arm

Open UCX on Arm

Pavel Shamis / Pasha

Principal Research Engineer

SC 2018

Scaling

Announcing the **GW4** Tier 2 HPC service, 'Isambard': named after Isambard Kingdom Brunel

System specs:

- Cray CS-400 system
- **10,000+** ARMv8 cores
- HPC optimised software stack
- Technology comparison:
  - x86, KNL, Pascal
- To be installed March-Dec 2017
- £4.7m total project cost over 3 years

Simon McIntosh Smith, simonm@cs.bris.ac.uk, @simonmc

Sandia's NNSA/ASC ARM Platforms
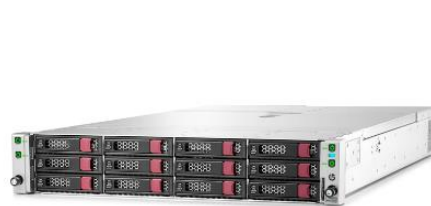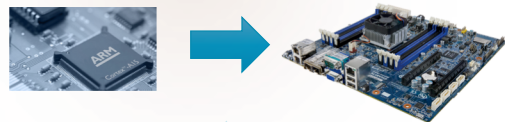
- **HPE Apollo 70:** This dense, scalable platform – and HPE's first ARM-based HPC system – brings more choice and flexibility to HPC customers. It provides easy access to HPC technology with support for standard HPE provisioning, cluster management and performance software. The Apollo 70, using Cavium's 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.
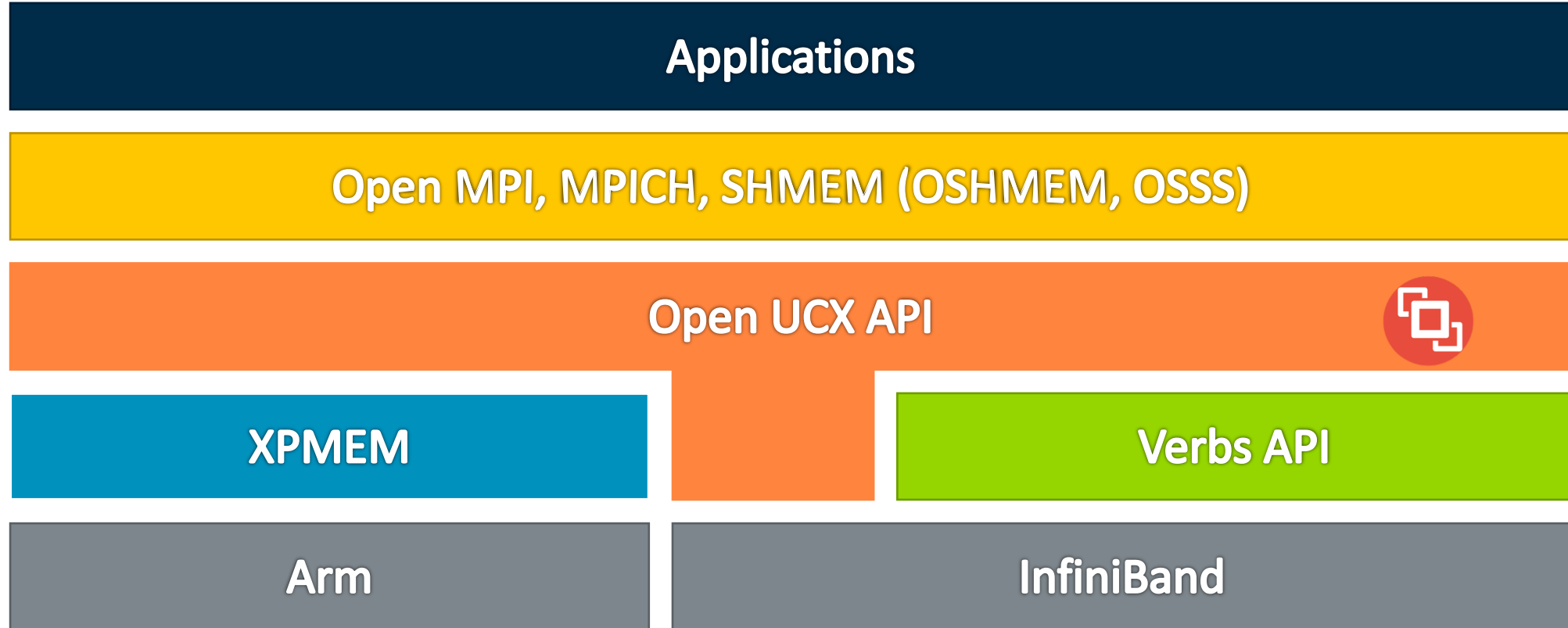
Open UCX on Arm

