HPC Network Stack on Arm

Pavel Shamis/Pasha

Principal Research Engineer

Mvapich User Group Meeting, 2017

Annapolis, MD

Arm Overview

arm

An introduction to Arm

Arm is the world's leading semiconductor intellectual property supplier.

We license to over 350 partners, are present in 95% of smart phones, 80% of digital cameras, 35% of all electronic devices, and a total of 60 billion Arm cores have been shipped since 1990.

Our CPU business model:

License technology to partners, who use it to create their own system-on-chip (SoC) products.

We may license an instruction set architecture (ISA) such as "ARMv8-A")

or a specific implementation, such as "Cortex-A72".

Partners who license an ISA can create their own implementation, as long as it passes the compliance tests.

…and our IP extends beyond the CPU

arm

A partnership business model

A business model that shares success

- Everyone in the value chain benefits
- Long term sustainability

Design once and reuse is fundamental

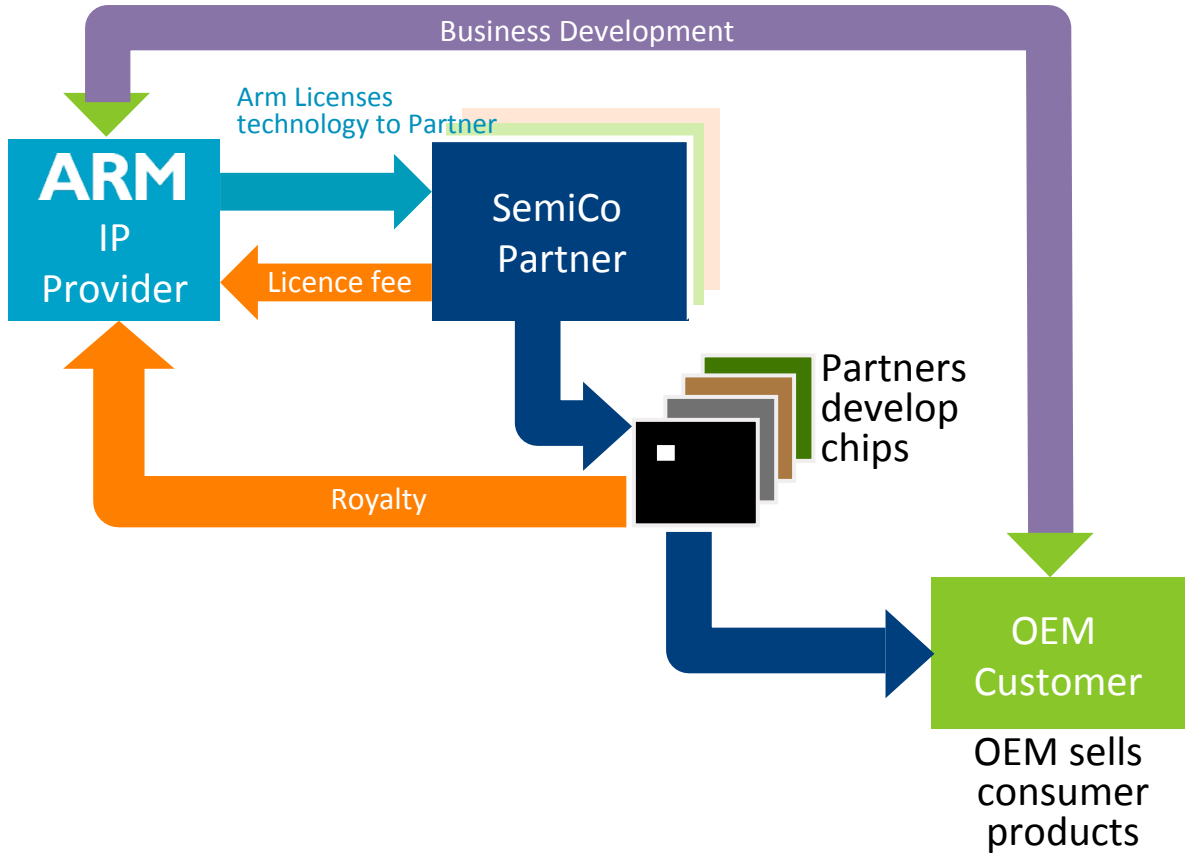
- Spread the cost amongst many partners
- Technology reused across multiple applications
- Creates market for ecosystem to target
  - Re-use is also fundamental to the ecosystem

Upfront license fee

- Covers the development cost

Ongoing royalties

- Typically based on a percentage of chip price
- Vested interest in success of customers



Approximately 1350 licenses
Grows by ~120 every year

More than 440 potential royalty payers

14.8bn+ Arm-powered chips in 2015

arm

Range of SoCs addressing infrastructure

XILINX ZYNQ UltraSCALE+

NXP QorIQ® Layerscape 2080A

Mellanox TECHNOLOGIES BlueField

socionext SC2A11

applied micro X-Gene 3™

QUALCOMM Centriq 2400

ALTERA Stratix 10 FPGA ▪ SoC

CAVIUM THUNDERX2

One size does not fit all

arm

Serious Arm HPC deployments starting in 2017

Two big announcements about Arm in HPC in Europe:

insideHPC

Bull Atos to Build HPC Prototype for Mont-Blanc Project using Cavium ThunderX2 Processor

January 16, 2017 by staff

Today the Mont-Blanc European project announced it has selected Cavium's ThunderX2 ARM server processor to power its new HPC prototype.

The new Mont-Blanc prototype will be built by Atos, the coordinator of phase 3 of Mont-Blanc, using its Bull expertise and products. The platform will leverage the infrastructure of the Bull sequana pre-exascale supercomputer range for network, management, cooling, and power. Atos and Cavium signed an agreement to collaborate to develop this new platform, thus making Mont-Blanc an Alpha-site for ThunderX2.

Bull
atos technologies

University of BRISTOL

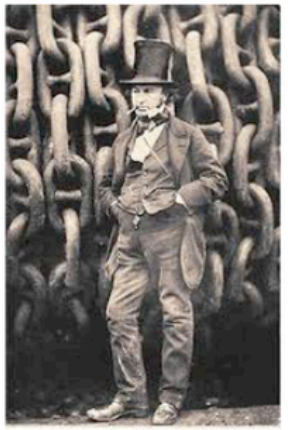
GW4

January 17th 2017

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



I.K.Brunel 1804-1859

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

5

bristol.ac.uk

arm

Japan

Post-K: Fujitsu HPC CPU to Support ARM v8

ARM FUJITSU

Post-K fully utilizes Fujitsu proven supercomputer microarchitecture

Fujitsu, as a lead partner of ARM HPC extension development, is working to realize ARM Powered® supercomputer w/ high application performance

ARM v8 brings out the real strength of Fujitsu's microarchitecture

HPC apps acceleration feature	Post-K	FX100	FX10	K computer
FMA: Floating Multiply and Add	✔	✔	✔	✔
Math. acceleration primitives*	✔Enhanced	✔	✔	✔
Inter core barrier	✔	✔	✔	✔
Sector cache	✔Enhanced	✔	✔	✔
Hardware prefetch assist	✔Enhanced	✔	✔	✔
Tofu interconnect	✔Integrated	✔Integrated	✔	✔

* Mathematical acceleration primitives include trigonometric functions, sine & cosines, and exponential...

slides from Fujitsu at ISC'16

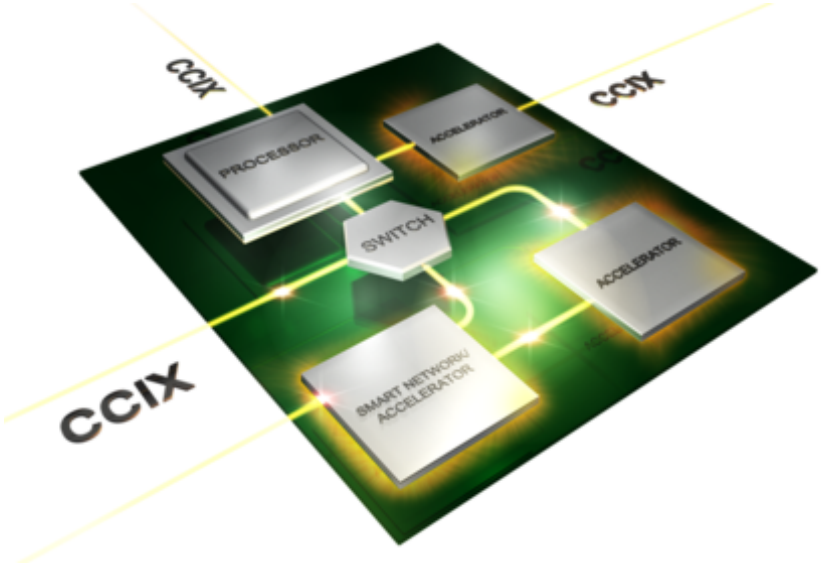
arm

Integration with Network Interconnects

arm

CCIX

CCIX Cache Coherent Interconnect for Accelerators



Accelerators and Network (NIC/HCA/etc.) as a first class "citizen" in the system

Seamless process and accelerator hardware cache coherence support

Low-latency and high-bandwidth

Allow in-line acceleration

- Bump in the wire processing (network packet processing, storage acceleration, etc.)

