Международная лаборатория суперкомпьютерного атомистического моделирования и многомасштабного анализа

Школа-семинар

«Поиск эффективных суперкомпьютерных архитектур в пост-Муровскую эру»

11 декабря 2017 МИЭМ НИУ ВШЭ

2017 год – начало активной конкуренции серверных процессоров

В.В. Стегайлов

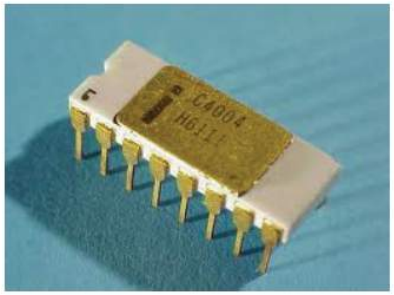
План доклада

- Проблематика пост-Муровской эры
- Новые процессоры Intel Xeon Skylake.
- Пропускная способность памяти и баланс
- Новые процессоры AMD, Qualcomm, Cavium, IBM и ускорители Nvidia
- Перспектива экзаскейла
- Опыт суперкомпьютера Theta и отмена постройки суперкомпьютера Aurora
- Японские процессоры для экзаскейла
- Заключение

Закон Мура

50 лет экспоненциального роста

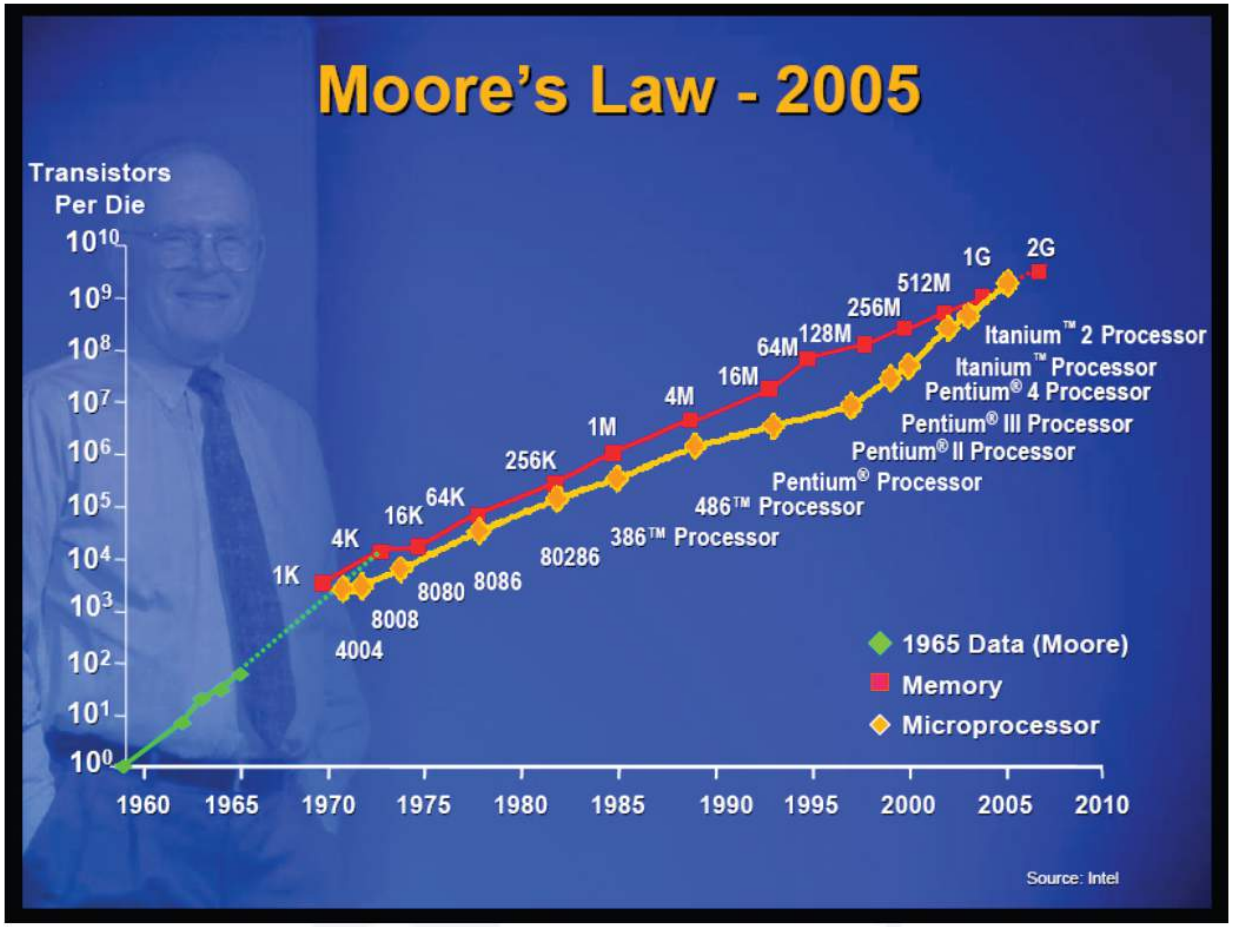
Intel 4004, 1971



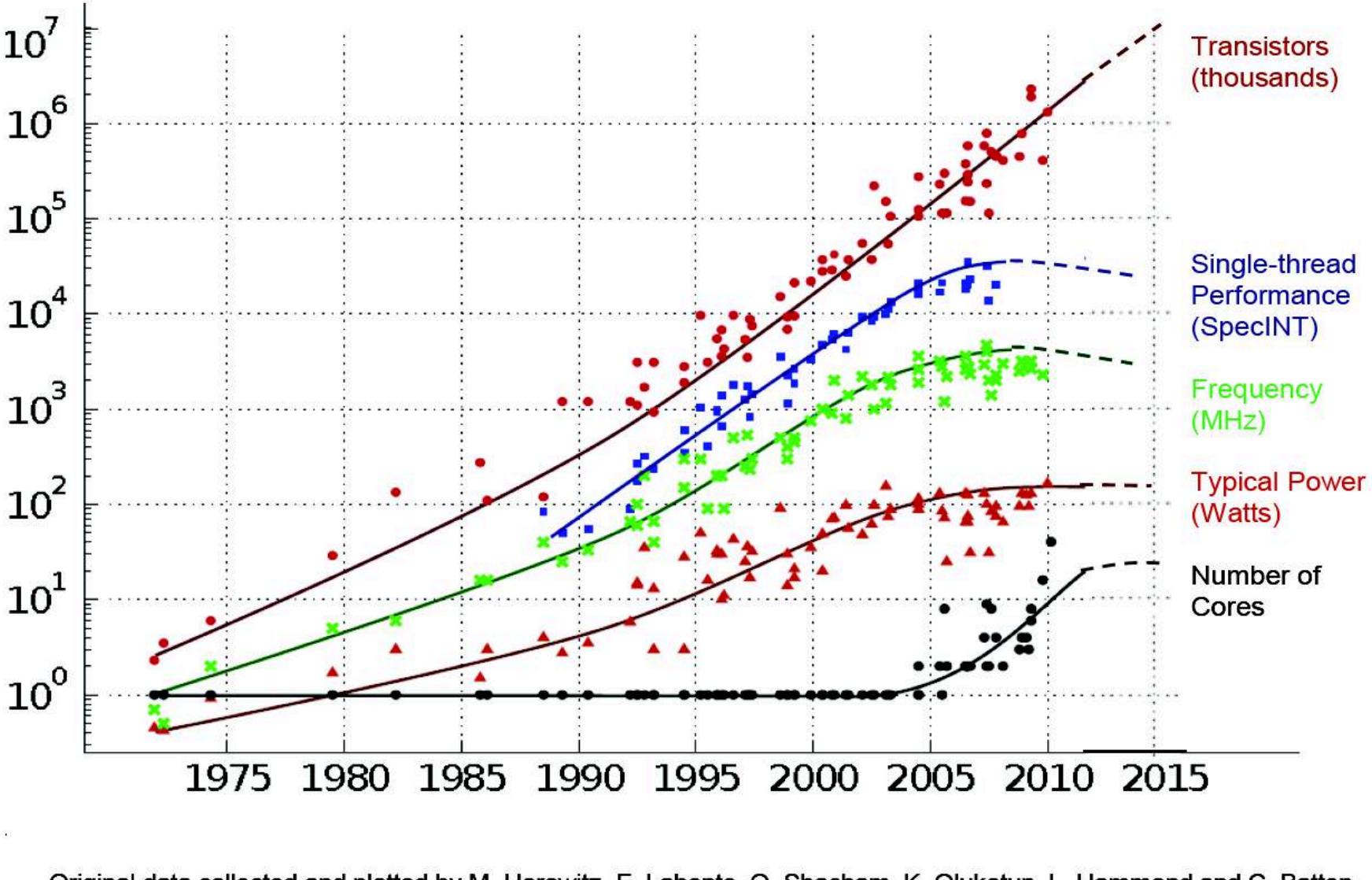
92,000 ops/sec

Intel Xeon, 2014

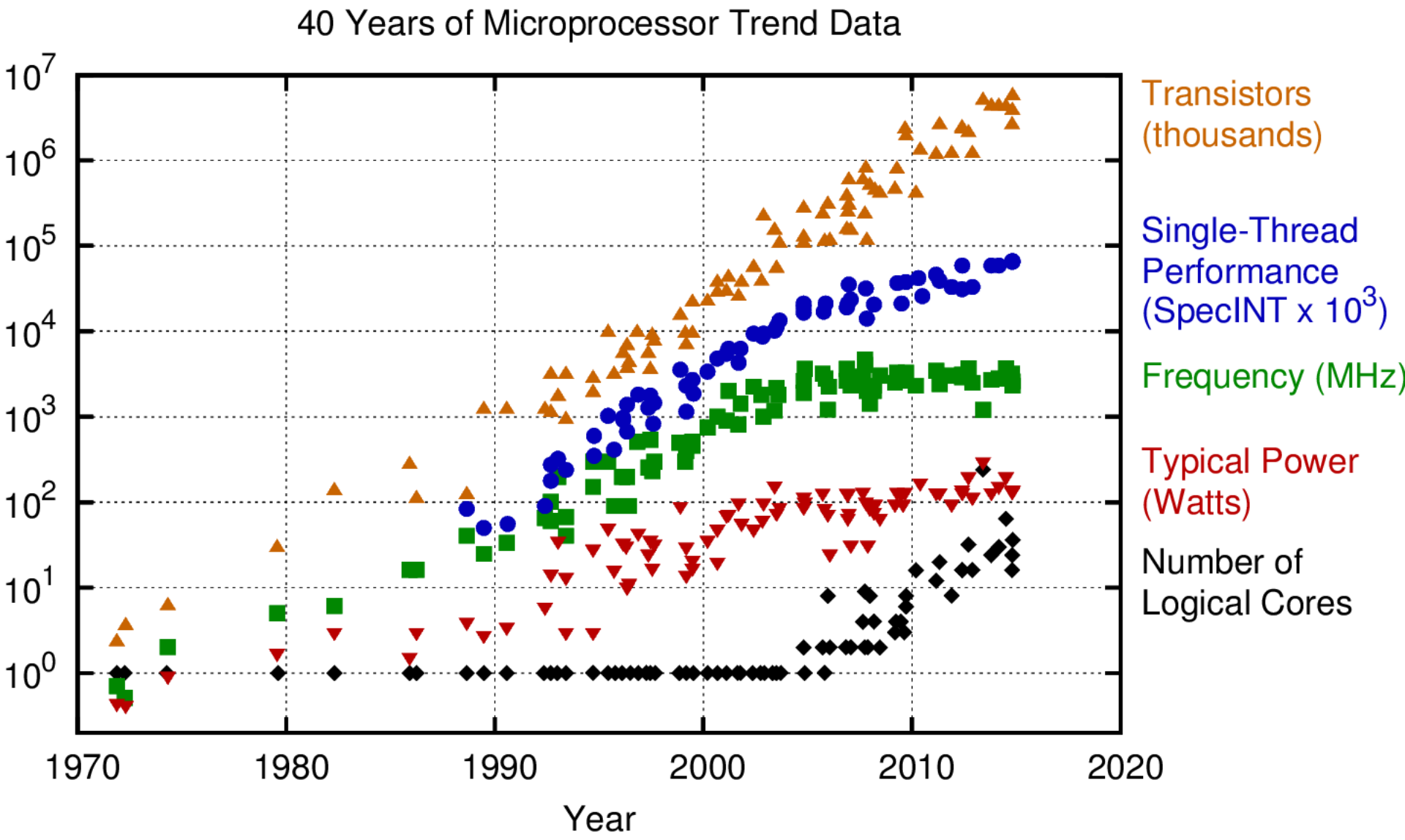
266,000,000,000 ops/sec



Дело не только в росте числа транзисторов…



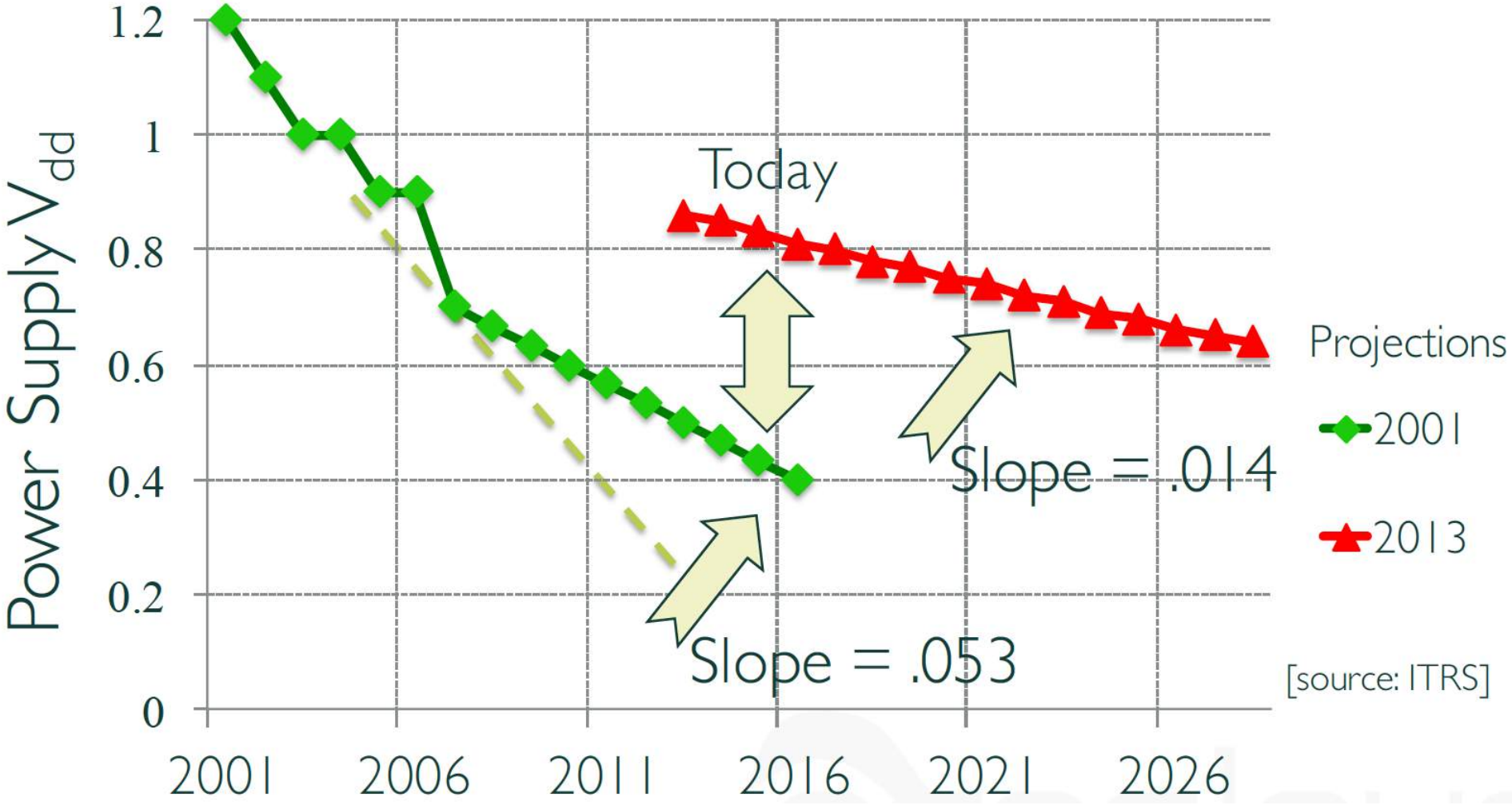
Дело не только в росте числа транзисторов…



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

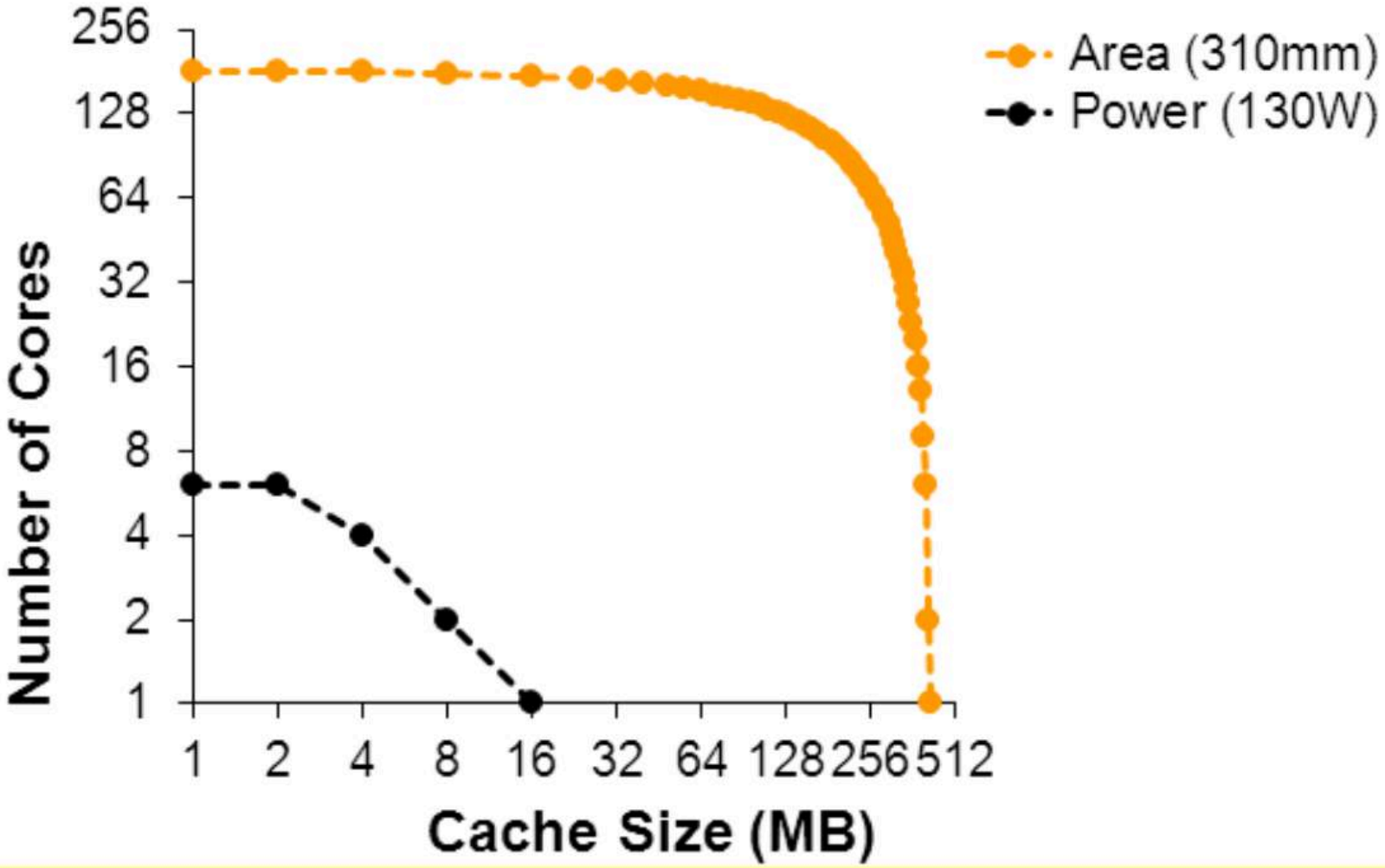
Закон масштабирования Деннарда

перестал выполняться примерно в 2006 году



“Dark silicon”

Area vs. Power Envelope



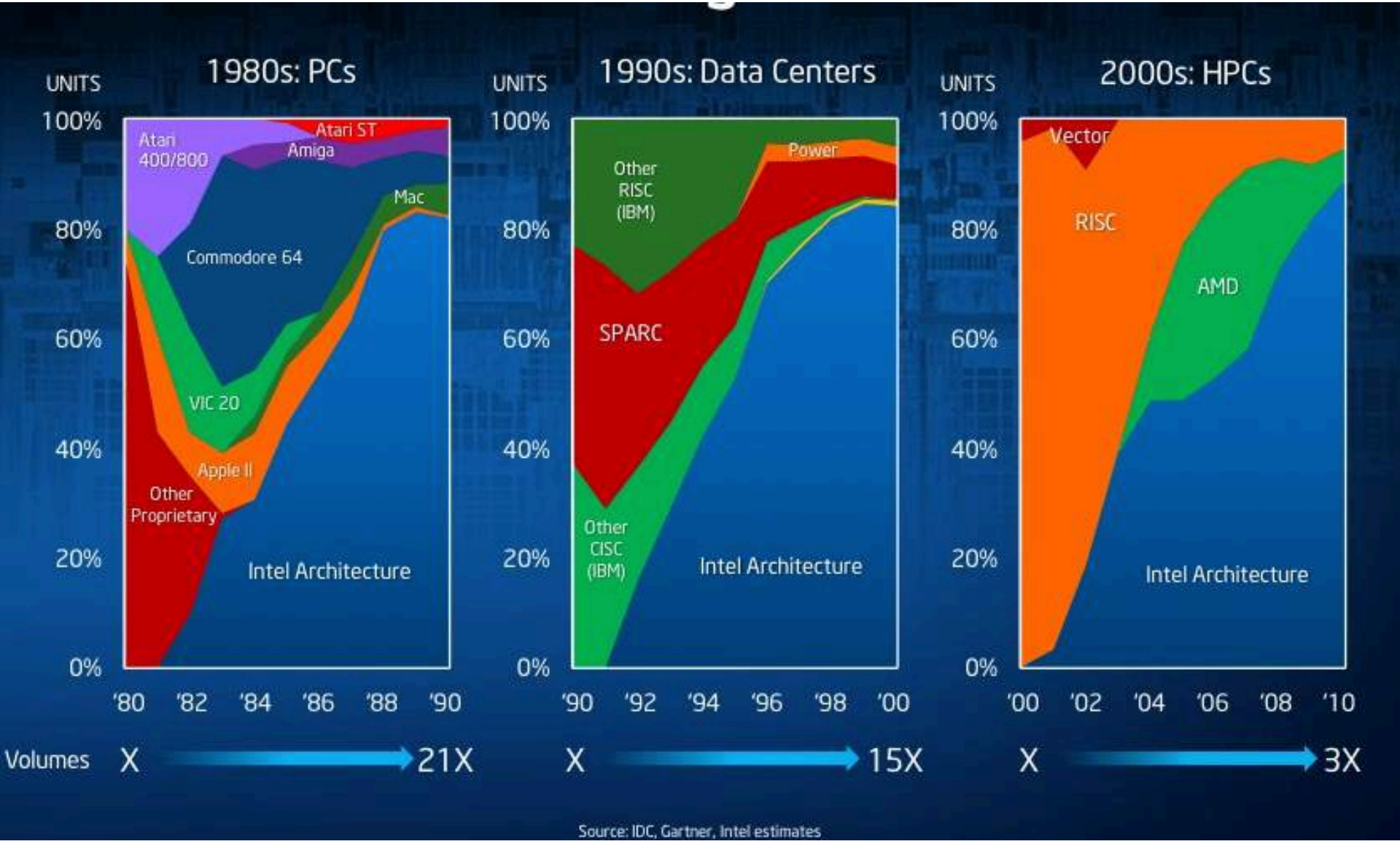
Good news: can fit 100’s cores. Bad news: cannot power them all

