Диверсификация и импортозамещение при разработке российских суперкомпьютеров

экза- и зетта– уровня

Л.К.Эйсымонт

(журнал “Открытые системы”,
26 апреля 2017)

Цели - прогнозы

Вариант А (общепризнанный)

10^{18} flops , 10^{18} byte {NUMA, PGAS}, 10-7 нм,
20 MW (или 50 Gflops/W), ~ 100000 mP,
mP - 10-20 Tflops , 1 Tbyte , 200W

Вариант В (пример частной постановки)

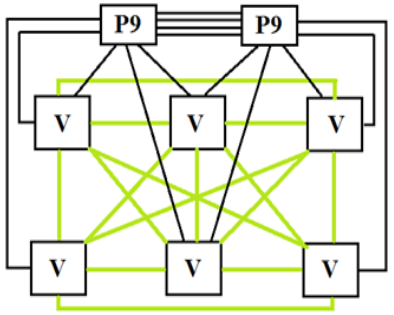
10^{18} ops , 10^{14} byte ,
1 MW (или 13000 Gops/W), ~ 30000 mP,
mP - 150 Tops, 4GB, 20W, 22-14 нм, <100 мм^2

Year	Feature size	Scalar parallelizm	Stream parallelizm	PIM parallelizm	Clock rate GHz	FMAs	GFLOPS (Scalar)	GFLOPS (Stream)	GFLOPS (PIM)	mNode (TFLOPS)	mNodes per Node	Node (TFLOPS)	Nodes per system	Total (PFLOPS)
2012	22	16	512	0	2	2	128	1,024	0	1	2	2	10,000	23
2020	12	54	1,721	0	2.8	4	1,210	4,819	0	6	2	12	20,000	241
2023	8	122	3,873	512	3.1	4	3,026	12,006	1,587	17	4	68	20,000	1,330
2030	4	486	15,489	1,024	4	8	31,104	61,966	8,192	101	16	1,616	20,000	32,320

ORNL
Проект Summit. 2018
6 PB, 23 MW

GFLOPS (Scalar)	GFLOPS (Stream)	GFLOPS (PIM)	mNode (TFLOPS)	mNodes per Node	Node (TFLOPS)	Nodes per system	Total (PFLOPS)
1,600	38,400	0	~ 40	1	~ 40	4600	184

Узел (Node) проекта Summit



P9 - Power 9, 24 ядра SMT4, 96 тредов, 14 нм
V - GPU Volta, 12 нм

Главные тезисы доклада

Прогноз развития в мире - достижение к 2020+ году предела развития кремниевых технологий после освоения технологий 7 нм. Далее некоторое время (до 2030 + года ?) неизбежна интенсификация работ по разного рода проблемно-ориентированным и специализированным СБИС для суперкомпьютеров высшего диапазона производительности (СКСН) . Представляет интерес также освоение нового класса приборов - мемристоров. Такой подход будет постепенно внедрен для вычислительных устройств других уровней производительности, вплоть до мобильных устройств. Так будет продолжаться до освоения пост-Муровских технологий, скорее всего, сверхпроводниковых (в том числе с учетом возможностей QCA-схемотехники) и нанотрубок.

Стратегия - через диверсификацию процессорных СБИС и других компонентов, учитывая требуемую проблемную ориентацию выйти на настоящее импортозамещение в области суперкомпьютеров класса СКСН.

Тактика - проводить значительно больше исследований по архитектуре, затруднения в массовой организации таких работ компенсировать активизацией информационно-аналитической работы с активным привлечением молодежи

Диверсификация архитектур в области суперкомпьютеров утверждена законом США от 2004 года ".. о возрождении HEC в DOE…".

Public Law 108–423
108th Congress

An Act

Nov. 30, 2004
[H.R. 4516]

To require the Secretary of Energy to carry out a program of research and development to advance high-end computing.

Department of Energy High-End Computing Revitalization Act of 2004.
15 USC 5501 note.

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

SECTION 1. SHORT TITLE.

This Act may be cited as the "Department of Energy High-End Computing Revitalization Act of 2004".

15 USC 5541.

SEC. 2. DEFINITIONS.

In this Act:

(1) CENTER.—The term "Center" means a High-End Software Development Center established under section 3(d).

(2) HIGH-END COMPUTING SYSTEM.—The term "high-end computing system" means a computing system with performance that substantially exceeds that of systems that are commonly available for advanced scientific and engineering applications.

(3) LEADERSHIP SYSTEM.—The term "Leadership System" means a high-end computing system that is among the most advanced in the world in terms of performance in solving scientific and engineering problems.

(4) INSTITUTION OF HIGHER EDUCATION.—The term "institution of higher education" has the meaning given the term in section 101(a) of the Higher Education Act of 1965 (20 U.S.C. 1001(a)).

(5) SECRETARY.—The term "Secretary" means the Secretary of Energy, acting through the Director of the Office of Science of the Department of Energy.

15 USC 5542.

SEC. 3. DEPARTMENT OF ENERGY HIGH-END COMPUTING RESEARCH AND DEVELOPMENT PROGRAM.

(a) IN GENERAL.—The Secretary shall—

(1) carry out a program of research and development (including development of software and hardware) to advance high-end computing systems; and

(2) develop and deploy high-end computing systems for advanced scientific and engineering applications.

(b) PROGRAM.—The program shall—

(1) support both individual investigators and multidisciplinary teams of investigators;

!!!

(2) conduct research in multiple architectures, which may include vector, reconfigurable logic, streaming, processor-in-memory, and multithreading architectures;

.....

(c) LEADERSHIP SYSTEMS FACILITIES.—

(1) IN GENERAL.—As part of the program carried out under this Act, the Secretary shall establish and operate 1 or more Leadership Systems facilities to—

(A) conduct advanced scientific and engineering research and development using Leadership Systems; and

(B) develop potential advancements in high-end computing system hardware and software.

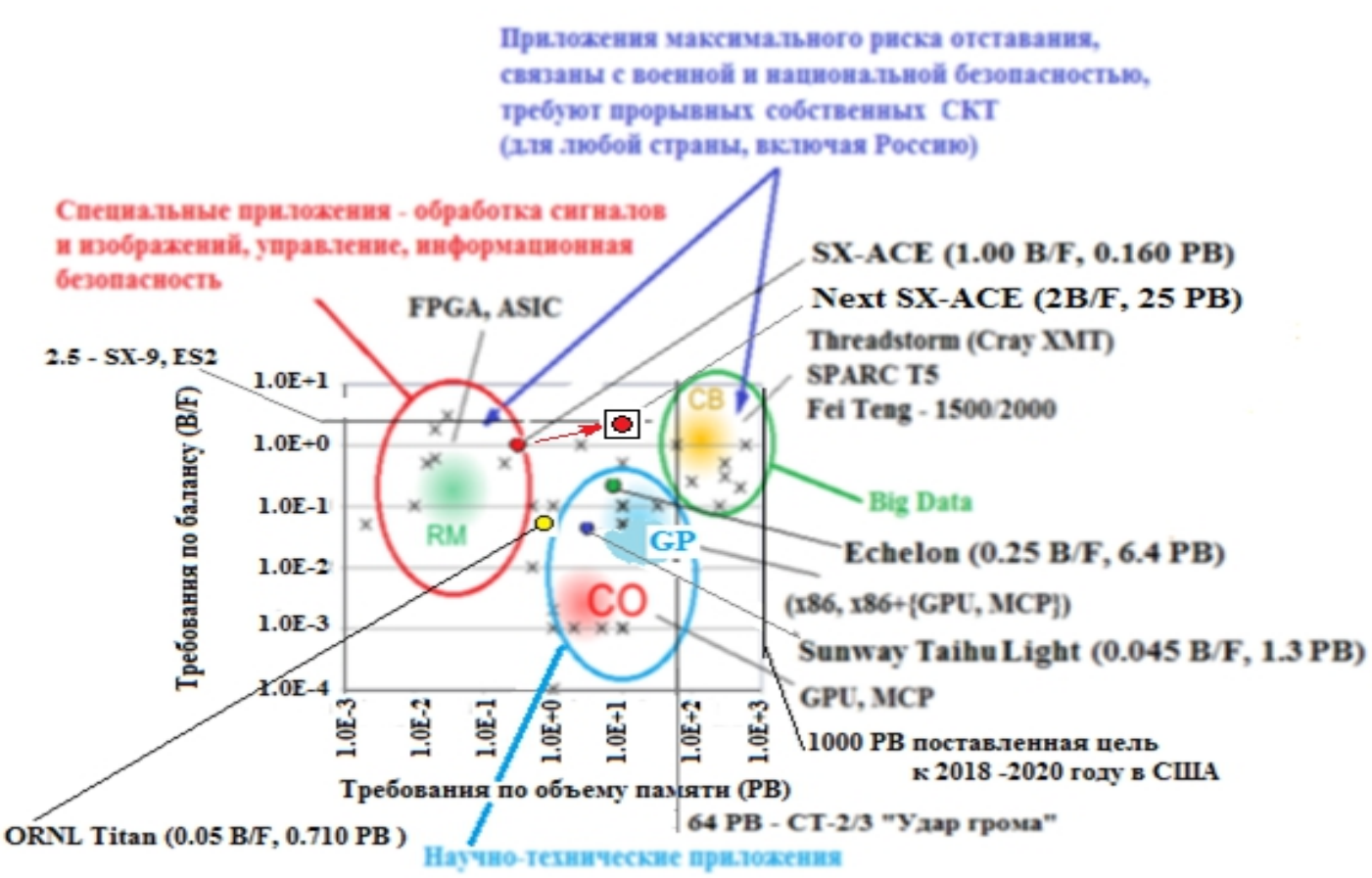
.....

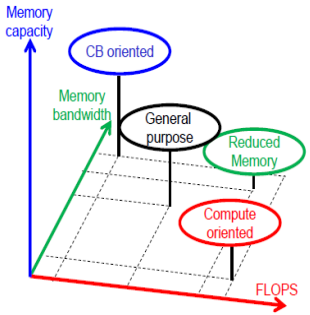
(d) HIGH-END SOFTWARE DEVELOPMENT CENTER.—

(1) IN GENERAL.—As part of the program carried out under this Act, the Secretary shall establish at least 1 High-End Software Development Center.

.....

