

Welcome!



European Exascale Processor & Memory Node Design



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 671578

Paving the way towards a highly energy-efficient and highly integrated compute node for the Exascale revolution: the ExaNoDe approach.

Kevin Pouget

Virtualization Engineer
Virtual Open Systems (ExaNoDe partner)

August 30th, 2017

Disclaimer: This presentation does not represent the opinion of the EC and the EC is not responsible for any use that might be made of information appearing herein.

Paving the way towards a highly energy-efficient and highly integrated **compute node** for the Exascale revolution: the ExaNoDe approach

Outline:

1. ExaNoDe Overview
2. System Architecture and Integration
3. ExaNoDe Prototype
4. Application and Software

Paving the way towards a highly **energy-efficient** and highly **integrated** **compute node** for the **Exascale** **revolution**: the ExaNoDe approach

Outline:

1. ExaNoDe Overview
2. System Architecture and Integration
3. ExaNoDe Prototype
4. Application and Software

The new European HPC research landscape



Source: Panagiotis TSARCHOPOULOS SC'15



Project Start: Autumn 2015

The new European HPC research landscape



Source: Panagiotis TSARCHOPOULOS SC'15



Project Start: Autumn 2015

The new European HPC research landscape

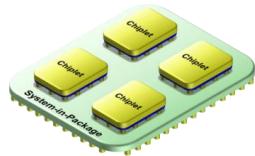


Source: Panagiotis TSARCHOPOULOS SC'15



Project Start: Autumn 2015

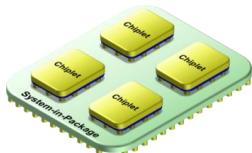
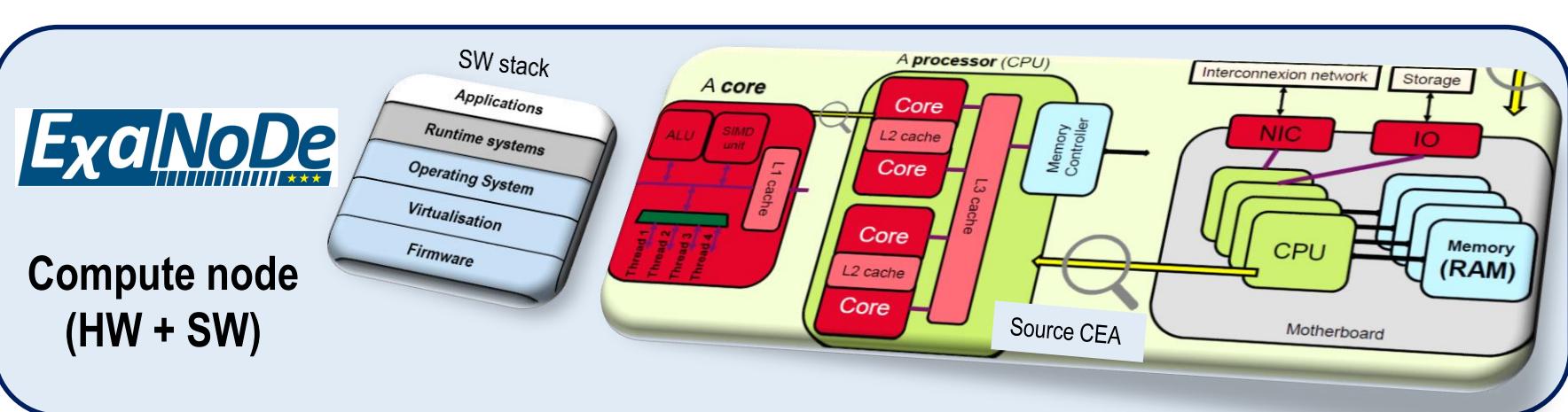
ExaNoDe technological context



Architecture

(ended in Jan. 17)

ExaNoDe technological context



Architecture

(ended in Jan. 17)

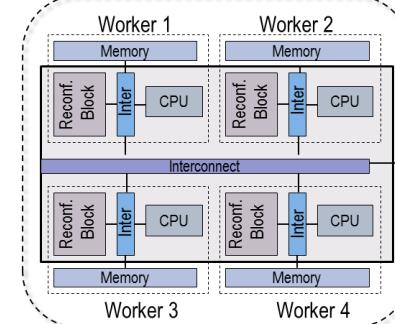
ExaNoDe technological context



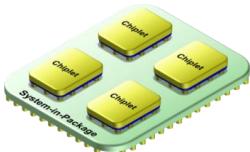
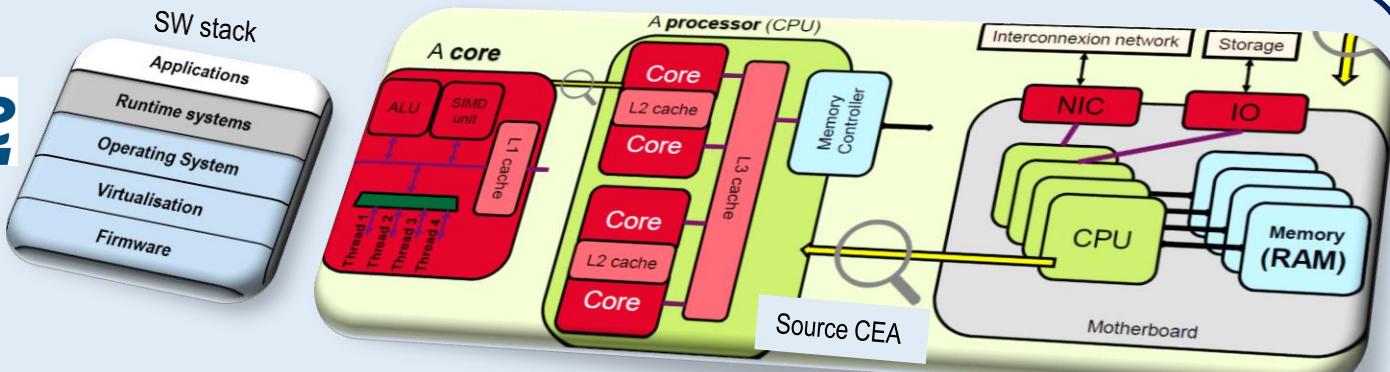
Interconnect
Storage



FPGA acceleration



Compute node
(HW + SW)



EURO
SERVER

Architecture

(ended in Jan. 17)



2017-08-30 DSD Conference

Copyright © 2017 Members of the ExaNoDe Consortium

Kevin Pouget

Ecosystem from European Projects

ARMv8,
UNIMEM for
micro-servers



EURO
SERVER

Architecture

Compute
node

ExaNoDe

Compute node
(HW & SW)



Acceleration
with FPGA



Interconnect,
storage &
cooling

ExaNoDe as part of a global strategy



EURO
SERVER

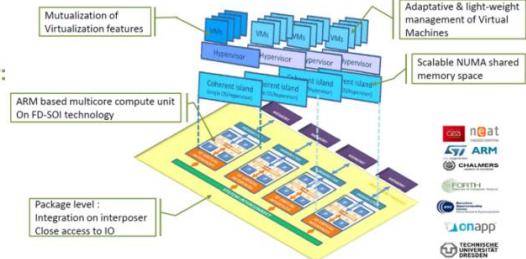
redesigns the enterprise server:

Lower cost through system integration

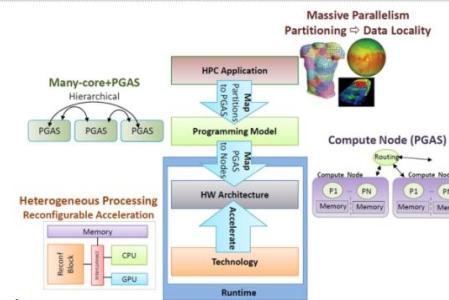
Energy efficiency : low-power 64 bit processor and more efficient software

Mutualization of I/O resources

Source: Isabelle Dor



focuses on acceleration



Source: Iakovos Mavroidis

European Exascale System Interconnect and Storage - www.exanest.eu



Storage,
Interconnect,
Cooling

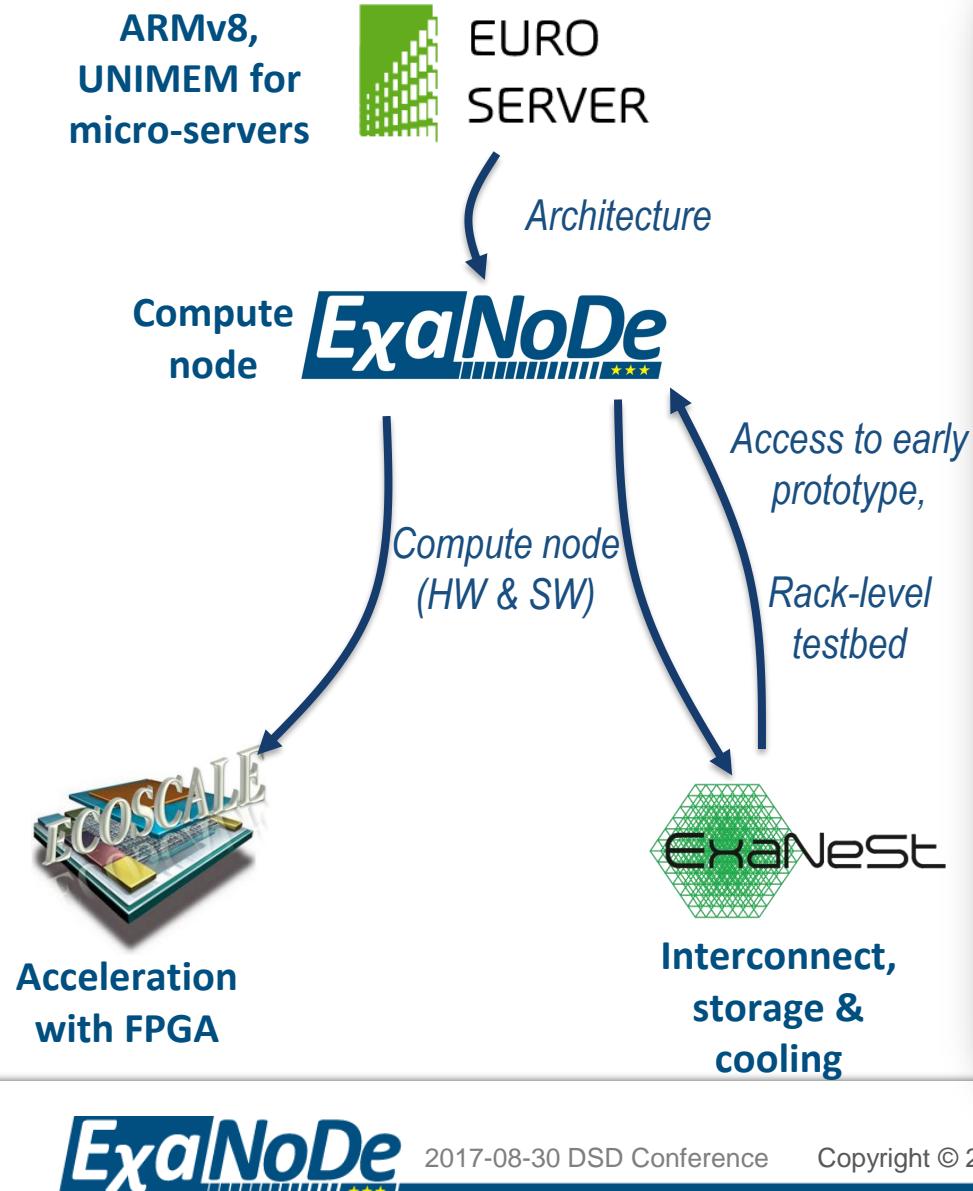
Storage: fast, distributed in-node non-volatile memory
Interconnect: low-latency, unified compute & storage traffic
Packaging: advanced, liquid-cooled
App's: real, scientific and datacenter
Prototype: 1000+ ARM cores from *EuroServer*: ARM nodes with UNIMEM address space & shared I/O from *ExaNoDe*: Chiplets, Si Interposer with *ECOSCALE*: Heterog. ARM+FPGA's



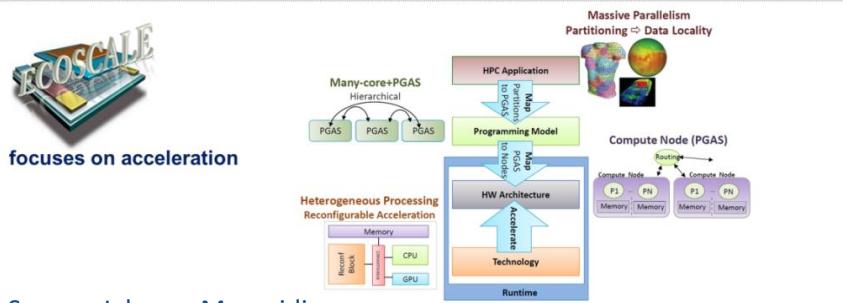
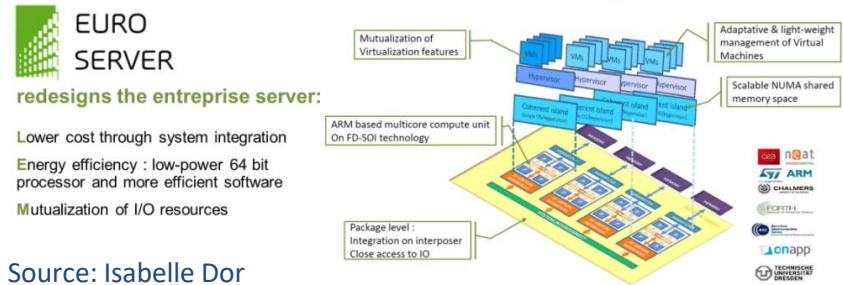
Iceotope Ltd:
Fully Immersed
Cooling Technology

Source: Manolis Marazakis

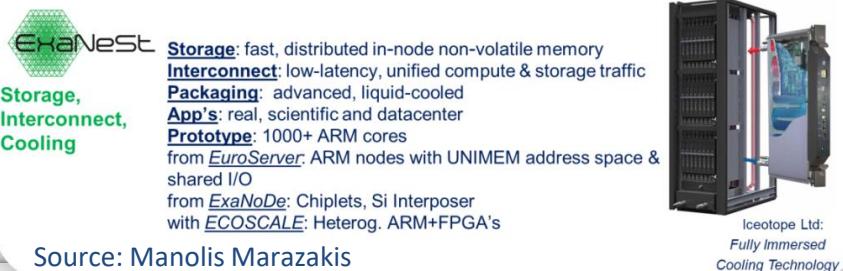
Ecosystem from European Projects



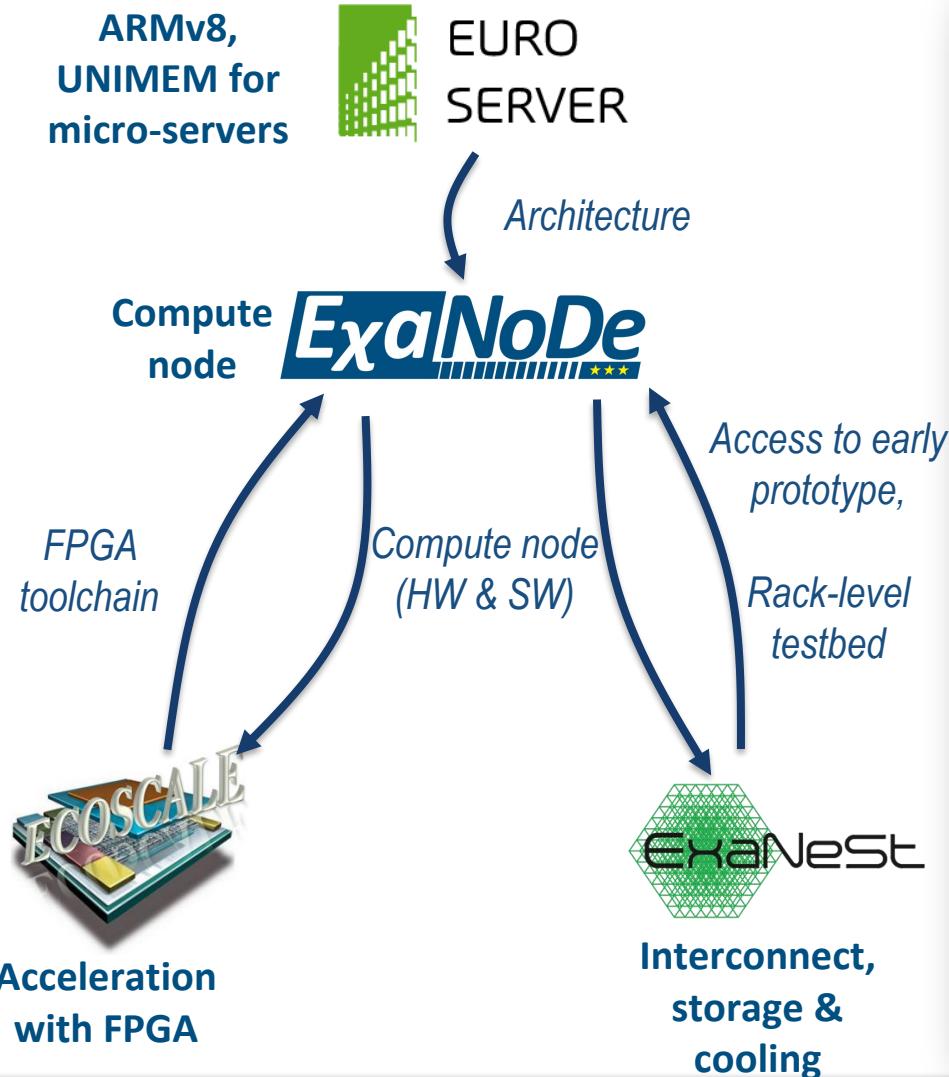
ExaNoDe as part of a global strategy



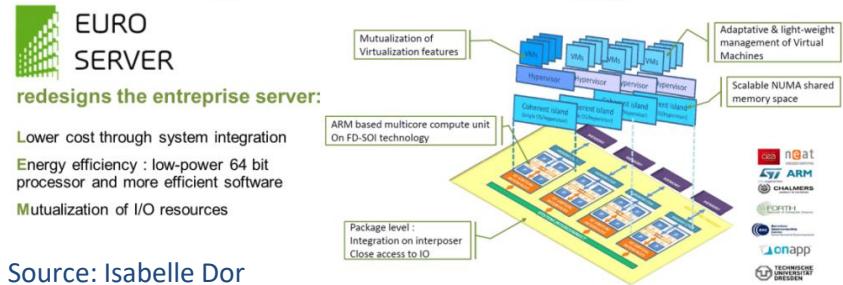
European Exascale System Interconnect and Storage - www.exanest.eu



