iSCSI at CERN

Andras.Horvath@cern.ch, 2009

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

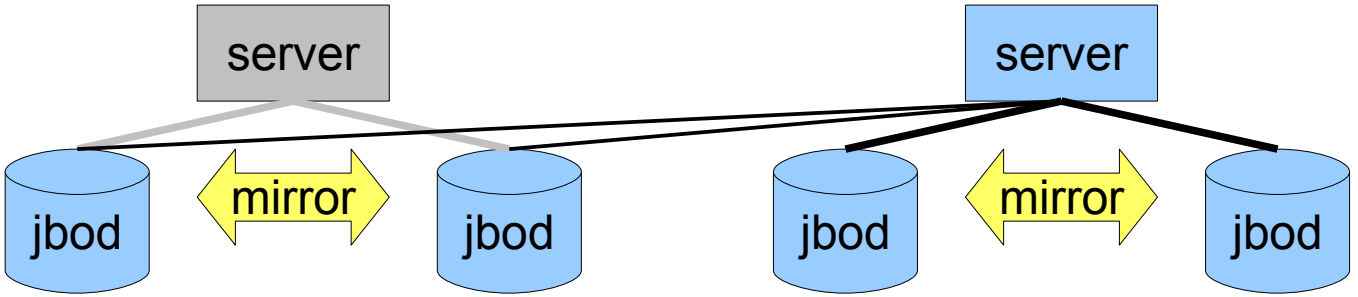
Focus

- "Good Enough (tm)" performance...
- Large-scale deployment
  - Reliability and robustness
  - Automatisation
  - Monitoring
- Cost
  - hardware
  - operational

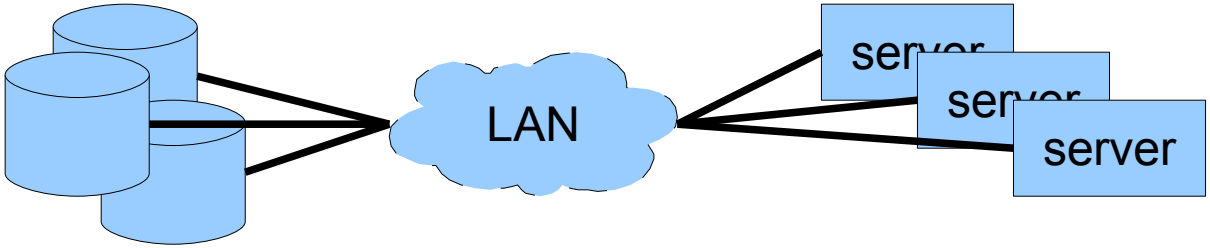
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Motivation

