www.evolve-h2020.eu

evolve

Leading the Big Data Revolution

EBDVF 2021

The Evolve Project : A Convergence of Machine Learning and HPC to address Big Data Challenges

Evolve Partners

EBDVF 2021, November 29, 2021

Agenda

- ❑ Introduction to Evolve (DDN/Jean-Thomas Acquaviva, 15’)
- ❑ Hardware Platform (ATOS/Huy-Nam Nguyen, 15’)
- ❑ Software Development (FORTH/Angelos Bilas, 15’)
- ❑ Applications (NEUROCOM/Vassilis Spitadakis, 15’)
- ❑ Proof-of-Concept (VIF/Alexander Stocker, 15’)
- ❑ Q&A (Evolve, 15’)

evolve

www.evolve-h2020.eu

EBDVF 2021

The Evolve Hardware Platform

Evolve Partners

EBDVF 2021, November 29, 2021

Objectives

❑ Motivations and Objectives

- Evolution of System Architecture for integration of Hw Accelerators
  - Flexible, scalable and interoperable exploitation of Hw accelerators
  - Acceleration of Computing and Data Transfer
  - Enabling HPC-BD/AI Convergence
- Validation on a wide spectrum of applications

❑ Main Technical Challenges

- Scalable approach to heterogeneous computing technologies
- Combination/Convergence of acceleration software stacks
- Convergence of HPC, BigData and AI
- Expansion of HPC's scope to the Cloud/Edge Computing
- HPC Features: Availability/Reliability/Accessibility/Security

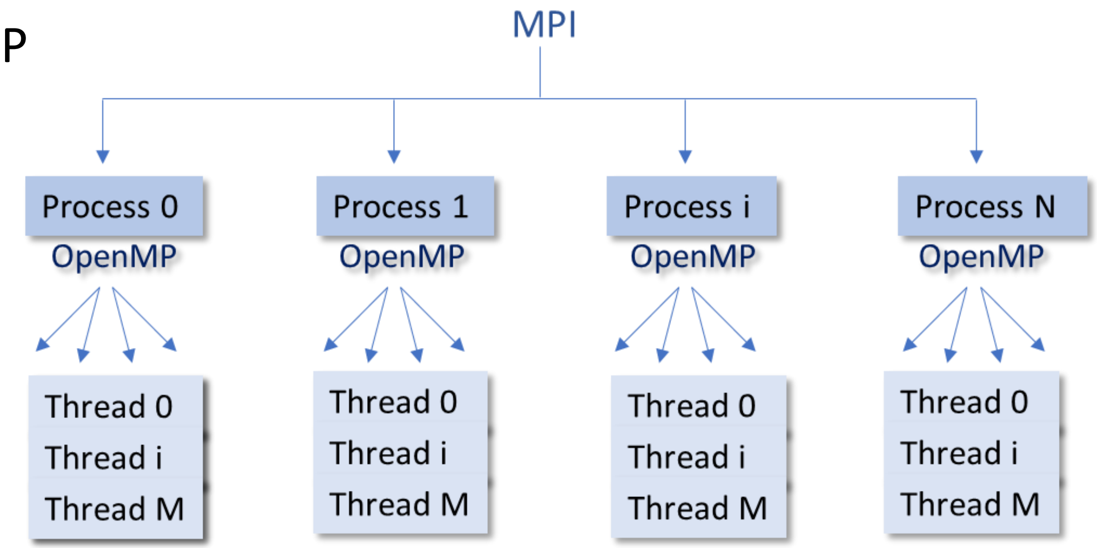
Computing Technologies

❑ **CPUs** (Intel/Broadwell-Haswell-SKX)

- Homogeneous Numa Multi-core
- Performance = #AddMult*#FVect*#Cores*Frequency
- Frequency, #Cores, Cache size, Integrated GPU
- Cache Coherence / Memory Consistency Protocol
- Compute Node size, Node Controller, Numa Factor

❑ **Programming Models** (Hybrid OpenMP+MPI)

- Combinations from pure MPI to MPI+OpenMP
- Adequation with system architecture
- Memory management, Load Balancing
- Performance vs Memory occupation
- Criterions : NUMA factor, Message size, Synchronization (e.g. locks, race)
- MPI-3 shared memory programming
- Accelerator support in OpenMP4.0



MPI

Process 0
OpenMP
Thread 0
Thread i
Thread M

Process 1
OpenMP
Thread 0
Thread i
Thread M

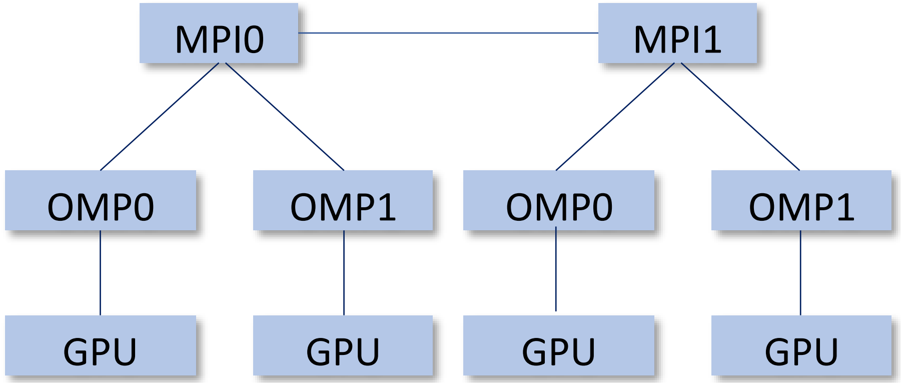
Process i
OpenMP
Thread 0
Thread i
Thread M

