IBM и HPC

Алексей Перевозчиков
IBM Server Solutions Product Manager

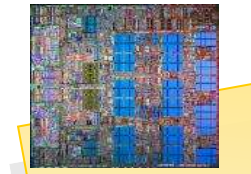
82189117@ru.ibm.com

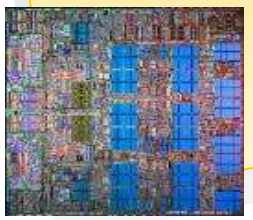
МГУ, 28 июня 2016г.

Power Systems

Планы развития процессора POWER

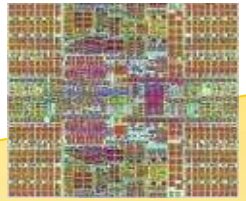
IBM

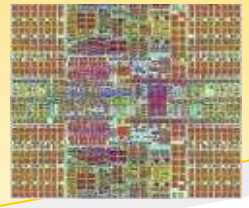




POWER5/5+
130/90 nm

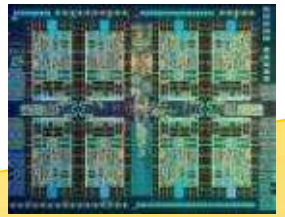
- ✓Dual Core
- ✓Enhanced Scaling
- ✓SMT
- ✓Distributed Switch +
- ✓Core Parallelism +
- ✓FP Performance +
- ✓Memory Bandwidth +
- ✓Virtualization

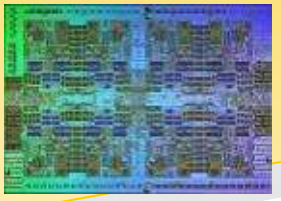




POWER6/6+
65/65 nm

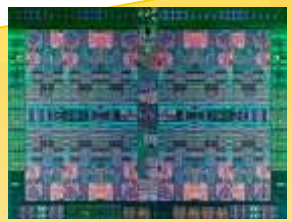
- ✓Dual Core
- ✓High Frequencies
- ✓Virtualization +
- ✓Memory Subsystem +
- ✓Altivec
- ✓Instruction Retry
- ✓Dynamic Energy Mgmt
- ✓SMT +
- ✓Protection Keys





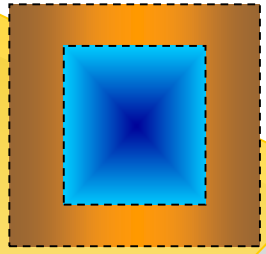
POWER7/7+
45/32 nm

- ✓Eight Cores
- ✓On-Chip eDRAM
- ✓Power-Optimized Cores
- ✓Memory Subsystem ++
- ✓SMT++
- ✓Reliability +
- ✓VSM & VSX
- ✓Protection Keys+



POWER8
22nm

- ✓More Cores
- ✓SMT+++
- ✓Reliability ++
- ✓FPGA Support
- ✓Transactional Memory
- ✓PCIe Acceleration

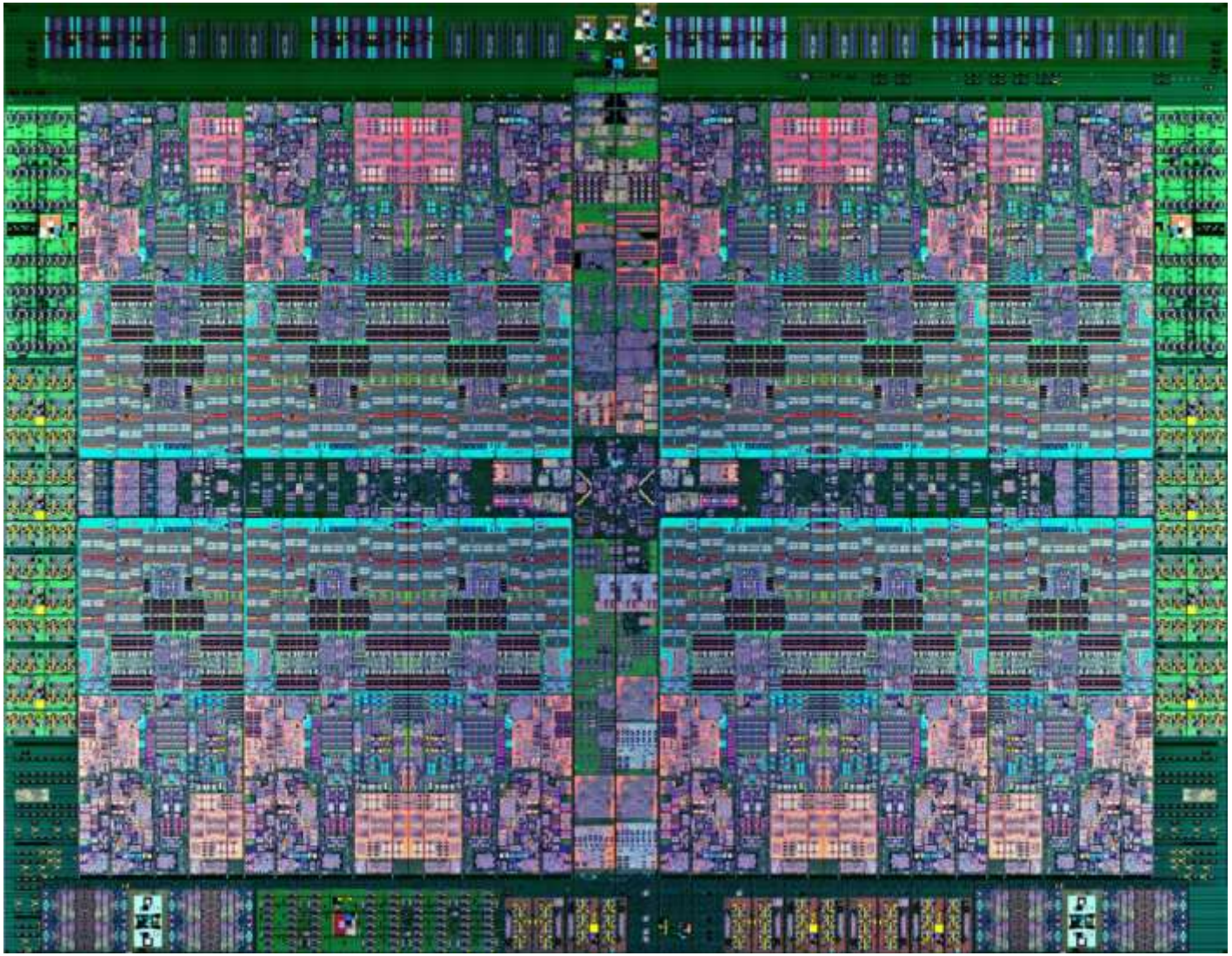


POWER9

- ✓Extreme Analytics Optimization
- ✓Extreme Big Data Optimization
- ✓On-chip accelerators

