

# Всё меняется. И НРС тоже...

**Алексей Перевозчиков**  
Server Solutions Product Manager

[82189117@ru.ibm.com](mailto:82189117@ru.ibm.com)



## 2010 Patent Leadership (18 лет лидерства)

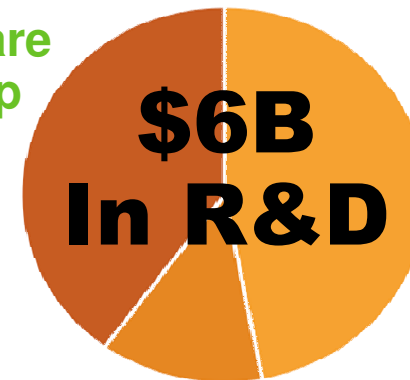
|                | Totals |
|----------------|--------|
| IBM            | 5,896  |
| Samsung        | 4,551  |
| Microsoft      | 3,094  |
| Canon          | 2,552  |
| Panasonic      | 2,482  |
| Toshiba        | 2,246  |
| Sony           | 2,150  |
| Intel          | 1,653  |
| LG Electronics | 1,490  |
| HP             | 1,480  |

\* Source: IFI Patent Intelligence



Software  
Group

Systems &  
Technology  
Group



Research

IBM Austin the Home of Power Systems  
development: 950 Patents  
#1 IBM location for 8<sup>th</sup> year

# 2012 Patent Leadership (20 лет лидерства)

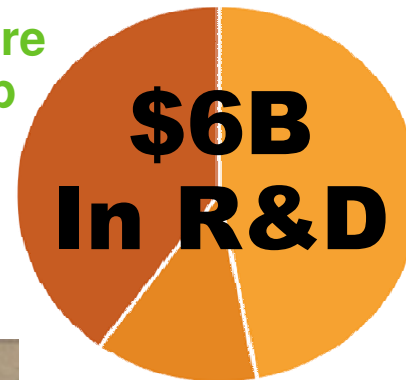
## Totals

|                   |       |
|-------------------|-------|
| IBM               | 6,478 |
| Samsung           | 5,081 |
| Canon             | 3,174 |
| Sony              | 3,032 |
| Panasonic         | 2,769 |
| Microsoft         | 2,613 |
| Toshiba           | 2,447 |
| Hon Hai Precision | 2,031 |
| General Electric  | 1,652 |
| LG Electronics    | 1,624 |



Software  
Group

Systems &  
Technology  
Group



Research

Streaming  
Analytics



Source: IFI Patent Intelligence

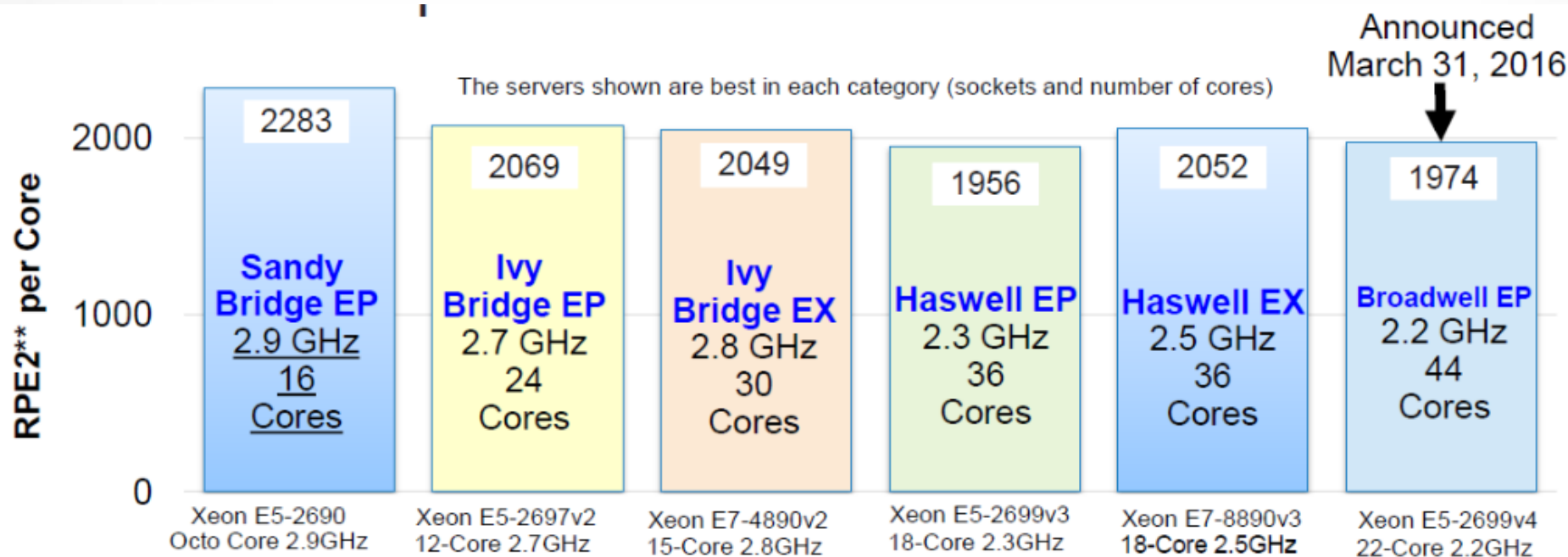
HP #15 1,394

Intel #18 1,290

Oracle is not in the top 50



## RPE2



\*\*Gartner RPE2 Details:  
<http://www.gartner.com/technology/research/RPE2-methodology-details.jsp>

RPE2\*\* numbers are derived from the following six benchmark inputs:

SAP SD Two-Tier, TPC-C, TPC-H, SPECjbb2006 and two SPEC CPU2006 components

The data on this chart is derived from RPE2 from Gartner, Inc's Competitive Profile tool. © 2016 Gartner, Inc. and/or its affiliates. All rights reserved.





# Intel отказалась от своей знаменитой стратегии «ТИК-ТАК»

Техника

[Основная версия](#)

24.03.2016, ЧТ, 11:57, Мск , Текст: Сергей Попсулин

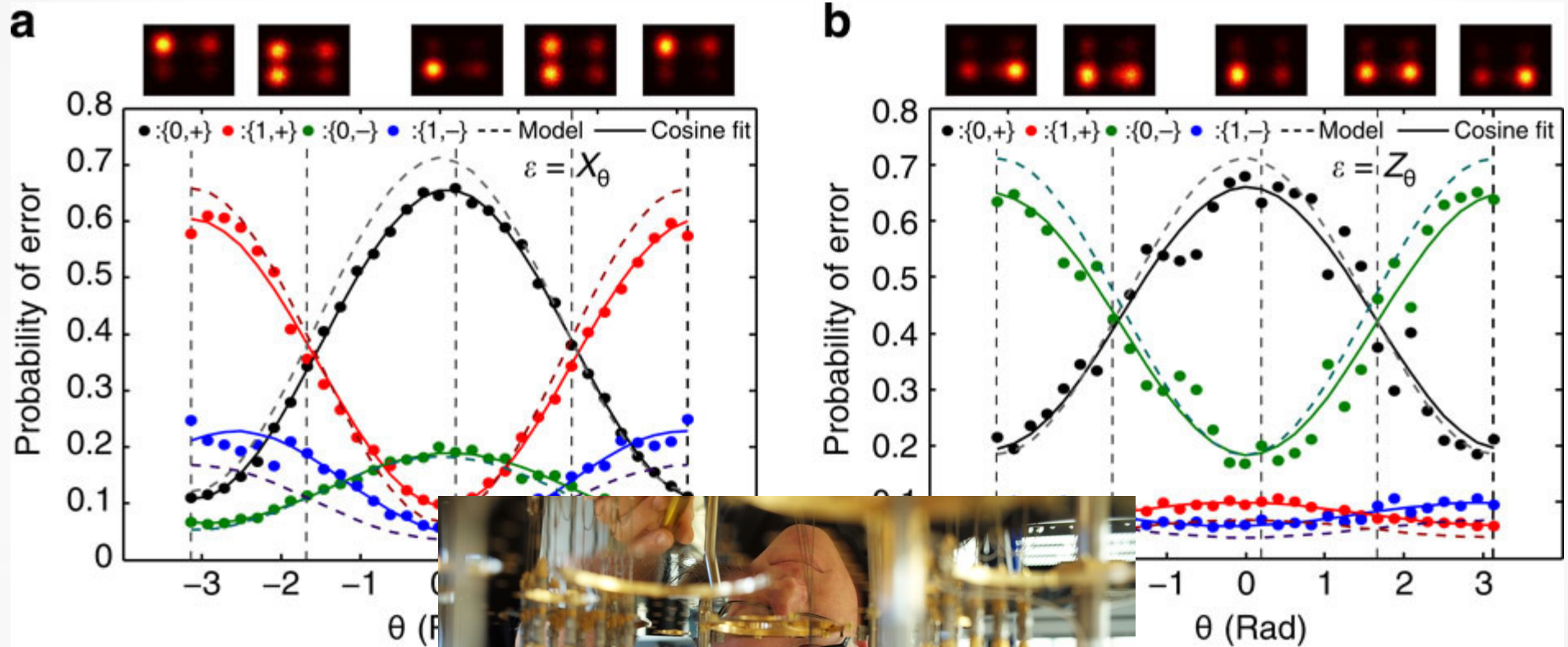
**Intel решила сбавить темп смены технологического процесса и увеличить время выпуска процессоров с 14-нм транзисторами до нестандартных трех лет.**

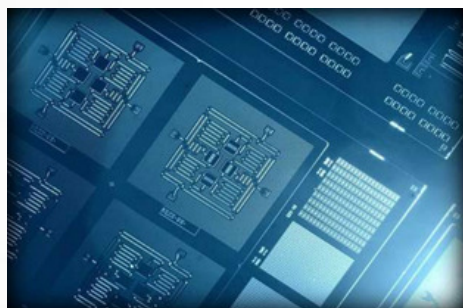
## Отказ от стратегии

Intel отказалась от стратегии «тик-так» и вместо нее перейдет на стратегию из трех этапов: техпроцесс-архитектура-оптимизация. Так будет на протяжении по крайней мере двух следующих технологических процессов, сообщает AnandTech со ссылкой на годовой отчет, в котором корпорация изложила суть перемен.

Изменение стратегии приведет к тому, что переход на новую технологическую норму теперь будет происходить не так часто, как прежде. Вместе с тем Intel обещает продолжить выпускать новые продукты ежегодно.

А тем временем... **IBM** сообщила об открытии способа контроля одновременно обоих типов квантовых ошибок и ...





