The OptIPuter: High-Performance, QoS-Guaranteed Network Service for Emerging E-Science Applications

Nut Taesombut[1], Frank Uyeda[1], Larry Smarr[1,2], Thomas A. DeFanti[4,2], Phil Papadopoulos[2,3], Jason Leigh[4], Mark Ellisman[5], John Orcutt[6] and Andrew A. Chien[1]
[1]Department of Computer Science and Engineering, University of California, San Diego (UCSD)
[2]California Institute for Telecommunications and Information Technology, UCSD
[3]San Diego Supercomputer Center, UCSD
[4]Electronic Visualization Laboratory, University of Illinois at Chicago
[5]National Center for Microscopy and Imaging Research, Center for Research in Biological Systems, UCSD
[6]Scripps Institution of Oceanography, UCSD

Abstract

Emerging large-scale scientific applications have a critical need for high bandwidth and predictable-performance network service. The OptIPuter project is pioneering a radical, new type of distributed application paradigm which exploits dedicated optical circuits to tightly couple geographically-dispersed resources. These private optical paths are set up on-demand, and combined with end resources to form a Distributed Virtual Computer (DVC). The DVC provides high-quality, dedicated network service to applications. In this article, we compare the OptIPuter's approach (DVC) which exploits network resources to deliver higher-quality network services with several alternative service models (intelligent network and asynchronous file transfer). Our simulations show that there are significant differences amongst the models in their utilization of resources and delivered application services. Key takeaways include that the OptIPuter approach provides applications with superior network service (as needed by emerging E-science applications and performance-critical distributed applications), at an expense in network resource consumption. The other approaches use fewer network resources, but provide lower quality application service.

I. Introduction

Emerging E-science posits wide-area scientific collaborations [1] with the ability to interactively share, process and visualize distributed data. Such collaborations are emerging in virtually every scientific front, including geosciences, biomedical informatics and nuclear physics, which aim to enhance understanding of complex systems. Typically, these applications require access to massive collections of distributed data objects (as large as several terabytes), which must be transferred with reliability and timeliness. To support these transfers, underlying networking infrastructures must deliver high quality of services, including extreme bandwidth (10's or even 100's Gbps) and controlled jitter/delay. These requirements, however, cannot be met by traditional shared, routed networks which offer only best-effort services.

Continuing advances in optical networking are producing networks with lower bandwidth-per-unit cost and predictable performance. Recently, Dense Wavelength Division Multiplexing (DWDM) has emerged as an efficient technique that increases dramatically the number of optical circuits that can be provided on a physical optical resource. DWDM enables each fiber to carry multiple wavelengths (or *lambdas*), increasing the aggregate throughput on each fiber to several terabits per second. Because each lambda is independent, network characteristics, such as bandwidth, jitter and delay, can be planned and controlled, enabling high-quality network service. Furthermore, recent advances in network control plane and middleware projects [2,3] are enabling dynamic (on-demand) provisioning of these lambdas (optical circuits). Dynamic provisioning not only allows applications to obtain dedicated use of high-speed optical circuits "on demand", but also enables efficient sharing of lambdas.

The emergence of configurable optical networks is paving a way for the next-generation E-science, moving from a network-constrained world into a network-rich world. With these dedicated high-speed connections, widely-dispersed resources, such as scientific instruments, federated data repositories, and computational resources, can be integrated tightly and virtually perceived as they are in a single machine room. This enables scientists from geographically distant institutions to collaboratively share, process and visualize their data and results, thereby improving the quality of scientific data processing analysis.

In the following, we describe two pioneering E-science research efforts that could benefit from configurable optical network technologies.

