

Requirements and Design of a Dynamic Grid Networking Layer

George Clapp, Joel W. Gannett, Ronald Skoog

Telcordia Technologies, Inc., 331 Newman Springs Road, Red Bank, New Jersey 07701

{clapp, jgannett, rskoog}@research.telcordia.com

Abstract

We address the requirements and design of bandwidth on demand networks in the context of grid services. Regardless of the deployment scenario for grid services (e.g., commercial or research), there is a need for efficient use of network facilities and a need to meet the performance requirements of the grid services users. We present quantitative analysis showing how these needs can be met.

1. Introduction

Grid Services and Web Services are becoming increasingly important to mainstream commercial information technology (IT). The two technologies are converging in the Open Grid Services Architecture (OGSA) to create standards and software that will provide greatly increased flexibility in carrying out computational tasks. Innovative Grid computing services promise substantial savings for organizations and new opportunities for service providers because data centers can be consolidated or significantly reduced, while the costs for computing can be made variable and based on usage. In addition, new computing models will be enabled that will stimulate the development of applications that go well beyond present day capabilities. For example, in the research and education (R&E) community today, visualization and collaboration efforts involving large databases (e.g., particle physics experiments) take place simultaneously across multiple geographic sites, but laborious and time consuming network setup is required to make this happen. In the computing and networking environment we envision, such massive allocation of resources would be done automatically and the resources used would be part of a commercially viable infrastructure.

The “on demand” aspect of Grid computing is fundamental to this model of computing. An organization may maintain a minimal computation

capacity and demand additional resources from a remote service provider as the need arises. Implicit in the additional computing capacity is the need for additional network capacity, and Grid computing is critically dependent on dynamic network services such as Bandwidth on Demand (BoD) and Virtual Private Networks (VPNs). Bandwidth on Demand gives an organization the ability to dynamically vary the bandwidth available on a network connection or between two or more sites. Virtual Private Networks provide connectivity to multiple sites of a subscriber and allow the subscriber to dynamically reconfigure both the bandwidth and the connectivity of the sites.

The full value of Grid computing can be realized only with dynamic network services. If organizations wish to work with extremely large data sets in real time and view the results in the form of data-intensive visualizations, then the bandwidth requirements of Grid applications might be many gigabits and even terabits per second. Despite the alleged “bandwidth glut,” bandwidth remains expensive, and it is simply infeasible for organizations to purchase high bandwidth network connectivity at their sites. Consequently, organizations make unwanted compromises between cost and performance and forego many promising applications. Advances in technology have created the means by which the needed dynamic network services can be offered. Optical networking provides very large, low cost, and flexible bandwidth through Wavelength Division Multiplexing (WDM) transmission and optical switching. The Generalized Multi-Protocol Label Switching (GMPLS) protocols of the Internet Engineering Task Force (IETF) and the optical User-to-Network Interface (UNI) and optical control plane of the Optical Internetworking Forum (OIF) enable dynamic and automated management of bandwidth in IP and optical networks. The Global Grid Forum is defining Grid network services that will enable computer applications to acquire, modify, and release bandwidth without human intervention. This paper advances our understanding of these network services by constructing a model for Bandwidth on

Demand and exploring the relationship between bandwidth granularity and network capacity.

In line with the view that is emerging from the R&E community [1], we envision that the grid networking layer will be a hybrid packet-optical infrastructure (HOPI). This means that the networking layer will provide various levels of bandwidth granularity for bandwidth management. This can range from the very fine granularity of best-effort IP packet networking to a very coarse granularity of full wavelengths and groups of wavelengths. In this paper we divide the bandwidth granularity into fine and coarse. Fine granularity relates to establishing connections with IP/MPLS paths or SONET/SDH channels using virtual concatenation and allocating bandwidth increments in the range of 1Mbps to 150 Mbps. Coarse granularity refers to full wavelengths, which would typically support bandwidths in the range of 1 to 10 Gbps in today's networks.