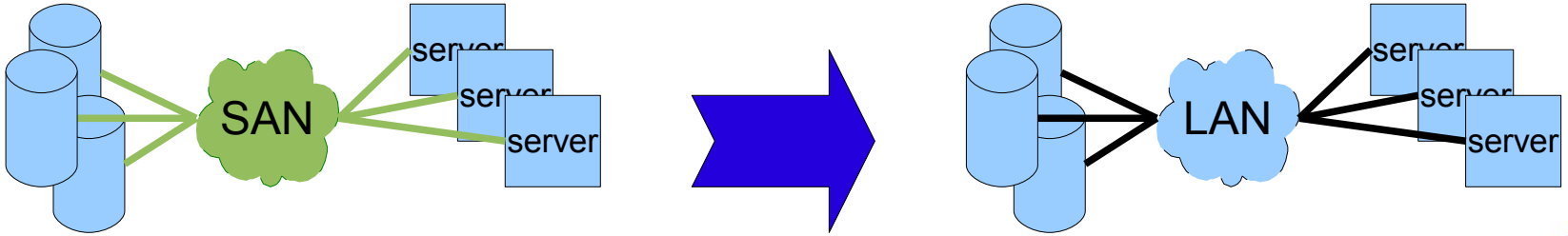
increase data availability



more flexibility in scaling



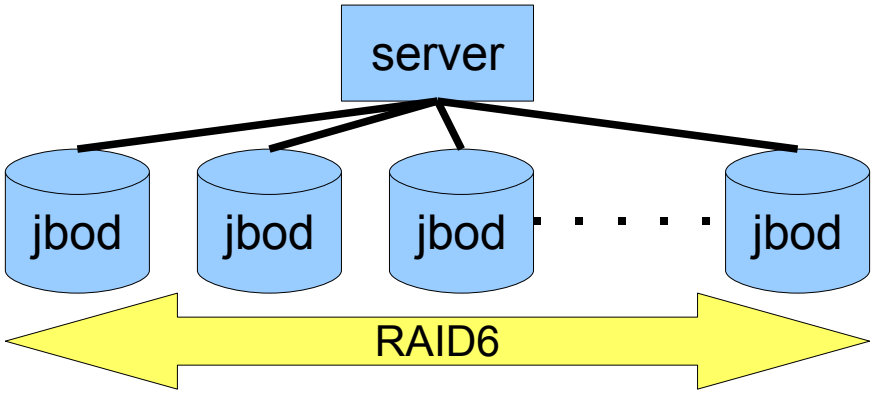
consolidate network infrastructure

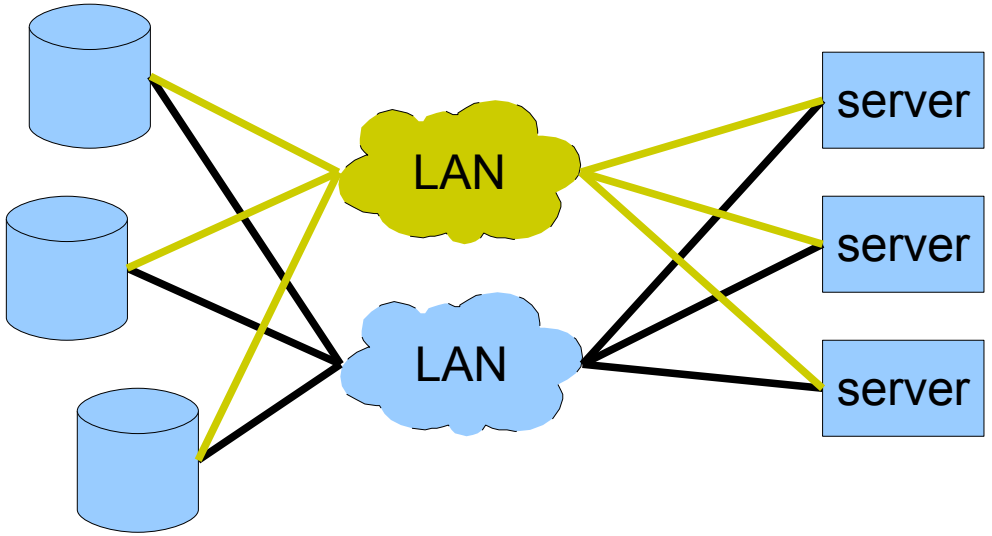


CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Use cases

- Bulk storage
- Databases (RAC)
- Virtualisation
- Backup space
- Lustre OST/MDT
- ....





CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Software

- Linux initiators: OK in SL4 and SL5
- Linux targets:
  - default RHEL5 stack (tgtd): single-threaded
  - IET target stack: multithreaded, faster
- Robustness
  - client can survive target reboots (!)
  - zero fsprobe errors so far
  - watch out for persistent connections

