Foundation for Research and Technology–Hellas (FORTH)
Institute of Computer Science (ICS)

NUMA-like architecture for Microservers

EUROSERVER

Iakovos Mavroidis (jacob@ics.forth.gr)
FORTH-ICS, Greece

MPSoC'14, July 8, Margaux, France

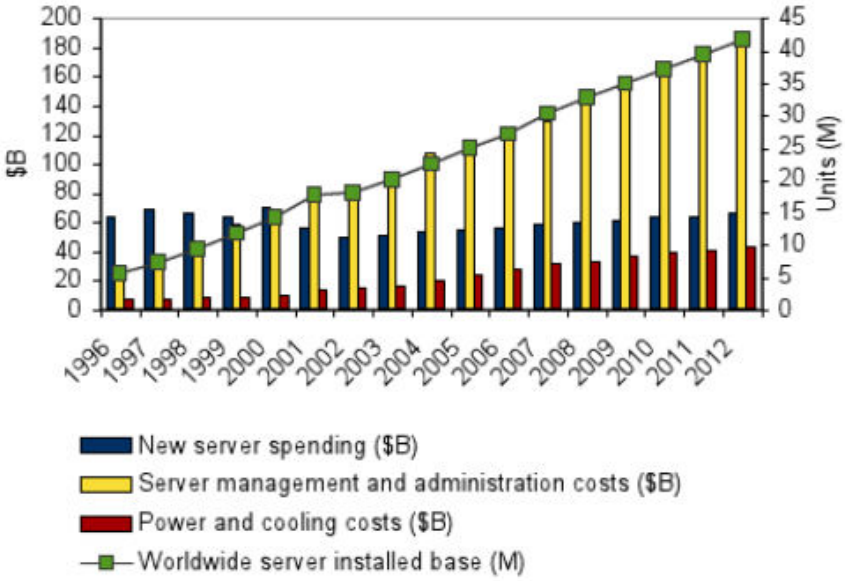
Outline

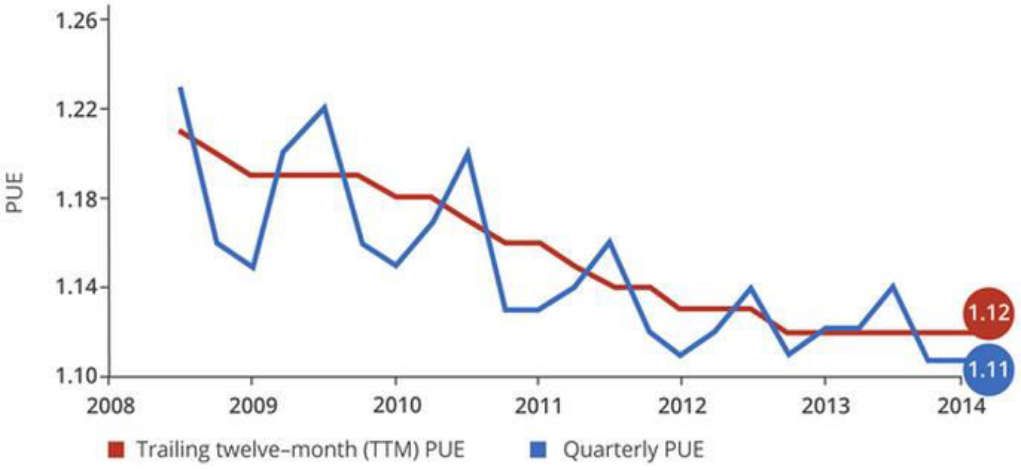
- Characteristics and requirements of today's Datacenters
  - Importance of Energy Efficiency
  - Energy Proportionality
  - Small form-factor
- Microservers
  - Energy Efficient Architecture
  - Intel or ARM?
- EUROSERVER approach
  - EUROSERVER Architecture
  - Unimem Architecture
- Testing Environment
  - FMC Fan-Out Daughtercard

Why is Size/Power/Energy Efficiency Important?

- Utility costs
  - ↑ Management Cost ☞ Improve Size
  - ↑ Power and Cooling costs ☞ Improve PUE

    (Power Usage Effectiveness)
- Electricity growth
  - 56% increase 2005-2010 (US increase 36%)
  - 19% increase 2011-2012
  - 7% increase in 2013
  - 1.1%-1.5% of global electricity 2010 (US 1.7-2.2%)
  - Google report
    - PUE=1.16 in 2010
    - PUE=1.14 in 2011
- Environmental friendliness programs
  - Energy Star (US), TopRunner (Japan), FOE (Switzerland)

