

CHEP '09 Summary

Jason A. Smith

What Is CHEP?

- Computing in High Energy and Nuclear Physics
- International forum to exchange information on computing experience and needs for the High Energy and Nuclear Physics community, and to review recent, ongoing, and future activities.
- Held in roughly 18 month intervals.
- 17th Conference in Prague, Czech Republic on March 21-27, 2009.

Main Themes

- Grid vs. Cloud Computing
 - Many people questioning if Grid has delivered on its promises and if there is a better alternative.
- Virtualization (Clouds, CernVM, Batch)
- Benchmarks (SI2K, SI2K-LCG, HEP SPEC06)
- Misc (Data Centers & Cooling, Lustre, SL5, Multi-cores, Intel Atom, Software Installation)

Belle Monte-Carlo production on the Amazon EC2 cloud

Martin Sevier, Tom Fifield (University of Melbourne)
Nobuhiko Katayama (KEK)



Hradcany Castle and Charles Bridge, Prague

17th International Conference on Computing in High Energy and Nuclear Physics
21 - 27 March 2009 Prague, Czech Republic

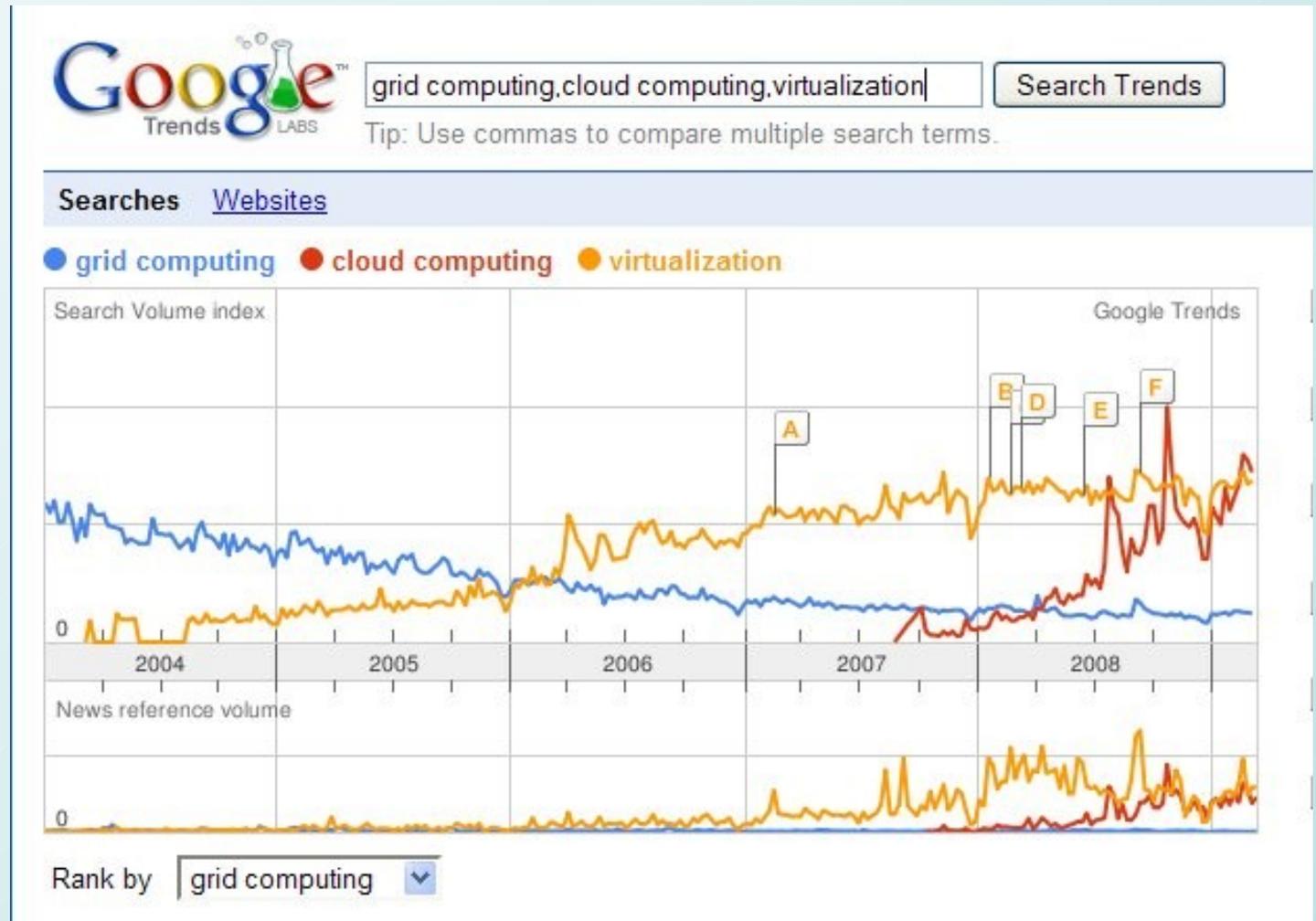
“Cloud Computing”

Cloud Computing has captured market interest

Cloud computing makes large scale computing resources available on a commercial basis

A simple SOAP request creates a “virtual computer” instance with which one can compute as they wish

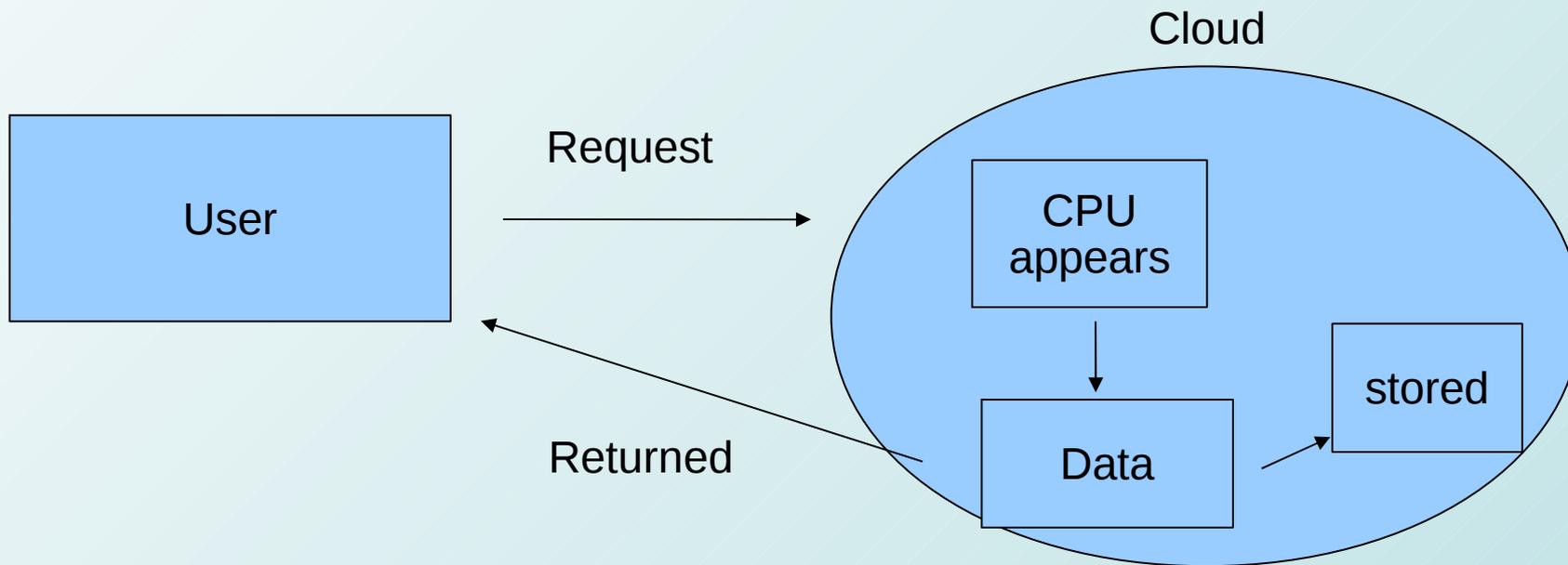
Internet Companies have massive facilities and scale, order of magnitude larger than HEP



Can we use Cloud Computing to reduce the TCO of the SuperBelle Computing?

Cloud Computing

Internet companies have established a Business based on CPU power on demand, one could imagine that they could provide the compute and storage we need at a lower cost than dedicated facilities.

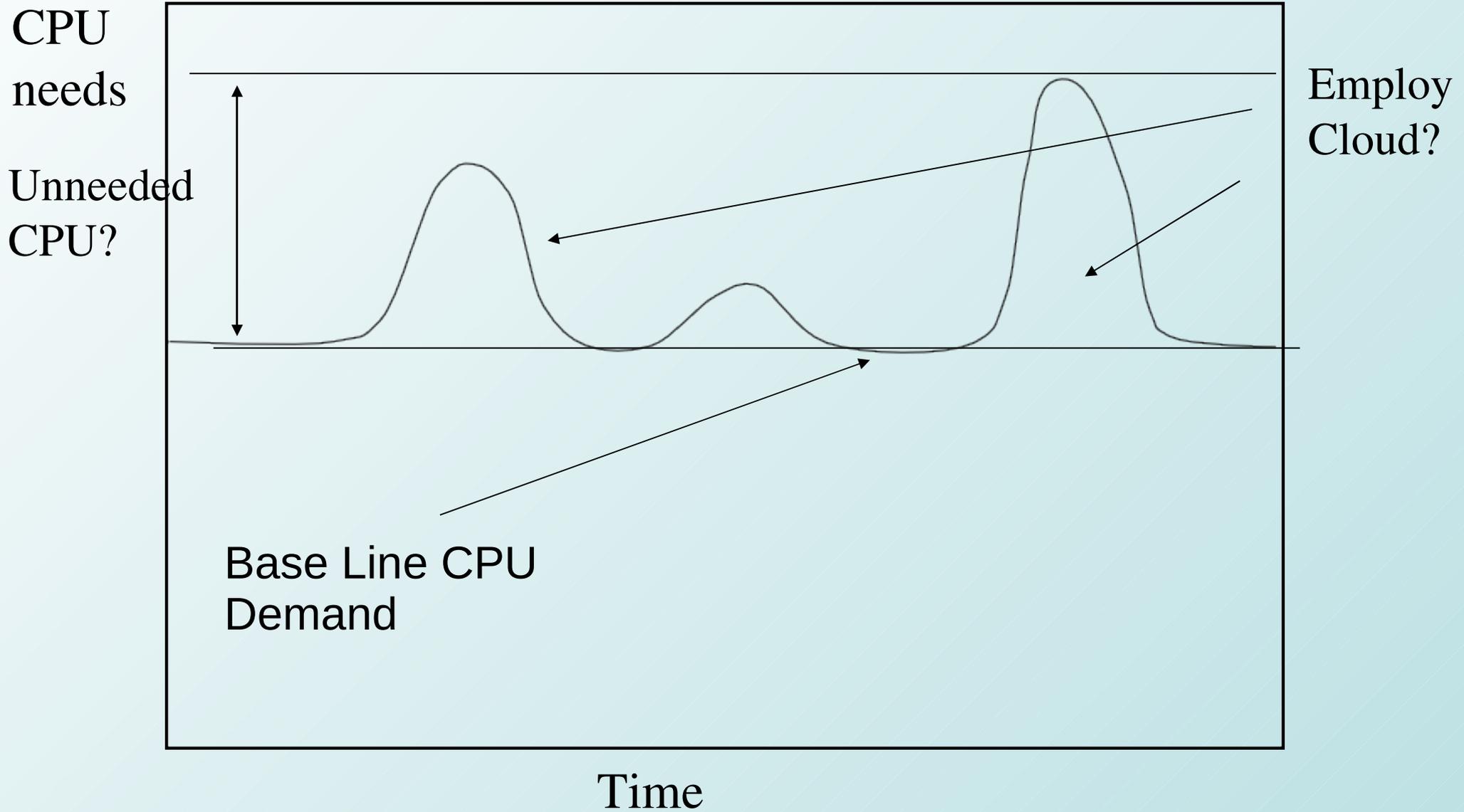


Resources are deployed as needed.

Pay as you go.

MC Production is a large fraction of HEP CPU - seems suited to Cloud

Particularly useful for Peak Demand



Costs

- Managed to do 752,233 events in time for this presentation
- CPU cost: \$80
 - ◆ 20 Instances, 4 hours 57minutes
- Storage cost: \$0.20
 - ◆ Storage on S3: Addbg 3.1Gb, pgen 0.5Gb, results 37Gb, \$6.08/month or \$0.20/day
- Transfer cost: \$6.65
 - ◆ Addbg, pgen in: \$0.36, mdst out: \$6.29
- **Total Cost: \$86.85**

Naïve early estimate without automation and storage overhead ~\$40

Need to get equivalent times for GRID production of MC data

Conclusions

- Value Weighted Output – metric to estimate the present time value of CPU
- Can make a few more tweaks to minimize costs
- Charged for the period of time we claim the AMIs
- Keep AMIs active!
- Cloud is promising for MC production
- Can deliver Peak Demand if needed
- Transfer speeds from S3 to AMI likely too slow for large scale data analysis for HEP
- On-demand creation of virtual machines is a flexible way of utilizing Computational Grids

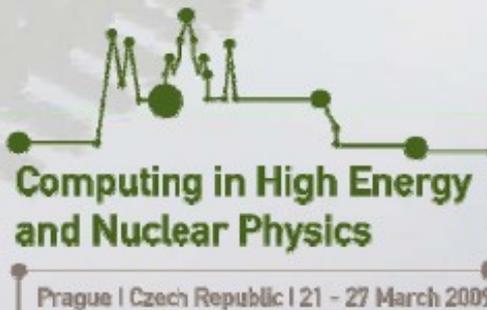
Thank you!



