Track 6
Computing Fabrics

CHEP 2004 / Interlaken

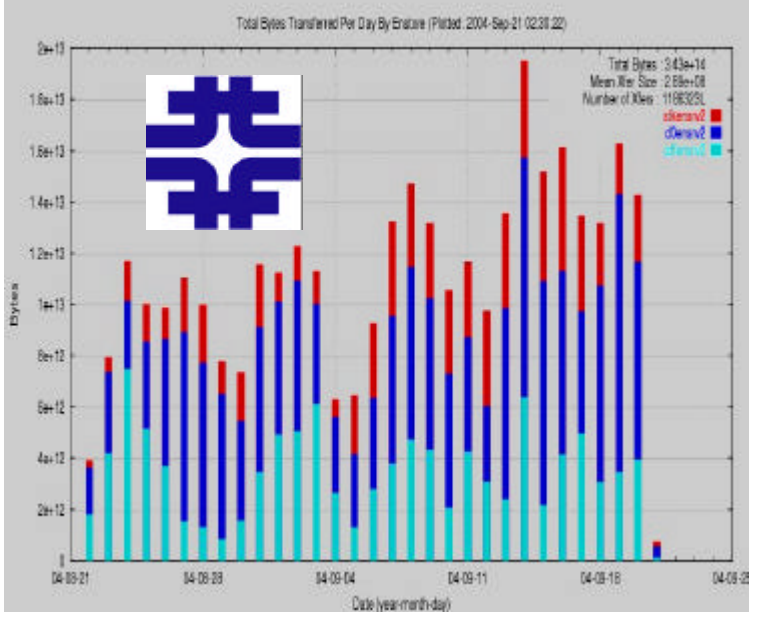
Ian Fisk, Tim Smith

Overview

- 24 oral and 18 poster
  - Not a comprehensive recital
    - Just pick out some themes
    - Or at least hot topics when no consensus
- Tiers for the LHC
- SW techniques
- HW technologies
- Concluding Remarks

Scale

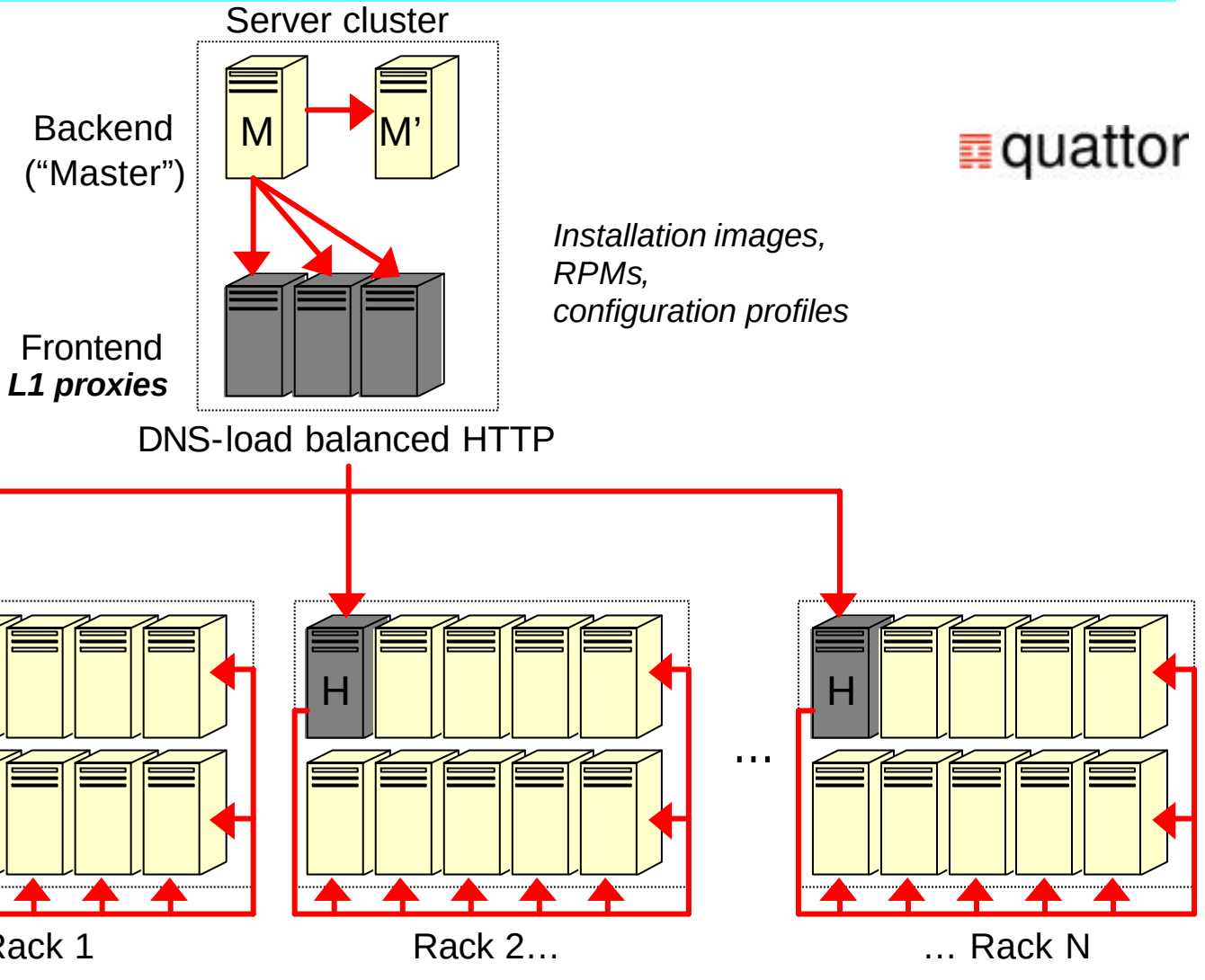
- GridKa: 500 dual CPU nodes, 220 TB disk, 400 TB tape
- BNL: 1300 nodes
  - PHENIX: 300MB/s
- Belle: 540 nodes
- FNAL: 110 tape servers
  - 1.9 PB
- CERN: 370 disk servers
  - 3 PB, 25M files
  - COMPASS 120 MB/s



