

**Simulation studies of
Gigabit Ethernet
vs.
Myrinet
using
Real Application Cores**

Pete Wyckoff and Helen Chen

Distributed Computing Research
Sandia National Laboratories
Livermore, CA

Overview

→ *Choose appropriate network for cluster*

- Evaluate available networks
- Consider end applications

- Interconnect technologies
- Simulation methodologies
- Parallel code algorithms
- Results

Myrinet

- 1.28 Gb/s full duplex
- Programmable NIC
- Source-based routing
- Switch
 - 16-port crossbar
 - cheap and simple
 - connect processors or other switches
 - dollar tradeoff
- Wormhole routing
- Cable length restriction
 - SAN 10'
 - LAN 35' (or 60' at half-speed)
 - Fiber $\$1800 \times 2$ for 550m multi at 1.0 Gb/s
 - Kills “interesting” topologies
- $256 \times \$1700$ NIC + $32 \times \$5000$ switch
+ $12 \times 32 \times \$200$ cable = \$670k

Gigabit ethernet

- Popularity
- 802.3x standard pause
- Conventional routers
 - Non-scalable
 - Crossbar or bus design
 - 64 port max
 - Spanning tree \Rightarrow no mesh
 - Store and forward for compatibility
- $256 \times \$700$ NIC + $5 \times \$30\,000$ switch
+ $280 \times \$75$ cable = \$350k

Avici terabit switch router

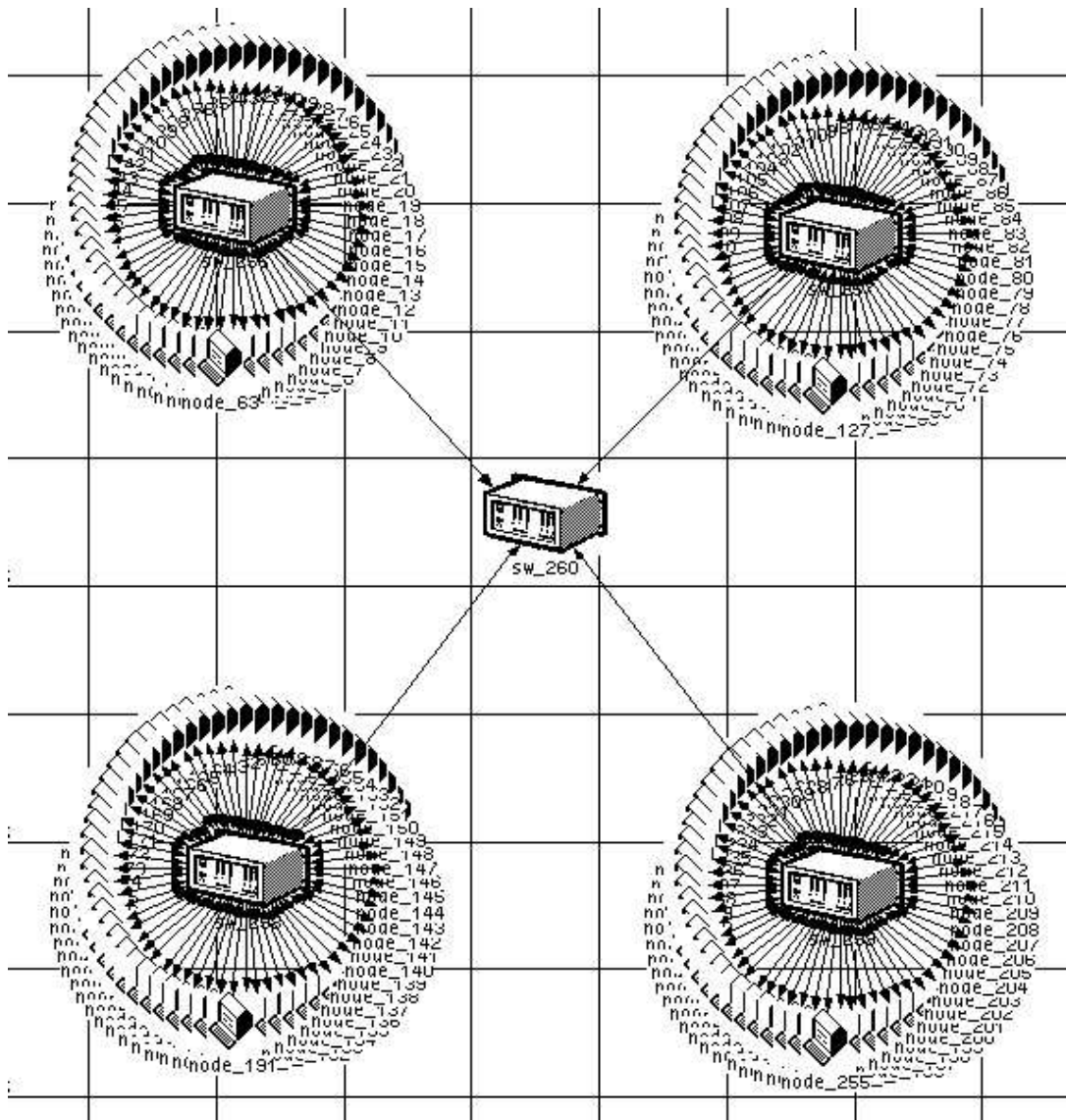
- Fancy switch
- Designed for telcos
- Dual 3-D torus mesh
- Internal 20 Gb/s links
- Maximum size $14 \times 16 \times 5 = 1120$ line cards
- 17 920 compute nodes
- Wormhole routing
- Line cards store and forward
- $256 \times \$700$ NIC + $\$250\,000$ switch
+ $256 \times \$75$ cable = $\$450k$