Типы суперкомпьютеров по балансу пропускной способности памяти и производительности (B/F), объему памяти (PB)





ORNL Titan - 27 PF, peak
200 стоек
9 MW
NUDT Tianhe-2 - 54.9 PF, peak
162 стойки
17.8 MW
Sunway TianhuLight - 125.4 PF, peak
40 стоек
15.3 MW

2018+

Echelon - 1258 PF, peak
200 стоек
23 MW
Next SX-ACE - 100 PF
(~ 1000 PF)
400 стоек
20-30 MW

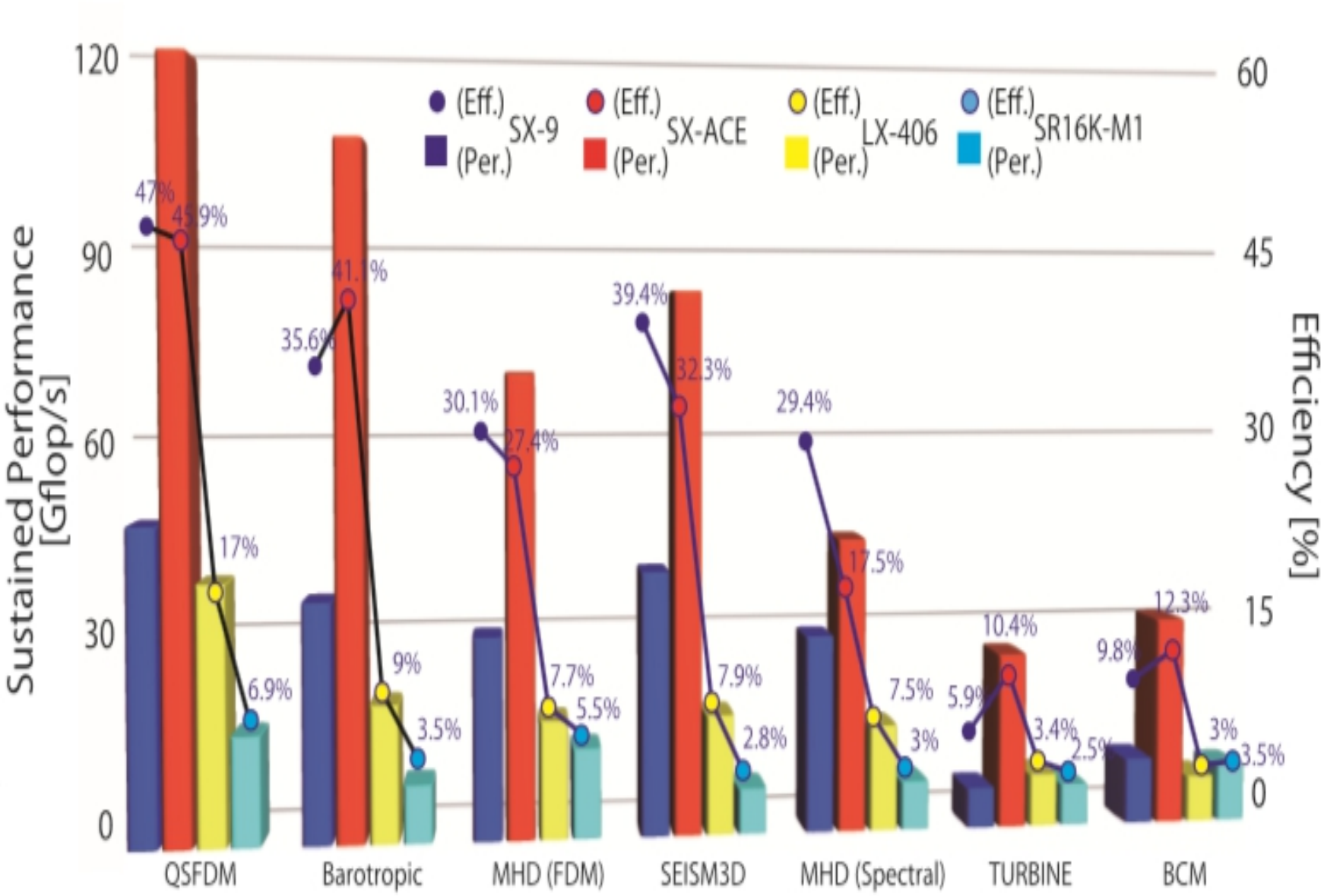
RM - reduced memory
CO - compute oriented
CB - communication bandwidth
GP - general purpose

Сравнительное оценочное тестирование процессоров SX-ACE, SX-9, Ivy Bridge, Power 7

Processor	Gflop/s/CPUs (Gflop/s/core, #. Cores)	Mem. GB/s	On-Chip Mem.	System B/F
SX-ACE	256 (64, 4 cores)	256	1MB ADB/core	1.0
SX-9	102.4 (102.4, 1 core)	256	256KB ADB/core	2.5
LX 406 (Ivy Bridge)	230.4 (19.2, 12 cores)	59.7	256KB L2/core 30MB shared L3	0.26
SR16000M1 (Power7)	245.1 (30.6, 8 cores)	128	256KB L2/core 32MB shared L3	0.52

Applications	Method	Memory access	Mesh Size	Code B/F	Actual B/F
QSFDM GLOBE	Spherical 2.5D FDM	Sequential	4.3×10^7	2.16	0.78
Barotropic	Shallow water model	Sequential	4322×2160	1.97	1.11
MHD (FDM)	FDM	Sequential	$2000 \times 1920 \times 32$	3.04	1.41
Seism3D	FDM	Sequential	$1024 \times 512 \times 512$	2.15	1.68
MHD (Spectral)	Pseudospectral Method	Stride	$900 \times 768 \times 96$	2.21	2.18
TURBINE	DNS	Indirect	$91 \times 91 \times 91 \times 13$	1.78	5.47
BCM	Navier Stokes Equation	Indirect	$128 \times 128 \times 128 \times 64$	7.01	5.86

Результаты сравнительное оценочного тестирования



Проблема N1:
“стена памяти ” (memory wall)

.... в сравнении с операциями процессора,
операции с памятью выполняются слишком долго ...

Базовые решения

- векторные и/или мультитредовые архитектуры
- архитектура с разделением обработки и доступа к данным (DAE-архитектура)
- 3D модули памяти с высокой пропускной способностью (HBM - память)
- встроенные в модули памяти процессоры (PIM - процессоры)

Проблема N2: “стена энергопотребления” (power wall)

.... много лишних энергетических затрат на выполнение операции из команды в одном ядре, большие потери на доставку операций и данных в функциональные устройства ...

... энергетические затраты обращений к памяти слишком велики ....

.... после перехода к технологиям 28 нм не все процессорные ядра можно одновременно включить - сгорит СБИС ....

Базовые решения

- векторные и/или мультитредовые архитектуры
- новые микроархитектурные решения по оптимизации доставки команд и данных к функциональным устройствам
- гибридные архитектуры
- HDM - память
- энергоэффективные реконфигурируемые коммутаторы
- динамическое управление энергопотреблением
- внутри- и внекристальные соединения (нанофотоника, оптика)
- специализация

Проблема N3:
“стена низкого параллелизма выполнения команд из одного потока” (ILP- wall)

.... из одного потока команд трудно автоматически выбрать и выполнить параллельно много команд, такие алгоритмы автоматического распараллеливания вдобавок требуют больших энергетических затрат ....

Базовые решения

- векторные и/или мультитредовые архитектуры
- SIMD - ускорители (128-, 256-, 512-, 1024-)
- TTA - архитектура процессора
- спецускорители

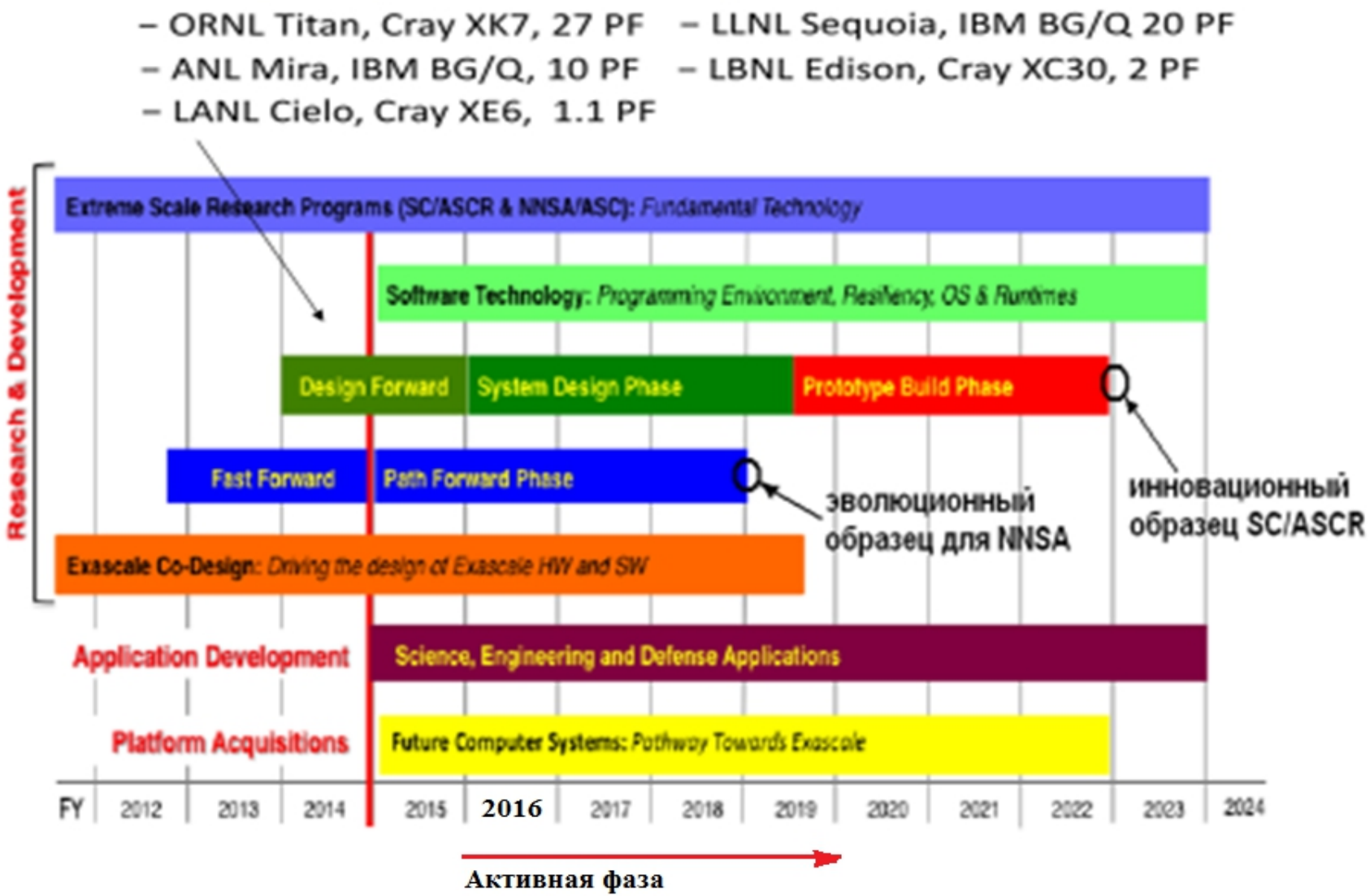
Проблема N4:
“проблема межплатных и межузловых коммуникаций” (communication wall)

... передача данных между платами и особенно вычислительными узлами требует фантастических затрат по времени и энергетике...