Dark Silicon Will Make Heterogeneity and Specialization More Relevant

Node	45nm	22nm	11nm
Year	2008	2014	2020
Area^{-1}	1	4	16
Peak freq	1	1.6	2.4
Power	1	1	0.6
		$(4 \times 1)^{-1} = 25\%$	$(16 \times 0.6)^{-1} = 10\%$

Source: ITRS 2008

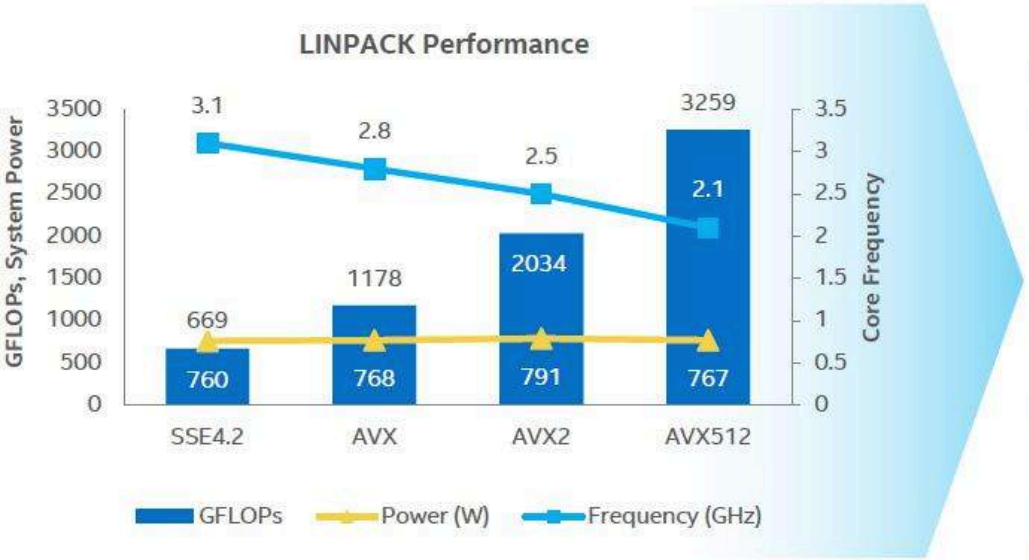
Intel Xeon Skylake



Source: IDC, Gartner, Intel estimates

Intel Xeon Skylake

Performance and Efficiency with Intel® AVX-512

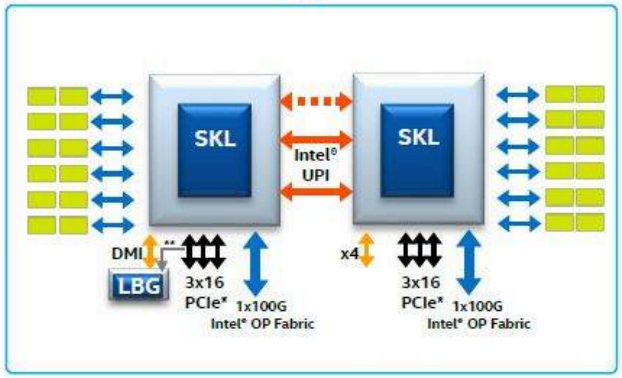


INTEL® AVX-512 DELIVERS SIGNIFICANT PERFORMANCE AND EFFICIENCY GAINS

Intel Xeon Skylake

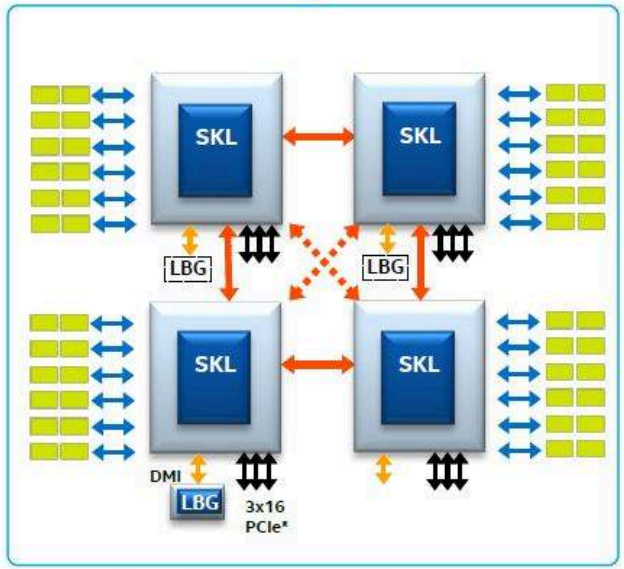
Platform Topologies

2S Configurations



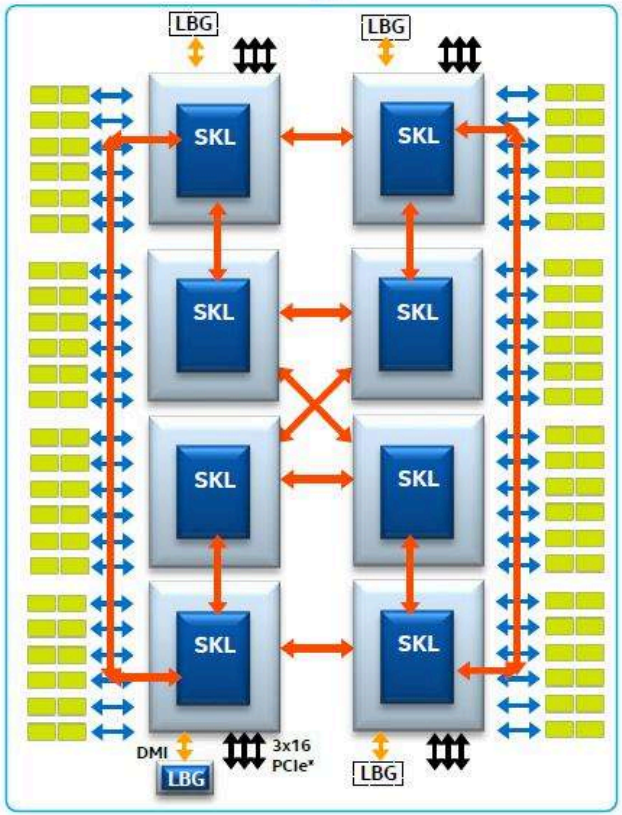
(2S-2UPI & 2S-3UPI shown)

4S Configurations



(4S-2UPI & 4S-3UPI shown)

8S Configuration



INTEL® XEON® SCALABLE PROCESSOR SUPPORTS CONFIGURATIONS RANGING FROM 2S-2UPI TO 8S

Суперкомпьютер МФТИ-60

vs Intel Xeon Skylake

7-ая редакция Топ50 от 25.09.2007

5	Москва Московский физико-технический институт (МФТИ) 2007 г.	272/544	узлов: 136 (2xXeon 5160 3 GHz 4 GB RAM) сеть: Myrinet/Gigabit Ethernet/Gigabit Ethernet	4.53	6.53	Hewlett-Packard, Институт системного программирования РАН (ИСП РАН)

Top500 от 06.2007

417	**MIPT-60** - Cluster Platform 3000 DL140G3, Xeon 51xx 3 Ghz, Myrinet , ISP RAS Moscow Institute of Physics and Technology Russia	544	4.5	6.5

Суперкомпьютер МФТИ-60

vs Intel Xeon Skylake

7-ая редакция Топ50 от 25.09.2007

5	Москва Московский физико-технический институт (МФТИ) 2007 г.	272/544	узлов: 136 (2xXeon 5160 3 GHz 4 GB RAM) сеть: Myrinet/Gigabit Ethernet/Gigabit Ethernet	4.53	6.53	Hewlett-Packard, Институт системного программирования РАН (ИСП РАН)

136 узлов
2x Xeon 5160 (2c, 3GHz) Rpeak=48 GFlops/sec
DDR2 667MHz TRIAD(4threads)=5.2 GByte/ses
SSE2

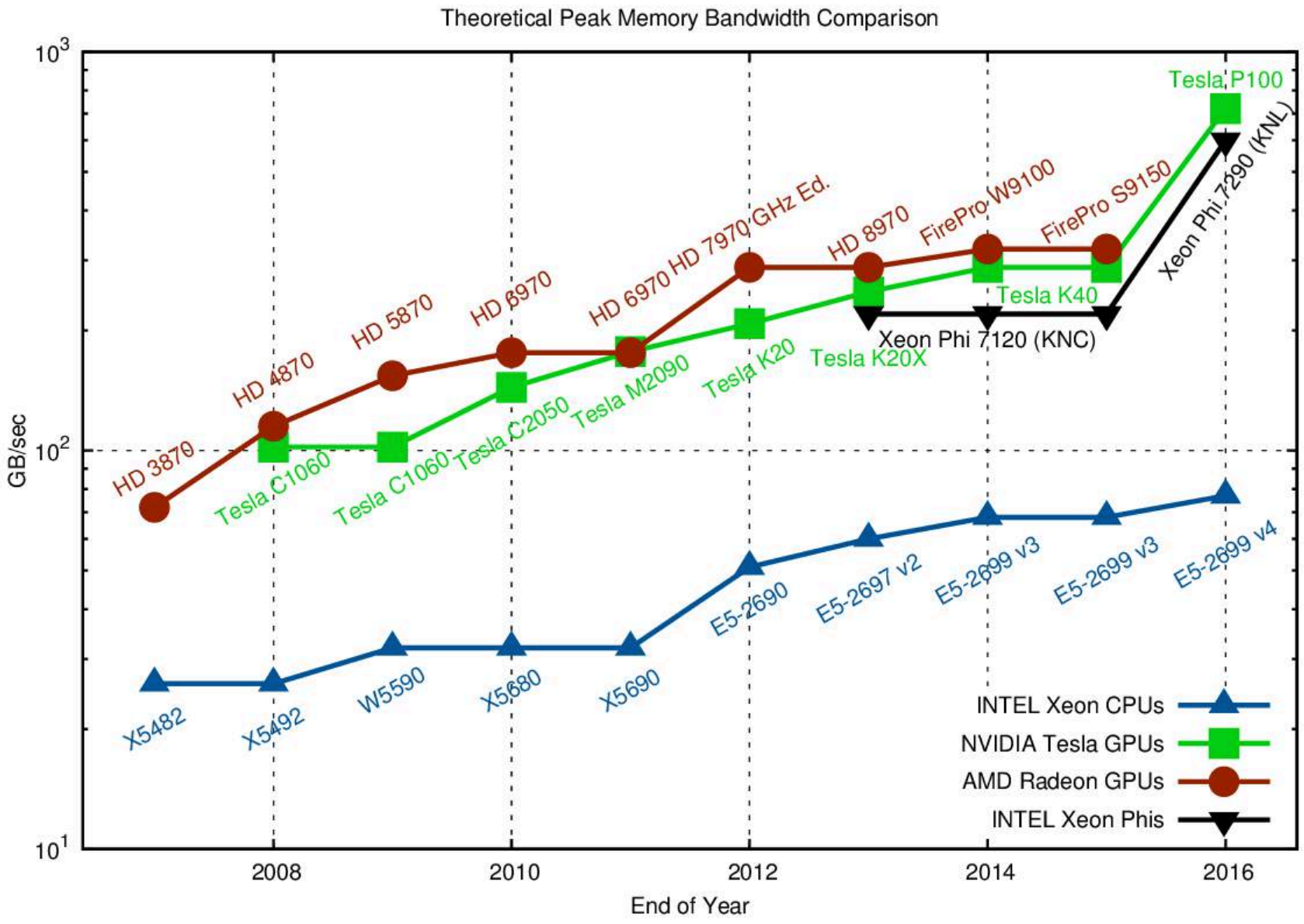
Core i7	Xeon Skylake
~70 GByte/sec	~100 GByte/sec
i7-7820x: 8c = 1 TFlops/sec	Xeon 8180: 28c = 2 TFlops/sec

