POWER8 для задач HPC

Алексей Перевозчиков
IBM Server Solutions Product Manager

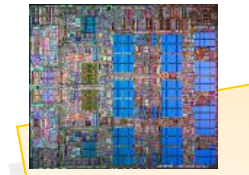
82189117@ru.ibm.com

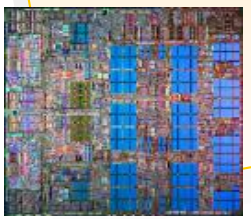
Дубна, 26 мая 2015г.

Power Systems

Планы развития процессора POWER

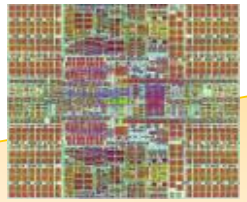
IBM

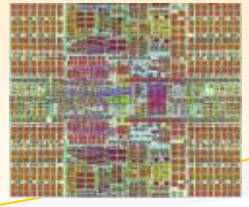




POWER5/5+
130/90 nm

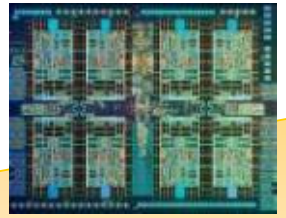
- ✓Dual Core
- ✓Enhanced Scaling
- ✓SMT
- ✓Distributed Switch +
- ✓Core Parallelism +
- ✓FP Performance +
- ✓Memory Bandwidth +
- ✓Virtualization

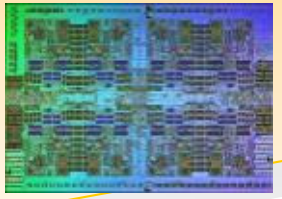




POWER6/6+
65/65 nm

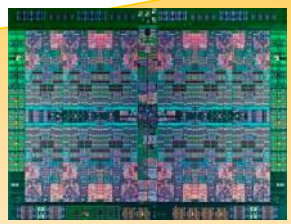
- ✓Dual Core
- ✓High Frequencies
- ✓Virtualization +
- ✓Memory Subsystem +
- ✓Altivec
- ✓Instruction Retry
- ✓Dynamic Energy Mgmt
- ✓SMT +
- ✓Protection Keys





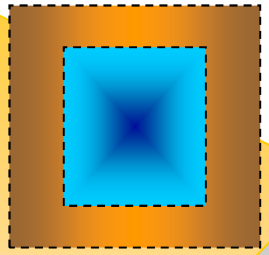
POWER7/7+
45/32 nm

- ✓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

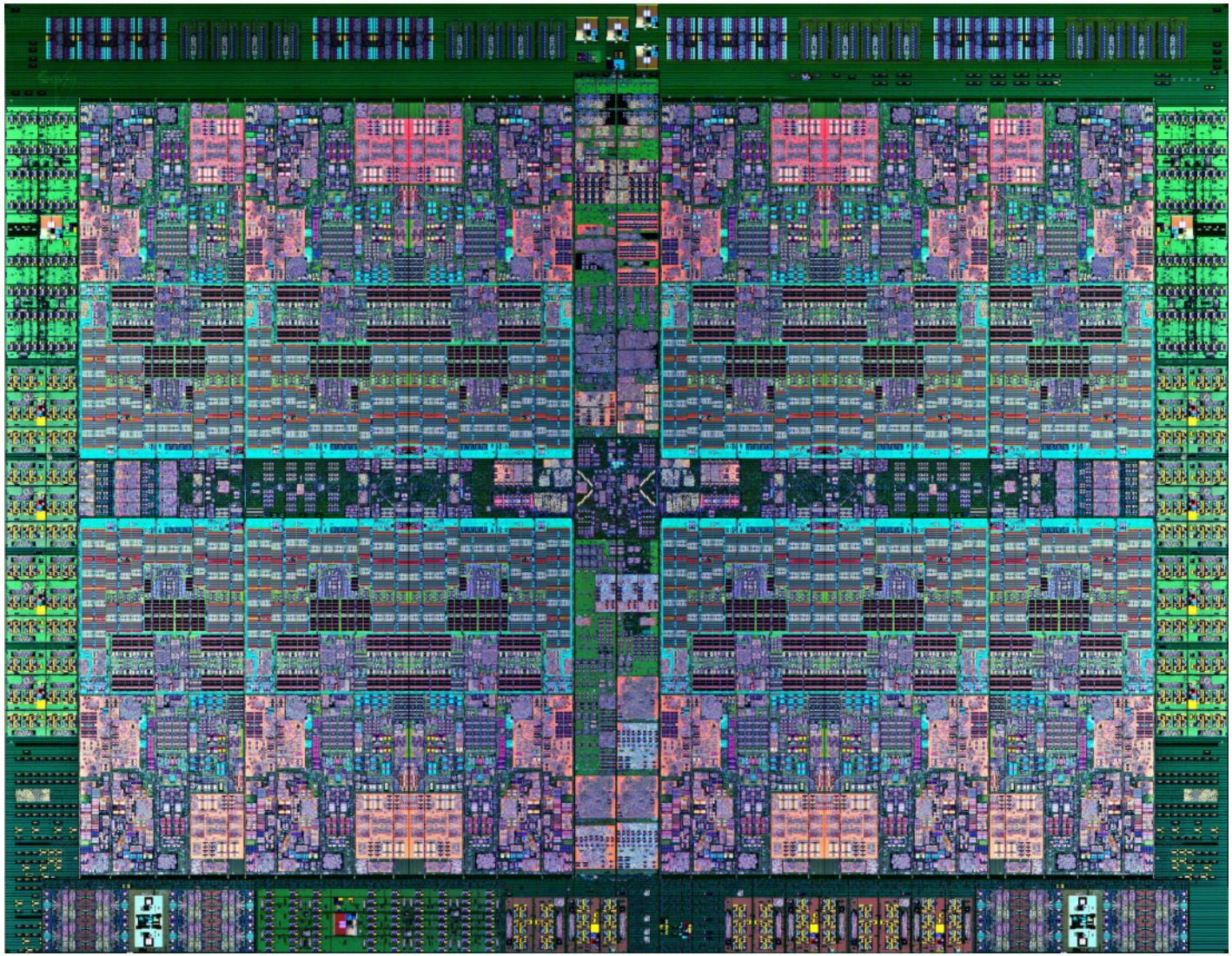
- ✓Extreme Analytics Optimization
- ✓Extreme Big Data Optimization
- ✓On-chip accelerators

