ONTC PRISM Newsletter

Dear ONTC Members,

A warm welcome from the Editorial desk of the ONTC's newsletter "Prism". It is our pleasure to bring to you this second edition of PRISM – ONTC's quarterly newsletter. The current times have been challenging for both industry and academia as well as exciting from an innovations perspective. New technologies must be innovated to be able to meet the needs of telecom users. Telecommunications is slated to be a 2.7 trillion USD industry by 2013 and to get there, a whole new set of innovations are required. 100 Gigabit Ethernet and data centers are two areas that would be critical in terms of innovation and business case to get our industry to those levels.

The current month is also an important one in terms of dissemination of technology in our community. The industry just concluded the IEEE 802 Standards Plenary in Orlando – marking 30 years since the first 802 standardization work was adopted. The academic conference Infocom also concluded in San Diego this week. From the optical communications industry perspective, our most important event, the OFC conference starts Sunday, March 21. This year's conference is dedicated to Dr. Charles Kao, industry pioneer of the optic fiber and last year's Nobel Prize winner.

In this issue of Prism we bring to you 3 articles on a wide range of topics. The first article is by John Ambrosia, Chair of the IEEE P802.3ba or 100Gbps Ethernet working group. As John puts it to the community, the 802.3ba is on its home stretch with it being finalized sometime in the next 6-8 months. This is perhaps the most important standard for this decade and it will mark an important step in our industry. As Bob Metcalfe put it eloquently in his 2008 OFC keynote, Ethernet and SONET/SDH have been playing tic-tac-toe, with Ethernet multiplying itself 10x and SONET/SDH 4x in each new avatars. At 10Gbps, the two technologies have rendezvoused together thereby bringing the telecom and the computer world at cross roads. In its next avatar Ethernet makes a gigantic leap to 100Gbps and this work will likely become the future of metro and access transport and spilling into the long-haul. John Ambrosia in his article talks about the evolution of 100Gbps Ethernet and the different thoughts, approaches etc. to making this technology happen. From the editorial desk of Prism, we do invite folks from the TDM world to send in articles on their thoughts to the next generation of high-speed transport. In particular, we seek folks to talk about OTN technologies and how to map OTN at 100Gbps.

The second article is by Weiqiang Sun and his group at Shanghai Jiaotong University in China and it talks about exploring the potential of GMPLS for future application. GMPLS has some interesting potential and along with ASON, forms the backbone of the network-wide control plane. Whether this technology can be adapted to newer applications, like clouds, IT virtualization etc., is something that we have to only wait and see.

The third article is by Christian Rothenberg from the University of Campinas in Brazil. This article is about Re-architecting cloud data center networks. The article compares different approaches to data center interconnection. Data centers from an IT application perspective are becoming very important and both CAPEX as well as OPEX issues need to be considered. Apart from these functionality within the data center environment is important, and this is where the article highlights the different approaches that have been presented in literature. From the editorial desk, we do ask industry to also communicate on their approaches to the data center with the hope that there is good match between the requirements of the industry and what academia has to offer.

The editorial desk also from this issue onwards highlight events, call for papers and new standards work – all of which would be relevant information for our readers.

We also do hope readers would send in their thoughts on how to make Prism better – we would be happy to publish their messages even if all of these cannot be adopted at the same time!

The next issue of PRISM would have its primary focus on large collaborative projects. We hence invite prospective authors to cover ongoing or recently concluded large collaborative projects. Interested authors who would like to send in an article, are welcome to send a 4 page (single column, 10 point font, with all one-inch margins) to submissions@ontc-prism.org. The deadline for reception of articles is May 15, 2010.

On behalf of the TAB we are thankful to the IEEE Communication Society as well as to the ONTC officers Byrav, Suresh, Admela and Dominic who have supported us in making this newsletter happen. A new member joins the TAB – Wael William Diab, from Broadcom. Wael is also actively involved in the standards.

It is our hope that the newsletter would bring the community together and identifying areas of growth and common interest.

Ashwin Gumaste, IIT Bombay.

Message Board

Standardization:

IEEE SIEPON	April 14-16, Shanghai	http://grouper.ieee.org/groups/1904/1/
IEEE Interim Meeting (802)	May 24-28, Geneva	http://www.ieee802.org/1/meetings/#may10gen
IETF	March 21-26, Anaheim	http://www.ietf.org/meeting/77/index.html
IEEE Plenary	July, 11-16, San Diego	http://ieee802.facetoface-events.com/future
ITU SG15		http://www.itu.int/ITU-T/studygroups/com15/index.asp

Academic Conferences

IEEE/OSA OFC 2010	Conf March 21-26. San Diego	http://www.ofcnfoec.org/
IEEE Globecom 2010	CFP March 31 Conf Dec 6-10. Miami	http://www.ieee-globecom.org/
IEEE LCN 2010	CFP April 12. Conf Oct 11-14 Denver, CO.	http://www.ieeelcn.org/
IEEE ICC 2010	May 23-27 , South Africa	http://www.ieee-icc.org
IEEE ICC 2011	CFP Sept. 7: June 5-9, 2011, Kyoto, Japan.	http://www.comsoc.org/confs/icc/2011/index.php
IEEE ANTS 2010	CFP July 15: Conf: Dec 15-17, Bombay, India	www.ieee-ants.org
HPSR 2010	Conf June 14-16, UT-Dallas, Richardson, TX	http://opnear.utdallas.edu/activ/hpsr2010/index.html
ICTON 2010	CFP March 31, Conf June 27-July 1, Munich, Germany.	http://www.nit.eu/konf/icton/2010/
Infocom 2011	CFP July 30, Conf April 10-15, Shanghai, China.	http://www.ieee-infocom.org/2011/

We would be happy to include more conferences in the above list, if readers email editor@ontc-prism.org a CFP of the conference. The conference must be at least technically supported by ONTC or ComSoc to be included in the list above and follow the ONTC endorsement policy.