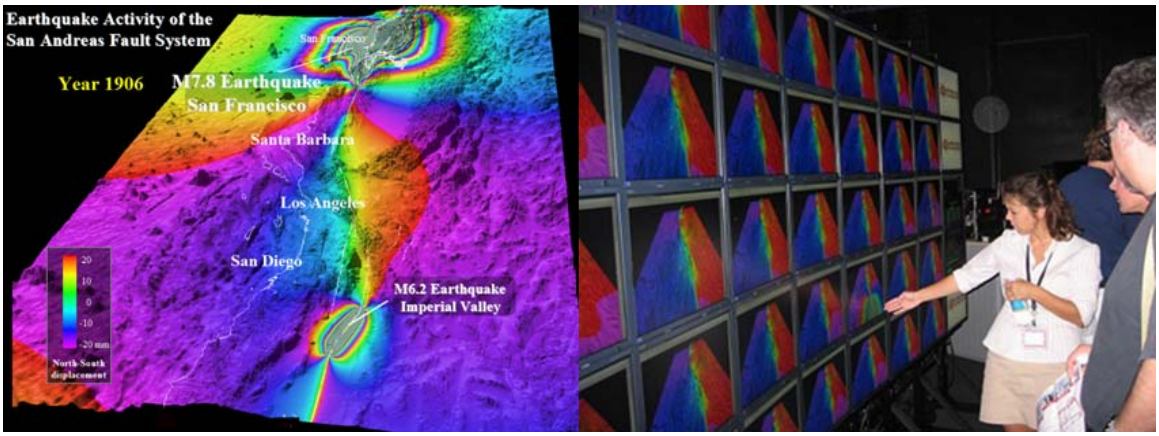


FIGURE 1. Scientists are studying the development of earthquakes in California using the 3D theoretical models of deformation along the San Andreas Fault from year 1901 to 2004 on the 100-megapixel tiled display (Images created by IGPP/SIO/UCSD)

EarthScope

The EarthScope (www.earthscope.org) is an initiative within the National Science Foundation (NSF) that aims at developing a national cyberinfrastructure to support the study of the structures and evolution of the Earth's crusts in North America. Originating as an effort to assess natural resources and mitigate risks from geological hazards, the project assembles geophysical data measurements from several distributed observing systems, including remote satellites and seismic sensing devices, and provides scientists from various disciplines access to these data for refined analysis.

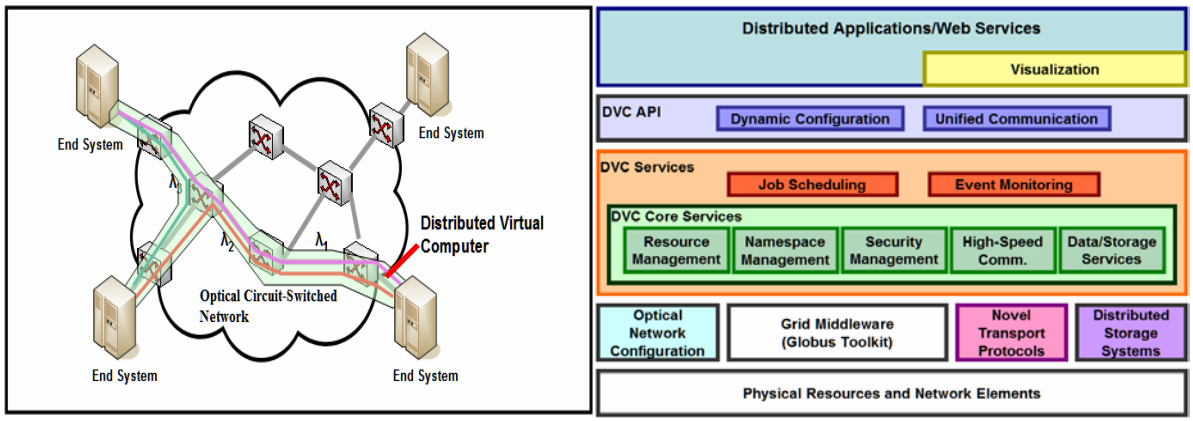
With technological advances in observational techniques and equipments, geophysical data is being collected at unprecedented high volume and quality. To enhance the study of complex geological systems, work is underway to develop tools that enable multi-dimensional visualization of these data which can be interactively explored and analyzed at high resolution. This ability enables scientists to examine a visualized geological sample at great details and in different dimensions. For example, researchers at the Scripps Institution of Oceanography (SIO) are using the datasets from EarthScope to study the activity of the San Andreas Fault in California. As shown in Figure 1, they use a visualization package called 'Fledermaus' (www.ivs3d.com) to visualize the 3D strain fields resulting from deformation along the San Andreas Fault and interactively explore them. Parallel visualizations of these datasets on a multi-tiled display, such as LambdaVision (www.evl.uic.edu/cavern/lambdavision), allow scientists to perform time-series analysis of the activity of earthquakes across years, thus enhancing understanding of their development.