Cascading Tiers

- Differentiating factor between tiers:
  - Not necessarily scale of HW
    - Multi-disciplinary T2s with large farms
  - Scale of support
  - Diversity of services
- Upper Tiers with large teams want tools to coordinate large and diverse services
  - Emphasis on coordinated information
  - Central servers to orchestrate the automation
- Lower Tiers with fractions of FTEs
  - Turnkey solutions
  - Low maintenance central services
    - *"Challenge is to run cheap HW with minimal staff and moderate expertise"*

Tier0 Installation Servers



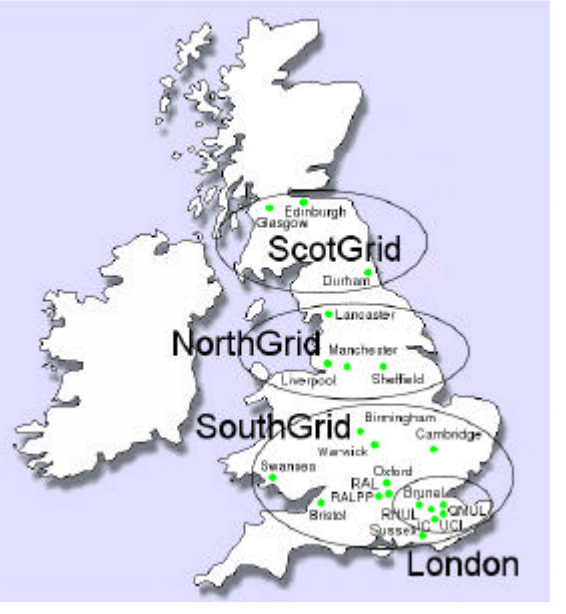
Fabrics @ Tier1

- *"Transforming from a local resource to a global one"*
- *"Gone are the independent kingdoms where emperors can setup what ever they want"*
  - Grid era restricts such freedoms – adhere to interfaces, which often reach right down to individual farm nodes
  - Compatibility and interoperability
- *"Centralised management to allow to reassign and redeploy resources"*

- [496] Developing and Managing a large Linux farm -- the Brookhaven Experience
- [195] The CMS User Analysis Farm at Fermilab

Fabrics @ Tier2

- Variety…
  - Different HEP experiments
  - Different Tier1/0s
  - Different sciences on the same campus
- Distributed T2s
  - Ensure matching of disk/CPU per part
- Infrastructure of automation also needs to be lightweight and easy
- Portals for remote management



