Making High Bandwidth But Low Revenue Per Bit Network Applications Profitable

Abstract

IMPLICITLY, ALL PREVAILING 'NEW' NETWORK TECHNOLOGY MARKETING, REGARDLESS OF THE ACRONYMS AND BUZZWORDS USED, AVOID ADDRESSING THE VERY FUNDAMENTAL REASON FOR LOWER THAN EXPECTED PROFITABILITY BUILT-IN WITH ALL CURRENT PACKET-SWITCHED (E.G. IP, MPLS OR ETHERNET PROTOCOL BASED) NETWORK SERVICES.

TO DISPEL THE HYPE, OCS PROVIDES A BUZZWORD-FREE LOOK AT WHY, AND OFFERS A VIEW ON HOW TO REACH THE REQUIRED NEW STANDARD OF EFFICIENCY.

It is well known that packet oriented network protocols such as Internet Protocol (IP) cause highly time-variable (bursty) traffic loads between the communicating IP nodes over the networks that deliver the IP packets. However, less attention has been paid to the fact that so far, bandwidth allocation at the physical networks between the IP nodes is statically provisioned, even though packet traffic loads are anything but static.

As IP video, multimedia and transaction data etc. service bandwidth volumes keep exploding while the revenue per bit of network services continues to decrease, this mismatch between the prevailing packet traffic dynamics and the static partitioning of the delivery network is causing unsustainable levels of inefficiencies built-in with all prevailing network technologies and services that rely on statically provisioned network physical layer capacity allocation.

Thus the questions:

- ◳ *Why is network physical layer bandwidth allocation still static in prevailing network implementations?*

- ◳ *And will this have to change in order to make a sustained business case for the emerging service mix dominated by high-bandwidth multimedia etc. applications offering relatively low revenue, while demanding high Quality of Service (QoS)?*

Internet Bandwidth Allocation Is Still Statically Partitioned at the Network Physical Layer - Why?

To realize the level of structural inefficiency that static partitioning of network capacity causes, let's recall how IP itself serves its application protocols:

With today's IP, the L3 packets are generated and sent between IP host nodes when and where demanded by the L4+ client applications, for reason of network usage efficiency -- it is obvious that a fixed IP connectivity model would be coarsely wasteful: in such a scenario each given IP node would send fixed-length IP packets (filled with client application, or idle, bytes depending on the momentary demand) at fixed intervals to all other IP nodes with which the given node may at times have traffic to send, causing the IP server layer bandwidth (BW) to be most of the time either severely insufficient vs. application demands, or simply wasted as idle packets.
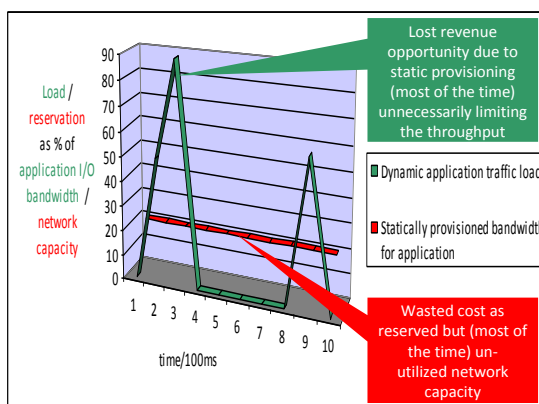
At network physical layers (L1/0), the conventional technologies (fiber/ WDM/SDH/OTN etc.) are limited to similar architectural inefficiency when carrying packetized L2+ (MPLS or direct IP over PPP, Ethernet etc.) traffic, as would be the case with fixed IP transmission for carrying variable BW L4+ applications. The industry has dealt with this structural inefficiency mainly by over-provisioning the (fixed capacity allocation based) L1/0 networks -- and even though such over-provisioning is not necessarily apparent in the light of average traffic loads, there in reality is plenty of room for efficiency gain when looking at the networks at the level that they really operate, which is at packet by packet level (this is analogous with the scenario of fixed-BW IP mesh).

