Четвертый Московский Суперкомпьютерный Форум
(23 октября 2013)

Обзор проектов экзафлопсных суперкомпьютеров за рубежом и в России, ограничения и перспективы роста

В.С.Горбунов, Г.С.Елизаров, Л.К.Эйсымонт
(ФГУП «НИИ «Квант»)

Публикации, в которых изложены основные положения доклада

1. **В.Горбунов, Г.Елизаров, Л.Эйсымонт. Экзафлопсные суперкомпьютеры: достижения и перспективы. Открытые системы, N7, 2013.**
2. **Д.Андрюшин, В.Горбунов, Л.Эйсымонт. Перспективные особенности Tianhe-2. Открытые системы, N8, 2013.**
3. **В.Горбунов. HPC. Оценить, измерить, оптимизировать. Суперкомпьютеры, 3(15) осень 2013, стр.50-55.**
4. **В.Горбунов, Л.Эйсымонт. Комплексная методика тестирования производительности суперкомпьютеров, профессиональный подход. Вычислительные методы и программирование. 2013 (в печати).**

Общая картина в области СКТ (на примере США)

- Внедрение результатов программы DARPA HPCS (2002-2010), коммерческие образцы и заказные суперЭВМ (2013-2017)
- Выполнение программы DARPA UHPC (2010-2020) и программ DoE по экзамасштабным технологиям и суперЭВМ экза-уровня
- Выполнение программы DARPA STARNet (с 2013 года) по оптимизации использования КМОП-технологий и разработки технологий пост-Муровской эры, зетта- и йотта-уровень

Экзафлопсные проекты США

DARPA UHPC – проекты Runnemede (Intel), Echelon (NVIDIA, Cray), X-Calibr(Sandia), Angstrom (MIT,Tilera)

DoE – проекты FastForward, C— Design Centers, X-Stack, OS/R, проекты крупных лабораторий и центров

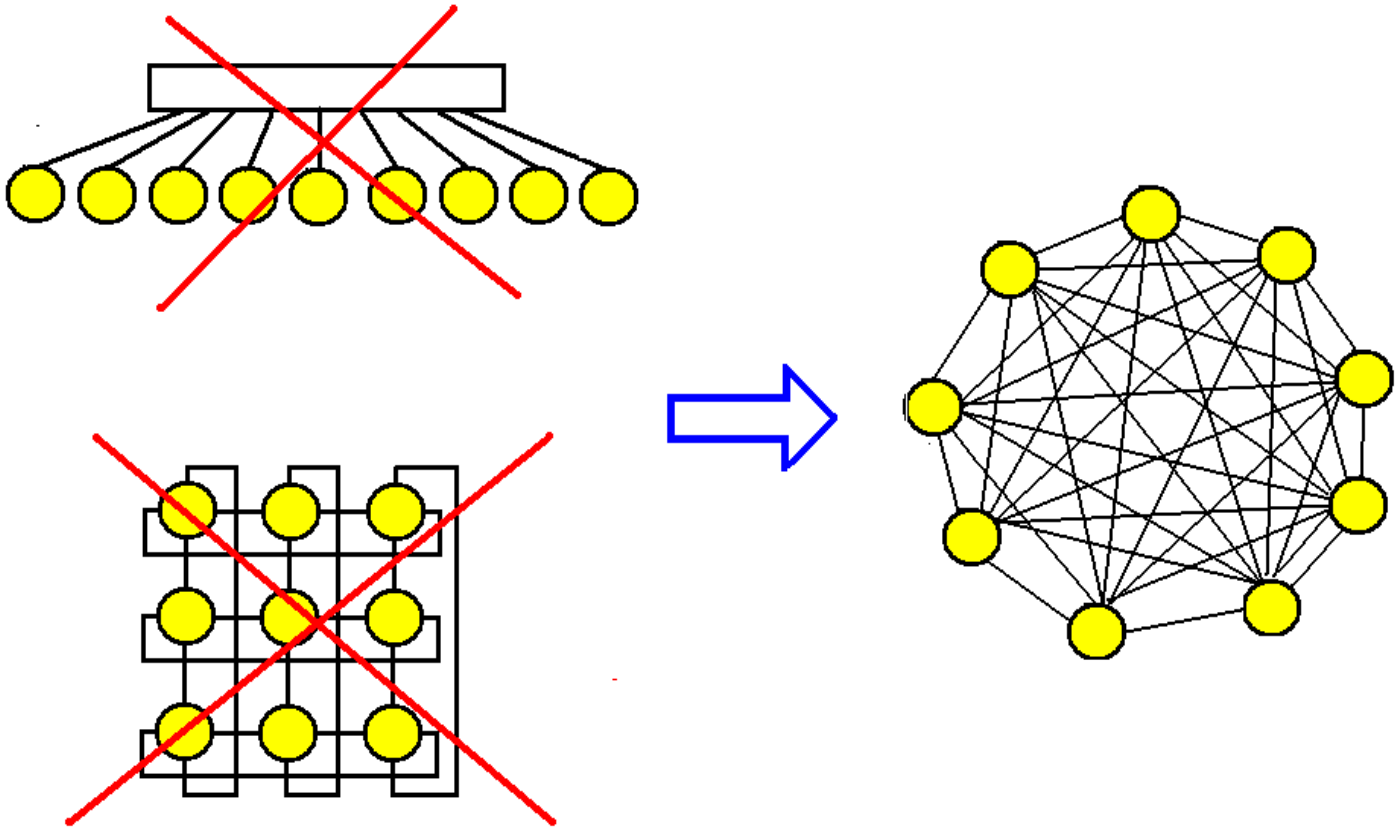
Специальные проекты, выполняемые крупными фирмами – Intel, Cray, IBM, HP, Google…

Ожидаемые суперЭВМ экза- и выше уровня в США

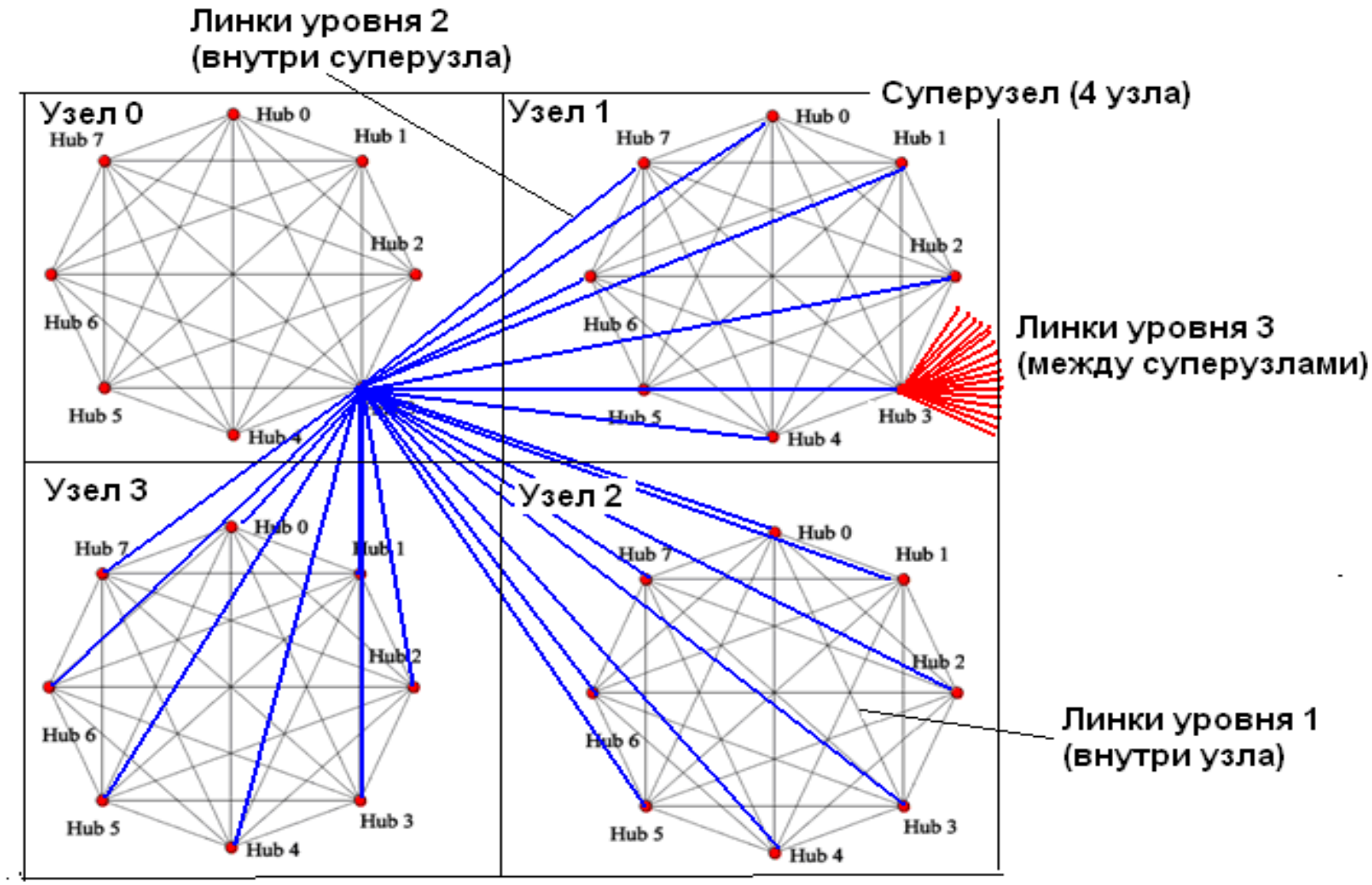
-2015-2017 – заказные суперкомпьютеры экза-уровня, оптимизация под CF- и DIS-задачи

-2018-2020 – эволюционная суперЭВМ экзафлопсного уровня NNSA DoE
- после 2022 – инновационная суперЭВМ экзафлопсного уровня OS/ASCR DoE
- после 2020 - заказные специализированные суперкомпьютеры зетта-уровня (~ 2020) и йотта-уровня (~ 2024), технологии RSFQ, QCA и квантовые аналогово-спиновые (типа D-Wave)

