HPC Network Stack Update

Pavel Shamis/Pasha, Principal Research Engineer

MVAPICH User Group Meeting 2018
Columbus, OH

RDMA Update

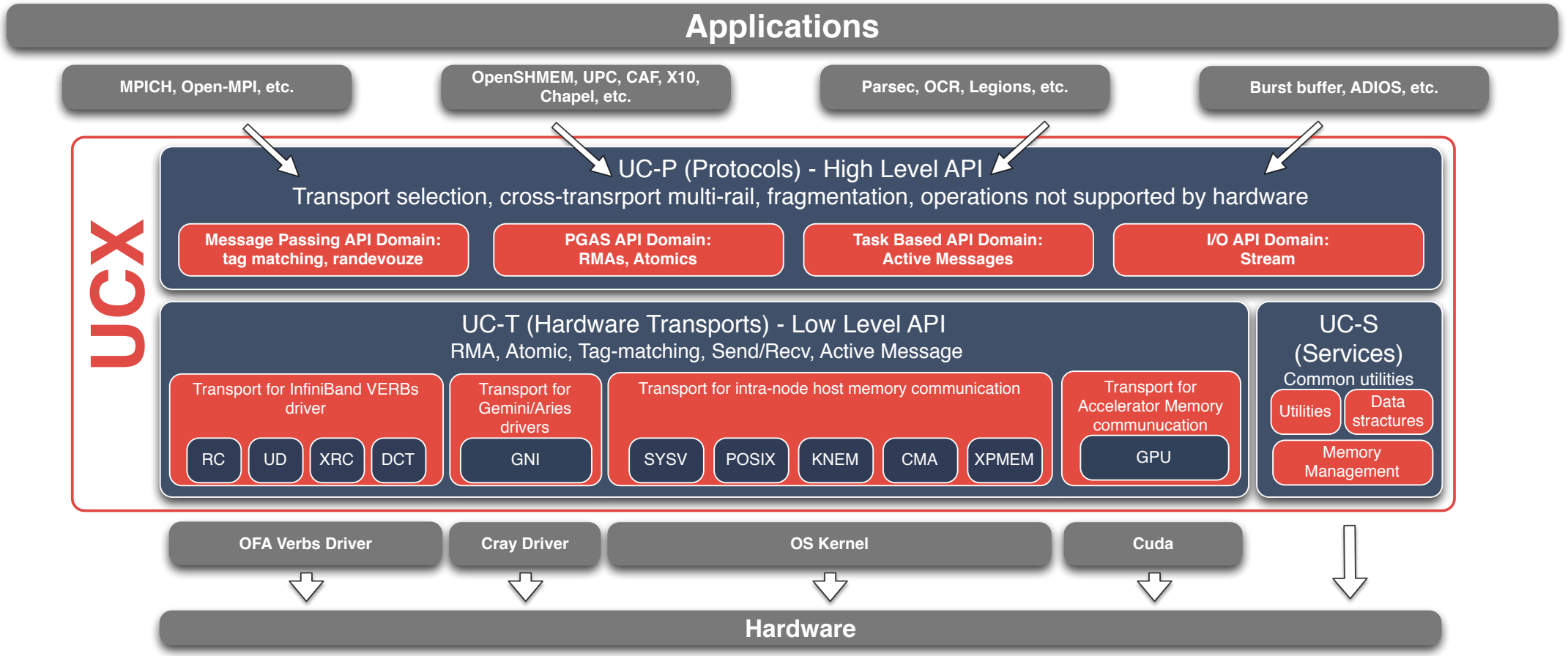
arm

VERBs API on Arm

- Besides bug fixes not much work was required
- Mellanox OFED 2.4 and above supports Arm
- Linux Kernel 4.5.0 and above (maybe even earlier)
- Linux Distribution Support – on going process
- OFED – no official ARMv8 support

arm

OpenUCX



WWW.OPENUCX.ORG

https://github.com/openucx/ucx

arm

UCX 1.3: https://github.com/openucx/ucx/releases/tag/v1.3.0

- **Multi-rail** support for eager and rendezvous protocols
- Added **stream-based communication** API
- Added support for **GPU** platforms: Nvidia CUDA and AMD ROCM software stacks
- Added API for **Client-Server** based connection establishment
- Added support for **TCP** transport (Send/Receive semantics)
- Support for InfiniBand **hardware tag-matching** for DC and accelerated transports
- Added support for **tag-matching communications with CUDA** buffers
- Initial support for **Java bindings**
- **Progress engine** optimizations
- Improved scalability of **software tag-matching** by using a hash table
- Added transparent **huge-pages** allocator
- Added **non-blocking flush and disconnect** semantics
- Added registration cache for **KNEM**

arm

UCX Roadmap

v1.4 – end of ~~July~~ August

- Bitwise atomics support
- Improvements for message injection (medium message size)
- Client/server connection establishment to support large address
- Support multiple connections between same pair of endpoints
- CUDA-IPC support

v1.5 – November 2018

- Bugfixes and optimizations
- Active Message API
- New Client-Server API
- Full functionality over TCP
- Full functionality over legacy RDMA devices
- Full functionality over uGNI API