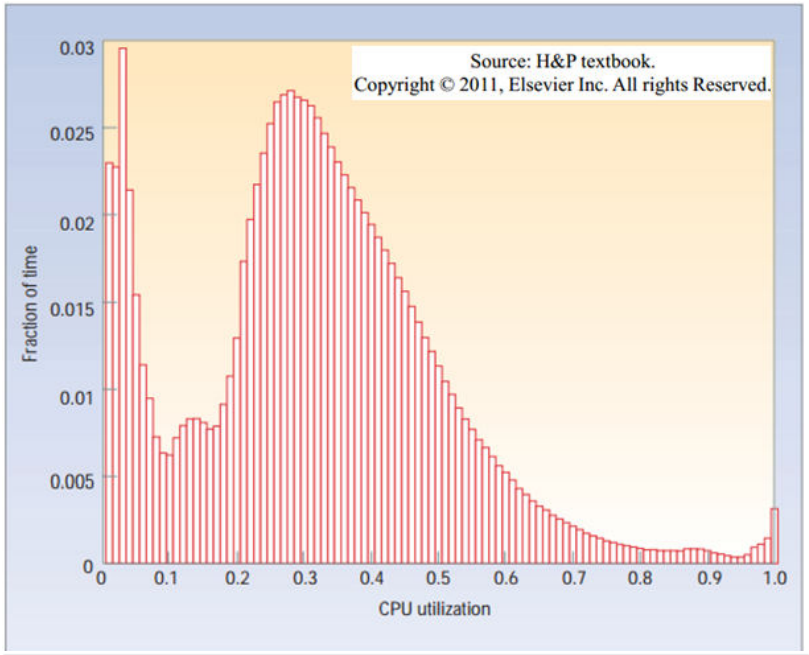


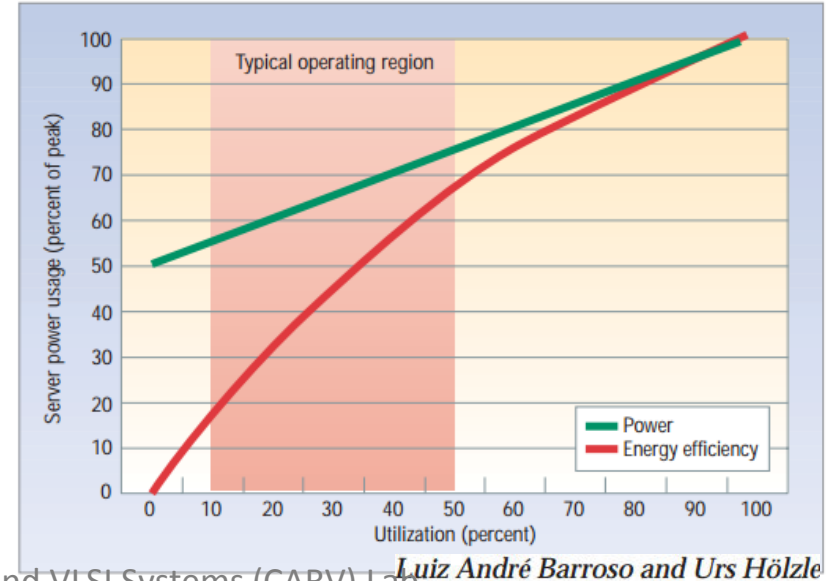


Energy Proportionality in Datacenters

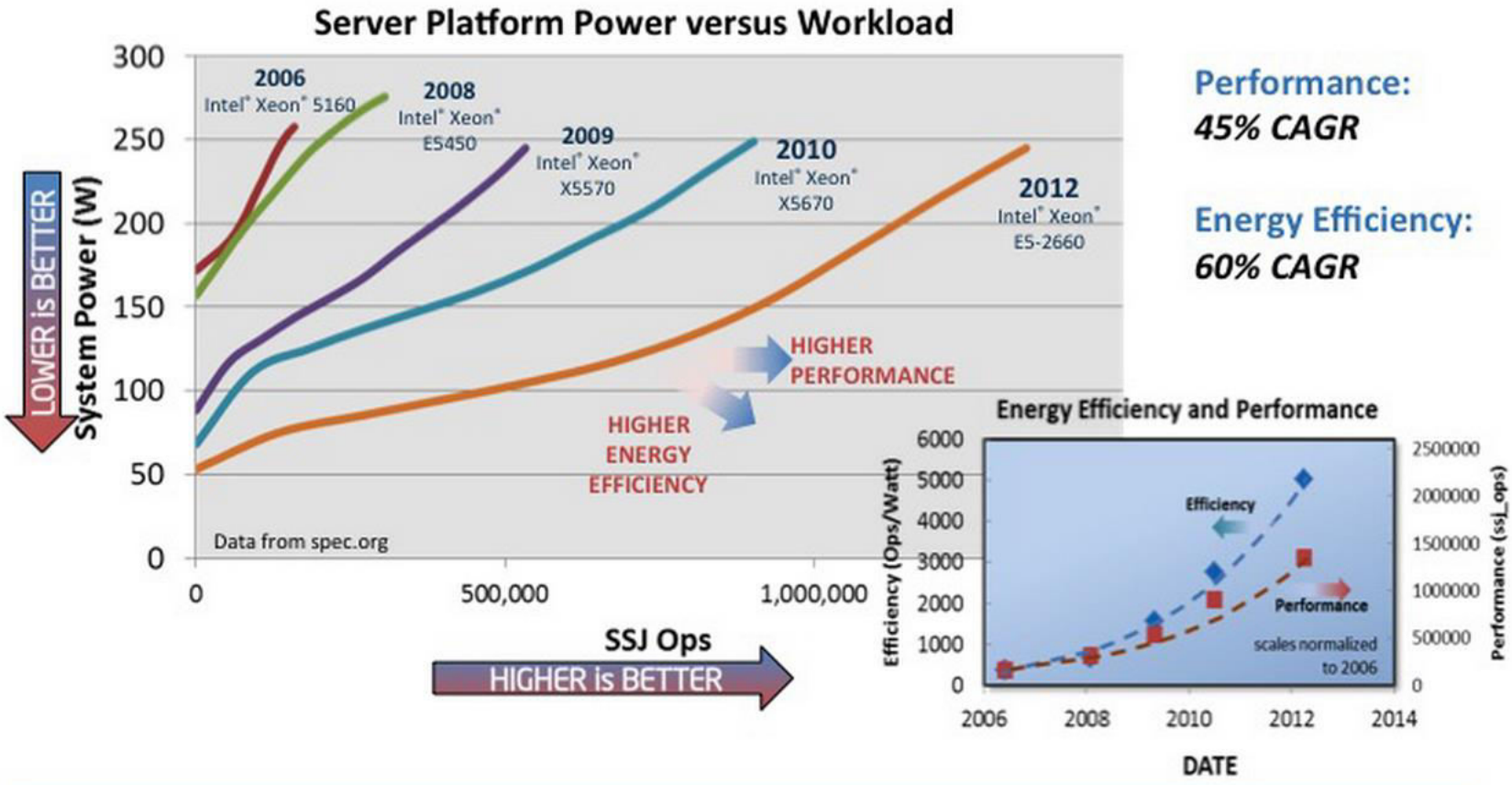
- Most of the time at 10 – 50%
- Challenge:
  - Power not proportional to utilization
  - Server underutilized

- Two approaches:
  - Turn off hardware when not used
    - Dynamic Voltage Scaling (DVS)
    - Clock Gating
  - Keep CPU utilization high
    - ☞Multiple Virtual Machines
    - ☞Overprovisioning
    - ☝ QoS guarantees ?





Aligning energy use with workloads



Xeon E5-2600 : Higher Performance and Energy Efficiency

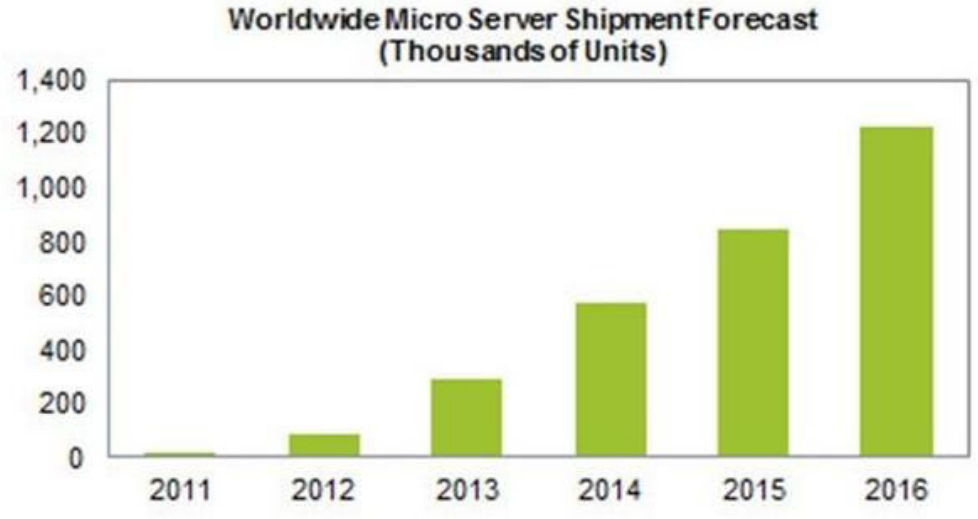
WINSTON SAUNDERS
Intel

Why many small cores?

- **Scale out** applications require large number of cores no brute processors
- Smaller cores more **power-efficien**t for several workloads
  - static web page serving, entry dedicated hosting, and basic content delivery, among others
- **Less power** consumption (sub-10W levels)
  - Lower running costs, lower PUE
- **Energy proportionality**
  - Easy to turn off idle cores (parts of the system)
- **Easier maintenance** and management
  - Small form factor allows tightly packed clusters and less physical space
- Easier more efficient implementation?
  - CPU partitioning instead of sharing (no sharing overhead)
- But…
  - No compute power for single-threaded application
  - Hard to parallelize an application
  - Not so efficient for HPC domain?

Energy-efficient architecture: Microservers

- Low-power components
  - CPU (ARM, Intel Atom)
  - Memory (HMC)
  - Storage (NVM)
- Small form factor
  - Small CPUs
  - Fast interconnections (high-speed serial links)
  - High integration
- Microservers are still in their infancy



Source: IHS iSuppli Research, February 2013

Intel or ARM in Microservers?

- Diversity of ARM ecosystem
  - Custom microservers using ARM-based SoCs
  - ☞ Hundreds of customers
- More than 50 variations of Intel Atom and Xeon
  - Xeon E3 suitable for webscale applications, online gaming, cloud
  - Atom C2000 suitable for lightweight scale-out workloads
  - ☝ Hard to compete hundreds of chip-makers (Samsung Exynos, AMD Opteron A1100 with 8 A-57, APM's X-Gene, Google, Facebook, …)
- However
  - Intel first released 64-bit SoC with ECC (Atom Avoton)
  - Intel 3-D technology
    - ☞ smaller die area
    - ☞ less energy consumption
  - Most datacenter software run on x86 (porting on ARM in progress)

Calxeda: ARM-based servers didn't have the software support or hardware needed to win enterprise customers

