MANCHESTER 1824

GridPP UK Computing for Particle Physics

Multicore

Alessandra Forti

GridPP32, Pitlochry

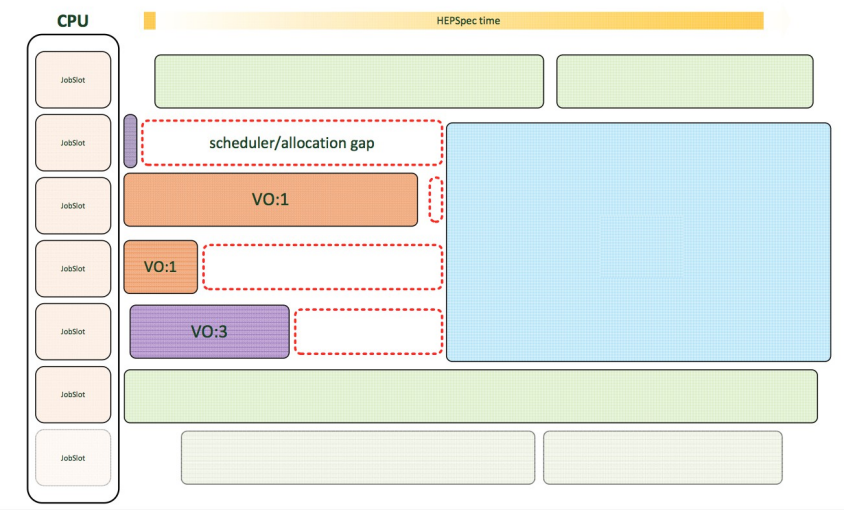
25 March 2014

Layout

- Multicore job scheduling
- Approaches: CMS and ATLAS
- Sites
- RAL&KIT
- Backfilling
- Atlas
- Accounting
- Summary

Multicore scheduling

- Multicore on the WNs is "easy"
  - Jobs request all slots on 1 node
    - No need of MPI.
- The problem is running them without wasting resources.
  - The main factor in wasting resources is draining slots for multicore jobs.

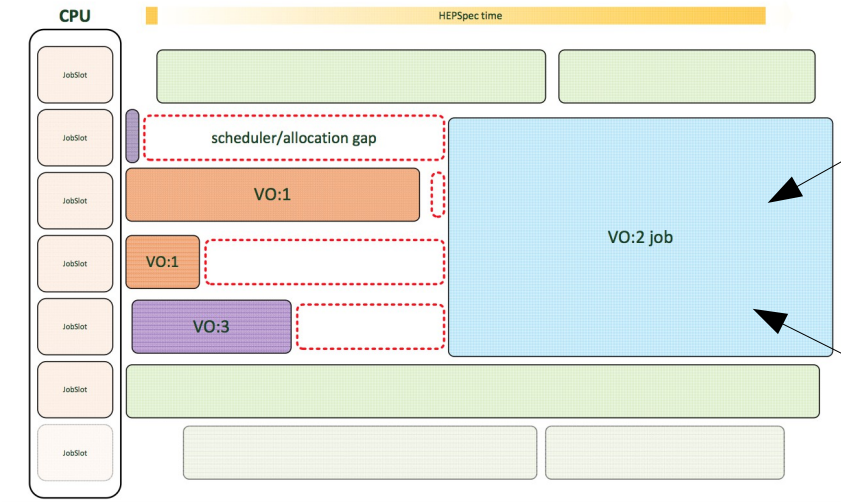


Thomas Hartmann plots

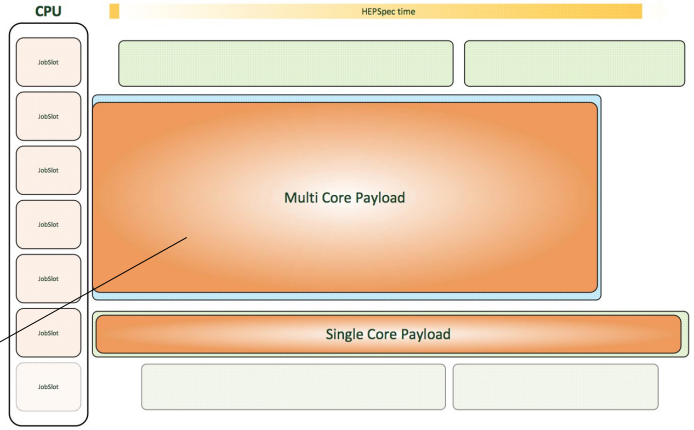
Approaches

1) Scheduling multicore is **only** a site problem
   - Just request 1 node and 8 cores in the JDL
   - Blah parser will convert the request into batch system requirements
   - Batch system will handle multicore/single core jobs race conditions
2) Scheduling multicore is **also** an experiment problem
   - Pilots run different payloads until pilot walltime is exhausted
     - Dynamic scheduling (i.e. internal backfilling) done internally
   - Aim at avoiding sites allocating dedicated resources
   - Draining still present minimised by length of pilot
     - Requires quite long pilots to be effective

