National Aeronautics and Space Administration



TOSS4 Migration: A Retrospective and Current Status

June 27, 2023

NASA Advanced Supercomputing Division

TOSS4 Migration
(TOSS4 = RHEL-based Tri-Lab Operating System Stack version 4)

- Timeline of TOSS4 Migration
  - Aug 15, 2022: Select early user testing of Rome TOSS4 nodes
  - Aug 25, 2022: Testing of Rome TOSS4 nodes open to all users
  - Sep 09, 2022: All new Rome jobs running TOSS4; existing TOSS3 jobs allowed to complete

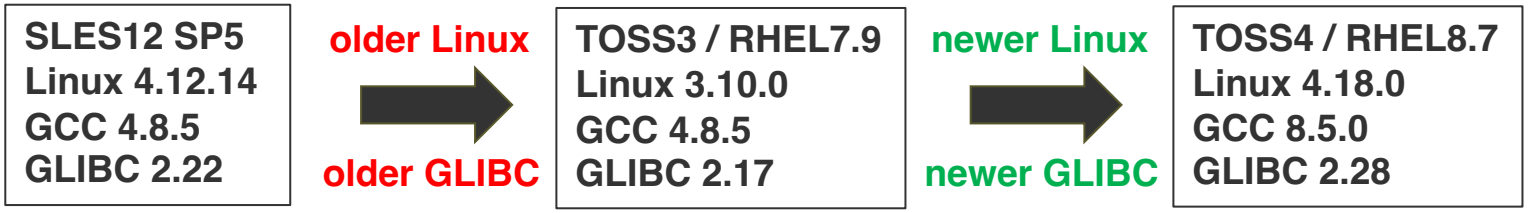
  - Apr 11, 2023: NAS Weekly Digest announcing transition of remaining systems to TOSS4
  - Apr 21, 2023: Announcement of Extended Dedicated times for TOSS4 transition
  - May 05-17, 2023: Dedicated time for Pleiades
  - May 17-31, 2023: Dedicated time for Electra, Aitken Cascade Lake, and GPU nodes
  - May 17, 2023: Pleiades and Endeavour back in production running TOSS4
  - May 31, 2023: Electra and Aitken Cascade Lake back in production running TOSS4

4+ months of testing 

- Timeline of TOSS3 Migration
  - Aug 10, 2021: General Announcement of planned migration to TOSS3
  - Aug 20, 2021: Select early user testing
  - Sep 17, 2021: Testing of TOSS3 open to all users
  - Dec 20, 2021: TOSS3 becomes default OS for new job submissions
  - Jan 25, 2022: SLES12 OS no longer available

Original plan (as of Aug, 2022) was for full migration to TOSS4 by the Fall of 2023, but the deadline was accelerated to May 31, 2023 due to security requirements. There was no time allotted for user testing of TOSS3 vs. TOSS4 on Pleiades, Electra, or Aitken Cascade Lake.

TOSS4 Migration: "Expectation" vs. Reality



- Backward compatibility means most applications built on TOSS3 should run on TOSS4 without recompiling
- Migration of Rome nodes from TOSS3 to TOSS4 was relatively seamless -- main issue was getting a TOSS4 node with internet access for some complicated builds

- **Staff testing of TOSS4 on limited set of Ivy Bridge nodes April 12 - 28, 2023**
- **April 12, 2023 surprise:**
  - Only HPE MPT 2.23 (and a test version of OpenMPI 4.1.4) was capable of running across Ivy Bridge nodes.
  - None of the newer HPE MPT versions could run across nodes
  - Many versions of MPT work fine on a single node, but once IB is involved, the job would error out with:

  MPT ERROR: IB UD Thread Domain create failed

- **April 25, 2023: Obtained a modified MPT 2.28 from HPE MPT developers that works across Ivy Bridge nodes**

