# ATLAS Computing in Run-2

Borut Paul Kersevan

WLCG Workshop - Copenhagen, November 2013

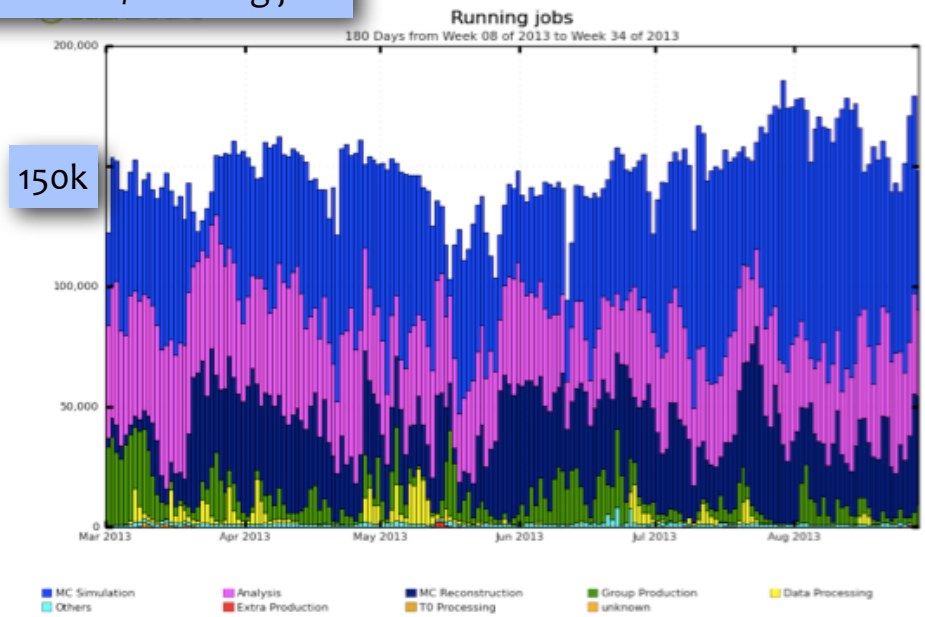
# ATLAS Resource Utilization in 2013

## ATLAS RESOURCE USAGE IN FIRST HALF OF 2013 (RRB)

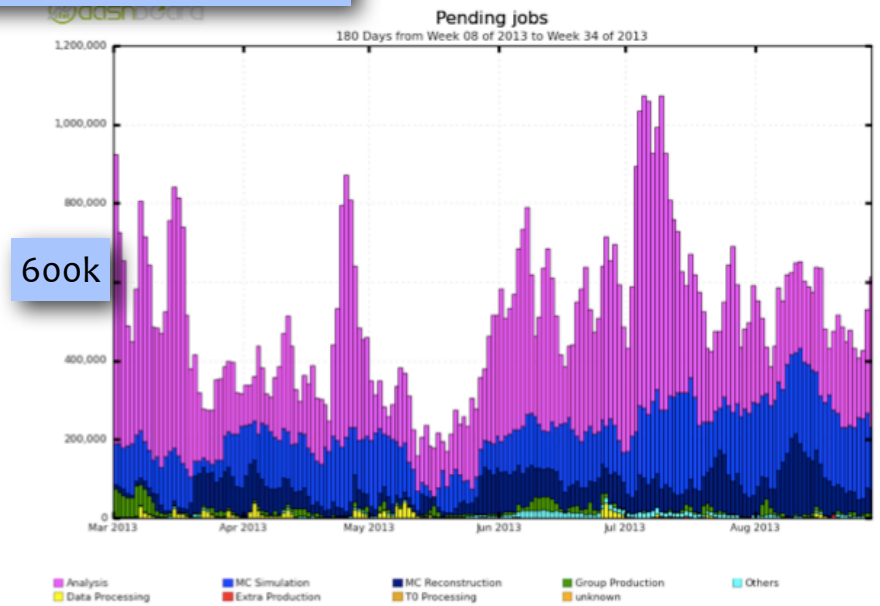
| | Location | Requested | Used |
|---|---|---|---|
| *CPU [kHS06]* | CERN | 111 | 111 |
| | Tier-1 | 316 | 435 |
| | Tier-2 | 360 | 713 |
| *Disk [PB]* | CERN | 9.6 | 8.9 |
| | Tier-1 | 35 [38] | 35 |
| | Tier-2 | 51 [52] | 48 |
| *Tape [PB]* | CERN | 25 | 29 (incl. 9 PB of ESD) |
| | Tier-1 | 42 | 33 |

- ATLAS has utilized the computing resources in its Tiers well in the last year: *many thanks to sites for resources and excellent operating!*
  - **We manage to provide a timely throughput of analyses to meet the physics requirements.**
- An ongoing effort in software development to optimise the resource utilization by reducing the CPU consumption, event sizes - for Run-2...

Tiers CPU / running jobs




Tiers CPU / pending jobs



# The Challenges of Run-2

## Constraints of 'flat budget'

- Both for hardware and for operation and development
- Hardware increase from Moore's law gain, estimated at factors of 1.2/year for CPU and 1.15/year for disk

## Data from Run-1

- Proper data preservation and integration with Run-2 data analysis

## New CPU architectues : less memory / core

## LHC operation

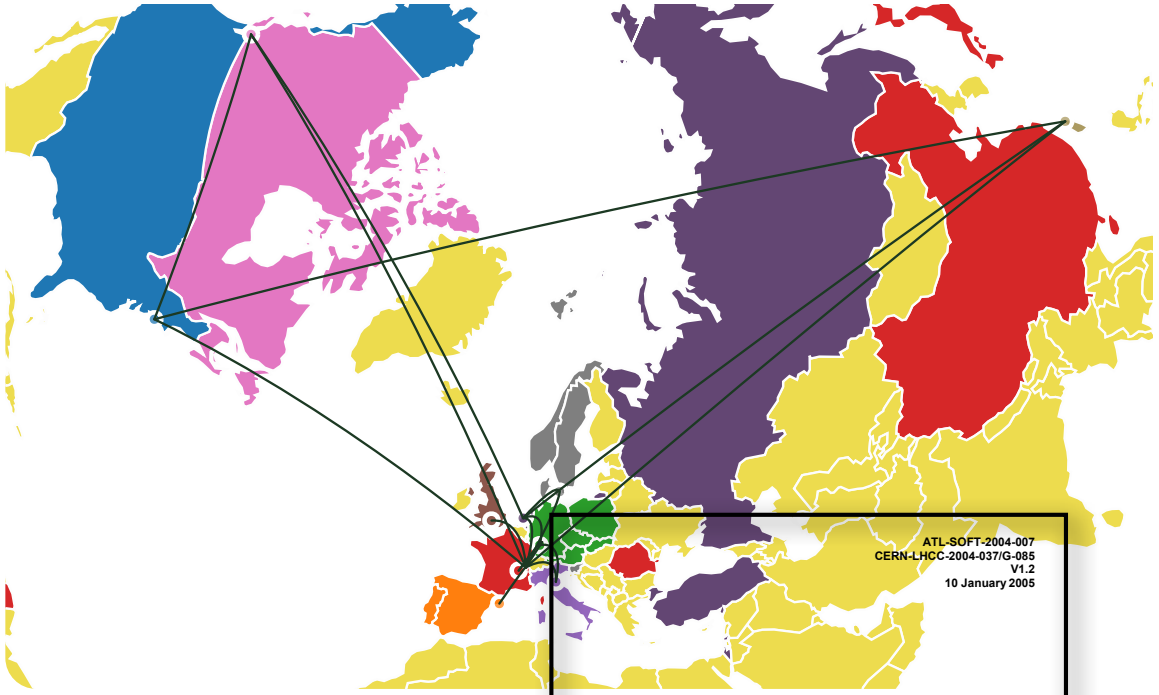
- HLT rate 1 kHz
- Pile-up up above 30
- 25 ns bunch spacing
- Centre-of-mass energy x ~2

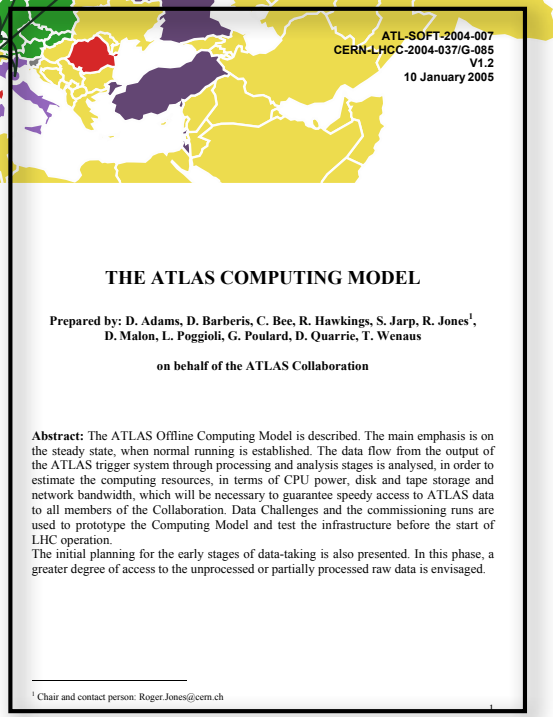
## ~new detector

- Needs to be incorporated in simulation and reconstruction.

