AR-PIN/PDC: Flexible Advance Reservation of Intradomain and Interdomain Lightpaths

Eric He, Xi Wang, Venkatram Vishwanath, Jason Leigh
Electronic Visualization Laboratory
Department of Computer Science
University of Illinois at Chicago
eric@evl.uic.edu

Abstract — **A collection of Grid computing resources interconnected by an application-configurable network of lightpaths is called a LambdaGrid. It provides data-intensive applications with the needed deterministic network bandwidth to transport data between grid instruments, high-performance storage systems, compute clusters and visualization systems, that is often needed for real-time interactive scientific exploration. Advance reservation is needed to guarantee the availability of network resources. Flexible scheduling affords users greater convenience while also improving resource utilization and acceptance rate. In this paper we propose a unified flexible advance reservation model called FARM and apply this model to the cross-domain Routing and Wavelength Assignment problem. We will present the architecture and implementation of a coordinated Interdomain and Intradomain optical control plane, which is capable of flexible advance reservations. Our simulation results show that by just relaxing the reservation constraint and providing flexibility on the starting time, the network can carry 39% more load, and resource utilization can improve by 41%.**

Keywords-LambdaGrid, Advance Reservation, Interdomain, Intradomain, RWA, Lightpath

I. Introduction

A collection of Grid computing resources interconnected by an application-configurable network of lightpaths is called a LambdaGrid [1]. This provides data-intensive applications with the necessary deterministic network bandwidth to transport data between grid instruments, high-performance storage systems, compute clusters and visualization systems, which is often needed for real-time interactive scientific exploration. Advance reservation is needed to guarantee the availability of network resources. The nature of resource reservations in Grid computing is quite different from those of telephone calls. For the latter, their durations are usually not known in advance and hence cannot be planned in advance. In contrast, resource allocations in Grid environments usually require large amount of different types of resources be acquired simultaneously. Therefore, they have to be reserved in advance, just like you need to reserve hotels, airlines, and cars in advance when you plan to have a pleasant travel.

For customers, the major performance parameter of resource reservations is *acceptance rate* or *blocking rate*, which is defined as the ratio of accepted (blocked) reservation requests of all submitted requests. For network operators, the major performance parameter is *resource utilization*, which is related directly to their revenue. In comparison to immediate reservations, *advance reservations* usually degrade the resource utilization and the acceptance rate due to the resource fragmentation [2]. In order to improve the network performance, fragmentation must be avoided. Allowing flexibility in defining the advance reservations can result in better resource utilization while offering greater convenience to users. In this paper we will examine, through simulations, the degree by which flexibility affects performance.

The OptIPuter [3] is a National Science Foundation funded project to interconnect distributed storage, computing and visualization resources using photonic networks at tens of gigabits per second. The main goal of the project is to exploit the trend that network capacity is increasing at a rate far exceeding processor speed, while at the same time plummeting in cost. This allows one to experiment with a new paradigm in distributed computing - where the photonic networks serve as the computer's system bus and compute clusters, taken as a whole, serve as the peripherals in a potentially planetary-scale computer. We consider photonic networks as all-optical networks comprised of optical fibers and 3D MEMS (Micro-Electro-Mechanical Systems) optical switching devices. There is no translation of photons to electrons in photonic networks, hence we can avoid electronic bottlenecks. MEMS optical switches are controlled by special control software called Photonic Domain Controller (PDC), that allows applications to request and acquire end-to-end lightpaths.

An increasing number of organizations are buying dark fiber or wavelengths, and they want to share their resources with each other, similar to how they might share computing resources in Grid environments. Photonic Interdomain Controller (PIN) is software that allows applications to provision and control multi-domain lightpaths [4]. PIN specializes in the interdomain routing and signaling schemes over heterogeneous optical network domains. In a multi-domain environment, security management and policy administration are also critical. Our collaborator, University of Amsterdam, has done some pioneering research on Authorization, Authentication and Accounting (AAA) and we are leveraging it within PIN software [12].

