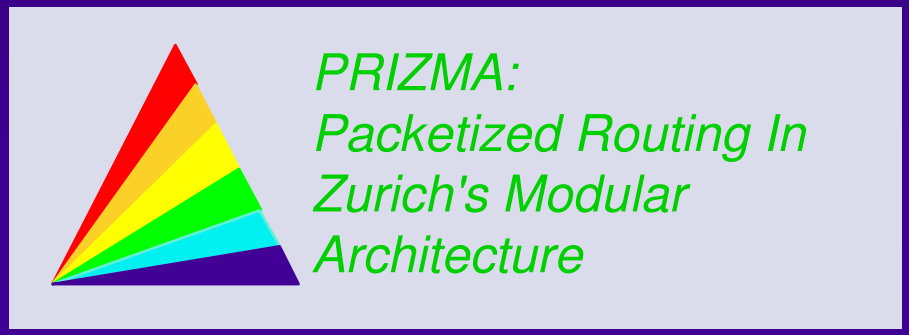
A new simulation approach for HPC interconnects

Ronald Luijten (lui@zurich.ibm.com)
IBM Rueschlikon
11 March 2009

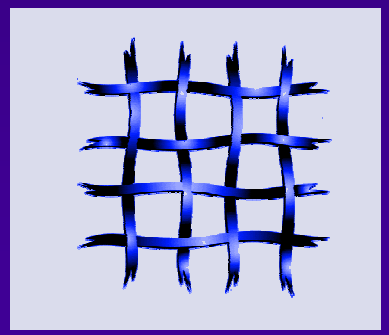

Andreas Doering*
Krzysztof Kryszczuk*
Phillip Stanley Marbell
Gergely Paljak
Alessandra Scicchitano
German Rodriguez Herrera
HansPeter Ineichen (borrowed from SCS)
Rolf Clauberg
Wolfgang Denzel
Cyriel Minkenberg
Mitch Gusat
Francois Abel *
Patricia Sagmeister*
Maria Gabrani*
Lydia (Yiyu) Chen
Florian Auernhammer

Alternative title:

A Journey from Telco Switching to Computer Interconnects…



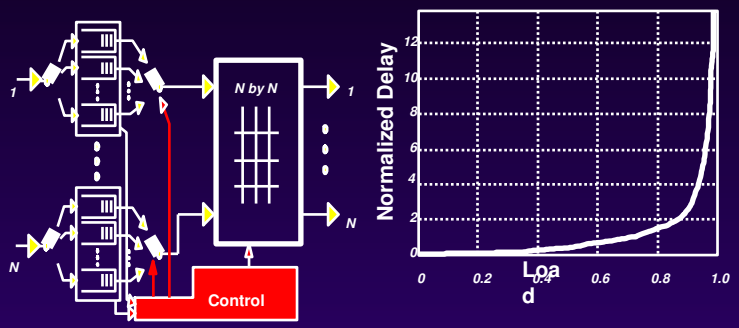
Ended 2003



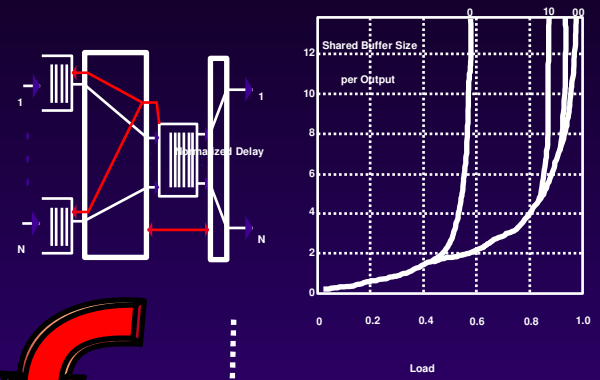
2009

PRIZMA's Robust Performance

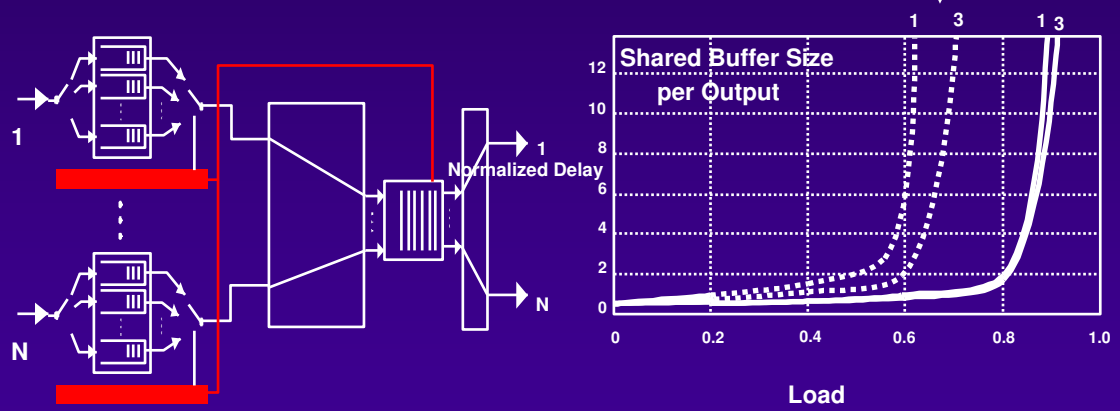
Virtual Output Queuing



Shared output Queueing



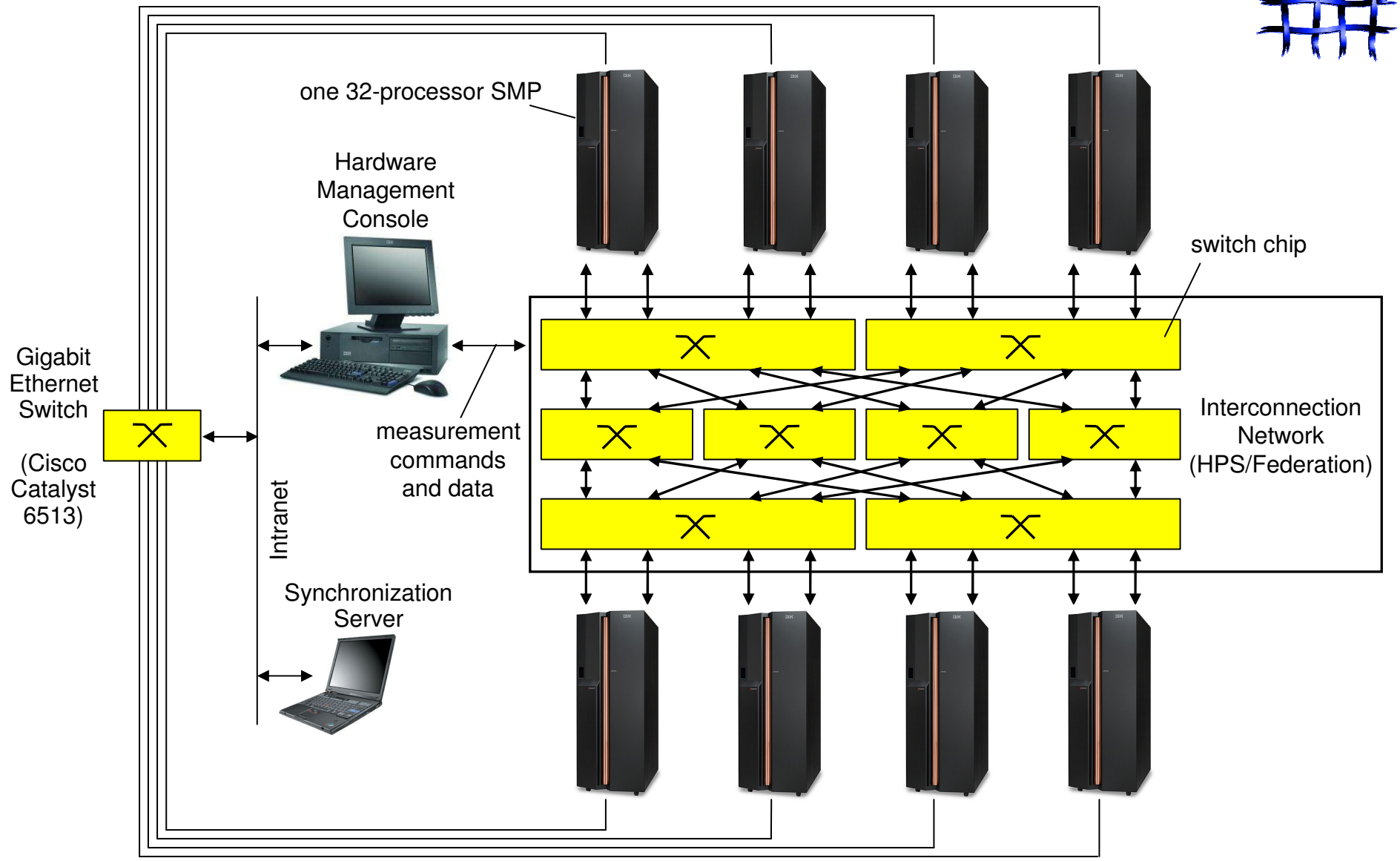
Virtual Output Queuing with Output Buffering