To appreciate the new levels of efficiency that adaptive L1 can bring to networking, it should be realized that the current practice of using non-adaptive network physical layer is really just an artificial limitation caused by that the conventional L1/0 equipment cannot support adaptive BW physical layer channeling, and that therefore, if technically feasible, **also L1 BW allocation should be adaptive to realtime L2+ packet traffic load variations, for the same reasons that L3 IP packet transmission is adaptive to the L4+ application demands**.

Thus, to assert that L1 connection capacity allocation may well remain fixed while carrying packetized L2+ traffic loads without any significant loss of efficiency is equal to saying that L3 IP BW connectivity should be fixed when carrying packet oriented L4+ applications:

The use of fixed (or non-adaptive) BW L1 capacity allocation for exchanging packet based L2+ flows is similar in its (in)efficiency as it would be to have a constant BW, pre-scheduled L3 packet transmission between a given IP node and each other IP destination/source that the given node may at times have any L4+ packets to exchange with: If the L3 send (receive) capacity of a given node would be e.g. 1Gbps, and it would have say 1,000 other IP end-nodes that it communicates with at times, to work analogous to the prevailing L1/0 technologies, it should statically partition its 1Gbps of IP communications capacity to 1000 fixed BW IP tunnels (of 1kbps each in average) -- although the node at any given time will have active L4+ communications sessions with only a few of its source/destination IP nodes and is receiving or transmitting a packet to/from only one of IP source/destination, and could thus do it at 1Gbps. Static L3 capacity allocation in this scenario would thus reduce the effective inter-node throughput by a rate of 1000:1 or such, achieving but 0.1% of the possible throughput.

Though the above example may not be directly applicable for all inter-node IP communications scenarios, it nevertheless serves to illustrate the fundamental reason why IP connectivity had to be made packet based, and **the drastic loss of efficiency that a static partitioning of server layer resources creates for packet oriented (chatty, bursty) applications -- whether apparent or hidden -- and in equal manner, whether in case of carrying L4+ traffic over L3 connectivity, or L2+ packets over L1/0 connections**.

Common examples of the prevailing static partitioning of physical network capacity allocation, even when carrying variable BW traffic between a group of packet-switching end-points such as IP nodes/routers include:
- Use of separate fibers (on same corridors);
- Use of separate WDM channels on same fiber, whether of not using a packet based or carrier signal such as Ethernet;
- Use of separate (non-adaptive) TDM channels on same WDM wavelength or fiber strand, whether based on traditional TDM protocols such as SDH/SONET or emerging techniques such as OTN, including ODU flex and such.

So why, after several years of increasingly packet based network application traffic, the network physical layer protocols are based on pre-provisioned, (semi-)fixed BW connections, instead of packet-by-packet adaptive connectivity?

As a reaction, one could try to argue that adaptive L1 is not needed when using L2 'soft-circuits' (e.g.. MPLS-TE LSPs, or PW, Ethernet equals) to allow L3 source-destination flows to more flexibly share L1/0 connection capacities. However, such argument is illogical at least for the following reasons:
- Unlike on inter-router L1/0 circuits, on L2 packet-switched paths the traffic encounters intermediate packet-switching hops, each of which increases delay, jitter [1] [2] and packet loss probability, as well as packet processing and related overhead (power, OAM) costs, degrading network cost-efficiency and scalability.