# Initial Computing Model (2005)

- Derived from MONARC ('99) model
- CERN-T0 the center
- 10 T1s connected by dedicated 10Gb/s links (LHCOPN)
- O(100) T2s each attached to a T1
- The data flows along the hierarchy
- Insufficient networking assumed
- Hierarchy of functionality and capability





ATL-SOFT-2004-007
CERN-LHCC-2004-037/G-085
V1.2
10 January 2005

THE ATLAS COMPUTING MODEL

Prepared by: D. Adams, D. Barberis, C. Bee, R. Hawkings, S. Jarp, R. Jones[1], D. Malon, L. Poggioli, G. Poulard, D. Quarrie, T. Wenaus

on behalf of the ATLAS Collaboration

**Abstract:** The ATLAS Offline Computing Model is described. The main emphasis is on the steady state, when normal running is established. The data flow from the output of the ATLAS trigger system through processing and analysis stages is analysed, in order to estimate the computing resources, in terms of CPU power, disk and tape storage and network bandwidth, which will be necessary to guarantee speedy access to ATLAS data to all members of the Collaboration. Data Challenges and the commissioning runs are used to prototype the Computing Model and test the infrastructure before the start of LHC operation.
The initial planning for the early stages of data-taking is also presented. In this phase, a greater degree of access to the unprocessed or partially processed raw data is envisaged.

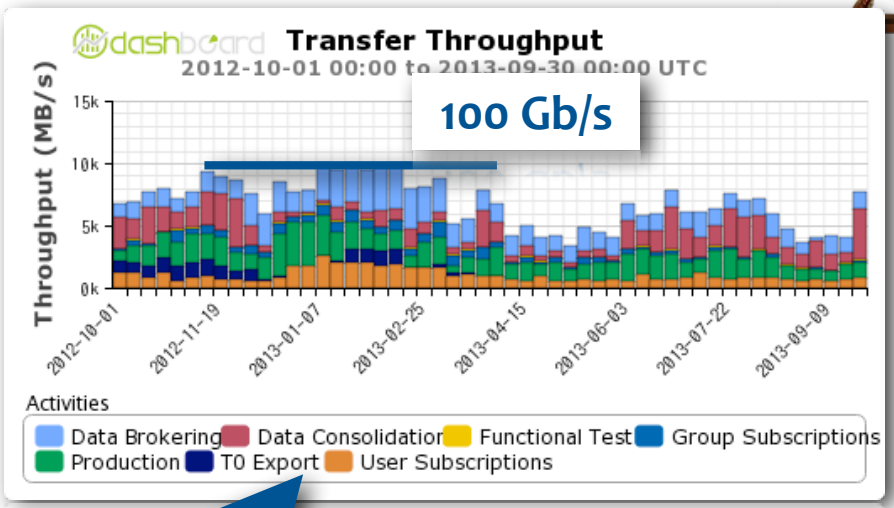
[1] Chair and contact person: Roger.Jones@cern.ch

http://monarc.web.cern.ch/MONARC/

Models of Networked Analysis at Regional Centres for LHC Experiments

# From 2010 to 2013: some of the many changes

- Hide grid complexity from users, **simplifications**, less middleware dependence
- **Caching** opposed to centralized DB
  - Conditions data access from any site, not only at T1s (Frontier)
  - No more need to pre-install software releases at sites (CVMFS)
- **Dynamic** data placement and deletion based on **popularity**
  - Better usage of disk space
  - Reduced job waiting times
- T2→N-T1s & T2↔T2 exchanges

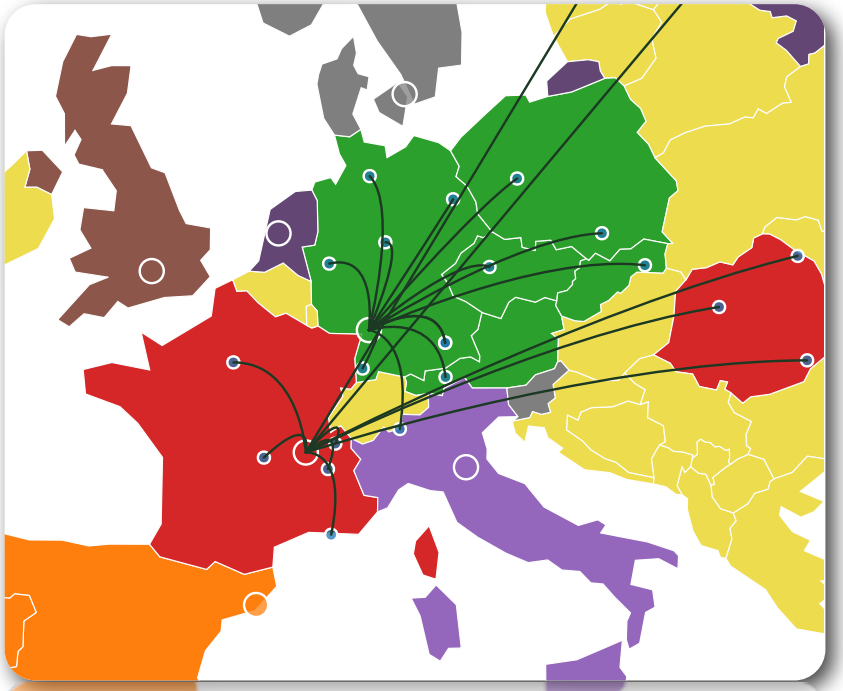



**Network performing over expectations**

**2011 LHCONE :**
**Dedicated network between (some) WLCG sites**

# 2010

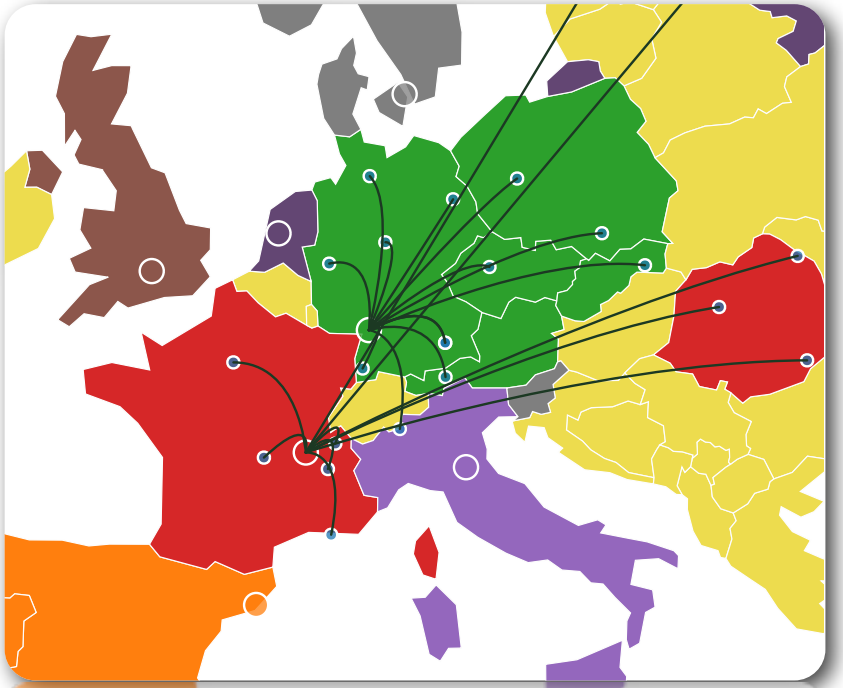
**Planned data distribution**
**Jobs go to data**
**Multi-hop data flows**
**Poor T2 networking across regions**



