# Aktuální trendy ve zpracování dat

(Zpráva o konferencích CHEP 2013 a HEPIX)

*Jiří Chudoba*

13.11.2013

Seminář Sekce Elementárních částic

Fyzikální ústav AV ČR, Praha

# Osnova

- CHEP 2013
- Hepix Fall 2013

organized by NIKHEF in collaboration with partners

# The 20th CHEP Conference

## Amsterdam 14-18 Oct 2013

CHEP ´13 AMSTERDAM

KPMG

SURF SARA

JUNIPER NETWORKS

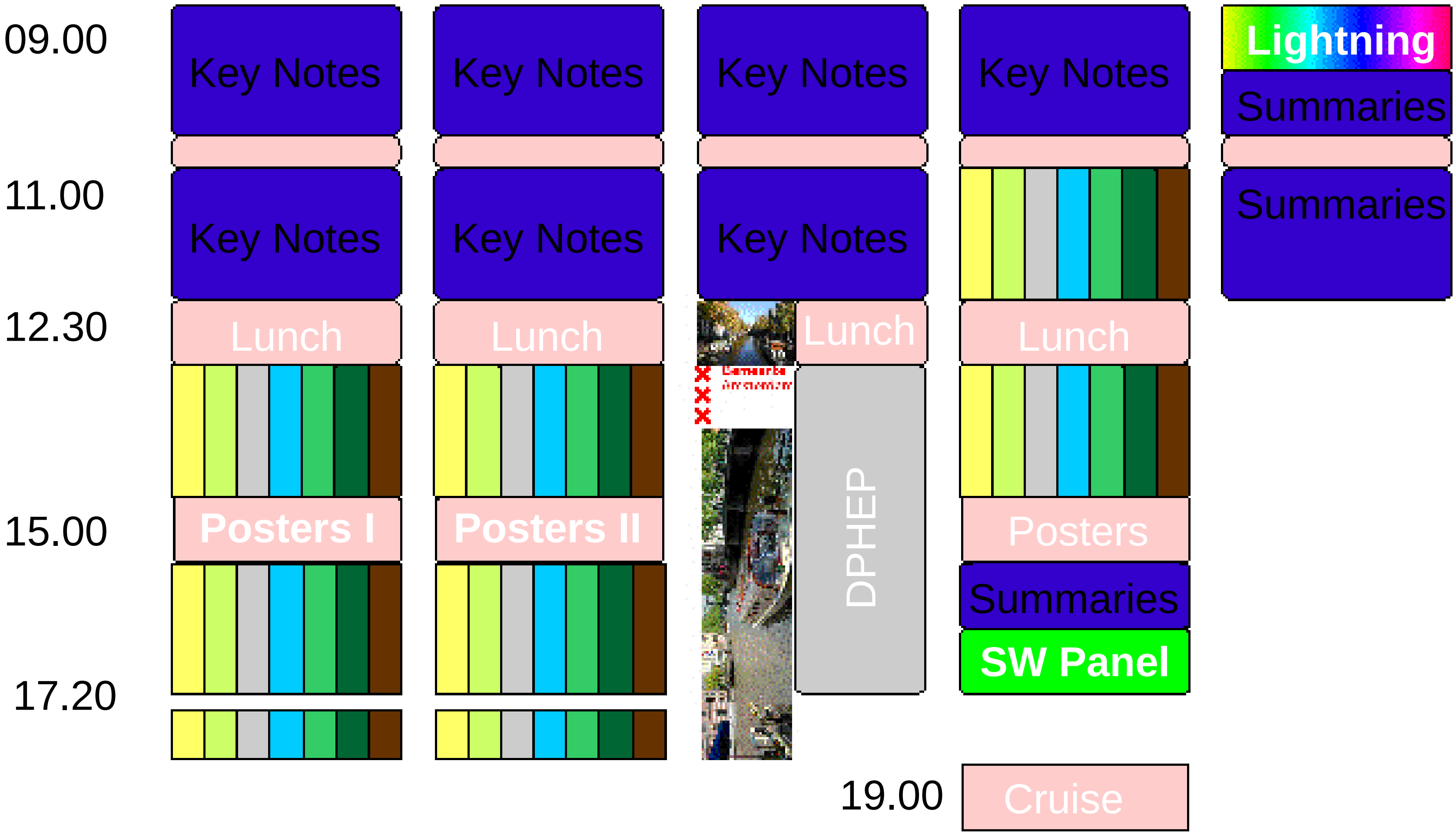
NIKHEF

EQUINIX

FOM

CHEP '85

# Programme at a Glance




Rozpočtová krize USA, úterní plenární přednášky z Fermilab

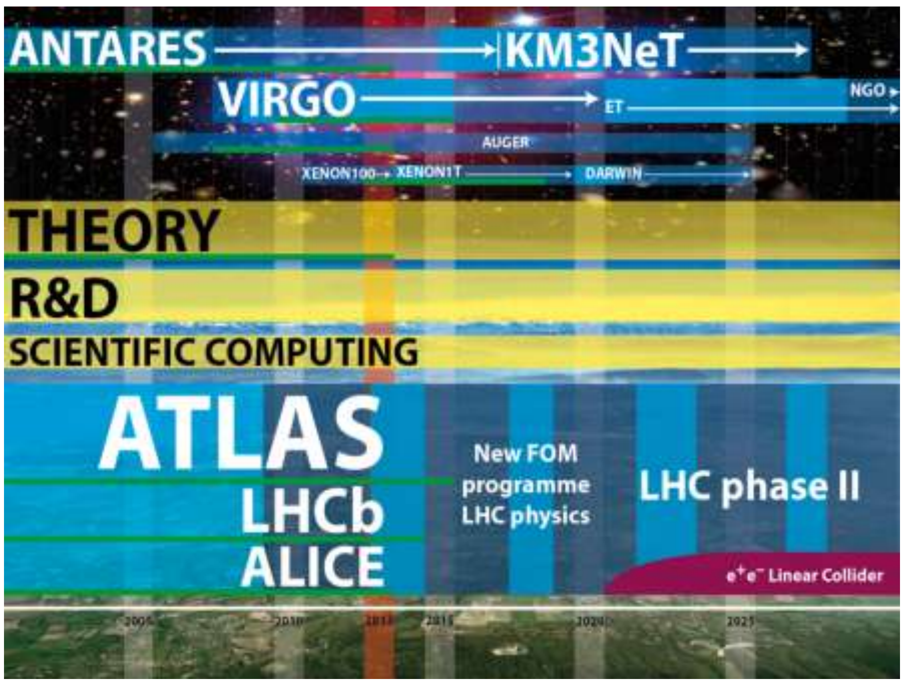
# Paralelní sekce

1. Data acquisition, trigger and controls
2. Event Processing, Simulation and Analysis
3. Distributed Processing and Data Handling
   - A: Infrastructure, Sites, and Virtualization
   - B: Experiment Data Processing, Data Handling and Computing Models
4. Data Stores, Data Bases, and Storage Systems
5. Software Engineering, Parallelism & Multi-Core
6. Facilities, Production Infrastructures, Networking and Collaborative Tools

# NIKHEF

**Frank Linde**, ředitel, člen ATLAS

~300 people
~30 M€/year





BIG GRID - rozpočet 30 ME, ukončeno minulý rok, žádají o další projekt

# **Robert Lupton** (Princeton)
# Writing Stellar Software: Preparing for the LSST

SDSS the Sloan Digital Sky Survey

Three mirrors: an 8.4m primary, a 3.4m secondary, and a 5m tertiary.

3.2 GPixels every 17s; c. 400 MB/s
20 TB per night; 60 PB over 10 years for the raw data and 15 PB for the catalog database.

# Roberts' Paradox

Unfortunately I'm naming it not for me, but for Eric Roberts at Stanford who in 2000 wrote a report for the US National Academy with the blessing of the ACM. The paradox is that:

- There are unemployed software engineers
- There is a shortage of software engineers

The resolution is that the shortage is of the best engineers, not the median:

> *If the best software developer can do the work of 10, 20, or even 100 run-of-the-mill employees, a single-person company that attracts such a superstar can compete effectively against a much larger enterprise*
> *[...]*
> *In some cases, software developers who fall at the low end of the productivity curve may be essentially nonproductive or even counterproductive*

# Lesson 10: Find some way to reward people working on the project

In SDSS we did this by promising them early access to the data via a proprietary period. Not only is this impossible for publicly funded projects, but it doesn't really work very well. One problem is that the promise of data in the distant future doesn't help a post-doc much; another is that the community (at least in the US) doesn't value work on the technical aspects of a large project. I don't think that the solution `Hire Professional Programmers' is viable (although hiring a significant number of *competent* software professionals is a good idea. My experience has been that we cannot afford to hire good programmers).

<hobbyhorse> My personal belief is that the only long term way out of this is to integrate instrumentation (hardware and software) into the astronomy career path, much the way that the high-energy physicists appear to have done (at least from the outside). </hobbyhorse>

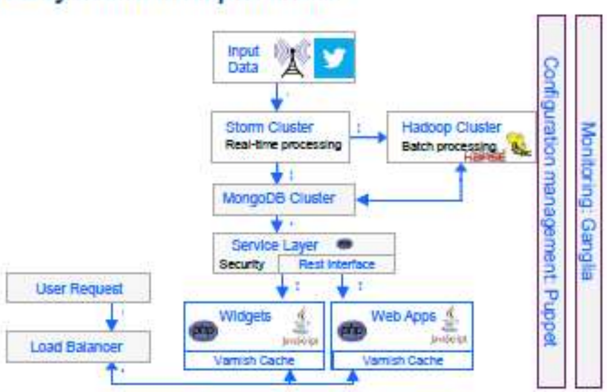
CHEP '13 AMSTERDAM

# Sander Klous: KPMG

## KPMG Data & Analytics

**Organized as a start-up within KPMG**

- Core team of Data Scientists
- Separated from the rest of the organization
- Our own P&L targets (*i.e.* not by the hour)
- A strong focus on improving society
- Building solutions, preferably on our own platform
- Ecosystem with partners

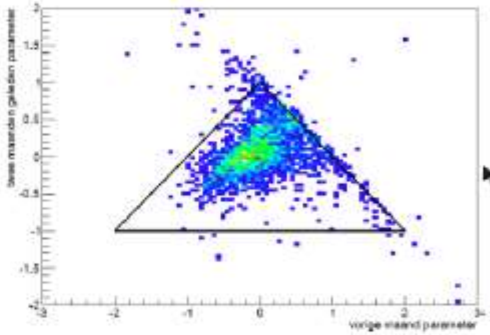

- **2 Vacancies (Data Scientist, Data Architect)**



4

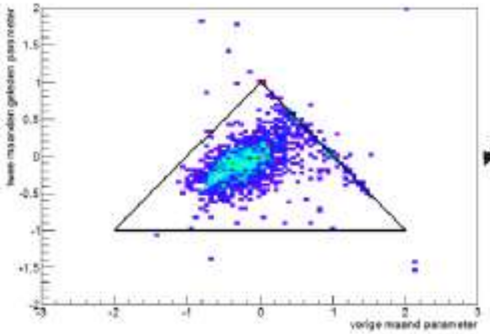
# Modeling with linear differential equations to describe behavior

## Pattern recognition, financial health prediction

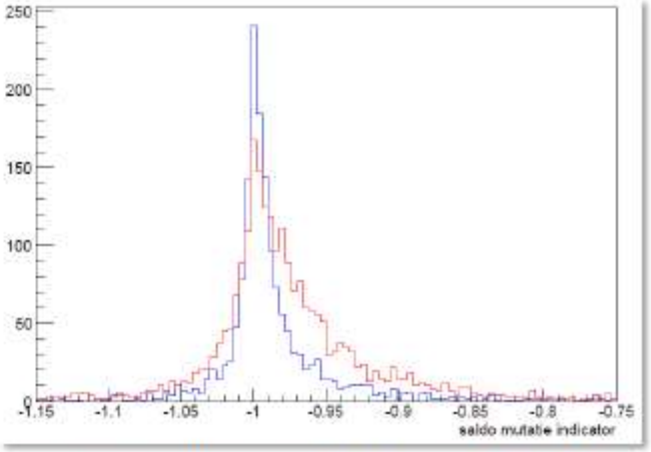



Parameters a & b towards Financial Health Dept.

Expenses within the range for sustainable behavior




Parameters a & b towards Non-Financial Health Dept.




