The imagination driving Australia's ICT future.

NATIONAL ICT AUSTRALIA LIMITED

Comprehensive Throughput Evaluation of LANs in Clusters of PCs with Switchbench

or

How to Bring Your Switch to Its Knees

Felix Rauch

National ICT Australia

felix.rauch@nicta.com.au

Australian Government
Department of Communications, Information Technology and the Arts
Australian Research Council

NICTA Members

THE AUSTRALIAN NATIONAL UNIVERSITY
UNSW
Department of State and Regional Development
First for Business
BusinessACT ACT GOVERNMENT

Victoria The Place To Be
The University of Melbourne
The University of Sydney
Queensland Government
Griffith University
Queensland University of Technology
The University Of Queensland Australia

NICTA Partners

CLUSTERS OF PCS

Harness the power of many compute nodes coupled together.

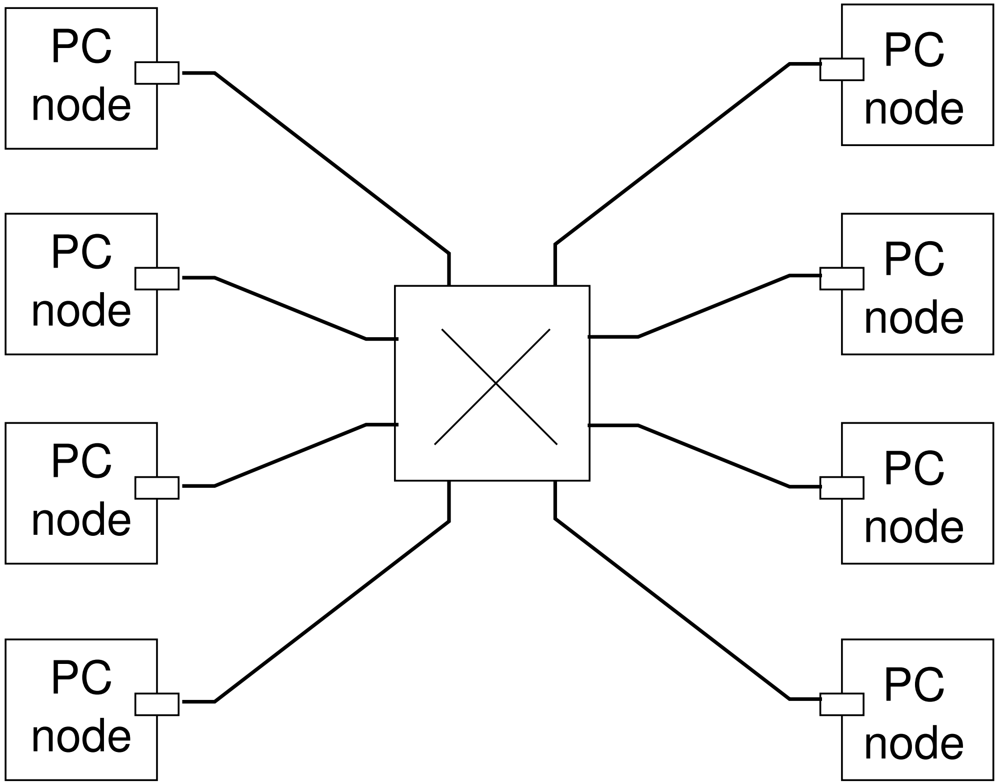
Rack-mounted compute cluster

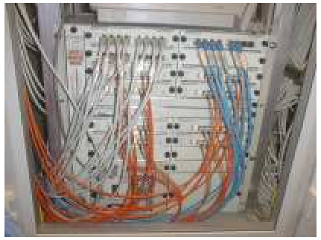
Network of workstations

Successful because:

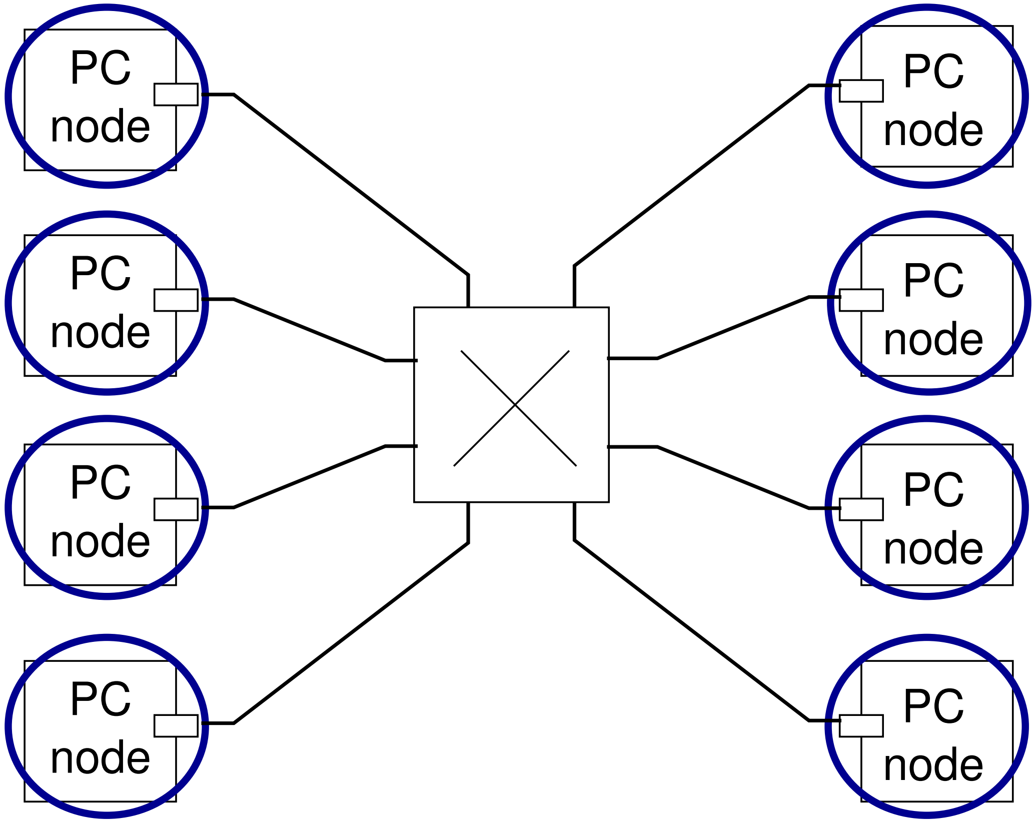
- Commodity off-the-shelf components (PCs, LAN)
- Often do-it-yourself approach
- Cost-effective high-performance computing

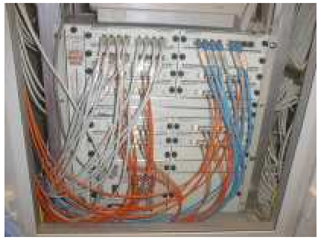
Understanding Performance in Clusters of Commodity PCs



