Multicore

Passing parameters to BS and memory handling

Alessandra Forti

Antonio Perez-Calero Yzquierdo

On behalf of the multicore TF

GDB

12 November 2014

Passing parameters to BS

- Would simplify batch system job in allocating resources
  - Instead of relying on queue parameters which are usually really large
  - Would enable backfilling
  - Would enable jobs to request the memory the need making limits less important
- Works at some sites but not all

CEs&BS

- 3 type of CEs
  - ARC-CE
  - **CREAM-CE**
  - HTCondor CE (US)
- 5 (main) batch systems
  - Torque/Maui
  - HTCondor
  - SGE
  - SLURM
  - LSF
- Several possible parameters
  - Not possible nor necessary to use them all

CREAM-CE

- Current setup is EGEE legacy
  - Framework for written with BDII in mind
    - ForwardOfRequirementsToTheBatchSystem
    - Too flexible for the good of anyone
    - Introduces a concept of minimum resource that the batch system can't handle and needs to be converted to a Max.
- *_local_submit.sh scripts to be written by sites admins
  - There are ~3 scripts around
    - Nikhef: Torque
    - EGI rpm: SGE another SGE in use at FZK
    - CERN: LSF
  - Never really agreed on a common format although some commonalities between 2 main scripts circulating for Torque and SGE the one written by CERN for LSF is completely different

Glue1 or Glue2

- Another dimension of the problem is what to pass to the CEs.
  - Need to match what is in the BDII?
    - BDII is going away for LHC still need to think to smaller Vos.
    - ARC-CE and HTCondor CE don't use Glue to pass parameters
  - US sites still use Glue1 in their IS
    - Different system different CEs not clear they'll be affected if experiments pass whatever parameter to CREAM-CE
    - OSG Ops now involved in the TF
  - CREAM-CE currently uses Glue1
    - It add a suffix to a _Min or _Max depending on the operator used
    - Should work with any string but haven't tried yet

Starting from the BS

- Reduced the number of params to 5
  - Check which parameters correspond to each batch system
  - Check what they do (do they behave in the same way)
  - Match them to whatever string the CE requires from the user after agreeing on a uniform meaning understood by sys admins and users

Batch Sys	corecount	Memory (RSS)	Vmem	CPU time	Wall time
Torque/maui	ppn	mem	vmem	cput	walltime
*GE	-pe	s_rss	s_vmem	s_cpu	s_rt
HTCondor (*)	RequestCpus	RequestMemory	Recipe	Recipe	Recipe
SLURM	?	?	?	?	?
LSF	?	?	?	?	?

(*) ARC-CE has a HTCondor backend with *Limit parameters which make it simpler

Virtual Memory

- Many sites limit vmem because they want to limit RSS+swap
  - Kernels are changing and vmem doesn't mean RSS+swap anymore
- Standard tools do not report the memory correctly anymore
  - 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
- Nor are able to limit RSS+swap
  - 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

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
  - 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