2004 2007 2010 2014

Процессор POWER8



Процессор POWER8

Технология

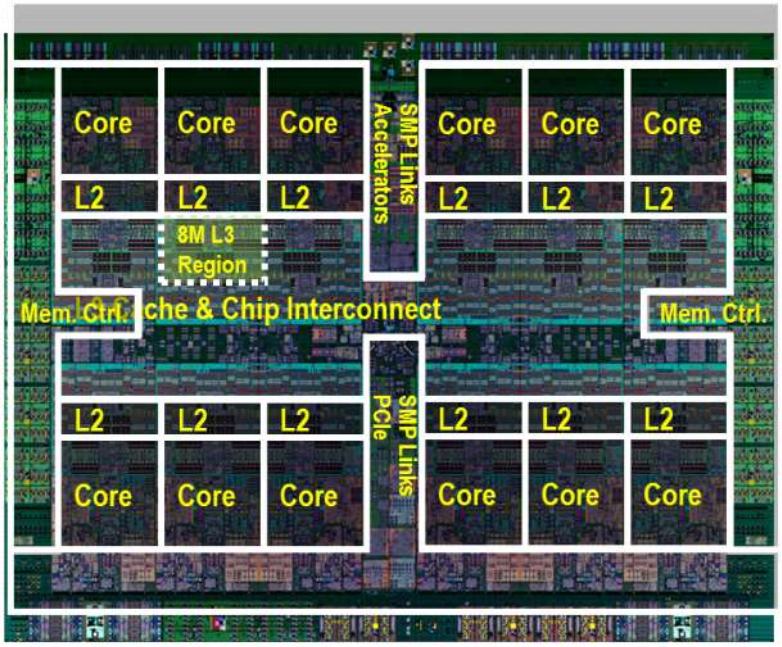
22nm SOI, eDRAM, 650mm2, 4.2B transistors

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

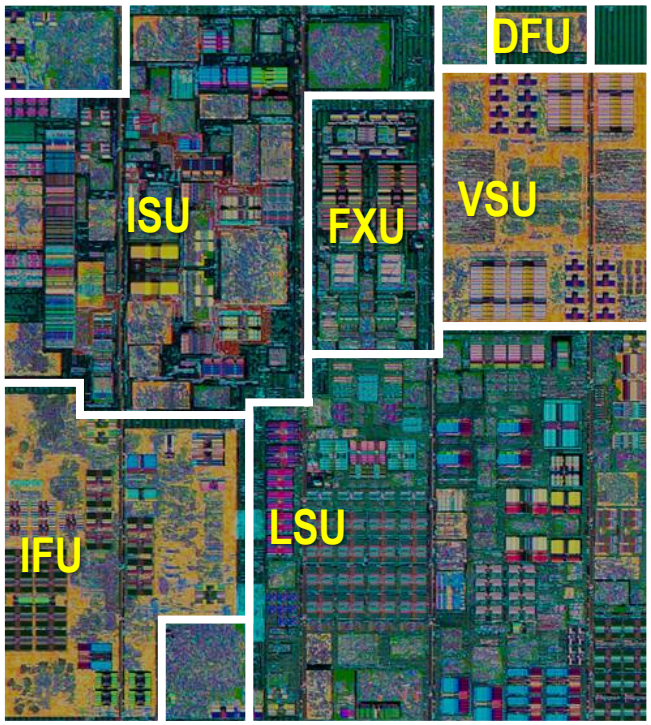
- **Up to 230 GB/s sustained bandwidth**

Шинные интерфейсы

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

Ядро POWER8 (относительно POWER7)

- **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

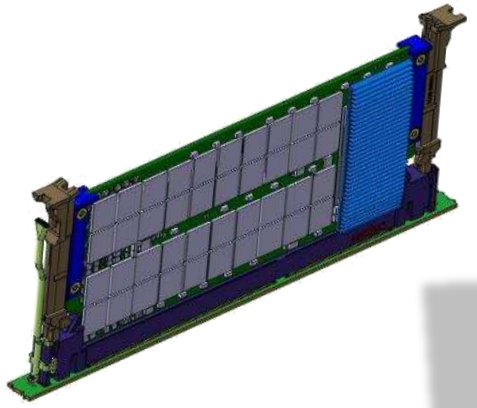
Enhanced Prefetch

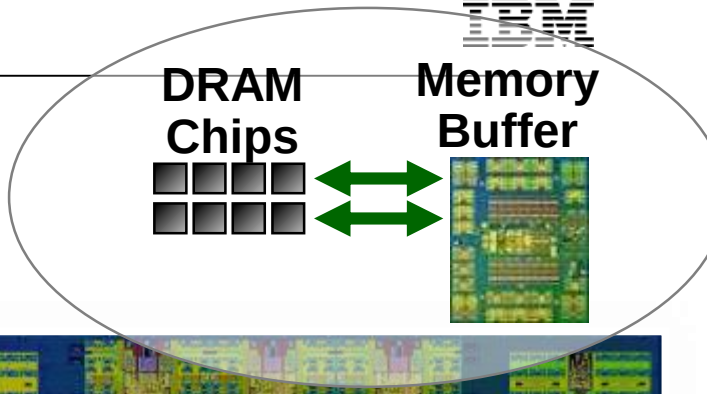
- **Instruction speculation awareness**
- Data prefetch depth awareness
- Adaptive bandwidth awareness
- Topology awareness

Power Systems

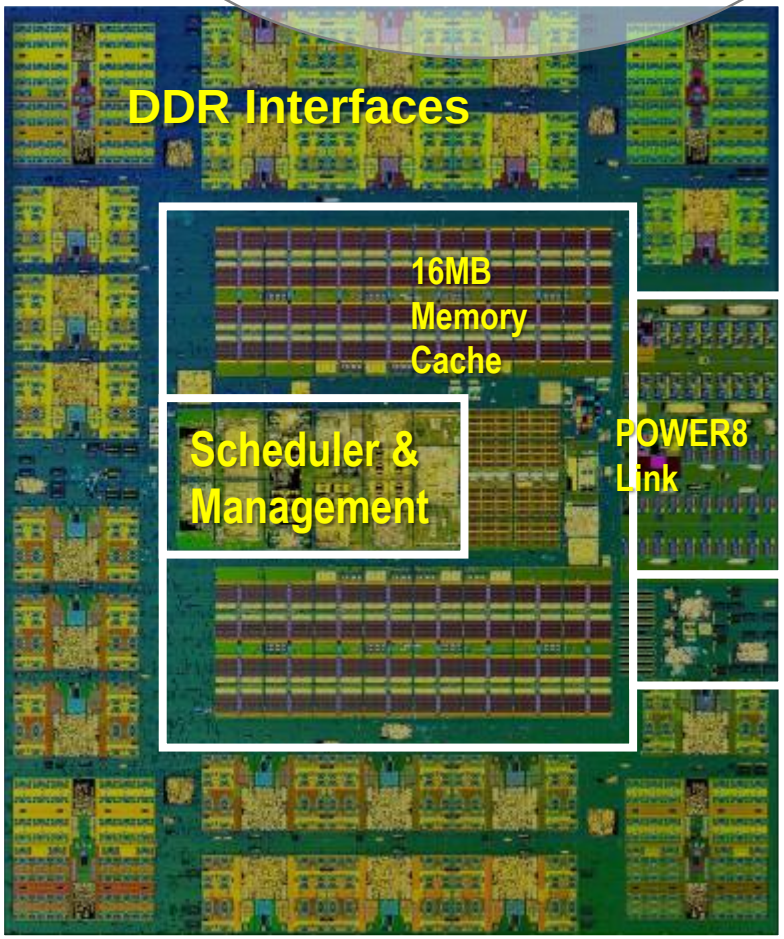
Memory Buffer Chip
…*with 16MB Cache…*

IBM





“L4 cache”

