



# CHEP'10 Report

Armando Fella  
XV SuperB Workshop, Caltech 14-17 December, 2010

# Layout

- CHEP10 generalities
- Worldwide LHC Computing Grid
- Networking
- Many core hardware and software
- Data access, memory supports in HEP
- Cloud computing
- RestFull systems, Frontier
- Conclusion and references

# CHEP 2010

- SuperB presentations
  - “Fast Simulation for SuperB” (D. Brown, poster)
  - “Distributed Production System for SuperB” (A. Fella, poster)
  - “Computing for Flavor Factories” (A. Fella, plenary)
- Parallel streams I covered
  - Distributed Processing and Analysis
  - Grid and Cloud Middleware
  - Computing Fabrics and Networking Technologies

# Factoids

- Hosted by Academia Sinica Grid Computing (ASGC), Taipei – Taiwan
- ~430 subscribers
- 12 plenary sessions, 7 parallel tracks
- 251 oral presentations, 191 posters
- 7/7 rainy days, 2 typhoons



# Parallel tracks

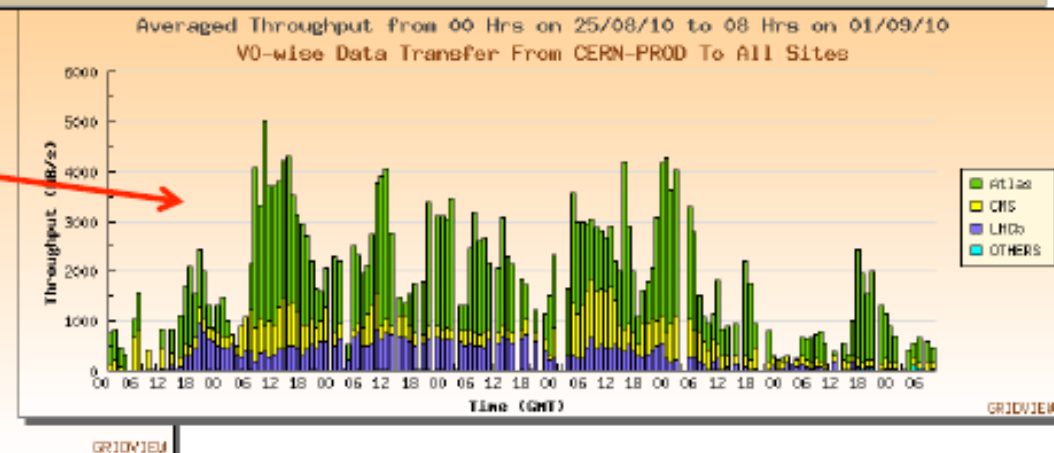
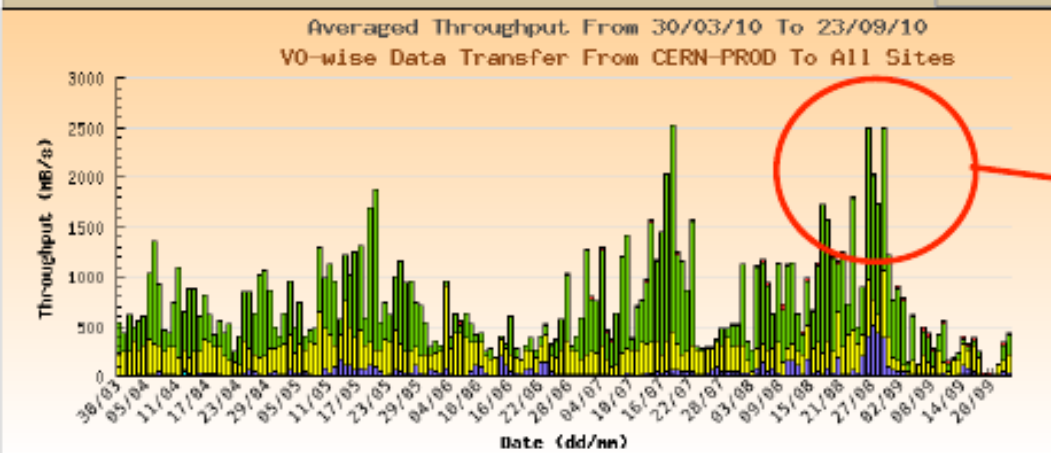
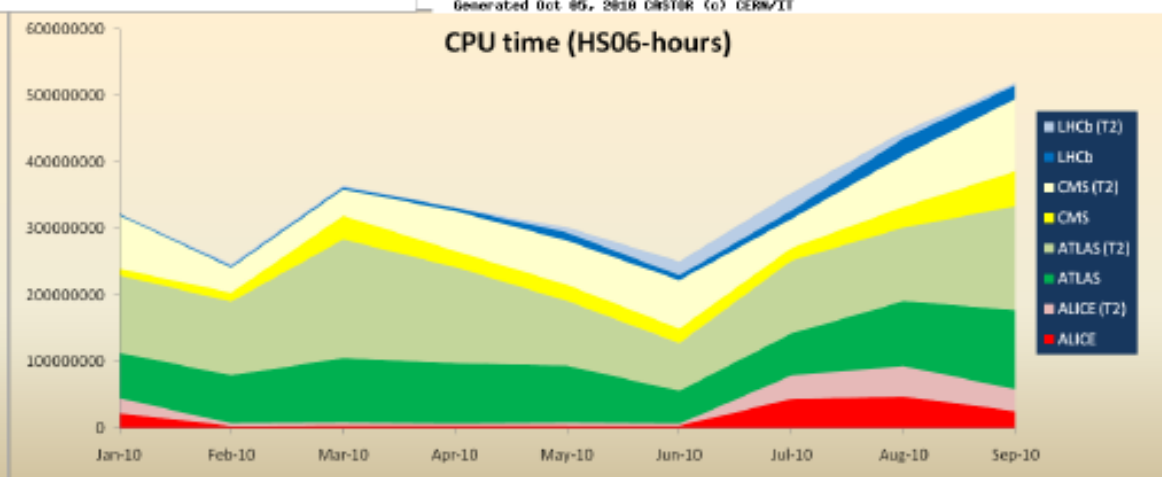
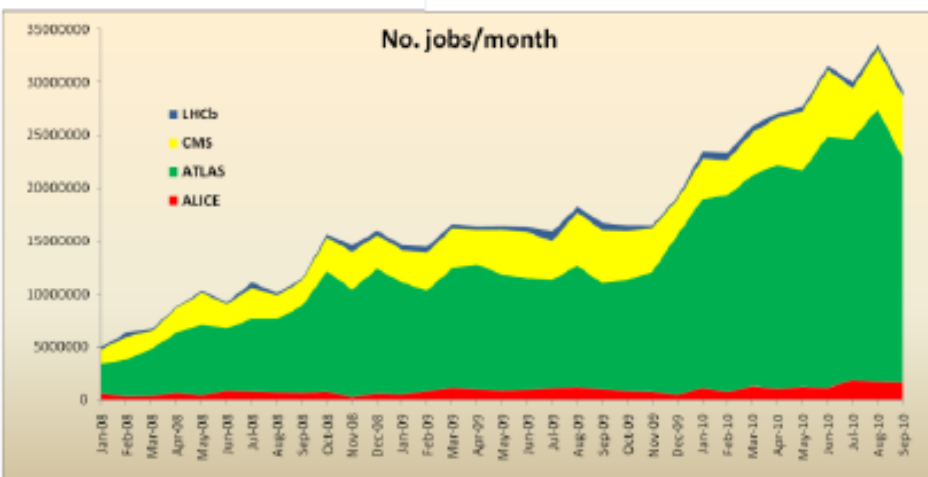
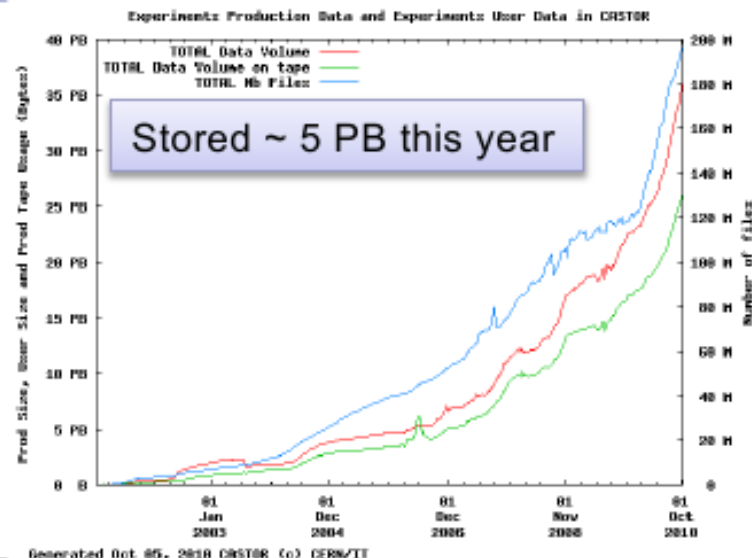
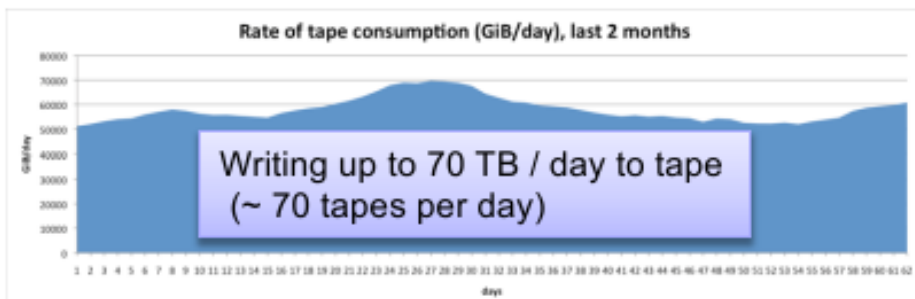
- **Distributed Processing and Analysis**
  - P.Kreuzer (CERN), K.Chen (NTU), F.Rademakers (CERN), T.Wenaus (BNL), I.Fisk (FNAL)
- **Event Processing**
  - F.Cossuti (INFN), R.Itoh (KEK), O.Gutsche (FNAL)
- **Grid and Cloud Middleware**
  - M.Schulz (CERN), A.Di Meglio (CERN)
- **Online Computing**
  - T.Johnson (SLAC), R.Schwemmer (CERN), E.Gottschalk (FNAL), A.Gupta (Jammu U), R.Mommsen (FNAL), N.Katayama (KEK), J.Stelzer (DESY)
- **Computing Fabrics and Networking Technologies**
  - T.Wong (BNL), J.Gordon (RAL), H.Newman (Caltech), T.Cass (CERN), D.Duellmann (CERN), H.Sakamoto (Tokio U)
- **Software Engineering, Data Stores and Databases**
  - M.Cattaneo (CERN), S.Roiser (CERN)
- **Collaborative Tools**
  - P.Galvez (Caltech), M.Lokajicek (FZU Prague)



