RZ 3742 (# 99752) 12/21/2009
Computer Science 9 pages

Research Report

Flow and Congestion Control for Datacenter Networks

M. Gusat*, C. Minkenberg*, G.J. Paljack*‡

*IBM Research – Zurich, Rüschlikon, Switzerland

‡Budapest University of Technology and Economics, Budapest, Hungary

Research
Almaden • Austin • Beijing • Delhi • Haifa • T.J. Watson • Tokyo • Zurich

Flow and Congestion Control for Datacenter Networks

M. Gusat, C. Minkenberg, G. J. Paljak
IBM Research, Zurich Research Laboratory
Budapest University of Technology and Economics, Hungary
mig@zurich.ibm.com

Abstract: The limits of power dissipation and Moore's law are leading toward increasing parallelism and a shift of focus from CPUs to interconnection networks. This trend is also reflected in the rise of blade-based datacenters, which cluster server and storage units packaged as blades, with several networks. We begin with the trends and requirements of datacenter interconnection networks. Next, we show that lossless link-level flow control is a necessary feature of such networks, required for correctness and performance. However, such flow control schemes have a side-effect: saturation tree congestion, potentially causing catastrophic performance collapse. We argue that the ongoing trends toward increased efficiency, consolidation, and virtualization will escalate the likelihood of congestive collapse. This amplifies the need for congestion management with prevention and recovery mechanisms. Solutions established in best-effort networks (TCP/IP and ATM) are not directly suitable, mainly because they assume a lossy link layer. We advocate the need for research in flow and congestion control in lossless interconnects to meet the challenges of ubiquitous parallelism.

1. Pervasive Parallelism and Trends in Datacenters

Advances in CPU and computer architecture in the last few years both in commercial-off-the-shelf products (e.g. multi-core chip multiprocessors) and in large-scale specialized systems (e.g. High-Performance Computing, HPC) renewed interest in parallelism, exploited at all levels: from VLSI to server packaging and from instruction- to wide-area system-level grids. A fast growing example is scalable datacenters based on blade servers, described in [16] built from commodity components: high-volume processors and off-the-shelf network equipment. Symmetric Multi Processor (SMP) blades built with multi-core chip multiprocessors (CMPs), system area networks (SAN) clusters made of dense server racks, and Storage Area Networks (StAN) are only a few examples.

The performance of such systems is not primarily determined by the features of a single unit, but by how well balanced the integration of many such units is. These systems are moving away from a processor-centric to an interconnect-centric architecture, where an increasing number of processors, and storage units are interconnected by high-performance *packet-switched* fabrics, referred to as interconnection networks (ICTNs) or datacenter networks (DCNs).

The scalable datacenter architecture is based on hierarchically clustering commodity servers around several DCNs. Following the HPC model of a supercomputing center, each cluster inside a datacenter is used and managed as a *single* entity. The multi-tier hierarchy of a typical datacenter involves three layers:

1. *Appliance* - servers dedicated to a specific function, e.g., firewall, caching, load balancing.
2. *Application* - servers dynamically provisioned; may host a variety of applications.
3. *Database* - tightly coupled SMP servers of highest performance and reliability.

The trend toward such scalable datacenters is sometimes referred to as *scale-out* [16]. Scale-out can be summarized as building scalable hardware platforms that combine the performance and ease of programming of *scale-up* systems with the reliability of distributed systems, at commodity cost points. Scale-up refers to large monolithic systems employing specialized hardware components for maximum performance, typically mainframes and high-end SMP and non-uniform memory architecture (NUMA) machines. The success of clusters has established the scale-out as dominant trend in supercomputing and in IT systems, while scale-up systems are restricted to a decreasing share.

1.1. Trends in data centers

We observe the following ongoing trends.

T1. Power constraints The persistent increases in clock frequencies of the late 1990s and early 2000s is coming to an end. Designers have to face the challenges of the so-called power wall; higher frequencies lead to higher power consumption, which, in turn, raises a multitude of cost, cooling, reliability issues. The strict upper-bounds of overall datacenter energy consumptions are being challenged by large-scale datacenters consisting of hundreds of thousands of nodes.

T2. Federation For sustainable growth, we anticipate consolidation at all layers - applications, servers, storage and, in particular, of networks; closely associated with the need for *virtualization and partitioning* as elaborated in [16]. Consolidation is required because of the following reasons: (i.) the growing number of users, rising application load requires larger, higher performance systems; (ii.) new complex requirements of functionality, performance and stability arise; (iii) demand for high utilization, less surplus resources.

Current systems (see Fig.1), there are at least three disjoint external networks – commonly a LAN, a SAN, and a StAN – which implies multiple: protocol stacks, management schemes, cables, adapters and routers. This is a costly proposition both in terms of the upfront costs of the hard- and software as well as the running costs of managing and maintaining multiple nets. These nets are often running at very low average utilization, which drives the cost-benefit ratio even higher.

