IA64 as disk server (WAN / LAN)

Andreas Hirstius, CERN openlab

Andras Horvath, CERN IT-ADC

How to deal with the LHC datarate ?

- Distribute the work (“Grid”)
  - The data will be transferred to Tier 1 centers
    - Large regional centers (e.g. FNAL, BNL, IN2P3 ...)
    - Most physics analysis is done at the Tier 1 centers
  - At the moment 12 Tier 1 centers worldwide
    - The closest one is ~100km away
    - The one farthest away is in Japan
    - Continuous datarates of **~2–7Gbit/s *per site***
  - Total datarate (estim. Sept. 2004): **~55Gbit/s**
  - 10Gb link to each Tier 1 center foreseen

(Disk)-Storage Subsystem tests

- Used in the “Robust Data Transfer Challenge”
  - Goal: reliable high speed data paths to Tier 1 centers
  - Currently one 10Gb lightpath to Europe + one to US
  - First goals for datarates (mid 2005)
    - 500MB/s disk-to-disk sustained over 2 weeks to one site
    - Max. possible data rate disk-to-disk sustained to all sites
    - N ∗ 100MB/s “tape-to-tape”
  - Other objectives
    - Improving stability & reliability of the used software

Network attachement

- Two possibilities:
  - Aggregation of 1Gb connections (current model)
    - \+ Simple and straightforward setup
    - \- Large Number of machines necessary
    - \- Max. single stream transfer rate limited to 1Gb

      (will be a problem when tapes can deliver >100MB/s ... next year??)
  - Directly attached 10Gb
    - \+ Very high transfer rates possible
    - \+ Smaller number of machines necessary
    - \- Expensive hardware

10Gb LANs

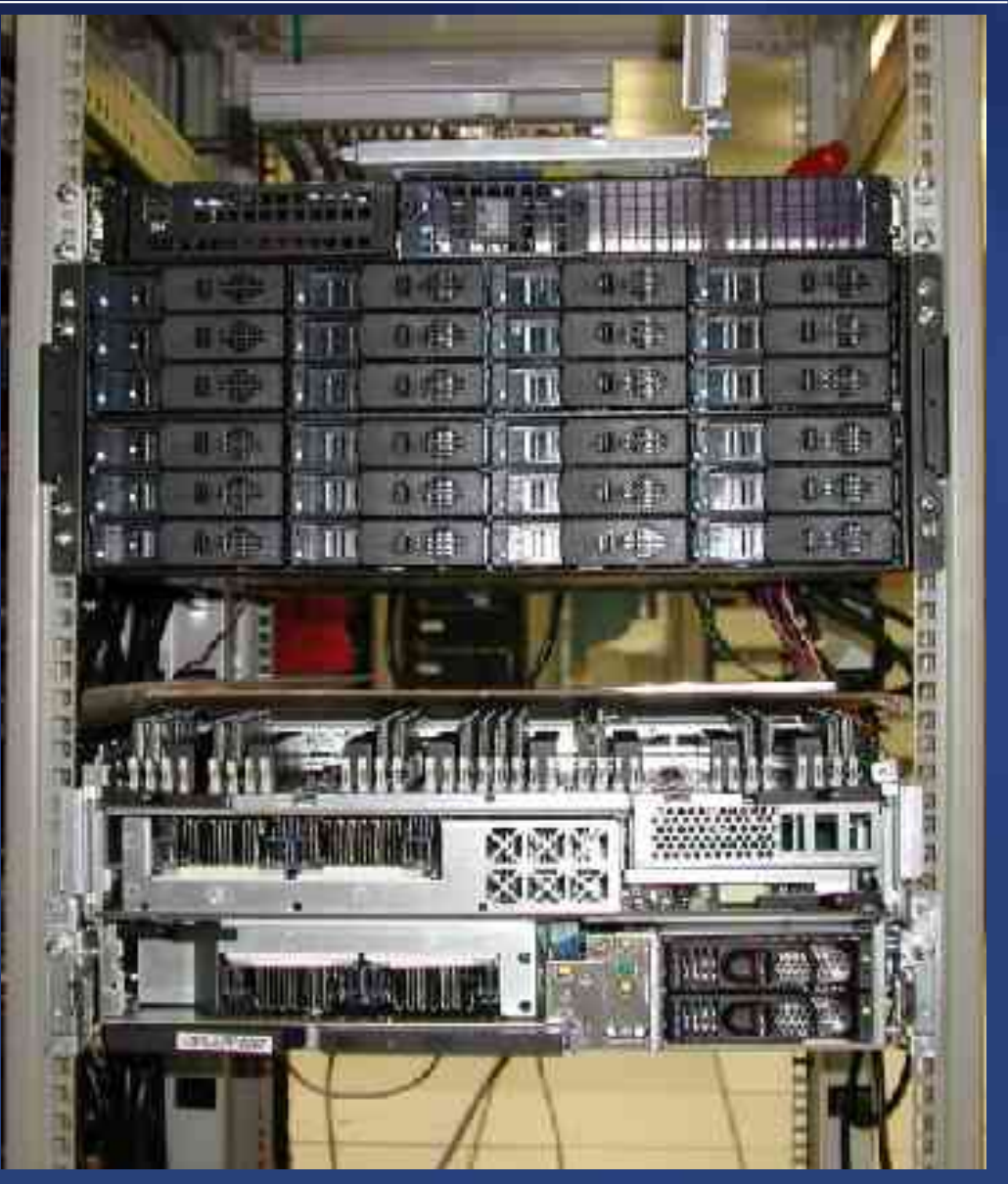
- Several 10Gb NICs from Intel (SR and LR)
  - Stress testing Enterasys X-Series
    (and other network equipment)
  - Back-to-Back performance tests
  - Unfortunately no adequate 10GE infrastructure (yet)
- InfiniBand
  - Awaiting 96-port switch from Voltaire
  - \+ Very low CPU overhead at linespeed
  - \+ Home grown protocol (RFIO) is ported to IB
  - \- Still expensive, but costs dropping fast

