Los Alamos
NATIONAL LABORATORY
EST.1943

LA-UR-16-26965
Approved for public release; distribution is unlimited.

Title: Using OFI libfabric on Cori/Edison

Author(s): Pritchard, Howard Porter Jr.

Intended for: presentation during NERSC site visit

Issued: 2016-09-12

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Using OFI libfabric on Cori/Edison

Howard Pritchard (LANL)
Sung-Eun Choi (Cray)
August 22, 2016

LA-UR-XX-YYYYY

What we'll cover

- Building/installing libfabric
- OpenMPI using libfabric
- MPICH using libfabric
- Open SHMEM (Sandia)
- Other applications
- What's Next?

Building libfabric for Cori/Edison

- Use the GNU compiler (PrgEnv-gnu) - the more recent the better.
- build instructions on the libfabric-cray Github wiki: https://github.com/ofi-cray/libfabric-cray/wiki/GNI-provider-building-it
- Until Edison CLE is upgraded, need to add the following to your .bashrc.ext:

```
export PKG_CONFIG_PATH=~hpp/opt/cray/gni-headers/default/lib64/pkgconfig:${PKG_CONFIG_PATH}
export PKG_CONFIG_PATH=~hpp/opt/cray/ugni/default/lib64/pkgconfig:${PKG_CONFIG_PATH}
export LD_LIBRARY_PATH=~hpp/opt/cray/ugni/default/lib64:${LD_LIBRARY_PATH}
```

- release tarballs are at https://github.com/ofiwg/libfabric/releases

Reporting Issues with libfabric for Cori/Edison

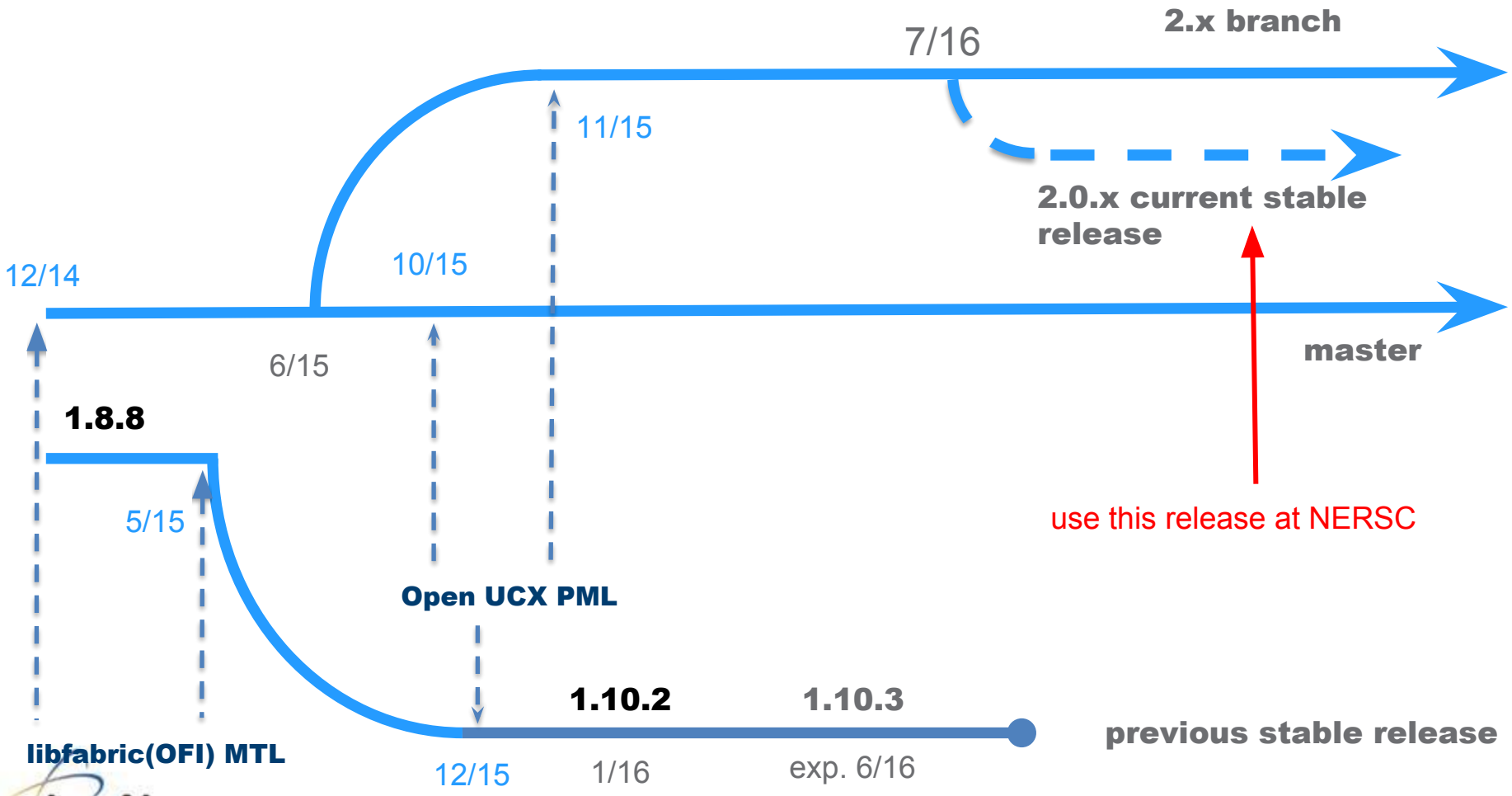
- Good place to start is the libfabric-users mail list
- If the problem appears to be GNI provider specific, open an issue on the libfabric-cray Github repo: https://github.com/ofi-cray/libfabric-cray/issues