**~20 AOD copies distributed worldwide**

## 2010

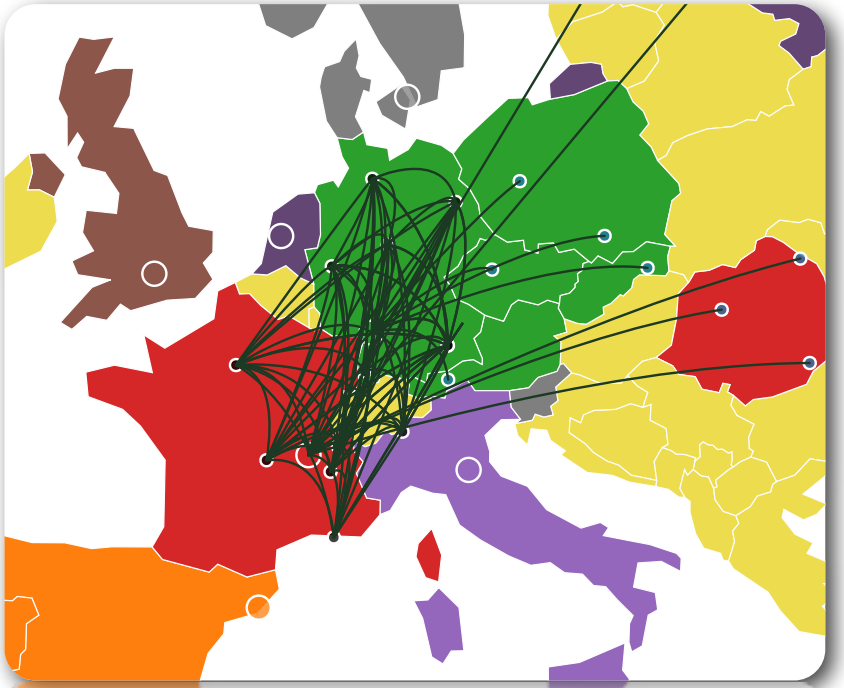
Planned data distribution
Jobs go to data
Multi-hop data flows
Poor T2 networking across regions



~20 AOD copies distributed worldwide

## 2013

Planned *& dynamic* distribution data
Jobs go to data *& data to free sites*
*Direct* data flows *for most of T2s*
*Many T2s connected to 10Gb/s link*



4 AOD copies distributed worldwide

# Networking Potential

- Networking is the one item that will most probably continue its progress & evolution further..
- ... and reduce the cost fastest.
  - In terms of bandwidth increase.
  - In terms of new technologies (NaaS - Network as a (virtual) Service ? )

ESnet

Software-Defined Networks (SDN):
Bridging the application-network divide

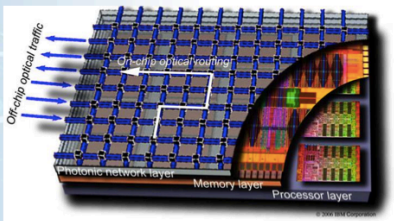
Inder Monga

Chief Technologist and Area Lead,
Engineering, Research and Software development

CHEP 2013

A fun peek into the future…just imagine

ESnet




With silicon photonics integration, each chip will have a network interface

That implies each chip could be network addressable

If so, we could design servers without needing NIC cards – no difference between communication within the motherboard or outside.

With HEP applications like FAX, file systems or memory can be mounted remotely to my chip while 'streaming data for analysis.'

With SDN, can effectively route IP and non-IP protocols (like ROCE)

**SDN could revolutionize how computing is done, are we ready for that?**

US LHCNet

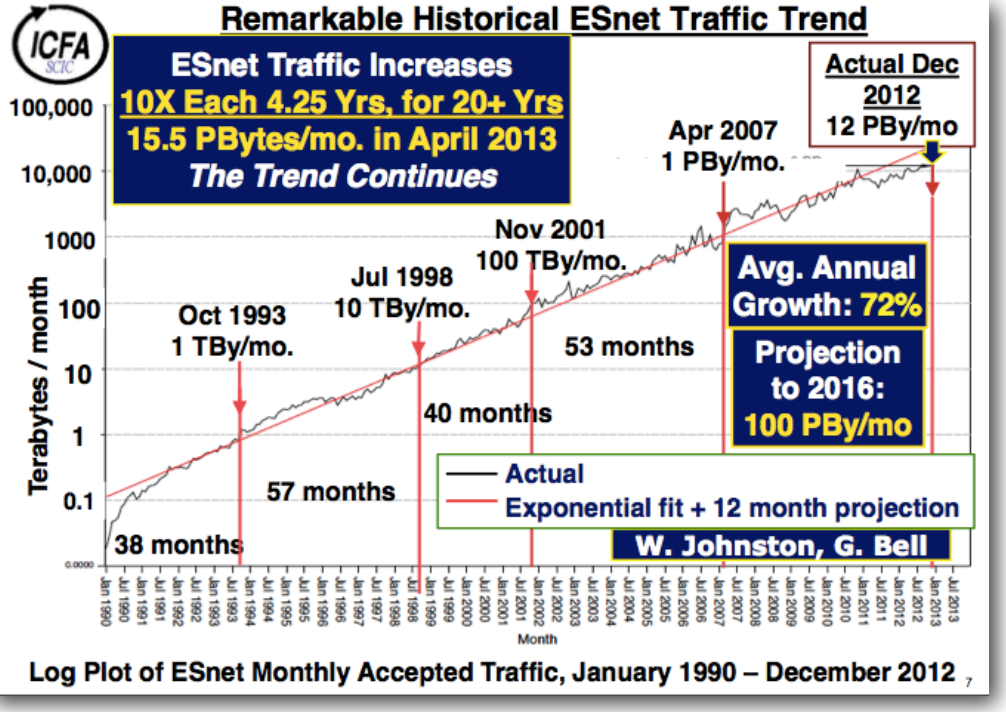
**Advanced Networking for HEP, Research and Education in the LHC Era**

**Harvey B Newman**
**California Institute of Technology**

**Remarkable Historical ESnet Traffic Trend**

ICFA SCIC

**ESnet Traffic Increases**
**10X Each 4.25 Yrs, for 20+ Yrs**
**15.5 PBytes/mo. in April 2013**
***The Trend Continues***




**Log Plot of ESnet Monthly Accepted Traffic, January 1990 – December 2012**

https://indico.cern.ch/sessionDisplay.py?sessionId=0&confId=214784#20131017

# How can we profit from network evolution?

- Two interesting ATLAS initiatives ongoing:
  - **data federations (FAX - xrootd federation, http federation)** - remote file access over WAN.
  - **Event Service** - passing single events for processing from/to storage.

Computation Institute

## FAX Commissioning Issues

Rob Gardner

Computation and Enrico Fermi Institutes

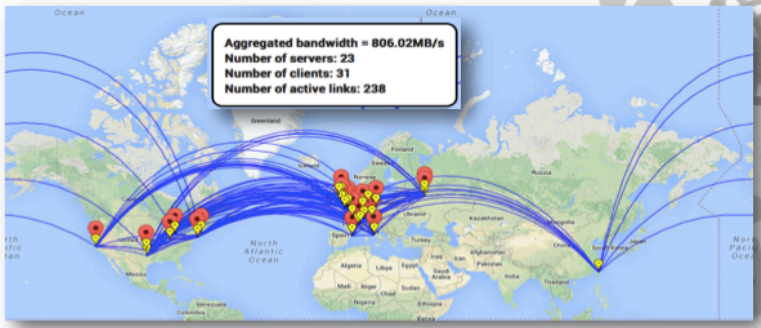
Univers

ATLAS S

Octobe

## A deployed federation

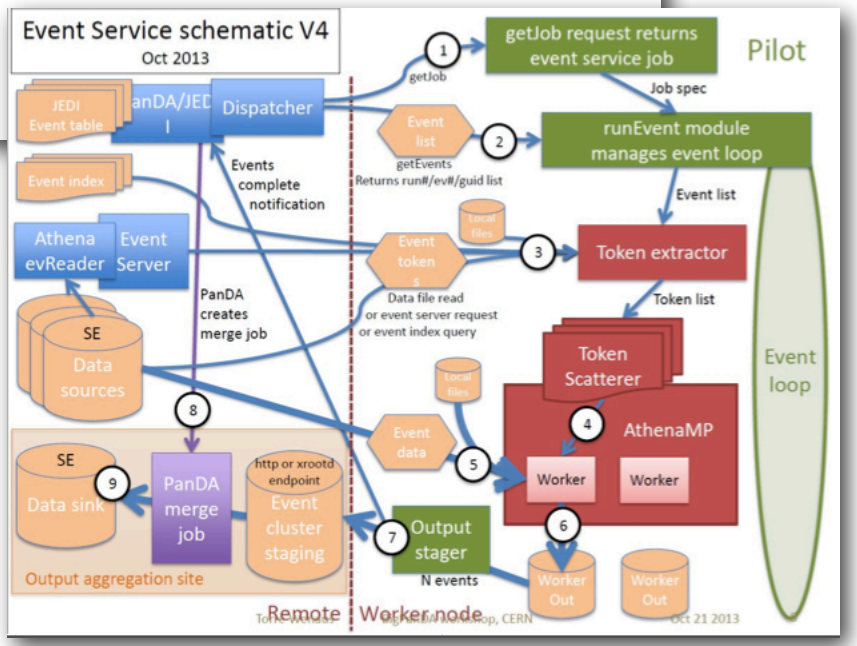
Provides a global namespace
Unifies dCache, DPM, Lustre/GPFS, Xrootd storage backends
Xrootd an efficient protocol for WAN access
Fallback use case in production in some queues
Regional redirection network provides lookup scalability
Accumulates WAN IO data for a brokering cost matrix




