Clusterstats

Demo

Clusterstats

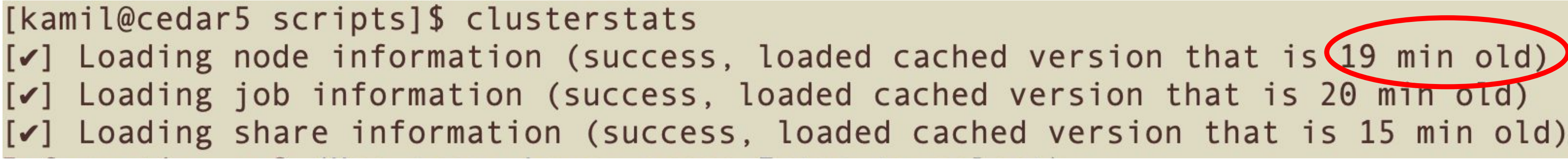
Run it by typing in “clusterstats”

It will be using a cached version of cluster information.

 Job information may not include newly submitted jobs

 Node state information can be a few minutes out of date.

If the cached info is old it may take a few minutes to run to get fresh data.



```
[kamil@cedar5 scripts]$ clusterstats
[✔] Loading node information (success, loaded cached version that is 19 min old)
[✔] Loading job information (success, loaded cached version that is 20 min old)
[✔] Loading share information (success, loaded cached version that is 15 min old)
```

We do this for 2 main reasons

1. We have cached information because it is much faster to run clusterstats from the cache.
2. Querying the scheduler in this detail is quite taxing and slows scheduler responsiveness when other people run commands.

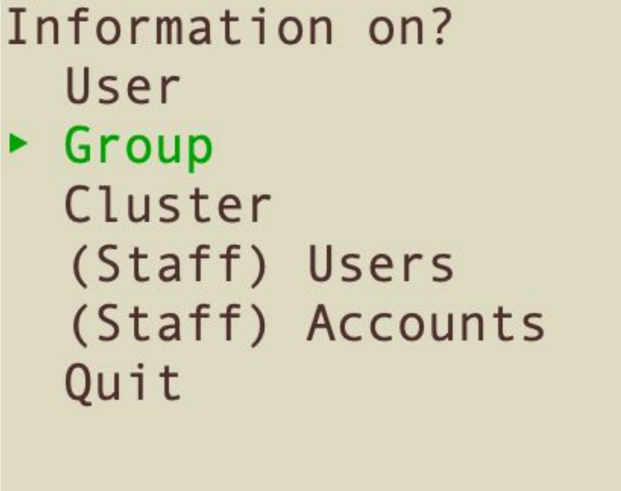
Main Menu

You will have 3 main options; select via arrow keys and enter.

User - Contains info on your jobs and your usage of different accounts/groups

Group - Contains info on your group(s) and other group members usage.

Cluster - Contains info on the cluster state, partitions and nodes



```
Information on?
  User
▸ Group
  Cluster
  (Staff) Users
  (Staff) Accounts
  Quit
```

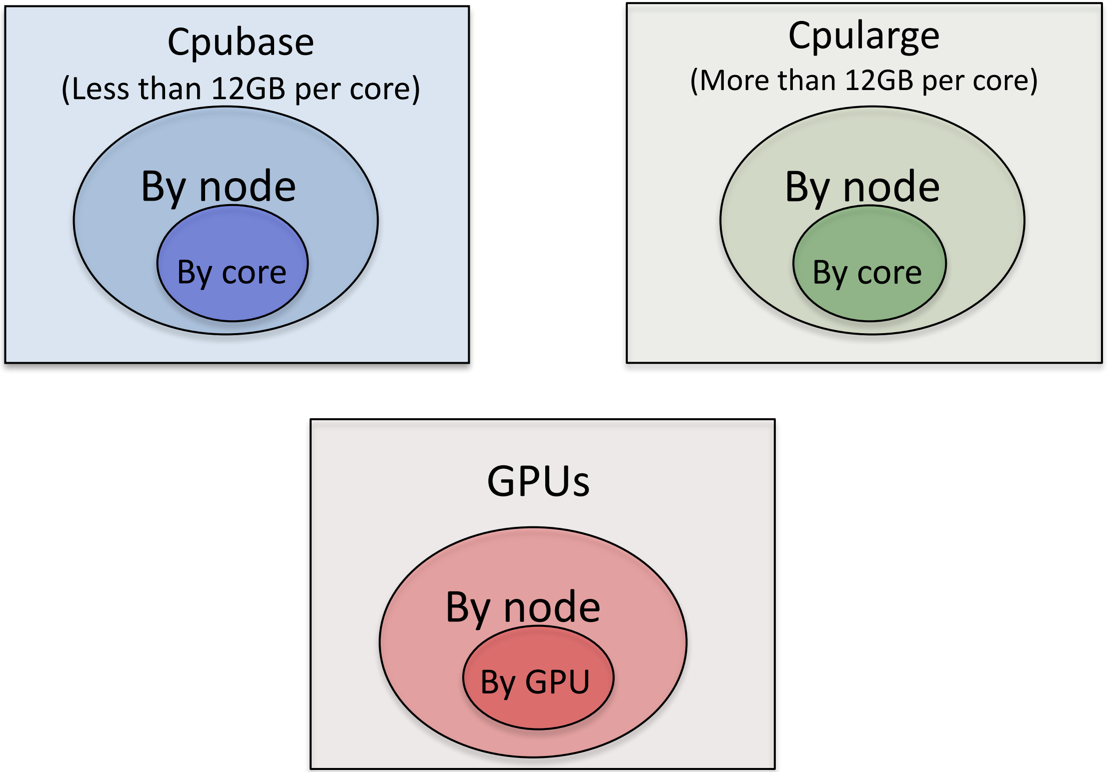
Compute Canada staff members have 2 additional options:

(Staff) Users - Contains a list of users with jobs on the cluster, and when a user is selected shows the user information above for the selected user.

(Staff) Accounts Contains a list of Accounts/Groups with jobs on the cluster and when an account is selected, displays the same group information as described above.

