Master Informatics Eng.

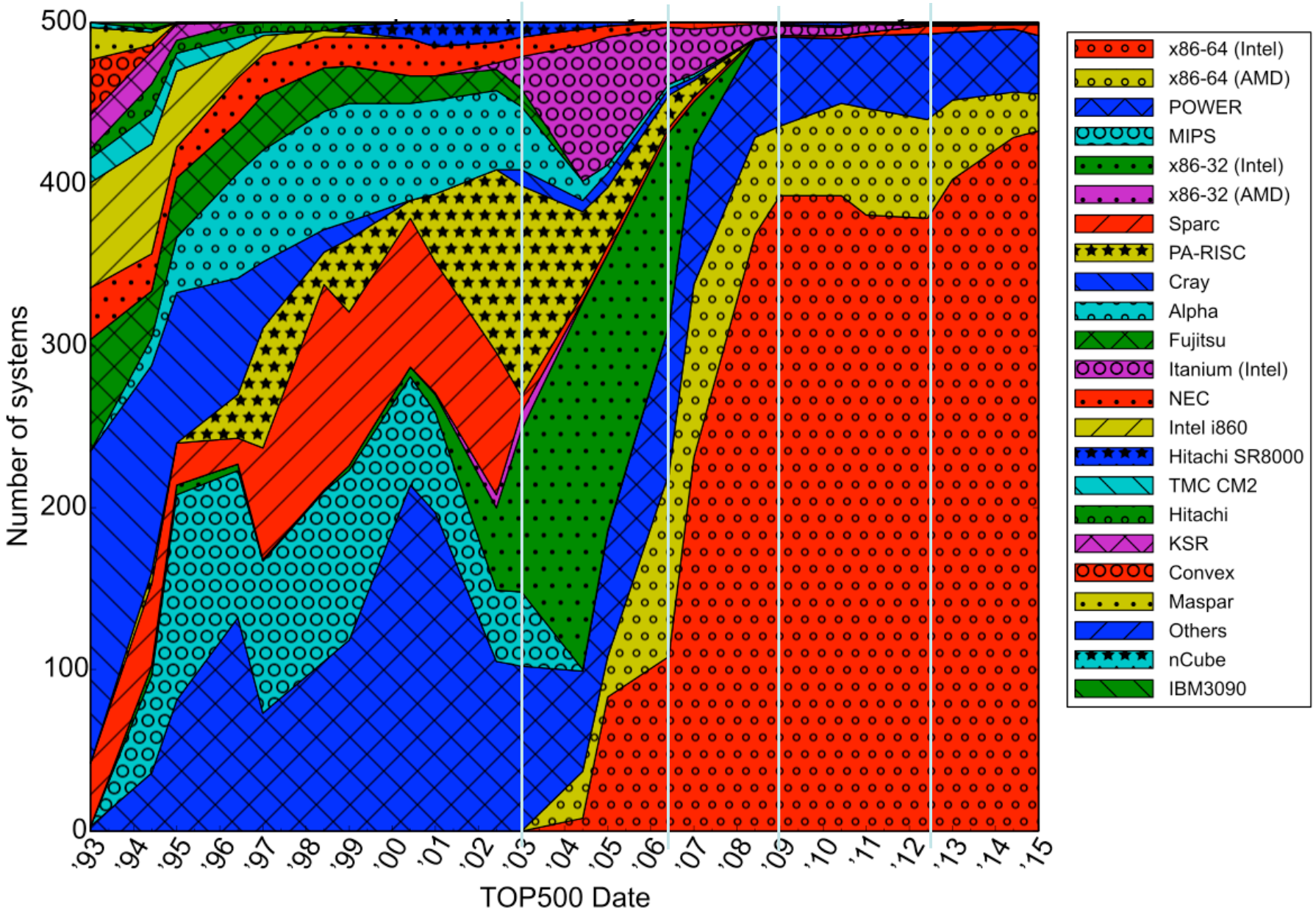
2017/18

A.J.Proença

Next generation of HPC systems

(most slides are borrowed)

Past: processor family distribution of all systems

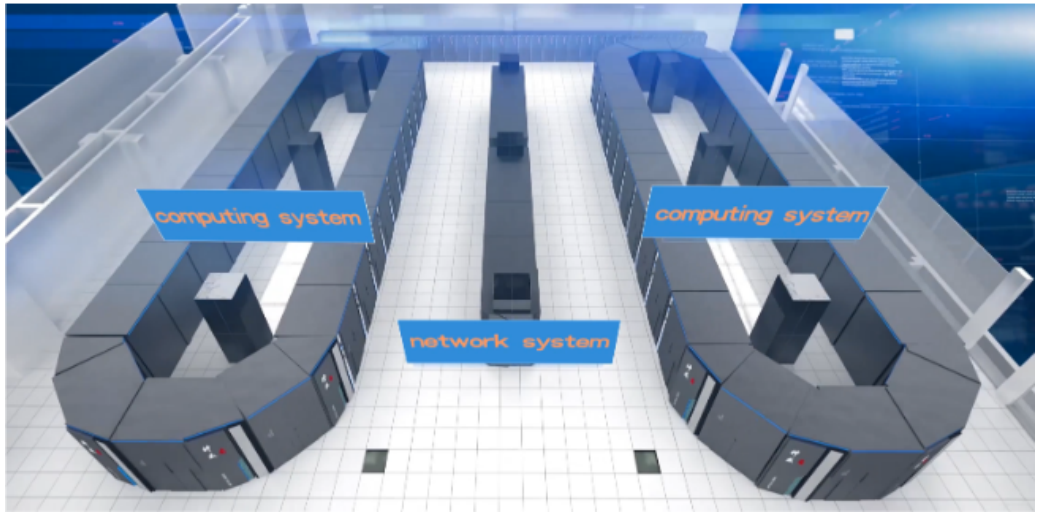


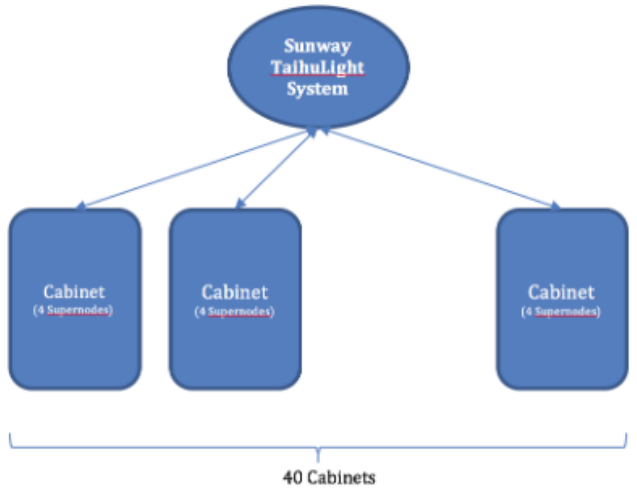
Current top 10 HPC systems
Nov'17 TOP500

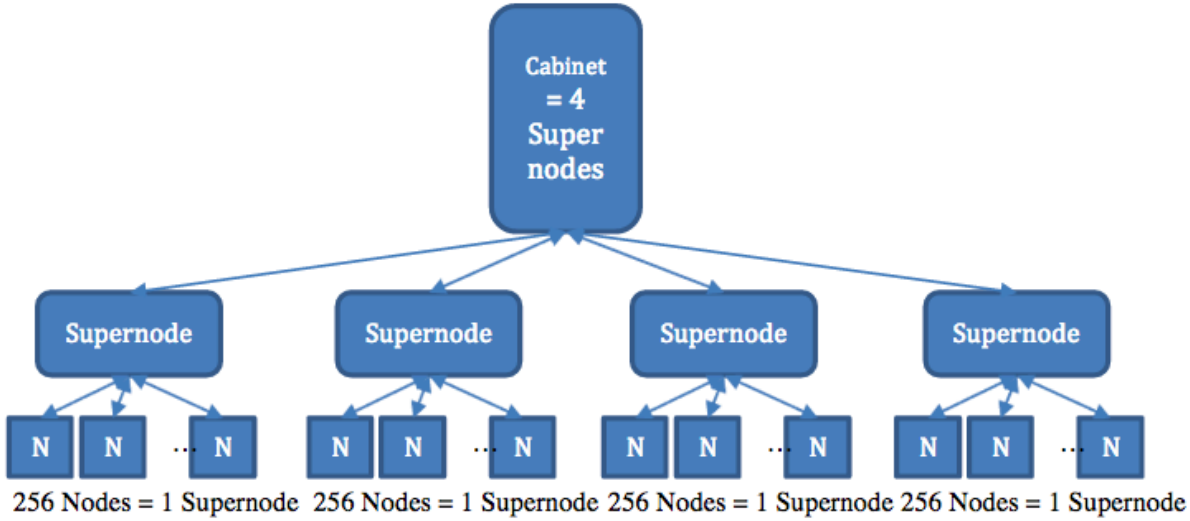
Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
2	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P , NUDT National Super Computer Center in Guangzhou China	3,120,000	33,862.7	54,902.4	17,808
3	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	361,760	19,590.0	25,326.3	2,272
4	**Gyoukou** - ZettaScaler-2.2 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz , ExaScaler Japan Agency for Marine-Earth Science and Technology Japan	19,860,000	19,135.8	28,192.0	1,350
5	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x , Cray Inc. DOE/SC/Oak Ridge National Laboratory United States	560,640	17,590.0	27,112.5	8,209
6	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom , IBM DOE/NNSA/LLNL United States	1,572,864	17,173.2	20,132.7	7,890
7	**Trinity** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , Cray Inc. DOE/NNSA/LANL/SNL United States	979,968	14,137.3	43,902.6	3,844
8	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , Cray Inc. DOE/SC/LBNL/NERSC United States	622,336	14,014.7	27,880.7	3,939
9	**Oakforest-PACS** - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path , Fujitsu Joint Center for Advanced High Performance Computing Japan	556,104	13,554.6	24,913.5	2,719
10	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect , Fujitsu RIKEN Advanced Institute for Computational Science (AICS) Japan	705,024	10,510.0	11,280.4	12,660

#1 since June'16 TOP500: Sunway TaihuLight

Overview of the Sunway TaihuLight System



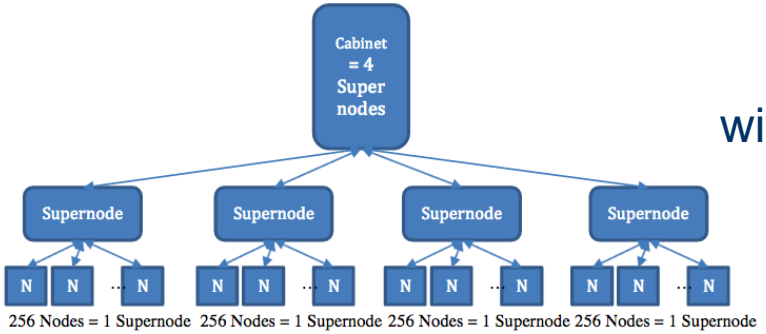




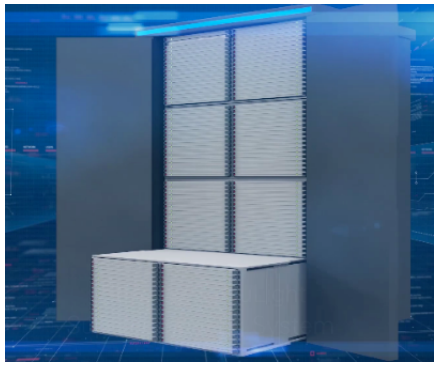
AJProença, Advanced Architectures,

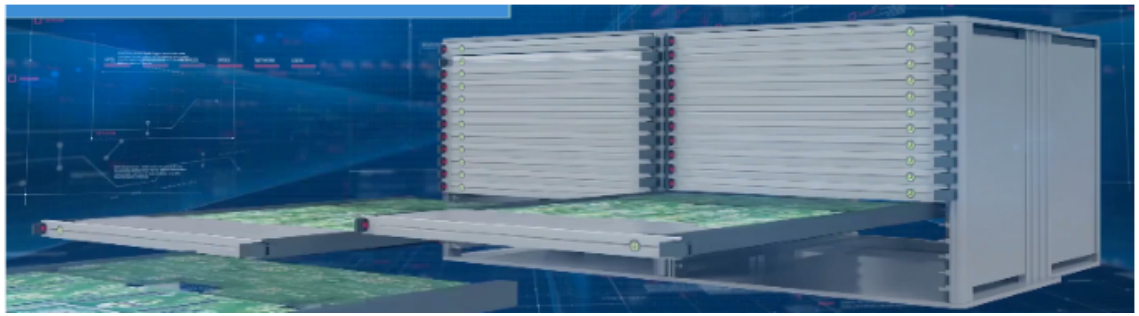
http://www.netlib.org/utk/people/JackDongarra/PAPERS/sunway-report-2016.pdf

