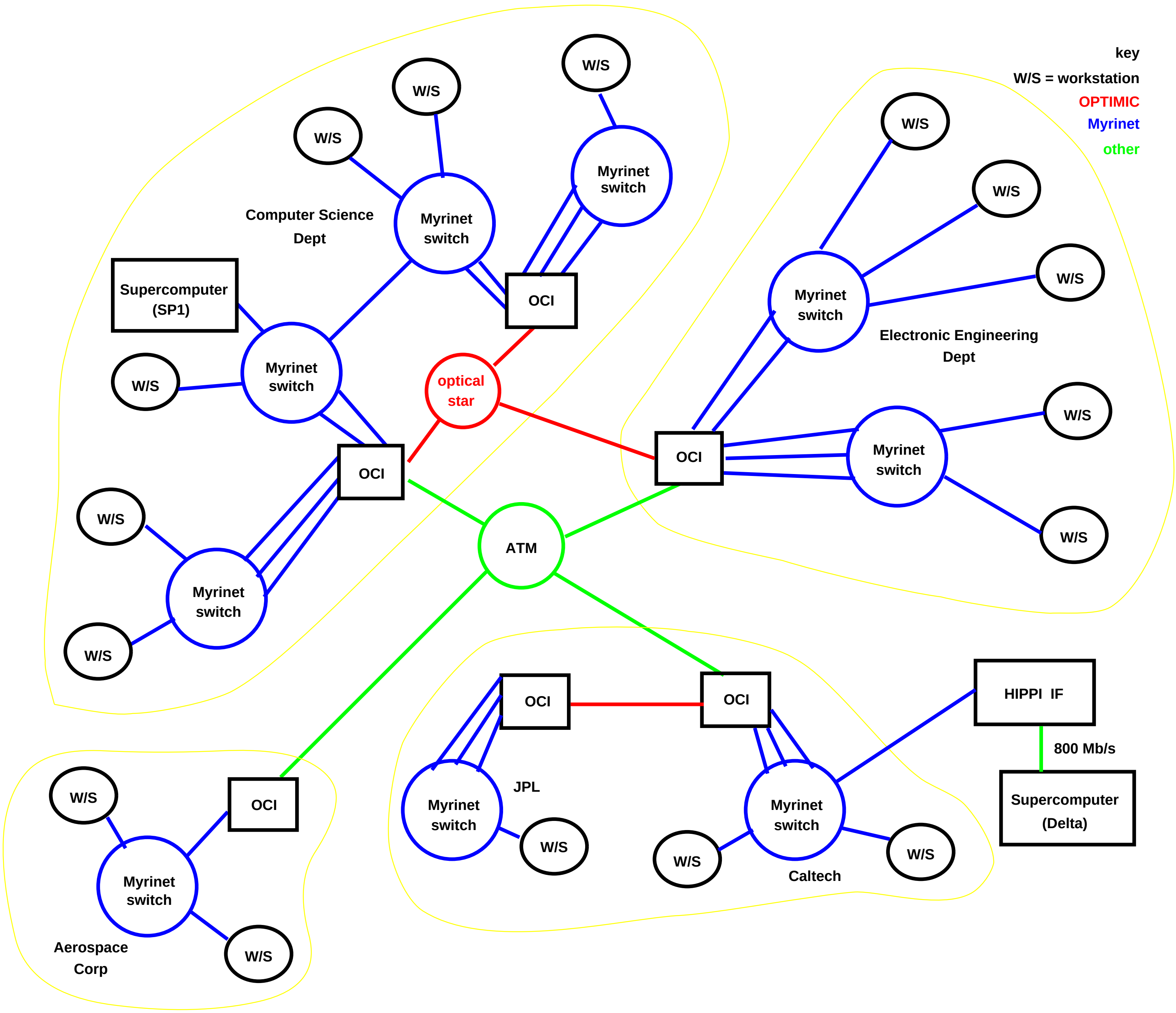
Multicasting Protocols for High-Speed, Wormhole Routing Local Area Networks

