



**GridPP**

UK Computing for Particle Physics



UNIVERSITY  
of  
GLASGOW

# Future Directions for Computing in Particle Physics

RAL@50

13th Nov 2014

Prof. David Britton  
GridPP Project leader  
University of Glasgow



# Foreword

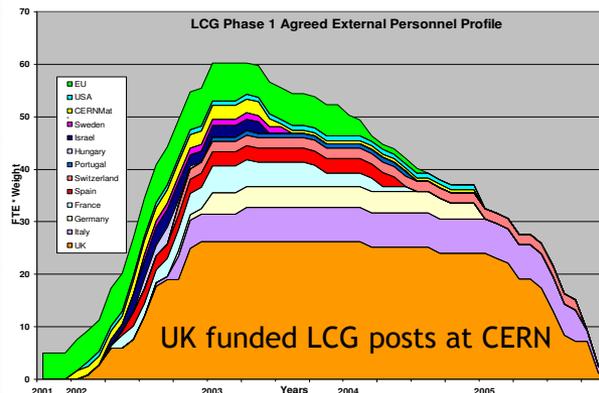
- In terms of funding and physicists, Particle Physics is currently dominated by the LHC which has produced the largest data-sets in the field.
- Computing for Particle Physics is, therefore, also dominated by the LHC computing, so this is the focus of this talk.
- LHC computing in the UK is provided by the GridPP project; we also support many (~30) other groups but the LHC accounts for 90% of what we do.



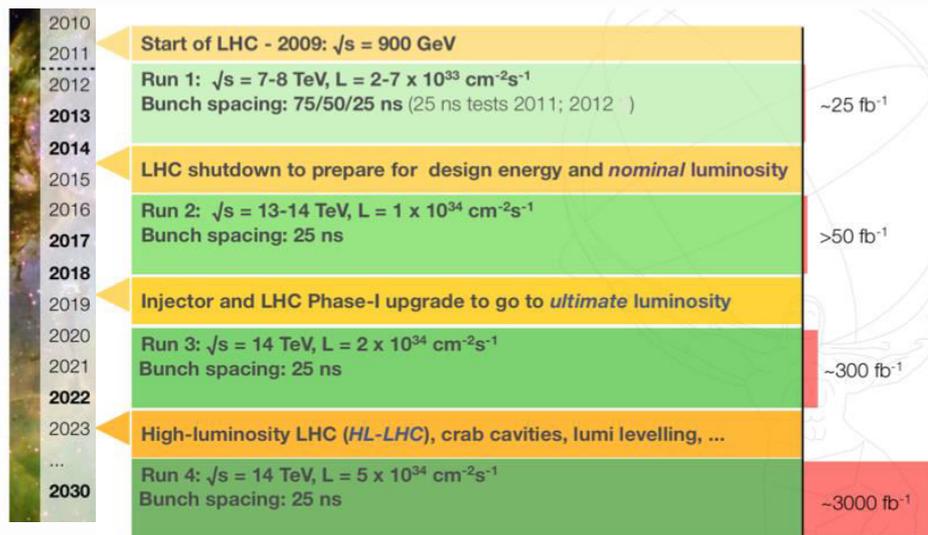
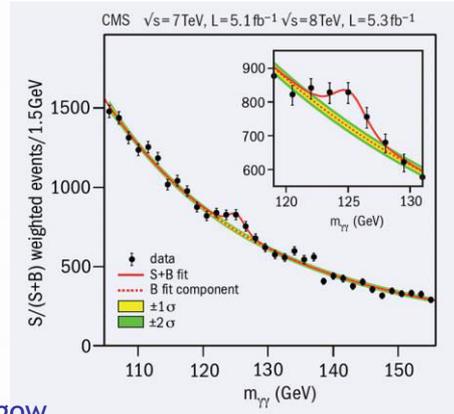
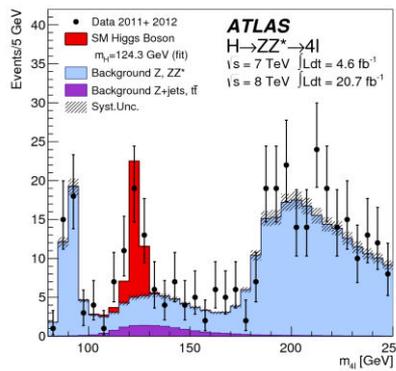
# GridPP Mission

Mission: To deliver resources to the UK and worldwide particle physics community in accordance with the WLCG MOU, by means of a large-scale computing Grid in the UK.

- 2001 GridPP1 - *From Web to Grid*
- 2004 GridPP2 - *From Prototype to Production*
- 2007 GridPP2+ (6-month extension)
- 2008 GridPP3 - *From Production to Exploitation*
- 2011 GridPP4 - *Computing in the LHC era*
- 2015 GridPP4+ (One year extension)
- 2016 GridPP5 - *Computing beyond the Higgs*



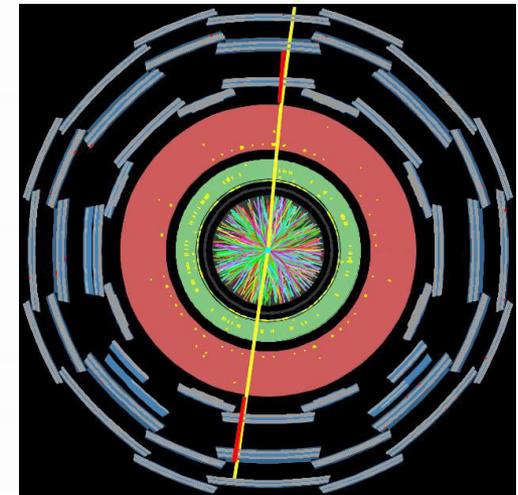
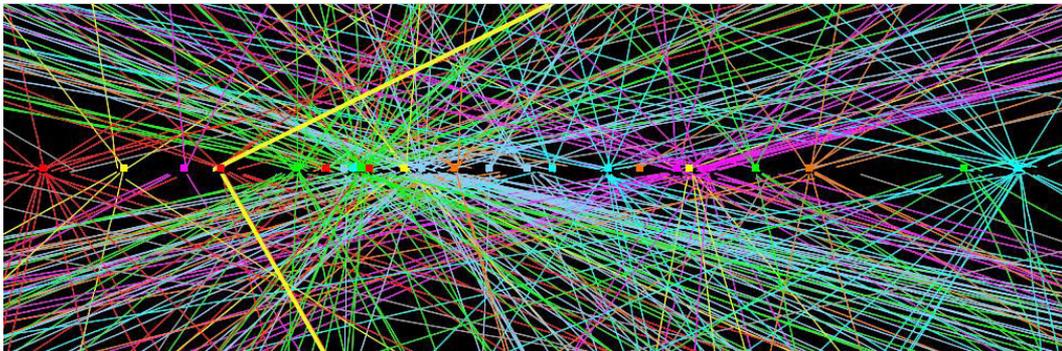
The UK kick-started WLCG in 2002 with a £5.6m investment.



- Current detectors designed for 10 years with  $\langle\mu\rangle=23$ .
- Performing well beyond this specification ( $\langle\mu\rangle=40$  expected in Run-2), but have limited life and will not handle HL-LHC  $\langle\mu\rangle=140$ .