Partitions on Cedar, Graham and Beluga

Type



Cpubase
(Less than 12GB per core)

By node

By core

Cpularge
(More than 12GB per core)

By node

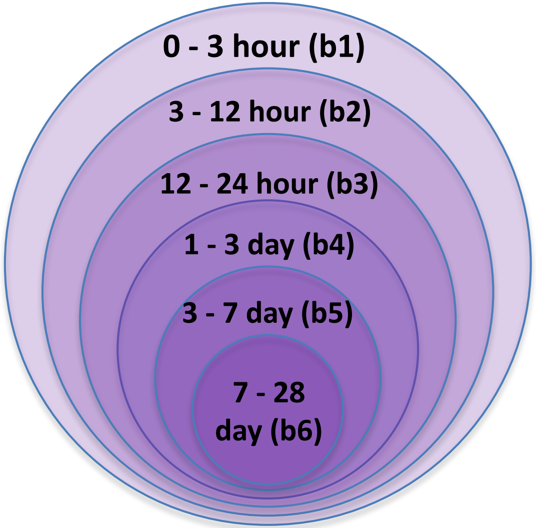
By core

GPUs

By node

By GPU

Walltime



0 - 3 hour (b1)

3 - 12 hour (b2)

12 - 24 hour (b3)

1 - 3 day (b4)

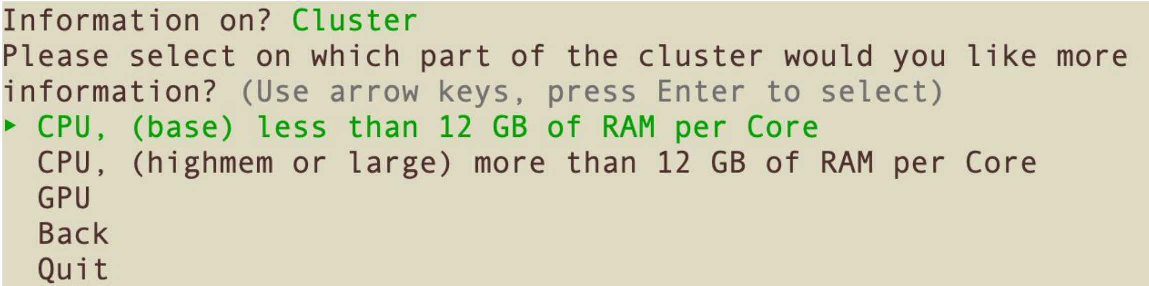
3 - 7 day (b5)

7 - 28
day (b6)

Cluster Menu

Back - goes back to the main menu.

Quit - Exits cluster stats



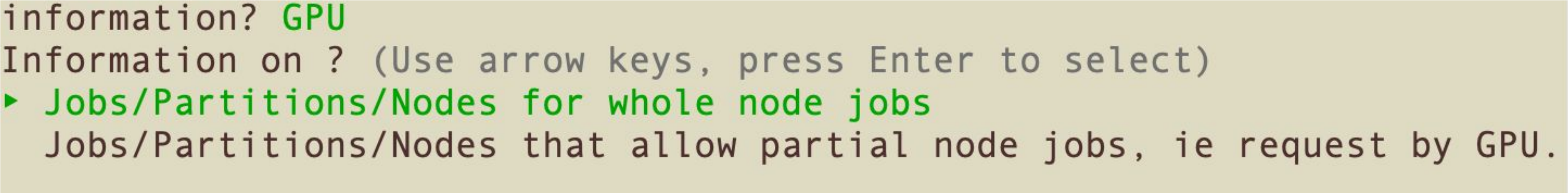
```
Information on? Cluster
Please select on which part of the cluster would you like more
information? (Use arrow keys, press Enter to select)
‣ CPU, (base) less than 12 GB of RAM per Core
  CPU, (highmem or large) more than 12 GB of RAM per Core
  GPU
  Back
  Quit
```

GPU - Select this to see more info on the nodes and partitions with GPUs

CPU (high mem) - Select this to see more info on the nodes and partitions with larger amounts of memory, those that run jobs with 12 GB of RAM per core or more.

CPU (base) - Select this to see more info about the regular most common nodes types, and partitions without GPUs or large amounts of memory.

Cluster Menu (part 2)



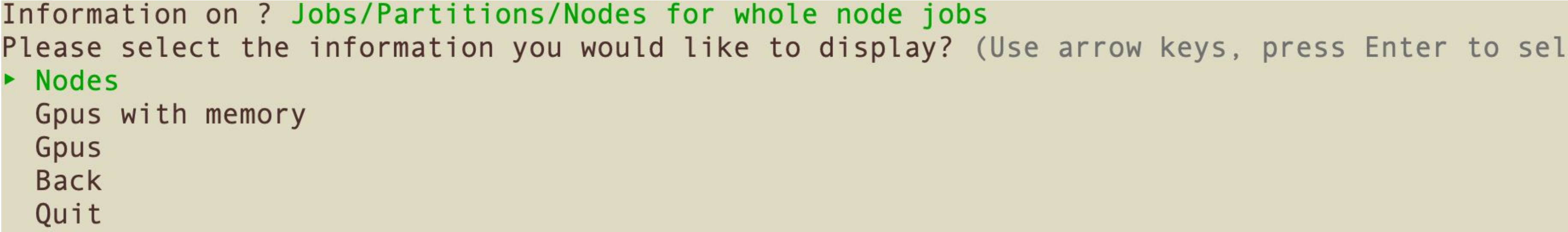
```
information? GPU
Information on ? (Use arrow keys, press Enter to select)
▸ Jobs/Partitions/Nodes for whole node jobs
  Jobs/Partitions/Nodes that allow partial node jobs, ie request by GPU.
```

Certain partitions and nodes are reserved for jobs that take up whole nodes.

Select **Jobs/Partitions/Nodes for whole node jobs** to see all the nodes and partitions that can run whole node jobs.