2004 2007 2010 2014

Power Systems

Процессор POWER8

IBM



Power Systems

Процессор POWER8

IBM

Технология

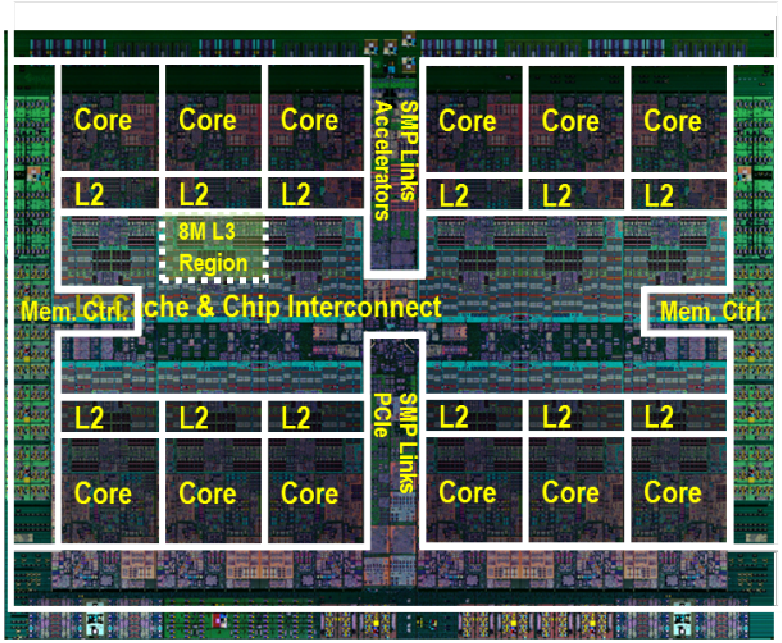
22nm SOI, eDRAM, 650mm2, 4.2B transistors

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**

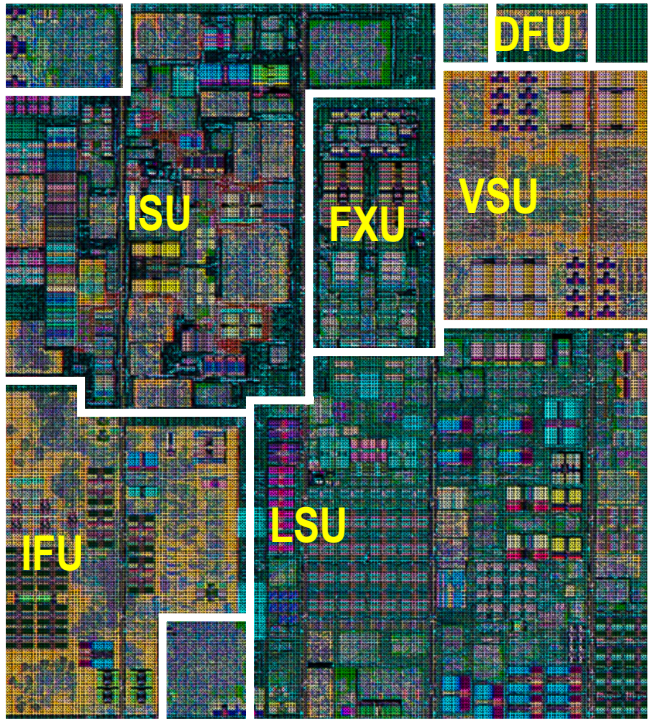
Шинные интерфейсы

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

POWER8 ядро

Execution Improvement vs. 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



Larger Caching Structures vs. POWER7

- 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

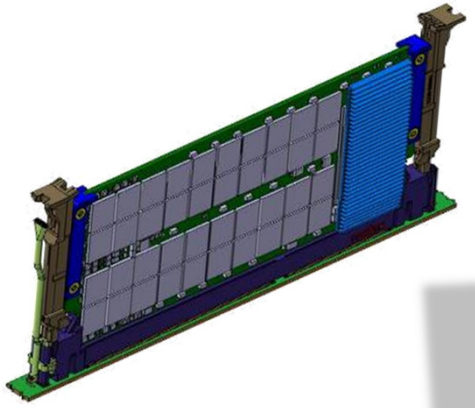
Enhanced Prefetch

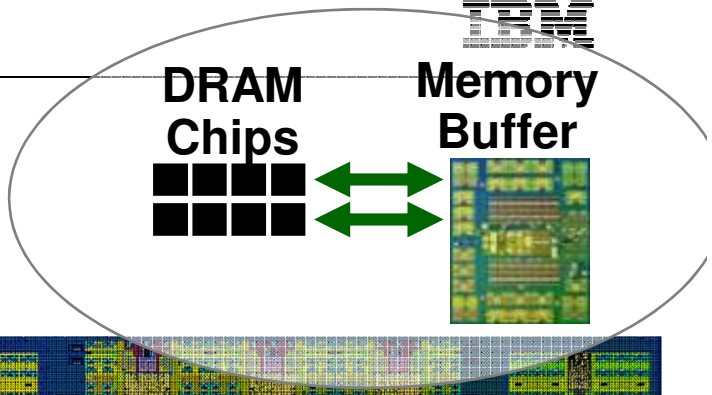
- **Instruction speculation awareness**
- Data prefetch depth awareness
- Adaptive bandwidth awareness
- Topology awareness

