Admission Control in Data Transfers over Lightpaths

Wojciech Golab and Raouf Boutaba, *Senior Member, IEEE*

***Abstract*— The availability of optical network infrastructure and appropriate user control software has recently made it possible for scientists to establish end-to-end circuits across multiple management domains in support of large data transfers. These high-performance data paths are typically provisioned over 10 Gigabit optical links, and accessed using Ethernet encapsulation at Gigabit and 10 Gigabit rates. The resulting mixture of circuit sizes gives rise to resource conflicts whereby requests to allocate bandwidth partitions are blocked despite vast underutilization of the optical link. In an attempt to remedy this problem, we investigate intelligent admission control policies that consider the long-term effects of admission decisions. Using analytic techniques we show that the greedy policy, which accepts requests to allocate bandwidth partitions whenever sufficient bandwidth exists, is suboptimal in a pertinent scenario. We then consider dynamic online computation of the optimal admission control policy and show that the acceptance ratio of requests to establish end-to-end circuits can be improved by up to 19% on a fifteen-node network where the behaviour of each link is governed by a local optimization effort.**

***Index Terms*— Admission control, stochastic knapsack, user-controlled networks, optical networks.**

I. Introduction

A. Motivation

CIRCUIT switching has recently emerged as a viable method of providing high-capacity end-to-end connections in support of large transfers of scientific data. Circuits offer scientists an attractive alternative to the Internet as they intrinsically provide guaranteed bandwidth and minimal delay, while avoiding the costly electronics associated with high-speed queueing and scheduling hardware. Optical circuits, subsequently referred to as lightpaths, are particularly promising thanks to their superior capacity, low error rate, and low signal attenuation characteristics, as well as their favourable cost following a period of massive over-provisioning in the late 1990's. In fact, there is a strong trend for research institutions, schools, and large enterprises to purchase optical wavelengths or entire strands of optical fibre in order to connect to each other, or to Internet service providers, using their own switching equipment [1], [2].

User-owned optical networks not only offer significant cost savings over conventional carrier-managed services, but also make it possible for users to flexibly control and manage their infrastructure according to their particular needs. For example, using the recently-developed software for user control of lightpaths (see `www.canarie.ca`), users can create lightpaths on demand and across multiple management domains. In particular, this can be done through a client application using an interface based on emerging Web services standards, for example a modified GridFTP client [3]. Thus, bandwidth-guaranteed connections can be automatically created in support of individual data transfer sessions.

The hardware deployed in user-controlled optical networks comprises a variety of optical and electronic devices. The most popular switching elements are digital cross-connect systems (DCSs) implementing the Synchronous Optical Network (SONET) or Synchronous Digital Hierarchy (SDH) specifications, which provide powerful traffic grooming, performance monitoring, and protection switching capabilities. Data-bearing connections are typically created by encapsulating Ethernet frames, whereby remote local area networks are bridged using a SONET/SDH circuit. The wide-spread deployment of Gigabit Ethernet interfaces, along with the emergence of 10 Gigabit interfaces, makes these technologies a powerful basis for high-performance wide-area data networking.

Digital switching elements are complemented by all-optical Wavelength Division Multiplexing (WDM) hardware, which enables concurrent transmission of multiple optical carriers (i.e. wavelengths) over a common fibre. Switching devices based on micro-electro-mechanical systems (MEMS) make it possible to perform all-optical switching at wavelength granularity at a much lower cost than the corresponding digital hardware, but such devices have not yet seen wide deployment. Thus, we consider that user control over lightpaths is exercised primarily through creation and tear-down of cross-connections in digital switching fabrics.

The current state of circuit-switching technologies leads us to investigate the resource allocation problems associated with formation of digitally multiplexed lightpaths. Specifically, in this paper we consider the problem of admission control from a combinatorial perspective, where we attempt to maximize link utilization by considering the long-term effects of packing bandwidth partitions of various sizes within a common parent lightpath. This aspect of admission control is especially important in the user-controlled scenario where the bandwidths involved are large (i.e., comparable to link capacities), since in that case requests to allocate bandwidth partitions are more prone to rejection even when the link is vastly underutilized.

B. Organization and Contributions

The remainder of this paper is organized as follows. In Section II we review relevant prior work. In Section III we describe concrete scenarios under which the greedy admission

Manuscript received January 11, 2006; revised April 14, 2007. This research is partially supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

W. Golab is at the Department of Computer Science, University of Toronto, Canada (email: wgolab@cs.utoronto.ca).

R. Boutaba is at the David R. Cheriton School of Computer Science, University of Waterloo, Canada (email: rboutaba@cs.uwaterloo.ca).

Digital Object Identifier 10.1109/JSAC-OCN.2007.024806.

control policy is suboptimal, and present proofs based on closed-form expressions for the expected utilization ratio. In Section IV we present a novel policy optimization scheme along with simulation-based performance analysis. Our results are novel with respect to prior work in that we use online computation of the optimal policy based on measurements of the offered traffic, and that we consider an entire network in addition to the much simpler case of a single link. Furthermore, our semi-Markov decision process (SMDP) model uses a novel reward function that addresses a limitation of the one considered in [4]. Thus, our contribution is to combine variations of prior theoretical results into a concrete system where performance management is partially automated in a novel way, and to compare the performance of this system against the less sophisticated approach used in existing systems.

II. Background and Related Work

The topic of admission control has been studied in a variety of QoS scenarios. The problem is simplified in circuit-switched networks, where one does not need to contend with the complexities of user mobility or variable bit rate connections. Instead, the success of an incoming call or connection is assured at admission time through hard reservation of bandwidth on an end-to-end basis. The problem of finding the optimal admission policy can then be formulated as a stochastic knapsack problem, where a link with capacity F is used to service K classes of calls with bandwidths $b_1, \ldots, b_K$, arriving according to independent Poisson processes with rates $\lambda_1, \ldots, \lambda_K$, and departing according to independent general distributions with means $1/\mu_1, \ldots, 1/\mu_K$ (i.e. with departure rates $\mu_1, \ldots, \mu_K$). In our context, a call corresponds to a request to create a partition of a certain size in a parent lightpath in support of an end-to-end lightpath.