- Works seamlessly on Arm v8

  - https://developer.arm.com/products/software-development-tools/hpc/resources/porting-and-tuning/building-openmpi-with-openucx

- uArch level optimizations for ThunderX2 and other Arm implementations

- Optimized for NEON SIMD

- Optimized memory pulling routines for Arm

- Tested with multiple Arm v8 implementations

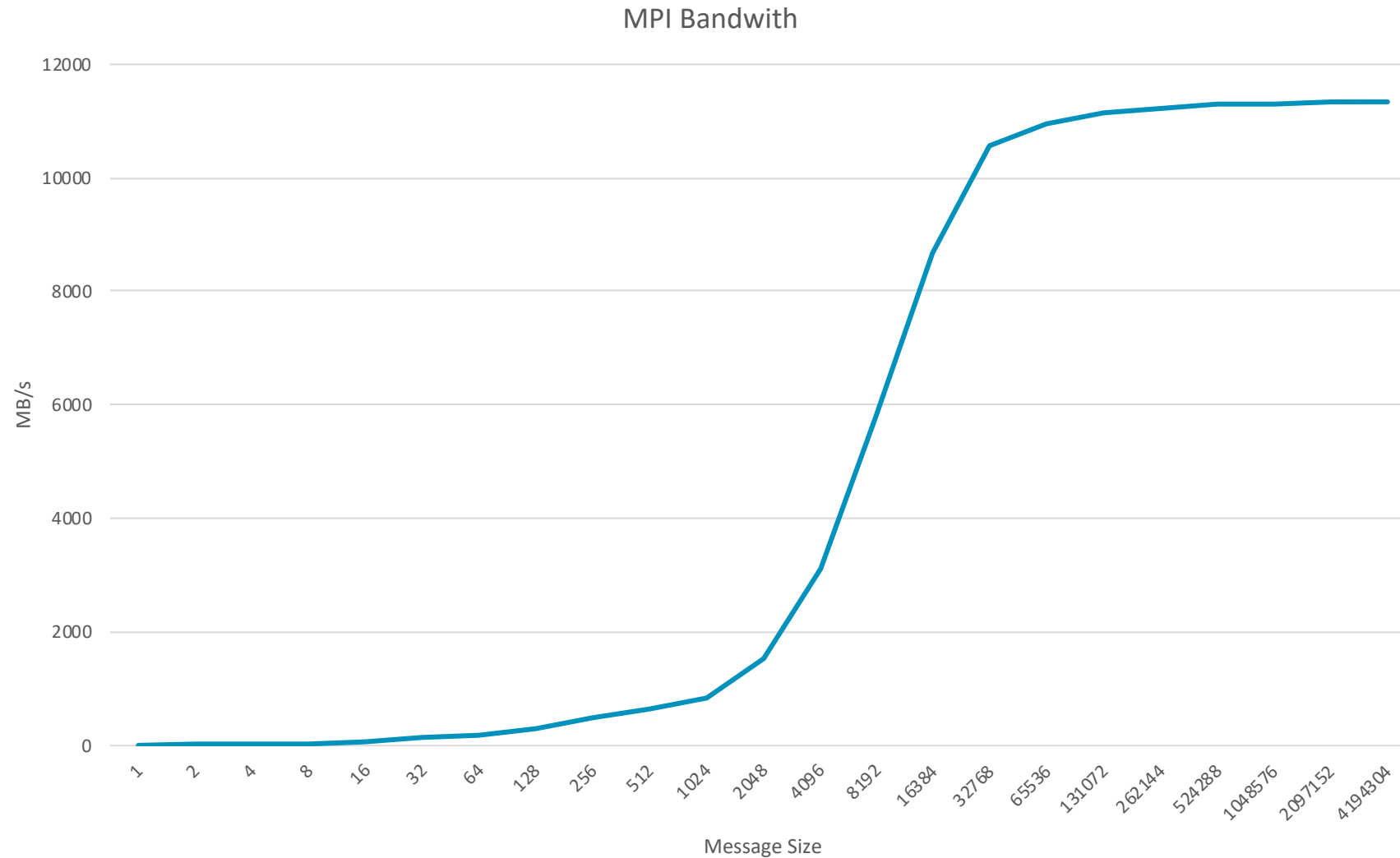- Tested with InfiniBand, ROCE, TCP, shared memory

arm

Hardware Software Stack Example

Applications

Open MPI, MPICH, SHMEM (OSHMEM, OSSS)