Today, a major challenge facing EarthScope is a phenomenal amount of data being produced and collected. For example, EarthScope's modern digital seismic arrays can produce high-resolution images of North America's continental crust and its supporting layer, each of which can easily exceed 50GB [1]. These images can be produced on timescales of days to decades, thus producing the total seismic data being assembled per year to exceed 40TB. To support a time-series analysis of the deformation of the Earth's crust on a 55-panel tiled display (55 3D images), as shown in Figure 1, each display panel currently consumes 2GB of the datasets. In order to maintain an interactive environment, these datasets must be transferred within 5 seconds, thereby creating a bandwidth demand of 176Gbps. With today's networking infrastructures, it is not possible for scientists to transfer these data quickly to support real-time data analysis.

Biomedical Informatics Research Network (BIRN)

The BIRN (www.nbirn.net) is a National Institute of Health (NIH)-supported project to enable large-scale distributed collaborations for medical research in the neurosciences. The project is pioneering in the development of tools and infrastructures which will enable scientists at distant locations to seamlessly share, visualize and analyze multi-scale image data, behavioral data and genomic data in an innovative fashion.

The initial focus of the project is on brain mapping of human neurodegenerative diseases and associated mouse animal models. These datasets are generated by a variety of medical imaging tools such as MRI, light microscopy and high-energy electron microscopy, which can produce multi-scale, multi-dimensional images ranging from whole organs to cells to subcellular structures. The availability of images of a related specimen at multiple modalities allows scientists to visualize and examine a biological sample progressively at every structural level and thus permits refined analysis of its structures and their interrelations.



(a) **(b)**

FIGURE 2. (a) Distributed Virtual Computer (DVC) allocates a private network for the duration of application execution; (b) OptIPuter software architecture to enable DVC abstractions (shaded boxes indicate OptIPuter research areas)

To conduct multi-modality brain mapping experiments, scientists need a way to simultaneously visualize multiple high-resolution images. Currently, researchers at the Electronic Visualization Laboratory, University of Illinois at Chicago are developing a suite of tools [4] which allow simultaneous display of multiple 2D and 3D images and live video conferencing on an ultra-high-resolution multi-tiled display, such as LambdaVision. With these tools scientists can simultaneously explore digital montages of a slice of a rat cerebellum at different scales, while interacting with their collaborators at remote sites.

With traditional networking technologies, supporting real-time interactive large-scale scientific collaborations is virtually impossible. Due to limited availability of sophisticated imaging instruments and storage devices, brain data is collected and distributed across sites. Collaborative environments require the ability to disseminate, share and visualize this data quickly and in real-time, which demands high-speed network service. However, BIRN's datasets are massive: individual 2D brain images can be as large as 1GB, while high-resolution 3D images can easily exceed 100GB in size [1]. To support a multi-modality brain mapping experiment, it will eventually require the ability to process, share and interactively visualize multiple 100GB datasets – a network bandwidth requirement of several terabits per second. Today, to simultaneously visualize and explore eight 3D images on a LambdaVision, it requires 64Gbps of network bandwidth to fetch new regions of data (8x1GB) for every interactive "zoom" or "pan" operation.

Configurable Optical Network Services for E-Sciences

The OptIPuter project [5] is an National Science Foundation (NSF) funded research project to exploit the availability of dynamic, high-speed optical paths to provide revolutionary capabilities for emerging E-sciences. OptIPuter envisions dedicating sets of lambdas to small groups of applications for secure, high-performance and reliable execution; a decidedly different goal from the Internet – a shared network resources for millions of users. Specifically, an OptIPuter is a middleware-enabled, configurable collection of private optical networks and Grid resources [6]. The OptIPuter project is researching and prototyping a wide range of innovative middleware to realize this vision. To simplify application use of these resources, our approach is based on the idea of a Distributed Virtual Computer (DVC) [7] – a resource abstraction which provides applications with a simple usage and performance model. DVC's enable applications to conveniently acquire distributed resources and dedicated lambdas, and use them as a private resource context to manage both application functionality and performance. Within a DVC, applications can have secure, high-speed and reliable access to remote resources such as compute, storage, or display devices.

Configurable optical networks admit a wealth of possible service models, varying in optimization objectives, levels of application visibility, and granularity of network allocation. We describe several of the most popular models below:

- Intelligent network (INET): The network monitors the traffic and automatically creates optical circuits when high-speed flows are detected, and tears them down when the flow ends.
- Asynchronous File Transfer (AFTP): Applications submit file transfer requests and the network infrastructure schedules and transfers the requested files through short-lived, dynamic optical circuits.
- Distributed Virtual Computer (DVC): Applications explicitly request desired network connectivity, and use the resulting dedicated network to achieve predictable, high performance.

