UBM

CCIX: a new coherent multichip interconnect for accelerated use cases

Jeff Defilippi, Senior Product Manager, Arm

Millind Mittal, Sr Director FPGA Architect, Xilinx

Jon Masters, Computer Architect, Red Hat

#ArmTechCon

© 2017 Arm Limited

arm TechCon

Interconnects for different scale

SoC interconnect

- Connectivity for on-chip processor, accelerator, IO and memory elements.

Server node interconnect - ‘scale-up’

- Simple multichip interconnect (typically PCIe) topology on a PCB motherboard with simple switches and expansion connectors.

Rack interconnect - ‘scale-out’

- Scale-out capabilities with complex topologies connecting 1000’s of server nodes and storage elements.

© 2017 Arm Limited

Multichip capability landscape

Coherent ↔ Non-coherent

Scale Up ↔ Scale Out

Short Reach ↔ Long Reach

Direct Connected ↔ Switch Connection

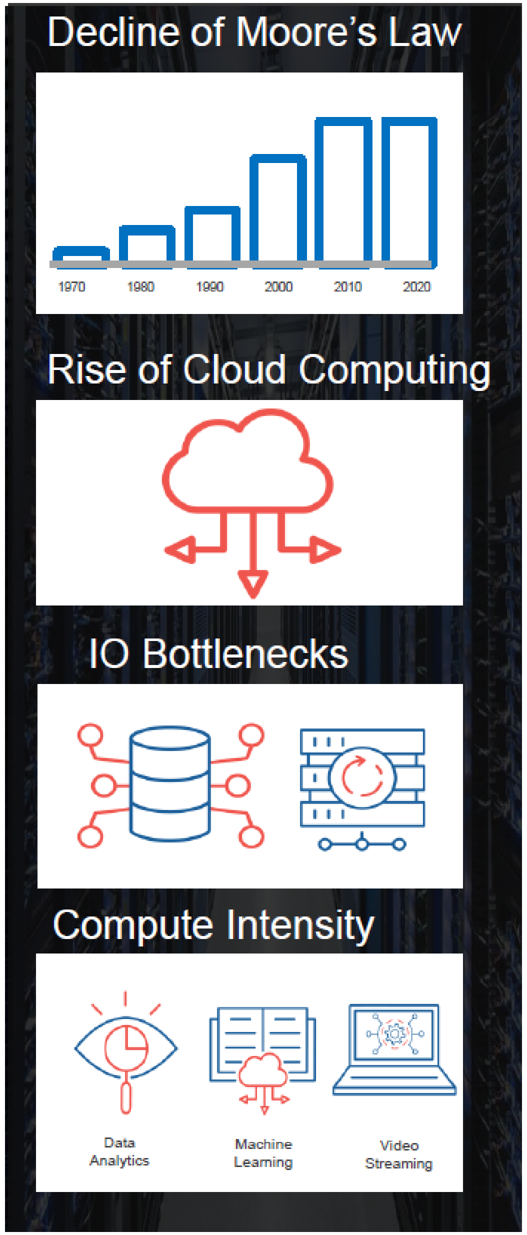
Shared Memory ↔ Message Passing

Hardware Terminated ↔ Software Terminated

© 2017 Arm Limited

Key drivers for interconnect technology

- Decline of Moore's law forcing more heterogeneous compute
- Big data analytics growing at 11.7% CAGR
- 5G wireless applications requiring 10x more bandwidth, 10x lower latency by 2021
- Increase in distributed data forcing more network intelligence at faster data rates (10GbE -> 100GbE -> 400GbE)
- Data bandwidth and sharing growth projected at 10x-50x increase vs present PCIe by 2021



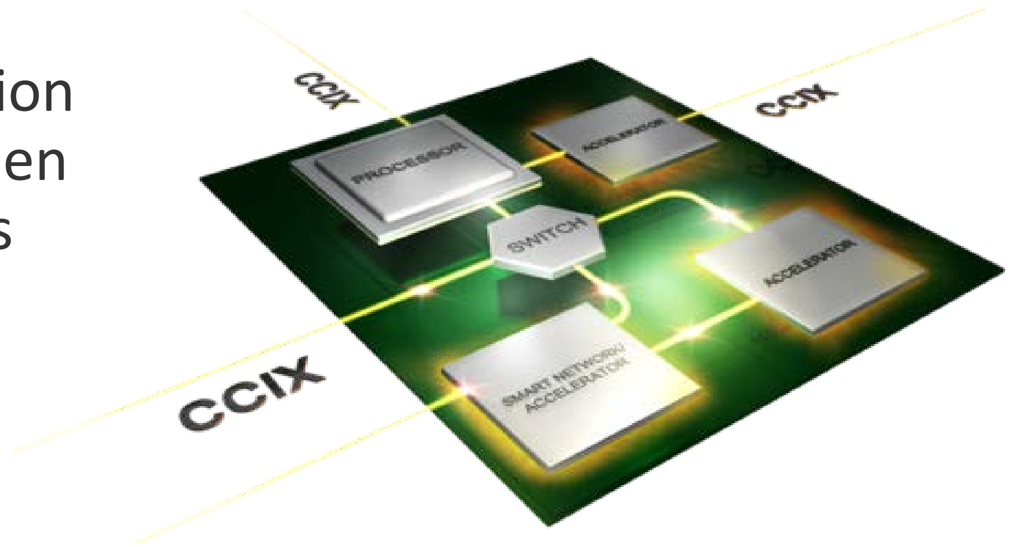
© 2017 Arm Limited

CCIX^{TM} cache coherent interconnect for accelerators

New class of interconnect for accelerated applications

Mission of the CCIX Consortium is to develop and promote adoption of an industry standard specification to enable coherent interconnect technologies between general-purpose processors and acceleration devices for efficient heterogeneous computing.

https://www.ccixconsortium.com/



© 2017 Arm Limited

CCIX Consortium Inc

- Formed January 2016, incorporated in February 2017
- Complete ecosystem with 38 members and growing
- Hardware specification available for design starts for member companies
- CCIX pronounced: (c’ siks)

Promoters

Contributors

Adopters

Arteris, Inc. Guizhou Huaxintong Semiconductor Technology Co. Ltd.
INVECAS INC Netronome Phytium Technology Co., Ltd. PLDA
Shanghai Zhaoxin Semiconductor Co., Ltd. Silicon Laboratories Inc.
SmartDV Technologies India Private Ltd.

