Medium Term Issues for the Data Challenges

Available Hardware

Commodity off-the-shelve

Dual processor PCs with INTEL CPUs (PIII ~1GHz, 512MB)

+ Fast Ethernet controller → CPU server
(ATX housing in racks, ~ 2KSFr per box + 0.3 KSFr infrastructure)

+ Gigabit Ethernet controller, EIDE RAID controller
1 TB EIDE disks → Disk server
(4U rack mounted, ~ 11KSFr per box + 1.8 KSFr infrastructure)

+ Gigabit Ethernet controler, SCSI or Fiber channel controller
one or two tape drives → Tape server

Tape infrastructure

STK 10 silos with 55000 cartridges
28 x 9940 drives (10 MB/) used by all experiments

New acquisition in Q3/Q4
→ ~ 20 drives dedicated to the LCG project
higher capacity tapes and higher performance drives

Network infrastructure

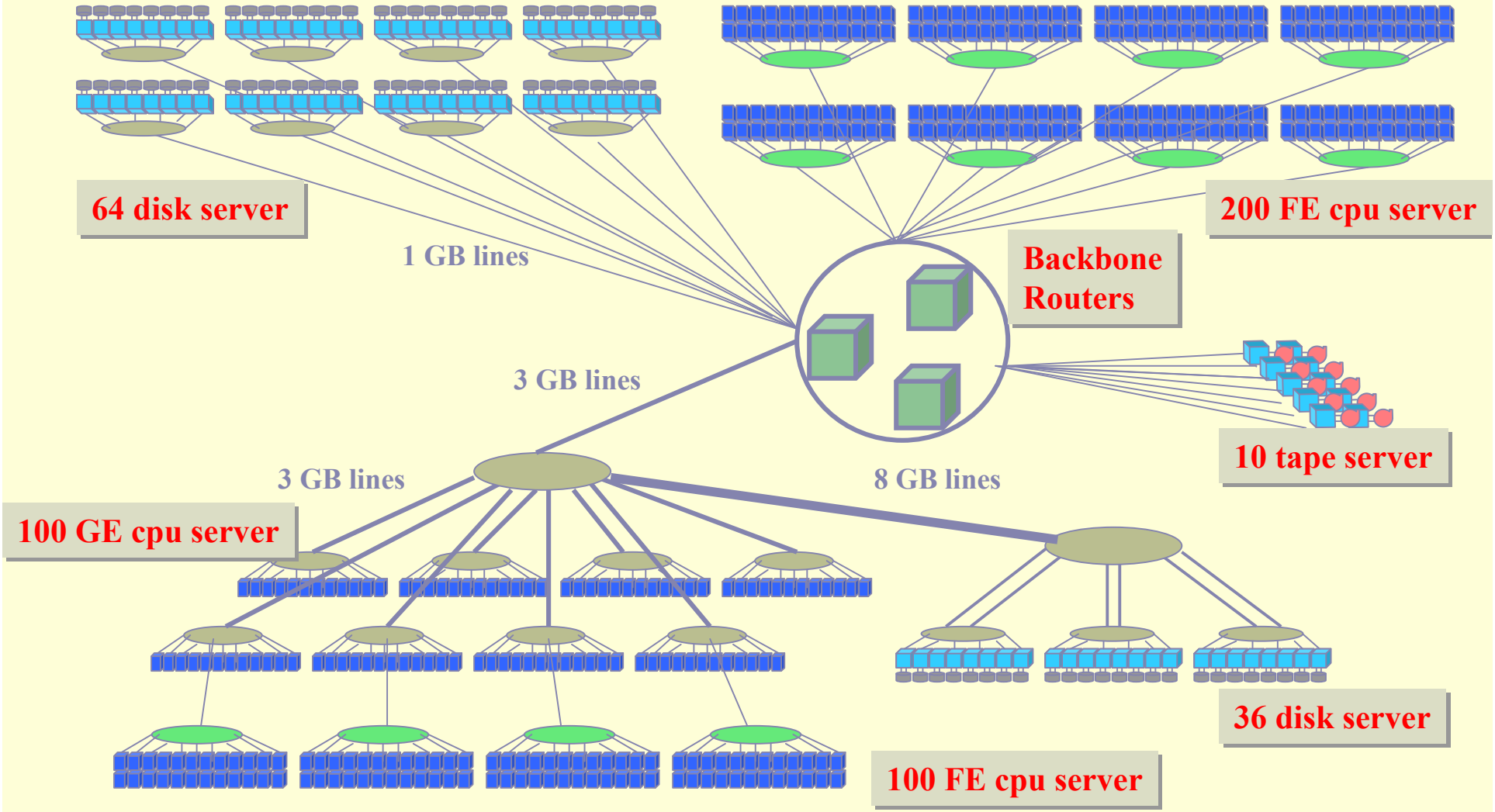
3COM Fast Ethernet and Gigabit Ethernet switches

