Rethinking Networking in a Non-volatile, Heterogenous World

Satyajayant Misra and Mai Zheng
Computer Science, New Mexico State University

The increase in the number of sources of data collection and data precision is resulting in a constant and rapid increase in science data and its transfer volumes. If we look at anticipated data volumes at three levels: new data set size, LAN transfer requirements, and WAN transfer requirements–the projections show that in the near term of 0-2 years these values will be 6 TB/day, 5 TB/day, and 200 TB/day respectively. Looking at the 5+ years horizon these values would be 100 TB/day, 200 TB/day, and 3 PB/month at 10 GB/sec respectively [2]–the growth increase is almost ten-folds every two years. This growth rate supercedes that of semiconductors' and of the networking infrastructure and protocols, a significant future challenge to the networking and systems research community. Particularly, we envision two intertwined technology trends that will affect the scalability both positively and negatively:

What if every byte becomes non-volatile? DRAM and hard disks have been used for buffering and permanent storage respectively for decades. Existing networked systems have been iteratively optimized for this two-layered storage hierarchy. For example, a Linux server typically queues the IP packets received from the network interface controller (NIC) in a DRAM ring buffer. A user program has to explicitly invoke system calls to write to disk for data persistence. These communication procedures for transferring, buffering, and storing will change fundamentally with the emergence of non-volatile memory (NVM) technologies, which will provide both persistent storage and DRAM-like access latency. The NVM technologies will help simplify the memory hierarchy. The persistence will also negatively affect failure handling; all errors/corruptions during communications will be persistent. Restarting a server program or rebooting a crashed machine will no longer bring the distributed system into a valid state; sophisticated checkpointing and rollback mechanisms have to be designed for recovery.

What if CPU is no longer the central processor? Similar to graphics processing units (GPUs), modern solid state drives (SSDs) contain advanced processors and are capable of offloading complex data processing tasks from CPU. Moreover, aggressive PCIe-based device control and communication protocols are being standardized [1] to unify devices (including NIC) and make SSDs the center of data processing through device-to-device communication (i.e., bypassing the CPU).

What are the challenges? These new functionalities will surely help with the scalability problems, but to a limited extent. For a more robustly scalable solution, network challenges (both transport and network layers) have to be addressed at two scales: networking within a node—effective networking of the different device controllers, e.g., extending the network-on-a-chip perspective; and networking heterogeneous nodes/clusters over the network. For example, identifying networking paradigms that would make collaborations between nodes containing data and the ones performing collaborative computation efficient, not only within the respective groups but across these groups. The proposed solutions would need to have mechanisms for efficient large datasets transfer, coordinated data placement, and coordinated computation placement at the Internet scale. Drs. Misra and Zheng have a unique combination of expertise in systems and networks, and will be able to approach the problem holistically by looking at the interaction between the devices and the network. This workshop will give them a chance to interact with the DoE Science stakeholders, the scientists and researchers, who dictate the broad requirements for this system-network collaboration.

References

[1] NVM Express. `http://www.nvmexpress.org/`.

[2] J. Zurawski. Bridging the technical gap: Science engagement at esnet. In *Great Plains Network Annual Meeting*, 2015.