EUROSERVER Challenges and Approach

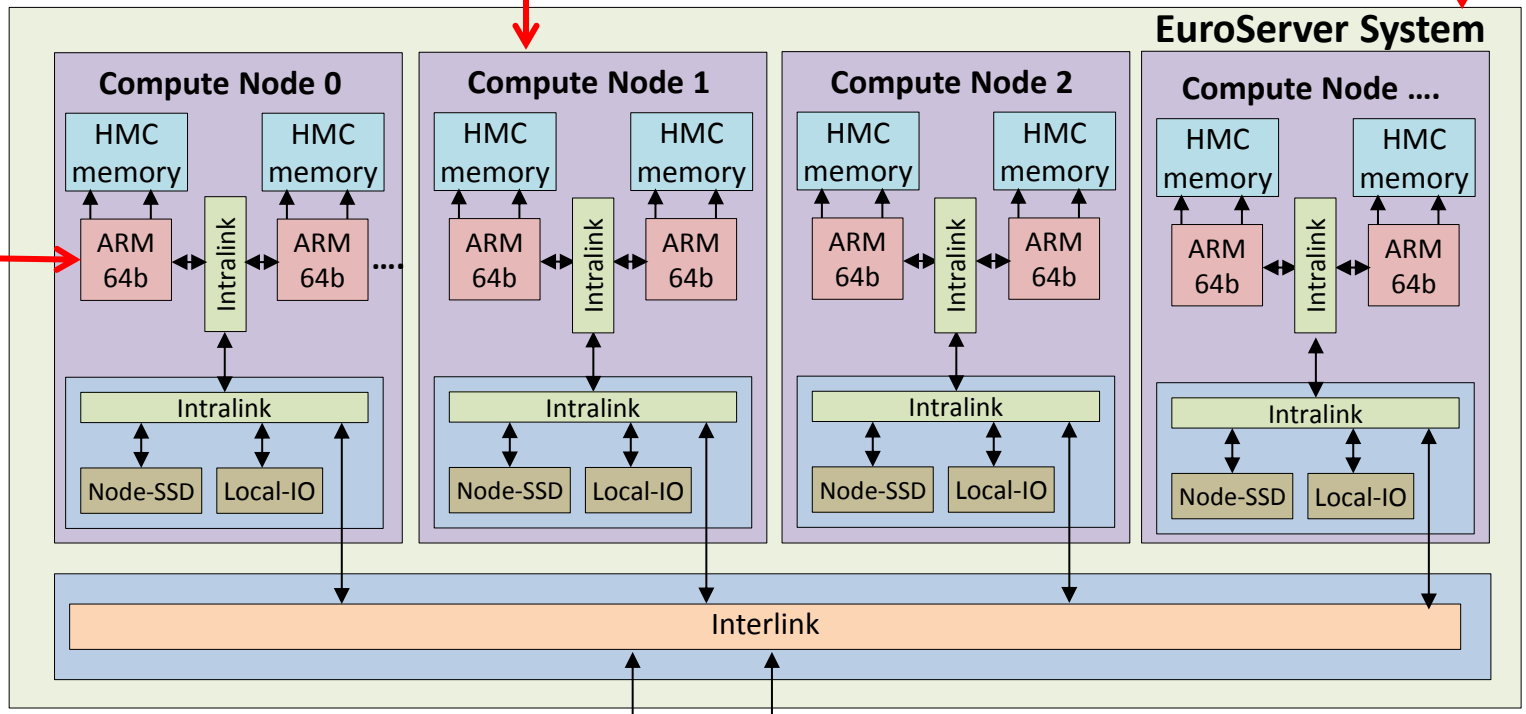
- **Energy-efficient architecture**
  - Use of highly-integrated, high-performance, energy-efficient components in a Microserver arcitecture
    - Many low-power ARMv8
    - 3D - Interposer Technology
    - HMC main memory
    - NVM memory for storage
- **Suitable from cloud data-centers to embedded applications**
  - Unimem Architecture (Focus of this presentation)
  - Take advantage of fast communication
- **Scalable architecture**
  - Many coherent islands
  - Global Address Space
- **Facilitate maintenance and management**
  - Small form factor
- **Energy proportionality**

EUROSERVER Architecture

Chiplet:
Cores+L0 Coherent Interconnect
1 coherence island

Node:
Chiplets+L1 Interconnect
Shared IO and Storage

µServer:
Nodes+L2 Interconnect
Scale-out or HPC

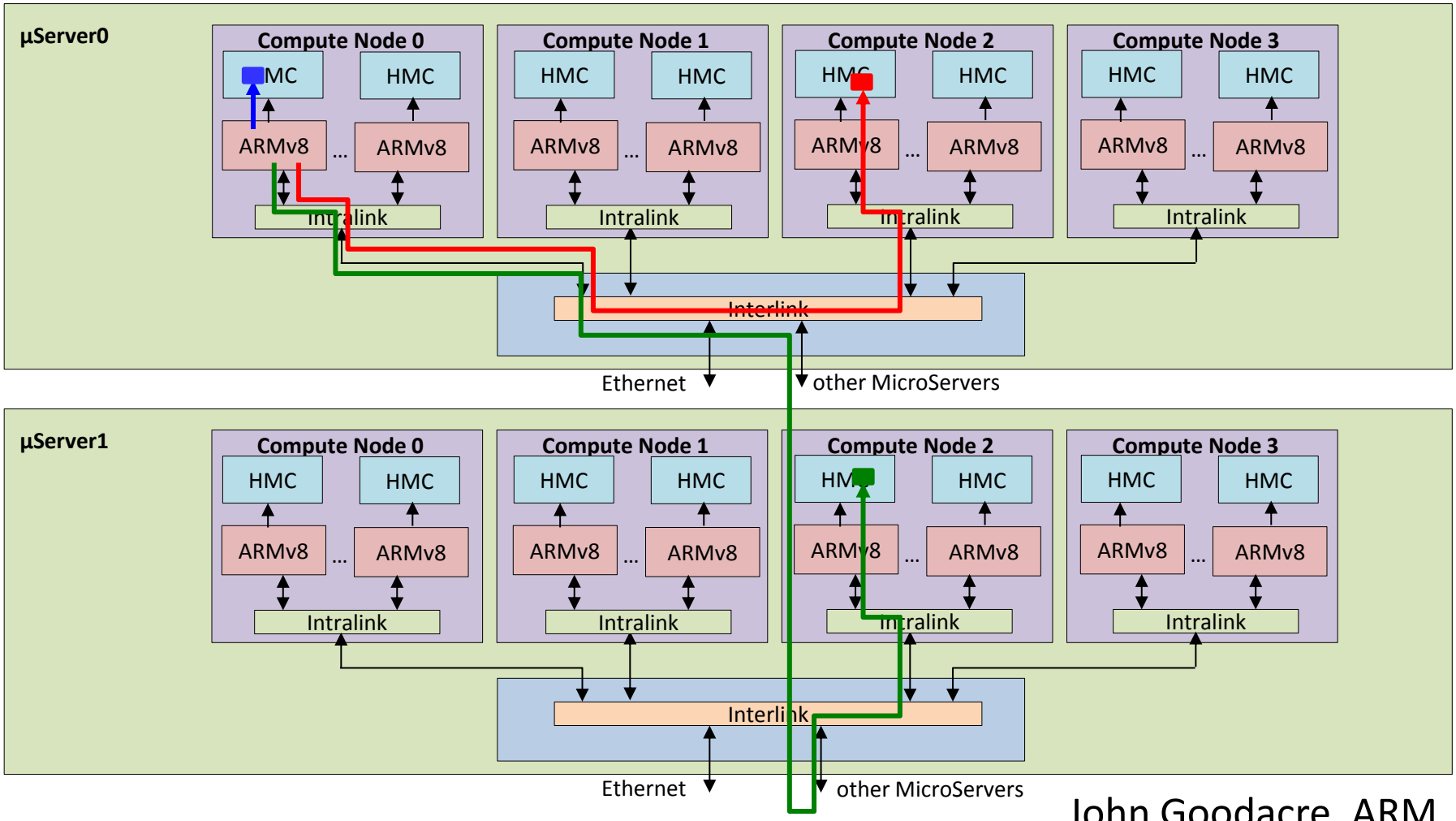


System:
Nodes+L3 Interconnect

John Goodacre, ARM

- Clustered Architecture: Coherence Islands communicating through multi-level Interconnect
- Shared IO's
- Each Coherence Island has its own local independent global (coherent) address space (GAS^L)