ENTERASYS high end backbone routers

10 GB routers are currently being tested and will be incorporated in the Testbed soon

LCG Testbed Structure

100 cpu servers on GE, 300 on FE, 100 disk servers on GE (~50TB), 10 tape server on GE



Data Challenge Types

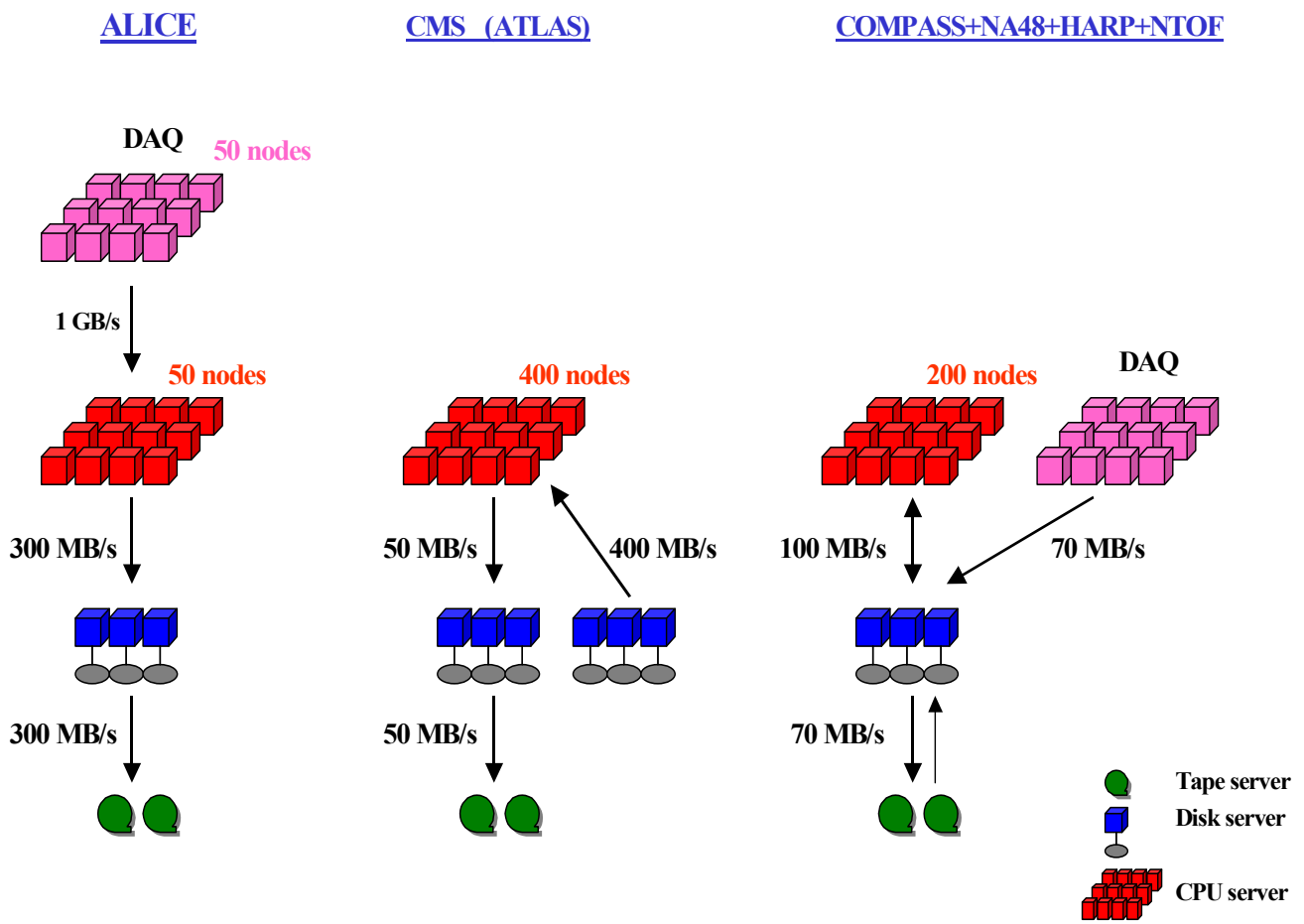
ALICE DAQ tests event building, processing and storage
→Goal for this year 200 MB/s into CASTOR sustained for one week (peak 300 MB/s)