Key journals reporting results in the optical networking area:

IEEE/OSA Journal of Optical Communication and networks (JOCN)
http://www.opticsinfobase.org/jocn/journal/jon/author.cfm
IEEE/OSA Journal of Lightwave Technology: http://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=50
IEEE/ACM Transactions on Networking: http://www.ton.seas.upenn.edu/
IEEE Communications Magazine: http://mc.manuscriptcentral.com/commag-ieee
IEEE Network: http://dl.comsoc.org/ni/

The Next Generation of Ethernet

John D'Ambrosia
Senior Scientist, CTO Group, Force10 Networks
Chair, IEEE P802.3ba 40Gb/s and 100Gb/s Ethernet Task Force

Introduction

In December of 2007 the IEEE Standards Association approved the formation of the IEEE P802.3ba Task Force, which was chartered with the development of 40 Gigabit Ethernet (40 GbE) and 100 Gigabit Ethernet (100GbE). The shift from the traditional Ethernet approach of 10x leaps only in speed reflected the differing growth rates between computing / server applications and network aggregation applications. Networking applications, whose bandwidth requirements are doubling approximately every 18 months, have greater bandwidth demands than computing applications, where the bandwidth capabilities for servers are doubling approximately every 24 months. The impact of this difference in bandwidth growth is illustrated in Figure 1.

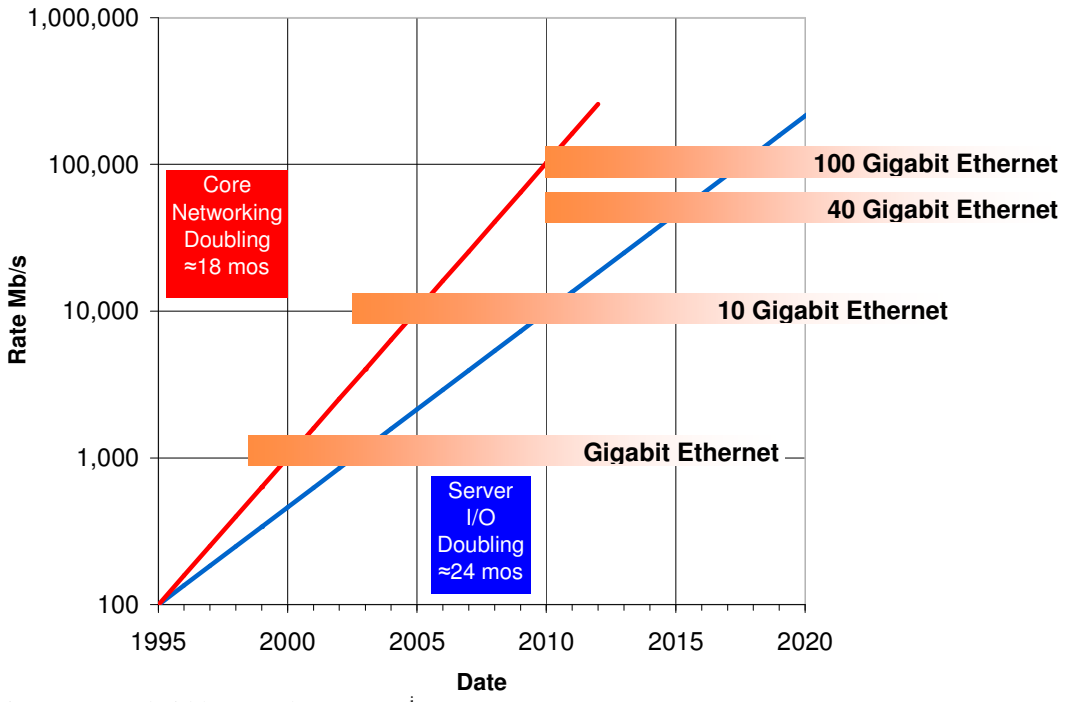

Figure 1- Bandwidth Growth Forecasts[i]

The physical layer specifications selected for each rate target the distance requirements for the intended applications. For computing and server applications at 40 Gb/s, there are three distance objectives: at least 1m over a backplane; at least 7m over a copper cable assembly; and at least 100m on OM3 multimode fiber (MMF). For core networking and aggregation applications at 100 Gb/s, there are four distance objectives: at least 7m over a copper cable assembly; at least 100m on OM3 MMF; at least 10km on single-mode fiber (SMF); and at least 40km on SMF. Additionally, the need for the transmission of Ethernet over optical transport networks (OTN) was recognized, and an objective to provide appropriate support for OTN was adopted.

Physical Layer Specifications

Table 1 provides a summary of the different physical layer specifications that were ultimately targeted by the task force with their respective port type names.

Port Type	Reach	40 GbE	100 GbE	Description
40GBASE-KR4	At least 1m backplane	√		4 x 10 Gb/s
40GBASE-CR4	At least 7m cu cable	√	√	"n" X 10 Gb/s

100GBASE-CR10				
40GBASE-SR4 100GBASE-SR10	At least 100m OM3 MMF	√	√	”n” x 10 Gb/s (Use of Parallel Fiber)
40GBASE-LR4	At least 10km SMF	√		4 x 10 Gb/s
100GBASE-LR4	At least 10km SMF		√	4 x 25 Gb/s
100GBASE-ER4	At least 40km SMF		√	4 x 25 Gb/s

Table 1 – Summary of IEEE P802.3ba Physical Layer Specifications

Below is a description of each of the different PHYs in the 40GbE / 100GbE family:

- The 40GBASE-KR4 PMD supports backplane transmission over four channels in each direction at 40 gigabit per second. The effective date rate per differential pair is 10 gigabit per second. It leverages the 10GBASE-KR architecture, already developed channel requirements and PMD.

