End-System Awareness and Transport-Layer Convergence

Nathan Hanford Chrisopher Nitta Dipak Ghosal Matthew K. Farrens
Venkatesh Akella

1 Introduction

Network throughputs are continuing to grow. A 400 Gbps Converged Ethernet standard is on the horizon (IEEE P802.3bs). Features are being added to commodity multicore processor architectures at a brisk pace. Power-saving throttling and turbo-boosting allows the speed of a processor to reflect the incoming workload. GPUs are also being included on commodity processor cores. Memory, including non-volatile memory (NVRAM) is becoming less expensive, faster, and increasingly ubiquitous. Scale-out systems, including HPCs, are becoming increasingly internally heterogeneous. Finally, miniaturization is leading to the convergence of multiple discrete micro- and macro-architectures into fewer and fewer structures. For example, in cloud computing, storage has converged with compute resources in clustered storage. Naturally, this leads to a convergence of multiple paradigms. For example, separate storage and data transfer networks have converged into a single data-plane.

2 Science DMZ Challenges

The future of Science DMZ will be shaped by the scientists who use the underlying networks, and their home institutions. An increasing number of institutions are converging their own HPC capabilities across departments into on-campus HPC centers. While this allows for excellent opportunities for collaboration, it means that network data-plane demands are changing significantly. By 2025, it means that there will be a demand for more data, with more throughput, and less latency, stored in more places. It will be challenging to maintain data integrity, security, and availability. Through convergence, there will also be the opportunity to get data from experimental equipment to HPC computer models, and return the results in a fraction of the current time.

3 Proposed Research

Flexible, end-system aware protocols are the key to meaningful performance enhancements for Science DMZ. Using burst-mode processing within an end-system, a distributed system architect can provide statistical service time guarantees for packet processing in poor conditions, where an end-system would typically be overwhelmed. Burst-mode processing allows for a CPU core to temporarily run at a multiple of its typical clock speed. In an environment where data transfer and data processing is occurring on the same core (for purposes such as integrity checking), this can allow an end-system to avoid a cascading slowdown of protocol and application processing. RDMA over Converged Ethernet (RoCE) and Intel's Data Plane Development Kit (DPDK) are examples of lower-layer frameworks which we can build upon to provide high performance, while still maintaining data integrity and network survivability. These frameworks allow us to create transport-layer protocols-as-applications that are capable of taking advantage of the latest commodity hardware advancements. By 2025, Integrated Network Interface Controllers (INICs) will most likely achieve wide deployment. When one views the protocol semantics of database, storage, on-chip network, memory, I/O, storage network, LAN, and WAN protocols, it becomes plausible that a single transport-layer protocol and a single session-layer protocol could, if flexible enough, allow for the convergence of all protocol processing onto a single processor type. The result would finally tie together software-defined networking (SDN), software-defined storage, and flexible I/O architecture design in a way that will not only allow for the next generation of big-data convergence, but also for the next generation of big-data peripherals. If ESnet and ScienceDMZ provided this, it would serve as a hardware and systems blueprint not only for distributed scientific computing, but any big-data application, including smart grid and beyond.