## Event Service

Vakho Tsulaia
LBNL

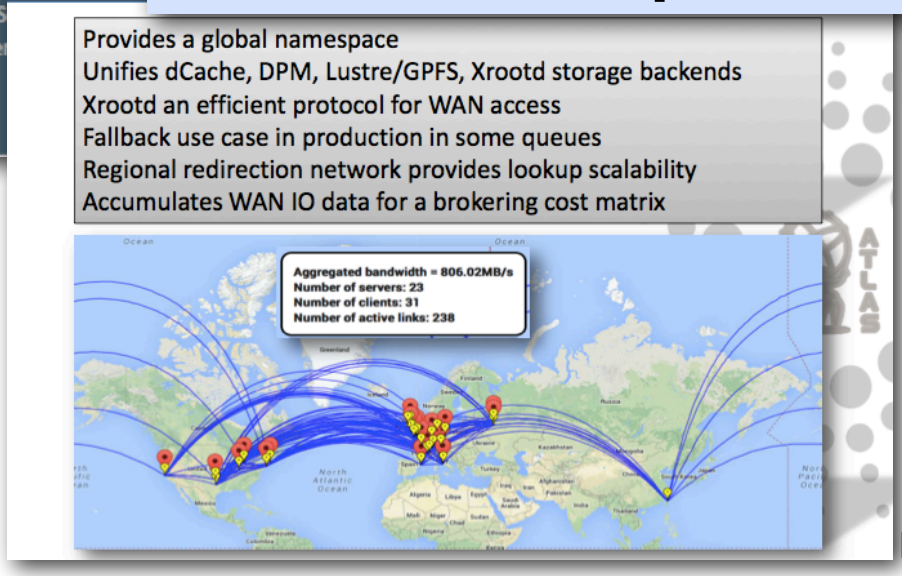
*For P. Calafiura, K. De, T. Maeno, D. Malon, P. Nilsson, P. Van Gemmeren, R. Vitillo, T. Wenaus, and other contributors*

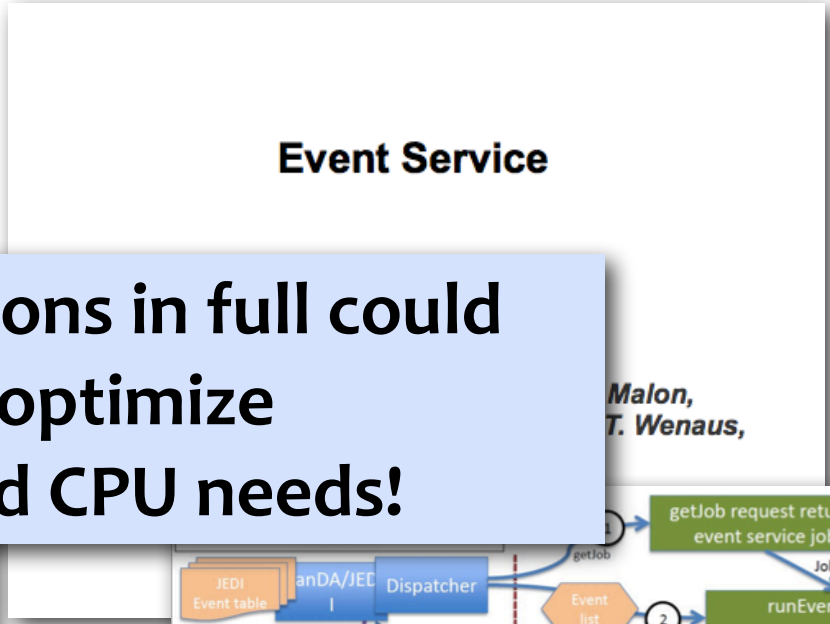


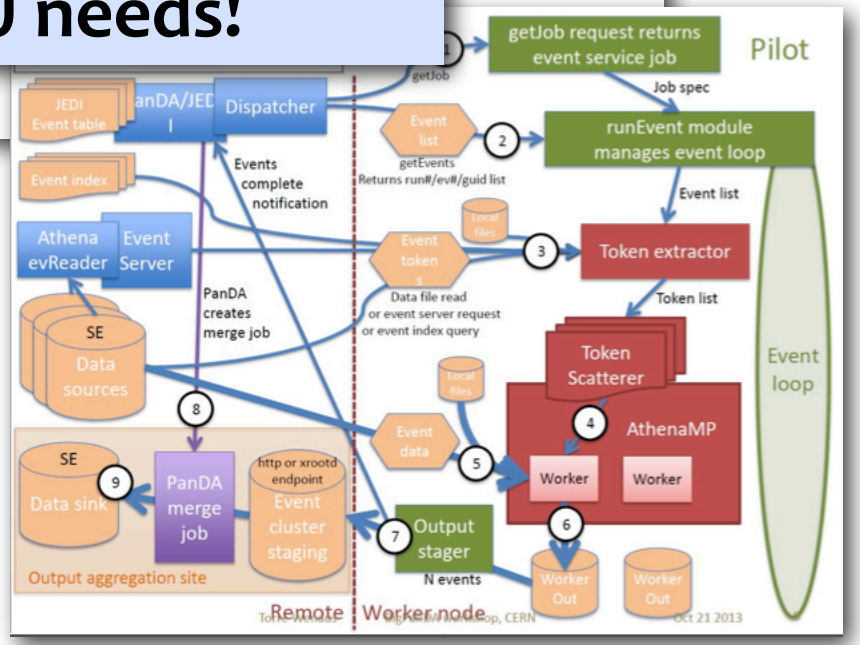
# How can we profit from network evolution?

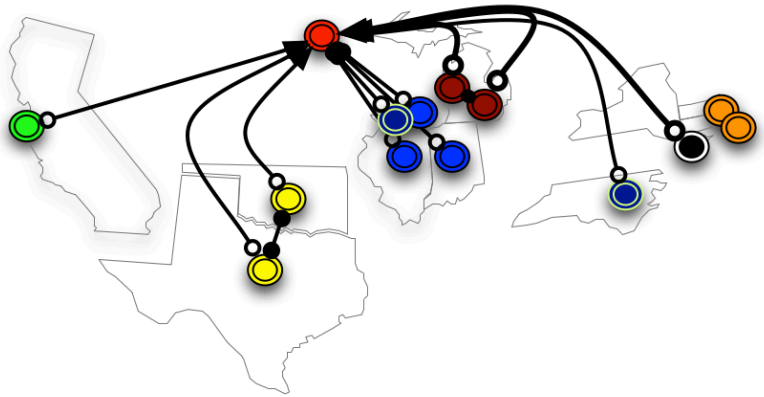
- Two interesting ATLAS initiatives ongoing:
  - **data federations (FAX - xrootd federation, http federation)** - remote file access over WAN.
  - **Event Service** - passing single events for processing from/to storage.

**Adopting such solutions in full could substantially optimize our disk space and CPU needs!**

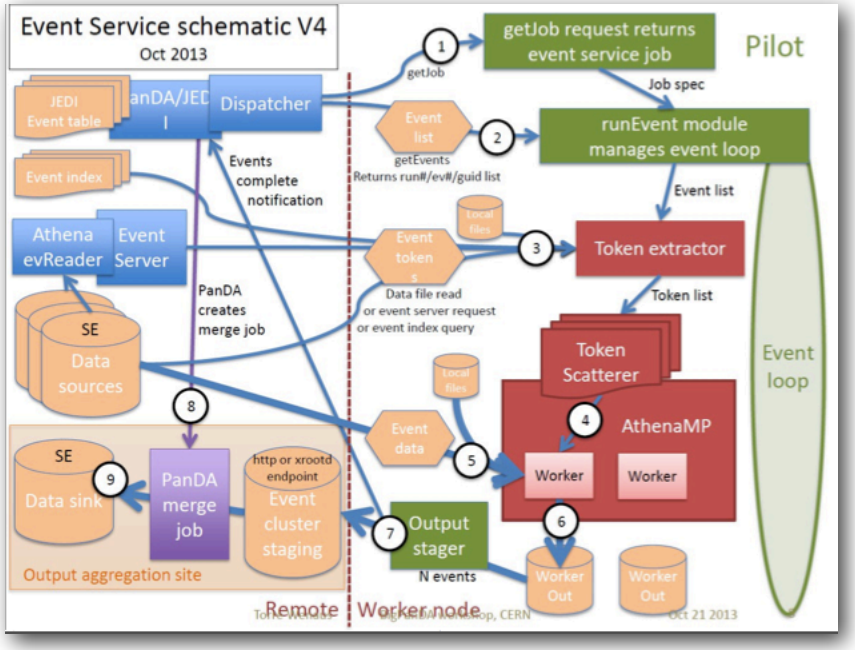





# DISTRIBUTED STORAGE / REMOTE ACCESS

- ★ Jobs access data on shared storage resources via WAN
- ★ For a better usage of storage resources (disk prices!)
- ★ Bandwidth and stability needed
- ★ FAX (Federating ATLAS data stores using Xrootd) demonstrator, job fail-over in case of access failure for first implementation
- ★ http protocol also considered