---

[1] The major reasons for jitter on L2 networks are not so much the variations in packet-processing delays at intermediate nodes, but the fact that when transmitting packets between L3/2 nodes over non-channelized L1/0 connections, at most one packet may be transmitted at any given time, while all other packets, no matter how short and how high priority, directed over the same L1/0 interface will have to wait until the present packet, no matter how long and how low priority, being sent has been transmitted in its entirety over the shared L1 port. For instance, a 64kB jumbo frame on 1Gbps interface blocks all other traffic, regardless of priority, for 0.5ms, per each such shared transmission interface between the source and destination nodes. This should be compared against typical business network SLA jitter specifications, e.g. max 2ms total, meaning that even just four such shared 1GbE switch interfaces between the customer nodes have the potential to reach the maximum jitter tolerance, leaving no budget for jitter caused by packet processing, switching and congestion buffering, the latter of which can be substantial. In practice, the consequence is that, unlike L1 switched networks, L2 switched networks can only have a limited, such as maximum of 3, aggregation/switching nodes between the customer interfaces, and that L2 switched networks will thus end up having to be expensively over provisioned (kept at low average utilization) for any delay and jitter sensitive traffic, incl. the higher revenue applications such as voice, video conferencing, IPTV, multimedia and business applications such as data replication and transaction processing.

[2] Note that with L1 channelization, multiple packets can be sent in parallel over same L1 ports, and with adaptive L1 channelization, the highest priority and highest load packet flows getting most bandwidth dynamically.

- Since most higher revenue applications require minimum throughput and maximum jitter and latency guarantees, L2 'soft circuits' need physical BW reservations, which all other traffic must honor at all times whether or not these static BW reservations are being used. ***Thus L2 'virtualization' of L1/0 resources amounts to nothing more elegant than fixed BW allocation, just done at packet-layer rather than via L1 TDM or L0 WDM*** (and thus requiring complex packet-layer QoS policing at network nodes).

L2 'soft circuits', just like traditional non-adaptive L1 circuits, thus impose artificially low BW caps for the packet traffic flows, unavoidably reducing the available throughput to/from any given L3 node.

To take the discussion to a more concrete level of switching and transport network design, let's consider a case of a routing node with an assumed 10Gbps full duplex interface throughput (this could be any other bit rate just as well) over which the router exchanges traffic with eight other peer routers (and this again could be any other number typical for router adjacencies). While the 10Gbps router interface is L1-unchannelized 'fat pipe', it is partitioned at packet-layer to traffic engineered L2 'soft circuits' (e.g. LSPs, called here "flows"), and for the sake of simplicity of illustration we assume there to be one such flow per each of the 8 peer routers (in each direction). Fig. 1 shows this network diagram studied.
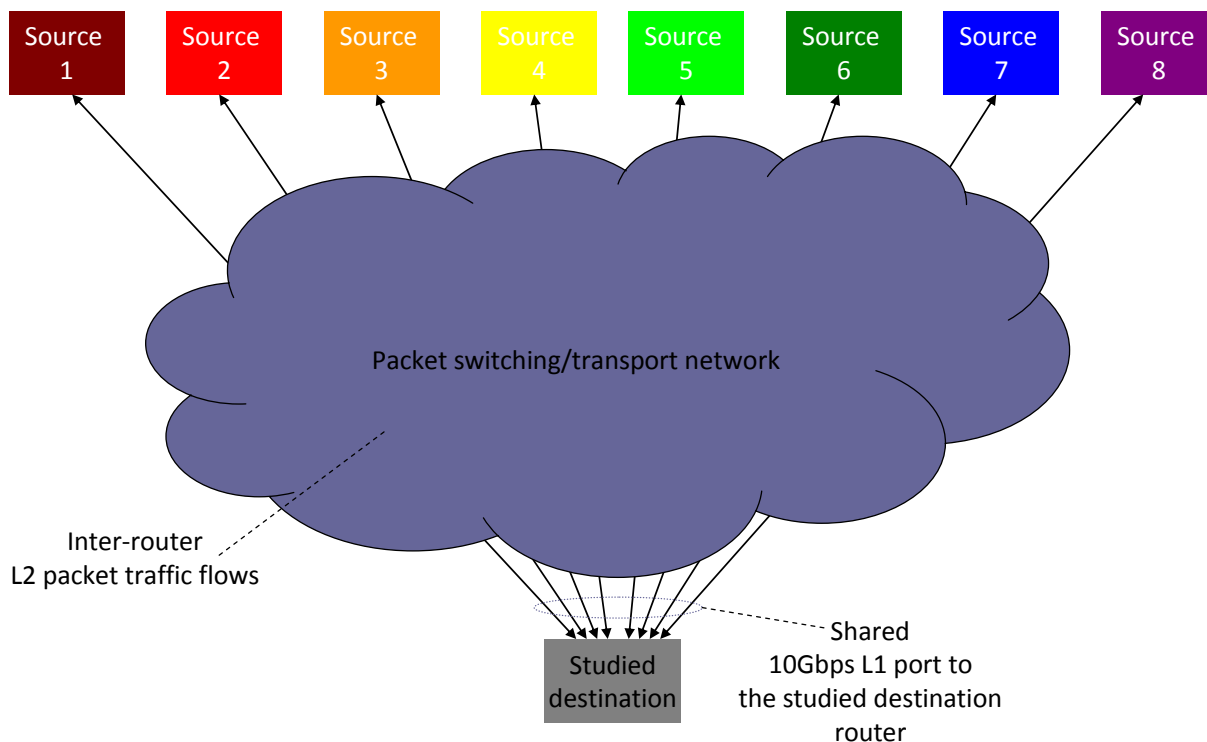
Fig. 1. A generic diagram for analyzing the packet transport network partitioning dilemma. Arrows represent the L2 flows that can be traffic engineered for committed bandwidth rates, to share the network capacity toward the studied destination.

