Cloudy, With a Chance of Dynamic Nodes

Nick Ihli
SchedMD

Slurm User Group Meeting 2022

Agenda - US Mountain Time (UTC-6)

Time	Speaker	Title
9:00 - 9:50	Jason Booth	Field Notes 6: From The Frontlines of Slurm Support
10:00 - 10:20	Ole Nielsen (DTU)	Pathfinding into the clouds
10:30 - 10:55	Nate Rini	OCI Containers, and scrun
11:00 - 11:20	Wei Feinstein (LBNL)	LBNL Site Report
11:30 - 11:55	Nick Ihli	Cloudy, With A Chance of Dynamic Nodes
12:00 - 12:20	Kota Tsuyuzaki (NTT)	Burst Buffer Lua Plugin for Lustre
12:30 - 12:55	Tim Wickberg	Slurm 22.05, 23.02, and Beyond

Welcome

- Seven separate presentations, seven separate streams
- Presentations are available through the SchedMD Slurm YouTube channel
  - https://youtube.com/c/schedmdslurm
- Or through direct links from the agenda
  - https://slurm.schedmd.com/slurm_ug_agenda.html

Asking questions

- Feel free to ask questions throughout through YouTube's chat
- Chat is moderated by SchedMD staff
  - Tim McMullan, Ben Roberts, and Tim Wickberg
  - Also identified by the little wrench symbol next to their name
- For SchedMD presentations:
  - Questions will be relayed to the presenter by the moderators
  - Some may be deferred to the end if they cannot be relayed in a timely fashion
  - Or some may be answered by the moderators in chat directly
- For community presentations:
  - Please ask questions in the live chat
  - The presenter (if available) may respond through chat
  - Or SchedMD staff may try to answer in their absence

Cloudy, With a Chance of Dynamic Nodes

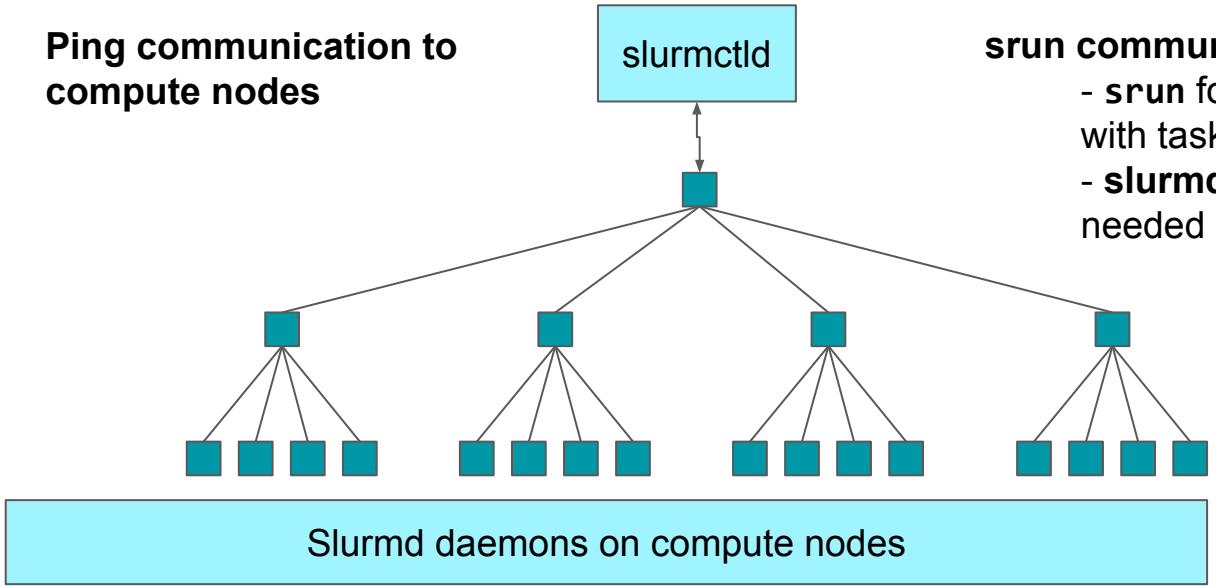
Nick Ihli
SchedMD

Adding Static Nodes

- Recommended process for adding a node
  - Stop the slurmctld daemon
  - Update the slurm.conf file on all nodes in the cluster
  - Restart the slurmd daemons on all nodes
  - Restart the slurmctld daemon
- Hierarchical communication with configurable fanout
  - Better efficiency for large clusters and large MPI jobs
  - Less overhead on the controller

Fanout

Ping communication to compute nodes



srun communication

- **srun** forwards credential with task info to **slurmd**
- **slurmd** forwards request as needed (per fanout)

(Hierarchical communications with configurable fanout)

Enter - Dynamic Nodes

- Nodes added/deleted from system without adding them into **slurm.conf**, restarting **slurmctld** or **slurmd**
- Use Cases
  - Multiple dynamic clusters, where nodes are added/removed frequently
  - Temporary addition of a new node(s)
  - Cloud services adding/removing nodes

Dynamic Nodes Configuration

- slurm.conf configuration
  - TreeWidth=65533
    - Fanout must be disabled
    - Dynamic nodes rely on alias_list for communications

Dynamic Nodes Configuration

- slurm.conf configuration
  - MaxNodeCount
    - Sets the number of nodes that can exist in the system
    - Minimally set to nodes read from slurm.conf
    - If not set, MaxNodeCount will be set to the number of nodes read in from slurm.conf
  - SelectType=select/cons_tres
    - Dynamic nodes are only supported with cons_tres
  - The “**cloud_dns**” SlurmctldParameter must **NOT** be set as this disables the alias list.