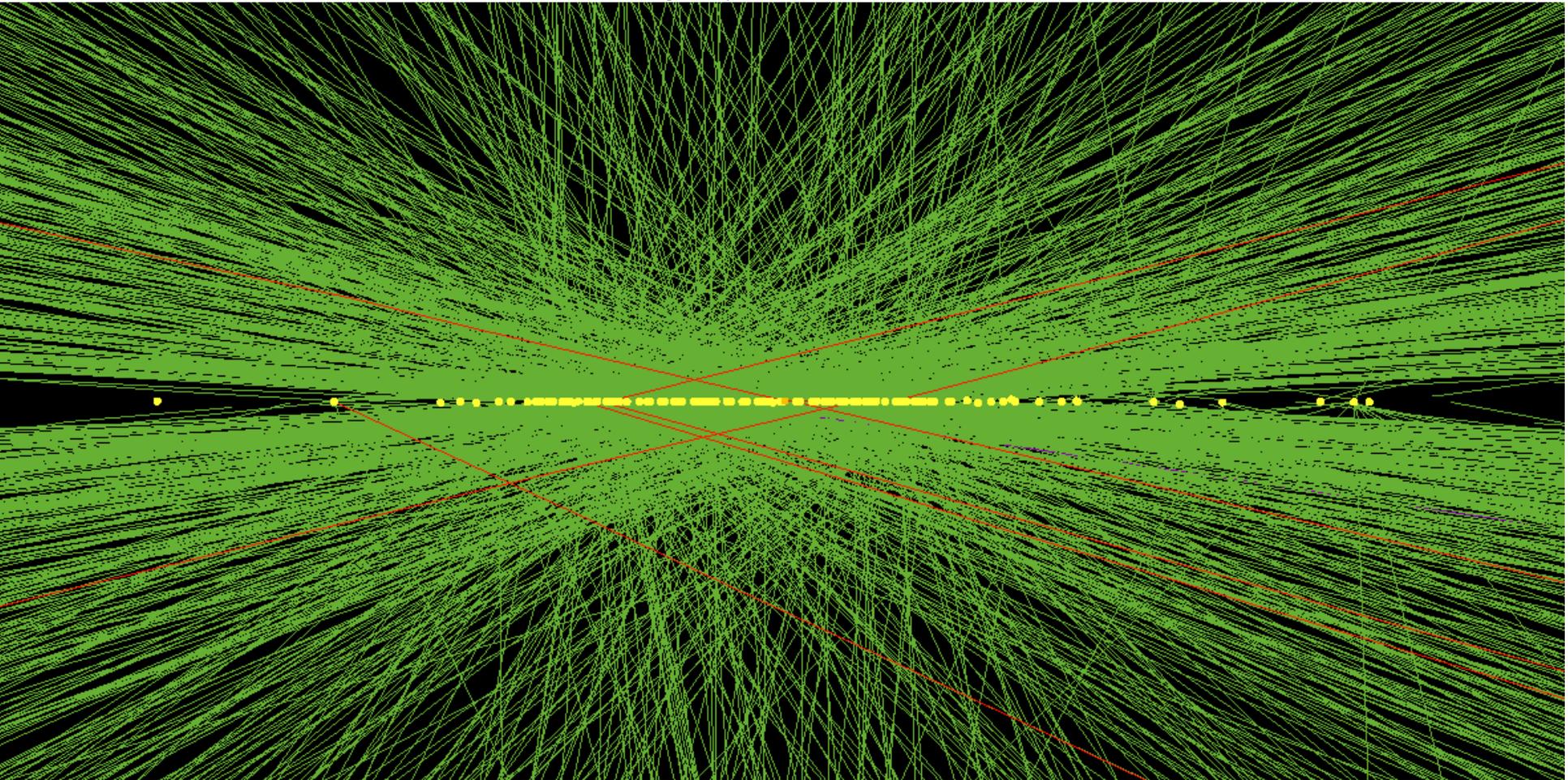


$Z \rightarrow \mu\mu$  decay with 25 vertices (April 15th 2012)



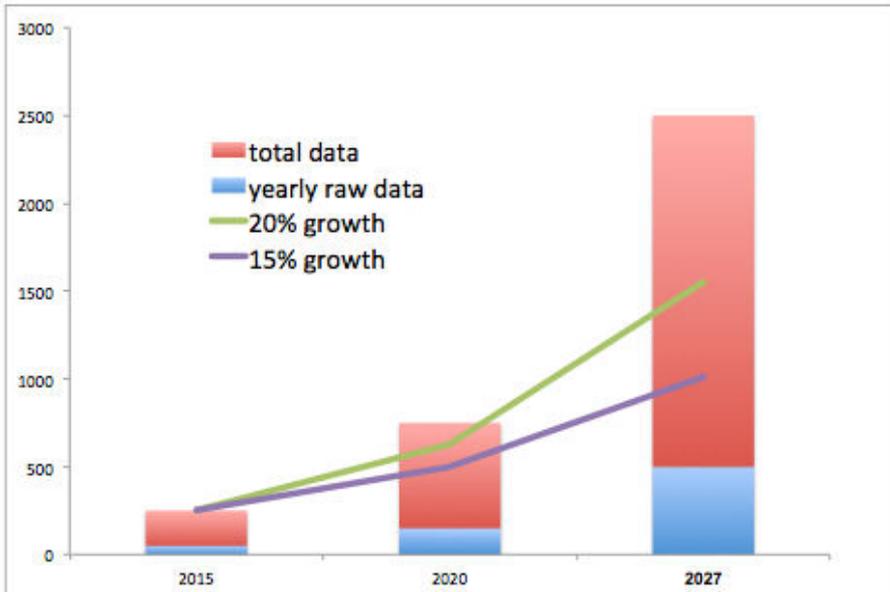


- Simulated Event Display at 140 PU (102 Vertices)

