#### Western Digital.

# **OmniXtend: Scalability and LPC**

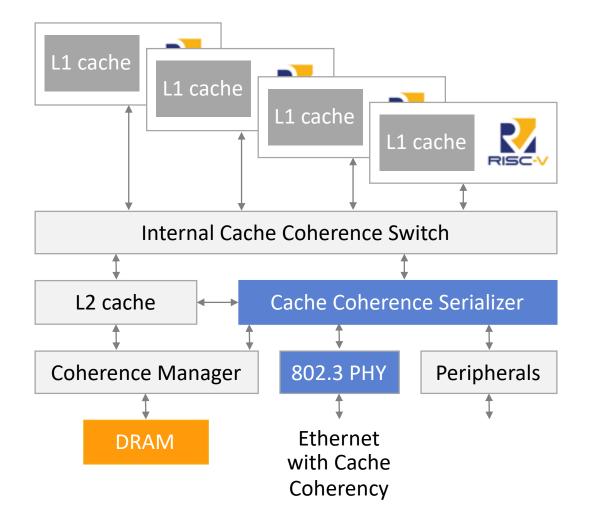
Jaco Hofmann, Tu Dang, Dejan Vucinic

### Why do we need memory disaggregation?

- Space in racks is limited
- Memory utilization is low Ethernet Switch Fast shared memory (000) Access Remote Accelerator Access Remote Memory Processor only nodes 000 000 000 000 000 000 ၀ုဝ Lowest Point of Application **Application** Application Coherence (LPC): Server 1: Only Server 2: Server 3: Memory only Processor Processor and Processor, Accelerator Memory and Memory

#### **OmniXtend Overview**

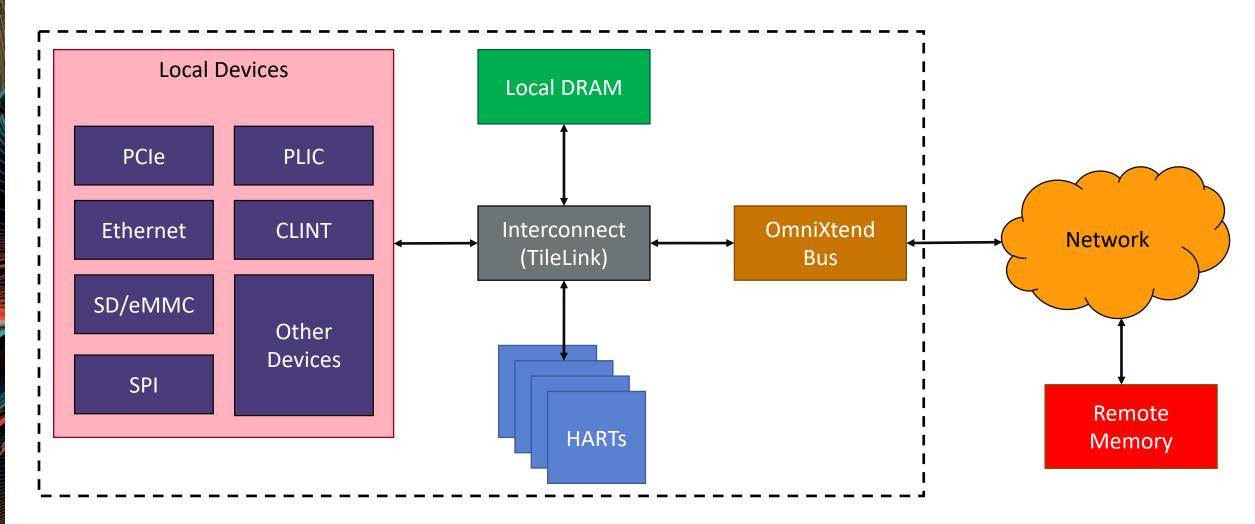
- OmniXtend is based off TileLink
  - TileLink is an open, coherent bus used to connect Cores with Memory



OmniXtend enhances TileLink and serializes it over Ethernet

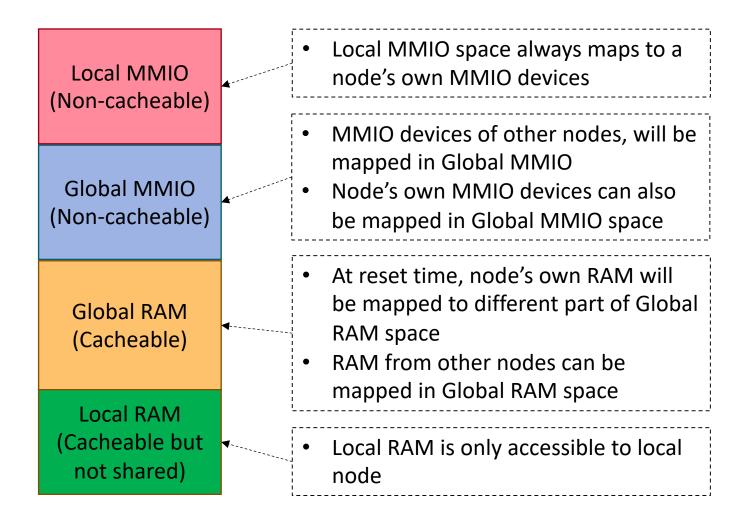
#### **An OmniXtend Compute Node**

**High-level view of each compute node** 



#### **Compute Node Address Space**

High-level view of physical address space



## **OmniXtend 1.0.3 to 1.1**

Western Digital. © 2021 Western Digital Corporation or its affiliates. All rights reserved.