#1 since June'16 TOP500: Sunway TaihuLight

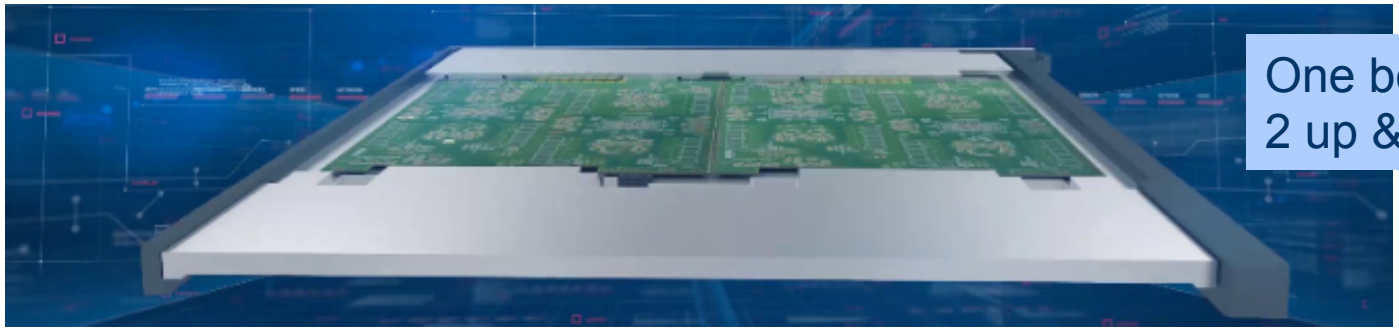


One cabinet with 4 Supernodes





One Supernode with 32 boards

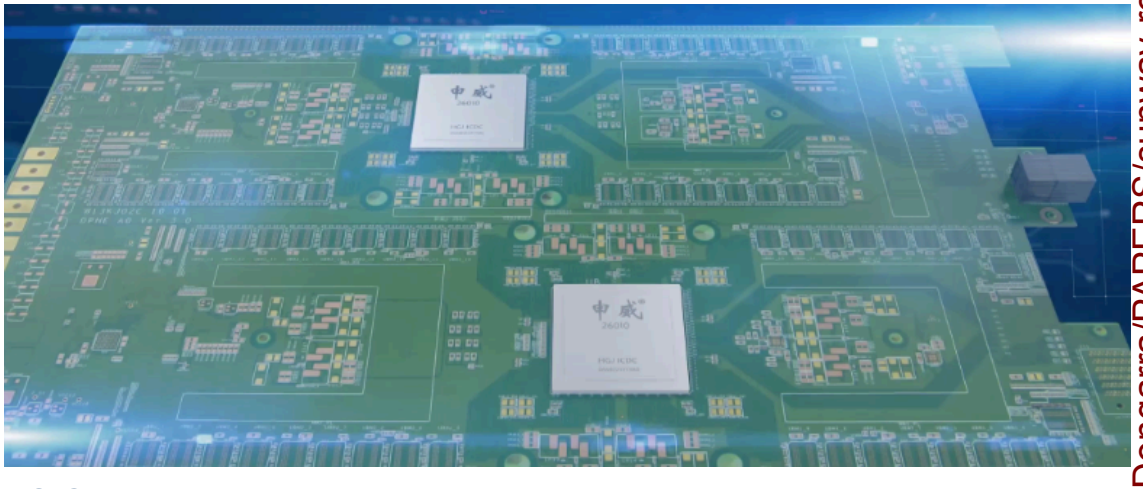


One board with 4 cards, 2 up & 2 down

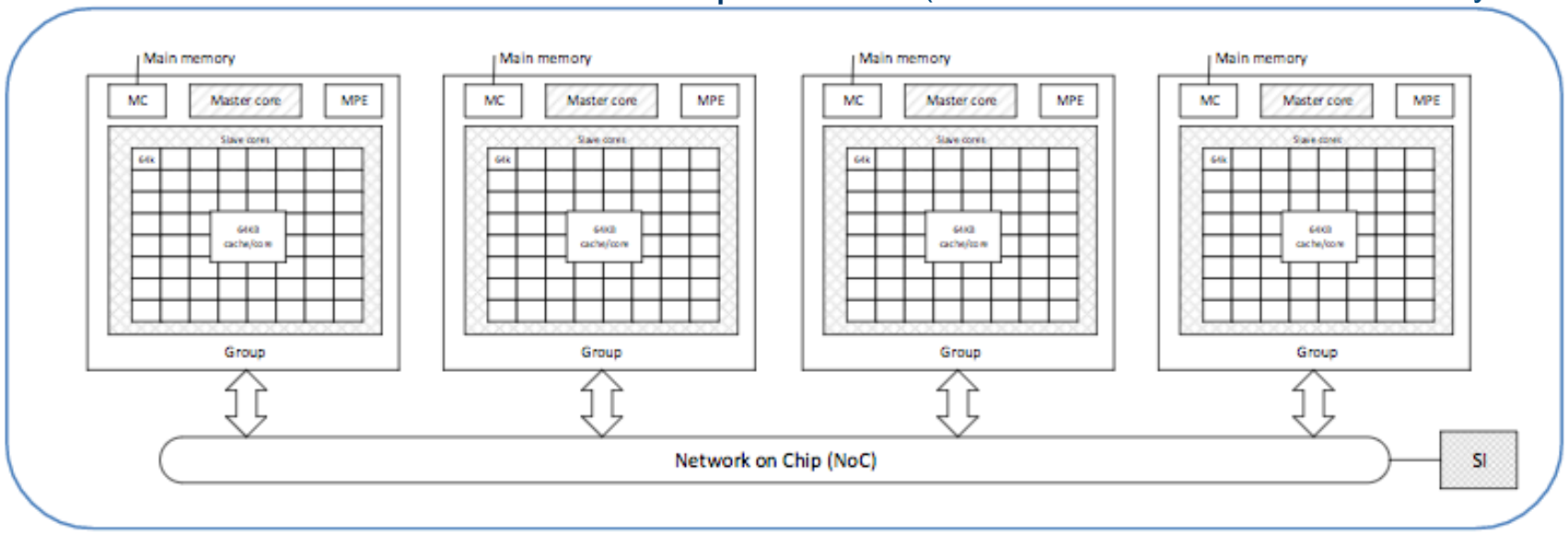
http://www.netlib.org/utk/people/JackDongarra/PAPERS/sunway-report-2016.pdf

#1 since June'16 TOP500: Sunway TaihuLight

One card with two nodes
(two SW26010 chips)

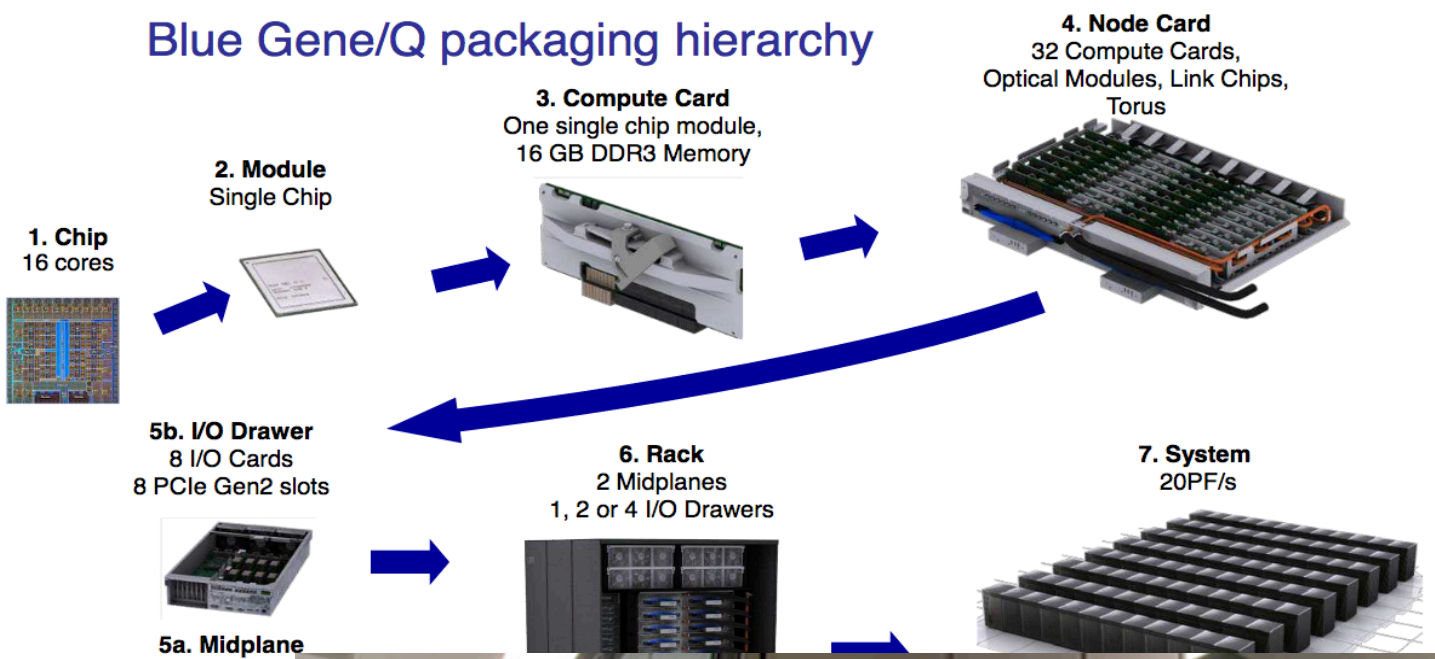


SW26010: the 4x64-core 64-bit RISC processor (w/ **256-bit vector** instructions & only cache L1)



http://www.netlib.org/utk/people/JackDongarra/PAPERS/sunway-report-2016.pdf

IBM Power BlueGene/Q Compute (Sequoia system)

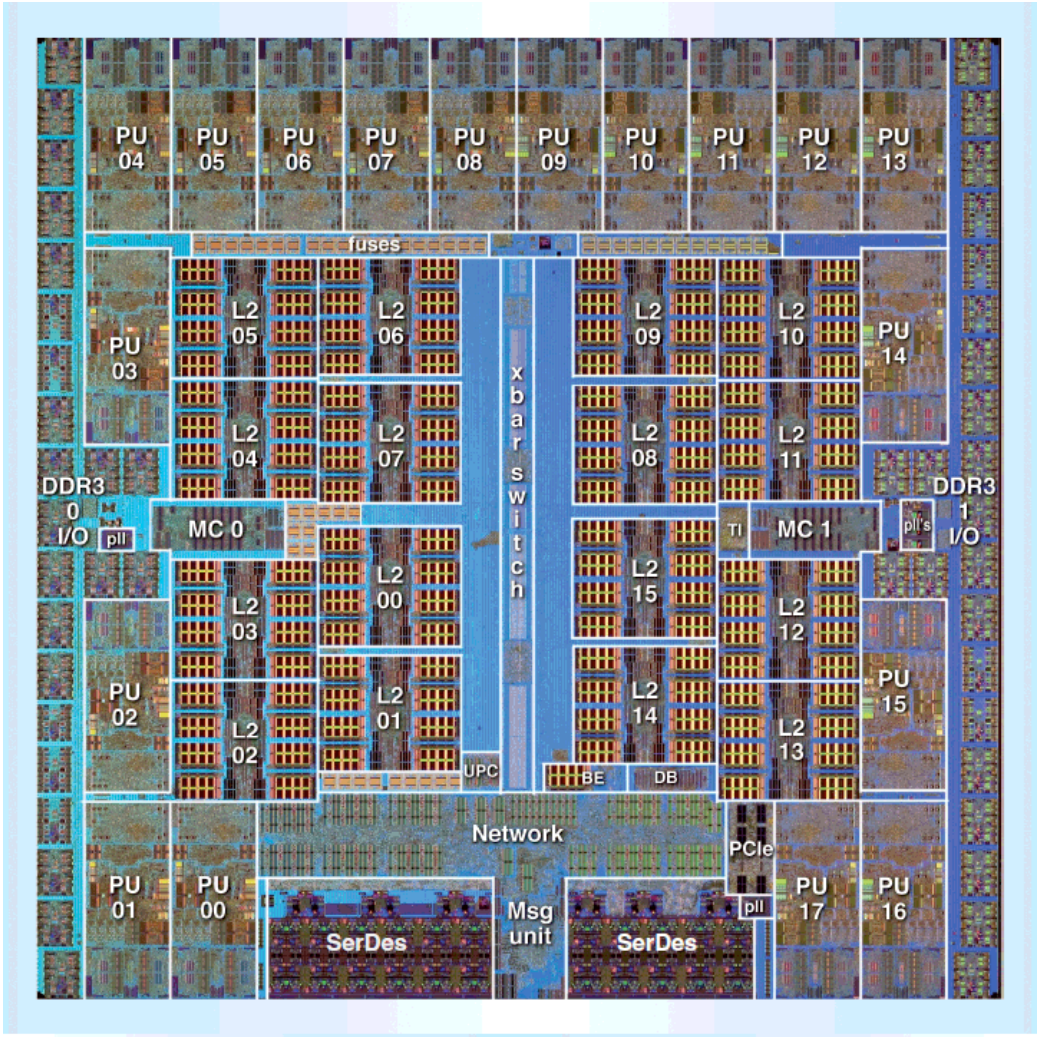


Ref: SC2010

TOP500 The List.