ИЕРАРХИЧНОСТЬ

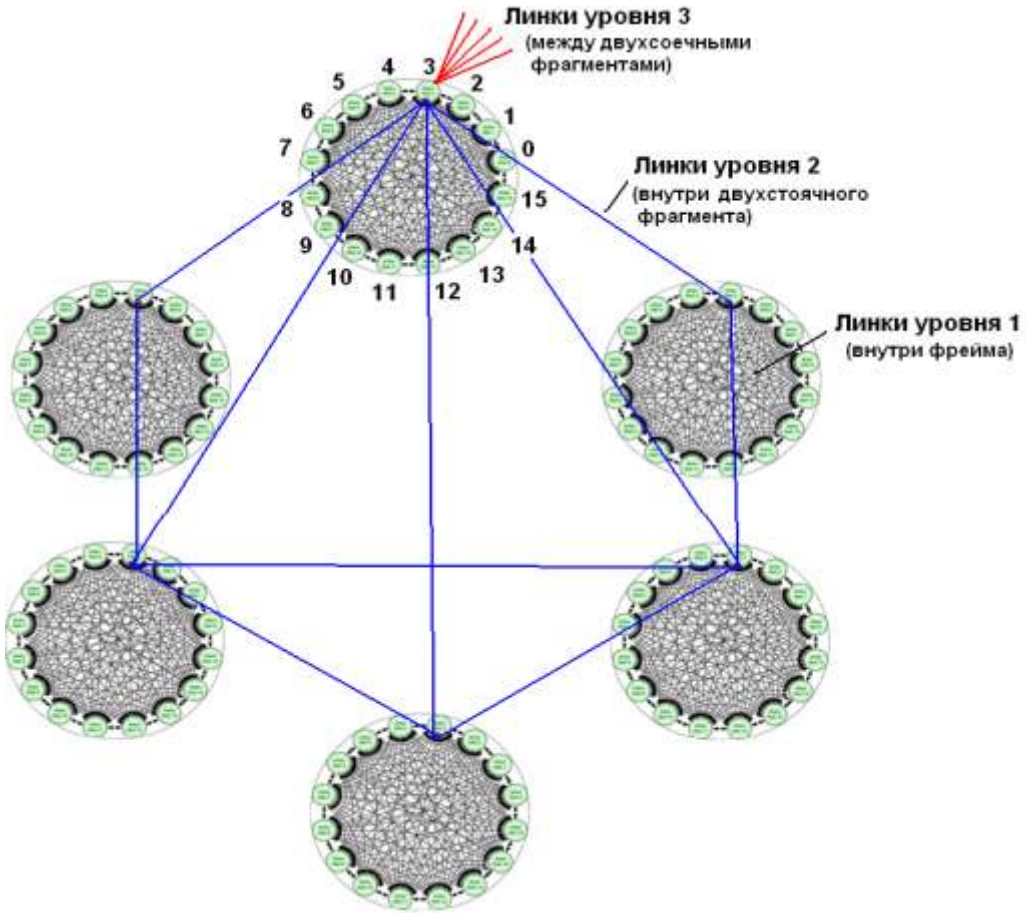


Многоуровневая сеть PERCS суперкомпьютера Power 775

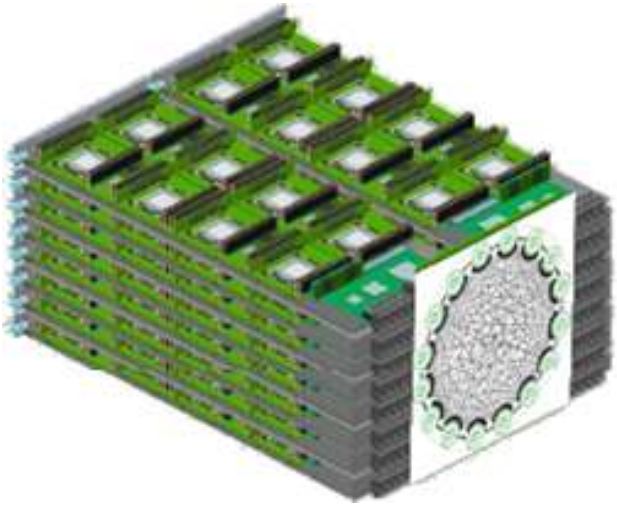


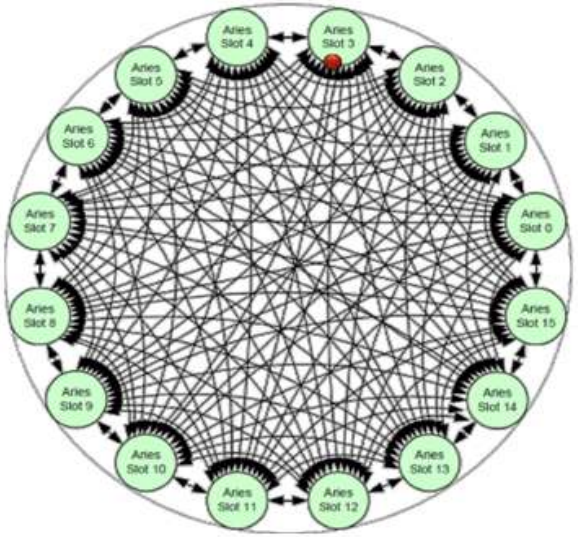
Многоуровневая сеть суперкомпьютера Cray XC30

Двухстоечный фрагмент

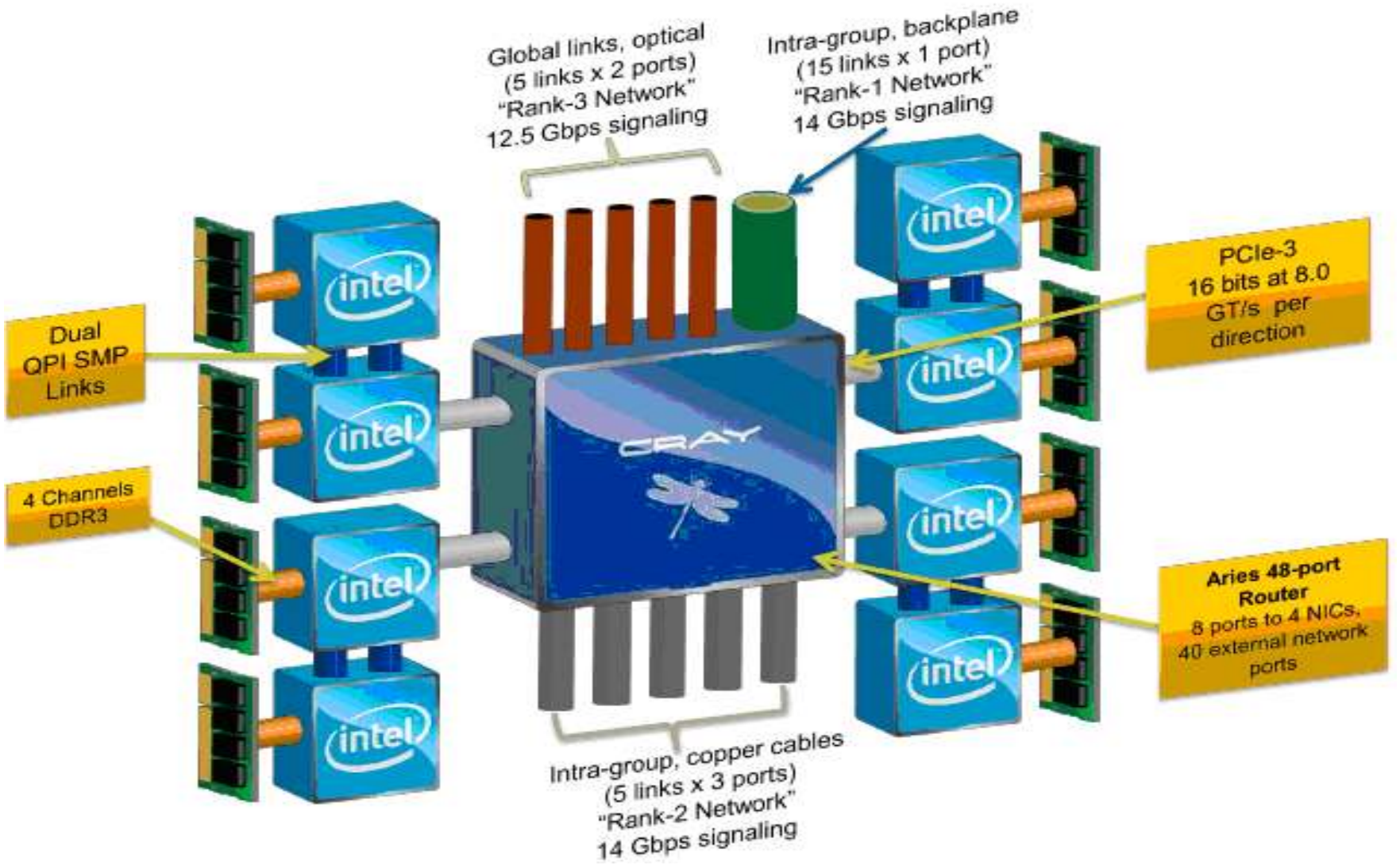


Фрейм





Вычислительное лезвие Cray XC30 с маршрутизатором сети Aries



Повышение производительности за счет перехода к сети Aries (256-512 узлов)

Tests (Units)	XE-Interlagos	XC	XC/XE
HPL (Tflops)	~81%	~86%	106%
Star DGEMM (Gflops)	~87%	~102%	117%
STREAMs (Gbytes/s/node)	72	78	108%
RandomRing (Gbytes/s/rank)	~0.055	~0.141	256%
Point-to-Point BW (Gbytes/s)	2.8-5.6	>8.5	157% - 314%
Nearest Node Point-to-Point Latency (usec)	1.6-2.0	<1.4	116% - 145%
GUPs	2.66	15.6	525%
GFFT (Gflops)	628	2221	354%
HAMR Sort (GiElements/sec)	9.4	36.6	390%

Рекорд (ноябрь 2012) - 2021 GUPS, P7(14 суперузлов, 1792 процессора) - Power 775+GPU

ГЕТЕРОГЕННОСТЬ

(специализированные фрагменты)

Tianhe-1A - 2048 FT-1000 (8 ядер, 64 треда,1 GHz)

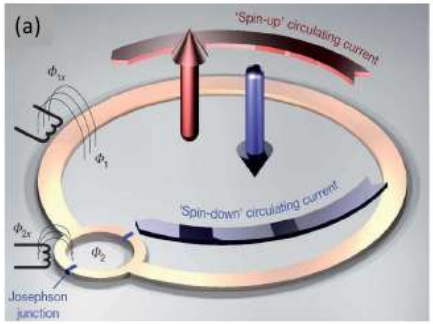
Tianhe-2 - 4096 FT-1500 (16 ядер, 256 тредов, 1.8 GHz, 144 GFlops, 60 W)

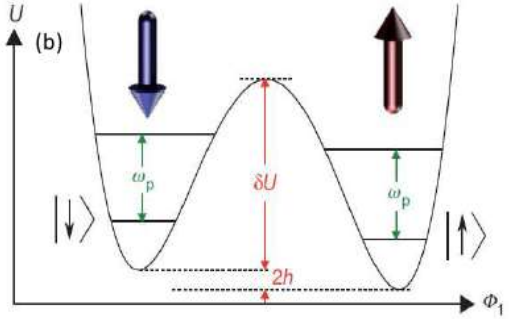
Специализированные блоки, даже аналоговые – например, D-Wave

Квантовый аналогово-спиновый суперкомпьютер D-Wave - 1

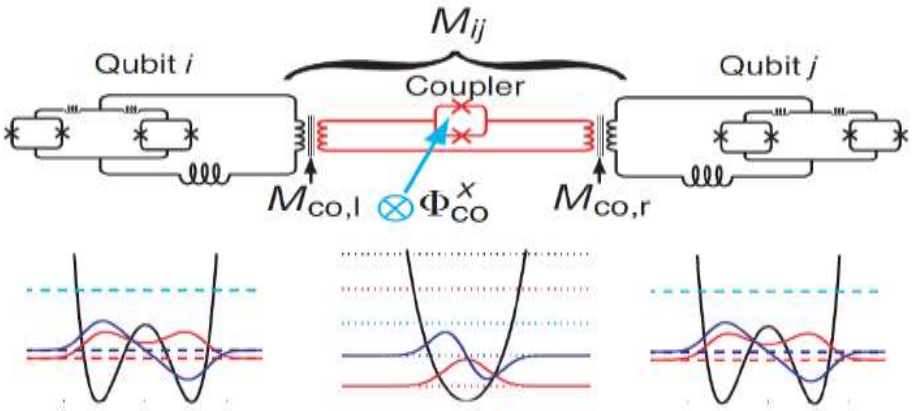
$$E_{\mathrm{Ising}}(s_1,\ldots,s_N) = -\sum_{i=1}^{N} h_i s_i + \sum_{\langle i,j\rangle \in \mathrm{E}} J_{i,j} s_i s_j$$

Вычисление, которое может выполнять D-Wave, si –спины, +1 или -1, hi и Ji,j – настроечные коэффициенты