Mario Gerla, Prasasth Palnati, Simon Walton,
(University of California, Los Angeles)

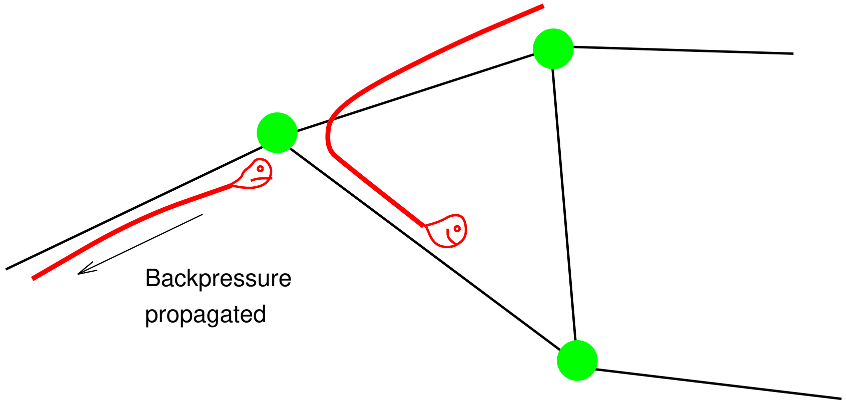
THE SUPERCOMPUTER SUPERNET

University of California at Los Angeles, Jet Propulsion Laboratory, Aerospace Corporation

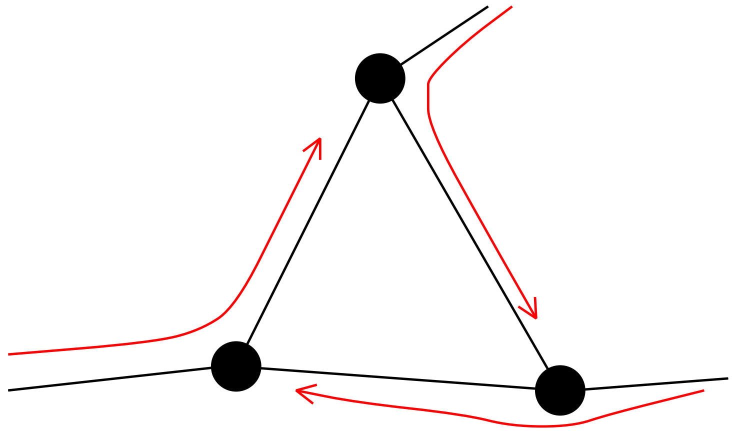


Wormhole Routing

- Wormhole routing employs cut-through and link-by-link flow-control
- The Myrinet wormhole LAN employs source routing – the packet destination address is a sequence of switch port numbers



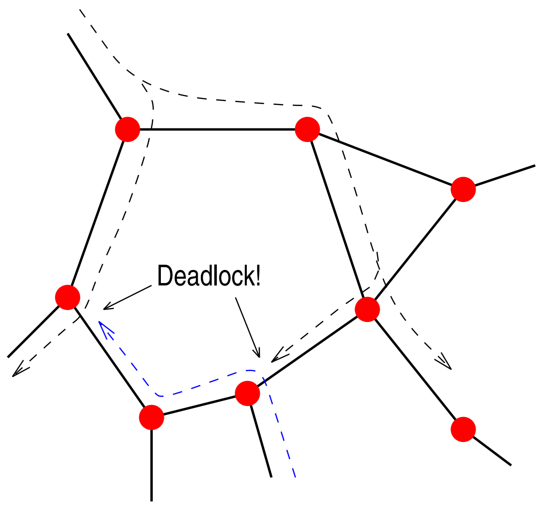
Deadlock in Wormhole Networks



- Deadlock in wormhole networks can occur when cycles of contention for switch output ports arise

- This is prevented in the case of Myrinet by the use of a restricted routing scheme known as up/down routing.