V2.0 – 2019

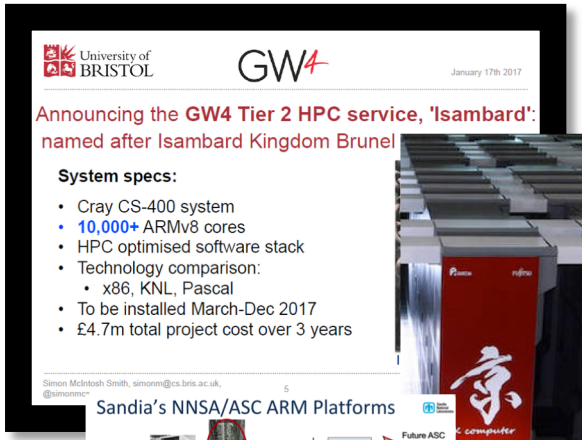
- Updated API – not backward compatible with 1.x
- Binary distribution will provide v1.x version of the library (in addition for 2.x) for backward compatibility
  - All codes should work as it is

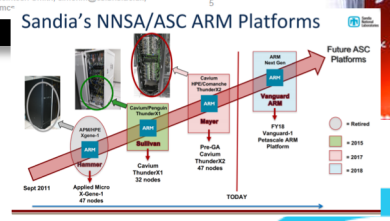
arm

MPI

arm

Scaling





- **HPE Apollo 70:** This dense, scalable platform – and HPE's first ARM-based HPC system – brings more choice and flexibility to HPC customers. It provides easy access to HPC technology with support for standard HPE provisioning, cluster management and performance software. The Apollo 70, using Cavium's 64-bit ARMv8-A ThunderX2™ Server Processor, is purpose-built for memory intensive HPC workloads and delivers up to 33 percent more memory bandwidth than today's industry standard servers. The Apollo 70 also provides access to HPE's partnership ecosystem delivering key HPC components including Red Hat® Enterprise Linux®, SUSE® Linux Enterprise Server for ARM, and Mellanox® high-speed InfiniBand & Ethernet fabric solutions.