Memory Buffer Chip
…*with 16MB Cache…*

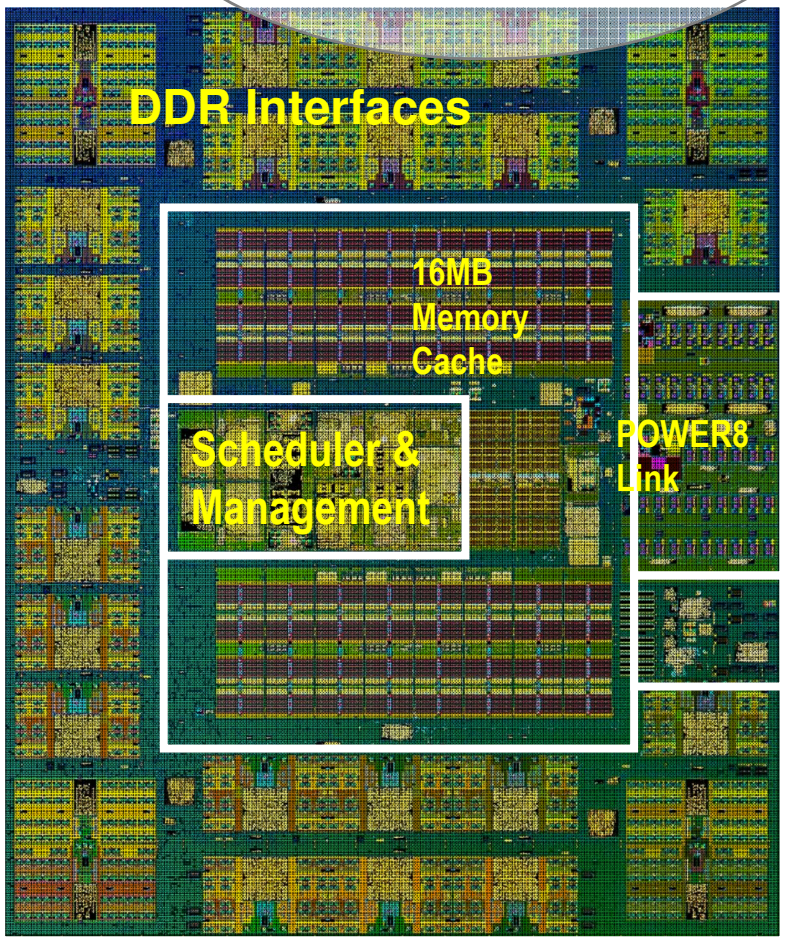
Power Systems

IBM





“L4 cache”



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

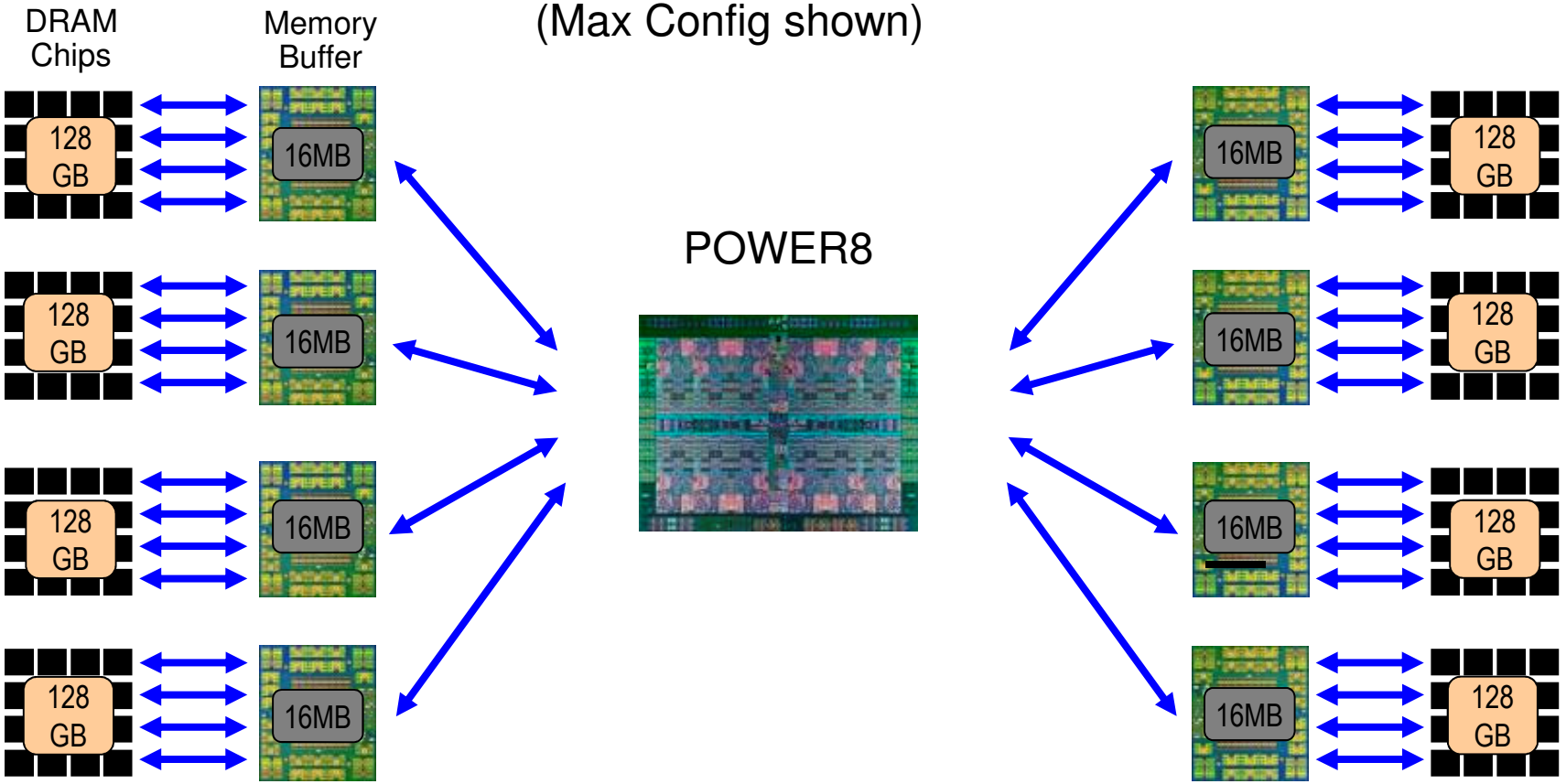
Оптимизированный интерфейс
- 9.6 GB/s high speed interface
- Интеллектуальная надежность
- Изоляция сбоев на лету