Базовые решения

- high-radix коммуникационные сети
- использование оптических соединений
- PGAS - память

Дорожная карта создания экзафлопсных суперкомпьютеров DoE США (2014 г)



Проекты США

ORNL - Summit, IBM&NVIDIA&Mellanox, 150-300 Пфлопс, 2017-2018,
4600 узлов, память - >6 PB + HBM,
файловая система - 250 PB (2.5 TB/s), 23 MW

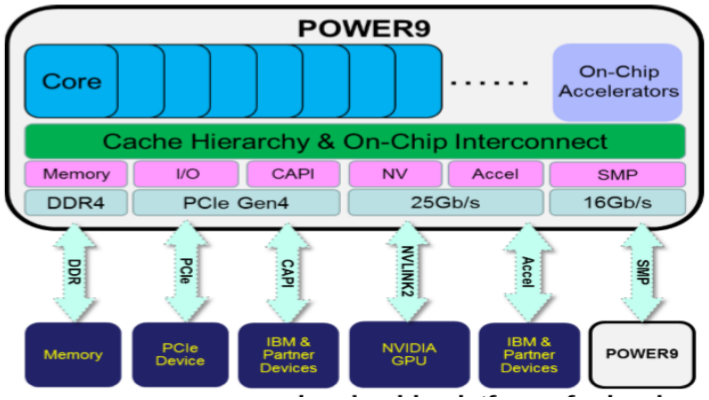
сеть - dial-rail Mellanox EDR или HDR Infiniband, Fat Tree,
узел - 2 IBM Power9 и 6 GPU NVIDIA Volta
NVLink, DDR4 - 0.5 ТБ, +HBM, NVRAM - 0.8 ТБ,
40TF, сетевой интерфейс EDR - 25 GB/s,
HDR - 48 GB/s

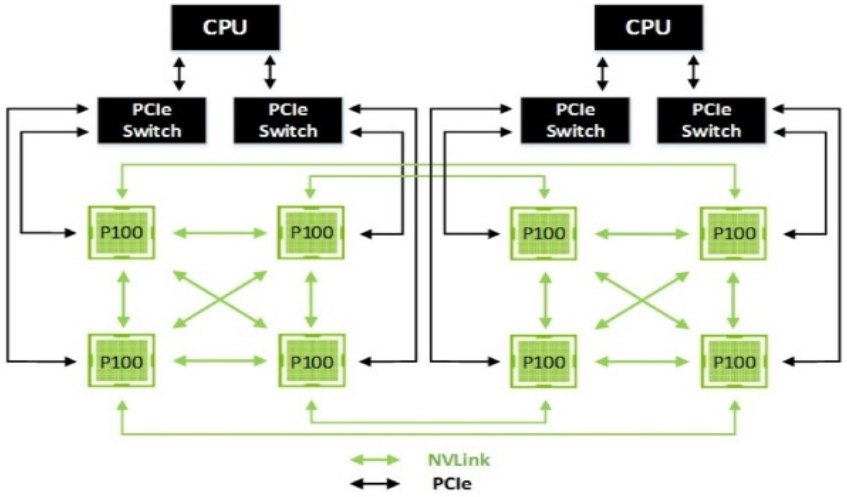
Power 9 - 14 нм (возможно 10 нм), 12 ядер SMT8 или
24 ядра SMT4, 96 тредов, 0.85 - 1 TF

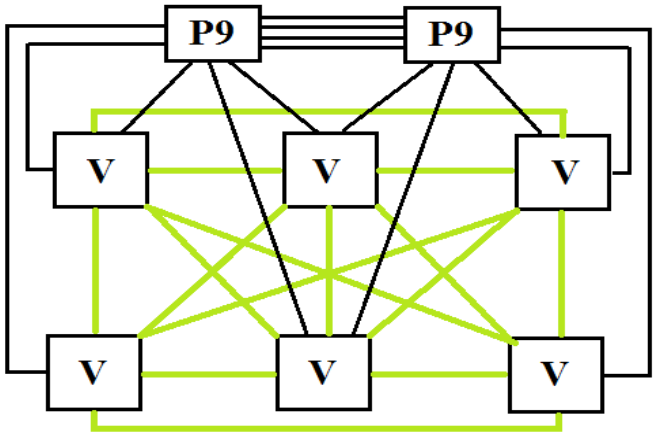
Было - Power 8 - 22 нм, 12 ядер, 96 тредов, 0.5-0.7 TF
Power 7 - 45 нм, 8 ядер, 32 треда, 0.25 TF

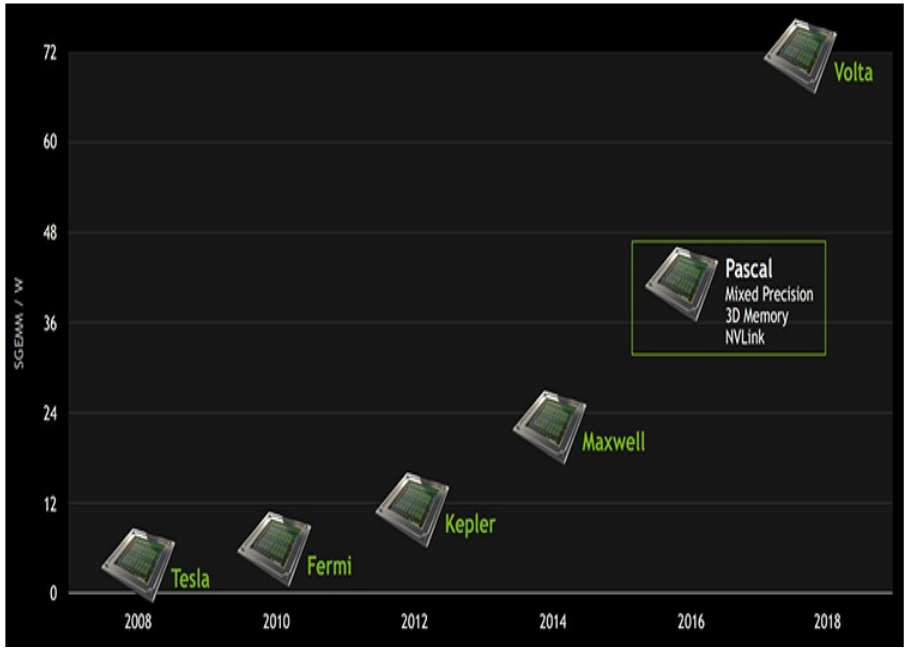
LLNL - Sierra, 100+ Пфлопс, остальное типа Summit
ANL - Aurora, Cray&Intel, 180 Пфлопс,
Xeon Knight Landing, сеть Intel OmniPath

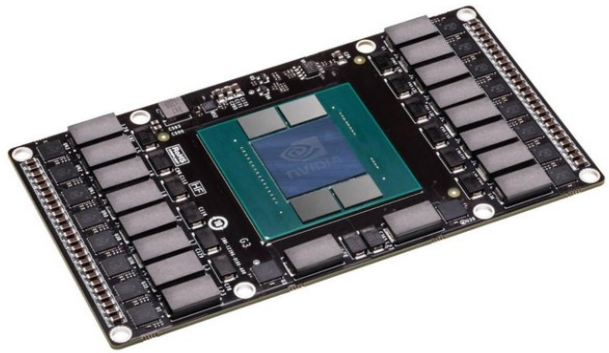
IBM Power 9 – подключение ускорителей





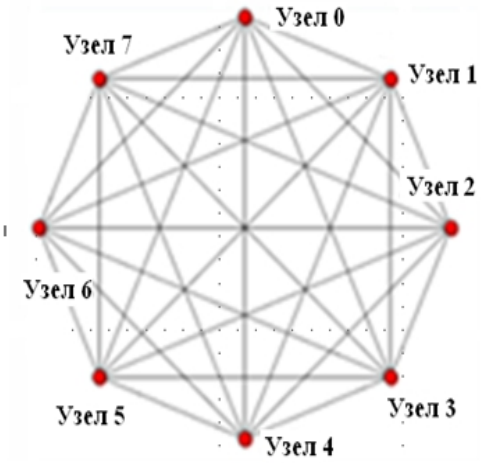




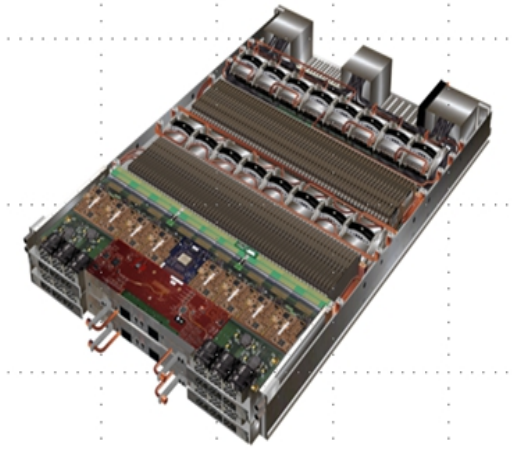


История – тесты одного узла (QCM) суперкомпьютера Power 775 с четырьмя микропроцессорами Power 7, использовался только режим SMT2

