



Reducing power consumption on demand

Greg Corbett, Alastair Dewhurst

{greg.corbett, alastair.dewhurst}@stfc.ac.uk

Motivation

- The UK's national electricity provider has a scheme called Demand Side Balancing Reserve (DSBR)[1,2].
 - Aim is to reduce peak time demand in the event there is a lack of generation available.
 - Between 4pm – 8pm during winter months.
 - Will pay for a 1MW drop with 15 minutes notice.
 - Different price bands depending on severity.
- WLCG VOs are starting to make significant use of opportunistic resources.
- Being energy efficient is a good thing!



Reducing Power on Demand

- Power monitoring and benchmarking
- Hibernating Idle Machines
- Introducing Preempt-able Jobs
 - Efficient draining of nodes
- Killing Preempt-able Jobs to fast hibernate machines



Power Monitoring and Benchmarking



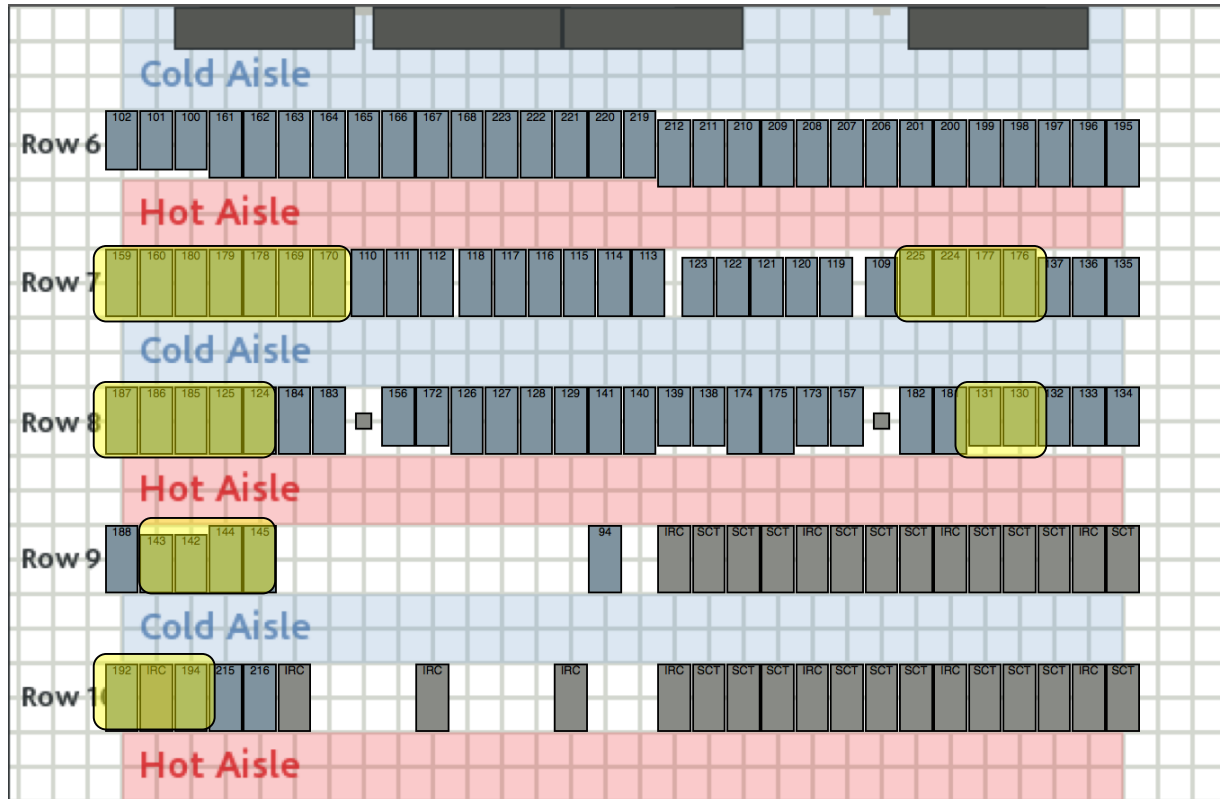
Science & Technology
Facilities Council

Site Setup

- RAL provides large-scale HPC facilities for science
 - Power consumption increased ~25% in last 2 years
 - Currently using ~1.5MW
 - PUE ~1.55
- Batch system at the RAL Tier-1
 - HTCondor scheduler[3]
 - ~600 worker nodes, ~15,500 slots
 - ~200kW power consumption when full



