Ian Bird

LHCC Open Session

CERN, 23rd September 2015

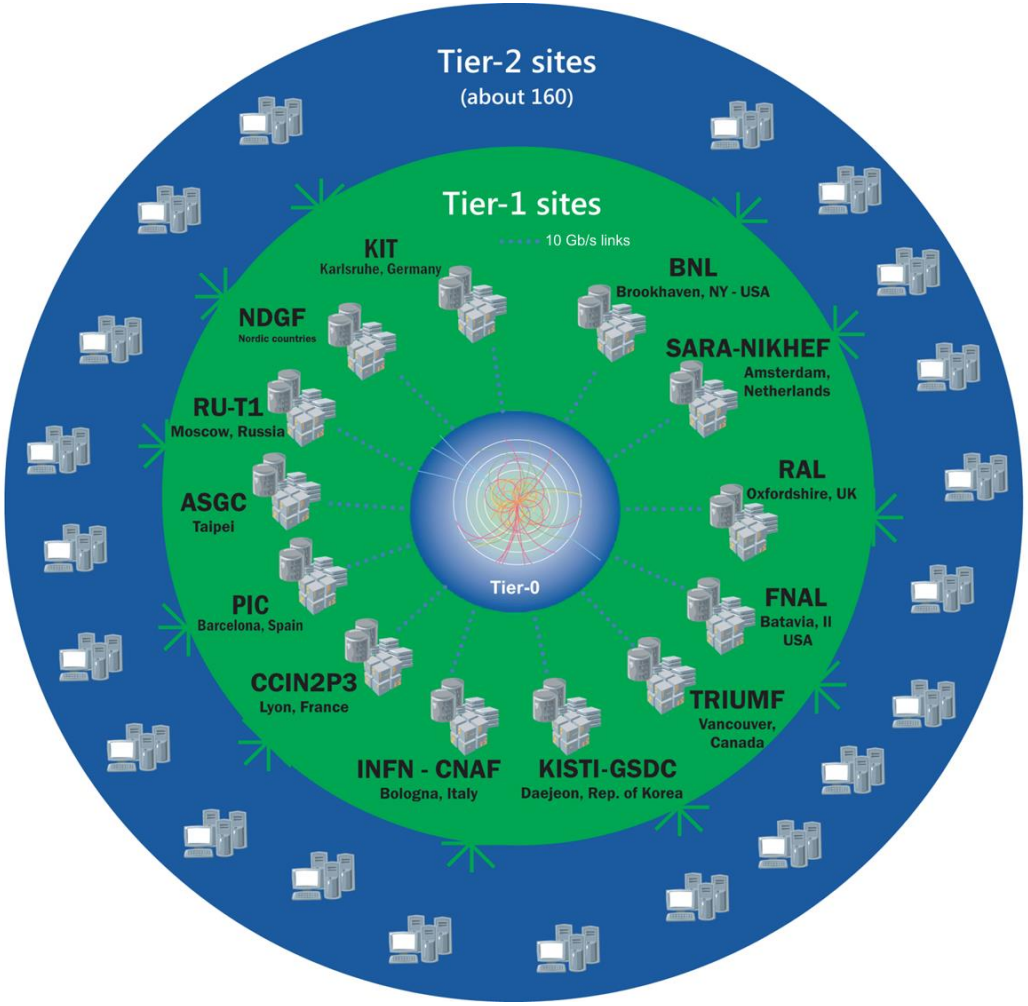
WLCG Status Report
Readiness for Run 2

The Worldwide LHC Computing Grid

Tier-0 (CERN): data recording, reconstruction and distribution

Tier-1: permanent storage, re-processing, analysis

Tier-2: Simulation, end-user analysis



nearly 170 sites, 40 countries

~350’000 cores

500 PB of storage

> 2 million jobs/day

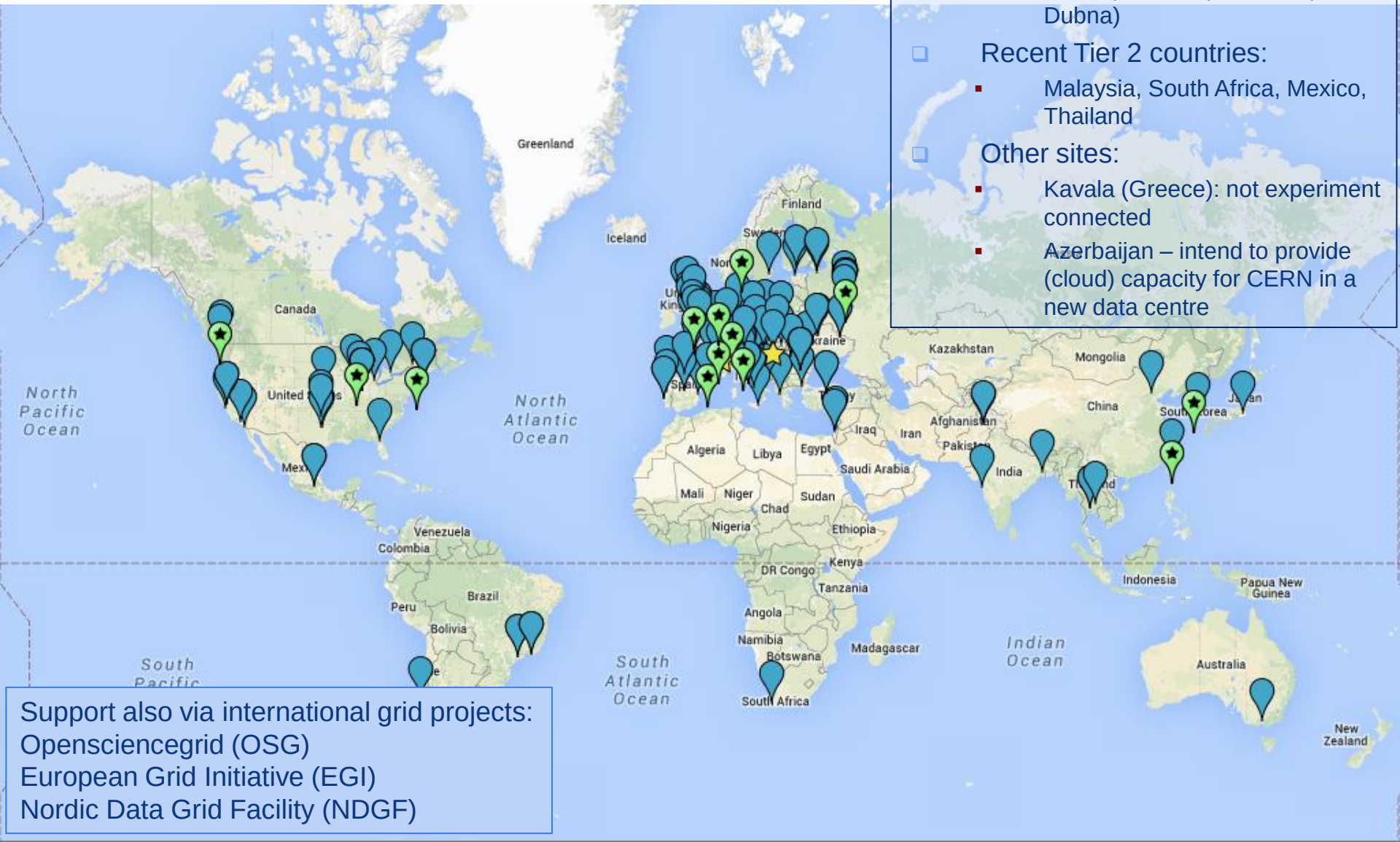
10-100 Gb links

WLCG:
An International collaboration to distribute and analyse LHC data

Integrates computer centres worldwide that provide computing and storage resource into a single infrastructure accessible by all LHC physicists

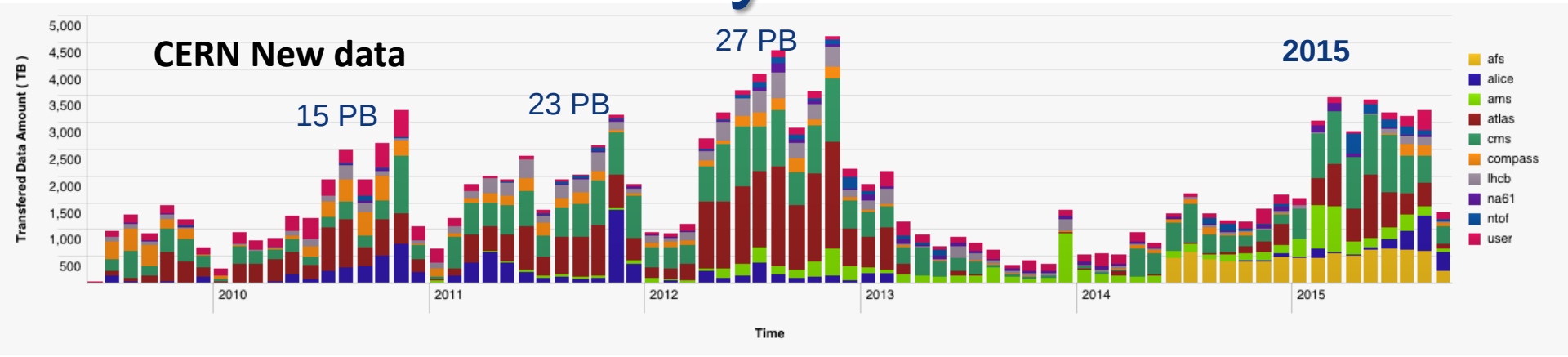
World-wide infrastructure

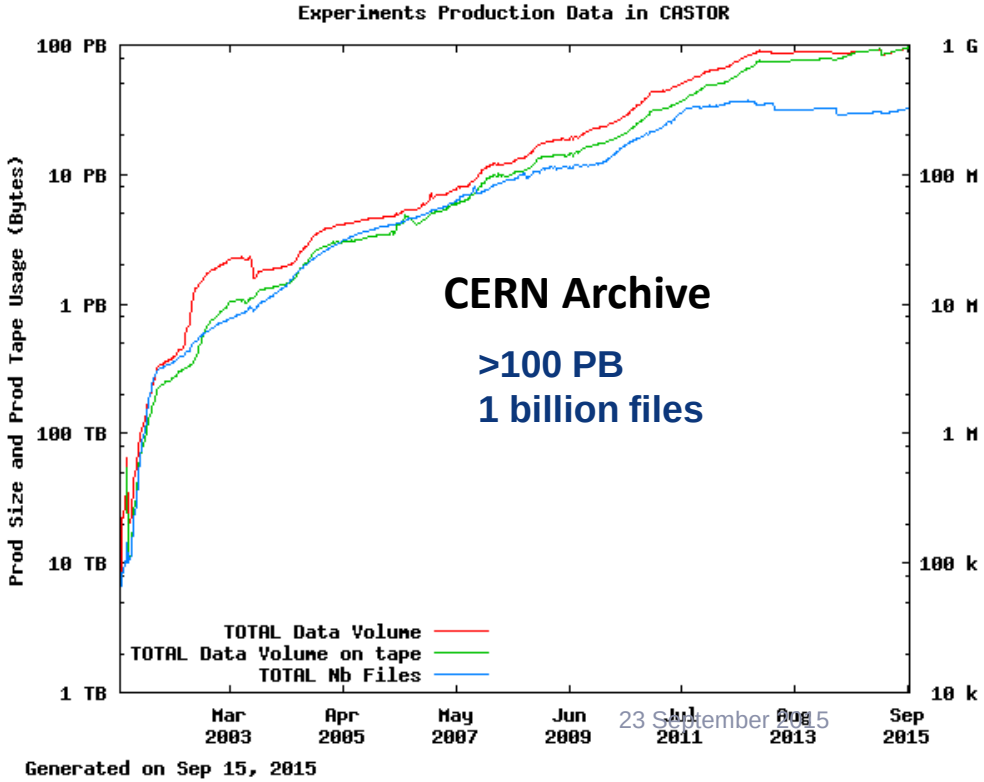
- New Tier 1s:
  - KISTI (S. Korea), Russia (NRC, Dubna)
- Recent Tier 2 countries:
  - Malaysia, South Africa, Mexico, Thailand
- Other sites:
  - Kavala (Greece): not experiment connected
  - Azerbaijan – intend to provide (cloud) capacity for CERN in a new data centre

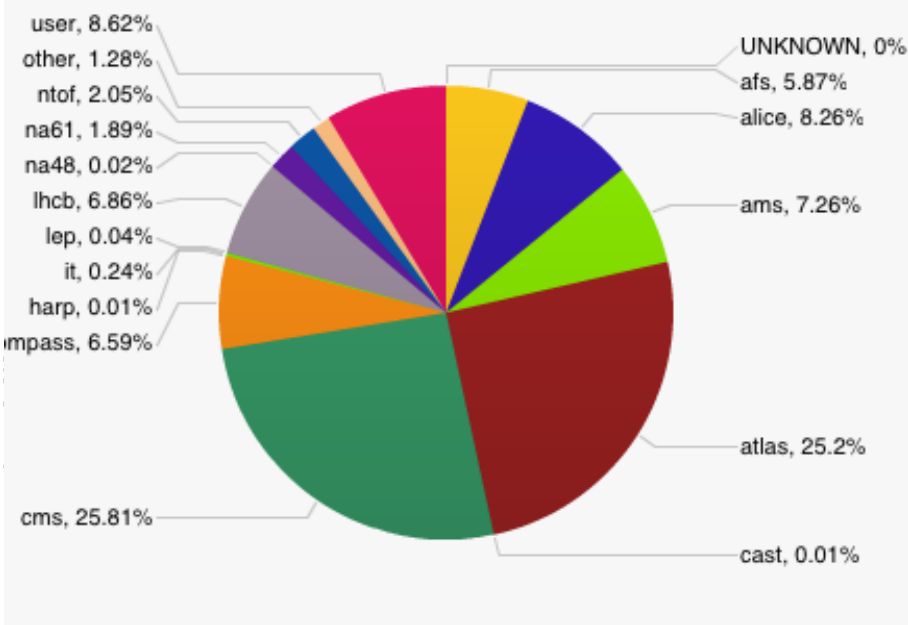


Support also via international grid projects:
Opensciencegrid (OSG)
European Grid Initiative (EGI)
Nordic Data Grid Facility (NDGF)

