Virtualizing and Scheduling Optical Network Infrastructure for Emerging IT Services [Invited]

Pascale Vicat-Blanc Primet, Sebastien Soudan, and Dominique Verchere

***Abstract*—Emerging IT service providers that aim at delivering supercomputing power available to the masses over the Internet rely on high-performance IT resources interconnected with ultra-high-performance optical networks. To adjust the provisioning of the resources to end-user demand variations, new infrastructure capabilities have to be supported. These capabilities have to take into account the business requirements of telecom networks. This paper proposes a service framework to offer Internet service providers dynamic access to extensible virtual private execution infrastructures, through on-demand and in-advance bandwidth and resource reservation services. This virtual infrastructure service concept is being studied in the CARRIOCAS project and implemented thanks to the scheduling, reconfiguration, and virtualization (SRV) component. This entity handles the service requests, aggregates them, and triggers the provisioning of different types of resources accordingly. We propose to adapt to envisioned heterogeneous needs by multiplexing rigid and flexible requests as well as coarse or fine demands. The goal is to optimize both resource provisioning and utility functions. Considering the options of advanced network bandwidth reservations and allocations, the optimization problem is formulated. The impacts of the malleability factor are studied by simulation to assess the gain.**

***Index Terms*—Network optimization; Assignment and routing algorithms**.

I. MOTIVATIONS

Three current trends are envisioned to strongly influence the development of the future Internet: (i) the convergence of the communication, computation, and storage aspects of the Internet; (ii) the deployment of ultra-high-capacity interconnection networks with predictable performance; (iii) coarse-grain web servers like Google, Yahoo, and Amazon providing the content, control, storage, and computing resources for the users. These three trends raise major research issues in networking and services, requiring a new vision of the network. Indeed, the current Internet stack (TCP/IP) and its associated simple network management protocol are not consistent with the evolution of the network infrastructure components and its use by emerging services that aim to deliver supercomputing power available to the masses over the Internet. The coordination of networking, computing, and storage requires the design, development, and deployment of new resource management approaches to discover, reserve, coallocate, and reconfigure resources and schedule and control their use.

Indeed, many emerging large-scale applications and services want time-limited but ultra-high-bit-rate network services of the order of the transmission capacity of the network infrastructures. For example, large-scale distributed industrial applications require use of ultra-high-performance network infrastructure and multiple types of end capacities such as computational, storage, and visualization resources. Collaborative engineers need further to interact with massive amounts of data to analyze the simulation results over high-resolution visualization screens fed by storage servers. To illustrate the variety of requirements, Table I classifies the use cases specified by developers and users contributing to the CARRIOCAS project [1]. The applications are listed depending on their connectivity characteristics, capacity requirements, localization constraints, performance constraints, and scenarios between networks and IT services offered to end users. These use cases combine the following requirements: (i) real-time remote interactions with constant bandwidth requirements, (ii) many large data file transfers between distant sites whose localization is known in advance (distributed storage and access to data centers) and with sporadic bandwidth requirement, (iii) streaming of many data files between anonymous sites (e.g., multimedia production)

Manuscript received October 30, 2008; revised January 8, 2009; accepted January 24, 2009; posted January 24, 2009; published June 25, 2009 (Doc. ID 103471).

P. Vicat-Blanc Primet and S. Soudan are with INRIA, University of Lyon, France (e-mail: Pascale.Primet@inria.fr).

D. Verchere is with Alcatel-Lucent Bell Labs, France.

Digital Object Identifier 10.1364/JOCN.1.00A121

1943-0620/09/02A121-12/$15.00

TABLE I
CLASSIFICATION OF DISTRIBUTED APPLICATIONS

Type	Characteristics	Capacity Required	Localization Constraints	Performance Constraints	Applications
Type A Parallel distributed computing tightly synchronized computation	Scientific applications involving tightly coupled tasks	High	Low	Computation, connectivity latency	Grand scientific problems requiring huge memory and computing power
Type B High-throughput low-coupling computation	Great number of independent and long computing jobs	High	Low	Computation	Combinatory optimization, exhaustive research (crypto, genomics), stochastic simulations (e.g., finance)
Type C Computing on demand	Access to a shared remote computing server during a given period	Medium	Low	Computation, high bit rate, connectivity latency	Access to corporate computing infrastructure, load balancing between servers, virtual prototyping
Type D Pipeline computing	Applications intensively exchanging streamed or real-time data	Medium	Average	Computation, bit rate, latency, storage	High-performance visualization, real-time signal processing
Type E Massive data processing	Treatment of distributed data with intensive communication between computer and data sources	Low	High	High bit rate and connectivity latency	Distributed storage farms, distributed data acquisitions, environmental data treatment, bioinformatics, finances, high-energy physics
Type F Largely distributed computing	Research, modification on distributed databases with low computing and data volume requirements	Medium	High	Connectivity latency	Fine-grained computing applications, network congestion can cause frequent disruptions between clients
Type G Collaborative computing	Remote users interacting on a shared virtual visual space	Medium	High	Bit rate and latency (for interactivity)	Collaborative visualization for scientific analysis or virtual prototyping

requiring statistical guarantees, and (iv) data and file transfers between locations carried on best-effort network services. For example, applications of type A or B can be executed on any server but have high bandwidth and latency requirements. Applications of type F or G involve specific servers and instruments but are not constrained much by the bandwidth.

However, the current Internet, built around the IP protocol, routers, and links, is a best-effort system with an unpredictable service and was not designed for competing flows on ultra-high-end optical links, nor for highly variable traffic. Recently, pilot experiments in optical/photonic switching and hybrid networking [2–4] have shown that predictable connectivity with low jitter and massive dedicated throughput is possible. These quality of service (QoS) properties can be delivered to the application only through careful direct control of the networking resources.

We then argue that exposing optical bandwidth as well as processing and storage capacities within the network will help in supporting the ever-growing spectrum of communication patterns and ways to use the Internet. We propose to decouple the physical layer from the service level to increase the flexibility and the efficiency in the usage of network infrastructures. We introduce the concept of a virtual infrastructure that corresponds to the time-limited interconnection of end capacities with virtual private networks. This concept is being explored in the CARRIOCAS project. The idea is to expose the composition and the access to virtual infrastructures as an intermediate service. CARRIOCAS explores in particular the technical aspects of the hybrid packet/optical network virtualization.

CARRIOCAS also studies the commercial needs of Internet service providers and those of network operators. Indeed, a commercial interface to invoke a virtual infrastructure service will include the maximum cost customers are willing to sustain. On the other side, network providers will publish their services according to their policies and will negotiate to maximize utilization rates. However, the granularity a realistic and efficient capacity (for example, optical bandwidth) reservation service can offer may not be flexible enough to meet the heterogeneity of customers' requirements. This may lead to poor resource uti-

lization, overprovisioning, and unattractive service pricing. Therefore, this paper investigates models, algorithms, and software for flexibly sharing and scheduling the optical infrastructure considering the lambda path as the first class resource.

This paper is organized as follows: Section II gives an outline of the CARRIOCAS project. Its specific virtual infrastructure management component, scheduling, reconfiguration, and virtualization (SRV), is presented in Section III. Section IV defines the model and formulates the problem for flexible bandwidth allocation. In Section V, simulation results are presented to demonstrate the impact of the patience and the influence of the proportion of flexible requests in the mix. Related works are reviewed in Section VI. Finally, we conclude in Section VII.