Select **Jobs/Partitions/Nodes that allow partial node jobs,** ex) jobs requesting individual CPU cores, GPUs to see all the nodes and partitions that can these type of jobs

Cluster Menu (part 3)



```
Information on ? Jobs/Partitions/Nodes for whole node jobs
Please select the information you would like to display? (Use arrow keys, press Enter to sel
▸ Nodes
  Gpus with memory
  Gpus
  Back
  Quit
```

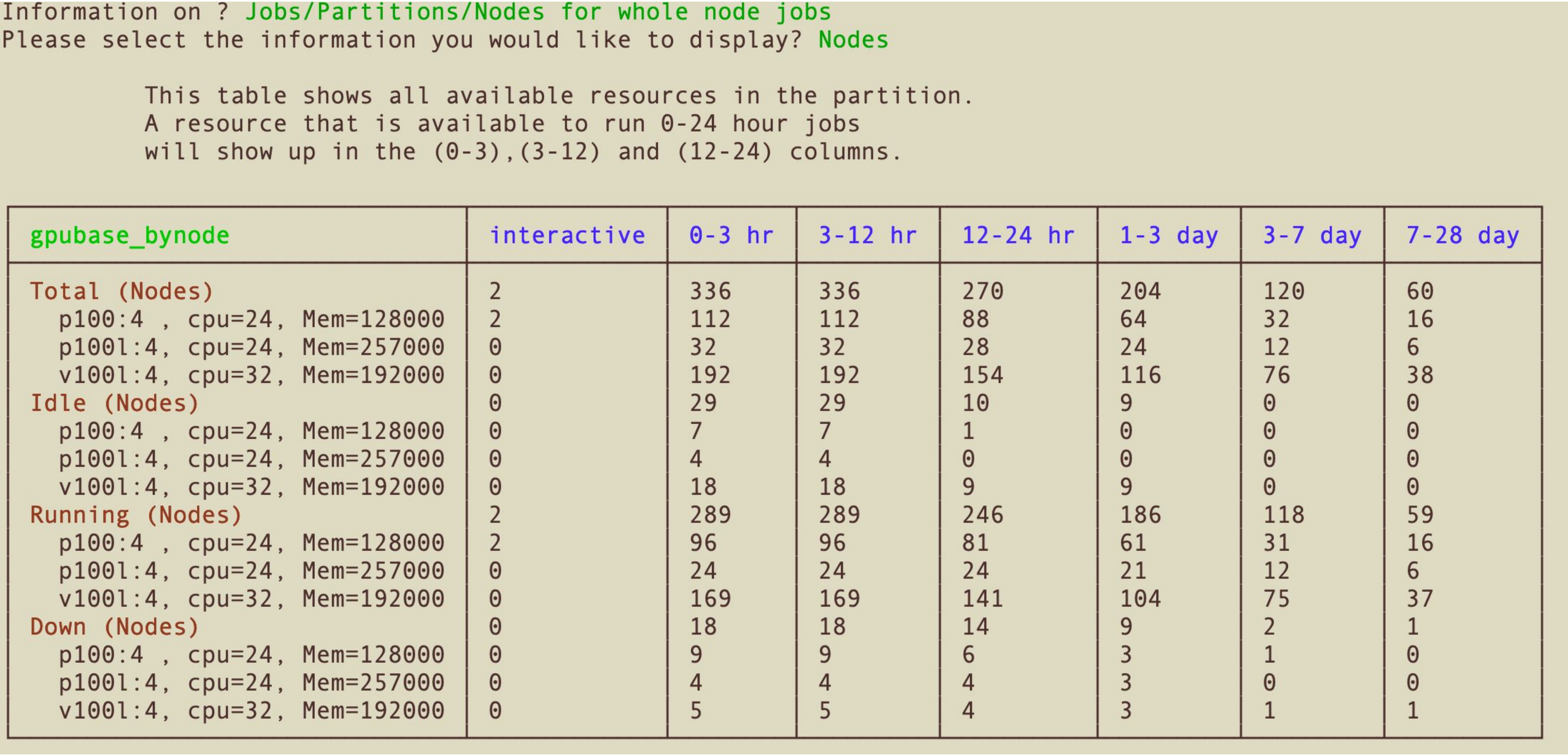
The final cluster menu asks what type of information do you wish to see:

Nodes - Will display node information

GPUs with Memory - Will display available GPUs or CPU cores that also have memory available

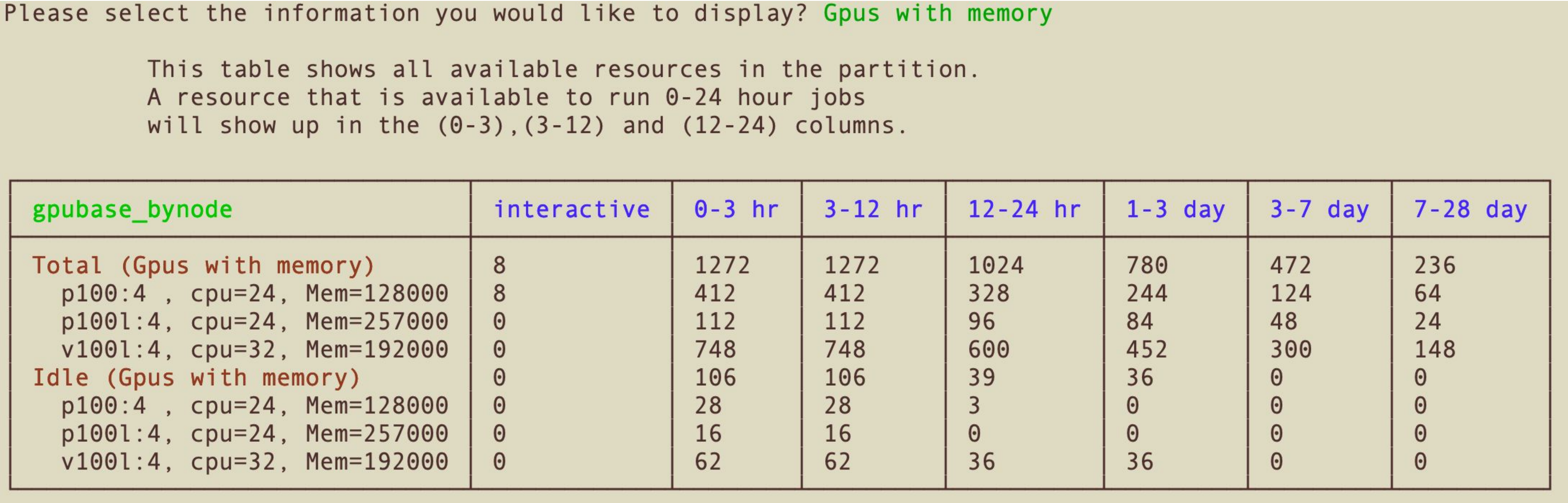
GPUs - Will display available GPUs or CPU cores regardless of memory availability