Scale of data today …

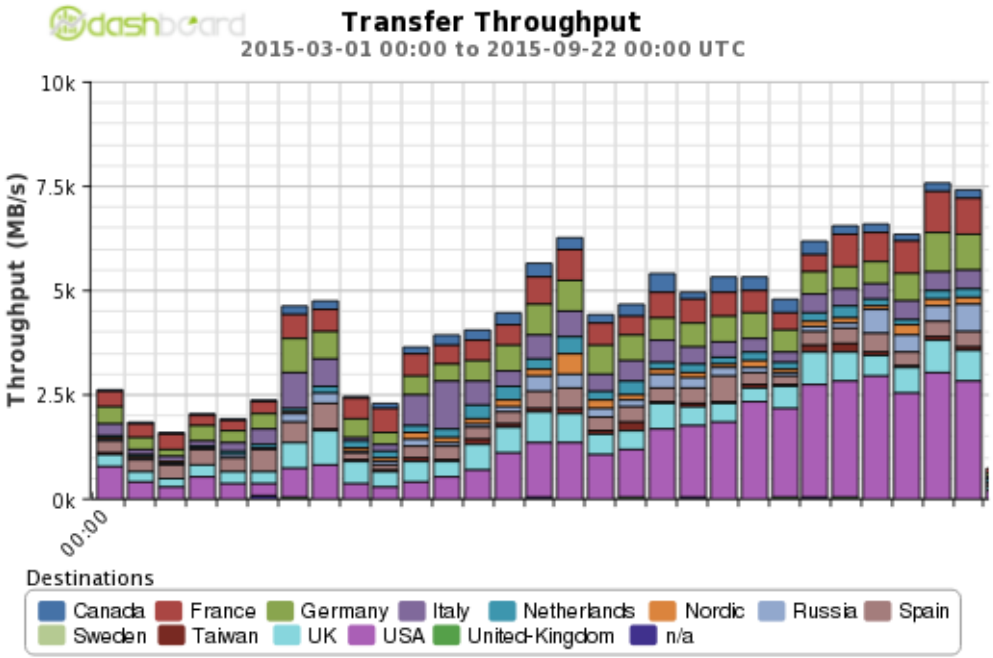


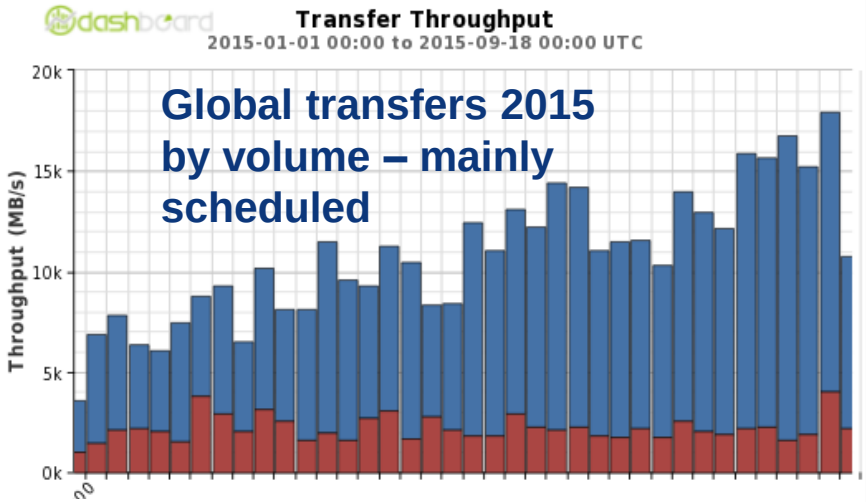


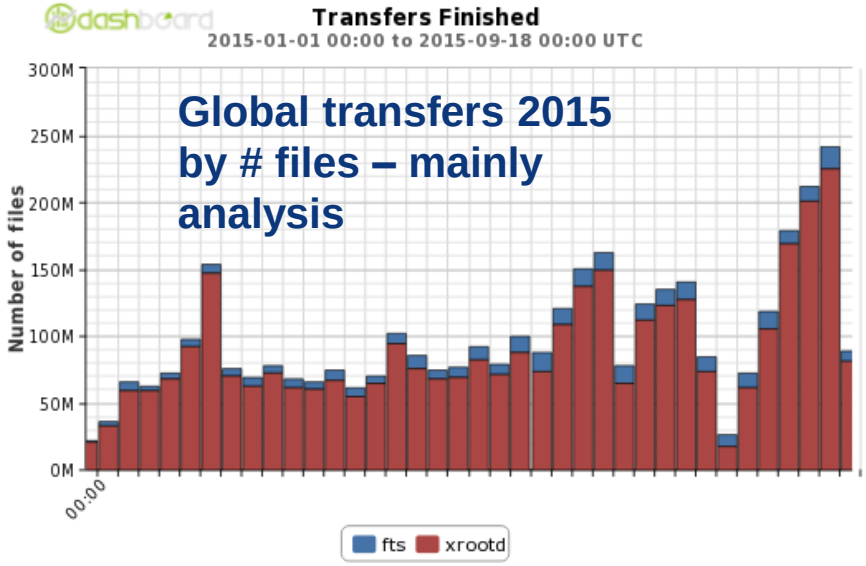


Data transfers

- CERN export rates driven (mostly) by LHC data export

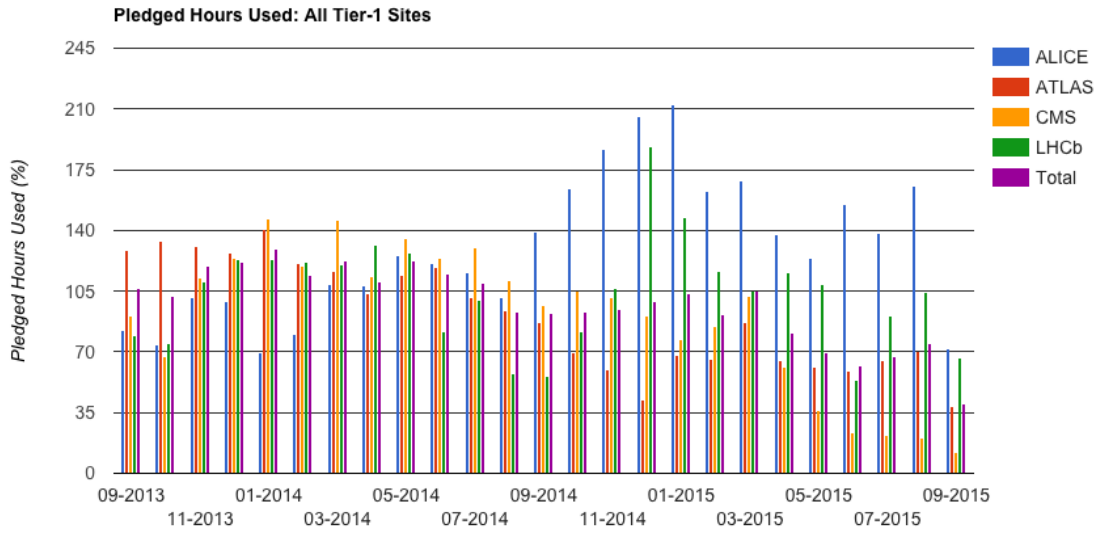


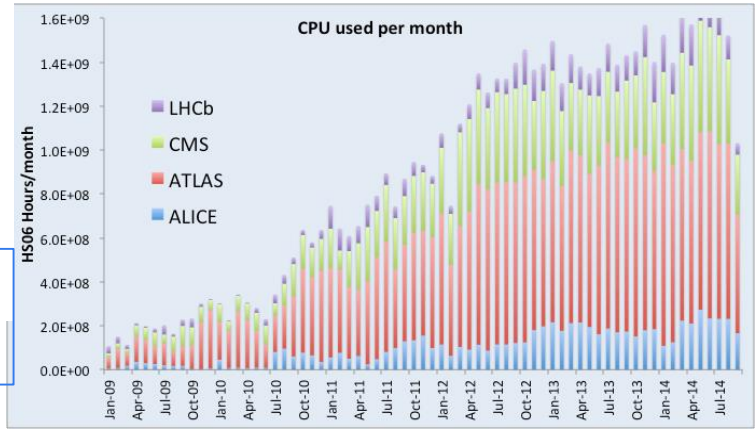




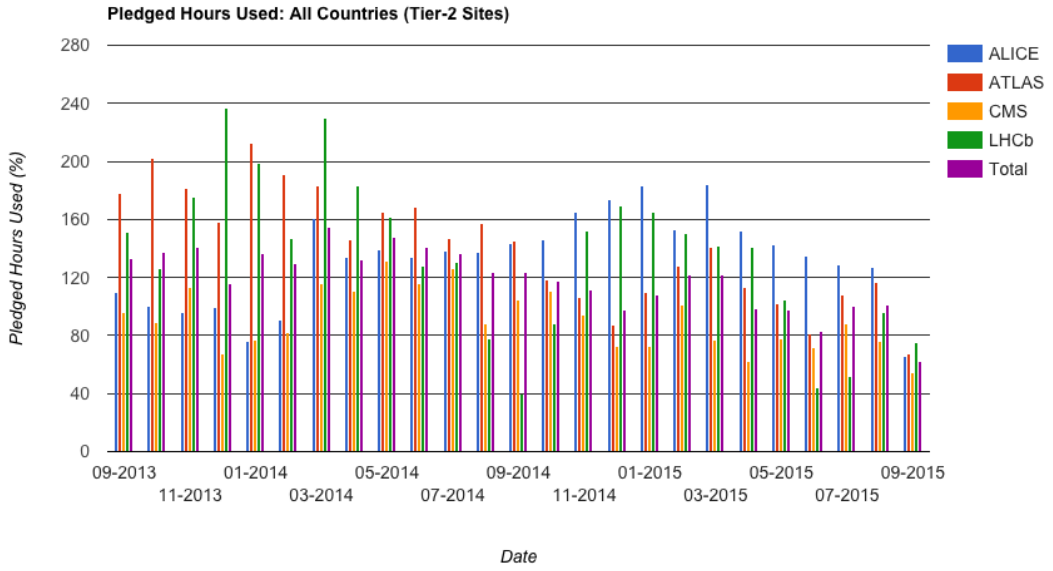
- Global transfer rates are always significant (12-15 GB/s) – permanent on-going workloads