II. Overview of the CARRIOCAS Project

The CARRIOCAS project aims at delivering commercial services for accessing virtual private IT infrastructures over ultra-high-speed (40 Gb/s) networks. The goal is to provide frameworks enabling users to reserve (dynamically or in advance) computing and storage resources but also network resources to connect them for composing customized virtual infrastructures. We envision the emergence of a capacity service provider entity, between Internet service providers and network operators with the central role of exposing and interconnecting virtual resources located at data centers and company customers and managing more dynamically the network services (see Fig. 1).

The CARRIOCAS project defines and develops the virtual infrastructure concept and designs the (SRV) (in the middle of Fig. 1) module to manage them. Indeed to implement the envisioned capacity service, which we name virtual infrastructure service, a specific management component has to be introduced. New management and control functions are required to adapt existing telecom network infrastructures interconnected in the current Internet to deliver such services to ISPs or company customers. The SRV module process requests complex virtual infrastructure composition and reservation. These requests formulate explicit reservation of network resources with other types of resources (e.g., computational, storage).

Because the connectivity service needs are heterogeneous (see Table I), several classes of service with high-performance QoS parameter values in terms of bandwidth and edge-to-edge latency are exposed. The connection configuration must be dynamic to adapt to different usage patterns. The network services must be adjusted in response to the changing customer environments such as new organizations joining and then connecting on network and data center infrastructures (e.g., a car designer is going to be connected to a complex fluid dynamics simulation application next month for two weeks). The network has to pro-

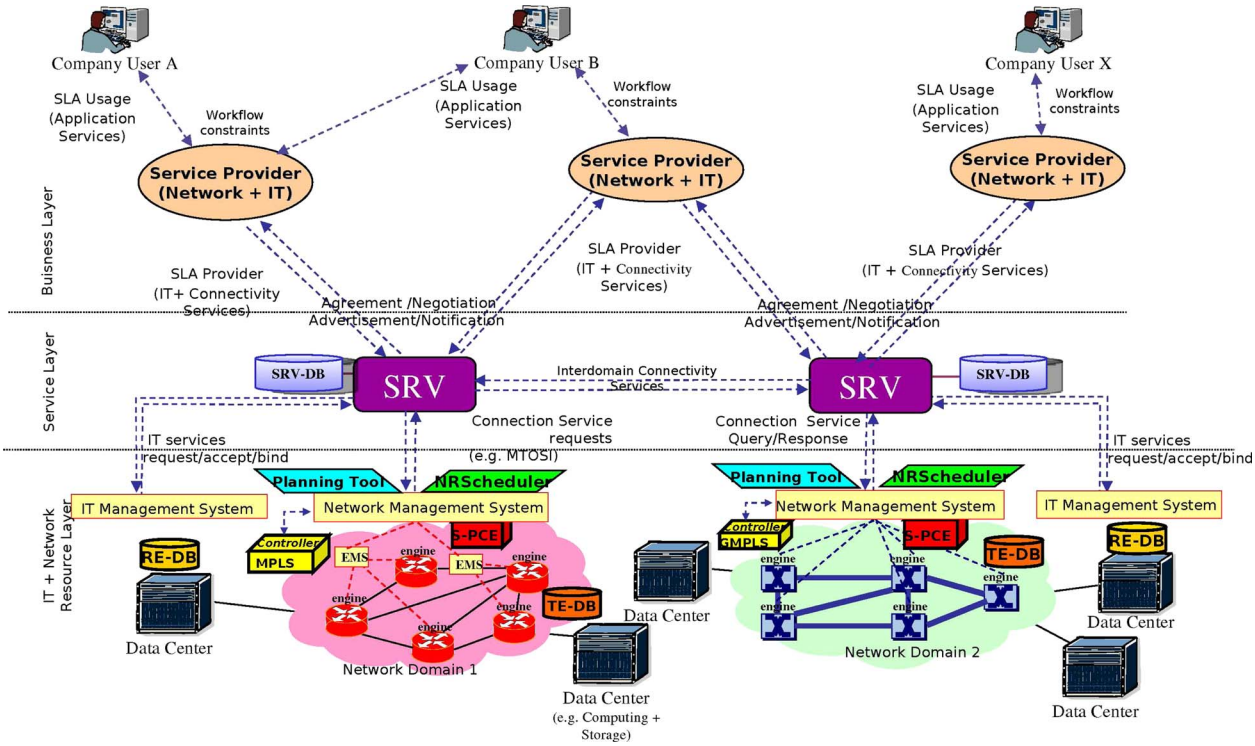


Fig. 1. (Color online) SRV architecture.

vide a dynamic and automatic (re)configuration of the network device features.

At the upper level, corresponding to the north interface of the SRV, the connectivity service classes describe the service constraints from the external customer points of view, and on the south interface of the SRV, the connection service classes describe the service attributes to allow the resource provisioning by the network management system. The network services are structured according to the information and data model as recommended by the multitechnology operations systems interface (MTOSI) [5].

The CARRIOCAS project focuses on layer 2 connections, leveraging the cost advantages of carrier-grade Ethernet services. Layer 2 virtual private network services (L2-VPN) can provide secure and dedicated data communications over telecom networks, through the use of standard tunneling, encryption, and authentication functions. These L2-VPNs are contrasted from leased lines configured manually and allocated to one company. An L2-VPN is a network service dedication achieved through logical configuration rather than dedicated physical equipment with the use of the virtual technologies. For example, in the CARRIOCAS pilot network, three classes of carrier Ethernet connectivity services called Ethernet virtual circuits (EVCs), are proposed to customers: point-to-point EVC, multipoint-to-multipoint EVC, and rooted multipoint EVC. The description of each EVC includes the user-to-network interface (UNI) corresponding to the reference point between the provider edge (PE) node and customer edge (CE) node. At a given UNI more than one EVC can be provisioned or signaled according the multiplexing capabilities.

The SRV, acting as an autonomous functional entity, reconfigures automatically the provisioning of VPNs. It accommodates addition, deletion, moves, and/or changes of access among data center sites and company members with no manual intervention.

III. Architecture of the SRV Module

The SRV module is in charge of the management of composite services to provide virtual infrastructures, or VXI, built from the composition of computing, storage, and networking resources. Further specific large-scale software applications can be associated as service elements of the composite service too.

The SRV functions are positioned between the external service provider (SP) and the network resource management and control functions as well as the IT resource management functions of the data centers.

The SRV module can be owned by an administrative entity that is external or internal to the network infrastructure owner and operator. In the CARRIOCAS architecture, the SRV is part of the network infrastructure operations and is defined over a single administrative network domain. SRV integrates different service interfaces according to the role of the entities it interacts with.

A. SRV North Interface and Internal Functions

The SRV architecture (Fig. 2) integrates the functional blocks required to deliver connectivity services as well as IT resource services fulfilling the service level agreement (SLA) requests from an SP customer. The SRV functional components interacting with the customers are the service publication, service negotiation, and service notification. These components respectively allow publishing, negotiating, and notifying the connectivity services from the SRV towards external SP customers. This interface (called the request handler in Fig. 2) is based on the VXDL (virtual infrastructure description language) [6], compliant with the service specifications of the open grid service architecture (OGSA), and based on RDF/XML. VXDL has the advantages of enabling the composition of heterogeneous service elements (e.g., a computing service combined with a connectivity service). A request (VXI request) formulated in VXDL can even include, for example, the amount of IT and network resources required, their type, the class of end devices required to deliver the services with performance, duration, or other timing constraints.