Incorporating flexible advance reservation into PIN/PDC is not trivial. Because PIN/PDC is based on all-optical networks, one main problem that PIN/PDC have to solve is Routing and Wavelength Assignment (RWA). The RWA problem is a NPhard problem. Usually it can be simplified by decoupling the problem into two sub problems: the routing problem and the wavelength assignment problem. The routing problem can be solved by Fixed Routing, Fixed Alternate Routing, or Adaptive Routing algorithms. Adaptive Routing is considered to be able to achieve the best performance by feeding the wavelength assignment status back to the routing algorithm [6]. The flexible advance reservation introduces a new temporal dimension into the resource allocation. The wavelength resources along the path have to maintain both wavelength and temporal continuity. It is difficult and computationally expensive to summarize meaningful resource status back to the routing algorithm. In this paper we will compare the performance of the three routing algorithms and found that the fixed alternate routing achieves the best performance.

For interdomain distributed control, the addition of a temporal dimension makes the resource state of each domain too large to disseminate to other domains. Therefore, only the relatively static topology summary information of each domain is disseminated to other collaborating domains. The Grid community consists of many Virtual Organization (VO) based collaborations, which means that the resource of each domain is usually not open for all the world, instead, each domain wants to define their own collaborators and individual access policy. We think that the peer-to-peer publish/subscription model is more effective in this regard and more scalable for interdomain topology exchange. The On-Demand Parallel Probe (ODPP) algorithm is used to find the best timeslot and wavelength two-dimensional match domain by domain. The new version of PIN/PDC with Advance Reservation (AR) is called AR-PIN/PDC.

The remaining paper is organized as follows. In section II we quickly go through some related work. In section III we describe a unified flexible advance reservation model. In section IV we elaborate the architecture of AR-PDC and AR-PIN. In section V we provide simulation result to show how flexibility improves network performance and compare the performance of different routing algorithms. Then the paper is concluded by section VI.

II. Related Work

Recently some significant research has been done on circuit-based intradomain or interdomain lightpath provisioning. User Controlled LightPath (UCLP) is a web service based software starting deployment in CA*net4 networks [7]. Bandwidth on Demand (BoD) concentrates on multi-domain policy-based access control [5, 12]. Circuit-switched High-speed End-to-End Transport arcHitecture (CHEETAH) provides end-to-end circuit connectivity by concatenating high-speed ethernet segments [8]. All of aforementioned work assumed the physical network is based on SONET or Ethernet segments and didn’t incorporate RWA algorithms. PIN and PDC assume that border switches are OEO switches and there is no wavelength continuity constraint between domains, and therefore only considered RWA algorithms within domains [4].

Advance reservation has been widely studied in networks other than all-optical networks. Zheng and Mouftah [9] studied the design of RWA algorithms for different types of advance reservations. Guerin and Orda [10] investigated the computational complexity of routing algorithms when supporting different models of advance reservations. Greenberg et al. [11] proposed a call admission control algorithm that occasionally allows a call in progress to be interrupted in order to efficiently share resources among book-ahead (BA) calls and non-BA calls. The Globus Architecture for Reservation and Allocation (GARA) is a toolkit used to implement advance reservations of grid resources in Globus software [13, 14].

The Dynamic Resource Allocation via GMPLS Optical Networks (Dragon) project is trying to build an interdomain lightpath resource management system and leverage Generalized Multi-Protocol Label Switching (GMPLS) standard as intradomain control plane [15]. It takes advance scheduling and AAA into consideration during the end-to-end path computation. In the process of state exchange, the topology or topology summary, Label Switching Path (LSP) reservation information, and AAA policy information of each domain will be disseminated to all other domains. This puts a huge amount of load on the control plane network which usually has relatively low bandwidth.