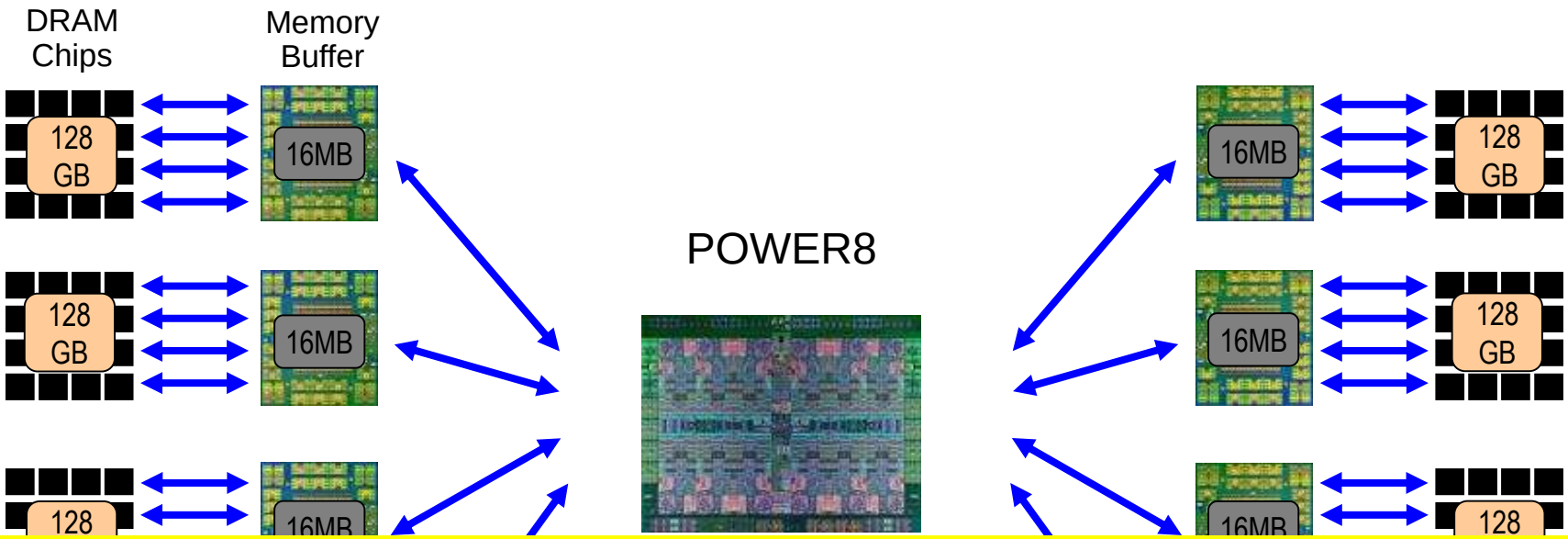


Модули памяти наполняются интеллектом
- Умная система кэширования
- Оптимизация энергии
- Надежность

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)

Организация памяти в POWER8



- 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)

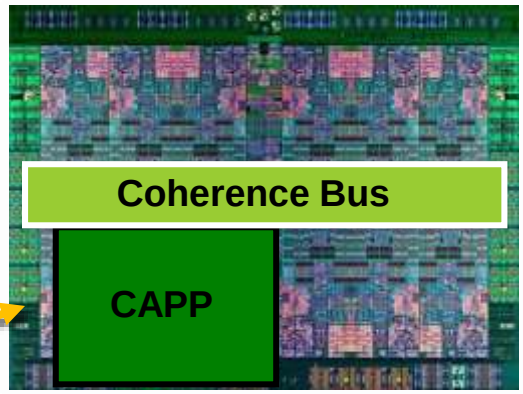
➡ ***До 32 портов DDR выдающих в пике 410 ГБ/с (на уровне DRAM)***

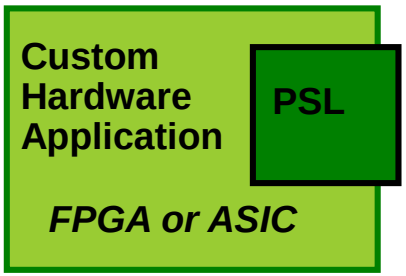
➡ ***До 1 ТБ памяти на сокет (для старших версий – до 2 ТБ на сокет)***

CAPI *(Coherent Accelerator Processor Interface)*

Virtual Addressing
•Ускоритель работает напрямик с разделяемой памятью
•Обмен данными с кэшем процессора.
•Исключает накладные расходы ОС и драйверов.

Hardware Managed Cache Coherence
•Стандартный механизм блокировок.