Computing for the RHIC Experiments

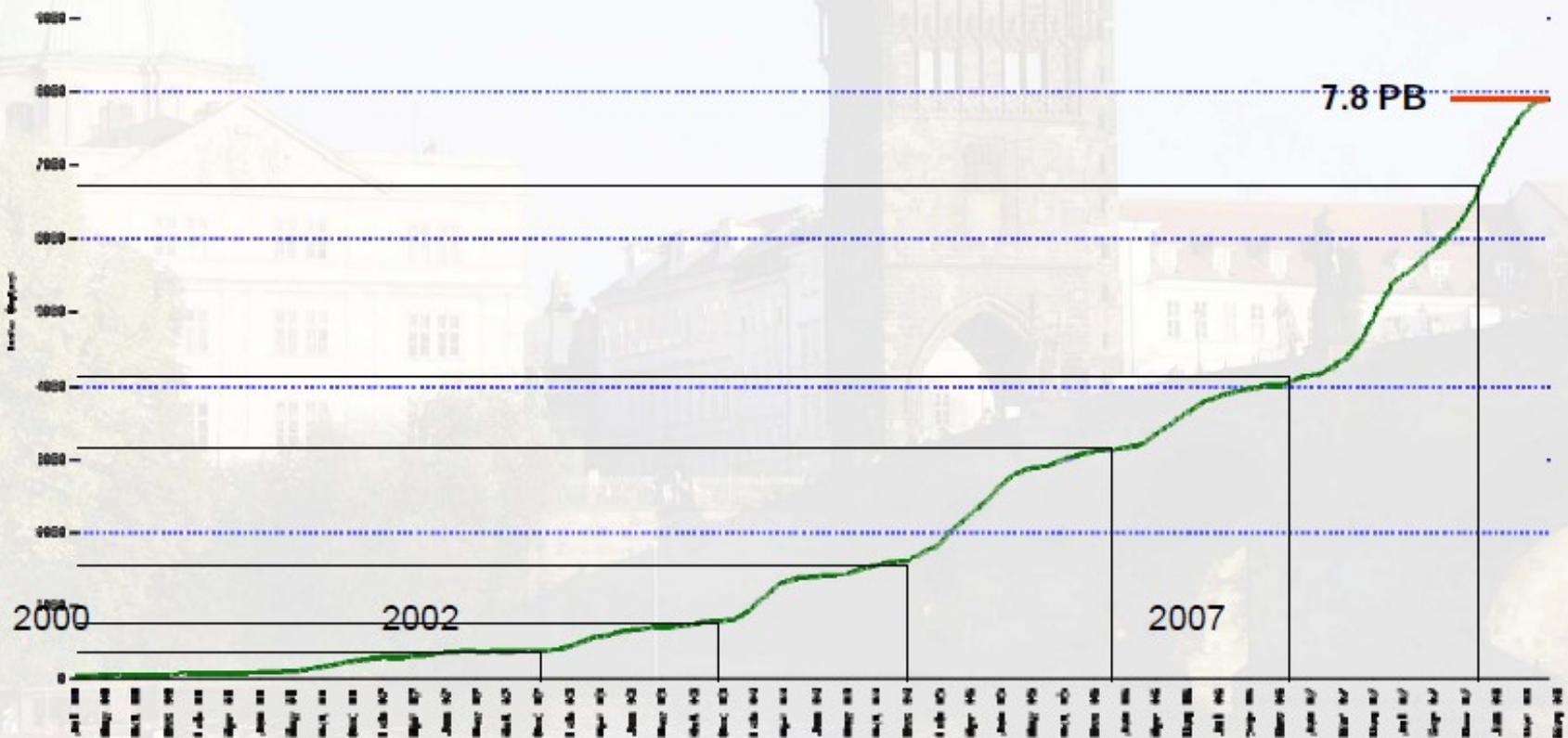
Jérôme LAURET

Brookhaven National Laboratory
CHEP 2009



Data so far, ...

Data Volume archived at the RACF (managed by HPSS)



U.S. DEPARTMENT OF
ENERGY

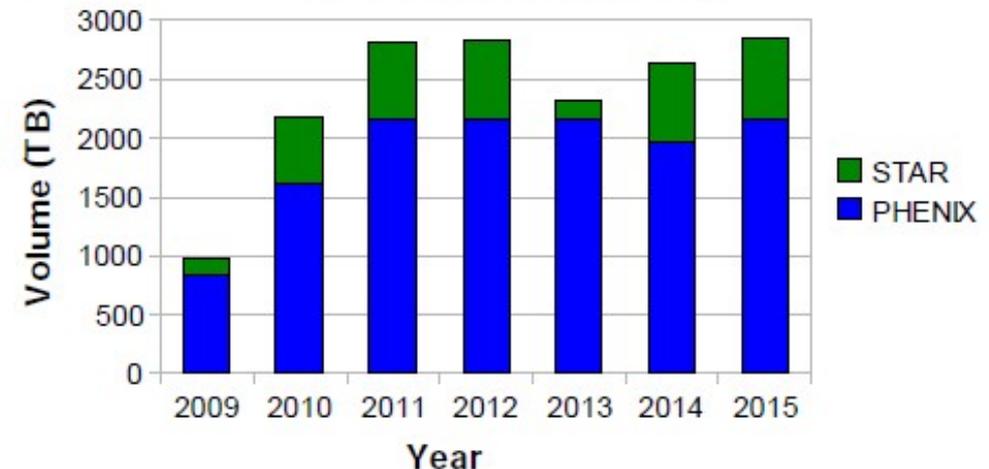
Jérôme LAURET for RHIC
CHEP 2009, March 21-27 - Praha / Czech Republic

BROOKHAVEN
NATIONAL LABORATORY

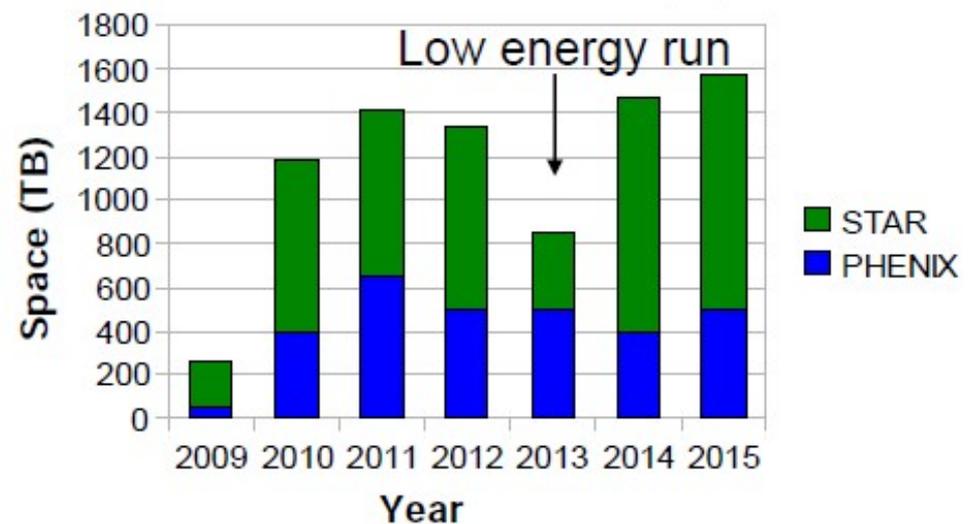
Data growth outlook

- Initial model:
 - fraction of data from previous years on disk and/or analyzed?
 - **WRONG!**
- RHIC Experience:
 - nearly all data from all years are being constantly analyzed, cross-compared, merged (analysis)
- ~ 1/2 of the cost in storage (tracked 2005-2008)

Estimated raw data volume (TB)
for PHENIX and STAR exp.



Derived data volume (TB)



Facility and relation to experiment

■ BNL/Tier0 Facility – RACF

- Mission: *Online Recording of Raw Data, Production reconstruction of Raw Data, Primary Facility for Data Selection and Analysis, Long time Archiving and Serving of all Data*
- Share, leverage, consolidate, focus on robust solutions
- Maximize CPU cycles – Shared (queues) if not used (cross experiments, EOL)

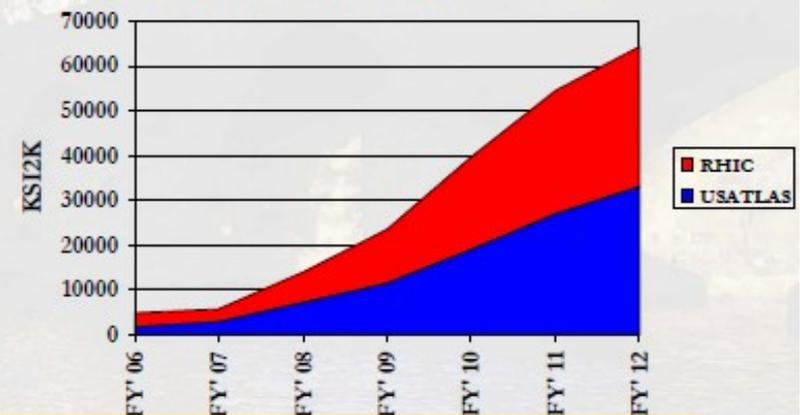
■ Procurement cycles

- Base funding for equipment shared by the experiments and the facility
- Cycle: 5 years plans, long term projections
- Common pool for facility + experiments

■ Issue

- Facility + experiment shared pool of money Zero sum principle ⇔ balance
- Storage: tape is a fixed cost
- CPU needed for processing
 - # of passes @ RHIC have been low (<< 3)
 - Implied Out-sourcing from the start

Processing Power - 65 MSI2k in 2012. Numbers may change with RHIC revised plan



Distributed computing



U.S. DEPARTMENT OF
ENERGY

Jérôme LAURET for RHIC
CHEP 2009, March 21-27 - Praha / Czech Republic

BROOKHAVEN
NATIONAL LABORATORY

Grid-ing or not Griding?

- What are the RHIC experiments doing Grid-wise (data movement apart)?
 - STAR: only active experiment to routinely run jobs on Grids (+dev)
 - **So, what is/are the problems if any?**

- Are Grids usable?
 - Outstanding efficiencies – efficiency > 97%
 - Operation support from Grid projects (OpenScience-Grid)
 - Justified to move all STAR Monte-Carlo productions on Grid (2006)
 - ✓ **USABLE**

- Where are the problems for production environments?
 - Grids are complex and too dynamic for production environment
 - Troubleshooting is simply inadequate (globus error # anyone?)
 - VO mainly using dedicated sites with pre-installed software stack
 - Little to no opportunistic use



Open Science Grid



Or is it Clouding or ...?

■ Are Cloud usable?

- STAR Use *Amazon/EC2* / Elastic Cloud Computing (Nimbus / Test in 2007/2008)
- Scale & Performance: ~ 300 jobs at all times, weeks long
 - Similar efficiencies than normal Grids measured so far
 - 5 MB/sec data transfer / WN – for simulation, enough
 - **NOT A SILVER BULLET** (under the hood, still the grid stack)

ContribID # 516

ContribID # 475

✓ **USABLE**

- Status: STAR run on EC2 to handle MC production (event generator + response simulator + full reconstruction) – Emergency request
 - **Results have been used for analysis to be presented for Quark Matter 2009**
[real practical use of Clouds helping science deliverables]

■ Economics of Clouds remain puzzling (within range of facility costs to first order)

- Cons: MSS unlikely on Clouds, Network performance low
- Pros : Truly opportunistic used at reach, software provisioning is immediate to any site
- **IMMEDIATE benefits, LEAST efforts, MAXIMAL confidence**

■ Prospects? Technology rapidly changing ...

- Grid and clouds are **NOT** orthogonal – VM provide on the fly resources
- Integrating technology on OSG, enhance/complement grids
- Truly opportunistic implies network dynamic circuit provisioning?
- ...

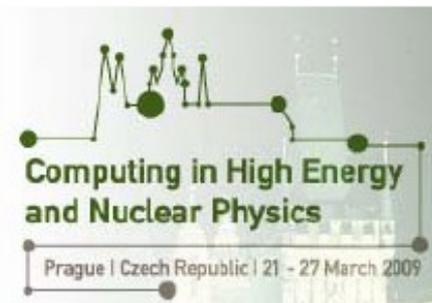


■ Are we ready?





Grid Middleware for WLCG



Where are we now – and where do we go from here?

Prague, 24th March 2009

Ian Bird, James Casey,
Oliver Keeble, Markus Schuler
& thanks to Pere Mato

CERN





Evolution has been

- Simplifying grid services
 - Experiment software has absorbed some of the complexity,
 - Computing models have removed some of the complexity,
- Grid developments have not delivered:
 - All the functionality asked for
 - Reliable, fault tolerant services
 - Ease of use
- But requirements surely were overstated in the beginning
 - And grid software was less real than we had thought ...
 - And as Les Robertson said, technology has moved on



What works?

- Single sign-on – everyone has a certificate, we have a world-wide network of trust
 - VO membership management (VOMS), also tied to trust networks
- Data transfer – gridftp, FTS, + experiment layers;
 - Demonstrate full end-end bandwidths well in excess of what is required, sustained for extended periods
- Simple catalogues – LFC
 - Central model – sometimes with distributed read-only copies (ATLAS has a distributed model)
- Observation: The network – probably the most reliable service – fears about needing remote services in case of network failure probably add to complexity
 - i.e. Using reliable central services may be more reliable than distributed services



What else works

- Databases – as long as the layer around them is not too thick
 - NB Oracle streams works – but do we see limits in performance?
- Batch systems and the CE/gateway
 - After 5 years the lcg-CE is quite robust and (is made to) scales to today's needs ... But must be replaced (scaling, maintenance, architecture, ...). Essentially a reimplement of the Globus gateway with add-ons
- The information systems – BDII – again a reimplement of Globus with detailed analysis of bottlenecks etc.
 - GLUE – is a full repository of experience/knowledge of 5 years of grid work – now accepted as an OGF standard
- Monitoring, accounting
 - Today provides a reasonable view of the infrastructure
- Robust messaging systems – now finally coming as a general service (used by monitoring ... Many other applications)
 - Not HEP code!