Jun'12: #1
Nov'12: #2
Jun'13: #3
Nov'13: #3
Jun'14: #3
Nov'14: #3
Jun'15: #3
Nov'15: #3
Jun'16: #4
Nov'16: #4
Jun'17: #5
Nov'17: #6

IBM Power BlueGene/Q Compute (chip)

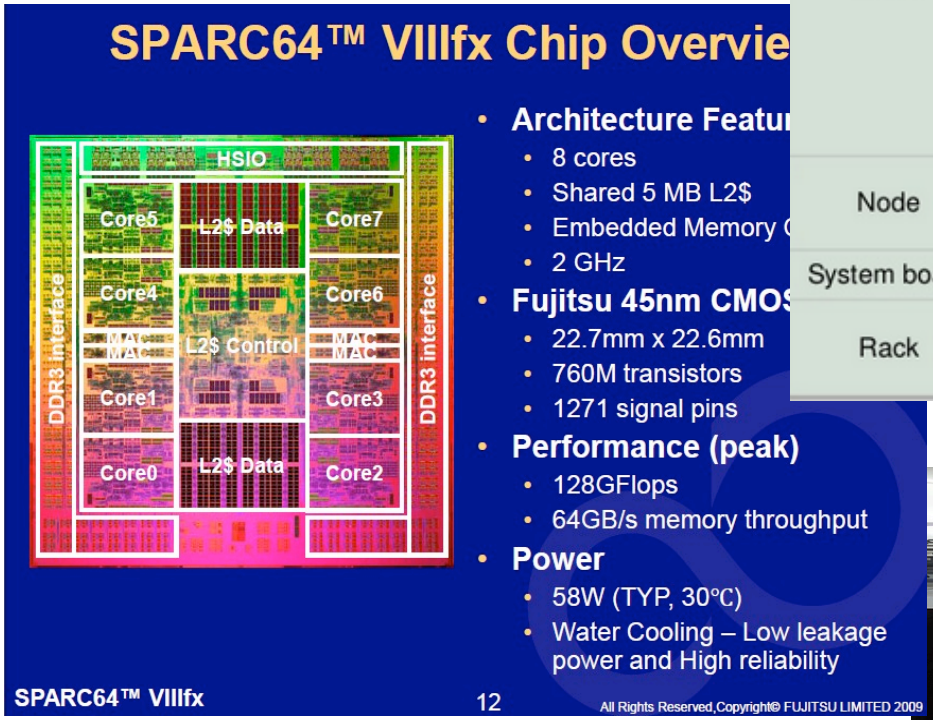


Features:

- launched in 2010/11
  (TOP500: #1 in Jun12, #4 in Jun16)
- 18-cores
  - 16 compute,
    1 OS support, 1 redundant
  - 64 bits PowerISA
  - 1.6 GHz
  - L1 I/D cache => 16 kB / 16 kB
  - **each core: quad-FPU**
    (4-wide double precision SIMD)
  - each core: 4-way multi-threaded
- shared L2 cache: 32 MB
- dual memory controller
- IBM ended development of BlueGene project in 2015...

Fujitsu K computer

(the Japanese word "kei" (京), means 10 quadrillion, 10^{16})



		K computer
CPU	Name	SPARC64™ VIIIfx
	Performance	128GFlops@2GHz
	Architecture	SPARC V9 + HPC-ACE extension
	Cache configuration	L1(I) Cache:32KB/core, L1(D) Cache:32KB/core
		L2 Cache: 6MB(shared)
	No. of cores/socket	8
	Memory band width	64 GB/s.
Node	Configuration	1 CPU / Node
	Memory capacity	16 GB
System board	Node/system board	4 Nodes
Rack	System board/rack	24 System boards
	Performance/rack	12.3 TFlops

TOP500 The List.

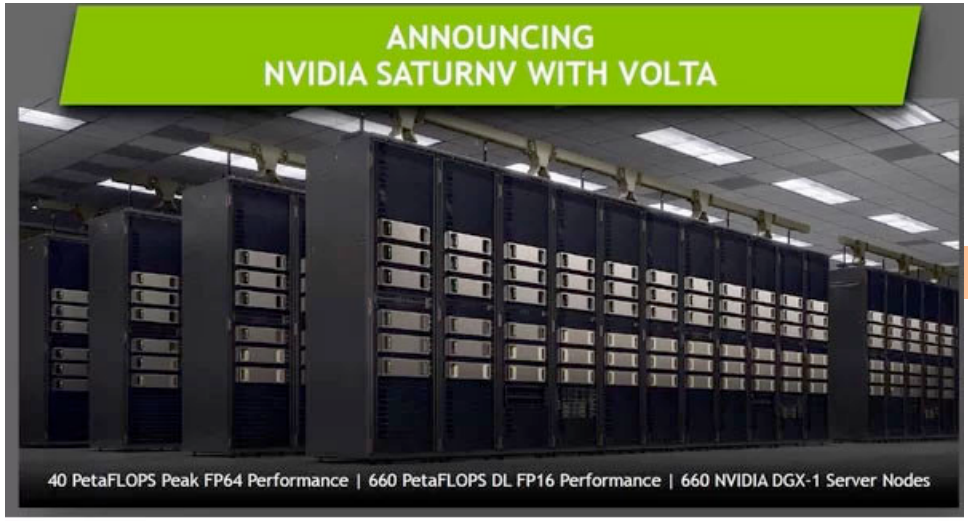
Jun'11: #1
Nov'11: #1
Jun'12: #2
Nov'12: #3
Jun'13: #4
Nov'13: #4
Jun'14: #4
Nov'14: #4
Jun'15: #4
Nov'15: #4
Jun'16: #5
Nov'16: #7
Jun'17: #8
Nov'17: #10

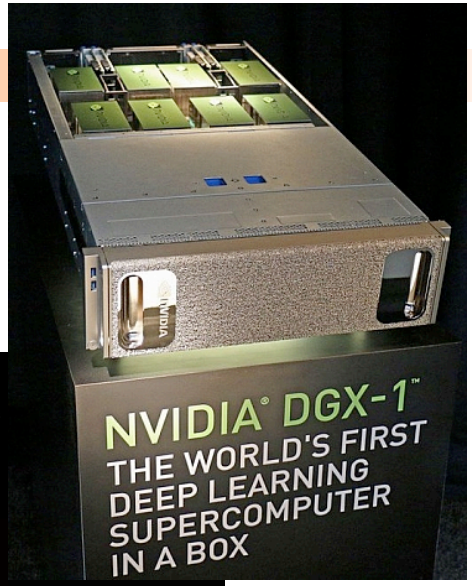
The GREEN 500

Current top 10 greener-HPC systems Nov'17 Green500

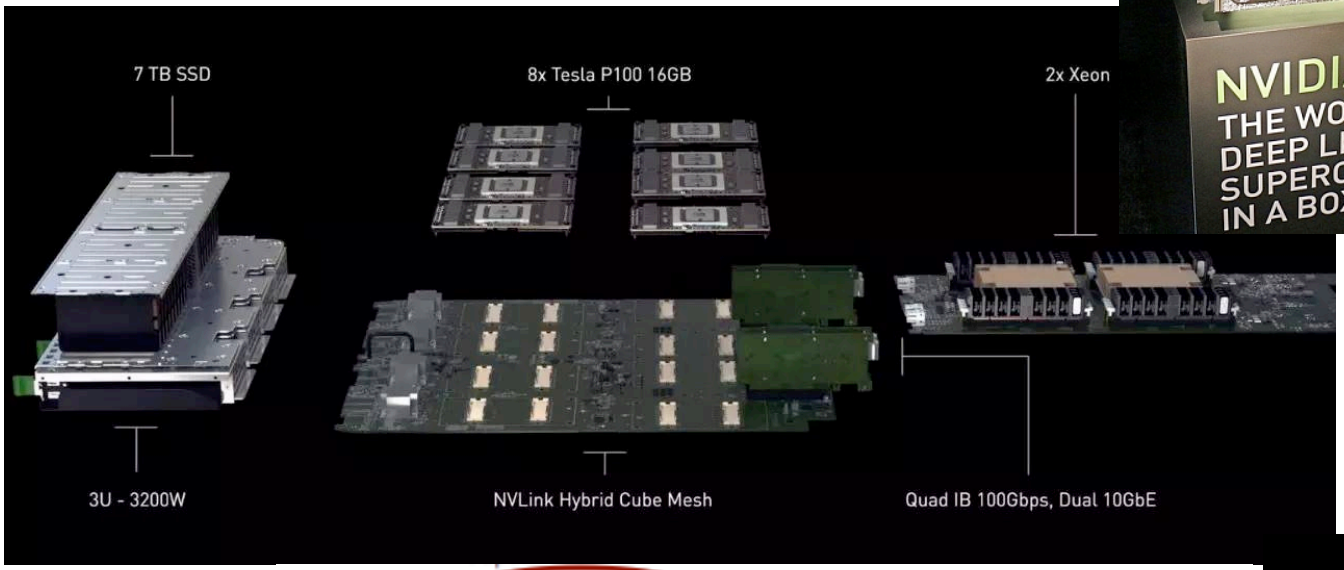
Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Power (kW)	Power Efficiency (GFlops/watts)
1	259	**Shoubu system B** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. Advanced Center for Computing and Communication, RIKEN Japan	794,400	842.0	50	17.009
2	307	**Suiren2** - ZettaScaler-2.2, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. High Energy Accelerator Research Organization /KEK Japan	762,624	788.2	47	16.759
3	276	**Sakura** - ZettaScaler-2.2, Xeon E5-2618Lv3 8C 2.3GHz, Infiniband EDR, PEZY-SC2 , PEZY Computing / Exascaler Inc. PEZY Computing K.K. Japan	794,400	824.7	50	16.657
4	149	**DGX SaturnV Volta** - NVIDIA DGX-1 Volta36, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla V100 , Nvidia NVIDIA Corporation United States	22,440	1,070.0	97	15.113
5	4	**Gyoukou** - ZettaScaler-2.2 HPC system, Xeon D-1571 16C 1.3GHz, Infiniband EDR, PEZY-SC2 700Mhz , ExaScaler Japan Agency for Marine-Earth Science and Technology Japan	19,860,000	19,135.8	1,350	14.173
6	13	**TSUBAME3.0** - SGI ICE XA, IP139-SXM2, Xeon E5-2680v4 14C 2.4GHz, Intel Omni-Path, NVIDIA Tesla P100 SXM2 , HPE GSIC Center, Tokyo Institute of Technology Japan	135,828	8,125.0	792	13.704
7	195	**AIST AI Cloud** - NEC 4U-8GPU Server, Xeon E5-2630Lv4 10C 1.8GHz, Infiniband EDR, NVIDIA Tesla P100 SXM2 , NEC National Institute of Advanced Industrial Science and Technology Japan	23,400	961.0	76	12.681
8	419	**RAIDEN GPU subsystem** - NVIDIA DGX-1, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , Fujitsu Center for Advanced Intelligence Project, RIKEN Japan	11,712	635.1	60	10.603
9	115	**Wilkes-2** - Dell C4130, Xeon E5-2650v4 12C 2.2GHz, Infiniband EDR, NVIDIA Tesla P100 , Dell EMC University of Cambridge United Kingdom	21,240	1,193.0	114	10.428
10	3	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	361,760	19,590.0	2,272	10.398

NVidia DGX-1 SaturnV

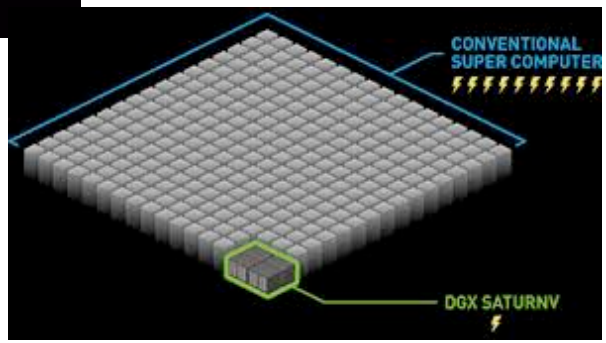




$149,000



4	149	DGX SaturnV Volta - NVIDIA DGX-1 Volta36, Xeon E5-2698v4 20C 2.2GHz, Infiniband EDR, NVIDIA Tesla V100 , Nvidia NVIDIA Corporation United States	22,440	1,070.0	97	15.113



TOP500: HPCG vs. HPL

HPCG UPDATE: ISC'17

Jack Dongarra
Michael Heroux
Piotr Luszczek

HPCG Snapshot

- High Performance Conjugate Gradients (HPCG).
- Solves *Ax=b*, *A* large, sparse, *b* known, *x* computed.
- An optimized implementation of PCG contains essential computational and communication patterns that are prevalent in a variety of methods for discretization and numerical solution of PDEs

- Patterns:
  - Dense and sparse computations.
  - Dense and sparse collectives.
  - Multi-scale execution of kernels via MG (truncated) V cycle.
  - Data-driven parallelism (unstructured sparse triangular solves).
- Strong verification (via spectral properties of PCG).

