Private /tmp for each job using SPANK

Magnus Jonsson
<magnus@hpc2n.umu.se>
http://www.hpc2n.umu.se/
About me

• HPC2N, Umeå University
• NeIC (contracted on 50%)
• First time in HPC in 1999
• Back in HPC from 2012
• Involved in two different accounting projects related to HPC
About HPC2N

- One of six national super computer centre in Sweden
- Located in Umeå
- Provide academic computing to Sweden
About NeIC

- Nordic e-Infrastructure Collaboration
- NeIC is an organisation that facilitates the development and operation of high-quality e-Infrastructure solutions in areas of joint Nordic interest.
- http://neic.nordforsk.org/
Background

- Abisko Cluster
- 332 nodes
- 48 cores / node
- Up to 48 jobs / node
- Cleaning of /tmp after each job
Problem with cleaning /tmp

- Which files to clean?
- Remove all files when there are no more running jobs for a user
- Not removing running jobs files
- Ending up with a full /tmp
Our solution

- Giving each job a private /tmp
- Just setting $TMPDIR does not help from creative users.
Basic concept

• Slurm SPANK
• Linux mount namespace
• mount --bind
Outline...

- mkdir
- mount --make-rshare
- unshare CLONE_NEWNS
- mount --make-rslave
- mount --bind
- rm -rf
Create new /tmp

- $SLURM_TMPDIR = $TMPDIR/slurm.$SLURM_JOB_ID.$SLURM_RESTART_COUNT
- mkdir $SLURM_TMPDIR
Make / shareable

- mount --make-rshare /
- Make / share all mounts to children
Private namespace

• Using the “unshare -m” command create a copy of the mount namespace.
Make / a slave

- mount --make-rslave /
- Make / “listen” to mounts of parent
mount --bind

• mount --bind /tmp
  $SLURM_TMPDIR
After job is completed

- `rm -rf $SLURM_TMPDIR`
- This is done in our epilog script.
Does it work?

• In production on all our clusters for about 1.5 years.
Problems

• Solved
  – NFS mounts outside the clone
  – NFS mounts inside the clone
  – Restartable jobs

• Unsolved
  – autofs unmount
Source code

- http://github.com/hpc2n/spank-private-tmp
- Use it, share it, patch it...
Thanks for me

• Questions?