

# CHEP2010

---

Rain and lecturing in Taipei  
Sam Skipsey

With thanks to Wahid Bhimji and Stuart Purdie for comments on parallel sessions I was unable to attend.

# Itinerary

- What is CHEP?
- Themes overview
  - A Summary of Summaries
  - Common themes
  - Presentation highlights
- Conclusion

# CHEP - 2010

- The big Computational High-Energy Physics Conference
- Hosted by Academia Sinica, Taipei, Taiwan
- Opened by the Vice-President of Taiwan, Vincent Siew (蕭萬長)
  - Good luck getting Nick Clegg in the UK!

# A brief slide about Taipei.

- Capital of Taiwan (Republic of China)
- Pronounced “Taibei” (臺北)



- Has an international baseball team.
- Has tons of Taoist, Buddhist and folk Temples.
- Filled with “Night Markets”
- Oh, and typhoon season is around October...



# A brief slide about Taipei.

- Capital of T
- Pronounced



We can't stop here; this is Typhoon Country!

# An overview of Summaries

- 7 Tracks of Parallel Sessions in total
  - Event Processing; Online Computing; Software Engineering & Data; Distributed Processing & Analysis; Computing Fabrics & Networking Tech.; Grid & Cloud Middleware; Collaborative Tools\*
- Each was summarised in 30 minutes on Friday.
- We'll start with a summary of the summaries:

\*we implicitly continue the agenda of Newton in pretending rainbows have 7 colours...

# Event Processing

- GEANT<sub>4</sub> (simulation)
- FAIR plans (GEANT<sub>4</sub>, event analysis trains)
- Analysis and I/O (GPU, etc)
- Software standardisation
- Alignment and calibration (brilliant pictures)





# Event Processing



## Conclusions

- ▶ LHC:
  - ▶ In general, everything is working remarkably well
  - ▶ Current data taking conditions are under control
  - ▶ The future will show how the experiments will cope with the increasing PileUp conditions
- ▶ Beyond LHC
  - ▶ FAIR experiments with non-traditional beam conditions mark a new frontier of challenges
  - ▶ New experiments rewrite or develop new frameworks, software standardization helps them ramping up more quickly
- ▶ Performance
  - ▶ **Multi-Core, GPUs and Vectorization:** buzz words of the processing world
  - ▶ Applications show significant speed increase, but usage still very dependent on specific situations
  - ▶ New experiments are designing their software for multi-core, multi-thread execution environments and also consider specialized hardware solutions like GPUs

# Online Computing

- All LHC experiments ++good with real data taking. (DAQ eff > 90%)
- Emphasis on storage perf (at To)
- Integration and automation!
- Data quality monitoring - and web access!
- Lessons learned: uniform stack (on and offline)

# Online Computing

## Summary

- The startup of the LHC experiments has been a tremendous success: DAQ efficiencies well over 90% and over all efficiencies for physics well over 80%
- Sophisticated tools for data quality monitoring allow remote and local experts to react and flag quickly
- Modern web-technologies make experiment info available everywhere
- But...
  - The LHC is improving and in some areas we are already beyond initial design parameters
  - Upgrades are coming: more data, new detectors, faster readout & storage

# Software Engineering, Data\*

- Heterogeneity of talks (also mentioned in other summaries)
- multithreading cool
- performance monitoring tools
- lots of “new” cool things: go, svn, html5, CVMFS etc
- software recycling for small exps.
- Data archiving and preservation



# Software Engineering, Data \*

## Conclusions

- The software frameworks for LHC are in very good shape
  - Processes and tools are in place
  - A lot of efforts on the performance side are underway
    - We need more specialists in this area
- Other experiments should be able to profit from the work
  - We need for more collaboration across experiments
  - Implementation islands are getting bigger

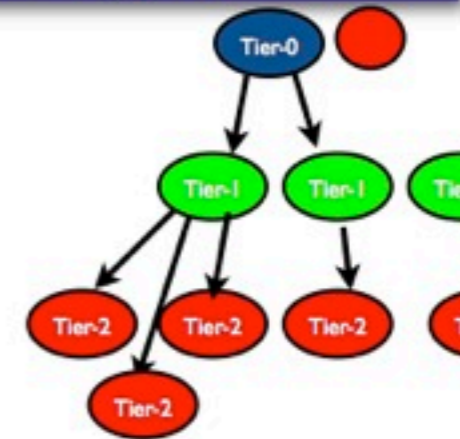
# Distributed Processing and Analysis

- Successes
  - First year LHC stuff, organised processing, phenix, ATLAS DDM
- Future Architectures
  - FAIR, SuperB, Belle2, CDF, Fermi Space telescope
- Improvements
  - Global FS, ARC, VMs,

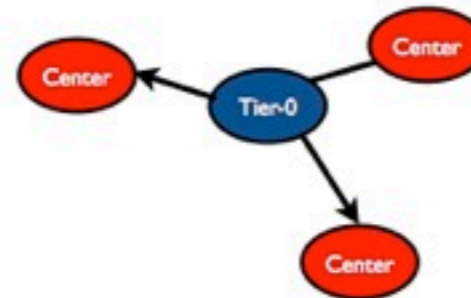
# Distributed Processing and Analysis

## Distributed Processing and Analysis

- ▶ Basically Breaks into 3 main themes
- ▶ Successfully Processing and Analysis in Distributed Environments
- ▶ Start-up of LHC Computing

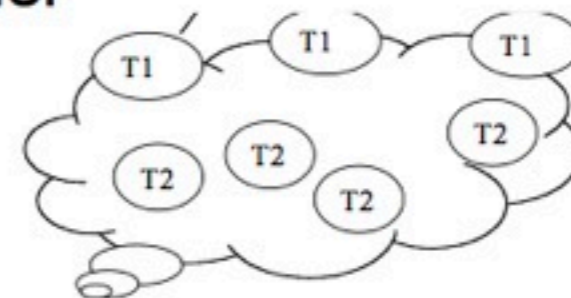


- ▶ Architectures for Future Facilities



- ▶ Improvements in infrastructure and Services for Distributed Computing

- ▶ VMs, Clouds, Storage
- ▶ Monitoring, Simulation, and Infrastructure



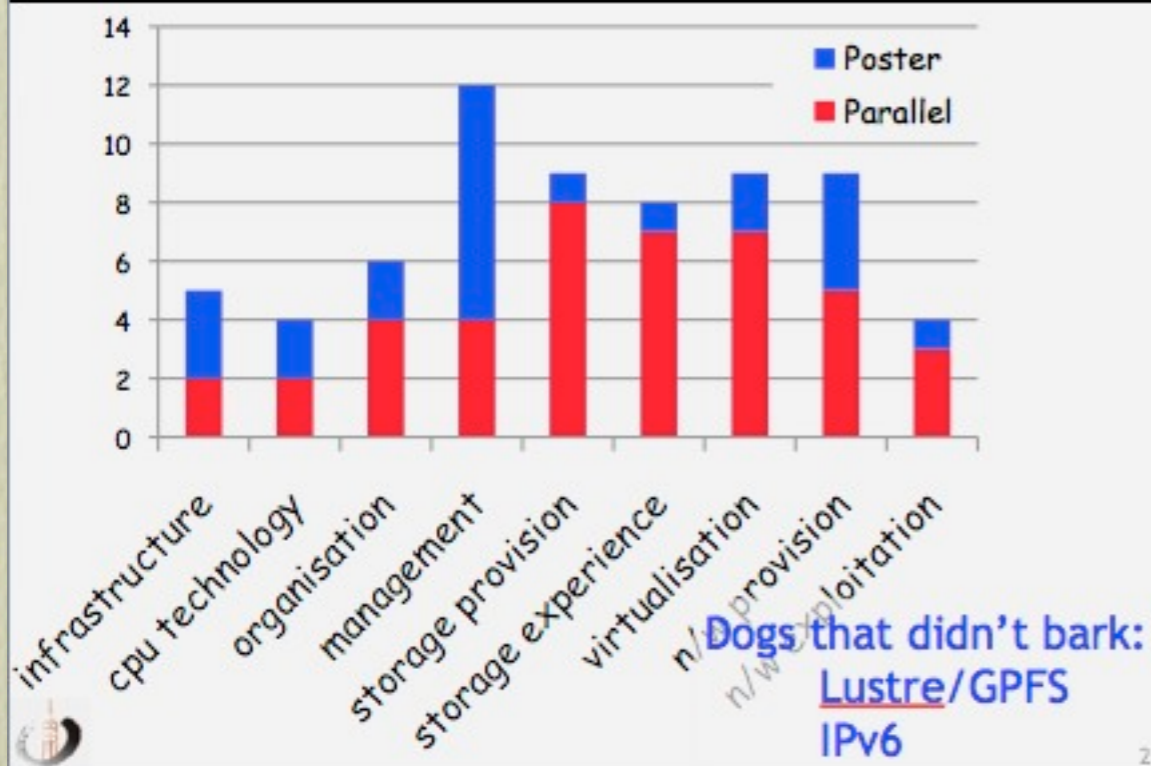


# Computing Fabrics and Networking Tech

- Storage, Storage, vms, management, multicore
- No: lustre (as a sole topic), IPv6
- Fabric management (puppet)
- Data management architecture reworking
- NFs4.1, EOS, CPU scaling, Clouds (expensive)

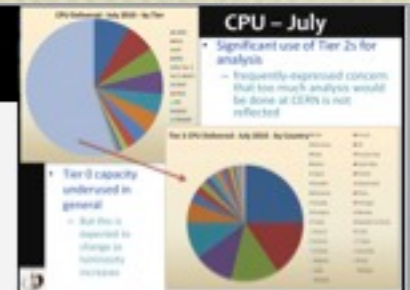
# Computing Fabrics and Networking Tech.

## What was covered?



## Summary Summarised

- ◆ Fabrics working well!
- ◆ Many interesting presentations
- ◆ Well attended
  - Thanks to all those who braved the rain!
- ◆ Virtualisation topics split across 3 tracks
  - Dedicated track for CHEP '12?
  - » or will it all be routine by then?
- ◆ We seem to be addressing many of Ian's concerns but...
  - wheels are often reinvented
  - developments sometimes occur in isolation
- ◆ Still scope for improved collaboration between sites and between different work areas.




• Tony Cass

# Grid and Cloud Middleware




- Operations and Monitoring
- Data Management - EMI plans
- Clouds -HiData (ceph!), Boinc+CVM+CoPilot
- Virtualisation, messaging, integration
- Pilot jobs (improvement of)

# Grid and Cloud Middleware



**Main Topics**


- Operational Experience
- **Operations and Monitoring**
- **Data Management**
  - **THE challenge**
- Workflow Management
- Security
- **Clouds and Virtualization**

Markus.Schulz@cern.ch 3



- Markus Schulz

# Grid and Cloud Middleware




**Cross Topic**

- **Virtualization**
- **Messaging as a foundation technology**
- **Integration and Interoperation**
  - Security, Middleware Stacks
  - Storage, Grids and Clouds
- **Managing change**
  - Coexistence old/new



Markus.Schulz@cern.ch



4

# Collaborative Tools

- Outreach Plenary
- Web2.0 internal/external communications (ATLAS), CMS collab. infrastructure
- Inspire, ATLAS Live, Glance information system
- EVO
- CERN Lecture archiving system,
- AbiCollab (like google Docs)

# Collaborative Tools

## Overview



- **Increasing areas in the CT field**
  - HD videoconferencing systems, Outreach and Inreach activities, Rich Media Content, Information systems, etc.
- **Representative examples covering the activities in the HEP community**
  - 1<sup>st</sup> session dedicated to Policies and New initiatives
  - 2<sup>nd</sup> session dedicated to SW systems and Collaborative Tools
- **Plenary Talk (Lucas Taylor, FNAL/CMS)**
  - Overview about the importance of the outreach activities for the HEP community
    - Contract with the Society
    - Need of a defined strategy: HQ messages, defined relation with the Media and use of latest multimedia technologies
    - Everyone needs to be involved

- Joao Fernandes – CERN, Philippe Galvez – Caltech, Milos Lokajicek – FZU Prague

# So, the summary of summaries:

- The LHC works!
- Data Management is hard
- Multi-core, GPGPU are exciting
- Virtualisation and Clouds are hot topics.
- Talking to people is good.



# Presentation Spotlights

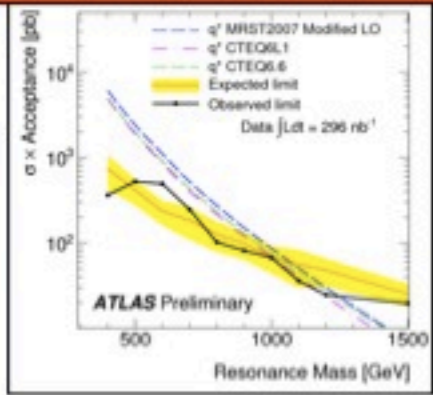
- A selection of the presentations that were most interesting.
- Subjective! Caveat Auditor!

# The LHC works!

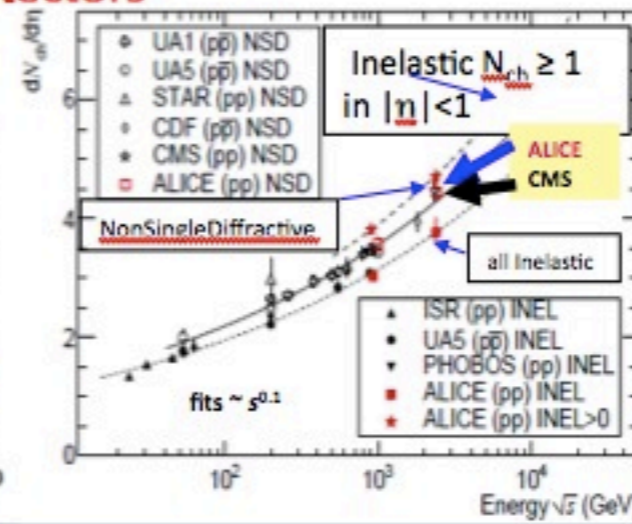
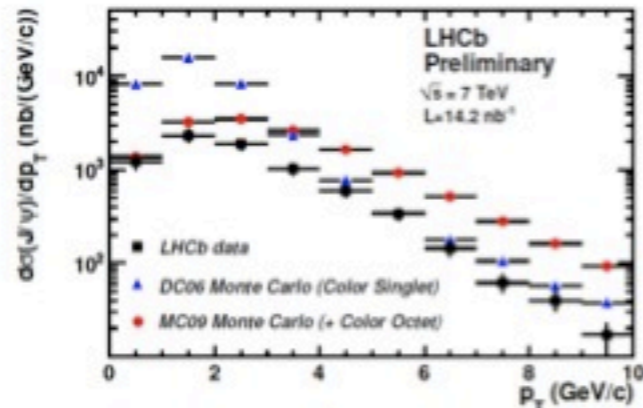
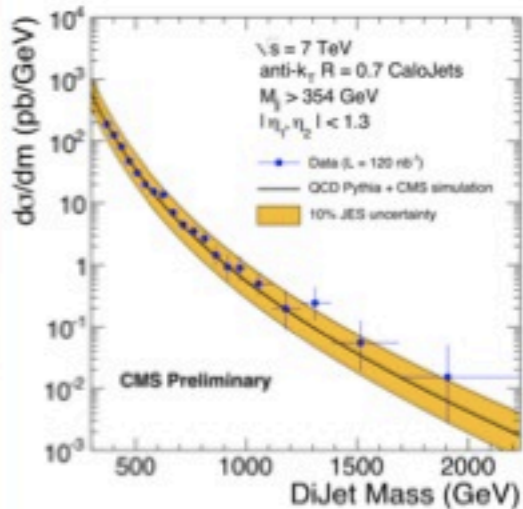


## One Giant Leap.....

$0.4 < M(\alpha^*) < 1.29 \text{ TeV}$  excluded at 95% C.L.



- To begin at the end:
  - The first year of operations has been a great success
    - The experiments produced fantastic results at ICHEP in Paris, often only days after the data was taken
    - The description of the detectors was astonishingly good

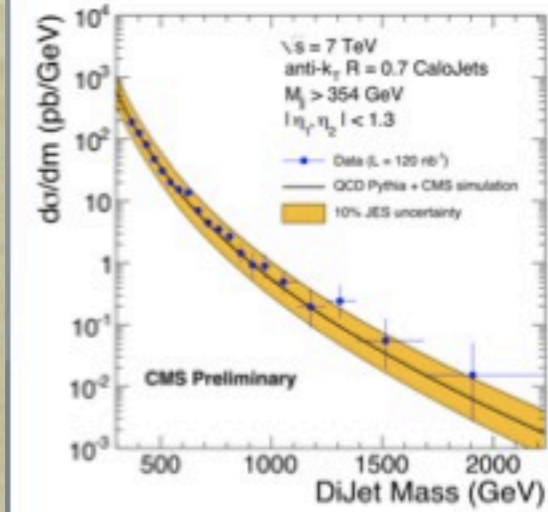
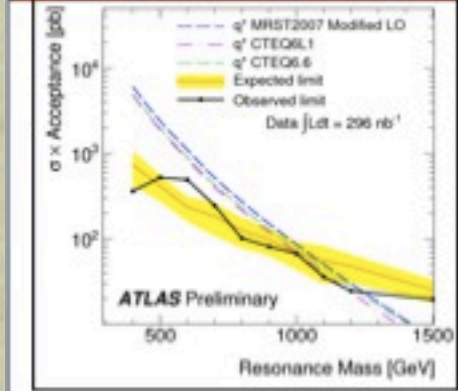


# The LHC works!

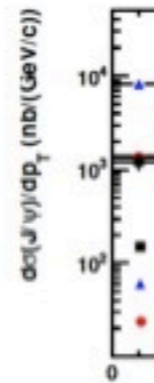
LANCASTER UNIVERSITY 10/18/10 RWL Jones CHEP2010 2

## One Giant Leap.....

0.4 < M ( $\alpha^*$ ) < 1.29 TeV excluded at 95% C.L.