Cluster Menu - Nodes table



```
Information on ? Jobs/Partitions/Nodes for whole node jobs
Please select the information you would like to display? Nodes

         This table shows all available resources in the partition.
         A resource that is available to run 0-24 hour jobs
         will show up in the (0-3),(3-12) and (12-24) columns.

 gpubase_bynode              interactive  0-3 hr  3-12 hr  12-24 hr  1-3 day  3-7 day  7-28 day

 Total (Nodes)               2            336     336      270       204      120      60
   p100:4 , cpu=24, Mem=128000  2         112     112      88        64       32       16
   p100l:4, cpu=24, Mem=257000  0         32      32       28        24       12       6
   v100l:4, cpu=32, Mem=192000  0         192     192      154       116      76       38
 Idle (Nodes)                0            29      29       10        9        0        0
   p100:4 , cpu=24, Mem=128000  0         7       7        1         0        0        0
   p100l:4, cpu=24, Mem=257000  0         4       4        0         0        0        0
   v100l:4, cpu=32, Mem=192000  0         18      18       9         9        0        0
 Running (Nodes)             2            289     289      246       186      118      59
   p100:4 , cpu=24, Mem=128000  2         96      96       81        61       31       16
   p100l:4, cpu=24, Mem=257000  0         24      24       24        21       12       6
   v100l:4, cpu=32, Mem=192000  0         169     169      141       104      75       37
 Down (Nodes)                0            18      18       14        9        2        1
   p100:4 , cpu=24, Mem=128000  0         9       9        6         3        1        0
   p100l:4, cpu=24, Mem=257000  0         4       4        4         3        0        0
   v100l:4, cpu=32, Mem=192000  0         5       5        4         3        1        1
```

Cluster Menu - GPU or Cores with memory table

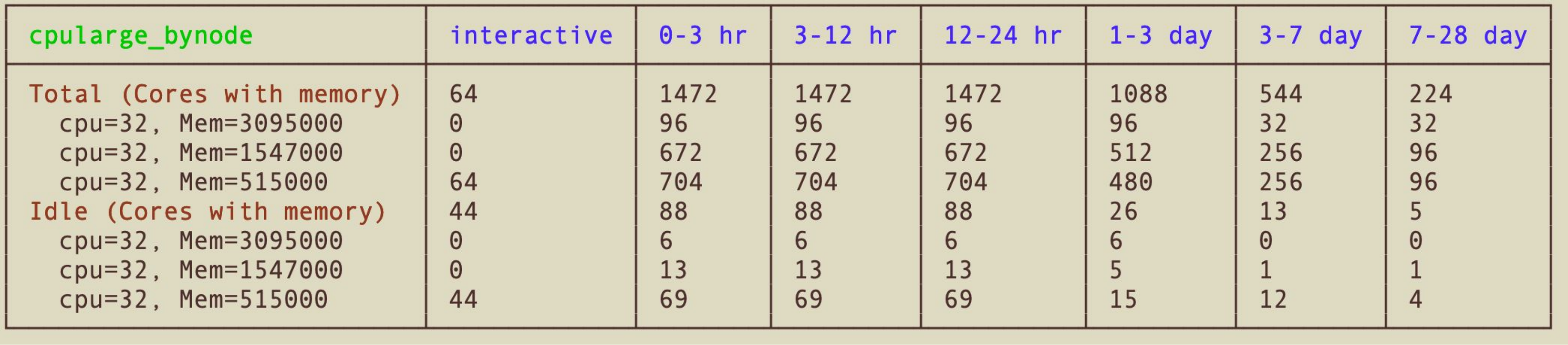


```
Please select the information you would like to display? Gpus with memory

         This table shows all available resources in the partition.
         A resource that is available to run 0-24 hour jobs
         will show up in the (0-3),(3-12) and (12-24) columns.

┌────────────────────────────────┬─────────────┬────────┬──────────┬───────────┬──────────┬──────────┬──────────┐
│ gpubase_bynode                 │ interactive │ 0-3 hr │ 3-12 hr  │ 12-24 hr  │ 1-3 day  │ 3-7 day  │ 7-28 day │
├────────────────────────────────┼─────────────┼────────┼──────────┼───────────┼──────────┼──────────┼──────────┤
│ Total (Gpus with memory)       │ 8           │ 1272   │ 1272     │ 1024      │ 780      │ 472      │ 236      │
│   p100:4 , cpu=24, Mem=128000  │ 8           │ 412    │ 412      │ 328       │ 244      │ 124      │ 64       │
│   p100l:4, cpu=24, Mem=257000  │ 0           │ 112    │ 112      │ 96        │ 84       │ 48       │ 24       │
│   v100l:4, cpu=32, Mem=192000  │ 0           │ 748    │ 748      │ 600       │ 452      │ 300      │ 148      │
│ Idle (Gpus with memory)        │ 0           │ 106    │ 106      │ 39        │ 36       │ 0        │ 0        │
│   p100:4 , cpu=24, Mem=128000  │ 0           │ 28     │ 28       │ 3         │ 0        │ 0        │ 0        │
│   p100l:4, cpu=24, Mem=257000  │ 0           │ 16     │ 16       │ 0         │ 0        │ 0        │ 0        │
│   v100l:4, cpu=32, Mem=192000  │ 0           │ 62     │ 62       │ 36        │ 36       │ 0        │ 0        │
└────────────────────────────────┴─────────────┴────────┴──────────┴───────────┴──────────┴──────────┴──────────┘
```

P100:4, cpu=24, Mem=128000

Means that each node or computer has 4 GPUs of type p100, 24 CPU cores and 128,000 MiB of RAM. There are 28 “idle” GPUs on this nodetype with memory available to run an up-to-12-hour job, however if the job is 24 hours long there are only 3 and if it is longer there are none.