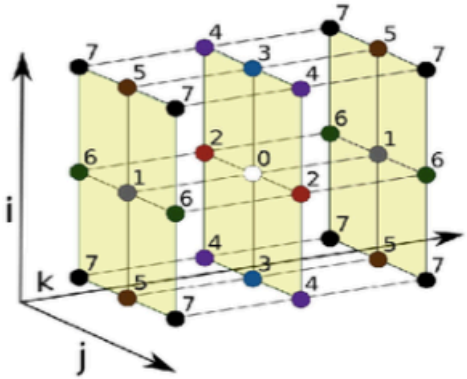
TOP500: HPCG vs. HPL

HPCG UPDATE: ISC'17
Jack Dongarra
Michael Heroux
Piotr Luszczek

hpcg-benchmark.org

Model Problem Description

- Synthetic discretized 3D PDE (FEM, FVM, FDM).
- Zero Dirichlet BCs, Synthetic RHS s.t. solution = 1.
- Local domain: $(n_x \times n_y \times n_z)$
- Process layout: $(np_x \times np_y \times np_z)$
- Global domain: $(n_x * np_x) \times (n_y * np_y) \times (n_z * np_z)$
- Sparse matrix:
  - 27 nonzeros/row interior.
  - 8 – 18 on boundary.
  - Symmetric positive definite.



27-point stencil operator

TOP500: HPCG vs. HPL

HPCG UPDATE: ISC'17

Jack Dongarra
Michael Heroux
Piotr Luszczek

Merits of HPCG

- Includes major communication/computational patterns.
  - Represents a minimal collection of the major patterns.
- Rewards investment in:
  - High-performance collective ops.
  - Local memory system performance.
  - Low latency cooperative threading.
- Detects/measures variances from bitwise reproducibility.
- Executes kernels at several (tunable) granularities:
  - nx = ny = nz = 104 gives
  - nlocal = 1,124,864; 140,608; 17,576; 2,197
  - ComputeSymGS with multicoloring adds one more level:
    - 8 colors.
    - Average size of color = 275.
    - Size ratio (largest:smallest): 4096
  - Provide a "natural" incentive to run a big problem.
- Full performance discussion:
  - http://www.hpcg-benchmark.org -> "Performance Overview" tab.

TOP500: HPCG vs. HPL

HPCG Benchmark

The High Performance Conjugate Gradients (HPCG) Benchmark project is an effort to create a new metric for ranking HPC systems. HPCG is intended as a complement to the High Performance LINPACK (HPL) benchmark, currently used to rank the TOP500 computing systems. The computational and data access patterns of HPL are still representative of some important scalable applications, but not all. HPCG is designed to exercise computational and data access patterns that more closely match a different and broad set of important applications, and to give incentive to computer system designers to invest in capabilities that will have impact on the collective performance of these applications.

HPCG is a complete, stand-alone code that measures the performance of basic operations in a unified code:

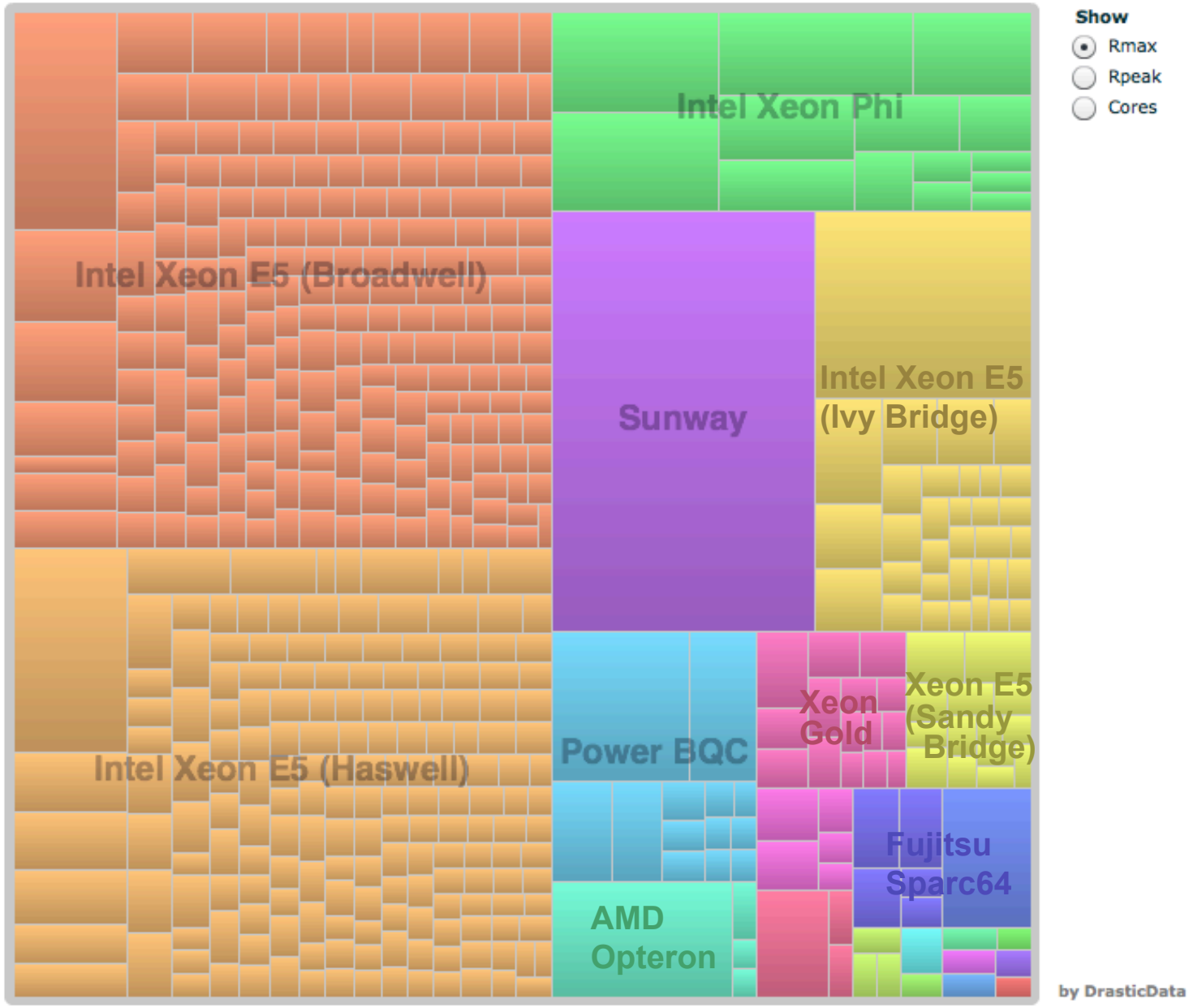
- Sparse matrix-vector multiplication.
- Vector updates.
- Global dot products.
- Local symmetric Gauss-Seidel smoother.
- Sparse triangular solve (as part of the Gauss-Seidel smoother).
- Driven by multigrid preconditioned conjugate gradient algorithm that exercises the key kernels on a nested set of coarse grids.
- Reference implementation is written in C++ with MPI and OpenMP support.

http://www.hpcg-benchmark.org

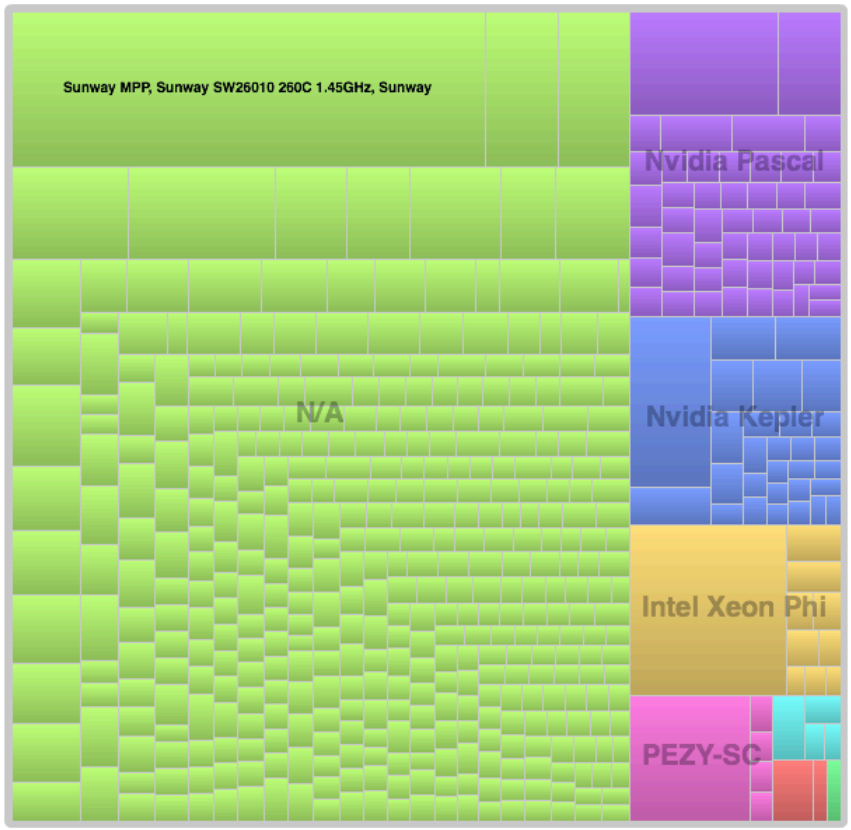
Current top 10 HPC systems using HPCG instead of HPL