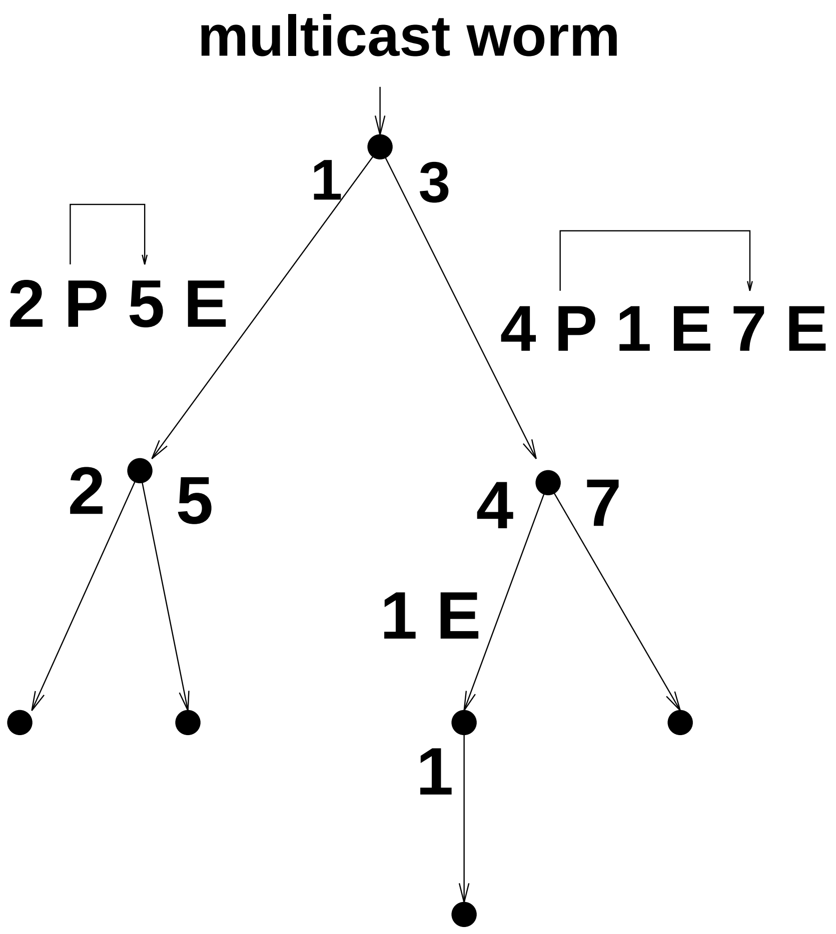
Multicasting within the Network Fabric



- It is possible to extend cut-through routing in the switch to achieve multicast at the switch level.

- The multicast worm defines a tree through the switch mesh. Backpressure applied to any outgoing branch must be propagated back.

- Difficulties include increased complexity of source route encoding and possibility of deadlock amongst different branches of same worm.