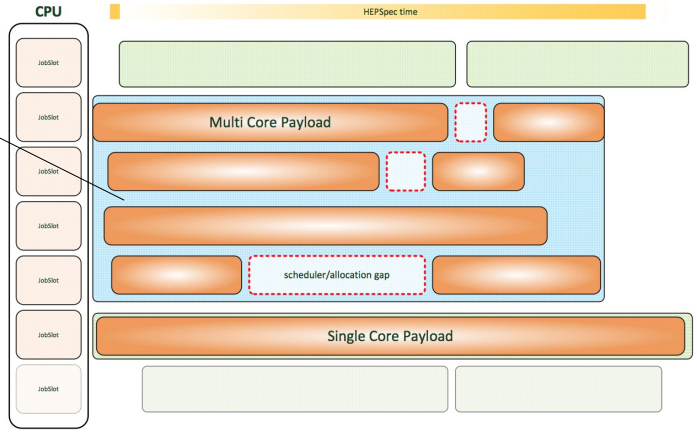
Atlas model (1)



Inside a scheduler



Atlas model

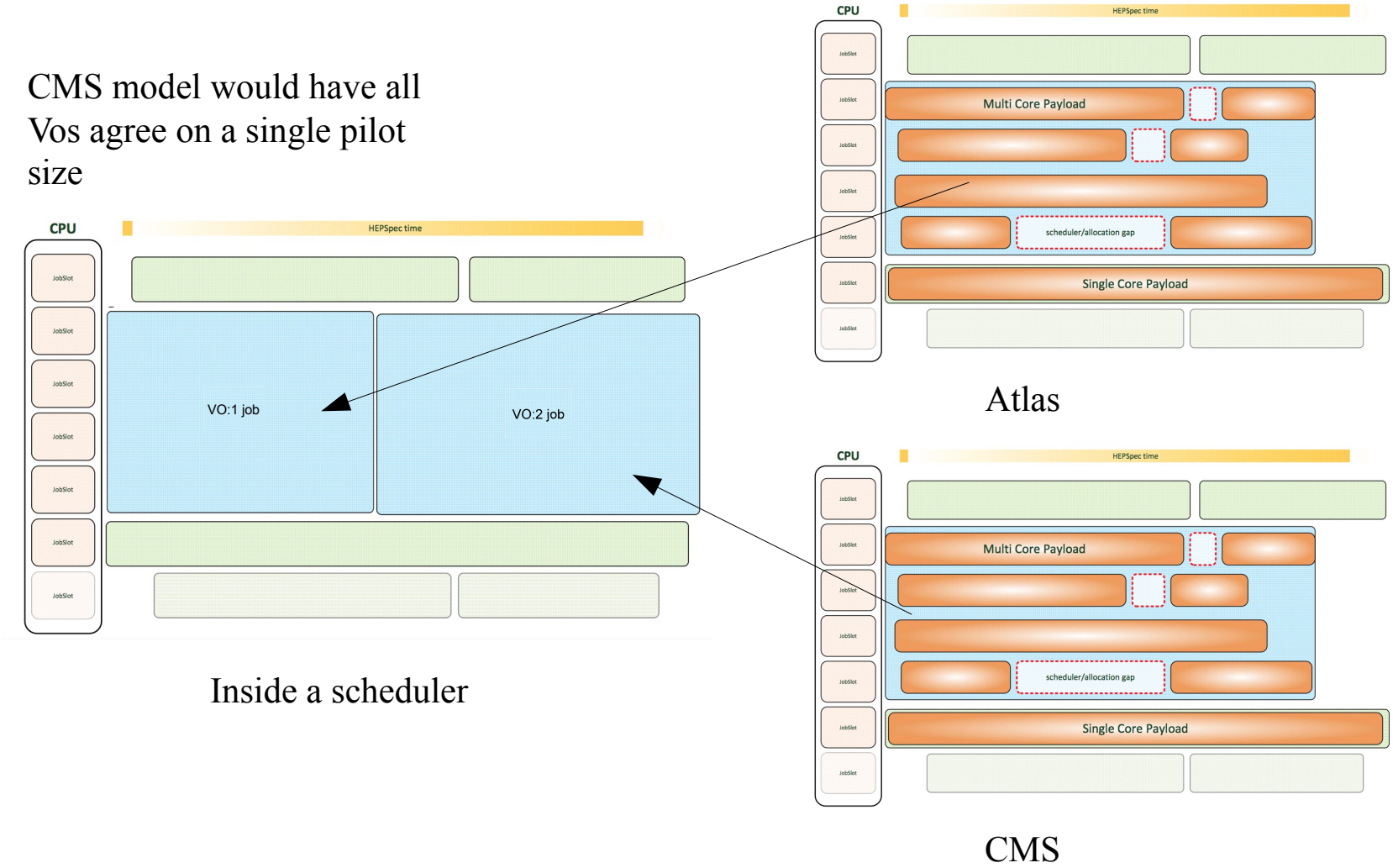


CMS model

Thomas Hartmann plots

CMS model (2)

CMS model would have all Vos agree on a single pilot size



Inside a scheduler

Atlas

CMS

Thomas Hartmann plots

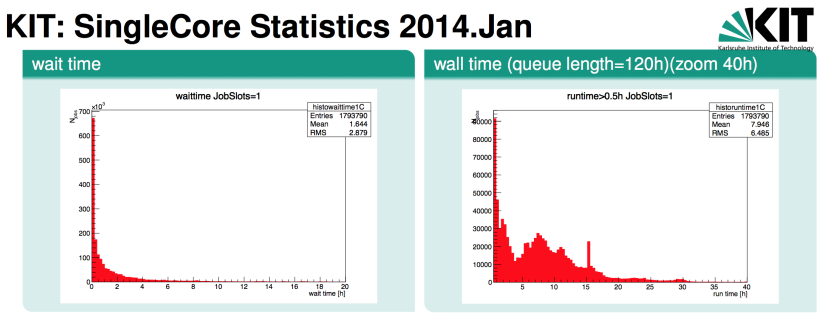
MANCHESTER 1824

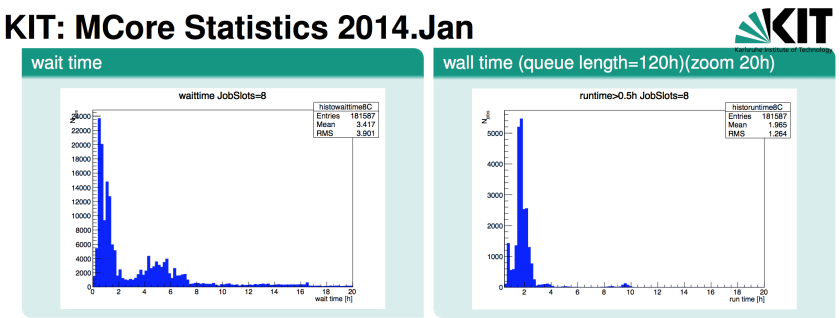
GridPP UK Computing for Particle Physics

Sites

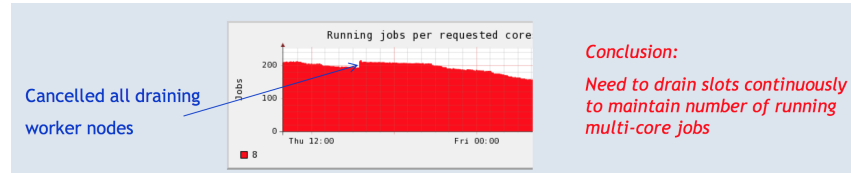
- Atlas gave a push to run mcore before Xmas
- Several sites now accept multicore jobs
  - Static dedicated resources
  - Dynamic scheduling
- Different batch systems have different approaches
  - Batch system capabilities reviewed by WLCG TF
  - Torque/maui, SGE, Htcondor first review

