Cost/Performance Tradeoffs in Network Interconnects for Clusters of PCs

Felix Rauch
Christian Kurmann and Thomas M. Stricker

Laboratory for Computer Systems, ETH Zürich

CoPs Project
http://www.cs.inf.ethz.ch/CoPs/

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

April 22, 2003

Commodity Clusters vs. Supercomputers

One important difference is the network.

Supercomputer network:

- Balanced
- Full bisection
- Remote deposit

➥ Built by design

Commodity cluster network:

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

➥ Built by shopping

Which commodity cluster networks are a bargain?

➡ We evaluate **commodity networks** with **supercomputer criterions**

Xibalba Commodity Cluster

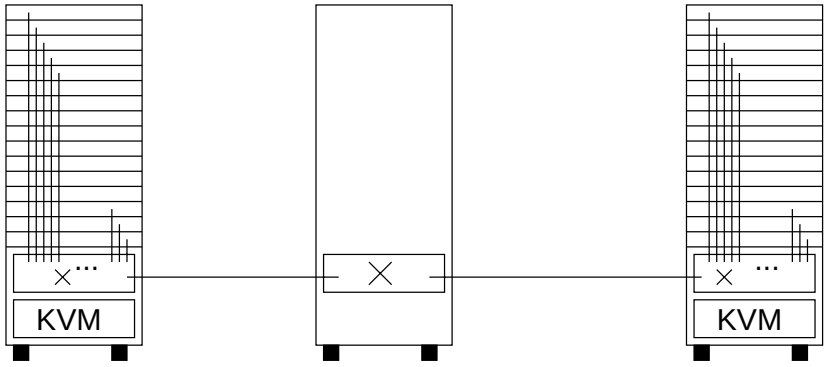
Experimental cluster used by 4 groups at computer science department of ETH Zurich, in service since september 2001:

- 128 nodes with 196 CPUs, 1 GHz PentiumIII
- 512 MB ECC-SDRAM per CPU
- 4×20 GB disk space
- 64 bit / 66 MHz PCI bus
- 2 Fast Ethernet adapters
- Enterasys Fast Ethernet switches
- Partially equipped with Myrinet adapter

Xibalba Network Options

Multiple installed networks:

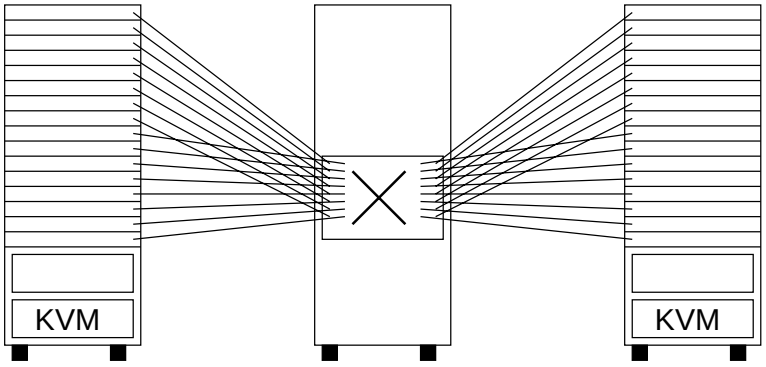
- **Maintenance network**
  (Fast Ethernet, small 24-port switches)



Xibalba Network Options

Multiple installed networks:

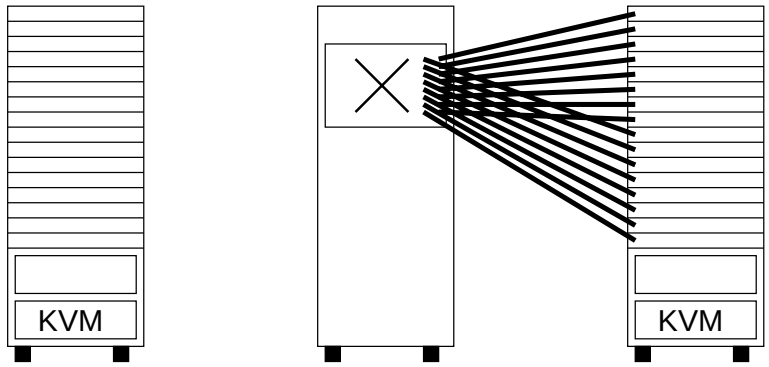
- Maintenance network
  (Fast Ethernet, small 24-port switches)
- **Full bisection network**
  (Fast Ethernet, Matrix E7 and X-pedition ER16)



Xibalba Network Options

Multiple installed networks:

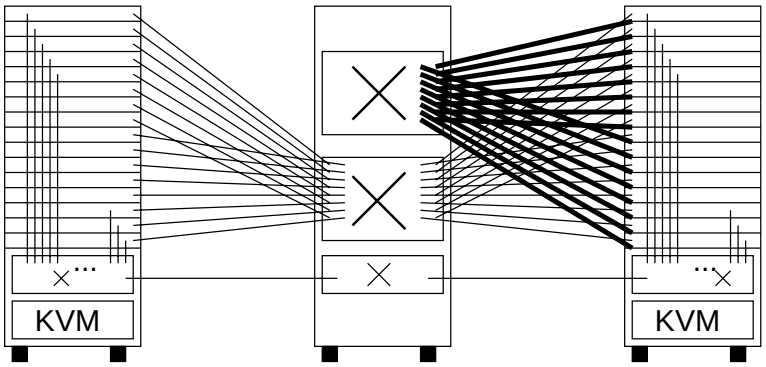
- Maintenance network
  (Fast Ethernet, small 24-port switches)
- Full bisection network
  (Fast Ethernet, Matrix E7 and X-pedition ER16)
- **High-performance network**
  (Myrinet 2000, only 32 dual nodes)



Xibalba Network Options

Multiple installed networks:

- Maintenance network
  (Fast Ethernet, small 24-port switches)
- Full bisection network
  (Fast Ethernet, Matrix E7 and X-pedition ER16)
- High-performance network
  (Myrinet 2000, only 32 dual nodes)



Evaluation Principles

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?

➡ Full bisection **bandwidth**

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.

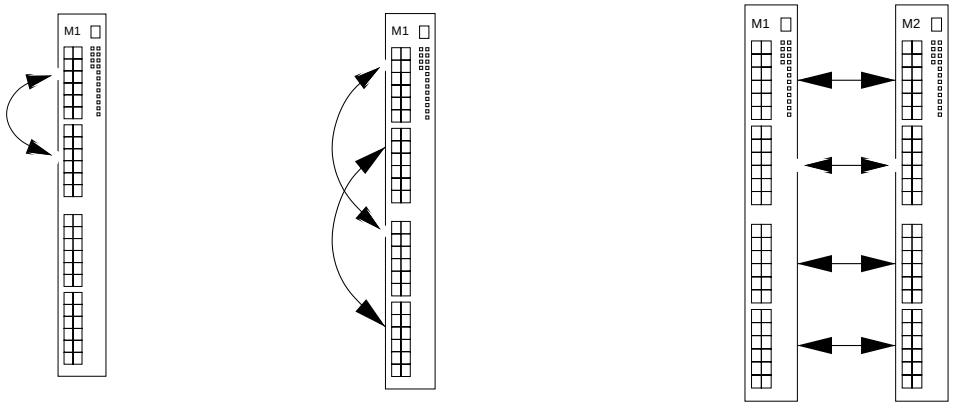
Cluster networks with full bisection:

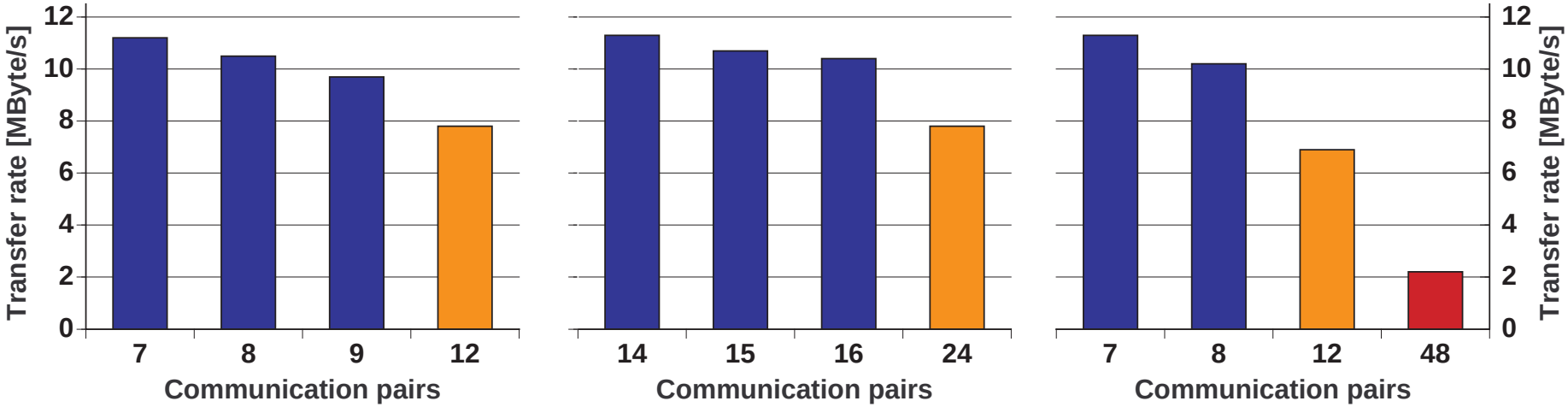
- Myrinet up to 128 nodes
  (expensive solutions for more than 128 nodes exist)
- Ethernet fat trees?
- Big Ethernet switches ?

Pairwise Communication on Matrix E7

Detailed measurement to find limiting bisections.

48 port switch modules (6H302-48):



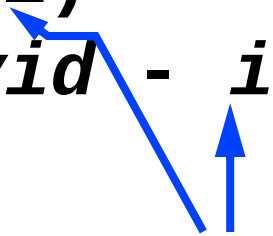


Pairwise tests explain good and bad AAPC

AAPC Communication Requiring Full Bisection Bandwidth

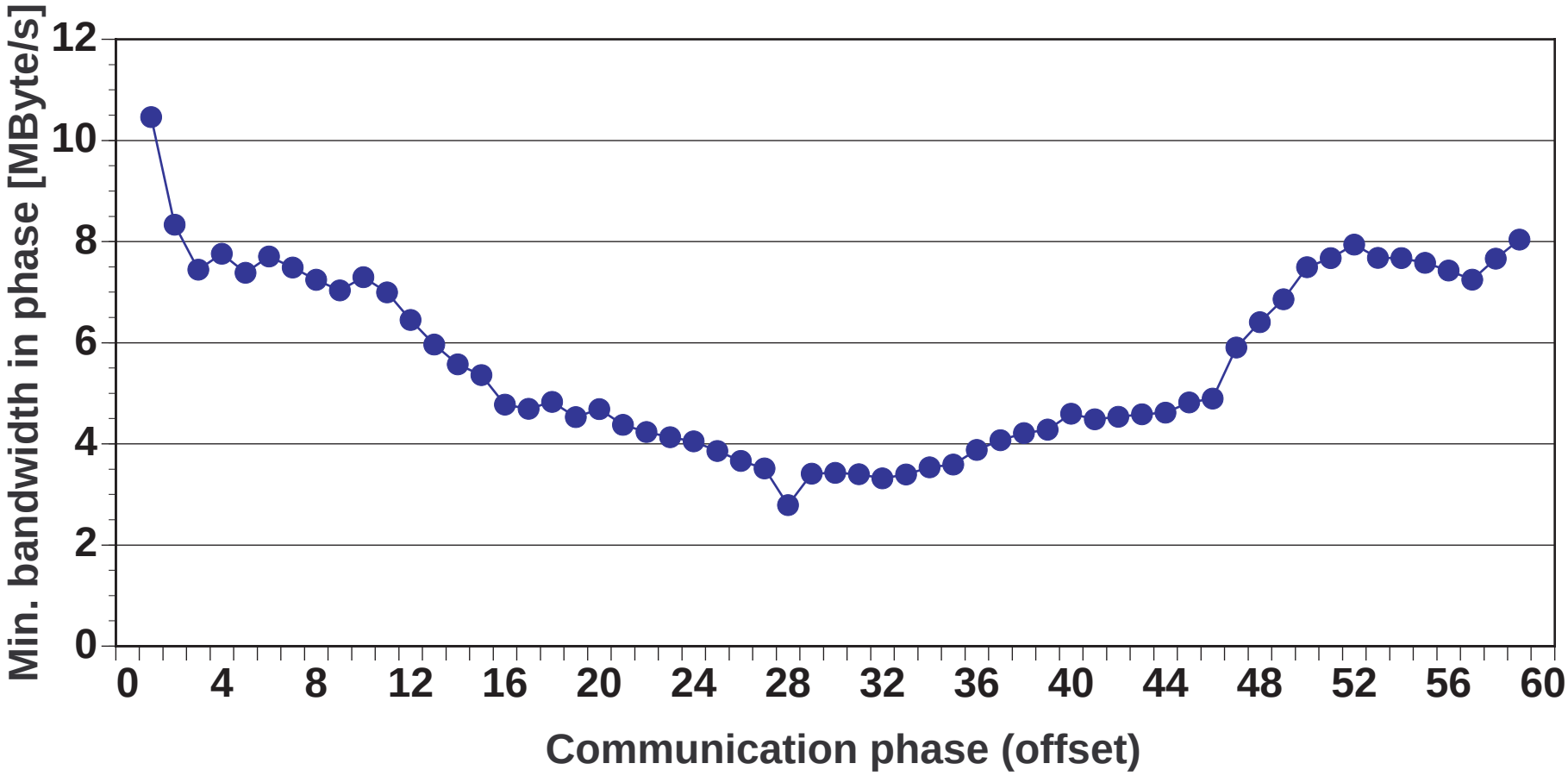
Congestion-controlled phased AAPC as microbenchmark:

```
Parallel algorithm all-to-all
for_all_nodes:
  for i = 1 to n do
    concurrently:
       send_to(myid + i)
       receive_from(myid - i)
    wait for barrier
```