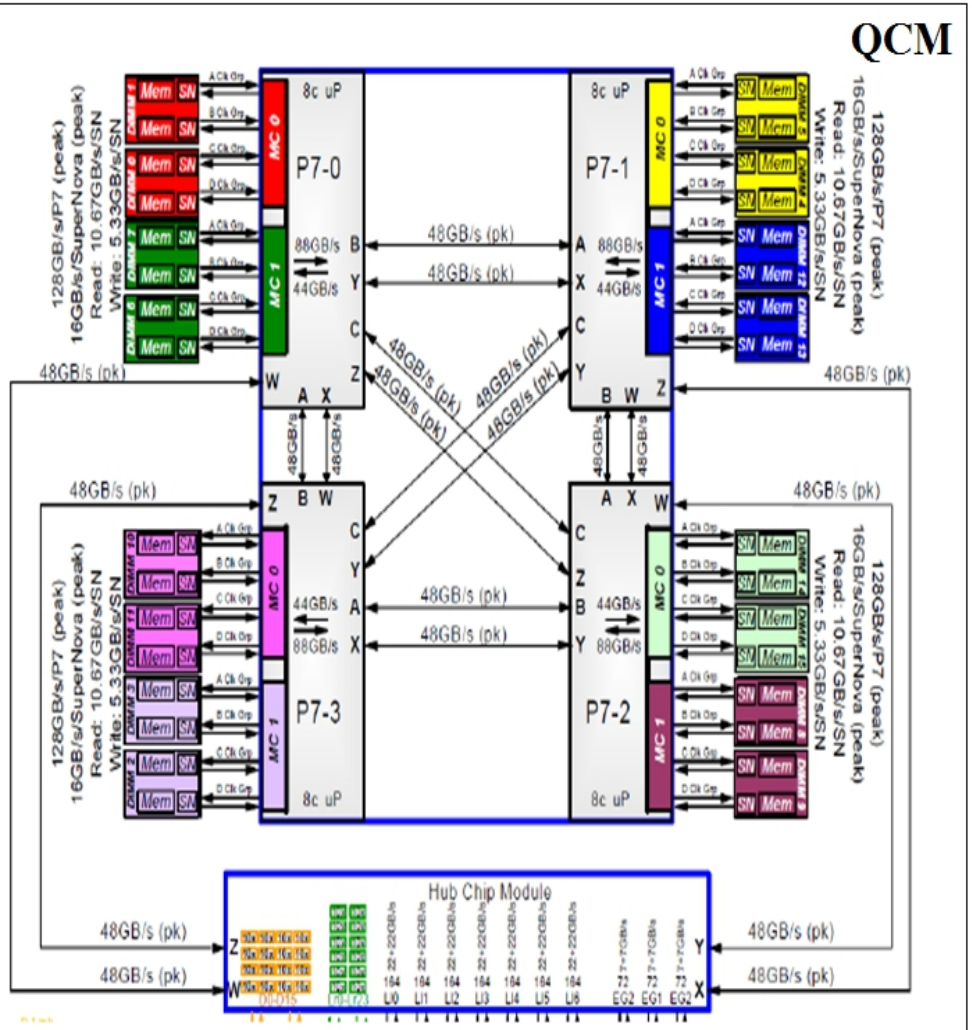
Узел и макроузел Power 775



Внутренняя сеть макроузла

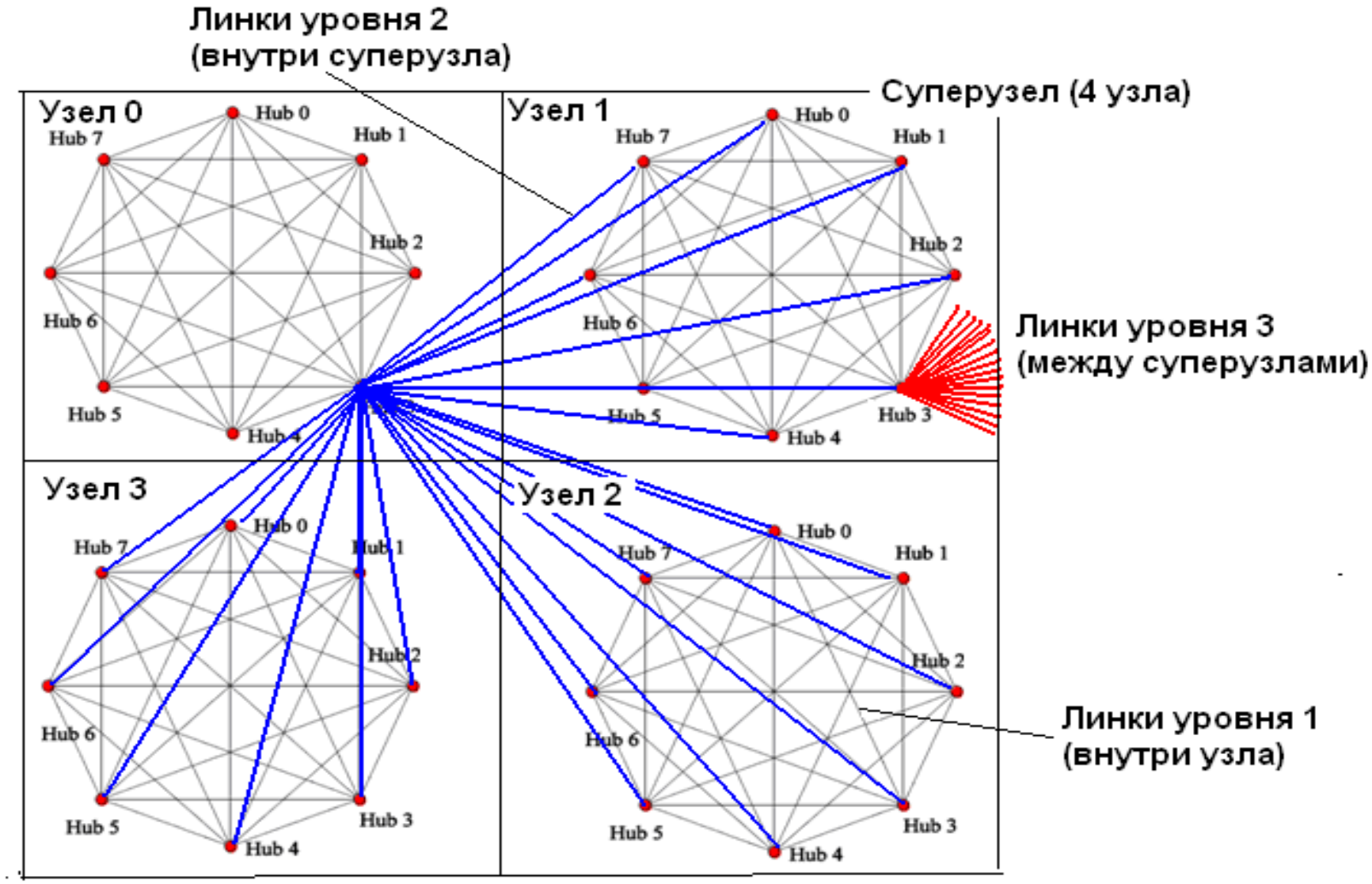


Внешний вид макроузла

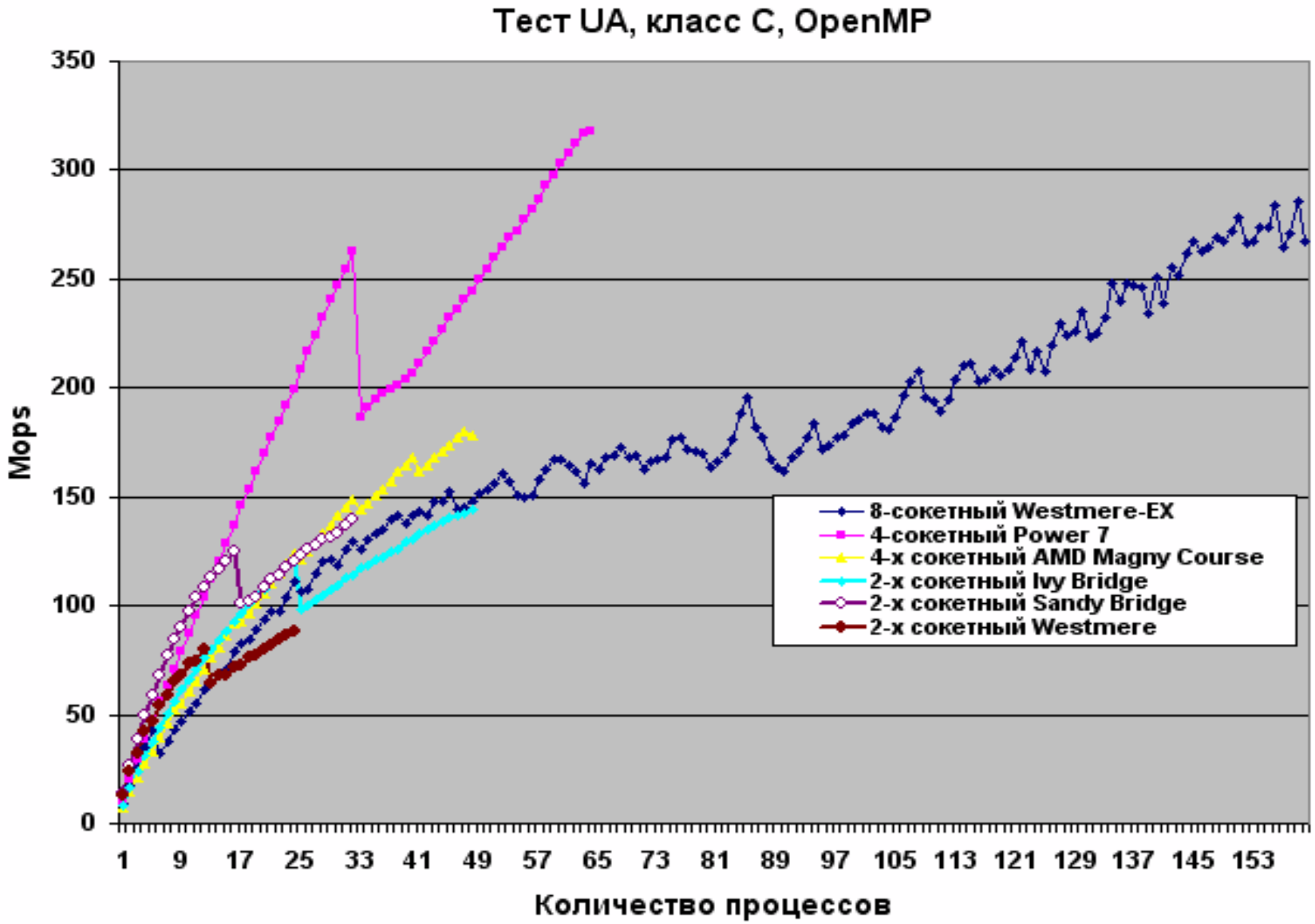


4-х процессорный узел (4 Power7 + HUB)

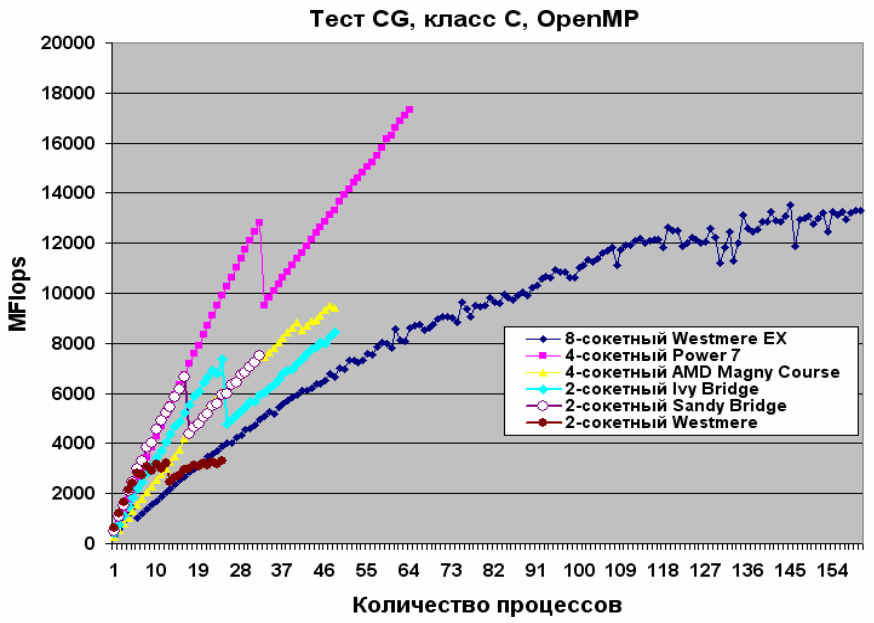
Многоуровневая сеть PERCS суперкомпьютера Power 775