Using the above formulation, several useful results are presented in [5], [6], [4], [7]. First, for the special class of coordinate-convex policies, the equilibrium behaviour can be analytically derived and follows a simple form. Let a state of the system be defined by a vector $n = (n_1, \ldots, n_K)$ where n_i is the number of calls of type i in progress. Then, a coordinate-convex policy is a set of states Ω such that two conditions are satisfied. First, for all k, a class k call is accepted in state n if and only if $n + e_k \in \Omega$, where e_k is a unit vector with a one in the k'th position. Second, if $n \in \Omega$ and $n_k > 0$, then $n - e_k \in \Omega$ as well. Given such a policy, the equilibrium probability of being in state n is

$$P_\Omega(n) = \frac{\prod_{k=1}^{K} q_k(n_k)}{G(\Omega)}$$

where

$$q_k(n_k) = \frac{1}{n_k!} \left(\frac{\lambda_k}{\mu_k} \right)^{n_k}$$

and

$$G(\Omega) = \sum_{n \in \Omega} \prod_{k=1}^{K} q_k(n_k). \tag{1}$$

Note that the formula holds for general holding time distributions [4], [5].

Two special members of the family of coordinate-convex policies are the complete sharing (CS) and complete partitioning (CP) policies [5]. In complete sharing, calls are accepted whenever there is sufficient spare bandwidth. We shall subsequently also refer to this as the greedy policy. In complete partitioning, each class of calls is allotted a fixed amount of bandwidth and no sharing of bandwidth occurs between classes. Complete partitioning is optimal in the limit as arrival rates approach infinity. Complete sharing, on the other hand, is sensible when arrival rates approach zero. The optimal complete partitioning policy can be obtained in $\mathcal{O}(F^2 K)$ time using dynamic programming, where F is the link capacity.

In the general case, the optimal admission policy is difficult to compute, and need not be coordinate-convex. For the special case of $K = 2$, the optimal coordinate-convex policy is of threshold type, where calls of one class are only admitted if the number of calls in progress of the same class is sufficiently low [5]. When call holding times follow exponential distributions, a semi-Markov decision process (SMDP) formulation can be applied and the optimal policy obtained using linear programming or policy-value iteration [4], [7]. A state in the SMDP formulation of [7] corresponds to the arrival or departure of a call. Given K call types, a state corresponding to a call arrival event is defined by a vector $(n_1, \ldots, n_K, b_k)$ where n_i represents the number of calls of type i in progress, and b_k is the size of the incoming call of type k. The corresponding action space has two elements, representing call acceptance and rejection decisions. A general reward r_k is considered upon acceptance of a call of type k. The approach of [4] is similar, but incorporates the size of an arriving call into the action space instead of the state space. Thus, the state vector is of the form $(n_1, \ldots, n_K)$, and the action space is a subset of $\{b_1, \ldots, b_K\} \times \{\text{true}, \text{false}\}$. Another difference compared to the formulation of [7] is that arrivals of rejected calls are not considered as events. Consequently, a state never makes a transition back to itself.

III. The Greedy Policy is Not Always Optimal

In this section, we describe a scenario under which the greedy admission control policy is suboptimal with respect to resource utilization. Consider an initially idle parent lightpath that is subjected to a sequence of partition allocation requests. Without loss of generality, let the bandwidth of the parent lightpath be $N > 1$ units. Moreover, consider the special case where partitions are requested in sizes of either one unit, or N units. Following the nomenclature of SONET, we shall discuss the parent lightpath as being composed of N channels. For example, if $N = 8$ then this corresponds to an STS-192 SONET circuit that can be partitioned into eight STS-24 children, each capable of carrying Gigabit Ethernet traffic. Finally, suppose that requests for creation of partitions of size N and 1 arrive according to independent Poisson processes with rates $\lambda_N > 0$ and $\lambda_1 > 0$. Similarly, request holding times for the two partition sizes follow independent general distributions with means $1/\mu_N$ and $1/\mu_1$, where $\mu_N, \mu_1 > 0$.

First, let us consider the N-only policy, which accepts all feasible requests for size N partitions and rejects all others.

This is a coordinate-convex policy, and applying equation 1 the utilization ratio is

$$\frac{\frac{\lambda_N}{\mu_N}}{1+\frac{\lambda_N}{\mu_N}}. \tag{2}$$

Similarly, the utilization ratio of the greedy policy is

$$\frac{\frac{\lambda_N}{\mu_N}+\sum_{k=1}^{N}\frac{k}{N}\left(\frac{1}{k!}\right)\left(\frac{\lambda_1}{\mu_1}\right)^k}{1+\frac{\lambda_N}{\mu_N}+\sum_{k=1}^{N}\frac{1}{k!}\left(\frac{\lambda_1}{\mu_1}\right)^k}. \tag{3}$$

Taking $\alpha = \frac{\lambda_N}{\mu_N}$, $\beta = \sum_{k=1}^{N}\frac{k}{N}\left(\frac{1}{k!}\right)\left(\frac{\lambda_1}{\mu_1}\right)^k$, and $\gamma = \sum_{k=1}^{N}\frac{1}{k!}\left(\frac{\lambda_1}{\mu_1}\right)^k$, the above ratios can be rewritten as $\frac{\alpha}{1+\alpha}$ and $\frac{\alpha+\beta}{1+\alpha+\gamma}$. Subtracting the two expressions we get

$$\begin{aligned}\frac{\alpha+\beta}{1+\alpha+\gamma}-\frac{\alpha}{1+\alpha} &= \frac{\alpha+\beta+\alpha^2+\alpha\beta-\alpha-\alpha^2-\alpha\gamma}{(1+\alpha+\gamma)(1+\alpha)}\\ &= \frac{\beta+\alpha(\beta-\gamma)}{(1+\alpha+\gamma)(1+\alpha)},\end{aligned} \tag{4}$$

where $\beta < \gamma$. Since the denominator is positive, for any combination of β and γ the entire expression is negative when α is sufficiently large. Thus, the N-only policy outperforms the greedy policy provided that requests for partitions of size N arrive at a sufficiently high rate relative to the corresponding departure rate.

Another result that follows from the above analysis is that the optimal coordinate-convex policy is greedy with respect to requests for partitions of size N. To prove this, it suffices to show that

$$\frac{\alpha+\beta'}{1+\alpha+\gamma'}-\frac{\beta'}{1+\gamma'} > 0 \tag{5}$$

where α is as defined above and $0 < \beta' < \gamma'$. Here $\frac{\beta'}{1+\gamma'}$ represents the expected utilization ratio of any policy that rejects all requests for size N partitions (i.e., is coordinate-convex but is not greedy with respect to such requests). Similarly, $\frac{\alpha+\beta'}{1+\alpha+\gamma'}$ represents the expected utilization ratio of the same policy modified to accept requests of size N whenever they are feasible. Simplifying the left side of equation 5 we arrive at

$$\frac{\alpha(1+(\gamma'-\beta'))}{(1+\alpha+\gamma')(1+\gamma')} \tag{6}$$

which is positive since $\beta' < \gamma'$ and since all other terms are positive.