In this paper, we describe a coordinated intradomain and interdomain control plane, taking into account both cross-domain RWA and advance reservation. We propose a publish/subscription model and On-Demand Parallel Probe (ODPP) algorithm to achieve the scalability of interdomain information dissemination. The intradomain control plane can work on not only GMPLS-enabled switches, but also bare MEMS switches. Through simulations, we found that flexibility in advance reservations can improve performance dramatically. We also compared the performance of different path computation algorithms when dealing with advance reservation requests.

III. Unified Flexible Advance Reservation Model (FARM)

An advance reservation request is typically characterized by a source node, a destination node, a bandwidth demand, a specified starting time, and a specified duration. In all-optical circuit switching networks, the bandwidth granularity is a wavelength. A request which needs multiple wavelengths can be decomposed into multiple requests wherein each request provisions a single wavelength. Therefore, in our scheme, we consider only single wavelength reservations.

Next we need to decide how to express the flexibility in the reservation requests. In [9], Zheng and Mouftah classified advance reservations into three types: specified starting time and specified duration (STSD), specified starting time and unspecified duration (STUD), and unspecified starting time and specified duration (UTSD). And they use different wavelength assignment algorithms for each request type. Actually there is another possibility that both starting time and duration are unspecified and only a range is specified with an earliest time and latest time (UTUD). We want to use this to express the flexibility because all three above types can be expressed by UTUD and some constraints such as the earliest, the longest. Once the RWA algorithm finds that there are multiple choices, we can use the constraints to select the best one.

Therefore, a unified flexible advance reservation is defined as follows:

$$R = (S, D, E, L, Dmin, Dmax) \quad (1)$$

Where S is the source node, D is the destination node, E is the earliest time, L is the latest time, Dmin is minimum duration, and Dmax is maximum duration. Dmax is an optional parameter.

IV. AR-PIN/PDC Architecture

AR-PIN and AR-PDC are interdomain and intradomain lightpath control software respectively. These two pieces of software work together to provision end-to-end interdomain lightpaths in advance. In this paper, we sometimes omit AR, so PIN and AR-PIN can be interchanged, and PDC and AR-PDC can be interchanged.

The system architecture is shown in Figure 2. We use an example to show the interaction sequences of users, PIN and PDC. The following steps will be executed when client A in domain 1 sends a reservation request to the PIN/PDC system:

- Periodically, all the collaborating domains exchange topology summary with each other.

1. Client A sends a lightpath reservation request to its local interdomain agent AR-PIN1.
2. AR-PIN1 computes the domain-level paths.
3. The source domain starts probe resource availability on multiple domain-level paths in parallel.
4. At each hop, each queried AR-PDC checks its own AAA policy, resource database, then returns the timeslot-wavelength availability matrix. The matrix will be intersected with the global matrix.
5. At the destination domain 3, the best switch-level path is selected. Then the reservations of all involved domains are done in parallel.
6. Within the reservation time window, the lightpath provisioning is triggered by committing the reservation. To do that, the device drivers send TL1 commands to switches to set up the end-to-end lightpath.

Next we will explain each important component in detail in remaining part of this section.

A. Interdomain Topology Summary and Exchange

A domain is an independently managed network cloud exposing a set of ingress and egress points and links with service specifications. Each link is controlled and managed by a single domain. The separation points between neighboring domains are switches. We call these switches as **border switches**. Ports of border switches can terminate links of multiple different domains. Every border switch needs a globally unique address or name for addressing purposes.

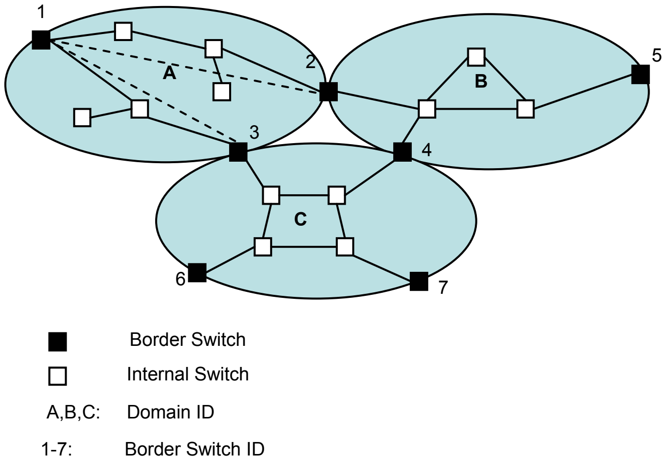

Figure 1. Structure of multiple photonic domains.