At the mediation layer, the bandwidth and resource allocator and scheduler selects resources and schedules them to serve the requests. Cross-optimization is performed through the different types of resources reserved. Heterogeneous criteria including the computational capability of the server (e.g., operating system type, computing capacity, memory space) and the available networking capability (interfaces, band-

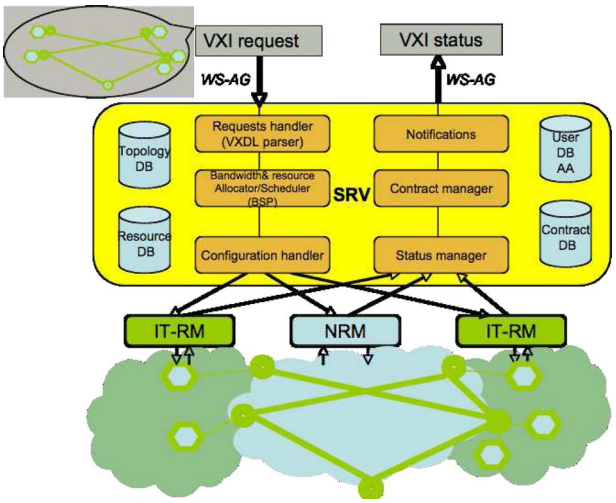


Fig. 2. (Color online) SRV internal functions.

width, latency) on the selected path are taken into account. The contract manager is in charge of verifying the correct execution of the customer's contract in terms of resource usage (customer side) as well as resource provisioning (provider side).

B. SRV Interfaces With Network and IT Infrastructures

The configuration manager in Fig. 2 interacts directly with the physical infrastructure. Towards the network infrastructure, the SRV southbound interface is based on two options: SRV requests for switched connections through the provider UNI (UNI-N) if it is a generalized multiprotocol label switching (GMPLS) controlled network [7] or SRV requests for permanent connections or soft-permanent connections through the network management interface. The connection service description is based on web-service description language (WSDL) as recommended by MTOSI from the Tele-Management Forum [5]. Connections are dynamically established [provisioned through the network management system (NMS) or signaled through UNI] with guaranteed QoS parameters defined from the SLA provider.

The connectivity services require a pool of resources to be explicitly reserved within the network. Now, the resource pools are computed by the network resource planning functions, i.e., network planning tool (NPT) (see Fig. 1). Today network services are provisioned in the infrastructure and are not automatically related to how the business rules operate during the publication, the negotiation, and the service notification to the customers. In another way, to enable a more autonomic connection provisioning, the NMS must cope with the scheduled service deliveries. The NMS has to be extended to ensure that the infrastructure delivers the on-demand or scheduled network services asked for by external entities.

The SRV East–West interface supports edge-to-edge connectivity service covering multiple routing domains. The SRV interfacing the SP manages the network service at the ingress network domain and the interdomain connectivity services through peer operation information exchanges with the SRVs involved in the chain [8].

IV. Dynamic Bandwidth Provisioning Approach

Due to the large variety of distributed applications as reported in Table I, it is considered that the subsequent application demands for IT and connectivity services can be very disparate in terms of resource requirements and very sporadic in terms of time windows. Each service demand can be either allocated explicitly on a defined set of reserved resources or several service demands can be multiplexed on a pool of resources.

The different actors have their respective objective to maximize their utility functions. For an end customer, the objective is to execute the submitted jobs at low costs within a defined time window. For a service provider it means to deliver connectivity services and the other IT services at the highest quality including performance and security and lowest cost. Each resource operator wants to maximize infrastructure utilization (CAPEX) with the minimal operation efforts (OPEX) to finally obtain the best possible return on investment (ROI).

The SRV management component handles requests, aggregates them, and provisions the resources accordingly. Considering the special case of the optical networks and dynamic bandwidth allocations, we address the problem of discrepancy between realistic and efficient (optical) bandwidth granularity and customer requirements by mixing different allocation strategies within the SRV itself. Rigid reservations (for real-time video and audio conferencing applications, for example) offer determinist bandwidth provisioning. Flexible reservations provide only time guarantees that a specified amount of data streaming is transferred within a strict time window. The scheduling component of the SRV is in charge of aggregating the heterogeneous requests and of provisioning the bandwidth on an edge-to-edge path accordingly. Consequently Internet service providers can flexibly express their bandwidth amount requests with minimal rate and maximal deadline, but also with volume and maximal achievable rate or even bandwidth profile as detailed below. The solution proposed in [9] carrying out and grouping giant transfer tasks (with volume higher than several gigabytes) in specified time intervals exemplifies this idea. Such an approach considers that most company customers and Internet service providers will carry on transferring large amounts of data in a limited and predictable time frame rather than requiring exactly a fixed rate for a given time window. This relieves the complex bandwidth reservation and allocation burden, while ensuring the flow completion time, the most useful metric [10] for a user or a data processing application.

In the rest of the paper we demonstrate that the flexibility proposed increases the efficiency in lambda path usage. In particular, we study how the virtual infrastructure management component, integrating this scheduling approach, can optimally serve different request pattern distributions mixing rigid and malleable resource requests.

A. Model and Problem Formulation

Let us consider a network model defined as an opaque cloud, owned by a network operator, exposing a dynamic bandwidth provisioning service. This service provides on-demand lambda path or provisioned

links between a set of access points (end points) and eventually peering points with other network clouds. The network is defined by its set of points of presence $s \in \{s_1, \ldots, s_S\}$. We assume here that any two points exposed by one cloud can be the source and destination of a network reservation. The basic service offered here by the simplified SRV is bandwidth leasing between two end points for a specific time window or the guaranteed transfer of a massive data set from one site to an other with strict completion deadlines. The bandwidth scheduling component, which schedules users' transfers with the help of the updated virtual topology and resource databases, tries to maximize the resource usage. We consider here that this internal component does not have access to the routing plane. It only manages movement capacity or bandwidth capacity between the network points of presence. For the sake of simplicity, we distinguish this entity, called the bandwidth service provider (BSP), from the network operator (NO) providing and managing the physical infrastructure. The three main actors in this model are then (i) the users, who aim to rent fixed bandwidth for a time window or to transfer data volumes from one site to another with strict time constraints; (ii) the BSP, which schedules users' requests on network paths and tries to maximize the resource usage; and (iii) the NO, which provides the connectivity service (i.e., constant bandwidth between two end points for a time slot). This service can be realized, for example, through point-to-point EVC service or lambda path service.

Figure 3 shows the relations between actors and the request format. Users are issuing transfer requests to the BSP, which groups them and issues connectivity requests to the NO based on the connectivity and time window requirements of each group. As shown in this section, users to BSP requests can have different forms.