Machine Room

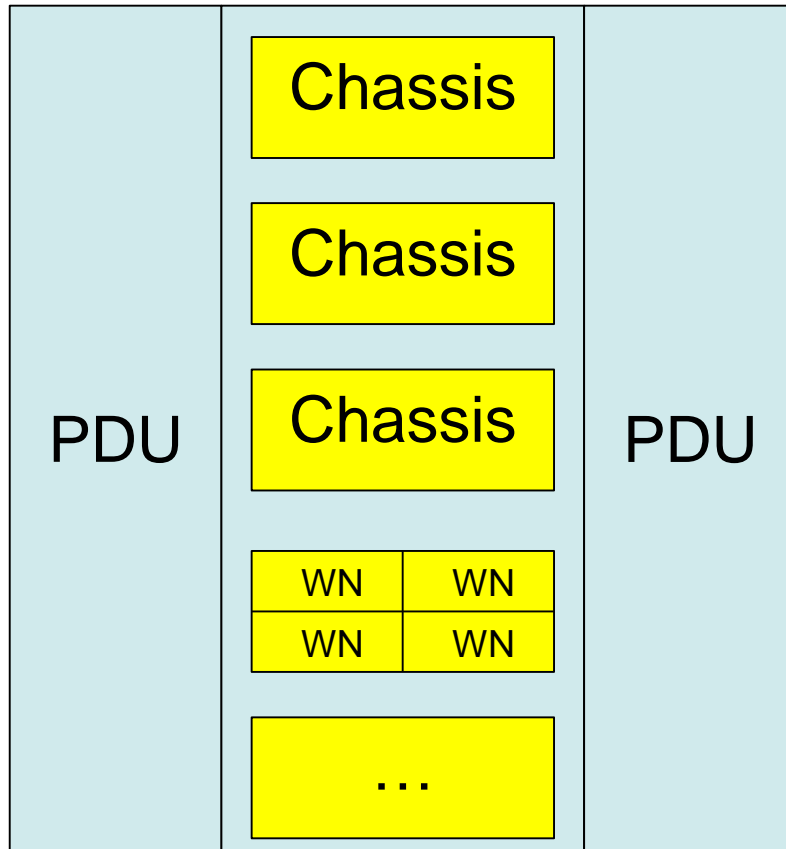


High Power Density Room

Worker Node Racks



Rack Setup

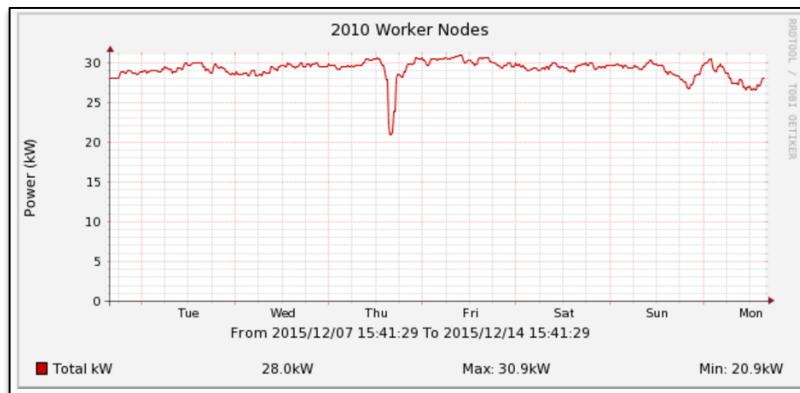


- Each Chassis contains four worker nodes
- Each worker node is fed from both PDUs
- Only meaningful measurement is the total power use of rack.



