12th ANNUAL WORKSHOP 2016

OPEN MPI AND RECENT TRENDS IN NETWORK APIS

#OFADevWorkshop

HOWARD PRITCHARD (HOWARDP@LANL.GOV)
LOS ALAMOS NATIONAL LAB

LA-UR-16-22559

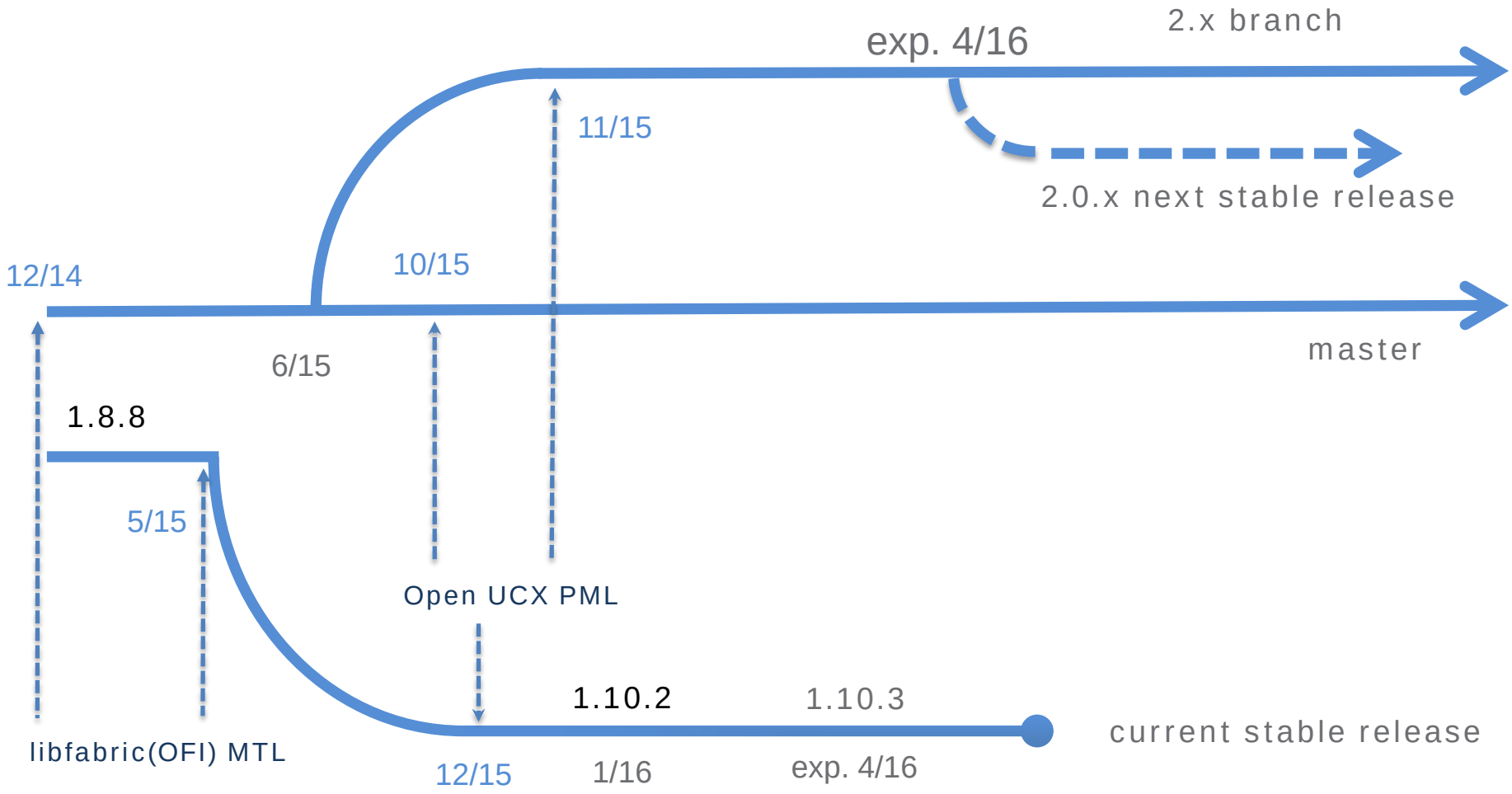
OUTLINE

- **Open MPI background and release timeline**
- **Open MPI Internals and Network APIs**
- **Lessons learned**
- **Advertisement**

