SHARCNET General Interest Webinar Series

All about job wait times in the Graham queue

James Desjardins
High Performance Computing Consultant
SHARCNET, Brock University

April 25th, 2018

Common questions

Why does a job take a long time to start?

Is there anything that can be done to make a job start more quickly?

Why does a job start time estimate keep moving into the future?

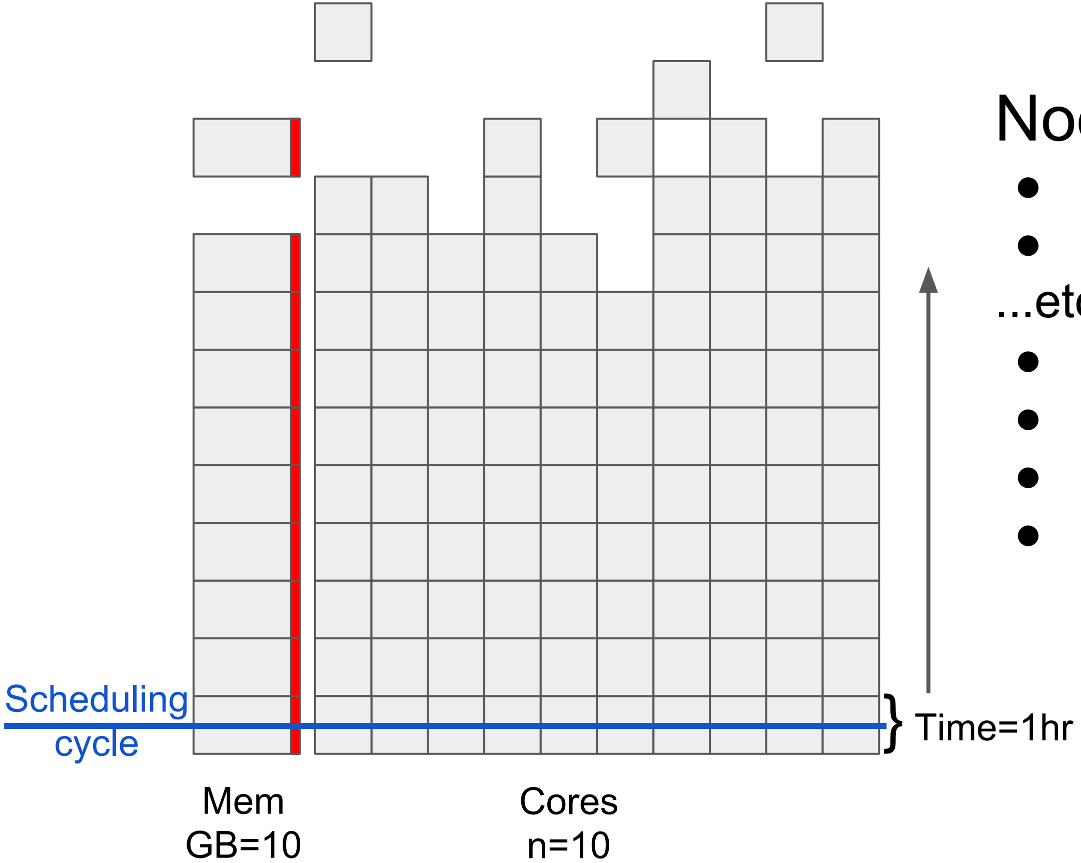
General factors to consider

Scheduling basics: node resources and resource requests (jobs)

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

Cluster resource basics: segmentation of nodes in the cluster (partitions)

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



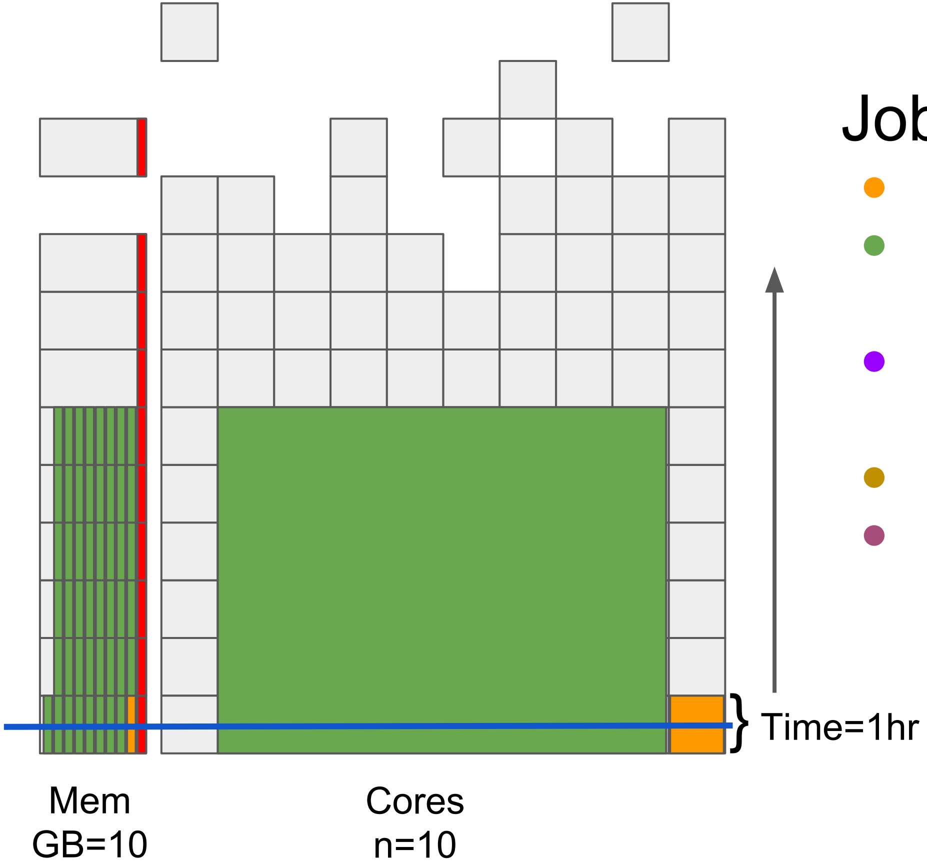
Node resources

- Cores
- Memory

...etc

- GPUs
- Software licenses
- …
- Time

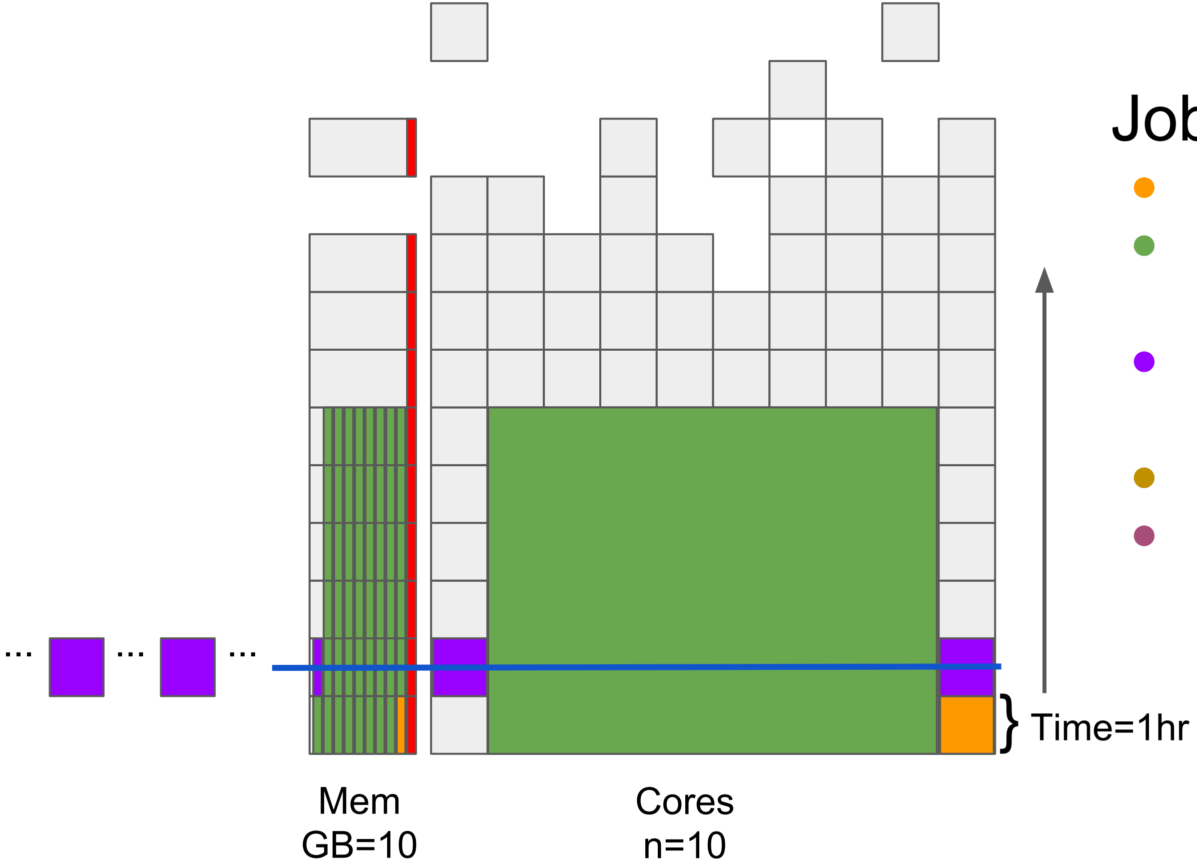
Scheduling basics: node resources and resource requests (job queue)



Job size

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

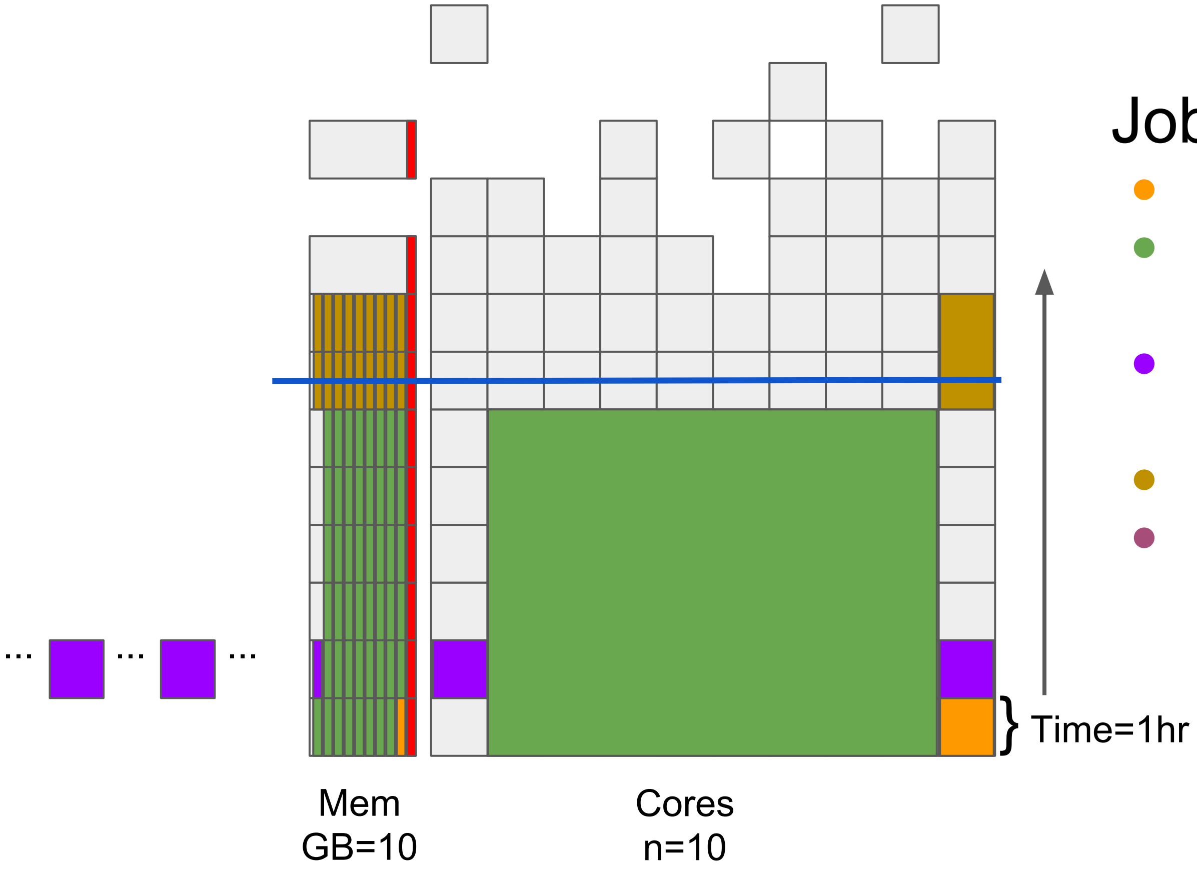
Scheduling basics: node resources and resource requests (job queue)