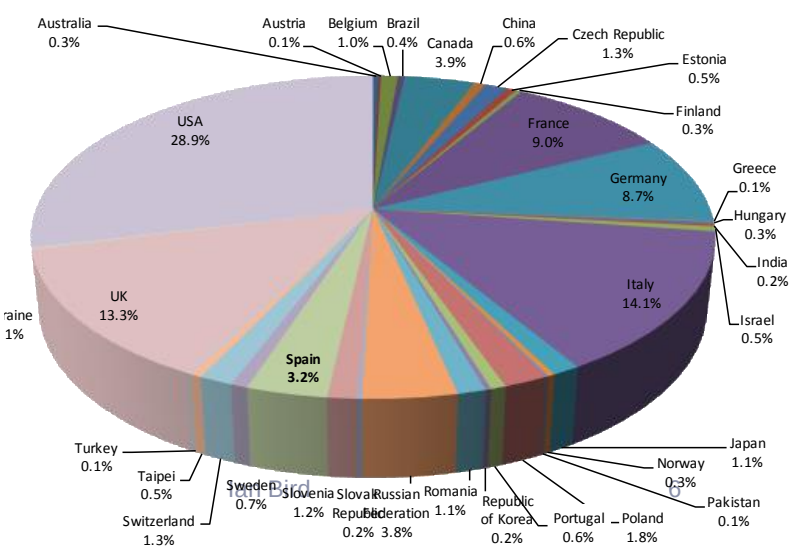
Resource usage





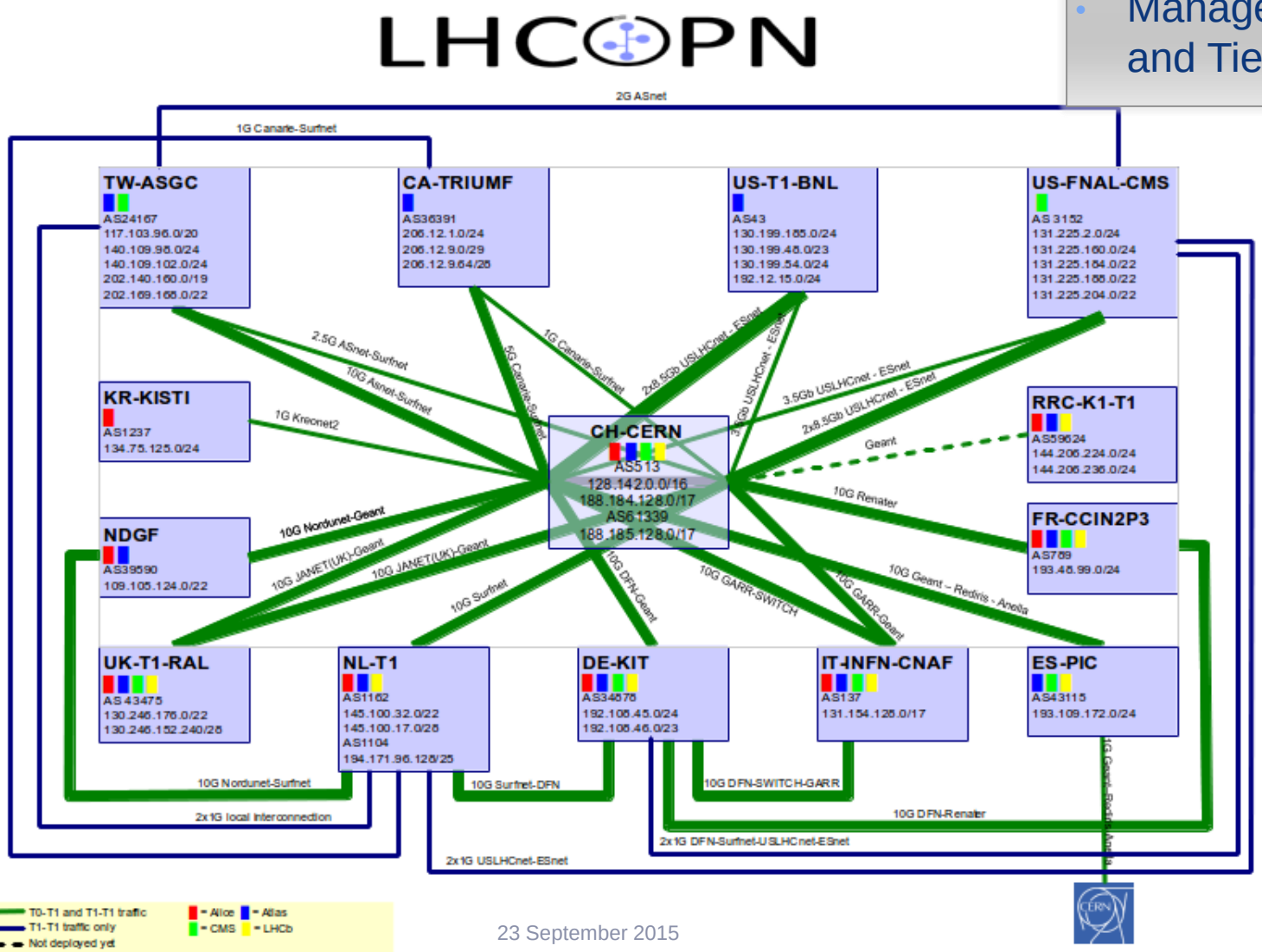
Experiments are able to use resources (significantly) exceeding those formally pledged



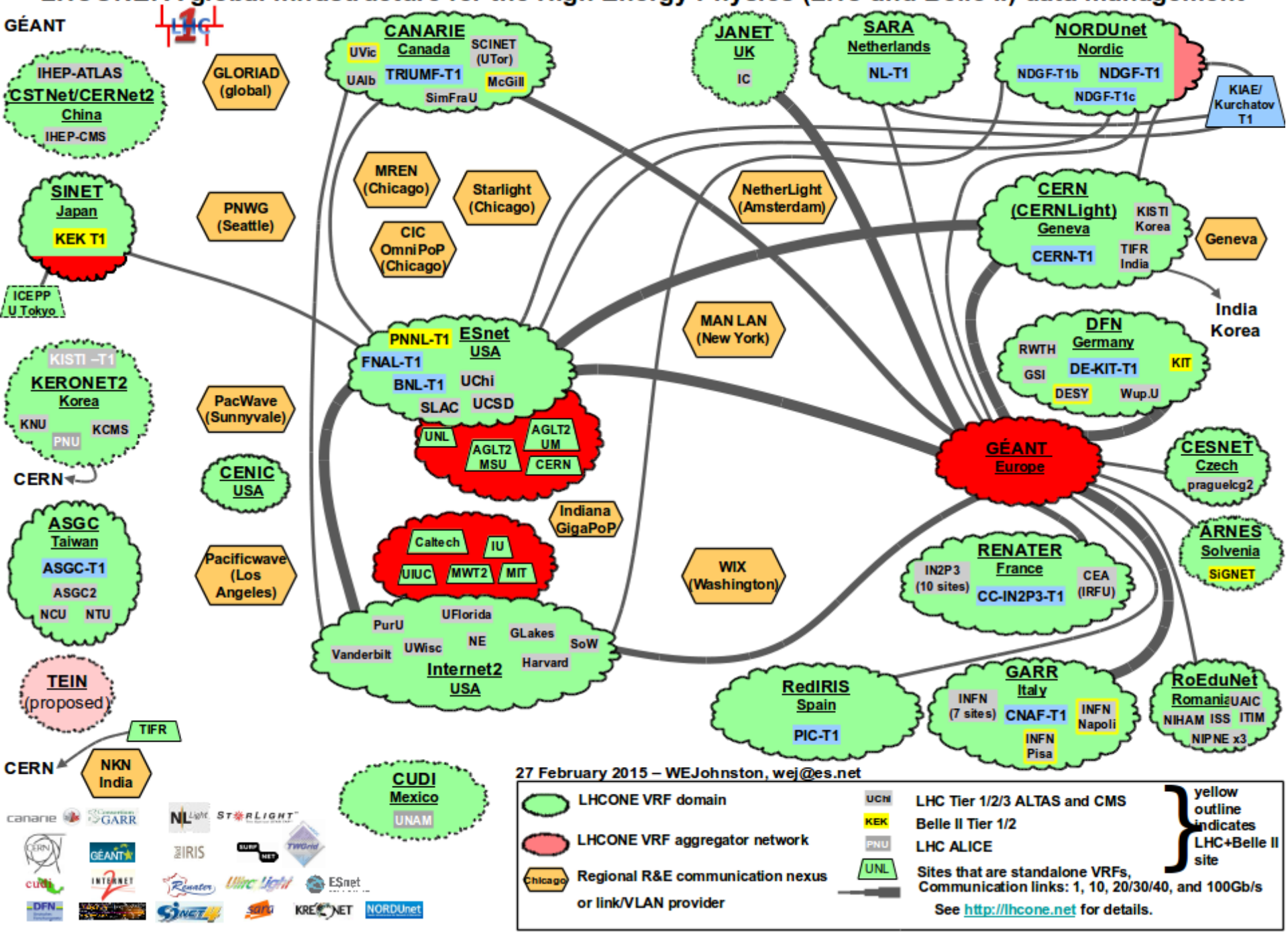


LHC OPN

- Optical Private Network
- Support T0 – T1 transfers
- Some T1 – T1 traffic
- Managed by LHC Tier 0 and Tier 1 sites



LHCONE: A global infrastructure for the High Energy Physics (LHC and Belle II) data management



Preparations for Run 2

Computing model update

- Scope:
  - In preparation for the data collection and analysis in LHC Run 2, the LHCC and Computing Scrutiny Group of the RRB requested a detailed review of the current computing models of the LHC experiments and a consolidated plan for the future computing needs. This document represents the status of the work of the WLCG collaboration and the four LHC experiments in updating the computing models to reflect the advances in understanding of the most effective ways to use the distributed computing and storage resources, based upon the experience gained during LHC Run 1

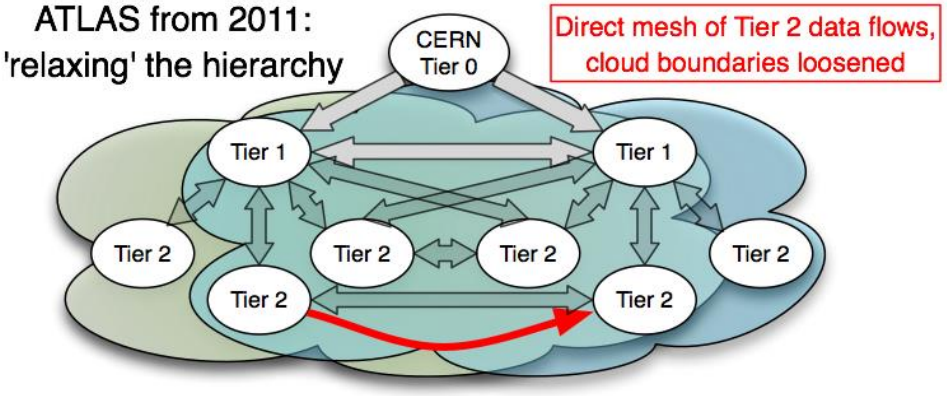
- Document was published early 2014
  - http://cds.cern.ch/record/1695401
  - CERN-LHCC-2014-014

Computing model changes

- Planned for significant changes in the trigger output rates and detector occupancy at higher energies and luminosities
- In Run 1 the experiments consistently used all of the pledged resources, and benefitted from the availability of additional resources above the pledges.
  - effort was invested to make use of non-dedicated resources (such as the HLT farms, external HPC resources, etc.).
- Changes in the computing models in each experiment, which represent significant efforts for the collaborations
  - Significant efforts invested in core software to improve the overall performance, and
  - Optimising reprocessing passes, number of data replicas stored, etc.,
- The community is investing significant effort in software development,
  - adapting HEP software for modern CPU architectures. This will be important for helping to ensure efficient use of available resources.

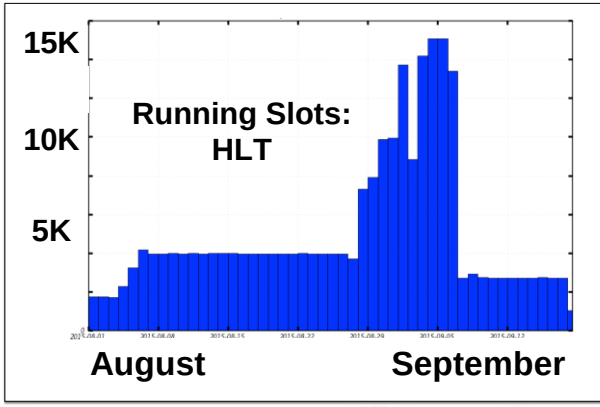
Distributed model

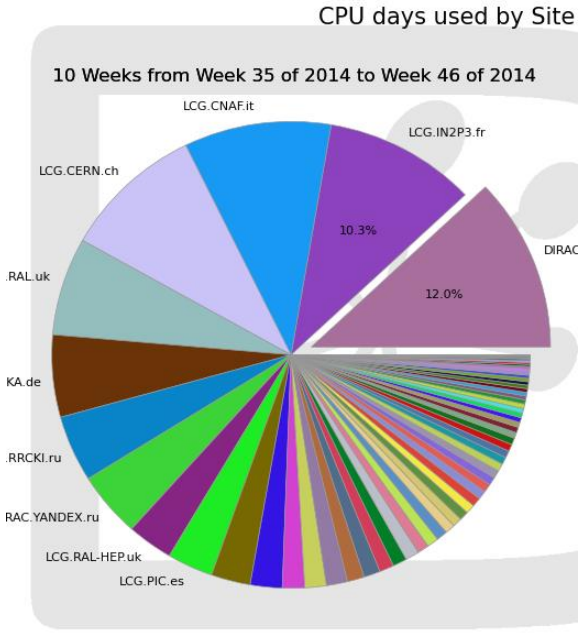
- Performance & reliability of the networks has exceeded early expectations or fears
  - 10 Gb/s → 100 Gb/s at large centres
  - 100 Gb/s transatlantic links now in place
  - Many Tier 2s connected at 10 Gb/s or better
  - NB. Still concern over connectivity at sites in less-well connected countries
- Strict hierarchical model of Tiers evolved even during Run 1 to make best use of facilities available
  - Move away from the strict roles of the Tiers to more functional and service quality based
  - Better use of the overall distributed system
- Focus on use of resources/capabilities rather than “Tier roles”
  - Data access peer-peer: removal of hierarchical structure



Common themes

- Optimisation of the processing:
  - Reduced number of (re-)processing passes
  - Automation of calibration and alignment procedures
  - Significant reductions in data formats, contents, and replicas
- Use of HLT for various tasks
  - Initially simulation, but now other tasks too
  - ALICE/CMS/ATLAS use openstack (cloud) to manage dynamic configuration between DAQ/offline
  - LHCb dynamically re-allocate HLT capacity to offline
  - Significant CPU resources – rely on Tier 0 for storage
  - Intend to use when feasible – not just in machine stops





Common themes – 2

- Better data management tools
  - Data federations introduced: allows access to remote data
  - More intelligent data placement algorithms, better clean up of unused data (“data popularity”)
  - Significantly improved File Transfer Service (FTS) – now a (more) central service
- Significant improvements in software performance, in order to keep resource requirements manageable and realistic
- Updated production and analysis tools
  - Move to multi-core and multi-threaded applications
- Simplification of “grid” services
  - Reduction of complexity, more centralisation