2. Fine Granularity Bandwidth Management

In this section we examine the fundamental forces that will drive the deployment decisions and network design methods regarding the network resources used to provide fine granular BoD capability for grid networks. Specifically, we consider how grid user sites would use fine granular BoD capabilities, and we identify the key parameters and issues that should be considered in network design so as to achieve efficient BoD facility usage while meeting grid network performance requirements (e.g., latency, jitter and loss). We envision that grid networks will emerge as both commercial enterprises and as research and education endeavors. Thus, the WAN that is used to provide the fine granular grid connectivity might be provided by either a telecom service provider or by customer-owned facilities. In either case, a choice between dedicated vs. BoD capabilities should be made to achieve high utilization and economic efficiency.

We first look at the potential grid user sites for BoD capability and examine how they would make economic tradeoffs between using dedicated connections (e.g., traditional private line service) and BoD connections. We then examine the problem of providing efficiently utilized transport facilities for the traffic flows generated by the BoD usage.

The main results of this section show:

- In the majority of cases, the most cost-effective grid network design is to use a mix of dedicated and BoD connections. The dedicated bandwidth connections are used to provide for a

'base level of bandwidth' that is known to be needed between grid user sites. The BoD is used to provide bandwidth capacity that exceeds the 'base level bandwidth.'

- To achieve the efficiencies needed to make BoD a cost effective capability, the network design needs to aggregate grid user BoD traffic flows so each transmission section (link) supports many BoD traffic streams. Our studies show that the number of traffic streams, rather than the size of the streams, is the critical parameter.
- The cost and utilization of the *dedicated* facilities used to access BoD has a significant impact on the amount of BoD that will be used. As the cost of dedicated BoD access becomes a larger portion of the total BoD costs, the average utilization of that dedicated access must increase for the use of BoD to be cost effective.
- The bandwidth granularity used for BoD has a significant impact on the cost efficiency of facilities used to provide BoD, with finer granularity providing greater cost efficiency (i.e., higher average utilization).

2.1 Network and Traffic Model

Figure 1 illustrates the network model used. Consider a grid network with multiple user sites. The grid network has bandwidth requirements between sites i and j that vary stochastically with user activity. We approximate the varying bandwidth required by the grid users by assuming bandwidth requests come in a fixed discrete bandwidth capacity unit (e.g., a bandwidth capacity unit could be a SONET STS-1 or VT1.5, an IP/MPLS LSP allocated 1 Mbps, etc.). Each requested bandwidth unit is used for a random holding time (its average is denoted by h). The idea is to provide the bandwidth needed over time to meet performance objectives, and we measure that bandwidth in multiples of the discrete bandwidth unit. The discrete bandwidth unit is called the granularity of the BoD capability. We assume the aggregate user behavior is random (e.g., user behavior is not correlated) so the discrete bandwidth unit requests arrive as a Poisson process.¹ The average arrival rate

¹ Extensive work in recent years (e.g., [4 – 6]) has shown that data network traffic is not Poisson and exhibits long-range dependence. We use the Poisson assumption here to keep the analysis tractable so we

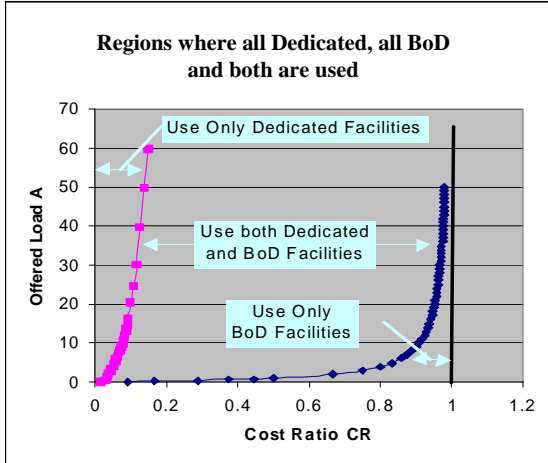


Figure 2. Customer Decision Regions for BoD

2.3 The Problem of Providing Facilities for BoD

Figure 3 shows the situation that needs to be considered to provide adequate capacity for BoD. Consider a transmission section (link) on which BoD channels need to be provided. The figure illustrates that there are N site-pairs that would need BoD capacity on this transmission section. Each site-pair load is first offered to the dedicated capacity, and if all dedicated capacity is busy it overflows to the BoD channels. The BoD provisioning problem is to determine how much bandwidth is needed to meet the BoD blocking objective of 1% blocking. The overflow load from each site-pair is a bursty non-Poisson arrival process. To size the required bandwidth to provide for a number of overflow streams as shown in Figure 3, we use the Wilkinson Equivalent Random Method [2].

