Crystal Ball Panel

Can academic ideas and new concepts still make a difference in high performance computing?

Prof. Thomas M. Stricker

Institute for Computer Systems
Swiss Institute of Technology
ETH Zürich

About our projects: www.cs.inf.ethz.ch/CoPs

Eidgenössische Technische Hochschule Zürich

March 6, 2003
SOS7 Workshop

My own experience in HPC

Golden Age of High Performance Computing (1985-1995)

A single research idea could make or break a new parallel machine.

- Systolic architectures - SIMD - Fat Trees - KSR - DASH.
- SUNMOS, OSF1, Software DSMs.

Dark Age of High Performance Computing (1995-2005)

- Everything goes commodity. Working with old broken hardware. Commercial constraint dominate over new ideas.
- Bad Time for Academics - Bad Time for HPC Companies - Great time for Customers in National Research Labs.

New Age of High Performance Computing (2005-)

- Single ideas will make a difference again.
- Integrated hardware - software co-designs will ultimately win.

Still doing research in dark ages...

Architecture of the Xibalba Cluster

- A multi-functional / multiboot cluster with all commodity parts. (2001: 128 Dual Xeons, Full Bisection Ethernet, part Myrinet)
- PXE, Intel Bios, Network, Maintenance Network, Cost/Performance were issues - not so sure how much that is research?

Increasing the Software Efficiency for Gigabit Ethernet Drivers

- The commodity Ethernet adapter has not changed for 15 years!
- A whole new game of system software tricks were needed to make drivers, efficient with zero copy architectures.

Managing Software Installations in Clusters and Desktop Grids

- Computers in Student Computing Rooms and on Desks.
- Fast cloning - partition cast - Dolly models.

The Xibalba Cluster
Department of Computer Science

Prof. G. Alonso (Cluster Programming)
Prof. K. Nagel (Traffic Simulation)
Prof. H.J.Schek (Power Databases)
Prof. T. Stricker (System Architect)

Xibalba mean the hell in Aztec culture.

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

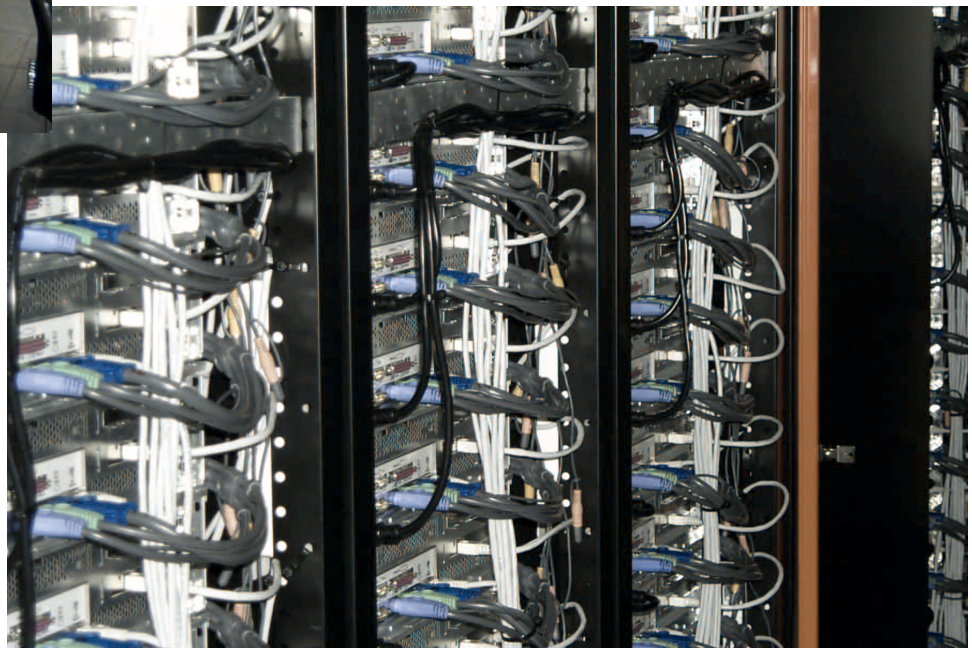
Aztec hell has four rooms

Multiboot / Multipurpose - WinNT/Linux
Fast SCSI disks for database research
128 nodes/Dual Xeon 1GHz/1GB RAM

