TeraPaths: End-to-End Network Resource Scheduling in High-Impact Network Domains

Dimitrios Katramatos[1], Xin Liu[1], Kunal Shroff [2], Dantong Yu[1], Shawn McKee[3], Thomas Robertazzi[4]

[1] Computational Science Center, Brookhaven National Laboratory, Upton, NY 11973, USA
[2] National Synchrotron Light Source II, Brookhaven National Laboratory, Upton, NY 11973, USA
{dkat, xinliu, shroffk, dtyu}@bnl.gov
[3] Department of Physics, University of Michigan, Ann Arbor, MI 48109, USA
smckee@umich.edu
[4] Department of Electircal and Computer Engineering, Stony Brook University, Stony Brook , NY 11794, USA
tom@ece.sunysb.edu

Abstract— **The TeraPaths project at Brookhaven National Laboratory is pioneering a framework that enables the scheduling of network resources in the context of data-intensive scientific computing. Modern wide area networks, such as ESnet and Internet2, have recently started providing network resource reservation capabilities in the form of virtual circuits. The TeraPaths framework utilizes these circuits and extends them into end-site local area networks, establishing end-to-end virtual paths between end-site hosts. These paths are dedicated to specific users and/or applications and provide guaranteed resources, minimizing or eliminating the adverse effects of network congestion. In this article, we present an overview of TeraPaths and examine issues raised by the end-to-end resource reservation-based networking paradigm as well as implications and benefits for end users and applications. We also discuss scalability issues and optimization techniques for wide area network circuit reservations.**

Keywords—End-to-end QoS networking, hybrid networks, network virtualization, virtual circuit reservation optimization.

I. INTRODUCTION

This article is an extended and revised version of the INTERNET 2009 conference paper entitled: “Establishment and Management of Virtual End-to-End QoS Paths Through Modern Hybrid WANs with TeraPaths” [1].

Modern data intensive scientific applications, including high energy and nuclear physics, astrophysics, climate modeling, nanoscale materials science, and genomics, will soon be capable of generating data on the order of exabytes per year [2]. This data must be transferred, visualized, and analyzed by geographically distributed teams of scientists, imposing unprecedented demands on computing and especially networking resources. While such applications can capitalize on modern high-performance networking capabilities, they can also be critically sensitive to the adverse effects of unpredictably occurring network congestion. Because network capacity is finite, competition among data flows may cause applications to suffer severe performance degradation and eventual disruption. When data delivery must conform to specific deadlines or application components need to interact in real time, the standard best-effort networking model may not always be sufficient. To work effectively, these applications may require resource availability guarantees. In the case of network, the requirement primarily translates to bandwidth guarantees, however, other Quality of Service (QoS) parameters may also be included, i.e., delay, jitter, etc. The Department of Energy (DOE) Office of Science identifies QoS as one the five top ranked issues essential to the success of distributed science [3].

The next section discusses the motivation behind TeraPaths, while section 3 describes two key projects that constitute the framework for the advance resource reservation model. Section 4 focuses on the differences between the two kinds of dedicated network paths through WAN domains supported by this framework, while Section 5 presents techniques necessary for the effective utilization of these dedicated WAN paths. Section 6 examines fault tolerance issues and Section 7 discusses related work. Finally, Section 8 presents our conclusions and future work directions.

II. HIGH-IMPACT NETWORK DOMAINS

As noted in the title, TeraPaths targets “High-Impact” network domains (sets of related users and systems connected by networks) and so we provide some background on what we mean by this. Typical network use for a given system characteristically utilizes a few-to-many, small bandwidth, short duration network flows: email, web browsing, and the occasional file-transfer are common examples. However, there is a much smaller set of systems which regularly transfer large amounts of data over the network. Typically, this may involve bandwidth-intensive applications or large files (data, movies, games, HD video-conferencing, etc.) and may use a significant fraction of the available bandwidth along a network path. More importantly, some of these large flows may have additional requirements regarding packet loss, delay, and jitter, as well as overall deadline scheduling needs that are critical to the specific user or application. We characterize high-impact domains as those sets of users and systems who need to transfer large amounts of data through the network and who may require additional control over network related characteristics of their critical flows (such as “real-time or interactive flows”, e.g., video-conferencing, real-time instrument control, conference audio/visual streaming, etc.).

The high-impact domains TeraPaths envisions supporting are in the e-Science area where significant amounts of data need to be shared across wide-area networks (WANs) and additional important considerations regarding timeliness of some data transfers and their corresponding flow characteristics are important to the success of the applications involved [4]. In particular, grid-computing infrastructures in science are already broadly deployed and could be considered synonymous with high-impact domains. Virtual organizations built upon grids would significantly benefit from end-to-end predictability of network paths interconnecting their shared resources [5]. While small in number (by relative count of users or end-sites), these domains can have a disproportionally disruptive effect on the network and thus are "high-impact".

We would further make the case that not all large-scale flows are of equal importance or criticality. On today's Research and Education networks one may see large scale flows corresponding to high-energy physics data transfers, eVLBI astronomy, bio-informatics and life sciences as well as peer-to-peer traffic sharing movies, applications, music, and other multimedia content. Even within a networked collaboration of users, some large scale transfers may have significantly different importance but are currently treated equivalently by the best effort network. Part of the motivation behind TeraPaths is to give researchers the tools they need to most effectively utilize the resources they have access to.

III. BACKGROUND

Several available networking technologies, such as the Differentiated Services (DiffServ) [6], Integrated Services (IntServ) [7], Multi-Protocol Label Switching (MPLS) [8], and Generalized MPLS (GMPLS) [9] architectures, have the capability to address the issue of providing resource guarantees. In practice, however, the scope of network connections utilized by distributed applications spans multiple autonomous domains. These domains typically have different levels of heterogeneity in administrative policies and control plane and data plane technologies, making it difficult or impossible to provide network QoS guarantees using a single architecture across all domains. For example, Differentiated Services Code Point (DSCP) packet markings, used in the DiffServ architecture, are by default reset at ingress points of network domains. As such, the DiffServ architecture is ineffective across domains without prior inter-domain Service Level Agreements (SLAs) in effect and proper configuration of involved network devices.

Recent networking research and development efforts [10] – [13] adopt a hybrid solution to the problem, with individual network segments utilizing different underlying technologies. From the end user perspective, however, these technologies are seamlessly tied together to ensure end-to-end resource allocation guarantees. This hybrid solution creates a new networking model that transparently co-exists but fundamentally differs from the standard best-effort model. Under the new model, it is possible to allocate network resources through advance reservations and dedicate these resources to specific data flows. Each such flow (or flow group) is steered into its "own" virtual network path, which ensures that the flow will receive a pre-determined level of QoS in terms of bandwidth and/or other parameters. Virtual paths can comprise several physical network segments and span multiple administrative domains. These domains need to coordinate to establish the virtual path. Coordination takes place by means of interoperating web services. Each domain exposes a set of web services that enable the reservation of resources within a domain's network. Authorized users of these services, which can be another domain's services, can reserve network resources within the domain and associate them with specific data flows. When reservations activate across all domains between a flow's source and destination, a dedicated end-to-end virtual path spanning these domains is assembled. This path offers to the flow of interest a predetermined level of end-to-end QoS. The coordination of multiple network domains through web services is essentially a loosely coupled Service Oriented Architecture (SOA) for the network control plane, a network "service plane" [14].

End-to-end virtual paths can be viewed as consisting of three main segments: two end segments, one within each end site Local Area Network (LAN), and a middle segment spanning one or more Wide Area Network (WAN) domains. In this article, we consider the establishment of end-to-end virtual paths from the perspective of end sites. User applications run on end site systems, communicate with the rest of the world through end site LANs, and are subject to end site administrative policies. In the standard networking model, traffic through the WAN is subject to pre-existing SLAs between adjacent network domains. In the new advance resource reservation model, such SLAs are essentially dynamic, allowing end sites to utilize and – indirectly – manage WAN capabilities in a way that maximizes the benefit to the end user.

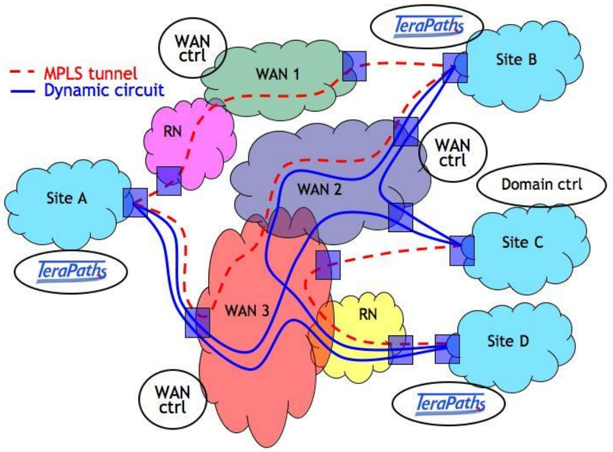


Figure 1. The framework for establishing end-to-end paths; TeraPaths-controlled sites are interconnected with WAN MPLS tunnels and/or dynamic circuits; some paths pass through regional networks that have long-term static configurations to accommodate QoS.