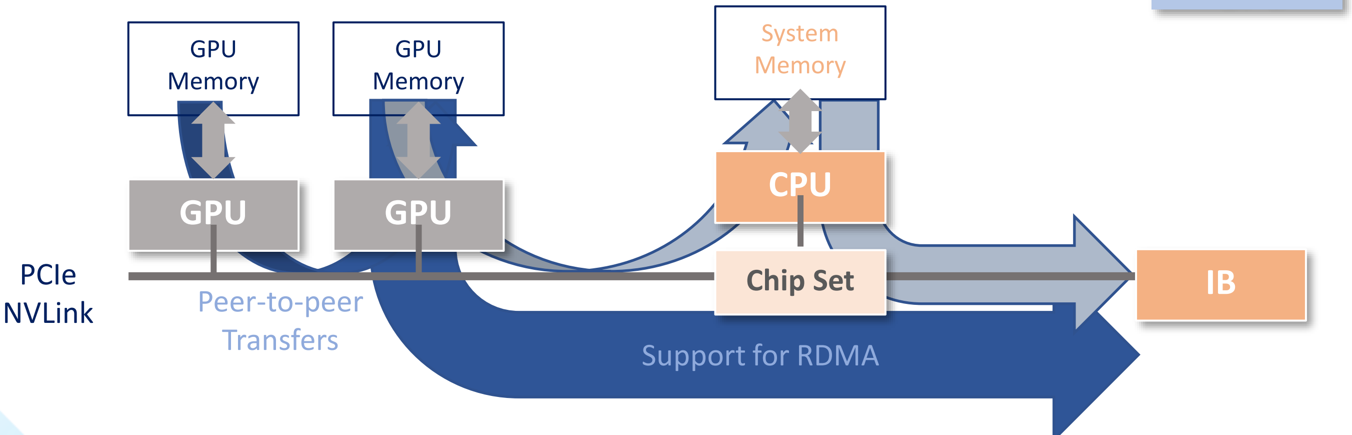
Process N
OpenMP
Thread 0
Thread i
Thread M

Computing Technologies (cont'd)

❑ GPUs (Intel embedded GPU, Nvidia Tesla K20, P40, V100)

- ○ Performance = #AddMult*#SM*#C_{SM}*Frequence
- ○ #Streaming Multiprocessor/#cores, Thread parallelism, Frequency
- ○ Memory Type, Size, Bandwidth and repartition
- ○ Programming Models : TF/MatLab, OpenMP/OpenAcc, CUDA/OpenCL
- ○ Interaction with CPUs
  - Hybrid MPI + OpenMP/CUDA-OpenCL
  - GPU scalability
    - ○ Mapping MPI tasks to GPUs
    - ○ Scaling code on multiple GPUs
    - ○ Monitoring GPUs in a heterogeneous cluster
  - Multiple CUDA-aware MPI implementation



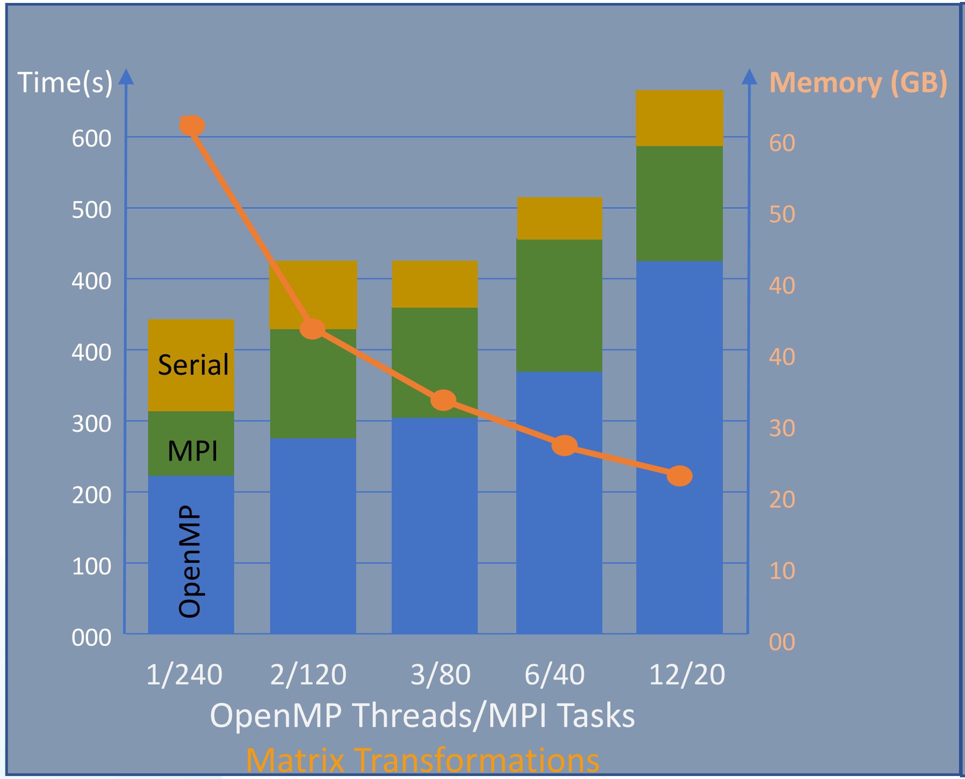


Computing Technologies (cont’d)