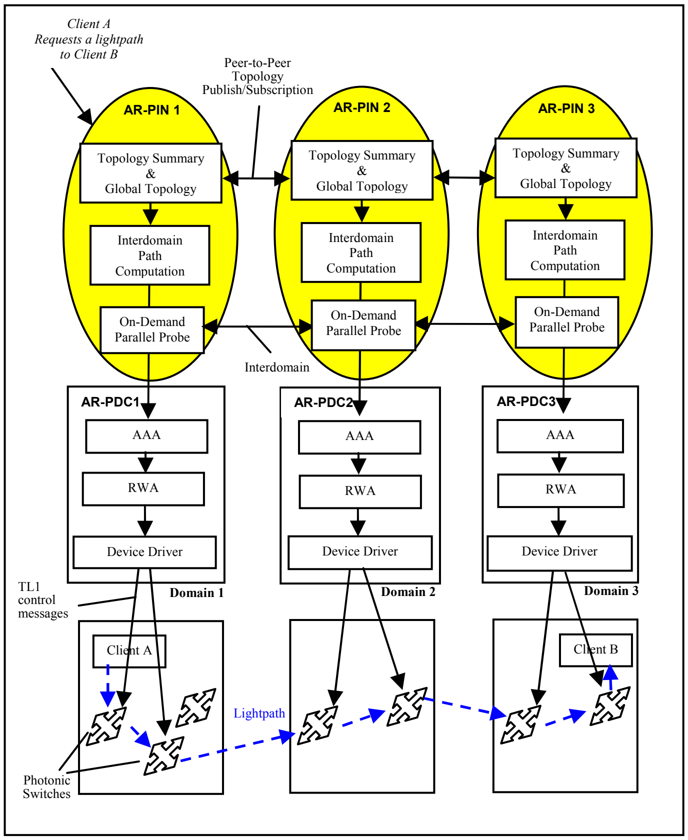

Figure 2. AR-PIN/PDC system architecture.

In the topology advertisement that a domain sends to other collaborating domains, it is unnecessary to include the details such as internal switches and internal links. Instead, it will just send out a topology summary of its own domain. The topology summary description consists of only border switches and abstracted links. For example, the advertisement from domain A will be:

Switch 1-2: wavelength w1, w2, w3, w4.

Switch 1-3: wavelength w1, w2, w3, w4, w5, w6, w7, w8.

These two abstracted links are shown as dotted lines in the Figure 1. The abstract link is actually an abstraction of a bunch of consecutive physical links in the same domain. The topology summary can be generated manually or automatically from the intradomain topology database. The topology summary generation is a maximum-flow problem and it can be solved by the Ford-Fulkerson method [16].

PIN runs a peer-to-peer publish/subscription based routing protocol to exchange topology summaries among different domains. The peer-to-peer exchange mode is more suitable than blinded flooding because it is possible that a domain may want to selectively advertise different sets of resources to different domains. The information exchange is based on subscription style. Every domain maintains a list of collaborating domains (subscriber). The information exchange is triggered by any change of the interdomain topology of the domain. In other words, whenever the topology changes in a domain, the topology summary will be regenerated, the PIN in this domain will update the new topology summary to all subscribed domains (push model) or just send a change notification and let the domains to request the update by themselves (pull model). Of course, the pull mode should always be supported in case of newly started domains or out-of-sync domains. After receiving the topology summaries from all collaborating domains, each domain can composite its own global topology. Because each domain gets different topology summaries from different collaborating domains, every domain has its own unique global topology database. In this global view, each node is a border switch, each link is an abstract link managed by a domain.