The framework for establishing end-to-end QoS-aware network paths encompasses web service-based systems that properly configure end site LAN and WAN domains (see Figure 1). The capability for advance resource reservation is currently available between sites interconnected through the ESnet [15] and Internet2 [16] networks. In this section we give background information on the two projects that constitute this framework, the TeraPaths project and the OSCARS project.

A. The TeraPaths Project

The DOE-funded TeraPaths project [10] at Brookhaven National Laboratory (BNL) combines DiffServ-based LAN QoS with WAN MPLS tunnels and dynamic circuits to establish end-to-end (host-to-host) virtual paths with QoS guarantees. These virtual paths prioritize, protect, and regulate network flows in accordance with site agreements and user requests, and prevent the disruptive effects that conventional network flows can bring to one another.

Providing an end-to-end virtual network path with QoS guarantees (e.g., guaranteed bandwidth) to a specific data flow requires the timely configuration of all network devices along the route between a given source and a given destination. In the general case, such a route passes through multiple administrative domains and there is no single control center able to perform the configuration of all devices involved. The TeraPaths system has a fully distributed, layered architecture (see Figure 2) and interacts with the network with the perspective of end-sites of communities. The local network of each participating end-site is under the control of an End-Site Domain Controller module (ESDC). The site's network devices are under the control of one or more Network Device Controller modules (NDCs). NDCs play the role of a "virtual network engineer" in the sense that they securely expose a very specific set of device configuration commands to the ESDC module. The software is organized so that NDCs can be, if so required by tight security regulations, completely independently installed, configured, and maintained.

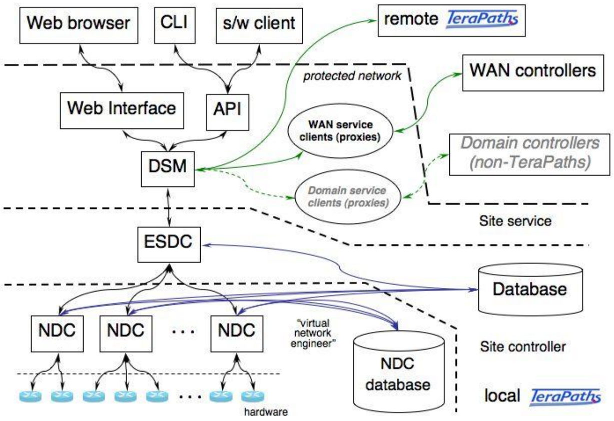


Figure 2. The software architecture of TeraPaths. Services of remote network domains are invoked through "proxy" server modules.

An NDC encapsulates specific functionality of a network device and abstracts this functionality through a uniform interface while hiding the complexity of the actual configuration of heterogeneous hardware from higher software layers. A site's ESDC and NDC(s) are complemented by a Distributed Services Module (DSM), which is the core of the TeraPaths service. The DSM has the role of coordinating all network domains along the route between two end hosts (each host belonging to a different end-site) to timely enable the necessary segments and establish an end-to-end path. The DSM interfaces with all ESDCs (local and remote) to configure the path, starting within the end-site LANs (direct control) and proceeding to arrange the necessary path segments through WAN domains (indirect control). To interface with non-TeraPaths domain controllers, primarily for WAN domains but also for end-sites that are using other controlling software (e.g., Lambda Station [11]), the DSM uses auxiliary modules that encapsulate the functionality of the targeted domain controller by invoking the required API but exposing a standardized abstract interface. As such, these auxiliary modules appear to a DSM as a set of "proxy" WAN or end-site services with a uniform interface. It should be noted that the responsibility of selecting and engineering the path within a WAN domain belongs to the controlling system of that domain. TeraPaths can only indirectly affect such a path by providing preferences to the WAN controlling system, if that system offers such a capability.

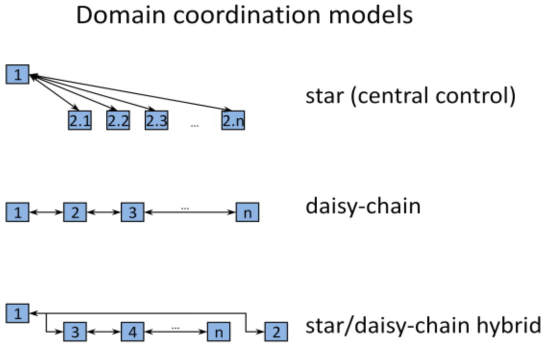


Figure 3. Coordination models. Each square represents a site's controller.

Currently, TeraPaths follows a hybrid star/daisy chain coordination model where the initiating end-site first coordinates with the target site and then indirectly sets up a WAN path by contacting its primary WAN provider and relying on that provider's domain to coordinate, if necessary, with other WAN domains along the desired route (see Figure 3). The hybrid coordination model was adopted as the most feasible since end-site and WAN systems need only to interface/coordinate. Thus, no unified communication protocol is required, as in the case of the daisy chain model, and there is no centralization of control, as in the case of the star model. The hybrid model essentially splits the network in two large segments: the end-sites and the WAN domains, with each segment coordinating with the other to setup a path.

The result of the domain coordination process is the establishment of dynamic Service Level Agreements (SLAs) between all network domains along an end-to-end path. TeraPaths is responsible for the two end-sites and OSCARS for one or more peering WAN domains. The Message Sequence Chart (MSC) in Figure 4 shows the messaging sequence taking place in the current system implementation: initiating end-site A negotiates with the other end-site B to reach a consensus based on the resource availability of both sites. Then, site A send the negotiated request to the WAN domain manager, in this case, OSCARS, which responds with a success or failure message.

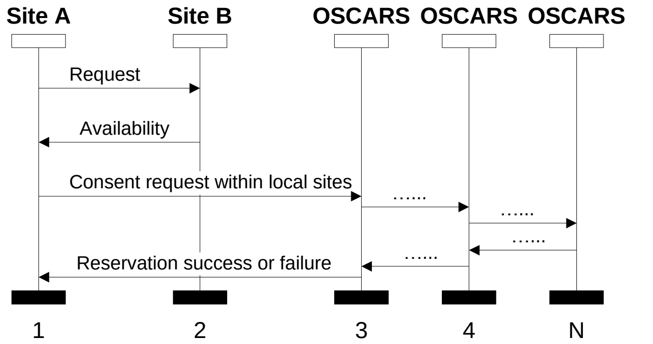

Site A
Site B
OSCARS
OSCARS
OSCARS
Request
Availability
Consent request within local sites
……
……
……
……
Reservation success or failure
1
2
3
4
N

Figure 4. A Message Sequence Chart for the coordination of network domains controlled by TeraPaths and OSCARS.

B. OSCARS

The DOE-funded On-demand Secure Circuit Advance Reservation System (OSCARS) [13] is a project initiated by ESnet. Initially, OSCARS could dynamically provision secure layer-3 (L3) circuits with guaranteed bandwidth in the form of MPLS tunnels, only within the ESnet domain.

Through collaboration between ESnet and Internet2, OSCARS evolved into a more general Inter-Domain Controller (IDC), a WAN domain controller, enabling adjacent WAN domains to interoperate and establish secure circuits spanning multiple domains via the use of a special protocol specifically developed for domain interoperation. While still capable of providing MPLS tunnels within ESnet, OSCARS can additionally provide guaranteed bandwidth layer-2 (L2) circuits within and between ESnet’s Science Data Network (SDN) and Internet2’s Dynamic Circuit Network (DCN). SDN and DCN are interconnected at New York and Chicago and bring together DOE laboratories and Universities across the United States.

Access to OSCARS circuit reservations is offered via a web interface. Additionally, the system’s functionality is exposed through a web services API for automatic invocation from programs. The API includes basic primitives for establishing and managing circuit reservations (create, cancel, query, list) and L2-specific primitives to signal and teardown dynamic circuits. TeraPaths utilizes a client module to automatically submit circuit reservation requests and further manage these reservations on behalf of end site users/applications. The selection of the actual WAN path is currently left at the discretion of OSCARS for simplicity and maximum flexibility in satisfying a request. The path provisioned by an OSCARS reservation is expected to satisfy the bandwidth requirements, however, the end-sites do not participate in routing decisions. The latest versions of OSCARS include support for obtaining topology information and specifying preferred path in reservation request. Selecting inter-domain paths is desirable from the end-site perspective for reserving, e.g., lower latency routes. However, it adds another dimension of complexity to reserving a path, as end-sites need to pull topology information and decide on which route they prefer based on certain criteria, while the chances of successfully reserving a path are probably decreasing as OSCARS is presented with a less flexible request. Nevertheless, we plan to explore such capabilities in our future work.