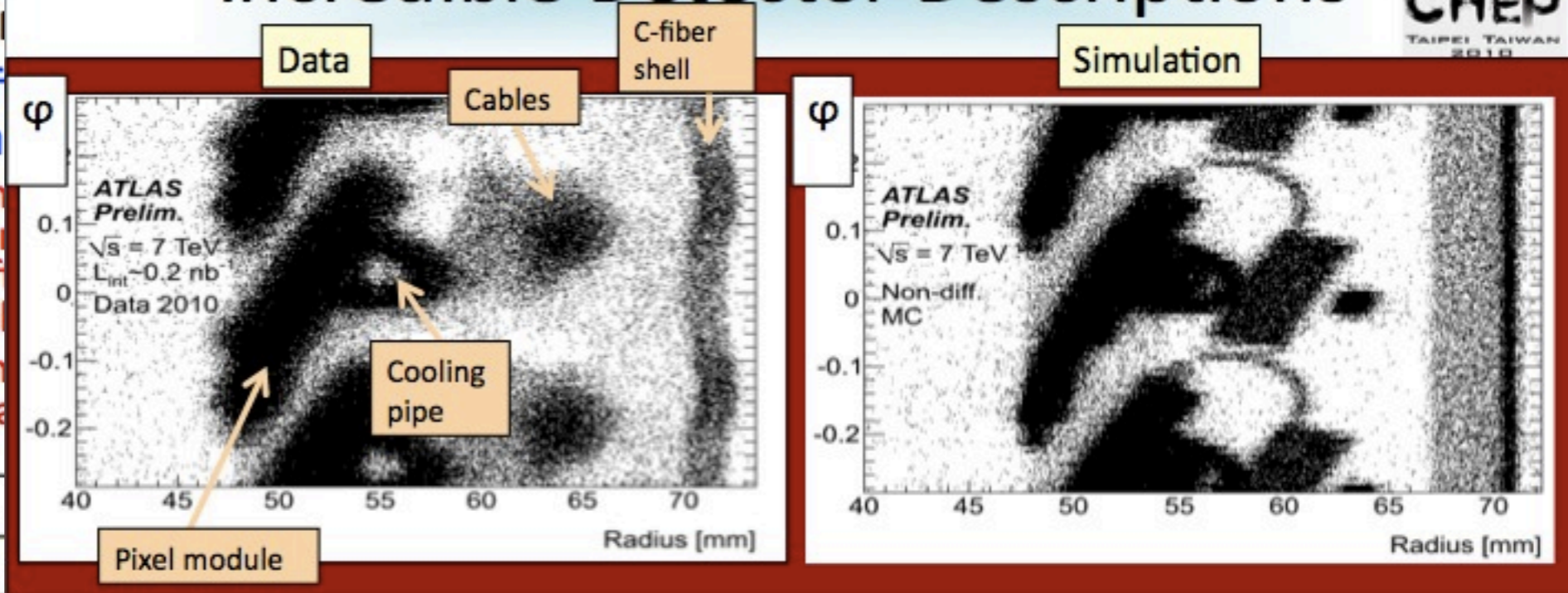


- To begin
- The f
- been
- The fa
- of ta
- The wa



LANCASTER UNIVERSITY 10/18/10 RWL Jones CHEP2010 3

## Incredible Detector Descriptions



- This is partly because of the long lead-time & the unexpected extra year of cosmics!
  - This has had implications in the analysis patterns – more later!
- But there have been lessons to be learned
  - And we have just started on a treadmill, which will require continual development

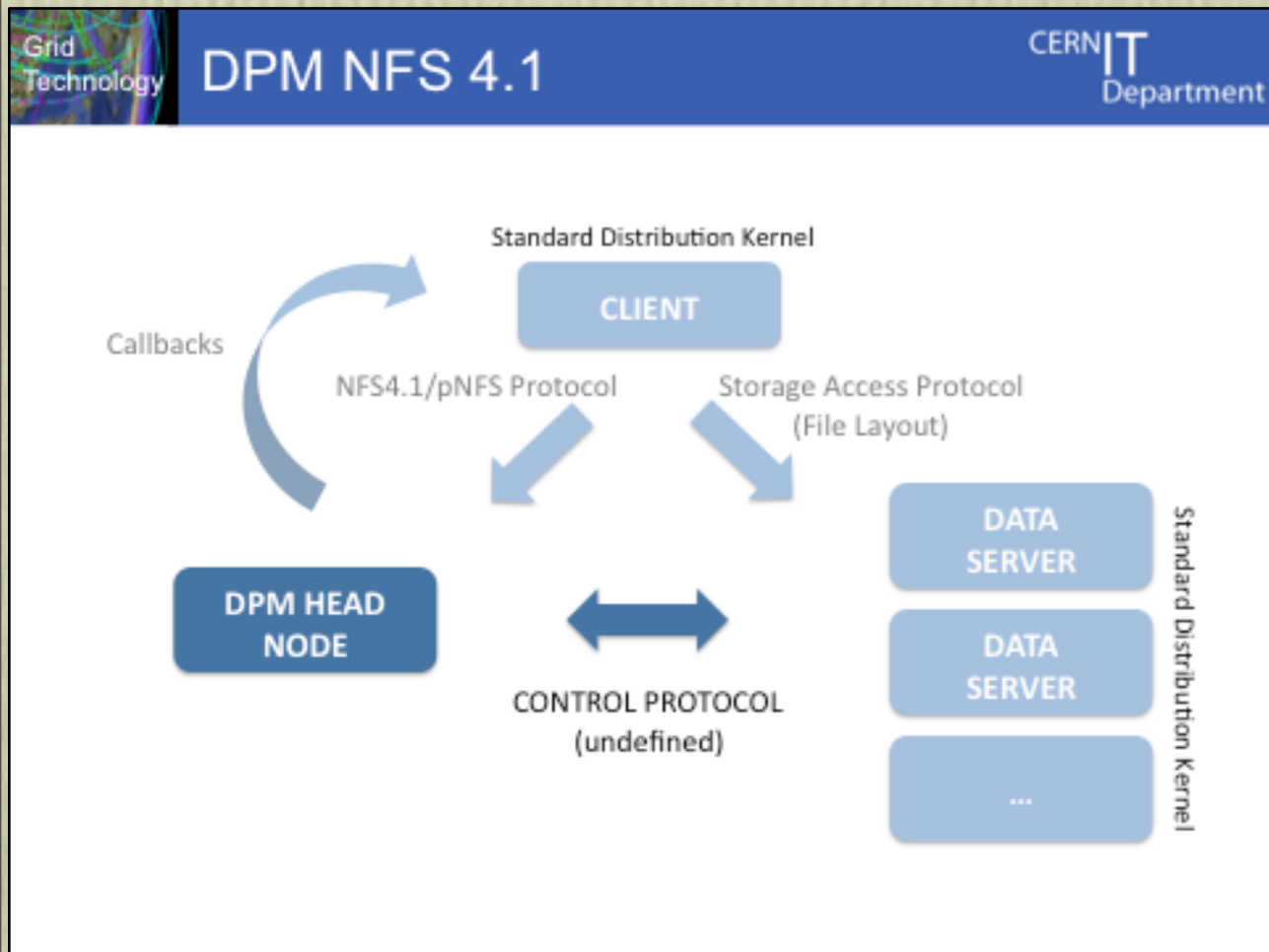
# Data Management

- WLCG Data Management meeting 16 June.
  - Many presentations sprung from this.
- New storage technologies evaluated.
  - SSDs, Ceph, CERNVMM-FS
- Archiving and long-term storage.

# Amsterdam Themes

- Storage themes:
  - Dynamic data and Caches
  - Consistency - Messaging
  - Global filesystems (xrootd, mostly)
  - NFS4.1/pNFS
  - Archiving!

# Amsterdam Themes



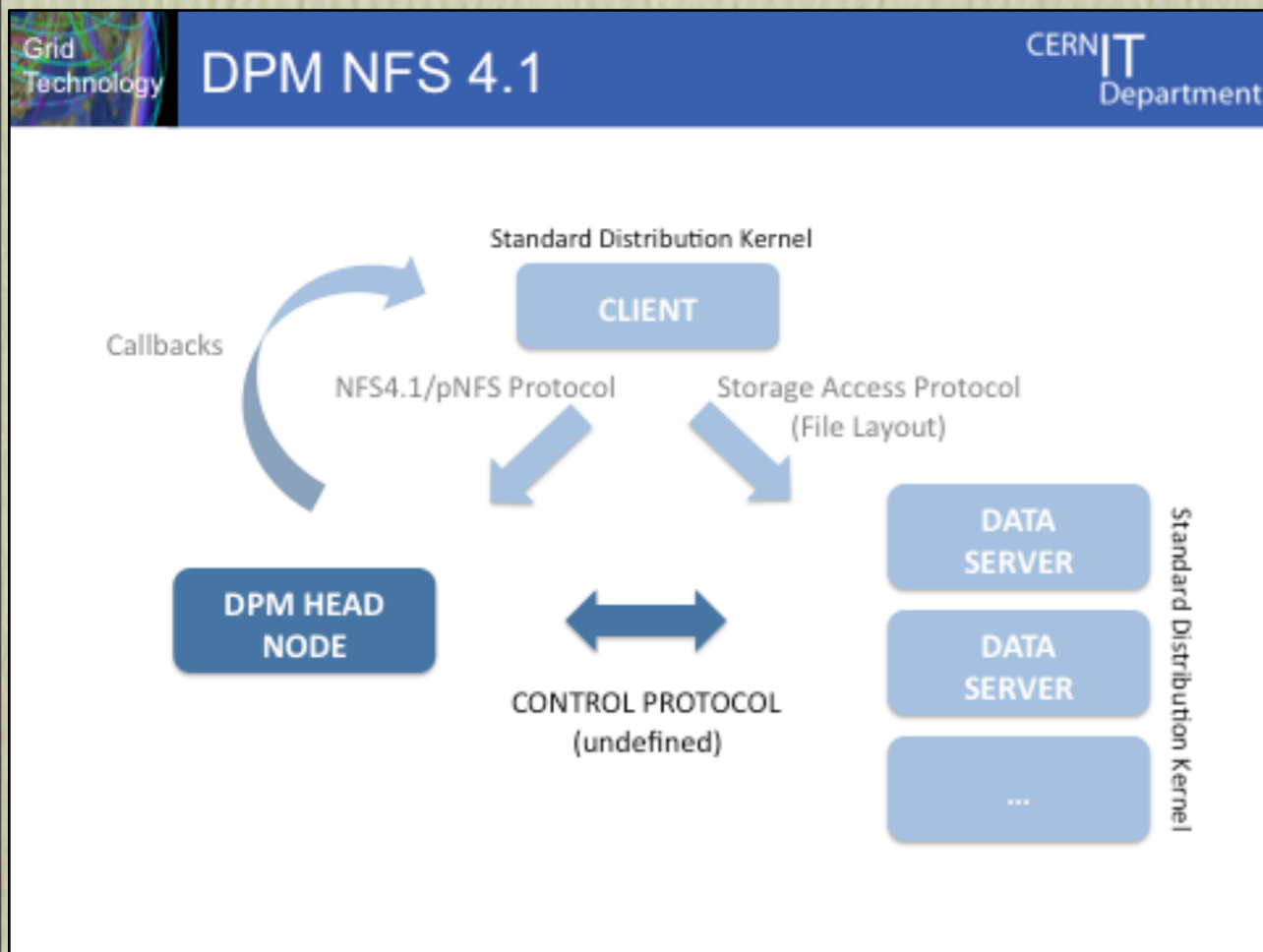
- NFS4.1/pNFS
- Archiving!

Caches

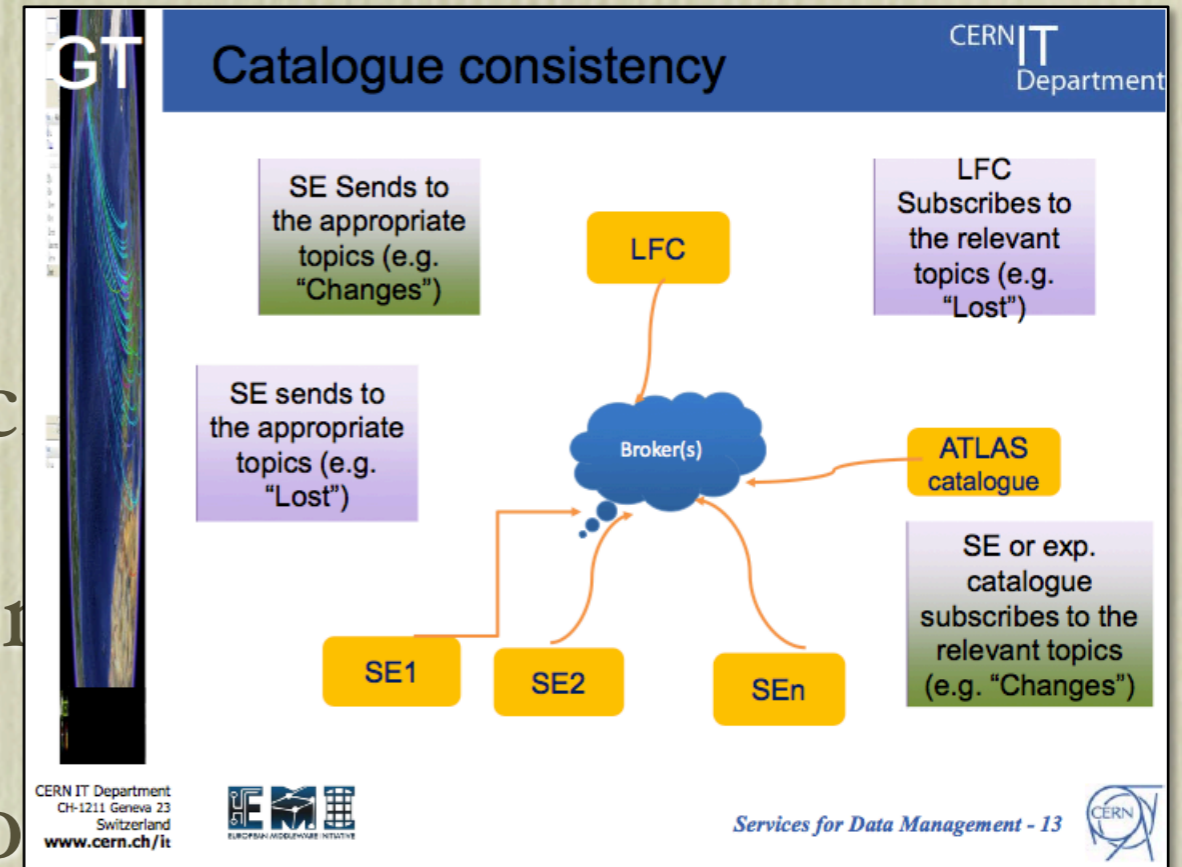
essaging

(xrootd, mostly)

# Amsterdam Themes



- NFS4.1/pNFS
- Archiving!

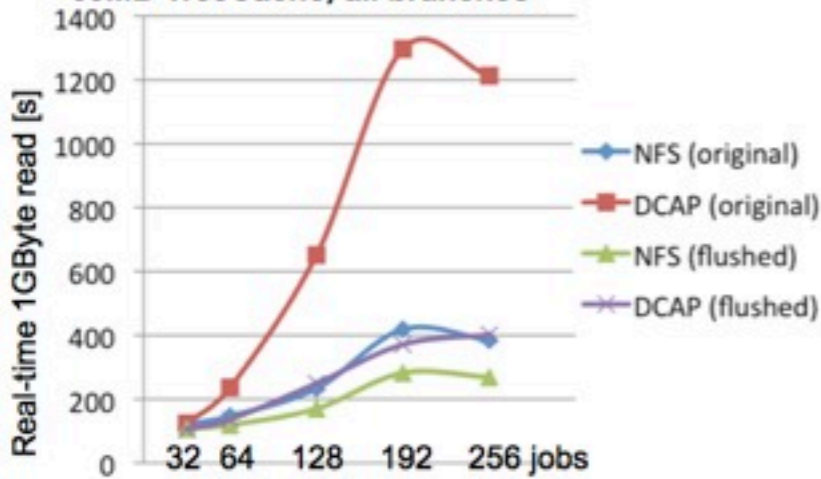


# Amsterdam Themes

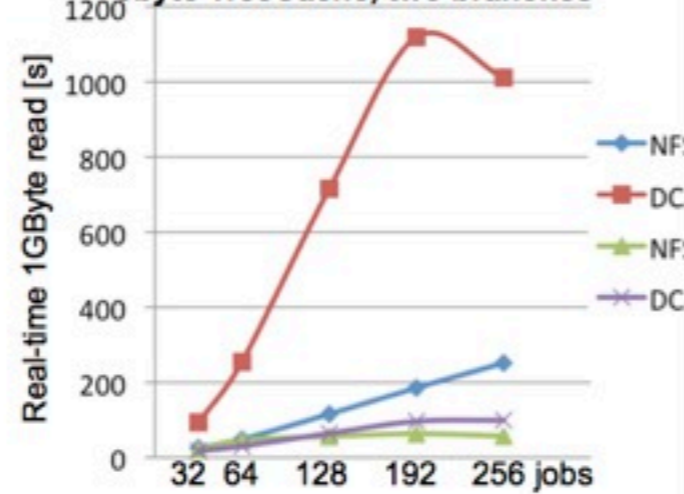
Standard Distribution Kernel

## Half-Synthetic ROOT tests: Results

60MB TreeCache, all branches

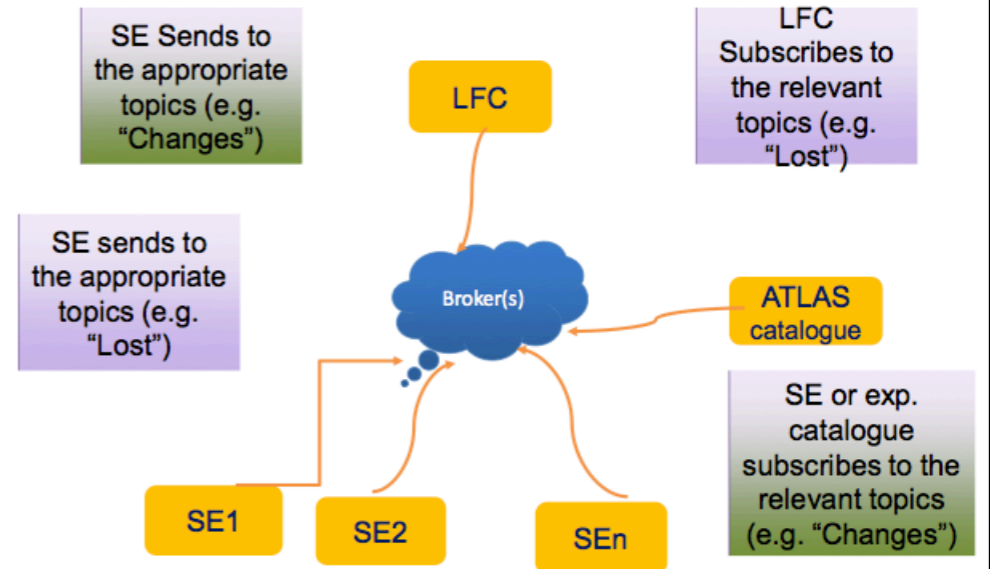


100MB TreeCache, two branches



- > NFS better for original and flushed files than dCap
  - Flushed: not much difference, original: Large difference
- > TreeCache helps, NFS adds additional speed
- > Peak at 192 clients not understood
- > Remember: Just going through events and doing nothing ... not really representative for analysis

Yves Kemp | LHC analysis using NFSv4.1 (pNFS) | 10/20/2010 | Page 18



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



Services for Data Management - 13





# Amsterdam Themes

Grid Technology

## DPM NFS 4.1

CERN IT Department

Standard Distribution Kernel

GT

## Catalogue consistency

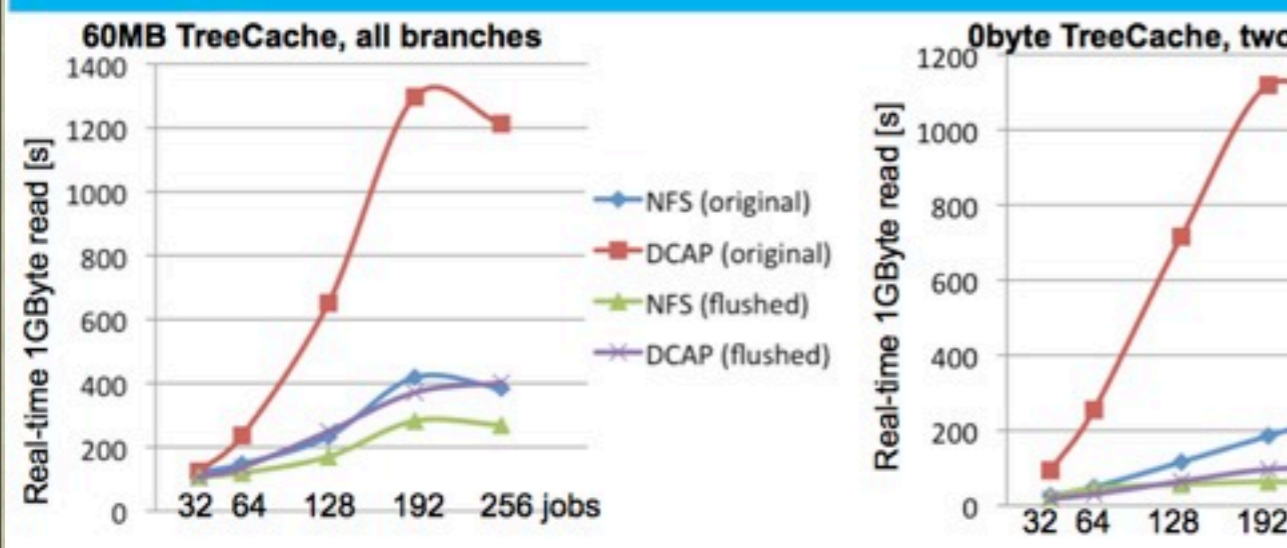
CERN IT Department

SE Sends to the appropriate topics (e.g. "Changes")