Adding Dynamic Nodes

- Two ways to add a Dynamic Node
  - 1. dynamic registrations
    - slurmd -Z --conf=”xyz”
      - **-Z** - Tells Slurm the node is registering as a dynamic node
      - **--conf** - Defines additional parameters of a dynamic node using the same syntax and parameters used to define nodes in the slurm.conf.
      - Not allowed in --conf=""
        - NodeName=
      - If no hw topology specified, slurmd will use hw configuration (slurmd -C)
      - If any hw topology specified, then slurmd will use what’s specified and not add to it.

Adding Dynamic Nodes

- If slurmd -C reports:

```
NodeName=node1 CPUs=16 Boards=1 SocketsPerBoard=1 CoresPerSocket=8 ThreadsPerCore=2 RealMemory=31848
```

- These --conf specifications will generate the corresponding node definitions:

```
--conf "Gres=gpu:2"
```

```
NodeName=node1 CPUs=16 Boards=1 SocketsPerBoard=1 CoresPerSocket=8
ThreadsPerCore=2 RealMemory=31848 Gres=gpu:2
```

```
--conf "CPUs=16 RealMemory=30000 Gres=gpu:2"
```

```
NodeName=node1 CPUs=16 RealMemory=30000 Gres=gpu:2"
```

Adding Dynamic Nodes

- Two ways to add a Dynamic Node
  - 2. scontrol
    - scontrol create NodeName= [conf syntax]
      - only State=cloud and State=future supported
    - No node is actually registered or started with this method, but a new node object is created that could be “Resumed” using the cloud Power_Save plugin or added as a Future node.

```
> scontrol create NodeName=node[0-99] CPUs=16 Boards=1 SocketsPerBoard=1
CoresPerSocket=8  ThreadsPerCore=2 RealMemory=31848 Gres=gpu:2
State=CLOUD
```

Slurm Configuration Files

- Configless or local/shared slurm.conf still work as before
  - When using configless:
    - gres.conf - recommend using “autodetect=nvml” in the central gres.conf, otherwise it would require all future dynamic nodes with a gres listed in gres.conf
    - Other option is to have a local gres.conf with autodetect=nvml or the node configuration

Adding Node to Partitions

- By default nodes aren't added to any partition
  - Two methods to automatically add dynamic nodes to a partition
    - 1. Nodes=All
      - If configured in the partition definition, the partition will always have all nodes in the partition, even new dynamic nodes

```
PartitionName=open Nodes=ALL MaxTime=INFINITE Default=Yes  State=Up
```

Adding Node to Partitions

- By default nodes aren't added to any partition
  - Two methods to automatically add dynamic nodes to a partition
    - 2. Nodesets
      - Create nodesets, add the nodeset to the partition. When registering the dynamic node, configure it with a feature to add it to the nodeset.

```
Nodeset=ns1 Feature=f1
Nodeset=ns2 Feature=f2

PartitionName=all Nodes=ALL
PartitionName=p1 Nodes=ns1
PartitionName=p2 Nodes=ns2
PartitionName=p3 Nodes=ns1,ns2
```

```
> slurmd -Z --conf="Feature=f1"
```

Deleting Dynamic Nodes

- To remove a dynamic node you must manually delete the node
  - scontrol delete nodename=<nodelist>
    - Nodes can't be deleted unless they are idle
    - Clear node from reservations
  - Stop the slurmd on the compute node

Cloudy Things

sreport + cloud

- Cloud+PoweredDown now show as Planned Down in database

```
Usage reported in CPU Minutes
--------------------------------------------------------------------------------
  Cluster Allocate     Down PLND Down     Idle  Planned Reported
--------- -------- -------- --------  -------- -------- --------
 smallgpu        0        0   230400        0        0   230400
```

Configless - Include file support

- **Include** file configuration now pushed along with other Slurm conf files
- Makes cloud configurations easier, especially hybrid!
  - Right now the slurm.conf is either baked into the image or on a shared file system
  - Most Cloud configurations with use an “Include” in the slurm.conf

```
> slurm.conf

ClusterName=cluster1
SlurmCtldHost=cntrlnode
...
...

Include cloud.conf
```

Cloud Partner update

- Google
  - HPC Toolkit integrated with v4
  - v5 scripts released in May - https://github.com/SchedMD/slurm-gcp
    - Improved error handling and debugging capabilities of Resume/Suspend scripts
    - Accounting data dumped to BigQuery
    - Terraform is source of truth for Infrastructure and Slurm configuration
    - Turnkey hybrid configuration
  - Sponsored development for next version of slurm-gcp scripts

Cloud Partner update

- AWS
  - ParallelCluster 3.2.0 supports memory-based scheduling - https://docs.aws.amazon.com/parallelcluster/latest/ug/slurm-mem-based-scheduling-v3.html
  - Next ParallelCluster release:
    - Upgrading to Slurm 22.05
    - Easier to enable Slurm Accounting
    - Greater flexibility in mapping Slurm nodes to heterogeneous EC2 instances
    - Support for cost and capacity optimized strategies
  - Sponsored development projects for Slurm 23.02

Cloud Partner update

- Microsoft
  - Continuing collaboration for Slurm on Azure/CycleCloud
  - Official Github Repo
    - https://github.com/Azure/cyclecloud-slurm

Questions?

End Of Stream

- Thanks for watching!