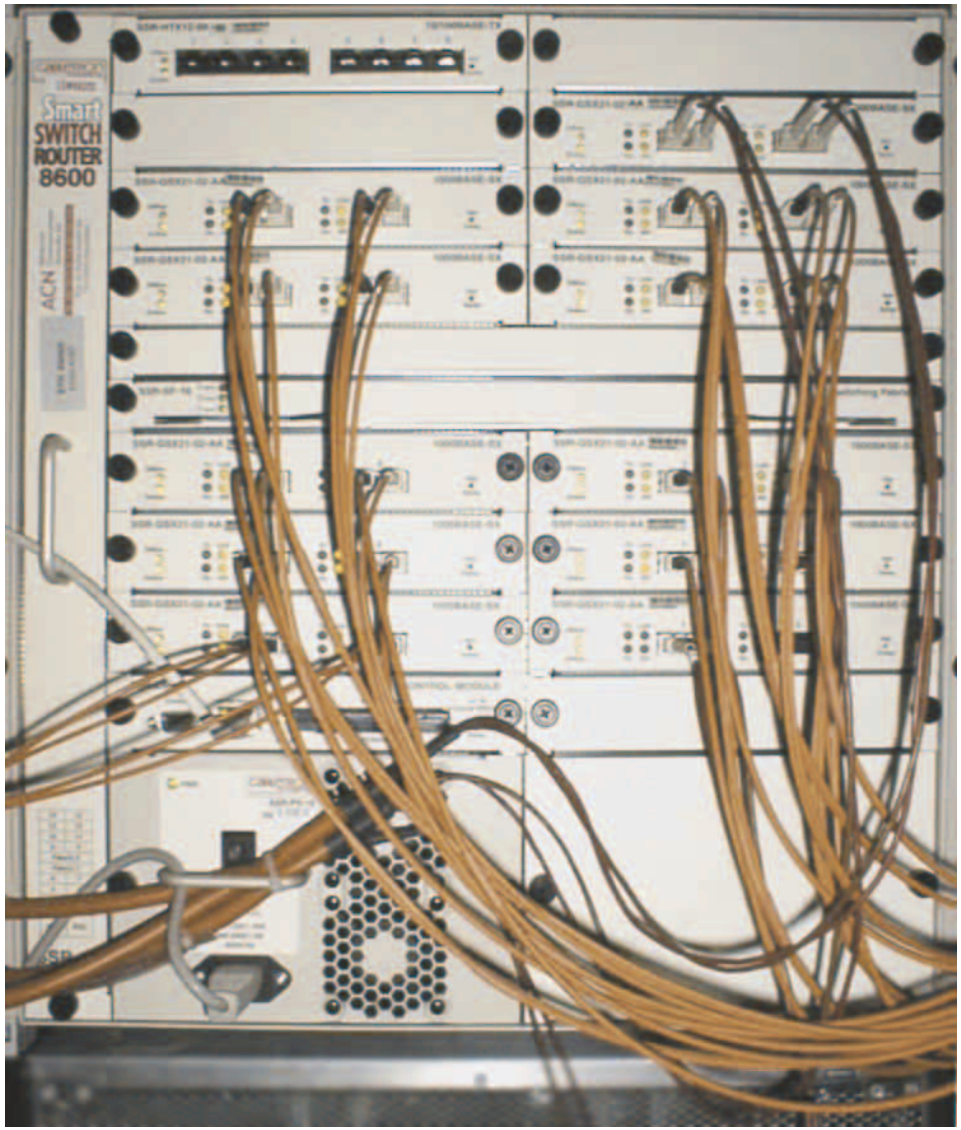
Xibalba has four compartments 32 nodes each

