Recommendations for sites

Alessandra Forti

ATLAS Jamboree

18 January 2017

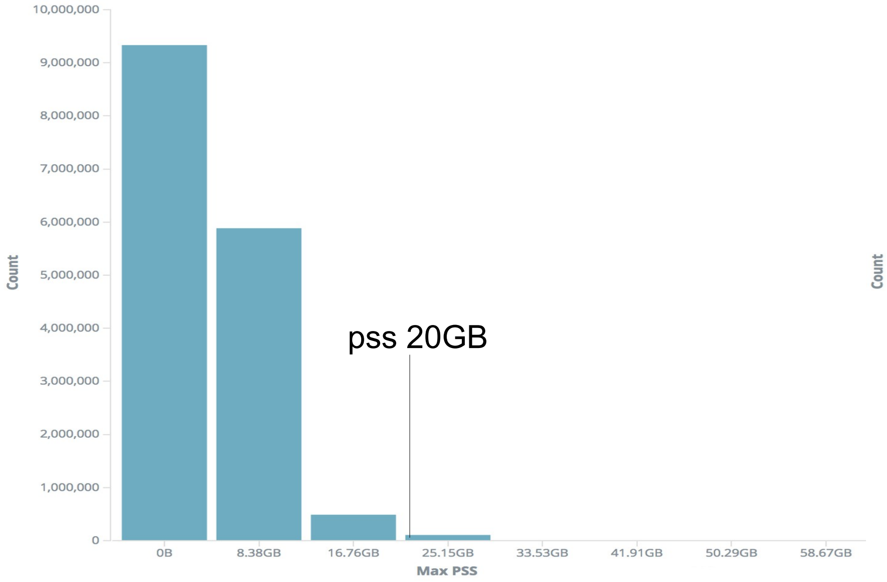
Introduction

- Memory
- Batch systems
- Shares
- CEs
- WNs hardware
- Storage
- Agis
- Squid
- Traceability
- Harvester
- Hammer Cloud

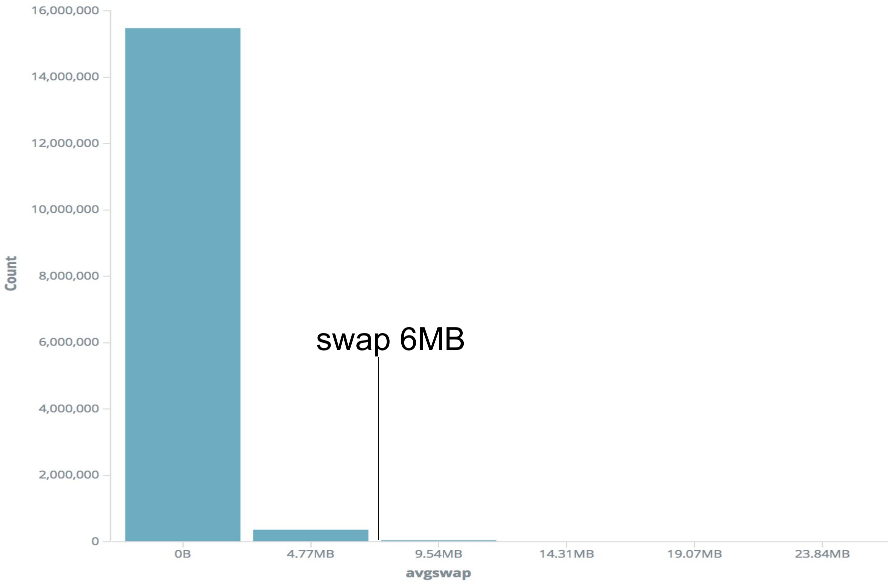
Memory

- Vmem: memory mapping in 64bit can be several times the actual memory used it doesn't mean it gets used. ✘
- Smaps RSS: physical memory used by a job double counting the memory shared with other jobs ✘
  - ≠ from cgroups RSS
- **Smaps PSS:** physical memory used by a job without double counting ✓
- **cgroups RSS:** physical memory used by the jobs without double counting ✓
  - Quantitatively similar smaps PSS