© 2017 Arm Limited

Applications benefiting from CCIX

4G and 5G base station

Data-center Search

Embedded Computing

High Performance Computing / Supercomputing

In memory database processing

Intelligent network acceleration

Machine / Deep Learning

Mobile Edge Computing

Video analytics

© 2017 Arm Limited

CCIX multichip connectivity

High performance, low latency

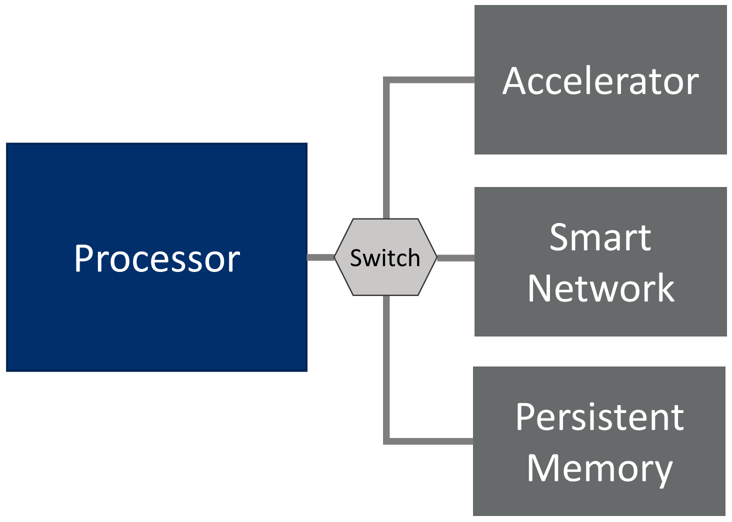
- CCIX defines 25GT/s (3x performance*)
- Examining 56GT/s (7x performance*) and beyond
- Enabling low latency via light transaction layer

Flexible, scalable interconnect topologies

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

Seamless integration

- Runs on existing PCIe transport layer and management stack
- Supports all major instruction set architectures (ISA)



© 2017 Arm Limited

Coherent virtual memory eliminates data transfer overhead

Non-coherent system without Shared Virtual Memory (SVM)

Software must manage cache maintenance and data copying

Cache coherent system with Shared Virtual Memory (SVM)

Hardware managed cache maintenance, shared address space with direct memory access

© 2017 Arm Limited

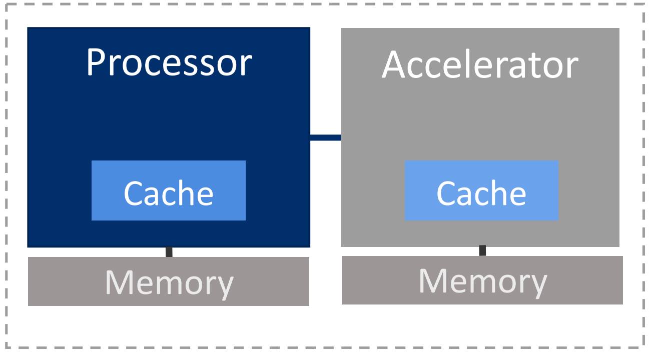
Benefits of virtualized, coherent accelerators

Simplified software development, eliminates difficult debug issues

Improved efficiency with true peer-processing and simpler mapping of job to processing element

Reduced latency translating to more transactions per second, faster response time

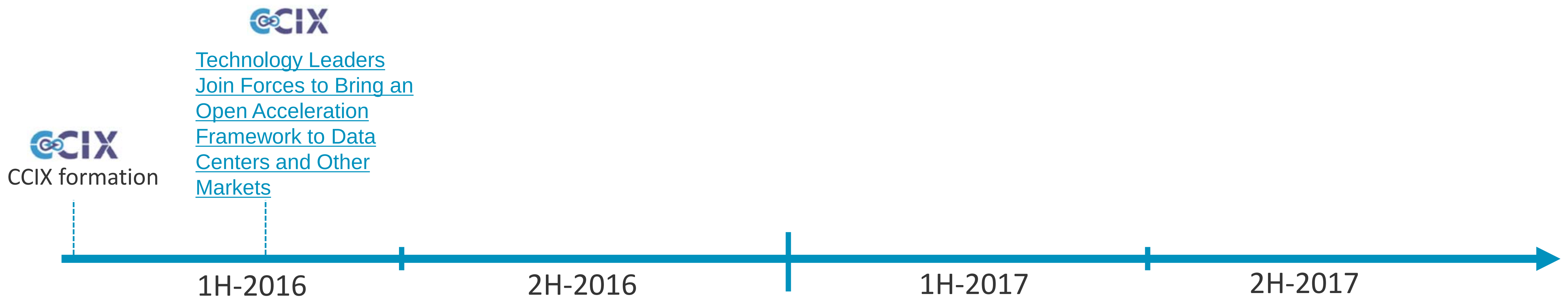
Improved fine grain data sharing, shared table updates with non-blocking, free flowing data transfers



Shared virtual memory

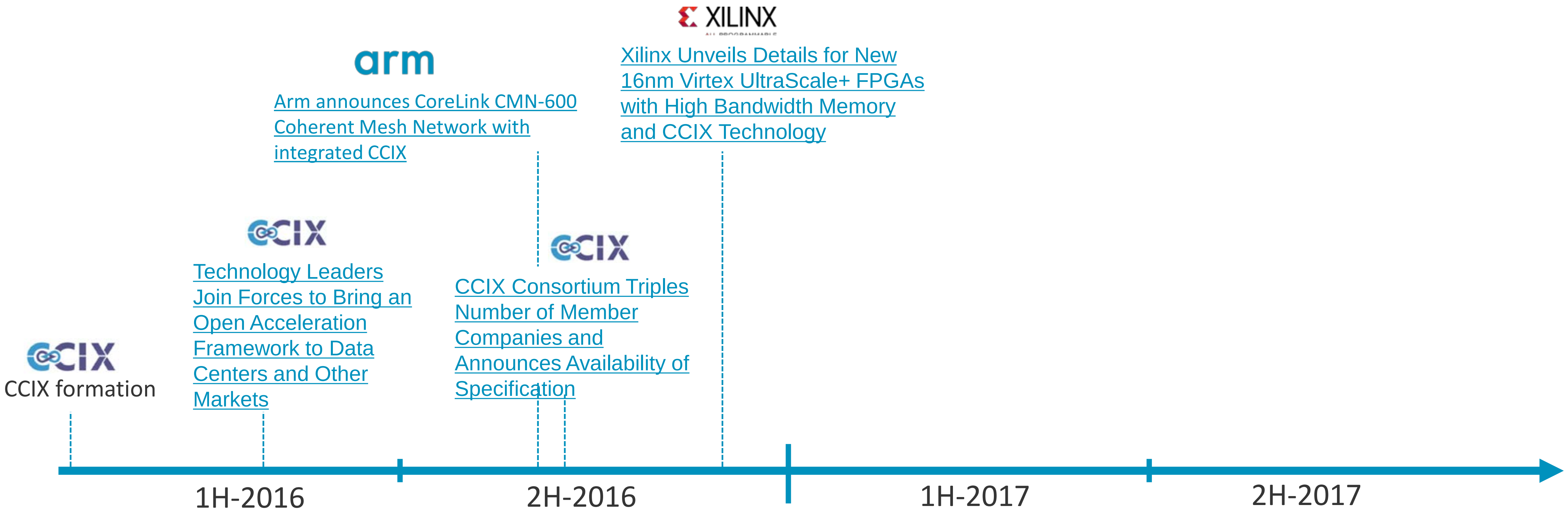
© 2017 Arm Limited

CCIX formation to ecosystem in record time



© 2017 Arm Limited

CCIX formation to ecosystem in record time



© 2017 Arm Limited

CCIX formation to ecosystem in record time

1H-2016

- CCIX formation
- Technology Leaders Join Forces to Bring an Open Acceleration Framework to Data Centers and Other Markets