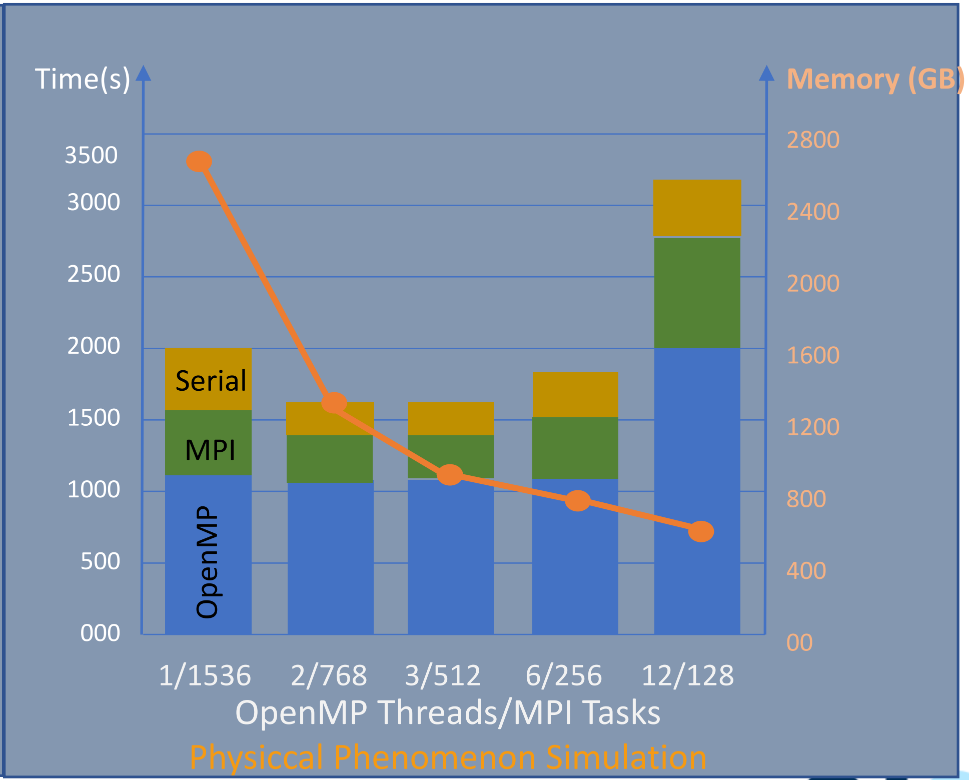
❑ Hybrid Programming Models

- ○ Combination of OpenMP + MPI
- ○ CPU vs OpenAcc vs CUDA
- ○ CUDA-aware MPI

FFT Time (s)	CPU	GPU OpenACC	GPU CUDA
Exec	53.426	3.767	1.08
Transfer	0	3.342	4.719
Total	53.426	7.109	5.799
SpeedUp	1	7.5	9.2



Matrix Transformations

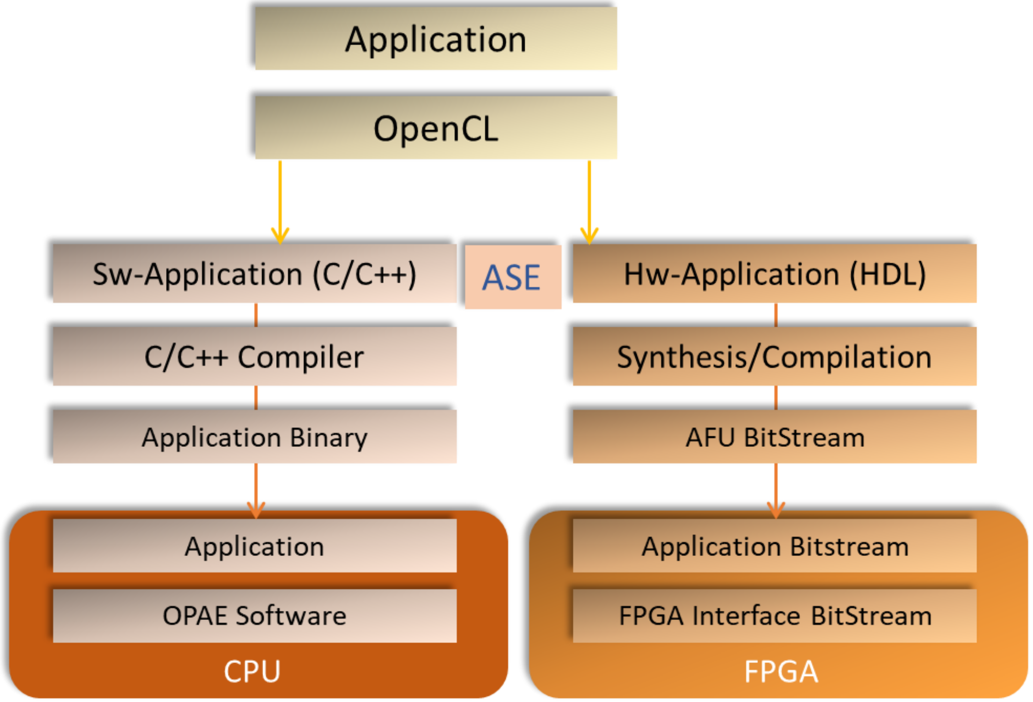


Physiccal Phenomenon Simulation

Computing Technologies (cont’d)

