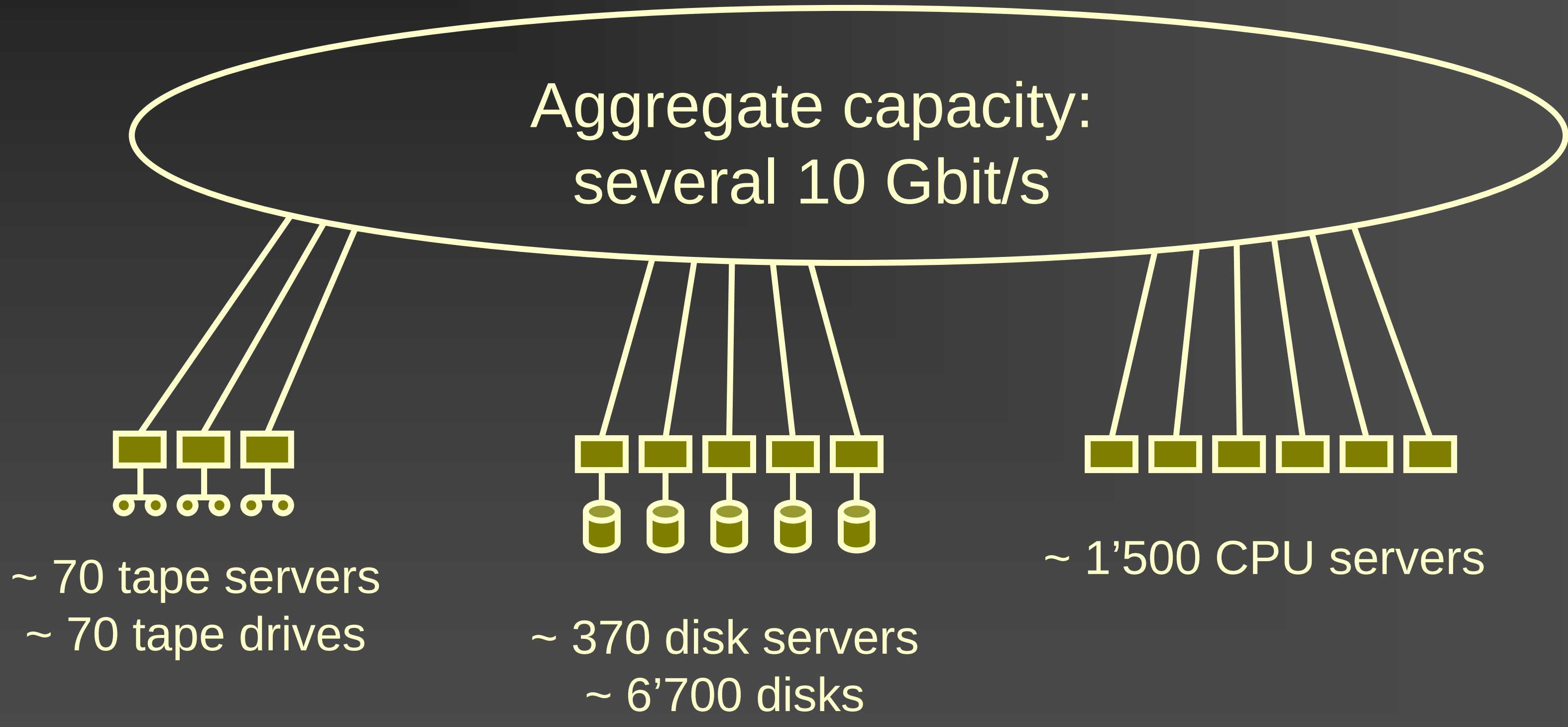
# Disk storage technology for the LHC T0/T1 centre at CERN

