AstroWiSE presents:

TERAPIX hardware: The next generation

Emmanuel Bertin (IAP/Obs.Paris)

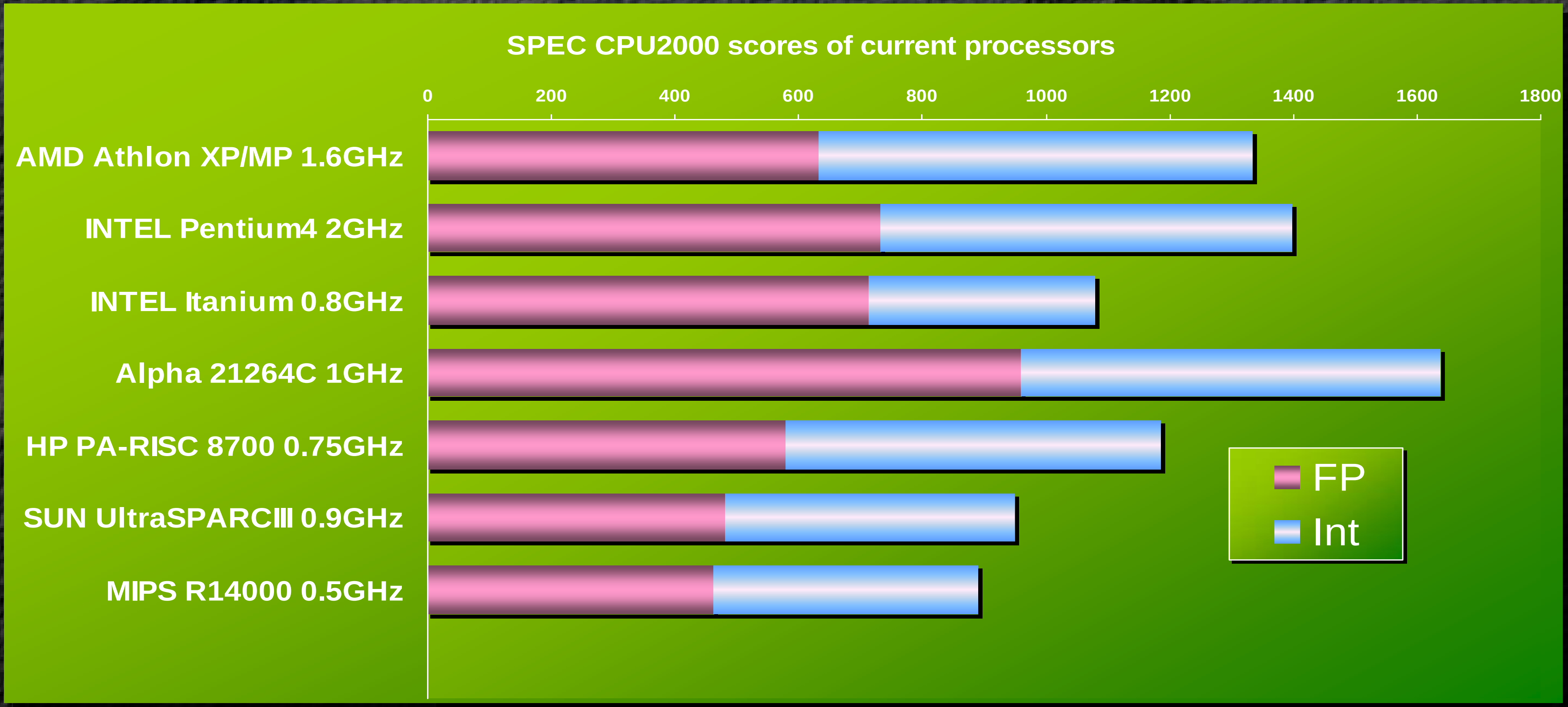
Our Current Hardware

The context

- Demise of Alpha
- Rise of fast, low-cost 32bit PC-Linux servers
  - Popular, well-documented environment for years to come
  - Distributed computing made easy (clusters of PCs)
  - Typical lifespan of a competitive system is 2 years
  - Cheap 64bit machines should appear in 2002 (AMD Hammer)
- Coarse grain parallel processing (as at CFHT)
  - Maximum flexibility
  - ☠ Constraints on network bandwidth different from your typical "Beowulf" PC cluster
    - Very high bandwidth, latency not critical
    - ☞ Evaluation of custom cluster architectures is required
  - ⌛ Our first series of machines has just arrived! (January 2001)

Which CPU?