## **IBM открывает эпоху публичных квантовых вычислений**

Автор: Андрей Колесов  
05.05.2016

IBM Research сообщило о запуске публичного бесплатного облачного сервиса IBM Quantum Experience, с помощью которого можно на практике познакомиться с возможностями

## Энергоэффективность

# GOOGLE'S ENERGY CONSUMPTION

16 February, 2013 · by Arnfinn Oines · in Energy, News.

Googles 13 data centres continuously draws 260 million watts. Below is an infographic made by CO2 Sense that highlights what it takes to power Google. It also includes statistics regarding the company's investment in renewable energy.

## IT now 10 percent of world's electricity consumption, report finds

**New analysis finds IT power suck has eclipsed aviation**

By Jack Clark, 16th August 2013



Follow

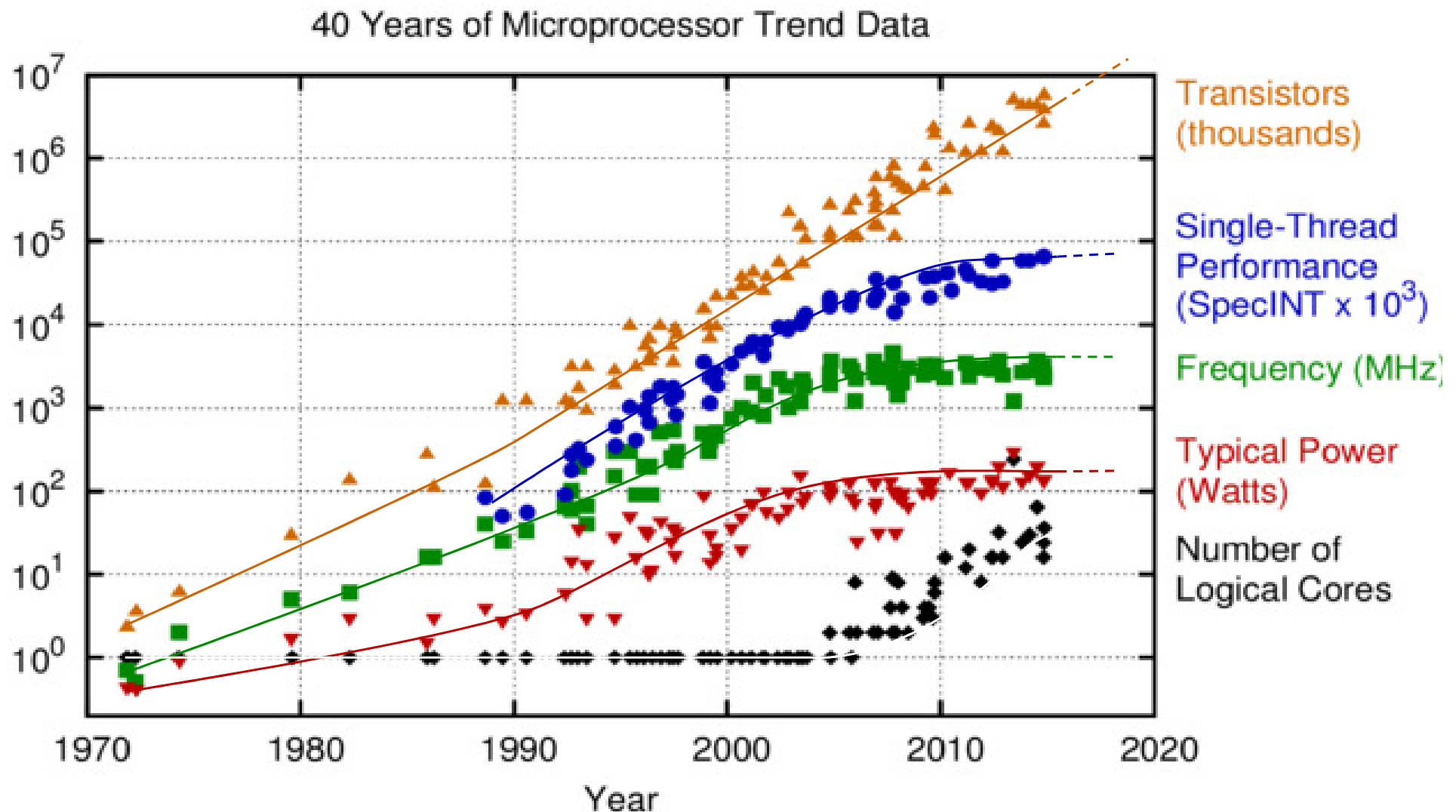
3,798 followers

# POWER8



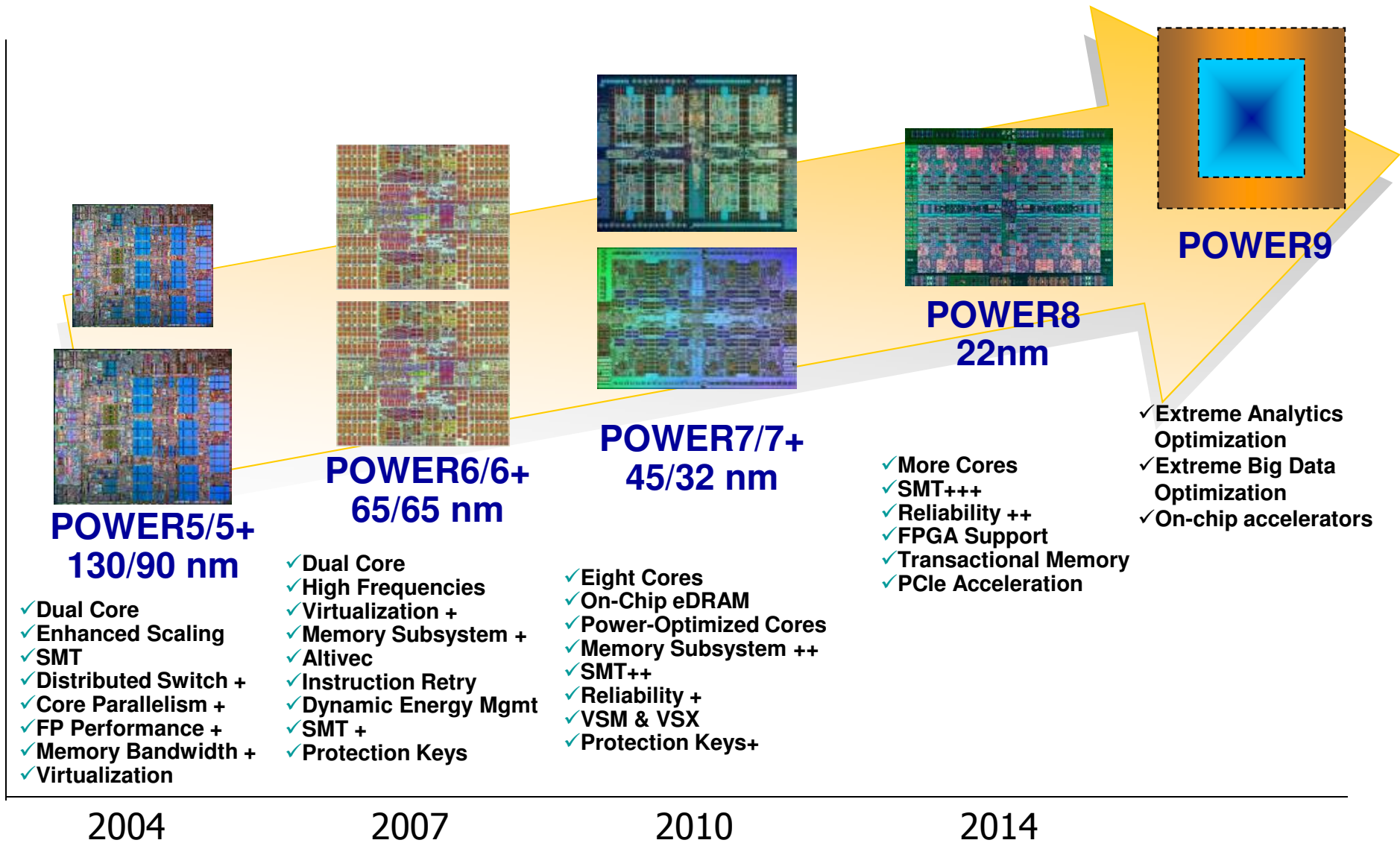


# 40 лет микропроцессорам: тенденции

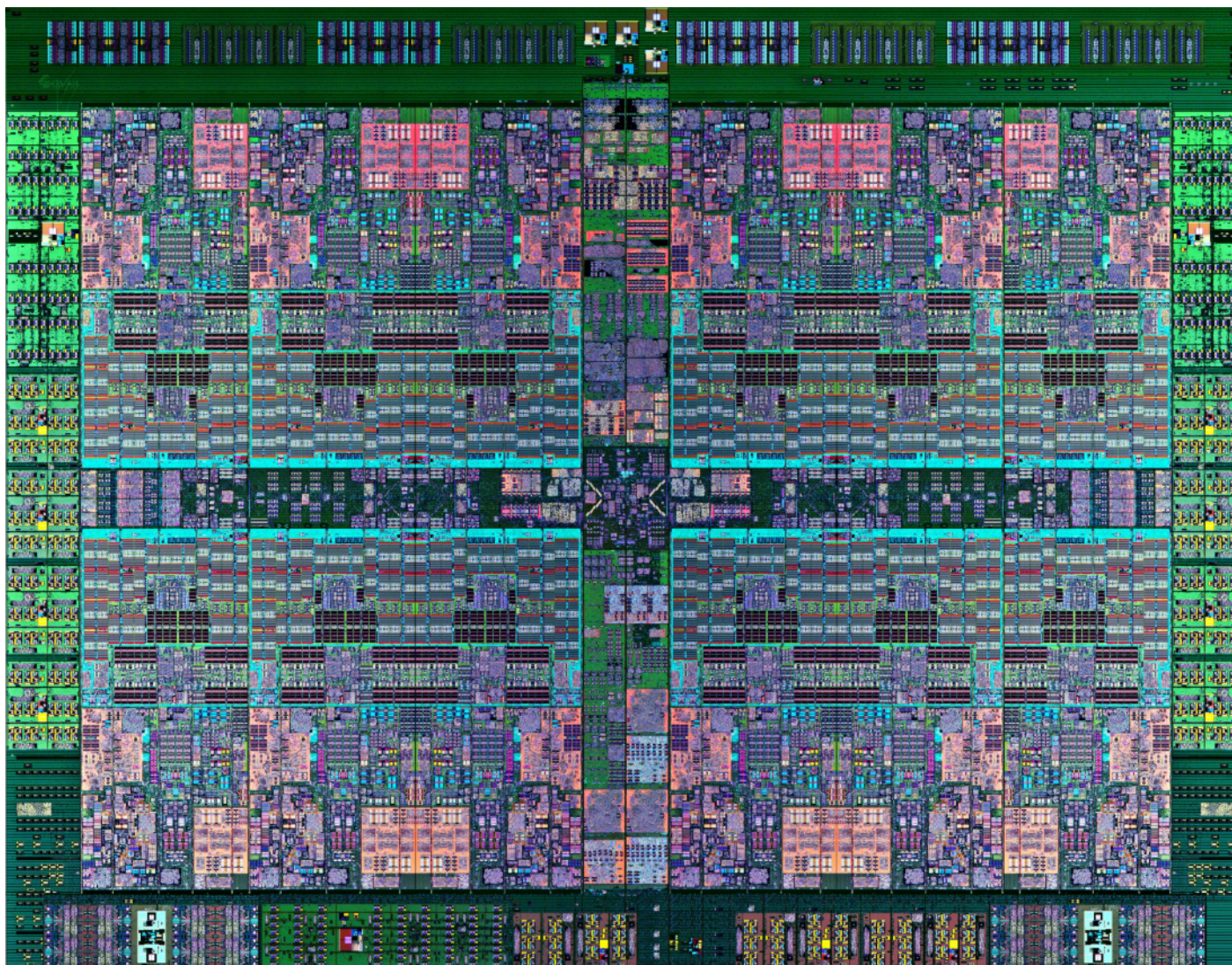


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten  
 New plot and data collected for 2010-2015 by K. Rupp











## Технология

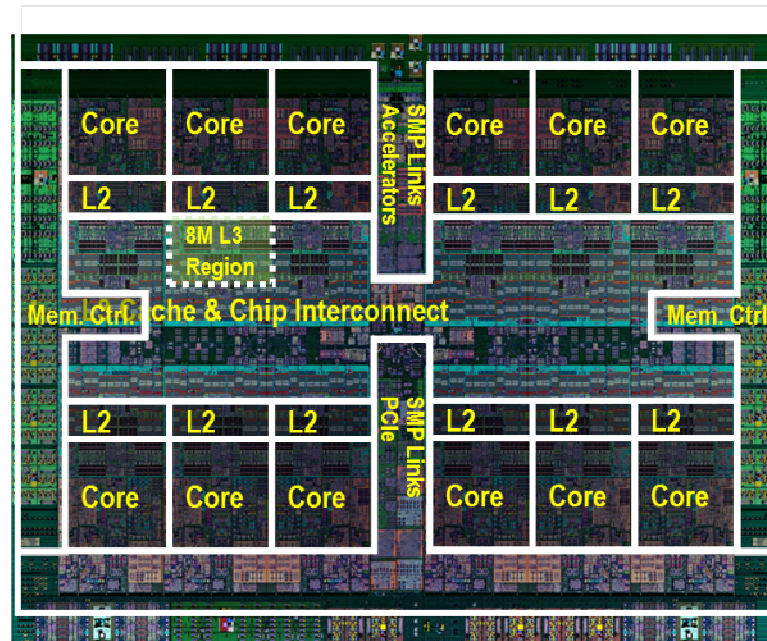
22nm SOI, eDRAM, 650mm<sup>2</sup>, 4.2B transistors

