of this integer problem must be given. 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 core node (i_2, r_2, e_2) and $w_p = 0$ otherwise. Also, let z_p be a 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

$$v = \min \sum_{p\in D_1} g_p w_p + \sum_{p\in D_2} h_p z_p \qquad (16)$$

subject to

$$\begin{aligned}
&\sum_{p\in D_1\in Orig_j} -Q_p w_p + \sum_{p\in D_2\in Orig_j} Q_p z_p \le \epsilon_{wj} \\
&\qquad\qquad \forall\, origin\ j \in (D_1 \cup D_2) \quad (17)\\
&\sum_{p\in D_1\in Orig_j} Q_p w_p - \sum_{p\in D_2\in Orig_j} Q_p z_p \le \epsilon_{zj} \\
&\qquad\qquad \forall\, origin\ j \in (D_1 \cup D_2) \quad (18)\\
&\sum_{p\in D_1\in Dest_k} -Q_p w_p + \sum_{p\in D_2\in Dest_k} Q_p z_p \le \eta_{wj} \\
&\qquad\qquad \forall\, destination\ k \in (D_1 \cup D_2) \quad (19)\\
&\sum_{p\in D_1\in Dest_k} Q_p w_p - \sum_{p\in D_2\in Dest_k} Q_p z_p \le \eta_{zj} \\
&\qquad\qquad \forall\, destination\ k \in (D_1 \cup D_2) \quad (20)\\
&\qquad\qquad w_p, z_p \in \{0, 1\}, \\
&\qquad\qquad \forall\, p \in (D_1 \cup D_2). \quad (21)
\end{aligned}$$

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 the core node (i_1, r_1, e_1) and the core node (i_2, r_2, e_2), respectively. η_{wk} and η_{zk} are the surplus capacities of the links between the core node (i_1, r_1, e_1) and the core node (i_2, r_2, e_2), respectively, and the destination edge node k. Definitions are given as follows:

$$\begin{aligned}
g_p &= \beta d_{i_2 p} Q_p - \beta d_{i_1 p} Q_p, \quad \forall\, p \in D_1\\
h_p &= \beta d_{i_1 p} Q_p - \beta d_{i_2 p} Q_p, \quad \forall p \in D_2\\
\epsilon_{wj} &= C_{\text{channel}} W s_{r_1} - \sum_{p\in D_1\in Orig_j} Q_p,\\
&\quad \forall\, origin\ j \in\ (D_1 \cup D_2)\\
\epsilon_{zj} &= C_{\text{channel}} W s_{r_2} - \sum_{p\in D_2\in Orig_j} Q_p,\\
&\quad \forall\, origin\ j \in\ (D_1 \cup D_2)\\
\eta_{wk} &= C_{\text{channel}} W s_{r_1} - \sum_{p\in D_1\in Dest_k} Q_p,\\
&\quad \forall\, destination\ k \in\ (D_1 \cup D_2)\\
\eta_{zk} &= C_{\text{channel}} W s_{r_2} - \sum_{p\in D_2\in Dest_k} Q_p,\\
&\quad \forall\, destination\ k (D_1 \cup D_2).
\end{aligned}$$

The objective (16) is to minimize the cost of the packing. Equations (17) and (18) 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. Equation (19) and (20) are surplus capacity constraints for the links between the core nodes (i_1, r_1, e_1) and (i_2, r_2, e_2) respectively and each destination edge node k. Equation (21) indicates that the variables w_p and z_p are binary.

The matching cost for this case is finally

$$\begin{aligned}
v_{III} = {}& 2|M| W s_{r_1} \gamma^{(s_{r_1}-1)} P + f_{r_1} + 2\phi(W) F s_{r_1} \sum_{j\in M} \delta_{i_1 j}\\
&+ \beta \sum_{p\in D_1} d_{i_1 p} Q_p + 2|M| W s_{r_2} \gamma^{(s_{r_2}-1)} P + f_{r_2}\\
&+ 2\phi(W) F s_{r_2} \sum_{j\in M} \delta_{i_2 j} + \beta \sum_{p\in D_2} d_{i_2 p} Q_p + v. \qquad (22)
\end{aligned}$$

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

$$c_{n_1+n_2+s, n_1+n_2+t} = \begin{cases} 2(2|M| W s_{r_1} \gamma^{(s_{r_1}-1)} P + f_{r_1} \\ +2\phi(W) F s_{r_1} \sum_{j\in M} \delta_{i_1 j} \\ +\beta \sum_{p\in D_1} d_{i_1 p} Q_p), \\ \quad \text{if}\, t = s \\ \min\{v_I, v_{II}, v_{III}\}, & \text{otherwise} \end{cases}$$

where v_I, v_{II} and v_{III} are given by (14), (15), and (22), respectively.

