IP Bandwidth on Demand and Traffic Engineering via Multi-Layer Transport Networks

Greg M. Bernstein
Grotto Networking
Fremont, USA
gregb@grotto-networking.com

Abstract— **This paper investigates the advantages of multi-layer methods of supplying moderate to high IP bandwidth either on demand for end users or for traffic engineering purposes over a shared, heterogeneous, wide area network infrastructure. In particular we show how emerging data plane and control plane mechanisms can be combined to deliver such services in an efficient and cost effective manner.**

Keywords- Bandwidth on Demand; GMPLS; MPLS; WDM; Ethernet; VCAT/LCAS; Control Plane;

I. INTRODUCTION

This article focuses on a multi-layer approach to Bandwidth on Demand (BoD) and IP traffic engineering over a wide area network (WAN). By multi-layer we specifically mean the active use of an additional technology layer such as MPLS, SDH or G.709 between the IP services layer and a WDM layer consisting of either lambda switches or wave band switches such as reconfigurable optical add/drop multiplexers.

In the case of BoD services we note that the "hold times" (the time duration that the communication flow is needed) can be significantly shorter from that of the timescales over which high bandwidth telecom/datacom services have traditional been provisioned. For example, a user may want 1Gbps of IP bandwidth between two points for only two hours as opposed to a period of weeks or months.

A. Single layer approach to IP traffic management and BoD

When we only have control over the IP layer to provide bandwidth on demand (BoD) or to traffic engineer our network we deem this a single layer approach. In a general high capacity, but not inefficiently over provisioned IP network, an effective way to allocate capacity is via the optimization of parameters governing the behavior of an IP interior gateway protocol (IGP). For example, near ideal route optimization for practical IP networks via the adjustment of OSPF [1] routing protocol link weights was shown in [2]. In particular, an optimization algorithm is run considering average and peak traffic flows through the network and a set of new IGP link weights are calculated. These new link weight values are then distributed to the routers "in charge" of those links from the IGP point of view. Then this updated information concerning the link weights is distributed via the IGP to all the other routers participating in the IGP. Finally all the routers participating in the IGP must recalculate their routing table based on all of these updates.

Needless to say this process is protocol and computationally intensive and disruptive of traffic carrying ability, since all routers in an IP network must have consistent routing tables, in order to avoid lost or looped packets [3]. In reference [4] it is shown how to minimize the number of link weights that need to be changed during such an optimization procedure. However, current practice does not utilize changing link weights for even diurnal (day/night) usage changes [4] and hence would not be an optimum method for supplying the two hour allocation of bandwidth on demand we previously mentioned.

B. A Two layer approach to IP TE and BoD

Given the existing difficulties in re-optimizing routing at the IP layer for BoD services, we need to either change the IP network topology (add remove links at the IP layer) or figure out a way to increase the size of the existing links. Given two agile layers, IP and WDM (based on reconfigurable optical add drop multiplexers) one would expect to take advantage of this additional flexibility to meet customer service demands. It is this "virtual topology" design problem that [5] focuses on, and, in particular, coming up with reasonable heuristic algorithms (since the problem in general is computationally hard) that also minimize the number of changes to the topology. Reference [5] calls this process "reconfiguration migration" which focuses on the operations such as removing or adding an edge to the IP layer network by reconfiguring the WDM network in small operational steps.

Unfortunately each of these topology changes is disruptive to the IP layer both the control and data planes. Recent work on IP routing protocols, [6], has looked at accelerating convergence times of IGPs such as OSPF and IS-IS with sub-second response times for single link removal deemed possible even for relatively large IP networks. However, this may still be far from acceptable with regard to impacts to guaranteed IP services or when multiple changes to the topology is required.

II. INTERMEDIATE LAYER DATA PLANE TECHNOLOGY

We can apply traffic engineering techniques to the IP layer alone or combined IP/WDM layers, however both techniques

are computationally intensive and can be disruptive to traffic flows. A third option would be to, if possible, increase the bandwidth available between IP routers, i.e., to an IP layer link. This would not preclude either IP IGP weight adjustments or IP virtual topology reconfiguration, but would provide a mechanism for dealing with BoD or shorter time scale traffic engineering in a non-disruptive manner.

