

Computing on Low Power SoC Architecture

Daniele Cesini – INFN-CNAF

Andrea Ferraro – INFN-CNAF

Lucia Morganti – INFN-CNAF

+ Outline

- Modern Low Power Systems on Chip
- Computing on System on Chip
 - ARM CPU
 - SoC GPU
- Low Power from Intel
- Conclusion

+ Low-Power System on Chip (SoCs)

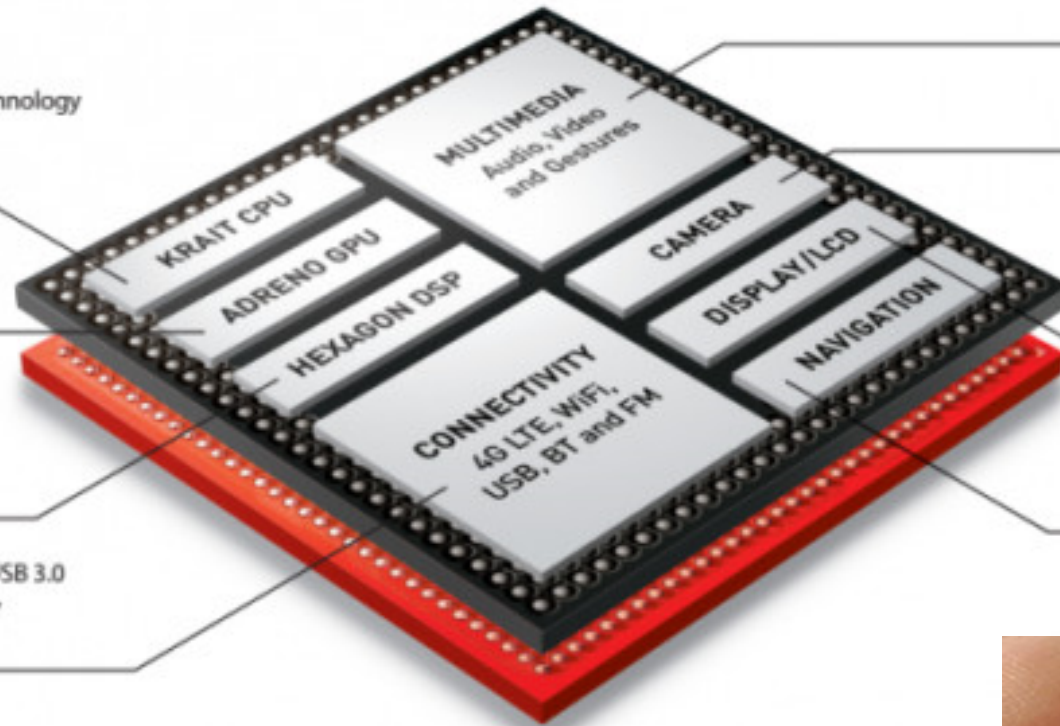
800 PROCESSOR

Krait 400 CPU
features 28HPm process technology
superior
2GHz+ performance

Adreno 330 for
advanced graphics

Hexagon QDSP6
for ultra low power
applications and custom
programmability

Integrated LTE⁺, 802.11ac⁺, USB 3.0
and BT 4.0 offers broad array
of high speed connectivity



Ultra HD Capture
and Playback
DTS-HD and Dolby
Digital Plus audio
Expanded Gestures

55MP with dual ISP

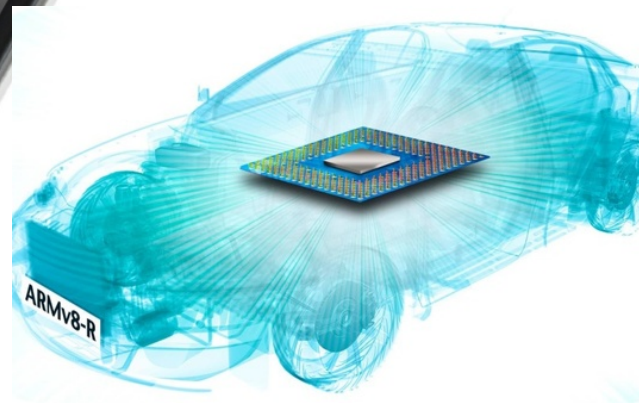
Support for up
to 2560x2048 display
Miracast 1080p
HD support

IZat GNSS with
support for three
GPS constellations

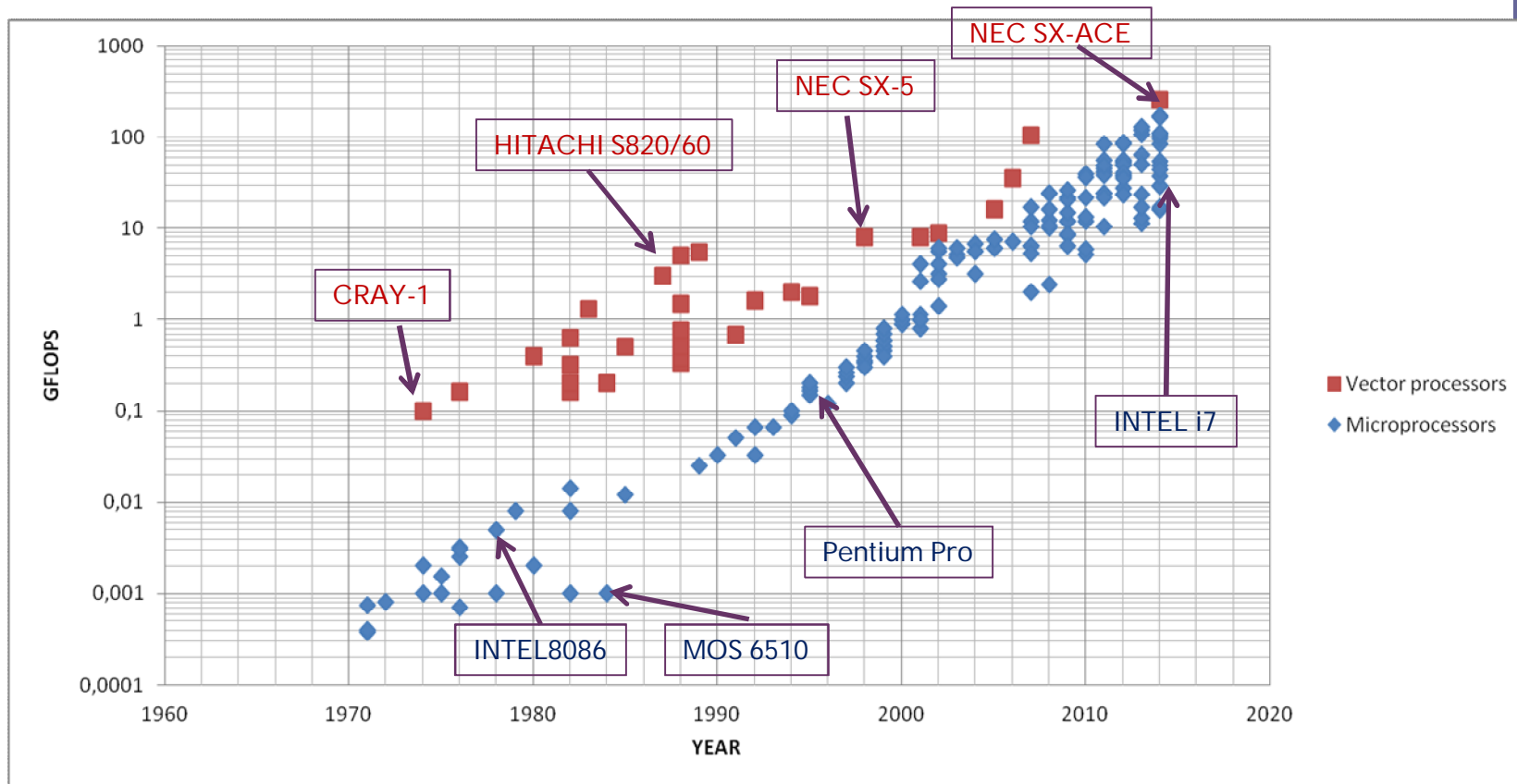


+ Where do I find a SoC?

- Mobile
- Embedded

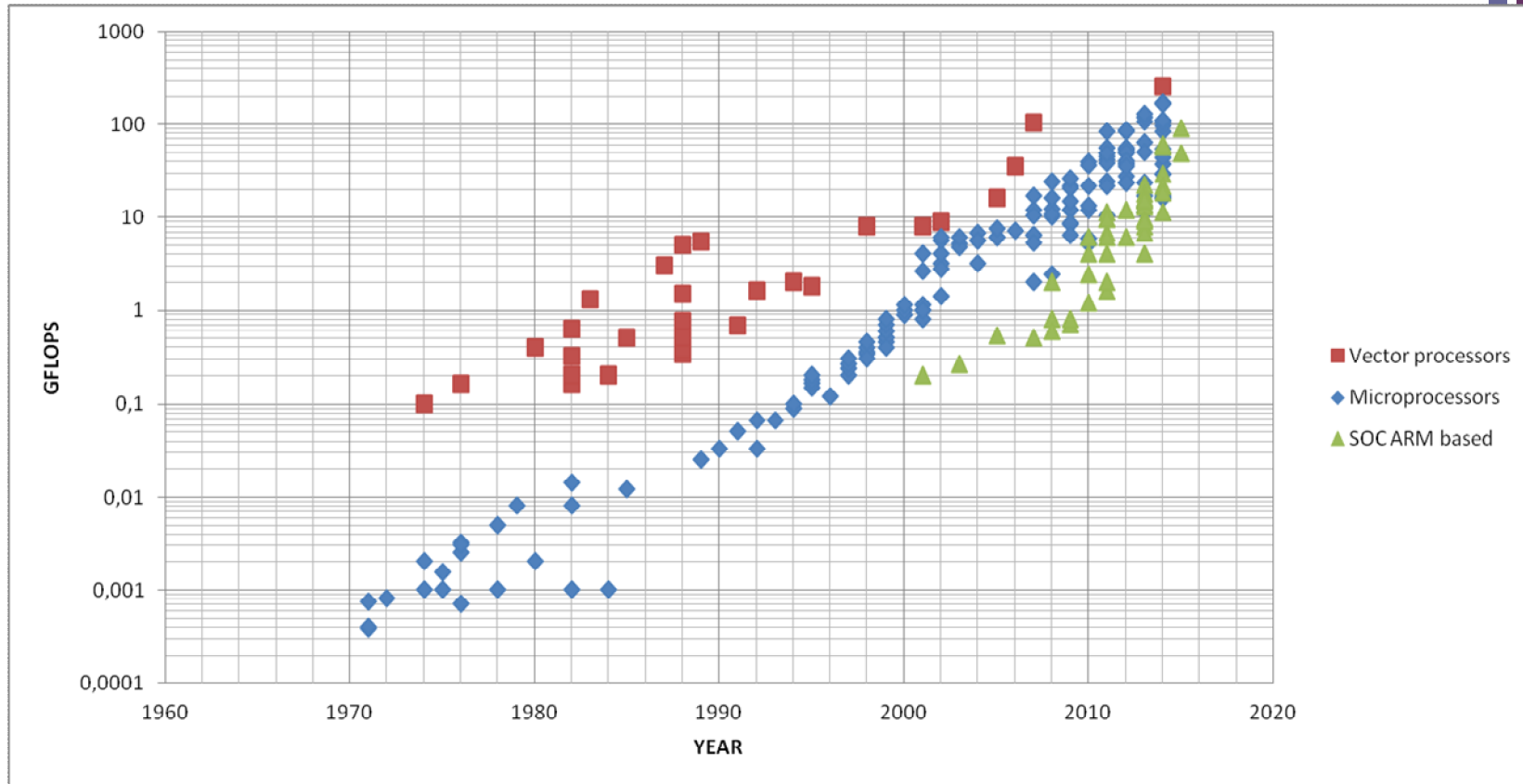


+ Vector vs Micro computing power



- Why did microprocessors take over?
 - They have never been more powerful...
 -but they were cheaper, highly available and less power demanding