Simulation methodologies

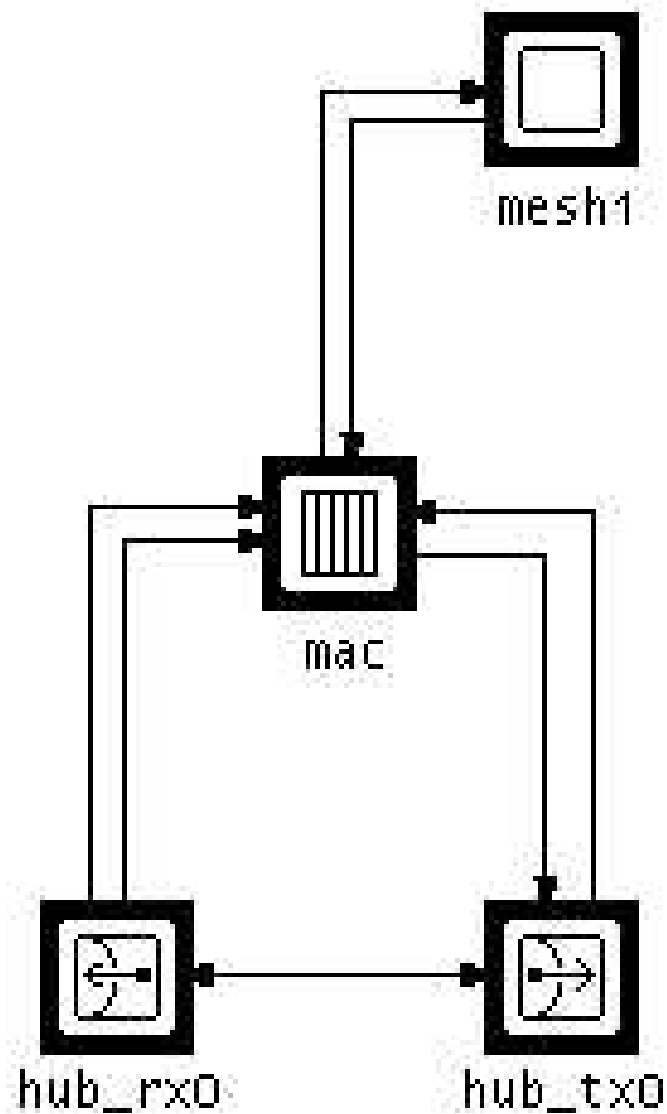
- Identical interface layer
- Three different simulation engines
- 256-node cluster
 - represents average user job
- No host effects
- No software effects

MIL3 Opnet

- Big simulation package
- Many predefined ethernet features
- Cascaded switches, star topology