Users can send transfer requests to the BSP. They are defined as follows: Each request r is associated

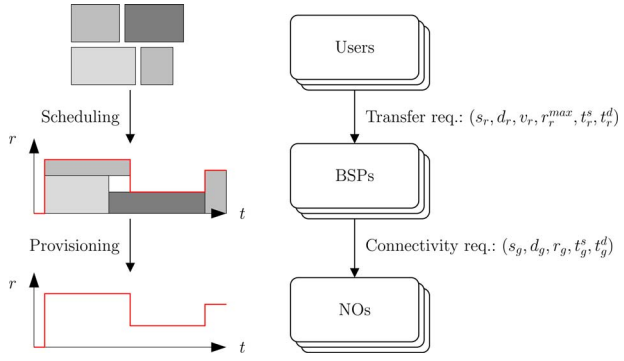


Fig. 3. (Color online) Transfer and connectivity requests exchanged between actors. BSPs group users' requests in connectivity requests issued to the NO.

with a 6-uple $(s_r, d_r, v_r, r_r^{max}, t_r^s, t_r^d)$, where s_r is the source, d_r is the destination, v_r is the volume to transfer, and r_r^{max} is the maximum instant rate used to carry v_r. Transfer can only start after t_r^s and must be finished before t_r^d. t_r^a is the arrival date of request r and t_r^r is the date of the request acceptance decision. Furthermore, for a request expressed with a volume, r_r^{min} is defined as $r_r^{min} = v_r/(t_r^d - t_r^s)$ and patience P_r as $P_r = r_r^{max}/r_r^{min}$. The request cannot be served and is invalid if $P_r < 1$ due to constraint 1, which is defined thereafter.

Users can also ask for another kind of service, which is bandwidth based compared with the previous one, which was volume based. This can be achieved by tailoring the transfer request so that there is no flexibility and requests can only be served using the requested bandwidth. Rigid requests have $P_r = 1$. To obtain rate R between t_r^s and t_r^d, this kind of rigid request is expressed as the 5-uple $(s_r, d_r, R, t_r^s, t_r^d)$ rewritten as the 6-uple $[s_r, d_r, (t_r^d - t_r^s)R, R, t_r^s, t_r^d]$. This request rewriting can be done internally. Therefore BSPs expose two different services with two request formats: one for malleable bulk transfers with transfer requests and one for bandwidth on demand with rigid requests.

At a low level, the connectivity requests specify what BSPs will ask of NOs. This kind of request is not malleable and is defined as follows: a connectivity request g is expressed as a 5-uple $(s_g, d_g, r_g, t_g^s, t_g^d)$, where s_g is the source, d_g is the destination, r_g is the requested rate, t_g^s is the start time, and t_g^e is the end of the reservation. Similarly, to transfer requests, t_g^a is the arrival date of request g. We assume that connectivity service is requested at $t_g^s - A$ (A in advance) and by slots of duration D (i.e., $t_g^d - t_g^s = D$).

Assumption 1. *Connectivity request issued for slot n is constrained to have $t_g^a \leq t_g^s - A$, $t_g^s = nD$, and $t_g^e = (n+1)D$.* If this request is issued by BSP *sp* to NO *no* at $t_g^a = mD - A$ with $m \leq n$, it will be noted $g_{sp,no}^{m,n}$. Final connectivity request for slot n is thus noted: $g_{sp,no}^{n,n}$. To avoid overestimation of in-advance connectivity reservations, we assume (Assumption 2) that a BSP can only reprovision for a given slot, by increasing the rate demand.

Assumption 2. *At time $m'D - A$, when updating advance connectivity requests previously made at $mD - A$ for slot n, BSP can only increase requested bandwidth. More formally, $\forall m < m' \leq n, r_{g_{sp,no}^{m,n}} \leq r_{g_{sp,no}^{m',n}}$.*

This is illustrated in Fig. 4, where new connectivity requests for slots n and $n+1$ issued at time $nD - A$ are shown as solid lines while old connectivity requests are dashed lines.

Assumption 3. *BSPs can only rent end-to-end connectivity resources to the network operator. BSPs do not have routing facilities.*

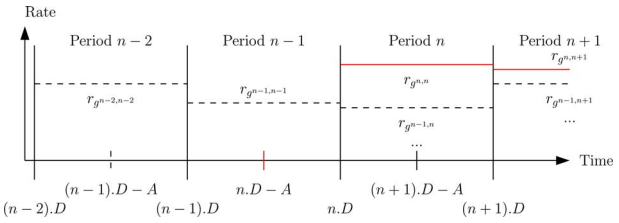


Fig. 4. (Color online) Connectivity service's slots $n-2$ to n as seen at time $nD-A$.

Assumption 3 is based on a realistic situation in which network operators do not provide routers nor direct access to routing to their customers. This implies that there is no routing opportunity from the BSP's point of view. Routing issues will be addressed by the NOs. This also avoids the need to expose detailed topological information of the core network.

We define the bandwidth allocation profile of r as a step function $t \mapsto p_r(t)$ defining the rate allocated to this transfer over time. A valid bandwidth allocation profile p_r must verify constraints 1, 2, and 3:

$$\forall t \in [t_r^s, t_r^d], \quad 0 \leq p_r(t) \leq r_r^{max}, \tag{1}$$

$$\forall t \notin [t_r^s, t_r^d], \quad p_r(t) = 0, \tag{2}$$

$$\int_{t_r^s}^{t_r^d} p_r(t)\mathrm{d}t = v_r, \tag{3}$$

where a_r^s is the actual start time of transfer r and a_r^f is its actual finish time. More formally, $a_r^s = \min\{t \,|\, p_r(t) \neq 0\}$ and $a_r^f = \max\{t \,|\, p_r(t) \neq 0\}$.

In this model, BSPs do connectivity reservations with the same source and destination sites as transfer requests. Connectivity reservations have to satisfy the following constrains to support transfer requests. Let us consider a set of N transfer requests $R = \{r_1, \ldots, r_N\}$ and a set of M nonoverlapping [i.e., $\forall g, g' \in G, (t_g^s, t_g^e) \cap (t_{g'}^s, t_{g'}^e) = \emptyset$] connectivity requests $G = \{g_1, \ldots, g_M\}$; the validity constraints are (in addition to 1, 2, and 3 for each request in R) the following:

$$\forall g \in G, \quad \forall t \in [t_g^s, t_g^e], \quad \sum_{r \in R} p_r(t) \leq r_g, \tag{4}$$

$$\forall r \in R, \quad \forall t \notin \bigcup_{g \in G} [t_g^s, t_g^e], \quad p_r(t) = 0, \tag{5}$$

$$\forall r \in R, \quad s_r = s_g, \tag{6}$$

$$\forall r \in R, \quad d_r = d_g. \tag{7}$$

Figure 5 shows a group of transfer requests and the corresponding connectivity request to serve them. Due to Assumption 3, and without loss of generality, the rest of this paper focuses on one source–destination pair as transfer requests with a different source or destination that do not interact.

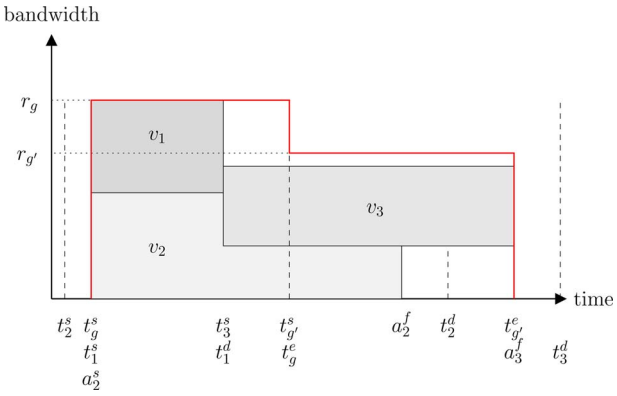


Fig. 5. (Color online) Transfer requests ($r \in \{1,2,3\}$) grouped in connectivity requests g and g'.