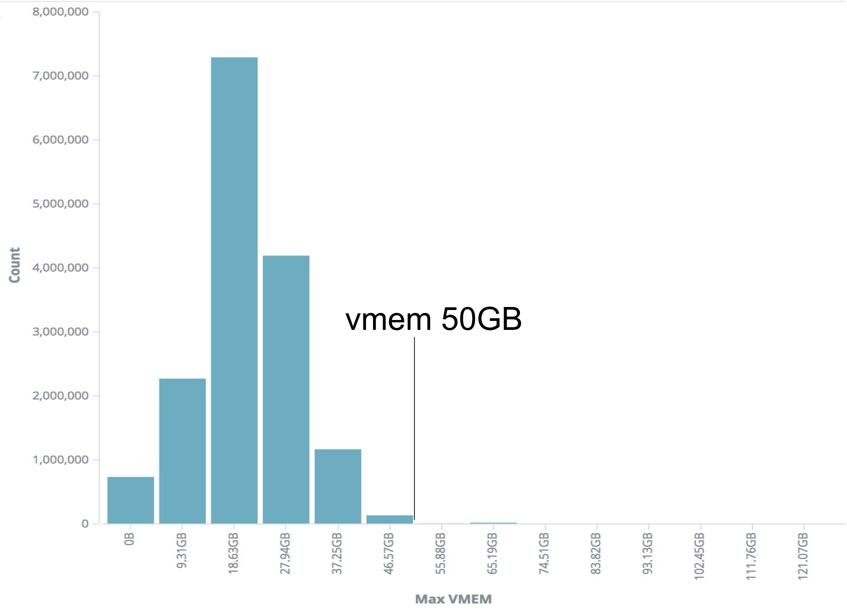
Memory plots



pss 20GB



swap 6MB



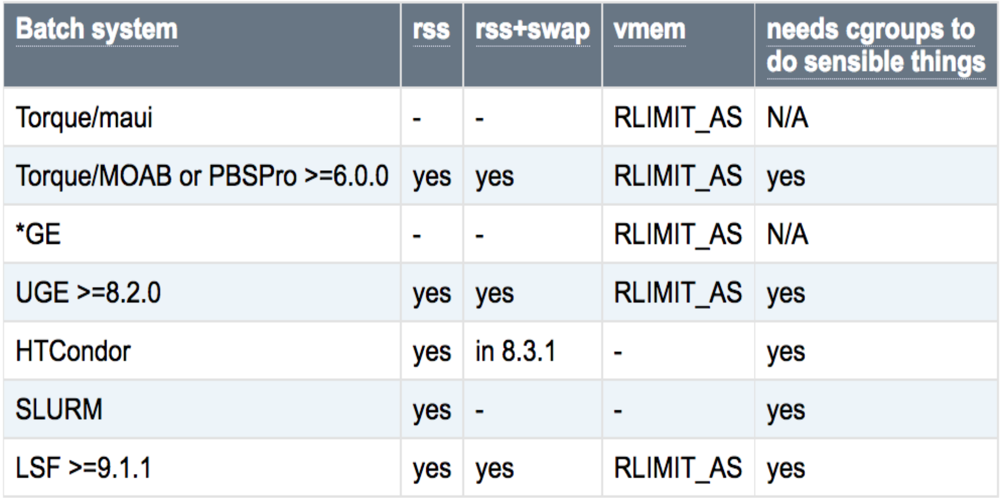
vmem 50GB

What batch systems do?

- Batch systems **without** cgroups
  - See the same RSS as reported in smaps
  - Kill on vmem which is **NOT** a physical memory measure
    - If you insist on this you need to set it **at least 3 times the RAM** requested by the job
  - If you kill with the scheduler it is likely to the same problem
- Sites with cgroups
  - Can setup soft and hard limits on the values the job reports
  - Soft limit allows the kernel to decide if the job can keep on using the extra RAM or has to swap
  - Hard limit will kill the job based on RAM
    - Often set to 2 or 3 times the RAM requested by the job

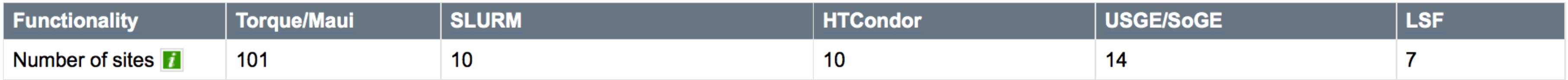
Which batch system

- ATLAS recommends sites move to a BS supporting cgroups and other more modern features.
- Community expertise for
  - **HTCondor**
  - **SLURM**