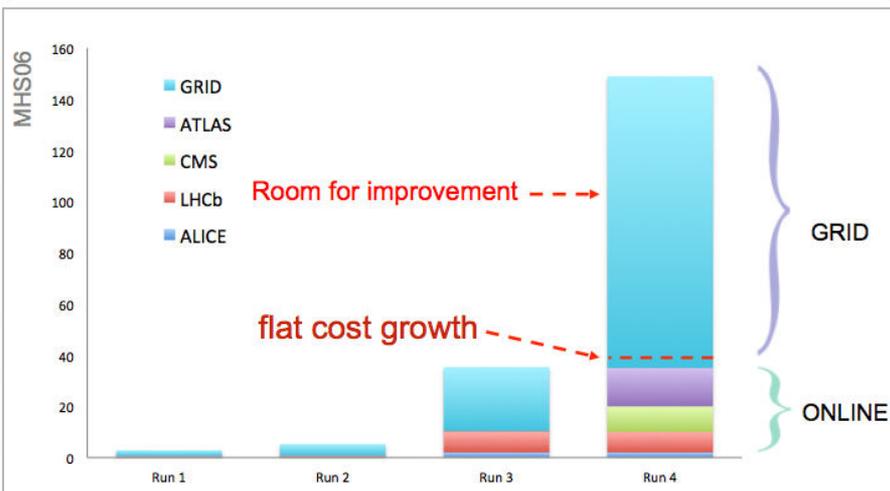




# The Challenge



**Volume:** >200 PB at present (on disk) will grow by factor of >10x by Run-4.



**Complexity:** Pile up increasing from 23 to 140 by Run-4 increases computational problem super-linearly.

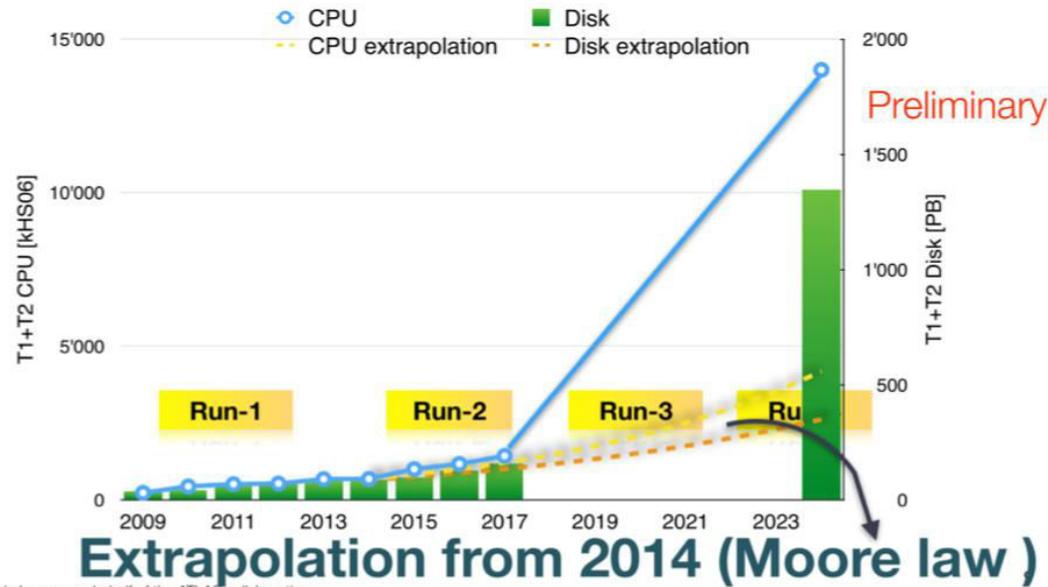


# The Challenge



## ATLAS: projections Run-4 (with 2014 performances)

ATLAS resource needs at T1s & T2s



Eric Lançon on behalf of the ATLAS collaboration

- Need to worry about disk and CPU usage for HL-LHC as well as access to disk (IO and capacity!).



# How to meet the future requirements?

- W
- 
- W
- 
- G
- fr
- 

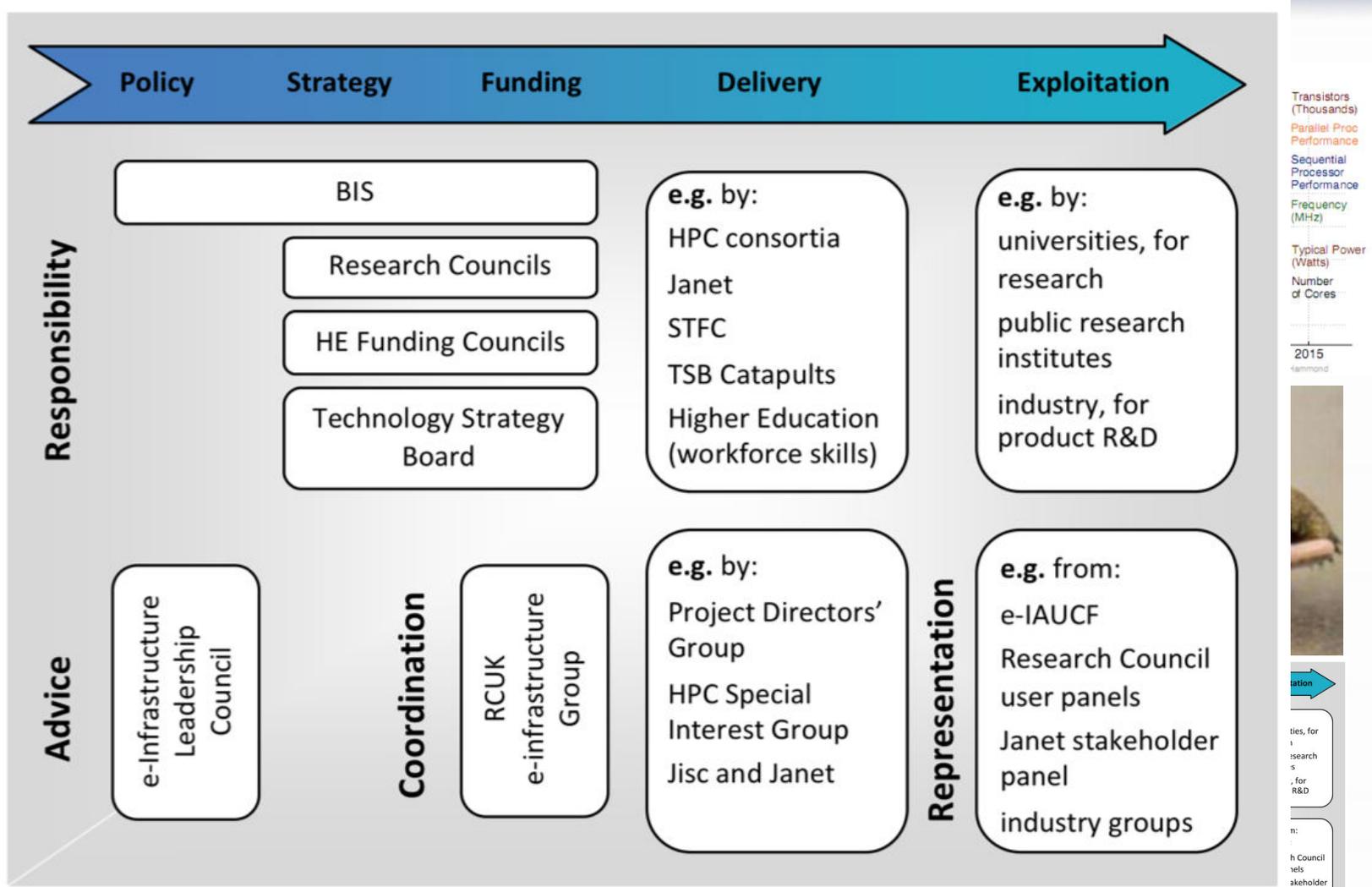
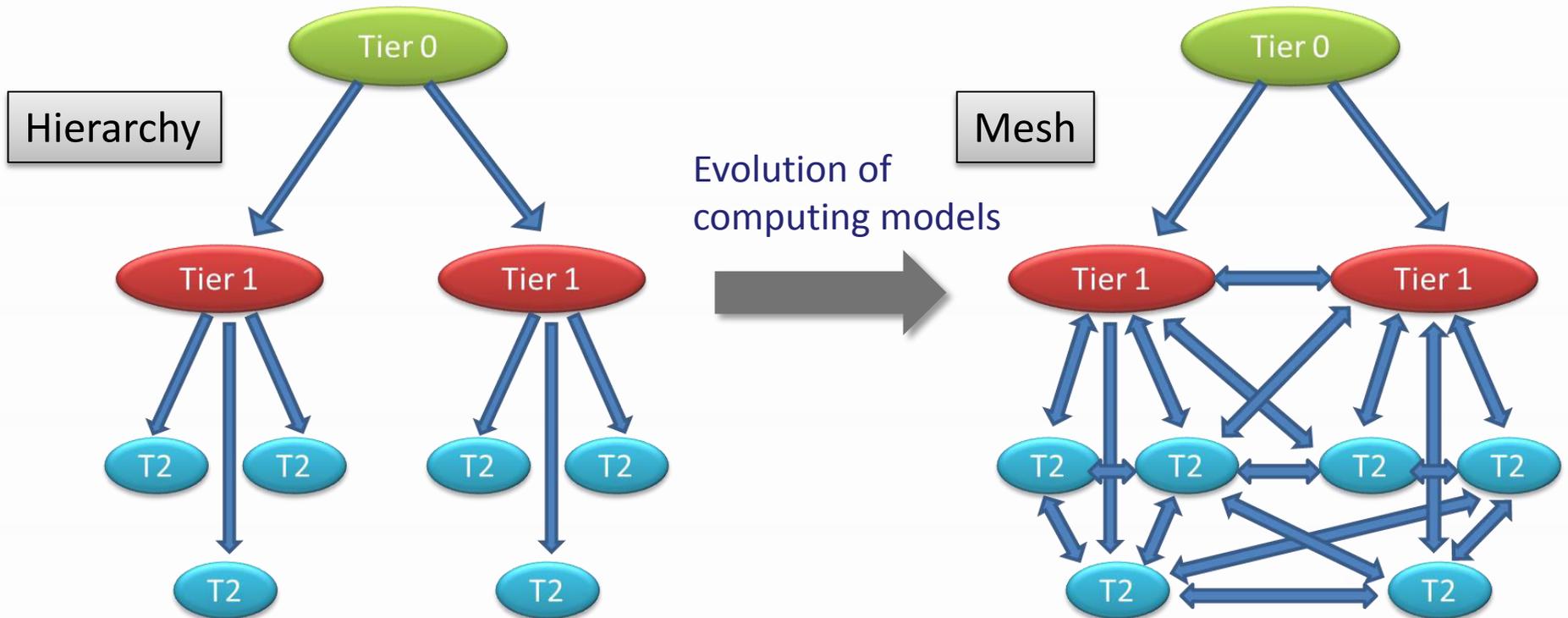