❑ FPGA (Intel/Arria10-Stratix10)

- #CLBs/LEs/LUTs, RAMs, DSPs, Interconnect, SERDES/Transceivers
- FPG-based Programming Models : HDL (Vhdl/SVerilog), OpenCL
- Discrete vs Integrated implementation
- Challenge/Evolution trend : Clocking, Multi-FPGA



Application

OpenCL

Sw-Application (C/C++)

ASE

Hw-Application (HDL)

C/C++ Compiler

Synthesis/Compilation

Application Binary

AFU BitStream

Application

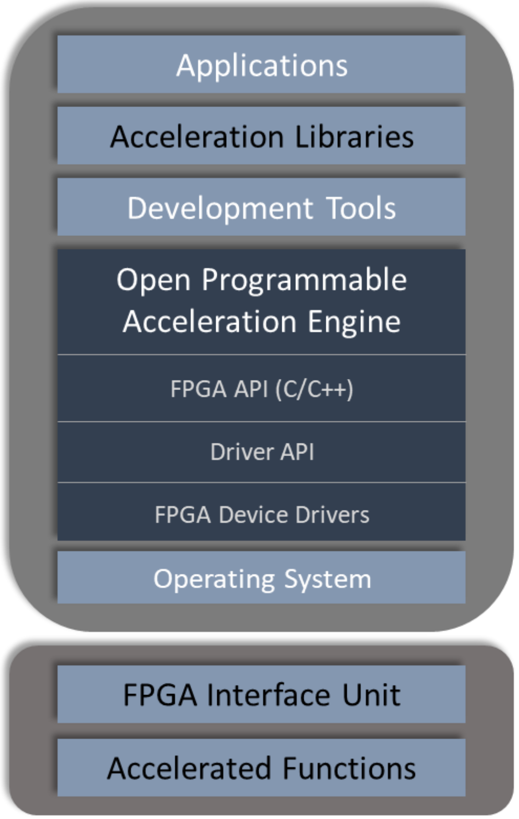
OPAE Software

CPU

Application Bitstream

FPGA Interface BitStream

FPGA



Applications

Acceleration Libraries

Development Tools

Open Programmable Acceleration Engine

FPGA API (C/C++)