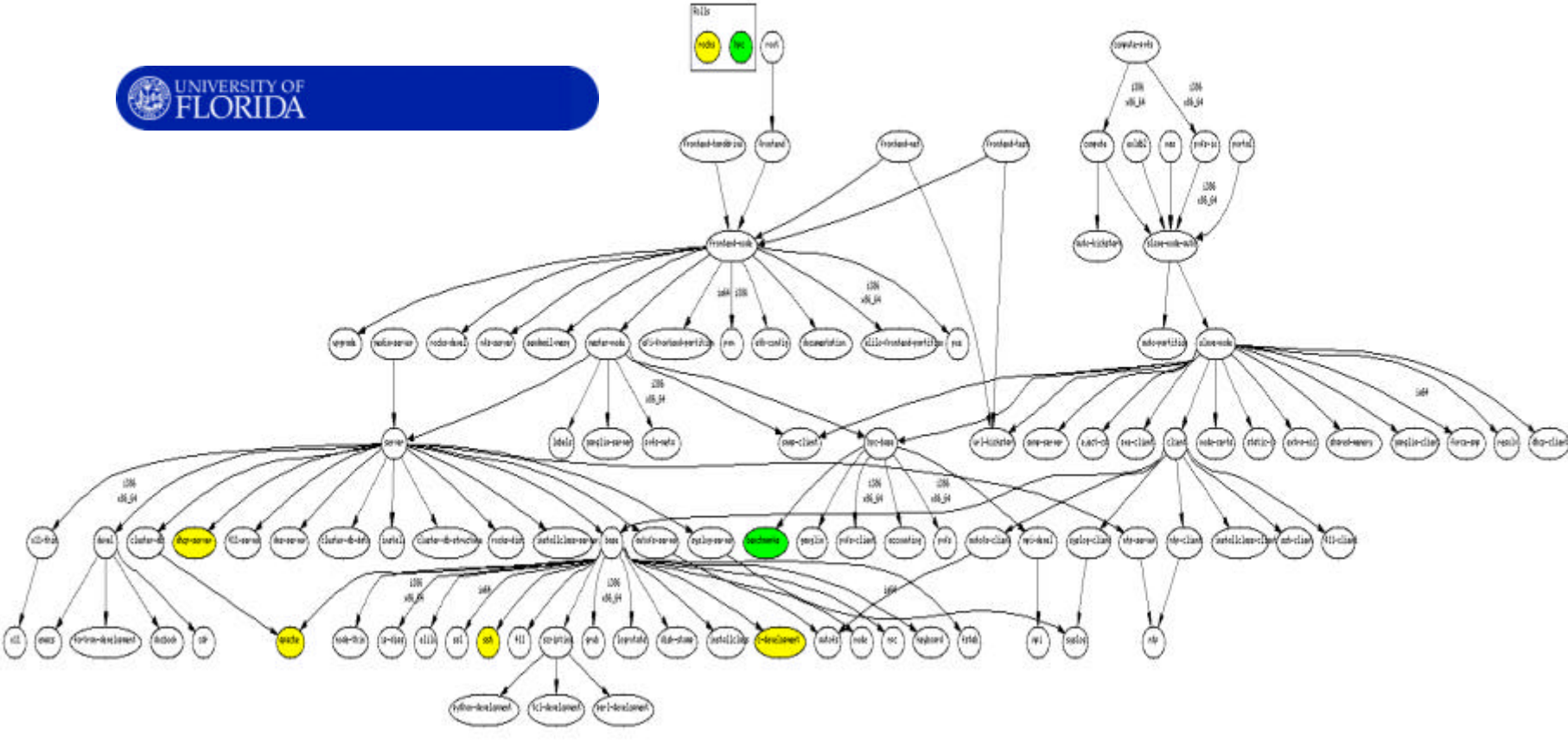
- [207] ScotGrid: A prototype Tier 2 centre
- [330] A Regional Analysis Center at the University of Florida
- [304] The Design, Installation and Management of a Tera-Scale High Throughput Cluster for Particle Physics Research

Installation / Maintenance

- Philosophies
  - Rocks: Reproducible installations
    - Reinstall to update
  - Quattor: Actively manage the running environment
    - Live sync with desired configuration
- Configuration description
  - Hierarchies and dependency graphs

- [489] Current Status of Fabric Management at CERN
- [496] Developing and Managing a large Linux farm -- the Brookhaven Experience
- [180] A database prototype for managing computer systems configurations