Figure 8: The e-infrastructure pipeline



# Working Smarter



- Network capabilities and data access technologies have significantly improved our ability to use resources independent of location.

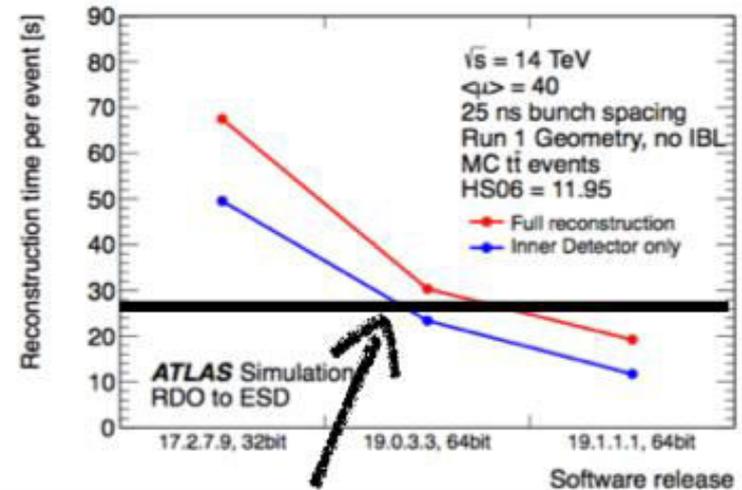
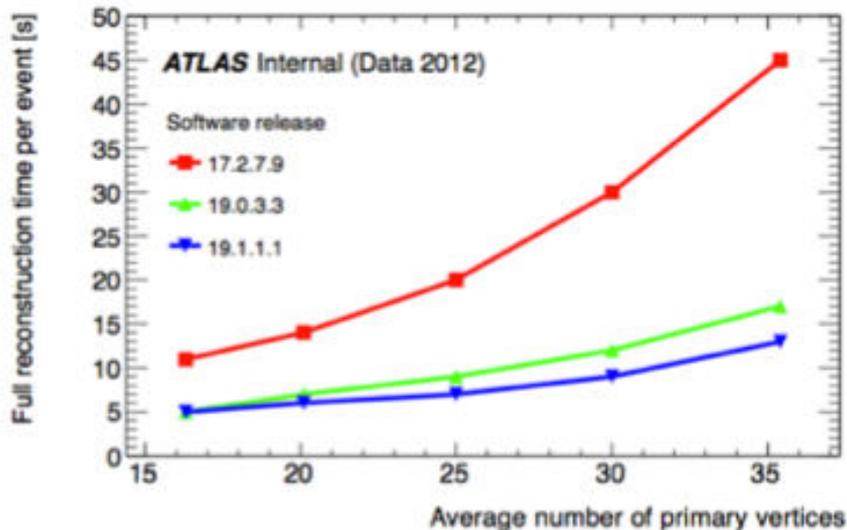


# Improvements: ATLAS Example

## Run 2 Challenges: Reconstruction

(Graeme Stewart)

- Run 2 sees the HLT output rate rise to 1kHz
  - Need to maintain single lepton triggers
- Tier-0 budget is around 11k cores

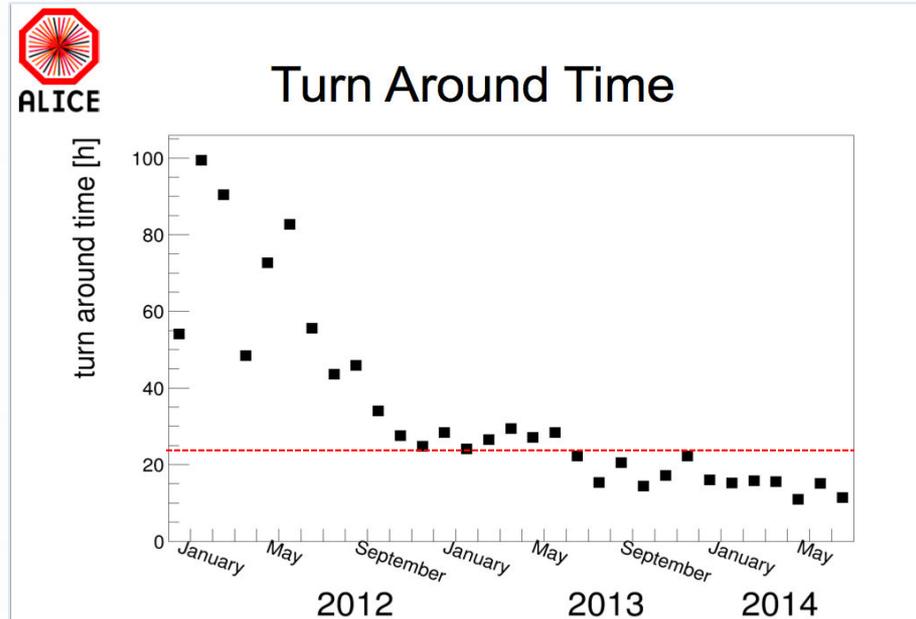
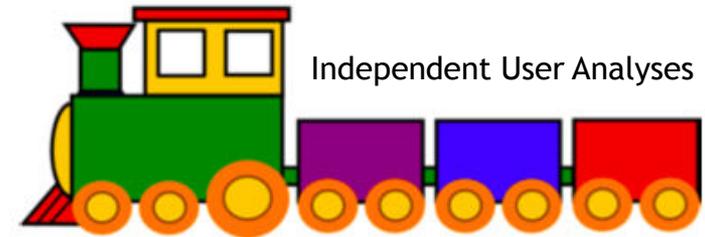


Tier-0 budget ( $\mu=40, 0.4$ )



# Analysis Trains

- In Run-1, we read the data 50x to produce 50 derived data-sets!
- Combine separate user analyses into a Centrally managed ‘Train’.
- Each ‘carriage’ is tested before departure and removed if it fails.
- Central book-keeping makes derived data more accessible to all in a standard way.
- Can be multiple trains (eg Fast and Slow).
- Reduces “chaotic” user analysis.