2H-2016

- Arm announces CoreLink CMN-600 Coherent Mesh Network with integrated CCIX
- CCIX Consortium Triples Number of Member Companies and Announces Availability of Specification
- Xilinx Unveils Details for New 16nm Virtex UltraScale+ FPGAs with High Bandwidth Memory and CCIX Technology

1H-2017

- CCIX incorporation
- CCIX Tech Demo Proves 25Gbps Performance over PCIe
- Cadence Introduces First Interface and Verification IP Solution for CCIX to Advance New Class of Datacenter Servers
- Synopsys' Complete CCIX IP Solution Enables Cache Coherency for High-Performance Cloud Computing SoCs
- Avery Design Systems Targets Accelerator Applications With Verification Solutions for CCIX, AMBA 5 CHI, and PCIe 4.0

2H-2017

© 2017 Arm Limited

CCIX formation to ecosystem in record time

1H-2016

- CCIX formation
- Technology Leaders Join Forces to Bring an Open Acceleration Framework to Data Centers and Other Markets

2H-2016

- Arm announces CoreLink CMN-600 Coherent Mesh Network with integrated CCIX
- CCIX Consortium Triples Number of Member Companies and Announces Availability of Specification
- Xilinx Unveils Details for New 16nm Virtex UltraScale+ FPGAs with High Bandwidth Memory and CCIX Technology

1H-2017

- CCIX incorporation
- CCIX Tech Demo Proves 25Gbps Performance over PCIe
- Cadence Introduces First Interface and Verification IP Solution for CCIX to Advance New Class of Datacenter Servers
- Synopsys' Complete CCIX IP Solution Enables Cache Coherency for High-Performance Cloud Computing SoCs

2H-2017

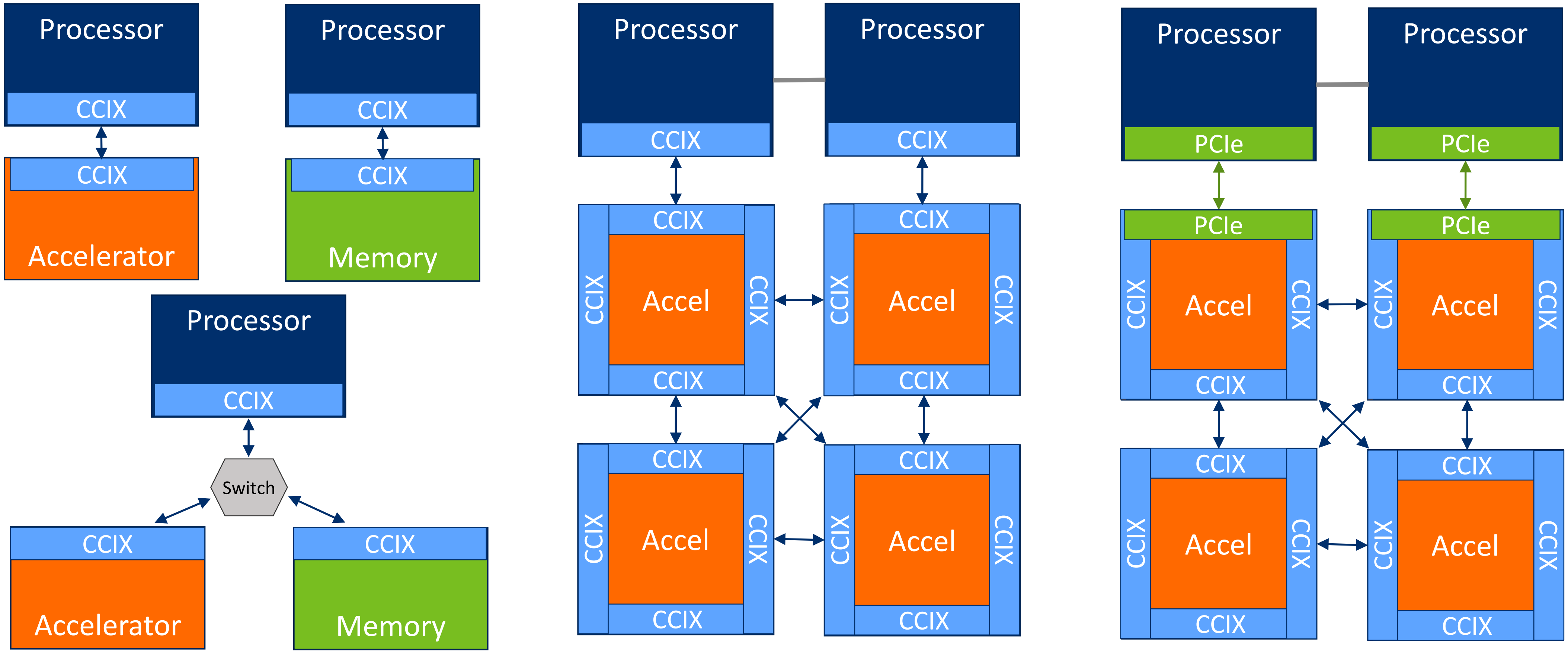
- Avery Design Systems Targets Accelerator Applications With Verification Solutions for CCIX, AMBA 5 CHI, and PCIe 4.0
- Xilinx, Arm, Cadence, and TSMC Announce World's First CCIX Silicon Demonstration Vehicle in 7nm Process Technology
- CCIX production specification

© 2017 Arm Limited

Hardware architecture

© 2017 Arm Limited

System topology examples

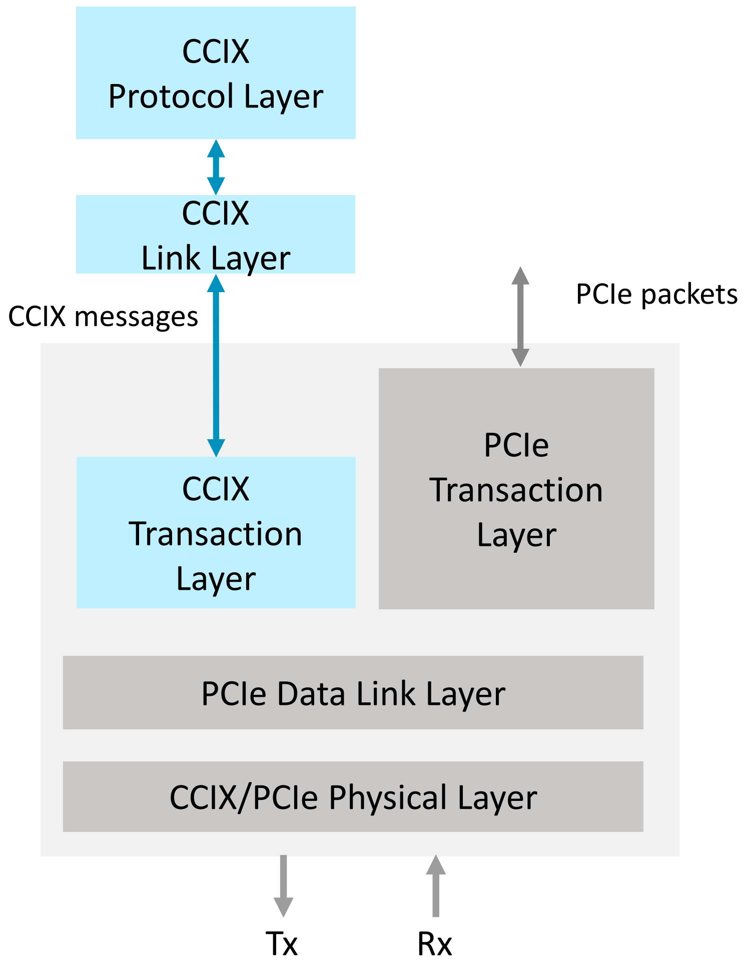


Direct attached, daisy chain, mesh and switched topologies

© 2017 Arm Limited

CCIX layered architecture

- Protocol Layer – coherency protocol, memory read & write flows
- Link Layer – formats CCIX messages for target transport
- Transaction Layer – Adds optimized packets, manages credit based flow control
- Physical Layer – Dual mode PHY to support extended data rates