Allows "off-line" acceleration (co-processor model)

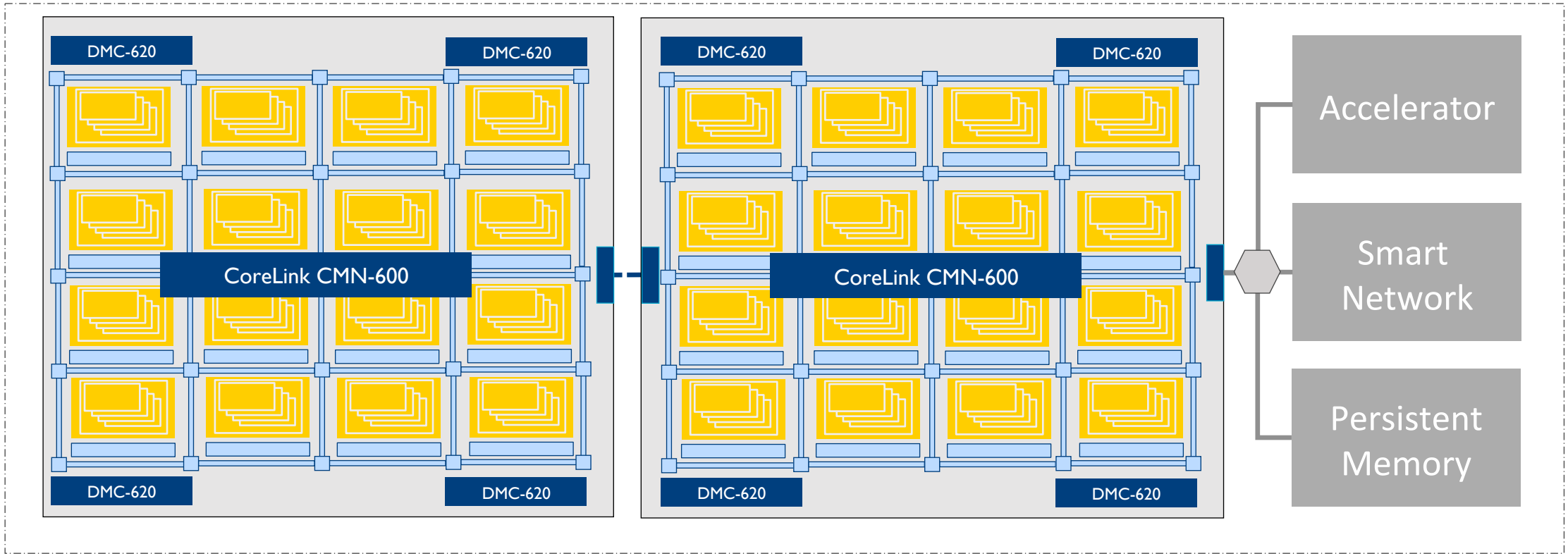
Driver-less / interrupt-less usage model

http://www.ccixconsortium.com

AMD Amphenol ARM Arteris
avery design systems BROADCOM Bull atos technologies cadence
CAVIUM HUAWEI IBM IDT
KEYSIGHT TECHNOLOGIES Mellanox TECHNOLOGIES Micron NETSPEED SYSTEMS
QUALCOMM redhat SYNOPSYS tsmc
TELEDYNE LECROY Everywhereyoulook XILINX ALL PROGRAMMABLE

arm

Scale-up server node



Shared virtual memory system

arm

CCIX multichip connectivity and topologies

New class of interconnect providing high performance, low latency for new accelerators use cases

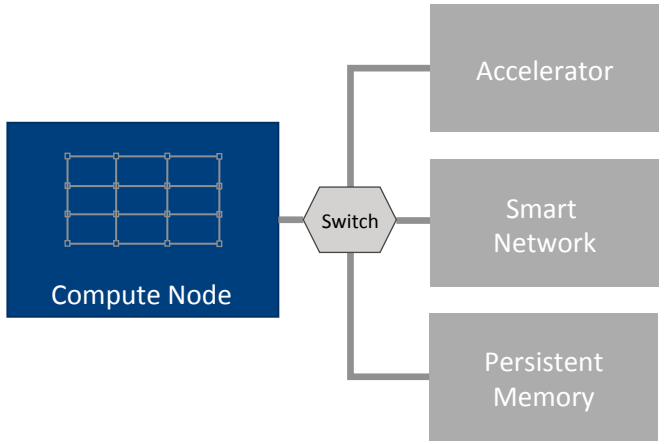
- CCIX defines 25GT/s (3x performance*)
- Examining 56GT/s (7x performance*) and beyond
- Enabling low latency via light transaction layer

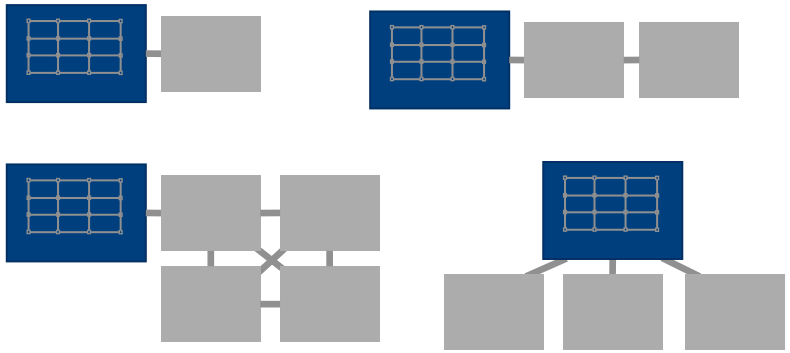
Flexible, scalable interconnect topologies

- Flexible point-to-point, daisy chained and switched topologies

Simplified deployment by leveraging existing PCIe hardware and software infrastructure

- Runs on existing PCIe transport layer and management stack
- Coexist with legacy PCIe designs





* Note: Based on PCIe Gen3 Performance

arm

Building CCIX devices

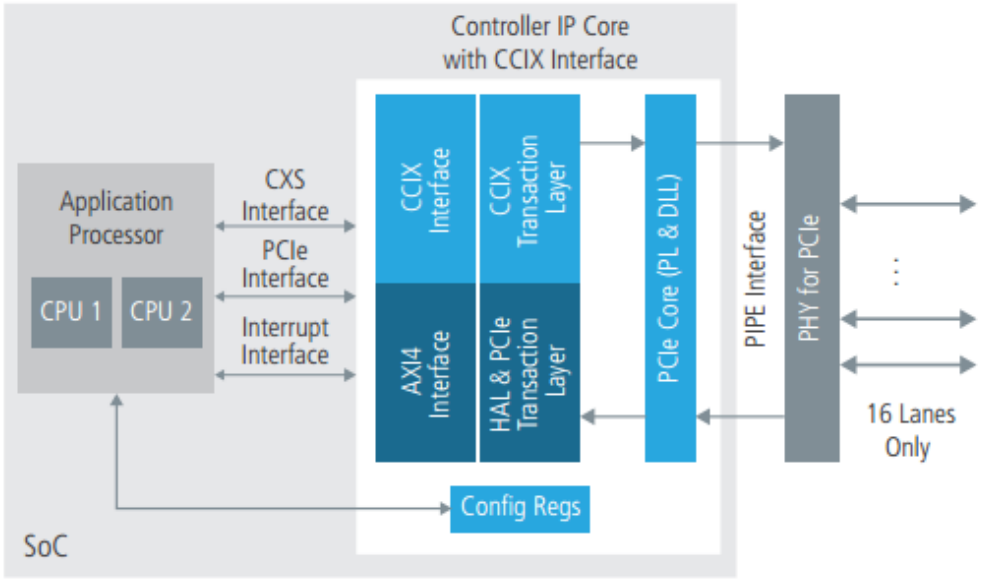
Cadence® IP for CCIX

- Built upon silicon proven PCIe solutions

IP products:

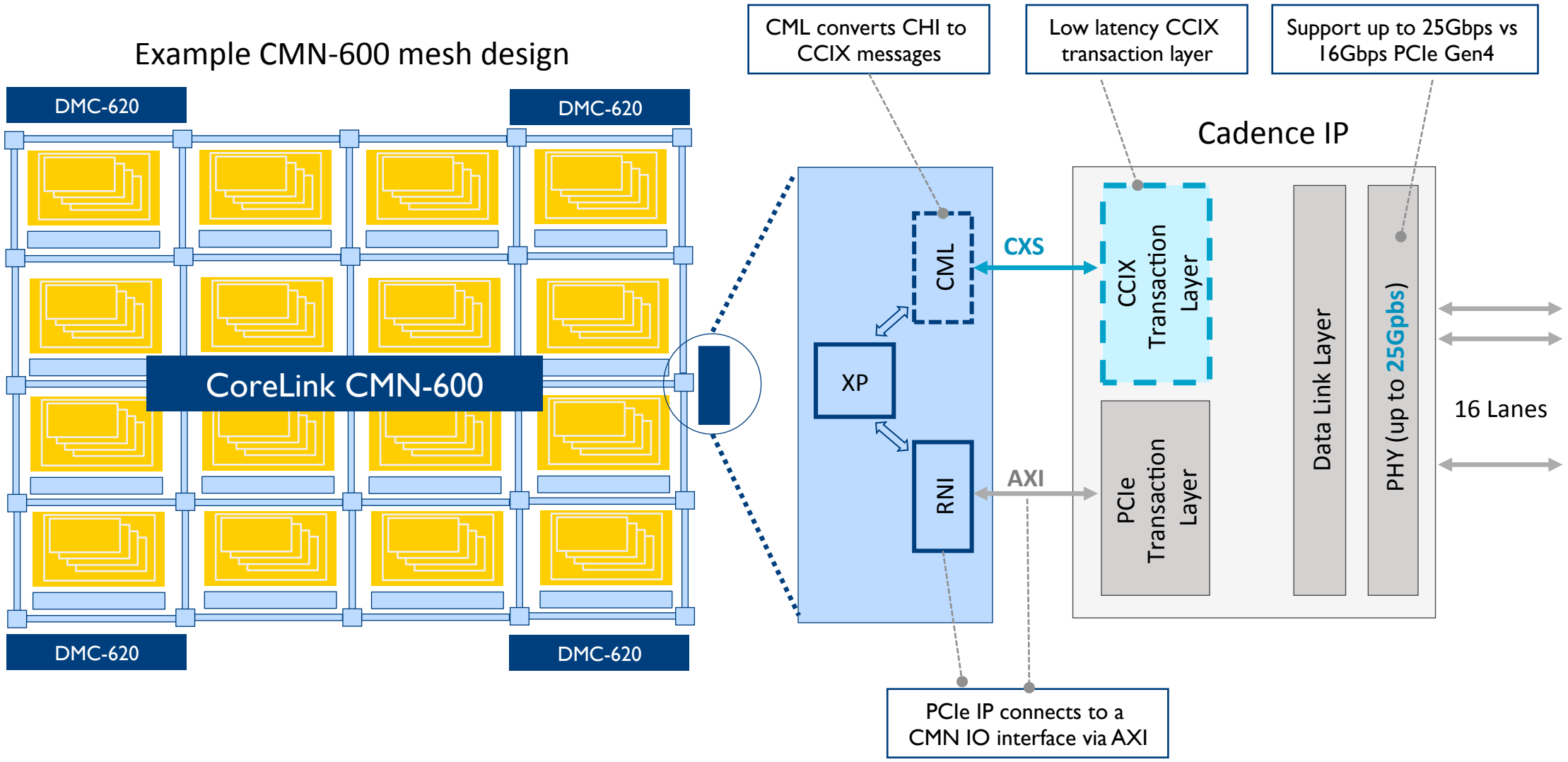
- Controller IP – Provides the CCIX transaction and data link layers.
- PHY IP – Provides the high performance SERDES physical layer supporting speeds up to 25Gpbs.
- Verification IP – Provides the necessary test infrastructure to verify CCIX designs.