The above results can be extended to the general case of $K > 1$ partition sizes provided that the bandwidth of the parent lightpath is one of the valid partition sizes. Let $\hat{n}$ be the state corresponding to one such partition being occupied, and let Ω be the greedy policy. Then, applying equation 1 (recall the definition of $q_k(n_k)$) the expected utilization is

$$\frac{\alpha+\beta}{1+\alpha+\gamma} \tag{7}$$

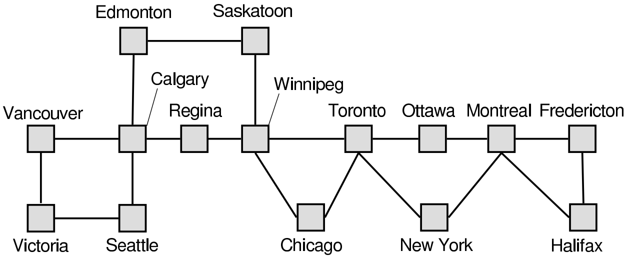

Fig. 1. Physical network topology used in our simulation environment.

where

$$\begin{aligned}\alpha &= \frac{\lambda_N}{\mu_N} && (8)\\ \beta &= \sum_{n\in\Omega, n\neq\hat{n}}\left(\frac{1}{N}\sum_{k=1}^{K}n_k b_k\right)\left(\prod_{k=1}^{K}q_k(n_k)\right) && (9)\\ \gamma &= \sum_{n\in\Omega, n\neq\hat{n}}\left(\prod_{k=1}^{K}q_k(n_k)\right). && (10)\end{aligned}$$

Once again we have $\beta < \gamma$, and it follows that the N-only policy outperforms the greedy policy when α is sufficiently high. Similarly, removing α makes the utilization ratio lower, which implies that the optimal coordinate-convex policy in the general case is greedy with respect to size N partitions.

IV. Dynamic Policy Optimization

In this section we build on the results of the last section by considering online computation of the optimal partitioning request admission policy in order to improve utilization. We consider two techniques of policy computation, applied at the level of individual link bundles (i.e. sets of lightpaths between physically adjacent nodes), and compare their performance through simulation. Our results are novel with respect to prior work in that we use online computation of the optimal policy based on measurements of the offered traffic, and that in our simulation environment we consider an entire network in addition to the much simpler case of a single link. Furthermore, our SMDP model uses an enhanced reward function that addresses a limitation of the one considered in [4].

A. Simulation Environment

1) Network Model: We consider a network based on the CA*net4 physical topology, presented in Figure 1. In our simulations, each edge in the physical topology corresponds to a pair of bidirectional OC-192 links. Although CA*net4 currently uses single OC-192 links, we believe that the presence of multiple parallel links is becoming an immediate concern due to the growing popularity of WDM technology. We consider that lightpaths are automatically partitioned in order to service end-to-end lightpath establishment requests, following the design of the University of Waterloo UCLP software [8]. In our simulation environment we assume that contiguous concatenation is used and that STS-24 and STS-192 are the circuit sizes used to carry Gigabit and 10 Gigabit

Ethernet traffic, respectively. Hardware constraints on starting positions of these circuit types are modeled after the Cisco ONS 15454 and Nortel OME 6500 platforms, which in our simulation environment behave identically.

In the establishment of end-to-end lightpaths, shortest path routing is used over a logical topology where each lightpath or partition thereof is considered a logical link and is represented by a graph edge. Note that each physical link may translate into multiple logical links of varying sizes depending on how it was partitioned. Path length is defined in terms of the number of physical hops, and the path computation is applied on the subset of logical links having sufficient free bandwidth to support the amount requested by the user. Ties among shortest paths are broken by using a sophisticated weight function as proposed in [9]. Specifically, each lightpath l is weighted according to

$$W(l) = 1 + \frac{B(l) - \text{bandwidth requirement}}{|V| \max_{k \in E} B(k)},$$

where $B(l)$ is the bandwidth of l, V is the set of cross-connect devices (vertices), and E is the set of lightpaths (edges).

2) Traffic Generation: We consider that the network is subjected to end-to-end lightpath establishment requests in response to data transfer operations made using the appropriate software. Due to the heavy-weight nature of lightpath establishment, we suppose that the set of files to be transferred in a single session is prepared in advance and that a single lightpath is accordingly allocated per session. We consider that sessions are human-initiated and arrive according to a Poisson process with rate λ (in units of requests/hour), which has been found to be an accurate model in [10], [11], [12]. Similarly, we model session size using a log-normal distribution, following the findings of [11] based on analysis of FTP traffic. In our experiments, we use a fixed mean session size of 10 TB with a standard deviation of 10 TB. Due to lack of concrete operational data, these parameters were selected based on anecdotal reports of large data transfer activities [13], [14], [15]. We also note that the offered load is determined by the combination of arrival rate and session size, so having fixed the distribution of the latter we can generate a variety of load conditions by varying the former.

Once the arrival time and session size Z of a lightpath establishment request are determined, we compute the remaining parameters as follows. First, a node pair is selected based on a relative traffic demand matrix, which is generated by filling 0.6 of the entries with uniformly random values on the interval [0.2, 1.2], and setting the remaining entries to zero, as in [16]. Next, the capacity of the equipment used to access the lightpath is determined by assuming that a proportion p of users are equipped with Gigabit hardware and the remaining $1 - p$ use 10 Gigabit hardware. For simplicity, we do this independently of Z, although in reality users with 10 Gigabit equipment may be more likely to initiate larger data transfers. The choice of Gigabit and 10 Gigabit capacities is motivated by the popularity of the corresponding Ethernet technologies, which are commonly used in conjunction with SONET encapsulation.

