TRES

Brian Christiansen
SchedMD

SLUG 2019

- What is a TRES
- Job Usage
- Limits
- Priority
- Billing TRES
- Reporting

TRES

TRES

- Usage Tracking and Limits
  - Previously only cpu and energy were tracked and limited.
- Trackable RESource (TRES)
  - BB (burst buffers)
  - Billing
  - CPU
  - Energy
  - FS (filesystem)
  - GRES
  - IC (interconnect)
  - License
  - Mem (Memory)
  - Node
  - Pages
  - VMem (Virtual Memory/Size)

TRES

- Configuration
  - slurm.conf
    - AccountingStorageTRES=
  - Default: Billing, CPU, Energy, Memory, Node, FS/Disk, Pages and VMem
    - Can't unset
    - AccountingStorageTRES=gres/gpu:tesla,license/iop1,bb/cray
  - Slurmctld pushes TRES to Slurmdbd
    - sacctmgr show tres

TRES

```
$ sacctmgr show tres
    Type            Name     ID
-------- --------------- ------
     cpu                      1
     mem                      2
  energy                      3
    node                      4
 billing                      5
      fs            disk      6
    vmem                      7
   pages                      8
    gres             gpu   1001
    gres       gpu:tesla   1002
 license           ansys   1003
    gres         gpu:k80   1004
      ic            ofed   1005
      fs          lustre   1006
 license         jobsize   1007
```

TRES: Job Usage

TRES: Job Usage

- scontrol show jobs
  - ReqTRES until running, AllocTRES after running
    - e.g. Requesting 1 cpu on a 2 threaded core with CR_CORE
  - Billing TRES not figured out until allocation
- sacct -o
  - ReqTRES
  - AllocTRES
  - TRESUsage<In|Out><Ave|Min|Max|Tot>
  - TRESUsage<In|Out><Min|Max>Node
  - TRESUsage<In|Out><Min|Max>Task

TRES: Limits

TRES: Limits

- Association, User
  - <Grp|Max>TRESMins
  - <Grp|Max>TRESRunMins
  - <Grp|Max>TRES
- QOS
  - GrpTRES
  - <Grp|Max>TRESMins
  - MaxTRESPerAccount
  - MaxTRESPerJob
  - MaxTRESPerNode
  - MaxTRESPerUser
  - MinTRES

TRES: Limits

- set/clear with sacctmgr
  - sacctmgr modify user bob set grptres=cpu=10,memory=200
  - sacctmgr modify user bob set grptres=cpu=-1,memory=-1
- Good to know
  - Limits stored in SlurmDBD
  - Usage tracked and stored (state files) in slurmctld
  - Limits enforced by Slurmctld
    - scontrol show assoc_mgr
      - [<users|qos|accounts>=<name1>[,...,<nameN>]]
      - flags=<users,assoc,qos>

TRES: Limits

```
$ scontrol show assoc users=brian accounts=stuff flags=assoc
...
ClusterName=lappy Account=stuff UserName= Partition= Priority=0 ID=3
   SharesRaw/Norm/Level/Factor=2147483647/0.00/5/0.00
   UsageRaw/Norm/Efctv=20612.12/1.00/1.00
   ParentAccount=root(1) Lft=86 DefAssoc=No
   GrpJobs=N(40) GrpJobsAccrue=N(29)
   GrpSubmitJobs=N(69) GrpWall=N(69.59)
   GrpTRES=cpu=500(80),mem=N(16000),energy=N(0),node=N(10),billing=N(160),
   GrpTRESMins=cpu=1000(171),mem=N(34350),energy=N(0),node=N(69),billing=N(343),...
   GrpTRESRunMins=cpu=N(80),mem=N(16000),energy=N(0),node=N(40),billing=N(160),...
   ...
ClusterName=lappy Account=stuff UserName=brian(1003) Partition= Priority=0 ID=4
   SharesRaw/Norm/Level/Factor=1/0.20/5/0.20
   UsageRaw/Norm/Efctv=20612.12/1.00/1.00
   ParentAccount= Lft=91 DefAssoc=Yes
   GrpJobs=N(40) GrpJobsAccrue=N(29)
   GrpSubmitJobs=N(69) GrpWall=N(69.59)
   GrpTRES=cpu=N(80),mem=N(16000),energy=N(0),node=N(10),billing=N(160),...
   GrpTRESMins=cpu=N(171),mem=N(34350),energy=N(0),node=N(69),billing=N(343),...
   GrpTRESRunMins=cpu=N(80),mem=N(16000),energy=N(0),node=N(40),billing=N(160),...
   MaxJobs= MaxJobsAccrue= MaxSubmitJobs= MaxWallPJ=
   ...
```

TRES: Priority

TRES: Priority

- PriorityWeightTRES
  - A comma separated list of TRES Types and weights that sets the degree that each TRES Type contributes to the job's priority
  - PriorityWeightTRES=CPU=1000,Mem=2000,GRES/gpu=3000
- By default, normalized against Partition's on-node resources (e.g. cpu, memory, gres) and against global resources (e.g. licenses, bb)
- In 19.05, PriorityFlags=NO_NORMAL_TRES was added to not normalize
  - NO_NORMAL_ALL
  - NO_NORMAL_ASSOC
  - NO_NORMAL_PART
  - NO_NORMAL_QOS