Cadence Controller & PHY IP



arm

Cadence CCIX integration



arm

Gen-Z

All data is accessed by some form of a Read or a Write

Example of reads: DDR Row + Column Read, PCI DMA Read, SCSI Write, Socket Read, File Read, RDMA Read

Example of writes: DDR Row + Column Write, PCI DMA Write, SCSI Read, Socket Write, File Write, RDMA Write

The Goal: Simplify world to memory semantic Reads & Writes

Gen-Z Overview

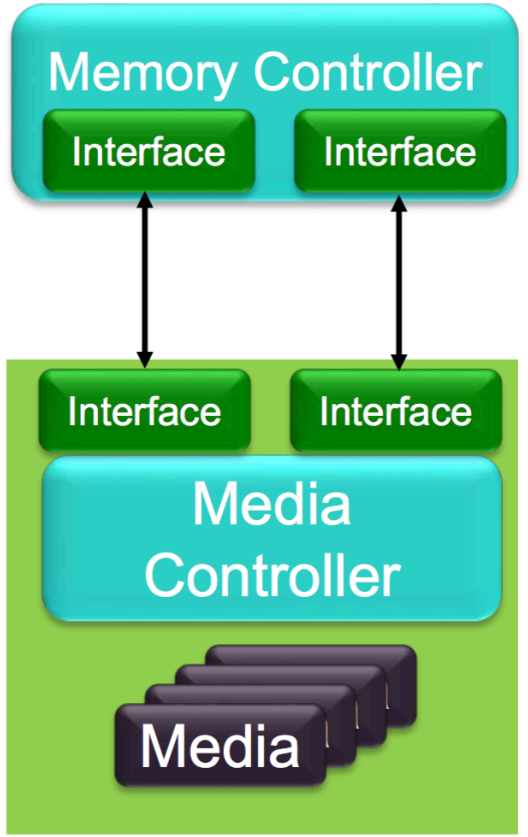
An open, standards-based, scalable, system interconnect and protocol.

Optimized to support memory semantic communications

Breaks Processor-Memory Interlock

Split controller model

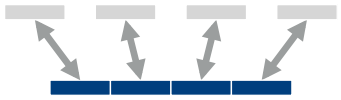
- Memory controller
  - Initiates high-level requests—Read, Write, Atomic, Put / Get, etc.
  - Enforces ordering, reliability, path selection, etc.
- Media controller
  - Abstracts memory media
  - Supports volatile / non-volatile / mixed-media
  - Performs media-specific operations
  - Executes requests and returns responses
  - Enables data-centric computing (accelerator, compute, etc.)



arm

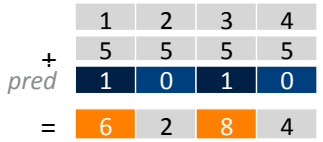
Introducing the Scalable Vector Extension (SVE)

A vector extension to the ARMv8-A architecture with some major new features:



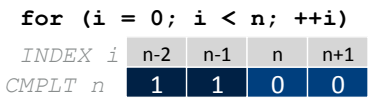
Gather-load and scatter-store

Loads a single register from several non-contiguous memory locations.



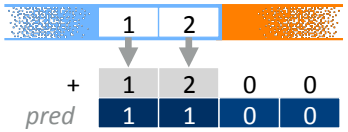
Per-lane predication

Operations work on individual lanes under control of a predicate register.



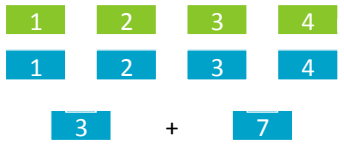
Predicate-driven loop control and management

Eliminate scalar loop heads and tails by processing partial vectors.



Vector partitioning and software-managed speculation

First Faulting Load instructions allow memory accesses to cross into invalid pages.



Extended floating-point horizontal reductions

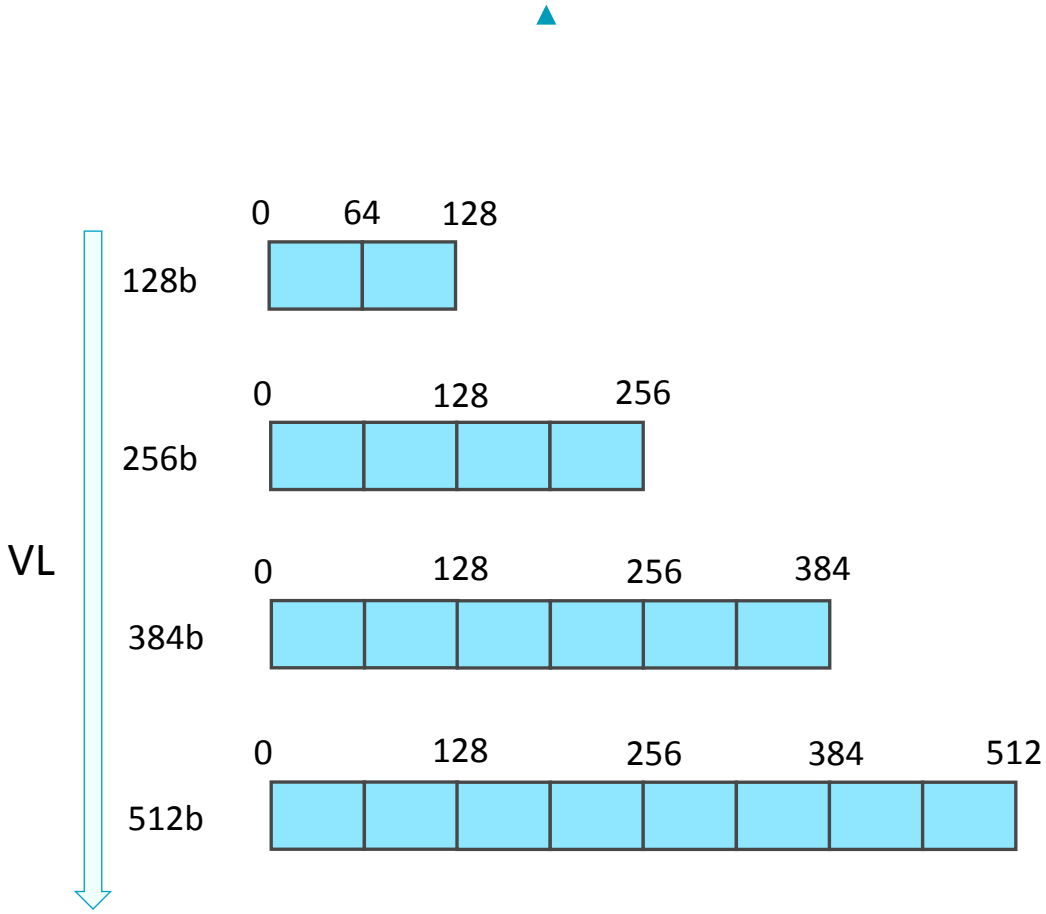
In-order and tree-based reductions trade-off performance and repeatability.

arm

What's the vector length?

There is **no** preferred vector length

- Vector Length (VL) is the CPU implementor's choice, from 128 to 2048 bits, in increments of 128
- Adopting a Vector Length Agnostic (VLA) code generation style makes code portable across all possible vector lengths.
- VLA is made possible by the per-lane predication, predicate-driven loop control, vector partitioning and software-managed speculation features of SVE.
- No need to recompile, or to rewrite hand-coded SVE assembler or C intrinsics

