Heterogeneous Resources and MPMD
(aka Job Pack)

Slurm 2015 User Group : Rod Schultz, Atos
Martin Perry, Atos
Matthieu Hautreaux, CEA
Yiannis Georgiou, Atos

What Is Slurm?

- Simple
- Linux
- Utility
- Resource
- Manager

Simple

- Slurm provides a SPMD (Single Program Multiple Data) environment.

  srun –N4 –Cgreen –gres=gpu myapp

- All nodes in an allocation have identical resources
  - 4 nodes are allocated, all have the feature green, all have a gpu

- All tasks execute the same application
  - myapp launched on all nodes.

Not Quite so Simple

- Slurm currently has limited MPMD (Multiple Program Multiple Data) support.
- The –multi-program option allows multiple programs to be executed, but the allocation is still homogenous.

Definitely Not Simple

- In some cases it is desirable to have nodes with different characteristics as part of the same step.
  - A node with lots of memory for the serial startup/wrapup phase.
  - Lots of nodes with GPU for the parallel phase.
  - Nodes with Fast I/O to store the results.
  - And these nodes run different executables that are part of the same MPI_Comm_World

- Kind of like multiple sruns scheduled so they run at the same time.
  - We do this by 'packaging' a set of jobs, or a ***Job-Pack***

Introducing Job Pack

Stand-Alone-Srun

- srun –N 1 –Cbig ./controller : -N 1000 –gres=gpu ./worker : -N 10 –pIO ./saver
- Colon separated list of Job Descriptions

  srun job_description_0 : job_description_1 : job_description_2
- A Job Description is essentially the full set of options available for a normal srun
  - (some ‘reasonable’ restrictions)

Introducing Job Pack …

Internally we ‘extend’ -- dependency

- srun -dpack –N 1 –Cbig ./controller
  - Gets job_id=101
- srun –dpack –N 1000 –gres=gpu ./worker
  - Gets job_id=102
- srun –dpackleader:101:102 -N 10 –pIO ./saver

Job Pack …

- Each job description defines a member job.
  - Each member job is a separate job from a scheduling point of view (unique jobid)
  - The first job member is called the pack_leader.
  - All resources allocated when the pack_leader is allocated. The other members are in a pending state
  - Job descriptors have an associated index called a pack_group. The leader is 0, increasing by 1, left to right on the command line.

```
srun –JLdr –pt96big controller : -JMbr1 -N2 –n4 –pt96gpu worker : -JMbr2 –N2 –t96iopx saver &
squeue
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
             47959   t96iopx     Mbr2    slurm  R       0:06      2 trek[8-9]
             47960    t96gpu     Mbr1    slurm  R       0:06      2 trek[4-5]
             47961    t96big      Ldr    slurm  R       0:06      1 trek7
```

Job Pack …

- All steps of srun are launched at the same time.

```
srun -JLdr -pt96big controller : -JMbr1 -N2 -n4 -pt96gpu worker : -JMbr2 -N2 -pt96iopx saver &
squeue
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
             47959   t96iopx     Mbr2    slurm  R       0:06      2 trek[8-9]
             47960    t96gpu     Mbr1    slurm  R       0:06      2 trek[4-5]
             47961    t96big      Ldr    slurm  R       0:06      1 trek7
```

- One task executing **controller** on trek7 (the pack_leader)
- Two tasks executing **worker** on both trek[4-5] (member_1)
- One task executing **saver** on both trek[8-9] (member_2)

salloc and sbatch

- Colon separated list of Job Descriptions

```
salloc –JLdr –pt96big {command} : -JMbr1 -N2 –n4 –pt96gpu : -JMbr2 –N2 –pt96iopx
sbatch –JLdr –pt96big script : -JMbr1 -N2 –n4 –pt96gpu : -JMbr2 –N2 –pt96iopx
```

- Both create allocations of multiple jobs.
- The salloc command or sbatch script is only allowed on the leader, and executes on the first node of the pack leaders allocation.
- salloc without a command opens a terminal session on the first node of the pack leaders allocation.
- The script or terminal session then execute ‘step launch sruns’

Step Launch Srun

- A salloc/sbatch script may contain multiple step launch sruns.
- srun step_description : step_description : step_description
- A step description is **--pack-group=[#,#-#] command** which specifies that *command* executes on the allocations of the set of *pack_group* jobs
  - # are pack_groups
  - A pack group can be specified on more than one step_description.

Step Launch Srun ...

```
sbatch -JLdr -pt96big doit.sh : -JMbr1 -N2 -n4 -pt96gpu : -JMbr2 -N2 -pt96iopx

Script doit.sh contains
   srun -pack-group=0 controller : --pack-group=[0-1] worker : --pack-group=2 storer

Assume    pack_group=0 is allocated trek7
          pack_group=1 is allocated trek[4-5]
          pack_group=2 is allocated trek[8-9]

trek7 will have 1 task running controller from step_description_0
trek7 will have 1 task running worker from step_description_1
trek[4-5] each have 2 tasks running worker from step_description_1
trek[8-9] each have 1 task running storer from step_description_2
```

Example

Here is an example of running three different MPI executables on three different resource requirements

```
$ srun -JLdr -w trek7 ./controller : -JMbr1 --gres=gpu -N2 --tasks-per-node=2 ./worker
                                   : -JMbr2 -pt96iopx -N2 ./storer &

$ squeue
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
             61661   t96iopx     Mbr2    slurm  R       0:03      2 trek[8-9]
             61662   trekall     Mbr1    slurm  R       0:03      2 trek[4-5]
             61663   trekall      Ldr    slurm  R       0:03      1 trek7

This is the controller, Name=Ldr Id=61663 MPI Rank 0 of 7, Running on host trek7
This is a workaholic, Name=Mbr1 Id=61662 MPI Rank 1 of 7, Running on host trek4
This is a workaholic, Name=Mbr1 Id=61662 MPI Rank 2 of 7, Running on host trek4
This is a workaholic, Name=Mbr1 Id=61662 MPI Rank 3 of 7, Running on host trek5
This is a workaholic, Name=Mbr1 Id=61662 MPI Rank 4 of 7, Running on host trek5
This is a storer, Name=Mbr2 Id=61661 MPI Rank 5 of 7, Running on host trek8
This is a storer, Name=Mbr2 Id=61661 MPI Rank 6 of 7, Running on host trek9
```

Status

- Job Packs will be available in the next release (16_5)
- We have a working prototype that has most of the features implemented for error free requests and error free jobs.

Thanks

For more information please contact:
T+ 33 1 98765432
F+ 33 1 88888888
M+ 33 6 44445678
rod.schultz@atos.net, yiannis.georgio@atos.net

Bull
atos technologies