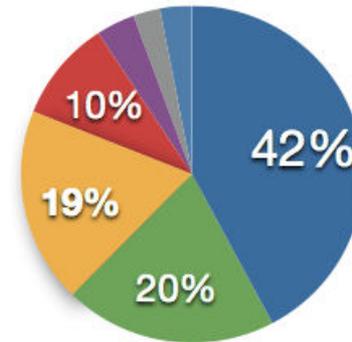




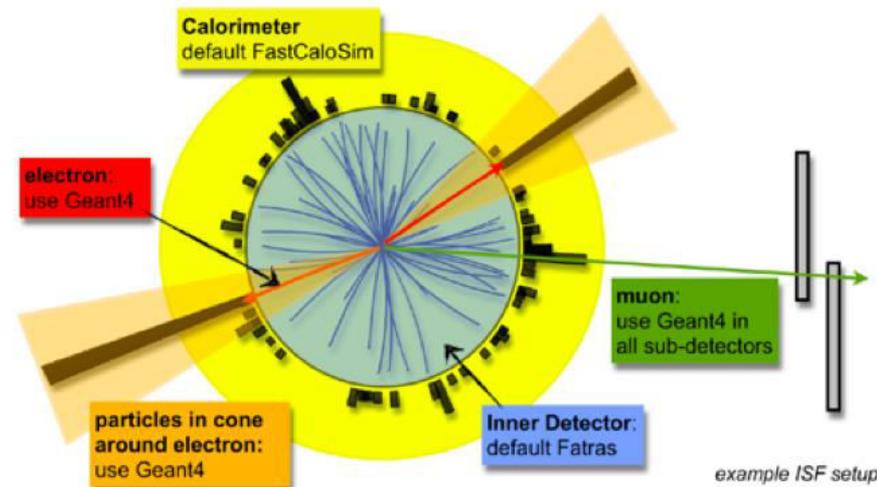
# Monte Carlo

- GEANT-4 full simulation is very expensive: e.g. up to 1000s/evt for ATLAS.
- Fast simulations use techniques such as Smearing, Frozen-Showers, Parametric Responses but are not good enough for all situations.
- Development of an Integrated Simulation Framework allows an appropriate mixture of Fast and Full in the same event and x100 speed-ups are possible.

## GRID CPU Consumption

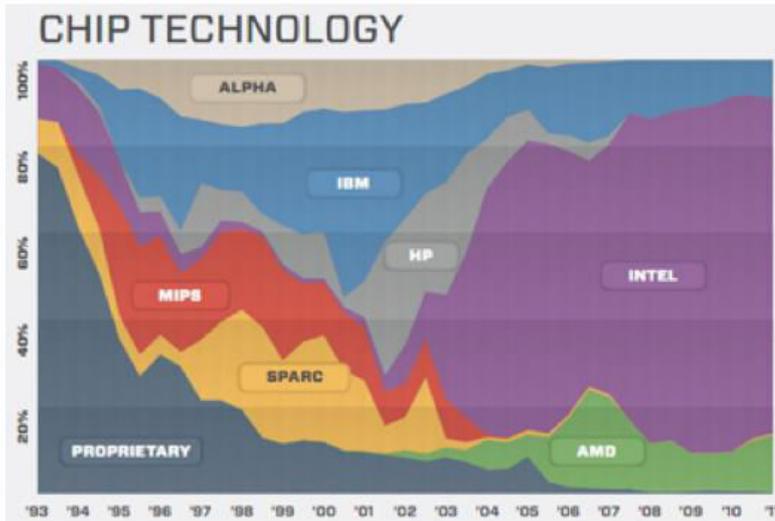


- MC Simulation
- MC Reconstruction
- Final Analysis
- Group Production
- Group Analysis
- Data Reconstruction
- Others

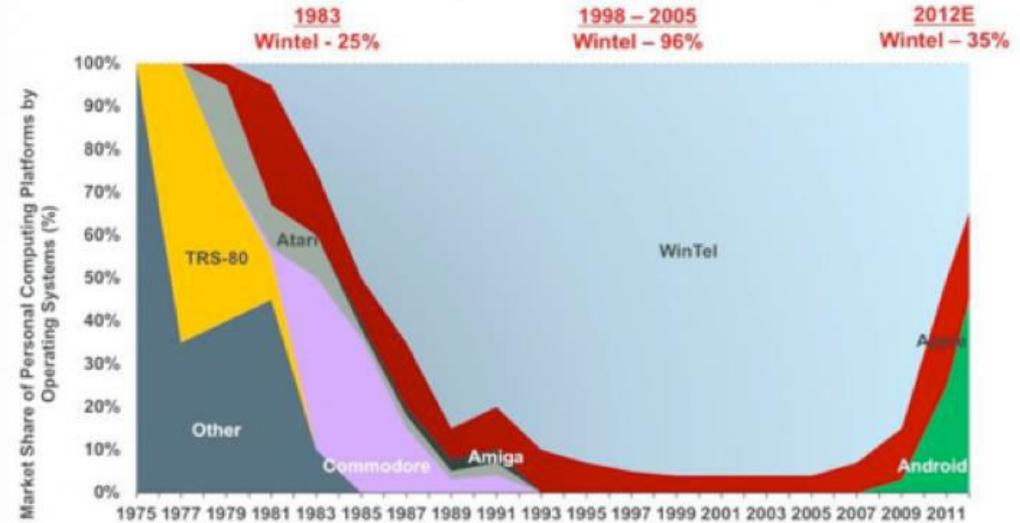




# Hardware Trends



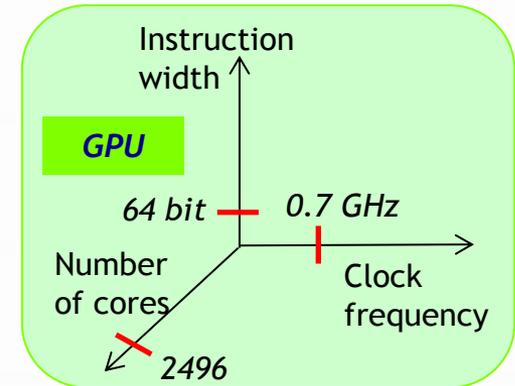
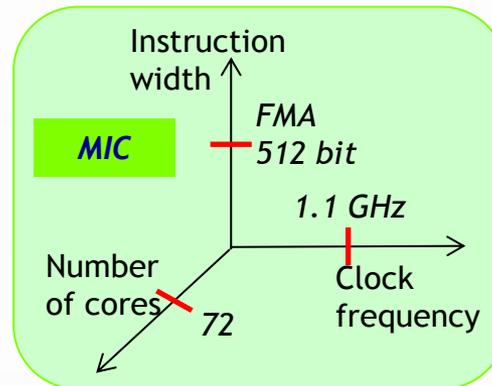
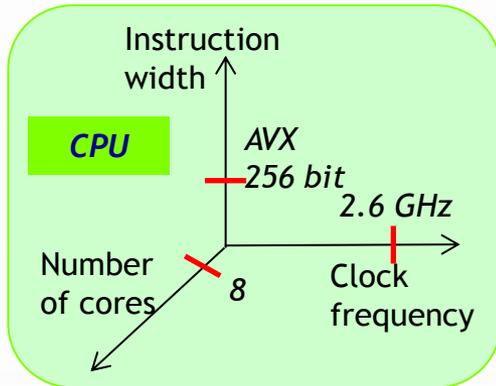
Global Market Share of Personal Computing Platforms by Operating System Shipments, 1975 – 2012E



- HEP utilises commodity hardware which is driven by external forces.
- Key driver is now power-consumption for both:
  - Portable devices, which are then supported by *Reduced clock speed*
  - Backend data-centre machines providing compute and data services.
- X86 mono-culture is breaking down.
  - ARM64, PowerPC, GPUs, etc.