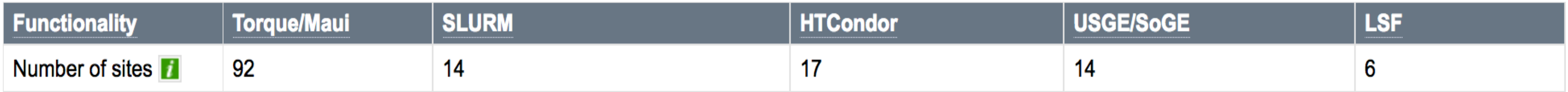
Batch system	rss	rss+swap	vmem	needs cgroups to do sensible things
Torque/maui	-	-	RLIMIT_AS	N/A
Torque/MOAB or PBSPro >=6.0.0	yes	yes	RLIMIT_AS	yes
*GE	-	-	RLIMIT_AS	N/A
UGE >=8.2.0	yes	yes	RLIMIT_AS	yes
HTCondor	yes	in 8.3.1	-	yes
SLURM	yes	-	-	yes
LSF >=9.1.1	yes	yes	RLIMIT_AS	yes

- 13/04/2014



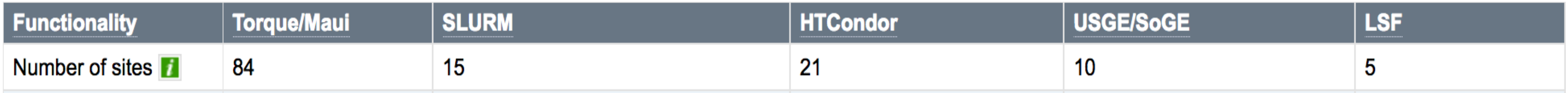
Functionality	Torque/Maui	SLURM	HTCondor	USGE/SoGE	LSF
Number of sites ℹ	101	10	10	14	7

- 26/01/2016



Functionality	Torque/Maui	SLURM	HTCondor	USGE/SoGE	LSF
Number of sites ℹ	92	14	17	14	6

- 17/01/2017



Functionality	Torque/Maui	SLURM	HTCondor	USGE/SoGE	LSF
Number of sites ℹ	84	15	21	10	5

CEs

- 3 CEs
  - ARC-CE
    - Most used at new sites and sites moving to HTCondor at EGI sites
    - Well integrated with SLURM
  - HTCondor CE
    - Most used in the US
    - If you have HTCondor batch system is just an additional layer of configuration
  - CREAM-CE
    - Most used in EGI for legacy reasons but ATM is best integrated with older batch systems like torque/maui and SGE
- If you change batch system you may want to consider reviewing also your CE

Shares at sites

- Requested fair shares
  - **Analysis: 25%**
  - **Production: 75%**
    - SCORE: 20%
    - MCORE: 80%
- Overall: 40% SCORE and 60% MCORE
  - 50-50 good enough
- However this share is **not constant** in time
  - Sometimes more MCORE
    - MCORE is still struggling to run at few sites
    - Static partition setup is NOT recommended
    - Reminder that recipes for more dynamic approaches for 3 batch systems can be found
      - WLCG Multicore TF pages: Torque, HTCondor, SGE