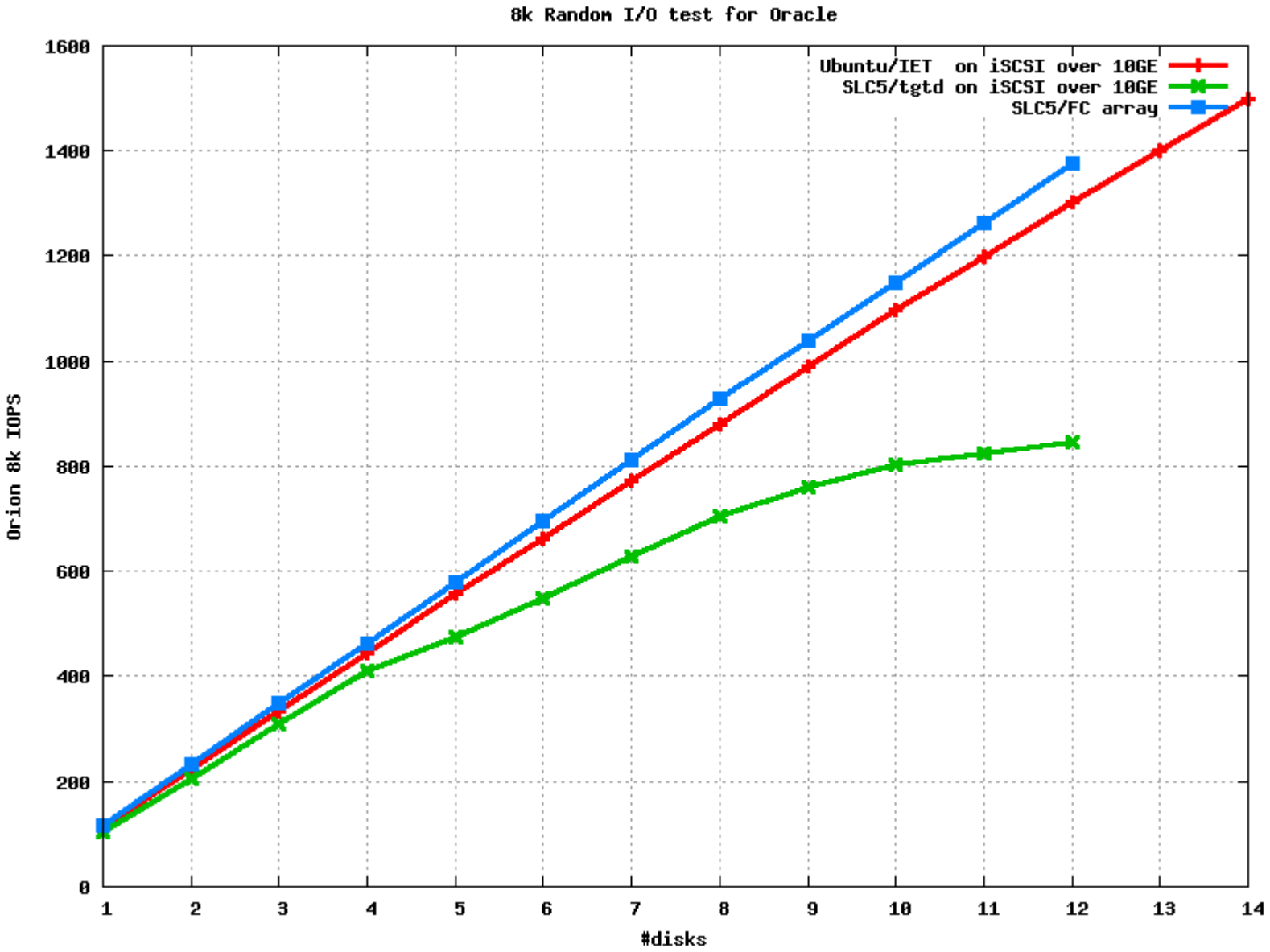
CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Target hardware

