COSA project update: new low power architectures and the Intel KNL

Daniele Cesini – INFN-CNAF
(On behalf of the COSA collaboration)

http://www.cosa-project.it

INFN
Istituto Nazionale
di Fisica Nucleare

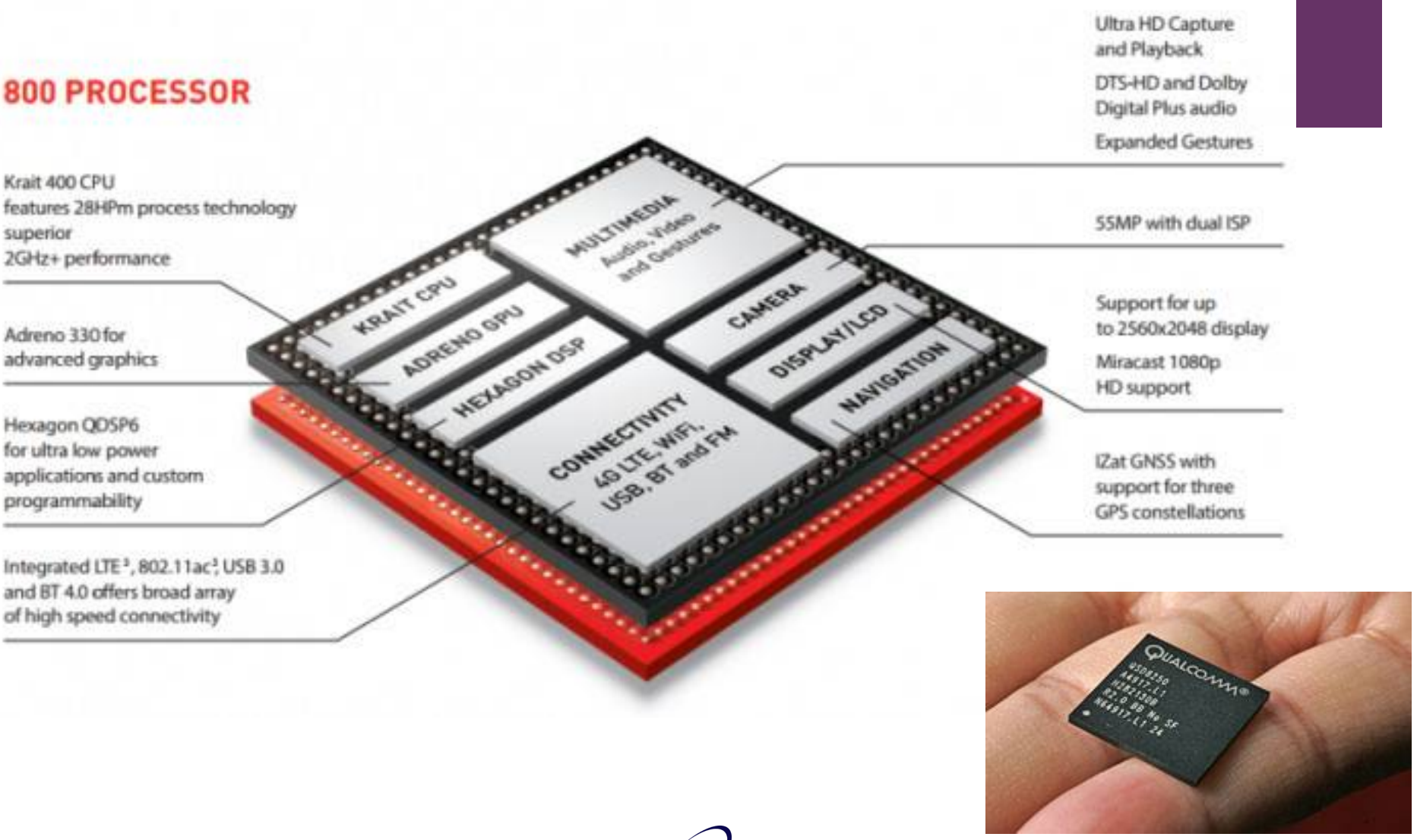
INFN COSA project

- COSA: Computing On SOC Architecture
- Duration: 3 years from January 2015
- Departments: 7 INFN
  - CNAF, PI, PD, ROMA1, FE, PR, LNL
- BUDGET :51.5 kEuro Year1, 42kEuro Year2
  - Funded by INFN CSN5

Objectives

- Acquire know-how
  - Porting and benchmarking of low power/low cost System on Chip
  - Operations of Linux system on SoCs
  - Benchmarking hybrid architectures
- Unification of INFN HW testing activities
  - Continuation of the COKA project
    - Computing on Knights Architecture
    - Porting on traditional accelerator (GPU/MIC)
  - Continuation of the HEPMARK projects
    - X86 benchmarking
- Study of custom low latency interconnection built with ARM+FPGA devices
- Prepare H2020 proposals on LowPower computing calls

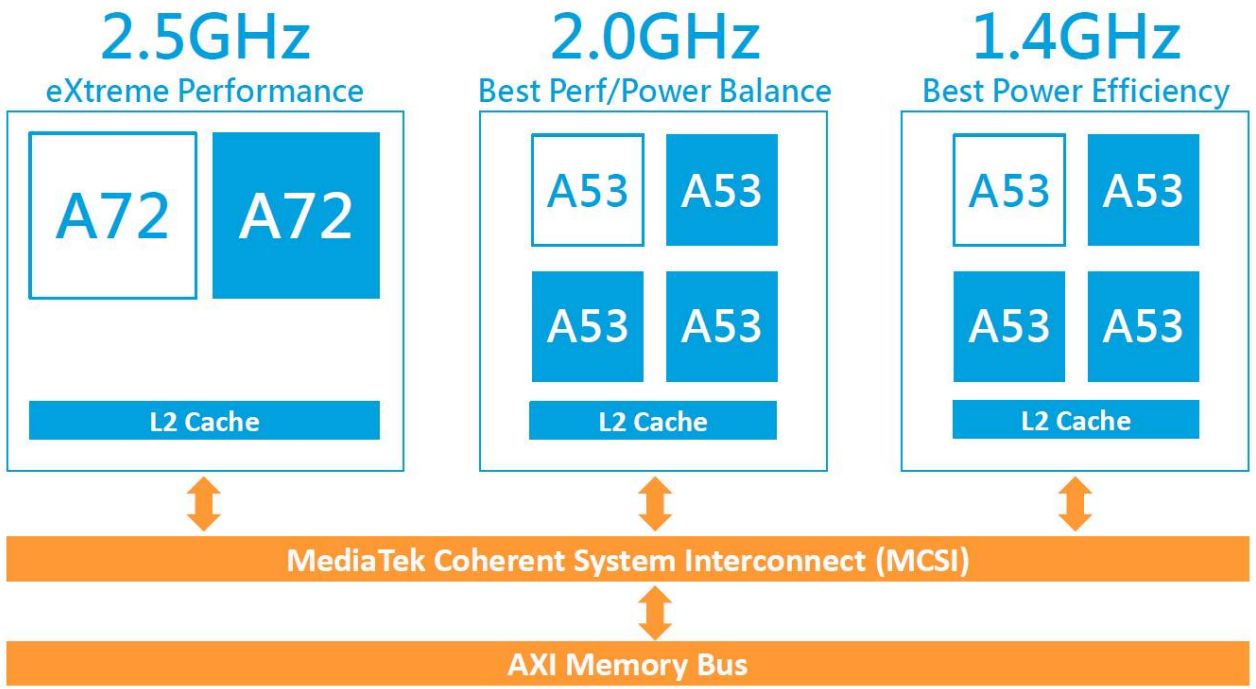
Low-Power System on Chip (SoCs)



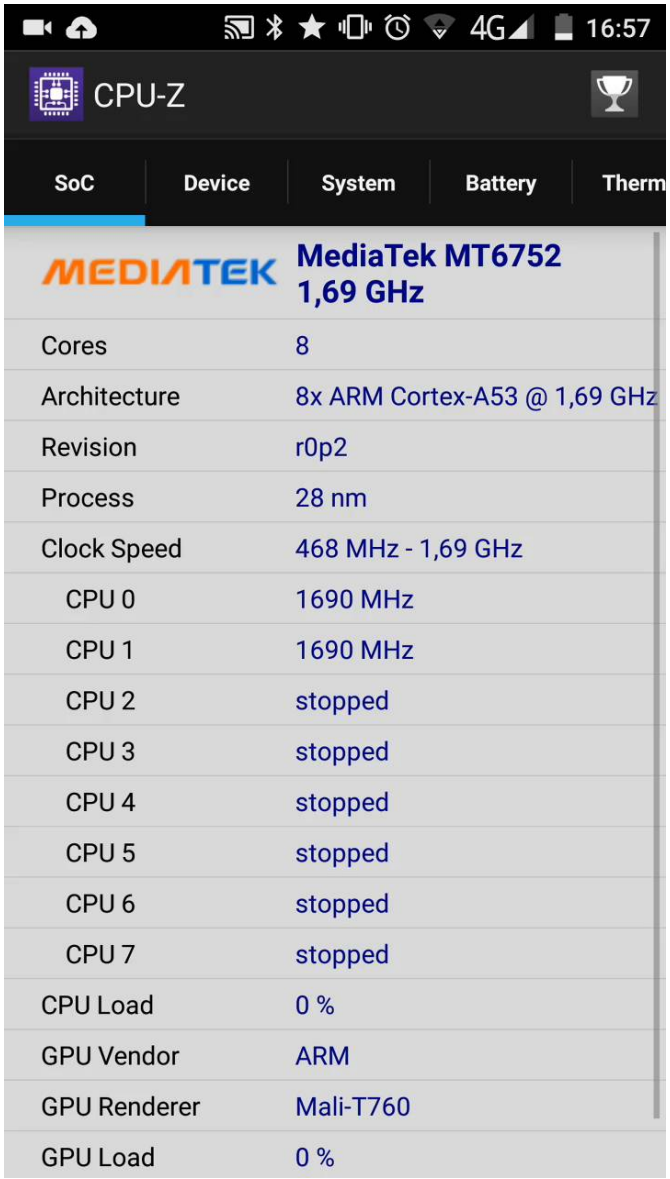
SoC Multicore Madness

big.Medium.LITTLE

Deca/10-Core CPU Architecture



Modern SoCs

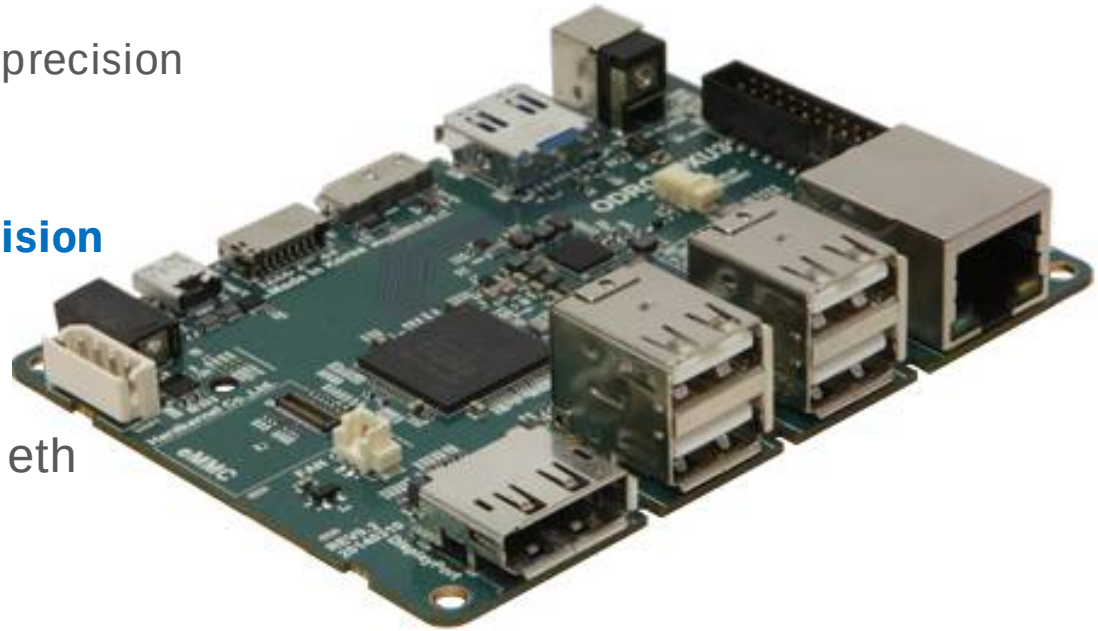

16:57
CPU-Z
SoC
Device
System
Battery
Therm
MEDIATEK
MediaTek MT6752
1,69 GHz
Cores 8
Architecture 8x ARM Cortex-A53 @ 1,69 GHz
Revision r0p2
Process 28 nm
Clock Speed 468 MHz - 1,69 GHz
CPU 0 1690 MHz
CPU 1 1690 MHz
CPU 2 stopped
CPU 3 stopped
CPU 4 stopped
CPU 5 stopped
CPU 6 stopped
CPU 7 stopped
CPU Load 0 %
GPU Vendor ARM
GPU Renderer Mali-T760
GPU Load 0 %

Ok, but then….an iPhone cluster?

- NO, we are not thinking to build an iPhone cluster
- We want to use these processors in a standard computing center configuration
  - Rack mounted
  - Linux powered
  - Running scientific application mostly in a batch environment
- ..... Use development board...

ODROID-XU3

- Powered by ARM® big.LITTLE™ technology, with a **Heterogeneous Multi-Processing (HMP)** solution
  - 4 core ARM A15 + 4 cores ARM A7
- Exynos 5422 by Samsung
  - ~ 20 GFLOPS peak (32bit) single precision
- **Mali-T628 MP6 GPU**
  - **~ 110 GFLOPS peak single precision**
- 2 GB RAM
- 2xUSB3.0, 2xUSB2.0, 1x1000Gbs eth
- Ubuntu 14.4
- HDMI 1.4 port
- 64 GB flash storage



Power consumption max ~ 15 W

Costs 150 euro!

Other nice boards...

...during the old good times of ARM 32bit

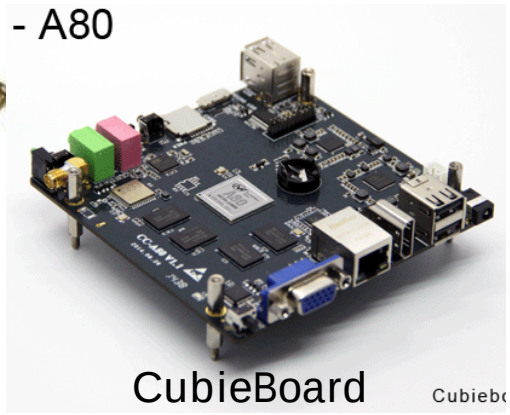
WandBoard

Rock2Board

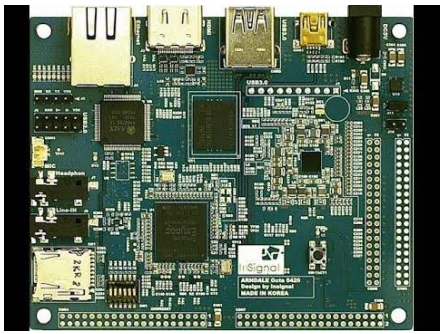
PandaBoard

DragonBoard

SabreBoard



CubieBoard



Arndale OCTA Board

Texas Instruments EVMK2H

http://elinux.org/Development_Platforms

- ...and counting...

NVIDIA JETSON TK1

- First **ARM+CUDA programmable SoC based** Linux development board
- 4 cores ARM A15 CPU
- 192 cores NVIDIA GPU
  ➔ 300 GFLOPS (peak sp)
- ~ **21 GFLOPS/W (sp)**
- ... for less than 200 Euros
- 32bit
  - 64bit version announced

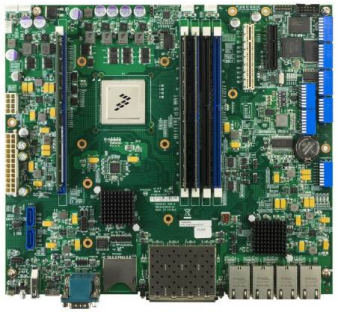
ARMv8 64bit boards...

...harder times

Server Grade platform

ARM Juno Board
r1: 2xA57 + 4xA53
r2: 2xA72 + 4xA53
DRAM: 8 Gbytes
4 PCI-E (Gen.2, 4x)
r1: 5000$
r2:7000$

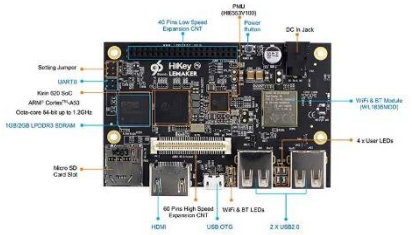
Gigabyte MP30-AR0
AppliedMicro X-Gene1 8core
DRAM:max128GB
2 x 10GbE SFP+
2 x 1GbE LAN ports
2 x PCI-Express slots (Gen.3, 8x)
700eu



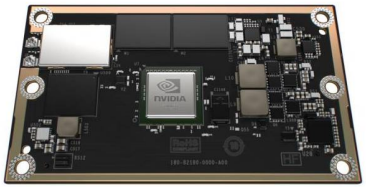
FreescaleQorIQ LS2085A
8 x Cortex-A57 cores
DRAM:max 16GB
PCI Gen3 (x8)
4 x 10 GbE SFP
4 x 10 GbE RJ45
About 3000$

AMD Opteron A1100
16GB RAM
2x10Gbs
Cost 2000$

Embedded platform

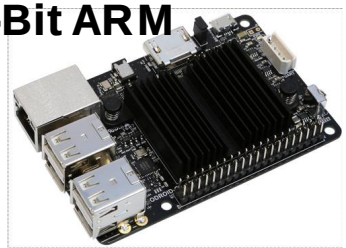


HiKey 96boards
1/2GB LPDDR3 SDRAM
8 x Cortex-A53 cores
Cost: $100 (2GB)

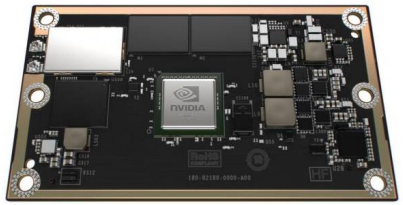


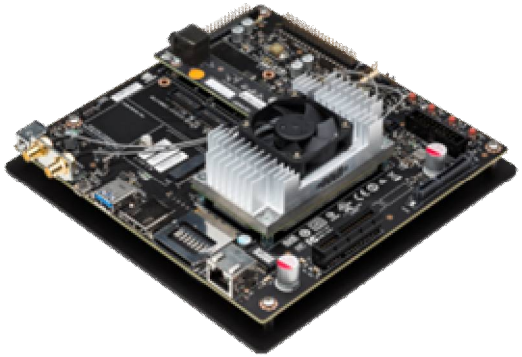
NVIDIA Jetson TX1
4x A57 2 MB di L2; 4x A53 512 KB di L2
256 core di GPU NVIDIA Maxwell
600$

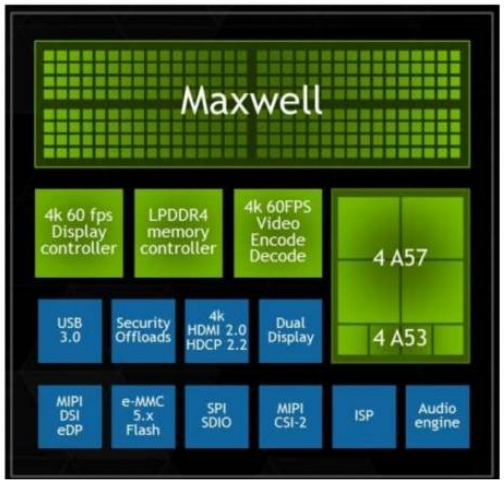
ODROID-C2 64-Bit ARM
4xA53@2GHz
Mali™-450 GPU
2GB RAM
1Gbs ETH



