DWDM-RAM: An Architecture for Data Intensive Services Enabled by Next Generation Dynamic Optical Networks

D. B. Hoang[1], T. Lavian[2], S. Figueira[3], J. Mambretti[4], I. Monga[2], S. Naiksatam[3], H. Cohen[2], D. Cutrell[2], F. Travostino[2].

[1]University of Technology, Sydney, [2]Nortel Networks Labs, [3]Santa Clara University, [4]iCAIR Northwestern University.

dhoang@it.uts.edu.au, {tlavian, imonga, hcohen, dcutrell, travos}@nortelnetworks.com,

{sfigueira,snaiksatam}@scu.edu, j-mambretti@northwestern.edu.

Abstract: **An architecture is proposed for data-intensive services enabled by next generation dynamic optical networks. The architecture supports new data communication services that allow for coordinating extremely large sets of distributed data. The architecture allows for novel features including algorithms for optimizing and scheduling data transfers, methods for allocating and scheduling network resources, and an intelligent middleware platform that is capable of interfacing application level services to the underlying optical technologies. The significance of the architecture is twofold: 1) it encapsulates "optical network resources" into a service framework to support dynamically provisioned and advance scheduled data-intensive transport services, and 2) it establishes a generalized enabling framework for intelligent services and applications over next generation networks, not necessarily optical end-to-end. DWDM-RAM[1] is an implementation version of the architecture, which is conceptual as well as experimental. This architecture has been implemented in prototype on OMNInet, which is an advanced experimental metro area optical testbed that is based on novel architecture, protocols, control plane services (Optical Dynamic Intelligent Network-ODIN[2]), and advanced photonic components. This paper presents the concepts behind the DWDM-RAM architecture and its design. The paper also describes an application scenario using the architecture's data transfer service and network resource services over the agile OMNInet testbed.**

1   INTRODUCTION

Although the existing packet-switching communications model has served well for transporting short data packets with burst transmission, e.g., for consumer oriented email and general web applications, it has not been sufficiently adaptable to meet the challenge of large scale data flows, especially those with variable attributes. Yet many emerging applications, especially within Grid environments, require the transport of such large scale data. For example, High Energy Physics (HEP) projects like Large Hardon Collider (LHC) at CERN [8] are projected to generate PetaBytes of data.

Packet switching has several limitations. The switching time required between packets for packets of 1500 bytes is around 10ms for a 1Mbps link, and about 100ns for a 100Gbps link. Therefore, it is difficult for silicon-based processing to support required switching times for extremely massive data flows. For example, sending a 100 Mbyte file over a 100Mbps link could take about 10s, but sending a 100 TeraByte file over a 100 Mbps link could take about 60 days. Furthermore, the transfer time in this simple calculation assumes only one hop between the data

---

source and the data sink. When large amounts of data are broken into packets and routed over multiple hops in the Internet, delays can be intolerable. An acceptable alternative may be dynamic wavelength switching, based on new optical technologies. One reason to explore this alternative is that many current and emerging large scale applications often require cooperation of resources that are distributed over many heterogeneous systems at many locations. To enable new classes of powerful distributed application, the Open Grid Services Architecture (OGSA) is being developed to allow ready access to multiple resources, such as advanced computation capabilities, extremely large data sets, and various "network resources". In order to realize this potential, an enabling framework is necessary to support application services that can intelligently schedule resources, for example, to provide for large scale data transfer among multiple geographically distributed data locations, interconnected by paths with different attributes. This paper presents a generalized enabling framework for intelligent services for high performance applications provisioned over next generation optical networks. In order to fully utilize the available bandwidth of the underlying optical network, applications must become aware of the network as an explicitly negotiated and scheduled resource. Traditionally, the designers of networks have not needed to concern themselves with long-term scheduling issues. The architecture proposed here is innovative in its premise that network services should move beyond notions of the network as an externally managed resource by allowing them to also include capabilities for dynamic on-demand provisioning and advance scheduling.

DWDM-RAM, described here, is an implementation of one type of architecture for data-intensive services enabled by next generation dynamic optical networks. A novel feature is its explicit representation of data transfer scheduling and network resource scheduling models. Another consists of a set of protocols for managing dynamically provisioned wavelengths. DWDM-RAM is experimental as well as conceptual. It is being implemented in prototype on OMNInet [10], an advanced optical networking testbed.