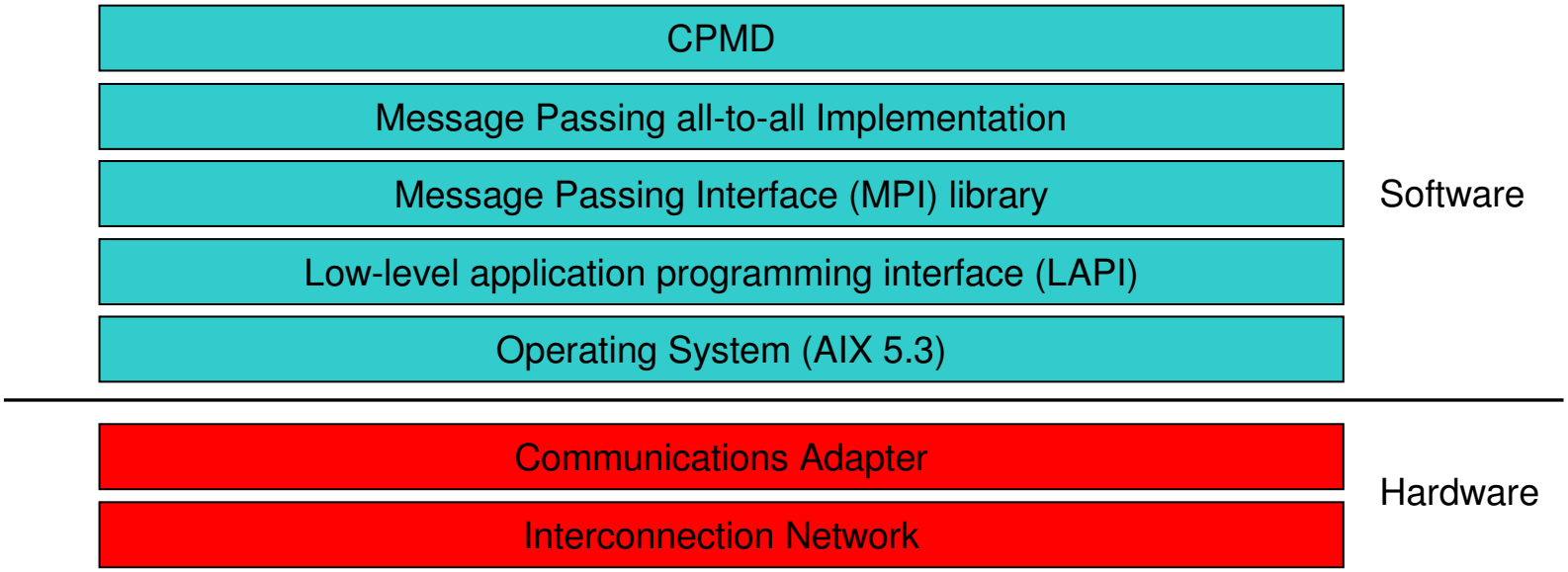
IP

- Robust switch performance for extreme traffic conditions
- memory used for contention resolution only - not for performance
- more scalable solution compared to VOQ - crossbar approach
- Documented: IEEE Comm.Mag. Dec. 2000 - Best Comm. Mag. Paper 2000

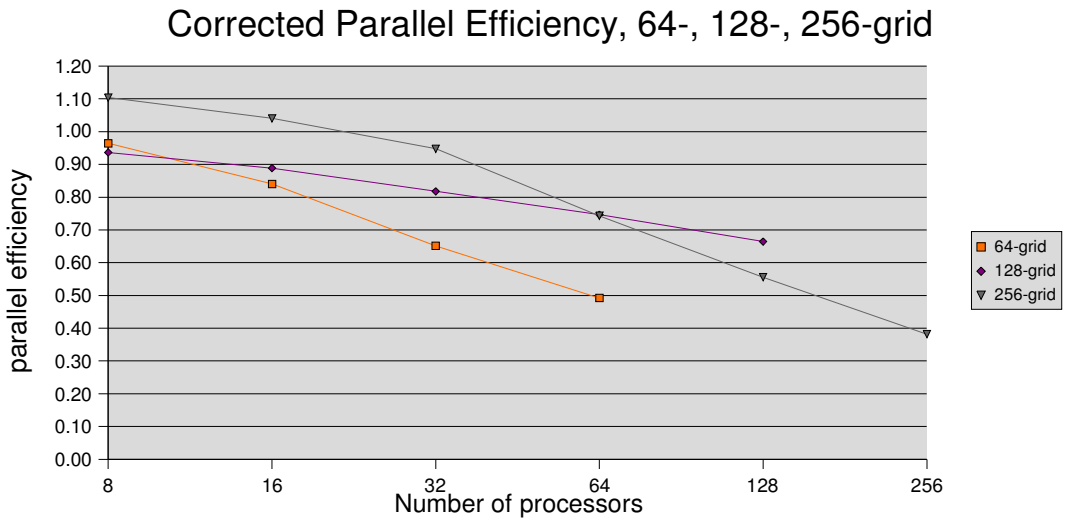
Measurement on ZRL Regatta w/ Federation



Regatta CPMD Communication stack

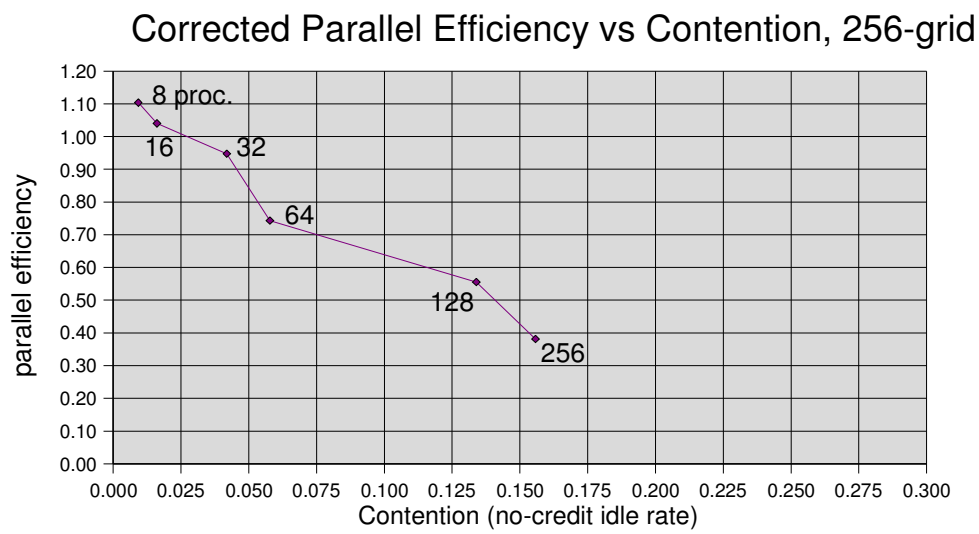


Measurement Results (Parallel Efficiency)