TRES: Priority

- Ex. If a partition has 80 cpus and a job uses 8, then the priority factor is .1 (or 10%)
- AccountingStorageTRES=cpu,mem,gres/gpu
- PriorityWeightTRES=cpu=1000,gres/gpu=3000

```
$ sprio
           JOBID    PRIORITY        AGE  FAIRSHARE                TRES
               3         625          0        500              cpu=125
               5         600          0        500              cpu=100
               6         812          0        500   cpu=12,gres/gpu=75
```

TRES: Billing

TRES: Billing

- Billing TRES
  - Billing is a value that represents multiple TRES
  - Previously, only cpu was accounted for in fairshare.
    - Jobs only cpu usage even if used 1 cpu and all the memory on the node
  - Added as a TRES in 17.11
    - Limits and Usage
- Calculated on a per-partition basis
  - TRESBillingWeights="CPU=1.0,Mem=0.25G,GRES/gpu=2.0"
- Two methods of calculating billable TRES
  - MAX_TRES
  - SUM of TRES

TRES: Billing

- SUM of TRES
  - Default
  - SUM(<TRES> + <TRES Weight>, …)
- MAX_TRES
  - PriorityFlags=MAX_TRES
  - MAX(Node TRES) + SUM(Global TRES)

TRES: Billing

- TRESBillingWeights=CPU=1.0,Mem=0.25
- 16CPU, 64GB nodes

SUM of TRES:

```
       CPUs        Mem GB
Job1: (1 *1.0) + (60*0.25) = (1 + 15) = 16
Job2: (16*1.0) + (1 *0.25) = (16+.25) = 16.25
Job3: (16*1.0) + (60*0.25) = (16+ 15) = 31
```

MAX_TRES:

```
           CPUs       Mem GB
Job1: MAX((1 *1.0), (60*0.25)) = 15
Job2: MAX((15*1.0), (1 *0.25)) = 15
Job3: MAX((16*1.0), (64*0.25)) = 16
```

TRES: Billing

- The Billing TRES is calculated from a partition's TRESBillingWeights. Though TRES weights on a partition may be defined as doubles, the Billing TRES values for a job are stored as integers. This is not the case when calculating a job's fairshare where the value is treated as a double.

TRES: sreport

TRES: sreport

- Need more memory? Or less cpus?

```
$ sreport -tminper cluster utilization --tres="cpu,mem" start=2015-09-01T00:00:00
--------------------------------------------------------------------------------
Cluster Utilization 2015-09-01T00:00:00 - 2015-09-01T23:59:59
Use reported in TRES Minutes/Percentage of Total
--------------------------------------------------------------------------------
  Cluster      TRES Name           Allocated                Down          PLND Down             Reserved                Idle             Reported
--------- -------------- -------------------- ------------------ ------------------ -------------------- ------------------ --------------------
    compy            cpu       253440(20.00%)           0(0.00%)           0(0.00%)             0(0.00%)     1013760(80.00%)     1267200(100.00%)
    compy            mem    4582306080(90.00%)           0(0.00%)           0(0.00%)    509145120(10.00%)           0(0.00%)  5091451200(100.00%)
```

TRES: sreport

- GPUs being used?

```
$ sreport -tminper cluster utilization --tres="cpu,gres/gpu" start=2015-09-02T00:00:00
--------------------------------------------------------------------------------
Cluster Utilization 2015-09-02T00:00:00 - 2015-09-2T23:59:59
Use reported in TRES Minutes/Percentage of Total
--------------------------------------------------------------------------------
  Cluster      TRES Name           Allocated                Down          PLND Down             Reserved                Idle             Reported
--------- -------------- -------------------- ------------------ ------------------ -------------------- ------------------ --------------------
    compy            cpu      1140480(90.00%)           0(0.00%)           0(0.00%)       126720(10.00%)           0(0.00%)      1267200(100.00%)
    compy       gres/gpu        63360(20.00%)           0(0.00%)           0(0.00%)              0(0.00%)     253440(80.00%)       316800(100.00%)
```

TRES: sreport

- Which GPUs are being used most?

```
$ sreport -tminper cluster utilization --tres="gres/gpu:k40,gres/gpu:k80" start=2015-09-02T00:00:00
--------------------------------------------------------------------------------
Cluster Utilization 2015-09-02T00:00:00 - 2015-09-2T23:59:59
Use reported in TRES Minutes/Percentage of Total
--------------------------------------------------------------------------------
  Cluster      TRES Name           Allocated                Down          PLND Down             Reserved                Idle            Reported
--------- -------------- -------------------- ------------------ ------------------ -------------------- ------------------ --------------------
    compy   gres/gpu:k40        63360(20.00%)           0(0.00%)           0(0.00%)             0(0.00%)     253440(80.00%)      316800(100.00%)
    compy   gres/gpu:k80       190080(60.00%)           0(0.00%)           0(0.00%)             0(0.00%)     126720(40.00%)      316800(100.00%)
```

TRES: sreport

- In sreport, the "Reported" Billing TRES is calculated from the largest Billing TRES of each node multiplied by the time frame. For example, if a node is part of multiple partitions and each has a different TRESBillingWeights defined the Billing TRES for the node will be the highest of the partitions. If TRESBillingWeights is not defined on any partition for a node then the Billing TRES will be equal to the number of CPUs on the node.

Questions?

- https://slurm.schedmd.com/tres.html