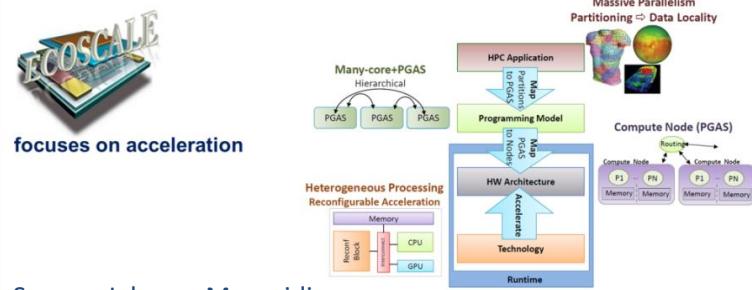
Ecosystem from European Projects



ExaNoDe as part of a global strategy

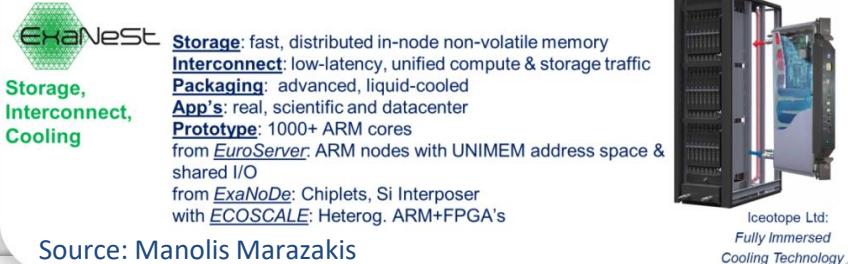


Source: Isabelle Dor



Source: Iakovos Mavroidis

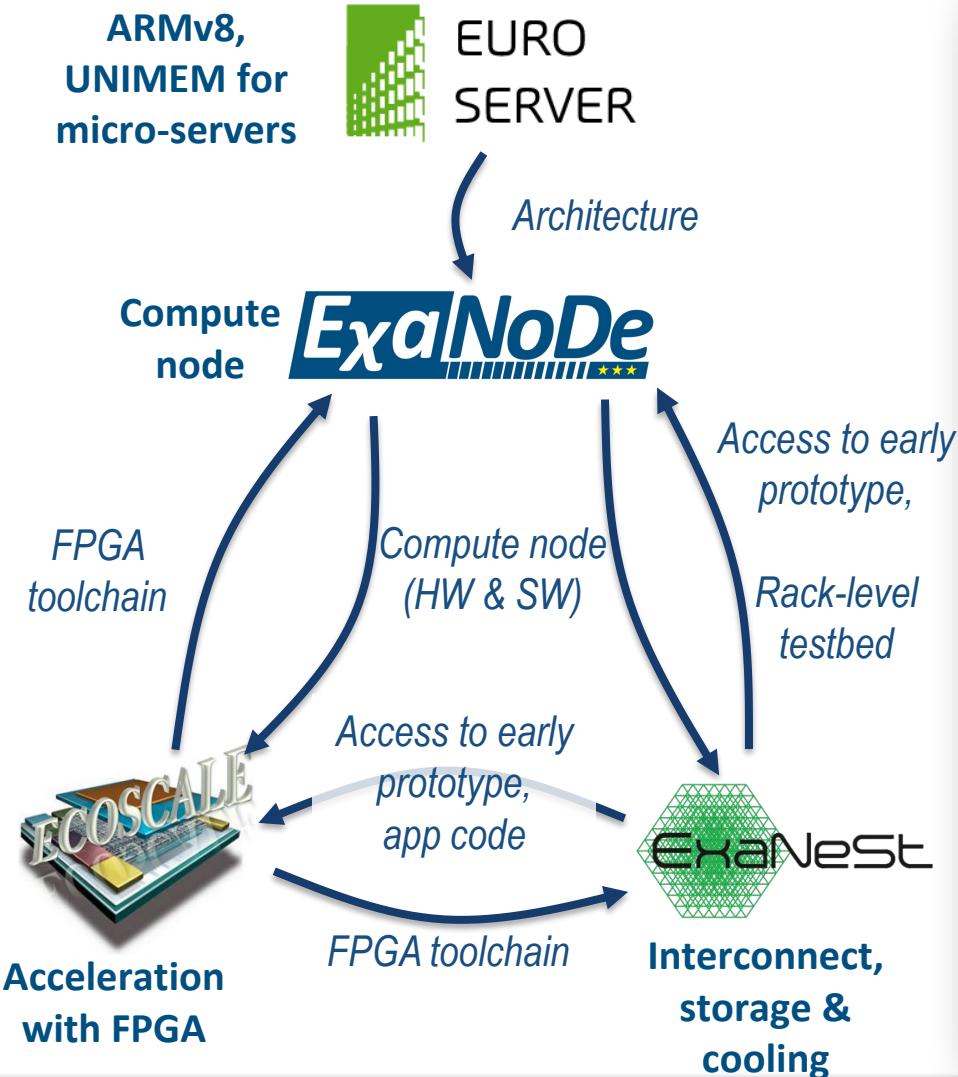
European Exascale System Interconnect and Storage - www.exanest.eu



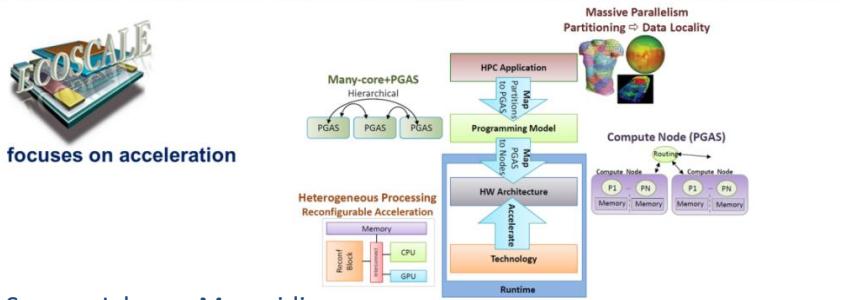
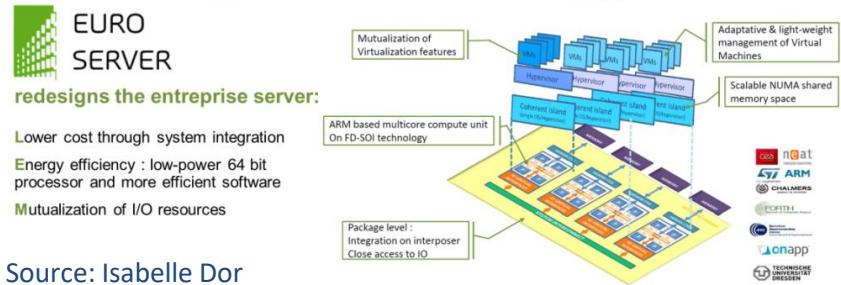
Source: Manolis Marazakis



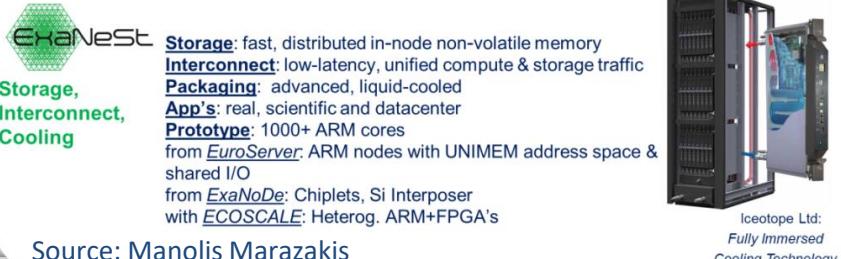
Ecosystem from European Projects



ExaNoDe as part of a global strategy



European Exascale System Interconnect and Storage - www.exanest.eu



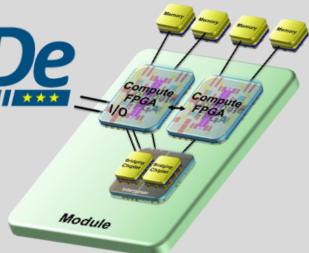
ExaNoDe Project Implementation

Start: October 1st, 2015

Duration: 3 years

ExaNoDe

8.6 M€



ExaNoDe

2017-08-30 DSD Conference

Copyright © 2017 Members of the ExaNoDe Consortium

Kevin Pouget

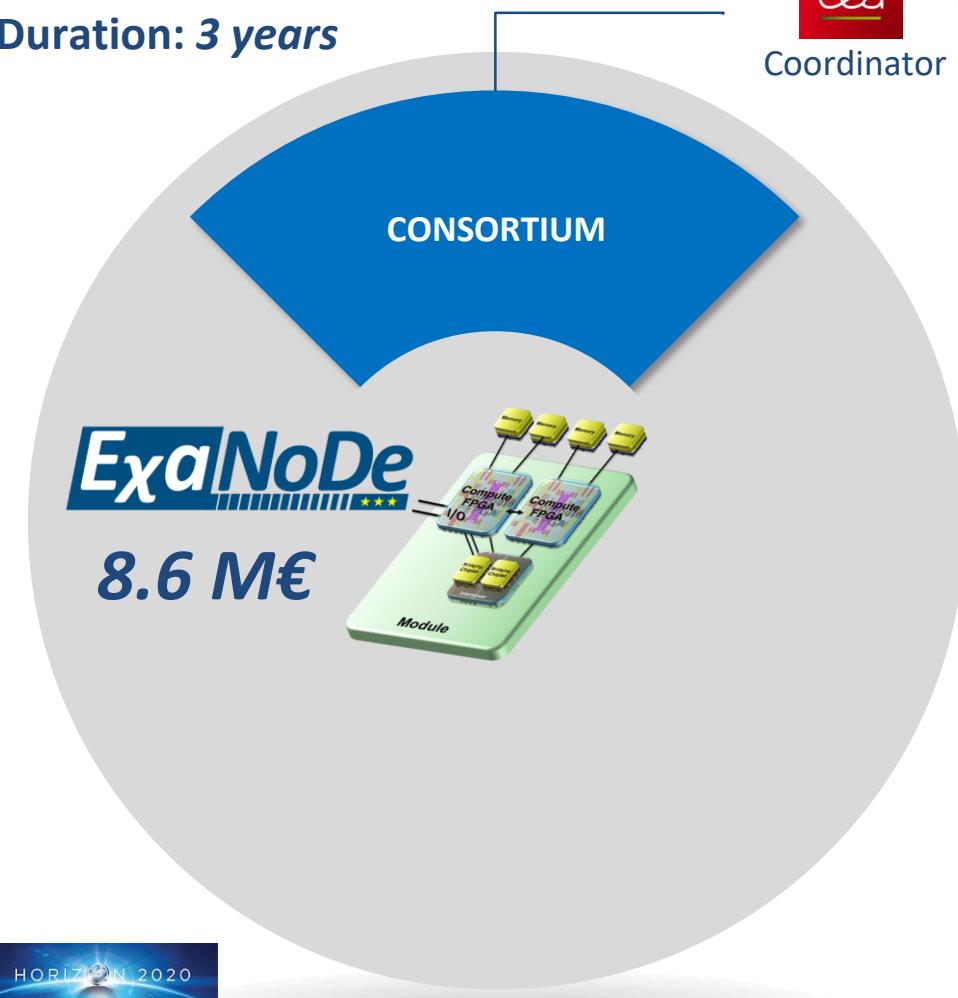
14

contact: denis.dutoit@cea.fr, project coordinator

ExaNoDe Project Implementation

Start: October 1st, 2015

Duration: 3 years



Coordinator



Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Bull
atos technologies



ExaNoDe Project Implementation

Start: October 1st, 2015

Duration: 3 years



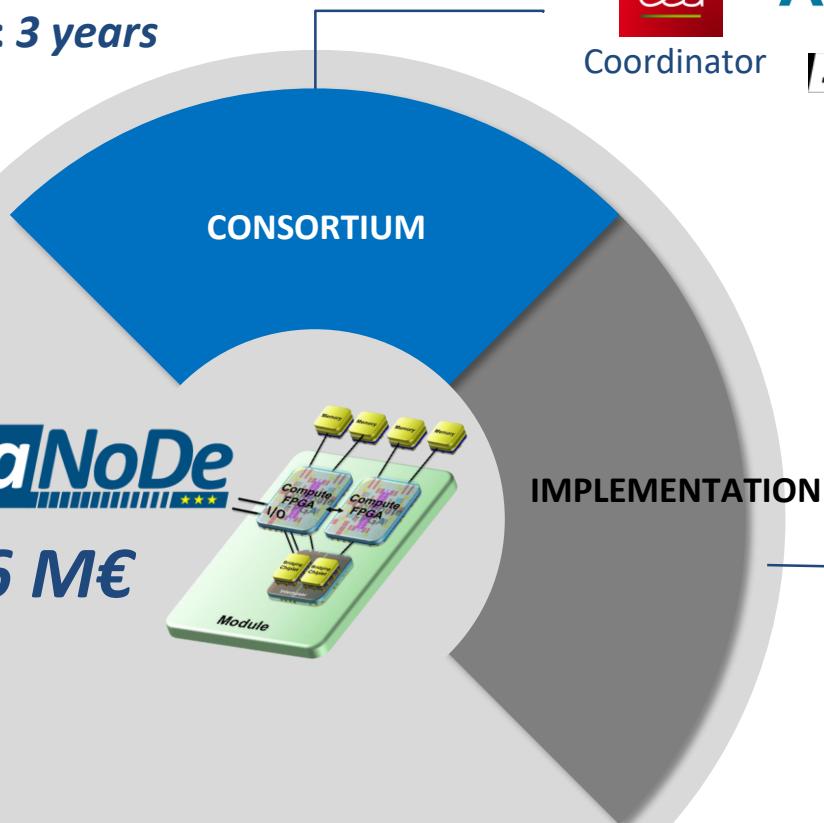
Coordinator



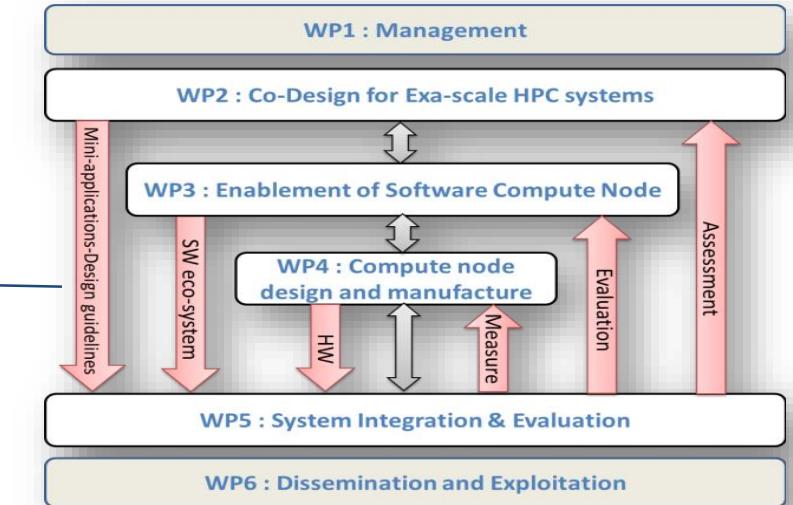
Barcelona Supercomputing Center
Centro Nacional de Supercomputación



