Towards Reconfigurable Rack-Scale Networking

Tyler Szepesi Bernard Wong Tim Brecht Sajjad Rizvi
School of Computer Science, University of Waterloo

Datacenters have traditionally been designed and deployed as a collection of server racks, where each rack consists of approximately 42 units of servers connected to one or two edge switches. In this model, computation is generally localized within a single server, and the edge switches provide 10 or 40 Gbps of non-blocking bandwidth between servers in the same rack.

Although the traditional model is easy to deploy, application writers are increasingly finding that a unit of computation does not conveniently fit within a single server. To address this issue, a new datacenter model, known as rack-scale computing, has been proposed. In rack-scale computing, a rack is not a collection of servers, but is instead a single system consisting of multiple interconnected subsystems, such as CPU, memory, and storage, that are disaggregated into separate locations within a rack. Depending on the design, a subsystem can be made up of components of the same type, where a component is essentially a mini-server with an abundance of a particular resource, or it can be made up of just the raw resources with the network serving as a rack-scale system bus. In this paper, we will focus on subsystems that are made out of components. Through disaggregation, components can be upgraded independently, and datacenter operators can customize the balance between different subsystems without having to replace or upgrade every component.

The key barrier to realizing subsystem disaggregation is the absence of commodity, high-speed rack-scale networks. Current 40 Gbps networks may not have sufficient bandwidth to satisfy the data requirements between different subsystems. Simply increasing the network's bandwidth is not a sustainable solution because it places an extreme demand on the edge switches. For example, a 100 Gbps network connecting 100 different components in a rack-scale computer, where a component is a unit of CPU, memory, or storage, requires an edge switch with a switching throughput of 10 Tbps. Distributed switching or server-centric networks can address this problem, but they introduce additional switching latency that can significantly reduce the performance of latency-sensitive applications.

In order to achieve the goals of rack-scale computing, we believe there is a critical need for reconfigurable rack-scale networks that can dynamically change their topology to provide high-bandwidth, direct connectivity between different components in response to changing workloads. Although there has been a significant body of work on reconfigurable networks [1, 2, 3], they have all approached the problem by reconfiguring the connectivity between switches to reduce or eliminate the impact of network oversubscription. We believe that, for rack-scale computers, a more effective approach is to have more edge switches, where each switch is responsible for only a small group of components, and to dynamically reconfigure the group membership of the switches, as illustrated in Figure 1. This approach provides non-blocking connectivity between components within the same group, and allows group membership to be adjusted in response to changing workloads. It is motivated by the observation that there is far stronger communication locality between small groups of components than there are between larger groups. Therefore, changing group membership is often more effective than changing switch connectivity.

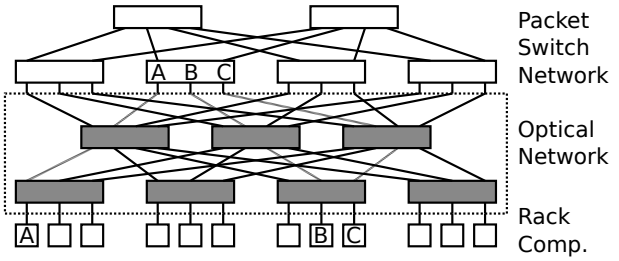

Figure 1: The components A, B, and C form a group by connecting to the same edge switch through the reconfigurable optical circuit switch network.

To implement dynamic group membership changes, we propose using a 2-stage Clos network of optical circuit switches to connect the rack components to an oversubscribed 2-tier packet-switched network. The optical network enables a rack component to be directly connected to any edge switch, and the 2-stage design allows for the use of smaller, commercially available optical circuit switches. By providing group membership reconfigurability, the 2-tier packet-switched network can be highly oversubscribed; near-saturation of the oversubscribed links can used as a signal to trigger a group membership change. An alternative design is to create a completely packet-switched folded-Clos network. However, this would increase packet switching latency within a group. Furthermore, at 100 Gbps, optical circuit switches are less expensive than packet switches.

Comparing the cost of our design with OSA [1], our design requires more optical circuit switches, but does not require any optical WSS or MUX/DEMUX hardware. More importantly, because our design requires less bandwidth between switches, each edge switch can be nearly half the size compared to OSA. This is a significant cost reduction, and more than covers the cost of the additional optical circuit switches.

References

[1] K. Chen, A. Singlay, A. Singhz, K. Ramachandranz, L. Xuz, Y. Zhangz, X. Wen, and Y. Chen. OSA: An Optical Switching Architecture for Data Center Networks with Unprecedented Flexibility. In *NSDI*, 2012.

[2] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat. Helios: A hybrid electrical/optical switch architecture for modular data centers. In *SIGCOMM*, 2010.

[3] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. E. Ng, M. Kozuch, and M. Ryan. c-through: Part-time optics in data centers. In *SIGCOMM*, 2010.