Driver API

FPGA Device Drivers

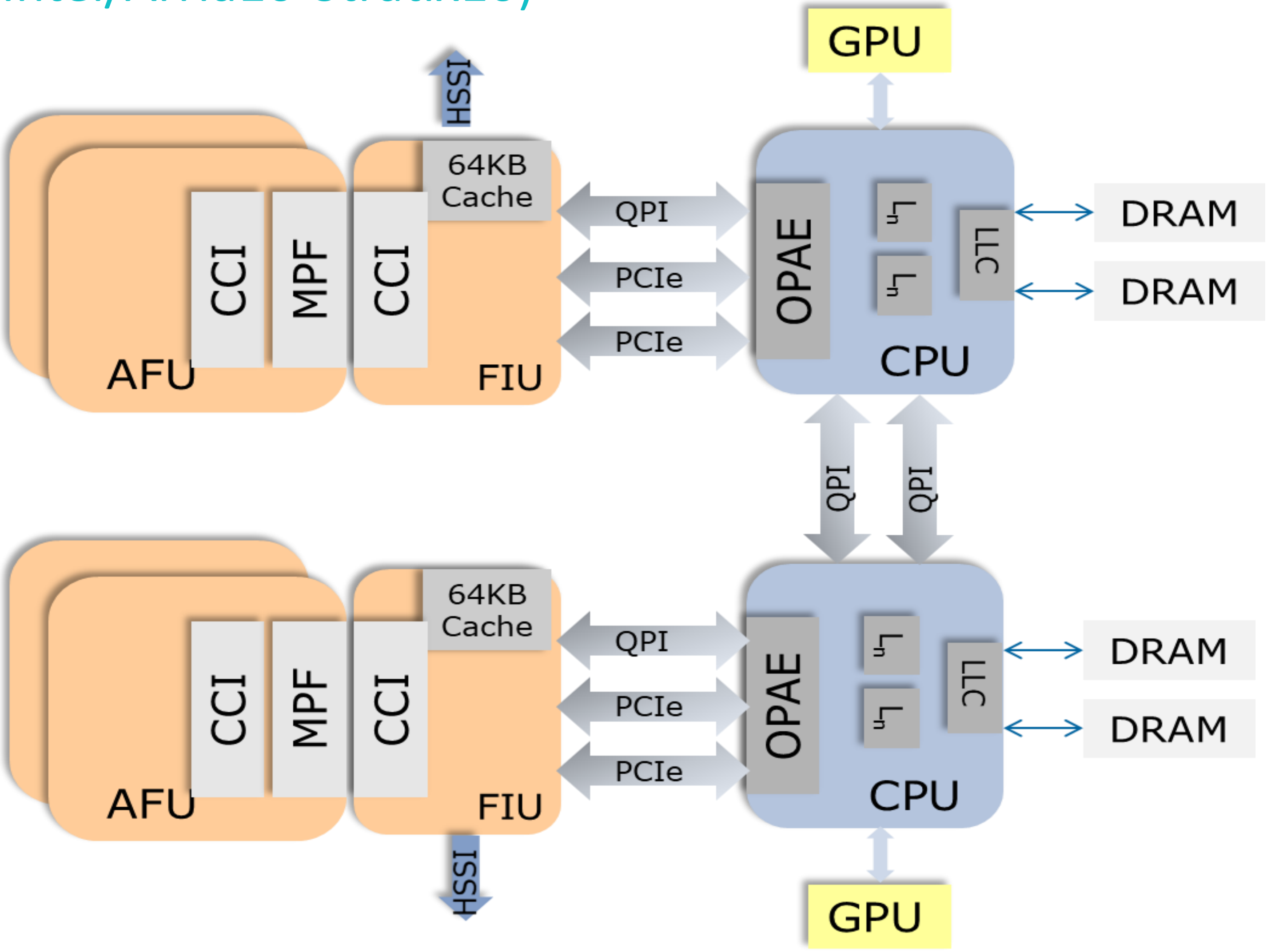
Operating System

FPGA Interface Unit

Accelerated Functions

Computing Technologies (cont'd)

❑ **FPGA** (Intel/Arria10-Stratix10)



Computing Technologies (cont'd)

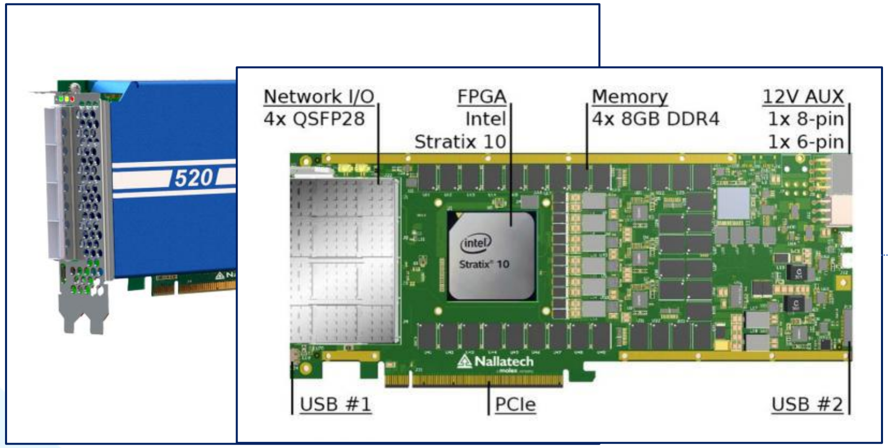
❑ Acceleration of Computation

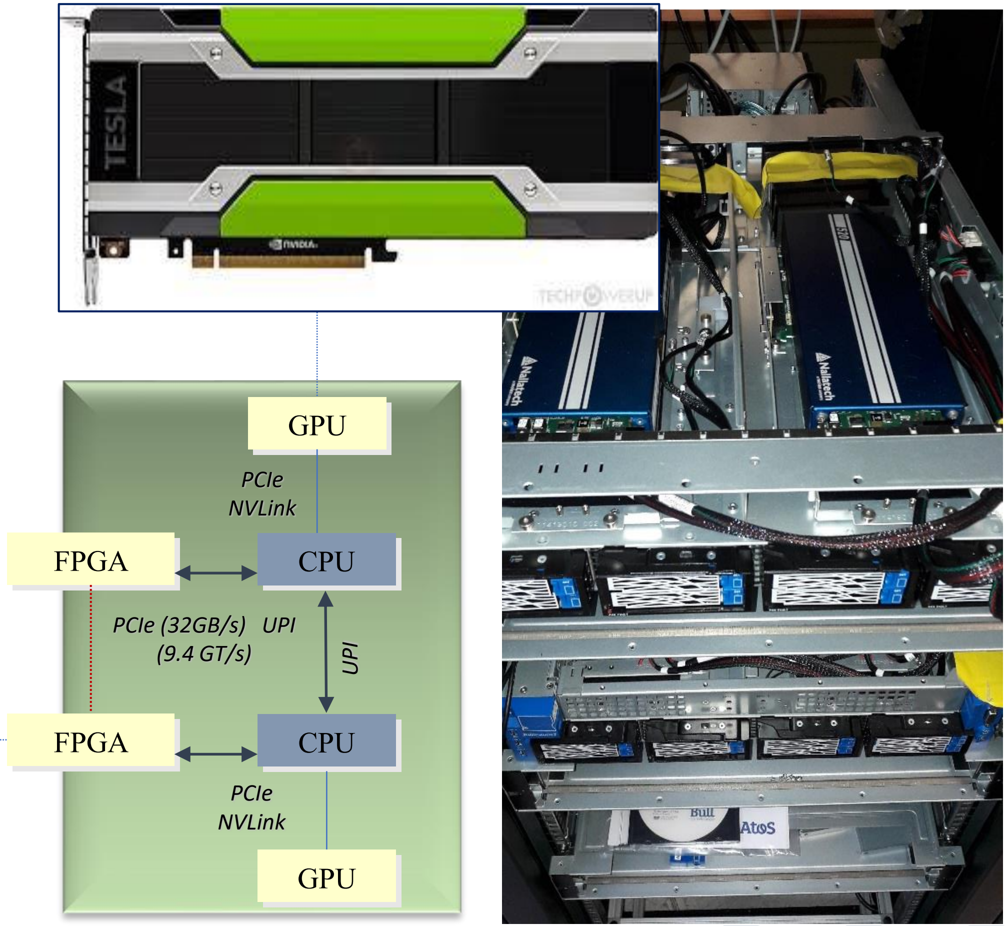
Nvidia/P40 : 3840 Cores, 24 GB GDDR5, PCIe 3.0

