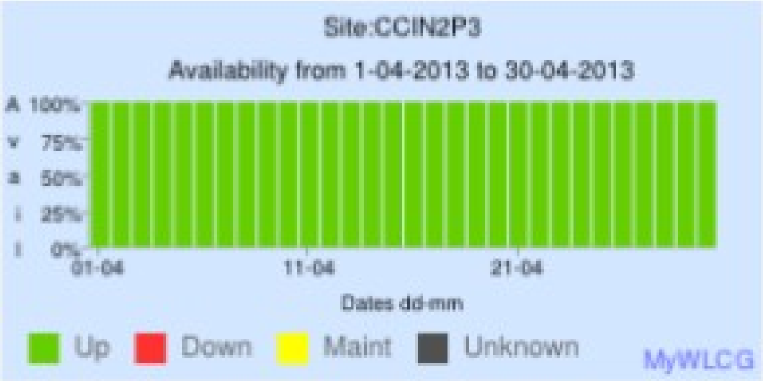
Reunion T1-AF

(16.05.2013)

Renaud Vernet

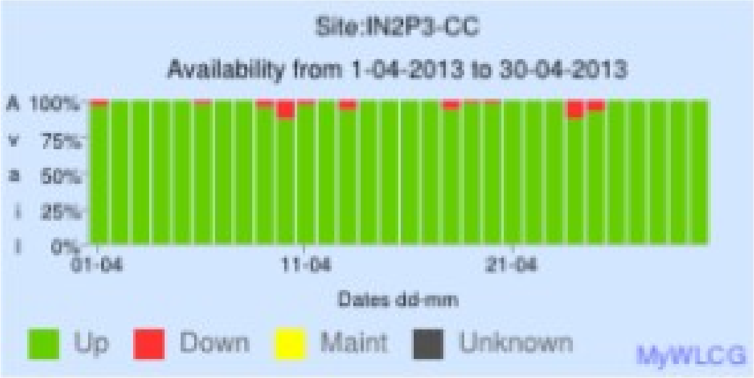
Disponibilite du site (avril)