Job size

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

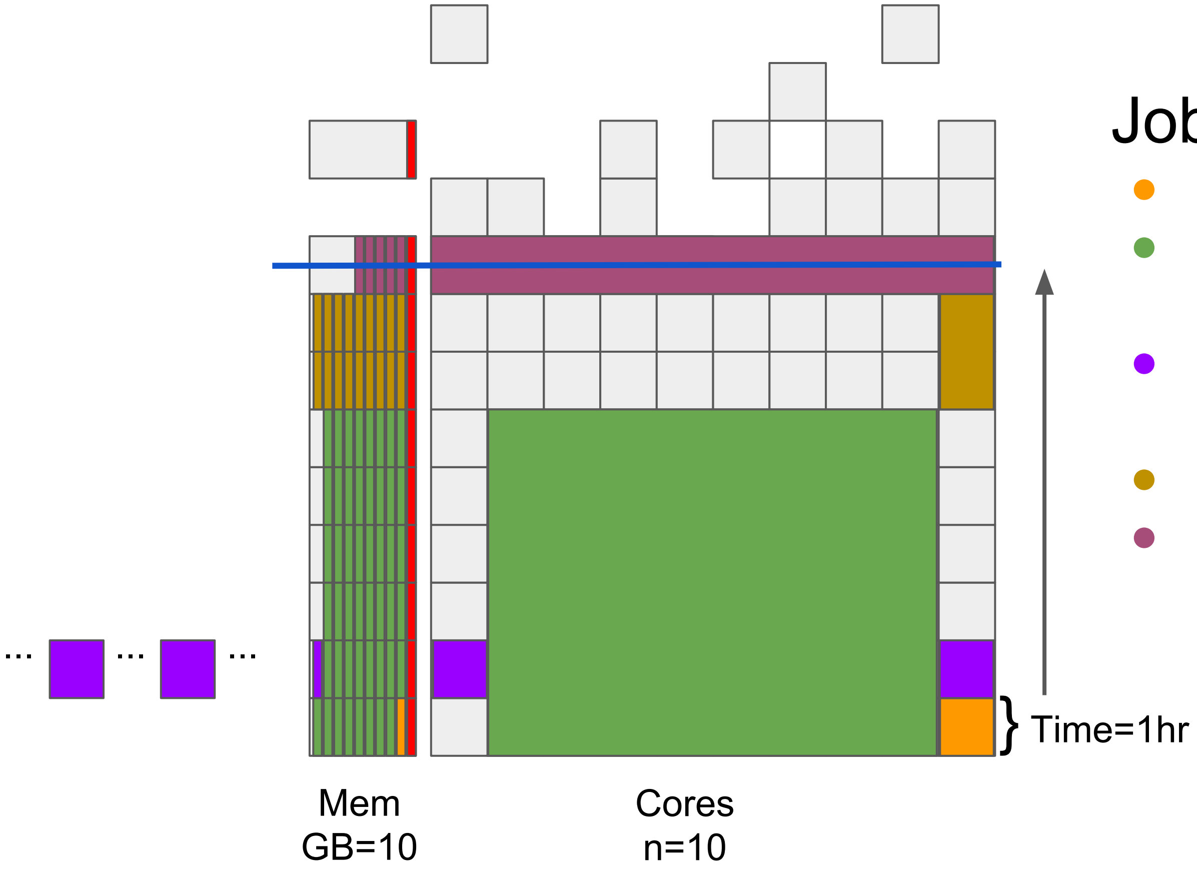
Scheduling basics: node resources and resource requests (job queue)



Job size

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

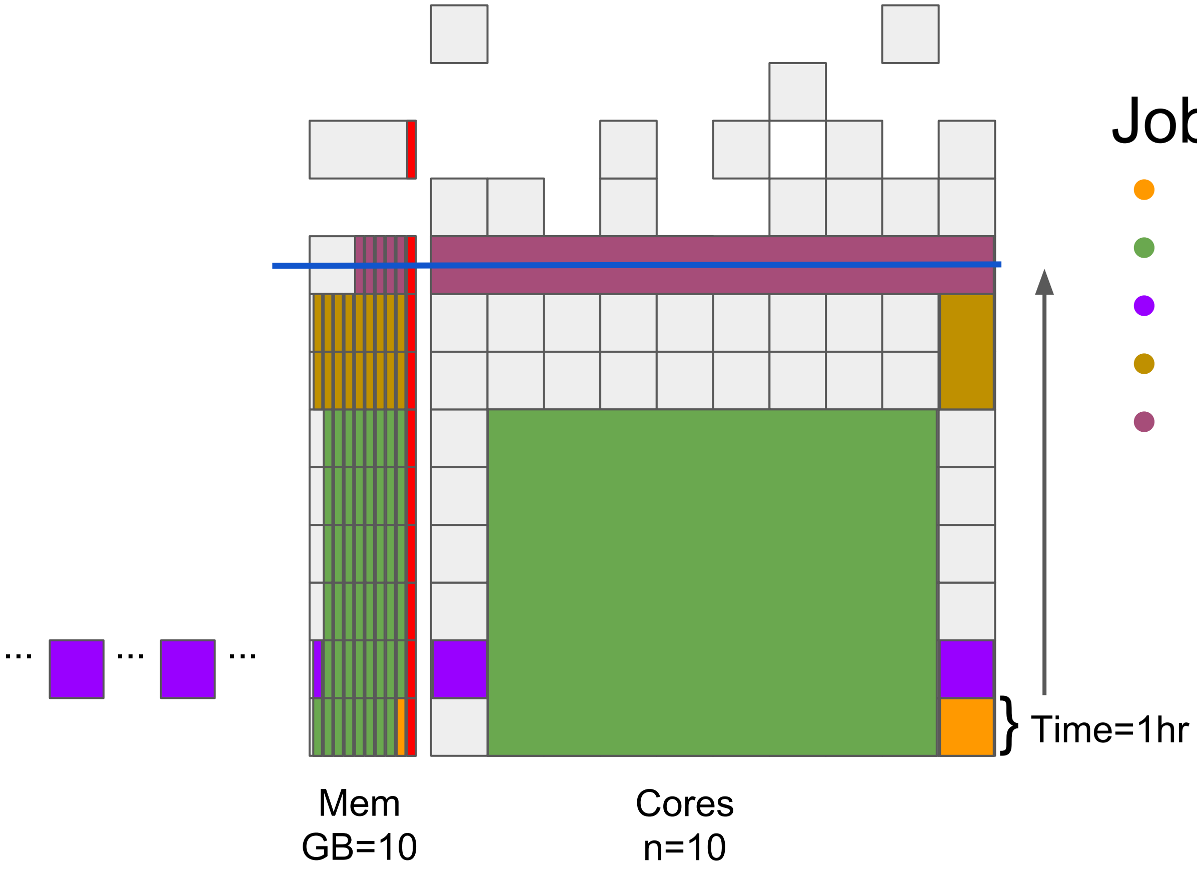
Scheduling basics: node resources and resource requests (job queue)



Job size

- ● --time=1:00 --mem=1G
- ● --time=6:00 --mem=8G --cpu-per-task=8
- ● --time=1:00 --ntasks=10 --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
- ● 10 core
- ● 9 core
- ● 10 core

Graham heterogeneous node shape and billing

32 core per node (ntasks-per-node=32 for bynode partition MPI)

128G, 256G, 512G

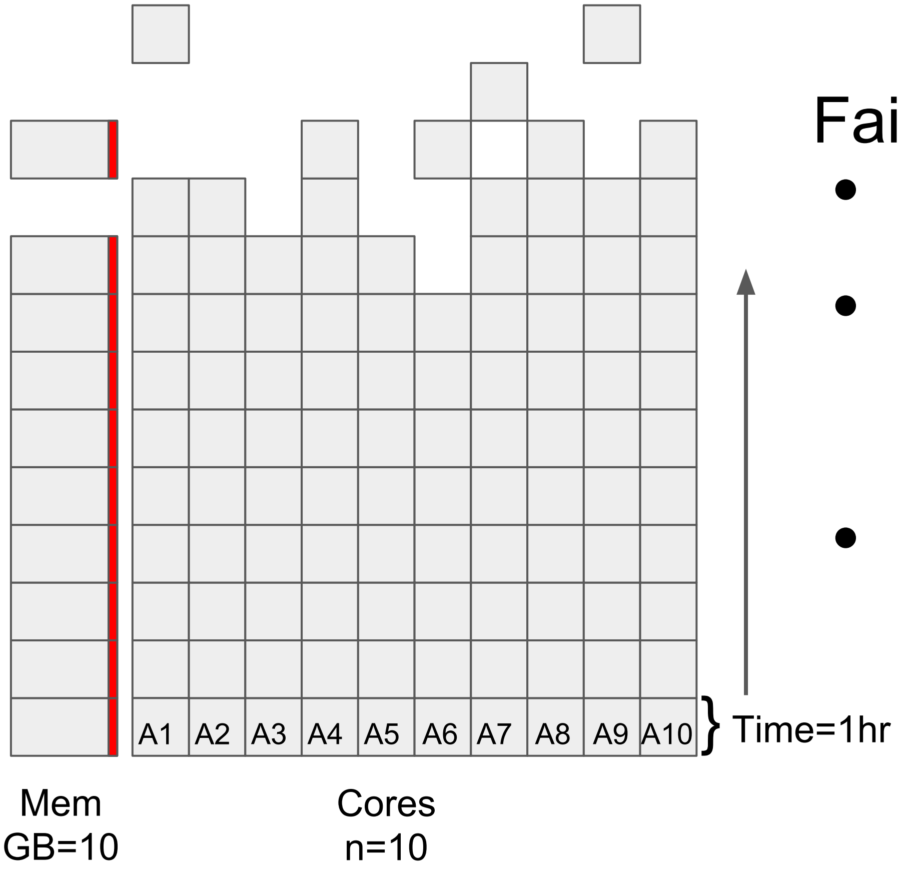
64 cores per node 3T memory nodes (3)

32 cores per node, 128G memory, 2 NVIDIA P100 Pascal GPUs

Core year equivalent billing is 4G memory = 1 core

Large parallel system Niagara is now online

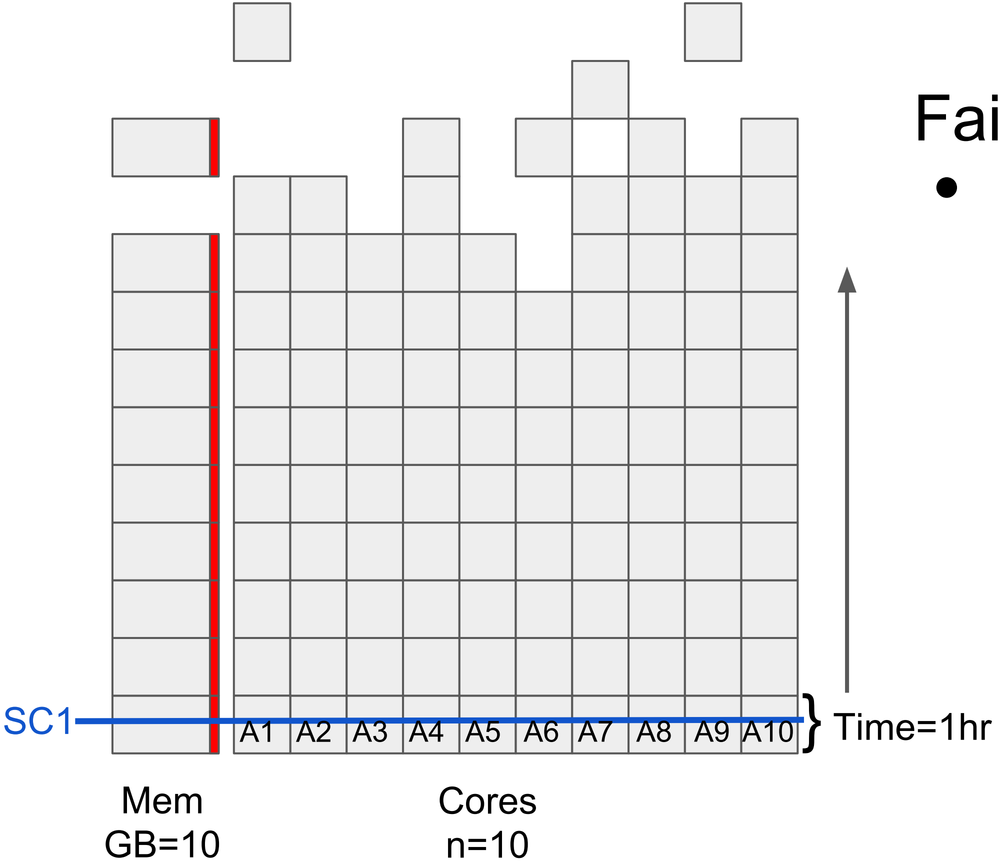
Job queue basics: factors that affect the order of jobs in queue (priority)



Fair-share

- Each account has a usage share target
- When account usage (resources reserved) is above the target, priority goes down.
- When account usage (resources reserved) is below the target, priority goes up.