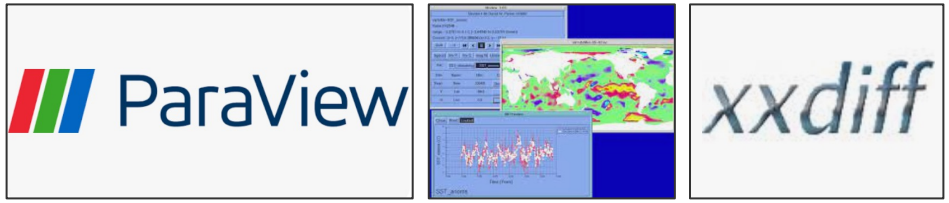
TOSS4 Migration: Working versions of HPE MPT

Sandy Bridge, Ivy Bridge, Haswell, Broadwell MOFED 4.96 IB HCA: ConnectX-3	Skylake, Cascade Lake MOFED 4.96 IB HCA: ConnectX-5	Rome MOFED 5.81 IB HCA: ConnectX-6
2.23 (or earlier)	2.23	
		2.25
	2.26	2.26
	2.27_toss4	2.27_toss4
	2.28_toss4	2.28_toss4
2.28_25Apr23_rhel87	2.28_25Apr23_rhel87	2.28_25Apr23_rhel87

When loading HPE MPT, use the version-less mpi-hpe/mpt module. This will ensure you always load the currently recommended version.

TOSS4 Migration: Updated Modules

- Some GUI-based applications needed to be rebuilt to work over X11 or VNC:
  - Paraview 5.11.1 has been built on TOSS4 (module: paraview/5.11.1)
  - Paraview 5.9.1 has been re-built on TOSS4 (module: paraview/5.9.1)
  - ncview 2.1.9 has been built on TOSS4 (module: ncview/2.1.9)
  - xxdiff 5.1 has been built on TOSS4 (module: xxdiff/5.1)



- Working versions of Intel MPI are under /nasa/modulefiles/testing:

Intel Nodes:		AMD Nodes:
mpi-intel/2019.5.281		mpi-intel/2019.5.281
mpi-intel/2020.0.166	patched May 23	
mpi-intel/2021.3.0	patched May 24	mpi-intel/2021.3.0
mpi-intel/2021.7.0	patched May 24	mpi-intel/2021.7.0

To tell the module command to see modulefiles in the testing area:

```
pfe27:~> module use -a /nasa/modulefiles/testing
```

TOSS4 Migration: Best Practices



- Use the version-less mpi-hpe/mpt. It will load the currently recommended version of MPT.
- Do not add the pkgsrc lib directory to your LD_LIBRARY_PATH except, possibly, in your PBS script to change the runtime LD_LIBRARY_PATH. Several Linux commands (i.e., ssh) will break if they use the libs from pkgsrc.
- If the updated TOSS4 GCC compiler is unable to compile your source code, consider using an older GCC compiler (**module load gcc/6.5**) or switching to the Intel compiler. If using the mpicc/mpicxx compiler wrappers from HPE MPT, need to set MPICC_CC=icc and MPICXX_CXX=icpc, otherwise those wrappers will use the GCC compilers.
- When using the Intel compilers, if there are problems accessing a license or there is a lag in accessing the license, consider using a version 2021 or newer Intel compiler. Those newer Intel compilers no longer require a license.
- For bash users, make sure that you're using the right module command:

```
pfe27:~> which module
module ()
{
    _module_raw "$@" 2>&1
}
```

Best Practices: Running X11 Apps

If X11 apps from pfe do not display on your local workstation, add both "ForwardX11 yes" and "ForwardX11Trusted yes" to the sfe and pfe blocks in your ~/.ssh/config file (on your local workstation)