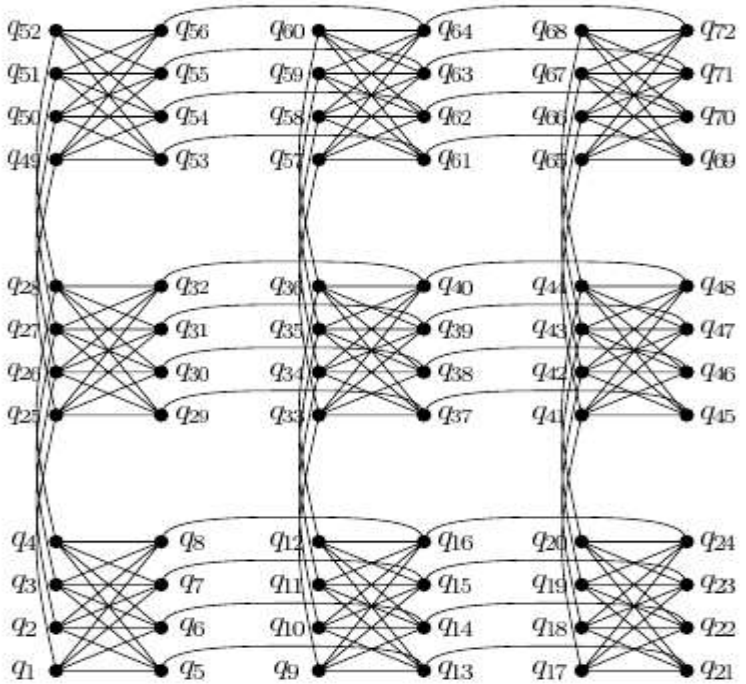




Один q-бит

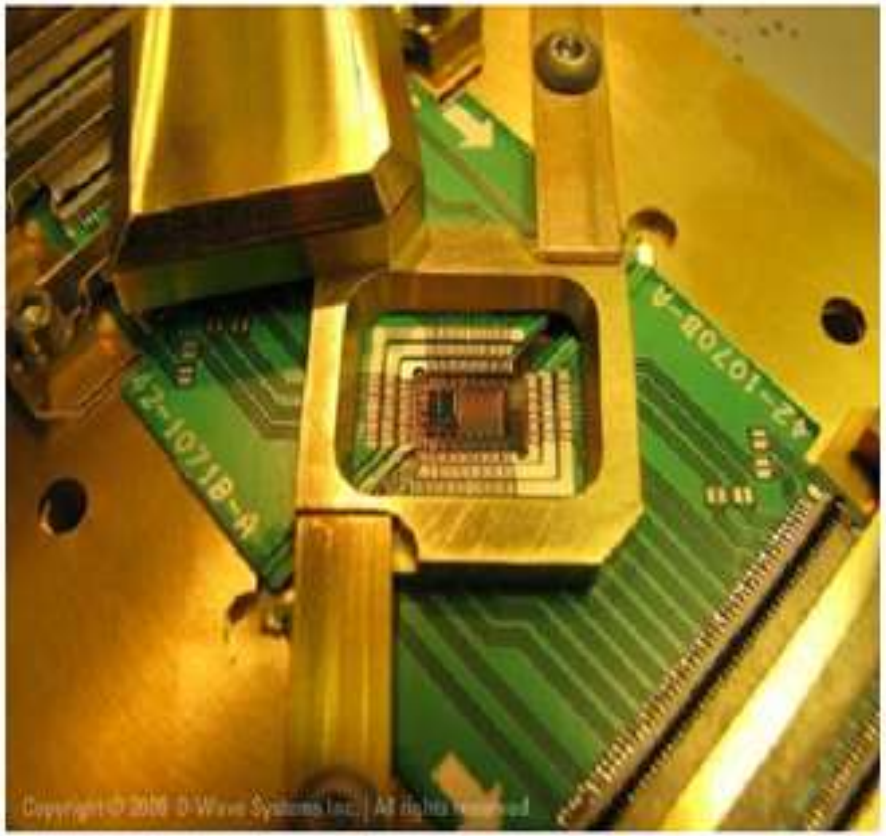


Соединение двух q-битов

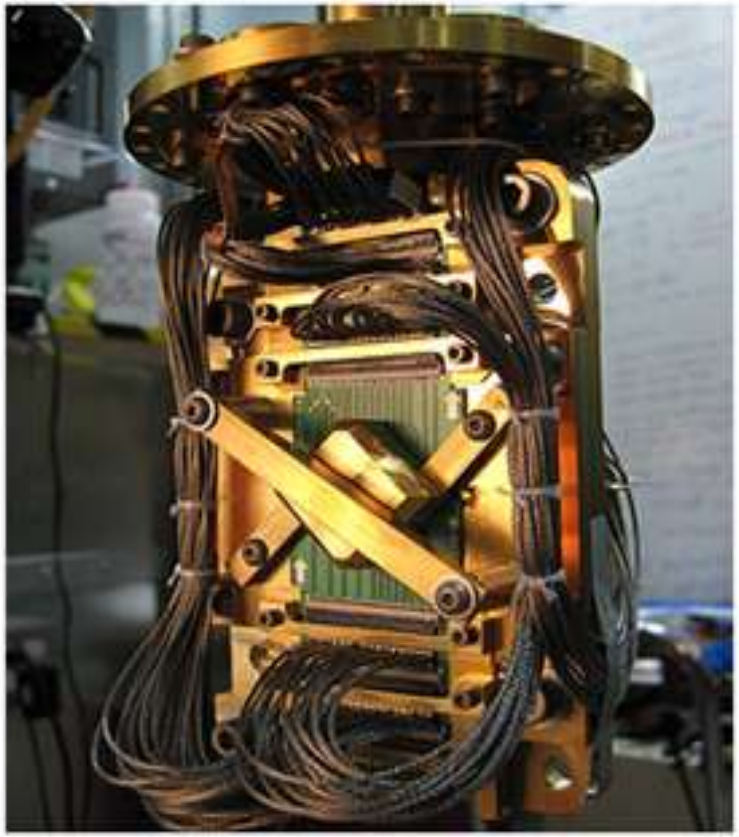


Логическое соединение 128 q-битов

Квантовый аналогово-спиновый суперкомпьютер D-Wave - 2



A. Процессор в камере непосредственного жидкостного охлаждения



B. Процессор в конструктиве охлаждения и защиты

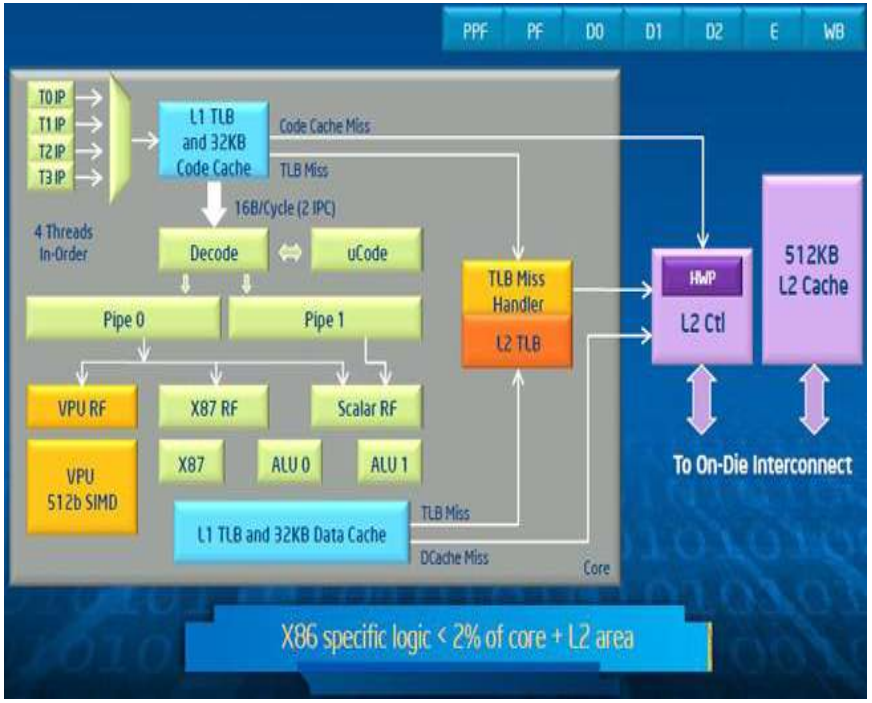
Рабочая температура ~ 20 mK

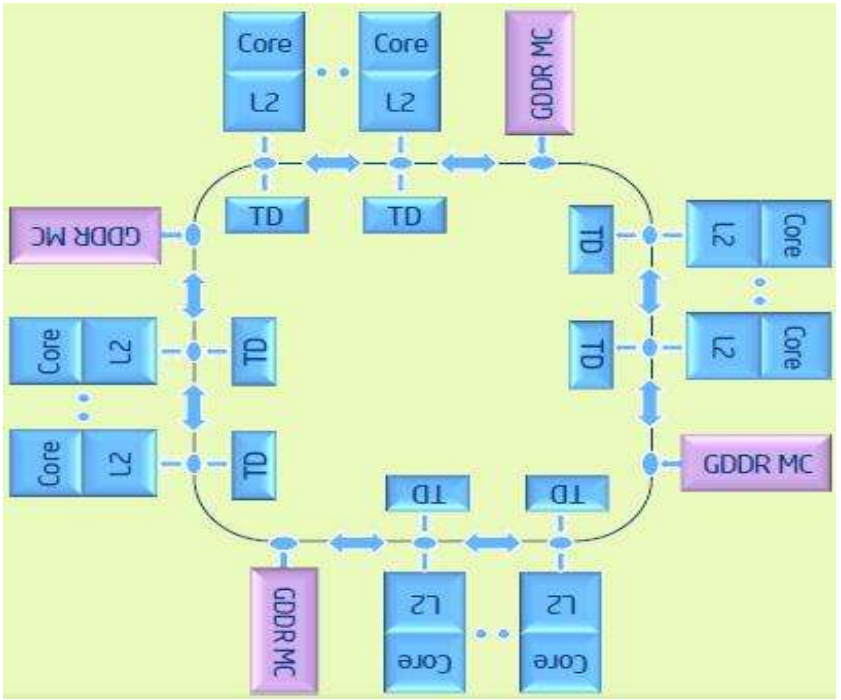
Квантовый аналогово-спиновый суперкомпьютер D-Wave - 3

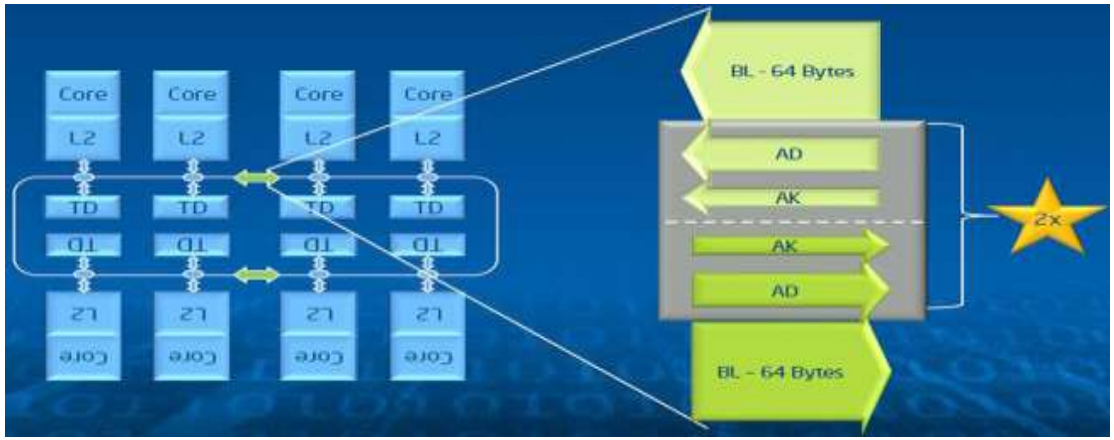
ГИБРИДНОСТЬ

Массово-многоядерные мультитредовые микропроцессоры, однородные и гибридные

Xeon Phi, микроархитектура

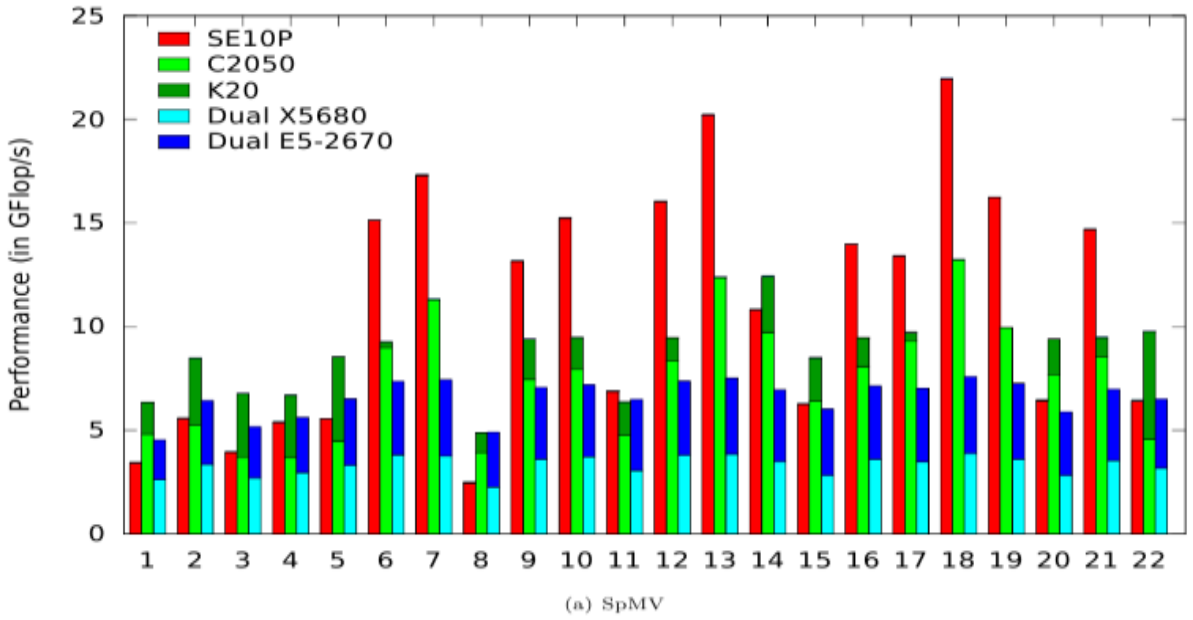






HPCG (SpMV) против HPL(Top500)

Установка	Реальная производительность на SpMV (% от пиковой)
2 x Intel Xeon X5680 (Westmere)	0,78 - 1,09
2 x Intel Xeon E5-2670 (Sandy Bridge)	1,36 - 2,12
NVIDIA Tesla C2050 (Fermi).	0,68 - 2,52
Tesla K20 (Kepler)	0,5 - 1,3
Intel Xeon Phi	0,25 - 2,25



2xW - 320 GF

2xSB – 330 GF

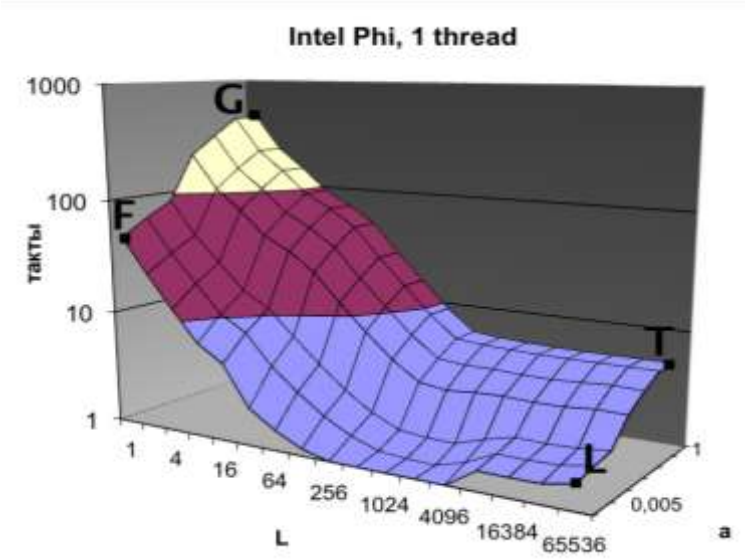
F - 512 GF

K - 1 TF

Phi - 1 TF

Saule E. *et al*. Performance Evaluation of Sparse Matrix Multiplication Kernels on Intel Xeon Phi. 5 Feb 2013, 19 pp

