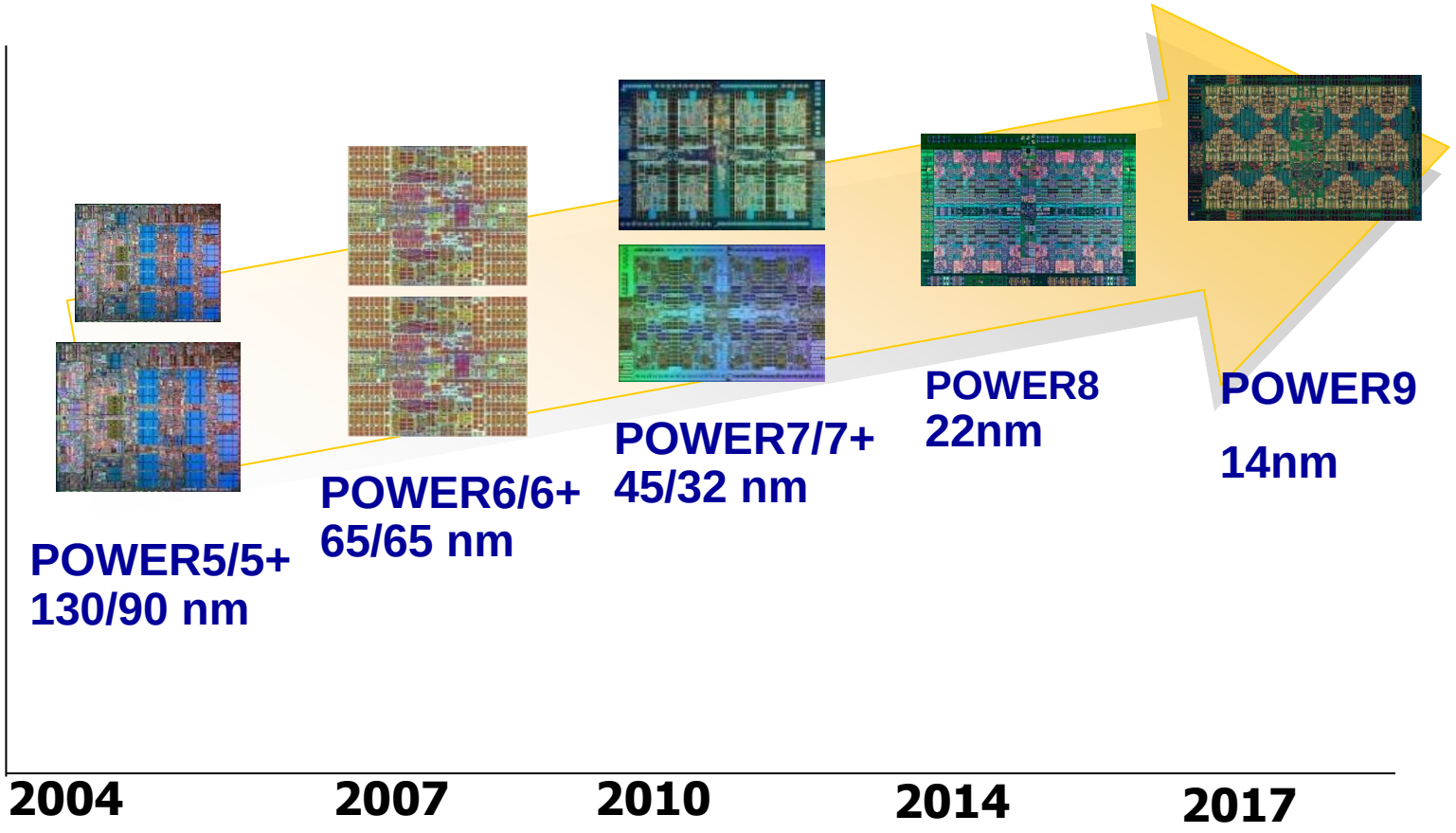
IBM Power Systems: бескомпромиссная надежность и производительность