NVIDIA Jetson Tegra X1







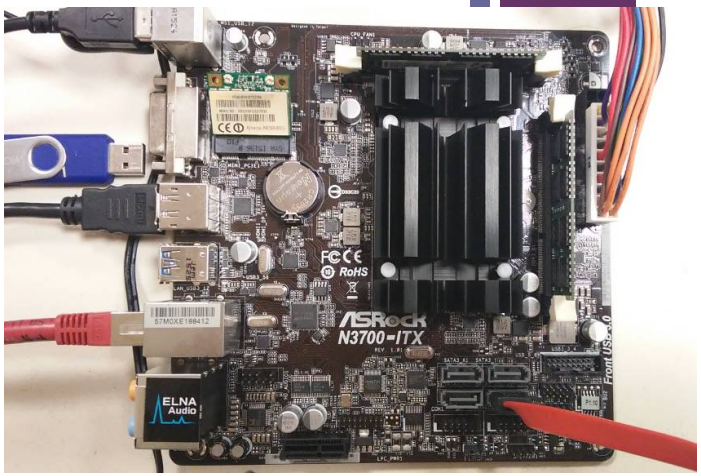
- ARM+CUDA programmable SoC based Linux development board
- 4 cores 64 bit-ARM A57 cpu
  - 1.73GHz
  - K1 was 2.32 GHz
- 256 cores NVIDIA GPU
  ➔ 1024 GFLOPS (peak hp)
- TDP: 6.5 W – 15 W
- ~ **100 GFLOPS/W (hp)**
- ... for 600 Euros

That's not fair…

≠

Low power from Intel

Product Name	Intel® Pentium® Processor N3700 (2M Cache, up to 2.40 GHz)	Intel® Pentium® Processor J3710 (2M Cache, up to 2.64 GHz)	Intel® Pentium® Processor N3710 (2M Cache, up to 2.56 GHz)
Code Name	Braswell	Braswell	Braswell
Processor Number	N3700	J3710	N3710
Cache	2 MB L2 Cache	2 MB L2 Cache	2 MB L2 Cache
Instruction Set	64-bit	64-bit	64-bit
Embedded Options Available	No	No	Yes
Lithography	14 nm	14 nm	14 nm
Recommended Customer Price	TRAY: $161.00	N/A	N/A
Datasheet	Link		Link
Conflict Free	Yes	Yes	Yes
Additional Information URL	Link		Link
Performance			
# of Cores	4	4	4
# of Threads	4	4	4
Processor Base Frequency	1.6 GHz	1.6 GHz	1.6 GHz
Burst Frequency	2.4 GHz	2.64 GHz	2.56 GHz
TDP	6 W	6.5 W	6 W
Scenario Design Power (SDP)	4 W		4 W
Memory Specifications			
Max Memory Size (dependent on memory type)	8 GB	8 GB	8 GB
Memory Types	DDR3L-1600	DDR3L-1600	DDR3L-1600
Max # of Memory Channels	2	2	2
ECC Memory Supported ‡	No	No	No
Graphics Specifications			
Processor Graphics ‡	Intel® HD Graphics	Intel® HD Graphics 405	Intel® HD Graphics 405
Graphics Base Frequency	400 MHz	400 MHz	400 MHz
Graphics Burst Frequency	700 MHz		700 MHz



INTEL N3700

- 4 cores / Intel HD Graphics
- 6W
- Airmont microarchitecture (64 bit, No AVX/AVX2)
- 16GB
- SATA ports / PCIe 2.0 1x
- Fanless
- 100 euro !!!

Low power from Intel - 2

Product Name	Intel® Core™ m5-6Y54 Processor (4M Cache, up to 2.70 GHz)	Intel® Core™ i7-6500U Processor (4M Cache, up to 3.10 GHz)	Intel® Xeon® Processor D-1540 (12M Cache, 2.00 GHz)	Intel® Atom™ Processor C2750 (4M Cache, 2.40 GHz)
Code Name	Skylake	Skylake	Broadwell	Avoton
Essentials				
Status	Launched	Launched	Launched	Launched
Launch Date	Q3'15	Q3'15	Q1'15	Q3'13
Processor Number	M5-6Y54	i7-6500U	D-1540	C2750
Cache	4 MB Intel® Smart Cache	4 MB Intel® Smart Cache	12 MB	4 MB
Instruction Set	64-bit	64-bit	64-bit	64-bit
Instruction Set Extensions	SSE4.1/4.2, AVX 2.0	SSE4.1/4.2, AVX 2.0	AVX 2.0	
Embedded Options Available	No	No	No	No
Lithography	14 nm	14 nm	14 nm	22 nm
Recommended Customer Price	TRAY: $281.00	TRAY: $393.00	TRAY: $581.00	TRAY: $171.00
Datasheet	Link	Link	Link	Link
Product Brief	Link	Link	Link	
Scalability			1S Only	
Performance				
# of Cores	2	2	8	8
# of Threads	4	4	16	8
Processor Base Frequency	1.1 GHz	2.5 GHz	2 GHz	2.4 GHz
Max Turbo Frequency	2.7 GHz	3.1 GHz	2.6 GHz	2.6 GHz
TDP	4.5 W	15 W	45 W	20 W

CORE M i7 Mobile XEON D AVOTON

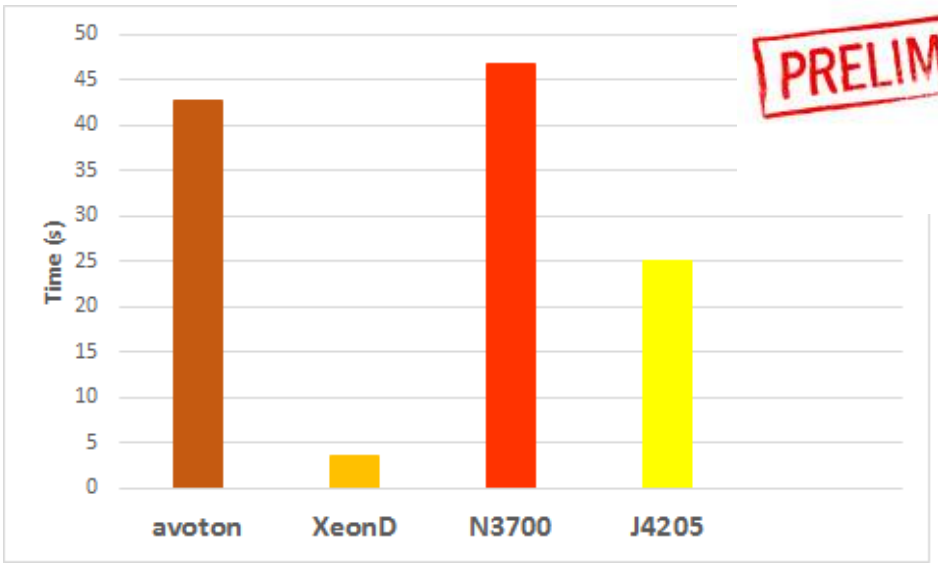
Intel® HD Graphics 515/520

INTEL Apollo Lake J4205

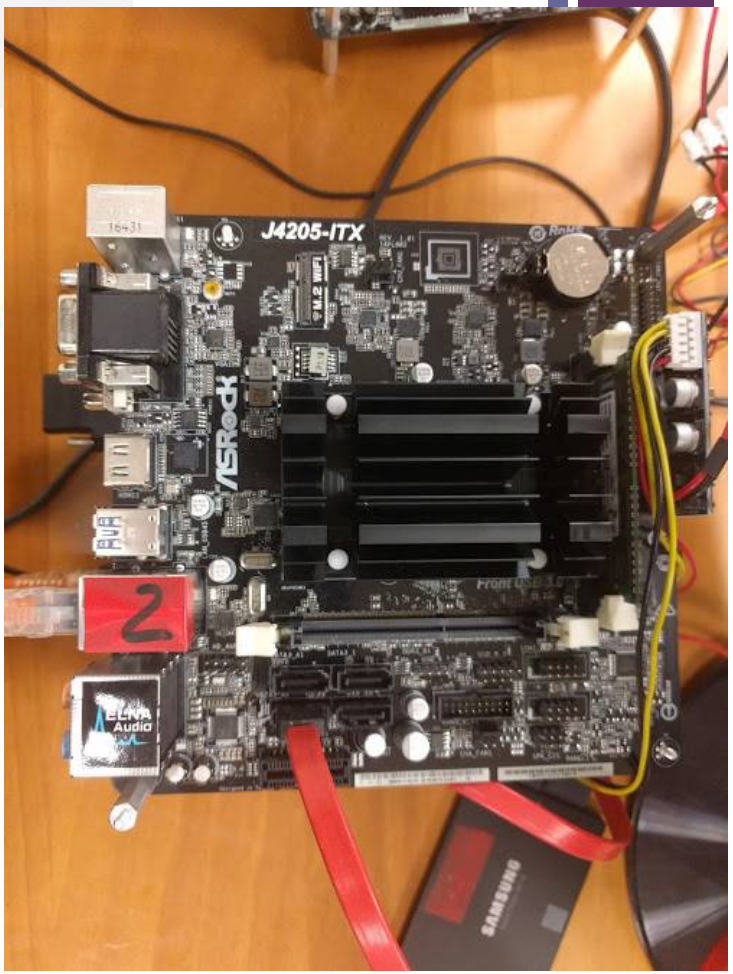
Intel® Pentium® Processor J4205

2M Cache, up to 2.6 GHz

- 4 core, 1.5 GHz base frequency, TDP 10 W



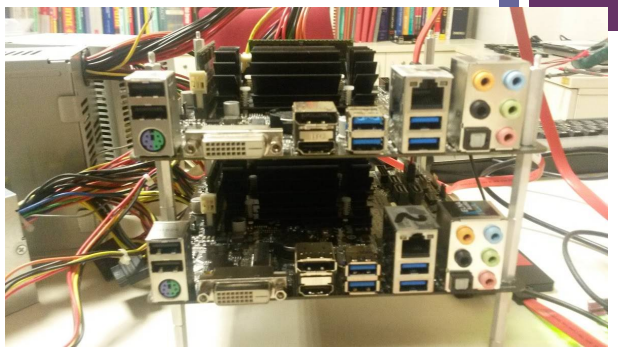
PRELIMINARY



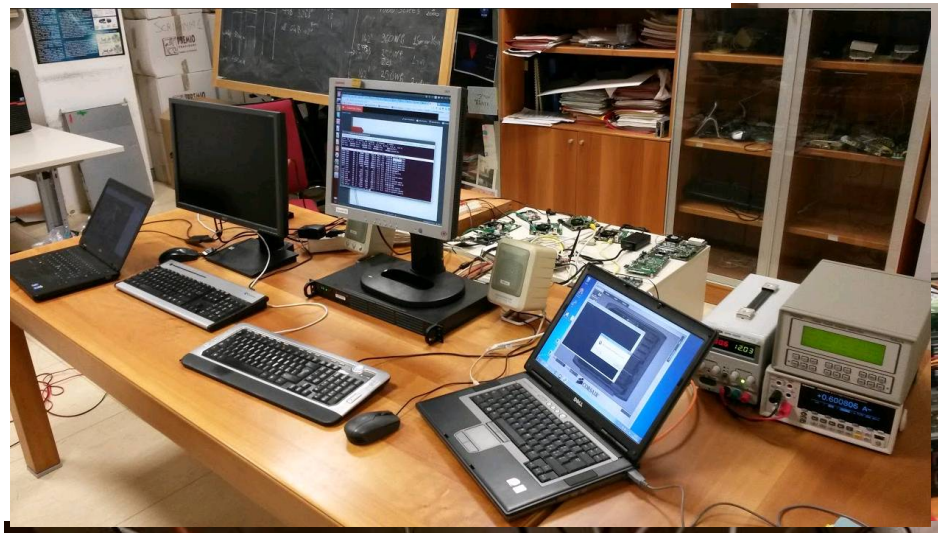
Low Power COSA Clusters@CNAF

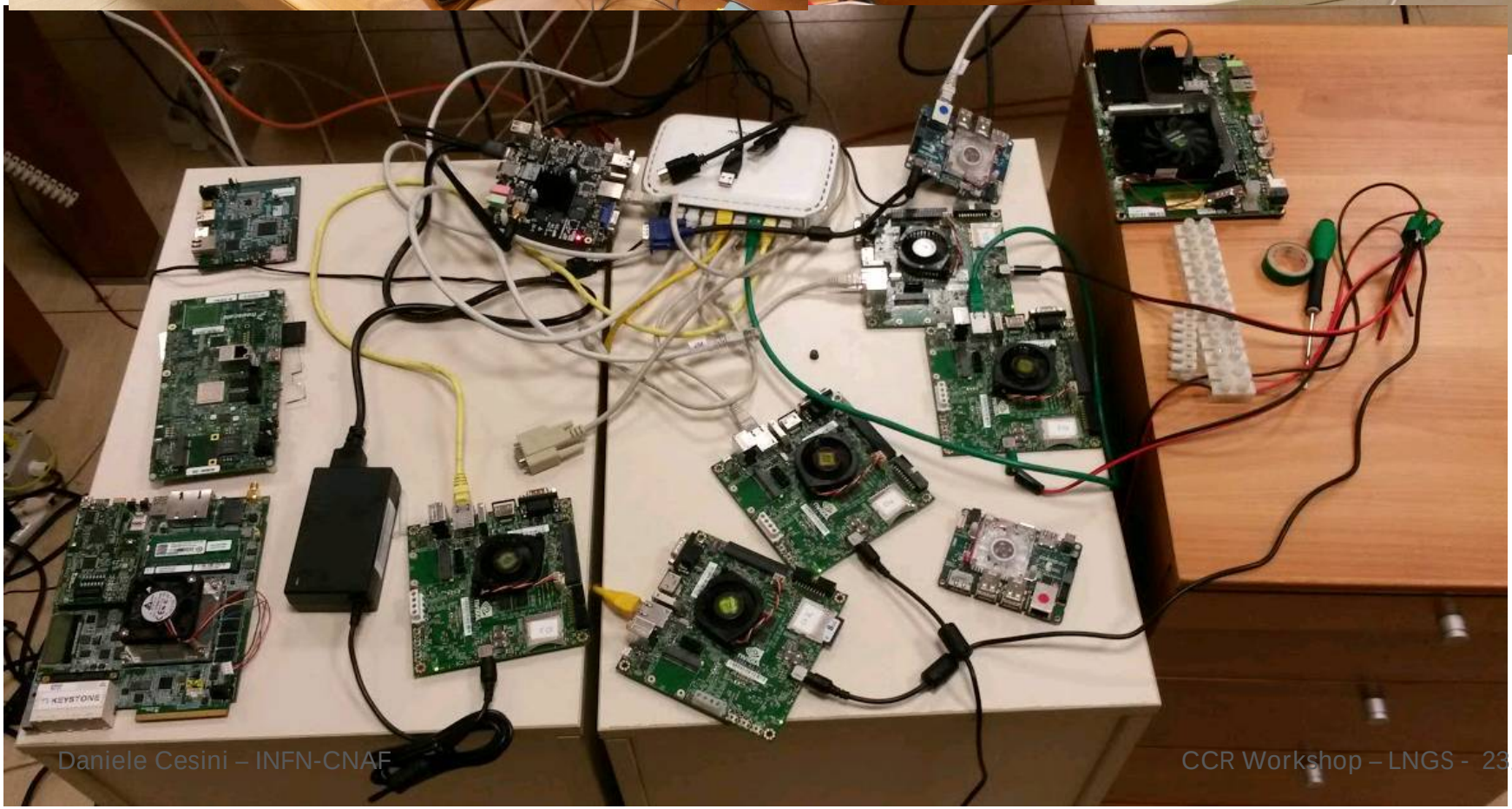
16xARMv7
8xARMv8

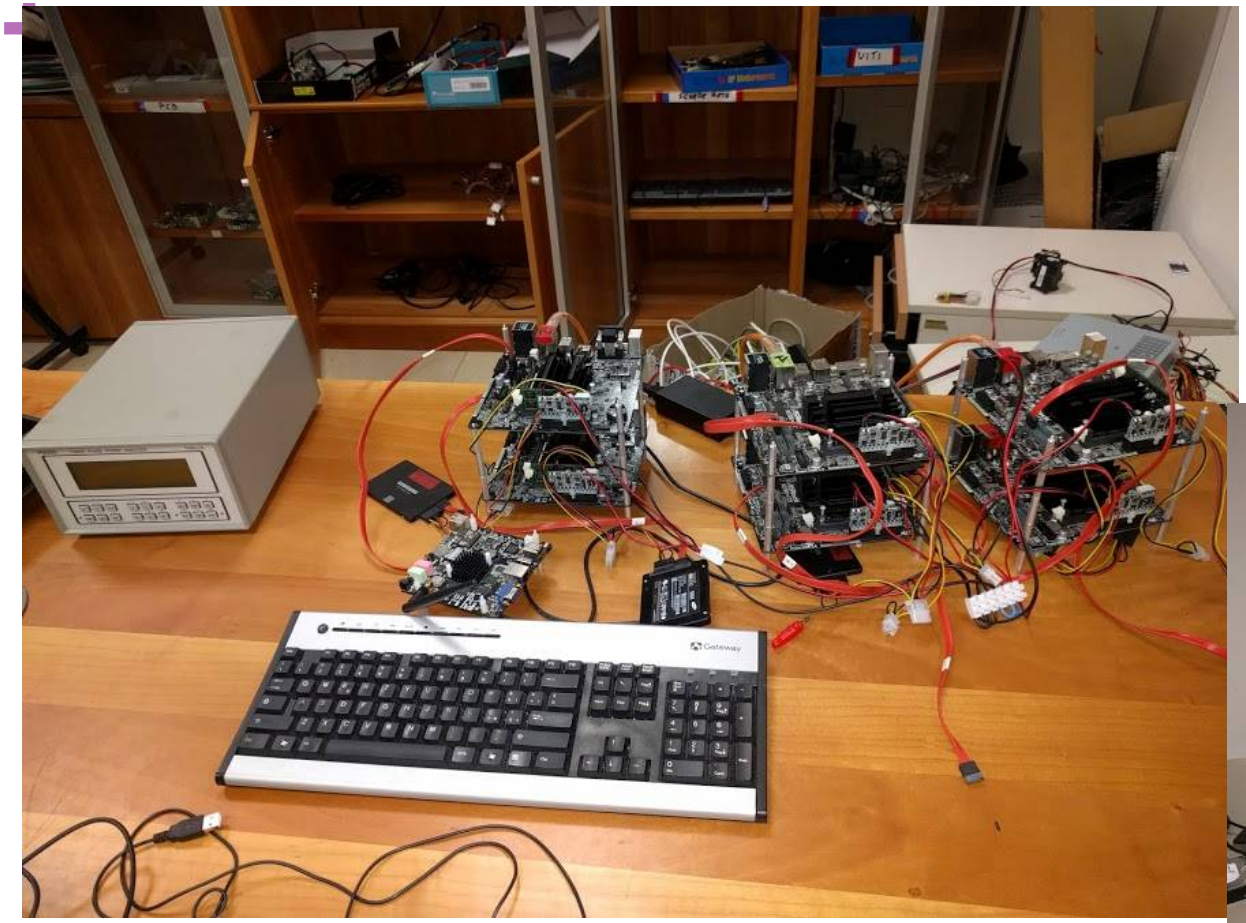
4xINTEL AVOTON C-2750
4xINTEL XEOND-1540

