Research Statement

Meg Walraed-Sullivan
PostDoc, Mobility and Networking Research Group
Microsoft Research
megwalraedsullivan@gmail.com

At the most fundamental level, my research focuses on co-designing topologies and corresponding communication protocols for novel network architectures. At the heart of this lies my passion for creating efficient and fast distributed protocols that leverage key properties of the underlying topology, for instance, structural symmetry or high path diversity. During the course of my post doctoral work at Microsoft Research, my goal has been to consider various graph structures and topologies that can serve as the foundations for a variety of current and emerging network architectures, and to design protocols that allow applications to leverage symmetries and other features of these topologies. Throughout this time, I've retained the approach that I began in graduate school when designing distributed protocols for large-scale networks, that is, to combine theory with experimental evaluation and systems building. I aim to both formally derive a distributed protocol from a simple, fundamental problem abstraction, and also to build and measure a system running real software on real hardware so as to verify its practical applicability. Working with this intersection between theory and practice continues to provide me with deeper insight into the complex problems that face systems researchers today.

As new applications continue to appear and as existing workloads grow, network architectures must also evolve to keep up with ever-increasing demands and requirements. For instance, there has been substantial growth in the size of data centers in recent years as more and more computation has moved to the cloud. At the same time, concerns over last-mile performance have led to the emergence of edge networks and the more recently coined micro data centers (mDCs). On one hand, we've seen the introduction of highly specialized hardware in parts of the data center, tailored to perform a single critical workload significantly more quickly than would be possible with a standard server. On the other hand, some data center operators are moving towards rack-scale architectures, in which thousands of small, identical compute nodes are packed into each rack. Each of these trends generates a myriad of open questions, including those of how to best structure and communicate over the underlying network.

Aspen trees

A key issue in today's data centers is quick and efficient failure recovery, but the multi-rooted fat tree topologies that form the basis for many data center networks do not immediately lend themselves to quick failure reaction. In these trees, a single link failure can have devastating consequences, effectively disconnecting a set of end hosts while updated routing information is disseminated to every switch in the topology. This can take a significant amount of time; global re-convergence of broadcast-based routing protocols such as OSPF and ISIS can be in the tens of seconds due to the use of under powered switch CPUs and conservatively-set protocol timers. During this time, packets sent to affected end hosts are lost, crippling applications until connectivity is restored. This type of disruption is unacceptable in the data center, where the highest levels of availability are required.

To address this problem, we introduced Aspen trees, a set of multi-rooted tree topologies with the ability to react to failures locally, and a corresponding failure notification protocol for quick

activation of backup paths. An Aspen tree includes strategically placed areas of denser interconnect than would be found in a traditionally-defined fat tree, at the expense of either scalability (in terms of the number of end hosts supported) or network cost (in terms of the number of switches and links in the topology). In our introduction of Aspen trees[3] we precisely quantify this fundamental tradeoff between fault tolerance, scalability and network cost, and we provide a taxonomy for understanding the range of possible Aspen trees given constraints such as budget limitations or requirements for end host support.

After designing this clean-slate network architecture, I next considered the effects of augmenting a more traditional data center interconnect with highly-specialized processing hardware.

Catapult in the Data Center

The Catapult project[1] introduces a reconfigurable, composable fabric designed to accelerate the most critical portions of a particular data center workload. A key consideration in integrating such a fabric into an existing data center is determining how to interconnect the newly introduced processing nodes. Should these new nodes communicate with one another over the existing data center network that interconnects servers or should a more performant secondary network be built to support out-of-band communication among fabric nodes? If a secondary network is built, can it ultimately replace the original network or should the two co-exist?

I investigated these questions in the context of Bing search workloads within Microsoft's data centers. Through this study, I was fortunate to learn a tremendous amount about practical concerns within an actual data center, from issues of monetary cost to those of physically wiring routers and switches in a space-efficient and maintainable manner. This experience proved invaluable as I began to consider another emerging network architecture, the ultra-dense data center.

Ultra-Dense Data Centers

Recent trends to pack data centers with more CPUs per rack have led to scenarios in which each individual rack may contain hundreds, or even thousands, of small compute nodes such as systems-on-chip (SoCs). At this scale, the traditional rack-level topology composed of a top-of-rack (ToR) switch as a hub and servers as leaves is no longer feasible in terms of monetary cost, physical space, and oversubscription. This is because a ToR that connects thousands of SoCs to one another and to the rest of the data center would require thousands of ports. The technology to build high-radix ToRs is becoming less feasible and more costly as link speeds increase. A ToR with thousands of high-speed ports would be prohibitively expensive to build, would occupy a large portion of the physical space within its rack, and would likely lack sufficient ports to provide an acceptable oversubscription ratio for traffic leaving the rack.