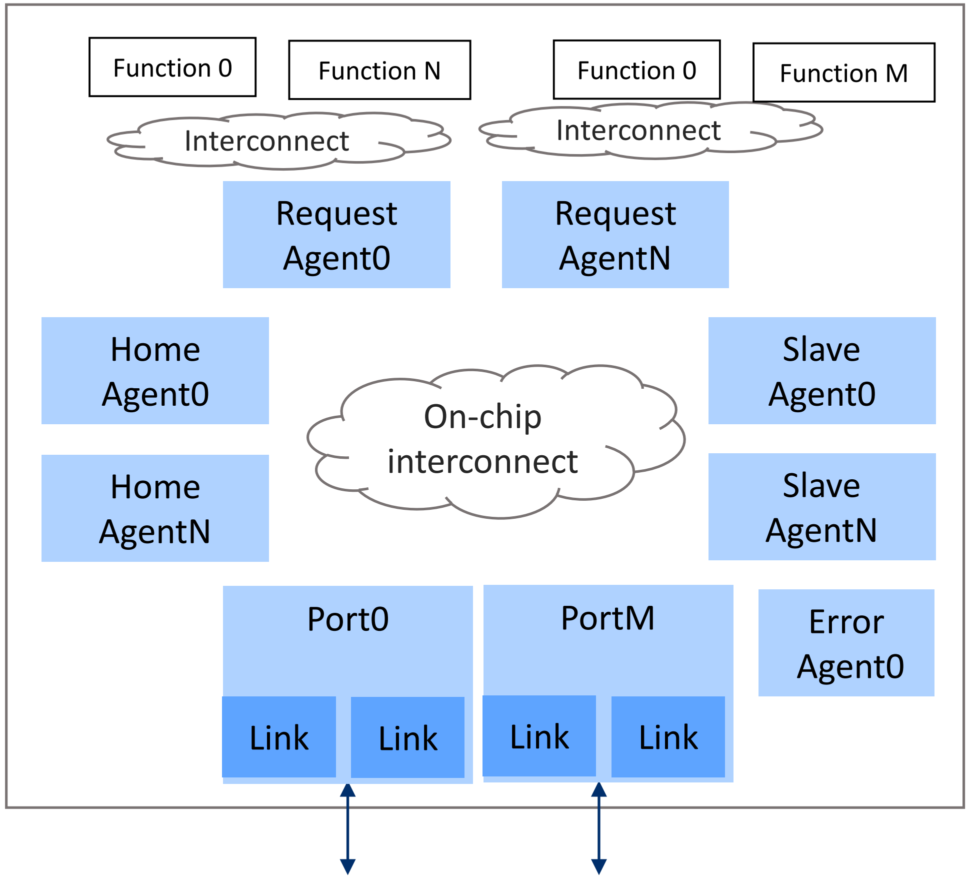


© 2017 Arm Limited

CCIX coherency layer architecture model

- Portable protocol to other transports
- Support for port aggregation, multiple link agents
- CCIX agent types:
  - Request Agent (RA) - single (implementation specific) function or proxy for multiple functions
  - Home Agent (HA) - point of coherency for a given address
  - Slave Agent (SA) - used for memory expansion
  - Error Agent (EA) – receives and processes protocol error messages

CCIX protocol agents



© 2017 Arm Limited

CCIX coherency protocol

CCIX provides a simple mapping to Arm AMBA CHI

Optional support for partial cache states

Supported CCIX transactions

- Read and writes
- Atomics
- Cache maintenance including persistence

Cache States	
I	Invalid
UC	Unique Clean
UCE	Unique Clean Empty
UD	Unique Dirty
UDP	Unique Dirty Partial
SC	Shared Clean
SD	Shared Dirty

© 2017 Arm Limited

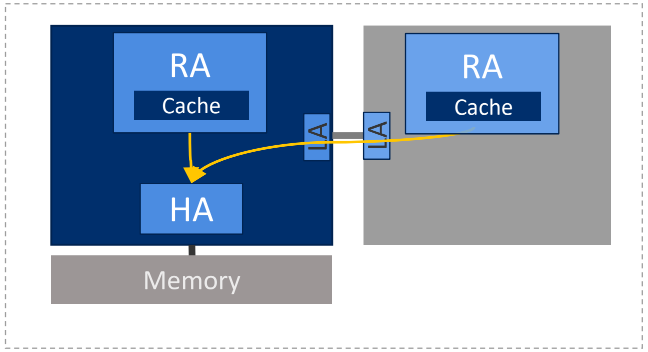
CCIX optimization for multichip

- Header format - options for PCIe or optimized versions
- Eliminate messages where possible (ex: no compACK)
- Message packing - combine multiple CCIX messages in a single packet
- Request and Snoop Chaining - chain request to the subsequent address of the previous message
- Port aggregation – increase bandwidth by aggregated multiple CCIX ports