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

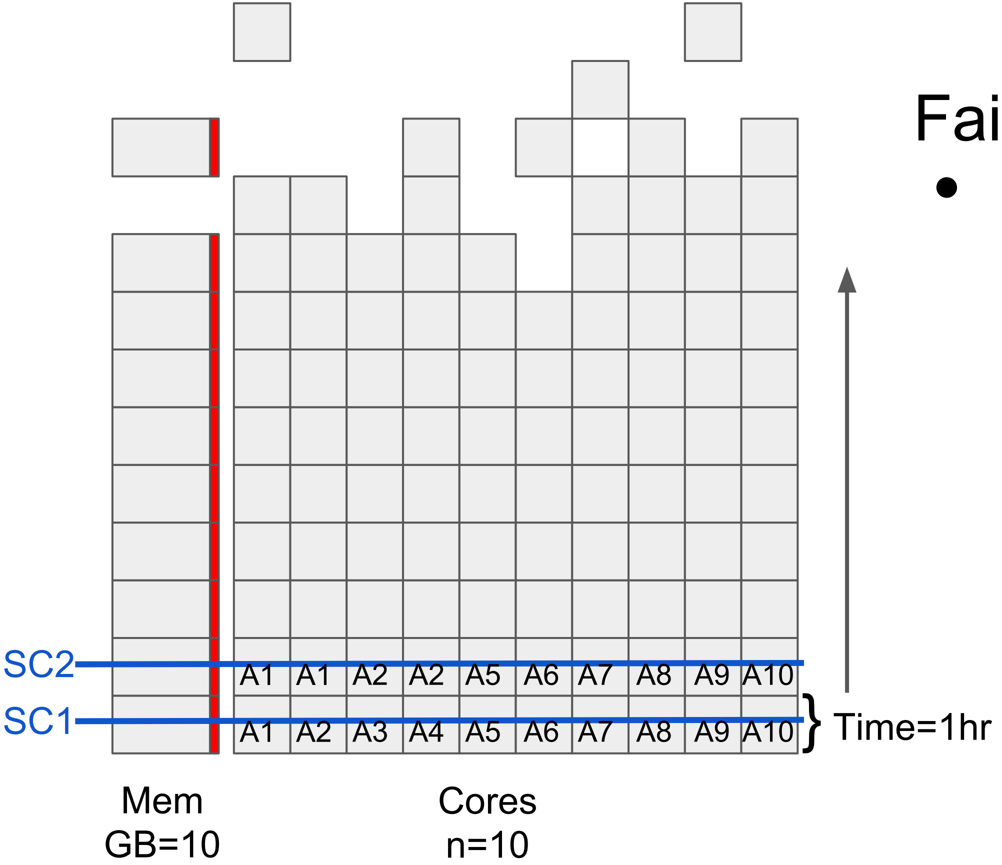


Fair-share queue sorting

- 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)

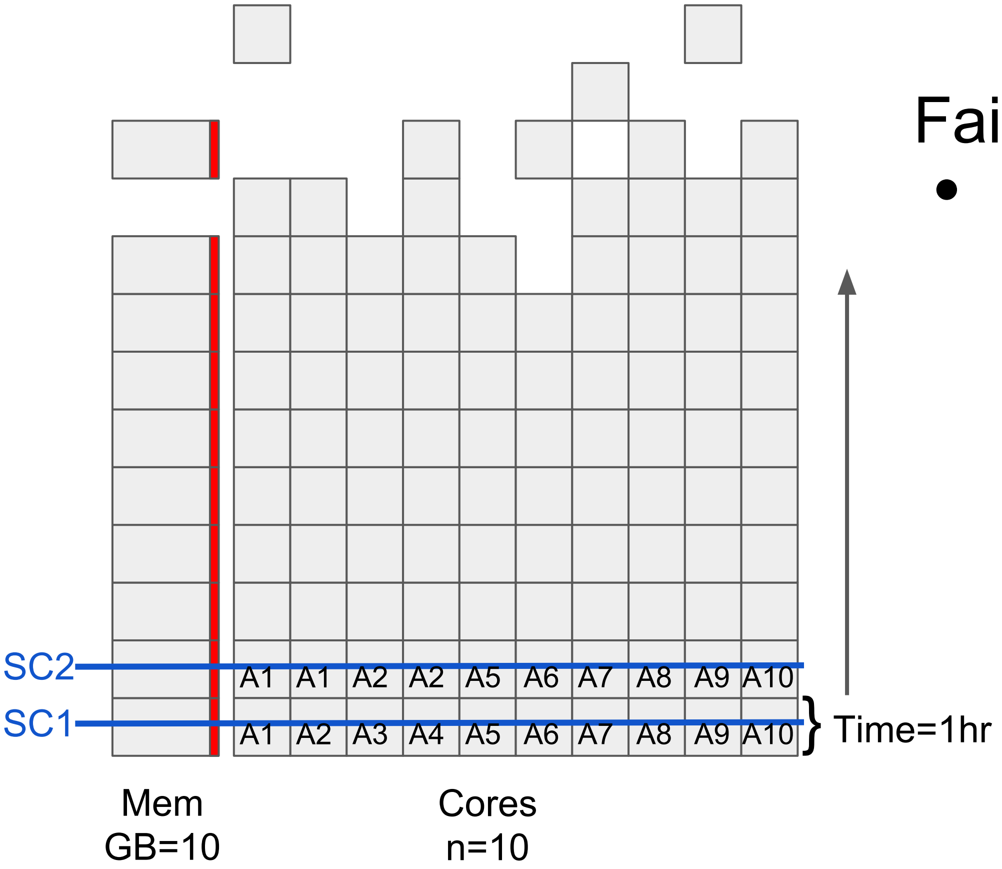


Fair-share queue sorting

- 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)

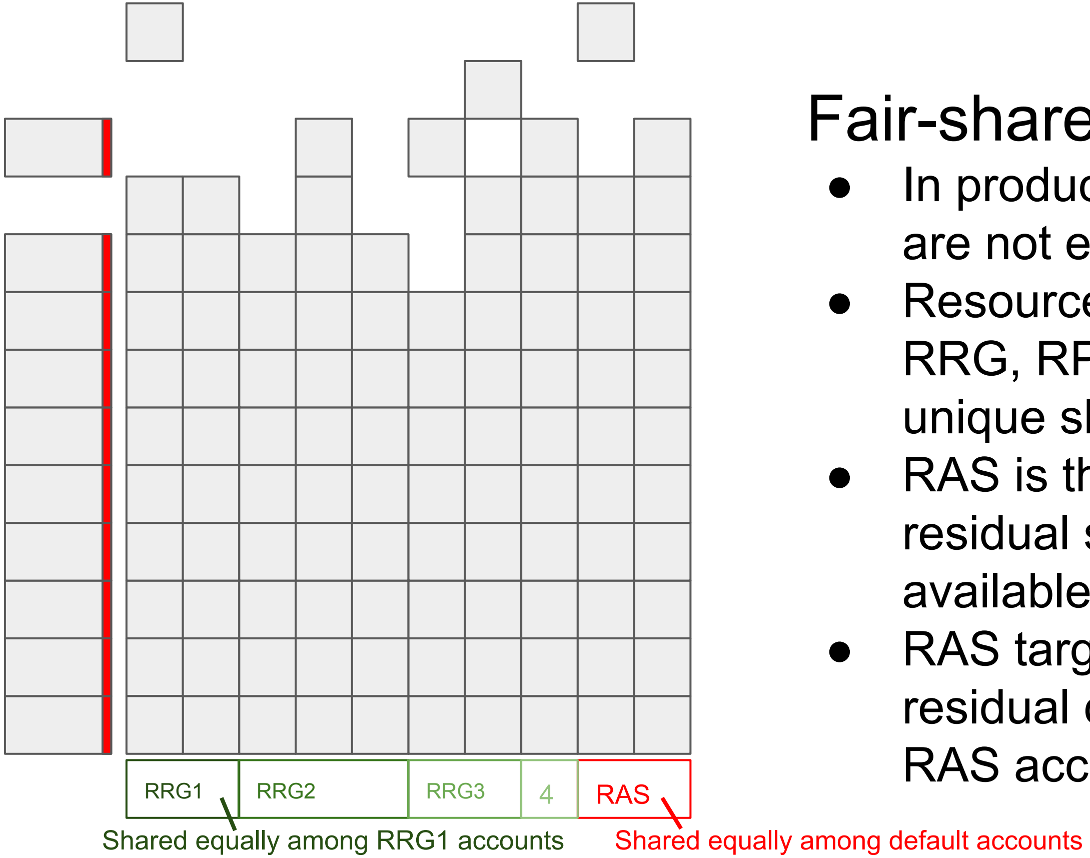

SC2
SC1
A1 A1 A2 A2 A5 A6 A7 A8 A9 A10
A1 A2 A3 A4 A5 A6 A7 A8 A9 A10
Time=1hr
Mem
GB=10
Cores
n=10

Fair-share queue sorting

- 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)

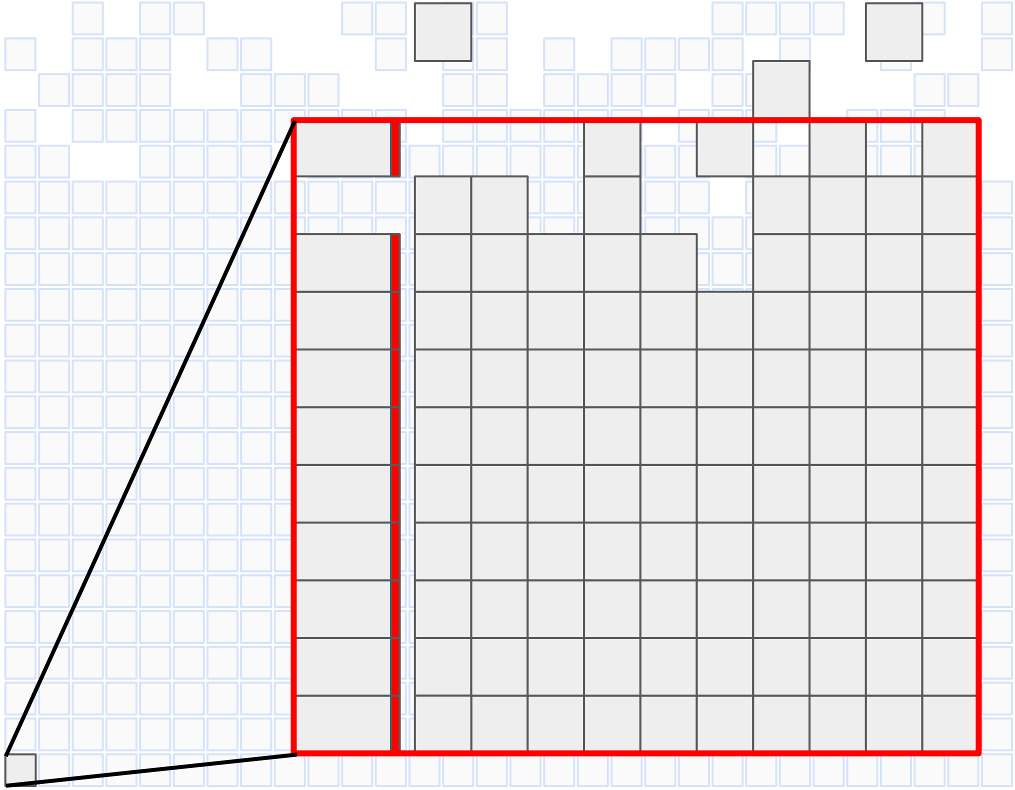
Job queue basics: factors that affect the order of jobs in queue (priority)


RRG1
RRG2
RRG3
4
RAS
Shared equally among RRG1 accounts
Shared equally among default accounts