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

Power Systems

POWER8 Memory Organization

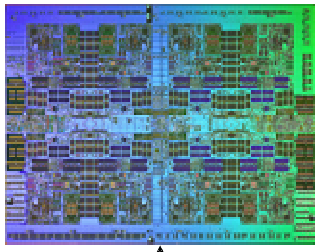
IBM



- ***Up to 8 high speed channels, each running up to 9.6 Gb/s for up to 230 GB/s sustained***
- ***Up to 32 total DDR ports yielding 410 GB/s peak at the DRAM***
- ***Up to 1 TB memory capacity per fully configured processor socket***

POWER8 Integrated PCI Gen 3

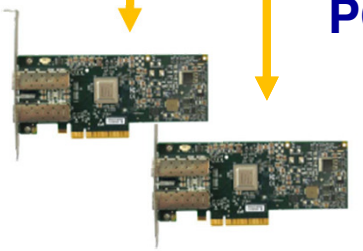
POWER7



GX Bus

I/O Bridge

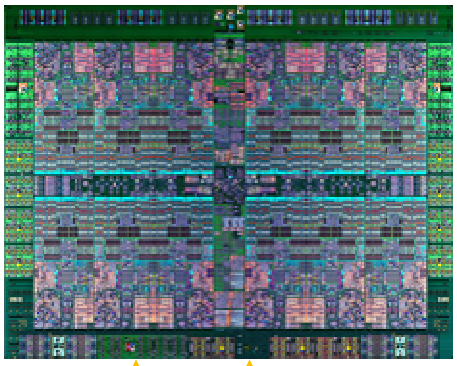
PCIe Gen2



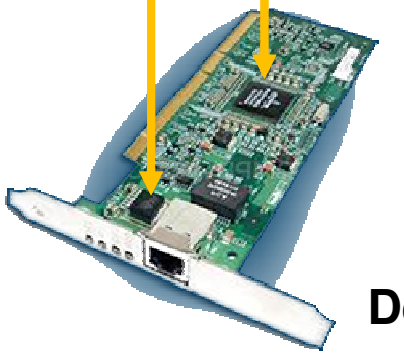
PCI Devices

Встроенная в кристалл поддержка PCIe Gen 3
-Замена GX/Bridge
-Минимальная задержка
-Gen3 x16 bandwidth (16 Gb/s)