MANCHESTER 1824
The University of Manchester



JÜLICH
FORSCHUNGZENTRUM



2017-08-30 DSD Conference

Copyright © 2017 Members of the ExaNoDe Consortium

Kevin Pouget

16

contact: denis.dutoit@cea.fr, project coordinator

ExaNoDe Project Implementation

Start: October 1st, 2015

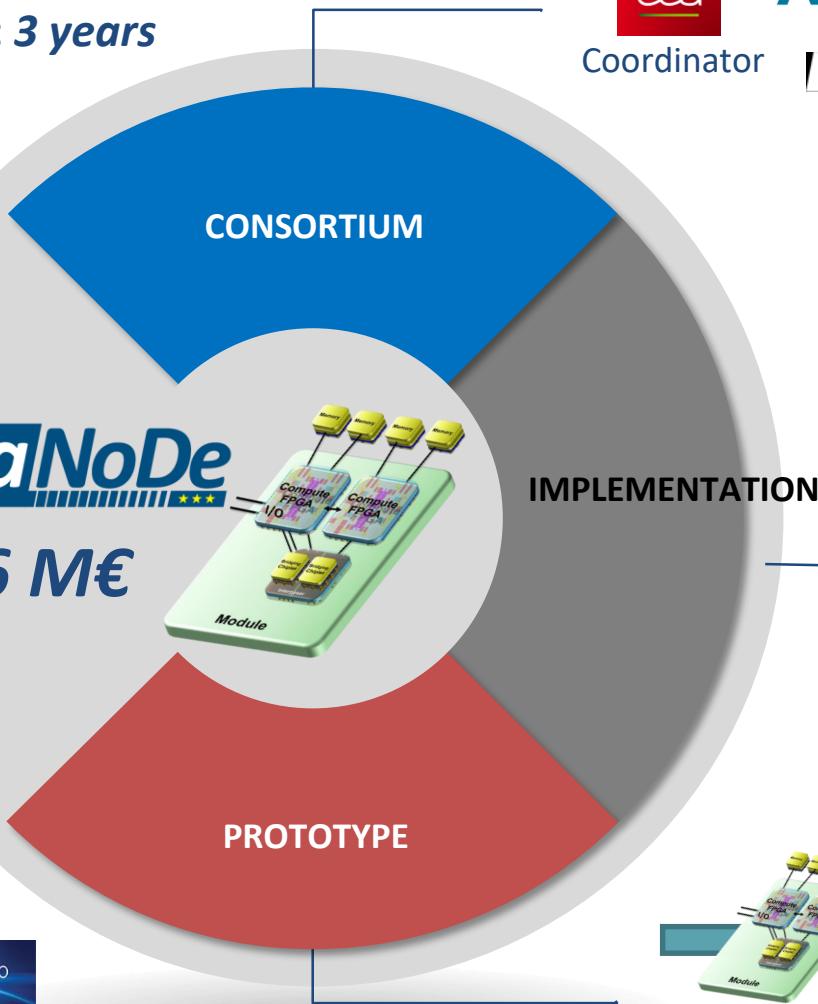
Duration: 3 years



Coordinator



Barcelona Supercomputing Center
Centro Nacional de Supercomputación



zürich



scapos



FORTH
Institute of Computer Science

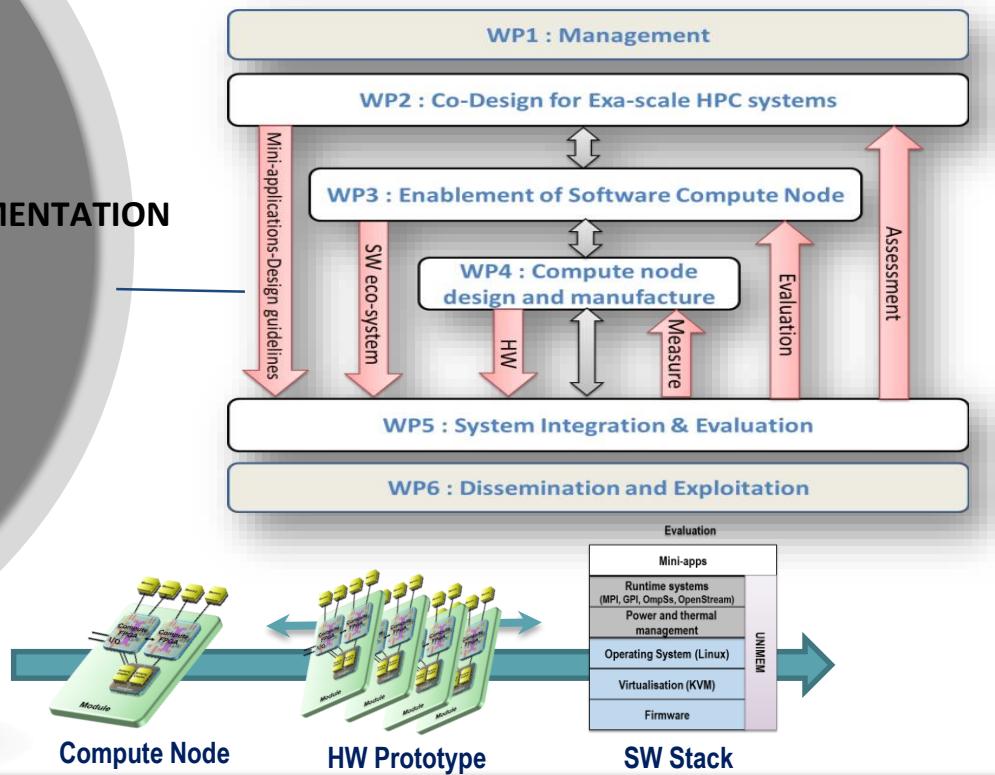


KALRAY



MANCHESTER
1824

The University of Manchester



ExaNoDe at a glance

Key technologies for compute nodes towards a future **Exascale capability**

Key
technologies

Exascale
requirements

ExaNoDe at a glance

Key technologies for compute nodes towards a future Exascale capability

	System Architecture	
Key technologies	<ul style="list-style-type: none">• ARMv8• Coherent islands• Global Address Space	
Exascale requirements	 Energy efficiency  Scalability	

ExaNoDe at a glance

Key technologies for compute nodes towards a future Exascale capability

	System Architecture	Silicon Integration	
Key technologies	<ul style="list-style-type: none">• ARMv8• Coherent islands• Global Address Space	<ul style="list-style-type: none">• 3D Integration:<ul style="list-style-type: none">• Chiplet• Active Interposer• Multi-Chip-Module:<ul style="list-style-type: none">• FPGA, Memory	
Exascale requirements	<ul style="list-style-type: none">➢ Energy efficiency➢ Scalability	<ul style="list-style-type: none">➢ Design/manufacturing costs➢ Heterogeneity/Specialization	

ExaNoDe at a glance

Key technologies for compute nodes towards a future Exascale capability

	System Architecture	Silicon Integration	Software
Key technologies	<ul style="list-style-type: none">• ARMv8• Coherent islands• Global Address Space	<ul style="list-style-type: none">• 3D Integration:<ul style="list-style-type: none">• Chiplet• Active Interposer• Multi-Chip-Module:<ul style="list-style-type: none">• FPGA, Memory	<ul style="list-style-type: none">• FW, OS• Virtualization• Programming models• Runtimes• Mini-apps
Exascale requirements	<ul style="list-style-type: none">➢ Energy efficiency➢ Scalability	<ul style="list-style-type: none">➢ Design/manufacturing costs➢ Heterogeneity/Specialization	<ul style="list-style-type: none">➢ Co-design➢ Scalability

All combined in an integrated prototype

2. System Architecture and Integration

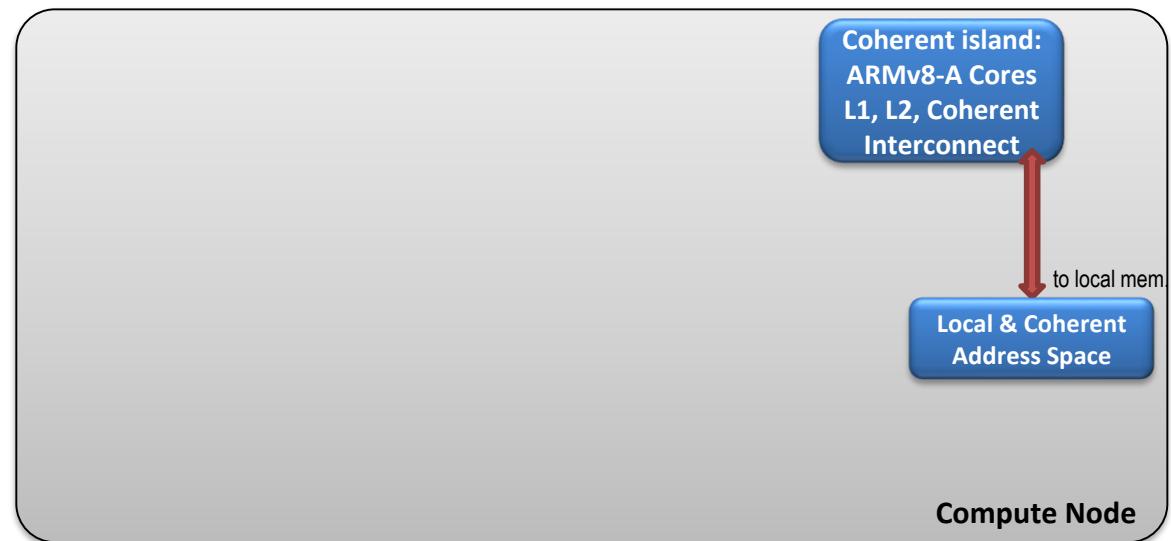
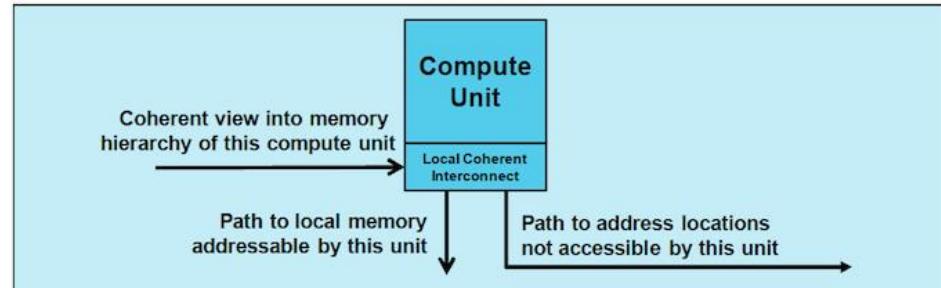
ExaNoDe Architectural Approach

Source: John Goodacre – DATE'13



■ Compute unit (coherent island)

- Several ARMv8 cores
- Local coherent interconnect
- Path to local memory
- Path “to remote” memory locations
- Path “from remote” compute units



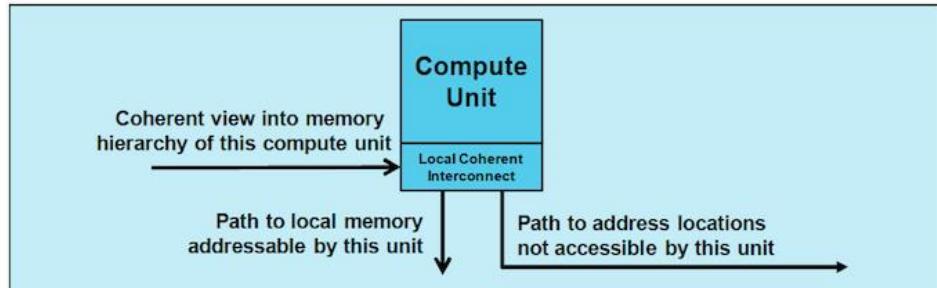
ExaNoDe Architectural Approach

Source: John Goodacre – DATE'13



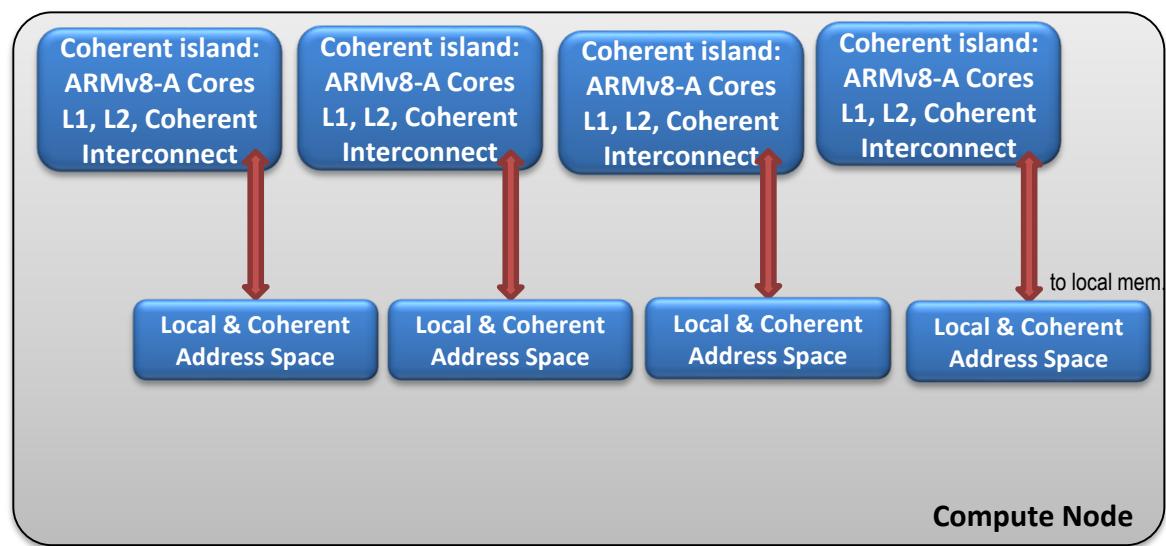
■ Compute unit (coherent island)

- Several ARMv8 cores
- Local coherent interconnect
- Path to local memory
- Path “to remote” memory locations
- Path “from remote” compute units



■ Compute Node

- Scale-out ...



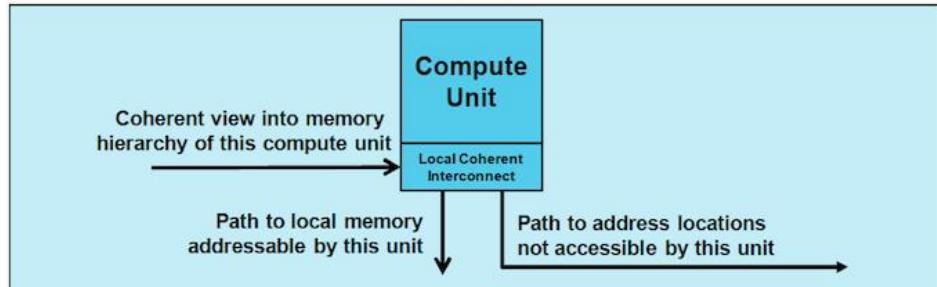
ExaNoDe Architectural Approach

Source: John Goodacre – DATE'13



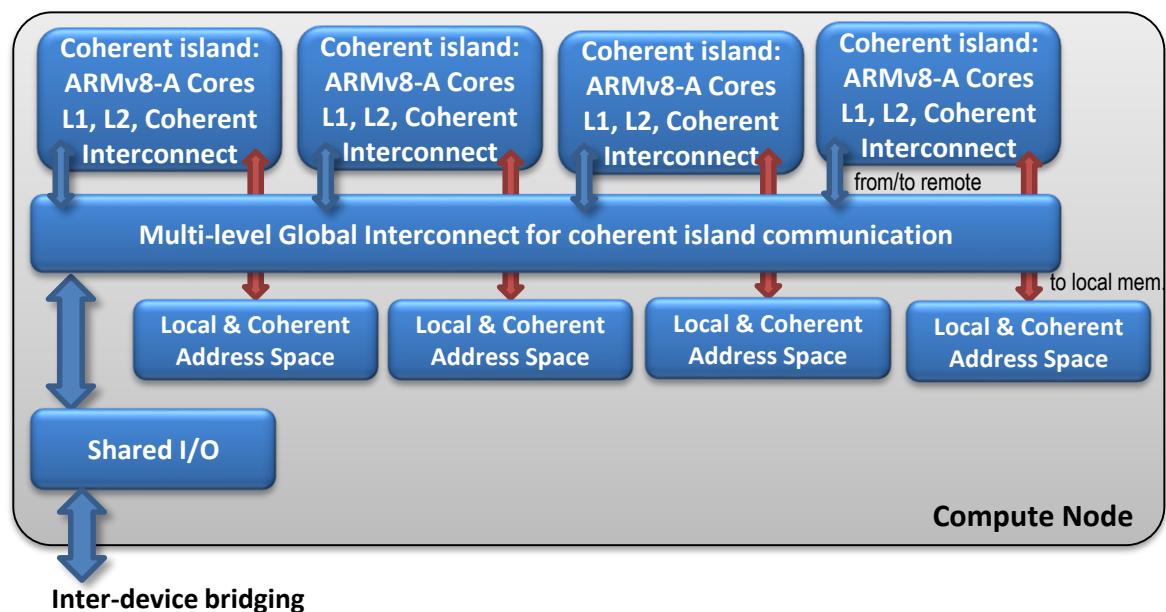
▪ Compute unit (coherent island)

- Several ARMv8 cores
- Local coherent interconnect
- Path to local memory
- Path “to remote” memory locations
- Path “from remote” compute units



▪ Compute Node

- Scale-out ...
- ... with “from” and “to” remote port connections to the global network



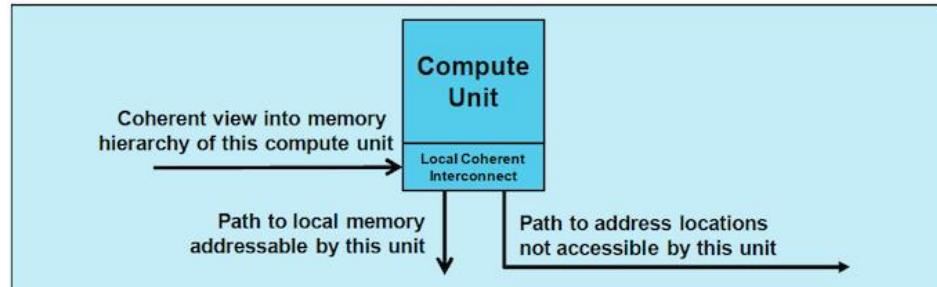
ExaNoDe Architectural Approach

Source: John Goodacre – DATE'13



▪ Compute unit (coherent island)

- Several ARMv8 cores
- Local coherent interconnect
- Path to local memory
- Path “to remote” memory locations
- Path “from remote” compute units