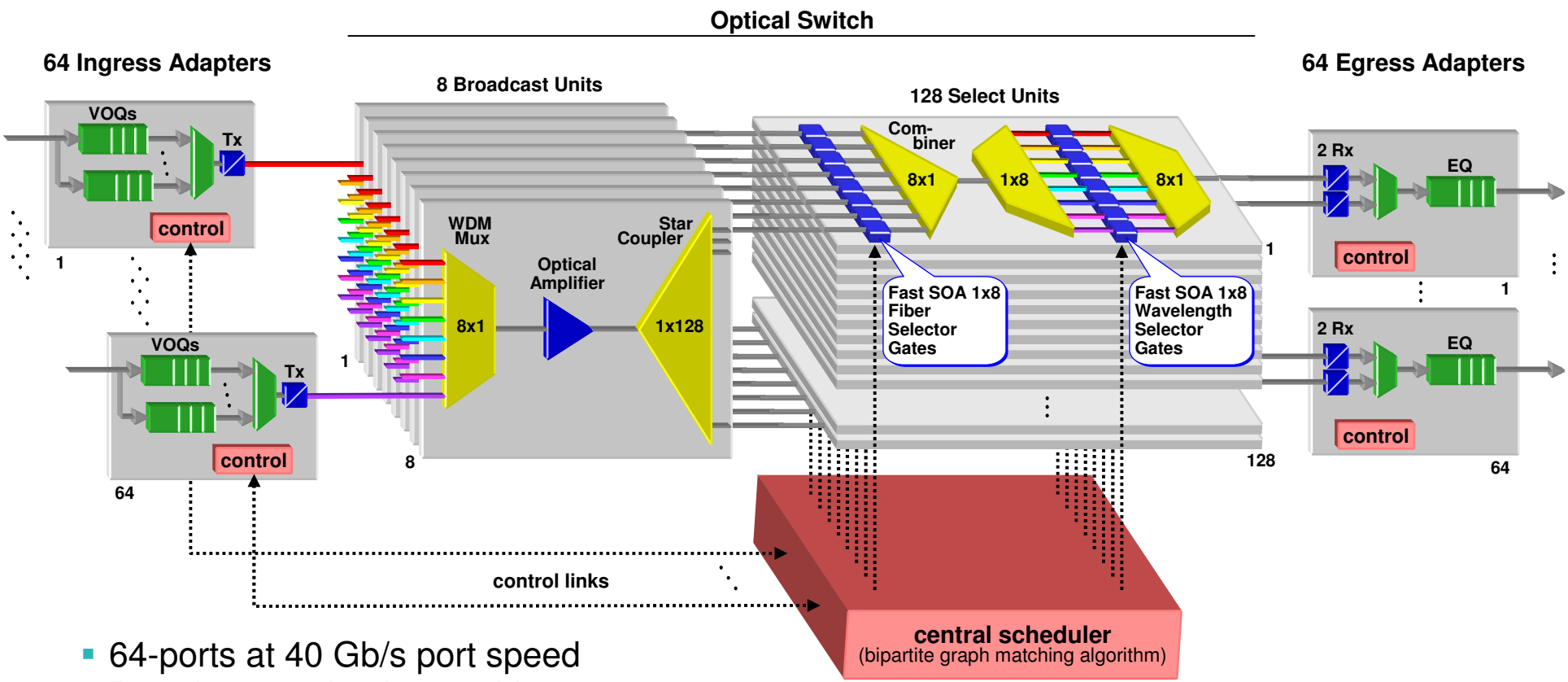
- Measurement config:
  - IBM MPI
  - LAPI
  - IBM SMA3 adapter
  - user-space mode
  - HPS
- Ideally: parallel efficiency is constant (1.00)

- Strong evidence that **contention in the interconnect causes decrease in parallel efficiency**
- Interconnect only ~20% used
- **Suspect effect of interconnect contention on inefficiency in the software layers**



OSMOSIS Demonstrator System - *Overview*

(Optical Shared MemOry Supercomputer Interconnect System)



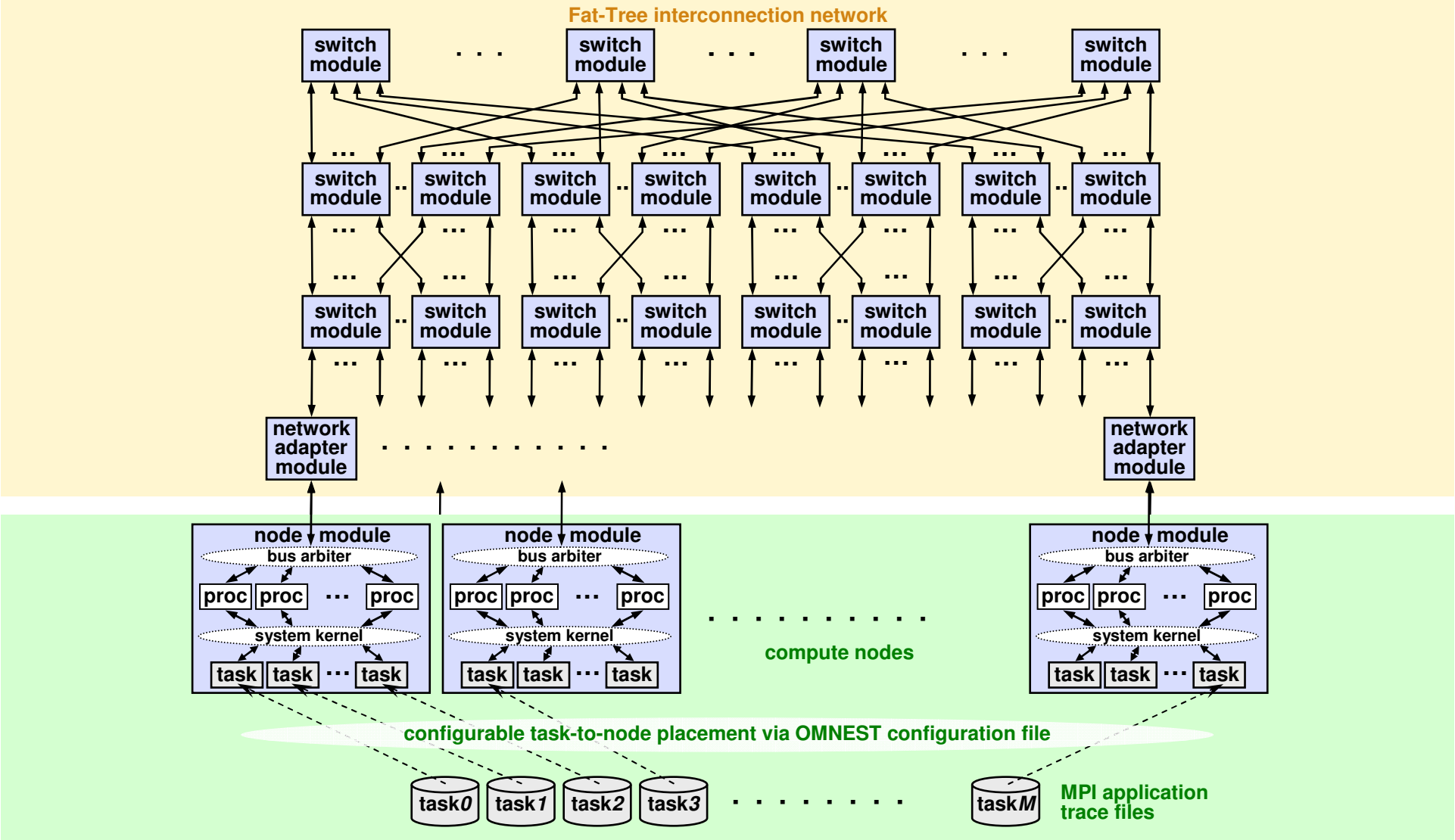
- 64-ports at 40 Gb/s port speed
- Broadcast-and-select architecture
- Combination of wavelength and space division multiplexing
- fast switching based on SOAs