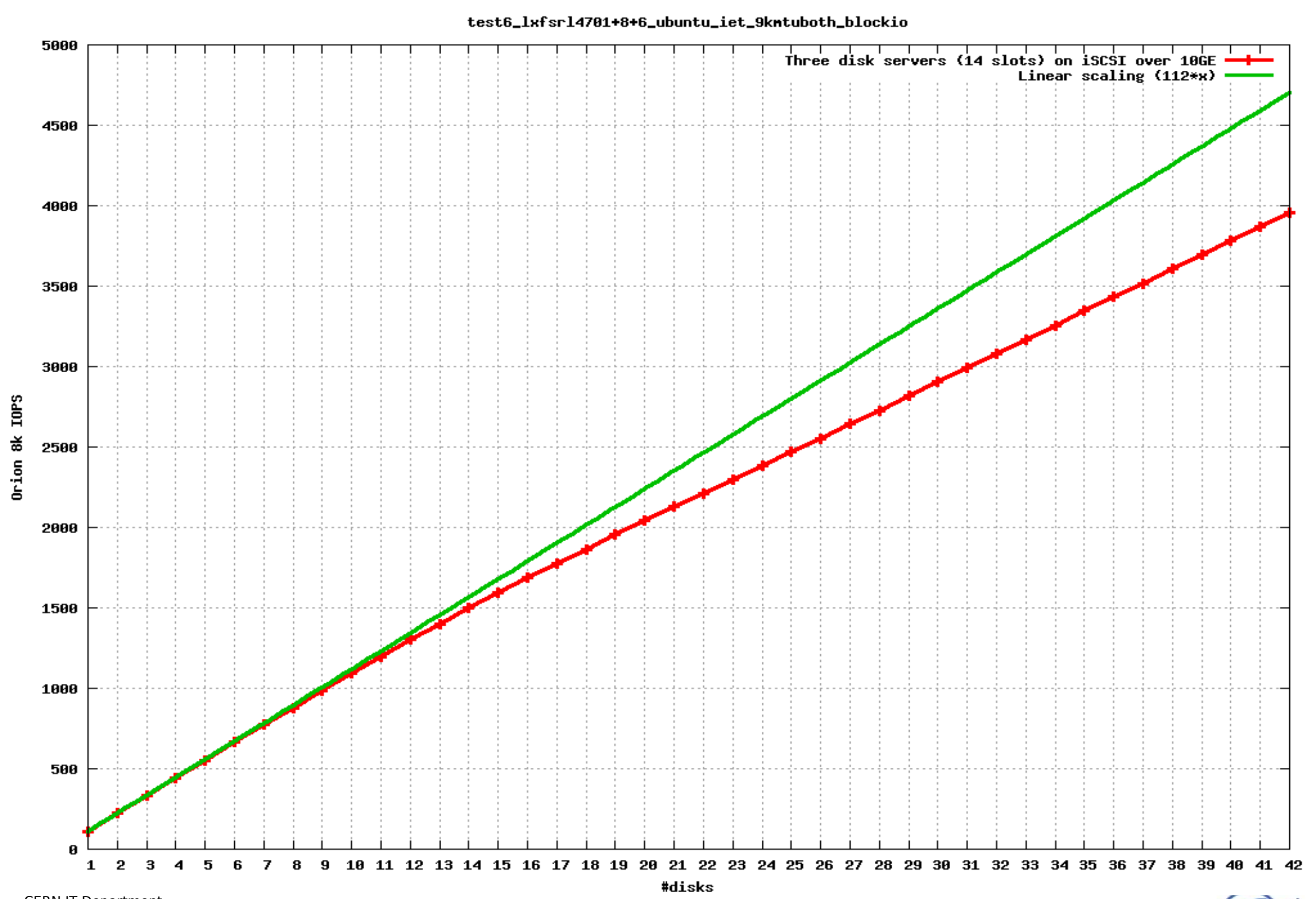
- Infortrend low-end box
  - inexpensive, 1GE only
- Dell EqualLogic “medium” box
  - 3x1GE, redundant everything, load balancing
- Storage-in-a-box server
  - GE or 10GE
  - Ubuntu Jaunty (IET stack) or SLC5 (tgtd)
  - 3ware RAID card in JBOD mode
  - 2x Clovertown CPUs, 8G RAM, 14+2 disks

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Oracle IOPS



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it



CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Ongoing efforts

- Test and learn corner cases (MD RAID..)
- Quattor NCM components
  - initiator- and target-side components
  - ncm-filesystems/blockdevices will be reusable
  - Linux-HA integration
- Monitoring
  - MD monitoring is available
- Recovery procedures and tools
- Resource management

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Timeline

1) Proprietary targets, pilot production
2) Linux target, simplest (mirrored) case
   - w/o failover
3) Oracle test setup
   - JBODs, Oracle manages data
4) Mirror with automated failover
   - Linux-HA integration
5) Lustre OST
   - Test for Lustre, production for iSCSI

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it

Long-term..?

- Improve MD rebuild speeds
  - write-intent bitmap (on SSD)?
  - parallel rebuild? (~RAID5E / 5EE / 6E)
- Data integrity
  - T10 DIF (data integrity)?
  - read/check parity on read?

CERN IT Department
CH-1211 Genève 23
Switzerland
www.cern.ch/it