OPEN MPI BACKGROUND

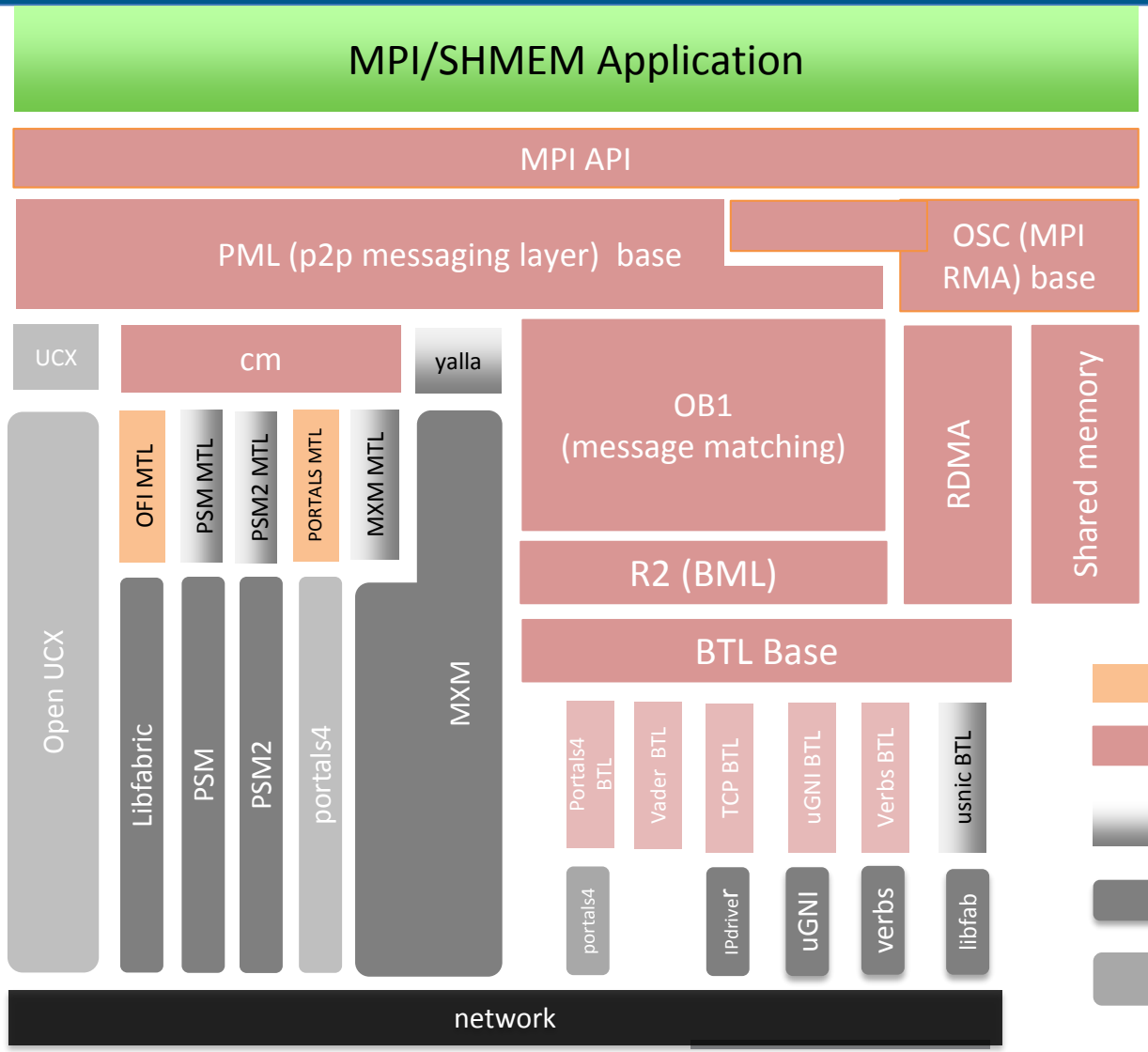
- **Open source implementation of MPI on github: https://github.com/open-mpi/ompi**
- **Developer community includes industry partners, government labs, and universities**
- **Lots of users. Widely used on IB clusters.**
- **Check the github wiki page for info on contributing to the project**