Configuration

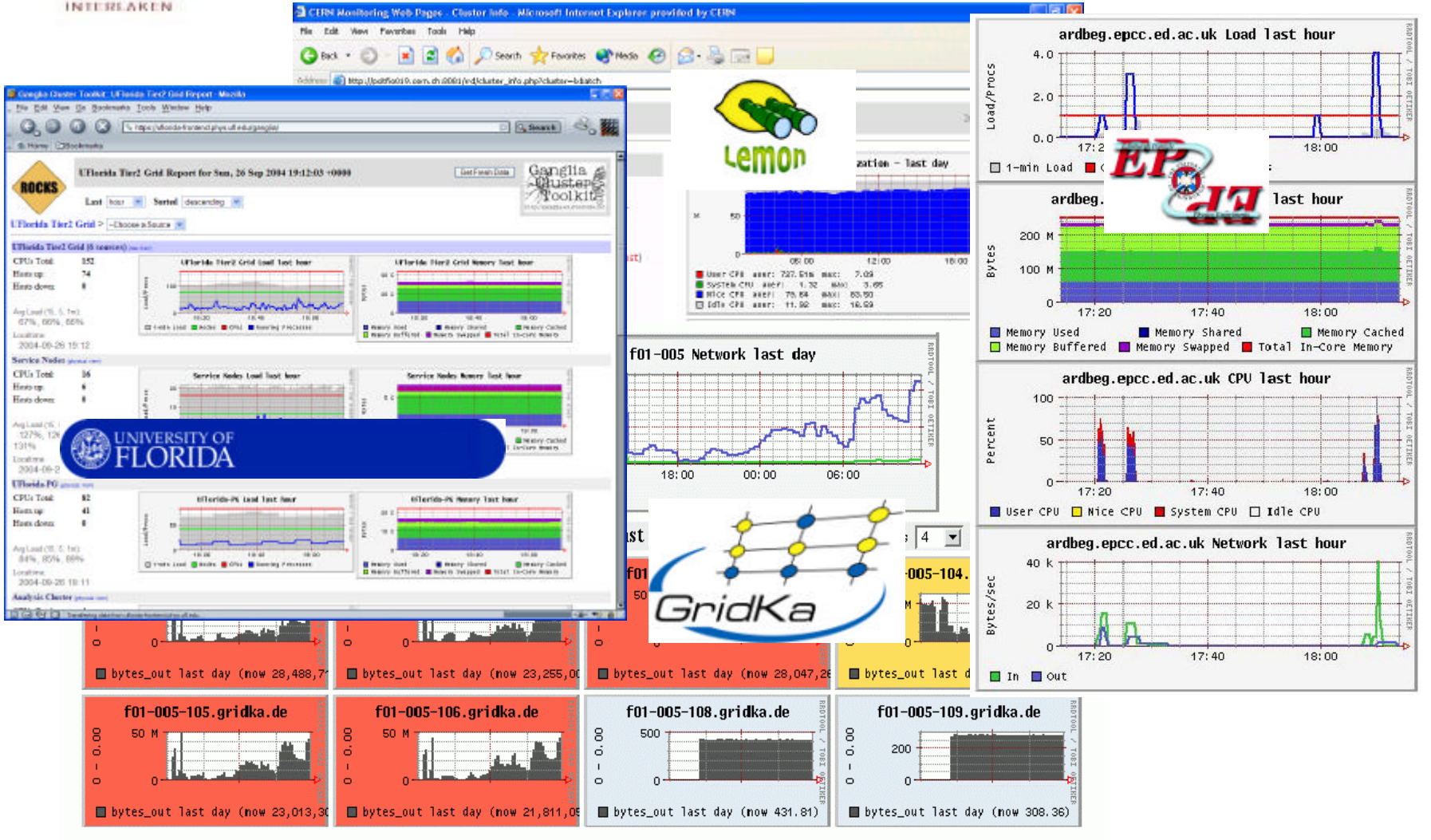


Installing Grid SW

- Grid *under*ware is complex to install
  - *"Gained valuable LCG experience, in both installation and maintenance"*
- ...including the grid installation server itself
  - *"installation of the "LCFG"-Server itself takes most of the time, thus hindering widespread use"*

- [207] ScotGrid: A prototype Tier 2 centre
- [151] Simplified deployment of an EDG/LCG cluster via LCFG-UML
- [152] InGRID - Installing GRID

Monitoring


CERN Monitoring Web Pages - Cluster Info - Microsoft Internet Explorer provided by CERN
Lemon
ardbeg.epcc.ed.ac.uk Load last hour
Load/Procs
1-min Load
ardbeg.epcc.ed.ac.uk Memory last hour
Bytes
Memory Used
Memory Shared
Memory Cached
Memory Buffered
Memory Swapped
Total In-Core Memory
ardbeg.epcc.ed.ac.uk CPU last hour
Percent
User CPU
Nice CPU
System CPU