In Fig. 1, how should the inter-router network BW be allocated[3], assuming that the external application generated traffic loads for the inter-router network can be managed (e.g. via packet-layer QoS policing) at most on 1 second intervals?

---

[3] Though we here focus on BW allocation from multiple sources toward a chosen destination packet-switch/router, the analysis that follows would be similar when studying the transmit direction BW allocation from a given packet-switching source node toward a number of destinations.

To study that question using a specific (but representative and practical scenario), let us further assume for the case of Fig. 1 that the 1s-average BW quotas of the L2 traffic flows (#1 - #8) from the source routers 1 through 8 respectively over the shared 10Gbps capacity to the studied destination are provisioned per Fig. 2.
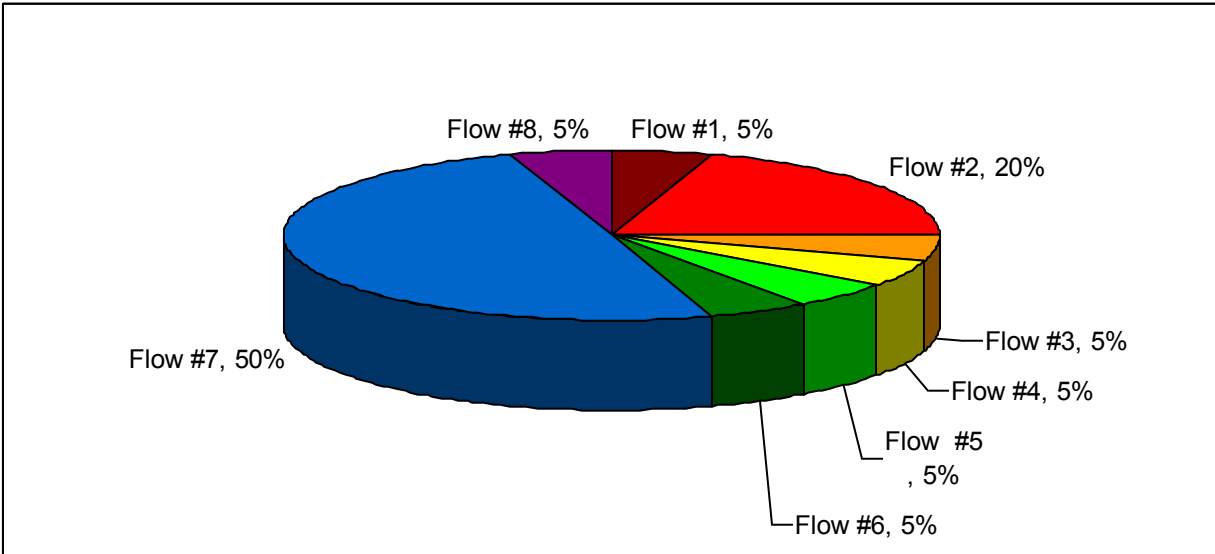
Fig. 2. Example partitioning of the L1 network capacity toward the studied destination router 10Gbps port among the L2+ traffic flows from the eight source routers sharing that physical port.

