CERN
openlab for DataGrid applications
Developing Solutions for the Data-Intensive Science of the Large Hadron Collider

10 Gbit/s Challenge inside the Openlab framework

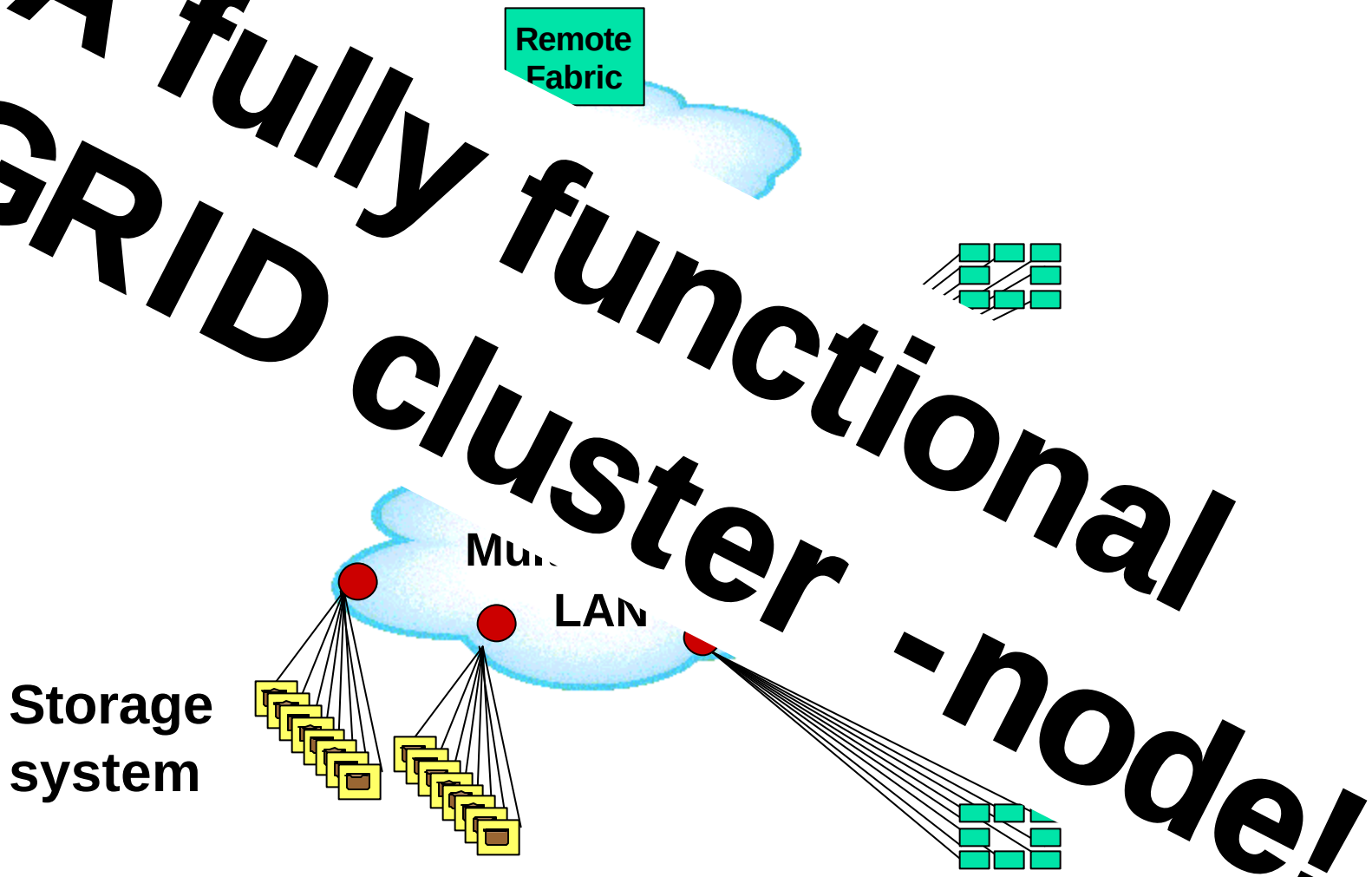
Sverre Jarp
IT Division
CERN

CERN
openlab for DataGrid applications
Developing Solutions for the Data-Intensive Science of the Large Hadron Collider

Agenda

- Introductions
  - All
- Overview
  - Sverre
- Feedback
  - Enterasys
  - HP
  - Intel
- Further discussions
- Elaboration of plan
  - Deliverables
  - Time line

Opencluster vision



A fully functional GRID cluster -node!