# Event service



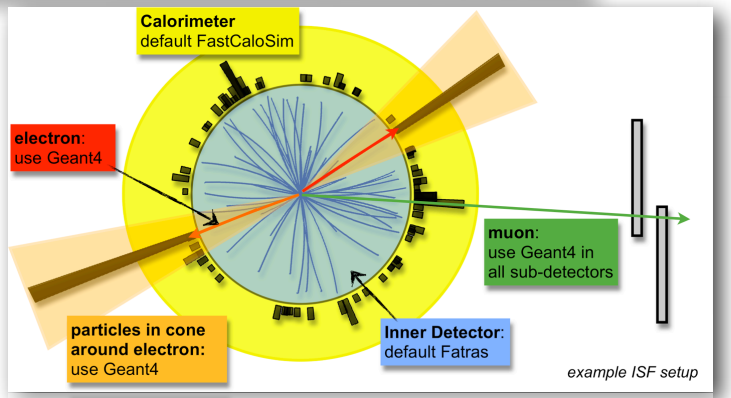

- In development : software and distributed computing effort
- Feed Virtual Machines with short jobs (simulate one single event)
- Usages :
  - Backfilling of HPC centers
  - Opportunistic use of commercial clouds
  - Volunteer computing (ATLAS@home)

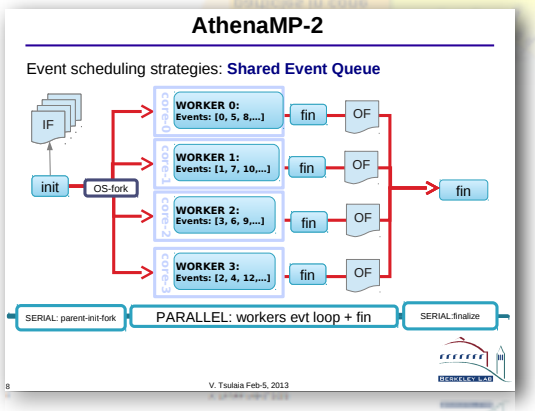
# Some Limitations of Current Model & Tools

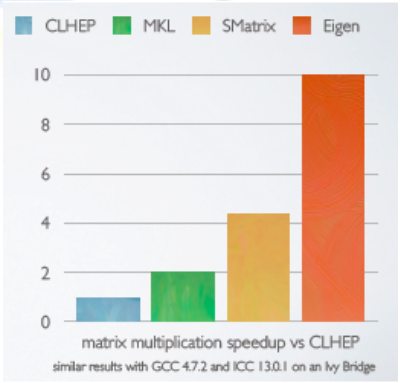
- Partitioning of resources:
  - Analysis vs Central Production,
  - T1s vs T2s.
- Difficulties of current Data Distribution Management & production systems to accommodate new use cases and technologies.
- Memory increase of MC pile-up digitization & reconstruction.
- Multitude of data format for analysis.
- Full reprocessing once a year.

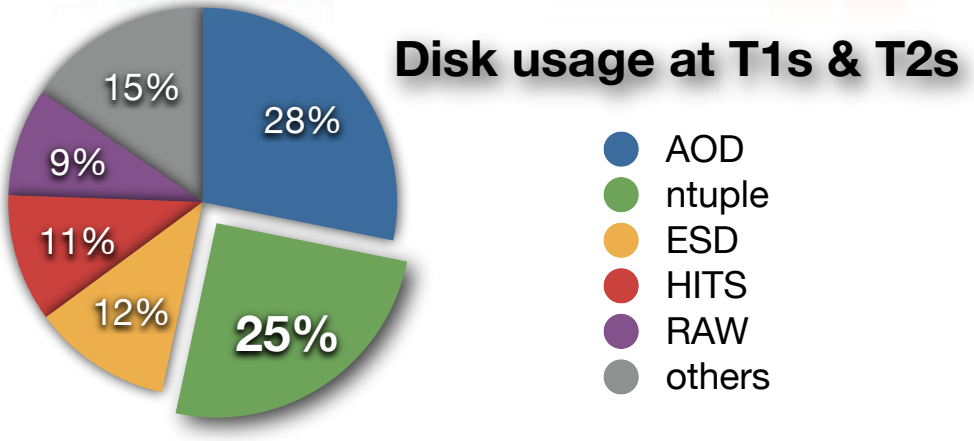
# Working towards Solutions

- **Simulation : CPU**
  - Integrated Software Framework.
- **Reconstruction : Memory & CPU**
  - Parallelism, code speedup.
  - MP solution to reduce memory footprint.
- **Analysis Model : multiplication of data formats**
  - Common analysis data format, xAOD.
  - Streamlining the analysis flow.



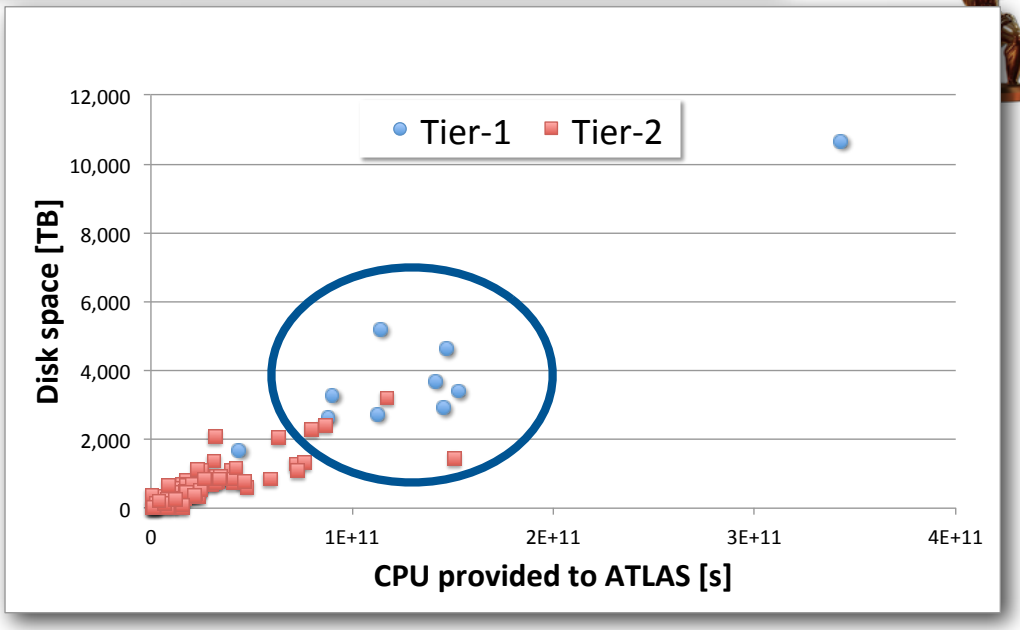



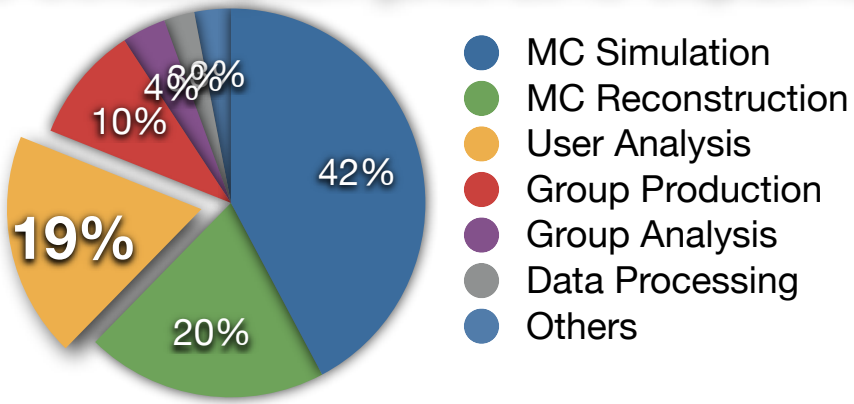
# Data Processing

- Optional extension of first pass processing from T0 to T1s in case of resource shortage at T0
- T1s and some T2s used for the most demanding workflows : **high memory and I/O intensive tasks**
- **Data reprocessing & MC reconstruction** also performed at some T2s
- Still one full reprocessing from RAW / year, but multiple AOD2AOD reprocessings / year
- Derivation Framework (train model) to centrally produce TB size data samples for analyses




**Some T2s are equivalent to T1s in term of disk storage & CPU power**

**CPU Consumption [Oct. 2012-Sep.2013]**



# Data Placement in Run-2

- Initially **2 copies** of analysis data formats (xAOD: one at T1s, one at T2s):
  - Already being implemented to gain disk space.
- **Non popular** data will be archived **to tape** at T1s:
  - Further refinement of the popularity monitoring and data placement:
    - In October we managed to recover 9% of disk space occupied by data never accessed over last 9 months.
    - Minimal number of copies on disk not guaranteed.
    - User access to data on tape granted through centralized tools.

