The path toward ExaScale: motivations, technologies, open issues and proposed solutions

Piero Vicini (INFN Rome)

XIII Seminar on Software for Nuclear, Subnuclear and Applied Physics

Alghero (SS)

June 7th, 2016

HPC definition

High Performance Computing i.e. Supercomputer
From Wikipedia page (https://en.wikipedia.org/wiki/Supercomputer):

- A supercomputer is a computer with a high-level computational capacity compared to a general-purpose computer.
- Introduced in 1960 (Cray): from a few computing nodes to current *MPP* Mssively Parallel Processors with 10^4 "off-the-shelf" nodes
- It's motivated by the search for solutions of *grand challenges* computing applications in many research fields (see PRACE scientific case for Europe HPC...)
  - quantum mechanics, weather forecasting, climate research, oil and gas exploration, molecular modeling, physical simulations...

HPC i.e. parallel computing system

High Performance Computing i.e. Supercomputer
From Wikipedia page (https://en.wikipedia.org/wiki/Supercomputer):

- Performance of a supercomputer is measured in floating-point operations per second (FLOPS) instead of million instructions per second (MIPS).
  - T(era)Flops (10^{12}), P(eta)Flops (10^{15}), ExaFlops (10^{18}), Z(etta)Flops (10^{21})
  - Today $n * 10$PFlops vs single workstation less than 1 TFlops...
- *Parallelism* is the key implemented with different approaches:
  - hundreds or thousands of discrete computers (e.g., laptops) distributed across a network (e.g., the Internet) devote some or all of their time to solving a common problem;
  - huge number of dedicated processors are placed in proximity and tightly connected to each other working in a coordinated way on a single task and saving time to move data around.

Performance...

In the past, scaling to high(er) performances was an "easy" game...

ENIAC 1943:
first programmable "mainframe"
18000 tubes == 5000 transistors

DOWNSIZING AND UPGRADING

The inception of computing inspired a remarkable race for faster, smaller, lighter, cheaper hardware

	ENIAC	Intel Core Duo chip
Debut	1946	2006
Performance	5,000 addition problems/sec	21.6 billion ops/sec
Power use	170,000 watts	31 watts max
Weight	28 tons	negligible
Size	80' w x 8' h	90.3 sq. mm
What's inside	17,840 vacuum tubes	151.6 M transistors
Cost	$487,000	$637

ENIAC vs current CPU

Performance...

In the past, scaling to high(er) performances was an "easy" game...

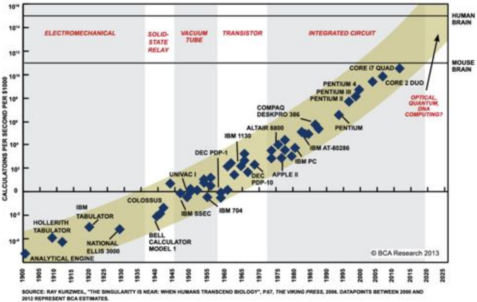
ENIAC 1943:
first programmable "mainframe"
18000 tubes == 5000 transistors

DOWNSIZING AND UPGRADING

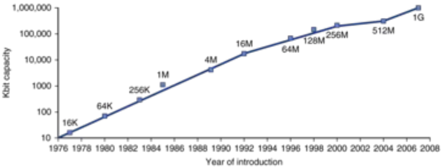
The inception of computing inspired a remarkable race for faster, smaller, lighter, cheaper hardware

	ENIAC	Intel Core Duo chip
Debut	1946	2006
Performance	5,000 addition problems/sec	21.6 billion ops/sec
Power use	170,000 watts	31 watts max
Weight	28 tons	negligible
Size	80' w x 8' h	90.3 sq. mm
What's inside	17,840 vacuum tubes	151.6 M transistors
Cost	$487,000	$637

ENIAC vs current CPU

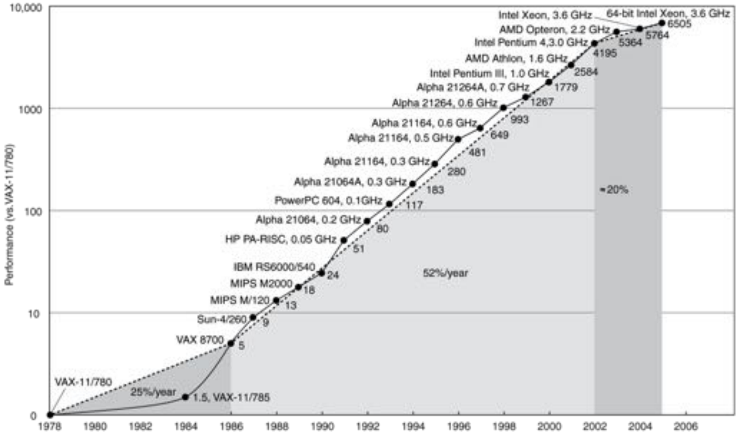


Moore's law: transistor density (i.e. computer performance) doubles every 18-24 (!) months



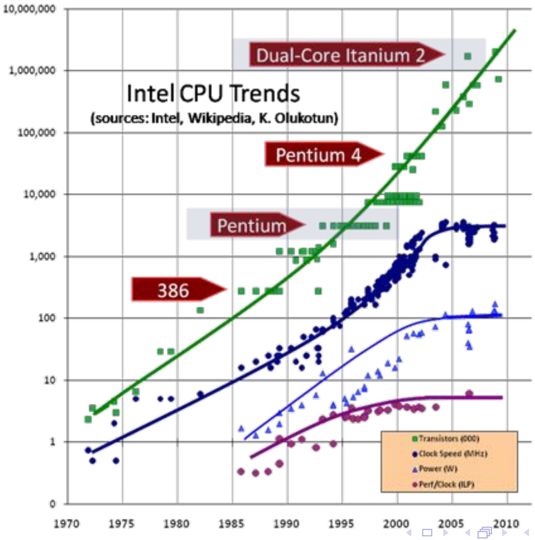
Memory density

- Once upon a time, and thanks to the Moore's law, performance scaled with the processor clock frequency...



"Free lunch is over..."

- From mid of 2000's it's no more true....



"Free lunch is over...": the Power Wall

The Power Wall

- CMOS technology: current through junctions flows ONLY when (and during) transistor changes state

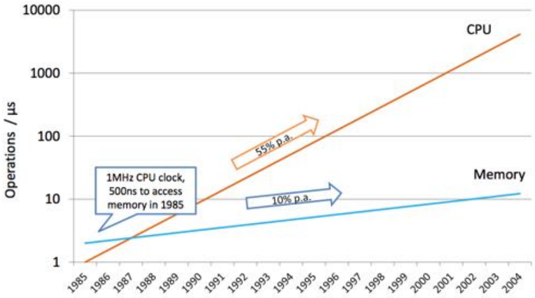
$$P = CxV^2x(\alpha f)$$

C = capacitance, V = voltage, f the switching frequency and α the fraction of gates switching per unit of time

- It exists a technological limit to surface power density. As a consequence:
  - processor clock frequency can not scale up freely...
  - supply voltage can not decrease too much (impact of leakage and errors due to fluctuations...)