B. Request State Diagram

Users' transfer requests or rigid requests are received and processed by BSPs. The BSP's transfer request state diagram comprises four states: new, scheduled, granted, and rejected.

A request r is (i) new when the request has just been received and is valid but has been neither accepted nor rejected; (ii) scheduled when it has been accepted, but allocated profile $t \mapsto p_r(t)$ can still be changed; (iii) granted when it cannot be changed anymore; (iv) and finally rejected when the request is not accepted by the BSP. These states and allowed transitions are depicted in Fig. 6. The state of request r is changed from scheduled to granted at time $t_r^s - a$ in order to give the sender some time before the transfer start time. Transitions from state new to scheduled or rejected depend on the decision made by the BSP when first scheduling this request. Note that once the request has been marked as scheduled, its state cannot be changed to rejected.

Assumption 4. *In this work, we assume that they are never rejected.*

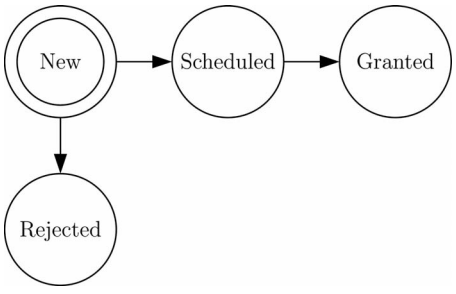


Fig. 6. Transfer request state diagram: once scheduled a transfer cannot be rejected anymore, but the profile can still change until the transfer reaches the granted state.

Assumption 4 implies that in-advance transfer requests can all be accepted. The BSP has infinite resources as demonstrated in the next section.

C. Scheduling and Provisioning

In the remaining, interactions between one given BSP and one given NO are considered, so *sp* and *no* subscripts will thus be omitted in notations.

For every slot n at $nD-A$, the BSP has to decide for each source–destination pair which connectivity request to issue for the next slot to accommodate the already-known transfer requests R_n ($\forall r \in R_n$, $t_r^a \leq nD-A$) and how to schedule these transfers. The problem presented in this section addresses this issue when rejection is not allowed and the objective is to minimize the aggregated provisioned capacity.

According to the previously defined state diagram, R_n is partitioned into three subsets $R_n = R_n^{new} \cup R_n^{sched.} \cup R_n^{granted}$, where $R_n^{granted}$ is the set of requests already granted during slots before $nD-A$, R_n^{new} contains new valid requests that have not yet been scheduled (or rejected) and $R_n^{sched.}$ contains transfer requests that can still be rescheduled. It can be noted that requests in $R_n^{sched.} \cup R_n^{granted}$ were already in R_{n-1}. Figure 7 summarizes this.

Let $G_n^{final} = \{g^{i,i} | 1 \leq i \leq n-1\}$ be the set of connectivity reservation made for slots up to $n-1$ (they cannot be modified anymore), $G_n^{prev.} = \{g^{n-1,i} | n \leq i \leq M\}$ and $G_n^{new} = \{g^{n,i} | n \leq i \leq M\}$, where M is the greatest slot utilized by a transfer request.

Using previously defined validity constraints for connectivity requests and constraints on increasing connectivity requests, we formulate the problem as

$$BP(n)\text{: minimize: } \sum_{g \in G_n^{final} \cup G_n^{new}} r_g$$

subject to:

$$\forall r \in R_n, \quad \forall t \in [t_r^s, t_r^d], \quad 0 \leq p_r(t) \leq r_r^{max},$$

$$\forall r \in R_n, \quad \forall t \notin [t_r^s, t_r^d], \quad p_r(t) = 0,$$

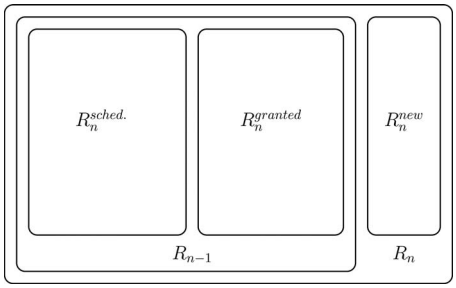


Fig. 7. Partitioning of R_n.

$$\forall r \in R_n, \quad \int_{t_r^s}^{t_r^d} p_r(t)\mathrm{d}t = v_r,$$

$$\forall g \in G_n^{final} \cup G_n^{new}, \quad \forall t \in [t_g^s, t_g^e], \quad \sum_{r \in R_n} p_r(t) \leq r_g,$$

$$\forall t \notin \left(\bigcup_{g \in G_n^{final} \cup G_n^{new}} [t_g^s, t_g^e] \right), \quad \forall r \in R_n, \quad p_r(t) = 0,$$

$$\forall i, \quad n \leq i \leq M, \quad r_{g^{n-1,i}} \leq r_{g^{n,i}}.$$

Let I be the set of time intervals defined by dividing the time axis on all t_g^s, t_g^e, t_r^s, and t_r^d. In this case, $\delta_{r,i}$ will be 1 if request r can be served on interval i and 0 else, $\gamma_{g,i}$ equals 1 if connectivity request g covers interval i and 0 else (all i are supposed to be covered by a connectivity request g possibly with $r_g = 0$), l_i is the length of interval i, and $p_{r,i}$ is the constant rate of $p_r(t)$ on interval i. Then the problem BP(n) can be rewritten as a linear program:

$$BPLP(n)\text{: minimize: } \sum_{g \in G_n^{final} \cup G_n^{new}} r_g$$

subject to:

$$\forall r \in R_n, \quad \forall i \in I, \quad 0 \leq p_{r,i} \leq r_r^{max},$$

$$\forall r \in R_n, \quad \forall i \in I, \quad (1 - \delta_{r,i}) p_{r,i} = 0,$$

$$\forall r \in R_n, \quad \sum_{i \in I} \delta_{r,i} p_{r,i} l_i = v_r,$$

$$\forall i \in I, \quad \forall g \in G_n^{final} \cup G_n^{new}, \quad \gamma_{g,i} \sum_{r \in R_n} p_{r,i} \leq r_g,$$

$$\forall j, \quad n \leq j \leq M, \quad r_{g^{n-1,j}} \leq r_{g^{n,j}},$$

where the variables are $\{p_{r,i} | r \in R_n^{new} \cup R_n^{sched.}, \ i \in I\}$ and $\{r_g | g \in G_n^{new}\}$. $\{p_{r,i} | r \in R_n^{granted}, \ i \in I\}$ is not part of the variable as granted request profiles cannot be changed anymore.

It can be proved that provided requests in R_n^{new} are in-advance requests, meaning they have their start time t_r^s after nD, and BPLP(n) has a solution.

The whole procedure for provisioning and scheduling transfers is depicted in Algorithm 1.

V. Performance Evaluation

In the previous section we unified malleable transfer requests and rigid requests in a common format. We formulated the dynamic bandwidth provisioning problem BPLP. Then we proposed and proved an optimal solution for any given set of malleable and fixed transfer requests. The goal of the performance evalu-

ation is to explore how the malleability of requests impact the resource provisioning and utilization. In order to evaluate the impact of requests' characteristics, we performed simulations with different workloads. The proposed provisioning algorithm has been implemented in jBDTS [11], which is used for the simulations.