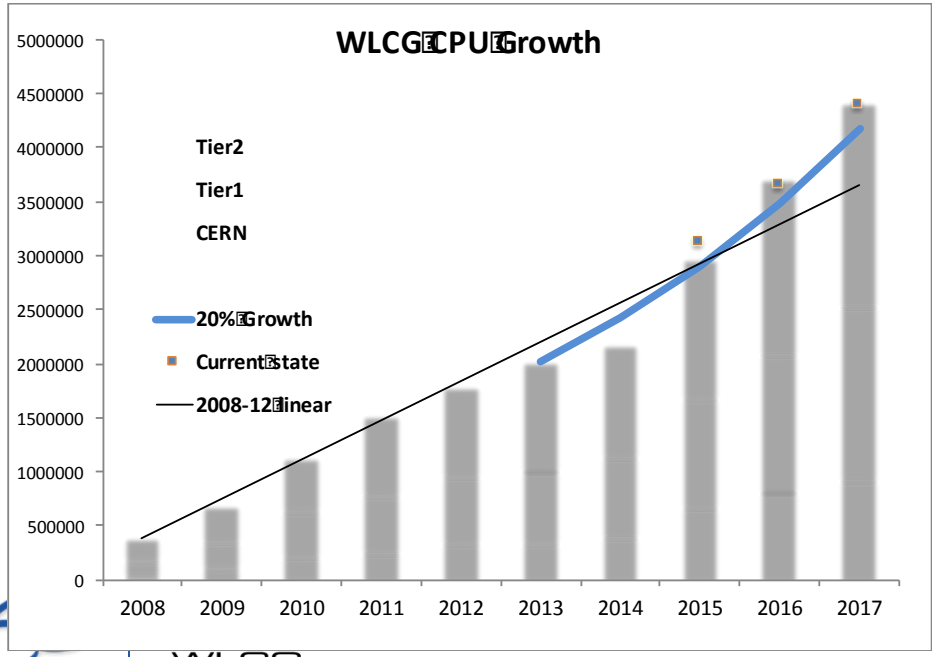
Evolution of requirements

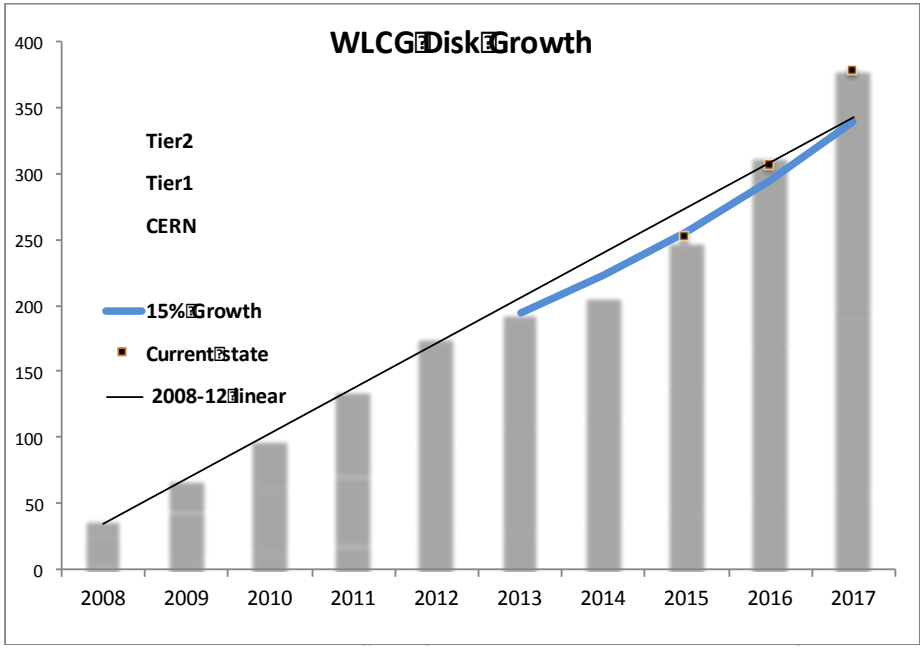
The reliability of resource predictions is continually improving, the largest uncertainties being the LHC running conditions.

Funding guidance: flat budgets for computing

Estimated evolution of requirements 2015-2017

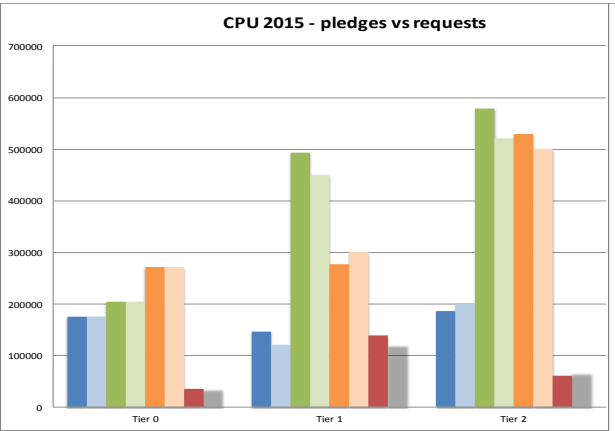
2008-2013: Actual deployed capacity

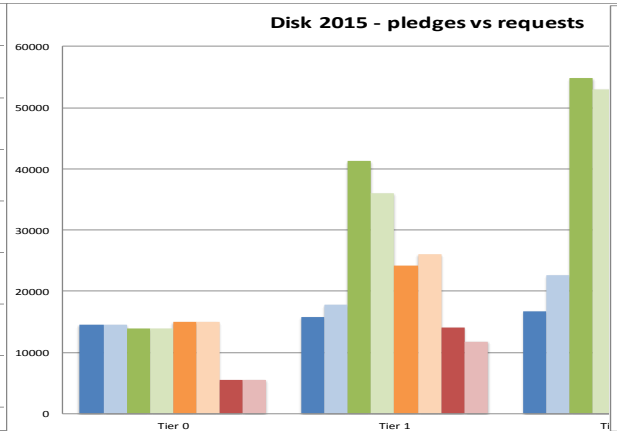


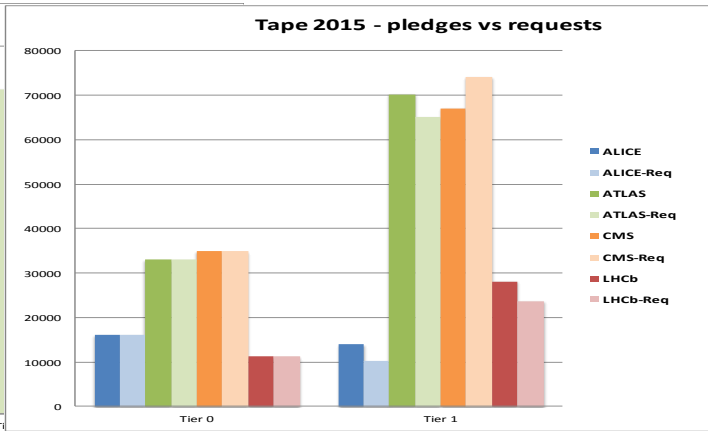


Run 2 readiness

- WLCG sites were prepared for Run 2
- Some delays in deploying new resources, but all were in place for 1st collisions
- Major procurement exercise has essentially doubled existing installations
  - Large scale procurement activities – that took significant effort and time
- WLCG resources fully deployed at the levels requested for 2015, 2016 requests are also on track to be satisfied







Evolution of the CERN Data Centre

- Networks allow us to break the boundaries –
- Existing CERN Geneva Data Centre reached power capacity
- New annexe procured …

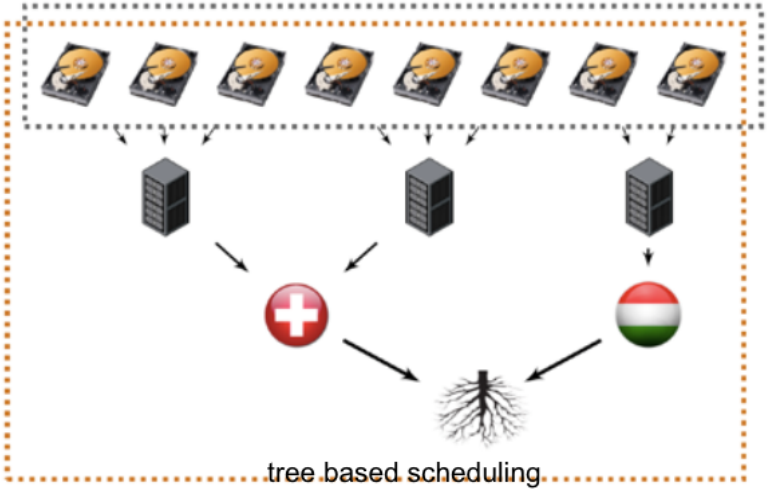
- In Budapest
  - Connected at 2 x 100 Gbps

- Operated as an integral part of the CERN DC

Data scheduling in Tier 0

- Meyrin & Wigner have 22 msec latency
- Initial strategy was a replica in each centre
  - Batch jobs tagged with GEO location and data access prefers local replica

- Newer EOS versions will permit 3 strategies:
  - scattered — max. distance — e.g max. distance regarding <site>
  - hybrid — e.g min. distance regarding <site>, but max. distance in <rack>
  - co-located — min. distance — e.g min. distance regarding <site>



Scale of CERN Data Centre

Overview: Data Centre — a month ago to a few seconds ago

MEYRIN DATA CENTRE

	last_value
Number of Cores in Meyrin	111,363
Number of Drives in Meyrin	68,846
Number of Memory Modules in Meyrin	75,415
Number of 10G NIC in Meyrin	4,479
Number of 1G NIC in Meyrin	21,035
Number of Processors in Meyrin	20,281
Number of Servers in Meyrin	10,972
Total Disk Space in Meyrin (TB)	118,211
Total Memory Capacity in Meyrin (TB)	436

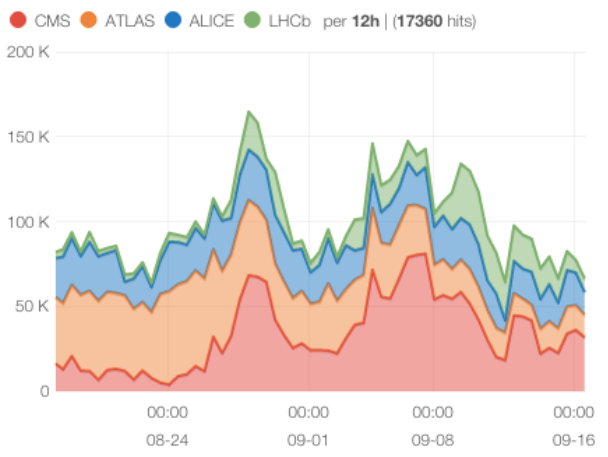
WIGNER DATA CENTRE

	last_value
Number of Cores in Wigner	43,264
Number of Drives in Wigner	23,171
Number of Memory Modules in Wigner	21,606
Number of 10G NIC in Wigner	1,399
Numer of 1G NIC in Wigner	5,059
Number of Processors in Wigner	5,410
Number of Servers in Wigner	2,708
Total Disk Space in Wigner (TB)	71,722
Total Memory Capacity in Wigner (TB)	172

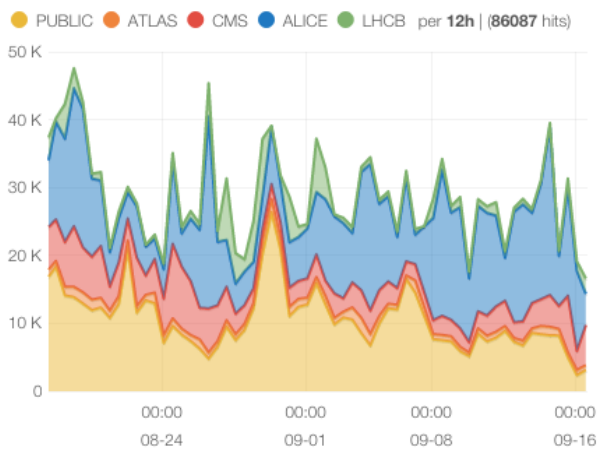
NETWORK AND STORAGE

	last_value
Tape Drives	104
Tape Cartridges	26,502
Data Volume on Tape (TB)	116,706
Free Space on Tape (TB)	48,348
Routers (GPN)	134
Routers (TN)	29
Routers (Others)	94
Star Points	641
Switches	3,547