*Customers who were never considered by the Financial Health Dept. balanced their spending with their income.*

# Computing for the LHC: The next step up

Torre Wenaus, BNL/ATLAS

October 15, 2013

CHEP 2013

Amsterdam

BROOKHAVEN NATIONAL LABORATORY

a passion for discovery

Office of Science U.S. DEPARTMENT OF ENERGY

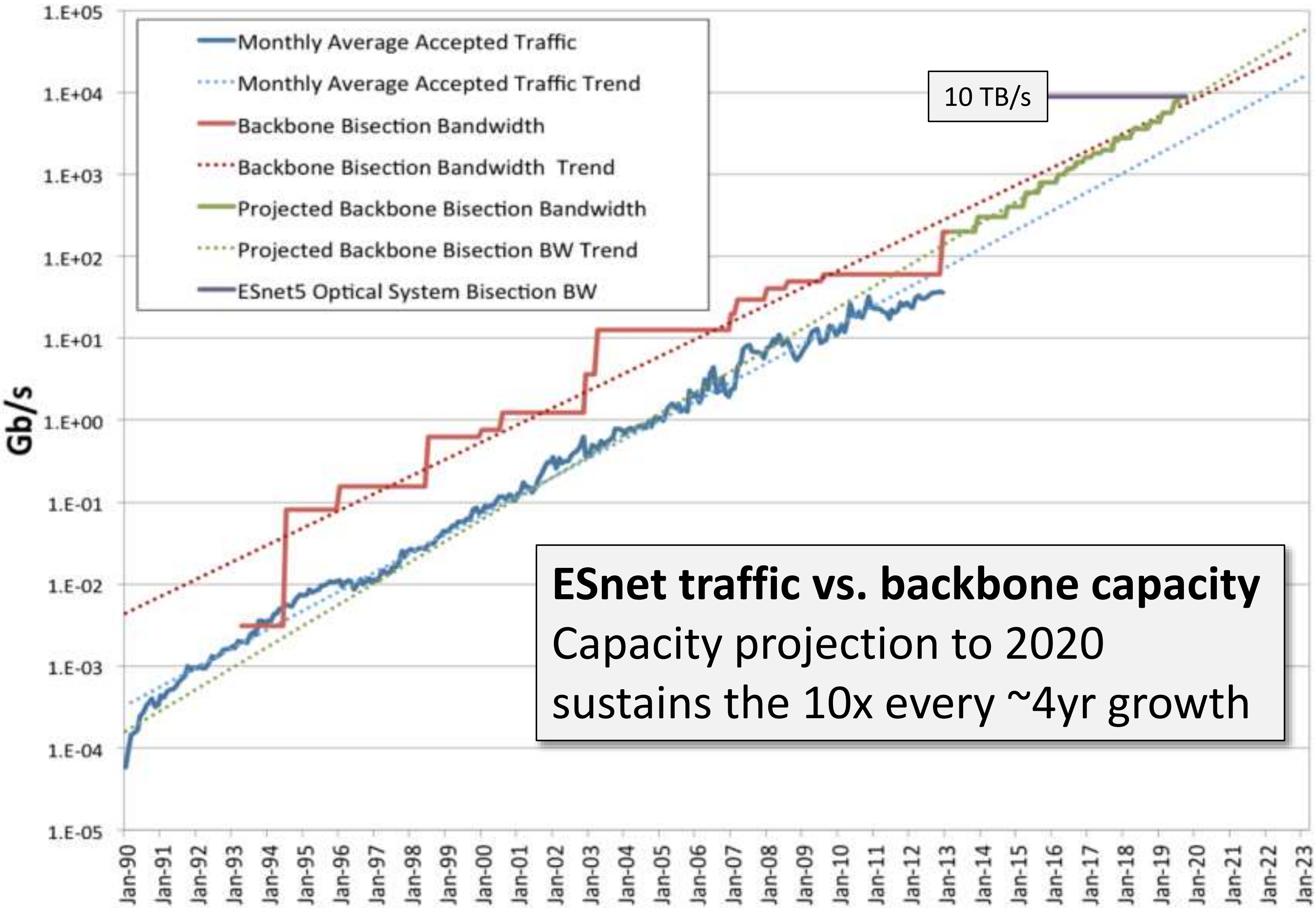
# LHC Computing in Run 2 and Beyond

- **Storage and processing extrapolations lead to unacceptable costs (flat budget assumption)** – we must work on performance and efficiency
- **Storage is largest cost**, e.g. ATLAS spends ~60% more money on disk than on CPU

Most LHC CPU cycles go to simulation (60-70%) – a lot to gain

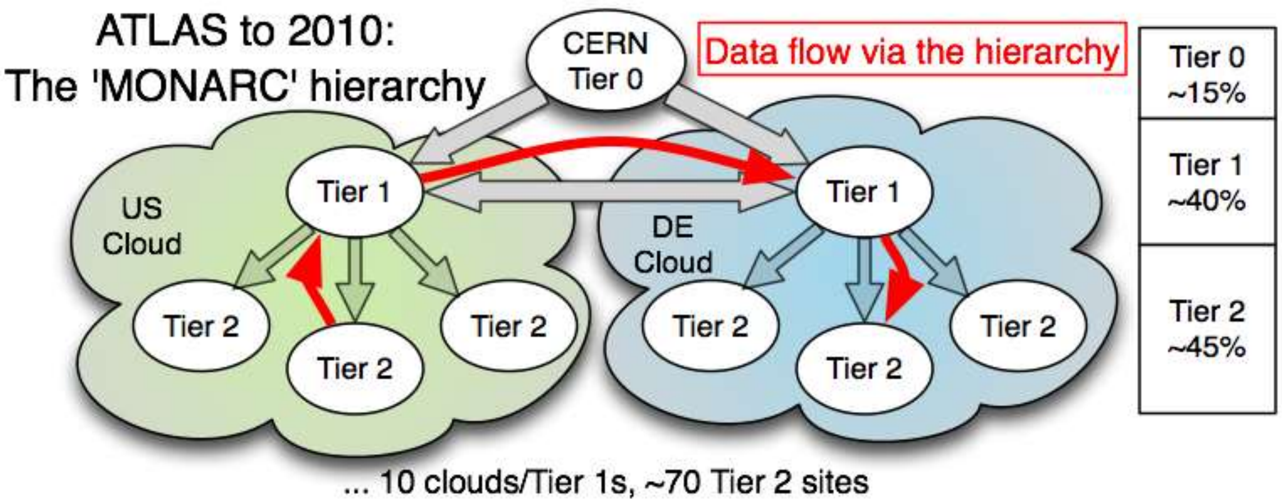
**In general it's much cheaper to transport data than to store it**

# Planned capacity growth sustains the trend




**ESnet traffic vs. backbone capacity**
Capacity projection to 2020
sustains the 10x every ~4yr growth

# Networking has been a critical enabler for evolving LHC computing models – ATLAS as example




**Original model:**

Static strict hierarchy

Multi-hop data flows

Lesser demands on Tier 2 networking

Virtue of simplicity

**Designed for <~2.5 Gb/s within the hierarchy**

**Today:**

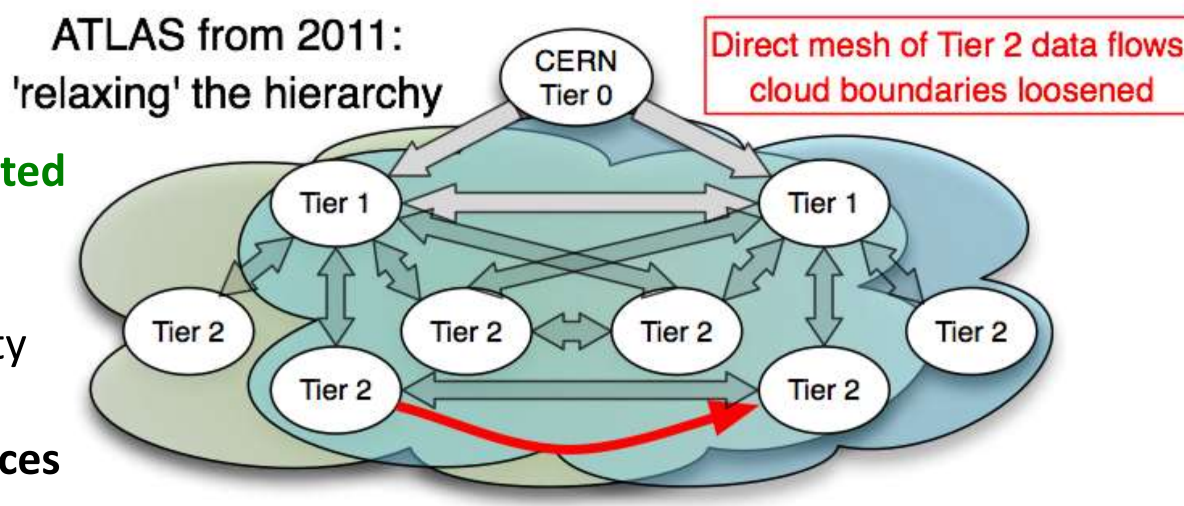
**Bandwidths 10-100 Gb/s, not limited to the hierarchy**