What about...

- Workload management?
 - Grand ideas of matchmaking in complex environments, finding data, optimising network transfer etc
 - Was it ever needed?
 - Now pilot jobs remove the need for most (all?) of this
 - Even today the workload management systems are not fully reliable despite huge efforts
- Data Management
 - Is complex (and has several complex implementations)
 - SRM suffered from wild requirements creep, and lack of agreement on behaviours/semantics/etc.



A view of the future

- WLCG could become a grid of cloud-like objects:
- Still have many physical sites
- But hide the details with virtualisation –
- What else is useful?
 - Virtualisation
 - Pilot jobs
 - File systems
 - Scalable/Reliable messaging services
 - Remote access to databases
 - Simplified data management interfaces (is Amazon too simple?)



The facility ...

- Goal to decouple the complexities and interdependencies:
- Ability to run virtual machines
 - Still need the batch systems – fairshares etc
 - Need to be able to manage VMs (LSF, VMWare, ...)
 - Tools for debugging (e.g. Halt and retrieve image?)
- Entry point:
 - CE? – but can now be very simple
 - Mainly needs to be able to launch pilot factories (may even go away?)
 - Need to be able to communicate fully with the batch system – express requirements and allow correct scheduling
 - Information published by site directly to a messaging system (rather than via 3rd party service)
 - Probably need caching for delivery of software environments etc
- The complexities of OS/compiler vs middleware vs application environment vs application interdependencies goes away from the site (to the experiment!)



Virtual machines at a site



Site installs and maintains:

- OS, compiler
- Middleware

VO at every site installs:

- App environment

Complex dependencies between all layers

Site installs and maintains:

- bare OS

Experiment installs (~once!):

- pilot VM

- Imagine that sw env installed in pilot via cache at site

- Almost no dependencies for site

Site could also provide VM for apps that want a "normal" OS environment, need tools to manage this. This is like Amazon – the app picks the VM it needs, either a standard one, or its own



Conclusions

- We have built a working system that will be used for first data taking
 - But it has taken a lot longer than anticipated ... and was a lot harder ... and the reality does not quite match the hype ...
- We now have an opportunity to rethink how we want this to develop in the future
 - Clearer ideas of what is needed
 - And must consider the risks, maintainability, reliability, and complexity
- It was always stated that ultimately this should all come from grid providers
 - Not quite there yet, but a chance to simplify ?

Will / Can Clouds Replace Grids?

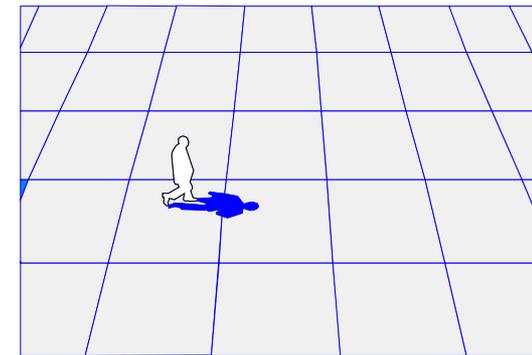
A Three-Point Checklist

Jamie.Shiers@cern.ch

Grid Support Group, IT Department, CERN

What is Grid Computing?

- Today there are many definitions of **Grid computing**:
- The definitive definition of a Grid is provided by [1] Ian Foster in his article "**What is the Grid? A Three Point Checklist**" [2].
- The three points of this checklist are:
 1. Computing resources are not administered centrally;
 1. Open standards are used;
 - 1. Non-trivial quality of service is achieved.**



What is Cloud Computing?

- a. The latest in a series of hype;
- a. Yet another form of utility computing;
- a. Grid Computing but with a business model;
- a. Where the action (money) is currently at;
- a. All of the above?

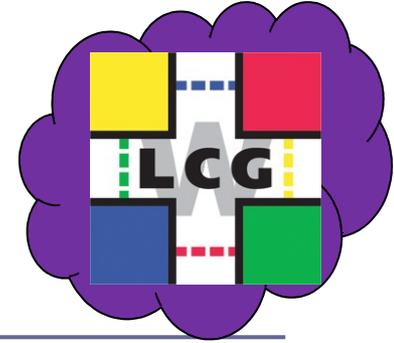
Remaining Questions

- Are Grids too complex?
- *Do Grids have to be too complex?*
- Are Clouds too Simple?
- *Do Clouds have to be too simple?*

IMHO we can learn much from the strengths and weaknesses of these approaches, particularly in the key (for us) areas of data(base) management & service provision. This must be a priority for the immediate future....

Can Clouds Replace Grids? - The Checklist

- We have established a short checklist that will allow us to determine whether clouds can replace – or be used in conjunction with – Grids for LHC-scale data intensive applications:
 - 1. Non-trivial quality of service must be achieved;**
 - 1. The scale of the test(s) must be meaningful for petascale computing;**
 - 1. Data Volumes, Rates and Access patterns representative of LHC data acquisition, (re-)processing and analysis;**
 - 1. Cost (of entry; of ownership).**



Conclusions

- **We cannot afford to ignore major trends in the computing industry**
 - Some may turn out to be dead-ends
 - Some may die only to be reborn in a different guise

- **We have established – through a long series of challenges – a well-proven mechanism for determining whether a (set of) computing service(s) satisfies an agreed set of requirements**

- **Not evaluating cloud computing for at least some HEP Use Cases would appear to be the one option we cannot afford to take...**

LHC Data Analysis will start on the Grid

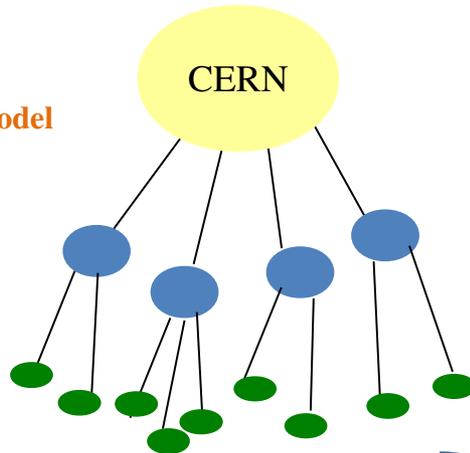
What's Next?



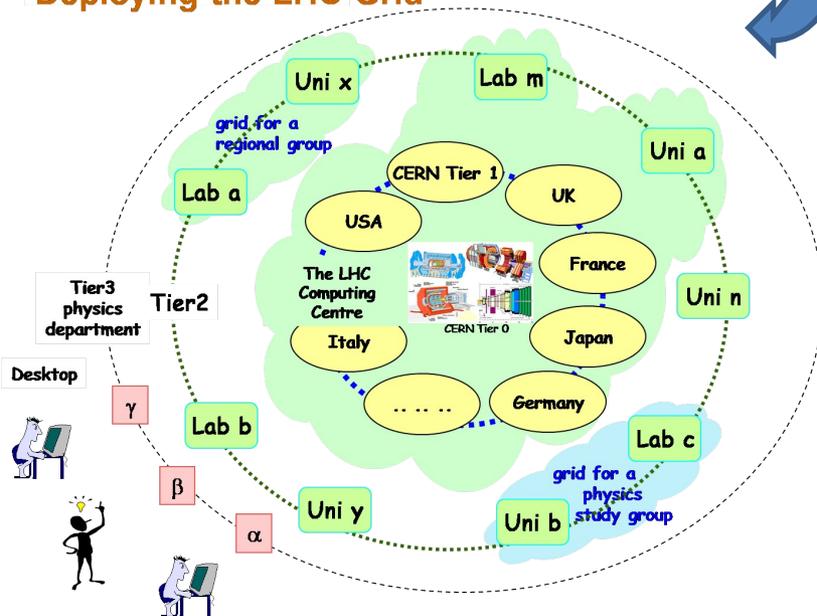
Then came the GRID

- 1999 – Grid
 - More flexible → easier to use, adapt to the reality of data analysis
 - But – more complex to build and to manage
 - However – the basics were already there!
 - Prospect of a general science grid (memories of research networks)
 - With expectation of non-HEP funding
 - during development
 - and for long term operation

The MONARC Model



Deploying the LHC Grid



- Consensus on this approach emerged during CHEP 2000 in Padova

2000s - the Decade of the GRID

- E-Science in fashion
- Many grid/science/physics projects funded in Europe and US
 - Stimulated international collaboration – open to all LHC sites
 - WLCG could operate on top of these multi-science infrastructure grids – EGEE, OSG, ..
- de-facto standards
 - Significant non-HEP funding was made available to LHC groups and centres – supporting operation, tools and middleware, application development and adaptation to grids
- Significant industrial interest came
 - added confusion and went
- But many other sciences and also some industries have ported applications to the HEP style of Grid

As LHC starts, data handling depends on a grid But is the model of a general science grid sustainable?

- WLCG operates on top of multi-science grids
 - With short-term funding cycles <> incompatible with long-term services
 - Hard to find other sciences outside physics that *depend* on these grids
 - Proposal in Europe for a long-term infrastructure (EGI), but still some way from agreement and approval, and EGEE ends next April
right in the middle of the first LHC run
 - Open Science Grid with a similar role to EGEE in US has 5-year funding from NSF and DoE through 2010
- Could be problematic - but ..
 - Tier-1 sites are still at the heart of EGEE and OSG operations
 - HEP institutes, collaborators and experiments are to a large extent responsible for the middleware
- So WLCG and LHC funding agencies can and **surely will** take on the necessary operational responsibility if EGEE and/or OSG close down

Energy

- “Data centres consumed 1 per cent of the world’s electricity in 2005. By 2020 the carbon footprint of the computers that run the internet will be larger than that of air travel, a recent study by McKinsey and the Uptime Institute predicted.”

Times Online - September 2008

- Even if the cost of oil is down at ~\$50, if this growth rate really continues -
power-efficient data centres and cheap renewable energy must be essential components of any infrastructure that is being planned today
- ☺ **The distributed (grid) model enables us to incorporate data centres wherever they may be located, and whoever is running them**

Grids versus Clouds



Cloud v. Grid

- Clouds aim at efficient sharing of the hardware
 - low-level execution environment, Isolation between users
 - Operated as a homogeneous, single-management domain
 - Straight-forward i/o and storage
 - Expose only a high-level view of the environment - scheduling, data placement, performance issues are hidden from the application and the user
- Grids aim at collaboration
 - Add your resources to the community, but retain management control
 - Expose topology - location of storage, availability of resources
 - Choice of tools to hide the complexity from the user, and the application can write its own tools
- **Both need complex middleware to function**
 - Grids had a problem in trying to provide a universal high-functionality environment (OS, data management,), with intersecting collaborations and a naturally competitive environment
 - Clouds have an advantage in offering a simpler base environment, leaving much of the functionality to the application - where universal solutions are not necessary - and what they do have to provide can be decided within a single management hierarchy
- As the names suggest -
the grids are transparent and the clouds are opaque

Grids and Clouds



.. and Mobility

- ADSL at 20 Mbps, WiFi/WiMAX/3G
 - We are close to having good bandwidth data connections almost everywhere we go
 - And we already have a powerful high capacity computer in the backpack
- **This is where end-user analysis is going to be done**
- The physicist's notebook must be integrated with the experiment environment, the physics data, and the grid resources
- Without burdening the notebook or its user
- The grid environment is too complex to be extended to the notebook
- Ganga does a good job of bridging these environments
- ☺ **The approach of cernVM looks like the right direction for analysis, enabling the end-user to cache the data she needs and extend her environment on to the grid, or cloud, or ...**

Summary

- Grids are all about sharing.
 - groups distributed around the world can pool their computing resources
 - large centres and small centres can all contribute
 - users everywhere can get equal access to data and
- Grids are also flexible
 - place the computing facilities in the most effective and efficient places
 - exploiting funding wherever it is provided
- HEP and others have shown that
 - grids can support computational and storage resources on a massive scale
 - that can be operated around the clock
 - running hundreds of thousands of jobs every day
- The grid model has stimulated high energy physics to organise its computing
 - in a widely distributed way
 - building a collaboration involving directly a large fraction of the LHC members and their institutes
- This will be the workhorse for production data handling for many years and as such must be maintained and developed through the first waves of data taking

But - the landscape has changed dramatically over the past decade

- The Web, the Internet, powerful PCs, broadband to the home, ...
 - have stimulated the development of new applications that generate a massive demand for computing remote from the user
 - that is being met by giant, efficient facilities deployed around the world
 - and creates a market for new technologies capable of operating on a scale equivalent to that of HEP
 - Whether or not commercial clouds become cost-effective for HEP data handling is only a financial and funding-agency issue
- BUT
- Exploiting the associated technologies is an obligation

Could there be a revolution here for physics analysis?