LFC

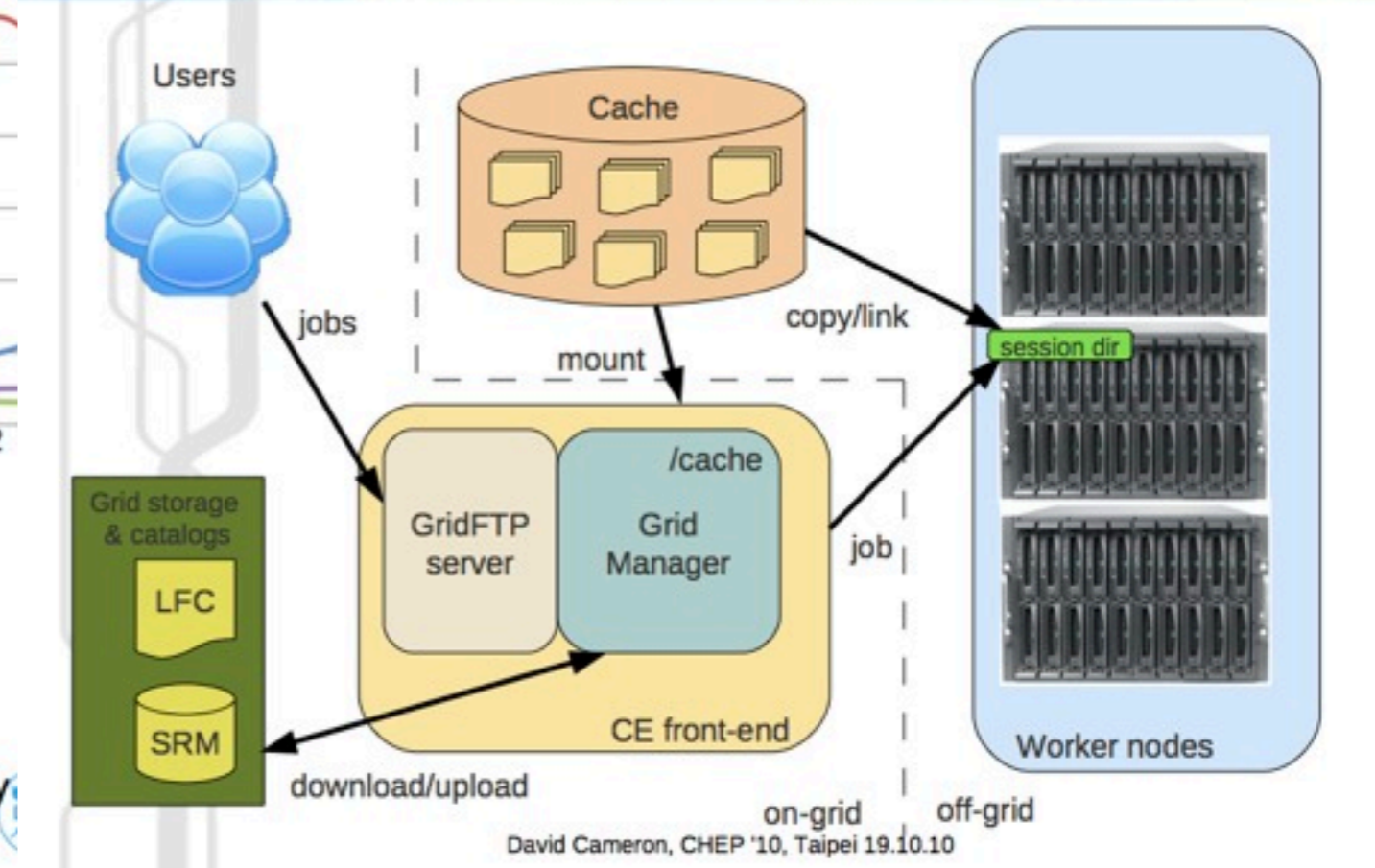
LFC Subscribes to the relevant topics (e.g. "Lost")

## Half-Synthetic ROOT tests: Results



- > NFS better for original and flushed files than dCap
    - Flushed: not much difference, original: Large difference
  - > TreeCache helps, NFS adds additional speed
  - > Peak at 192 clients not understood
  - > Remember: Just going through events and doing nothing ... not really representative for analysis
- Yves Kemp | LHC analysis using NFSv4.1 (pNFS) | 10/20/2010 | Page 18

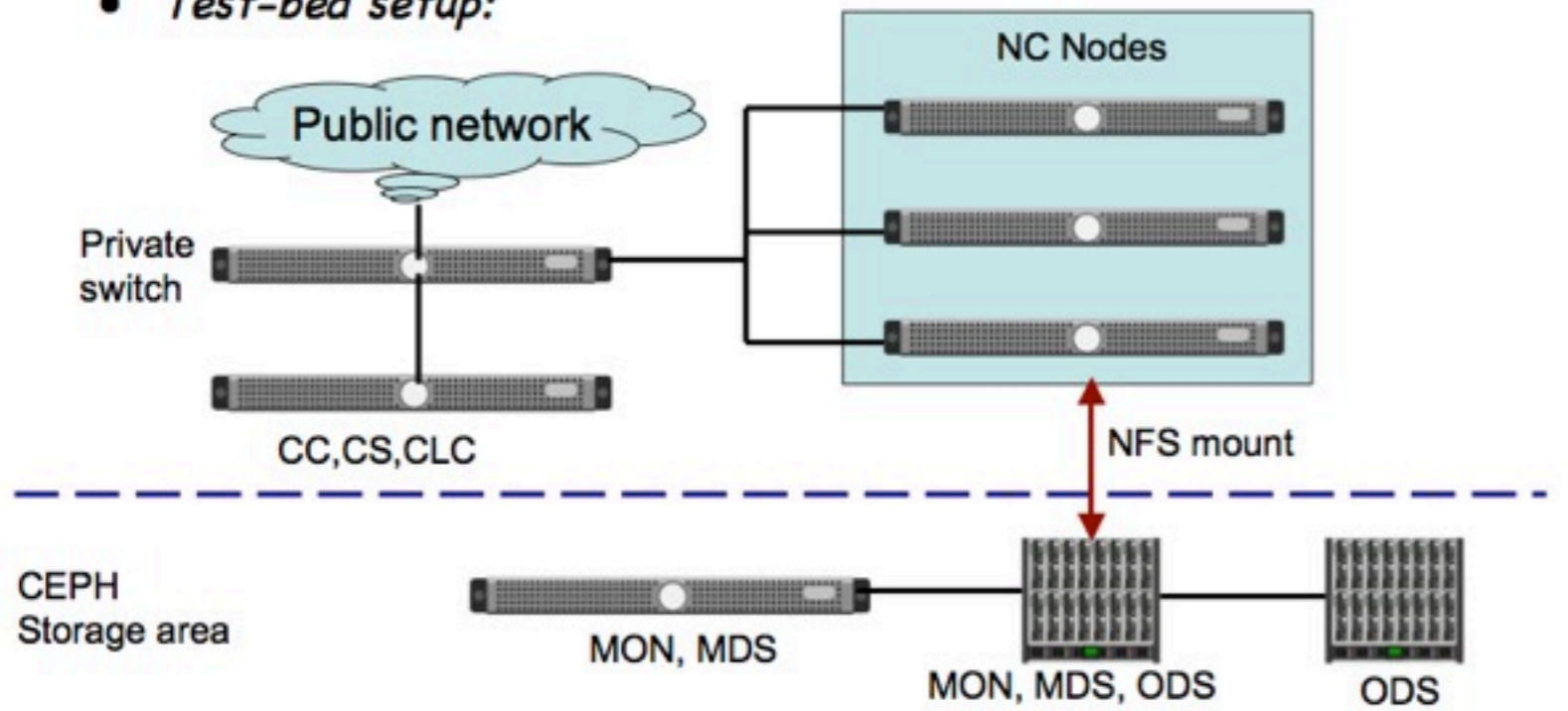
## ARC Data Management Architecture



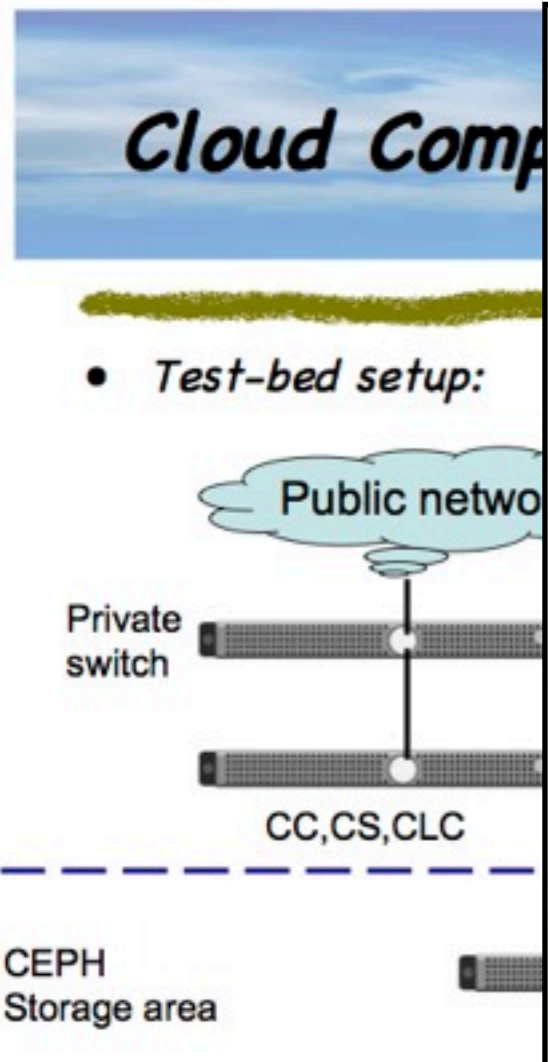
# “New” themes

## Cloud Computing test configuration

- *Test-bed setup:*



# “New” themes



## Results (FileStager)

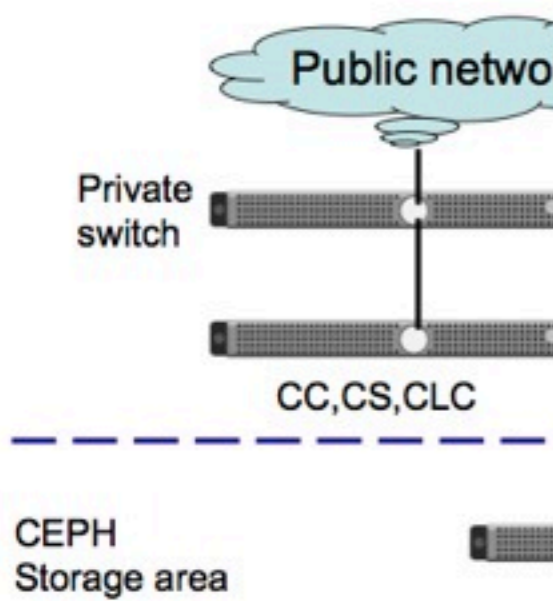
Jobs:Cores	Storage	Efficiency	Throughput
Standard Node			
8:8	1xKingston Value SSD	60%	4.5
8:8	1xSATA HDD	75%	5.5
8:8	1xIntel X25 SSD	80%	6
8:8	2xSATA HDD (RAID 1)	83%	6.6
8:8	2xSATA HDD (RAID 0)	90%	7
Magny-Cours Node			
24:24	1xIntel X25 SSD	50%	12
24:24	2xSATA HDD (RAID 0)	86%	21
Single Occupancy Efficiency (Measured)			
1	SATA HDD	90%	0.9



# “New” themes

## Cloud Comp

- *Test-bed setup:*



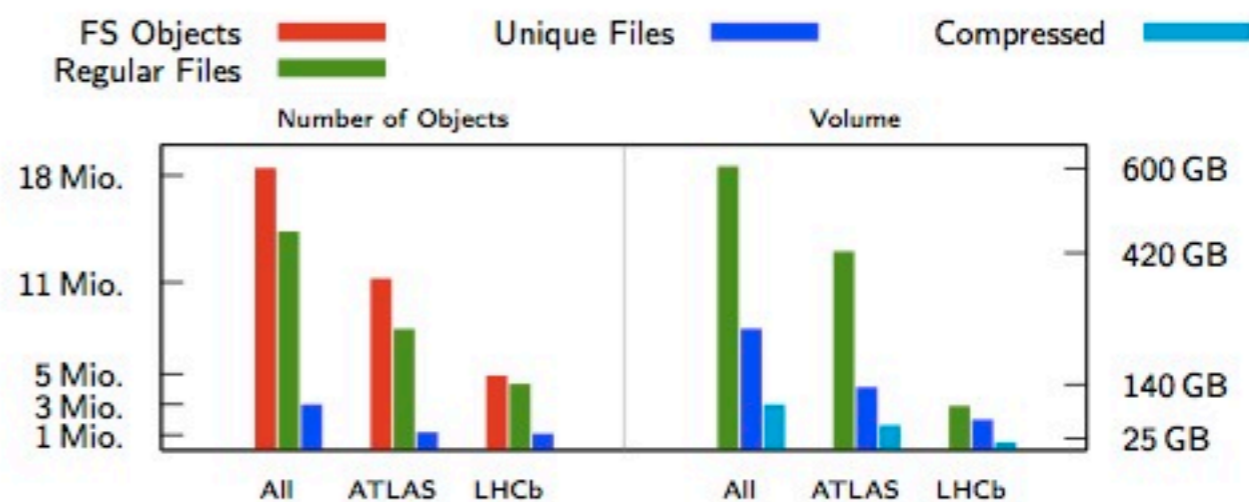
Jobs:Cores	Res
8:8	1xKir
8:8	1x
8:8	1x
8:8	2xSAT
8:8	2xSAT
24:24	1x
24:24	2xSAT
	Single O
I	

## Repository Statistics I

Repositories at CernVM:

ATLAS, CMS, LHCb, ALICE, LCD, NA61, H1, BOSS  
HEPSOFT, Grid UI, LCG Externals

Ongoing: ATLAS Nightlies, ATLAS Conditions Database



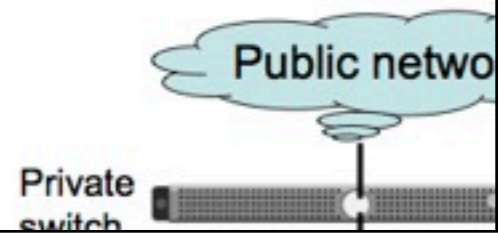
**Overall:** 600 GB, 18.5 Mio. File System Objects  
**Repository Core:** 97 GB (16%), 3.3 Mio. File System Objects (18%)  
 (+ 40 GB Archive Data)



# “New” themes

## Cloud Comp

- *Test-bed setup:*



LHC data access and storage  
Application tuning  
Filesystem / protocol tuning  
Alternative technologies  
Coordination  
Summary / Future plans

### What's in Store?

Tuning grid storage resources for LHC data analysis

Wahid Bhimji  
and the GridPP Storage Crew

University of Edinburgh

CHEP  
19th October 2010



## Res

Jobs:Cores	
8:8	1xKir
8:8	1x
8:8	1x
8:8	2xSAT
8:8	2xSAT

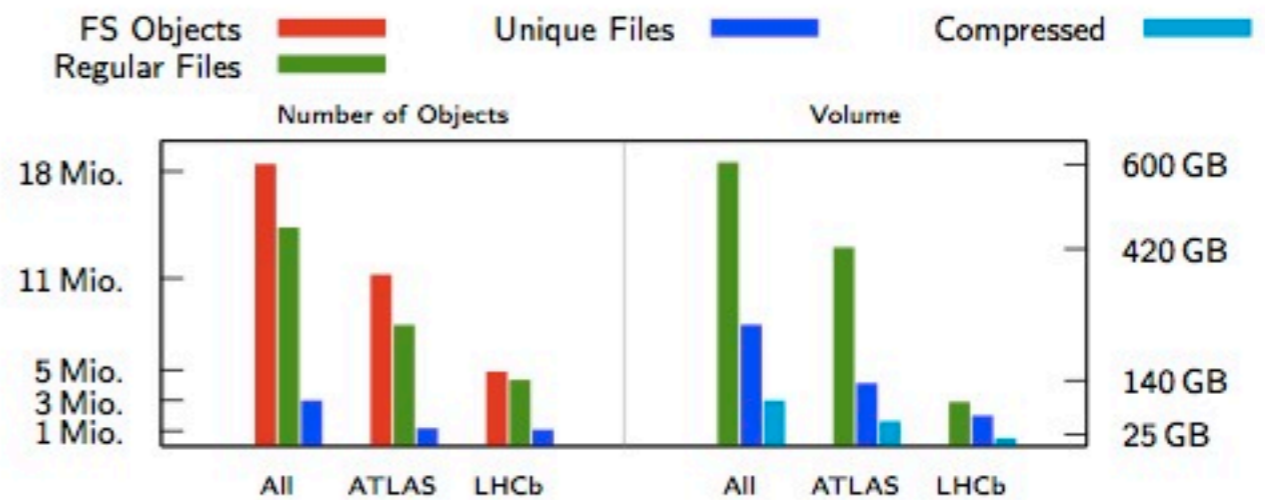
## Repository Statistics I

Repositories at CernVM:

ATLAS, CMS, LHCb, ALICE, LCD, NA61, H1, BOSS

HEPSOFT, Grid UI, LCG Externals

Ongoing: ATLAS Nightlies, ATLAS Conditions Database



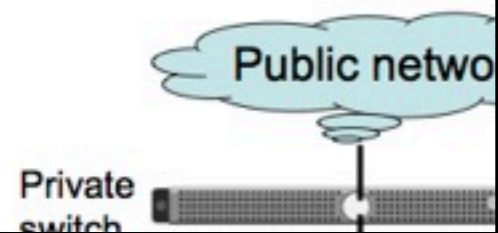
**Overall:** 600 GB, 18.5 Mio. File System Objects  
**Repository Core:** 97 GB (16%), 3.3 Mio. File System Objects (18%)  
 (+ 40 GB Archive Data)



# “New” themes

## Cloud Comp

- *Test-bed setup:*



Jobs:Cores	Res
8:8	1xKir
8:8	1x
8:8	1x
8:8	2xSAT
8:8	2xSAT

### Repository Statistics I

Repositories at CernVM:  
 ATLAS, CMS, LHCb, ALICE, LCD, NA61, H1, BOSS  
 HEPSoft, Grid UI, LCG Externals  
 Ongoing: ATLAS Nightlies, ATLAS Conditions Database

Legend:  
 FS Objects (red), Unique Files (blue), Compressed (cyan), Regular Files (green)

Number of Objects: 18 Mio.

Volume: 600 GB, 420 GB, 140 GB, 25 GB

Categories: All, ATLAS, LHCb

Mio. File System Objects  
 Mio. File System Objects (18%)

LHC data access and storage  
 Application tuning  
 Filesystem / protocol tuning  
 Alternative technologies  
 Coordination  
 Summary / Future plans

LHC data access and storage  
 Application tuning  
 Filesystem / protocol tuning  
 Alternative technologies  
 Coordination  
 Summary / Future plans

Problems  
 Solutions  
 Tests performed

## Solutions?

Many ways to tackle this:

### Application

Some things that can be done in ROOT layer centrally by experiment. Others need coordination/ education of many physicists.

### Filesystems/Protocols

Many alternatives. Each can require specific tuning.

### Hardware

Eg. SSDs - requires more cash  
 Worth it? see Sam Skipsey's talk

### Coordination

Currently a strong interplay between the above - so sites need to be aware of experiment changes and feedback experiences.

