SLURM

Pre-GDB 11.3.2014

- Ulf Tigerstedt CSC / NeIC / NDGF-T1
- Magnus Jonsson HPC2N / NeIC / NDGF-T1

SLURM

- Originally by LLNL (.gov) but now headed by SchedMD (.com) but active developers from around the world, including all the nordic countries.

SLURM and the nordics

- 100% penetration in scientific clusters in Finland, even used in the Cray supercomputer.
- Sweden is fast moving towards SLURM, and all new clusters will use it.

Multicore jobs

- Slurm supports multicore jobs well and defaults to MPI-type communication between the spawned threads. Non-MPI multicore jobs need to specify that all cores need to be on the same motherboard for them to function properly.

- The scheduling-performance of SLURM is heavily debated. For some it'sthe worst thing ever, for others it's the best thing ever. Generally theadmins with large clusters and multihour multicore jobs are happy. Running really short single core jobs clogs the default scheduler.

Releases

- Release schedule is quite nice: Major revisions almost twice a year with bugfix releases when needed. However, old versions are not updated oncea new major revision has been published.
- Normally quick response to questions/bug reports. Happy for patches both for bug fixes and for improvements.> Not released with RHEL/EPEL. Outdated version available in Ubuntu.
- Usage requires compilation, but it's easy with rpmbuild or direct/custom install with ./configure

Upgrades

- Upgrades can be a hassle, since version 2.x is not compatible with 2.(x+1). Minor version upgrades OK (2.x.y to 2.x.(y+1)).
- Downgrades will probably not work if the accounting database is used, since it will upgrade the schema.

Plugins

- Prolog,Epilog,TaskProlog and TaskProlog offers easy ways to handle administrative tasks for each job.
- For advanced tasks, the SPANK API allows plugins to hook into every part of the process from job submission to job control. This is used at HPC2N to give every job an isolated tmp directory.
- Most part of Slurm is extendible via plugins (logging, accounting, scheduling, ...)

Negative

- No IPv6 support
- All nodes need to have the same config file
- Prefers a shared filesystem
- Mixing of multi and single core jobs can mess up the default fairshare+backfill combination.