Xeon Phi – SandyBridge: APEX-поверхности



Пиковые производительности ядер

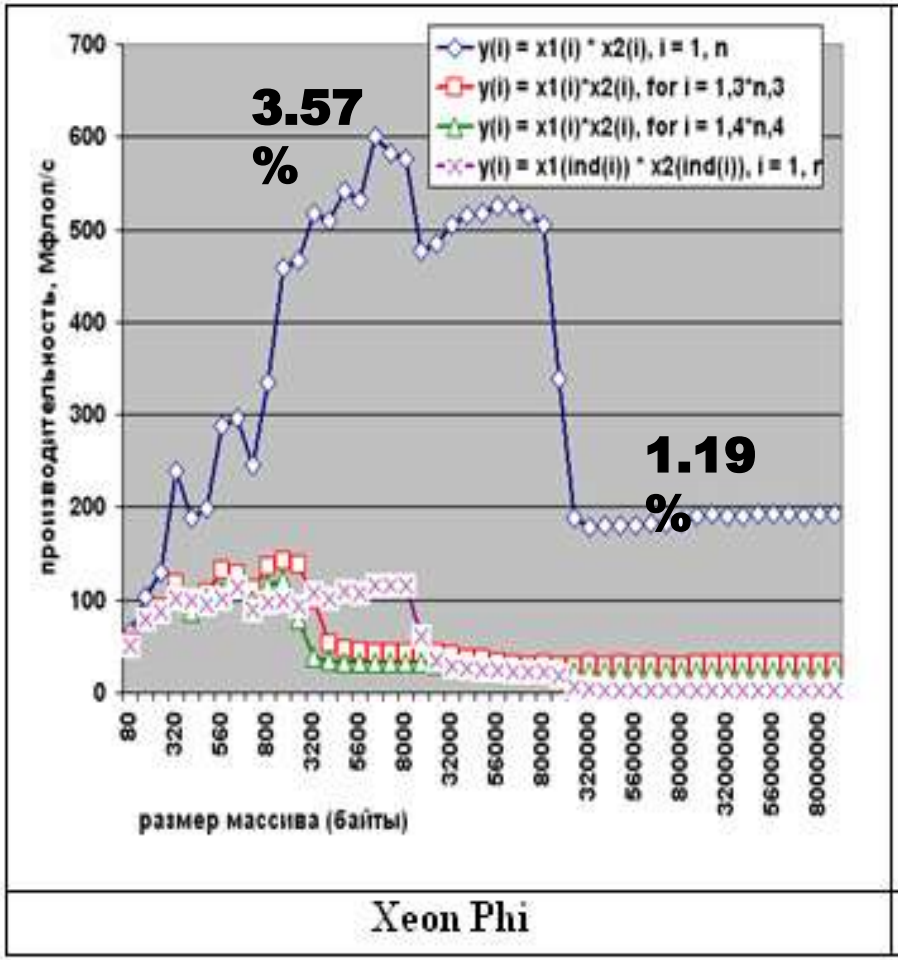
Xeon Phi – 16.8 GFlops

Sandy Bridge – 17.6 GFlops

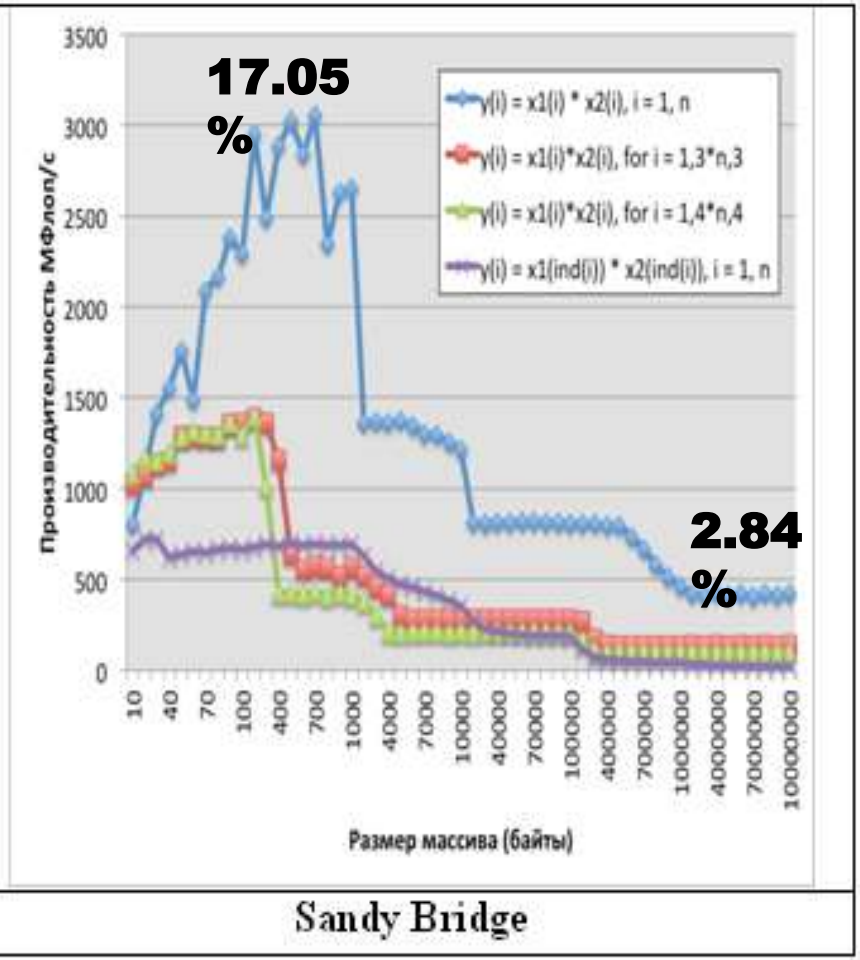
Таблица. Задержки выполнения обращений на считывание в предельных режимах

Точки локализации	Xeon Phi			Xeon Sandy Bridge		
	1 тред	60 тредов	120 тредов	1 тред	16 тредов	32 треда
Точка L	1,7	0,2	0,3	1,1	0,1	0,1
Точка G	442,8	8,6	4,6	229,8	15,4	15,9
Точка F	48,1	2.0	0,8	7,5	0,8	0,5
Точка T	5,1	0,2	0,2	1,9	0,4	0,4

Xeon Phi – SandyBridge: тесты усложнения доступа к памяти

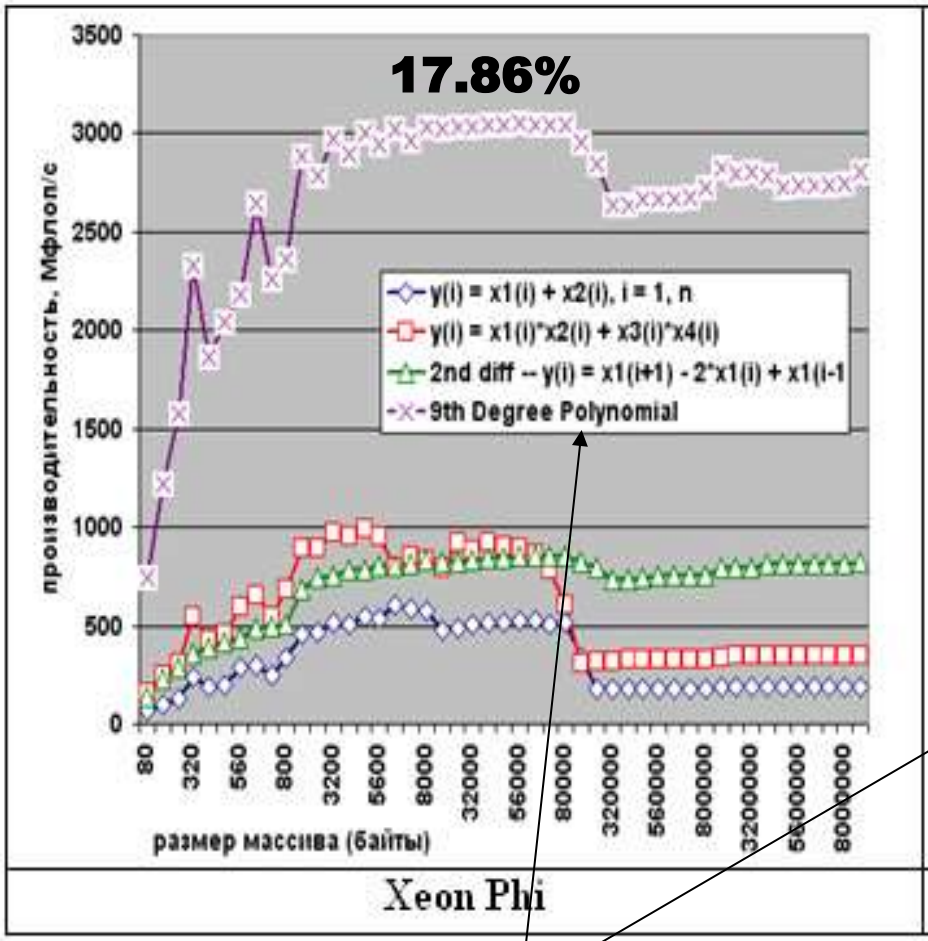


Xeon Phi

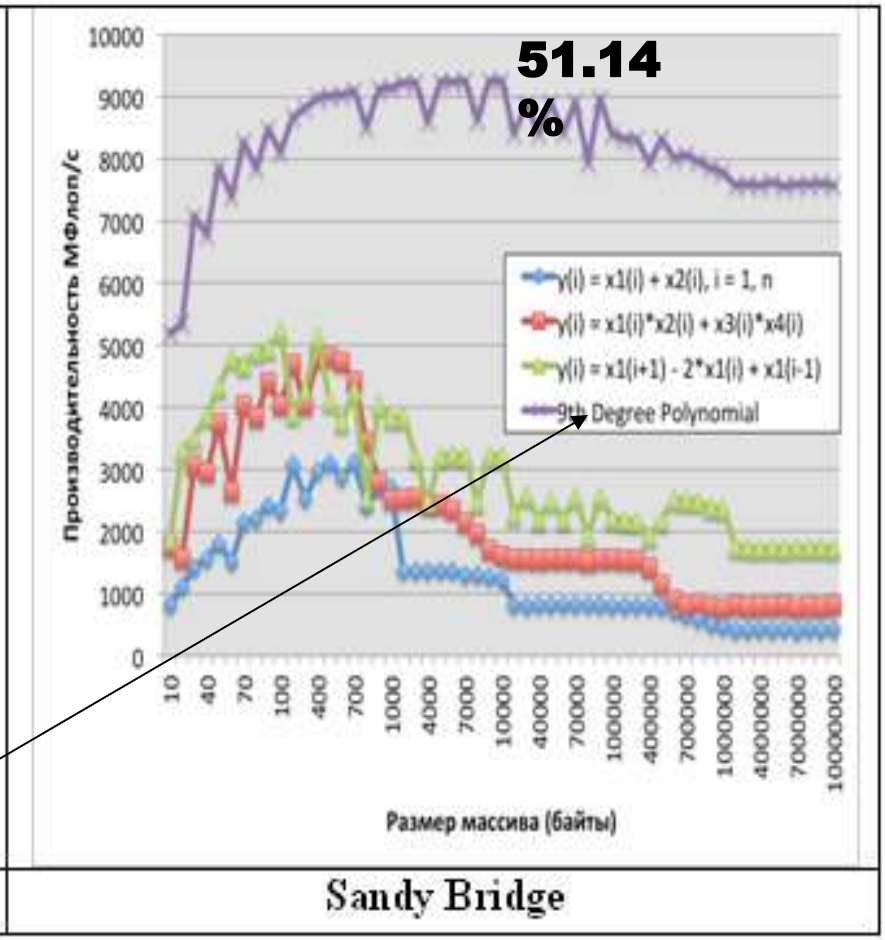


Sandy Bridge

Xeon Phi – SandyBridge: тесты повышения доли вычислений



Xeon Phi



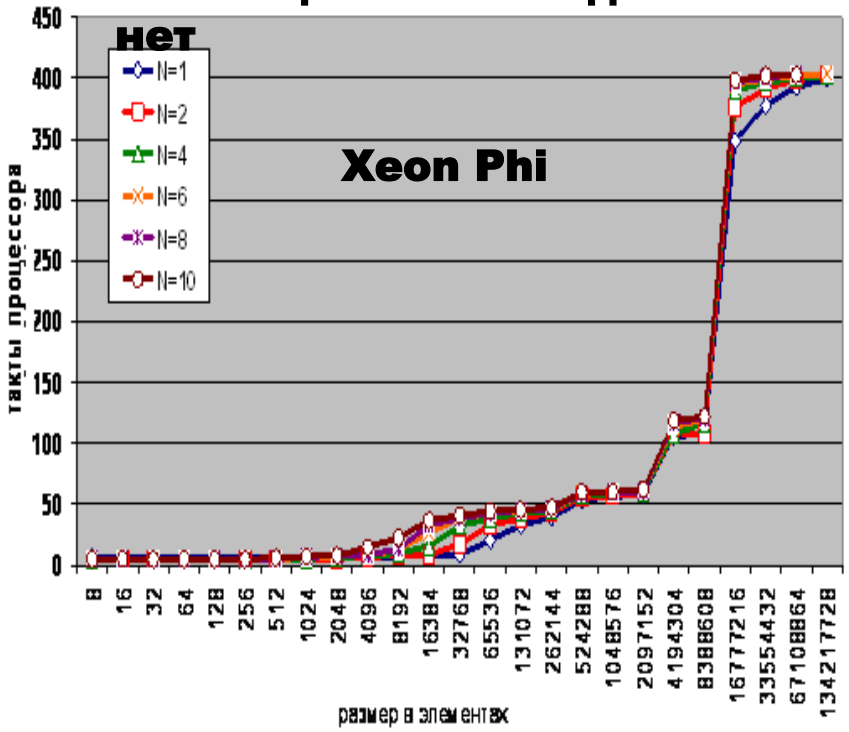
Sandy Bridge

$$a_3 x^3 + a_2 x^2 + a_1 x + a_0 \longrightarrow x\,(\,x\,(\,a_3 x + a_2\,) + a_1\,) + a_0$$