Block 4: **Matching a kit of L_3 with an inactive core node of L_1.** Note that this is a particular case of block 6. Let $((i_1, r_1, e_1), D_1)$ be the sth kit of L_3 and (i, r, e) be the tth core node of L_1. The inactive core node (i, r, e) can be seen as an active core node with no assigned pair: $(i, r, e) = ((i_2, r_2, e_2), \emptyset) \in L_3$. The matching cost is then:

$$c_{n_1+n_2+s, t} = \min\{v_I, v_{II}, v_{III}\},$$

where v_I, v_{II} and v_{III} are given by (14), (15) and (22).

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 sth kit of L_3 and q be the tth pair of L_2 with origin/destination (j, k). Two cases must be considered:

Case 1) The unassigned edge node pair can be assigned to the core node (i_1, r_1, e_1). Then D_1 becomes $D_1 \cup \{q\}$.

This case is possible if: the link capacity between the origin edge node j 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 the destination edge node k on the other hand, are respected.

The matching cost for this case is (now $q \in D_1$)

$$c_{n_1+n_2+s, n_1+n_2+t} = \begin{cases} 2|M| W s_{r_1} \gamma^{(s_{r_1}-1)} P + f_{r_1} \\ \quad +2\phi(W) F s_{r_1} \sum_{j\in M} \delta_{ij} \\ \quad +\beta \sum_{p\in D_1} d_{ip} Q_p, \\ \quad \text{if} \sum_{p\in D_1\in Orig_j} Q_p \le C_{\text{channel}} W s_{r_1}, \\ \quad \forall\, origin\ j \in D_1, \\ \quad \text{and} \sum_{p\in D_1\in Dest_k} Q_p \le C_{\text{channel}} W s_{r_1}, \\ \quad \forall\, destination\ k \in D_1. \end{cases}$$

Case 2) The unassigned edge node pair cannot be assigned to the core node (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 is built as an empty set. $D_1 = D_1 \cup \{q\}$ and $D_2 = \emptyset$.

We solve the problem (16)–(21) without considering the constraints (18) and (20) where now w_p is defined as being equal to 1 if the pair $p \in D_1$ is detached from the core node (i_1, r_1, e_1) and becomes an unassigned pair of L_2, and 0, otherwise. Also, $z_p = 0$, $g_p = \mathcal{M} - \beta d_{i1p} Q_p$, $\forall\ p \in D_1$ $h_p = 0$. Note that the surplus capacity ϵ_{wj} and η_{wk} can be negative. The set $\overline{D_1} \subset D_1$ corresponds to the edge node pairs assigned to the core node (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

$$\begin{aligned} c_{n_1+n_2+s,n_1+t} = {} & 2|M|Ws_{r_1}\gamma^{(s_{r_1}-1)}P + f_{r_1} \\ & + 2\phi(W)Fs_{r_1}\sum_{j\in M}\delta_{i_1j} \\ & + \beta \sum_{p\in\overline{D_1}} d_{i_1p}Q_p + (n_1 - \overline{n_1})\mathcal{M}. \end{aligned}$$

At last, the matching cost for the block 5 is

$$c_{n_1+n_2+s,n_1+t} = \begin{cases} 2|M|Ws_{r_1}\gamma^{(s_{r_1}-1)}P + f_{r_1} \\ +2\phi(W)Fs_{r_1} \sum_{j\in M} \delta_{i_1j} \\ +\beta \sum_{p\in D_1} d_{i_1p}Q_p, \\ \quad \text{if} \sum_{p\in D_1\in Orig_j} Q_p \le C_{channel}Ws_{r_1}, \\ \quad \forall\ origin\ j \in D_1, \\ \quad \text{and} \sum_{p\in D\in Dest_k} Q_p \le C_{\text{channel}}Ws_{r_1} \\ \quad \forall\ destination\ k \in D_1 \\ 2|M|Ws_{r_1}\gamma^{(s_{r_1}-1)}P + f_{r_1} \\ +2\phi(W)Fs_{r_1} \sum_{j\in M} \delta_{i_1j} \\ +\beta \sum_{p\in\overline{D_1}} d_{i_1p}Q_p + (n_1 - \overline{n_1})\mathcal{M}, \\ \quad \text{otherwise.} \end{cases}$$

ACKNOWLEDGMENT

The authors would like to thank F. Blouin and M. Beshai of Nortel Networks for proposing the problem and for stimulating discussions on this topic, J. Dégila for his suggestions, M. Forbes for his explanations related to his thesis work, and the anonymous referees for their thorough review and suggestions.

REFERENCES

[1] R. Vickers and M. Beshai, “Petaweb architecture,” presented at the 9th Int. Telecommun. Network Planning Symp. (Networks 2000): Towards Natural Networks, Toronto, Canada, 2000.

[2] M. Beshai, F. Blouin, and R. Krishnan, “Petaweb—Building block for a yottabit-per-second network,” DARPA Next Generation Internet, Technology Investment Agreement TIA.