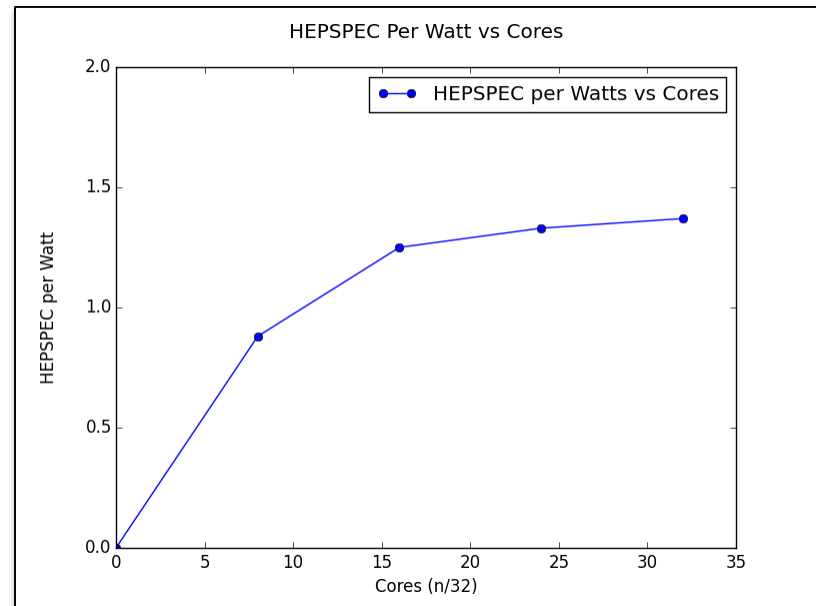
Power Monitoring

- Can aggregate into racks or generations
- Snmp get commands to each PDU
- Cacti automatically collects these responses



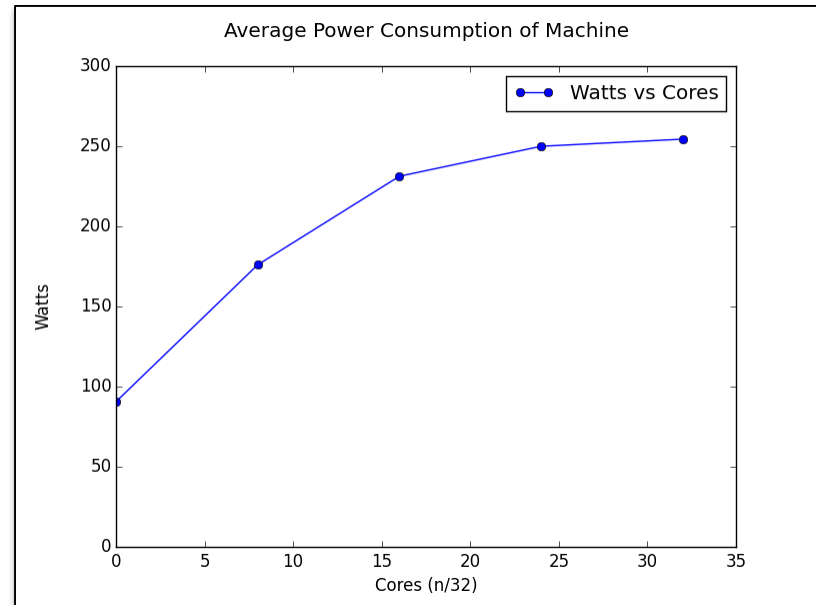
Power Benchmarking

- Measured using a multi-meter connected to chassis
- Used HEPSPec2006 as a measure of useful work
- Most useful work per Watt when machine is using all it's cores



Power Benchmarking

- A machine that is at 50% capacity uses 90% of the power of a fully loaded machine.
- Idle uses 35%
- Fewer, Fuller Worker Nodes is best.