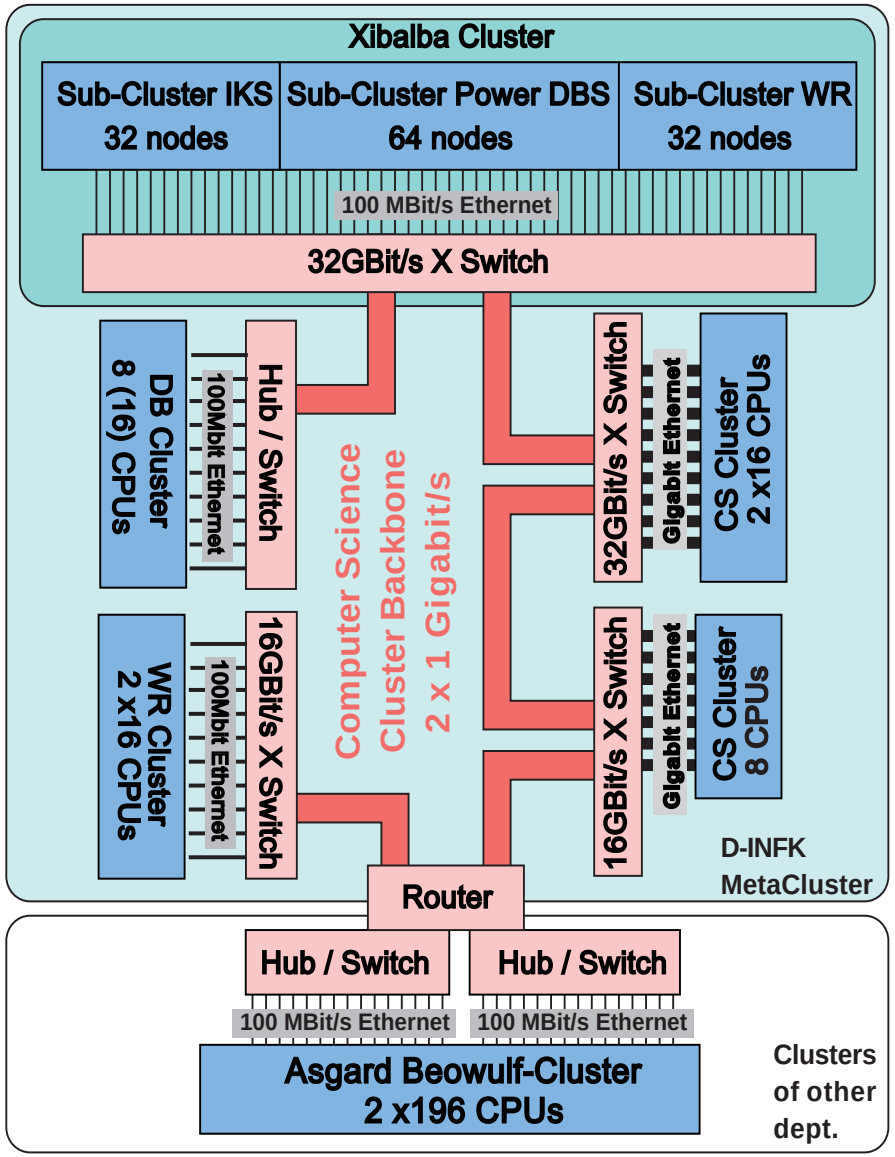
Full bisection single backplane Fast Ethernet. 32 nodes with Myrinet (traffic).

The Xibalba Cluster - Frontside/Backside view.