[3] F. Blouin, A. Lee, A. Lee, and M. Beshai, “A comparison of two optical-core networks,” *J. Opt. Networking*, vol. 1, no. 1, pp. 56–65, 2002.

[4] J. Dégila and B. Sansò, “Design optimization of a next generation Yottabit-per-second network,” in *Proc. IEEE Int. Conf. Communications (ICC 2004)*, Jun. 2004, pp. 1227–1231.

[5] J. Dégila and B. Sansò, “Topological optimization of a yottabit-per-second lattice network,” *IEEE J. Sel. Areas Commun.*, vol. 22, no. 9, pp. 1613–1624, Sep. 2004.

[6] J. Klincewicz, “Hub location in backbone/tributary network design: A review,” *Location Sci.*, vol. 6, pp. 307–335, 1998.

[7] M. Balinski, “On finding integer solutions to linear programs,” in *Proc. IBM Scientific Computing Symp. Combinatorial Problems*, 1966, pp. 225–248.

[8] M. Rönnqvist, S. Tragantalerngsak, and J. Holt, “A repeated matching heuristic for the single-source capacitated facility location problem,” *Eur. J. Oper. Res.*, vol. 116, pp. 51–68, 1999.

[9] M. Rönnqvist, K. Holmberg, and D. Yuan, “An exact algorithm for the capacitated facility location problems with single sourcing,” *Eur.J. Oper. Res.*, vol. 113, pp. 544–559, 1999.

[10] Y. Huei and P. Keng, “Framework for shared time-slot TDM wavelength optical WDM networks,” *J. Opt. Networking*, vol. 5, pp. 564–567, 2006.

[11] M. Forbes, J. Holt, P. Kilby, and A. Watts, “A matching algorithm with application to bus operations,” *Austral. J. Combinatorics*, vol. 4, pp. 71–85, 1991.

[12] M. Engquist, “A successive shortest path algorithm for the assignment problem,” in *Proc. INFOR*, 1982, vol. 20, pp. 370–384.

[13] R. Jonker and A. Volgenant, “A shortest augmenting path algorithm for dense and sparse linear assignment problems,” *Computing*, vol. 38, pp. 325–340, 1986.

[14] U.S. Census 2000 Population and Housing. U.S. Dept. Commerce and Administration, 2002 [Online]. Available: http://www.census.gov/main/www/cen2000.html

[15] The National Atlas of the United States of America. U.S. Dept. Interior Geological Survey, Washington, D.C., 1970.

[16] Geographic Co-ordinates. Faculty of Environmental Studies, Dept. Geography, Univ. Waterloo, Canada, Sep. 15, 2003 [Online]. Available: http://www.fes.uwaterloo.ca/crs/geog165/gcoords.htm

[17] *Airmaps Inc.* Minneapolis, MN: Cartographic Lab., Dept. Geography, Univ. Minnesota, 1989.

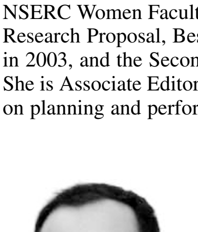
Anne Reinert received the M.Sc. degree from the École Polytechnique de Montréal, Montréal, QC, Canada, and the Engineering degree from the École Supérieure de l’Electricité (Supélec), Paris, France.

Her fields of interest are telecommunications and operational research.

Brunilde Sansò (M’92) received the degree in Electrical Engineering from Universidad Simón Bolívar, Caracas, Vanazuela, in 1981, and the M.Sc.A. degree in electrical engineering and the Ph.D. degree in applied mathematics from the École Polytechnique de Montréal, Montréal, QC, Canada, in 1985 and 1988, respectively.

After being a post-doc and a researcher at the CRT and the GERAD centers, respectively, she joined the Faculty of the École Polytechnique in 1992, where she has been a full Professor since 1997. Her interests are in reliability, design, performance, quality of service, routing and operational planning of telecommunication networks.

Dr. Sansò has been the recipient of several awards and honors such as the NSERC Women Faculty Award, FCAR Young Researcher Award, AQTR Best Research Proposal, Best Paper Awards from IEEE/ASME in 1995 and DRCN in 2003, and the Second Prize in the CORS OR practice competition in 2003. She is Associate Editor of *Telecommunication Systems* and editor of two books on planning and performance.



Stefano Secci (S’05) received the M.Sc. degree in telecommunication engineering from Politecnico di Milano, Milan, Italy, in 2005. He is currently working toward a double Ph.D. degree at TELECOM ParisTech (ENST), Paris, France, and Politecnico di Milano, Milan, Italy.

He performed a research internship with the École Polytechnique de Montréal. Previously, he was an Assistant Researcher with Politecnico di Milano and a Service Engineer with Fastweb Italia S.p.a. His interests include quality-of-service routing and traffic engineering, network design, and dimensioning for optical and IP networks.