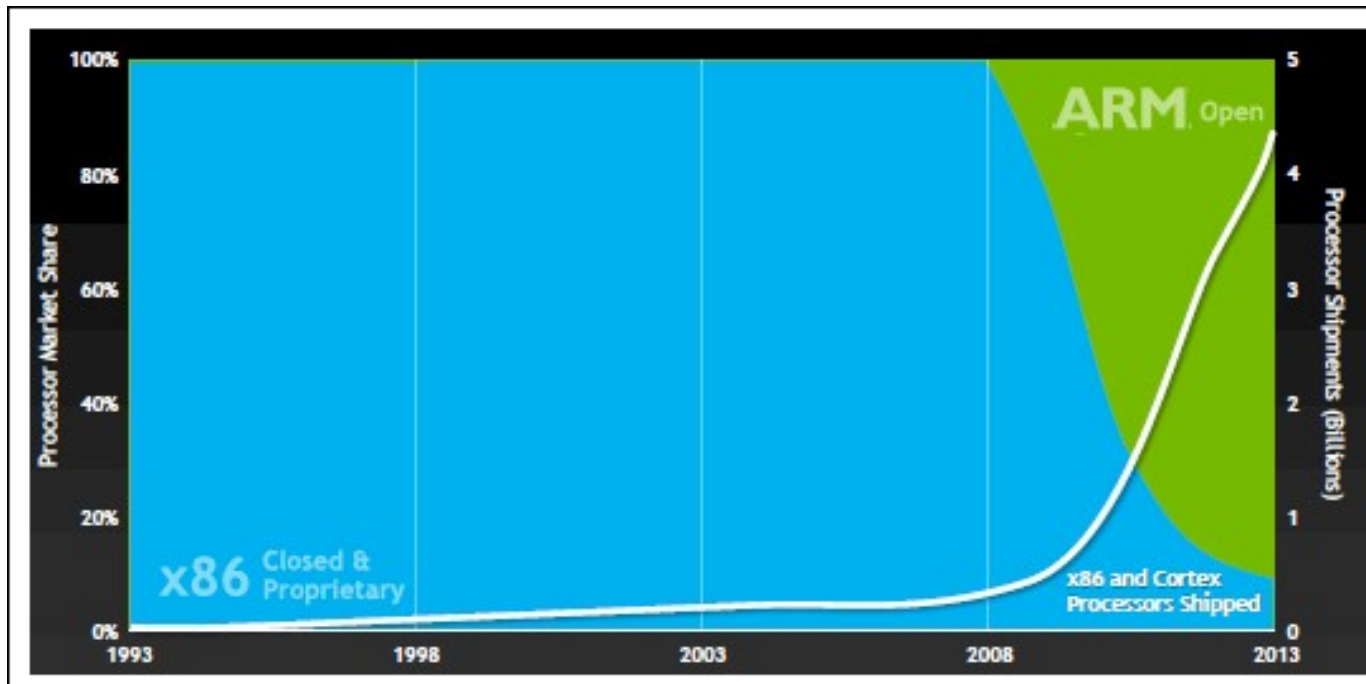
+ Vector vs Micro vs ARM based



■ Is history repeating?

+ ARM based processor shipment

7

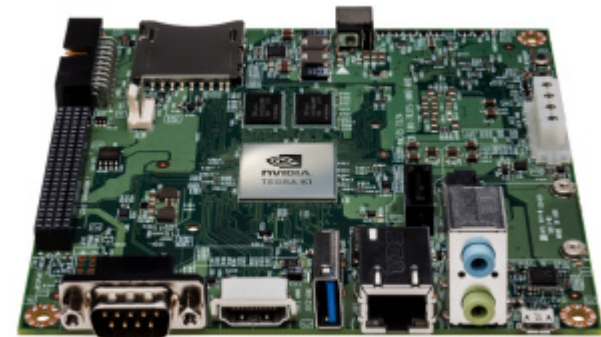


- ARM based processors are shipped in billions of units
 - ARM licences the Intellectual Properties to manufactures
 - ...many manufactures
 - Samsung (Korea), MediaTek (China), Allwinner (China), Qualcomm (USA), NVIDIA (USA), RockChip (China), Freescale (USA), Texas Instruments (USA), HiSilicon(China), Xilinx (USA), Broadcom(USA), Apple(USA), Altera(USA), ST(EU) ,WanderMedia(Taiwan), Marvel(USA), AMD(USA)etc..

+ Ok, but then....an iPhone cluster?

8

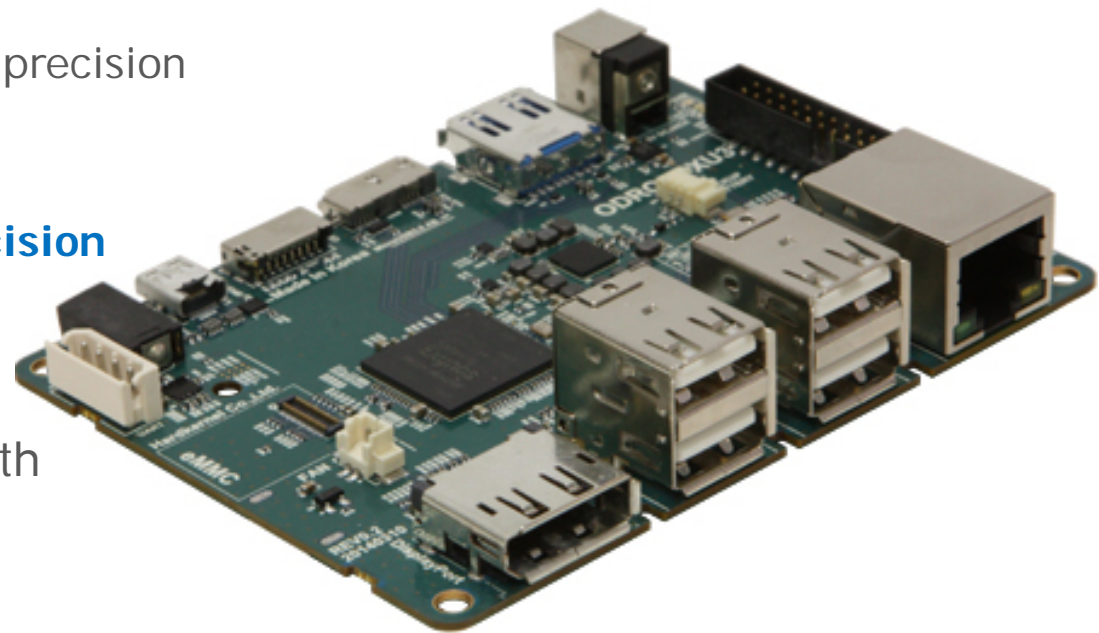
- NO, we are not thinking to build an iPhone cluster
- We want to use these processors in a standard computing center configuration
 - Rack mounted
 - Linux powered
 - Running scientific application mostly in a batch environment
- Use development board...



+ ODROID-XU3

9

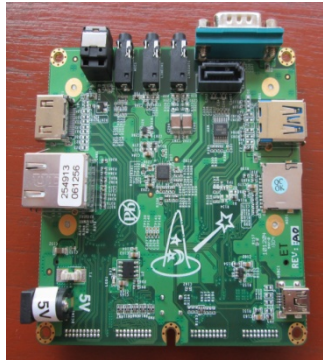
- Powered by ARM® big.LITTLE™ technology, with a **Heterogeneous Multi-Processing (HMP)** solution
 - 4 core ARM A15 + 4 cores ARM A7
- Exynos 5422 by Samsung
 - ~ 20 GFLOPS peak (32bit) single precision
- **Mali-T628 MP6 GPU**
 - ~ **110 GFLOPS peak single precision**
- 2 GB RAM
- 2xUSB3.0, 2xUSB2.0, 1x10/100 eth
- Ubuntu 14.4
- HDMI 1.4 port
- 64 GB flash storage



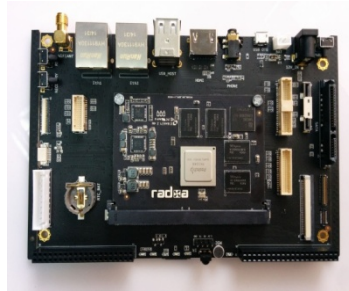
Power consumption max ~ 15 W

Costs 150 euro!