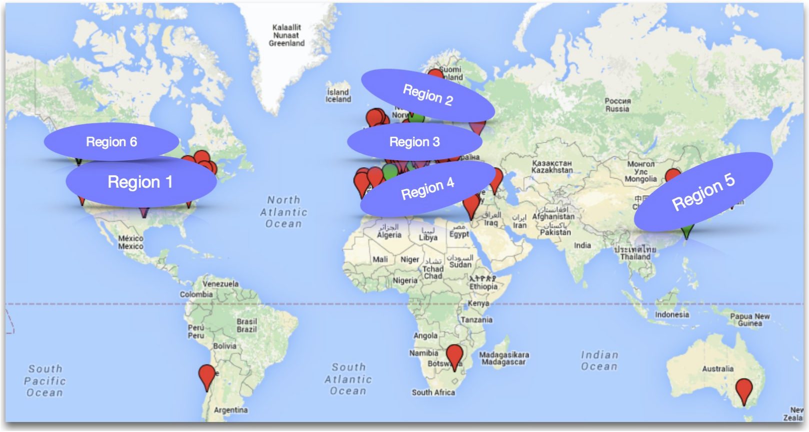
WN Hardware

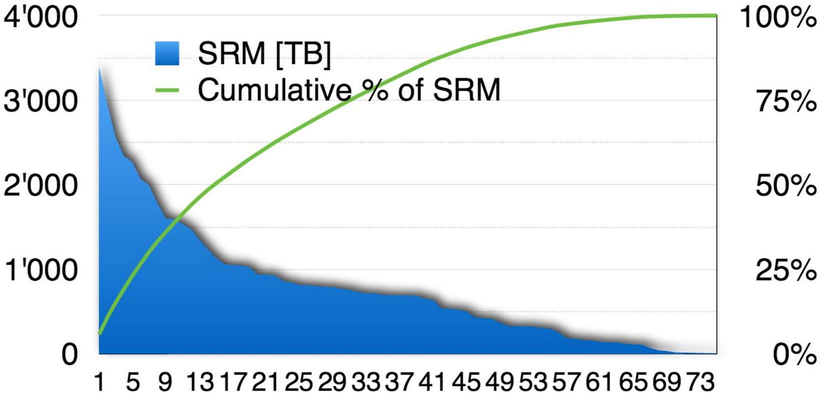
- A node should typically provide the following amount of hardware resources per single-core job slot
  - About 20 GB of disk scratch space
    - For an 8 core MCORE slot, ~80-100GB is sufficient
  - At least 2 GB of (physical) RAM
    - Having 3-4 GB would be beneficial
  - Enough swap space such that RAM + swap >= 4 GB
  - As a rule of thumb, about 0.25 Gbit/s of network bandwidth
    - Might want higher for more powerful CPUs.

Storage

- Change of storage topology
  - Bigger sites (T1 and T2) with satellites indepently from location
    - Evolution of sites towards caches or federations
- Consolidate storage
  - 75% of storage at ~30 sites
  - Small sites <400TB discouraged from buying storage unless they can go above or aggregate with other sites

Possible evolutions of computing model



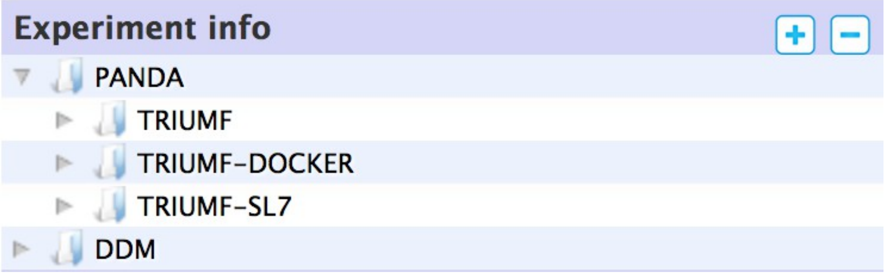


E. Lancon presentation

Master/slave queues

- Master queues
  - is_default = True
    - SW releases are fully validated
    - HC tests them frequently
    - Determine blacklisting status of all the dependent queues.
    - There has to be **one for production and one for analysis** in the main PandaSite

PandaSite



Experiment info

- PANDA
  - TRIUMF
  - TRIUMF-DOCKER
  - TRIUMF-SL7
- DDM

- Slave queues
  - All the other queues usually requesting more resources: HIMEM, MCORE

PQs obsolete parameters

- Plan is to remove them
- In the meantime here is a list that don't need to be set anymore some are already marked as deprecated on the PQs pages
  - gatekeeper, lfchost, lfcpath, lfcprodpath, lfcregister, minmemory, maxmemory, allowdirectaccess, cmd, cmtconfig, jdl, jdladd, name2, proxy, queue, datadir
  - setokens, seprodpath, sepath, seopt, se, seinopt, sein, copytool, copytoolin, copysetup, copysetupin, copyprefix, copyprefixin,
  - copyprefixin_fax_direct, copyprefixin_fax_xrdcp, copysetup_fax_direct, copysetup_fax_xrdcp, copysetupin_fax_direct, copysetupin_fax_xrdcp

