SHARCNET General Interest Webinar Series

How jobs are scheduled to run on Graham and Cedar

James Desjardins
High Performance Computing Consultant
SHARCNET, Brock University
July 19th, 2017

Overview

Documentation and getting help

Scheduling basics: node resources and resource requests (jobs)

Job queue basics: factors that affect the order of jobs in queue (priority)

Cluster resource basics: categorization of resources that affect priority (partitions)

Monitoring jobs, the queue and the cluster

Documentation and getting help

Slurm Documentation

- https://slurm.schedmd.com/
- https://slurm.schedmd.com/pdfs/summary.pdf

Compute Canada wikis

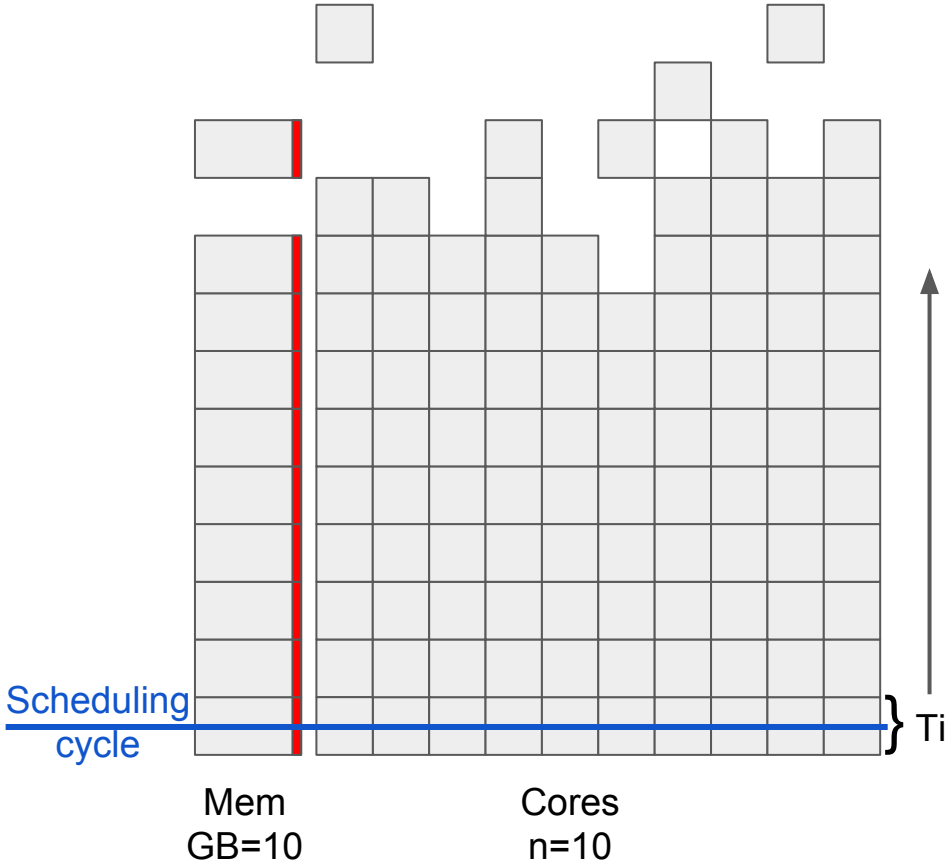
- https://docs.computecanada.ca/wiki/Graham https://docs.computecanada.ca/wiki/Cedar
- https://docs.computecanada.ca/wiki/Running_jobs
- https://docs.computecanada.ca/wiki/Job_scheduling_policies
- https://docs.computecanada.ca/wiki/Known_issues

Compute Canada YouTube introduction playlist

- https://www.youtube.com/channel/UC2f3cwviToj-mazutBNhzFw

support@computecanada.ca

Scheduling basics: node resources and resource requests (job queue)