Cluster Menu - Table cores with memory for large memory partitions



```
┌───────────────────────────┬─────────────┬────────┬────────┬─────────┬─────────┬─────────┬──────────┐
│ cpularge_bynode           │ interactive │ 0-3 hr │ 3-12 hr│ 12-24 hr│ 1-3 day │ 3-7 day │ 7-28 day │
├───────────────────────────┼─────────────┼────────┼────────┼─────────┼─────────┼─────────┼──────────┤
│ Total (Cores with memory) │ 64          │ 1472   │ 1472   │ 1472    │ 1088    │ 544     │ 224      │
│   cpu=32, Mem=3095000     │ 0           │ 96     │ 96     │ 96      │ 96      │ 32      │ 32       │
│   cpu=32, Mem=1547000     │ 0           │ 672    │ 672    │ 672     │ 512     │ 256     │ 96       │
│   cpu=32, Mem=515000      │ 64          │ 704    │ 704    │ 704     │ 480     │ 256     │ 96       │
│ Idle (Cores with memory)  │ 44          │ 88     │ 88     │ 88      │ 26      │ 13      │ 5        │
│   cpu=32, Mem=3095000     │ 0           │ 6      │ 6      │ 6       │ 6       │ 0       │ 0        │
│   cpu=32, Mem=1547000     │ 0           │ 13     │ 13     │ 13      │ 5       │ 1       │ 1        │
│   cpu=32, Mem=515000      │ 44          │ 69     │ 69     │ 69      │ 15      │ 12      │ 4        │
└───────────────────────────┴─────────────┴────────┴────────┴─────────┴─────────┴─────────┴──────────┘
```

Here we are looking at the large memory partitions and how many cores with memory are available. On the nodes with **32 cores and 3 TiB** (3,095,000 MB) **of RAM**: 6 cores with memory in a partition that allows 3-day long jobs. On such a node, each core has 96 GiB of RAM, 6 CPU cores are sitting idle with 6 * 96 = 576 GB of RAM.

Possible analysis with this information: It may be possible to run a 3-day, 6-core, 576-GiB memory job. However, we don't know the reason that the resources are idle; there may be a high-priority job that has requested the whole node scheduled to run when the currently running jobs finish. This may take place in a few hours and only a shorter job could be run in the currently idle resources.

Fairness between groups and users

or FairTree Fairshare Tree

Group Name	Group's share	Group's use of resources	User	Users share in Group	User used % of Group use	User used % of total cluster resources
Alberta	50%	**70%**	Alice	50%	**0%**	**0%**
			Albert	50%	**100%**	**70%**
Brazil	50%	**30%**	Betty	50%	**66%**	**20%**
			Bob	50%	**33%**	**10%**

Alice and **Betty** have jobs in the queue.

Discuss:

Whose job should run first?

Why?

Is this fair?

Fairness between groups and users

or FairTree Fairshare Tree

Group Name	Group’s share	Group’s use of resources	User	Users share in Group	User used % of Group use	User used % of total cluster resources
Alberta	50%	**70%**	Alice	50%	**0%**	**0%**
			Albert	50%	**100%**	**70%**
Brazil	50%	**30%**	Betty	50%	**66%**	**20%**
			Bob	50%	**33%**	**10%**

Alice and **Betty** have jobs in the queue.

Compute Canada’s answer:

The group’s usage is always more important, Betty’s job has higher priority.

This is done via the Fairtree Fairshare Tree algorithm.

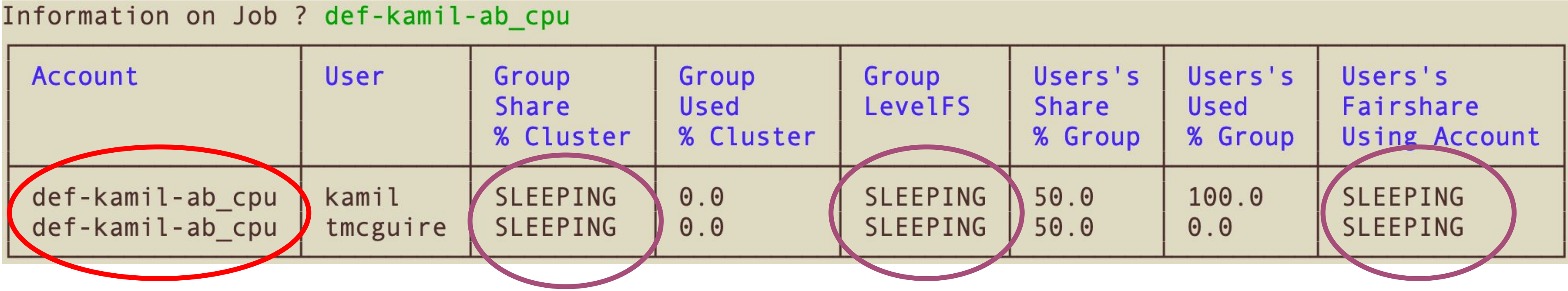
If all 4 users have jobs in the queue, then the users’ jobs in order of priority would be: Bob, Betty, Alice, Albert

Group menu, Group table

From the Group menu, select the Group account.

Groups that begin with **def** are default groups, groups that begin with rrg or rpp are allocated by the RAC (resource allocation competition) process.