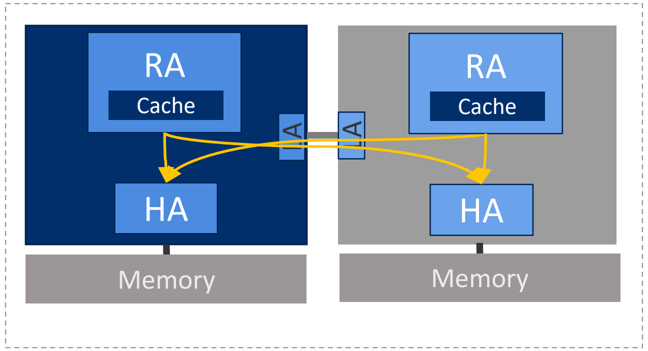
© 2017 Arm Limited

CCIX example request to home data flows

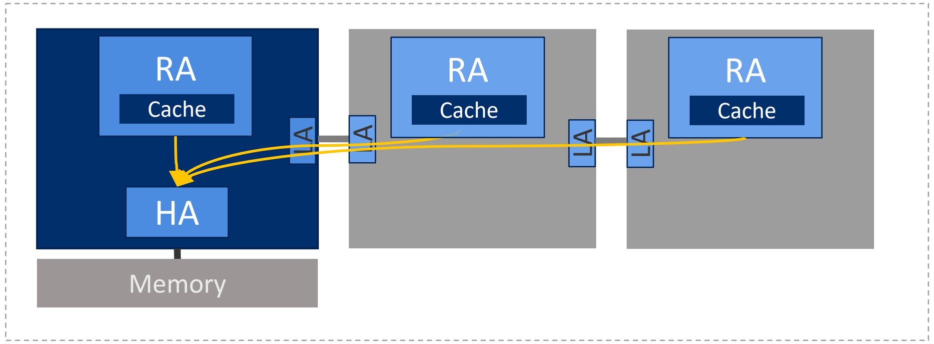
Accelerator shares processor memory



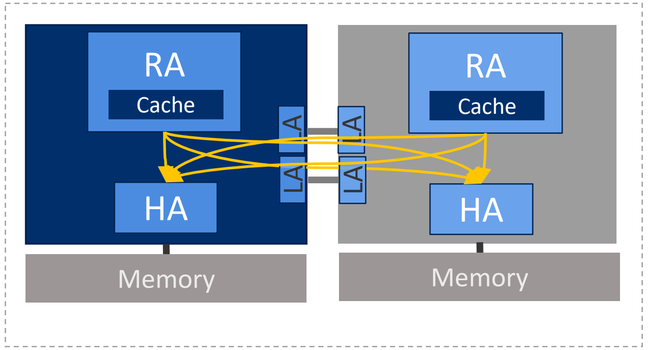
Shared processor and accelerator memory



Daisy chain to shared processor memory



Shared memory with aggregation



© 2017 Arm Limited

CCIX port aggregation to boost bandwidth and transactions

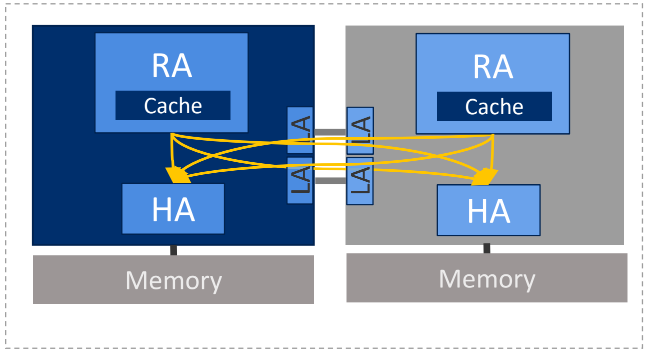
CCIX defines a hashing function to steer requests across multiple links

Aggregation effectively multiplies the bandwidth

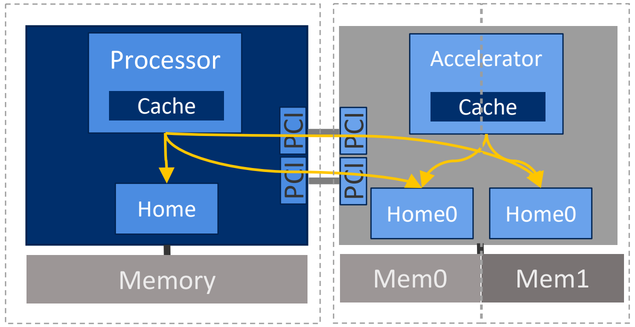
Aggregation could also be used to increase number of transactions (eg 50GT/s vs 25GT/s)

PCIe requires separate address spaces, requests can not be hashed

CCIX with Port Aggregation



PCIe with Aggregation

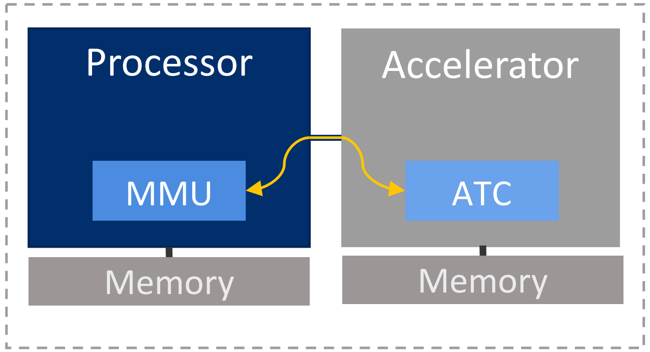


© 2017 Arm Limited

arm TechCon

Shared virtual memory with translation service

- CCIX expands the host centric SVM to include accelerator attached memory as part of system memory
- PCIe Address Translation Service is used for VA to PA translation
  - Use of ATS makes translation service ISA independent
- Translation service request is enhanced to provide additional memory attributes options than current PCIe specification
  - Attributes types defined are- WB with “no LLC allocate” hint , WB with “LLC allocate hint”, Non-cacheable, Device nRnE, Device nRE, Device RE
- CCIX Devices are required to ensure that accelerator function can not bypass access control enforced by ATC usage



Shared virtual memory

© 2017 Arm Limited

CCIX 25Gbps PHY technology

- 3X faster transfer speed with CCIX vs existing PCIe Gen3 solutions
- Transferring of a data pattern at 25 Gbps between two FPGAs
- Channel comprised of an Amphenol/FCI PCI Express CEM connector and a trace card
- Transceivers are electrically compliant with CCIX
- Fastest data transfer between accelerators over PCI Express connections

Xilinx and Amphenol FCI first public CCIX technology demo

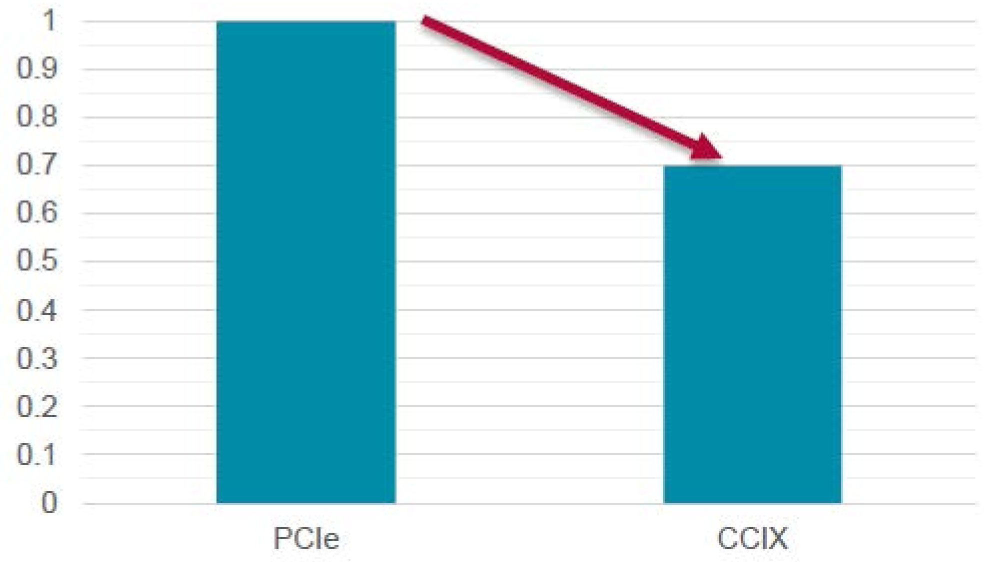
https://forums.xilinx.com/t5/Xcell-Daily-Blog/CCIX-Tech-Demo-Proves-25Gbps-Performance-over-PCIe/ba-p/767484