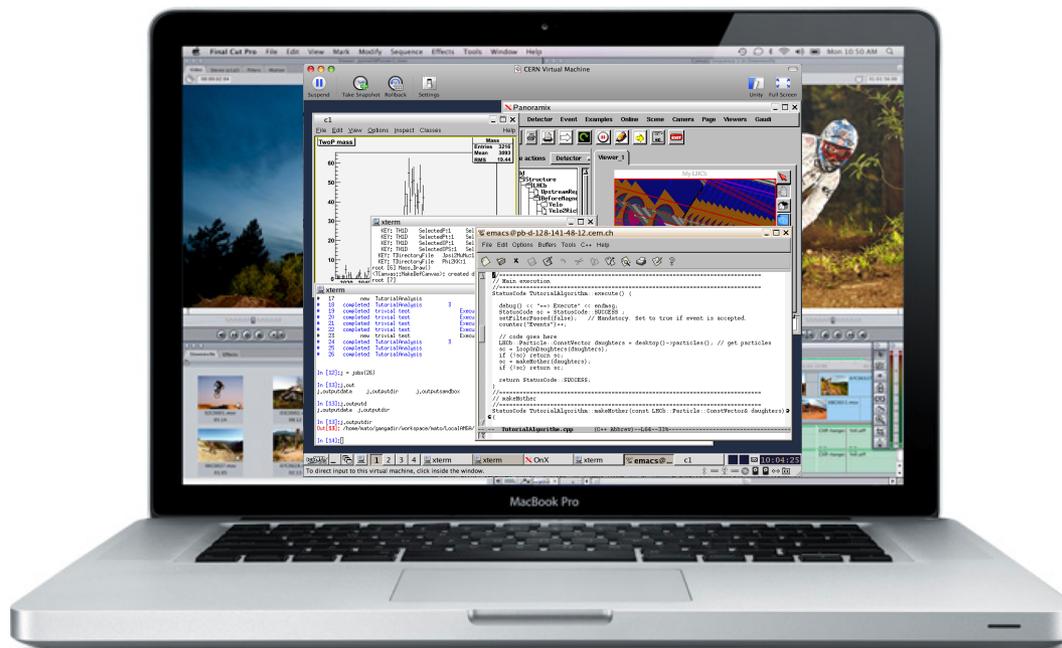
CernVM - a virtual software appliance for LHC applications

C. Aguado-Sanchez ¹⁾, P. Buncic ¹⁾, L. Franco ¹⁾, A. Harutyunyan²⁾,
P. Mato ¹⁾, Y. Yao ³⁾

- 1) CERN, Geneva,
- 2) Yerevan Physics Institute, Yerevan,
- 3) LBNL, Berkeley

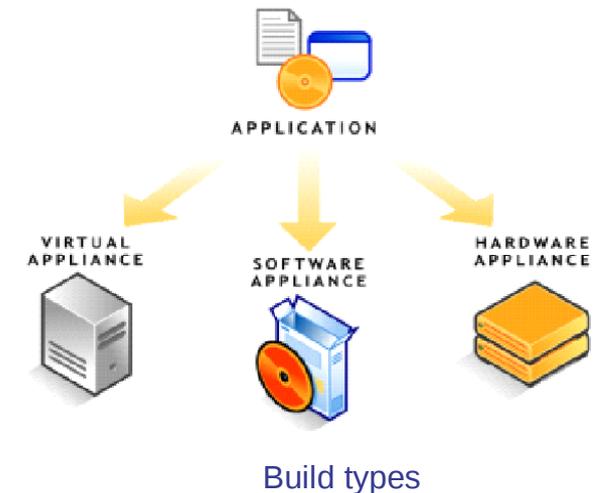
Predrag Buncic (CERN/PH-SFT)

- Portable Analysis Environment using Virtualization Technology (WP9)
 - Approved in 2007 (2+2 years) as R&D activity in CERN/PH Department
 - Started January 2008
 - Sister project to Multicore R&D
- Project goals:
 - Provide a complete, portable and easy to configure user environment for developing and running LHC data analysis locally and on the Grid independent of physical software and hardware platform (Linux, Windows, MacOS)
 - Decouple application lifecycle from evolution of system infrastructure
 - Reduce effort to install, maintain and keep up to date the experiment software
 - Lower the cost of software development by reducing the number of compiler-platform combinations



- A complete Data Analysis environment available for each experiment
 - Code check-out, edition, compilation, local small test, debugging, ...
 - Grid submission, data access...
 - Event displays, interactive data analysis, ...
- No software installation required
- Suspend/resume capability

- rBulder from rPath (www.rpath.org)
 - A tool to build VM images for various virtualization platforms
- rPath Linux 1
 - Slim Linux OS binary compatible with RH/SLC4
- rAA - rPath Linux Appliance Agent
 - Web user interface
 - XMLRPC API
 - Can be fully customized and extended by means of plugins (#401)
- CVMFS - CernVM file system
 - Read only file system optimized for software distribution
 - Aggressive caching
 - Operational in offline mode
 - For as long as you stay within the cache



- Installable CD/DVD
- Stub Image
- Raw Filesystem Image
- Netboot Image
- Compressed Tar File
- Demo CD/DVD (Live CD/DVD)
- Raw Hard Disk Image
- VMware® Virtual Appliance
- VMware® ESX Server Virtual Appliance
- Microsoft® VHD Virtual Appliance
- Xen Enterprise Virtual Appliance
- Virtual Iron Virtual Appliance
- Parallels Virtual Appliance
- Amazon Machine Image
- Update CD/DVD
- Appliance Installable ISO

Integration of Virtualized Worker Nodes in Standard-Batch-Systems

CHEP 2009 Prague

Oliver Oberst



Forschungszentrum Karlsruhe
in der Helmholtz-Gemeinschaft



Universität Karlsruhe (TH)
Forschungsuniversität • gegründet 1825



Outline



- General Description of Virtualization / Virtualization Solutions
- Shared HPC Infrastructure
- Virtualization in High Performance Computing
- Dynamic Partitioning of a shared Computing Cluster
 - Concept
 - KIT Implementation (KVM and Maui/Torque)
 - DESY Implementation (Xen, SGE and vmimagemanger)

Shared HPC Infrastructure



- **Compromise is not desirable/possible in some cases:**
 - **Incompatibilities** between **software and operating systems (OS)** within the needs of the different groups.
 - Some groups want to participate in a Grid environment (may lead to point above).
 - The **Grid environment should be isolated** from the local users (security...)

- Department or user groups run a shared computing infrastructure:

- **Pros:**

- Administration can be centralised (e.g. at the Computing Centre of a University)
- Shared funding may lead to a favourable hardware price
- Load-balancing

- **Cons:**

- Set-up has to be a compromise!

Virtualization

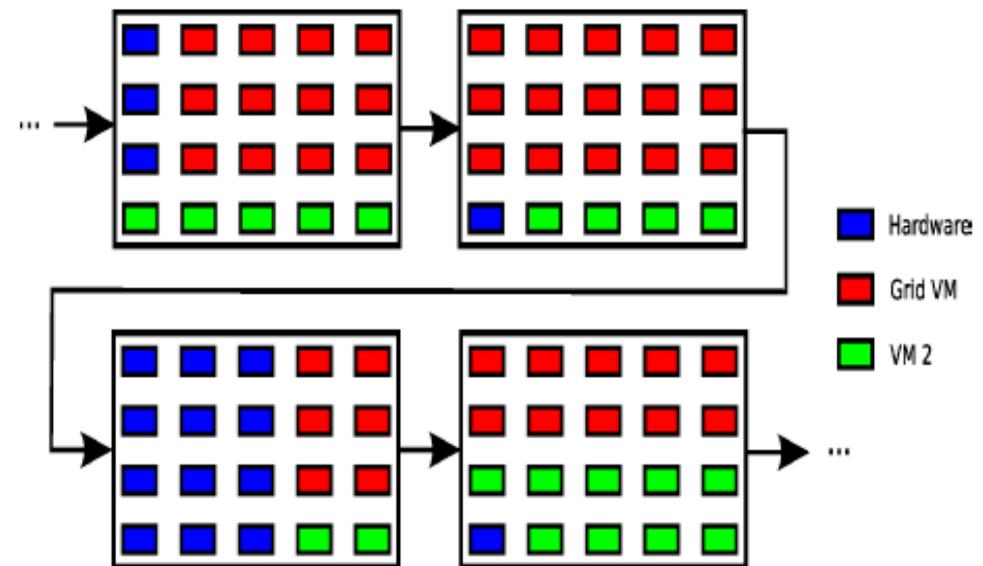
Dynamic Partitioned Cluster



- A cluster where multiple OS are needed can be partitioned dynamically with virtualization

Dynamic Partitioned Cluster:

- Virtual machines are deployed just as needed
- Scheduling with fair-share
- Virtualization techniques hidden from the user, sees only standard batch system and partition queues



vmimagemanager.py



■ Deployment

■ Scripted Configuration

- LVM
- image management

■ Low dependencies

■ Mostly automated (YAIM)

■ Functions

■ Fresh OS per job

- Easy to change to just reboot the OS

■ Easy to extend to user defined jobs.

- Hooks available

■ History

■ Deployed for debugging

- vmimagemanager.py
 - Years of use @ DESY
- Test Glite Batch Queue

■ Plans (in progress)

■ KVM, Qeum, Xen

- Through libvirt
- Kpartx: full disk handling
 - Mounting virtual hosts

■ SA3 certification

- WLCG
- Mass Deployment
- NIKEF interested in testing

<http://vmimagemanager.wiki.sourceforge.net/>



Conclusion

- There are use cases where a bare-metal use of worker nodes in a shared cluster is not possible
- Virtualization
 - Allows dynamic partitioning of a Cluster
 - customised software environments for all user groups
 - load-balancing
 - Performance overhead is acceptable
- Our approaches (DESY/KIT):
 - Do **NOT** need a modification of the used **batch system** to be “VM aware” (VM is seen as job)
 - Light-weight and transparent
 - Intelligent scripts and the standard-batch-system configuration do the job



HEPiX Benchmarking Group
Michele Michelotto at pd.infn.it



A comparison of HEP code with SPEC
benchmark on multicore worker nodes

- Since about 2004 several HEPiX users were presenting measurements on performances and benchmarking
- Anomalies in performances between application code and SI2K
- In 2006 a Working Group, chaired by Helge Meinhard (CERN) was setup inside HEPiX to address those issues
- We requested an help from the major HEP experiments

What is SPEC?

- SPEC
 - “www.spec.org : a non profit corporation that establish maintains and endorses a set of computer related benchmarks”
- SPEC CPU
 - “Designed to provide performance measurements that can be used to compare compute-intensive workloads on different computer systems“
- History
 - Before SPEC: CERN UNIT, MIPS, VUPS (Lep Era)
 - After SPEC: SPEC89, CPU92, CPU95, CPU2000, CPU2006

- Since SPEC CPU 92 the HEP world decide to use INT as reference instead of FP (Floating Point)
- HEP programs of course make use of FP instructions but with minimal impact on benchmarks
- I've never seen a clear proof of it

- SPEC CPU INT 2000 shortened as SI2K
- The “Unit of Measure”
 - For all the LHC Computing TDR
 - For the WLCG MoU
 - For the resources pledged by the Tier [0,1,2]
 - Therefore used in tender for computer procurements

- Results taken from www.spec.org for different processors showed good linearity with HEP applications up to ~ Y2005
- HEP applications use Linux + gcc
- SPEC.org makes measurements on Linux/Win + Intel or Pathscale compiler
- If you run SPEC on Linux+gcc you obtain a smaller value (less optimization)
- Is it proportional to SPEC.org or to HEP applications?

Too many SI2K?

- Too many definition of SI2K around
- E.g. take a common processor like an Intel Woodcrest dual core 5160 at 3.06 GHz
- SI2K spec.org: 2929 – 3089 (min – max)
- SI2K sum on 4 cores: 11716 - 12536
- SI2K gcc-cern: 5523
- SI2K gcc-gridka: 7034
- SI2K cern + 50%: 8284

- The use of the SI2K-LCG was a good INTERIM solution
- In 2006 SPEC published CPU 2006 and stopped the maintenance on CPU 2000
- Impossible to find SI2000 from SPEC for the new processor
- Impossible to find SI2006 for old processor
- Time to move to a benchmark of CPU 2006 family?

- What's new:

- Larger memory footprint: from ~200MB per core to about 1GB per core in 32bit environment
- Run longer (1 day vs 1 hour)
- CPU 2000 fitted too much in L2 caches
- INT: 12 CPU intensive applications written in C and **C+**
+
- FP: 17 CPU intensive applications written in C, **C++** and Fortran

- SPECint2006 (12 applications)
 - Well established, published values available
 - HEP applications are mostly integer calculations
 - Correlations with experiment applications shown to be fine
- SPECfp2006 (17 applications)
 - Well established, published values available
 - Correlations with experiment applications shown to be fine
- SPECcall_cpp2006 (7 applications)
 - Exactly as easy to run as is SPECint2006 or SPECfp2006
 - No published values (not necessarily a drawback)
 - Takes about 6 h (SPECint2006 or SPECfp2006 are about 24 h)
 - Best modeling of FP contribution to HEP applications
 - Important memory footprint
- Proposal to WLCG to adopt SPECcall_cpp 2006, in parallel and to call it **HEP SPEC06**



Addressing the Challenges of High Performance Computing with IBM Innovation and iDataPlex:

“Take Advantage of Cooler, Denser, and More Efficient Compute Power”

Gregg McKnight

Vice President

Distinguished Engineer

System x and BladeCenter Development

IBM Corporation

March 2009



Introducing System x iDataPlex

■ An Innovative x86 Solution from IBM to address:

- Total Cost of Ownership (TCO) from Acquisition to OPEX
- Data center density, scalability, serviceability, manageability
- Individual customer requirements



■ iDataPlex is:

- A half-depth server design
- Optimized for maximum energy and
- An Industry-standards based server platform
- Designed to minimize utilization of floor space, energy and cooling
- Easily maintainable front access solution
- Custom preconfigurable for compute, storage, or I/O needs and



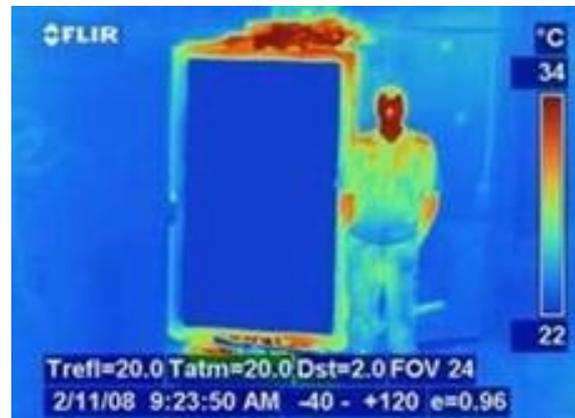
Cool Blue for Cool Savings

IBM Rear Door Heat eXchanger for iDataPlex

- 75%-95% greater efficiency than air cooling
- Completely eliminates rack heat exhaust
- No moving components or auxiliary fans
- No condensation
- Moves thermal transfer from CRAC to back of rack
- Can eliminate supplemental AC and raised floors



54° C – Cool Blue Off



16° C – Cool Blue On



DATA CENTER REVOLUTION

A magnifying glass with a gold handle and a black frame is positioned over the word "REVOLUTION" in the title, highlighting it.