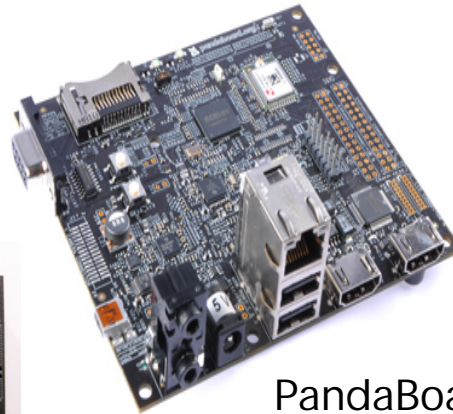
+ Other nice boards



WandBoard



Rock2Board



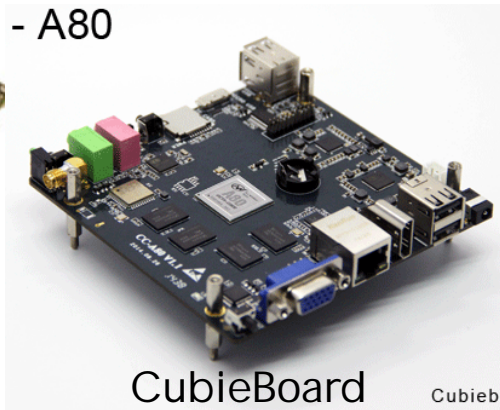
PandaBoard



DragonBoard



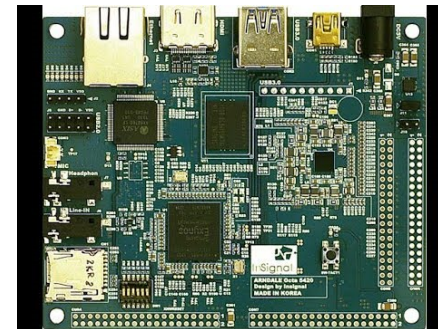
SabreBoard



- A80

CubieBoard

Cubieboard



Arndale OCTA Board



Texas Instruments EVMK2H

http://elinux.org/Development_Platforms

■ ...and counting...

+ Some specs

BOARD	soc				GFLOPS (CPU+GPU)	Eth
	Model	ARM IP	GPU IP	DSP IP		
FREESCALE (Embedded SoC) SABRE Board	Freescall i.MX6Q	ARM A9(4)	Vivante GC2100 (19.2GFlops)		25	1Gb
ARNDALE (Mobile SoC) Octa Board	Samsung Exynos 5420	ARM A15(4) A7(4)	ARM Mali-T628 MP6 (110Gflops)		115	10/100
HARDKERNEL (Mobile SoC) Odroid-XU-E	Samsung Exynos 5410	ARM A15(4) A7(4)	Imagination Technologies PowerVR SGX544MP3 (51.1 Gflops)		65	10/100
HARDKERNEL (Mobile SoC) Odroid-XU3	Samsung Exynos 5422	ARM A15(4) A7(4) (HMP)	ARM Mali-T628 MP6 (110 Gflops)		130	10/100
INTRINSIC (Mobile SoC) DragonBoard	Qualcomm Snapdragon 800	Qualcomm Krait(4)	Qualcomm Adreno 330 (130Gflops)		145	1Gb
TI (Embedded SoC) EVMK2H	TI Keystone 66AK2H14	ARM A15(2)		TI MS320C66x (189Gflops)	210	1Gb (10Gb)

**TDP between 5W and 15W
(EVMK2H > 15W)**


+ NVIDIA JETSON K1

12



TEGRA K1
192-core
Kepler-Class Chip

One Chip – Two Versions

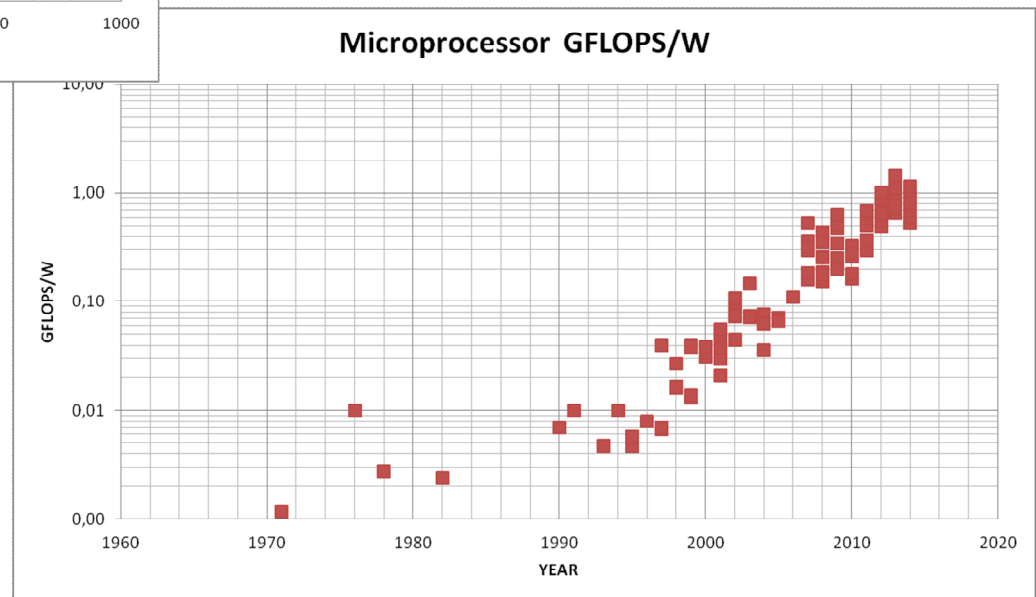
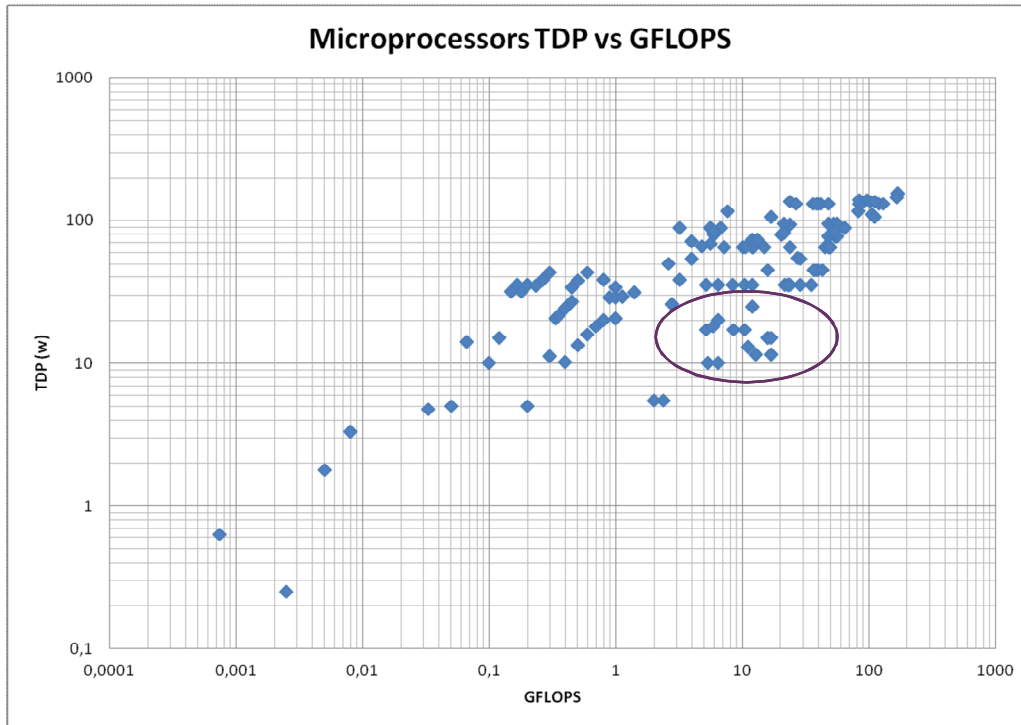


Quad A15 CPUs	Dual Denver CPUs
32-bit	64-bit
3-way Superscalar	7-way Superscalar
Up to 2.3GHz	Up to 2.5GHz
32K+32K L1\$	128K+64K L1\$

NVIDIA

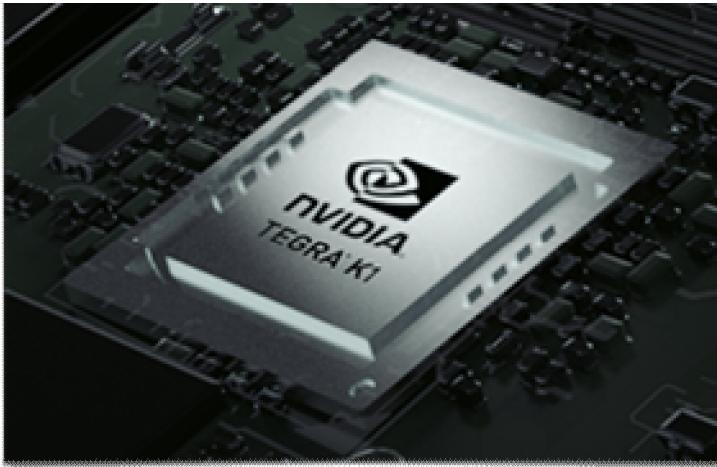
- First **ARM+CUDA programmable** GPU-accelerated Linux development board!
- 4 cores ARM A15 CPU
- 192 cores NVIDIA GPU → 300 GFLOPS (peak sp)
- ...for less than 200 Euros

+ CPU GFLOPS/Watt



+ GPU acceleration in K1

14



4 core ARM A15 ~ 18 GFLOPS
Kepler SMX1 192 core ~ 300 GFLOPS

~ 15 Watt

~ **21 GFLOPS /W**



N.B. Single precision – 32 bit architecture

~ **1.5 GFLOPS /€ (0.67 €/GFLOPS)**

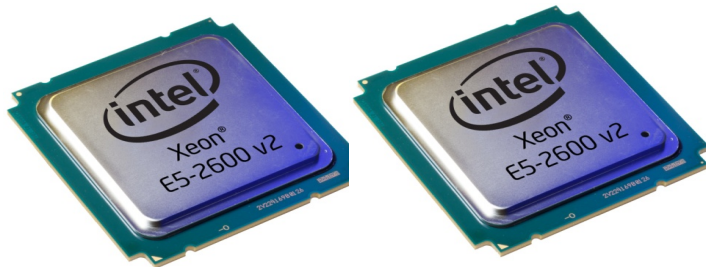
**2xE5-2640+1xK40 ~ 3 GFLOPS/W dp
~ 9 GFLOPS/W sp**

+ How do you program them? (in a Linux environment)

15

- GCC+OpenMP+MPI available for ARM architectures
- OpenCL for the GPU
 - If you are lucky enough to find working drivers
- CUDA available only on the Jetson K1
 - Computing capability 3.2 (vs 3.5)
- Cross compilation
- GCC5+OpenMP4 tests ongoing...