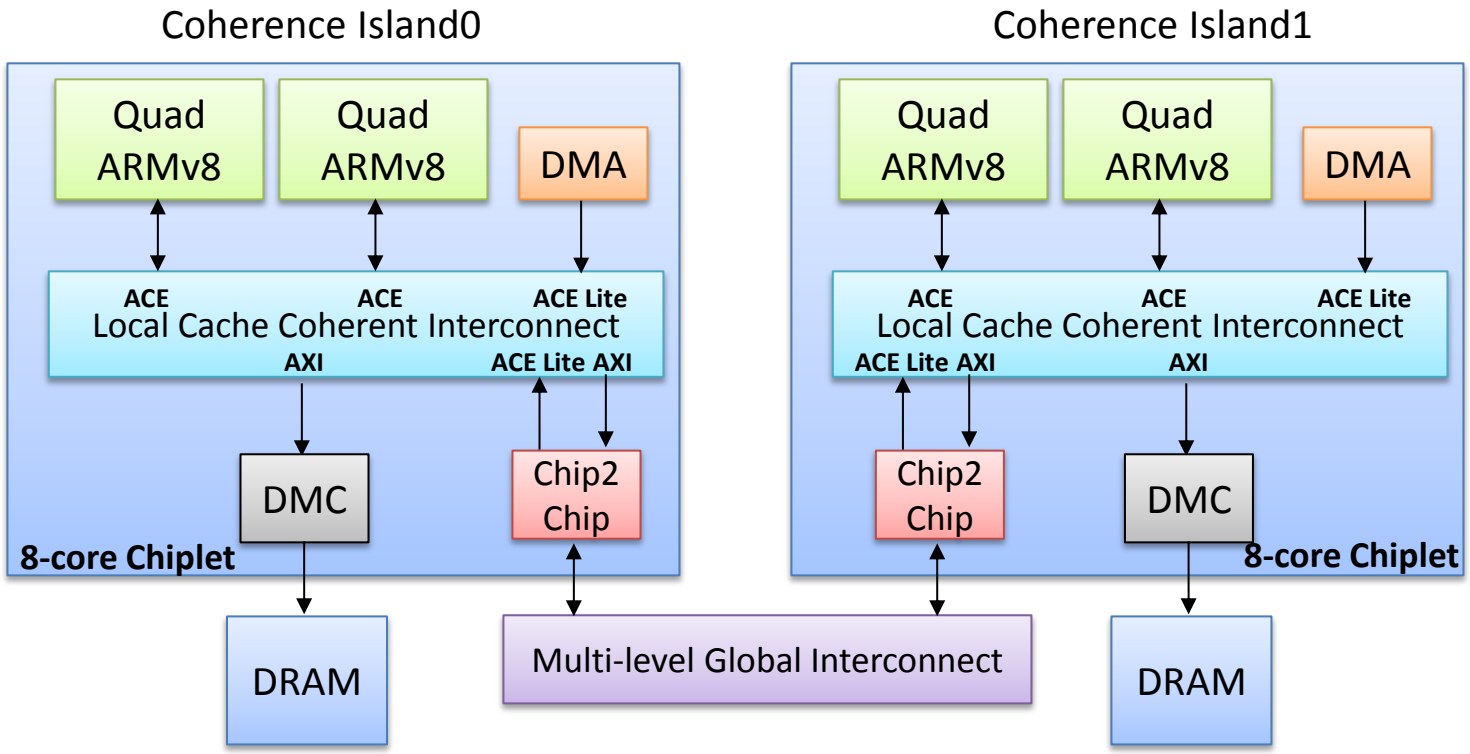
Unimem Architecture



John Goodacre, ARM

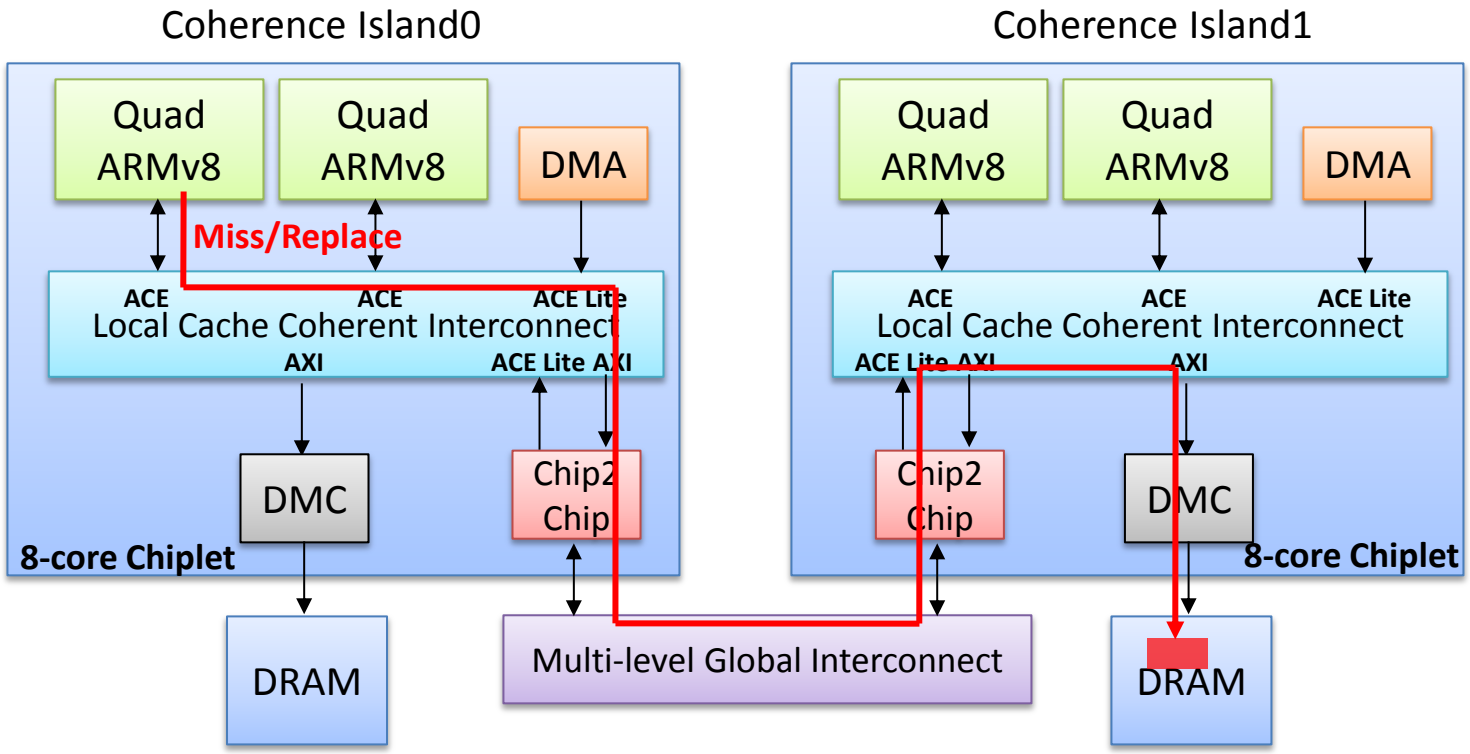
- Every memory page has a single owner (coherence island)
- A processor can access any page in the system through the page owner's coherent interconnect
- Every page can be cacheable either locally (single borrower) or remotely (owner) – but not both

EUROSERVER environment



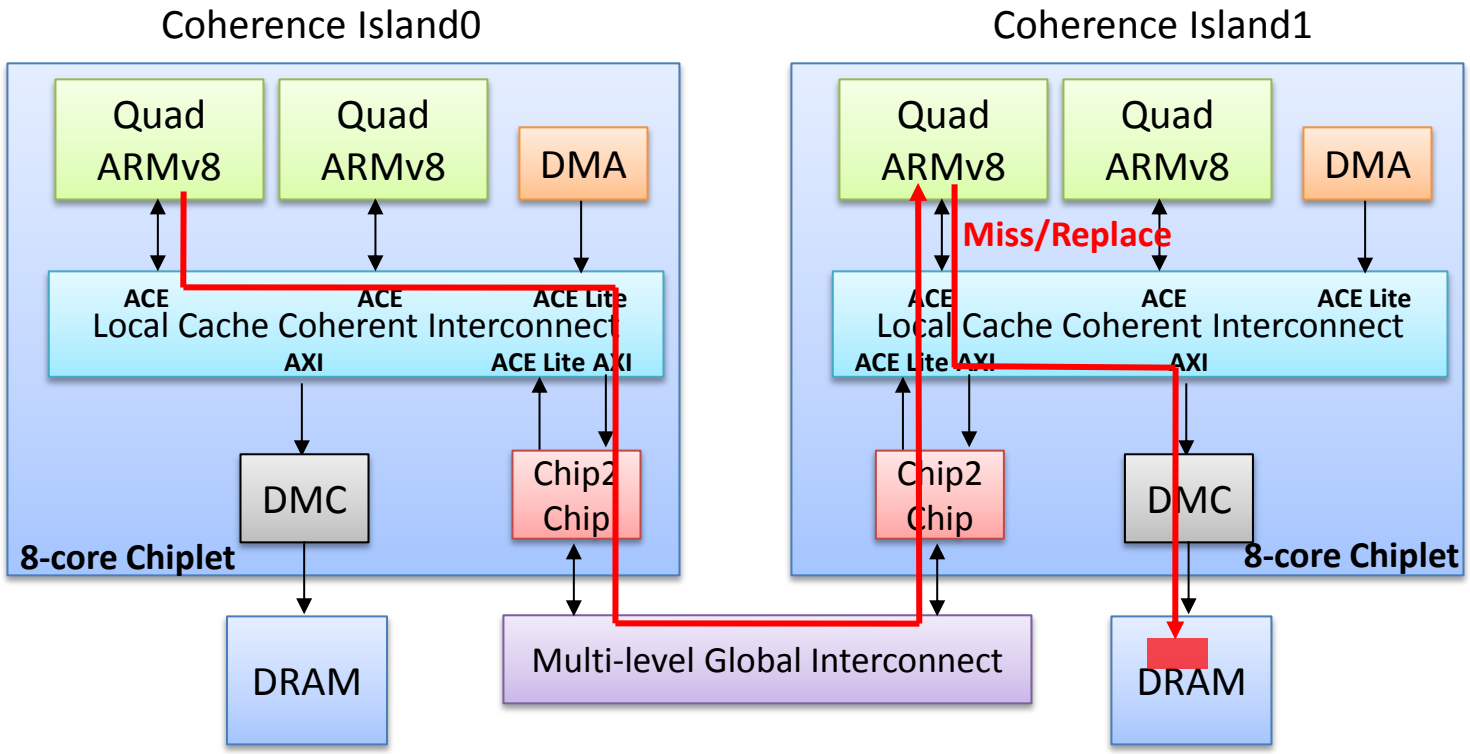
- Two coherence islands might belong in the same Compute Node (intralink communication) or not (intralink + interlink communication)