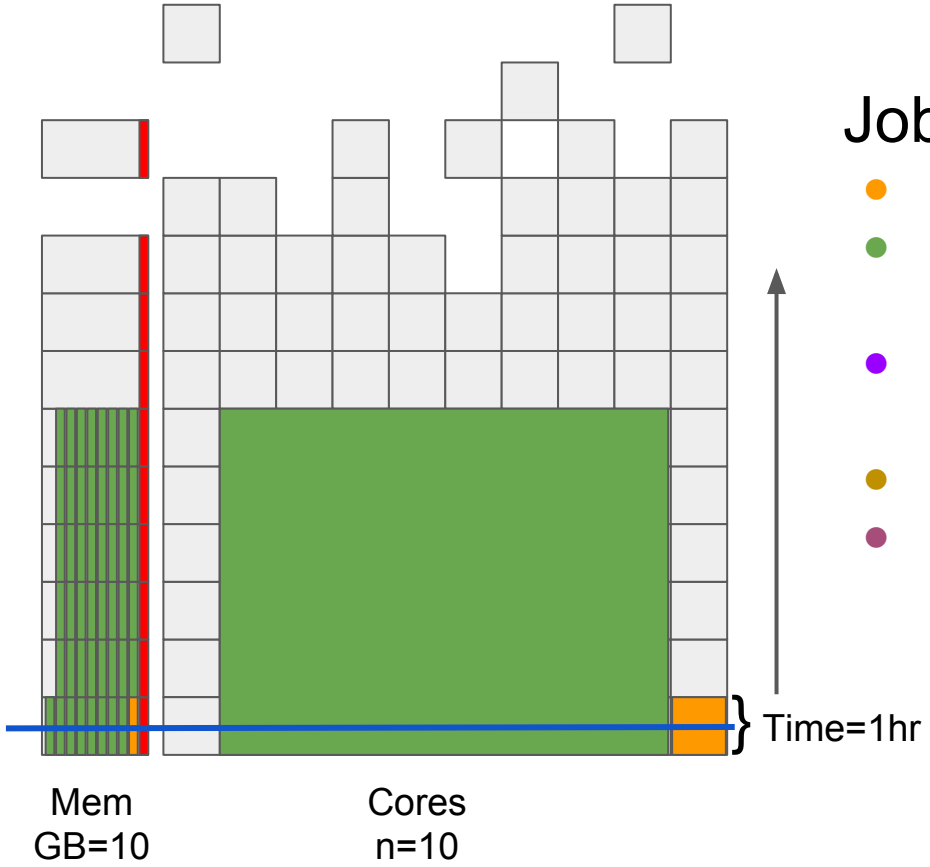
Node resources

- Cores
- Memory

...etc

- GPUs
- Software licenses
- …
- Time

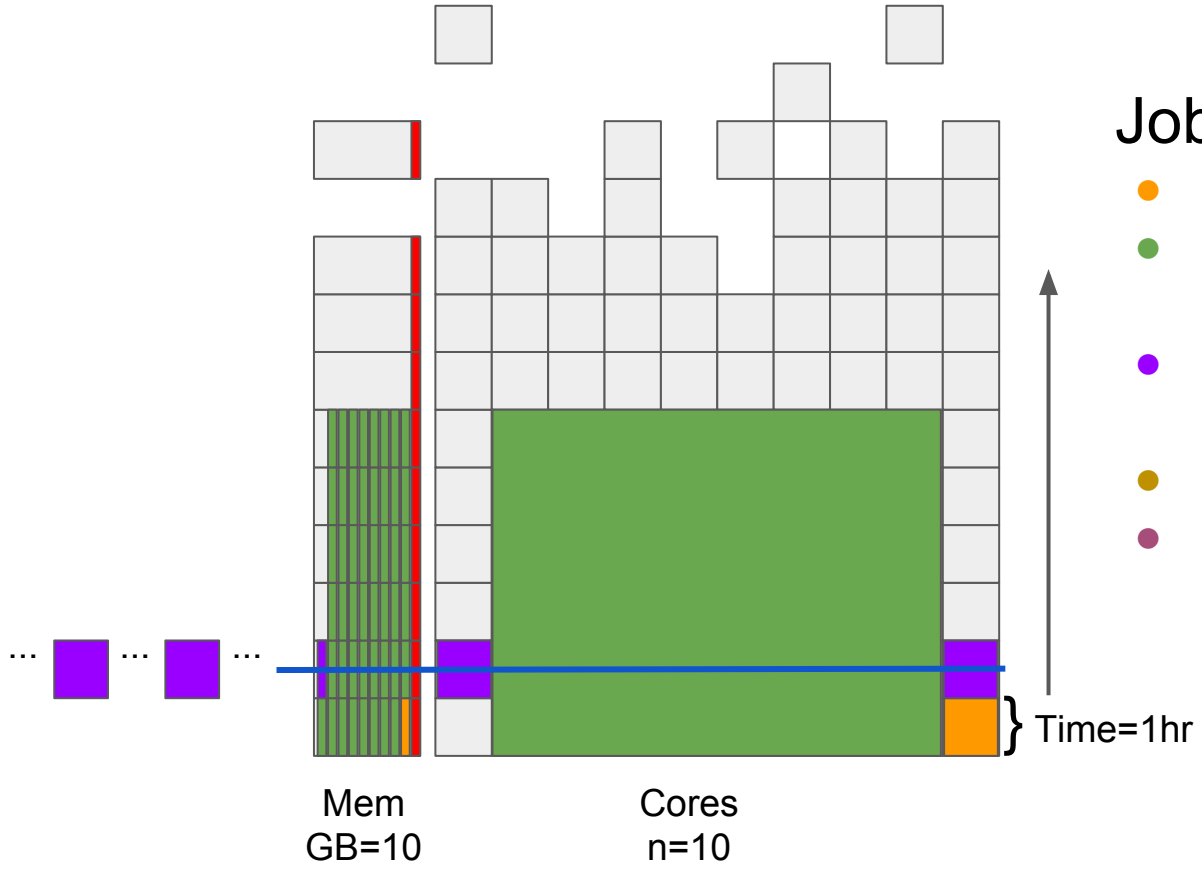
Scheduling basics: node resources and resource requests (job queue)



Job size

- --time=1:00 --mem=1G
- --time=6:00 --mem=8G --cpus-per-task=8
- --time=1:00 --ntasks=8 --mem-per-cpu=400
- --time=2:00 --mem=9G
- --time=1:00 --nodes=1 --n-tasks-per-node=10 --mem-per-cpu=400

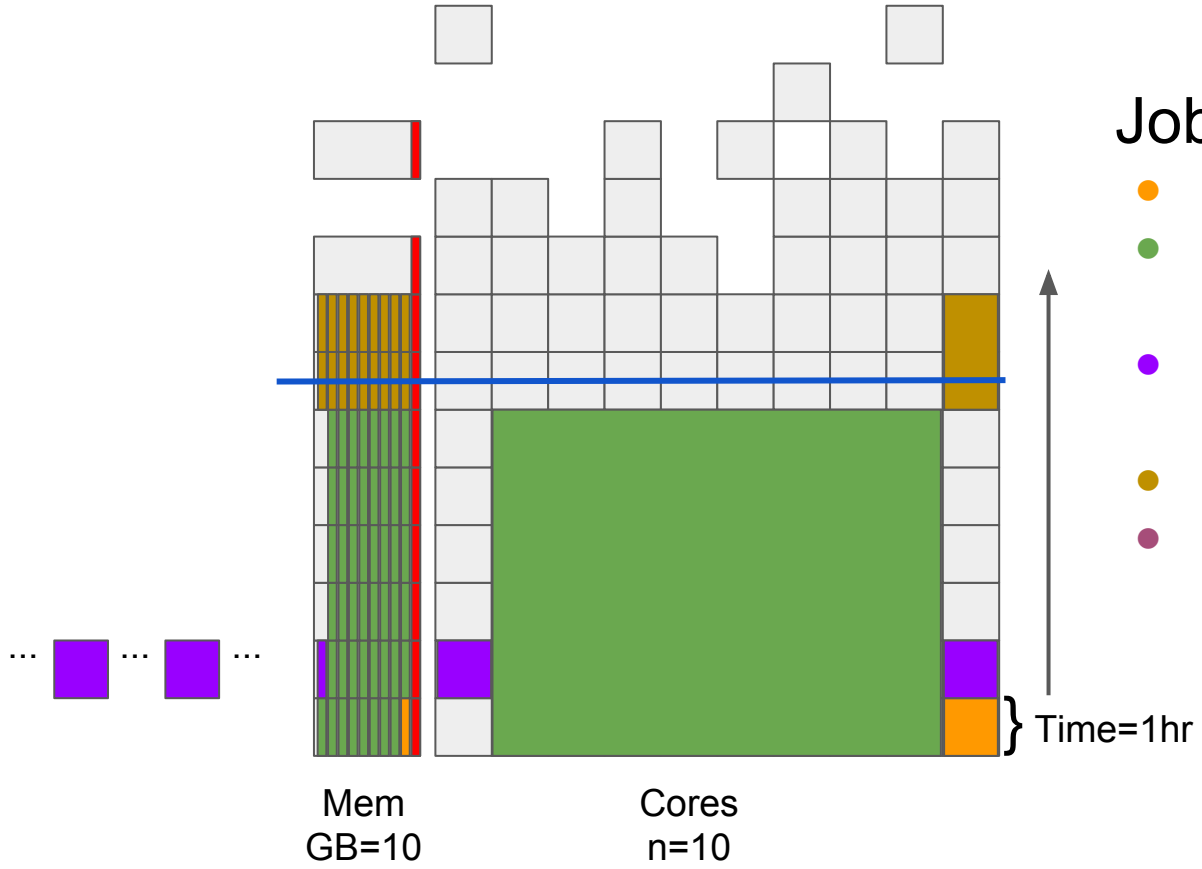
Scheduling basics: node resources and resource requests (job queue)



Job size

- --time=1:00 --mem=1G
- --time=6:00 --mem=8G --cpus-per-task=8
- --time=1:00 --ntasks=8 --mem-per-cpu=400
- --time=2:00 --mem=9G
- --time=1:00 --nodes=1 --n-tasks-per-node=10 --mem-per-cpu=400