## What's in Store

Tuning grid storage resources for

Wahid Bhimji  
 and the GridPP Storage

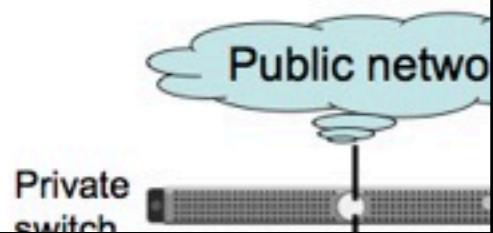
University of Edinburgh

CHEP  
 19th October 2010

# “New” themes

## Cloud Comp

- Test-bed setup:



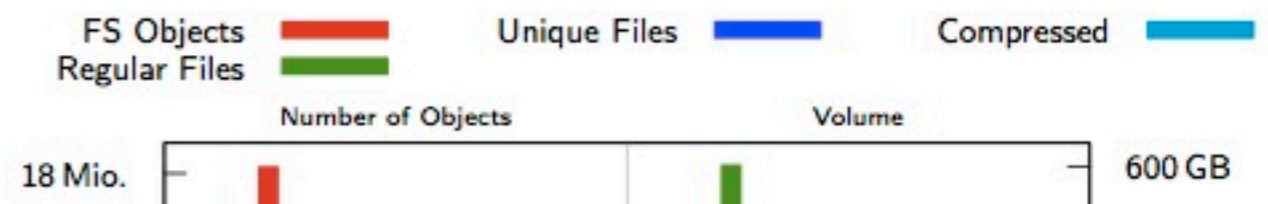
Jobs:Cores	Res
8:8	1xKir
8:8	1x
8:8	1x
8:8	2xSAT
8:8	2xSAT

## Repository Statistics I

Repositories at CernVM:

ATLAS, CMS, LHCb, ALICE, LCD, NA61, H1, BOSS  
HEPSOFT, Grid UI, LCG Externals

Ongoing: ATLAS Nightlies, ATLAS Conditions Database



LHC data access and storage  
Application tuning  
Filesystem / protocol tuning  
Alternative technologies  
Coordination  
Summary / Future plans

LHC data access and storage  
Application tuning  
Filesystem / protocol tuning  
Alternative technologies  
Coordination  
Summary / Future plans

LHC data access and storage  
Application tuning  
Filesystem / protocol tuning  
Alternative technologies  
Coordination  
Summary / Future plans

ROOT IO

## What's in Store

Tuning grid storage resources for

Wahid Bhimji  
and the GridPP Storage

University of Edinburgh

CHEP  
19th October 2010

## Solutions?

Many ways to tackle this:

### Application

Some things that can be done in ROOT  
Others need coordination/ education of

### Filesystems/Protocols

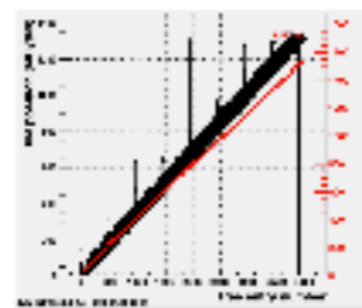
Many alternatives. Each can require specific tuning.

### Coordination

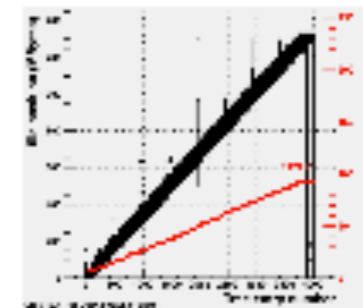
Currently a strong interplay between the aware of experiment changes and feedb

## Basket ordering - remote reading example

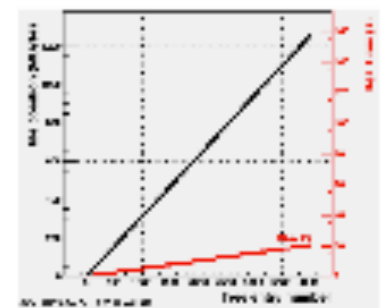
- ROOT test on these AODs, output from TTreePerfStats.
- GPFS running on same site as DPM.
- Ordering makes a much bigger impact.



Unordered - DPM (Rfio)  
Disk Time ≈ 1500s  
Wall Clock ≈ 1700s



Unordered - GPFS  
Disk time ≈ 100s  
Wall Clock ≈ 230s

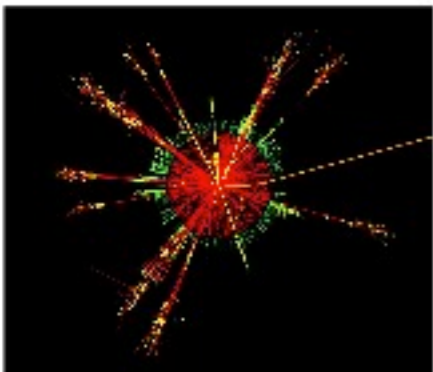


Ordered - DPM (Rfio)  
Disk time ≈ 20s  
Wall Clock ≈ 160s

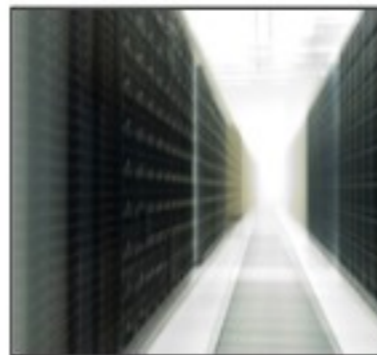
# Multi-Core, GPGPU and all that Jazz

## CHEP 2010

### How to harness the performance potential of current Multi-Core CPUs and GPUs



Sverre Jarp  
CERN  
openlab  
IT Dept.  
CERN



Taipei, Monday 18 October 2010

### Today: Seven dimensions of multiplicative performance



#### First three dimensions:

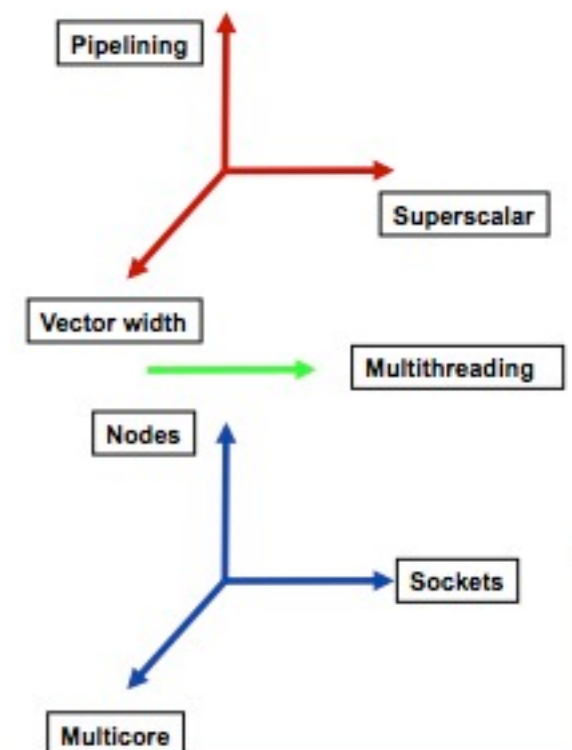
- Pipelined execution units
- Large superscalar design
- Wide vector width (SIMD)

#### 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



# CPU scaling

## **Parallelizing Atlas Reconstruction and Simulation: Issues and Optimization Solutions for Scaling on Multi- and Many-CPU Platforms**

**Charles Leggett<sup>1</sup>**

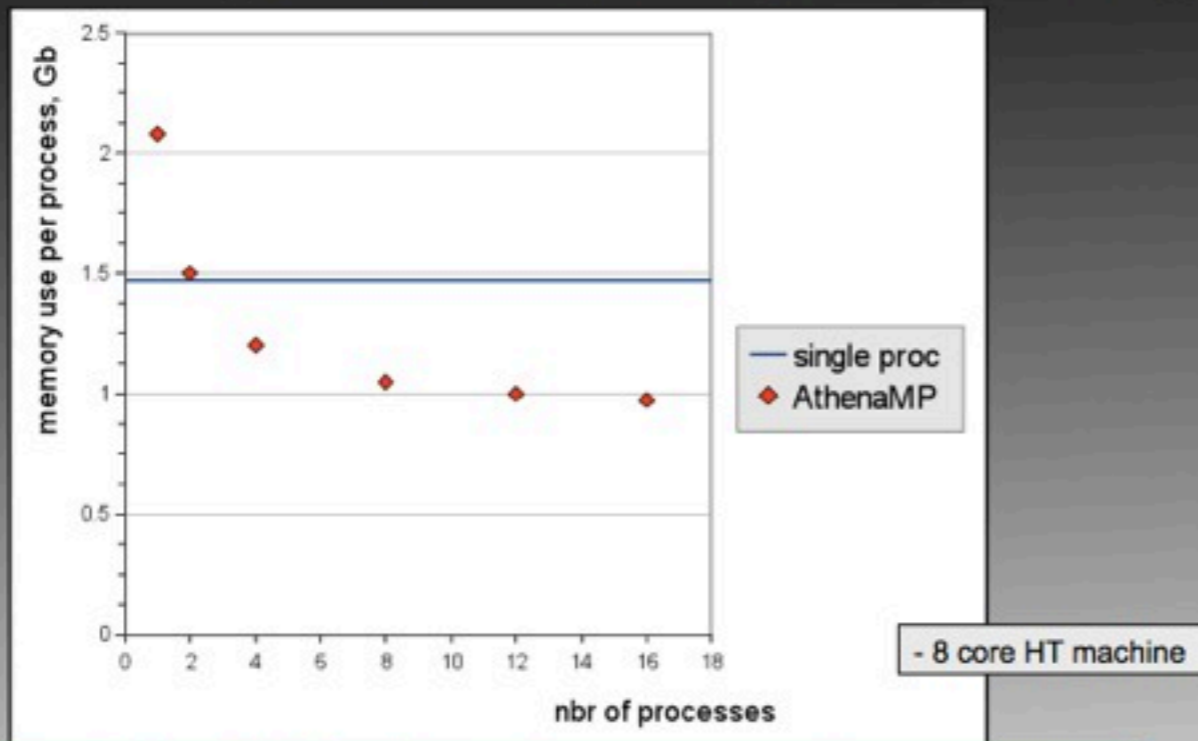
Sebastien Binet<sup>2</sup>, Paolo Calafiura<sup>1</sup>, Keith Jackson<sup>1</sup>, David  
Levinthal<sup>3</sup>, Mous Tatar khanov<sup>1</sup>, Yushu Yao<sup>1</sup>

<sup>1</sup>Lawrence Berkeley National Lab, <sup>2</sup>LAL, <sup>3</sup>Intel

# CPU scaling

## Memory Usage

- Single reco process uses ~1.5 Gb
- We can save significant memory through OS level sharing



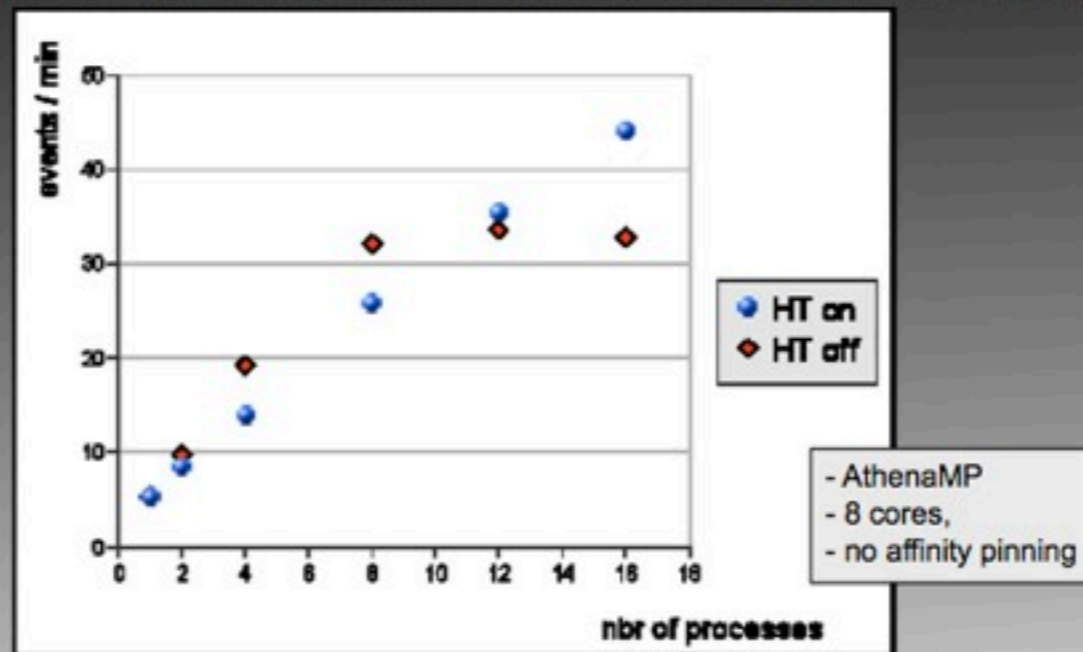
**AthenaMP ~0.5 Gb physical memory saved per process**

# CPU scaling

## Memory Usage

## Hyper-Threading

- Nature of instruction pipeline allows instructions to be interleaved
  - OS sees "virtual cores" as just another processor
  - Only effective until one of the threads saturates a shared resource and stalls



- Turn on HT when you have more processes than cores

# CPU scaling

## Memory Usage

## Hyper-Threading

## Conclusions: Longer Term Solutions

- We (HEP) are not the only ones facing these problems
  - Oracle, IBM, Google all have similar issues
    - They're only just now beginning to realize it
- We need new tools and analysis techniques
  - Current tools fail to show where the problem are, or are not suited to large scale deployment
- We need to drive these optimization techniques into the compilers and linkers themselves
- Changes at the hardware level would also improve the situation
  - It's already happening: new counters are being included in Intel's Westmere and Sandybridge chips which make profiling more useful
  - If we can show Intel exactly what's wrong, and what it will take to fix it in hardware, **they will listen.**

# CPU scaling

ng: What Next-Generation Languages Can Teach Us About HENP Frameworks in the Manycore Era



Author: *Sebastien Binet*

Institute: *LAL/IN2P3*

Date: *2010-10-19*

Conf: *CHEP-2010*

- We (HEP) are not the only ones facing these p
  - Oracle, IBM, Google all have similar issues
    - They're only just now beginning to realize it
- We need new tools and analysis techniques
  - Current tools fail to show where the problem are, or to large scale deployment
- We need to drive these optimization techniques into the compilers and linkers themselves
- Changes at the hardware level would also improve the situation
  - It's already happening: new counters are being included in Intel's Westmere and Sandybridge chips which make profiling more useful
  - If we can show Intel exactly what's wrong, and what it will take to fix it in hardware, **they will listen.**

# CPU scaling

## ng: What Next-Generation Languages Can Teach Us About HENP Frameworks in the Manycore Era

### Next-gen (and not so next-gen) languages

- C1X + GCD/libdispatch (closures + work queues)
- C++0x (lambda functions, `std::thread`)
- Python/Cython + PyCSP + multiprocessing + mpi4py + ...
- Vala/Genie
  - ▶ <http://live.gnome.org/Vala>
  - ▶ <http://live.gnome.org/Genie>
- Haskell, Erlang
  - ▶ is HEP ready for functional programming ?
- go
  - ▶ <http://golang.org>

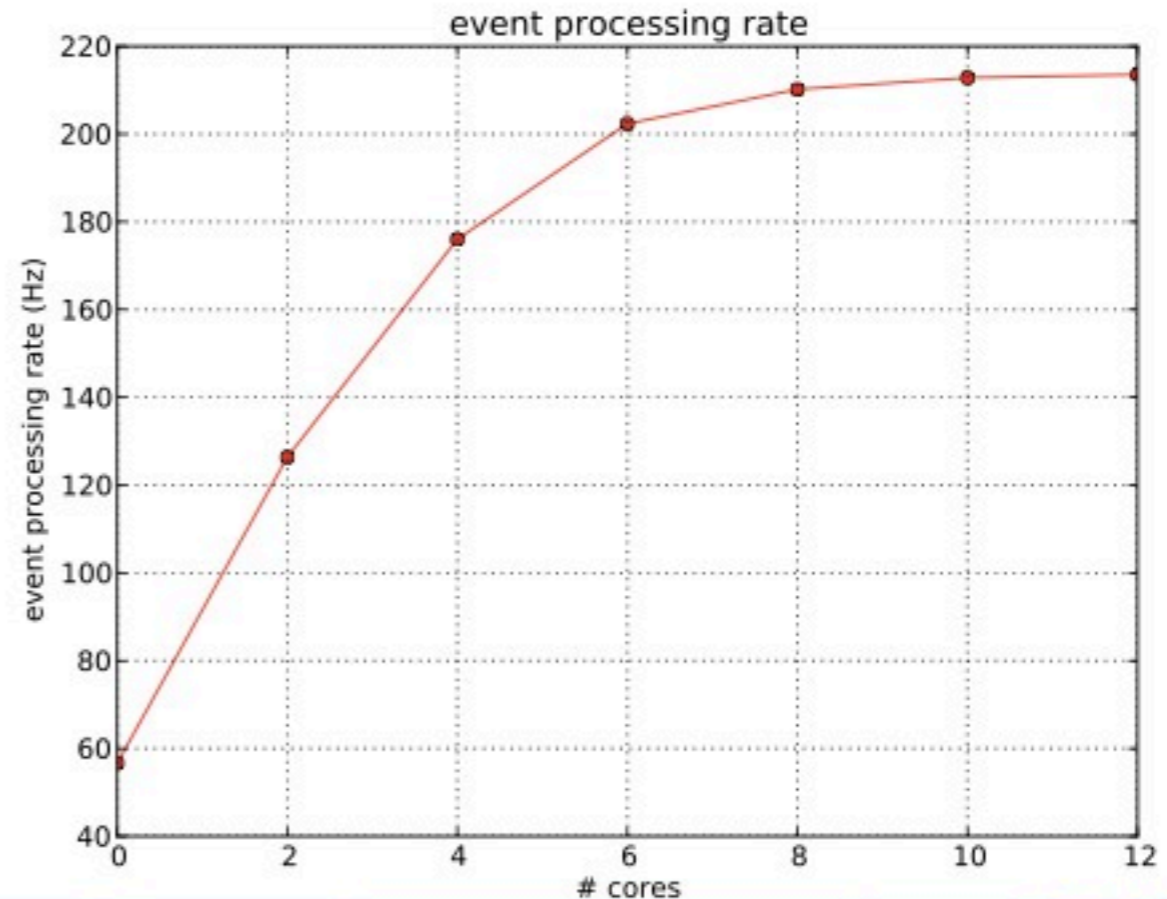
- We (HEP) are not the only ones facing these p
  - Oracle, IBM, Google all have similar issues
    - They're only just now beginning to realize it
- We need new tools and analysis techniques
  - Current tools fail to show where the problem are, or to large scale deployment
- We need to drive these optimization techniques into compilers and linkers themselves
- Changes at the hardware level would also improve the situation
  - It's already happening: new counters are being included in Westmere and Sandybridge chips which make profiling easier
  - If we can show Intel exactly what's wrong, and what it will take to fix it in hardware, **they will listen.**

# CPU scaling

ng: What Next-Generation Languages Can Teach Us About HENP Frameworks in the Manycore Era

Next-gen (and not so next-gen) languages

## Results



- We (HEP) are not the only ones facing these p
  - Oracle, IBM, Google all have similar issues
    - They're only just now beginning to realize it
- We need new tools and analysis techniques
  - Current tools fail to show where the problem are, or to large scale deployment
- We need to drive these optimization techniques into compilers and linkers themselves
- Changes at the hardware level would also improve the situation
  - It's already happening: new counters are being included in Westmere and Sandybridge chips which make profiling easier
  - If we can show Intel exactly what's wrong, and what to do about it in hardware, **they will listen.**