Fair-share targets

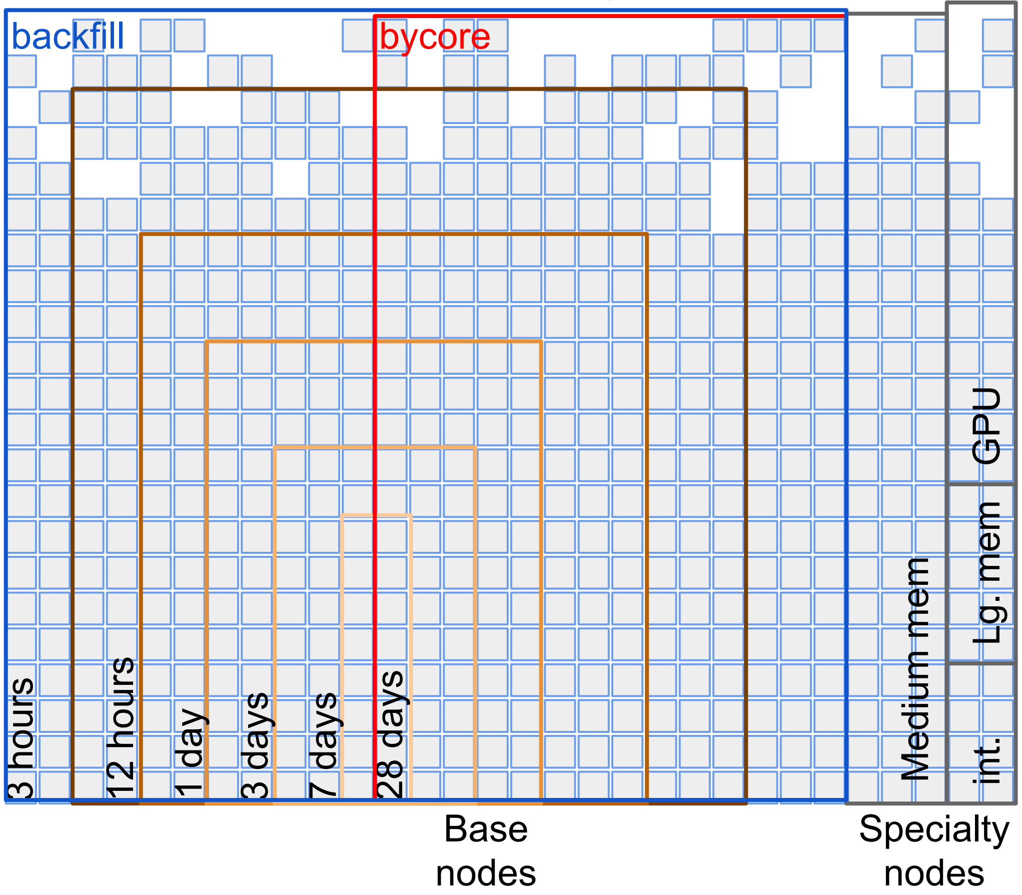
- In production target shares are not equal
- Resource allocations (e.g. RRG, RPP) are defined by unique share targets.
- RAS is the equally shared residual system resources available beyond allocations
- RAS target is the number of residual cores / number of RAS accounts.

Cluster resource basics: segmentation of nodes in the cluster (partitions)



Partitions

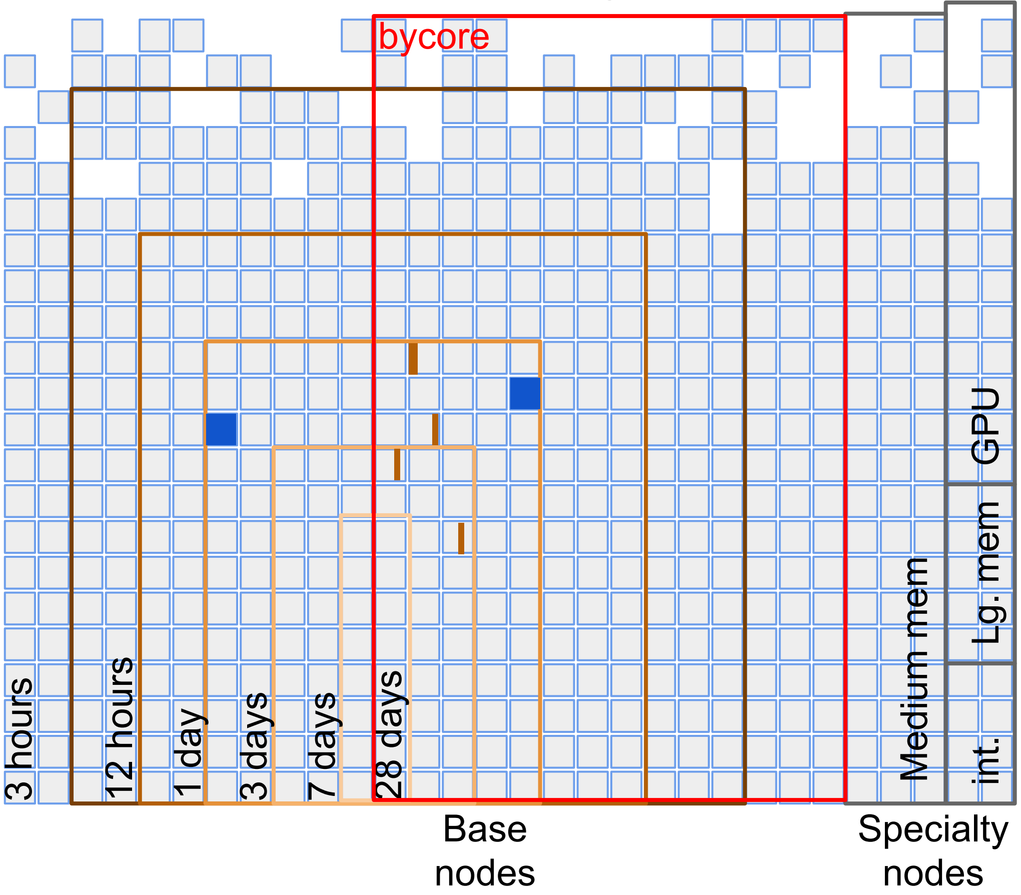
Cluster resource basics: segmentation of nodes in the cluster (partitions)



Partitions

- Restrict jobs of specific shapes to node sets
- Full node jobs can run on most any node (bynode)
- Jobs 3 hours and shorter can run on most any node
- Longer run time jobs have access to fewer nodes
- Partial node jobs (bycore) have access to fewer nodes
- Backfill jobs can run on most any node

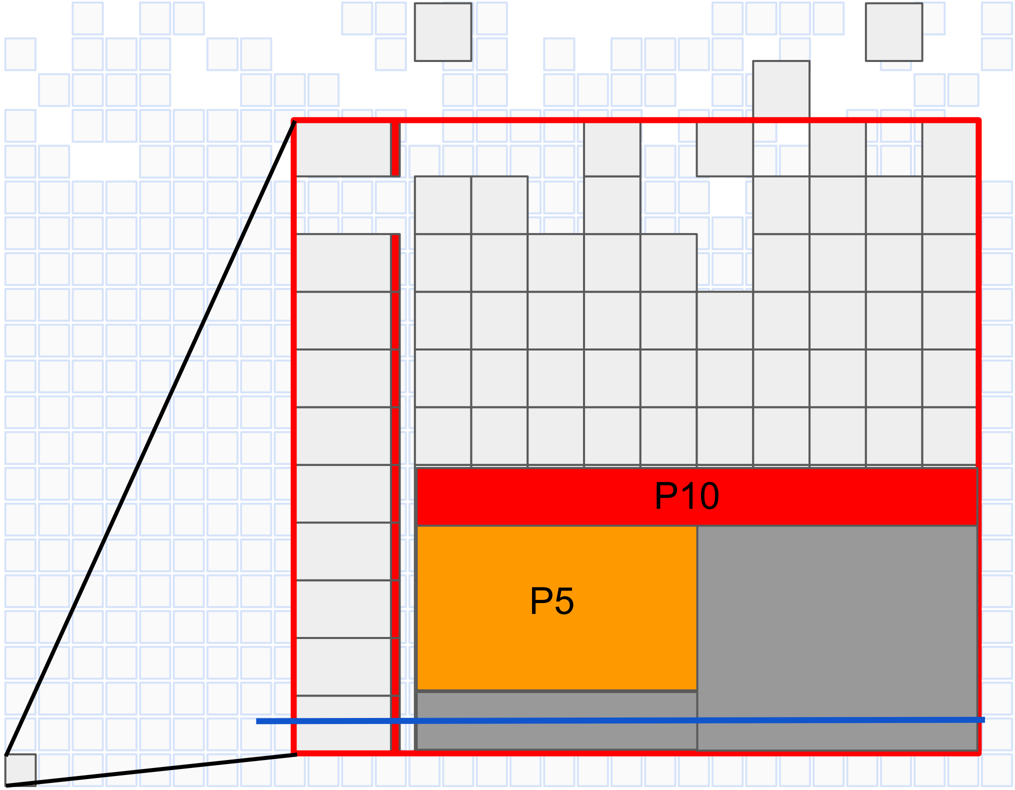
Cluster resource basics: segmentation of nodes in the cluster (partitions)



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=32
- --time=3-00:00 --ntasks=32

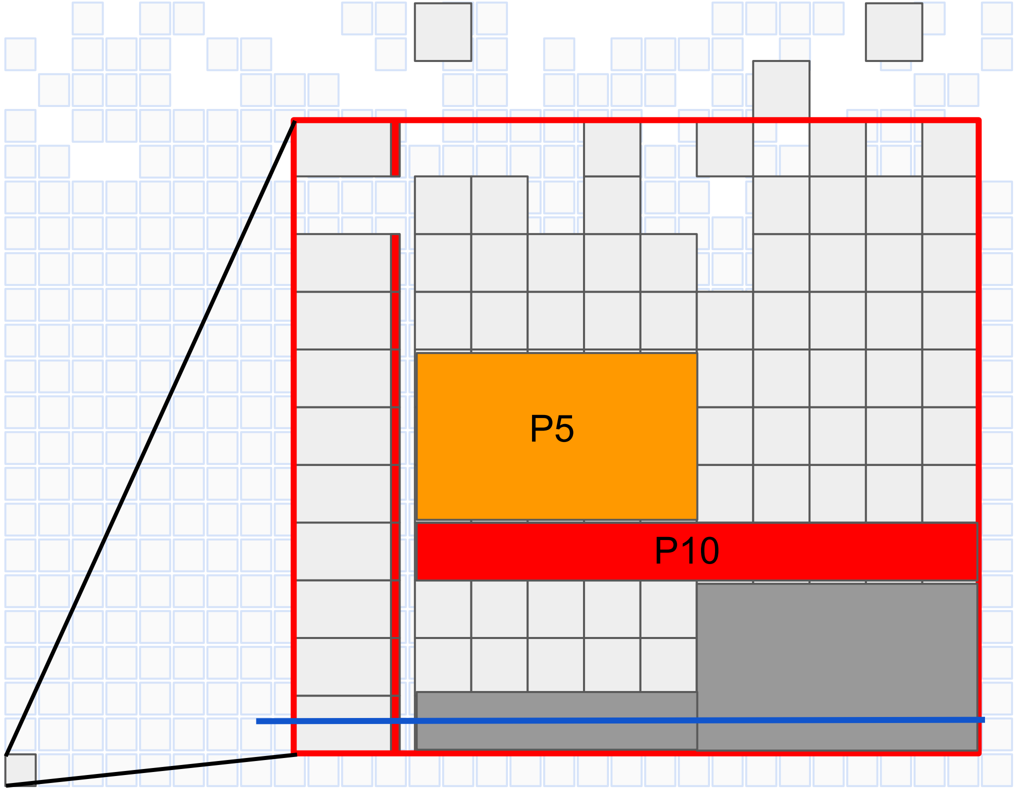
Cluster resource basics: segmentation of nodes in the cluster (partitions)



Backfill

- Jobs can start before higher priority jobs if they complete before the higher priority job can begin