# Hardware Directions



DRAM		Multi-GB Main Memory	200 cycles
Level 3 Cache		8MB Cache (shared)	40-70 cycles
Level 2 Cache		1MB Cache	10 cycles
Level 1 Cache		32kB (Data and Instruction)	4 cycles
Core	Core	... 32 x 64bit registers	1 cycle

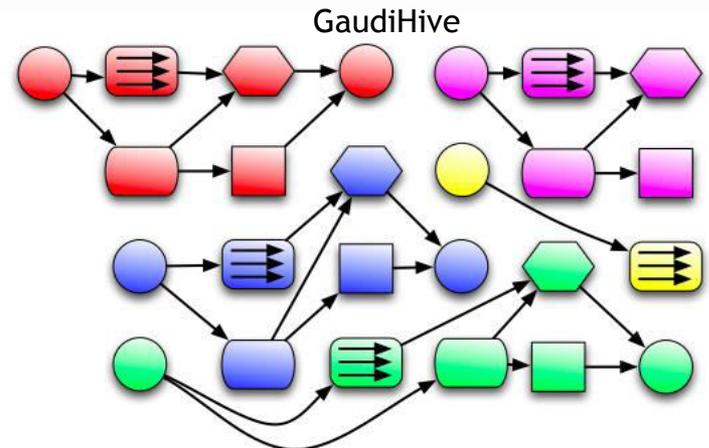
Costly to move data from main memory - can far outweigh benefits of SIMD. The software design needs to understand what data is needed in the caches. No abstract solution works for all data on all hardware.

# Using More Cores

- Xeon Grid Servers: 2-4 GB/core
  - Hyper-threaded (2): 1-2 GB/core
- Xeon-phi (MIC): 256 MB/core
  - Hyper-threaded (4): 64 MB/core
- Tesla K40 (GPU): 4MB/core

- Reduce memory use via “threading”
- Memory savings can be huge (“heap” is shared)
- But programming more difficult (races, deadlocks etc)
- Back-porting to millions of lines of legacy code hard.

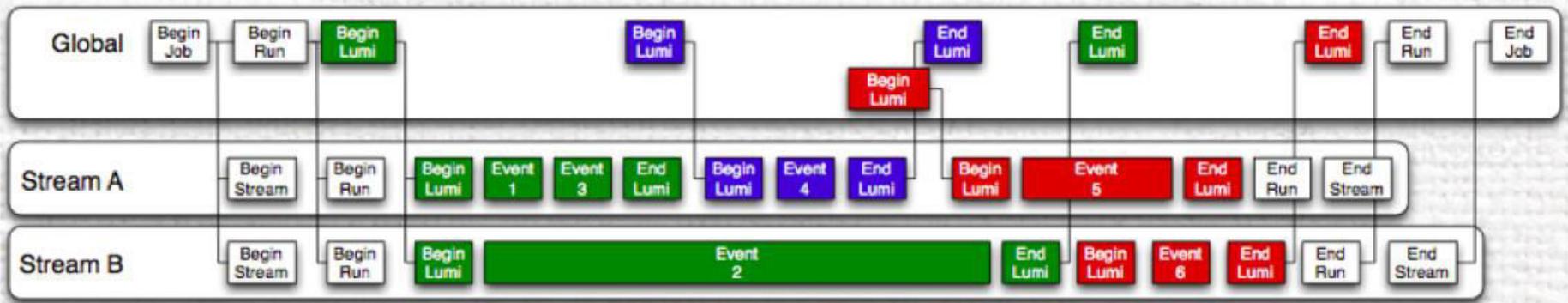
“Memory Wall”  
Typical HEP job requires  
~2GB memory. Can’t  
simply run independent  
jobs on each core.



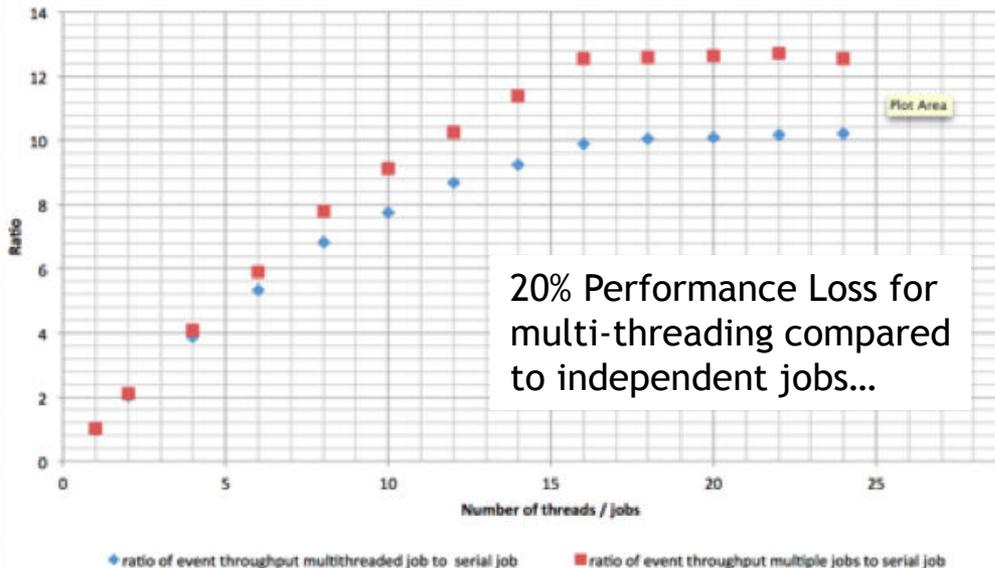
Colours represent different events,  
shapes different algorithms



