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

```bash
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.
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.
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`
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`

<table>
<thead>
<tr>
<th>JOBID</th>
<th>PARTITION</th>
<th>NAME</th>
<th>USER</th>
<th>ST</th>
<th>TIME</th>
<th>NODES</th>
<th>NODELIST(REASON)</th>
</tr>
</thead>
<tbody>
<tr>
<td>47959</td>
<td>t96iopx</td>
<td>Mbr2</td>
<td>slurm</td>
<td>R</td>
<td>0:06</td>
<td>2</td>
<td>trek[8-9]</td>
</tr>
<tr>
<td>47960</td>
<td>t96gpu</td>
<td>Mbr1</td>
<td>slurm</td>
<td>R</td>
<td>0:06</td>
<td>2</td>
<td>trek[4-5]</td>
</tr>
<tr>
<td>47961</td>
<td>t96big</td>
<td>Ldr</td>
<td>slurm</td>
<td>R</td>
<td>0:06</td>
<td>1</td>
<td>trek7</td>
</tr>
</tbody>
</table>
All steps of `srun` are launched at the same time.

```
srun -JLdr -pt96big controller : -JMbr1 -N2 -n4 -pt96gpu worker : -JMbr2 -N2 -pt96iopx saver & squeue
```

<table>
<thead>
<tr>
<th>JOBID</th>
<th>PARTITION</th>
<th>NAME</th>
<th>USER</th>
<th>ST</th>
<th>TIME</th>
<th>NODES</th>
<th>NODELIST(REASON)</th>
</tr>
</thead>
<tbody>
<tr>
<td>47959</td>
<td>t96iopx</td>
<td>Mbr2</td>
<td>slurm</td>
<td>R</td>
<td>0:06</td>
<td>2</td>
<td>trek[8-9]</td>
</tr>
<tr>
<td>47960</td>
<td>t96gpu</td>
<td>Mbr1</td>
<td>slurm</td>
<td>R</td>
<td>0:06</td>
<td>2</td>
<td>trek[4-5]</td>
</tr>
<tr>
<td>47961</td>
<td>t96big</td>
<td>Ldr</td>
<td>slurm</td>
<td>R</td>
<td>0:06</td>
<td>1</td>
<td>trek7</td>
</tr>
</tbody>
</table>

- 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
  
  ```
  alloc -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 alloc command or sbatch script is only allowed on the leader, and executes on the first node of the pack leaders allocation.

- alloc 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’
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.
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
Here is an example of running three different MPI executables on three different resource requirements

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

$ squeue

<table>
<thead>
<tr>
<th>JOBID</th>
<th>PARTITION</th>
<th>NAME</th>
<th>USER</th>
<th>ST</th>
<th>TIME</th>
<th>NODES</th>
<th>NODELIST(REASON)</th>
</tr>
</thead>
<tbody>
<tr>
<td>61661</td>
<td>t96iopx</td>
<td>Mbr2</td>
<td>slurm</td>
<td>R</td>
<td>0:03</td>
<td>2</td>
<td>trek[8-9]</td>
</tr>
<tr>
<td>61662</td>
<td>trekall</td>
<td>Mbr1</td>
<td>slurm</td>
<td>R</td>
<td>0:03</td>
<td>2</td>
<td>trek[4-5]</td>
</tr>
<tr>
<td>61663</td>
<td>trekall</td>
<td>Ldr</td>
<td>slurm</td>
<td>R</td>
<td>0:03</td>
<td>1</td>
<td>trek7</td>
</tr>
</tbody>
</table>
```

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

Atos, the Atos logo, Atos Consulting, Atos Worldgrid, Worldline, BlueKiwi, Bull, Canopy the Open Cloud Company, Yunano, Zero Email, Zero Email Certified and The Zero Email Company are registered trademarks of the Atos group. August 2015. © 2015 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.