The Memory Wall

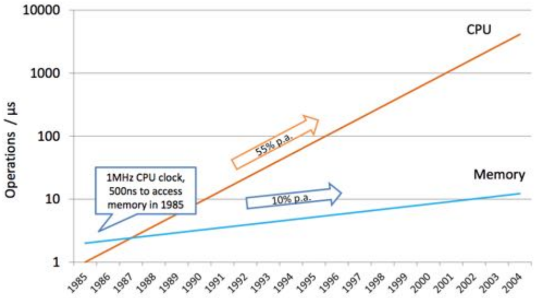
- The GAP between CPU performance and memory devices bandwidth is growing
- The majority of application workloads are *memory limited* so the memory access is a real bottleneck



"Free lunch is over...": *Memory Wall* e *ILP Wall*

The Memory Wall

- The GAP between CPU performance and memory devices bandwidth is growing
- The majority of application workloads are *memory limited* so the memory access is a real bottleneck



ILP Wall

- Instruction-level parallelism (ILP) is a measure of how many of the operations in a computer program can be performed simultaneously.
- The implicit parallelism in a single computing thread of a processor is quite limited.
  - Try to reorder instructions, reduce to sequence of micro-instructions, aggressive branch predictive but...
  - ... you can't feed the computing units if you are waiting for data memory
  - Additionally, adding functional units to exploit ILP parallelism increase HW complexity —> increase the power dissipation (Power Wall)

Power wall + ILP wall + Memory wall → Serial hardware Game over...

Game over for serial/monolithic hardware.

Power wall + ILP wall + Memory wall → Serial hardware Game over...

- Use concurrency as much as you can → parallel architectures: *multiprocessors*, *multi*-core, *many*-core
  - multi/many computing cores with "low" clock frequency
  - many multi/many cores processors interconnected by efficient networks
  - new programming model able to cope with parallel systems and able to distribute the workload in parallel
- Warning: effective parallel programming (performance next to the theoretical peak) is a BIG issue...
  (luckily not fully covered in this talk ;-)

Amdahl's law gives the theoretical *speedup* of the execution of a task at fixed workload that can be expected of a system whose resources are improved.

- *If P is the fraction of a computer program that can be parallelized on N computing nodes ($1 - P$ is the non-parallizable part) , the execution time, $T(N)$, is:*
$$T(N) = T(1)(\frac{P}{N} + (1 - P))$$
- so the speedup is $S(N) = \frac{T(1)}{T(N)}$ is equal to
$$S(N) = \frac{1}{(1-P)+\frac{P}{N}}$$

Amdahl's law gives the theoretical *speedup* of the execution of a task at fixed workload that can be expected of a system whose resources are improved.

- *If P is the fraction of a computer program that can be parallelized on N computing nodes ($1 - P$ is the non-parallizable part) , the execution time, $T(N)$, is:*

$$T(N) = T(1)(\frac{P}{N} + (1 - P))$$

- so the speedup is $S(N) = \frac{T(1)}{T(N)}$ is equal to

$$S(N) = \frac{1}{(1-P)+\frac{P}{N}}$$

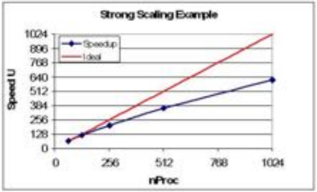
- Es: compute P to get 90% of speedup if the number of processing units of the computing system increases from 1 to 100.

$$90 = \frac{1}{(1-P)+\frac{P}{100}} \rightarrow P = 0.999$$

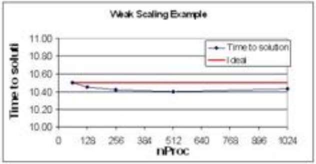
- The sequential part (not-parallizable) of program is to be less than 0.1% (!!!)

- Scalability Scalability of a system respect to an application measures how well latency and bandwidth scale with the addition of more processors.

- Strong scaling: given a fixed size computing problem the *strong scaling* represents its time to solution as a function of (increasing) number of execution processors.

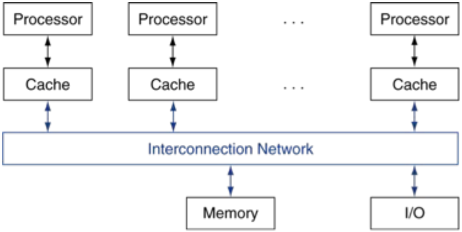


- Weak scaling: *weak scaling* is the time to solution of a fixed size *per processor* problem varying the number of processors.



- System scalability is affected from *load balancing*
  - If the average computational load of a single processor is 2x, speedup may be halved...

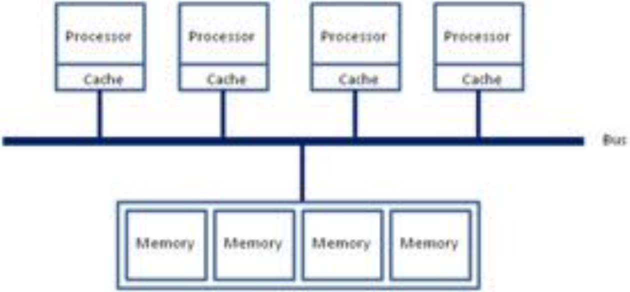
- Parallel hardware architecture: all processors share *a single memory space*
- Parallel execution (coordination) and data exchange through *shared variables* located in shared memory.
  - Synchronisation primitives (*locks*, *barriers*) to handle memory access contention.



- Two different types of memory access:
- UMA: Uniform Memory Access vs NUMA Non-Uniform Memory Access

UMA vs NUMA

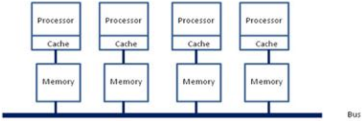
- UMA: i.e *Symmetric Multi-Processing* architecture
- Any processor (core) can access any memory location with the same access time / latency (!!!)



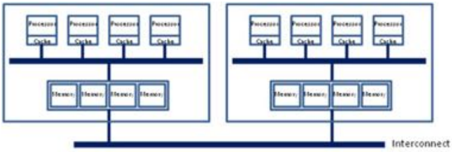
- SMP is effective and easy to program but scaling is limited to few processors... (Multi-Core)
- It's really complex (almost unfeasible?) to ensure "uniform access" when hundreds of CPU compete to access data in memory...

UMA vs NUMA

- NUMA
- Every processor (core) owns its (private) *local* memory and can access in parallel *remote* memory of others processors (cores), using high performance (hopefully, low latency) inteconnection network.