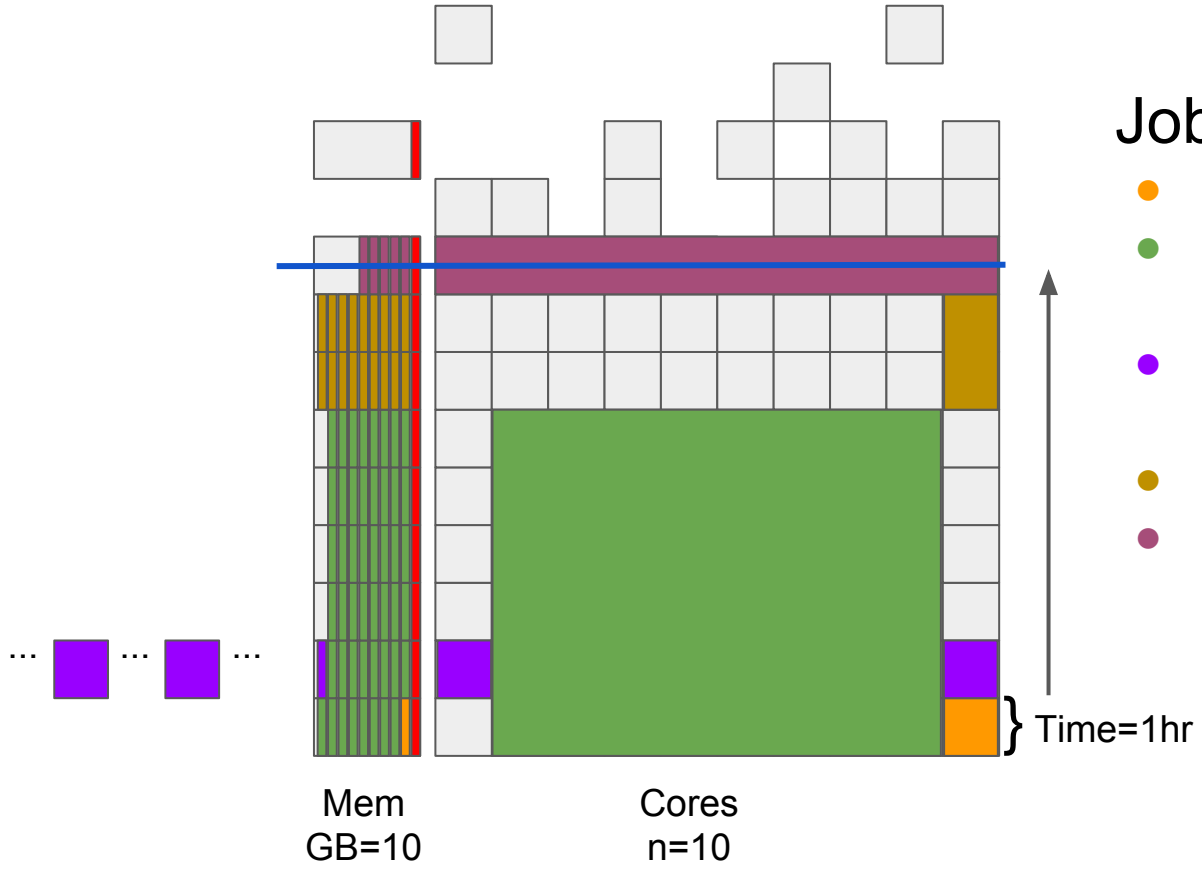
Scheduling basics: node resources and resource requests (job queue)



Job size

- --time=1:00 --mem=1G
- --time=6:00 --mem=8G --cpus-per-task=8
- --time=1:00 --ntasks=8 --mem-per-cpu=400
- --time=2:00 --mem=9G
- --time=1:00 --nodes=1 --n-tasks-per-node=10 --mem-per-cpu=400

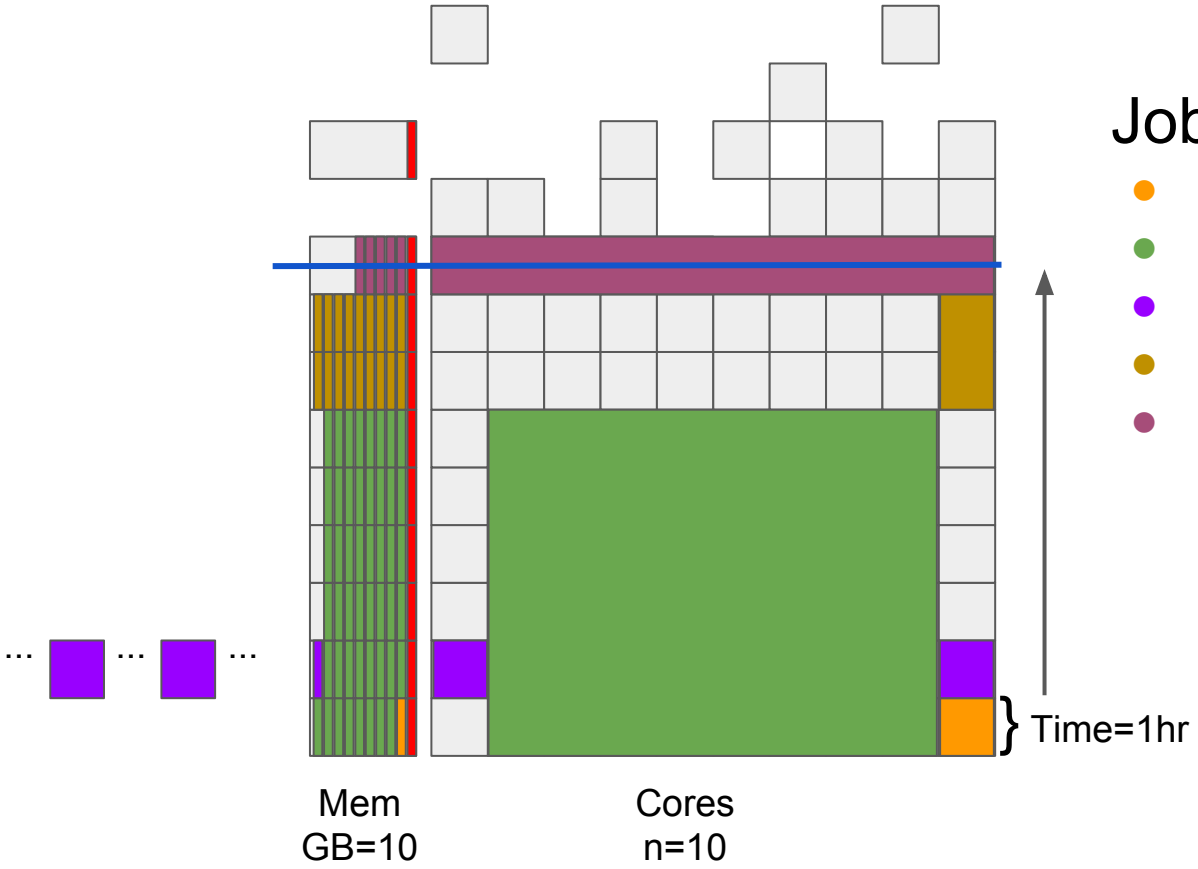
Scheduling basics: node resources and resource requests (job queue)