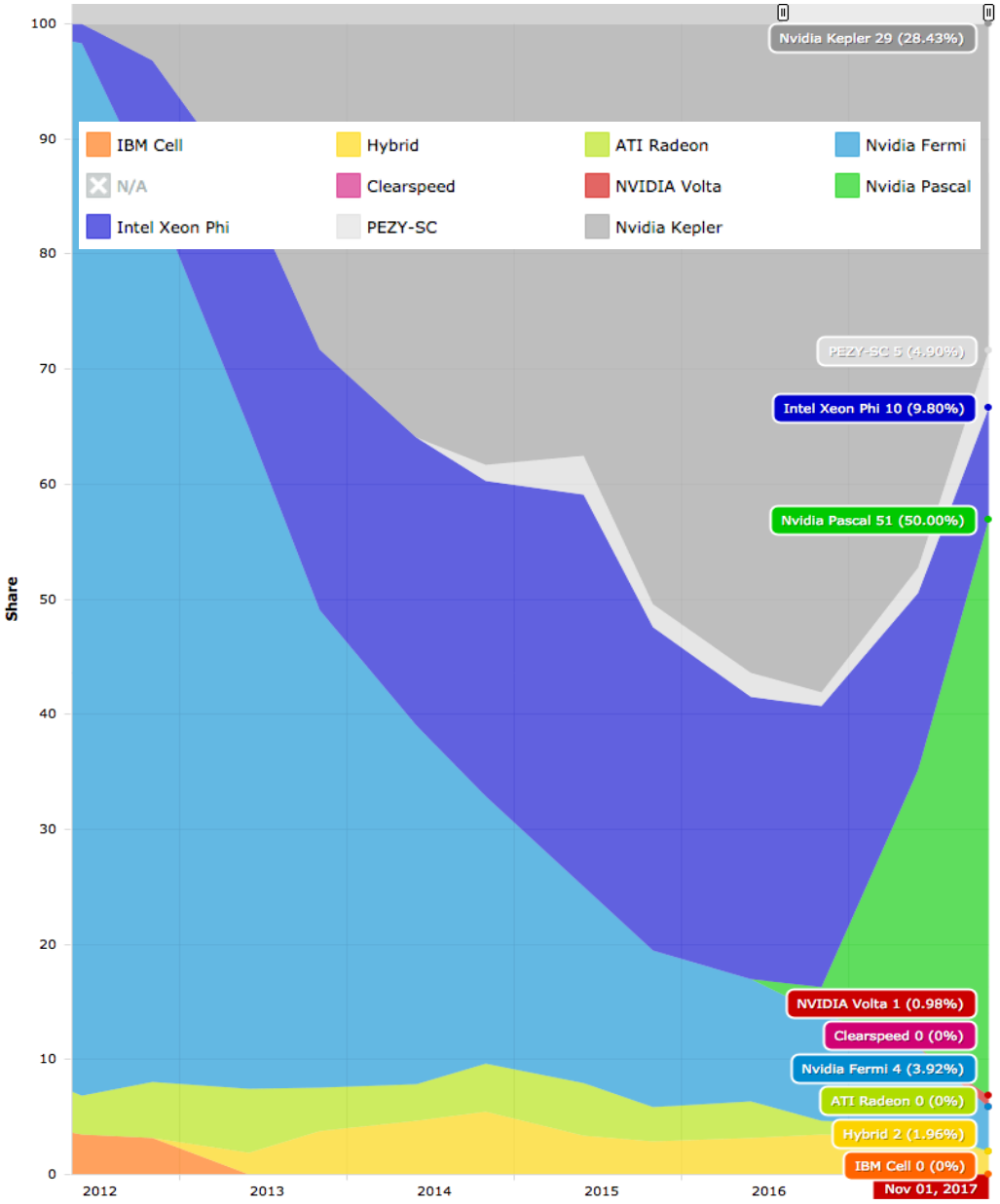
Rank	TOP500 Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	HPCG (TFlop/s)
1	10	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect , Fujitsu RIKEN Advanced Institute for Computational Science (AICS) Japan	705,024	10,510.0	11,280.4	602.736
2	2	**Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P , NUDT National Super Computer Center in Guangzhou China	3,120,000	33,862.7	54,902.4	580.109
3	7	**Trinity** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , Cray Inc. DOE/NNSA/LANL/SNL United States	979,968	14,137.3	43,902.6	546.124
4	3	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	361,760	19,590.0	25,326.3	486.398
5	1	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	480.8
6	9	**Oakforest-PACS** - PRIMERGY CX1640 M1, Intel Xeon Phi 7250 68C 1.4GHz, Intel Omni-Path , Fujitsu Joint Center for Advanced High Performance Computing Japan	556,104	13,554.6	24,913.5	385.479
7	8	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , Cray Inc. DOE/SC/LBNL/NERSC United States	622,336	14,014.7	27,880.7	355.442
8	6	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom , IBM DOE/NNSA/LLNL United States	1,572,864	17,173.2	20,132.7	330.373
9	5	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x , Cray Inc. DOE/SC/Oak Ridge National Laboratory United States	560,640	17,590.0	27,112.5	322.322
10	13	**TSUBAME3.0** - SGI ICE XA, IP139-SXM2, Xeon E5-2680v4 14C 2.4GHz, Intel Omni-Path, NVIDIA Tesla P100 SXM2 , HPE GSIC Center, Tokyo Institute of Technology Japan	135,828	8,125.0	12,127.1	188.6

Processor generations in November'17



Accelerator family distribution over all systems
Nov'17





Next PU generations for HPC: Intel® Xeon® Processor Scalable Family

(formerly code-named Skylake-SP)

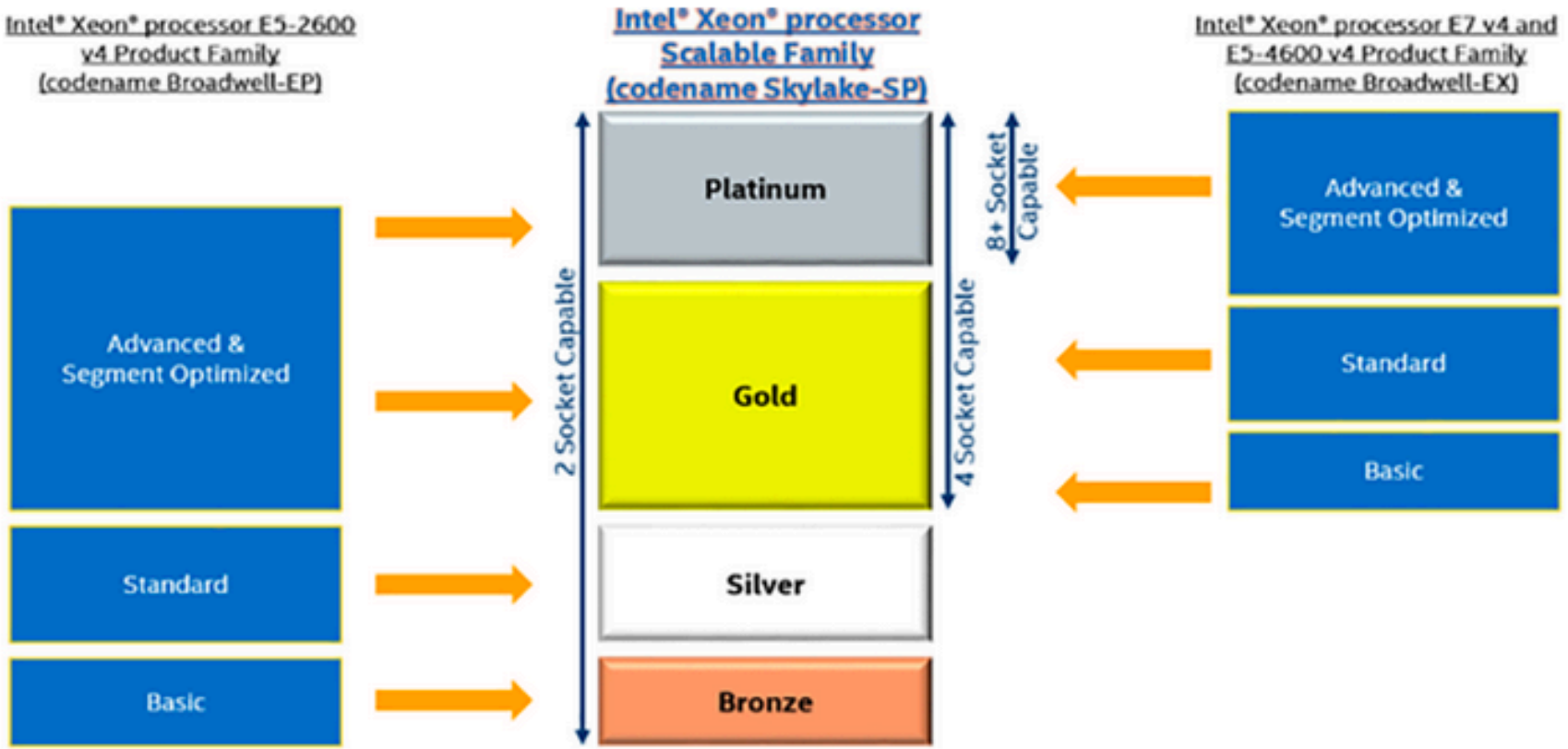


Figure 2. New branding for processor models.

https://software.intel.com/en-us/articles/intel-xeon-processor-scalable-family-technical-overview

Next PU generations for HPC: Intel® Xeon® Processor Scalable Family

(formerly code-named Skylake-SP)

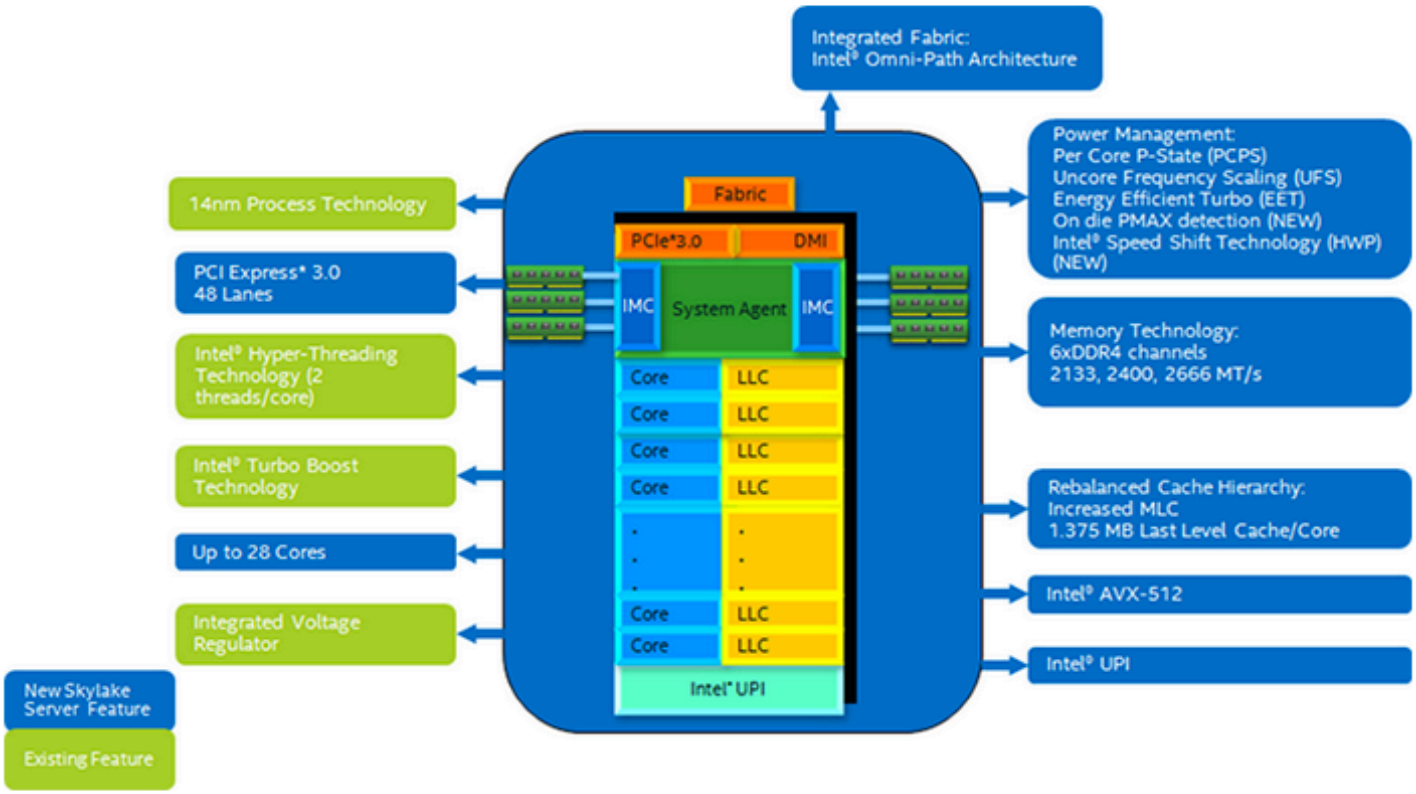


Figure 3. Block Diagram of the Intel® Xeon® processor Scalable family microarchitecture.

https://software.intel.com/en-us/articles/intel-xeon-processor-scalable-family-technical-overview