- Programming NUMA may be more difficult than programming UMA
- Very good scaling for "hybrid" architectures like NUMA SMP processors
  - New generation multi-proc of multi-core, AMD Hypertransport, INTEL QPI,...



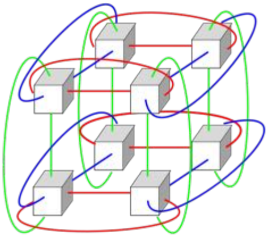
- MP: methods for NUMA *Distributed Memory* architectures
  - Each processor owns its private memory space and data is shared through explicit *data send* and *data receive* calls.
  - Message passing primitives for system coordination (*message send* , *message receive*).
- Pros
  - easy to implement
  - Low level message passing can also be implemented in hardware using shared memory....
  - *MPI* (Message Passing Interface) de facto standard: dominant model used in HPC today.

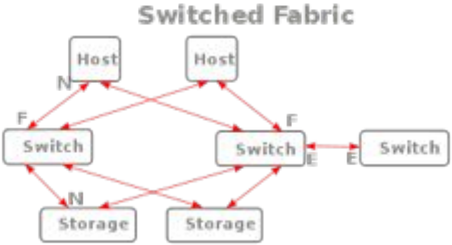
Issues for scalability

- *Access latency* to remote data memory.
  - MP introduces time overhead typically 10-100 uS equivalent of 10^5 - 10^6 FP operations...
- Communication *bandwidth* has to be large enough for feeding processors
  - first order evaluation: 3 64 bit words per operation –> $10^{11} Flops * 24 Bytes = 2 TB/s$
  - to be multiplied for the number of nodes......
  - many technics (algorithmic and technological) to mitigate the effects but not enough
- porting of sequential applications may be not easy since every data movement is explicit and source/target has to be identified

Network Topology is how the processors are connected (Direct point-to-point and switched).

- 2D or 3D *Torus mesh*: is simple and ideal for programs with mostly nearest-neighbour communication.
- *Hypercube*: minimizes number of "hops" between processors, many wires/channels
- *Switched network*: all processors connected through hyerarchical layers of high-speed switches. Overhead but quite fast especially for limited number of computing nodes

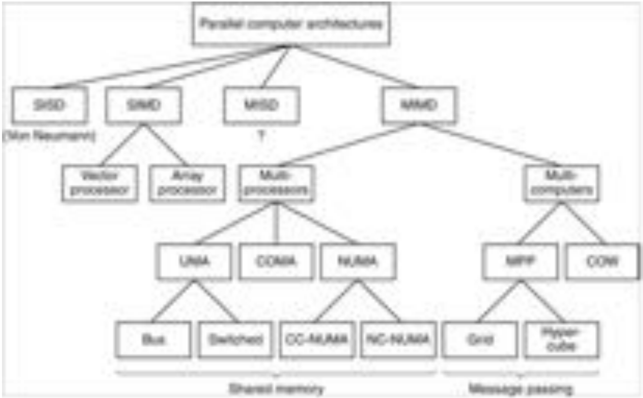




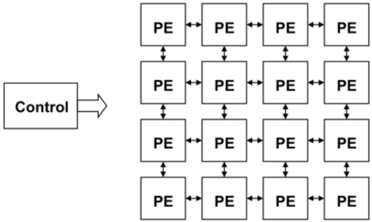
- Performance
  - Latency per message (unloaded network)
  - Throughput
    - Link bandwidth
    - Total network bandwidth
    - Bisection bandwidth
  - Congestion delays (depending on traffic)
- Cost
- Power
- Routability in silicon

- *Cluster* is a network of independent computer systems communicating through message passing protocol. Sometimes are called *Beowulf*
- Each computer has its own private memory space and, in principle, different OS
- Commodity network to move data
- Scalable by default, cost effective, robust and fault tolerant, cloud-izable...
- Most of TOP500 ranking list are clusters with main differences due to:
  - CPUs (x86 multi-core, Power,)
  - FP accelerators (GPGPU, MIC, FPGA,...)
  - network technology (ethernet, Infiniband, Myrinet,...)
  - network topology (fat-tree, torus,...)

- Flynn's Taxonomy provides a simple, but very broad, classification of architectures for high-performance computers:



- SISD Single Instruction Single Data: Von Neumann architecture.
- MISD Multiple Instruction Single Data: ???????
- MIMD Multiple Instruction Multiple Data streams: multiprocessor
- SIMD Single Instruction Multiple Data streams: vector processor.



- A single control unit and multiple cheap, simple datapaths (PE) executing in parallel
- Usually PEs are interconnected via mesh o torus network
- PEs execute *same* program on their *local* dataset

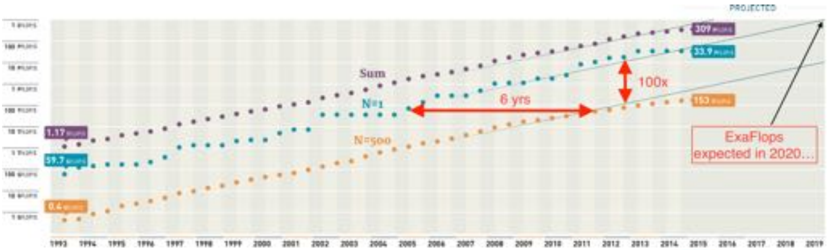
- It's a good way to scale to huge number of processing units and MPP (Massively Parallel System)
- Control and synchronization is easy...
- Effective is the application exposes high level of *data parallelism*: only half of HPC applications...
  - Es. special set of instructions MMX and SSEn in x86 archs
  - GPUs and many-core architectures
  - Vector processors
  - INFN APE systems dedicated to LQCD...

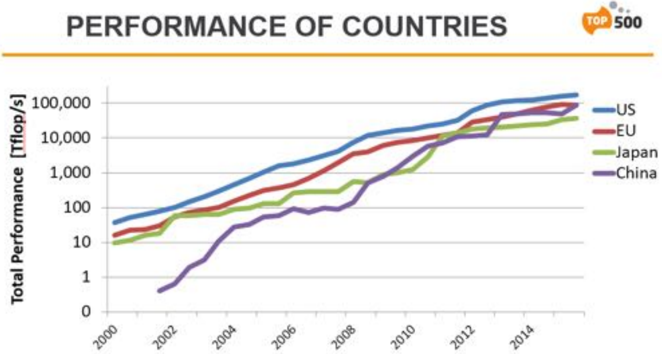
HPC measure and compare: the TOP500 list

Web site: *www.top500.org*

- Based on a common benchmark: LinPack, a package for linear algebra
- *www.top500.org/resources/posters − and − materials/* for a bit of history (1993-...)