Presented at Computing in High Energy & Nuclear Physics

CHEP - Prague, Czech Republic

March 25, 2009

Dean Nelson

Sr. Director, Global Lab & Datacenter Design Services (GDS)

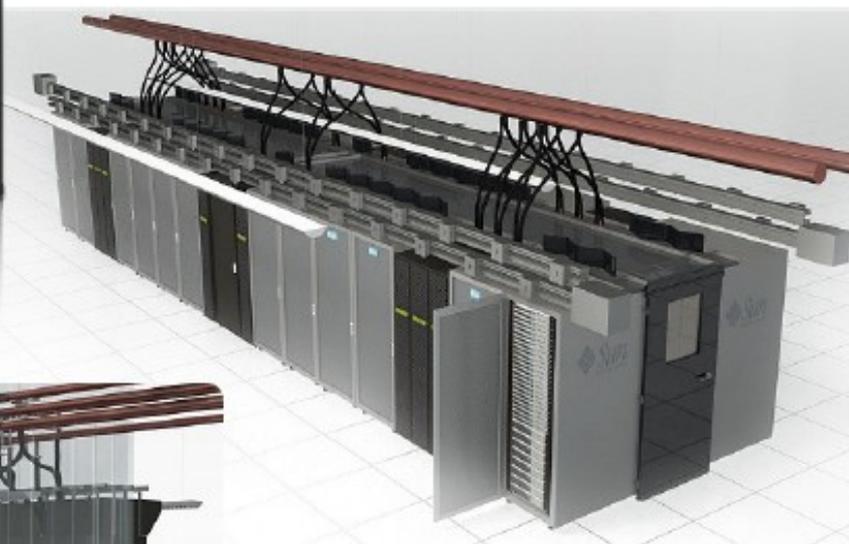
A moment of silence...



- Raised Floors Are Dead
 - > No longer required
 - > Go against physics
 - > Increasingly cumbersome
 - > Expensive
- Next Generation equipment requires a new way of thinking...

Pod Architecture

Modular Data Center Building Blocks
Container and/or Brick & Mortar



Sun Pod Architecture



Floating Data Centers



- Tier1-Tier3 ECO datacenters at US and international ports
- Capacity: 4000 racks and over 350 SunMDs
- 75MW of power, free cooling from ocean water
- Six months time to market, up to 40% less than traditional build



- At the end of a dock instead of the end of a street

Air Conditioning and Computer Centre Power Efficiency The Reality

Christophe Martel

Tony Cass

Basic Housekeeping is essential!



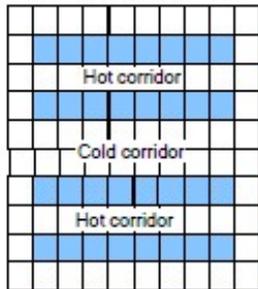
Basic Housekeeping is essential!

Mind the gaps!



High density starts at 6KW/rack

Typical room

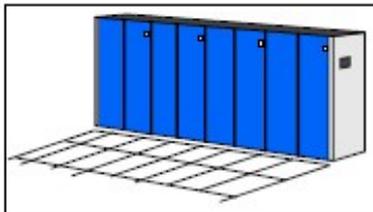


30% of the area for racks

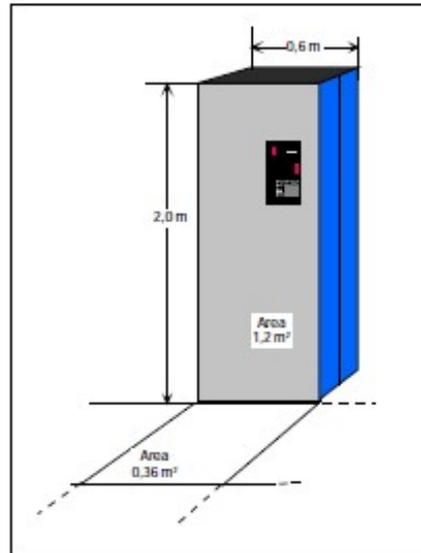
Basic data

Standard air velocity: max 0.25 m/s
Temperature in/out IT : 8°C

Area equivalence:
1 rack front side = 3.33 floor tiles



Typical rack



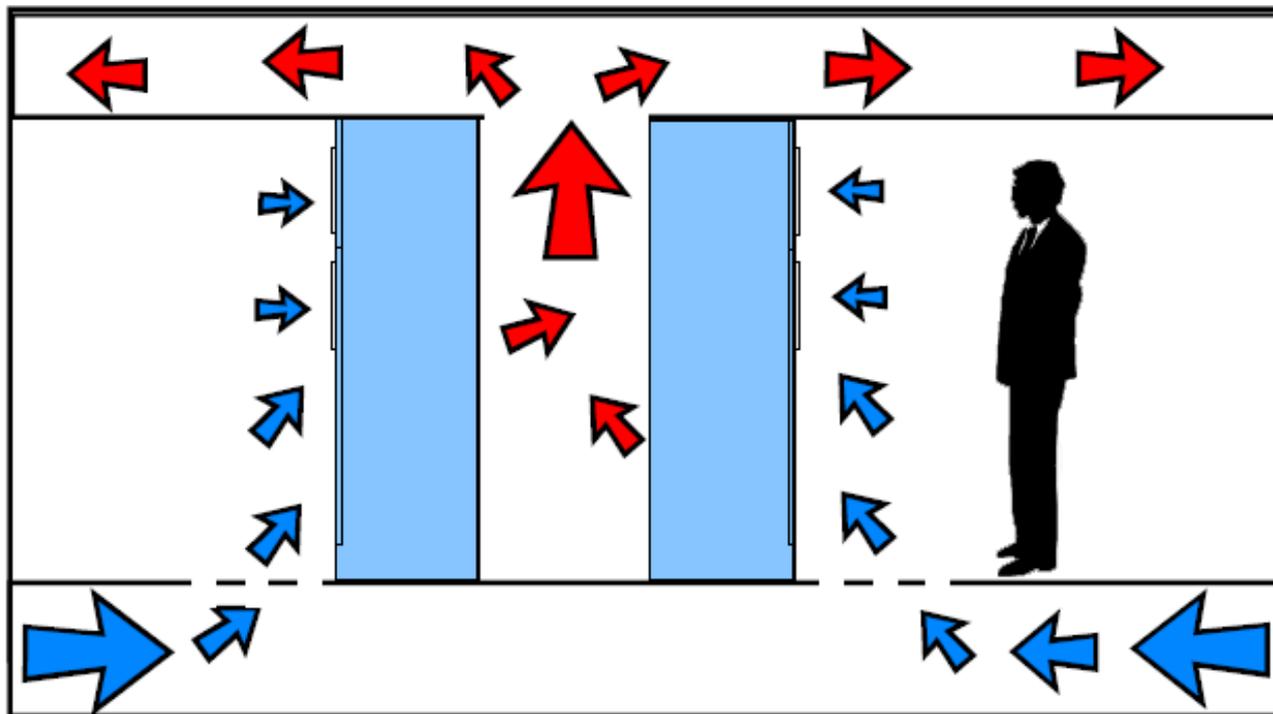
	Air speed [m/s]	Air flow rate [m3/h]	Cooling power [W]
Rack front side (1.2m ²)	0.25	1080	2938
	0.5	2160	5875
Floor tile (0.36m ²)	0.25	324	881
	0.5	648	1763



0 < low density < 2kW/m² < high density
Central air cooling ≤ 2kW/m² ≤ Local air cooling

Reducing air speed

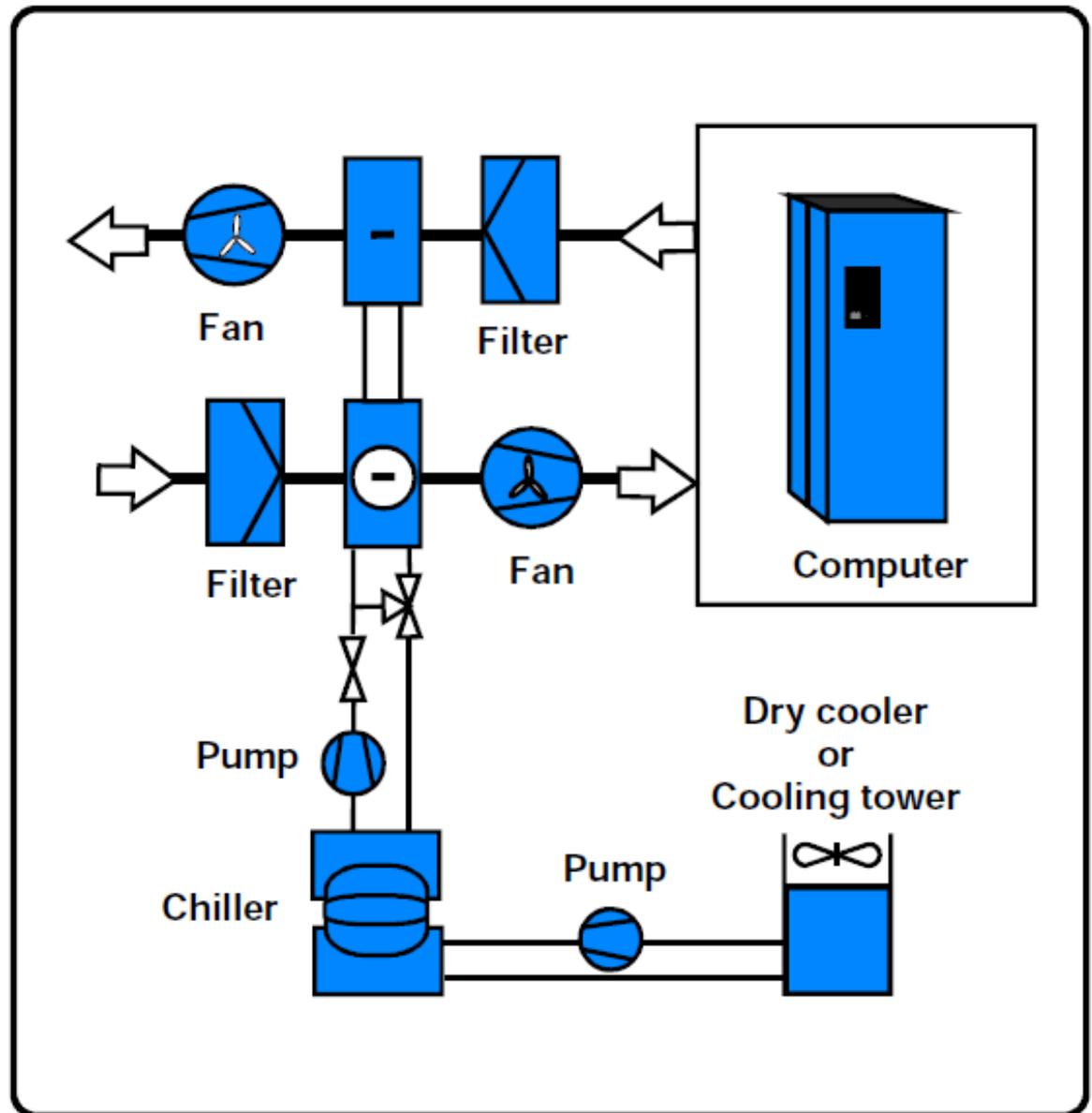
Cold aisle closed



False floor. False ceiling for return air.
Cold aisle closed. Hot aisle open.

Conclusion

- Optimise...
- ... everywhere!