Helge Meinhard / CERN-IT

CHEP Interlaken / 2004-09-29

Presenting work of IT-FIO and IT-ADC

# Current model



# Current disk storage: HW

- 370 disk servers: Storage in a box
  - Dual Intel PIII or Xeon
  - 1 or 2 GB of memory
  - Gigabit Ethernet
  - Hardware RAID controller (PCI cards)
  - 12…26 EIDE disks in hot-swap trays
  - Standard CERN Linux, CERN tools for installation, configuration and monitoring (ELFms)
- 6’700 disks in total
  - 544 TB before RAID-ing

# Current disk storage: HW

July 2003 (tender),
January 2004 (delivery):

- 8U rackmount
- 3 RAID cards
- 22 data disks @120GB
- 2 system disks @80GB

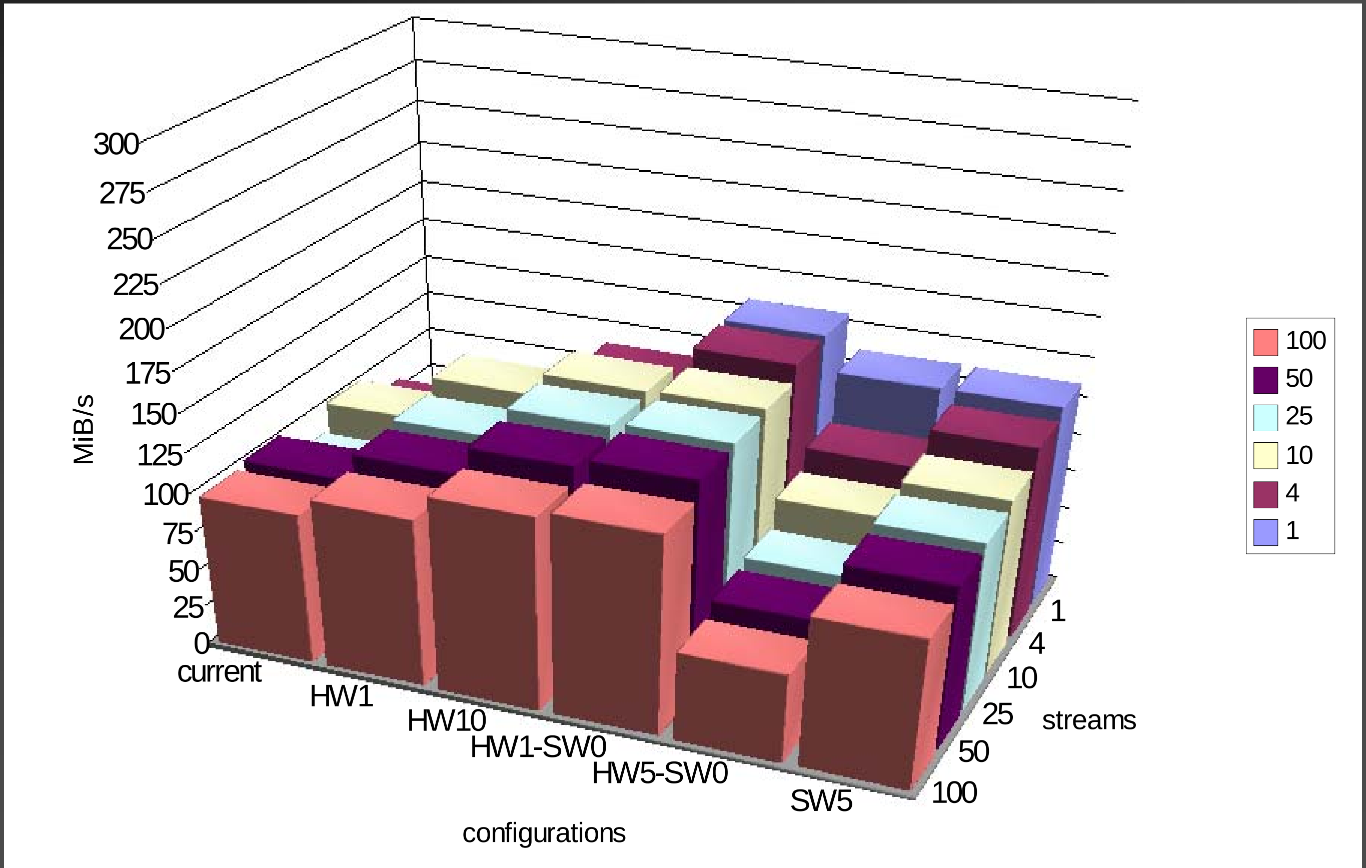
# RAID options and file systems

- Until spring 2004:
  - RAID 1 (mirroring) over two disks
  - One ext2 or ext3 file system per mirror
- Drawbacks:
  - Expensive in terms of capacity loss
  - Sub-optimal performance if fewer streams than mirrors
- Detailed performance studies in highly dimensional phase space has resulted in …
  - Hardware RAID 5 over all disks of one controller
  - Software RAID 0 (stripe)
  - xfs filesystem
  - Linux kernel: New elevator / VM tuning parameters