Test setup at CERN

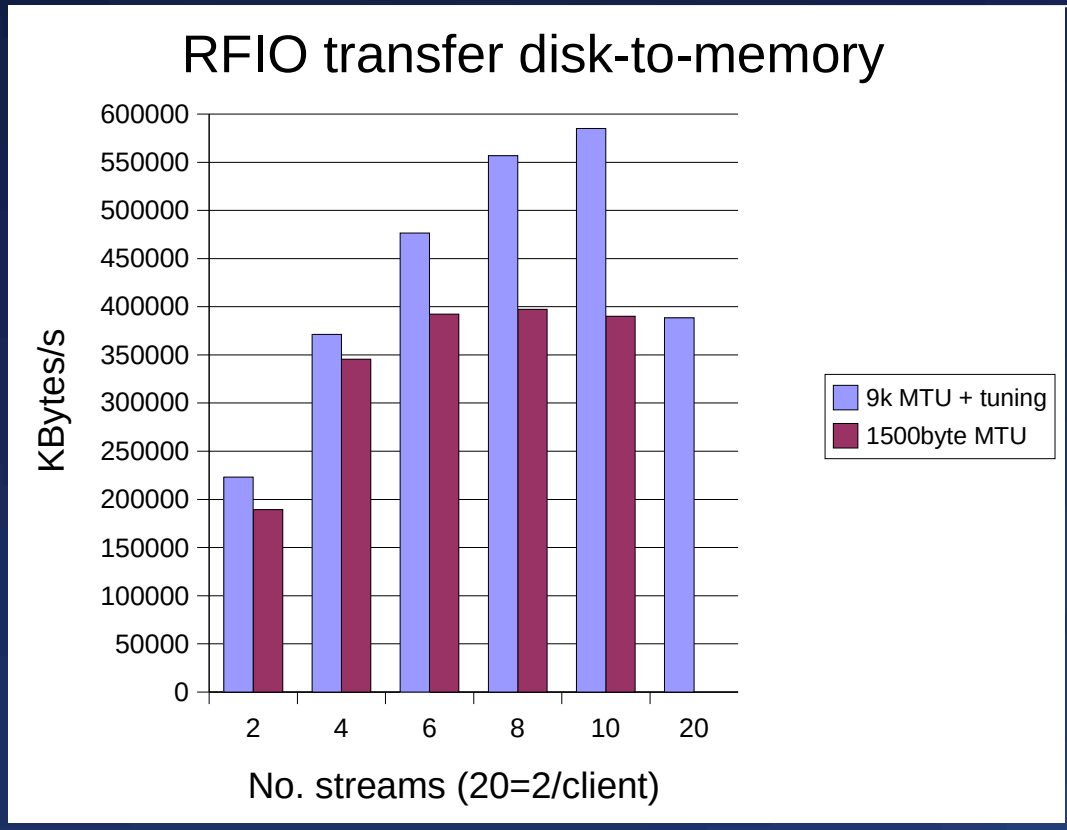
- ten 2-way@1.5Ghz HP rx2600 Itanium2 clients (LAN tests)
- one 4-way@1.5Ghz HP rx4640 Itanium2 disk server
  - 16GB RAM
  - 24 SATA disks and two 3ware 9500 controllers
  - Scientific Linux CERN with vanilla 2.6.9-rc1 kernel
  - Software RAID0 over hardware RAID0, XFS
- Intel 10GE and GE adapters / (Voltaire IB HCAs)
- Foundry MG8 10GE switch / (Voltaire ISR9600)
- RFIO, home-grown transfer protocol, part of CASTOR HSM
  - Infiniband port by Dr Ulrich Schwickerath @ FZK