Докладчик: Андрей Калита
Дата:15.06.2017

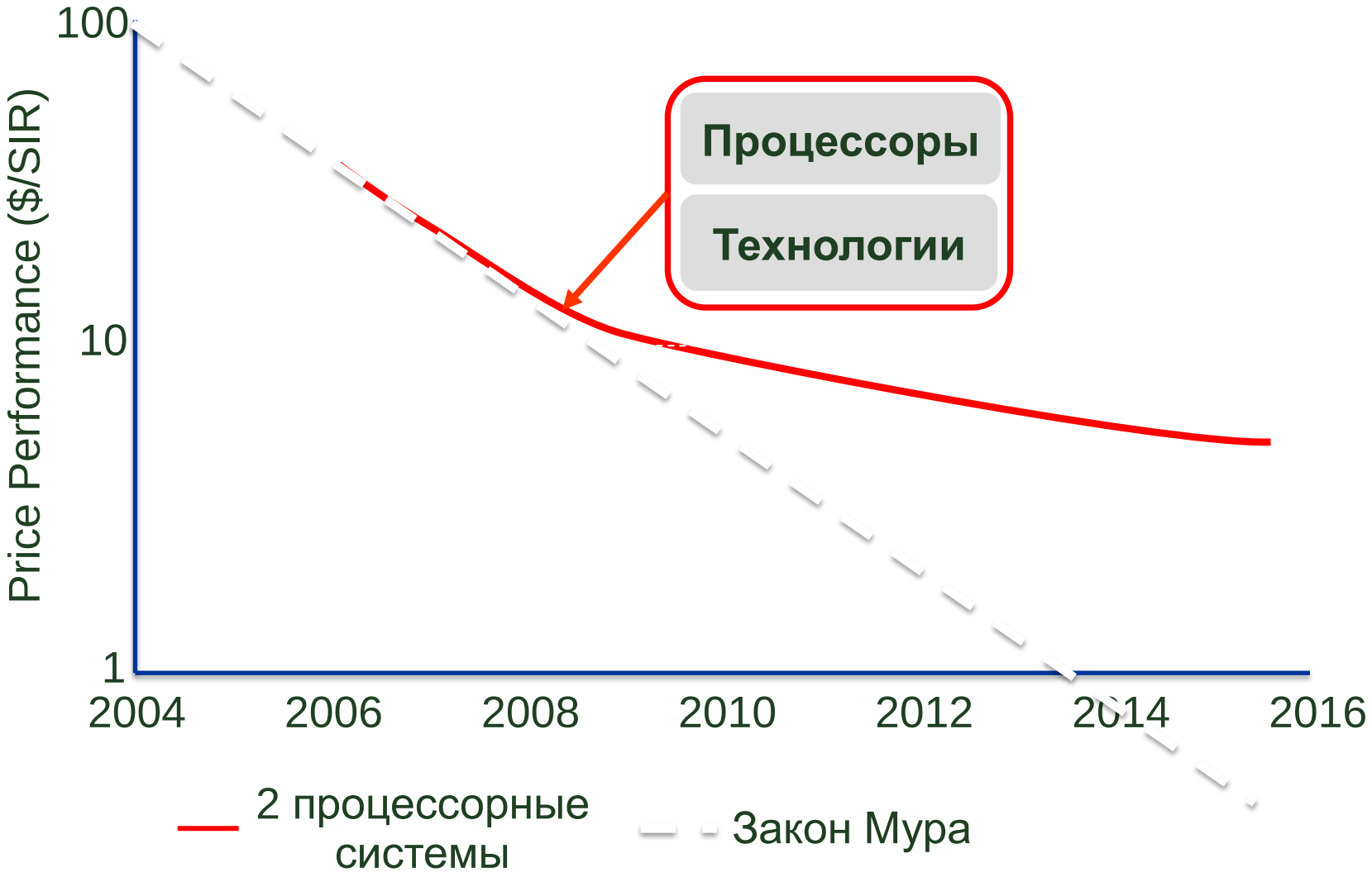
IT.Integrator

Процессор IBM Power 8

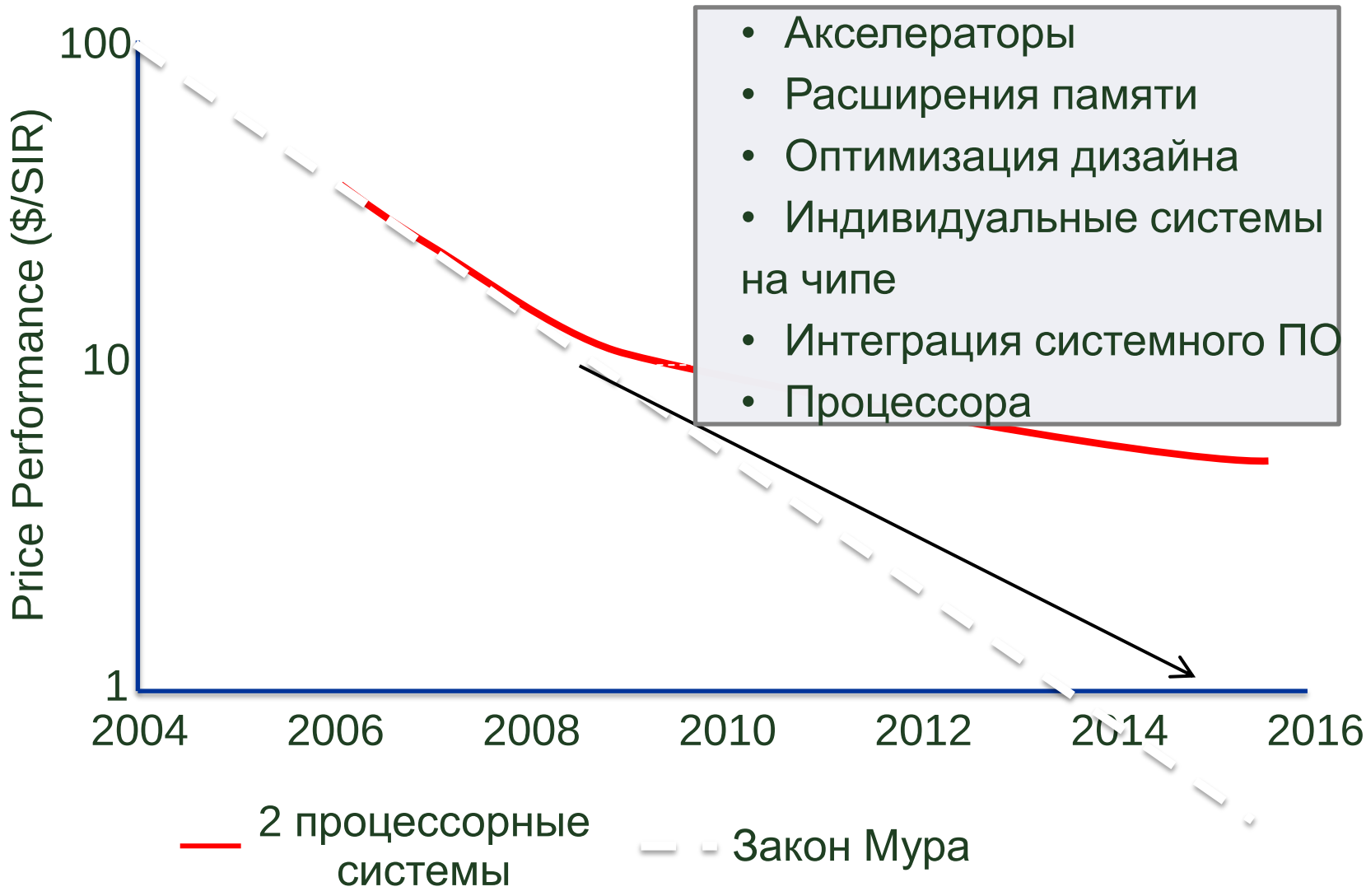
Процессор IBM Power



Закон Мура


100
10
1
Price Performance ($/SIR)
2004
2006
2008
2010
2012
2014
2016
Процессоры
Технологии
2 процессорные системы
Закон Мура

Закон Мура

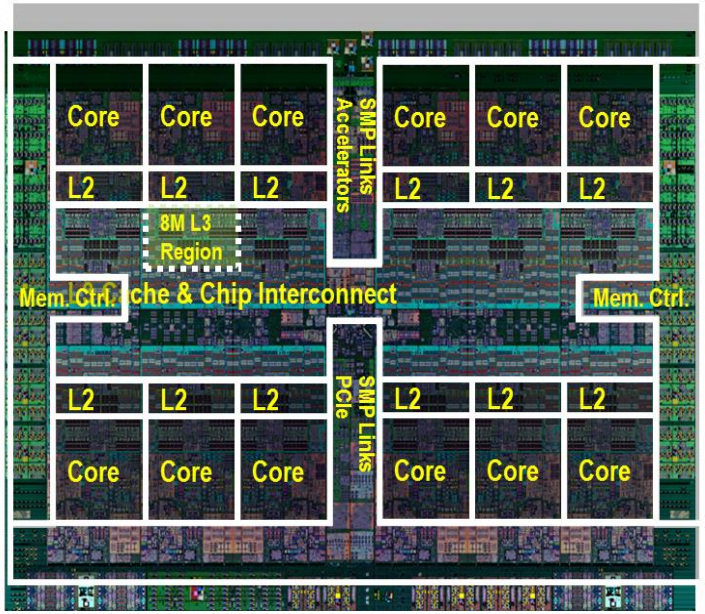

100
10
1
Price Performance ($/SIR)
2004
2006
2008
2010
2012
2014
2016
• Акселераторы
• Расширения памяти
• Оптимизация дизайна
• Индивидуальные системы на чипе
• Интеграция системного ПО
• Процессора
2 процессорные системы
Закон Мура