Idle CPU
ardbeg.epcc.ed.ac.uk Network last hour
Bytes/sec
In
Out
f01-005 Network last day
UNIVERSITY OF FLORIDA
GridKa
f01-005-105.gridka.de
f01-005-106.gridka.de
f01-005-108.gridka.de
f01-005-109.gridka.de
bytes_out last day (now 23,013,30
bytes_out last day (now 21,811,09
bytes_out last day (now 431.81)
bytes_out last day (now 308.36)

Monitoring

- Maturing field
  - Ganglia: widespread use
  - LEMON: similar look and feel
  - *“Expose 100s of quantities on open web as believe users can help debug problems in complex distributed environment”*
- Self-healing fabrics

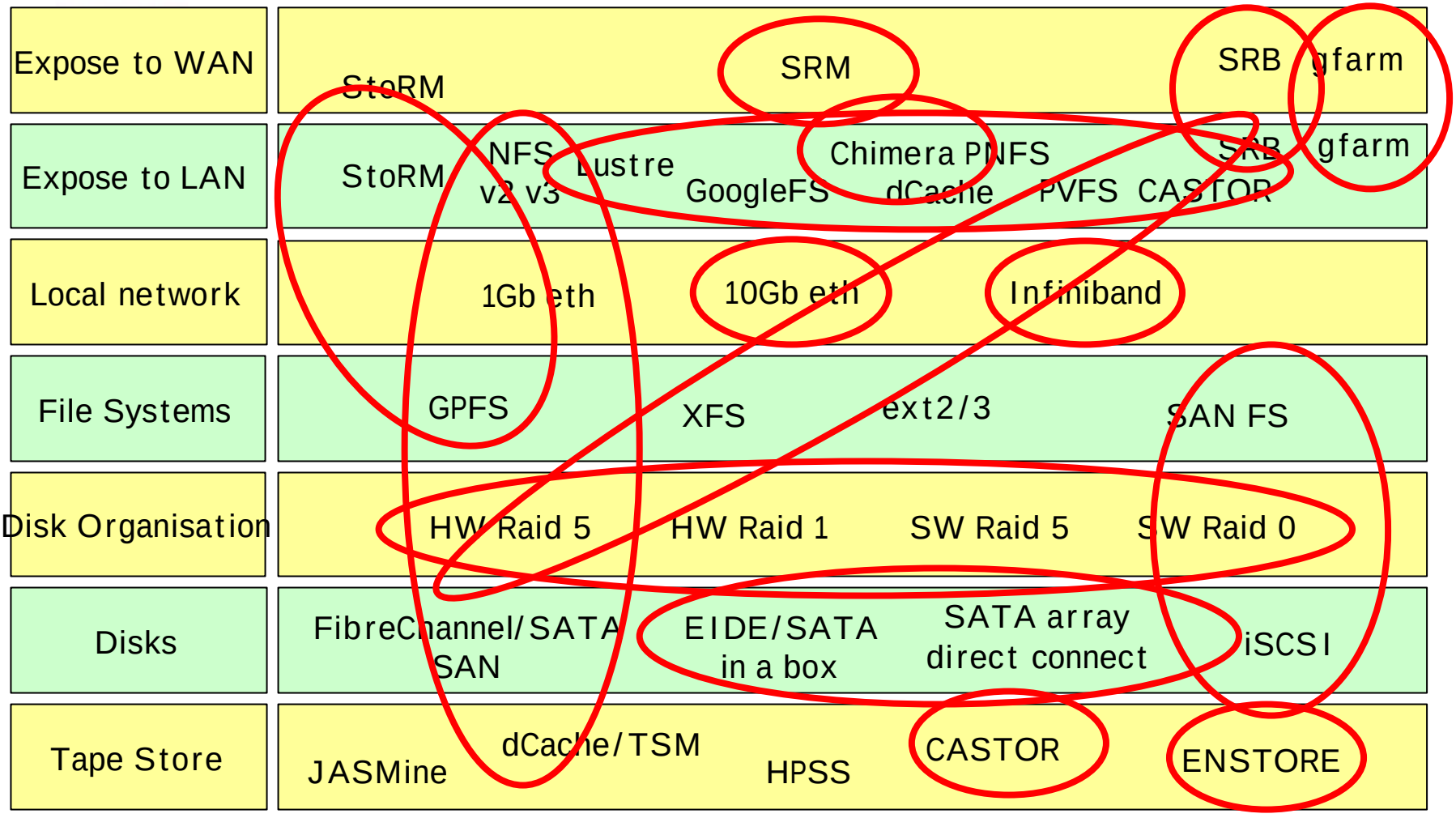
- [489] Current Status of Fabric Management at CERN
- [496] Developing and Managing a large Linux farm -- the Brookhaven Experience
- [474] Experiences Building a Distributed Monitoring System
- [484] Monitoring the CDF distributed computing farms