# WLCG

The Worldwide LHC Computing Grid (WLCG) is a global collaboration of more than 140 computing centres in 34 countries, the 4 LHC experiments, and several national and international grid projects.

Usage consistently high



# WLCG Operations

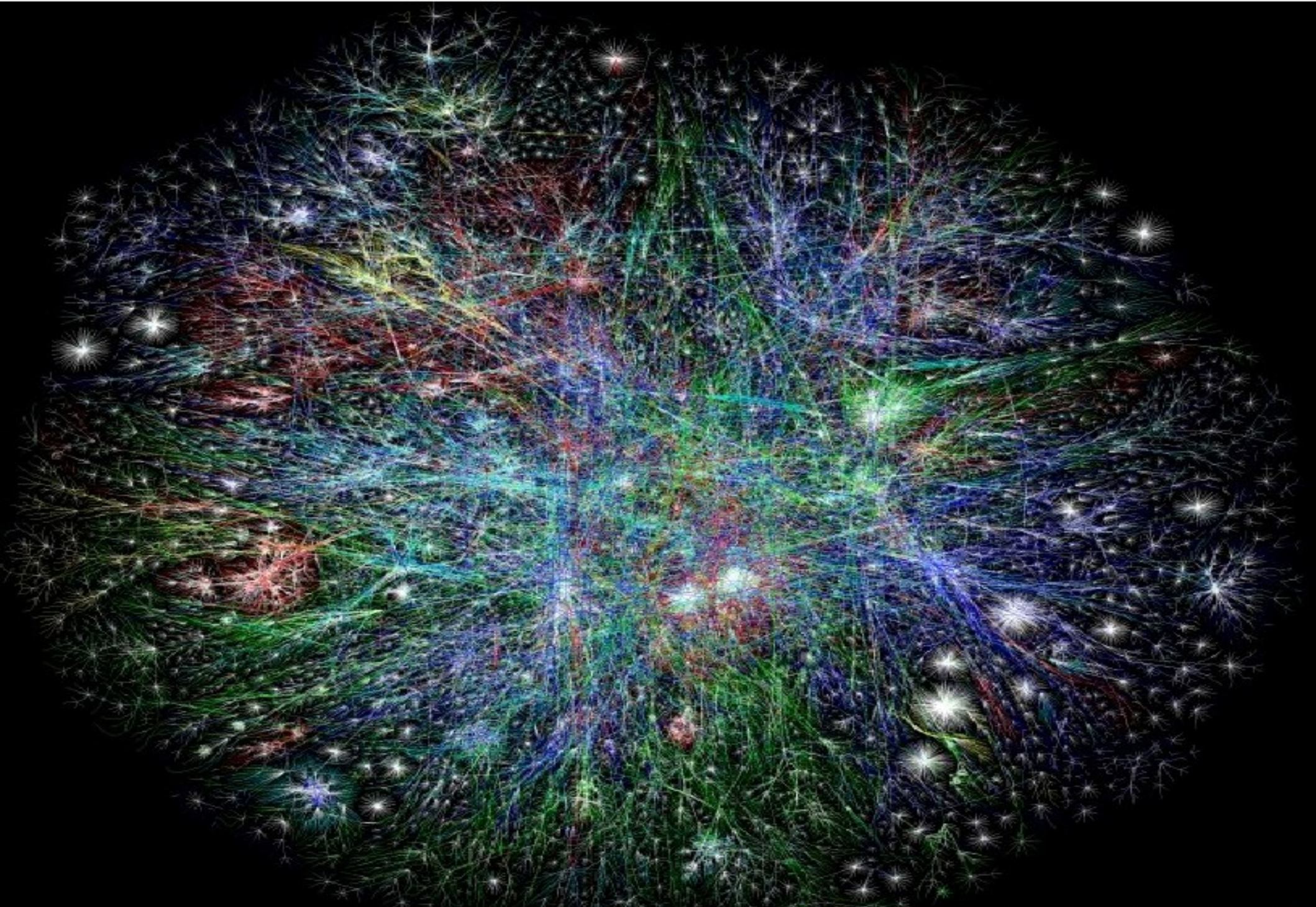
- Availability of grid sites is hard to maintain...
- Hardware is not reliable, no matter if it is commodity or not
  - We have 100 PB disk worldwide – something is always failing
- Problems are (surprisingly?) generally not middleware related ...
- Missing: (global) prioritisation, ACLs, quotas, ...
- Actual use cases today are far simpler than the grid middleware attempted to provide for
  - – E.g. advent of “pilot jobs” changes the need for brokering
- Deployment of upgrades/new services is very slow
- Inter-dependencies application-middleware-OS is a nightmare
- Have we (HEP) really understood how to use a distributed architecture?