Lustre File System Evaluation at FNAL

Stephen Wolbers

for

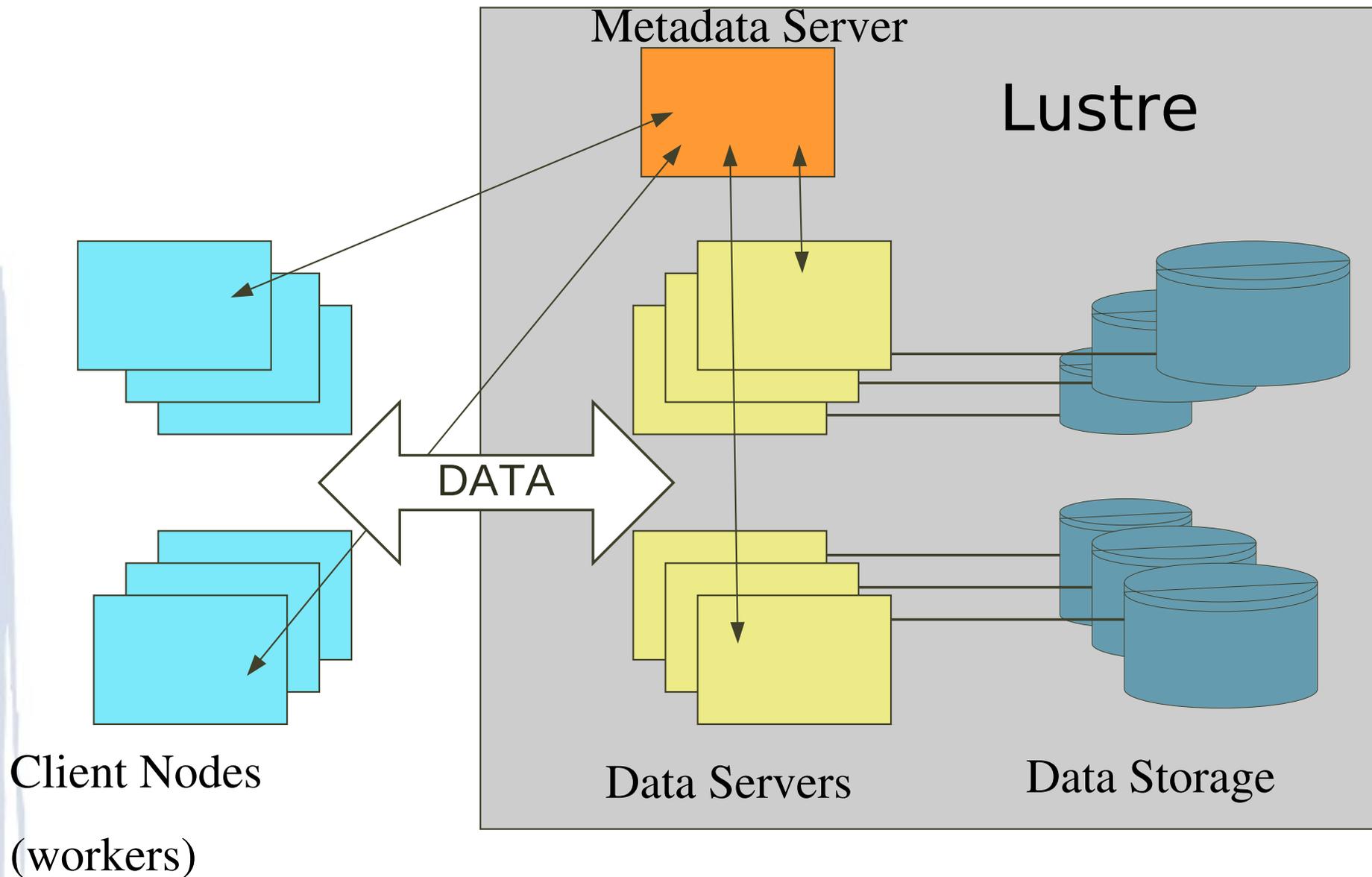
Alex Kulyavtsev, Matt Crawford, Stu Fuess, Don Holmgren,
Dmitry Litvintsev, Alexander Moibenko, Stan Naymola,
Gene Oleynik, Timur Perelmutov, Don Petravick, Vladimir Podstavkov,
Ron Rechenmacher, Nirmal Seenu, Jim Simone

Fermilab

CHEP'09, Prague

March 23, 2009

What is Lustre?



Lustre Experience - HPC

- From our experience in production on Computational Cosmology Cluster (starting summer 2008) and limited pre-production on LQCD JPsi cluster (December 2008) the Lustre File system:
 - Lustre doesn't suffer the MPI deadlocks of dCache
 - direct access eliminates the staging of files to/from worker nodes that was needed with dCache (Posix IO)
 - improved IO rates compared to NFS and eliminated periodic NFS server "freezes"
 - reduced administration effort

Lustre HSM Feature

- Lustre does not yet have HSM feature. Some sites implement simple tape backup schemes
- HSM integration feature is under development by CEA and Sun

HSM version v1.0

- “Basic HSM” in a future release of Lustre — beta in fall 2009 ?
- Integration with HPSS (v1), others will follow
- Metadata scans to select files to store in HSM v1
 - File store on close() on-write in HSM v2

Integration work

- Work specific to the HSM is required for integration

Conclusions - HEP

- Lustre file system meets and exceeds our storage evaluation criteria in most areas, such as system capacity, scalability, IO performance, functionality, stability and high availability, accessibility, maintenance, and WAN access.
- Lustre has *much* faster metadata performance than our current storage system.
- At present Lustre can only be used for HEP applications not requiring large scale tape IO, such as LHC T2/T3 centers or scratch or volatile disk space at T1 centers.
- Lustre near term roadmap (about one year) for HSM in principle satisfies our HSM criteria. Some work will still be needed to integrate any existing tape system.

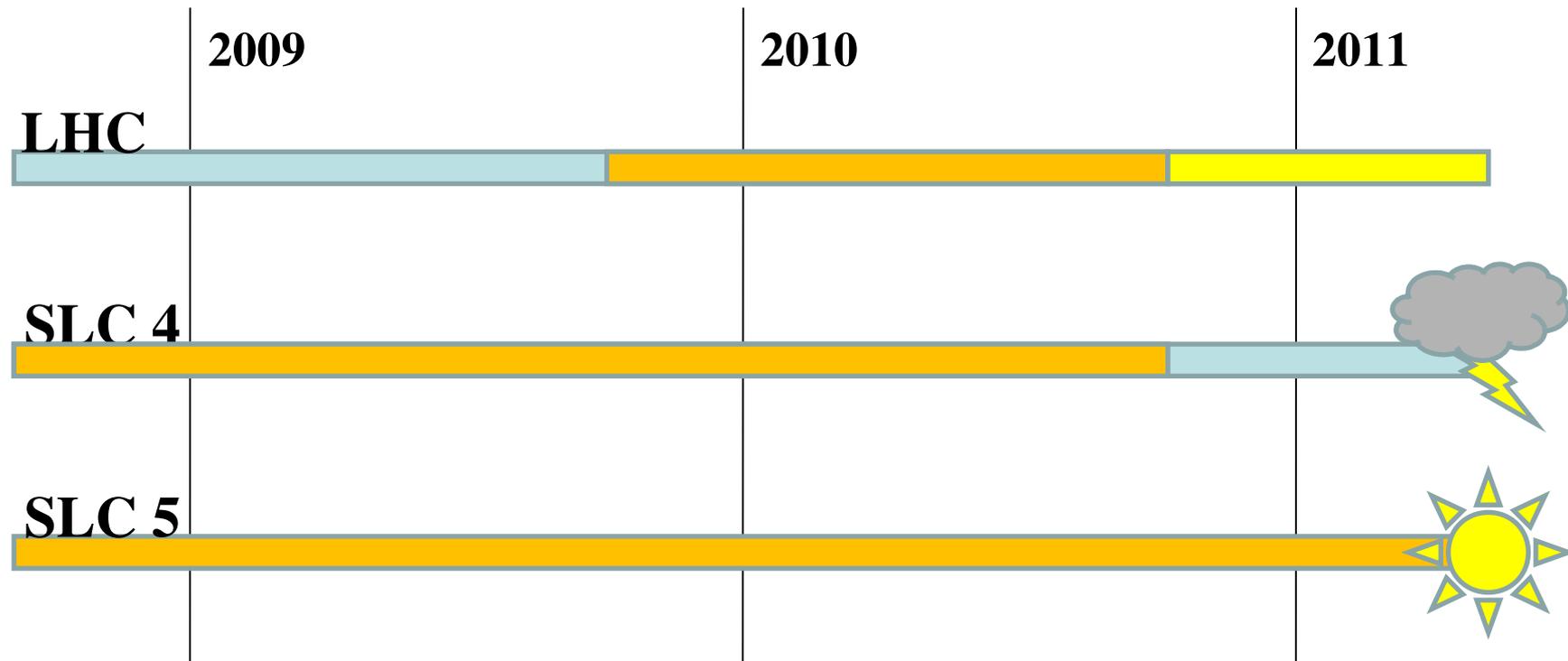
SL(C) 5 Migration at CERN

CHEP 2009, Prague

Ulrich SCHWICKERATH

Ricardo SILVA

CERN, IT-FIO-FS

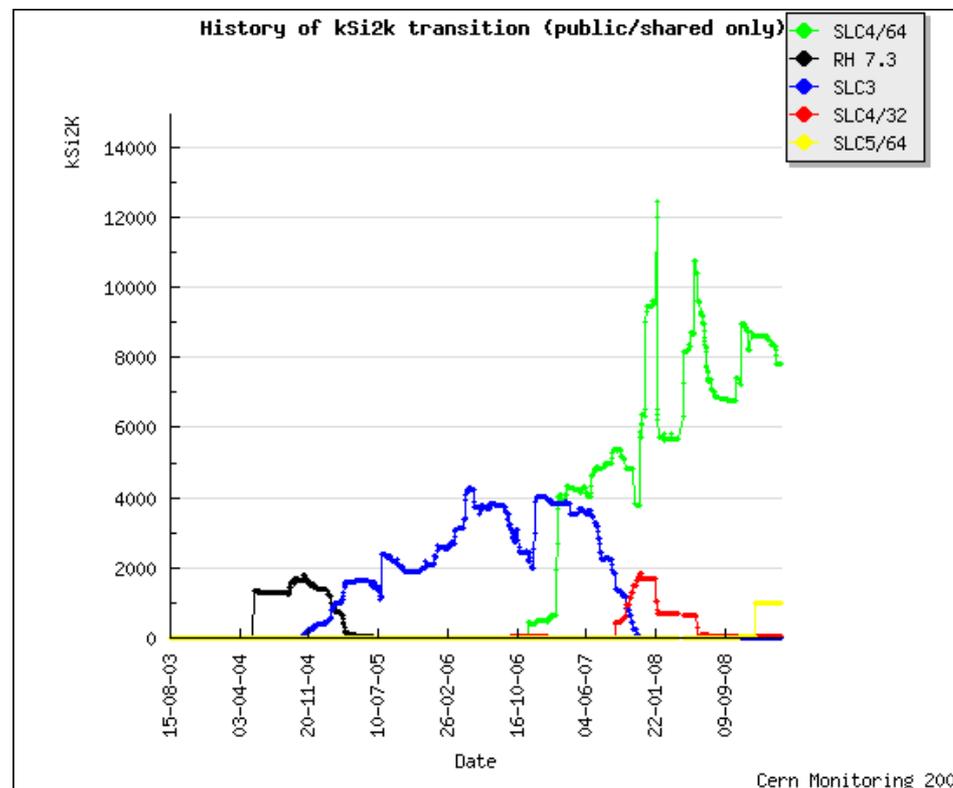


- We want stability during the LHC run period!

- Newer versions of software which include new features and bug fixes are not available for SL(C) 4
 - Bug fixes and increased performance in the XFS code
- Security
 - Some security fixes need to be back-ported to our versions of the software
- Virtualization support
 - No support for virtualization tools in SL(C)4
 - We want to increase the use of VMs for consolidation of resources

- Change of Ixplus (interactive cluster) alias
 - Until this switch is done code is compiled on SLC4; after that point software will by default be compiled on SLC5
 - Should happen once the **majority of the resources** are migrated on WLCG
 - All the VOs are confident they can move to SLC5 by the summer

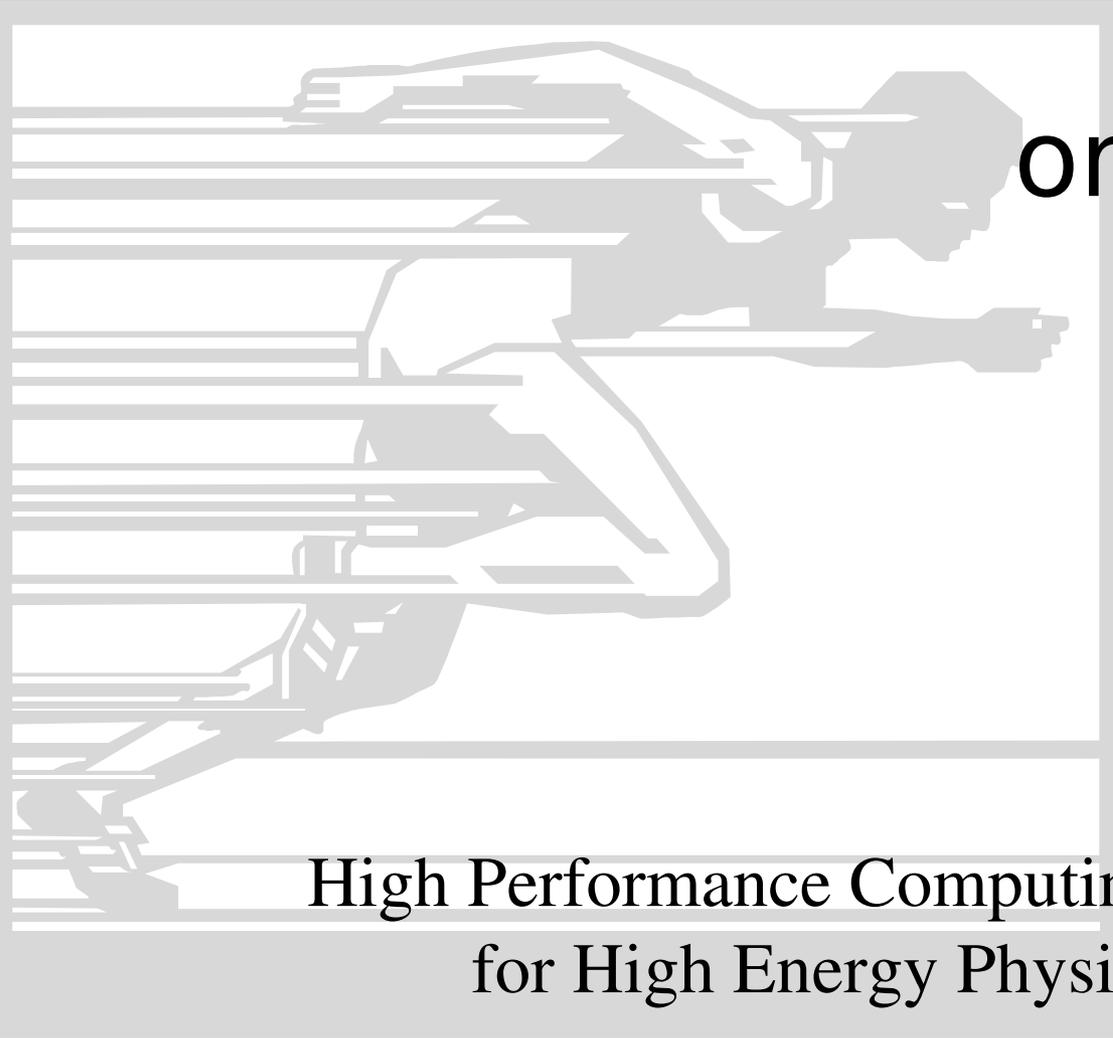
- All pledged CPU capacity for the LHC experiments for 2009 will be on SLC5 by the summer
 - SLC4 resources will be kept for other user communities at CERN
 - No “big bang” approach