RANK	SITE	SYSTEM	CORES	RMAX (TFLOP/S)	RPEAK (TFLOP/S)	POWER (KW)
1	National Super Computer Center in Guangzhou China	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P NUDT	3,120,000	33,862.7	54,902.4	17,808
2	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560,640	17,590.0	27,112.5	8,209
3	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1,572,864	17,173.2	20,132.7	7,890
4	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705,024	10,510.0	11,280.4	12,660
5	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786,432	8,586.6	10,066.3	3,945
6	DOE/NNSA/LANL/SNL United States	**Trinity** - Cray XC40, Xeon E5-2698v3 16C 2.3GHz, Aries interconnect Cray Inc.	301,056	8,100.9	11,078.9	


PROJECTED
Sum
N=1
N=500
6 yrs
100x
1.17
59.7
0.4
309
33.9
153
ExaFlops expected in 2020...
1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019


PERFORMANCE OF COUNTRIES
TOP 500
Total Performance [Tflop/s]
100,000
10,000
1,000
100
10
1
0
2000
2002
2004
2006
2008
2010
2012
2014
US
EU
Japan
China

TOP500 Web site: *www.green500.org*

The Green500 List

Listed below are the June 2014 The Green500's energy-efficient supercomputers ranked from 1 to 10.

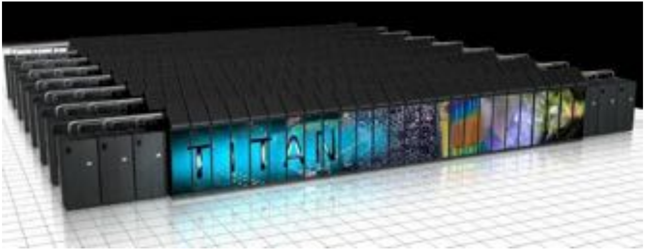
Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	4,389.82	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	34.58
2	3,631.70	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
4	3,459.46	SURFsara	Cartesius Accelerator Island - Bullx B515 cluster, Intel Xeon E5-2450v2 8C 2.5GHz, InfiniBand 4× FDR, Nvidia K40m	44.40
5	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x Level 3 measurement data available	1,753.66
6	3,131.06	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41
7	3,019.72	CSIRO	CSIRO GPU Cluster - Nitro G16 3GPU, Xeon E5-2650 8C 2GHz, Infiniband FDR, Nvidia K20m	86.20
8	2,951.95	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.93GHz, Infiniband QDR, NVIDIA K20x	927.86
9	2,813.14	Exploration & Production - Eni S.p.A.	HPC2 - iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, NVIDIA K20x	1,067.49
10	2,678.41	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	54.60

TOP500: Tianhe-2

- 16000 nodes (2Xeon+3PHI);
- Peak: 54,9 PFlops, Sust: 33,8 PFlops; $\epsilon = 62\%$; Power: 17,8MW

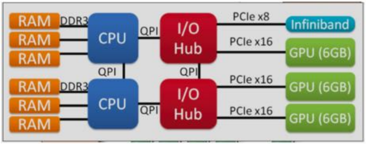
TOP500: Titan

- 18000 nodes (1 Xeon+1 K20x);
- Peak: 27,1 PFlops, Sust: 17,6 PFlops; $\epsilon = 65\%$; Power: 8,2MW



Hybrid Supercomputer: CPU + Accelerators

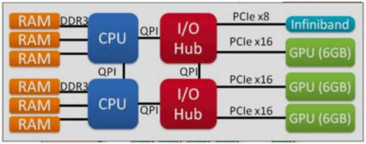
Most high-end HPC systems are characterized by *hybrid architecture*



- ASIP, FPGA or commodity components (GPGPU...)
- Better $/*PeakFlops*: offload cpu task to accelerator able to perform faster
- May consume less energy and may be better at streaming data.

Hybrid Supercomputer: CPU + Accelerators

Most high-end HPC systems are characterized by *hybrid architecture*



- ASIP, FPGA or commodity components (GPGPU...)
- Better $/*PeakFlops*: offload cpu task to accelerator able to perform faster
- May consume less energy and may be better at streaming data.
- —> warning!!!:
  - computing efficency ϵ (Sustained/Peak) not impressive
  - it's a function of accelerator and network...

	Nazione	Score	Numero Nodi	Tipologia Acc.	Peak Perf(Pflops)	Linpack Perf (Pflops)	Efficiency	Power (MW)	Interconnect
Tianhe-2	China	1	16000 (2CPU+3PHI)	Xeon + PHI	54,9	33,8	62%	17,8	Proprietary
Titan	USA(Oak R.)	2	18000(1CPU+1K20x)	Opteron + K20x	27,1	17,6	65%	8,2	Cray Gemini
Piz Daint	Switzerland	6	5272(1CPU+1K20x)	Xeon + K20x	7,8	6,2	79%	2,3	Cray Aries
Stampede	USA (TACC)	7	6400	Xeon + PHI	8,5	5,1	60%	4,5	Infiniband

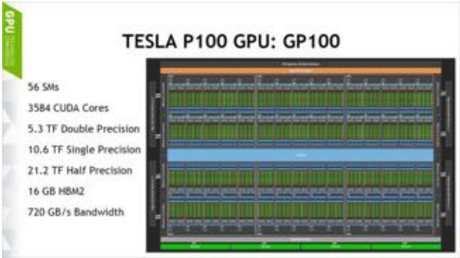
- Heterogeneous CPU/GPU systems: CPU for sequential code, (GP)GPU for parallel code
- Impressive use of state-of-the-art technologies
  - Example NVidia Tesla: 3D stacked mem, Proprietary GPU-GPU interconnect (NVLink), multi (10) TFlops/Proc, power effective...
- Processing is highly data-parallel (i.e. good for data parallel applications)
  - GPUs are SIMD-like and highly multithreaded: many parallel threads (up to 10^3...) distributed on many cores ($10^2 - 10^3$)
  - Graphics memory is wide ($N * 10^2$ bits) and high bandwidth ($N * Ghz$ per bit line).
- Programming languages standard (DirectX, OpenGL, OpenCL) or proprietary (NVidia Compute Unified Device Architecture (CUDA))

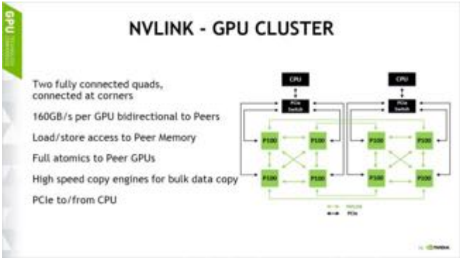
INTEL Westmere
+many caches - few processing

NVidia Fermi GPU
many computing units!!!