Remote Page Borrowing



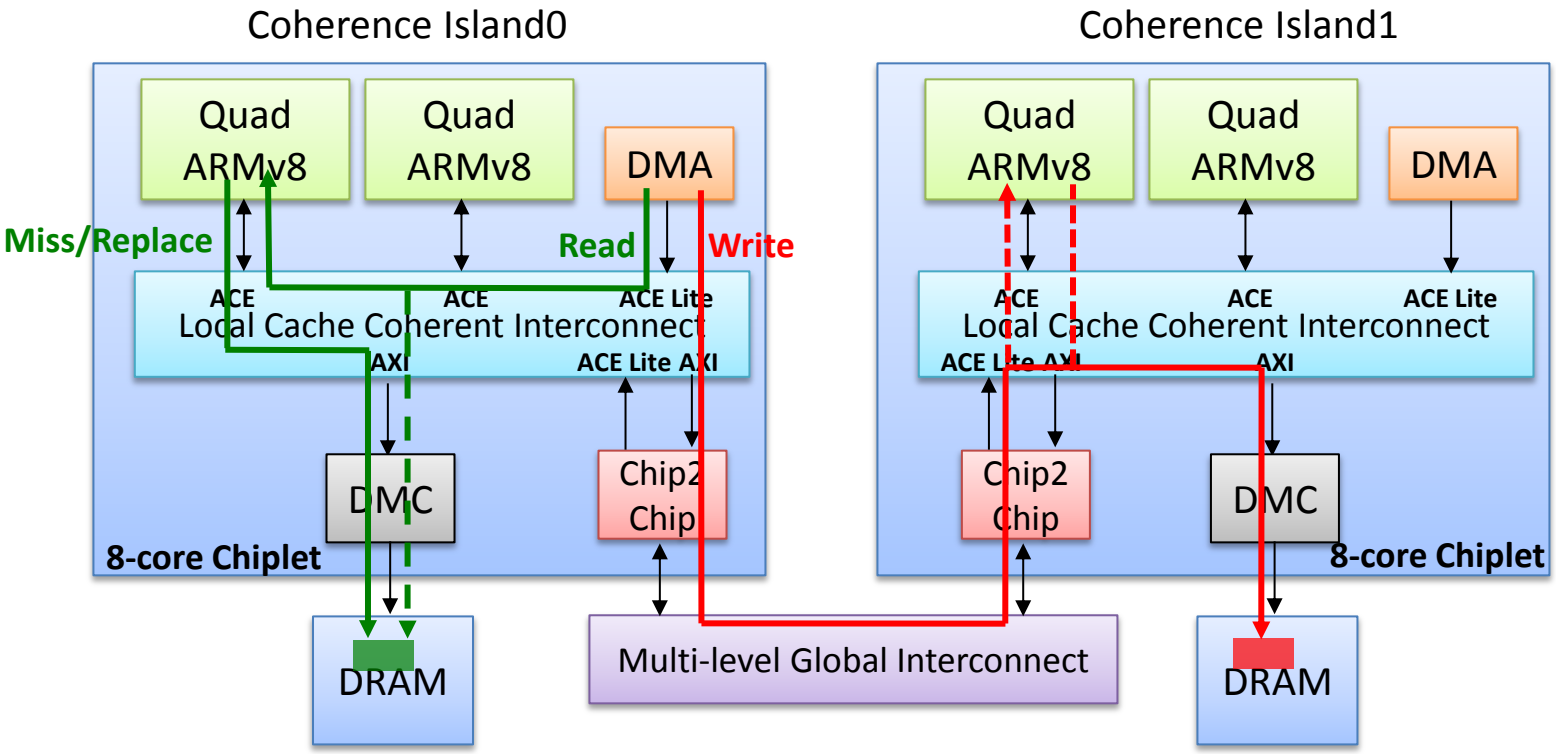
- Locally cacheable (initiator's cache)

Shared Memory



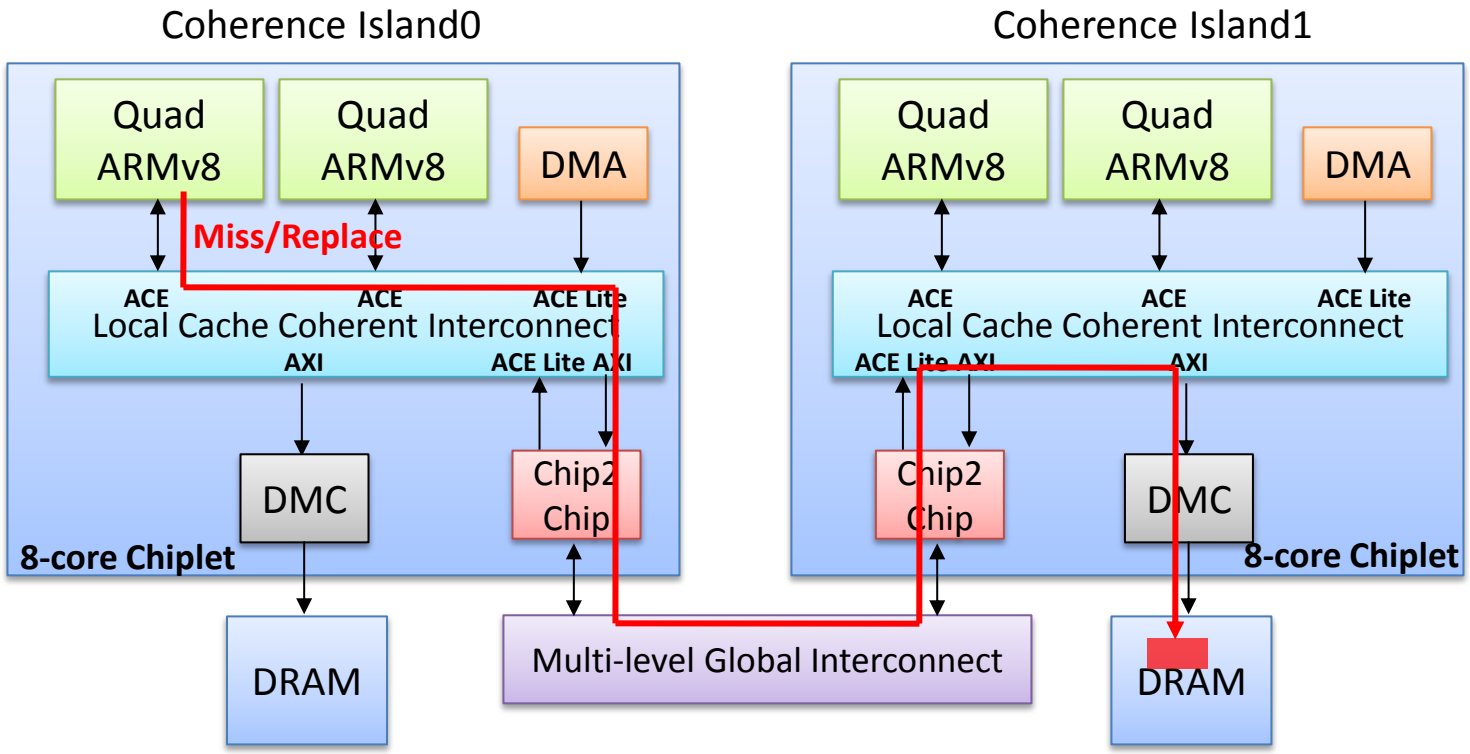
- Remotely cacheable (owner's cache)