Desirable properties that a potential intermediate layer data plane technology would provide include:

- Pipes of different sizes at a granularity smaller than that provided by the WDM layer
- A mechanism for changing the size of pipes without impacting the IP layer control plane
- A mechanism for changing the size of pipes without impacting the IP layer data plane
- A mechanism for the extraction of bandwidth scattered around from a mesh transport network

By providing "pipes" of different sizes smaller than the optical switching granularity, i.e., sub-wavelength in the case of a wavelength switching network we can use an appropriate amount of bandwidth for the desired flow leaving unused bandwidth available to other flows.

We separate the IP impacts into control plane and data plane since not all impacts to the IP data plane result in control plane action. For example, an optical or SONET layer protection switching, less than 50ms, action would result in an impact to some flows in the IP data plane but should not cause the IP control plane (routing protocols) to take action.

It may also arise is that no single link between the source and destination requiring bandwidth may have sufficient available bandwidth to meet the demand, but sufficient capacity exists across the network in the aggregate. Hence, a technique is desired, that is transparent to IP, for extracting bandwidth from an underlying mesh network. The general process of assembling multiple lower speed channels into a combined higher speed channel is called *inverse multiplexing*. This can either be accomplished at layer 1 or layer 2, i.e., circuit switched inverse multiplexing or packet based data link layer inverse multiplexing, respectively.

A. Candidate Technologies

In Table I we give a rough assessment of the data plane capabilities of Ethernet, MPLS, and circuit switch based Virtual Concatenation (VCAT) with and without the *link capacity adjustment scheme* (LCAS).

TABLE I. CANDIDATE TECHNOLOGIES AND THEIR PROPERTIES.

Tech	*Granularity*	*Impact on IP control plane*	*Impact on IP data plane*	*Extract BW from Mesh*
Ethernet	Continuous	No	No	Partial
MPLS	Continuous	No	No	Partial
VCAT w/o LCAS	Fixed increments	No	Yes	Yes
VCAT w LCAS	Fixed increments	No	No	Yes

1) Ethernet

Switched (Bridged) Ethernet with virtual LAN (VLAN) capabilities [7] from a data plane perspective has a number of features useful for both BoD and traffic engineering. First a mesh Ethernet network can be decomposed into separate VLANs with possibly distinct spanning trees, hence allowing for effective allocation of network bandwidth. Second, Ethernet provides "link aggregation" [8], which allows individual (point to point) Ethernet links of the same size to be combined to provide a higher capacity link. Note that this is a local, not network wide form of inverse multiplexing and can't help us extract bandwidth scattered around a mesh network. Another restriction on Ethernet Aggregation comes from its desire to preserve packet ordering to higher layer protocols such as TCP. Ethernet's link aggregation mechanisms preserves ordering in a somewhat restrictive way, i.e., they restrict "conversations" to the same port. They distinguish "conversations" by any combination of: source MAC address, destination MAC address, reception port, type of destination address, higher level protocol identification, etc… This may not meet the requirements for some high bandwidth applications.

Ethernet's control plane consisting of the learning bridge technique and multiple spanning trees is less well suited to BoD applications hence the recent experimental networks [9] and interest at standards bodies in applying a Generalized Multi-Protocol Label Switching (GMPLS) like control plane to Ethernet along with other extensions [10].

2) MPLS

Multi-Protocol Label Switching (MPLS) is a connection oriented form of packet switching that can use its own native data plane or can apply its control plane to other connection oriented packet switching technologies such as ATM. With MPLS' explicit routing capability, MPLS can make effective use of network bandwidth. However, MPLS does not currently support any native inverse multiplexing scheme which limits its ability to extract spare capacity from a mesh network. MPLS supports hitless regrooming but can not necessarily guarantee in order delivery of IP packets directly after the transition from one LSP to the "re-groomed" LSP.