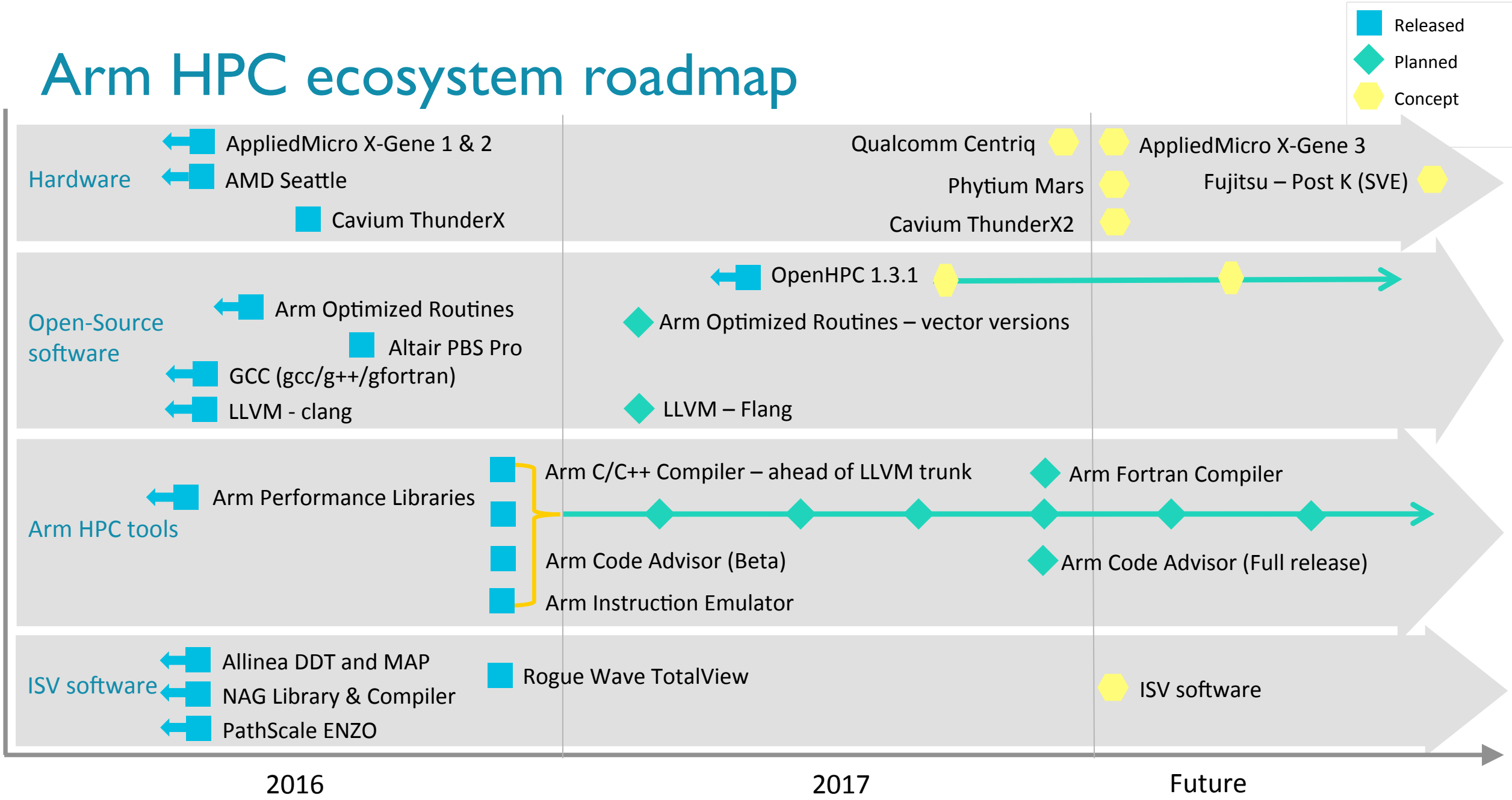


arm

Software Eco-system

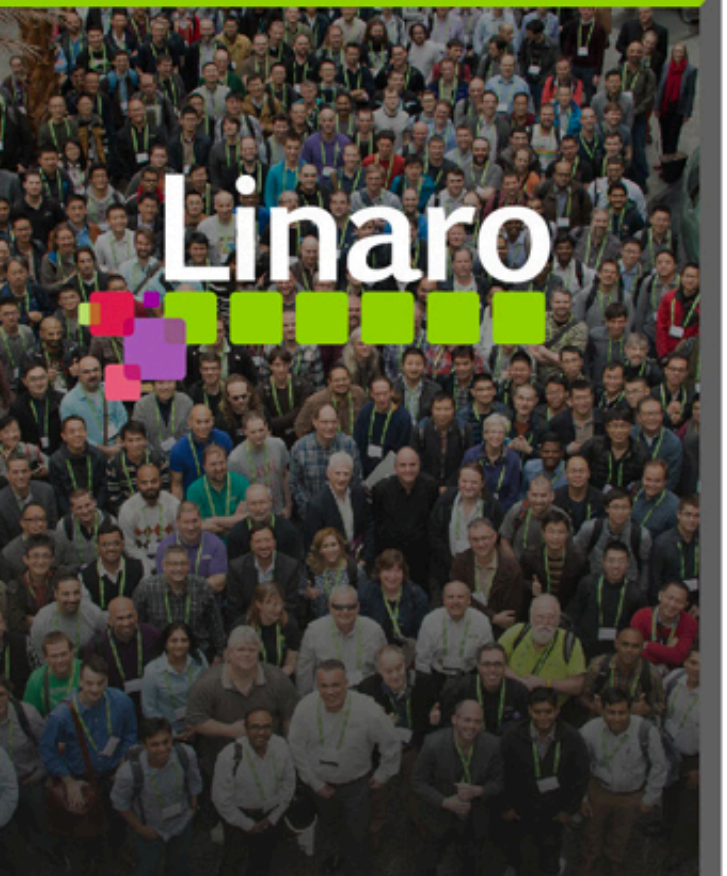
arm

Arm HPC ecosystem roadmap



arm

Celebrating 5 years of Open Source Engineering on ARM



24TB

data from June 2014 - May 2015
615,000 downloads from **>100 countries**

16 Connects
14 Cities on 3 continents

32 member companies
Six members at launch

4,410
gallons consumed at Linaro Connects

1,141,014
minutes of videos showing demos, talks and training sessions watched

More than **220 Engineers**
from seed of twenty

#3
company contributor for
Linux Kernels 3.11 - 3.18

11,589
patches upstream since 2011

~20
hardware platforms

www.linaro.org
www.96boards.org

50,217
Wiki pages

>50
Monthly Software Releases

openembedded
yocto PROJECT
redhat.

>1 Million
website users

"The ARM situation has just improved tremendously over the last several years. It used to be a major pain to me, it has gone to almost being entirely painless."

- Linus Torvalds
May 2015

Note:
Linus Torvalds Image from Linux Foundation. Icons made by Freepik. Logos & trademarks remain the property of their respective owners and represent a range of products and services supported by Linaro.

Linux / FreeBSD w/ AARCH64 support

Debian 8 adds AARCH64 – April 2015

12.04LTS & 14.04LTS ← Also 14.10 & 15.04 releases released

Fedora 22 released – May 2015
Fedora 23 released – Nov 2015

Red Hat Enterprise Linux Server for Arm
7.2 BETA – Sept, 2015

CentOS Linux 7 for AArch64
GA – August 2015

OpenSUSE 13.2 – Nov 2014

SUSE Launches Partner Program to Bring SUSE Linux Enterprise 12 to 64-bit Arm
July 2015 @ ISC

← Engaged with FreeBSD foundation / Semi-half & Cavium to get FreeBSD on ARMv8
FreeBSD Beta version demo'd by Semihalf – Nov. 2015

arm

Open source and commercial compilers

- GCC
  - C, C++, Fortran
  - OpenMP 4.0

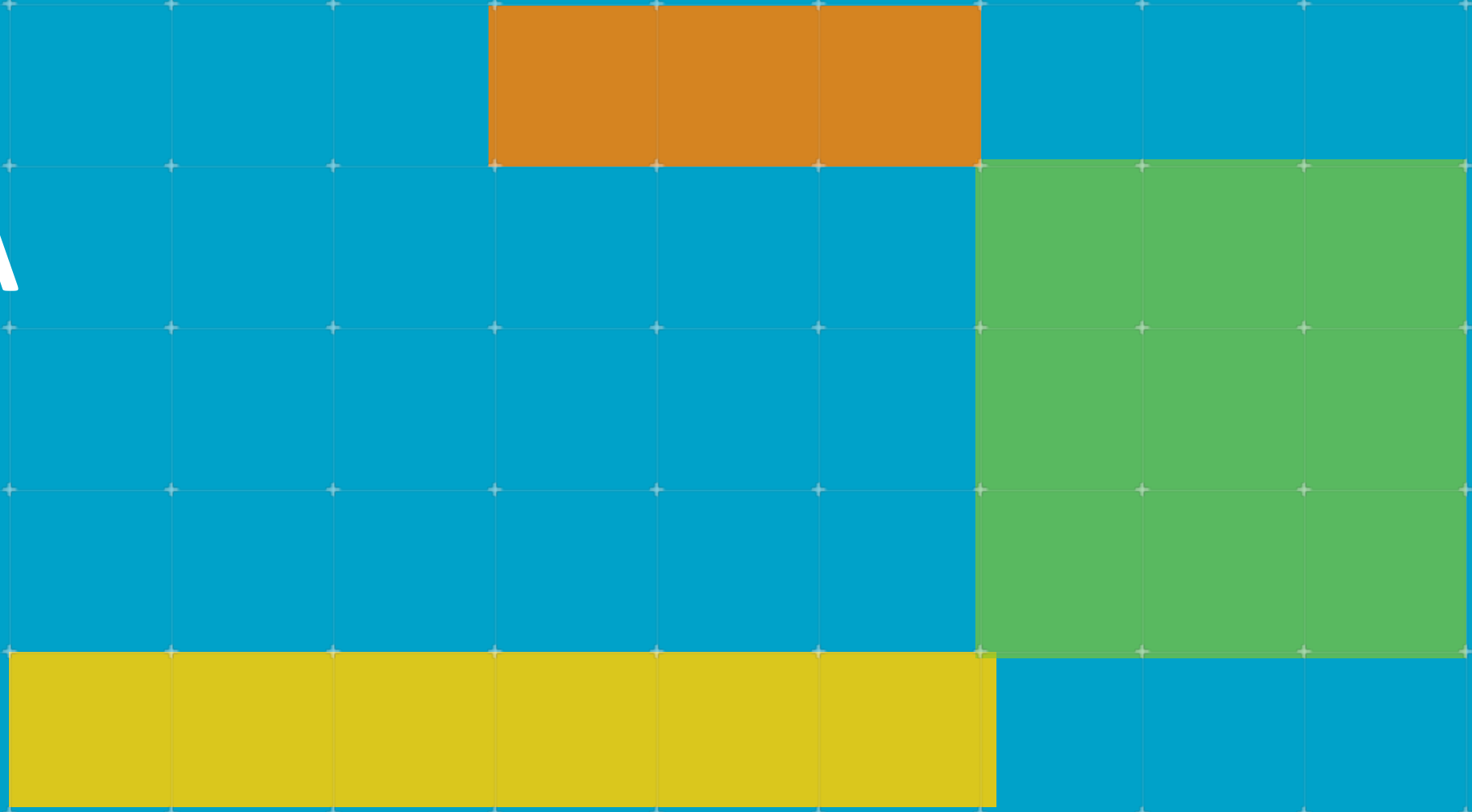
- NAG
  - Fortran
  - OpenMP 3.1

- LLVM
  - C, C++, Fortran
  - OpenMP 3.1, (4.0 coming soon)
  - Fortran coming Q1 2017