The challenge of adapting HEP physics software applications to run on many-core cpus

CHEP, March '09

Vincenzo Innocente
CERN



High Performance Computing
for High Energy Physics

The 'three walls'

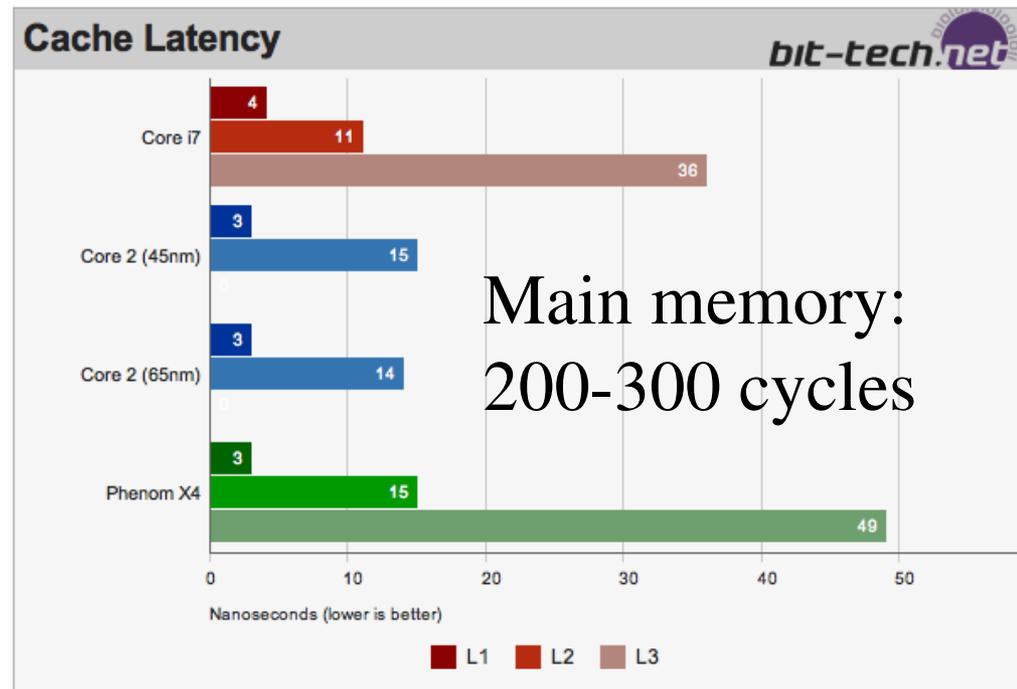
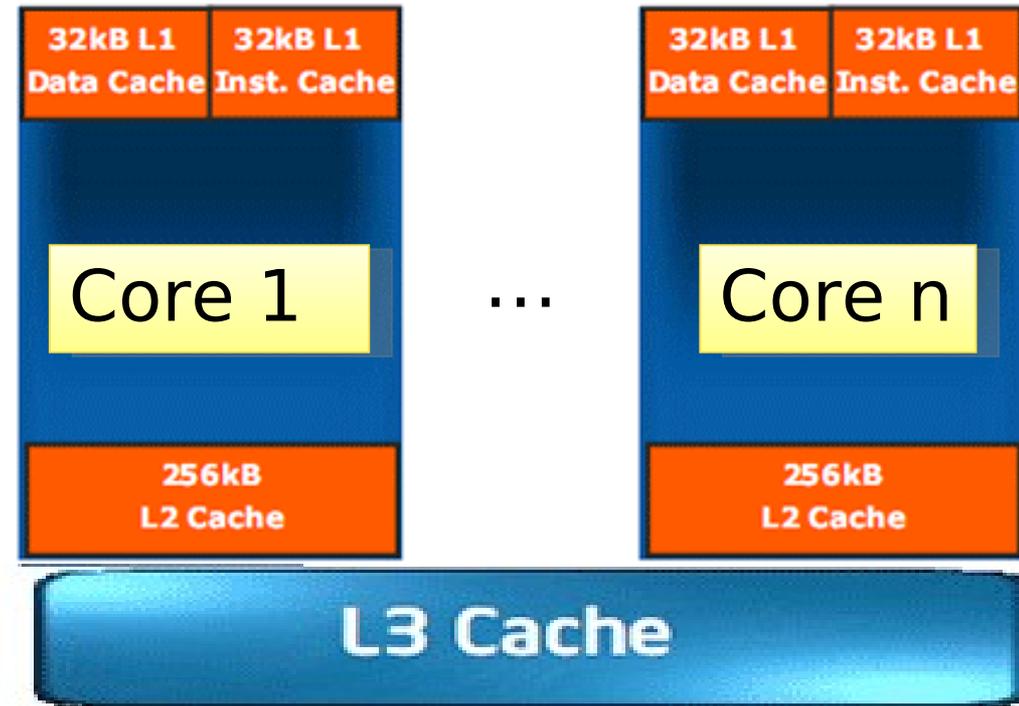
While hardware continued to follow Moore's law, the perceived exponential growth of the "effective" computing power faded away in hitting three "walls":

- The memory wall
- The power wall
- The instruction level parallelism (micro-architecture) wall

A turning point was reached and a new paradigm emerged: **multicore**

The 'memory wall'

- Processor clock rates have been increasing faster than memory clock rates
- larger and faster “on chip” cache memories help alleviate the problem but does not solve it.
- Latency in memory access is often the major performance issue in modern software applications



The 'power wall'

- Processors consume more and more power the faster they go
- Not linear:
 - » 73% increase in power gives just 13% improvement in performance
 - » (downclocking a processor by about 13% gives roughly half the power consumption)
- Many computing center are today limited by the total electrical power installed and the corresponding cooling/extraction power.
- How else increase the number of instruction per unit-time: **Go parallel!**

Where are WE?

See talks by P.Elmer, G.Eulisse, S. Binet

- HEP code does not exploit the power of current processors
 - » One instruction per cycle at best
 - » Little or no use of vector units (SIMD)
 - » Poor code locality
 - » Abuse of the heap
- Running N jobs on N=8 cores still efficient but:
 - » Memory (and to less extent cpu cycles) wasted in non sharing
 - “static” condition and geometry data
 - I/O buffers
 - Network and disk resources
 - » Caches (memory on CPU chip) wasted and trashed
 - Not locality of code and data
- This situation is already bad today, will become only worse in future architectures

HEP software on multicore: a R&D effort

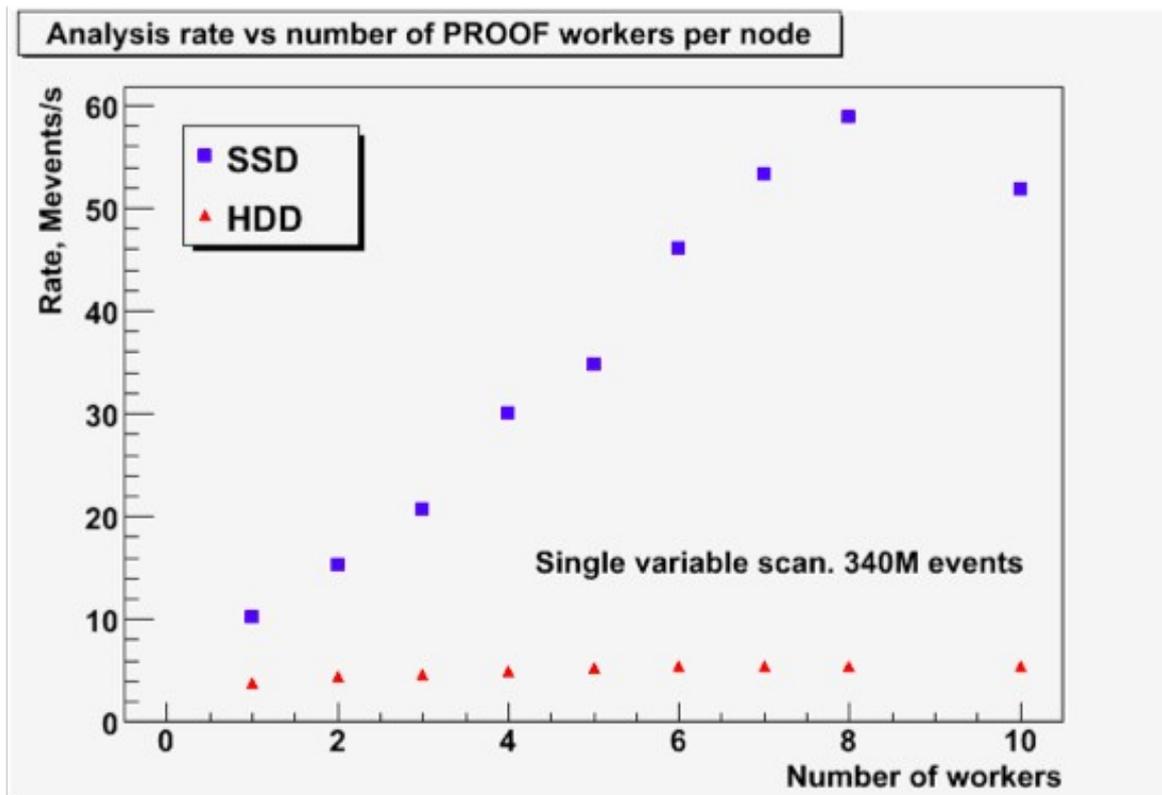
- Collaboration among experiments, IT-departments, projects such as OpenLab, Geant4, ROOT, Grid
- Target multi-core (8-24/box) in the short term, many-core (96+/box) in near future
- Optimize use of CPU/Memory architecture
- Exploit modern OS and compiler features
 - » Copy-on-Write
 - » MPI, OpenMP
 - » SSE/AltiVec, OpenCL
- Prototype solutions
 - » *Adapt legacy software*
 - » *Look for innovative solution for the future*

Exploit “Kernel Shared Memory”

- KSM is a linux driver that allows dynamically sharing identical memory pages between one or more processes.
 - » It has been developed as a backend of KVM to help memory sharing between virtual machines running on the same host.
 - » KSM scans just memory that was registered with it. Essentially this means that each memory allocation, sensible to be shared, need to be followed by a call to a registry function.
- Test performed “retrofitting” TCMalloc with KSM
 - » Just one single line of code added!
- CMS reconstruction of real data (Cosmics with full detector)
 - » No code change
 - » 400MB private data; 250MB shared data; 130MB shared code
- ATLAS
 - » No code change
 - » In a Reconstruction job of 1.6GB VM, up to 1GB can be shared with KSM

SSD vs HDD on 8 Node Cluster

See Sergey Panitkin's talk



Solid State Disk:
120GB for 400Euro

- Aggregate (8 node farm) analysis rate as a function of number of workers per node
- Almost linear scaling with number of nodes

Algorithm Parallelization

- Ultimate performance gain will come from parallelizing **algorithms** used in current LHC physics application software
 - » Prototypes using posix-thread, OpenMP and parallel gcclib
 - » Effort to provide basic thread-safe/multi-thread library components
 - Random number generators
 - Parallel minimization/fitting algorithms
 - Parallel/Vector linear algebra
- Positive and interesting experience with Minuit
 - » Parallelization of parameter-fitting opens the opportunity to enlarge the region of multidimensional space used in physics analysis to essentially the whole data sample.

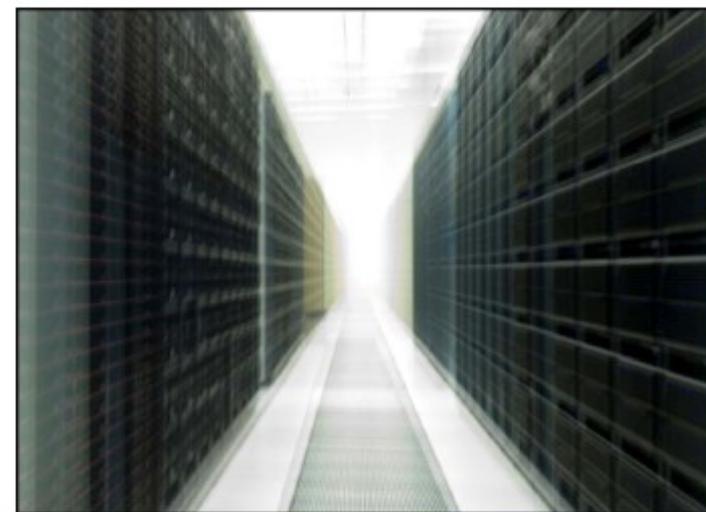
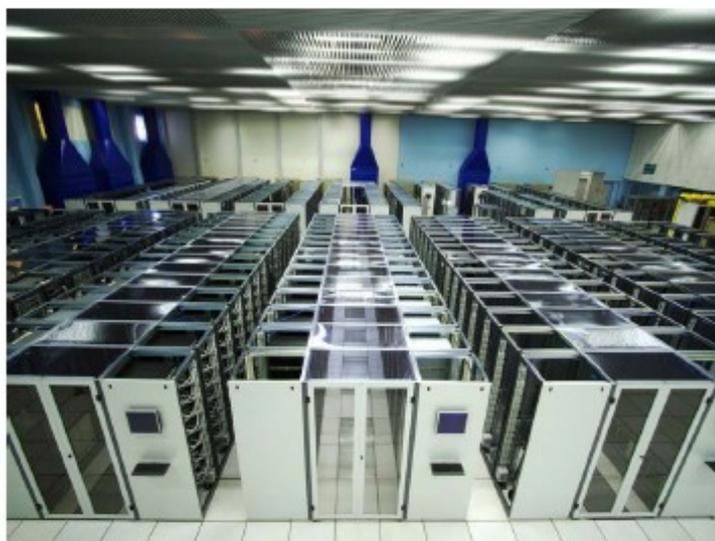


Is the Atom (N330) processor ready for High Energy Physics?