The requested bandwidth B in an end-to-end lightpath establishment request is taken to be the minimum of the randomly-determined capacities supported by the two end users. Then, the corresponding session holding time is taken to be Z/B. For simplicity, we consider here that the throughput during the data transfer is B, which corresponds to circuit sizes of STS-24 in the Gigabit case and STS-192 in the 10 Gigabit case. In reality, however, one must consider protocol overhead in all relevant network layers, the impact of congestion control [17], and disk performance [15]. For example, throughput can be limited to less than 90% of the raw bit rate of a circuit when using TCP/IP traffic with Ethernet over SONET technology [3]. The rationale behind our simplifying assumption is that throughput is a large and approximately constant fraction of bandwidth, hence relative performance numbers (i.e., optimized v.s. unoptimized network) remain valid.

The initial state of the simulation is an idle network. The simulation period is 300 days. We measure the end-to-end lightpath establishment request acceptance ratio, subsequently referred to simply as the request acceptance ratio, by averaging over ten simulation runs, each run corresponding to a randomly generated traffic demand matrix. Averaging over ten repetitions yields a standard error of the mean of approximately 1%. We vary the request arrival rate λ and the proportion p of users having Gigabit circuit access hardware in order to evaluate the relative performance of the admission control policies under consideration over a range of operating conditions.

B. Policy Optimization Scheme

In the interest of automation and flexibility, we base our optimization of the partitioning request admission policy on online measurements of offered traffic rather than relying on an externally-specified traffic demand matrix. To this end, we divide each 300-day simulation run into a series of ten thirty-day training periods, between which we recompute the policy for each link bundle (equivalently each physically adjacent node pair) based on traffic measurements from the last training period. In addition to the fifteen-node network topology from Figure 1, we shall also consider a two-node network connected by a single bundle of two OC-192 links.

Measuring offered traffic online is challenging in our simulation environment due to the use of online route computation, whereby the shortest path is chosen subject to resource availability. Since lightpaths that do not have sufficient spare capacity are filtered out when an end-to-end lightpath establishment request is serviced, only feasible partition allocation requests are offered to the selected subset of lightpaths. Thus, the set of partitioning requests seen by a link bundle is not a direct basis for measuring the amount of offered traffic as it does not count rejected requests. At the same time, it does not make sense to count requests at the filtering stage since all lightpaths are checked for resource availability but only some lie on a shortest path, and no traffic is actually offered to the others. Instead, we estimate the effective rate at which requests for partitions of a particular size arrive at a particular link bundle by tracking the number M of feasible requests seen over a training period, and the total time T spent in states where such requests were feasible. Then, M/T estimates the number of partitioning requests that would have been seen had the link bundle been sufficiently underutilized during the entire training period, and is taken to be the effective arrival rate of requests.

Note that by nature of equation 1, if requests for a certain partition size are not seen at all during a training period, then no bandwidth is assigned to that partition size in the event that a complete partitioning policy is adopted. In fact, we enforce similar behaviour in all policies, meaning that requests for partitions of a size unseen during training are always rejected. This feature promotes stability in our policy optimization scheme, which is a type of feedback mechanism since measurements of policy-influenced traffic patterns are used as input to the policy optimizer. Without some form of stabilization, oscillations may occur between the greedy policy and a CP policy. Consider what happens when we begin with the greedy policy, and a policy that rejects all STS-24 requests is adopted after some training period. In the next training period, only STS-192 requests are seen, and consequently the greedy policy is optimal with respect to the measured request arrival rates. However, if a truly greedy policy is chosen, then in the next training period STS-24 requests are admitted once again, and we return to a state similar to two training periods earlier.

C. Policy Computation

We consider three admission control policies in our performance comparison, subsequently denoted *greedy*, *greedy-CP*, and *SMDP*. Each policy is applied at the level of a link bundle joining a pair of nodes. Greedy is the policy that accepts partition allocation requests whenever there is sufficient idle bandwidth. This policy corresponds to the current state of the art in user-controlled lightpath management systems, except for the novel stabilization mechanism described at the end of Section IV-B, which in fact applies to all three policies under consideration. The other two policies introduce more sophisticated admission control. The hypothesis behind our performance evaluation is that these policies will lead to a performance gain over the greedy policy by managing bandwidth in a way that maximizes long-term reward.

Greedy-CP is the optimal policy chosen from the subset of coordinate-convex policies consisting of the greedy policy and the set of complete partitioning (CP) policies. To compute greedy-CP, we first determine the optimal CP policy using the dynamic programming technique of [5], and then compare the expected utilization of this policy against the greedy policy using equation 1. Thus, greedy-CP represents an inexpensive approximation of the optimal coordinate convex policy by choosing the best among previously studied special cases.

Finally, SMDP refers to the optimal policy approximated using policy-value iteration applied to a novel variant of the semi-Markov decision process formulation of [4]. The novelty here consists of a modified reward function that vastly improves performance in our simulation environment, as discussed in Section IV-D. SMDP is the most costly policy to compute, but at the same time is the closest approximation of the optimal policy under certain statistical assumptions on the pattern of network traffic. We discuss the assumptions behind and computation of greedy-CP and SMDP in more detail below.

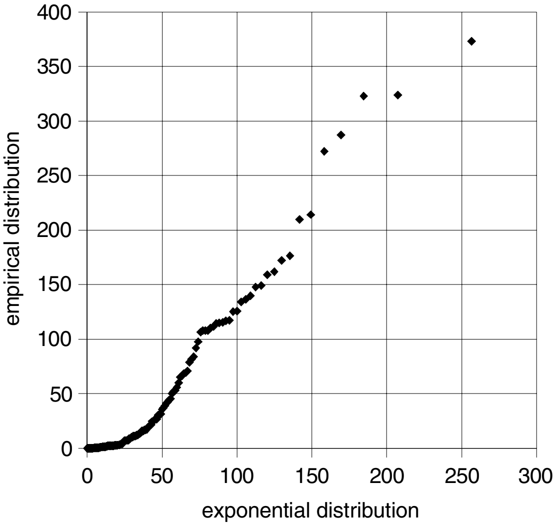

Fig. 2. Quantiles of the empirical distribution of STS-24 request inter-arrival times (Montreal-Halifax link, $p = 0.9$, $\lambda = 5$/h) v.s. quantiles of an exponential distribution.

The computation of the greedy-CP policy is based on predicting the expected utilization ratio using equation 1, which is correct for any coordinate-convex policy under the assumption of Poisson arrivals and arbitrary holding times (see Section II). We consider that this technique is well-suited to our simulation environment, where holding times do not follow the exponential distribution frequently assumed in analytic work. However, the assumption of Poisson arrivals may not necessarily hold at the level of a link bundle, even though end-to-end lightpath establishment requests are generated this way at network level.