CORNING

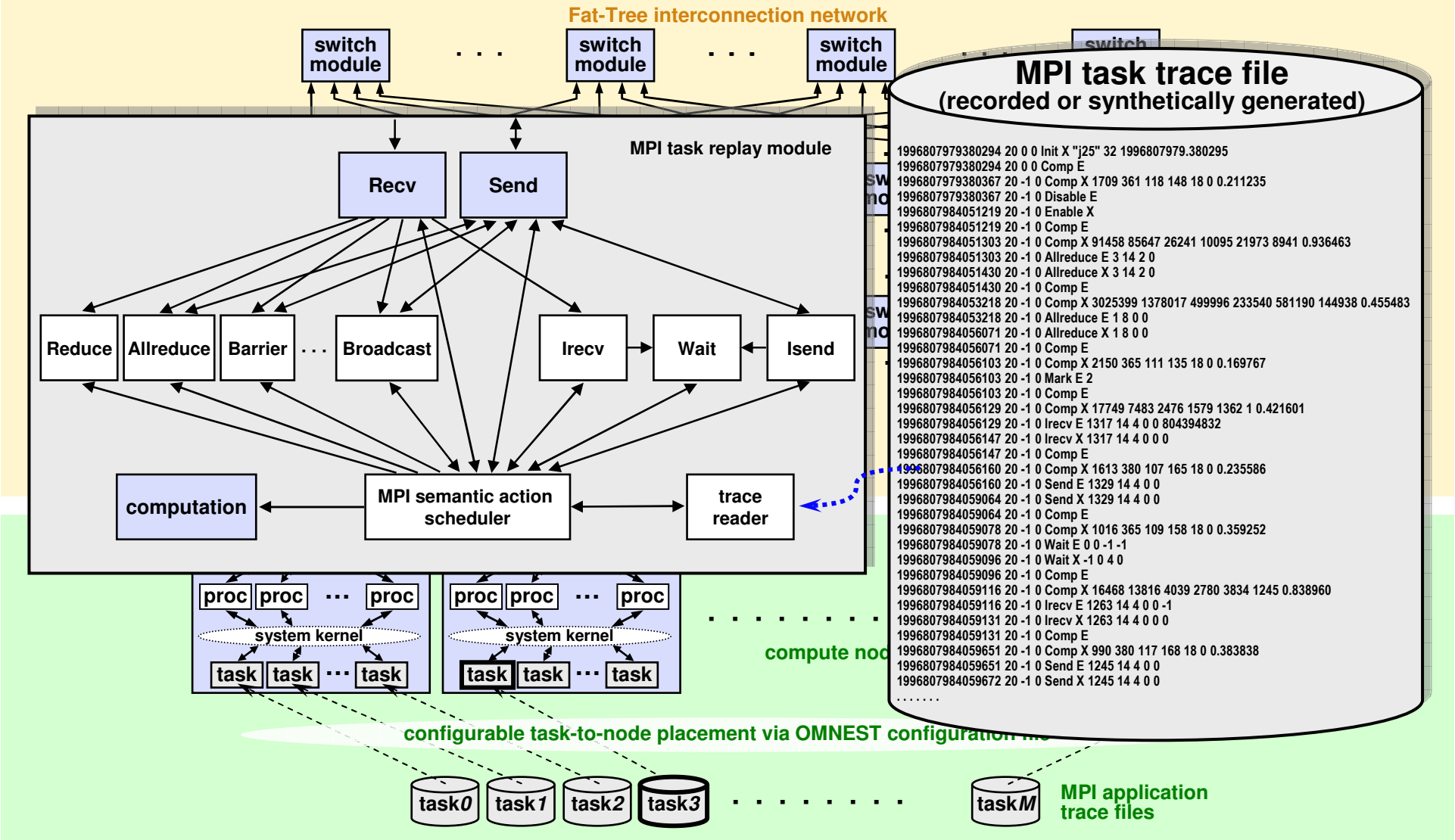
Insights gained 3 years after our transition

