



SLURM for WLCG in the Nordics

Mattias Wadenstein
Hepix 2012 Fall Meeting
2012-10-18, Beijing

- LFC decommissioned this summer
 - ATLAS migration downtime 2nd - 3rd July
 - Which was right before ICHEP :)
- Accidentally makes data federation work
 - `preferredpattern="srm://srm.ndgf.org|.si$|.se$"`
 - If preferred PFN unavailable, try other PFNs
 - Makes [some] pool downtime invisible to users
- Other data federation work
 - ARC remote cache reads – devel in progress

ATLAS Grid Monitor

2012-10-13 CEST 10:50:31



Processes: ■ Grid ■ Local

| Country | Site | CPU | Load (processes: Grid+local) | Queueing |
|--------------|---|--------------|---|--------------------|
| Denmark | Steno Tier 1 (DCSC/KU) ● | 5168 | ■ 864+2935 ■ | 1296+1 |
| | Tier1 (BCCS/UiB) | 372 | ■ 0+234 | 0+1 |
| Norway | Titan A (UiO/USIT) ● | 9856 | ■ 886+2216 ■ | 282+8 |
| | Titan C (UiO/USIT) ● | 9856 | ■ 0+3103 | 0+8 |
| Slovenia | Arnes ● | 1636 | ■ 1558+0 ■ | 1003+0 |
| | SiNET ● | 2092 | ■ 2027+0 ■ | 684+0 |
| Sweden | Alarik (SweGrid, Luna> ● | 3328 | ■ 114+2315 ■ | 52+0 |
| | Grad (SweGrid, Uppmax) | 512 | ■ 353+0 ■ | 59+2 |
| | Ritsem (SweGrid, HPC2> ● | 544 | ■ 322+0 ■ | 243+0 |
| | Siri (SweGrid, Lunarc) | 512 | ■ 332+138 ■ | 240+50 |
| | Smokerings (NSC) | 520 | ■ 416+80 ■ | 2525+0 |
| | Smokerings TEST (NSC) | 520 | ■ 0+496 (queue inactive) | 0+2519 |
| Switzerland | Bern ATLAS T3 | 532 | ■ 456+0 ■ | 150+0 |
| | Bern UBELIX T3 | 1216 | ■ 448+0 ■ | 396+1 |
| | Geneva ATLAS T3 | 278 | ■ 27+193 ■ | 77+357 |
| | Manno PHOENIX T2 | 2176 | ■ 252+1916 ■ | 53+403 |
| | Manno PHOENIX T2 | 2176 | ■ 245+1921 ■ | 54+402 |
| TOTAL | 17 sites | 41294 | 8300 + 15547 | 7114 + 3752 |

- IJS
- LUNARC
- HPC2N

- Native OS is Gentoo
- ATLAS computing run in SL5 (hopefully soon SL6) chroot
- Slurm setup using schroot to run the job inside chroot when requested in the xrsi
- Need to bump `/proc/sys/kernel/pid_max`

- Nicer than torque
- Information about running jobs lacking
 - ARC integration issue to get mem/cputime etc
 - Running jobs sstat, finished jobs sacct
 - And sometimes info is missing depending on config
- task/affinity cgroup not good for fat nodes with many jobs
- cgroups for memory containment on rmem
 - Works much better than vmem etc
- Defaults are not always good

- slurm and ARC 2.0 on our new WLCG resource Alarik
- Lot of bugs in the submit-SLURM-job.sh
- Better control on running jobs
 - enabling more jobs to run without interference on the same node
- Would be nice if job constraints like "-N 1 -n 16" could be expressed using XRSL or similar job description languages
- Modern, resource efficient, job isolation

- Big HPC system first out
 - cgroups essential
 - Very scalable and efficient
 - Fairly stable
 - And “fairly” is way better than torque+maui
- Then we reinstalled our grid resource
 - Because #jobs was larger → less stable maui
 - ARC 1.1 interface works, but with quirks
 - Infosys publishing could be much better
 - fairshare=parent for pool accounts etc
- Strong sanity checks of nodes



Questions?