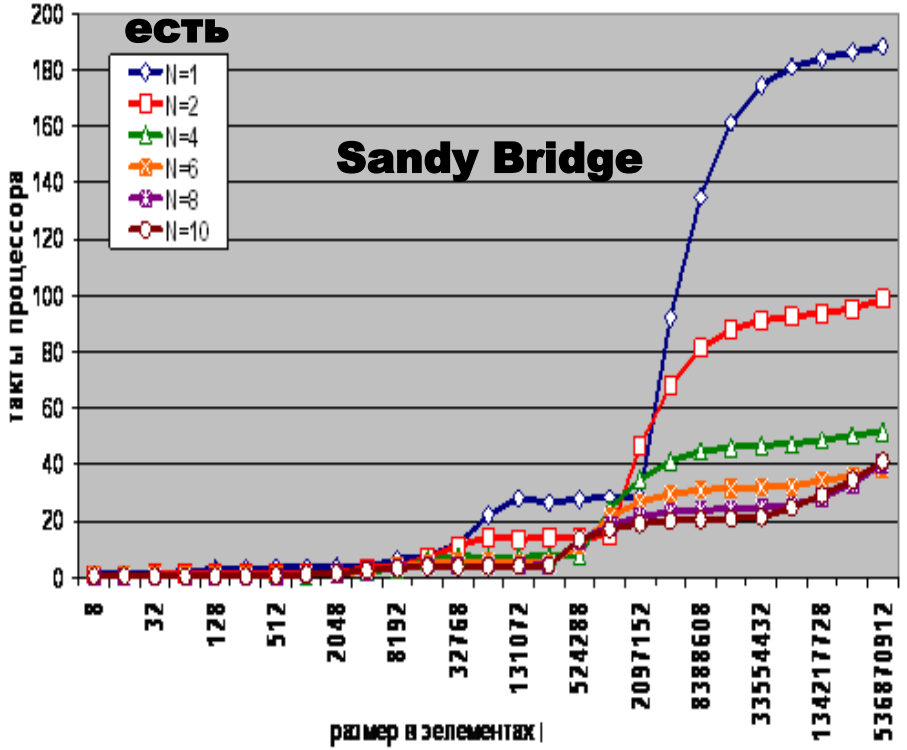
Xeon Phi – SandyBridge: исследование задержек - 1

```
while(count-- > 0)
  { list1  = list1.next;
    list2  = list2.next;
    ...
    listN  = listN.next }
```

Совмещения команд потока нет



Совмещение команд потока есть



Xeon Phi – SandyBridge: исследование задержек - 2

Поток 1

```
while(count-- > 0)
  { list1  = list1.next }
```

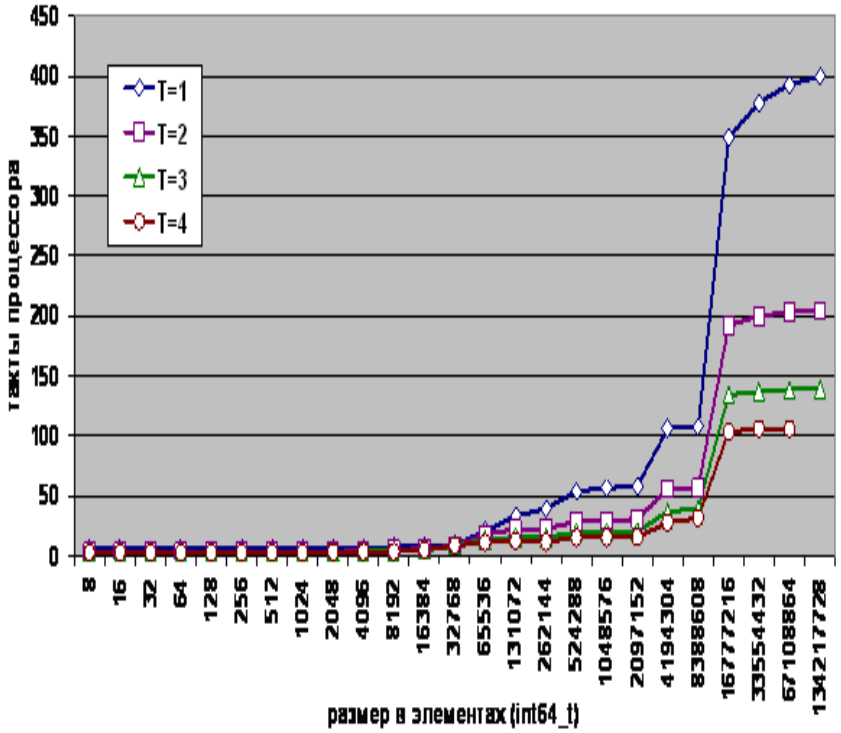
Поток 2

```
while(count-- > 0)
  { list2  = list2.next }
```

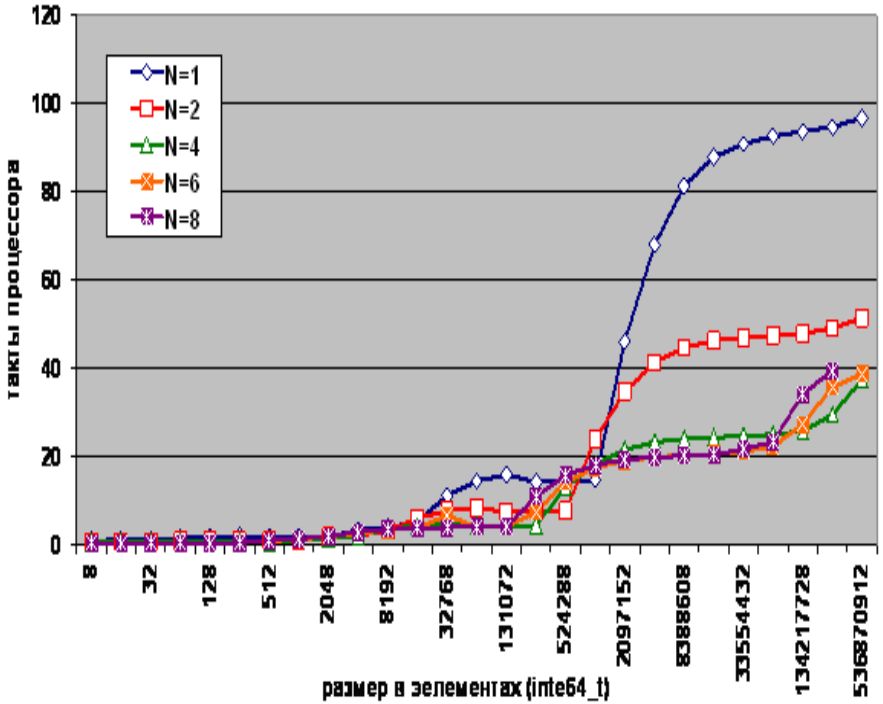
. . .