- Level of problems is still fairly significant: 5-6 / month require formal analysis
  - Site problems –
    - Power/cooling
    - Hardware issues (and related: procurement delays)
    - Database problems

## Successes:

- ✓ We have a working grid infrastructure
- ✓ Experiments have truly distributed models
- ✓ Has enabled physics output in a very short time
- ✓ Network traffic close to that planned – and the network is extremely reliable
- ✓ Significant numbers of people doing analysis (at Tier 2s)
- ✓ Today resources are plentiful, and no contention seen ... yet
- ✓ Support levels manageable ... just

Providing reliable data management is still an outstanding problem



**The Internet 2009**

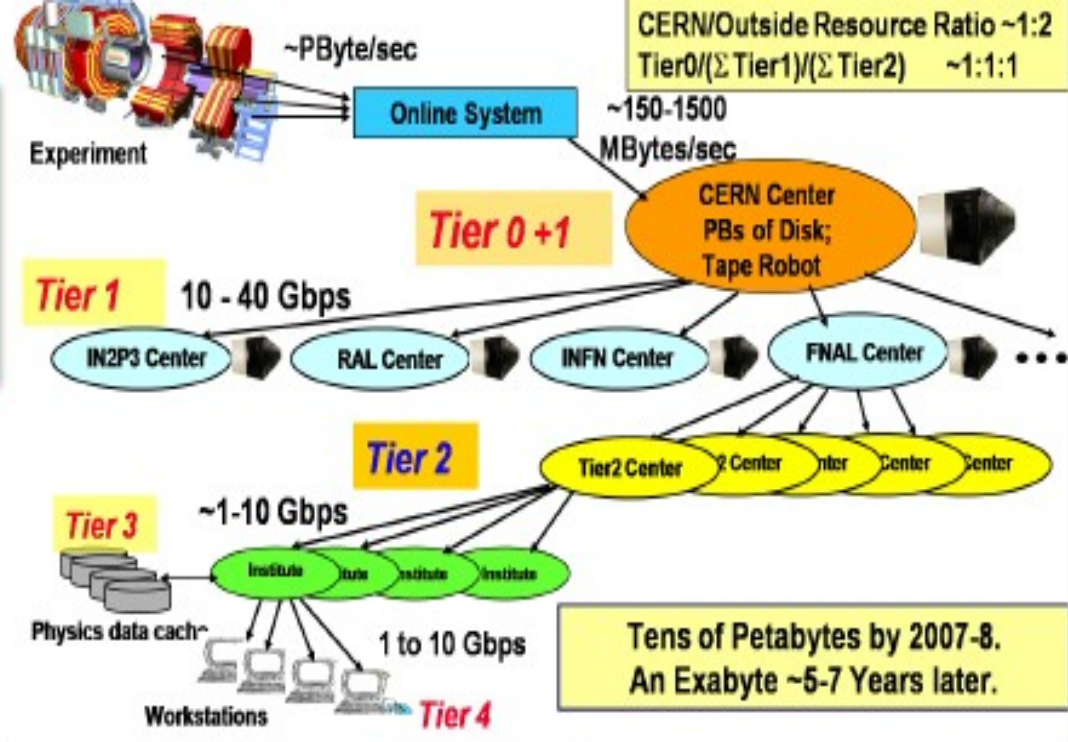




# Networking

The models are based on the MONARC model

Now 10+ years old



## What do we need?

□ **Service:** monitoring, operational support

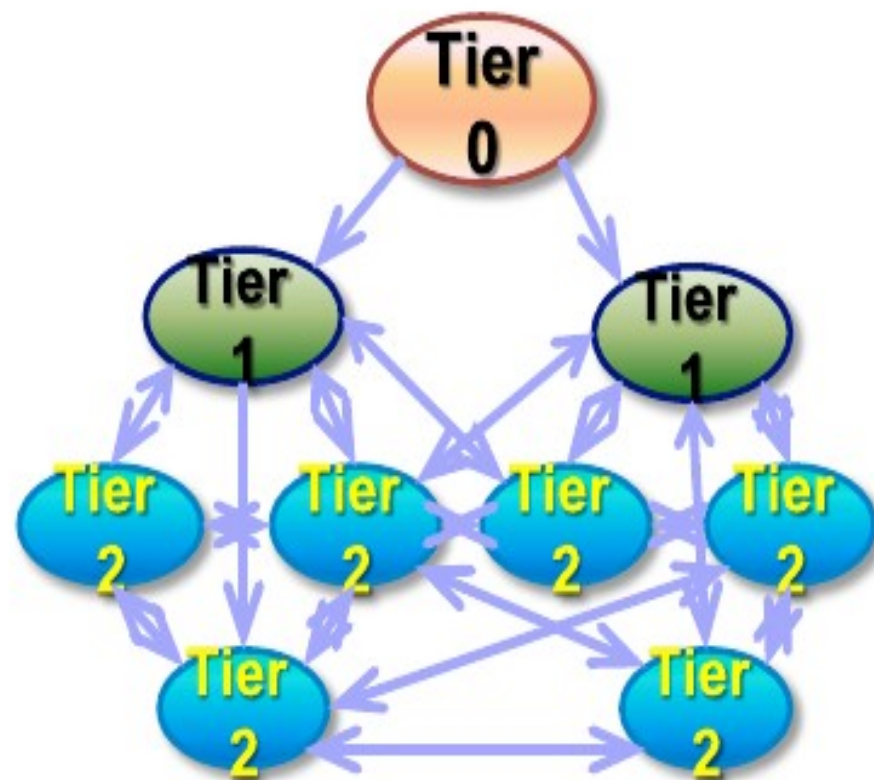
■ **Monitoring:** we have a hard time to understand if there is a network problem.