## Ядра

- **12 ядер (SMT8)**
- 8 dispatch, 10 issue, 16 exec pipe
- **2X internal data flows/queues**
- Enhanced prefetching
- **64K кэш данных, 32K кэш инструкций**

## Акселераторы

- **Криптография**
- **Расширение памяти**
- **Транзакционная память**
- **Поддержка VMM**
- **Перемещение данных / VM**



## Energy Management

- On-chip Power Management Micro-controller
- Integrated Per-core VRM
- **Critical Path Monitors**

## Увеличенные кэши

- 512 KB SRAM L2 / core
- **96 MB eDRAM shared L3**
- **Up to 128 MB eDRAM L4 (off-chip)**

## Память

- Up to 230 GB/s sustained bandwidth

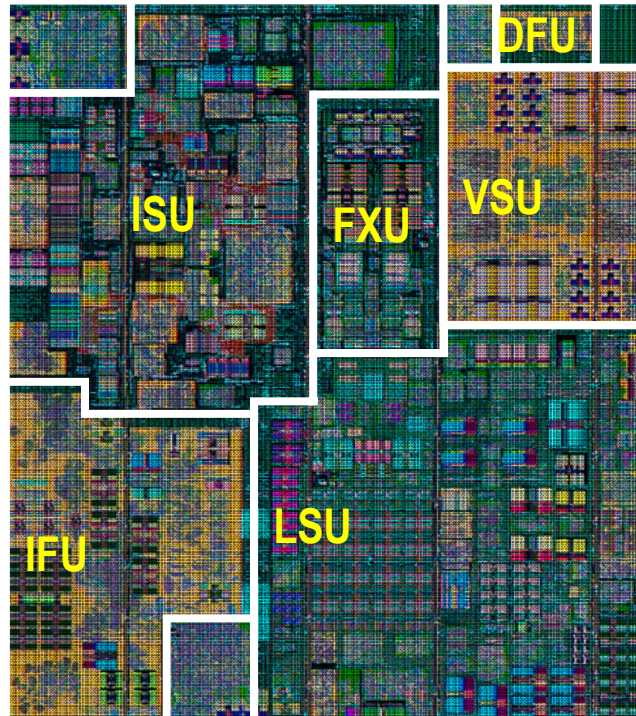
## Шинные интерфейсы

- Durable open memory attach interface
- **Интегрированный PCIe G3**
- SMP Interconnect
- **CAPI (Coherent Accelerator Processor Interface)**



- **SMT4 → SMT8**

- 8 dispatch
- 10 issue
- **16 execution pipes:**
  - 2 FXU, 2 LSU, **2 LU**, 4 FPU,
  - 2 VMX, 1 Crypto, 1 DFU,
  - 1 CR, 1 BR
- Larger Issue queues  
(4 x 16-entry)
- Larger global completion,  
**Load/Store reorder**
- Improved branch prediction
- Improved unaligned storage access



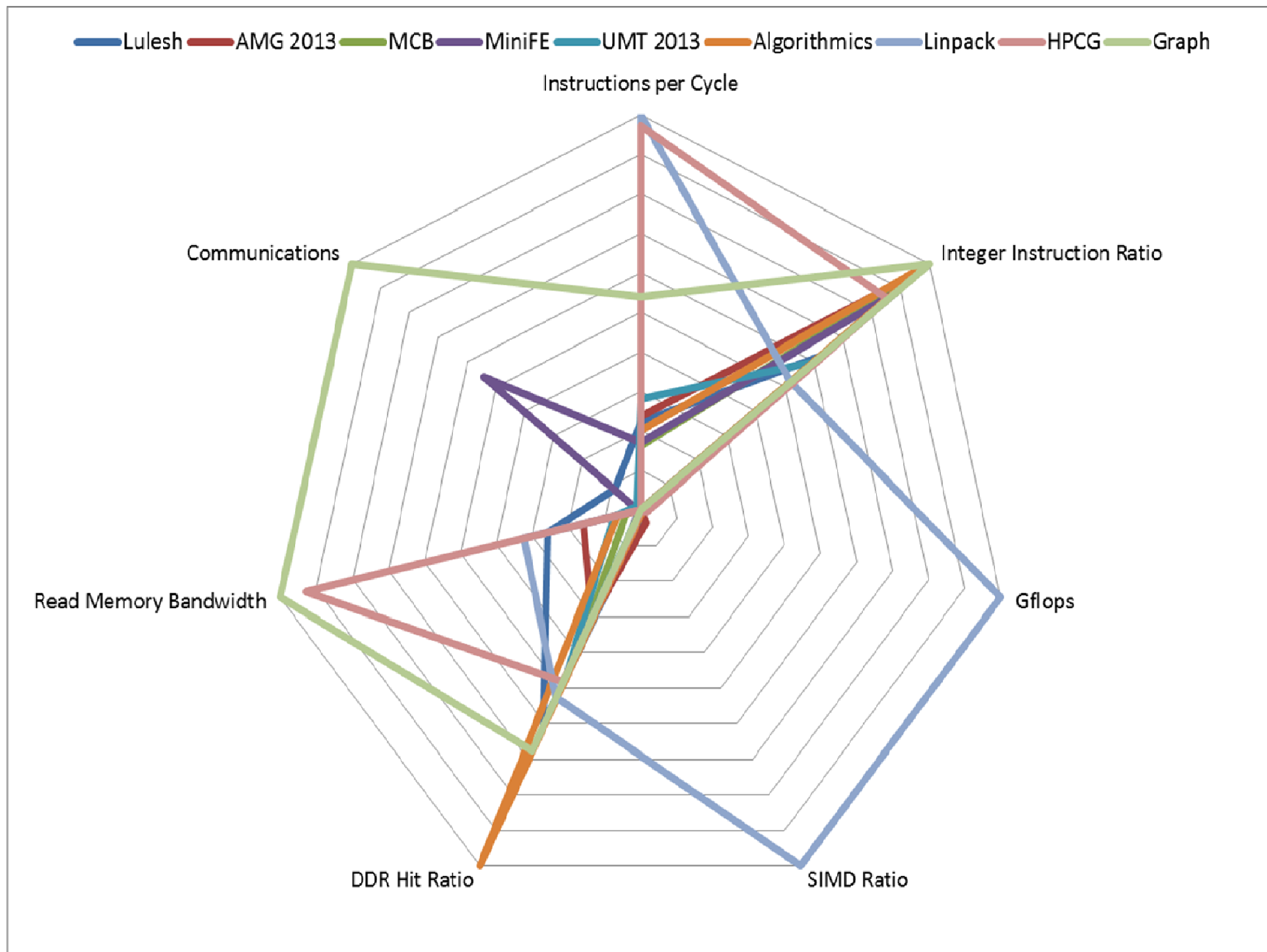
- 2x L1 data cache (64 KB)
- 2x outstanding data cache misses
- 4x translation Cache