Процессор IBM Power 8

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

Cores
- 12 cores (SMT8)
- 8 dispatch, 10 issue, 16 exec pipe
- 2X internal data flows/queues
- Enhanced prefetching
- 64K data cache, 32K instruction cache

Accelerators
- Crypto & memory expansion
- Transactional Memory
- VMM assist
- Data Move / VM Mobility



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

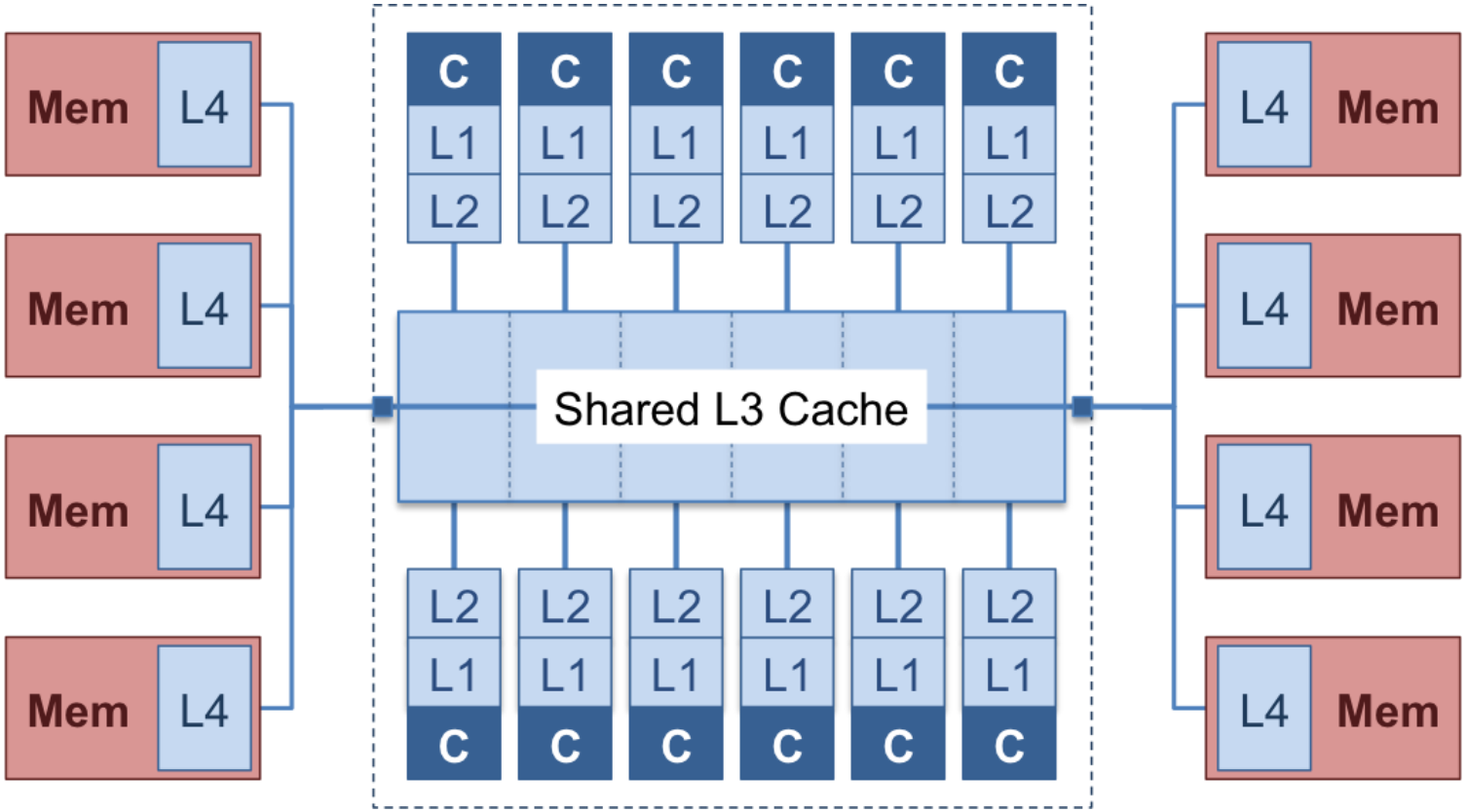
Larger Caches
- 512 KB SRAM L2 / core
- 96 MB eDRAM shared L3
- Up to 128 MB eDRAM L4 (off-chip)

Memory
- Up to 230 GB/s sustained bandwidth

Bus Interfaces
- Durable open memory attach interface
- Integrated PCIe Gen3
- SMP Interconnect
- CAPI (Coherent Accelerator Processor Interface)