Rates proposed by the NO are generally discrete. Therefore, in the following simulations Algorithm 1 is compared with a modified version that uses BPLP as a linear relaxation of the mixed integer linear problem with a discrete value for r_g. In this modified version, rounding of r_g to the upper discrete value is performed just before sending connectivity requests at line 11 of Algorithm 1. We used discrete steps of 100 Mbps.

The first experiment used a set of default parameters parameterized by the slot duration D:

- slot duration: D;
- provisioning advance: $A=D/2$;
- granting advance: $a=D/20$;
- transfer request interarrival: $D/12$;
- number of requests: 2000;
- requests' attributes:
  - $t_r^d - t_r^s = d = D/2$,
  - $t_r^s - t_r^a = 2D$,
  - $r_r^{min} = R = 53$ Mbps ($v_r = Rd$),
  - $P_r = 2$.

Algorithm 1 Schedule and provision at $t=nD-A$

Input: R_n^{new}, $R_n^{sched.}$, $R_n^{granted}$, G_n^{final}, $G_n^{prev.}$

Output: $R_{n+1}^{sched.}$, $R_{n+1}^{granted}$, R_{n+1}^{final}, $R_{n+1}^{prev.}$, $\{t \mapsto p_r(t) | r \in R_{n=1}^{sched.} \cup R_{n+1}^{granted}\}$

//Initialize set of req. for next slot

1: $R_{n+1}^{sched.} \leftarrow \emptyset$

2: $R_{n+1}^{granted} \leftarrow R_n^{granted}$

//Determine profiles for transfer req. and connectivity req.

3: Solve BPLP (n)

//Update transfer req.'s states

4: **for all** $r \in R_n^{new} \cup R_n^{sched.}$ **do**

5: **if** $t_r^s - a \leq (n+1)D - A$ then

6: $R_{n+1}^{granted} \leftarrow R_{n+1}^{granted} \cup \{r\}$

7: **else**

8: $R_{n+1}^{sched.} \leftarrow R_{n+1}^{sched.} \cup \{r\}$

9: **end if**

10: **end for**

//Send connectivity req. to NO

11: Issue connectivity requests $g \in \{G_n^{new}\}$ to NO.

//Update connectivity req. sets for next slot.

12: $G_{n+1}^{final} \leftarrow G_n^{final} \cup \{g^{n,n}\}$

13: $G_{n+1}^{prev.} \leftarrow G_n^{new} \setminus \{g^{n,n}\}$

14: **return** $R_{n+1}^{sched.}$, $R_{n+1}^{granted}$, G_{n+1}^{final}, $G_{n+1}^{prev.}$, $\{t \mapsto p_r(t) | r \in R_{n+1}^{sched.} \cup R_{n+1}^{granted}\}$

A. Impact of Patience

Given that most of the time the duration and the volume of a request will be fixed by the application semantic, we want to explore the impact of the patience factor, which represents the request's malleability. In our model, patience is higher or equal to 1; otherwise requests would not be accepted. If patience equals 1, requests cannot be reshaped and have to be scheduled as one single rectangle at r_r^{max} during $[t_r^s, t_r^d]$. The impact of this factor can be observed in Fig. 8, where extra reserved bandwidth decreases quickly as P_r increases. When P_r is greater than 1.5, performance is constant. This means that the patience factor has a strong impact.

B. Impact of Proportion of Flexible Requests

In this set of experiments, we used two classes of requests: one with patience equal to 1 (rigid requests) and one with varying patience (malleable requests).

Requests used here have the following parameters: There are 30 clients each acting as an ON–OFF source with an exponential OFF time of mean 2 hours. ON times are given by the request windows $[t_r^s, t_r^e]$. Each client submits 60 requests. The volumes of each request are derived from a Pareto distribution with a minimum volume equal to 100 GB and a shape parameter equal to 1.5, leading to an average volume of 300 GB per request. Each source has a maximum sending rate of 100 Mbps or 1 Gbps randomly chosen with equal probability. Similarly each transfer can be flexible and have a varying patience ($P_r \in \{1, 1.1, 1.2, 1.5, 2, 3\}$) or can have no flexibility ($P_r = 1$); the probability of being a flexible request is p. The slot duration is equal to 24 hours. As an example, with a sending rate of 100 Mbps, a 300 GB transfer has a minimum completion time (with patience equal to 1) of 6.67 hours. In the overall experiment the mean actual duration of transfer was observed to be 6.33 hours. In this experiment, as the maximum rate

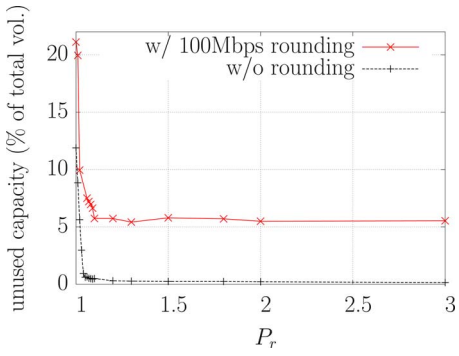


Fig. 8. (Color online) Provisioned but unused bandwidth as a function of patience without and with rounding to the upper 100 Mbps.

is fixed, increasing patience decreases minimum rate r_r^{min} and thus increases the transfer's potential duration.

We can observe on Fig. 9 that provisioning of this considered set of requests lead to an important waste of bandwidth when there is no flexibility ($P_r=1$ or $p=0$). But the gain is linear with the proportion of flexible requests because r_r^{min} of flexible requests is P_r times smaller when the request is flexible while the number of flexible requests is p times the total number of requests.

The second observation is that increasing the patience is an efficient way to decrease the overprovisioning. Increasing the patience of 10% from $P_r=1$ to $P_r=1.1$, which actually also increases transfer duration by 10%, increases the utilization ratio by 20% when all requests are flexible. However, this gain is not linear with patience because going from $P_r=1.1$ to $P_r=3$ only improves the provisioning by another 20%.

VI. Related Works

One of the directions recently investigated in large-scale distributed computing and data processing systems is the capability of dynamically establishing dedicated connection-oriented circuit switching or a lambda path [2–4]. The current state of the art is that these circuit services are configured manually either by fax, phone, or e-mail or in selected cases with first-generation web services. As an example of these new services, UCLP (user-controlled light paths) [3] allows users and applications to partition and configure the resources of an optical network. The CHEETAH (circuit-switched high-speed end-to-end transport architecture) project [12] is also towards an end-to-end connection-oriented circuit in a gigabit Ethernet over synchronous optical netork (SONET) transport network with GMPLS control functions at the network elements. The goal of the DRAGON (dynamic resource allocation via GMPLS optical networks) project [13] is to develop functions such as a network-aware resource broker (NARB) and an application-specific topology builder (ASTB) required by the network infrastructure for performing immediate and in-advance reservations of network resources for connections in a heterogeneous and multidomain transport network. Reference [14] suggests making network reconfigurations available to application users by making visible all the resources and allowing them to send signaling messages in carrier networks. This approach can be applied to specific distributed applications like e-science deployed over the National Research and Educational Networks (NREN).