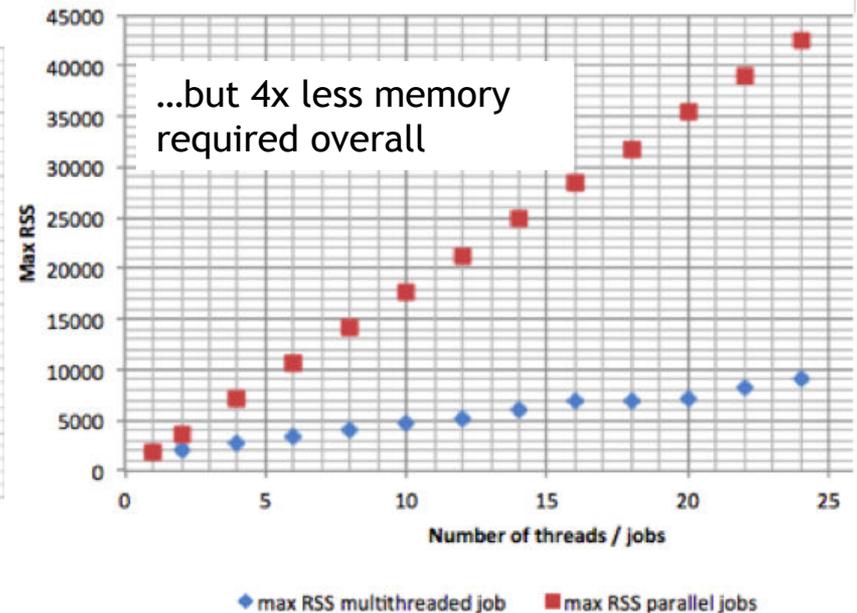
# CMS Multi-Threading



Ratio of throughput

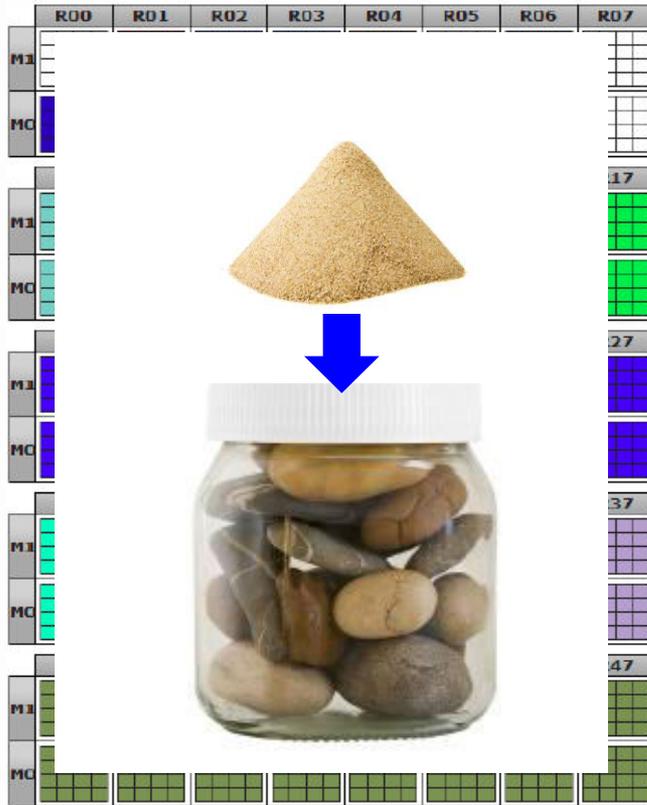


Max RSS





# Getting Help



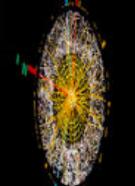
“Full” HPC machine: Only 85% utilised because shortest job in queue requires larger partition than is available.

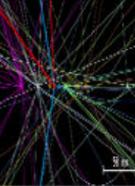




## Pilot stress tests on Titan

- **In May 2014 we ran the first 24 hour continuous job submission test via PanDA@EC2 with pilot in backfill mode, with MPI wrappers for two workloads from ATLAS and ALICE**
  - Stable operations
  - ~22k core hours collected in 24 hours
  - Observed encouragingly short job wait time on Titan ~4 minutes
- **Ran second set of tests in July 2014, with pilot modifications that were based on information obtained from the first test**
  - Limit on number of nodes removed in pilot
  - Job wait time limit introduced – 5 minutes
  - 145763 core hours collected
  - Average wait time ~70 sec
  - Observed IO related effects that need to be understood better
- **Final tests have been conducted in August**
  - Were able to collect ~ 200,000 core hours
  - Max number of nodes per job – 5835 (93360 cores)
    - Close to 75% ATLAS Grid in size!
  - Used ~2.3% of all Titan core hours or ~14.4% of free core hours





Our community are involved in many of the new UK/European/Global initiatives:

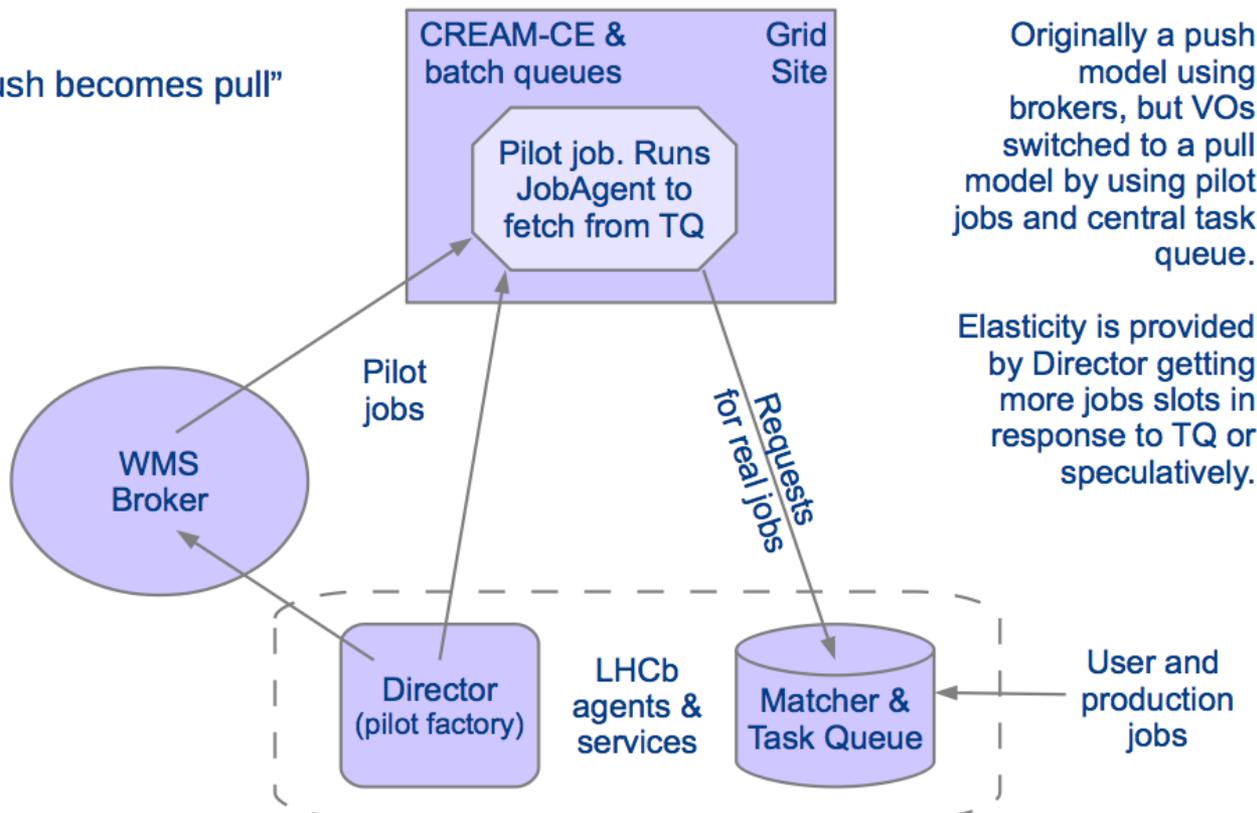
- EGI Engage (H2020)
- VLData (H2020)
- Zephyr (H2020)
- HEP Software Foundation
- EU-T0 Initiative
- ...and more.





## The Grid

“Push becomes pull”



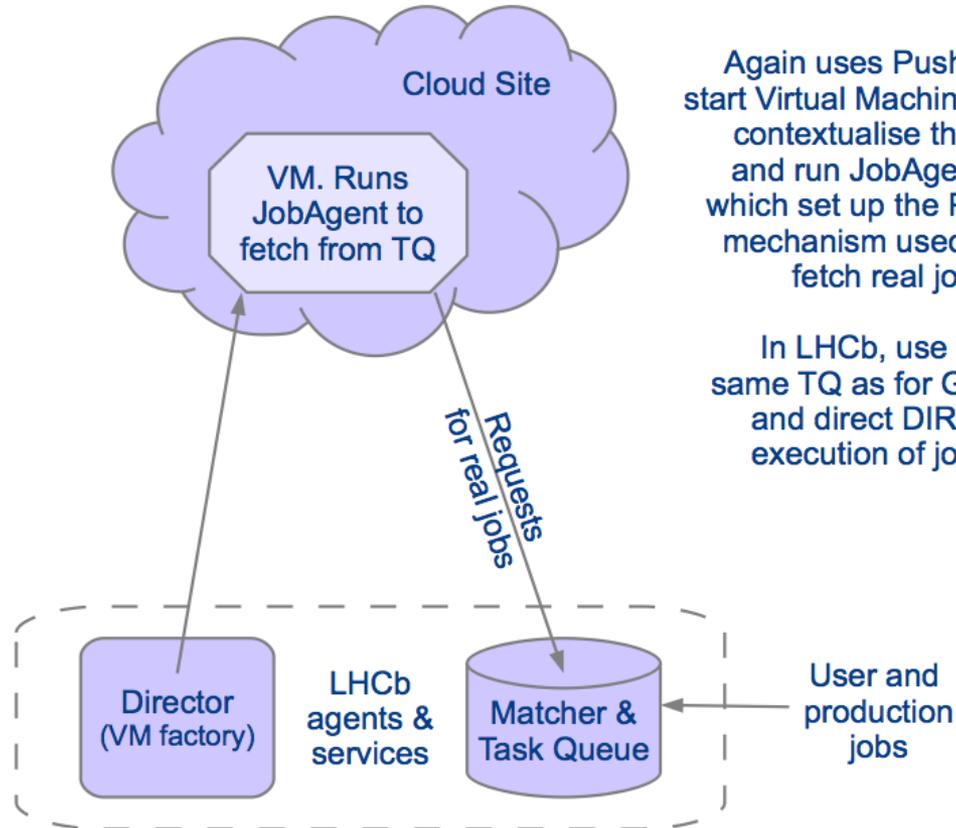
Originally a push model using brokers, but VOs switched to a pull model by using pilot jobs and central task queue.