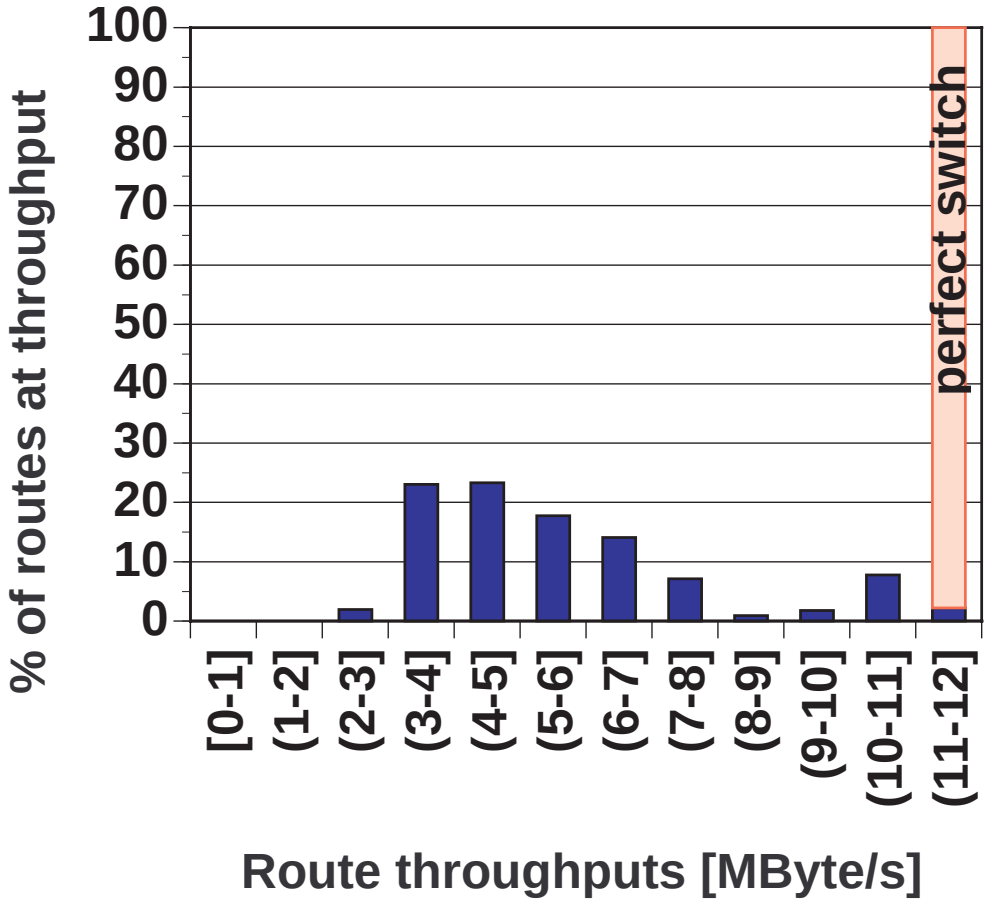
i: Communication with increasing distance

Example: AAPC on Matrix E7 Switch Phase Bandwidths



Problem with larger offsets: Inter-module communication

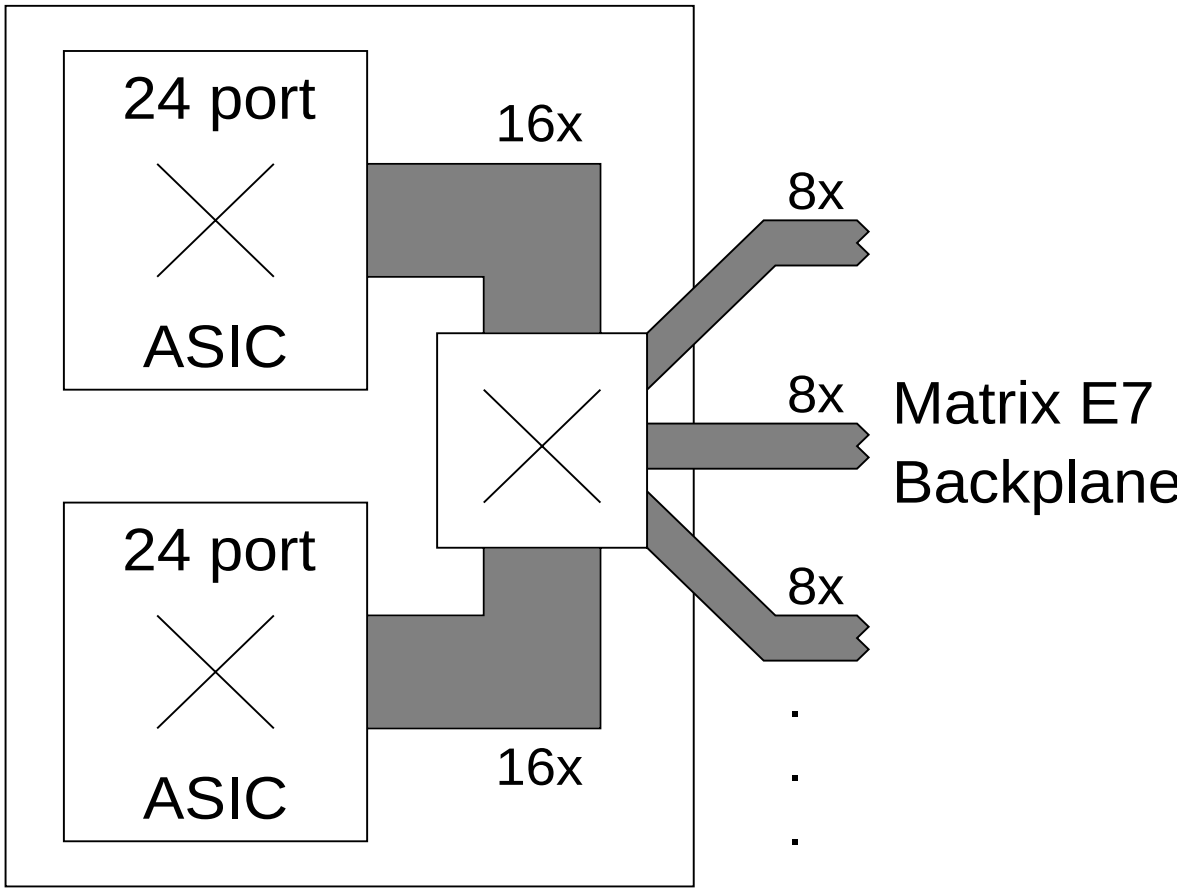
Example: AAPC on Matrix E7 Switch Route Histograms



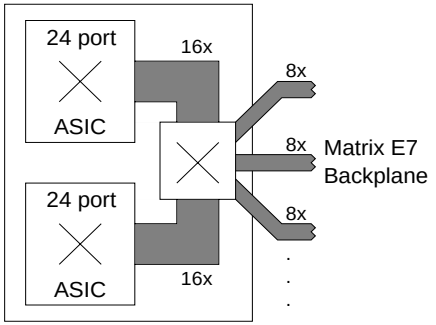
Bad routes result in bad overall performance