Squid

- Condition data and software are accessed through squid
  - Frontier & CVMFS
- Sites are requested to install at least one squid
  - Two for redundancy and load balancing
- Frontier squid or OS squid?
  - Frontier squid has some patches to boost performance. It is also a higher version with bug fixes.
  - OS squid is easier to maintain because is there by default.
    - T2s can get away with the OS squid but ATLAS recommends to use the Frontier version
- Monitoring:
  - http://wlcg-squid-monitor.cern.ch/snmpstats/indexatlas.html
  - SSB squid
  - If you are using the old ATLAS mrtg please change it

Traceability

- Glexec has been dropped
  - WLCG Traceability TF working on other tools and models
- Singularity container solution being tested at CERN and in OSG
  - 1 single executable doesn't need a daemon
  - Can isolate payload from pilot environment
  - Cannot do traceability that will have to be done at VO level
    - ATLAS already does this
      - Site: time/(IP|host) -> VO
      - VO: time/host -> user+payload
  - Being integrated in slurm and Htcondor

Harvester requests for sites

(Alessandro De Salvo)

Number of local queues

- Sites with batch systems and CEs supporting dynamic description of the jobs should limit where possible the number of static local queues
  - The batch system should be able to handle at least the requests about the number of desired cores and memory
  - Most batch systems can handle this, although the slot fragmentation could be an issue
  - The enforcement of the limits can be done via cgroups (preferred) or just via the pilot
  - The site CE flavour should be able to pass the parameters correctly to the batch system, in case of Grid sites
  - Ideally, we could eventually have just one local queue for each site

Monitoring

- Site monitoring
  - Detailed monitoring of the local activities can still be performed, but we have to use centralized tools to have a full view
    - The Analitics Platform offers all the needed tools to aggregate activities and fully monitor our sites
- Redefining the concept of activity monitoring
  - In general we need to redefine/update our concept of activities and aggregate sub-activities via the available tools to have the real view
  - For example you can monitor all the Analysis activities via the Analytics Platform
    - Sites may prefer to see the analysis jobs components (user+group)
    - Funding agencies may check the Analysis overall workload by adding Derivations too (Derivations+User+Group)
  - More info in the following talks

Harvesting site description

- Harvester will need to have a deep knowledge of the resources which are part of the clusters in grid sites
  - It is desirable to have read access to inner batch system information in order to create a node map in Harvester with informations that are normally available to batch systems for scheduling purposes
  - Examples are the node type, the number of cores/slots, the available memory, the load, the batch queue configuration, etc
    - Those information should be available at least to the pilot, so with a unprivileged access, or exposed via the CEs
    - Most of those parameters are available to the pilot already, but some are more tricky, like the number of slots and the batch queue configuration
  - Not all the informations should be frequently refreshed
    - For example the node description could be refreshed slowly or on-demand (semi-static), but e.g. the load should be frequently updated (dynamic)
  - Sites may start exposing the semi-static informations and, later, the dynamic ones
  - Do the sites already have the possibility to expose such informations?
  - The ATLAS central team is available to help creating these kind of “information providers”, working together with the sites

HammerCloud Update

(Jaroslava Schovancova)

Achievements

- Improved internal service monitoring
  - No more silent crashes in the middle of the blacklisting script run!
- Topology: all taken from AGIS
  - No more hardcoding!
- Blacklisting: improved topology handling (master vs. slave)
- HammerCloud extension in production since November 2016
  - Much easier and more flexible template definition, no restriction w.r.t what type of workload, what input parameters/how many different input datasets
  - Decommissioned DQ2 in HC (client/API end of life by end of 2016)
  - Decommissioned Ganga in ATLAS HC (outdated application, no dev support)
- Commissioning
  - Components or sites: New pilot mover, srm-less storage, batch ramp-up speed, …