- Arm C/C++/Fortran Compiler
  - LLVM based
  - Includes SVE

arm

RDMA

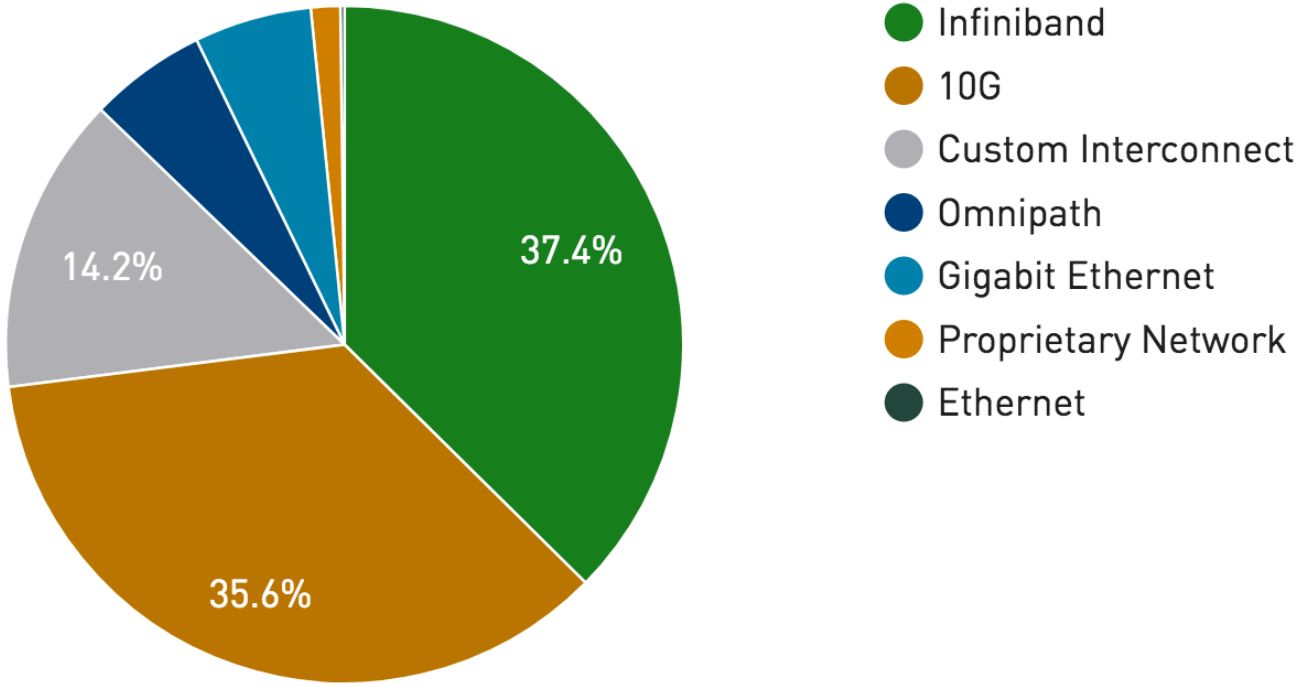


arm

RDMA Networks

Remote Direct Memory Access (RDMA) – popular hardware network technology

- InfiniBand – 37% of systems in TOP500



https://www.top500.org

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 ARMv8 support

ubuntu

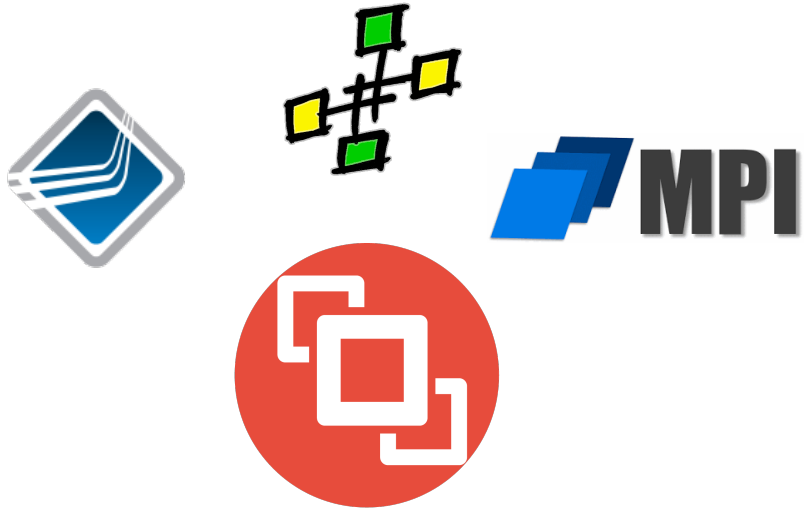
redhat

SUSE

arm

OpenUCX v1.2

- The first official release from Open UCX community
  - https://github.com/openucx/ucx/releases/tag/v1.2.0
- Features
  - Support for InfiniBand and RoCE
    - Transports RC, UD, DC
  - Support for Accelerated Verbs – 40% speedup on Arm compared to vanilla Verbs
  - Support for UGNI API for Aries and Gemini
  - Support for Shared Memory: KNEM, CMA, XPMEM, Posix, SySV
  - Support for x86, ARMv8, Power
  - Efficient memory polling – 36% increase in efficiency on Arm
  - UCX interface is integrated with MPICH, OpenMPI, OSHMEM, ORNL-SHMEM, etc.



Pavel Shamis, M. Graham Lopez, and Gilad Shainer. "Enabling One-sided Communication Semantics on Arm", HIPS 2017

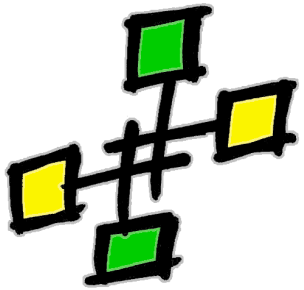
Programing models

MVAPICH 2.3b – works on ARMv8 !

Open MPI – works on ARMv8

MPICH – compiles and runs on ARMv8

OSHMEM – compiles and runs



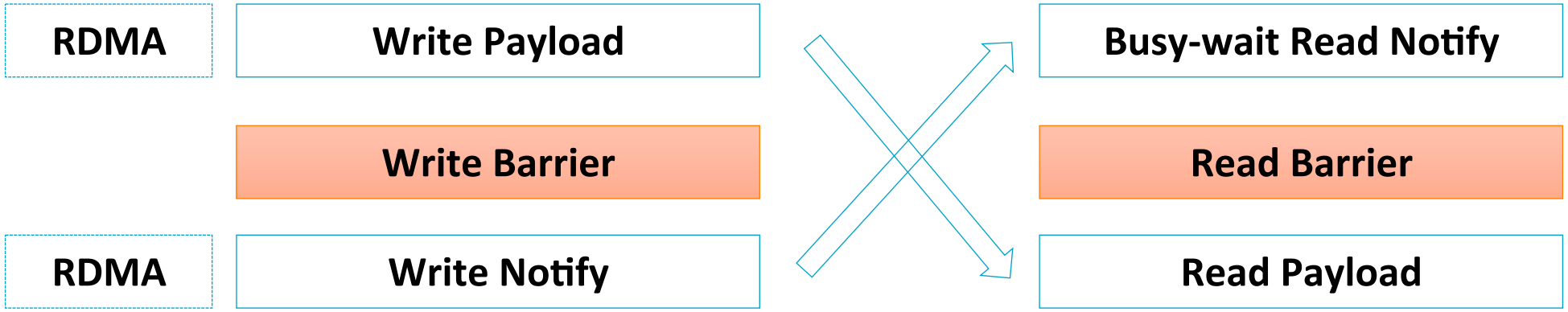
Lessons Learned

Memory Barriers

```
#define ucs_memory_bus_fence()       asm volatile ("dsb sy" ::: "memory");
#define ucs_memory_bus_store_fence() asm volatile ("dsb st" ::: "memory");
#define ucs_memory_bus_load_fence()  asm volatile ("dsb ld" ::: "memory");

#define ucs_memory_cpu_fence()       asm volatile ("dmb sy" ::: "memory");
#define ucs_memory_cpu_store_fence() asm volatile ("dmb st" ::: "memory");
#define ucs_memory_cpu_load_fence()  asm volatile ("dmb ld" ::: "memory");
```

- 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

Lessons Learned – continued

Not all cache-lines are 64Byte !