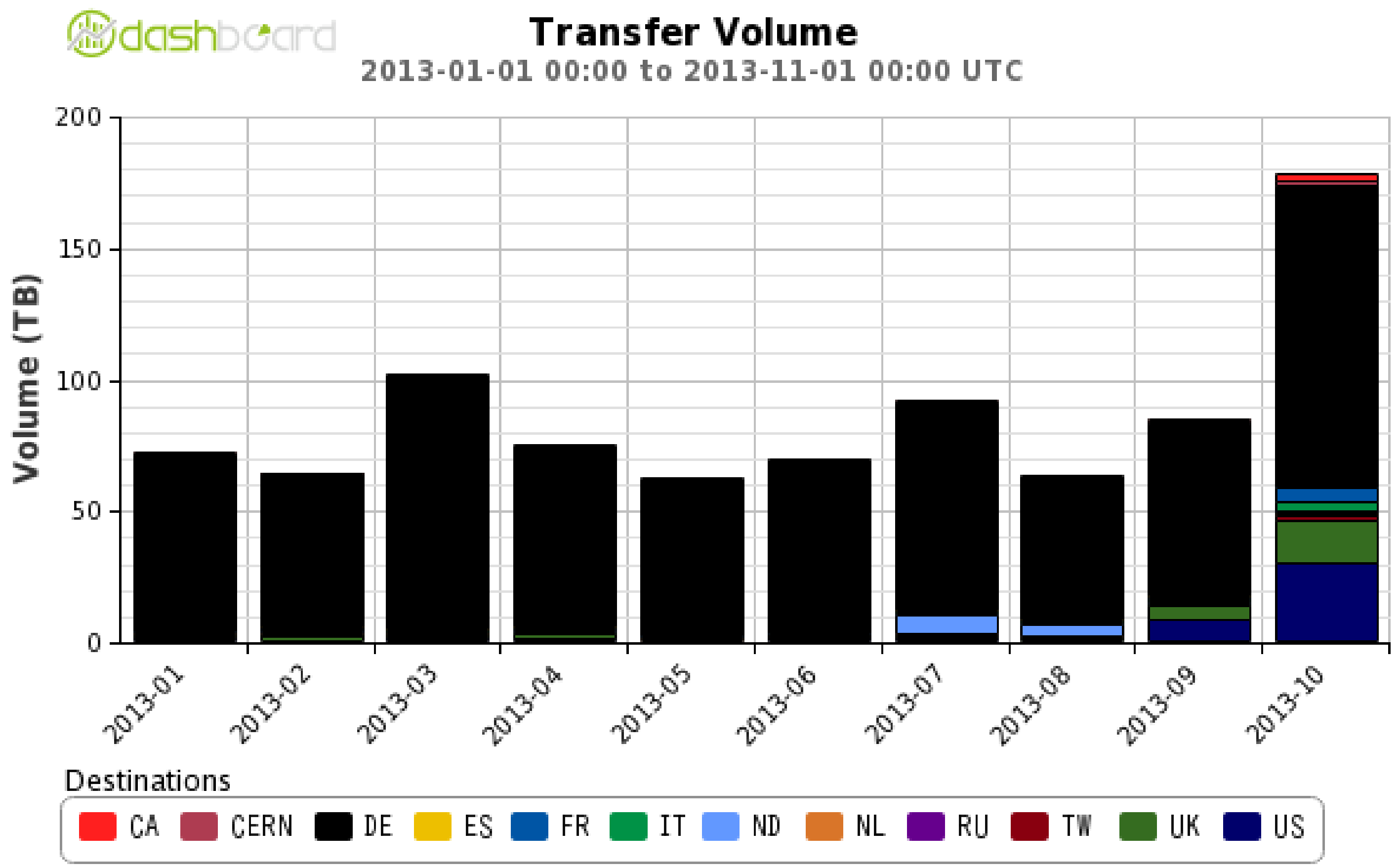
Flatter, mostly a mesh

Sites contribute based on capability

**Greater flexibility and efficiency**

**More fully utilize available resources**

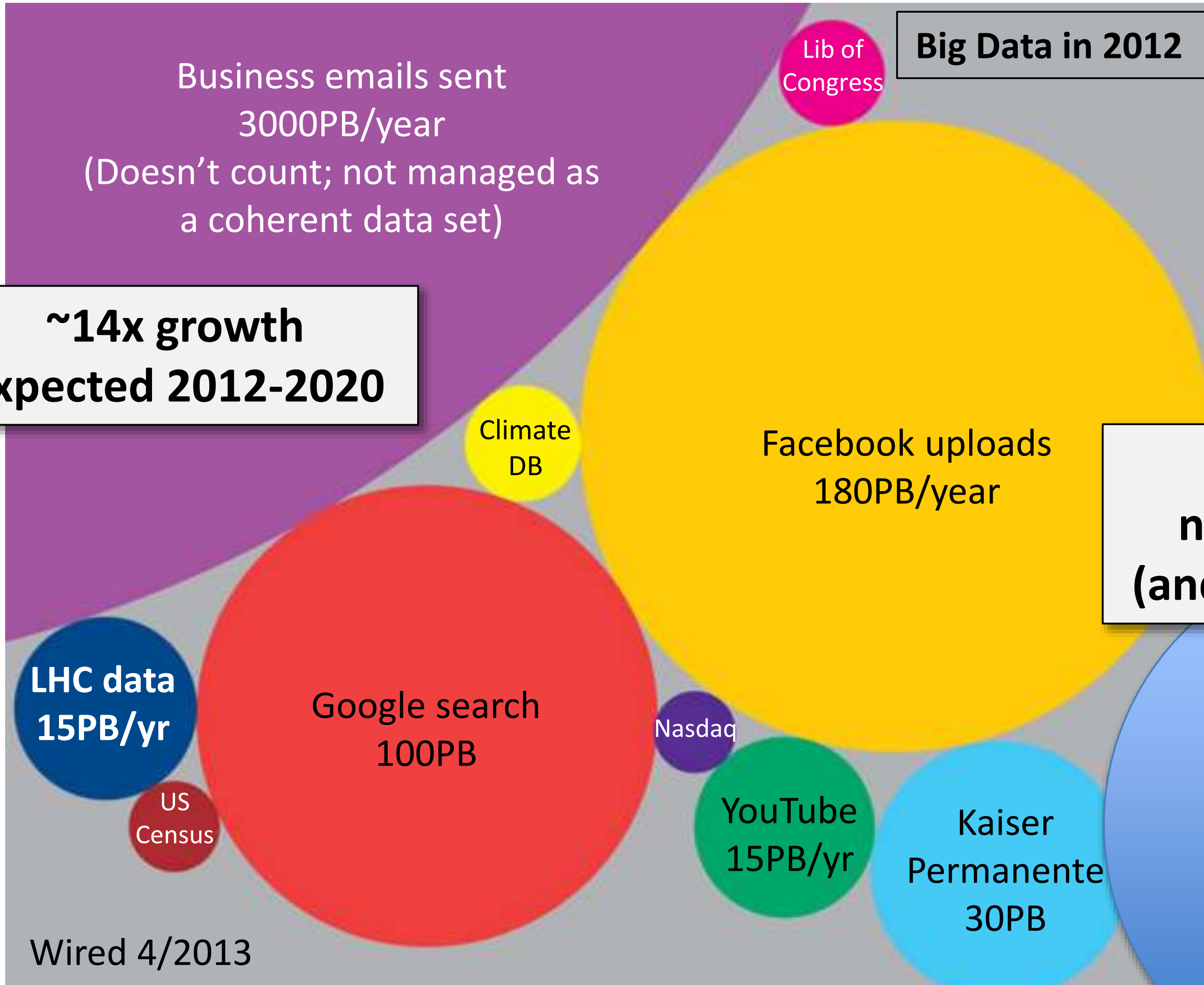





Objem dat přenesených z Tier-2 ve FZÚ do jiných středisek. Dominují přenosy do DE oblasti, jejíž jsme součástí. Po zapojení LHCONE na začátku července lze vidět přímé přenosy do jiných oblastí s výrazným nárůstem v říjnu, kdy se Tier-2 ve FZÚ zařadilo mezi T2D střediska.

# Data Management

## Where is LHC in Big Data Terms?




In 2012: 2800 exabytes created or replicated
1 Exabyte = 1000 PB

Reputed capacity of NSA's new Utah center: 5000 ExaB (50-100 MW)

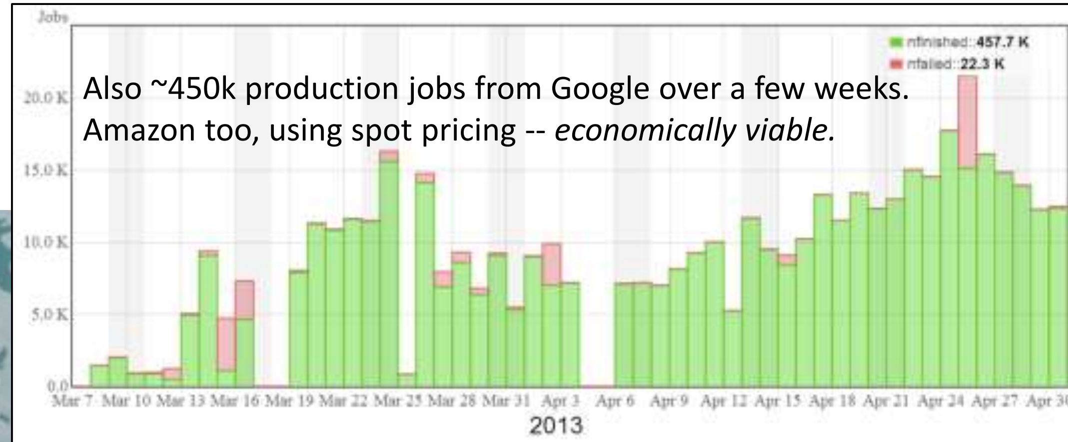
**~14x growth expected 2012-2020**

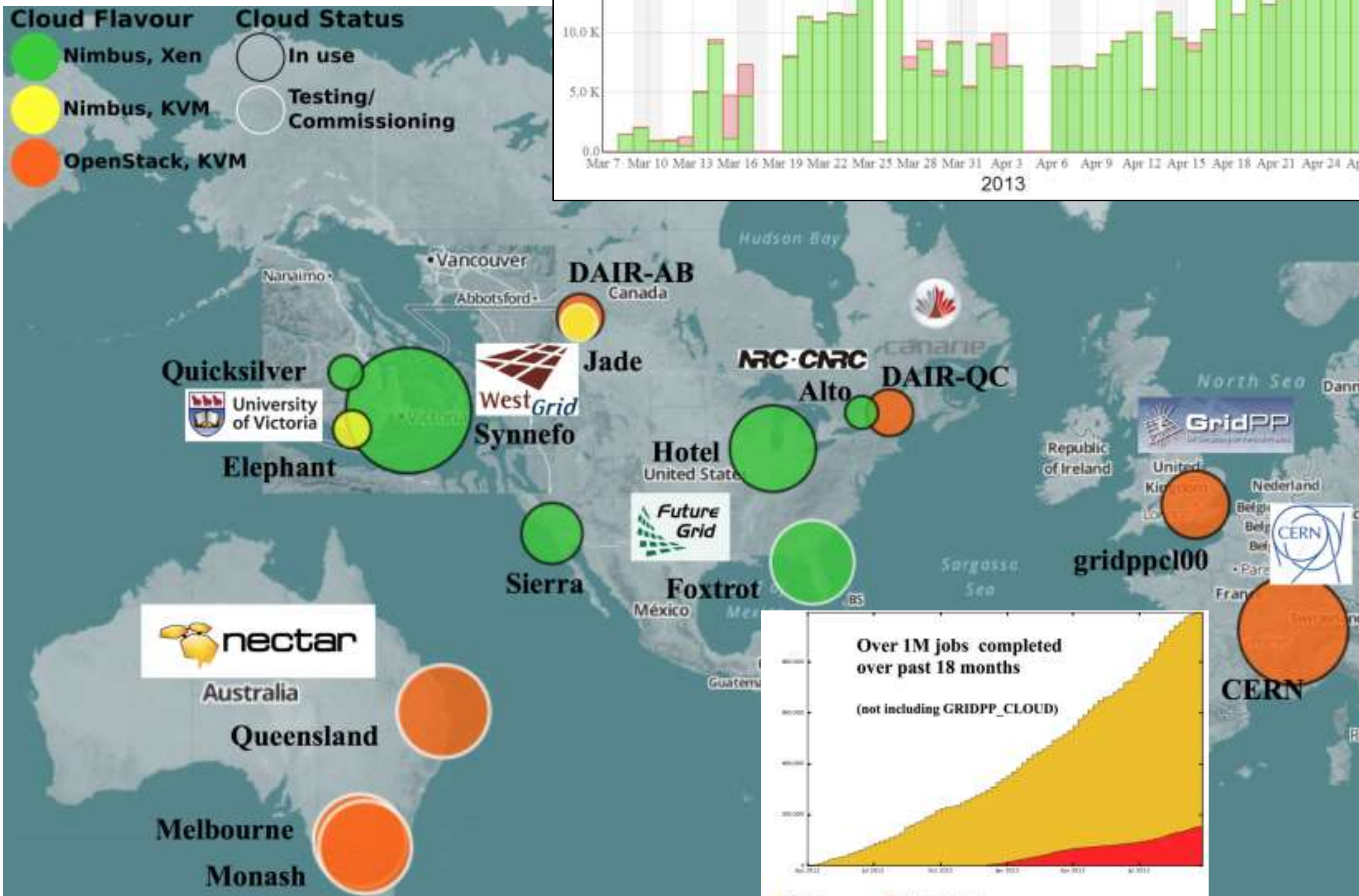
**We are big…**
**not NSA-big, but big**
**(and more cost efficient)**

**Current ATLAS data set, all data products: 140 PB**

http://www.wired.com/magazine/2013/04/bigdata/

# “Grid of Clouds” used by ATLAS



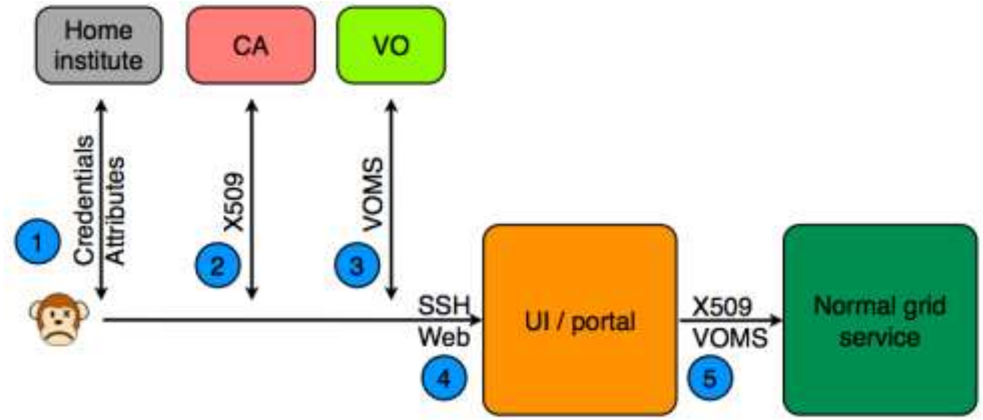
# Opportunistic Resources – HPCs

- HPC (supercomputing) resources can be valuable to HEP computing
- They have cycles open to us – many – even though we wouldn't build the machines that way if we were paying for them (the point is, we aren't)
  - They have holes we can fill: cycles instead of sitting idle would be going to high profile science
  - The *current* US national HPC allocation for HEP is comparable to global CMS+ATLAS computing in 2012, ~1.5B hours
- Also there is increasing convergence, making our apps more appropriate
  - HPC has a growing number of data intensive use cases, future architectures will have to take this into account
  - **More concurrency, leveraging architectures used in HPCs make our applications more suited to HPC**
- We're porting appropriate applications (generators, simulation) and extending workflow and data management systems to support them
- We've begun to put HPC facilities into production

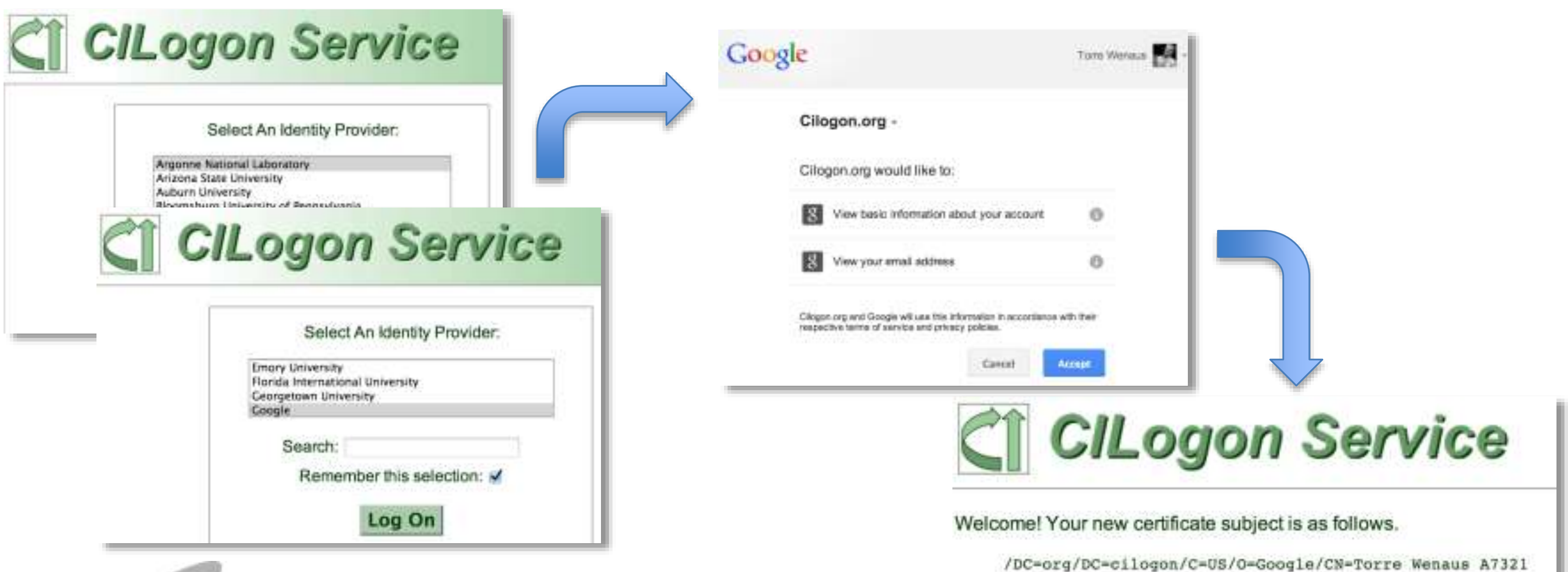
# Ease of Use – Improving on Grid Certificates

Universal authentication is at the root of the grid's success, and yet it's imperfect…

The current bad old days:




WLCG and are pursuing an easy to use (and manage) CILogon.com based service. Objective: A certificate-less grid