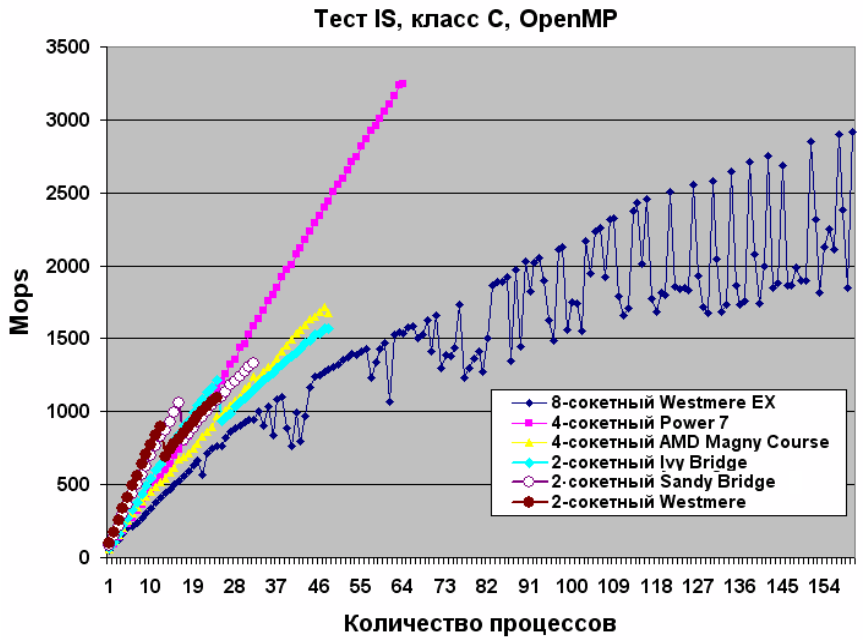


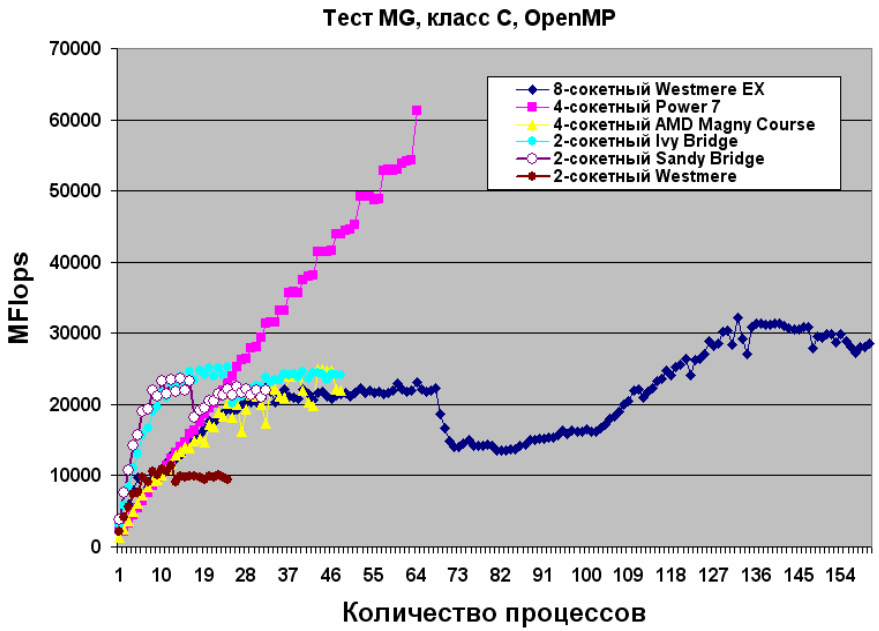
IBM Power 7 - тест UA (класс C)

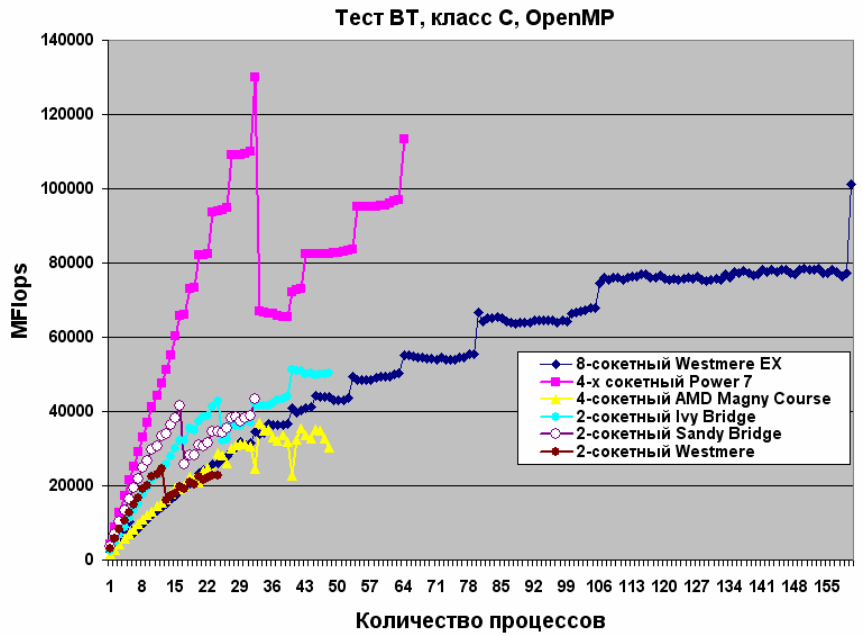


IBM Power 7 - тесты CG, IS, MG, BT



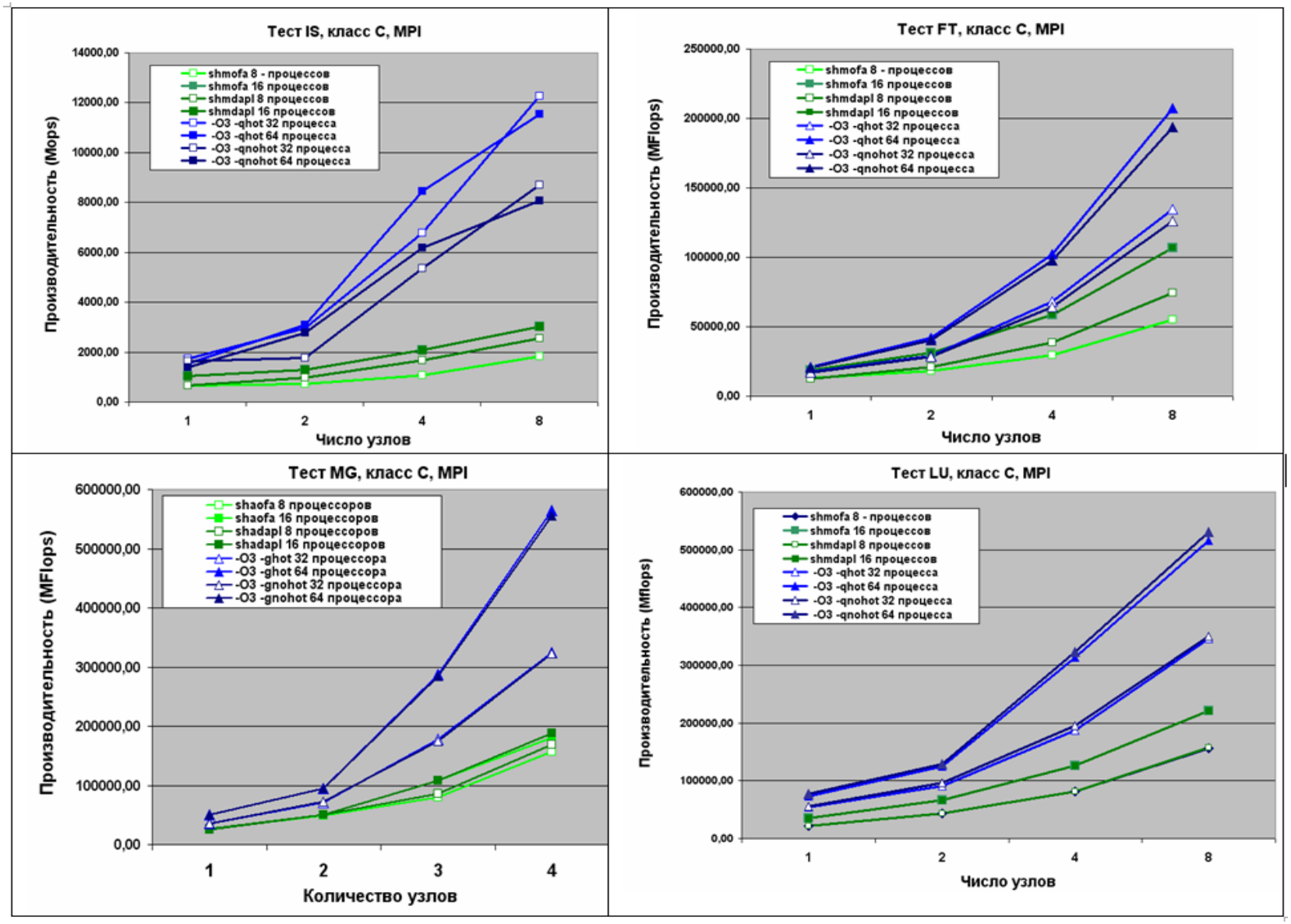




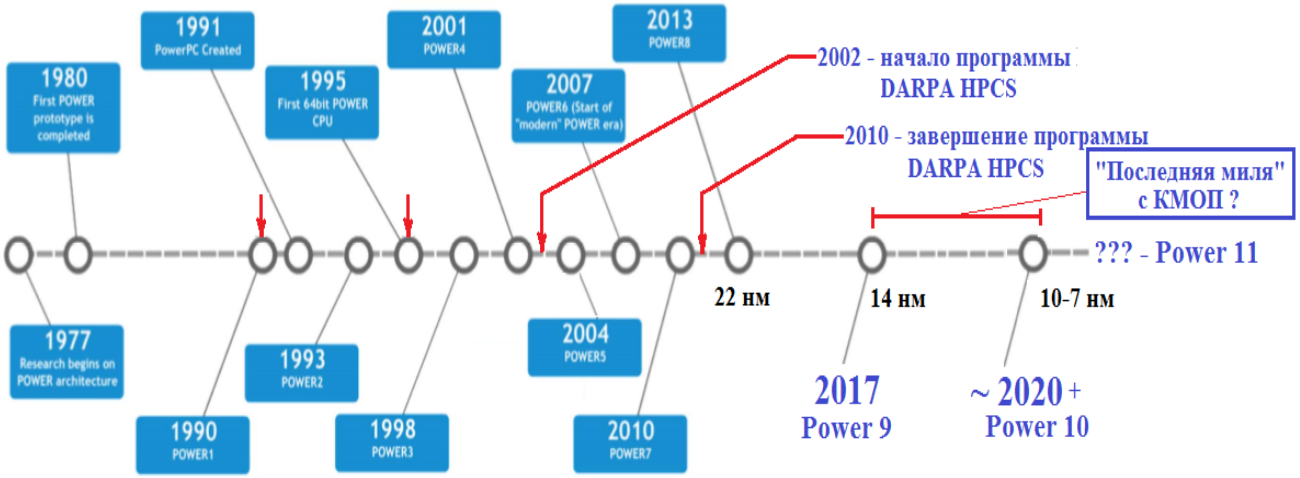


История – тесты одного макроузла суперкомпьютера Power 775 с восемью QCM (32 микропроцессора Power 7), использовался только режим SMT2