AVX512

Top500 от 06.2007

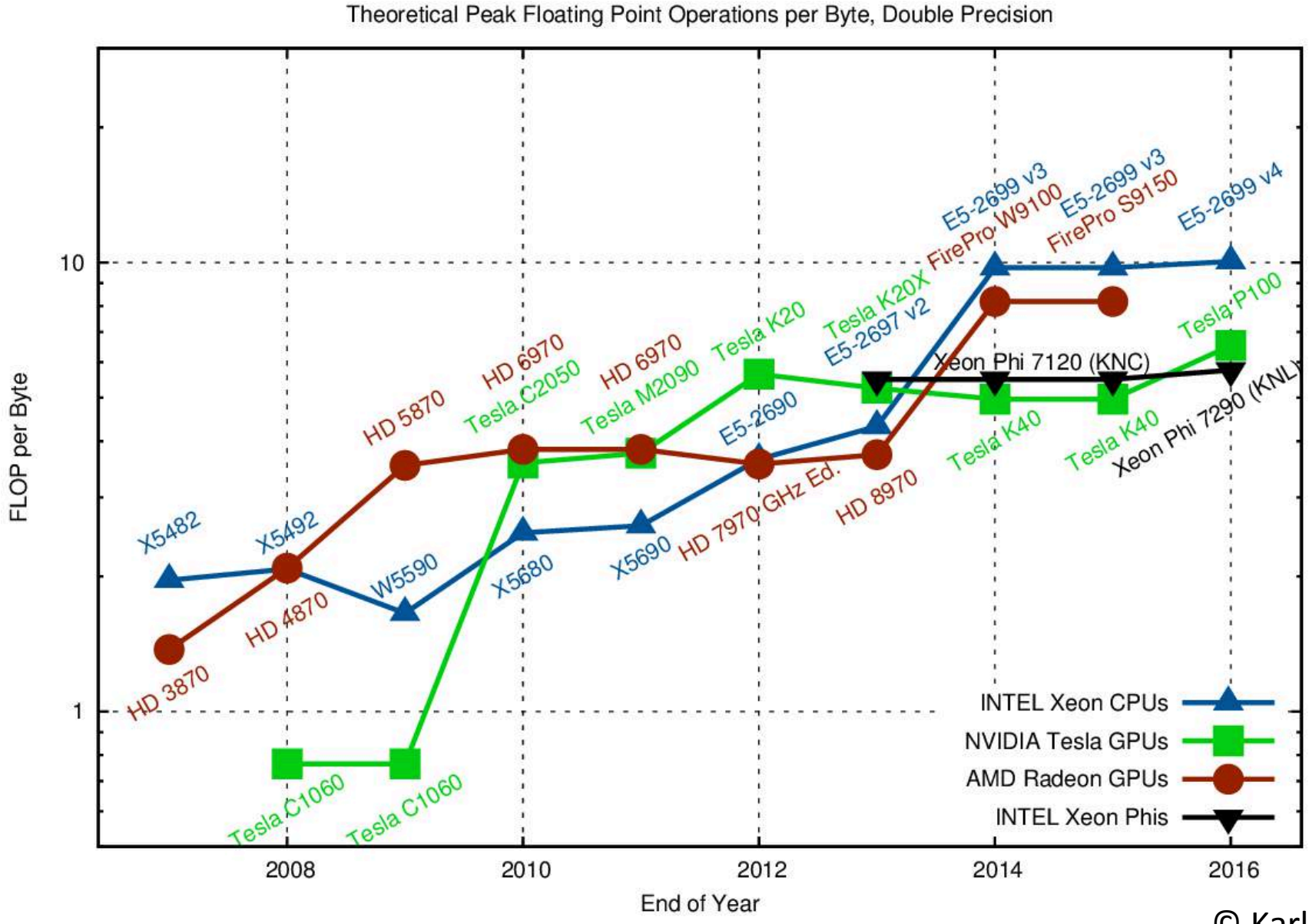
417	**MIPT-60** - Cluster Platform 3000 DL140G3, Xeon 51xx 3 Ghz, Myrinet , ISP RAS Moscow Institute of Physics and Technology Russia	544	4.5	6.5

Memory bandwidth



Balance

Flops per Byte

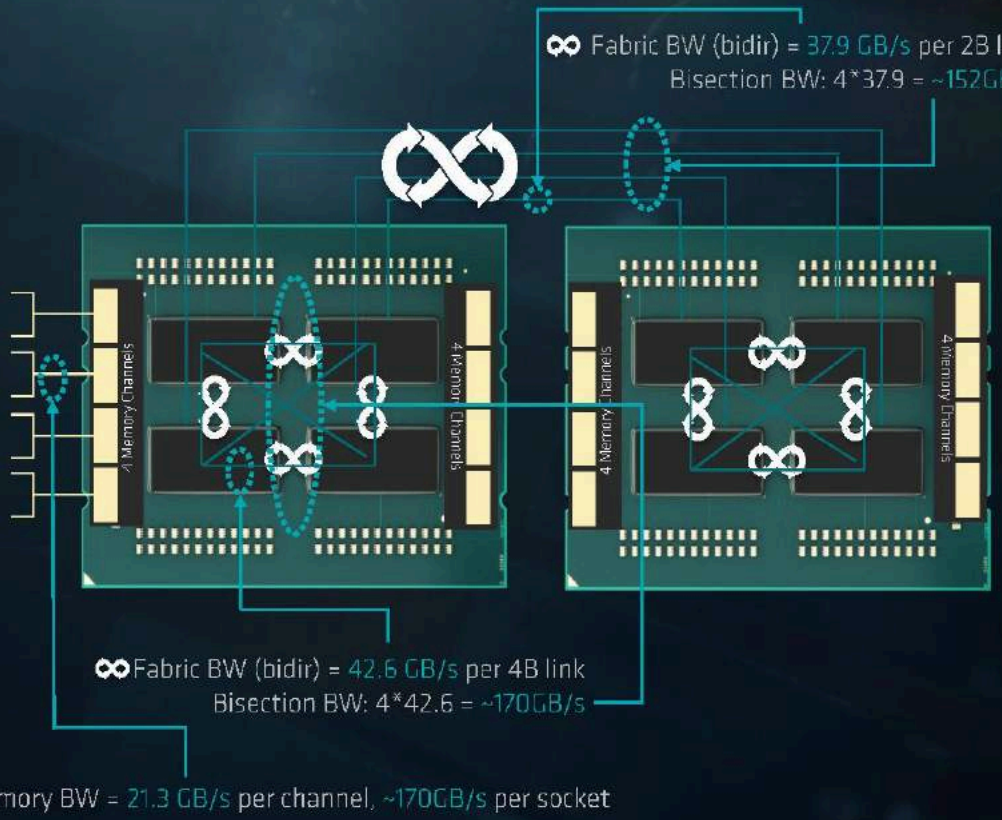


AMD Epyc

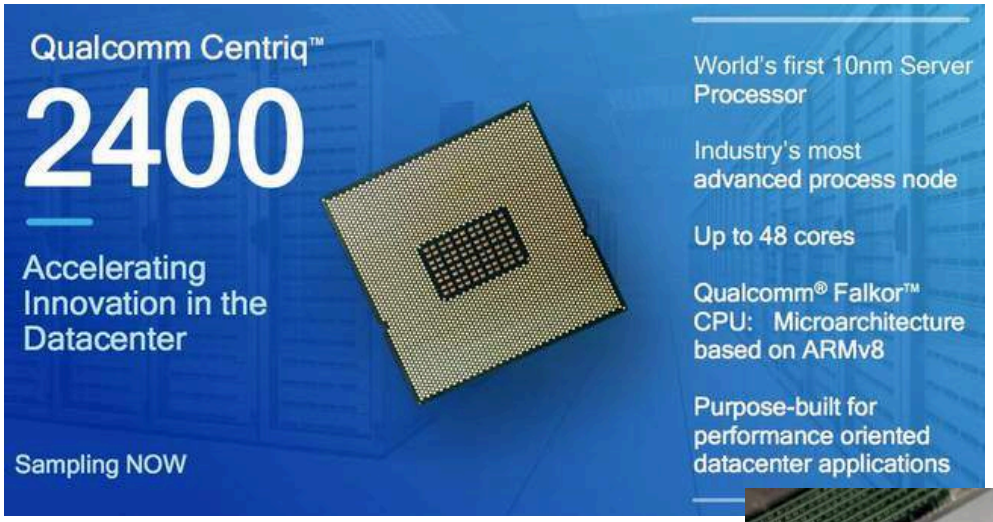
PUTTING IT TOGETHER: 2P PEAK BANDWIDTH WITH DDR4-2667

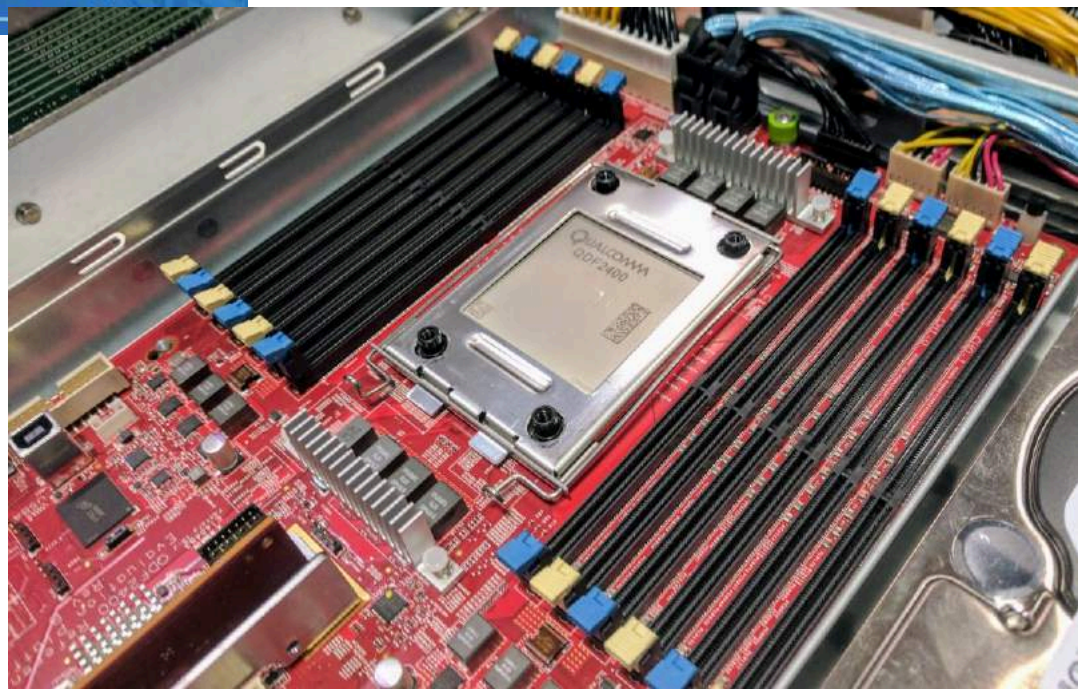
- High performance cores, I/O require strong system balance
  - Avoid bottlenecks
- Bisection bandwidth
  - 2x required within Socket
  - Matched between Sockets
- Low latency
  - Purpose built Infinity Fabric Links