Hibernating Idle Machines



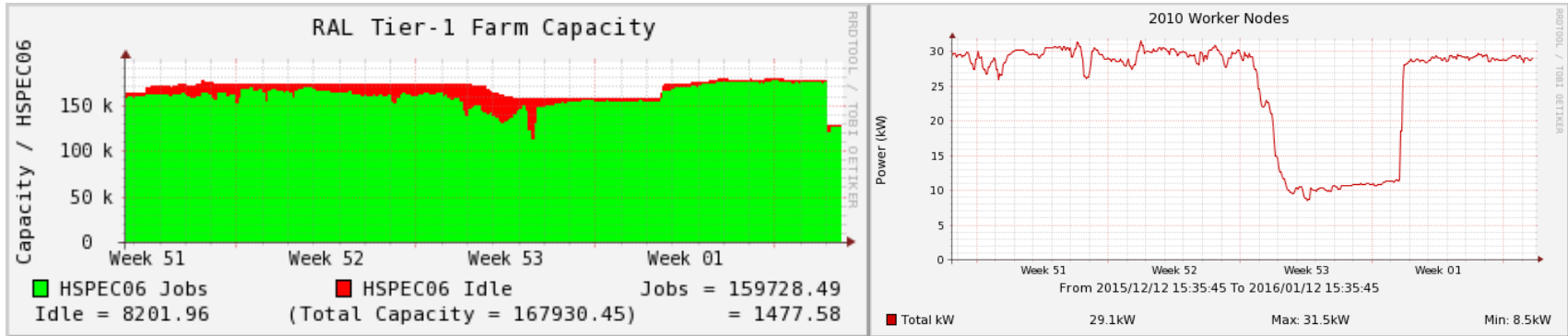
Science & Technology
Facilities Council

Hibernate Idle Machines

	Machine Online	Machine Offline
Jobs starting promptly	Hibernate if: <ul style="list-style-type: none">• Machine has been idle for 5 minutes• At least 30 minutes since last wakeup• Machine COULD start jobs	Do Nothing
Jobs waiting in queue		Wake machine if: <ul style="list-style-type: none">• At least 30 minutes since last hibernation



Hibernation of idle nodes



- Over new year there were insufficient jobs.
- 72 out of 104 WN from 2010 generation were hibernated (1728 job slots).
- Expect energy saving of ~£1000 a year.



Introducing Preempt-able Jobs



Science & Technology
Facilities Council

Preempt-able?

- Amazon spot pricing will kill jobs when price rises:
 - Only charged for whole hours used.
 - One of the reasons behind development of “ATLAS Event Service[4]”
- For WLCG VOs running on the Grid it is difficult to pause and restart jobs easily.
 - Resubmission of failed jobs is easy.
 - Prefer batch farms not to kill jobs as they lose useful debugging information.



Use cases?

- Nodes need to be drained:
 - Partially to create headroom for multicore jobs
 - Fully to reset the node
- Not possible to back fill slots with normal jobs
- Cloud resources:
 - Batch farm can expand into unused cloud resources
 - Cloud users need immediate access to resources
- Reduce Power on demand!



Draining

- Machines chosen based on how fast we expect them to drain:

$$\text{Rank} = \frac{[\text{Slots Running Jobs}]}{[\text{Required Slots}]} = \frac{[\text{Total Slots}] - [\text{Free Slots}]}{[\text{CPUs requested}] - [\text{Free Slots}]}$$

- Modified to include preempt-able jobs:

$$\text{Rank} = \frac{[\text{Total Slots}] - [\text{Preemptable Jobs}] - [\text{Free Slots}]}{[\text{CPUs requested}] - [\text{Preemptable Jobs}] - [\text{Free Slots}]}$$

- Draining machines only allowed to start preempt-able jobs.
- Exact number to be drained changed dynamically depending on demand.