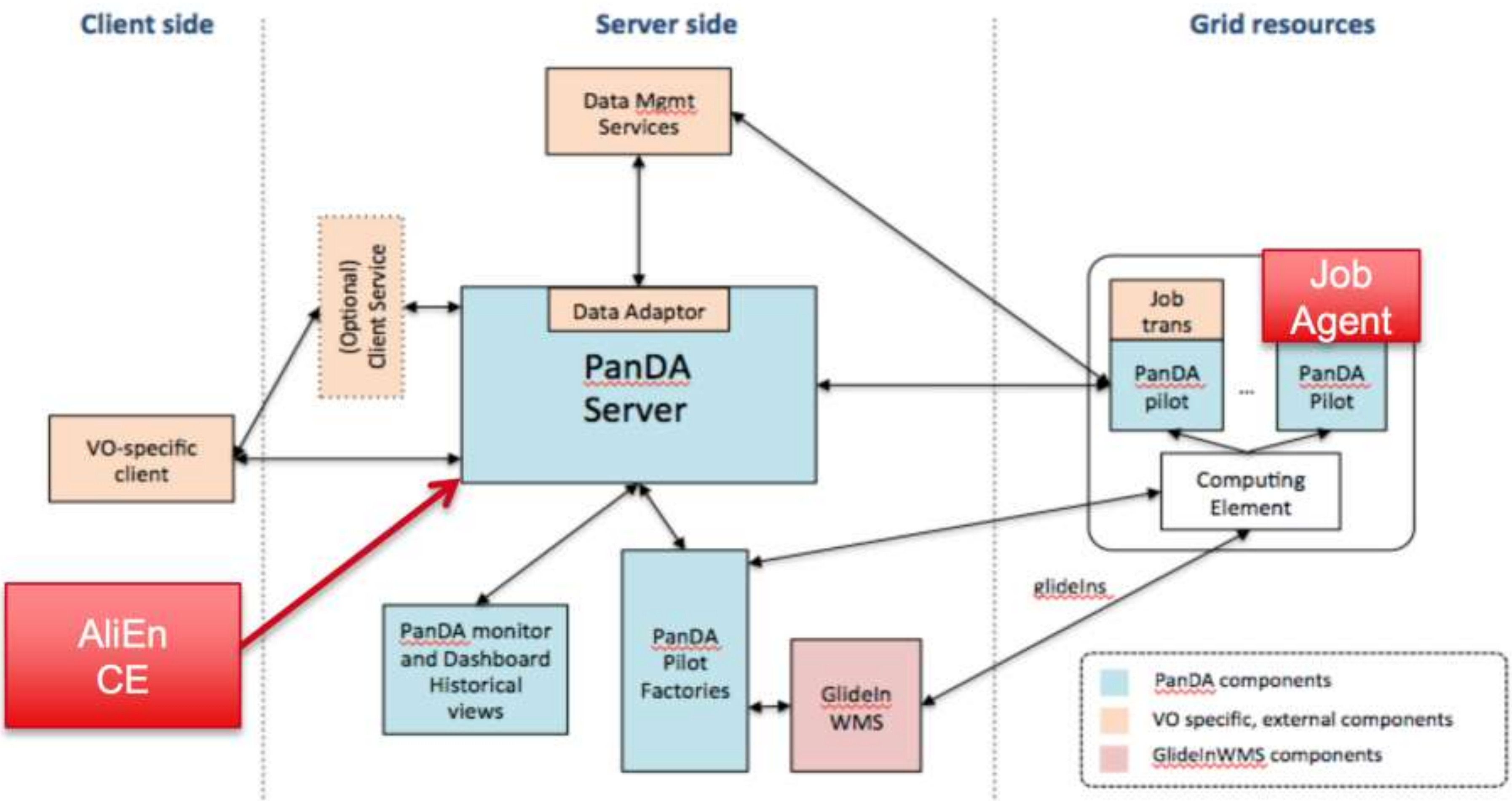


# CMS, ALICE integration with PanDA

PanDA core

Refactoring for CMS (et al)

AliEn integration



# Designing the Computing for the Future Experiments

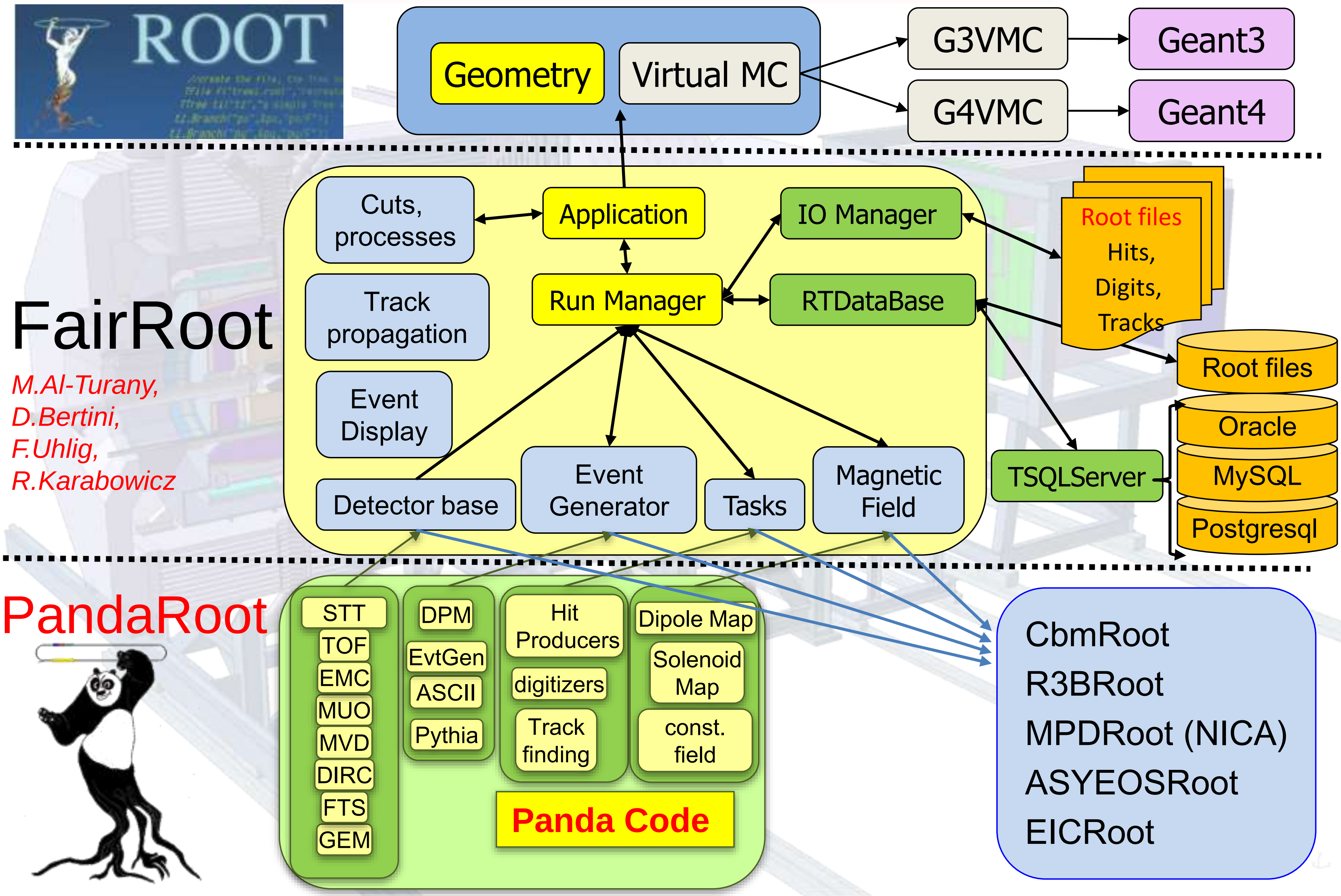

Stefano Spataro

UNIVERSITÀ DEGLI STUDI DI TORINO
ALMA UNIVERSITAS TAURINENSIS

ISTITUTO NAZIONALE DI FISICA NUCLEARE
Sezione di Torino


Tuesday, 15th October, 2013

# The PandaRoot Code Design



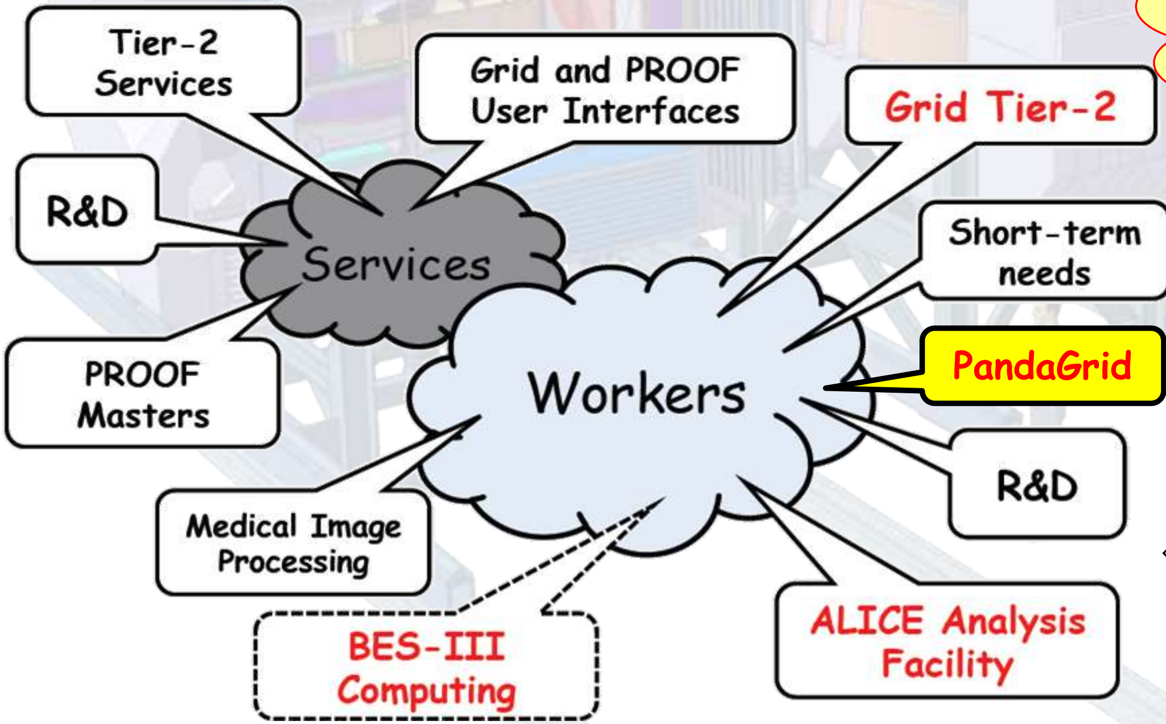
# What is now the future of our distributed computing?

No further developments for Alien2 ⇒

- Alien3?
- PanDa?
- Big batch farm?

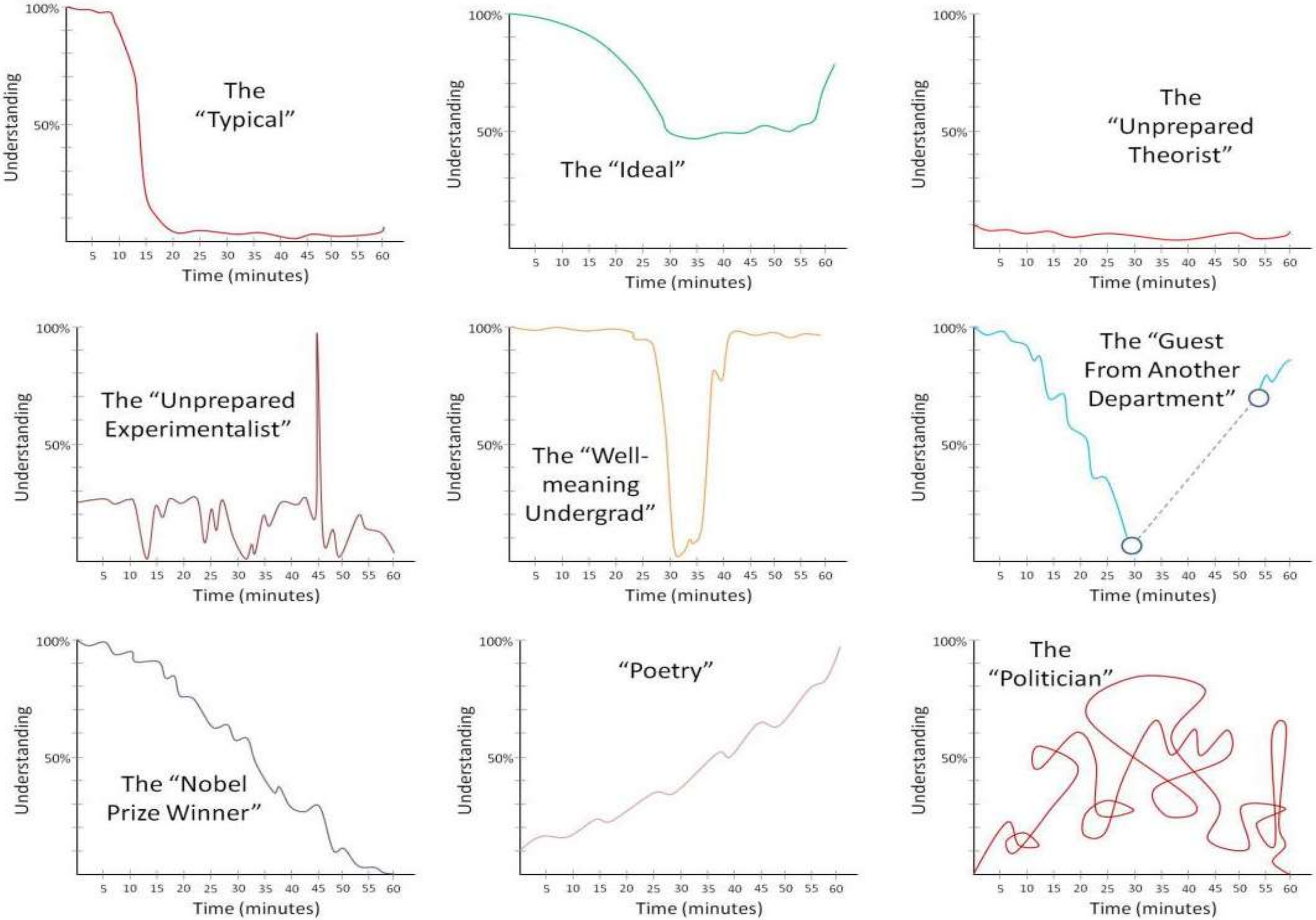
Still some time before taking decisions for PandaGrid

Cloud computing will help us




Torino Private Cloud
(S. Bagnasco et al.)