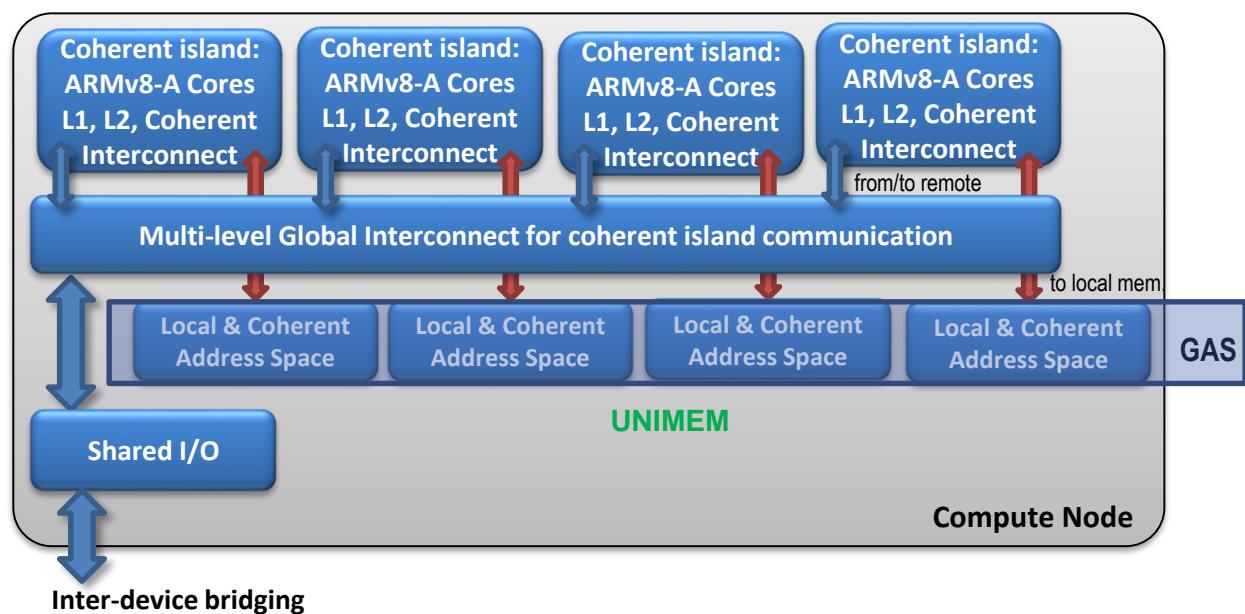


▪ Compute Node

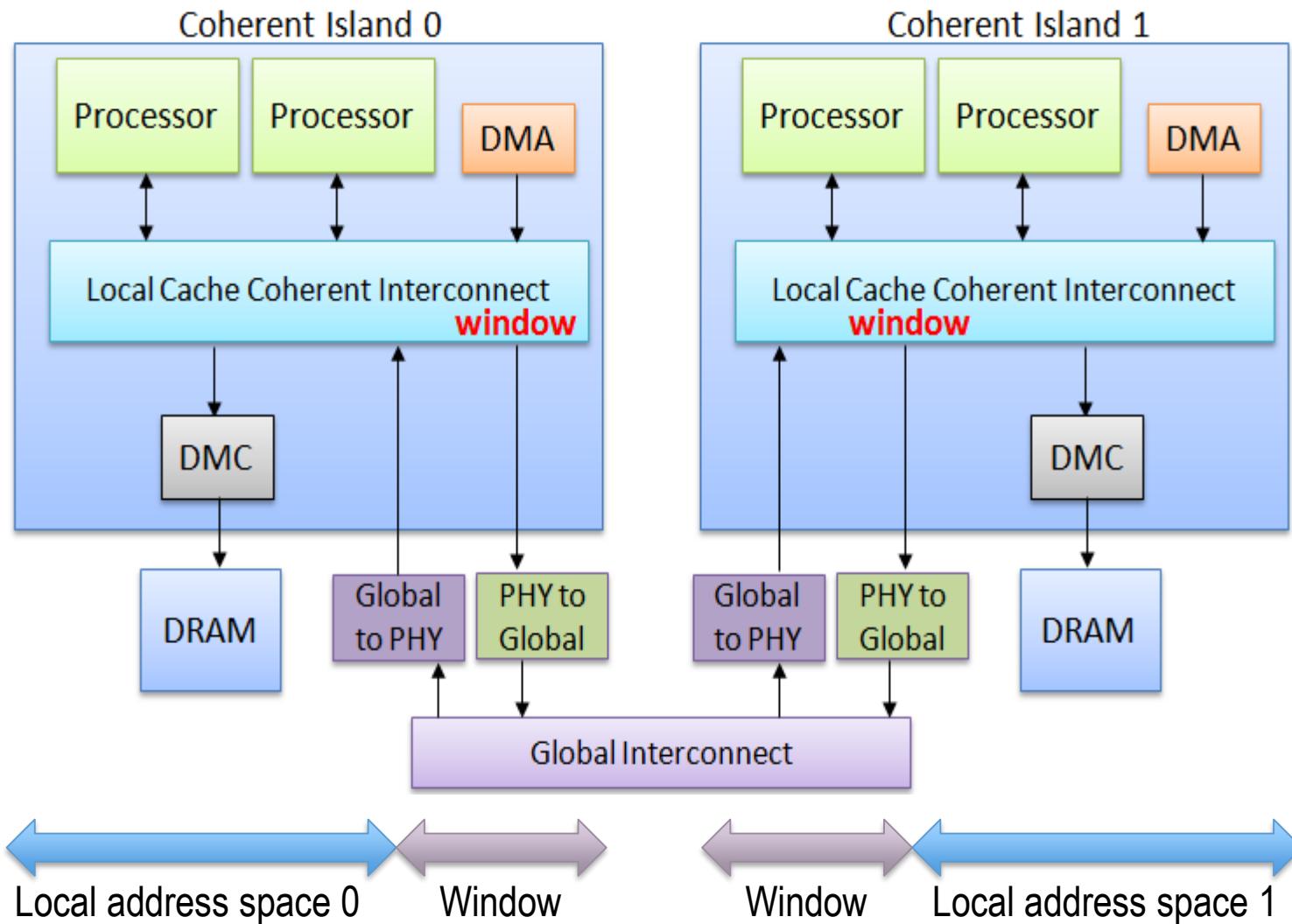
- Scale-out ...
- ... with “from” and “to” remote port connections to the global network

▪ Memory scheme

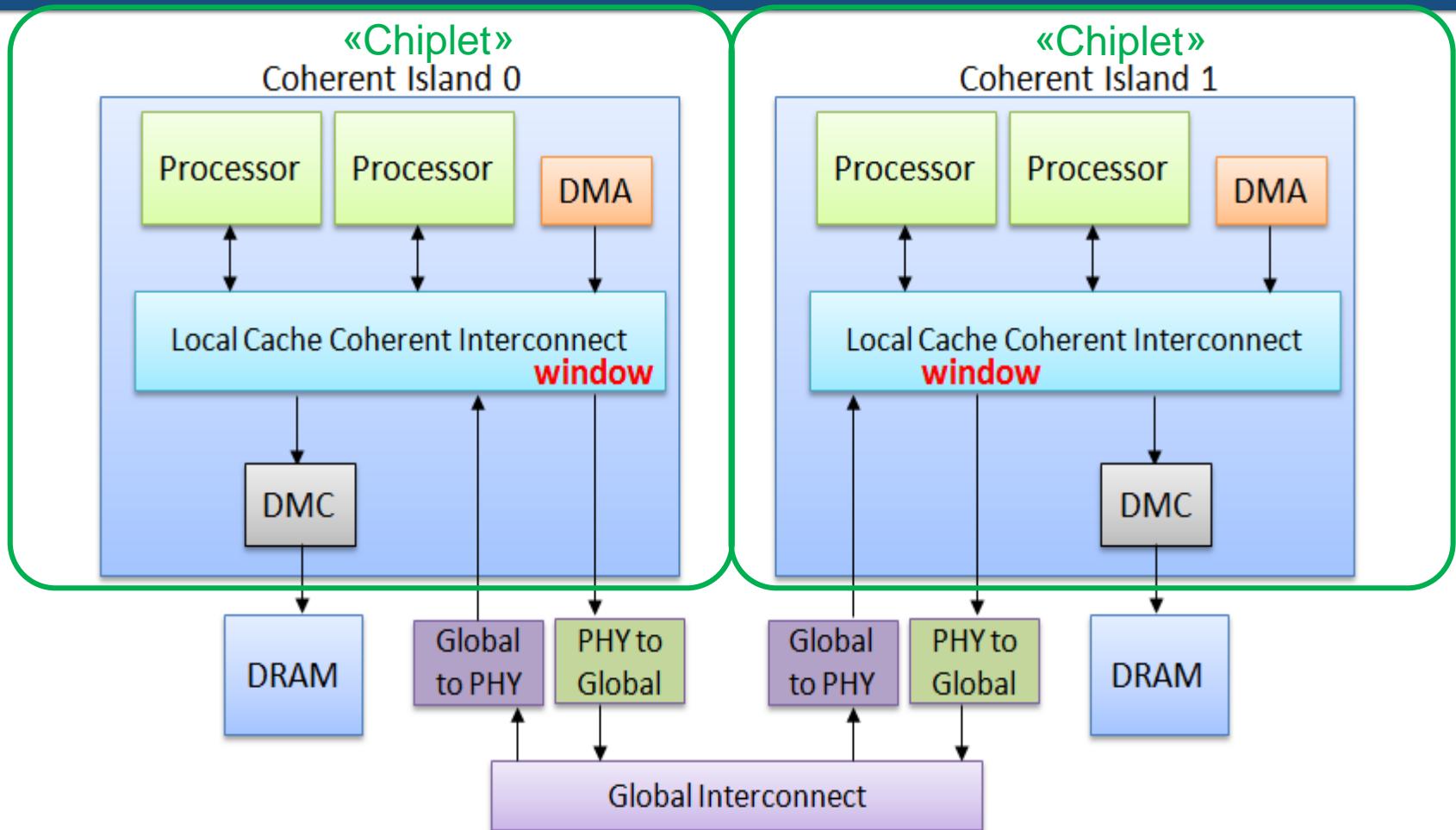
- Global Address Space and remote coherent access with UNIMEM



UNIMEM: Partitioned Global Address Space

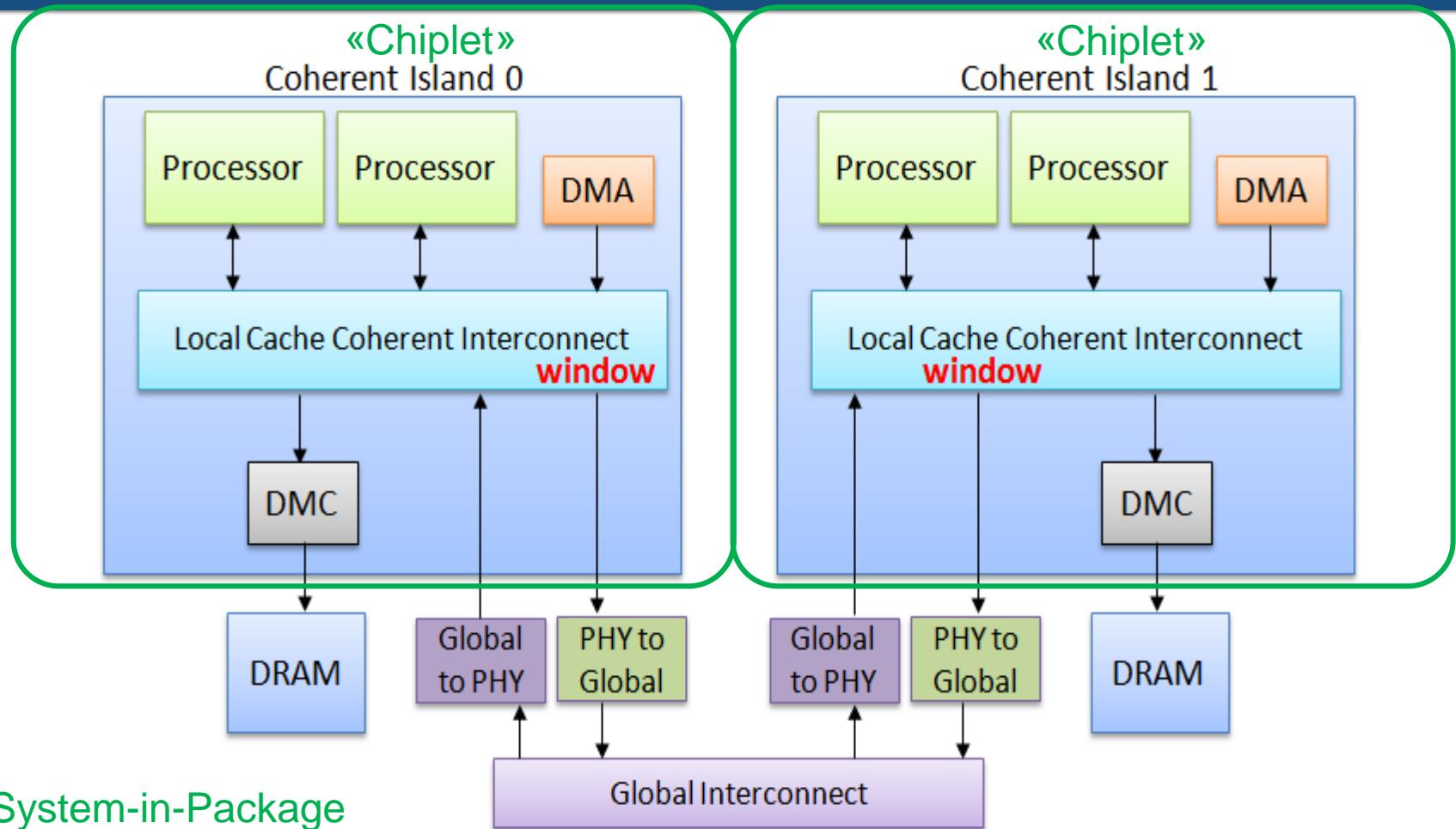


Combined Architecture/Integration



ExaNoDe

Combined Architecture/Integration

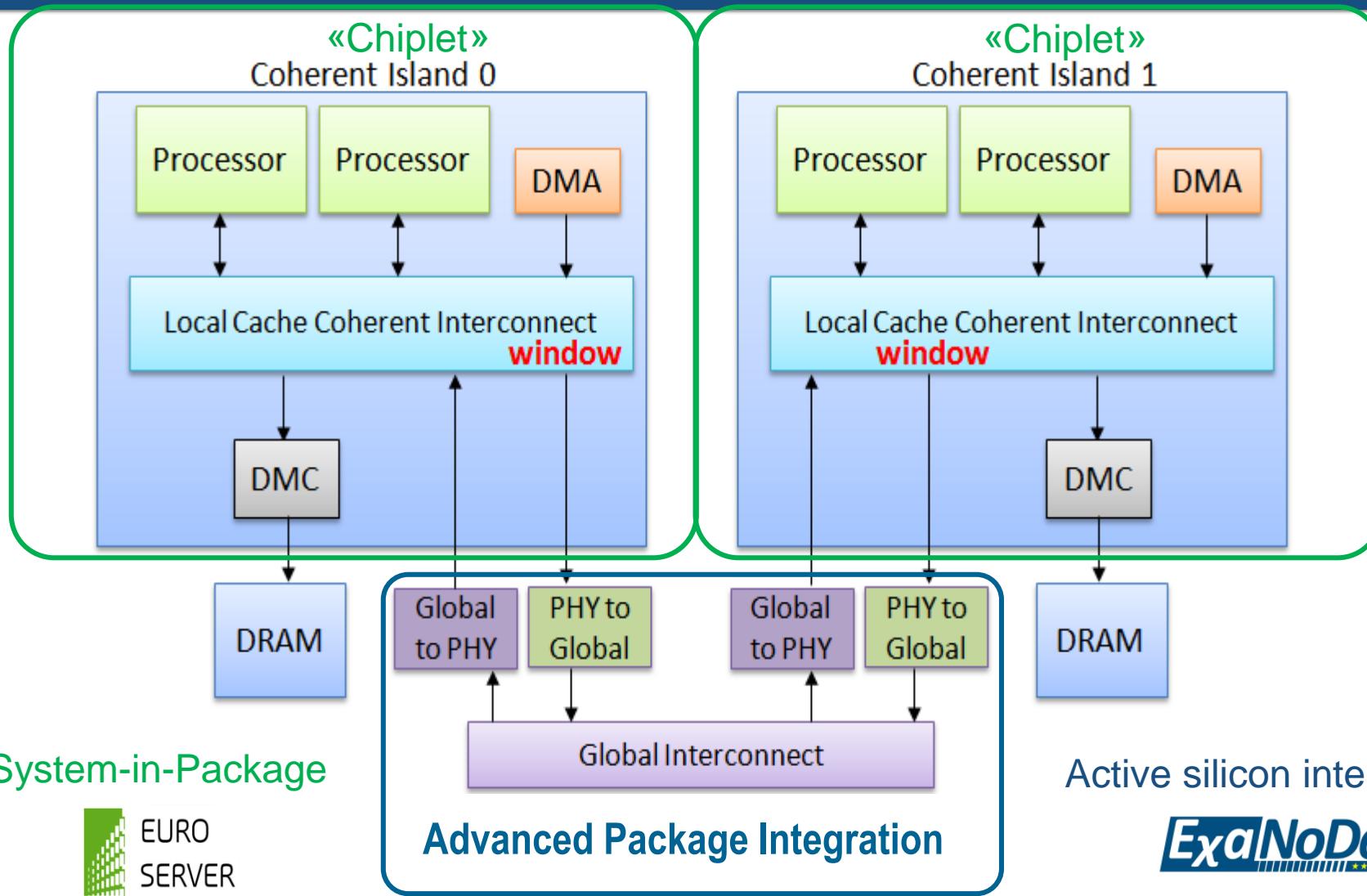


System-in-Package



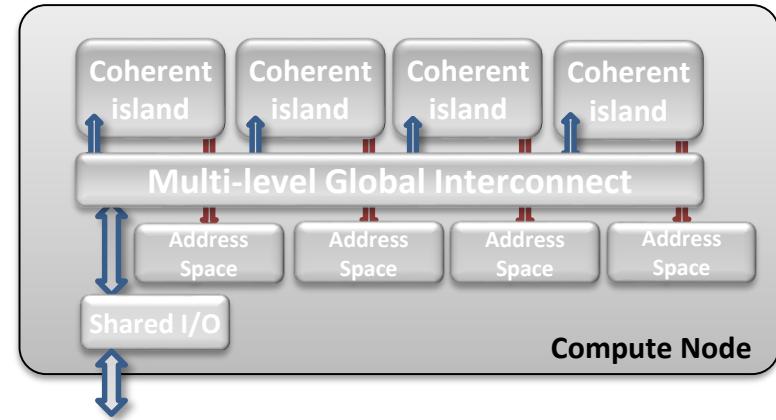
ExaNoDe

Combined Architecture/Integration



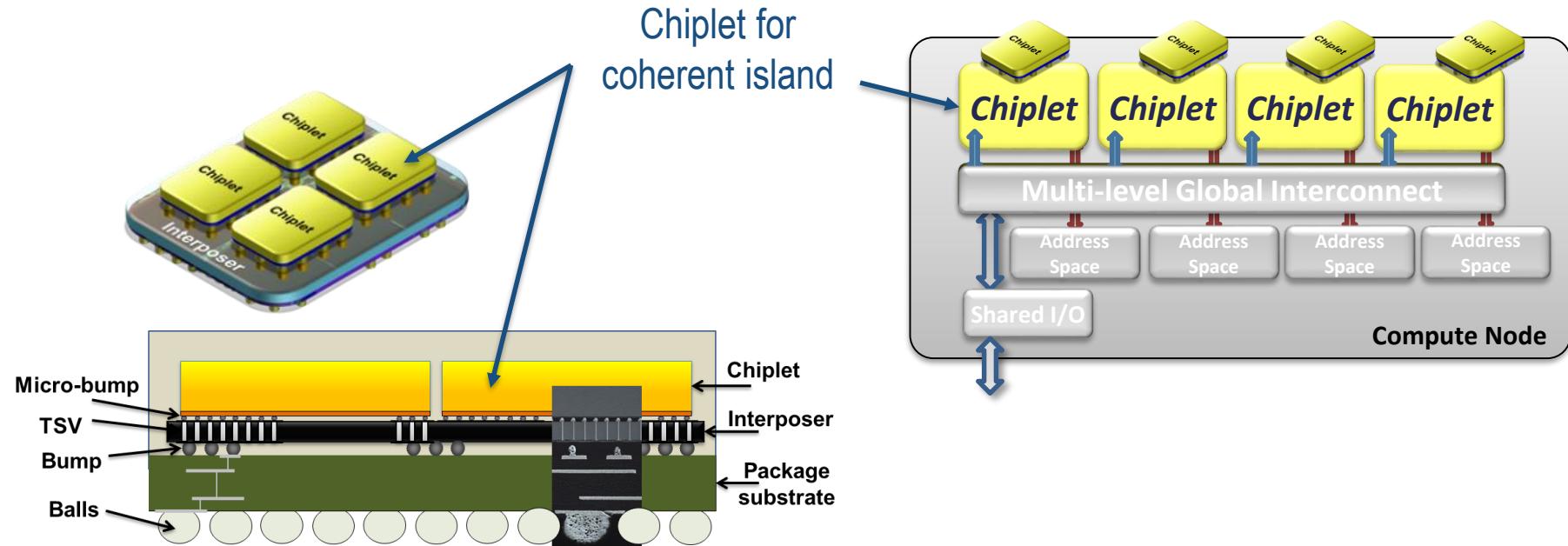
- 3D integration with active silicon interposer and chiplets