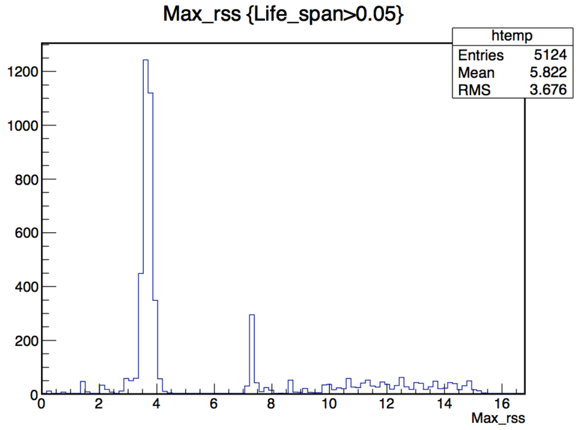
Memory multicore case

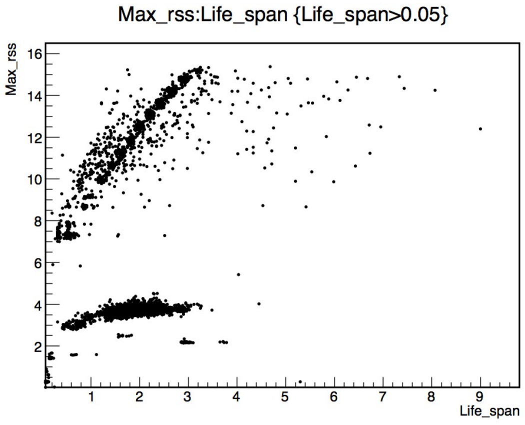
- To the previous slide we need to add that multicore (v)memory is wrong by default because the shared memory is accounted multiple times.
    - Even without counting the experiments asking for more to cover the 5 minutes peaks
- Some sites limiting the (v)memory had to increase the limit
    - Problem 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

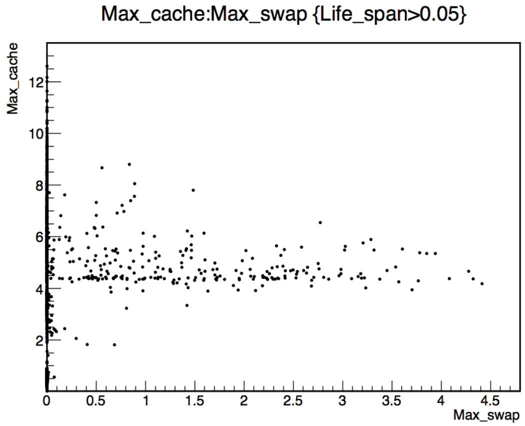
Memory and cgroups

- Some sites are enabling cgroups.
  - Allows more accurate monitoring (see plots next slide)
  - Allows smart soft limit without allocating memory
    - If jobs exceed this the kernel pushes them back to a smaller value
  - Allows hard limit job gets killed

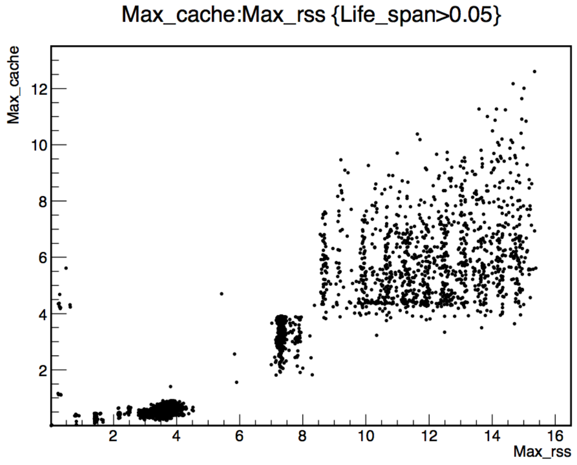
Memory and cgroups (2)


Max_rss {Life_span>0.05}


Max_rss:Life_span {Life_span>0.05}


Max_cache:Max_swap {Life_span>0.05}

- Glasgow last 10 days
  - Stats taken from memory.stat every minute
  - Global values not associated with PID
  - There maybe recipes to collect the memory metrics for each job.
- Results from real jobs confirm those presented by Andrej at ADC weekly


Max_cache:Max_rss {Life_span>0.05}

cgroups and BS

- Can it work everywhere?
  - Really easy to enable in Htcondor
  - Supported in SLURM
  - UGE has been patched
  - SoGE/OGE no support
    - Most GE sites use this I think
  - torque/maui no support
    - At last count still 100 sites
- Sites moving away from torque should look into it though
  - HTCondor recipe really easy
  - SLURM probably easy too

smaps

- Can we use /proc/$PID/smaps instead?
  - smaps reports things correctly but there are no standard toools
- It was suggested to write something for monitoring
  - Can we write something for limiting the memory?
  - Are recipes or scripts circulating?
  - Would it be to in-the-house-solution?
- Further discussion is needed

Summary

- Passing parameters to the batch systems discussion is progressing
  - Needs more discusssion to better define each quantity and to make it uniform across the board
- Vmem discussion is related but right now it seems more difficult to solve as it may require a radical change of the infrastructure
  - Cgroups is the OS solution
    - Currently it is not going to work on most sites due to batch system limitations
  - Need to discuss how to approach this
    - Some other bout of creativity from sys admins to use either cgroups or smaps (?)