■ **Support:** Who owns the problem? How can we (user) track progress?

□ **Evolution**

■ 10X in usage every 47 months in ESnet over 18 years

■ Dynamic Circuit Networks



# LHC Experiments' Future Networking Requirements Working Group

- A Requirements Working Group was formed in June 2010 among the experiments and network providers, to investigate future network requirements
  - **Following the Workshop on Transatlantic Networking for the LHC Experiments**
  - **More intensive use of Tier2s & Tier3s; more complex flows. That are proving to be more agile and effective**
  - **Complete a plan in ~2011, in advance of “STEP13” (2012)**
- The new data and computing models incorporate greater reliance on network **performance**; will rely more on network **infrastructure bandwidth and robustness**
- Requirements need to be based on a complete operational model that includes all significant flows across the networks
- Harvey Newman (US LHCNet)      Bill Johnston (ESnet)  
Jerry Sobieski (NORDunet)      Klaus Ullmann (DFN, DANTE)  
David Foster (CERN)      Ian Fisk (CMS)  
Kors Bos (ATLAS, Chair)      Artur Barczyk (US LHCNet)  
Eric Boyd (Internet2)

# Networking for High Energy Physics

Artur Barczyk  
California Institute of Technology  
CHEP 2010 conference  
Taipei, October 19th, 2010



# Dynamic Bandwidth Provisioning



- **Separate high impact data flows from commodity traffic**
- **Create user/experiment specific end-to-end topologies using different technologies**
- **Provide Quality of Service guarantees**
  - Bandwidth, latency, jitter, availability
- **Hybrid: support various technologies**
  - Optical ( $\lambda$ -switched)
  - Packet switched
  - Routed (IP/MPLS, GMPLS)
- **Advance reservation of network resources**



# Summary



- **40Gbps & 100Gbps networking is reality**
  - 40GE switching/routing & NICs beginning to appear
  - 40G continental links and transoceanic links beginning to appear
  - Next generation 40/100G transport technology (OTN) starting to be deployed in backbone and metro networks; 40G in production
  - Widespread production by ~2012 (continental) & ~2013 (transoceanic)
- **New network services become available in advanced NRENs**
  - Dynamic bandwidth allocation
  - End-to-end monitoring
- **Evolving HEP data and computing models can benefit from these developments**
  - If taken into account in the overall designs, and integrated into the data management software
- **Applications capable of using these services are emerging**
  - FDT
  - LambdaStation, Terapaths
  - StorNet, ESCPS

NRENs - National Research and Education Networks

# Evaluating the Scalability of HEP Software and Multi-core Hardware

S. Jarp, Alfio Lazzaro, J. Leduc, A. Nowak  
CERN openlab

International Conference on Computing in  
High Energy and Nuclear Physics 2010  
(CHEP2010)