 # BATCH JOBS



 # EOS ACTIVE DATA TRANSFERS



 # VM CREATED

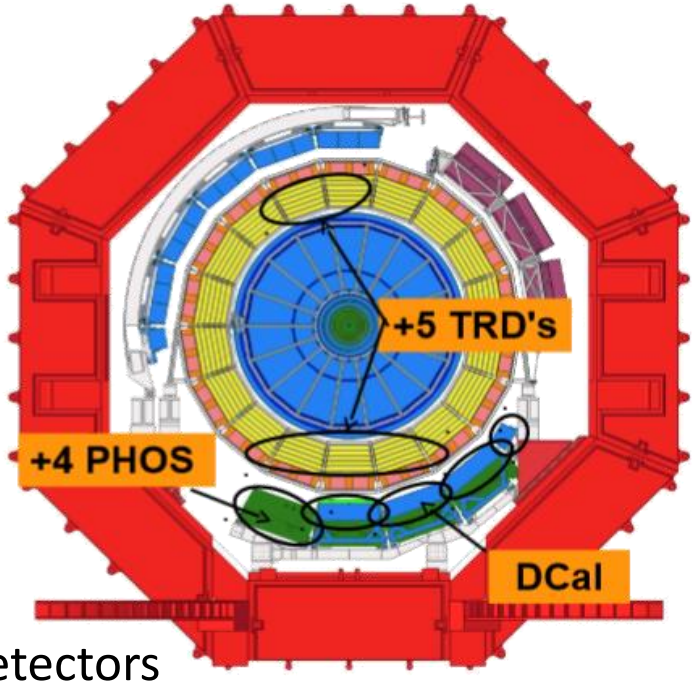
CERN Private cloud

- CERN capacity is now configured as a private cloud (IaaS)
  - Using openstack to provision resources
- Important to aid more flexible resource provisioning especially for services
  - Needed a mechanism to help us scale out the size of the Tier 0 by many factors without additional staff
  - Mechanism for dynamic extension of the resources to Wigner, or to commercial clouds (see later)
  - Also same system used to manage HLT farms

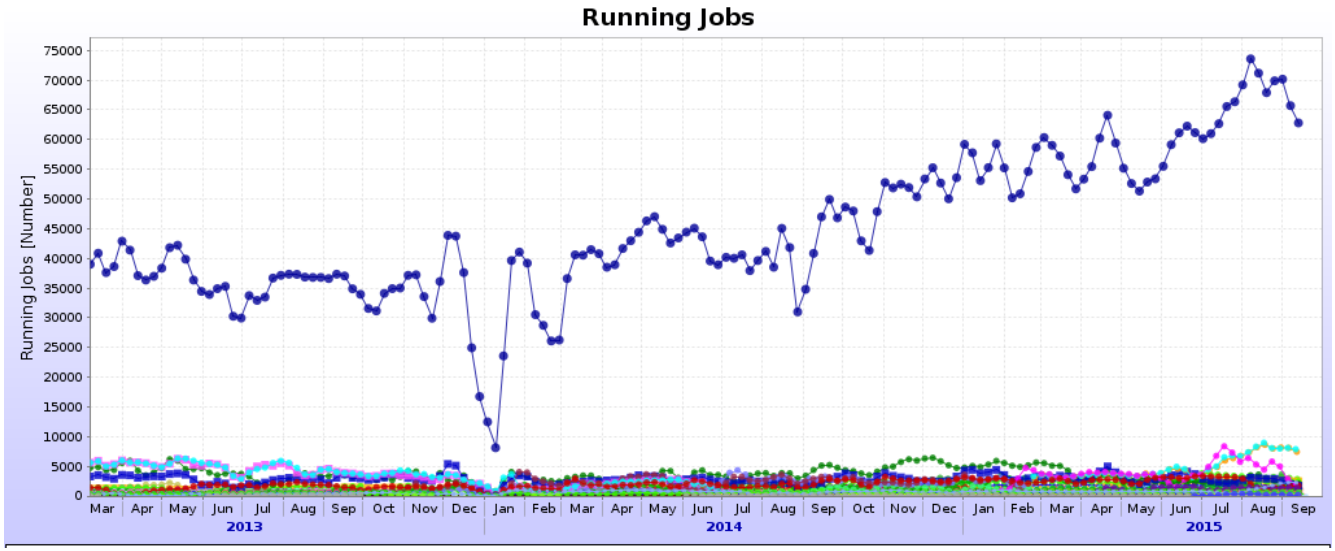
Experiment preparations for Run 2

ALICE during LS1

- Detector upgrades
  - TPC, TRD readout electronics consolidation
  - TRD full azimuthal coverage
  - +1 PHOS calorimeter module
  - New DCAL calorimeter
- Software consolidation
  - Improved barrel tracking at high p_T
  - Development and testing of code for new detectors
  - Validation of G4
- Re-processing of RAW data from RUN1
  - 2010-2013 p-p and p-A data recalibrated
  - All processing with the ***same*** software
  - General-purpose and special MC completed

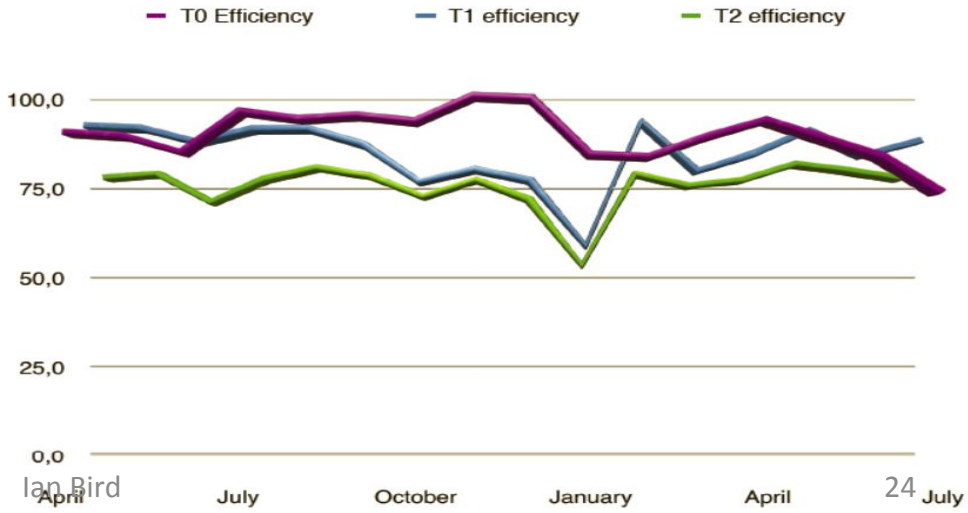


Grid utilization: end of RUN1 - today



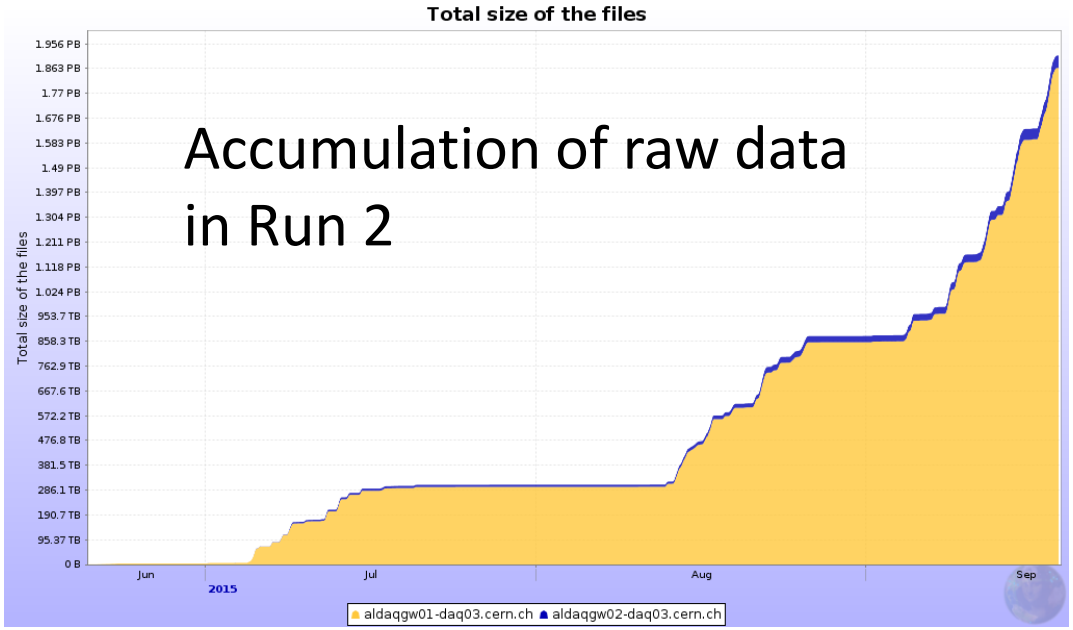
- Reaching new highs
- 96K parallel jobs

- Consistent and very good efficiency in all computing centres



Run2 progress

- Steady data taking
  - Fast calibration and reconstruction
  - Quasi-online data processing
    - Without backlog

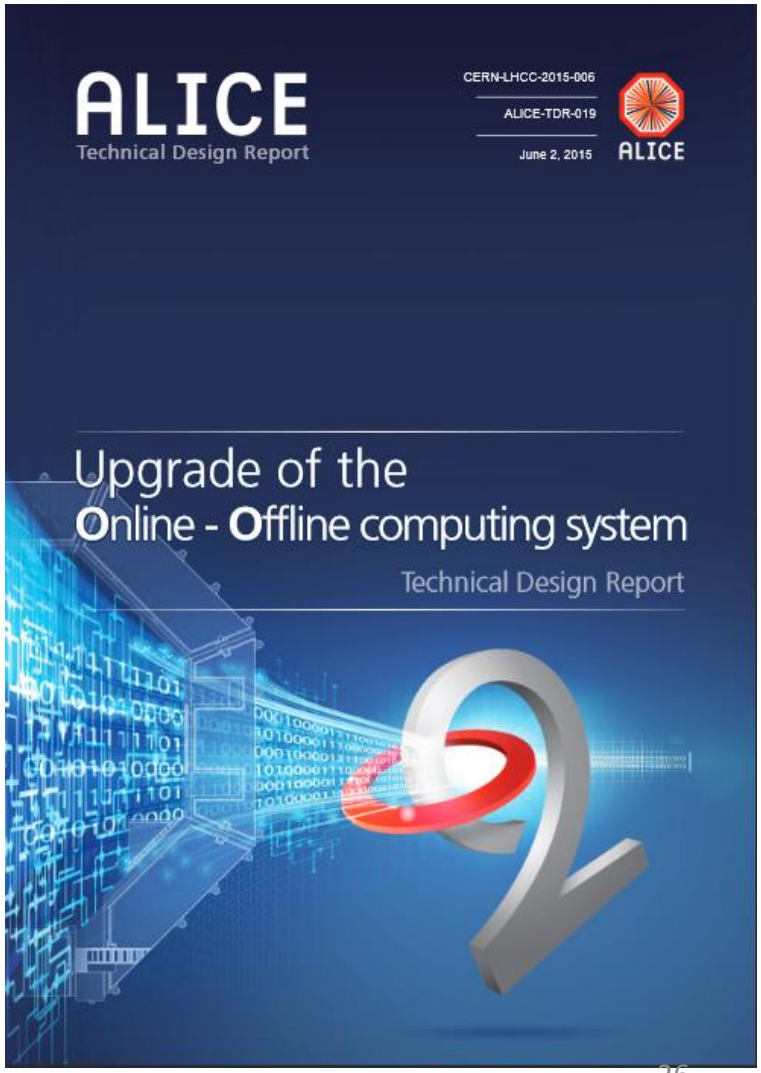


Production	Description	Status	Run Range	Runs	Chunks	Size	Chunks		Size		Events
LHC15i_pass1	LHC period LHC15i - Full production pass 1	Running	235196 - 235886	25	101,081	164.8 TB	83,245	82%	8.665 TB	6%	62,007,426
LHC15h_pass1	LHC period LHC15h - Full production pass 1	Completed	232914 - 234050	68	327,386	544.9 TB	293,861	89%	38.96 TB	7%	213,863,587
LHC15g_pass1	LHC period LHC15g - Full production pass 1	Completed	228855 - 230292	31	26,567	37.65 TB	26,171	98%	6.125 TB	16%	20,766,687
LHC15f_pass1	LHC period LHC15f - Full production pass 1	Completed	224895 - 226532	45	18,857	21.9 TB	16,542	87%	12.1 TB	62%	84,564,615
LHC15e_pass1	LHC period LHC15e - Full production pass 1	Completed	223270 - 224772	59	15,648	9.16 TB	11,595	74%	1.685 TB	24%	73,262,707
LHC15d_pass1	LHC period LHC15d - Full production pass 1	Completed	220139 - 222966	100	6,148	5.513 TB	5,234	85%	656.5 GB	13%	29,817,237
					495,687	**783.9 TB**	**436,648**		**68.17 TB**		**484,282,259**

Beyond Run2 – the O^2 project

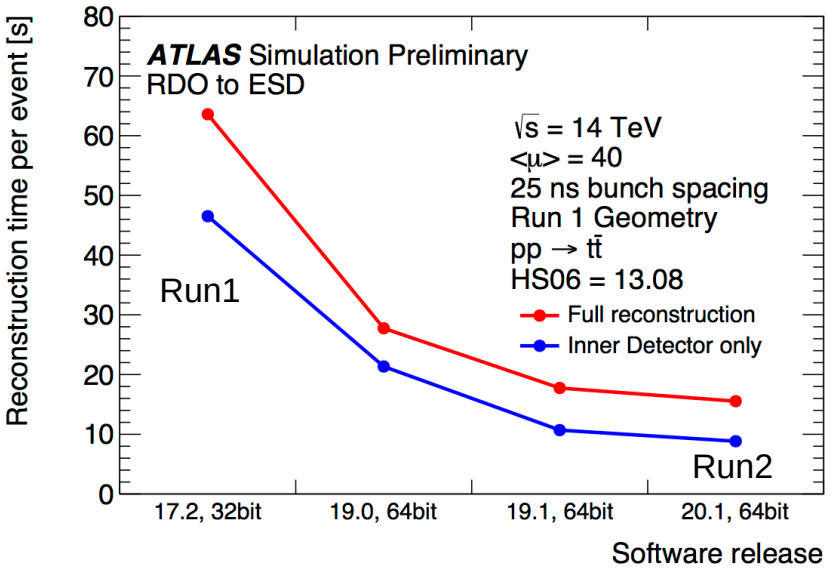
Technical Design Report for the Upgrade of the Online-Offline Computing System

