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.