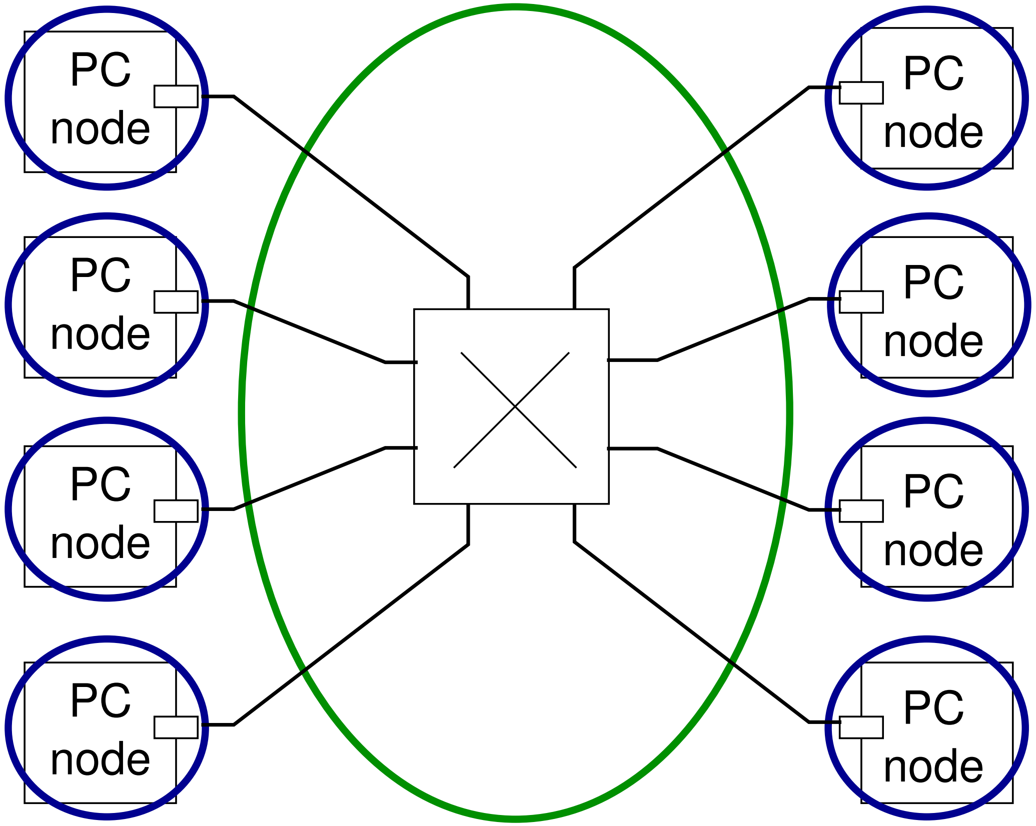


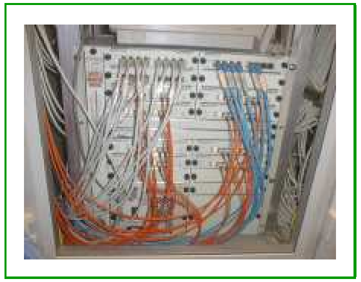
Understanding Performance in Clusters of Commodity PCs



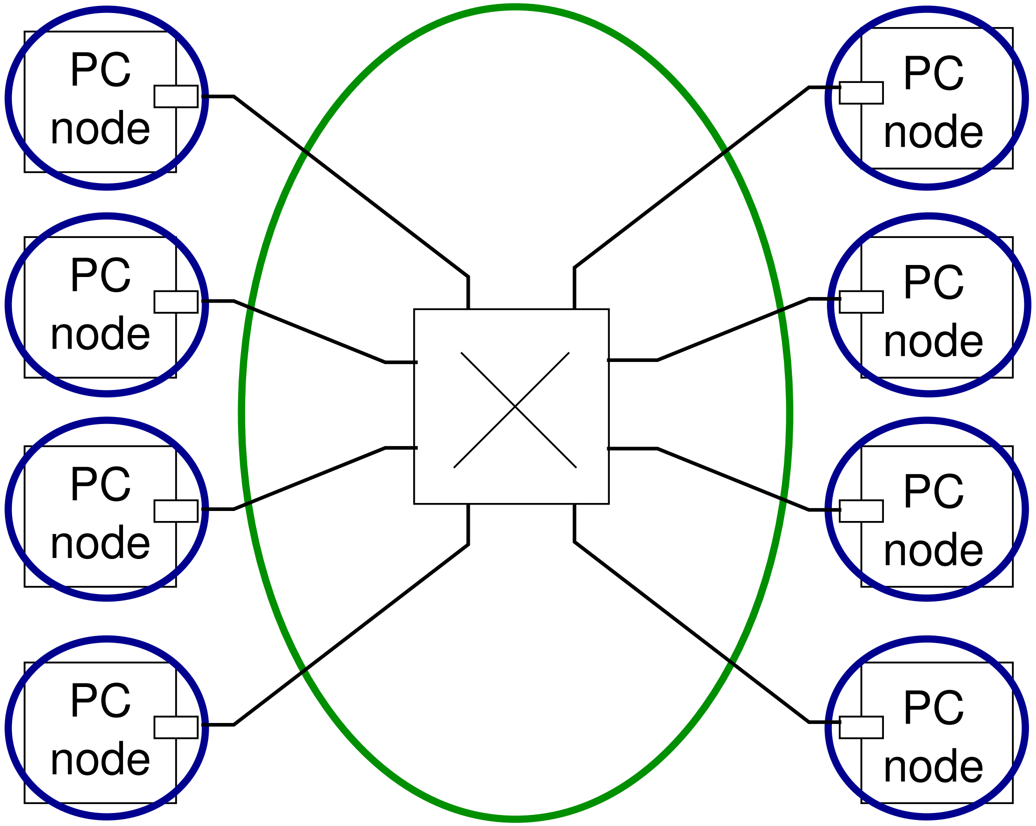


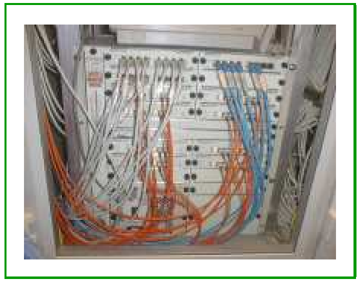
UNDERSTANDING PERFORMANCE IN CLUSTERS OF COMMODITY PCS





Understanding Performance in Clusters of Commodity PCs





Switchbench measures the overall network performance.

OVERVIEW

- Introduction
- Network Performance
- Evaluation principles
- Switchbench microbenchmarks with evaluation examples
- Conclusions

Network Performance in Clusters of PCs

Supercomputers:

- Balanced
- Full bisection
- Remote deposit

➜ Built by design

Commodity Clusters:

- Cheap (commodity) parts
- One-fits-all (LAN)
- Sometimes hacks to improve performance

➜ Built by shopping

NETWORK PERFORMANCE IN CLUSTERS OF PCS

Supercomputers:

- Balanced
- Full bisection
- Remote deposit

➜ Built by design

Commodity Clusters:

- Cheap (commodity) parts
- One-fits-all (LAN)
- Sometimes hacks to improve performance

➜ Built by shopping

Problems when choosing commodity components (they are all different!):

- make sure products adhere to specifications (not all do!)
- know performance characteristics (they differ widely!)

Network Performance in Clusters of PCs

Supercomputers:

- Balanced
- Full bisection
- Remote deposit

➜ Built by design

Commodity Clusters:

- Cheap (commodity) parts
- One-fits-all (LAN)
- Sometimes hacks to improve performance

➜ Built by shopping

Problems when choosing commodity components (they are all different!):

- make sure products adhere to specifications (not all do!)
- know performance characteristics (they differ widely!)

➜ Need benchmark tools for comprehensive evaluation.

Related Work: Performance Evaluation in Clusters

Analytic models:

- LogP (Culler 1993)
- LogGP (Alexandrov 1995)

Overall benchmark for parallel machines:

- High-Performance Linpack (Dongarra 1979)

Point-to-point network benchmarks:

- Netperf (Jones)
- NetPIPE (Turner)
- TTCP (PCAUSA)

Distributed network benchmark framework:

- IPbench (Wienand 2004)

BANDWITH VS. LATENCY

How to evaluate networks / switches?