B. Virtual Concatenation enabled Circuit Switching

A more recent high speed form of circuit based inverse multiplexing is *virtual concatenation* which has been applied to SONET/SDH, the Optical Transport Network (OTN) , and even the legacy PDH hierarchy [11]. As an example in SONET/SDH virtual concatenation allows for the "gluing together" of either approximately 50Mbps (STS-1/VC-3) or 150Mbps (STS-3c/VC-4) SONET/SDH signals. The standard allows up to 256 of the component signals to be combined allowing up to approximately 12Gbps and 40Gbps based on VC-3 or VC-4 based virtual concatenation respectively. In addition, the differential delay (caused by the components signals taking different paths) can be up to 256 ms, so different paths across a network can be used by the component signals to decrease the risk of a failure affecting the entire group. In the case of OTN (G.709) signals component signals

are either 2.5, 10, or 40Gbps with aggregate signals defined with up to 10Tbps bandwidth capacity.

An important supplementary capability to virtual concatenation is the link capacity adjustment scheme (LCAS). This low level protocol allows the coordinated addition and removal of component signals from a virtual concatenation group in a manner that is hitless to packet data services utilizing standard mappings. This is very important for our purposes here since this capability allows us to hitless modify the bandwidth served to the IP layer from the optical layer. More information on virtual concatenation, LCAS and packet layer mappings can be found in reference [12].

III. Multi-Layer Networking for IP TE and BoD

To use the transport networks multiple layer capability to allocate bandwidth appropriately at the IP layer we need to take into account the following:

- The WDM layer network topology,
- The intermediate layer network topology
- The IP layer network topology,
- The resources available in the intermediate layer that can be allocated to the IP layer, and
- In the case of Bandwidth on Demand the allocation of IP bandwidth to the appropriate IP flows.

In the following, for concreteness, we give examples utilizing circuit switched VCAT/LCAS technology as the intermediate layer since it currently is the only one of the candidate technologies that includes network wide inverse multiplexing and can require additional intermediate layer subnetwork connections to be established.

A. Optical Network Topology and Intermediate Layer Topology

The optical network topology can differ significantly from the IP network topology. For example in Figure 1 we show a possible (though not necessarily probable) optical topology for the network in Figure 3. The optical connectivity here could be either a single OC-48/192 (STM-16/64) SONET/SDH link or WDM links supporting multiple SONET/SDH links.

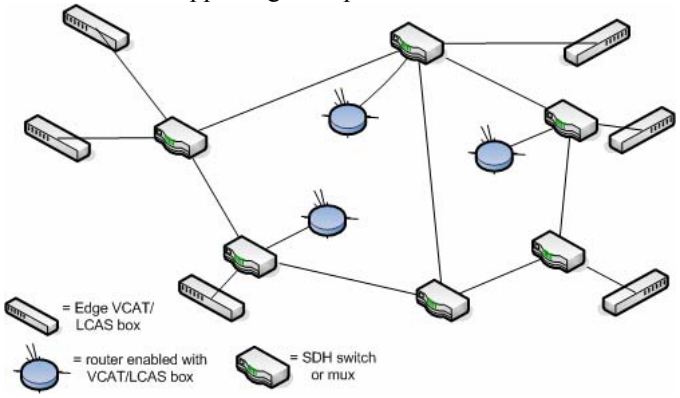

Figure 1. *View into the optical cloud. Note "meshy" optical network. Multiple paths between each IP router.*

B. IP Network Topology

Given a set of IP edge systems and core routers one can talk of the IP network layer topology on top of the intermediate network technology (circuit switched VCAT/LCAS) as shown in Figure 2. Here the specific link layer and physical layer technology is ignored and only the IP layer connectivity considered. For example, in Figure 2 we show the IP layer connectivity from the access/edge boxes to the routers and the adjacencies between peer (IP layer) routers.

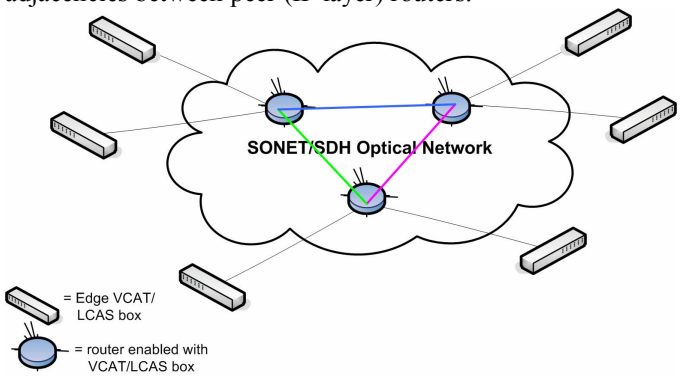

Figure 2. *Edge to router homing and routing adjacencies. Not necessarily optical paths or VCAT paths to or between routers.*