In order to test the validity of the Poisson arrival assumption on partitioning requests, we performed an experiment whereby the empirical distribution of request inter-arrival times was measured. Specifically, we modified the simulation environment by replacing the link bundle between one pair of nodes with a dummy link of very high capacity, and recorded the arrival times of partitioning requests issued to that link over a 1 000 day period. Figure 2 and Figure 3 demonstrate the correspondence between the empirical distribution of request inter-arrival times with the exponential distribution having the same mean as the empirical distribution. The results in these figures are representative of a broader series of runs where measurements were taken from the Montreal-Halifax and Winnipeg-Toronto links for parameter value combinations $(p, \lambda) \in \{0.01, 0.1, 0.9\} \times \{0.5/h, 2/h, 10/h, 20/h\}$. The somewhat-linear pattern indicates reasonable correspondence between the empirical and exponential distributions of request inter-arrival times.

The SMDP policy is computed using a technique that assumes exponential request holding times, in contrast to the theory related to coordinate-convex policies. Although this assumption is not justified in our simulation environment due to the use of log-normal holding times, we consider the SMDP formulation in an effort to approximate the optimal policy.

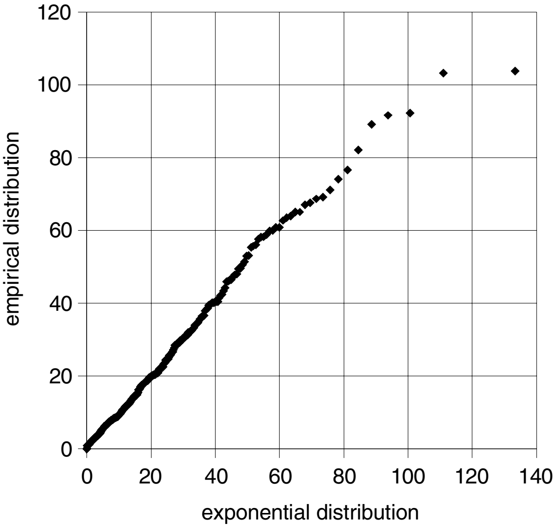

Fig. 3. Quantiles of the empirical distribution of STS-192 request inter-arrival times (Winnipeg-Toronto link, $p = 0.1$, $\lambda = 10$/h) v.s. quantiles of an exponential distribution.

Despite its rigidity, the SMDP formulation may still yield policies more optimal than the greedy policy and greedy-CP policies, because it is not restricted to coordinate-convex policies.

We adopt the compact SMDP formulation of [4] (see Section II), but instead of using linear programming we apply the simpler policy-value iteration technique to approximate the optimal policy. We use a discounting factor of 0.99, and we stop iteration when the difference between consecutive expected reward values in all states is less than 10^{-10} of the maximum value in any state, corresponding to between 1 000 and 10 000 iterations. In order to maximize network utilization, it is natural to consider that the rate $R(s)$ of reward generation in a state s is the total amount of bandwidth allocated. If $T(s)$ is the expected duration of s, then it makes sense to define the reward accumulated in s as the weighted rate of reward generation $R(s)T(s)$, which is the idea used in the linear programming formulation of [4]. However, a problem with this approach is that $T(s)$ is dependent on the policy (since rejected requests do not translate into state transitions), and in the idle state s_0 where $R(s_0) = 0$, there is no incentive to select the policy that minimizes $T(s_0)$. At the same time, s_0 is a frequently visited and consequently important state in our simulation environment, due to the coarse granularity of partitions considered.

In order to address the above issue, we propose the novel approach of fixing $R(s) = -1$ for all states s. In that case, the reward of being in state s is $-T(s)$, and the policy-value iteration procedure selects the policy that minimizes the expected amount of time required to make a constant number of state transitions. Since transitions in our SMDP formulation correspond only to admission and departure of feasible requests, this equivalently minimizes the amount of time taken to accept a constant number of requests. Finally, since the expected volume of traffic associated with each accepted request is constant and independent of the bandwidth, our reward definition maximizes the rate at which traffic is admitted, and hence maximizes utilization.

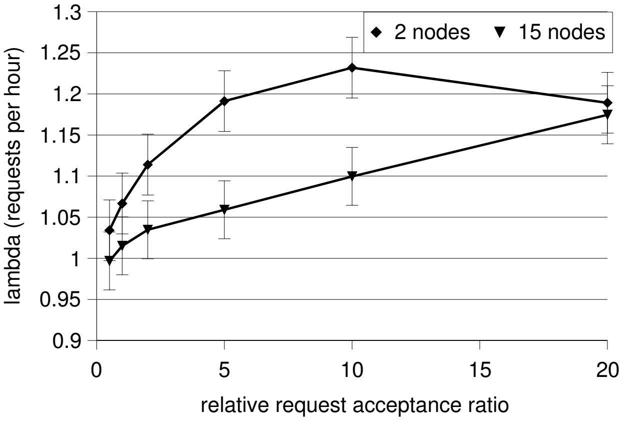

Fig. 4. Performance gain using greedy-CP policy, $p = 0.01$.

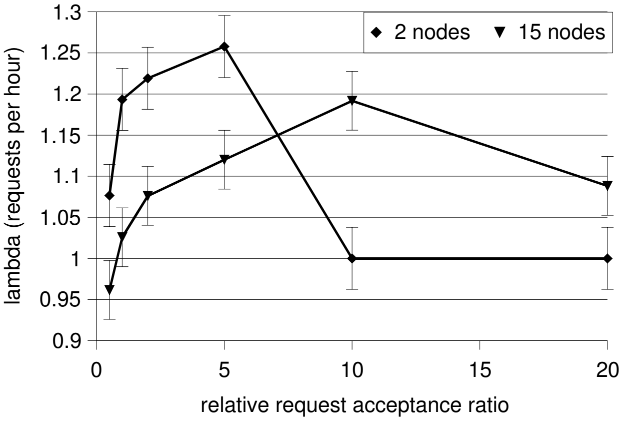

Fig. 5. Performance gain using greedy-CP policy, $p = 0.05$.