https://youtu.be/JpUSAcnn7VA

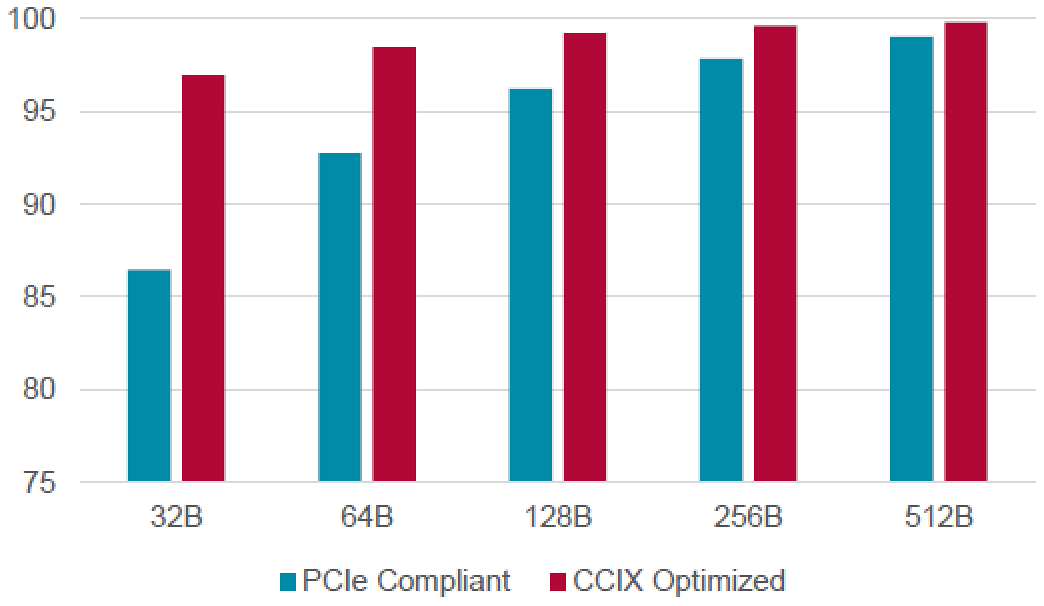
© 2017 Arm Limited

Improved efficiency with CCIX transaction layer

Reduced latency with light weight transaction layer



Improved packet efficiency with optimized CCIX header



© 2017 Arm Limited

Software architecture

© 2017 Arm Limited

DMA Engines: The problem with traditional accelerators

- Operating System vendors are interested in the opportunity for workload-optimized accelerators
- Traditional DMA approach is to provide a special (Linux) kernel driver for every unique accelerator
    - Requires skilled kernel developers (a driver for each accelerator), failure mode is catastrophic (system crash/downtime)
- Operating Systems used tomorrow have already been deployed. Updates are 9-12 months apart
    - Drivers must be in “upstream” Linux before we support them, a year+ turnaround for every accelerator

'Trilby”: DMA Engine driven FPGA based workload accelerator built by Jon Masters for research into the barriers to adoption in the Enterprise, uses traditional approach of kernel driver and Operating System hacks.

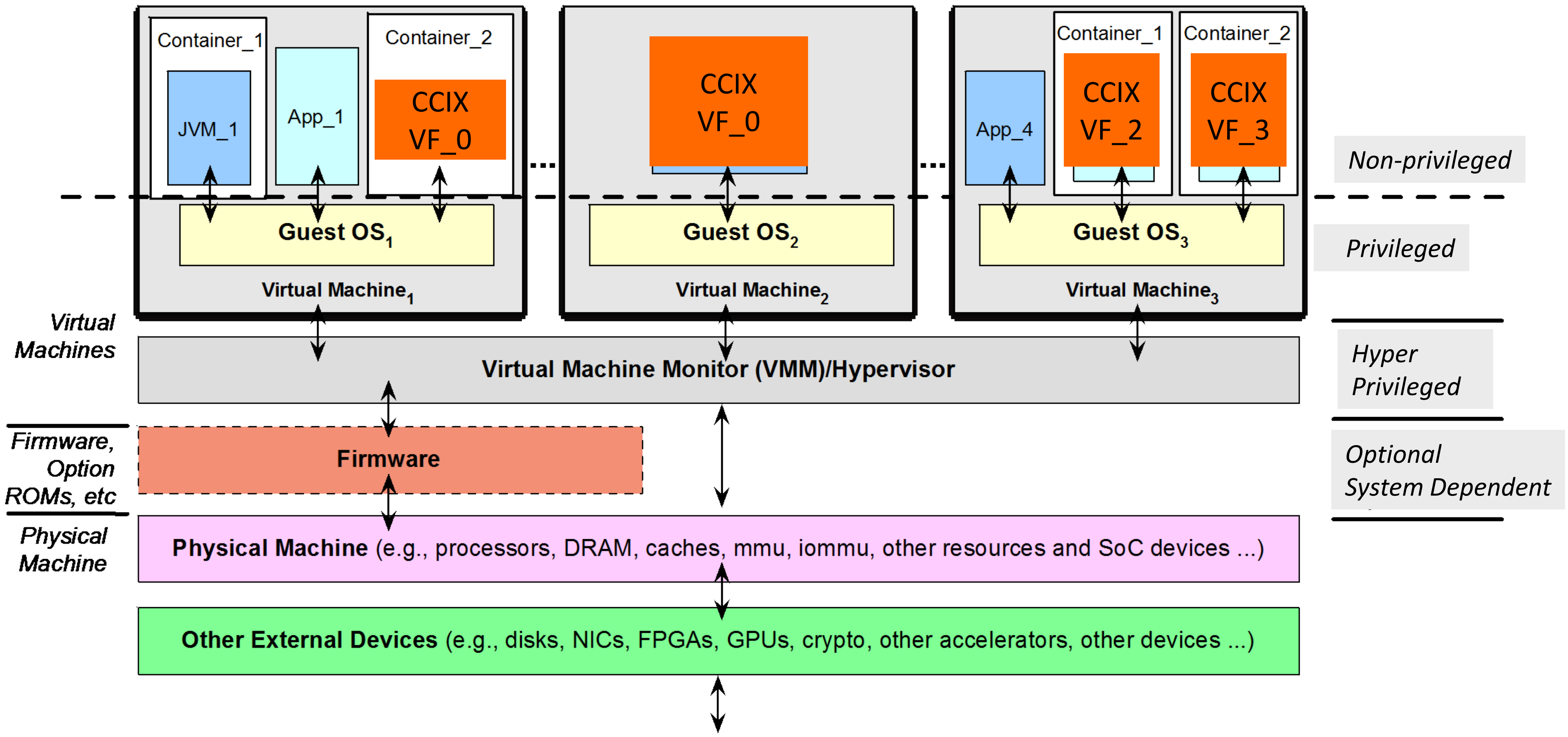
© 2017 Arm Limited

Shared Virtual Memory (Driverless) model used by CCIX

- CCIX capable devices behave similarly to nodes in existing NUMA systems
  - Memory based approach leverages existing Operating System capabilities
  - Enabled by coherent shared virtual memory – it's all “just memory”
- Minimal OS changes required, mostly for optional/enhanced capabilities
    - e.g. one OS driver for power management, firmware-first error handling, etc.
    - No Operating System drivers required for individual accelerators
- Acceleration Framework (SW framework for offloading)
  - Simple software library approach for applications running within VMs/Containers
  - Developer writes regular application software in any language with full toolset