Matrix E7 Bottlenecks

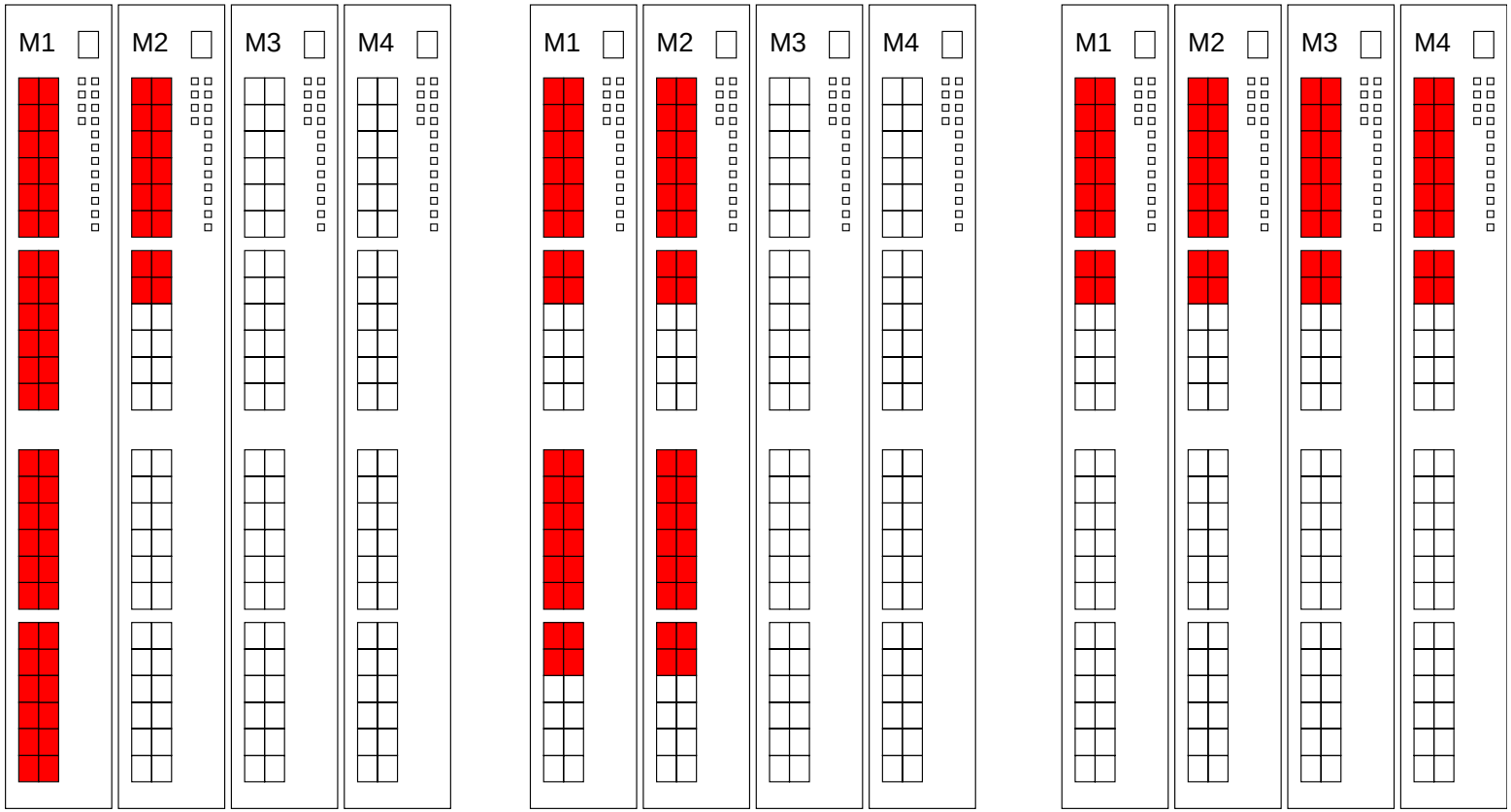
Our analysis revealed several bottlenecks:



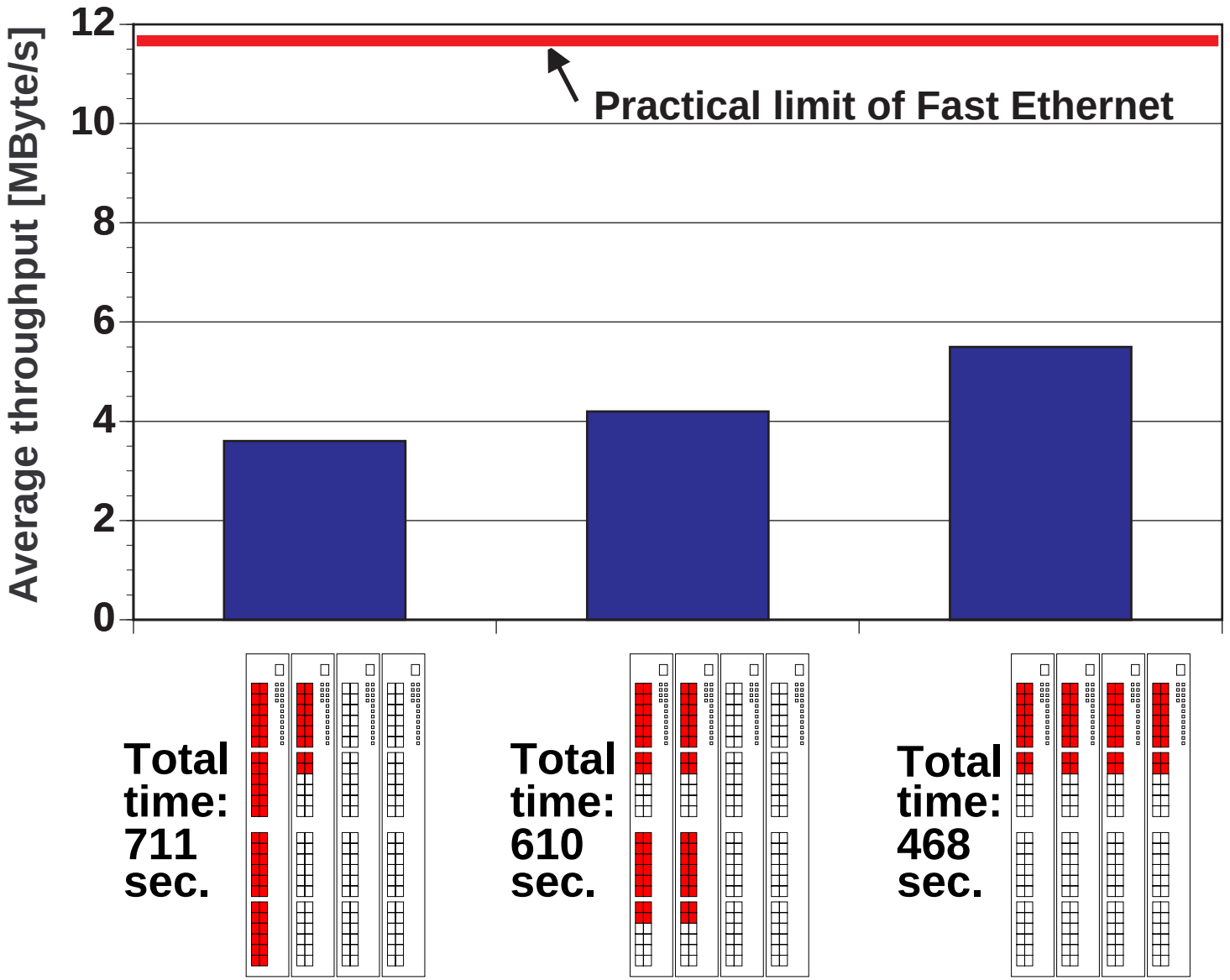
Matrix E7 Workarounds

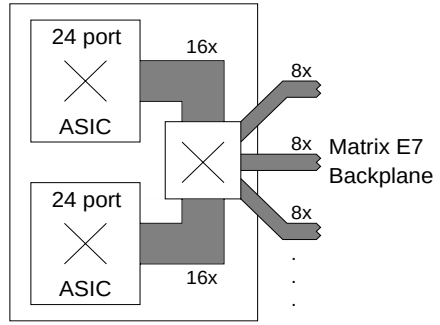


Different configurations:



Matrix E7 AAPC Performance





➡ E7 could **not** deliver advertised cost / performance ratio.

Comparison of Different Options

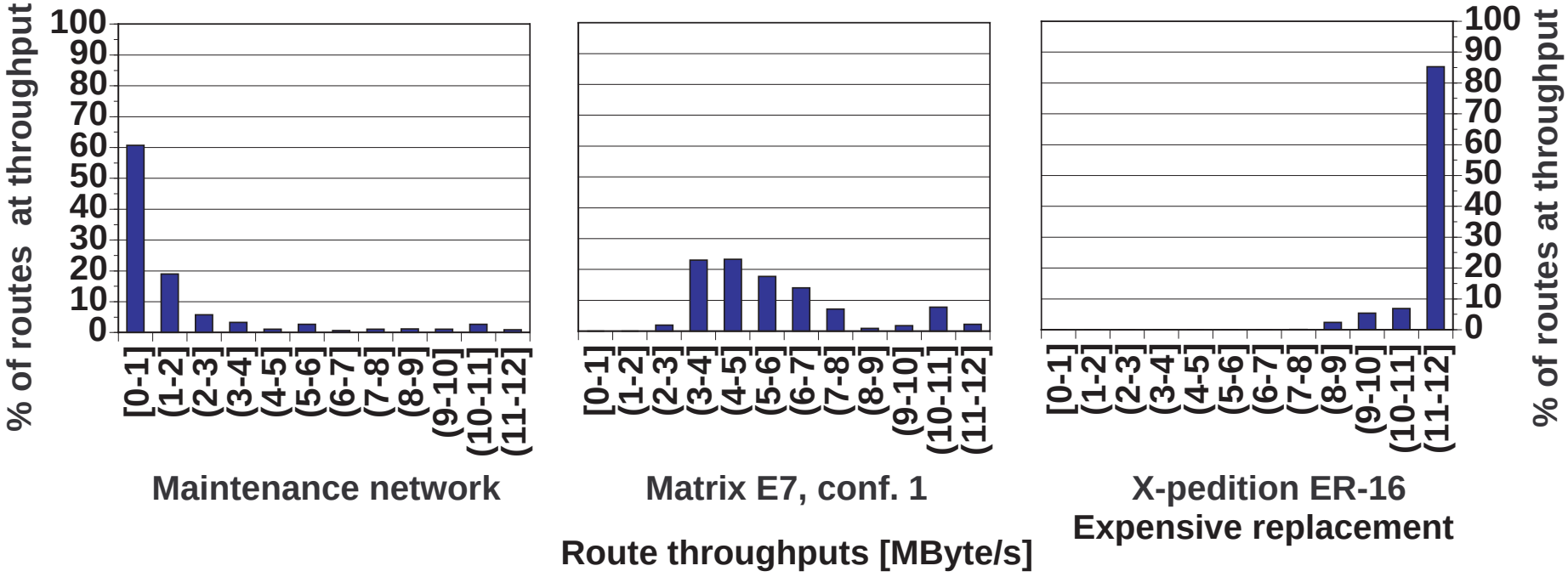
After a free upgrade to the more expensive **X-pedition ER16** switch, we could compare different Ethernet options:

Network	Cost per port [US$]	Predicted Performance
Maintenance	**145**	**low**
Matrix E7	**480**	~~**high**~~
X-pedition ER16	**810**	**high**

Comparison of Different Networks

All-to-all communication on 64 nodes = 64 × 64 routes.

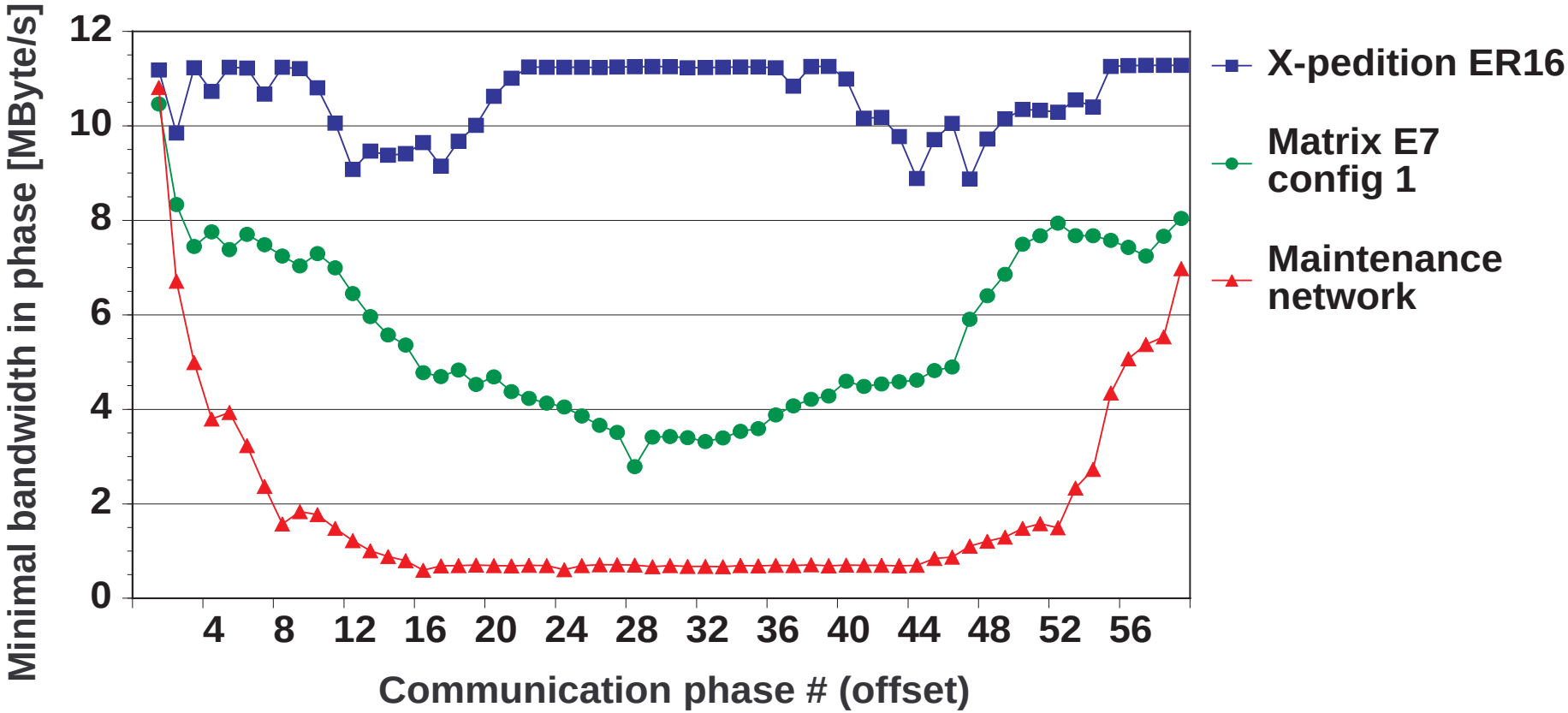
Histograms of route throughputs:



Route performance reflects global communication capabilities of the networks.

AAPC Microbenchmark on Ethernets

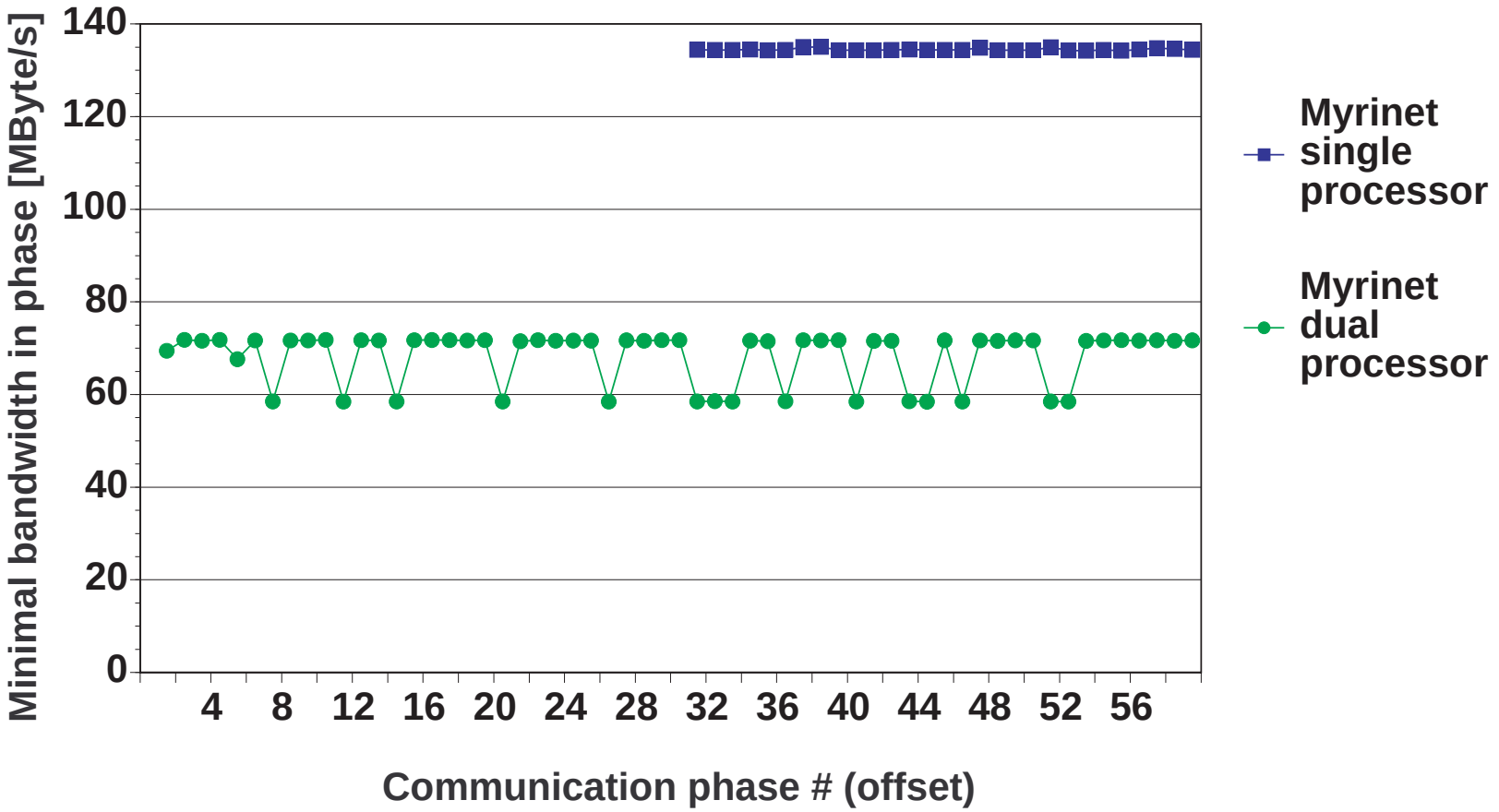
Minimal bandwidths for each phase:



Bad routes result in bad AAPC performance.