Производительность

Кэш процессора Power 8

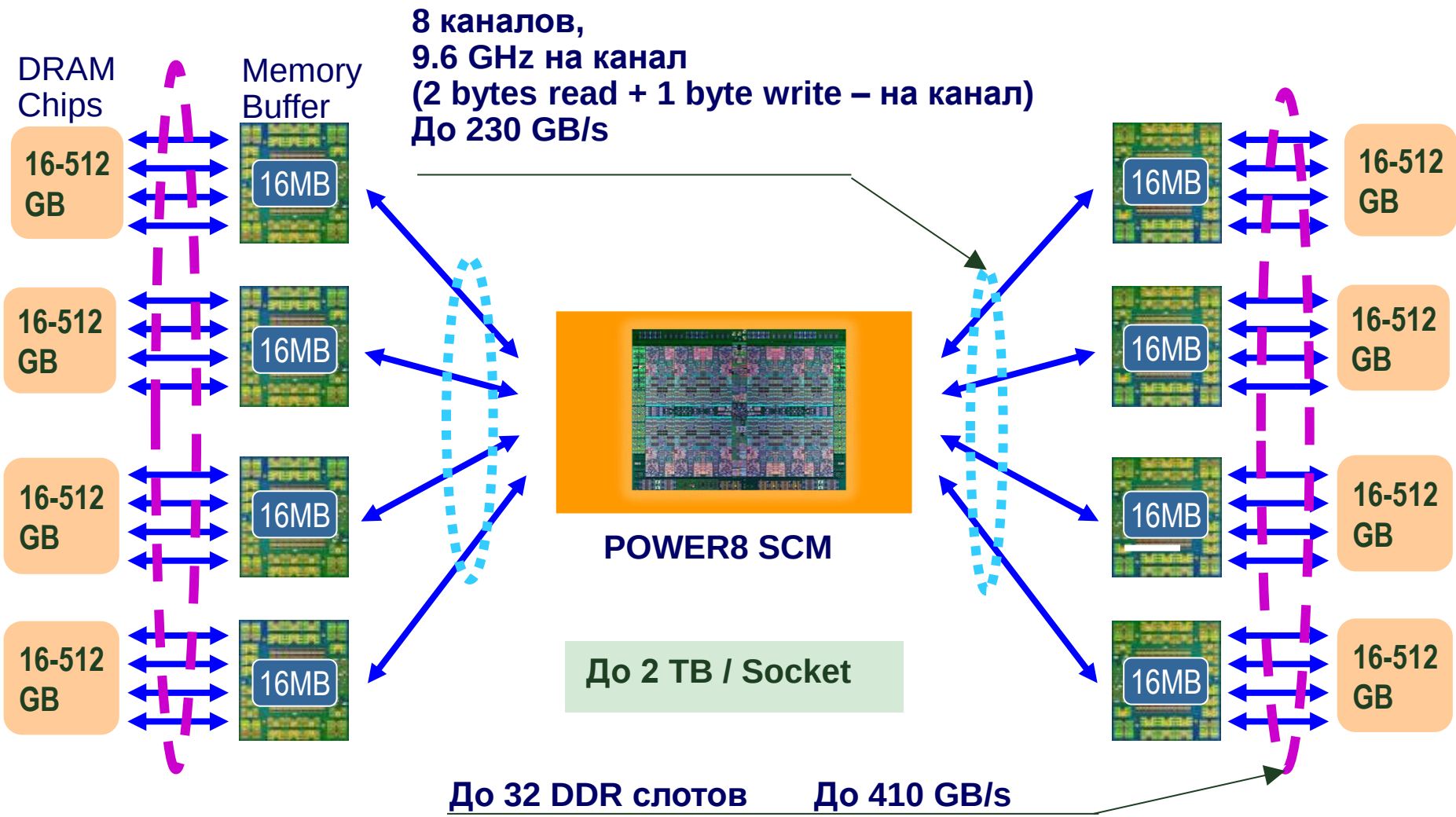


- Большие нагрузки
- Многопоточный

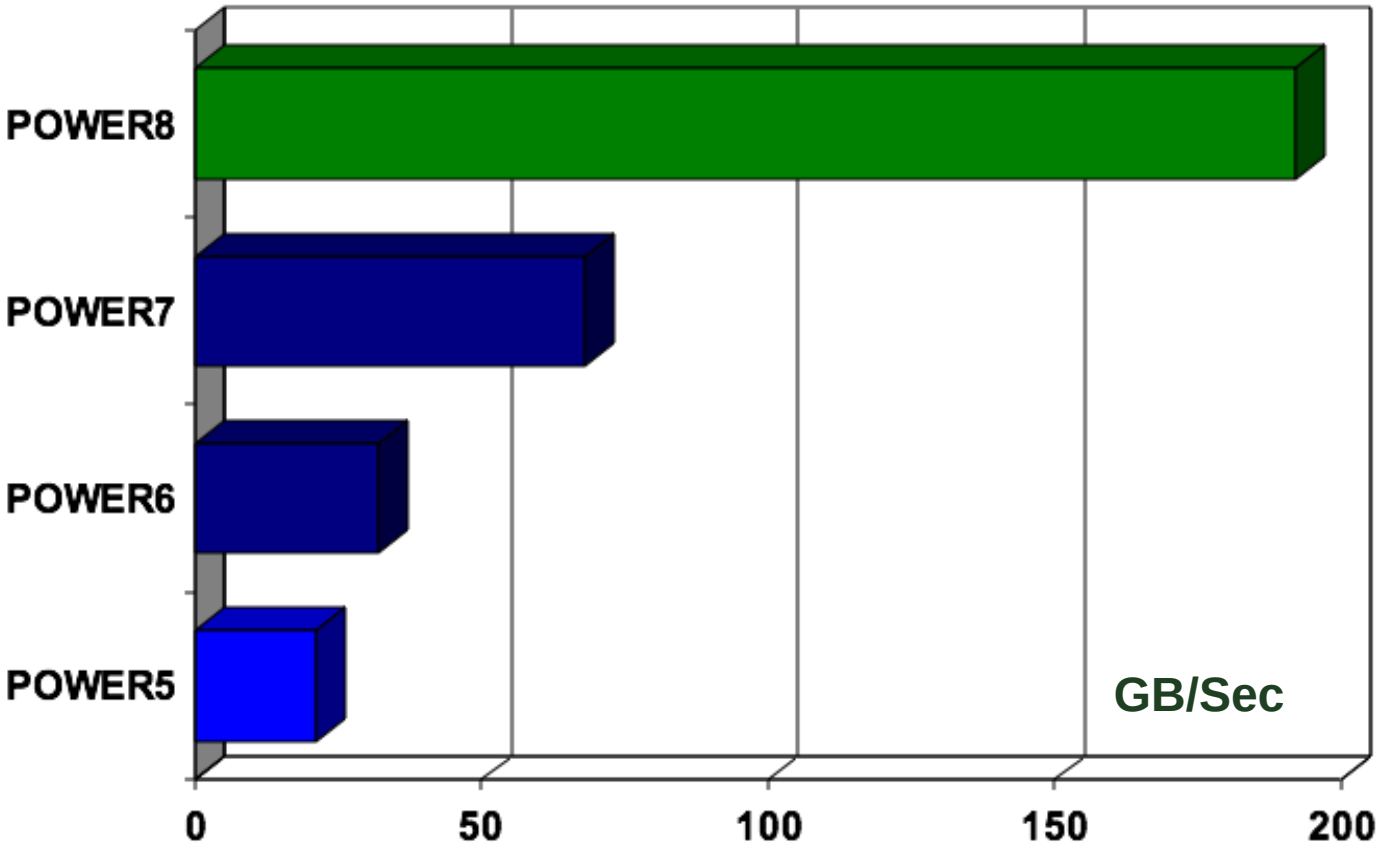
- Виртуализация
- Общие данные

- Запись пакетов данных
- На 55% ниже задержки
- Смешанное чтение и запись

Организация памяти

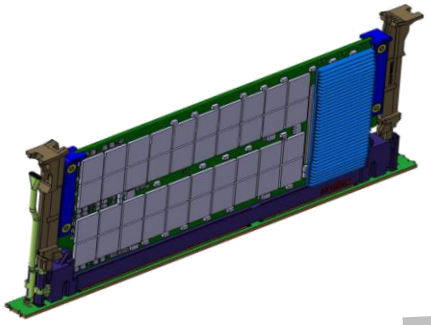


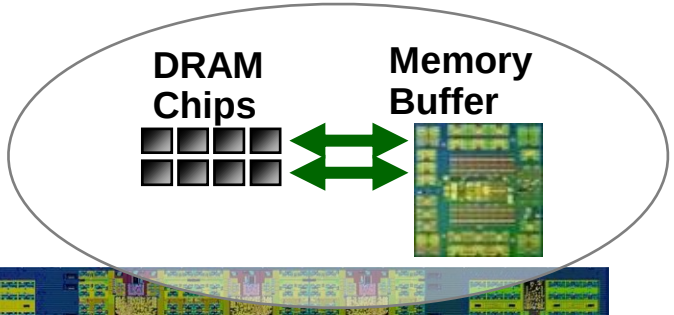