GCC

nag®

OpenMP

MPI
OpenMP

MPI
OpenMP

MPI

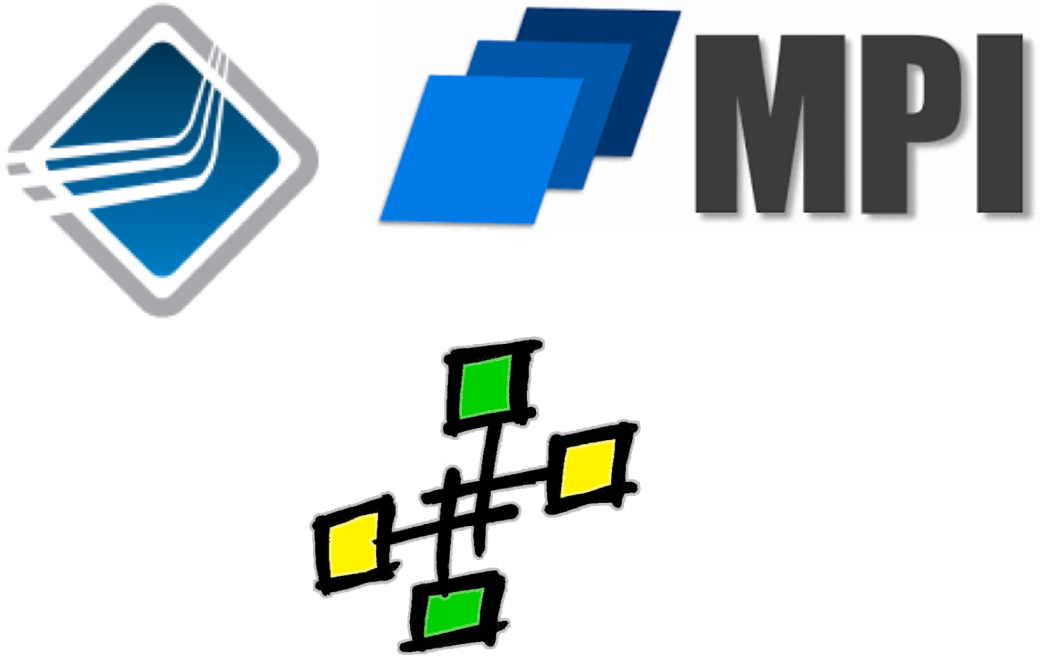
Programing models

MVAPICH 2.3 – works on ARMv8

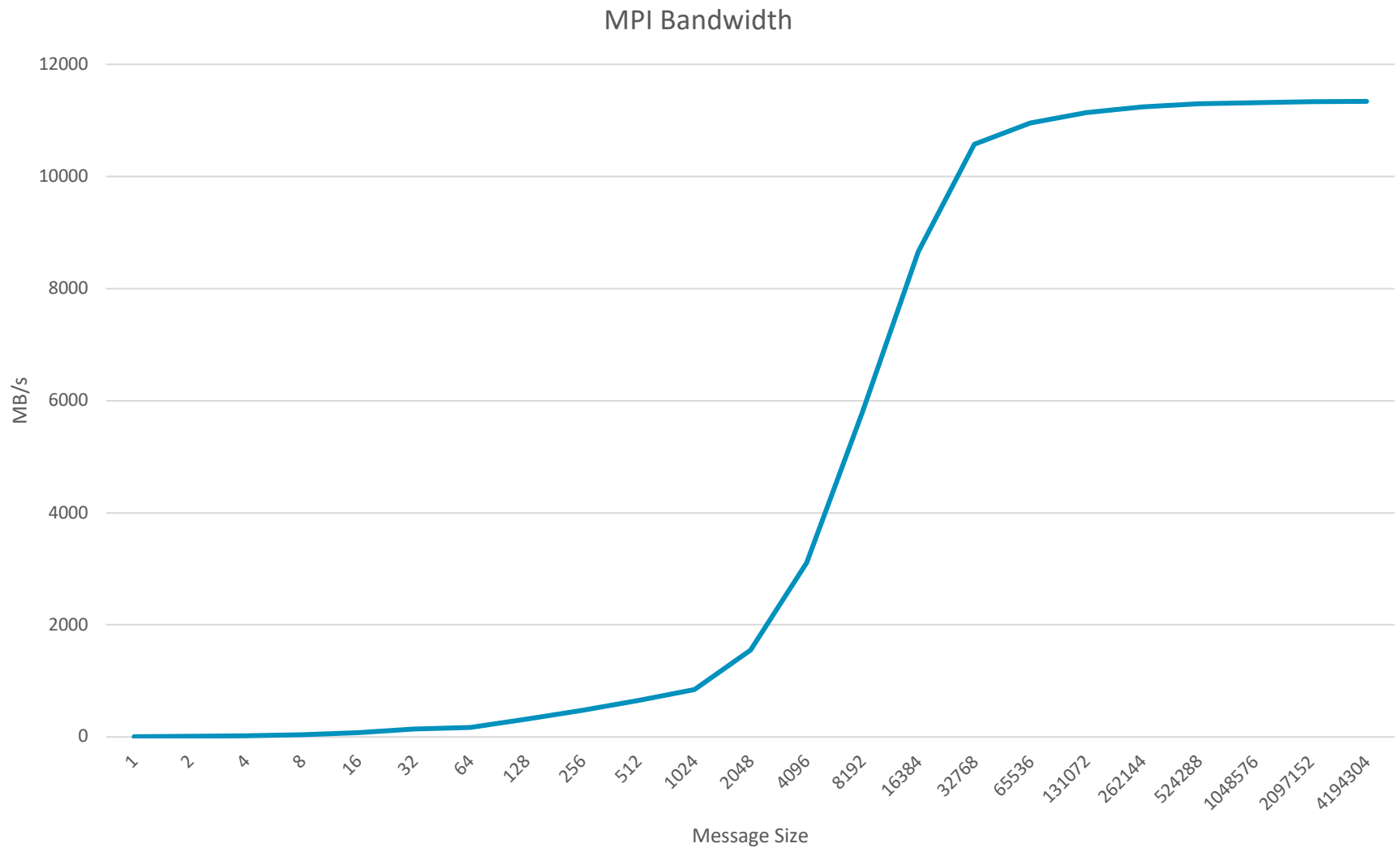
MPICH 3.3b – works on ARMv8