Thus, over time windows such as 1 second, it may be possible to control the applications feeding the inter-router flows to not exceed their committed rates, such as 0.5Gbps for flow #3 from the source router 3 to the destination of Fig. 1. However, the real traffic materializes as individual packets that are always transmitted at full physical port bit rates, which in the case of Fig. 1 where the routers are assumed to have 10Gbps ports to the packet transport cloud, means that each of the flows from source routers of Fig 1, at any given time instance, either would have a packet to transmit at the full link rate of 10Gbps or otherwise will idle at 0bps.

Moreover, the collective packet traffic flows from the individual users and applications constituting the inter-router flows are not directly controllable by the network, and therefore at finer time granularity, e.g. at 100ms windows, the optimal (revenue maximizing) allocation of network bandwidth [4] [5] [6]

between the individual flows of Fig. 1, while averaging to their committed rates over time per Fig. 2, may look like the example shown in Fig. 3.

---

these techniques moreover are known to demonstrably work in test networks.

[5] A key reason why adaptive L1 channelization per above references is able to optimize network physical BW allocation according to packet byte traffic load variations of the flows, while honoring all minimum flow BW quotas (whenever actually needed by flow traffic loads), is the patented, destination node driven distributed hardware logic algorithms (per above patent references) that keep the network control plane in byte-timeslot accurate sync with the dynamically channelized data plane.

[6] One of the reasons why conventional L3/2 traffic engineering methods cannot do this packet traffic load adaptive optimization is that, being software based and thus non-synchronous with data plane, as well as not destination node driven, they at the source nodes lack the foresight of when and how much they could exceed their committed BW quota without blocking other flows. Thus, conventional L3/2 TE techniques have to rely on statically provisioned flow rate limiting, unnecessarily blocking flow bursts as well as keeping reserved capacity idling when the flow traffic did not materialize (illustrated in Fig. 4).
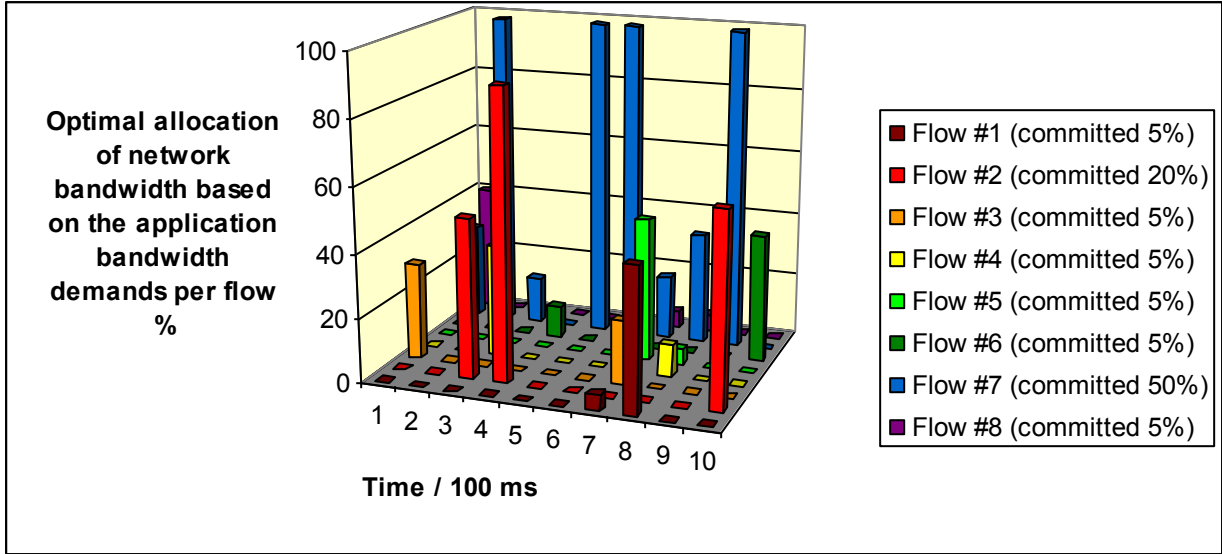
---