Производительность памяти



- *До 8 каналов памяти со скоростью до 9.6 Гб/сек каждый суммарная скорость до 230 GB/s Объем памяти до 2 TB memory на сокет*

Кэш 4 уровня



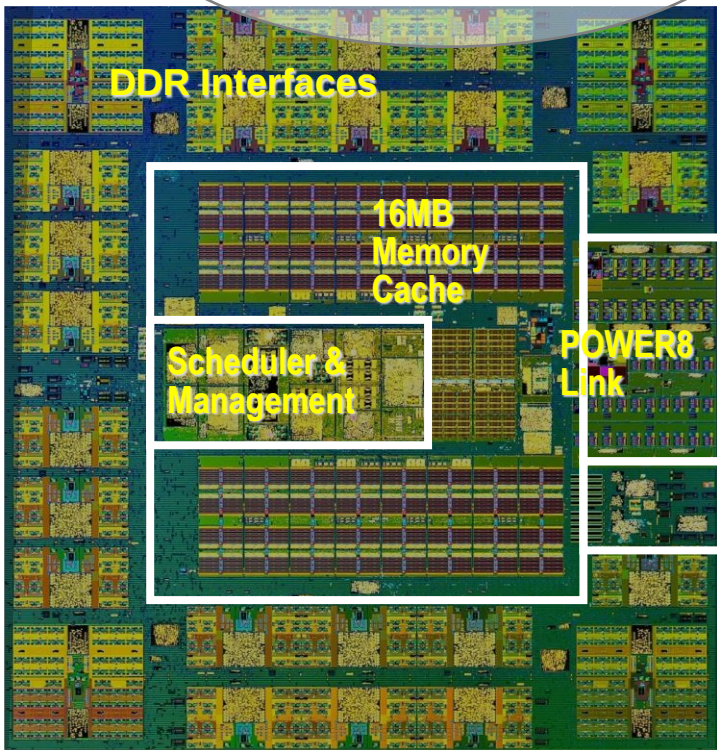


“L4 cache”

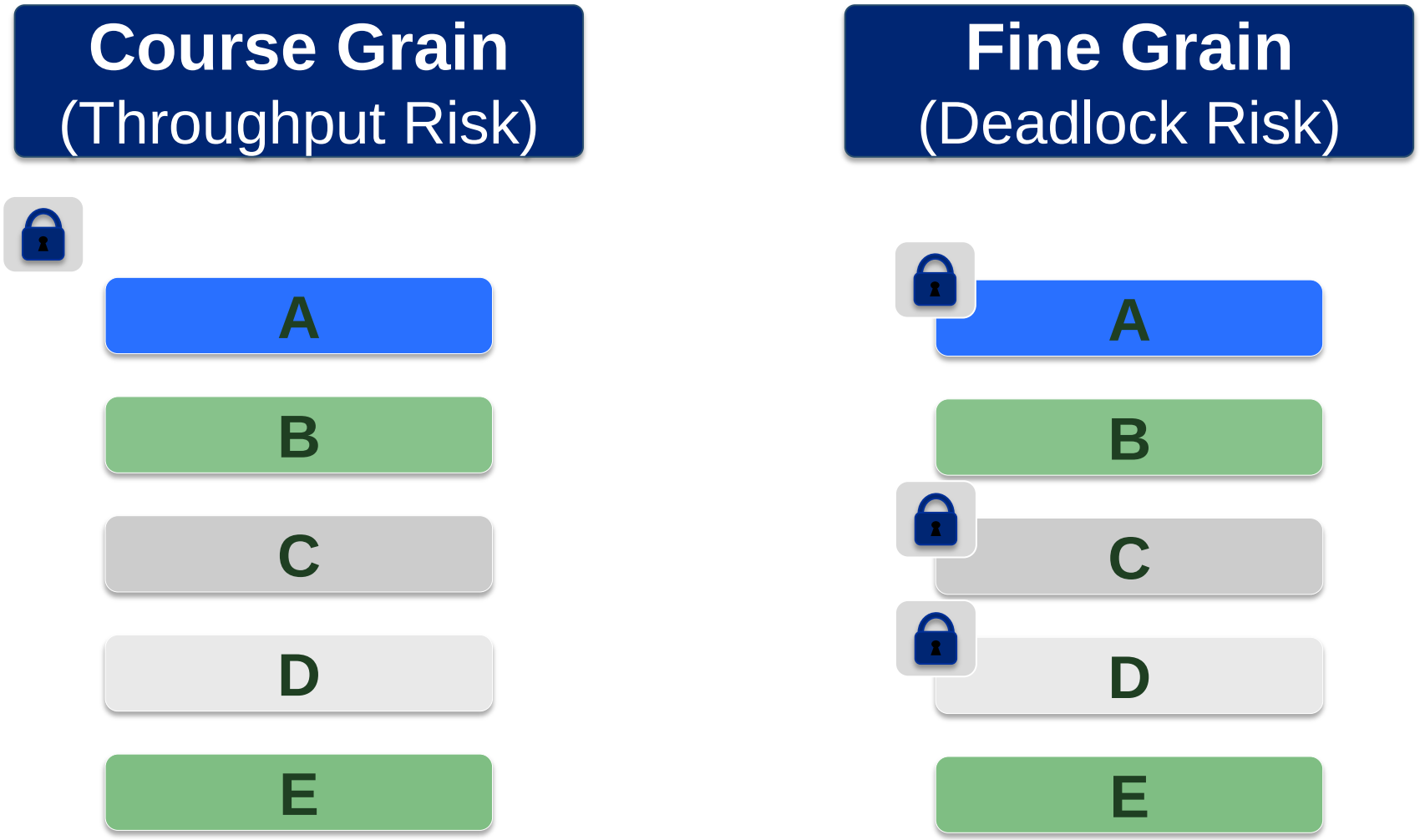
Модули памяти наполняются интеллектом
- Умная система кэширования
- Оптимизация энергии
- Надежность