To this end, we introduced Theia as an architecture for these new *ultra-dense* data centers (UDDCs). A key observation in Theia's design is that at such an enormous scale, oversubscription is unavoidable and therefore should be placed in a thoughtful and strategic manner. For instance, with thousands of compute nodes per rack, it may be possible to keep entire workloads or tasks rack-local. Therefore, we focus on designing Theia to enable extremely high performance and cost-efficiency within the rack while allowing for tunable out-of-rack connectivity. One of the primary architectural components of Theia is a low-latency, circuit-style optical patch panel that is used to connect SoCs within a rack to one another. This patch panel is passive and therefore introduces no queuing delay and draws no power; these are important factors in the overall cost of a UDDC. Careful selection of the topology implemented by this patch panel and thoughtful design of the corresponding routing protocols is critical to performance. Our initial efforts[2] focused on the circulant graph as a basis for rack-level connectivity; we have expanded to explore several other graphs as our investigation continues.

We have built a Theia prototype, based on the initial proposal of the circulant graph, and are evaluating it on vendor hardware. This experience has allowed us to understand better the properties of our circulant graph topology and corresponding communication protocols as well as to evaluate and inform the design of cutting edge hardware technology for this emerging network architecture.

Future Work

I have enjoyed the opportunities that I've had at Microsoft Research to consider new architectures for emerging networks, and I've been fortunate to have had the chance to combine the theoretical analysis of such networks with practical experiments and actual implementation. In the case of Theia, we combined an in-depth theoretical study of numerous rack-level topologies with an implementation and evaluation over brand new vendor hardware in order to come up with a practical architecture for UDDCs.

Moving forward, I plan to continue my study of topology and protocol co-design, and to continue to leverage this unique combination of theory, implementation and evaluation to provide efficient and scalable solutions for some of today's distributed systems problems. I look forward to iterating on network designs for UDDCs as new hardware becomes available and as requirements become more crisp. If my career path returns me to Microsoft Research, I will be able to build and evaluate new rack-level topologies for UDDCs, continuing Microsoft's partnership with the current hardware vendor. I will be able to inform the design of the next generation of UDDC racks, including the internals of the patch panel, the structure of the chassis that houses the rack's SoCs, tuning of the in-rack oversubscription levels, and so on. I will have the chance to study a wide variety of real workloads running over these topologies, so that I can determine the best suited topology for each type of workload that may ultimately run in a UDDC at Microsoft.

Alongside these efforts, I plan to explore our ideas for UDDC networks in the space of micro data centers; such mDCs could benefit from the availability of a relatively large number of compute nodes packed into a small physical footprint with low power and cooling requirements. The relatively self-contained nature of mDCs provides a perfect arena for exploring rack-level UDDC topologies and communication protocols. Additionally, mDCs are unique in that the division of in-rack versus out-of-rack bandwidth allocation may be somewhat more static than in a traditional data center rack, allowing a network architect to make this allocation decision a priori and then focus on rack-level performance and tuning.

My experience at Microsoft Research has allowed me to work with cutting edge network architectures and brand new hardware and technology, and has led me to expect continued increases in scale as well as the emergence of new network architectures in the next few years, as cloud applications continue to grow and evolve. I look forward to discovering and tackling the set of challenges introduced by the next generations of data center network architectures.

References

[1] A. Putnam, A. M. Caulfield, E. S. Chung, D. Chiou, K. Constantinides, J. Demme, H. Esmaeilzadeh, J. Fowers, G. Prashanth, G. Jan, G. Michael, H. S. Hauck, S. Heil, A. Hormati, J.-Y. Kim, S. Lanka, J. Larus, E. Peterson, S. Pope, A. Smith, J. Thong, P. Yi, and X. D. Burger. A reconfigurable fabric for accelerating large-scale datacenter services. In *Proceeding of the 41st Annual International Symposium on Computer Architecuture*, ISCA '14, pages 13–24, Piscataway, NJ, USA, 2014. IEEE Press.

[2] M. Walraed-Sullivan, J. Padhye, and D. A. Maltz. Theia: Simple and cheap networking for ultra-dense data centers. In *Proceedings of the 13th ACM Workshop on Hot Topics in Networks*, HotNets-XIII, pages 26:1–26:7, New York, NY, USA, 2014. ACM.

[3] M. Walraed-Sullivan, A. Vahdat, and K. Marzullo. Aspen trees: Balancing data center fault tolerance, scalability and cost. In *Proceedings of the Ninth ACM Conference on Emerging Networking Experiments and Technologies*, CoNEXT '13, pages 85–96, New York, NY, USA, 2013. ACM.