+ GPU acceleration in scientific computation



2 x (E5-2673v2 (IvyBridge) 8 cores)
~ 2 x 100 = 200 GFLOPS (double precision)
2 x 110 Watt = 220 W
~ 1 GFLOPS/W



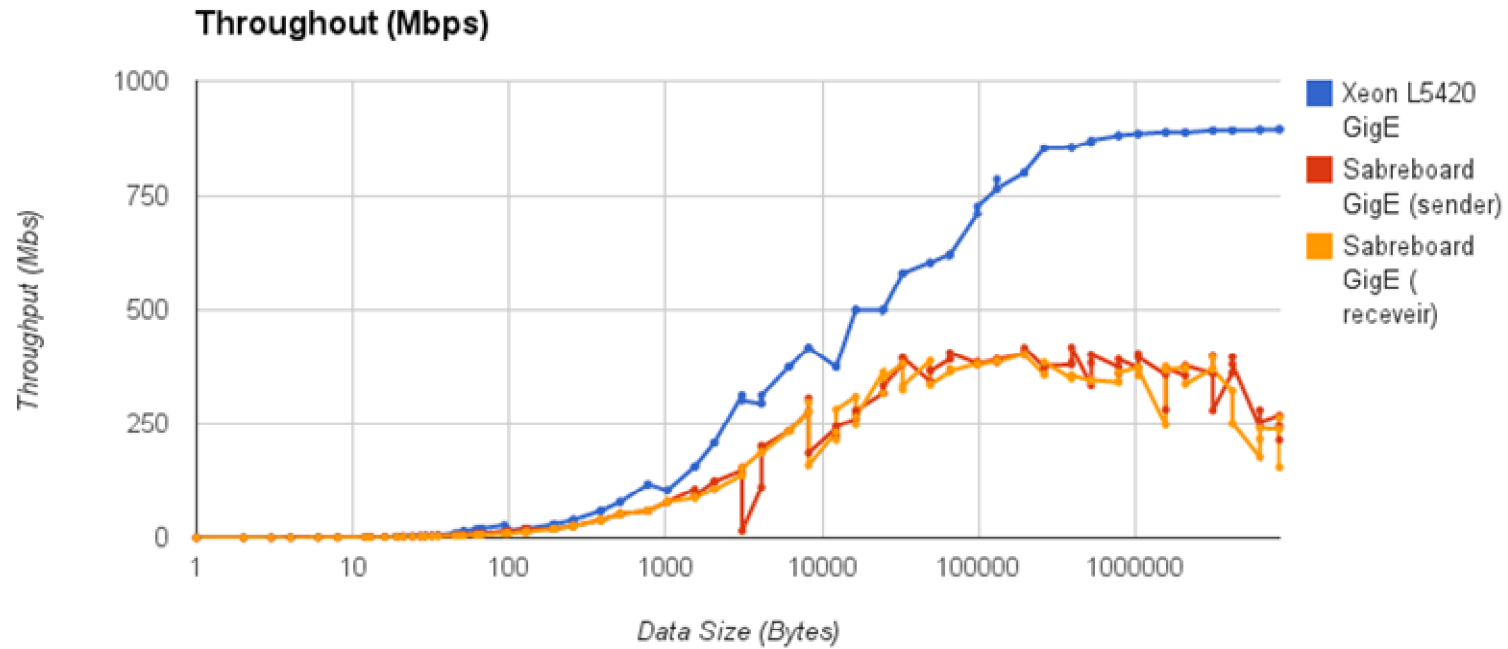
1xNVIDIA TESLA K40
2880 cores
12 GB RAM
~ 1400 GFLOPS (double precision)
~ 4300 GFLOPS (single precision)
235 Watt
~ 6 GFLOPS / W dp
~ 18 GFLOPS/W sp

CPU+GPU ~ 3 GFLOPS/W dp
~ 9 GFLOPS/W sp

+ Limitations

- Commodity SoCs and development boards have a number of limitations:
 - 32 bit
 - Small caches
 - Small RAM size in the boards (O(2GB))
 - However modern SoCs can address 40bit
 - No ECC memory
 - Frequent failures and system crashes
 - Slow connections (10/100Mb eth) in many cases
 - Ethernet via USB in same boards
 - HW bugs

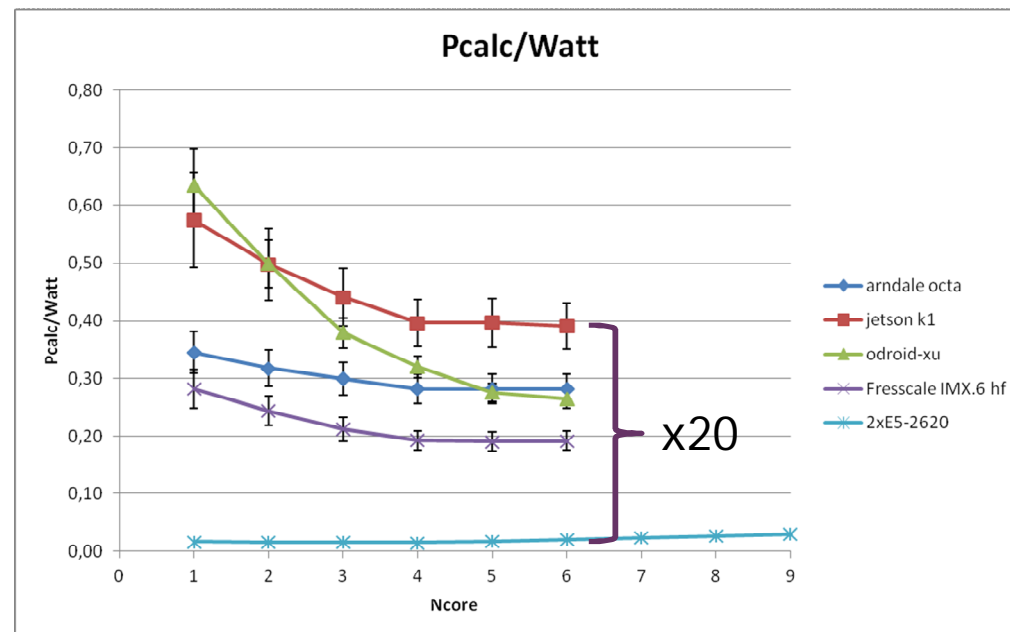
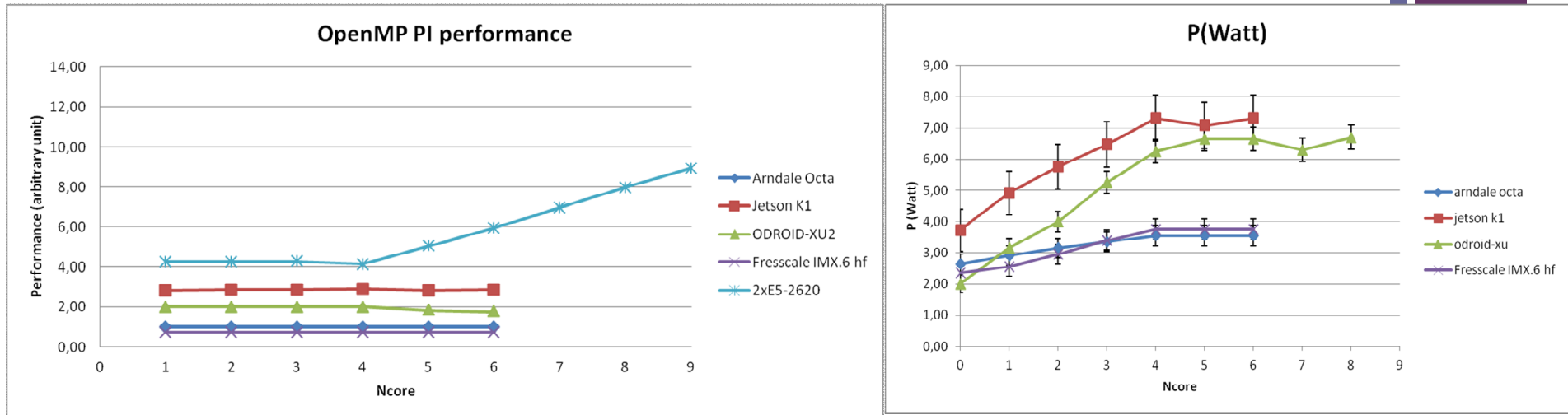
+ Gbit Ethernet



While latency was comparable to a server class 1Gb ethernet card (50/75 us)