Open UCX API

XPMEM	Verbs API

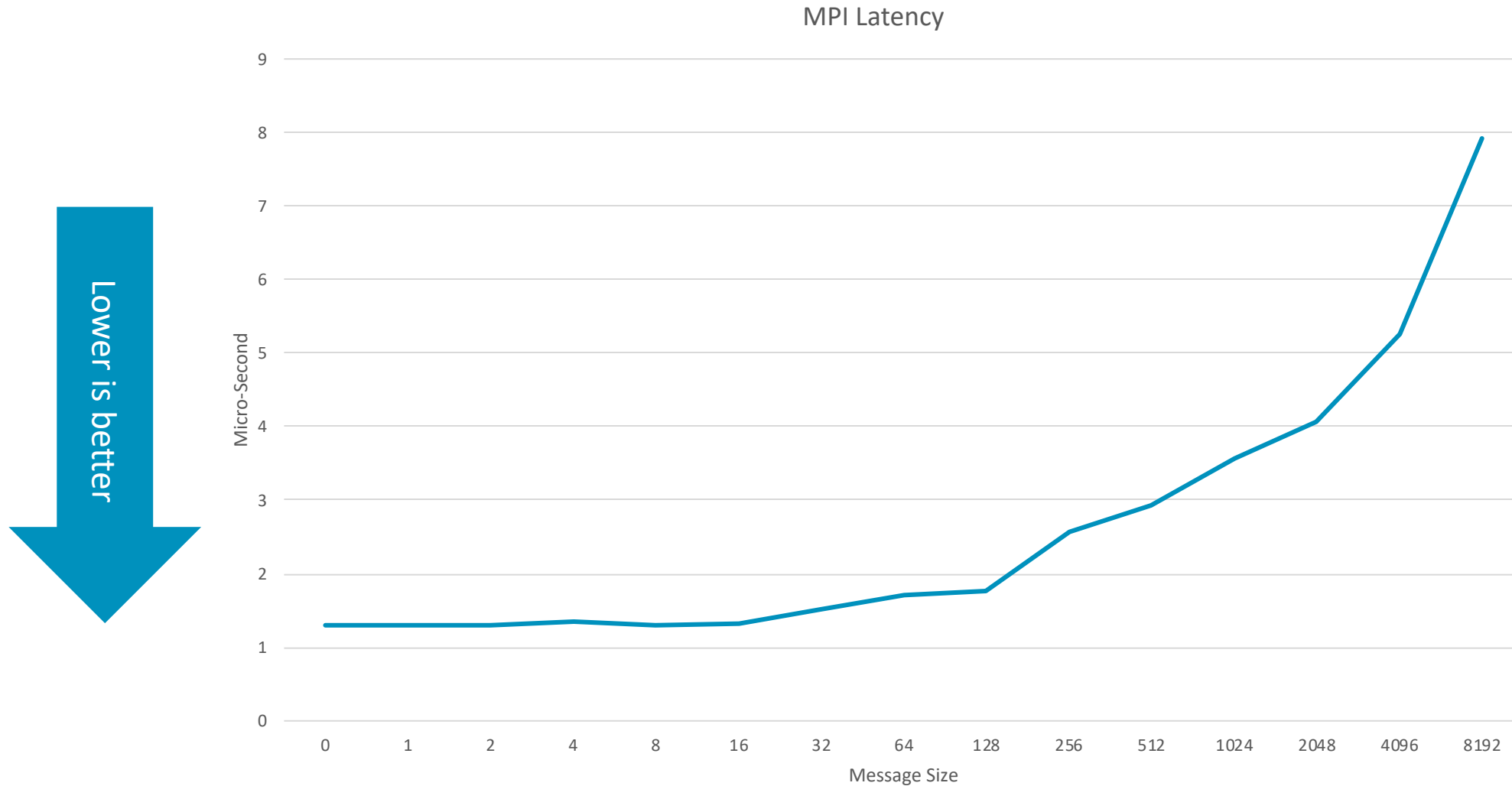| Arm	InfiniBand

arm

Testbed

- HPE Comanche platform with Cavium ThunderX2 (pre-production)