B. *Interdomain Path Computation*

When a lightpath reservation request comes in, the local domain will compute a domain-level path based on its own view of global topology. This path includes only border switches. Source routing will be used to compute the path. There are several possible path computation algorithms.

1) Fixed Routing: The simplest approach to routing a connection is to always choose the same fixed route for a given source-destination pair. One example of such an approach is fixed shortest-path routing. The shortest-path route for each source-destination pair is calculated beforehand using Dijkstra's algorithm.

2) Fixed-Alternate Routing: In fixed-alternate routing, for each source-destination pair, an ordered list of multiple fixed routes is calculated in advance. For example, these routes may include the shortest-path route, the second-shortest-path route, the third-shortest-path route, etc. Different routes are link-disjointed, which means that different routes for the same source-destination pair do not share any links. It may also be used to provide some degree of fault tolerance upon link failures.

3) Least-Load-Path Routing: Least-Load-Path Routing is one form of adaptive routing. In adaptive routing, the wavelength availability information will be fed back to the path computation algorithm. In order to decrease the blocking probability, we should choose routes with the least load in the requested time window.

For each lightpath request, we assign a weight w(i, j, t) to each network link (i, j) at reservation time window t, which is defined as:

$$w(i,j,t) = (1 - \frac{a(i,j,t)}{W}) \cdot l(i,j) \qquad (2)$$

Where a(i, j, t) is the average number of available wavelengths on the link (i, j) at time window t, which is the defined in equation (3). W is the number of total wavelengths on the link (i, j), l(i, j) is the link length.

$$t = L\text{-}E \qquad (3)$$

Where E is the earliest time of the reservation, L is its latest time.

We will compare the performance of these algorithms in the situation of flexible advance reservations in section V. Through simulation, we found fixed alternate routing has the best performance.

C. *On-Demand Parallel Probe (ODPP) and Reservation*

Although fixed alternate routing has the best performance, it takes a long time to iterate k fixed routes among multiple domains serially. Therefore, we will select k shortest disjointed paths and probe the resources **in parallel**. After the domain-level paths are decided, the next step is to translate the domain-level paths into switch-level paths by interacting with the PDCs along the paths. The detail of PDC will be explained in next subsection.

For example, if the domain level path is 1A2B5 in Figure 1. The PDCs in domain A and domain B need to compute an intradomain switch-level path from switch 1 to 2 and from switch 2 to 5, respectively. Not only that, PDC will also return a timeslot-wavelength availability matrix to PIN. Then PIN will then form a new timeslot-wavelength availability matrix by executing the intersection operation on the current availability matrix and the PDC returned one. The new matrix will be passed along downstream.

At the destination domain, after the intersection operation, if there are more than one candidate paths, the best one can be selected according to user preference such as longest reservation time, earliest reservation or shortest path, etc. Another option is to send back these choices to the user and let the user select the best one manually.

The next step is reservation. The reservation will be initiated by the destination domain to all other domains in parallel in order to reduce the end-to-end response time. After a successful reservation is made, each domain will send an acknowledgement message to the source domain. After collecting positive acknowledgements from all domains, this request is successfully fulfilled.

D. AR-PDC: Intradomain Control Plane

AR-PDC provisions intradomain lightpaths. Reservation requests may come from local domain users or its interdomain control plane AR-PIN. During the ARPIN On-Demand Parallel Probing process, it relies on AR-PDC to extend the domain-level path into a switch-level path and check the wavelength availability status.

a) Authorization, Authentication and Accounting (AAA)

When a reservation request comes from foreign domains, they need to go through the AAA mechanism to ensure the foreign user is authenticated. Then according to the identity of the user and the local access policy, the network resources will be filtered and a virtual topology will be generated and it will be used in the following Routing and Wavelength Assignment (RWA) operation.

b) Routing and Wavelength Assignment (RWA)