HC configuration in AGIS

- PanDA queue object
  - I.e CERN-PROD-all-prod-CEs AGIS link
- Important fields:
  - **state:** ACTIVE vs. DISABLED
  - **type:** production/analysis/special
  - **is_default:** master (true) vs. slave (false)
  - **capability:** score vs. mcore vs. himem/mcore_himem/lowmem
  - **hc_param:** AutoExclusion, OnlyTest, OnlyExclusion, False
  - **hc_suite:** PFT/AFT/PFT_MCORE
- HC Tutorial for site admins: Integration with AGIS

Configure Queues for HC Testing

- **state:** ACTIVE
- **hc_param:** AutoExclusion or OnlyTest
- **hc_suite:** PFT/AFT/PFT_MCORE
- **type:** production/analysis/special
- **is_default:** master (true) vs. slave (false)
- **capability:** score or *mcore*

Configure Queues for blacklisting

- Define 1 master queue per PanDA site per activity
  - i.e. max 1 master per production per PanDA site, max 1 master per analysis per PanDA site
- Configure the master queue for HC tests
  - hc_param in AutoExclusion or OnlyExclusion
- Configure slaves for HC tests (hc_param dtto)
- Blacklisting: based on test results of master
  - Slaves are excluded only if master fails, whitelisted when master succeeds. This will evolve in 2017.
- HC Tutorial for sys admins: blacklisting rules

Hot topics 2017

- JEDI integration
  - e.g. Event Service based tests
- Blacklisting (explore possibility to use analytics to blacklist)
- Cloud computing resources pre-commissioning
- Re-introduce Nightlies tests
- RTT (ATLAS RunTimeTests)
- HC infrastructure: CC7, 3rd party SW update, CERN SSO for login

Get in touch

- Links:
  - HammerCloud Tutorial for ATLAS site admins
  - HammerCloud portal

- Do you have a question, feedback, idea, testing need?
- In need of a stress test?
- Would you like to contribute?

- Get in touch with ATLAS HammerCloud team:
  - <**atlas-adc-hammercloud-support@cern.ch**>

Backup slides

Virtual Memory

- Many sites limit vmem because they want to limit RSS+swap
  - Kernels have changed years ago and vmem doesn't mean RSS+swap anymore it's the size of the address space
    - SCORE 32bit vmem=RSS+swap was still negligible in first approximation
    - 64bit address space much larger difference will increase
- Standard tools do not report the memory correctly anymore nor are able to limit RSS+swap
  - Processes may look like they are using 40GB of vmem but if one looks at RSS+swap with other tools the same processes don't go above 20GB (see plots next slide)
    - Swap for multicore jobs is negligible
  - ulimit used to be able to distinguish for example it could limit RLIMIT_RSS now it limits only RLIMIT_AS which affects all memory allocation and mapping functions

Memory multicore case

- To the previous slide we need to add that multicore RSS is wrong by default because the shared memory is accounted multiple times.
  - Even without counting the experiments asking for more to cover the 5 minutes peaks
- Some sites limiting the (v)memory had to increase the limit
  - Problem when limit = allocation of resources
- Some sites are oversubscribing the memory by a factor
  - Useful particularly for multicore when most of the time the memory is not used.
  - Recipes for maui and HTcondor exist

Computing consolidation

- Reduce the variety of CE/BS combinations
  - OSG consolidating on HTCondor-CE/HTcondor
    - HTCondor-CE is a configuration of HTCondor
  - ARC-CE/HTCondor deployment is increasing
    - ARC-CE in general has some advantages for ATLAS
      - ARC-CE cache mechanism for sites that don't want a full blown storage
      - aCT solves the “one size fits all problem
        - works only with ARC-CE
    - HTCondor advantages
      - Use opportunistic resources when they become available
      - Has better support for virtual WNs
      - Better integrated with Linux resource management (cgroups, docker...)

Computing consolidation (2)

- Cont....
  - ARC-CE/SLURM also well supported in ATLAS
    - SLURM advantages
      - node health checks disables bad nodes and reenables them if they are sane again.
      - using chroot, containers is relatively simple and the OS can be dynamically chosen by the jobs.
      - very efficient backfilling mechanism maximizes the cluster usage
      - support for massively parallel jobs (designed for HPCs) and property management (eg additional resources such as GPUs)
  - ARC-CE/Other BS
    - Other batch system supported but not as well integrated as these two.
- These are recommendations, not requirements, for sites that want to move away from their current setup