# GPGPU use

GPU Computing  
Z Finder  
Kalman Filter  
Summary

## Algorithm Acceleration from GPGPUs for the ATLAS Upgrade

Computing in High Energy and Nuclear Physics 2010

Andrew Washbrook  
on behalf of the ATLAS Collaboration

University of Edinburgh

21st October 2010



1/22



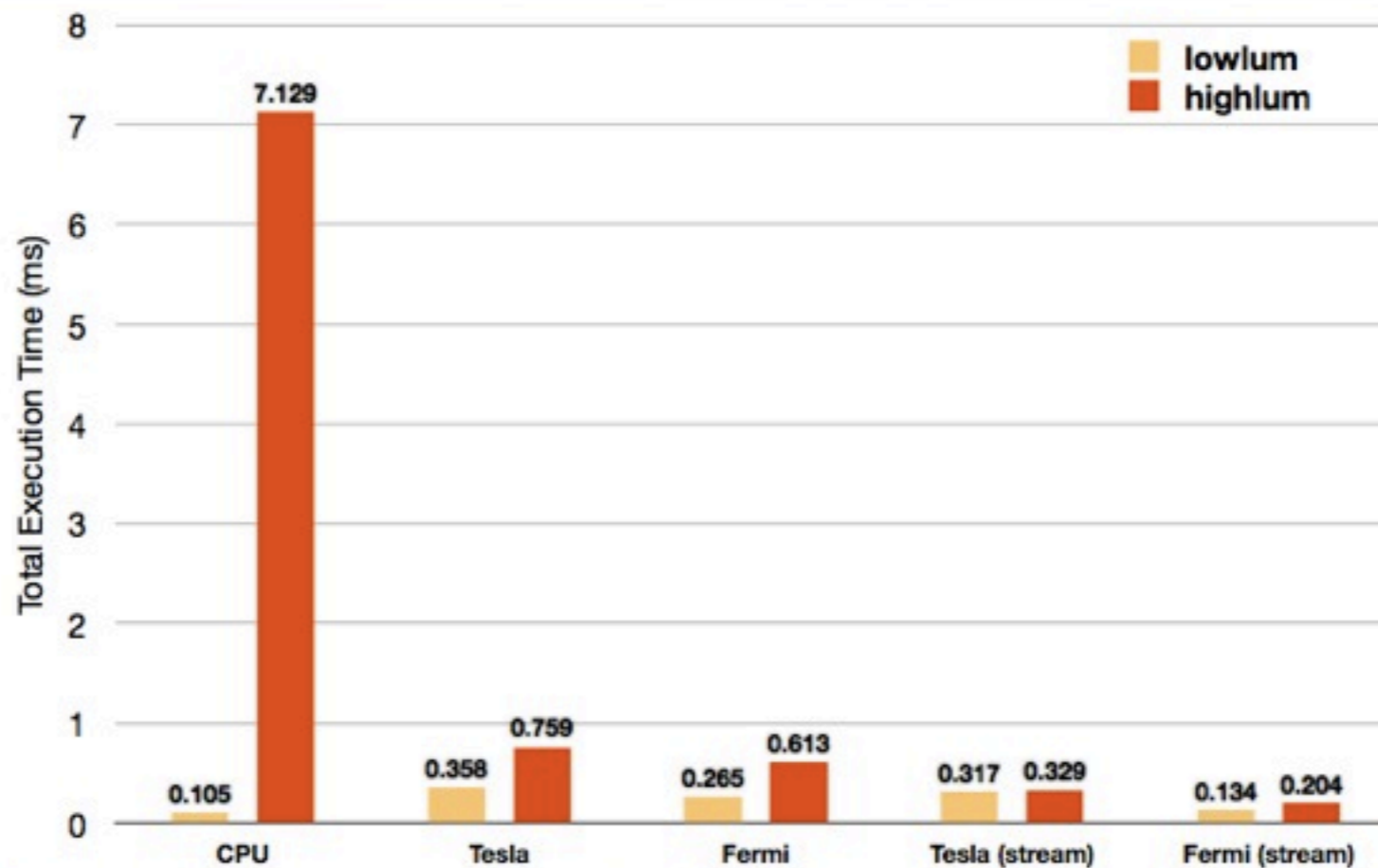
# GPGPU use

GPU Computing

GPU Computing  
Z Finder  
Kalman Filter  
Summary

GPU Motivation  
Algorithm and Test case  
Z Finder Kernel  
Timing Results

## Algorithm Timing Results



- Results for spacepoint pairs show up to 35x speed-up (Fermi).
- Initial results for spacepoint *triplets* also show speed-up.

# GPGPU use

GPU Computing

GPU Computing

GPU Motivation

GPU Computing

Z Finder

Kalman Filter

Summary

GPU Motivation

Track Reconstruction in ATLAS

Kalman Filter for CUDA

Algorithm

Timing

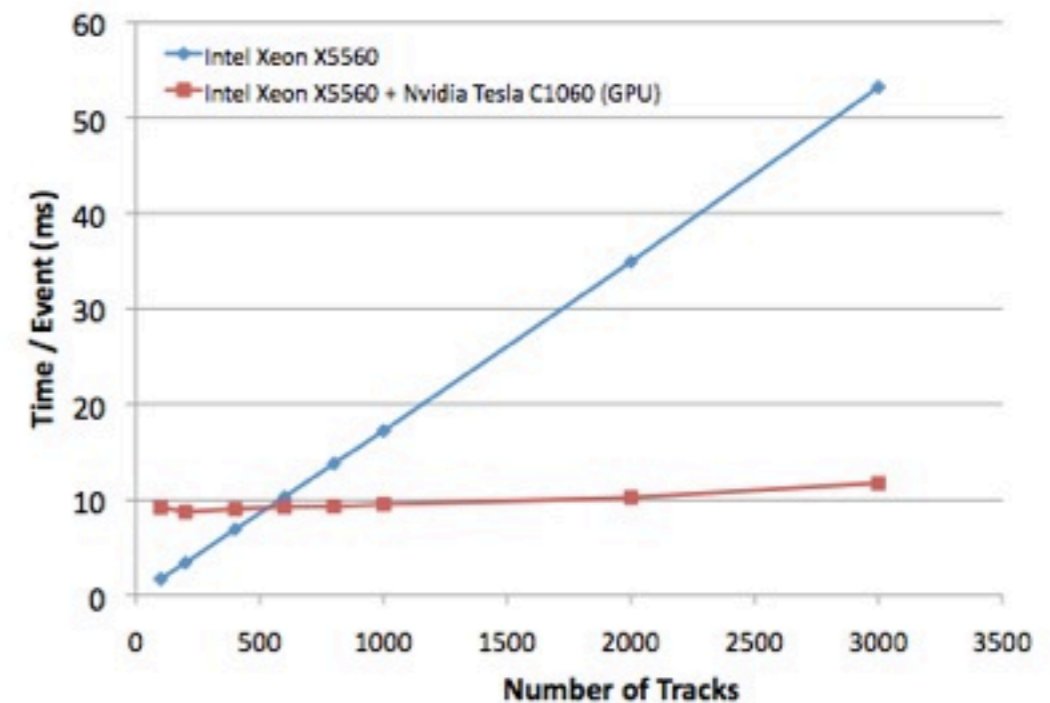
## Kalman Filter for CUDA

D. Emelianov

Con

Total Execution Time (ms)

- Standalone version successfully ported to C.
  - Structs of arrays used to store track data.
  - Vector data types (e.g. *float4*) for compact representation of data.
  - One GPU thread per track.
  - Modification of smoothing algorithm required for single precision arithmetic.
- R
  - In



Muon tracks,  $p_T=10\text{GeV}$ , full MC simulation

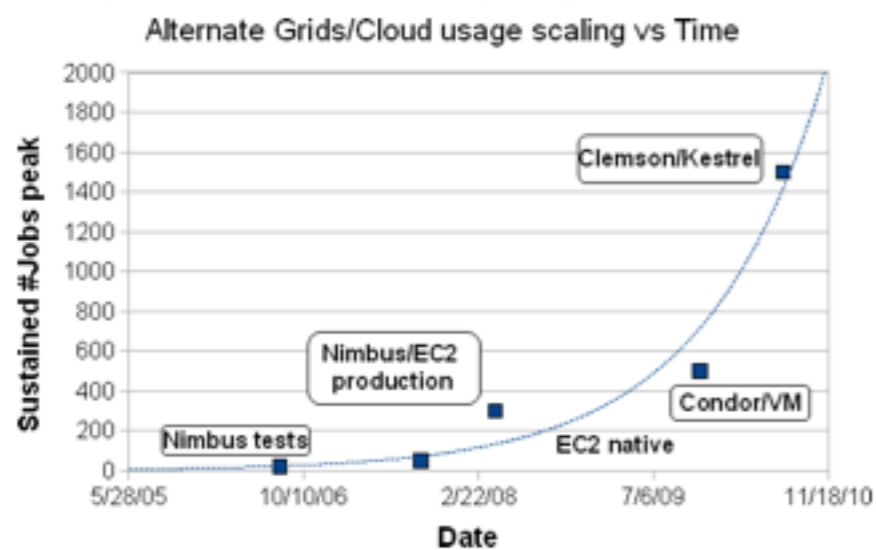
- Over 5x speed-up seen at 3000 tracks.

# Virtualisation and Clouds

- Virtual machines promising
  - but there are security and config implications
- Use of Clouds - works, but you still have to pay more (and aren't they just VMs in a datacenter?).



## Testing scale & usage growth



Today, 1,000 jobs stable on Cloud or Hybrid ("virtualized Grids") possible (some challenges with stability / scalability at times)

**10 k to 100 k jobs needed for STAR – Should be "routine" by beginning of 2011**



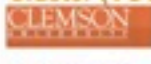
Promising ... OSG ~ 13 M jobs at times ... some way to go ...

## Testing scale & usage growth



## Models – 10-100 job scale series



Model	Scale, Effic. & Scope	Observations	Key features & observations
 Amazon/EC2 Xen	Eff. > 99% Medium instance Raw simulation workflow (event generator)	<ul style="list-style-type: none"> <li>IO rather inadequate for large scale efforts (5 MB/sec per node)</li> <li>Good for simulations and simple workflows: little "I", not that much "O" in IO</li> <li>Normal instances unlikely suitable for HPC/HTC or large data mining (did not test the HPC instance. Has anyone?)</li> </ul>	<ul style="list-style-type: none"> <li>Amazon has a concept of VM repository (needed)</li> <li>Amazon AAA rudimentary (lacking?)</li> <li>Amazon has a simple and competitive pricing model: \$0.09 / hour - 300 jobs, week long cost ~ \$5,600. A year long CPU @ 100 jobs saturation ~ 79k\$</li> </ul>
 Nimbus/EC2 Xen	Eff. ~ 85% (1 <sup>st</sup> ) and ~ 97% (2 <sup>nd</sup> ) Raw simulation workflow (event generator)	<ul style="list-style-type: none"> <li>Initial drop of efficiency due to batch system "inside"</li> <li>More "natural" to handle submission (Cloud / Virtualization behind, Grid-like in front)</li> <li>First REAL usage for Physics</li> </ul>	<ul style="list-style-type: none"> <li>Contextualization needed at startup (GK not known a-priori, batch need to know topology)</li> <li>OSG stack inside, GK+WN - virtual space looks like "another OSG site" – good attempt to unify</li> </ul>
Virtual Org. Cluster (VOC)  [ACAT 2010] KVM	Eff 100% (?) Did not lose a single job Raw simulation workflow (event generator)	<ul style="list-style-type: none"> <li>Looked like a "grid" site – submit job to it (pull or push), VM appeared on demand due to a "subscribe" mechanism for batch</li> <li>Transfer limited to University / National lab line</li> <li>Required IPs (some scale problem; thought of overlaid network ...)</li> </ul>	<ul style="list-style-type: none"> <li>Contextualization remains a site specific (time) overhead</li> <li>Interface is standard Grid – user is agnostic of technology</li> <li>Performance improved by caching image locally OR directing changes to local disk – impossible on EC2. Final overhead &lt; 1% (no NFS read)</li> </ul>

## Testing scale & usage growth





15 11 10 100 1 1 1 1



## Models – larger scales ...



Model	Scale, Effic. & Scope	Observations	Key features & observations
<b>Condor/VM</b> 	Scale 500+ jobs  Eff. unclear [80,85%]  <i>Raw simulation workflow (event generator)</i>	<ul style="list-style-type: none"> <li>10% of the VM never started, 15% stopped (crashed), 5% net loss for long simulation jobs (VM reboot every 24 hours). Need to be able to extend lease?</li> <li>Data transfer mechanism was through common SE and separate from workflow</li> <li>Results used for an analysis PoP</li> </ul>	<ul style="list-style-type: none"> <li>Interface remains grid-like but only starts VM; No real job get "inside" – "pull model" (via cron)</li> <li>As many VMs as one wants: nearly no contextualization (apart from SE) reduce overheads on local staff, condor steering</li> <li>IP space is local – no connection to outside – transfer of data out hard but SE &amp; Cloud may be the path ...</li> </ul>
<b>Clemson/Kestrel</b> 	Scale 1000 jobs  Eff. Unclear, cruising average 90%  <i>Full simulation with track reconstruction, detector efficiency correction using the full STAR framework</i>	<ul style="list-style-type: none"> <li>Job are "lost" if problem (communication, dead program, dying VM)</li> <li>FIRST time we mixed VMs from Clemson + CERN (true "Cloud" idea)</li> <li>Full database access as a service within the VM</li> <li>Second massive usage of cloud targeting a conference, physics publication</li> </ul>	<ul style="list-style-type: none"> <li>VM packaged everything</li> <li>Globus and MyProxy inside for transfer out – no real problems</li> <li>Job do not get "inside" – simple command trigger a script (external workload manager needed)</li> <li>Jobs start/stop were managed using a common Jabber-like client</li> </ul>

STAR framework btw is a single purpose framework for simu, reco, user analysis

## Testing scale & usage growth



Model 1: 10,000 : 1 : 1 : 1

Model 2: 10,000 : 1 : 1 : 1

## Clemson/Kestrel model



### Generation

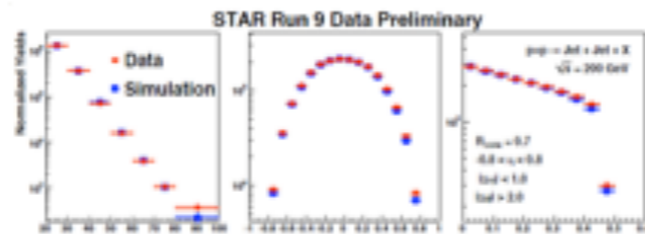
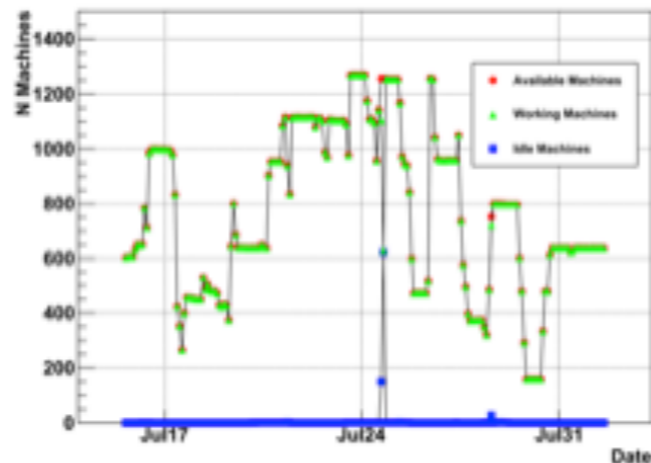
- 12 Billion PYTHIA events were generated – LARGEST sample produced we know off
- Used over 400,000 CPU hours on 1,000 CPUs at Clemson (+CERN) over the course of a month

### Comparison to normal operation

- Cloud allowed STAR to expand its computing resources by 25%. Student thesis work possible
  - Available #of CPU per users ~ 50
- A year long science wait time.

### Achievement for this analysis

- 4 orders of magnitude increase in number of events used in similar analysis in STAR
- Near elimination of all uncertainties caused by statistics
- Un-ambiguously demonstrated good agreement between our data sample and simulation
- Results presented at Spin 2010 conference (October)



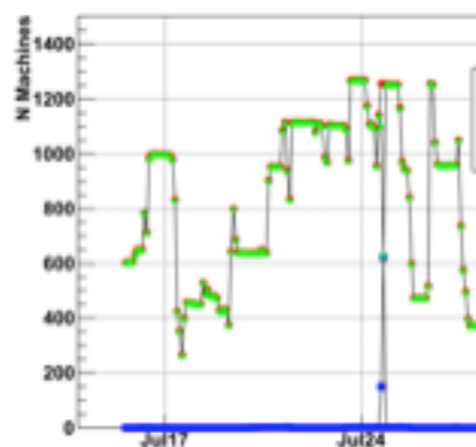
## Testing scale & usage growth



M 1 1 10 100 : 1 : 1 : 1

## Clemson/Kestrel model

- **Generation**
  - 12 Billion PYTHIA events were generated – LARGEST sample produced we know off
  - Used over 400,000 CPU hours on 1,000 CPUs at Clemson (+CERN) over the course of a month
- **Comparison to normal operation**
  - Cloud allowed STAR to expand its computing resources by 25%. Student thesis work possible
    - Available #of CPU per users ~ 50
  - A year long science wait time.
- **Achievement for this analysis**
  - 4 orders of magnitude increase in number of events used in similar analysis in STAR
  - Near elimination of all uncertainties caused by statistics
  - Un-ambiguously demonstrated good agreement between our data sample and simulation
  - Results presented at Spin 2010 conference (October)



**CernVM**  
Software Appliance

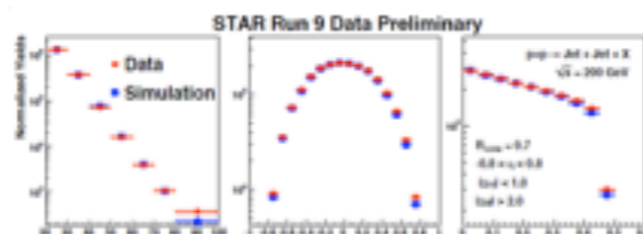
## CernVM Users



CHEP 2010

Predrag.Buncic@cern.ch

Taipei, 19 October 2010 - 3







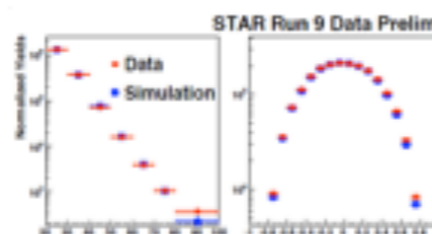
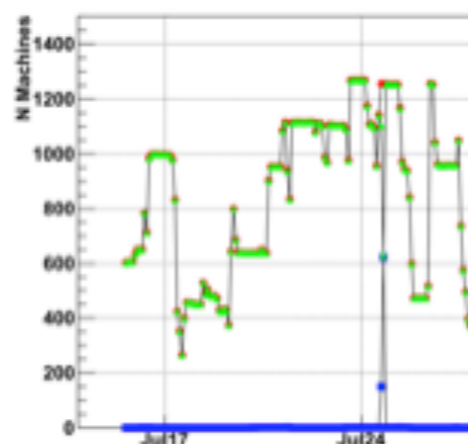
## Testing scale & usage growth

May 11 10 100 1 1 1

Model 1 1 1 1 1 1

## Clemson/Kestrel model

- Generation**
  - 12 Billion PYTHIA events were generated – LARGEST sample produced we know off
  - Used over 400,000 CPU hours on 1,000 CPUs at Clemson (+CERN) over the course of a month
- Comparison to normal operation**
  - Cloud allowed STAR to expand its computing resources by 25%. Student thesis work possible
    - Available #of CPU per users ~ 50
  - A year long science wait time.
- Achievement for this analysis**
  - 4 orders of magnitude increase in number of events used in similar analysis in STAR
  - Near elimination of all uncertainties caused by statistics
  - Un-ambiguously demonstrated good agreement between our data sample and simulation
  - Results presented at Spin 2010 conference (October)