Open MPI 3.x – works on ARMv8

OSHMEM – work on ARMv8



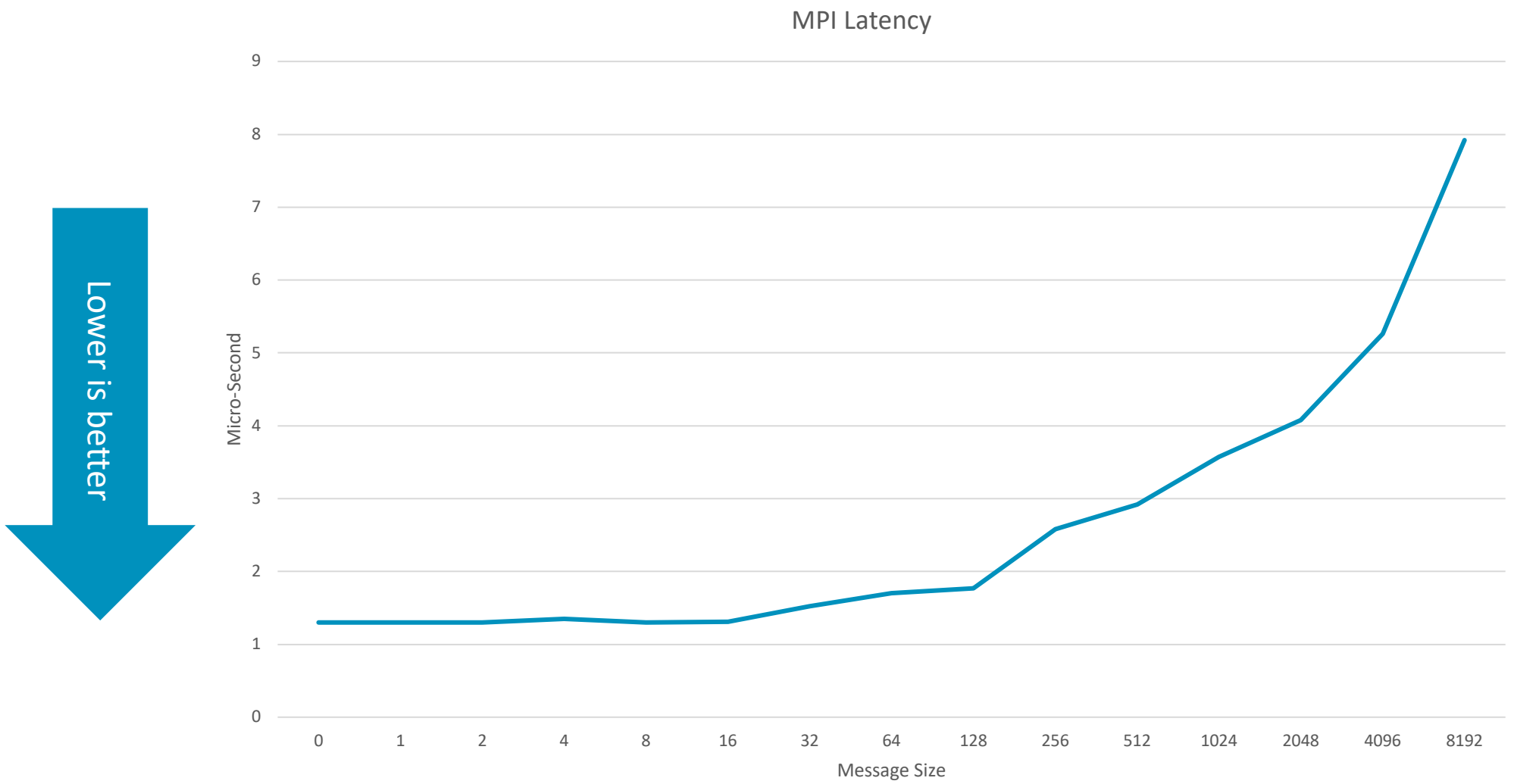
HPE Comanche (Apollo 70) with Cavium Thunder X2 SINGLE core, Mellanox ConnextX-4 100Gb/s (EDR) - Bandwidth