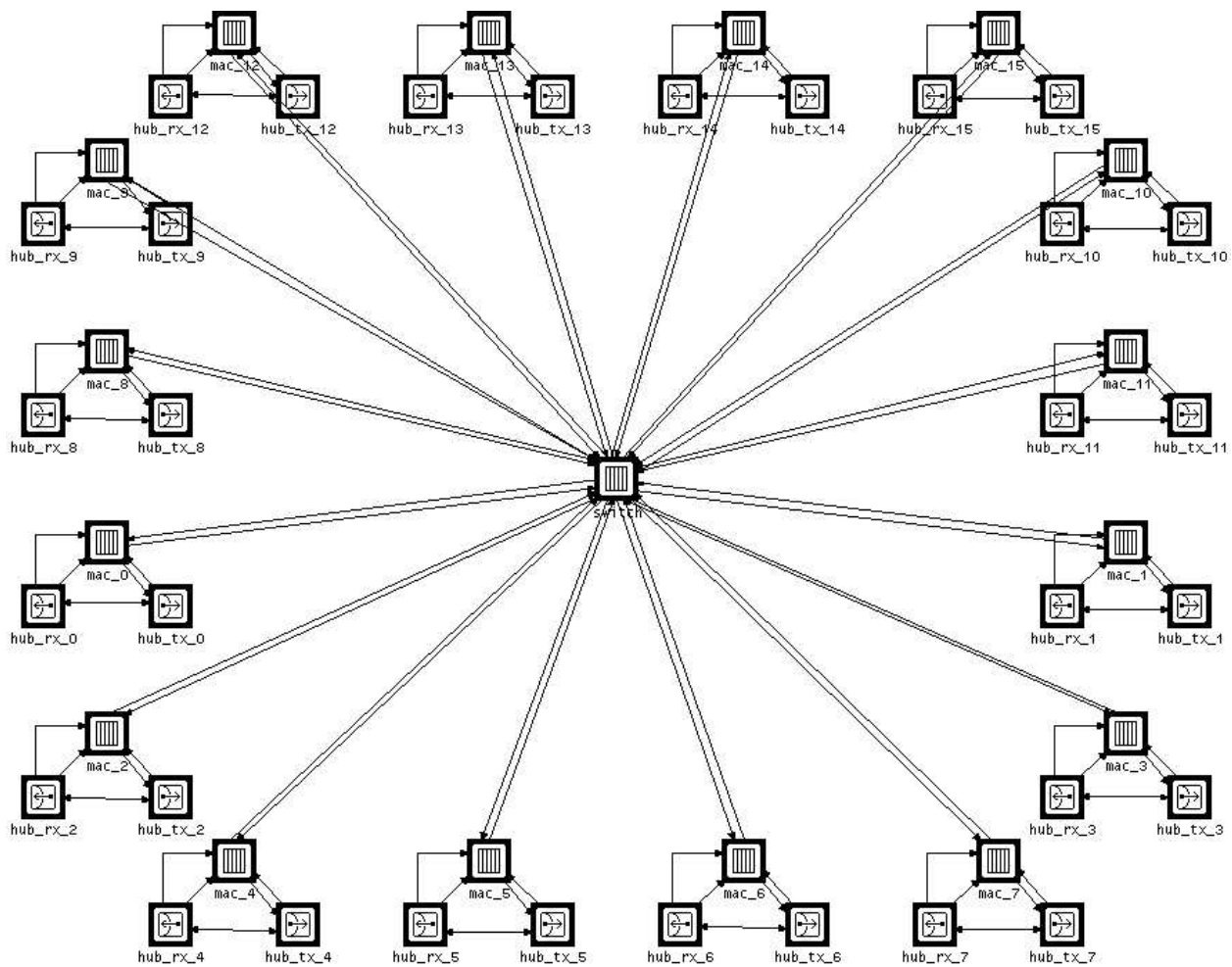


End system node



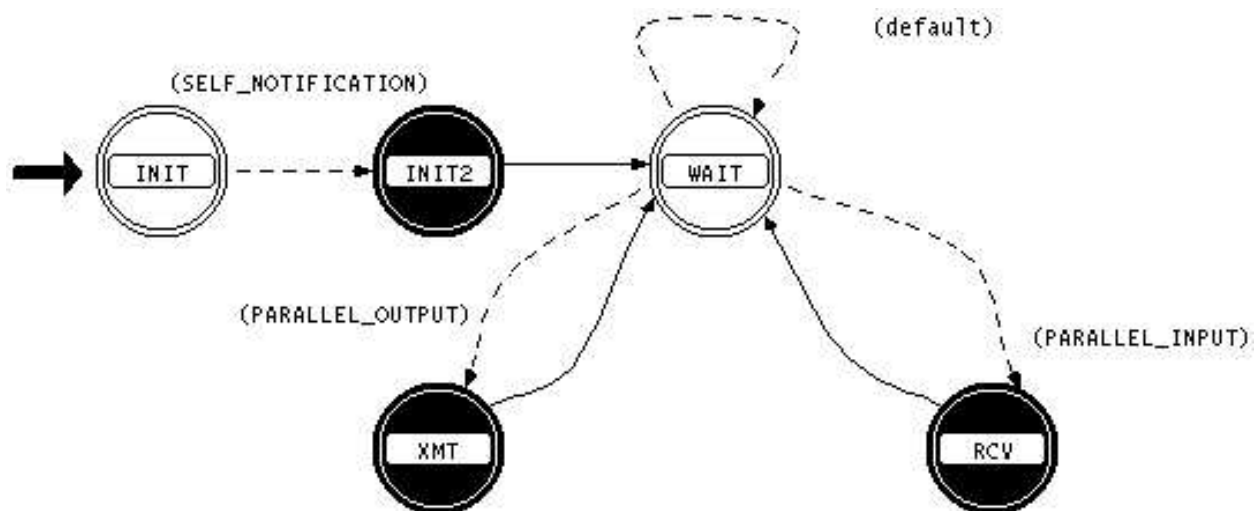
Switch node

- 16-port example

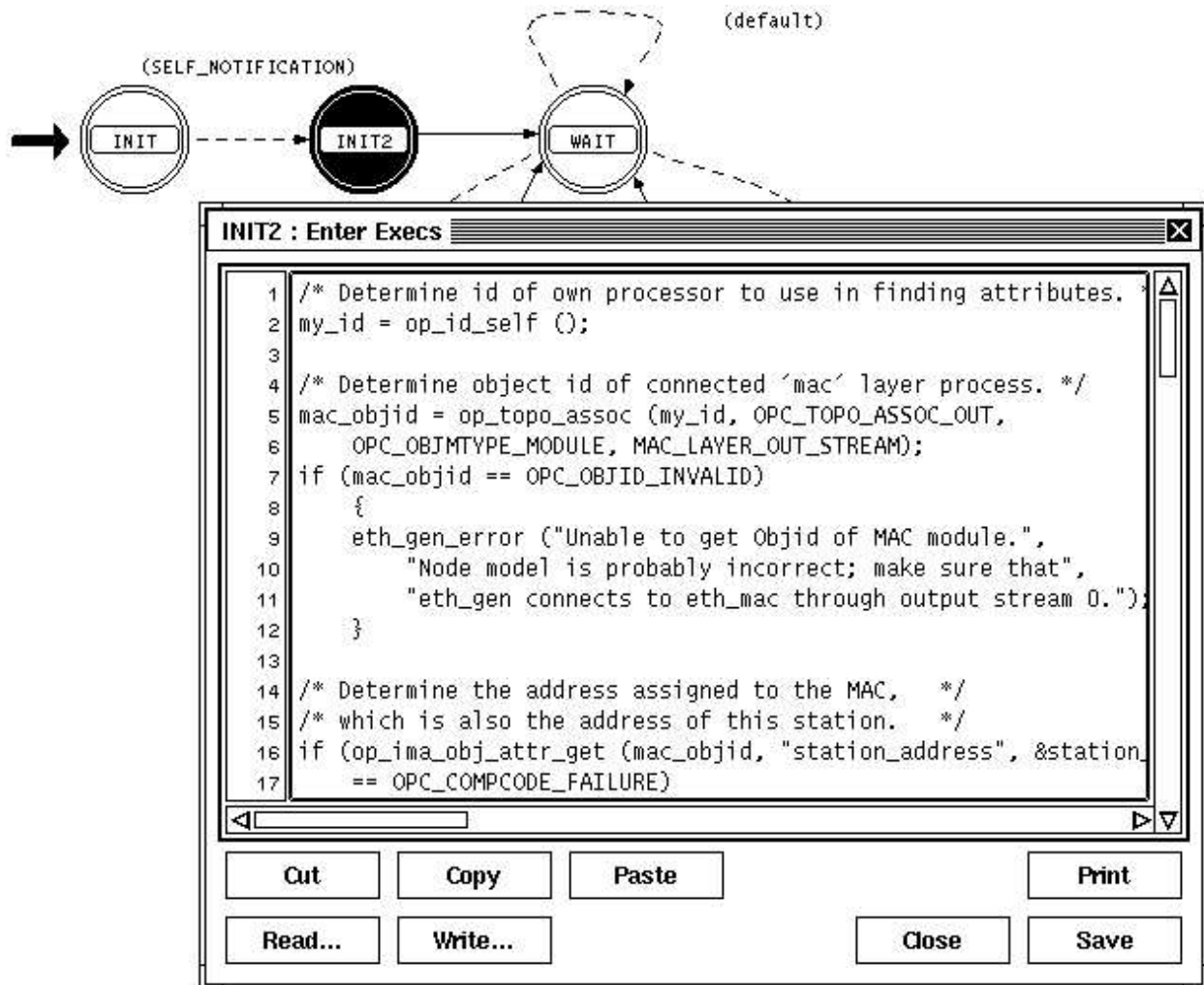


Process

- Opnet provides
 - state machine
 - packet manipulation
 - event scheduling
- We write
 - parallel code details
 - statistics