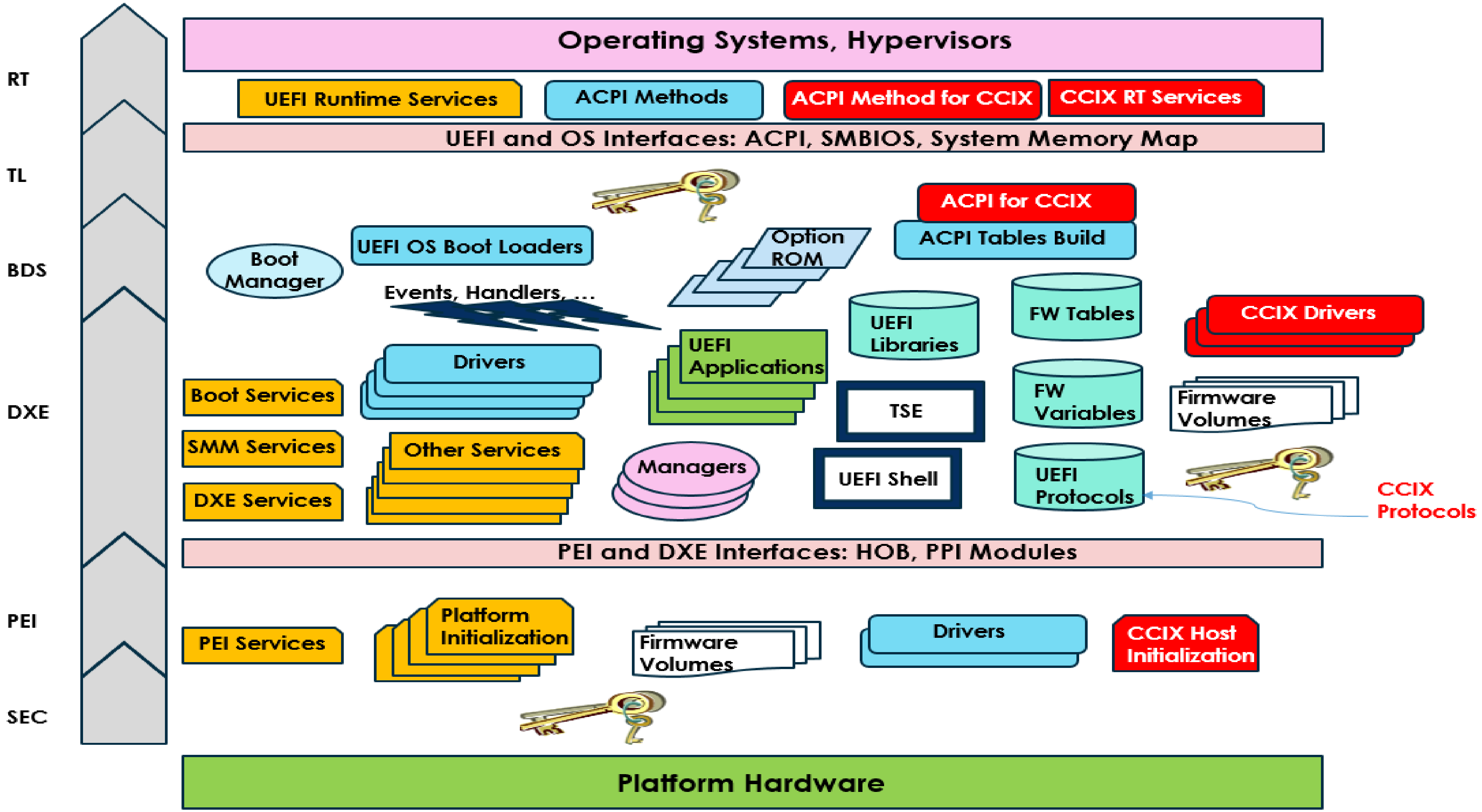
© 2017 Arm Limited

Application stack with virtual CCIX acceleration functions



© 2017 Arm Limited

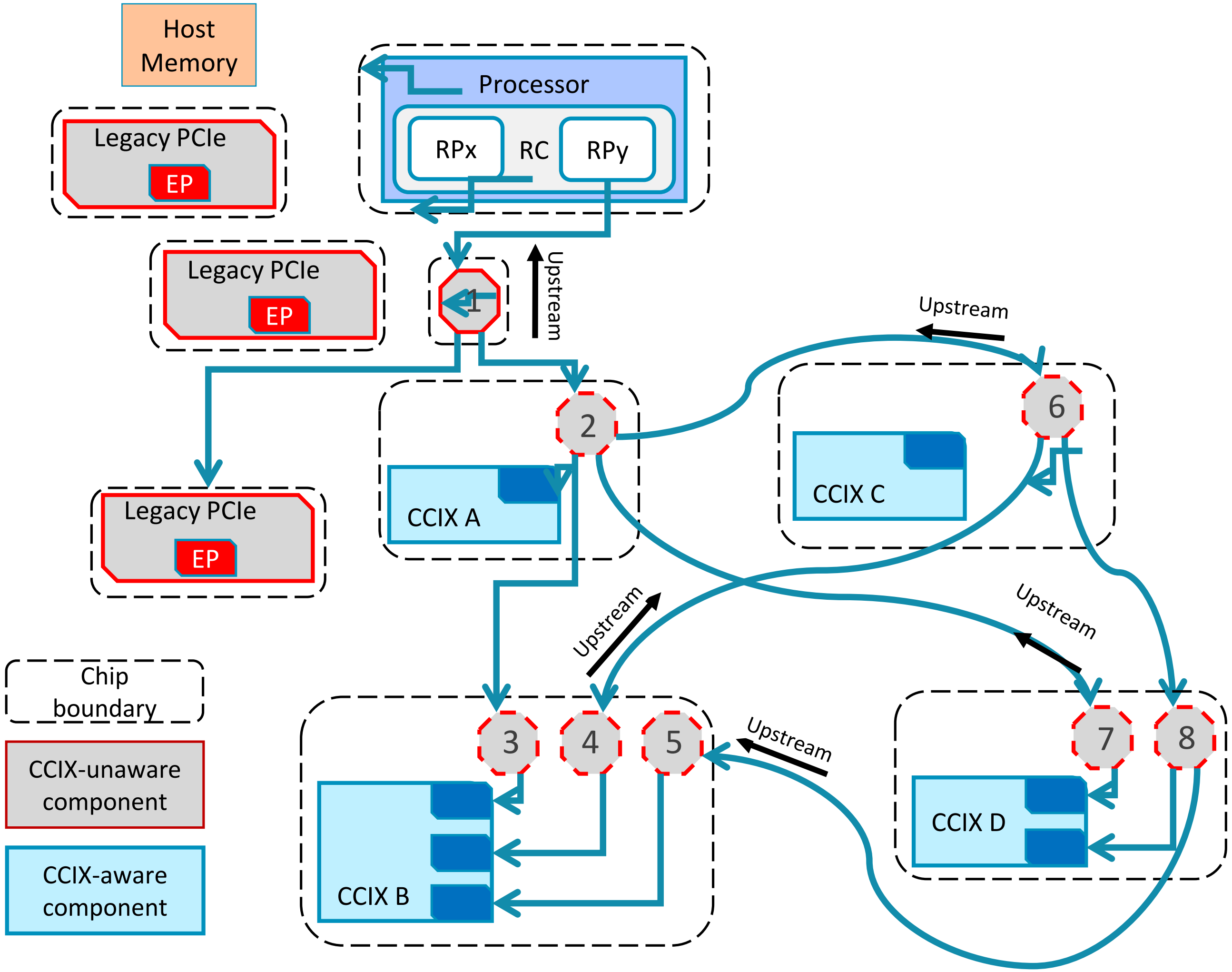
UEFI stack with CCIX extensions



© 2017 Arm Limited

CCIX discovery and initialization for complex system

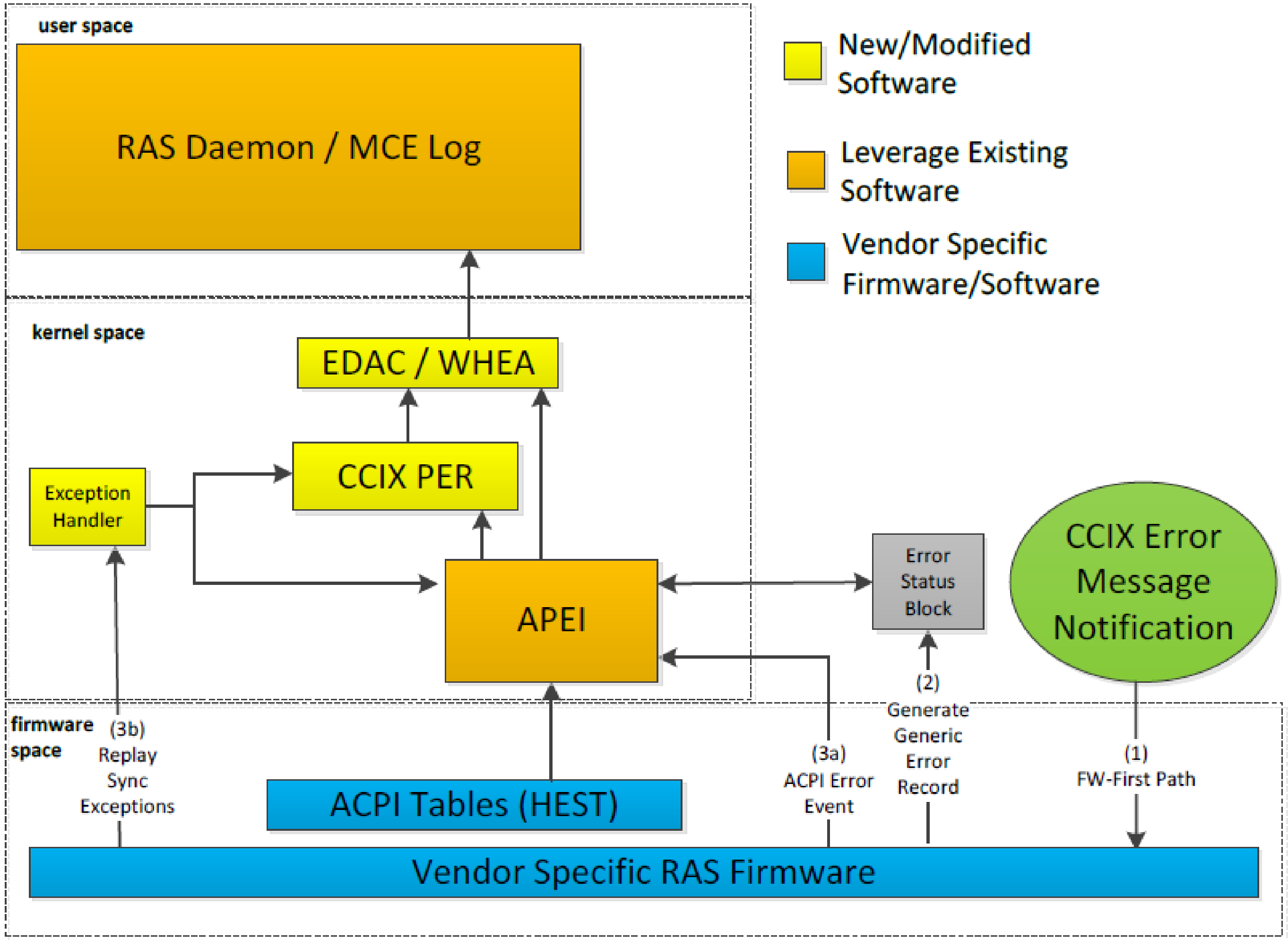
- System (UEFI) firmware enumerates and configures the CCIX topology at start of day before the OS boots
  - Walks tree assigning HAIDs/RAIDs, creates global memory tables, and programs devices
- Memory (HA/SA) mapped into global Physical Address Space (G-*SAM)
- Legacy PCIe (non-CCIX protocol aware) devices pass CCIX via Vcs
  - Top level switch can be a traditional server PCIe switch, CCIX topology drops in below



© 2017 Arm Limited

Error handling

- CCIX Errors (e.g. RAS, protocol, link credit…) signaled via PER (Protocol Error Record) message
- Handled using “firmware first” approach that allows unmodified OSes to operate. Slots into ACPI
- Operating System uses standard APEI handlers to log messages. An enlightened OS can provide greater handling if required



© 2017 Arm Limited

CCIX Software Roadmap

- We are driving the software specification for the CCIX Consortium
  - Includes DVSEC (hardware discovery/control), RAS
- Creating a firmware reference for ease of implementation
  - Firmware specification document provides guidance
  - Reference UEFI (Tianocore) firmware with ACPI
- Collaborating with industry on standardization of accelerator framework
  - Goal is a standard software library for applications (multi-language bindings)

© 2017 Arm Limited

CCIX: Seamless Acceleration

CCIX benefits accelerated applications such as machine learning, smart networks, and big data analytics with increased bandwidth, lower latency and more efficient data sharing

Shared virtual memory enables CCIX accelerator functions that just work in the cloud

Easy adoption and simplified development by leveraging today's data center infrastructure

© 2017 Arm Limited

Trademark and copyright statement
The trademarks featured in this presentation are registered and/or unregistered trademarks of CCIX Consortium, Inc in the EU and/or elsewhere. All rights reserved. All other marks featured may be trademarks of their respective owners.

Copyright © 2017 CCIX Consortium, Inc

Thank You!

© 2017 Arm Limited