Compute node architecture



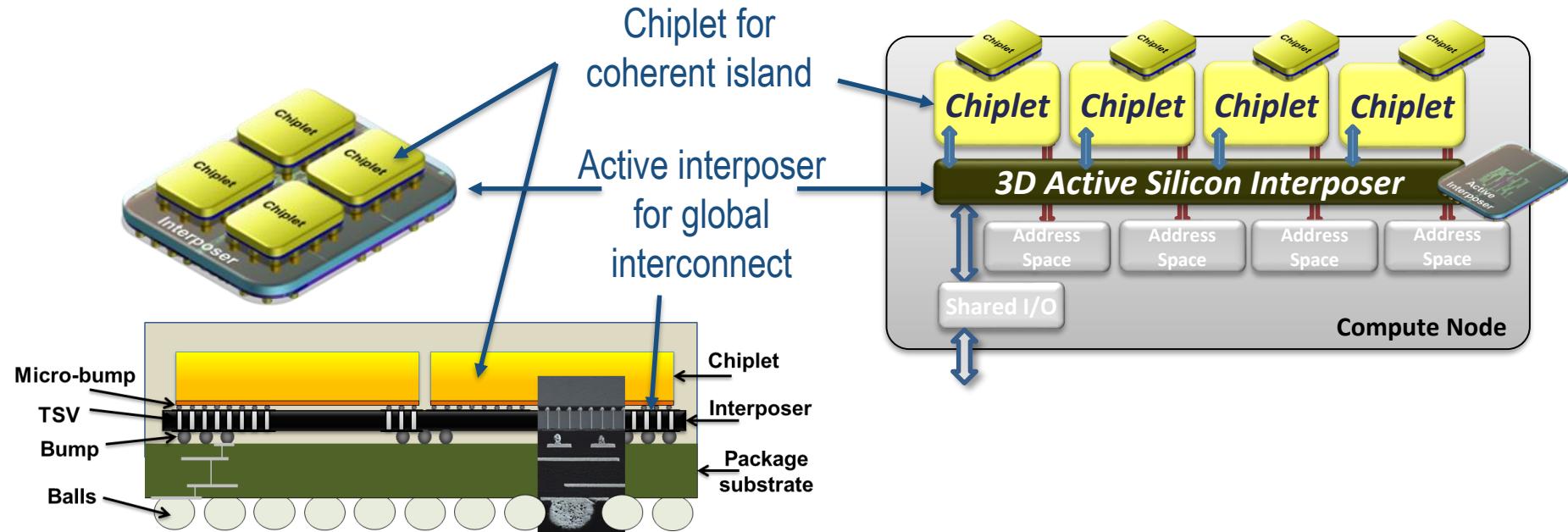
ExaNoDe Integration Approach

■ 3D integration with active silicon interposer and chiplets



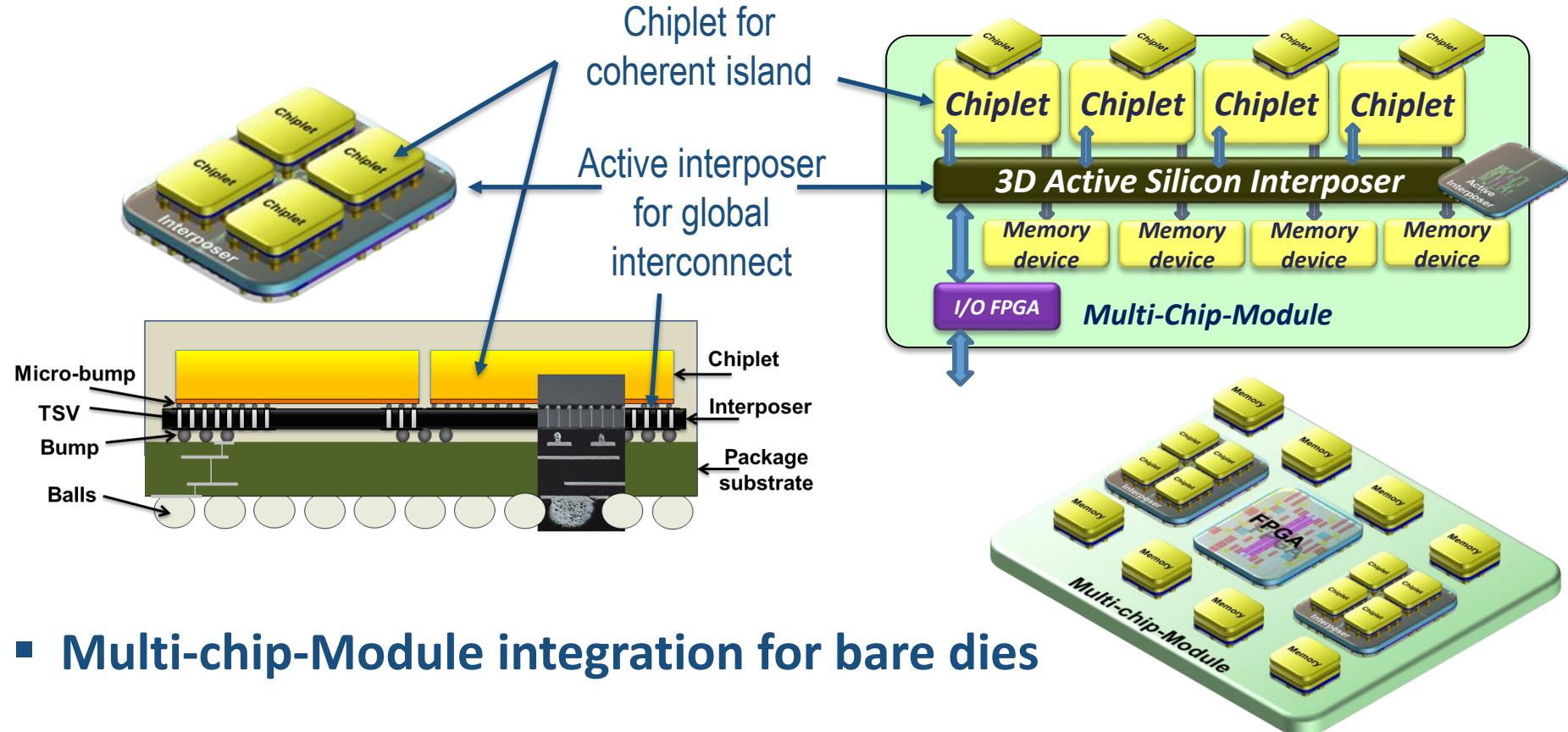
ExaNoDe Integration Approach

■ 3D integration with active silicon interposer and chiplets



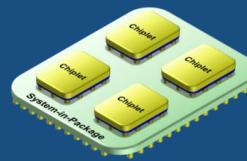
ExaNoDe Integration Approach

- 3D integration with active silicon interposer and chiplets



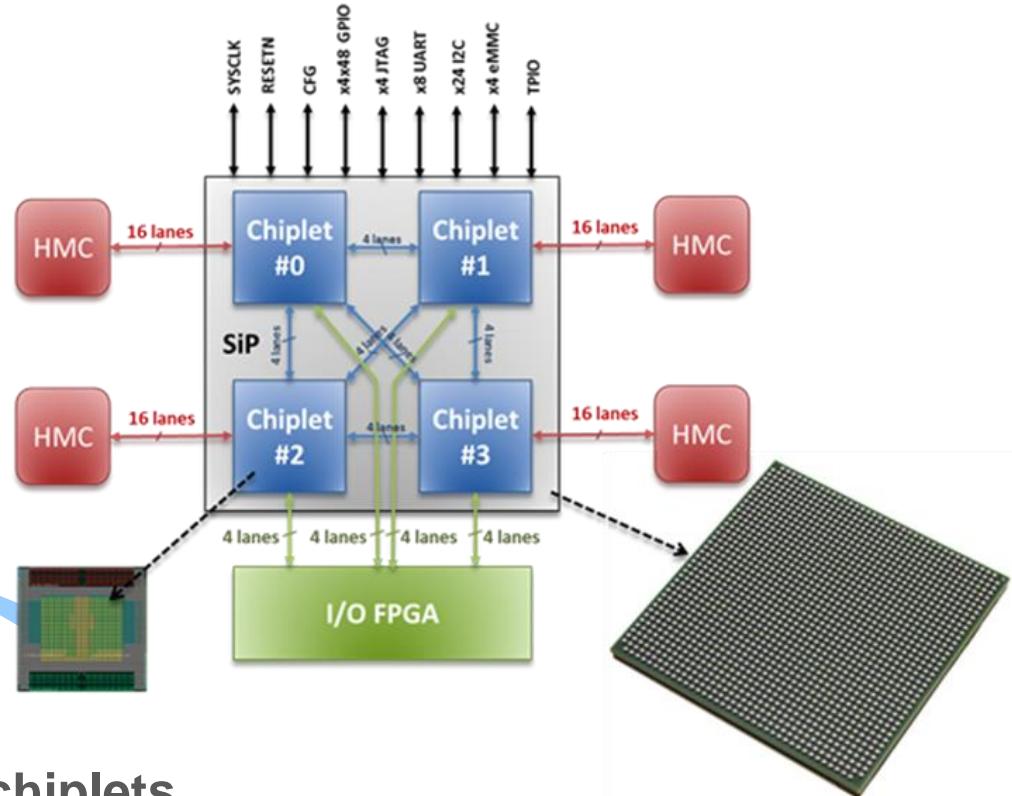
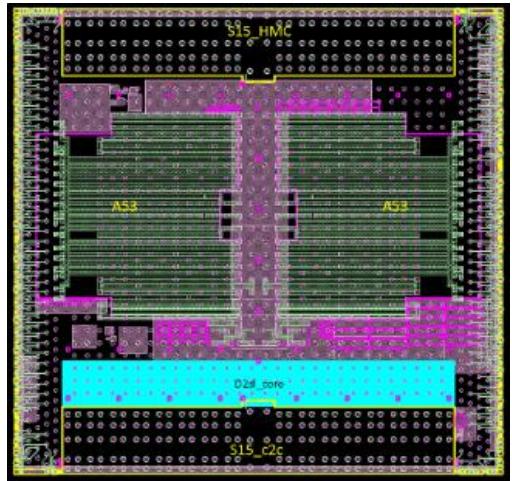
- Multi-chip-Module integration for bare dies

EUROSERVER: Package



EURO
SERVER

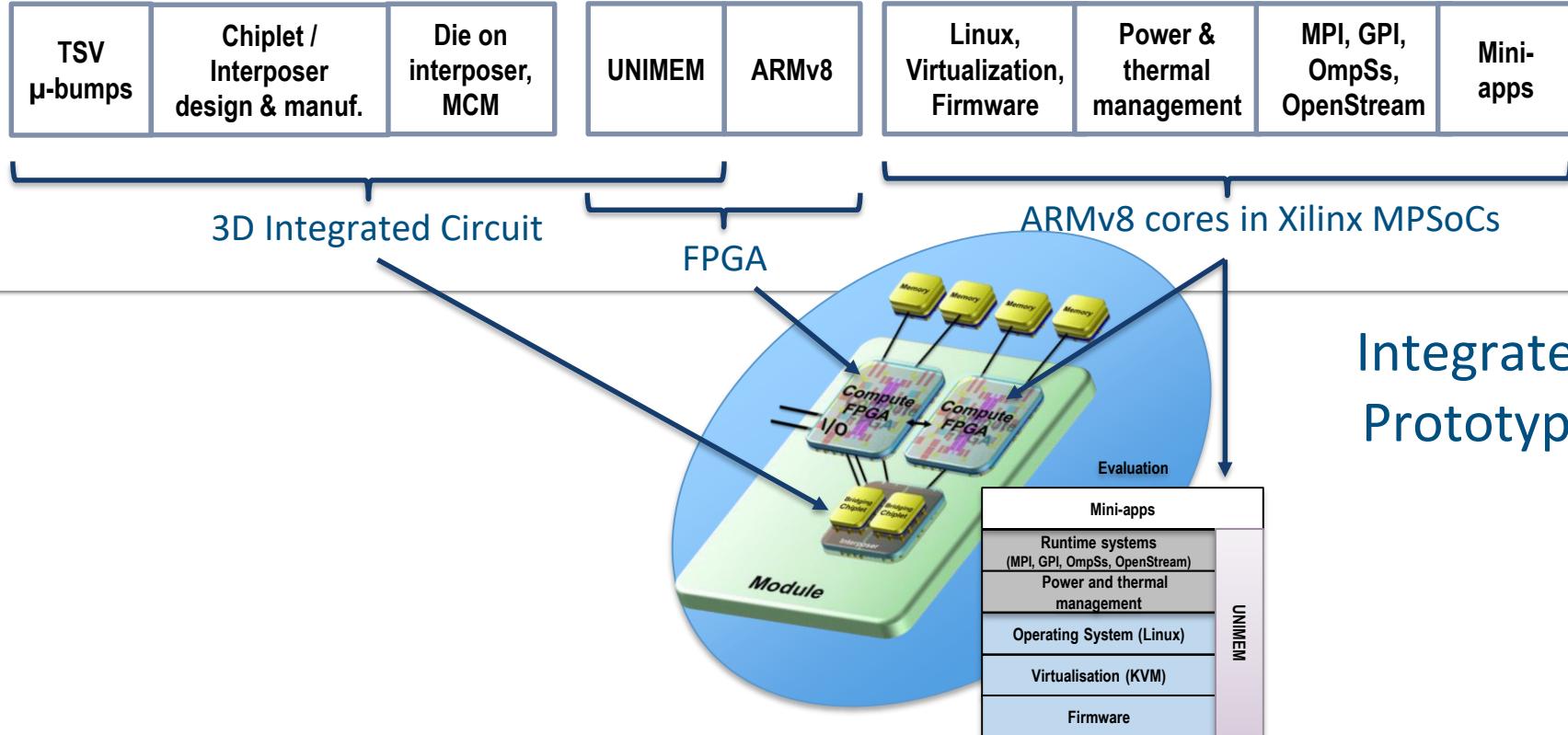
- Multi-Core System with four chiplets in a 40x40 package



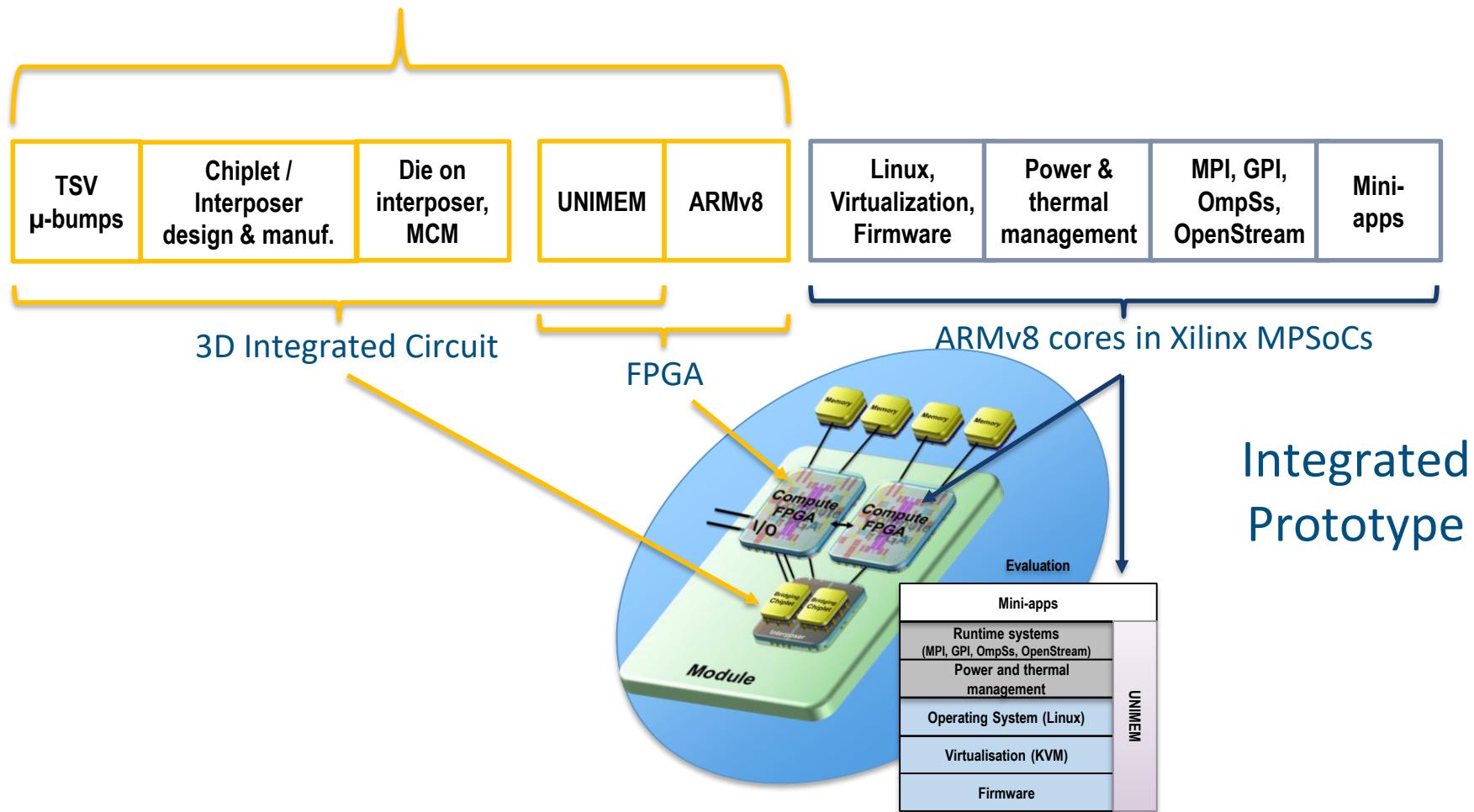
- A package contains 4 chiplets
- Each chiplet contains 2 quad-core ARMv8 A53
- System of 32 A53 cores in a Package