PERCS

PERCS phase-2 full-system model



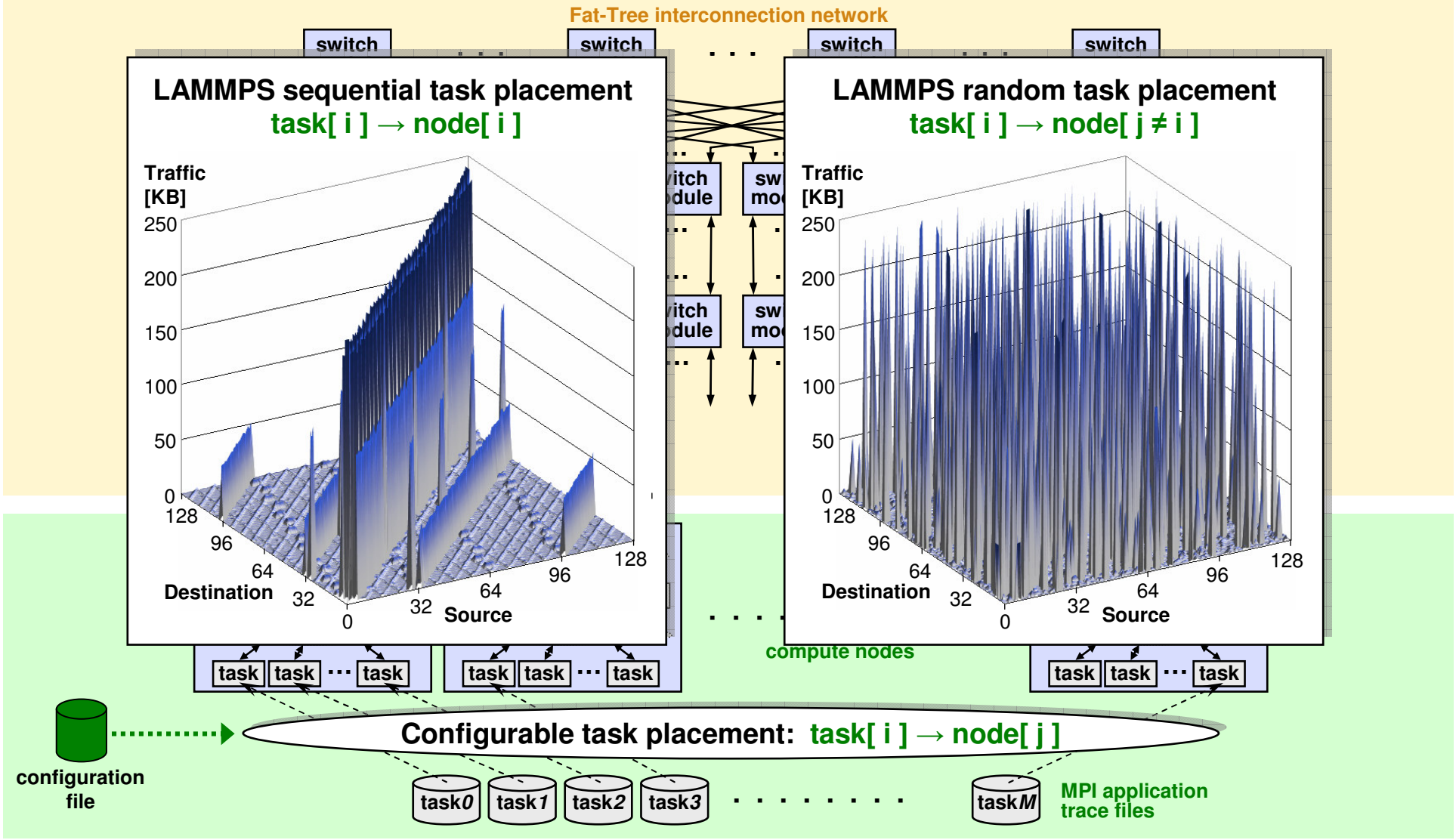
MARS trace replay



MPI task trace file
(recorded or synthetically generated)

```
1996807979380294 20 0 0 Init X "j25" 32 1996807979.380295
1996807979380294 20 0 0 Comp E
1996807979380367 20 -1 0 Comp X 1709 361 118 148 18 0 0.211235
1996807979380367 20 -1 0 Disable E
1996807984051219 20 -1 0 Enable X
1996807984051219 20 -1 0 Comp E
1996807984051303 20 -1 0 Comp X 91458 85647 26241 10095 21973 8941 0.936463
1996807984051303 20 -1 0 Allreduce E 3 14 2 0
1996807984051430 20 -1 0 Allreduce X 3 14 2 0
1996807984051430 20 -1 0 Comp E
1996807984053218 20 -1 0 Comp X 3025399 1378017 499996 233540 581190 144938 0.455483
1996807984053218 20 -1 0 Allreduce E 1 8 0 0
1996807984056071 20 -1 0 Allreduce X 1 8 0 0
1996807984056071 20 -1 0 Comp E
1996807984056103 20 -1 0 Comp X 2150 365 111 135 18 0 0.169767
1996807984056103 20 -1 0 Mark E 2
1996807984056103 20 -1 0 Comp E
1996807984056129 20 -1 0 Comp X 17749 7483 2476 1579 1362 1 0.421601
1996807984056129 20 -1 0 Irecv E 1317 14 4 0 0 804394832
1996807984056147 20 -1 0 Irecv X 1317 14 4 0 0 0
1996807984056147 20 -1 0 Comp E
1996807984056160 20 -1 0 Comp X 1613 380 107 165 18 0 0.235586
1996807984056160 20 -1 0 Send E 1329 14 4 0 0
1996807984059064 20 -1 0 Send X 1329 14 4 0 0
1996807984059064 20 -1 0 Comp E
1996807984059078 20 -1 0 Comp X 1016 365 109 158 18 0 0.359252
1996807984059078 20 -1 0 Wait E 0 0 -1 -1
1996807984059096 20 -1 0 Wait X -1 0 4 0
1996807984059096 20 -1 0 Comp E
1996807984059116 20 -1 0 Comp X 16468 13816 4039 2780 3834 1245 0.838960
1996807984059116 20 -1 0 Irecv E 1263 14 4 0 0 -1
1996807984059131 20 -1 0 Irecv X 1263 14 4 0 0 0
1996807984059131 20 -1 0 Comp E
1996807984059651 20 -1 0 Comp X 990 380 117 168 18 0 0.383838
1996807984059651 20 -1 0 Send E 1245 14 4 0 0
1996807984059672 20 -1 0 Send X 1245 14 4 0 0
.......
```

MARS task placement

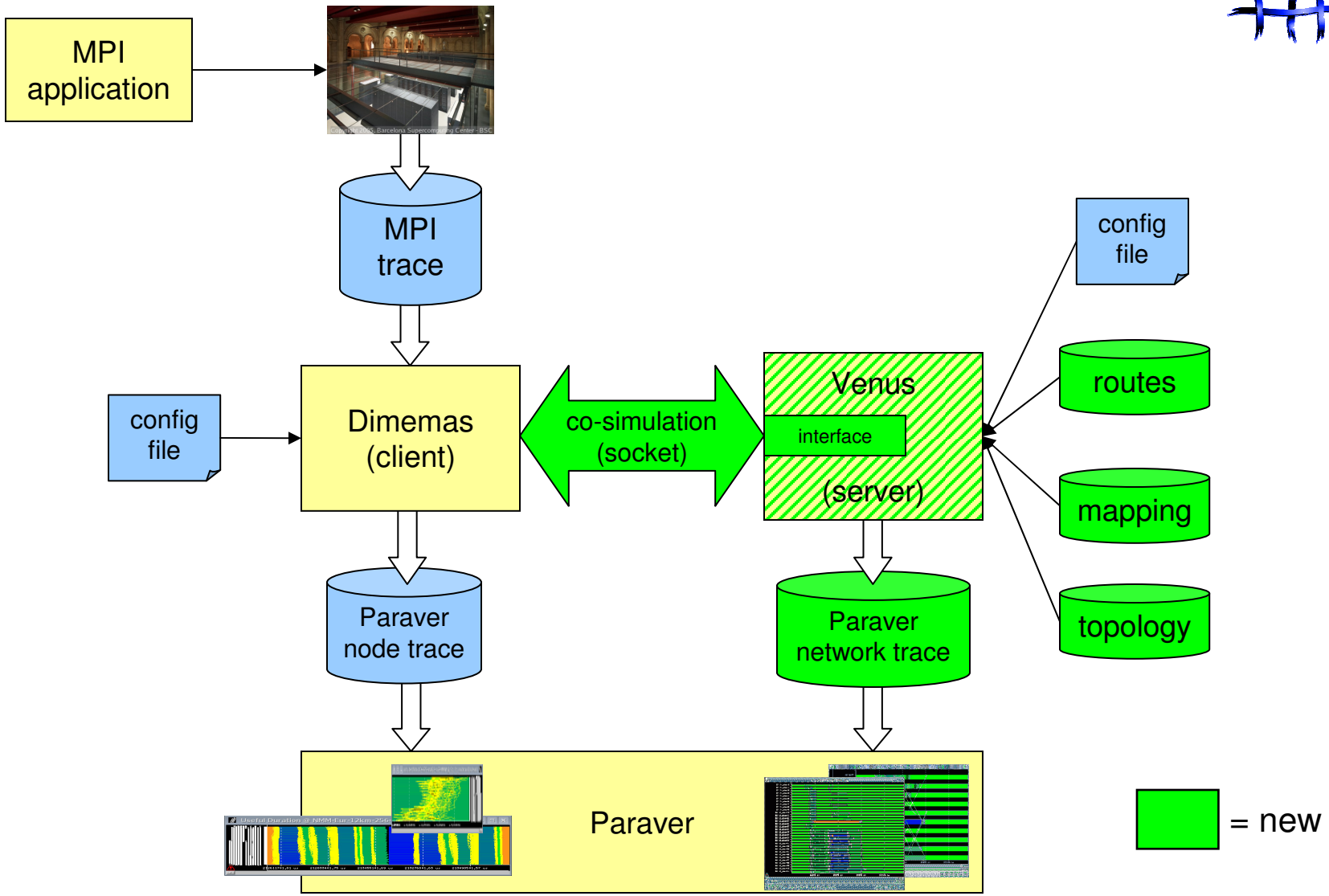


Good new direction… However

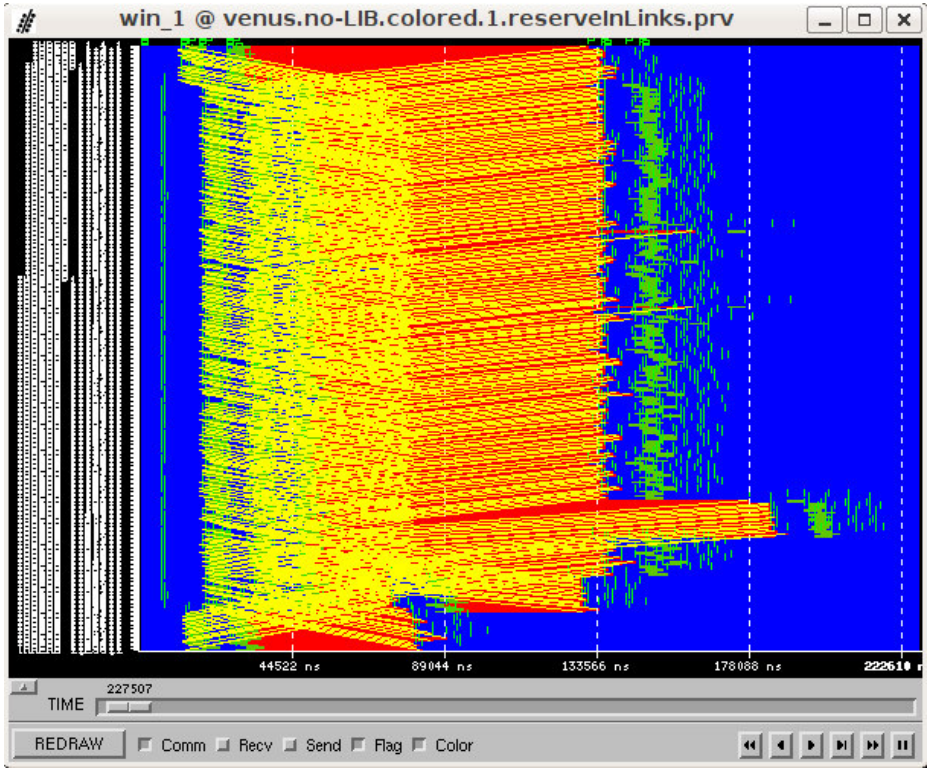
Where do I get good traces?