[4] Note that it is technically possible to keep network physical BW allocation optimized per packet by packet load variations, without any significant offsetting cost factors, per specifications of US patent application 12/363,667 and related patents 7,558,260 and 7,333,511;

Fig. 3. An example of what the optimal allocations of bandwidth for the flows of Fig:s 1 and 2 could be at timescales finer than what QoS policing can operate.

To analyze the impact of the static partitioning of the inter-router network BW allocation on the network on-time throughput i.e. its revenue generation potential, let's take a detail look at how the real per-flow application BW demands and their committed network BW allocations can compare over a slice of time such as 100ms, e.g. the 600ms to 700ms time window from Fig. 3, presented below in Fig. 4.
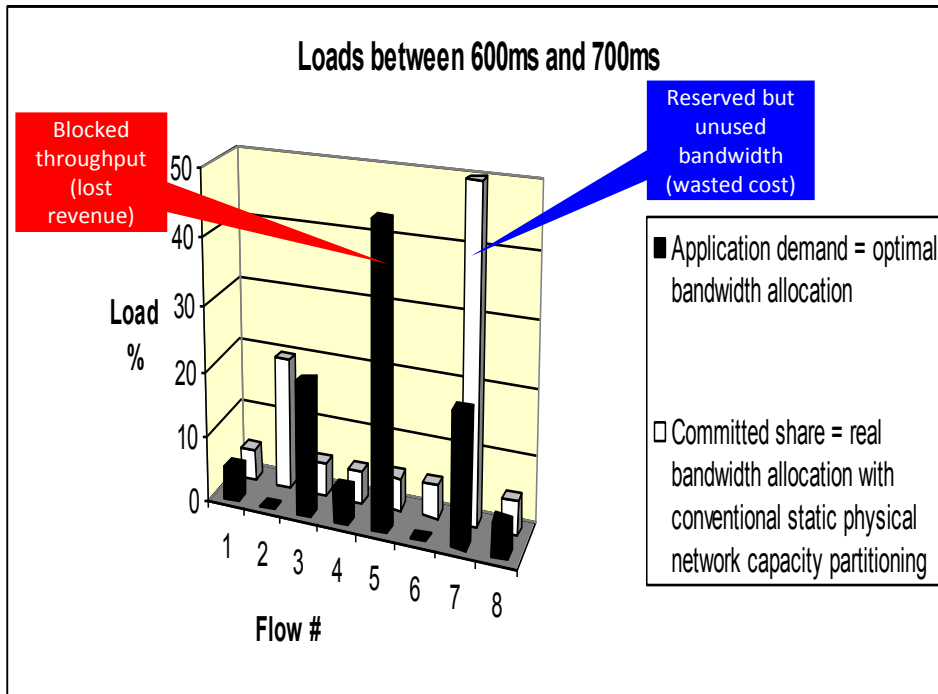

Fig. 4. A representation of a continuously occurring mismatch between actual application demands for and average-load based allocation of inter-router BW in case of (semi-)static partitioning of physical network BW: While static partitioning of L1 BW can be seemingly correct vs. average L2+ flow traffic distribution, at actual packet timescales, any static BW allocation will be completely off vs. actual packet by packet loads effectively all the time, directly eroding service provider profitability.