POWER8

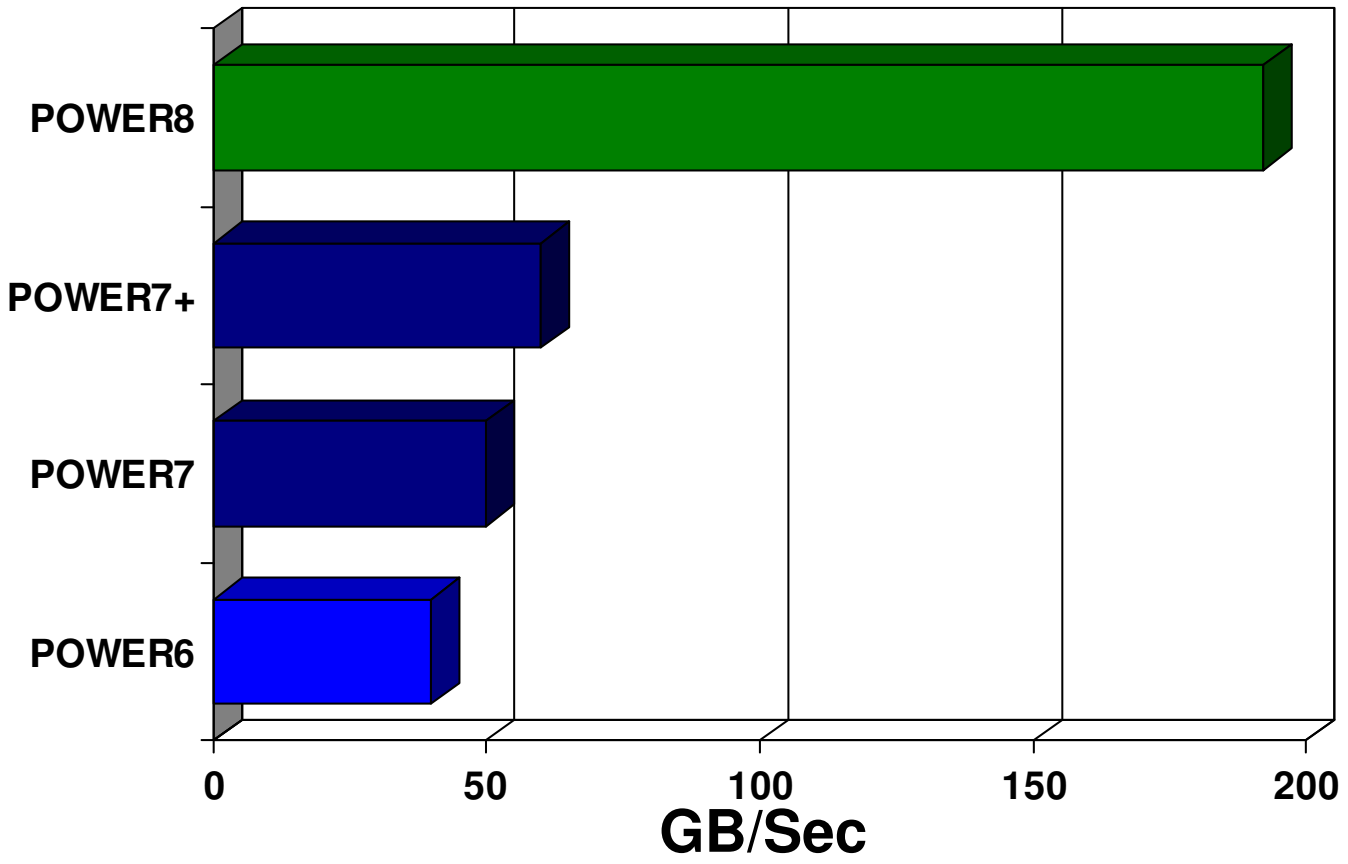


PCIe Gen3



PCI Device

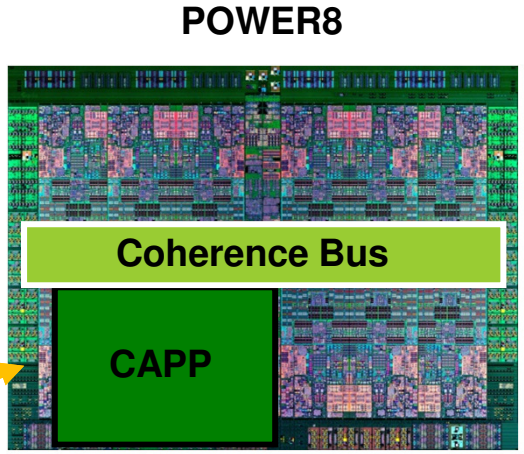
IO Bandwidth

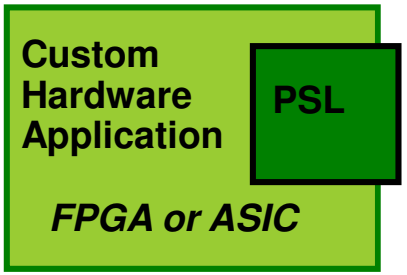


CAPI *(Coherent Accelerator Processor Interface)*

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.