Scalable middleware tests from the DataGrid project

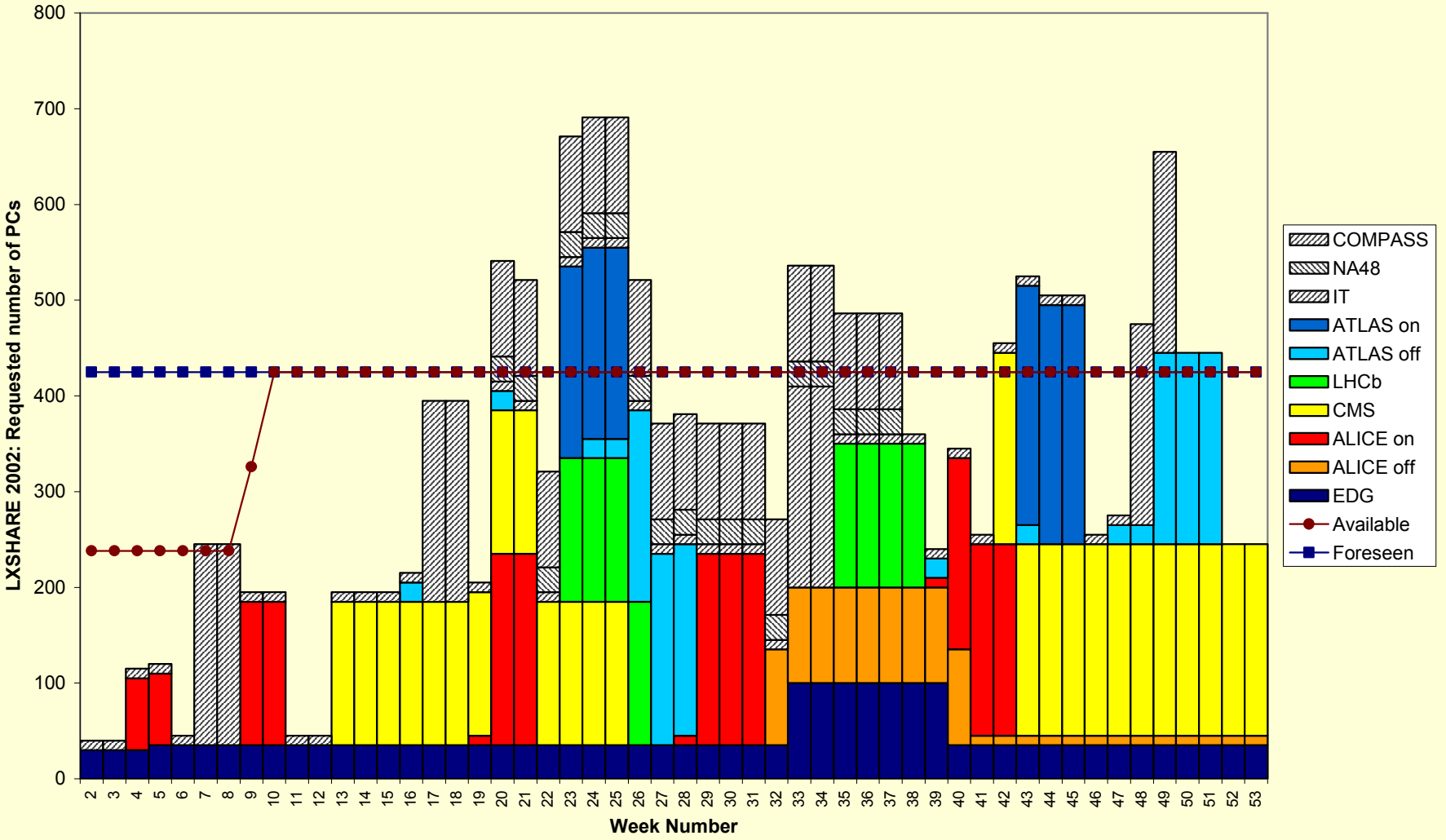
Large scale productions for the physics TDRs (CMS, ATLAS)

Installation, configuration and monitoring of large farms
→Scalability and robustness
(the whole LCG facility will be used by quite a number of experiments with different environment needs
→ reconfigurations

Requirements for different challenges and productions



Draft schedule for the node allocation in 2002 (clear priority guidelines have been approved)



Problems and Solutions (1)

LINUX

Stability and performance has improved considerably during the last two years (MTBF disk server > 200 days, cpu server >100 days scheduled interruptions included)

IO performance to be watched
Kernel 2.2.x → 2.4.x showed improvements by ~ 60%
Kernel variations still to be explained 2.4.x → 2.4.y : 20 – 30 %

Sometimes hard to follow developments and changes

Problems and Solutions (2)

Network

Working well, needed a few firmware upgrades in the beginning, only one major bug in a high end router

(Obviously we are using/stressing the equipment like nobody else does)