ALICE



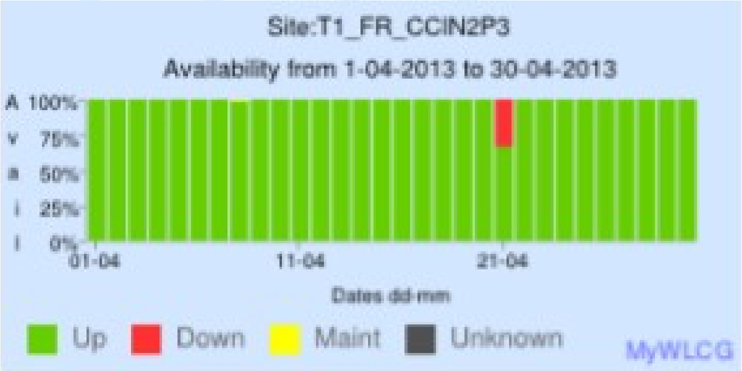
CCIN2P3 Avail: 100% Unkn: 0%

ATLAS



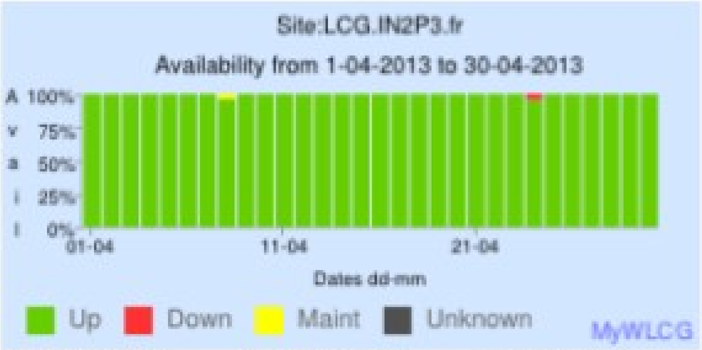
IN2P3-CC Avail: 98% Unkn: 5%

CMS



T1_FR_CCIN2P3 Avail: 99% Unkn: 0%

LHCb

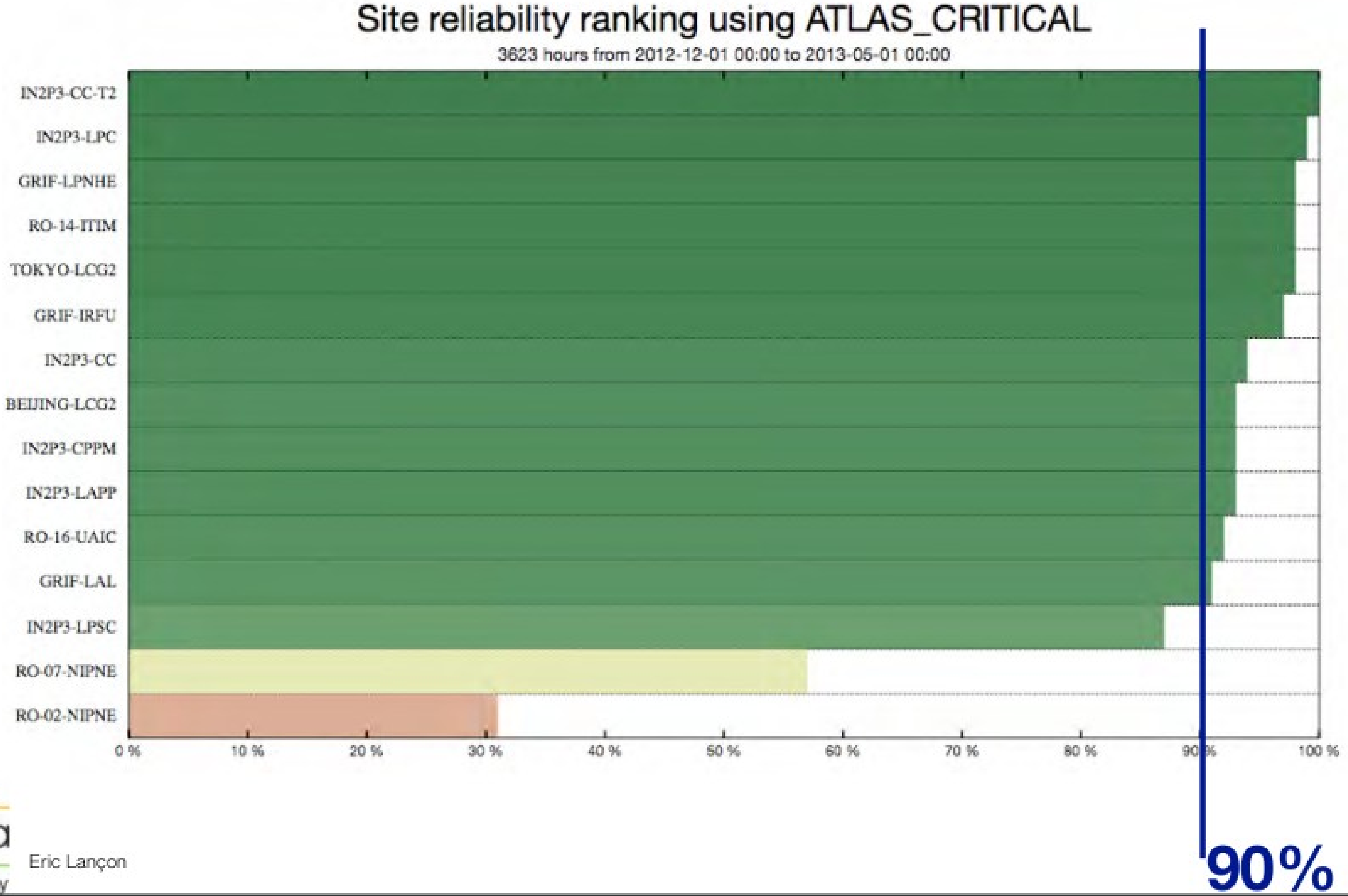


LCG.IN2P3.fr Avail: 100% Unkn: 0%

E. Lançon : prospectives ATLAS

Should be ~100%

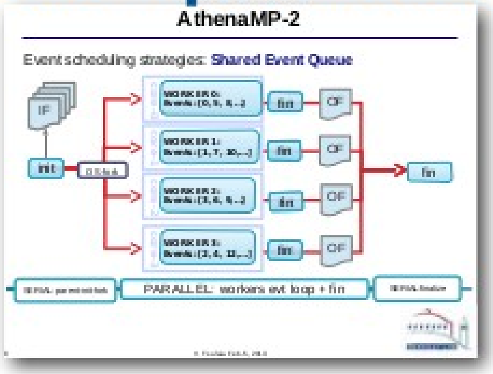
Reliability of FR-cloud sites



Software changes