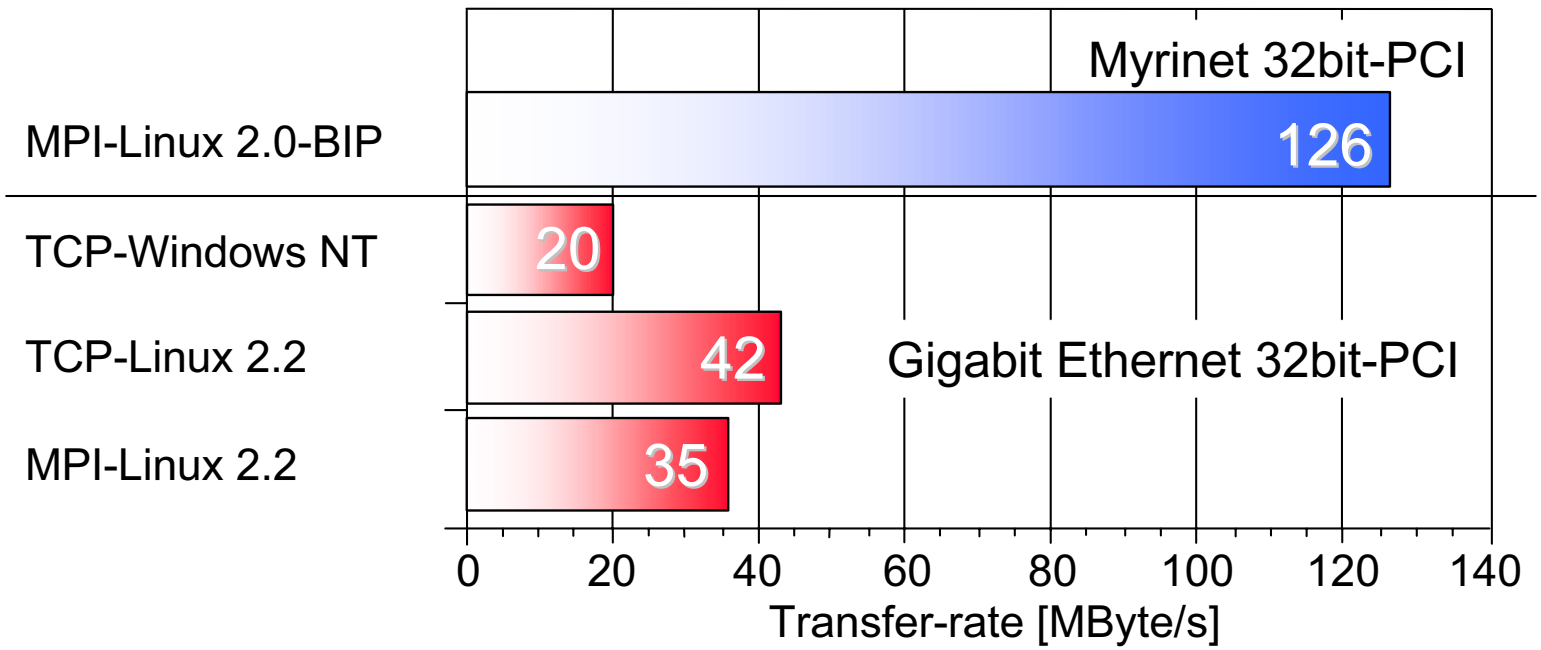


The Cluster - Gigabit Backbone





Communication Software Efficiency (1998-1999)



Message passing over Myrinet reaches a full Gigabit/s (125 MByte/s) with specialized hardware software (BIP MPI)

Regular TCP/IP communication over Gigabit Ethernet does not perform satisfactorily due to operating system overheads

Why is this situation so bad...
Analysis of overheads in TCP/IP protocol stacks

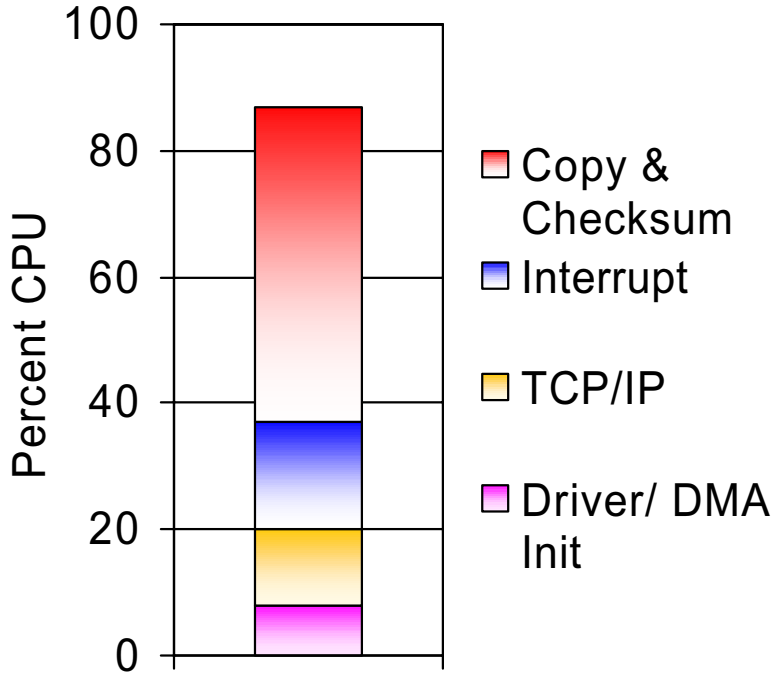
- largest contributing factors: **Copies** and **checksumming** against transmission errors.

  Zero copy solution required!

