Shoal: Network Fabric for Next Generation High-density, Disaggregated Racks
Vishal Shrivastav †, Asaf Valadarsky *, Hitesh Ballani †, Paolo Costa †, Ki Suh Lee †, Han Wang †, Rachit Agarwal †, Hakim Weatherspoon †
† Cornell University * The Hebrew University of Jerusalem ✦ Microsoft Research Cambridge, UK

Enabling Technology via High-density Racks
- Comprises 100s-1000s of micro-servers connected via an internal network fabric
- Leverages hardware technologies like SoCs, which are extremely space and power efficient, to achieve high density

New Cloud Scaling via Resource Disaggregation
- Decoupling of CPU, memory, and storage, interconnected via an internal network fabric
- More efficient (fine-grained) resource provisioning
- Circumvent “Memory-capacity Wall”

Goal
Build a high-performant network fabric for high-density racks with disaggregated resources, at low power, cost and complexity

Challenges
- How to connect ~1000 nodes subject to space and power constraints of a rack?
- How to achieve predictable and low latencies to support converged traffic (comprising IP, storage, and memory traffic)?

Fabric Design
- At physical layer, fabric comprises simple circuit switches with no buffers, no arbitration and no packet inspection mechanisms
- Very precise rack-wide time and frequency synchronization
- Nodes send fixed-sized packets and switches reconfigure circuits, according to a static schedule

Routing & Congestion Control
- 2-hop routing scheme, emulating a load-balanced Birkhoff-von Neumann switch
- Packets are buffered at the intermediate node in the forwarding buffers, one for each destination
- Achieves bounded worst-case system throughput of 50% across any traffic pattern
- Very simple back-pressure based congestion control mechanism
- Achieves fair bandwidth allocation, bounded queuing and hence a lossless network

Prototype
- Built a 10G network with 4 nodes using FPGA-based custom NICs and layer-1 circuit switches

Simulation Results (512 nodes)
Shoal is equipped with 2x bandwidth, but the estimated total cost of the fabric is still less than that of an equivalent packet switch network