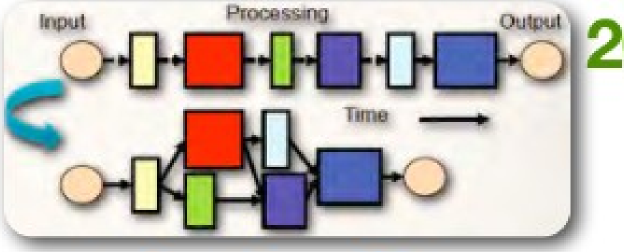
- All experiments embarked in profound software changes
- Geant team as well…
- Reduction of memory footprint
- Revision of data models
- Multithreating (memory sharing)
- Vectorisation (to exploit new architectures)
- I/O is a major concern

event parallelism



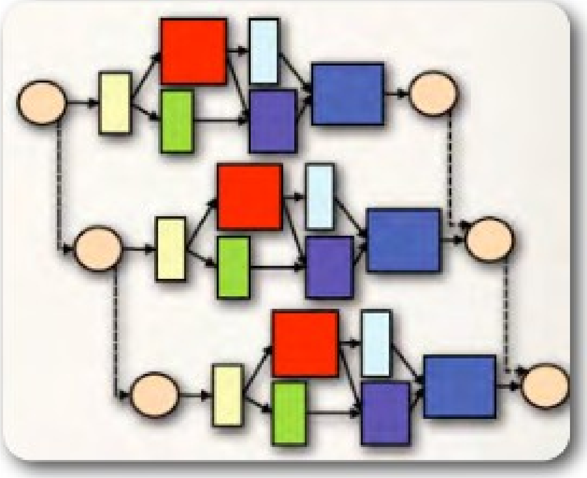
Today

algorithm parallelism



2015-2016 ?

event & algorithm parallelism



LS2 or before?