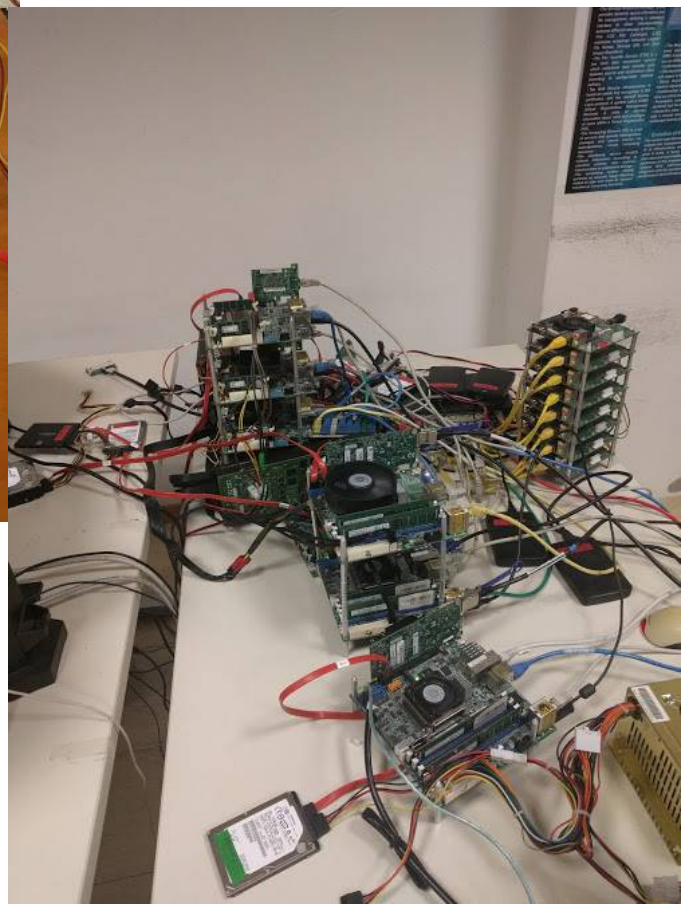


2xINTEL N3700
4xINTEL N3710
2XINTEL J4205

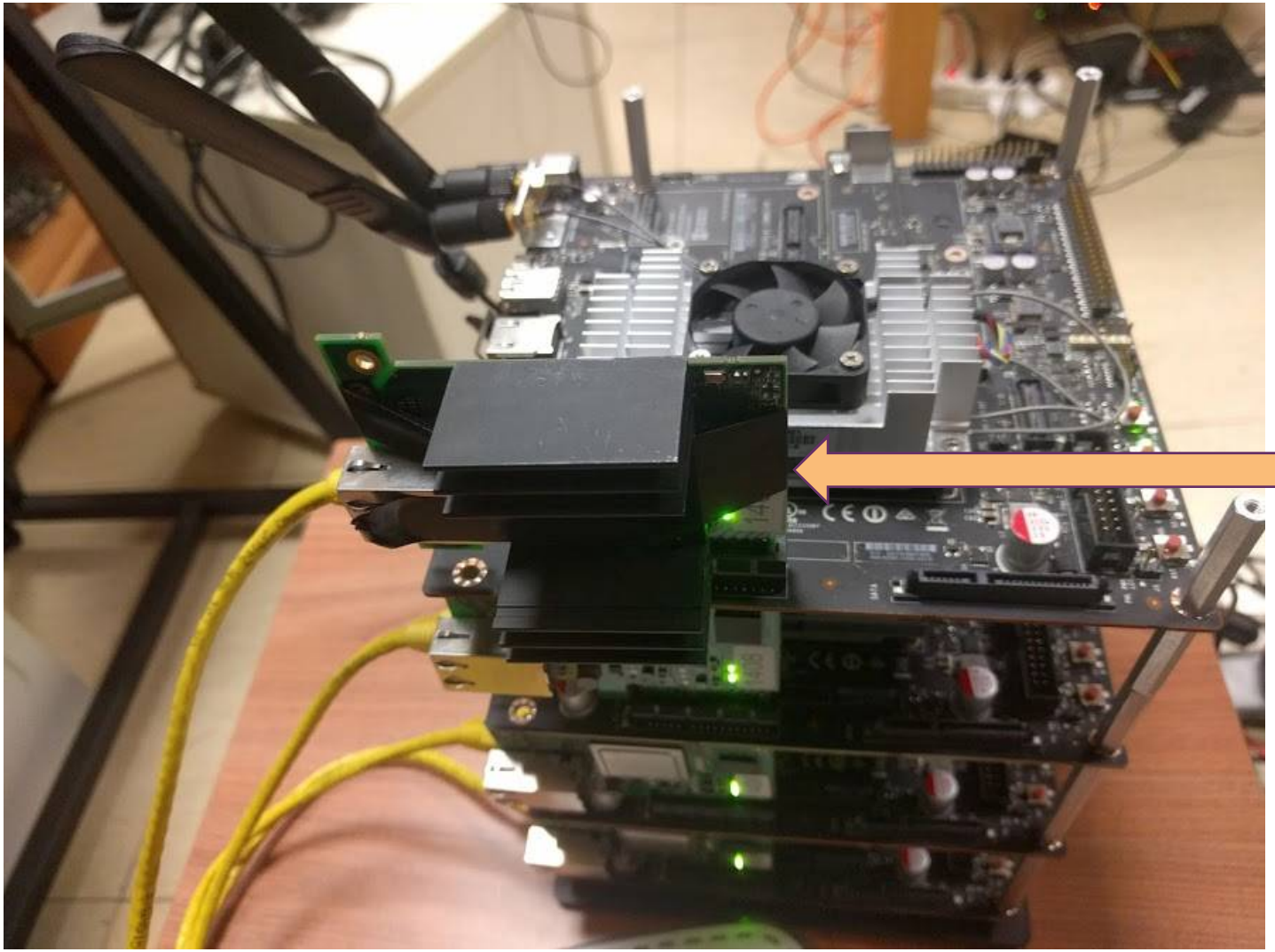






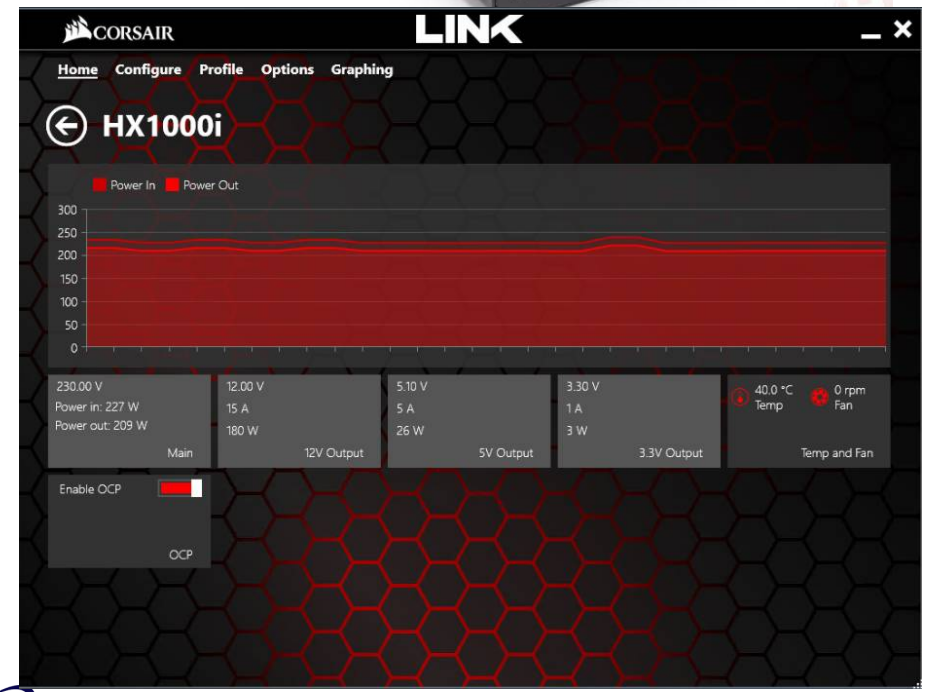


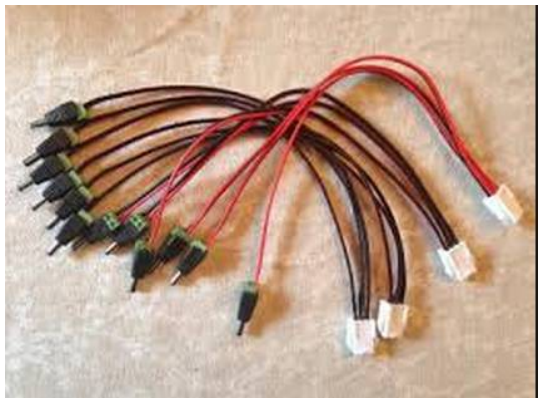
Jetson TX1 rack



PSU&Cables

- PSU HX1000i
  - 12 lines@12V (Jetson+Intel)
  - 6 lines@5V (other boards)
  - 2lines@3V (n3700)
- GRIDSEED Cable
  - 1 MOLEX ➔ 3 BARREL

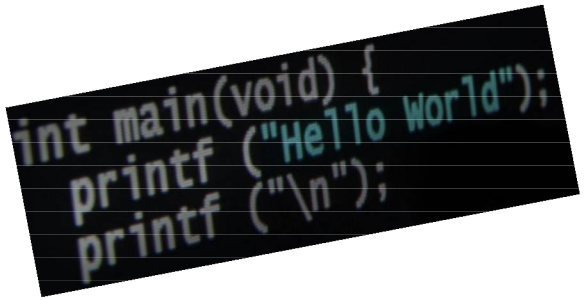




Applications

- **Experimental Physics**
  - Montecarlo and analysis of LHC experiments
  - Applications needing portable systems: Computed tomography

- **Theoretical Physics**
  - Parallel applications commonly run in HPC systems:
    - Relativistic astrophysics
    - Lattice Quantum Chromo-Dynamics simulations
    - Lattice Boltzmann fluid dynamics
    - Monte Carlo simulations of Spin-Glass systems

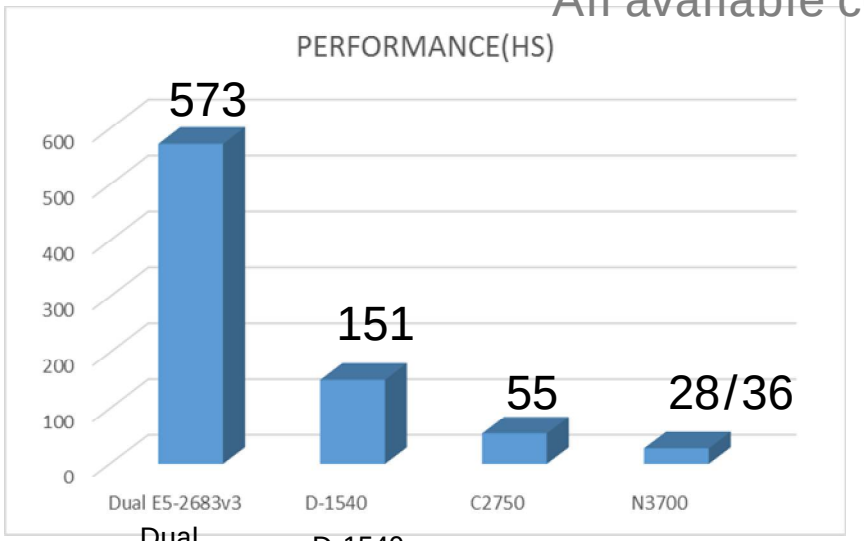


- **Systems biology**
  - Space-aware stochastic simulator

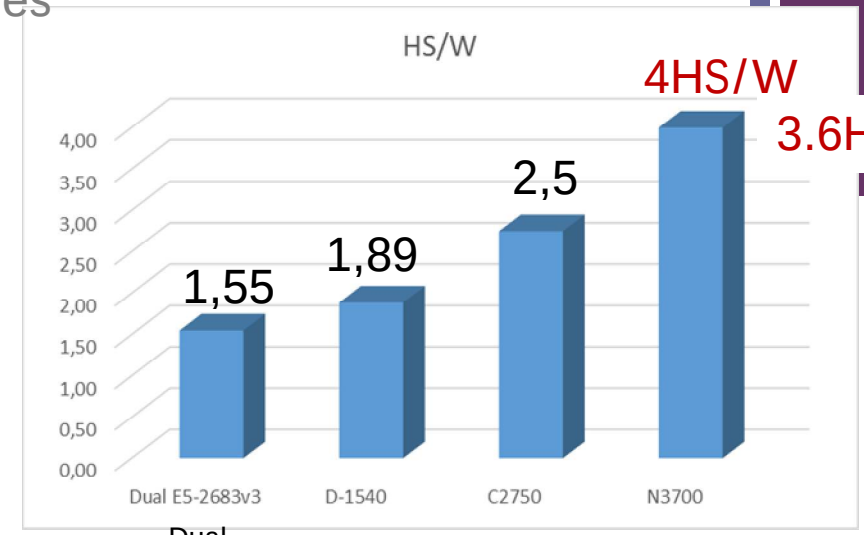
- **Deep learning and neural networks**
  - Image classification and segmentation

HS06 on Intel platforms

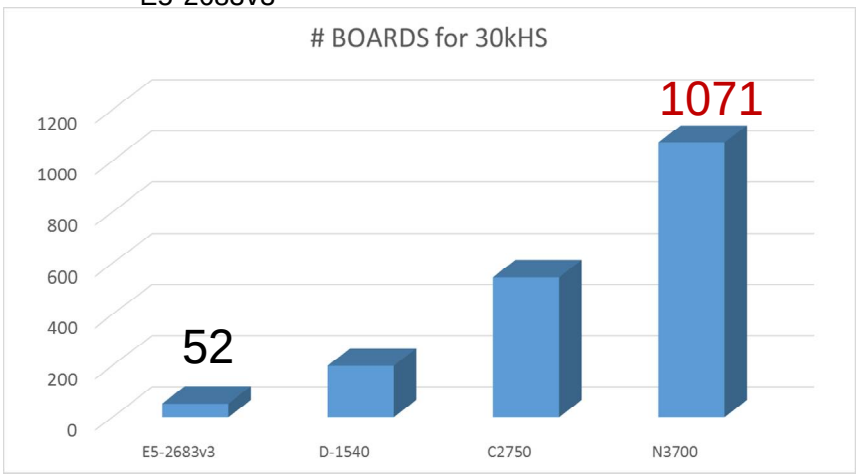
All available cores

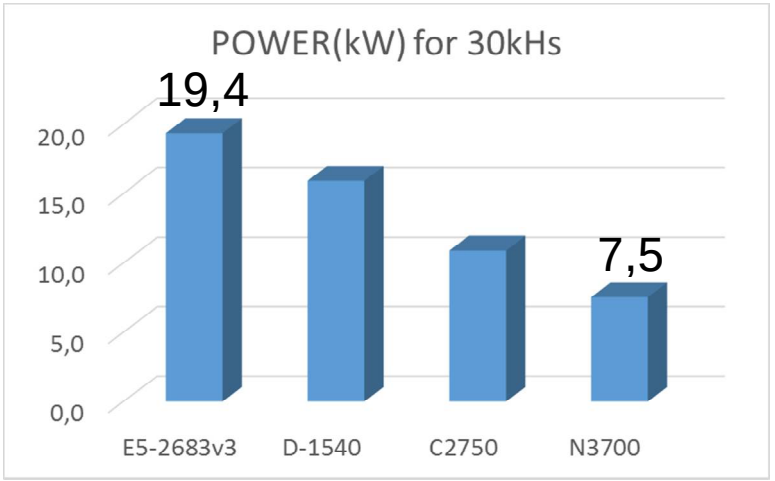


Dual E5-2683v3 | D-1540 | C2750 | N3700/J4205



Dual E5-2683v3 | D-1540 | C2750 | N3700/J4205



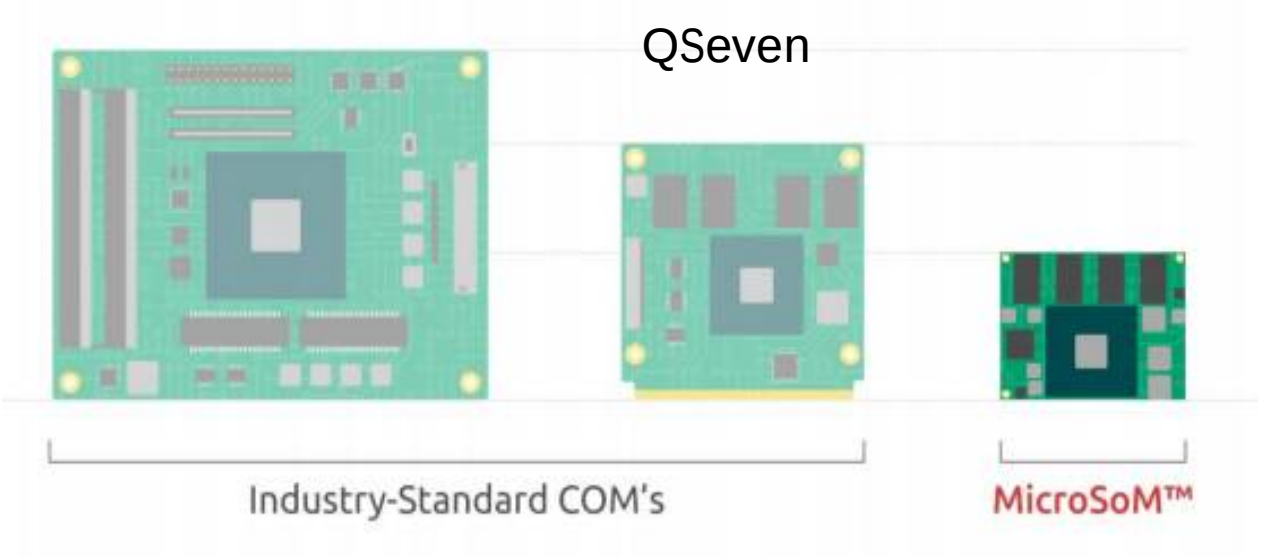


EXTREME INTEGRATION NEEDED!!