Масштабирование производительности при увеличении процессов и узлов

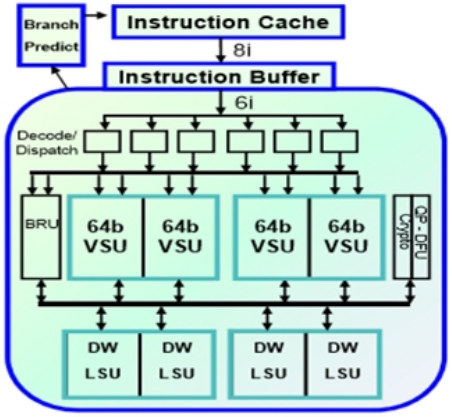


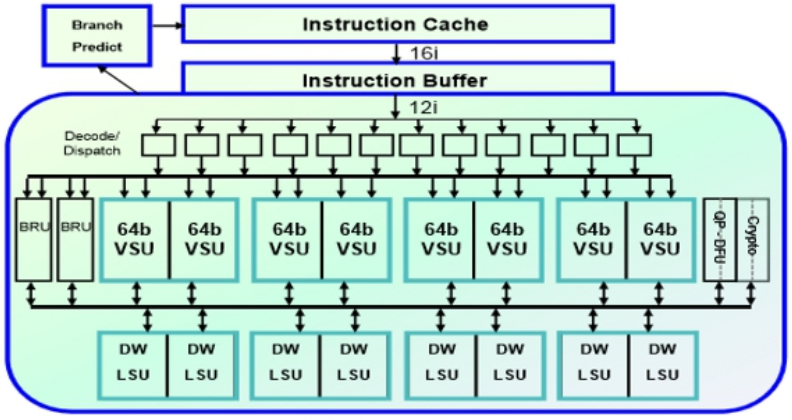
Микропроцессоры IBM Power - Roadmap



Focus on Enterprise Technology and Performance Driven				Focus on Scale-Out and Enterprise Cost and Acceleration Driven					Future
POWER6 Architecture		POWER7 Architecture		POWER8 Architecture		POWER9 Architecture		Partner Chip POWER8/9	POWER10
2007 POWER6 2 cores 65nm New Micro-Architecture New Process Technology	2008 POWER6+ 2 cores 65nm+ Enhanced Micro-Architecture Enhanced Process Technology	2010 POWER7 8 cores 45nm New Micro-Architecture New Process Technology	2012 POWER7+ 8 cores 32nm Enhanced Micro-Architecture New Process Technology	2014 POWER8 12 cores 22nm New Micro-Architecture New Process Technology	2016 POWER8 w/ NVLink 12 cores 22nm Enhanced Micro-Architecture With NVLink	2017 P9 SO 24 cores 14nm New Micro-Architecture Direct attach memory New Process Technology	TBD P9 SU TBD cores 14nm Enhanced Micro-Architecture Buffered Memory	2018 - 20 P8/9 SO 10nm - 7nm Existing Micro-Architecture Foundry Technology	2020+ New Micro-Architecture New Technology
High Frequency Enhanced RAS Dynamic Energy Management		Large eDRAM L3 Cache Optimized VSX Enhanced Memory Subsystem		Optimized for Data-Centric Workloads Integrated PCIe CAPI Acceleration / I/O		Scale-Out Datacenter TCO Optimization Scale-up performance Optimization Acceleration Enhancements to CAPI and NVLINK Modularity for OpenPOWER		OpenPOWER Ecosystem Design Targeting Partner Markets & Systems Leveraging Modularity	New Features and Functions
Price, performance, feature and ecosystem innovation →									

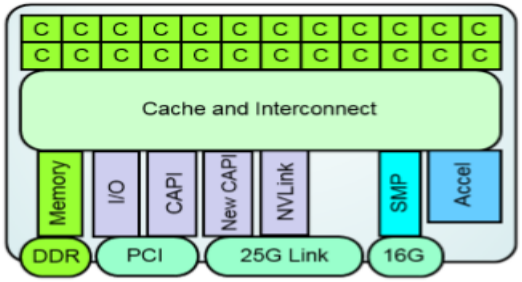
IBM Power 9 – разнообразие ядер

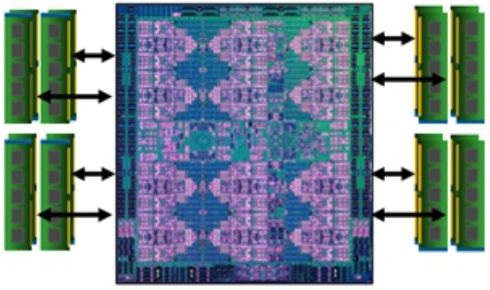


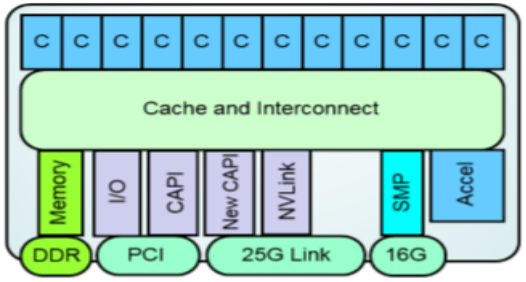


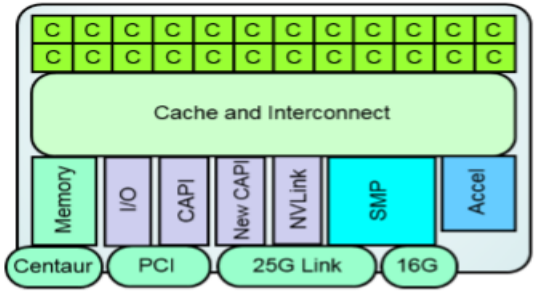
MPI-модель параллельных вычислений

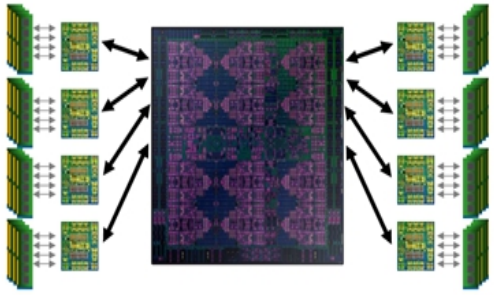
OpenMP или PGAS – модель параллельных вычислений

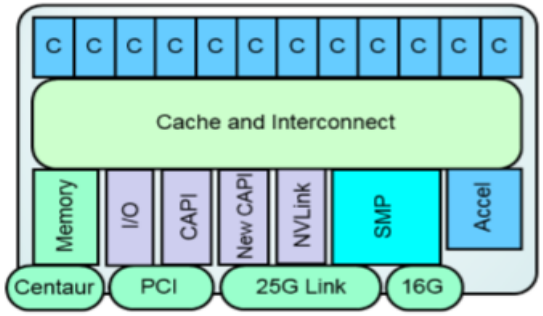




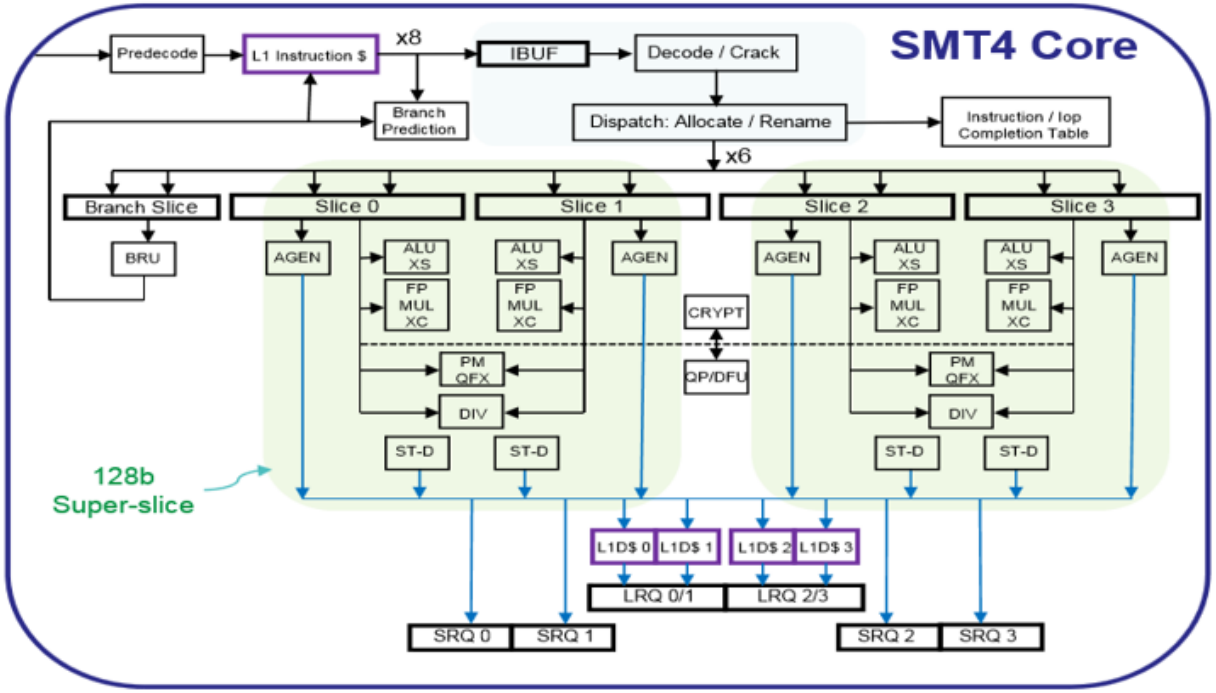








IBM Power 9 - регулярность VSU



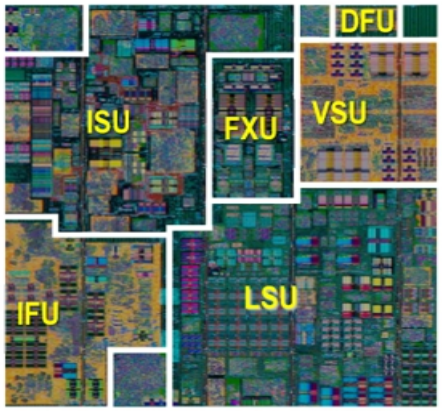
SMT4 Core Resources

Fetch / Branch
- 32kB, 8-way Instruction Cache
- 8 fetch, 6 decode
- 1x branch execution