PRACE, the Partnership for Advanced Computing in Europe

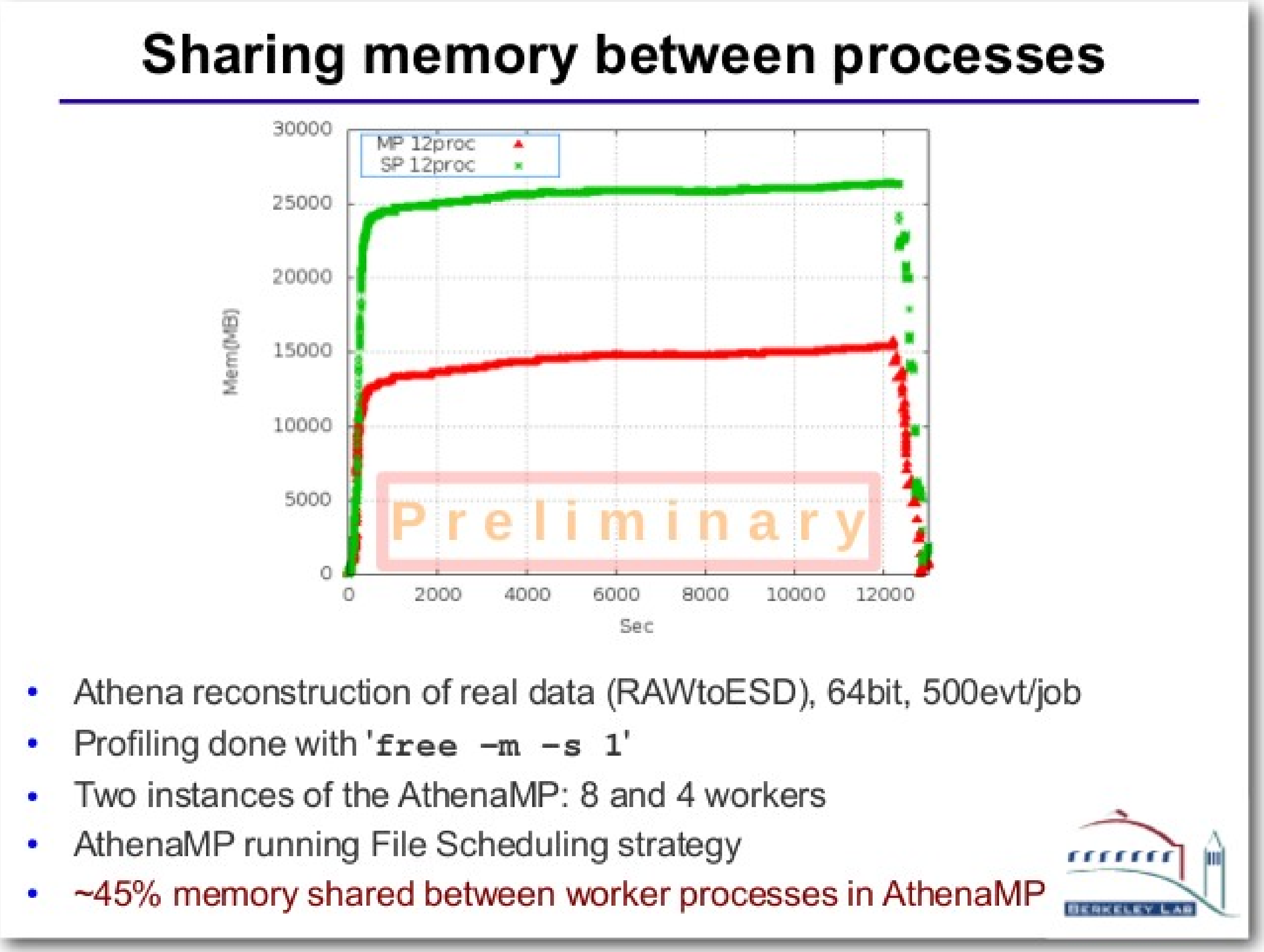
SuperMUC a PRACE Tier-0 centre :
155,000 Sandy Bridge cores, 2.8M HS06

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

Opportunistic use of HPC (High-Performance Computing) resources

- Latest competitive supercomputers are x86 based (familiar linux cluster)
- ATLAS & CMS projects to use idle CPU cycles at HPC centers in US (Argonne, San Diego) & DE (Munich)
- Demonstrators working for simulation
- Difficult to use HPC centers for I/O intensive applications
- Outbound connectivity of HPC centers may be an issue