C. The TeraPaths Testbed

The TeraPaths project utilizes a multiple-site testbed for research, software development, and testing. Currently, the testbed encompasses subnets at three sites, BNL, University of Michigan (UMich) and Boston University (BU) (see Figure 5). Each site runs its own instance of the TeraPaths service.

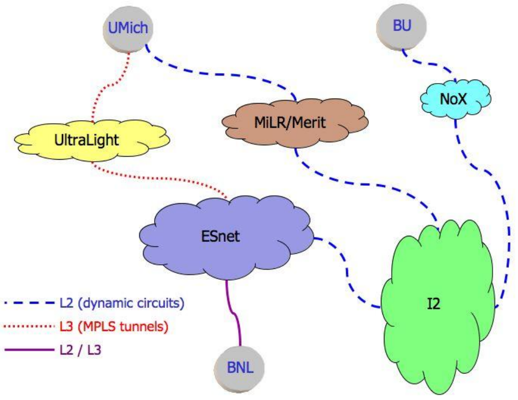


Figure 5. The TeraPaths testbed encompasses subnets at BNL, UMich, and BU. Only BNL is directly connected to ESnet.

All instances can interface with OSCARS interdomain controllers to setup MPLS tunnels through ESnet and dynamic circuits through ESnet and Internet2. Future end-sites will have similar interconnecting capabilities depending on which WAN they subscribe to (ESnet supports both L2 and L3 circuits, while Internet2 only L2). Figure 6 presents the results of traffic tests between BNL and UMich. The target host at UMich, the same for all traffic streams, has a maximum capacity of 10 Gbits/second. Priority traffic between BNL and UMich is competing against other inter-site traffic and traffic local to UMich. The desired rate of the priority traffic is 700 Mbits/second, achieved only when a TeraPaths reservation is active. The rate of the competing traffic drops by approximately 500 Mbits/second, which is gained by the priority traffic for the duration of the reservation.

TeraPaths instances can regulate and guarantee the bandwidth of multiple flows between the testbed sites. These flows may utilize individual WAN circuits or may be grouped together, based on source and destination, into the same WAN circuit (which accommodates the aggregate bandwidth). Figure 7 shows a demonstration of flow bandwidth regulation for multiple periodic data transfers as monitored by Internet2's perfSONAR system. The aggregate bandwidth passing through circuits between BNL, UMich, and BU is displayed. Two transfers take place during each period, with each transfer maintained at a guaranteed bandwidth level. The second transfer (2) starts later than the first (1) and continues after the latter finishes. Each flow is policed to its guaranteed bandwidth level preventing competition within the circuit. Use of DiffServ QoS in the end site LANs and dynamic WAN circuits ensures that presence of any other traffic does not affect the regulated flows. In the particular example, transfer (2) is being policed even after transfer (1) is over. In the general case, it is possible to alter the policing rules to allow the continuing transfer to use all the bandwidth of the circuit. The QoS guarantee provided by the TeraPaths and OSCARS systems is at the network device level, i.e., network devices are configured to recognize specific packet flows and offer them a different level of service as determined by the coordinated system reservations. The quality of the guarantee mainly depends on the implementation DiffServ, MPLS, and GMPLS technologies in the network devices along a path. During our experiments we have observed a bandwidth variance of less than 10%, depending also on the load conditions of the network. Specifically for the end sites where DiffServ is used, the highest level of guarantee is achieved when utilizing the Expedite Forward (EF) class of service, as traffic belonging to this class is typically serviced by strict priority queuing schemes.

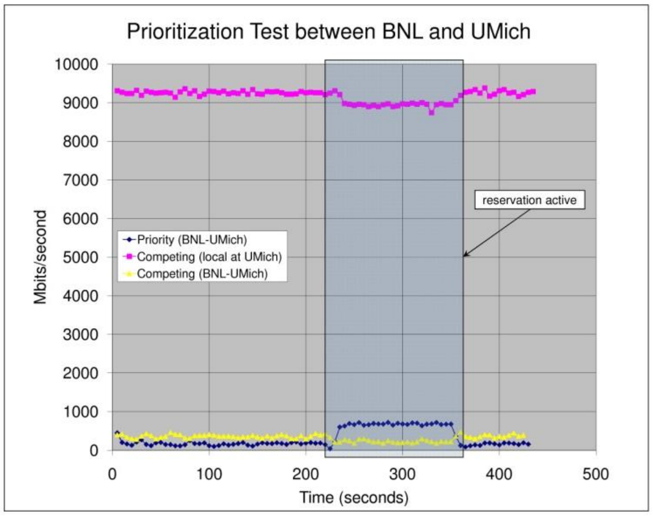


Figure 6. Traffic tests between BNL and UMich: priority inter-site traffic competing against (a) local and inter-site traffic (b) local traffic.

IV. Layer-3 vs. Layer-2

From the perspective of end sites, the requirements for utilizing a L2 or a L3 circuit are significantly different. In this section we discuss these requirements and related issues.

A. *MPLS Tunnels (L3)*

In the case the path through one or more WAN domains is established in the form of an MPLS tunnel (see Figure 8a), admission control into the tunnel is done at the ingress device of the MPLS tunnel on the WAN side. Packets that belong to an authorized flow or group of flows are recognized based on source and destination IP address and possibly additional selection criteria (e.g., port numbers). The source end site essentially hands over all packets to the WAN but only those that belong to authorized flows enter their corresponding tunnel. The MPLS tunnel maintains the packet DSCP markings so that flows emerging at the egress of the tunnel receive differential treatment within the destination end site LAN.

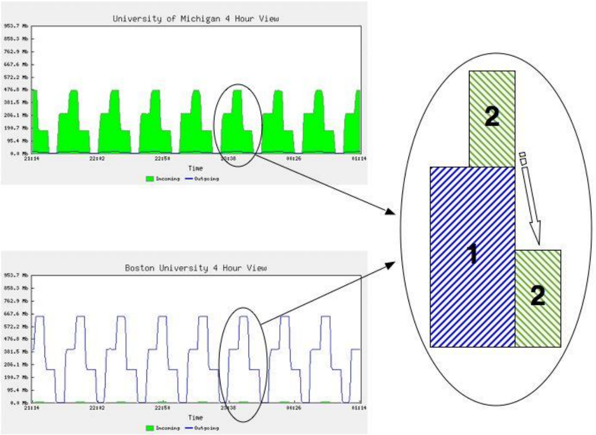


Figure 7. Demonstration of flow bandwidth regulation at SuperComputing 2007 and Joint Techs winter 2008.

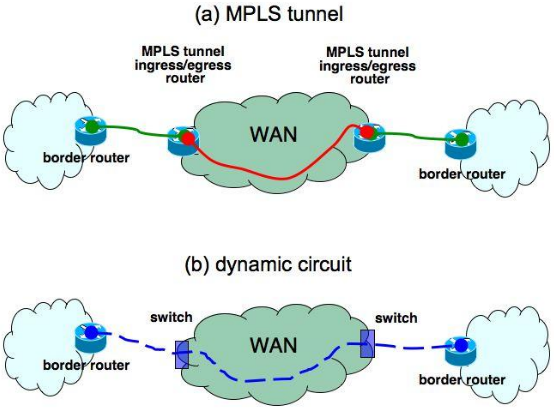


Figure 8. WAN circuits: (a) MPLS tunnels vs. (b) L2 dynamic circuits.

B. *Dynamic Circuits (L2)*

The infrastructure for the utilization of dynamic L2 circuits is quite different (see Figure 8b). In this case, the

WAN circuit established between two end sites makes those sites members of the same Virtual LAN (VLAN). The interfaces of the end site border routers participating in the connection appear as if connected directly with a patch cable, i.e., there is a single hop between them. Forwarding authorized traffic to the VLAN assigned to the circuit is the responsibility of each end site's border router. Each router uses Policy Based Routing (PBR) to selectively forward authorized flow packets (identified by source and destination IP addresses and possibly other criteria, e.g., ports) into this VLAN. For bidirectional traffic through a circuit, the border routers have to be configured in a mirrored configuration so that the destination site's border router appears as the next hop to the source site's border router and vice versa.