Control and management

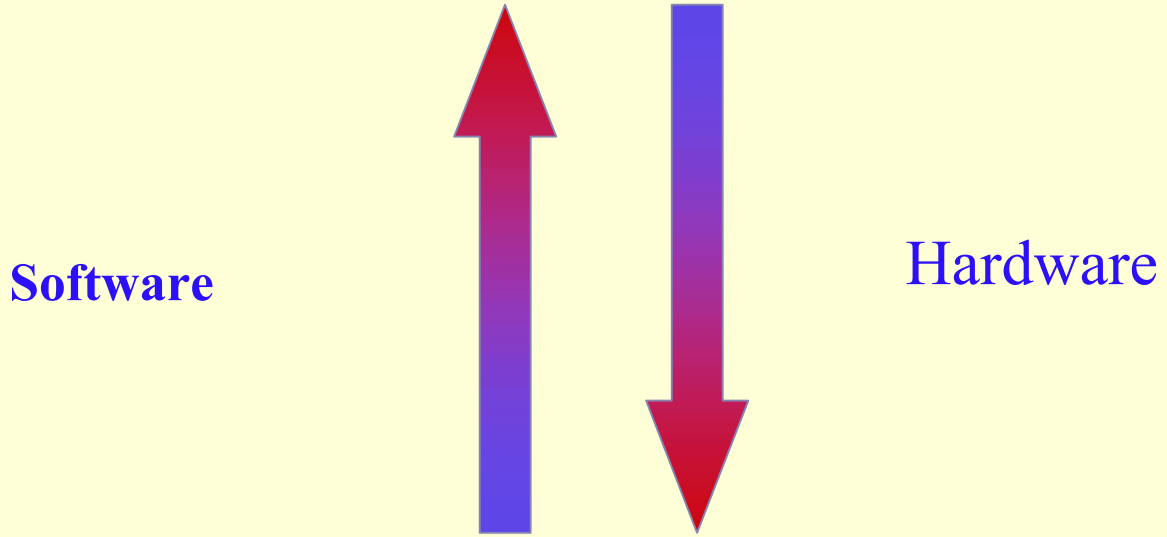
Installation, configuration and monitoring
→Some prototypes used, close collaboration with the DataGrid

Low level fabric infrastructure planned for Q2/Q3
(console, reset, diagnostics)
→ No PC standard

Problems and Solutions (3)

How does the software (middleware, application) cope with Hardware problems ? TCO considerations

Level of redundancy, error recovery, fault tolerance



e.g. Lack of fail safeness in the software has to be compensated by complexity in the hardware

Problems and Solutions (4)

Disks

Problems with certain IBM disk model , high error rate
Finally fixed by firmware upgrade of ~800 disks
Only seen with certain data access patterns
Regular updates of RAID controller firmware

Tapes

CASTOR HSM successful tested, but questions about load balancing and scalability still need investigations on larger scales

General Architecture is based on fully distributed and asynchronous
- only few tape drives 'Impedance' problem when coupling to disk servers,
IO performance of disk server should >> tape drive performance
→ Higher end disk servers, LINUX IO improvements,…
→ introduction of more disk cache levels
→ replace tapes with disks

Tapes versus disks, Some 'naive' calculations (1)

1 PB of tapes with 20 drives
Our current installation is from STK, 9940 drives with 60 GB cassettes, ~ 15 MB/s read/write performance per drive

Costs for silos, drives, servers, tape media, maintenance over 4 years
→ ~ 4.1 SFr/GB (1PB with 0.3 GB/s aggregate throughput)

With the new types of drives announced (STK, IBM, etc.) (Q3/Q4), an estimation would be the following
→ ~ 2.6 SFr/GB (1PB with 0.6 GB/s aggregate throughput)

Tapes versus disks, Some 'naive' calculations (2)

1 PB of disk

1 TB of disk space per server, EIDE disk server
current standard type (120 GB per disk, 10 disks)
→ ~11 SFr/GB (50 GB/s aggregate throughput)

10 TB of disk space per server, ~ 60 disks (160 GB)
Currently not easy with EIDE channels, maybe Firewire or USB 2.0
→ ~5.5 SFr/GB (5 GB/s aggregate throughput)

We assume that there are already quite some CPU servers around (2200 nodes) Each node is upgraded with 3 x 160 GB disks Just need to compensate for 10% CPU performance (== 5 MB/s extra IO per node)
→ ~3.8 SFr/GB (11 GB/s aggregate throughput)

Tapes versus disks, Some 'naive' calculations (3)

Tapes 2.6 – 4.1 SFr/GB → Disks 3.8 – 11 SFr/GB

But need to consider :

Reliability tape versus disks → double disk copies needed

Can the software cope with this kind of large distribution of disk space ??

Influence from the persistency model, data storage Model, HSM system

Other issues

- **When is the correct time to move to IA64 ?**
- **Is SAN really an alternative solution ?**
- **What is the Analysis model ?**
- **Blade systems are interesting, but still very expensive (x4)**
- **Full LHC computing is 6 years away**
  → Paradigm changes ?? (PDA, set-top boxes, 'Xbox', eLiza,….)