Need to encode multicast tree as linear list for source routing

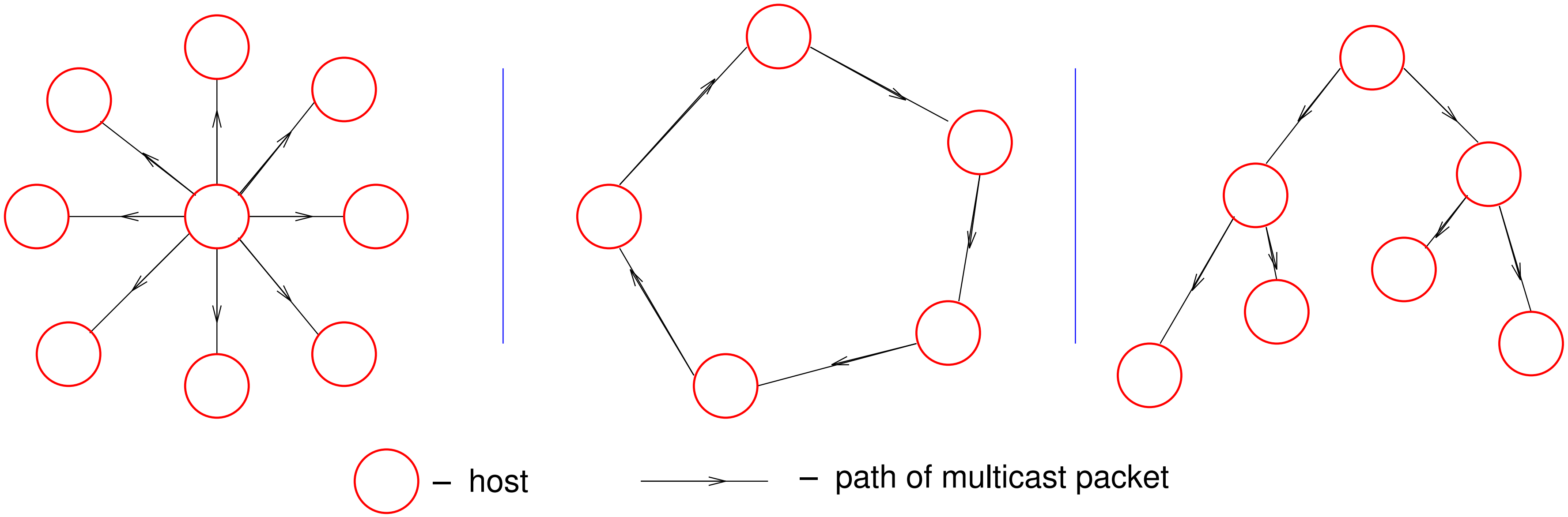
Use leftmost, depth-first traversal

'P' is pointer, encoded as byte offset

'E' is end marker

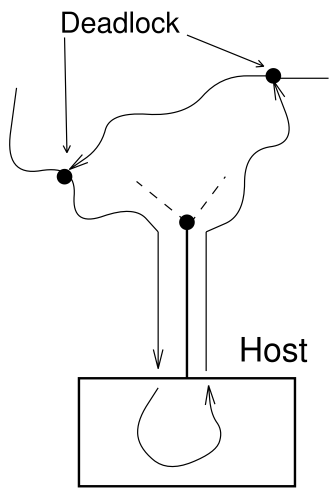
Encoded source route: 1 P 2 P 5 E 3 P 4 P 1 E 7 E

Multicasting by Host Forwarding

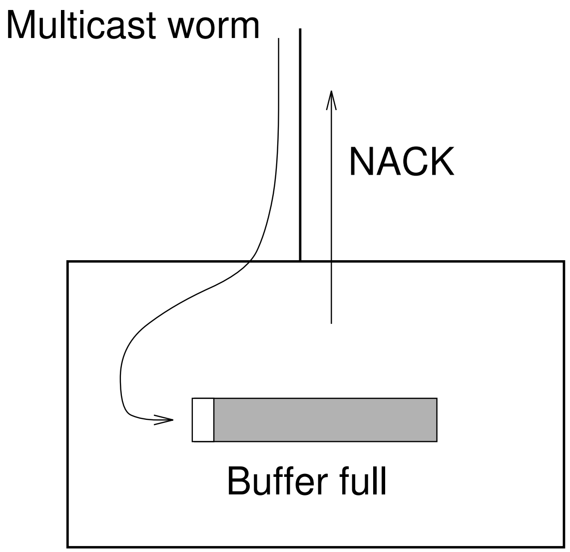


- Use multiple host forwarding to achieve multicasting
- Various schemes shown above – forwarding from source host, forwarding around Hamiltonian cycle, forwarding down binary tree
- Worm may be stored and forwarded or cut-through

Deadlock Prevention with Buffer Reservation



- Worm forwarding can reintroduce deadlock
- Use buffer reservation protocol at each hop to prevent deadlock within switch fabric.