C. Related Issues

When an end site gains access to a WAN domain through a Regional Network (RN) that cannot be dynamically configured through a domain controller, it is necessary to statically configure the RN's devices so that (a) DSCP markings are not reset at the boundaries and (b) VLANs are extended through the RN. The same techniques need to be used within an end site LAN for network devices that are along routes used by end-to-end paths but are not under direct TeraPaths control. The static configuration is applied only to those specific device interfaces that interconnect TeraPaths-controlled devices with WAN devices. We call such statically configured network segments "pass-through" segments, in the sense that they honor DSCP markings and allow extension of VLANs through them. Figure 9 gives an example of a "pass-through" setup.

In both L2 and L3 circuit cases, scalability issues must be considered because both technologies require all involved network devices to be configured to recognize specific data flows. Both MPLS tunnels and dynamic circuits are technologies well suited to establish special connections between WAN endpoints and accommodate qualifying traffic between sites connected to these endpoints. However, dedicating an MPLS tunnel or a dynamic circuit to each individual flow between a pair of end sites may cause severe scalability problems, especially in the case of dynamic circuits. With MPLS tunnels, scalability depends on the limitations and efficiency of the WAN hardware, while reserved bandwidth is allocated only when qualifying flows are present. MPLS tunnels are unidirectional, so bidirectional flows require two separate WAN reservations, one for each direction. With L2 dynamic circuits, additional restrictions apply. Because a circuit behaves as an Ethernet-based VLAN, a fundamental requirement is the utilization of the same VLAN tag along the entire route covered by the circuit. All network devices along the path must use the same VLAN tag. This is a severe restriction as current devices support a total of roughly 4,000 tags with several tag ranges reserved for device use and for administrative reasons. Therefore, only a small fraction of the overall tag range is actually available for utilizing dynamic circuits, furthermore, each domain may have its own tag subset. The establishment and utilization of a circuit between two end sites requires all domains along the path to have a common subset of tags. In the current implementation of TeraPaths, this is required so that no tag conflicts exist when setting up a circuit. This requirement may be relaxed in the future by exploiting VLAN renaming capabilities.

In the TeraPaths testbed there is an agreement that 50 VLAN tags, 3550-3599, are reserved for dynamic circuit use. Ensuring that no tag conflicts exist within the testbed is relatively easy, because all testbed sites are serviced by ESnet and Internet2, which form a composite domain that can be configured by contacting a single OSCARS instance. Thus, it is possible to rely on OSCARS to select an available VLAN tag within a range suitable for the end sites involved.

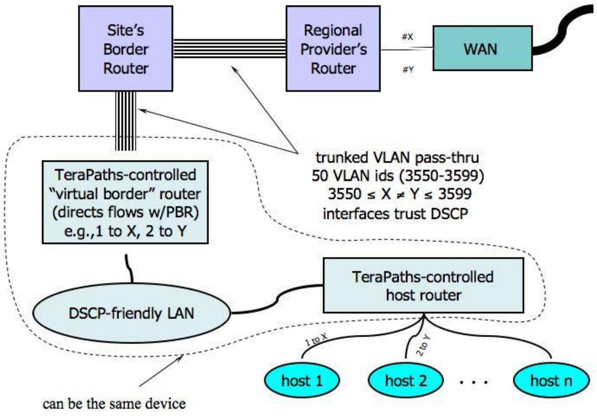


Figure 9. Example pass-through configuration for the end site's regional network and border router. The router where circuit VLANs terminate plays the role of a "virtual border" router. If only one router is controlled by TeraPaths, this router both conditions and forwards authorized traffic.

The limitation in the number of available VLAN tags and the additional properties of circuits to reserve bandwidth regardless of the presence of qualifying traffic and to be bidirectional make evident the need to treat L2 dynamic circuits as an "expensive" resource requiring sophisticated techniques to maximize utilization efficiency. Clearly, such circuits need to be viewed as "highways" between end sites. Flows with matching source and destination need to be grouped together and forwarded through common circuits, configured so that they accommodate the aggregate bandwidth of the grouped flows.

V. MANAGING WAN RESERVATIONS

Grouping together individual data flows or flow groups with common source and destination and forwarding them to a common WAN circuit with enough total bandwidth and duration to accommodate all flows can drastically reduce the number of circuits that are needed between a pair of end sites simultaneously and increase the availability of the dedicated paths. The first step of this approach is to decouple the end site reservations with the WAN reservations. End sites still reserve resources for individual flows, however multiple end site reservations can be accommodated by a single WAN circuit reservation as long as the aggregate duration and bandwidth can be determined. The level of reservation consolidation (or unification) needs to be controlled by suitable criteria to minimize waste of resources. Figure 10 shows an example of such criteria. If all reservations #1 through #5 were to be associated with a single encompassing WAN reservation, the resource waste would be significant because of the short but high-bandwidth reservation #4 and the distance in time between #4 and #5. Therefore, limits in the maximum difference in bandwidth between reservations (Δbw) and the time period between the end of one reservation and the beginning of the next (Δt) have to be taken into account when selecting which reservations should be consolidated.

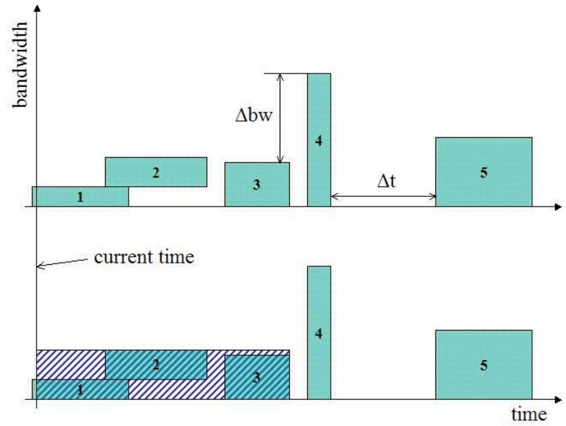


Figure 10. Example of reservation consolidation. Unifying reservations #1, #2, and #3 is feasible, #4 has too big Δbw, #5 is too distant in the future.

The initiating ESDC needs to handle the WAN reservations on the one hand, and the configuration of both end sites on the other. Although basic WAN reservation primitives can be used for consolidating reservations, additional primitives may be necessary to streamline the process and make it effective. Using basic primitives, the ESDC can create a new WAN reservation (for a dynamic L2 circuit this requires at least one VLAN tag to be available) to accommodate a newly arrived reservation that fulfills the criteria to use a specific circuit. If the circuit is pending, the consolidated WAN reservations can be immediately cancelled. However, if the circuit is already active, all relevant traffic must be switched to the new VLAN before the cancellation. With L3 circuits, this switching is not necessary. A problem with this technique is that the submission of the new WAN reservation may fail due to lack of available bandwidth occupied by reservations that will be cancelled. A new WAN primitive, allowing the submission of a reservation while taking into account the simultaneous cancellation of a set of existing ones would greatly increase the efficacy of the technique.

If the WAN domain controller allows modification of its reservations to a certain degree, it is possible to extend a reservation time-wise and/or to modify its bandwidth. While time-wise modifications are straightforward and are contingent on resource availability, bandwidth modifications need to be considered not only with regard to when they should take place within active or pending reservations, but also with regard to what the repercussions will be for existing connections through an active circuit which may be interrupted during reconfiguration.

We consider here two optimization and consolidation techniques for WAN reservations. We assume that initially WAN reservations correspond 1-to-1 to end site reservations. However, committing a reservation and deactivating a reservation are events triggering an optimization and consolidation phase for the WAN reservations. In both event cases, active or pending reservations within specific time “distance” before the beginning and/or after the end of a new reservation can be selected for consolidation. These techniques are roughly analogous to disk buffering or caching, i.e., “read ahead” and “write behind”. The goal of disk caching is to maximize the utilization of the disk and speed up access by buffering as much data as possible with read operations and before write operations. In a similar sense, selecting WAN reservations based on optimization criteria (e.g. reduce waste of resources) and consolidating them maximizes the utilization of a circuit and reduces the number of expensive create and teardown operations. We thus call these two techniques “create ahead” and “teardown behind.”

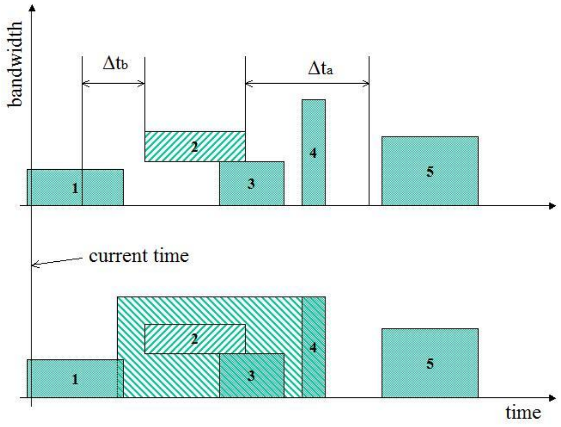


Figure 11. An example of "create ahead". #2 is a new reservation. Circuit corresponding to #1 is modified to accommodate #2, #3, and #4 with a single reservation. #5 is too distant.