- The 40GBASE-CR4 / 100GBASE-CR10 PHY supports transmission of 100GbE over 7m of twinaxial copper cable across four differential pair in each direction for 40GbE and across ten differential pair in each direction for 100GbE. The effective date rate per differential pair is 10 gigabit per second. The PHY leverages past work from the Backplane Ethernet project by utilizing the 10GBASE-KR architecture, channel budget, and Physical Medium Dependent Sub-layer. (PMD).

- The 40GBASE-SR4 / 100GBASE-SR10 PHY is based on 850nm multimode fiber (MMF) optical technology and supports transmission of 40GbE across four parallel fibers in each direction and 100GbE across ten parallel fibers in each direction. The effective date rate per lane is 10 gigabit per second. Optical Multimode 3 (OM3) grade fiber, which has an effective modal bandwidth of 2000MHz/km, can support reaches up to at least 100m, while Optical Multimode 4 (OM4) grade fiber, which has a an effective modal bandwidth of 4700MHz/km, can support reaches up to at least 125m. An optional non-retimed electrical interface, CPPI (100 Gigabit Parallel Physical Interface), has been designed to support chip-to-module applications, and is optimized based on the 40/100GBASE-SR physical layer specification.

- The 40GBASE-LR4 PMD is based on 1310nm, Coarse Wave Division Multiplexing (CWDM) technology and supports transmission over at least 10km over Single-mode fiber (SMF). The grid is based on the ITU G.694.2 specification, and the wavelengths used are 1270, 1290, 1310, and 1330nm. The effective data rate per lambda is 10 gigabit per second, which will help maximize re-use of existing 10G PMD technology. Therefore, the 40GBASE-LR4 PMD supports transmission of 40 Gigabit Ethernet over 4 wavelengths on each SMF in each direction.

- The 100GBASE-LR4 PHY is based on Dense Wave Division Multiplexing (DWDM) technology and supports transmission of at least 10km over a pair of single-mode fibers (SMF). The four center wavelengths are 1295nm , 1300nm, 1305nm, and 1310nm. The center frequencies are spaced at 800 GHz, and are members of the frequency grid for 100 GHz spacing and above defined in ITU-T G.694.1. The effective data rate per lambda is 25 gigabit per second. Therefore, the 100GBASE-LR4 PMD supports transmission of 100GbE over 4 wavelengths on a single SMF in each direction.

- The 100GBASE-ER4 PHY is also based on DWDM technology and supports transmission of at least 40km over a pair of single-mode fibers. The four center wavelengths are 1295nm, 1300nm, 1305nm, and 1310nm. The center frequencies are spaced at 800 GHz, and are members of the frequency grid for 100 GHz spacing and above defined in ITU-T G.694.1. The effective data rate per lambda is 25 gigabit per second. Therefore, the 100GBASE-LR4 PMD supports transmission of 100GbE over 4 wavelengths on a single SMF in each direction. To achieve the 40km reaches called for, it is anticipated that implementations may need to include semiconductor optical amplifier (SOA) technology.

The IEEE P802.3ba Architecture

Figure 2 illustrates the overall IEEE P802.3ba architecture that supports both 40 Gigabit Ethernet and 100 Gigabit Ethernet. While all of the PHYs have a Physical Coding (PCS) sub-layer, Physical Medium Attachment (PMA) sub-layer, and a Physical Medium Dependent (PMD) sub-layer, only the copper cable (-CR) and backplane (-KR) PHYs have an Auto-Negotiation (AN) sub-layer and an optional Forward Error Correction (FEC) sublayer.

This architecture is the true innovation in the standard, as it is a flexible and scalable architecture that can support both 40GbE and 100GbE, the multiple PHYs being developed as part of the IEEE P802.3ba standard, as well as physical layer specifications that may be developed by future task forces. The keys to this flexible architecture reside in the lane distribution scheme of the Physical Coding (PCS) sub-layer and the multiplex / de-multiplex functionality of the Physical Medium Attachment (PMA) sub-layer.

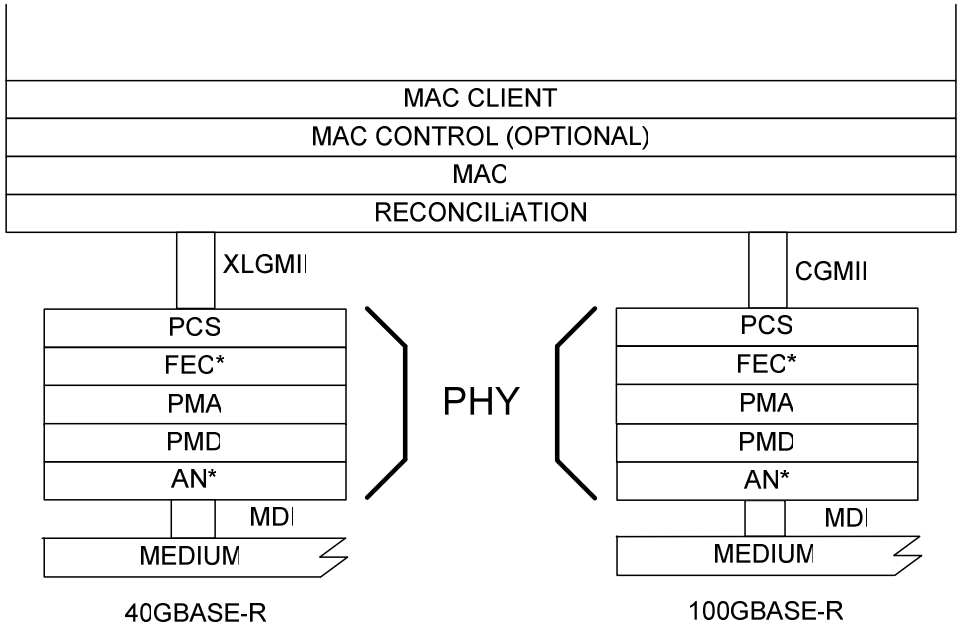

Figure 2 - IEEE P802.3ba Architecture

The PCS sub-layer couples the respective Media Independent Interface (MII) to the PMA sub-layer. The aggregate stream coming from the MII into the PCS sub-layer undergoes the 64B/66B coding scheme that was used in 10 Gigabit Ethernet. Using a round robin distribution scheme, 66-bit blocks are then distributed across multiple lanes, referred to as "PCS Lanes," each with a unique lane marker, which is periodically inserted. This is illustrated in Figure 3. For 40GbE there are four PCS lanes and for 100GbE there are twenty PCS lanes.