Mare Incognito

Tool Enhancement & Integration

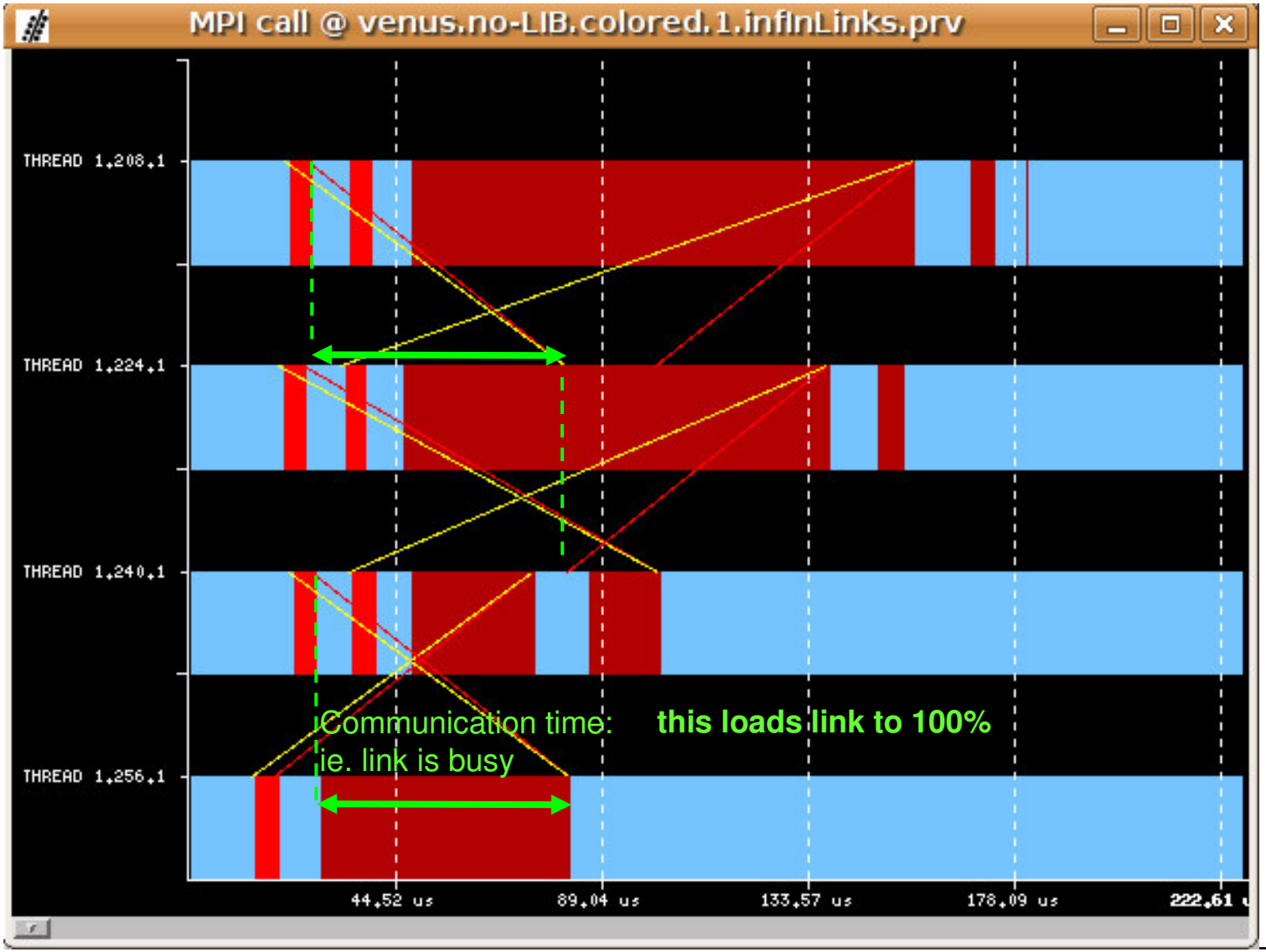


WRF application example

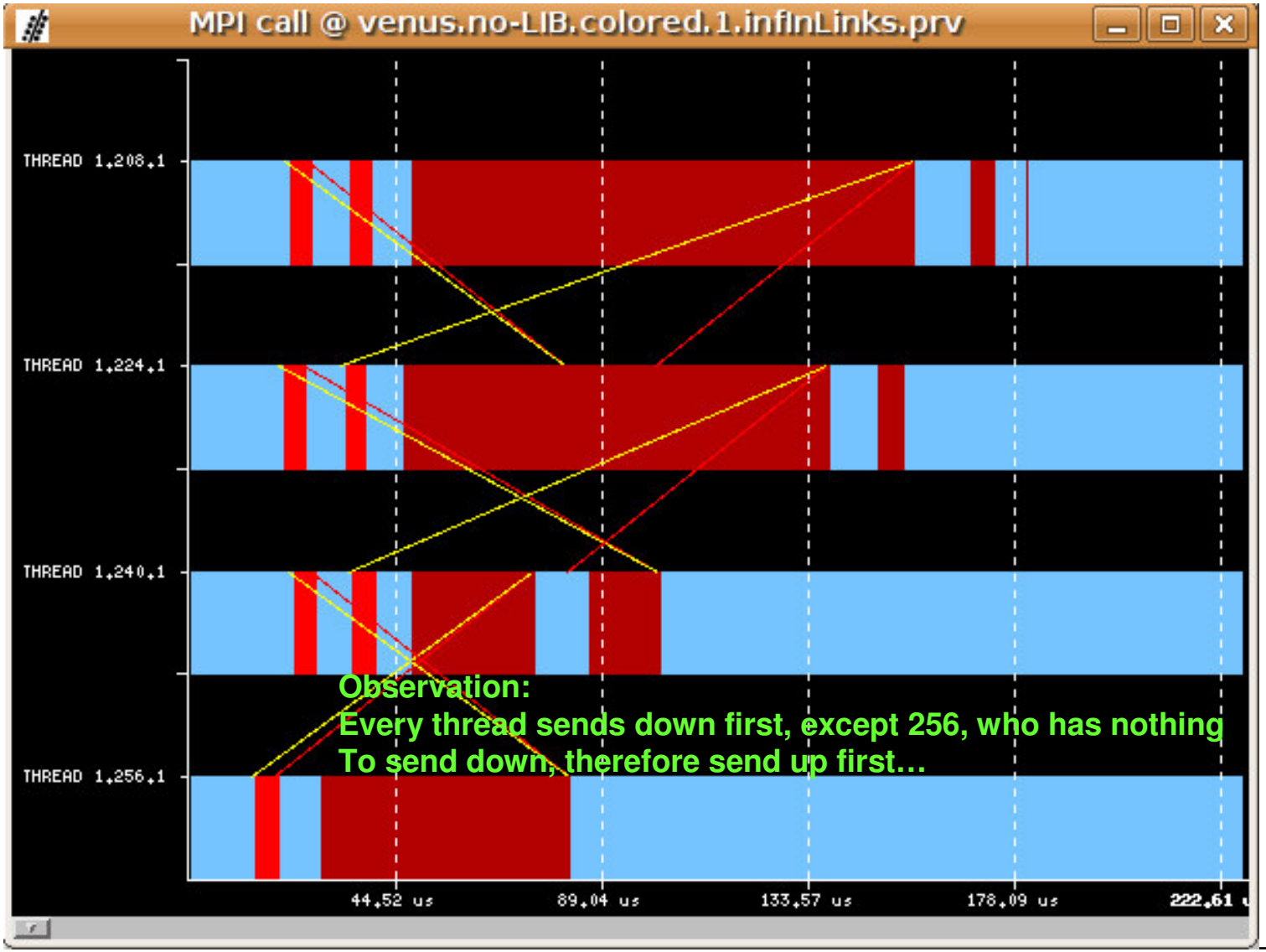


- WRF: 256 threads, sim time ~ 2hours
  - Routes with no conflicts
  - Simulation by Venus
  - Load Balanced
- Some threads take longer:
  - Dependency chains