- **Per packet processing** overhead in particular overhead of interrupt handling is far to big.

  Batch processing and interrupt coalescing is required

PII 400MHz, Linux 2.2



Overheads in TCP/IP over Gigabit Ethernet

TCP/IP: Transmission Control Protocol
Internet Protocol (Cerf/Kahn)

Commercial Gigabit Ethernet Adapters

- Technology of “zero-copy” is **only supported** by “intelligent” network adapters **with co-processors on board** (ATM, Silicon TCP).
- The off-the-shelf Ethernet Adapters are **too simple** and **too cheap** for zero copy:

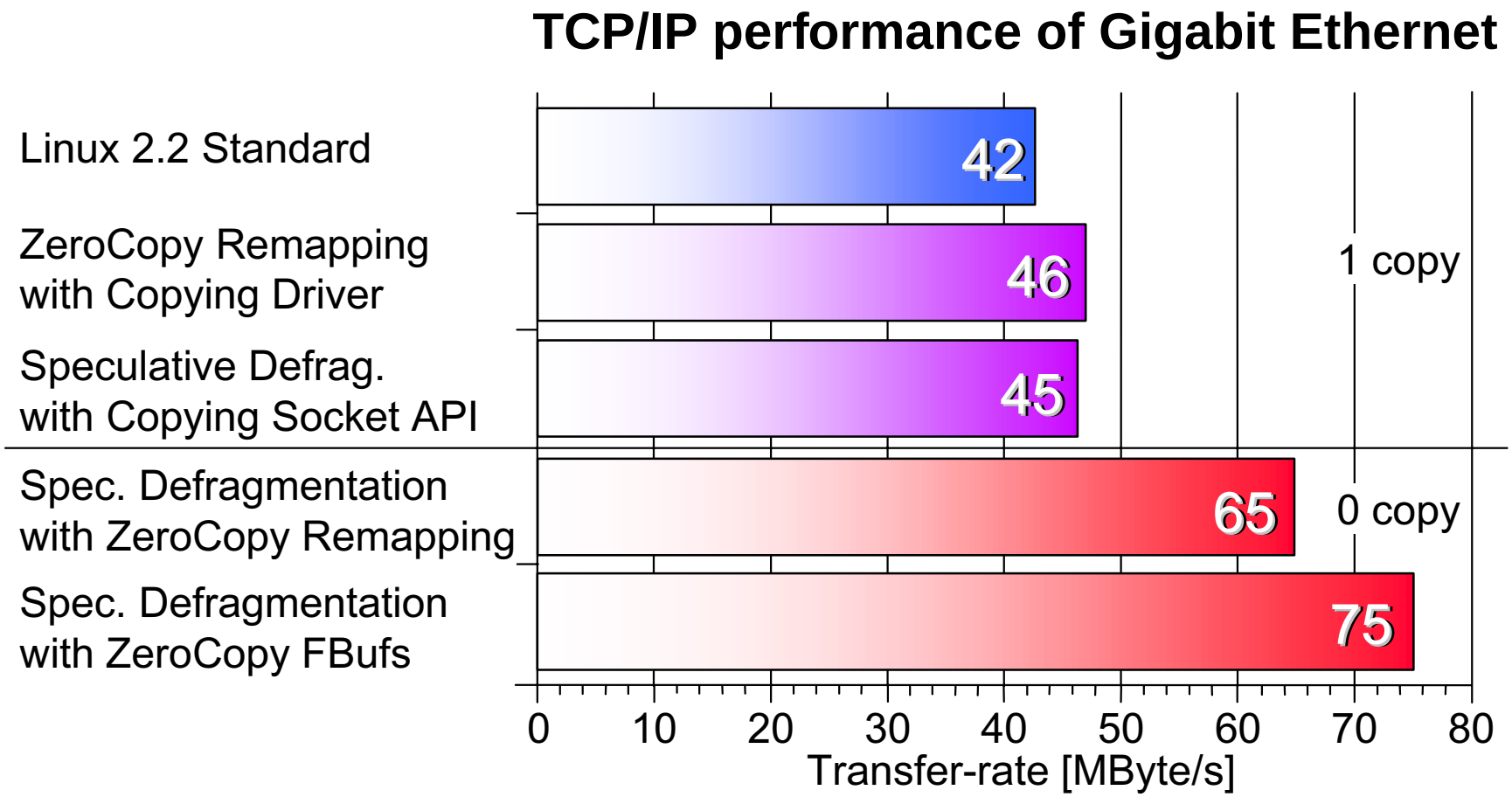
A **deterministic solution** of “zero-copy” TCP/IP with proper de-fragmentation seems **impossible** with this hardware!