Latency vs. bandwidth:

- Latency mostly "given by nature".
  Addressed with latency hiding techniques.
- One can purchase (additional) bandwidth.

There are more interesting cost/performance tradeoffs for additional bandwidth than for lower latency.

➜ Focus on bandwidth

How to measure bandwith of entire networks?

NETWORK LIMITATIONS

Three main limitations:

End nodes

Hardware: Network interface controller, CPU, memory, I/O bus.

Software: Communication protocol stack.

Switches

Processing limit (number of packets per second).

Internal bandwidth limitation.

Bisection bandwidth

Network architecture (topology).

FULL BISECTION BANDWIDTH

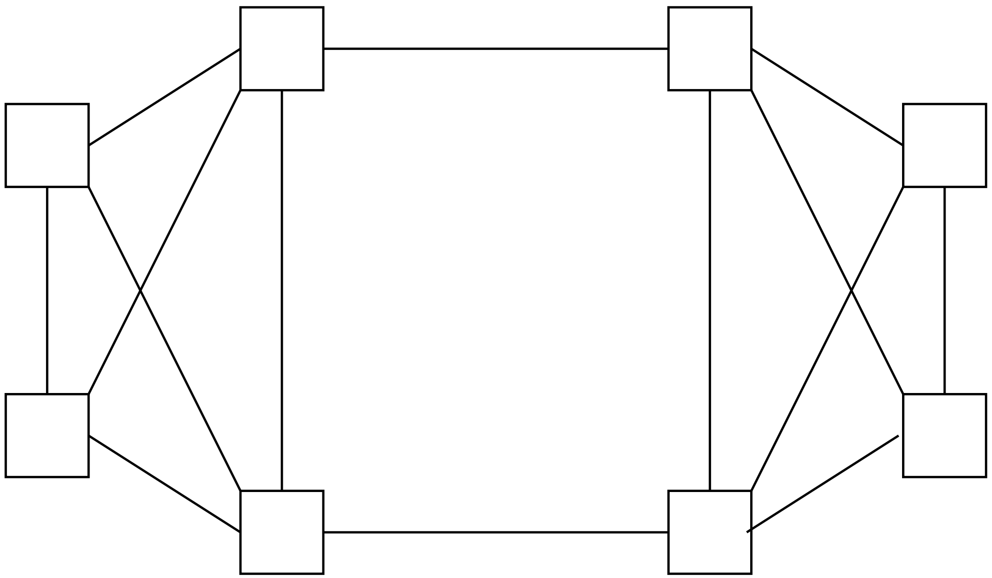
A network with N nodes has full bisection bandwidth if the sum of the link bandwidths between any two halves of the network is $N/2$ times the bandwidth of a single link.

$\Leftrightarrow$ Nodes of any two halves can communicate at full speed with each other.

Full Bisection Bandwidth

A network with N nodes has full bisection bandwidth if the sum of the link bandwidths between any two halves of the network is $N/2$ times the bandwidth of a single link.

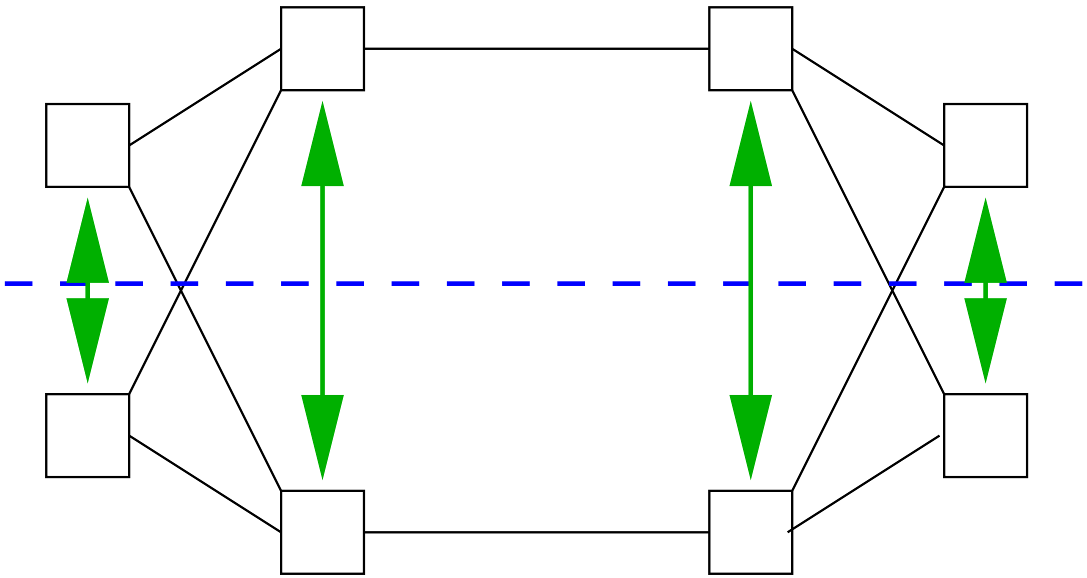
$\Leftrightarrow$ Nodes of any two halves can communicate at full speed with each other.



Full Bisection Bandwidth

A network with N nodes has full bisection bandwidth if the sum of the link bandwidths between any two halves of the network is $N/2$ times the bandwidth of a single link.

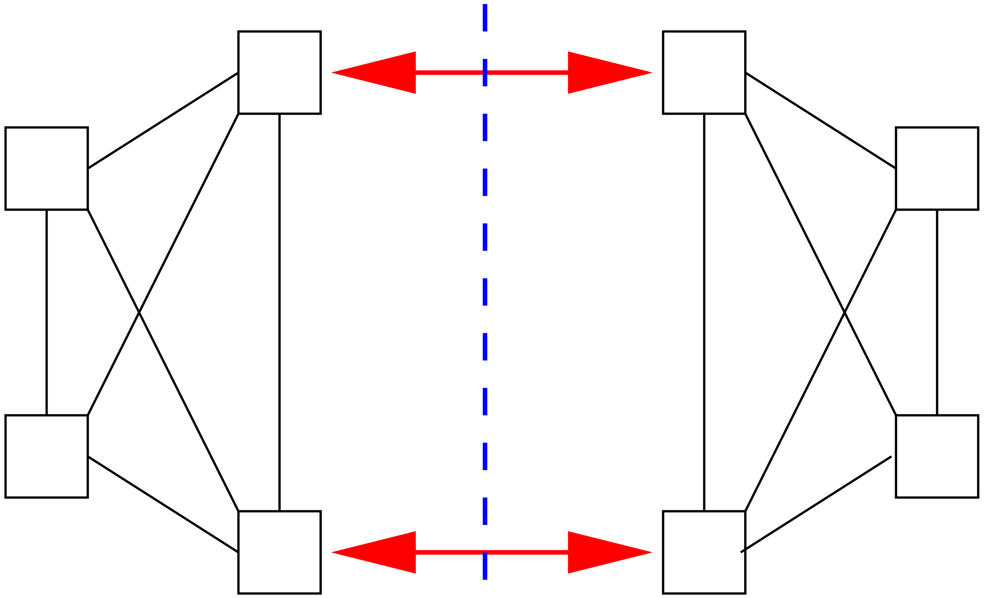
$\Leftrightarrow$ Nodes of any two halves can communicate at full speed with each other.



Full Bisection Bandwidth

A network with N nodes has full bisection bandwidth if the sum of the link bandwidths between any two halves of the network is $N/2$ times the bandwidth of a single link.

$\Leftrightarrow$ Nodes of any two halves can communicate at full speed with each other.