Code

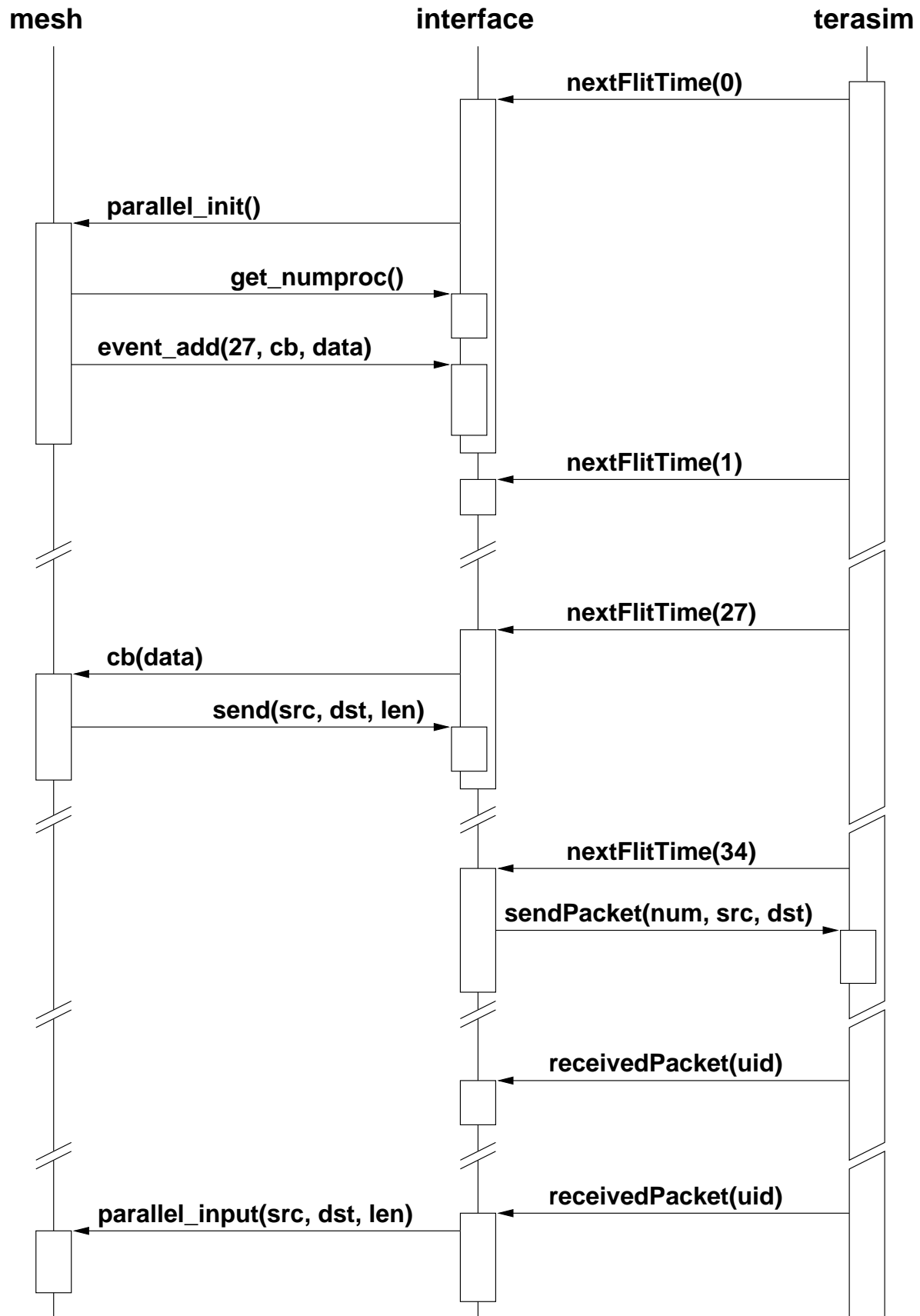


Avici simulator

- “Terasim” hardware planner
- Generation 1 vs. 2
- Terasim provides
 - data motion
 - timing
 - queue management
- We write
 - line card interface
 - ethernet delays
 - state machine
 - packet manipulation
 - event scheduling
 - parallel code details
 - statistics

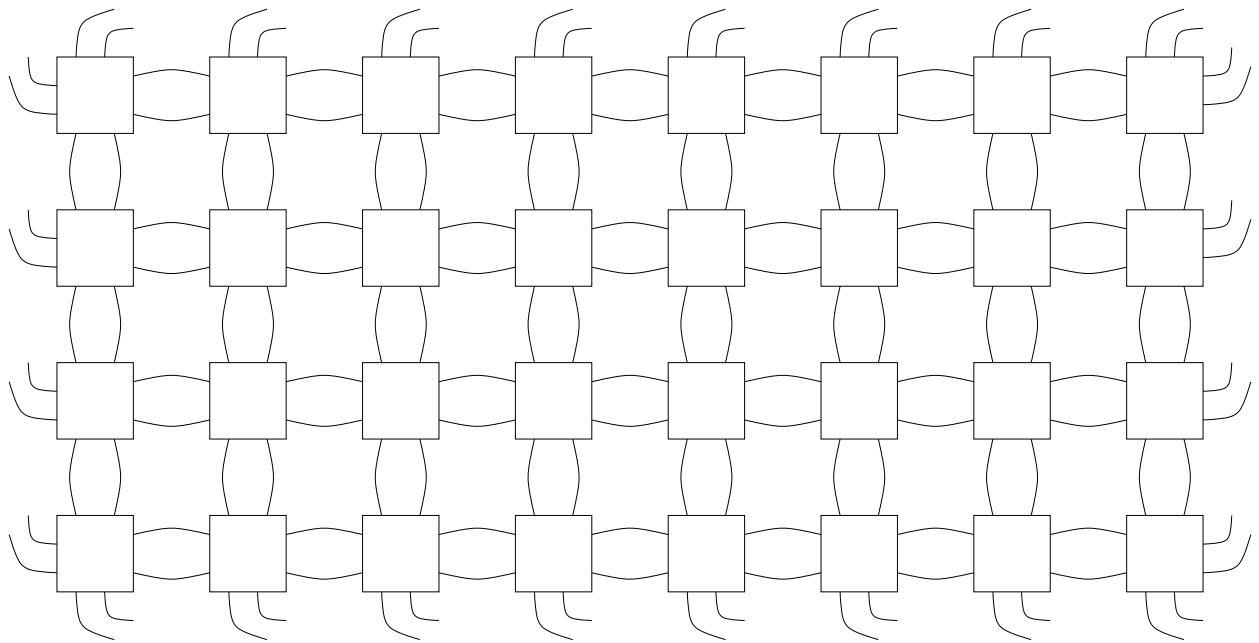
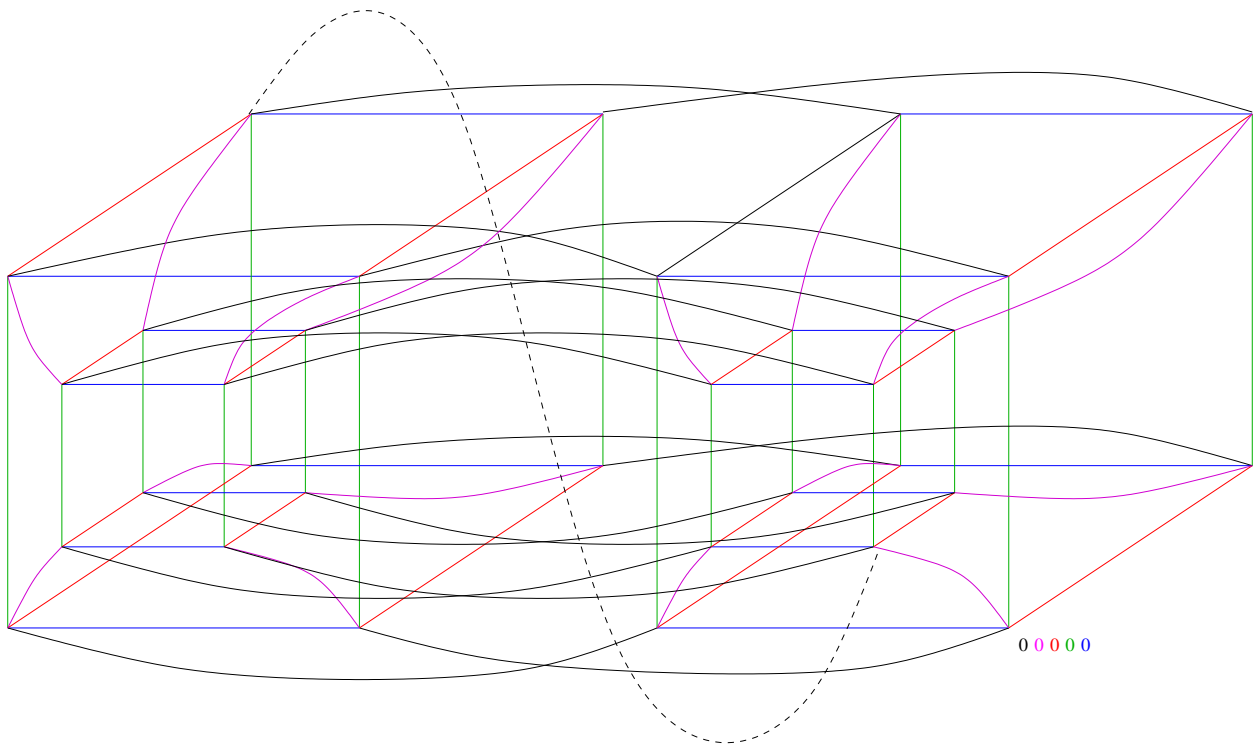
Interactions

