How to Modify Tempo for Large Scale ICE Installations

By Jim Karellas
Agenda

• Overview of Pleiades
• Pleiades Tempo Configuration
• Tempo Modifications
• Experience After Modifications
• Future Modifications
Overview of Pleiades

- 168 Compute Racks
- 10752 Nodes - 100,352 Cores
- 14 Front End Nodes
- 1 PBS Server
- 1 License Server
- 1 DNS Server
- 1 Performance Collection Server
- 4 Infiniband Subnet Manager Servers (IB0=Compute Fabric, IB1=Storage Fabric)
- 4 Bridge Node Servers (Bridge Columbia and Pleiades)
- 63 Lustre OSS/MDS Servers
- 7 Lustre Scratch Filesystems (Approx 4PB Raw Storage)
- 1 NFS Home Filesystem (Nexus 9000)
Pleiades Tempo Configuration

- Turn off post-discovery creation of the image tar files on RLCs.
- On the admin node, service, and compute nodes, remove pdsh-mod-dshgroup rpm. Use pdsh-mod-genders instead.
- On service nodes with external interfaces, we “chattr +i /etc/hosts”.
- After Tempo clones a compute image, it copies the admin node's /root/.ssh/ and /etc/ssh information into the image. We copy our files back.
We disable many things using the `/etc/opt/sgi/conf.d/exclude` file, most of them because we do equivalent setting changes another way.

<table>
<thead>
<tr>
<th>80-enable-sysrq</th>
<th>80-md5-password-encryption.rhel6</th>
</tr>
</thead>
<tbody>
<tr>
<td>80-fixup-zypp-product.sles10</td>
<td>80-md5-password-encryption.sles</td>
</tr>
<tr>
<td>80-increase-arp-cache-sizes</td>
<td>80-modprobe</td>
</tr>
<tr>
<td>80-increase-ssh-max-startups</td>
<td>80-named-init-fix.rhel5</td>
</tr>
<tr>
<td>80-ipmi-kernel-modules</td>
<td>80-network-kernel-tuning</td>
</tr>
<tr>
<td>80-kdump-diskfull</td>
<td>80-nscd-invalidate-hosts-cache</td>
</tr>
<tr>
<td>80-kudzu.rhel</td>
<td>80-ntp-sysconfig</td>
</tr>
<tr>
<td>80-limits-core-files</td>
<td>80-postfix</td>
</tr>
<tr>
<td>80-limits-mpi</td>
<td>80-serial-console-setup</td>
</tr>
<tr>
<td>80-make-gnome-default</td>
<td>80-service-distro-services</td>
</tr>
</tbody>
</table>
Pleiades Tempo Configuration

• In addition, the first run of bcfg2 after a service node is discovered chkconfig these services off:
  – 00-update-tempo-configs
  – 15-network-setup
  – 20-name-resolution
  – 80-csn-distro-services
Pleiades Tempo Configuration

- When not discovering, we disable sgi-esphttp and sgi_espd xinetd services because they cannot handle the weekly Security port scan and flood the logs with error messages.
Tempo Modifications

- We want to disable most updating that Tempo does. Most of the stuff that Tempo updates becomes very redundant and causes a lot of time to be lost and generates a lot of extra network traffic when dealing with a large system. We modify the discover-rack and update-configs scripts to curb the number of updates.
Tempo Modifications

- /opt/sgi/lib/discover-rack (admin node)

pladmin3 /opt/sgi/lib # diff -c discover-rack discover-rack.orig
*** discover-rack     Tue Dec 14 15:22:13 2010
--- discover-rack.orig  Thu Jan 20 07:02:31 2011
***************
*** 170,181 ****
# unlock($lock_fh);
- # update Tempo configs (does its own locking)
- # NAS mod to limit updates to single rack leader
- $ENV{NAS_LEADER} = $leader;
- run_cmd("cmd_update_configs");
- # end mod
-
# sync with ESP/ log ESP change event (ESP does its own locking)
run_cmd("cmd_esp --setup ice_system --no_rmt_subscr --rack $rack");