Open MPI

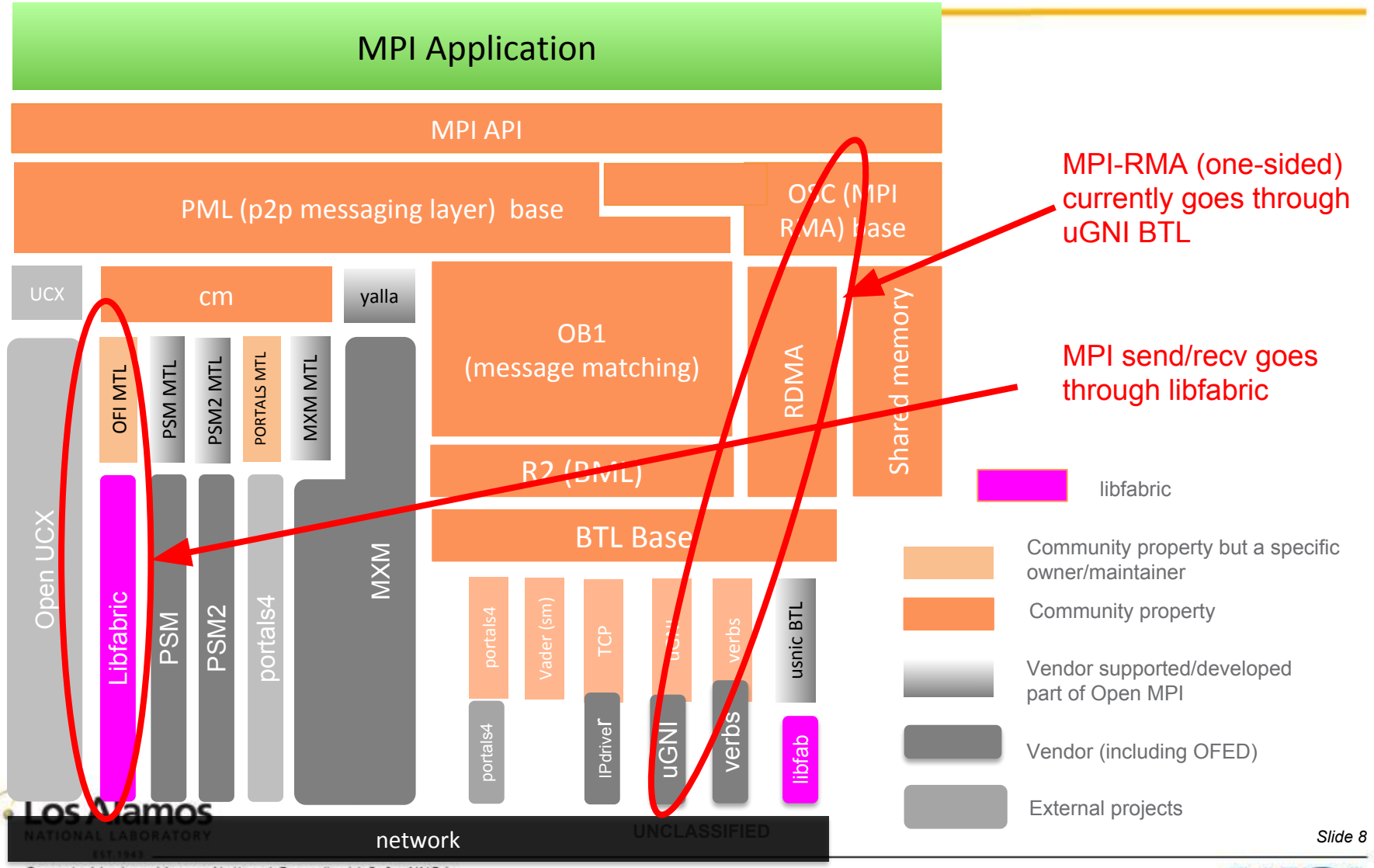
Open MPI and Libfabric

- Intel introduced an OFI (libfabric) MTL in Open MPI 1.10. Known to work with following providers -
  - Intel Truescale and Omnipath (PSM1/2)
  - Cray XC (GNI)
  - Sockets provider (OS-X, etc.)
- Cisco also introduced an OFI BTL targeting Cisco usNIC ethernet cards in Open MPI 1.10
- Both the OFI MTL and BTL have been carried forward in to the 2.X Open MPI release stream
- Current Open MPI release is 2.0.0 (2.0.1 releasing today!)
- The use of the Cray XC HSN can be confusing...

Open MPI Release Timeline



Open MPI Use of Cray uGNI



Los Alamos NATIONAL LABORATORY EST. 1943
Operated by Los Alamos National Security, LLC for NNSA
NNSA

Building Open MPI to use libfabric for Cori/Edison

- Use the GNU compiler (PrgEnv-gnu) - the more recent the better.
  - ➢ Has been built with Intel 16 compilers
- Has only been tested using Cray PMI (automatically detected by *configure* for Open MPI 2.x or master)
- build instructions on the libfabric-cray Github wiki: https://github.com/ofi-cray/libfabric-cray/wiki/Building-and-Running-OpenMPI
- Especially for KNL nodes, want to disable other Open MPI components using uGNI:

```
export OMPI_MCA_btl=self,vader,tcp
```

may not need with Open MPI 2.0.1 or higher

MPICH

What's going on with MPICH

- Two versions of MPICH to consider
  - older CH3 device - more mature but a dead-end for libfabric
  - CH4 device - what's going to be deployed on ANL Aurora system
- MPICH CH3 using libfabrics has been tested with DOE mini-apps on Cray XC KNL systems
  - Reasonable performance for applications
  - However, testing has revealed scaling limitations of the CH3 device using libfabric - (MPI_Finalize)
- Just starting to use CH4 device using libfabrics

Building MPICH CH3 to use libfabric on Edison/Cori

- can be patched to use Cray PMI