RDMA



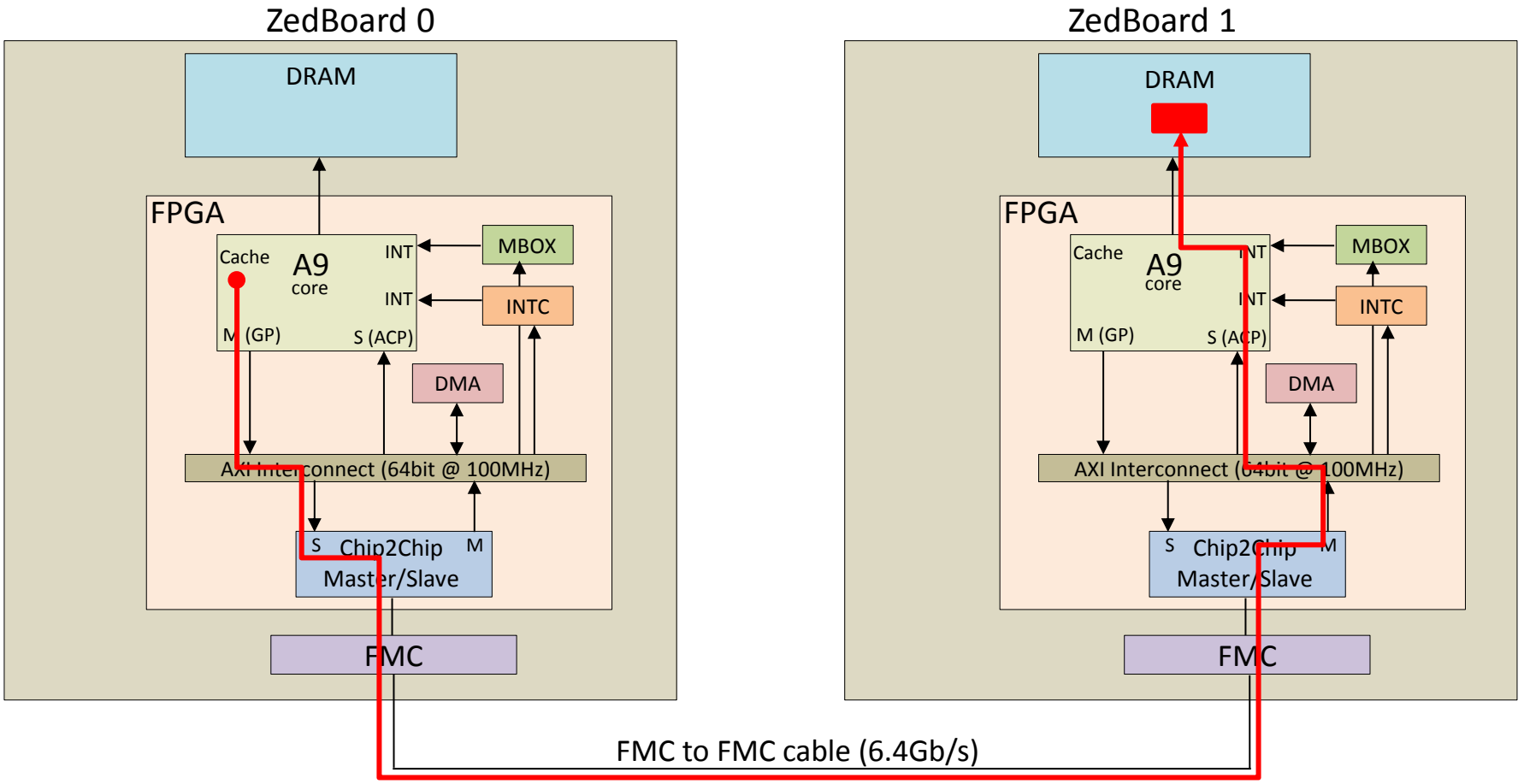
- DMA reads from (or writes to) DRAM on Coherence Island0 and writes to (or reads from) DRAM on Coherence Island1
- Accesses can also be uncacheable locally or cacheable remotely (dashed lines)

NUMA-aware linux



- Borrow unused remote memory instead of page faulting
- Fast shared memory and MPI communication
- NUMA-aware memory allocator and garbage collector

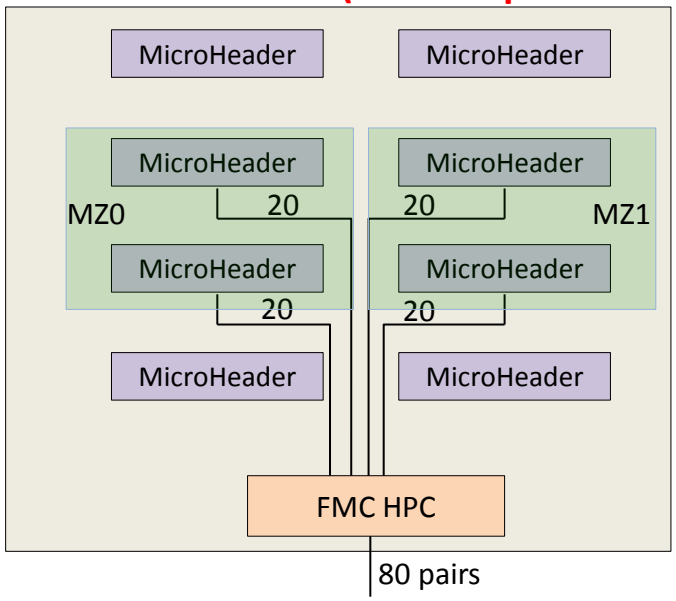
Initial Testing Environment using A9-based boards



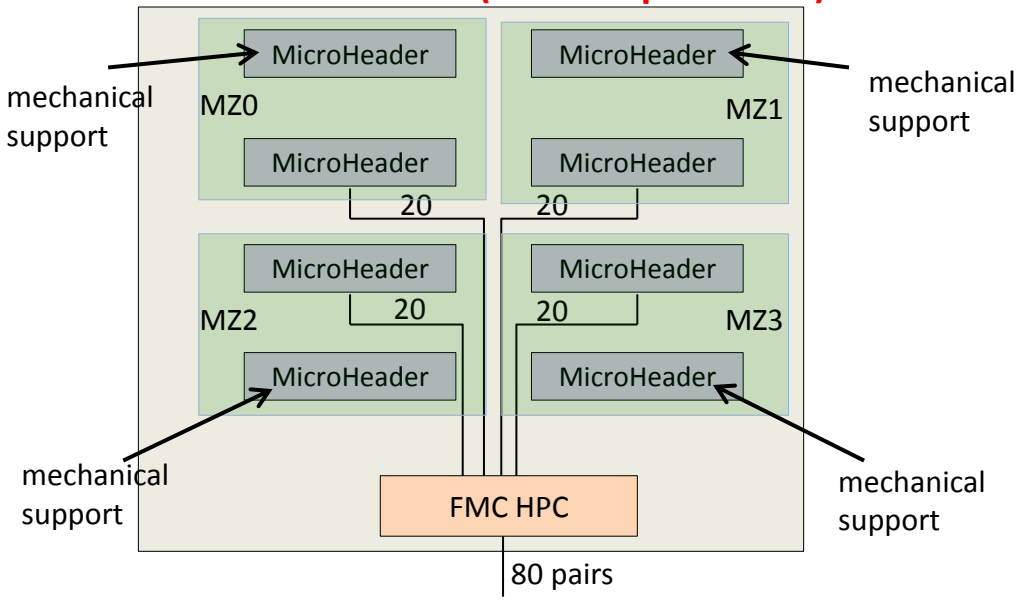
- Can we interconnect more A9 processors? (see next slides)

