New Statistics Using TRES

Bill Brophy
Martin Perry
Thomas Cadeau

26-09-2017

© Atos

Bull
atos technologies

New Statistics Using TRES

- TRES overview
- New TRES for Lustre & Ofed
- Conclusion

Bull
atos technologies

What’s TRES ?

TRES: Trackable Resources

- TRES documentation
  - SchedMD HTML on Trackable RESources (TRES)
  - SchedMD HTML on Resource Limits
  - Slurm User Group Meeting 2015 : Brian Christiansen and Danny Auble
  - Other web pages contain additional documentation

- Concept introduced in Slurm for 4 primary reasons:
  - Supports limiting of resources besides cpu/memory/nodes
  - Provides new factors for computing priority
  - Provides new factors to be used for billing
  - Supports tracking of other resources

Current TRES

- Current TRES Types are:
  - BB (burst buffers)
  - CPU
  - Energy
  - GRES
  - License
  - Mem (Memory)
  - Node
- Default: CPU, Energy, Memory and Node
- Configuration: AccountingStorageTRES
  - Example: AccountingStorageTRES=gres/craynetwork,license/iop1,bb/cray
- sacctmgr commands are used to establish TRES limits

Bull
atos technologies

Example of Current TRES Display

- To display the allocated resources for a job:

```
> sacct -j 409 --format=alloctres%20

           AllocTRES
--------------------
               node=6
cpu=8,mem=0,node=1
cpu=6,mem=0,node=6
```


Bull
atos technologies

TRES for Lustre and Ofed

- Why ? Customers request !
  - Lustre filesystem accounting
  - OFED interconnect accounting
  - Profiling already available

- How ? TRES
  - New TRES can be easily added into Slurm (Developper)
  - “Simplifies” the introduction of new accounting information (Admin)
  - Slurm print functions (scontrol, squeue, sacct) ready for any TRES (User)

Bull
atos technologies

TRES : Lustre and Ofed

Lustre & Ofed Statistics though Slurm

- Only/already available with Profiling configured

- Lustre statistics requires configuring with acct_gather_filesystem/lustre
  - Statistics are obtained by the API from a file populated by the filesystem
    - /proc/fs/lustre filesystem (if it is mounted)

- For Ofed statistics requires configuring with acct_gather_interconnect/ofed
  - Statistics are obtained using MAD services (Management Datagram services)

Bull
atos technologies

Development part

- new TRES introduction
  - usage_disk (to replace existing local disk statistics)
  - usage_fs_lustre (lustre file system)
  - usage_ic_ofed (interconnect ofed)
- Account Gather Plugins expansion
  - Account Gather Filesystem
    - function to return Lustre statistics
  - Account Gather Interconnect
    - function to return Ofed statistics
- Job Account Gather Plugin
  - Modified to obtain Lustre statistics
  - Modified to obtain Ofed statistics

Bull
atos technologies

Database changes

- New Accounting statistics in step_table_fields for the TRES

  - tres_ave_usage_in (total usage/ # tasks) in mb
  - tres_max_usage_in (for a task) in mb
  - tres_max_usage_in_taskid
  - tres_max_usage_in_nodeid

  - tres_ave_usage_out (total usage / # tasks) in mb
  - tres_max_usage_out (for a task) in mb
  - tres_max_usage_out_taskid
  - tres_max_usage_out_nodeid

Bull
atos technologies

New Statistic Display

- New TRES Statistics can be displayed
  - sstat
    - by default when no options are designated
    - explicitly using --format options
  - sacct
    - only explicitly using --format options
- New --format options for both sstat and sacct
  - MaxDiskRead[Ø,Node,Task]
  - MaxDiskWrite[Ø,Node,Task]
  - AveDisk[Read,Write]
  - MaxUsageIn[Ø,N,T]Tres
  - MaxUsageOut[Ø,N,T]Tres
  - AveUsage[In,Out]Tres

Configuration

- To collect Lustre filesystem statistics
  - AcctGatherFilesystemType=acct_gather_filesystem/lustre
  - (default is AcctGatherFilesystemType=acct_gather_filesystem/none)
- To collect OFED infiniband statistics
  - AcctGatherInfinibandType=acct_gather_infiniband/ofed
  - (default is AcctGatherInfinibandType=acct_gather_infiniband/none)
- usage_disk statistics are collected by default (no configuration requirements)
- Everything already there if profiling activated !

Bull
atos technologies

Current status

Display Example

- sacct -j 264 --format=MaxUsageOutTres%78,MaxUsageoutNTres%78,
  MaxUsageOutTTres%78

```
                                                  MaxUsageOutTres
                                                  MaxUsageOutNTres
                                                  MaxUsageOutTTres
-------------------------------------------------------------------
usage_disk=24,usage_fs_lustre=16,usage_ic_ofed=6
usage_disk=1,usage_fs_lustre=28,usage_ic_ofed=3
usage_disk=3,usage_fs_lustre=1,usage_ic_ofed=18
```

Project Status

- Available to our Customers (Beta version)
  - installed on Bull & Customer test systems
- Targeted for release in an upcoming version of Slurm
  - ongoing discussions in Bugzilla
- Future enhancements
  - Support of other networks
    - BXI: Bull eXascale Interconnect

- Addition of new statistics to database greatly simplified
- Display of new TRES information almost transparent
- Customers are pleased with this new functionality

Bull
atos technologies

Thanks

For more information please contact:
Thomas.Cadeau@atos.net

Atos, the Atos logo, Atos Codex, Atos Consulting, Atos Worldgrid, Bull, Canopy, equensWorldline, Unify, Worldline and Zero Email are registered trademarks of the Atos group. September 2017. © 2017 Atos. Confidential information owned by Atos, to be used by the recipient only. This document, or any part of it, may not be reproduced, copied, circulated and/or distributed nor quoted without prior written approval from Atos.

Bull
atos technologies