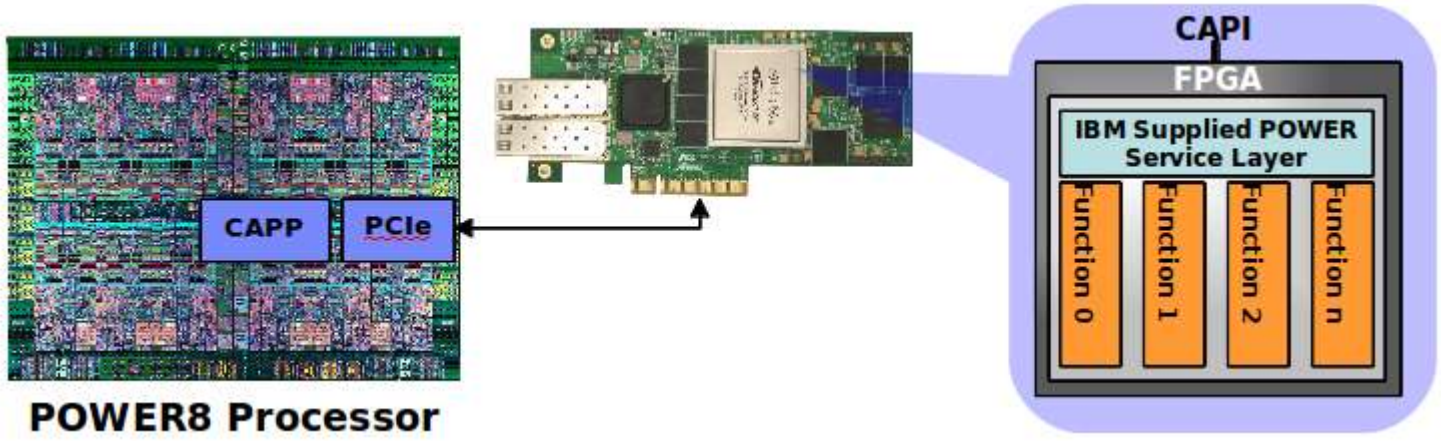




PCIe Gen 3
Transport for encapsulated messages

Специализированные контроллеры
Программные ускорители

Coherent Accelerator Processor Interface (CAPI) Flow



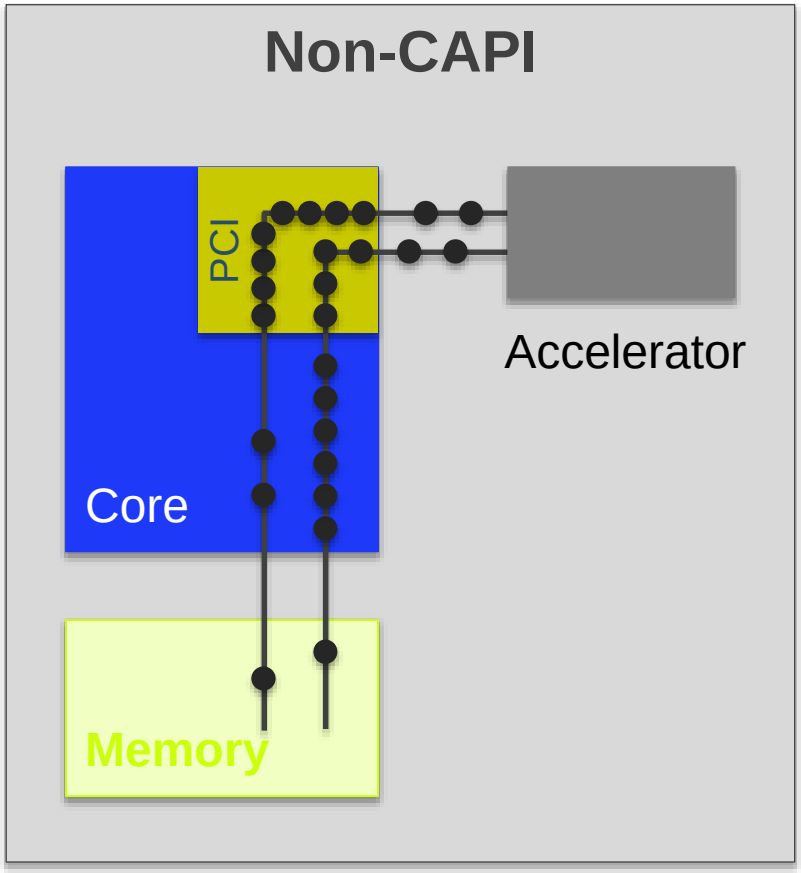
Типичный процесс работы I/O

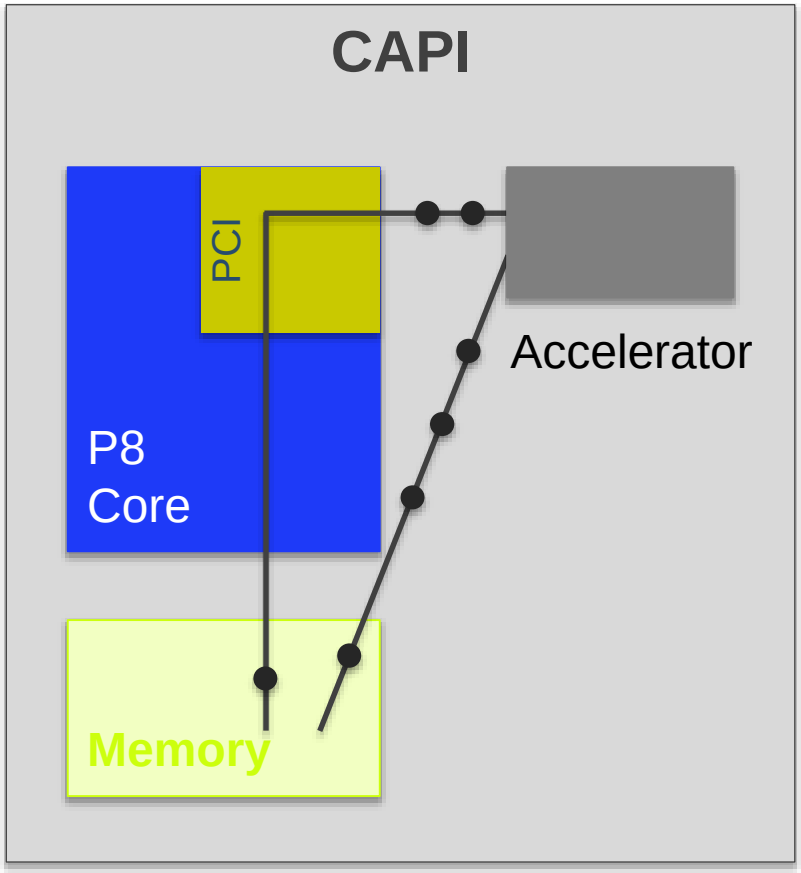
DD Call → Copy Source Data → MMIO Notify Accelerator → Acceleration → Poll / Int Completion → Copy Result Data → Ret. From DD Completion

Процесс при использовании когерентной памяти

Shared Mem. Notify Accelerator → Acceleration → Shared Memory Completion

Coherent Accelerator Processor Interface





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.