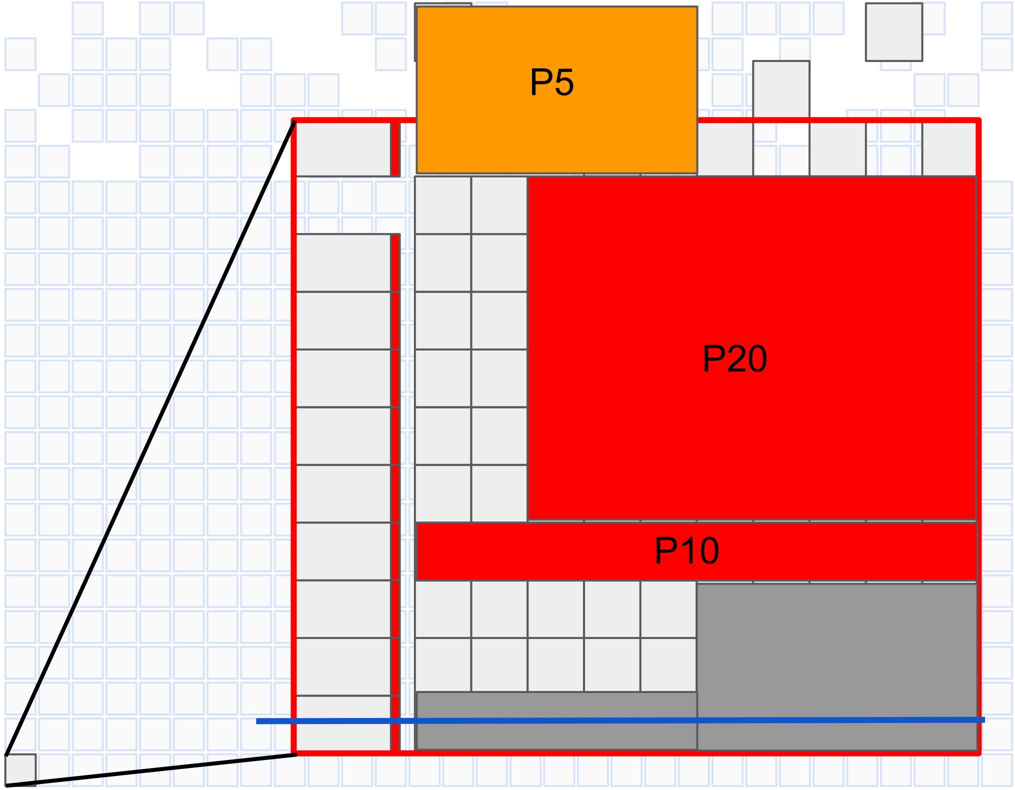
Cluster resource basics: segmentation of nodes in the cluster (partitions)



Start time estimates

- Scheduler start time estimates are constantly changing
- Due to unknown job termination times

Cluster resource basics: segmentation of nodes in the cluster (partitions)



Start time estimates

- Scheduler start time estimates are constantly changing
- Due to unknown job termination times
- And introductions of new high priority jobs in the queue.

Monitoring jobs, the queue and the cluster

Show properties of all jobs on the system since a stated date:

```
sacct -aX -S 2018-04-25 -o account%4,partition%32,submit,start,end,elapsed,timelimit,reqmem,ncpus,nnodes,state
```

Show all of the jobs in the queue sorted by their current priority:

```
squeue -P --sort=-p,i --states=PD -o "%.4a %P %.8C %m %V %e %l %r %t %S" | less
```

Show partition properties:

```
partition-stats
```

```
sinfo
```

```
scontrol show partition
```

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

Gaming the scheduler at SHARCNET demonstration

- https://www.youtube.com/watch?v=IVxdVb5Gw4E

support@computecanada.ca

What can be done about wait times? (conclusions)

Job resource footprint (shape of the job on the cluster)

Decrease job footprint: minimize accurate requests, checkpointing, dependent queuing

Consider the compressed vs distributed footprint of MPI jobs.

Load on the system (relative to resources available)

Users have no control over the load on the system (by others) but there are methods to view the state

The contribution model gives users the ability to influence the resource pool

Account target share (fair-share priority)

Be efficient about usage (both in terms of job numbers and footprint)

Apply for a resource target allocation

Conclusions (other)

The scheduling policy is prioritizing account target consumption and system utilization.

Job submission should prioritize the optimal running of the procedure (profiling, scaling tests, etc) and feasibility within the scheduling policy.

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

Do not hesitate to open support tickets regarding job shape and queue properties by email us at:

support@computecanada.ca

Thank you!

Monitoring jobs, the queue and the cluster

```
[jdesjard@gra-login4 ~]$ sacct -aX -S 2018-04-20 -o account%4,partition%32,submit,start,end,timelimit,reqmem,ncpus,nnodes,state

…
rrg+                cpubase_bycore_b2 2018-04-24T13:11:21 2018-04-24T21:56:56             Unknown   12:00:00            256Mc               1         1         RUNNING
rrg+                cpubase_bycore_b2 2018-04-24T13:11:21 2018-04-24T21:56:56             Unknown   12:00:00            256Mc               1         1         RUNNING
rpp+                cpubase_bycore_b2 2018-04-24T21:57:02 2018-04-24T21:57:09 2018-04-24T21:59:52   06:00:00            4Gn       1         1         FAILED
def+      cpubase_bycore_b2,cpubackfill 2018-04-24T21:57:03           Unknown             Unknown   05:00:00            4Gn       1         1         PENDING
def+                cpubase_bycore_b6 2018-04-24T21:57:09             Unknown             Unknown 10-00:00:+            32Gn      16        1         PENDING
def+      cpubase_bycore_b1,cpubackfill 2018-04-24T21:57:09           Unknown             Unknown   03:00:00            4Gn       1         1  COMPLETED
def+                cpubase_bycore_b1 2018-04-24T19:56:06 2018-04-24T21:57:09 2018-04-24T21:59:42   03:00:00            4Gn       1         1  COMPLETED
def+                cpubase_bycore_b1 2018-04-24T19:56:06 2018-04-24T21:57:09 2018-04-24T21:59:42   03:00:00            4Gn       1         1  COMPLETED
def+                cpubase_bycore_b1 2018-04-24T19:56:06 2018-04-24T21:57:09 2018-04-24T21:59:46   03:00:00            4Gn       1         1  COMPLETED
def+                cpubase_bycore_b1 2018-04-24T19:56:06 2018-04-24T21:57:09 2018-04-24T21:59:46   03:00:00            4Gn       1         1  COMPLETED
def+                cpubase_bycore_b1 2018-04-24T19:56:06 2018-04-24T21:57:09 2018-04-24T21:59:50   03:00:00            4Gn       1         1  COMPLETED
rpp+                cpubase_bycore_b2 2018-04-24T21:57:11 2018-04-24T21:57:11             Unknown   06:00:00            4Gn       1         1         RUNNING
rpp+                cpubase_bycore_b2 2018-04-24T21:57:15 2018-04-24T21:57:22             Unknown   06:00:00            4Gn       1         1         RUNNING
def+      cpubase_bycore_b1,cpubackfill 2018-04-24T21:57:18           Unknown             Unknown   00:05:00            256Mc               1         1         PENDING
rpp+                cpubase_bycore_b2 2018-04-24T21:57:20 2018-04-24T21:57:22             Unknown   06:00:00            4Gn       1         1         RUNNING
...
```

Monitoring jobs, the queue and the cluster

```
squeue -P --sort=-p,i --states=PD -o "%.4a %P %.8C %m %V %e %l %r %t %S" | less

ACCO PARTITION         CPUS MIN_MEMORY SUBMIT_TIME END_TIME TIME_LIMIT REASON ST START_TIME
...
def- cpubackfill       256 125G 2018-03-16T15:58:38 N/A 2:30:00 Resources PD N/A
def- cpularge_bynode_b1       256 1T 2018-02-07T17:23:29 N/A 2:30:00 Resources PD N/A
def- cpubackfill       256 1T 2018-02-07T17:23:29 N/A 2:30:00 Resources PD N/A
def- cpubase_bycore_b1        3600 2G 2018-03-16T15:13:26 N/A 10:00 Resources PD N/A
def- cpubackfill       3600 2G 2018-03-16T15:13:26 N/A 10:00 Resources PD N/A
def- cpubase_bycore_b1        1728 2G 2018-03-16T16:16:45 N/A 5:00 Resources PD N/A
def- cpubackfill       1728 2G 2018-03-16T16:16:45 N/A 5:00 Resources PD N/A
def- cpubase_bynode_b2        256 256M 2018-01-19T07:33:47 N/A 3:30:00 Resources PD N/A
def- cpubackfill       256 256M 2018-01-19T07:33:47 N/A 3:30:00 Resources PD N/A
def- cpubase_bycore_b2        3840 30G 2018-04-13T11:15:31 N/A 12:00:00 Resources PD N/A
def- cpubackfill       3840 30G 2018-04-13T11:15:31 N/A 12:00:00 Resources PD N/A
def- cpubase_bycore_b2        3840 30G 2018-04-13T11:26:57 N/A 12:00:00 Resources PD N/A
def- cpubackfill       3840 30G 2018-04-13T11:26:57 N/A 12:00:00 Resources PD N/A
def- cpubase_bynode_b1        32 125G 2018-02-09T18:05:06 N/A 2:20:00 Resources PD N/A
def- cpubackfill       32 125G 2018-02-09T18:05:06 N/A 2:20:00 Resources PD N/A
rpp- cpubase_bycore_b6        2 100G 2018-04-23T18:02:27 2018-05-04T20:37:01 7-12:00:00 Resources PD 2018-04-27T08:37:01
rrg- cpubase_bycore_b5        60 8000M 2018-04-23T23:10:30 2018-05-02T19:03:14 7-00:00:00 Resources PD 2018-04-25T19:03:14
rrg- cpubase_bycore_b5        60 8000M 2018-04-23T23:11:12 2018-05-05T00:13:54 7-00:00:00 Priority PD 2018-04-28T00:13:54
...
rrg- cpubase_bycore_b5        60 8000M 2018-04-24T14:07:54 2018-05-05T00:13:54 7-00:00:00 Priority PD 2018-04-28T00:13:54
def- cpubase_bycore_b1        4 2024M 2018-04-18T18:09:47 N/A 3:00:00 Dependency PD N/A
def- cpubackfill       4 2024M 2018-04-18T18:09:47 N/A 3:00:00 Dependency PD N/A
def- cpubase_bycore_b1        4 2024M 2018-04-20T15:53:57 N/A 3:00:00 Dependency PD N/A
…
```

Monitoring jobs, the queue and the cluster

```
[jdesjard@gra-login4 ~]$ sinfo
PARTITION              AVAIL  TIMELIMIT  NODES  STATE NODELIST
cpubase_interac        up      3:00:00        1       mix gra800
cpubase_interac        up      3:00:00        1  alloc gra796
cpubase_interac        up      3:00:00        3   idle gra[797-799]
cpubase_bynode_b1      up      3:00:00        15 drain* gra[222,732,988-997,1020,1030,1040]
cpubase_bynode_b1      up      3:00:00        16   drng gra[13,33,37,39,46,60,67-68,71,79,87,115,120,130,135,343]
cpubase_bynode_b1      up      3:00:00 144    mix
gra[44,47,91,100-101,116,118,124,138-139,225,236,263,284-286,291,293,295,299-300,309,314,321-323,325-331,333-340,342,344-352,354-
355,357,360-368,370,372-375,377-379,381,384,387-389,391,393-396,401,428,433,447,506,509,542,547,550,568,584-585,608,616,622,625-6
26,634-635,640,643-644,647,650-651,668-669,701-702,720,724,727,738-739,741-745,998-1002,1005-1011,1013-1014,1016,1018,1026,1031-1
036,1042]
cpubase_bynode_b1      up      3:00:00 687  alloc
gra[1-12,14-32,34-36,38,40-43,45,48-59,61-66,69-70,72-78,80-86,88-90,92-99,102-114,117,119,121-123,125-129,131-134,136-137,140-22
1,223-224,226-235,237-262,264-283,287-290,292,294,296-298,301-308,310-313,315-320,324,332,341,353,356,358-359,369,371,376,380,382
-383,385-386,390,392,397-400,402-427,429-432,434-446,448-505,507-508,510-541,543-546,548-549,551-567,569-583,586-607,609-615,617-
621,623-624,627-633,636-639,641-642,645-646,648-649,652-667,670-700,703-719,721-723,725-726,728-731,733-737,740,746-795,1003-1004
,1012,1015,1017,1019,1027,1037-1038,1041,1108-1127]
cpubase_bynode_b1      up      3:00:00        9   idle gra[1021-1025,1028-1029,1039,1043]
cpubase_bynode_b2      up    12:00:00  15 drain* gra[222,732,988-997,1020,1030,1040]
cpubase_bynode_b2      up    12:00:00  16   drng gra[13,33,37,39,46,60,67-68,71,79,87,115,120,130,135,343]
cpubase_bynode_b2      up    12:00:00  144    mix
gra[44,47,91,100-101,116,118,124,138-139,225,236,263,284-286,291,293,295,299-300,309,314,321-323,325-331,333-340,342,344-352,354-
355,357,360-368,370,372-375,377-379,381,384,387-389,391,393-396,401,428,433,447,506,509,542,547,550,568,584-585,608,616,622,625-6
26,634-635,640,643-644,647,650-651,668-669,701-702,720,724,727,738-739,741-745,998-1002,1005-1011,1013-1014,1016,1018,1026,1031-1
036,1042]
cpubase_bynode_b2      up    12:00:00  667  alloc
gra[1-12,14-32,34-36,38,40-43,45,48-59,61-66,69-70,72-78,80-86,88-90,92-99,102-114,117,119,121-123,125-129,131-134,136-137,140-22
1,223-224,226-235,237-262,264-283,287-290,292,294,296-298,301-308,310-313,315-320,324,332,341,353,356,358-359,369,371,376,380,382
-383,385-386,390,392,397-400,402-427,429-432,434-446,448-505,507-508,510-541,543-546,548-549,551-567,569-583,586-607,609-615,617-
621,623-624,627-633,636-639,641-642,645-646,648-649,652-667,670-700,703-719,721-723,725-726,728-731,733-737,740,746-795,1003-1004
```