D. Performance Evaluation and Analysis

Simulation results for the greedy-CP policy on the two-node and fifteen-node network shown in Figure 1 are presented in Figures 4–6 and Figures 8–9, corresponding (respectively) to the following values of p: 0.01, 0.05, 0.1, 0.5 and 0.9. The parameter λ (requests per hour) was varied from 0.5 to 20, corresponding to a lightly loaded and heavily loaded network, respectively. Each point plotted represents an average over ten simulation runs of the performance gain associated with using the greedy-CP policy versus the greedy policy. The standard error of the mean in the measurements is no more than 1%. The error bars plotted represent three standard errors of the mean.

Analogous results were collected for the SMDP policy. The performance difference between the greedy-CP and SMDP policies is less than 0.1% in 90% of test cases. Larger differences occur only when $p = 0.1$, and the corresponding numbers are shown in Figure 7. Even in those cases the differences are relatively small (i.e., less than three standard errors of the mean). Consequently, unless stated otherwise, subsequent observations apply to both the greedy-CP and SMDP policies. For comparison, the policies obtained using the SMDP approach behave very poorly if we use the reward function of [4] instead of our novel reward function described

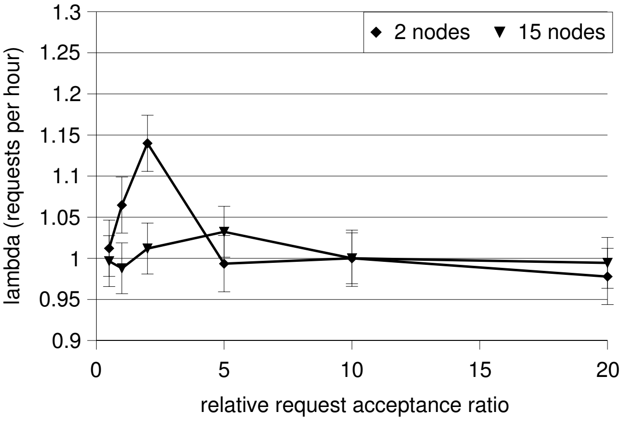

Fig. 6. Performance gain using greedy-CP policy, $p = 0.1$.

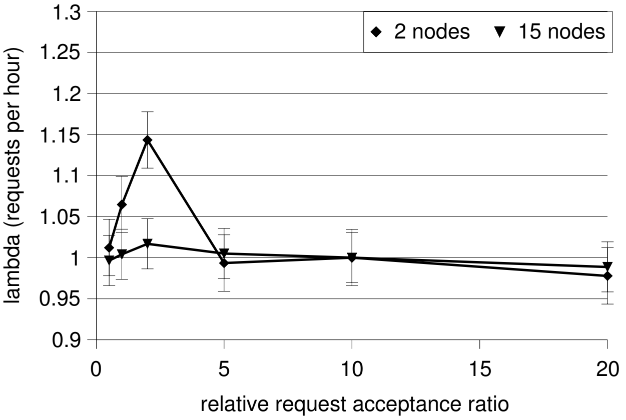

Fig. 7. Performance gain using SMDP policy, $p = 0.1$.

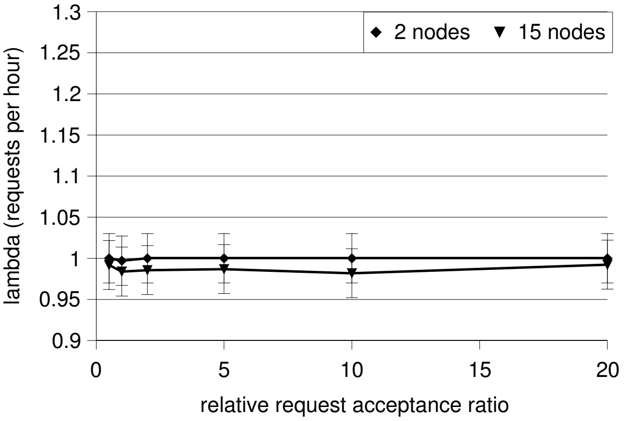

Fig. 8. Performance gain using greedy-CP policy, $p = 0.5$.

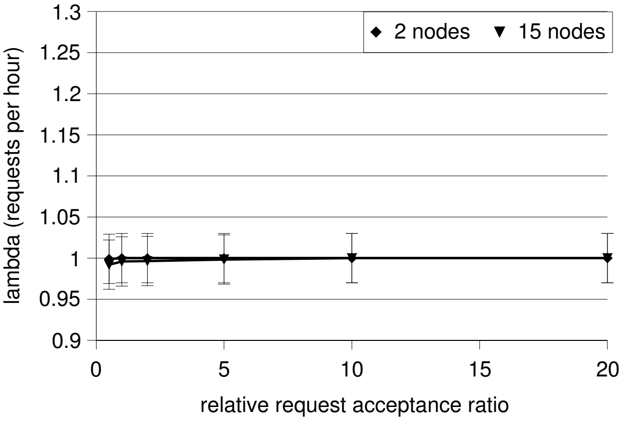

Fig. 9. Performance gain using greedy-CP policy, $p = 0.9$.

earlier. Specifically, performance levels 50% below the greedy policy are common, due to the "optimal" policy unwisely rejecting STS-192 (i.e., 10 Gb) requests in the idle state (see discussion in Section III).

Figures 4–6 exhibit a significant performance peak with respect to λ, the location of which moves to the left (i.e. toward smaller λ) as p increases. This is expected since greedy-CP is most effective relative to the greedy policy when requests for STS-24 partitions arrive at a particular link bundle at such a rate compared to STS-192 requests that poor utilization occurs and STS-192 requests are frequently rejected for lack of free bandwidth. In contrast, for small and large λ, greedy performs as well as greedy-CP since either STS-192 requests are rarely rejected, or they are sometimes rejected but utilization is high nevertheless. Thus, the performance of greedy-CP relative to greedy peaks under conditions when greedy accepts requests for small partitions that give small short-term rewards but preclude larger future rewards. This scenario is approximately characterized by a constant value of the product $p \times \lambda$, which occurs for smaller λ as p tends to unity, as observed. The exact value of λ at which the performance peak occurs also depends on the size of the network, since the greater the number of links the fewer requests for lightpath partitions arrive on average at a particular link. Consequently, the performance peak always occurs at a smaller value of λ in the two-node case than in the fifteen-node case. Finally, Figures 8–9 show that the greedy-CP policy does not perform better than the greedy policy for $p \in \{0.5, 0.9\}$ and the values of λ under consideration.