- Implementation dependent
- 128Byte and 64Byte

http://xeroxnostalgia.com/duplicators/xerox-9200/

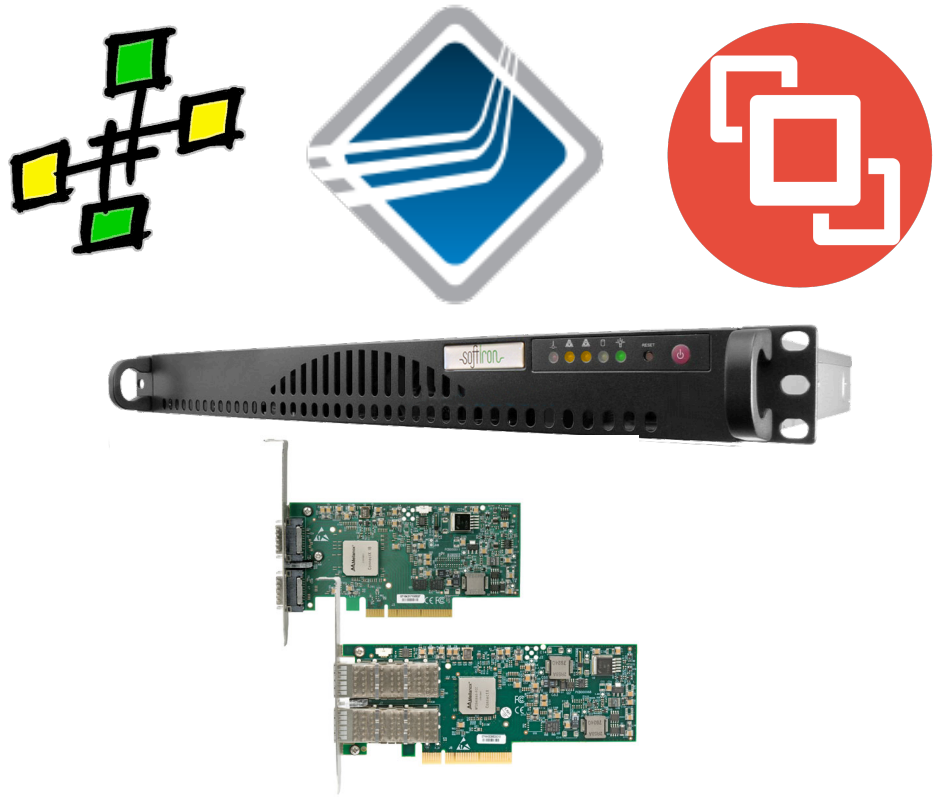
arm

Preliminary Results

arm

Testbed

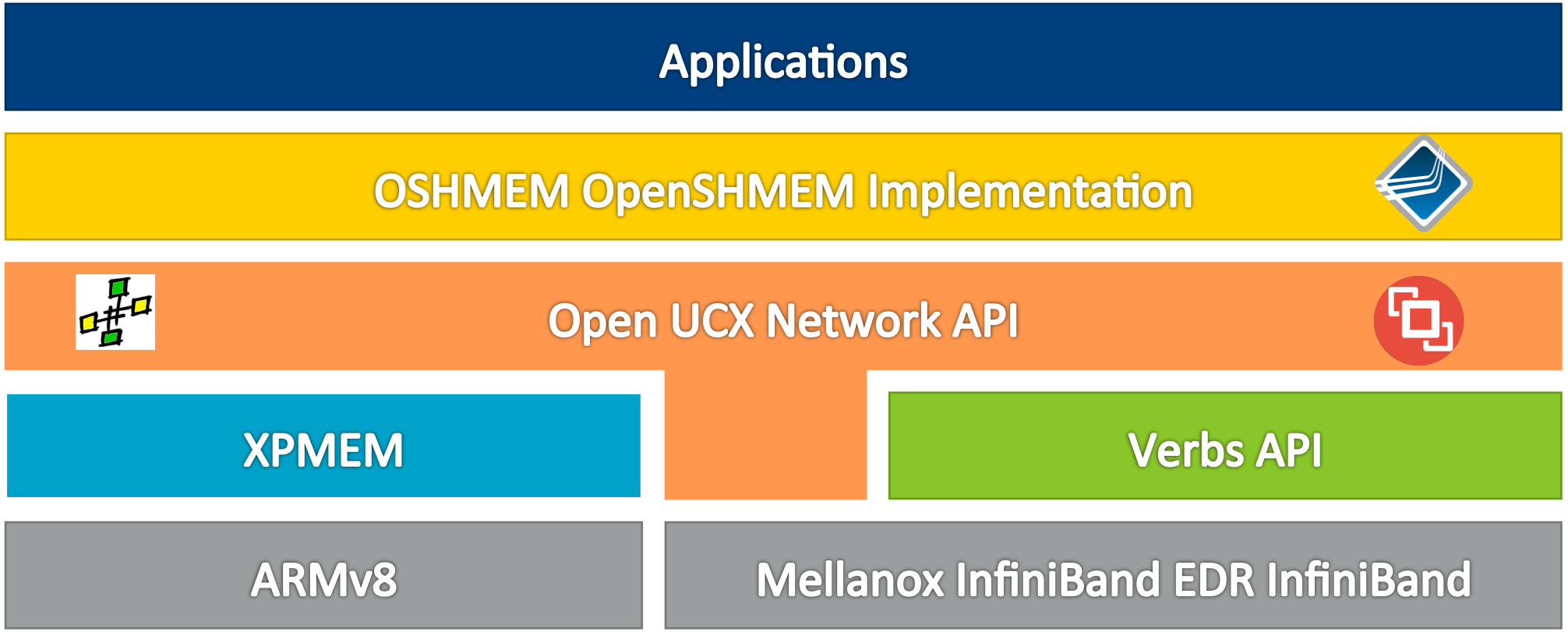
- 2 x Softiron Overdrive 3000 servers with AMD Opteron A1100 / 2GHz
- ConnectX-4 IB/VPI EDR (PCIe gen2 x8)
- Ubuntu 16.04
- MOFED 3.3-1.5.0.0
- UCX [0558b41]
- XPMEM [bdfcc52]
- OSHMEM/OPEN-MPI [fed4849]



Pavel Shamis, M. Graham Lopez, and Gilad Shainer. "Enabling One-sided Communication Semantics on Arm"

arm

Hardware Software Stack Overview



arm

XPMEM

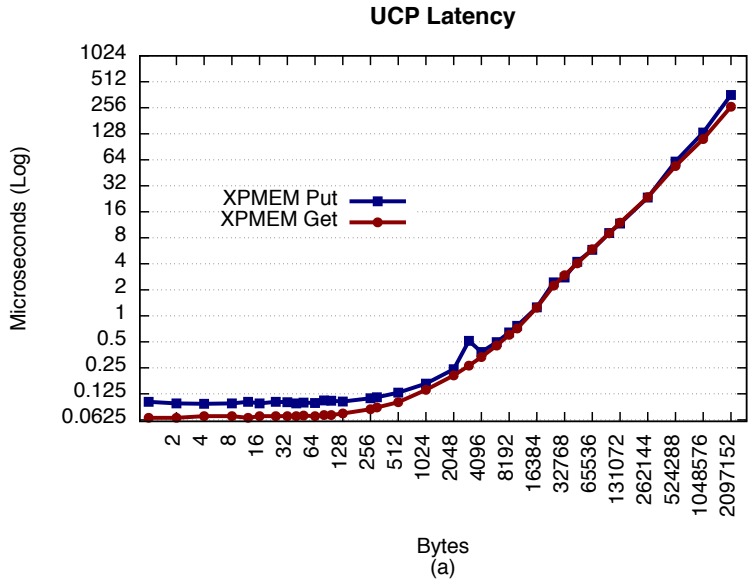
	XPMEM	KNEM	CMA	MMAP
System Call	No	Yes	Yes	No
Low Latency	V	X	X	V
High Bandwidth	V	V	V	V
Any Virtual Address	V	V	V	X
Open Source	V	V	V	V
Upstream Linux	X	X	V	V

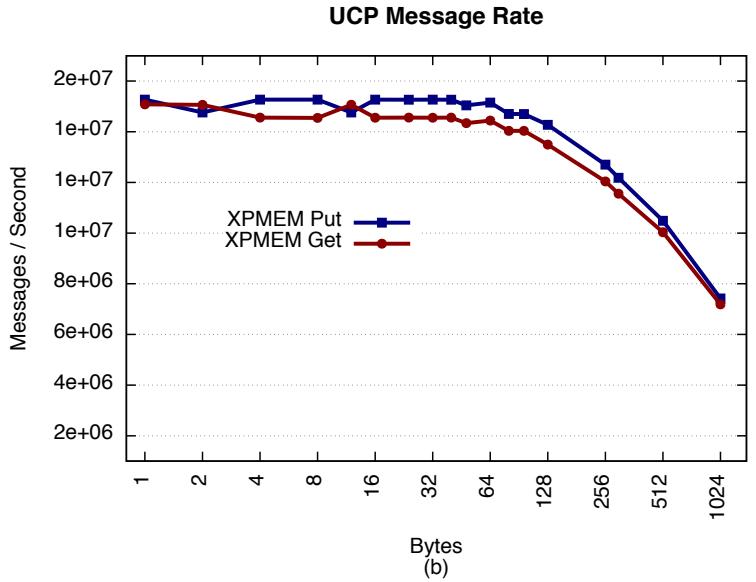
XPMEM - Cross Memory Attach

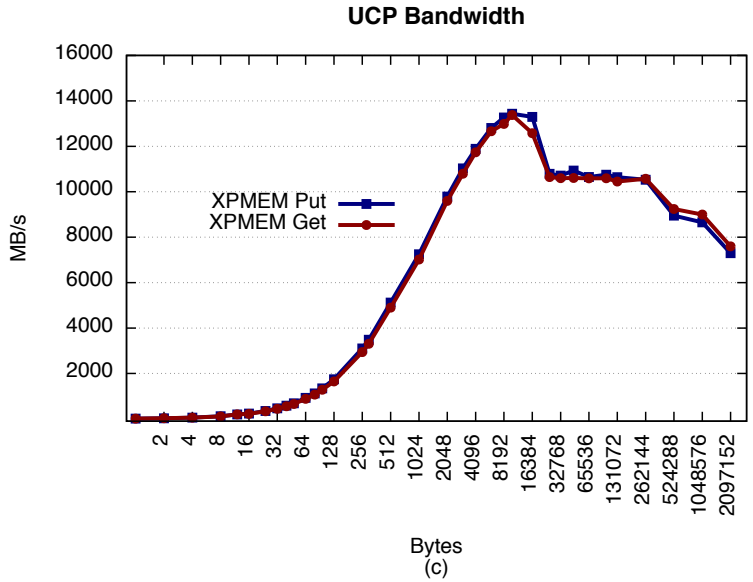
- The code was updated to compile and run on ARMv8 (https://github.com/hjelmn/xpmem)
- 3rd party kernel module based on SGI/Cray XPMEM and maintained by LANL