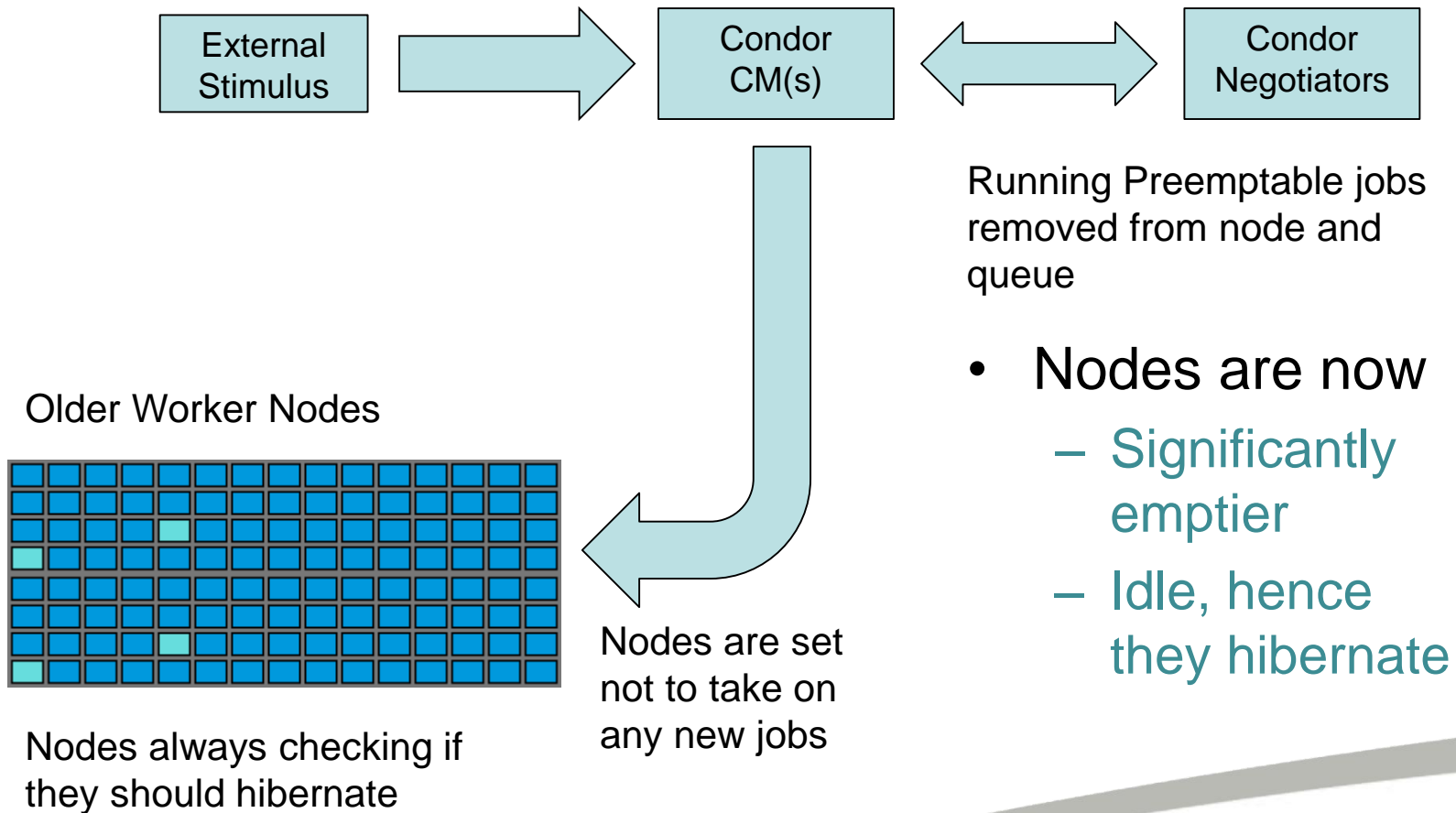


Killing Preempt-able Jobs to Fast Hibernate Machines



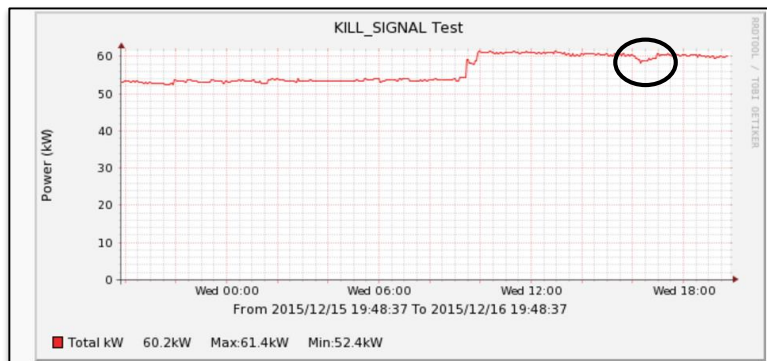
Science & Technology
Facilities Council

Reducing Power on Demand



Testing

- Tests of reducing power on demand performed by running ATLAS Hammer Cloud jobs[5]:
 - Small scale so far as farm is busy with 2015 LHC data!
- Demonstrated functionality
 - Need to increase fraction of preempt-able jobs on WN.
 - Payment of £0.5 /kWh saved would have net benefit for RAL. 2nd lowest price bands on offer.



Summary and future plans

- Hibernating Idle machines should save ~£1000 / year
 - Will be looking at impact on hardware.
- Preempt-able jobs will allow 3% of the farm previously being wasted to be used.
 - Happy to share with other HTCondor users
- Significant increase in preempt-able slots from cloud resources will become available soon.



Thanks for Listening



Science & Technology
Facilities Council

References

- [1] <http://www2.nationalgrid.com/UK/Services/Balancing-services/System-security/Contingency-balancing-reserve/>
- [2] <http://www.theguardian.com/business/2015/nov/04/national-grid-issues-urgent-call-for-extra-power>
- [3] <https://research.cs.wisc.edu/htcondor/HTCondorWeek2014/presentations/LahiffA-RalTier1.pptx>
- [4] <http://iopscience.iop.org/article/10.1088/1742-6596/664/6/062065/pdf>
- [5] <http://hammercloud.cern.ch/hc/app/atlas/test/20072375/>



Hibernate Idle Machines

- HTCondor configuration file that defines ShouldHibernate