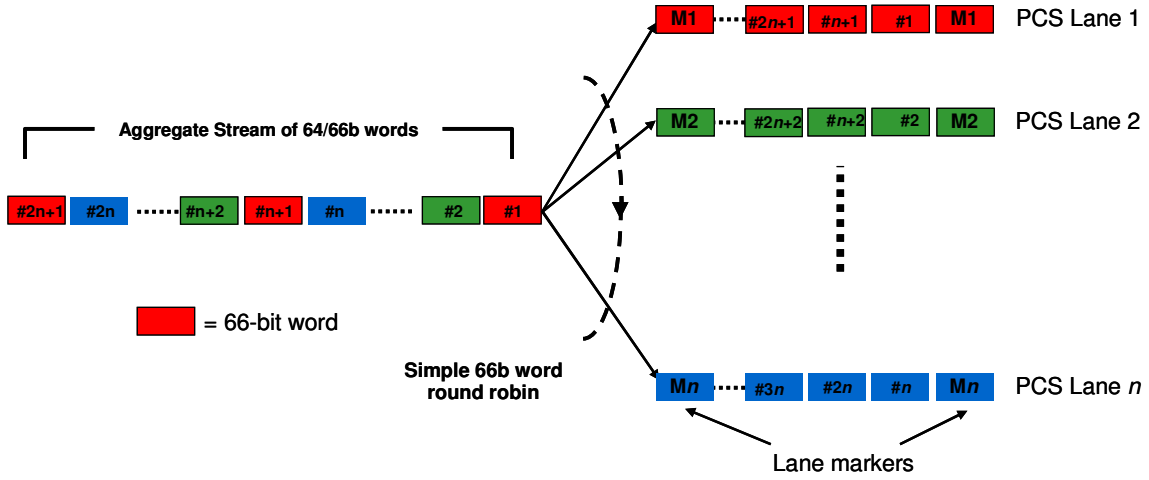

Figure 3 – PCS Lane Distribution Concept [ii]

The PMA sub-layer, which is the intermediary sub-layer between the PCS and the PMD, provides the multiplexing function that is responsible for converting the number of PCS lanes to the appropriate number of lanes or channels needed by the PMD.

Additionally, the PMA sub-layer plays an integral role in the implementation of the respective Attachment Unit Interface (AUI), which is an optional physical interface based on 10Gb/s electrical signaling, as a PMA sub-layer will exist on both sides of the respective AUI. These interfaces are used for portioning the system design, and are used for chip-to-chip and chip-to-module applications. For 40 Gigabit Ethernet, the AUI is called "XLAUI" ("XL" is the Roman numeral for 40), and there are four transmit pairs and four receive pairs.

Conclusions

While the upcoming ratification of the IEEE P802.3ba standard, scheduled for June of 2010, will be a major accomplishment, it is anticipated that the 100GbE family will undergo further development. For example, at the time of writing this column the IEEE 802.3 has formed a new study group that will examine the need for the development of a 40Gb/s Ethernet single-mode fiber PMD that is optimized for client applications in the carrier environment. It is also anticipated that future developments in electrical and optical signaling will be applied to the IEEE P802.3ba architecture, which was developed for just this purpose. Furthermore, the IEEE P802.3ba architecture was conceived as an architecture that could scale to support future speeds of Ethernet, and as noted earlier, based on the observation that bandwidth requirements for core networking applications are doubling approximately every 18 months, core networking applications are forecasted to need Terabit Ethernet in 2015.

Exploring the potentials of GMPLS for future applications

Weiqiang Sun, Yaohui Jin, Wei Guo and Weisheng Hu
Shanghai Jiao Tong University

GMPLS at a glance

Generalized Multiprotocol Label Switching, or GMPLS, is a suite of protocols to enable automated resource discovery, automated service provisioning and automated failure recovery. Driven by the benefit of improved network reliability (through protection or fast failure recovery) and reduced network OPEX, an increasing amount of GMPLS enabled networks are now being deployed in metro area and even in national backbones. In the short run, the deployment of such networks will enable network operators to provide new value added services such as Bandwidth on Demand (BoD) with a reduced OPEX. In the long run, GMPLS networks have the potential of carrying a big variety of services. In this article we try to highlight the current status in the research community and challenges for such a trend.

Examples of applying GMPLS for novel services

Starting from 2001, the time when the key concepts and features are gradually being standardized, a considerable number of efforts have been seen in the area of putting GMPLS into advanced networking/service environments. This is exemplified by the various research programs in the US, Europe and Asia (see [1] and [2] for an overview). In such programs, GMPLS has not only been used as a way to reduce management complexity and increase reliability, like the industry is doing right now, but also it is used as a new way for service provisioning. For example, the GMPLS control plane is often integrated with the application to realize seamless on-demand circuit provisioning, so that dynamic data intensive applications may be served with dedicated bandwidth pipes in an efficient manner.

In a more recent effort, Vincent W. S. Chan et al. argue that GMPLS can be one of the enabling technologies for realizing large scale dynamic wavelength services to the masses [3]. By combining statistical multiplexing in access networks and GMPLS provisioned quasi-static light-paths in core networks, the researchers claim that future all optical networks may be able to serve up to 10^7 users in a metro area network (MAN), each at a data rate that a wavelength can provide.

In another work, W. Sun et al. demonstrated cross-layer circuit provisioning to satisfy different application needs [2]. An application, instead of an end system as a whole, can be the entity to request a light-path. In the demonstration, the initiation of light-path setup is seamlessly integrated with the TCP 3-way handshake process, resulting in a high performance yet transparent transmission service. When implemented in a private network, this way of circuit provisioning may be an interesting candidate for future demanding applications, data center inter-connection being one of them.