AAPC Microbenchmark on Myrinet

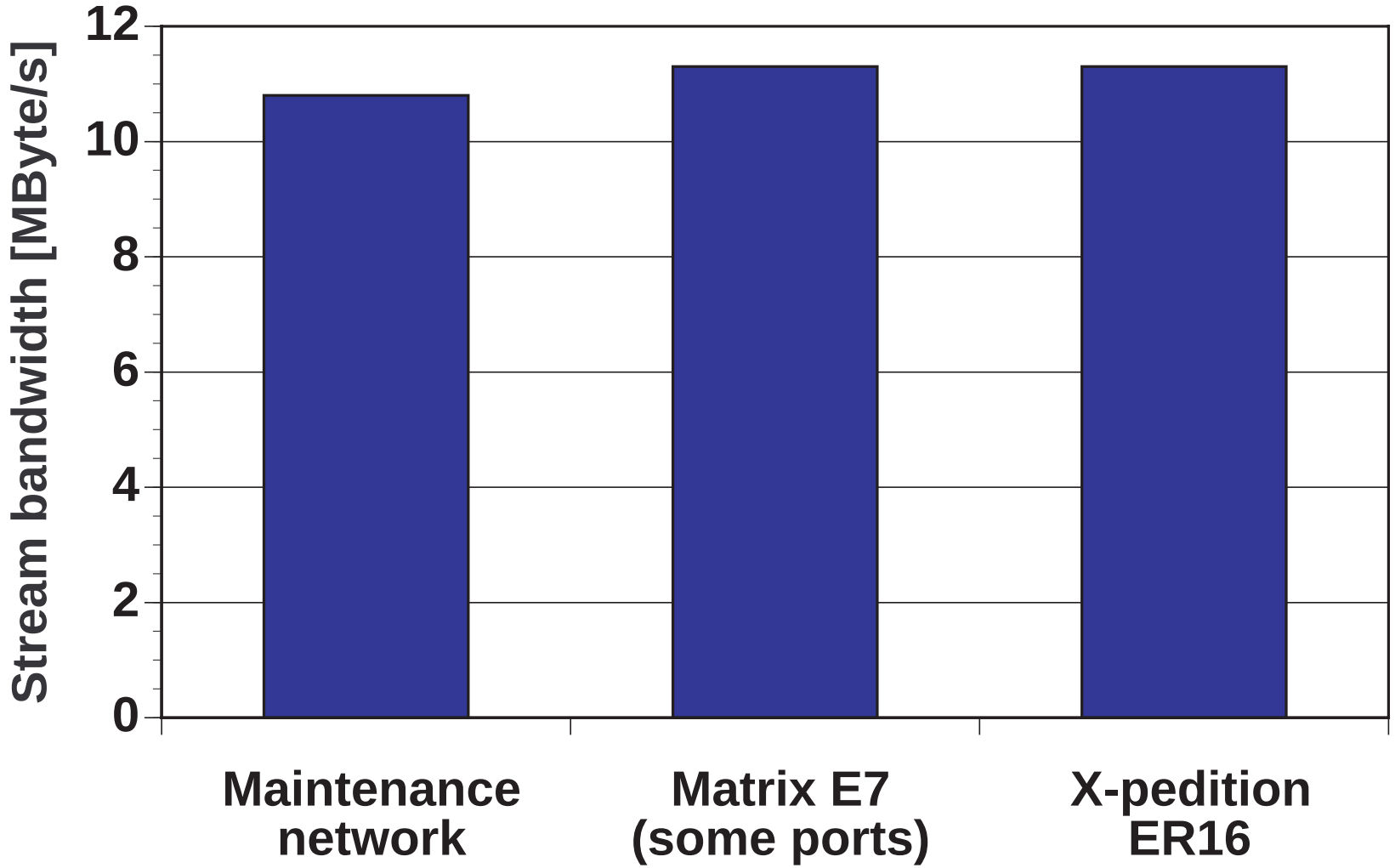
Dual configuration = 60 CPUs, single conf. = 30 CPUs



Perfect performance in single CPU configuration.

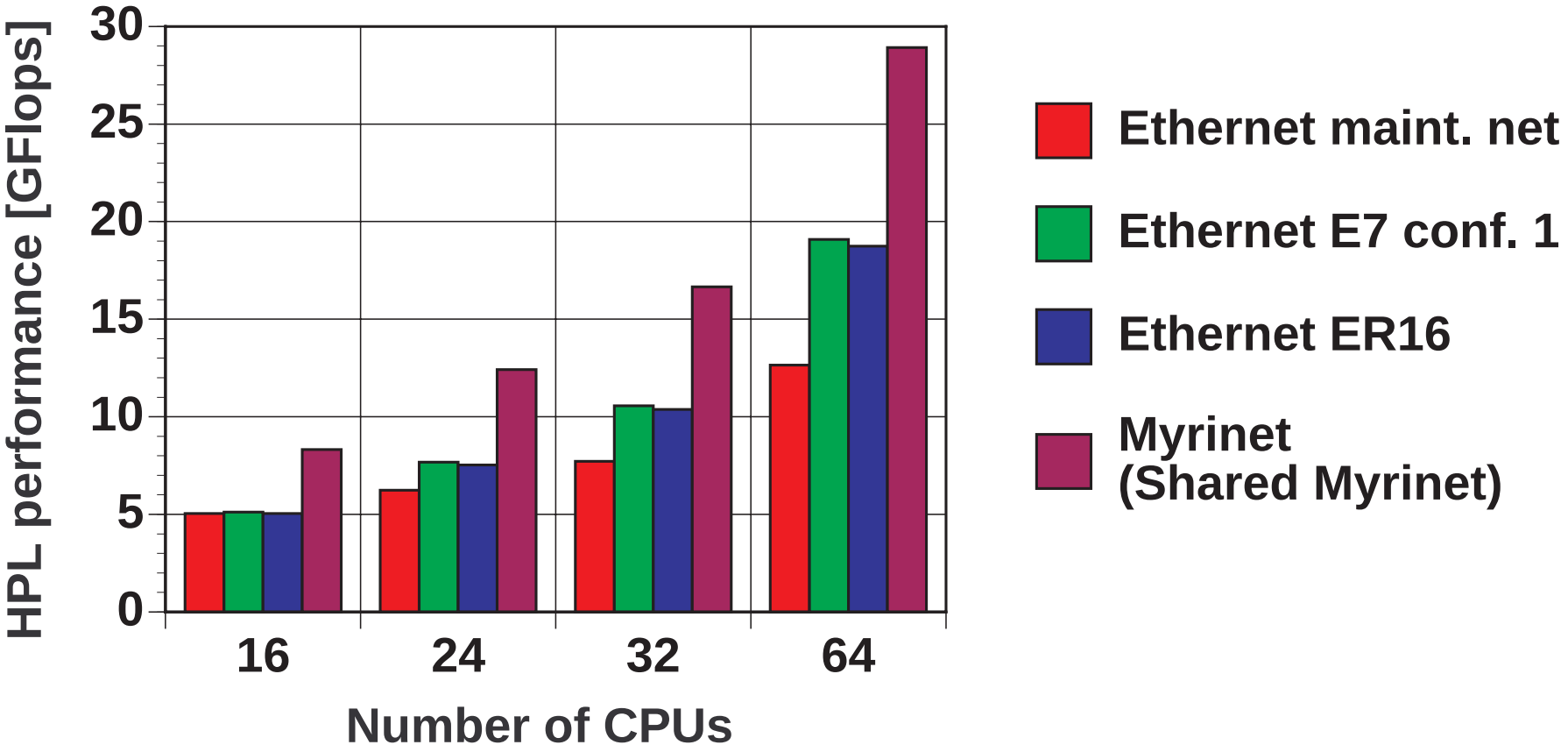
Application Benchmark with Dolly

Data-casting tool Dolly, next neighbor communication.