CernVM

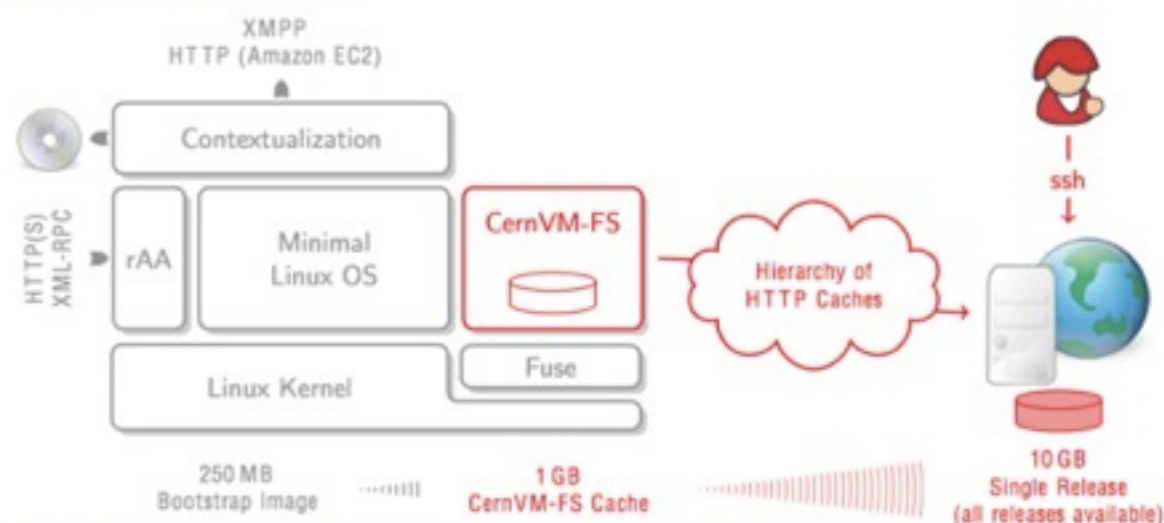
CernVM

CernVM  
Software Appliance

CernVM Users

Part #1: Minimal OS image

Part #2: CernVM-FS



- Experiment software is changing frequently and we want to avoid need to frequently update, certify and redistribute VM images with every release
- Only a small fraction of software release is really used
- CernVM-FS: Read-only, network (HTTP) file system optimized for efficient software delivery. See: J.Blomer ([PS06-5-434](#))

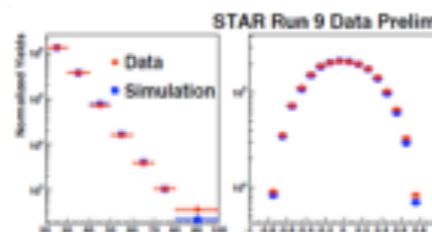
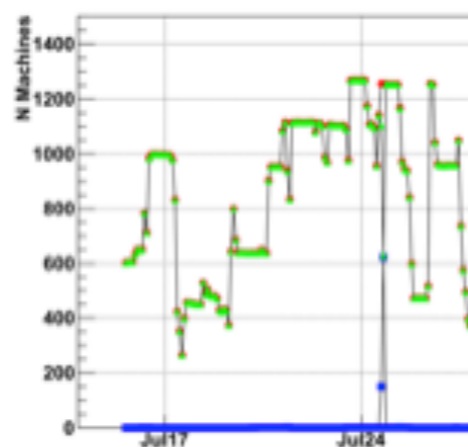
## Testing scale & usage growth

May 11 10 100 1 1 1 1

Model 1 1 1 1 1 1 1

## Clemson/Kestrel model

- **Generation**
  - 12 Billion PYTHIA events were generated – LARGEST sample produced we know off
  - Used over 400,000 CPU hours on 1,000 CPUs at Clemson (+CERN) over the course of a month
- **Comparison to normal operation**
  - Cloud allowed STAR to expand its computing resources by 25%. Student thesis work possible
    - Available #of CPU per users ~ 50
  - A year long science wait time.
- **Achievement for this analysis**
  - 4 orders of magnitude increase in number of events used in similar analysis in STAR
  - Near elimination of all uncertainties caused by statistics
  - Un-ambiguously demonstrated good agreement between our data sample and simulation
  - Results presented at Spin 2010 conference (October)



CernVM

CernVM

CernVM

CernVM

## CernVM Users

Part #1: Minimal OS image

Part #2: CernVM-FS

As easy as 1,2,3

1. Login to Web interface

2. Create user account

3. Select experiment, appliance flavor and preferences

Taipei, 19 October 2010 - 18

# Talking to people.



**A unique collaboration system,  
designed for the LHC**

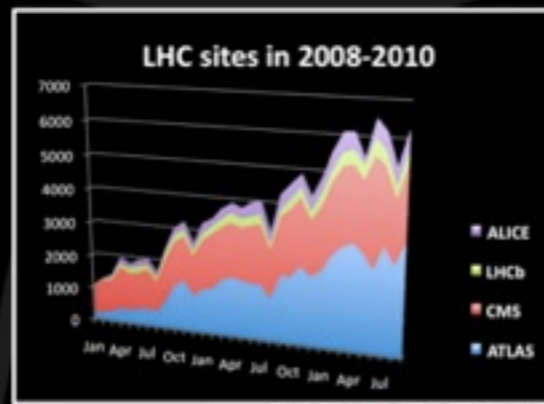
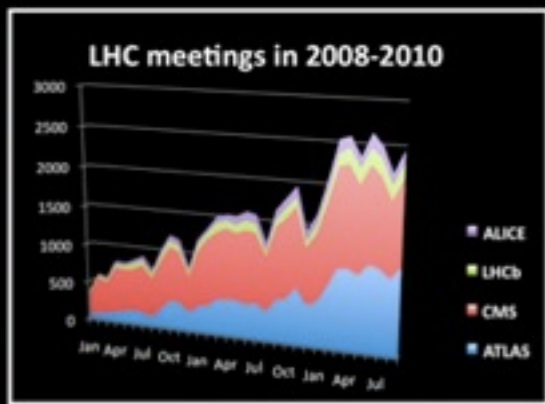
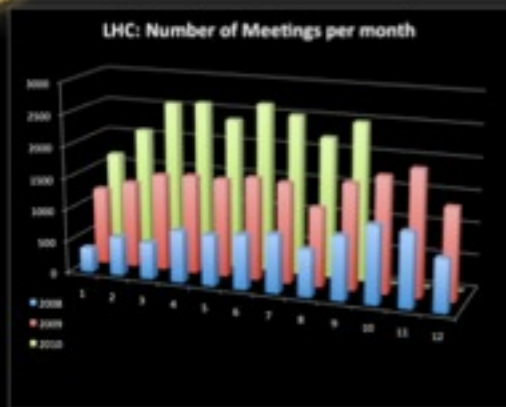
Philippe Galvez, Caltech

CHEP2010, Taipei 10/21/2010

# Talking to people.



## EVO Usage



EDU

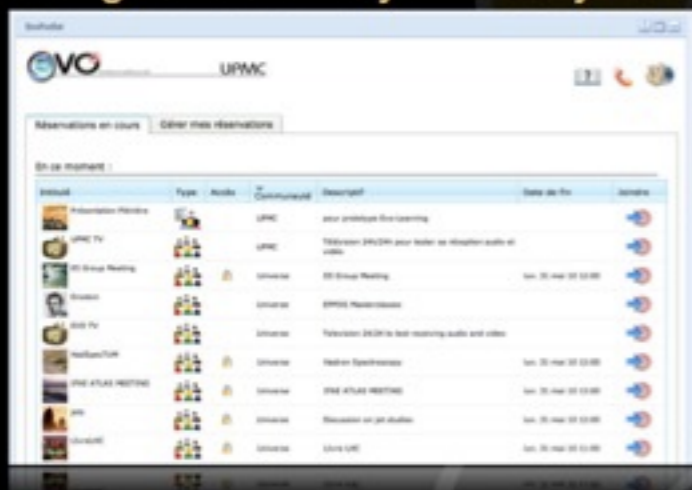
CALIFORNIA INSTITUTE OF TECHNOLOGY

# Talking to people.

## EVO Usage

## EVO Web Portal

Web Interface to allow users to access information on EVO meetings and click to join directly from the portal



Observation	Type	Media	Comments	Description	Date de Fin	Actions
Observation 1	UPMC			your language for learning		
UPMC TV	UPMC			Telepresence 2009/09 your leader in observation audio in video		
UPMC Meeting	UPMC			UPMC Meeting	Mon, 20 Mar 09 00:00	
UPMC	UPMC			UPMC Meeting		
UPMC TV	UPMC			Telepresence 2009/09 for learning audio and video		
UPMC Meeting	UPMC			UPMC Meeting	Mon, 20 Mar 09 00:00	
UPMC	UPMC			UPMC Meeting	Mon, 20 Mar 09 00:00	
UPMC	UPMC			UPMC Meeting	Mon, 20 Mar 09 00:00	
UPMC	UPMC			UPMC Meeting	Mon, 20 Mar 09 00:00	

Integration with Shibboleth and others authentication mechanisms will be added as we deploy the portal

EVO.CALTECH.EDU

# Talking to people.



## EVO Usage



## EVO Web Portal

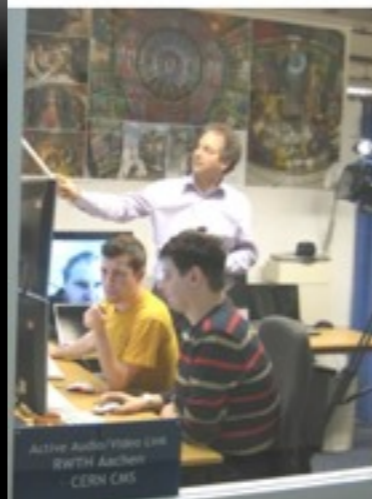
Web Interface to allow users to access information on EVO meetings and click to join directly from the portal

Meeting	Type	Hosts	Community	Description	Date & Time	Details
Introduction Meeting	UPMC	UPMC	UPMC	your language for learning		
UPMC TV	UPMC	UPMC	UPMC	Telepresence 2010/11 your leader in education with an		
UPMC TV	UPMC	UPMC	UPMC	UPMC TV	Mon, 20 Nov 2010 10:00	
UPMC TV	UPMC	UPMC	UPMC	UPMC TV	Mon, 20 Nov 2010 10:00	
UPMC TV	UPMC	UPMC	UPMC	UPMC TV	Mon, 20 Nov 2010 10:00	
UPMC TV	UPMC	UPMC	UPMC	UPMC TV	Mon, 20 Nov 2010 10:00	
UPMC TV	UPMC	UPMC	UPMC	UPMC TV	Mon, 20 Nov 2010 10:00	
UPMC TV	UPMC	UPMC	UPMC	UPMC TV	Mon, 20 Nov 2010 10:00	
UPMC TV	UPMC	UPMC	UPMC	UPMC TV	Mon, 20 Nov 2010 10:00	

Integration with Shibboleth and others authentication mechanisms will be added as we deploy the portal

EVO.CALTECH.EDU

CMS Centres Worldwide  
A New Collaborative Infrastructure



Lucas Taylor  
Fermilab and CMS

# Talking to people.

## Origins

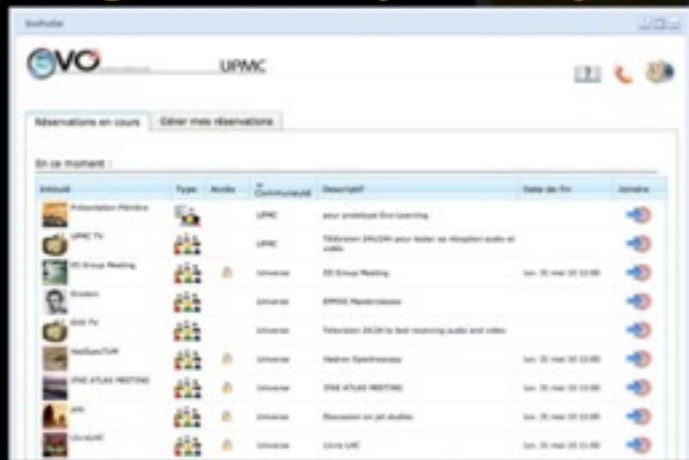
- ▶ 2007 – Fermilab sets up a CMS Remote Operations Centre
- ▶ 2008 – CMS Centre @CERN on main site
  - Opposite side of LHC to CMS Control Room
- ▶ **2009** – 16 CMS Centres Worldwide



## EVO Usage

## EVO Web Portal

Web Interface to allow users to access information on EVO meetings and click to join directly from the portal



Integration with Shibboleth and others authentication mechanisms will be added as we deploy the portal

EVO.CALTECH.EDU

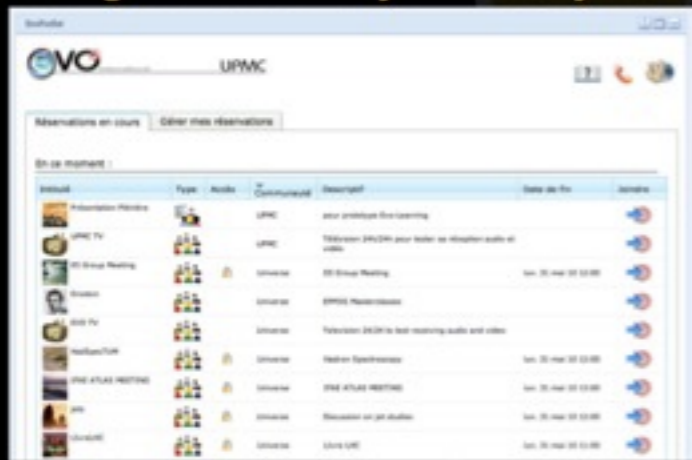


# Talking to people.

## EVO Usage

## EVO Web Portal

Web Interface to allow users to access information on EVO meetings and click to join directly from the portal



Observation	Type	Month	Comments	Description	Date on file	Details
Observation Meeting	UPMC	UPMC	your language for learning			
UPMC TV	UPMC	UPMC	UPMC	UPMC		
UPMC Meeting	UPMC	UPMC	UPMC	UPMC		
UPMC	UPMC	UPMC	UPMC	UPMC		
UPMC TV	UPMC	UPMC	UPMC	UPMC		
UPMC Meeting	UPMC	UPMC	UPMC	UPMC		
UPMC	UPMC	UPMC	UPMC	UPMC		
UPMC	UPMC	UPMC	UPMC	UPMC		
UPMC	UPMC	UPMC	UPMC	UPMC		

Integration with Shibboleth and others authentication mechanisms will be added as we deploy the portal

EVO.CALTECH.EDU

## Origins

## What are they used for ?

- ▶ People working together
  - Locally and via video
- ▶ Remote operations
  - Computing shifts, DQM, now some detector shifts
- ▶ Outreach and Education
  - Attracting students
  - VIP and media visits, events
    - E.g. "First 7 TeV collisions"
    - ~ 300 written articles
    - ~ 100 radio broadcasts
    - ~ 75 TV broadcasts



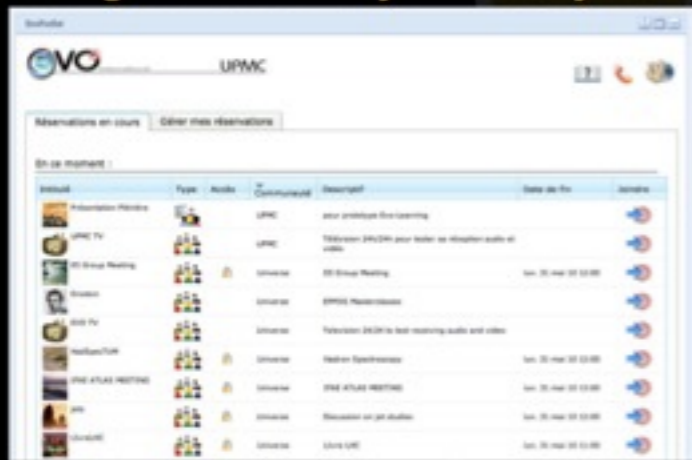


# Talking to people.

## EVO Usage

## EVO Web Portal

Web Interface to allow users to access information on EVO meetings and click to join directly from the portal



Observation	Type	Priority	Comments	Description	Date due to	Assignee
Observation 1	UPMC	High	your language for learning			
Observation 2	UPMC	Medium	Telepresence 2009/10 your leader in education with an			
Observation 3	UPMC	Low	UPMC Meeting		Nov. 20, 2009 02:00:00	
Observation 4	UPMC	Low	UPMC Meeting		Nov. 20, 2009 02:00:00	
Observation 5	UPMC	Low	UPMC Meeting		Nov. 20, 2009 02:00:00	
Observation 6	UPMC	Low	UPMC Meeting		Nov. 20, 2009 02:00:00	
Observation 7	UPMC	Low	UPMC Meeting		Nov. 20, 2009 02:00:00	
Observation 8	UPMC	Low	UPMC Meeting		Nov. 20, 2009 02:00:00	
Observation 9	UPMC	Low	UPMC Meeting		Nov. 20, 2009 02:00:00	
Observation 10	UPMC	Low	UPMC Meeting		Nov. 20, 2009 02:00:00	

Integration with Shibboleth and others authentication mechanisms will be added as we deploy the portal

EVO.CALTECH.EDU

Origins

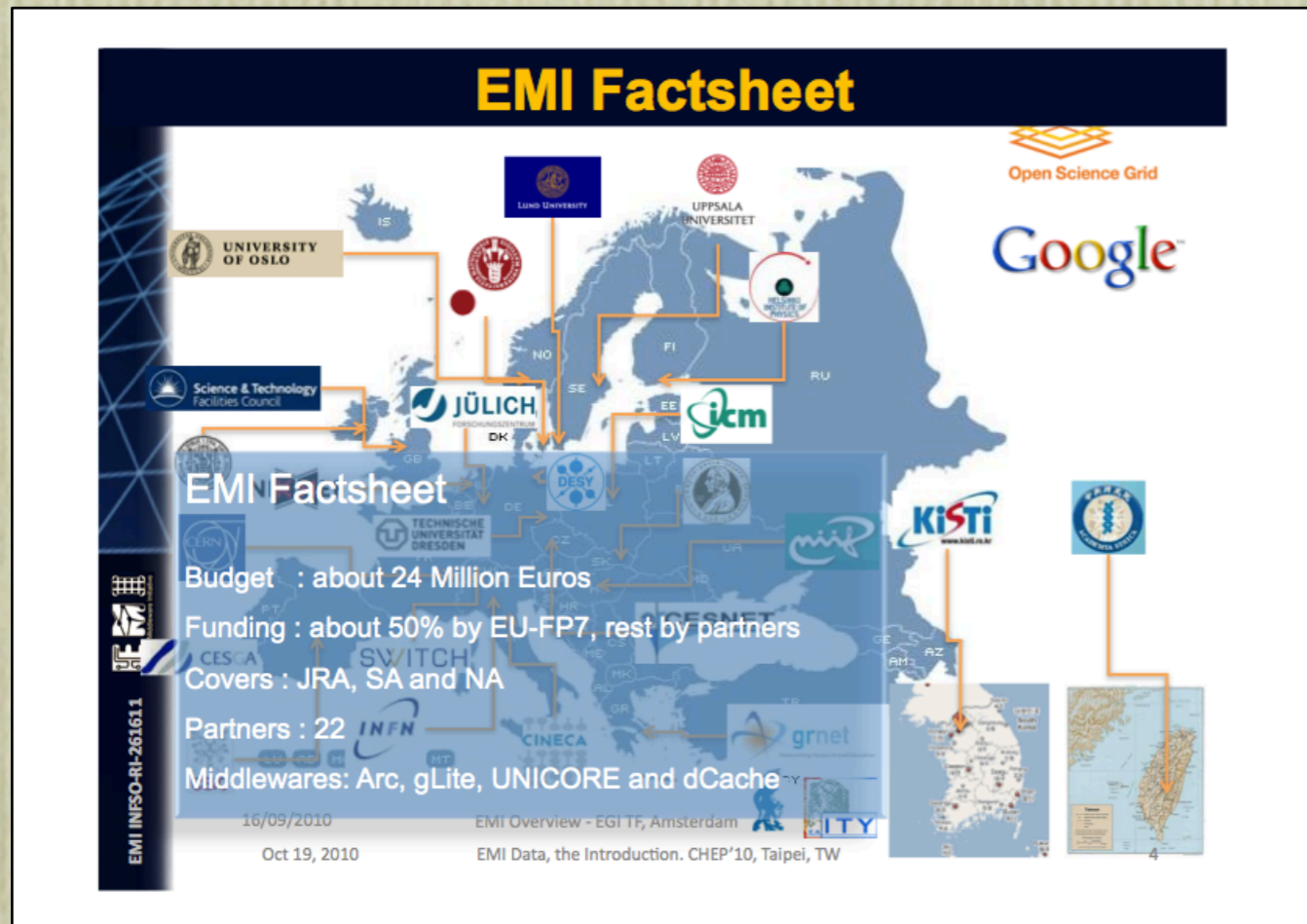
What are they used for?