Dynamic provisioning performance of GMPLS

Before GMPLS can be fully utilized to its potential, it is important that we have good ways to characterize and measure its performance. What seem obvious performance measures include LSP dynamic provisioning performance, failure recover performance, singling and routing scalability etc. What seems less obvious is the consistency between control plane and data plane. Standardization on some of these performance measures has started [4-7] and still has a long way to go.

From 2004, we performed a number of tests on several SDH based GMPLS networks/testbeds. The main finding of the tests include: i) Implementation of GMPLS is getting mature over time. We find that the delay to setup a 2-hop LSP decreased steadily from the initial 600+ milliseconds to less than 100 milliseconds. ii) LSP setup delay exhibits high variance, especially under high traffic load. This may have very important implications to applications when LSP setup/release is very dynamic. iii) LSP setup delay can be highly implementation dependent. Although randomness is observed in all measurements, we find that different implementations may exhibit totally different statistical behavior. iv) the delay in cross connection programming and data plan/control plane interaction can be a dominating part in the overall LSP setup delay. Small but non-zero setup failure probability is observed in many tests.

These results indicate that although current state of GMPLS impelmentations may well meet the current provisioning requirements, it is far from enough for future applications. As both the interest and understanding in this topic are increasing in the community, it is expected that more issues will be identified. Further engineering of existing protocols and implementations will be necessary for significant performance improvement.

This research is supported in part by STCSM under grant 09QA1403200 and NSFC.

References

[1] I. W. Habib et al., "Deployment of the GMPLS control plane for grid applications in experimental high-performance networks," IEEE Commun. Mag., vol. 44, no. 3, pp.65-73, 2006.
[2] W. Sun et al., "A cross-layer optical circuit provisioning framework for data intensive IP end hosts," IEEE Communications Magazine, vol. 46, no. 2, pp.S30-37, 2008.
[3] Vincent W. S. Chan et al., Optical Flow Switching, http://www.mit.edu/~medard/papersnew/WOBS Final.pdf, accessed Jan.2010
[4] W. Sun et al., "Label Switched Path (LSP) Dynamic Provisioning Performance Metrics in Generalized MPLS Networks," Internet draft, draft-ietf-ccamp-lsp-dppm-11.txt, Dec. 2009, work in progress.
[5] W. Sun et al., "Label Switched Path (LSP) Data Path Delay Metric in Generalized MPLS/MPLS-TE Networks," draft-sun-ccamp-dpm-00.txt, Internet draft, work in progress
[6] S. Poretsky et al., "Benchmarking Terminology for Protection Performance," Internet Draft, draft-ietf-bmwg-protection-term-07.txt, Nov. 2008, work in progress.
[7] R. Papneja et al., "Methodology for Benchmarking MPLS Protection Mechanisms," Internet draft, draft-ietf-bmwg-protection-meth-06.txt, Nov. 2008, work in progress.

Re-architecting the Cloud Data-Center Networks

Christian Esteve Rothenberg

University of Campinas (Unicamp) – Brazil
chesteve@dca.fee.unicamp.br

Large-scale Internet data centers (DC) are empowering the new era of *cloud computing*, a still evolving paradigm that promises infinite capacity, no up-front commitment and pay-as-you-go service models. Ongoing research [3] towards providing low-cost powerful *utility computing* facilities includes large-scale (geo)-distributed application programming, innovation in the infrastructure (e.g., energy management, packing), and re-thinking how to interconnect thousands of commodity PCs. In this article, we focus on the latter and review developments that are taken place in architecting data center networks (DCN) to meet the requirements of the cloud.

Introduction - In contrast to traditional enterprise DCs built from high-prize "scale-up" hardware devices and servers, cloud service DCs consist of low-cost commodity servers that, in large numbers and with appropriate software support (e.g., virtualization), match the performance and reliability of traditional approaches at a fraction of the cost. However, the networking fabric within the data center has not evolved (yet) to the same levels of commoditization [1]. Today's DCs use expensive enterprise-class networking equipment that require tedious network and IT management practices to provide efficient Internet-scale data center services. Consolidated on converged IP/Ethernet technologies, current DCNs are constrained by the traditional L2/L3 hierarchical organization which hampers the *agility* to dynamically assign services provided by virtual machines (VM) to any available physical server. Moreover, IP subnetting and VLAN fragmentation end up yielding poor server-to-server capacity even when relying on expensive equipment at the upper layers of the hierarchy [5].

Resource usage in the highly virtualized Cloud is very dynamic due to the nature of cloud services, causing unpredictable traffic patterns [11] for which common enterprise traffic engineering practices or intra-domain networks are not well suited and often result in over-subscription rates as high as 1:240 [4]. While not critical in enterprise networks, two main limitations of traditional Ethernet adversely affect its use in DCs: (1) scalability limits of ARP-broadcasting-based bridged spanning tree topologies; and (2) means to alleviate congestion without increasing latency. As a result, Ethernet-based store and forward switching potentially cause unacceptable high latencies in addition to dropped or reordered packets and excessive path failure recovery times even in the rapid versions of the spanning tree protocol (STP). An additional network management issue is concerned with the requirement of tweaking network path selection mechanisms to force the traffic across an ordered sequence of middleboxes (e.g., firewall, WAN opt., DPI, LB).

These and other shortcomings have made traditional Ethernet switching generally unsuitable for large-scale and high-performance computing needs of the cloud DCN. Industry efforts have been undertaken towards Data Center Ethernet extensions to provide QoS, enhanced bridging (IEEE 802.1 DCB), multipathing (IETF TRILL), Fibre Channel support, and additional Convergence Enhanced Ethernet (CEE) amendments. In the following, instead of delving into the market-driven incremental path of DC Ethernet solutions, we focus on the overarching requirements identified by industry and academia:

· Resource Pooling. The illusion of infinite computing resources available on demand requires means for elastic computing and agile networking. Hence, statistical multiplexing of physical servers and network paths needs to be pushed to levels higher than ever. Such degree of *agility* is possible (i) if IP addresses can be assigned to any VM within any physical server, and (ii) if all network paths are enabled and load-balanced.