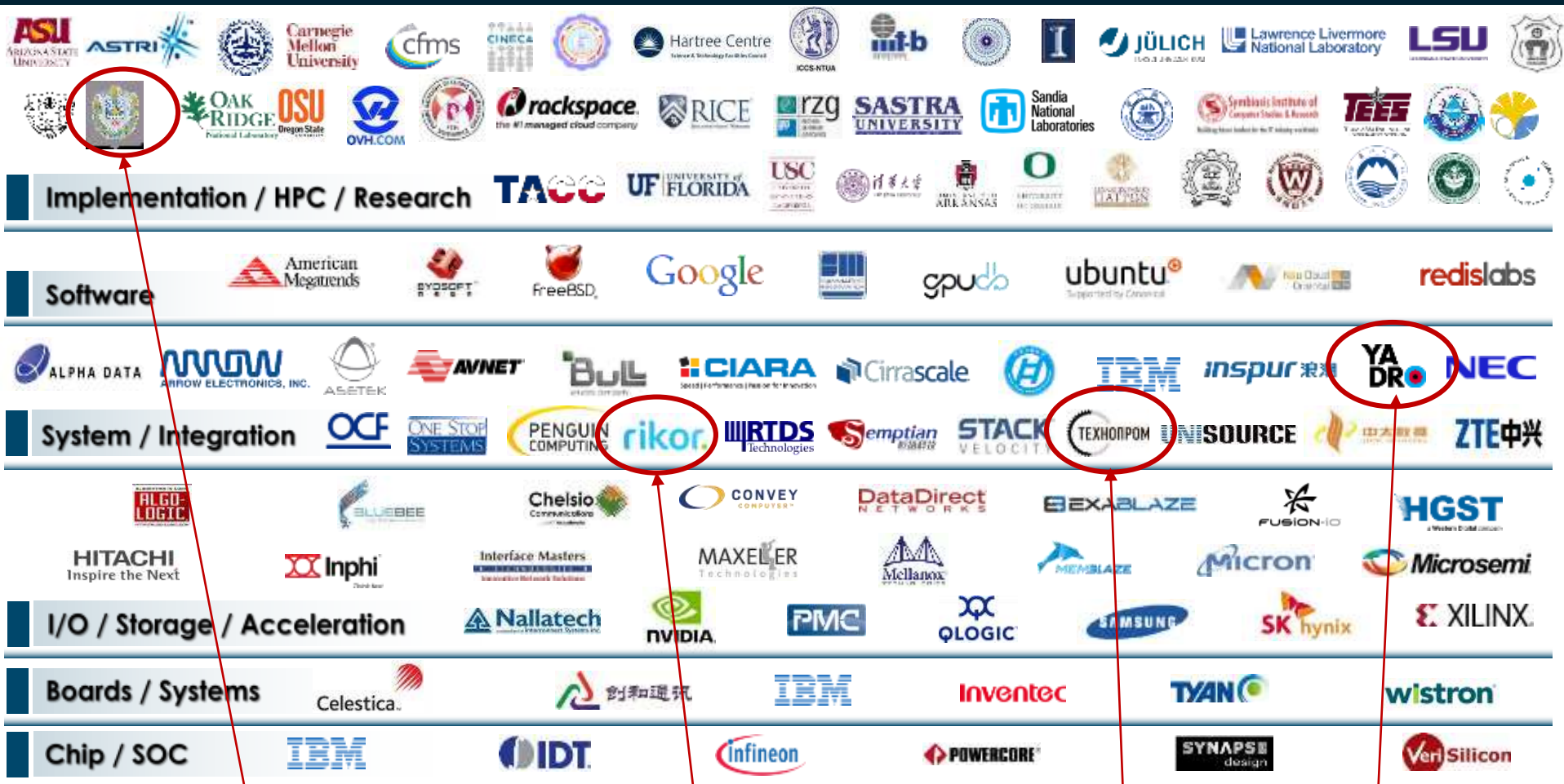
OpenPOWER™

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

Март 2015 г.
113 членов консорциума

Открытое сообщество разработчиков

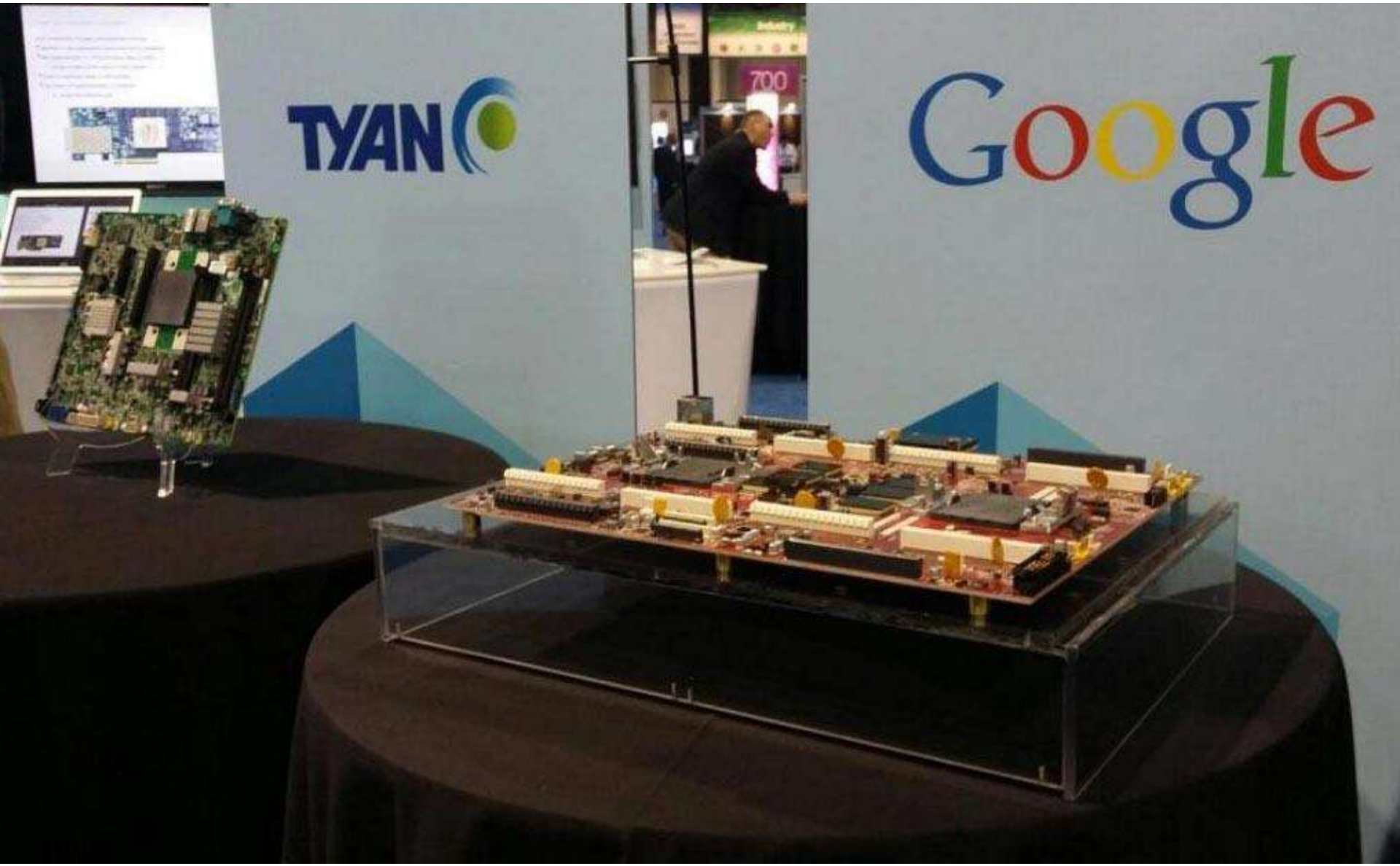
190+ участников консорциума OpenPOWER

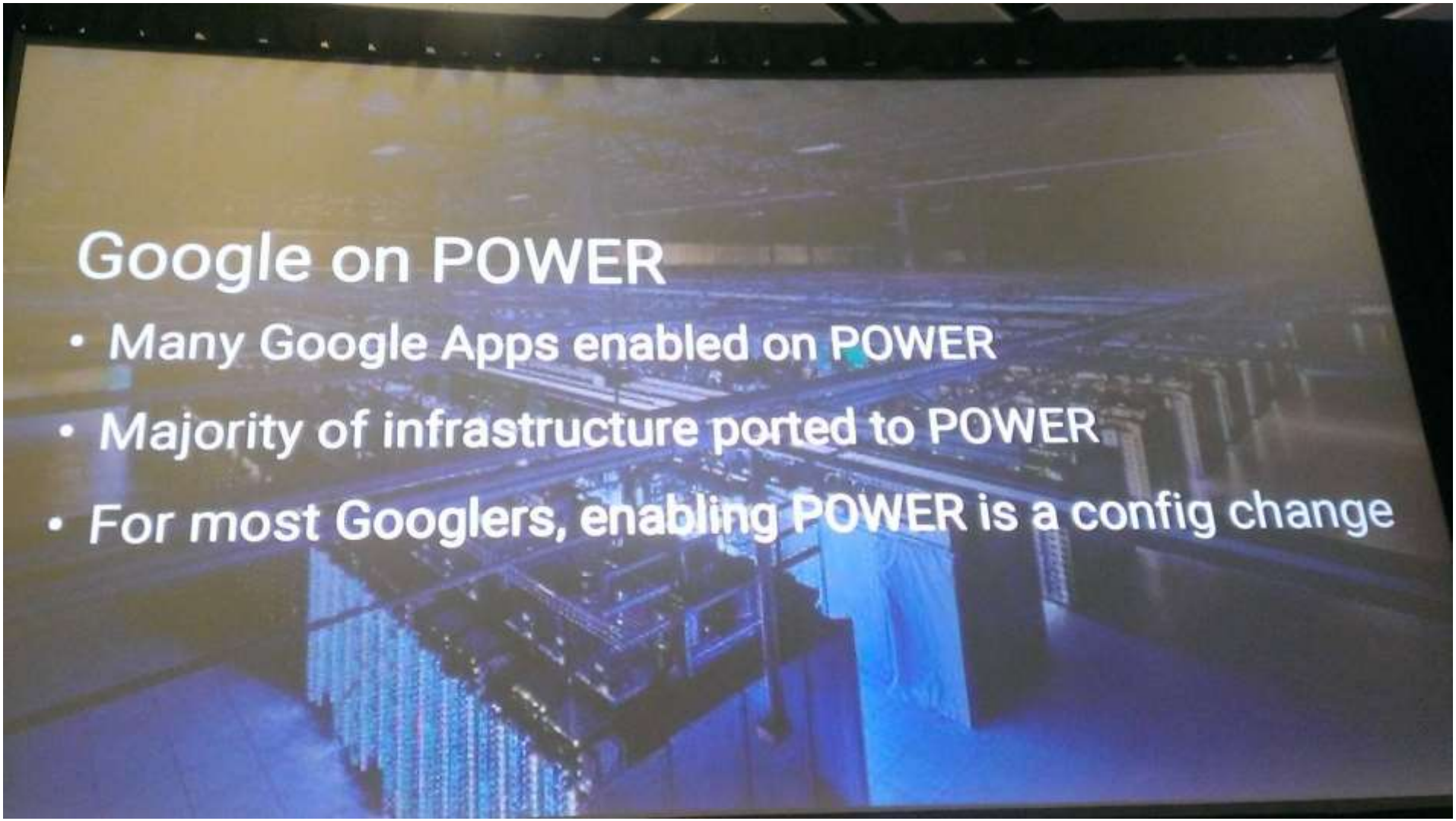


26 университетов

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

Несколько слов о стратегии

Развитие стратегии аппаратных средств для HPC

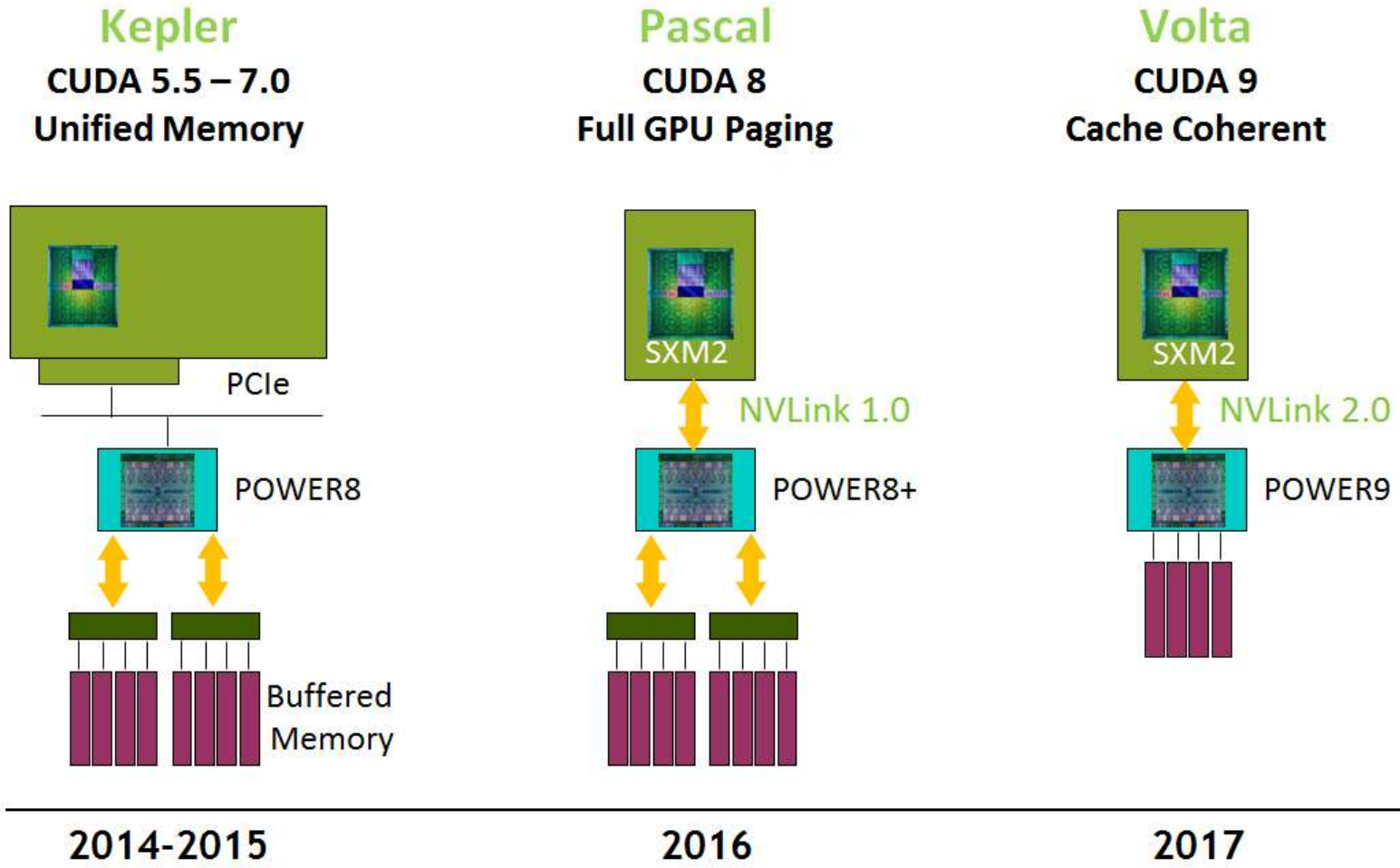
- Общий дизайн платформы для высокопроизводительных вычислений и высокопроизводительной аналитики
- Углубление отношений с технологическими партнёрами