3. *ExaNoDe Prototype*

3. ExaNoDe Prototype



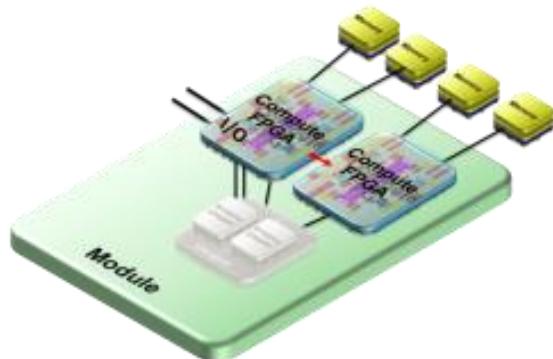
ExaNoDe Hardware Prototype



ExaNoDe Objective: Deliver a Multi Chip Module

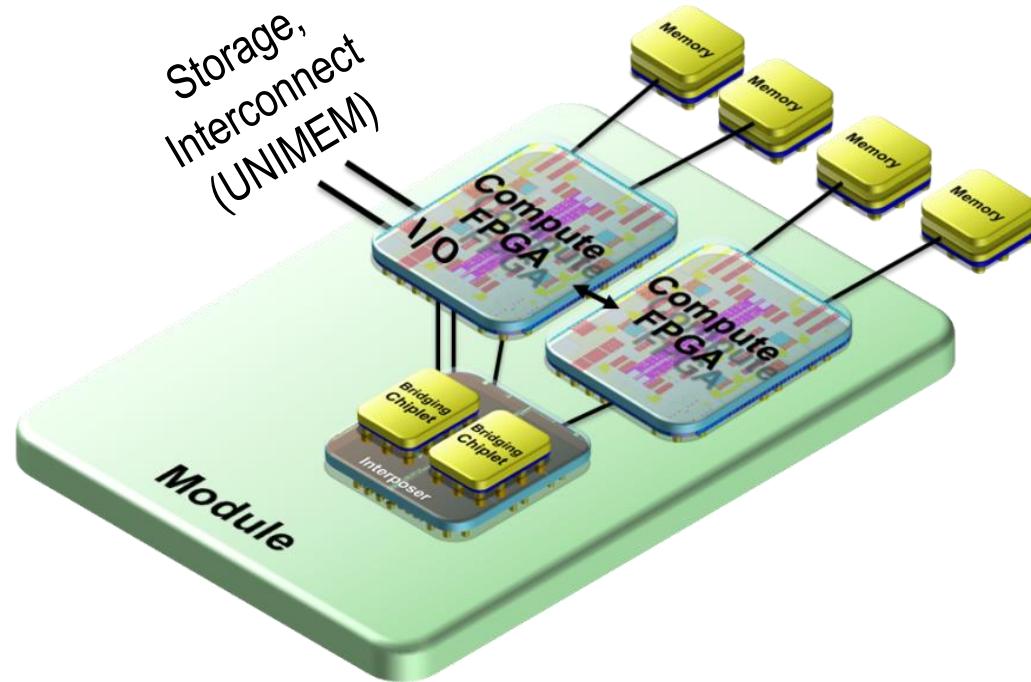
MCM Step 1: Q4-2017

- Without 3D-IC
- UNIMEM FPGA

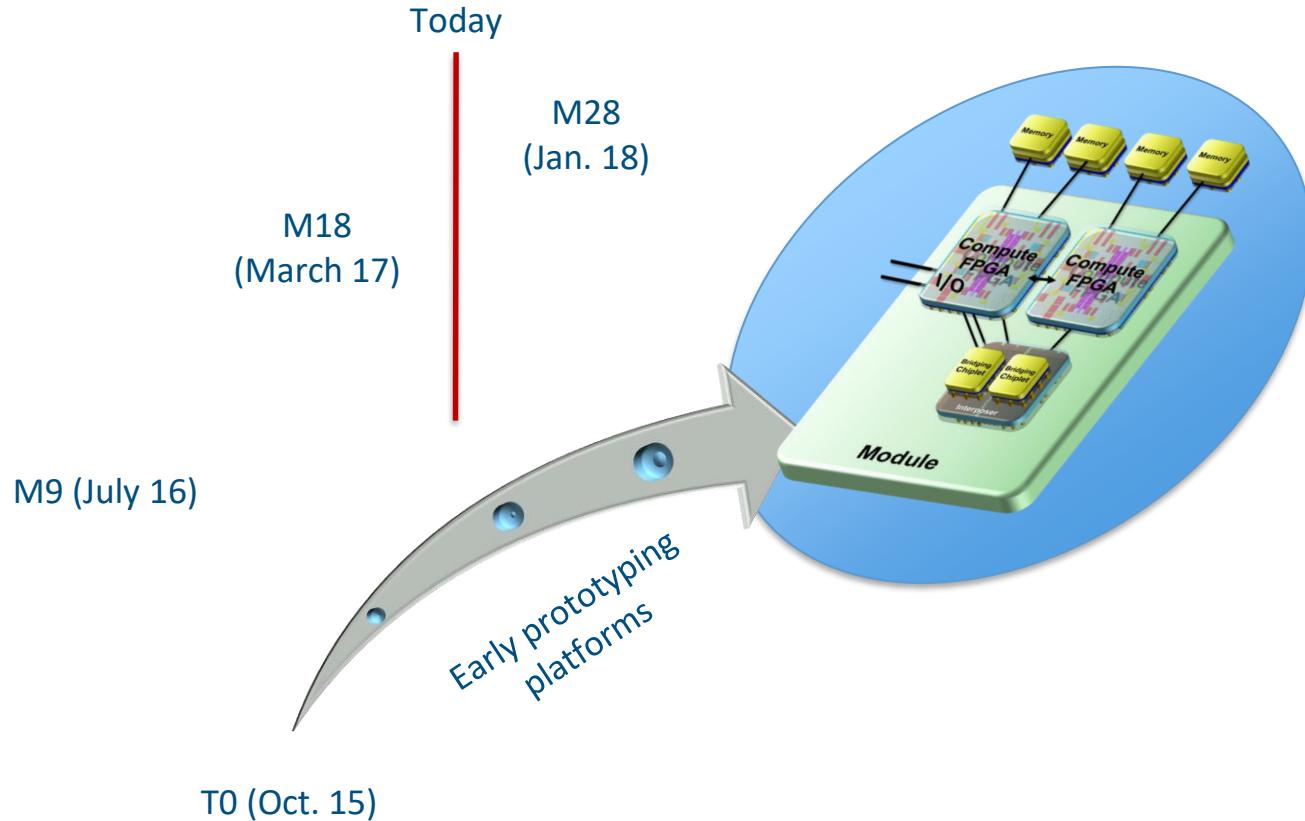


MCM Step 2: Q3-2018

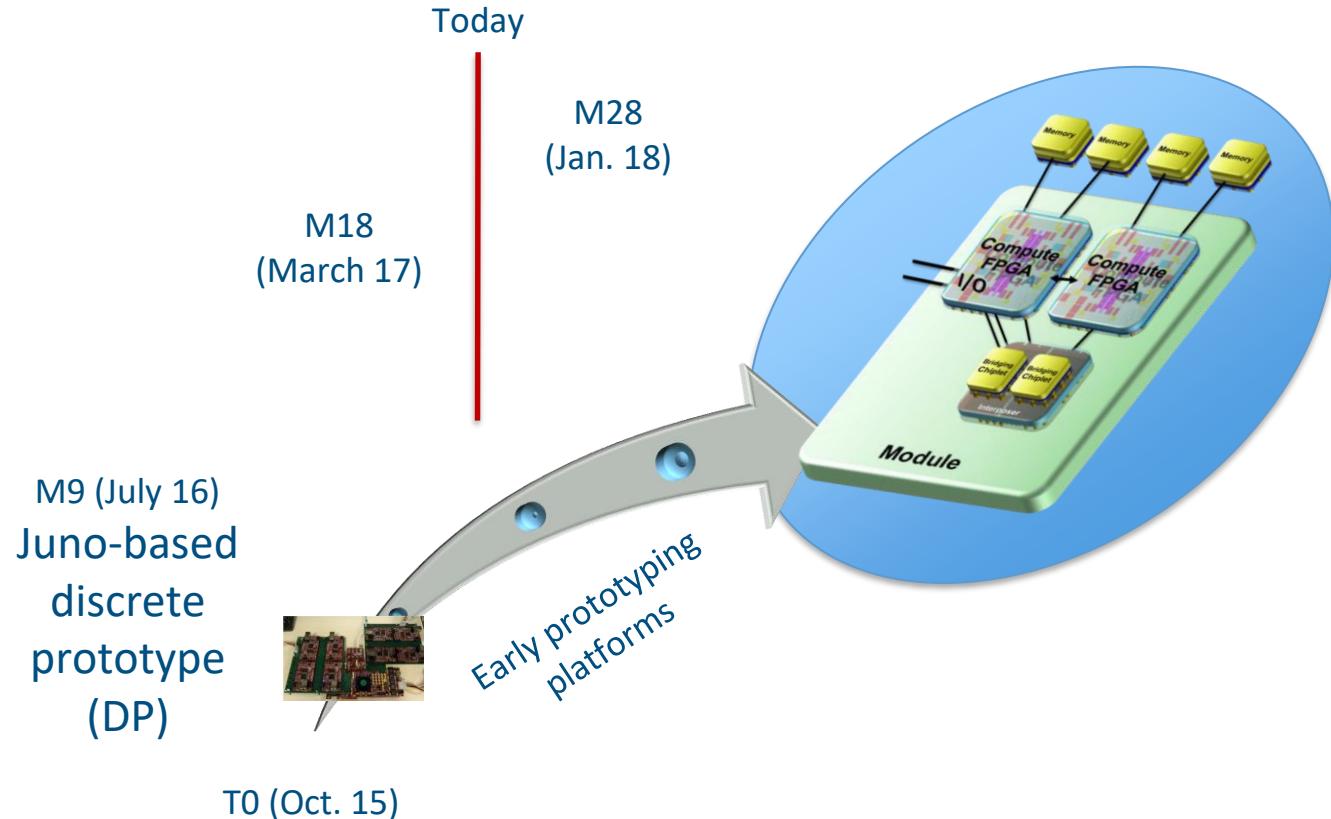
- With 3D-IC
- UNIMEM-capable 3D-IC



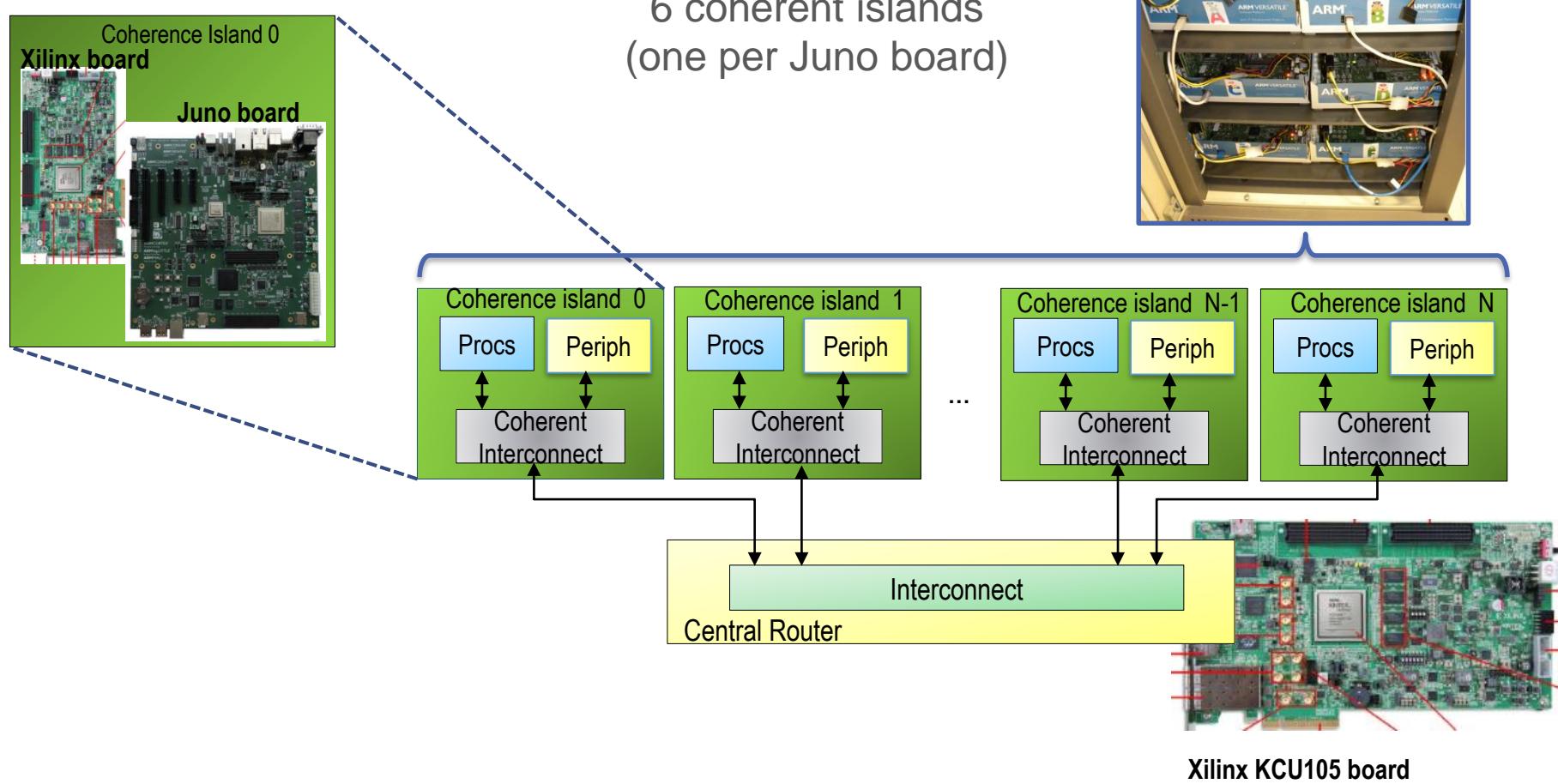
Early prototyping platforms



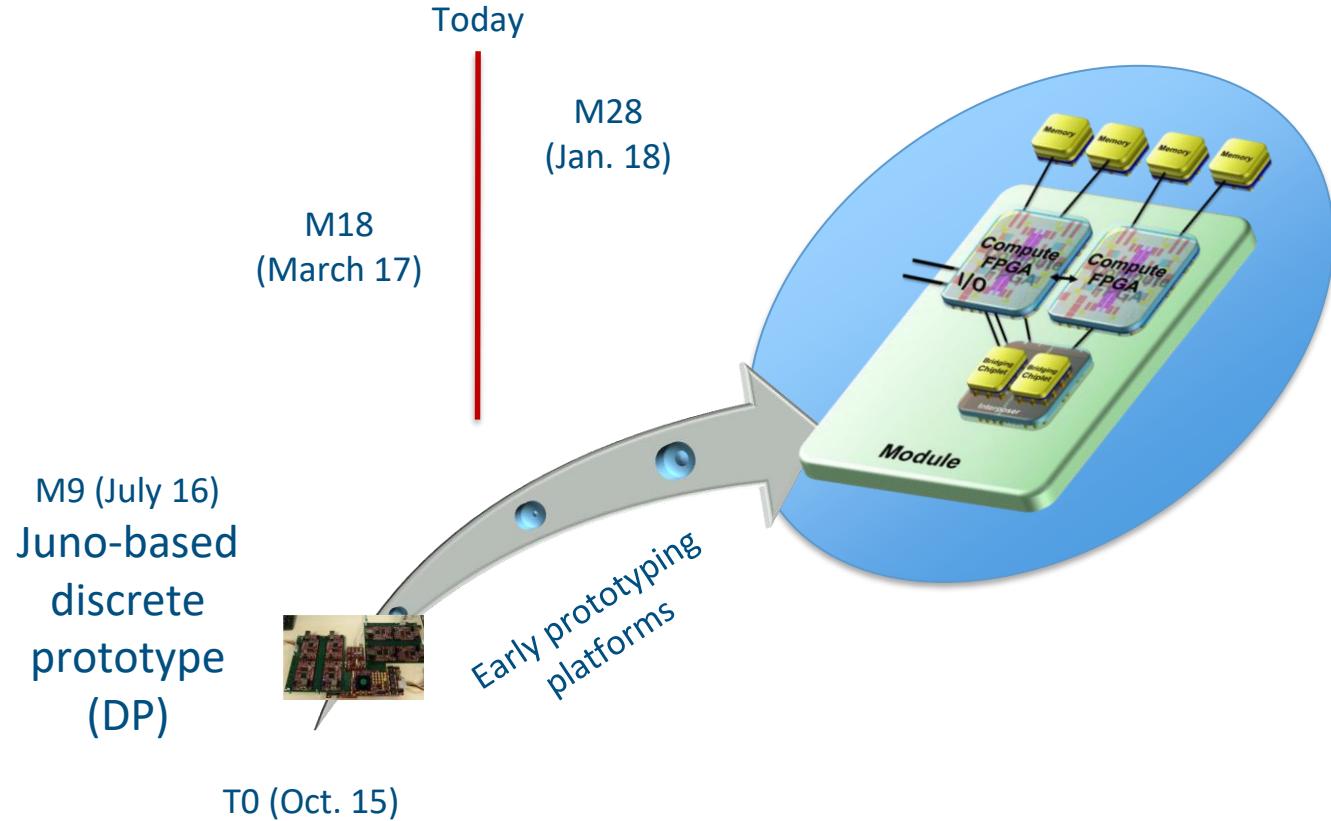
Early prototyping platforms



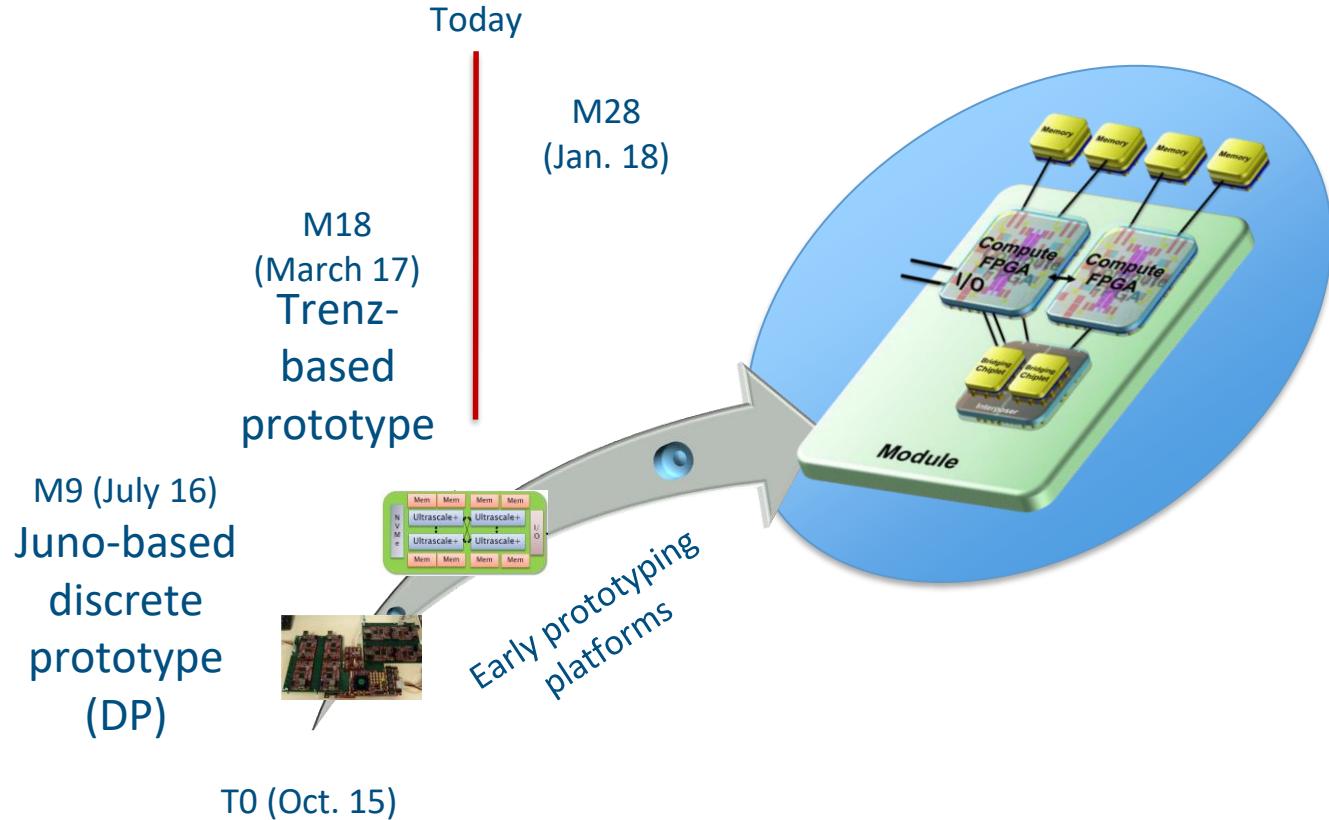
High-level Architecture of 64-bit Discrete Prototype (DP)