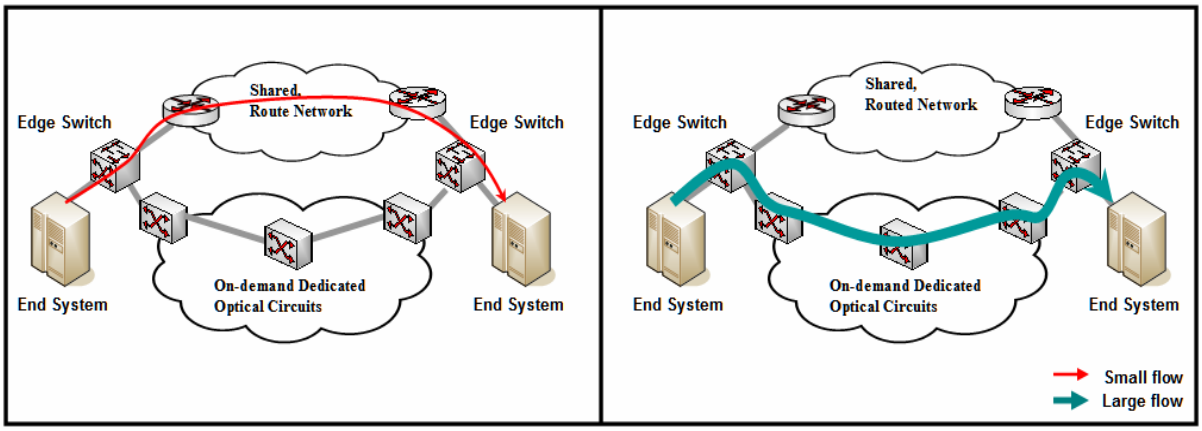


FIGURE 3. Integrating shared and configurable optical circuit-switched networks

A successful model must not only enable applications and efficient resource use, but also be attractive for network operators. To compare and evaluate these three network service models, (INET, AFTP, DVC), we use synthetic workloads, and transfer-level network simulations. Metrics cross the two key constituencies – application users and network operators – and include application performance, resource utilization, and network setups.

The remainder of this article is organized as follows. First, we present the OptIPuter approach (DVC) of use of configurable optical networks. Second, we present two alternative models (INET and AFTP). Third, we describe the simulation model used in this study and provide a detailed comparative analysis of the three models. Finally, we summarize the conclusions and outline their key cost/performance tradeoffs.

II. The OptIPuter Approach (DVC): Dedicating Private Networks for High-Performance Applications

The OptIPuter approach is to enable innovative scientific applications to exploit the Grid and configurable optical networks to couple petabyte data collections across wide-area networks. Our project is driven by two leading E-science efforts (EarthScope and BIRN), providing "application pull", or requirements, for development of novel distributed computing and networking infrastructures and enabling middleware.

Specifically, the OptIPuter is a revolutionary distributed infrastructure that dedicates optical networks for individual applications, enabling them to achieve unprecedented high quality of service. In the OptIPuter, private end-to-end lambdas are configured on-demand between distributed resources. Unlike connections in virtual private networks (VPNs), these optical paths are truly dedicated; there is no sharing of the optical circuits. Each circuit is a direct, secure and congestion-free path between end resources, enabling applications to achieve high performance and guaranteed quality of service.

While a DVC [7] provides applications with dramatically higher capabilities, there are significant questions about how to realize each DVC's private network configuration. We view a DVC as a resource abstraction which provides applications simple use and controllable performance, shielding them from the complexity of underlying software and hardware infrastructures. Operationally, to create a DVC, an application describes its resource needs, and requests a DVC which may include end resources (storage, compute) with a set of dynamically configured switches and optical circuits. In response, the DVC middleware matches the application requirements with appropriate network and end system resources. These resources are configured and reserved for dedicated use for the duration of the application. Within DVC's, the applications make use of these resources as a private resource context to achieve secure, high-performance and reliable execution (See Figure 2(a)).

As shown in Figure 2(b), the DVC abstractions are realized by a wealth of OptIPuter system software efforts in advanced distributed computing, network control planes, high-speed network protocols, and distributed storage. Development of the OptIPuter middleware relies on existing Grid technologies in many areas, including basis security, resource access and communication. We leverage existing Grid middleware, being innovative to leverage the novel capability of dedicated optical circuits to applications.

III. Alternative Service Models for Configurable Optical Networks

We present two alternative approaches to the DVC model, highlighting their distinguishing features and underlying assumptions.