+ OpenMP π computation

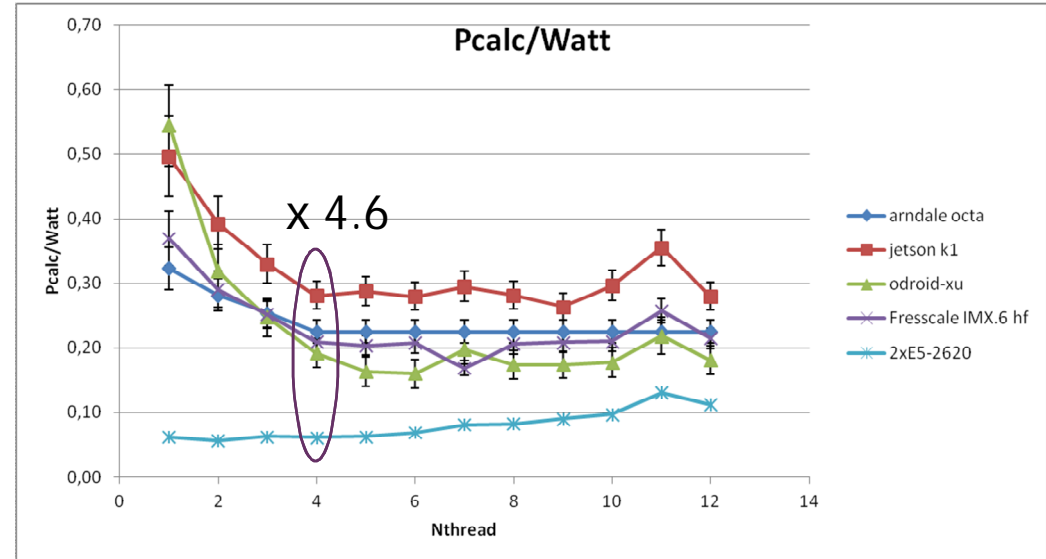
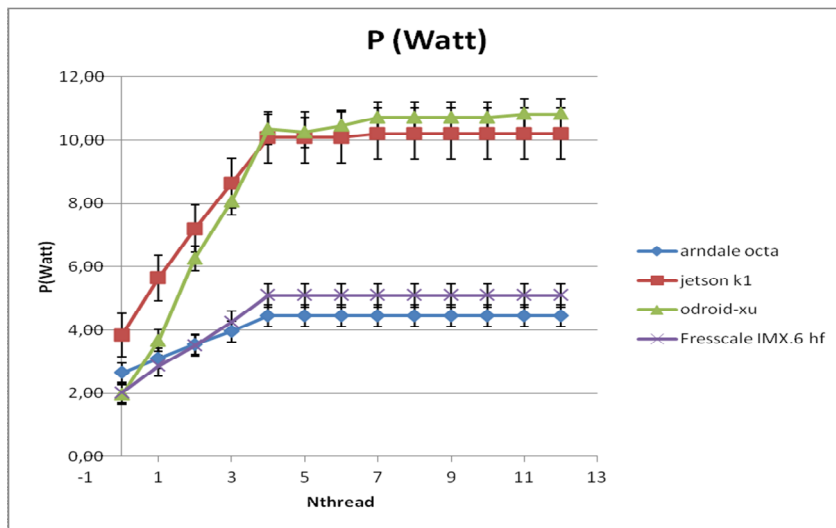
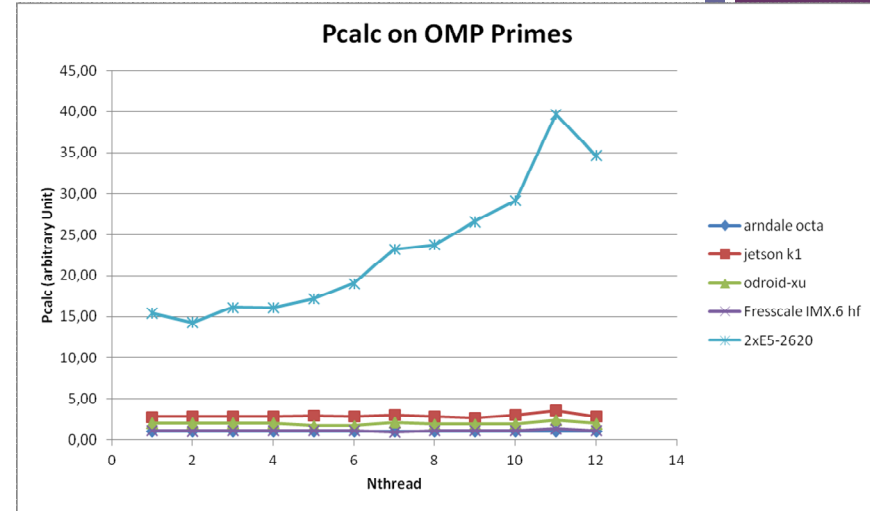
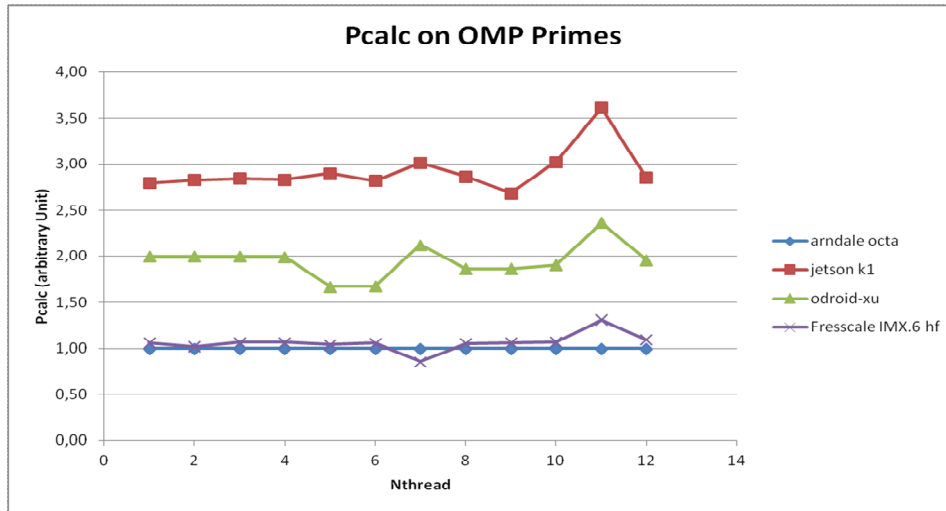
CPU ONLY





Prime numbers computation

CPU ONLY



+ CMS 2014 results

CPU ONLY

High Energy Physics MonteCarlo simulations

Table 1. Results of run time tests for single core CMSSW (GEM-SIM) and a multi-threaded version of the Geant4 benchmark “FullCMS” with 4 threads (G4MT).

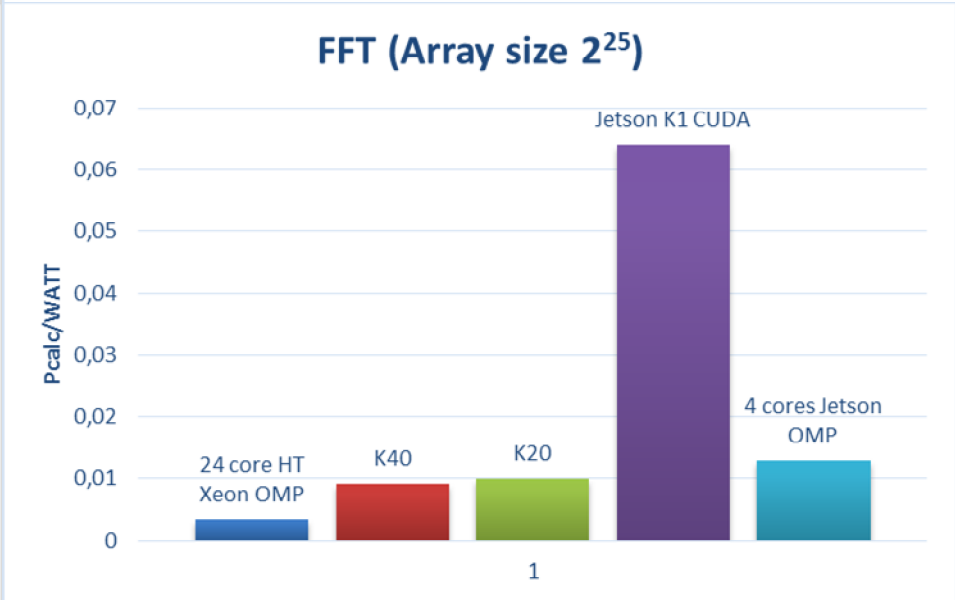
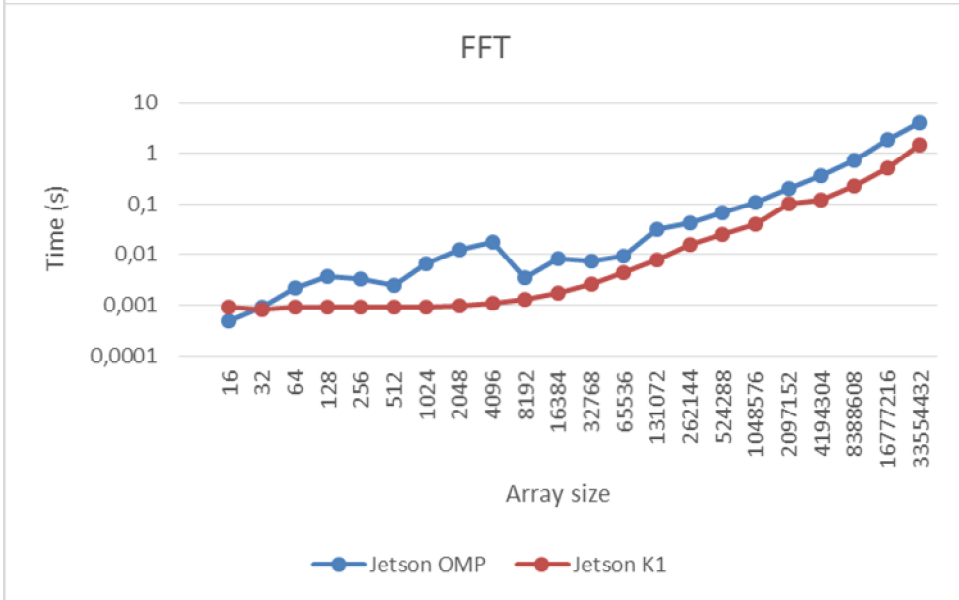
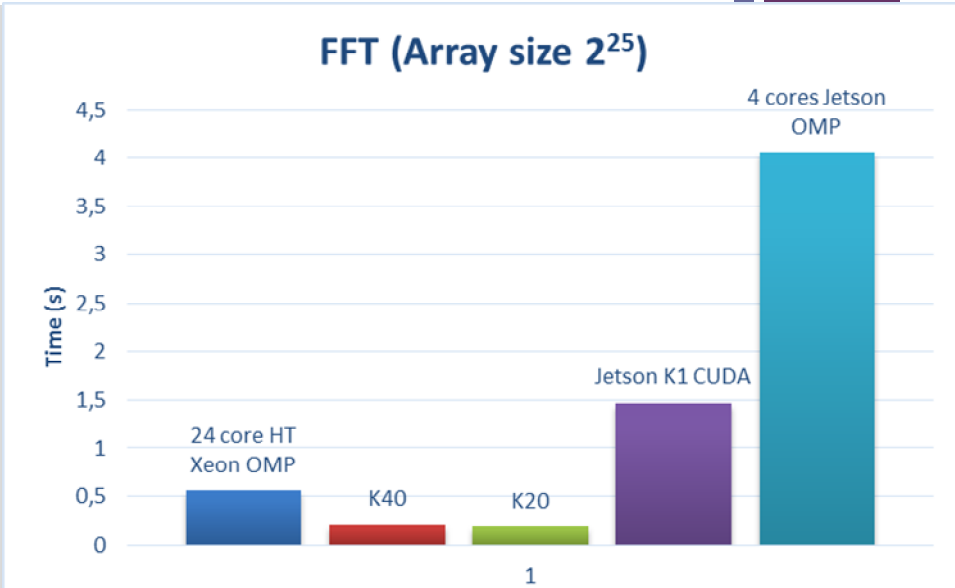
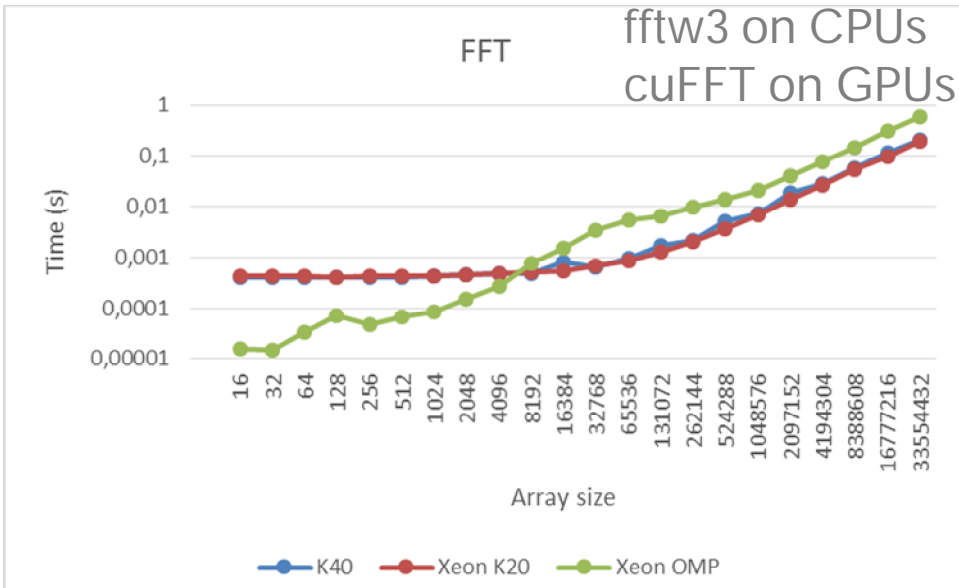
Type	Cores	Power (TDP)	GEN-SIM Events /minute /core	GEN-SIM Events /minute /Watt	G4MT Events /minute (threads)	G4MT Events /minute /Watt
ODROID U2	4	4W	1.08	1.08	34.2 (4)	8.6
ODROID XU+E	4/4	5W ?	1.47	1.07	47 (4)	9.4
dual Xeon L5520 @2.27GHz	2 × 4	120W	3.37	0.22	307.2 (16)	2.6
dual Xeon E5-2630L @2.0GHz	2 × 6	120W	3.46	0.35	N/A	N/A