- Although Alpha processors are still the fastest for scientific computing, cheap competitors have almost filled the gap

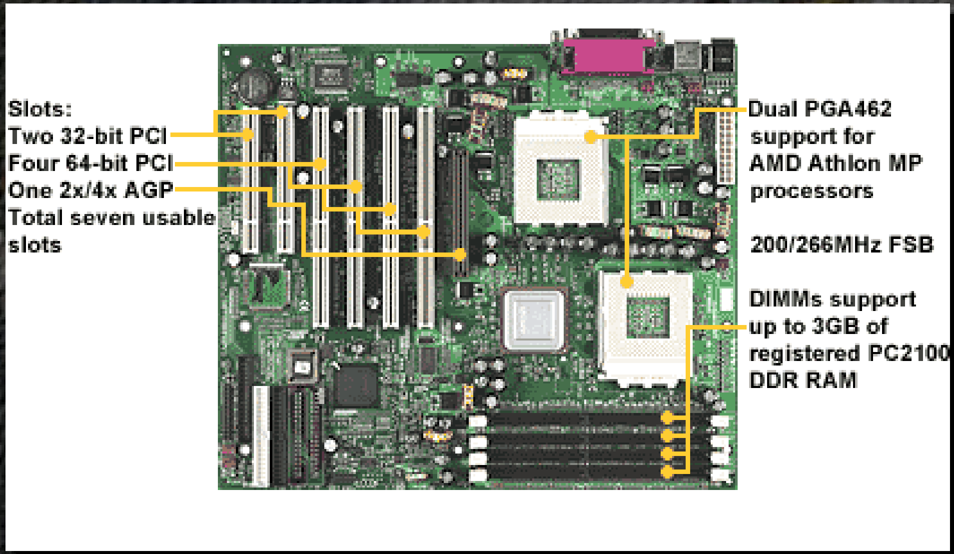


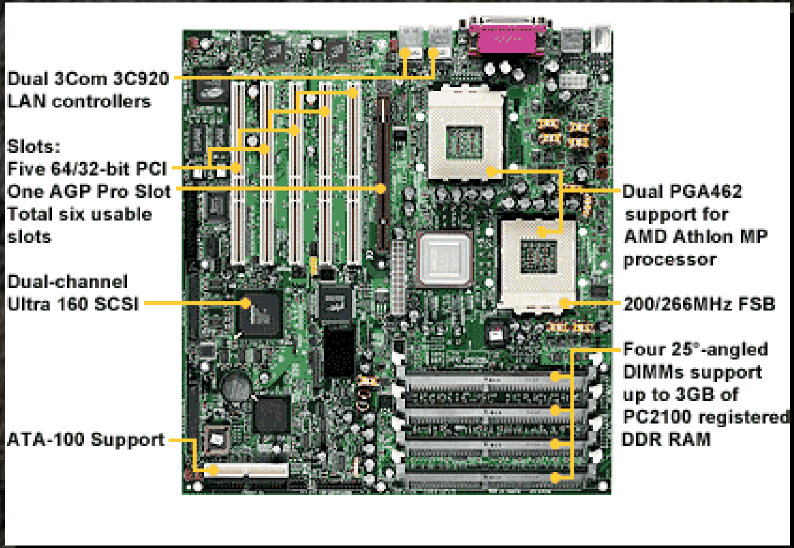
Which CPU? (cont.)

- All the fastest CPUs exhibit almost similar performances (within 20%)
  - ☞ Buy the cheaper ones (AMD Athlons@1.53GHz), but buy many!!
  - ☝ Cheap architectures have some shortcomings:
    - Addressable memory space limited to 3GB in practice with 32bit CPUs
    - Limitations of x86 motherboards:
      - Slower PCI bus (32bit@33MHz = 130MB/s)
      - Less IRQs and DMA channels available
    - ☞ Do not neglect motherboard performance
    - ☞ Go for bi-processors
      - ☞ More efficient in a distributed-computing environment, even for mono-processing (handling of system tasks e.g. I/O software layers)

Which CPUs? (cont.)