- Submitted to LHCC
- Framework and code development is well under way

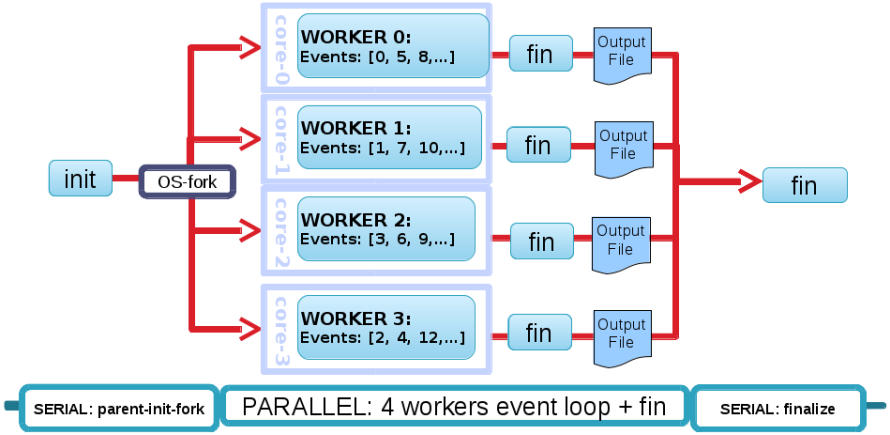


Improvements in LS1

- Speedup of reconstruction: **factor 4**
- Simulation **20% faster**
- Software moved to **multi-core** (parallel event processing)
- New **data management** and **production** systems; deployed late 2014
- New **analysis model** : New *ROOT readable* data format with analysis outputs made using a train model
- + **Many others**…



Schematic View of ATLAS AthenaMP

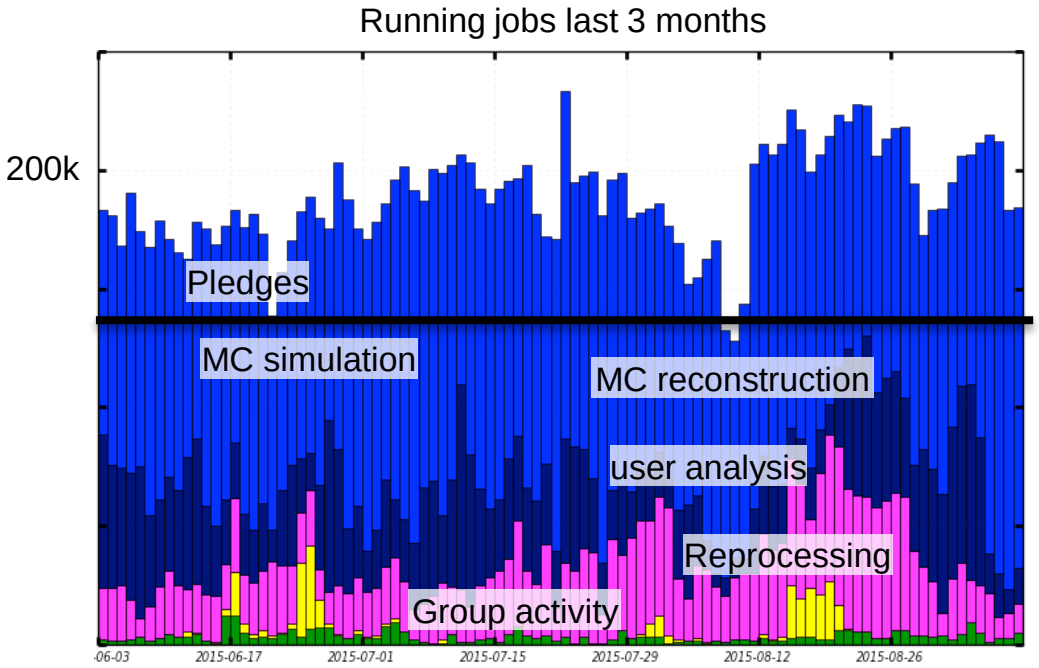


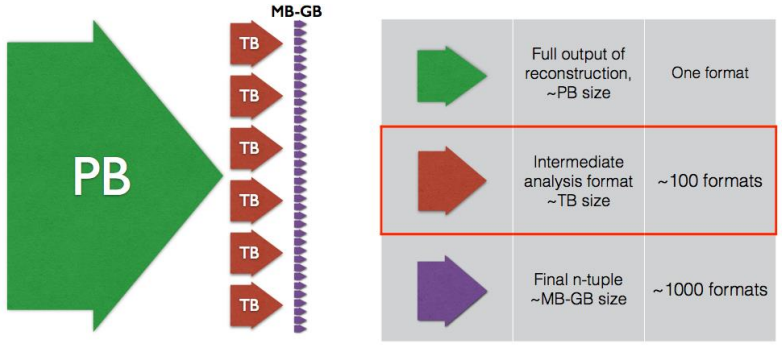
Parallel event processing

ATLAS EXPERIMENT
http://atlas.ch

Computing

- Grid utilisation at full :
  - MC simulation: 2.8B simulated events produced
  - ~5B events reconstructed for 50 ns & 25 ns conditions
- No issue with data transfer and data processing. 2015 data have been reprocessed twice. Major software update for summer 2016 only.
- New **analysis model** : group data format DxAOD made using a train model
  - Production of 83 DxAOD species on the grid via 17 trains
  - Within 24h after data reconstruction at Tier-0
  - Successful and popular





ATLAS Summary

- 2015 data have been processed, distributed and analysed without major issue
- New analysis model is working
- 2017 resource requests have been reevaluated in light of updated LHC running parameters
- Software upgrade for Run 3 and beyond is on track
- Sizeable computing resources will be needed for TDRs for ATLAS upgrades

Computing Evolution Changes for More Agile Operations

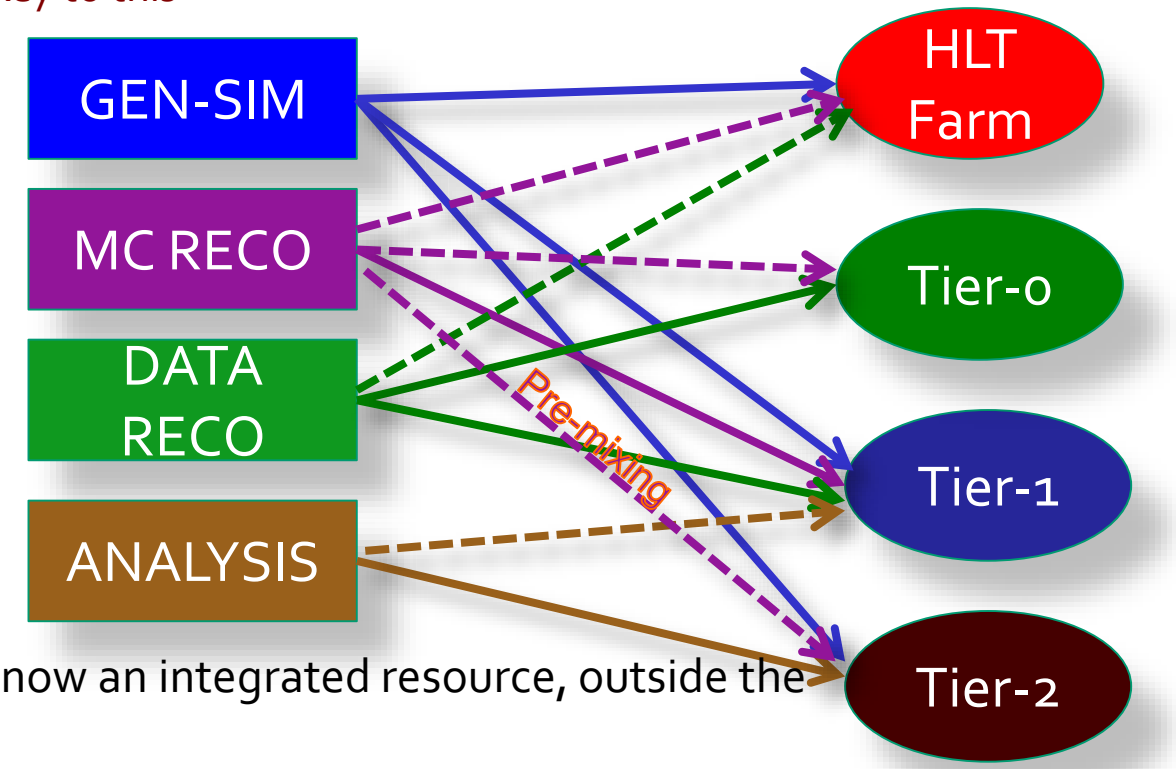
- **Anydata, Anywhere, Anytime (AAA)**
  - CMS applications can read data efficiently over wide-area networks
  - Relaxes constraints on locations of datasets and workflows
- **Disk-tape separation at Tier-1 sites**
  - Greater control over what datasets are available on disk
  - Through AAA, allows T1 data to be used in workflows anywhere
- **Dynamic Data Management**
  - Automatic transfers of datasets on creation, deletion when not needed
  - More agile and efficient use of disk space
- **Global Pool for resource provisioning via GlideInWMS**
  - Allows central control of job priorities, simplified infrastructure
  - Demonstrated scaling to operate all T1/T2/opportunistic resources in single pool
- **Ability to provision cloud infrastructures via glideinWMS**
  - Allows use of HLT and potentially opportunistic and commercial clouds
  - Ability to burst into extra resources if necessary
- **Establishment of 100 Gbps transatlantic network link via ESnet**
- **MiniAOD: new analysis format for Run2**
  - Compact format: ~30-50Kb/event (10% the size of the Analysis Datasets used in Run1) that can serve ~80% of all CMS analyses

Flexibility for Facilities and Workflows

- Use excellent facilities in more flexible, heterogeneous ways
  - Improved networks have been key to this

- **Data Federation** will make CMS datasets available transparently across the Grid

- One **central queue** for all resources and all workflows will facilitate prioritization between analysis and production

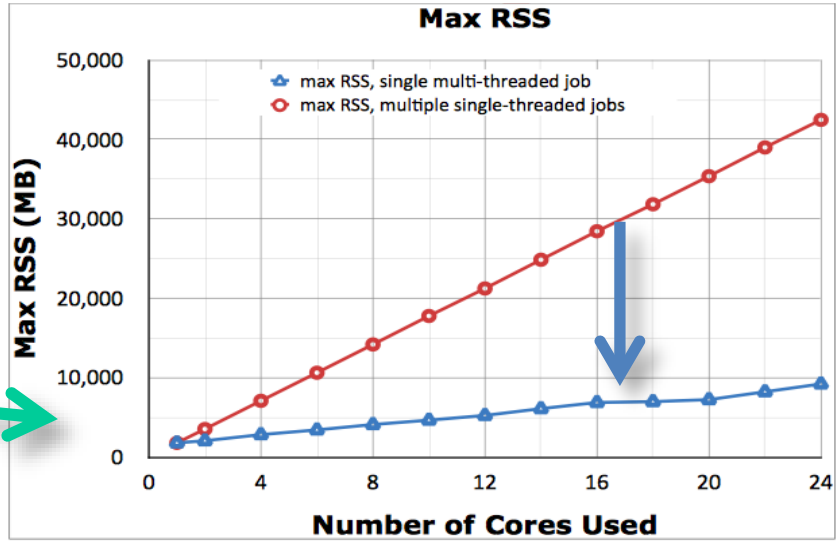


- The **HLT farm** (~size of Tier-0) is now an integrated resource, outside the LHC running
- Commission Tier-2's to do reconstruction tasks previously limited to Tier-1's
- Allow analysis jobs to run at more sites

The more places work can run, the faster the work goes!

