Supercomputers

Interconnects

- *Type:*
  - Infiniband
  - Myrinet (Declining)
  - Custom (Blue Gene)
  - Gig E/10 Gig E
- *Use: Zero-copy architectures*
  - TCP/IP? RDMA/VIA?

Interconnect Speed

- GigE: 1 Gbps, 40µs, $40/port
- 10GigE: 10Gbps, 40µs?, $1000/port
- Infiniband
  - 8Gbps 4µs, $600/port
    16Gbps 4µs, $700/port
    32Gbps, 4µs, $1200/port

What's different?

- Hardware-offloaded reliability, in-order delivery, encapsulation, etc.
  - You could do this with TOE/iWARP now
- End-to-end credit-based flow control
  - (Prevents "Incast" problem...)
- Different application use model -- RDMA and VIA

VIA/RDMA

- VIA: Virtual Interface Architecture
  - Per-process memory-mapped access to network card. => Can send (and recv...) without kernel intervention or context switch or copy. Awesome if hardware takes data and sends it.
  - Note many copies already avoidable - *sendfile*, page swapping, etc., if application cooperates
- Remote Direct Memory Access
  - Put/get data directly into remote process mem
  - Pretty cool for DSM, file server access, etc.

Or the BlueGene Approach

- Max pkt size: 256 bytes
- To send data: Load into 2 Floating Point registers, direct load to network FIFO SRAM
- Hardware ACKs and flow control