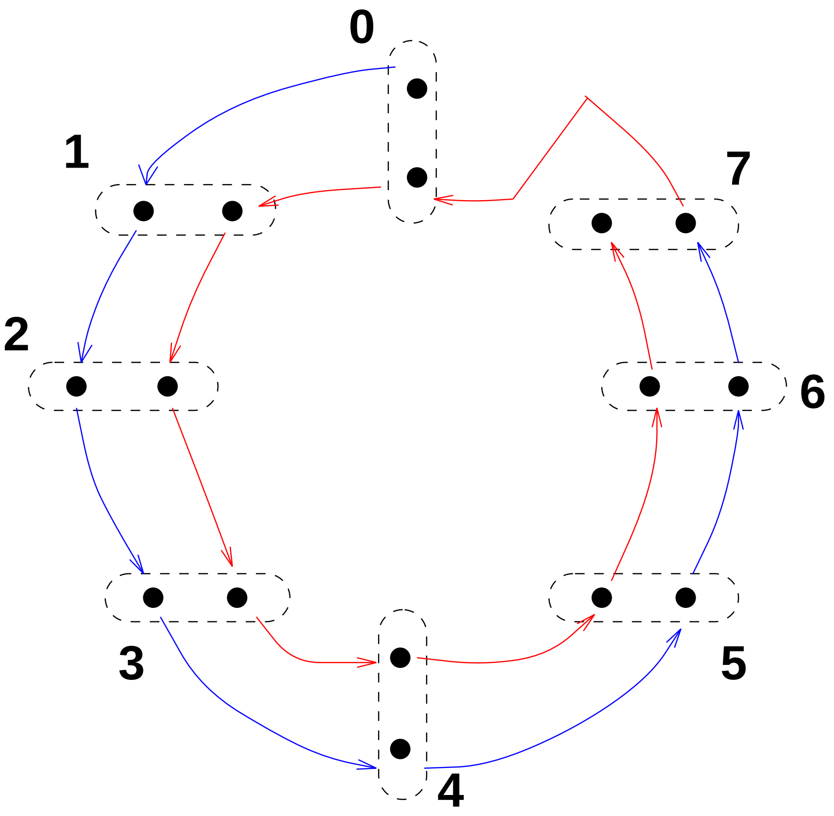


- Each host has buffer for largest multicast worm-size.
- Can reduce latency with ‘optimistic’ reservation protocol – worm transmitted immediately; NACK returned by next hop if buffer space not available

Deadlock Prevention by Buffer Reservation

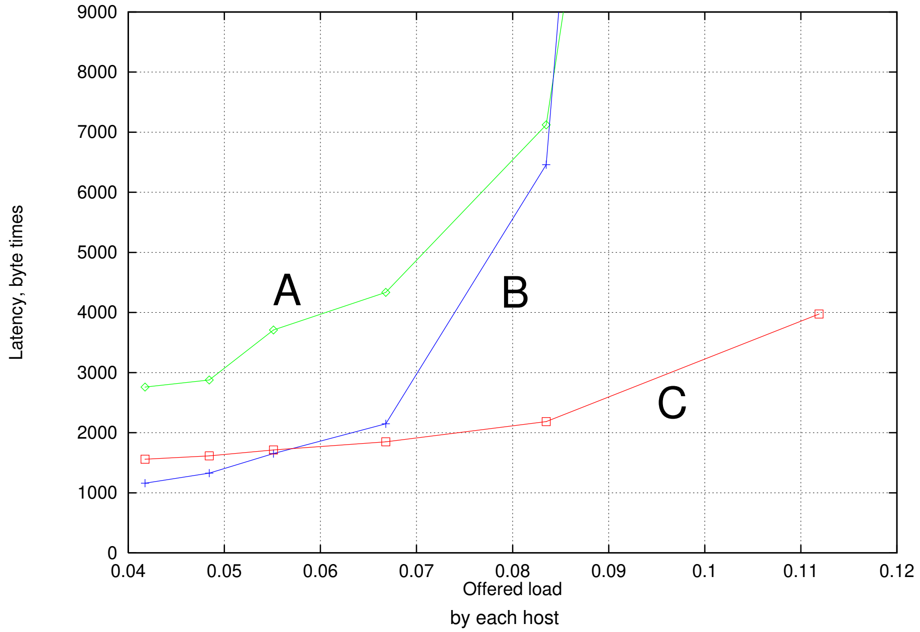
- Buffer reservations to prevent network deadlock