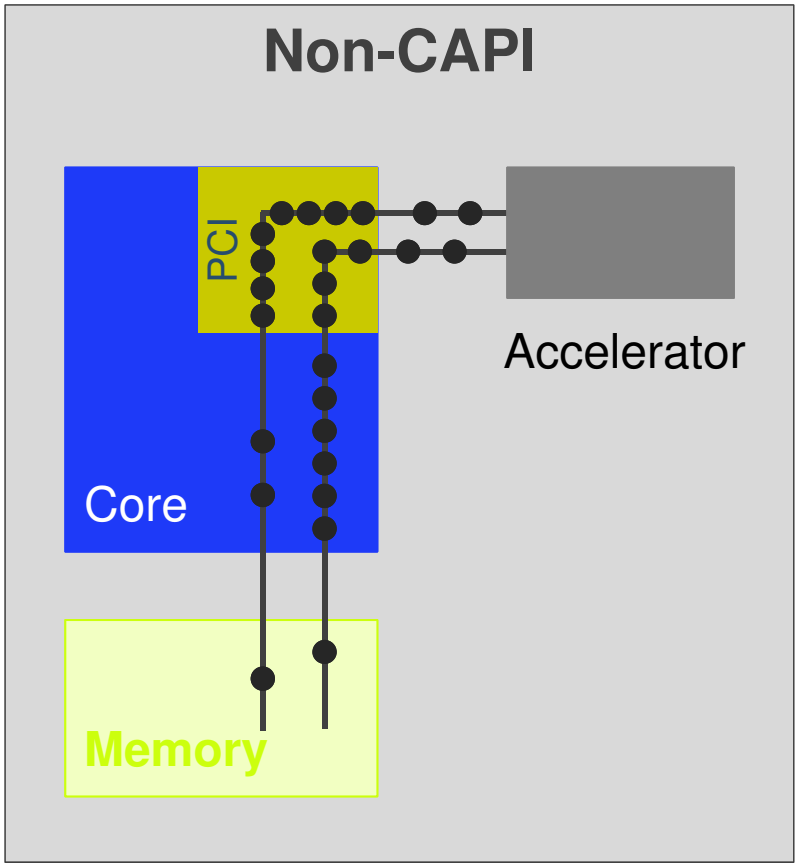


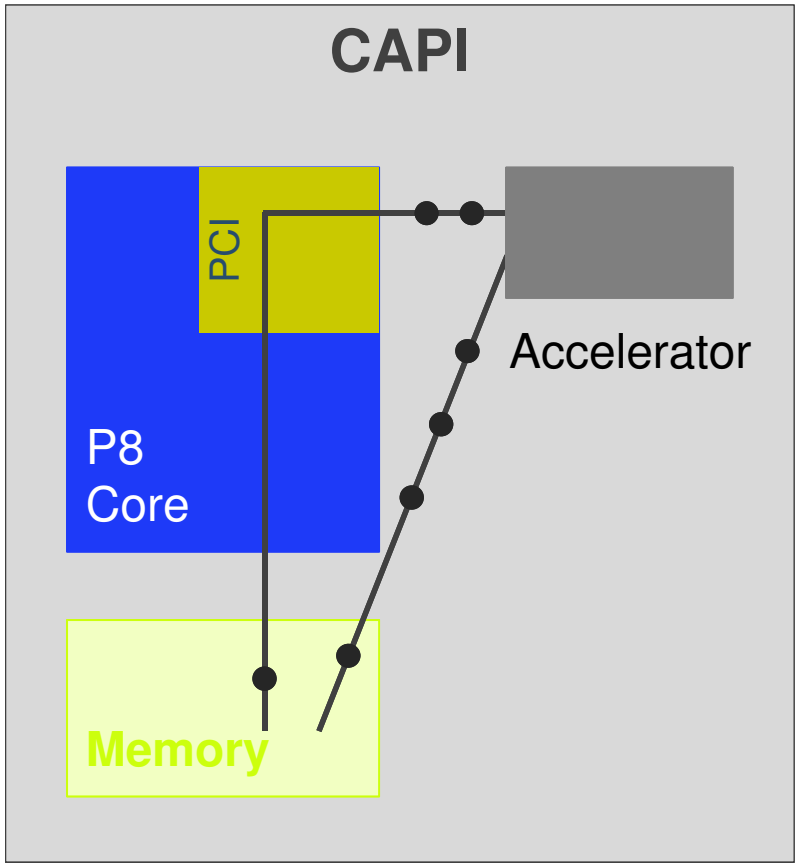


PCIe Gen 3
Transport for encapsulated messages

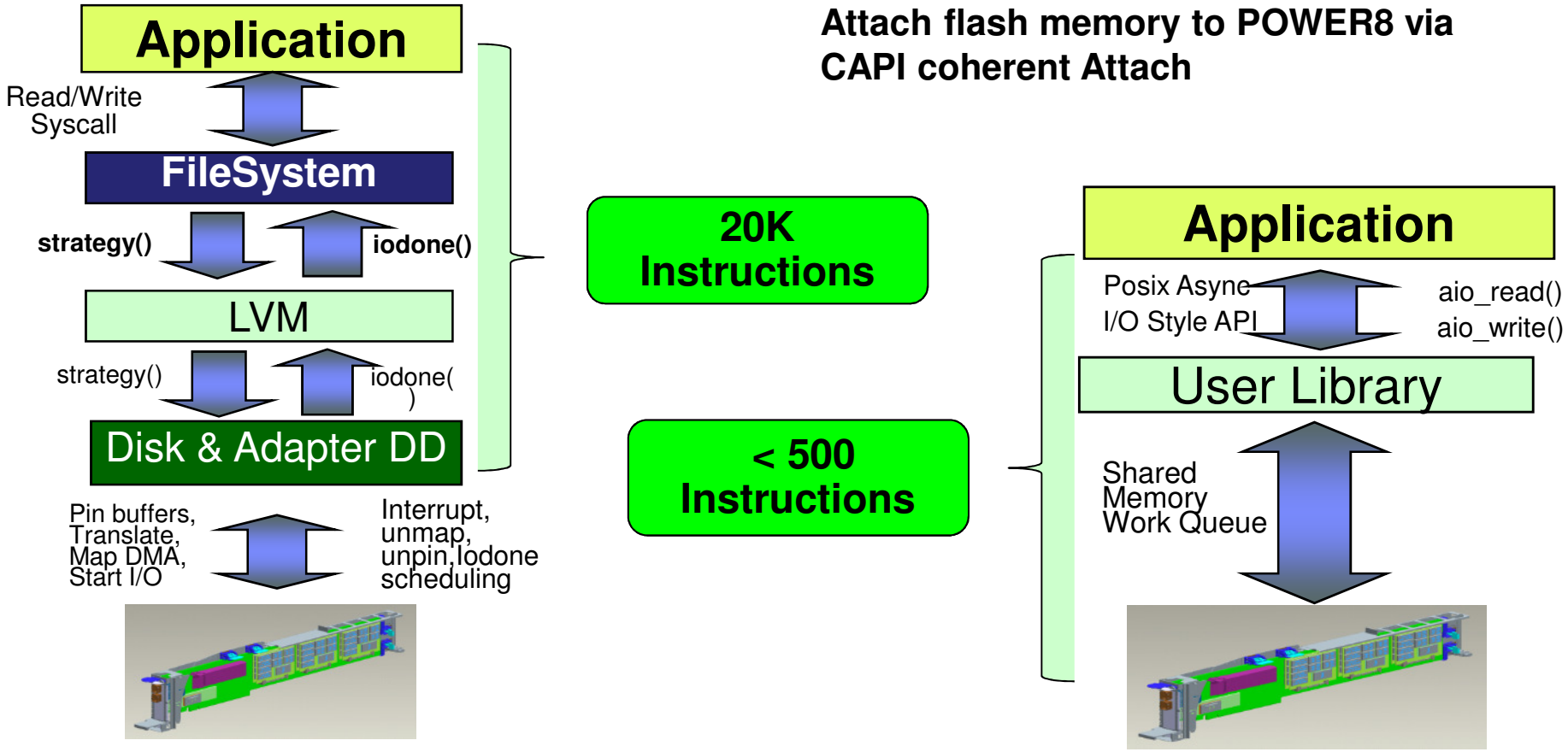
Специализированные контроллеры
Программные ускорители