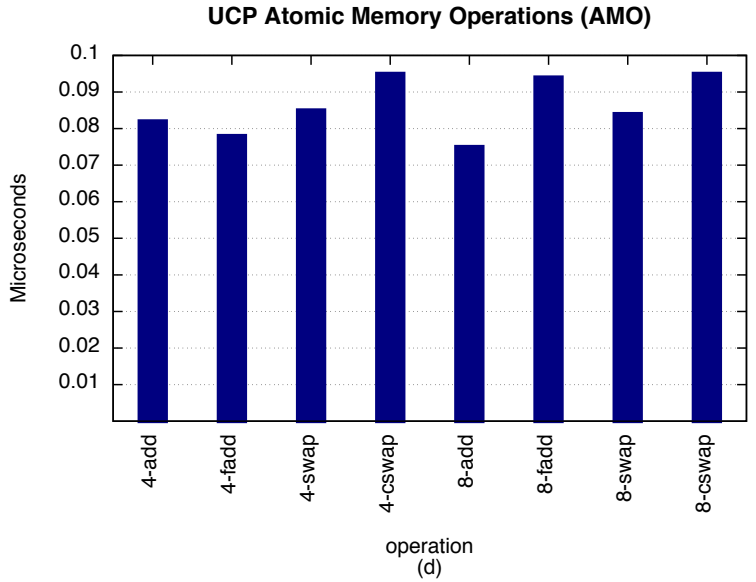
arm

OpenUCX: XPMEM

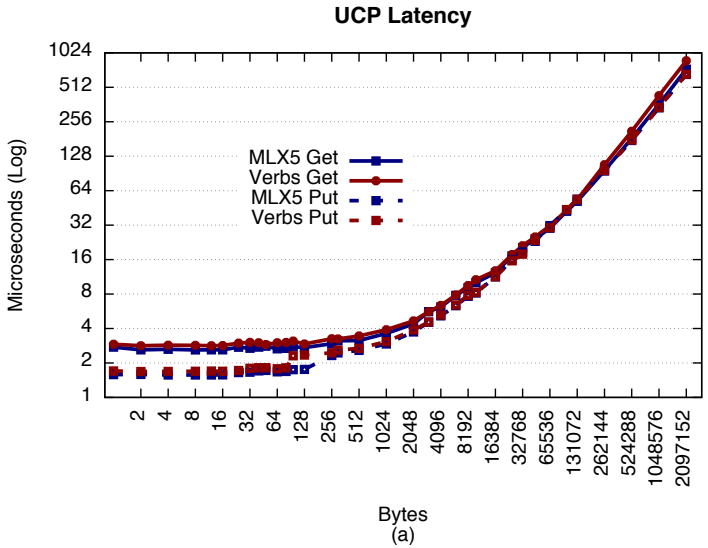


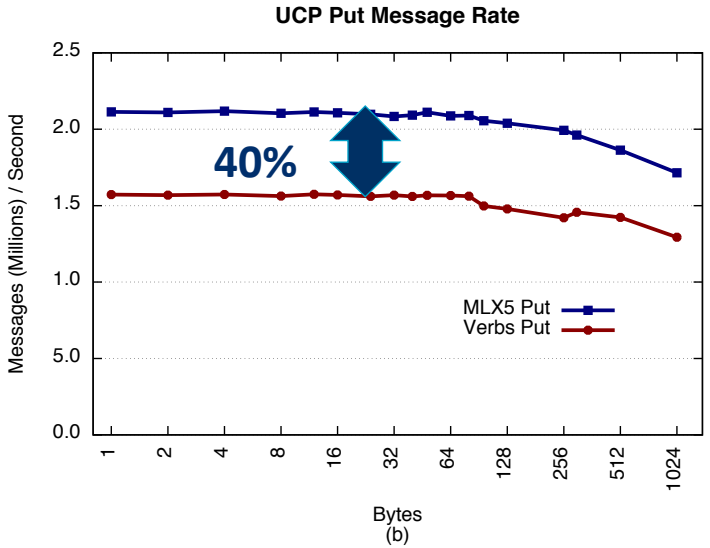


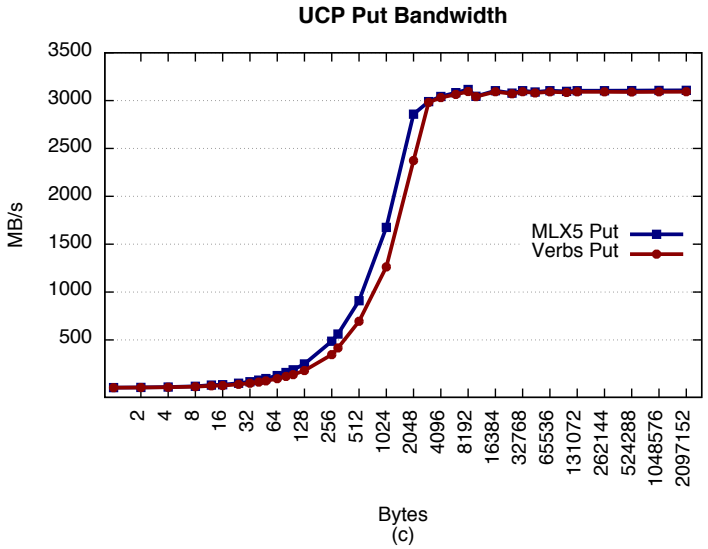


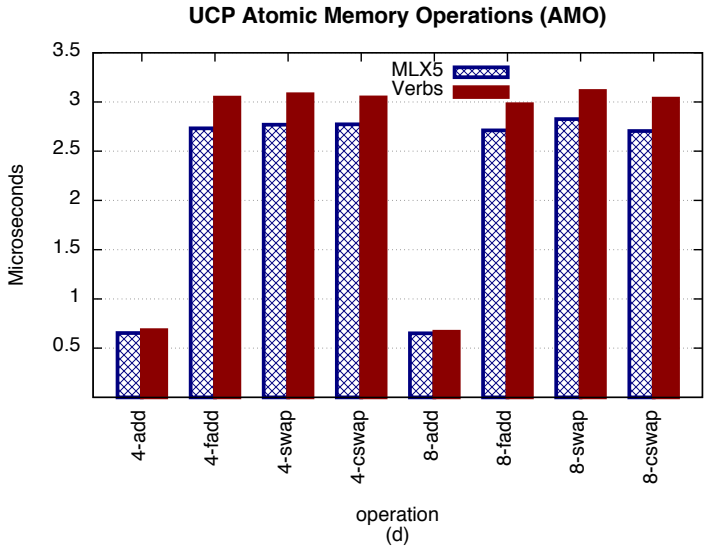


OpenUCX IB: MLX5 vs Verbs



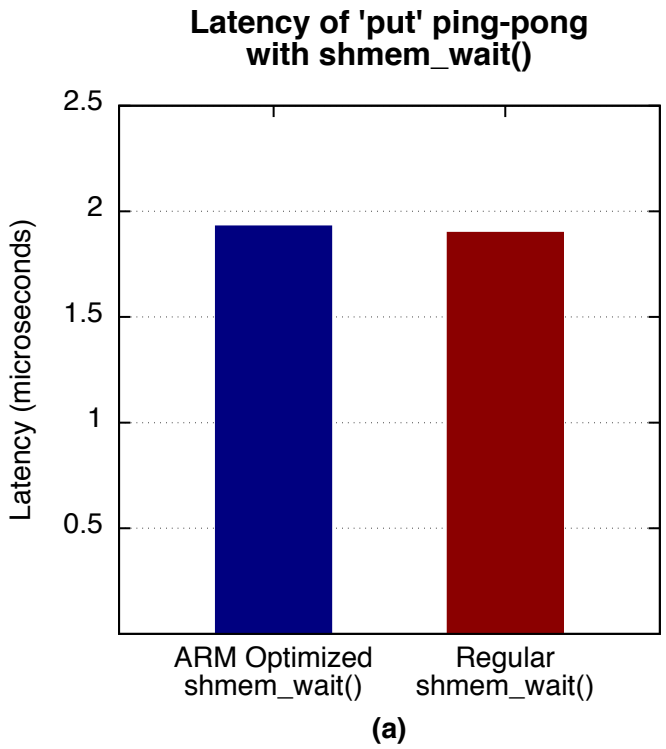


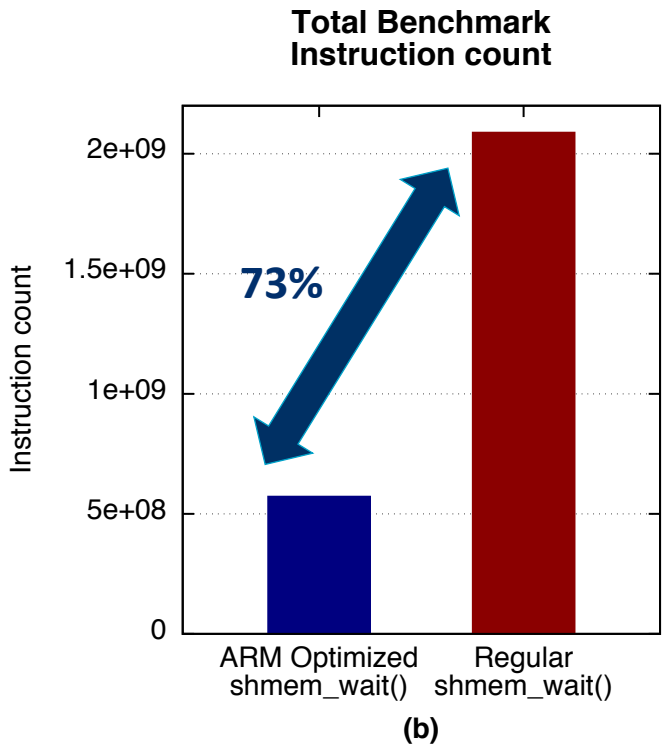


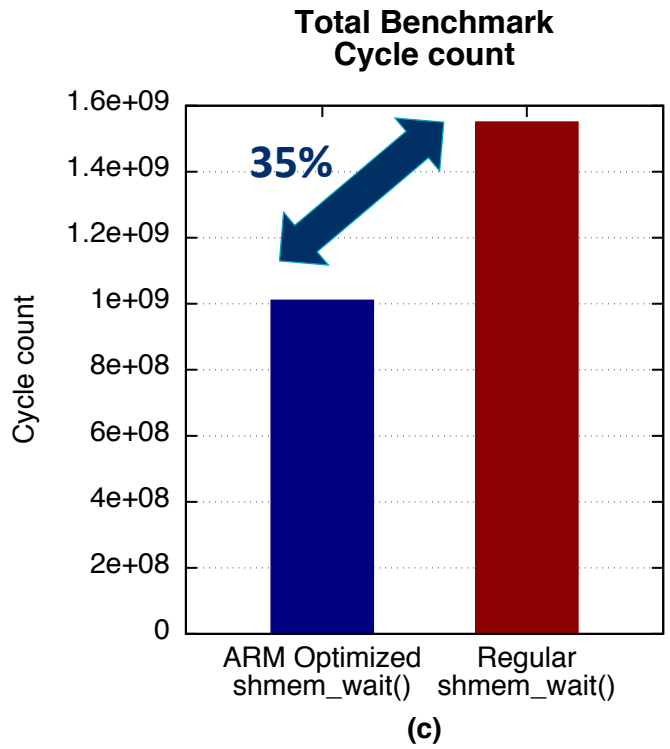


arm

SHMEM_WAIT()

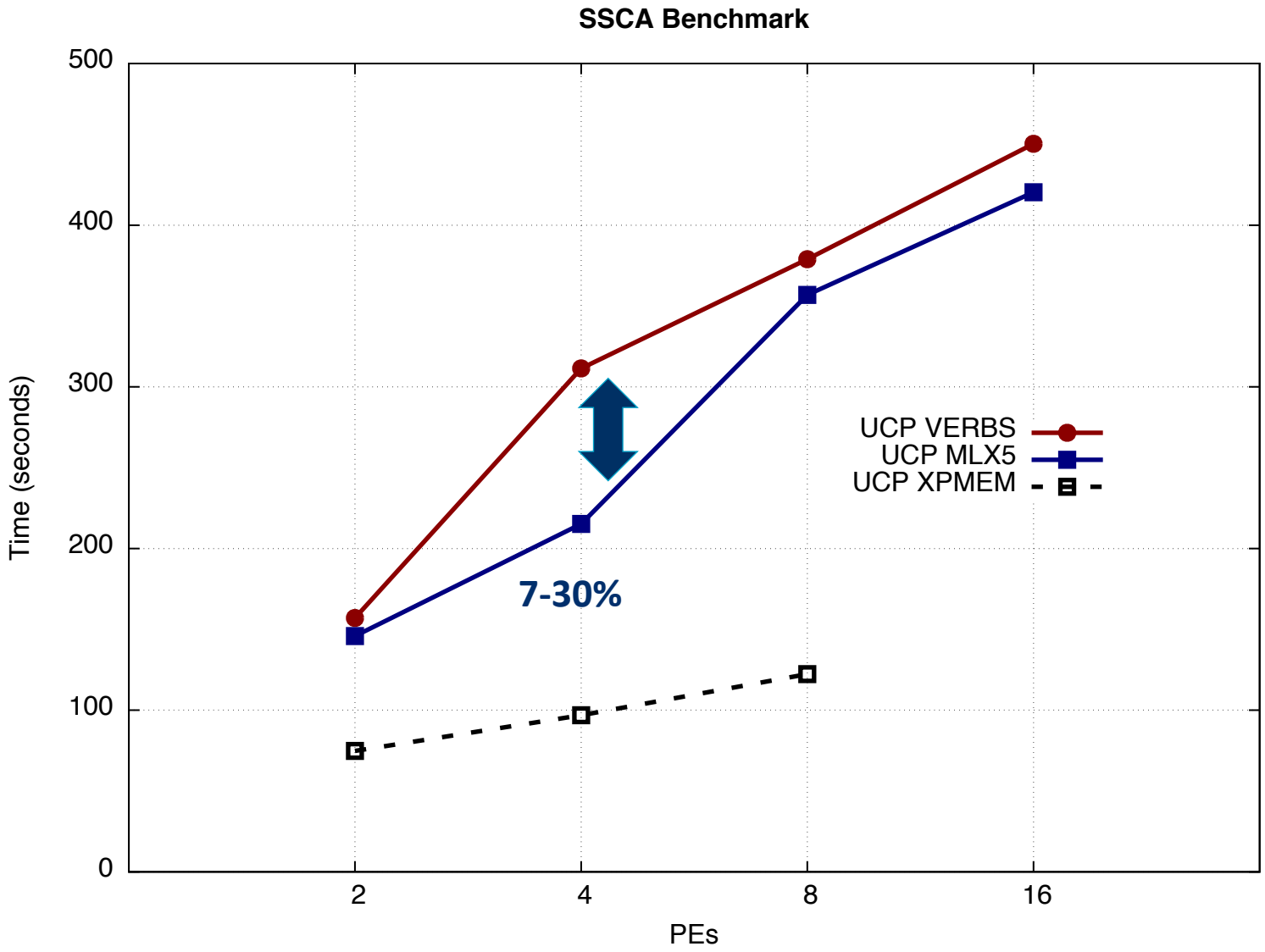