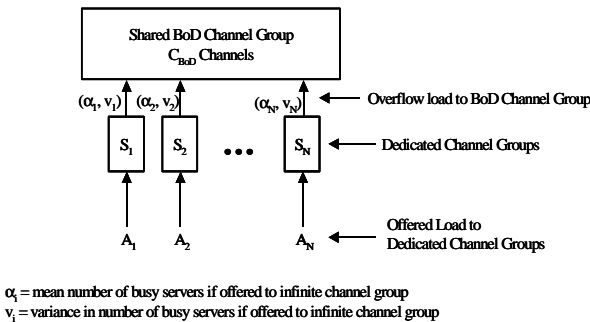


Figure 3. Model of BoD Carrier Load

Figure 4 shows the average utilization of the BoD installed bandwidth as a function of the number of customer point-pairs providing overflow load to the BoD facilities. We have assumed here all point-pairs have the same offered load, A . Utilization curves are shown for the point-pair offered loads of 10, 20 and 30 Erlangs. The main conclusion that can be drawn from this result is that there needs to be a large number of point-pairs providing overflow traffic if reasonable utilizations (> 0.7) are to be achieved. Also, it is seen that the value of the offered load A has little effect on the utilization. That is, the number of point-pairs overflowing is the critical parameter that determines a BoD carrier's bandwidth efficiency. This is because as A is increased, most of the increased offered load is carried by dedicated facilities, and the overflow load grows much slower than A 's growth. Thus, the BoD network needs to be designed so that many BoD traffic flows share the same transmission sections.

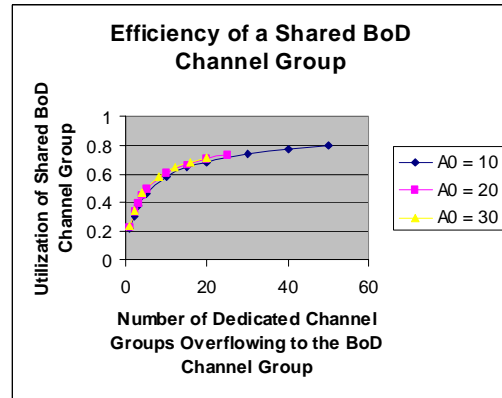


Figure 4. Utilization of WAN BoD Facilities

2.4 The Effect of Dedicated BoD Access Costs

The above results assumed the dedicated BoD access cost was negligible. If the dedicated BoD access cost is considered, we have shown that the previous results can be used with the cost ratio being suitably modified to reflect the access costs. The basic conclusion that can be drawn from these results is that as the BoD access costs become more significant, the higher the utilization of those dedicated access facilities need to be for BoD to be economic. Thus if access costs are significant, infrequent use of BoD (e.g., for failure or overload conditions) will not be cost effective.

3. Bandwidth Granularity and Network Capacity Issues

The bandwidth granularity used to provide BoD channels is an important consideration. If the bandwidth increments are chosen to be very small, then the bandwidth assigned to a site-site path can closely match the actual load and there is little “stranded capacity;” however, frequent bandwidth adjustments will need to be made. If bandwidth increments are chosen to be large, the frequency of bandwidth assignments will be reduced, but the assigned bandwidth will exceed the amount of bandwidth actually needed. The excess assigned capacity is called “stranded capacity” because it is unavailable for use by other site-site loads, and as a result there will be higher blocking of BoD channel requests.