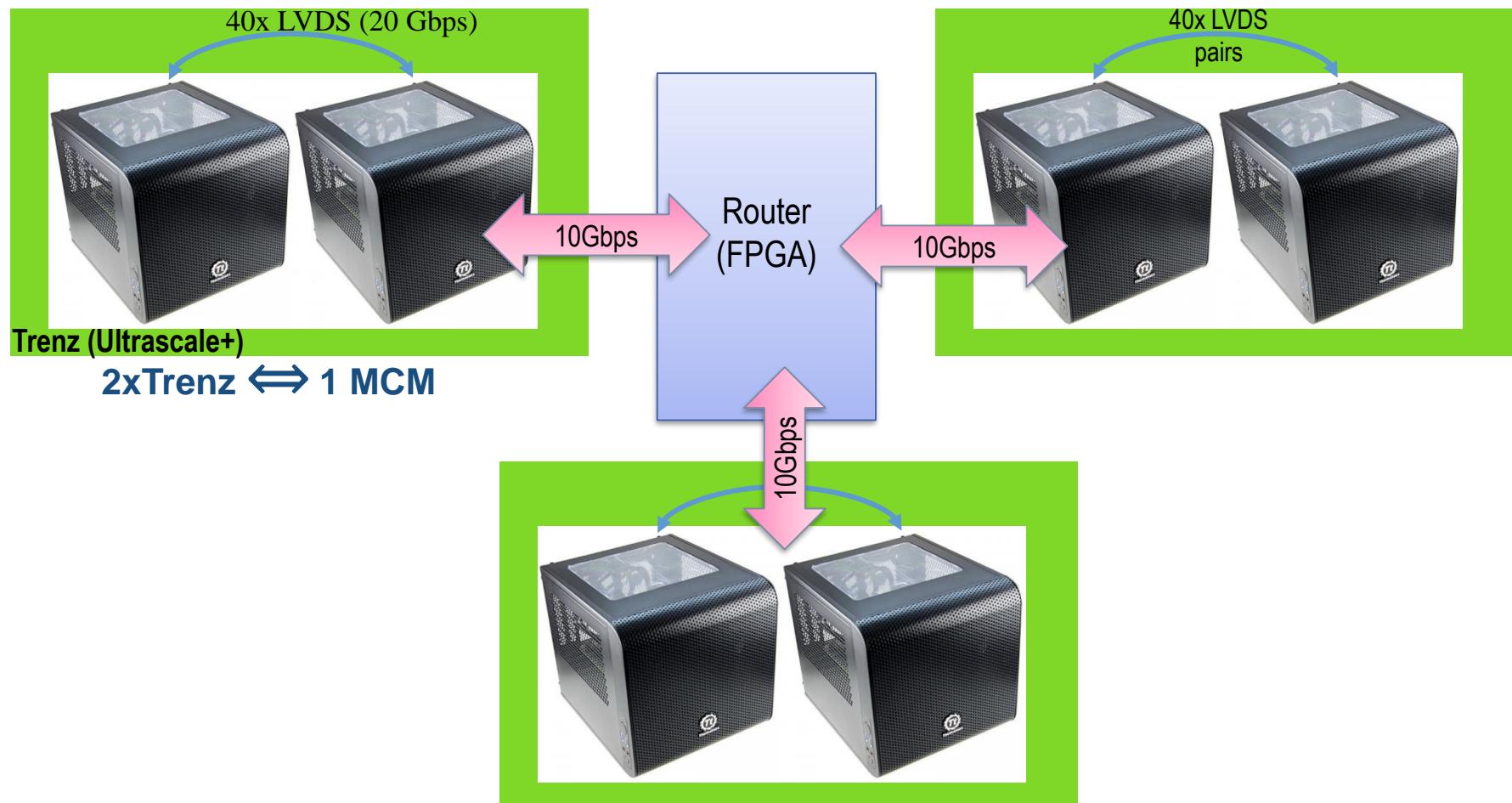
Early prototyping platforms



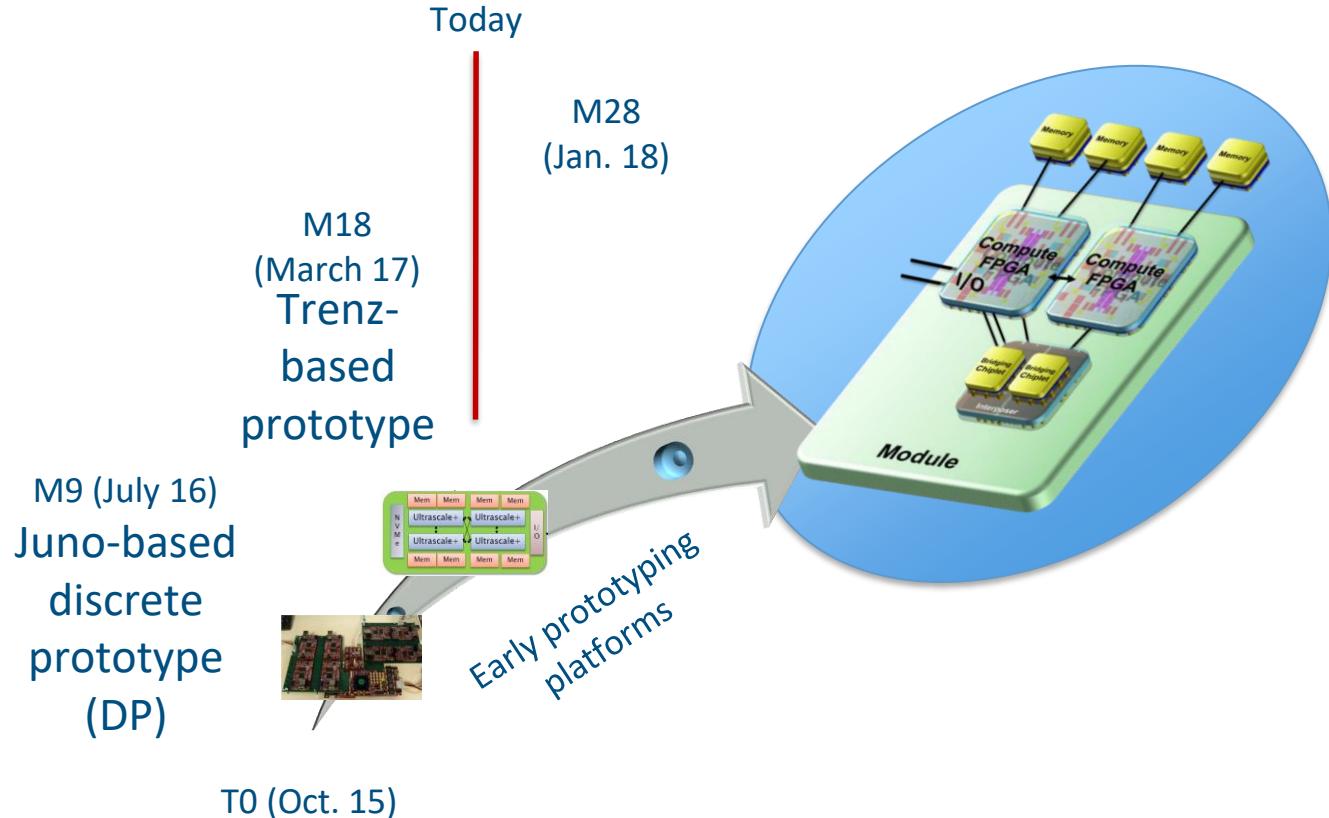
Early prototyping platforms



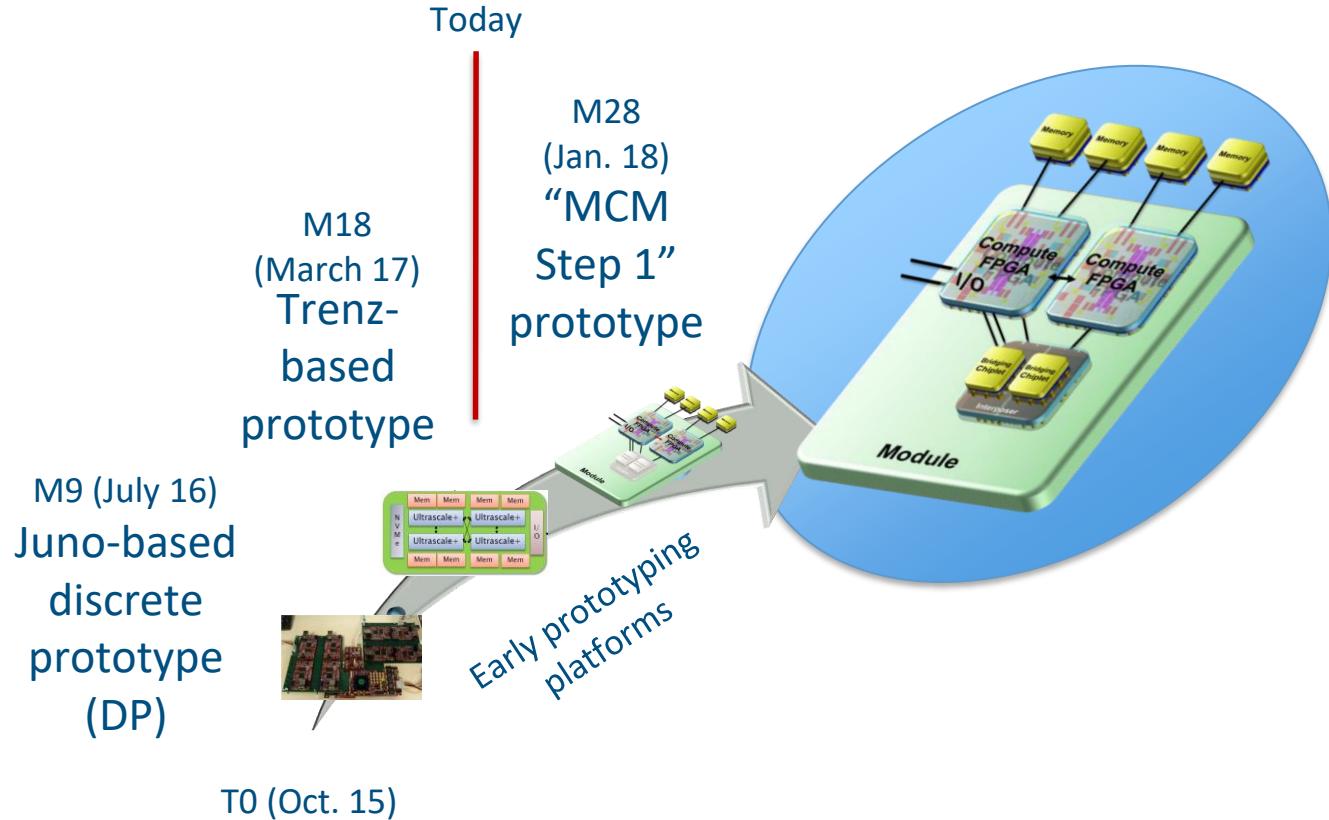
Trenz-based Prototype



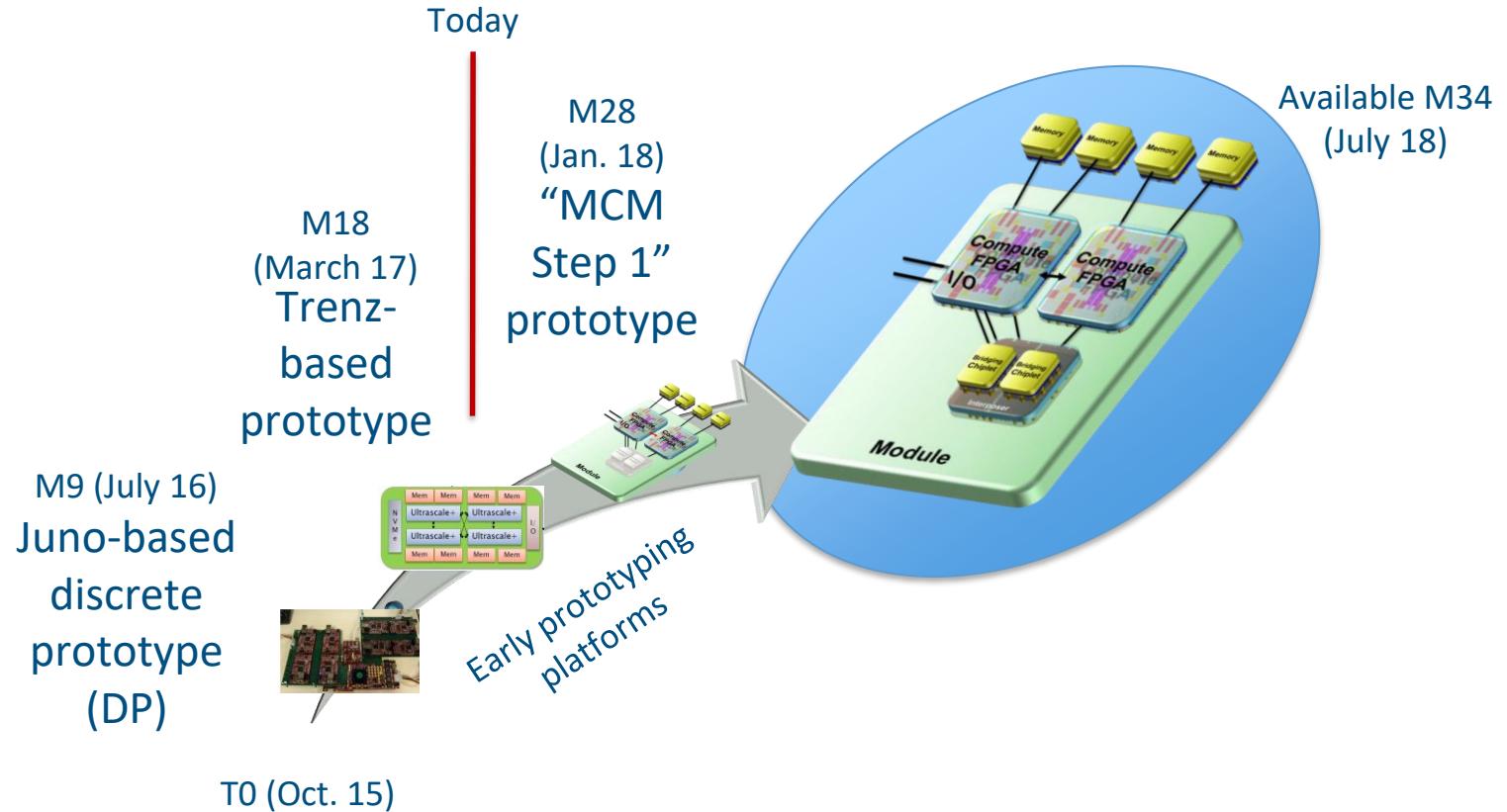
Early prototyping platforms



Early prototyping platforms



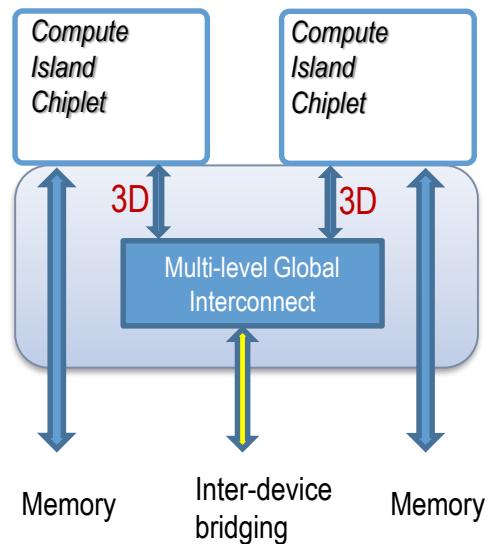
Early prototyping platforms



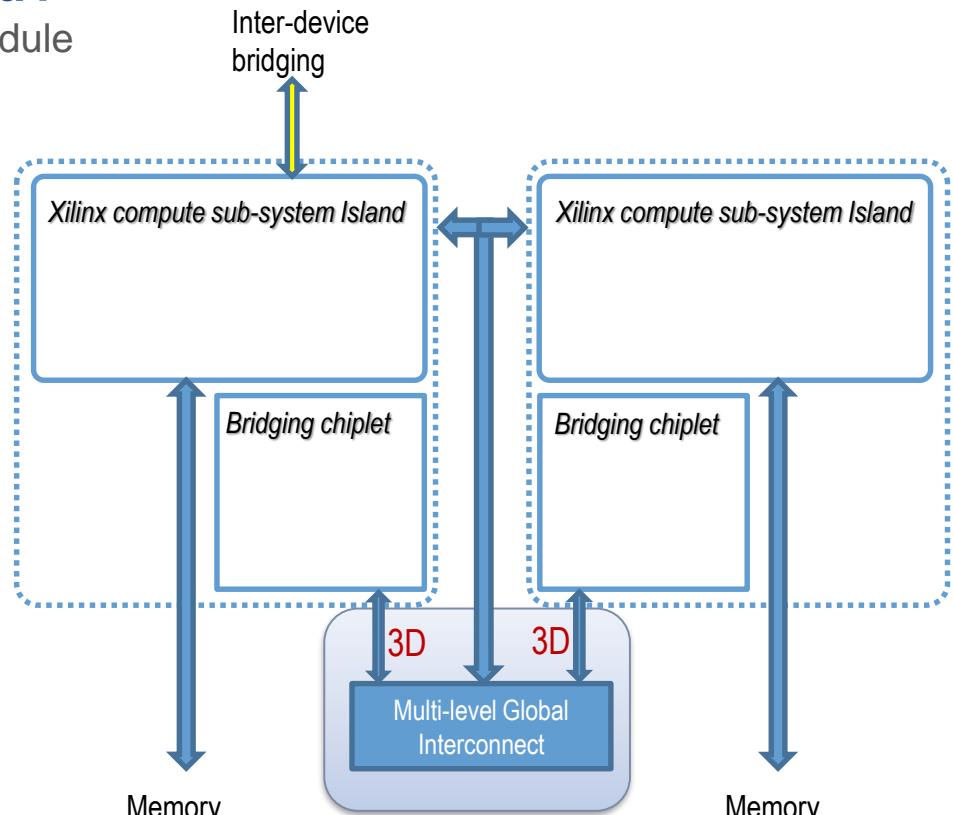
From approach to prototype

➤ What has to be demonstrated?