FMC Fan-Out Daughtercard v.1

2MicroZed boards (40LVDS per board)



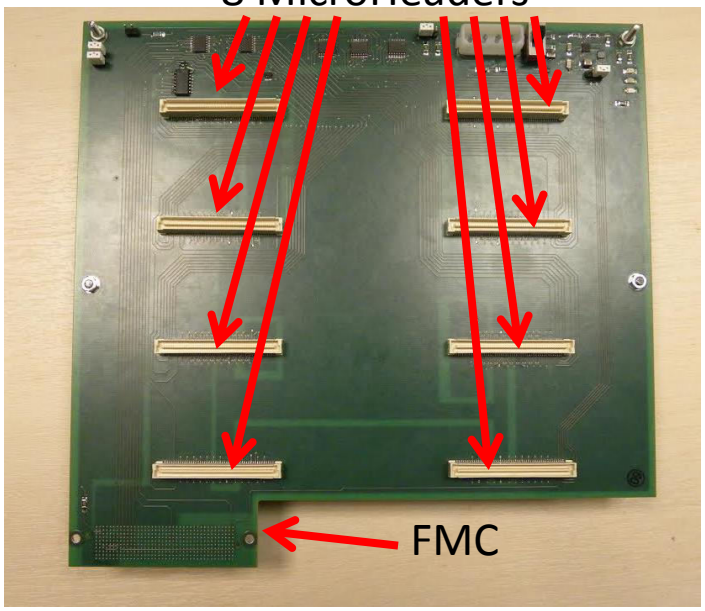
4 MicroZed boards (20 LVDS per board)

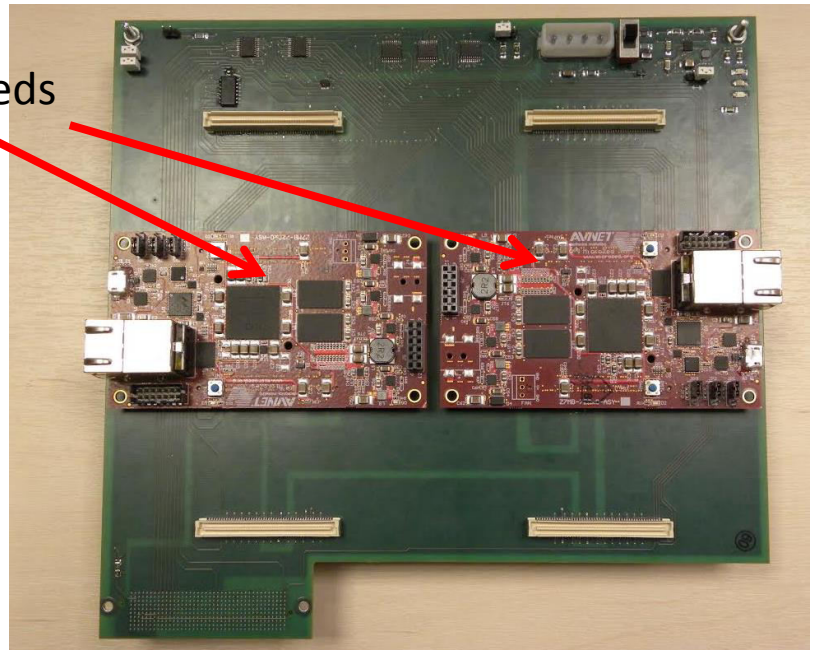


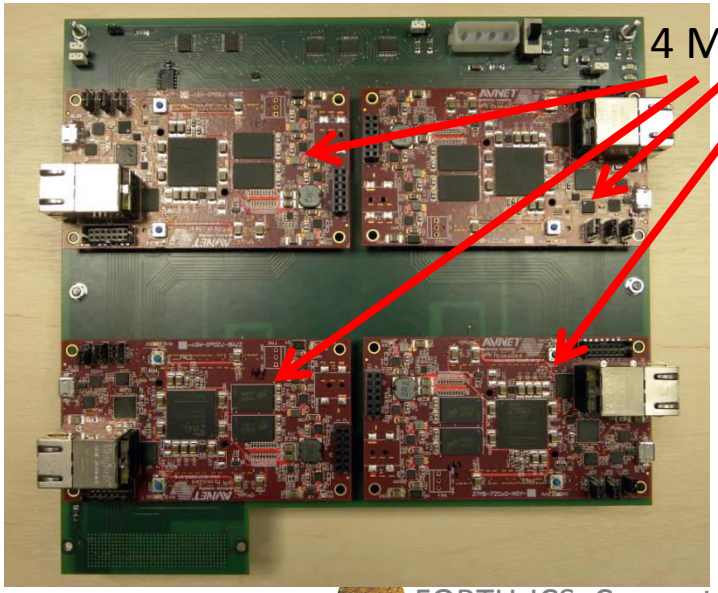
- Top and bottom MicroHeaders are mainly used for mechanical support.
- Two connectivity modes: **support for 1 to 4 MicroZed boards**

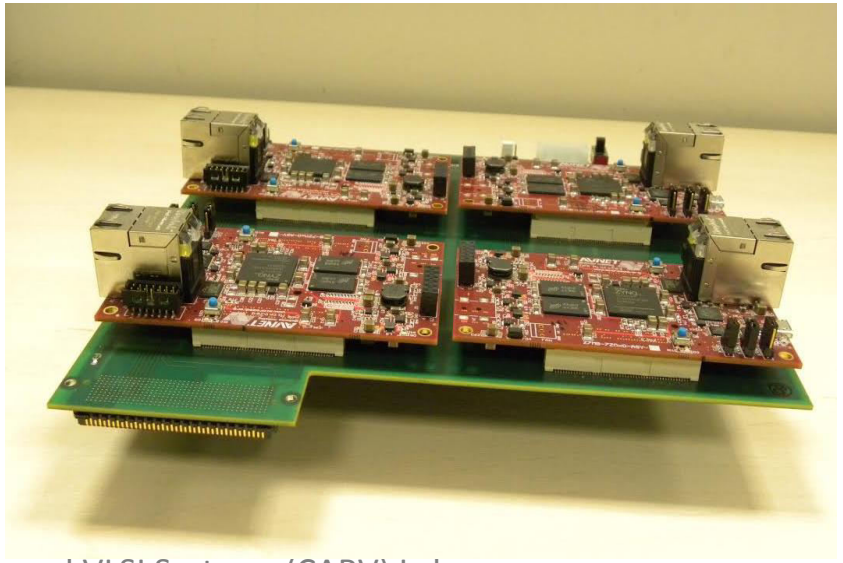
✓ **PCB design and fabrication done**
✓ **Testing done**
- Version 2 in progress