T3. Service Level Agreements (SLAs): SLAs recapitulate the extra-functional requirements against a system; these Quality of Service (QoS) metrics and their strict enforcement are closely associated with the growing importance of media-rich traffic in servers, storage and across networks, and the enforcements of SLAs. From the user perspective, congestion control and adaptive routing are required to create a lossless network within the datacenter and also for communication out of the datacenter.

For example, for a financial institution communication with exchanges and their clients, losslessness within the datacenter alone is considered insufficient. Also real time (RT) features are essential QoS components; future datacenter will implement more of these, e.g., the ability to be not interrupted while running on CPU is necessary in many datacenter applications; it also implies that the network communications associated with this thread must have RT priority across the DCN.

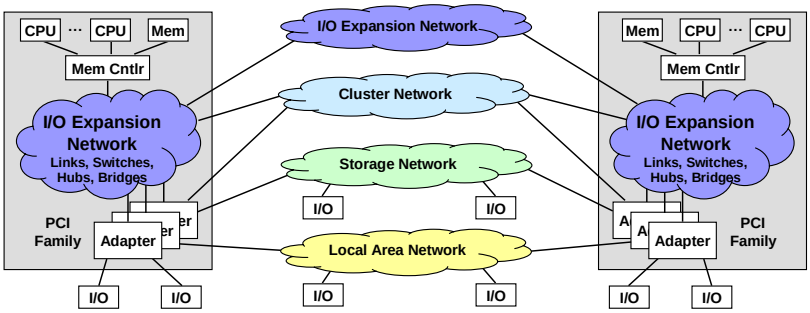
T4. High Availability and Fault-Tolerance: minimal or no system outages are expected. The capacity for IT to offer reliability, redundancy, disaster resilience and recovery is reduced if the application, server, storage, and network RAS needs are decoupled from each other. E.g., network's 1+1 redundancy must be continued and properly terminated inside server's hardware - all the way up to application level. Nonetheless, building and managing such highly-available datacenter systems is a complex challenge.

Next we will derive the implications that these trends have on the datacenter DCN.

2. Datacenter Network: Requirements.

2.1. Correctness and Performance Requirements

Computer systems traditionally have communicated (i.) internally between chips on the motherboard and to/from IO peripherals using legacy busses such as ISA, PCIe, AGP, IDE, SCSI; (ii.) externally with the outside world using various LAN/WAN protocols, predominantly IP over Ethernet. This is set to change because neither busses nor legacy LAN/WANs meet future datacenter requirements. This is also the reason why specialized proprietary interconnects are being used for HPC clustering, or Fibre Channel for StANs, and why standardized DC network protocols have been developed and are being introduced, like IEEE 802 Datacenter Bridging (DCB).



	PCI Family	Ethernet	Fiber Channel	FC-AL	SCSI & (S)ATA	InfiniBand	Vendor Custom	IHV
I/O Expansion	●	-	-	-	-	-	●	-
Storage - Host	-	●	●	-	-	-	●	-
Storage - Device	-	-	-	●	●	-	•	-
Local Area Networks	-	●	-	-	-	-	-	-
Cluster	-	●	-	-	-	•	●	●

Figure 1. Datacenter networks

Successful adoption of datacenter systems hinges on the scale-out infrastructure, in particular, on the communication response time, throughput, reliability, and cost. Besides performance specifications, a DCN must also meet a level of reliability and stability unprecedented in earlier HPC and LAN networks.

DCN traffic can be highly bursty, because when any *single* source needs full access to bandwidth to achieve the lowest latency in the absence of congestion, bursts are injected without traffic shaping or policing. This is a different environment then Internet-type of networks, where aggregation of many low-speed flows (like FTP, web-browsing) determines the traffic characteristics.

More formally, the main requirement categories for DCNs are *correctness* and *performance*.

Table 1. Latency / Bandwidth / RTX Sensitivity: DCNs from OCNs to WANs

	OCN	SMP	Mem	IO	SAN	StAN	LAN	MAN	WAN/Grid
Latency [µs]	0.	1…5		1…10			10	10	10
Bandwidth [Gb/s]	10	10	5…100				0.1…10		1
Type of RTX (link,e2e,h/w,s/w)	LL FEC preferred	tolerated		LL is common			end-to-end in s/w (e.g. TCP)		
							end-to-end in h/w (e.g. TOE)		

Correctness encompasses three core requirements:

C1. **Losslessness**. It comprises of two terms:
   a. Strong: losslessness specification, on which layer are retries to be executed, see Table 1.
   b. Soft: the Bit Error Ratio (BER) is suitable for a DCN application; e.g., SMP links require BER < 10^{-20} [12], whereas 10GigE offers 10^{-12} to 10^{-15} and therefore on average one retry per minute.

C2. **Independent virtual partitions with dead-/livelock-free operation**. The DCN must provide hardware facilities to segregate traffic into classes with independently assigned resources (buffer space, bandwidth slices), so that each class can make progress without interference [4] from or with the others.

C3. **Stable operation**. Independent of adverse traffic patterns and workload characteristics, neither congestion collapses, nor oscillations should happen; the congestion management (CM) mechanism should fairly and efficiently control any congestive situations. Entails prevention and recovery from congestion.