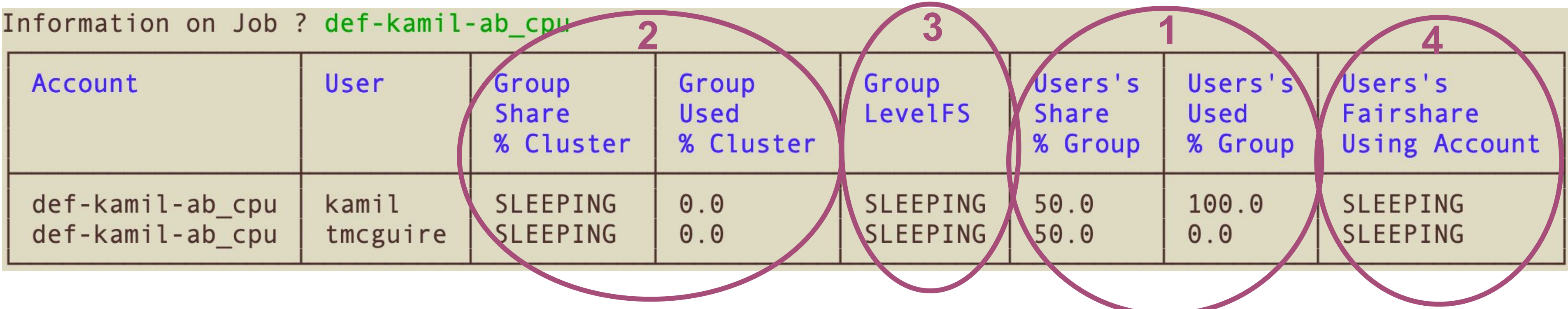
Default groups without jobs in the queue are **sleeping** and don’t get an allocation; active default groups get an equal share of unallocated resources which is about ~20% of each cluster.



```
Information on Job ? def-kamil-ab_cpu

 Account           User        Group      Group       Group      Users's   Users's   Users's
                               Share      Used        LevelFS    Share     Used      Fairshare
                               % Cluster  % Cluster              % Group   % Group   Using Account

 def-kamil-ab_cpu  kamil       SLEEPING   0.0         SLEEPING   50.0      100.0     SLEEPING
 def-kamil-ab_cpu  tmcguire    SLEEPING   0.0         SLEEPING   50.0      0.0       SLEEPING
```

Group menu, Group table

1) One can see 2 users here and kamil has an equal share in the group but has used all the resources so far, the group itself has used almost no resources.
2) A group’s cluster usage vs its share is vastly more important in determining jobs priority than the use or share of the individual.
3) LevelFS is the group’s share divided by the usage in the recent past.
4) The priority from Fairshare algorithm using this accounting group.

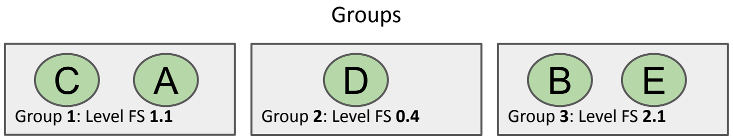


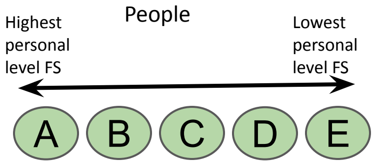
Information on Job ? def-kamil-ab_cpu

Account	User	Group Share % Cluster	Group Used % Cluster	Group LevelFS	Users's Share % Group	Users's Used % Group	Users's Fairshare Using Account
def-kamil-ab_cpu	kamil	SLEEPING	0.0	SLEEPING	50.0	100.0	SLEEPING
def-kamil-ab_cpu	tmcguire	SLEEPING	0.0	SLEEPING	50.0	0.0	SLEEPING

FairTree Fairshare Tree

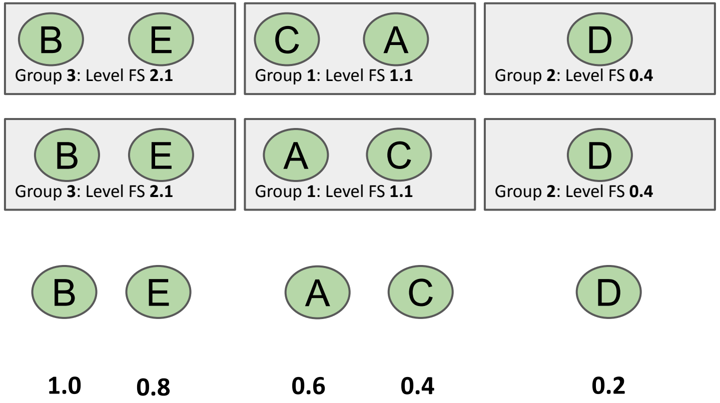
or fairness between groups and users


Groups
C
A
Group **1**: Level FS **1.1**
D
Group **2**: Level FS **0.4**
B
E
Group **3**: Level FS **2.1**



Giving users and accounts priority:

1. Sort the groups by Level FS
2. Sort within the group by usage/share
3. Rank users using the accounts
4. Assign priority according to the ranking
5. Schedule jobs according to priority


B
E
Group **3**: Level FS **2.1**
C
A
Group **1**: Level FS **1.1**
D
Group **2**: Level FS **0.4**
B
E
Group **3**: Level FS **2.1**
A
C
Group **1**: Level FS **1.1**
D
Group **2**: Level FS **0.4**
B
E
A
C
D
1.0
0.8
0.6
0.4
0.2