Nvidia/V100 : 5120 Cores, 32 GB HBM2, Pcie Gen3

Nallatech/520N

- Intel/Stratix 10 GX 2800
- 4x QSFP28s for 400Gbps
- BittWare-optimized OpenCL BSP for Intel SDK





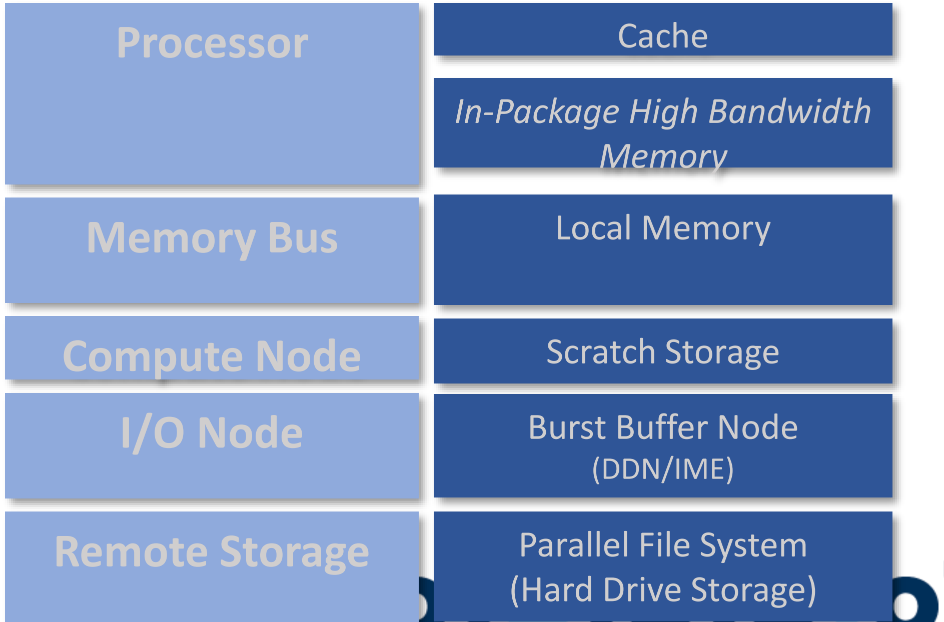
Memory/Storage Hierarchy

❑ Memory Sizing and Distribution

- Memory requirements of production HPC Applications
  - HPL/HPCG benchmarks : **2 GB (resp. 1GB) / x86 (resp. PowerPC) core**
  - UEABS benchmarks : 10^2 MB/core
  - Increase with #cores : 10^6cores => 10GB/core
- Reduction of Application to/from Network latency

○ Storage

- DDN's IME® scale-out, software-defined, flash storage platform
  - Streamlines the data path for application I/O
  - Realizes flash-cache economics with the storage
  - 8 IME120 systems. For a total of 48TB flash available
- Global vs Local storage
  - 60 disks (1.8Gb each): 98TB
  - 24 disks (1.8Gb each): 47TB
  - Local storage : 1TB/Compute node
- Revisited/New Features
  - System/Users CheckPointing/Restart Procedures
  - Local data Compression



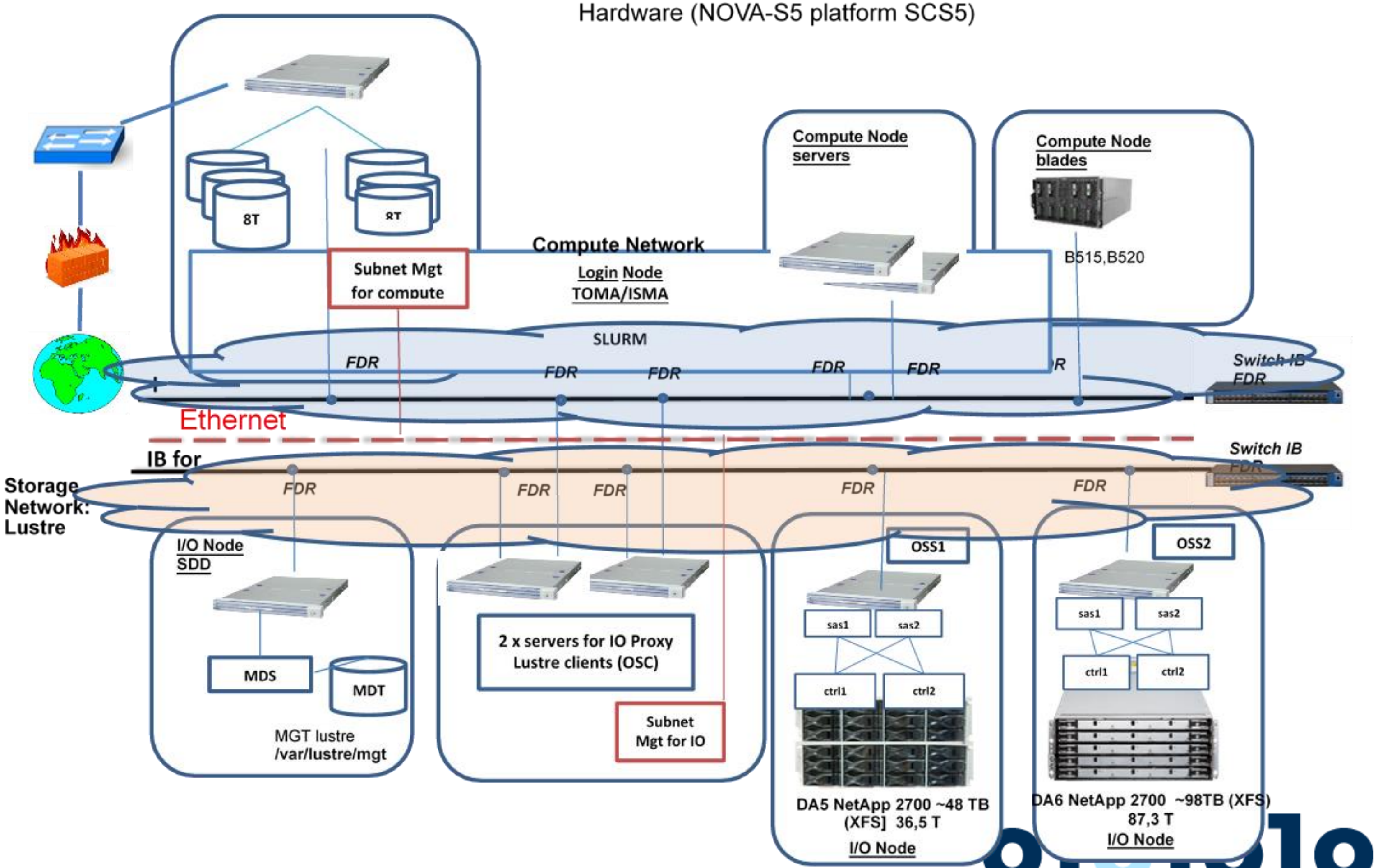
Processor	Cache *In-Package High Bandwidth Memory*
Memory Bus	Local Memory
Compute Node	Scratch Storage
I/O Node	Burst Buffer Node (DDN/IME)
Remote Storage	Parallel File System (Hard Drive Storage)