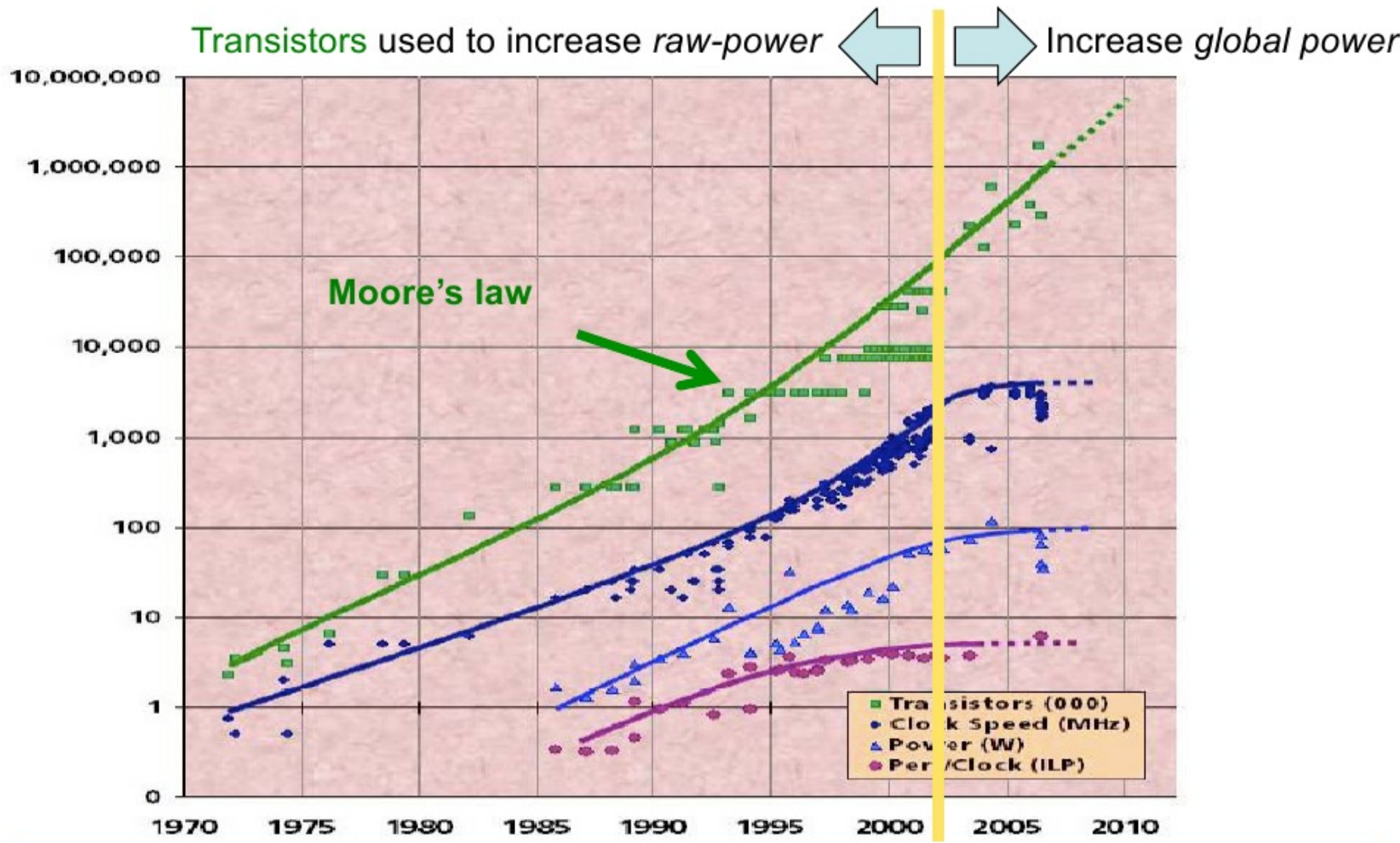
October 18<sup>th</sup>, 2010

Academia Sinica, Taipei



Presentation on behalf of A. Nowak

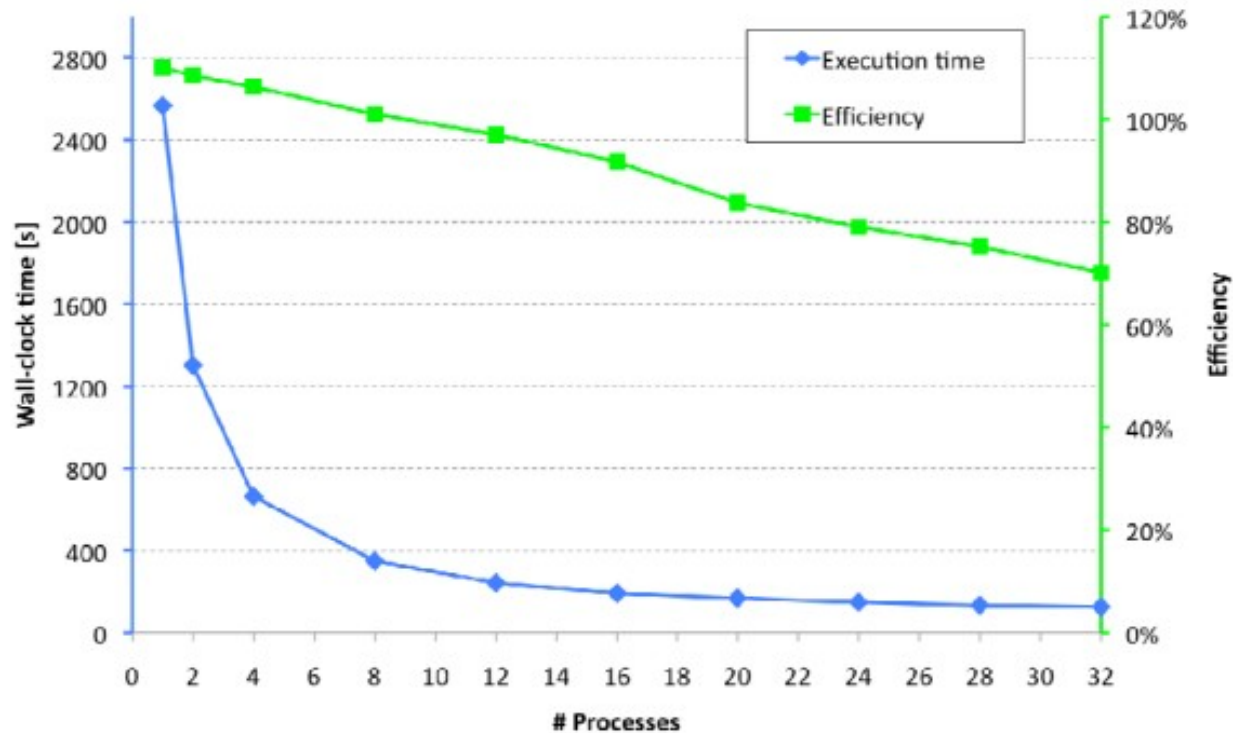






- Currently available nodes with up to 8 cores (4-cores dual-socket)
  - Soon this number will increase up to 48 cores
- Poor usage of multi-threading software
  - A machine with  $N$  cores is considered as  $N$  independent slots for  $N$  independent applications
  - No shared memory among the applications on the node
    - Memory usage increases linearly with  $N$ !
- Poor usage of hardware multi-threading (SMT), usually switched off by default
  - Current CPU can handle 2 hw-threads per core
  - For sequential applications the benefit of the SMT (10% - 30%) is small if compared to memory requirement (100% more memory required), but it is compute power for free in case of parallel applications!

1. **HEPSPEC06 performance**
  - a standard HEP benchmark
2. **Multi-threaded Geant4 prototype scalability** (J. Apostolakis et al, *Multithreaded Geant4: Semi-automatic transformation into scalable thread-parallel software*, Europar 2010)
  - parallel implementation of the test40 example from Geant4
    - 200 random events per thread
  - ParFullCMSmt, a full CMS simulation ported to a parallel model
    - 100 pi- events per thread @ 300 GeV
3. **MPI Parallel Maximum Likelihood (ML) fit with ROOT/RooFit** (A. Lazzaro and L. Moneta, *MINUIT package parallelization and applications using the RooFit package*, *J. Phys.: Conf. Ser.* **219** 042044)
4. **Power consumption vs performance**
5. **NUMA aspects (Nehalem-EX)**

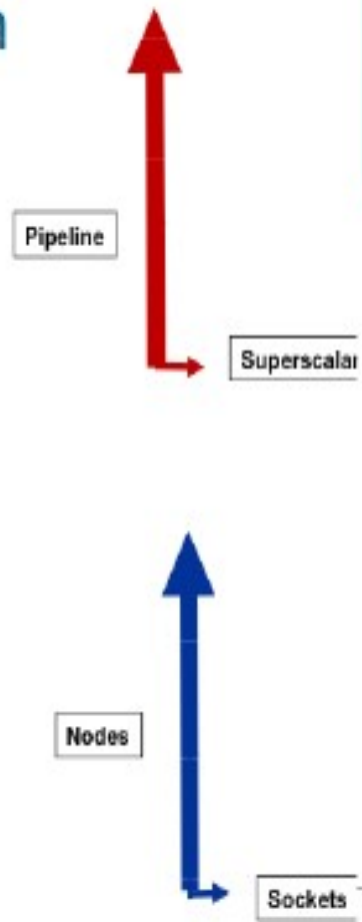


- Strong scaling:
  - fraction of execution time spend in code we can parallelize is 98.7%
  - Scaling as predicted by Amdahl's law
- Test done with Turbo Mode on
  - Efficiency calculated wit respect to 1 process with Turbo Mode off

# Hardware/CPU

## In the days of the Pentium

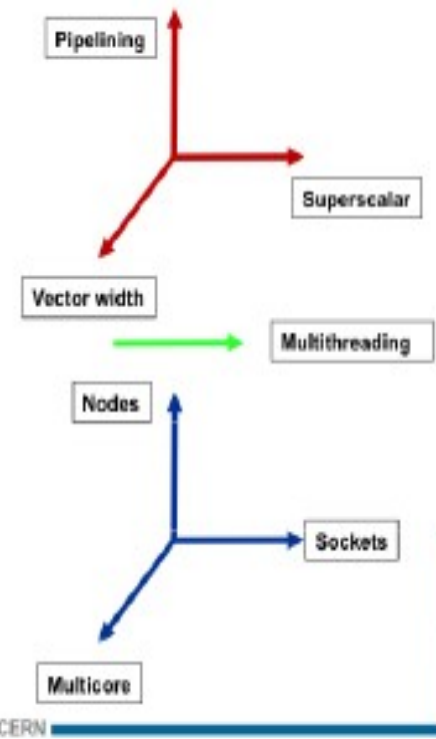
- Life was really simple:
  - Basically two dimensions
    - The frequency of the pipeline
    - The number of boxes
  - The semiconductor industry increased the frequency
  - We acquired the right number of (single-socket) boxes