The case of Fig. 4 shows that the prevailing static partitioning of physical network BW allocation severely reduces network profitability, both through i) unnecessarily limiting network revenue generating service volume via blocking traffic flows beyond their average committed quotas and ii) requiring multiple times more spending on the network capacity due to the under-utilized flow BW reservations than what would be necessary (with packet-by-packet adaptive L1 BW allocation) for the achieved throughput. *In practice, with adaptive L1 channelization, e.g. 40Gbps of interface capacity can thus support more revenue generating service volume than a statically partitioned 100Gbps interface.*

To round up the discussion and to provide context for above illustrations:

- For clarity of illustration, there is always assumed to be enough application traffic, e.g. data back-up etc. background transfers, available to utilize the studied network capacity. In reality, there could be more (oversubscription) or less (under-utilization), but in any case a static partitioning of the transport network BW would reduce network profitability for the reasons i) and/or ii) above, just in different proportions, but with same end result.
- Applications are assumed to not tolerate delays (or jitter) more than 100ms i.e. the network cannot just buffer a burst of traffic to be delivered over time on a lower BW path; such delayed traffic would be same as lost from the applications' point of view.
  - The study above could be taken to ever finer time granularity, e.g. to microsecond level (e.g. a 1kB packet consumes roughly 1 microseconds at 10Gbps) where software based QoS policing is not feasible in any manner, still producing similar insights.
- At any instant in time, there is always some allocation of physical network BW among the inter-router flows that maximizes the total network utility, e.g. its revenue for the network service provider, and while the averaged BW allocation can over time windows such as 1 second or longer be fairly static (from a 1 second window to the next), at the packet-by-packet granular time windows (microseconds) the revenue-maximizing optimum BW allocation is highly dynamic, as it is based on the application demands for network BW, which always are highly time variable as they are made of intermittent full link rate packet transmissions alternating with idle periods (even when over longer time periods the flows can be policed to stay within their committed BW levels).

Thus, in order to reach the next level of network efficiency necessary for a profitable business case for emerging high-BW and high-QoS, but relatively low revenue/bit, services, the packet-load adaptive network allocation efficiencies have to be taken all the way to the network physical layer. Moreover, BW optimization at network physical layer eliminates the need for the complex packet-layer stat-muxing done in current networks to gain BW efficiencies, and which when (as typically is the case) done across different client applications causes major everything-affects-everything type QoS and security problems. The transition from prevailing (semi-)static to traffic load adaptive physical layer BW allocation thus appears to be one of those efficiency gain measures that does not increase complexity elsewhere, but in fact overall simplifies network implementations and services delivery.

Above analysis should prompt decision makers responsible for network economics to ask: ***Is there a valid business case for continuing to spend on networks based on non-adaptive physical layer capacity allocation?*** The answer (for same reasons why IP had to be packet based) is 'no', provided that packet traffic load adaptive physical layer connectivity is technically feasible--which in turn raises the question: ***Is overhead-free, realtime adaptive bandwidth L1 connectivity technologically possible?*** A known working implementation exists: Adaptive-Mesh by OCS – *http://www.ocsipholding.com*

Summary:

- Static partitioning of physical layer network bandwidth is but an artifact of traditional network equipment limitations; however, technically or economically fixed bandwidth inter-node connectivity over shared lower layer media makes no more sense at L1 for packetized L2+ traffic, than it would at L3 (IP) for dynamic L4+ applications.
- As packet traffic load adaptive network physical layer connectivity is now a technological reality, any further spending on non-adaptive physical layer based networks (whether Eth/ SDH/OTN/WDM) is directly away from the profits that could be made with adaptive L1 optimization.