The imagination driving Australia’s ICT future.

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
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

Switchbench measures the overall network performance.

Switchbench — How to Bring Your Switch to Its Knees
Understanding Performance in Clusters of Commodity PCs

Switchbench — How to Bring Your Switch to its Knees
The imagination driving Australia’s ICT future.

UNDERSTANDING PERFORMANCE IN CLUSTERS OF COMMODITY PCs

Switchbench measures the overall network performance.
Switchbench measures the overall network performance.
OVERVIEW

- Introduction
- Network Performance
- Evaluation principles
- Switchbench microbenchmarks with evaluation examples
- Conclusions
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!)
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.
The imagination driving Australia’s ICT future.

**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)
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 bandwidth of entire networks?
The imagination driving Australia’s ICT future.

**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).
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.
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.
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.

$\Rightarrow$ 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.
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
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
Cluster with 16 nodes:

- 2 Intel Pentium III, 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

The imagination driving Australia’s ICT future.

SWITCHBENCH — HOW TO BRING YOUR SWITCH TO ITS KNEES
Any duplex communication pattern for increasing number of nodes.
Any duplex communication pattern for increasing number of nodes.
Any duplex communication pattern for increasing number of nodes.
Any duplex communication pattern for increasing number of nodes.
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.
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)
The imagination driving Australia’s ICT future.

**Evaluation with Pairwise Streaming**

Detailed measurement to find limiting bisections on Matrix E7 switch.

Pairwise tests show severe inter-module bottleneck.
The imagination driving Australia’s ICT future.

**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 $n_{self} + i \mod n$
   and receive data from node $n_{self} - i \mod n$
3. wait for barrier

→ Communication with increasing distance.
The imagination driving Australia’s ICT future.

**Benchmark: Congestion-Controlled AAPC**

Phase 1

Switchbench — How to Bring Your Switch to Its Knees
The imagination driving Australia’s ICT future.

**Benchmark: Congestion-Controlled AAPC**

Phase 2
The imagination driving Australia’s ICT future.

**Benchmark: Congestion-Controlled AAPC**

Phase 4
The imagination driving Australia’s ICT future.

**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
The imagination driving Australia’s ICT future.

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

- **Matrix E7 switch**
  - AAPC execution time: 610 s
  - Bandwidth: 4.2 MB/s

- **X-pedition ER16 switch**
  - AAPC execution time: 249 s
  - Bandwidth: 10.3 MB/s

- **Maintenance network**
  - AAPC execution time: 830 s
  - Bandwidth: 3.1 MB/s
The imagination driving Australia’s ICT future.

**EVALUATION WITH ALL-TO-ALL: PHASES**

Minimal bandwidth for each phase:

- **X-pedition ER16**
- **Matrix E7**
- **Maintenance network**

---

**SWITCHBENCH — HOW TO BRING YOUR SWITCH TO ITS KNEES**
The imagination driving Australia’s ICT future.

EVALUATION WITH ALL-TO-ALL: PHASES

Minimal bandwidth for each phase:

- **Myrinet (Gbit/s)**
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.
The imagination driving Australia’s ICT future.

**QUESTIONS?**

Switchbench download page:

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

National ICT Australia (**NICTA**)
The imagination driving Australia’s ICT future.

**APPLICATION BENCHMARK:**
**HIGH-PERFORMANCE LINPACK (HPL)**

Popular benchmark for supercomputers and clusters

<table>
<thead>
<tr>
<th>Number of CPUs</th>
<th>HPL performance [GFlops]</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td>32</td>
<td>10</td>
</tr>
<tr>
<td>64</td>
<td>25</td>
</tr>
</tbody>
</table>

- Ethernet maint. net
- Ethernet Matrix E7
- Ethernet X-pedition ER16
- Myrinet (shared Myrinet)

**SWITCHBENCH — HOW TO BRING YOUR SWITCH TO ITS KNEES**
The imagination driving Australia’s ICT future.

APPLICATION BENCHMARK: QTP PLAN LARGE-SCALE TRAFFIC SIMULATION

Switchbench — How to Bring Your Switch to Its Knees 39