- Investigate data access patterns to provide information to sites to optimize the site hardware configurations (and cost):
  - Low access/high access on disk (caching of popular data)
  - Low access/high access on tape (different tape technologies & cost)

# New production system: ProdSys-2

**PanDA+JEDI+DEfT**

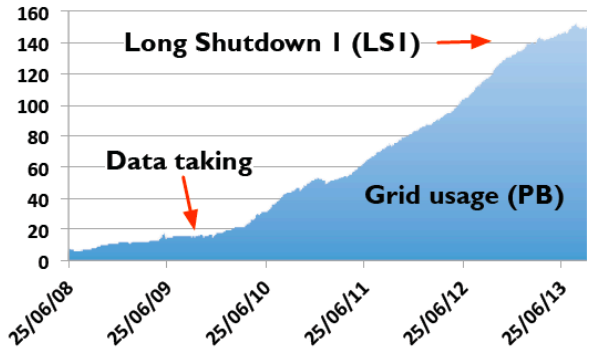
- Same engine for analysis and production
  - Currently analysis vs production shares managed by sites not by ATLAS.
  - Better reactivity to analysis load
- Data traffic minimized.
- Optimized job to resource matching.

# Data Distribution Management & databases

- New Data Distribution Management system: Rucio will replace DQ2
  - New scalable architecture.
  - File level functionality instead of dataset.
  - Built-in data replication policy for space and network optimization.
  - Multi-protocol (http,… ).
- Database infrastructure: simplification and streamlining.

The current DDM system Don Quijote 2 (DQ2) has demonstrated very large scale data management

- 150 PB
- 130 grid sites
- 800 users
- +40 PB per year
- +1 M files per year
- 0.6 M downloaded files per day




DQ2 will simply not continue to scale for LHC Run-2

http://rucio.cern.ch

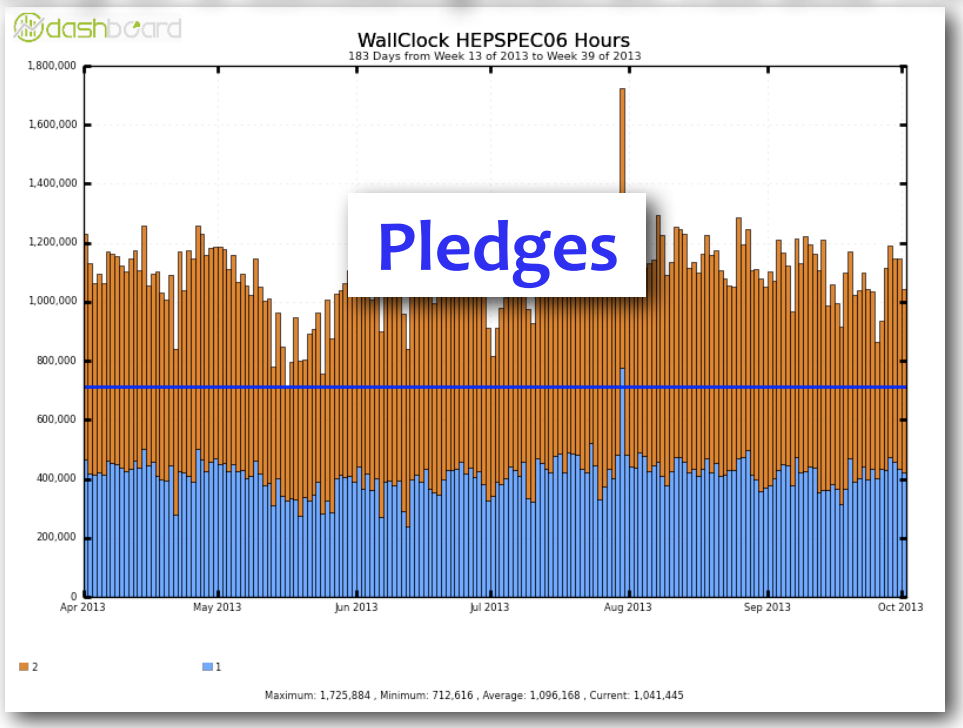
# Opportunistic resources

CPU consumption above pledges both at T1s and T2s

- **Additional solutions :**
  - HLT farm at P1.
  - Cloud computing.
  - Large HPC (High Performance Computing) centers.
  - Volunteer computing: ATLAS@home.

**Sites and Funding Agencies provide more than pledged resources (thank you!)**

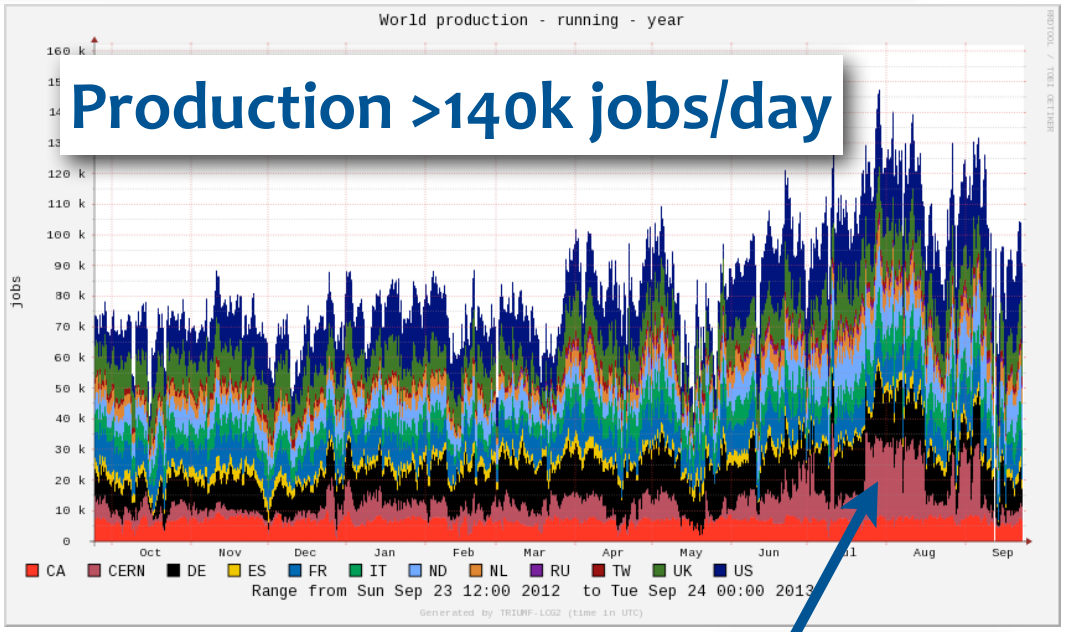
**2013 CPU consumption at T2s and T1s**



# HLT farm at P1

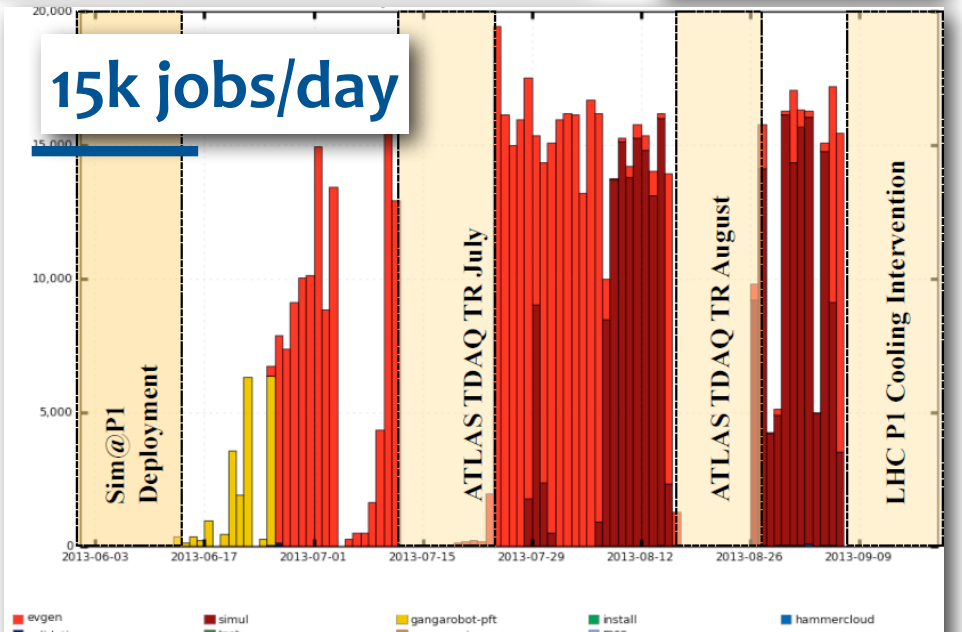
- HLT farm 'cloudified' mid-2013
  - Reached >15k concurrent simulation jobs
- Switch between trigger and simulation mode operational.
- **Availability in Run-2:**
  - for **MC production** during **shutdowns** or **LHC technical stops if/when no other TDAQ activities.**
  - ~30% over a Run-2 year?