The rest of the paper is organized as follows. Section 2 presents our proposed architecture and its principal components. Section 3 describes an application scenario. Section 4 describes a current DWDM-RAM implementation. Section 5 briefly discusses related work on network resource management services and dynamic optical networks. Section 6 concludes the paper with suggestions for future work.

2 AN ARCHITECTURE FOR DATA INTENSIVE SERVICE ENABLED BY DYNAMIC OPTICAL NETWORKS

This proposed architecture, integrated with a dynamic underlying optical network, will support: 1) both on-demand and scheduled data retrieval, 2) a meshed wavelength switched network capable of establishing an end-to-end

lightpath in seconds, 3) bulk data-transfer facilities using lambda-switched networks, and 4) out-of-band tools for adaptive placement of data replicas.

In this architecture, the distributed virtual services control various sets of resources, including sets of data residing on data nodes, compute nodes and storage nodes connected by a set of access channels and optical lambda channels. The topology is unconstrained. Given any two points, the optical network can be asked to reserve or construct a path on demand, including light paths, with some parameters, such as QoS parameters. In order to perform the schedule optimizations discussed later, it is necessary that the optical network accept some input as to choice of paths, when multiple choices are possible. The data-intensive service architecture handles all aspects from discovering and acquiring appropriate resources regardless of their locations, coordinating network resource allocations and data resource allocations, making plans for optimal transfer, initiating and monitoring the execution of the resources and the transfer and notifying the client of the status of the task.

The proposed architecture is shown in Figure 1. It is flexible such that it can be implemented as layers or modules via an object oriented approach. As layer architecture, its components can be realized and organized hierarchically according to their service layer. A component at a lower layer provides services to a component at the layer immediately above it. As modules via an object-oriented approach, the architecture provides a more flexible interaction among its modules. Each module is a service, which can be composed of other services. Interface between services will follow specifications; say for Grid services, so that they can interact in a well defined manner.

Conceptually, the architecture supports data-intensive services by separating itself into 3 principal service layers: a Data Transfer Service layer, a Resources Service layer and a Data Path Control Service layer, over a Dynamic Optical Network as shown in Figure 1.

2.1   Data Transfer Service Layer
This layer presents an interface between a system and an application. Currently, it has only one module, the Data Transfer Service module; however, for a new class of applications, other modules can be accommodated.

The Data Transfer Service (DTS) consists of two closely coupled services: a Basic Data Transfer Service and a Data Transfer Scheduler Service. The Basic Data Transfer Service is mandatory and the Data Transfer Scheduler Service is necessary for a more complex and intelligent DTS. The data Transfer Scheduler Service thus can be developed independent of the Basic Data Transfer Service.
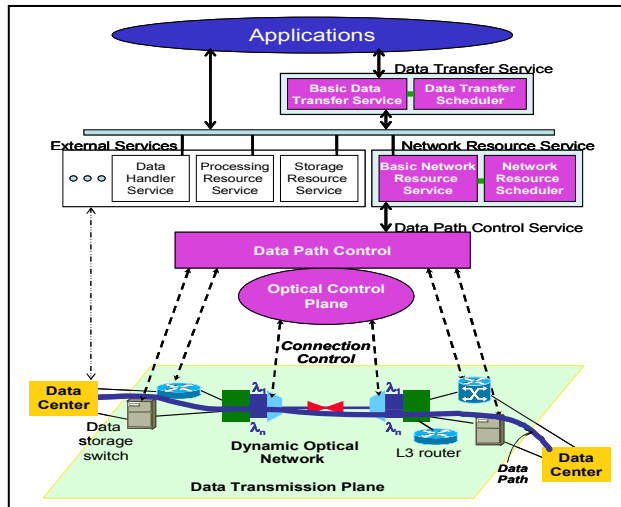
Figure 1 – Data-intensive service architecture

As the top layer service, the DTS receives high-level client requests, policy-and-access filtered, to transfer specific named blocks of data with specific advance scheduling constraints. DTS has a complete knowledge of the available network, processing, storage, data resources and the current schedules maintained by various Resources Services and Schedulers. It employs an intelligent strategy to schedule an acceptable action plan that balances user demands and resource availabilities. The action plan involves advance co-reservation of network, processing, and storage resources. Those resources are also used by other entities that are not managed by the DTS. A degree of distribution of the planning intelligence may be provided by a cooperative planning protocol, which allows peers to negotiate a schedule in the absence of global information. Extensions of this service can add support for replication and versioning of data objects. Various models for scheduling, priorities, and event synchronization can be employed.