Fairness between groups and users

or FairTree Fairshare Tree

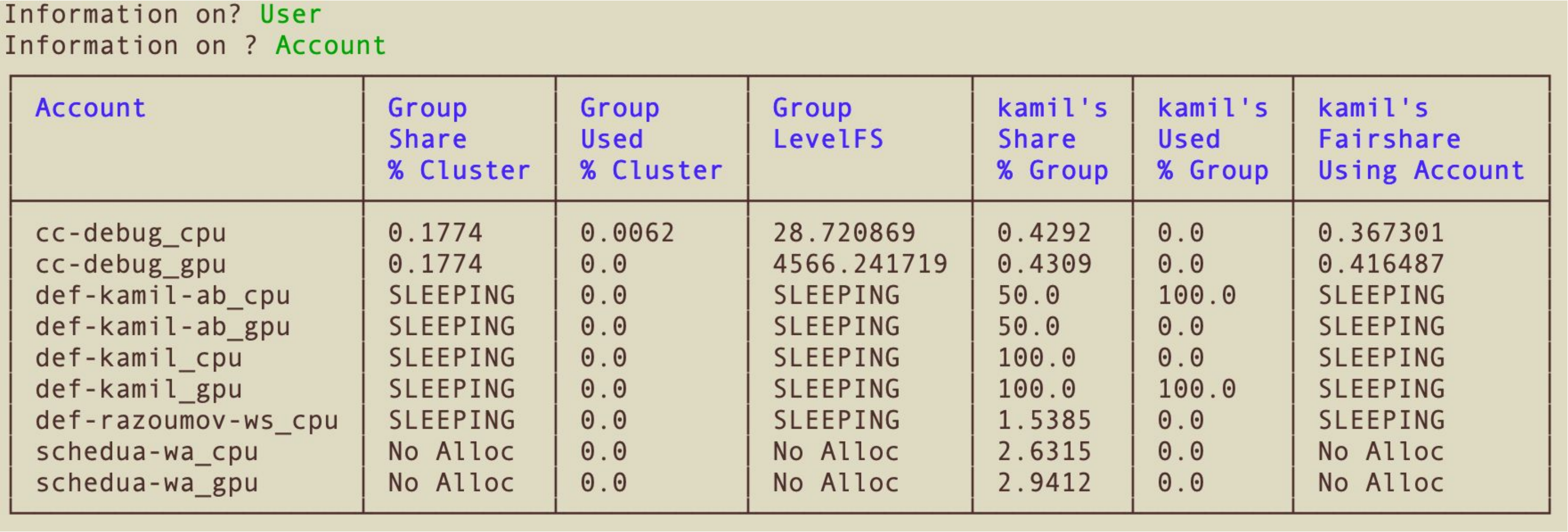
Your group’s usage and share matter the most.

Since there are a small number of users within your group and large number of groups, usage within a group matters little except when your jobs are directly competing with another group member over resources.

With such small differences in priority between group members, the member with the slightly easier-to-run job will run first.

Tip: check if a group member is using all the group’s resources.

User menu - Account - Table

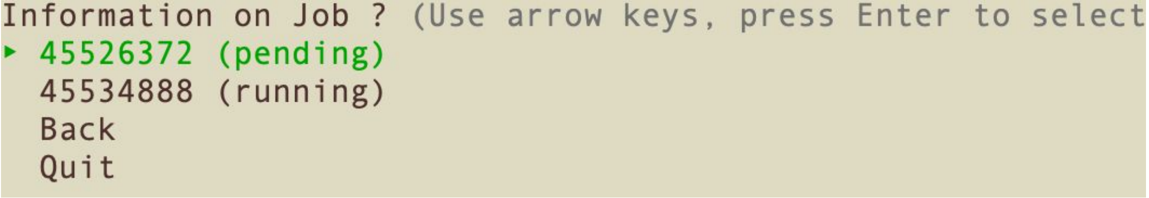


```
Information on? User
Information on ? Account
┌──────────────────────┬────────────┬────────────┬──────────────┬──────────┬──────────┬────────────────┐
│ Account              │ Group      │ Group      │ Group        │ kamil's  │ kamil's  │ kamil's        │
│                      │ Share      │ Used       │ LevelFS      │ Share    │ Used     │ Fairshare      │
│                      │ % Cluster  │ % Cluster  │              │ % Group  │ % Group  │ Using Account  │
├──────────────────────┼────────────┼────────────┼──────────────┼──────────┼──────────┼────────────────┤
│ cc-debug_cpu         │ 0.1774     │ 0.0062     │ 28.720869    │ 0.4292   │ 0.0      │ 0.367301       │
│ cc-debug_gpu         │ 0.1774     │ 0.0        │ 4566.241719  │ 0.4309   │ 0.0      │ 0.416487       │
│ def-kamil-ab_cpu     │ SLEEPING   │ 0.0        │ SLEEPING     │ 50.0     │ 100.0    │ SLEEPING       │
│ def-kamil-ab_gpu     │ SLEEPING   │ 0.0        │ SLEEPING     │ 50.0     │ 0.0      │ SLEEPING       │
│ def-kamil_cpu        │ SLEEPING   │ 0.0        │ SLEEPING     │ 100.0    │ 0.0      │ SLEEPING       │
│ def-kamil_gpu        │ SLEEPING   │ 0.0        │ SLEEPING     │ 100.0    │ 100.0    │ SLEEPING       │
│ def-razoumov-ws_cpu  │ SLEEPING   │ 0.0        │ SLEEPING     │ 1.5385   │ 0.0      │ SLEEPING       │
│ schedua-wa_cpu       │ No Alloc   │ 0.0        │ No Alloc     │ 2.6315   │ 0.0      │ No Alloc       │
│ schedua-wa_gpu       │ No Alloc   │ 0.0        │ No Alloc     │ 2.9412   │ 0.0      │ No Alloc       │
└──────────────────────┴────────────┴────────────┴──────────────┴──────────┴──────────┴────────────────┘
```

You can see your and your group’s share and usage for all the group accounts in which you are a member.

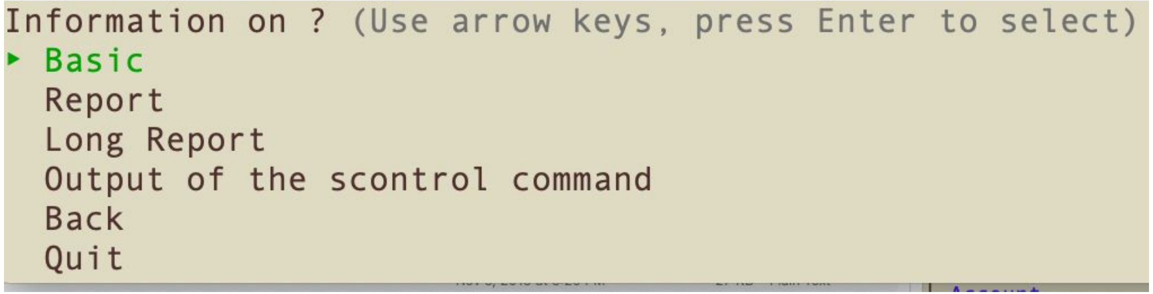
User menu -> Jobs Menu

The Jobs menu is located in the user menu, and you can select which job to get more information on.