OPEN MPI RELEASE TIMELINE



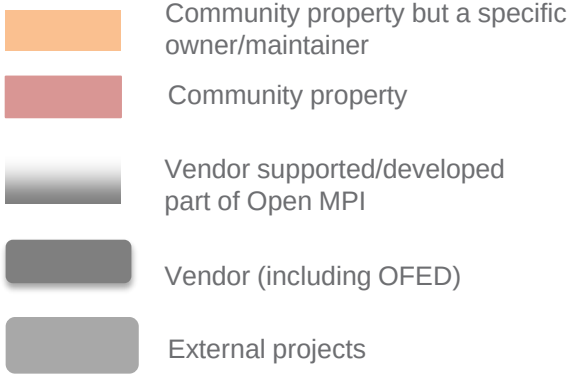
OPEN MPI INTERNALS AND NETWORK APIS

OPEN MPI STRUCTURE

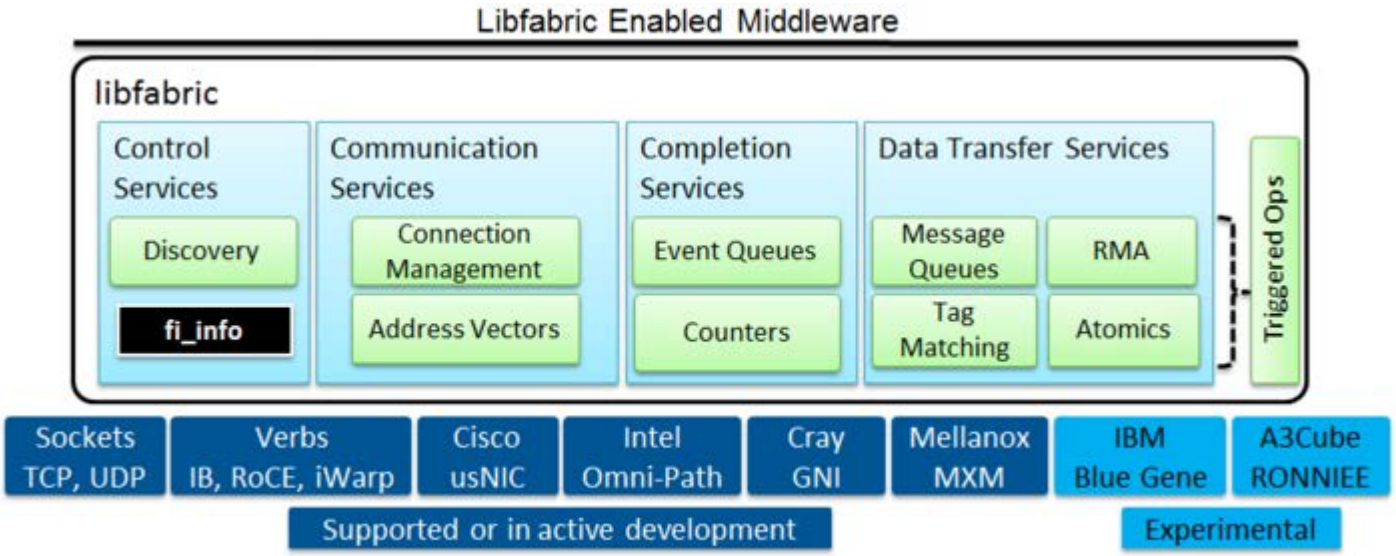


ORTE component not shown

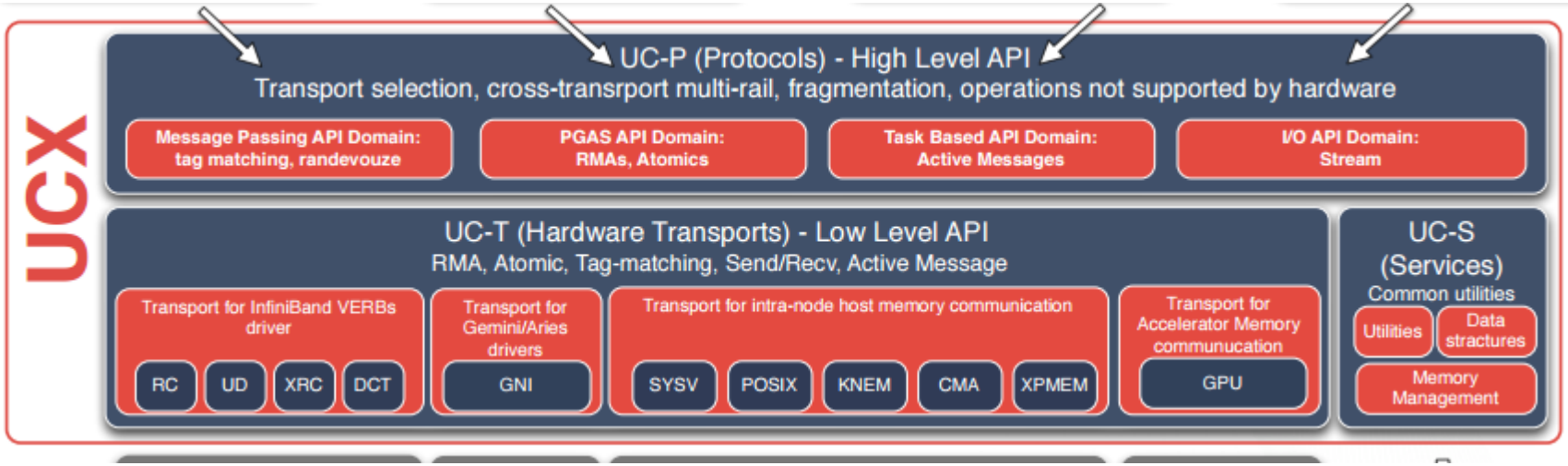
OSHMEM component not shown



LIBFABRIC IN BOXES



OPEN UCX IN BOXES



Accelerated verbs for MLX5

LESSONS LEARNED

FOR NETWORK API DEVELOPERS

- **Great for these folks – lots more tests, even if the API is only exercised through certain paths**
- **Helps a lot if the developer can toggle between different underlying network transports – helps discriminate between bugs in a provider (libfabric speak) vs bugs in the Open MPI usage of the API**

FOR OPEN MPI DEVELOPERS…

- **There have been some issues which needed solving (memory hooks interfering), there probably will be others. This may become a potential issue if the interference is due to a vendor proprietary component.**
- **Useful for providing feedback to Network APIs – new features, etc. that would be great to have**

FOR END USERS HOWEVER…

- **Unlike previous network APIs like PSM, verbs, etc. multi-network software like libfabric and Open UCX can be confusing to users and administrators**
