# Running research at PDC

## Sharing the PDC systems with other users

There are many researchers who run their programs on PDC's computer systems (which are also known as clusters), so we need a way to decide when different people's programs are run and for how long. To do this we use a job scheduling system known as Slurm.

When someone wants to run a program at PDC, they submit a request that goes into a queue of jobs. As computing resources become available on the PDC systems, the Slurm Workload Manager decides which job to start on the resources that are currently available.

The goal of our job scheduling system is to make sure that the computing resources are allocated fairly between the people using the systems and that the machines are used as fully as possible.

There are some things you can do to give your jobs a better chance of being run. To do this, you need to understand a little about how the scheduling system works.

## How jobs are scheduled

The SLURM scheduler uses two main methods to decide which jobs are run.

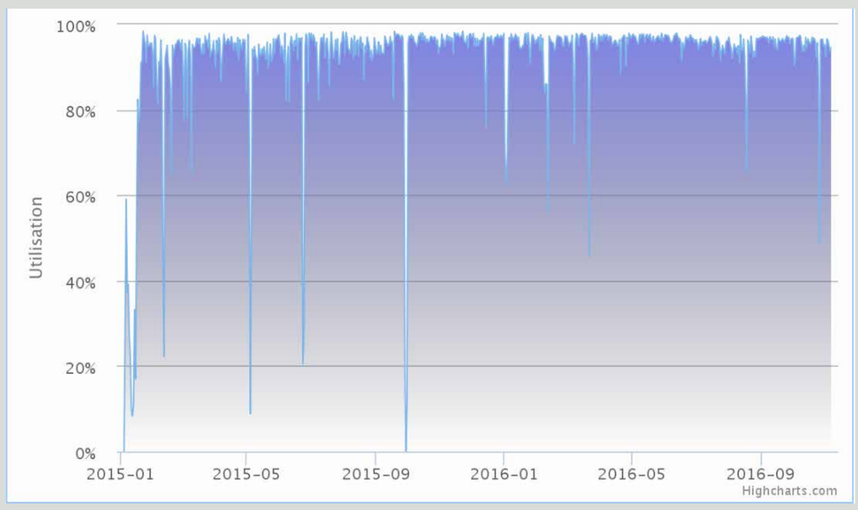
### Fair-share

The goal of the fair share algorithm is to make sure that all projects can use their fair share of the available resources within a reasonable time frame. The priority that a job (belonging to a particular project) is given will depend on how much of that project's time quota has been used recently in relation to the quotas of jobs belonging to other projects - the effect of this on the priority declines gradually with a half-life of 14 days. So jobs submitted by projects that have not used much of their quota recently will be given high priority, and vice versa.

### Backfill

As well as having a main queue to ensure that the systems are as full as possible, the job scheduling system also implements "backfill". If the next job in the queue is large (that is, it will need lots of nodes to run), the scheduler collects nodes as they become free until there are enough to start running the large job. Backfill means that the scheduler looks for smaller jobs that could start on nodes that are free now, and which **would finish** before there are enough nodes free for the large job to start. For backfill to work well, the scheduler needs to know how long jobs will take. So, to take advantage of the possibility of backfill, you should set the maximum time your job needs to run as accurately as possible in your submit scripts.

This graph shows the percentage of the nodes on Beskow that were in use on different dates from early 2015 till late 2016. You can see how the scheduler makes good use of Beskow as nearly all of the available nodes are being used all the time.




Note: All researchers sharing a particular time allocation have the same priority. This means that if other people in your time project have used up lots of the allocated time recently, then any jobs you (or they) submit within that project will be given the same low priority.

## Example of scheduling

Project A (Anders & Anna)

total time allocation awarded to project

time used in the last 30 days by Anders

Project B (Barbro & Björn)

total time allocation awarded to project

no time used in the last 30 days by Barbro or Björn

Now two new jobs (both needing the same amount of time) are submitted by Anna and Björn.

time needed

time needed

Of course both Anna and Björn would like their jobs to be run as soon as possible.

However, in the current situation, the scheduler will give priority to Björn's job as his project (B) has not used as much of its time allocation recently as project A has used of their allocation.

The fact that Anna has not used any time herself does not make any difference as it is the total amount of time recently used by each project that is taken into consideration when deciding which job will be scheduled next.

**Access QR code or visit www.pdc.kth.se for more information.**

**PDC Center for High Performance Computing** SNIC