Next PU generations for HPC: Intel® Xeon® Processor Scalable Family

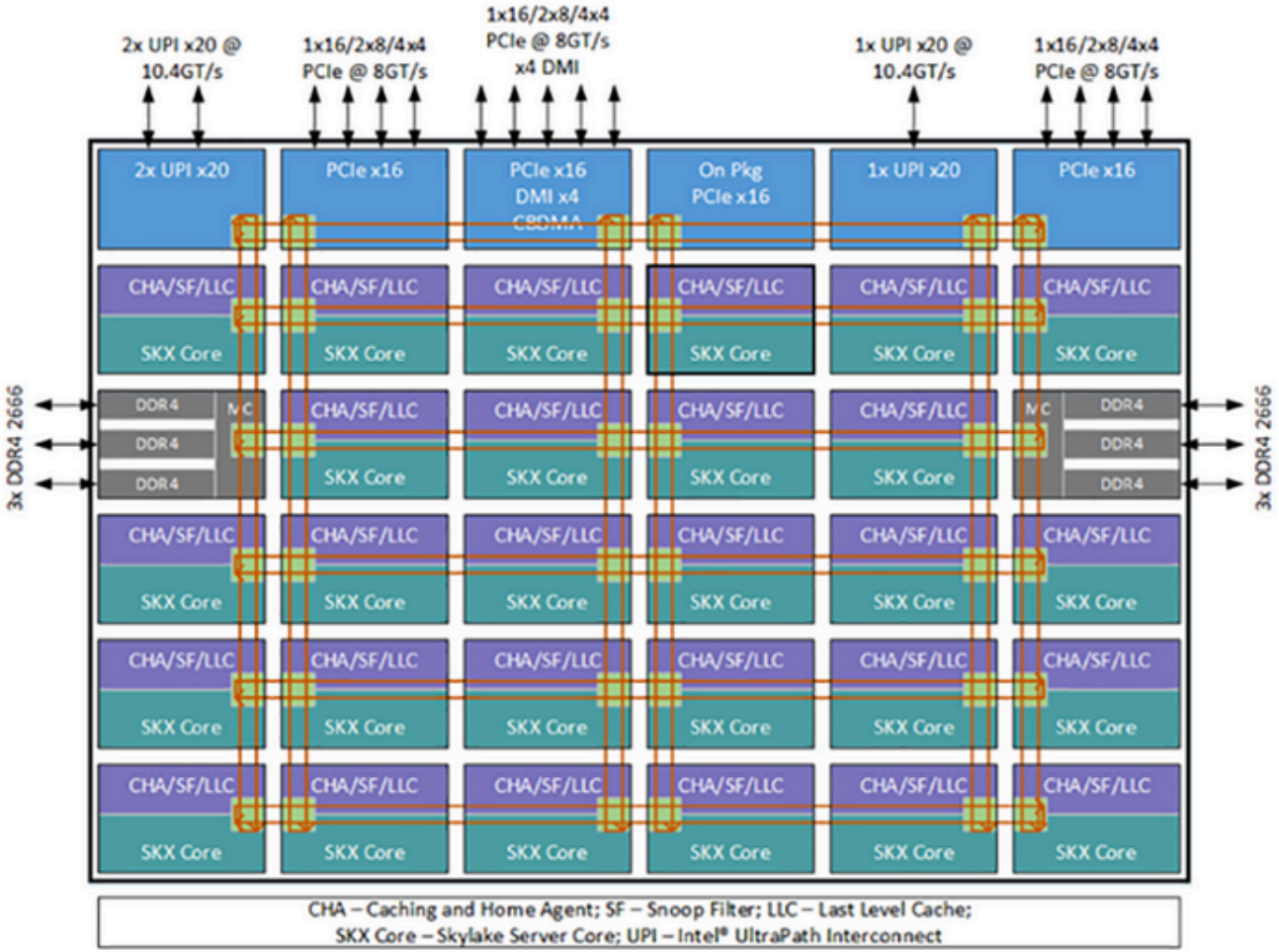


Figure 5. Intel Xeon processor Scalable family mesh architecture.

https://software.intel.com/en-us/articles/intel-xeon-processor-scalable-family-technical-overview

Next PU generations for HPC: Intel® Xeon® Processor Scalable Family

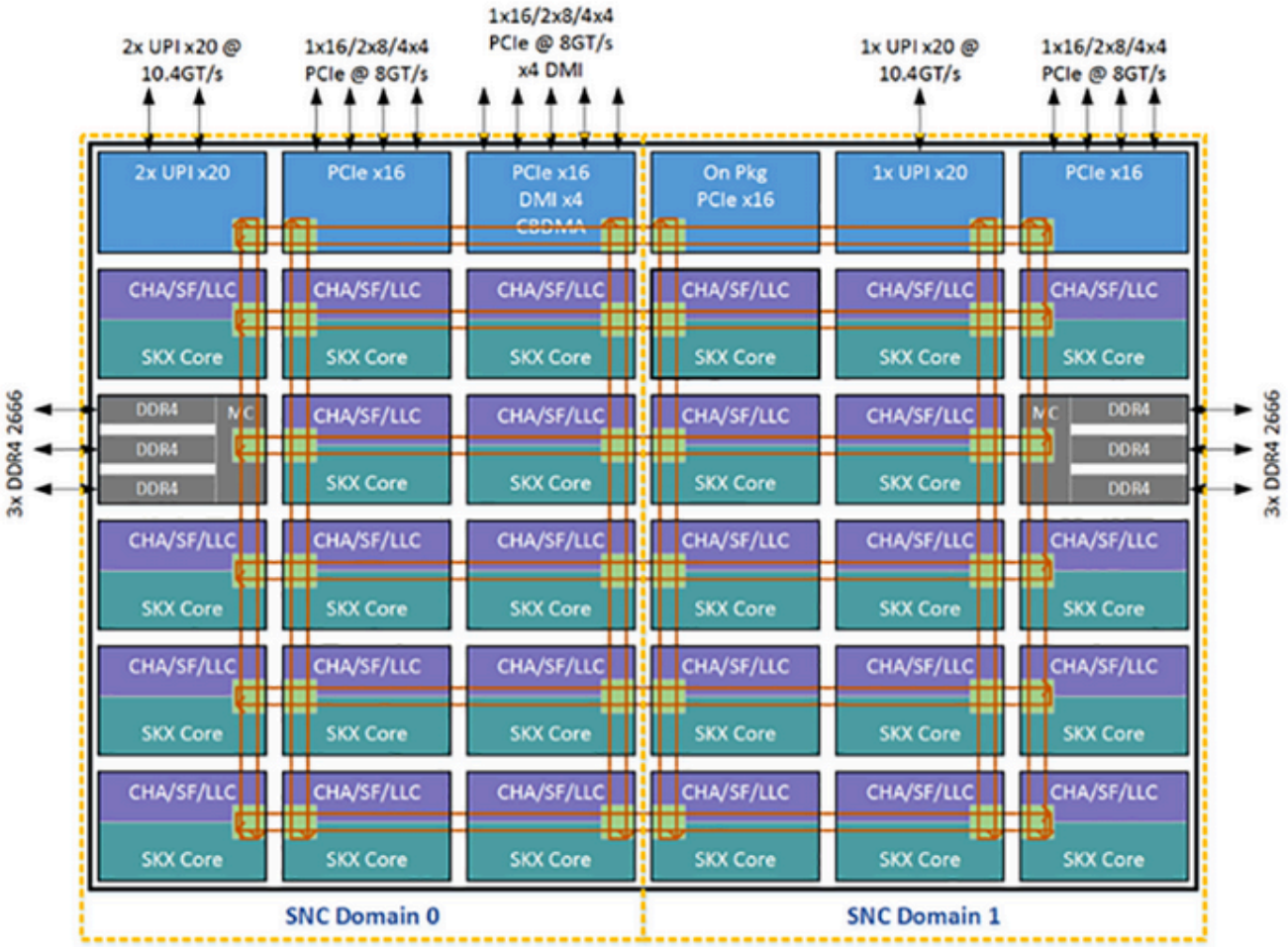


Figure 10. Sub-NUMA cluster domains.

https://software.intel.com/en-us/articles/intel-xeon-processor-scalable-family-technical-overview

Next PU generations for HPC: Intel® Xeon® Processor Scalable Family

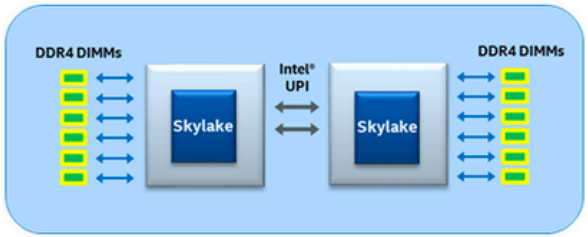


Figure 6. Typical two- socket configuration.

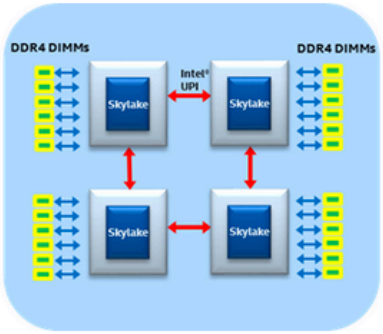


Figure 7. Typical four-socket ring configuration.

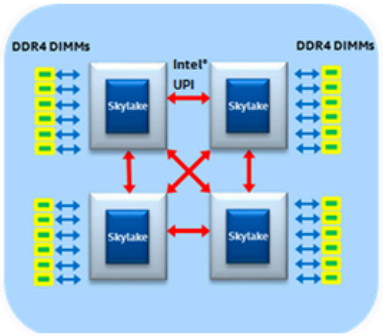


Figure 8. Typical four-socket crossbar configuration.

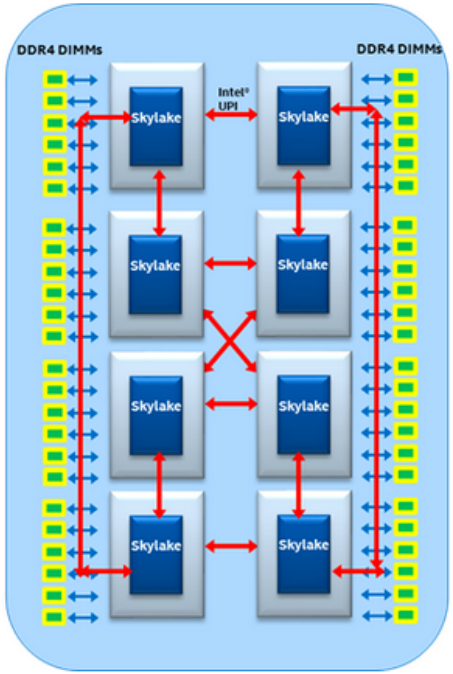


Figure 9. Typical eight-socket configuration.

https://software.intel.com/en-us/articles/intel-xeon-processor-scalable-family-technical-overview

Next PU generations for HPC: Intel® Xeon® Processor Scalable Family

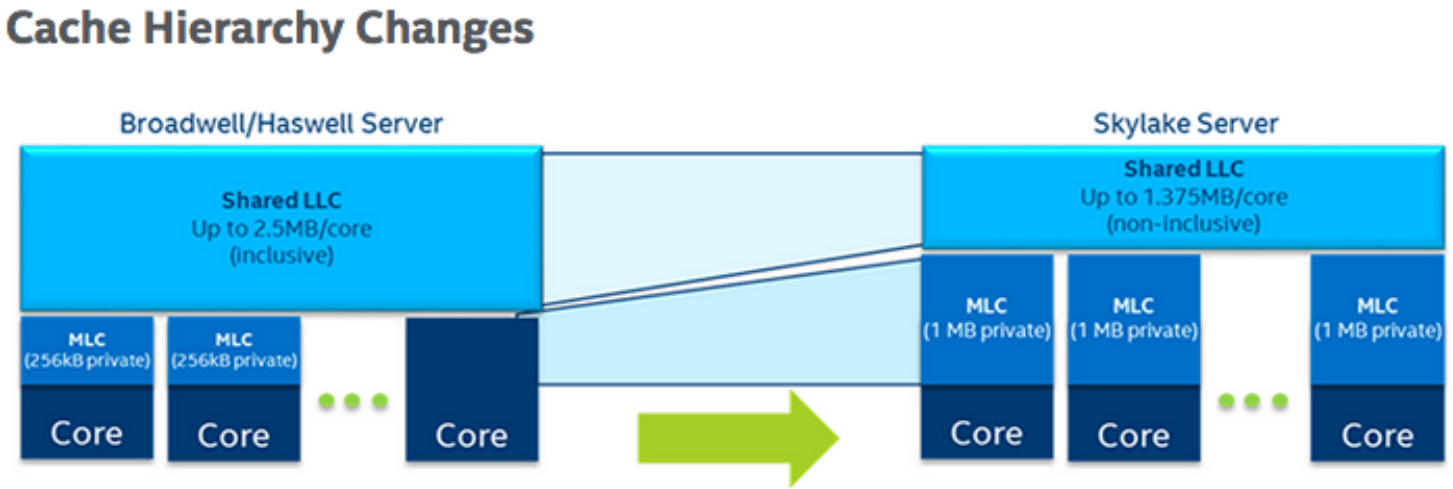


Figure 11. Generational cache comparison.