Поток N

```
while(count-- > 0)
  { listN = listN.next }
```

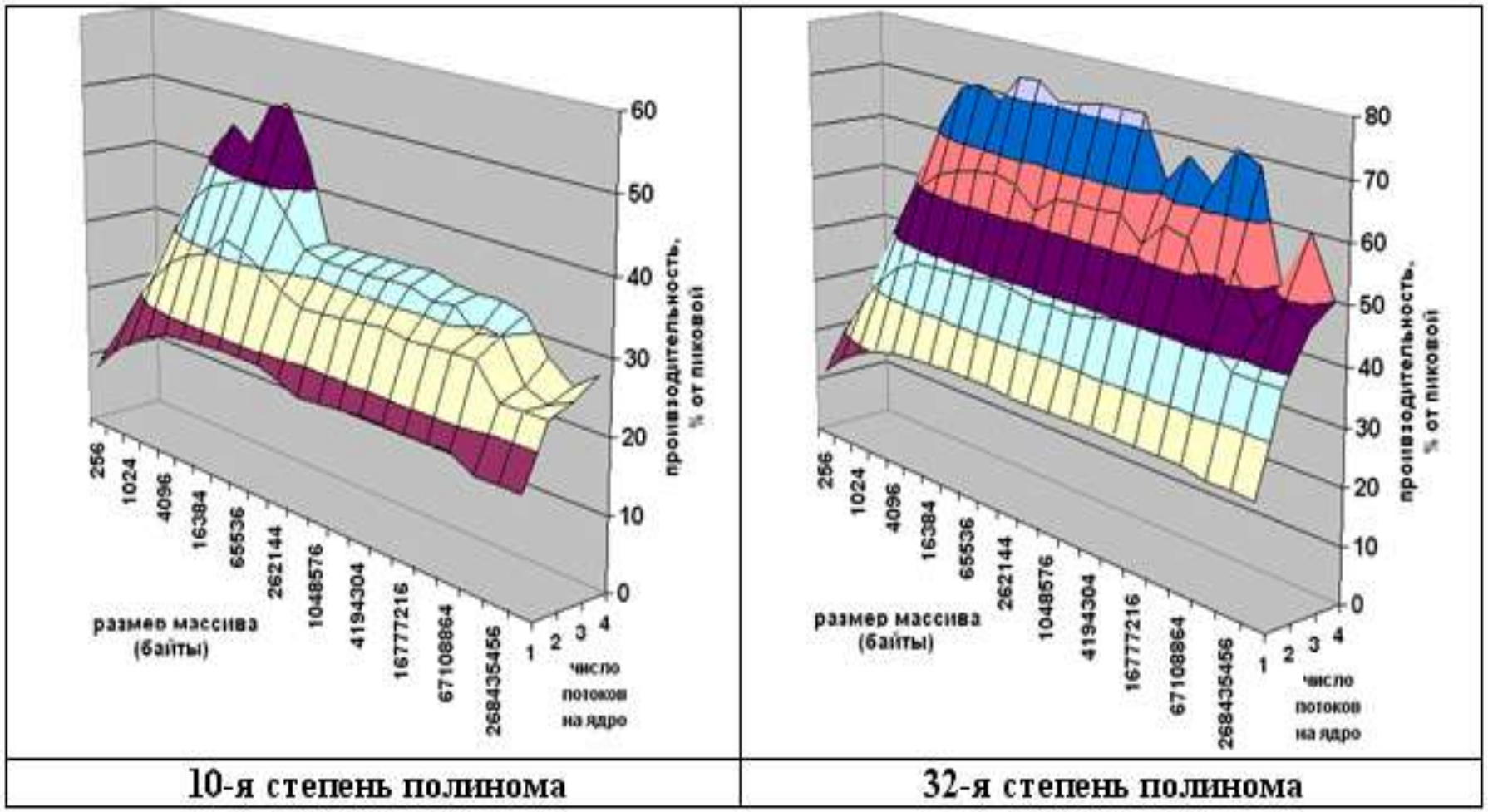


Xeon Phi



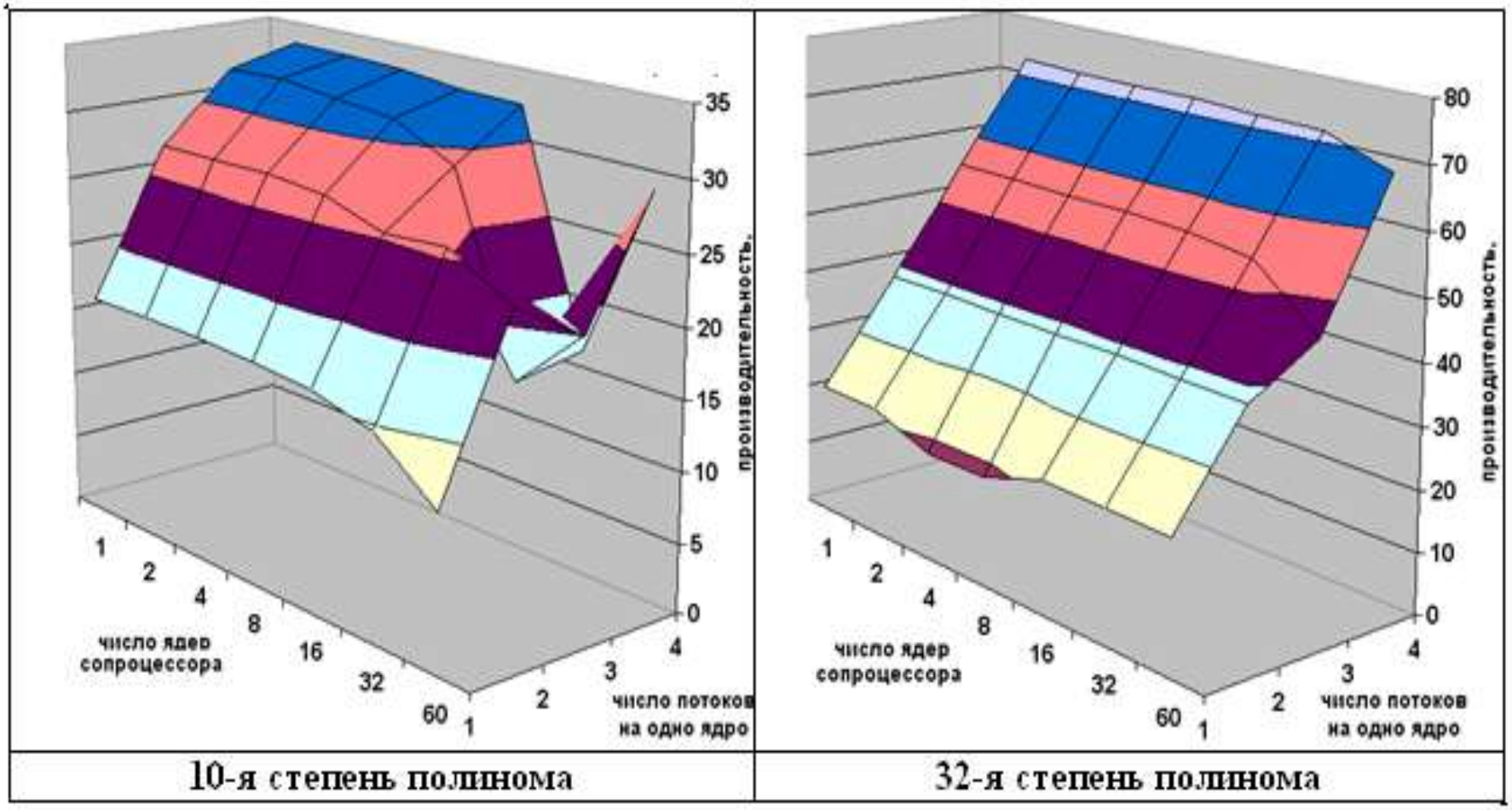
Sandy Bridge

Xeon Phi:одно ядро, большая доля вычислений, увеличение длины вектора и тредов на ядро



10-я степень полинома | 32-я степень полинома

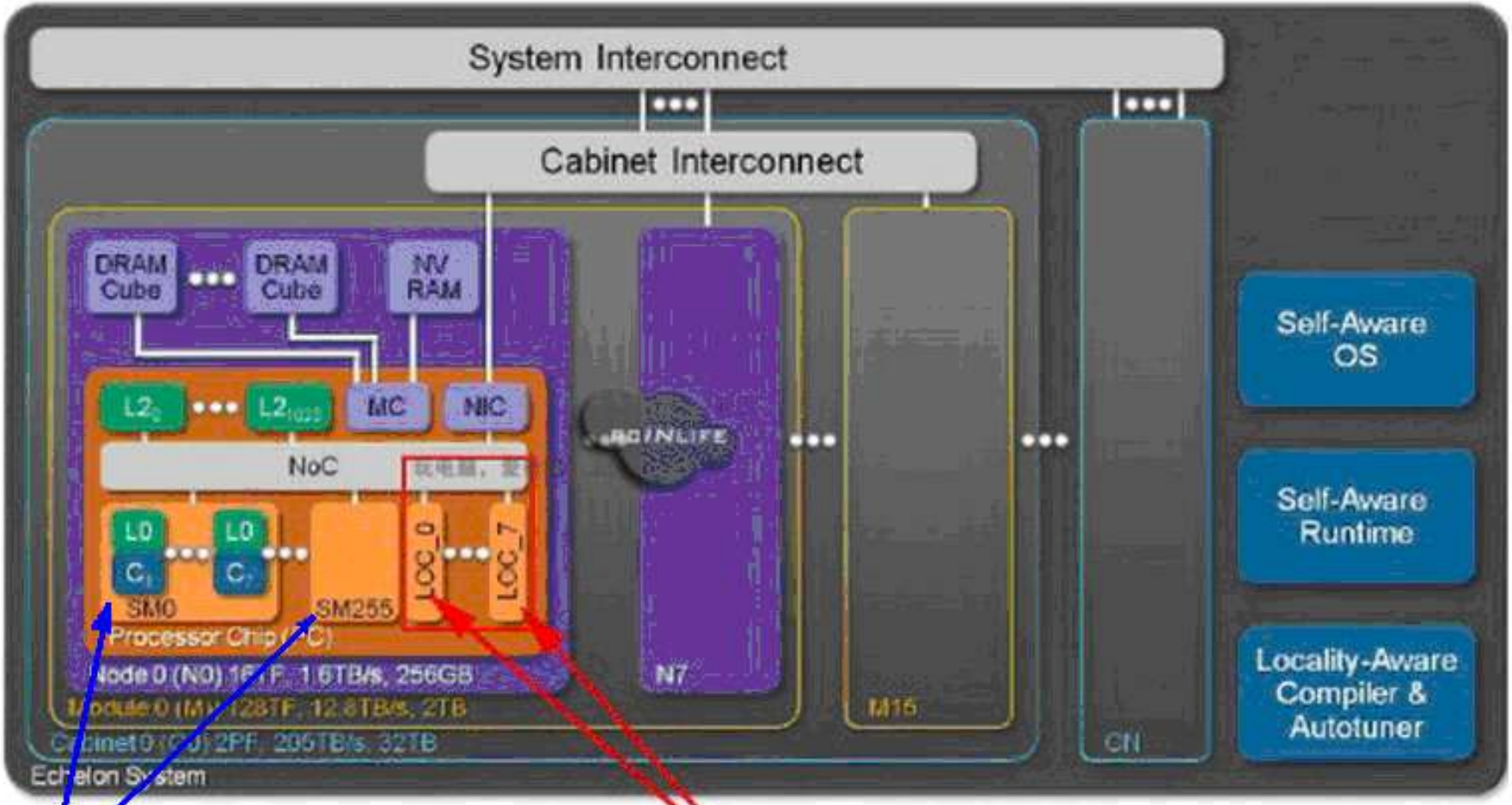
Xeon Phi:
вектор 1Мбайт, увеличение ядер и тредов



10-я степень полинома

32-я степень полинома

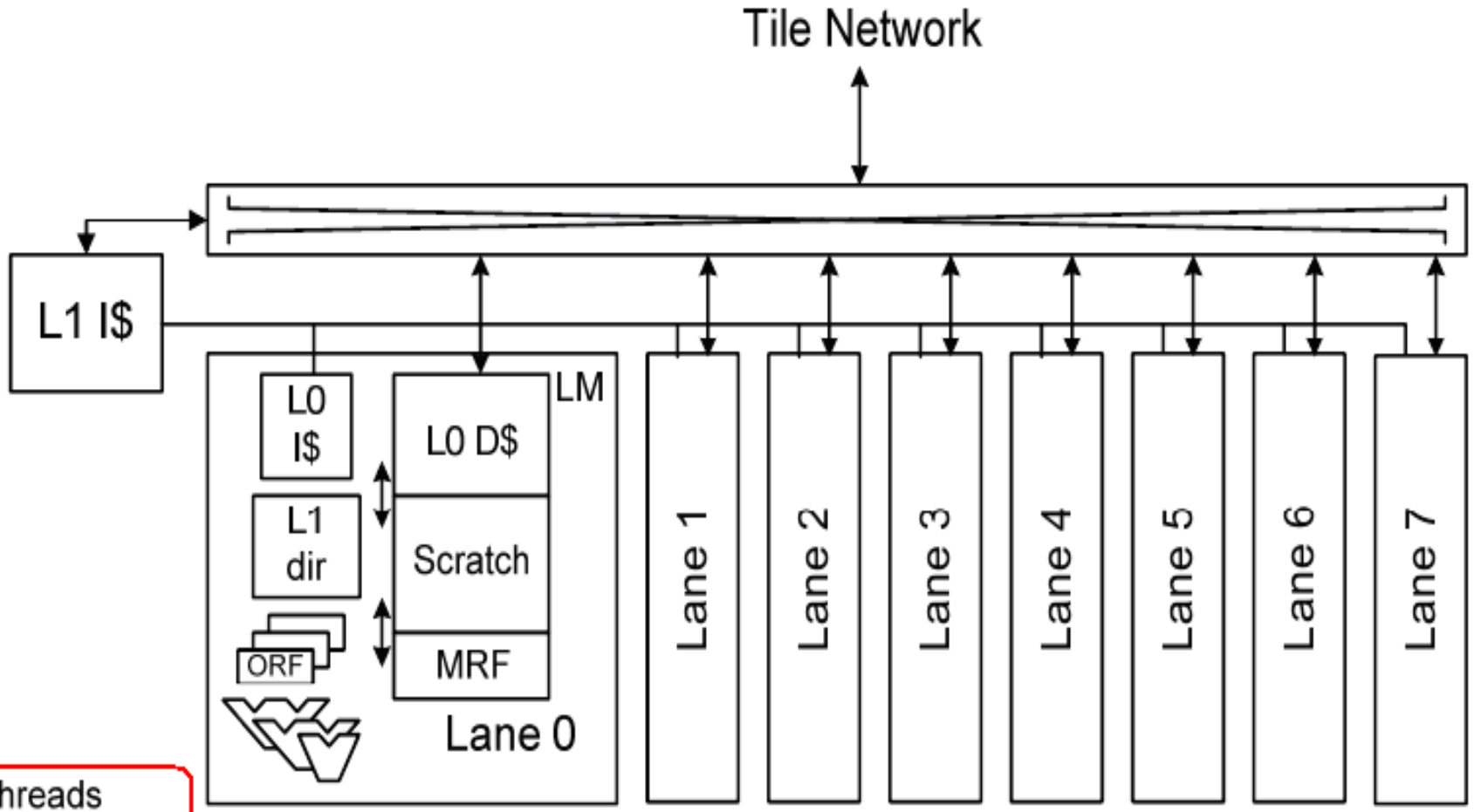
Проект Echelon (NVIDIA, Cray, ORNL, Lockhead Martin, 8 университетов)



Массово-мультитредовые ядра

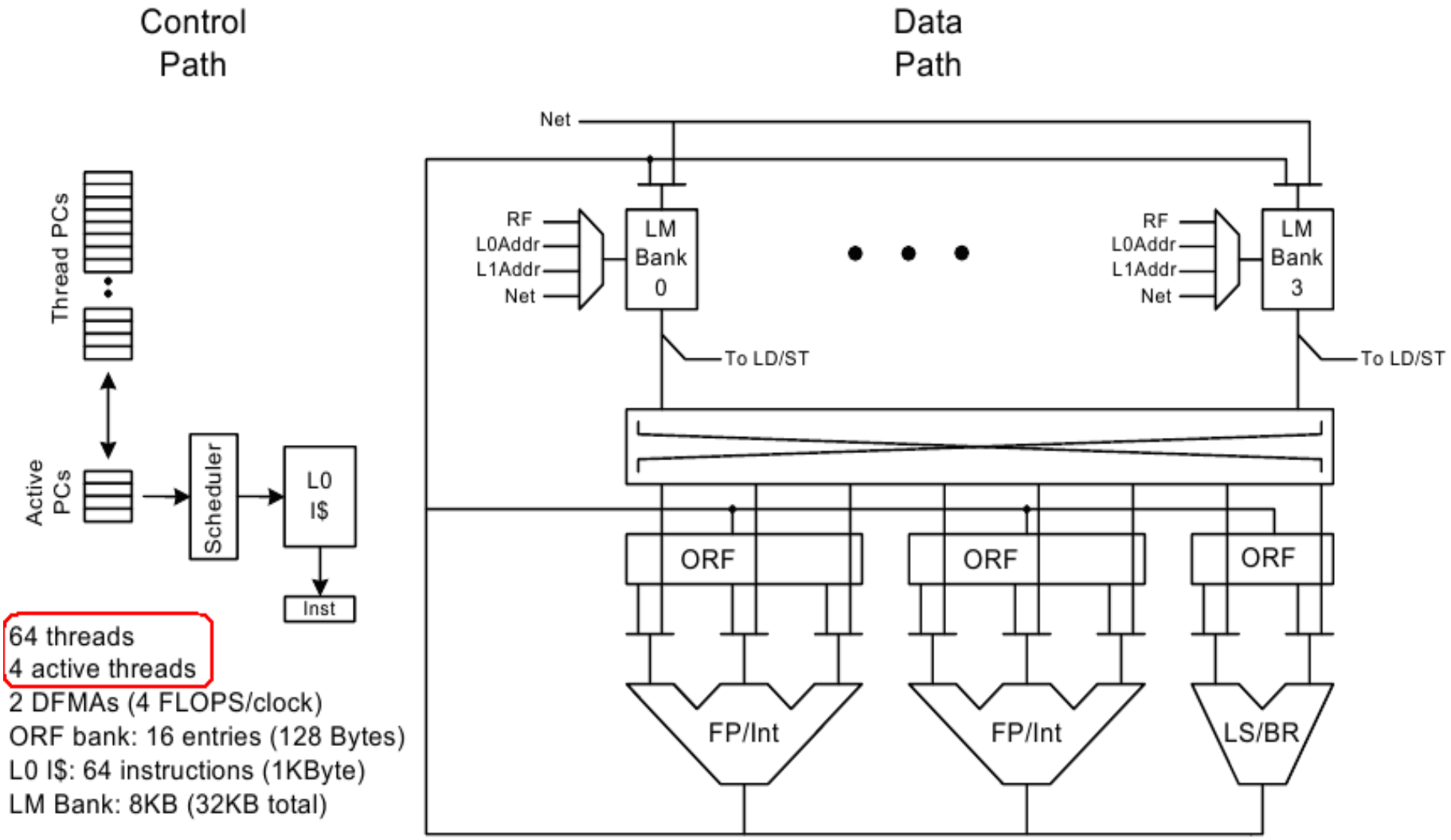
Суперскалярные ядра

Структура SM-ядра



512 threads
32 active threads
16 DFMAs (32 FLOPs/clock)
L1 I$: 2K instructions (32KB)
RF/Scratch/D$: 256KB
L0 caches in other lanes form L1 cache