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

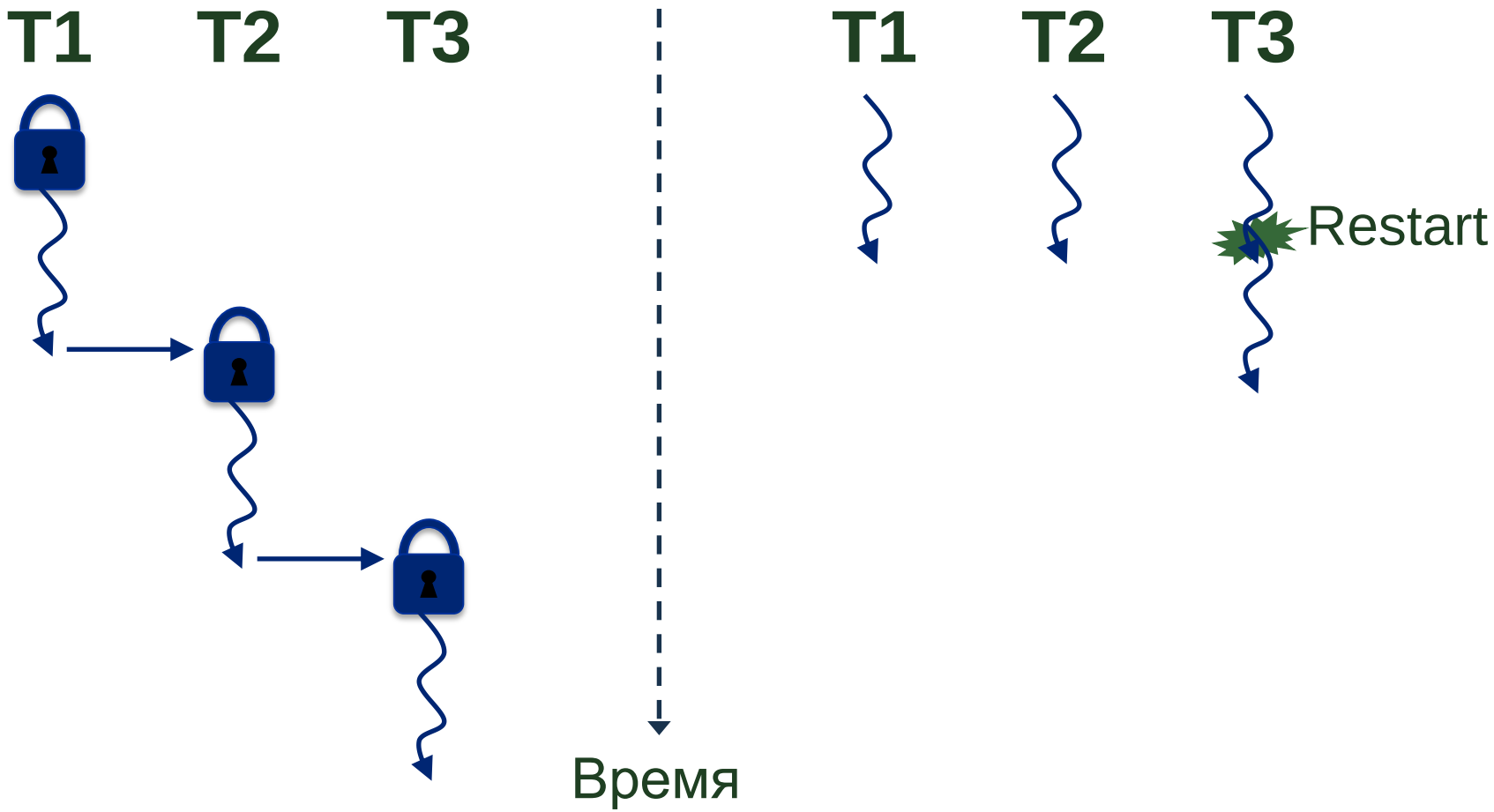
Уникальная производительность
- Уменьшенные задержки
- Cache → latency/bandwidth, partial updates
- Логика предсказания
- 22nm SOI for optimal performance / energy
- 15 metal levels (latency, bandwidth)