Higher is better

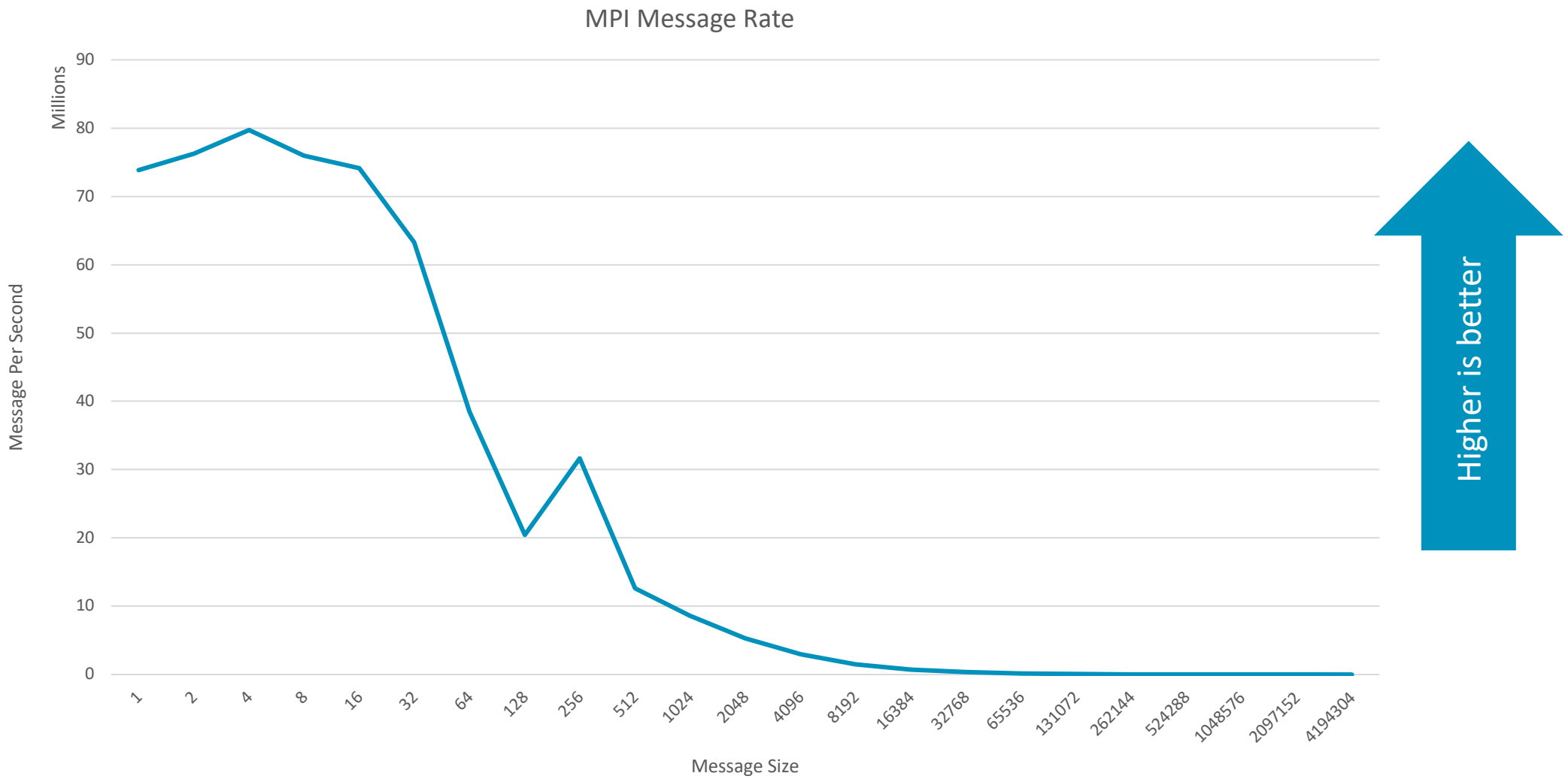
arm

HPE Comanche (Apollo 70) with Cavium Thunder X2, Mellanox ConnectX-4 100Gb/s (EDR) – Latency/Ping Pong



arm

HPE Comanche (Apollo 70) with Cavium Thunder X2, Mellanox ConnectX-4 100Gb/s (EDR) – MPI Message Rate (28 cores)



arm

MVAPICH Update

arm

Building with Arm compiler

Example:

my_cc=armclang

my_cxx=armclang++

my_fc=armflang

../configure CC=${my_cc} CXX=${my_cxx} F77=${my_fc} FC=${my_fc} --prefix=$INSTALL_DIR --with-device=ch3:mrail --with-rdma=gen2 --enable-cxx --enable-fc

Post configure fix for libtool (it does not get the right flags for armflang):

sed -i -e 's#wl=""#wl="-Wl,"#g' libtool

sed -i -e 's#pic_flag=""#pic_flag=" -fPIC -DPIC"#g' libtool

Arm Allinea Studio 18.4 release is now available on developer.arm.com

arm

Weakly Ordered Memory Model

Weakly ordered memory access means that changes to memory can be applied in any order as long as *single-core* execution sees the data needed for program correctness

Benefits:

- The processor can make many optimizations to reduce memory access
  - This has power (pushing bits is expensive) and memory bandwidth benefits
- The optimizations are transparent to single-core execution

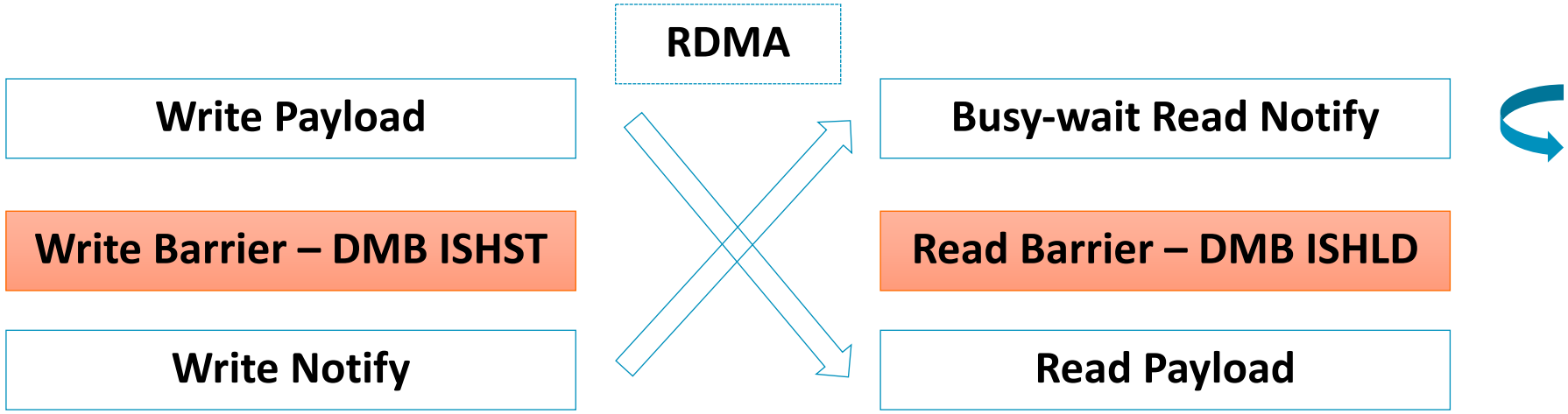
Challenges:

- Synchronization of data between cores must be explicit
- Popular legacy architectures (EG: x86_64, x86) provide “almost” strongly ordered memory access
  - This means that existing multi-core codes may be dependent on strongly ordered accesses

Memory Barriers on Arm

Memory Barriers

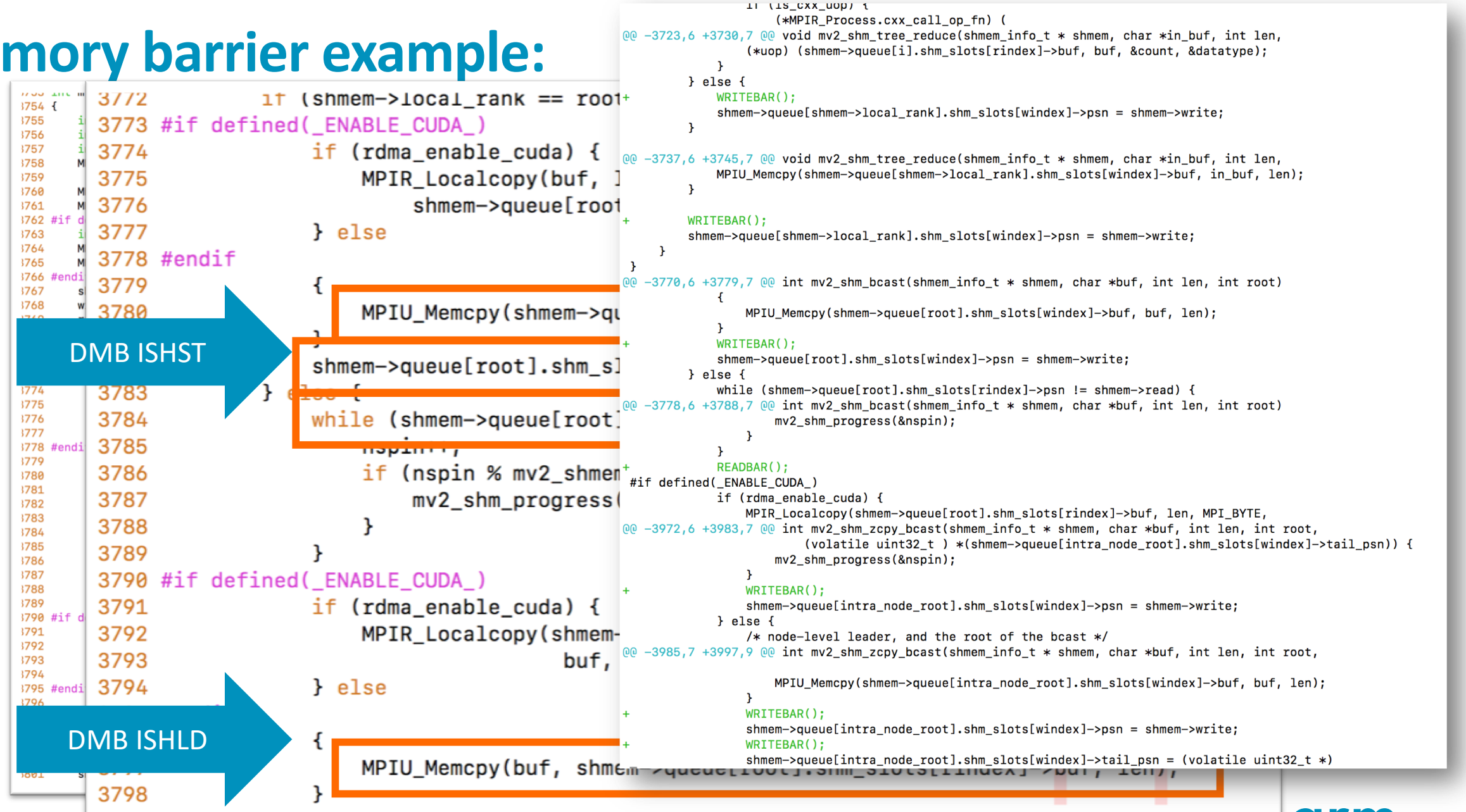
- Multithread environment
- Software-hardware interaction
- Examples https://github.com/openucx/ucx/blob/master/src/ucs/arch/aarch64/cpu.h#L25
- You can “fish” for these bugs in MPI implementations around Eager-RDMA and shared memory protocols



