Science and Technology Facilities Council

Running multiple experiment workflows on heterogeneous resources

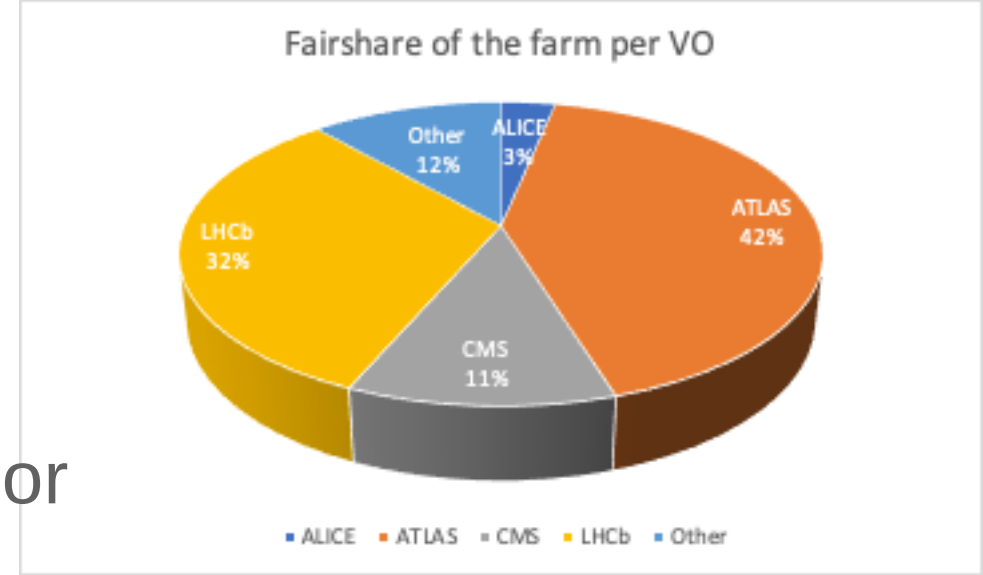
Alastair Dewhurst, on behalf of RAL Tier-1

About RAL

Harwell Science and Innovation Campus is 270 Hectares in size, located 30km south of Oxford.

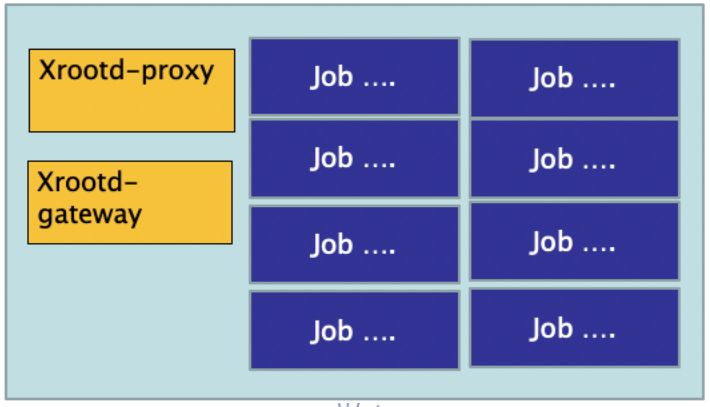
RAL Tier-1

- The RAL Tier-1 primarily supports the 4 LHC VOs.
  - Increasing number of other VOs with significant requirements.
- We provide:
  - 50 000 Job slot HTCondor farm.
  - 40PB data written to disk.
  - 70PB data written to tape.
- The Tier-1 is run with 15.5FTE total.
- 3 members of staff work with HTCondor
  - Thomas Birkett
  - Jose Caballero Bejar
  - James Adams



Batch Farm setup

- RAL batch farm consists of:
  - 5 Schedd (8.8.12) running ARC CEs (6.12.0)
  - HA HTCondor managers
  - ~700 Worker Nodes all running SL7
- To run a job:
  - JOB_TRANSFORM to convert (almost) all jobs to docker jobs.
  - Wrapper script on WNs to start start the Docker container with custom settings.
  - Jobs can have SL6 or SL7 containers.
  - Docker container to provide access to Echo/Ceph.