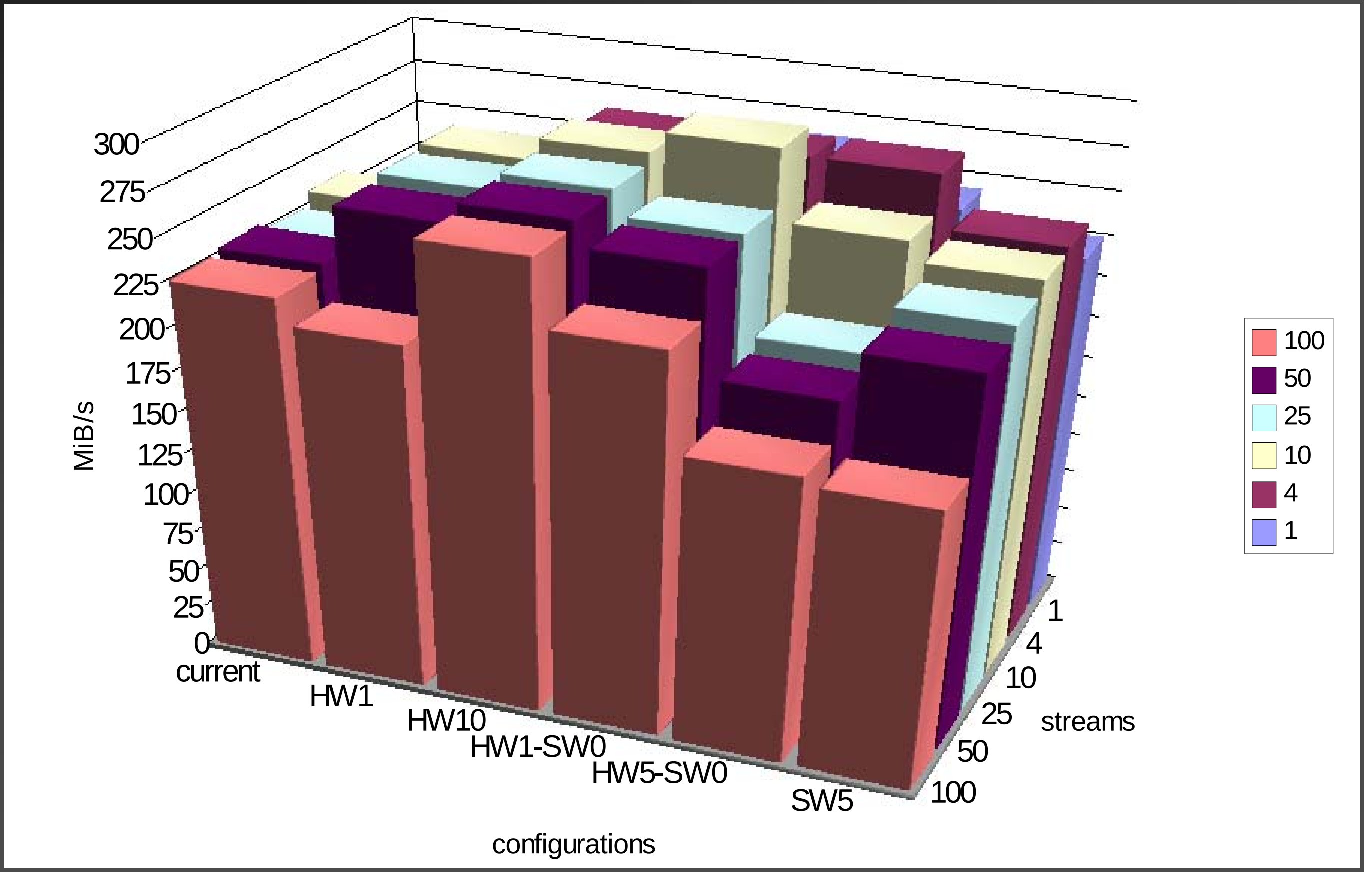
# RAID options

- Comparison of various RAID options
  - Using xfs as file system
  - Tuned elevator and vm kernel parameters
  - iozone benchmark
    - Testing transfers between memory and disk
    - No network involved