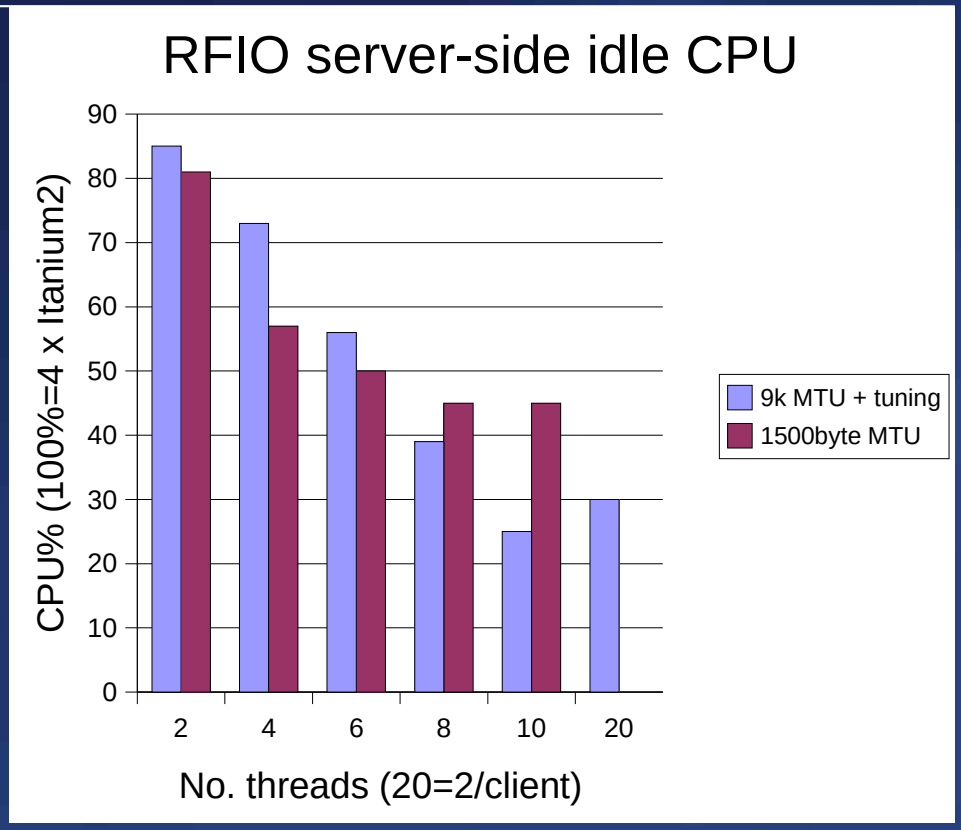
Disk ⇒ Memory results (using iperf)

- Internal data rates (ms+ss)
  - ~770MB/s read
  - ~350MB/s write
- 10Gb back-to-back (ms+ss)
  - ~710MB/s read (NIC is the limitation)
  - ~350MB/s write
- 10Gb to California (ss)
  - **~660MB/s** read :-))
    (~1300GB in 2000s)



Disk ⇒ Memory results (using rfio)





Outlook (not that Outlook. Just future development forecast)

- Performance improvements for RFIO
  - Analyze differences to iperf
  - e.g. use sendfile()
- Infiniband tests (have started, new driver for 2.6.9 kernel available)

Already reported at Spring HEPix 2004 (cached mem to /dev/null):