- Mellanox ConnectX-4 EDR (100Gb/s) HCA

- Mellanox 100Gb/s Switch

- OpenMPI 3.0 with OpenUCX 1.3 (MLNX_OFED_LINUX-4.3-1.0.1.0)
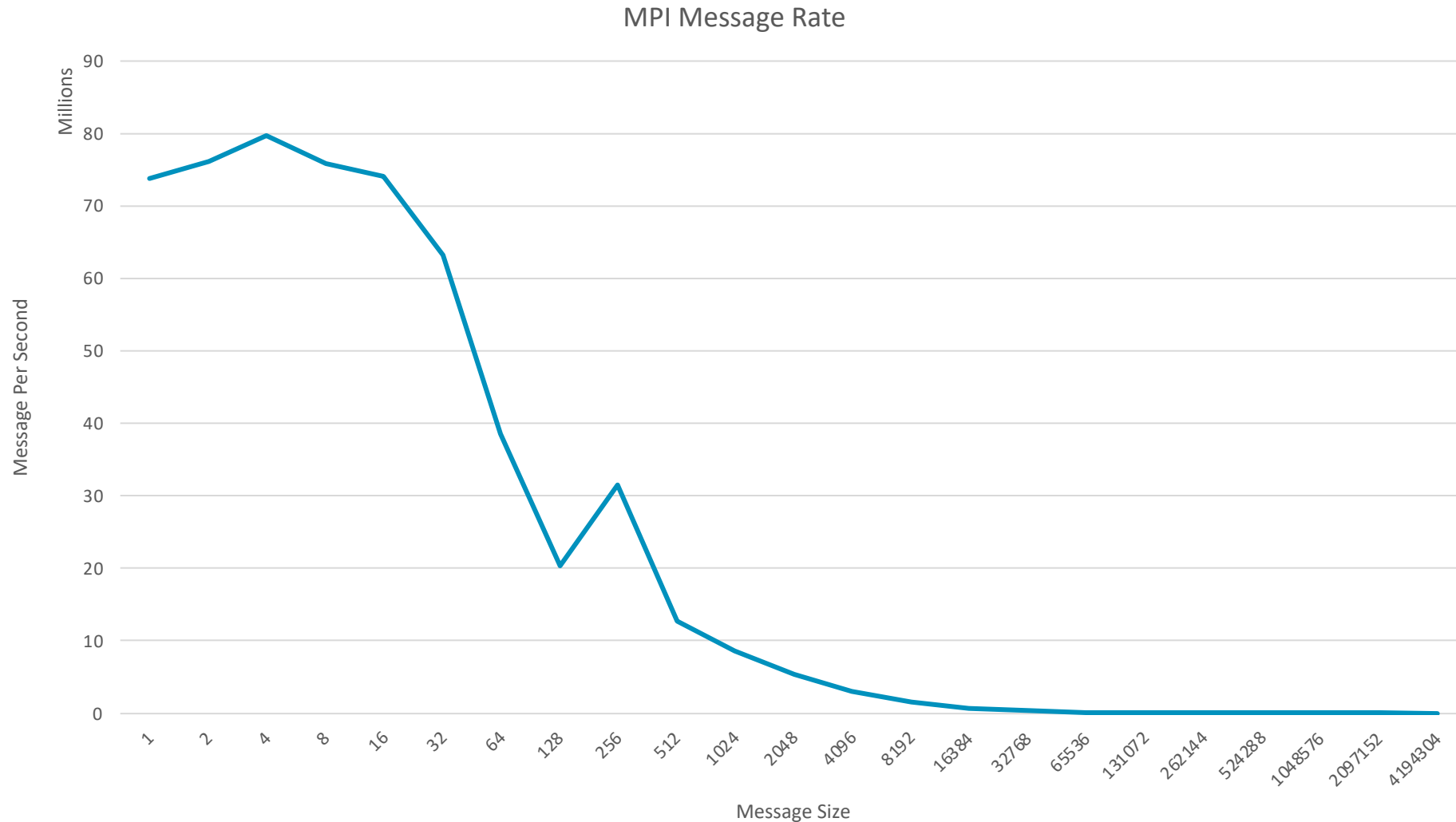
- OSU Micro-Benchmarks

arm

Cavium Thunder X2 SINGLE core InfiniBand Bandwidth

MPI Bandwith

arm

Cavium Thunder X2 MPI Ping-Pong Latency with InfiniBand

MPI Latency

Lower is better

Micro-Second

0	1	2	4	8	16	32	64	128	256	512	1024	2048	4096	8192

Message Size

arm

Cavium Thunder X2 MPI Message Rate with InfiniBand (28 cores)

MPI Message Rate

8   © 2018 Arm Limited

arm

arm

SHMEM_WAIT()

Latency of 'put' ping-pong with shmem_wait() (a); Total Benchmark Instruction count (b); Total Benchmark Cycle count (c)

arm