Business model



- ▶ Franchise business model – each institute pays
- ▶ Standard design with commodity systems

# And another thing: EMI



# And another thing: EMI

## EMI Factsheet

**EMI Factsheet**

- Budget : about 24 Million Euros
- Funding : about 50% by EU-FP7, rest by partners
- Covers : JRA, SA and NA
- Partners : 22
- Middlewares: Arc, gLite, UNICORE and dCache

16/09/2010      EMI Overview - EGI TF, Amsterdam  
 Oct 19, 2010      EMI Data, the Introduction. CHEP'10, Taipei, TW

## What is EMI doing

### EMI Middleware Evolution

Stolen from Alberto Di Meglio

**Before EMI**      **3 years**      **After EMI**

**Before EMI:** KnowARC, gLite, UNICORE, dCache

**3 years:** EMI EUROPEAN MIDDLEWARE INITIATIVE

**After EMI:** Applications Integrators, System Administrators; Specialized services, professional support and customization; EMI Reference Services; Standards, New technologies (clouds) Users and Infrastructure Requirements; redhat, ubuntu

Oct 19, 2010      EMI Data, the Introduction. CHEP'10, Taipei, TW      7

# And another thing: EMI

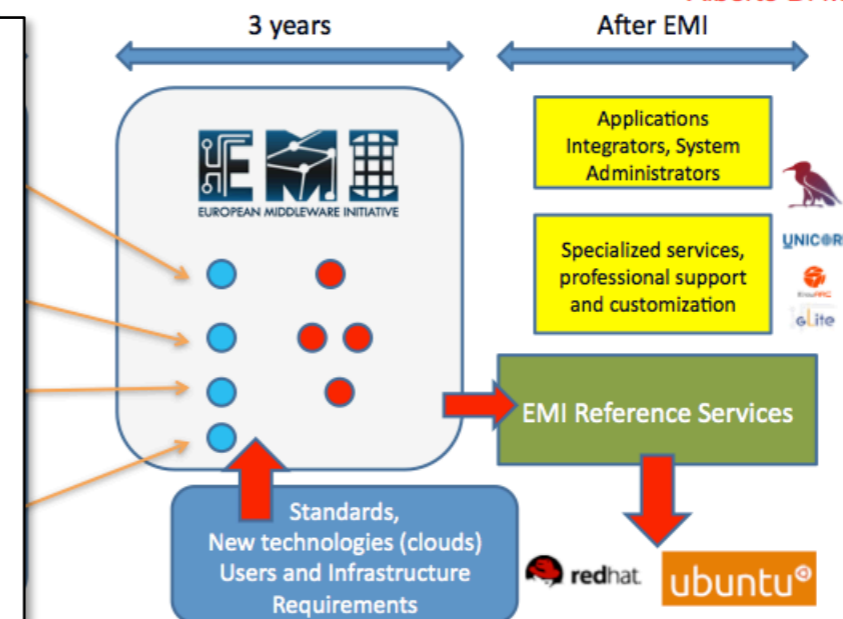
## EMI Factsheet



## What is EMI doing

### EMI Middleware Evolution

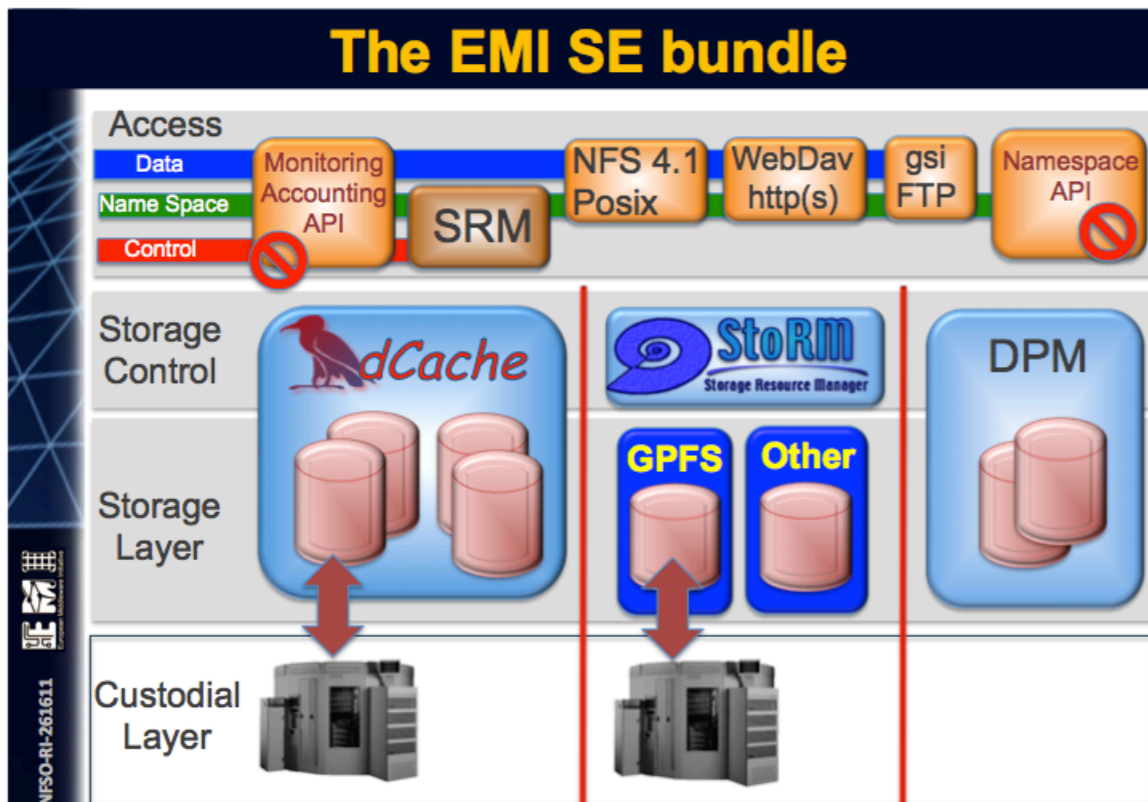
Stolen from Alberto Di Meglio



EMI Data, the Introduction. CHEP'10, Taipei, TW

7

## The EMI SE bundle



Oct 19, 2010

EMI Data, the Introduction. CHEP'10, Taipei, TW

14

EMI INFO-RI-261611

# And another thing: EMI

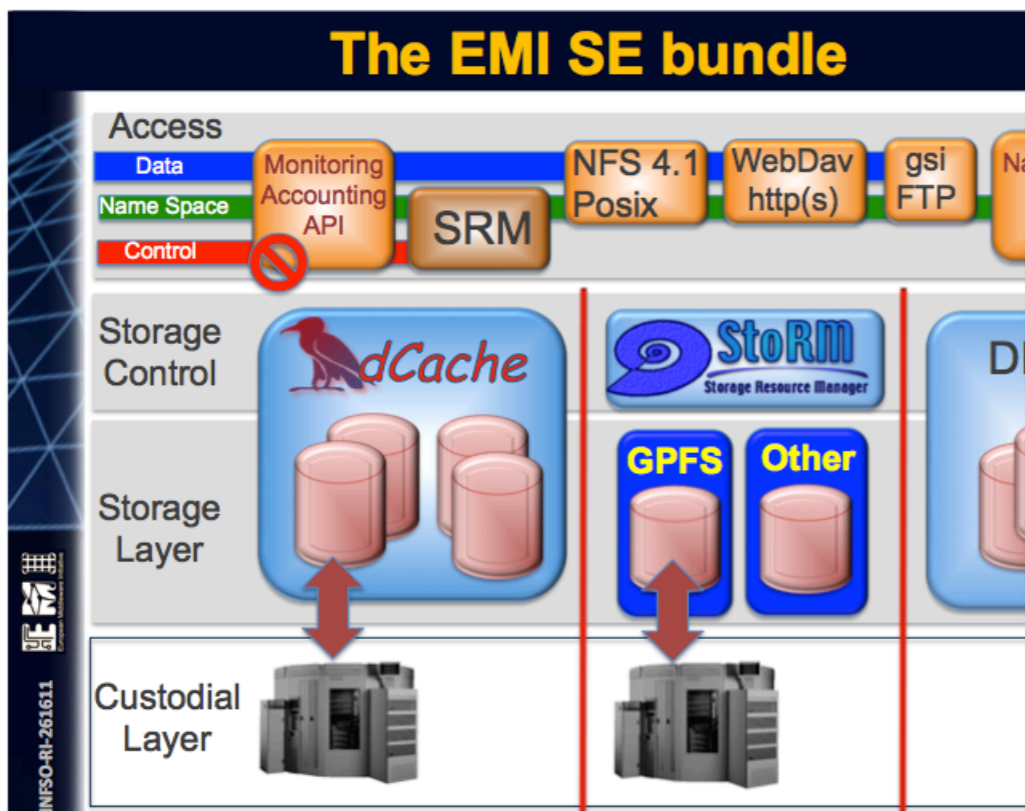
### EMI Factsheet

### What is EMI doing

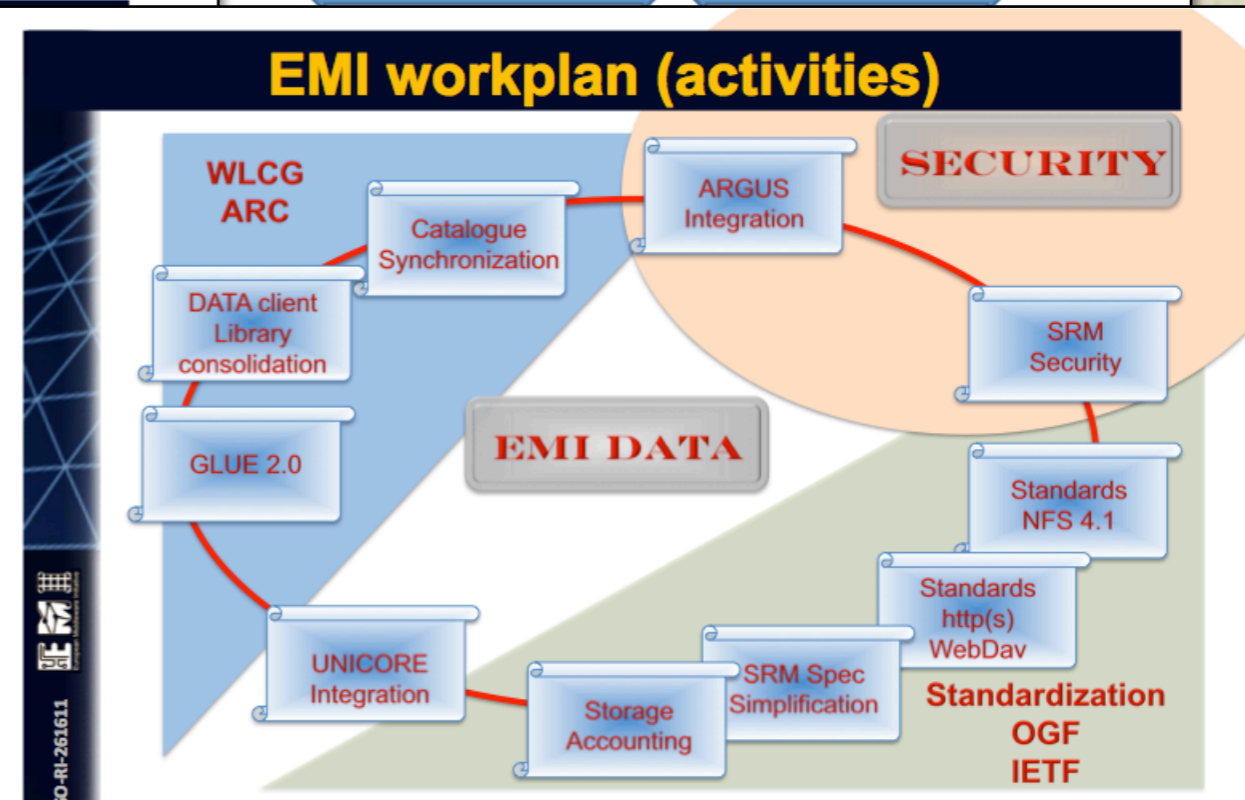
## EMI Middleware Evolution

Stolen from Alberto Di Meglio

3 years      After EMI



Oct 19, 2010      EMI Data, the Introduction. CHEP'10, Taipei, TW



Oct 19, 2010      EMI Data, the Introduction. CHEP'10, Taipei, TW

# The best plenary

- Lucas Taylor's CERN Outreach talk.



ATLAS, CMS and New Challenges for  
**Public Communication**



LIVE Geneva  
This Local

**CERN COLLIDER**  
Giant sub-atomic particle collider restarts  
WORLD NEWS ID OF THE CONFLICT WITH







**Lucas Taylor**  
Fermilab and CMS  
Head of Communications

with ... **Dave Barney**  
CERN and CMS

**Steve Goldfarb**  
Michigan and ATLAS

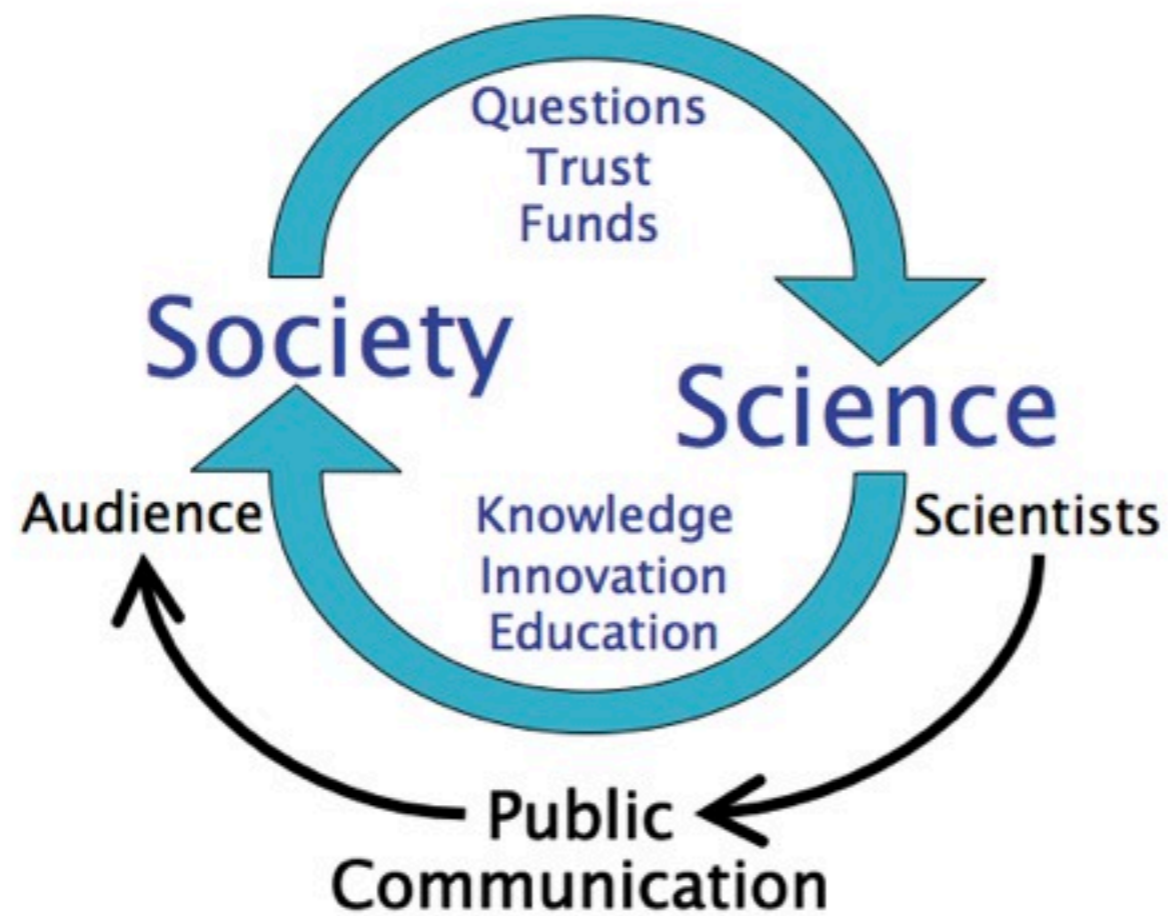
**Claudia Marcelloni**  
Berkeley and ATLAS



# The best plenary

- Lucas Taylor's CERN Outreach talk.

## Means of Communication



- (1) **Traditional media** (TV, radio, newspapers...)
- (2) **New media** (Web 2.0, Twitter, Facebook...)

# The best plenary

- Lucas Taylor's CERN Outreach talk



A majority of European citizens agree  
“Scientists do not put enough effort into  
informing the public about new  
developments in science and  
technology”

[http://ec.europa.eu/public\\_opinion/archives/ebs/ebs\\_340\\_en.pdf](http://ec.europa.eu/public_opinion/archives/ebs/ebs_340_en.pdf)



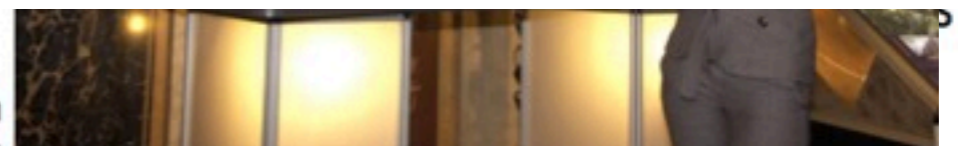
CMS Centre in the Austrian  
Parliament, Vienna

# The best plenary

- Lucas Taylor's CERN Outreach talk.

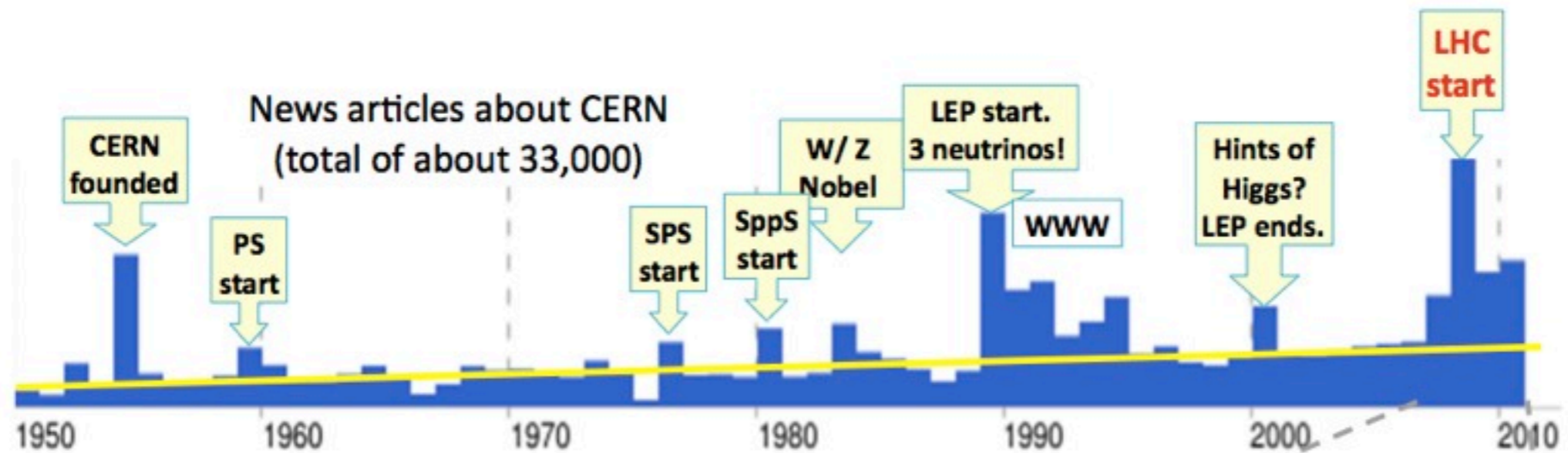
## LHC Communications Strategy

1. Coherent and high-quality messages
2. Open engagement with traditional media
3. Exploitation of new media (Web 2.0)

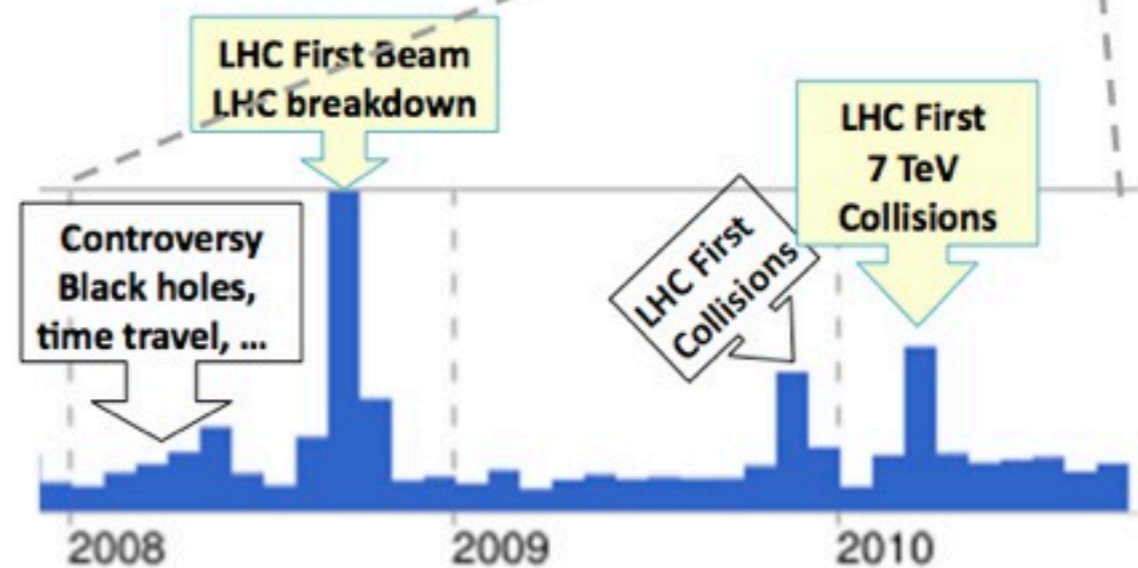


# The best plenary

How well are we doing?