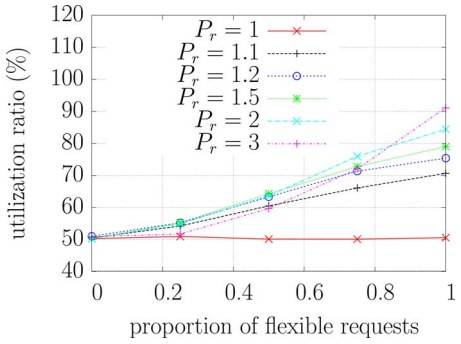


Fig. 9. (Color online) Provisioned but unused bandwidth as a function of p for different patience ($P_r \in \{1, 1.1, 1.2, 1.5, 2, 3\}$) with rounding to the upper 100 Mbps.

To develop and adapt these facilities into an industrial context, CARRIOCAS proposes new management and control functions to evolve existing telecom network infrastructures to deliver commercial services for company customers. Commercial services have to be built on an abstracted view of the infrastructures and resource signaling has to go through policy and business procedures before triggering network reconfiguration or resource reservations. In the CARRIOCAS approach, company customers can access a well-defined abstracted view of the services exposed by the SRV. This approach allows network operators to hide the real topology and availability of their networks. G-Lambda [2] also considers the case of lambda paths as a commercial service (single provider or multiprovider cases) but concentrates on the user-to-network interface specifications with a grid network services–web-service interface (GNS–WSI). Most of these current hybrid network implementation connection services are permanently or semipermanently configured and managed by a proprietary network resource provisioning system (NRPS). Comparable service-oriented approaches are also explored in [5,15].

Several works expand the dynamic provisioning approach to IT resources. In [16] two different integrations of the service plane and network control plane are proposed. The first one is grid-enabled GMPLS (G^2MPLS), which implements GMPLS control functions and extends them to include nonnetwork resources as new switching capabilities. The second is a session initiation protocol (SIP) over an optical burst switching network (SIP-based OBS network), where the SIP is used as a service-level signaling protocol and SIP proxies are used to access the OBS transport plane. They both suppose that network infrastructure topology, its capability, and its states are accessible to the end users. CARRIOCAS implements the virtual infrastructure concept by decoupling the management of the physical infrastructure (network element inventory, equipment and network maintenance and administration, performance monitoring, etc.) from the management of the services that can be delivered to external service providers.

A dynamic bandwidth scheduling scheme that exploits the quasi-flexible nature of connectivity reservations and considers, as we do, two broad classes of generated network traffic, streaming and elastic, with the same type of QoS requirements that have been proposed in [17]. This problem was also studied by Burchard in [18], where the concepts of malleable reservations to address bandwidth fragmentation was proposed. Our approach is similar to both proposals. However, unlike in these related works, this paper formulates the problem and proposes a linear programming approach that offers an optimal solution.

In previous works [19,20], multistep allocation of time-constrained transfer requests were studied under fixed capacity constraints. This work extends that and demonstrates how the approach applies to any in-advance connectivity reservation service.

Reference [21] proposes an architecture relying on OBS that maps jobs to optical bursts. This solution is limited to jobs that can be put in bursts and does not require interactivity or an established circuit between fixed entities, such as video over IP. Our proposed approach can serve such requests through rate-based requests.

In the context of self-sizing networks adaptively dimensioned as traffic changes, several dynamic bandwidth allocation strategies have been proposed [22–24]. These approaches are all based on predictors. This paper considers that a large fraction of the traffic is known in advance and is malleable. The bandwidth provisioning service is exposed to users. Our solution benefits from the traffic malleability and the in-advance knowledge to adapt the resource reservation to its needs (or objective) and to provision the network equipment accordingly.

VII. Conclusion

This paper has presented virtual infrastructures to help emerging IT service providers facing service demand variations and the need for services for real-time access to right-sized capacities, which are currently not supported by telecom network infrastructures. This paper advocates for the design, development, and deployment of new infrastructure management functions to discover, select, reserve, coallocate, and reconfigure resources and schedule, monitor, and control in real time their usage. These functions were studied and implemented in the SRV module of the CARRIOCAS project. We proposed a flexible approach for dynamic bandwidth provisioning suitable for telecom network operations. The interface enables us to specify malleable transfer requests and rigid requests in a common format. The corresponding dynamic bandwidth provisioning problem has been formulated. It provides a solution for in-advance malleable and rigid request scheduling and the provisioning of the underlying network infrastructure in case this infrastructure has infinite resources. The proposed objective function is to minimize the total amount of network resources reserved for the application workflows as requested by service providers. It is an important objective for service providers who have to pay for the reserved resources. Performance evaluation demonstrates how time slots and service demand rate granularity as well as malleability affect resource provisioning and its utilization.

We are currently doing experiments on the automatic provisioning of connection services by interfacing the SRV module with the network equipment deployed in the CARRIOCAS pilot network. Thanks to its evolutive architecture, the dynamic reconfiguration functions of wavelength switched connections will be explored through tunable and reconfigurable add–drop multiplexer (T&R-OADM) nodes and GMPLS-based control functions. The node controllers will be upgraded to establish automatically the wavelength-switched optical connections on demand by communicating connectivity service end points to client carrier Ethernet nodes embedding dynamic protocol message exchanges in the form of RSVP-TE messages through a UNI interface to the photonic cross-connect servers.

Acknowledgments

This work has been funded by the French Ministry of Education and Research, INRIA, and CNRS, via ACI GRID's Grid'5000 project, Aladdin ADT, and the CARRIOCAS project (pôle Systém@tic IdF).

References

[1] CARRIOCAS website, 2008, http://www.carriocas.org.

[2] S. R. Thorpe, L. Battestilli, G. Karmous-Edwards, A. Hutanu, J. MacLaren, J. Mambretti, J. H. Moore, K. S. Sundar, Y. Xin, A. Takefusa, M. Hayashi, A. Hirano, S. Okamoto, T. Kudoh, T. Miyamoto, Y. Tsukishima, T. Otani, H. Nakada, H. Tanaka, A. Taniguchi, Y. Sameshima, and M. Jinno, "G-lambda and enlightened: wrapped in middleware co-allocating compute and network resources across Japan and the US," in *GridNets '07: Proc. 1st Int. Conf. Networks for Grid Applications*, Brussels, Belgium, 2007, pp. 1–8.

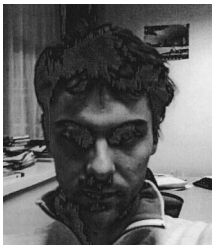
[3] UCLP: User Controlled Lightpaths website, 2008, http://www.uclp.ca.

[4] G. Zervas, E. Escalona, R. Nejabati, D. Simeonidou, G. Carrozzo, N. Ciulli, B. Belter, A. Binczewski, M. Stroiski, A. Tzanakaki, and G. Markidis, "Phosphorus grid-enabled GMPLS control plane (G2MPLS): architectures, services and interfaces," *IEEE Commun. Mag.*, vol. 46, no. 6, pp. 128–137, June 2008.

[5] "Multi-Technology Operations Systems Interface (MTOSI) 2.0, TMF Forum," 2008, http://www.tmforum.org/browse.aspx?catid=2319.