Interconnect Fabric

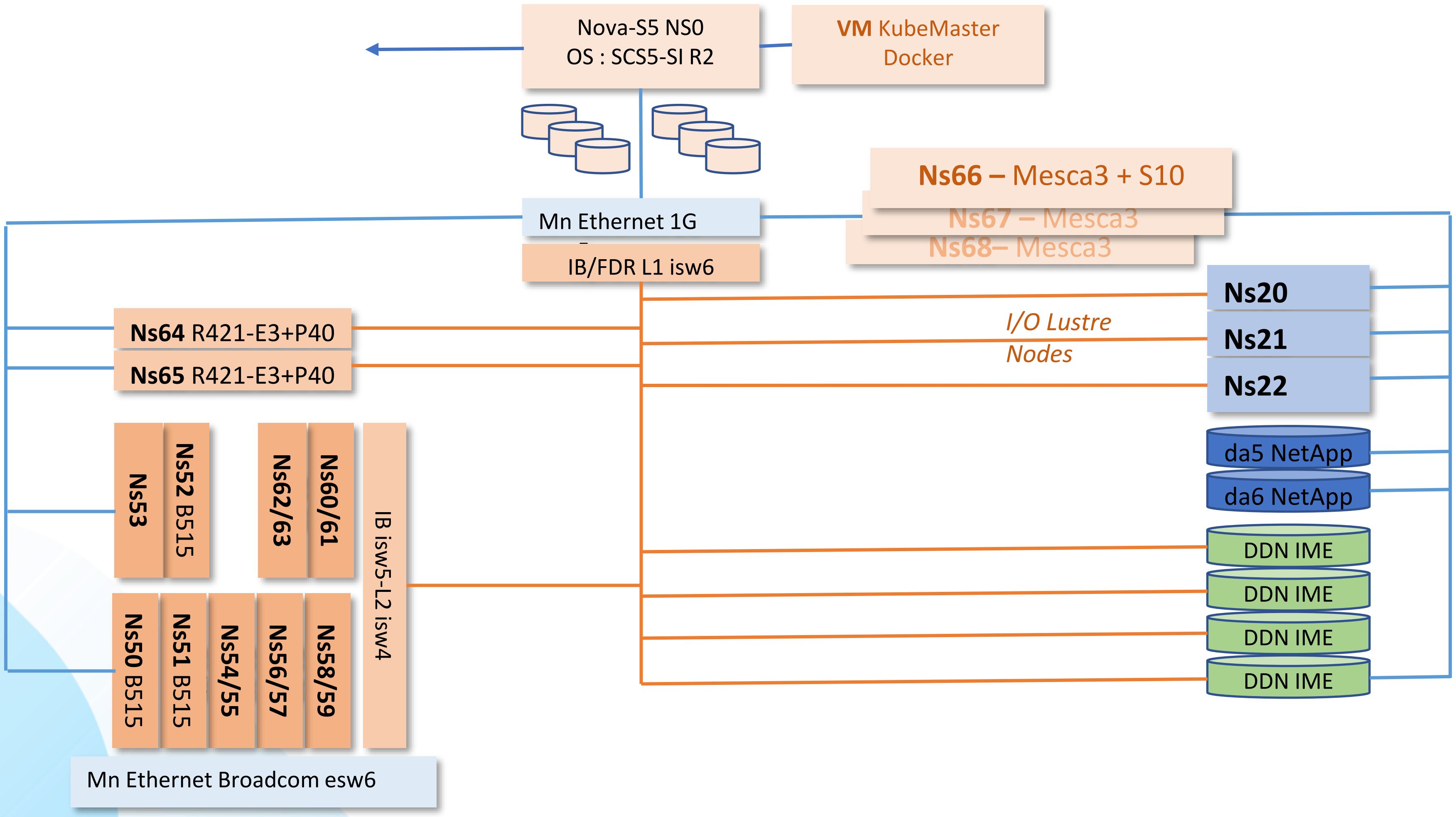
❑ Interconnect Fabric

- IB/FDR (ISR 9024D and ISR 9024D-M) @56 Gb/s
- Cisco catalyst 3560G 44 ports – private Ethernet
- Fat-Tree Topology