### Wider Load/Store

- 32B → 64B L2 to L1 data bus
- 2x data cache to execution dataflow

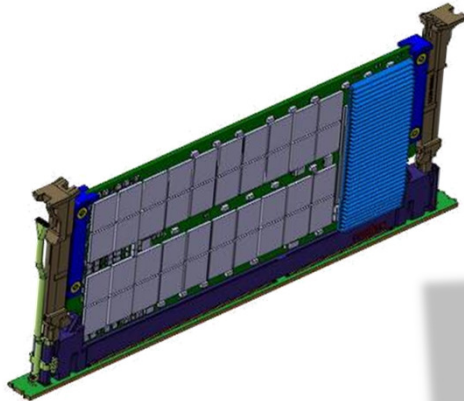
### Enhanced Prefetch

- **Instruction speculation awareness**
- Data prefetch depth awareness
- Adaptive bandwidth awareness
- Topology awareness





# Memory Buffer Chip ...with 16MB Cache...



“L4 cache”

## Модули памяти наполняются интеллектом

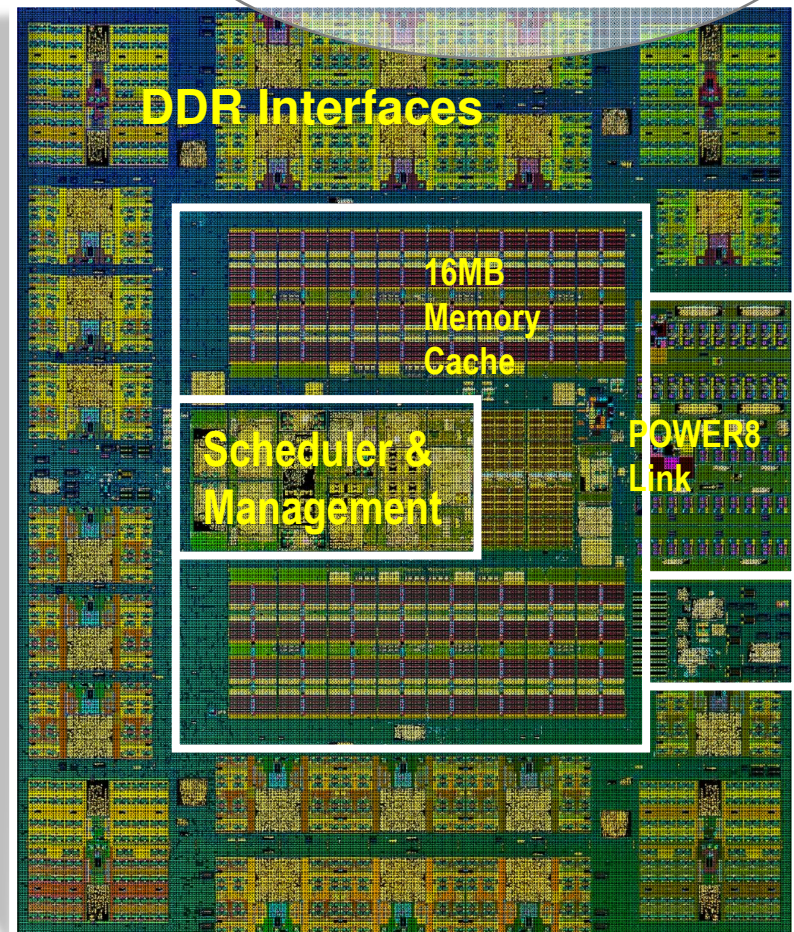
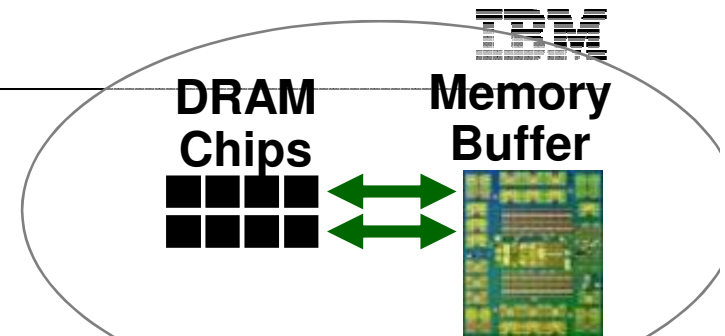
- Умная система кэширования
- Оптимизация энергии
- Надежность

## Оптимизированный интерфейс

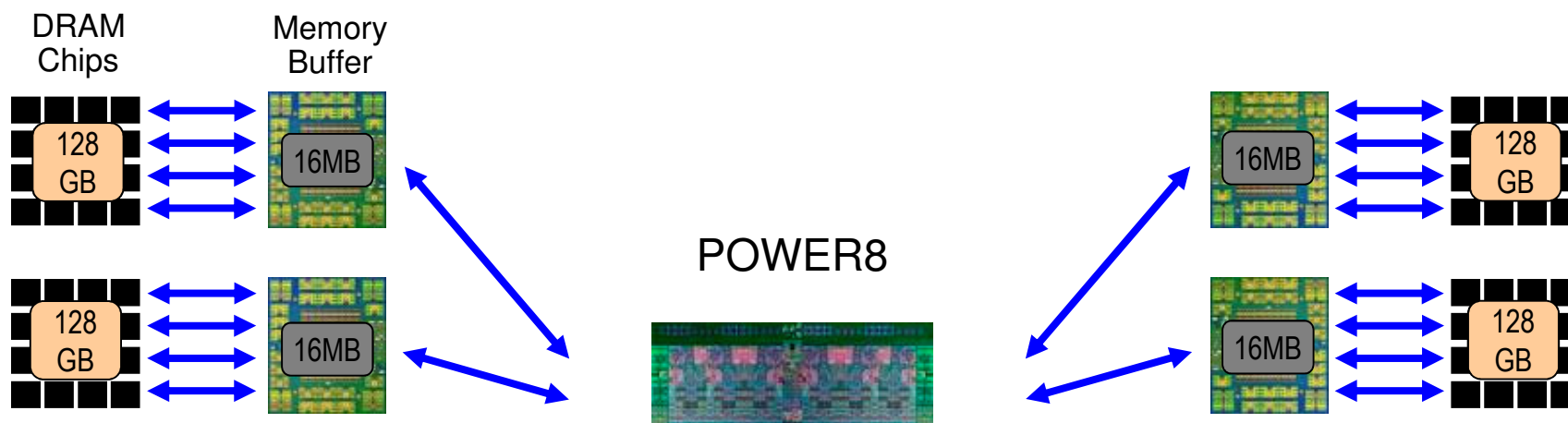
- 9.6 GB/s high speed interface
- Интеллектуальная надежность
- Изоляция сбоев на лету

## Уникальная производительность

- Уменьшенная латентность fastpath
- Cache → latency/bandwidth, partial updates
- Логика предсказания
- 22nm SOI for optimal performance / energy
- 15 metal levels (latency, bandwidth)







- У Intel нет L4 и они показывают цифры “to the DIMM”
  - Наши 230 ГБ/с вполне достижимы в реальных условиях
  - Цифры “to-DIMM” теоретические, реально достижимые намного ниже (из-за используемых протоколов DIMM, это справедливо для всех производителей)
- ➔ 8 скоростных каналов, каждый до 9.6 Гб/с  
до **230 ГБ/с в устойчивом режиме (sustained)**
  - ➔ До 32 портов DDR выдающих в пике **410 ГБ/с (на уровне DRAM)**
  - ➔ **До 1 ТБ памяти** на сокет (для старших версий – до 2 ТБ на сокет)



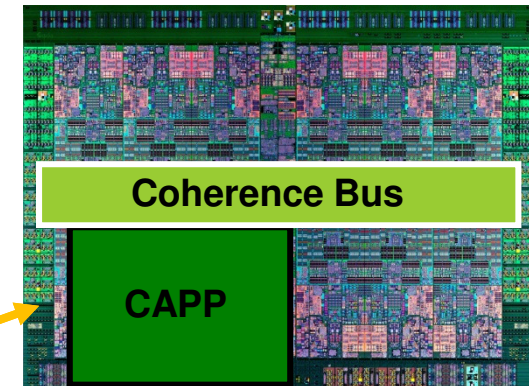
## Virtual Addressing

- Ускоритель работает напрямую с разделяемой памятью
- Обмен данными с кэшем процессора.
- Исключает накладные расходы ОС и драйверов.

## Hardware Managed Cache Coherence

- Стандартный механизм блокировок.

POWER8

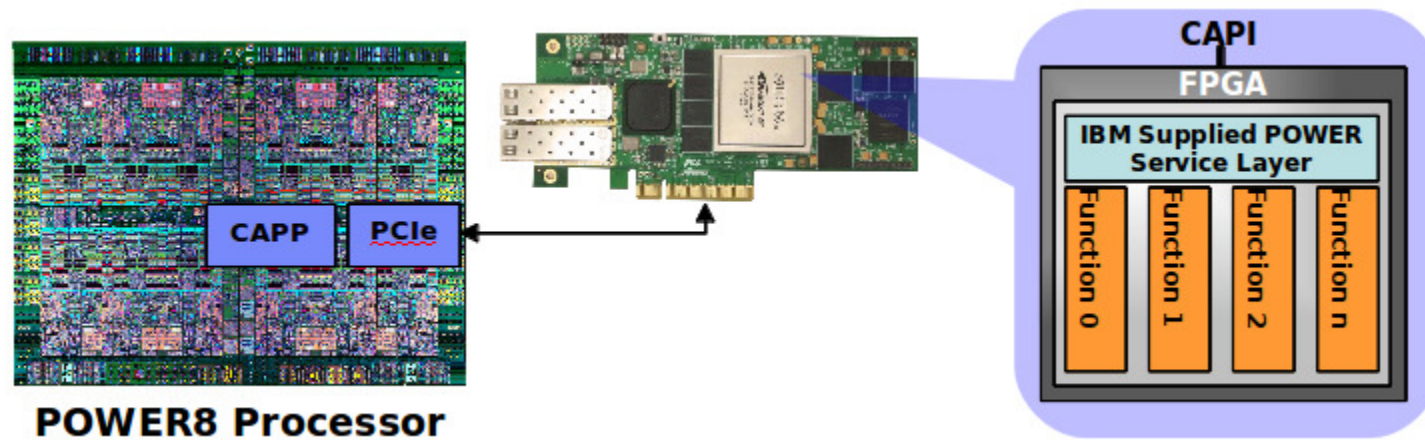


PCIe Gen 3

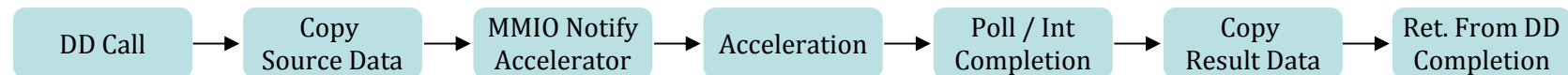
*Transport for encapsulated messages*

**Специализированные контроллеры  
Программные ускорители**

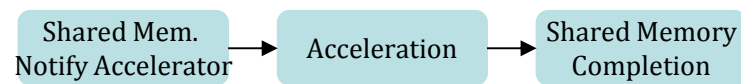
## Coherent Accelerator Processor Interface (CAPI) Flow