- **Which network APIs got built in to my Open MPI install?**
- **Which lower level API(s) (provider in case of libfabric, TL in case of Open UCX) am I using?**

MAKE CONFIGURY OUTPUT CLEARER

```
Open MPI configuration:
-----------------------
Version: 3.0.0a1
Build MPI C bindings: yes
Build MPI C++ bindings (deprecated): no
Build MPI Fortran bindings: mpif.h, use mpi
MPI Build Java bindings (experimental): no
Build Open SHMEM support: yes
Debug build: no
Platform file: (none)

Miscellaneous
-----------------------
CUDA support: no

Transports
-----------------------
Cray uGNI (Gemini/Aries): no
Intel Omnipath (PSM2): no
Intel SCIF: no
Intel TrueScale (PSM): yes
Mellanox MXM: no
Open UCX: no
OpenFabrics Libfabric: yes
OpenFabrics Verbs: yes
Portals4: no
Shared memory/copy in+copy out: yes
Shared memory/Linux CMA: no
Shared memory/Linux KNEM: no
Shared memory/XPMEM: no
TCP: yes

Resource Managers
-----------------------
Cray Alps: no
Grid Engine: no
LSF: no
Slurm: yes
ssh/rsh: yes
Torque: no
```

This feature is not in 1.10.2 and 2.0.0 releases.