“Create ahead” (see Figure 11) selects WAN reservations within Δt_b before the start of a new reservation and Δt_a after the end of a new reservation for consolidation, if additional limits in bandwidth differences and time distance are met. To reduce waste of resources, the second technique “teardown behind” (see Figure 12) modifies a unified reservation to conform to the bandwidth requirements at the time when the corresponding end site reservation expires by consolidating WAN reservations within Δt_a after the expiration of the end site reservation. The net result of the combination of the two techniques is to reduce the number of required circuits and the frequency of circuit creation and teardown operations for circuits between the same end sites while also reducing the waste of WAN resources.

In the remainder of this section, we formulate the reservation consolidation problem and devise an algorithm to apply the above techniques to minimize the request blocking rate. We consider both the offline case, where a set of reservation requests are given in a batch, and the online case, where a new request is serviced with possible reconfiguration of existing reservations. Extensive simulation results show the tradeoff between bandwidth utilization and VLAN ID utilization.

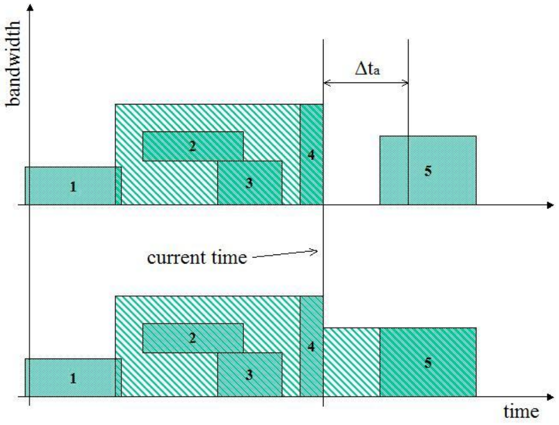


Figure 12. An example of "teardown behind". When #4 expires, the circuit servicing #2, #3, and #4 is not torn down, but instead modified to accommodate #5.

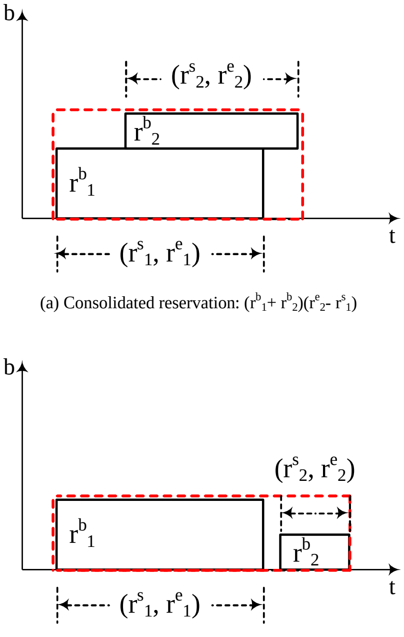
A. Models and Assumptions

An advance reservation request can be represented by a 3-tuple $r_i = (r_i^s, r_i^e, r_i^b)$, which asks for a reservation with bandwidth r_i^b within an active window (r_i^s, r_i^e), where r_i^s is a future starting time. The main challenging issue is, when given a request or a set of requests, to find the most cost-effective way to allocate bandwidth for each circuit and map each request to a circuit. In our model, one circuit has to be established with a constant bandwidth during its life since bandwidth-varying circuit reservations are not supported in the WAN. However, more than one reservation can be consolidated at the end site and then be carried on one circuit. This flexibility intuitively leads to two benefits: saving VLAN IDs and reducing the number of tear-down and setup operations. These two benefits are important because the number of VLAN IDs can be very limited in practice and the tear-down and setup operations can be costly. The downside of consolidating reservations with different bandwidth requests and active windows is that not all reserved bandwidths are used for the actual data transfer during certain intervals, which translates to lower resource utilization. In the following, we will study the tradeoffs between bandwidth utilization and circuit management efficiency.

B. Bandwidth Allocation and Circuit Assignment (BACA)

1) Offline case

We first study the problem of how, given a set R of requests $r_i, i \in \{1, 2, ..., m\}$, to allocate bandwidths and assign requests to circuits such that the maximum number of requests can be satisfied. In this way, the service provider can accommodate as many requests as possible or in other words, achieve high availability.



(a) Consolidated reservation: $(r^b_1 + r^b_2)(r^e_2 - r^s_1)$

(b) Consolidated reservation: $r^b_1(r^e_2 - r^s_1)$

Figure 13. Illustration of reservation.

More specifically, we need to make decisions on 1) the bandwidth allocation c_j^b and active duration (c_j^s, c_j^e) for each circuit $c_j, j \in \{1, 2, ..., n\}$, and 2) the assignment of reservations to circuits $x_{ij}, i \in \{1, 2, ..., m\}, j \in \{1, 2, ..., n\}$. The objective is to satisfy as many requests as possible, while observing the following constraints:

- Each reservation is assigned to a circuit.
- The total bandwidth used at any time is bounded by a given capacity *C*.
- If a reservation is assigned to a circuit, its active window must be within the active window of that circuit.
- Within one circuit, the maximum simultaneous data transmission rate must be bounded by the bandwidth allocated for that circuit.
- The bandwidth utilization in each circuit must be higher than a given value β.
- The number of available circuit IDs are constrained by a given value.

2) *Efficient Heuristics for the BACA Problem*

First, we order requests by their start times such that $r_i^s < r_j^s, i < j$. Second, if two reservations are not overlapping but are close enough to justify consolidation against additional tear-down and setup operations, we also consider them "overlapping", which makes them subject to consolidation too. Last, we perform admission control. That is, if $r_i^b > C$, we reject (and remove) the request by setting $x_{ij} = 0, \forall j \in \{1, 2, ..., n\}$. Before we describe the heuristic, we define the following:

- One-to-one assignment: allocate a circuit *c* for a request *r* by setting $c^b = r^b, c^s = r^s, c^e = r^e$ and set $x_{rc} = 1$.
- Consolidated reservation: If two reservations are overlapping, $r^v = (r_1^b + r_2^b)(\max(r_1^e, r_2^e) - \min(r_1^s, r_1^s))$, which is illustrated in Figure 13(a), where the *x* axis is time *t* and *y* axis is bandwidth *b*. If two reservations are not overlapping but very close, $r^v = \max(r_1^b, r_2^b)(\max(r_1^e, r_2^e) - \min(r_1^s, r_1^s))$ as illustrated in Figure 13 (b).
- Minimum bandwidth utilization guarantee: If $r^v \le \frac{1}{\beta}[r_1^b(r_1^e - r_1^s) + r_2^b(r_2^e - r_2^s)]$ is satisfied.

Now, we describe the algorithm as shown in Figure 14.

```
initiation.
while $R$ is not empty and $k \le n$ do
    select request r with earliest start time from R. Set consolidation = false.
    let $R'$ be the set of reservations that will start (have not yet started) earlier than r and also overlaps with r
    for each reservation $r' \in R'$ in increasing order of consolidated reservation volume. do
        if the consolidated reservation (from r and r′) meets the minimum bandwidth utilization guarantee and does not violate the total capacity constraint then
            consolidate r and r′.
            set consolidation = true
        end if
    end for
    if consolidation = false then
        if all circuit IDs are used then
            reject r and return failure;
        else
            find first available ID k.
            if assigning r to circuit k in a one-to-one fashion violates the total capacity constraint then
                reject r and return failure;
            else
                assign r to circuit k in a one-to-one fashion.
    end if
end while
```

Figure 14. Proposed BACA algorithm.

3) *Online case*

The above algorithm can be easily adapted for use with an online case, where a new request is serviced without the information of future reservation requests. More specifically, given a new request, we retrieve its adjacent reservations within a predefined "optimization window" and form a set of reservations *R* (including the newly arrived one) for re-optimization. We then can use the above algorithm to reconfigure existing reservations in order to maximize the number of satisfied reservations. However, if the reconfiguration rejects existing reservations, we will reject *r* instead. In other words, only when the reconfiguration can reserve all the requests in *R* do we actually commit the new configurations in the reservation table. In addition, those reservations in *R* that have already been in effect will not be reconfigured. However, we need information about them in the re-optimization in order to obtain the current bandwidth and VLAN ID usage.

C. *Qualitative Analysis*

In general, if we require a higher bandwidth utilization β when we optimize bandwidth allocation and circuit assignment using reservation consolidation, more VLAN IDs will be used. In the extreme case when $\beta = 100\%$, each reservation uses a distinguished VLAN ID. In this way, we limit the bandwidth waste in each circuit (as shown in Figure 12) so that the total capacity consumption is lower. The above qualitative analysis or hypothesis is summarized in Table 1:

Bandwidth utilization β in one circuit	VLAN ID Consumption	Capacity Consumption
high	high	Low
low	low	High

Table 1. Qualitative analysis summary.