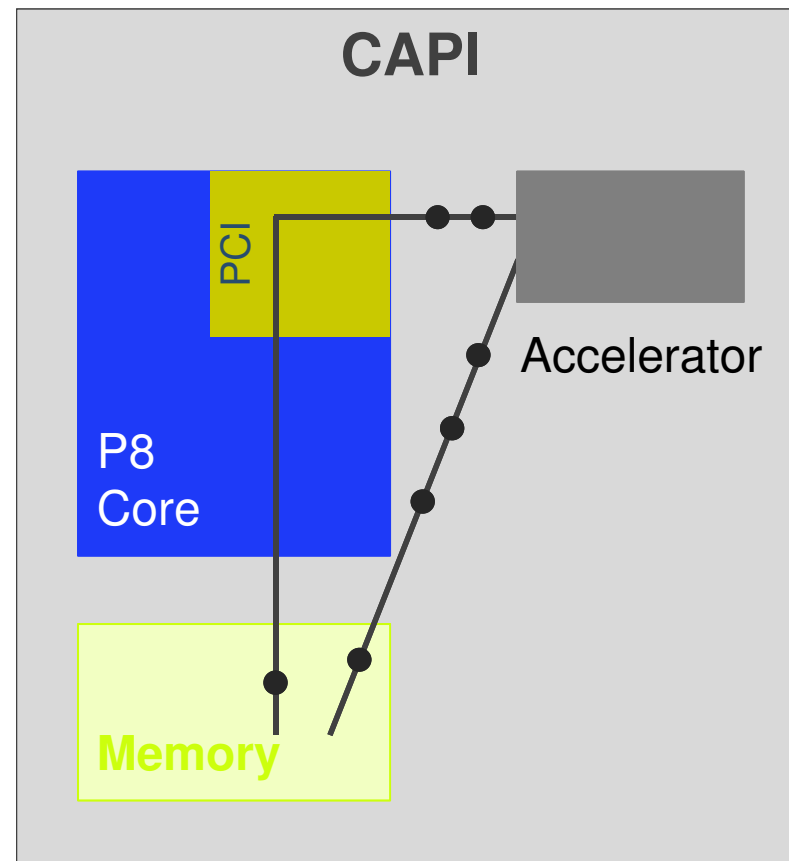
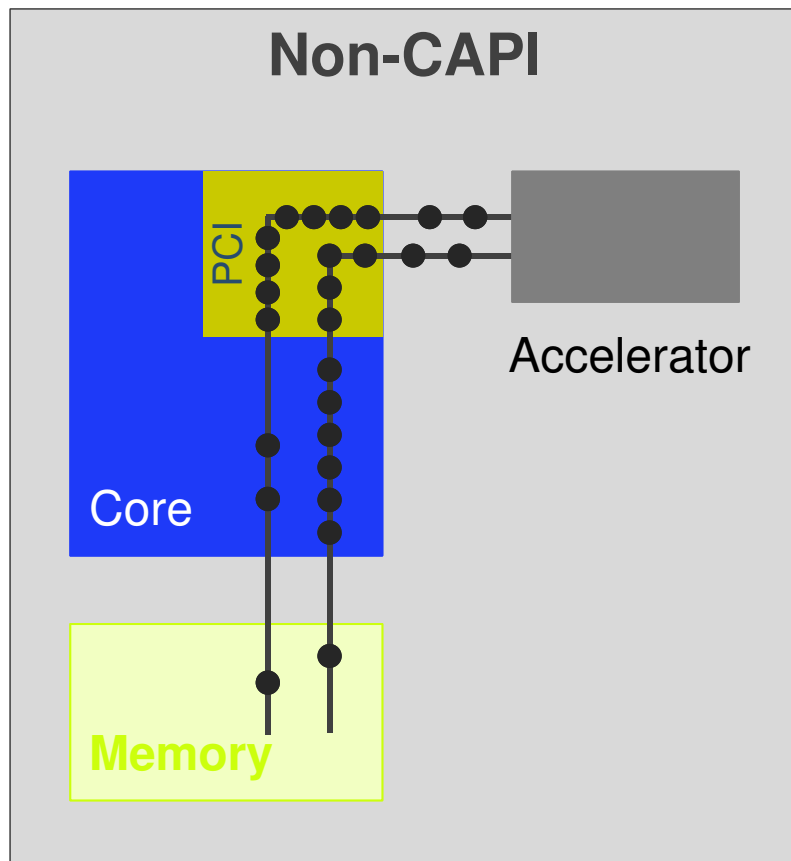
### Типичный процесс работы I/O



### Процесс при использовании когерентной памяти



# Coherent Accelerator Processor Interface





# Несколько слов о стратегии в области НРС



## Развитие стратегии аппаратных средств для HPC

- Общий дизайн платформы для высокопроизводительных вычислений и высокопроизводительной аналитики
- Углубление отношений с технологическими партнёрами
- Серверы для данного сегмента в основном 2 сокета
- Усиление поддержки InfiniBand и Ethernet
- Большая часть производительности на операциях с плавающей точкой будет достигаться за счёт GPU
- Стандартные промышленные стойки и корпуса
  - Варианты воздушного и водяного охлаждения

## Стратегия развития процессоров архитектуры POWER

- Консолидация усилий и фокус на одном процессоре (чипе) общего назначения для каждого поколения
  - ❖ Дизайн для более плотной интеграции с вспомогательным оборудованием
  - ❖ Множественный дизайн модулей обеспечивает различные комбинации памяти и шин I/O
- Использование ускорителей подключаемых к процессору для соответствующих платформ и приложений
  - ❖ FPGA для коммерческих задач, таких как Java, СУБД, аналитика
  - ❖ GPU для научных и вычислительных задач

# OpenPOWER Foundation – что, как, зачем.



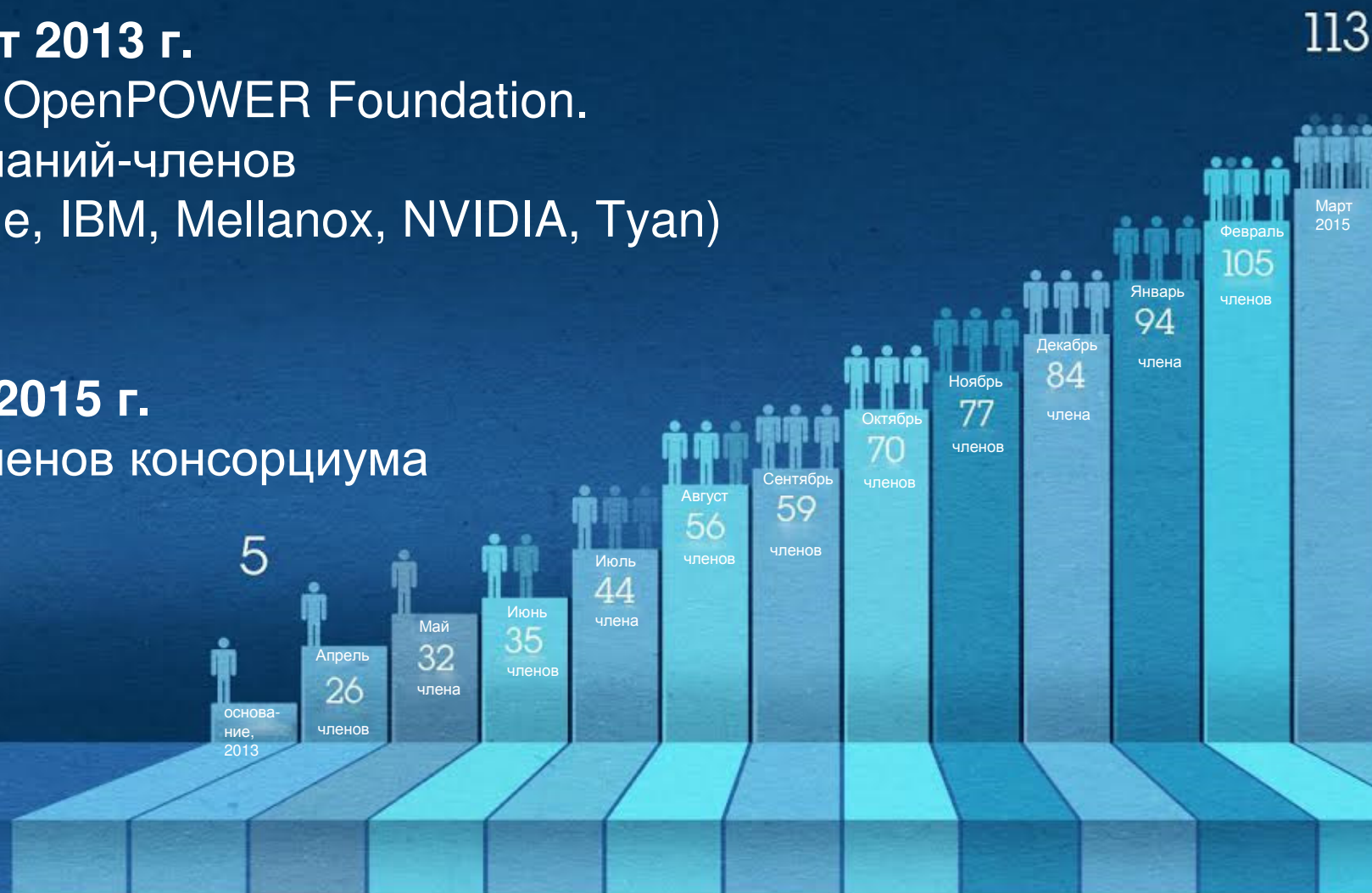
## Основные особенности OpenPOWER

- Это общественная организация, деятельность которой не регулируется кем бы то ни было. Ни коммерческими, ни государственными структурами
- Идея близка к концепции ПО с открытым кодом, но в применении к аппаратуре
- Отличие от мира СПО – участники консорциума кооперируются, а не конкурируют.
- Каждый участник делает свою часть или создаёт свои изделия используя наработки остальных участников сообщества.