https://software.intel.com/en-us/articles/intel-xeon-processor-scalable-family-technical-overview

Next PU generations for HPC: Fujitsu SPARC64 & ARMv8

Post-K: Powered by Fujitsu-designed CPU & Tofu

FUJITSU

- Fujitsu CPU cores support the ARM SVE instruction set architecture
- Post-K Fujitsu CPU cores & Tofu maintain the programming models and provide high application performance
- ARM's standard frameworks (SBSA, etc.) assure compatibility among platforms
- FP16 ("giant vector throughput") for supercomputers

	Functions & archtecture	Post-K	FX100	FX10	K
CPU Core	Instruction set architecture	ARMv8-A	SPARC V9		
	SIMD width	512bit	256bit	128bit	128bit
	Double precision (64bit)	✔	✔	✔	✔
	Single precision (32bit)	✔	✔	✔	✔
	Half precision (16bit)	✔	-	-	-
Interconnect	Tofu interconnect	Enhanced	Tofu2	Tofu	Tofu

Next PU generations for HPC: Cray & BSC are betting on ThunderX2 (ARMv8)

CRAY

NEWS RELEASE

CRAY CATAPULTS ARM-BASED PROCESSORS INTO SUPERCOMPUTING

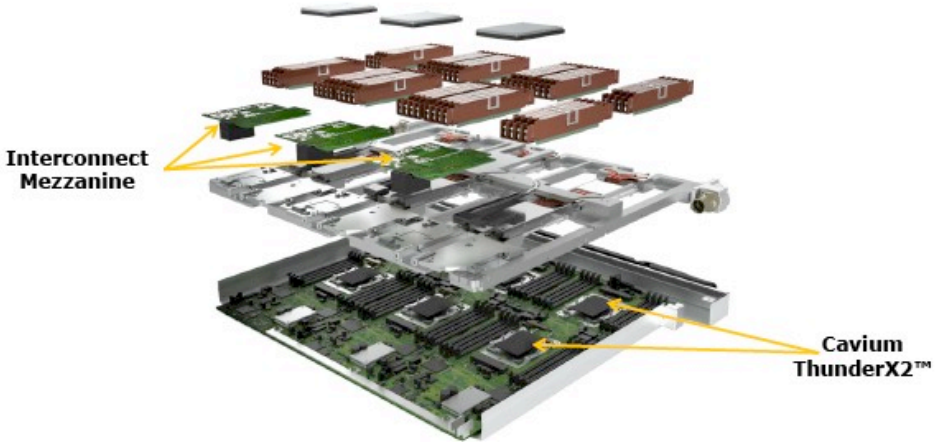
Cray Adds Arm Processors with Complete Software Stack to the Cray XC50 Supercomputer

SEATTLE, Nov. 13, 2017 (GLOBE NEWSWIRE) -- Global supercomputer leader Cray Inc. (Nasdaq:CRAY) today announced the Company is creating the world's first production-ready, Arm®-based supercomputer with the addition of Cavium (Nasdaq:CAVM) ThunderX2™ processors, based on 64-bit Armv8-A architecture, to the Cray® XC50™ supercomputer. Cray customers will have a complete Arm-based supercomputer that features a full software environment, including the Cray Linux Environment, the Cray Programming Environment, and Arm-optimized compilers, libraries, and tools for running today's supercomputing workloads.

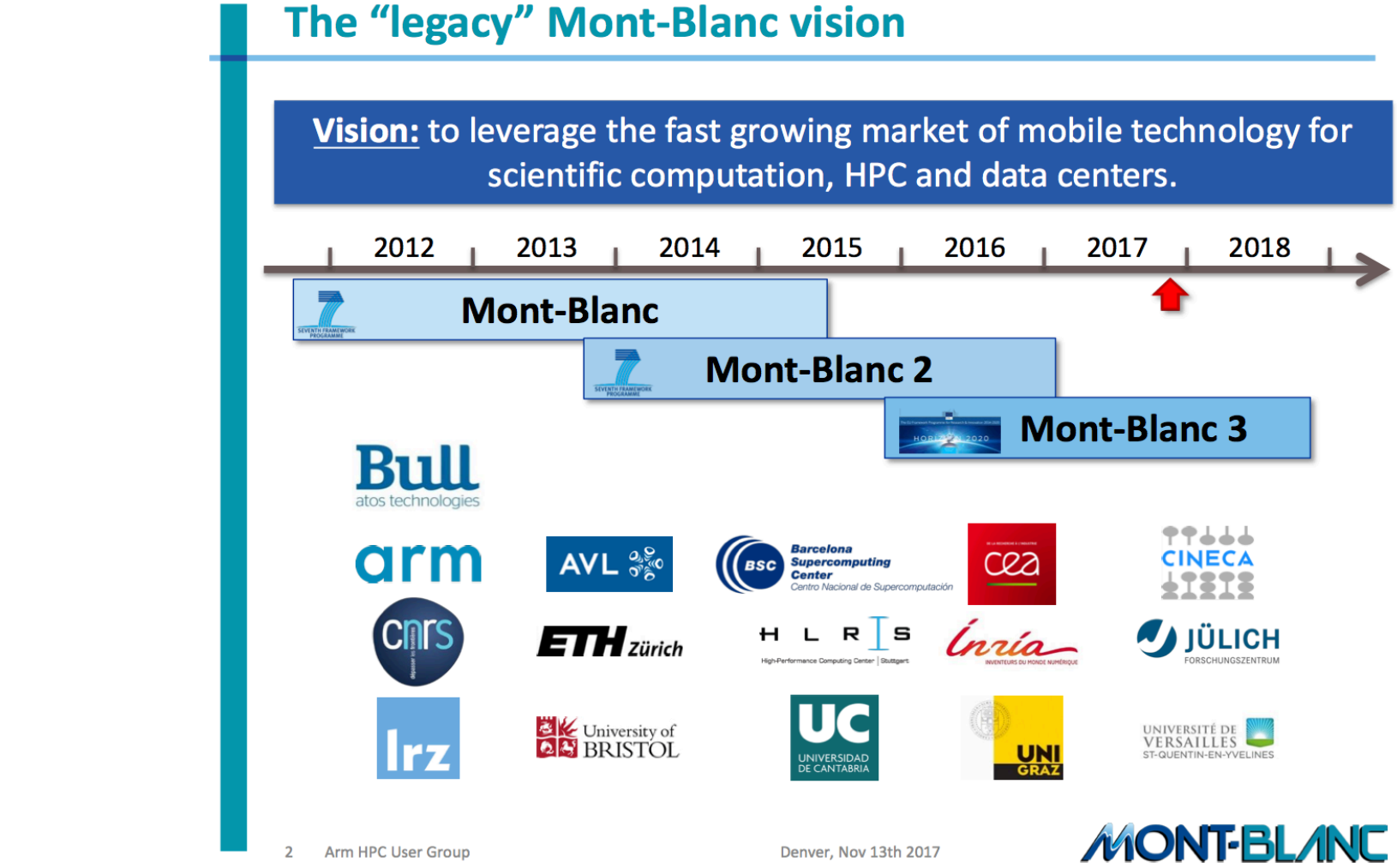
Bull sequana compute blade: X1310

Cavium ThunderX2™ – ARMv8 processor

MONT BLANC



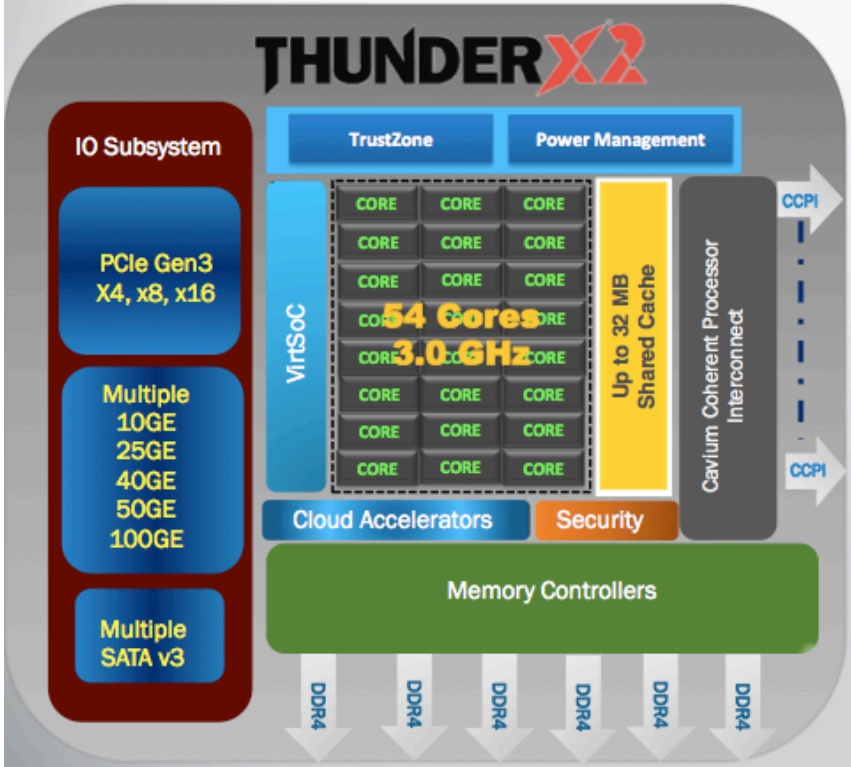
Next PU generations for HPC: Mont-Blanc roadmap at BSC
(ARMv8)



Next PU generations for HPC: Cavium ThunderX2 (ARMv8)

THUNDERX2 Product Family

CAVIUM



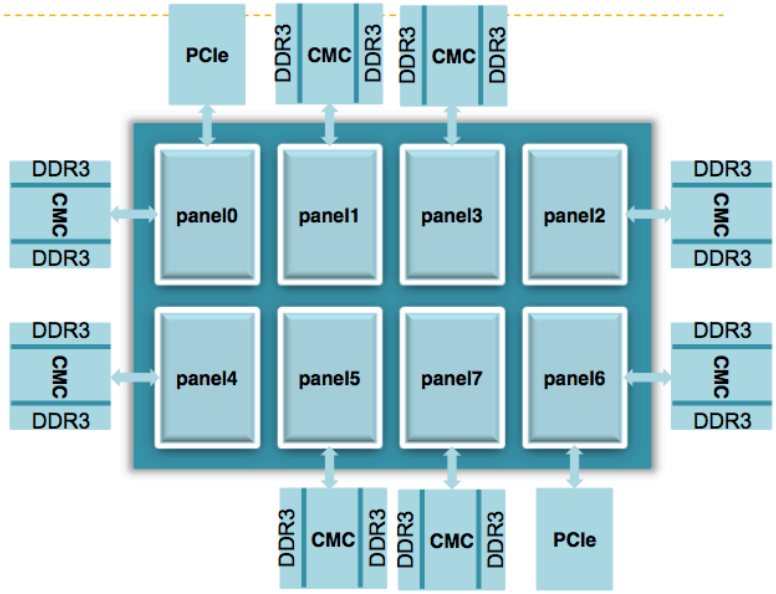
Delivers up to 2-3X the performance

- Up to 54 custom ARM cores @ 3.0GHz
- 64K-I cache and 40K-D cache
- 1S and 2S configuration
- Up to 6 DDR4 Memory Controllers
- Improved RAS features
- Integrated Ethernet with 25G KR ports
  - Supporting 10/25/40/50/100GE
- Integrated Cloud Accelerators
- Integrated SATAv3, PCIe Gen3
- Integrated Boot, Legacy IO, Manageability Engine

Next PU generations for HPC: Chinese HPC-chip, Phytium Mars (ARMv8)

Mars Overview

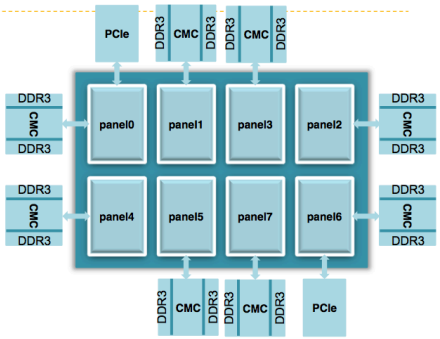
- Architecture Features
  - 64 Xiaomi cores, ARMv8 compatible
  - Hardware-maintained global cache coherency
  - Panel-based data affinity architecture
  - Mesh topology on chip network
  - 32MB L2 cache
  - 8 Cache & Memory Chips (CMC)
    - 128MB L3 cache
    - 16 DDR3-1600 channels
  - Two 16-lane PCIE3.0 i/f
  - ECC and parity protection on all caches, tags and TLBs



