High speed adaptive rack-scale fabrics

Omer S. Sella
University of Cambridge
Omer.Sella@cl.cam.ac.uk

Andrew W. Moore
University of Cambridge
Andrew.Moore@cl.cam.ac.uk

Noa Zilberman
University of Cambridge
Noa.Zilberman@cl.cam.ac.uk

1 ABSTRACT

Rack-scale systems contain thousands of densely packed connected components. While a data center may accommodate a fully provisioned network, rack-scale systems demand a more compact and versatile network that would even up within a heavily populated system. Unless the critical path between communicating hosts is made faster, distributed rack-scale applications cannot scale. We present adaptive rack-scale fabrics, an architecture that uses *Physical Layer Primitives*, coupled with a *Closed Ring Control*. The resulting fabric uses pre-fetching techniques, but at the physical layer of the interconnect, to optimize performance within strict power-budget limitations.

2 MOTIVATION

Rack-scale systems do not necessarily follow the cpu-board-centric architecture that traditional racks use [4]. Instead of using regular server blades, we strip down the components and redesign according to the relevant metric - NVMe for fast storage, significant amount of DRAM for caching etc. This leads to a layout of hundreds and even thousands of interconnected nodes in a single rack. The meaning is that within a single rack we find a network as sophisticated and complex as in a data center, only much more constrained. In particular two problems arise: latency and power consumption.

Figure 1 shows the latency a packet experiences by traversing multiple hops through layer 2 cut-through switches. It also shows that the delay due to the media, (e.g., fiber) is negligible relative to the use of packet switching. The conclusion is that in the scale of a rack, it is packet switching that prevents distributed rack-scale applications from scaling. As an example, consider a MapReduce operation that requires transmission from all nodes. Since a reducer has to wait for data from all mappers, the slowest link pulls down the performance of an entire system.

Power budget is also a constraint, since rack-scale systems inherit the power budget of a traditional rack, and is factored into our proposed architecture as shown in figure 2. Three key points of the architecture are:

- Backwards compatibility - No restructuring of the network layer is needed. In particular, existing applications benefit from the architecture with no required change.

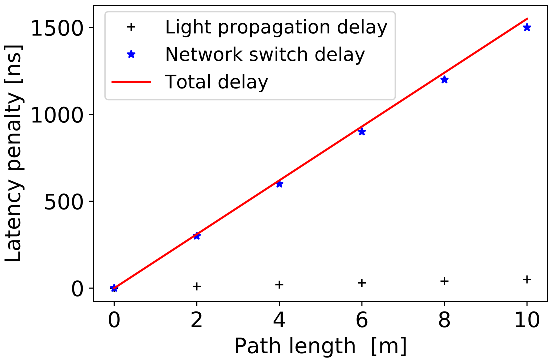

Figure 1: The latency due to propagation of packets in the media vs. the latency due to packet traversing a layer 2 state-of-the-art cut through switch. We assume a switch every 2 meters. In the scale of a rack, the latency due to packet switching is dominant, and hence is bottlenecking scalability.

- Media agnostic - the specific underlying media is irrelevant. We only expect it to provide some subset of the *Physical Layer Primitives* that we define.
- Forward compatibility and fast adoption - Novel physical layer advancements could be easily integrated into a system already running our *CRC*.

3 PROPOSED ARCHITECTURE

Configurable interconnect has seen many advances in recent years. Both on the optics side as in [2], as well as the electrical side as in [3]. While these solutions are different in the underlying media (optics vs. electrical) as well as in configuration times, they could be treated as functionalities that were added to the (already existing) physical layer. We place these extensions to the physical layer under a single framework, which we call *Physical Layer Primitives (PLP).* In turn, these *PLP* are orchestrated by a control mechanism, that also schedules flows according to the availability of *PLP*'s. The control part of the architecture, called *Closed Ring Control (CRC)*, uses feedback from the interconnect such as latency, power consumption etc., to tag each link with a cost function. In this way, both routing as well as changes to the topology, are subject to the tools of control theory. By detaching the

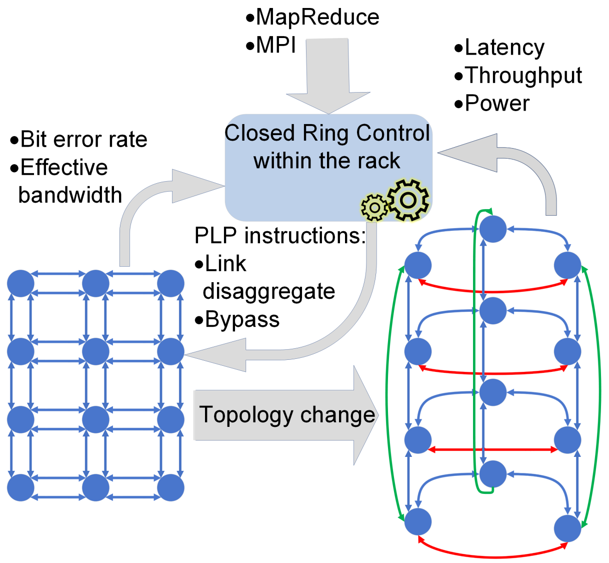

Figure 2: An example of the adaptive rack-scale network operation. Initially, the rack is configured using a grid topology of two lanes per link. Internal indications are fed to the *Close Ring Control - CRC*, that issues commands to the *Physical Layer Primitives - PLP*. These result in a torus topology running at one lane per link.

