Dull Networks are Interesting

October 4, 1994

Greg Papadopoulos
Chief Scientist
Server Systems Engineering
Sun Microsystem Computer Corporation

How to Make a Boring Network

- Total Overhead of the order of cache misses
- Total Latency of the order of local DRAM
- LogP valid: only end-point contention matters

Looks, smells and tastes like a crossbar memory switch...

Realities

- Overhead dominates other latencies
- Total Latency >> local cache miss
- Topologies are too interesting

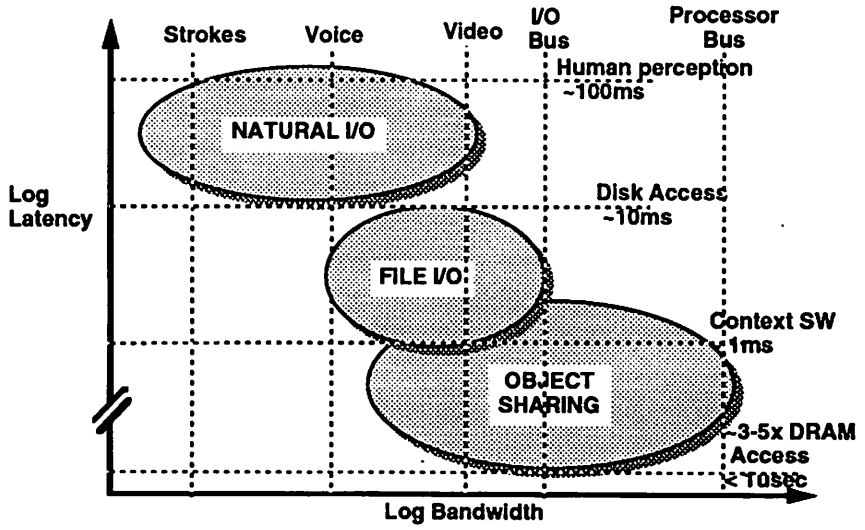
Reducing Overhead

- No kernel calls in the common case
  - just like VM: kernel for set-up and faults; user for dereference
- Watch out for "hidden" flow control costs
  - uncachable reads to check status
  - vagaries of NI buffering
- NI must look more like memory and less like PIO
  - strong ordering and precise bus exceptions increasingly expensive!
  - can have a big affect on "g"
- No kernel calls in the common case

Most Topologies of Interest are Too Interesting

- "Bad" communications patterns are commonplace:
  - N-to-M I/O reordering
  - sorting
  - transposition
  - cshifts
- Meshes (particularly DOR), Butterflies, Hypercubes are vulnerable to one or more of the above
- Randomness plays a crucial role in destroying singularities
- BTW, composition across patterns and users must be uninteresting as well!

Communications Spectrum



Two Fixpoints

- "Globally sharable PCI busses"
  - Nodes can perform block I/O operations on some other node's I/O bus.
  - BW tracks I/O busses
  - Latency dominated by transmission time for big blocks
- "Globally sharable DRAM"
  - Nodes can perform cache-line size operations on other nodes memory banks
  - BW tracks memory bank (processor bus) bandwidth
  - Latency tracks DRAM access times (3-5x)

KEY: Depend upon and exploit solutions to latency tolerance for uniprocessors.

Flames

- It's the overhead, stupid
- Being able to tolerate latency is no excuse for introducing it.
- Keep interconnect total latency to small multiplier of local latencies.
- Make sure that interconnect is otherwise uninteresting.

Bottom line:
Scalable I/O systems seem possible with switched-LAN latencies
Scalable memory systems require aggressive engineering, but seem possible assuming solutions for uniprocessor latency tolerance.