In the following, given the relative magnitude of available number of VLAN IDs and available capacity, we conduct simulation to obtain the bandwidth utilization β that leads to lowest (or desired) job blocking rate.

D. *Numerical Study*

In this section, we simulate a large number of come-and-go jobs (i.e., the online case) and evaluate the proposed BACA algorithm considering a variety of cases. To facilitate the presentation, we define a ratio r_{cb}, which is used to govern the magnitude of average bandwidth of requests compared to the total capacity and traffic intensity. The traffic intensity is defined to be the product of average request arrival rate and average reservation duration. In the simulation, we use r_{cb} to generate various jobs with different average bandwidth requests as follows:

$$\text{Average bandwidth} = \frac{\text{total capacity}}{(\text{traffic intensity} \times r_{cb})}$$

We now present results for the following cases:

1) *Case 1: Sufficient VLAN IDs and varying bandwidth requests*

As shown in Figure 15 (assuming 10 VLAN IDs and varying r_{cb}), higher bandwidth utilization leads to a lower blocking rate in all cases. Therefore we can verify that reservation consolidation wastes bandwidth and result in higher blocking rate when bandwidth resource is scarce. More than 10 VLAN IDs will not make any difference. Therefore, 10 IDs are considered sufficient.

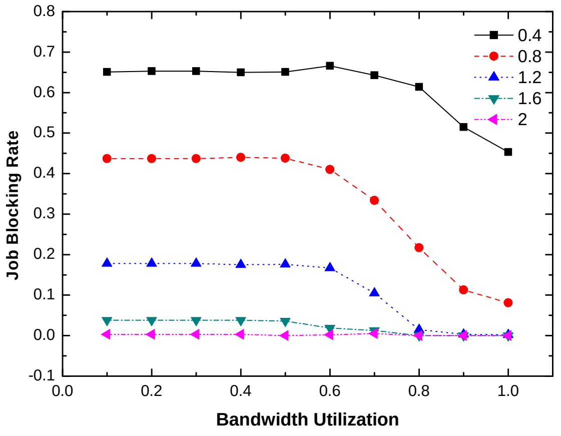


Figure 15. Sufficient (10) VLAN IDs, varying r_{cb}

2) *Case 2: Sufficient capacity and varying number of available VLAN IDs*

Figure 16 shows that reservation consolidation reduces the job blocking rate greatly when we have sufficient capacity (assuming r_{cb} =2) and varying number of available VLAN IDs in all cases. By “sufficient capacity”, we mean r_{cb} is large enough so that a job will not be blocked due to the capacity constraint. Any value of r_{cb} larger than 2 will not make any difference.

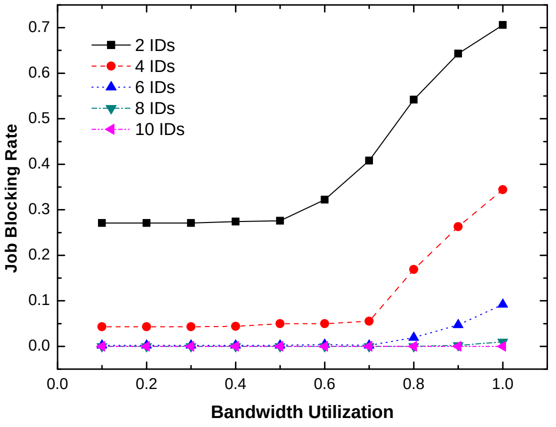
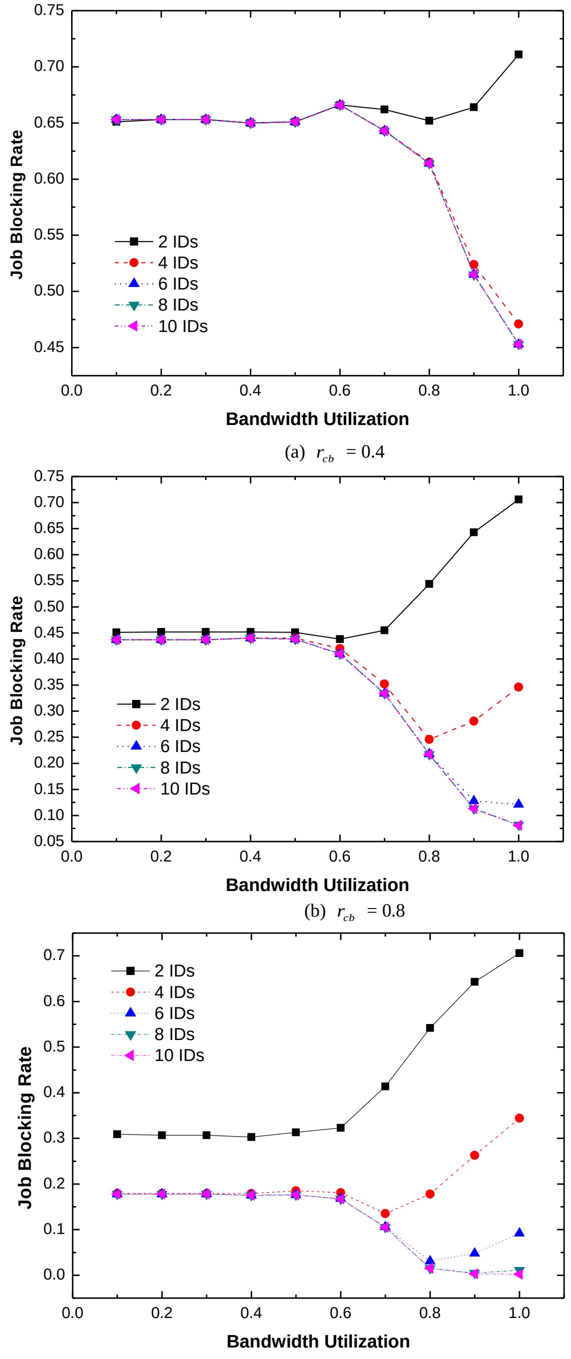


Figure 16. Sufficient capacity (r_{cb} =2), varying VLAN IDs

3) *Case 3: Limited number of available VLAN IDs with different bandwidth requests*

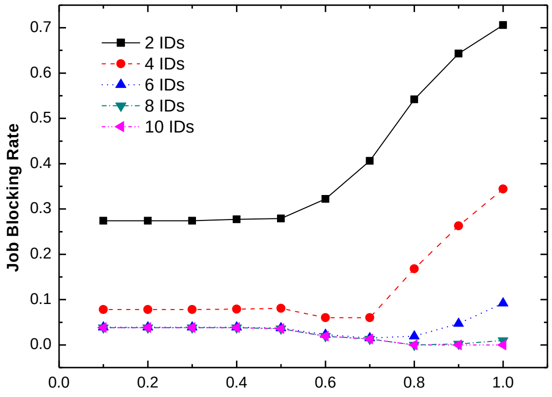
We further examine other cases. In each subfigure of Figure 17, we fix one value of r_{cb} and evaluate the job blocking performance with varying number of available VLAN IDs. For example, when $r_{cb} = 1.2$ and the bandwidth utilization is larger than 0.6, the blocking rate in the case of 2 available IDs begins to increase as in Case 1. However, we see a drop in blocking rate in other cases when we have more IDs. The uses of available IDs (by reducing circuit consolidation) can compensate for limited bandwidth.



(a) r_{cb} = 0.4

(b) r_{cb} = 0.8

(c) r_{cb} = 1.2



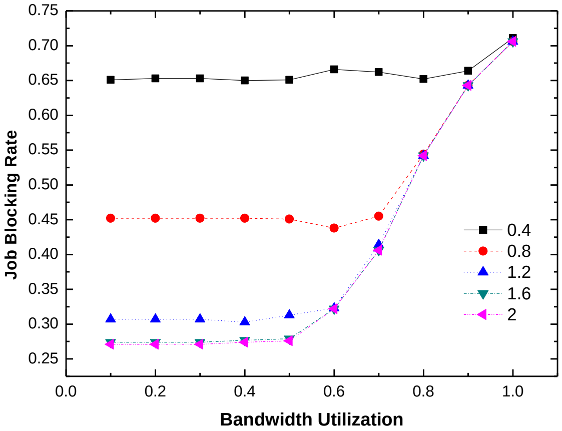
(d) r_{cb} = 1.6

Figure 17. Job blocking rate when r_{cb} = 0.4, 0.8, 1.2, 1.6

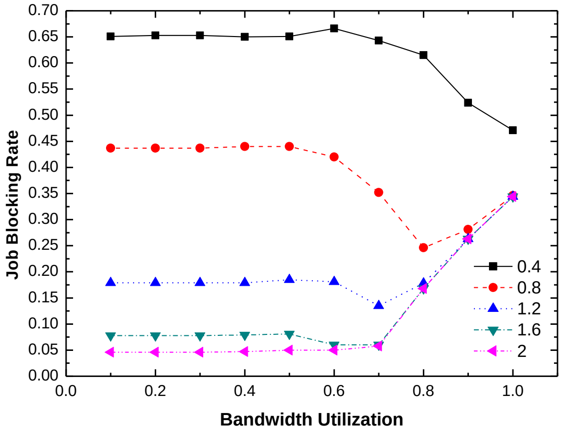
When bandwidth utilization increases further, we can see that all IDs are used up and then the blocking rate begins to increase again.