--- 170,175 ----
Tempo Modifications

• /opt/sgi/lib/update-configs (admin node)

pladmin3 /opt/sgi/lib # diff -c update-configs update-configs.orig
*** update-configs Wed Dec 15 14:46:19 2010
--- update-configs.orig Thu Jan 20 07:03:25 2011
***************
*** 39,54 ****
   my $update_tempo_configs = "/etc/opt/sgi/conf.d/00-update-tempo-configs";
   my $pdsh_leader_group = "/etc/dsh/group/leader";
   my $pdsh_service_group = "/etc/dsh/group/service";
   - # NAS mods to disable most updating
   - my $pdsh_leaders = "echo pdsh -g leader $update_tempo_configs";
   - if (defined($ENV{NAS_LEADER})) {
   -     $pdsh_leaders = "pdsh -w $ENV{NAS_LEADER} $update_tempo_configs";
   - }
   - my $pdsh_service = "echo pdsh -g service $update_tempo_configs";
   - if (defined($ENV{NAS_SERVICE})) {
   -     $pdsh_service = "pdsh -w $ENV{NAS_SERVICE} $update_tempo_configs";
   - }
   - # End of NAS tweaks
     my $lock;
     my $lock_fh;
     my $update_flags = ";
--- 39,44 ----
Experience During Last Tempo Upgrade

- Total time to upgrade Pleiades: 3 days and change
- 15 of first 50 RLC's had bmc issues. Stopped counting after that.
- CMC issues galore.
- Partial failures on rack with no indication from Tempo that there was a failure. (hosts come up without names, etc..)
- One rack re-imaged itself after crashing and rebooting.
- cimage –push-rack hasn't worked since approximately after 40 racks were installed.
Future Changes

• Definitions
  – tempo_current = current version of tempo
  – tempo_upgrade = version of tempo we are upgrading to
  – admin_current = current admin node
  – admin_upgrade = ad
Future Changes

- On current admin node, clone current slot to upgrade slot:
  - `root@admin`# clone-slot -source 1 -dest 2
- Boot into slot2
- Add repos
  - `root@admin`# yume -prepare -repo /tftpboot/SGI/tempo-upgrade
  - etc..(foundation, OS, propack)
- (may need to run udevadm trigger to create device files)
Future Changes

• Mount slot 2 on the admin node
  – # mkdir /a
  – # mount LABEL=sgiroot2 /a
  – # mount LABEL=sgiboot2 /a/boot

• Mount slot 2 on the RLCs
  – # pdsh -g rlc mkdir /a
  – # pdsh -g mount LABEL=sgiroot2 /a
  – # pdsh -g mount LABEL=sgiboot2 /a/boot
Future Changes – Mount proc, etc...

- Mount /proc, /sys, /dev on admin_current
  - `# mount -o bind /proc /a/proc`
  - `# mount -o bind /sys /a/sys`
  - `# mount -o bind /dev /a/dev`

- Mount /proc, /sys, /dev on RLCs
  - `# pdsh -g rlc mount -o bind /proc /a/proc`
  - `# pdsh -g rlc mount -o bind /sys /a/sys`
  - `# pdsh -g rlc mount -o bind /dev /a/dev`
Future Changes

- Upgrade tempo

- Set the default slot to “slot 2”

- Umount /a/proc, /a/sys, /a/dev and reboot admin node (leave leaders alone for now)

- After system reboots, we are now in the admin_upgrade server. Run the DB upgrade script:
  - `/etc/init.d/sgi-database-update start`
Future Changes

• May need to run the following scripts:
  – /opt/sgi/lib/reset-admin-network
  – /opt/sgi/lib/update-configs
  – /opt/sgi/lib/cluter-configuration

• Add new repos (we did this already, but on the slot 1 admin server)
  – root@admin# crepo --del (foundation, OS, propack)
  – root@admin# crepo --add (foundation-upgrade, OS-upgrade, propack-upgrade)
Future Changes

• Back to leaders
• At this stage all leaders should be up with slot 2 mount at /a
  – pdsh -g rlc chroot /a env PBL_SKIP_BOOT=1 yume -y
    --nopugins --repo http://admin/repo/tftpboot/2.1-
    upgrade/tempo --repo http://admin/repo/tftpboot/2.1-
    upgrade/foundation --repo http://admin/repo/tftpboot/
    2.1-upgrade/sles11sp1 upgrade
• Umount everything
  – pdsh -g rlc umount /a
Final Thoughts

- New patch puts a mysql db on all RLCs.