- Серверы для данного сегмента в основном 2 сокета
- Усиление поддержки InfiniBand и Ethernet
- Большая часть производительности на операциях с плавающей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

Стратегия развития процессоров архитектуры POWER

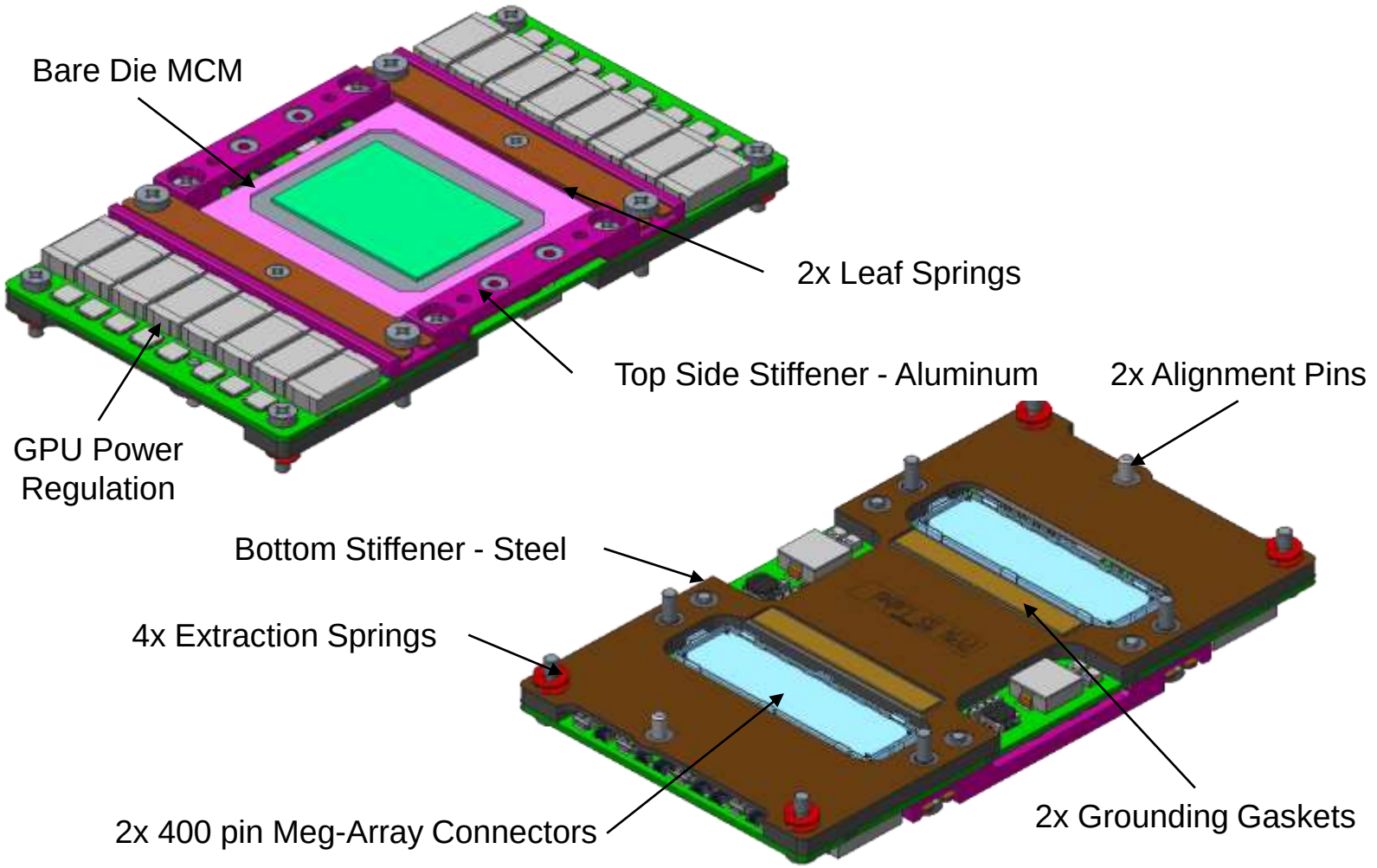
- Консолидация усилий и фокус на одном процессоре (чипе) общего назначения для каждого поколения
  - Дизайн для более плотной интеграции с вспомогательным оборудованием
  - Множественный дизайн модулей обеспечивает различные комбинации памяти и шин I/O
- Использование ускорителей подключаемых к процессору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