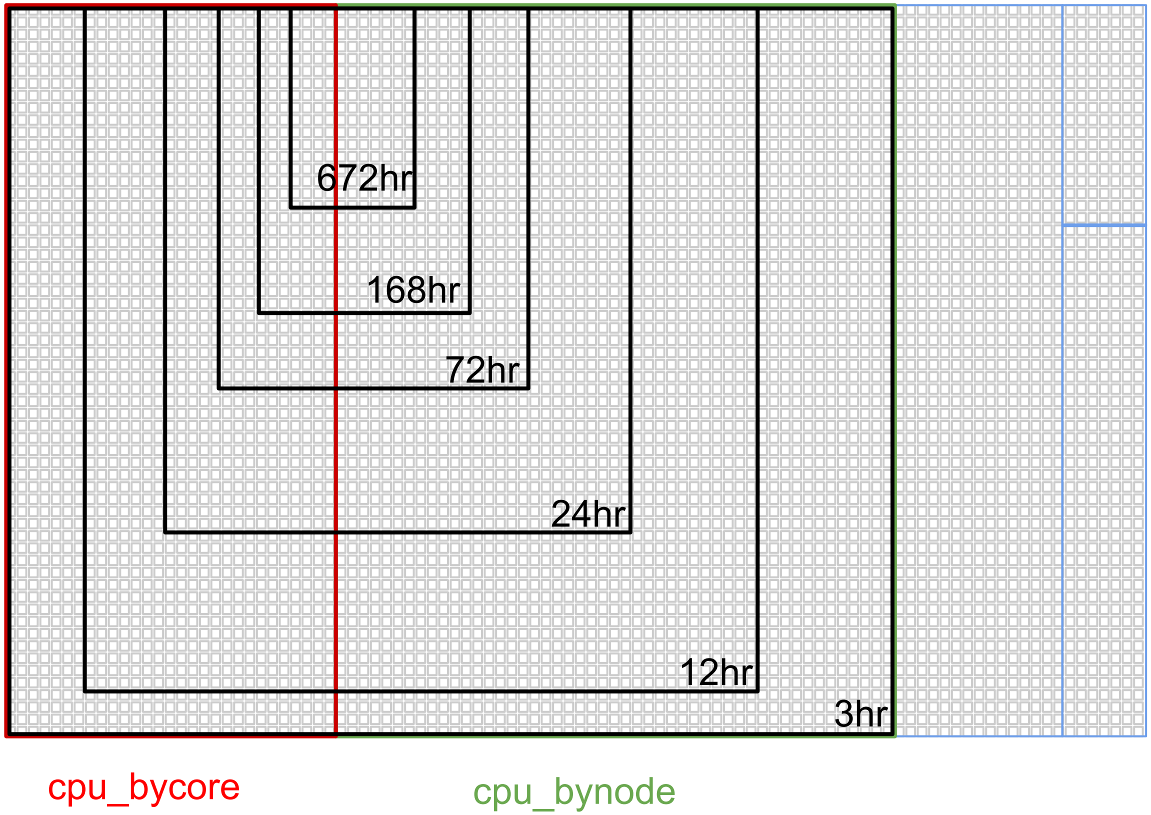
Monitoring jobs, the queue and the cluster

```
[jdesjard@gra-login4 ~]$ partition-stats

Node type |                   Max walltime
        |   3 hr   |  12 hr  |  24 hr  |  72 hr  |  168 hr |  672 hr |
----------|-------------------------------------------------------------
        Number of Queued Jobs by partition Type (by node:by core)
----------|-------------------------------------------------------------
Regular   |   29:179 |   7:5492|  69:293 | 238:724 |   1:945 |   3:118 |
Large Mem |    1:0   |   0:0   |   0:0   |   0:9   |   0:6   |   2:2   |
GPU     |      0:101 |   0:10  |   0:44  | 181:23  | 412:35  |   1:0   |
----------|-------------------------------------------------------------
        Number of Running Jobs by partition Type (by node:by core)
----------|-------------------------------------------------------------
Regular   |   43:76  |  14:1437|  73:204 | 106:250 |   7:960 |  24:110 |
Large Mem |    0:0   |   0:0   |   0:0   |   0:1   |   0:1   |   0:2   |
GPU     |      0:18  |   1:36  |  15:53  |  49:39  |   0:7   |   0:2   |
----------|-------------------------------------------------------------
        Number of Idle nodes by partition Type (by node:by core)
----------|-------------------------------------------------------------
Regular   |    1:0   |   1:0   |   1:0   |   1:0   |   0:0   |   0:0   |
Large Mem |    3:1   |   3:1   |   0:0   |   0:0   |   0:0   |   0:0   |
GPU     |   13:0   |  13:0   |   7:0   |   0:0   |   0:0   |   0:0   |
----------|-------------------------------------------------------------
        Total Number of nodes by partition Type (by node:by core)
----------|-------------------------------------------------------------
Regular   |  871:431 | 851:411 | 821:391 | 636:276 | 281:164 |  90:50  |
Large Mem |   27:12  |  27:12  |  24:11  |  20:3   |   4:3   |   3:2   |
GPU     |  156:78  | 156:78  | 144:72  | 104:52  |  13:12  |  13:12  |
----------|-------------------------------------------------------------
```

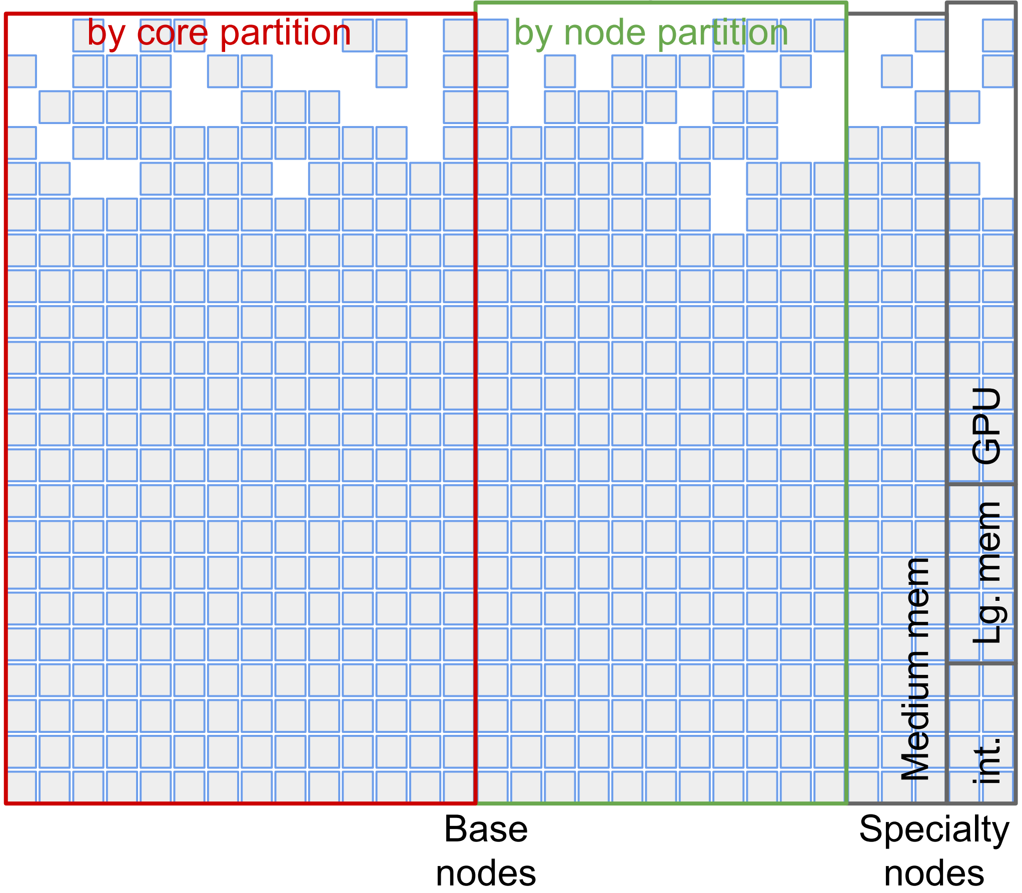
Monitoring jobs, the queue and the cluster

```
[jdesjard@gra-login4 ~]$ scontrol show partition
PartitionName=cpubase_interac
   AllowGroups=ALL AllowAccounts=ALL AllowQos=ALL
   AllocNodes=ALL Default=NO QoS=N/A
   DefaultTime=01:00:00 DisableRootJobs=NO ExclusiveUser=NO GraceTime=0 Hidden=NO
   MaxNodes=UNLIMITED MaxTime=03:00:00 MinNodes=1 LLN=NO MaxCPUsPerNode=UNLIMITED
   Nodes=gra[796-800]
   PriorityJobFactor=1 PriorityTier=1 RootOnly=NO ReqResv=NO OverSubscribe=NO
   OverTimeLimit=NONE PreemptMode=OFF
   State=UP TotalCPUs=160 TotalNodes=5 SelectTypeParameters=NONE
   DefMemPerCPU=256 MaxMemPerNode=UNLIMITED
   TRESBillingWeights=CPU=1.0,Mem=0.25G

PartitionName=cpubase_bynode_b1
   AllowGroups=ALL AllowAccounts=ALL AllowQos=ALL
   AllocNodes=ALL Default=NO QoS=N/A
   DefaultTime=01:00:00 DisableRootJobs=NO ExclusiveUser=NO GraceTime=0 Hidden=NO
   MaxNodes=UNLIMITED MaxTime=03:00:00 MinNodes=1 LLN=NO MaxCPUsPerNode=UNLIMITED
   Nodes=gra[1-795,988-1043,1108-1127]
   PriorityJobFactor=12 PriorityTier=1 RootOnly=NO ReqResv=NO OverSubscribe=NO
   OverTimeLimit=NONE PreemptMode=OFF
   State=UP TotalCPUs=27872 TotalNodes=871 SelectTypeParameters=NONE
   DefMemPerCPU=256 MaxMemPerNode=UNLIMITED
   TRESBillingWeights=CPU=1.0,Mem=0.25G

PartitionName=cpubase_bynode_b2
   AllowGroups=ALL AllowAccounts=ALL AllowQos=ALL
   AllocNodes=ALL Default=NO QoS=N/A
...
```