 - Machine has been idle for 5 minutes
 - At least 30 minutes since last hibernation
 - Machine COULD start jobs
- Also alters the nodes start expression
 - `START = $(START) && ifThenElse(Offline =?= undefined, true, ((CurrentTime - QDate) >= 900))`



Unhibernate Offline Machines

- Condor Rooster used
 - Wakes up 4 machines at a time
- Machine must have been:
 - Down 30 minutes
 - Matched in the last 5 minutes
- Python script converts machine class ad into `condor_power` command



PREEMPTABLE_ONLY

- New draining “state”
- If true, node will only take preemptable jobs
 - `$(START) && ifThenElse($(PREEMPTABLE_ONLY), isPreemptable =?= True, True)`
- All nodes draining identified by
`condor_status -constraint PREEMPTABLE_ONLY==True`



Killing Jobs to Reduce Power

- Python script and Node HTCondor configuration file
- Configuration file defines:
 - `KILL_SIGNAL = False`
 - `START = $(START) && ($(KILL_SIGNAL) == False)`
- Configuration file also allows Central Managers to change value of `KILL_SIGNAL`



£1000 / Year ?

- Assumes 3% of farm in hibernation
 - 18 machines
- Hibernation saves 95W per machine
- We pay £0.07 / kWh
- $18 * 0.095 * 356 * 24 * 0.07 = \sim£1022$



£0.50 / kWh Payment

- Batch farm hardware costs £2 million.
 - Spread over 5 years
 - Farm per hour usage costs £46.
- Hardware cost per kWh of energy is ~£0.30.
 - The overall power usage of the farm is ~200kW
- Price bands in steps of 25p.
- We run services very cheaply!