- interface → terasim
 - sendPacket
 - finish
- terasim → interface
 - receivedPacket
 - nextFlitTime
 - runEnd
- mesh → interface
 - get_numproc
 - get_current_time
 - send
 - event_add
 - end_simulation
- interface → mesh
 - parallel_init
 - parallel_input
 - parallel_end



Myrinet simulator

- Utah Paint simulator module
- Utah provides
 - data motion
 - timing
- We write
 - (similar to terasim)
- Parameters
 - numOfProcessor, numOfSwitch
 - propDelay, fallThruDelay
 - switch buffer limits
- Topology file
 - S2: P16 P17 ... S1.9 S3.8 ...
- Routes
 - p0 p14 b6
 - p0 p40 9990
- Language for connectivity description



Algorithms

- Token pass
 - large token \rightarrow bandwidth measurement
 - small token \rightarrow latency measurement
 - no contention
- Fan in/out
 - gather/scatter aspects
 - data output, restart files, initialization
- Mesh

finite element calculation

$$\frac{\partial \mathbf{f}}{\partial t} = F \left(t, \mathbf{f}, \frac{\partial \mathbf{f}}{\partial x}, \frac{\partial \mathbf{f}}{\partial y} \right)$$

$$\mathbf{f}^{n+1} = \mathbf{f}^n + \Delta t F \left(t^n, \mathbf{f}^n, \frac{\partial \mathbf{f}^n}{\partial x}, \frac{\partial \mathbf{f}^n}{\partial y} \right)$$

2D or 3D

mesh or torus space

Characterization results

	Min	Avg	Max	Stdev
Myrinet	0.388	0.427	0.869	0.093
Avici fabric	0.180	0.186	0.330	0.024
Avici GigE	1.380	1.386	1.530	0.024
Conventional GigE	1.564	1.595	3.532	0.244

- No contention, jitter is distance
- No conventional switch latency, 10–30 μ s
- Fabric has no store-and-forward delay

	Theoretical	Simulation
Myrinet	1280.000	1279.947
Avici GigE	980.000	975.610
Conventional GigE	980.000	974.301

Fan results

- Fan-in

2048 byte timing (all data delivered)

	Duration (ms)
Myrinet	3.314
Avici GigE	4.350
Conventional GigE	4.345
Avici fabric	0.233

Line rate effect

- Fan-out

	Duration (ms)
Myrinet	3.319
Avici GigE	0.259
Conventional GigE	0.055
Avici fabric	0.233