Most Common Error
Could not find ':' in DISPLAY:
Error: Can't open display: localhost:45.0

```
Host sfe sfe?.nas.nasa.gov
   ForwardX11                yes
   ForwardX11Trusted         yes
   ServerAliveInterval       5m

# Add additional hosts to the list below as needed
Host *.nas.nasa.gov lou lou? lfe? pfe?? ?fe? *-last
   ForwardX11                yes
   ForwardX11Trusted         yes
   HostbasedAuthentication   no
   Protocol                  2
   ProxyCommand              ssh -ax -oCompression=no sfe ssh-proxy %h
   ServerAliveInterval       5m
   # Uncomment and replace <NAS_login_name> with your NAS username
   # if different than local username
   User                      your_NAS_username
```

~/.ssh/config

If X11 apps still do not show up, make sure you are doing "ssh pfe" from your local workstation, and not manually doing "ssh sfe" and then "ssh pfe" from the sfe. The latter will not enable X11 apps. It is recommended to set up SSH Passthrough as it streamlines this process (and the entire login sequence only requires your PASSCODE, once).

Best Practices: pkgsrc

- If "module avail" does not list a package that you need, check pkgsrc
- Searching for packages and libraries in pkgsrc:




```
pfe27:~> module avail pkgsrc
------------------------ /nasa/modulefiles/pkgsrc/toss4 -----------
pkgsrc/2022Q1-rome

pfe27:~> module help pkgsrc/2022Q1-rome
----------------------------------------------------------------------
Module Specific Help for /nasa/modulefiles/pkgsrc/toss4/pkgsrc/2022Q1-rome:

The pkgsrc module is a collection of software built from
NetBSD's pkgsrc system. Some of the key packages included are:

  apache-maven-3.8.3 arpack-ng-3.8.0    bioperl-1.7.7      boost-1.78.0
  cgal-4.14          clang-13.0.1       cmake-3.22.3       cppcheck-2.1
<... rest of key packages omitted ...>

For a complete list of the packages included with this module, run:
/nasa/pkgsrc/toss4/2022Q1-rome/sbin/pkg_info
```

Best Practices: pkgsrc (cont.)

- Pkgsrc contains an extensive list of packages:

```
pfe27:~> /nasa/pkgsrc/toss4/2022Q1-rome/sbin/pkg_info | wc -l
1660
```

- Grep for specific packages from the pkgsrc_info command:

```
pfe27:~> /nasa/pkgsrc/toss4/2022Q1-rome/sbin/pkg_info | grep -i cairo
cairo-1.16.0nb4     Vector graphics library with cross-device output support
cairo-gobject-1.16.0nb6 Vector graphics library with cross-device output
support
py39-cairo-1.20.1nb1 Python bindings for cairo
```

- When linking to libs from pkgsrc, use the -Wl,-R option to add a directory to the runtime library search path. For example, to link to the cairo library:

```
-L/nasa/pkgsrc/toss4/2022Q1-rome/lib -lcairo \
  -Wl,-R/nasa/pkgsrc/toss4/2022Q1-rome/lib
```

Despite ending with -rome, the pkgsrc/2022Q1-rome module can be used with both Intel and AMD node types.

TOSS4 Migration: Issues We Are Tracking …

1. Newer filesystems {/nobackupp27, /nobackupp28, /nobackupnfs1} have problems mounting on ~10-30% (??) of the nodes
2. Stuck sync processes on Broadwell nodes {bro and bro_ele} causing jobs to fail PBS prologue and go into Hold state
3. SEC OOM detection mechanism not functional and cannot notify users when their jobs run nodes out of memory
4. pkgsrc module needs to be updated
5. Instability of Lou front-ends
6. LDAN nodes not back to production
7. Pfes often slow or unresponsive
8. App that used to run on TOSS3 front-ends won't run on TOSS4 front-ends but can run on a TOSS4 compute node. Workaround: use a reserved front-end

Please do not hesitate to reach out or submit a ticket to support@nas.nasa.gov with other open issues you may encounter on NAS systems.

Questions?

Suggestions for future webinar topics are welcome