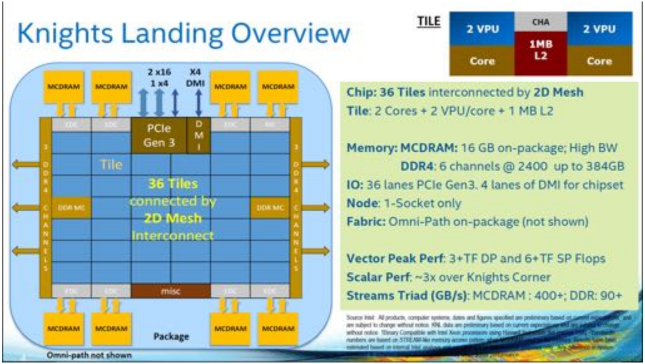
NVidia Pascal P100 recently announced...

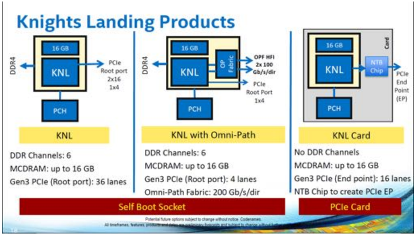
Tesla Products	Tesla P100	Tesla K80	Tesla K40	Tesla M40
GPU	GP100 (Pascal)	2 x GK210 (Kepler)	GK110 (Kepler)	GM200 (Maxwell)
SMs	56	26 (13 per GPU)	15	24
CUDA cores	3840	4992 (2 x 2496)	2880	3072
Base Clock	1328 MHz	560 MHz	745 MHz	948 MHz
GPU Boost Clock	1480 MHz	875 MHz	810/875 MHz	1114 MHz
Peak Double Precision	5.3 TFLOPS	2.91 TFLOPS	1.68 TFLOPS	.2 TFLOPS
Peak Single Precision	10.6 TFLOPS	8.73 TFLOPS	5.04 TFLOPS	7 TFLOPS
Memory Interface	4096-bit HBM2	2 x 384-bit GDDR5	384-bit GDDR5	384-bit GDDR5
Memory Size	16 GB	24GB (12GB per GPU)	12 GB	24 GB
Peak Bandwidth	720 GB/s	480 GB/s (240 GB/s per GPU)	288 GB/s	288 GB/sec
TDP	300 Watts	300 Watts	235 Watts	250 Watts
Transistors	15.3 billion	2 x 7.1 billion	7.1 billion	8 billion
GPU Die Size	610 mm²	2 x 561mm²	551 mm²	601 mm²
Manufacturing Process	16-nm	28-nm	28-nm	28-nm

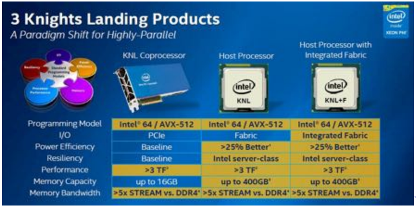




Accelerators: INTEL PHY (ex-MIC)

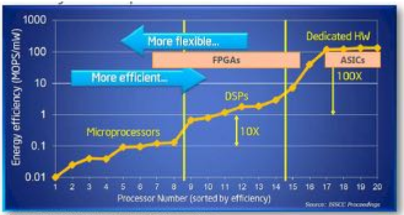




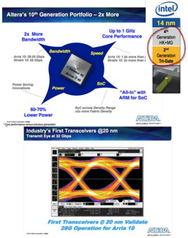


Accelerators: FPGA

- Stratix10 high-end, introduction 2016
- INTEL TriGate 14nm -> 30% less than old generation power consumption
- 96 transceivers @32Gbps (56Gbps?) for chip-to-chip interconnection and @28Gbps for backplane/cable interconnection
- Many industrial standards supported included CAUI-x (Nvlink)
- tons of programmable logic @1GHz
- ...and "for free"
  - 10 Tflops of DSP single precision FP
  - HMC (3D mem, high bandwidth) support
  - Multiple (4->8) ARM Cores (a53/57) @1.5GHz
- Similar in performance: XILINX Zynq UltraScale+ MPSoC Devices

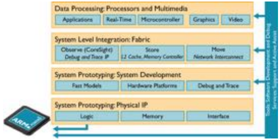


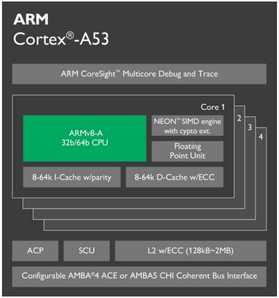
Source: Bob Broderson, Berkeley Wireless group



Low power CPU: ARM

- ARM is the only "European" CPUs maker
- Innovative business model: ARM sell Intellectual Properties hw/sw instead of physical chip;
  - 1100 licenses signed with over 300 companies and royalties received on all ARM-based chips
  - Pervasive technology: Android and Apple phones and tablets, RaspberryPI, Arduino, set-top box and multimedia, ARM-based uP in FPGA, ...
  - From 1990, *60 billion* of ARM-based chips delivered



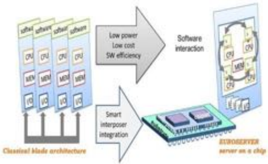


- Architecture specialised for embedded/mobile processors:
  - low power, low silicon area occupation, real-time, scalable, energy-efficient
- few generations of high end (64 bits) processors delivered
  - current Cortex Axx ARM V8-A enabling multi-core ARM-based processors
  - complete IP portfolio

ARM & HPC

Several attempts to use ARM low power processors in high end computing

- Server and micro-server ARM-based
  - AMCC X-gene 3, 32 v8-A cores@3GHz,
  - CAVIUM ThunderX SoCs up to 48 v8-A cores@2.4GHz
  - Broadcom/Qualcomm multi-core, Samsung SoC Exynos
- EU-funded projects
  - Mont-blanc project (BSC)
  - UniServer
  - ....



- INFN COSA project measured energy efficiency of low power architecture ARM based for scientific computing (Astrophysics, Brain simulation, Lattice-Boltzmann fluid-dynamics,..). On average:
  - $\sim$3x ratio x86 core / ARM core performances
  - but $\sim$10x ratio x86 core / ARM power consumption
  - –> ARM architectures *3x less* energy to solution for scientific applications

APE supercomputers

APE is a 25 years old project (!)