- Current motherboards with AMD760MP chipset: Tyan Tiger MP and Thunder K7
  - Stable but modest performance
- Faster motherboards based on the new AMD760MPX chipset now available from Abit and MSI





Optimizing parallel processing

- Amdahl's law: $$T(n) = T_s + \frac{T_p}{n}$$
  - The efficiency of parallel processing is limited by sequential tasks
    - Communication (latency, data throughput) between machines
      - Can be minimized with very coarse-grain parallelism and by limitating pixel data transfers
    - Synchronization of machines (MUTEX)
      - Can be minimized by working on independent tasks/fields/channels
    - Reading/writing data to a common file-server
      - Large transfer rate (high bandwidth) required if one wants to be able to initiate the processing rapidly
        - ☞ Gigabit (cheap) or Fiber Channel (expensive) link

How many machines?

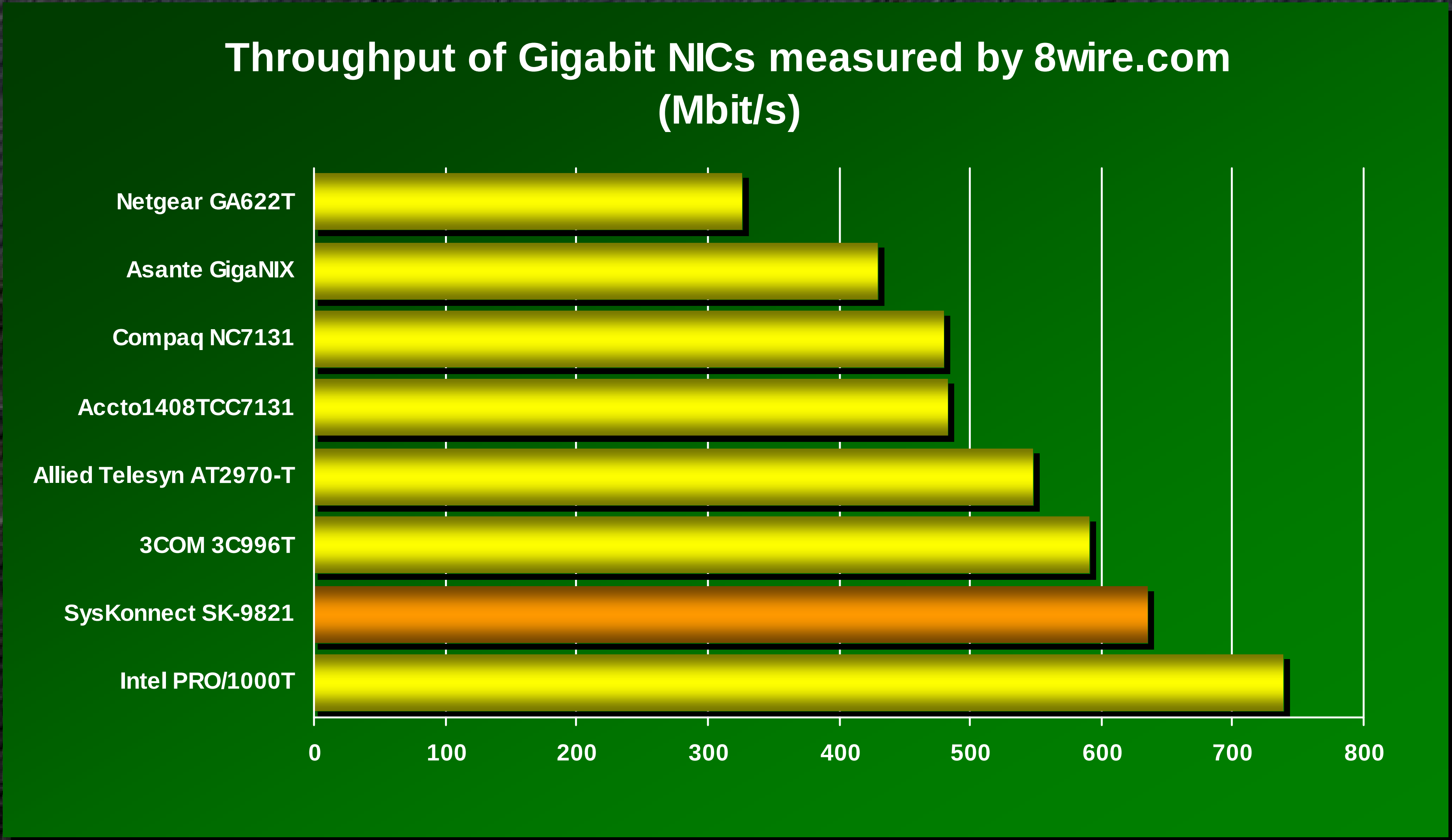
- Not much gain in speed above a number of machines $n_{\mathrm{max}} = t_p/t_s$
  - The slowest tasks (resampling) run at about **250kpix/s**, that is $\approx$ **4MB/s** (including weight-maps and reading+writing)
  - ☞ Hence if one manages to optimize the sharing of server bandwidth, assuming a sustained **80MB/s** total in full duplex (Gigabit+PCI buses), one gets a limit in the number of machines of $\boldsymbol{n_{\mathrm{max}} \approx 20}$
  - But:
    - Reading and writing to the server occurs in bursts, because of synchronization constraints in the pipeline
    - The cluster might be used for faster tasks than resampling
    - One may get an "internal speed-up" in using both processors at once
  - ☞ The practical n_{max} is probably closer to something like **8** machines or even less