2.2   Resource Service Layer

This layer consists of various services; some of them manage different types of resources (e.g., Network Resource Service, Processing Resource Service, and Storage Resource Service), some effectuate data transfer (e.g., Data Handler Service), and others perform additional services. This layer offers the Data Transfer Service layer all necessary information for effective application level scheduling and initiating the data transfer. For our intended architecture, the core service in this layer is the Network Resource Service (NRS). Other services are considered external and they can be called into service when necessary. Similar to the DTS, the NRS consists of two closely coupled services: a Basic Network Resource Service and a Network Resource Scheduler Service. The Basic Network Resource Service is mandatory and the Network Resource Scheduler Service is necessary for more complex and intelligent NRS. In general, the NRS should be able to:

- Interact with the control plane of the underlying network to discover the exact network topology, including a list of nodes, segments, and bandwidth parameters.

- Determine the current runtime utilization of specified segments.

- Make an on-demand reservation for a path by specifying the requested segments (and perhaps end-nodes and throughput parameters).

- Deallocate a route reserved as above, perhaps again through an explicit list of segments.

The NRS receives requests from the DTS, as well as requests from other services such as Grid services (both scheduled and on-demand). It maintains a job queue and allocates proper network resources according to its schedule. This service may be co-managed by the Grid Reservation and Allocation Manager [2] in OGSA. It relies on detailed up-to-date information about network resources (topology, bandwidths, etc.). A crucial feature of this layer is support for a bi-directional client callback interface, which is used to request that clients attempt to reschedule their previously scheduled jobs to achieve global optimization in the face of dynamically changing conditions.

2.3   Data Path Control Service Layer

The optical layer of the architecture is a dynamic lightpath service, implemented on an advanced operational wide area optical network, based on a novel architecture, protocols and control plane services: Data Path Control services. The Data Path Control Service layer resides between the resource service layer and various network resources. It receives resource requirement requests from the higher level service layers and matches those requests with network resources, such as path designations. It has complete understanding of network resource state information because it receives this information from lower level processes. The Data Path Control Services can establish, control, and deallocate complete paths across both optical and electronic domains.

2.3.1   Dynamic Optical Network

The architectural approach described here is compatible with basic Grid services concepts of resource discovery, integration, use, and release. This architecture extends this concept to network resources, particularly, lightpaths within optical networks. Currently, almost all optical channels within traditional carrier networks are statically provisioned. This type of provisioning, while useful for traditional types of communication services, is highly restrictive for many emerging types of applications. The architecture proposed here allows for dynamically provisioned optical paths, not only within data centers and metro areas but even across long haul spans.

3   AN APPLICATION SCENARIO

Consider an environment in which an individual client may request a certain large file to be transferred to its site under some constraints related to the actual time of the transfer operation. For example, a High Energy Physics group may wish to move a 100TB data block from a particular run or set of events at an accelerator facility to its local or remote computational machine farm for extensive analysis. An application client issues requests related to data named in an abstract namespace. In this model, a client may be associated with a unique data store node on the network, and the request is to obtain a copy of the data to that node. An important consideration is "Is the copy the 'best' copy?" What is "best" depends on a) the content, b) the network connectivity/availability, and c) the location in the context of the process. The client issues the request and receives a ticket in response which describes the resultant scheduling and provides a method for modifying and monitoring the scheduled job. The client does not know or care about the actual source of the data, which may come from any of the nodes of the network; indeed, the source might be any one of a number of replicas of the original data file, chosen by the Data Transfer Scheduling Service, in interaction with other Grid services.