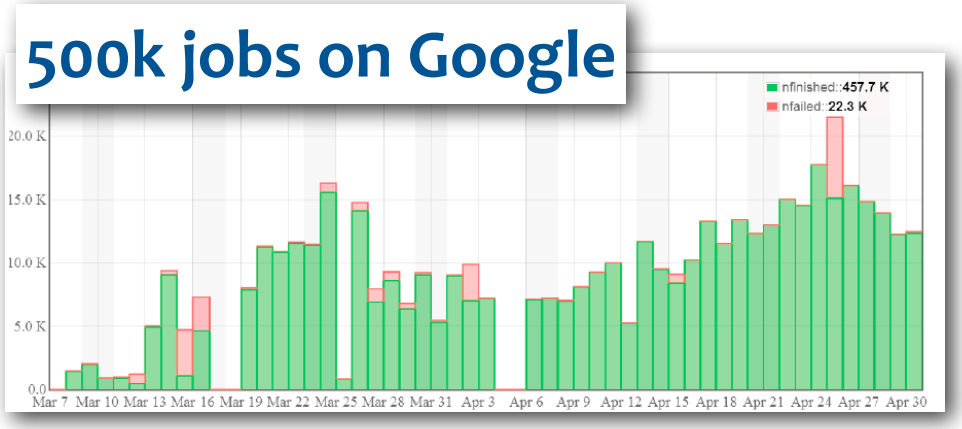
**BNL/T1+IT/SDC+ATLAS/TDAQ experts: many thanks!**

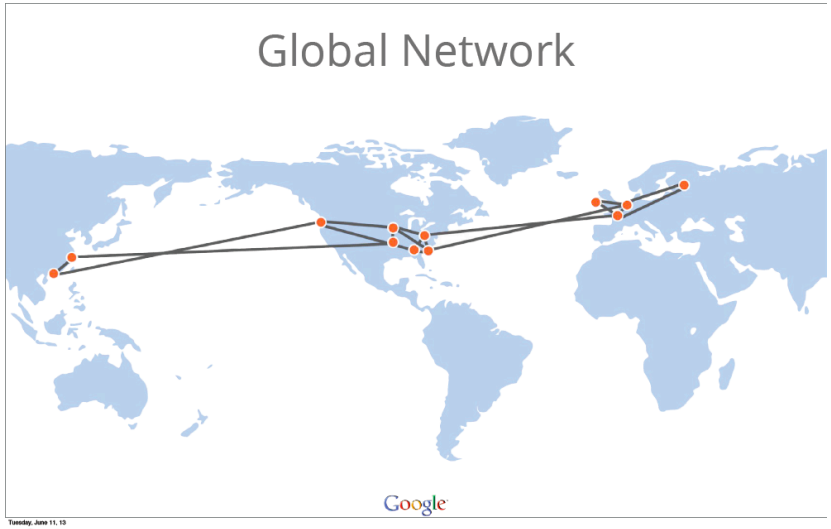



Equivalent to a T1 or a large T2

HLT farm






# Cloud computing

On going R&D on academic clouds and Amazon or Google (AUS,CA, US,...)
Issues with long jobs and I/O
Plan for use 'academic' clouds and opportunistic use of 'cheap' commercial is possible
Some cloud computing providers start to propose cost-competitive offers (with some limitations)

PRACE, the Partnership for Advanced Computing in Europe


**SuperMUC a PRACE Tier-0 centre :**
**155,000 Sandy Bridge cores, 2.8M HS06**

**WLCG 2013 T0/1/2 pledges ~2.0M HS06**

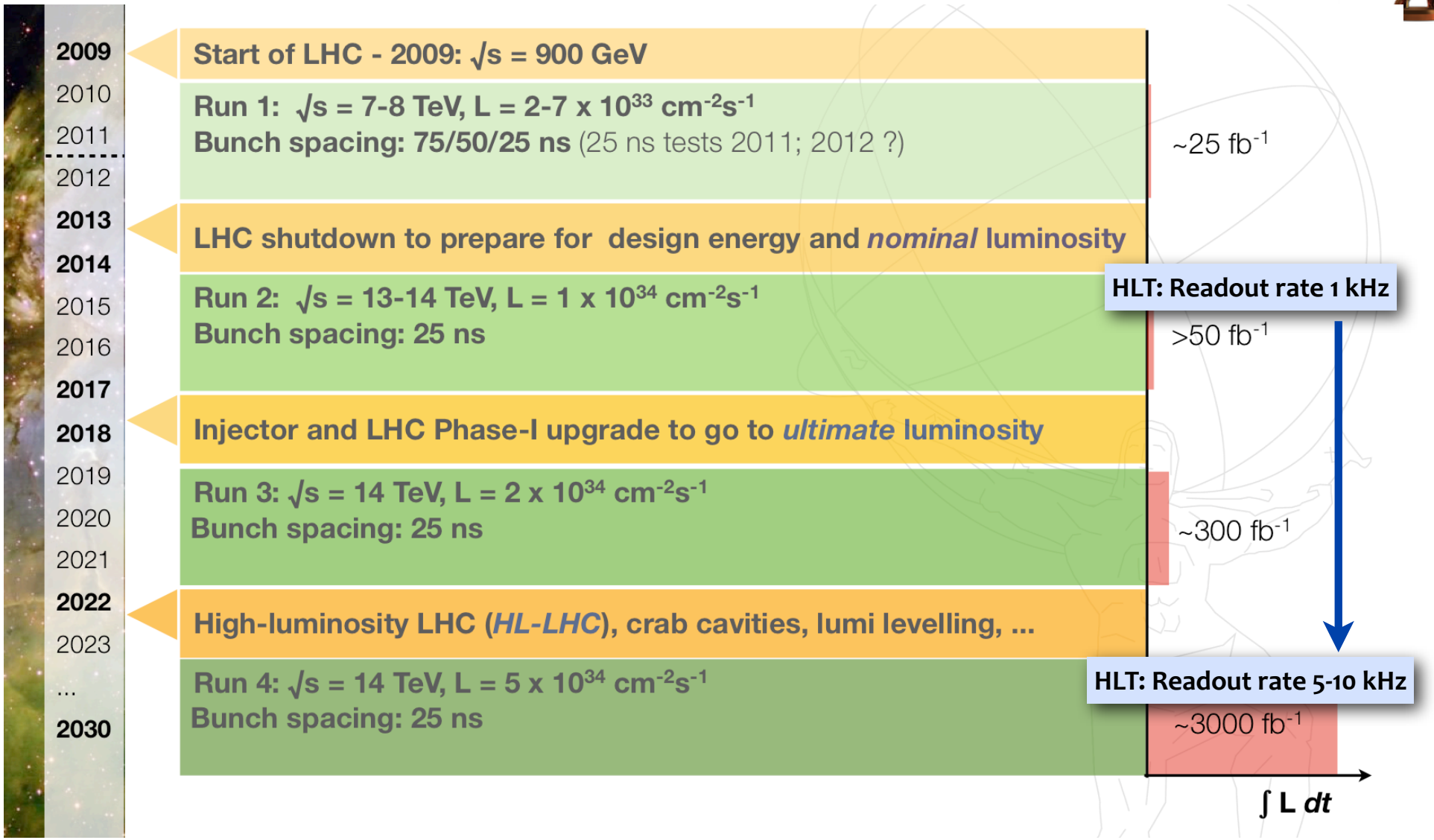
# HPC (High-Performance Computing) resources

- Large investments in many countries : from Peta to Exa scales initiatives[1]
- Latest competitive supercomputers are familiar Linux clusters
- Large number of spare CPU cycles are available at HPCs which are not used by 'standard' HPC applications
- Projects to use idle CPU cycles at HPC centers in US, China & DE
- Demonstrators working for simulation & event generation
- Difficult to use HPC centers for I/O intensive applications
- Outbound connectivity of HPC centers may also be an issue
- Some T2s plan to provide pledges resources on shared HPC facilities

***Might endanger traditional HEP computing budget***

[1] : http://www.eesi-project.eu/pages/menu/eesi-1/publications/investigation-of-hpc-initiatives.php

# LHC Upgrade Timeline - the Challenge to Computing Repeats periodically!



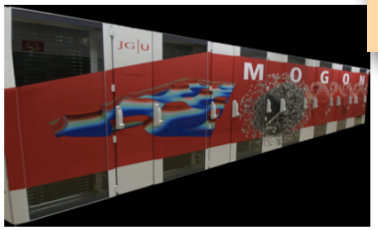
# Looking further in the future: in 10 years?