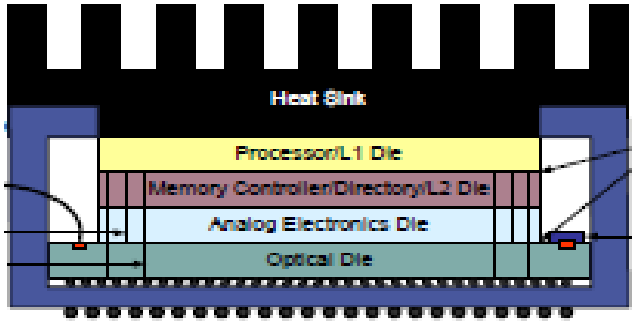
Полоса обработки (Lane) SM-ядра



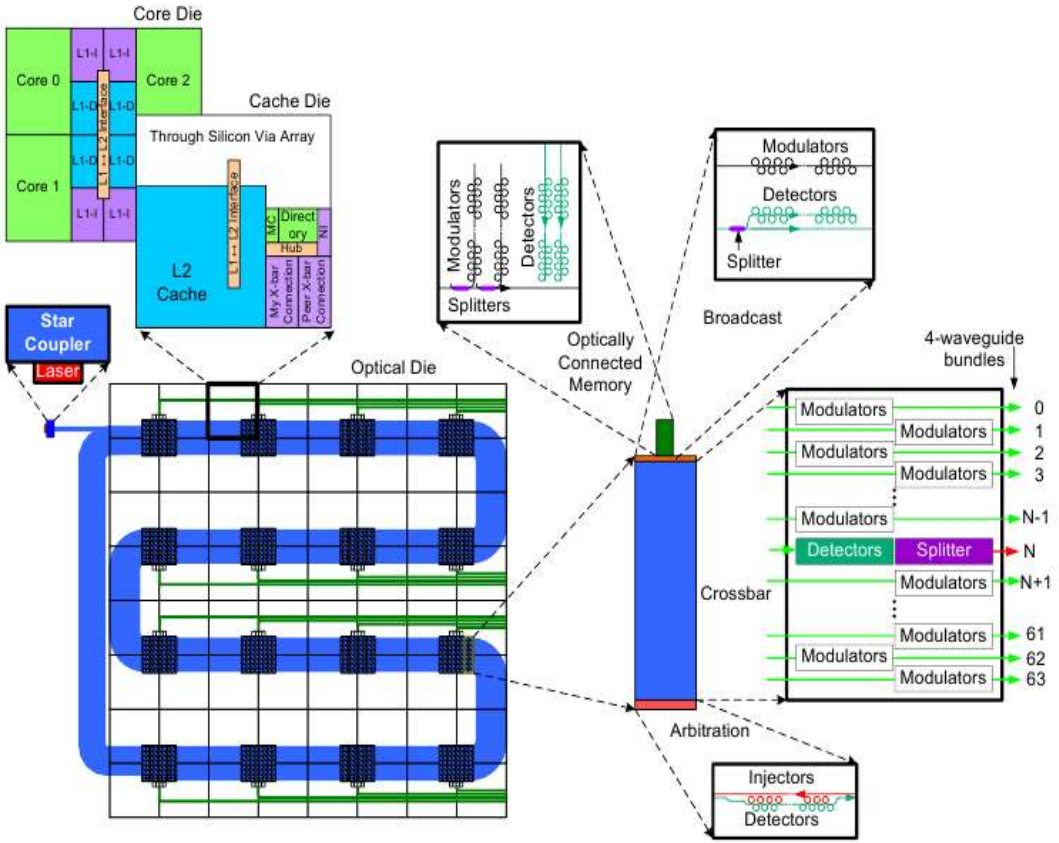
Проект Corona

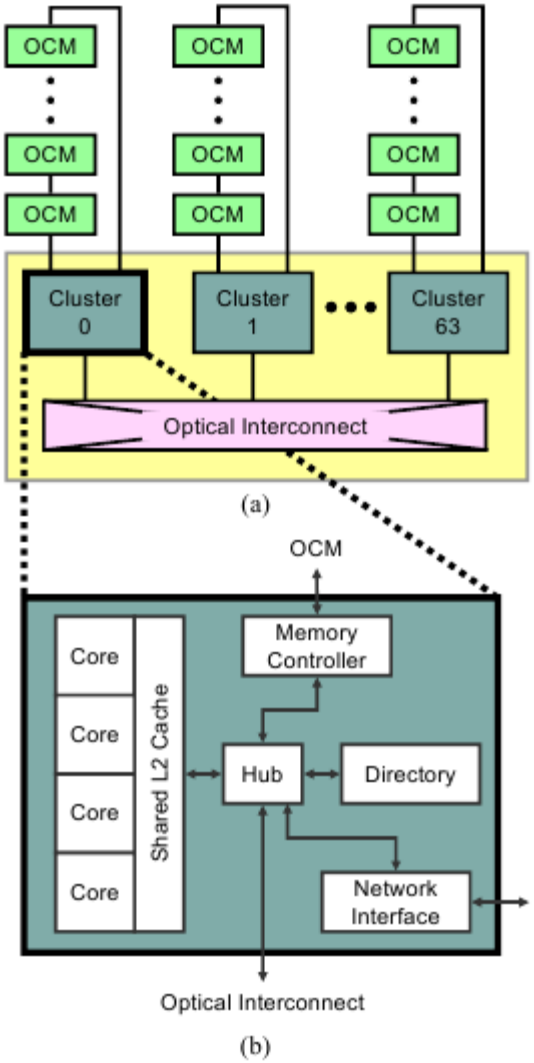
(Hewlett-Packard, University of Visconsin, University of UTAH)

3D-модуль процессора

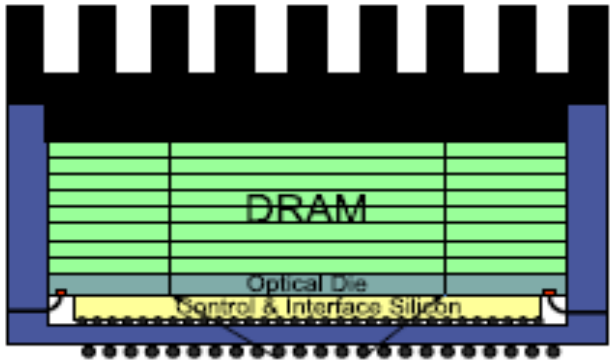


Corona 3D Package

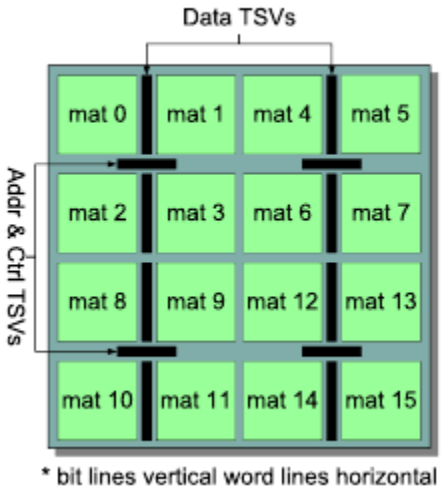




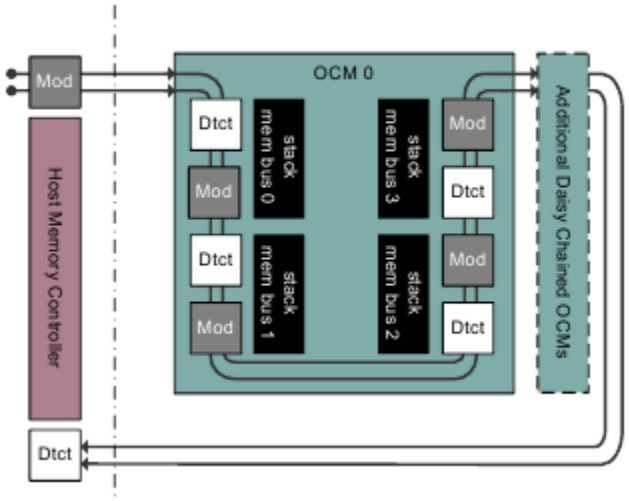
3D-модуль памяти

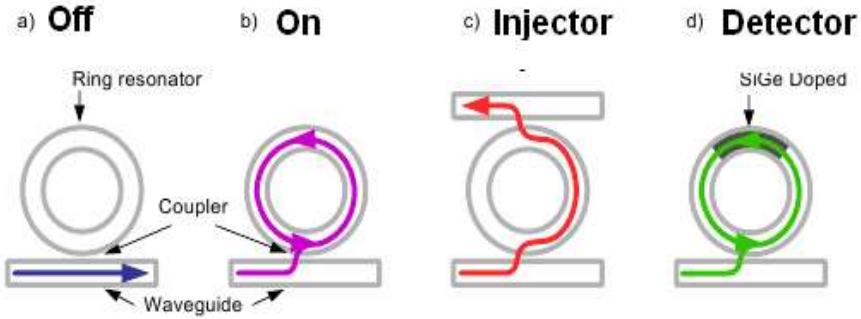


Optically Connected Memory



* bit lines vertical word lines horizontal





<u>10 TB/s – цена реализации</u>

ECM – 160W, 2 mW/Gb/s

OCM – 6W, 0.078 mW/Gb/s

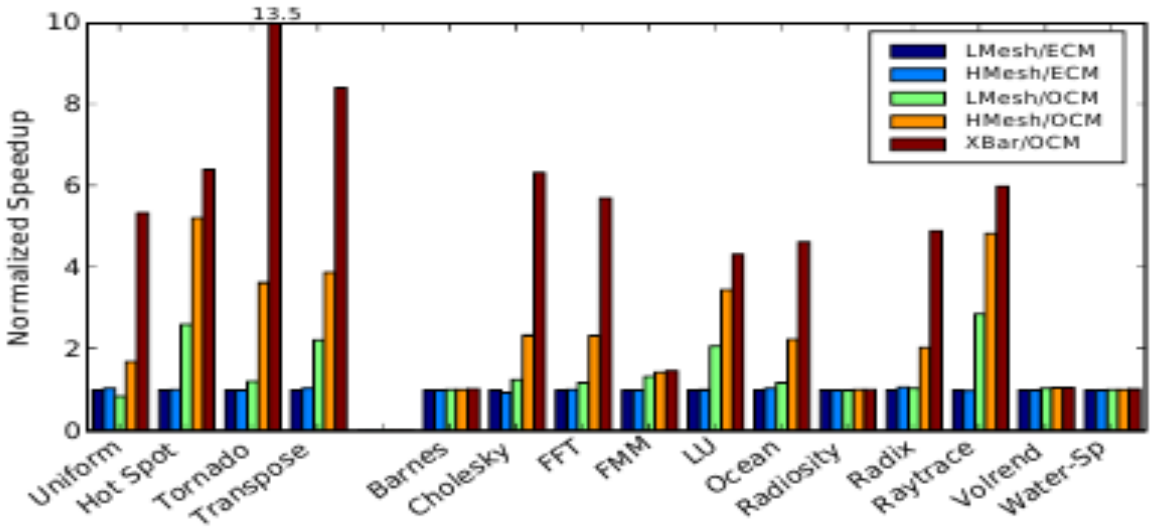
Результаты моделирования микропроцессора Corona – эффект новой памяти и сетей

Память

Resource	OCM	ECM
Memory controllers	64	64
External connectivity	256 fibers	1536 pins
Channel width	128 b half duplex	12 b full duplex
Channel data rate	10 Gb/s	10 Gb/s
Memory bandwidth	10.24 TB/s	0.96 TB/s
Memory latency	20 ns	20 ns

Внутрикристальная

- XBar – An optical crossbar (as described in Section 3.2), with bisection bandwidth of 20.48 TB/s, and maximum signal propagation time of 8 clocks.
- HMesh – An electrical 2D mesh with bisection bandwidth 1.28 TB/s and per hop signal latency (including forwarding and signal propagation time) of 5 clocks.
- LMesh – An electrical 2D mesh with bisection bandwidth 0.64 TB/s and per hop signal latency (including forwarding and signal propagation time) of 5 clocks.