Job size

- --time=1:00 --mem=1G
- --time=6:00 --mem=8G --cpus-per-task=8
- --time=1:00 --ntasks=8 --mem-per-cpu=400
- --time=2:00 --mem=9G
- --time=1:00 --nodes=1 --ntasks-per-node=10 --mem-per-cpu=400

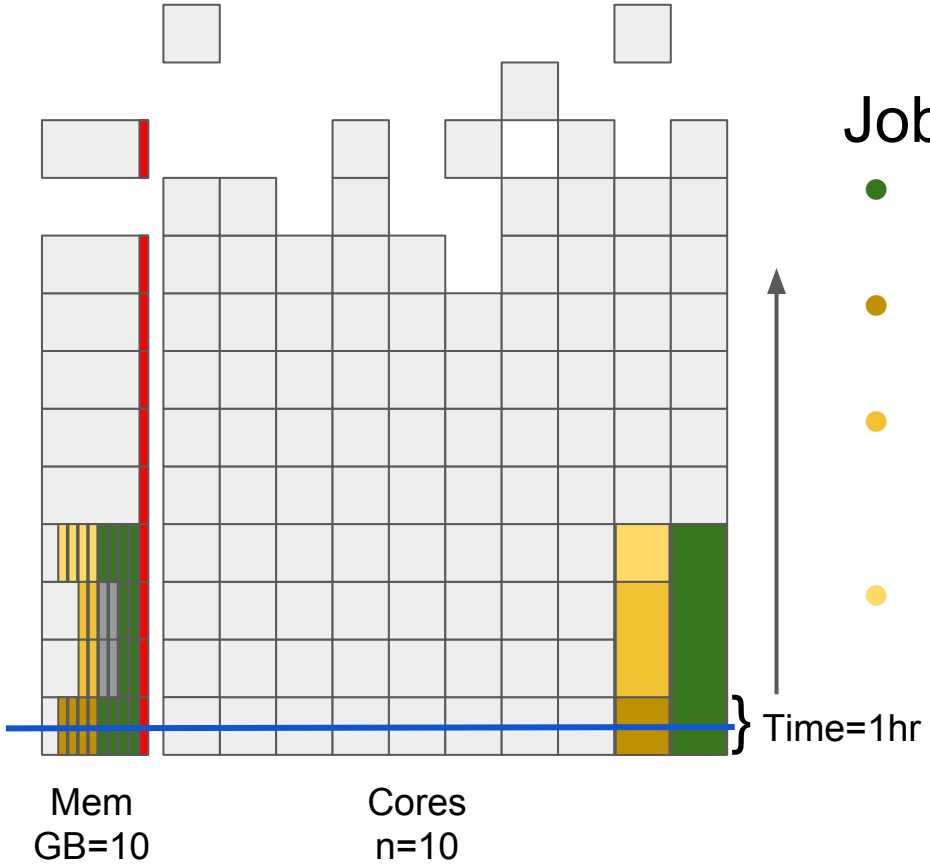
Scheduling basics: node resources and resource requests (job queue)



Job billing (by core & mem)

- 1 core
- 8 core
- 8 core
- 9 core
- 10 core

Scheduling basics: node resources and resource requests (job queue)



Job dependencies

- jobid 1
  - --time=4:00 --mem=4G
- jobid 2
  - --time=1:00 --mem=4G
- jobid 3
  - --time=2:00 --mem=2G --dependency=afterok:2
- jobid 4
  - --time=1:00 --mem=4G --dependency=afterok:3

Job queue basics: factors that affect the order of jobs in queue (priority)

Job size

- The shape of requested resources affects a job's priority

Age

- A jobs duration in the queue affects its priority (for FIFO this is the only factor)

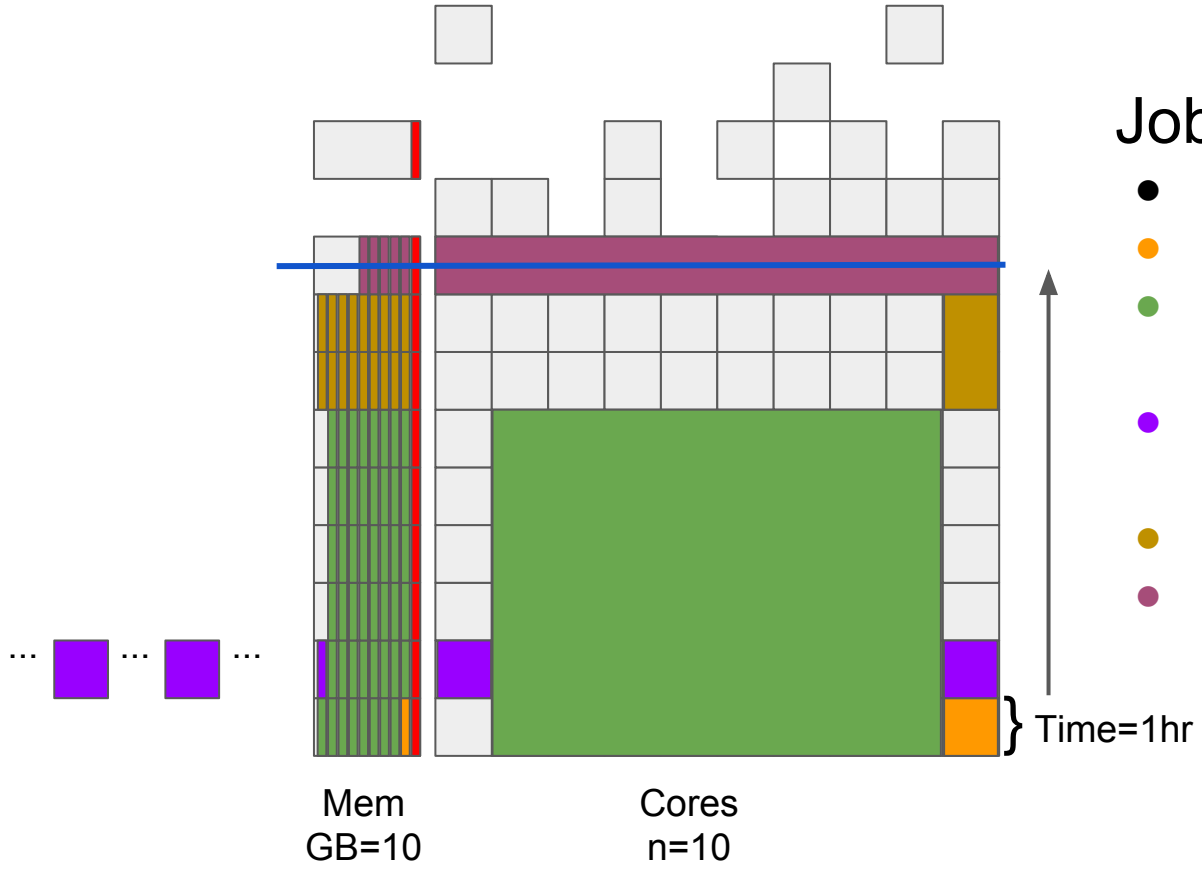
Fair-share

- An account's past usage affects the priority of queued jobs

Partition

- The classification of node sets interacts with job size in determining priority