Since we are interested in the connection between IP and circuit switched VCAT networks, we will be interested in IP edge boxes featuring connectivity to the optical WAN infrastructure via VCAT/LCAS. The enterprise or central office side of these same boxes would most likely feature Gigabit Ethernet or 10 Gigabit Ethernet interfaces. Similarly we are interested in routers featuring VCAT/LCAS interfaces, either directly or indirectly with extra conversion equipment.

C. Allocation of Intermediate layer bandwidth to an IP path

As we stated previously we are interested in BoD for IP on a timescale less that that typically used for IP traffic engineering. For IP service between any two edge boxes there will generally be a single route. For example in Figure 3 we show the edge to edge IP path, in terms of IP layer links, traversed between edge box A and edge box Z.

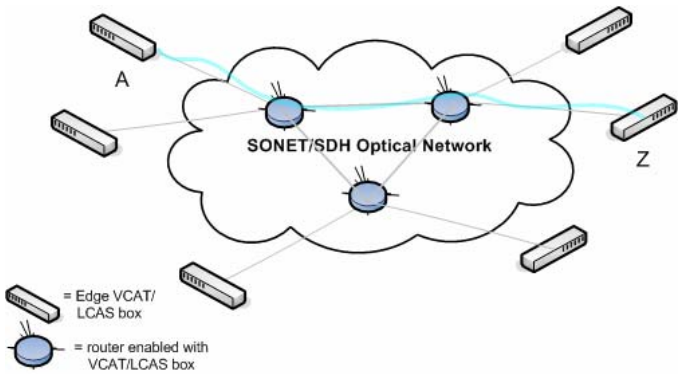

Figure 3. *IP Path between edge boxes A and Z.*

In Figure 4 we show the actual VCAT connections that support these IP layer links. The IP bandwidth given to a particular application between edge boxes A and Z is constrained by (1) the bandwidth on the VCAT connections between edge box A and router R1, router R1 and router R2, and between router R2 and edge box Z; (2) the other IP traffic using these same links.

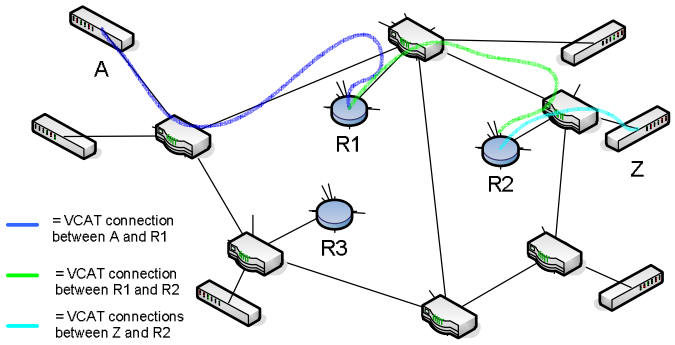

Figure 4. *Circuit switched VCAT connections supporting the IP layer along the path of from A to Z.*

Now we can use VCAT/LCAS to increase the bandwidth of each of these IP layer links without impacting the IP layer link weights (OSPF or IS-IS) or the IP layer routing protocols in any way.

In Figure 5 we show the addition of a VCAT component to each of the IP layer adjacencies along the IP path from A to Z. We see that these new VCAT components do not need to take the same path as the original component shown in Figure 4. This is the power of a network wide inverse multiplexing technology to extract bandwidth from a mesh.