Gyorgy Balazs
Sverre Jarp
Andrzej Nowak

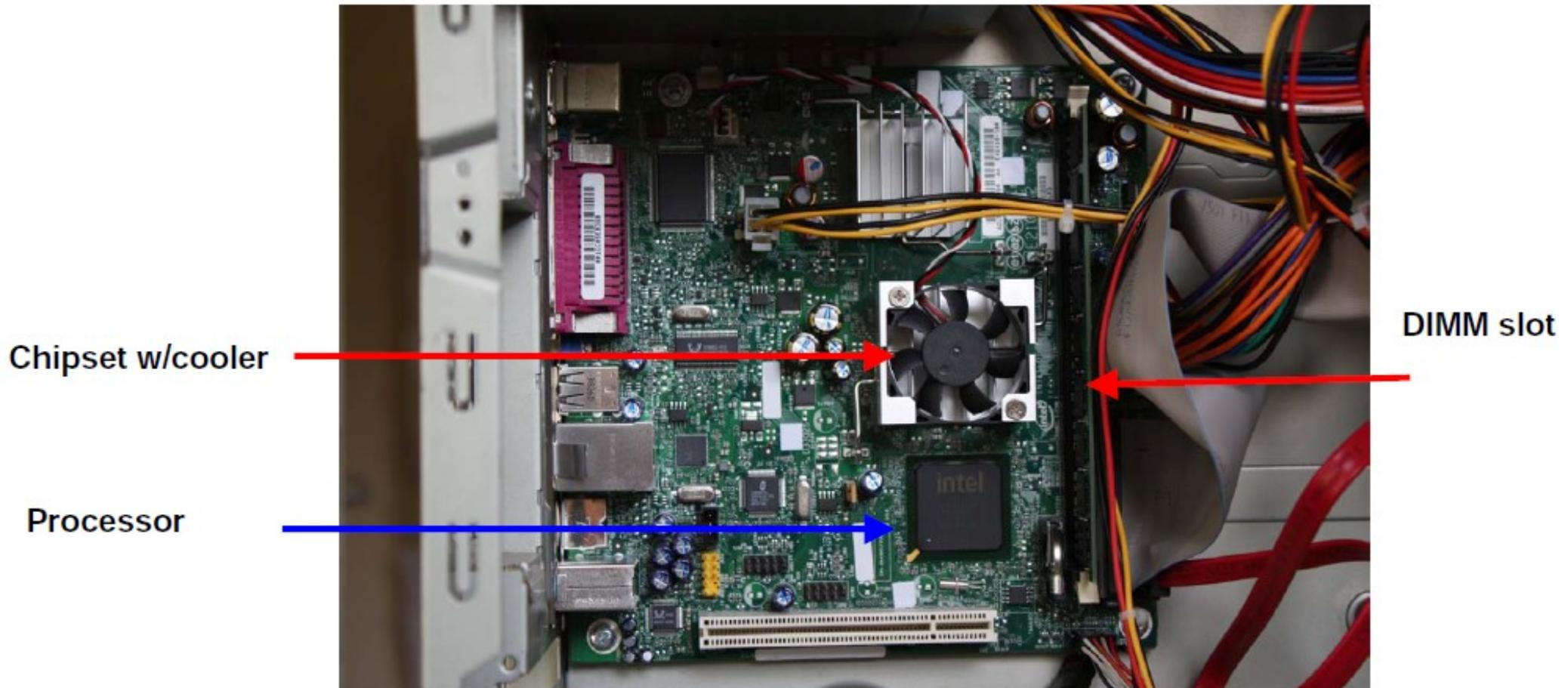
CERN openlab

CHEP09 – 23.3.2009



ATOM motherboard

- Note that the fan is not on the processor but on the D945GC chipset



Benchmark results (cont'd)

- **"test40" from Geant4 (in summary):**
 - Atom baseline: 1 process at 100% throughput at 53W
 - Atom peak: 4 processes at 302% throughput at 56W
 - Harpertown: 8 processes at 3891% throughput at 290W
- **In other words (Harpertown/Atom ratios):**
 - Cost ratio was: 16.5 (with adjusted memory)
 - 12.9x throughput advantage
 - 5.2x power increase
- **Atom N330 could be interesting in terms of performance/franc**
 - Currently uninteresting when looking at performance/watt

Main issues with Atom system

- **Memory:**
 - Need support for large memories
 - Or: HEP software that needs less memory per process

- **Power consumption:**
 - **Need a chipset with reduced consumption**
 - And: More efficient power supply



The (B)right future of software installation

S. Bagnasco, L. Betev, F. Carminati, F. Furano,
C. Grigoras, A. Grigoras, P. Mendez Lorenzo,
A. Peters, P. Saiz

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





- Current scenario

- One installation (per platform) per site

- On a dedicated area

- High bandwidth
 - Security risk

Do we need it?

- AliEn installation:

- New sites: running a script
 - Updating a site: triggered by a VO admin

- Experiment software:

- Install on demand according to the jobs
 - Before submitting JobAgents

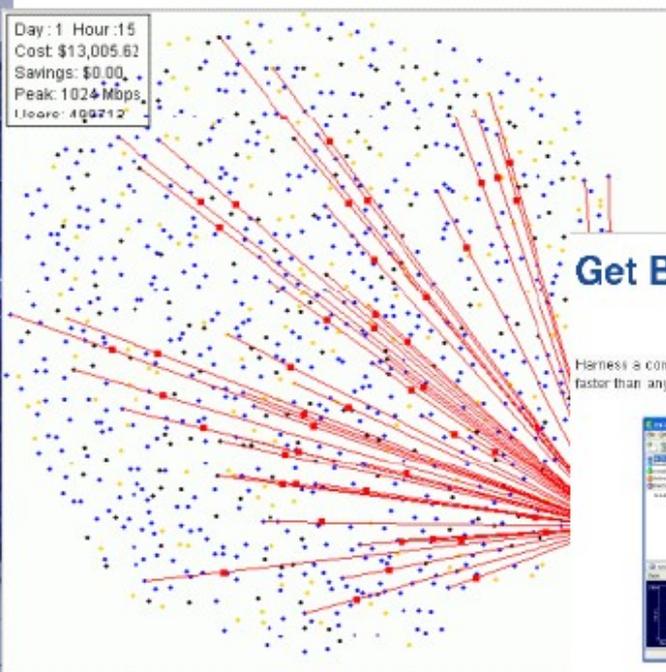


Can we do something better?



- Automatic installation on every worker node
- Automatic
- Self-contained
- User space
- Small software (300 MB)
- Job Agents (can run more than one small job)





More than 150 million users!!

Get BitTorrent

Harness a community of over 150 million users to deliver files to your PC faster than anything else.



The new BitTorrent 6 for Windows brings together BitTorrent's proven expertise in networking protocols with µTorrent's efficient implementation and compelling UI to create a better BitTorrent client. For questions about the BitTorrent client, take a look at the FAQ in our [support center](#), or visit the [client forums](#).

Advertisement

Get BitTorrent

Download other versions:

- [BitTorrent for Windows](#)
- [BitTorrent for MacOS X](#)
- [Linux, Source Code, and Older Versions](#)

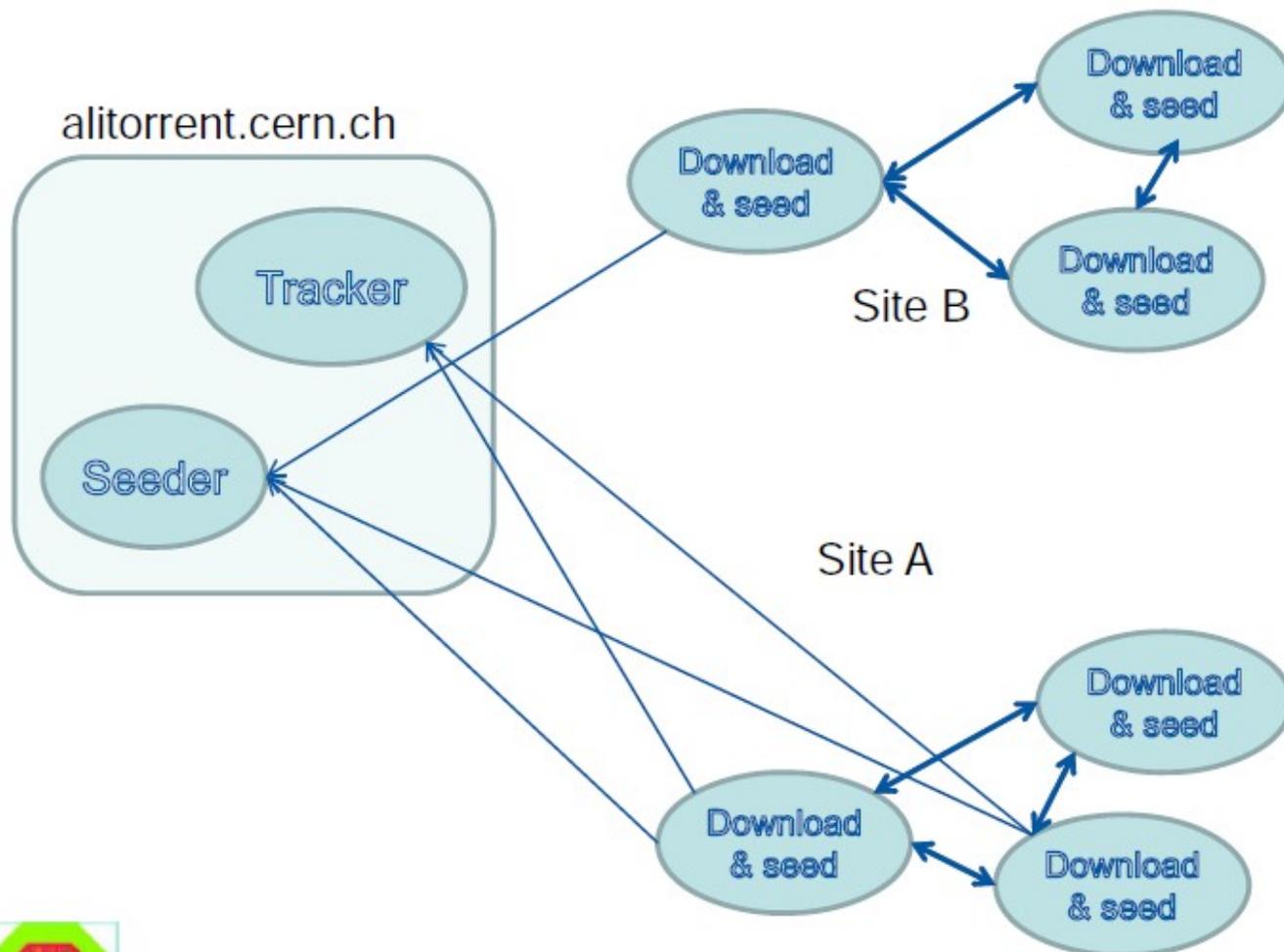
If you encounter any problems, please report them in the [Client Forums](#).

BitTorrent 6 Features

- Lightweight client
- Local peer discovery
- Configurable bandwidth scheduler
- Global and per-torrent speed limiting
- RSS Downloader
- Always Spyware-Free

<http://bittorrent.com>







- Torrent files created from the build system
- One seeder at CERN
 - Standard tracker and seeder.
- Get torrent client from ALICE web server
 - Aria2c
- Download the files and install them
- Seed the files while the job runs





- Working in collaboration with CERN Security team
- Peer-peer is allowed for professional usage
- Torrent files have checksums to detect corrupted files/wrong files
- Only VO admin can register files in tracker
- Signing the packages





CHEP 2009

An extremely biased personal view

Dario Barberis

CERN & Genoa University/INFN

Grids vs Clouds?

- We had this week the first reports of serious usage of cloud computing
 - Both in plenary and parallel sessions (and posters)
- It is an interesting concept but I am not sure it makes financial sense for data-intensive applications
 - CPU power may be cheap but data storage and transfer is very expensive
 - It may be an option if/when we are short of simulation power
- On the other hand, commercial companies are evidently able to provide the software environment the user needs, without many of the existing restrictions of the Grid



Grids AND Clouds!

- Virtualization is the keyword here!
- Can't we just do the same for our Grid sites?
 - Install "cloud middleware"...
 - And keep control of the resources and the data management tools
- All we need after all is a simple and reliable way to send jobs to where they can run fastest and most reliably
 - Virtual machines running user jobs can shield from local setup details
 - An existing implementation of Grid middleware (ARC) is already very close to these needs
- I know some of this may sound like heresy to some in this room
- On the first day we had a long list of people who got into trouble after lecturing in Prague!



Hardware trends

- One stumbling point until now: full use of 64-bit architectures saves some 10% of CPU power but almost doubles the memory footprint
 - What is the trade-off in terms of €/\$/£/¥/CHF?
 - Or is the solution in the use of a multi-threaded application?
 - How would that fit on the Grid?
 - Or in a virtual machine that runs on the Grid?
 - Good news at this CHEP (from CMS): VMEM increase can be reduced with custom linker script
- Multi-core 64-bit processors are nevertheless yesterday's technology.
 - We must think seriously of tomorrow's technology: many-core processors
- One possibility is to design
 - Parallel software that can occupy any number of cores in a machine
 - Virtual machines that can run in many-core processors
 - Grid middleware that allows the submission of this new kind of jobs
- These components must work simply and reliably together



**CHEP 2010
in Taipei, Taiwan
17-22 October, 2010**

Hosted by:

Academia Sinica Grid Computing (ASGC)