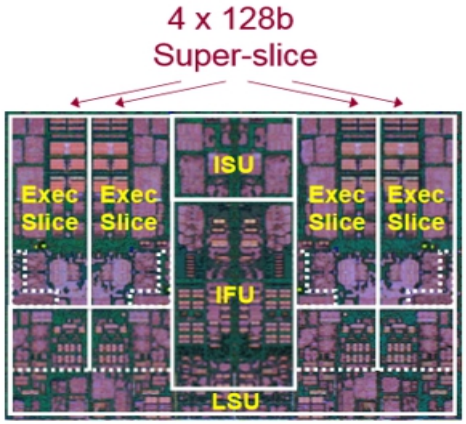
Slices issue VSU and AGEN
- 4x scalar-64b / 2x vector-128b
- 4x load/store AGEN

Vector Scalar Unit (VSU) Pipes
- 4x ALU + Simple (64b)
- 4x FP + FX-MUL + Complex (64b)
- 2x Permute (128b)
- 2x Quad Fixed (128b)
- 2x Fixed Divide (64b)
- 1x Quad FP & Decimal FP
- 1x Cryptography

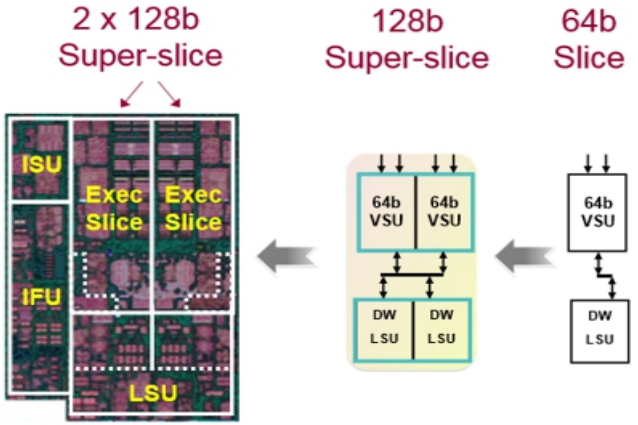
Load Store Unit (LSU) Slices
- 32kB, 8-way Data Cache
- Up to 4 DW load or store



POWER8 SMT8 Core



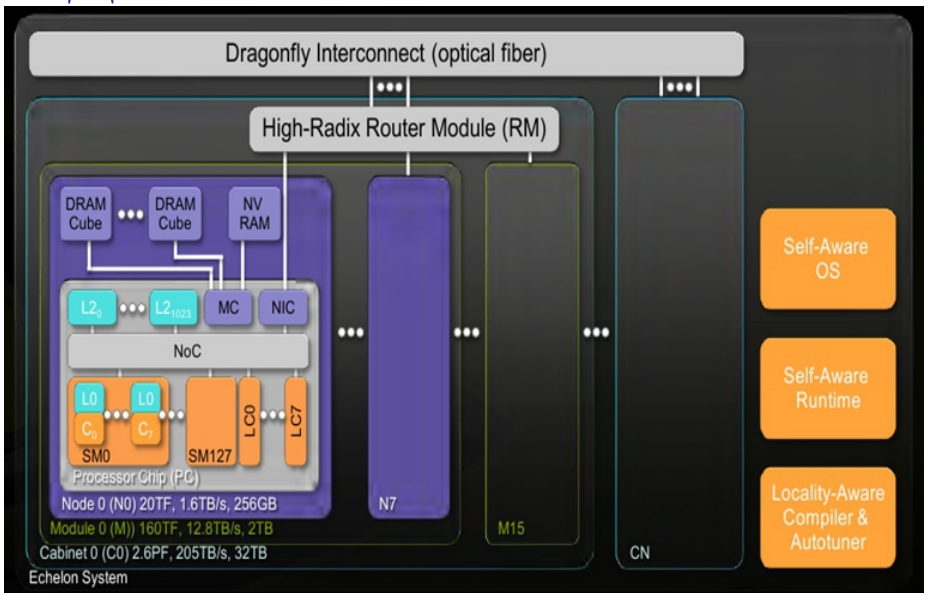
POWER9 SMT8 Core

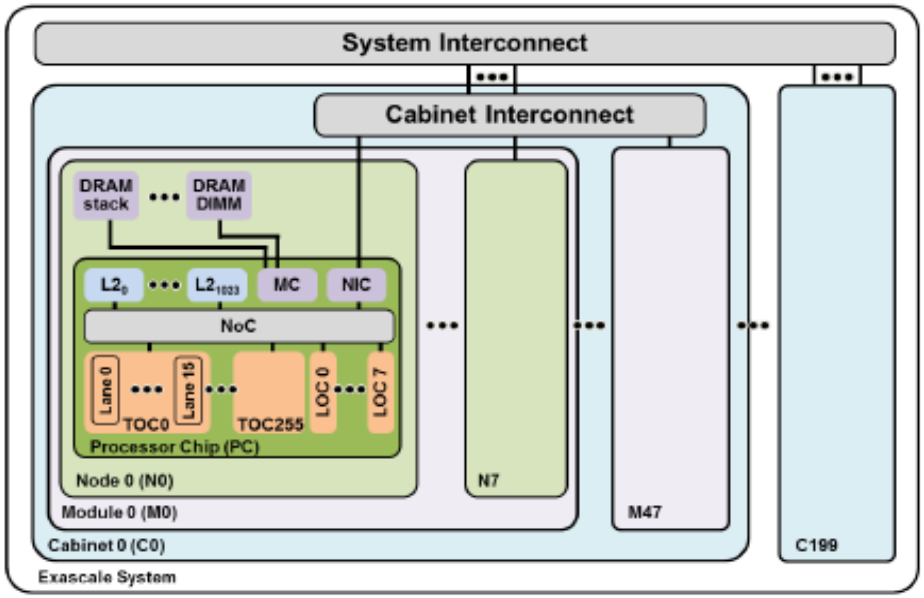


POWER9 SMT4 Core

Проект США
Echelon
гибридная массово-мультитредовая СБИС

Суперкомпьютер Echelon (NVIDIA) – видение 2012 и 2014 годов





Node Configuration	
Technology	7nm
Number of TOCs	512
Number of LOCs	8
LOC Maximum IPC	3.0
DP ALU per TOC	16
DP ALU Total	8,192
L2 per TOC	256 KB
Total L2	128 MB
Memory Controllers	64
Total Memory BW	4 TB/s
NIC Bandwidth	100 GB/s
LOC Frequency	2 GHz
TOC Frequency	1 GHz
TOCs Total DP Perf.	16 TFlops
Processor Area	650 mm^2
Processor peak power	230 W

Exascale System Configuration	
Number of Cabinets	200
Nodes per Cabinet	384
Number of Nodes	76,800
Number of Network Slices	4
Total Router Count	19,200
Peak DP PetaFlops	1,258
Max Node Power	300 W
Max System Power	23 MW

Прогнозируемые характеристики суперкомпьютера Echelon на тестовых приложениях.

Количество стоек - 200
Количество вычислительных узлов - 19200
Теоретическая пиковая производительность - 27 PF/s
Общая мощность потребления - 9 MW

Таблица 1. Характеристики на задачах суперкомпьютера типа ORNL Titan (2012) , узел - 16-ядерный AMD Interlagos + GPU Fermi.

APPS	PetaFlops	PetaOps	MWatts	GFlops/W	GOps/W
CNS	4.61	9.95	5.25	0.88	1.89
CoMD	1.65	13.93	5.32	0.31	2.62
LULESH	4.41	7.82	5.16	0.85	1.51
MiniFE	0.33	4.71	5.08	0.06	0.93
SNAP	0.54	9.15	5.20	0.10	1.76
XSBench	0.52	5.52	5.02	0.10	1.10
LINPACK	17.56	21.95	8.20	2.14	2.68

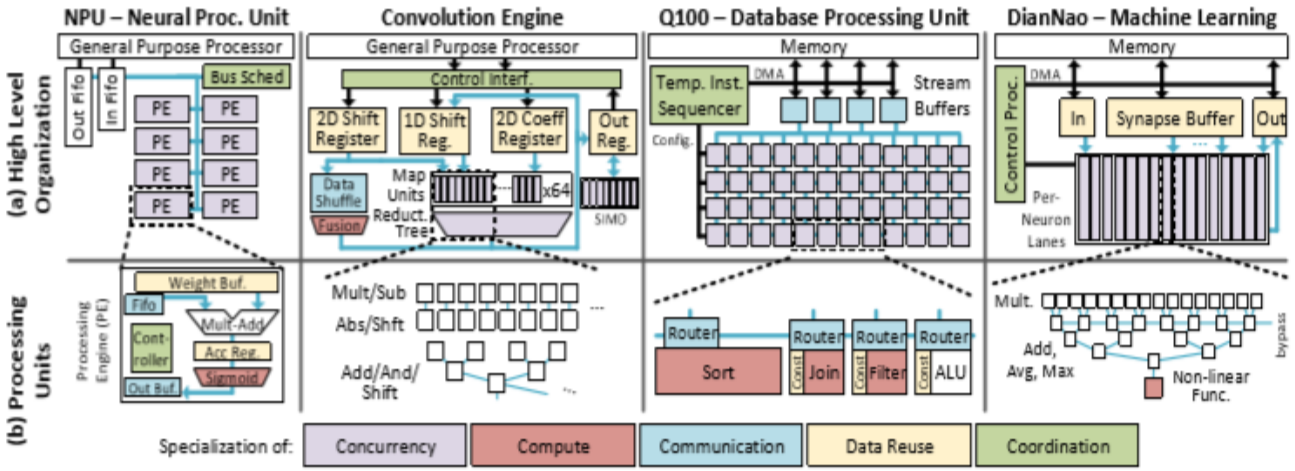
Количество стоек - 200
Количество вычислительных узлов - 76800
Теоретическая пиковая производительность - 1258 PF/s
Общая мощность потребления - 23 MW

Таблица 2. Характеристики на задачах суперкомпьютера Echelon, реализованного на технологии 7 нм (>2020).

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

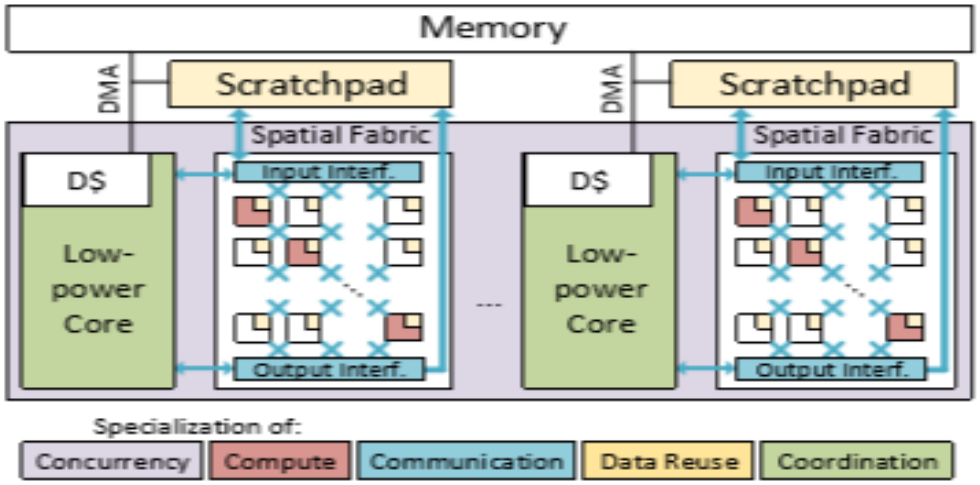
Проект США
DySer/LSSD
массово-многоканальная СБИС

Проблемно-ориентированные СБИС (DSA)и массово-многоканальная СБИС LSSD



4 разных DSA

(что пытаются заменить)



СБИС LSSD

(на что пытаются заменить)

Элементная база России

Реально - только процессоры Эльбрус и сеть Ангара.