PDC does the RWA job at the switch level. We also divide the RWA problem into two sub problems: routing and wavelength assignment. For the routing problem, PDC can use Fixed, Fixed Alternate or Adaptive algorithms – same as interdomain path computation. For the wavelength assignment problem, we execute an intersection operation on all hops from the ingress switch to the egress switch and return the result timeslot-wavelength availability matrix to PIN. When we use the Fixed Alternate algorithm, we can return the matrixes of all paths to PIN and let PIN choose the best one according to the intersection result with the matrix of the explored part of the path.

c) Device Driver

If a request gets reserved successfully, the user needs to commit the request when he(she) wants to activate the reservation. Then each domain along the path will send TL1 commands to the MEMS switches to set up cross connects. So far we have built device drivers for Calient DiamondWave PXC and Glimmerglass Reflexion 3D MEMS switches. PDC software has unified interface to different types of MEM switches.

V. SIMULATIONS

We ran simulations on the NSFNET topology with 14 nodes as shown in Figure 3. We assumed that each link is a single bi-directional fiber with 8 wavelengths. The entire topology was fully-optical without any wavelength converter. The workload consists of advance reservations of which starting time is a Poisson process and the reservation duration is a negative exponential distribution with mean T. All requests try to reserve a lightpath with exactly one wavelength. For flexible reservation requests, we fixed the reservation duration and left the starting time flexible. We ran the simulations on 5 different generated workloads and took the average. The network load is decided by call inter-arrival time and call holding time:

Network load (Erlang) = call holding time/call inter-arrival time

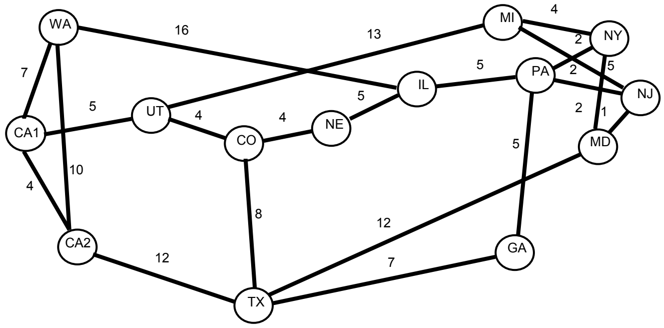

Figure 3. 14 node NSFNET topology.

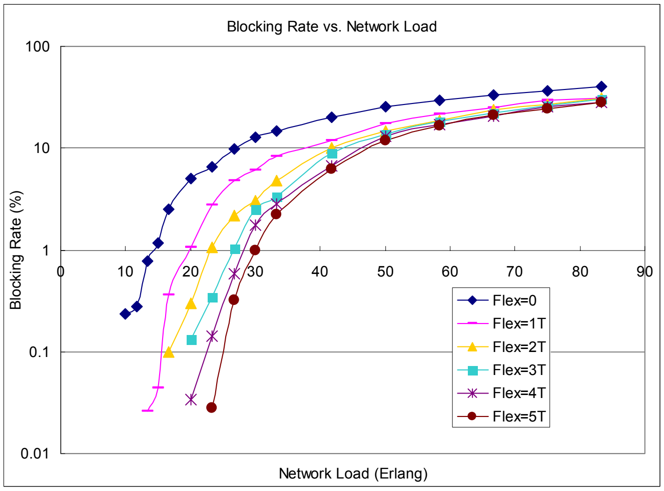

Figure 4. Blocking rate under different flexibility of advance reservations.

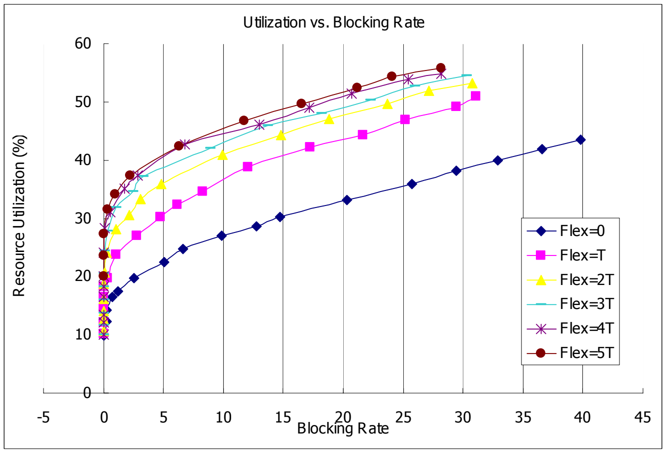

Figure 5. Resource utilization under different flexibility of advance reservations.