EPYC™ IS BALANCED
Delivers on-die, within-socket, and cross-socket scaling



Qualcomm Centriq 2400





Microsoft's Windows Server OS runs on ARM chips

Cavium ThunderX2

Cray XC50

Microsoft Project Olympus

Bull sequana compute blade: X1310

Cavium ThunderX2™ – ARMv8 processor

ThunderX2 CPU:
32 custom cores
ARMv8 ISA
8 mem channels

IBM Power9

Power Systems

POWER9 Processor – Common Features

IBM

New Core Microarchitecture

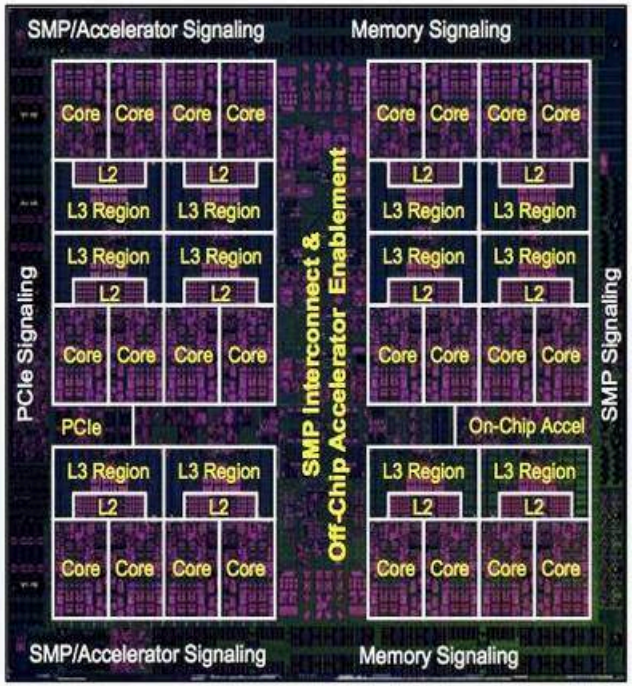
- Stronger thread performance
- Efficient agile pipeline
- POWER ISA v3.0

Enhanced Cache Hierarchy

- 120MB NUCA L3 architecture
- 12 x 20-way associative regions
- Advanced replacement policies
- Fed by 7 TB/s on-chip bandwidth

Cloud + Virtualization Innovation

- Quality of service assists
- New interrupt architecture
- Workload optimized frequency
- Hardware enforced trusted execution



14nm finFET Semiconductor Process

- Improved device performance and reduced energy
- 17 layer metal stack and eDRAM
- 8.0 billion transistors

Leadership Hardware Acceleration Platform

- Enhanced on-chip acceleration
- Nvidia NVLink 2.0: High bandwidth and advanced new features (BlueLink)
- CAPI 2.0: Coherent accelerator and storage attach (PCIe G4)
- New CAPI: Improved latency and bandwidth, open interface (BlueLink)

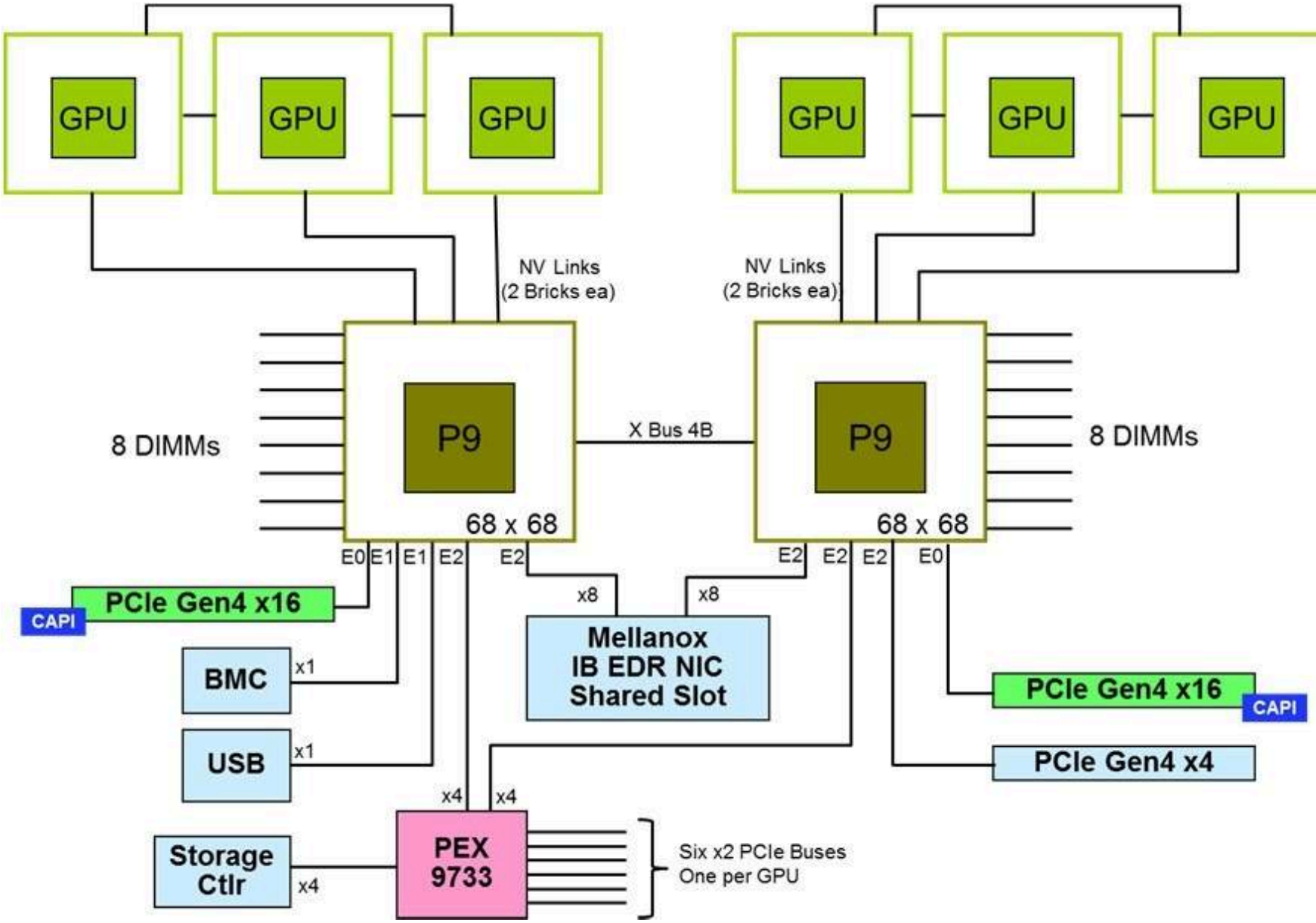
State of the Art I/O Subsystem

- PCIe Gen4 – 48 lanes

High Bandwidth Signaling Technology

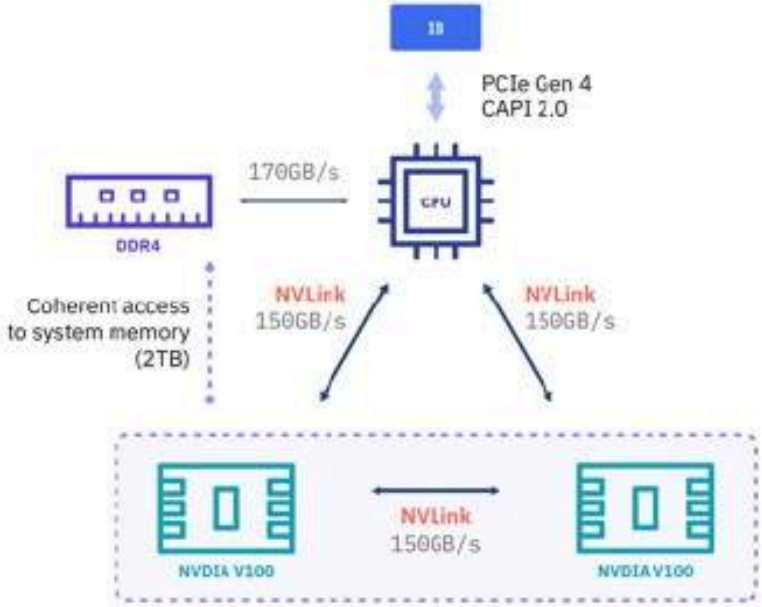
- 16 Gb/s interface
  - Local SMP
- 25 Gb/s IBM BlueLink interface
  - Accelerator, remote SMP

IBM Power9


GPU
GPU
GPU
GPU
GPU
GPU
NV Links
(2 Bricks ea)
NV Links
(2 Bricks ea)
8 DIMMs
P9
X Bus 4B
P9
8 DIMMs
68 x 68
68 x 68
E0
E1
E1
E2
E2
E2
E2
E2
E0
x8
x8
CAPI
PCIe Gen4 x16
BMC
x1
USB
x1
Mellanox
IB EDR NIC
Shared Slot
PCIe Gen4 x16
CAPI
PCIe Gen4 x4
x4
x4
Storage
Ctlr
x4
PEX
9733
Six x2 PCIe Buses
One per GPU

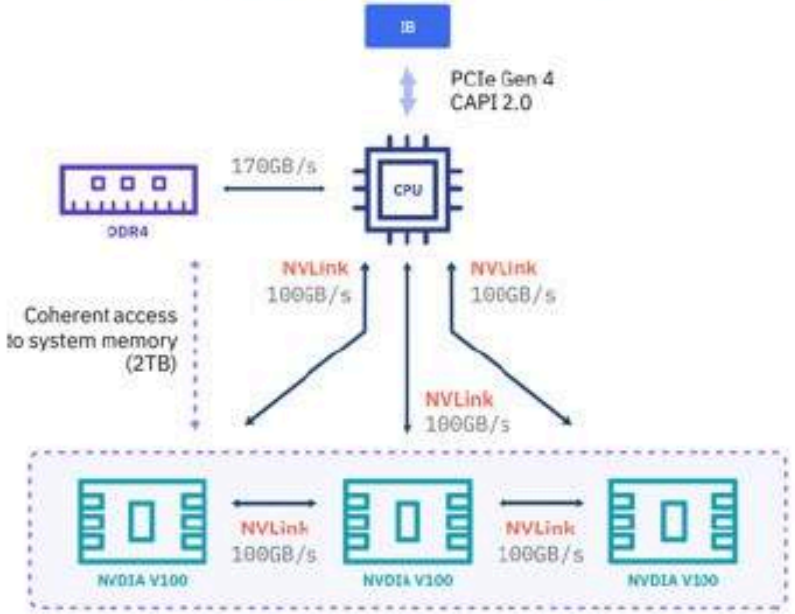
IBM Power9

4 GPUs - Air (4Q'17)/Water Cooled (2Q'18)



- Up to 4 GPUs, air/water cooled options
- 150GB/s of bandwidth from CPU-GPU

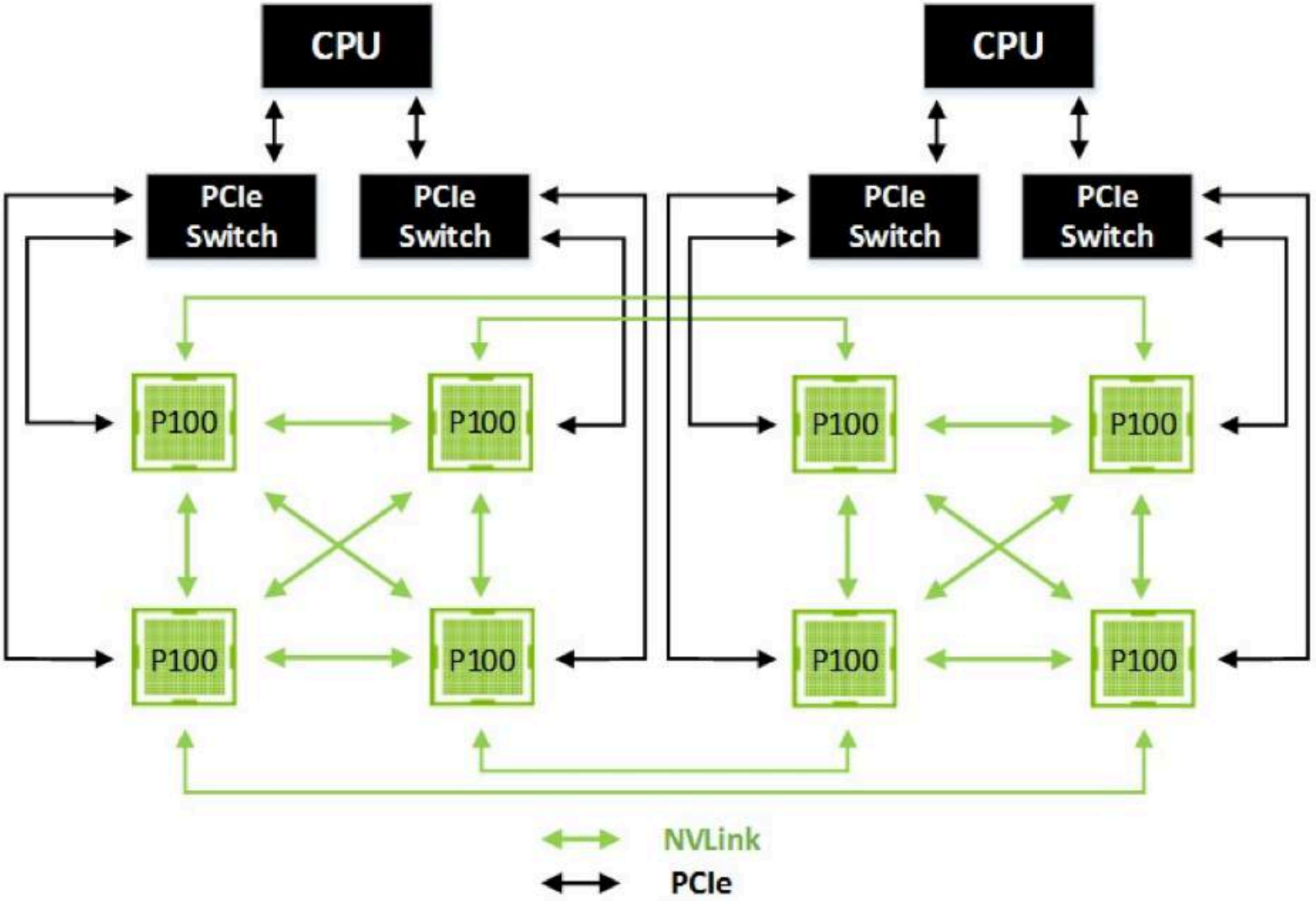
6 GPUs - Water Cooled (2Q'18)