Elasticity is provided by Director getting more jobs slots in response to TQ or speculatively.



## The Cloud

Elasticity is provided by Director's ability to request more VMs at sites, either in response to TQ or speculatively.



Again uses Push to start Virtual Machines, contextualise them and run JobAgents which set up the Pull mechanism used to fetch real jobs.

In LHCb, use the same TQ as for Grid and direct DIRAC execution of jobs.

Grid sites increasingly using Cloud Technology internally to manage resources (e.g. CERN T0 split between Geneva and Budapest). But cloud interface is not typically exposed externally.

Grid of Clouds?



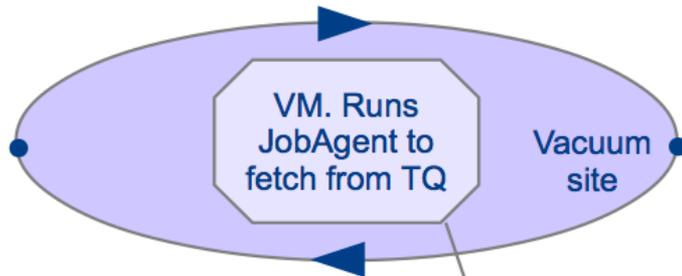
# Accessing Resources

## “The Vacuum”

Instead of being created by VOs, the Virtual Machines appear spontaneously “out of the vacuum” at sites.

As with the other models, the JobAgent runs and requests real jobs from the Matcher and normal Task Queue.

This is similar to DIRAC on the existing LHCb online farm, but with VMs added.



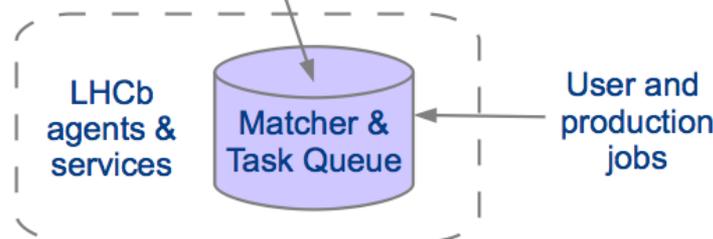
Hypervisors/hosts can run VMs for particular VOs depending on work available and target shares for each VO.

Elasticity comes directly from matching resources to demand from TQ.

VMs created and contextualised by the sites themselves.

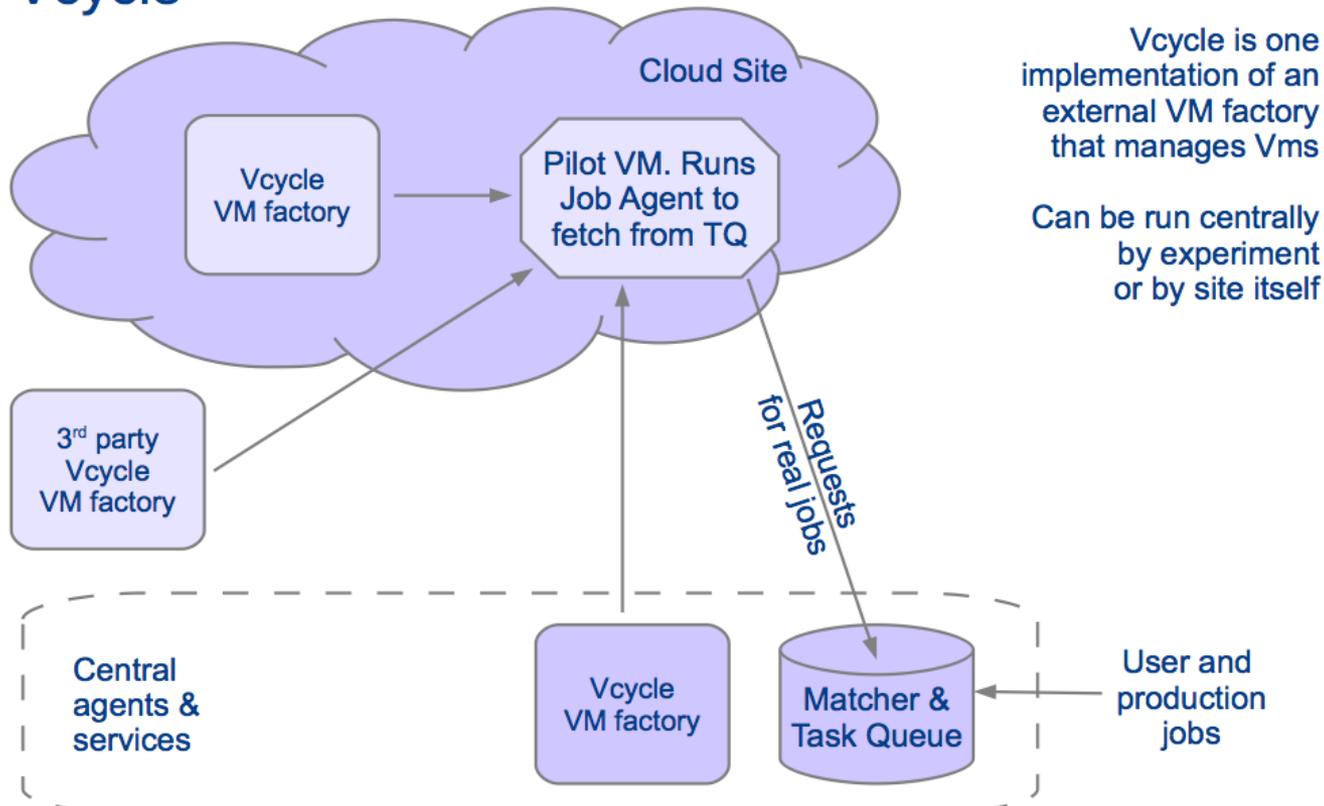
Suitable for sites dedicated to a few VOs.

Containers instead of Virtual Machines?





## Vcycle



Vcycle is one implementation of an external VM factory that manages Vms

Can be run centrally by experiment or by site itself

Applies VAC idea to Cloud Resources (OpenStack).

Can be run by site, experiment, or 3<sup>rd</sup> party.

Could we use VAC or Vcycle to flexibly integrate Tier-3 resources?



- HEP computing resource needs are large and will continue to grow to meet the volume and complexity of LHC data.
- Evolution of Computer Hardware is making it harder to realise the ‘Moore’s Law’ type growth.
  - We can cope with Run-2 until 2020
  - But we need to prepare for runs beyond that
  - A lot of work ongoing/needed to adapt frameworks
  - More work is needed; e.g. within Algorithms.
  - Work → Manpower which is increasingly expensive c.f. hardware
- Joining up the e-infrastructures can help.
- Evolving the Grid to use Cloud technologies where appropriate.