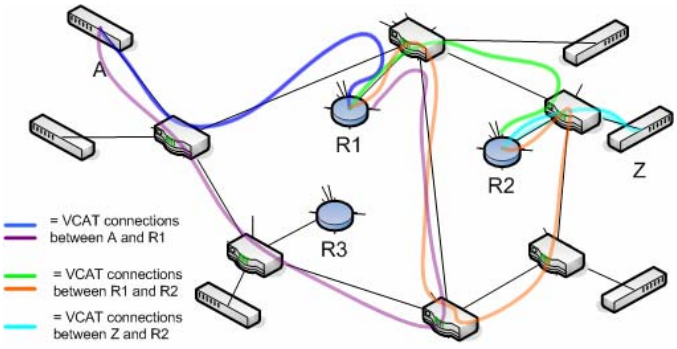

Figure 5. *Increasing end to end IP layer bandwidth via additional VCAT components on each IP layer link.*

D. Allocating the IP Bandwidth

Given that we can use an intermediate layer technique to adjust bandwidth available to an IP layer link, how do we insure that the appropriate users get to actually use this bandwidth? For this we need an IP layer mechanism. A number of methods have been proposed such as the integrated services internet (IntServ) and the differentiated services internet (DiffServ) [13]. In either case resources (bandwidth) for a flow are reserved on the routers along the path from source to destination. The packet flows may be identified by source-destination address and higher layer port (UDP/TCP) information in the case of IntServ "micro flows" or via the differentiated services code point (DSCP) in the IP header. Since the DiffServ mechanism deals with "aggregated flows", i.e., all those with the same DSCP, its scaling properties are significantly better for the "core" of the network.

The IP traffic is classified at the DiffServ domain boundary by its micro flow information and marked with a differentiated services code poin*t* (DSCP) in its IP packet header [14]. This is also where traffic can be restricted via policing of its average and peak rates and possibly other parameters. Once inside the differentiated service domain packets with a particular DSCP value are differentiated from each other with respect to how they are treated in queues encountered in routing nodes. This is known as the *per hop behavior* (PHB) given to a differentiated services aggregate. Connection admission control is needed to determine whether a new micro flow can be added to a DiffServ aggregate. In addition, if a micro flow is added to an aggregate, bandwidth allocation parameters may need to be adjusted. These are handled by the control and/or management plane. For the purposes of this paper we assume that suitable DSCP value and corresponding PHB has been implemented to meet the requirements of the application. Note that the collection of PHBs at a node can be viewed as an allocation of the available IP link bandwidth amongst the egress traffic on that link, but not as a method for increasing the available IP link bandwidth.

IV. Control Plane for BoD Services based on Multi-layer Transport Networks

To enable timely BoD services we need an appropriate control plane that can be used to provision flows and/or connections at multiple layers. In addition mechanisms must exist for the control planes at individual layers to interact to meet the overall service requirements. Currently MPLS and SONET/SDH networks via GMPLS have standardized control planes that include mechanisms for discovering topology and link resource status, as well as signaling for setting up explicitly routed connections. The first challenge lies in that most of these methods are specific to a single domain. The second challenge is how to link these mechanisms with emerging IP QoS signaling standards such as NSIS [15]. The framework for the Next Steps in Signaling (NSIS) [15] describes a new approach to signaling for IP networks that has been implemented via the protocols in [16, 17]. One of the advantages of NSIS over RSVP for IP QoS signaling is its design for use in a number of network scenarios beyond end-to-end such as end-to-edge and edge-to-edge. The NSIS signaling entity, particularly at a subnetwork border location, is ideally suited to interact with an intermediate layer technology such as VCAT/LCAS as needed.

The process would proceed as follows. We would start with an IP layer signaling mechanisms such as NSIS that would traverse the IP network requesting IP link bandwidth. At each IP link (interface) or IP subnetwork along the path, connection

admission control (CAC) is performed. If the IP link bandwidth is sufficient then CAC passes, if not we need to see if we can increase the bandwidth via the intermediate transport layer as previously described. Note that the lower layer bandwidth is allocated first, then the IP layer bandwidth. The provisioning order is important here since we don't want to open up the DiffServ bandwidth until after we've allocated the optical bandwidth. Otherwise our BoD IP flow may unduly impact other IP layer flows.

V. CONCLUSIONS AND DISCUSSION

In this paper we discussed intermediate layer data plane technologies such as MPLS, Ethernet and SONET/SDH VCAT/LCAS for use in supporting IP BoD applications in a manner that is not disruptive to the IP data or control plane. By intermediate layer we meant above the WDM layer but below the IP layer. We pointed out that existing single layer, IP only traffic engineering techniques are disruptive due to transient routing loops [3] and similarly for those that utilize both IP and WDM layer techniques. In addition we described how the IP control plane and the intermediate layer control planes could be integrated.

