Disaggregated Racks

Why Disaggregation?
- High compute density
- Fine-grained resource pooling and provisioning
- Seamless scaling and independent evolution of resources

Challenges for Disaggregated Rack Network
Connect typically an order of magnitude more nodes than traditional racks

1. Needs to be high performant
2. Needs to be power efficient

Strawman Network Designs
- Packet-switched Network
  - High performance but high power
- Direct-connect Network
  - Low power but poor performance

Shoal
Built out of a collection of fast circuit switches that can reconfigure within nanoseconds

Fabric Reconfiguration
- De-centralized, traffic agnostic reconfiguration
- Nodes send fixed-sized packets and switches reconfigure circuits according to a static, per-defined schedule
- Requires very precise rack-wide synchronization

Design
- 2-hop routing scheme: load-balancing followed by forwarding to final destination.
- Packets are buffered at the intermediate node in the forwarding buffers, one per destination
- Worst-case network throughput of 50% compared to an ideal packet-switched network
- Very simple back-pressure based congestion control mechanism
- Achieves fair bandwidth allocation, and bounded queuing across all traffic patterns

Prototype
- FPGA-based custom NICs and circuit switches
- Circuit switch can reconfigure in < 6.4 ns
  - Altera Stratix V FPGAs
  - Bluespec System Verilog

Evaluation (512-node rack)
Shoal consumes 3.5x less power, while achieving better network performance than packet-switched network

Circuit switches consume significantly less power than packet switches
Performance bottlenecked by reconfiguration