Adaptive Routing Strategies for Modern High Performance Networks

Patrick Geoffray
Myricom
patrick@myri.com

Torsten Hoefler
Indiana University
htor@cs.indiana.edu

28 August 2008
Hot Interconnect
Stanford, CA
Problem

• Vendors are liars.
• They claim full bisection networks.
  – Full bisection: Cut the world in half, any node in first half can communicate with any node of the second half, at full speed.
• Full bisection on paper only.
  – Number of links in any bisection is at least the number of pairs.
  – Clos networks, Fat trees.
• Poor effective bisection.
  – Head of Line blocking !
• Practical solutions to reduce HoL blocking ?
  – Adaptive routing.
**Clos Networks**

- Multiple paths between pair of nodes.
- Example: 3-hop rearrangeable non-blocking Clos network with 32-port crossbars.

> For any given bisection pattern, there is at least one set of non-blocking routes.
Context

• Source-routing:
  – Path in the network is chosen on the sender.
  – No routing decision at each hop.
  – Routes should be deadlock-free.
  – Routes can be changed on a per-packet basis.

• Backpressure flow-control:
  – Bounded per-port buffering on each crossbar.
    • Never big enough.
    • Don’t talk to me about QoS.
  – Ultimately, flow-control can propagate to sender NIC.
    • Cheap way to sense contention.
    • Hard to determine where the blocking is in the path.
Simple Routing Strategies

• Static routing:
  – Single route per destination.
  – Links globally load-balanced across routes.
  – Everything is in order on the wire.
  ➢ Very good for a few patterns, very bad for a few others, and not great for most.

• Random oblivious routing:
  – Multiple routes (16).
  – Route changes randomly for each packet.
  – Packets may not arrive in-order.
    • Higher level protocols should not be dumb enough to require order on the wire.
  ➢ Statistically average for all patterns.
Adaptative Routing Strategies

- **Adaptive routing:**
  - Multiple routes.
  - Contention is sensed with back-pressure.
  - Route changes after sensing contention on the current path.
  - New route is chosen randomly.
  - When low contention, converges to static routing. With high contention, degenerates into random oblivious routing.

- **Probing adaptive routing:**
  - Same as adaptive, but...
  - New route is first probed to check if path is free.
  - Similar to adaptive, but should converge faster to non-blocking set of routes, if it does.
Testbed/Benchmark

- 512-node Myrinet cluster at University of Southern California.
  - Single 21U 512-port switch, Clos network, 32-port crossbars.
  - One single-port Myri-10G NIC in each Xeon-class node.
  - MX-1.2.7 (16 routes per peer in route table).
  - Variable node counts (leaf crossbar granularity).

- Effective bisection benchmark.
  - Randomly split the nodes in two groups of equal size.
  - Randomly pair up nodes between both groups.
  - Measure the bandwidth for 50 iterations of MPI_Sendrecv of 1 MB messages (pair-wise exchange).
  - Lather, rinse, repeat 5000 times.

- Results: Min/Avg/Max of all pair-wise bandwidths, for several nodes counts.
Static Routing
Random Oblivious Routing

![Graph showing effective bisection bandwidth (%) vs number of nodes. The graph has three lines representing minimum, average, and maximum values. The number of nodes ranges from 16 to 512.]
Adaptive Routing

![Graph showing effective bisection bandwidth vs. number of nodes]
Probing Adaptive Routing

![Graph showing effective bisection bandwidth (%) against number of nodes. The graph includes minimum, average, and maximum values.]
Comparing Routing Strategies

[Graph showing the comparison of different routing strategies (static, naive, adaptive, probing) against the number of nodes (16, 32, 64, 128, 256, 512). The x-axis represents the number of nodes, and the y-axis represents the effective bisection bandwidth (%).]
Conclusions

• Static routing is bad.
  – InfiniBand, most Ethernet switches.

• Random routing is deterministic, better at scale.
  – Require decent protocols that do not require order.

• Adaptive routing is better.
  – Probing is necessary for good performance.

• Ultimately, probing adaptive routing does not scale for very large fabrics.
  – Per-hop routing decision, hardware support (Quadrics).

• Things we didn’t do:
  – How fast does the routing converge? Does it converge?
  – What about small/medium messages?
  – What about more than 3-hop Clos networks?
A bit of hope

- Topology-aware collectives:
  - Limit domain space, faster/consistent convergence.
  - Leaf crossbar granularity in Clos networks: bridge pattern.
Bridge pattern