- Up to 6 GPUs, water cooled only
- 100 GB/s of bandwidth from CPU-GPU

- Coherent access to system memory
- PCIe Gen 4 and CAPI 2.0 to InfiniBand
- Water cooled options available in 2Q'18

Nvidia Pascal / Volta GPUs
Nvlink & Unified Memory

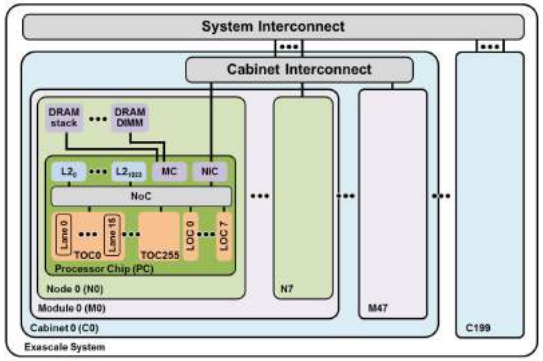


Что сможет дать 1 Эфлопс?

согласно оценкам 2014 года

TABLE V. PERFORMANCE, POWER, AND ENERGY EFFICIENCY OF A TARGET 7NM EXASCALE MACHINE.

APPS	PetaFlops	PetaOps	MWatts	GFlops/W	GOps/W
CNS	369.94	1065.98	16.71	22.13	63.78
CoMD	140.43	1158.51	15.12	9.29	76.61
LULESH	370.57	394.47	16.09	23.04	24.52
MiniFE	21.89	470.73	15.57	1.41	30.23
SNAP	22.02	251.66	16.58	1.33	15.18
XSBench	23.91	179.31	14.81	1.61	12.11
LINPACK	1019.22	1223.06	18.43	55.30	66.36



SC14: International Conference for High Performance Computing, Networking, Storage and Analysis

Scaling the Power Wall: A Path to Exascale

Oreste Villa, Daniel R. Johnson, Mike O'Connor, Evgeny Bolotin, David Nellans,
Justin Luitjens, Nikolai Sakharnykh, Peng Wang, Paulius Micikevicius, Anthony Scudiero,
Stephen W. Keckler and William J. Dally
Email: {ovilla, djohnson, moconnor, ebolotin, dnellans, jluitjens, nsakharnykh,
penwang, pauliusm, ascudiero, skeckler, bdally}@nvidia.com

Argonne Theta supercomputer

a 9.65 petaflops system based on the second-generation Intel Xeon Phi

Early Evaluation of the Cray XC40 Xeon Phi System 'Theta' at Argonne

Scott Parker, Vitali Morozov, Sudheer Chunduri, Kevin Harms, Chris Knight, and Kalyan Kumaran
Argonne National Laboratory, Argonne, IL, USA
{sparker, morozov, chunduri, harms, knightc, kumaran}@alcf.anl.gov

THETA – CRAY XC40 ARCHITECTURE

Nodes	3,624
Processor core	KNL (64-bit)
Speed	1100 - 1500 MHz
# of cores	64
# of HW threads	4
# of nodes/rack	192
Peak per node	2662 GFlops
L1 cache	32 KB D + 32 KB I
L2 cache (shared)	1 MB
High-bandwidth memory	16 GB
Main Memory	192 GB
NVRAM per node	128 GB SSD
Power efficiency	4688 MF/watt [7]
Interconnect	Cray Aries Dragonfly
Cooling	Liquid cooling

Argonne Theta supercomputer

Nekbone mini-app

(incompressible Navier-Stokes CFD solver based on spectral element method)

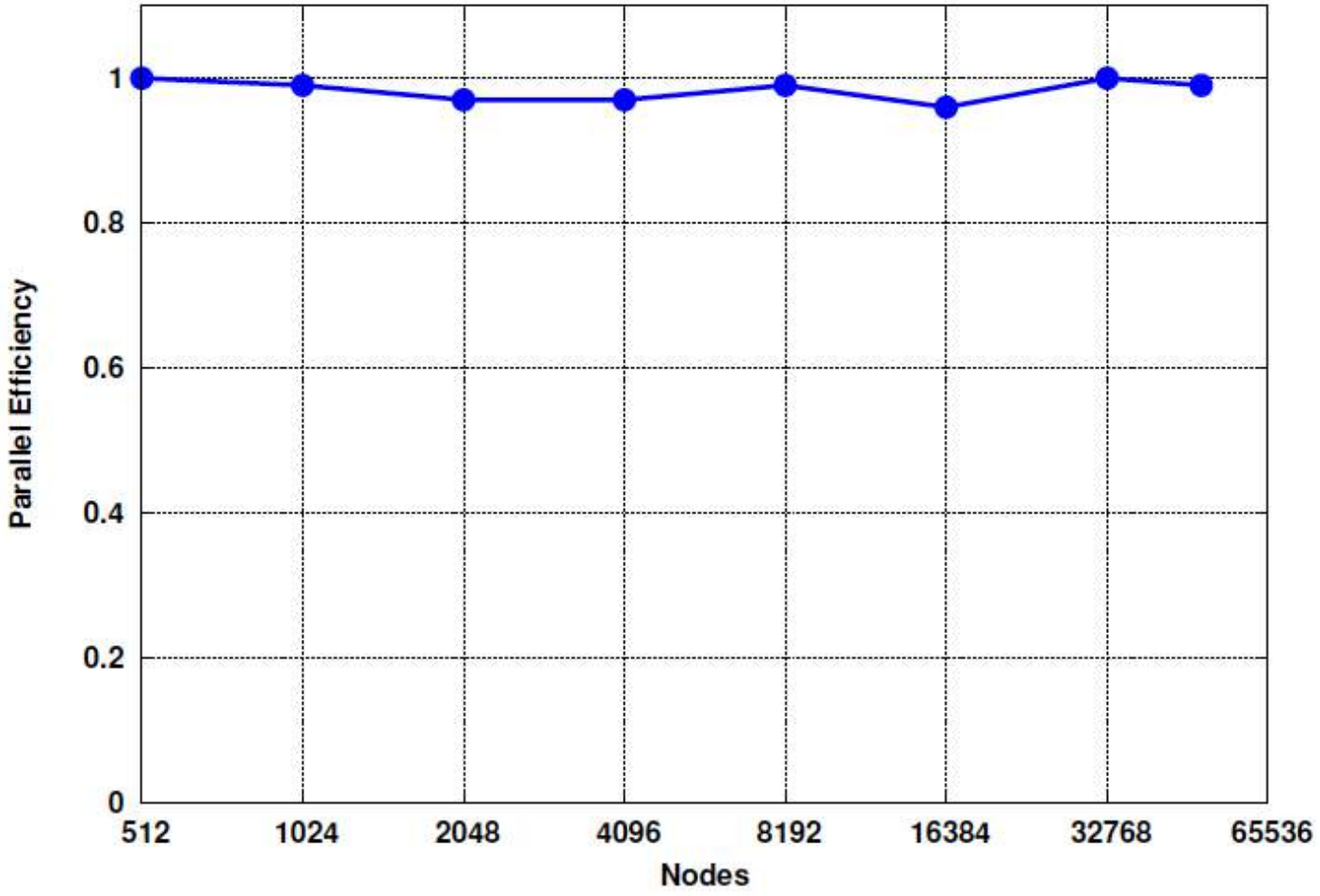


Fig. 15. Nekbone BG/Q weak scaling

Argonne Theta supercomputer

Nekbone mini-app

(incompressible Navier-Stokes CFD solver based on spectral element method)

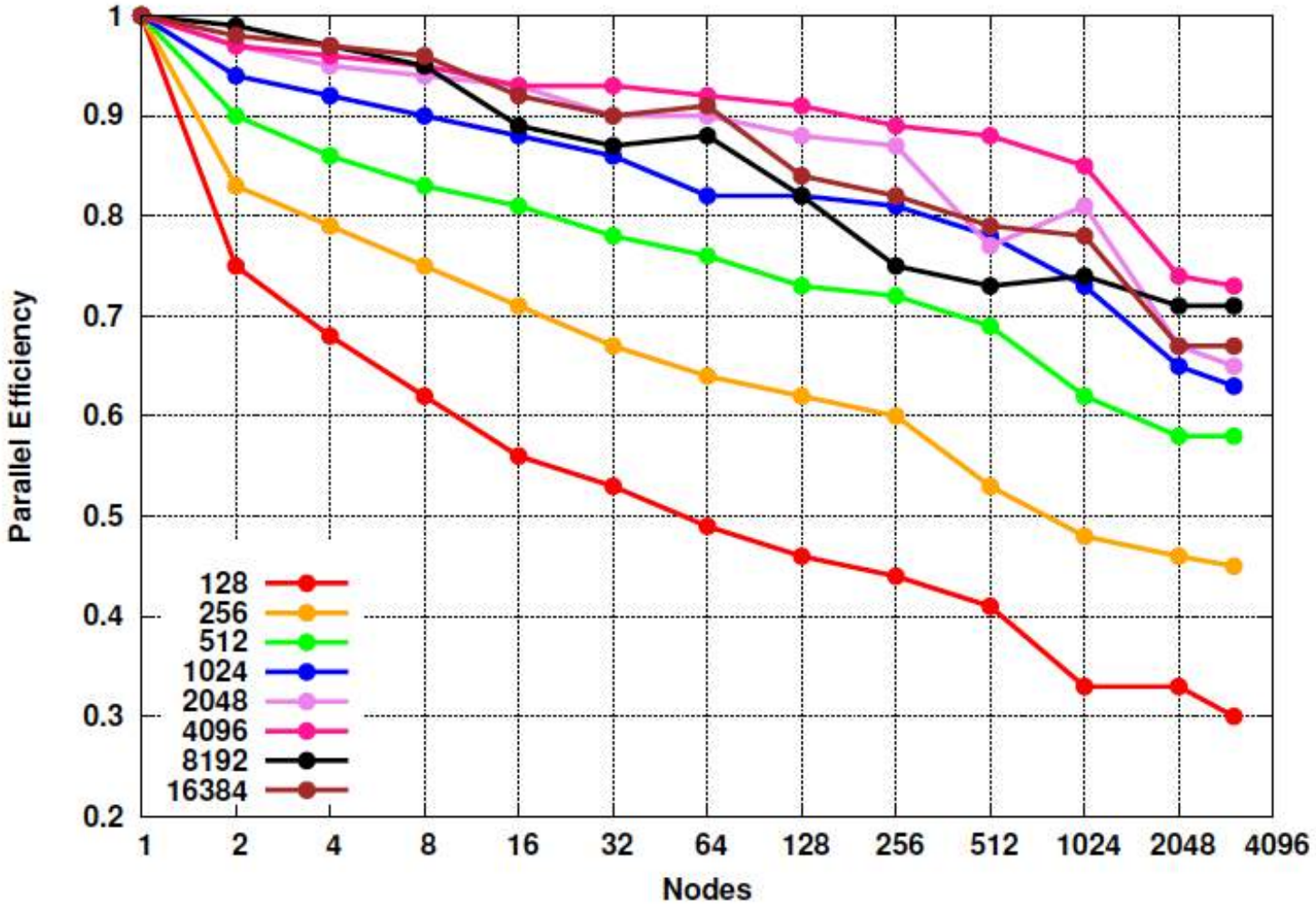


Fig. 14. Nekbone Theta weak scaling

Argonne Theta supercomputer

Nekbone mini-app

(incompressible Navier-Stokes CFD solver based on spectral element method)

Nekbone, itself is highly scalable, as shown in Figure 15, which shows near perfect weak scaling on the ALCF BG/Q system Mira up to 48k nodes when using 512 elements per node. Loss of parallel efficiency with weak scaling on Theta is attributable to either increased cost for point-to-point communication due to network contention or the increased cost of MPI_Allreduce operations as the rank count increases, since the workload per node remains otherwise same. Future work will examine the use of explicit rank mapping across the dragonfly network to optimize rank placement and minimize point-to-point communication contention.