Блокировки


Course Grain
(Throughput Risk)
A
B
C
D
E
Fine Grain
(Deadlock Risk)
A
B
C
D
E

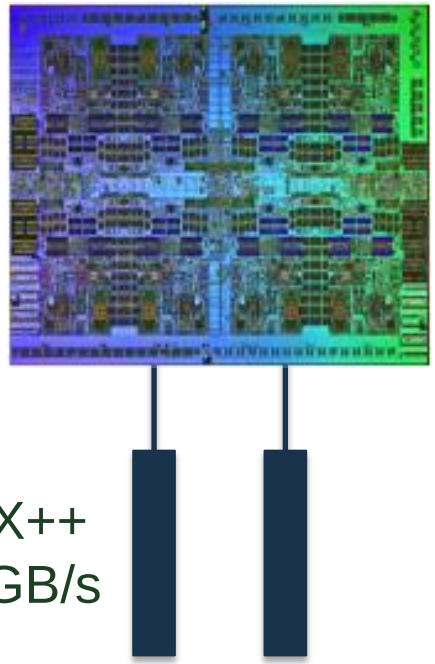
Транзакционная память



Традиционные блокировки **Транзакционная память**

Ввод-вывод

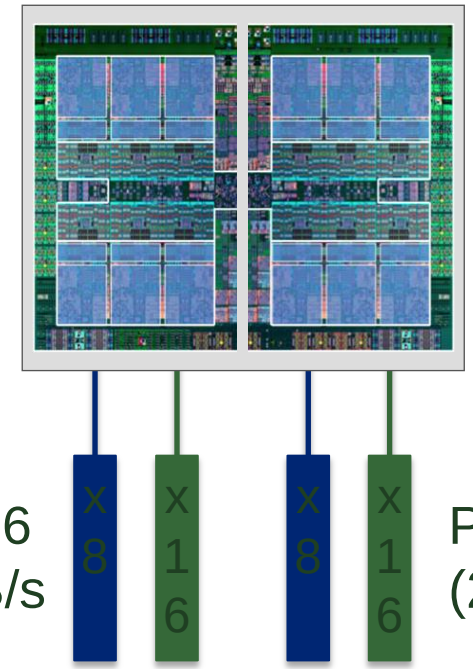
POWER7+



GX++
(2) 20 GB/s

40 GB/s
Peak Bandwidth

POWER8

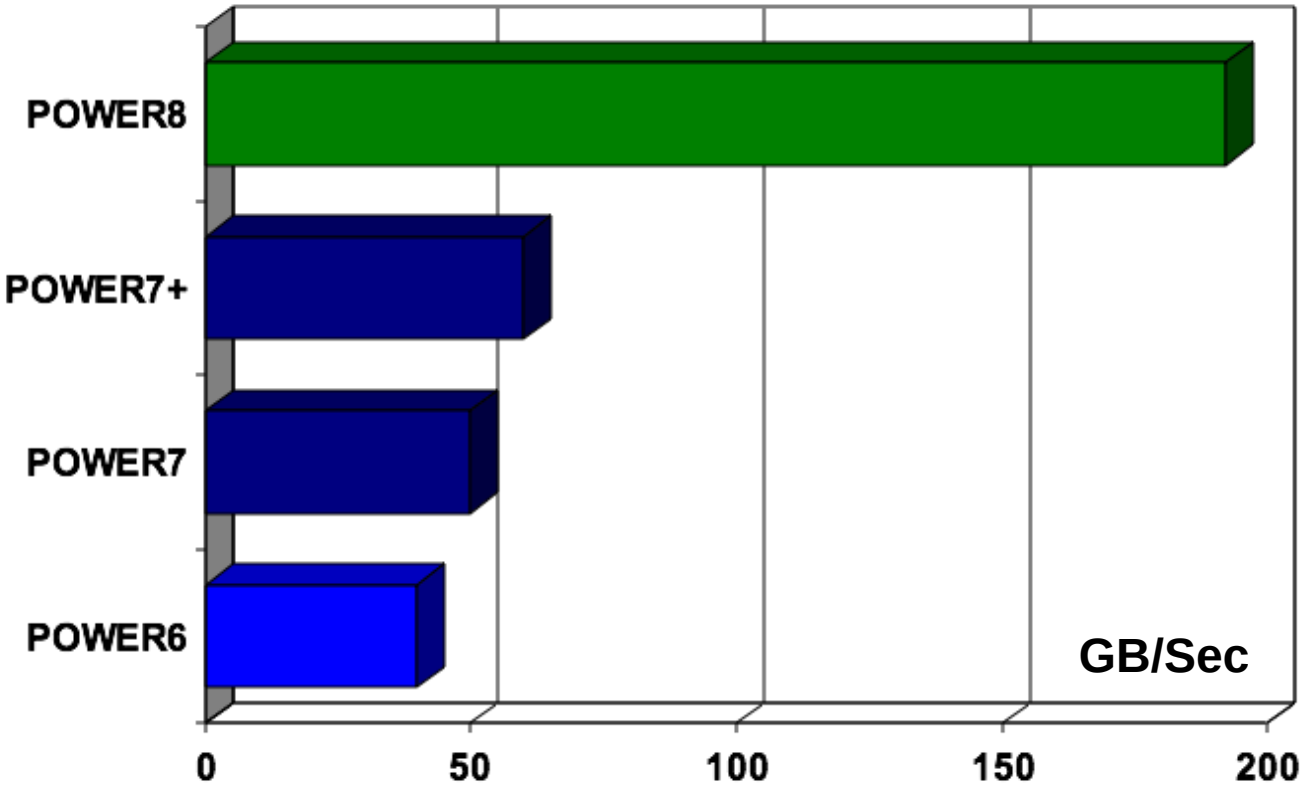


PCIe Gen3 x16
(2) 32 GB/s

PCIe Gen3 x8
(2) 16 GB/s

96 GB/s
Peak Bandwidth

Производительность I/O



- *Суммарная скорость до 192 GB/s для сервера Low End.*

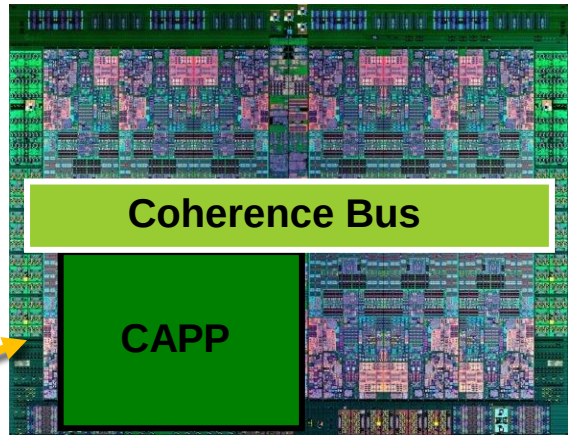
Интерфейс CAPI

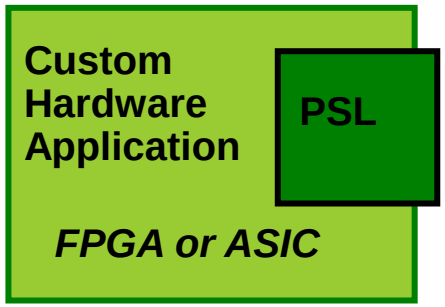
CAPI (Coherent Accelerator Processor Interface)