Client requests include scheduling specifications. A typical scheduling specification is "Copy data X to the local store on machine Y after 1:00 and before 3:00." At the application level, Data Transfer Scheduler Service creates a tentative plan for data transfers that satisfy multiple requests over multiple network resources distributed at various sites, all within one Administrative Domain. The scheduled plan is formed based on the knowledge of the requirements of the requests, the existence of the data and its size, its locations and availability. At the middleware level, a network resource schedule is formed based on the understanding of the dynamical lightpath provisioning capability of the underlying network and its topology and connectivity. Co-reservation of network resources occurs at this level. At the resource provisioning level, the actual physical optical network resources are provisioned and allocated at the appropriate time for a transfer operation. Finally, a Data Handler Service on the receiving node is contacted to initiate the transfer. At the end of the data transfer process, the network resources are deallocated and returned to the pool.

During the above procedure, network resources may be unavailable or better plans can be scheduled or better agreement can be negotiated with the clients, and a firm schedule will eventuate and be executed within the predetermined time window. Job requests which cannot be accommodated within the context of a current schedule may result in callbacks to requesting clients to ask them to reschedule their allocated jobs in order to better satisfy all

current requests or newer higher priority requests. In all, the system tries to satisfy all data transfer and network bandwidth requests while optimizing the network usage and minimizing resource conflicts.

4 THE DWDM-RAM ARCHITECTURE

The DWDM-RAM is a realization of the general architecture proposed in section 2. Figure 2 illustrates the architecture. Applications can interact directly with either the Network Resource Service (NRS) or the Data Transfer Service (DTS) or both, as well as with other Grid services. One important point here is the separation of network services from file transfer services. The latter depends on the former, but network services may be requested (on-demand or via advance scheduling) independent of any file transfer request.
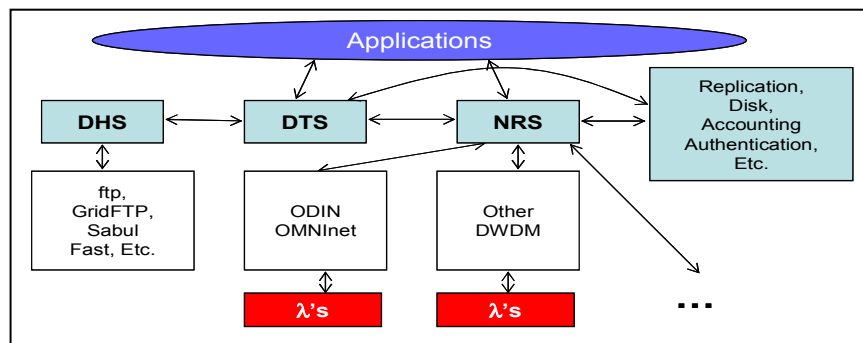

Figure 2 – The DWDM-RAM Architecture

A request for network bandwidth to NRS may be satisfied by its own network scheduling and routing modules (both end-to-end and segment-by-segment requests are allowed). NRS decides which underlying networks to use to satisfy a request. In the OMNInet testbed (described below), resources are controlled by ODIN software (described below). Other networks might also be available and NRS can hide their implementation details from upper layers and present a uniform interface to the application layer. A request is authenticated in terms of security, probably integrated with a policy server. Then the source data is verified: does the data exist, is it readable to this user, and how big is it? Data set size determines how long the transfer operation should take, given expected network speeds over the segments chosen for the transfer as well as IO capabilities of the end point machines. At the scheduled time for a data transfer, the NRS allocates the segment-by-segment path. The DTS then sends a request to the DHS running on the destination machine. When the transfer completes, DHS informs the DTS, which then tells the NRS to deallocate the path and return those resources to the pool available to service other requests. Initial results from a prototype implementation have been presented in [4, 17].

A. The OMNInet testbed [10] and the Optical Dynamic Intelligent Network Services (ODIN) [9]:

The OMNInet project is a multi-organizational partnership, which was established to build the most advanced metro area photonic network testbed. OMNInet is a wide area testbed consisting of four photonic nodes at widely separate locations in the Chicago metro area. These nodes are interconnected as a partial mesh with lightpaths provisioned with DWDM on dedicated fiber. Each node includes a MEMS-based (Micro-Electro-Mechanical Systems) Wave Division Multiplex (WDM) photonic switch, an Optical Fiber Amplifier (OFA), optical transponders/receivers (OTRs), and high-performance L2/L3 router/switches. The core photonic nodes are not commercial products but unique experimental research implementations, integrating state of the art components. The photonic switches are supported by Optera 5200 OFAs to compensate for link and switch dB loss. They are configured with eight ports capable of supporting 10Gbps optics. Application cluster and compute node access is being provided at three of the four locations by Passport 8600 L2/L3 switches, which are provisioned with 10/100/1000 Ethernet user ports, and 1GigE 1550XD trunks (i.e., dedicated fiber supporting 1550nm cross-connect trunks). The current OMNInet configuration is shown in Figure 3.