- 3D technology capability for HPC module
- UNIMEM capability
- Scalability:
 - in term of architecture
 - in term of performance

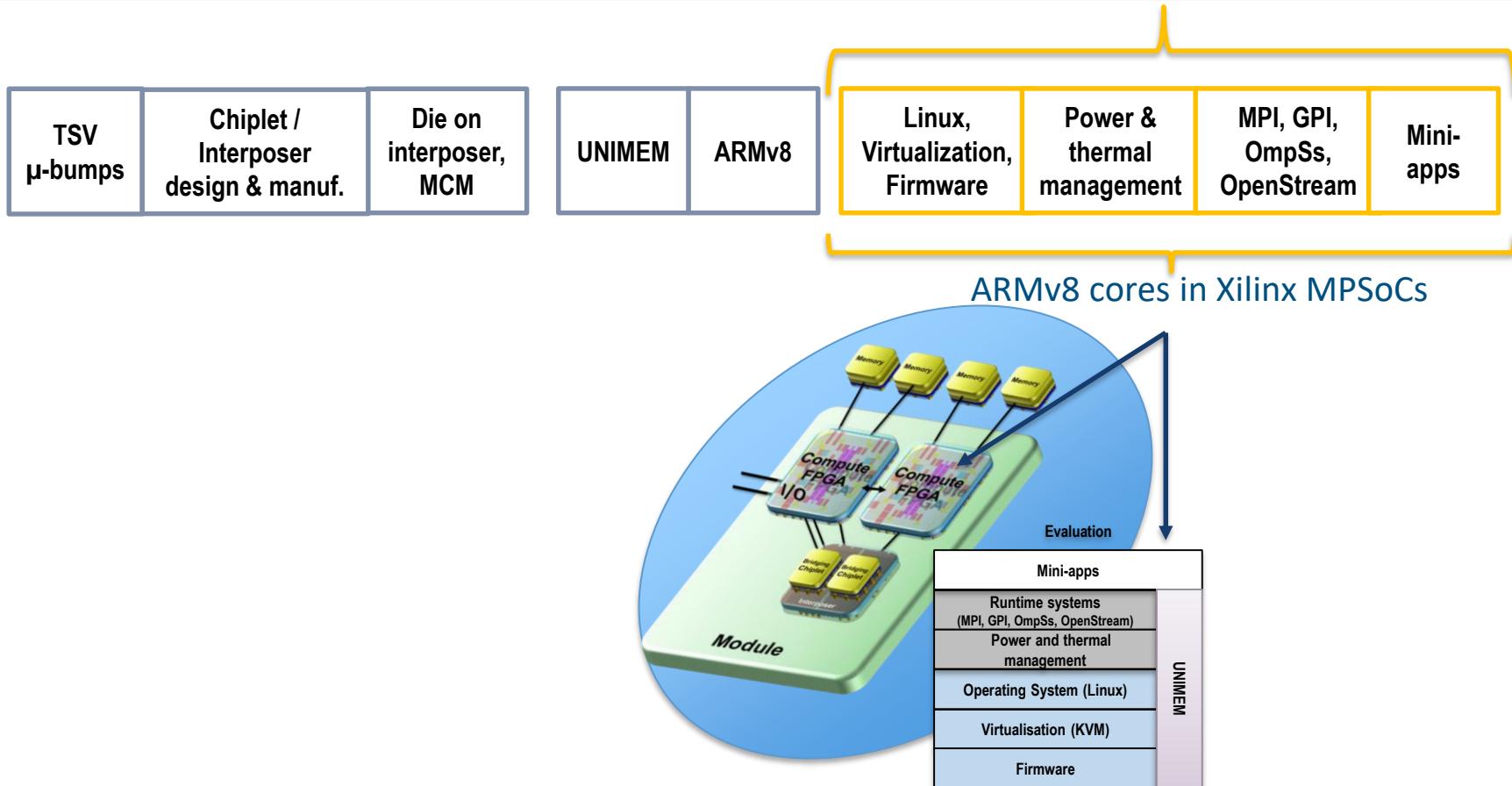


Approach



Prototype

4. Application and Software



ExaNoDe SW objectives

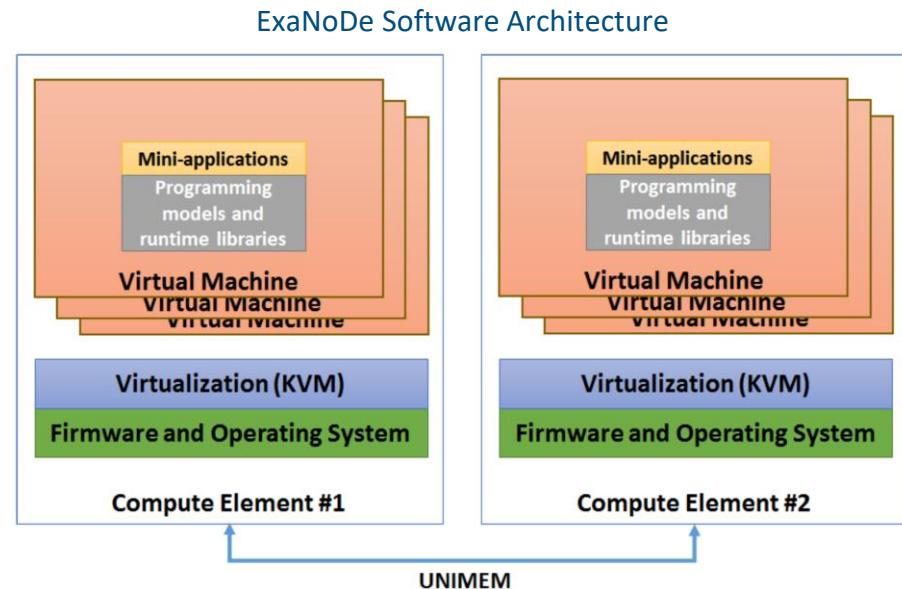
- **Mini-apps for co-design process**
 - Select of HPC applications to co-design the ExaNoDe architecture.

- **Software infrastructure**

- Deploy a software ecosystem for the ARM-based compute node ...
 - in conjunction with the UNIMEM system architecture.

- **Evaluation**

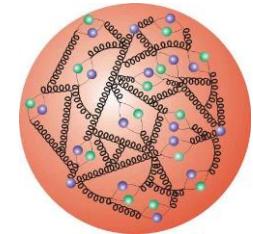
- Analyse and compare ExaNoDe architecture.



Application Portfolio

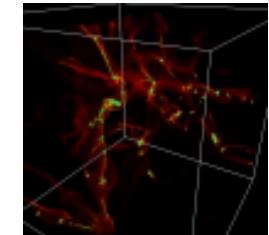
■ **BQCD**

- Simulation of Quantum Chromodynamics on a lattice (LQCD)



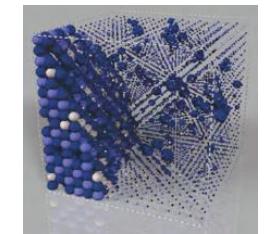
■ **HydroC**

- Simplified version of the astrophysical code RAMSES



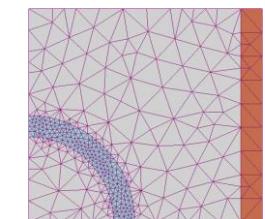
■ **KKRnano**

- Materials science application based on the Density Functional Theory (DFT) method



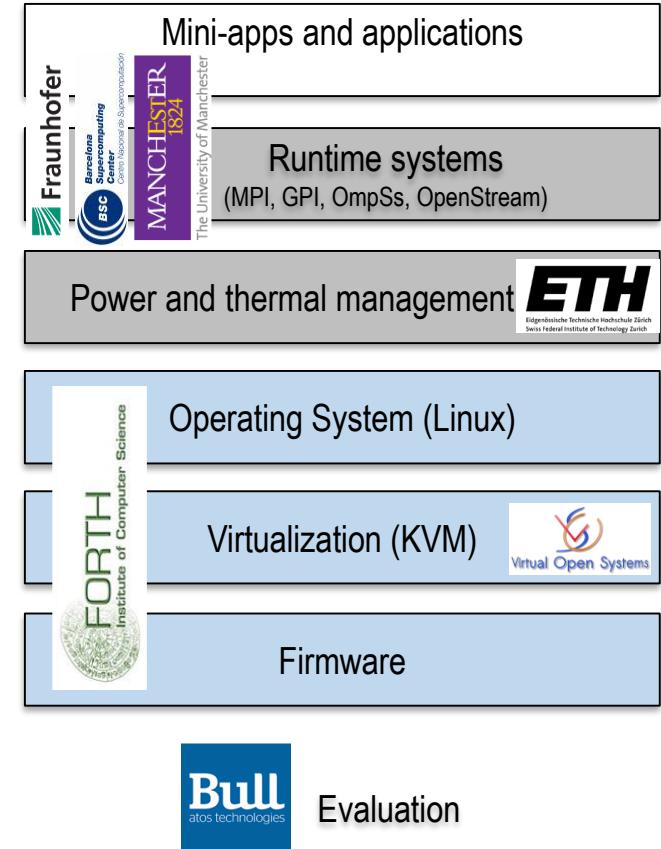
■ **MinIFE**

- Mini-application that mimics the finite element generation, assembly and solution for an unstructured grid problem



SW stack: deployment support

- **Deliver firmware and Numa-aware OS**
 - For multi-board and ExaNoDe prototypes
(Unimem data movement, memory protection and integration with peripheral devices; OS interface for light RDMA operations)
- **Provide virtualization layer**
(next slide)
- **Support programming models:**
 - Enable portable exploitation of UNIMEM
(MPI, GPI, OmpSs, OpenStream)
- **Evaluate UNIMEM architecture**
 - Latency, bandwidth, memory footprint, CPU usage, ...



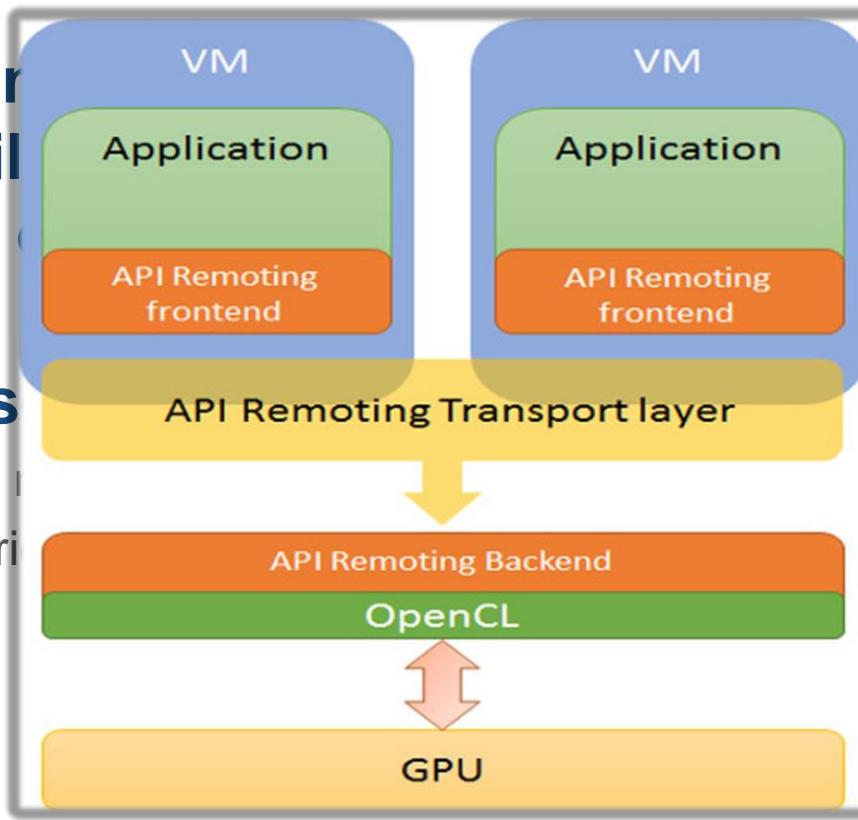
- For deployment ease,
compatibility,
efficiency of resource usage

- **For deployment ease, compatibility, efficiency of resource usage**
- **For an increased flexibility and reliability**
 - **Snapshot:** save now, restore later, maybe elsewhere
 - **Checkpoint:** periodic and incremental, to cope with HW or SW failures

- **For deployment ease, compatibility, efficiency of resource usage**
- **For an increased flexibility and reliability**
 - Snapshot: save now, restore later, maybe elsewhere
 - Checkpoint: periodic and incremental, to cope with HW or SW failures
- **For performance KVM + paravirtualization**
 - API Remoting: use accelerator APIs inside VM
UNIMEM atomics and RDMA, MPI, OpenCL ...

SW stack: Virtualization Layer

- For deployment compatibility and efficiency

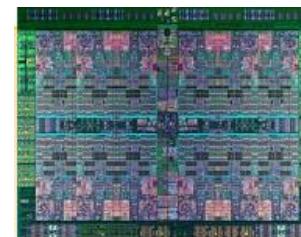
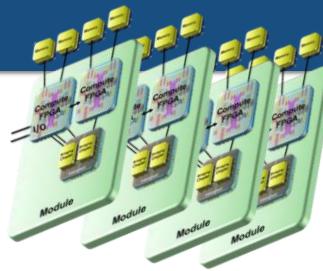


- For an increase in reliability
 - Snapshot: save state of the system
 - Checkpoint: periodically save state for SW failures

- For performance KVM + paravirtualization
 - API Remoting: use accelerator APIs inside VM
UNIMEM atomics and RDMA, MPI, OpenCL ...

Platform comparisons

- **ExaNoDe prototype**
- **EUROSERVER prototype based on Juno ARM**
- **Mainstream servers based on Intel Xeon**
 - Nehalem, Sandybridge, Haswell
- **Emerging server architectures**
 - Intel KNL
 - IBM POWER8
 - Cavium ThunderX, APM X-Gene
- **Low-end platforms**
 - Intel Atom
 - Raspberry Pi

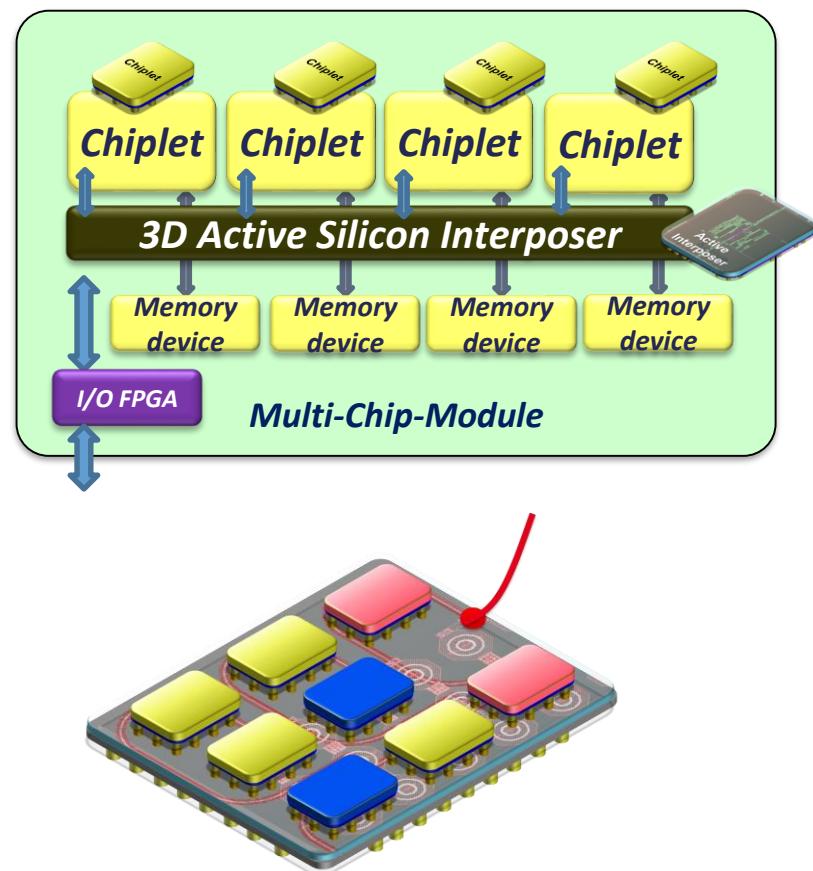


Conclusion

ExaNoDe: Conclusion

■ ExaNoDe concept:

- **Architecture:** Many simple cores instead of few complex ones
- **Integration:** Many simple heterogeneous chiplets instead of few complex System-on-Chip
- **Compute node interconnect:** Active silicon interposer



Thank you!



European Exascale Processor & Memory Node Design

Contact: denis.dutoit@cea.fr (Project coordinator)

k.pouget@virtualopensystems.com (Presenter)