Full Bisection Bandwidth

A network with N nodes has full bisection bandwidth if the sum of the link bandwidths between any two halves of the network is $N/2$ times the bandwidth of a single link.

$\Leftrightarrow$ Nodes of any two halves can communicate at full speed with each other.

Important for programs with global communication patterns.

Important communication pattern requiring full bisection:

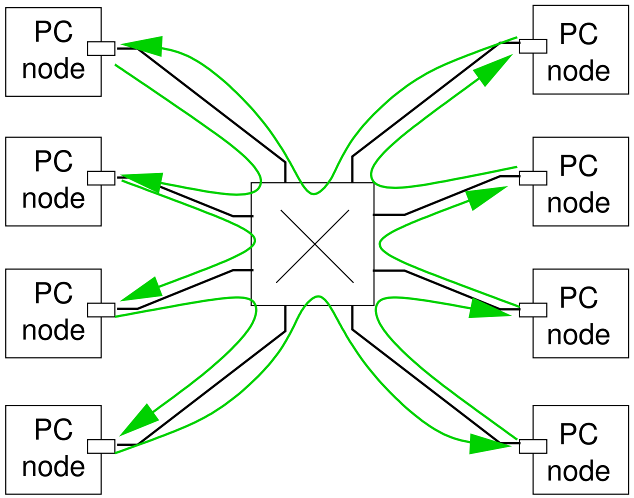
- All-to-all personalised communication (AAPC).
  Every node exchanges some data with every other node.

IMPLEMENTATION

- Based on earlier work done at ETH Zurich, together with C. Kurmann & T. Stricker.
- GNU public license.
- Core functionality in two small C programs.
- Shell scripts support:
  - starting programs on many nodes (by ssh)
  - specify node ranges
  - reordering of virtual node numbers to match physical layout
- Results in human-readable text file.
- Implemented and tested on GNU/Linux.

BENCHMARK: DAISY CHAIN

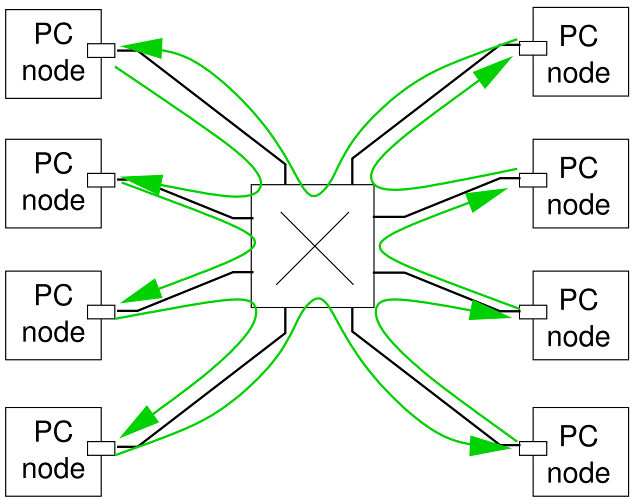
Virtual TCP daisy chain through an increasing number of nodes.



- ☑ Next-neighbour communication
- ☒ Bisection bandwidth not tested
- ☑ Full-speed duplex connections on all ports
- ☑ Limited by switch performance
- ☑ Increase load to find switch's limit

Benchmark: Daisy Chain

Virtual TCP daisy chain through an increasing number of nodes.



- ☑ Next-neighbour communication
- ☒ Bisection bandwidth not tested
- ☑ Full-speed duplex connections on all ports
- ☑ Limited by switch performance
- ☑ Increase load to find switch's limit

Result: Bandwidth of TCP chain.

Taken from Dolly partition-casting tool (disk cloning):

- Successfully used to install large clusters

Daisy-Chain Benchmark: Example Evaluation Platform

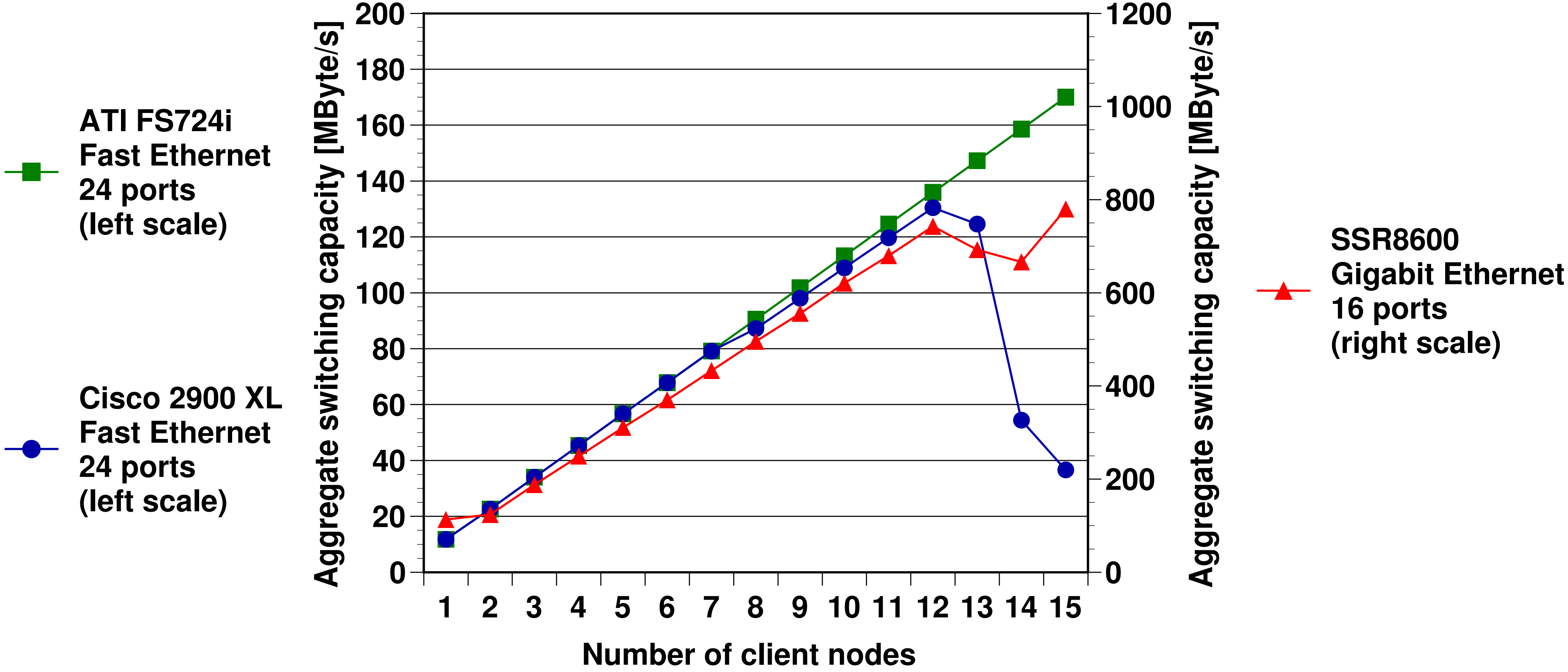
Cluster with 16 nodes:

- 2 Intel PentiumIII, 1 GHz
- 512 MByte RAM
- Intel Ethernet Pro 100, Fast Ethernet adapter
- Packet Engines G-NIC II, Gigabit Ethernet adapter

Experiments to compare performance characteristics of 3 different switches:

- Cisco 2900 XL Fast Ethernet switch (24 ports)
- ATI FS724I Fast Ethernet switch (24 ports)
- Cabletron SSR8600 Gigabit Ethernet switch (16 ports configured)

DAISY-CHAIN BENCHMARK: EXAMPLE EVALUATION

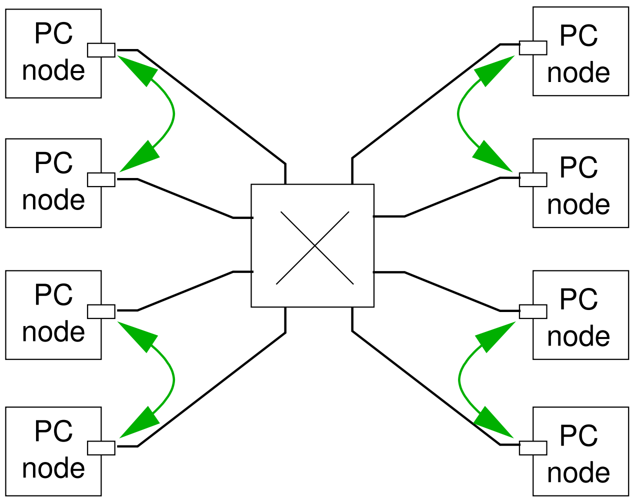


Benchmark: Pairwise Streaming

Any duplex communication pattern for increasing number of nodes.