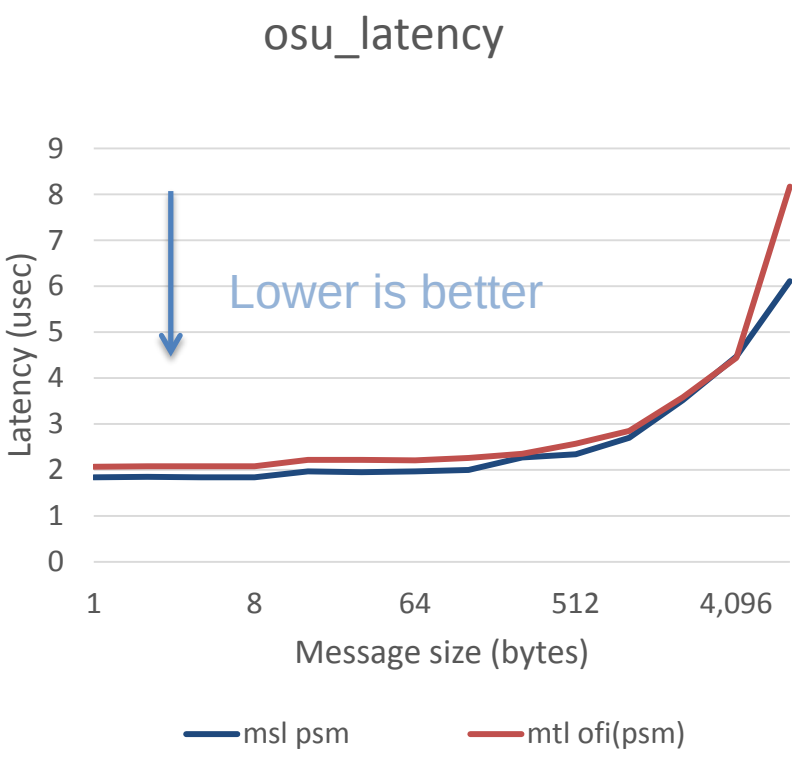
WHICH NETWORK API OPTION IS BEST FOR MY CLUSTER?

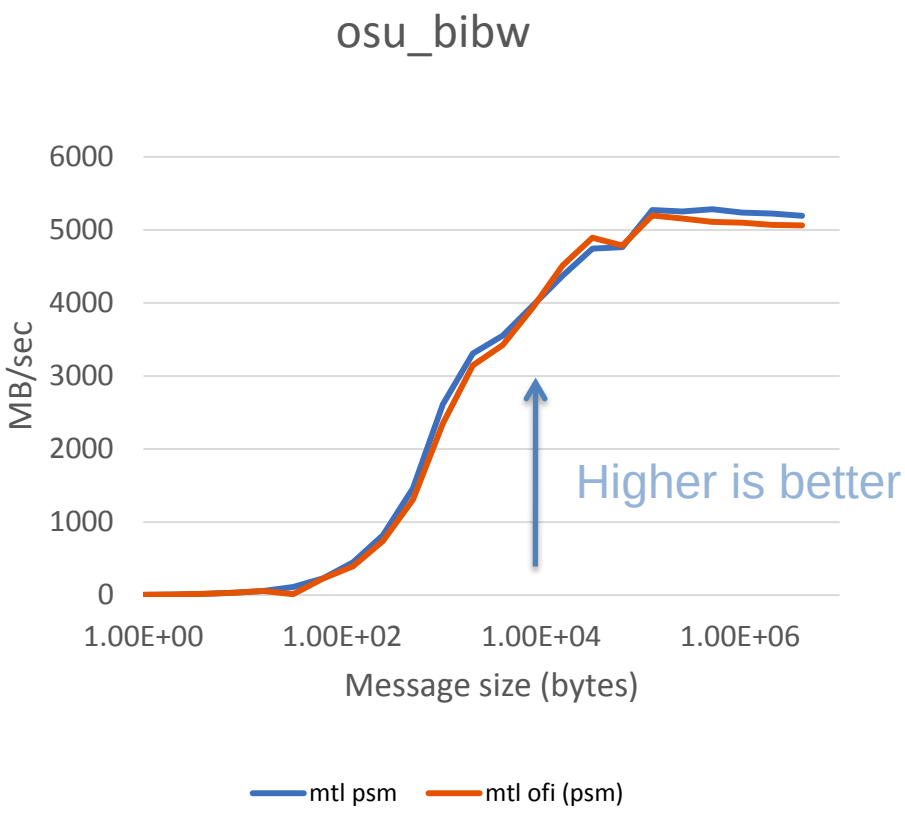
- **Not easy to tell from project descriptions**
- **Try some simple MPI benchmarks with your install.**

