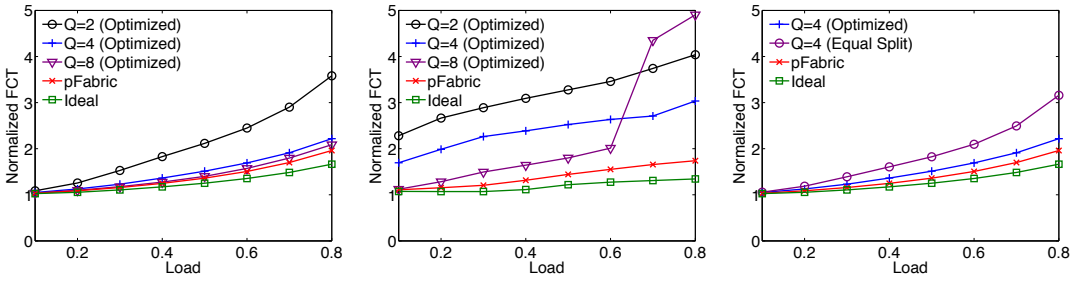

(a) Overall: Avg (b) (0, 100KB]: 99th prctile (c) Optimal thresholds vs equal split: Overall Avg

Figure 17: Web search benchmark with 2, 4, and 8 priority queues. Parts (a) and (b) show the average normalized FCT across all flows and the 99th percentile for the small flows. Part (c) compares the performance using the optimized thresholds with a heuristic which splits the flows equally in case of 4 queues.

be appropriate. This can easily be handled by operating the pFabric priority scheduling and dropping mechanisms within individual "higher-level" traffic classes in an hierarchical fashion. Traditional QoS mechanisms such as WRR are used to divide bandwidth between these high-level classes based on user-defined policy (e.g., a soft-real time application is given a higher weight than batch jobs), while pFabric provides near-optimal scheduling of individual flows in each class according to the class's priority scheme (remaining flow size, deadlines, etc).

Other datacenter topologies: We have focused on Fat-tree/Clos topologies in this paper as this is by far the most common topology in practice. However, since conceptually we think of the fabric as a giant switch with bottlenecks only at the ingress and egress ports (§3) we expect our results to carry through to any reasonable datacenter topology that provides uniform high throughput between ingress and egress ports.

Stability: Finally, the theoretical literature has demonstrated scenarios where size-based traffic prioritization may reduce the stability region of the network [20]. Here, stability is in the *stochastic* sense meaning that the network may be unable to keep up with flow arrivals even though the average load on each link is less than its capacity [10]. However, this problem is mostly for "linear" topologies with flows traversing different numbers of hops — intuitively it is due to the tradeoff between prioritizing small flows versus maximizing service parallelism on long routes. We have not seen this issue in our study and do not expect it to be a major concern in real datacenter environments because the number of hops is very uniform in datacenter fabrics, and the overall load contributed by the small (high-priority) flows is small for realistic traffic distributions.

8. CONCLUSION

This paper decouples the key aspects of datacenter packet transport — flow scheduling and rate control — and shows that by designing very simple mechanisms for these goals separately we can realize a minimalistic datacenter fabric design that achieves near-ideal performance. Further, it shows how surprisingly, large buffers or complex rate control are largely unnecessary in datacenters. The next step is to integrate a prototype implementation of pFabric with a latency-sensitive application to evaluate the impact on application layer performance. Further, our initial investigation suggests that further work on designing incrementally deployable solutions based on pFabric could be fruitful. Ultimately, we believe this can pave the path for widespread use of these ideas in practice.

Acknowledgments: We thank our shepherd, Jon Crowcroft, and the anonymous SIGCOMM reviewers for their valuable feedback. Mohammad Alizadeh thanks Tom Edsall for useful discussions regarding the practical aspects of this work.

9. REFERENCES

[1] M. Al-Fares, A. Loukissas, and A. Vahdat. A scalable, commodity data center network architecture. In *Proc. of SIGCOMM*, 2008.

[2] M. Al-Fares, S. Radhakrishnan, B. Raghavan, N. Huang, and A. Vahdat. Hedera: dynamic flow scheduling for data center networks. In *Proc. of NSDI*, 2010.