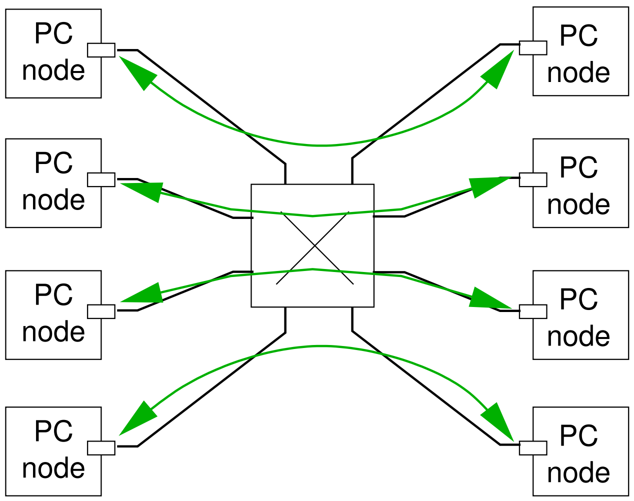
BENCHMARK: PAIRWISE STREAMING

Any duplex communication pattern for increasing number of nodes.



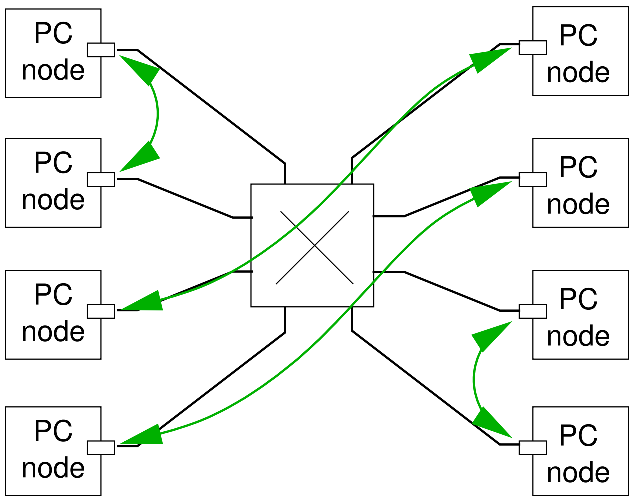
Benchmark: Pairwise Streaming

Any duplex communication pattern for increasing number of nodes.



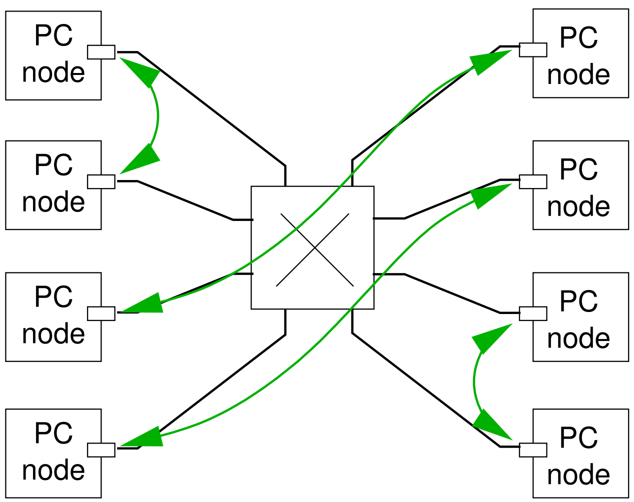
BENCHMARK: PAIRWISE STREAMING

Any duplex communication pattern for increasing number of nodes.



Benchmark: Pairwise Streaming

Any duplex communication pattern for increasing number of nodes.



- ☑ Great for debugging networks and switches
- ✗ Less automated
- ☑ Any pattern
- ✗ Cannot compare results

Result: Bandwidth of pairwise connections.

Successfully identified critical bottlenecks in commercial switches.

EXAMPLE EVALUATION PLATFORM

ETH “Xibalba” cluster with 128 nodes:

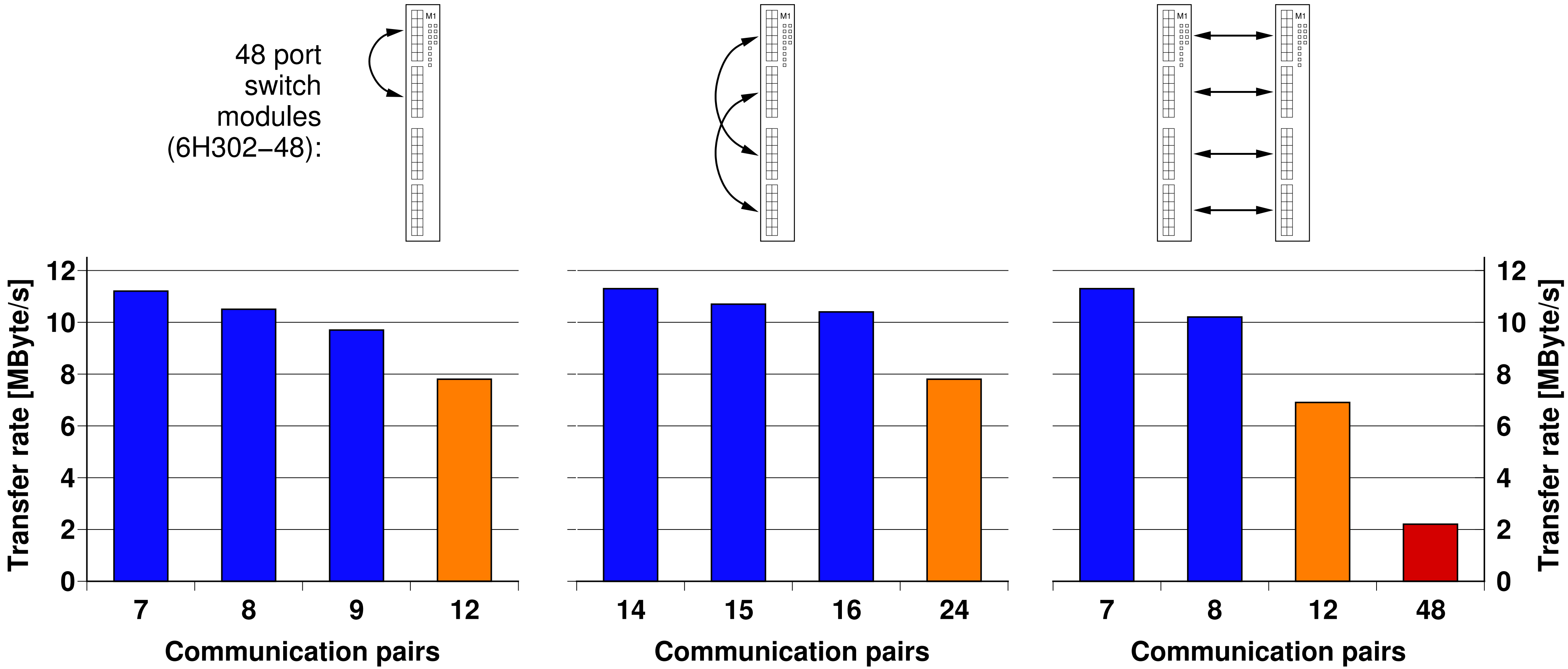
- 1–2 Intel PentiumIII, 1 GHz
- 256 MByte RAM per processor
- 2 Intel-based Fast Ethernet adapters
- Myrinet Gbit/s adapters (part.)