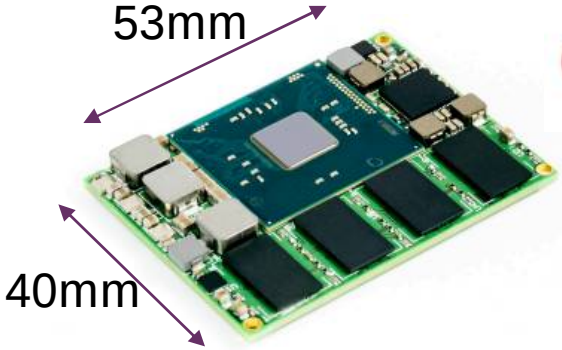
Integration with SoMs

- Can we integrate something using existing SoMs ?
  - Similar to HP moonshot with off-the-shelf components
  - Standard communication links
    - Easy to change the SoMs

PCI COM Express 10



SOMs and carriers with N3700 and J4205

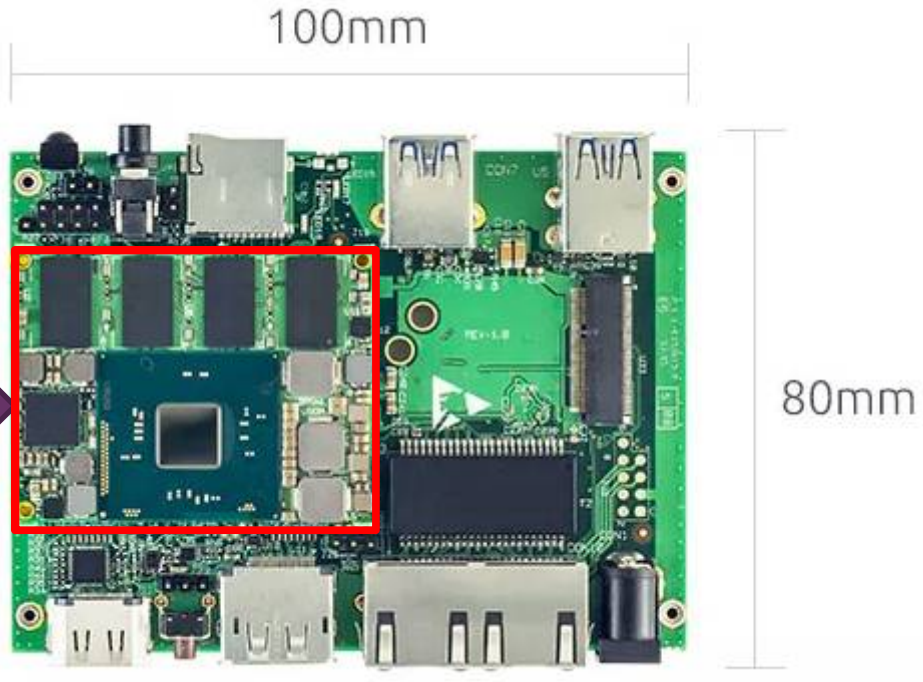


Braswell MicroSoM

Pentium N3710, 8GB, 1Gbe, SATA, PCI2

SolidRun
Do Your Thing

Proprietary bus



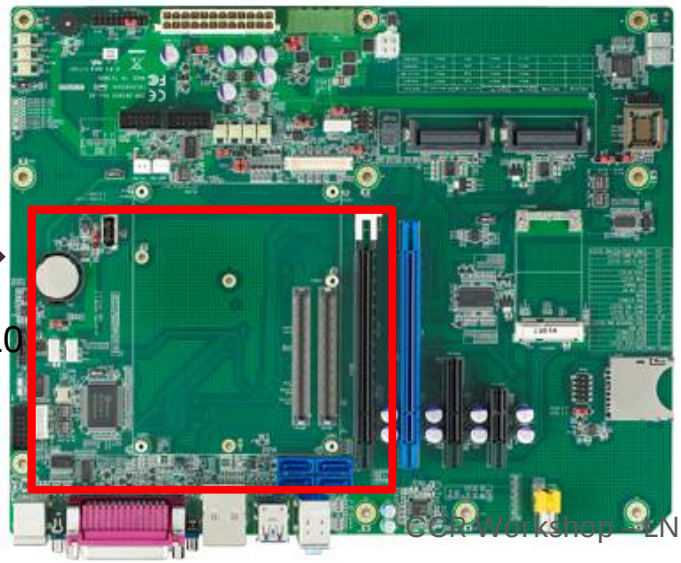
Carrier Board

SOM-7568

Pentium N3710, 8GB, 1Gbe, SATA, PCI2

ADVANTECH

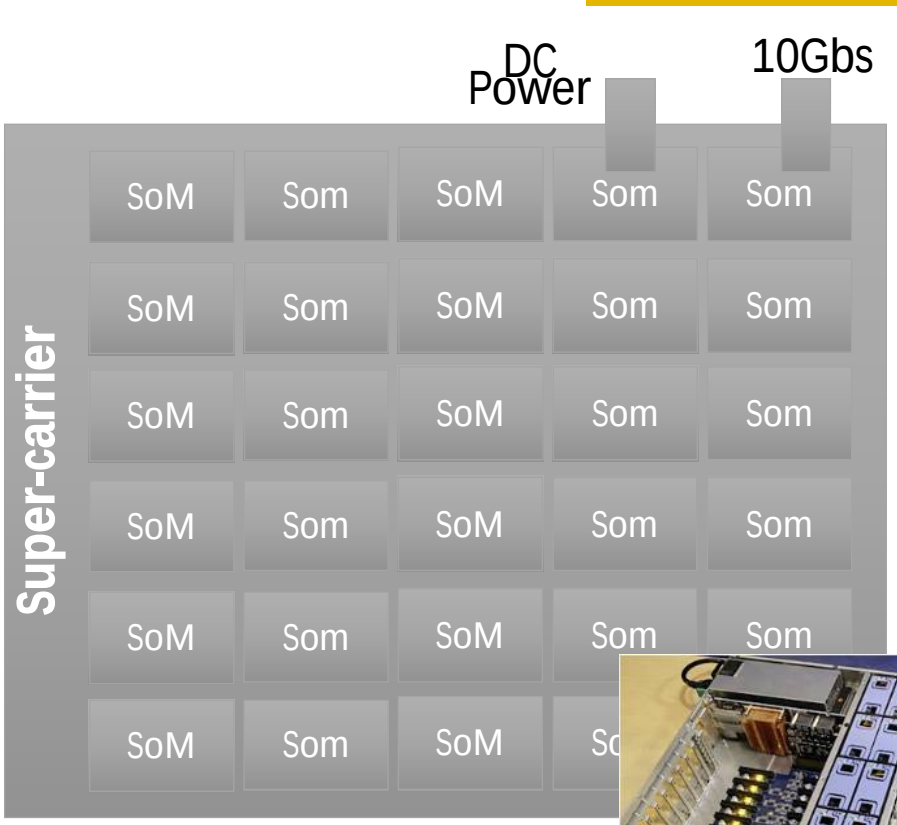
COM Express® Type 10



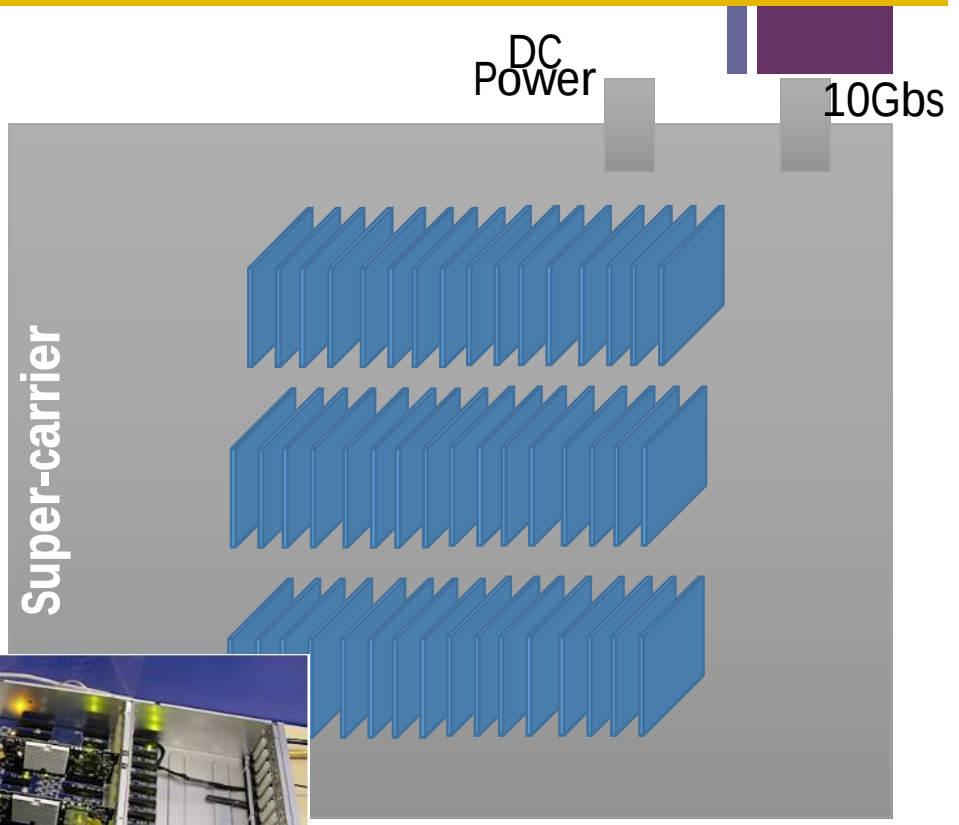
ATX form factor

SUPER-CARRIER (350W, 1344HEPSPEC)

- 48 SOMs (every SOM: 7W, 28HEPSPEC)
- 48 2.5'' HDDs (every HDD is face-to-face with every SOM)
- DC Power connector
- 10 Gbit connector



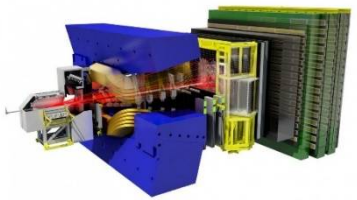
PLANAR SOMs

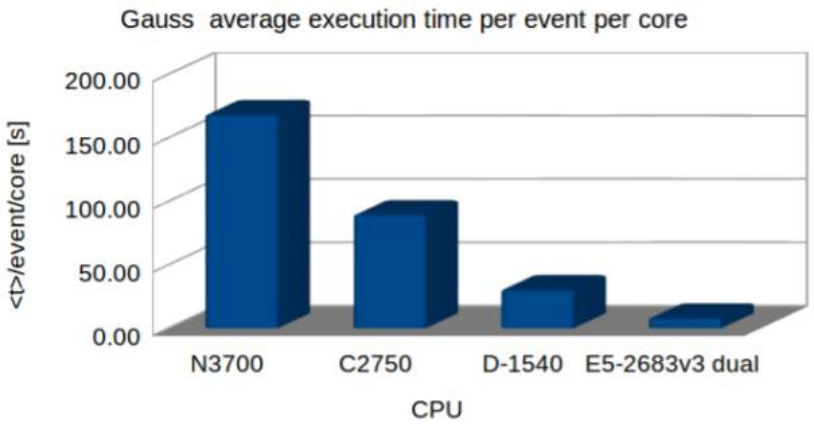


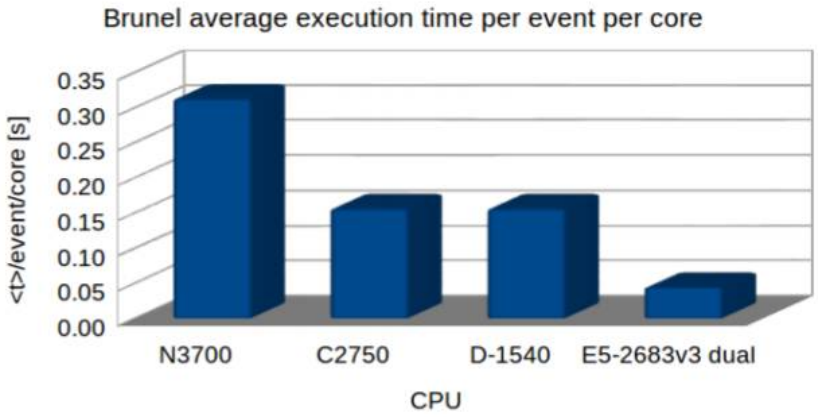
VERTICAL SOMs

NEC 368 avoton core in 2U

Experimental Physics: LHCb sw

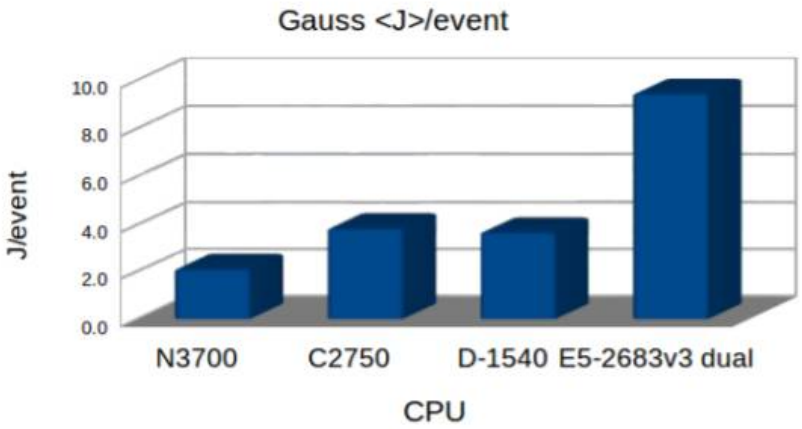


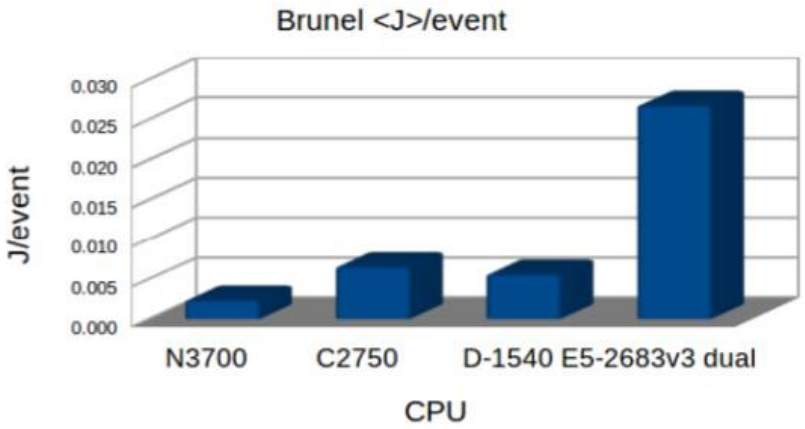




LHCb simulation (CPU bound)

Offline reconstruction (IO bound)

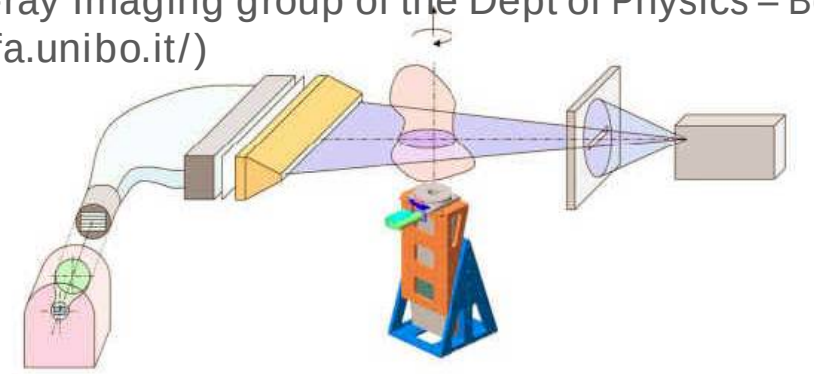


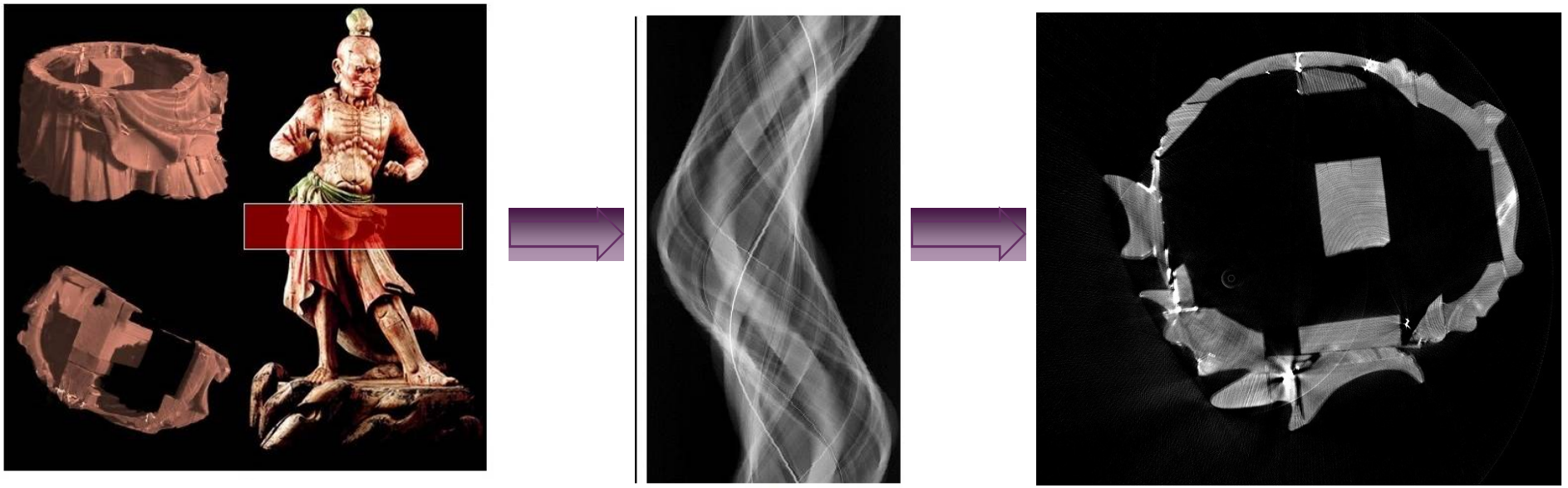


Computer tomography

Filtered Backprojection Algorithm

In collaboration with the X-ray Imaging group of the Dept of Physics – Bologna University (http://xraytomography.difa.unibo.it/)

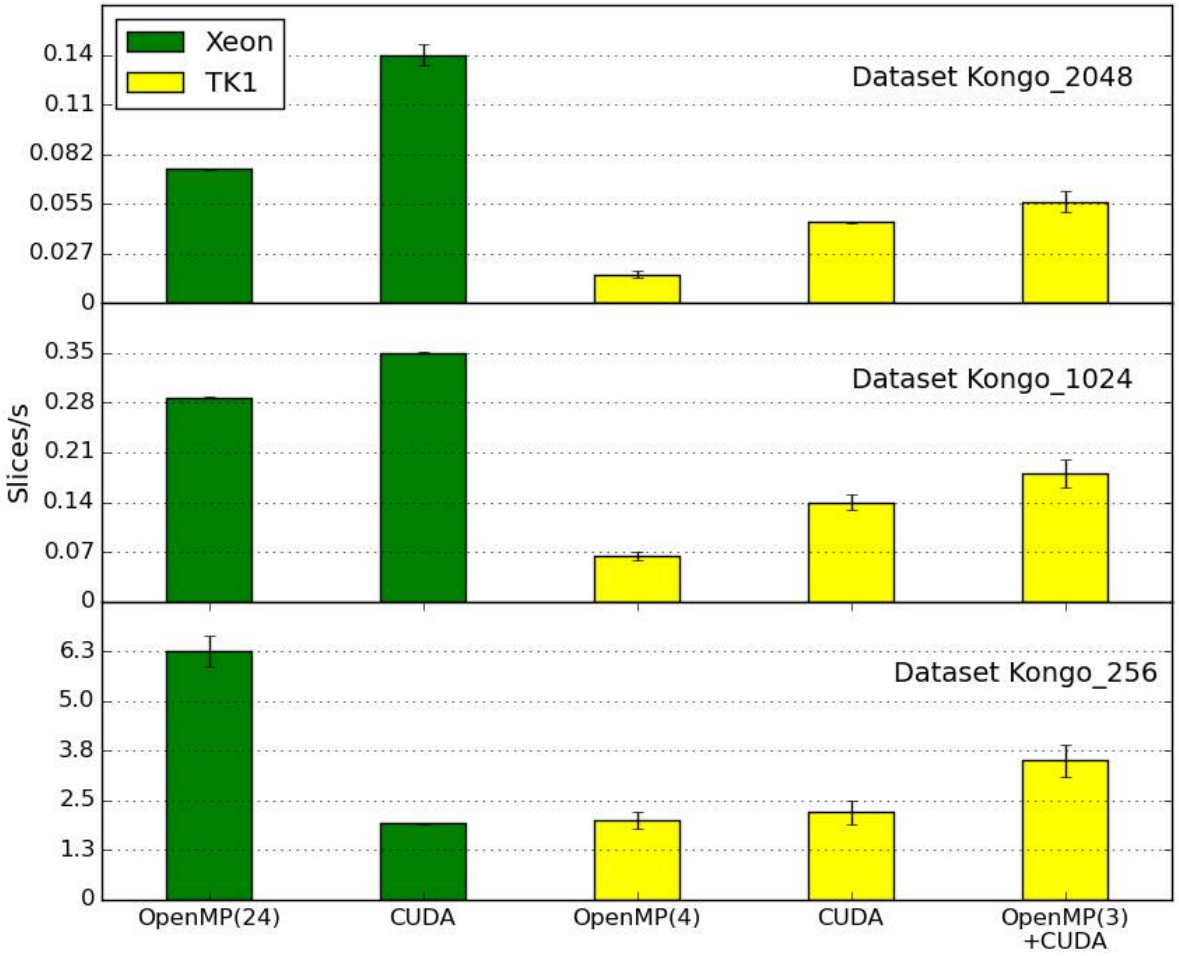




Real-Time Reconstruction for 3-D CT Applied to Large Objects of Cultural Heritage, R. Brancaccio, M. Bettuzzi, F. Casali, M. P. Morigi, G. Levi, A. Gallo, G. Marchetti, and D. Schneberk, IEEE TRANSACTIONS ON NUCLEAR SCIENCE, VOL. 58, NO. 4, AUGUST 2011