CHEP 2010, Taipei  
Today:  
Seven dimensions of multiplicative performance



- First three dimensions:**
  - Pipelined execution units
  - Large superscalar design
  - Wide vector width (SIMD)
- Next dimension is a “pseudo” dimension:**
  - Hardware multithreading
- Last three dimensions:**
  - Multiple cores
  - Multiple sockets
  - Multiple compute nodes



SIMD = Single Instruction Multiple Data

Sverre Jarp - CERN

- We are being “drowned” in transistors
  - More and more complex execution units, with 100s of new instructions; longer SIMD vectors; large number of cores; more hardware threading

How to profit? **Think parallel:**

- Data parallelism
- Task parallelism

# Data Preservation in HEP

*To Whom it may concern,*

*In the tape storage area we still have 4132 tapes of type 3840 containing HERA data.*

*We do not have a functioning reading device anymore and the storage area was polluted recently, so it is likely that the tapes are damaged.*

*Would you like us to send you these tapes or should we **destroy them directly**?*

*Yours Sincerely,*

*Tape admin. service [a large computing centre]*

*“We cannot ensure data is stored in file formats appropriate for long term preservation.*

*“We cannot ensure those data are still usable. The software for exploiting those data is under the control of the experiments.*

*“We are sure most of the data are (not easily) accessible!”*

- HEP has little or no tradition or clear current model of long term conservation of data in a meaningful and useful way
- It is likely that most older HEP experiments have in fact simply lost the data (In some cupboard, in the basement, or simply trashed...)
- The preservation of and supported long term access to the data is generally not part of the planning, software design or budget of a HEP experiment

# DPHEP: International Study Group on Data Preservation



**DPHEP** Study Group for Data Preservation and Long Term Analysis in High Energy Physics

- > Chair: **Cristinel Diaconu (DESY/CPPM)**
- > Working Groups
  - Physics Cases: **François Le Diberder (SLAC/LAL)**
  - Preservation Models: **D. South (DESY), Homer Neal (SLAC)**
  - Technologies: **Stephen Wolbers (FNAL), Yves Kemp (DESY)**
  - Governance: **Salvatore Mele (CERN)**
- > International Steering Committee
  - Participants from ee, ep and pp collider experiments
  - Associated computing centers at the labs
  - Some funding agencies
- > International Advisory Committee
  - Chairs: **Jonathan Dorfan (SLAC), Siegfried Bethke (MPIM)**
  - Advisers: **Gigi Rolandi (CERN), Michael Peskin (SLAC), Dominique Boutigny (IN2P3), Young-Kee Kim (FNAL), Hiroaki Aihara (IPMU/Tokyo), Alex Szalay (JHU)**

- > Group has grown since 2008 to around 100 contact persons
- > Endorsed by ICFA summer 2009

# Storage Management

- Big disk farms
- Organized Processing: **GEMSS explicitly mentioned and shown as an example 😊**
- From a data access perspective in 2010 data available over the **network** from a disk at a remote site may be closer than data on the local tape installation
- Work must be done on **access** and **management**

- LHC is no longer talking about 10% disk caches

	ALICE	ATLAS	CMS	LHCb
T0 Disk (TB)	6100	7000	4500	1500
T0 Tape (TB)	6800	12200	21600	2500
T1 Disk (TB)	7900	24800	19500	3500
T1 Tape (TB)	13100	30100	52400	3470
T2 Disk (TB)	6600	37600	19900	20
Disk Total (TB)	20600	69400	43900	5020
Tape Total (TB)	19900	42300	74000	5970

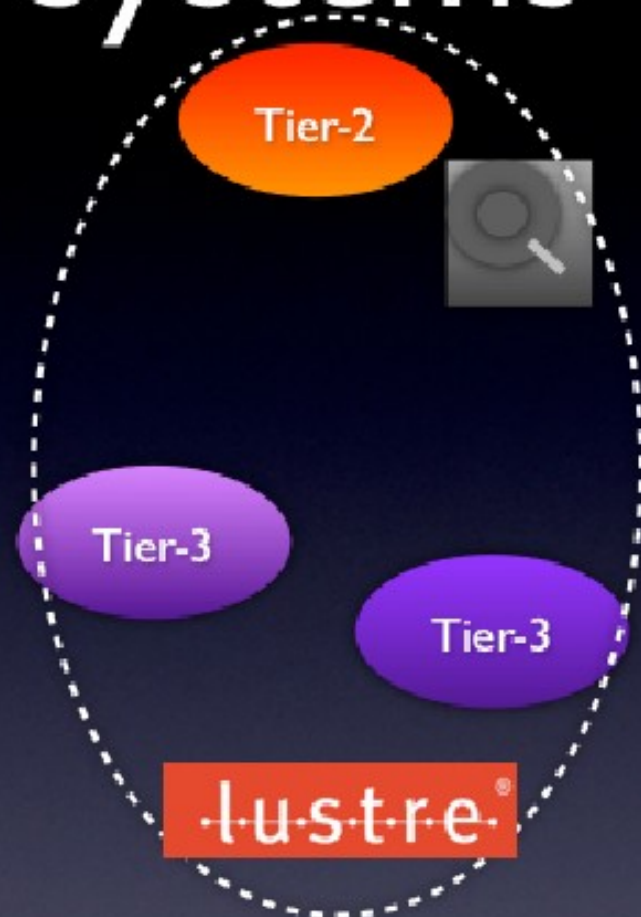
DZero	CDF
~500	~500
5900	6600

- In 2011 majority of the currently accessed data could be disk resident

- Analysis or skimming, are they substantially different from video streaming, once you have streams of objects and optimized I/O?  
→ web delivery of content