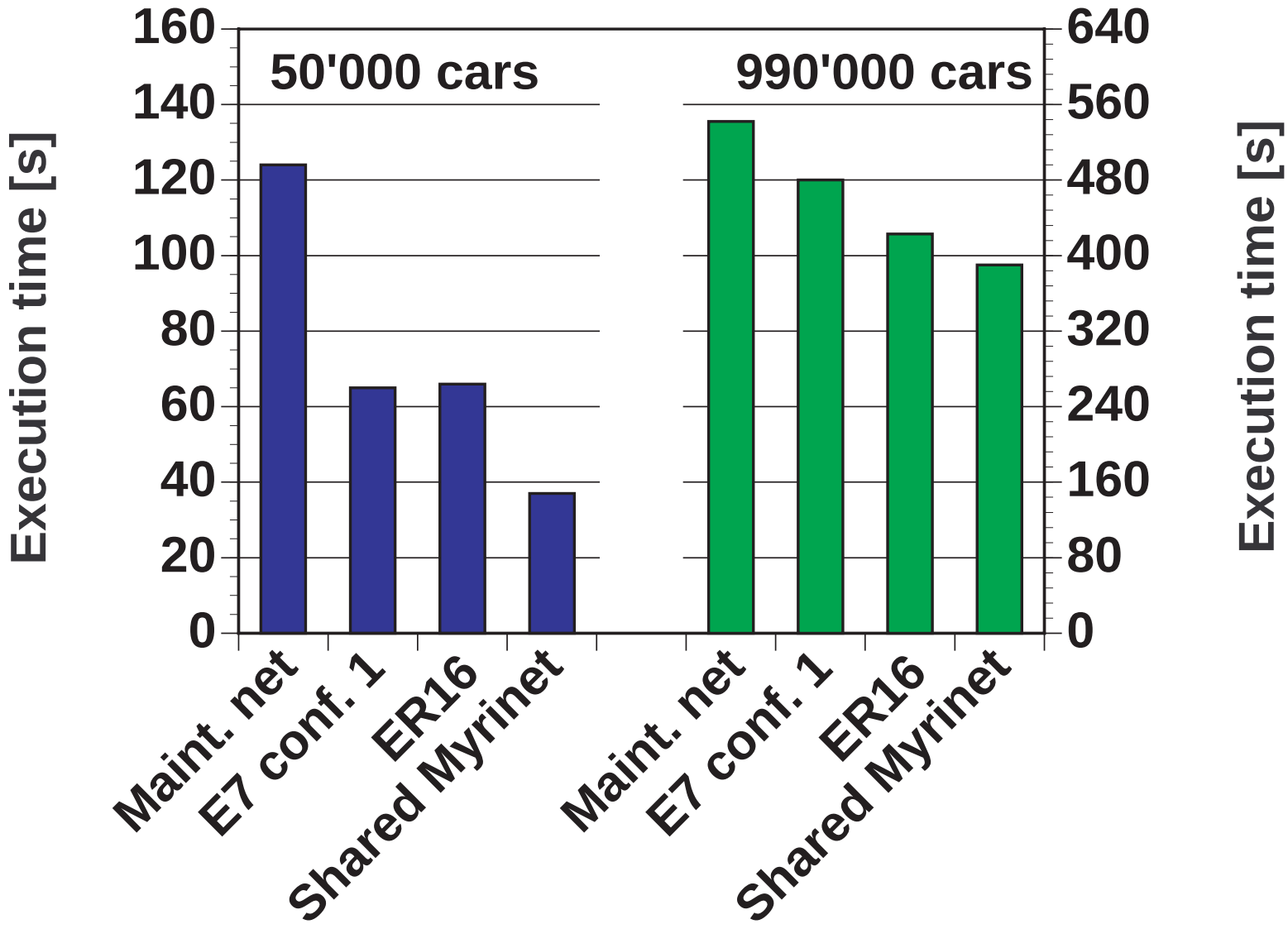


High-Performance Linpack (HPL)

Popular benchmark for supercomputers and clusters.

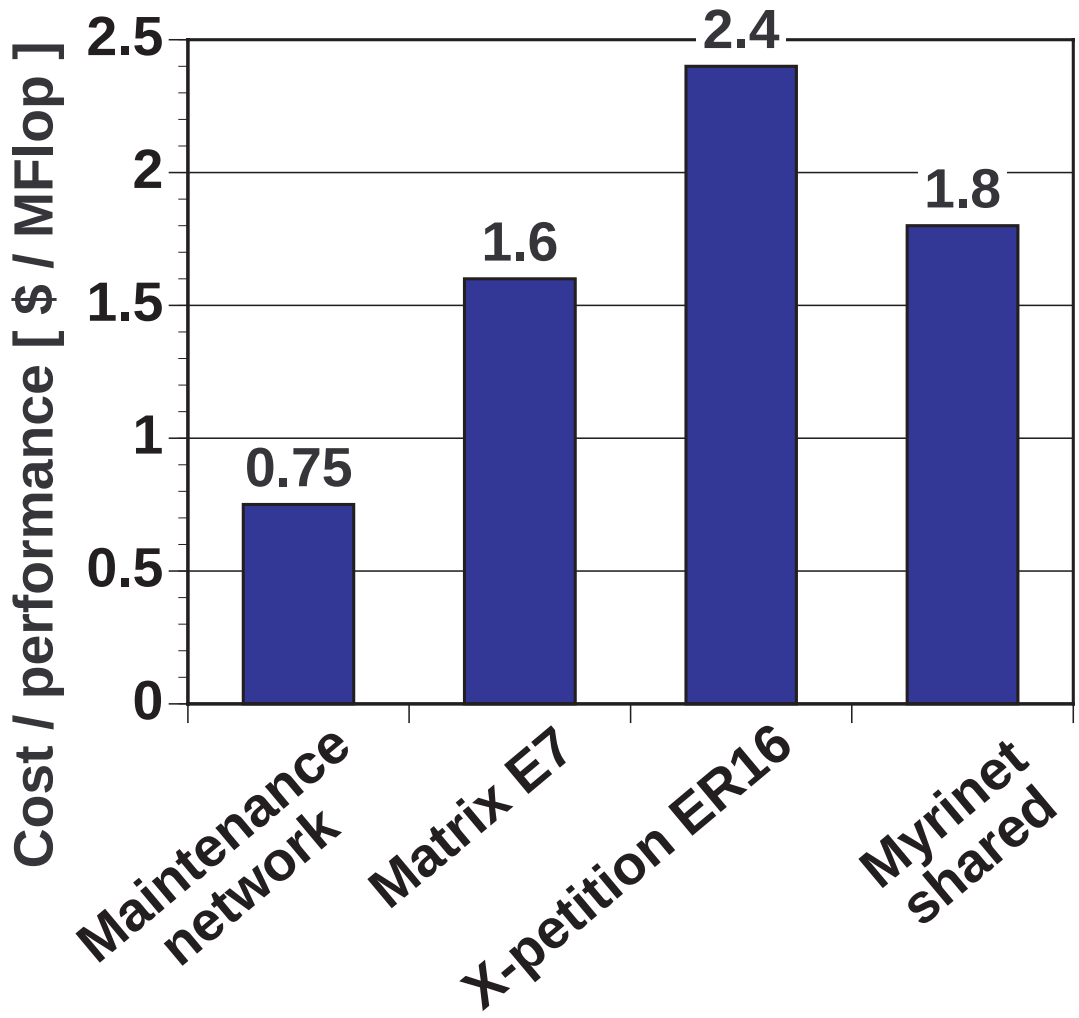


QTPlan Large Scale Traffic Simulation



Cost / Performance Ratios

Price / MFlop for HPL on different networks:



Conclusions

Cluster networking requirements are **different** from mainstream networking with LANs. **Do not trust** the data sheets of commodity network switches!

Cost / performance is best with low-end network.
Fast Ethernet switches with high performance and full bisection bandwidth increase costs unnecessarily.
High-speed LANs are not worthwhile for Beowulfs
➡ use **low-cost LAN** or **high-speed network** (Myrinet).

Performance of different networks seems highly significant at first sight. But **performance impact** on application codes is less **then expected** and depends on the **ratio of communication / computation**.

Questions?

CoPs - Project

Cluster of PCs

http://www.cs.inf.ethz.ch/CoPs/

Xibalba Cluster:

http://www.xibalba.inf.ethz.ch/