Pictures of FMC Fan-Out Daugthercard v.1

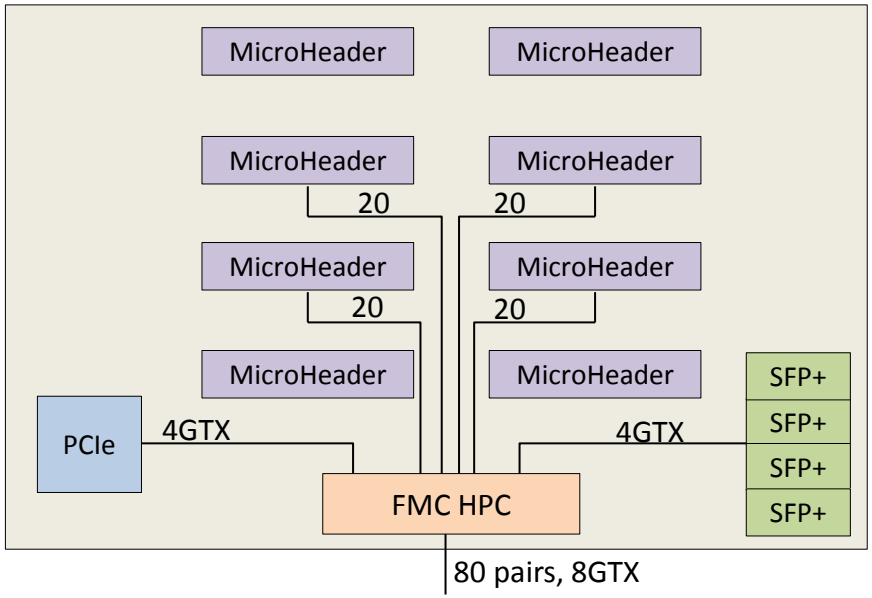






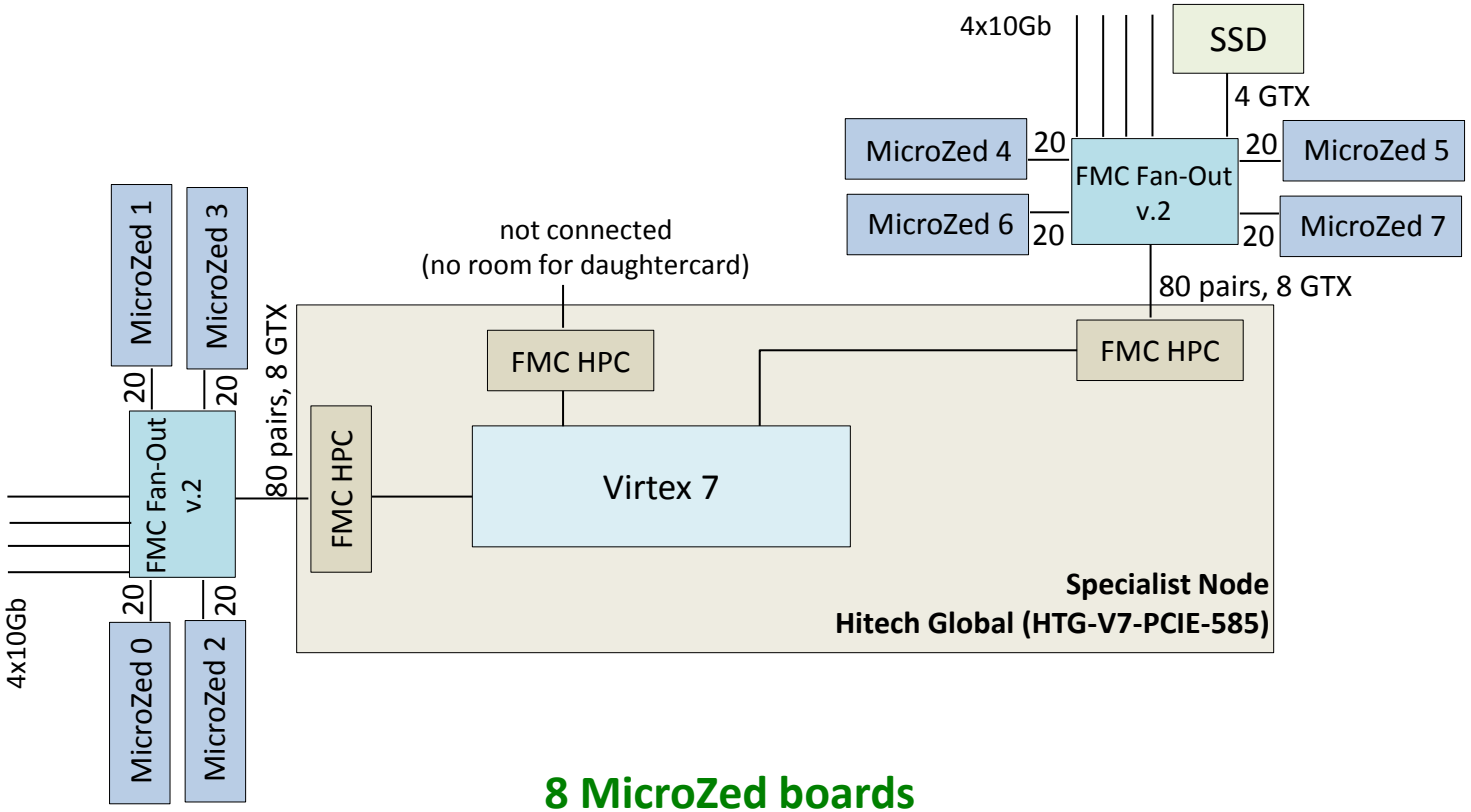


FMC Fan-out v.2 with 10GE and PCIe



- Support for:
  - Four 10Gb SFP+
  - 2.5'' PCIe socket

Initial Prototype using FMC Fan-Out v.2

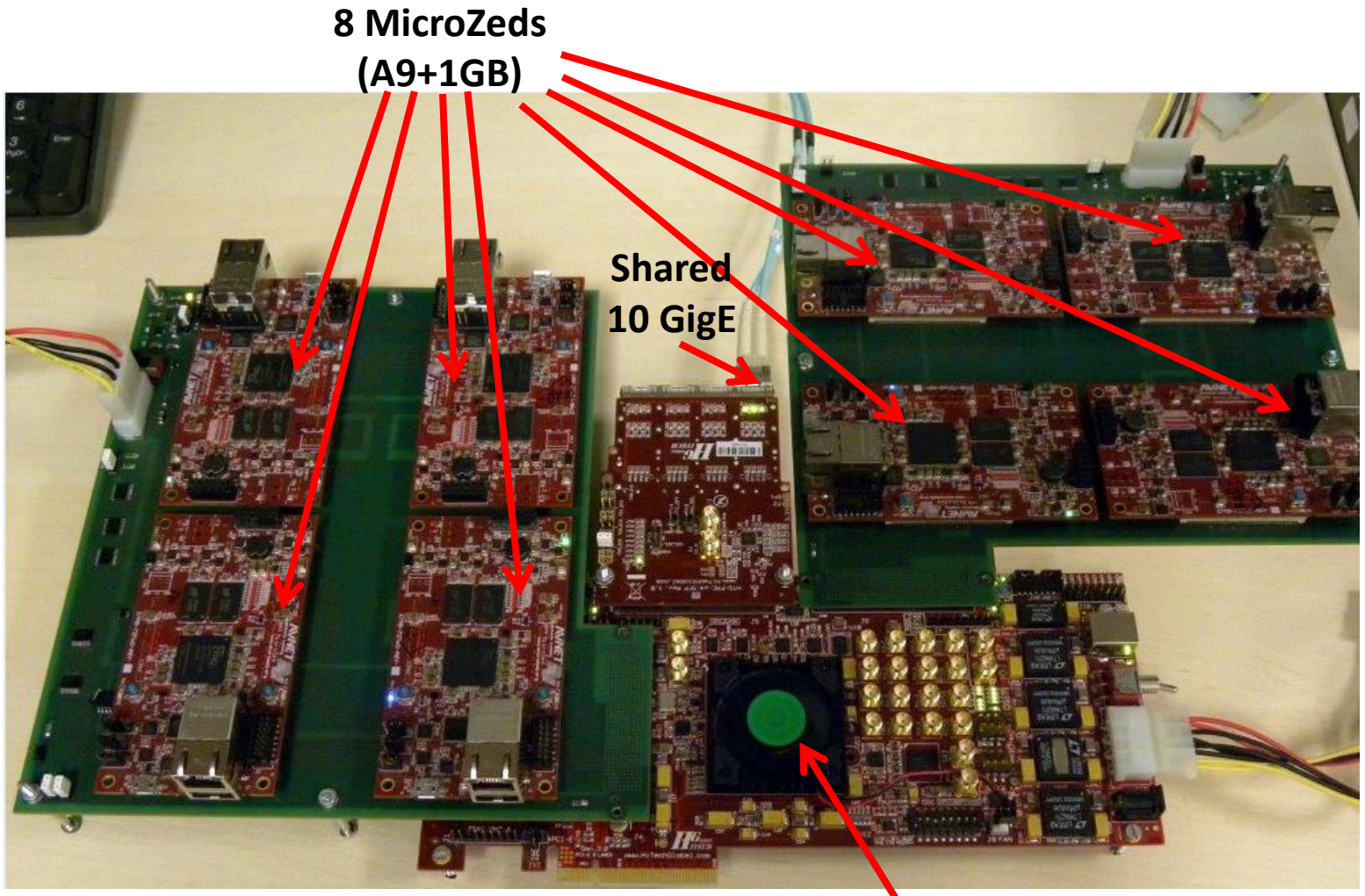


8 MicroZed boards

8 10Gb SFP+ ports

1-2 PCIe x4 SSD

Picture of testing environment using FMC Fan-Out v.1



Thank you!

Questions?

Iakovos Mavroidis
jacob@ics.forth.gr
FORTH-ICS