Перспектива - процессор Байкал-М;

- векторные процессоры NM-серии

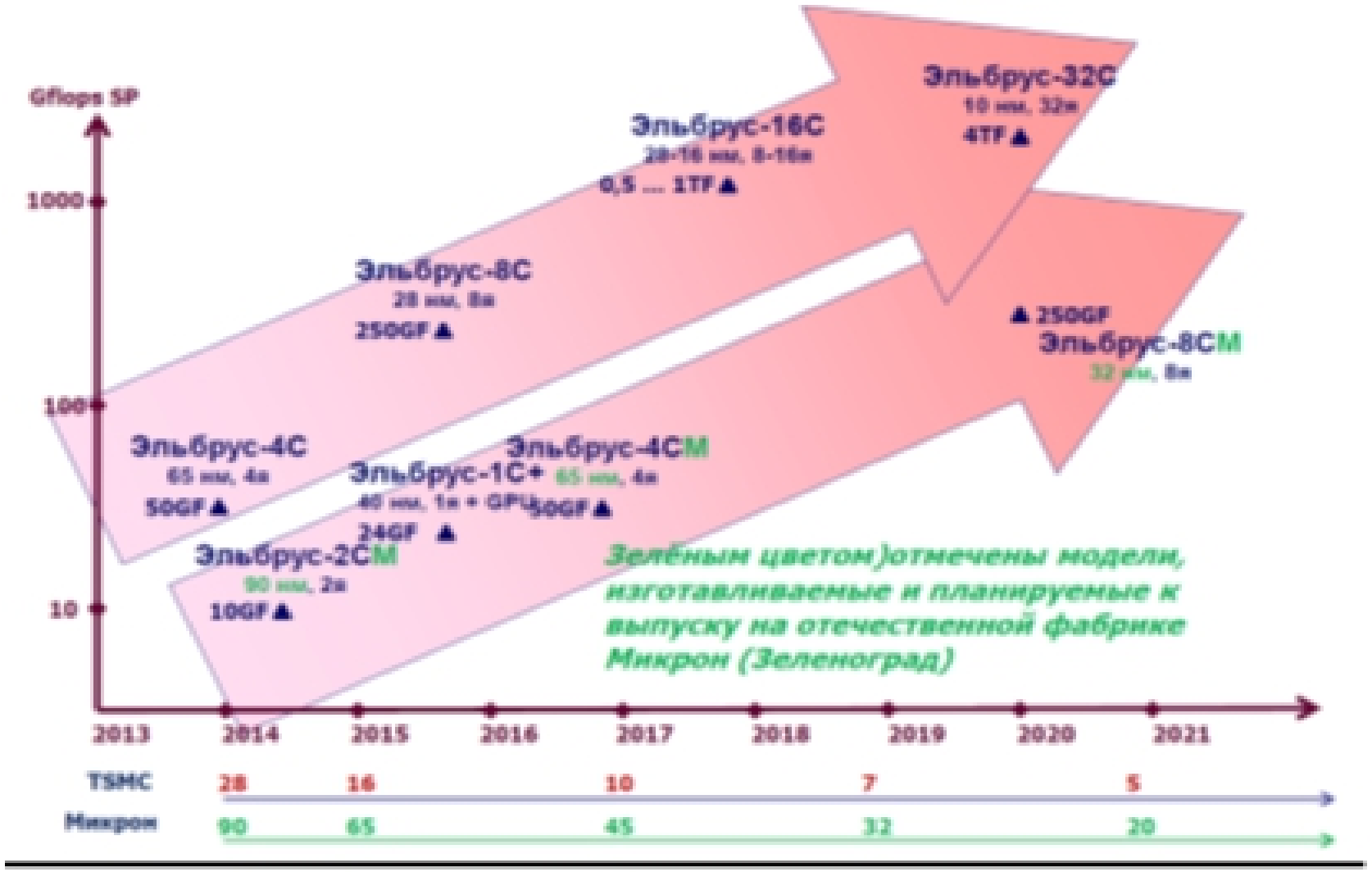
- специализированные процессоры - ускорители

Сравнительные характеристики высокопроизводительных зарубежных и отечественных микропроцессоров

Характеристика	Зарубежные					Отечественные					
	2014			2015		ЗАО "МЦСТ"			ЗАО "НТЦ "Модуль"	ГК "Т-платформы"	
Наименование	1	2	3	4	5	6	7	8	9	10	11
Тип	CPU	GPU	GPU	GPU	MCP	CPU	CPU	CPU	CPU +vect	CPU	CPU
Технология (нм)	22	28	28	28	14	90	65	28	28	28	28
Степени параллелизма потоков команд											
Количество ядер	10	44	15	256	72	2	4	8	21 + +vect	2	8
Количество тредов на ядро	2	10	64	256	4	1	1	1	1	1	1
Количество синхронных потоков в треде или ширина SIMD (разряды)	SIMD 256 р	64	32	SIMD 16х 32р	SIMD 2х 512 р	нет	нет	нет	SIMD 32х 64 р	SIMD 128 р	SIMD 128 р
Итого параллельных потоков команд	20	28160	30720	65536	288	2	4	8	21	2	8
Подсистема памяти											
Пропускная способность памяти (ГБ/с)	59.7	320	249.6	нет	120	12.8	38.4	51.2	32	12.8	34.1
Корпусированная с процессором 3D-память (ГБ/с (ГБ))	нет	нет	нет	512 (4)	500 (16)	нет	нет	нет	нет	нет	нет
Пиковая производительность											
Гфлопс-64	240	1408	1430	н/д	3000	8	23	125	128	9.6	64
Гфлопс-32	480	5632	4291	8900	6000	16	46	250	512	19.2	128
Гопс-32	120	1408	4291	н/д	6000	8	23	104	16	24	128

1 – E5-2690 v2 (Ivy Bridge EP), 3 GHz
2 – AMD Radeon R9 290X Hawaii, 1 GHz
3 – NVIDIA Tesla K40, 0.745 GHz
4 – AMD Fiji
5 – Intel Knight Landing, 1.3 GHz
6 – Эльбрус-2С+, 0.5 GHz
7 – Эльбрус-4С, 0.72, GHz
8 – Эльбрус-8С, 1.3 GHz
9 – NM6408MP, 1GHz
10 – Байкал-Т, 1.2 GHz
11 – Байкал –М, 2 GHz

Дорожная карта микропроцессоров Эльбрус



Сравнительные характеристики зарубежных и отечественных коммуникационных сетей

Характеристика	Зарубежные				Отечественные
	Заказные (недоступные)			Коммерческие	Коммерческое
	Tofu-2 (Fujitsu)	PERCS (IBM)	Aries (Cray)	FDR Infiniband (Mellanox)	Ангара (ОАО "НИЦЭВТ")
Топология	6D-тор	Иерархическая многосвязная	Иерархическая многосвязная	Толстое дерево	3D тор
BW интерфейса с маршрутизатором узла (ГБ/с)	20	96	16	16	8
Количество NIC-интерфейсов маршрутизатора (коммутатора)	1	4	4	36	1
BW линка маршрутизатора (ГБ/с)	12.5	24 5 10	5.25 15.75 13.4	6.8	8.0
Общая BW линков маршрутизатора на узел (ГБ/с)	10 x 12.5 = 125	7 x 24 + 24 x 5 + 16 x 10 = 448	15 x 5.25 + 5 x 15.75 + 5 x 13.4 = 224	6.8 x 36 = 245	6 x 8.0 = 48

Будущее,
что нам в России делать?

.... Мир скоро окажется перед пропастью отсутствия микроэлектронных технологий и думает, что в этом случае делать....

.... Наше преимущество - мы оказались перед этой пропастью раньше, даже это наше обычное состояние, поэтому давно придумываем решения, имеем наработки и опыт ...

Выводы

1. Для вычислительно емких специальных приложений с короткими, но много раз повторяющимися алгоритмами возможно повышение энергоэффективности программируемых процессоров в 20+ раз, если использовать двухуровневую память команд и регистров с локализацией нижнего уровня при функциональных устройствах. Это дает возможность заменить блоки с прямой реализацией таких алгоритмов программируемыми процессорами (фундаментальный результат, экспериментально получен в США и России)

2. Для переборных задач с интенсивным использованием коротких, много раз повторяющихся алгоритмов, возможно применение массово-многоканальных СБИС, каждый канал которых состоит из легкого энергоэффективного RISC-процессора с подключенным к нему ускорителем с динамически реконфигурируемой и локально программируемой проблемно-ориентированной архитектурой, в котором также применяются новые методы повышения энергоэффективности.

Выводы

3. Для обеспечения толерантности процессорных СБИС к задержкам выполнения операций с памятью или других длинных операций, а также повышения энергоэффективности за счет новых алгоритмов выборки и выполнения команд целесообразно применение мультитредовых архитектур.

4. Для эффективной реализации статически реконфигурируемых вычислительных устройств необходимы исследования по созданию коммутаторов с большим количеством портов (мемристоры?).

5. Суперкомпьютеры класса RM и CO могут быть построены на массово-многоканальных СБИС, в состав которых введены мультитредовые ядра для работы с внешней DDR-памятью, которая может применяться для хранения таблиц с константами или другими данными.

6. Суперкомпьютеры класса GP млогут быть построены на мультитредовых СБИС с поддержкой выполнения средних и тяжелых тредов, классифицируемых так по объему выполняемых в них вычислений с сравнении с операциями с памятью.

Выводы

7. Суперкомпьютеры класса СВ могут быть построены на мультитредовых СБИС с поддержкой легких тредов.

8. Необходима срочная организация работ по 3D модулям памяти высокой пропускной способности (HBM-типа) со встроенными в них процессорными элемантами (PIM-процессоры).

9. Отечественные разработки по сетевым СБИС необходимо переориентировать на маршрутизаторы многосвязных сетей (high-radix) и использование в них оптоэлектронных соединений.

10. Целесообразно организовать создание мощного универсального процессора-менеджера с большим количеством высокоскоростных интерфейсов для подключения разного рода СБИС-ускорителей. Возможно применение лицензирования зарубежных образцов.

Вопросы ?

Эйсымонт Леонид Константинович
verger-lk@yandex.ru