**Август 2013 г.**  
анонс OpenPOWER Foundation.  
5 компаний-членов  
(Google, IBM, Mellanox, NVIDIA, Tyan)

**Март 2015 г.**  
113 членов консорциума





Bauman Moscow State  
Technical University

Rikor.IT

Technoprom

KNS Group

26 университетов

22 страны

8 рабочих групп



## Google on POWER

- Many Google Apps enabled on POWER
- Majority of infrastructure ported to POWER
- For most Googlers, enabling POWER is a config change

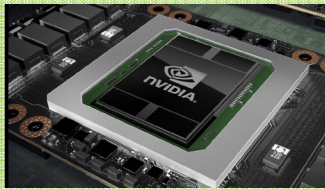
# Специализированный сервер для **HPC**





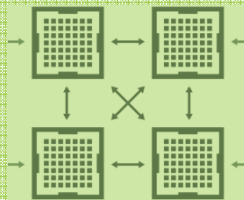
# TESLA P100 ACCELERATOR

## Pascal Architecture



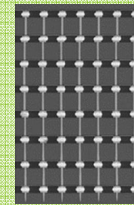
Highest Compute Performance

## NVLink



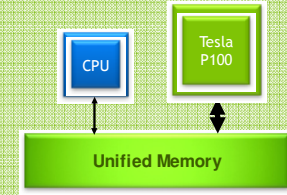
GPU Interconnect for Maximum Scalability

## CoWoS HBM2



Unifying Compute & Memory in Single Package

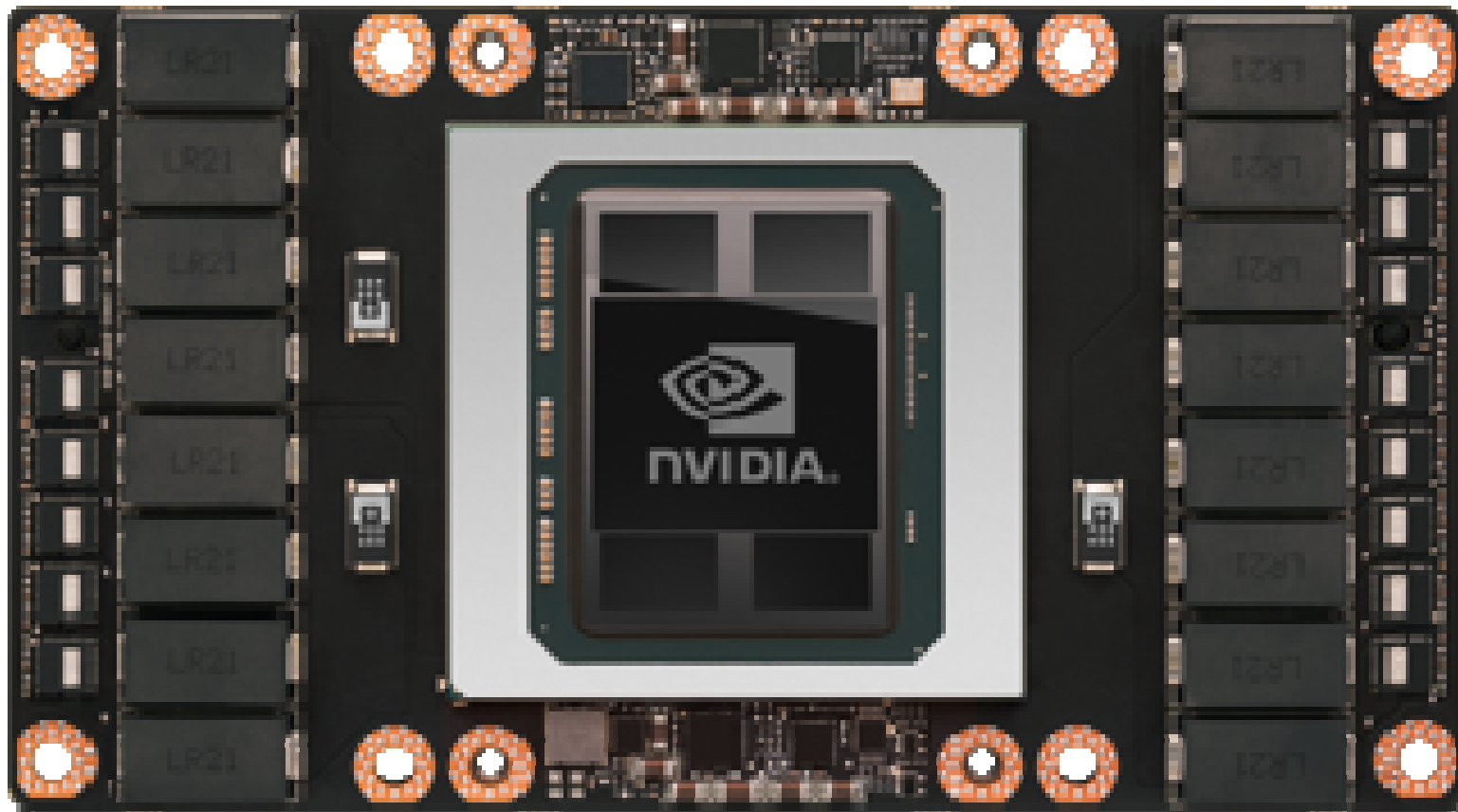
## Page Migration Engine



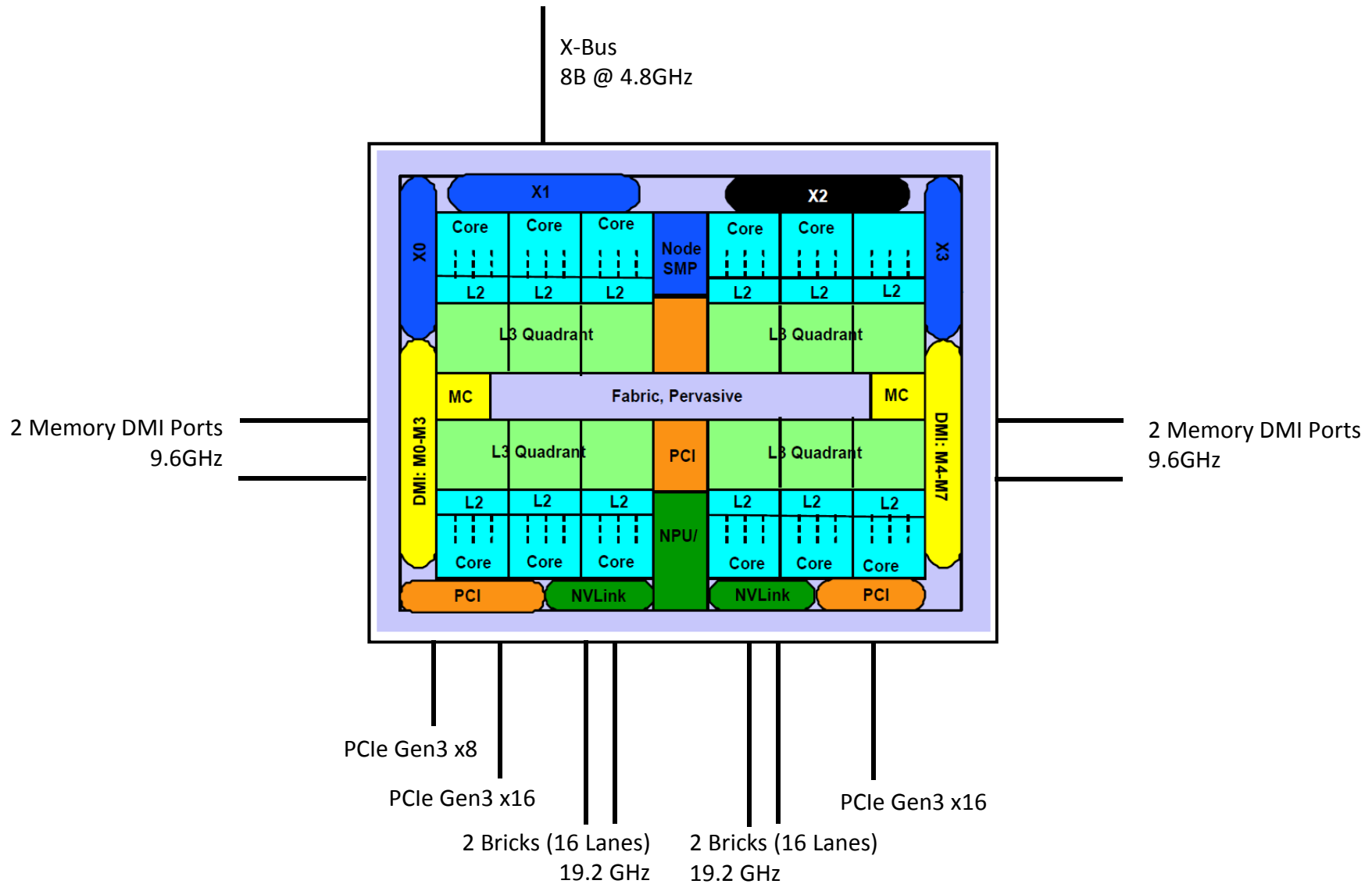
Simple Parallel Programming with Virtually Unlimited Memory Space

|                 |                                              |
|-----------------|----------------------------------------------|
| Compute         | 5.3 TF DP · 10.6 TF SP · 21.2 TF HP          |
| Memory          | HBM2: 720 GB/s · 16 GB                       |
| Interconnect    | NVLink (up to 8 way) + PCIe Gen3             |
| Programmability | Page Migration Engine<br>Unified Memory      |
| Availability    | Ships in IBM “Minsky” System: September 2016 |