# The 9 kinds of physics seminar




http://manyworldstheory.com/2013/10/03/the-9-kinds-of-physics-seminar/

# IPv6

- D. Kelsey et al.: WLCG and **IPv6** - the HEPiX **IPv6** working group
- D. Gutierrez et al.: Network architecture and **IPv6** deployment at CERN
- T. Kouba, J. Chudoba, M. Eliáš: Enabling **IPv6** at FZU - WLCG Tier2 in Prague
- A. Petzold: Deploying an **IPv6**-enabled grid testbed at GridKa


iCal export
More
America/Detroit
English
Login
UNIVERSITY OF MICHIGAN
HEPiX
Ann Arbor
Fall 2013
HEPiX Fall 2013 Workshop
from 28 October 2013 to 1 November 2013 (America/Detroit)
University Of Michigan
America/Detroit timezone
Overview
Scientific Programme
Timetable
Contribution List
Author index
Video Services
Remote Participation Info
Accomodations
Transportation
Area Map
Venue
Invitation Letters
Social Events and Info
Participants
Photos
Support
HEPiX Fall 2013 at University of Michigan, Ann Arbor
The HEPiX forum brings together worldwide Information Technology staff, including system administrators, system engineers, and managers from the High Energy Physics and Nuclear Physics laboratories and institutes, to foster a learning and sharing experience between sites facing scientific computing and data challenges. Participating sites include BNL, CERN, DESY, FNAL, IN2P3, INFN, JLAB, NIKHEF, RAL, SLAC, TRIUMF and many others.
Registration for the conference is now closed. We look forward to seeing everyone soon!
ATLAS Great Lakes Tier 2
UNIVERSITY OF MICHIGAN
AGLT2
MICHIGAN STATE UNIVERSITY
Hosted by ATLAS Great Lakes Tier 2 at the University of Michigan in Ann Arbor Michigan.
Starts 28 Oct 2013 08:00
Ends 1 Nov 2013 18:00
America/Detroit
University Of Michigan
Sponsors
Platinum
Gold
ADVANCED RESEARCH COMPUTING
UNIVERSITY OF MICHIGAN
Silver
DataDirect NETWORKS
NetApp
Bronze

# Attendees

- 115 registered participants
  - Including many first-timers!
  - 47 from North-America (including 27 attendees from 8 North-American universities), 48 from Europe, 3 from Asia, 2 from Australia, 15 (!) from companies
- 42 different affiliations
  - 13 from North America, 17 from Europe, 2 from Asia, 1 from Australia, 9 (!) companies
- Fortunate that the "man-made business continuity issue" (US budget crisis) was (temporarily) averted so that US DoE labs could largely participate

3 z FZÚ
2 z CESNET

# Tracks and Trends…

- Security and networking: 9 total, 3 on ***federations***
- Storage and file systems: 10 total, 3 on AFS, 2 on ***CEPH***, ***WD vendor talk***
- Grids/clouds: 7 total, 3 on ***production private clouds***
- Computing: 6 total, 4 on batch systems – ***HTCondor***
- IT facilities and business continuity: 3 total
- Basic IT services: 8 total, 4 on ***Puppet***, 2 on ***log analysis***
- End-user IT services and operating systems: 3 total

# CernVM-FS – Beyond LHC Computing

Ian Collier, Catalin Condurache

STFC RAL Tier 1

HEPiX Ann Arbor November 1st 2013

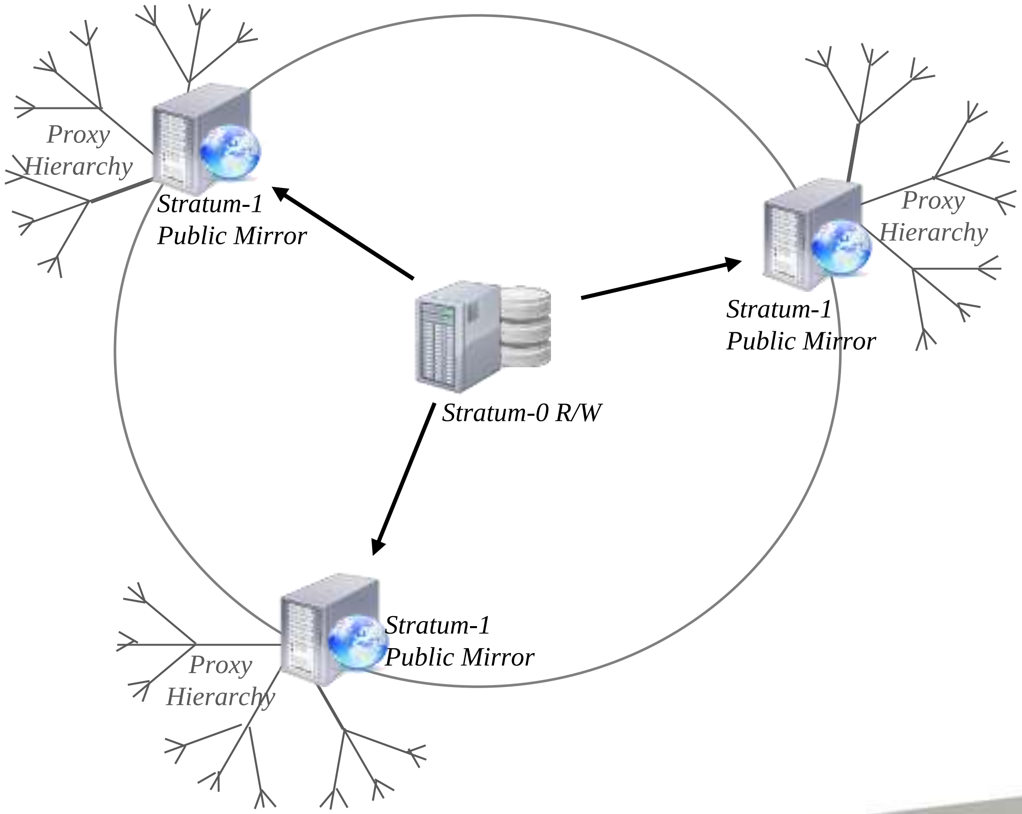
# What is CVMFS?

- Read-only, distributed filesystem, originally developed to get frequently changing VO software to VMs that might not have access to software servers.
- Data integrity and validity are ensured by the signed file catalog and access authentication for software server updates (done by Software Grid Manager or other privileged member of the VO).
- Built using standard technologies (fuse, sqlite, http, squid proxies and caches).
- Removes the need for local installation jobs and conventional software servers at sites & helps standardise the computing environment across the Grid.
- Once the signed catalog has been downloaded and mounted, metadata operations require no further network access. Together with the file based de-duplication this makes CernVM-FS efficient in terms of disk usage and network traffic.
- The software needs one single installation and then is available at any site with CernVM-FS client installed

Science & Technology
Facilities Council

# CernVM-FS WLCG deployment

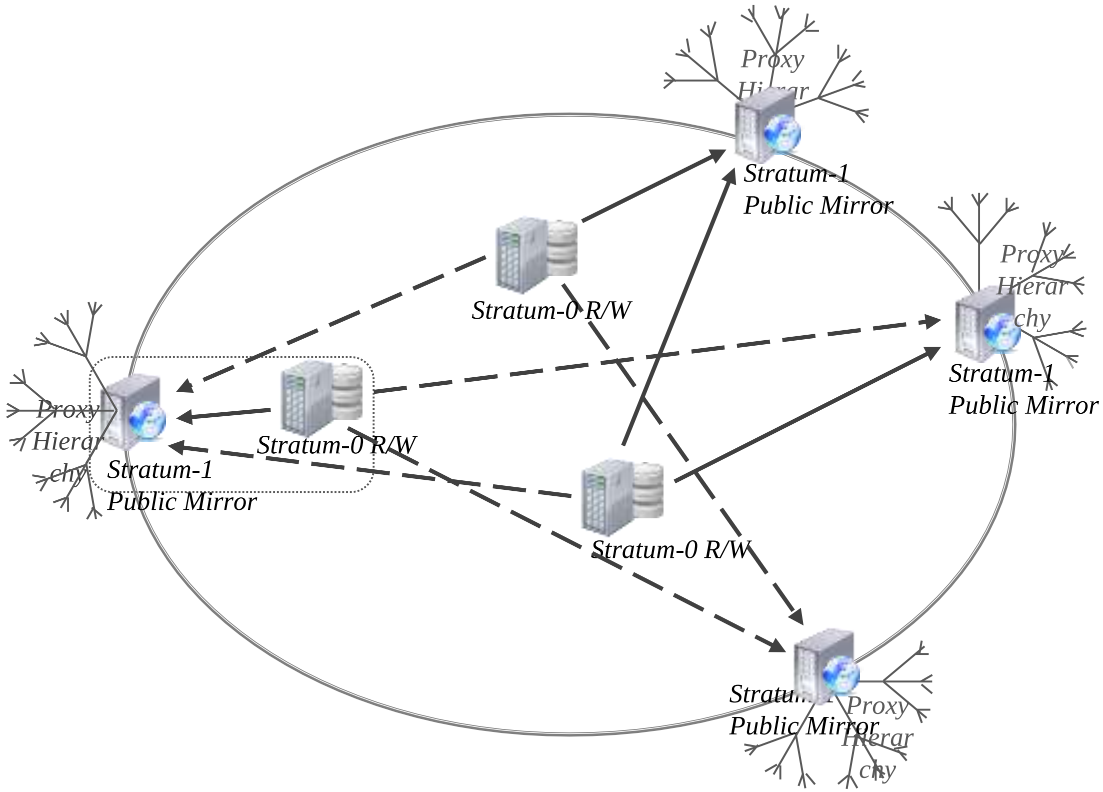
- Software is installed by LHC VOs at Stratum-0 hosted at CERN and replicated to Stratum-1 hosted by WLCG Tier-1 sites
- CernVM-FS clients connect to one of the Stratum-1 services (via local squid caches)
- Client manages transparent failover to other Stratum-1 in case of connection problems




Science & Technology Facilities Council

# CernVM-FS EGI deployment

- Stratum-0 (source repositories) and Stratum-1 (replicas) can be geographically co-located, or not
- Stratum-1 can replicate a whole Stratum-0 (solid), or can partially replicate (dotted) – the ***'relaxed'*** model



# CernVM-FS Stratum-0 Web Frontend

- Web application for CernVM-FS Stratum-0 uploads used as an alternative to installation jobs or 'power users'.
- Developed by a student on an Erasmus Programme placement at RAL-Tier 1 UK.
- Users can upload tarballs and unpack them within the /cvmfs/<repo_name> 'space', followed by synchronization with the real CernVM-FS Stratum-0 repository.
- Authenticates with X509 certificates (managed by a web server)
  - Further authentication mechanisms can be added
- Removes need for privileged roles and jobs at sites