BASIC BENCHMARK RESULTS

- **Used two clusters at LANL**
  - An AMD Opteron/Mellanox Connect X2 (pci-e gen2)
  - An Intel Sandy-Bridge/Truescale (pci-e gen2) system
- **Open MPI (master, aka 3.0.0a1), no special config options (if possible), except those needed to pick up libfabric and Open UCX installs**
- **Considerable –mca parameter specification was required to pick up the right PML and/or MTL. This is not bad.**
- **Used libfabric master@13f841c and Open UCX master@4103c00**

INTEL/TRUESCALE CLUSTER

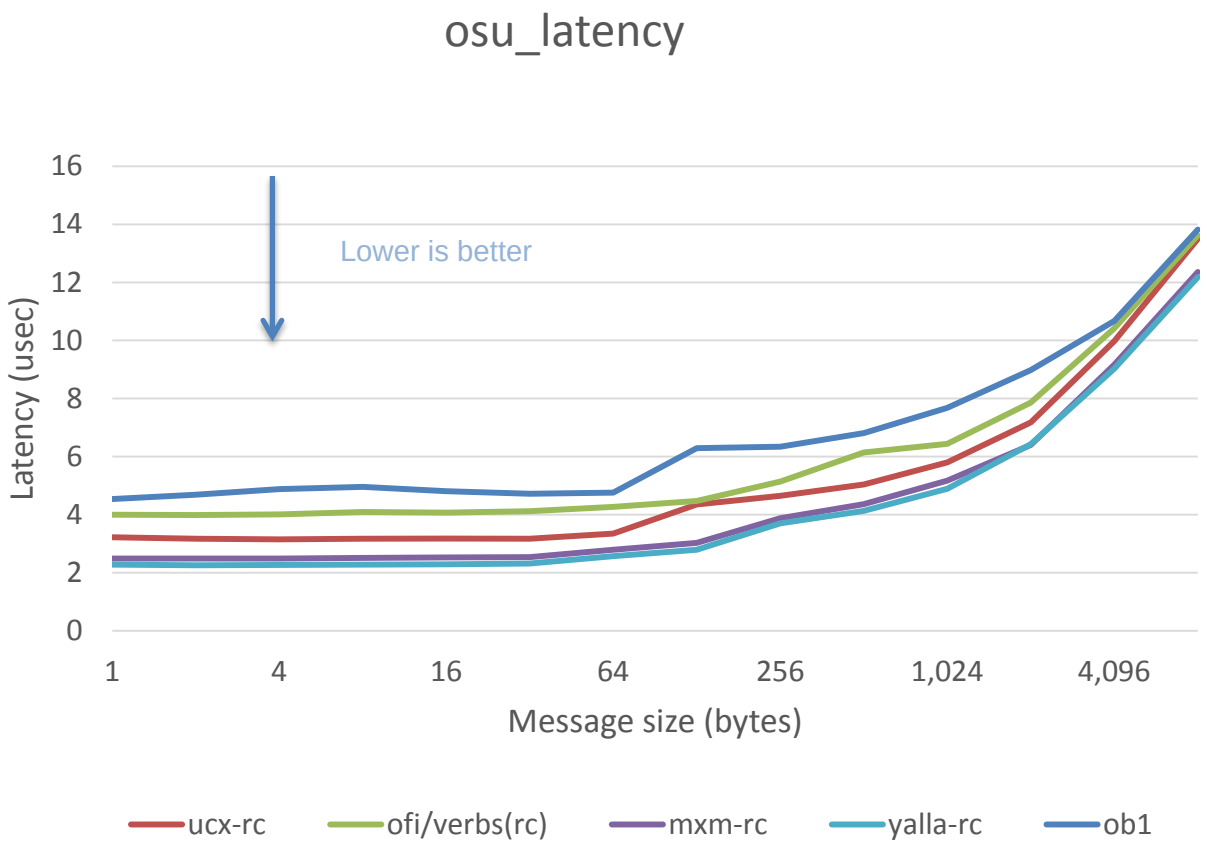




UCX needs work to support max inline data 0.