- Two buffer classes used to prevent buffer reservation deadlock



Host buffers are divided into two classes

Simulation Results for Host Multicasting



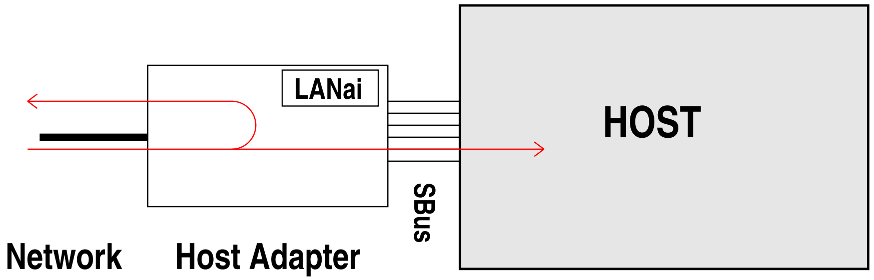
Simulations performed using Maisie simulator at byte level. Network modelled: 8x8 torus; 10 multicast groups, each with 10 members; offered load is fraction of link speed per host. Mixed multicast and unicast traffic.

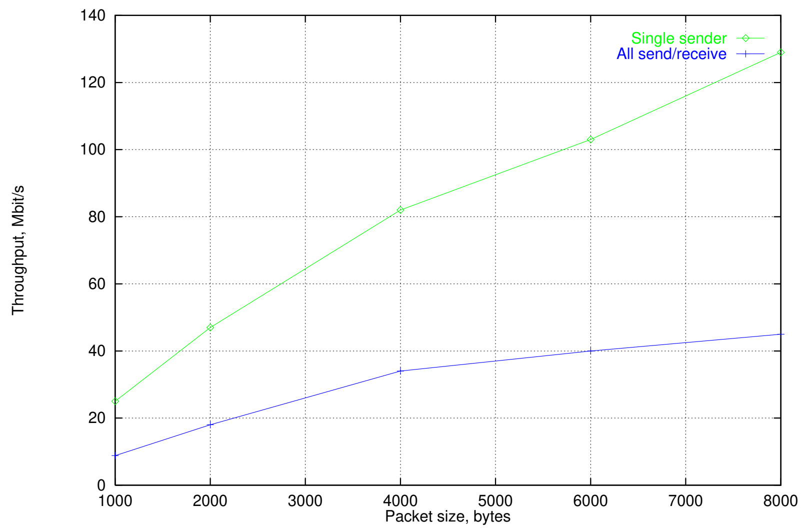
A – multicast around cycle, store and forward

B – multicast around cycle, cut-through at host

C – multicast over binary tree

Implementation of Multicasting in Myrinet





- Hamiltonian circuit multicast scheme implemented
- Worm forwarding done in host adapter
- Experimental results show that host adapter is bottleneck

Conclusions/Further Work

- Switch-fabric based solutions have the lowest latency, but impose a high degree of complexity upon the switches
- Multicasting based on host-forwarding is more amenable to implementation – main issue is deadlock prevention without degradation of performance
- Work in progress on evaluation by simulation of buffer deadlock prevention schemes

Contact Points

Mario Gerla `gerla@cs.ucla.edu`

Prasasth Palnati `palnati@alpo.casc.com`

Simon Walton `simonw@cs.ucla.edu`

Supercomputer Scalable Network

`http://millennium.cs.ucla.edu/~ssn`

Computer Science Dept., UCLA

`http://www.cs.ucla.edu`