OS choice

- Move to RHES3 / Scientific Linux
  - On exposed nodes (security issues)
  - For uniformity of support from diverse SW suppliers
- Still large demand for OS variety
  - Avoid problem by offering all on same machines: CHOS
    - But does this encourage slow porters?
  - UML

- [476] CHOS, a method for concurrently supporting multiple operating system

Storage Stack

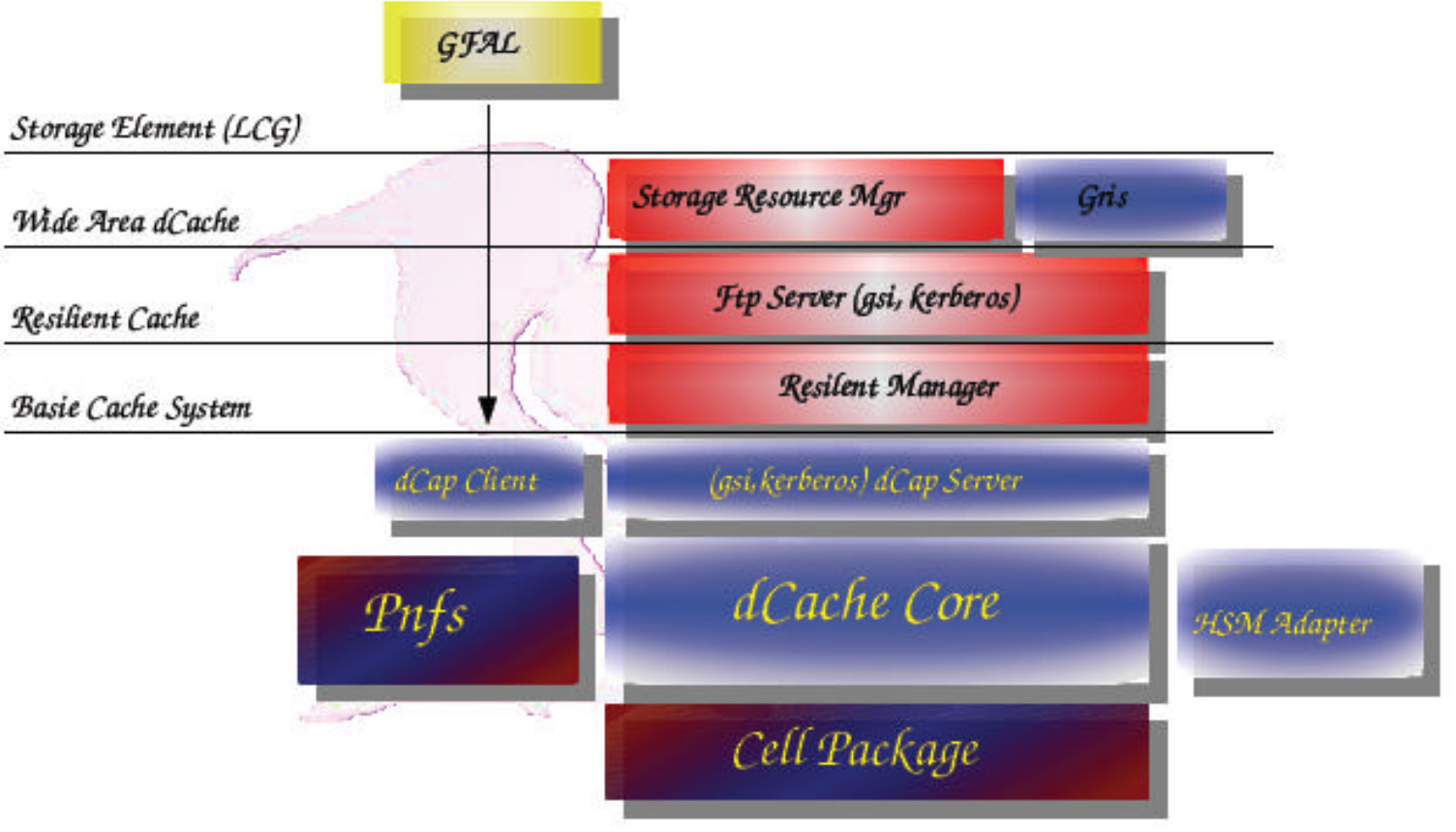


Storage Observations (I)

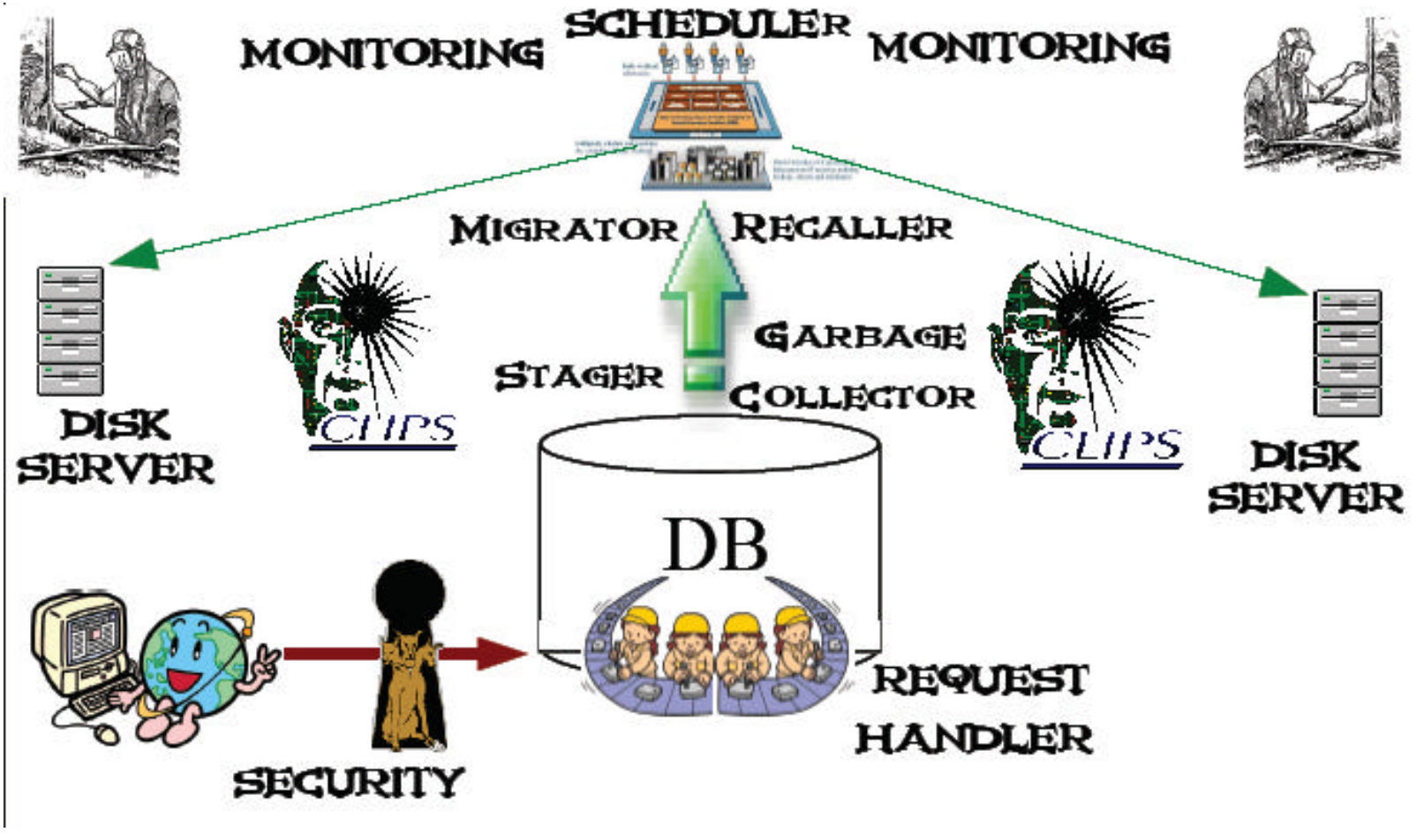
- CASTOR and dCache are in full growth
  - SW developments
  - Growing numbers of adopters outside the development sites
- SRM proliferating to support all major storage managers
  - SRB at Belle