TABLE III
MPI MESSAGE LATENCY IN US

Benchmark	Zero Bytes message	One Byte message
Ping Pong	3.07	3.22
Put	0.61	2.90
Get	0.61	4.70

Программа CORAL

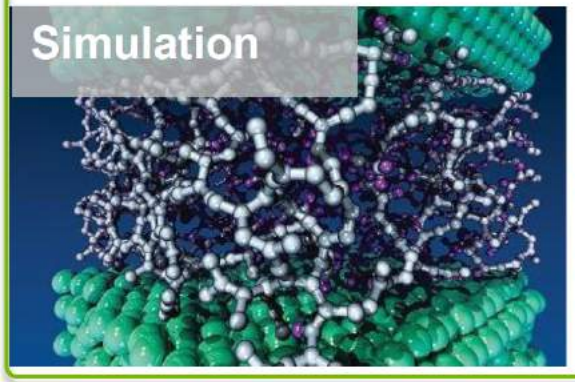
Почему отменили постройку суперкомпьютера Aurora?

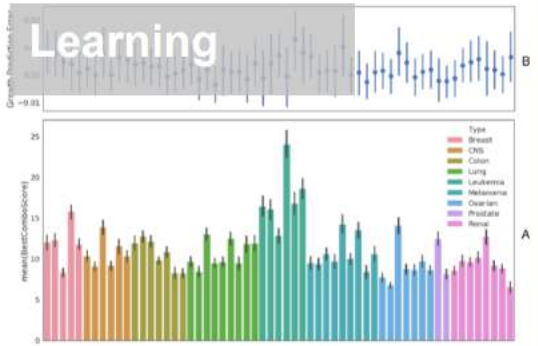
ALCF 2021 EXASCALE SUPERCOMPUTER – A21

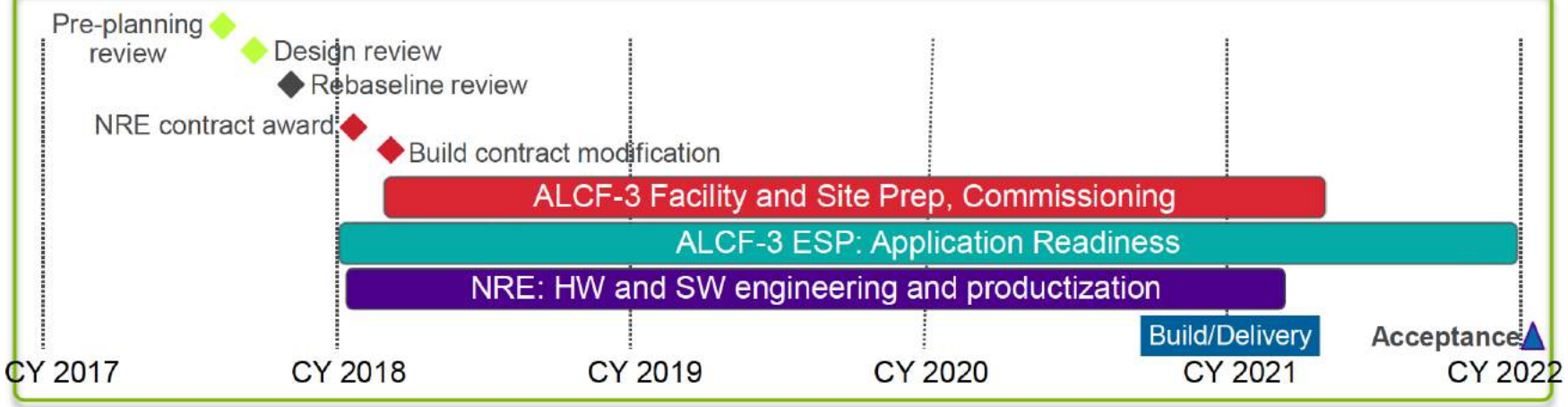
Intel/Cray Aurora supercomputer planned for 2018 shifted to 2021

Scaled up from **180 PF to over 1000 PF**

Support for three “pillars”

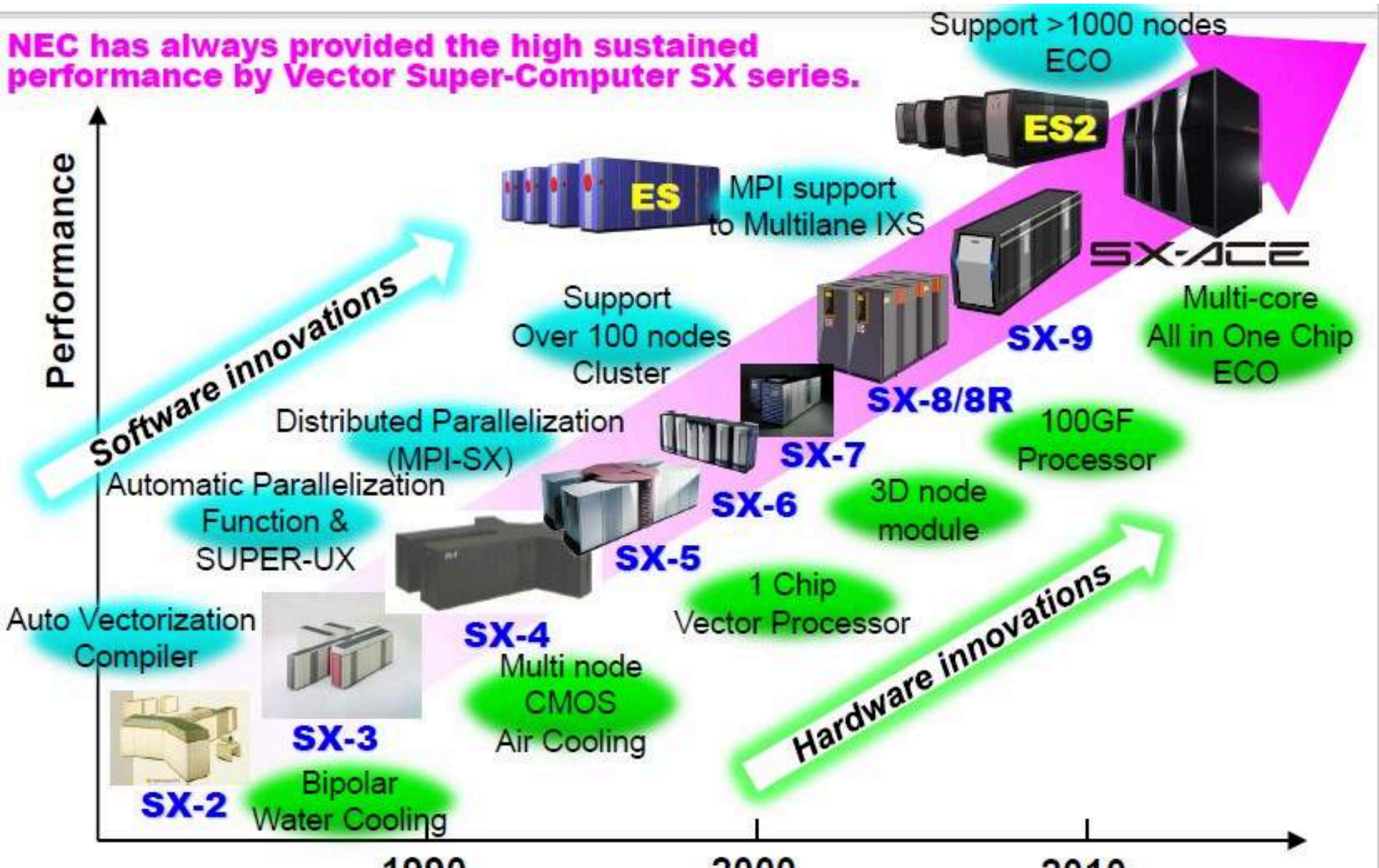






Японские процессоры

Nec Vector Processors: SX-ACE, SX-10, Aurora SX-10+…



Японские процессоры

Nec Vector Processors: SX-ACE, SX-10, Aurora SX-10+…

	SX-2	SX-3	SX-4	SX-5	SX-6	SX-7	SX-8	SX-9	SX-10	SX-10+
Launch Date	1983	1989	1994	1998	2001	2002	2004	2007	2014	2017
Chip Technology	Bipolar	Bipolar	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS
Process	-	-	350 nm	250 nm	150 nm	150 nm	90 nm	65 nm	28 nm	14 nm
Clock Speed	166 MHz	340 MHz	125 MHz	250 MHz	500 MHz	552 MHz	1.0 GHz	3.2 GHz	1.0 GHz	1.6 GHz
Core Count	1	1	1	1	1	1	1	1	4	8
DP Performance, Gigaflops	1.3	5.5	2.0	8.0	8.0	8.8	102.4	102.4	256.0	2,457.6
Memory Bandwidth, GB/sec	10.7	12.8	16.0	64.0	32.0	35.3	64.0	256.0	256.0	1,228.8