David Abdurachmanov *et al* 2014 *J. Phys.: Conf. Ser.* **513** 052008 doi:10.1088/1742-6596/513/5/052008

ARM slower by a factor 3 or 4 but...

...ARM better by a factor 3 or 5 on the power ratio

+ FFT on CPU and GPU

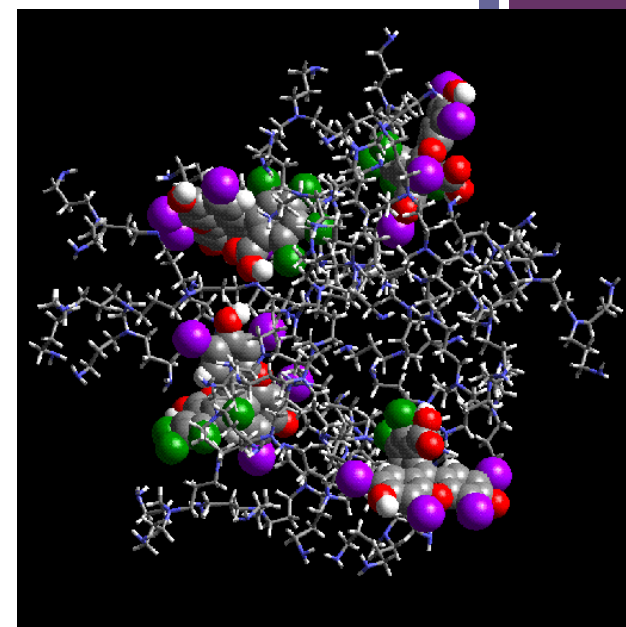
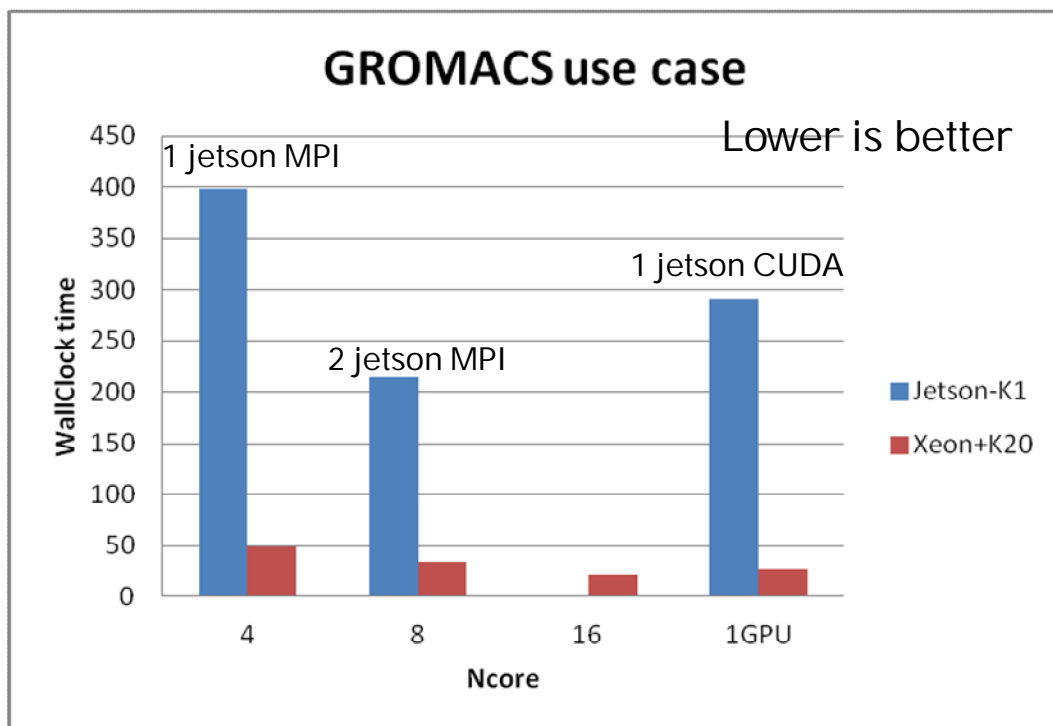


+ Molecular Dynamics on Jetson-K1

23

CPU and GPU

**Parallel application for CPU and GPU:
real life use case with GROMACS**

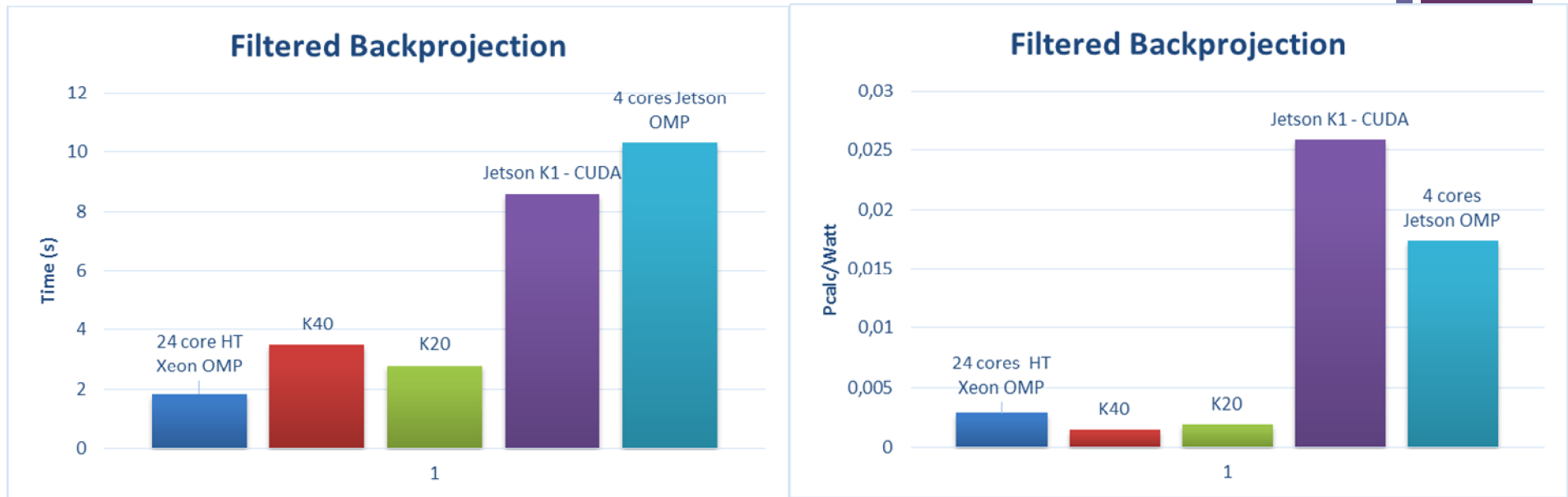


- Jetson-K1 about 10X slower using the same number of CPU cores
- Jetson-K1 about 10X slower using the GPU (vs. an NVIDIA Tesla K20)
 - Jetson-K1 13.5Watt
 - Xeon+K20 ~320Watt

+ Filtered Backprojection

CPU and GPU

24



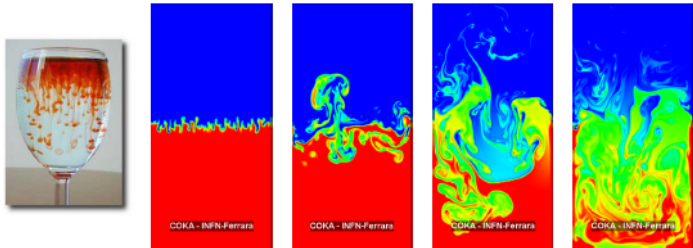
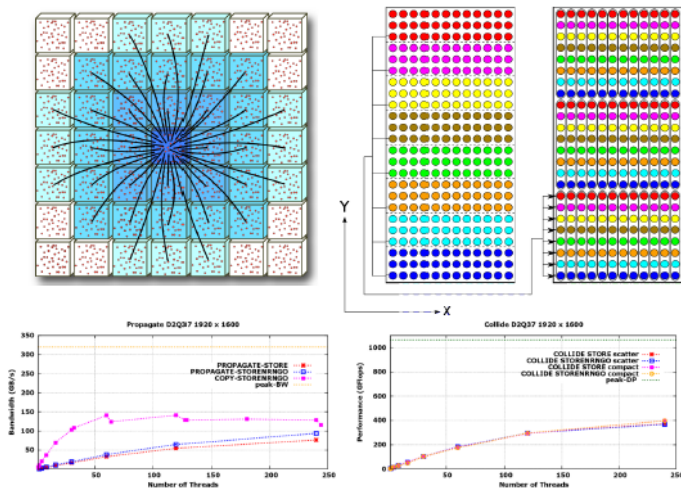
On (2xE5-2620+K20): 1956 images analyzed in 1 hour: 350Wh (GPU not fully loaded)