Performance also encompasses three core requirements:

P1. **Low latency**. Latency measured from memory to memory (application/user) is the key factor in the overall system performance, latency guarantees are generally required by SLAs. However, this requirement has remained the most costly to fulfill. (See Table 1.)

P2. **Efficiency**. High throughput and utilization. This term has two connotations: the raw bit rate of the links and the fraction of it that is actually used to transport user data. The latter we refer to as utilization. Also the power-efficiency is a crucial constraint (according to T1), along with [Transaction/s], [Transaction/W] is also to be optimized. The *maximum* utilization that can be achieved on a given network is determined by various factors, such as the overhead of the physical and protocol layers as well as flow control, congestion control, routing schemes, switch architecture, and scheduling.
   a. **Robustness** is a component of efficiency. These DCNs must be extremely robust with respect to rapidly changing traffic patterns, and it must be able to sustain much higher utilization.

P3. **Native Remote Direct Memory Access (RDMA) support**. In addition to the scatter/gather mechanism required to translate between virtual and physical addresses, it involves lighter-weight (than CP Offload Engines, TOEs) solutions for link-level reliable delivery, i.e. error-free, in-order, single-copy. Though arguable, we ascribe the RDMA functionality to performance requirements because of its impact on the DCN performance.

Table 2. Derivations of correctness requirements

Requirement	Solutions / Options	Drawback / Potential conflict with
Losslessness (C1)	(a) EDC and redundancy (FEC, multiple networks)	Against efficiency and low cost
	(b) Lossless LL-FC	Against low complexity
	(c) Recovery via end-to-end retry	Against low latency, efficiency
Losslessness (C1) & Indep. virtural part. (C2)	VL / VC system built on top of LL-FC	Medium complexity
Losslessness (C1) & Low latency (P1)	Lowest latency implies a LL-FC and LL-Reliable Delivery	Medium complexity
Losslessness (C1) & Native RDMA (P3)	Lossless LL-FC and fast end-to-end CM	High complexity (congestion management)
Indep. virtural part. (C2)	(a) Multiple physical DCNs	Against consolidation and low cost and low power
	(b) Single DCN with multiple virtual networks	Medium complexity
Indep. virtural part. (C2) & Stable operation (C3)	Lossy form of LL-FC (e.g., ATM VP/VC)	Against losslessness
Stable operation (C3)	Fast congestion management	Operation depends on the LL-FC

Here we show that the foremost corollary of requirements C1-C3 and P1-P3 is the necessity of lossless *link-level flow control* (LL-FC). As shown in Table 1, DCN losslessness, non-interference and stability can be implemented at different layers, with the corresponding trade-offs. As a general rule of DCN architecture, the lower the layer a feature is implemented at, the faster, costlier and less flexible the respective solution will be. E.g., physical and link-layer protocols are commonly built in hardware – hence are the fastest—whereas the network and transport layers are a mixture of software and hardware.

Table 3. Factors increasing the probability of congestion incidence		
Factor	**Current / Past**	**Emerging**
Over-provisioning (mainly commercial)	Drown the problem in bandwidth; it's cheap.	Doesn't scale with number of nodes. Consolidation adds load. Message sizes rising dramatically (XML)
Single use systems (mainly HPC)	One problem per machine → congestion is a bug	Virtualization on and of clusters => communication patterns are not algorithmically predictable
Other bottlenecks mask the DCN congestion	Slow IO, multiple copies & context switches, slow IRQs,	Fast IO (e.g. PCIe), 0-copy / RDMA protocols, TOE, IRQ coalescing, optimized OS directly expose the DCN.

In Table 2 we summarize consequences derived from the correctness requirements and their combinations. There must be strong error detection and correction, robust flow and congestion control, and fast retransmissions. Round-trip times (RTT) are small compared to WANs, while the maximum transmission units (MTUs) are comparable (2-8 KB). For latency and cost reasons, these DCNs call for shallow buffers. QoS must be supported in hardware to enable resource separation *and* traffic class at the link layer. Finally, cost must be low: under $100/port @ 10 Gb/s and proportional at 30, 40 and 100 Gb/s.

In the last column of Table 2 one can observe that the solution with the least conflicts is a *lossless* LL-FC with virtual lanes/virtual channels (VL/VCs). However, a lossless LL-FC is not only relatively complex to design, but also a radical departure from the established best effort networks, like Ethernet/IP and ATM. Nevertheless, given the volume and importance of datacenters for future computer architectures, a probable migration from best effort to lossless networks could be very consequential – and therefore, incremental.

Table 3 summarizes three main factors for the appearance of congestion, here we reason why current and past solutions are not applicable anymore:

First, over-provisioning does not scale. As the number of nodes increases, the fraction of each node's traffic required for congestion to form decreases proportionally.