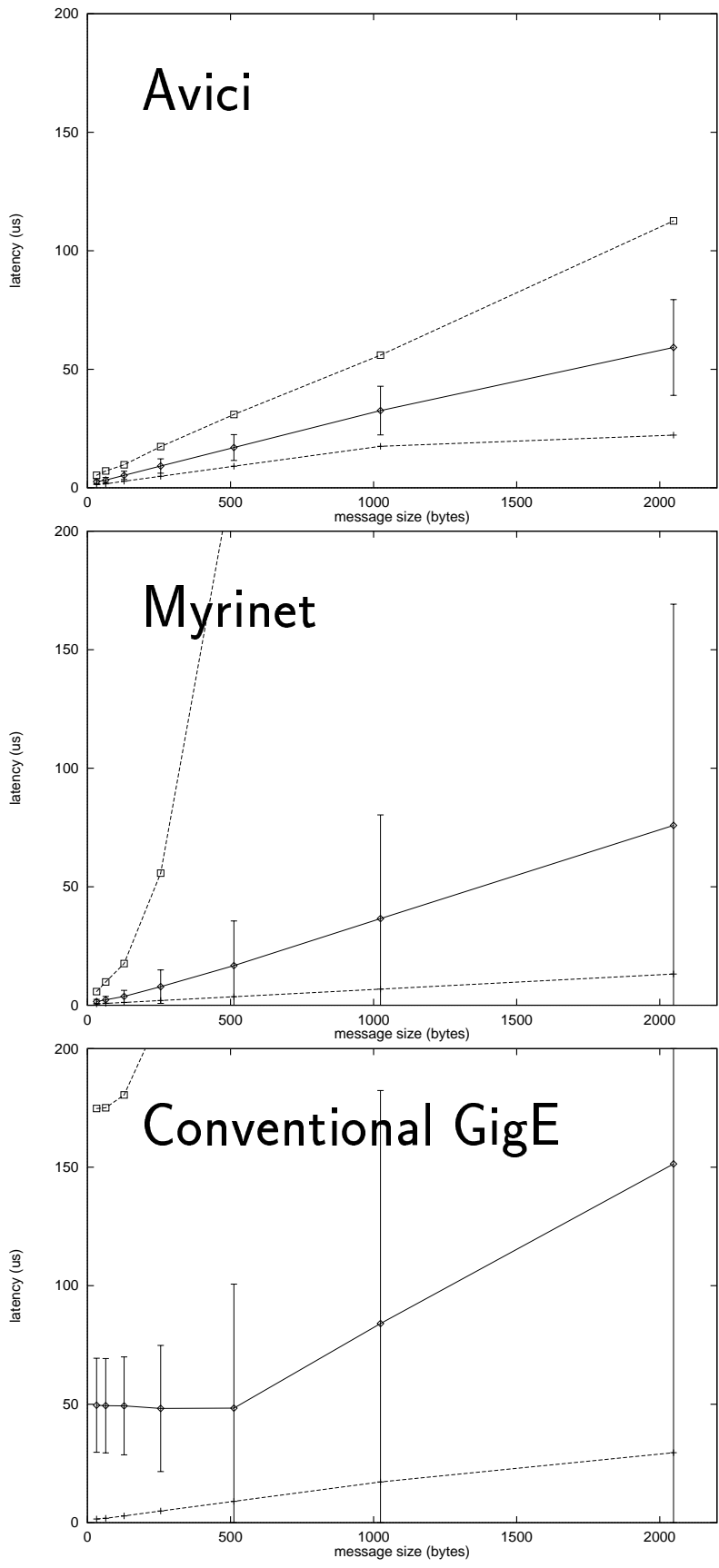
Conventional GigE multicast hardware

Avici better aggregate bandwidth

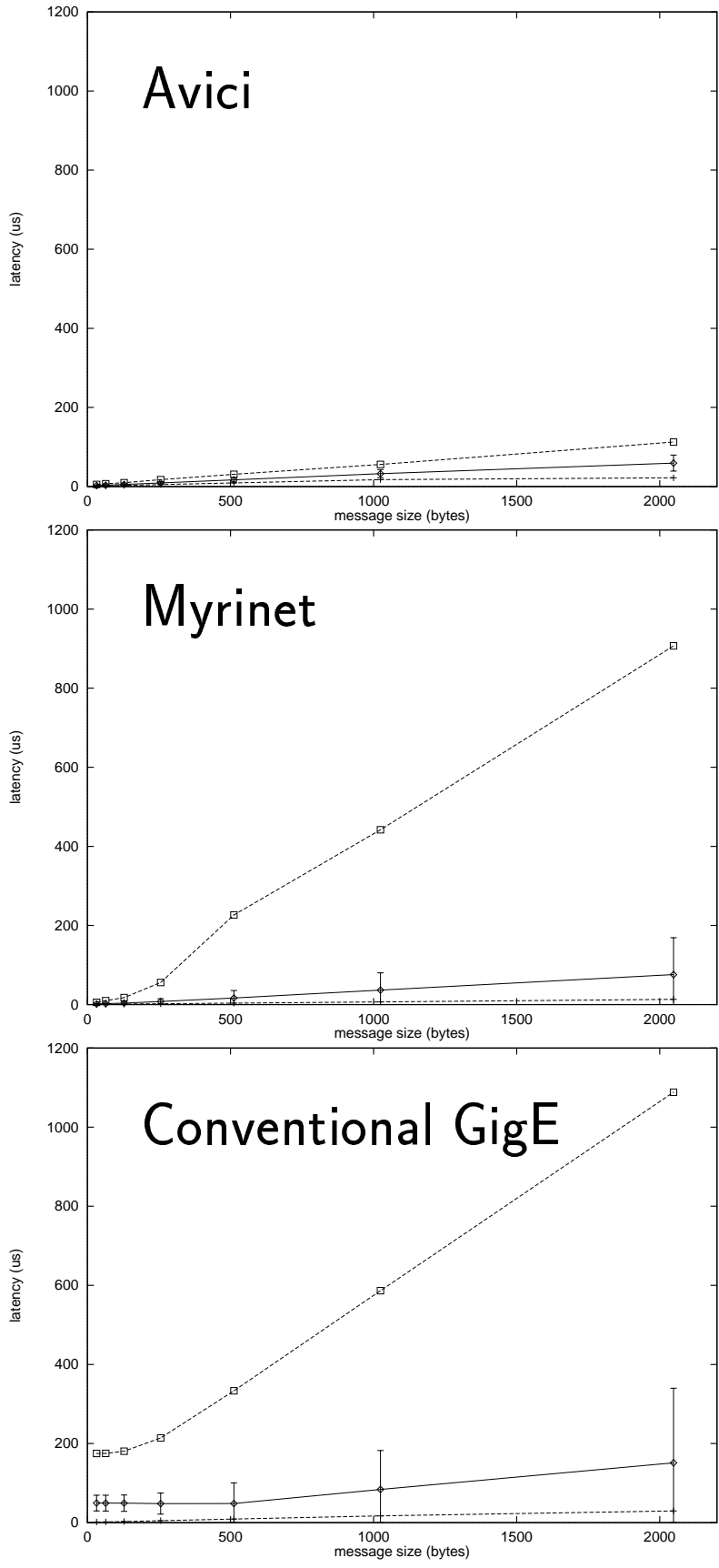
Mesh latency results

- Avici and Myrinet similar average
 - Avici $2\mu\text{s}$ core latency
 - Myrinet 300 ns per hop
 - Crossing at 256 byte messages
- Conventional gigE central switch contention
 - trunked links, fat tree
- Order of magnitude delay possible
 - Myrinet and conventional gigE switch contention
 - Avici output port contention only
- Virtual torus similar results

Mesh close-up



Mesh wide-angle



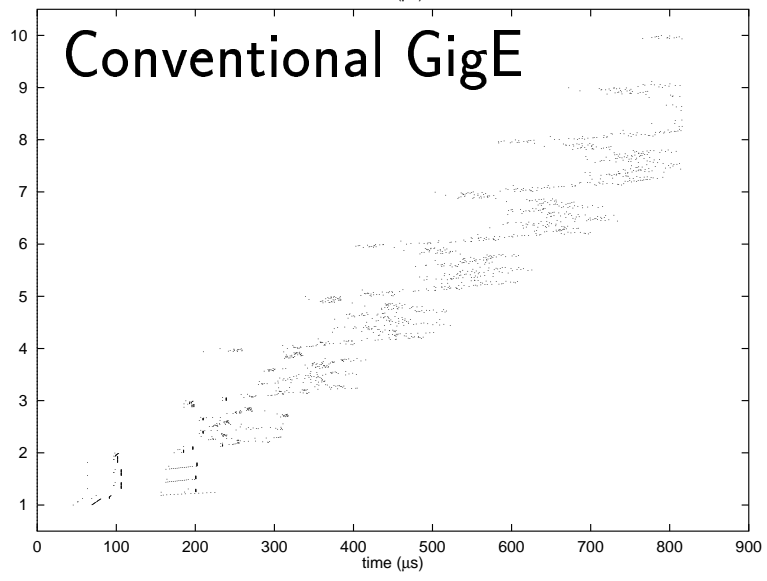
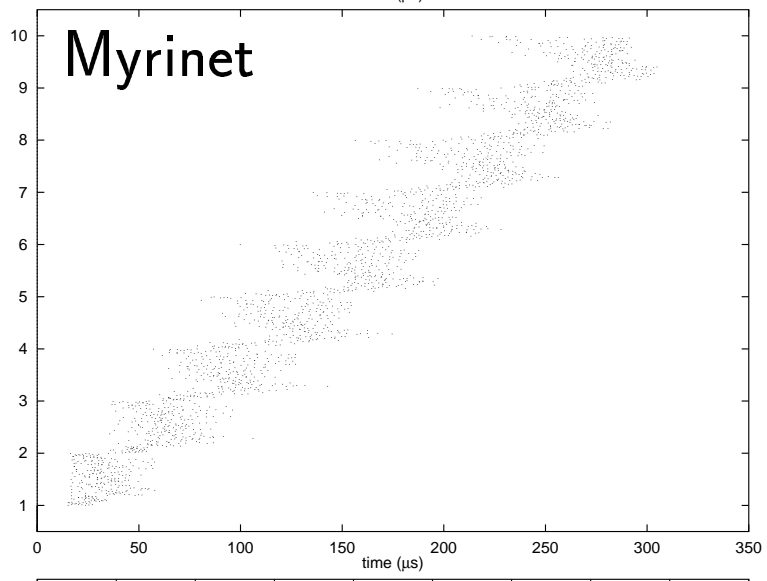
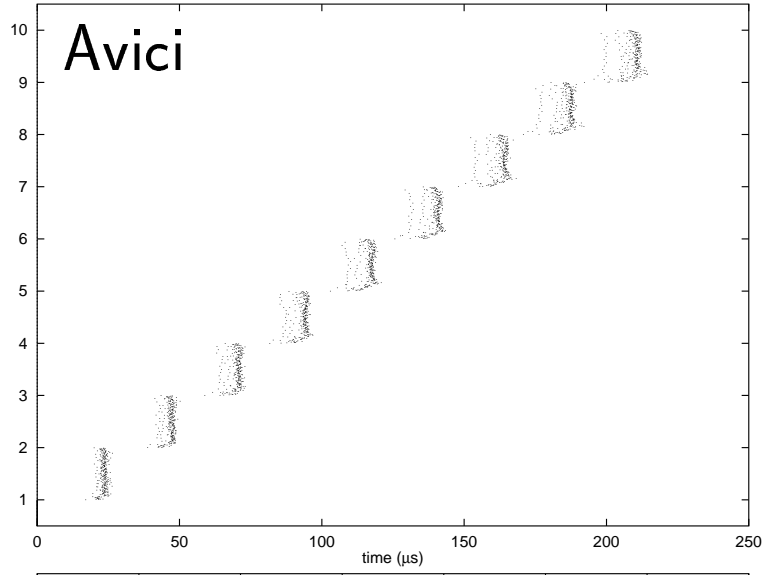
Mesh completion times