FBP Algorithm - Productivity

Number of reconstructed slices for time unit

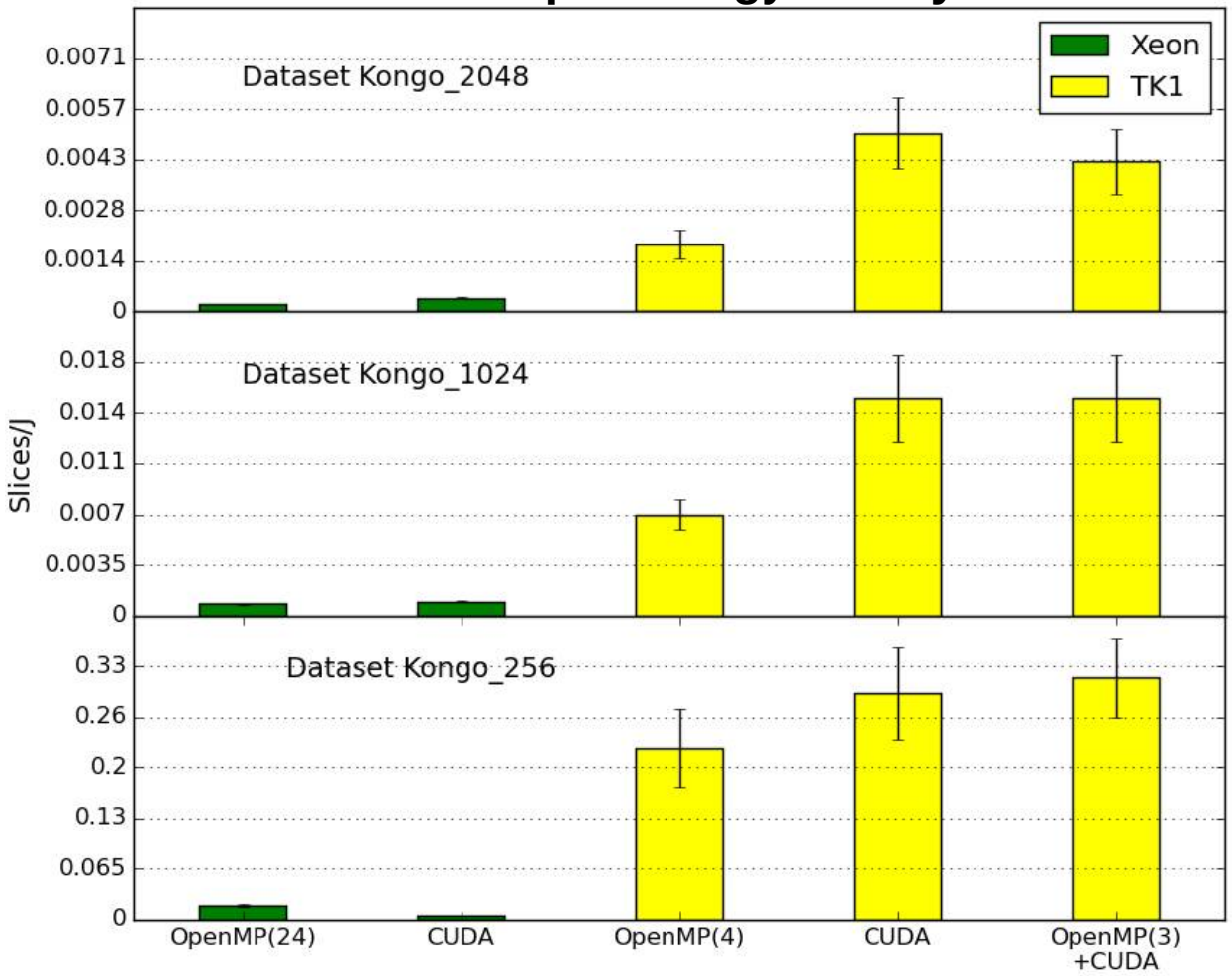


Xeon is a dual E5-2620 + NVIDIA K20
TK1 is the NVIDIA Jetson TK1

- Not surprisingly, the Xeon guarantees a higher speed than the SoC architecture
- The multi-threaded version of the algorithm is faster than the GPU version for small sizes of the slice when the application performances are broken by data transfer to and from device

FBP Algorithm - Energy efficiency

Reconstructed slices per energy unit by different runs



Xeon is a dual E5-2620 + NVIDIA K20
TK1 is the NVIDIA Jetson TK1

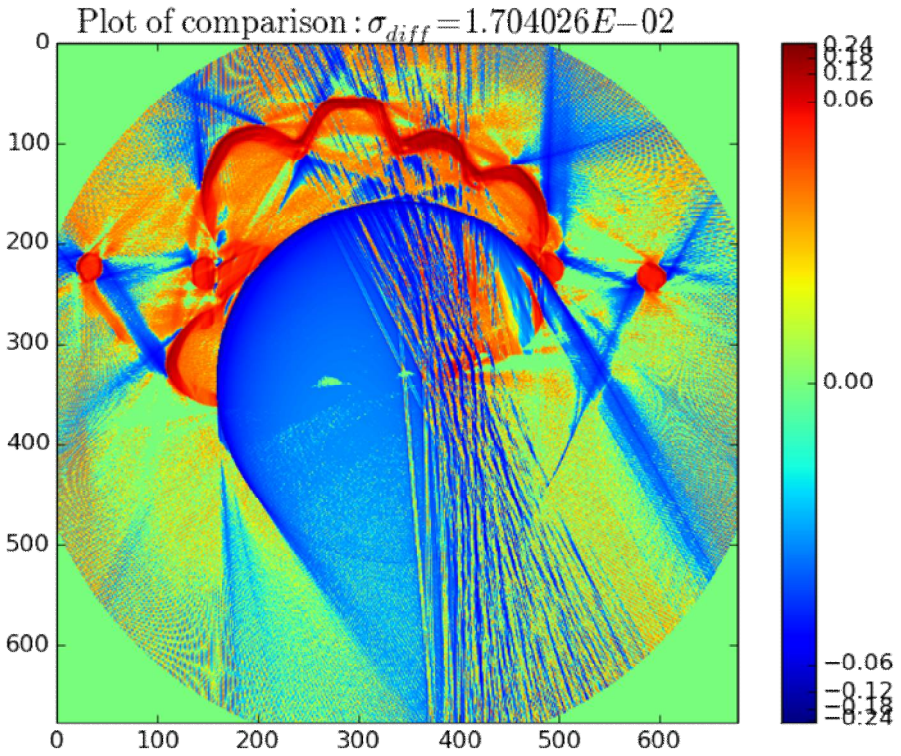
In one hour, considering the 1024x1024 slice and combining the CUDA and OMP runs on both architectures:

- **5** TK1:
  2340 slices consuming
  41W
- **2x**E5-2620+**1**K20
  2268 slices consuming
  350W

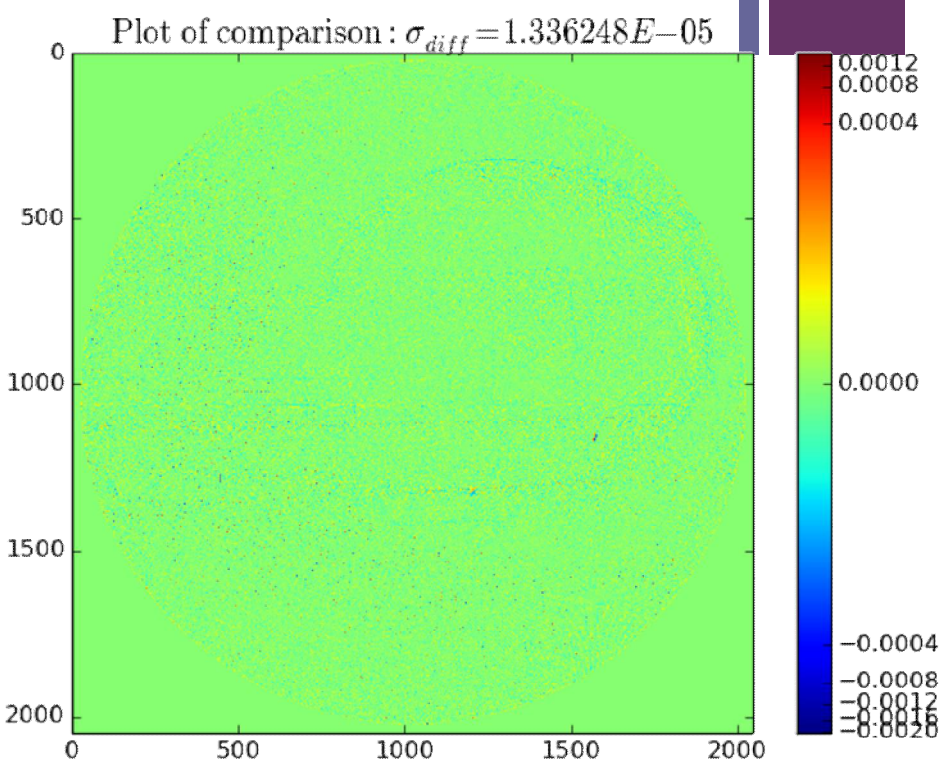
Master thesis of Elena Corni:
Implementazione dell'algoritmo Filtered Back-Projection (FBP) per architetture Low-Power di tipo Systems-On-Chip

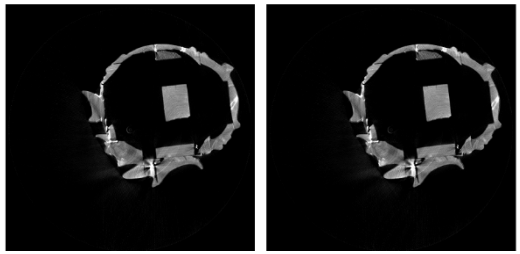
Numerical correctness

Wrongly reconstructed slice



Accurately reconstructed slice

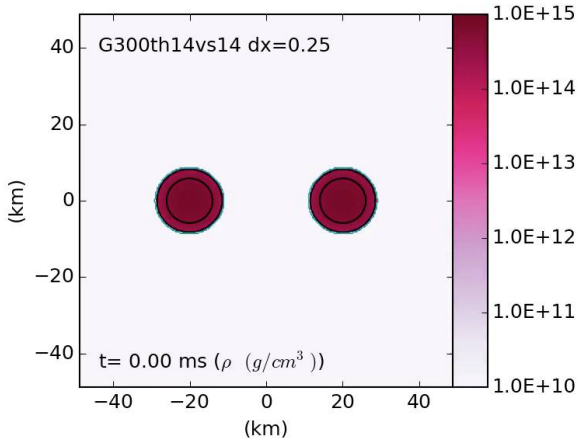


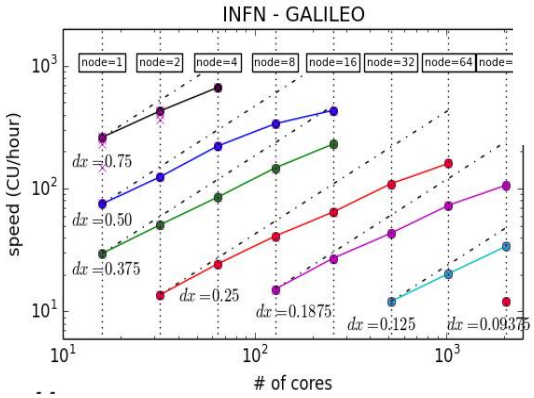


Einstein Toolkit

Roberto De Pietri , Roberto Alfieri - INFN Parma and Parma University

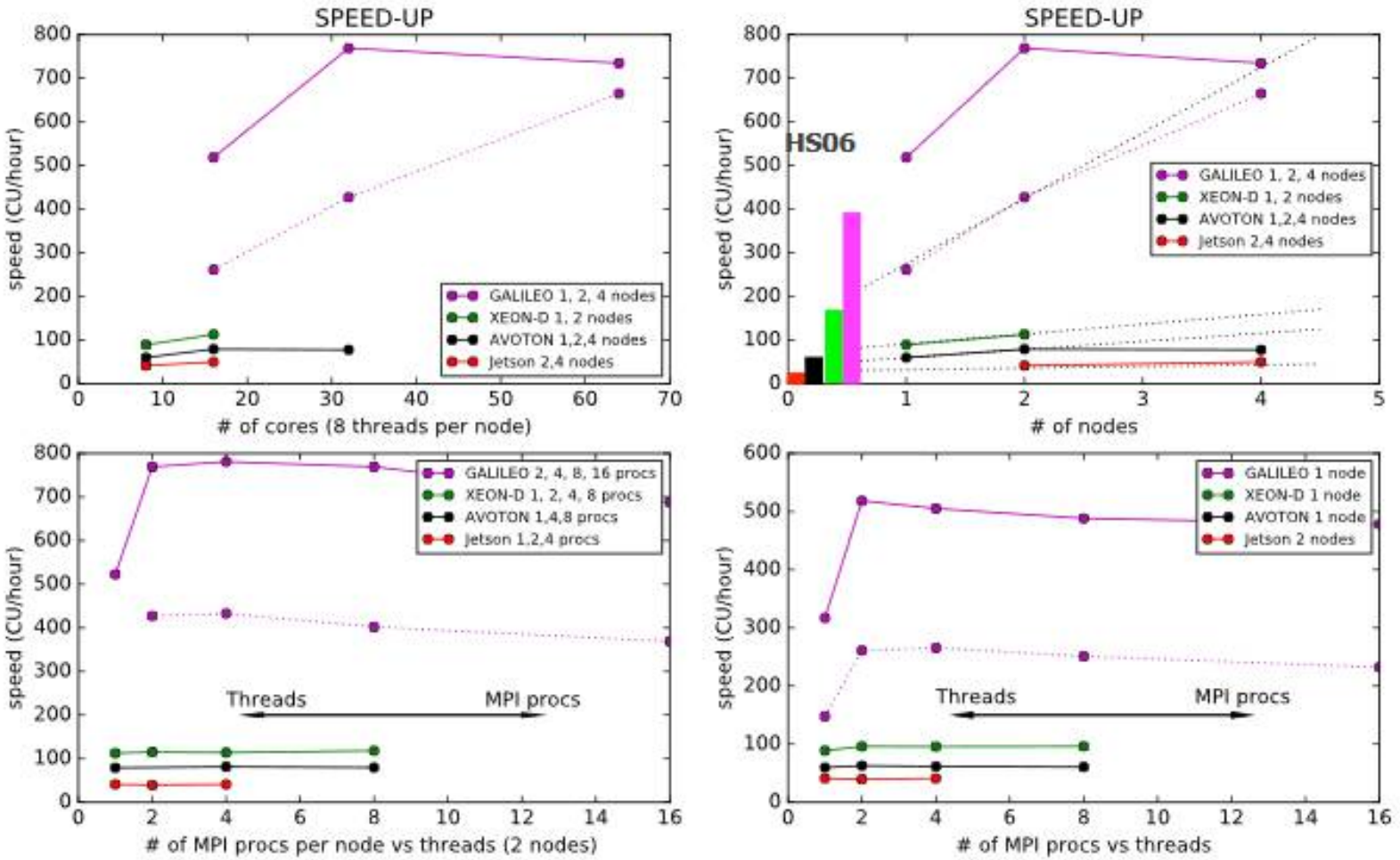
- **The scientific case: high resolution simulation of inspiral and merger phase of binary neutron stars system**
  - **one of source of the gravitational waves that are the observational target of the LIGO/VIRGO experiment**
- **Computation performed using The Einstein ToolKit**
- **Result obtained on Galileo at CINECA**
- **COSA low power systems**
  - **Basic performance analysis**
  - **Porting of the application**
  - **Comparative results analysis**





Einstein Toolkit on COSA low power systems

Roberto De Pietri , Roberto Alfieri - INFN Parma and Parma University



https://agenda.infn.it/getFile.py/access?contribId=19&resId=0&materialId=slides&confId=10434

Einstein Toolkit on COSA low power systems

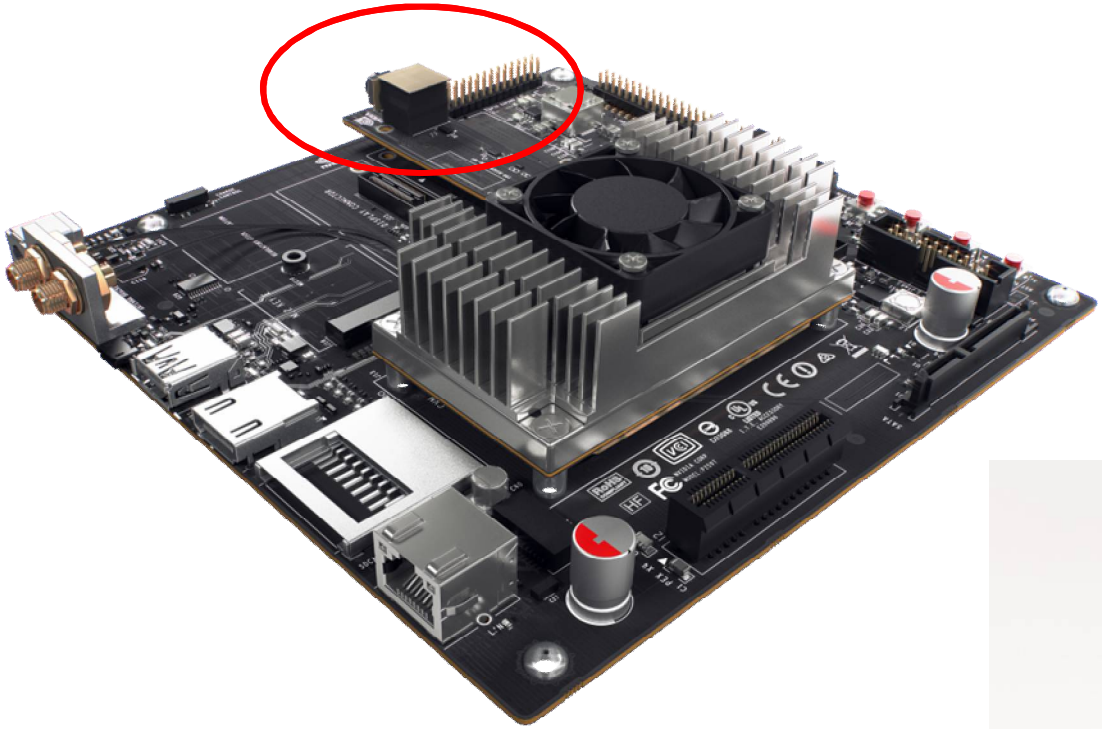
Roberto De Pietri , Roberto Alfieri - INFN Parma and Parma University

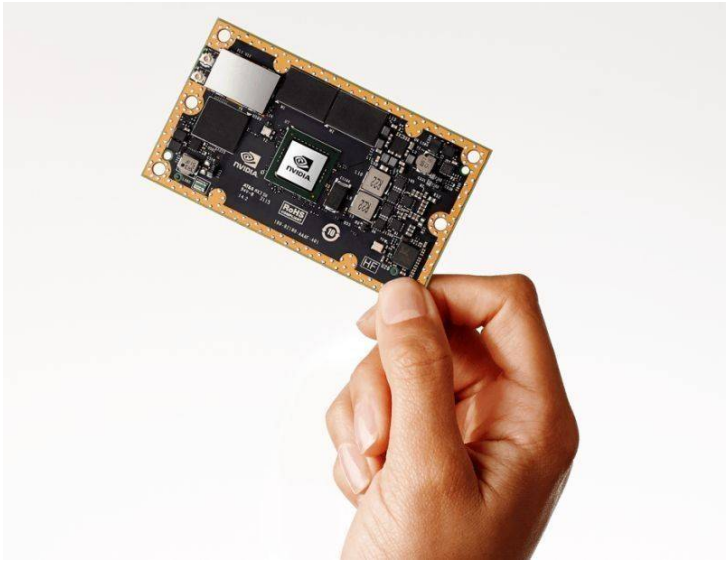
- GOOD NEWS: the framework **works** on LOW-POWER architectures
- BAD NEWS: performance not up to the par of traditional High-End Processor. Memory limitation would require an even higher number of nodes interconnected with **a high speed network**
- In order to run our application on Low Power architectures at production level **we need to exploit the accelerator present on the system (GPU)** in order to speed up the computation

....but.....