Recent work [18] on intra-domain IP routing protocols such as OSPF and IS-IS has shown that changes for network optimization or maintenance could be done in a manner that is non-disruptive if the routing protocols are modified in an appropriate fashion. Hence, if such changes are incorporated into standards the single layer approach could also be viable for SLA/QoS sensitive IP networks.

The choice of which intermediate layer candidate to use is as much situational and economic. For example MPLS is typically already integrated with IP routers; Ethernet Passive Optical Networks (PONs) may already be used in the access network; VCAT/LCAS can make very efficient use of existing WAN infrastructure; etc… Hence, a multi-layer approach such as one outlined here utilizing control planes for both IP and an intermediate layer can help provide cost effective BoD and traffic engineering to networks providing IP quality of service.

REFERENCES

[1] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[2] B. Fortz and M. Thorup, "Internet traffic engineering by optimizing OSPF weights," *Proc. IEEE INFOCOM*, 2000, pp. 519 - 528, March 2000.

[3] Urs Hengartner, Sue Moon, Richard Mortier and Christophe Diot, "Detection and analysis of routing loops in packet traces", IMW '02: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment, p. 107--112, 2002.

[4] Bernard Fortz and Mikkel Thorup Optimizing OSPF/IS--IS weights in a changing world, *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 4, May 2002, pp. 756 – 767.

[5] K. H. Liu, C. Liu, J. L. Pastor, A. Roy, and J. Y. Wei, "Performance and testbed study of topology reconfiguration in IP over optical networks," IEEE Trans. Communications, vol. COM-50, pp. 1662 - 1679, October 2002.

[6] Francois, P., Filsfils, C., Evans, J., and Bonaventure, O. 2005. "Achieving sub-second IGP convergence in large IP networks". *SIGCOMM Comput. Commun. Rev.* 35, 3 (Jul. 2005), 35-44.

[7] IEEE Std 802.1Q-2003, Virtual Bridged Local Area Networks, IEEE, 2003.

[8] IEEE Std 802.3-2002, Part 3: Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications Standard – Chapter 43 Link Aggregation, 2002.

[9] Ibrahim W. Habib, Qiang Song, Zhaoming Li and Nageswara S. V. Rao, "Deployment of the GMPLS Control Plane for Grid Applications in Experimental High-Performance Networks", IEEE Communications Magazine, vol. 44, no. 3, , March 2006.

[10] David Allan, Nigel Bragg, Alan McGuire and Andy Reid, "Ethernet as Carrier Transport Infrastructure", IEEE Communications Magazine, vol. 44, no. 2, , February 2006.

[11] Greg Bernstein, Diego Caviglia, Richard Rabbat and Huub van Helvoort, "Standards: VCAT/LCAS in a Clamshell", IEEE Communications Magazine, vol. 44, no. 5, , May 2006.

[12] G. Bernstein, B. Rajagopalan, and D. Saha, *Optical Network Control: Optical Network Control,* Addison Wesley, 2004.

[13] Bernet, Y., Ford, P., Yavatkar, R., Baker, F., Zhang, L., Speer, M., Braden, R., Davie, B., Wroclawski, J., and E. Felstaine, "A Framework for Integrated Services Operation over Diffserv Networks", RFC 2998, November 2000.

[14] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Service", RFC 2475, December 1998.

[15] R. Hancock, G. Karagiannis, J. Loughney and S. Van den Bosch, "Next Steps in Signaling (NSIS): Framework", RFC 4080, June 2005.

[16] H. Schulzrinne and R. Hancock, "GIST: General Internet Signaling Transport", Work in Progress, July 2006.

[17] J. Manner (ed.), G. Karagiannis, and A. McDonald, "NSLP for Quality-of-Service Signaling", Work in Progress, June 2006.

[18] Pierre Francois and Olivier Bonaventure, "Avoiding transient loops during IGP convergence in IP networks", IEEE INFOCOM 2005 - The Conference on Computer Communications, p. pp. 237 - 247 , March 2005.