Working in parallel: SWarp

Connecting the machines

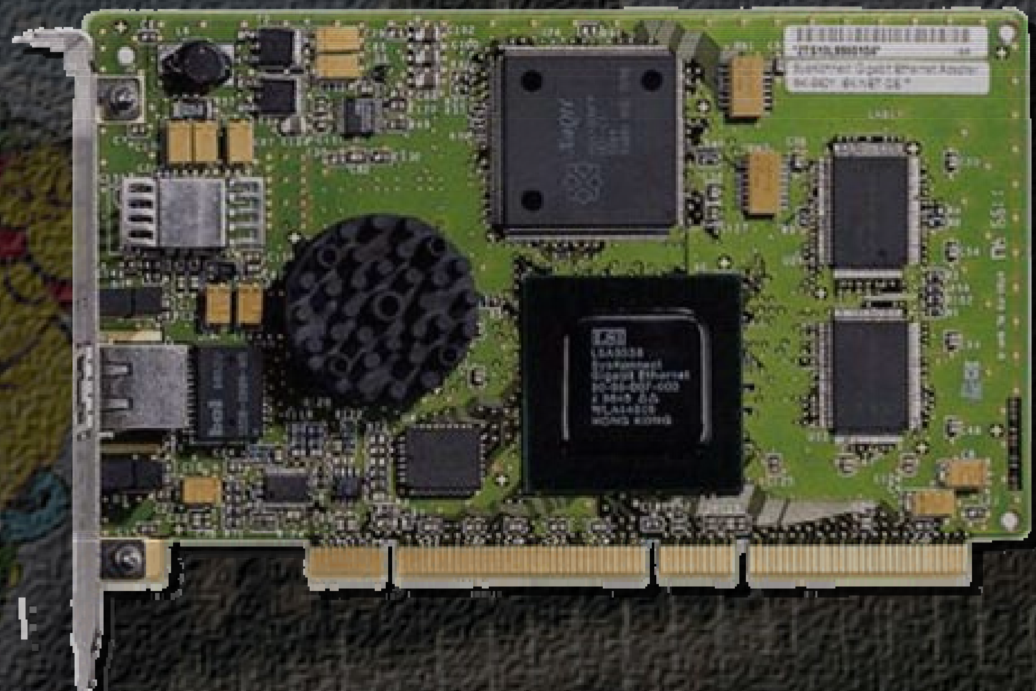
- Adopt TCP/IP protocol (portability, simplicity)
- The 12MB/s bandwidth offered by Fast Ethernet is too slow when it comes to transfer gigabytes of data between machines
- Faster technologies (except multiple Fast Ethernet) include GigabitEthernet, Myrinet, SCI, IEE1394, USB2.0
  - Gigabit Ethernet: bandwidth = 100MB/s, typical latency = 100μs
  - Myrinet: bandwidth = 100+MB/s, typical latency = 10μs
  - SCI: bandwidth = 800+MB/s, typical latency = 5μs
  - IEEE1394a: bandwidth = 50MB/s, typical latency = 125μs (?)
  - USB2.0: bandwidth = 60MB/s, typical latency = 120μs
- For the parallel image processing of TERAPIX, latency is not critical (few transfers), but bandwidth is (lots of bytes at each transfer)
  - TCP layers wastes latency anyway!
- Go for Gigabit Ethernet!
  - The price of 1000base-T Gigabit Ethernet NICs has fallen considerably in 2001 (from >1000 € to less than 140 €)
  - …but Gigabit switches are still fairly expensive (>1000 €)

Which Gigabit Ethernet adapter?