Memory gain ~45% for 12 processes



Sharing memory between processes

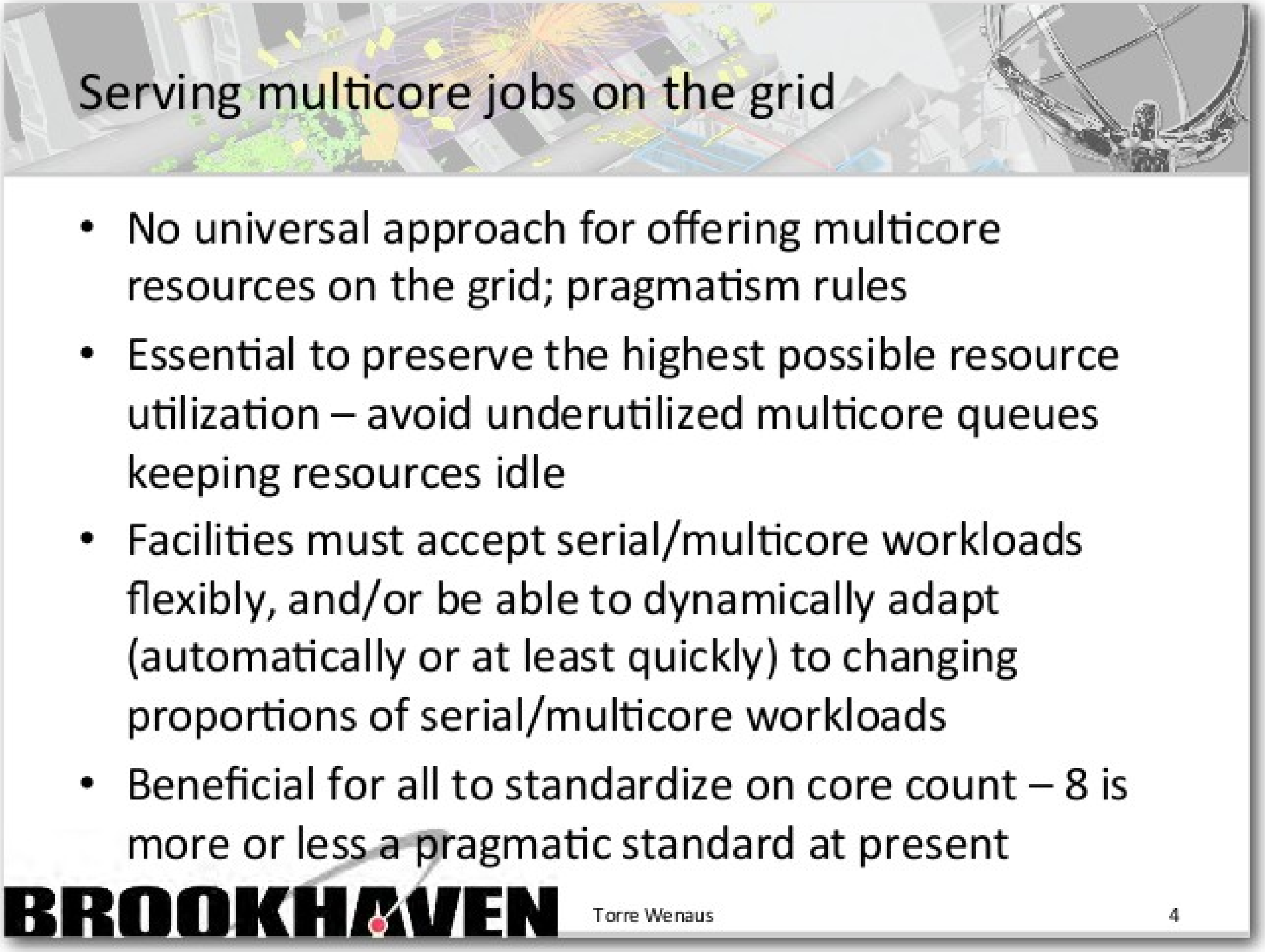
- Athena reconstruction of real data (RAWtoESD), 64bit, 500evt/job
- Profiling done with '`free -m -s 1`'
- Two instances of the AthenaMP: 8 and 4 workers
- AthenaMP running File Scheduling strategy
- ~45% memory shared between worker processes in AthenaMP

Multi-core : how to optimize resources?