RAL and KIT

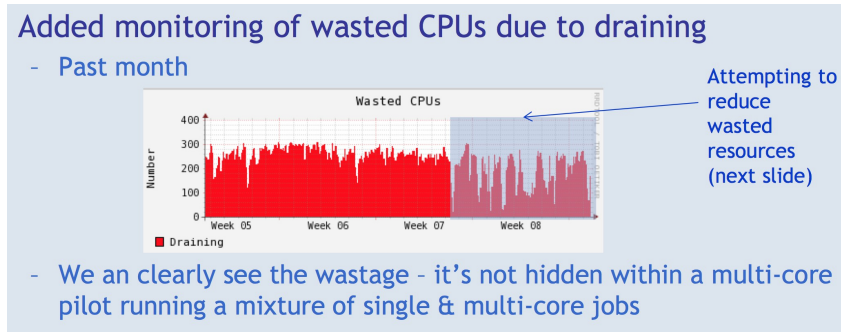




Longer wait times, shorter jobs



Need to drain constantly



Waste of CPU affected by submission patterns. Wavelike submission most disruptive

T. Hartmann, A.Lahiff plots

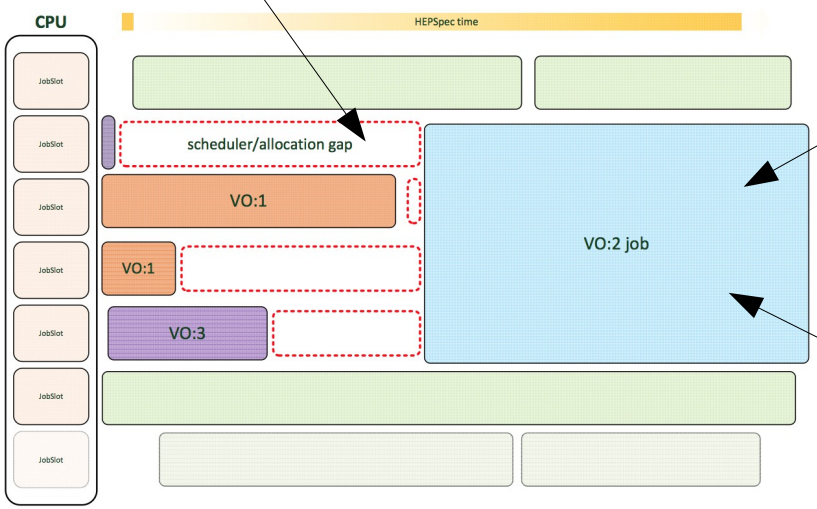
Backfilling

- Jobs of lower priority are allowed to utilize the reserved resources only if their prospective job end (i.e. the declared wallclock usage) is before the start of the reservation
  - Most batch system are designed to do this
  - Experiments for several reasons never passed parameters to batch systems
    - One reason is also that Cream CE doesn't pass the parameters.
      - A vintage problem: since 2006!
      - Only blah SGE scripts instrumented to do it
    - ARC-CE can do it

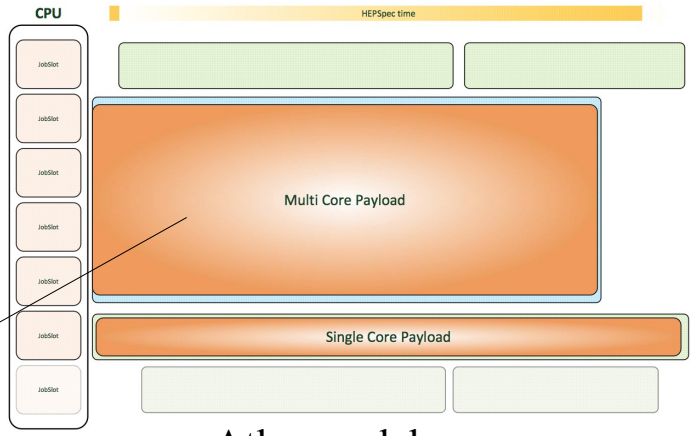
Functionality	Torque/Maui	SLURM	HTCondor	USGE/OSGE	Son of GE	LSF
Efficient Backfilling	tunable	tunable	not out-of-the-box, but similar behaviour can probably be configured	yes	yes	yes

Most sites would like....