Coherent Accelerator Processor Interface





CAPI: CAPI Attached Flash Optimization



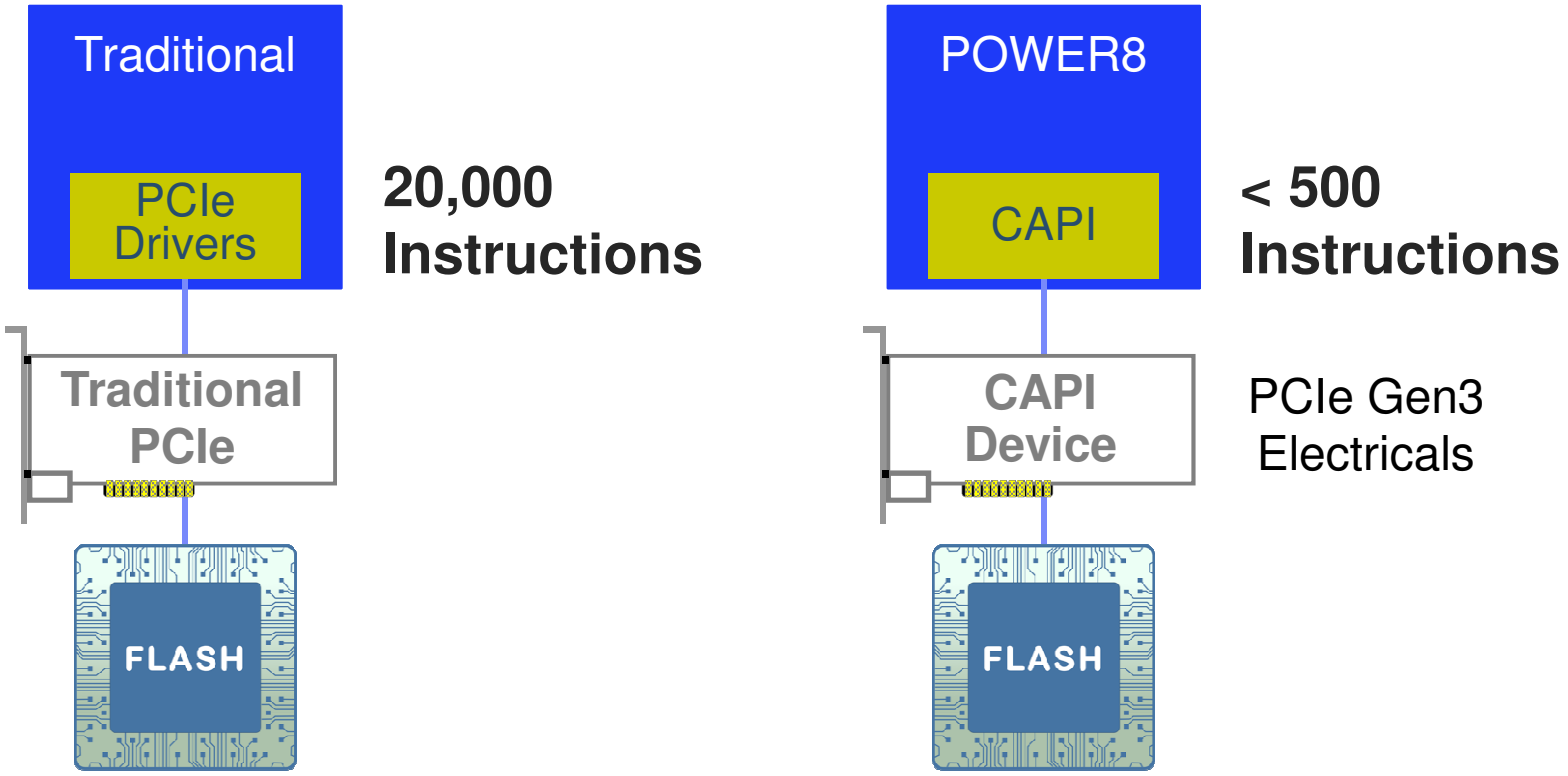
Attach flash memory to POWER8 via CAPI coherent Attach

Приложение непосредственно выдает инструкции Read/Write.

Уменьшение количества инструкций до 97%. (CAPI Flash controller Operates in User Space)