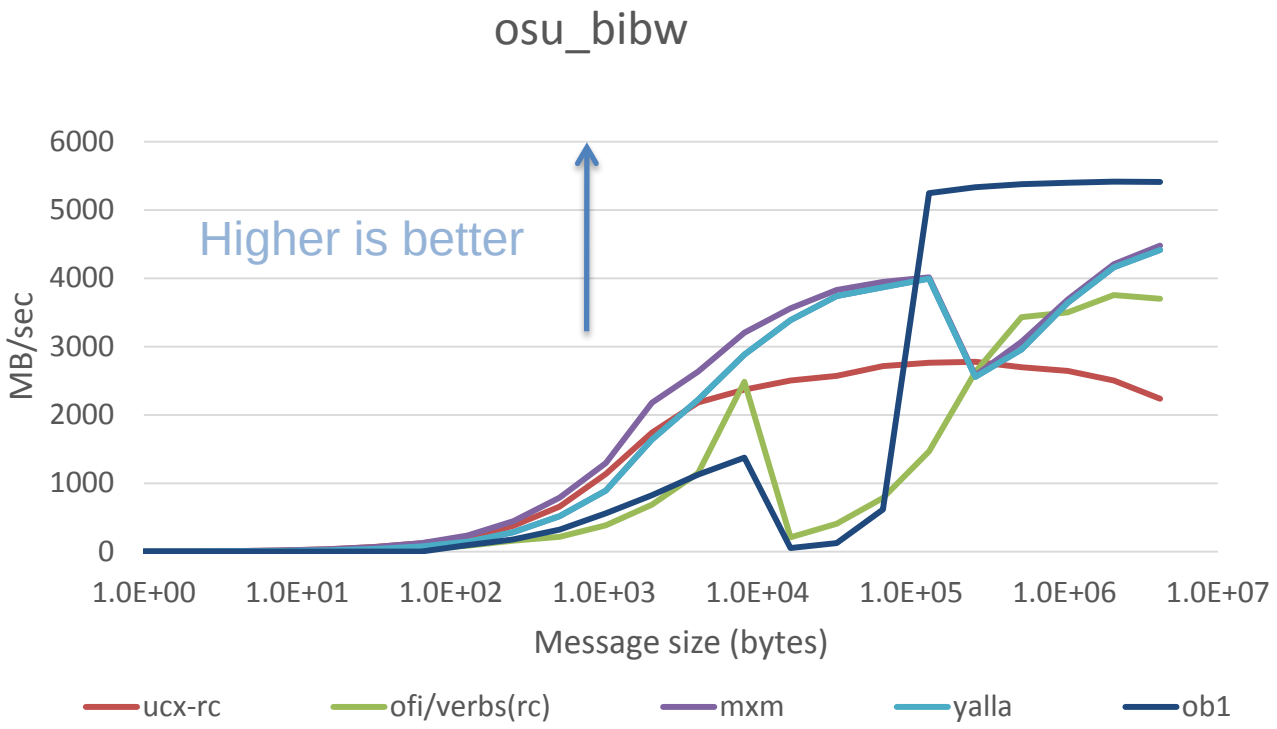
Performance of libfabric verbs provider not good for large messages – not unexpected though.