Японские процессоры

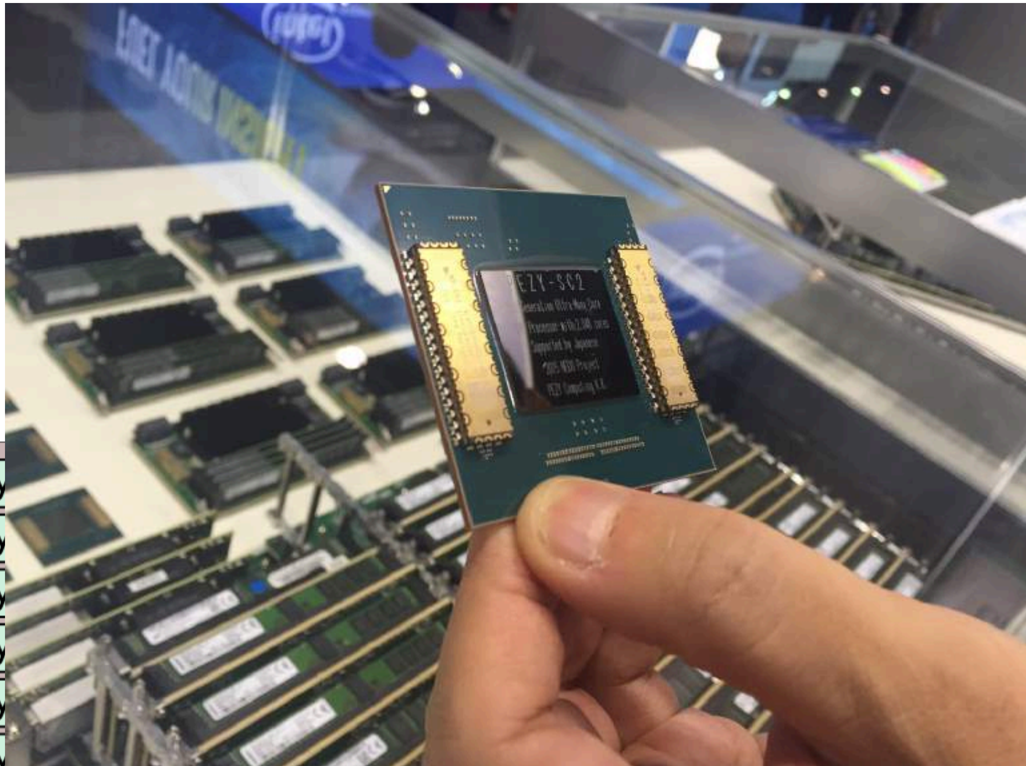
PEZY-SC2 - #1 в Green500

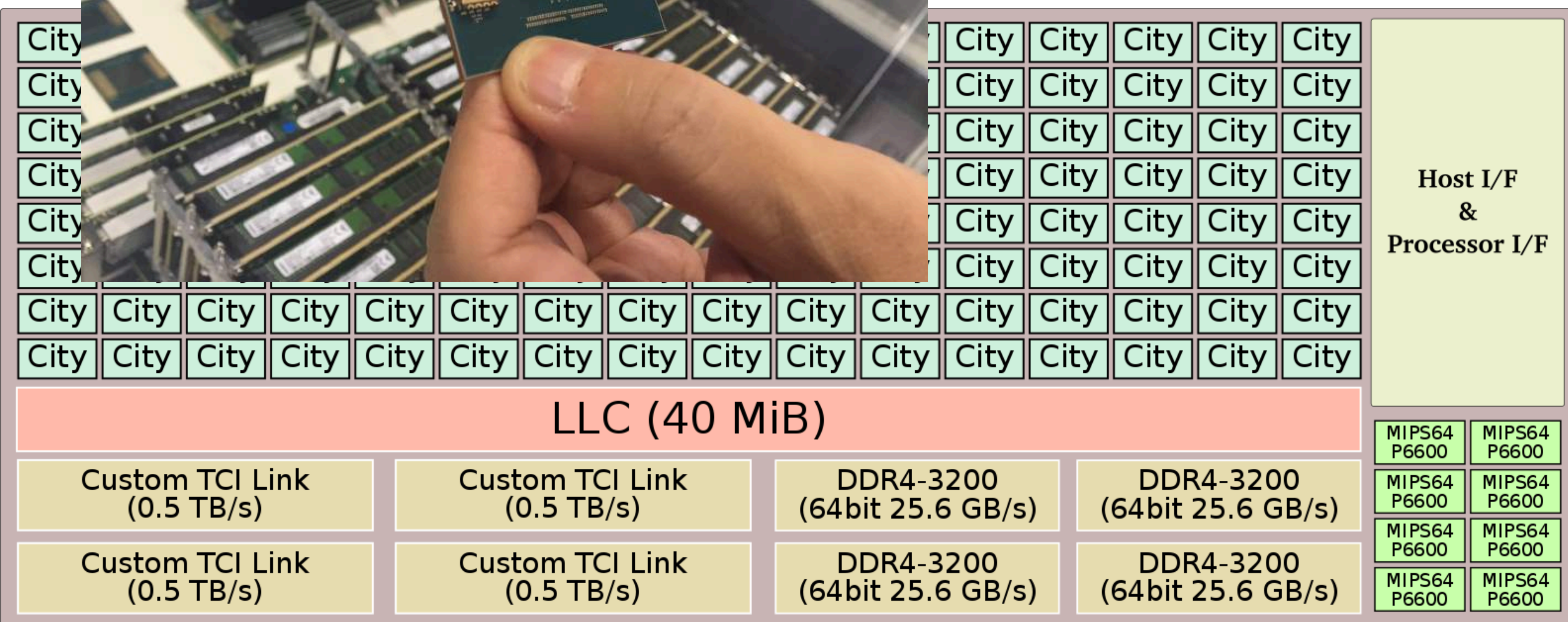
The ZettaScaler-2.2 at the Japan Agency for Marine-Earth Science and Technology (JAMSTEC)

#4 в списке Топ500 ноября 2017 года (Rmax=19 Пфлопс, Rpeak=28 ПФлопс) - 19,860,000 ядер!

Японские процессоры

PEZY-SC2 - #1 в Green500





Японские процессоры

PEZY-SC2 - #1 в Green500

PEZY-SCx Processor Roadmap

	PEZY-SC	PEZY-SC2	PEZY-SC3	PEZY-SC4
Process	28nm	16nm	7nm	5nm
Die Size	412mm2	620mm2	700mm2	740mm2
Number of Cores	1,024	2,048	8,192	16,384
Core Voltage	0.9V	0.8V	0.65V	0.55V
Core Clock	733MHz	1GHz	1.33GHz	1.6GHz
DRAM-IO	DDR4	DDR4	DDR4/5	DDR5
DDR Clock	2,133MHz	2,666MHz	3.6GHz	4GHz
Port	8	4	4	4
Wide-IO Clock	-	2GHz DDR	3 GHz DDR	3GHz DDR
Wide-IO Width		1,024bit	2,048bit	4,096bit
Wide-IO Ports		4	8	8
Memory Bandwidth	153.6GB/s	2.1TB/s	12.2TB/s	24.4TB/s
Peripheral IO	PCI3e Gen3	PCIe Gen4	Custom Optical	Custom Optical
Peripheral IO lane	24	32	128	512
Peripheral IO Bandwidth	32GB/s	64GB/s	256GB/s	1TB/s
DP Performance	1.5TFLOPS	4.1TFLOPS	21.8TFLOPS	52.5TFLOPS
SP Performance	3.0TFLOPS	8.2TFLOPS	43.6TFLOPS	105TFLOPS
HP Performance	-	16.4TFLOPS	87.2TFLOPS	210TFLOPS
Power Consumption	100W	200W	400W	640W
Power Efficiency	15GFLOPS/w	20.5GFLOPS/w	54.5GFLOPS/w	82.0GFLOPS/w
System Efficiency	6.7GFLOPS/w	15GFLOPS/w	40GFLOPS/w	60GFLOPS/w

U.S. DEPARTMENT OF ENERGY | Office of Science

Advanced Scientific Computing Research

Presented to the

Advanced Scientific Computing Advisory Committee

by

Barbara Helland
Associate Director

September 26, 2017

Office of Science FY 2018 President's Request

(Dollars in thousands)

	FY 2016 Enacted	FY 2016 Current w/SBIR-STTR[a]	FY 2017 Enacted	FY 2018 President's Request	FY 2018 Request vs. FY 2017 Enacted	
Science						
Advanced Scientific Computing Research	621,000	621,000	647,000	722,010	+75,010	+11.6%
Basic Energy Sciences	1,849,000	1,849,000	1,871,500	1,554,500	-317,000	-16.9%
Biological and Environmental Research	609,000	609,000	612,000	348,950	-263,050	-43.0%
Fusion Energy Sciences	438,000	438,000	380,000	309,940	-70,060	-18.4%
High Energy Physics	795,000	795,000	825,000	672,700	-152,300	-18.5%
Nuclear Physics	617,100	617,100	622,000	502,700	-119,300	-19.2%
Workforce Development for Teachers and Scientists	19,500	19,500	19,500	14,000	-5,500	-28.2%
Science Laboratories Infrastructure	113,600	113,600	130,000	76,200	-53,800	-41.4%
Safeguards and Security	103,000	103,000	103,000	103,000	...	...
Program Direction	185,000	185,000	182,000	168,516	-13,484	-7.4%
Subtotal, Science	5,350,200	5,350,200	5,392,000	4,472,516	-919,484	-17.1%
Rescission of Prior Year Balances	-3,200	-3,200	-239	...	+239	-100.0%
Total, Science Appropriation	**5,347,000**	**5,347,000**	**5,391,761**	**4,472,516**	**-919,245**	**-17.0%**

[a]The FY 2016 Enacted column printed in the FY 2018 Congressional Budget Justification (President's Request) includes SBIR/STTR funding in the program lines

and reflects programmatic updates through the end of the fiscal year.

[b]This column provides the Annualized CR amount (CR through April 28, 2017; P.L. 114-254). It is calculated by reducing the FY 2016 Enacted by 0.1901%

Суперкомпьютеры 1986-87 гг.

Thinking Machines CM-2:
16384 однобитовых процессора совместно с 512 арифметическими ускорителями Weitek

Meiko Computing Surface:
64 транспьютерных узлов с процессорами Intel i860

В. В. Стегайлов, Г. Э. Норман. Проблемы развития суперкомпьютерной отрасли в России: взгляд пользователя высокопроизводительных систем // Программные системы: теория и приложения: электрон. научн. журн. 2014. Т. 5, № 1(19), с. 111–152. URL: http://psta.psiras.ru/read/psta2014_1_111-152.pdf

… и параллельные алгоритмы для первопринципных расчетов

VOLUME 68, NUMBER 9 PHYSICAL REVIEW LETTERS 2 MARCH 1992

Ab Initio Theory of the Si(111)-(7×7) Surface Reconstruction: A Challenge for Massively Parallel Computation

Karl D. Brommer,[1] M. Needels,[2] B. E. Larson,[3] and J. D. Joannopoulos[1]
[1]*Department of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139*
[2]*AT&T Bell Laboratories, 600 Mountain Avenue, Murray Hill, New Jersey 07974*
[3]*Thinking Machines, Cambridge, Massachusetts 02139*
(Received 8 November 1991)

An *ab initio* investigation of the Si(111)-(7×7) surface reconstruction is undertaken using the state of the art in massively parallel computation. Calculations of the total energy of an ~700 effective-atom supercell are performed to determine (1) the fully relaxed atomic geometry, (2) the scanning tunneling microscope images as a function of bias voltage, and (3) the energy difference between the (7×7) and the (2×1) reconstructions. The (7×7) reconstruction is found to be energetically favorable to the (2×1) surface by 60 meV per (1×1) unit cell.

PACS numbers: 73.20.−r, 68.35.Bs, 68.35.Md

Thinking Machines CM-2

VOLUME 68, NUMBER 9 PHYSICAL REVIEW LETTERS 2 MARCH 1992

Ab Initio Total-Energy Calculations for Extremely Large Systems: Application to the Takayanagi Reconstruction of Si(111)

I. Štich, M. C. Payne, R. D. King-Smith, and J-S. Lin
Cavendish Laboratory (TCM), University of Cambridge, Madingley Road, Cambridge CB3 0HE, United Kingdom

L. J. Clarke
Edinburgh Parallel Computer Centre, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JZ, United Kingdom
(Received 8 November 1991)

We have implemented a set of total-energy pseudopotential codes on a parallel computer which allows calculations to be performed for systems containing many hundreds of atoms in the unit cell. Using these codes we have calculated the total energies and structures of the 3×3, 5×5, and 7×7 Takayanagi reconstructions of the (111) surface of silicon. We find that the 7×7 structure minimizes the surface energy and observe structural trends across the series which can be correlated with the degree of charge transfer between the dangling bonds on the adatoms and rest atoms.

PACS numbers: 68.35.−p, 31.20.−d, 71.45.Nt

Meiko Computing Surface

Выводы

- Переход в пост-Муровскую эру ведет к необходимости повышения эффективности существующих технологий и ко-дизайну.
- Новые технологии требую развития экосистемы.
- Процессоры Intel Xeon Skylake сместили баланс в область более высоких значений (для топового процессора B ~ 20) и требуют адаптации кодов под AVX-512.
  Для топовых AMD Epyc B ~ 3.
- К настоящему времени сформировалась экосистема ARM, выпуск двух типов серверных процессоров показывают перспективность этой архитектуры, как дающей большую степень вариабельности, чем x86_64.
- Технологии (экосистеме) Nvidia CUDA исполнилось 10 лет, и ее развитие в направлении Unified Memory обещает удобный инструмент достижения высокой производительности на большом числе приложений.
- Отмена третьего поколения Xeon Phi и суперкомпьютера Aurora – пример сложности развития в направлении экзаскейла.
- **Ключ к успеху в развитии экзаскейла – развитие методов параллельного программирования!**