# Wide Area File Systems

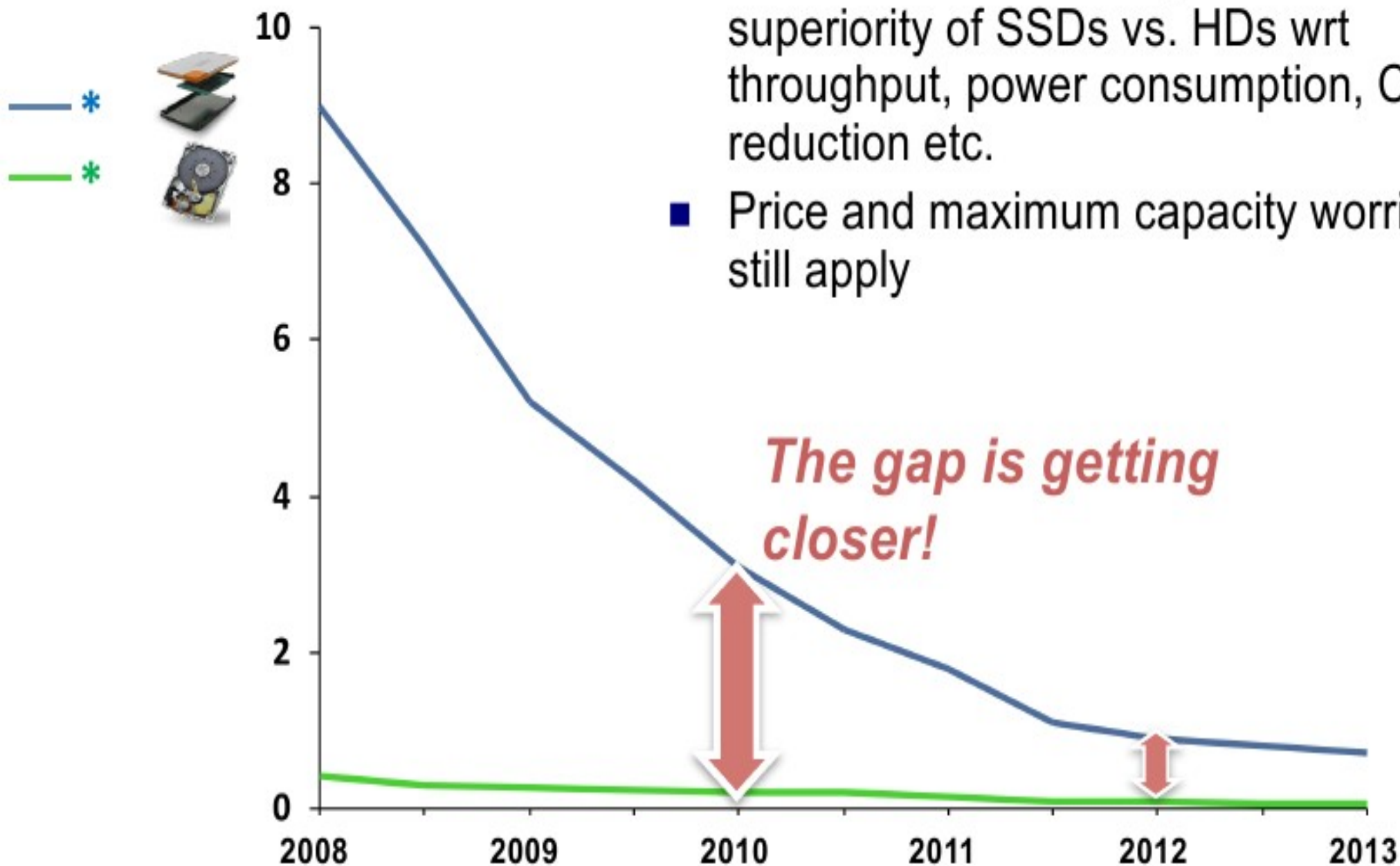
- ➔ Florida, FNAL, OSG, and TeraGrid have been working with wide area Lustre
  - Successfully demonstrated to serve data to Tier-3s in a geographical area
  - Wide area deployment for TeraGrid
  - Not yet at extremely large disk capacity
- ➔ Interesting technique that would simplify data management
- ➔ Wide Area NFS4 may have similar functionality





# SSDs

\$USD / per GB



- A lot of data shown to prove superiority of SSDs vs. HDDs wrt throughput, power consumption, CO<sub>2</sub> reduction etc.
- Price and maximum capacity worries still apply

*The gap is getting closer!*

# Cloud computing/virtualization

- Platform ISF adaptive cluster
- Open Nebula
- Amazon EC2
  - Poster by Vanderbilt U, about cloud CMS use
  - See talk about security problem in EC2
- Nimbus
  - See Sotomayor, Montero, Llorente, Foster, "An Open Source Solution for Virtual Infrastructure Management in Private and Hybrid Clouds", IEEE 2009, Special Issue on Cloud Computing for a comparison between Nimbus, OpenNebula, Eucalyptus et al.

Site	Events/ Core-Hour	USD/Core- Hour	USD/1000 events
Vanderbilt	244	\$0.024	\$0.098
Amazon	208	\$0.17 (varies)	\$0.817

## ■ Blessings:

- Deploy custom, user-owned and user-controlled environments on remote resources
- On-demand access
- Elastic processing
- Growth and cost management
- Capital expense -> operational expense

## ■ Challenges (some of)

- Appliance management
- Lack of reliability
- Elasticity, but how?
- Performance of deployment and runtime
- Cost

# Scaling HEP to Web Size with RESTful Protocols: The Frontier Example

Dave Dykstra, Fermilab

[dwd@fnal.gov](mailto:dwd@fnal.gov)



# REST background

- REpresentational State Transfer
- Defined by Roy Fielding in his PhD dissertation
- General architectural style derived from using a subset of http strictly according to http RFCs
  - Roy was a principal author of http RFCs
- Designed for Internet-sized scaling, that is, primarily for **caching**
  - Also for easy replication of servers
- Good for many purposes beyond web browsing
  - including many HEP tasks that need large scaling