- [230] CASTOR: Operational issues and new Developments
- [233] dCache, Grid Storage Element and enhanced use cases
- [107] Storage Resource Manager
- [216] SRB system at Belle/KEK

dCache


GFAL
Storage Element (LCG)
Wide Area dCache
Resilient Cache
Basie Cache System
Storage Resource Mgr
Gris
Ftp Server (gsi, kerberos)
Resilent Manager
dCap Client
(gsi,kerberos) dCap Server
Pnfs
dCache Core
HSM Adapter
Cell Package

CASTOR


MONITORING
SCHEDULER
MONITORING
MIGRATOR
REGALLER
GARBAGE
COLLECTOR
STAGER
DISK SERVER
CLIPS
CLIPS
DISK SERVER
DB
REQUEST HANDLER
SECURITY

Storage Observations (II)

- Not always going for largest disks (capacity driver), already choosing smaller for performance
  - Key issue for LHC
- Cluster file system comparisons
  - SW based solutions allow HW reuse

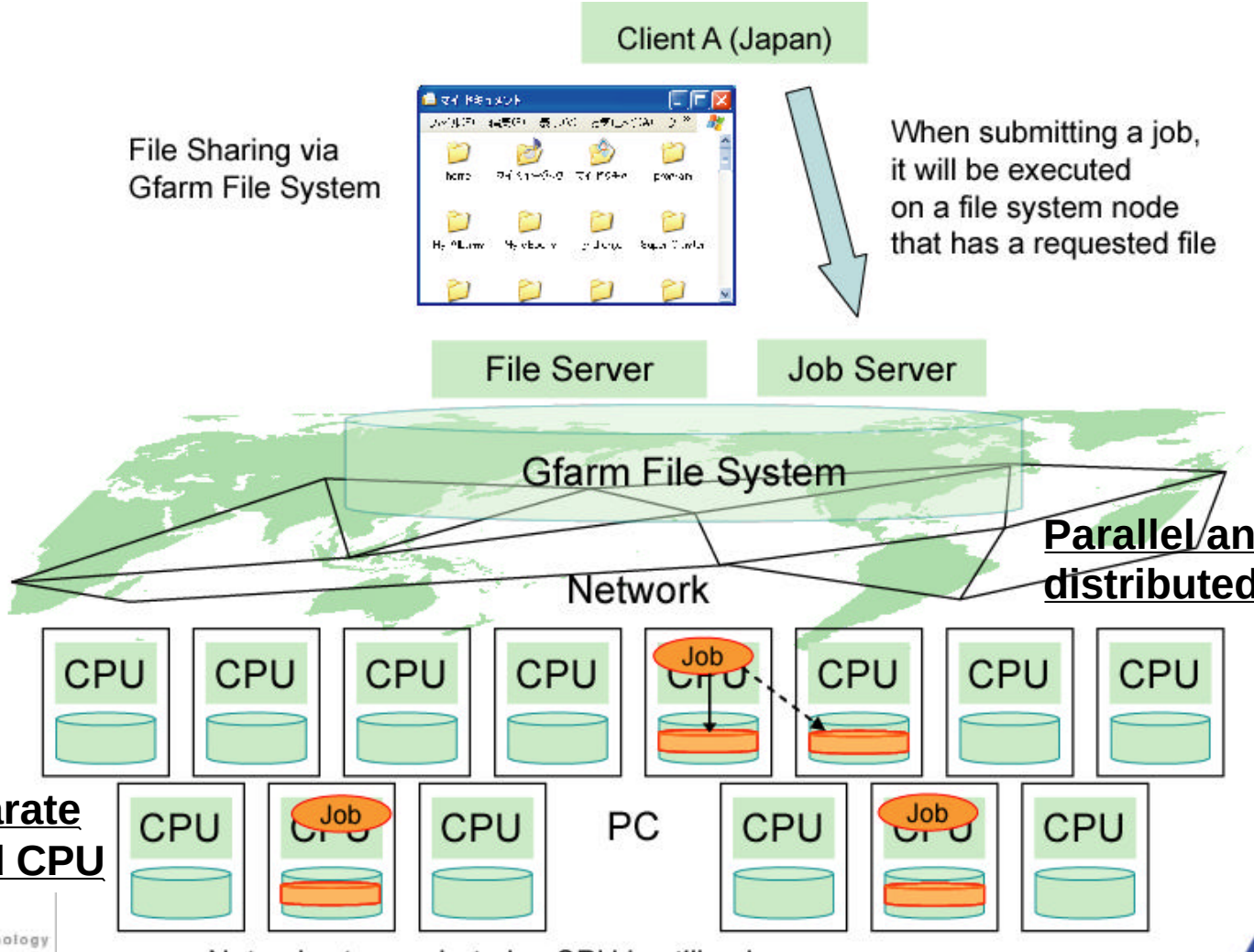
- [325] Disk storage technology for the LHC T0/T1 centre at CERN
- [72] Performance analysis of Cluster File System on Linux
- [187] Distributed Filesystem Evaluation and Deployment at the US-CMS Tier-1 Center

Architecture choices

- Balance of CPU to disk resources
- Security issues
  - Which servers exposed to users
  - Which servers exposed to WAN
- Separate or join compute and storage functions?
  - Destructive or economic to ask a busy CPU node to supply data elsewhere?
- Scale -> Cost factors drive choices
  - Move away from home grown solutions (sched/moni)
  - Move away from LSF as farm grows