New Ideas are required:

- Why go for a **general purpose** solution**?** Why not focus on the **common case** and handle the **special case** in software?

This is the basic idea of a speculative implementation.

Performance gain with successful speculation

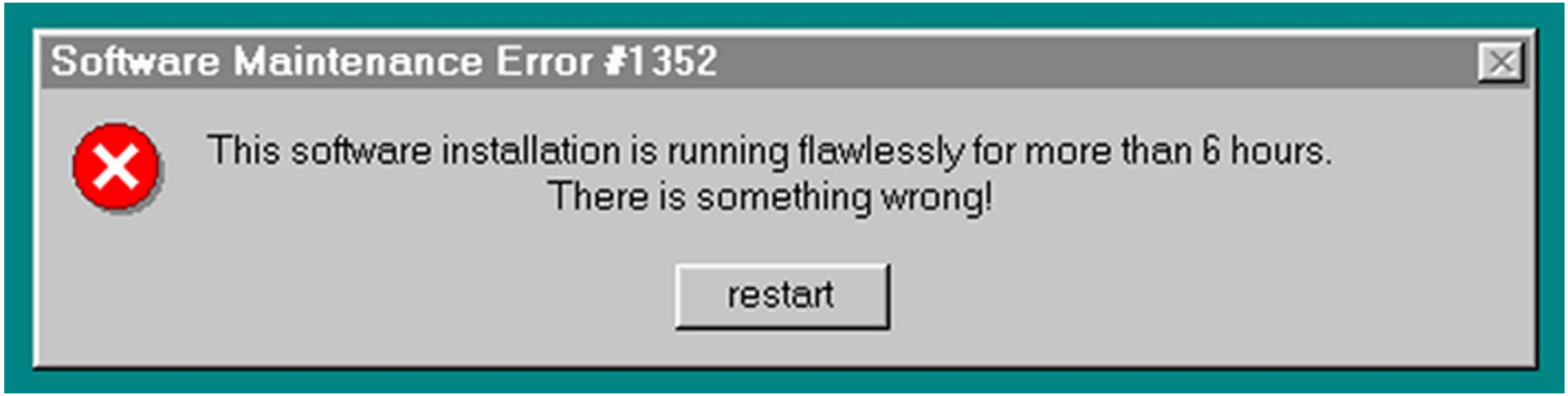


The performance gain is about 80% in throughput

The Ethernet/Commodity Roller-Coaster

- **6 out of 10MBit/s** on Ethernet in **1986**.
  Special purpose protocols based on speculation
- **90 out of 100Mbit/s** on Ethernet in **1998**.
  Problem solved with better bus and memory systems.
- **600 out 1GBit/s** on Ethernet in **1998**.
  General purpose speculative defragmentation.
- **900 out 1GBit/s** on Ethernet in **2001**.
  Problem solved with better bus and memory systems.
- **? out of 10GBit/s** on Ethernet in **2004**.
  Problem will be investigated again?

Common problem of software installations



Maintenance of software installations is difficult:

- Different operating systems/application sets in PC Clusters
- Temporary installations for tests, experiments, courses
- Systematic juvenation of software to combat software rots and prevent system crashes.

manual install: days, network install: hours, cloning: minutes.