[6] G. Koslovski, P. Vicat-Blanc Primet, and A. Charaoan, "VXDL: virtual resources and interconnection networks description language," in *GridNets 2008*, Beijing, China, 2008, pp. 138–154.
[7] "ITU-T recommendations G.807/Y.1302: Requirements for automatic switched transport networks (ASTN)," 2001.
[8] NSI-WG: OGF Network Service Interface WG (NSI-WG) website, 2008, http://www.ggf.org/gf/group_info/view.php?group=nsi-wg.
[9] B. B. Chen and P. Vicat-Blanc Primet, "Supporting bulk data transfers of high-end applications with guaranteed completion time," in *Int. Conf. Communication*, Glasgow, Scotland, pp. 575–580.
[10] N. Dukkipati and N. McKeown, "Why flow-completion time is the right metric for congestion control," *Comput. Commun. Rev.*, vol. 36, pp. 59–62, 2006.
[11] INRIA RESO team, "jBDTS: Bulk Data Transfer Scheduling service website," 2008, http://www.ens-lyon.fr/LIP/RESO/Software.html.
[12] X. Zheng, M. Veeraraghavan, N. Rao, Q. Wu, and M. Zhu, "CHEETAH: circuit-switched high-speed end-to-end transport architecture testbed," *IEEE Commun. Mag.*, vol. 43, no. 8, pp. s11–s17, Aug. 2005.
[13] T. Lehman, J. Sobiesky, and B. Jabbari, "DRAGON: a framework for service provisioning in heterogeneous grid networks," *IEEE Commun. Mag.*, vol. 44, no. 3, pp. 84–90, March 2006.
[14] A. Jukan and G. Karmous-Edwards, "Optical control plane for the grid community," *IEEE Commun. Surveys Tutorials*, vol. 9, pp. 30–44, 2007.
[15] F. Verdi, M. Magalhães, E. Cardozo, E. Madeira, and A. Welin, "A service oriented architecture-based approach for interdomain optical network services," *J. Network Syst. Manage.*, vol. 15, pp. 141–170, 2007.
[16] N. Ciulli, G. Carrozzo, G. Giorgi, G. Zervas, E. Escalona, Y. Qin, R. Nejabati, D. Simeonidou, F. Callegati, A. Campi, W. Cerroni, B. Belter, A. Binczewski, M. Stroinski, A. Tzanakaki, and G. Markidis, "Architectural approaches for the integration of the service plane and control plane in optical networks," *Opt. Switching Networking*, vol. 5, pp. 94–106, 2008.
[17] S. Naiksatam and S. Figueira, "Elastic reservations for efficient bandwidth utilization in lambdagrids," *FGCS, Future Gener. Comput. Syst.*, vol. 23, pp. 1–22, 2007.
[18] L.-O. Burchard, "Networks with advance reservations: applications, architecture, and performance," *J. Netw. Syst. Manage.*, vol. 13, pp. 429–449, 2005.
[19] B. Chen and P. Vicat-Blanc Primet, "Scheduling deadline-constrained bulk data transfers to minimize network congestion," in *Proc. 7th IEEE Int. Symp. Cluster Computing and the Grid*, Rio de Janeiro, Brazil, 2007, pp. 410–417.
[20] S. Soudan, B. Chen, and P. Vicat-Blanc Primet, "Flow scheduling and endpoint rate control in GridNetworks," *FGCS, Future Gener. Comput. Syst.*, to be published.
[21] M. De Leenheer, P. Thysebaert, B. Volckaert, F. De Turck, B. Dhoedt, P. Demeester, D. Simeonidou, R. Nejabati, G. Zervas, D. Klonidis, and M. O'Mahony, "A view on enabling-consumer oriented grids through optical burst switching," *IEEE Commun. Mag.*, vol. 44, no. 3, pp. 124–131, March 2006.
[22] Z. Sahinoglu and S. Tekinay, "A novel adaptive bandwidth allocation: wavelet-decomposed signal energy approach," in *Global Telecommunications Conf.*, 2001, pp. 2253–2257.
[23] Y. Liang and M. Han, "Dynamic bandwidth allocation based on online traffic prediction for real-time mpeg-4 video streams," *EURASIP J. Appl. Signal Process.*, vol. 2007, pp. 51–51, 2007.
[24] B. Krithikaivasan, Y. Zeng, K. Deka, and D. Medhi, "ARCH-based traffic forecasting and dynamic bandwidth provisioning for periodically measured nonstationary traffic," *IEEE/ACM Trans. Netw.*, vol. 15, pp. 683–696, 2007.

Pascale Vicat-Blanc Primet has been an IEEE member since 2004. She obtained the Ms.C. (1984) and Engineer diploma (1984) in computer science from INSA de Lyon (France) and the Ph.D. (1988) in computer science and HDR (Habilitation à Diriger les Recherches) (2002) from the University of Lyon (France). Her interests include high-speed and high-performance networks, Internet protocols, protocol architecture, quality of service, network and traffic measurement, network programmability and virtualization, grid computing, and grid networking and communications. From 1989 to 2004, she was an Associate Professor at the École Centrale de Lyon. Since 2005 she has been a Research Director at the National Institute of Research in Computer Science (INRIA). Since 2002, she has led the INRIA RESO team (22 researchers and engineers). She is a member of the Board of Directors of the École Normale Supérieure de Lyon and co-director of its Computer Science Laboratory (LIP). Dr. Vicat-Blanc Primet is a member of the scientific Networks and Telecoms expert committee of CNRS (National Research and Science Center—France). She is a member of the Grid5000's—French Computer Science Grid initiative—steering committee and participates in evaluation committees of the French National Research Agency (ANR). She has co-chaired the OGF Data Transport Research Group. She is currently the leader of the ANR HIPCAL project. She represents and leads the activity of INRIA in variety of international, European, and national grid projects including IST DataGRID, IST DataTAG, RNTL eToile, ACI GdX, IST EC-GIN, and ANR IGTMD CARRIOCAS. She is co-chairing the PFLDnet and ACM Gridnets conference steering committees and participates in numerous international program committees. She has published more than 100 papers in international journals and conferences in networking and grid computing.



Sebastien Soudan graduated from the École Centrale de Lyon, France, in 2005, received the M.S. degree in computer science in 2006, and the Ph.D. in 2009 from ENS Lyon. His main areas of research are bandwidth sharing, network scheduling, network economics, and game theory.

Dominique Verchere (M'98) holds a Ph.D. degree in computer science on performance evaluation and a master's degree in computer science from the University of Paris VI, France, since 1997 and 1994, respectively. Since 1998 he has been with Alcatel-Lucent Bell Labs France, first designing the OmniPCX Enterprise Call Server, then developing data path processing (scheduler and buffer management) of terabit-switching capacity systems within the Terabit IP Optical Router project in which the concepts of dynamic optical burst switching were introduced. Then his involvement was dedicated to GMPLS-based resilience functions for IP routers and optical cross-connects and carrier-grade Ethernet systems according to the GMPLS specifications based on RSVP-TE and OSPF-TE. Dr. Verchere actively contributed to several European projects including EuroNGI, TBones, and NOBEL1 and 2 and external projects such as VIOLA. He is currently working on service management functions oriented optical networks (based on TMF MTOSI 2.0), leading the subproject on Network Architectures and Protocols for computing service deliveries enabled by ultra-high-capacity optical networks (CARRIOCAS). He has published more than 40 papers and 30 patents, has been a member of the Alcatel-Lucent Technical Academy (ALTA) since 2003, and has co-chaired the ALTA French Chapter since 2006.