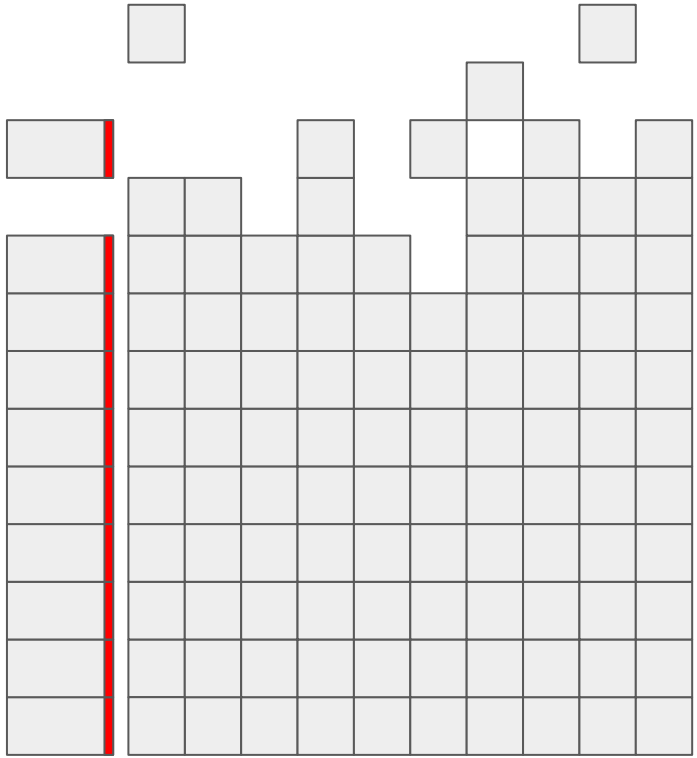
Job queue basics: factors that affect the order of jobs in queue (priority)



Job age

- FIFO (First In First Out)
- --time=1:00 --mem=1G
- --time=6:00 --mem=8G --cpus-per-task=8
- --time=1:00 --ntasks=8 --mem-per-cpu=400
- --time=2:00 --mem=9G
- --time=1:00 --nodes=1 --ntasks-per-node=10 --mem-per-cpu=400

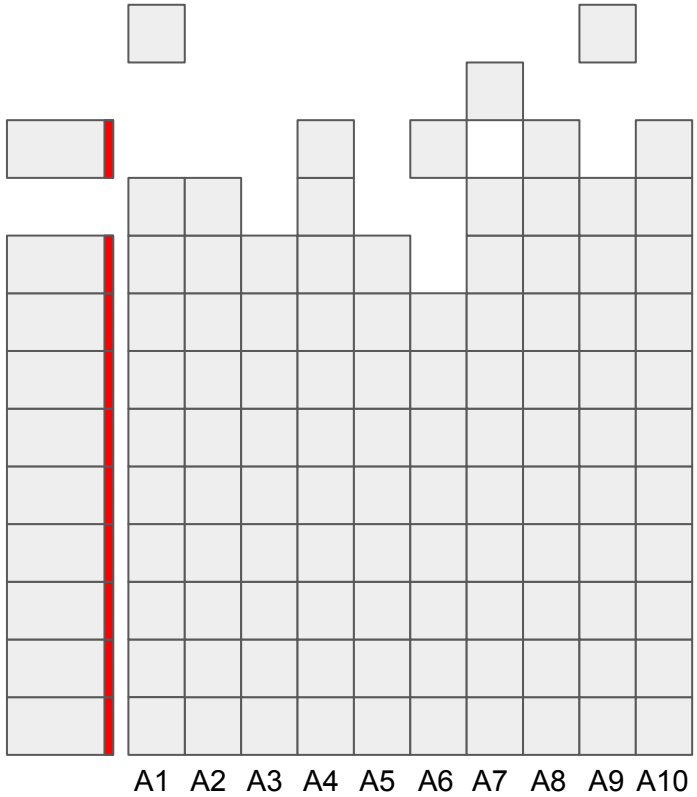
Job queue basics: factors that affect the order of jobs in queue (priority)



Job age

- FIFO (First In First Out)

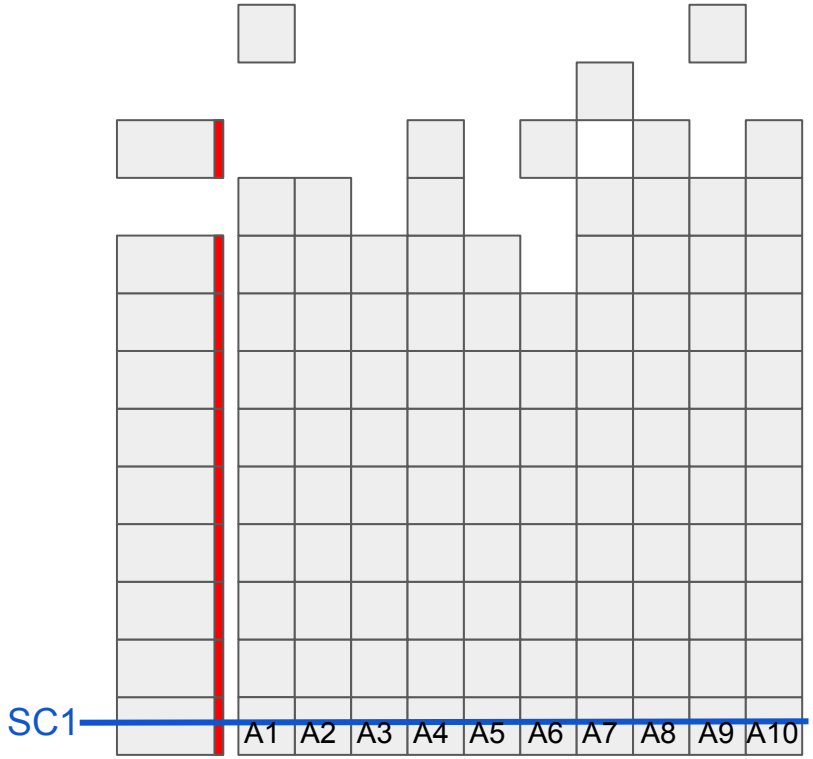
Job queue basics: factors that affect the order of jobs in queue (priority)



Fair-share tree

- Each account has a usage share target
- FairShare values range between 0 and 1
- .5 indicates that usage is on par with target
- Towards 0 indicates that usage is ahead of target
- Towards 1 indicates that usage is behind target

Job queue basics: factors that affect the order of jobs in queue (priority)