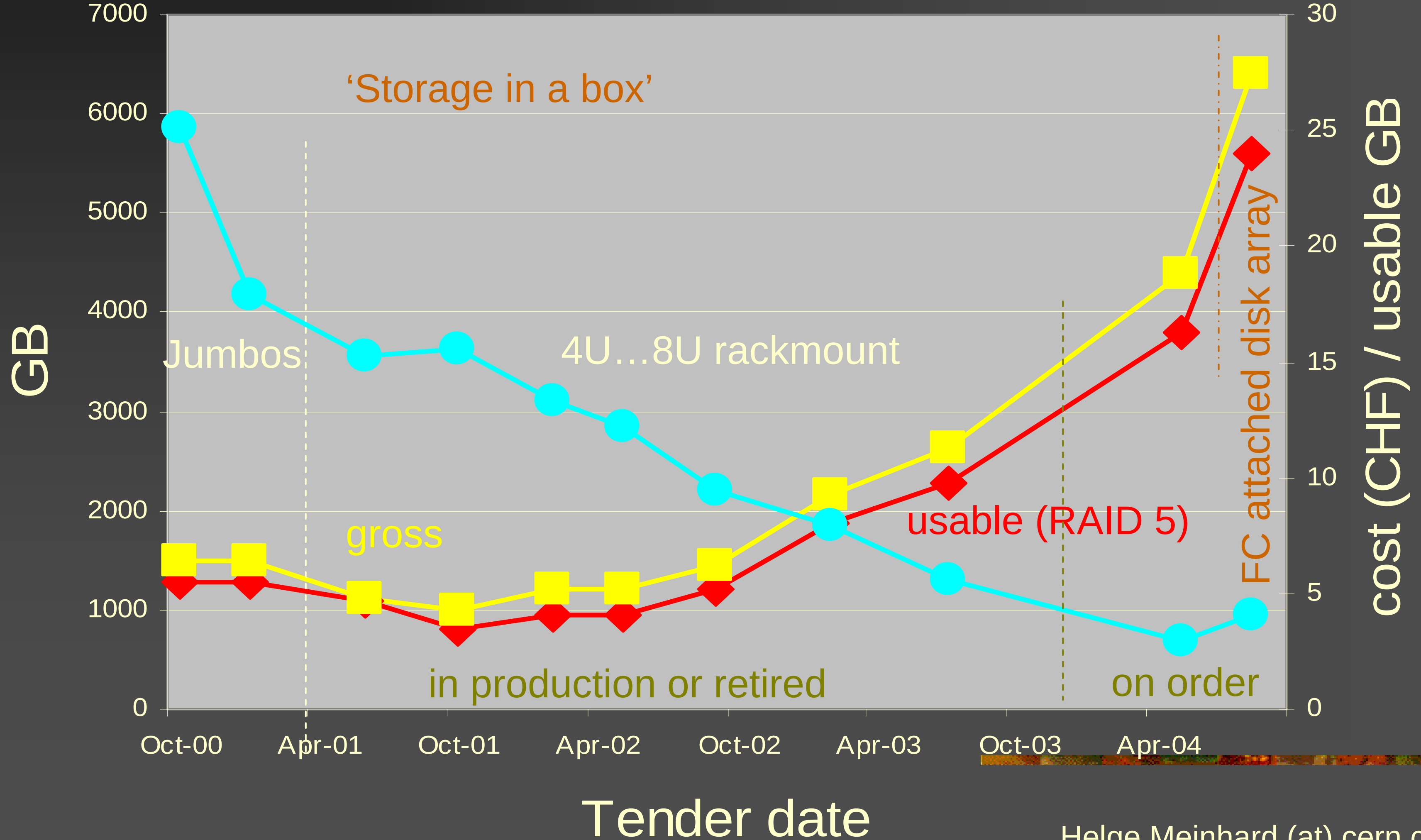
# RAID options – writing



# RAID options – reading



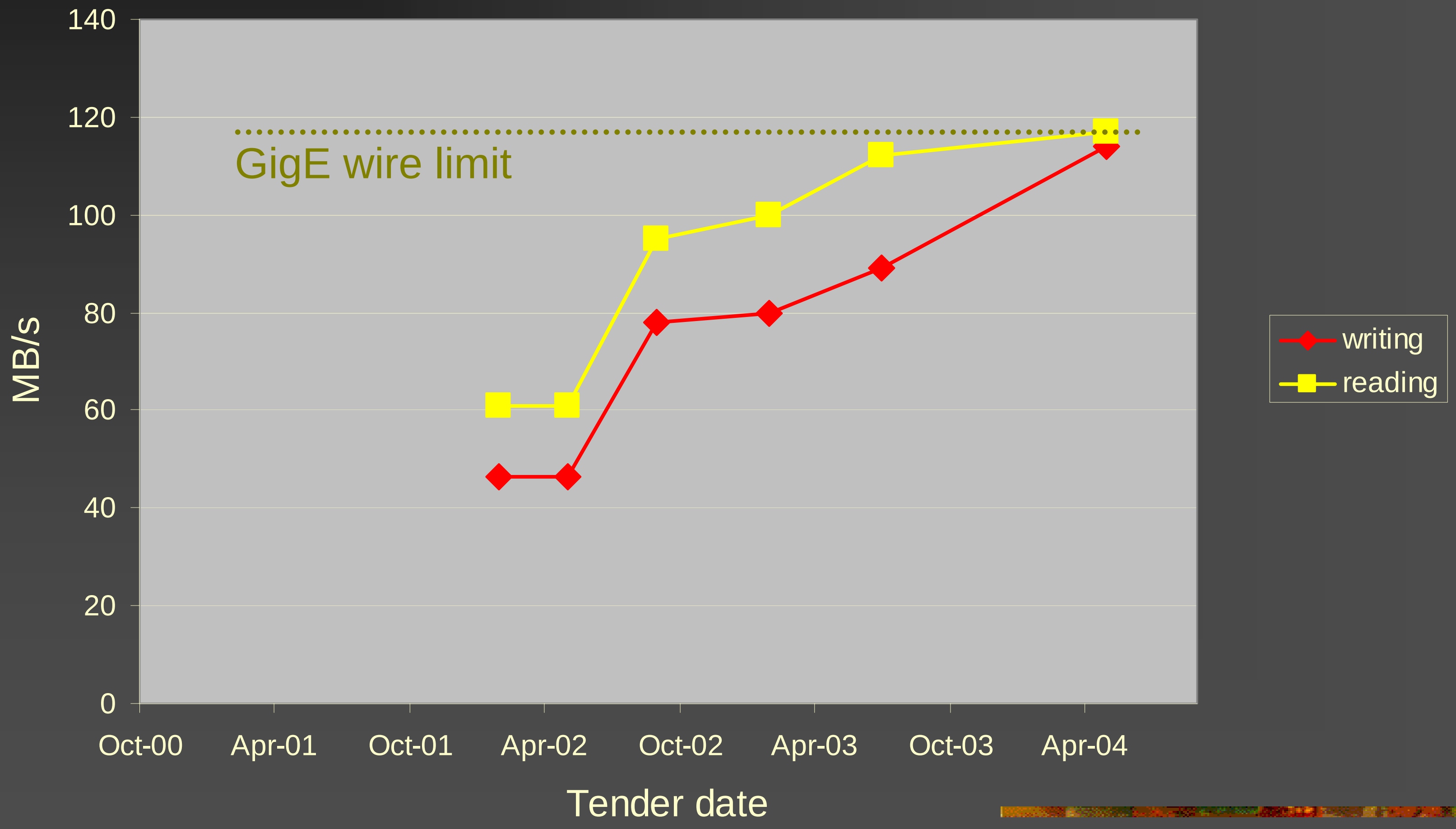
# Capacity per server, cost per GB



# Performance (1)

- Transferring >= 10 files of 2 GB size each
- Into or from the disk server
- Protocol: rfio
- Data path: single Gigabit line
- Disks: mirrored (RAID 1), ext2, no kernel tuning (~ previous configuration)

# Performance (2)



# Reliability, ease of management

- Detailed study under way (see Tim Smith's talk earlier in this session)
- Biggest problem: 51 servers delivered with 24 disks each of a bad batch (bad head construction)
  - All 1224 disks replaced by supplier after 10 months
  - Cages replaced as well
- Most worries (apart from failing disks): bad connectivity (trays and cages, cables)

# Future directions – short term (1)

- Disk technology: Move to SATA
  - Disk server tenders of 2004 have excluded EIDE disks
  - Getting SATA disks now