Hardware (NOVA-S5 platform SCS5)



The Hw Evolve Platform



Nova-S5 NS0
OS : SCS5-SI R2

VM KubeMaster
Docker

Ns66 – Mesca3 + S10

Ns67 – Mesca3

Ns68– Mesca3

Mn Ethernet 1G

IB/FDR L1 isw6

Ns64 R421-E3+P40

Ns65 R421-E3+P40

I/O Lustre
Nodes

Ns20

Ns21

Ns22

da5 NetApp

da6 NetApp

DDN IME

DDN IME

DDN IME

DDN IME

Ns53

Ns52 B515

Ns62/63

Ns60/61

IB isw5-L2 isw4

Ns50 B515

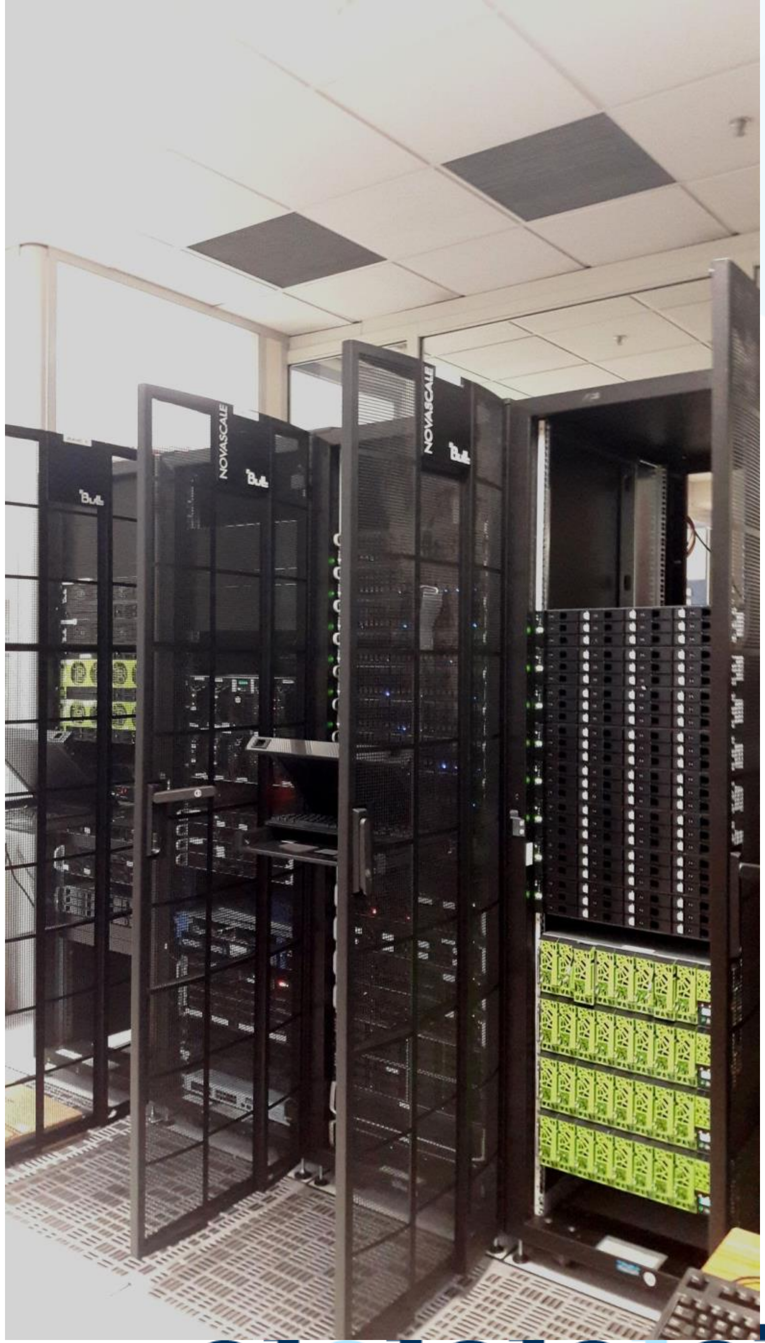
Ns51 B515

Ns54/55

Ns56/57

Ns58/59

Mn Ethernet Broadcom esw6



Conclusions

❑ **Investigation of some aspects of the HPC Evolution**

- ❑ General-purpose Computing Acceleration (GPU, FPGA)
- ❑ Restructuration of the Memory/Storage Hierarchy
- ❑ Co-Design

❑ **Future Development** (in the context of other projects)

- ○ AI-Acceleration
- ○ In-situ Computation
- ○ Co-Design
- ○ Evolution towards the HPC-Cloud/Edge Continuum

Consortium

DDN, BULL, IBM, FORTH, ONAPP,

ICCS, MEMO, WLT, LOBA, TAS, SPH,

CYB, NEURO, MEMEX, TIEMME, VIF,

AVL, BMW, KOOLA

www.evolve-h2020.eu

info@evolve-h2020.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825061