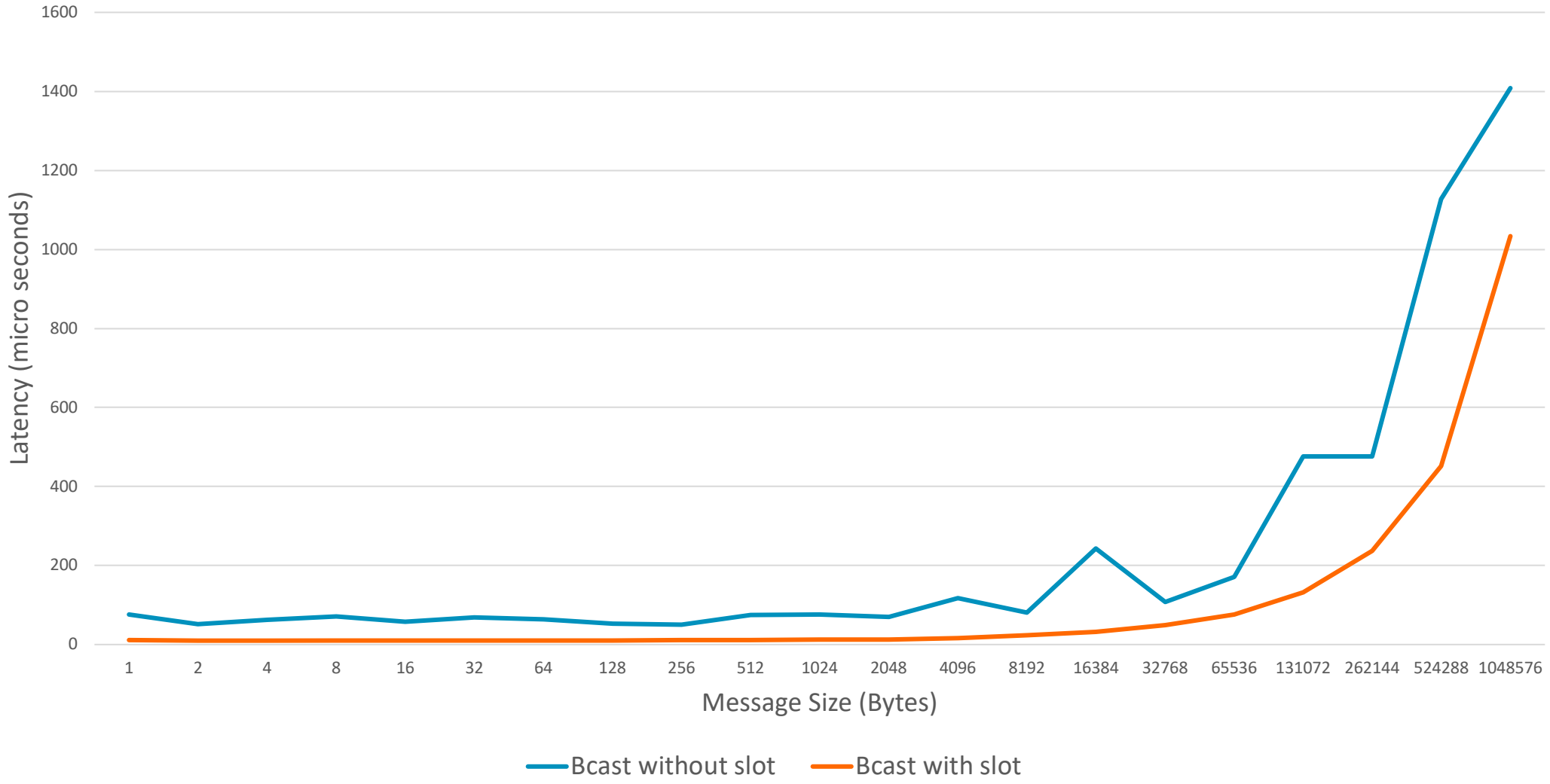
Maranget, Luc, Susmit Sarkar, and Peter Sewell. "A tutorial introduction to the Arm and POWER relaxed memory models." Draft available from http://www. cl. cam. ac. uk/~ pes20/ppc-supplemental/test7. pdf (2012).

arm

Memory barrier example:

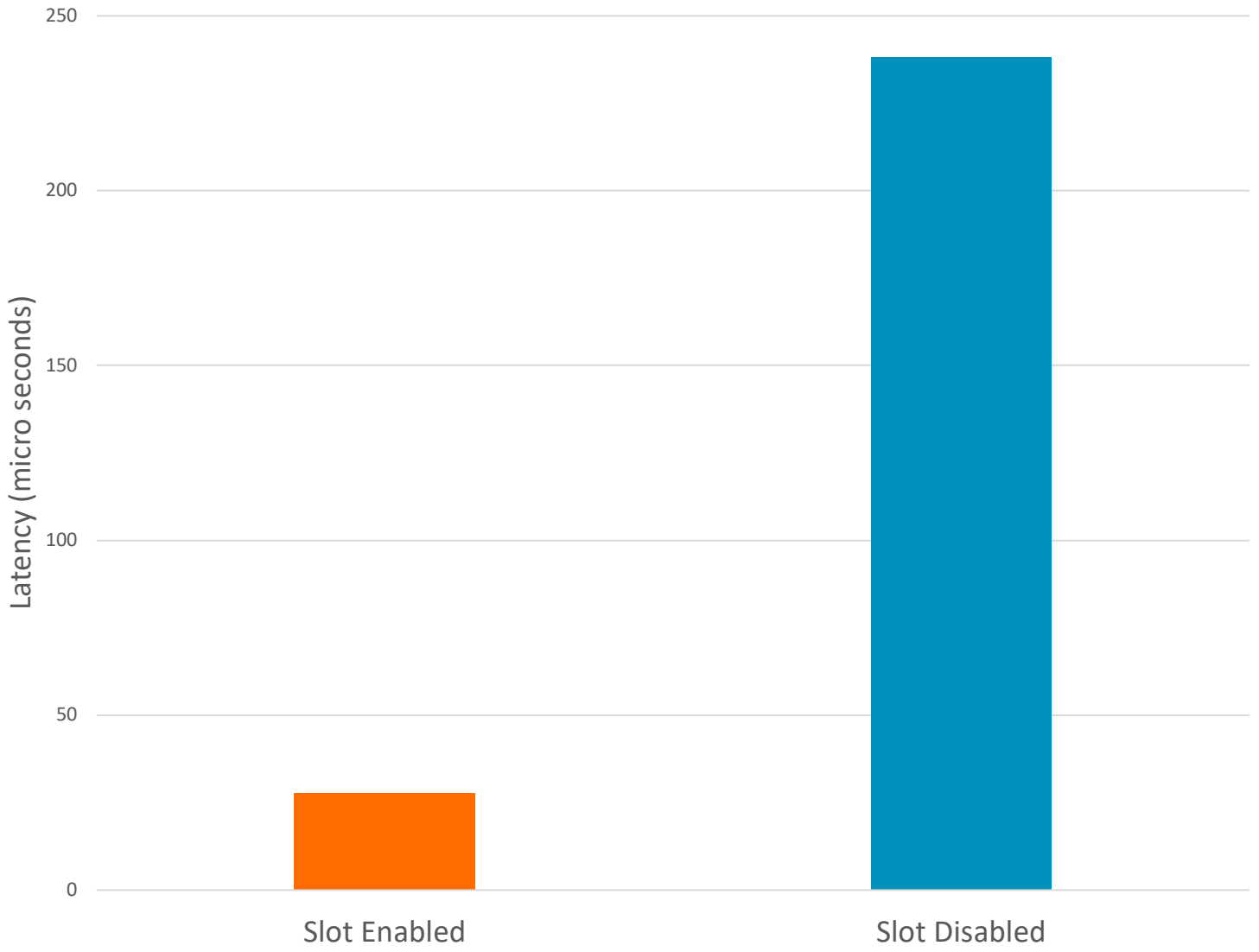


OSU MVAPICH BCAST (448 processes)



arm

OSU MVAPICH Barrier (448 processes)



arm

HPCAC - http://hpcadvisorycouncil.com/

Juno

- 1x GIGABYTE R270-T64 Chassis
  - 2 x Cavium ThunderX 48-core ARM processors
  - Memory: 64GB DDR4 2400 MHz
  - Mellanox ConnectX-4 EDR 100Gb/s InfiniBand/VPI adapter
  - SSD 480GB SATA 3
- 2x GIGABYTE MT30-GS0 Chassis
  - 1x Cavium ThunderX 32-core ARM Processor
  - Memory: 128GB DDR4 2400 MHz
  - Mellanox ConnectX-5 EDR 100Gb/s InfiniBand/VPI adapter
  - SSD 1TB SATA 3

Switch : Mellanox Switch-IB 2 SB7800 36-Port 100Gb/s EDR InfiniBand switches

Apply for System Access

ARM CAVIUM NETWORKS GIGABYTE Mellanox TECHNOLOGIES

arm

Thank You!
Danke!
Merci!
谢谢!
ありがとう!
Gracias!
Kiitos!
감사합니다
धन्यवाद

Backup