- MPP (APE1, APE100, APEmille, apeNEXT) & PC Cluster interconnection network (apeNET)
- FP Engine optimized for application + Smart dedicated 3D Torus interconnection network

APE1 (1988) 1GF, chipset Weitek

APE100 (1992) 25GF, SP, REAL "Home made" VLSI processors

APEmille (1999) 128GF, SP, Complex
Italy+France+Germany collaboration

apeNEXT (2004)
800GF, DP, Complex

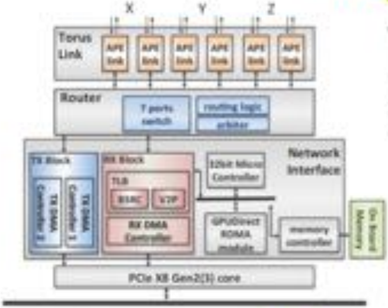
apeNET

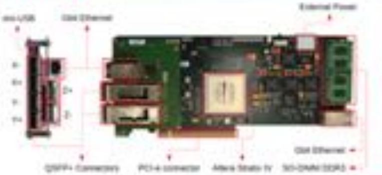
- APEnet: PC Cluster 3D torus network FPGA-based
  - Integrated routing and switching capabilities
  - High throughput, low latency, "light-weight" protocol
  - PCI Interface, 6 Links full-bidir on torus side

- History
  - 2003-2004: APEnet V3 (PCI-X)
  - 2005: APEnet V3+
    - same HW with RDMA API
  - 2006-2009: APEnet goes embedded
  - 2011: APEnet+
    - PCI Express, enhanced torus links

2

APEnet+





Project	Board	Comb. ALUT	Register	Memory (KB)
APEnet+	[illegible]	[illegible]	[illegible]	[illegible]

Logic Block	Comb. ALUT	Register	Memory (KB)
PCIe	[illegible]	[illegible]	[illegible]
Host TX block	[illegible]	[illegible]	[illegible]
GPU TX block	[illegible]	[illegible]	[illegible]
RX Logic	[illegible]	[illegible]	[illegible]
Nios II	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
TLB total	[illegible]	[illegible]	[illegible]
DNP Fault Manager	[illegible]	[illegible]	—
[illegible]	[illegible]	[illegible]	[illegible]
HARDWARE CUSTOMIZATION			
[illegible]	[illegible]	[illegible]	—
Router Inter-Port	[illegible]	[illegible]	—
Router Intra-Port	[illegible]	[illegible]	—
APElink	[illegible]	[illegible]	[illegible]

- APEnet+ Card:
  - 3D Torus: Scalable; Cost effective
  - Altera Stratix IV FPGA (PCIe X8 Gen2 board)
  - 6x Full bidir Channel (~400 Gbps; QSFP+)
- APEnet+ based on DNP:
  - Network Interface:
    - **RDMA** : Nios II (2011); TLB (2013); ASIP (2013)
    - **GPUDirect**: NVP2P (2012); RDMA (2013)
  - Router: XYZ ; VCT; 7 data flows @2.8GB/s
  - Torus Link (6x APElink):
    - Physical Layer: Altera Transceiver (PCS/PMA)
    - Data Layer: TCL; Word-Stuffing; Diagnostic Message

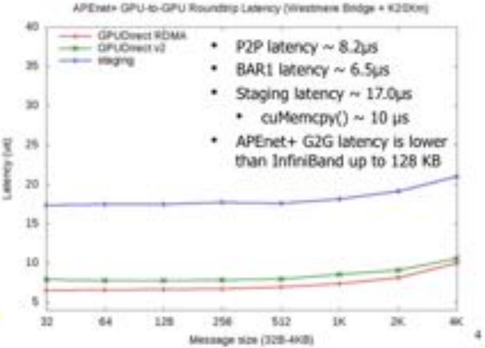
3

GPUDirect RDMA in APEnet+

Peer-to-Peer means:
- Data exchange on the PCIe bus
- No bounce buffers on host

APEnet+ P2P support
- cutting-edge HW/SW technologies developed jointly with Nvidia
  - APEnet+ board acts as a peer
  - APEnet+ board can read/write "directly" GPU memory

Direct GPU access
- Specialized APEnet+ HW block
- Latency saver for small size messages

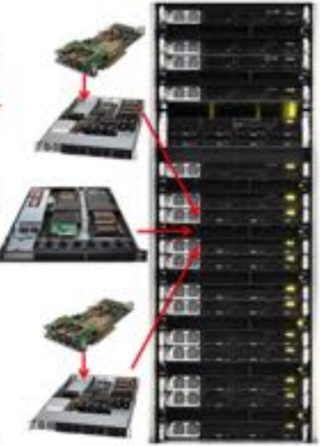


4

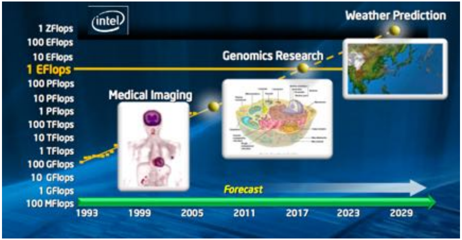
QUOnG: GPU+3D Network FPGA-based

QUonG (QUantum chromodynamics ON Gpu) is a comprehensive initiative aiming to deploy an GPU-accelerated HPC hardware platform mainly devoted to theoretical physics computations.

- Heterogeneous cluster: PC mesh accelerated with high-end GPU (Nvidia) and interconnected via 3-D Torus network
- Added value:
  - tight integration between accelerators (GPU) and custom/reconfigurable network (DNP on FPGA) allows latency reduction and computing efficiency gain
  - Huge hardware resources in FPGA to integrate specific computing task accelerators
    - ASIP, OpenCL (in the future..)
- Communicating with optimized custom interconnect (APEnet+), with a standard software stack (MPI, OpenMP, ...)
- Community of researchers sharing codes and expertise (LQCD, GWA, Laser-plasma interactions, BioComputing, Complex systems,...)



3



- HPC is mandatory to compare observations with theoretical models
- HPC infrastructure is the theoretical laboratory to test the physical processes.

Let's talk of Basic Science...

- High Energy & Nuclear Physics
  - LQCD (again...), Dark-energy and dark matter, Fission/Fusion reactions (ITER)
- Facility and experiments design
  - Effective design of accelerators (also for medical Physics, GEANT...)
  - Astrophysics: SKA, CTA
  - ...
- Life science
  - Personal medicine: individual or genomic medicine
  - Brain Simulation <– see P. Paolucci talk of this afternoon ...

Technological challenges for ExaScale systems

Just to name a few....

- Power efficiency and compute density
  - huge number of nodes but limited data center power and space
- Memory and Network technology
  - memory hierarchies: move data faster and closer...
  - increase memory size per node with high bandwidth and ultra-low latency
  - distribute data across the whole system node set but access them with minimal latency...
- Reliability and resiliency
  - solutions for decreased reliability (extreme number of state -of-the-art components) and a new model for resiliency
- Software and programming model
  - New programming model (and tools) needed for hierarchical approach to parallelism (intra-node, inter-node, intra-rack....)
  - system management, OS not yetready for ExaScale...
- Effective system design methods
  - CO-DESIGN: a set of a hierarchical performance models and simulators as well as commitment from apps, software and architecture communities