NVIDIA GPU Roadmap

NVIDIA

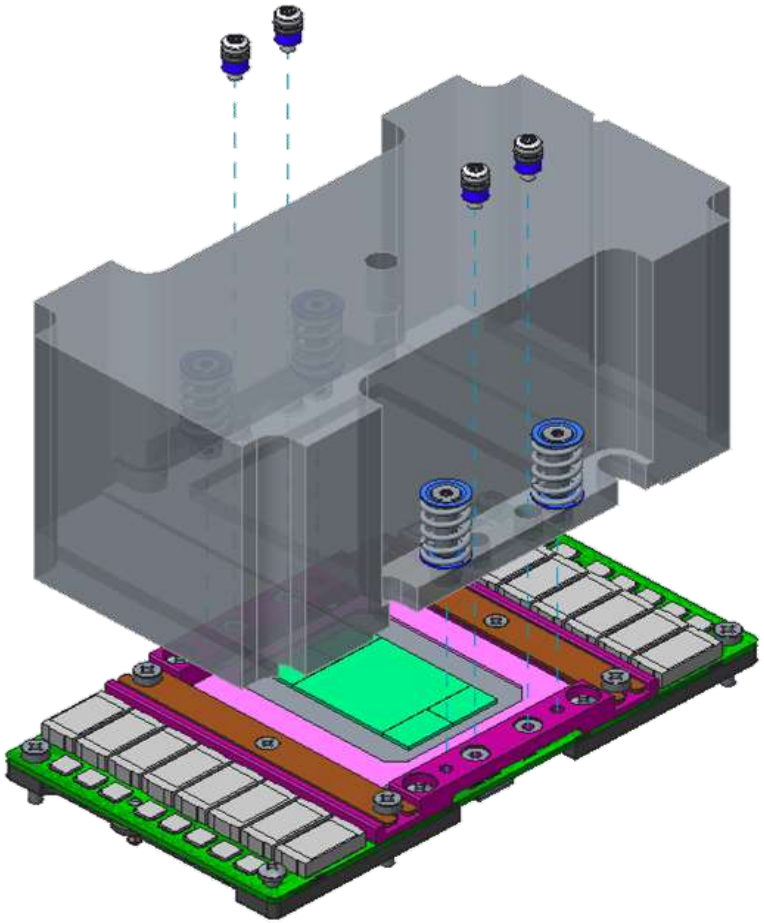


NVIDIA Pascal GPU



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

Power Supplies (2x)
- 1300W
- Common Form Factor Supply

Power 8 with NVLink (2x)
- 190W Sort
- Integrated NVLink 1.0

Service Controller Card
- BMC Content

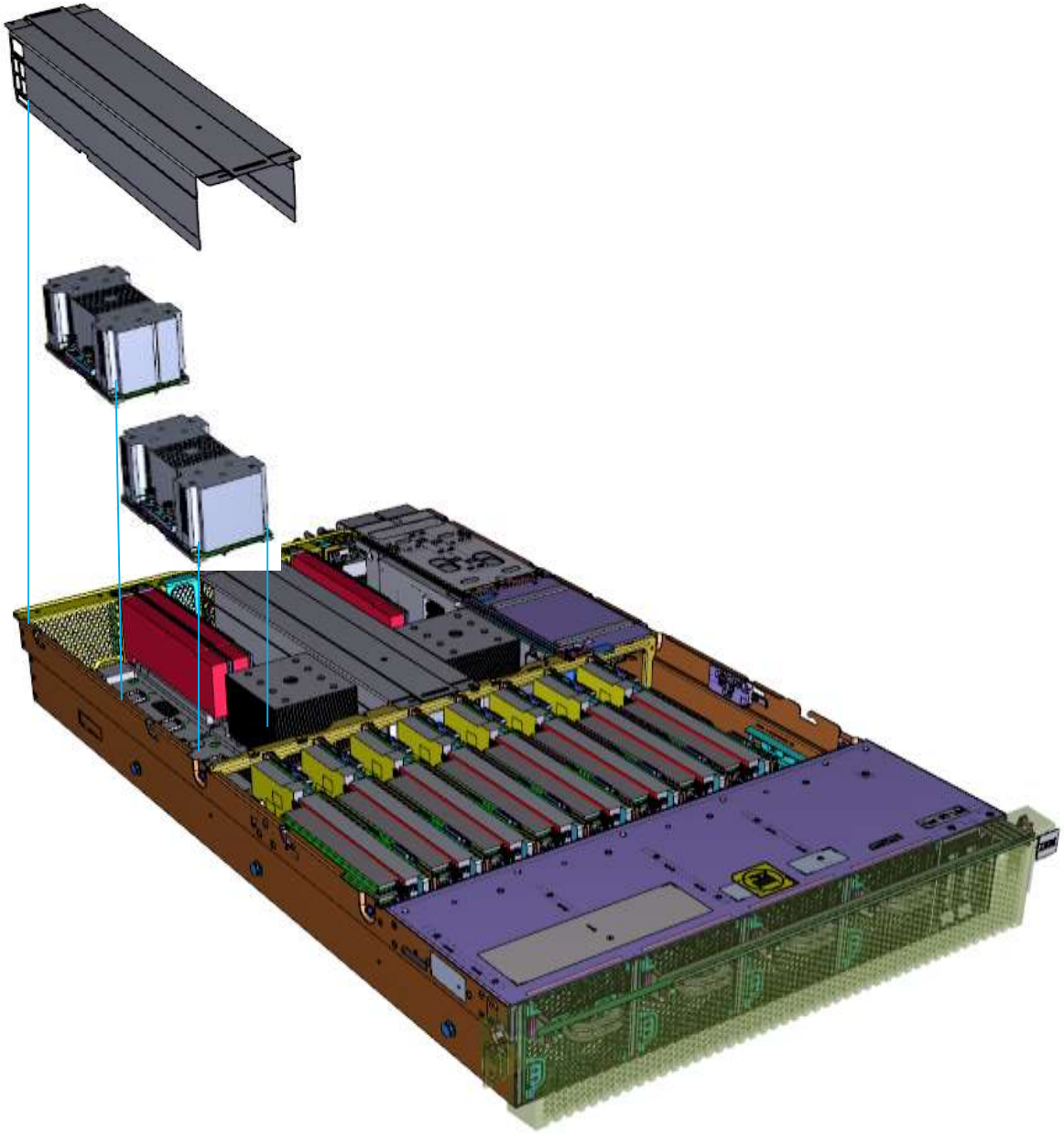
PCIe slot (1x)
- Gen3 PCIe
- HHHL Adapter