Physical
- ~180M instances
- 2.0GHz@28nm
- 120W

Performance
- Peak：512GFLOPS
- Mem BW：204GB/s
- I/O BW： 32GB/s

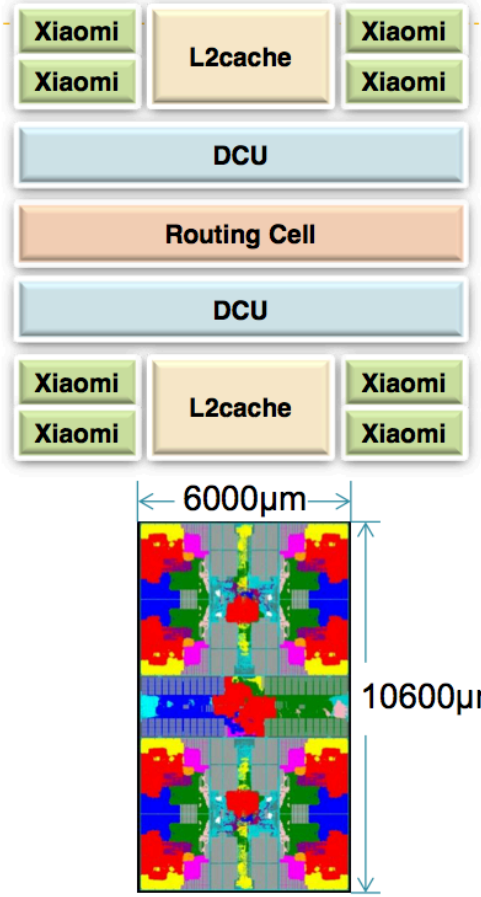
https://www.slideshare.net/insideHPC/phytium-64-core-cpu-preview?next_slideshow=1



Next PU generations for HPC: Chinese HPC-chip, Phytium Mars (ARMv8)

Panel Architecture

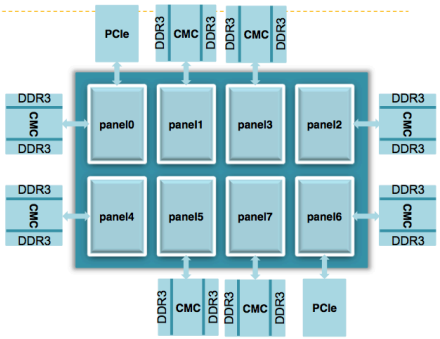
- Eight Xiaomi Cores
  - Compatible design with ARMv8 arch license
  - Both AArch32 and AArch64 modes
  - EL0~EL3 supported
  - ASIMD-128 supported
  - Adv. hybrid Branch Prediction
  - 4 fetch/4 decode/4 dispatch Out-of-Order superscalar pipeline
- Cache Hierarchy
  - Separated L1 ICache and L1 Dcache
  - Shared L2 cache, totally 4MB
- Directory-based cache coherency maintenance
  - Directory Control Unit (DCU)
- Routing Cell



7

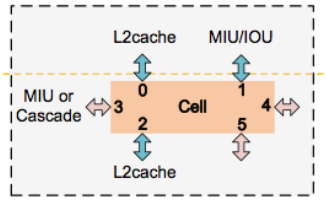
Phytium Technology Co., Ltd

https://www.slideshare.net/insideHPC/phytium-64-core-cpu-preview?next_slideshow=1

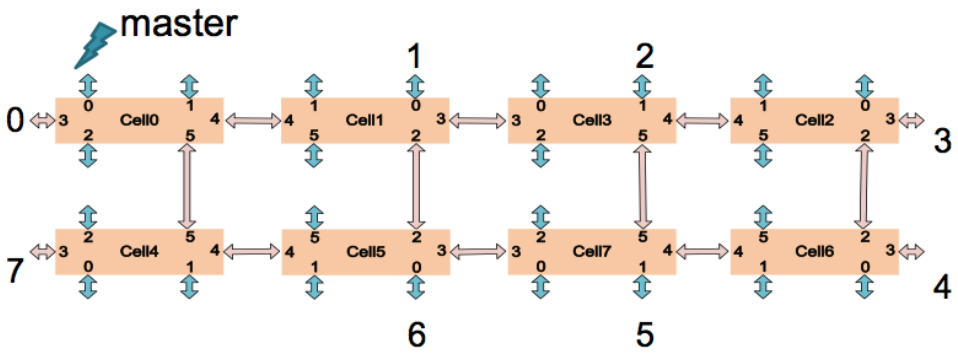


Next PU generations for HPC: Chinese HPC-chip, Phytium Mars (ARMv8)

Network on Chip



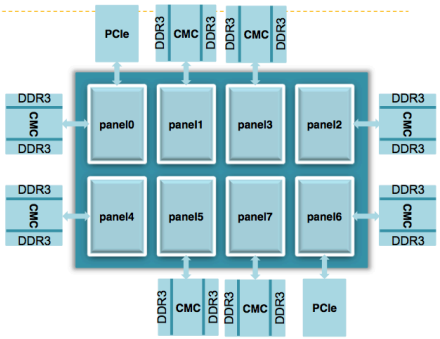
- 2D Concentrated Mesh Architecture
  - Cell based switch with 6 bidirectional ports
  - Uniform package format for each port, a port can be configured to be connected with a device or cascade cell
  - 4 physical channels for CC and 1 channel for debug, DOR Y-X routing
  - Low latency: 3 cycles for each hop
  - High bandwidth: 384GB/s each cell



Dest.	Lat. (cycles)
0	3
1	6
2	9
3	12
4	15
5	12
6	9
7	6
Avg.	9

Phytium Technology Co., Ltd

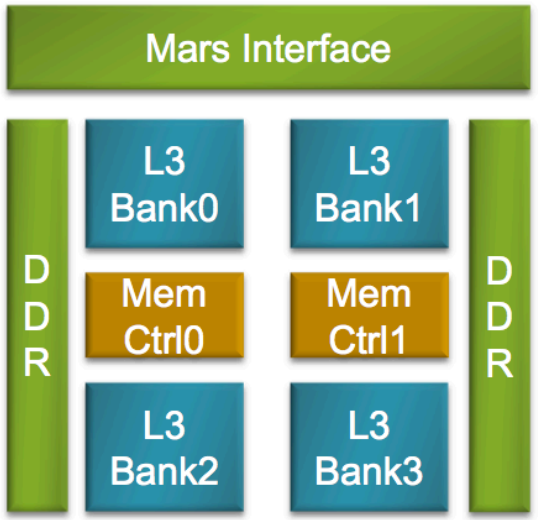
https://www.slideshare.net/insideHPC/phytium-64-core-cpu-preview?next_slideshow=1



Next PU generations for HPC: Chinese HPC-chip, Phytium Mars (ARMv8)

Cache & Memory Chip

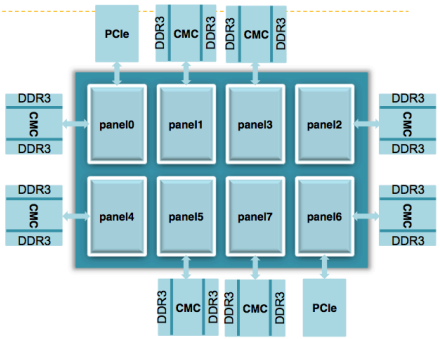
- L3 cache
  - 16MB Data Array
  - 2MB Data ECC
- DDR bandwidth
  - 2 x DDR3-800：25.6GB/s
- Proprietary interface between Mars & CMC
  - Parallel interface
    - Needs more pins, but lower latency than serdes
  - Separate write/cmd and read data channel
    - Effective read channel bandwidth：12.8GB/s
    - Effective write/cmd channel bandwidth：6.4GB/s



15

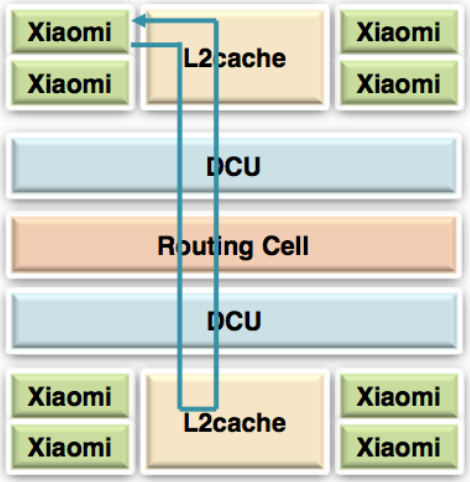
Phytium Technology Co., Ltd

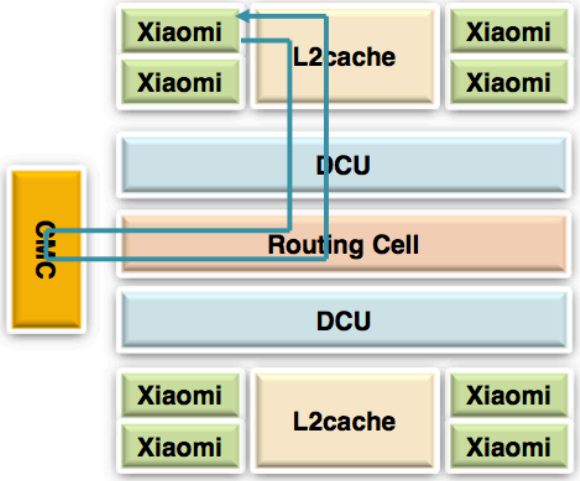
https://www.slideshare.net/insideHPC/phytium-64-core-cpu-preview?next_slideshow=1



Next PU generations for HPC: Chinese HPC-chip, Phytium Mars *(ARMv8)*

Latency of affinitive access





Memory access	latency(ns)
Local L1 cache hit	~2
Local L2 cache hit	~8
Affinitive L2 cache hit	~20
Affinitive L3 cache hit	~36
Affinitive DDR access	~70

- Panel : 2.0GHz
- NoC: 2.0GHz
- CMC: 1.5GHz

* PCB latency not considered

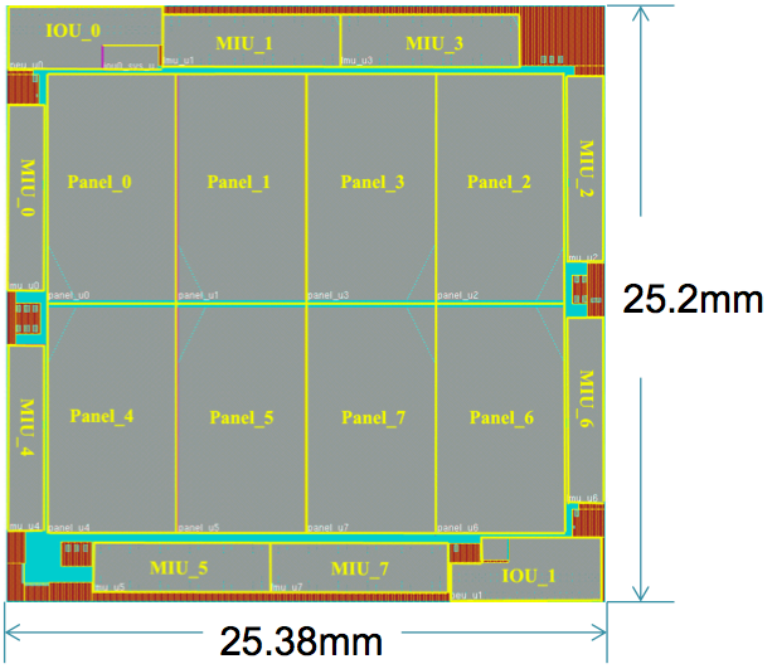
16

Phytium Technology Co., Ltd

https://www.slideshare.net/insideHPC/phytium-64-core-cpu-preview?next_slideshow=1

Next PU generations for HPC: Chinese HPC-chip, Phytium Mars (ARMv8)

Physical Design



- 28nm process
- 0.9v core/1.8v IO
- 10 metal layers
- ~180M instances
- 2.0GHz
- 120W
- 640mm^2 die size
- FCBGA
- ~3000 pins

19

Phytium Technology Co., Ltd

https://www.slideshare.net/insideHPC/phytium-64-core-cpu-preview?next_slideshow=1