On 5xJetson-K1: 2095 images analyzed in 1 hour: 41 Wh

+ Lattice Boltzmann on the Tegra K1

GPU only

Lattice Boltzmann Methods: D2Q37



(*) Schifano et al. ; A portable OpenCL Lattice Boltzmann code for multi- And many-core processor architectures;
 Procedia Computer Science Volume 29, 2014, Pages 40-49,
 doi: 10.1016/j.procs.2014.05.004

Daniele Cesini – INFN-CNAF

LBM Performance Comparison (*)

Code Version	Xeon-Phi 7120		Tesla K20Xm			i7-4930K	
	OCL	C ^(*)	OCL	CUDA SM_20	CUDA SM_35	OCL	C ^(*)
propagate T/iter [msec]	30.46	37.67	14.89	15.40	15.38	186.42	162.00
GB/s	76.42	61.8	156.33	151.16	151.36	12.48	14.54
\mathcal{E}_p	22%	17%	62%	60%	60%	21%	24%
bc T/iter [msec]	3.20	4.61	7.08	5.68	5.70	4.30	4.87
collide T/iter [msec]	72.79	79.14	93.27	83.33	43.06	440.18	307.42
GFLOPS (DP)	410	377	320	358	680	68	97
MLUPS	54.02	49.69	42.16	47.19	89.44	8.93	12.94
\mathcal{E}_c	34%	31%	24%	27%	52%	42%	59%
$\mu J / \text{site}$	5.55	6.03	5.57	4.98	2.63	14.55	10.04
T_{WC}/iter [msec]	106.45	121.42	115.24	104.42	65.03	630.90	489.98
MLUPS	36.94	32.38	34.12	37.65	60.46	6.23	8.12

On Tegra-K1
(preliminary)

15 GFLOPS
12 GB/s
 $P_e \sim 10 \text{ Watt}$

40x slower than
a K20m

- Porting easier than expected
- Performance under investigation



GDB - 11/02/2015

+ Only ARM based SoCs? And Intel?

- INTEL produce SoCs
 - Probably you have one in your laptop
- Some of them are low power
- Already 64bit
- Integrated GPU
 - CILK++ programmable
 - OpenCL programmable



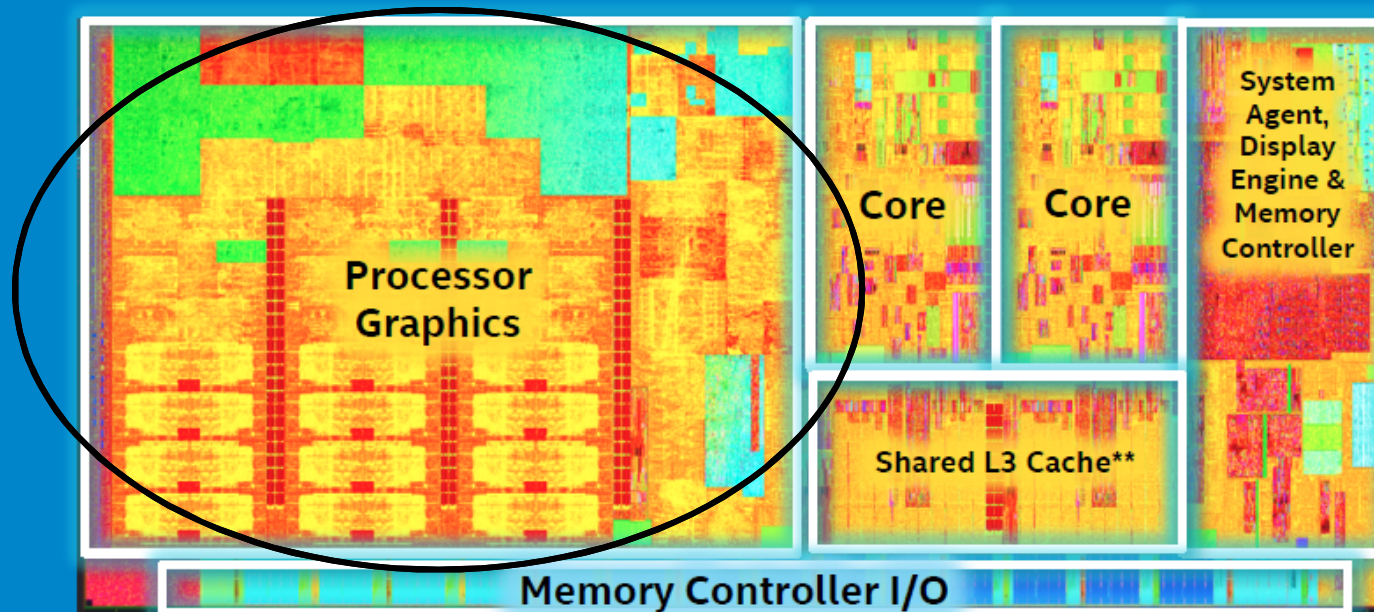


Some low power from Intel

Nome prodotto	Intel® Atom™ Processor E3845 (2M Cache, 1.91 GHz)	Intel® Core™ M-5Y71 Processor (4M Cache, up to 2.90 GHz)	Intel® Core™ i7-4578U Processor (4M Cache, up to 3.50 GHz)	Intel® Core™ i7-4702EC Processor (8M Cache, up to 2.00 GHz)	Intel® Atom™ Processor C2730 (4M Cache, 1.70 GHz)
Nome in codice	Bay Trail	Broadwell	Haswell	Haswell	Avoton
Informazioni di base					
Stato	Launched	Launched	Launched	Launched	Launched
Data di lancio	Q4'13	Q4'14	Q3'14	Q1'14	Q3'13
Numero di processore	E3845	5Y71	i7-4578U	i7-4702EC	C2730
Cache	2 MB L2 Cache	4 MB	4 MB	8 MB Intel® Smart Cache	4 MB
Set di istruzioni	64-bit	64-bit	64-bit	64-bit	64-bit
Opzioni integrate disponibili	Yes	No	No	Yes	No
Litografia	22 nm	14 nm	22 nm	22 nm	22 nm
Prezzo consigliato per il cliente	TRAY: \$52.00	TRAY: \$281.00	TRAY: \$426.00	TRAY: \$459.00	TRAY: \$150.00
Datasheet	Link	Link		Link	Link
Privi di minerali provenienti da zone di conflitto	Yes	Yes	Yes	Yes	
Estensioni set di istruzioni		AVX, SSE	SSE 4.1/4.2, AVX 2.0	SSE 4.1/4.2, AVX 2.0	
Tipo di Bus			DMI2	DMI	
Bus di sistema			5 GT/s	5 GT/s	
Prestazioni					
Numero di core	4	2	2	4	8
Numero di thread	4	4	4	8	8
Frequenza base del processore	1.91 GHz	1.2 GHz	3 GHz	2 GHz	1.7 GHz
TDP	10 W	4,5 W	28 W	27 W	12 W
Specifiche della grafica					
Grafica del processore †	Intel® HD Graphics	Intel® HD Graphics 5300	Intel® Iris™ Graphics 5100		
Frequenza di base grafica	542 MHz	300 MHz	200 MHz		
Frequenza di burst della grafica	792 MHz				
Intel® Quick Sync Video	Yes	Yes	Yes		
Numero massimo di schermi supportati ‡	2	3	3		
Frequenza dinamica massima grafica		900 MHz	1.2 GHz		

Intel® Core™ M Processor Die Map

14nm 2nd Generation Tri-Gate 3-D Transistors



Dual Core Die Shown Above

Transistor Count: 1.3 Billion

Die Size: 82mm²

4th Gen Core Processor (Y series): .96B

4th Gen Core Processor (Y series): 131mm²

** Cache is shared across both cores and processor graphics

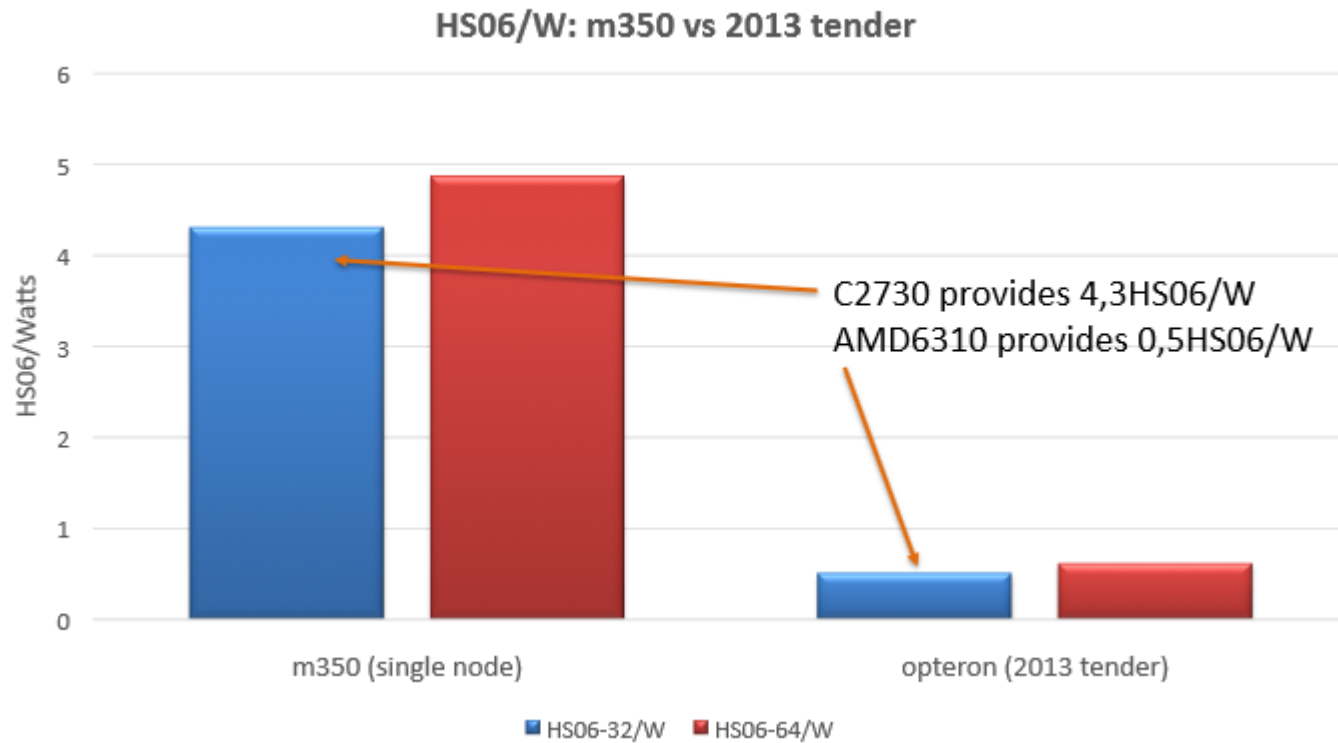
- 2 cores + GPU
 - Intel HD Graphics
 - OpenCL 2.0 Support
- 4.5 Watt (TDP)