- [496] Developing and Managing a large Linux farm -- the Brookhaven Experience
- [330] A Regional Analysis Center at the University of Florida

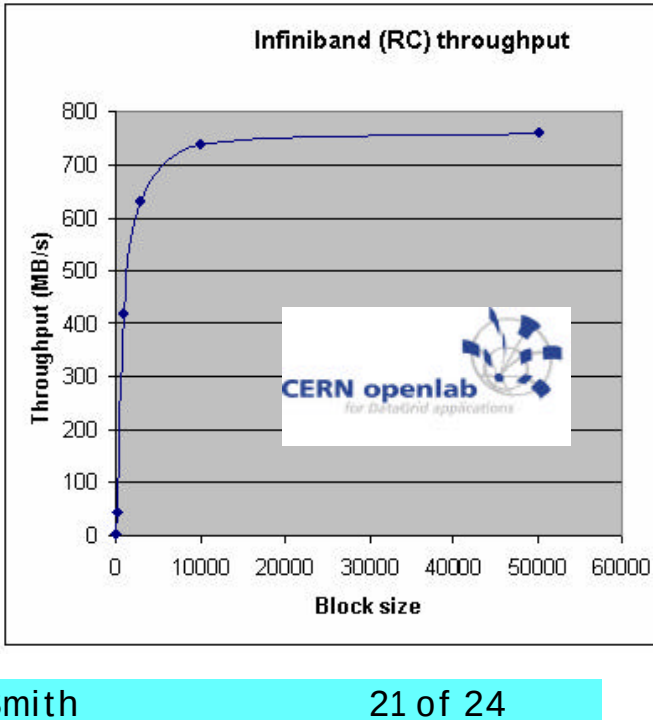
High-performance data access and computing support



Infiniband

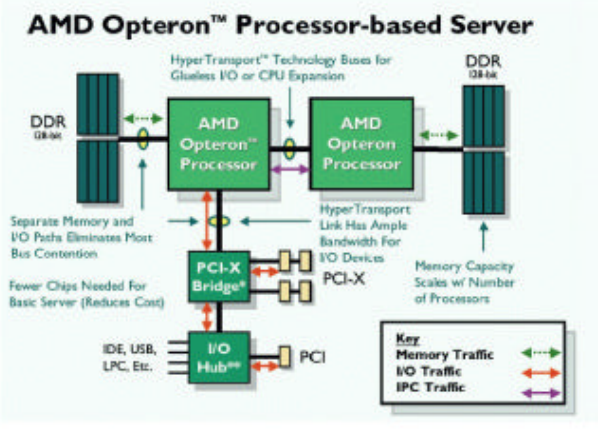
- Computational power increased faster than network bandwidth + growing data needs
- Low latency high bandwidth for HPC (open standards)
- Scalable, redundant network for data intensive computing
  - High data rate at low CPU
  - Ported RFIO, next ROOTIO
- Costs started near FibreChannel, now falling rapidly
  - On motherboard in 2005?

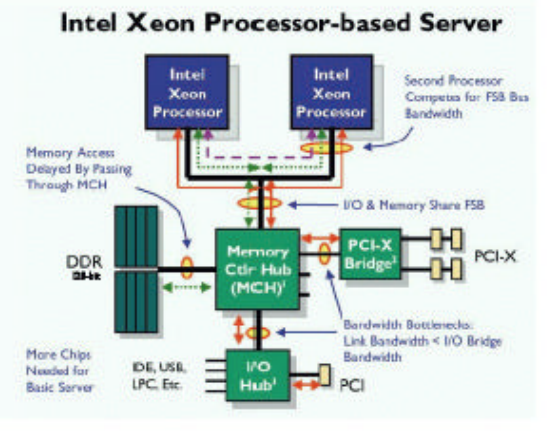
- [346] Lattice QCD Clusters at Fermilab
- [239] CERN's openlab for Datagrid applications
- [487] InfiniBand for High Energy Physics



64 bits (I)

- Debate:
  - AMD64 vs EM64T (vs IA64)
    - IBM Power, MAC G5, Sun SParc





  - Raw performance comparisons, scalability CPUs
  - Compiler choice
- [138] 64-Bit Opteron systems in High Energy and Astroparticle Physics
- [237] Future processors: What is on the horizon for HEP farms?

64 bits (I I)

- No debate:
  - 64 bits are coming soon and HEP is not really ready for it!
  - The writing is on the wall:
    - AMD/Intel stopped production of pure 32 bit chips
    - Memory explosion on farm nodes; 512MB/2002, 1GB/2003, now 2GB … but 4GB is the limit to pure 32-bit memory page addressing
  - HEP ready to profit from the new features?
    - *"Don't wait until it is critical to sort out your ints, longs and pointers"*
  - Must avoid wasteful compatibility modes (which condone slow code migraters)

Concluding Remarks

- *"Organised and systematic approach is the key thing in systems administration"*
- Already demonstrating the ability to manage LHC scale resources
  - Installation, tracking, support model
- Facing the same problems, but with different constraints and resources
  - Some sharing and collaboration, but Norman's 1983 observation still rings true:
    - "Don't do it better, do it the same!"