But even if the bandwidth of the channels is well-matched to the BoD requests (so the channels are well-utilized once they are established), and even if the number N of channels on each link is well-matched to the time-average offered load, more fundamental considerations show that a small N implies high blocking. This follows because 1 out of $N+1$ channel utilization states is the blocked state (i.e., all channels in use); hence, if N is small, the blocked state simply becomes more probable statistically because there are fewer unblocked states. The curves in this section quantify this effect.

Another aspect that needs to be considered in BoD network design is the network load carrying capacity. As the results in Figure 4 show, BoD facility utilization on a link increases as the number of site-site flows using that link increases. Thus, network designs that allow many site-site loads to share link capacity will yield high link utilization. However, to achieve that sharing, longer site-site paths may be required, and this results in more total bandwidth capacity being installed per site-site unit of load. Thus, an important property of a network design is the additional site-site load that can be handled by an increase of a unit of bandwidth capacity on each network link.

To examine these issues, we carried out simulations on three realistic LATA-like networks of sizes 19, 71 and 200 nodes (LATA stands for Local Access Transport Area). We assumed each link in the network has the same number of channels (the channel size is the BoD bandwidth granularity). The model for site-site bandwidth unit requests coming into the network used exponentially distributed (Poisson) inter-arrival times and exponentially distributed holding times. As each bandwidth request arrives to the network, its user site end points are assigned randomly such that the

expected time-average number of connections terminating at a node is proportional to the degree of that node.

The total load offered to the network is normalized to the number of channels per link; specifically, the network offered load is expressed as Erlangs per channel. For a fixed amount of link capacity, increasing the number of channels per link is equivalent to decreasing the size of the BoD bandwidth granularity. Thus, a large number of channels per link represents very fine bandwidth granularity, and thus the ability to closely match the assigned capacity to the required bandwidth. We approximated the fine-grained “infinitesimal” limit with a simulation of 2048 channels on each link.

Figure 5 shows the BoD network blocking probability as a function of the network offered load for different numbers of channels per link (i.e., different bandwidth granularity). The indicated load L_f is the limit value for the allowed offered load (Erlangs per channel) that keeps the blocking probability at 0.001 (0.1%) as the number of channels goes to infinity (infinitely fine granularity). This quantity measures the maximum network throughput that can be achieved with 0.1% blocking. If the bandwidth of a link is B bps, then the maximum network throughput is $L_f \times B$ bps. This limiting throughput is related to the connectivity of the network, as will be discussed below. For a smaller number, N , of channels per link (coarser granularity), the maximum offered network load $L(N)$ for a specified blocking probability (e.g., 0.01%) decreases as shown in Figure 5; thus, as the granularity becomes coarser the maximum network throughput $L(N) \times B$ bps decreases.

If we plot the normalized maximum network offered load $L(N)/L_f$ as a function of the granularity N , we obtain the “relative efficiency” curve shown in Figure 6. From these results we see that it takes a granularity of 109 channels per link to attain 90% of the maximum load limit L_f . It is also seen that 79% of the limit L_f can be obtained with a relatively coarse granularity of 32 channels per link. It is also seen that very coarse granularity (e.g., less than 10 channels per link) leads to a significant drop in relative efficiency.