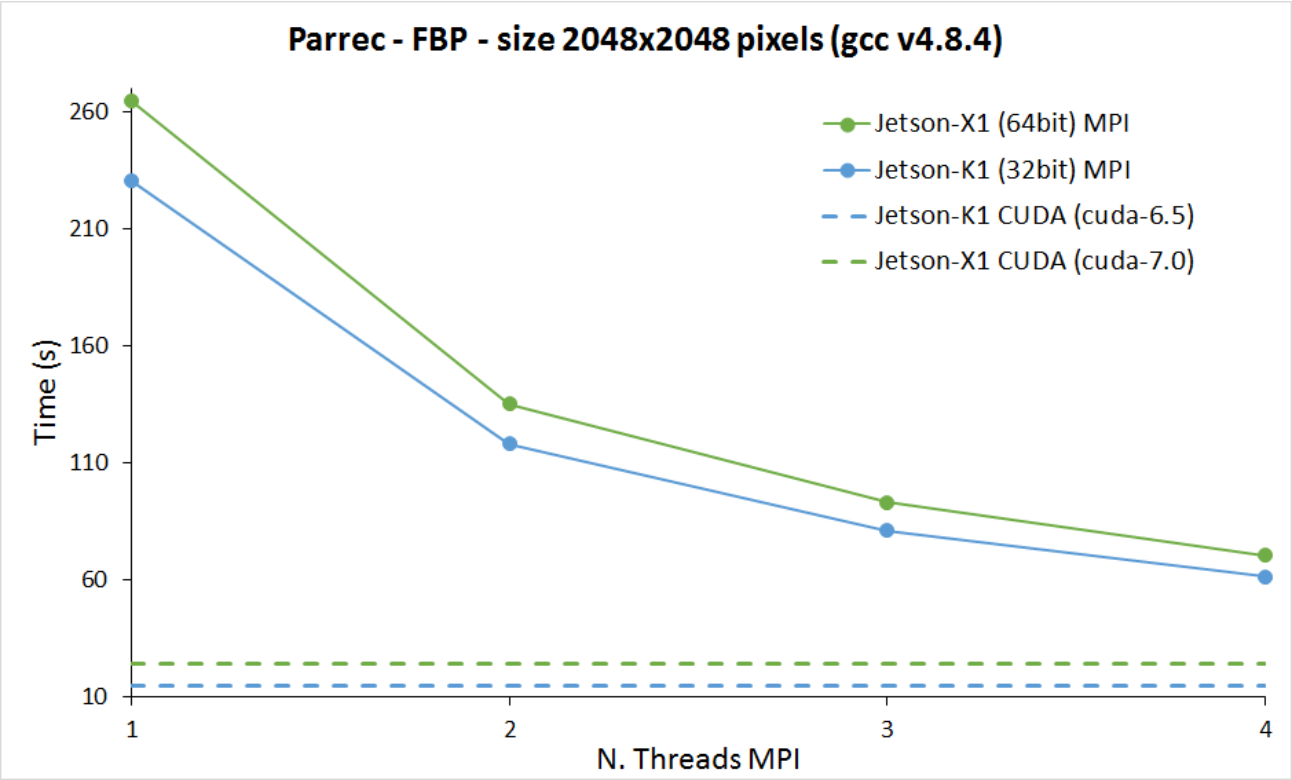
- We cannot change the code for every new hardware device

Jetson TX1





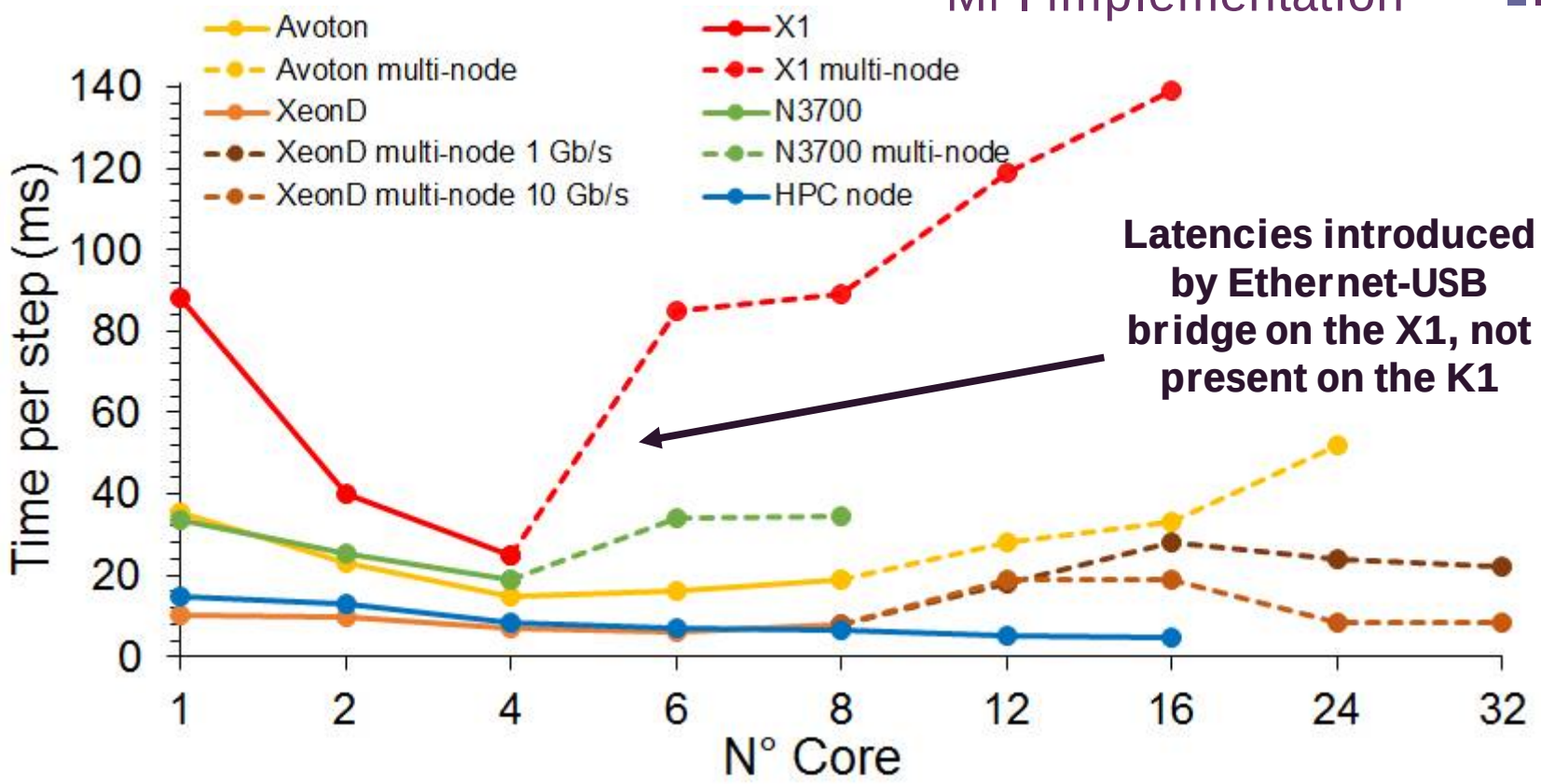
Jetson TX1 tests



- Slightly better CPU cores on the K1 (2.3GHz vs 1.73GHz)
- Better GPU on the X1

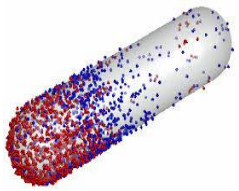
Space-aware stochastic simulator

MPI implementation



Latencies introduced by Ethernet-USB bridge on the X1, not present on the K1

Reaction-diffusion system composed of 4096 membranes



imati

CNR - ITB

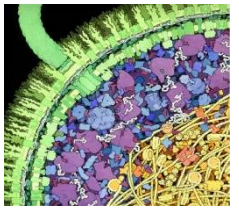
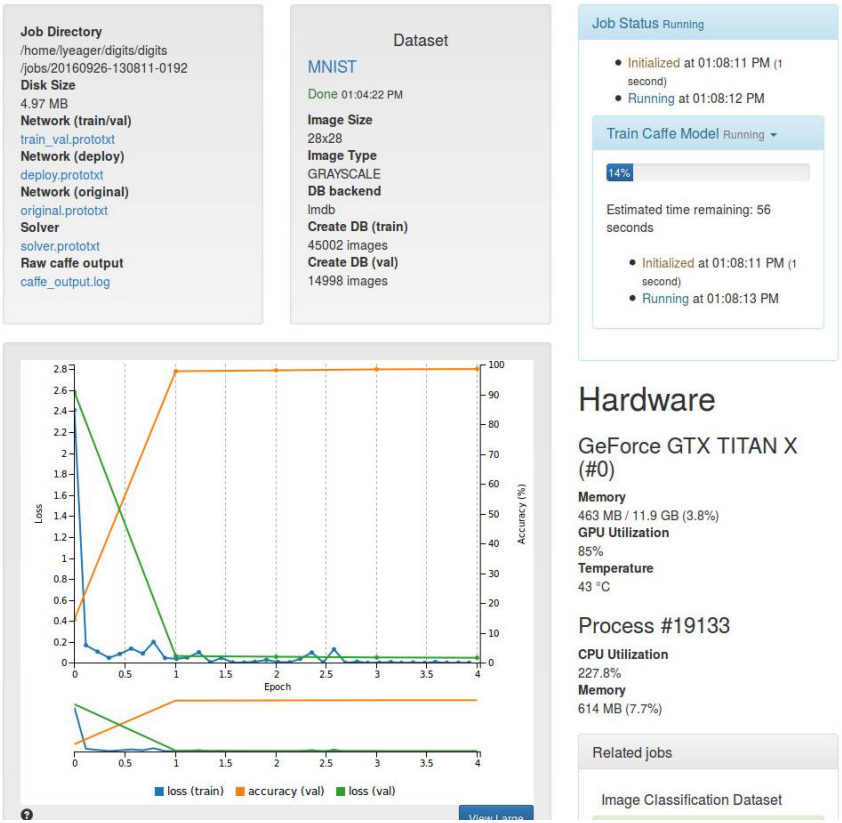


Image classification on the TX1

- **Dataset creation using ImageNet**
- **Training using DIGITS + Caffe on 2 K40 or XeonD+K20**



For comparison:

Training on CPU (Xeon E5-2640 v2, 16HT) -> 4 cycles in 1 day, then stopped

Training on GPU for the same model -> 90 cycles in 3h27m

@Tirocinio Gianluca Guidi

Image classification on the TX1

- **Dataset creation using ImageNet**
- **Training using DIGITS + Caffe on 2 K40 or XeonD+K20**
- **Inference on Jetson X1 using NVidia TensorRT**
- **Real time classification**

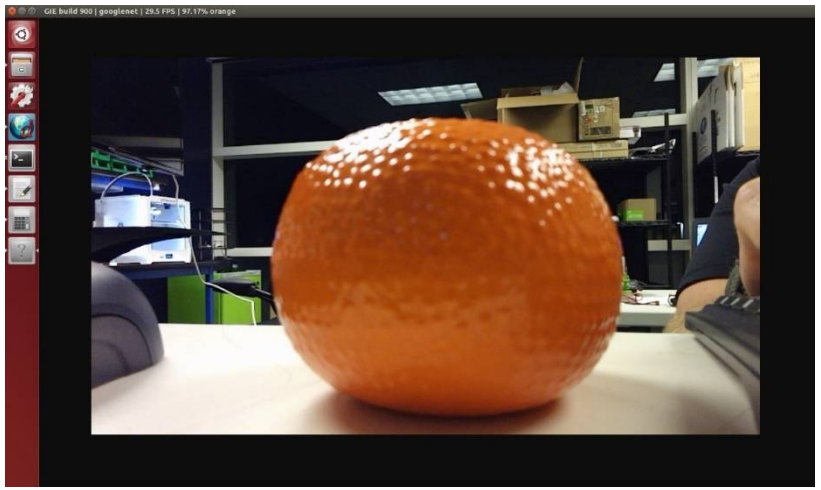


Image classification on the TX1

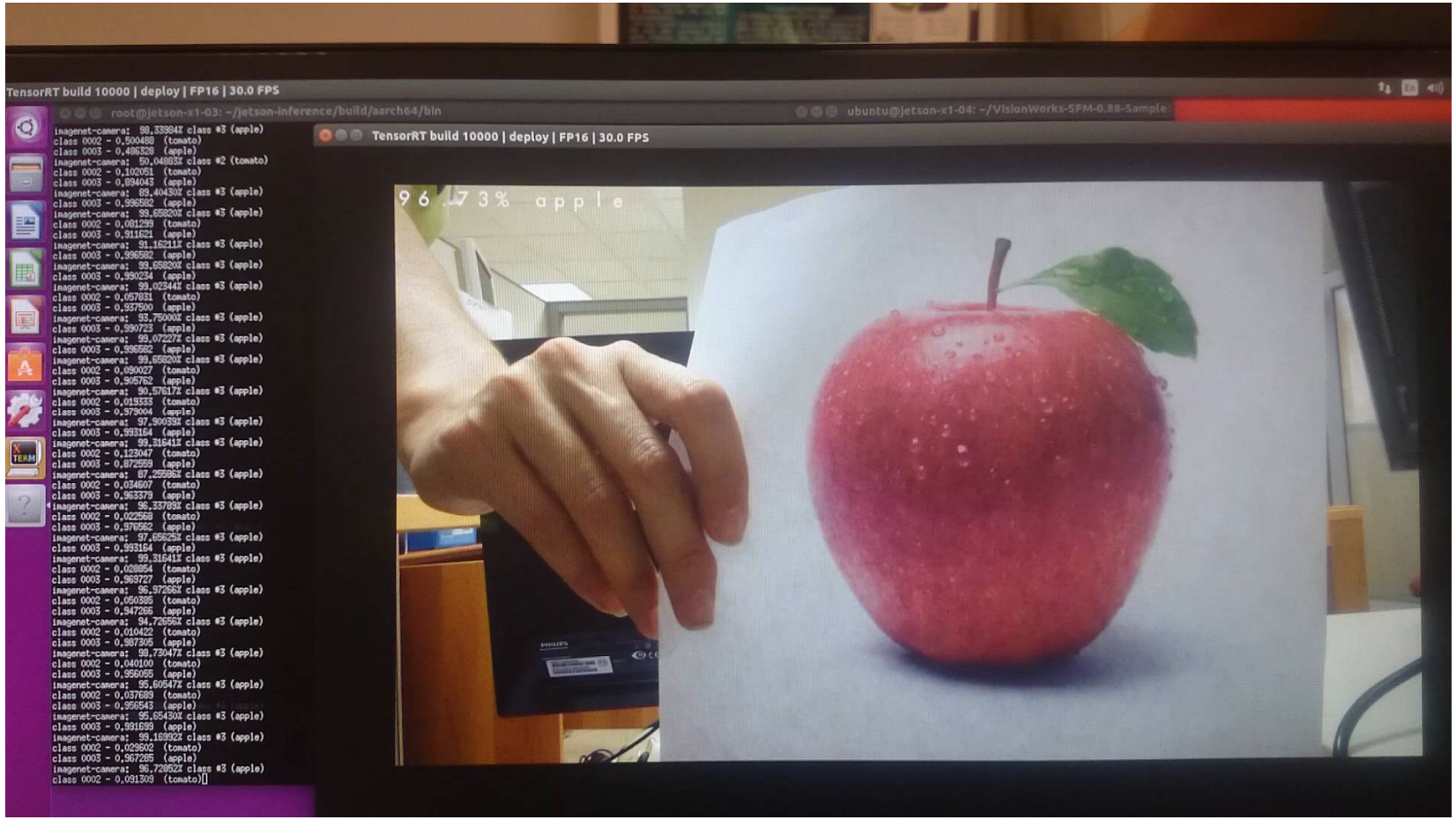
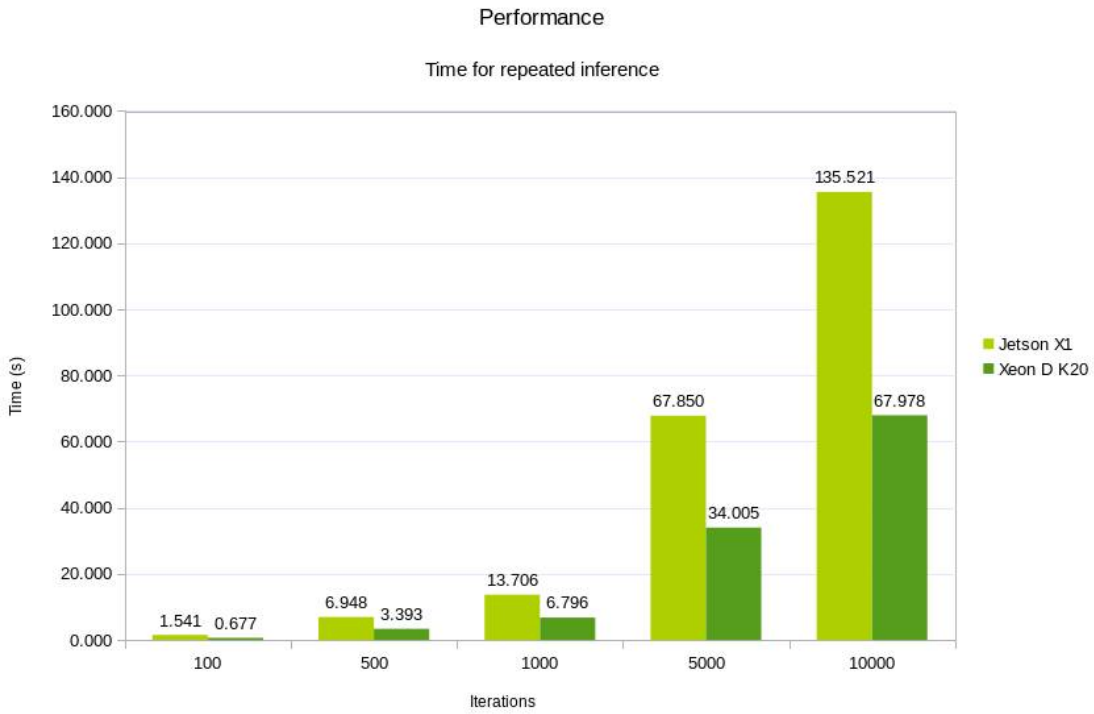
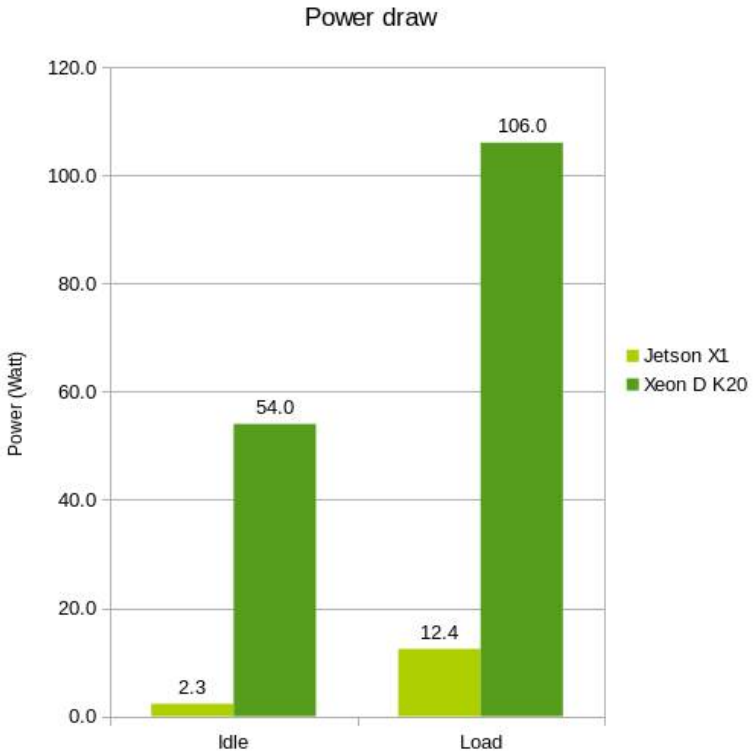