Analytic model and optimization of a system for software distribution by cloning

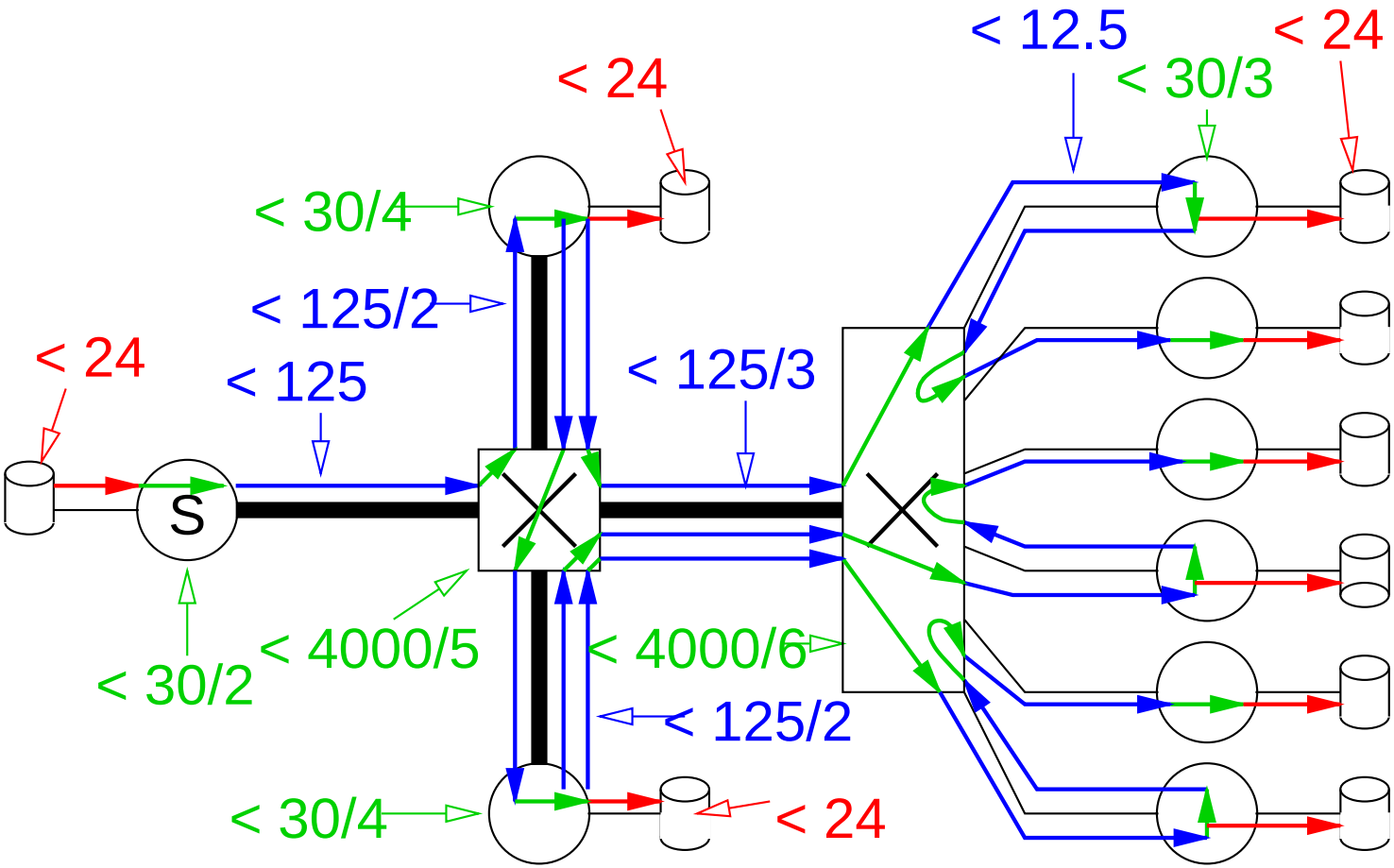
given parameters:

- **Physical network topology** (connectivity of the wires)
- **Resource limitations** (maximal throughput of the links or limiting factors in the switching nodes)

parameters to be found:

- Best **logical network topology** for data distribution
- Best **embedding** of the local network into the physical network?
- **Performance limitation** (throughput) for the distribution of large data sets (e.g. partition casts)

Example including performance figures



Cloning Tool Dolly

Open Source Distribution

New Age of High Performance Computing

I see **four areas** of academic research:

- **Complexity, Correctness** and **Efficiency** of the HPC software (OS, Runtime, Libraries) has to be addressed.
- **Hardware Software Codesign** for **Network Interfaces**, FPGA and other techniques will finally allow new kinds of network interfaces. Communication will again become a first order activity supported by the instruction set.
- **Compilers** will still **make or break** a machines. If we stop pushing the limits - we will not even have the people to write existing compiler. Same holds for **Operating Systems**.
- **Comprehensive management** of compute-, storage- & communication **resources** will be a topic (this is "**Grid**" **research,** but using a lot of locality instead of going needlessly global).