Datacenter networks have changed radically in recent years. Their bandwidth and latency has improved by orders of magnitude, and advanced network devices such as NICs with Remote Direct Memory Access (RDMA) capabilities and programmable switches have been deployed. The conventional wisdom is that to best use fast datacenter networks, distributed systems must be redesigned to offload processing from server CPUs to network devices. In this dissertation, we show that conventional, non-offloaded designs offer better or comparable performance for a wide range of datacenter workloads, including key-value stores, distributed transactions, and highly-available replicated services.

We present the following principle: The physical limitations of networks must inform the design of high-performance distributed systems. Offloaded designs often require more network round trips than conventional CPU-based designs, and therefore have fundamentally higher latency. Since they require more network packets, they also have lower throughput. Realizing the benefits of this principle requires fast networking software for CPUs. To this end, we undertake a detailed exploration of datacenter network capabilities, CPU-NIC interaction over the system bus, and NIC hardware architecture. We use insights from this study to create high-performance remote procedure call implementations for use in distributed systems with active end host CPUs.

We demonstrate the effectiveness of this principle through the design and evaluation of four distributed in-memory systems: a key-value cache, a networked sequencer, an online transaction processing system, and a state machine replication system. We show that our designs often simultaneously outperform the competition in performance, scalability, and simplicity.

Thesis Committee:
David G. Andersen, Chair
Garth A. Gibson
Michael Kaminsky
Miguel Castro, Microsoft Research