III.A Intelligent Network (INET)

Intelligent network (INET) is a network-centric approach to enhance application performance transparently, while sharing network resources efficiently across applications. Using network monitoring, the INET approach dynamically creates (or adjusts) optical circuits to improve performance. For example, an optical circuit might be configured to optimize a large, high-speed flow by cutting through part of its path through a shared network.

INET is typically proposed as an enhancement to a shared network. As shown in Figure 3, a service provider implementing INET would comprise both packet-switched and optical circuit-switched networks. Traffic between two end hosts would initially stream through the shared network, but when a large flow is detected, an optical circuit is created dynamically and the traffic is redirected to use it. Subsequently, if the observed flow rate drops below a specified threshold, the traffic will be reverted to the shared network and the circuit will be released. In our simulation, the flows which send at over 8 Mbps more than 10 seconds are classified as large flows, and the network tries to allocate a dedicated optical circuit for each such flow. If a circuit is not available, the flow continues to use the shared network until a path can be allocated or the flow completes.

III.B Asynchronous File Transfer (AFTP)

The asynchronous file transfer model (AFTP) provides an asynchronous communication service where applications submit file transfer requests in a fashion similar to the FTP service. The transfers are scheduled and the applications are notified on completion. The asynchrony in AFTP allows the system to collect transfer requests from a number of applications and exploit this information to optimize the transfer scheduling and use of high-speed optical circuits. The optical circuits are created and are held only as long as necessary to complete the extant transfers.

The AFTP approach has been studied intensively [3,8]. One of the most efficient ways to schedule AFTP transfers on optical circuits is Varying-Bandwidth List Scheduling (VBLS) [8]. In VBLS, admitted transfers are given varying bandwidth allocations throughout their transfer periods according to the current state of workload and resource contention. In our simulation, we implement AFTP using a simplified VBLS algorithm. This algorithm divides each optical path segment allocation into globally synchronized time slots of duration 1 second (including the network reconfiguration time for 100 msec). For each file transfer, the system scheduler constructs a transfer plan consisting of a list of time slots and bandwidths. If there is an available optical circuit between the target end resources, it is reserved for dedicated use until the transfer is completed. Otherwise, the transfer request is held in a system queue. Unlike the DVC approach, AFTP dynamically sets up a private optical path on a "per-flow" basis (instead of "per-application").

IV. Evaluating Network Service Models

IV.A Methodology

We compare and evaluate the three service models (INET, AFTP and DVC) using synthetic workloads, transfer-level network simulations and metrics across application performance, resource utilization and network setups.

The simulated network infrastructure consists of both packet-switched and optical circuit-switched networks (See Figure 3). Both DVC and AFTP utilize only an optical circuit-switched network, while INET uses it as an enhancement to the shared network. Our optical network topology models were derived from extant Internet Service Provider's networks (the MCI global backbone network [9]). The topology consists of 95 switches, 185 internal links and 2,245 end hosts. For each internal link, we assigned 50 lambdas – each at 1Gbps. On the other hand, we modeled a packet-switched network that supports transfers at 8 Mbps between any two end hosts. The end resources (host information) were generated by a statistical grid resource generator [10], which generates resource distributions matching currently deployed grid infrastructures. Each end host is connected to a 10Gbps uplink to the core network. Applications were assumed to send at 1Gbps, and each optical path setup takes 100ms.

Our workload is a synthetic trace of application file transfer requests. The transfer requests were generated using FONTS [11] – the Flexible Optical Network Traffic Simulator, a tool for simulating advance, on-demand or periodical requests of lightpaths in dynamically provisioned optical networks. These requests were grouped together to form individual applications. Each application has approximately 10 transfers amongst three to six randomly chosen end hosts. The average size of each transfer is 3.5 GB. We assumed two consecutive transfers to be dependent, meaning that the following transfer can begin only after the previous one has completed. In order to observe the system under different loads, we scaled the inter-arrival time between subsequent applications (or conversely the request rate). For each request rate, we used 5 traces, each with 63,000 applications shuffled in a random order. This number is high enough to ensure that the average and cumulative metrics are measured over simulations that spend greater than 95% of their time in a steady state.