- Distinct multi-core queues at sites (ncore in queue setup)
    - Real site is partitioned ?
    - How to optimize site resources? done by PanDA or by site ?
    - Output merging serial job
- Many questions remain open before going to production
- However having test queues at site (with limited number of WNs behind) help both ATLAS and site to understand the issues

No recipe yet

T. Wenaus March S&C week



Serving multicore jobs on the grid

- No universal approach for offering multicore resources on the grid; pragmatism rules
- Essential to preserve the highest possible resource utilization – avoid underutilized multicore queues keeping resources idle
- Facilities must accept serial/multicore workloads flexibly, and/or be able to dynamically adapt (automatically or at least quickly) to changing proportions of serial/multicore workloads
- Beneficial for all to standardize on core count – 8 is more or less a pragmatic standard at present

BROOKHAVEN Torre Wenaus 4

Renaming status

- A few T2s have already been completely migrated, in particular LAPP in the FR cloud (big thanks to Eric Fede for the debugging).
- Renaming at T1s is on-going at FZK and Lyon. Hope to have the 2 sites migrated before the end of May.
- We need more T2s to volunteer for the migration, in particular DPM sites :
  - How to deploy WebDAV on DPM is summarized here : https://svnweb.cern.ch/trac/lcgdm/wiki/Dpm/WebDAV/Setup
  - DPM 1.8.6 and lcgdm-dav-0.12.1-2 are needed.
- The plan is to have every sites migrated by the end of the year.

A. Filipcic : use of HPC for ATLAS

The Computing Challenge

- Computing sites are part of institutes which collaborate in ATLAS experiment
  - Some sites (10 Tier-1 centers) are big and provide additional services, tape storage
- ATLAS uses many additional resources, many of them are external
  - National computing centers
  - Cloud providers (Amazon, Google, academic clouds)
  - Supercomputer centers (US, France, Germany, Norway, Sweden)
- Opportunistic resources enabled the big success in 2012 – the Higgs discovery

Job properties

- High I/O jobs (20GB input, few GB output), relatively fast (up to few h), high memory (2-4GB)
  - Monte-Carlo and real-data reconstruction
  - User analysis
- Low I/O (<100MB), very long jobs (>1 day), low memory (<1GB)
  - Event generation
  - Monte-Carlo simulation (10 min /event)

Problems of trivially-parallel jobs

- Long running times of some job classes
- Inefficient memory usage (shared memory)
- So, ATLAS is developing some new ways:
  - AthenaMP – fork after initialization to share ½ of job memory among the processes
  - Event distribution: Master job + slave event processing jobs (good for opportunistic usage due to frequent commits)
  - Multi-threaded jobs on single event level (parallel tracking, algorithm parallelization)
- The complication is not artificial: The LHC operation at higher energies in 2015 will require much more demanding jobs which cannot be done in the current way efficiently.
- HPC modus operandi is becoming a necessity for ATLAS, but for further development not many ATLAS computing sites can provide the required capabilities

What does a site need to run ATLAS jobs

- Many complex scenarios, but the simplest one is:
- Provide one dedicated server with external connectivity and ARC grid middleware
  - Input file transfer to local shared file-system
  - Submission of batch jobs
  - Uploading of output files to ATLAS remote storage
- Provide node access to ATLAS software
  - Server with installed CVMFS which NFS-exports to computing nodes (/cvmfs mountpoint on the nodes)
- Provide access to ATLAS “detector calibration” databases through a local squid server
- Provide some system libraries on the nodes (depending on the OS)
  - libxml2, gcc libs, …
- Although a regular ATLAS grid site is fairly complex with custom node operating system and software, a “minimal” site can be configured with one grid server.