# P100

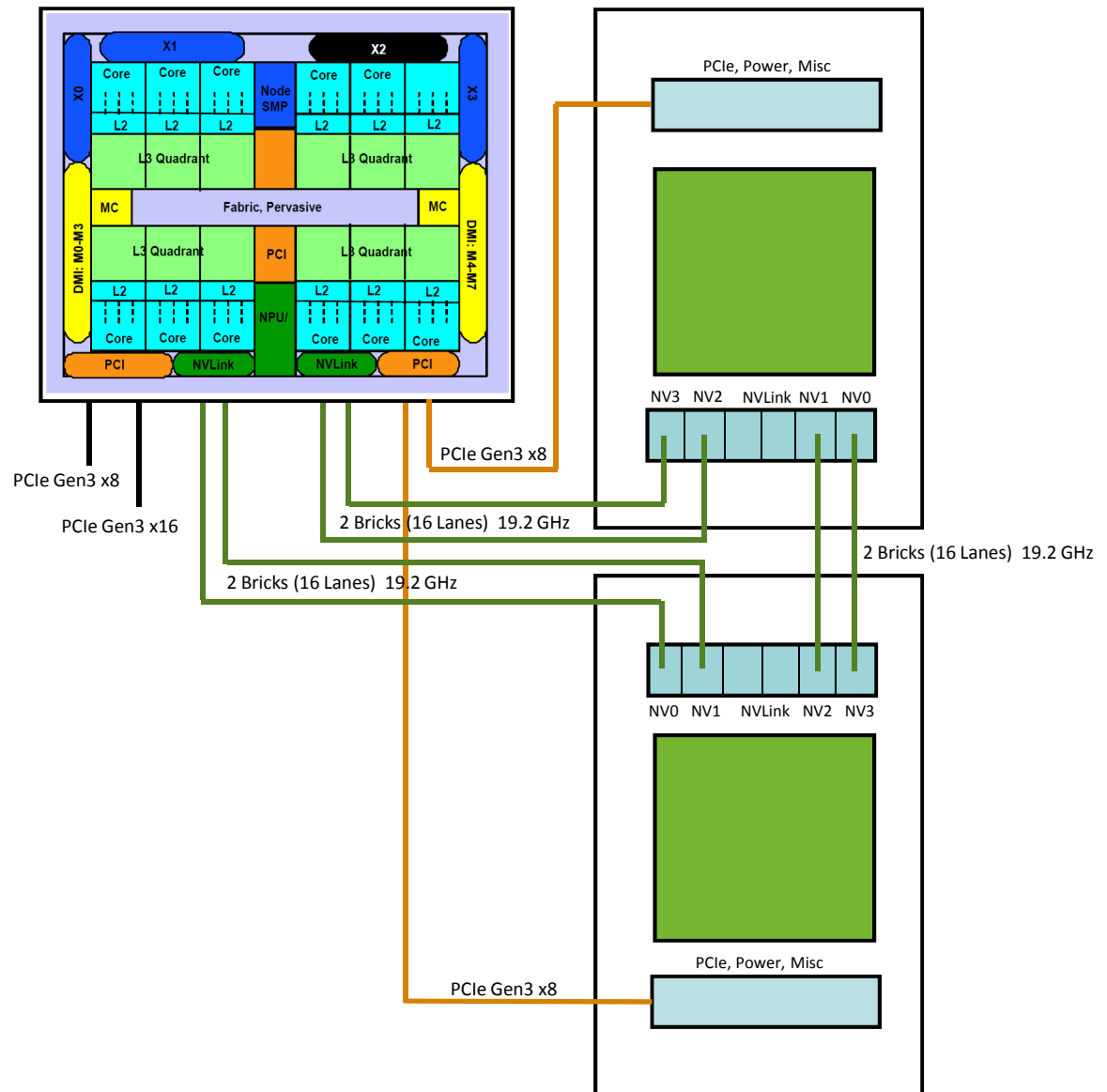


## POWER8 with NVLink Module Interfaces

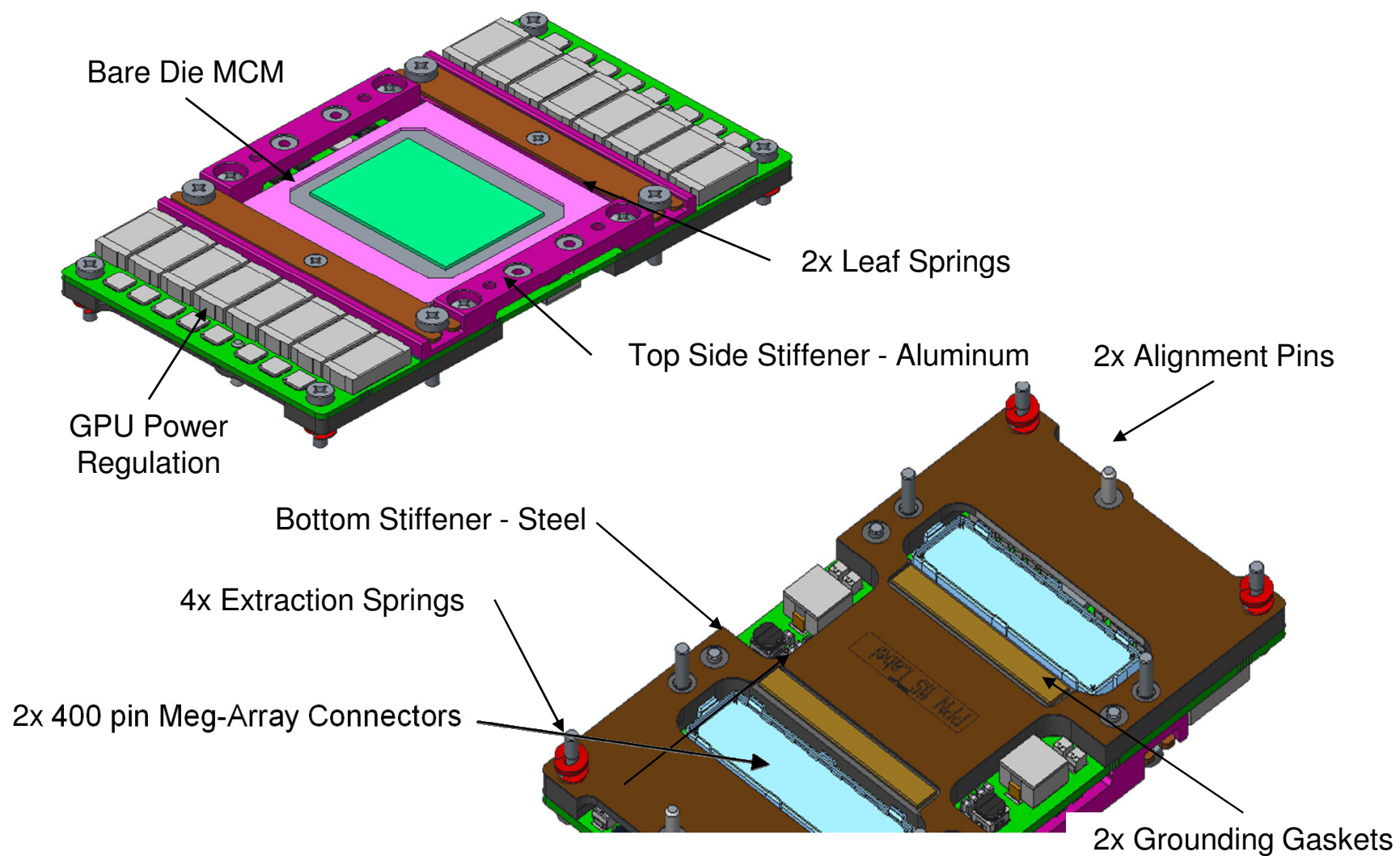


## GPU Interconnect

- NVLink Interface
  - High bandwidth interface far exceeding any existing or planned future PCIe interface
  - 16 Lanes CPU to GPU
  - 16 Lanes GPU to GPU
- PCIe Interface
  - Used for initiation, control and in band reporting of GPU status
  - PCIe x16 interface on the GPU, Garrison uses this interface in x8 mode