Openlab today

- **Industrial Collaboration**
  - Enterasys, HP, and Intel are our partners today
  - Additional partner(s) joining soon
    - Storage subsystem from 4th partner
  - Technology aimed at the LHC era
    - Network switches at 10 Gigabits
    - Rack-mounted HP servers
    - Itanium-2 processors

  - Cluster evolution:
    - 2002: Cluster of 32 systems (64 processors)
    - 2003: 64 systems ("Madison" processors)
    - 2004/05: Possibly 128 systems ("Montecito" processors)

CERN openlab for DataGrid applications
Developing Solutions for the Data-Intensive Science of the Large Hadron Collider

Opencluster - phase 1

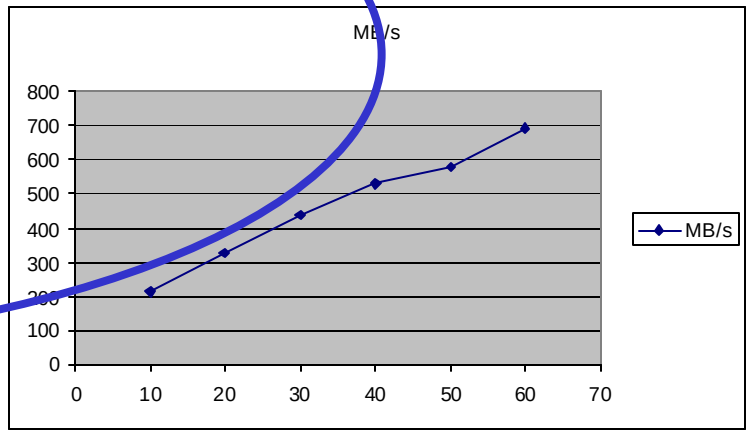
- **Integration of the cluster:**
  - **Fully automated network installations**
  - **32 nodes + development nodes**
  - **RedHat Advanced Workstation 2.1**
  - **OpenAFS, LSF**
  - **GNU, Intel, ORC Compilers**
  - **CERN middleware: Castor data mgmt**
  - **CERN Applications**
    - **Porting, Benchmarking, Performance improvements**
    - **CLHEP, GEANT4, ROOT, Sixtrack, CERNLIB, etc.**
  - **Database software (MySQL, Oracle?)**

Also: Prepare porting strategy for GRID phase

CERN

Opencluster - phase 1 (cont'd

- **Perform cluster benchmarks:**
  - **Parallel ROOT queries (via PROOF)**
    - **Observe scaling: 2 → 4 → 8 → 16 → 32 → 64 processors**
  - **"1 GB/s to tape" challenge**
    - **Opencluster as CPU servers**
    - **50 StorageTek tape drives in parallel**
  - **"10 Gbit/s network" challenge**
    - **Groups together all Openlab partners**
      - **Enterasys switch**
      - **HP servers**
      - **Intel processors and n/w cards**
      - **CERN Linux and n/w expertise**



CERN
openlab for DataGrid applications
Developing Solutions for the Data-Intensive Science of the Large Hadron Collider

Why such a challenge?

- **Demonstrate LHC-era technology**
  - **All necessary components inside Opencluster**
  - **Identify bottlenecks**
    - **And see if we can improve**
- **We know that Ethernet is here to stay**
  - **4 years from now 10 Gbit/s should be commonly available**
    - **Backbone technology**
    - **Cluster interconnect**
    - **Possibly also for iSCSI and RDMA traffic**

Advancing the state-of-the-art !