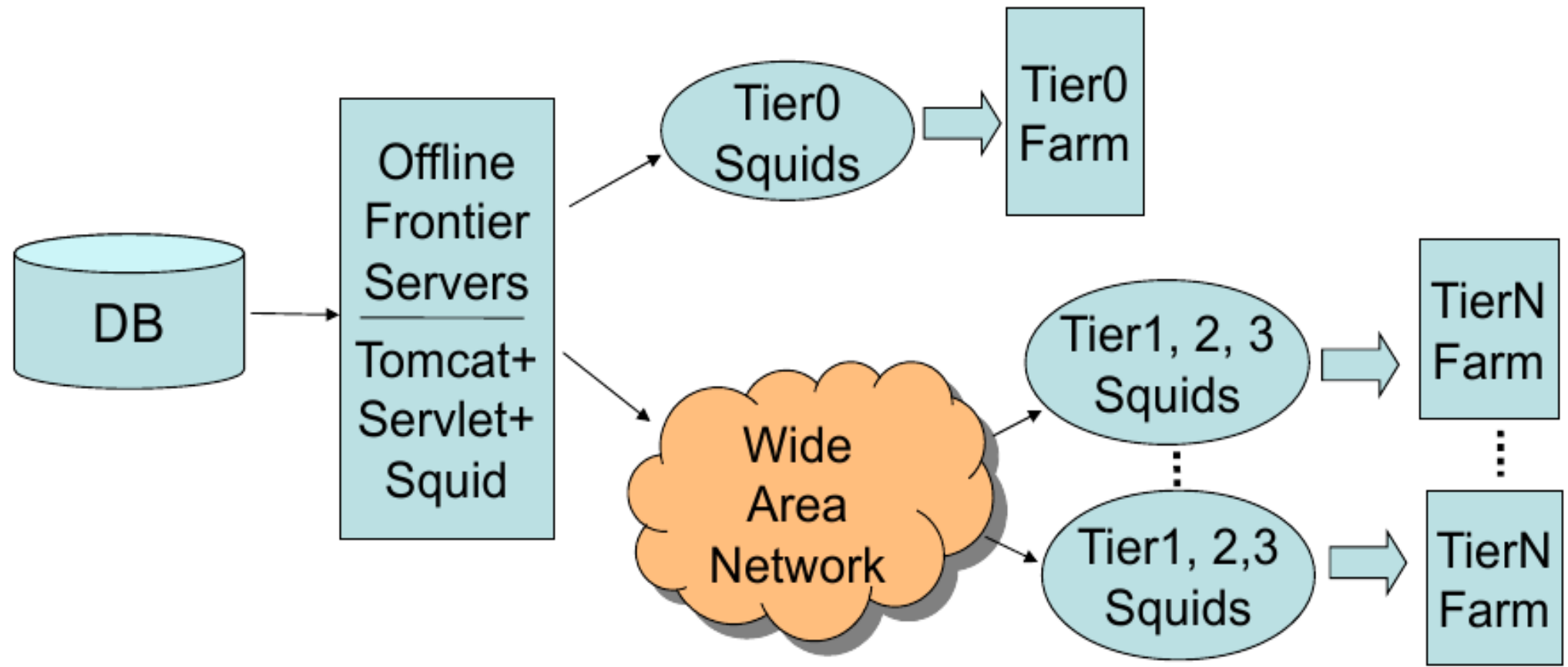


# REST Frontier example

- Distributes read-only database SQL queries
  - Updates are done with a different protocol (like most of the RESTful cacheable systems I have seen)
- Designed for HEP “Conditions” data with many readers of same data distributed worldwide
- Ideal for caching



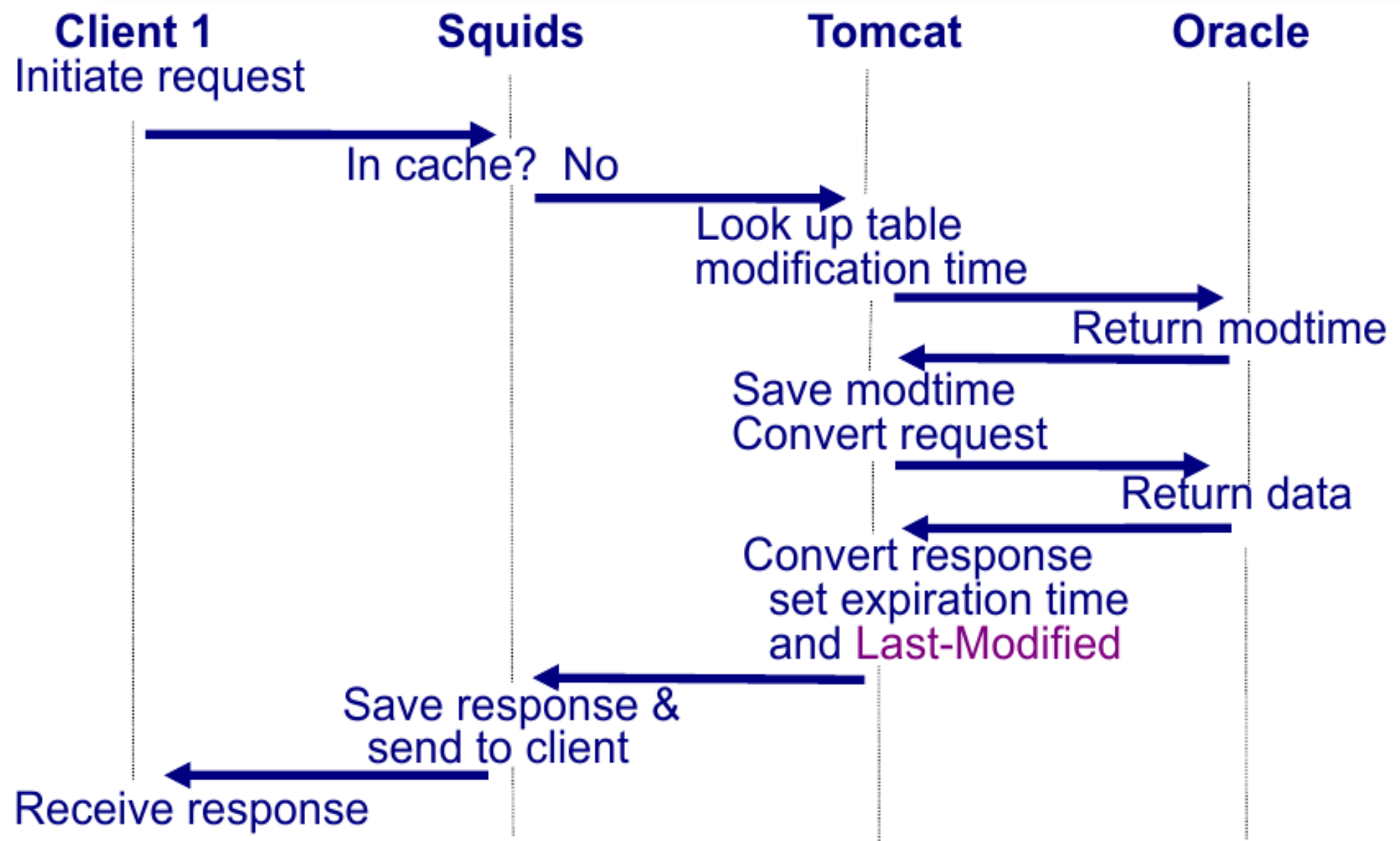
# CMS Offline Frontier example



- Many copies of `frontier_client` in jobs on the farms
- Jobs start around the world at many different times
- Cache expirations vary from 5 minutes to a year



# Filling cache in Frontier





# Summary

- Use REST!
  - Whenever the same information is needed in many places
  - Use locally deployed standard caches
    - Already deployed at most sites participating in LHC experiments



# Conclusions

- Too much interesting materials from chep10
  - Need to make a selection of subjects in main thread lines and try to make them fit in a Computing Model
- Very useful source of expertise
  - Need to contact people, cooperate as possible, participate in key groups
- Thanks to D. Salomoni for the many and excellent stolen slides
- See report from CHEP'10 by D.Brown at II Computing workshop:  
<http://agenda.infn.it/conferenceDisplay.py?confId=3104>

# Many other talks of interest

- “Ten Years of European Grids: What Have We Learnt?” by Stephen Burke
- “Establishing Applicability of SSDs to LHC Tier-2 Hardware Configuration” by Sam Skipsey
- “Computing at Belle II” by Thomas Kuhr
- “Modular software performance monitoring ” by Daniele Francesco Kruse
- “Hepsoft” by Stefan Roiser
- “WNoDeS, a tool for integrated Grid/Cloud access and computing farm virtualization” by A. Italiano and D.Salomoni
- “Reinforcing User Data Analysis with Ganga in the LHC Era: Scalability, Monitoring and User-support” by Johannes Elmsheuser