· Scalability. Dynamically networking a large pool of location-independent IP addresses (i.e., in the order of millions of VMs) requires a large scale Ethernet forwarding layer. Unfortunately, ARP broadcasts, MAC table size constraints, and STP limitations place a practical limit on the size of the system.

· Performance. Available bandwidth should be high and uniform, independent from the endpoints' location. Therefore, congestion-free routing is required for any traffic matrix, in addition to fault-tolerance (i.e., graceful degradation) to link and server instabilities.

Re-architecting approaches - Traditional DCN architectures consist of a tree of L2/L3 switches with progressively more specialized and expensive equipment moving up the network hierarchy. Unfortunately, this architectural approach is not only costly but results in the network becoming the bottleneck for cloud DC applications. Recent research in re-architecting DCNs has spurred creative designs to interconnect PCs at large, including shipping-container-tailored designs with servers acting as routers and switches as dummy crossbars [6] or re-thinking the flatness of MAC Ethernet addresses in favor of location-based pseudo MAC addresses [8].

The architectural approach of so-called next generation DCNs can be classified as *server-centric* or *network-centric*, depending on where the new features are implemented. The common goal is to provide a scalable, cost-efficient networking fabric to host Web, cloud and cluster applications. Many of these applications require bandwidth-intensive, one-to-one, one-to-several (e.g., distributed file systems), one-to-all (e.g., application data broadcasting), or all-to-all (e.g., MapReduce) communications among servers. Non-uniform bandwidth among DC nodes complicates application design (i.e., requires notion of data locality) and limits the overall system performance, turning the inter-node bisection bandwidth the main bottleneck in large-scale DCNs. The principal architectural challenges of DCNs are L2 scalability, limiting broadcast traffic, and allowing for multipath routing.

The rationale behind *server-centric* designs is to embrace the "end-host customization" and leverage servers with additional networking features. In a managed environment like the DC, servers are already commonly equipped with modified operating systems, hypervisors and/or software-based virtual switches to support the instantiation of networked VMs. Under a server-centric paradigm, routing intelligence is (sometimes solely) placed into servers handling also load-balance and fault-tolerance. Servers with multiple network interfaces act as routers (aka P2P networks) and switches do not connect to switches and act as crossbars. The approach is to leverage commodity hardware to "scale-out" instead of high-end devices to "scale up". The resulting server-centric interconnection networks follow the principles of e.g., mesh, torus, rings, hypercubes or *de Bruijn* graphs, well-known from the high performance computing (HPC) and peer-to-peer (P2P) fields.

Two remarkable examples from Microsoft Research branches are VL2 [4] and Bcube [6]. VL2 describes a large Virtual Layer 2 Ethernet DCN that builds upon existing networking technologies and yields uniform high capacity and traffic fairness by virtue of valiant load balancing (VLB) to randomize traffic flows throughout a 3-tiered switching fabric using IP-in-IP encapsulation and Equal Cost Multi-Path (ECMP). In order to support agility, VL2 uses flat addresses in the IP layer and implements address resolution (mapping of application IP address to location IP address) by modifying the end systems and querying a scalable directory service. Bcube [6] is a shipping-container-tailored DCN design where switches only interconnect servers acting as routers. Scalable, high-performance forwarding is based on source routing upon a customized shim header (additional packet header) inserted and interpreted by end-hosts, which are equipped with multiple-cores and programmable network interface cards (e.g., NetFPGA). Container-based modular DCs emerge as an efficient way to deliver computing and storage services by packing a few thousand servers in a single container. The notable benefits are the easy deployment (just plug-in power, network, and chilled water), the high mobility, the increased cooling efficiency, and foremost the savings in manufacturing and hardware administration. Challenges include high resilience to network and server failures, since manual hardware replacement may be unfeasible or not cost-effective.

On the other hand, *network-centric* designs aim at unmodified endpoints connected to a switching fabric such as a Clos network, a Butterfly or a fat-tree topology. For instance, the fat-tree topology is very appealing because it provides an enormous amount of bisection bandwidth (without over-subscription) while using only small, uniform switching elements [1, 2]. The key modification happens at the control plane of the network, leaving end hosts and the switch hardware untouched, exploiting the availability of an open API such as OpenFlow [7]. Network customization through switch programmability requires network-wide controllers to install the forwarding tables of switches, resolve IP identifiers to network locators in response to ARP requests intercepted at edge switches, which are programmed for the desired line-speed packet flow handling actions (e.g., header re-writings). For instance, PortLand [8] is a native layer 2 network based on translating Ethernet MAC addresses into position-based "pseudo" MAC addresses. Network equipment vendors have already begun building switches from merchant silicon using multi-stage fat-tree topologies internally [2].

If we abstract the details of proposed DCN architectures (see examples in Table 1), in addition to design for failure (breakdown of servers and switches assumed to be common at scale), the following design principles can be identified:

· Scale-out topologies. Similar to how HPC clusters have been using two and multi-layer Clos configurations for around a decade because of their nice properties (e.g., blocking probability, identical switching elements), scale-out topologies of cloud DCN commonly follow a 3-tier arrangement with a lower layer of top-of-rack (ToR) switches, a layer of aggregation switches, and an upper layer of core switches. However, as long as they offer large path diversity and low diameter, other scale-out topologies can be considered (e.g., DHT-like rings, Torus).

· Separating Names from Locations. Identifier-locator split is not only an issue of Internet routing research (cf. IRTF RRG, LISP) to overcome the semantic overload of IP addresses, but is the common approach in DCNs to enable scalability and resource pooling of IP addressable services. The lack of topological constraints when assigning IP addresses to physical servers and VMs, enables cloud services to expand or contract their footprint as required. In this context, IP addresses are not meaningful for packet routing, which is commonly based on a revisited (usually source-routing-based) packet forwarding approach.