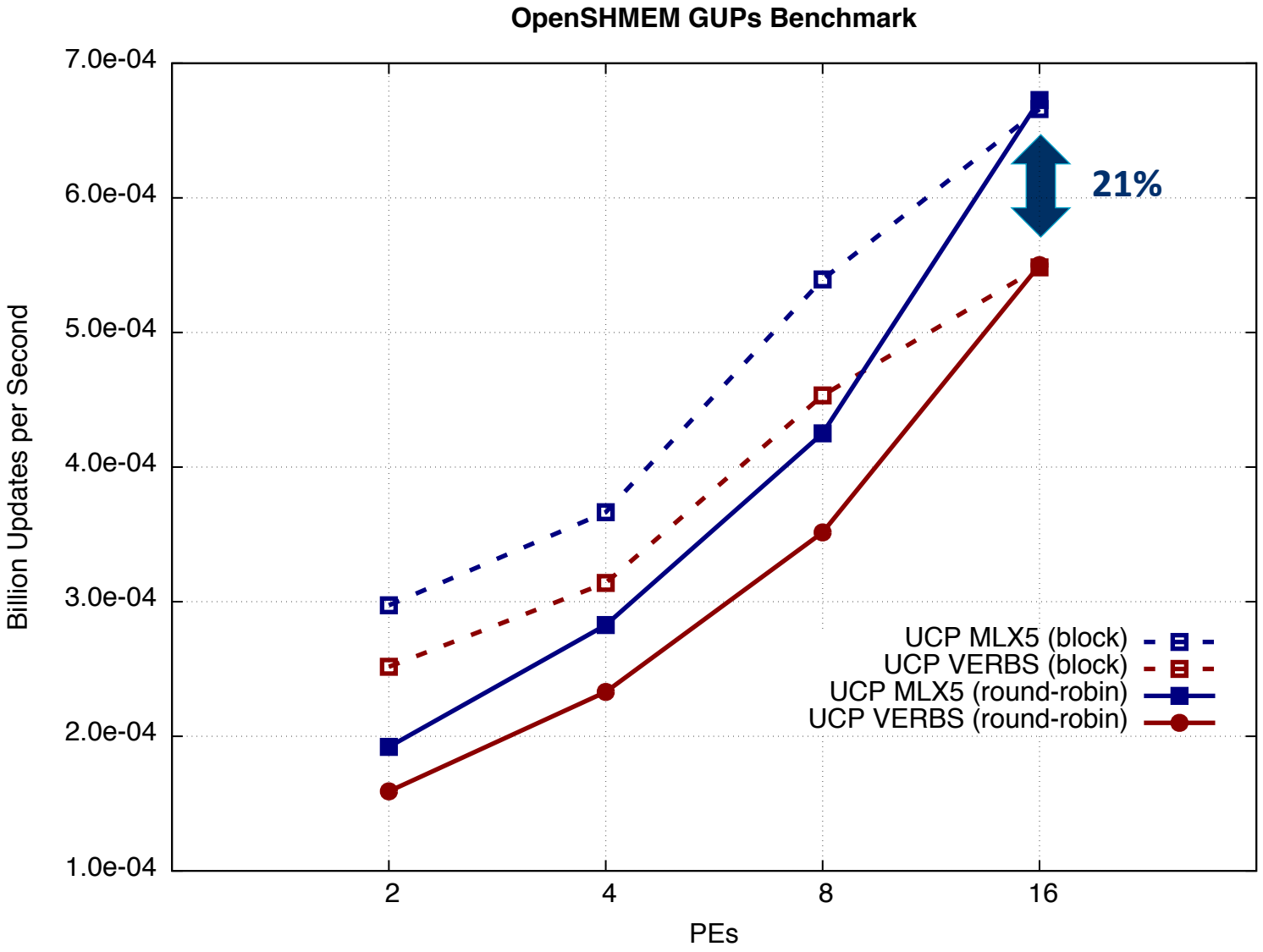
arm

OpenSHMEM SSCA



arm

OpenSHMEM GUPs



arm

Developer website : www.arm.com/hpc

A HPC-specific microsite

This is home to our HPC ecosystem offering:

- technical reference material
- how-to guides
- latest news and updates from partners
- downloads of HPC libraries
- third-party software recommendations
- web forum for community discussion and help

Participate and help drive the community

arm