development of *PLP* from innovation in *CRC* we obtain two goals: 1. Allowing new physical layer improvements to be coupled instantaneously with a control algorithm, and 2. Enabling faster data centre adoption of high cost disruptive technologies. A system that already uses our *PLP* will absorb seamlessly any physical layer advancement that could be characterized as a *CRC*.

3.1 *Physical Layer Primitives* - PLP

We assume that a physical link is made up from physical lanes. The canonical example is a 100Gbps link that is made from four 25Gbps physical links, but different wavelengths under wavelength division multiplexing is an equivalent example. Looking at [3] and [2], we can identify several *Physical Layer Primitives*, and in addition draw new ones:

1. Link breaking / bundling - separating a link of N lanes into two links of k and N-k lanes and vice versa.
2. High speed bypass - connecting two links at the lowest possible physical level.
3. Turning a link on or off.
4. Adaptive forward error correction.
5. Per-lane statistics such as: bit error rate, latency, and effective bandwidth.

3.2 *Closed Ring Control* - CRC

The *Closed Ring Control, or CRC* uses per-link price tags, with respect to metrics such as latency, congestion, link health etc. to allocate *PLP*'s and schedule flows. The problem that arises in all reconfigurable fabrics is finding the minimum flow size for which reconfiguration is worth the cost. This could be formulated as an optimization problem and solved distributively by the *CRC*. Further insights on rapid provisioning and reconfiguration, as well as traffic engineering for virtual switching can be found in Andromeda [1]. Figure 2 shows a *CRC* embedded in the rack. Upon receiving per-link statistics, the *CRC* issues *PLP* instructions to improve the target metric, e.g: latency, by reducing the amount of switching logic that a packet has to go through.

4 EVALUATION

Since rack-scale systems contain hundreds to thousands of connected nodes, a simulation is used to evaluate the solution. We chose omnet++ as our simulation framework. To be certain that a large scale simulation is sound and credible, we begin with a small scale simulation verified by a hardware proof of concept (POC). We intend to use the NETFPGA SUME platform [5] for the hardware POC. Once the small scale simulation is validated, the POC will be integrated into the large scale simulation.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their constructive comments and helpful suggestions. This work was partly funded by Microsoft Research through its PhD Scholarship Programme, the Leverhulme Trust (ECF-2016-289) and the the Isaac Newton Trust. Omer S. Sella gratefully acknowledges past support by the Samuel and Lottie Rudin Scholarship Foundation.

REFERENCES

[1] Michael Dalton, David Schultz, Jacob Adriaens, Ahsan Arefin, Anshuman Gupta, Brian Fahs, Dima Rubinstein, Enrique Cauich Zermeno, Erik Rubow, James Alexander Docauer, et al. 2018. .

[2] Monia Ghobadi, Ratul Mahajan, Amar Phanishayee, Nikhil Devanur, Janardhan Kulkarni, Gireeja Ranade, Pierre-Alexandre Blanche, Houman Rastegarfar, Madeleine Glick, and Daniel Kilper. 2016. Projector: Agile reconfigurable data center interconnect. In *Proceedings of the 2016 ACM SIGCOMM Conference*. ACM, 216–229.

[3] Vishal Shrivastav, Asaf Valadarsky, Hitesh Ballani, Paolo Costa, Ki Suh Lee, Han Wang, Rachit Agarwal, and Hakim Weatherspoon. 2017. *Shoal: A Lossless Network for High-density and Disaggregated Racks.* Technical Report.

[4] Georgios S Zervas, Fangsheng Jiang, Qianqiao Chen, Vaibhawa Mishra, Hui Yuan, Kostas Katrinis, Dimitris Syrivelis, Andrea Reale, Dionysios Pnevmatikatos, Michael Enrico, et al. 2017. Disaggregated compute, memory and network systems: A new era for optical data centre architectures. In *Optical Fiber Communication Conference*. Optical Society of America, W3D–4.

[5] Noa Zilberman, Yury Audzevich, G Adam Covington, and Andrew W Moore. 2014. NetFPGA SUME: Toward 100 Gbps as research commodity. *IEEE micro* 34, 5 (2014), 32–41.