## **OmniXtend 1.0.3 Features**

- What does OmniXtend provide right now?
- Cached, Uncached and Coherent Accesses
- Flow Control
- Out-of-sequence/dropped packet detection and handling

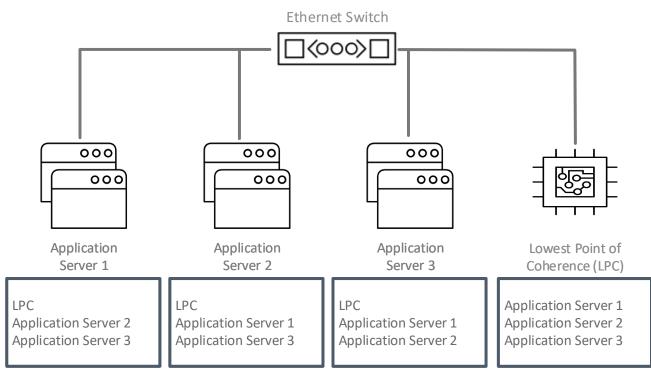
Ethernet MAC header (14 bytes)
TLoE frame header (8 bytes)
TileLink message 1
TileLink message 2
TileLink message m
Padding (P x 8 bytes)
TLoE frame mask (8 bytes)

E	Byte 7	Byte 6	Byte 5	Byte 4	Byte 3	Byte 2	Byte 1		Byte 0
63 62 61	60 59 58 57 56 55 54	53 52 51 50 49 48	47 46 45 44 43 42 41 40	39 38 37 36 35 34 33 32	31 30 29 28 27 26 25 24	23 22 21 20 19 18 17 16	5 15 14 13 12 11 10 9	8 7 6 5	4 3 2 1 0
VC	Res		Sequence_number	r	Seq	uence_number_ack	A C		Credit
(3)	(7)		(22)			(22)	к	(1) (3)	(5)

#### **OmniXtend 1.0.3 Scalability Concerns**

#### OmniXtend requires a statically set up system

- Resend/Flowcontrol mechanisms require state for each communication pair
  - 10s of sessions using SRAM, 100s to 1000s in DRAM with latency penalty



Permanent connection between all participants is not necessary

Western Digital. © 2021 Western Digital Corporation or its affiliates. All rights reserved.

### **OmniXtend 1.1 Dynamic Connections**

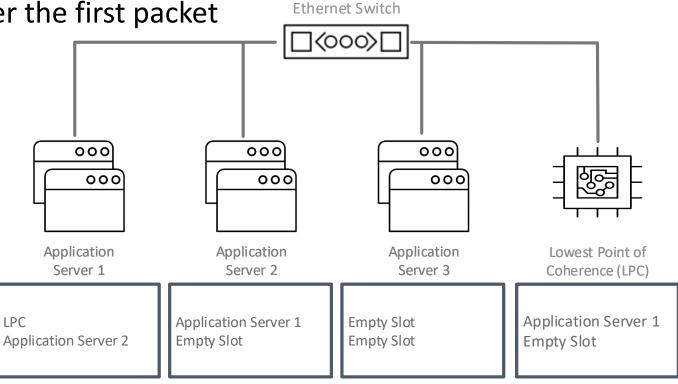
#### Goal: Connection establishment and termination based on existing OX mechanisms

$\begin{array}{c c} (2) \\ (4) \\ (2)$	(3)	Type Re			(5)
---	-----	---------	--	--	-----

- Three new message types indicated by OX header field
  - Establish Connection -> Starts with Sequence Number 0
  - Terminate Connection -> Indicate end of connection
  - ACK only

## **Connection Establishment**

- Utilizes existing fault tolerance mechanisms
  - Retry until success if communication partner does not answer
- In the best case: Zero additional latency
- No changes to the protocol after the first packet



#### **Connection Termination**

- Both parties can initiate connection termination
  - Termination can be delayed if necessary
- Termination can only be approved if there are no outstanding TileLink transactions
- Permissions for cache lines:
  - must be returned in a probe-based cache system
  - can be kept in directory-based cache systems
    - Requires connection reestablishment for permission changes

#### **Canals: Messages bypassing the resend logic**

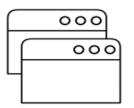
- Messages outside the fault tolerance mechanisms
- First canal message type: ACK only
  - A message that contains only an ACK for a previous message
  - Avoids congestion of the resend buffers
- Avoids a potential deadlock in high throughput, high latency scenarios
  - Both parties have full resend buffers and cannot send another ACK
  - Resend buffers remain full -> Deadlock

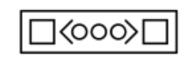
#### **OmniXtend Lowest Point of Coherence**

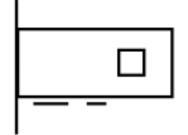
- Fully OmniXtend 1.1 compatible LPC for FPGA
- Written in Bluespec (Open-Source Compiler available)
- Designed for 10Gbit/s Ethernet
- Supports a variety of Xilinx FPGAs (using TaPaSCo for bitstream generation)
- Will be released as source and Verilog under Apache 2.0 license at Github
- Includes software implementation of the requester and full system simulation

#### **Demonstration**

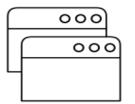
Software Requester 1







Software Requester 2



Software Requester 3 **Ethernet Switch** 

FPGA LPC

# Video

Western Digital. © 2021 Western Digital Corporation or its affiliates. All rights reserved.

# Western Digital.