[3] M. Alizadeh, A. Greenberg, D. A. Maltz, J. Padhye, P. Patel, B. Prabhakar, S. Sengupta, and M. Sridharan. Data center TCP (DCTCP). In *Proc. of SIGCOMM*, 2010.

[4] M. Alizadeh, A. Kabbani, T. Edsall, B. Prabhakar, A. Vahdat, and M. Yasuda. Less is more: trading a little bandwidth for ultra-low latency in the data center. In *Proc. of NSDI*, 2012.

[5] M. Alizadeh, S. Yang, S. Katti, N. McKeown, B. Prabhakar, and S. Shenker. Deconstructing datacenter packet transport. In *Proc. of HotNets*, 2012.

[6] M. Alizadeh, S. Yang, M. Sharif, S. Katti, N. McKeown, B. Prabhakar, and S. Shenker. pFabric: Minimal Near-Optimal Datacenter Transport. `http://simula.stanford.edu/~alizade/pfabric-techreport.pdf`.

[7] B. Atikoglu, Y. Xu, E. Frachtenberg, S. Jiang, and M. Paleczny. Workload analysis of a large-scale key-value store. In *Proc. of SIGMETRICS*, 2012.

[8] N. Bansal and M. Harchol-Balter. Analysis of SRPT scheduling: investigating unfairness. In *Proc. of SIGMETRICS*, 2001.

[9] A. Bar-Noy, M. M. Halldórsson, G. Kortsarz, R. Salman, and H. Shachnai. Sum multicoloring of graphs. *J. Algorithms*, 2000.

[10] T. Bonald and L. Massoulié. Impact of fairness on Internet performance. In *Proc. of SIGMETRICS*, 2001.

[11] A. Dixit, P. Prakash, Y. C. Hu, and R. R. Kompella. On the Impact of Packet Spraying in Data Center Networks. In *Proc. of INFOCOM*, 2013.

[12] 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. In *Proc. of SIGCOMM*, 2009.

[13] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris. *Fundamentals of Queueing Theory*. Wiley-Interscience, New York, NY, USA, 4th edition, 2008.

[14] C.-Y. Hong, M. Caesar, and P. B. Godfrey. Finishing Flows Quickly with Preemptive Scheduling. In *Proc. of SIGCOMM*, 2012.

[15] The Network Simulator NS-2. `http://www.isi.edu/nsnam/ns/`.

[16] J. Ousterhout, P. Agrawal, D. Erickson, C. Kozyrakis, J. Leverich, D. Mazières, S. Mitra, A. Narayanan, D. Ongaro, G. Parulkar, M. Rosenblum, S. M. Rumble, E. Stratmann, and R. Stutsman. The case for RAMCloud. *Commun. ACM*, 2011.

[17] C. Raiciu, S. Barre, C. Pluntke, A. Greenhalgh, D. Wischik, and M. Handley. Improving datacenter performance and robustness with multipath TCP. In *Proc. of the SIGCOMM*, 2011.

[18] B. Vamanan, J. Hasan, and T. N. Vijaykumar. Deadline-Aware Datacenter TCP (D2TCP). In *Proc. of SIGCOMM*, 2012.

[19] V. Vasudevan, A. Phanishayee, H. Shah, E. Krevat, D. G. Andersen, G. R. Ganger, G. A. Gibson, and B. Mueller. Safe and effective fine-grained TCP retransmissions for datacenter communication. In *Proc. of SIGCOMM*, 2009.

[20] M. Verloop, S. Borst, and R. Núñez Queija. Stability of size-based scheduling disciplines in resource-sharing networks. *Perform. Eval.*, 62(1-4), 2005.

[21] C. Wilson, H. Ballani, T. Karagiannis, and A. Rowtron. Better never than late: meeting deadlines in datacenter networks. In *Proc. of SIGCOMM*, 2011.

[22] D. Zats, T. Das, P. Mohan, D. Borthakur, and R. H. Katz. DeTail: Reducing the Flow Completion Time Tail in Datacenter Networks. In *Proc. of SIGCOMM*, 2012.