# Hard Disk Drive - Reliability Overview

**Dr. Amit Chattopadhyay**

**Sr. Engineering Manager, Recording Sub-Systems**

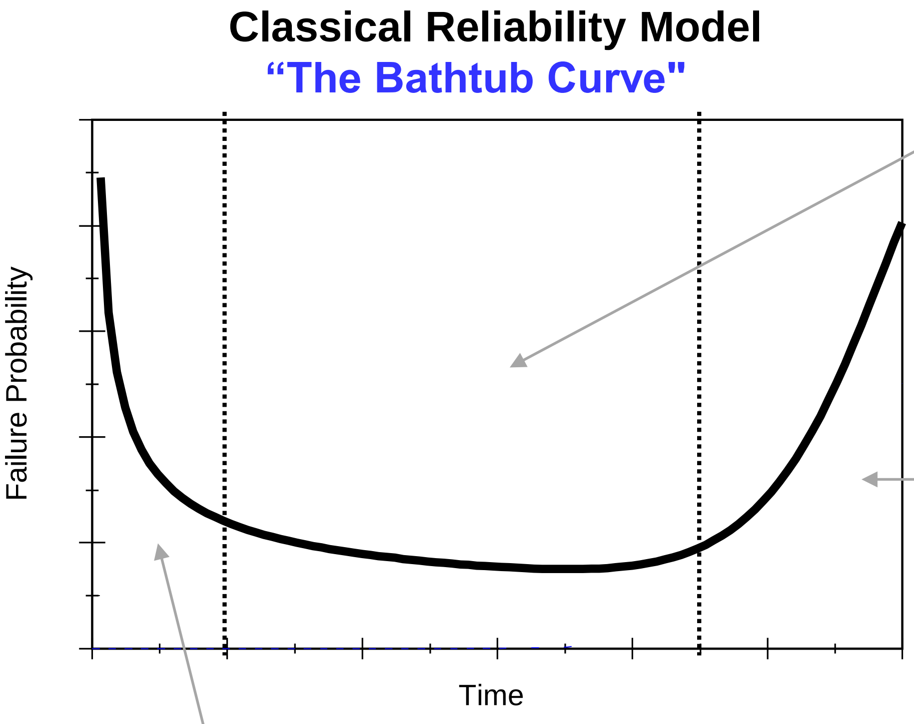
**Advanced Reliability Engineering**

**Western Digital, San Jose**




absolutely™ WD®

# Time to Failure: The “Bathtub” Reliability Model

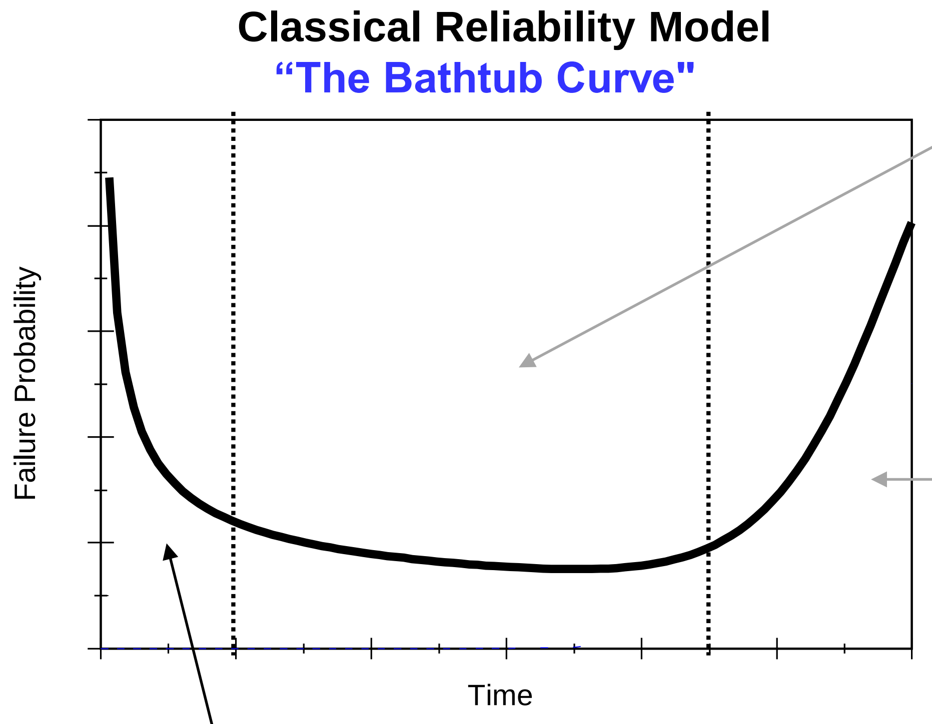



- **Steady State region**
  - After the weak drives are removed from the population, the failure rate reaches a fixed value for the service life of the drive
- **Wear-out Region**
  - At long times, one enters the wear-out region where normal wear and tear of the system components results in an increasing failure rate with time
- **Infant mortality region**
  - Failure rate decreases with increasing time
  - Result of defects etiher designed into, or inadvertently built into a product
    - Indicative of quality “escapes”
    - Marginal materials
    - Drives with the least margin for some critical design tolerance.
    - Manufacturing anomaly



absolutely

# Time to Failure: The "Bathtub" Reliability Model



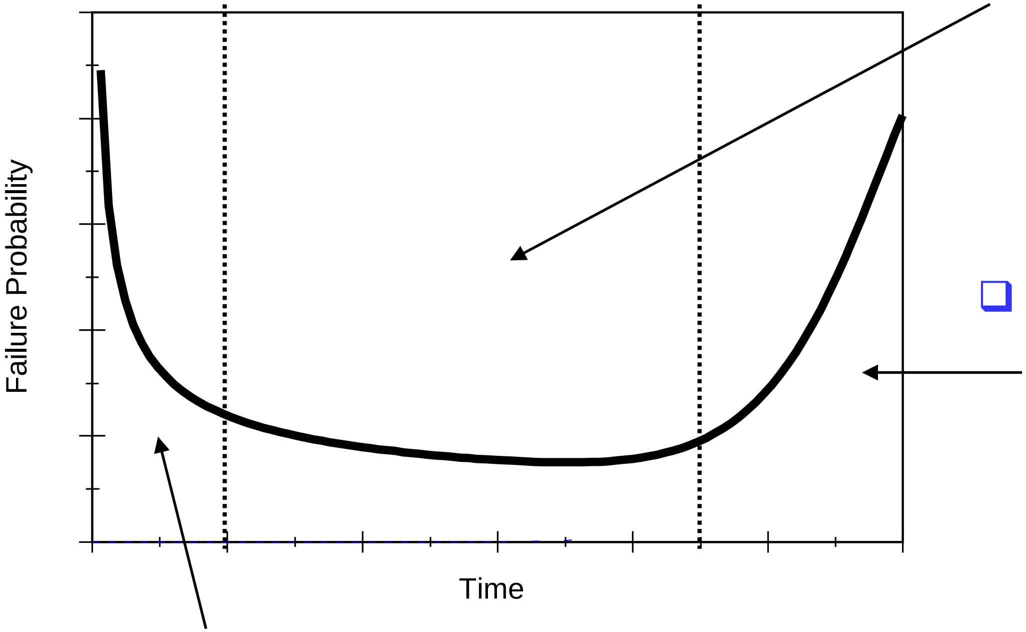

- **Steady State region**
  - After the weak drives are removed from the population, the failure rate reaches a fixed value for the service life of the drive
- **Wear-out Region**
  - At long times, one enters the wear-out region where normal wear and tear of the system components results in an increasing failure rate with time
- **Infant mortality region**
  - Failure rate decreases with increasing time
  - Result of defects etiher designed into, or inadvertently built into a product
    - Indicative of quality "escapes"
    - Marginal materials
    - Drives with the least margin for some critical design tolerance.
    - Manufacturing anomaly



absolutel

# Time to Failure: The "Bathtub" Reliability Model

**Classical Reliability Model**
**"The Bathtub Curve"**




- **Steady State region**
  - After the weak drives are removed from the population, the failure rate reaches a fixed value for the service life of the drive

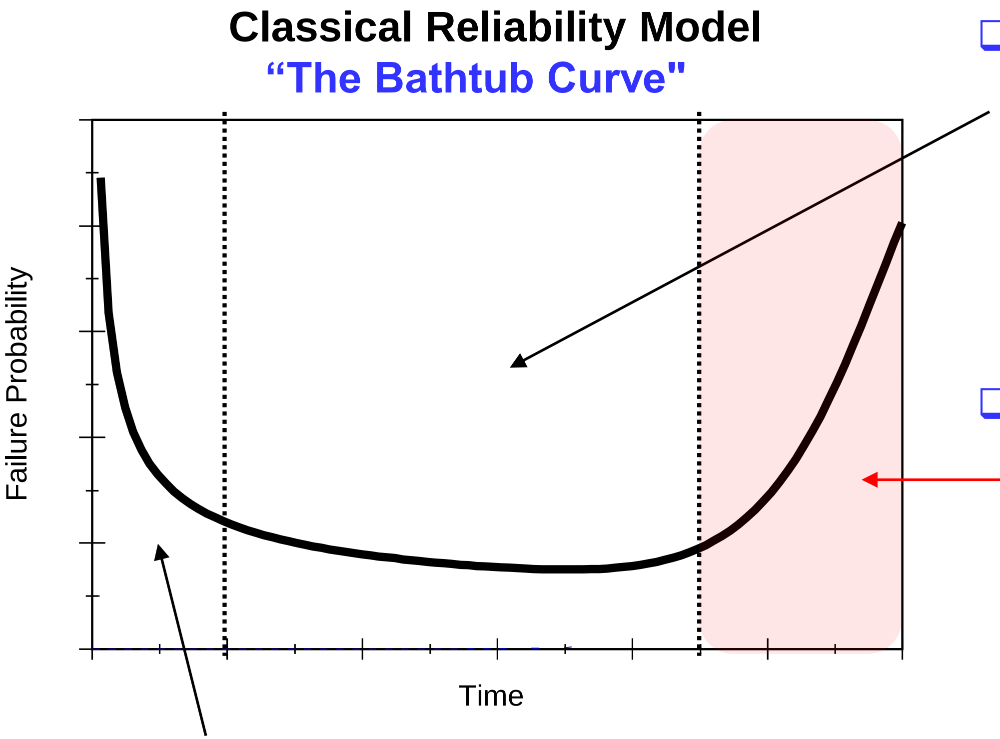
- **Wear-out Region**
  - At long times, one enters the wear-out region where normal wear and tear of the system components results in an increasing failure rate with time

- **Infant mortality region**
  - Failure rate decreases with increasing time
  - Result of defects etiher designed into, or inadvertently built into a product
    - Indicative of quality "escapes"
    - Marginal materials
    - Drives with the least margin for some critical design tolerance.
    - Manufacturing anomaly



absolutel

# Time to Failure: The "Bathtub" Reliability Model




❑ **Steady State region**

- After the weak drives are removed from the population, the failure rate reaches a fixed value for the service life of the drive

❑ **Wear-out Region**

- At long times, normal wear and tear of the system components results in an **increasing failure rate with time**
- This type of behavior results in costly excursions to both WD and our customers
- **This regime must be avoided at all costs**

❑ **Infant mortality region**

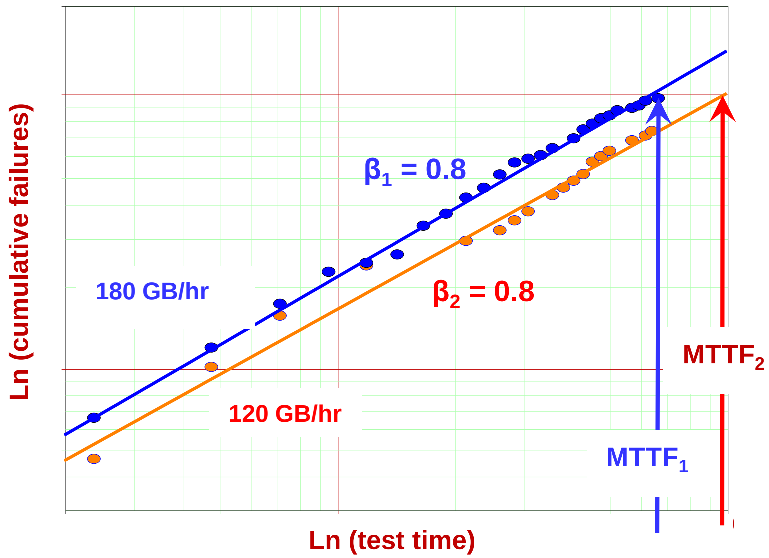
- Failure rate decreases with increasing time
- Result of defects either designed into, or inadvertently built into a product
  - Indicative of quality "escapes"
  - Marginal materials
  - Drives with the least margin for some critical design tolerance.

# Duty Cycle

- **Is the concept of "Duty Cycle" valid?**
  - DOE with same drives built at the same time
  - Two tests with equivalent duty cycles (>95%)
  - ……..but differing workloads (1.5:1)

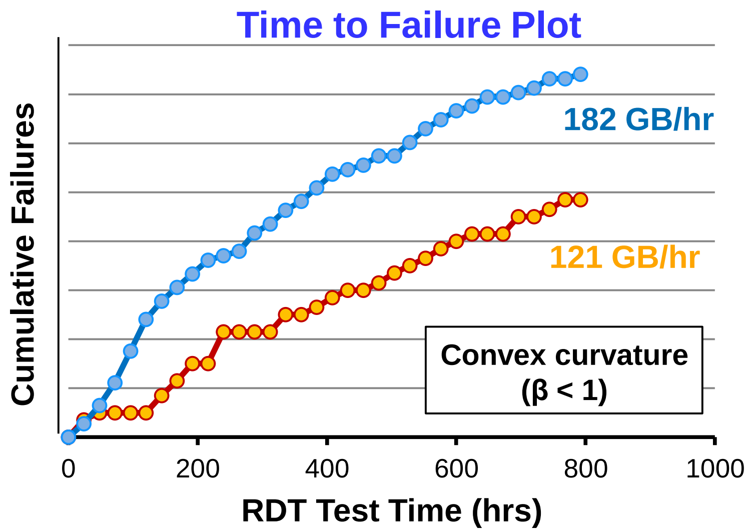
- Results clearly show that **failure rates scale with workload…..not duty cycle**