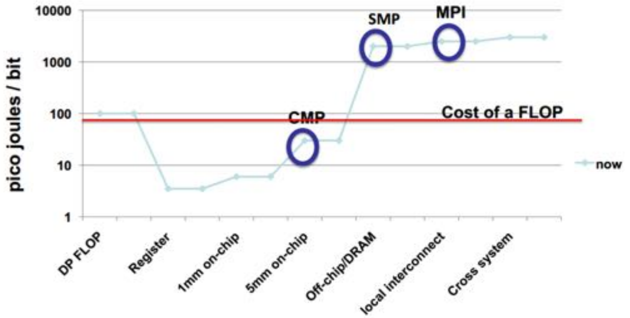
- General agreement on the fact that data center power budget is less than 20 MW
  - half for cooling -> only 10MW for active electronics
- Current processors performances are
  - multi-core CPU: 1 TFlops/100W
  - GPGPU: 5-10 TFlops/300W but worst sustained/peak (and needs CPU) so only a factor 1.5 better
  - add few tens of watt for distributed storage and memory per node
- ExaScale sustained (where $\epsilon = 50\% - 70\%$)
  - 10^6 computing nodes
  - 100 MW of power –> *low power* approach is needed

Big numbers, big problems: system packing and cooling

- Current computing node assembly:
  - 8 processors into 1U box
  - ~30 1Uboxes per 42U rack (25% of volume dedicated to rack services)
- Summing up
  - 4000 racks per ExaFlops sustained
  - 6000 m^2 of floor space
  - service racks (storage, network infrastructure, power&controls, chillers,...) not included (!!)

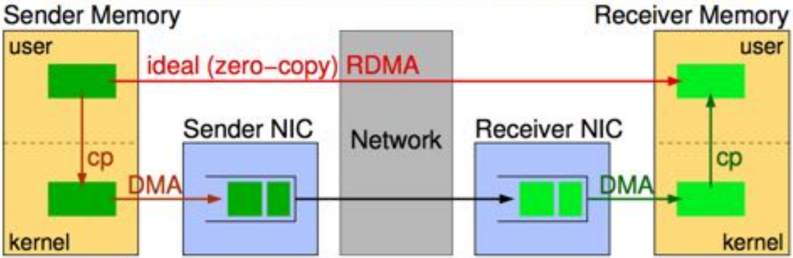
- It needs:
  - New mechanics for denser systems
  - New cooling technology (liquid/gas cooling) for reduce impact of cooling system on power consumption and data center space

Big numbers, big problems: data locality



- Needed improved hierarchical architectures for memory and storage
  - distributed hierachical memory
  - zero-copy through R(emote)DMA, P(artitioned)G(lobal)A(ddress)S(pace) leveraging on affinity to exploit data locality
- low latency, high bandwidth network

<u>Inefficiencies of traditional "Send" – *how to overcome them*</u>



<u>Up to 5x (!) inefficiencies</u>

- Receiver copy NIC→User – *rcv addresses visible to sender: RDMA – PGAS*
- Protection – *virtualized, user-level DMA initiation, IOMMU*
- Buffer Pinning for DMA – *allow RDMA to fail, like page faults for ld/st*
- Send before receive buffer allocated – *fix the API / Application*
- Send buffer reuse immediately after send – *fix the API / Application*

Katevenis & Chrysos - AISTECS 2016 - Exascale-Computing Interconnects 45

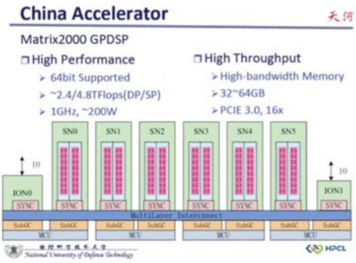
Big numbers, big problems: data locality

Wish List for an ideal Copy (RDMA) Engine

- User-Level RDMA Initiation:
  - Arguments to be full, arbitrary 64-bit *Virtual* Addresses
  - Control Registers to be virtualized and protected *per-process*
- No System Call necessary:
  - Virtual to Physical Address Translation via *HW MMU's* –not OS
  - Notification of Compl'n-Arrival: *per-process* Mailbox, not interrupt
- (true) Zero-Copy:
  - Any user page as source / destination
  - No need for pinning the src-dst pages in-memory: allow for translation failures during RDMA operation, resuting in notification of incomplete operation –like normal page-faults
  - Also useful for **Resilience**
- Exascale Global Addr. Space: full 64-b virtual addr. (+PID) throughout
- Performance: multi-channel engine; per-channel flow/rate control

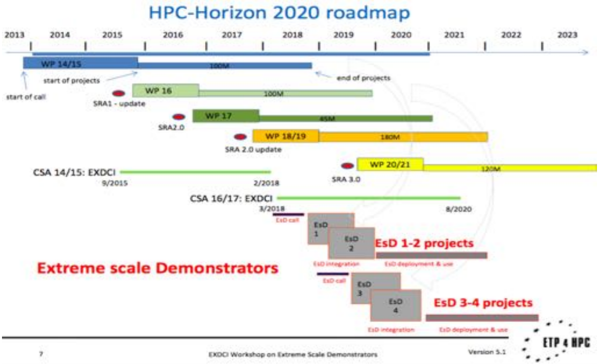
What next?

- US CORAL (Collaboration of Oak Ridge, Argonne, and Livermore) project, 525+M$ from DOE, for 3 100-200 PetaFlops systems in 2018-19 (Pre-Exascale system), ExaScale in 2023
  - *Summit/Sierra* OpenPower-based (IBM P9 + NVidia GPU + Mellanox) 150(300) PFLops/10MW
  - *Aurora* Intel-based (CRAY/INTEL, Xeon PHI Knights Hill, Omnipath) 180(400) PFlops/13MW
- JAPAN FLAGSHIP2020 RIKEN + Fujitsu
  - derived from Fujitsu K-computer, SPARC64-based + Tofu interconnect, delivered in 2020
- CHINA ??? , NUDT + Government
  - ShenWei and FeiTang CPUs plus proprietary GPU and network... delivered in 2020

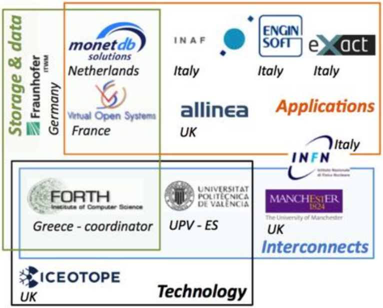


What next in Europe?

- Attempt to stimulate an European based HPC industry
- Mainly EU funded trough FPx and H2020 R&D programs: from PRACE to FETHPC and FET Infrasctructure projects
- Industrial partners involved in R&D: ARM(UK), ST(FR), ATOS-BULL(FR), Eurotech (IT), E4(IT) and many more on software/system side
- a pletora of Universities and Research Centers and an "obscure" cloud of keywords and acronyms ;)



ExaNeSt



ExaNeSt: European Exascale System Interconnection Network & Storage

- EU Funded project H2020-FETHPC-1-2014
- Duration: 3 years (2016-2018). Overall budget about 7 MEuro.
- Coordination FORTH (Foundation for Research & Technology, GR)
- 12 Partners in Europe (6 industrial partners)

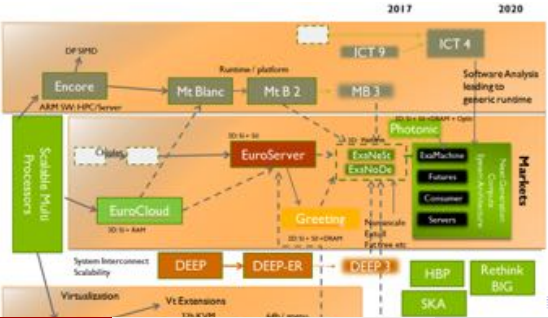
ExaNeSt objectives

"...Overall long-term strategy is to develop a European low-power high-performance Exascale infrastructure based on ARM-based micro servers..."