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



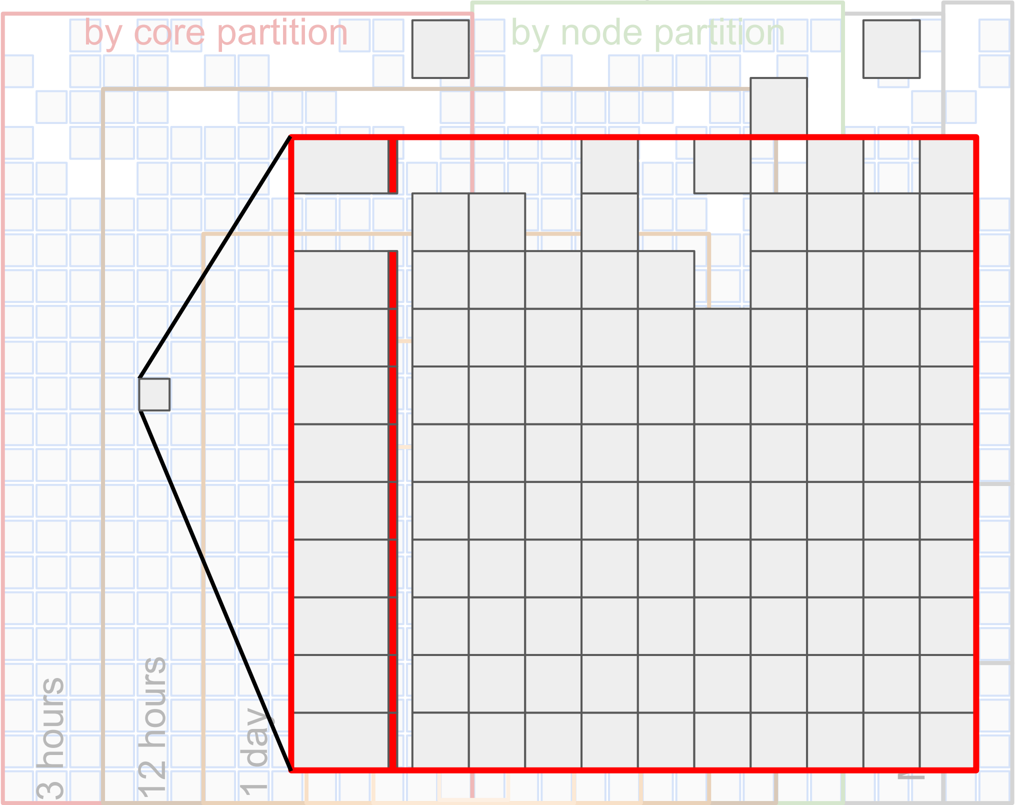
Cluster resource basics: categorization of resources that affect priority (partitions)



Partitions

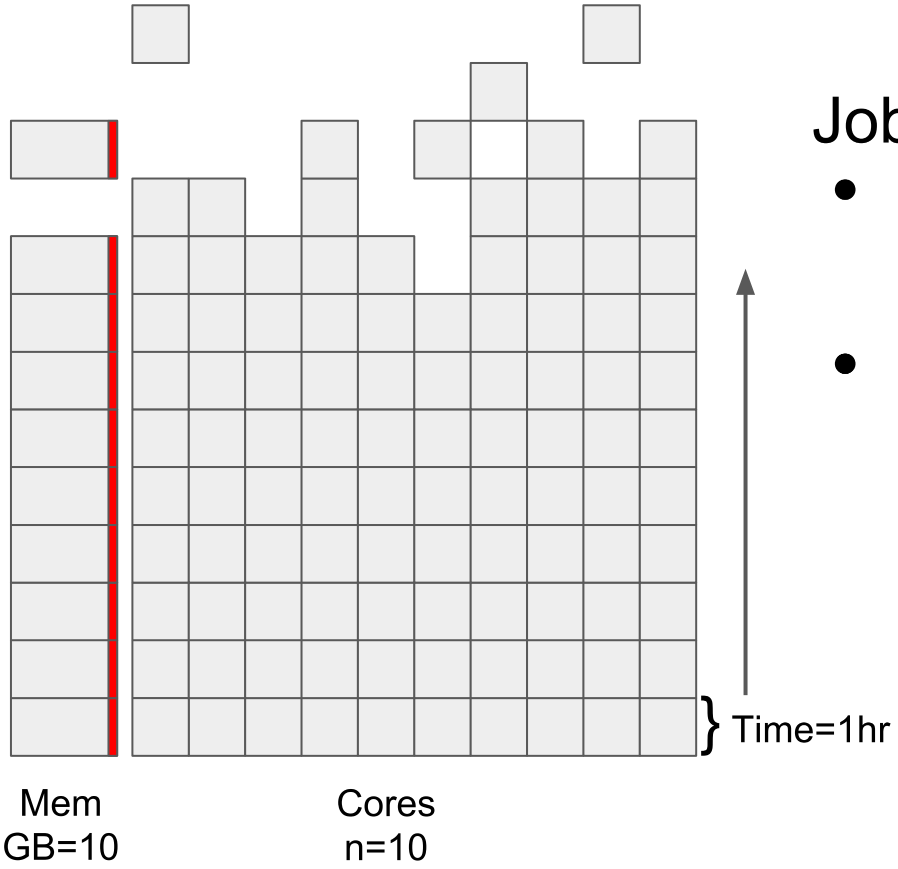
- By node
  - ntasks=32
    nodes=1
- By core
  - ntasks=32

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



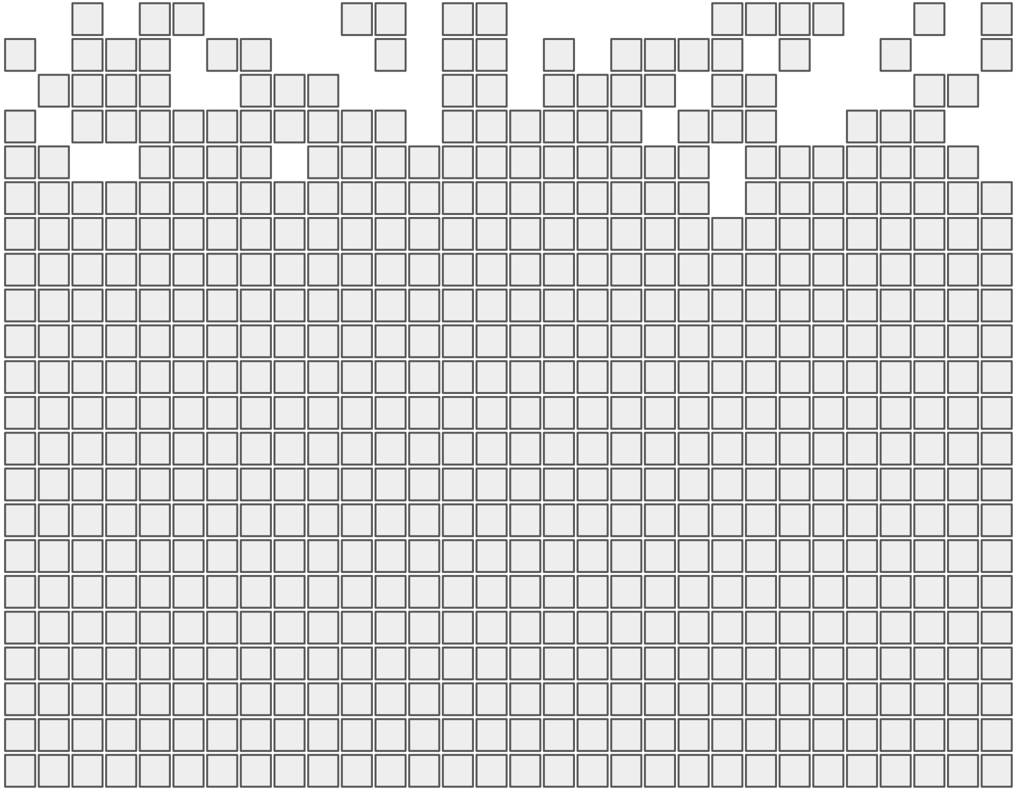
Backfill

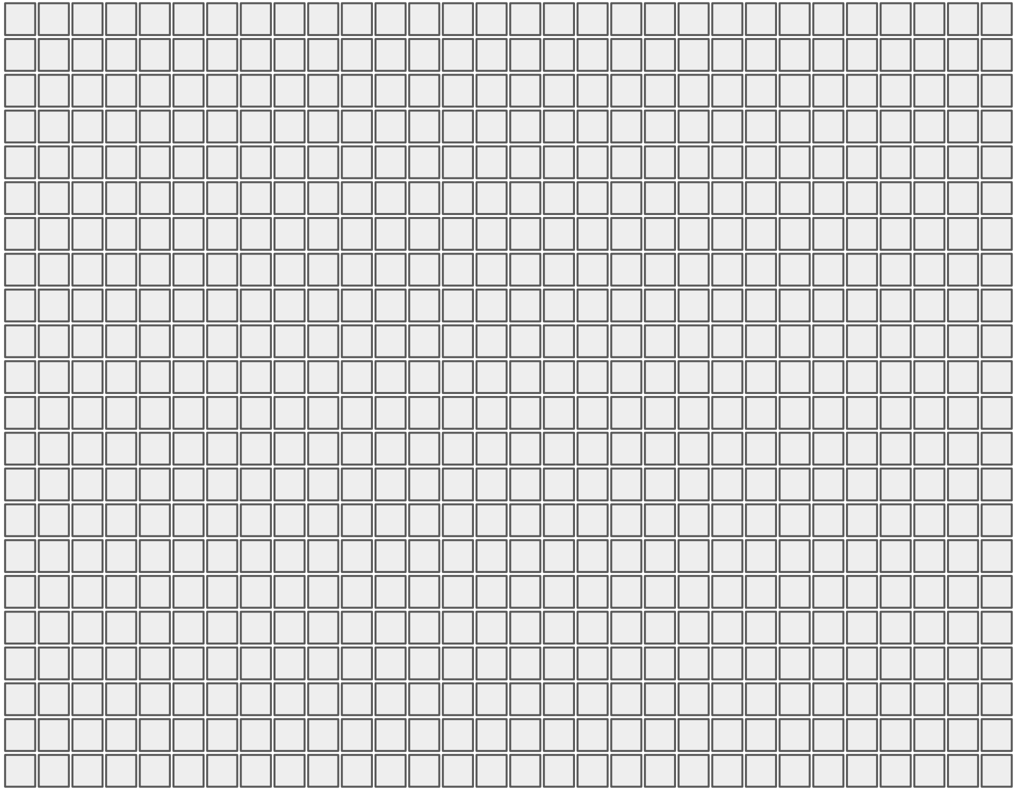
Scheduling basics: node resources and resource requests (job queue)



Job size

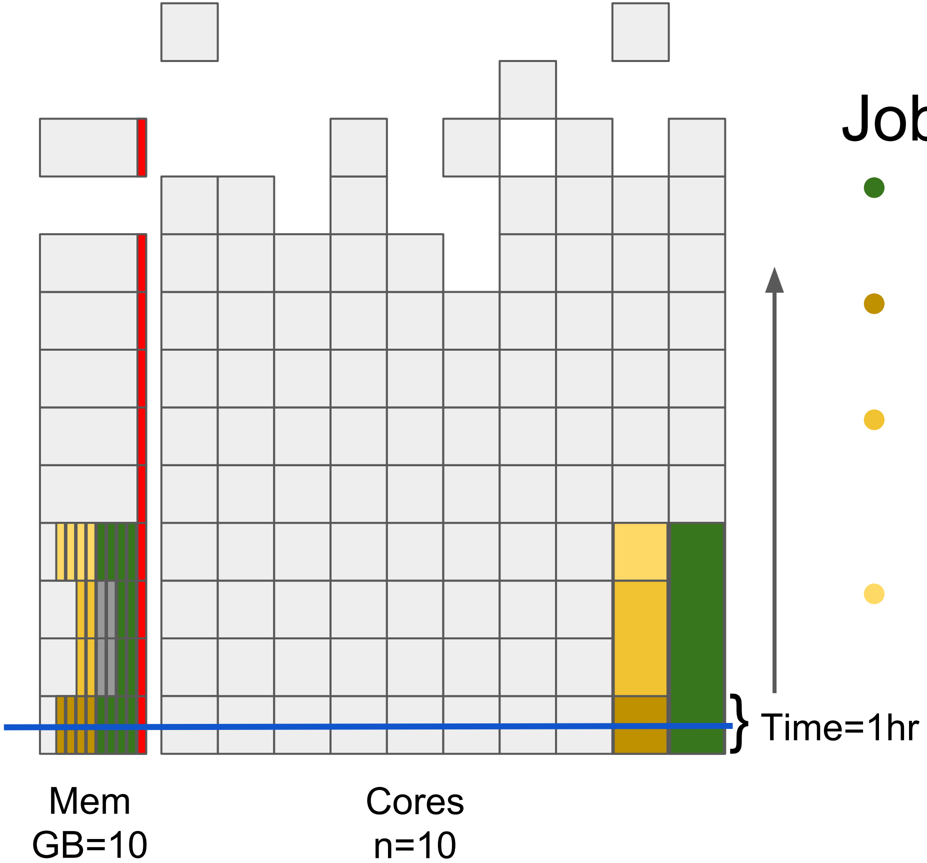
- Full node
  - MPI
  - Threaded
- By core
  - MPI
  - Threaded
  - serial





REMOVE PAGE

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

Factors contributing to job queue time

Job resource footprint (shape of the job on the cluster)

Load on the system (relative to resources available)