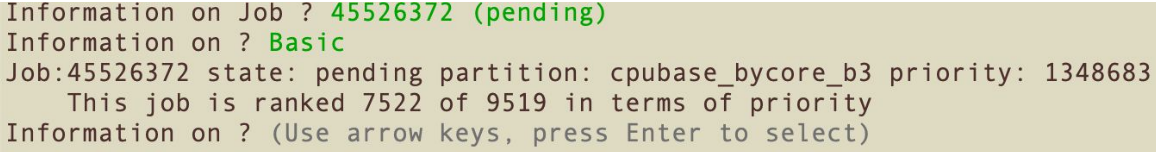
```
Information on Job ? (Use arrow keys, press Enter to select
▸ 45526372 (pending)
  45534888 (running)
  Back
  Quit
```

Basic information contains the job's priority and where it ranks compared to other jobs.



```
Information on ? (Use arrow keys, press Enter to select)
▸ Basic
  Report
  Long Report
  Output of the scontrol command
  Back
  Quit
```

scontrol output, showing job diagnostic information is also available.



```
Information on Job ? 45526372 (pending)
Information on ? Basic
Job:45526372 state: pending partition: cpubase_bycore_b3 priority: 1348683
   This job is ranked 7522 of 9519 in terms of priority
Information on ? (Use arrow keys, press Enter to select)
```

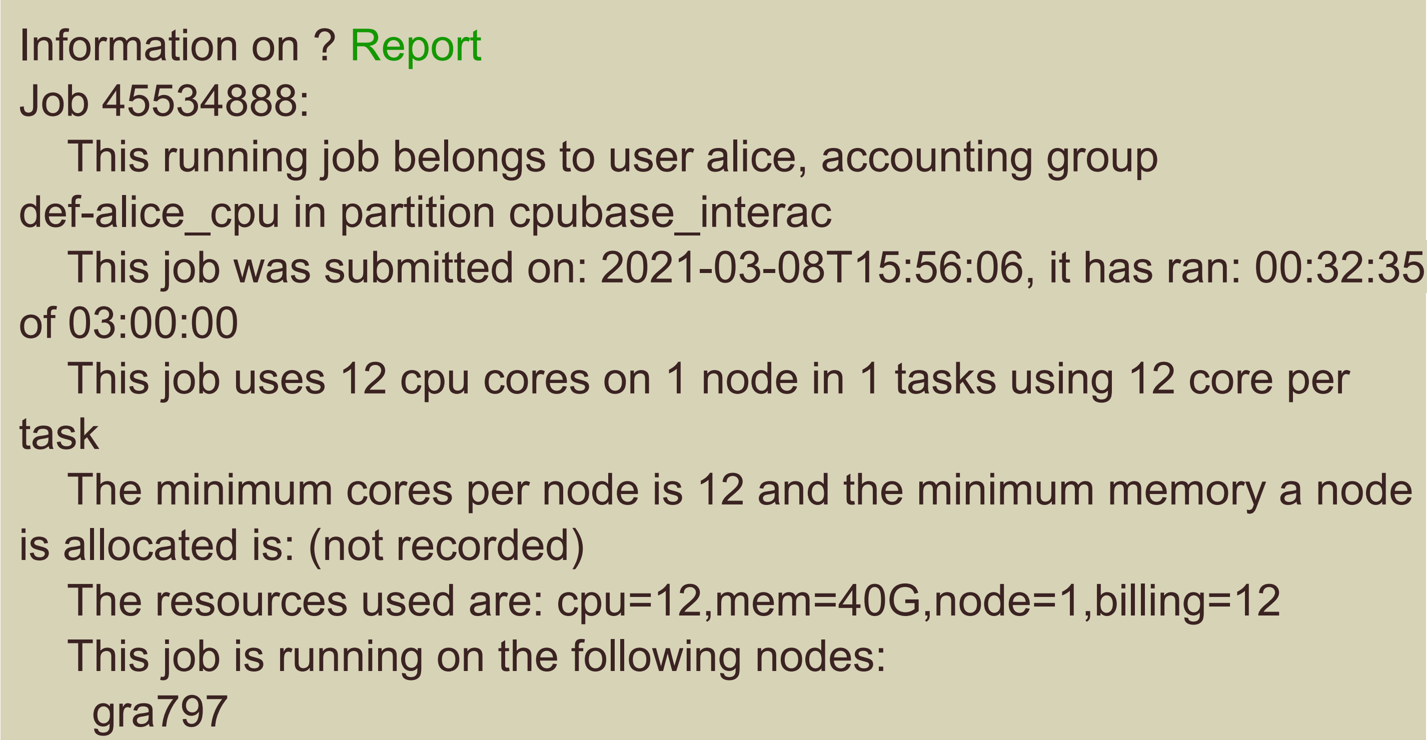
Jobs Menu Report

Job report shows more information, including the number and type of nodes that are available within the partition that your job is in.

Long Report has even more details.

```
Information on ? Report
Job 45526372:
    This pending job belongs to user Alice, accounting group def-alice_cpu in
partition cpubase_bycore_b3
    Nodes that can possibly run the job:
      Total: 627 Busy: 542 Down: 85 Idle: 0
        Node Type (cpu=32, Mem=128000):  Total 438 Down 85 Idle 0
        Node Type (cpu=32, Mem=256500):  Total 56 Down 0 Idle 0
        Node Type (cpu=44, Mem=191840):  Total 133 Down 0 Idle 0
      This job is ranked 7522 of 9519 in terms of priority on these nodes
```

Jobs Menu - Report on a Running Job



```
Information on ? Report
Job 45534888:
    This running job belongs to user alice, accounting group
def-alice_cpu in partition cpubase_interac
    This job was submitted on: 2021-03-08T15:56:06, it has ran: 00:32:35
of 03:00:00
    This job uses 12 cpu cores on 1 node in 1 tasks using 12 core per
task
    The minimum cores per node is 12 and the minimum memory a node
is allocated is: (not recorded)
    The resources used are: cpu=12,mem=40G,node=1,billing=12
    This job is running on the following nodes:
      gra797
```