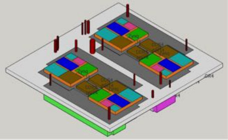
- System architecture for datacentric Exascale-class HPC
  - Fast, distributed in-node non-volatile-memory
  - Storage Low-latency unified Interconnect (compute & storage traffic)
    - RDMA + PGAS to reduce overhead
- Extreme compute-power density
  - Advanced totally-liquid cooling technology
  - Scalable packaging for ARM-based (v8, 64-bit) microserver
- Real scientific and data-center applications
  - Applications used to identify system requirements
  - Tuned versions will evaluate our solutions

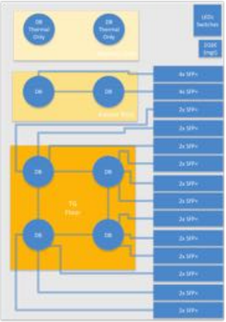
ExaNeSt ecosystem

- EuroServer: Green Computing Node for European microservers
  - *UNIMEM* PGAS model among ARM computing nodes
- INFN EURETILE project: *brain inspired* systems and applications
  - APEnet+ network on FPGA + brain simulation (DPSNN) scalable application
- Kaleao: Energy-efficient uServers for Scalable Cloud Datacenters
  - startup company interested in commercialisation of results
- *Twin* projects: ExaNode and EcoScale
  - ExaNode: ARM-based Chiplets on silicon Interposer design
  - EcoScale: efficient programming of heterogenous infrastructure (ARM + FPGA accelerators)



ExaNeSt prototypes





- Computing module based on Xilinx Zynq UltrScale+ FPGA...
  - Quad-core 64-bit ARM A53
  - ~1 TFLOPS of DSP logic
- ... placed on small Daugther Board (QFDB) with
  - 4 FPGAs, 64 GB DDR4,
  - 0.5-1 TB SSD,
  - 10x 16Gb/s serial links-based I/O per QFDB
- mezzanine(blade) to host 8 (16 in second phase) QFDBs
  - intra-mezzanine QFDB-QFDB direct network
  - lots of connectors to explore topologies for inter-mezzanine network

ExaNeSt prototypes

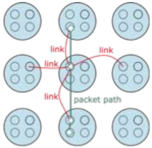
- ExaNeSt high density innovative mechanics...
  - 8(16) QFDBs per mezzanine
  - 9 blades per chassis
  - 8-12 chassis per rack
- ...totally liquid cooling
  - track 1: immersed liquid cooled systems based on convection flow
  - track 2: phase-change (boiling liquid) and convection flow cooling (up to 350 kW of power dissipation capability...)

- $\sim 7PFlops$ per racks and $20GFlops/W$
- Extrapolating from current technology, ExaNeSt-based Exascale system with 140 racks, 21M ARM cores and 50MW

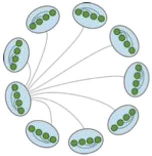
ExaNeSt Interconnect

ExaNeSt is working testbed FPGA-based to explore and evaluate innovative network architectures, network topologies and related high performance technologies.

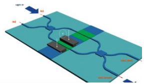
- Unified approach
  - merge interprocessor and storage traffic on same network medium
  - PGAS architecture and RDMA mechanisms to reduce communication overhead



- innovative routing functions and control flow (congestion managements)
- explore performances of different topologies
  - Direct blade-to-blade networks (Torus, Dragonfly,...)
  - Indirect blade-switch-blade networks



- All-optical switch for rack-to-rack interconnect (ToR switch)
- Support for resiliency: error/detect correct, multipath routing,...
- Scalable network simulator to test large scale effects in topologies



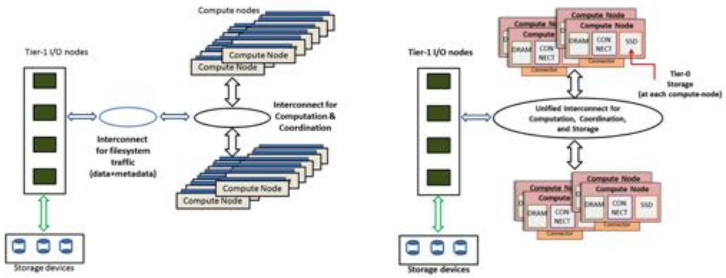
Co-design approach

- Applications define quantitative requirements for the system under design
- Applications evaluate the hw/sw system
  - 1st phase: synthetic benchmarks extracted by execution of target applications Traces; used to test I/O and network capability
  - 2nd phase: re-engineering of real applications through optimisation of data distribution, data communication and storage usage
- ExaNeSt will deliver a new generation of Exascale (almost...) ready applications

List of ExaNeSt applications

- Cosmological n-Body and hydrodynamical code(s) (INAF)
  - Large-scale, high-resolution numerical simulations of cosmic structures formation and evolution
- Brain Simulation (DPSNN) (INFN)
  - Large scale spiking behaviours and synaptic connectivity exhibiting optimal scaling with the number of hardware processing nodes (INFN).
  - Mainly multicast communications (all-to-all, all-to-many).
- Weather and climate simulation (ExactLab)
- Material science simulations (ExactLab and EngineSoft)
- Workloads for database management on the platform and initial assessment against competing approaches in the market (MonetDB)
- Virtualization Systems (Virtual Open systems)

ExaNeSt storage



- Distributed storage: NVM close to the computing node to get low access latency and low power access to data
- based on Bee**GFS** open source parallel filesystem with caching and replication extensions
- Unified interconnect infrastructure per storage and inter-node data communication
- Highly optimized I/O path in the Linux kernel

Conclusions

- Fundamental scientific and engineering computing problems needs ExaScale computing power
- The race toward ExaScale is started and Europe is trying to compete with established and emerging actors (USA, Japan, China,...)
- Many challenging issues require huge R&D efforts: power, interconnect, system packing and effective software frameworks
- ExaNeSt will contribute to the evaluation and selection of ExaScale enabling technologies, leveraging on Europe traditional expertise: embedded systems (ARM), excellence in scientific programming, design of non-mainstream network architecture
- *...not only paper*: ExaNest will deliver a fully working prototype able to be scaled up to the ExaFlops in the next five years