Second, algorithmic predictability of network traffic is eliminated as consolidation and virtualization becomes more common in datacenters. When multiple virtualized networks and nodes simultaneously inhabit the same physical hardware, it is no longer the case that any one algorithm controls the traffic pattern; it becomes an overlay of several different algorithms' patterns, varying with the deployment of the virtual resources. Since the virtualization managers themselves cannot be expected to understand all systems' traffic, it becomes unpredictable.

Third, as the cost/performance ratio and the power consumption have become the key metrics determining the success of a DCN, its resources must be used efficiently. In practice, this implies increasing link utilization and driving the network load closer to saturation. The saturation load can be increased by using sophisticated scheduling and adaptive routing [4, 6]. The broad move to XML-based messaging for commercial data is estimated to increase bandwidth requirements at least 20%; this is already beginning to strain installed communication facilities.

2.2 Best Effort and Lossless Interconnects

Aside from correctness, gaining performance from a lossless implementation is a quantitative and qualitative question about P1 and P2. First we compare qualitatively the candidate DCNs from a flow control *perspective*; second, in section 3.3, we give a quantitative comparison. The networks currently used to interconnect storage and servers can be categorized as follows:

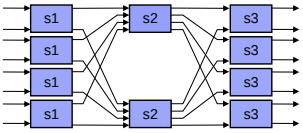
(a) Networks that take every possible measure in hardware not to lose any packet—commonly called "lossless"—like InfiniBand (IBA) [24], Fibre Channel, RapidIO [22], PCIe [23] and others. Lossless networks, prevent buffer overflows, offer faster response time in the case of corrupted packets, do not suffer from loss-induced throughput limitations, and allow bursty flows to start sending immediately at full bandwidth (which, however, may lead to transient congestion). Although packet drops due to errors can never be avoided, lossless networks employ link-level hardware for (i.) a flow control mechanism (LL-FC) to prevent buffer overflow and (ii.) a retry mechanism to ensure quick retransmission of corrupted packets. Various LL-FC schemes, e.g. on/off grants, credits, rate control, and others [6] are used to implement a closed-loop feedback per receive-transmit (RX-TX) pair. The manner in which the LL-FC loop informs the upstream TX about RX's buffer status differentiates between the LL-FC schemes in buffer size and performance.

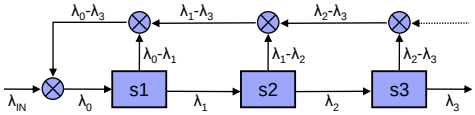
(b) Networks that occasionally may lose some packets are called best effort (BE) or lossy. Best known are TCP/IP over Ethernet and ATM networks. By design, the TPC allows—and even relies on—*packet loss*. BE networks typically employ end-to-end flow control in software (e.g., TCP). It is simple, cheap, and exports the problem from the network to the end nodes. However, recovery from packet drops incurs a significant latency. To prevent a source from being overly greedy and thus exacerbating this problem, TCP employs slow start with additive-increase-multiplicative-decrease (AIMD) window-adjustment algorithm [13,17-21,34]. In a datacenter environment, however, the slow start is undesirable as it introduces significant latency until the full link bandwidth can be used.

2.3. Performance Comparison of Best Effort and Lossless

We provide a quantitative performance comparison of a 10 Gb/s Ethernet-like BE network against its 'equivalent' lossless DCN, both used in a datacenter cluster as follows:

1. Three-stage fat tree multi-stage interconnection network (MIN), see Fig. 2a. Despite higher cost, fat trees provide full bisectional bandwidth and remove topology-induced bottlenecks.
2. Switches varying from 2x2 up to 256x256, currently the maximum switching degree supported in IBA 1.2. Of practical interest are switches between 8 and 64 ports, whereas the extreme degrees (2 and 256) only set the performance boundaries.
3. Link RTT < 1 MTU; an optimistic assumption that fairly favors both DCNs: BE has instant retries, whereas the LL-FC applies immediate backpressure (grants), respectively instant restart after backpressure (credits).
4. Traffic is uniformly distributed in space and time. While this contradicts the reality in datacenter applications (high temporal and spatial burstiness), it favors both DCNs by measuring their respective upper performance bounds.





Nodes	Switch degree	Throughput
2	2	0.52
32	8	0.41
512	32	0.38
2K	64	0.37
32K	256	0.37

Figure 2a. 3-stage fat-tree. **2b. Linearized BE model of (2a).** **2c. BE throughput**

Differences (potential sources of unfairness) are as follows.

Best Effort: We will determine analytically the impact of retries on the goodput of a BE network with a simple probabilistic model. Therefore we substitute the original fat tree with the *linear* 3-hop series system in Fig. 2b, by parameter lumping [6]. We disregard queuing effects due to PAUSE (disabled in most real nets) and model the packet contention per each hop with:

$$(1) \qquad \lambda_{i+1} = 1 - (1 - \lambda_{i+1} / n)^n \quad , n = \textit{switch degree}$$

Recursively applying Eq.1 to the network from Fig. 2b, we obtain the goodput values from Fig. 2c. We observe that for practical scenarios the achievable goodput converges to 38% of the offered load, which is roughly validated by the 35% load empirical rule of Ethernet. Multiple retries can lead to a quasi-collapse of overall network throughput when the load increases above a critical point, with losses due to overflowing buffers triggering retransmissions in a vicious cycle until goodput goes to virtually zero. In terms of latency, the first dependency lies again with retries, since the probability of a packet being dropped (and retransmitted) k-times is O(RTX fraction) = $O(\lambda_0-\lambda_3) / \lambda_0$), plotted in red in Fig. 3.