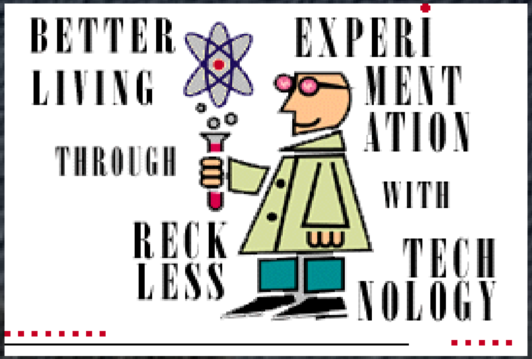


The SysKonnect SK-9821

- ☺ 200 €
- ☺ PCI 32/64bit, 33/66MHz
- ☺ Efficient Linux driver included in kernels 2.2 and above
- ☺ Excellent technical support for user
- ☹ Gigabit only
- ☹ Bulky radiator runs pretty hot
- ☹ “Old product”, the 3C1000-T might be a better bargain



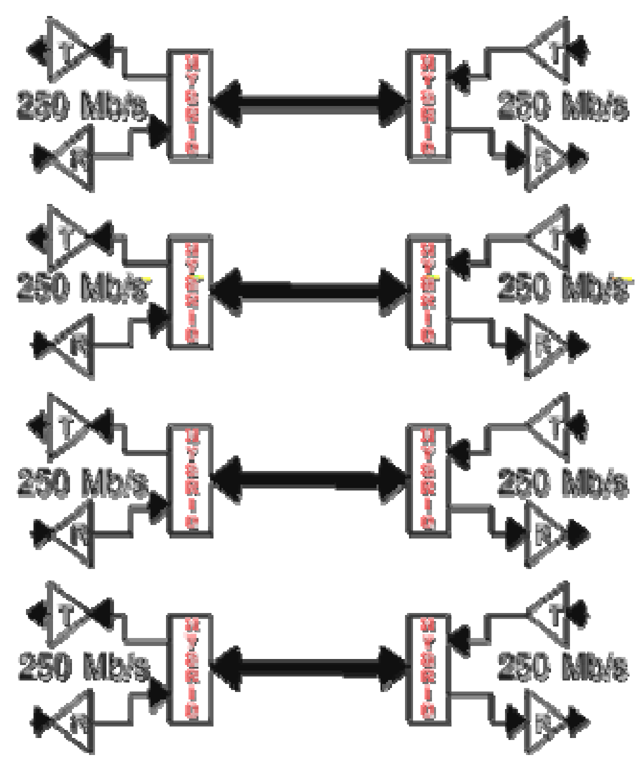
Getting rid of the hub



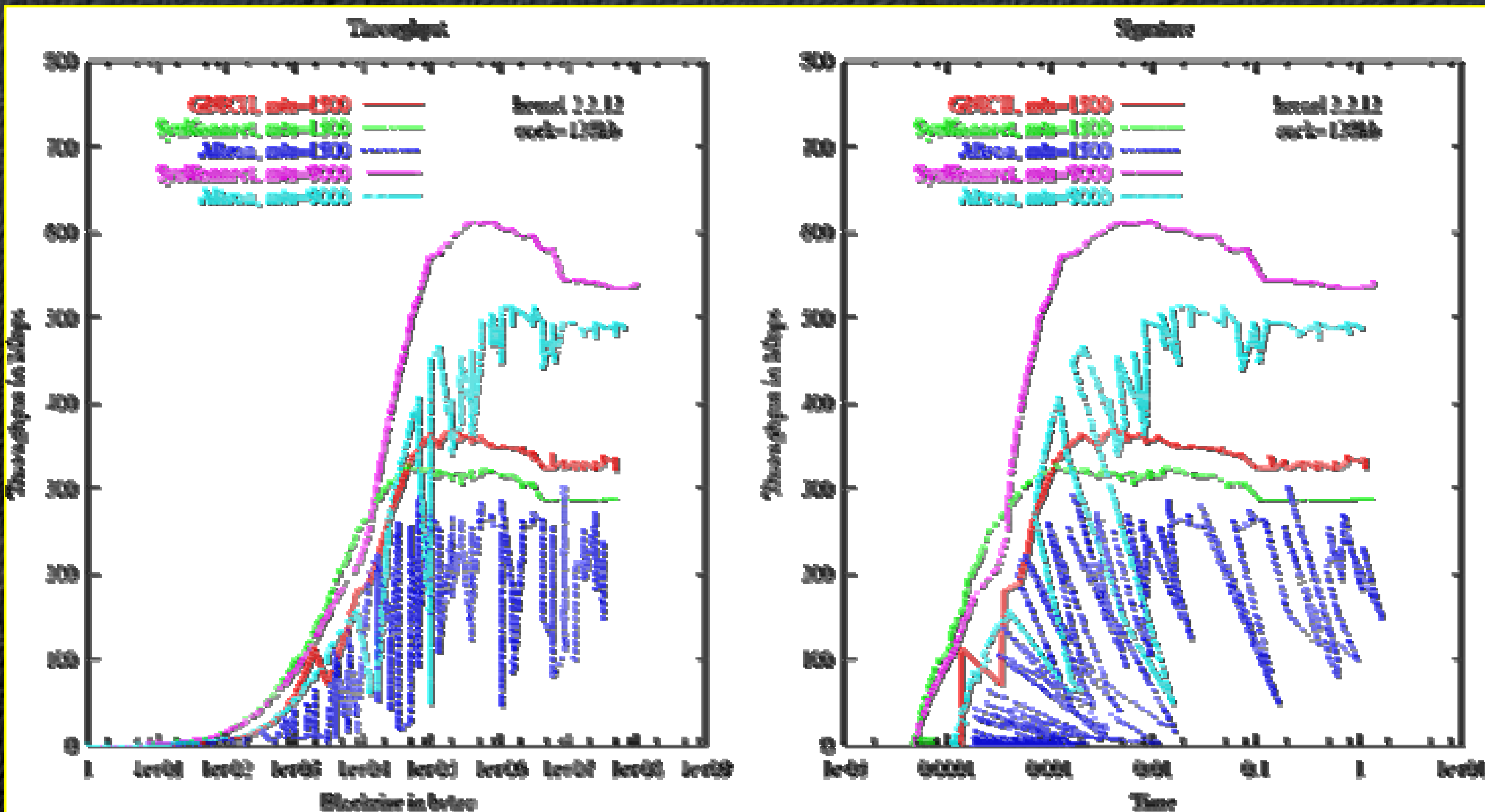
- A gigabit hub is as expensive as a PC equipped with a NIC!
- The connection to the file server has to be shared by the computing units
- Why not use direct Gigabit Ethernet "cross-links" between the server and the clients?
  - 1 NIC on the client side
  - 1NIC per client on the server side
    - Fairly common with Fast Ethernet NICs
    - Caution: IRQ sharing, PCI slots, power draw-out
    - Experimental stuff! If it does not work, we will buy a switch

Testing Gigabit cross-link connections

- 2 SysKonnect SK-9821 where used for the tests
- Gigabit cross-links are not crossed!
- Without tuning, a throughput of about **30MB/s** is reached (the ping is 0.1ms)
- After tuning (jumbo frames and TCP buffers increased), transfer speed is extremely dependent on the chipset.
  - We measure the following PCI bus throughputs:
    - VIA KT266: **56MB/s**
    - VIA694XDP : **85MB/s**
    - AMD761: **125MB/s**
  - Using the 2 last machines, we measure **63MB/s** sustained (ncftp+RAM disk, or IPerf), with **20%** of CPU usage
- The 64bit PCI bus of bi-Athlon motherboards helps a lot (>205MB/s)