Экономия 10 ядер на каждый 1млн. IOPs

CAPI Can Lower Flash Latency



Несколько слов о стратегии

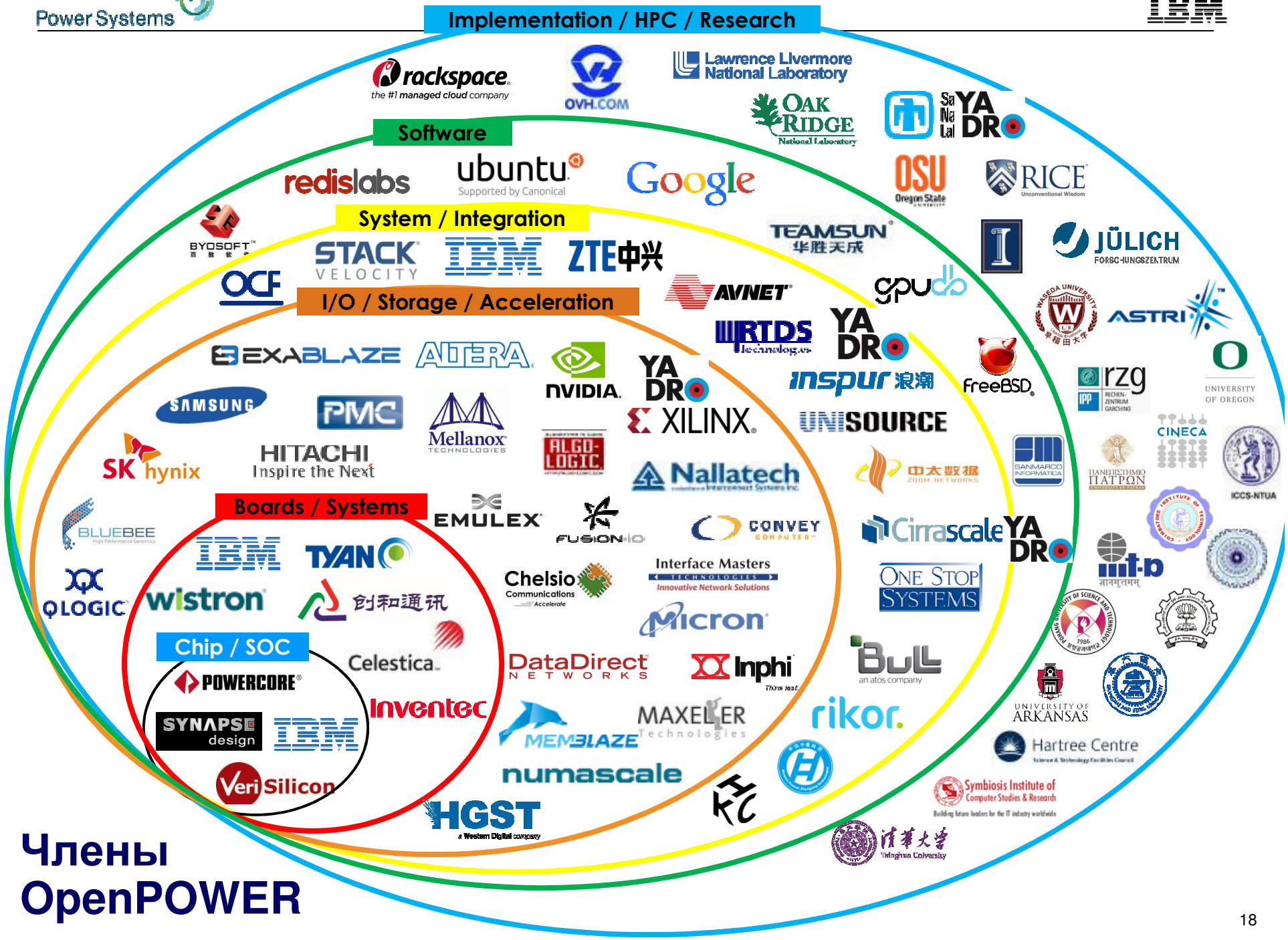
Развитие стратегии аппаратных средств для HPC

- Общий дизайн платформы для высокопроизводительных вычислений и высокопроизводительной аналитики
- Углубление отношений с технологическими партнёрами
- Серверы будут 2 и 4 сокета
  - Коммерческие системы будут масштабироваться до (close-coupling/NUMA) 4-х блоков
- 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 упрощена

- 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

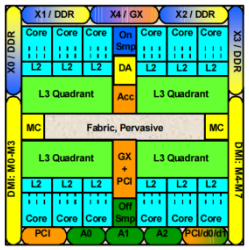
The OpenPOWER Foundation



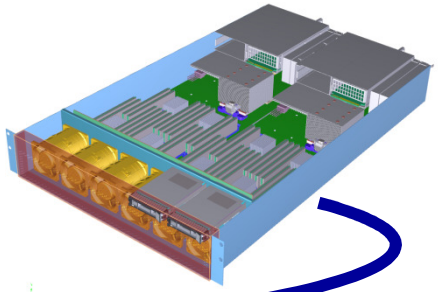
Члены OpenPOWER

Representative large DOE configuration 2015-2018

P8:
10 Cores (of 12)
8Threads/core
~ .27 DP TFlop
~ 3.4 GHz



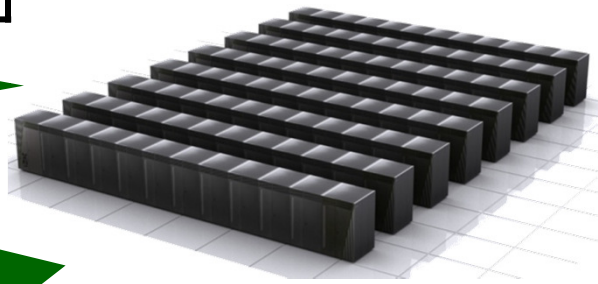
POWER8 2 Socket Server
2 P8 + 2 Kepler Duo GPU (@2.74 TF/s Boost)
256 GiB SMP Memory (8 GB DDR3 RDIMMs)
48 GiB GPU Memory (HBM stacks)



Cluster Network:
Mellanox IB4X EDR Switch

Racks	System
Compute	306
Storage	32

System



- Scalable system software and data architecture
- Technical compilers
- Water cooling with RDHX

Compute Rack:
20 Servers/rack
86 TFlop/rack
6.1 TB/rack
32.4 kWatts

Storage Drawers

- **Compute racks: 26.2PF, 1.86 PB**
- **Storage racks: 72 PB**
- **Management & Gateway racks**
- **10.3 MW**