· Traffic randomization. The burstiness and the unpredictability of DC traffic patterns [11] requires routing solutions that provide load balancing for all possible traffic patterns, i.e., demand-oblivious load balanced routing schemes. Oblivious routing has shown excellent performance guarantees for changing and uncertain traffic demands in the Internet backbones and more recently in DCN environments [4, 6]. For instance, VLB bounces off every flow to random intermediate switches and can be implemented via encapsulation (e.g., IEEE 802.1ah, IP-in-IP) or revisited packet header bit spaces (e.g., position-based hierarchical MAC addresses [8], Bloom-filter-based Ethernet fields [13]).

· Centralized controllers. In order to customize the DCN and achieve the meet control requirements, a direct networking approach based on logically centralized controllers is a common approach to transparently provide the networking functions (address resolution, route computation) and support services (topology discovery, monitoring, optimization). Implemented as fault-tolerant distributed services in commodity servers, centralized directory and control plane services have shown to scale well and be able to take over the network control, rendering flow-oriented networking, load balancing, health services, multicast management, and so on.

Table 1. Comparison of published architectural approaches for cloud data center networks.

	VL2 [4]	Monsoon [5]	Bcube [6]	Portland [8]	SiBF [13]
Topology	3-tier 5-stage Clos	3-tier 5-stage Clos	Hypercube	3-level fat-tree	Any
Routing & Forwarding	IP-in-IP encapsulation	MAC-in-MAC tunneling	Shim-header-based source routing	Position-based hierarchical MAC	Bloom-filter-based source routing MAC
Load balancing	VLB	VLB	Oblivious	Not defined	VLB
End-host modification	Yes	Yes	Yes	No	No
Programmable switches	No	Yes	No	Yes	Yes

Trends - Cloud DCs are like factories, i.e., the number one goal is to maximize useful work per dollar spent. Hence, many efforts are devoted to minimize the costs of running the large scale infrastructures [3], which requires bringing down the power usage effectiveness (PUE) levels and potentially benefiting from tax incentives for (near) zero-carbon-emission DCs. In this context, energy efficiency of photonic cross-connects outperform the electrical counterparts. However, before we assist to the first all-optical DCN, the price-per-Gbit of optical ports needs to sink at a higher rate than the electrical versions. Further technology market break even points that need to be monitored include high speed memory and solid state disks. Spinning-based hard disks offer the best bit-per-dollar ratio but are limited by their access time, which motivates the design of novel DC architectures [9] where information is kept entirely in low latency RAM or solid state flash drives, while legacy disks are deprecated to back-up jobs. Another ratio that may motivate the design of new (content-centric) inter-networking solutions is the *memory vs. transit* price, which may motivate DCNs (and routers) to cache every piece of data in order to reduce the costs of remote requests.

The so-called green networking trend favors connections to remote locations close to (cheap/clean) energy sources. Recent studies [10] in cost-aware Internet routing have reported 40% savings of a cloud computing installation's power usage by dynamically re-routing service requests to wherever electricity prices are lowest on a particular day, or perhaps even where the data center is cooler. Such green inter-networking approaches require routing algorithms that track electricity prices and take advantage of daily and hourly fluctuations, weighting up the physical distance needed to route information against the potential savings from reduced energy costs.

Finally, the following domains can be identified as distinctive areas of opportunities for optical technologies:

1) Intra-DCN with all-optical technology, potentially with multiple lambdas per port and WDM-based solutions. Innovation is call for to provide fast reconfigurable optical paths to circumvent congestions by dynamically setting up light paths between ToRs (cf. [12]), or novel configuration-less multicast-friendly optical switching, e.g., borrowing from the Bloom filter principle of the electrical domain (cf. [13]) to provide pure optical switching based on the presence of a certain combination of optical signal wavelengths.

2) Inter-DCN solutions to support the (live) migration of VM and data-intense computation jobs from the enterprise to the cloud and vice-versa, the so-called *cloud-bursting*. In addition to being bandwidth-hungry, cloud-bursting requires scalable networking solutions with built-in security and control mechanisms (aka Virtual Private Lan Services - VPLS) that provide addressing protocol and topology transparency over QoS capable virtual private clouds. In this context, multi-domain optical technologies may be an aid to the emergence of an Inter-Cloud, i.e., the inter-networking of Clouds (public, private, internal) for the dynamic creation of federated computing environments that promise to leverage the Internet to an even more consolidated global service platform.

References

[1] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," SIGCOMM CCR, vol. 38, no. 4, pp. 63–74, 2008.

[2] N. Farrington, E. Rubow, and A. Vahdat, "Data Center Switch Architecture in the Age of Merchant Silicon," in IEEE Hot Interconnects, New York, Aug. 2009.

[3] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," SIGCOMM CCR, vol. 39, no. 1, 2009.

[4] A. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "Vl2: a scalable and flexible data center network," SIGCOMM CCR, vol. 39, no. 4, pp. 51–62, 2009.

[5] A. Greenberg, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "Towards a next generation data center architecture: scalability and commoditization," in PRESTO '08. New York, NY, USA: ACM, 2008, pp. 57–62.

[6] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "Bcube: a high performance, server-centric network architecture for modular data centers," in SIGCOMM '09. ACM, 2009.

[7] N. McKeown, T. Anderson, H. Balakrishnan, G. Parulkar, L. Peterson, J. Rexford, S. Shenker, and J. Turner, "Openflow: enabling innovation in campus networks," SIGCOMM CCR, vol. 38, no. 2, pp. 69–74, 2008.

[8] R. Niranjan Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, and A. Vahdat, "Portland: a scalable fault-tolerant layer 2 data center network fabric," in SIGCOMM '09, 2009.