The CMS Tier-0 for 2015 is multithreaded

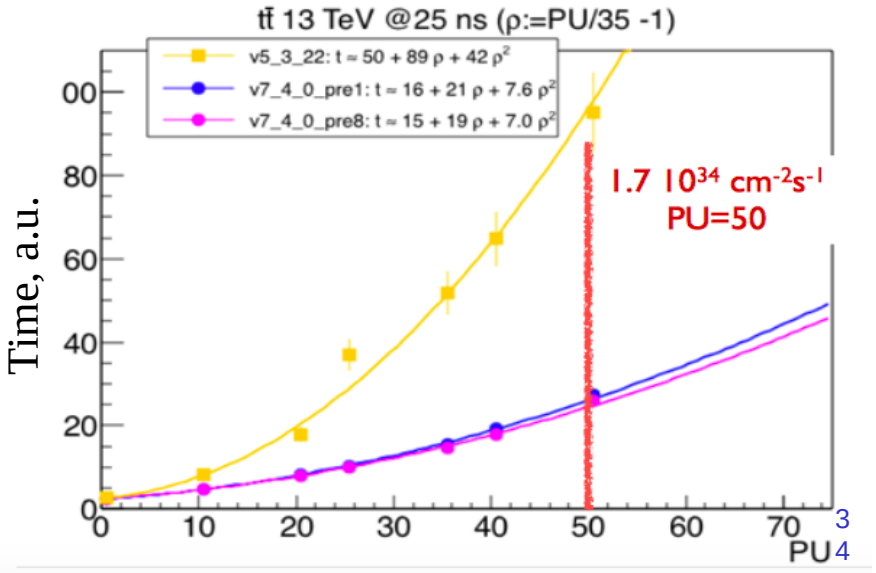
- The threaded framework and reconstruction algorithms developed during LS1 is now deployed in production at Tier-0
  - Efficiency process large trigger rates in Run 2 with low latency
  - Dramatic memory savings from threading
- The memory savings lets us produce a full suite of outputs directly from the Prompt Reco application
  - Reconstructed events in full format and two analysis formats (AOD and MiniAOD)
  - Monitoring histograms for data certification
  - Detector and physics skims
- On-going development priorities for 2015
  - Finalize 25 ns reconstruction configuration
  - Evolve MiniAOD format to follow analysis needs. Since MiniAOD can be rederived from the AOD format, CMS can quickly reproduce these data when needed



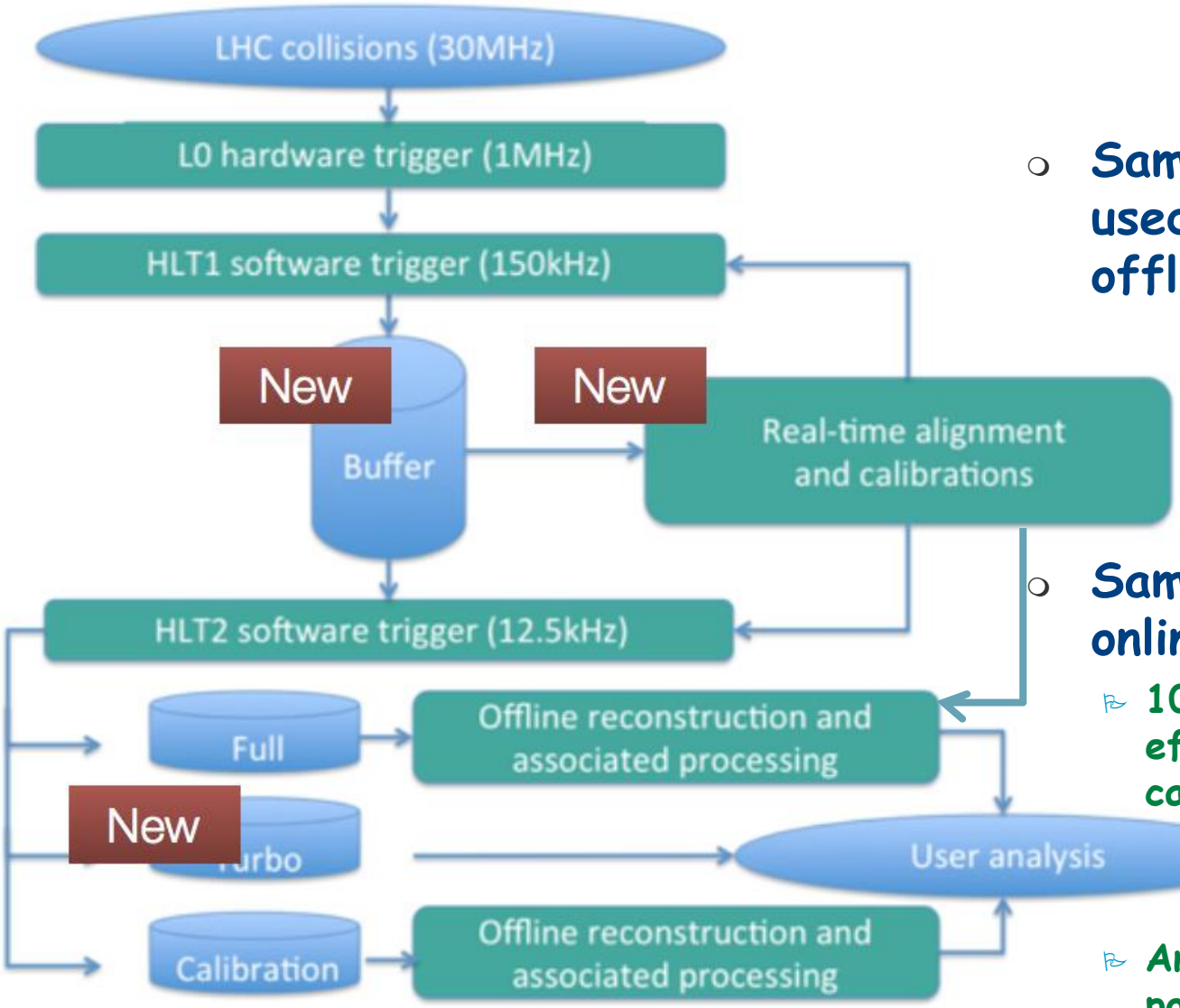
CPU for SIM/RECO: Ready for Run2

- Large technical performance gains achieved during LS1
  - Simulation: Factor of 2 gain in CPU utilization, primarily from Russian Roulette sampling algorithm to reduce time spent tracking low-energy particles in Geant4
    - Visible improvements already in the number events/month produced for CSA14(CMSSW6_2) and 2015 production RunIIWinter15GS (CMSSW7_1)
  - Reconstruction: Large gains, particularly in tracking area and algorithms appropriate for 25 ns conditions)

These achievements were essential to meet Run2 challenges within resource constraints

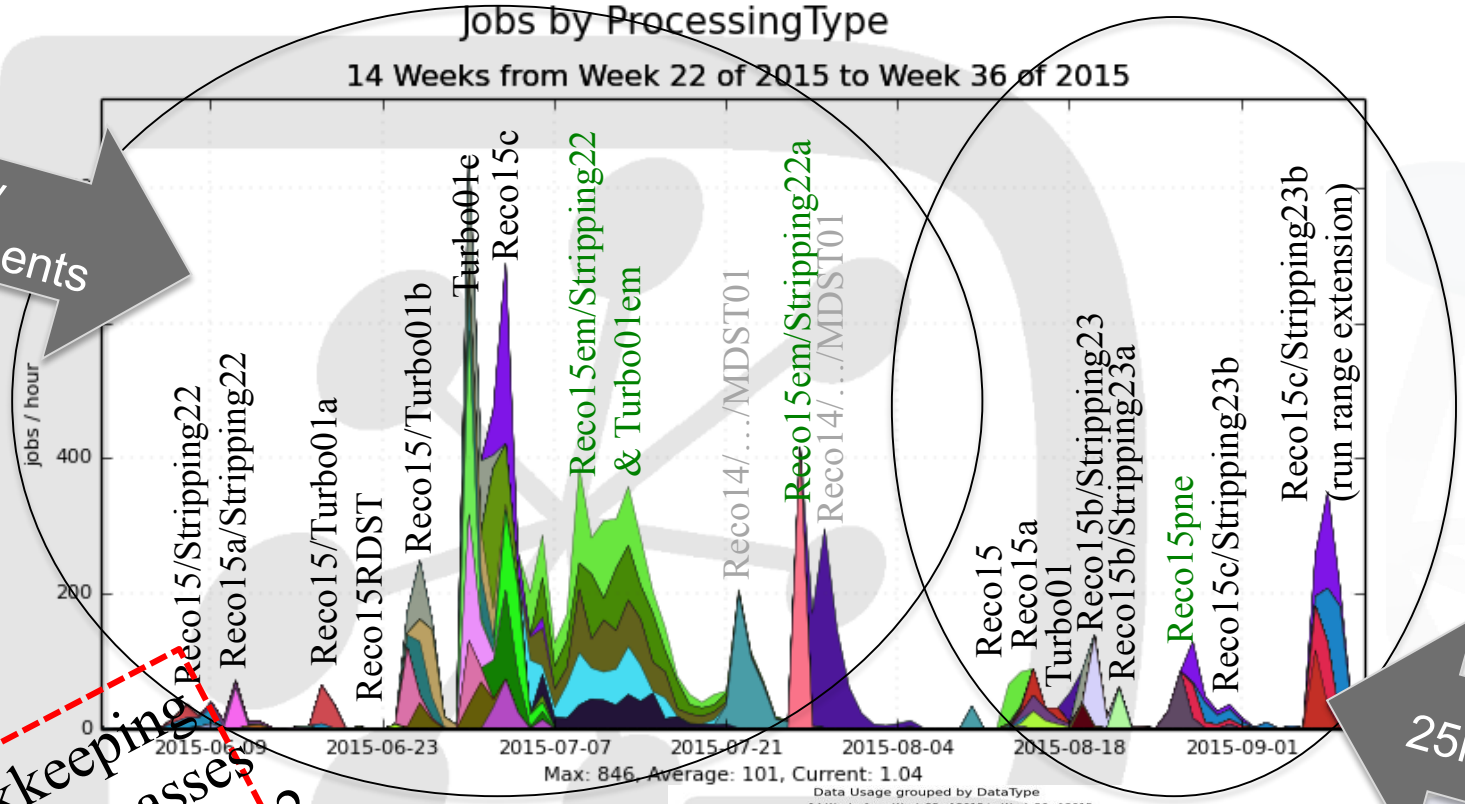


Split HLT in Run 2



- Same calibration used online and offline
  - No reprocessing
- Same reconstruction online and offline
  - 100% offline selection efficiency on trigger candidates
  - Analysis selections possible online (TURBO stream)

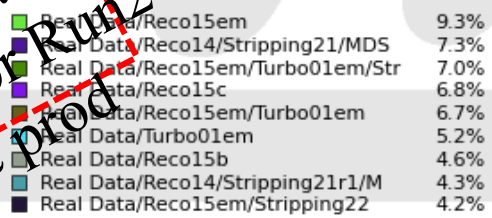
Run 2 Data Validation & Production

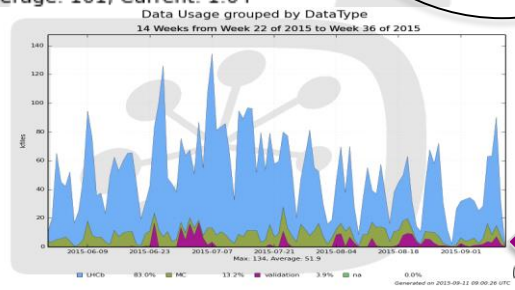


Early Measurements

25ns ramp

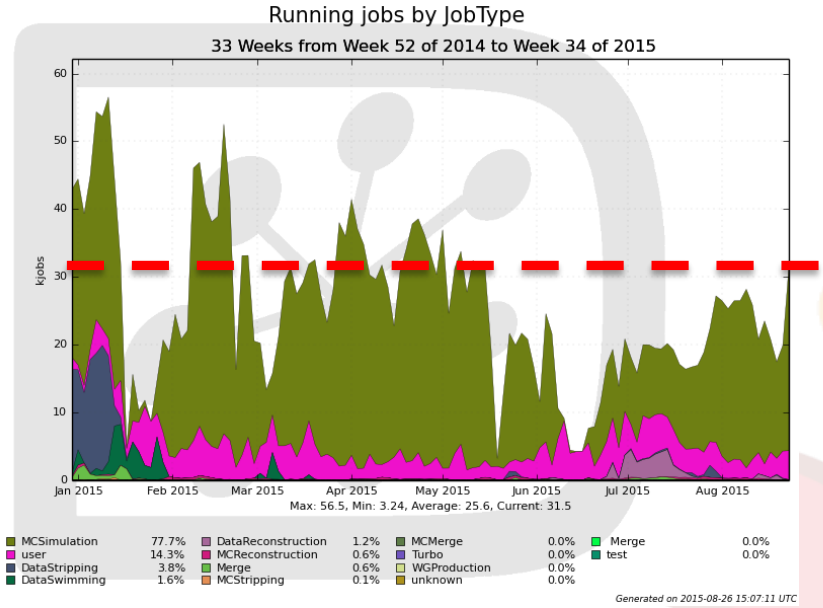
41 Bookkeeping processing passes produced for Run2 data val & prod

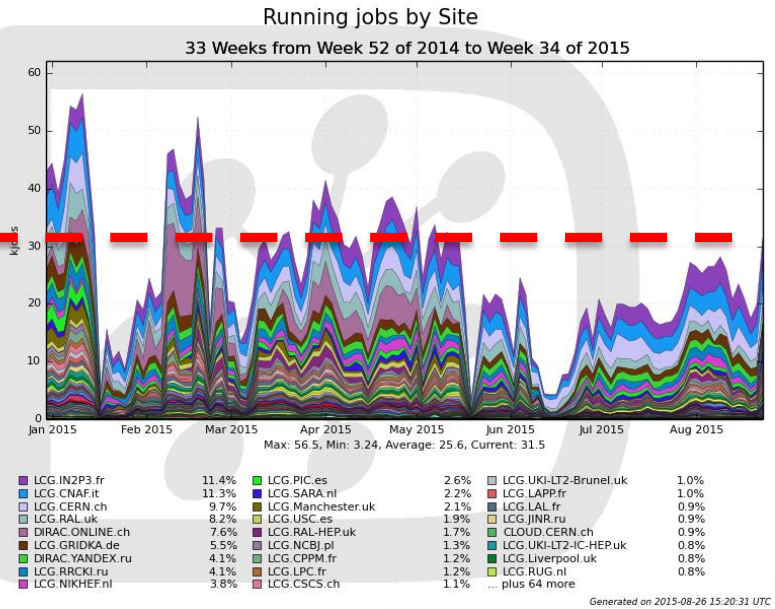




Validation popularity

Running jobs





- Usage consistent with pledges
- Somewhat below in Q3 due to reduced data-taking and preparation of new simulation cycle
- LHCb continues to use efficienctly the HLT farm and opportunistic resources (Yandex, OSC, Zurich and others)
- Expect usage in line with pledges until the end of 2015 WLCG year