The goal of the first set of experiments is to evaluate how the flexibility affects the blocking rate and resource utilization of advanced reservations. We changed the flexibility of the starting time of reservations from 0, 1T, …, to 5T. T is the mean duration of reservations. Figure 4 shows that how the blocking rate varies with network load. Different curves are under different flexibility degree. We can see that just introducing 1T flexibility improves the performance considerably, but more flexibility does not help as much. Consider when the blocking rate is 5%, the system load is improved from 19.8 to 27.5 by introducing 1T flexibility. It shows 39% improvement. Figure 5 shows that how the relation of resource utilization vs. network load is affected by different flexibility. Also when the blocking rate is 5%, the resource utilization is improved from 22% to 31% by introducing 1T flexibility, it shows 41% improvement.

In another set of simulations, we wanted to evaluate how different routing algorithms (heuristics) perform under flexible advance reservations. The flexibility of starting time of reservations is 1T in simulations. We compared the three routing algorithms: Fixed Routing (FR), Alternate Fixed Routing (AFR), and Least Load Path Routing (LLP). LLP is one form of adaptive routing. In most prior RWA research in which flexibility is not considered, adaptive routing showed superior performance to FR and AFR routing. However, from the simulation results in Figure 6, we can see the Alternate Fixed Routing (AFR) algorithm has the best performance and the Fixed Routing (FR) has the worst. The flexible advance reservation introduces a new temporal dimension into the resource allocation. The resource availability information is a three dimensional cube of hop, wavelength and time slot. The reserved units scatter within this three dimensional cube. The uncertainty of time parameters makes it difficult to filter out useful resource status and feed it back to the routing algorithm. This is why the LLP routing algorithm cannot perform well by just calculating average utilization of each wavelength within the reservation time window.

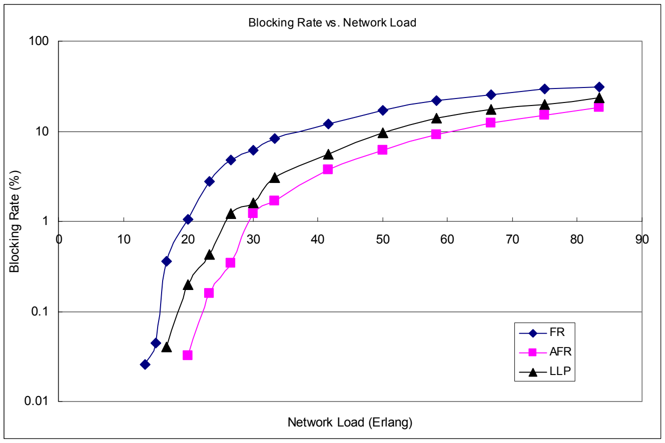

Figure 6. Blocking rate of different routing algorithms when flexibility of advance reservations is 1T.

VI. CONSLUSION

In this paper we described an architecture of intradomain and interdomain control planes for photonic networks which are capable of flexible advance reservations. A peer-to-peer based publish/subscription topology model is used to avoid huge amount of state flooding. The On-Demand Parallel Probe algorithm renders the periodic dissemination of time-based resource availability information unnecessary and hence makes the system more scalable. Through simulations, we found that by introducing some flexibility on the time parameters of advance reservations, the network performance can be improved dramatically. We also showed that the alternate fixed routing outperformed adaptive routing algorithm – Least Load Path routing. In the future we will implement the entire AR-PIN/PDC software, deploy them in real testbeds and measure important parameters such as resource utilization, end-to-end latency, etc.

