Communications-aware job scheduling in hybrid
SDN datacenters

Malathi Veeraraghavan∗, Nelson Fonseca†, Andrea Fumagalli‡
∗University of Virginia (UVA), Charlottesville, VA 22904, USA
†Institute of Computing of The University of Campinas, Campinas, Brazil
‡The Univ. of Texas at Dallas, Dallas, TX 75080, USA

Topic areas: Software-defined networking (SDN), and Cloud computing and data center networking

Background: Optical-switched networks consume lower levels of energy when compared to electronic-switching networks. Hybrid electronic-optical networks have been proposed for datacenters, such as Helios [1], c-Through [2], and OSA [3]. However, owing to the cost of DWDM networks, each lightpath (wavelength circuit) is typically operated at 40 or 100 Gbps. This constraint allows only a few top-of-rack (TOR) switch (typically Ethernet or InfiniBand) pairs to be interconnected via optical circuits at any given time. A new technology that supports lower circuit rates while simultaneously consuming lower energy, called Digital Subcarrier Multiplexing (DSCM), was proposed [4]. With DSCM, 100 subcarriers can be multiplexed on a single wavelength. Assuming 160 wavelengths can be carried on a single fiber, DSCM can support 16,000 channels per fiber. With OFDM, 512 subcarrier channels are carried on a 40Gb/s system, which implies that circuit rates on the order of tens of Mbps can be supported. Digital Subcarrier Cross-Connect (DSXC) switching times are in the sub-milliseconds range. In summary, the digital subcarrier technology has the potential to support higher-granularity circuits, while simultaneously consuming low energy, and is hence considered in this study for use in hybrid datacenter software-defined networks (SDNs).

Problem statement: In commercial data centers, applications use Hadoop MapReduce for analyzing large datasets. In scientific data centers, computing jobs that require thousands of cores are submitted to batch schedulers such as PBS, LSF, or SLURM. These schedulers assign the jobs corresponding to a single application to different hosts, which could be distributed across several racks. Highly parallelized applications are often limited by network throughput when using 1000s of cores. The problem statement of this study is to improve communication performance within datacenter networks, while simultaneously reducing energy consumption.

Solution approach: We plan to use an approach proposed by Wang et al. [5] in which application-awareness is leveraged to determine circuit allocations between ToR electronic (packet) switches. Specifically, network-aware job scheduling is proposed for Hadoop jobs with “rack-based bin-packing placement to aggregate tasks on to a minimum number of racks.” Such a job placement method then allows for ToR switches to be interconnected via an optical circuit for faster inter-task communications. An SDN controller allows for an easier integration of the Hadoop job scheduler with the scheduling and provisioning of optical circuits. Our DSCM technology allows for a greater number of racks to be interconnected simultaneously. Our project will study Hadoop job communication patterns and design a solution that integrates a job scheduler, SDN controller, and DSCM/WDM optical switches. A recent paper by Ren et al. [6] offers a public dataset characterizing Hadoop workloads, which will serve as a starting point for our communication pattern analysis. We also plan to run our own Hadoop Mapreduce software for network traffic analysis as was done by Lee et al. [7], and collect network usage statistics.

Other research topics: We are also working on the other two topic areas of the SwitchOn workshop, Cyber-infrastructure, and GENI and other federated testbeds. Specifically, an NSF MRI grant was used to deploy a distributed Dynamic Network System (DYNES). Leveraging the DYNES equipment, and equipment deployed through the NSF CC-NIE program grants, we are creating inter-domain dynamic Layer-2 (L2) virtual circuits between campuses. High-speed file transfer applications are being tested. We are also developing a solution for inter-domain Layer-1 (L1) end-to-end optical circuits and will be testing 10 Gb/s transfers on an Internet2 supported testbed. We would like to explore creating inter-domain L2, and if possible L1, circuits from UVA campus to Campinas campus.
REFERENCES