ODIN is server software that comprises of components that a) accept requests from clients for resources (the client requests a resource, i.e., implying a request for a path to the resource – the specific path need not be known to the client), b) determines an available path – possibly an optimal path if there are multiple available paths), c) creates the mechanisms required to route the data traffic over the defined optimal path (virtual network), and d) notifies the client and the target resource to configure themselves for the configured virtual network.
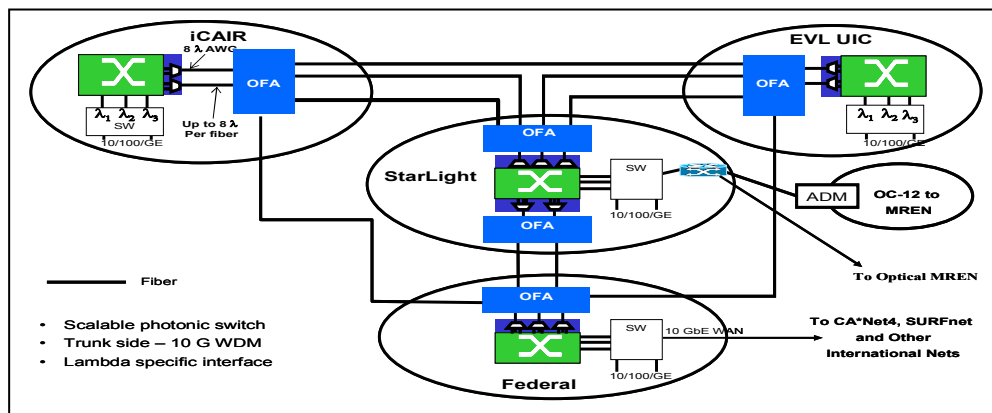

Figure 3 – OMNInet Testbed Configuration

5   RELATED WORK

TeraGrid [1] connects a grid network of supercomputers distributed in 4 remote locations from the Midwest to the West coast and exchanges data at 40Gbps transport rate (4xOC-192 10Gbps lambdas). However these are static

lambda connections while DWDM-RAM provides a dynamic setting of lambdas. TeraGrid requires L3 routing while DWDM-RAM provides dedicated optical path(s) to the destination. Therefore, the data transfer in Globus is done mainly in L3 environment, and specifically in GridFTP. These L3 routing limitations require doing specific optimization to the data transfer. In comparison, DWDM-RAM provides light path in the optical domain and not in L3 and layers above.

In the Grid architecture, a Resource Management Architecture for Metacomputing Systems [2] was proposed to deal with the co-allocation problem where applications have resource requirements that can be satisfied only by using resources simultaneously at several sites. This architecture, however, does not address the issue of advance reservations and heterogeneous resource types that are necessary for realizing end-to-end quality of service (QoS) guaranteed in emerging network-based applications [5]. To address this problem, the Globus Architecture for Reservation and Allocation (GARA) was proposed [6]. By splitting reservation from allocation, GARA enables advance reservation of resources, which can be critical to application success if a required resource is in high demand. Our architecture goes a step further by addressing one of the most challenging issues in the management of resources in Grid environments: the scheduling of dynamic and stateful Grid services where negotiation may be required to adapt application requirements to resource availability, particularly when requirements and resource characteristics change during execution. Recently, a WS-Agreement negotiation model was proposed [1] which uses agreement negotiation to capture the notion of dynamically adjusting policies that affect the service environment without necessarily exposing the details necessary to enact or enforce the policies.

The OptIPuter [3] research program is designed a new type of infrastructure based on a concept of enabling applications to dynamic create distributed virtual computers. This architecture will provide for a close integration of various resources, high performance computational processors, mass storage, visualization resources, and dynamically allocated distributed backplanes based on optical networks using advanced wavelength switching technologies. The first prototype is currently being implemented between StarLight and NetherLight. Other, more extensive, implementations should be more widely available around 2008. In contrast, DWDM-RAM has narrower scope, high performance data services over dynamic lightpaths within metro areas, and a shorter timeline toward more general implementations.