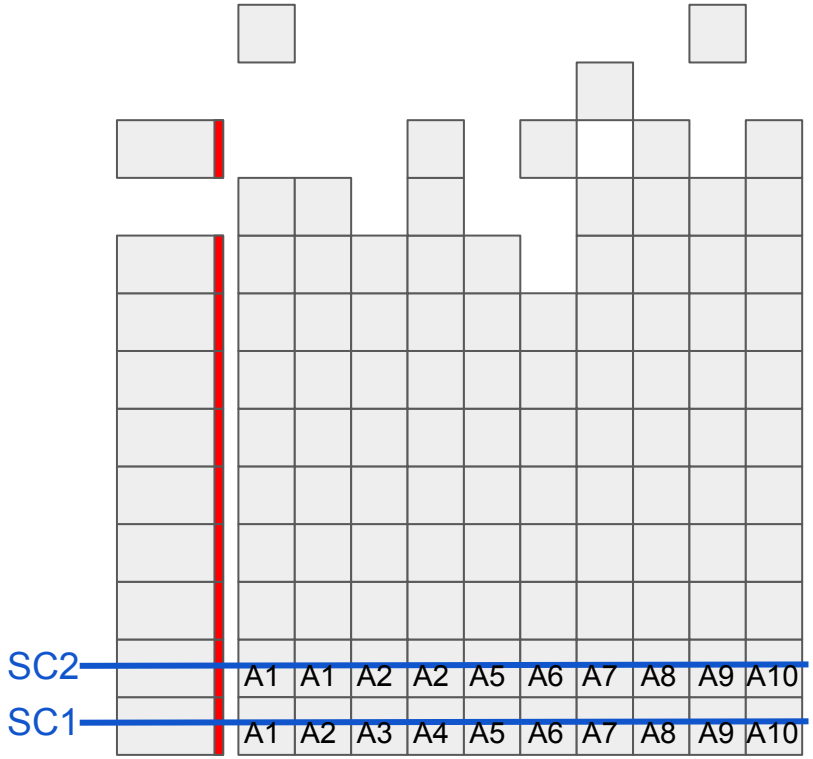


Fair-share tree

- Example: 10 accounts with equal shares of 1.

SC1
A1, .5
A2, .5
A3, .5
A4, .5
A5, .5
A6, .5
A7, .5
A8, .5
A9, .5
A10, .5
(FIFO)

Job queue basics: factors that affect the order of jobs in queue (priority)

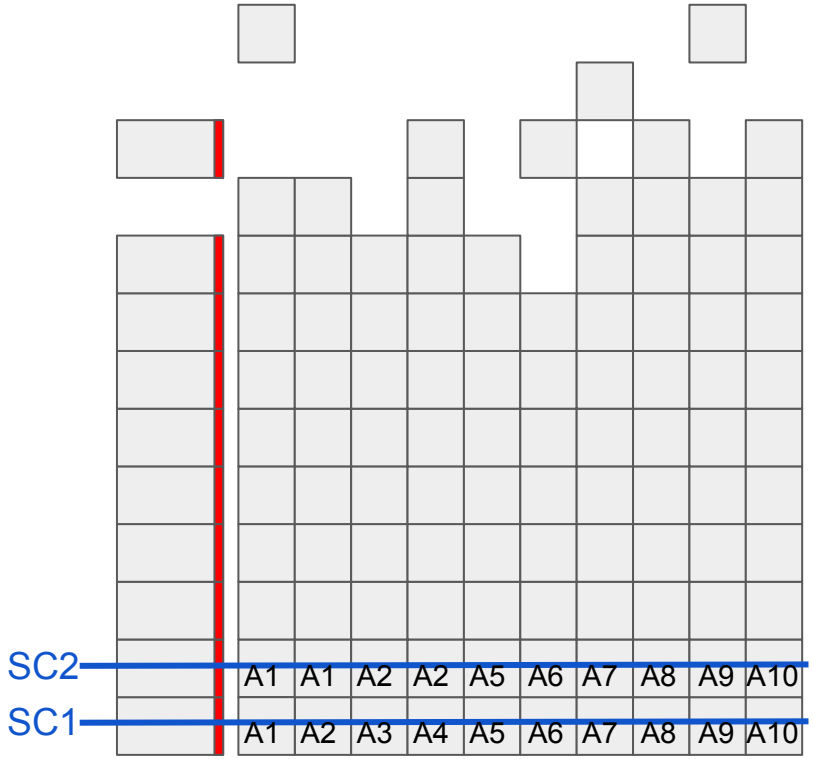


Fair-share tree

- Example: 10 accounts with equal shares of 1.

SC1	SC2
A1, .5	A1, .5
A2, .5	A1, .5
A3, .5	A2, .5
A4, .5	A2, .5
A5, .5	A5, .5
A6, .5	A6, .5
A7, .5	A7, .5
A8, .5	A8, .5
A9, .5	A9, .5
A10, .5	A10, .5
(FIFO)	(FIFO)

Job queue basics: factors that affect the order of jobs in queue (priority)

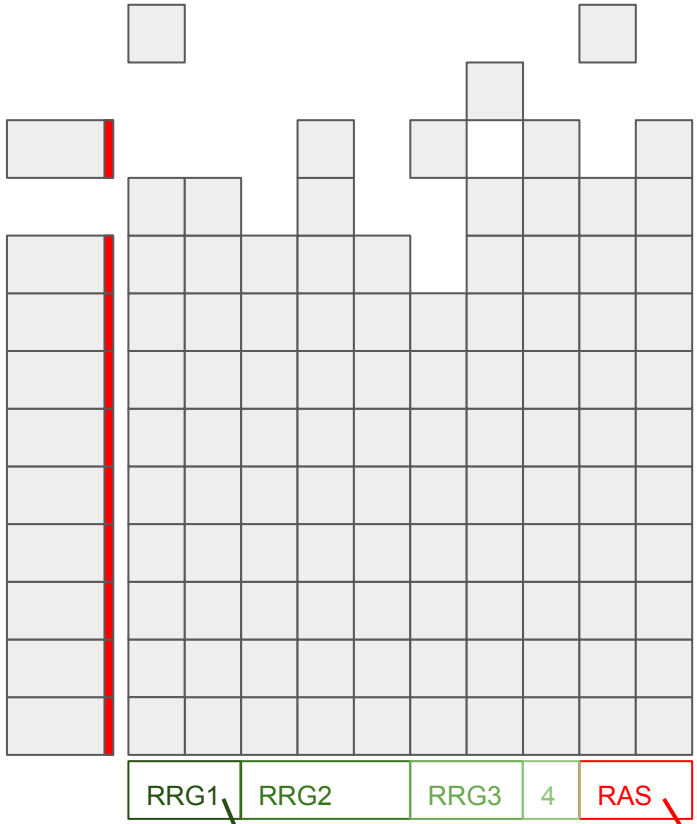


Fair-share tree

- Example: 10 accounts with equal shares of 1.