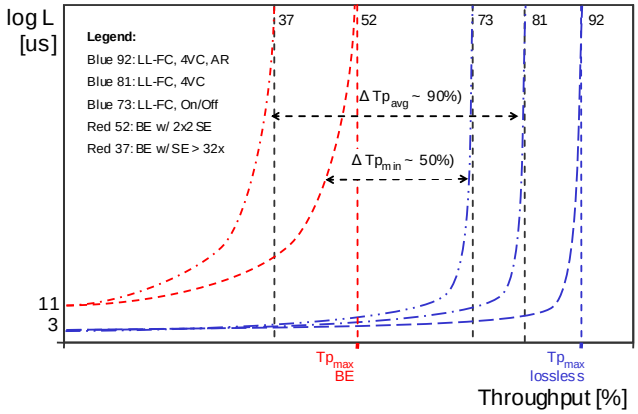


Figure 3.Throughput/Latency comparison: BE (red); Lossless (blue)

Lossless: For the lossless DCN we simulate the actual fat tree built of DCB- and IBA-like switches. Analytical models of LL-FCed networks with adaptive routing are rare and considerably more involved due to their non-linearity and saturation trees discussed in the next section. We simulate 3 models: (a) a simple grant-based LL-FC (73% curve); (b) replace grants with credits (to halve the buffers) and using 4 VLs [6] (81% curve); (c) as in (b), using advanced scheduling and adaptive routing [4,6] (92% curve).

Cases (b,c) favor the lossless DCN over the BE due to the advanced flow control and scheduling - not directly possible in the BE network. However, our comparative framework is meant to reveal the practical consequences in the datacenter environment.

Results (fig. 3): In addition to meeting C1, at the same buffer size a lossless DCN such as IB, can outperform BE networks by 50 to over 90%. The basic LL-FC (a) brings ca. 50% goodput improvement; hence the lower latency

and higher utilization, beneficial in power and cost. Another 30-40% improvement is possible because the advanced LL-FC (VL-based) scheme enables: (i.) independent virtual partitions (C2) over multiple VLANs; (ii.) T_{put} over 92% using load-balanced scheduling that is based on status info provided by the LL-FC.

Overall we observe that a carefully designed lossless LL-FC can deliver 1.5-2x better performance than a BE solution. The key needs, however, remain losslessness and latency (C1 & P1), solved by the introduction of a lossless scheme such as credit-based LL-FC.

3. Saturation Tree Congestion in DCNs

The majority of computer interconnects have adopted credited network operation, with the exception of Ethernet DCB ([25-27, 33].) which has Priority Flow Control (PFC) based on-off grants, because of the performance advantages discussed above. The main challenge that must be solved by all DCNs for a successful DC implementation is that lossless networks can experience high-order head-of-line blocking [3], saturation tree congestion [1, 2, 10], and possibly deadlocks. In this section, we will argue that *congestion management* (CM) is a critical component of DCNs in the datacenter, without which C2, C3, and P3 cannot be met.

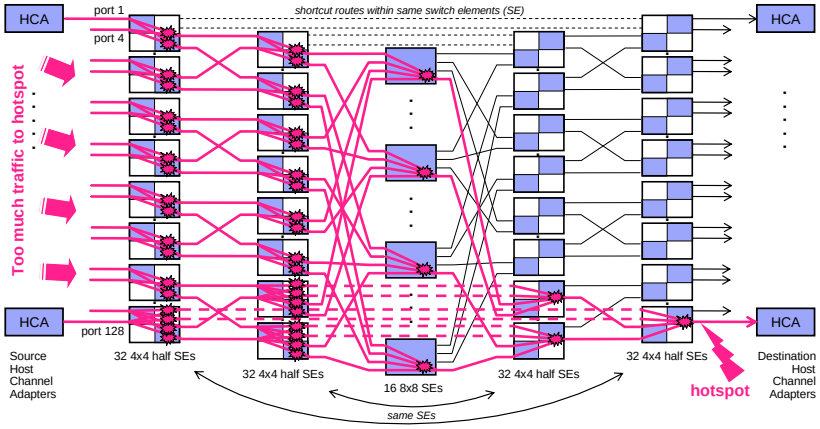


Figure 4. Hotspot saturation tree in a 5-stage fat tree

Figure 4 illustrates the problem (hotspot congestion box). If a sufficient fraction of all the inputs' traffic targets one of the outputs (in the figure, the output labeled 128), that output link can saturate: it becomes a hotspot (HS) that causes the queues in the switch feeding that link to fill up. If the traffic pattern persists, then, no matter what techniques are used to reassign buffer space, it is all ultimately exhausted. This forces that switch's LL-FC to *quickly* throttle back all the inputs feeding that switch. That in turn causes the previous stage to fill its buffer space. In a domino effect, the congestion eventually backs up all the way to the network inputs. This has been called *tree saturation* [1] or, in other contexts, *congestion spreading*.