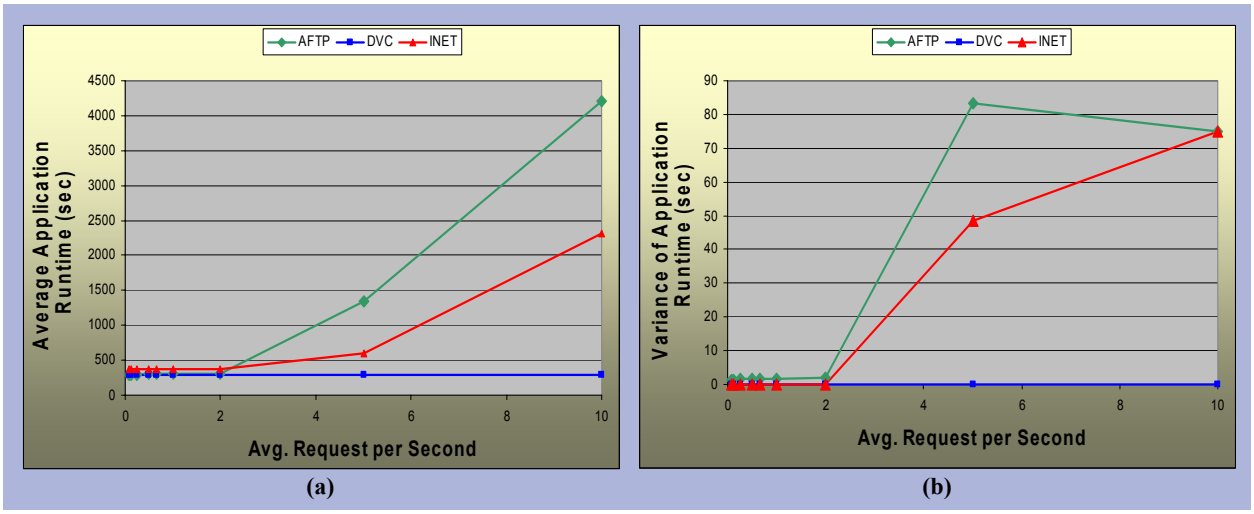


FIGURE 4. Comparison of application performance with different service models as a function of a load: a) average application runtime; and b) variance of runtime

IV.B Simulation Results

Application Performance

Perhaps the most important evaluation metric is how the network service model affects application performance. We characterize application performance by average runtime over a number of applications in the trace, and to characterize variability, we also report the standard deviation.

As shown in Figure 4(a), the DVC model delivers better performance than INET and AFTP for all the simulated workload cases. At low system utilization, we see the DVC consistently outperforms AFTP and INET by 4.38% and 23.10%, respectively. The advantages of DVC become clearer as system load increases. This is because all required networking resources are reserved for the entire application execution time. After a DVC allocation, the applications always have access to high-speed optical circuits without any additional circuit setup (switching) overhead. At high workload cases, INET outperforms AFTP because the transfers that are blocked from unavailable optical circuits can still utilize the shared network and thus finish sooner. In terms of quality of transport service offered to applications, DVC provides the most predictable network performance. As shown in Figure 4(b), there is no variation in application runtime with DVC regardless of the change in workload. This is attributed to the allocation of a private network for applications. On the other hand, we observe variations with AFTP and INET. With higher workloads, the file transfers may be delayed from unavailable lambdas.

Now, the DVC's superior application performance is real – applications do run much faster from start to finish. However, because the DVC model blocks application initiation when network requirements cannot be satisfied, the blocking time is an important factor. As load reaches high levels, DVC's average blocking time increases at a much faster rate than that of AFTP and INET. At the high workload of 10 requests per second, DVC's average blocking time grows up to ~50,000 seconds (or more than 13 hours). For DVC, an application needs to wait until all its required network resources become satisfied, while for AFTP it can start right after there is a lambda available for its first transfer event. For INET, because an application can always utilize a shared network, it can start right away.

Resource Efficiency

To measure the efficiency of resource use, we use two metrics: application lambda utilization is the fraction of time that an allocated lambda is sending application data and system lambda utilization is the fraction of lambdas in the system that are allocated for use. We report the averages of these metrics over the entire trace run.

Figure 5(a) show the application lambda utilization of the three models. Both INET and AFTP exploit nearly the full capacity of the allocated circuits (100% and 98.88%, respectively) regardless of the workload factor. These results reflect the philosophy that network resources are expensive, and these network service models manage them carefully. AFTP performs slightly worse than INET since its allocations are made in fixed length time slots. DVC produces much lower application lambda utilization (~16%) because it dedicates resources to applications throughout their entire executions, not to individual transfers. The application-level utilization for DVC on real workloads depends heavily on the private network request structure and actual application usage. Our workloads may represent a pessimistic case for DVC, as our application trace has only one active flow at a time.