[9] J. K. Ousterhout et al., "The case for ramclouds: Scalable high performance storage entirely in dram," SIGOPS Oper. Syst. Rev. 43, 4 (Jan. 2010), 92-105.

[10] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," in SIGCOMM '09. ACM, 2009.

[11] A. G. S. Kandula, Sudipta Sengupta and P. Patel, "The nature of data center traffic: Measurements and analysis," in ACM SIGCOMM IMC, November 2009.

[12] G. Wang, D. G. Andersen, M. Kaminsky, M. Kozuch, T. S. E. Ng, K. Papagiannaki, M. Glick, and L. Mummert, "Your data center is a router: The case for reconfigurable optical circuit switched paths," in Proc. of HotNets-VIII, 2009.

[13] C. Esteve Rothenberg, C. A. Macapuna, F. L. Verdi, M. F. Magalhães and A. Zahemszky, "Data center networking with in-packet Bloom filters", in 28th Brazilian Symposium on Computer Networks (SBRC), Gramado, Brazil, May 2010.

Bio

Christian Esteve Rothenberg is a research scientist at Fundação Centro de Pesquisa e Desenvolvimento (CPqD), Campinas , Brazil. His main technical interests include Cloud Computing, IMS/NGN, Service Delivery Platforms, OpenFlow, and Future Internet architectures. He works towards his PhD on compact forwarding methods in data-centric networks at University of Campinas (Unicamp), Brazil. He holds a Telecommunication Engineering degree from the Technical University of Madrid (UPM), Spain, and a German Diplom in Electrical Engineering and Information Technology from the Darmstadt University of Technology (TUD) for his thesis at Deutsche Telekom / T-Systems on IMS-based fixed mobile convergence and mobility management.

About PRISM (***www.ontc-prism.org***)

***Aim*:** To disseminate relevant content pertaining to optical networking and related growth areas across industry and academia. To promote the growth of optical networking activity by creation of a unified knowledge base. To create a communication bridge between industry and academia in terms of research frontiers and complementary strategies for future growth.

***Scope*:** The optical networking community stands at a point where its potential is not fully realized. The bandwidth offered by the fiber at price points that currently prevail is a fantastic business case for Internet services for providers the world over. Optical networking has transcended itself from a point-to-point communication service to a WDM based multi-point granular networking hierarchy. This journey was made possible through successful and important innovations in the optics and networking domain, bringing together a rich technology set for deployment in telecommunication networks. It would be fair to say that without optical networking, the scope of the Internet would not reach its global scale that it has presently reached. In the future, optical networking has the potential to impact the telecom world through new innovations in architecture, protocol and devices that would lead to new service offerings impacting human lives. Amongst these futuristic offerings are cloud computing, energy efficient systems, data-centers, 100 Gigabit Ethernet, WDM PON, multi-point communication systems, sub-wavelength grooming and transparent ROADM-based services. It is clear, and especially pronounced in Asia and parts of Europe that optical networking will play a very important role in the design of future networks. Whether it is the GENI project in the US or the Akari in Japan – optical networking finds a clear way into technological offerings for the future of the telecommunication industry. From a historical perspective, optical networking has offered significantly to the telecom industry – we distinctly note that after the telecommunication bubble burst, it was the area of metropolitan networks that led to the re-bounce of telecommunications the world over. It is always important to highlight such historical perspectives from industry leaders and pioneers to bring the optical community closer. We continue to exploit the latest advances in this area of telecommunications – delving on the research and development of optical networking solutions.

The scope of the newsletter is as follows:

- A forum that brings the optical networking community together, through leadership articles in technology and research.
- Bring to the fore issues that both industry and academia are working on, with the focus of being able to minimize this gap through interaction via the newsletter forum.
- Highlight important events related to the area of optical networking, in particular focus on consortiums, projects, awards, seminal breakthroughs, standards and industry related information.
- Research: Focus on research issues pertaining to optical networking. Showcase key growth areas (like data centers, metro ROADMs, 100GE, etc.).
- Consortiums and Projects: Focus on consortiums and projects relevant to optical networking, in which the primary entities are research focused (non-profit groups like universities etc.)
- Developing Economies: Focus on emerging economies and the networks there.
- Standardization activity: The newsletter will periodically discuss standard related activities especially when new drafts are circulated or a standard in form of an MSA is accepted. A standard pioneer will be invited to write about the standard. Our focus will be on the IEEE 802 working group, the ITU groups and FSAN groups in terms of coverage.
- Industry information: latest *technical* happenings will be reported from the industry. These will be critically based on demonstrations at international tradeshows such as OFC, ECOC and World Broadband Forum. Care will be taken not to report any company specific information and ensure vendor neutrality in the newsletter.
- Service provider focus: Since a key consumption point of our industry are service providers, it is most important to focus a section of the newsletter on them. We will in every newsletter focus on the latest happenings in the provider space – whether it is adoption of new technologies or new deployments or even network designs, we will cover these through neutral writings. In particular, we will ensure that no names are taken in the coverage, making it generic – for example, "a select provider in the North America has decided to deploy ROADM technology using WSS cross-connects [source]..".
- Periodically create a roadmap of technologies in different domains pertaining to optical networking. The roadmap would be a team effort by multiple experts in association with the editor.
- Optical Networking is Fun (ONiF): a section devoted to humor in optical networking – puzzles, crosswords and "did you know" for after-hours research.

Submit your article as a .pdf file to submissions@ontc-prism.org. Note that you must have a covering note that descibes the nature of the article from one of the above scope keywords. **The scope keywords are: consortiums, projects, growth areas, emerging economies, Standardization activity, Industry information, Service provider focus, roadmap of technologies and Optical Networking is Fun.**

Note to prospective authors: ONTC Prism follows strict policies mandated by the IEEE Code of Ethics. We will strongly enforce plagiarism and self-plagiarism as a review criteria. For more information visit:http://www.ieee.org/web/publications/rights/ID_Plagiarism.html.