CERN
openlab for DataGrid applications
Developing Solutions for the Data-Intensive Science of the Large Hadron Collider

Demonstration of Openlab partnership

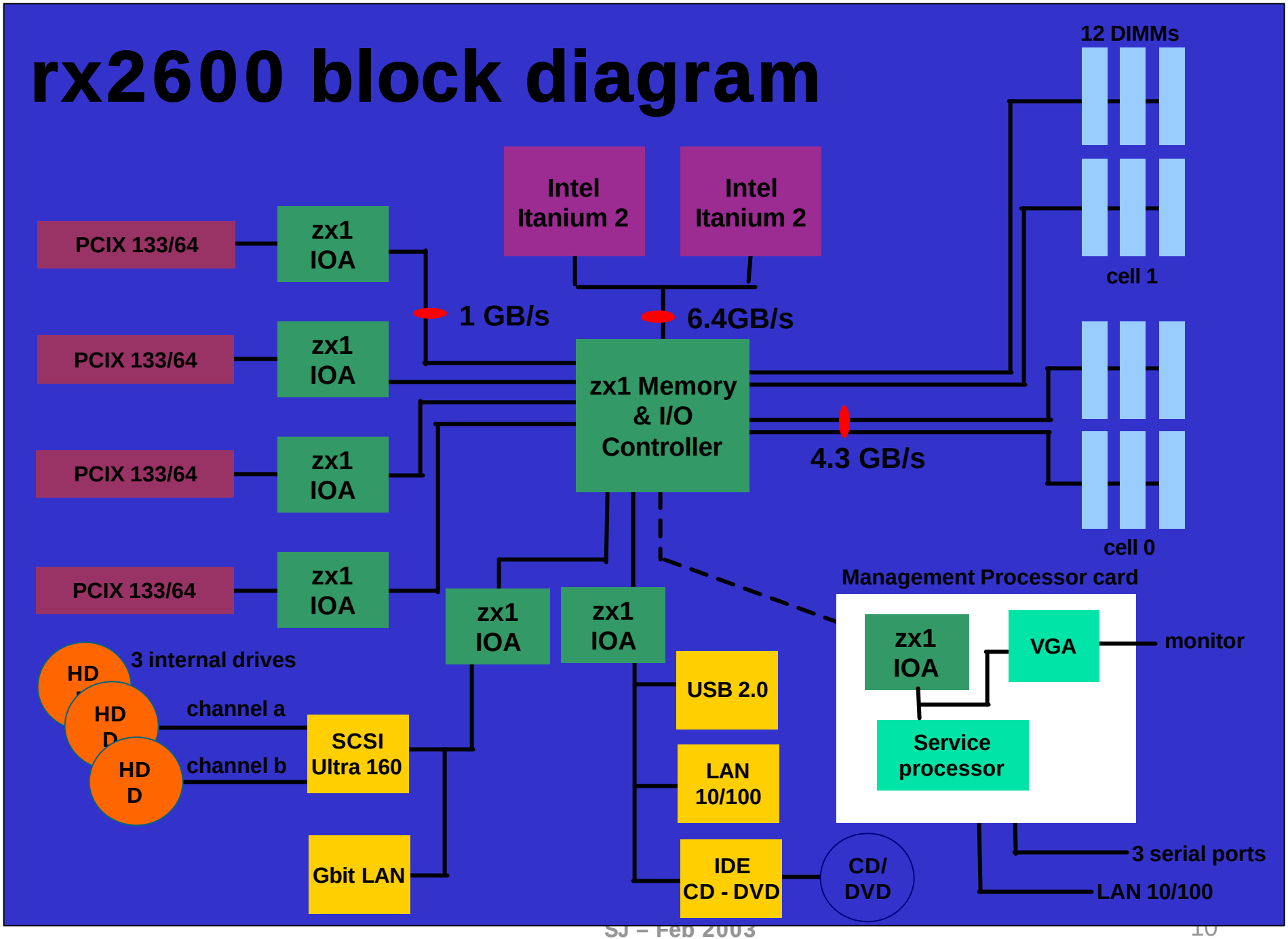
- **Everybody contributes:**
  - **Enterasys**
    - **10 Gbit switches**
  - **Hewlett-Packard**
    - **Servers w/PCI-X slots and memory bus**
  - **Intel**
    - **10 Gbit NICs**
    - **Processors (i.e. code optimization)**
  - **CERN**
    - **Linux kernel expertise**
    - **Network expertise**
    - **Project management**
    - **IA32 expertise**
      - **CPU clusters, disk servers on 1 Gbit infrastructure**