- ▶ LHC is well above CERN historical trend
- ▶ Media events are clearly successful



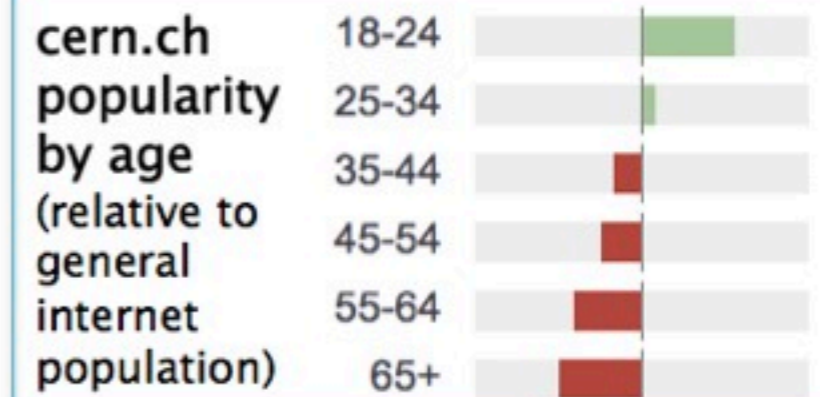
3

# The best plenary

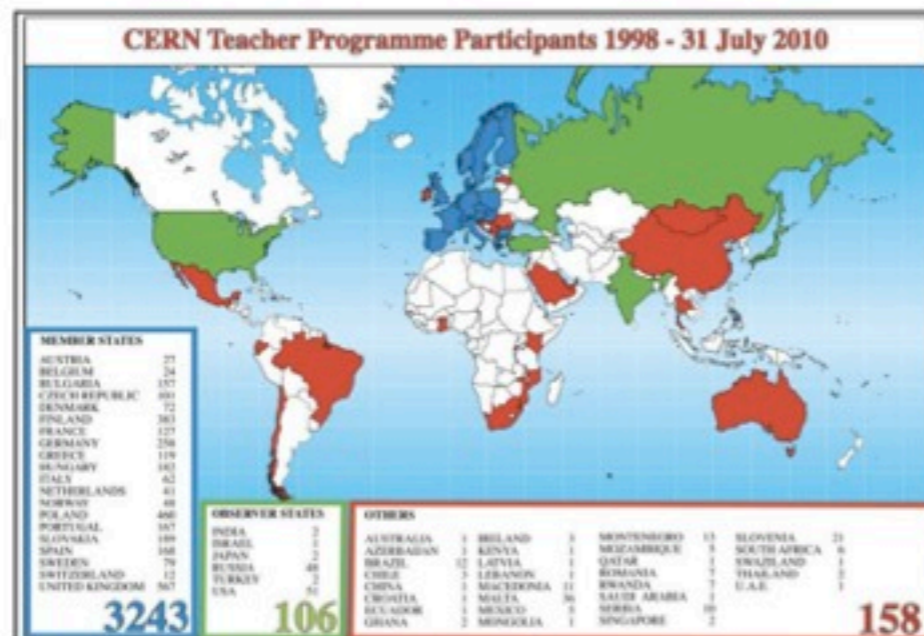
How about with younger people ?



Poll says not enough effort is made to reach young people



- ▶ 3,500 teachers went through CERN Teachers Programme and now teach O(100,000) school students at any time



# The best plenary

How well are we doing?

Language monitoring of online and print media

<http://www.languagemonitor.com/news/top-words-of-2009/>

## Top Phrases of 2009

1. King of Pop
2. Obama-mania
3. Climate Change
4. Swine
5. Too Large to Fail
- 6. Cloud Computing**
7. Public
8. Jai Ho!
9. Mayan Calendar
- 10. God Particle**

## Top Words of 2009

- 1. Twitter**
2. Obama
3. H1N1
4. Stimulus
5. Vampire
- 6. 2.0 (next gen.)**
7. Deficit
- 8. Hadron**
9. Healthcare
10. Transparency

## Top Names of 2009

1. Barack Obama
2. Michael Jackson
3. Mobama
- 4. Large Hadron Collider**
5. Neda Agha Sultan
6. Nancy Pelosi
7. M. Ahmadinejad
8. Hamid Karzai
9. Rahm Emmanuel
10. Sonia Sotomayor

with nothing at all LHC-related in 2008

# The best plenary

How well are we doing?  
What can you do ?



63% of respondents agree that

“scientists working at a university or government laboratories are the best qualified to explain scientific and technological developments”

[http://ec.europa.eu/public\\_opinion/archives/ebs/ebs\\_340\\_en.pdf](http://ec.europa.eu/public_opinion/archives/ebs/ebs_340_en.pdf)



10. God Particle

10. Transparency

10. Sonia Sotomayor

with nothing at all LHC-related in 2008

# The worst plenary (a moment of indulgence).

- ACER Marketing SSD talk.
- An object lesson in how marketing people will misuse graphs to attempt to mislead you.
- SLC (expensive) SSDs used for performance comparisons.
- MLC (cheap) SSDs used for price comparisons.
- Write performance carefully not explored much (15K HDDs and RAID0 both beat even some SLC SSDs at this).



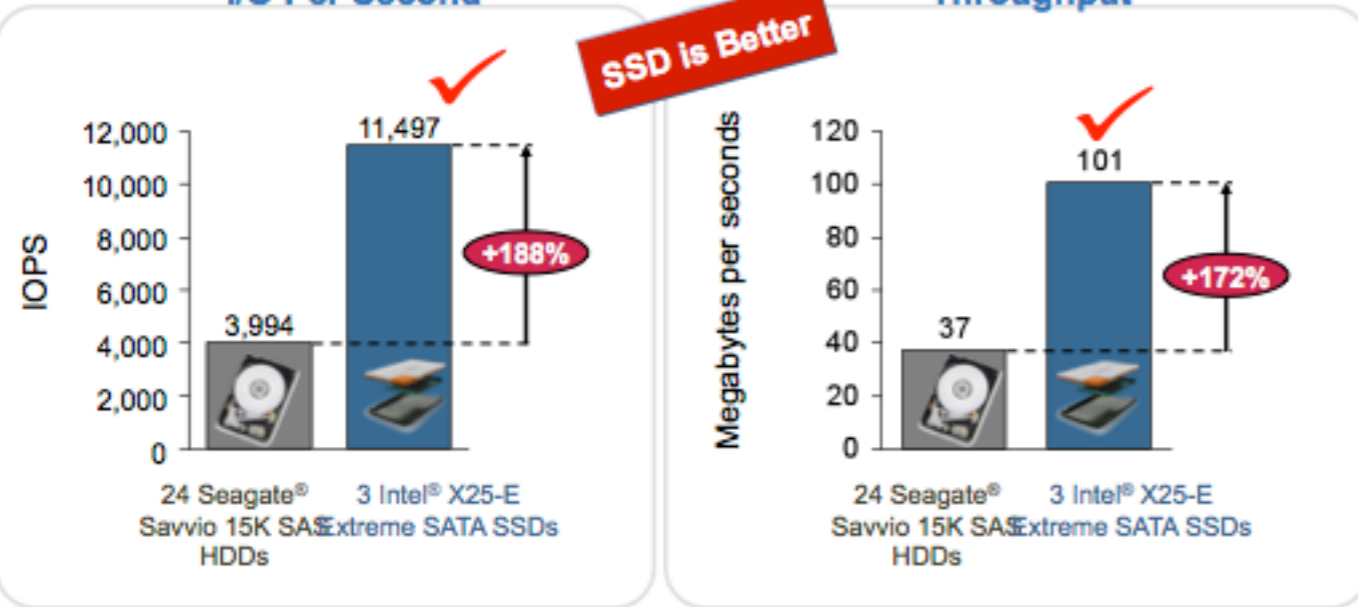
# The worst plenary (a moment of indulgence).

## HDD vs. SSD Throughput Performance

Price/performance?

What about writes?

I/O Per Second Throughput



- Hardware: a shelf of 3 Intel® X25-E Extreme SATA 32 GB SSDs and 24 Seagate® Savvio 15K SAS 73 GB HDDs. In both cases, Principled Tech. used a Newisys NDS-2240 enclosure.
- Software: Jetstress; Performance results in IOPS for the two storage configurations. A higher number of IOPS is better.
- Average throughput in MB per second for the two storage configurations. Higher throughput is better.

Source: Principled Technologies, 2009

talk.

marketing people will  
not to mislead you.

used for performance

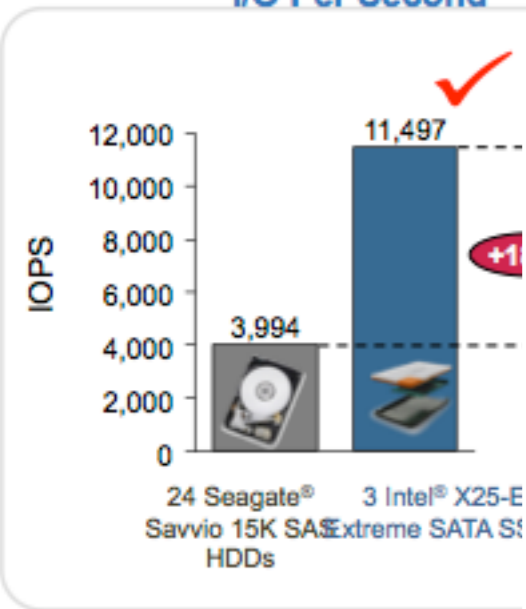
l for price comparisons.

- Write performance carefully not explored much (15K HDDs and RAID0 both beat even some SLC SSDs at this).

# The worst plenary (a moment of indulgence).

## HDD vs. SSD Throughput Performance

Price/performance? What about writes? **Throughput**



## HDD vs. SSD Read/Write Performance

If you ignore 15K HDDs and RAID arrays



Hardware: Intel® Core™2 Duo processor E8400 (3 GHz, 6 MB L2 cache, 1333 MHz FSB), 2 GB DDR2 Non-ECC SDRAM, 800 MHz  
 Software: IOMeter, measured as a secondary drive

- Write

much (15K HDDs and RAID0 both beat even some SLC SSDs at this).

alk.

people will  
you.

rmance

mparisons.

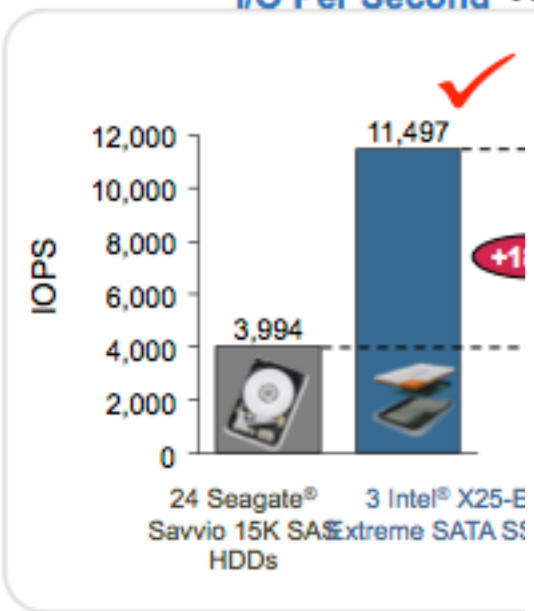
lored

Do both beat even

# The worst plenary (a moment of indulgence).

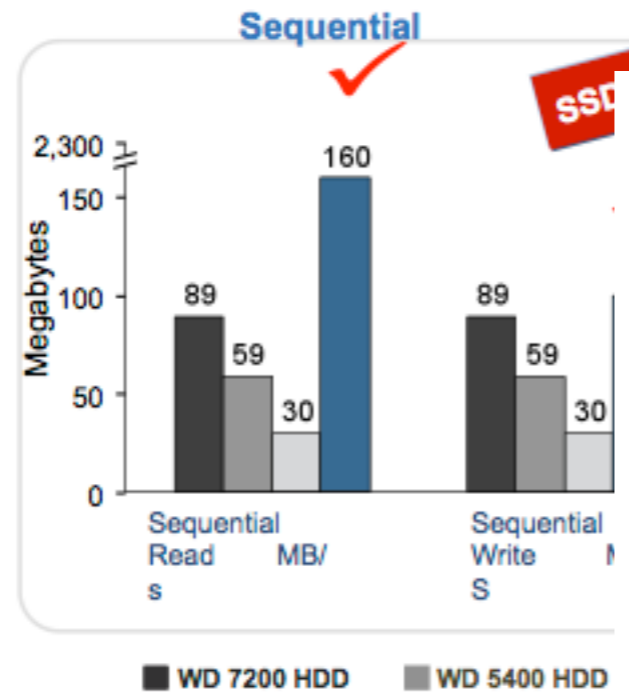
## HDD vs. SSD Throughput Performance

Price/performance? What about writes? **Throughput**



## HDD vs. SSD Read/Write Performance

If you ignore 15K HDDs and RAID arrays

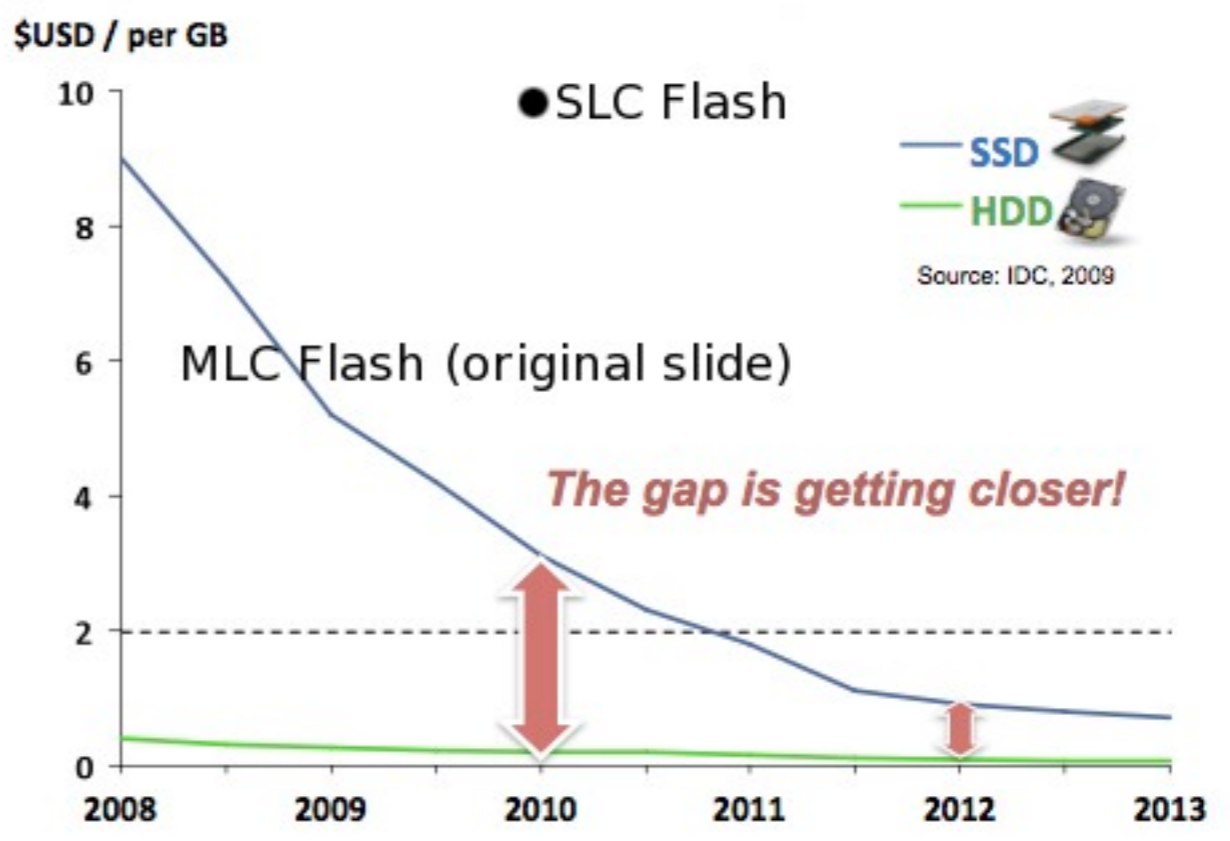


Hardware: Intel® Core™2 Duo processor E8400 (3 GHz), 2 GB DDR2 Non-ECC SDRAM, 800 MHz  
Software: IOMeter, measured as a secondary drive

alk.

people will you.

## Trends in Selling Prices



• Write much (15K HDDs and some SLC SSDs at

# The summarised Summary, concluded.

- Data Management is hard.
  - So we have to be cleverer about how we do it.
- But the LHC works.
- And lots of cool tech is waiting for us in the future, if we can use it.

# Conclusion

- 謝謝你們。你們要訊嗎？
- (Thank you. Do you have any questions?)

# List of talks referenced.

- **EMI, the Future of the European Data Management Middleware (PS15-1-463)**
- **Standard Protocols in DPM (PS15-3-443)**
- **LHC Data Analysis Using NFSv4.1 (pNFS): A Detailed Evaluation. (PS35-4-289)**
- **ng: What Next-Gen Languages Can Teach Us About HENP Frameworks in the Manycore Era [114]**
- **Parallelizing Atlas Reconstruction and Simulation: Issues and Optimization Solutions for Scaling on Multi- and Many-CPU Platforms (PS18-3-126)**
- **Algorithm Acceleration from GPGPUs for the ATLAS Upgrade [273]**
- **Adaptive Data Management in the ARC Grid Middleware (PS21-1-156)**
- **When STAR Meets the Clouds – Virtualization & Grid Experience (PS44-1-322)**
- **Tests of Cloud Computing and Storage System Features for Use in the H1 Collaboration Data Preservation Model (H1 Collaboration) (PS44-2-346)**
- **Establishing Applicability of SSDs to LHC Tier-2 Hardware Configuration (PS35-3-288)**
- **Distributing LHC Application Software and Conditions Databases Using CernVM File System (PS06-5-434)**
- **CernVM: Minimal Maintenance Approach to the Virtualization (PS29-4-432)**
- **ATLAS, CMS, New Challenges for the Public Communication (PL-08)**
- **One Small Step: A View of the LHC Experiments' Offline Systems After One Year of Data-Taking (PL -02)**
- **How to Harness the Performance Potential of Current Multi-Core CPUs and GPUs (PL-04)**
- **CMS Centres Worldwide - A New Collaborative Infrastructure (PS34-2-267)**