Image classification on the TX1

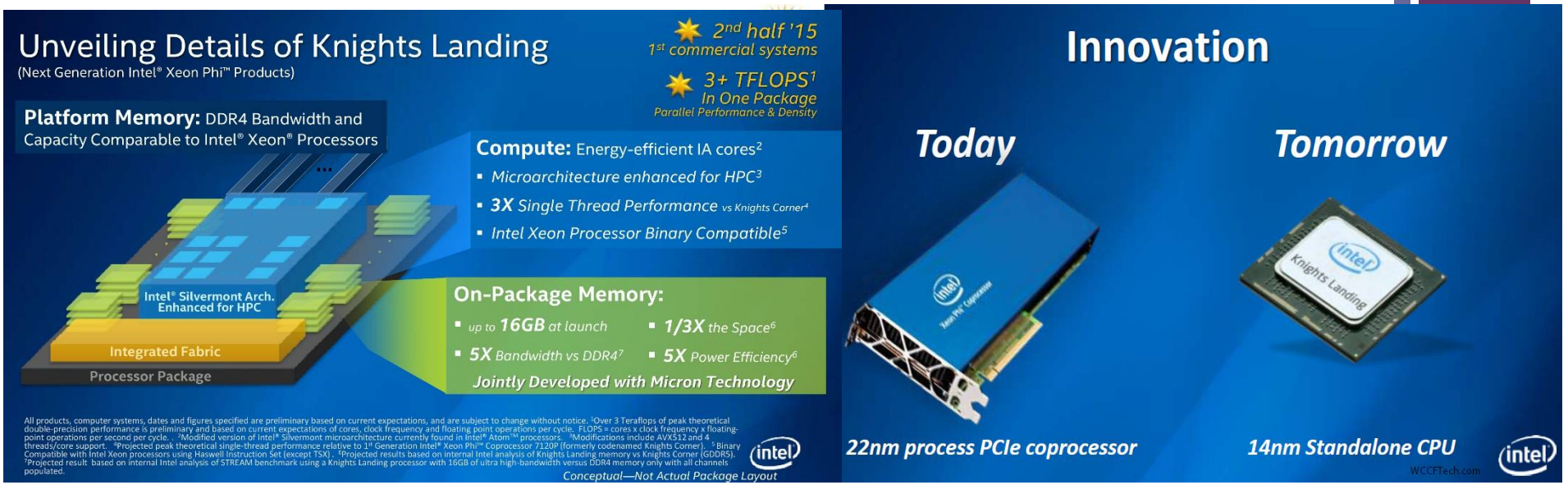




- XEOND + K20 is 2x faster, but…
- …TX1 is 10 times less power hungry

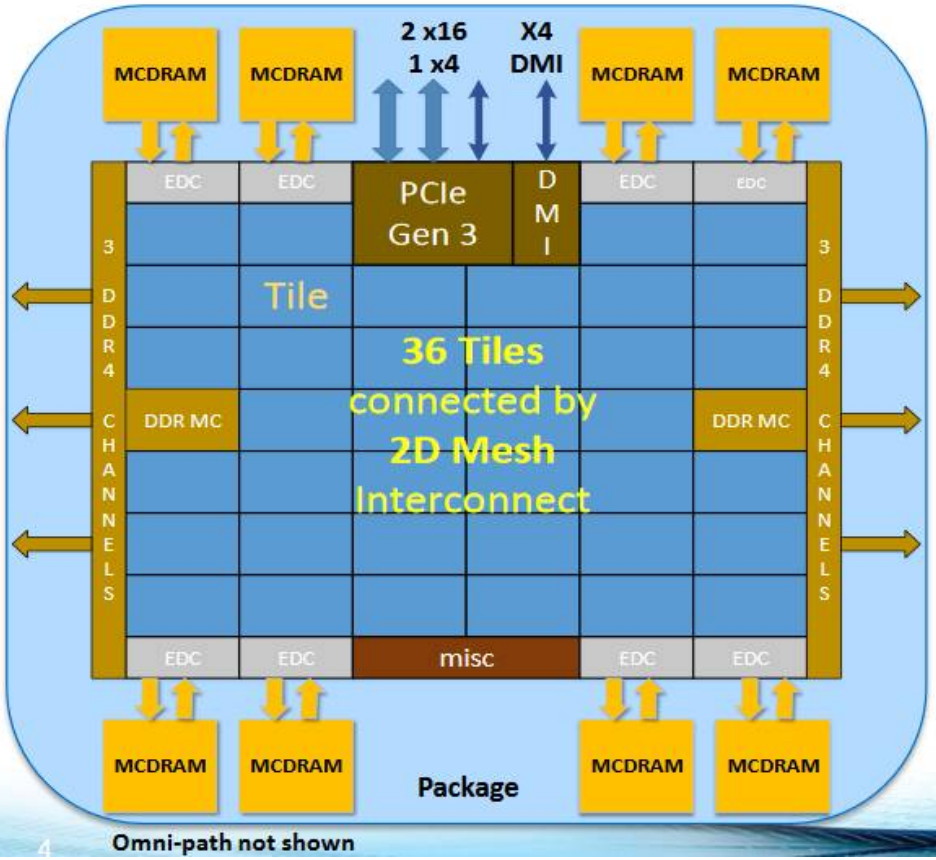
@Tirocinio Gianluca Guidi

Intel KNL XEON-PHI



- up to 72 Airmont (Atom) cores with four threads per core
- up to 384 GB of "far" DDR4 RAM and 8–16 GB of stacked "near" 3D MCDRAM
- Each core will have two 512-bit vector units and will support AVX-512 SIMD instructions
- Standalone version
- Onmipath fabric integrated (2x100gbs)

Knights Landing Overview



Chip: 36 Tiles interconnected by **2D Mesh**
Tile: 2 Cores + 2 VPU/core + 1 MB L2

Memory: MCDRAM: 16 GB on-package; High BW
DDR4: 6 channels @ 2400 up to 384GB
IO: 36 lanes PCIe Gen3. 4 lanes of DMI for chipset
Node: 1-Socket only
Fabric: Omni-Path on-package (not shown)

Vector Peak Perf: 3+TF DP and 6+TF SP Flops
Scalar Perf: ~3x over Knights Corner
Streams Triad (GB/s): MCDRAM : 400+; DDR: 90+

Source Intel: All products, computer systems, dates and figures specified are preliminary based on current expectations, and are subject to change without notice. KNL data are preliminary based on current expectations and are subject to change without notice. 1Binary Compatible with Intel Xeon processors using Haswell Instruction Set (except TSX). 2Bandwidth numbers are based on STREAM-like memory access pattern when MCDRAM used as flat memory. Results have been estimated based on internal Intel analysis and are provided for informational purposes only. Any difference in system hardware or software design or configuration may affect actual performance.

(*) @ http://www.hotchips.org/wp-content/uploads/hc_archives/hc27/HC27.25-Tuesday-Epub/HC27.25.70-Processors-Epub/HC27.25.710-Knights-Landing-Sodani-Intel.pdf

Flat MCDRAM SW Usage: Code Snippets

C/C++ (*https://github.com/memkind)

Allocate into DDR

```
float   *fv;
fv = (float *)malloc(sizeof(float)*100);
```

Allocate into MCDRAM

```
float   *fv;
fv = (float *)hbw_malloc(sizeof(float) * 100);
```

Intel Fortran

Allocate into MCDRAM

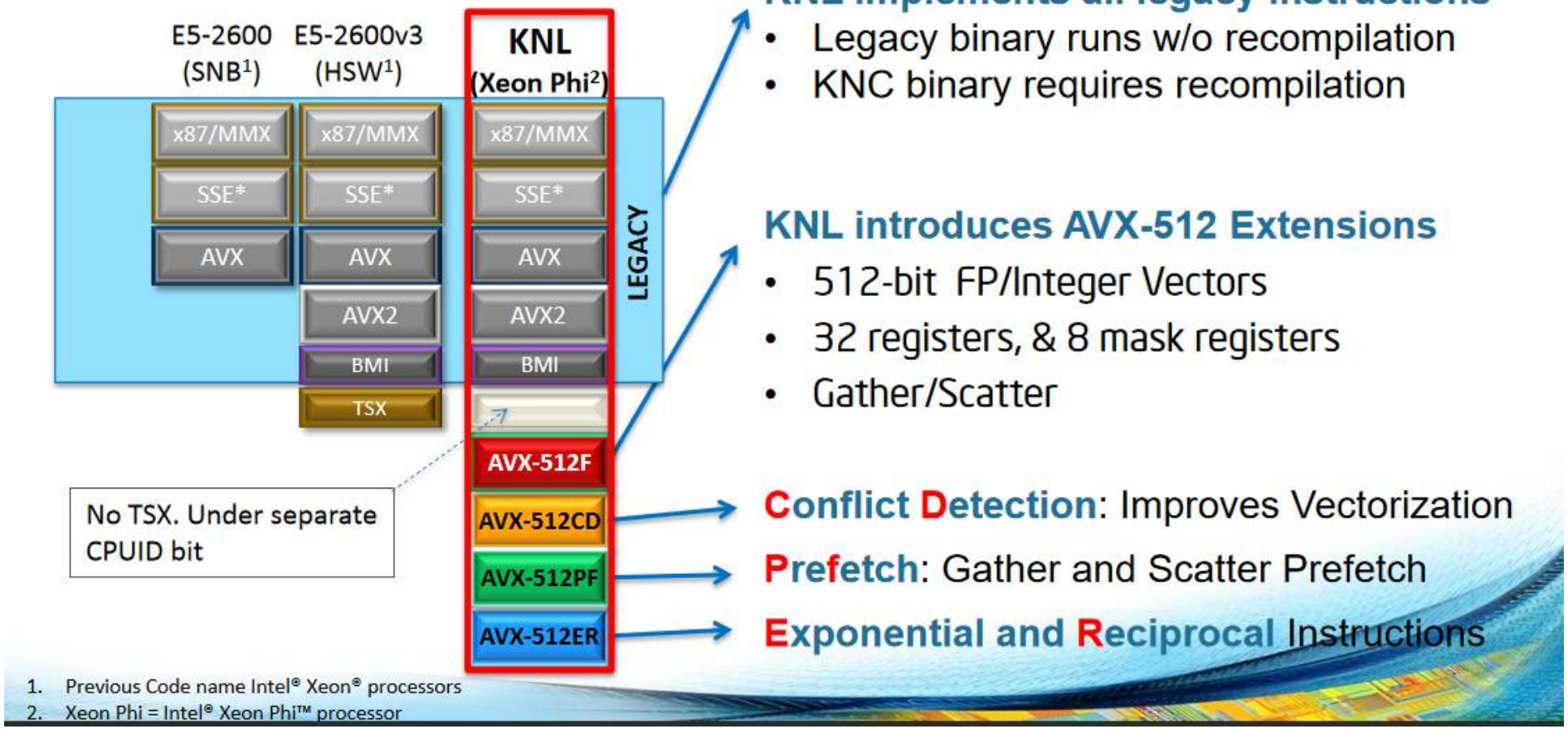
```
c      Declare arrays to be dynamic
       REAL, ALLOCATABLE :: A(:)

!DEC$ ATTRIBUTES, FASTMEM :: A

       NSIZE=1024
c      allocate array 'A' from MCDRAM
c
       ALLOCATE (A(1:NSIZE))
```

(*) @ http://www.hotchips.org/wp-content/uploads/hc_archives/hc27/HC27.25-Tuesday-Epub/HC27.25.70-Processors-Epub/HC27.25.710-Knights-Landing-Sodani-Intel.pdf

KNL ISA



(*) @ http://www.hotchips.org/wp-content/uploads/hc_archives/hc27/HC27.25-Tuesday-Epub/HC27.25.70-Processors-Epub/HC27.25.710-Knights-Landing-Sodani-Intel.pdf

Intel KNL XEON-PHI

- XEONPHI - 7230 - 64 core

Launch Date	Q2'16
Lithography	14 nm
Recommended Customer Price	$3710.00

Performance

# of Cores	64
Processor Base Frequency	1.30 GHz
Max Turbo Frequency	1.50 GHz
Cache	32 MB L2
TDP	215 W
VID Voltage Range	0.550-1.125V

KNL thread affinity

- 4 way Hiperthread
- Affinity control is important in case of numerical computation

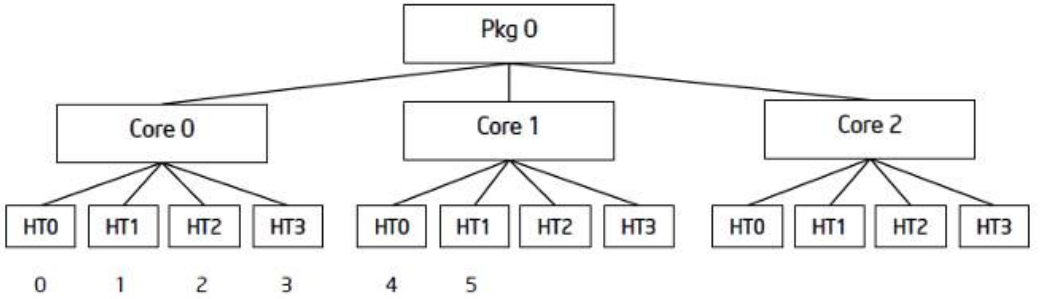


Figure 2:*KMP_AFFINITY=compact*

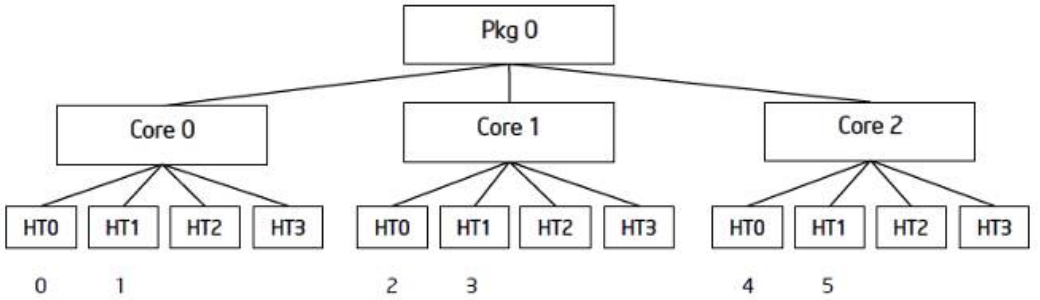


Figure 3:*KMP_AFFINITY=balanced*

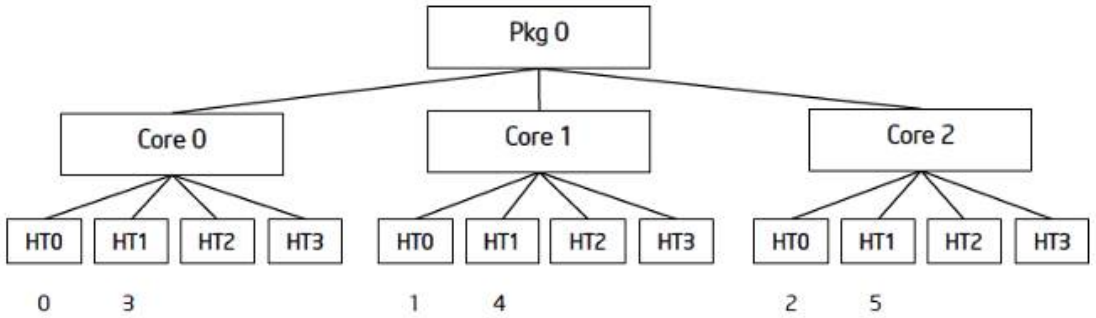
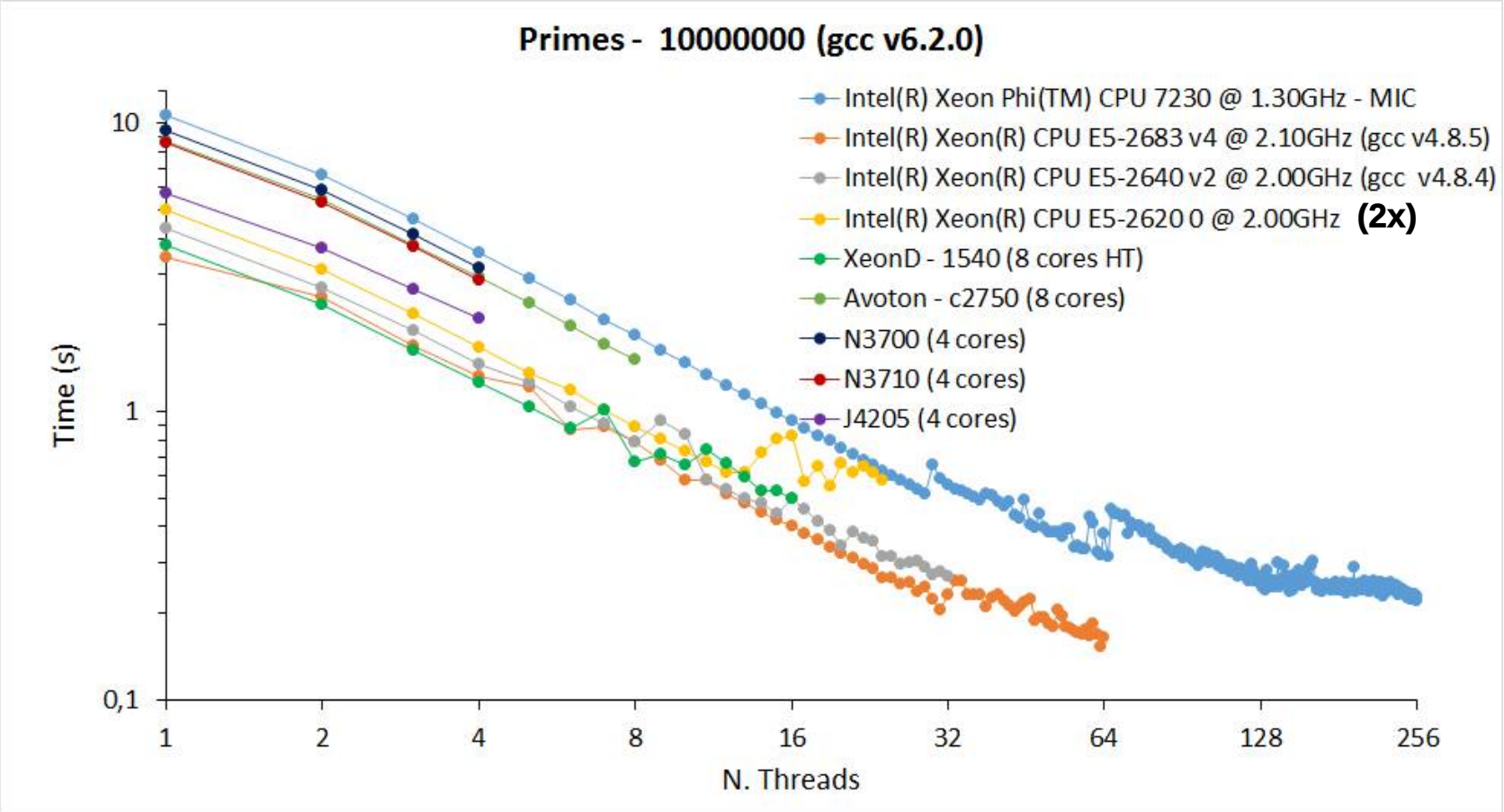