The GGF High-Performance Networking Research group also explores solutions towards an efficient and intelligent network infrastructure for the Grid taking advantage of recent developments in optical networking technologies [11].

6   CONCLUSION

This paper describes novel directions in advanced optical technology architecture. Generically, it establishes a generalized enabling framework for intelligent services and applications over next generation dynamically switched optical networks. It also presents a design based on those concepts - DWDM-RAM architecture as well as its implementation on the OMNInet optical network testbed. Specifically, DWDM-RAM is an architecture for data-intensive services enabled by a dynamically switched optical network. Further research is being conducted to provide for enhanced process performance measures, metrics and analysis. In addition, this research will be exploring closer integration among OGSA and WSRF compliant Grid services, DWDM-RAM architecture, and various supplemental processes such as Grid data and storage services. Also, other efforts are examining the relationship between optical networking protocols and higher level protocols.

REFERENCES

[1] K. Czajkowski, A. Dan, A., Rofrano, S. Tuecke, and M. Xu, "Agrement-based Grid Service Management (OGSI-Agrement)," *GWD-R draft-ggf-czajkowski-agrement-00*, June 2003.

[2] K. Czajkowski, I. Foster, N. Karonis, C. Kesselman, S. Martin, W. Smith, and S. Tuecke, "A Resource Management Architecture for metacomputing systems," *In the 4th Workshop on Job Scheduling Strategies for Parallel Processing, pp. 62-82. Springer-Verlag LNCS 1459, 1988.*

[3] T. DeFanti, M. Brown, J. Leigh, O. Yu, E. He, J. Mambretti, D. Lillethun, J. Weinberger, "OptIPuter Switching Middleware for the OptIPuter," *Invited Paper, Special Issue on Photonic IP Network Technologies for Next Generation Broadband Access, IEICE Transactions on Communications*, Vol. E86-B No 8 August 2003, pp., 2263-2272.

[4] S. Figueira, S. Naiksatam, H. Cohen, D. Cutrell, P. Daspit, D. Gutierrez, D. B. Hoang, T. Lavian, J. Mambretti, S. Merrill, F. Travostino, "DWDM-RAM: Enabling Grid Services with Dynamic Optical Networks," *IEEE CCGRID/GAN 2004, Workshop on Grid and Advanced Networks*, April 2004.

[5] I. Foster, M. Fidler, A. Roy, V. Sander, L. Winkler, "End-to-End Quality of Service for High-end Applications." *Computer Communications, Special Issue on Network Support for Grid Computing*, 2002.

[6] I. Foster, I., C. Kesselman, C. Lee, R. Lindel, K. Nahrstedt, and A. Roy, A., "A Distributed resource management architecture that supports advance reservation and co-allocation," *In Proceedings of the International Workshop on Quality of Service*, pp. 27-36, 1999.

[7] T. Lavian, D. Hoang, J. Mambretti, S. Figueira, S. Naiksatam, N. Kaushik, M.Inder, R. Durairaj, D.Cutrell, S. Merrill, H. Cohen, P. Daspit, F. Travostino, "A Platform for Large-Scale Grid Data Service on Dynamic High-Performance Networks," *Accepted for the First International Workshop on Networks for Grid Applications,* Oct 2004.

[8] LHC at CERN www.cern.ch/LHC.

[9] J. Mambretti, J. Weinberger, J. Chen, E. Bacon, F. Yeh, D. Lillethun, B. Grossman, Y. Gu, M. Mazzuco, "The Photonic TeraStream: Enabling Next Generation Applications Through Intelligent Optical Networking at iGrid 2002," *Journal of Future Computer Systems*, Elsevier Press, August 2003, pp.897-908.

[10] OMMInet: http://www.icair.org/omninet.

[11] D. Simeonidou (Ed), "Optical Network Infrastructure for Grid," *Draft submitted at GGF 9*, Chicago, Oct 5 – 8, 2003. (http://forge.gridforum.org/projects/ghpn-rg/).

[12] The TeraGrid: A Primer. http://www.teragrid.org/about/TeraGrid-Primer-Sept-02.pdf