- Time each processor completed each iteration
- 256 byte message size shown (32–2048 tested)
- Some processors always finish earlier (edges)
- Some processors many iterations ahead of others
- Patterns are internode distance effect

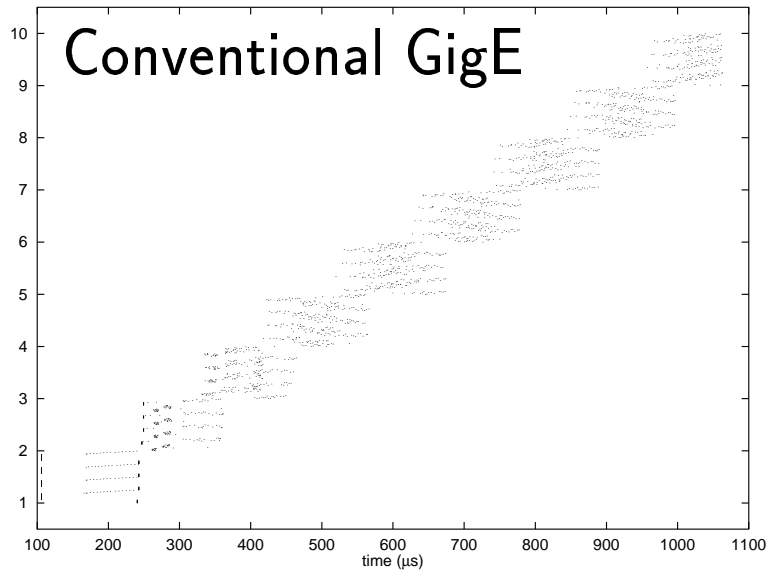
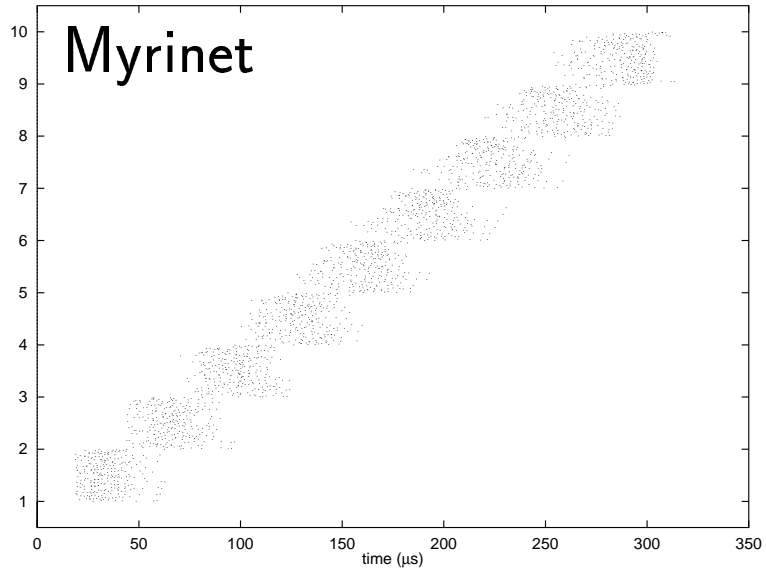
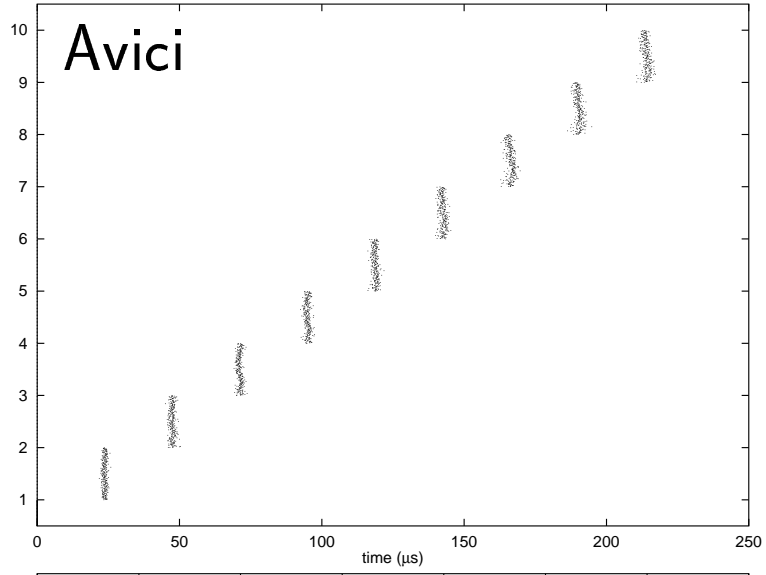
- Avici
 - 140 μ S to 900 μ S
- Myrinet
 - similar at small, 3 \times longer at large
- Conventional gigE
 - stripes 3 ms long at large