- The answer to this has strong political/financial components, which are hard to predict. Still..

R. Walker

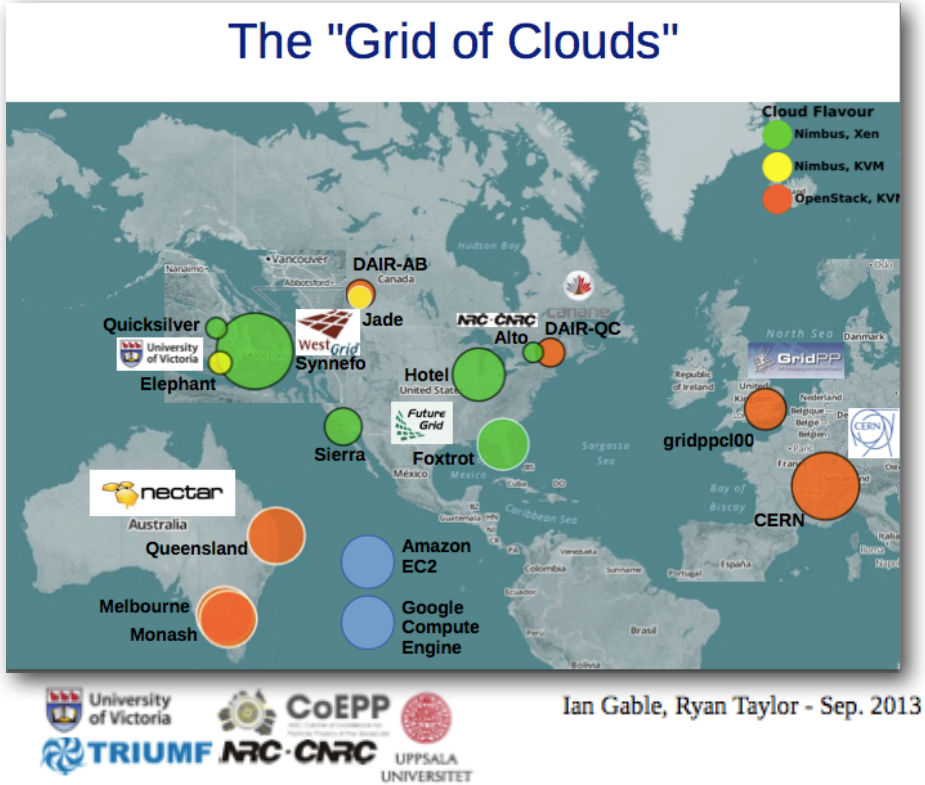
## HPC Resource Examples


- SuperMUC, Munich
  - 155,000 Sandy Bridge cores, 2.8M HS06
    - ATLAS 2013 T1/2 pledges ~ 730K HS06
  - Suse Enterprise Linux 11, 2GB/core
  - warm water cooling
    - 40°C inlet. 70°C outlet used to heat building
- Hydra, MPI, Munich
  - 'similar' cluster in spec and scale
    - due Summer 2013. 10k core integration system in place now
- MOGON, Mainz
  - 34k cores SL6



**Even today, our CPU capacities fit into one super-computing center. Will we get fractions of Super-Computer CPUs?**

**What will be the impact of IaaS (Cloud) techologies? Will we get cheques for commercial cloud use (again, super-computers..)?**




**The main item that does not have many solutions and gives a severe constraint is our data storage:**
**We need reliable and permanent storage under ATLAS control.**

**From another perspective, with the network evolution (and federated storage, event service..) 'local' becomes re-defined (again).**
**No need for local storage?**
**Consolidate to a fewer (20|10|..1?) main storage points? Cheaper?**

# Summary & Outlook

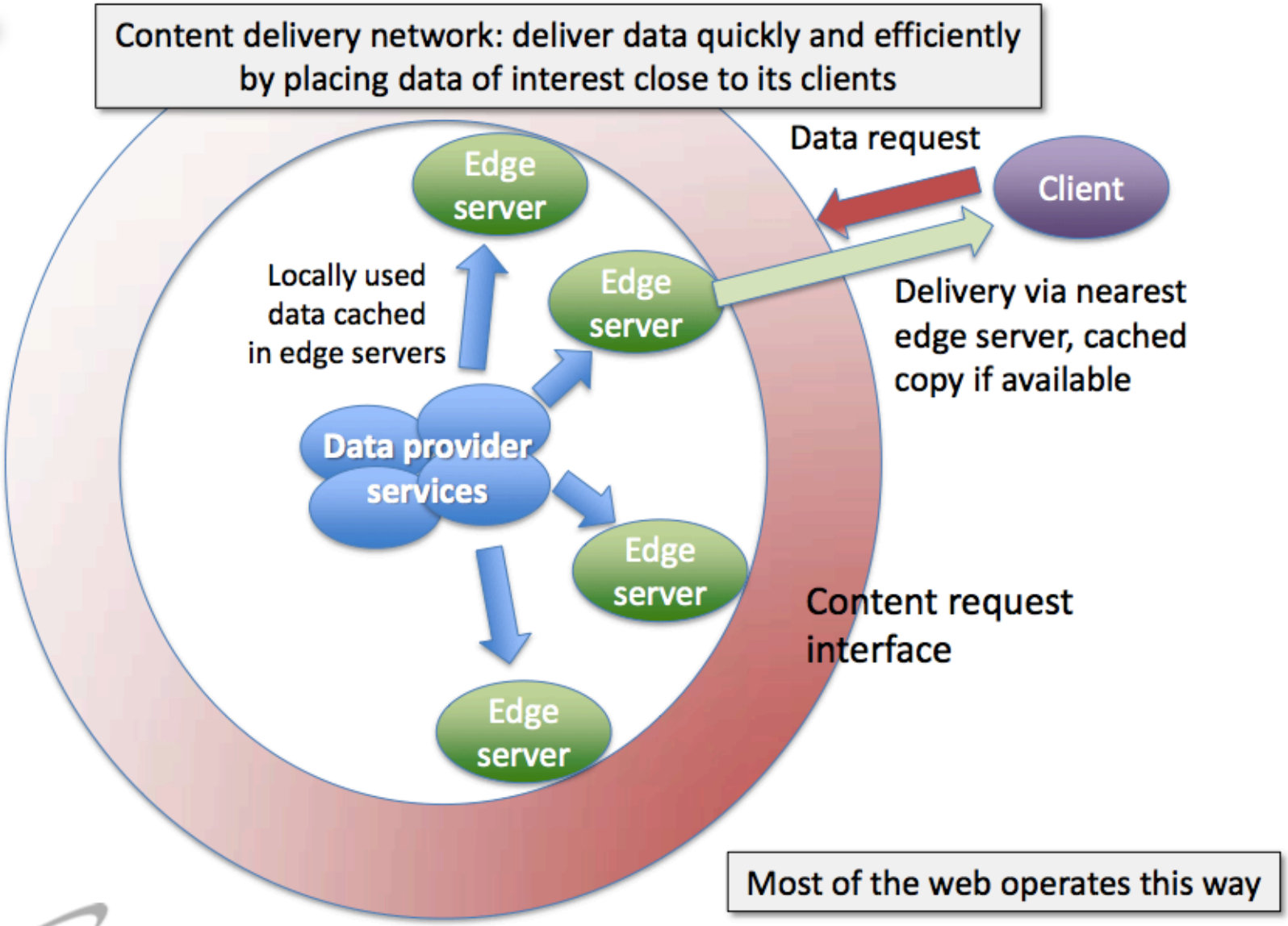
- A lot of experience acquired in 3 years of LHC data taking.
- Run2 will put high pressure on hardware and human resources.
- Solutions under development and manpower is needed.
- New computing model and its components will be tested during 2014 data challenge.
- LHC & ATLAS upgrades also mean resources for software & computing.

# Backup

# More on Future Data Acess…

## The Content Delivery Network Model

T. Wenaus



# More on Future Data Acess..

## The Content Delivery Network Model

A growing number of HEP services are designed to operate broadly on the CDN model

T. Wenaus

| Service | Implementation | In production |
|---|---|---|
| Frontier conditions DB | Central DB + web service cached by http proxies | ~10 years (CDF, CMS, ATLAS, …) |
| CERNVM File System (CVMFS) | Central file repo + web service cached by http proxies and accessible as local file system | Few years (LHC expts, OSG, …) |
| Xrootd based federated distributed storage | Global namespace with local xrootd acting much like an edge service for the federated store | Xrootd 10+ years<br>Federations ~now (CMS AAA, ATLAS FAX, …)<br>*See Brian's talk* |
| Event service | Requested events delivered to a client agnostic as to event origin (cache, remote file, on-demand generation) | ATLAS implementation coming in 2014 |
| Virtual data service | The ultimate event service backed by data provenance, regeneration infrastructure | Few years? |