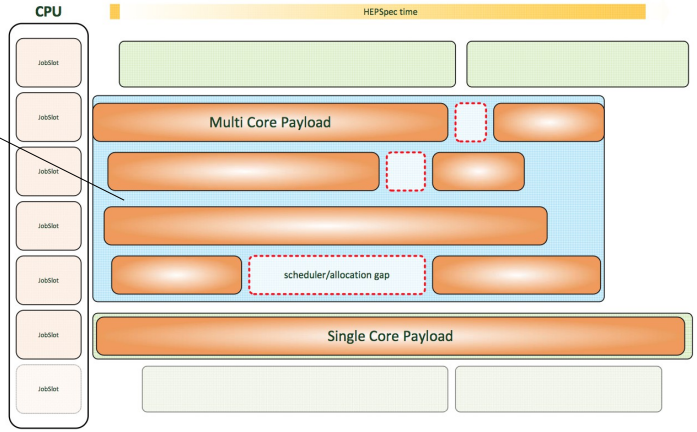
Without walltime high entropy of jobs so far was enough to fix the problem. Multicore require more organisation and backfilling with a guesstimate of walltime is needed to reduce gaps.



Inside a scheduler



Atlas model

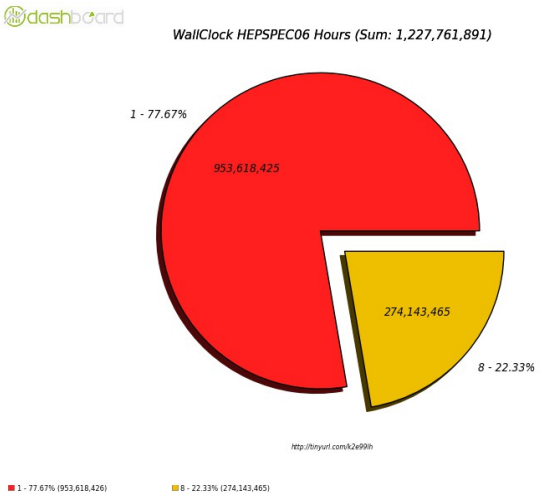


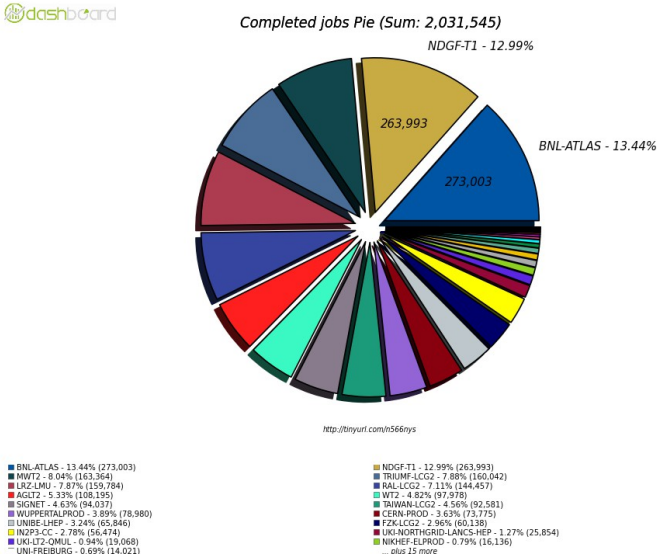
CMS model

Thomas Hartmann plots

Atlas

- 33 sites set up
  - Efficiently or not
- Mostly at T1s and US sites
  - Large and dedicated
  - Other sites contribute too though
- Quarter of MC production done on mcore





More dynamic approach

- Multicore not only use cases to pass parameters to the batch system
  - Create a panda queue every time there is the need to pass a parameter.
    - SITE_QUITESHORT_ABITMOREMEM_SOMECORES
    - ANALY_SITE_ABITLONGER_LOTSOFMEM_ALLCORES
  - Progressively messier site data structures
    - Difficult to maintain and error prone
- Task requests: walltime, memory usage, no of cores
  - Queues created dynamically
    - mcore jobs on demand on every site
    - + fast turnaround for short analysis
    - + custom jobs (high memory) can run everywhere
    - + long jobs correctly assigned to long queues

MANCHESTER 1824

GridPP UK Computing for Particle Physics

Accounting

- Still a problem in APEL with walltime and efficiency
  - Walltime not multiplied by number of cores yet
  - Correct but would like also the walltime multiplied
- EMI-3 cream CE needed to collect correct data
  - Sites haven't upgraded yet.
- Atlas accounting has been corrected
  - Currently still testing in the prototype dashboard

Summary

- Multicore is now ongoing
- Different experiments models
  - Aim is to find a compromise that maximise efficiency
  - Need testing with both experiments active
    - So far only Atlas
  - And shorter with walltimes declared ahead.
    - So far only BNL and Nikhef with short queues
    - Cream will be a problem for this though
- It is possible to run multicore though sites should start to play with it.
  - Some recipes already circulating