Ultimately, the traffic causing the hotspot will root one or more saturation trees partly caused by the inherent traffic distribution [1, 2, 10] and partly by flow interference [6, p.112] or high-order head-of-line (HOL) blocking [3]. Once the tree of saturated switches is fully formed, every packet must cross at least one saturated switch. As the time to exit a queue grows exponentially the further a switch is from the hot destination, a majority of the delay is incurred even if only a *single* switch must be crossed. Hence, the network as a whole suffers a catastrophic loss of throughput: its aggregate throughput is gated by the throughput of the single hot output.

Saturation spreads very quickly via LL-FC; according to the analysis of [28], the tree is filled in less than 10 traversal times of the network, far too quickly for software to react in time to the problem. Naturally, the problem also dissipates slowly because all the queues involved must be emptied. Hence, a hardware solution is required that reacts quickly enough to keep the tree from growing large. Clearly the network topology is irrelevant to this effect; saturation trees can be induced in any topology.

4. Congestion Management Solutions for DCNs

As discussed above, lossless LL-FC offers substantial performance benefits, but has the drawback, besides its complexity, of facilitating saturation tree congestion. Unless an efficient CM protocol is designed and implemented to control the fabric operation just below the saturation region and recover from the occasional crossovers, lossless DCNs will be increasingly exposed to saturation trees and congestion collapse. However, while the problem is long outstanding definitive solutions are not yet practically available.

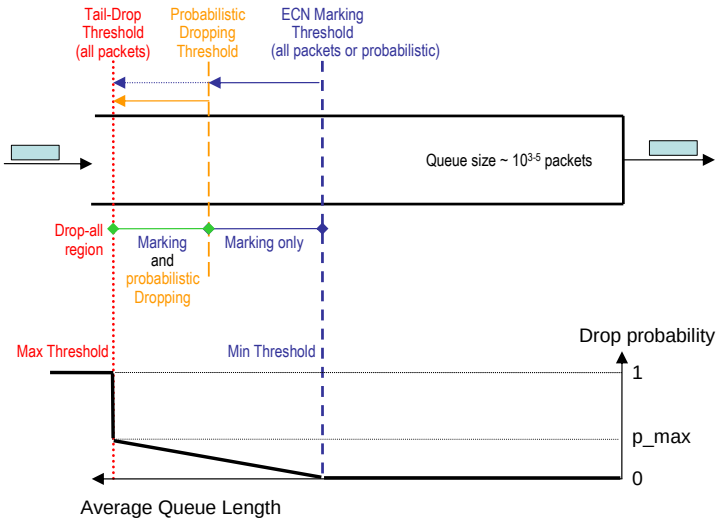


Figure 5. ECN Buffer Sizing

The DCN congestion phenomena are sufficiently different from the BE IP networks to invalidate transfer of TCP (even if explicit congestion notification, ECN, is added) – that is, without major adaptations– to the DCN environment. The main three reasons are:

1. **Losslessness**: TCP has been designed to operate based on *loss*; packet drops are the basic feedback mechanism which triggers the source reaction. Packet loss, however, contradicts C1. Next, recovery is very slow whenever the TCP window is smaller than 6 packets [34]. In smaller DCNs with large MTUs the TCP window size is mostly < 6 packets. TCP doesn't support C2.
2. **Double feedback loop**: Unlike TCP in IP networks, FCC mechanisms in DCNs are based on a *dual* closed-loop control system: (i) LL-FC and (ii) end-to-end CM. The former is the smaller and faster loop taking care of LL correctness and, sometimes, performance like e.g. advanced scheduling. CM involves a larger and slower loop with much longer time constants than the LL RTT; a complete CM solution may include congestion avoidance/prevention and control (after it happens). Since CM is inherently slower than its underlying LL-FC loops, it needs an aggregated view of the DCN status – whereas the LL-FC relies only on local status. Thus CM should compensate the inertia of its larger loop by (a) acquiring global feedback (ECNs, delays etc.) about traffic conditions and (b) elaborating a more complex source reaction that considers the *outdated* global view and ideally, tries to predict the traffic based on the trends acquired so far. Problem is that TCP does not assume the existence of a fast and lossless LL-FC layer; nor does TCP coexist well with other flow control schemes, as proven by TCP over ATM/ Available Bit Rate (ABR).
3. **Shallow buffers**: The alternative would be to *over-design* the switch buffers beyond the size mandated in [11] for lossless DCNs. This, however, is not only is practically impossible (see required buffer size ECN[1] box), but also *aggravates* the post-congestion phase by slowing its recovery [5].