Hardware

- New generations of hardware have SSD storage and far more cores per server.

	# Servers	HS06 per Server	CPU (Dual)	Cores per Server (Logical)	Memory per Server (GB)	Disk per Server (GB)	Total HS06	Total Cores (Logical)
XMA 15	124	372.01	Xeon E5-2640 V3 @2.60GHz	32	128	1800	46129	3968
HPE 15	164	339.19	Xeon E5-2630 V3 @2.40GHz	32	128	1800	55629	5248
Dell 16	48	420.3	Xeon E5-2630 V4 @2.30GHz	40	160	2400	19188	1920
Dell 17	72	772.28	Xeon Gold 6130 CPU @ 2.10GHz	64	192	1863	53287	4608
XMA 17	56	665.64	Xeon Gold 5120 CPU @ 2.20GHz	56	192	1863	37276	3136
XMA18	44	785.13	Xeon Gold 6130 CPU @ 2.10GHz	64	192	1863	33082	2816
Dell 19	96	1768.91	AMD EPYC 7452 @ 2.35GHz	128	512	3576	169815	12288
XMA 20	112	1765.29	AMD EPYC 7452 @ 2.35GHz	128	512	3576	197712	14336
XMA21	48	2938	AMD EPYC 7763 @ 2.45Ghz	256	1024	7680	141024	12288

STFC Cloud

- STFC provides an OpenStack Cloud for local users.
  - Rapidly growing, now exceeds number of CPUs compared to Tier-1
  - Also provides GPUs, FPGAs.
- STFC Cloud have developed a plugin called Coyote to allow Tier-1 batch to make use of spare resources.
  - Cloud usage is bursty, we have a constant supply of jobs.
- A python script which interacts with the OpenStack APIs and creates worker nodes and runs on a defined interval
  1. Delete virtual worker nodes which are either shutoff or in error state.
  2. Query OpenStack for amount of available vCPUs
  3. If available is greater than a buffer then create a worker node
  4. Worker Node boots and connects to HTCondor to start running jobs.
  5. After 1 week or no jobs HTCondor stops and the worker shuts down.

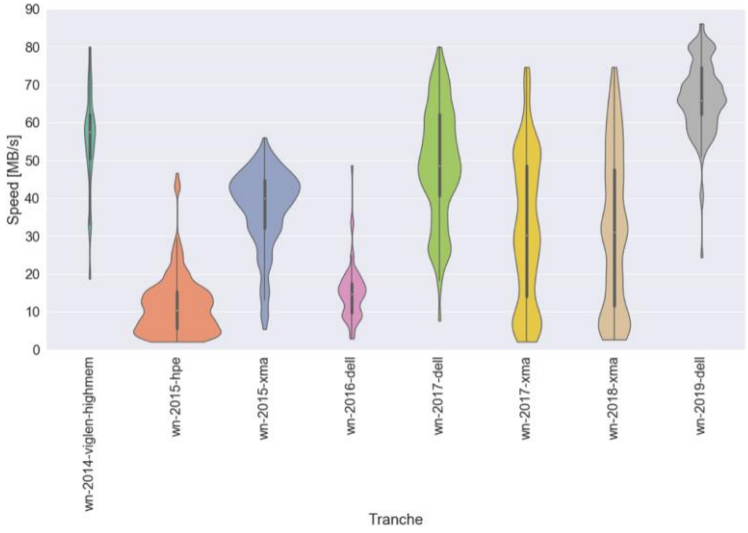
Disk I/O

- Since the introduction of WN with SSD, we have observed a significant improvement in efficiency of some jobs.
  - We need to include IOPs in scheduling decisions.
- Non-SSD nodes still make up a significant fraction of the farm for a few years to come.
  - Still perfectly good at running CPU intensive jobs.