Uniform

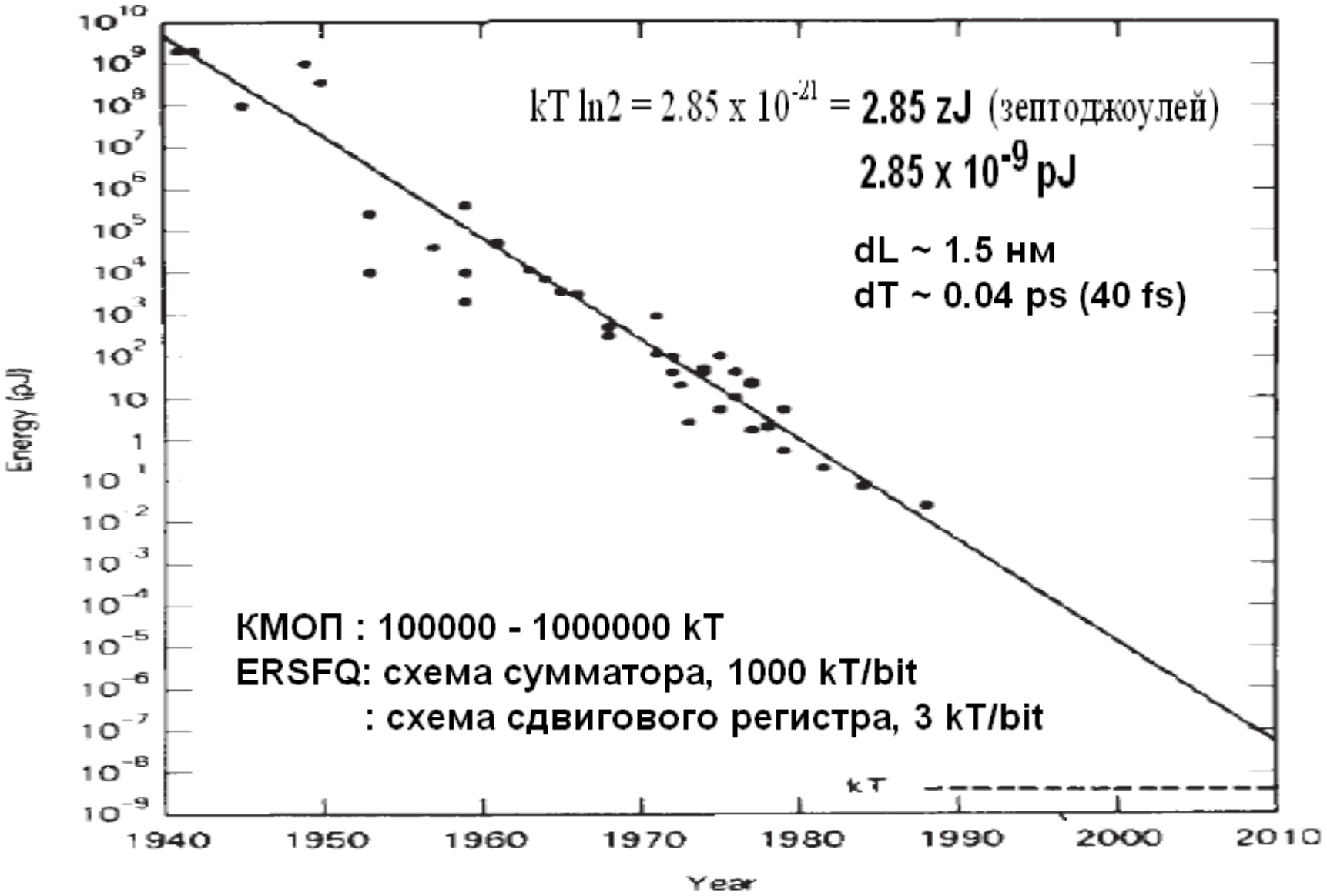
Transpose

FFT

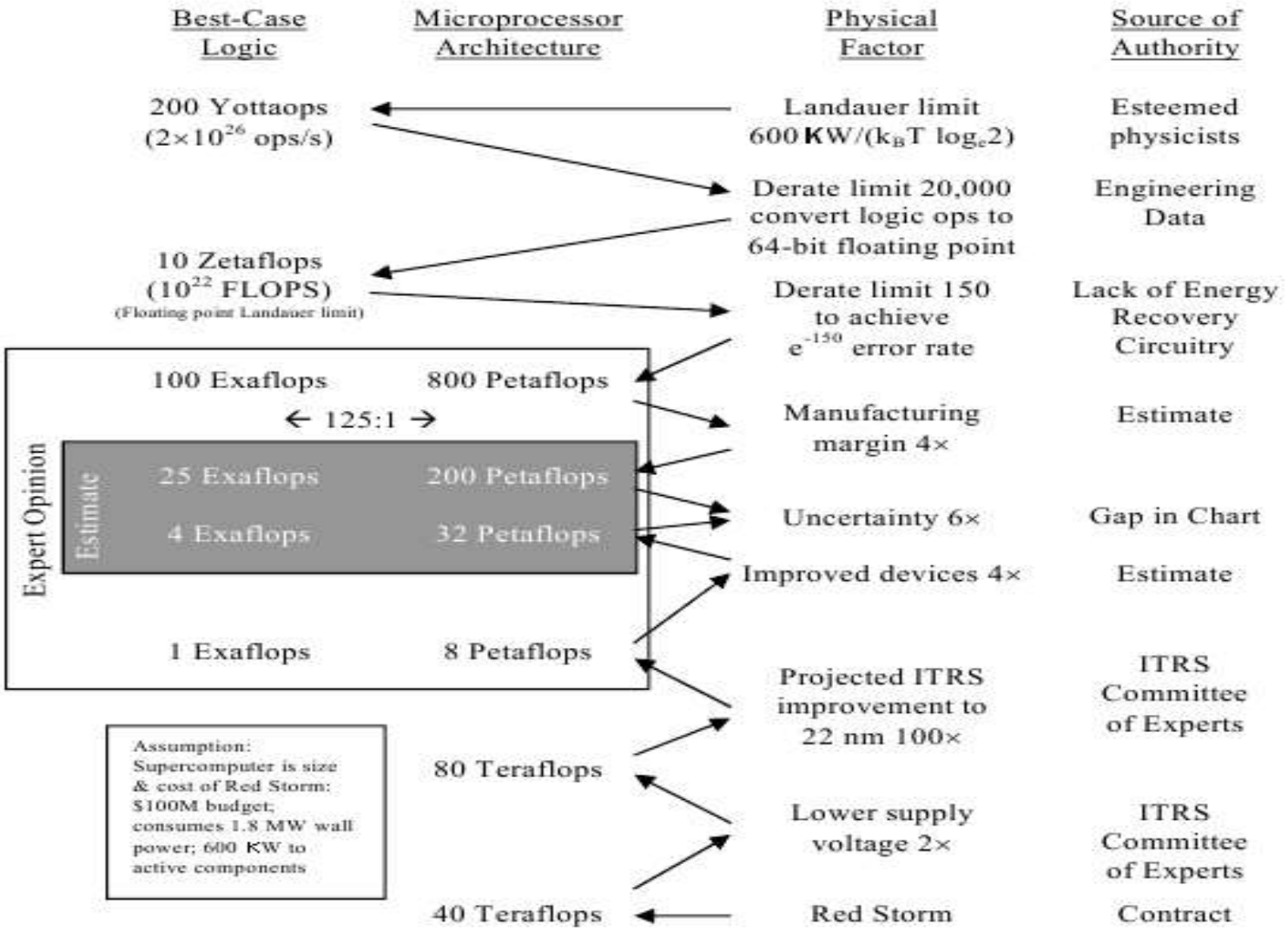
LU

Физические ограничения и пост-Муровская эра

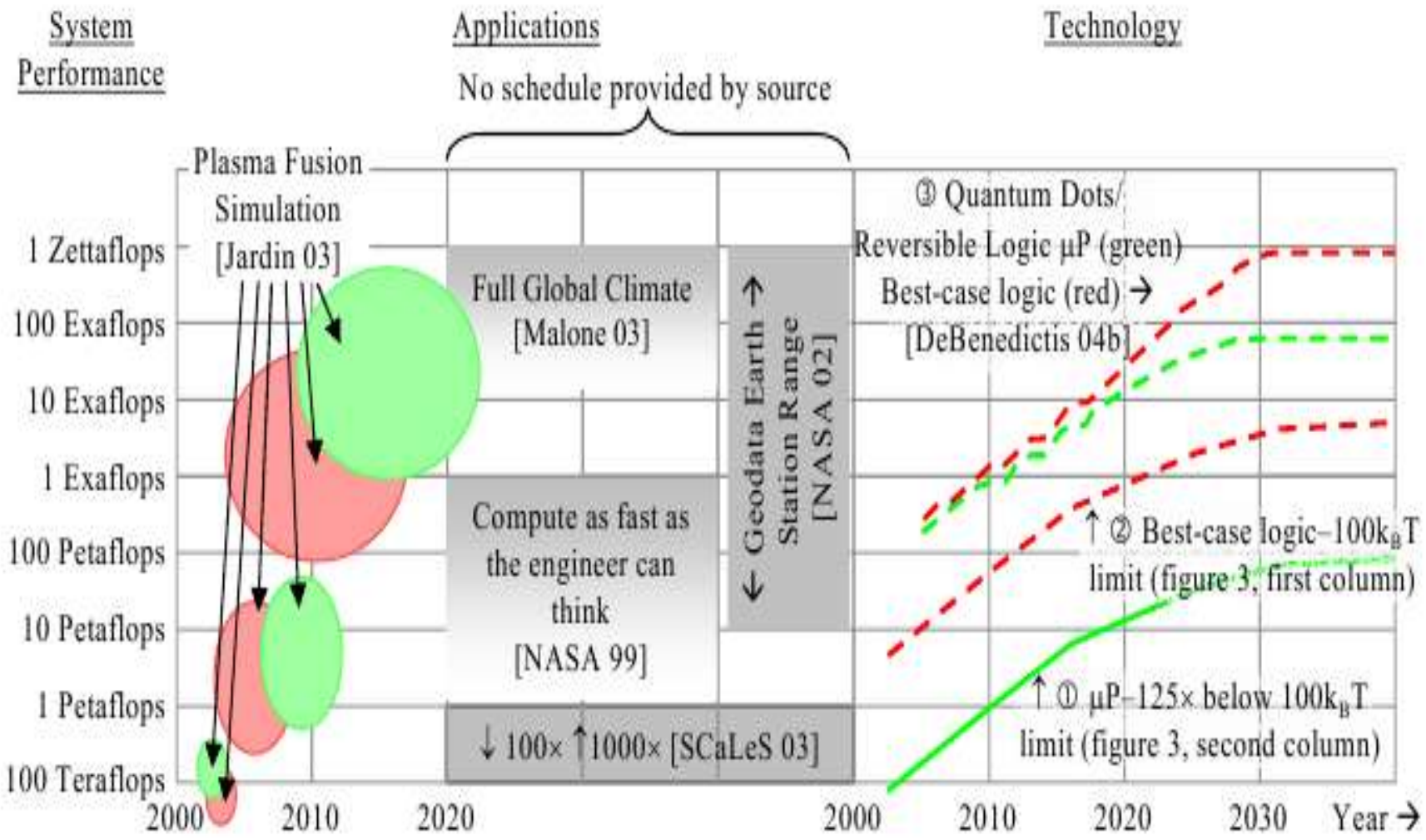
Ограничение Лэндауэра. Динамика снижения затрат на обработку одного бита



Физический предел производительности нереверсивных суперкомпьютеров – “точка Стерлинга”(для мощности 600 KW)



Прогноз роста потребностей производительности и возможностей создаваемых суперкомпьютеров



Выводы

1. Началось внедрение результатов зарубежных НИР прошлого десятилетия по петафлопсной тематике

2. Усилился объем и глубина исследований по экзамасштабной тематике и новой элементной базе пост-Муровской эры. Начало выполнения крупной экзафлопсной программы в DoE ожидается не ранее 2014 года

3. Важнейшие направления системных исследований и разработок: иерархические сети, гибридные массово-мультитредовые многоядерные микропроцессоры, гетерогенность (включая специализированные фрагменты, в том числе фрагменты аналогового типа)

4. Важнейшие направления по элементно - конструкторской базе: нанофотоника, сверхпроводниковая электроника, квантовая электроника (спиновая, клеточные автоматы, реверсивная логика), TSV-соединения, оптоволоконные соединения через матрицы линз и лазеров.

5. Новая элементно-конструкторская база позволит преодолеть зетта- и йотта- уровень производительности, это намечено в первую очередь в заказных специализированных зарубежных суперкомпьютерах

6. В России есть Концепция создания экзафлопсных систем. Ведется множество отдельных проектов, в том числе и в РАН (проект МГВС – ИПМ им.М.В.Келдыша РАН, ФГУП НИИ”Квант”, Физфак МГУ) .

7. Проект МГВС нацелен на поддержку ряда проектов по имитационному моделированию архитектур и нового программного обеспечения, включает также использование новых аналитических моделей.

Также рекомендуется для ознакомления:

The Technology Lane on the Road to a Zettaflops. SC'06, April 24, 2006, 13 pp.

Вопросы ?

Горбунов Виктор Станиславович (gorbunov@rdi-kvant.ru)

Елизаров Георгий Сергеевич (elizarov@rdi-kvant.ru)

Эйсымонт Леонид Константинович (eisymont@rdi-kvant.ru)