4) Case 4: Varying bandwidth requests under different number of available VLAN IDs

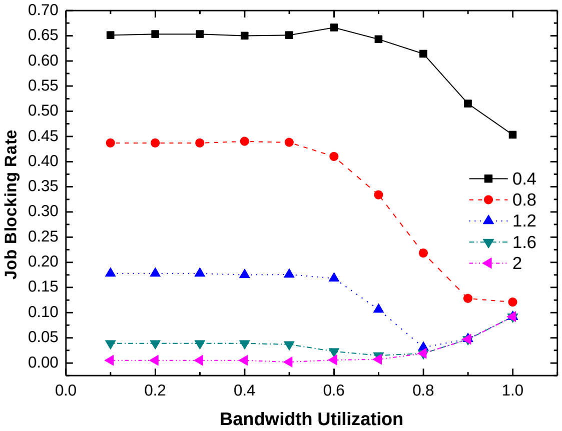
The graphs in Figure 18 also verify our hypothesis. In each subfigure below, we fix one value of available VLAN IDs and evaluate the job blocking performance with varying r_{cb}. These results can be explained by similar arguments as in Case 3.



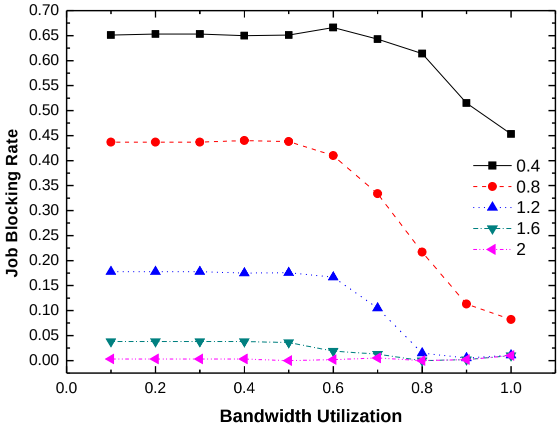
(a) 2 VLAN IDs



(b) 4 VLAN IDs



(c) 6 VLAN IDs



(d) 8 VLAN IDs

Figure 18. Job blocking rate when there are 2, 4, 6, 8 VLAN IDs

VI. FAULT TOLERANCE ISSUES

The survivability of a data transfer is crucial for data transfer applications. In TeraPaths, we view the survivability issue from a "do no harm" perspective. Because TeraPaths reserves an end-to-end path for better servicing the needs of an application, which may or may not be aware of the TeraPaths technology, our primary concern is to avoid situations where an application is disrupted because of a failure along the established end-to-end path. As such, we have started focusing on techniques to early detect and remedy configuration failures within end-sites network devices, and also handle WAN circuit failures.

In the event of a circuit failure, for any reason, flows that are being directed into that circuit will be interrupted, causing the corresponding applications to lose their connections. To prevent such situations, TeraPaths utilizes active circuit probing at the network device level. In this context, the end site network devices (border routers) that are the end points of a WAN circuit, periodically or on-demand exchange probes through that circuit for the duration of each related reservation. When a failure is detected, the immediate step is to stop forwarding traffic into the failed circuit and fall back to the standard IP network.

The next step is to attempt to acquire a new circuit and redirect traffic back into it (see Figure 19), while extending the reservations by the amount of time lost. The latter step is subject to WAN circuits becoming available again. Therefore, TeraPaths will keep trying for a pre-determined amount of time, after which the reservation will be considered failed.

With frequent periodic probes, it is possible to catch a circuit failure early and attempt to remedy the problem so that applications don't lose their connections. This approach is transparent to applications, however, it can impose significant load on the network hardware with increasing number of reservations. Thus, only highly critical reservations should be safeguarded with frequent periodic probing. A more scalable solution is to make applications aware of the probing/recovery capabilities (TeraPaths exposes these capabilities through its API) and enable them to trigger probing and recovery on-demand.

An alternative, albeit more resource-consuming, approach to recovery is to reserve in advance a backup circuit and, upon detection of failure, switch application traffic to it, instead of failing over to best effort and attempting to re-acquire the failed circuit. Steering traffic from one circuit to another is essentially instantaneous, once a failure is detected, therefore, the application should not notice anything more than a short-lived variation in bandwidth. We plan to explore this approach in our future work.

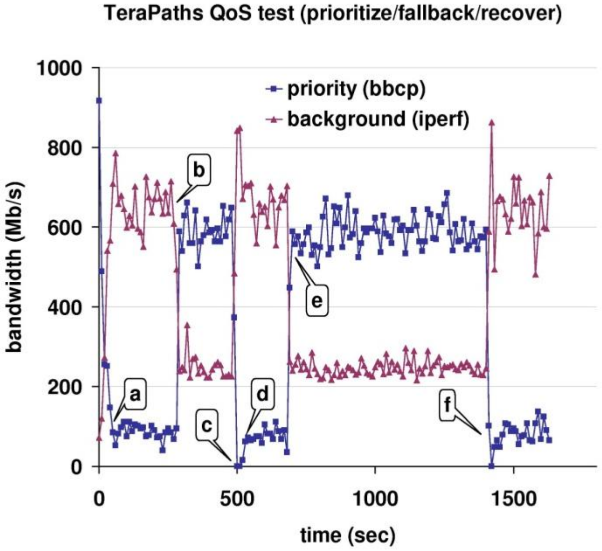


Figure 19. Demonstration of recovery: (a) competing traffic causes drop in bandwidth, (b) QoS/circuit reservation active, (c) circuit failure, (d) fall back to best effort, (e) recovery (acquired new circuit), (f) end of reservation.

VII. Related Work

The design parameters and goals of the TeraPaths project, i.e., provisioning of true end-to-end (host-to-host) virtual paths through direct configuration of end-site network devices and indirect configuration of WAN domains through tight interoperation with OSCARS, are, to the best of our knowledge, unique. In this section we compare our approach with several other systems, with which some similarities exist, in terms of design and implementation differences.

Lambda Station [11] is a Fermi National Accelerator Laboratory (FNAL) project with the goal to provide specific data intensive applications with alternate network paths between local production computing resources and advanced high performance networks. The Lambda Station service selectively forwards authorized data flows to alternate network paths, allowing such flows to utilize premium high bandwidth connections between end sites.

Phoebus [12], an Internet2 project, is a framework and protocol for high-performance dynamic circuit networks. The Phoebus approach is to split the end-to-end network path into distinct segments at "adaptation" points located at backbone ingress and egress points, then find and create an optimized network path for a specific application from each such point. Application-generated traffic between end sites is redirected to the circuit network via Phoebus Gateways.

While TeraPaths, Lambda Station, and Phoebus are all "consumers" of WAN circuits through OSCARS, TeraPaths is unique in that it uses DiffServ QoS and traffic conditioning at the edges to provide QoS guarantees to each individual flow within a group of flows going through the same WAN circuit and utilizes WAN circuit reservation consolidation techniques to practically address scalability issues.

Curti et al. [17] describes a system that can make advance reservations of lightpaths and MPLS-based layer-2 VPNs with QoS support in a large-scale network infrastructure. The authors mention possible approaches where users can scan the advertised resources of each domain and make a reservation by themselves in each administrative domain. However, synchronization problems may arise in the latter case if several reservation requests are processed at the same time.

Advance reservations have been studied in various scenarios and in different contexts. In the case of bulk data transfers, Rajah et al. [18] and Chen and Primet [19] have taken a centralized approach where resource reservation and allocation decisions are based on a global view of the network and on all job requests. As a result, it is possible to allocate network resources more efficiently. In order to improve the resource utilization, other approaches were considered in [20, 21]: a) transferring the data at time-varying bandwidth instead of constant bandwidth; b) using multiple paths for each job. For applications involving a large number of users and reserving resources from multiple domains, a distributed approach is expected to be more appropriate due to its better scalability and flexibility. In the case of distributed advance reservations, users and resource managers may need to negotiate on the reservation schedule in order to increase the success rate of submitted requests. For example, it was proposed in [20][21] that the resource manager should find another acceptable set of reservation characteristics and attach it to the resource allocation acknowledgment that is being returned to the requestor when rejecting a resource allocation request. Furthermore, if users are willing to negotiate a flexible reservation schedule (which is likely to happen in practice), the chance of satisfying requests is increased. Yuan et al. [22] proposes a probing mechanism to deal with requests that may have certain flexibility in starting time, duration or bandwidth (but only on one dimension). However, none of the above deals with the issues of providing connectivity with a specific service guarantee across heterogeneous network domains. In particular, previous studies have not studied the benefit of reservation consolidation.