CERN openlab for DataGrid applications
Developing Solutions for the Data-Intensive Science of the Large Hadron Collider

The compute nodes

- **HP rx2600:**
  - Rack-mounted (2U) systems
  - Two Itanium-2 processors
    - 900 or 1000 MHz
    - Field upgradeable to next generation
  - 2 or 4 GB memory (max 12 GB)
  - 3 hot pluggable SCSI discs (36 or 73 GB)
  - On-board 100 Mbit and 1 Gbit Ethernet
  - 4 PCI-X slots:
    - full-size 133 MHz/64-bit slot(s)
  - Built-in management processor
    - Accessible via serial port or Ethernet interface

rx2600 block diagram

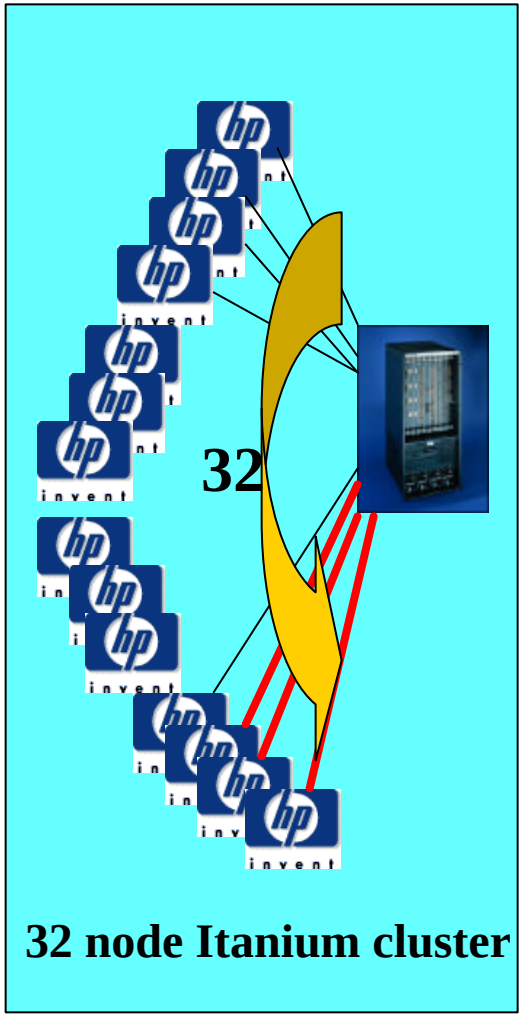


Can we reach 400 – 600 MB/s throughput?

- **Bottlenecks could be:**
  - **Linux**
    - **Kernel and driver optimization**
      - **Number of interrupts; tcp checksum; ip packet handling, etc.**
  - **Server hardware**
    - **Memory banks and speeds**
    - **PCI-X slot and overall speed**
  - **Switch**
    - **Single transfer throughput**
- **Aim:**
  - **identify bottleneck(s)**
  - **Measure**
    - **peak throughput**
    - **Corresponding cost: processor, memory, switch, etc.**

Cluster interconnect

- **Our switch:**
  - **Enterasys ER16**
- **Mix of 1 and 10 Gbit/s connections**
  - **Measure**
    - **CPU server ← → CPU server**
      - **Memory to memory**
    - **Disk server ← → CPU server**
      - **Disks to applications**
    - **@ 1 Gbit/s**
      - **Full speed without problems?**
    - **@ 10 Gbit/s**
      - **400 – 600 MB/s ?**



CERN
openlab for DataGrid applications
Developing Solutions for the Data-Intensive Science of the Large Hadron Collider

Next steps today

- **Feedback**
  - **Enterasys**
  - **HP**
  - **Intel**
- **Further discussions**
- **Elaboration of plan**
  - **Deliverables**
  - **Time line**