Water Cooling Access
- Removable panel for water line access

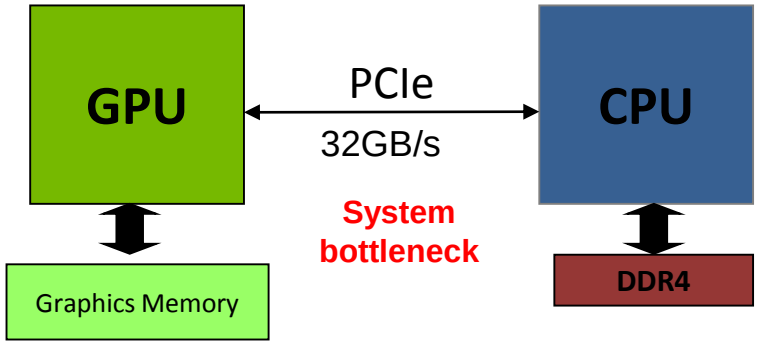
PCIe slot (2x)
- Gen3 PCIe
- HHHL Adapter

NVidia GPU
- SXM2 form factor
- NVLink 1.0
- 300 W
- Max of 2 per socket

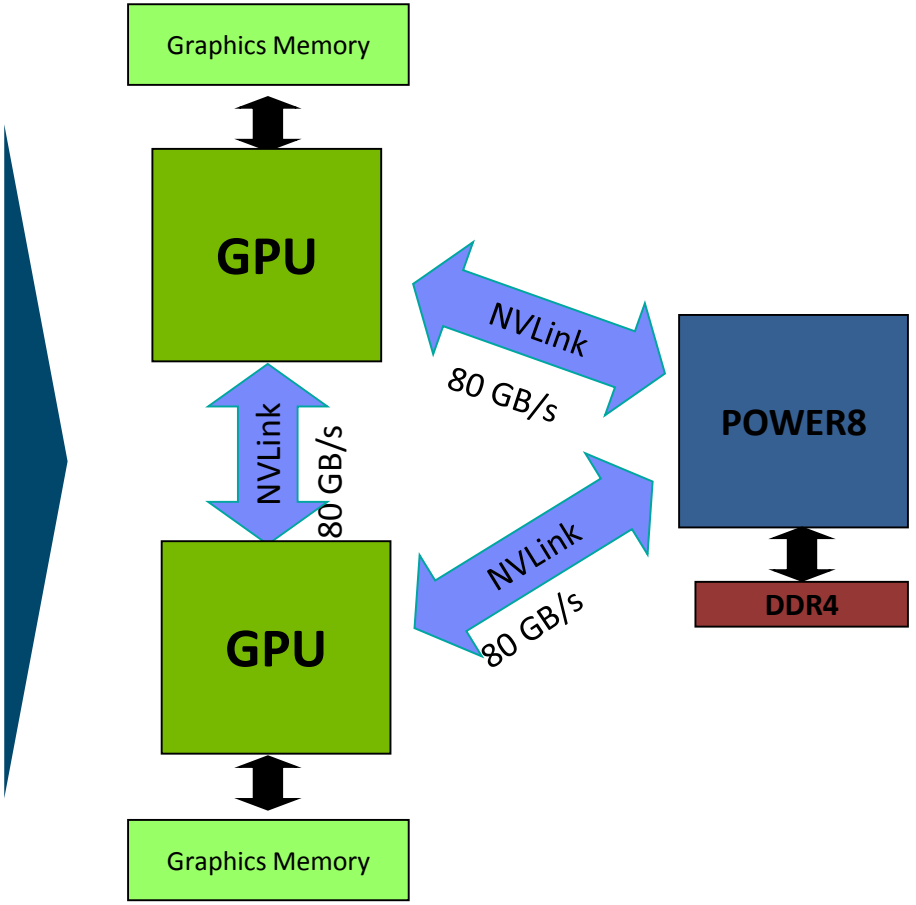
Установка GPU



POWER8 with NVLink: в 2.5 раза быстрее связь CPU-GPU



GPUs Limited by PCIe Bandwidth
From CPU-System Memory



NVLink Enables Fast Unified Memory Access
between CPU & GPU Memories

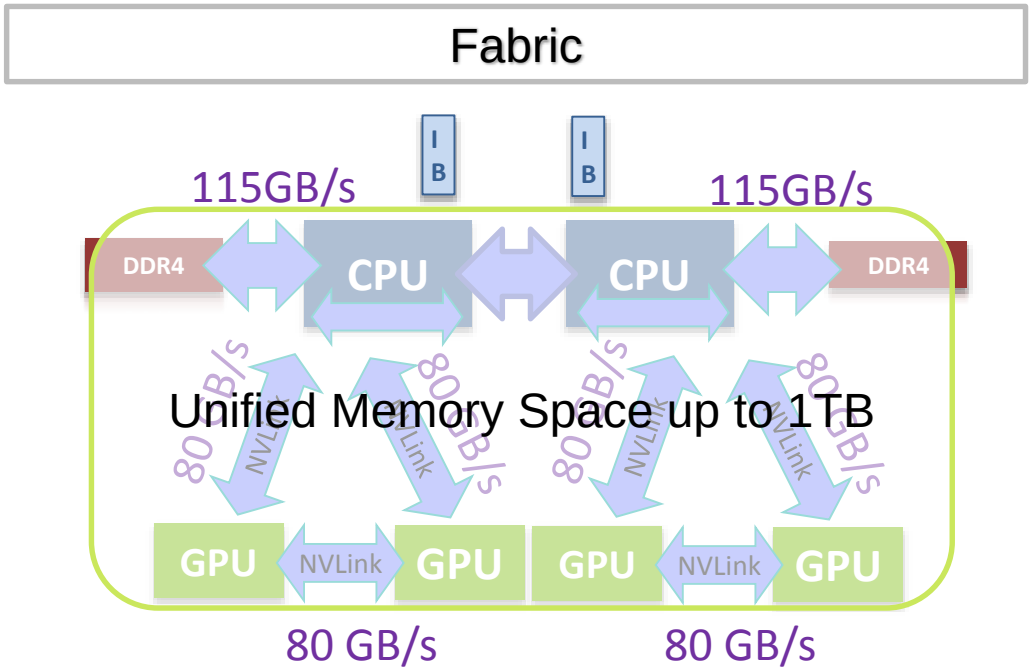
Design: Flat and Fat

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



Приложения в процессе тестирования их производительности на POWER8 с NVLink и NVIDIA P100

Watson Concept Insights	GPUdb	Apache Spark	STAC-A2
LatticeQCD	CPMD	OpenFOAM	SOAP3-dp
NAMD	Caffe	Nekbone	LULESH
AMG	FFT	HPL	HPCG

Спасибо за внимание!

Вопросы?

Алексей Перевозчиков
82189117@ru.ibm.com