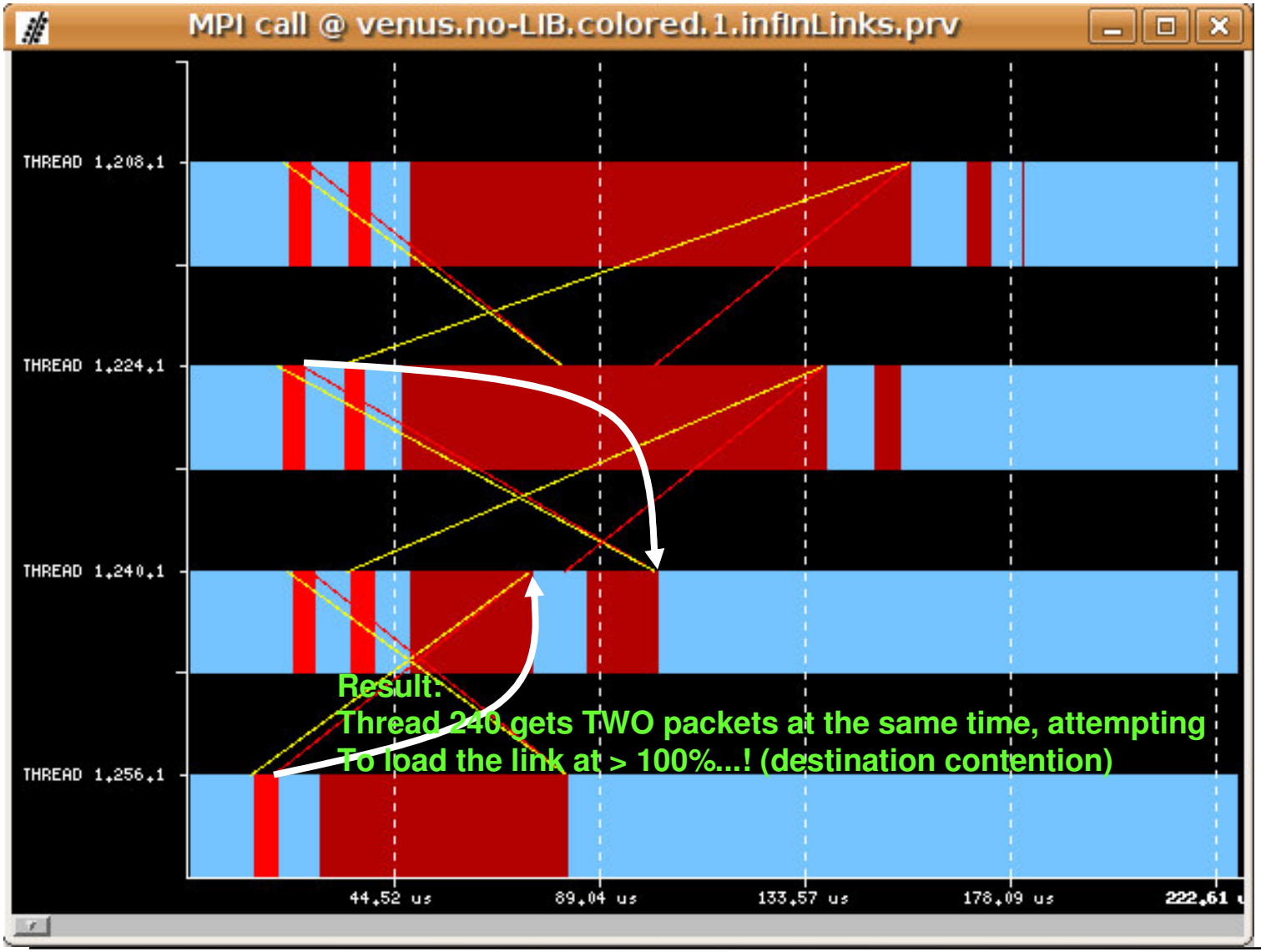
Communication Dependency chains



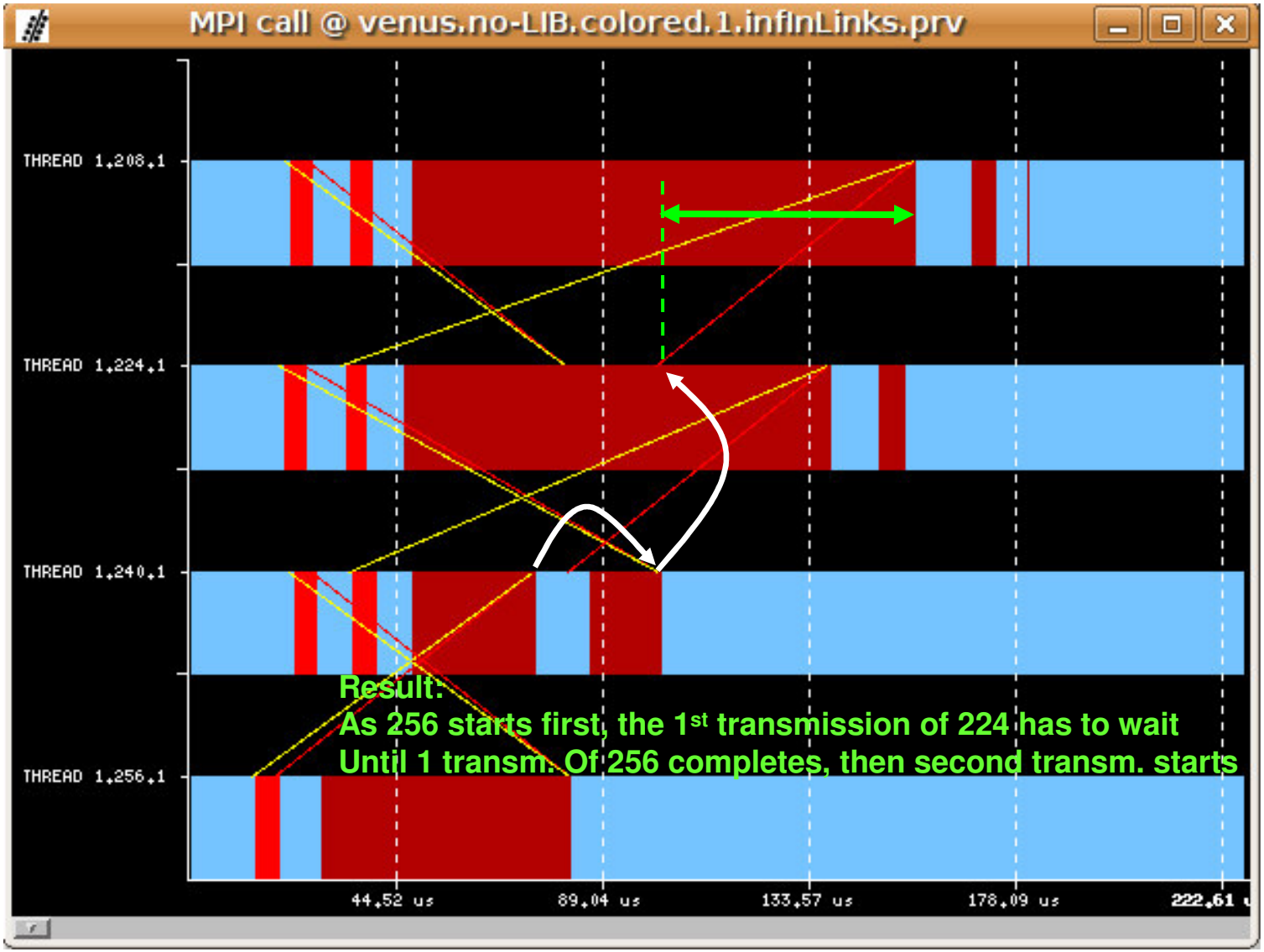
Communication Dependency chains



Communication Dependency chains



Communication Dependency chains



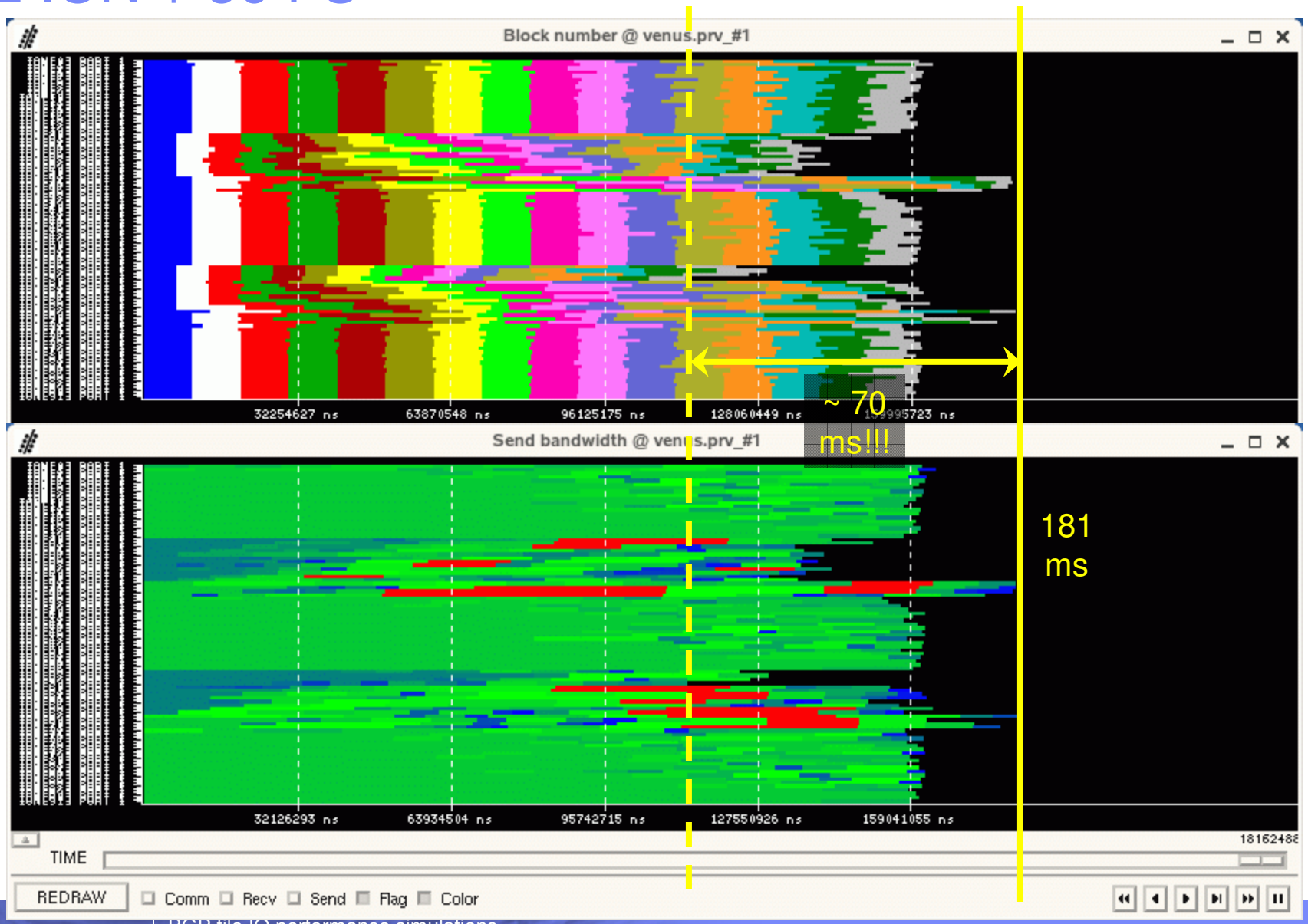
We also study file I/O performance…

Checkpointing on a BG/P machine, using GPFS

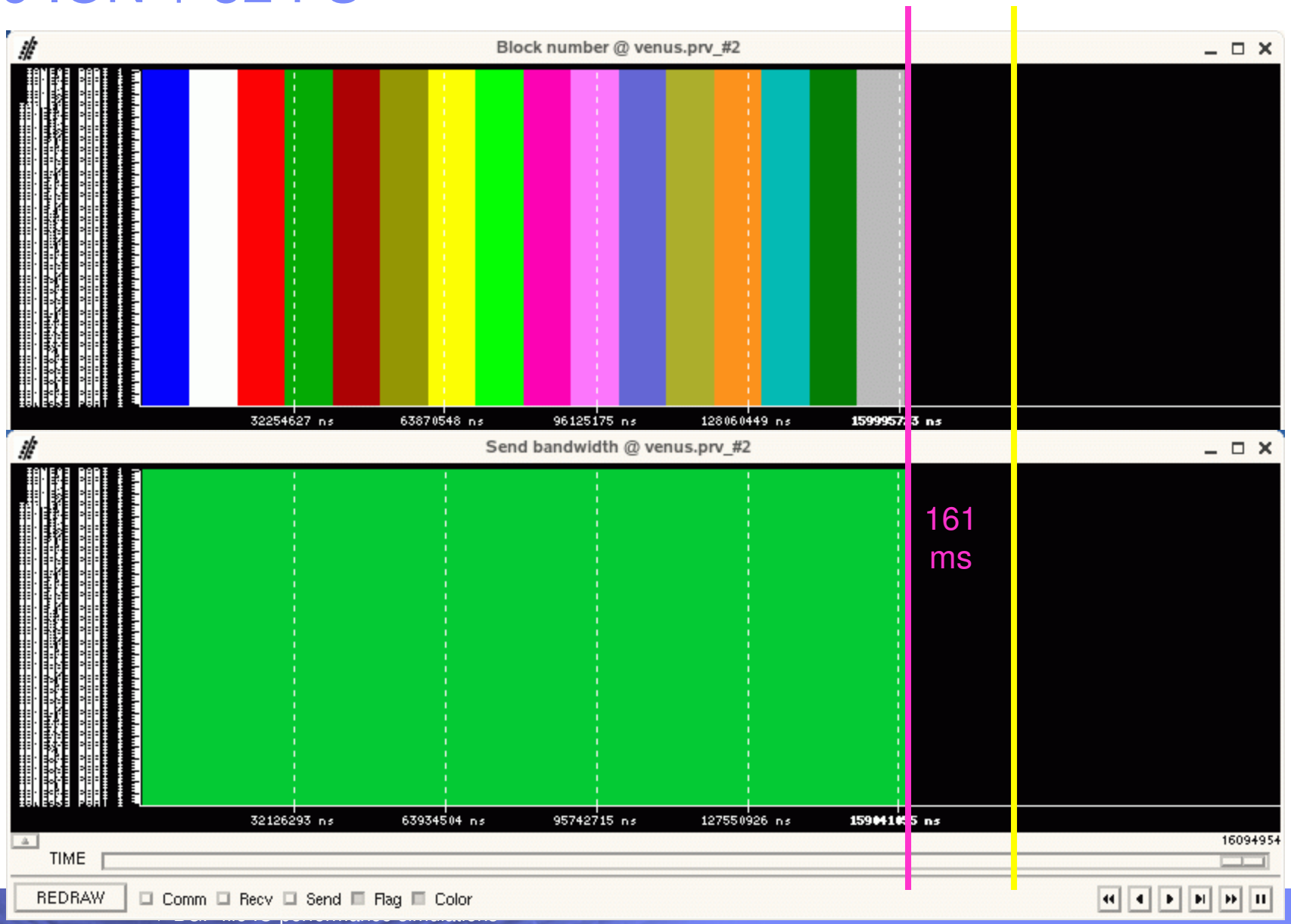
Simulation parameters

- Network
  - 128 nodes total
  - Three-level fat tree
  - 8-port switches: 32+32+16 = 80 switches
- Nodes
  - One compute node per IO node (ION)
  - 16 blocks of 4 MB each per node (64 MB per node)
  - ACK size = 512 B
- Myrinet
  - Flit size = 256 B
  - Flit duration = 204.8 ns (10 Gb/s)
  - Switch buffer size = 4 KB/port
  - No overhead (raw BW = user BW)
  - 4 KB segments; interleaved in groups of 4 segments by adapter
  - All latencies zero (link, switch, adapter)
- Three configurations
  - 92 ION + 36 FS
  - 96 ION + 32 FS
  - 100 ION + 28 FS

92 ION + 36 FS


Block number @ venus.prv_#1
32254627 ns
63870548 ns
96125175 ns
128060449 ns
~ 70 ms!!!
Send bandwidth @ venus.prv_#1
32126293 ns
63934504 ns
95742715 ns
127550926 ns
159041055 ns
181 ms
TIME
REDRAW
Comm
Recv
Send
Flag
Color

96 ION + 32 FS


Block number @ venus.prv_#2
32254627 ns
63870548 ns
96125175 ns
128060449 ns
Send bandwidth @ venus.prv_#2
32126293 ns
63934504 ns
95742715 ns
127550926 ns
161 ms
16094954
TIME
REDRAW
Comm
Recv
Send
Flag
Color

Insights gained 5 years after our transition

- Compute vs communicate
- What am I measuring, really?
- If a porsche were your hpc

Literature

- **"A Framework for End-to-end Simulation of High performance Computing Systems",** Wolfgang E. Denzel, Jian Li, Peter Walker, Yuho Jin. Simutools 2008
- **"Trace-driven Co-simulation of High-Performance Computing Systems using OMNeT++",** Cyriel Minkenberg, Germán Rodriguez Herrera. Simutools 2009