Whereas TCP (and ATM/ABR) were extensively studied and improved for BE networks [5,8,13,17-21,34], we still lack conclusive evidence of their applicability and sufficiency in DCNs. Furthermore, recent research invalidates TCP's use for certain types of middleware [14]. This, however, should not prevent us from using solutions and ideas from TCP/IP and ATM/ABR in lossless DCNs [15]. The large body of knowledge about congestion control developed for BE networks (as in [29-32]) may prove beneficial also for lossless DCNs, as proved in [7-10].

5. Conclusions and Outlook

Our main conclusions from the above are as follows.

I. When datacenter DCNs will adopt lossless LL-FC, congestion control becomes mandatory. We have shown that the core DCN requirements lead to lossless LL-FC. In return for its complexity and the exposure to saturation trees, the addition of credit-based LL-FC to a DCN provides multiple benefits.

- Losslessness: meets the *correctness* requirements of datacenter clustering.
- Latency: up to 2x better performance – by extension of the linear range of operation – than traditional BE networks.

[1] Recent additions to TCP such as ECN [17-21] provide useful hints before loss begins to occur in WAN / Internet routers, yet elicit that (i.) buffers are much larger than feasible in a DCN switch and (ii.) suitably long time constants of the e2e feedback loop, 2-4 orders of magnitude larger.

- Memory reduction: allows a reduction of the buffer size to the minimum imposed by the link RTT.
- Virtualization: can provide distinctly controlled VCs/VLs for virtual partitions.

However, lossless LL-FC introduces two challenges: first, the network exhibits a sharp knee in its delay – throughput characteristics. Therefore, when operating close to the saturation region, even a slight increase in load may induce hard-to-control congestion. Notice that, however, the start of this knee is often at a 2x higher load then for BE's. Second, in a lossless architecture, hotspots introduce saturation trees that hurt performance throughout the system. Both issues call for strong congestion management, including prevention as well as recovery. The need is further increased by the datacenter trends like consolidation and virtualization.

II. Simple transfer of congestion solutions from TCP/IP and ATM/ABR networks to DCNs is not advisable without due diligence. Therefore we have investigated whether a lighter mechanism (than TCP/ECN) is suitable for datacenter applications, then illustrated that such a scheme may guard against performance collapse - provided that the scheme is tuned with correctly set parameters. Deriving the 'right' set of parameters from the network configuration *and* traffic patterns remains subject of further research. Nevertheless, a deeper understanding of congestion management is needed to protect against arbitrarily adverse traffic patterns. Hence the importance of flow and congestion control topics for DCNs.

III. Finally, DCNs would benefit from acquiring a body of flow and congestion control knowledge of rigor and scope comparable to TCP. This includes architecting *native* congestion management solutions for datacenter DCNs, and in-depth analytical and simulation performance comparisons with the best methods from BE networks. The design of high performance native congestion management solutions draws upon non-linear control theory and also involves a good command of scheduling, flow control, switching, routing and queuing.

Outlook

Efficient CM schemes remain complex to design, tune, and validate. This was attested with TCP, which is being continuously improved by a large and active community. The topic of Flow and Congestion Control has not yet acquired the same recognition and status in the DCN community. However, the ubiquity of parallelism with its ensuing network-centrism, the rise of datacenters built out of modular servers, its power constraints, and the general IT trends toward consolidation and virtualization is expected to renew the interest in flow and congestion control.

ACKNOWLEDGMENT

We thank Wolfgang Denzel, Ton Engbersen, Ilias Iliadis, Andreas Kind, Ronald Luijten, Bernard Metzler, Thomas Mittelholzer, Fredy Neeser, Greg Pfister, Roman Pletka - and our colleagues from Systems and Technology Group, for their contributions to this paper. We are also grateful to the HOTI reviewers for their insightful suggestions.

REFERENCES

[1] G.F. Pfister and V.A. Norton, "Hotspot Contention and Combining in Multistage Interconnection Networks", IEEE Trans. on Computers, Vol C-34, No. 10, October 1985, pp. 933-938

[2] W. Vogel et al., "Tree-Saturation Control in the AC^3 Velocity Cluster Interconnect", *Hot Interconnects 8, A Symposium on High Performance Interconnects*, Stanford University, CA, August 16-18, 2000.

[3] M. Jurczyk and T. Schwederski, "Phenomenon of Higher Order Head-of-Line Blocking in Multistage Interconnection Networks under Nonuniform Traffic Patterns", *IEICE Transactions on Information and Systems, Special Issue on Architectures, Algorithms and Networks for Massively Parallel Computing*, Vol. E79-D, No. 8, August 1996, pp. 1124-1129.

[4] J. Duato et al., "Interconnection Networks, an Engineering Approach", IEEE Computer Soc. Press, 1997.

[5] Nagle, J, "On Packet Switches with Infinite Storage", *IEEE Trans. on Communications*, vol. 35, pp. 435-438, April 1987.

[6] W. Dally and B. Towles, Principles and Practices of Interconnection Networks, Elsevier MKP, ISBN: 0-12200-751-4, 2004.