    - 75 disk servers with 1’800 disks to be delivered next month
  - Hope: better reliability
    - Mechanical quality expected to be (at least) the same as EIDE disks
    - Easier connectivity
    - More professional cages and trays
      - SATA in widespread use
      - Replacing more and more SCSI and FC disks

# Future directions – short term (2)

- System architecture: FC attached space
  - Medium-size tender for FC attached disk arrays and hosts
    - 22 arrays of 16 disks of 400 GB each, to be delivered in November 2004
  - Advantages over disk servers:
    - System architecture more flexible
      - Possible to move to SAN
    - Storage can be made fully redundant
      - Only few applications need that
  - Drawback: higher price
  - Performance measurements ongoing, no conclusive results yet

# Future directions – longer term

- Distributed storage across CPU servers
  - Some testing done
  - Parallel file systems all not adequate today
  - Standard Castor-like usage
    - Not really a change of the big architectural picture
    - Could reduce cost of disk storage
    - Drawbacks: number of 'disk servers' much higher, CPU servers would become stateful

# Conclusions

- Current architecture: distinct tape, disk, CPU services interconnected by Ethernet / TCP-IP
  - Matches well current requirements
  - Is expected to scale such that requirements of LHC will be met as well
  - Has proved to be cost-effective and manageable
- Keeping eyes and ears open for possibilities to optimise performance, reliability, and/or cost
- Future will be evolutionary, not  revolutionary

# Network backbone capacity / load

- Now: 6 routers interconnected with 4 Gbit links each
  - Estimated capacity: ~ 10 Gbit/s
  - Used currently: 200…300 MBytes/s (~ 20%)
- Backbone designed for 2.5 Terabit/s in 2007/2008
  - Estimated usage: T0: 5…10 GBytes / s, the rest: 50 GBytes / s