Break down of ATLAS jobs success rate on the farm

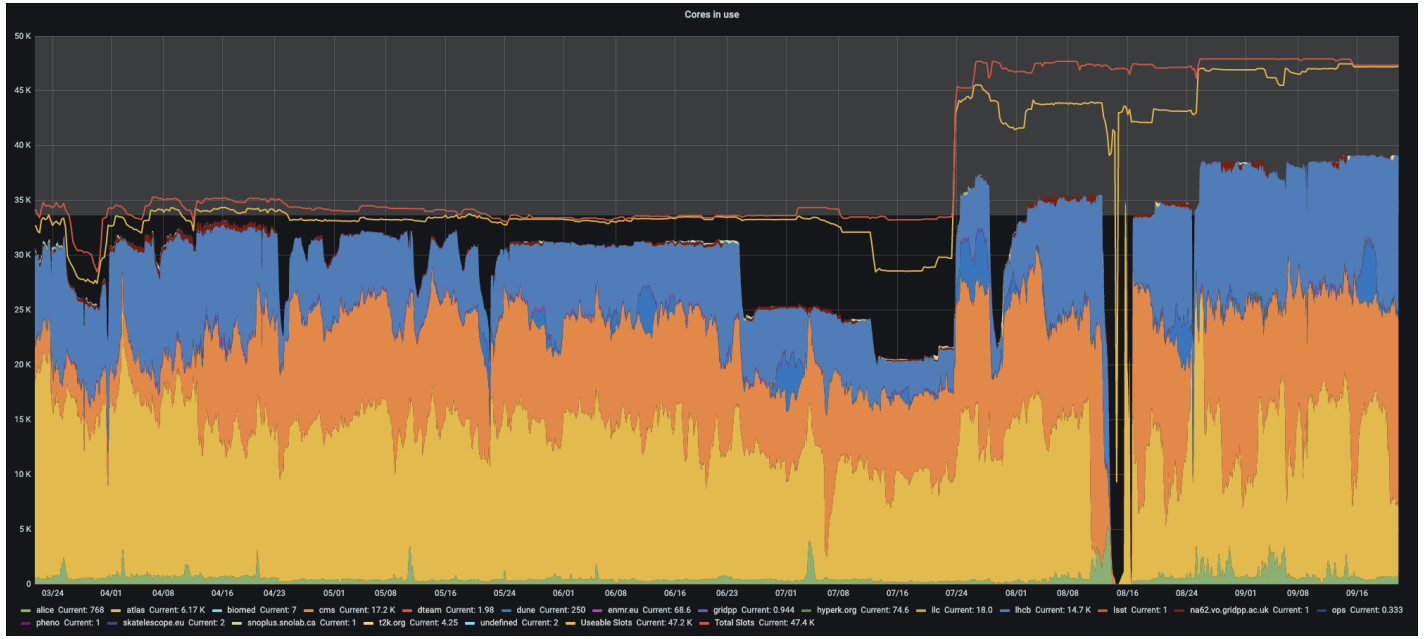
gshare	(Count, noSSD)	(Count, SSD)	(Success Fraction, noSSD)	(Success Fraction, SSD)
COVID	51	23	0.96	0.96
Data Derivations	2428	2442	0.90	0.98
Express	1905	856	0.76	0.71
Group Analysis	1882	1119	0.89	0.97
Group Exotics	2633	1054	0.91	0.96
Group Higgs	1258	456	0.83	0.90
Group SM	61	29	0.89	1.00
Group Susy	1643	805	0.92	0.93
MC 16	13523	14517	0.72	0.77
MC 16 evgen	49	3	0.39	0.67
MC 16 simul	629	765	0.98	0.97
MC Derivations	5730	5098	0.94	0.98
MC Other evgen	46	2	0.20	1.00
MC merge	4284	1883	0.89	0.99
Reprocessing default	240	34	0.83	0.97
Test	2016	215	0.99	1.00
Upgrade	17	8	0.94	1.00
User Analysis	71562	31154	0.86	0.93
Validation	13	15	0.92	1.00