SC1	SC2	SC3
A1, .5	A1, .5	A3, .75
A2, .5	A1, .5	A4, .75
A3, .5	A2, .5	A5, .5
A4, .5	A2, .5	A6, .5
A5, .5	A5, .5	A7, .5
A6, .5	A6, .5	A8, .5
A7, .5	A7, .5	A9, .5
A8, .5	A8, .5	A10, .5
A9, .5	A9, .5	A1, .25
A10, .5	A10, .5	A2, .25
(FIFO)	(FIFO)	(FS priority)

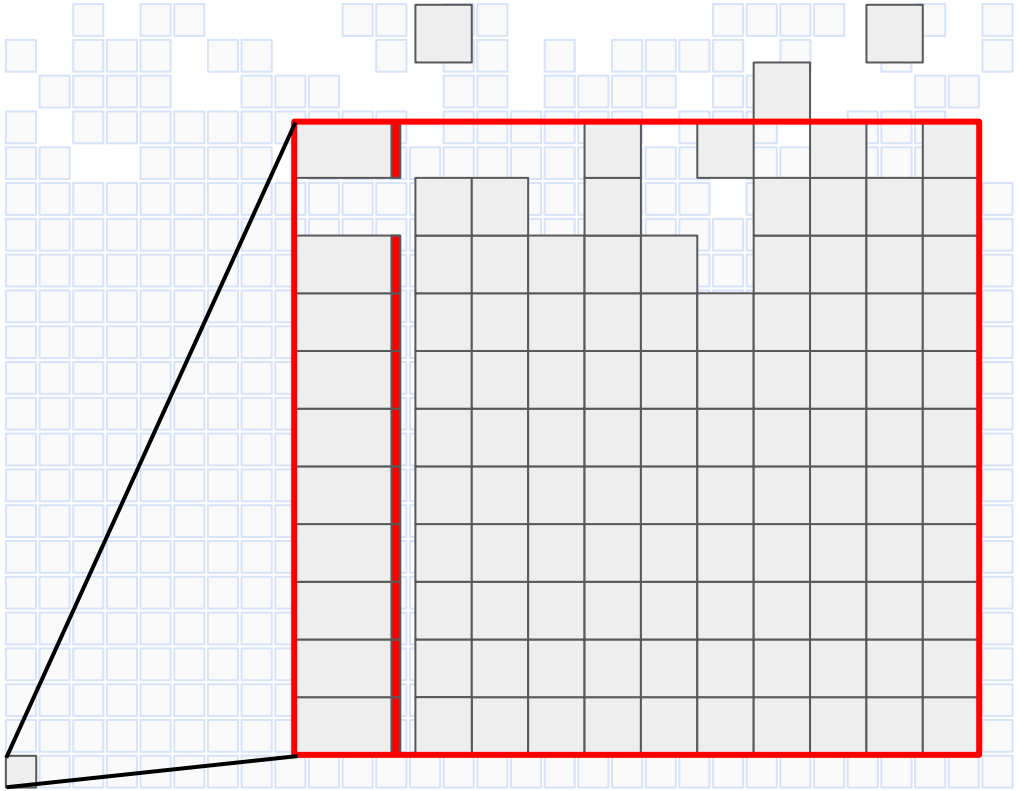
Job queue basics: factors that affect the order of jobs in queue (priority)



Fair-share tree

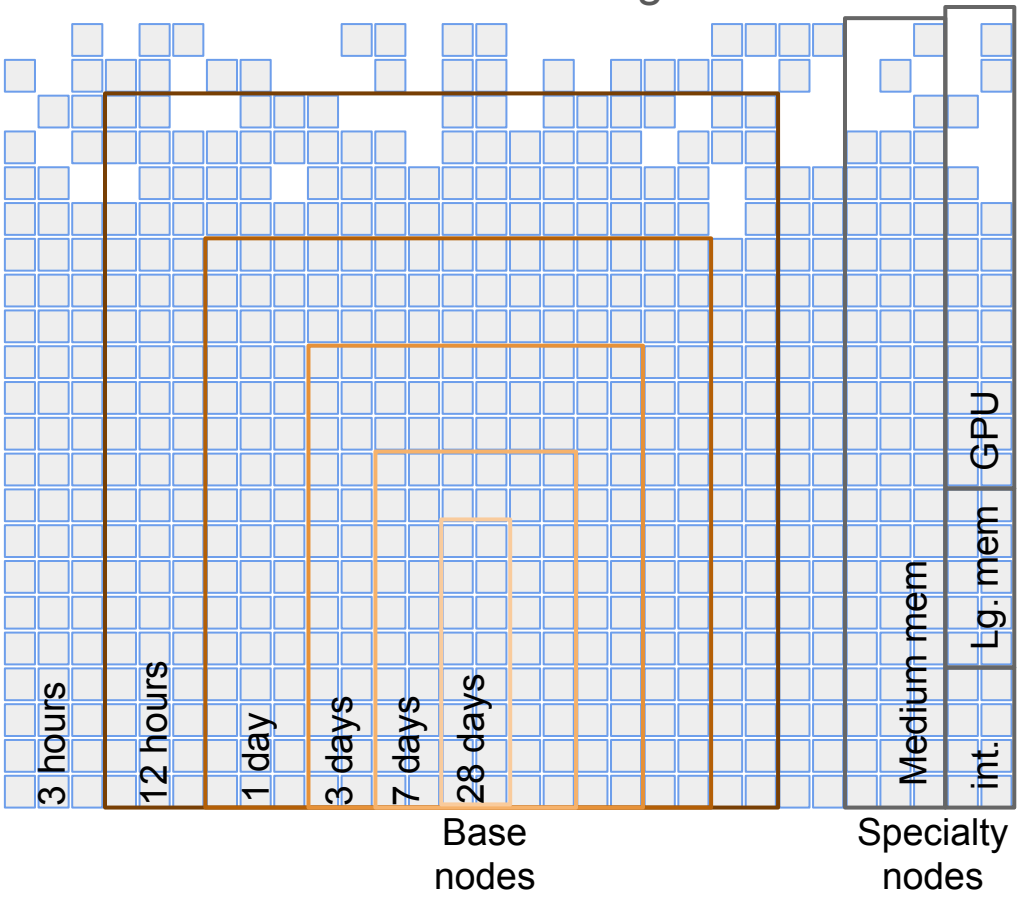
- In production shares are not equal
- Resource allocations (e.g. RRG) are defined by unique share targets.
- RAS is the equally shared residual system resources beyond allocations

Job queue basics: factors that affect the order of jobs in queue (priority)