Tuning for better Gigabit performance



Ong & Farrell 2000

Local disk storage

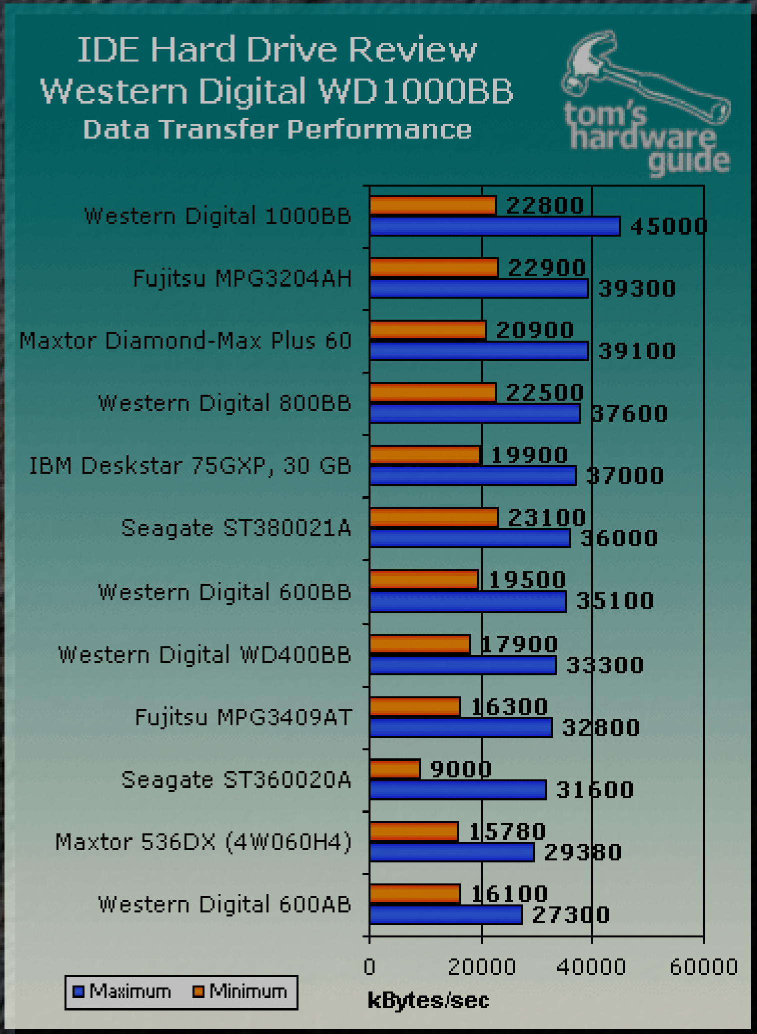
- On the computing units ("clients"): fast, local disk storage is required for data processing
  - Load raw/reduced images from the server only <u>once</u>
  - Scratch disk
  - Two sets of disks are needed: to read and to write from
  - Speed (transfer rate) is more important than reliability
  - ☞ Go for 2 RAID0 arrays
- Hard drive failure
  - At IAP (DeNIS, Magique, TERAPIX and 100 PCs): <5% per year
  - Downtime can be tolerated (No permanent storage on computing units)
- RAID0 controllers
  - For RAID0, sophisticated PCI RAID controllers are not required
  - Any bunch of disks can be operated in software RAID0 mode under Linux
  - Cheap (<200€) RAID controllers for 4 UDMA100 drives: Adaptec 1200A, HotRod 100 (Highpoint 370), Promise FastTrak 100:
    - The Fastrak 100 is the fastest (**80MB/s**). There is now support for Linux.
    - 4 disks per controller: 2 PCI RAID controllers are needed, for a total of 8 disks

Local disk storage (cont.)

- On the file server, securized disk storage is required
- RAID5 array:
  - Software RAID5 is very slow (<**10MB/s**) and resource-consuming under Linux
  - 3Ware Escalade 7850 RAID0/1/10/5/JBOD card:
    - ☺Hardware XOR: **50+MB/s** in RAID5 with 4% CPU usage! (measured in Windows2000)
    - ☺8 IDE master channels
    - ☺PCI 64bit, 33MHz
    - ☺Supported in Linux kernel 2.2 and above (…)
    - ☹Quite expensive (≈900€)

Which hard drives?

- RAID0 disks:
  - Raw transfer rate is important with 4 disks: 7200 RPM recommended
  - Highest capacity at 7200RPM: Western digital WD1000BB
    - ☺ High capacity: 100GB
    - ☺ Rather cheap: ≲300€
    - ☹ Long-term reliability unknown
- RAID5 disks:
  - Parity computations, dispatching and 8 disks: 5400 RPM is sufficient
  - Highest capacity: Maxtor 540DX
    - ☺ Very high capacity: 120GB
    - ☺ Rather cheap: ≲300€
    - ☹ Long-term reliability unknown



TERAPIX “pipeline” cluster

4 × Computing units:
- ✓ Bi-AthlonMP @1.53GHz
- ✓ 2GB of RAM 266MHz
- ✓ 2×400GB RAID0 arrays
- ✓ Gigabit Interface
- ✓ Fast Ethernet interface