- can use the SLURM PMI without the patch
- build instructions are at https://github.com/ofi-cray/libfabric-cray/wiki/Building-and-Running-MPICH-(CH3)

Building MPICH CH4 to use libfabric on Edison/Cori

- No Cray PMI patch, have to use SLURM PMI
- build instructions are at https://github.com/ofi-cray/libfabric-cray/wiki/Building-and-Running-MPICH-(CH4)
- Very experimental at this point - GNI provider needed fixes that haven't been pushed upstream yet
- We'd like to have MPICH CH4 functional with the GNI provider by libfabric release 1.4

Open SHMEM

Sandia Open SHMEM (SOS)

- Open source version of Open SHMEM
  - supports the 1.3 Open SHMEM standard
- Designed to give good performance, while also being simple enough to use for research tool
- https://github.com/Sandia-OpenSHMEM/SOS

Building Sandia Open SHMEM

- Need special config options for Cray XC
  - disable scalable memory registration
  - enable XPMEM or enable hard polling
- Can be used with both SLURM and Cray PMI
- SOS has a nice build check option
  - make check TEST_RUNNER=”srun -n 2”
  - configure option to test longer running performance tests
- Details for build instructions are on SOS wiki - https://github.com/Sandia-OpenSHMEM/SOS/wiki/OFI-Build-Instructions-Cray-XC

Other Applications and libfabrics

- GASNet
  - Doesn't work with GNI provider (yet)
  - Working on this
- Legion
  - Work being done at LANL to port to libfabric
- Mercury RPC package
  - https://github.com/mercury-hpc/mercury
  - adding a libfabric NA as part of Aurora

Next Steps

Next Steps

- Get a libfabric 1.4 out
- It would be nice to have libfabric 1.4 installed in system space on Cori
- Add libfabric modules on Cori
- Might be a good idea to have a SLURM PMI module to simplify its use when building/using Open MPI and MPICH built to use libfabric
- Upgrade Edison to CLE 5.2 UP04 or newer

Questions