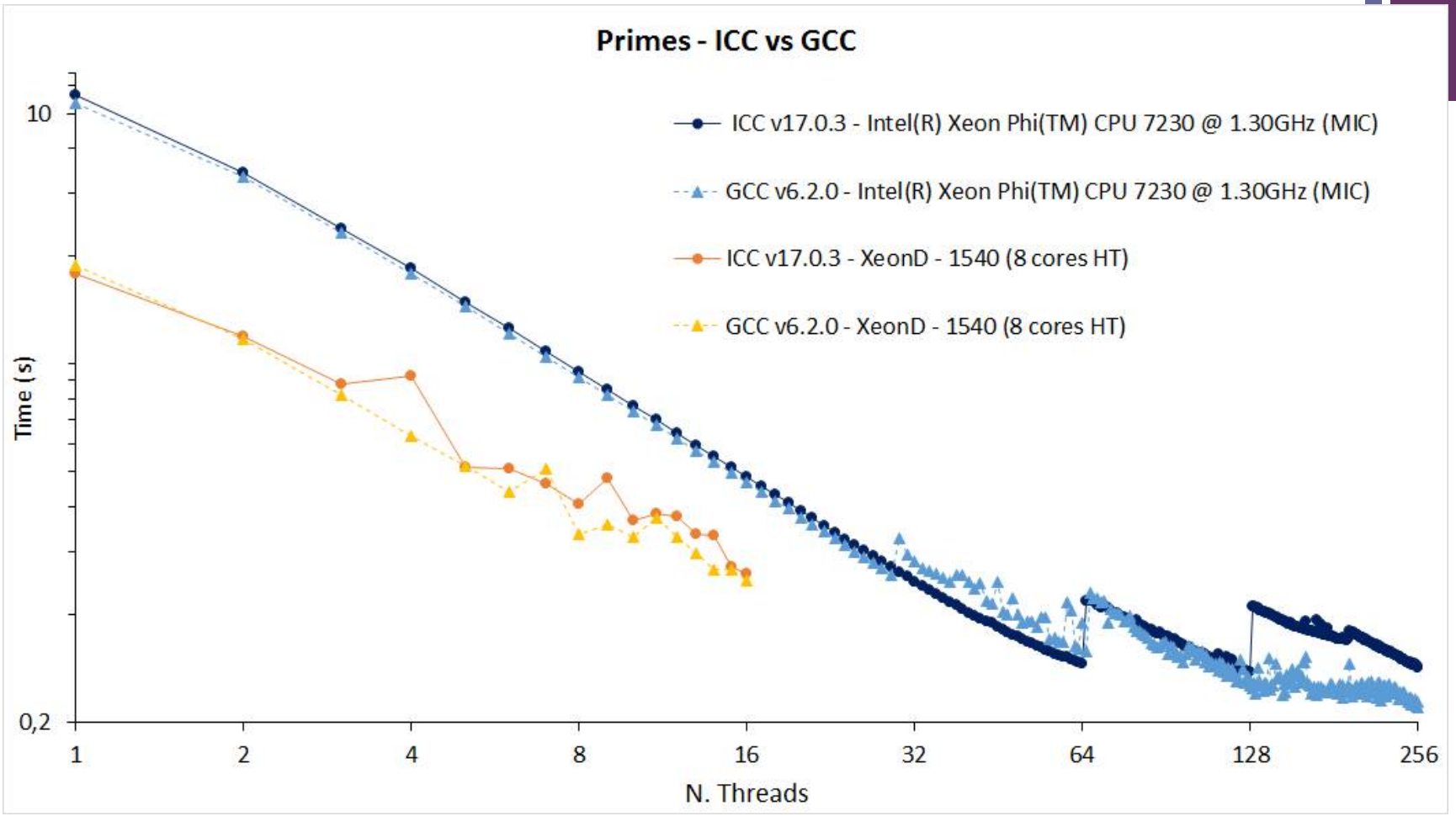
Figure 4:*KMP_AFFINITY=scatter*

KNL test with synthetic tests -Primes

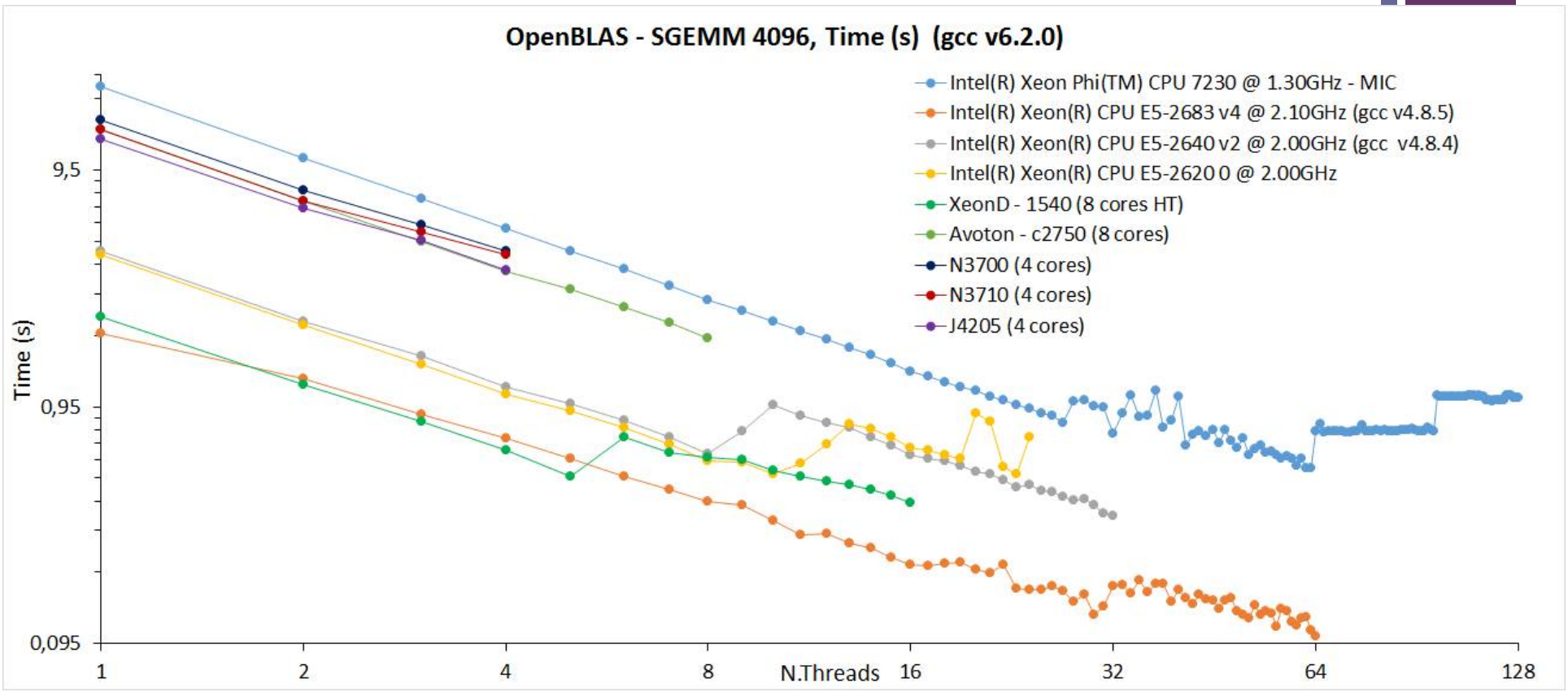


NO explicit use of AVX
NO explicit use of HBW memory

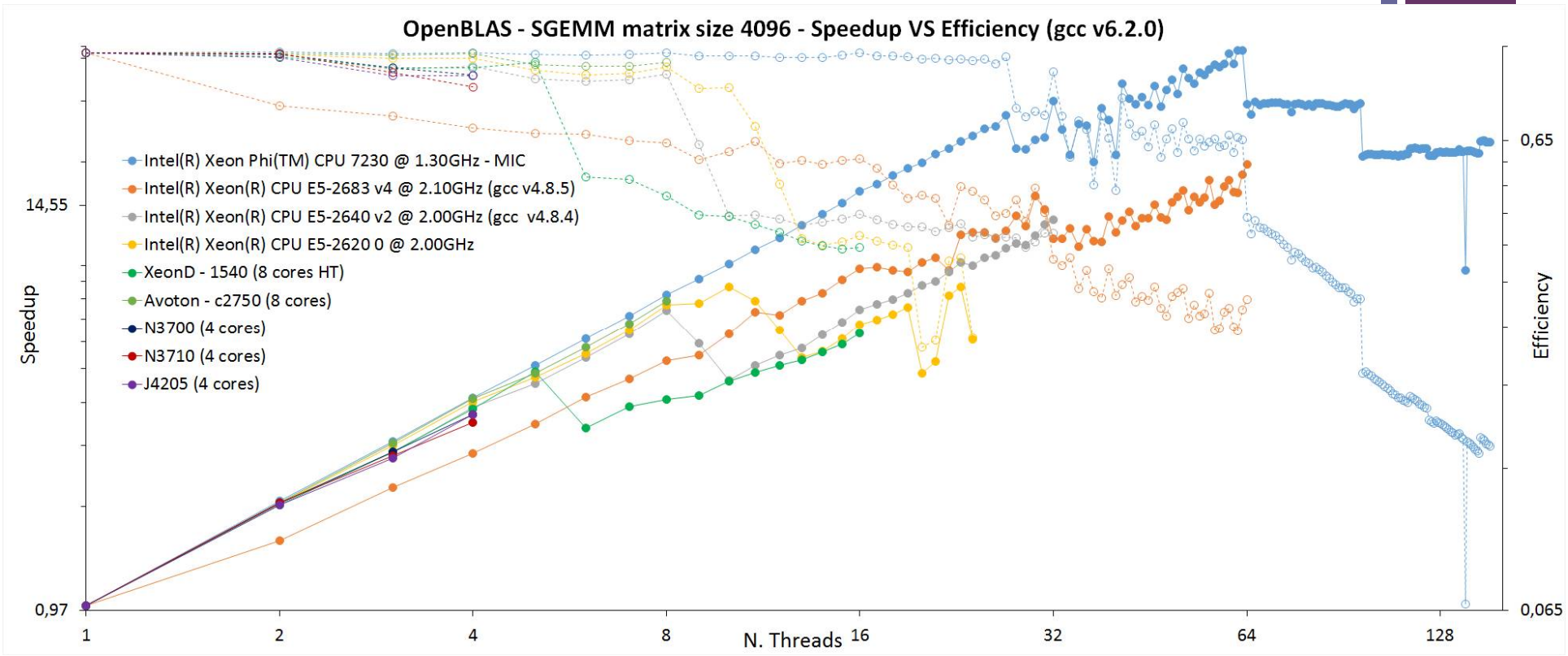
KNL test with synthetic tests using icc



KNL test with synthetic tests - SGEMM

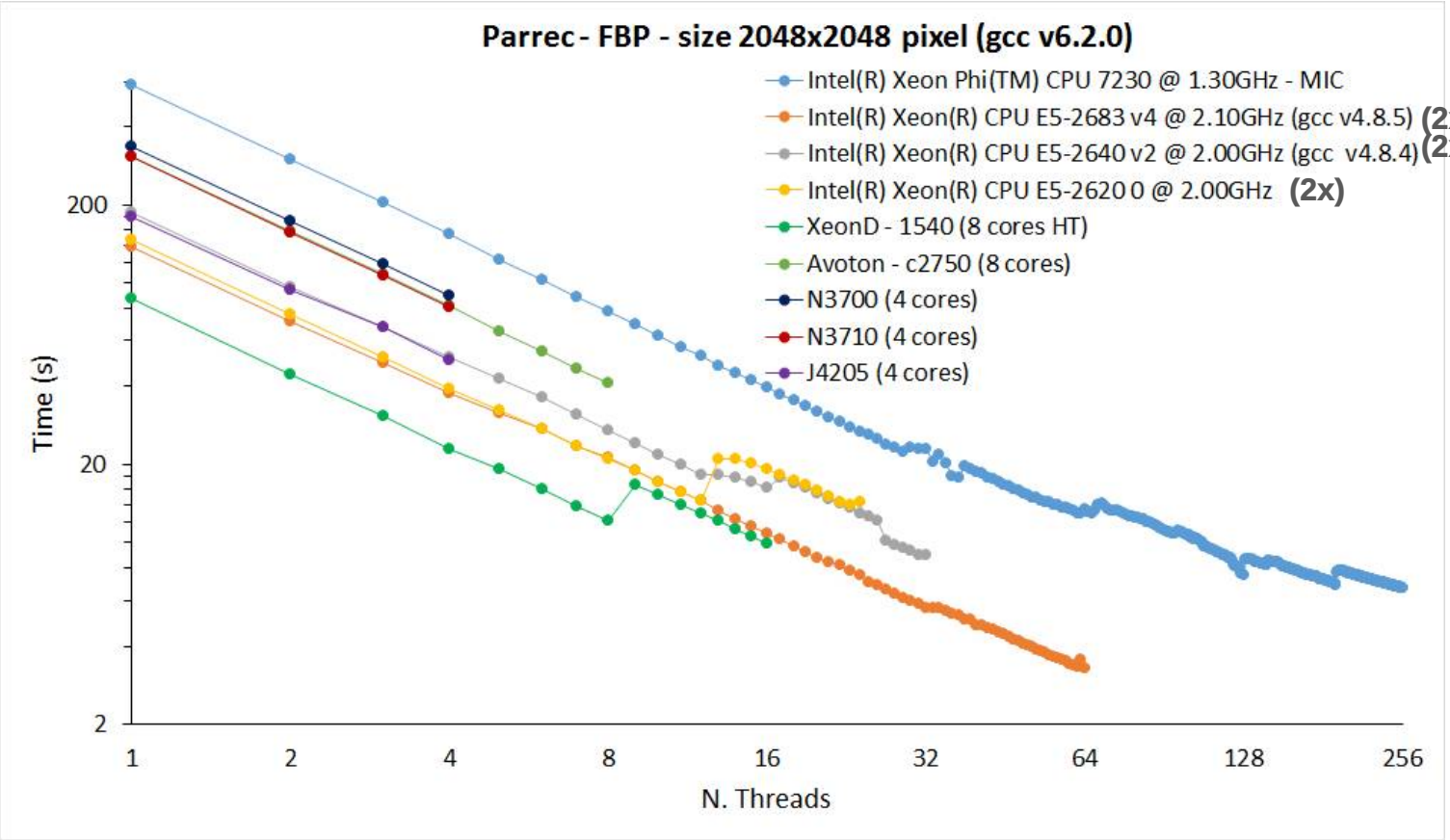


KNL test with synthetic tests - SGEMM



KNL with a real application

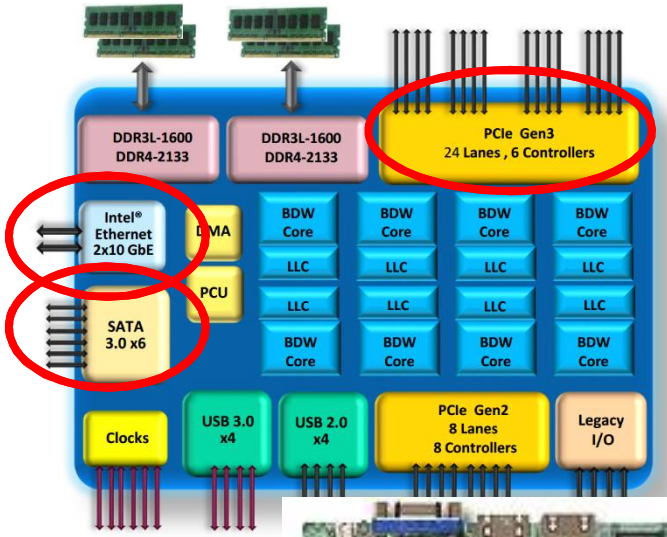
PARREC – FBP algorithm
NO explicit AVX
NO HBW memory



XEON D-1540 for storage bricks

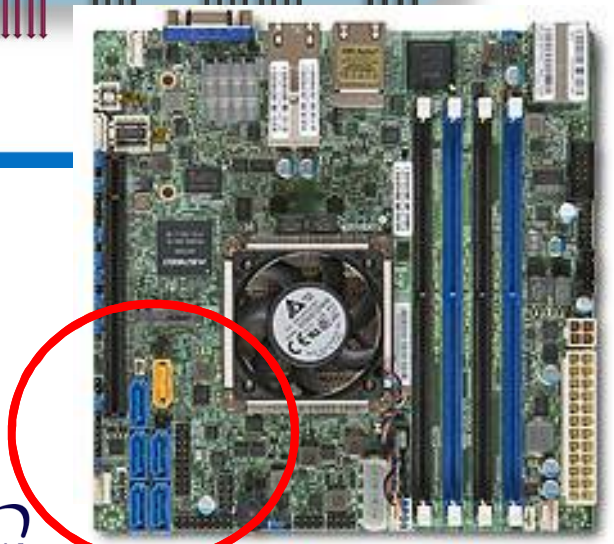
Intel® Xeon® Processor D - SoC Architecture

CPU	2-8 Core Intel® Xeon™ (14nm) CPUs
L1 cache	32K data, 32k instruction per core
L2 cache	256K per core
LLC cache	1.5MB per core
Addressing	46 bits physical / 48 bits virtual
Memory	DDR4 up to 2133 MT/s DDR3L up to 1600 MT/s Two Channels (2 DIMMs/Channel)
Memory Capacity	RDIMM: 128 GB (32 GB/DIMM) UDIMM/SODIMM: 64 GB (16 GB/DIMM)
DIMM Types	SODIMM, UDIMM, RDIMM with ECC and non-ECC
Memory RAS	Enhanced ECC Single bit Error Correction – Dual bit Error Detection (SEC-DED) covers address and data paths, DDR scrambler to reduce error rate.
PCI-E*	x24 PCIe Gen3 with up to 6 controllers x8 PCIe Gen 2 with up to 8 controllers
Integrated IO	Intel® Ethernet 2x10 GbE , x4 USB 3.0, x4 USB 2.0, and x6 SATA 3
Technologies	Intel® VT, Core RAPL, PECI over SMBUS, PSE
Power Management	FIVR, PCPS, EET, UFS Hardware PM
Legacy I/O	SPI for boot flash, SMBus, UART LPC, GPIO, 8259, I/O APIC, 8254 Timer, RTC



Supermicro X10SDV-TLN4F

Performance	
# of Cores	8
# of Threads	16
Processor Base Frequency	2 GHz
Max Turbo Frequency	2.6 GHz
TDP	45 W



Broadwell-DE, Xeon D 8-Core

HDD/SSDs

Conclusion

- COSA is testing two types of SoCs
  - Low-Power SoCs from the mobile/embedded world
    - still have many limitations for a production environment
  - Low Power SoCs from the server world
    - very expensive and in some cases not really low power
    - 10Gbs/Infiniband networking easier to obtain
- SoCs are becoming attractive for real life scientific applications
  - In particular if you manage to extract power from the integrated GPU – much easier on the NVIDIA based platforms
  - CPU porting was easier than expected
- Low-power/low cost dominated by ARM until last year, now INTEL is becoming competitive in this segment
  - No porting required for the CPU
- Advanced HW integration needed to maintain a reasonable size of the system
- KNL not satisfactory up to now
  - Need to test using AVX512 and HBW memory

Special thanks….

- CNAF: A. Ferraro, L. Morganti, E. Corni, F. Giacomini, M. Manzali, A.Falabella
- PR: R. Alfieri, R. De Pietri
- RM1: P. Vicini, A. Lonardo, P.S. Paolucci
- FE: F. Schifano, E. Calore
- PD: M. Michelotto
- PI: T. Boccali