Network infrastructure:

- Enterasys Matrix E7 Fast Ethernet switch (mid range)

Evaluation with Pairwise Streaming

Detailed measurement to find limiting bisections on Matrix E7 switch.



Pairwise tests show severe inter-module bottleneck.

BENCHMARK: ALL-TO-ALL

Congestion-controlled all-to-all personalised communication (AAPC):

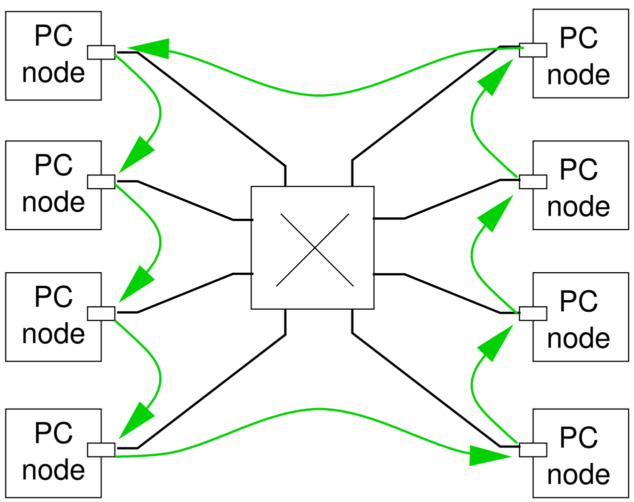
- Requires full bisection bandwidth
- Use phases to avoid congestion

parallel algorithm all-to-all

1 **for** $i = 1$ **to** $n - 1$ **do**
2 concurrently send data to node $n_{self+i \bmod n}$
and receive data from node $n_{self-i \bmod n}$
3 wait for barrier

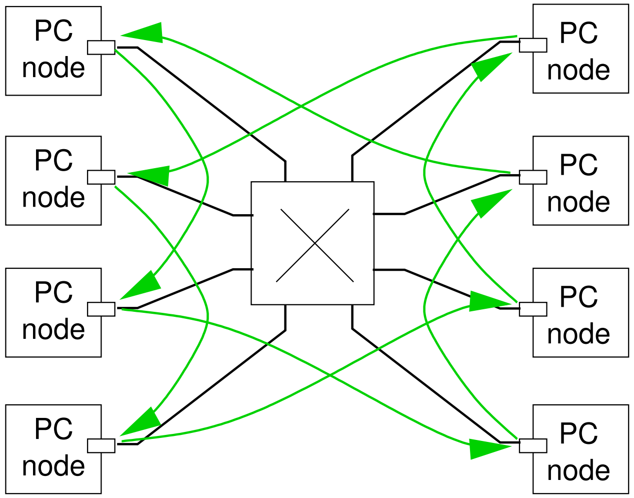
➜ Communication with increasing distance.

BENCHMARK: CONGESTION-CONTROLLED AAPC



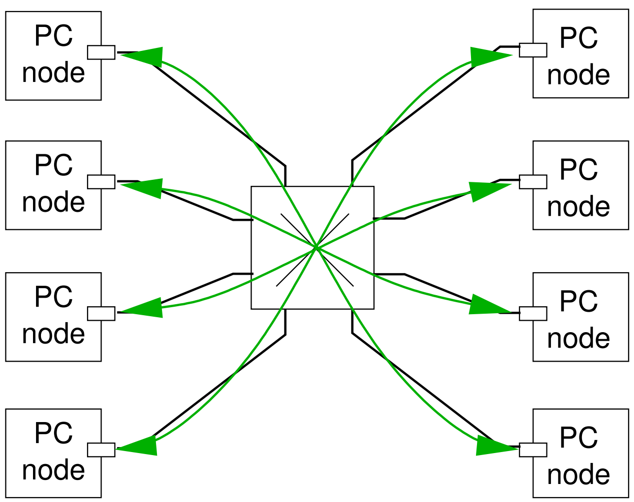
Phase 1

BENCHMARK: CONGESTION-CONTROLLED AAPC



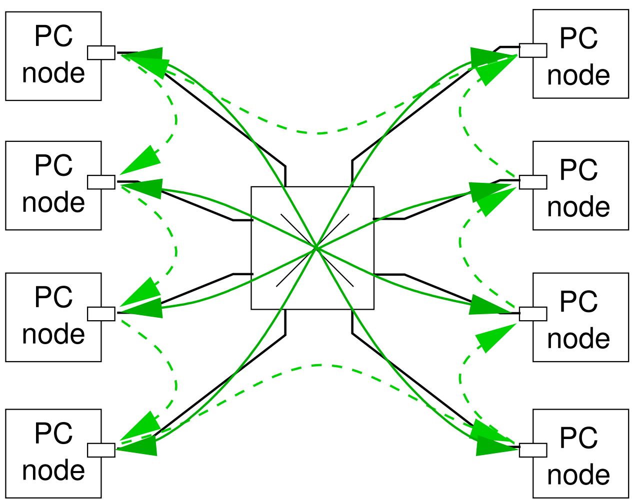
Phase 2

Benchmark: Congestion-Controlled AAPC



Phase 4

BENCHMARK: CONGESTION-CONTROLLED AAPC



- ☑ Automatic
- ☑ Comprehensively tests all communication distances
- ☑ More realistic communication pattern

- Simple result: Bandwidth for whole run
- More detailed results: Bandwidth for each phase

All-to-All Benchmark: Example Evaluation Platform

ETH "Xibalba" cluster with 128 nodes:

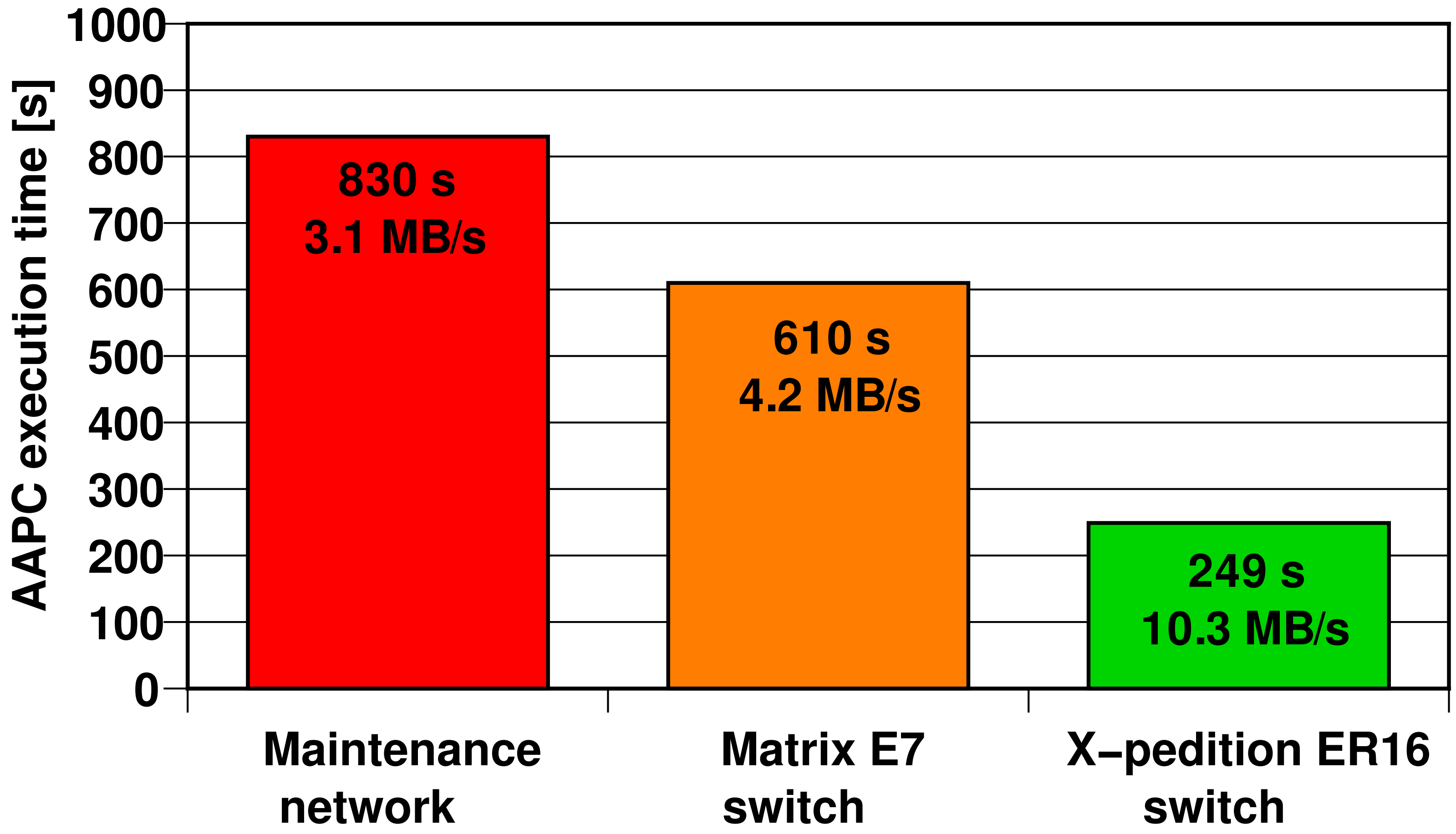
- 1–2 Intel PentiumIII, 1 GHz
- 256 MByte RAM per processor
- 2 Intel-based Fast Ethernet adapters
- Myrinet Gbit/s adapters (only 32 nodes)

Network infrastructure:

- Enterasys Matrix E7 Fast Ethernet switch (**mid range**)
- Enterasys X-pedition ER16 Fast Ethernet switch (**high end**)
- 8 Enterasys Horizon VH-2402 Fast Ethernet switches (**cheap DIY**)
- Myricom M3-E64 Gbit/s Myrinet switch (**Gbit/s class**)

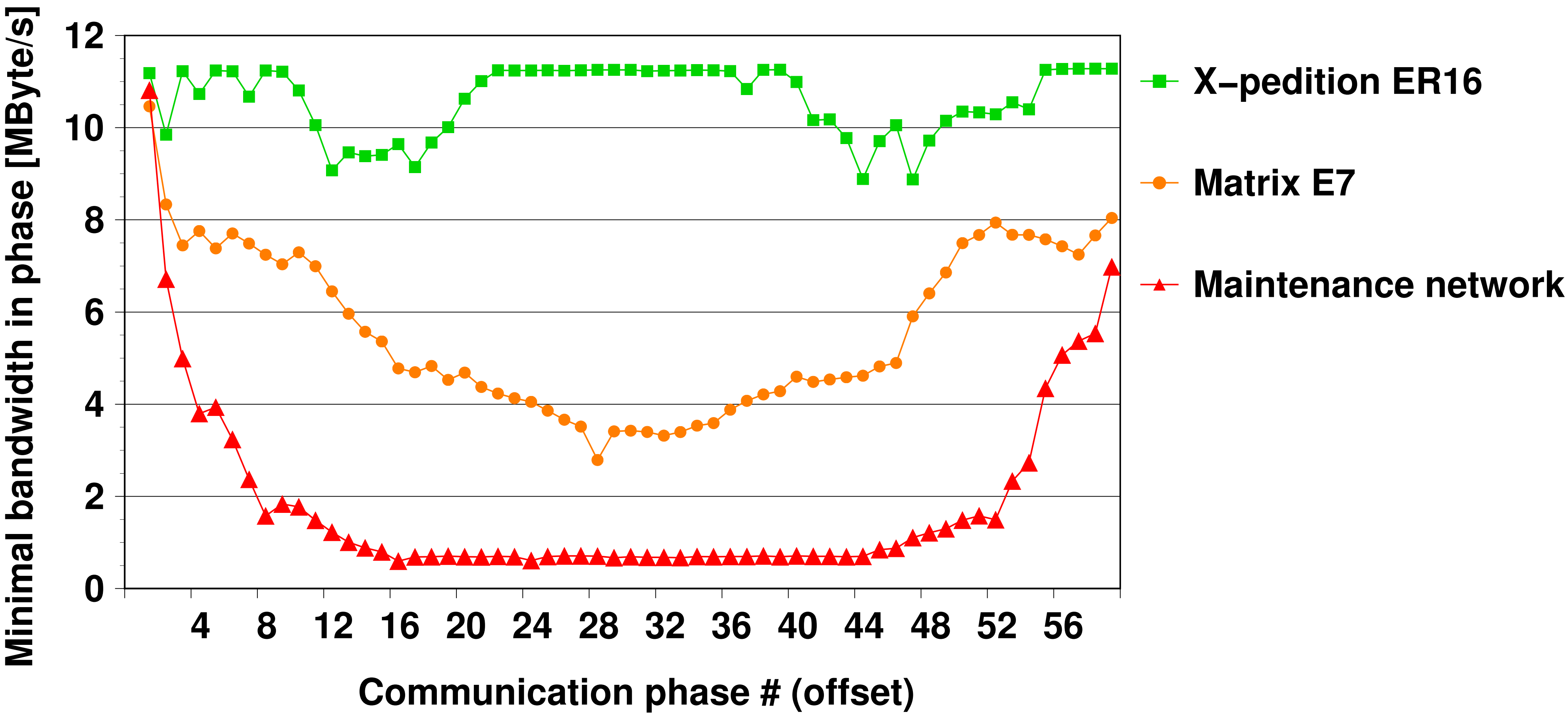
EVALUATION WITH ALL-TO-ALL: EXECUTION TIMES

Execution times of AAPC benchmark on different networks (60 CPUs):