- **Standard (time-based) Weibull Analysis**




**Same drive / same DC / two workloads = different MTTFs**

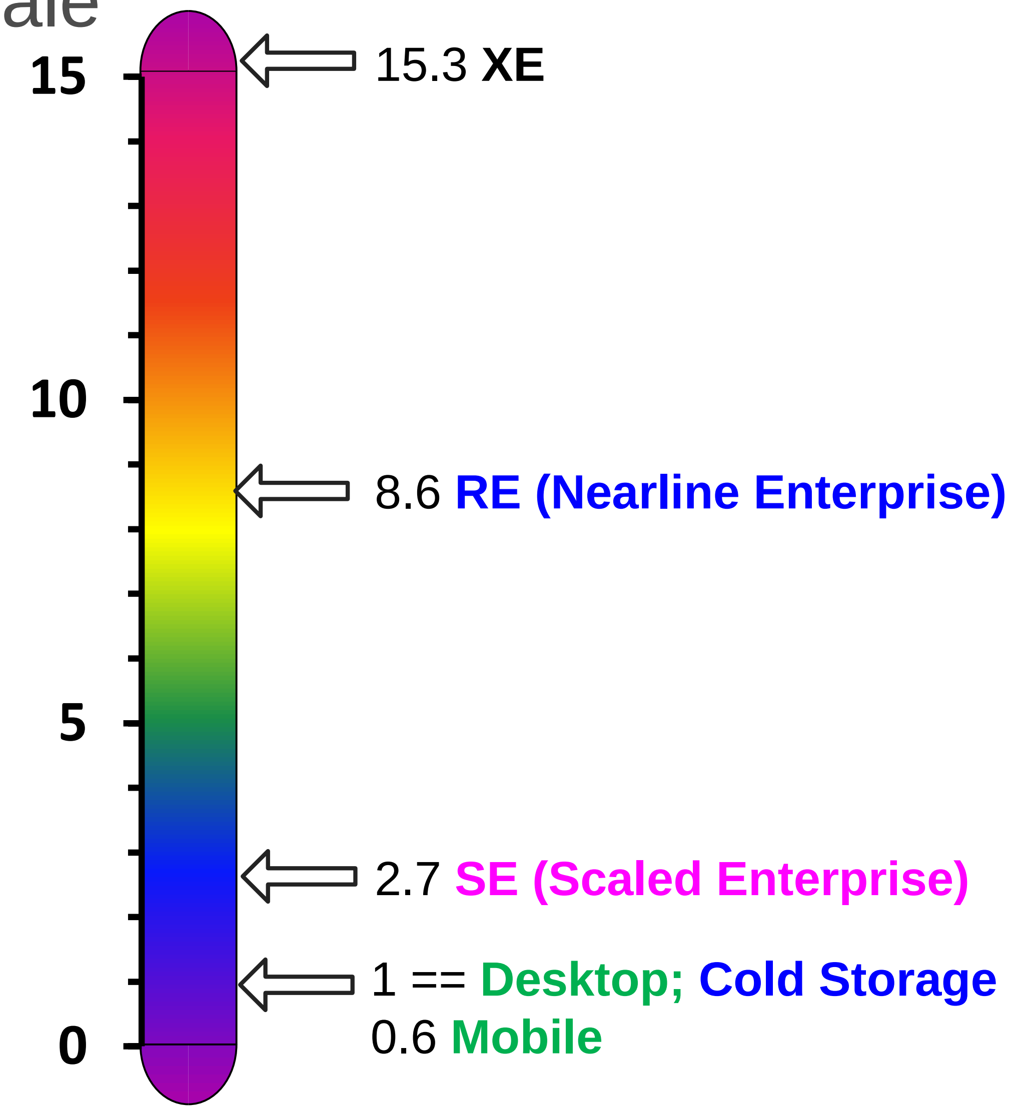

$$F(t,T) \propto \mathrm{DC}^{x} \mathrm{x}\ \mathrm{POH}^{\beta} \times \exp\left[-\frac{E_A}{kT}\right]$$




- **Conclusions**:
  - **MTTF typically used to specify reliability of HDDs**
  - **Since MTTF is not uniquely defined……**
  - **MTTF alone is an insufficient measure of drive reliability!**

# Rough Drive Quality Scale

- Introduced to allow comparison of basic quality requirements
- Priority list for enablers
- Reflective of
  - Intrinsic quality spec: MTTF
  - Workload
  - Temperature
- Normalized to Desktop







absolutely™ WD®

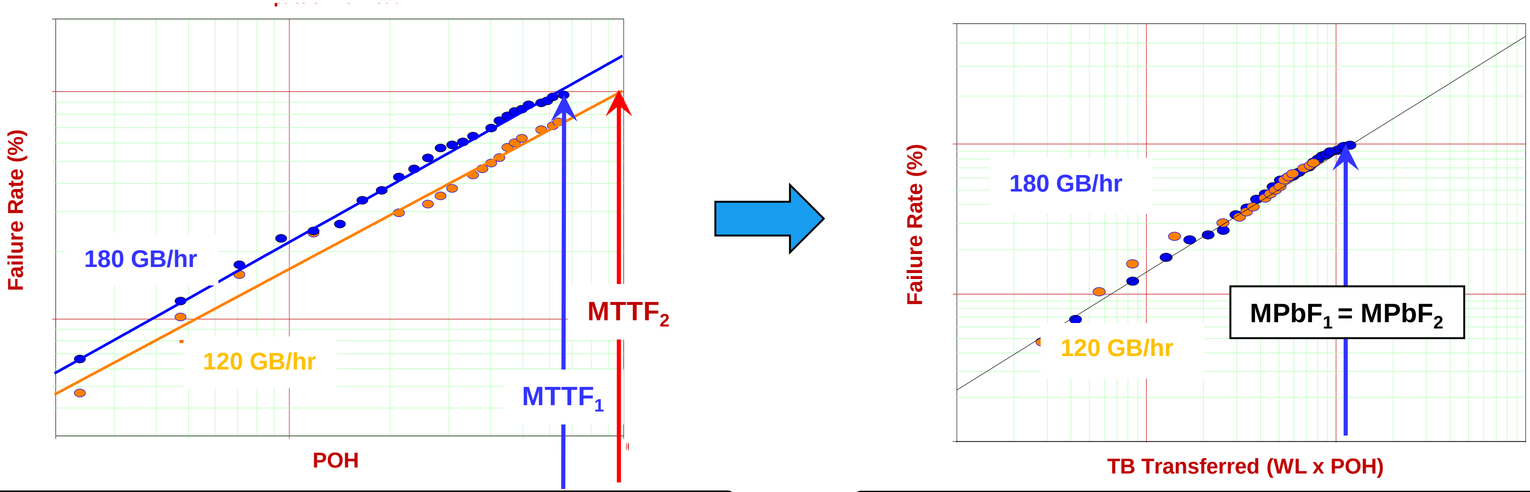
# Validation of Workload Impact on HDD reliability

- Failure rates scale with the total TB transferred

$$AFR \propto (TB)^{\beta}$$

- Weibull Analysis

**Standard (time-based) Treatment** **Workload-based Treatment**




**Same drive + 100% DC / two workloads = different MTTFs**

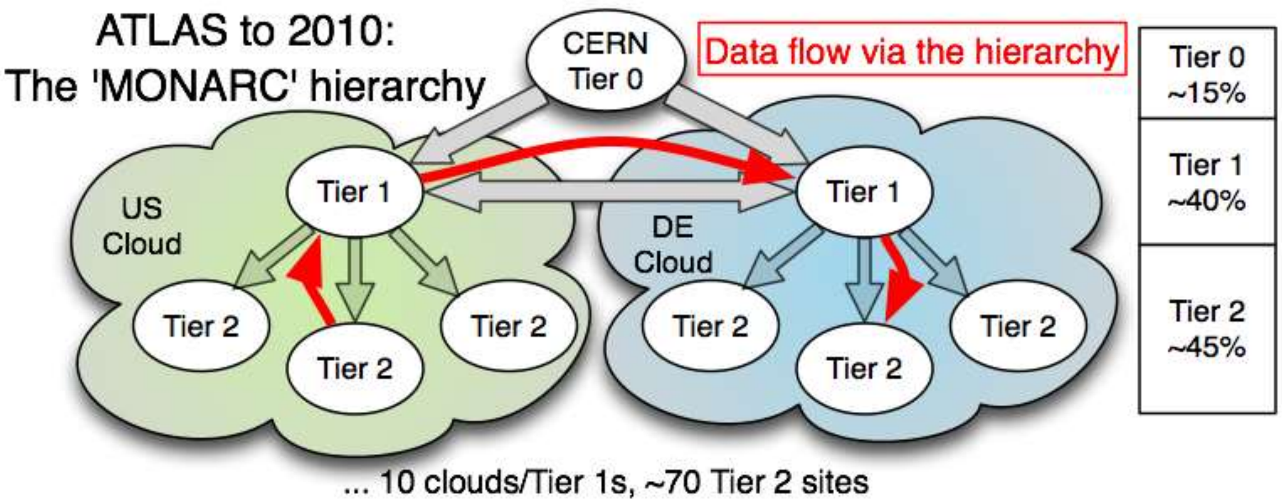
**Same drive +100% DC / two workloads = unique MPbF**

- Results demonstrate that TB transferred is the critical reliability parameter….not time POH
- **Natural reliability metric**: **Mean Petabytes to Failure (MPbF)**

  ➔ **This naturally leads to a DWM (Drive Workload Monitor) (like an odometer)**

- **Minimum requirement**: Simultaneously define **max workload spec and MTTF**
  - **This is now done by all HDD manufacturers**

# Networking has been a critical enabler for evolving LHC computing models – ATLAS as example




**Original model:**

Static strict hierarchy

Multi-hop data flows

Lesser demands on Tier 2 networking

Virtue of simplicity

**Designed for <~2.5 Gb/s within the hierarchy**

**Today:**

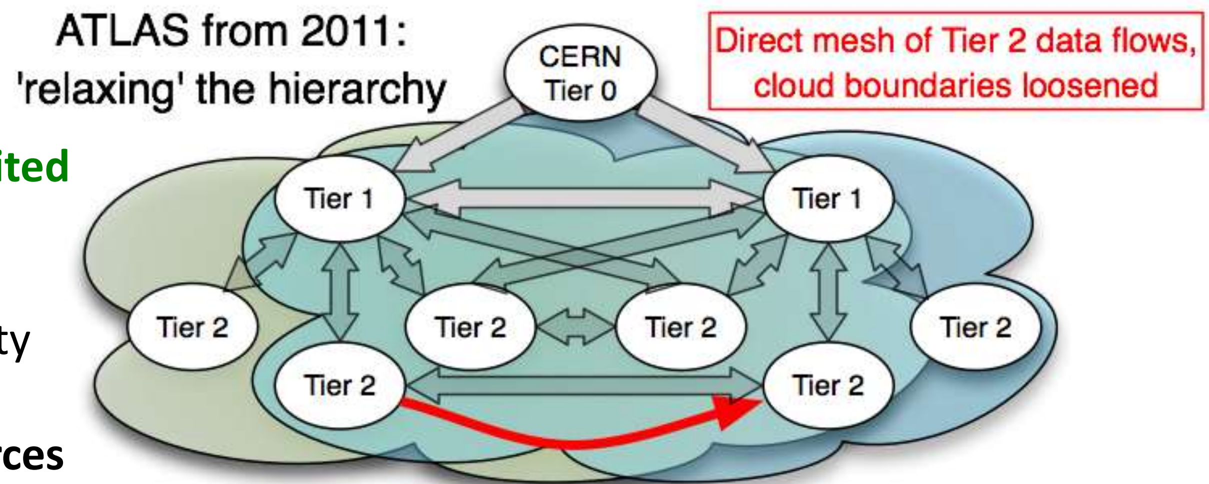
**Bandwidths 10-100 Gb/s, not limited to the hierarchy**

Flatter, mostly a mesh

Sites contribute based on capability

**Greater flexibility and efficiency**

**More fully utilize available resources**



# Impact on HEP labs?

- Politically motivated attacks and surveillance
  - Who owns your routers?
    - It is pretty difficult to determine
    - (Tip: setting your User Agent to "xmlset_roodkcableoj28840ybtide" gives instant root on many D-Link routers)
  - How can you protect your staff and users?
    - Data privacy is a significant concern
    - (And a marketable feature)
- Now facing extreme levels of sophistication (political/money)
  - Complex malware, complex infrastructures
  - Far too much expertise needed for an average site/system admin
- Important to have or be in touch with knowledgable experts
  - If not possible, then join existing efforts and contribute
  - Many groups of trusted experts always keen to help!

WLCG Worldwide LHC Computing Grid

Adv: Selling Iframe traffic in a huge amount JID#1: [redacted]@jabber.ru icq#1: [redacted] JID#2: [redacted]@jabber.org icq#2: [redacted]
Adv: [redacted].net - comfortable buying\selling iframe traffic with no limits. 256 countries. 24/7. Loads from 8%. Tell password "blackhole" and get +5% to the first order.

19

# Operating Dedicated Data Centers – Is It Cost-Effective?

HEPIX – University of Michigan

Tony Wong - Brookhaven National Lab

# Amazon EC2

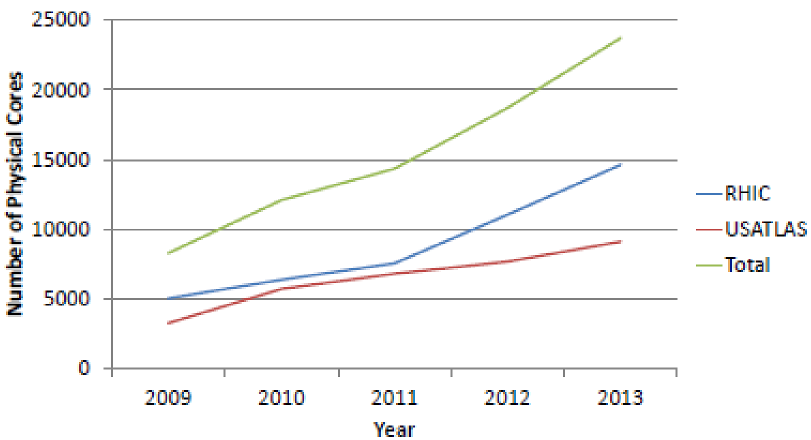
| | Type | ECU | RAM (GB) | Storage (GB) | Network I/O | Cost/hr (US$) |
|---|---|---|---|---|---|---|
| spot | m1.small | 1 | 1.7 | 160 | low | 0.007 |
| spot | m1.medium | 2 | 3.75 | 410 | moderate | 0.013 |
| On-demand | m1.medium | 2 | 3.75 | 410 | moderate | 0.12 |

- Full details at aws.amazon.com/ec2/pricing.
- Linux virtual instance
  - 1 ECU = 1.2 GHz Xeon processor from 2007 (HS06 ~ 8/core)
  - 2.2 GHz Xeon (Sandybridge) in 2013 → HS06 ~ 38/core
- Pricing is dynamic and region-based. Above prices were current on August 23, 2013 for Eastern US.