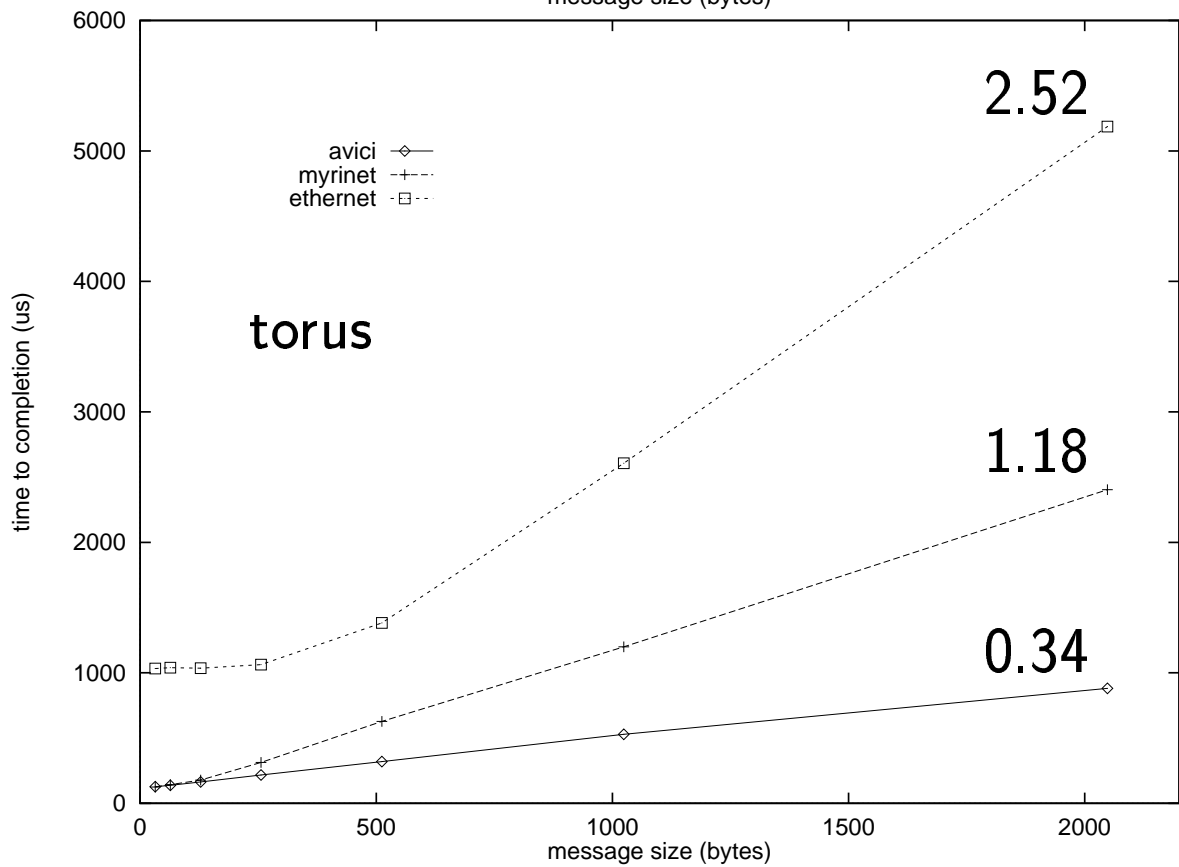
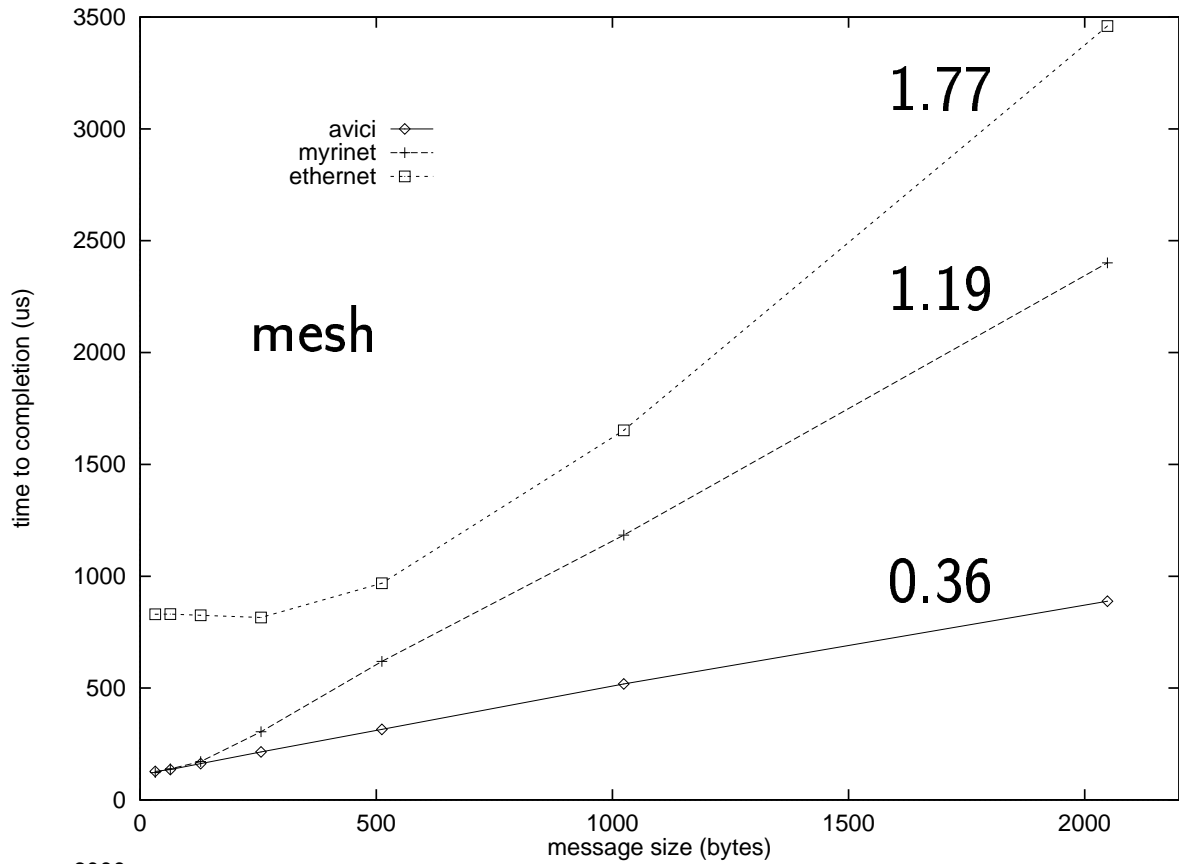
- Performance degradation with increasing message size
- Dictates how large parallel app may be

mesh



torus





Conclusion

- Message latency affected by
 - available bandwidth
 - presence of bottlenecks
- Fabric blocking bad to apps
- Conventional gigE
 - trunking option
 - scalability limit
- Myrinet
 - potential blocking at every switch
 - cable length problem
 - non-commodity
- Avici
 - only one switch, highly connected
 - historical ethernet, good and bad