Lecture 12 Summary

Two papers were presented in class: Safe and Effective Fine-grained TCP Retransmission for Datacenter Communication and I/O Performance Challenges at Leadership Scale. The first paper introduced a problem facing high-fan-in, high-bandwidth synchronized TCP workloads in datacenter Ethernets-TCP incast collapse. This problem is that application throughput drastically reduces when multiple senders communicate with a single receiver in high bandwidth, low delay networks using TCP, caused by highly bursty, fast data transmissions overfilling Ethernet switch buffer. And therefore, authors provided a practical, effective and safe solution to eliminate this problem. Before getting started, they defined some preconditions: 1. High-bandwidth, low-latency networks with small switch buffer; 2. Clients that issue barrier-synchronized requests in parallel. 3. Servers that return a relatively small amount of data per request. Then, the authors provide three solutions which require no re-implementation of some features in TCP: 1. Large switch buffers could delay the onset of incast collapse. 2. Ethernet flow control was effective when the machines were on a single switch, which was dangerous across inter-switch trunks because of head-of-line blocking. 3. Reduce TPC’s minimum RTO.

The second paper, the authors present a case study of the I/O challenges to performance and scalability on Intrepid, the IBM BG/P system. They introduced the system environment of Intrepid: I/O architecture and I/O software. For I/O software part, they also introduced some info about collective I/O, state avoidance, consistency and distributed metadata. The experiments were performed during an acceptance period before the production began, which will use PVFS 2.8.0. The results were gathered after tuning each component of the I/O stack. For component analysis, these results show that the I/O system has a reasonably balanced provisioning, where storage component peak rates are reached only when applications perform runs utilizing almost half of the machine. For scaling analysis, there were two benchmarks: IOR benchmark from LLNL to measure aggregate I/O throughput, and metarates benchmark from UCAR. For application benchmarks, there are 3 parts: BTIO, MADbench2 and Flash3 I/O. The results still show that systems continue to struggle with the difference between the large, contiguous, aligned I/O patterns. And storage systems of tomorrow need to handle wide variations in the I/O patterns and utilization of the storage system. Systems also face new challenges as unprecedented process counts and storage components increase the complexity of the storage system and the likelihood of component failures. Storage systems need to handle component faults seamlessly while file systems must be designed to avoid single points of contention.