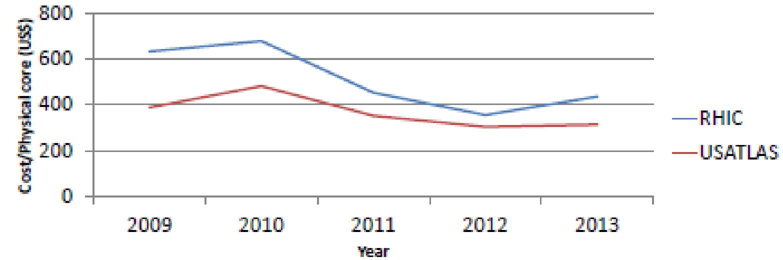
# BNL Experience with EC2

- Ran ~5000 EC2 jobs for ~3 weeks (January 2013)
  - Tried m1.small with spot instance
  - Spent US $13k
- Strategy
  - Declare maximum acceptable price, but pay current, variable spot price. When spot price exceeds maximum acceptable price, instance (and job) is terminated without warning
  - Maximum acceptable price = 3 x baseline → $0.021/hr
- Low efficiency for long jobs due to eviction policy
- EC2 jobs took ~50% longer (on average) to run when compared to dedicated facility

## Growth of RACF Computing Cluster

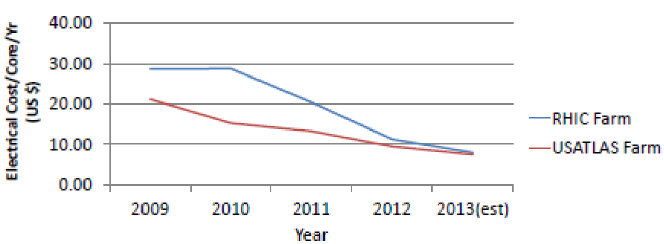



## Server Costs



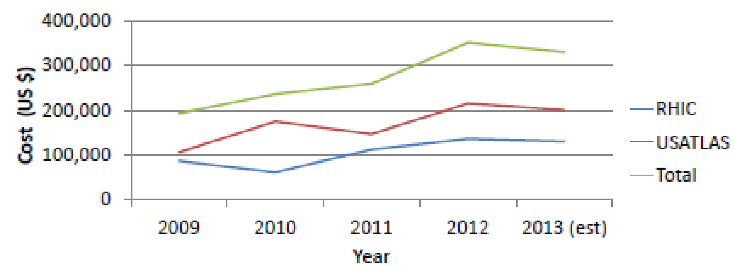

- Standard 1-U or 2-U servers
- Includes server, rack, rack pdu's, rack switches, all hardware installation (does not include network cost)
- Hardware configuration changes (ie, more RAM, storage, etc) not decoupled from server costs → partly responsible for fluctuations

## Electrical Costs




- Increasingly power-efficient hardware has decreased power consumption per core at the RACF in recent years
- RHIC costs higher than USATLAS due to differences in hardware configuration and usage patterns
- Average instantaneous power consumption per core was ~25 W in 2012

## Overall Data Center Space Charges




- Overhead charged to program funds to pay for data center infrastructure (cooling, UPS, building lights, cleaning, physical security, repairs, etc) maintenance—upward trend a concern
- Based on footprint (~13,000 ft² or ~1200 m²) and other factors
- USATLAS occupies ~60% of the total area.
- Rate reset on a yearly basis – not predictable

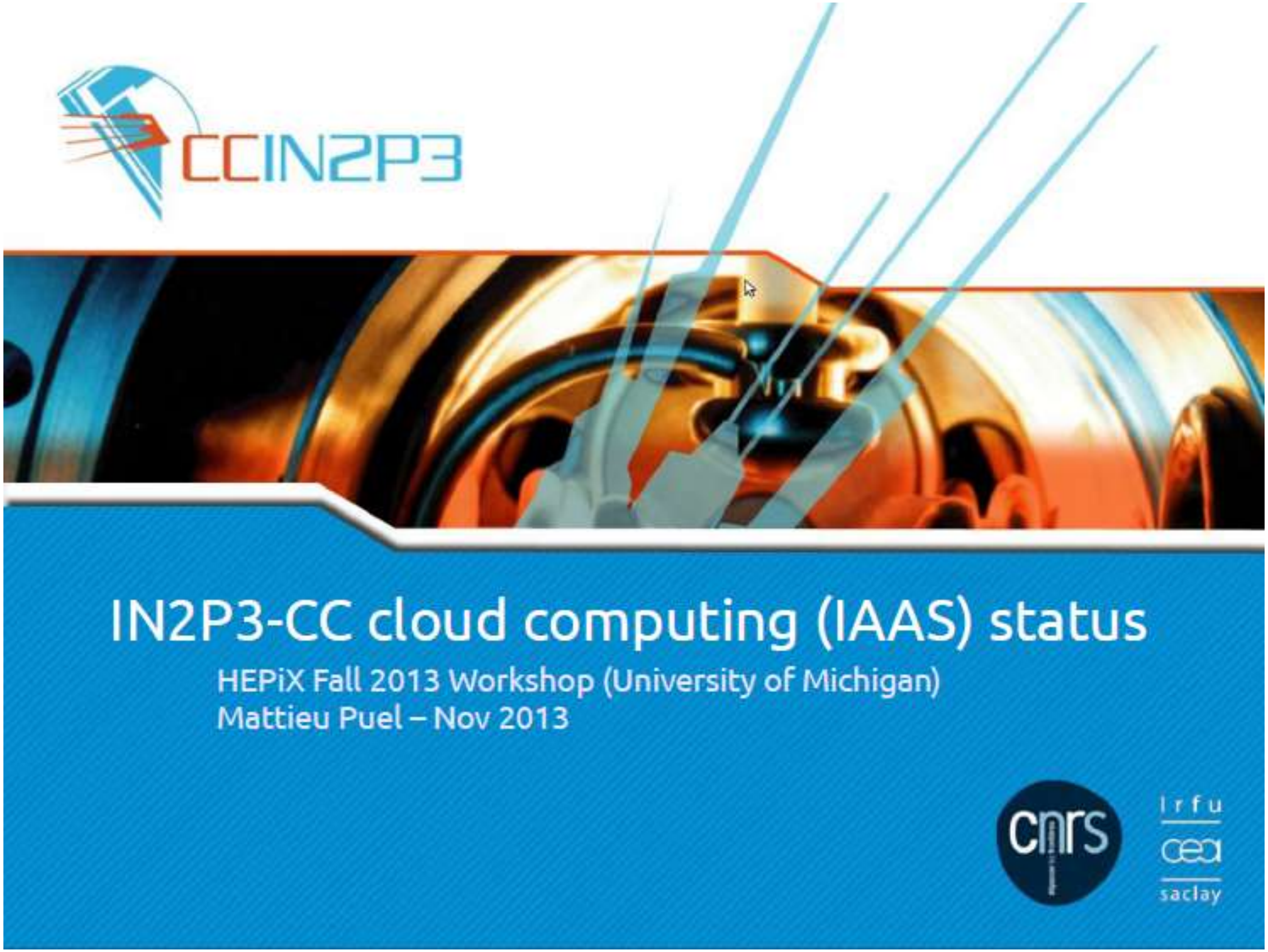
# Historical Cost/Core

| | USATLAS | RHIC |
|---|---|---|
| Server | $228/yr | $277/yr |
| Network | $28/yr | $26/yr |
| Software | $3/yr | $3/yr |
| Staff | $34/yr | $34/yr |
| Electrical | $12/yr | $16/yr |
| Space | $27/yr | $13/yr |
| Total | $332/yr ($0.038/hr) | $369/yr ($0.042/hr) |

- Includes 2009-2013 data
- BNL-imposed overhead included
- Amortize server and network over 4 or 6 (USATLAS/RHIC) years and use only physical cores
- RACF Compute Cluster staffed by 4 FTE ($200k/FTE)
- About 25-31% contribution from other-than-server

# Summary

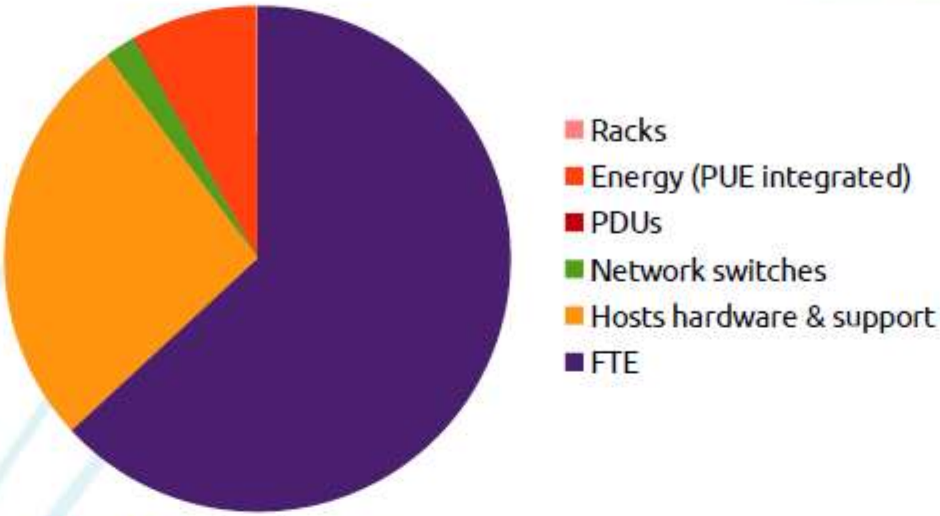
- Cost of computing/core at dedicated data centers compare favorably with cloud costs
  - $0.04/hr (RACF) vs. $0.12/hr (EC2)
  - Near-term trends
    - Hardware ↘
    - Infrastructure ↗
    - Staff ⇒
    - Data duplication ↗
- Data duplication requirements will raise costs and complexity – not a free ride
- This doesn't mean cloud computing isn't useful –it is– but dedicated resources can be competitively priced


CCIN2P3
IN2P3-CC cloud computing (IAAS) status
HEPiX Fall 2013 Workshop (University of Michigan)
Mattieu Puel – Nov 2013
cnrs
Irfu
cea
saclay

# Public cloud considerations : costs

CCIN2P3

## Cost per sector




Some assumptions / moderations :

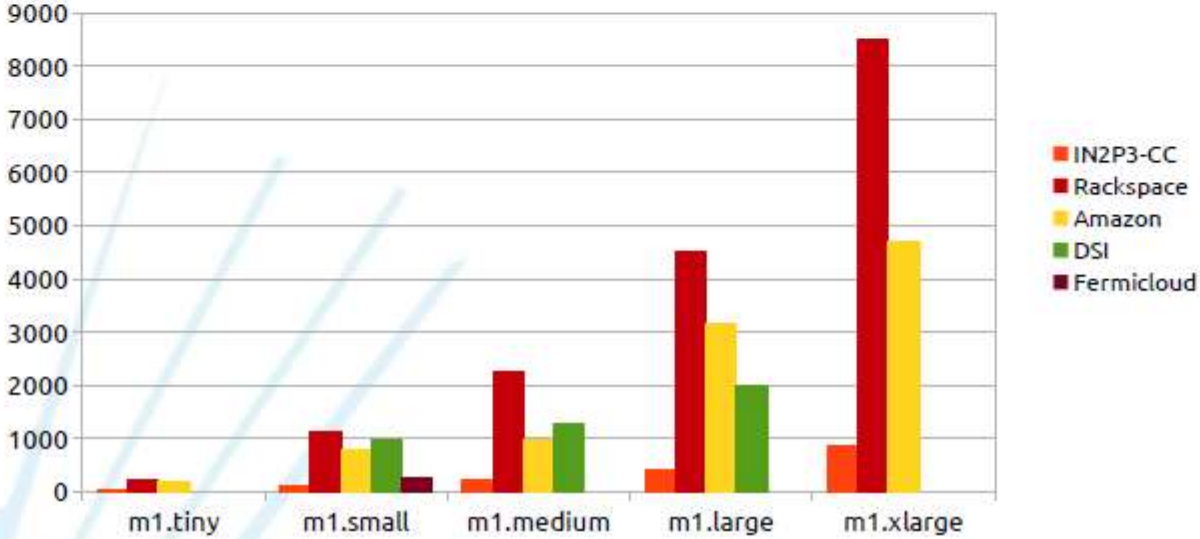
- Free software
- Costs evaluated for a 400 VM infra of ~400k$ (but the bigger, the cheaper right ?)
- French energy not that costly (is it ?) **0.0874$/kWh**
- French employment not that cheap... but public sector employee is (uhhh ME ?)
- One admin per 400 Vms

9

# Public cloud pricing

CCIN2P3

Pricing comparison ($/year)




Some assumptions / moderations :

- Based on memory capacity (often the lacking resource in virtualized envs)
- Disk is the cheapest ressource
- CPU is expensive, but is more shareable, depending on the SLA

10

# Další témata

- OpenAFS vs YFS
- IPv6
- Puppet
- Perfsonar
- HPC

# DĚKUJI ZA POZORNOST!

# HEPiX

- Site reports
- Security & Networking
- Storage & Filesystems
- Grid, Cloud & Virtualisation
- Computing & Batch Services
- IT Facilities & Business Continuity
- Basic IT Services
- End-user IT Services & Operating Systems
- Miscellaneous