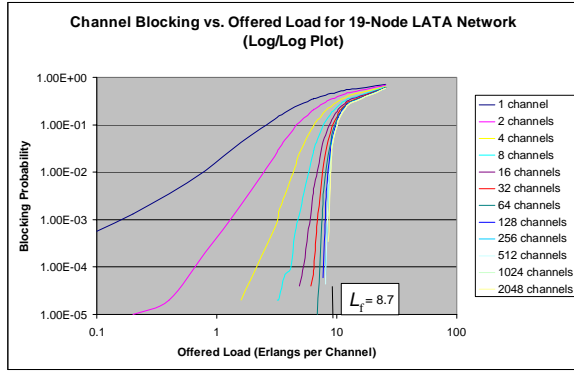


Figure 5. Blocking vs. offered load for the 19-node network

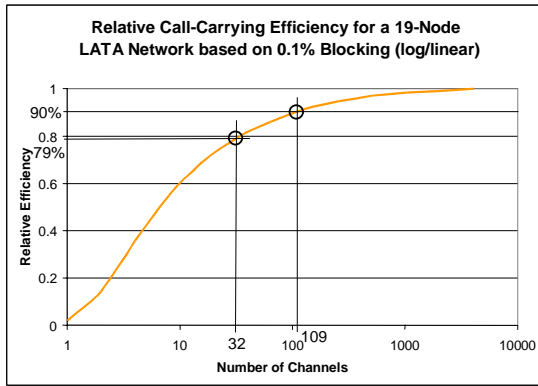


Figure 6. Relative efficiency vs. granularity for the 19-node network

Table 1 shows our simulation results for the three example LATA networks. For comparison, the results for a single link are included in the first row. The 19- and 71-node networks are based on real LATAs, while the 200-node network was generated using a proprietary Telcordia statistical LATA network generator.

Table 1. Simulation Results Summary

# Nodes / # Links	Average Nodal Degree	# Links in Minimal Balanced Cutset	L_f @ 0.1% Blocking (Erlangs/ch)	# of chs @ 90% Efficiency Ratio	Efficiency Ratio @ 32 Channels
2 / 1	1	1	1.001	559	0.57
19 / 31	3.3	5	8.7	109	0.79
71 / 138	3.9	6	7.5	121	0.79
200 / 361	3.6	13	14.6	164	0.71

The first column in Table 1 gives the number of nodes and number of links in the network, while the second gives the average nodal degree (i.e., the average number of links impinging on a node). The third gives

the size of the minimal balanced cutset of the network graph. That is, for each possible partitioning of the nodes into two equal-sized groups (or two groups whose size differs by one when the total number of nodes is odd), the set of links connecting the two groups is called a *balanced cutset*. A balanced cutset that has no more links than any other balanced cutset is called a *minimal* balanced cutset. The last three columns of Table 1 give, respectively, L_f the granularity (number of channels per link) needed to attain an efficiency ratio of 90%, and the efficiency ratio at a granularity of 32 channels per link.

It is notable in Table 1 that L_f seems to be tracking the minimal balanced cutset size. This is intuitively reasonable, since the minimal balanced cutset would tend to be the bottleneck of the network in the limit of infinitesimal granularity. On average, about 50% of the site-site channel requests offered to the network would have their A and Z sites lying on opposite sides of a given balanced cutset; this is a higher percentage than would be expected for any unbalanced cutset. Hence, a *minimal* balanced cutset would tend to be the bottleneck of the network in the limit of infinitesimal granularity.

While L_f applies in the limit of infinitesimal granularity, as we reduce the number of channels the chance of any particular link anywhere in the network becoming blocked increases. Local connectivity then becomes more important, and the number of channels available at a node becomes dominant. Hence, average nodal degree (a local topology metric) becomes more important than minimal balanced cutset size (a global topology metric) as the granularity becomes coarser. This observation is reflected in Table 1, where all three LATA networks have similar local connectivity (average nodal degree between 3.3 and 3.9) and we see that the finite-channel performance metrics in columns five and six are similar from row to row despite the differences in minimal balanced cutset size.

4. Coarse Granularity Bandwidth Management

As discussed in the introduction, the grid networking layer requires a range of bandwidth granularity choices. The previous section addresses the fine to medium range of granularity. In this section we address the coarse bandwidth management granularity, namely when lightpaths are used to establish site-site connectivity. The applications requiring a lightpath (or group of lightpaths) are those that involve very large file transfers that need to be completed in a relatively

short amount of time (e.g., so a collaborative group can exchange large volumes of data and have meaningful “near real-time” interactions). The problem is how to provide on demand lightpath capability in a cost-effective manner.