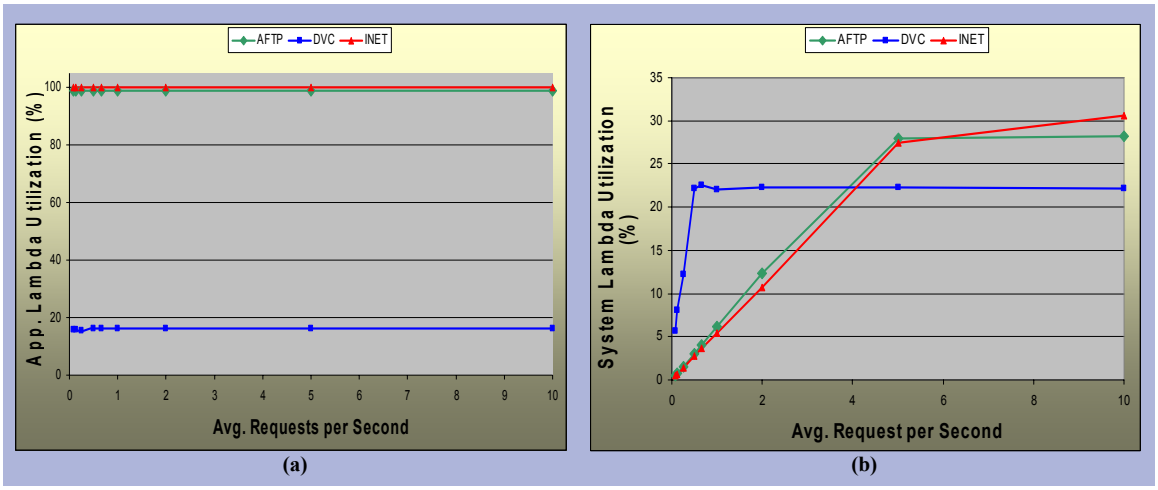


FIGURE 5. Comparison of lambda utilization with different service models as a function of a load: a) application lambda utilization; and b) system lambda utilization

Figure 5(b) illustrates the average system lambda utilization for all three network service models. For all three models, the system lambda utilization increases with higher loads. Because the DVC makes less efficient use of the lambdas, its utilization and need for system lambdas grow much faster than INET and AFTP for a fixed workload. However, when the request rates continue to increase and the resources become more congested, each model moves from linear growth to a flattened saturation region. DVC's saturation point is lower than that of INET and AFTP, because in DVC applications present more complex network requirements – multiple lambdas (between every pair of their target end hosts) must be simultaneously available. As a result, it's more difficult to allocate many applications (or relatively lambdas) to run at a given time.

Network Setup Cost

A new cost in dynamically configurable networks is the effort required to configure and remove connections. We categorize this cost as the number of optical circuit setups (and teardowns) that must be performed. For all three models the number of circuit setups remains constant regardless of the workloads. The numbers of circuit setups for AFTP, DVC and INET are 20,128,500, 390,600 and 541,800, respectively. AFTP has a dramatically larger number of connection setups because it uses VBLS scheduling which requires network reconfiguration at regular time slots. These time slots should be much shorter than transfers to achieve high network efficiency, yet long enough to help aggregate the network reconfiguration time. INET performed far fewer optical path setups than AFTP, but significantly more than DVC. INET creates a single connection for each application transfer that triggers automatic reconfiguration. On the other hand, DVC requires the fewest setups since it allocates its optical circuits once for the duration of the application. Compared to AFTP and INET, DVC avoids the costs of per time slot and per transfer connection setups. While DVC avoids these setups, this comes at the cost of poorer lambda utilization. These tradeoffs are visualized in Figure 6, a diagram that depicts the network transmission and network control cost space. It shows the region in which each network service model is attractive.