Transfers speeds when access data



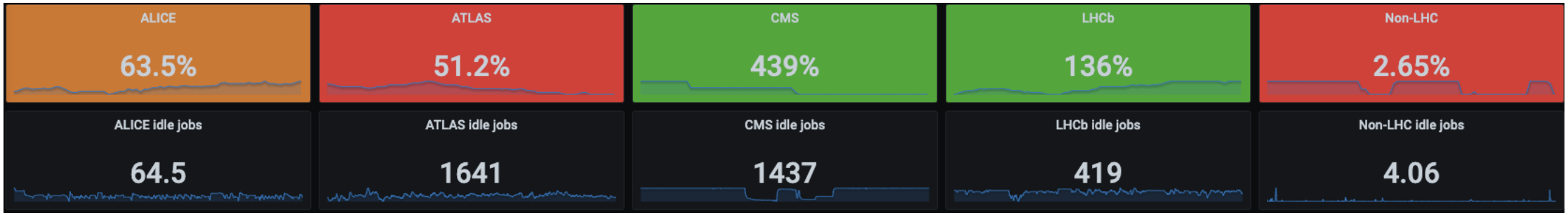
Updates in the last year

- Docker uplift from 18.04 to Docker 20.10.7
- HTCondor upgrade from 8.6.13 to 8.8.12 on Masters and Workers
- ARC-CE upgrade involving major refactor of config and custom scripts. Moving from ARC 6.6.0 to 6.12.0
- Introduction of 2020 generation of worker nodes



Job Scheduling

- WN are split into two types:
  - Those that can run any job
  - Those that can run Multi-Core only
- We switched off our efficient defrag.py because it broke.
  - We allow HTCondor to defrag nodes to schedule multi-core jobs.
- Maintaining the balance of jobs is hard as the job mix fluctuates.



Scheduling issues

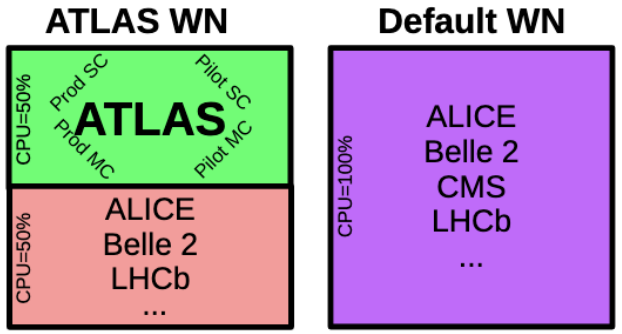
- We can break users into 3 categories:
  - Users that exclusively submit single core jobs (LHCb, ALICE, DUNE)
  - Users that exclusively submit multi-core jobs (CMS)
  - Users that submit a mixture of single and multi-core jobs (ATLAS)
- Main issues:
  - ATLAS often receive less than their fairshare as a result of changing their job mix.
  - CMS jobs which are the most I/O intensive are often concentrated on a small number of nodes.
- Now we have upgraded we have significantly more options to deal with these problems.

Possible solution

- KIT have similar job mix as RAL.
  - They have created a dedicated ATLAS partition and it works well.
- Doesn't feel very adaptable, what if more VOs run a mix of jobs?
- Doesn't allow us to take into account different hardware types.

ATLAS Unified SC/MC Queue Support

- Modified setup optimizing ATLAS unified SC/MC queue scheduling
  - Several issues caused by frequent MC-SC-MC transitions
  - Solution (since autumn 2020): dedicated ATLAS farm partition, still maintaining job mix by implementing 2 HTCondor partitionable slots on subset of WNs, serving either only ATLAS, or only the other VOs



pre-GDB 2021-07-13 Manfred Alef, Max Fischer: WN Setup and Procurements at GridKa Steinbuch Centre of Computing (SCC)

https://indico.cern.ch/event/876806/contributions/4400261/attachments/2281024/3875653/WN_Setup_and_Procurements_at_GridKa-2021-07-13.pdf

Future Work / Conclusions

- We haven't decided how best to solve our scheduling issues.
  - We want to make changes before the Run 3 starts.
  - We hope we will get many answers / ideas from this workshop.
- Newer Larger Worker nodes have meant that traditional job requirements such as memory are becoming less important while disk I/O and network connectivity are becoming more important.
  - The fact that most jobs come in as pilot means often the best thing we can do is to try and spread jobs from different VOs around as much as possible.
- Mixed workflows from ATLAS are difficult to schedule and maintain their fairshare. A dedicate solution for them will work but is not scaleable.

Science and Technology Facilities Council

Questions?