In the early stages of developing grid networking capabilities, it is expected that most of the connectivity requirements will be in the fine to medium bandwidth granularity. The number of users (applications) requiring lightpath connectivity will be relatively small, and the frequency that lightpath connections that would be required between a specific pair of sites will also be relatively small. Thus, in the context of Figure 3, lightpath connectivity would initially be well within the region where only BoD facilities are used. For lightpath connectivity, each grid site would need dedicated access to a core network that provides on-demand lightpath connectivity. Initially this core network might be hub-based with a few major interconnected hubs (e.g., like StarLight and NetherLight). Individual grid sites would have dedicated lightpath access to one of the hub nodes.

The cost and blocking performance of an individual site’s dedicated lightpath access is an important consideration. If a BoD capability to serve random arrival connection requests is desired, then the cost of the dedicated access can be quite high. For example, if a 10% (1%) blocking probability were desired, a single lightpath access connection could only be loaded to 10% (1%) utilization before a second access lightpath would need to be installed. Since these access facilities would need to reach a major hub node, they would be relatively expensive, and thus the access costs could become prohibitive for the random arrival/blocking mode of operation. An alternative is to use a scheduled access to the network. With scheduling it is possible to achieve very high utilization of an access facility and still meet the needs of the user community. The impact on the users is that they need to be willing to adjust their schedule to when facilities are available. Before significant demands for lightpath connectivity emerge, the core network (hubs and their interconnection facilities) may also need to be scheduled in order to achieve cost effective use of facilities. It is envisioned that the scheduling of the lightpath resources would be done using extensions of the grid scheduling and resource management capabilities that are being developed for computation and storage resources. In fact, an important area will probably be the need for coordinated scheduling between the computation, storage, and networking resources. Work along these lines is currently being done in the Global Grid Forum (e.g., [7] & [8]).

When the aggregate lightpath connection load out of a grid site becomes large enough, the scheduling discipline can be changed to a queueing discipline and achieve nearly the same cost savings. An important parameter in this case is the delay time the users would tolerate. An important distinction between scheduling and queueing is that with queueing the user does not have to plan far in advance; however, they need to be tolerant of some delay before their random arrival connection request can be met.

5. Conclusions

This paper explores how a dynamic grid networking layer would be used when cost efficiencies and commercial viability factors are considered. The results show that for fine granularity bandwidth management most grid sites will use BoD in conjunction with dedicated bandwidth services. The BoD will serve overflow traffic that exceeds dedicated capacity. The BoD network design needs to consider that it handles overflow traffic, which is bursty, and so the network must properly aggregate traffic to achieve sharing of capacity and efficient facility utilization. In the early stages of dynamic grid networking, coarse bandwidth granularity appears to be best managed through scheduling rather than random arrival/blocking disciplines.

6. References

- [1] J. Mambretti, *et al.*, “Hybrid Packet-Optical Infrastructure,” <http://www.internet2.edu/presentations/fall-03/20031015-Networks-Mambretti.ppt>
- [2] R.B. Cooper, *Introduction to Queueing Theory – Second Edition*, North Holland, 1981.
- [3] G. R. Ash, *Dynamic Routing in Telecommunications Networks*, McGraw-Hill, 1998.
- [4] W.E. Leland, M.S. Taquq, W. Willinger, D.V. Wilson, “On the self-similar nature of ethernet traffic,” *IEEE/ACM Transactions on Networking*, Vol. 2, No. 1, pp. 1-15, January 1994.
- [5] V. Paxton, S. Floyd, “Wide-area traffic: the failure of poisson modeling,” *IEEE/ACM Transactions on Networking*, Vol. 3, No. 3, pp.226-244, June 1995.
- [6] Special Issue : Advances in modeling and engineering of long-range dependent traffic, *Computer Networks*, (40), 2002.
- [7] V. Sander et al., "Networking Issues of Grid Infrastructures," Global Grid Forum draft, September 2003, work in progress, <http://forge.gridforum.org/projects/ghpnr-g/document/draft-ggf-ghpn-netissues-1/en/1>
- [8] D. Simeonidou et al., "Optical Network Infrastructure for Grid," Global Grid Forum draft, September 2003, work in progress, <http://forge.gridforum.org/projects/ghpnr-g/document/draft-ggf-ghpn-opticalnets-0/en/1>