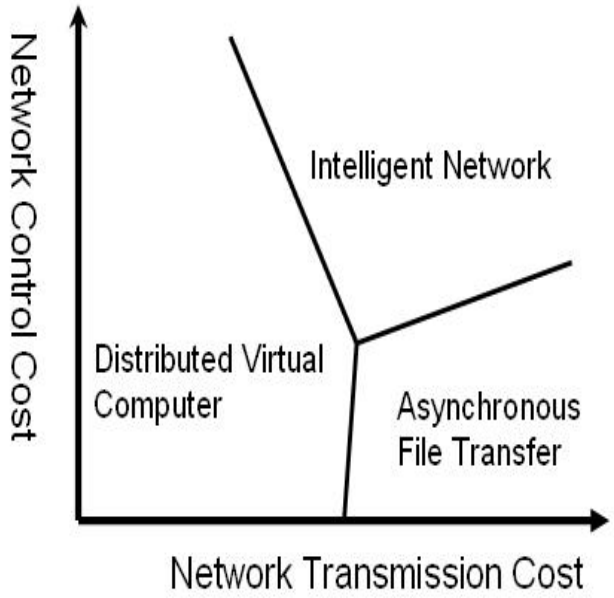


FIGURE 6. Cost balance for configurable network service models

V. Summary and Discussion

Emerging E-science and large-scale distributed applications have challenging network requirements. The OptIPuter project is exploiting configurable private optical networks based on the concept of Distributed Virtual Computer (DVC). With the emergence of plentiful, low-cost bandwidth enabled by DWDM, it has become possible to leverage excess network capacity for superior network service. The DVC enables applications to easily describe and acquire private high-speed networks, using them to achieve good performance throughout their execution.

For configurable network services to be broadly deployed, network service models must have demonstrable benefits for both applications and network service providers as well as acceptable costs and complexity. Our comparison of the OptIPuter approach (DVC) and two alternatives – Intelligent Networks (INET) and Asynchronous File Transfer (AFTP) – shows that the choice of model is indeed important, producing significant differences in usability, delivered performance, network configuration cost, and resource efficiency. The DVC model provides the best application performance, achieving both lower and deterministic application runtime. Both INET and AFTP achieve high system lambda utilization, suitable in networks where transmission costs are critical. In networks where network transmission cost is relatively low and resources are plentiful, DVC and other schemes which dedicate network resources to applications are most attractive and advantageous.

Acknowledgements

The work described in this paper is supported in part by the National Science Foundation under awards NSF Cooperative Agreement ANI-0225642 (OptIPuter), NSF CCR-0331645 (VGrADS), NSF ACI-0305390, NSF Research Infrastructure Grant EIA-0303622, and from NIH/NCRR for the National Center for Microscopy and Imaging Research (P41RR004050) and BIRN (U24 RR019701). Support from the UCSD Center for Networked Systems, BigBangwidth, and Fujitsu is also gratefully acknowledged.

References

[1] T. DeFanTi, et al., "Teleimmersion and Visualization with the OptIPuter," in *Proceedings of the 12th International Conference on Artificial Reality and Telexistence*, December, 2002.

[2] O. Yu, "Intercarrier Interdomain Control Plane for Global Optical Networks," in *Proceedings of IEEE ICC*, June, 2004.

[3] M. Veeraraghavan, X. Zheng, H. Lee, M. Gardner and W. Feng, "CHEETAH: Circuit-switched High-speed End-to-End Transport Architecture," in *Proceedings of the 4th Optical Networking and Communications Conference*, October, 2003.

[4] B. Jeong, et al, "Scalable Graphics Architecture for High-Resolution Displays," in *Proceedings of IEEE Information Visualization Workshop*, October, 2005.

[5] L. Smarr, A. A. Chien, T. DeFanti, J. Leigh, and P. Papadopoulos, "The OptIPuter," in *Communication of the ACM*, 46(11), November, 2003. http://www.optiputer.net.

[6] I. Foster and C. Kesselman, editor, "The Grid: Blueprint for a New Computing Infrastructure," Morgan Kaufmann, 1999.

[7] N. Taesombut and A. A. Chien, "Distributed Virtual Computer: Simplifying the Development of High-Performance Grid Applications," in *Proceedings of the Grids and Advanced Networks Workshop*, April, 2004.

[8] M. Veeraraghavan, H. Lee, E. K. P. Chong and H. Li, "A Varying-Bandwidth List Scheduling Heuristic for File Transfer," in *Proceedings of the IEEE International Conference on Communication (CC2004)*, June, 2004.

[9] MCI Global Network http://global.mci.com/about/network/global_presence/global/

[10] Y. Kee, H. Casanova, and A. A. Chien, "Realistic Modeling and Synthesis for Computational Grids," in *Proceedings of the ACM Conference on High Performance Computing and Networking (SC2004)*, November, 2004.

[11] S. Naiksatam, S. Figueira, S. A. Chiappari and N. Bhatnagar, "Analyzing the Advance Reservation of Lightpaths in LambdaGrids," in *Proceedings of the International Conference on Cluster Computing and Grid*, May, 2005.