[7] J. Jiang, R. Jain, "Analysis of backward congestion notification (bcn) for Ethernet in datacenter applications," May 2007, pp. 2456–2460

[8] R. Pan et al. Qcn: Quantized congestion notification. [Online]. Available: http://www.ieee802.org/1/files/public/docs2007/auprabhakar-qcn-description.pdf, May, 2007

[9] Y. Lu et al., "Congestion control in networks with no congestion drops," in Proc. 44th Annual Allerton Conference on Communication, Control, and Computing, Monticello, IL, Sep. 2006.

[10] B. Prabhakar et al. (2008, Nov.) Qcn: Averaging principle. [Online]. Available: http://www.ieee802.org/1/files/public/docs2008/au-prabhakar-ave-principle-1016.pdf

[11] M. Gusat et al., "A study of the stability degree of switches with finite buffers under non-negligible RTT," in *Microprocessor and Microsystems Journal*, Elsevier, 2003.

[12] A.F. Benner et al., "Exploitation of optical interconnects in future server architectures," *IBM Journal of Research and Development*, Volume 49, Number 4/5, 2005, pp. 755-775.[former 38]

[13] V. Jacobson, "Congestion avoidance and control," *ACM Computer Communication Review*, Proc. of Sigcomm '88 Symposium, Stanford, CA, August, 1988, 18, 4:314-329, 1988.

[14] P. Pietzuch and S. Bhola, "Congestion Control in a Reliable Scalable Message-Oriented Middleware," ACM/IFIP/USENIX Int. Middleware Conference 2003, Springer-Verlag.[former 39]

[15] T. Blackwell et al., "Credit-based control for ATM networks," Proc. of *First Annual Conference on Telecommunications R&D*, Massachusetts, 1994.

[16] BladeCenter papers in "IBM BladeCenter Systems", *IBM Journal of Research and Development*, Volume 49, Number 6, 2005. [former 40]

[17] S. Floyd and V. Paxson, "Difficulties in simulating the Internet," *IEEE/ACM Transactions on Networking*, vol. 9(4), pp. 392-403, 2001.

[18] S. Floyd, "TCP and explicit congestion notification," *Computer Communication Review 24*, 5 (October 1994), 8–23.

[19] S. Floyd et al., "Equation-based congestion control for unicast applications," In *SIGCOMM* (August 2000).

[20] S. Floyd, V. Jacobson, "Random Early Detection gateways for congestion avoidance," *IEEE/ACM Transactions on Networking* , 397–413., 1993
[21] K.K. Ramakrishnan et al., "The addition of Explicit Congestion Notification (ECN) to IP," Tech. Rep. RFC 3168, IETF, September 2001.
[22] RapidIO Specification, http://www.rapidio.org/specs/current
[23] PCI Express Specification, http://www.pcisig.com/specifications/pciexpress/
[24] InfiniBand Trade Association, "InfiniBand Architecture Specification, Volume 1, Release 1.2" [online document], October 2004, Available at http://www.infinibandta.org/specs/.
[25] D. Bergamasco. (2005) "Data center Ethernet congestion management: Backward congestion notification." [Online]. Available: http://www.ieee802.org/1/files/public/docs2005/new-bergamasco-backward-congestion-notification-0505.pdf
[26] G. Pfister, M. Gusat, W. Denzel, D. Craddock, N. Ni, W. Rooney, T. Engbersen, R. Luijten, R. Krishnamurthy, and J. Duato, "Solving hot spot contention using InfiniBand Architecture congestion control," in Proc. HP-IPC 2005, Research Triangle Park, NC, Jul. 24 2005.
[27] C. Minkenberg and M. Gusat, "Congestion management for 10g ethernet," in Proc. Second Workshop on Interconnection Network Architectures: On-Chip, Multi-Chip (INA-OCMC 2008), Goteborg, Sweden, Jan. 2008.
[28] G. Pfister, V.J. Kumar, "The Onset of Hotspot Contention," Proc. of *1986 International Conference in Parallel Processing*, August 1986
[29] Mascolo et al., "Tcp westwood: bandwidth estimation for enhanced transport over wireless links," in Proc. of MobiCom, Rome, Italy, 2001
[30] Kelly et al., "Rate control in communication networks: shadow prices, proportional fairness and stability," Journal of Operational Research Society, no. 49, pp. 237–252, 1998.
[31] Raynaud et al., "Towards a state-space approach to congestion and delay control in communication networks," in 7th European Control Conference, Cambridge, 2003.
[32] F. Paganini, Z. Wang, S. Low, and J. Doyle, "A new tcp/aqm for stable operation in fast networks," vol. 1, pp. 96–105, vol.1., April 2003
[33] D. Bergamasco, "Ethernet congestion manager (ECM) specification," Cisco Systems, Draft EDCS-574018, Feb. 2007.
[34] R. Morris, "Scalable TCP congestion control," Proc. of *IEEE INFOCOM 2000*, Volume 3, 26-30 March, 2000, pp. 1176-1183.