Performance peaks of the magnitude observed in Figures 4–6 can be predicted in the two-node case by applying equation 1 to compare the performance of the greedy policy and a policy that always rejects STS-24 requests and accepts STS-192 requests, the latter policy being one of the candidates from which greedy-CP is chosen. Note that despite the fact that equation 1 yields the expected utilization ratio, we are able to use it to predict the gain in the request acceptance ratio, on which Figures 4–9 are based; this is because the volume of data carried by a lightpath is chosen at random and independently of the lightpath's bandwidth in our simulation environment, and so the average utilization is proportional to the request acceptance ratio.

As noted above, there are significant performance gains associated with using greedy-CP or SMDP instead of the greedy policy in both the two-node and fifteen-node case when $p \leq 0.1$. The largest gains observed in our performance evaluation occur when $p = 0.05$, corresponding to 26% and 19% in the two-node and fifteen-node cases, respectively. The performance gains for the simple two-node network are generally larger than for the fifteen-node network. This behaviour is expected since in the latter case each pair of physically adjacent nodes performs a local optimization effort on the corresponding link bundle, without any coordination with neighbouring nodes, whereas each end-to-end lightpath is typically routed through a series of such link bundles. Still, we observe a significant (3% or more) performance gain in 33% of the test cases with the fifteen-node network, compared to 40% with the two-node network. A significant performance loss is observed only on the fifteen-node network, and only in one test case out of thirty, namely when $p = 0.05$ and $\lambda = 0.5$ (see Figure 5). The magnitude of the performance loss is only 4%. We discuss this case in more detail later.

In the two-node scenario with the greedy-CP policy, observed utilization levels differ by up to 20% from the expected values predicted by equation 1, but the differences are symmetrically distributed around zero. Such behaviour is expected since in the two-node case our simulation environment meets the statistical assumptions of the theory behind equation 1, namely Poisson arrivals and holding times following an arbitrary distribution with a well-defined mean. Surprisingly, the same correspondence is seen between observed utilization levels and the optimal levels predicted using the SMDP approach, despite the fact that the assumption of exponential holding times does not hold. Moreover, the utilization levels predicted using equation 1 and using the SDMP approach agree to within 0.1%. Based on this observation, we conclude that the performance of the SMDP approach in our simulation environment is not significantly impaired by the incorrect assumption of exponential holding times.

The optimal policies obtained using greedy-CP and SMDP are typically in agreement. In test cases exhibiting the highest performance gain due to policy optimization, both approaches select a policy that rejects all STS-24 requests, which is consistent with the intuition arising from the analytic results presented in earlier sections. Discrepancies between the two policy computation approaches occur on occasion, but only when the predicted performance difference between optimal and greedy policies is small. For example, if the greedy policy is expected to perform slightly better than the optimal complete partitioning policy, then the SMDP approach may select a policy that rejects STS-24 requests in the state where a single STS-192 partition is allocated on a link bundle, and rejects no other feasible requests. However, the predicted and observed utilization ratios are again only slightly better than with the greedy policy, and no significant performance gain is observed. Thus, we conclude that the broader policy space considered in the SMDP approach does not yield practical benefits in our simulation environment.

Our policy optimization system appears stable in the sense that performance levels settle after a small number of observation periods. An example of this is shown in Figure 10, where the request acceptance ratio stabilizes after approximately five training periods. Such behaviour is expected due to the design of our simulation environment, which was discussed in Section IV-A. Specifically, the fact that each policy rejects requests not seen in the previous training period has a tendency to maintain complete partitioning policies after they are first adopted, and consequently to stabilize the evolution of performance levels in time.

Upon closer investigation of the single test case leading to a performance loss, we find that it is not an anomaly. Instead, suboptimal policies were chosen due to random deviation in the measured request arrival rates from the expected arrival rates. In some training periods, only STS-192 requests were witnessed by some link bundles. In that case, the greedy policy is equivalent to the complete partitioning policy that devotes all the bandwidth to STS-192 partitions, and equation 1 in theory predicts the same utilization level for these two policies. In the simulation, however, we find that a slightly higher utilization is predicted for the CP policy due to rounding and representation error in the corresponding floating point calculations, and as a consequence the CP policy is adopted, which turns out to be suboptimal in the long-run.

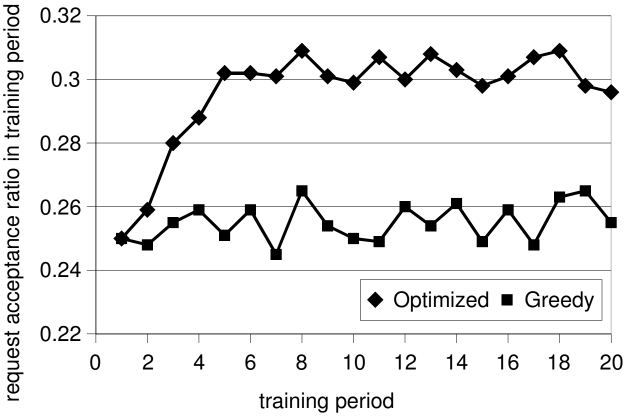

Fig. 10. Stabilization of request acceptance ratio, $\lambda = 10$, $p = 0.05$, 15-node network from Figure 1.

The small numerical errors described above are a potential problem in our performance optimization scheme due to the presence of the same feature that gives it stability. Specifically, when numerical error causes the CP policy to appear more optimal than the greedy policy, it causes a permanent transition into a CP policy. However, the decision itself is based on a crude estimate of the offered traffic pattern, and may not be optimal in the long-run. Thus, when in doubt, it is safer to adopt the greedy policy rather than committing to a CP policy. We considered realizing such a heuristic in our policy optimizer by applying a threshold to the policy selection stage, whereby the CP policy is only chosen if its predicted utilization is better by a factor z than the utilization under the greedy policy. However, with z between 1.01 and 1.1, the performance loss observed originally persists. As an alternative, we experimented with increasing the length of the training period from 30 days to 100 days, and were able to close the performance gap between the greedy policy and a dynamically optimized greedy-CP policy to less than 1% (i.e., one standard error of the mean) in the test case that originally led to a performance loss.

Another method of addressing the problem of permanent suboptimal policy selection decisions is to periodically return each link bundle to the initial greedy policy, and begin optimization anew. For example, in our simulation environment we measure the performance gain over ten training periods, starting with an idle network and with the greedy policy applied globally. Consequently, we would expect equally good performance results during continuous operation if the greedy policy were to be reapplied throughout the network after every ten training periods. Alternately, simulation can be used in parallel with the operation of the network in order to determine the optimal set of policies, and revert link bundles from a CP policy back to the greedy policy when this is deemed beneficial.