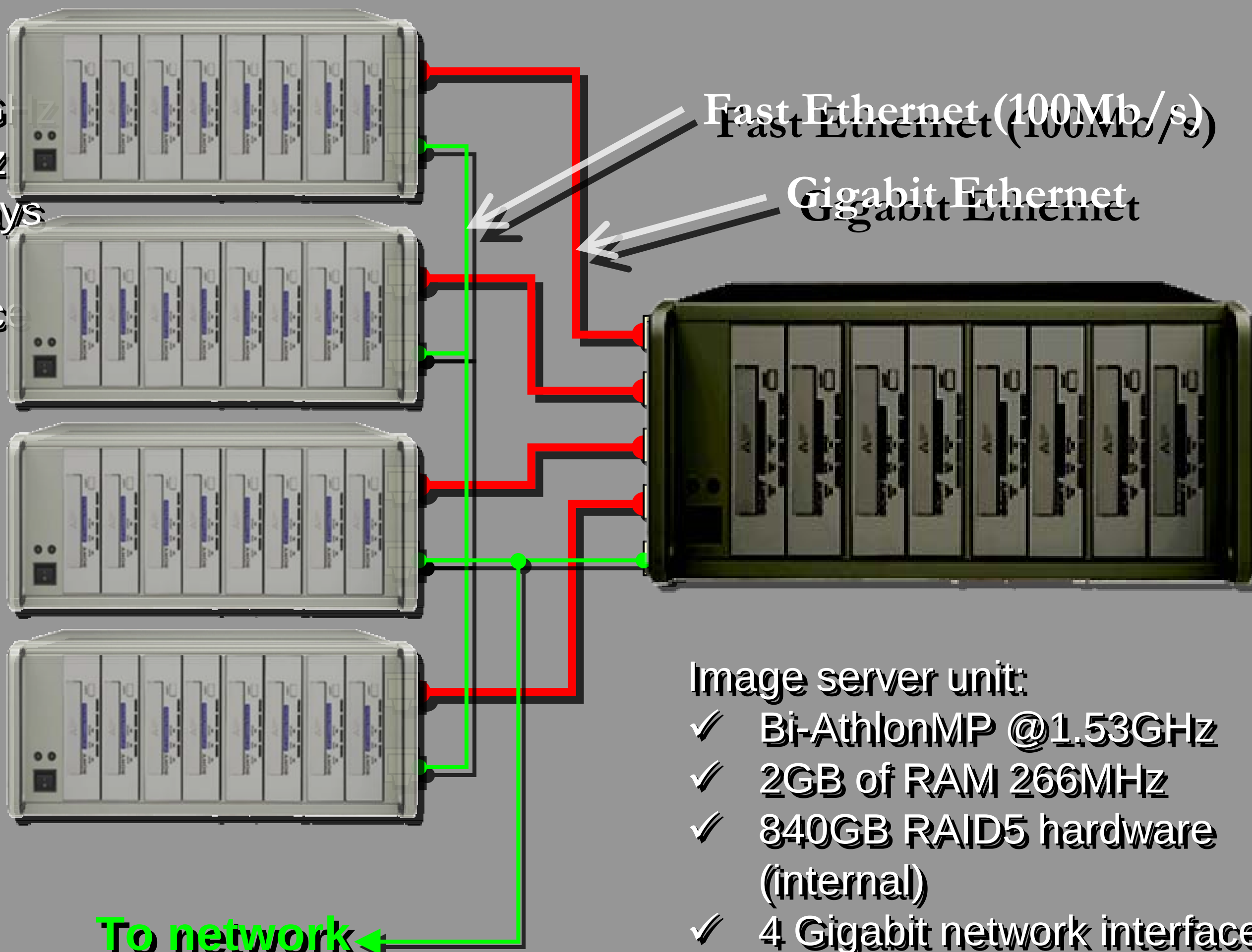


Image server unit:
- ✓ Bi-AthlonMP @1.53GHz
- ✓ 2GB of RAM 266MHz
- ✓ 840GB RAID5 hardware (internal)
- ✓ 4 Gigabit network interfaces
- ✓ Fast-Ethernet interface
- ✓ SCSI Ultra160 interface

Cost

- Computing units (assembled, 1 year warranty): 4 × 6k€
- Server (assembled, 1 year warranty): 7k€
- Rack, Switchbox, cables, 3kVA UPS: 2.5k€
- Total: 34k€ for 10 processors and 4TB of disk storage