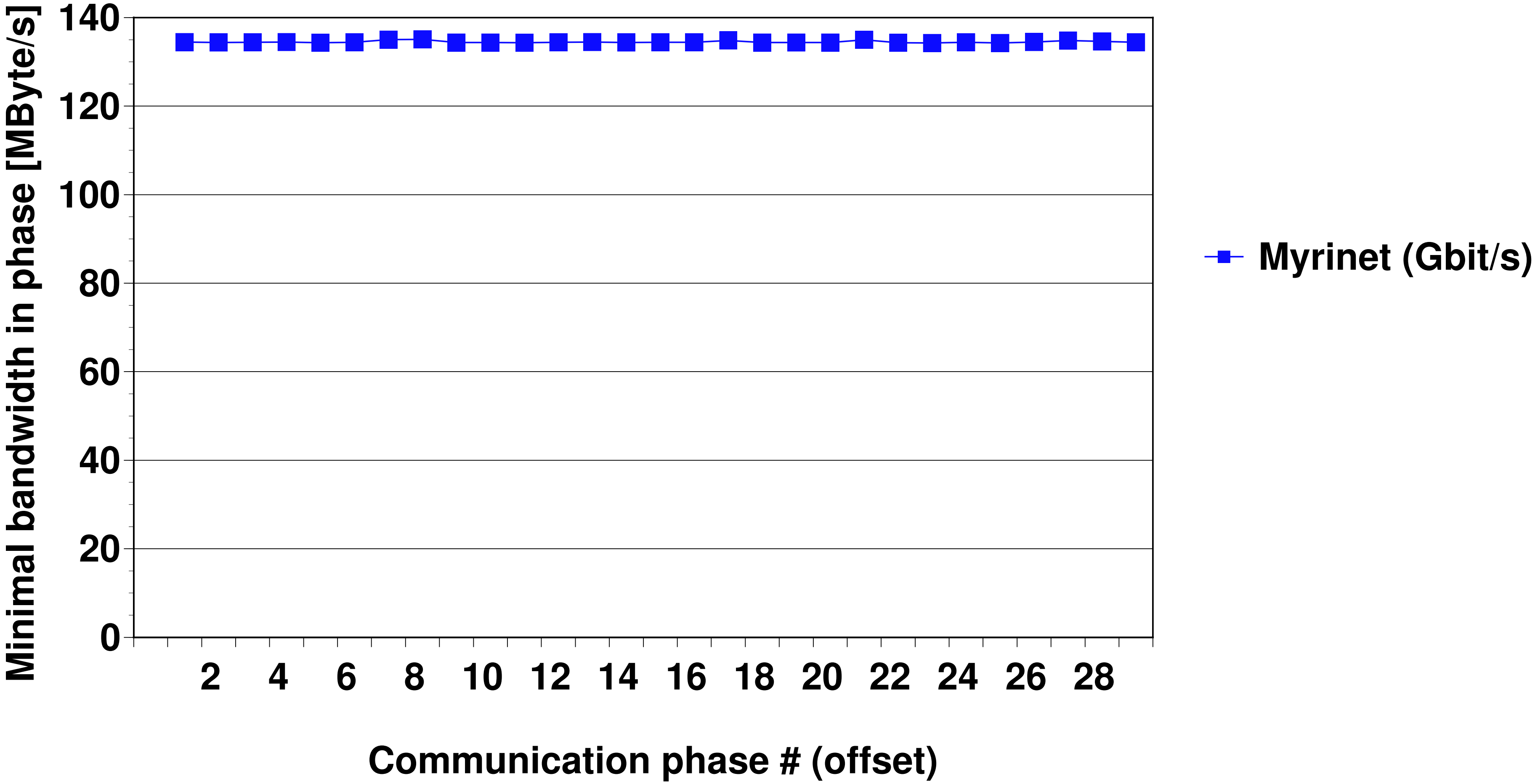
Evaluation with All-to-All: Phases

Minimal bandwidth for each phase:



EVALUATION WITH ALL-TO-ALL: PHASES

Minimal bandwidth for each phase:



CONCLUSIONS

Switchbench is a set of three microbenchmarks for measuring and debugging networks and switches.

Switchbench found:

- significant differences and variations in switch performance
- some data sheets are plain wrong!
  ➜ FREE switch upgrade from the producer

Switchbench is useful to:

- better understand performance
- better adapt applications to existing networks in clusters

Future work: Complete automatic performance characterisation.

Switchbench is a valuable tool to evaluate network performance.

QUESTIONS?

Switchbench download page:

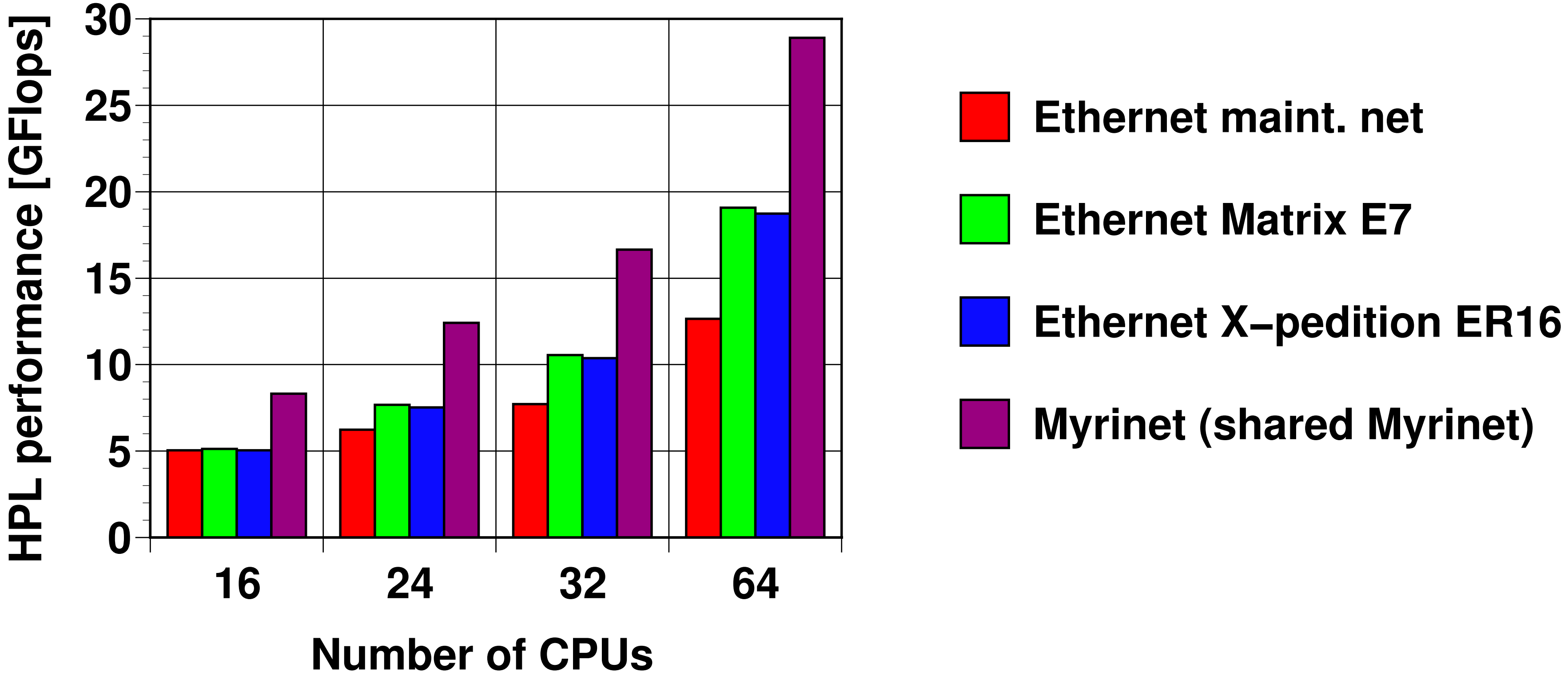
`http://www.ertos.nicta.com.au/Software/`

Embedded, Real-Time and Operating Systems (ERTOS)
research program,

National ICT Australia (NICTA)

APPLICATION BENCHMARK: HIGH-PERFORMANCE LINPACK (HPL)

Popular benchmark for supercomputers and clusters



APPLICATION BENCHMARK: QTPLAN LARGE-SCALE TRAFFIC SIMULATION