REFERENCES

[1] T. DeFanti, M. Brown, J. Leigh, O. Yu, E. He, J. Mambretti, D. Lillethun, J. Weinberger, "Optical Switching Middleware for the OptIPuter", IEICE Transactions on Communications, invited paper in special issue on Photonic IP Network Technologies for Next Generation Broadband Access. Vol. E86-B, No. 8, pp. 2263

[2] Lars-Olof Burchard, Hans-Ulrich Heiss, Cesar A. F. De Rose, "Performance Issues of Bandwidth Reservations for Grid Computing", Proc. Of the 15th Sym. On Computer Architecture and High Performance Computing (SBAC-PAD'03), 2003.

[3] Larry L. Smarr, Andrew A. Chien, Tom DeFanti, Jason Leigh, Philip M. Papadopoulos, "The OptIPuter", Communications of the ACM, Volume 46, Number 11 (2003), Pages 58-67.

[4] O. T. Yu, T. A. DeFanti, "Collaborative User-Centric Lambda-Grid over Wavelength-Routed Network", In Proceedings of the 2004 ACM/IEEE Conference on Supercomputing (Nov 06 - 12, 2004), Washington, DC.

[5] L. Gommans, C. de Laat, B. van Oudenaarde, and A. Taal, "Authorization of a QoS path based on generic AAA", Future Gener. Comput. Syst. 19, 6 (Aug. 2003), 1009-1016.

[6] Hui Zang, Jason P. Jue, and Biswanath Mukherjee, "A review of routing and wavelength assignment approaches for wavelength-routed optical WDM networks," SPIE Optical Networks Magazine, vol. 1, no. 1, Jan. 2000.

[7] Raouf Boutaba, Wojciech Golab, Youssef Iraqi, and Bill St. Arnaud, "Lightpaths on Demand: A Web-Services-Based Management System", IEEE Communications magazine, July 2004, pp 2-9.

[8] M Veeraraghavan, X Zheng, H Lee, M Gardner, W Feng, "CHEETAH: Circuit-switched High-speed End-to-End Transport ArcHitecture", Proc. of Opticomm 2003, 2003.

[9] Jun Zheng and Hussein T. Mouftah, "Routing and Wavelength Assignment for Advance Reservation in Wavelength-Routed WDM Optical Networks", IEEE International Conference on Communications (ICC), 2002.

[10] R Guerin, A Orda, "Networks with Advance Reservations: The Routing Perspective", IEEE INFOCOM 2000, Tel-Aviv, Israel, March 26-30, 2000.

[11] AG Greenberg, R Srikant, W Whitt, "Resource sharing for book-ahead and instantaneous-request calls", IEEE/ACM Transactions on Networking, Vol. 7, No. 1, Feb 1999.

[12] S Van Oudenaarde, Z Hendrikse, F Dijkstra, L. Gommans, C. de Laat, R. Meijer, "Dynamic paths in multi-domain optical networks for grids", Future Generation Computer Systems, Vol. 21, No. 4, Page 539-548, Apr. 2005.

[13] I. Foster, C. Kesselman, C. Lee, R. Lindell, K. Nahrstedt, A. Roy, "A Distributed Resource Management Architecture that Supports Advance Reservations and Co-Allocation**"**, Int'l Workshop on Quality of Service, 1999.

[14] C Curti, T Ferrari, L Gommans, S Van Oudenaarde, et al, "On advance reservation of heterogeneous network paths", Future Generation Computer Systems, Vol. 21, No. 4, Page 525-538, Apr. 2005.

[15] Xi Yang, Tom Lehman, Chris Tracy, Jerry Sobieski, Payam Torab, Shujia Gong, Bijan Jabbari, "Policy-Based Resource Management and Service Provisioning in GMPLS Networks", Adaptive Policy-Based Management workshop, Barcelona, Spain, April 28, 2006.

[16] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, "Introduction to Algorithms", Second Edition, The MIT Press, 2001.