In [5][23], a prototype of General-purpose Architecture for Reservation and Allocation (GARA) was implemented to support end-to-end QoS for high-end applications. The goal of the GARA framework was to support high bandwidth flows with different QoS specifications, provide advance reservation mechanisms, and facilitate application-level monitoring. In GARA, a resource manager works as a broker to reserve and manage various types of resources, such as bandwidth, CPU and disk. A major difference between GARA and TeraPaths is that in GARA the resource manager is deployed at each domain to control resources and only deals with layer 3 flows, whereas in TeraPaths a major challenge comes from the need to reserve resources across different domains (end-site LANs and multiple WAN domains in between) controlled by heterogeneous systems and deal with traffic in layer 2 and layer 3. As a result, TeraPaths selectively conditions and forwards layer 3 traffic into layer 2 to utilize dynamic WAN circuits and also addresses issues such as reservation consolidation and reservation negotiation across different domains to improve resource utilization and availability.

VIII. Conclusion and Future Work

New network capabilities enable the establishment of end-to-end QoS-aware paths across multiple domains, paths that can be dedicated to individual data flows. Although the overall framework is in its first steps, the technology is promising as it coexists with standard best-effort networking and is accessible transparently to specific data flows. We discussed issues involved with the utilization of WAN circuits from the perspective of end sites and presented techniques that the TeraPaths system utilizes for addressing the problem of scalability with increasing number of flows. We specifically focused on the problem of maximizing system availability (minimizing job blocking rate) constrained by limited VLAN IDs and bandwidth. This is a new problem, specifically encountered when utilizing L2 dynamic circuits, which we needed to address with novel heuristics. The effective resolution of this problem will make the technology applicable to an ever-growing number of data flows between end sites and will enable effective network scheduling. Our main approach, reservation consolidation, was shown to be effective in utilizing resources through extensive simulation studies.

The TeraPaths team continues the research and development effort to improve the functionality and reliability of the TeraPaths framework, in close collaboration with the OSCARS developers. Our near future plans include study and evaluation of an efficient negotiation protocol across multiple administrative domains to complement our BACA algorithm in providing end-to-end bandwidth guaranteed connections. This negotiation protocol considers flexible/negotiable user requests, suggestions of alternative reservations from services providers, and time-varying bandwidth within the same reservation in order to push the resource utilization as high as possible. We also plan to expand and improve upon the fault tolerance capabilities of TeraPaths, not only by pursuing early failure detection and recovery in a scalable way, but also by exposing services that make applications aware of such capabilities and enable them to request status checks and/or failover actions whenever they deem it necessary. In the longer term, we intend to incorporate the framework into a more general, application-centric network virtualization system. This system will provide individual applications with on-demand guaranteed network resources dedicated and tuned to their needs while isolating them from interference from other applications and strengthening security.

References

[1] D. Katramatos, D. Yu, K. Shroff, S. McKee, and T. Robertazzi. (2009, Aug). Establishment and Management of Virtual End-to-End QoS Paths through Modern Hybrid WANs with TeraPaths. Proceedings of the First International Conference on Evolving Internet (IARIA/IEEE INTERNET 2009), Cannes/La Bocca, French Riviera, France, August 23-29, 2009.

[2] High-Performance Networks for High-Impact Science. Report of the High-Performance Network Planning Workshop. August 2002. [Online] Available: http://www.doecollaboratory.org/meetings/hpnpw/finalreport/high-performance_networks.pdf (most recent access: May 2009).

[3] DOE Science Networking Challenge: Roadmap to 2008. Report of the DOE Science Networking Workshop (June 2003) [Online] Available: http://www.es.net/hypertext/welcome/pr/Roadmap/Roadmap%20to%202008.pdf (most recent access: May 2009).

[4] GGF4: Network QoS applied to GRIDs BOF. [Online]. Available: http://server11.infn.it/netgrid/ggf/ggf4-qos-bof/index.html (most recent access: December 2009).

[5] I. Foster, M. Fidler, A. Roy, V, Sander, and L. Winkler. (2004). End-to-End Quality of Service for High-end Applications. *Computer Communications*, 27(14):1375-1388, 2004.

[6] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. (1998, Dec.). An architecture for differentiated services. IETF RFC 2475. [Online]. Available: http://ietf.org/rfc/rfc2475.txt (most recent access: May 2009).

[7] R. Braden, D. Clark, and S. Shenker. (1994, June). Integrated Services in the Internet Architecture: an Overview. IETF RFC 1633. [Online]. Available: http://www.ietf.org/rfc/rfc1633.txt (most recent access: May 2009).

[8] E. Rosen, A. Viswanathan, and R. Callon. (2001, Jan.). Multiprotocol label switching architecture. IETF RFC 3031. [Online]. Available: http://www.ietf.org/rfc/rfc3031.txt (most recent access: May 2009).

[9] E. Mannie. (2004, Oct.). Generalized Multi-Protocol Label Switching (GMPLS) Architecture. IETF RFC 3945. [Online]. Available: http://www.ietf.org/rfc/rfc3945.txt (most recent access: May 2009).

[10] The TeraPaths End-to-End QoS Networking Project. [Online]. Available: http://www.terapaths.org (most recent access: May 2009).

[11] The Lambda Station project. [Online]. Available: http://www.lambdastation.org (most recent access: May 2009).

[12] The Phoebus project. [Online]. Available: http://e2epi.internet2.edu/phoebus.html (most recent access: May 2009).

[13] On-demand Secure Circuits and Advance Reservation System (OSCARS). [Online]. Available: http://www.es.net/oscars/ (most recent access: May 2009).

[14] T. Lehman, X. Yang, C. P. Guok, N. S. V. Rao, A. Lake, J. Vollbrecht, and N. Ghani. (2007, May). Control Plane Architecture and Design Considerations for Multi-Service Multi-Layer, Multi-Domain Hybrid Networks. INFOCOM 2007 IEEE. [Online]. Available: http://www.es.net/oscars/documents/papers/2007hsn-infocom-paper-lehman-etal.pdf (most recent access: May 2009).

[15] Energy Sciences Network (ESnet). [Online] Available: http://www.es.net/(most recent access: May 2009).

[16] Internet2. [Online] Available: http://www.internet2.edu/ (most recent access: May 2009).

[17] C. Curti, T. Ferrari, L. Gommans, S. van Oudenaarde, E. Ronchieri, F. Giacomini, and C. Vistoli. (2005). On advance reservation of heterogeneous network paths. *Future Generation Computer Systems*, vol. 21, no. 4, pp. 525 – 538, 2005. High-Speed Networks and Services for Data-Intensive Grids: the DataTAG Project.

[18] K. Rajah, S. Ranka, and Y. Xia. (2009, Nov.) Advance reservations and scheduling for bulk transfers in research networks. Parallel and Distributed Systems, IEEE Transactions on, vol. 20, pp. 1682–1697, Nov. 2009.

[19] B. B. Chen and P. V.-B. Primet. (2007). Scheduling deadline-constrained bulk data transfers to minimize network congestion. Cluster Computing and the Grid, IEEE International Symposium on, vol. 0, pp. 410–417, 2007.

[20] D. Ferrari, A. Gupta, and G. Ventre. (1997). Distributed advance reservation of real-time connections. *Multimedia Systems*, vol. 5, no. 3, pp. 187–198, 1997.

[21] A. Hafid, G. von Bochmann, and R. Dssouli. (1998). A quality of service negotiation approach with future reservations (nafur): a detailed study. *Comput. Netw. ISDN Syst.*, vol. 30, no. 8, pp. 777–794, 1998.

[22] L. Yuan, C.-K. Tham, and A. L. Ananda. (2003). A probing approach for effective distributed resource reservation," in *QoS-IP 2003: Proceedings of the Second International Workshop on Quality of Service in Multiservice IP Networks*, (London, UK), pp. 672–688, Springer-Verlag, 2003.

[23] V. Sander, I. Foster, A. Roy, and L. Winkler. (2000). A differentiated services implementation for high-performance TCP flows. Computer Networks, Vol. 34, No. 6, pp. 915-929, 2000.