V. Summary and Future Work

In our investigation of lightpath establishment request admission policies we found that the greedy policy is suboptimal in some pertinent cases. We were able to demonstrate both analytically and by simulation that for lightpaths supporting multiple partition sizes, under some conditions it is beneficial to reject requests for small partitions even when resource availability permits their admission. In our experiments, we found that network-level performance can be improved by up to 19% through dynamic optimization of the admission policy based on online traffic measurements, without sacrificing stability.

We envision further work on the topic of optimal non-greedy policies, particularly design and performance evaluation of dynamic policy optimization schemes. In addition, one can consider more sophisticated methods of inferring properties of the traffic pattern from empirical data, and of policy selection. In the latter case, it may be possible to improve on our dynamic scheme, which represents a series of local optimization efforts, by considering a greater scope of interactions between node pairs. Other open research problems include analyzing the fairness and stability of dynamic policy optimization schemes, as well as their robustness against deviations from assumptions such as Poisson request arrivals.

Acknowledgement

Sincere thanks to Youssef Iraqi and Majid Ghaderi for enlightening discussions on the topics of end-to-end lightpath establishment and stochastic models. We are also grateful to the anonymous reviewers for their insightful comments.

References

[1] B. St. Arnaud, “Frequently asked questions about customer owned dark fiber, condominium fiber, community and municipal fiber networks,” Mar. 2002. [Online]. Available: http://www.canarie.ca/canet4/library/customer/frequentlyaskedquestionsaboutdarkfiber.pdf

[2] B. St. Arnaud and J. Wu, “Customer-controlled and -managed optical networks,” *J. Lightwave Technol.*, vol. 21, no. 11, pp. 2804–2810, Nov. 2003.

[3] R. Boutaba, W. Golab, Y. Iraqi, T. Li, and B. S. Arnaud, “Grid-controlled lightpaths for high performance grid applications,” *J. Grid Computing*, vol. 1, no. 4, pp. 387–394, 2003.

[4] K. Ross and D. Tsang, “Optimal circuit access policies in an ISDN environment: a Markov decision approach,” *IEEE Trans. Commun.*, vol. 37, no. 9, pp. 934–939, 1989.

[5] K. Ross and D. Tsang, “The stochastic knapsack problem,” *IEEE/ACM Trans. Netwo.*, vol. 37, no. 7, pp. 740–747, 1989.

[6] K. Ross and D. Yao, “Monotonicity properties for the stochastic knapsack,” *IEEE Trans. Inf. Theory*, vol. 36, no. 5, pp. 1173–1179, 1990.

[7] E. Altman, T. Jiamenez, and G. Koole, “On optimal call admission control in a resource-sharing system,” *IEEE Trans. Commun.*, vol. 49, no. 9, pp. 1659–1668, 2001.

[8] R. Boutaba, W. Golab, Y. Iraqi, and B. S. Arnaud, “Lightpaths on demand: a Web services based management system,” *IEEE Commun. Mag.*, vol. 42, no. 7, pp. 101–107, July 2004.

[9] W. Golab and R. Boutaba, “Resource allocation in user-controlled circuit-switched optical networks,” in *Proc. IEEE HSNMC*, June 2004, pp. 776–787.

[10] V. Paxson and S. Floyd, “Wide-area traffic: The failure of Poisson modeling,” *IEEE/ACM Trans. Netw.*, vol. 3, no. 3, pp. 226–244, 1995.

[11] V. Paxson, “Empirically derived analytic models of wide-area TCP connections,” *IEEE/ACM Trans. Netw.*, vol. 2, no. 4, pp. 316–336, 1994.

[12] A. Feldmann, A. Gilbert, and T. Kurtz, “The changing nature of network traffic: Scaling phenomena,” *ACM SIGCOMM Computer Commun. Rev.*, vol. 28, no. 2, pp. 5–29, 1998.

[13] T. Koonen, H. de Waardt, J. Jennen, J. Verhoosel, D. Kand, M. de Vos, A. van Ardenne, and E. J. van Veldhuizen, “A very high capacity optical fibre network for large-scale antenna constellations: the RETINA project,” in *Proc. NOC*, June 2001, pp. 165–172.

[14] H. Newmann, M. Ellisman, and J. Orcutt, “Data-intensive e-science frontier research,” *Commun. of the ACM*, vol. 46, no. 11, pp. 68–77, 2003.

[15] B. Dobinson, R. Hatem, W. Hong, P. Golonka, C. Meirosu, E. Radius, and B. S. Arnaud, “Transatlantic native 10 Gigabit Ethernet experiments connecting Geneva to Ottawa,” in *Proc. IEEE HSNMC*, 2004.

[16] A. Gençata and B. Mukherjee, “Virtual-topology adaptation for WDM mesh networks under dynamic traffic,” in *Proc. IEEE INFOCOM*, vol. 1, 2002, pp. 48–56.

[17] C. Jin, D. Wei, S. Low, G. Buhrmaster, J. Bunn, D. Choe, R. Cottrell, J. Doyle, W. Feng, O. Martin, H. Newman, F. Paganini, S. Ravot, and S. Singh, “Fast TCP: From theory to experiments,” *IEEE Network*, vol. 19, no. 1, pp. 4–11, Jan/Feb 2005.

Wojciech Golab received an Honours Bachelor of Science degree in computer science from the University of Toronto in 2002. Upon graduation, he was awarded the Governor General’s silver academic medal. In 2004, he completed a Master of Math degree at the University of Waterloo, where he studied resource allocation problems in optical networks. Currently, he is a PhD student at the University of Toronto, studying the complexity of synchronization problems in shared memory multiprocessors.

Raouf Boutaba received the MSc. and PhD. Degrees in Computer Science from the University Pierre & Marie Curie, Paris, in 1990 and 1994 respectively.

He is currently a Professor of Computer Science at the University of Waterloo. His research interests include network, resource and service management in multimedia wired and wireless networks.

Dr. Boutaba is the founder and Editor-in-Chief of the IEEE Transactions on Network and Service Management and on the editorial boards of several other journals. He is currently a distinguished lecturer of the IEEE Communications Society, the chairman of the IEEE Technical Committee on Information Infrastructure and the IFIP Working Group 6.6 on Network and Distributed Systems Management. He has received several best paper awards and other recognitions such as the Premier’s research excellence award.