Account target share (fairshare priority)

Monitoring jobs, the queue and the cluster

cluster

- sinfo
- scontrol show

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)

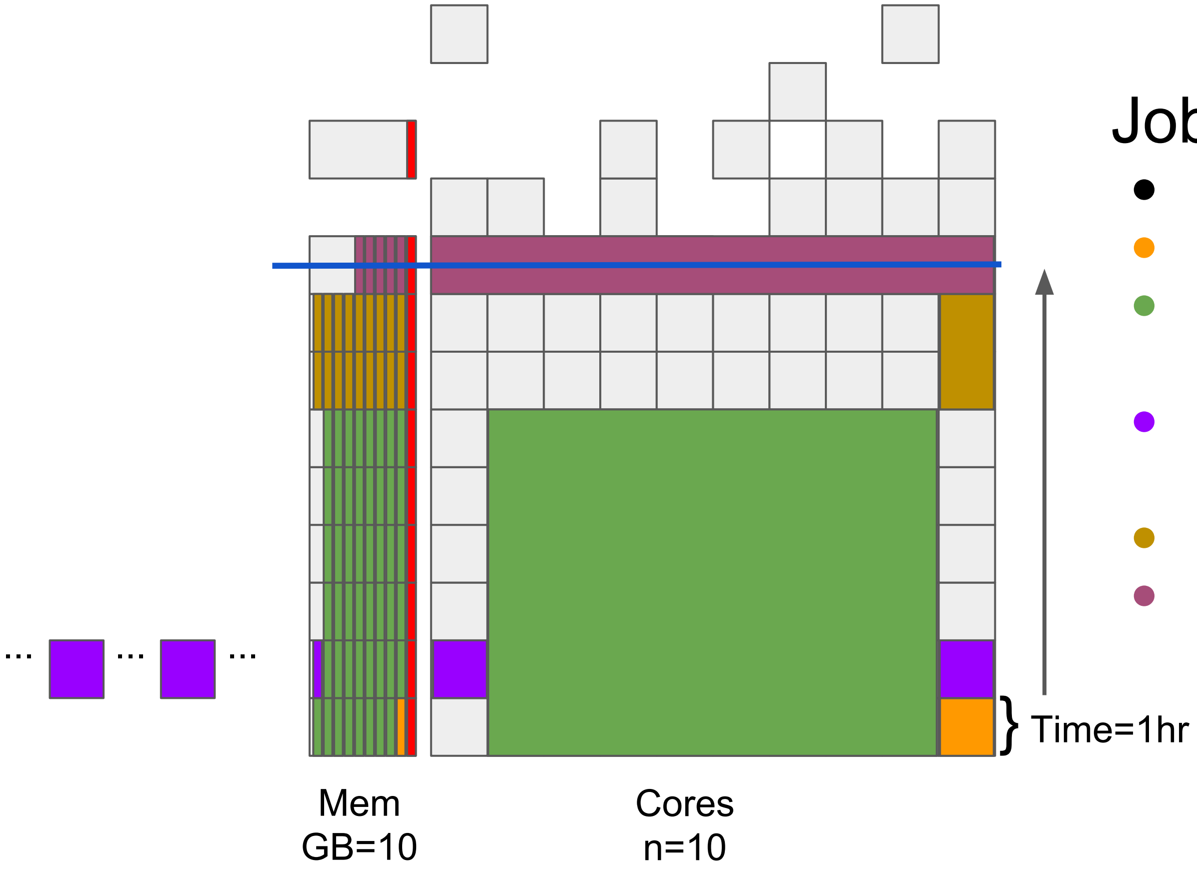
Fair-share

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

Partition

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

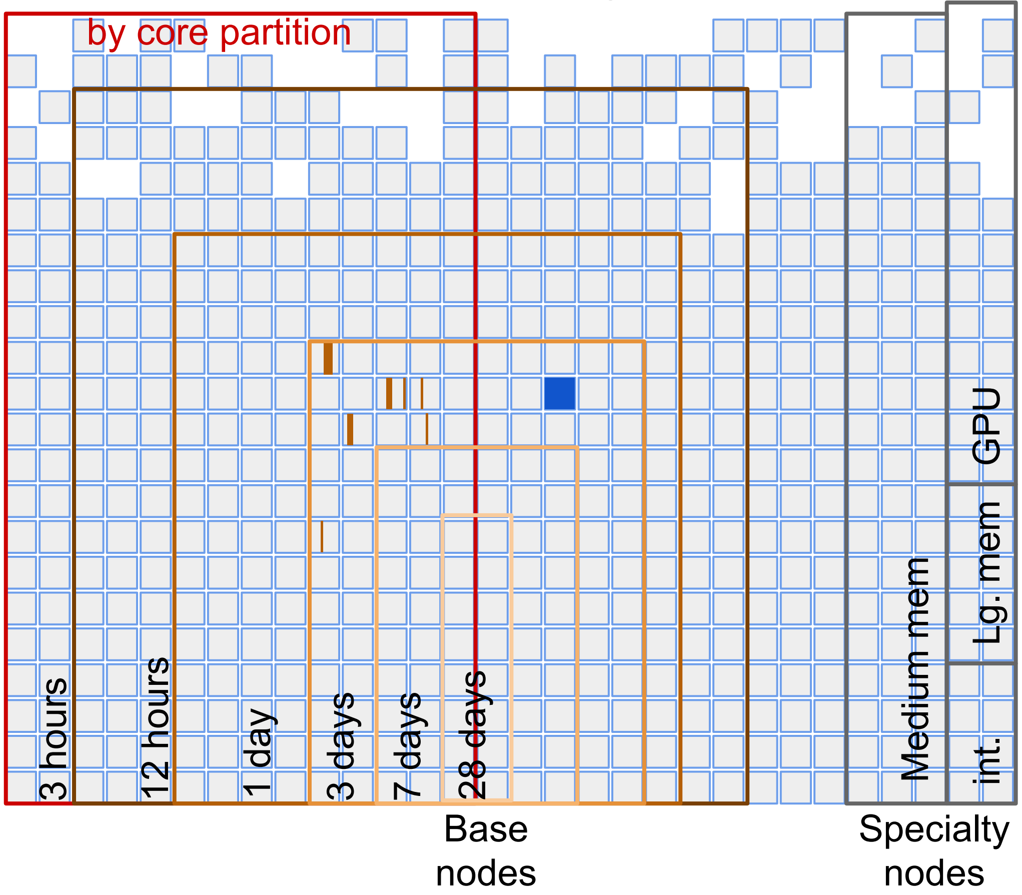
Job queue basics: factors that affect the order of jobs in queue (priority)



Job age

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

Cluster resource basics: segmentation of nodes in the cluster (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 --ntasks=32 --nodes=1
- --time=3-00:00 --ntasks=32 --nodes=1

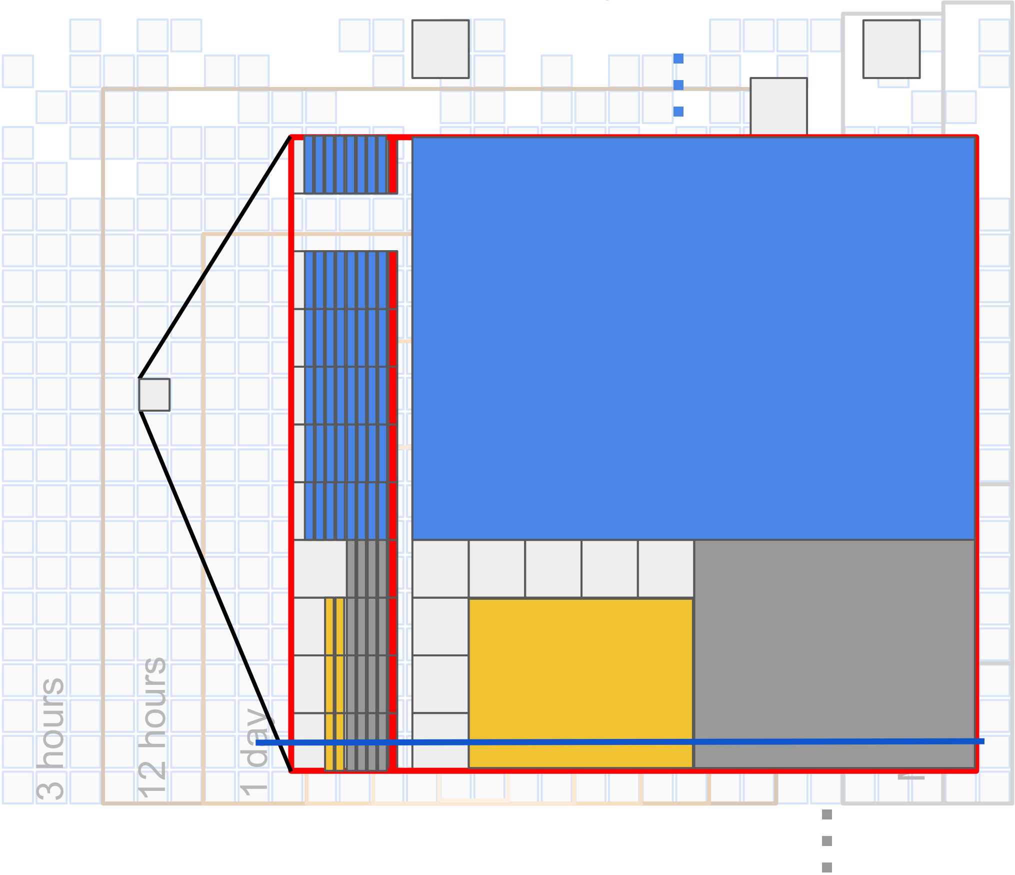
General purpose clusters

Traditionally SHARCNET systems were relatively homogeneous

The researcher chose a system based on fitting job resources to system specs

On Graham and Cedar the scheduler makes decisions about where a job runs on a heterogeneous system.

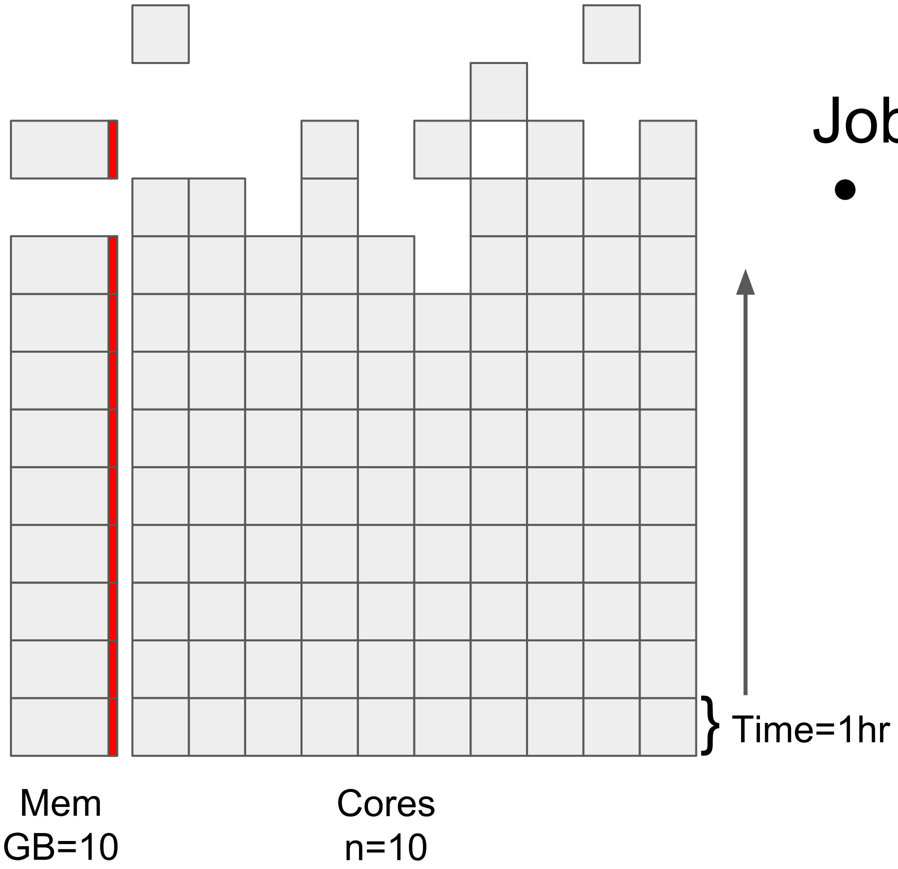
Cluster resource basics: segmentation of nodes in the cluster (partitions)



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

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



Job age

- FIFO (First In First Out)