Partitions

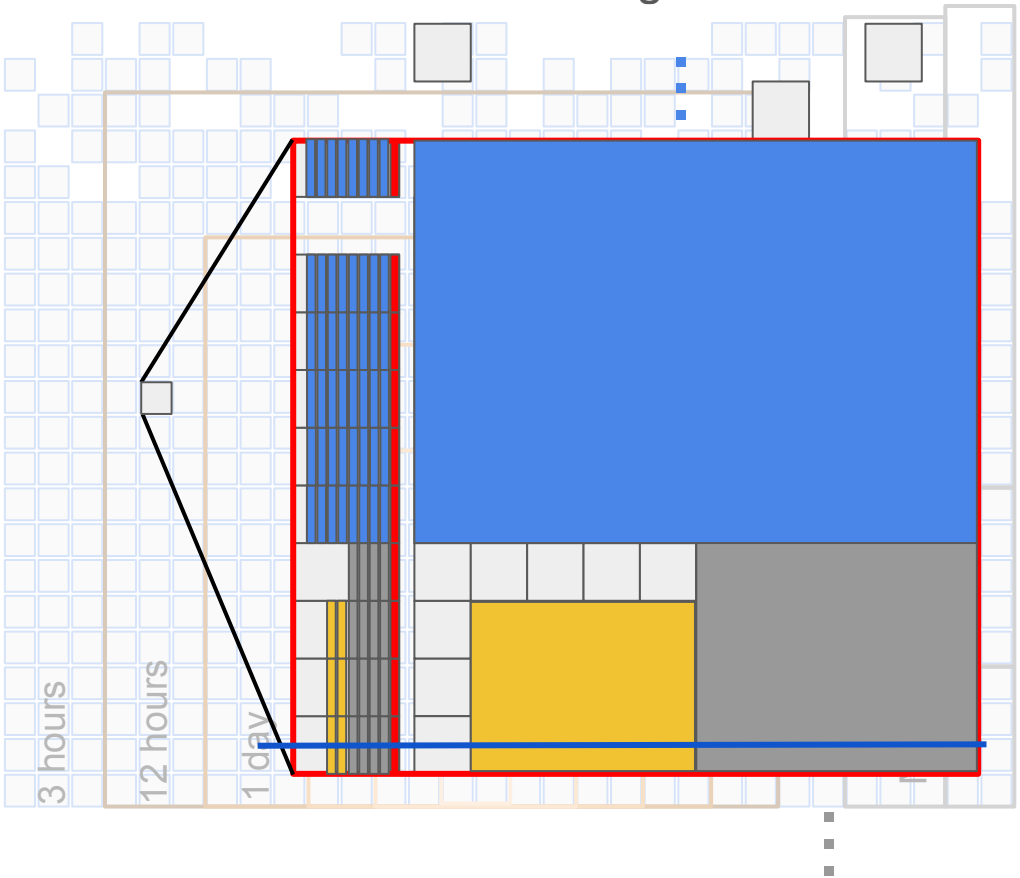
Cluster resource basics: categorization of resources that affect priority (partitions)



Partitions

- Allow for job shapes to interact with priority on subsets of nodes.

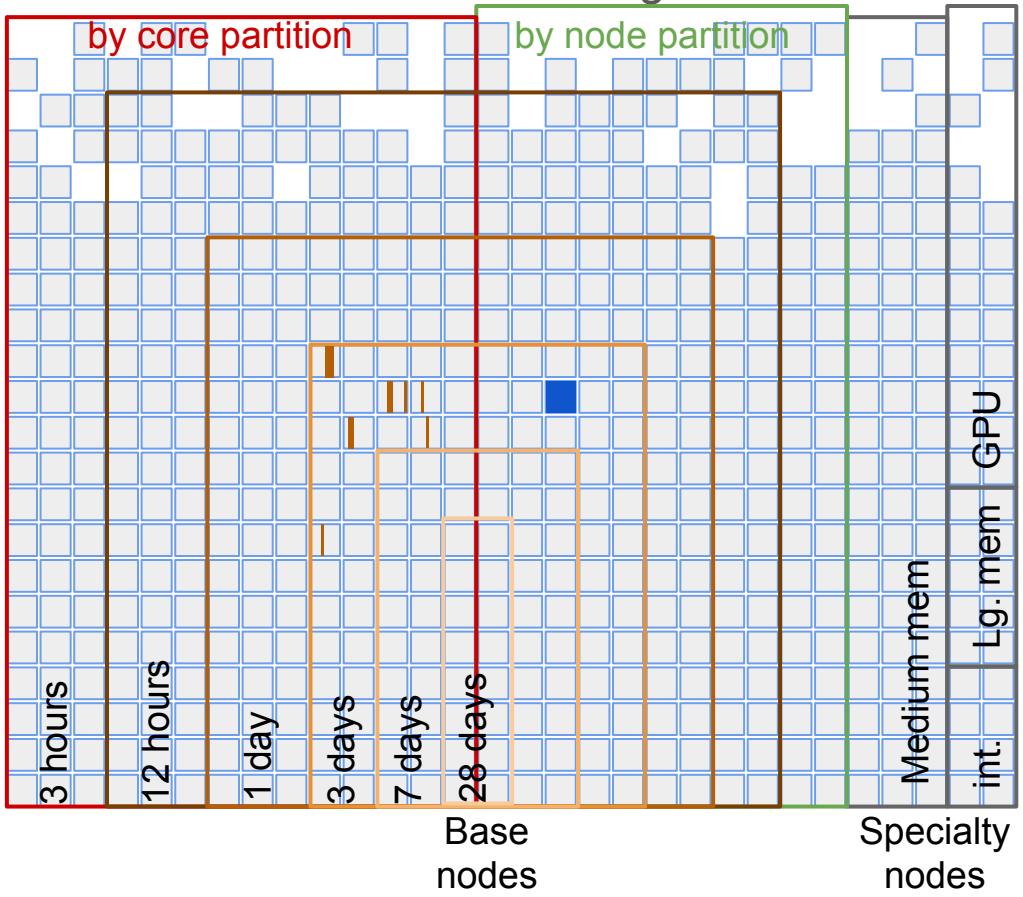
Cluster resource basics: categorization of resources that affect priority (partitions)



Backfill

- Running of lower priority jobs that can finish before any higher priority job can begin
- --time=12:00 --ntasks=1 --cpus-per-task=10 --mem=8G
- --time=12:00 --ntasks=1 --cpus-per-task=4 --mem=2G
- --time=3:00 --ntasks=1 --cpus-per-task=4 --mem=2G

Cluster resource basics: categorization of resources that affect priority (partitions)



Partitions

- By node vs by core
  - By node jobs can perform better
  - By core jobs have more opportunity to run
- --time=3-00:00 --nodes=1 --ntasks-per-node=10
- --time=3-00:00 --ntasks=10

Job properties

- squeue -u username
- squeue -j jobid
- sstat jobid
- sacct jobid

Queue properties

- squeue -u username
- squeue | less
- squeue -o ‘%V %S %b %C %D %e %L %m %M %p %r %t’ | less
- squeue -P --sort=-p,i --states=PD | less

Cluster properties

- scontrol show partition | less
- sinfo | less

Future directions

As the load on the systems balances out and continues to evolve, scheduling configuration policies (e.g. partition definitions, etc) will be adjusted to maximize utilization and performance.

As scheduling configuration properties settle to the workloads on the systems, the documentation of the scheduling policy at https://docs.computecanada.ca/wiki/Job_scheduling_policies will become more detailed.

Job profiling tools such as Remora (https://github.com/TACC/remora/wiki) can be explored in order to complement Slurm job monitoring tools for the estimation of a procedure's resource requirements

Conclusions

Jobs should be submitted with resource shapes that best match the optimal running of the procedure (profiling, scaling tests, etc).

The configuration of the cluster (partitions, etc) will be adjusted to best suit the system workloads defined by user job shapes maximizing both utilization and performance.

Do not hesitate to open support tickets regarding job profiling, job resource shape and queue properties by emailing us at:

support@computecanada.ca

Thank you for your attention!

Contact us at:

support@computecanada.ca