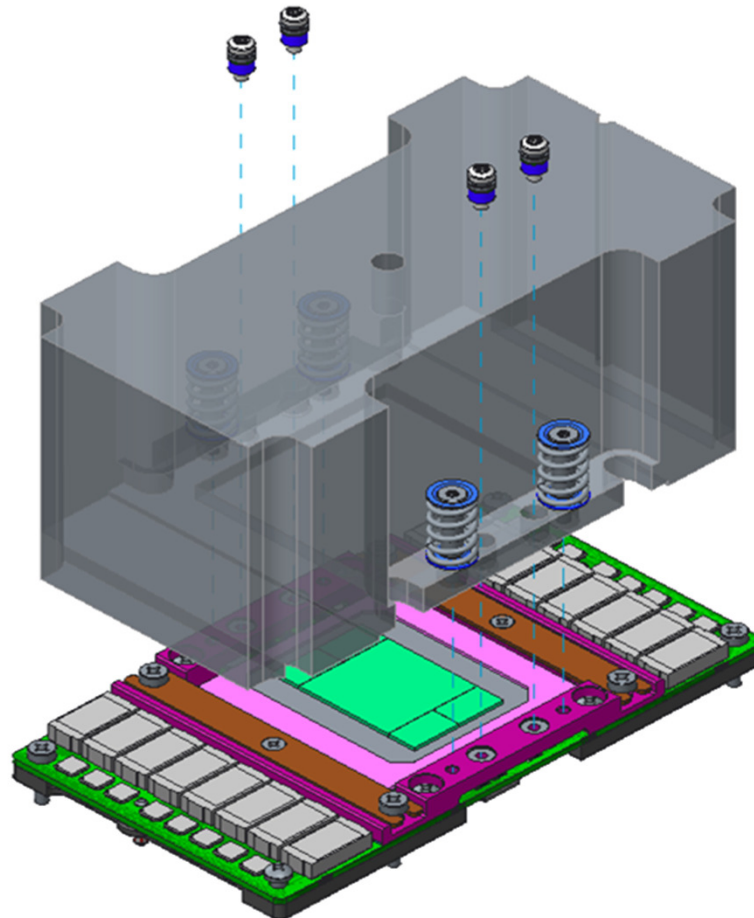
# NVIDIA Pascal GPU





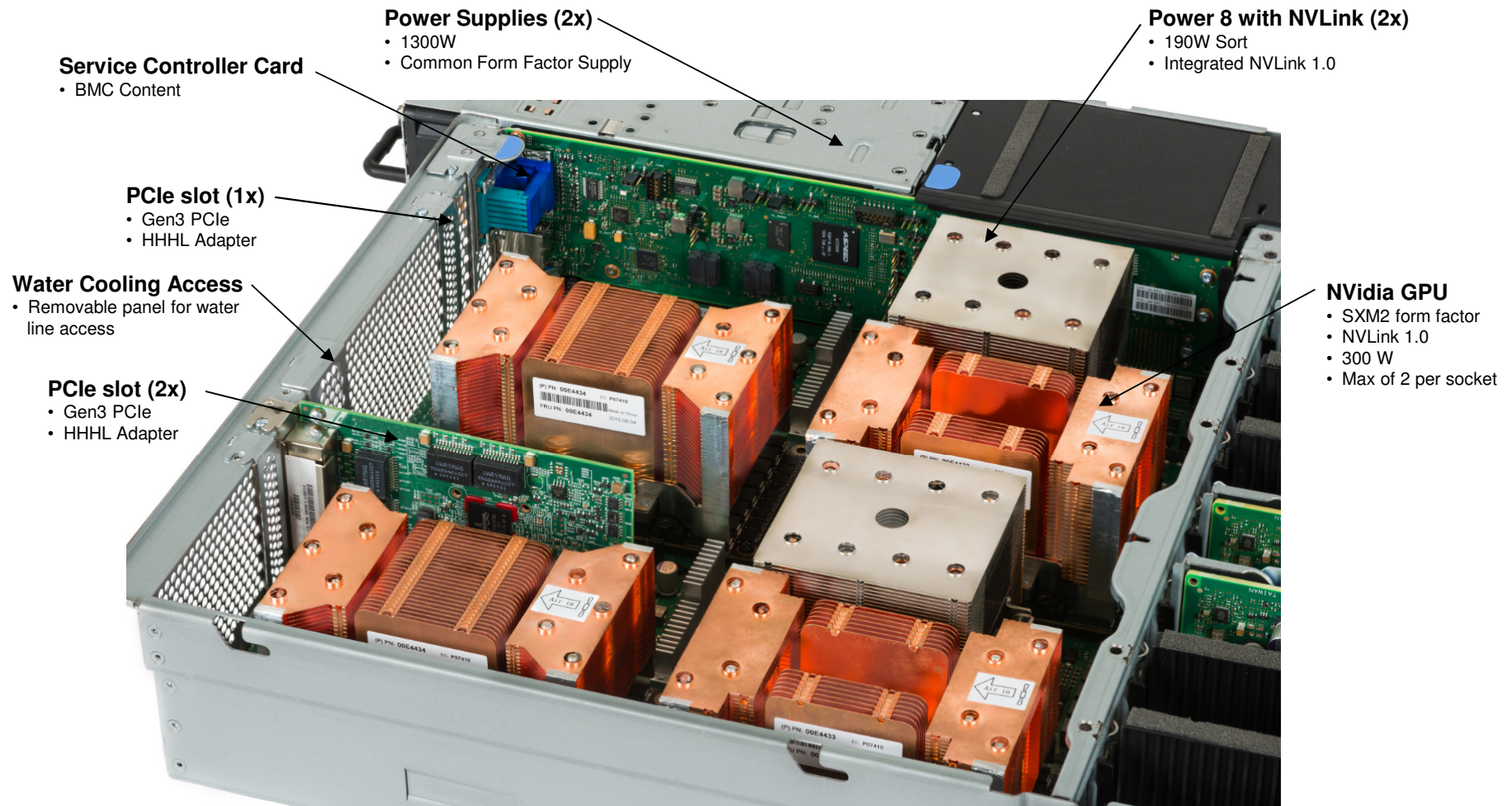
# NVIDIA Pascal GPU с радиатором

IBM FRU Creation from NVidia PPN

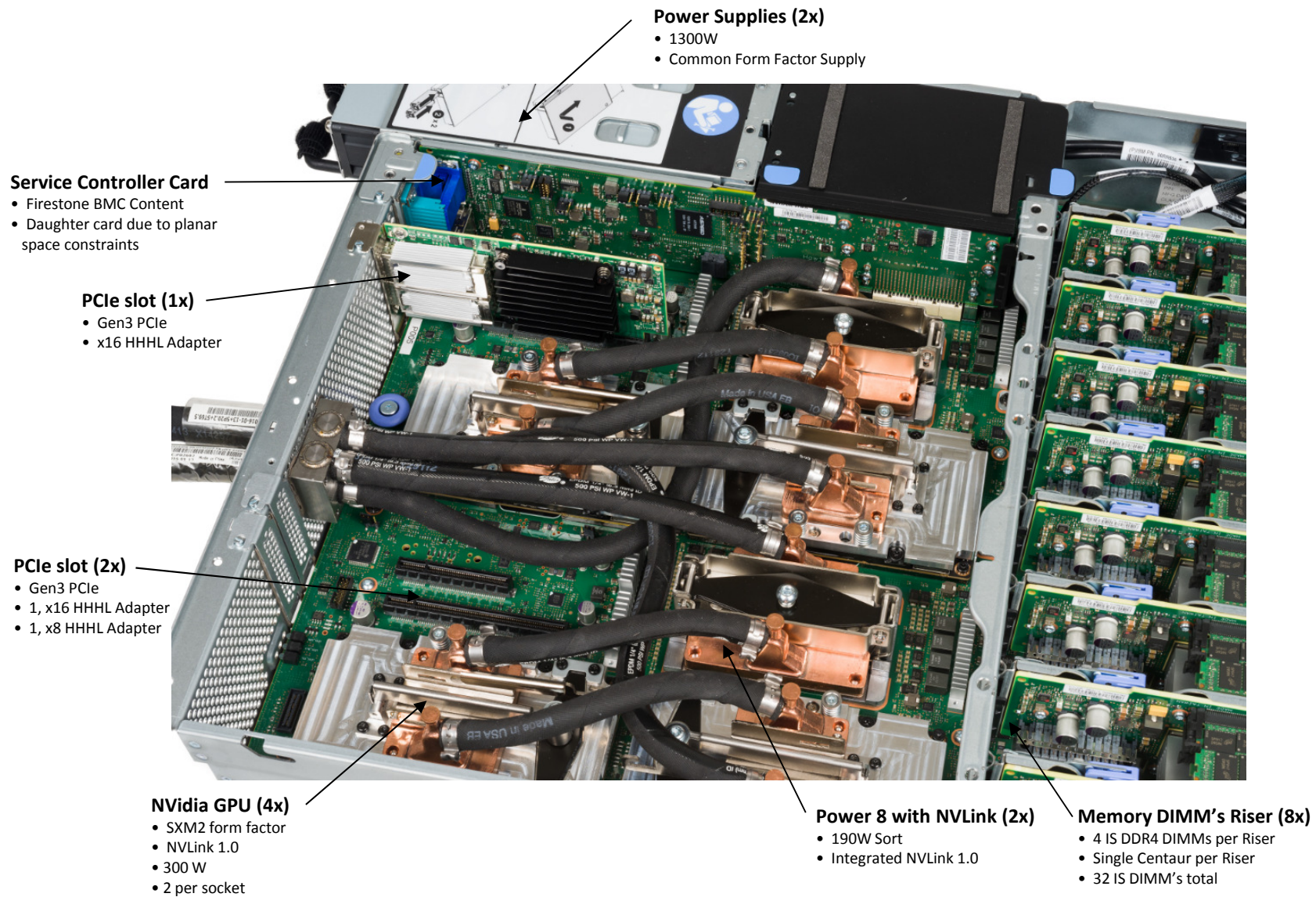


*“Assemble TIM, heatsink & NIFs to NVidia PPN”*

## 2 Socket P8 with NVLink + 4 GPU P100 = 21.5TFlops in 2U

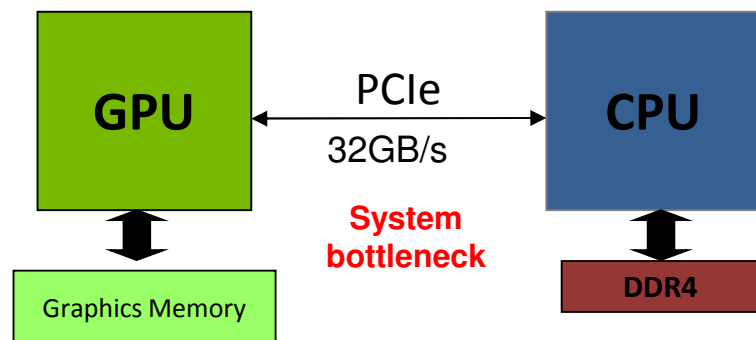




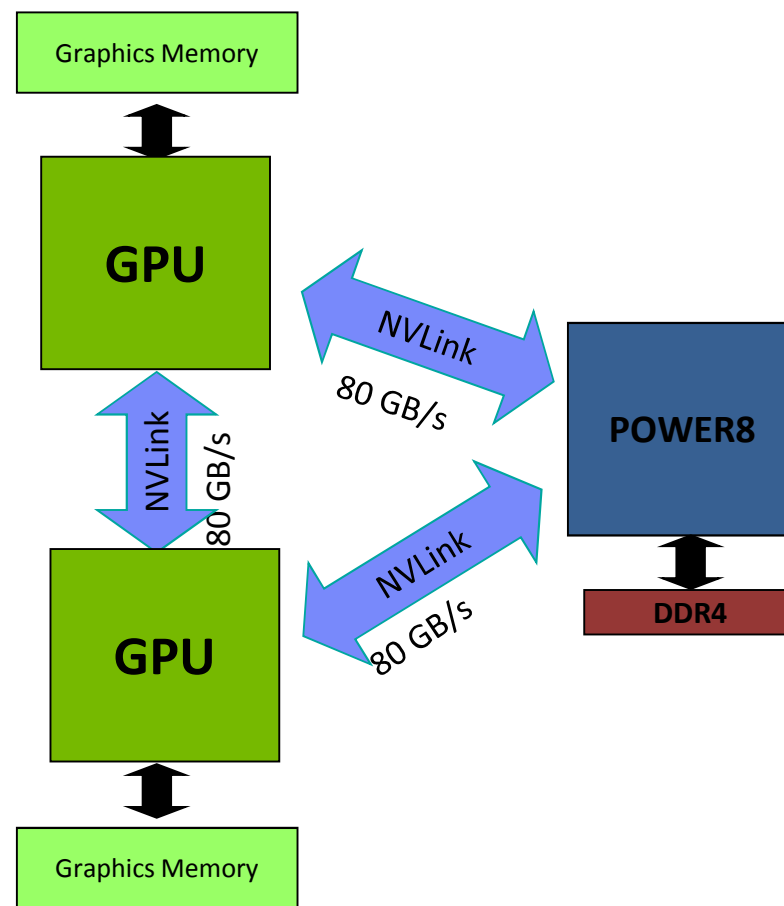


## NVLink: формально в 2.5 раза быстрее связь CPU-GPU

Реально: 3-4ГБ/с vs 17-18ГБ/с  
на PCIe на NVLink



GPUs Limited by PCIe Bandwidth  
From CPU-System Memory



NVLink Enables Fast Unified Memory Access  
between CPU & GPU Memories

## Design: Flat and Fat

### Дизайн “flat and fat”

- Данные свободно протекают в системе
- Полоса CPU: GPU почти такая же как Системная Память: CPU
- Широкие каналы между GPU подключенными к тому же сокету

Устраняет ограничения PCI-e для многих типов задач

- Пики на старте / сброс итогов
- Обеспечение непрерывного потока данных Host-Device
- Постоянные пересылки между 2 GPU
- Скрытые пересылки по шине в направлении Host-Device