Recent changes in the Computing Model

- Follow the recommendation of the CRSG: cancel second copy of ARCHIVE.
  - In case of tape losses, derived data would have to be regenerated.
  - Decrease of tape requests for 2016 and 2017.
- Do not store the content of RAW banks in FULL.DST.
  - ~2x saving on FULL.DST size
  - Decrease of tape requests.
- Re-balance generation of MC events in 2016 and 2017, in order not to impact physics analysis
  - (small) increase in CPU and storage
- Postpone parking of RAW data to 2018 (if really needed)

Longer term

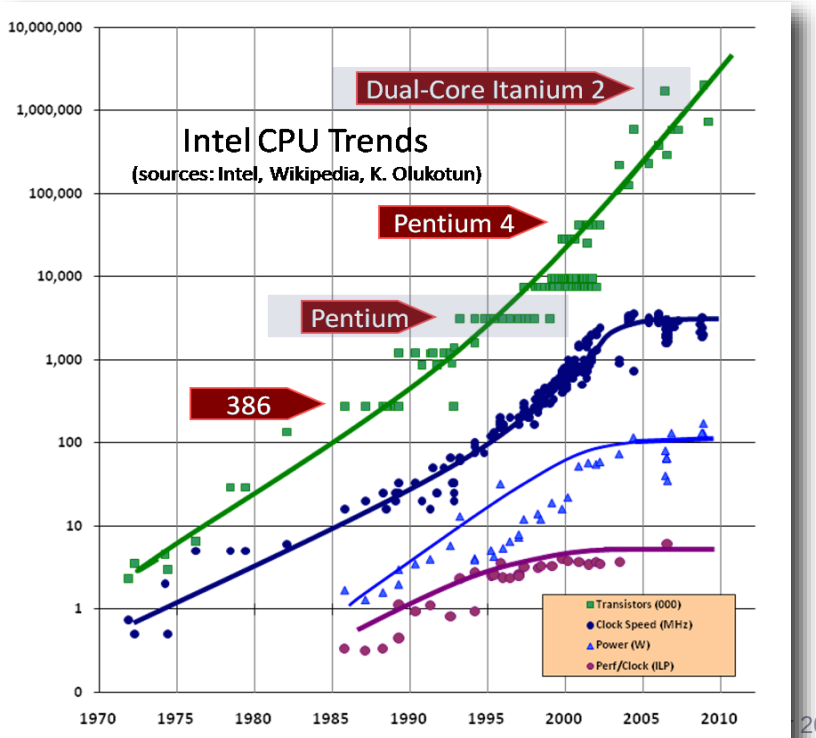
Evolution and challenges

- WLCG Grid → federated grid/cloud/other resources
  - Reduce operational effort so that WLCG Tiers can be self supporting (no need for external funds for operations)
  - Enable the experiments to easily make use of opportunistic resources
    - (Grid) clusters, clouds, HPC, …

- Challenges:
  - Huge increases in data volumes and processing needs
    - 25 PB/year in 2012 → 400 PB/year in 2024
  - Software complexity and performance
    - Modern CPU architectures require significant software re-engineering
  - Must live within ~flat budgets

Software

- Not just a HEP problem
- Transistors go into many cores, co-processors, vector units, etc.
- Memory access and I/O paths also become problematic
- Getting performant software requires significant investment in skills



The new realism: software runs slowly on supercomputers

No supercomputer runs real applications faster than five percent of its design speed. Robert Roe and Tom Wilkie report on recalibrating expectations of exascale and on efforts to tune software to run faster

director of the HLRS supercomputer centre in Stuttgart, Germany. The HLRS was not targeting exascale per se, he said, because the centre's focus was on the compute power it could actually deliver to its users.

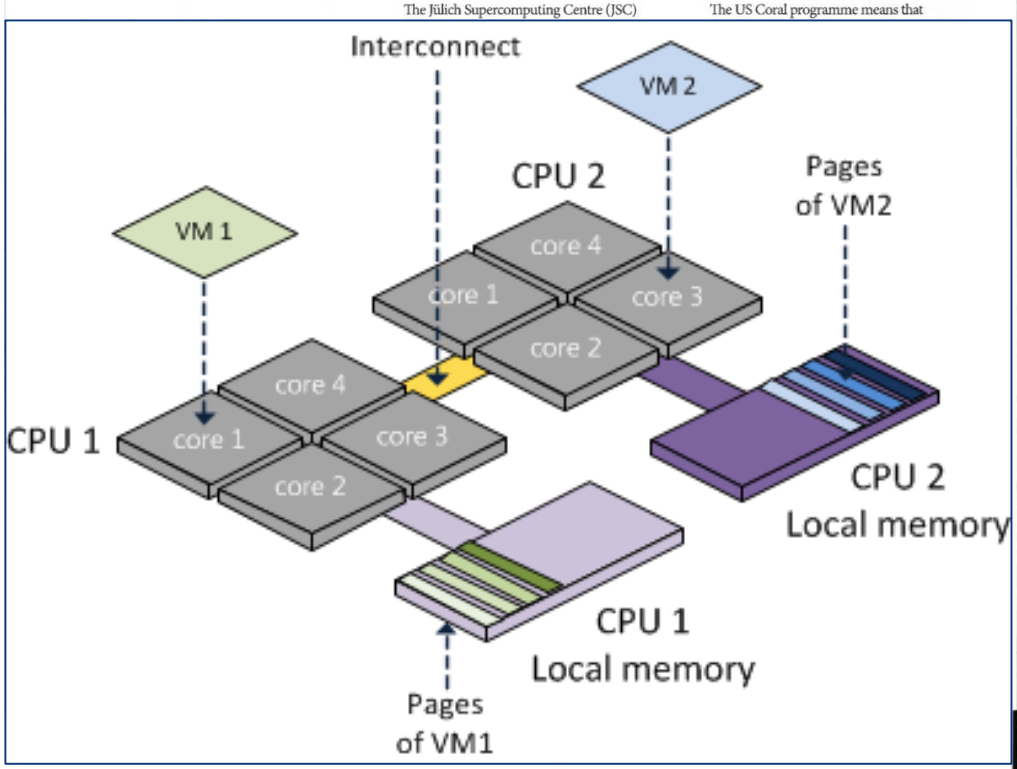
According to Resch, simple arithmetic meant that if an exascale machine achieved a sustained performance of only 1 to 3 percent, then this would deliver 10 to 30 Petaflops. So buying a 100 Petaflop machine that was 30 per cent efficient – which should be achievable, he claimed – would deliver the same compute power, for a much lower capital cost and about one tenth the energy cost of an exascale machine.

The Jülich Supercomputing Centre (JSC)

encourage our users to try and reach exascale readiness.'

At the Lawrence Livermore National Laboratory, effort is going into developing APIs and tools to create applications and effectively optimise how code is run on a cluster. Supinski stated that the LLNL's plan was to use a programming tool developed at Livermore called Raja. "The idea of Raja is to build on top of new features of the C++ standard. The main thing that we are looking at is improving application performance over what we are getting on Sierra, Sequoia, and Titan. Application performance requirements are what we really care about.'

The US Coral programme means that



HEP Software Foundation

- Organisation put in place to address the software performance concerns for HEP
- Have a framework to coordinate efforts and to prioritise common requirements
  - Help to bring in expertise and external funding
  - Coordination with e.g. CERN openlab (INTEL, etc.)
- Building on the work of the concurrency forum
- Have had 2 large workshops and regular meetings of a core team
- Aspects of work:
  - Technical forum
    - Place for technology discussion and dissemination of experiences
    - Help build expertise in the community
    - Interest group – optimising reconstruction software
  - Training
  - SW Knowledge Base to aid commonality of solutions
  - Performance optimisation – small scale activity already, hope to expand

Longer term planning

- Putting in place a “WLCG Technical Forum” to explore possible computing models for the HI-LHC era
  - Initial action will be to document some outline ideas for investigation
  - WLCG workshop in Feb 2016 is an opportunity for further input on this
  - In parallel look at a number of evolutionary topics
- Must be done in parallel with ramping up effort on software improvements (HSF)

Science Clouds

- Experiments and sites have made many explorations of use of private and commercial clouds:
  - Cloud infrastructures at many sites
  - Use of AWS, Google, Rackspace, etc by experiments, CERN, others
  - Helix Nebula EC project in Europe (together with other sciences)
  - Also testing real commercial procurements to understand cost
  - So far most use has been simulation, only now looking at data-intensive use cases: data federations help

- Pre-commercial Procurement (PCP) project has been approved and will start in January 2016
  - Derogation of CERN procurement rules to allow compliance with EC procurement (EC member states) agreed by CERN Council last week
  - Will allow a joint procurement across Tier 1s – understand if we can obtain economies of scale

HNSciCloud H2020 PCP Project

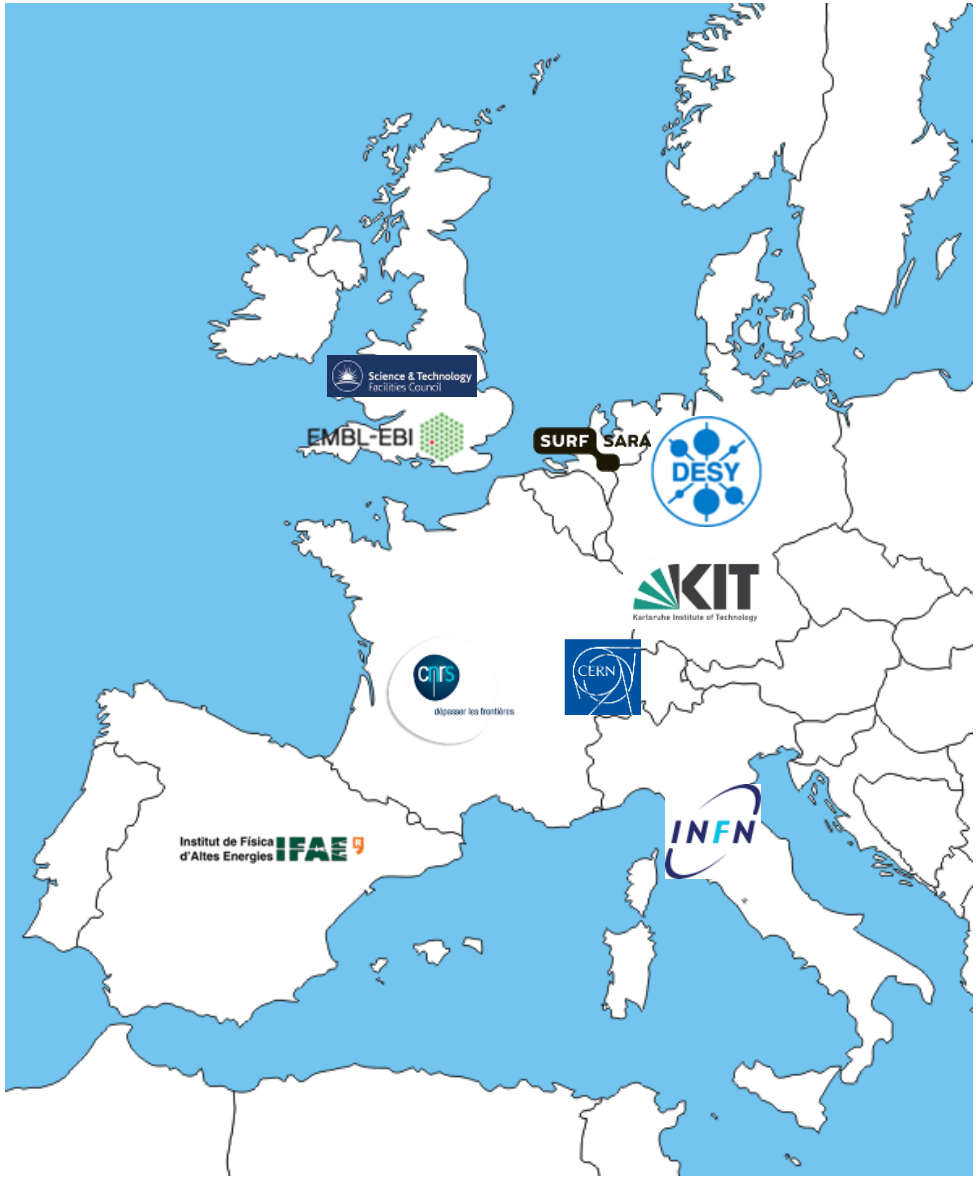
The group of buyers have committed

- ~1.6M€ of funds (generating ~6M€ total funds)
- Manpower
- Applications & Data
- In-house IT resources

To procure innovative IaaS cloud services integrated into a hybrid cloud model

- Commercial cloud services
- European e-Infrastructures
- In-house IT resources

Procured services will be made available to end-users from many research communities



Summary

- The experiences of Run 1 have resulted in significant changes in the LHC computing models
  - Evolving an operational system
- Run 2 has required significant additional resources and technical preparations
  - In production now and being demonstrated successfully
- The challenge for the coming years remains being able to evolve the system to adapt to changing technology and to fit within constrained budgets without compromising physics output