(*) <http://www.notebookcheck.net/Intel-Core-M-5Y70-Broadwell-Review.130930.0.html>

(**)

<http://www.intel.com/content/www/us/en/processors/core/core-m-processor-family-spec-update.html>

+ AVOTON on HP Moonshot - HS06

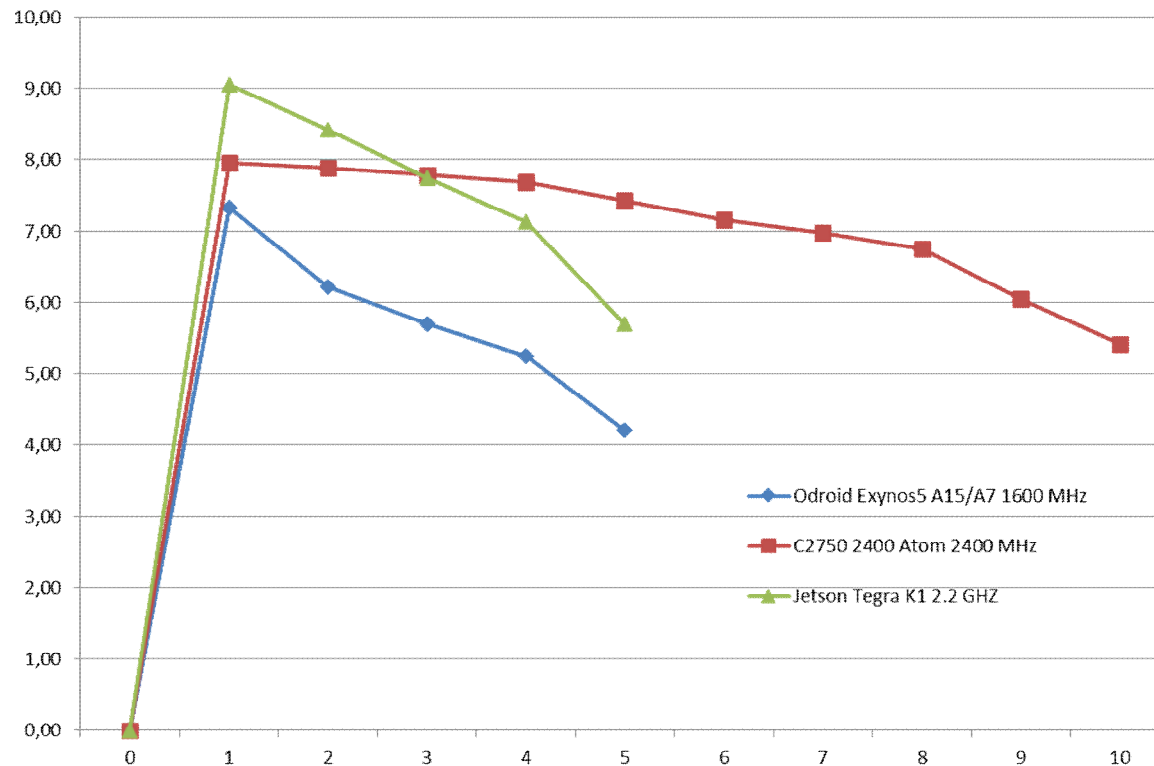


(data from A.Chierici@HEPIX

<https://indico.cern.ch/event/305362/session/2/contribution/22/material/slides/0.pdf>)

+ HS06

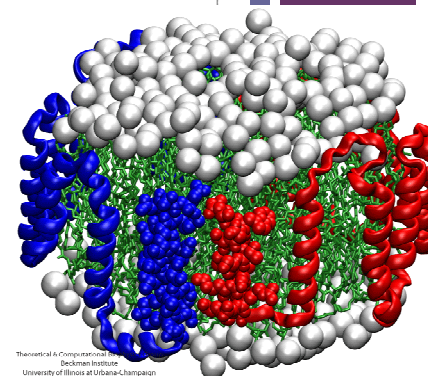
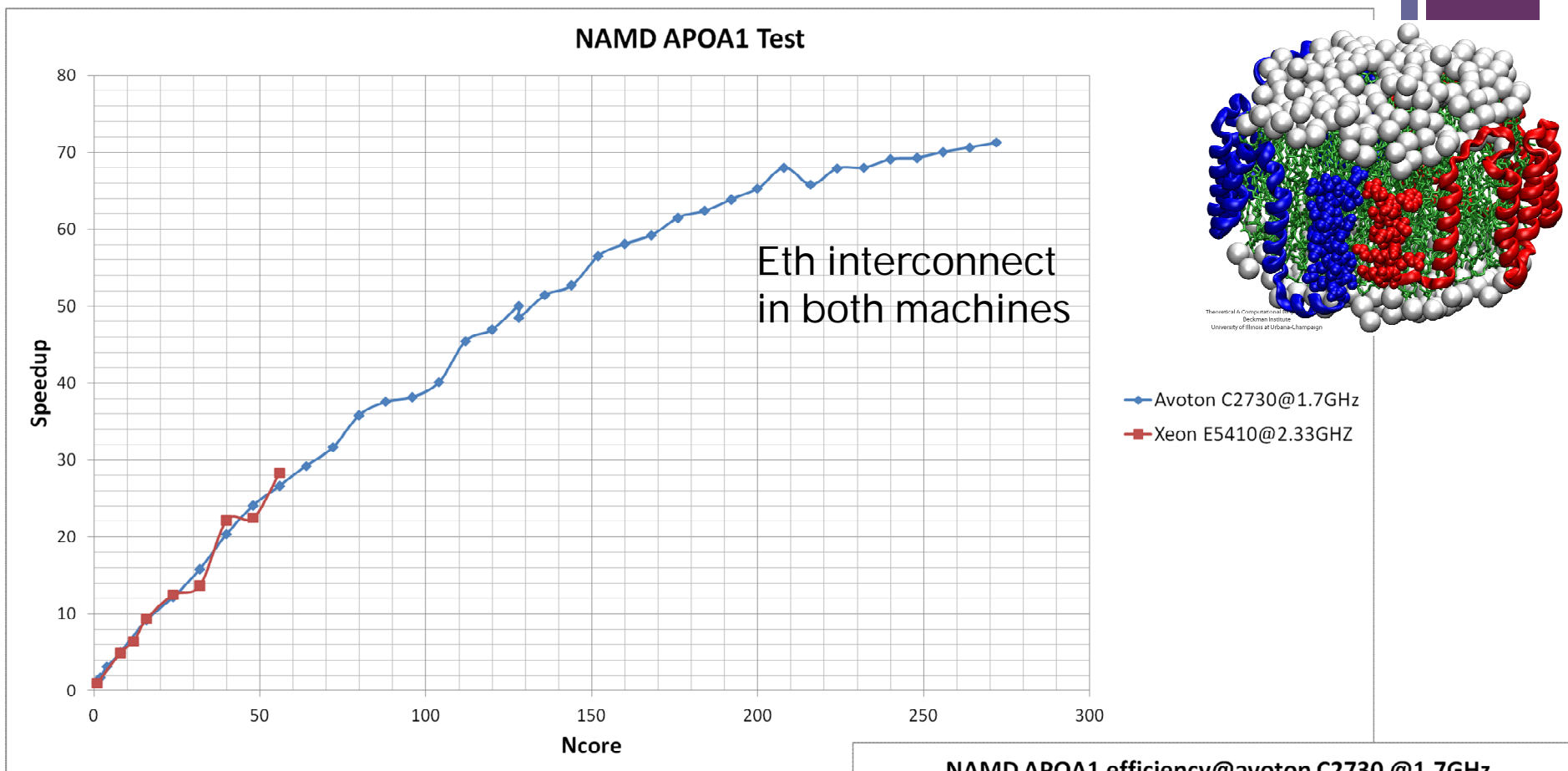
HS06 on Exynos5, TegraK1 and Atom C2750 – Per core loaded



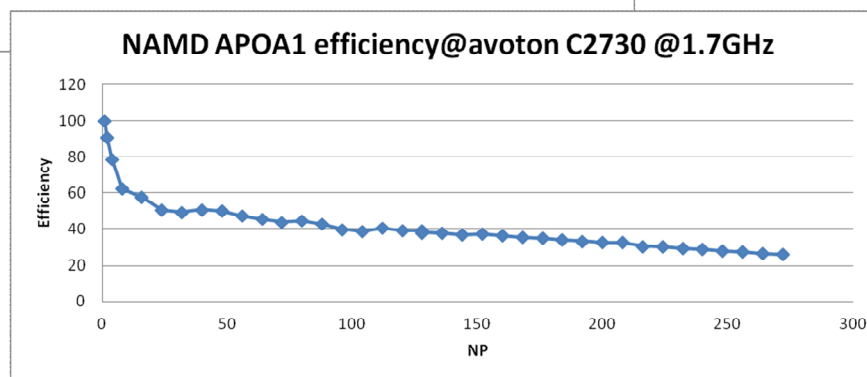
(data from M.Michelotto@HEPIX 2014

<https://indico.cern.ch/event/320819/session/3/contribution/30/material/slides/0.pptx>)

+ Test on Intel AVOTON



N.B. – Comparison with an old Penryn Xeon CPU



+ Conclusion

32

- Mobile and embedded low power System-on-Chip are becoming attractive for scientific computing
 - In particular if you manage to extract power from the GPU
 - For selected applications
 - Image processing
 - No high RAM/RAM bandwidth requirements
- They still have many limitations for a production environment
 - 32bit, no ECC, bugs, system stability, etc..
 - (BUT we used development boards - not server grade machines)
- NVIDIA K1 in our tests was the most powerful ARM based SoC
 - Easy to install and use
 - Easy to port CUDA based applications
- Intel has interesting low power SoCs
 - Avoton has a high HS06/W ratio
- Looking forward to development boards based on 64bit SoCs with an ARM CPU on board

+ Links and contacts

33

- <http://www.cosa-project.it>
- <http://montblanc-project.eu>
- <https://indico.cern.ch/event/320819/session/3/contribution/30/material/slides/0.pptx>
- <https://indico.cern.ch/event/305362/session/2/contribution/22/material/slides/0.pdf>