OPTERON/MLNX CLUSTER



ofi/verbs required fixes to work

OPTERON/MLNX CLUSTER



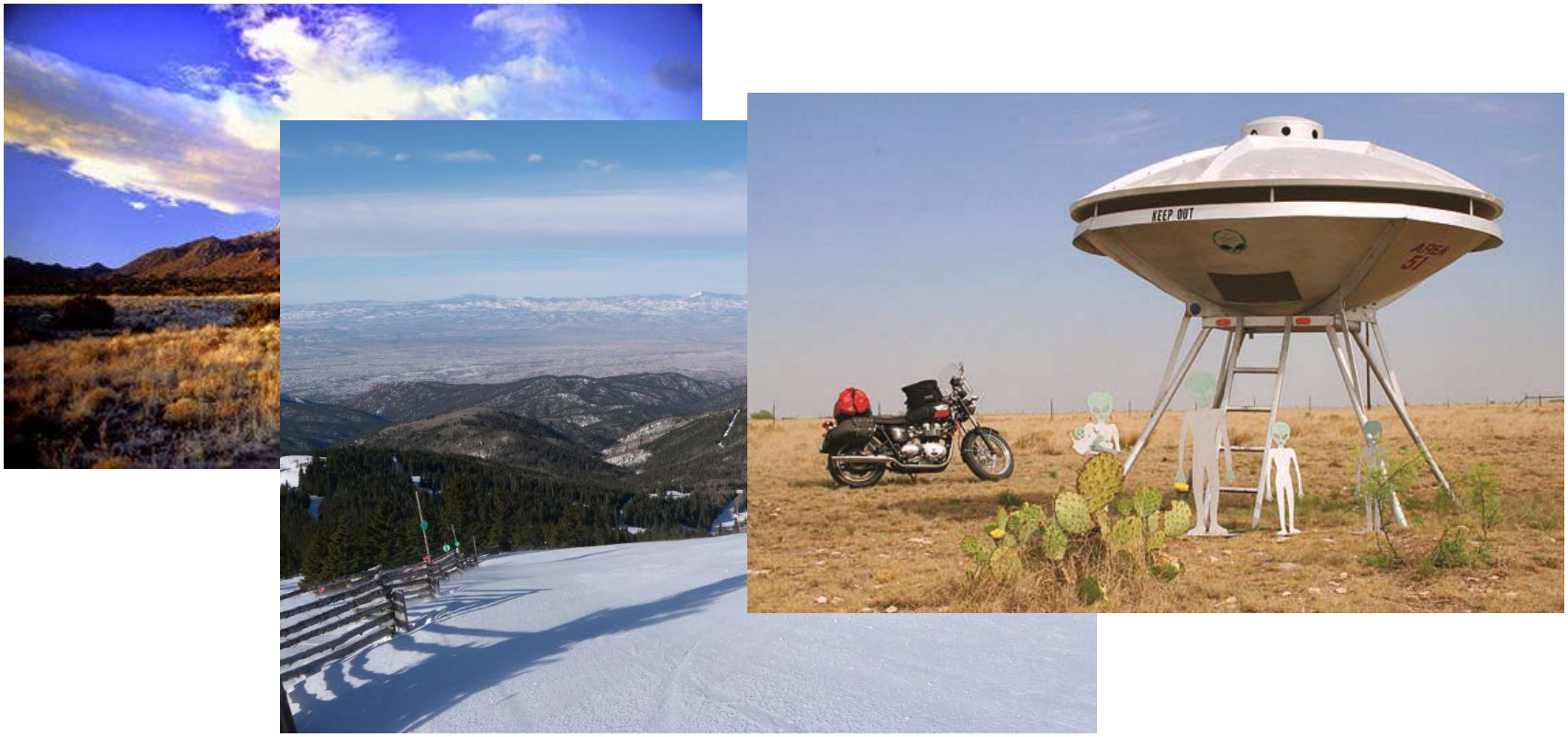
Currently can't coax UCX to use a zcopy path

ofi/verbs not able to use application's memory registrations.

A FEW CONCLUSIONS

- **Both libfabric and Open UCX are works in progress.**
- **In areas with vendor focus, things look promising**
- **For older interconnects, or those that may not be main focus of vendors, not so clear.**
- **Open source solutions help reduce risk of interoperability issues with other components of Open MPI**

LANL/NMC LOOKING FOR STUDENTS AND POST DOCS



http://www.lanl.gov/careers/career-options/student-internships/index.php
https://newmexicoconsortium.org/home/jobs

12th ANNUAL WORKSHOP 2016

THANK YOU