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

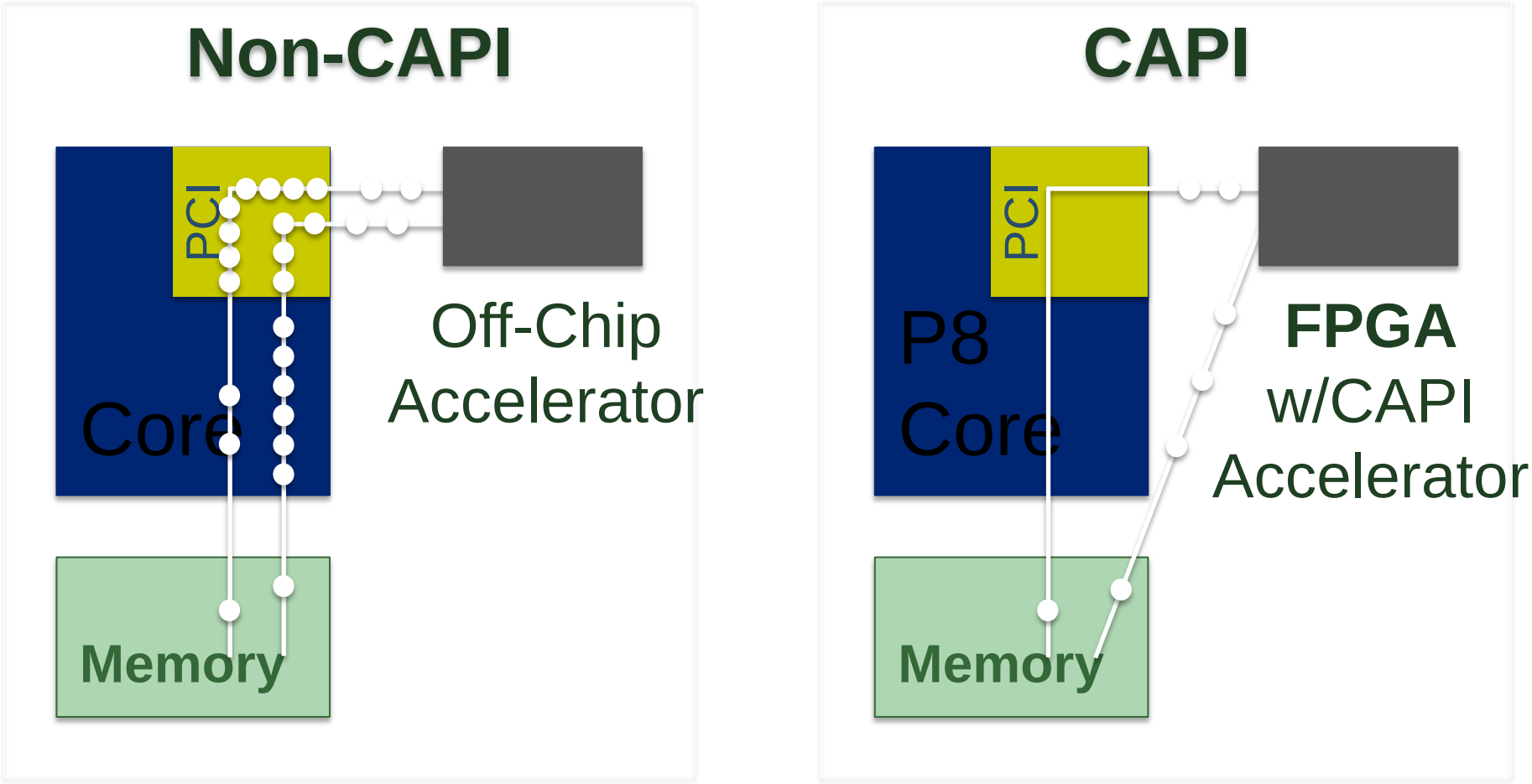




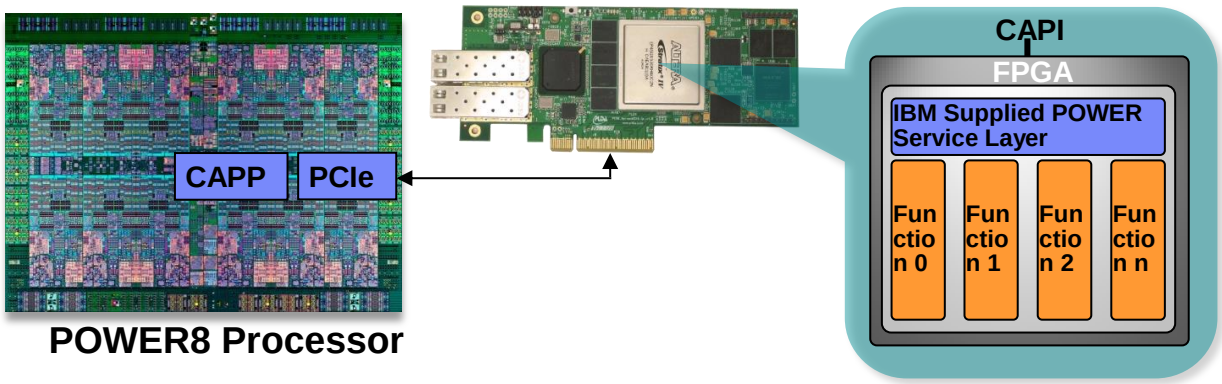
PCIe Gen 3
Transport for encapsulated messages

Специализированные контроллеры
Программные ускорители

Интерфейс CAPI



Интерфейс CAPI



Специализированные контроллеры
Программные ускорители

Typical I/O Model Flow

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

Flow with a Coherent Model

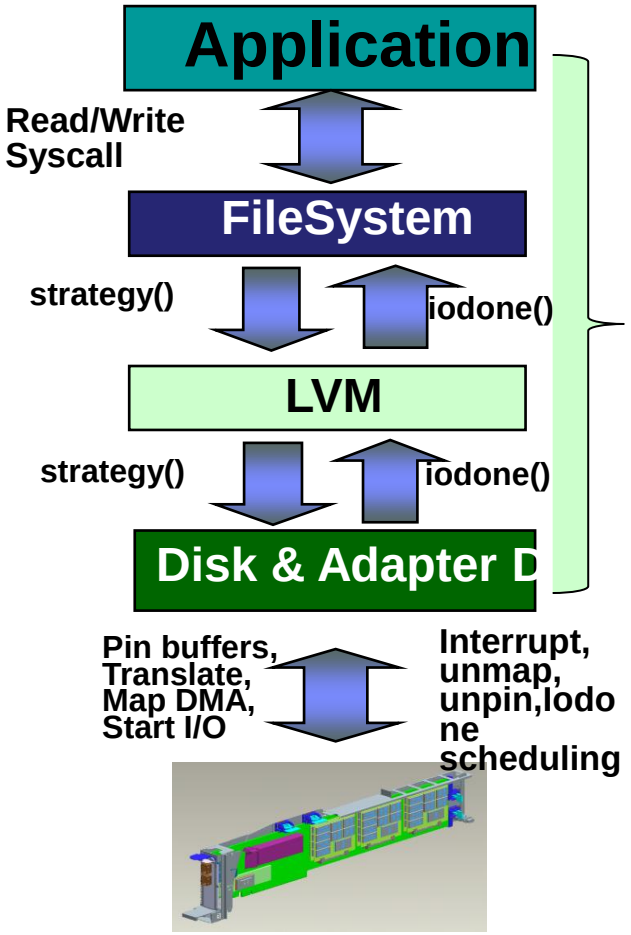
Shared Mem. Notify Accelerat → Acceleration → Shared Memory Completion

CAPI Technology

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.

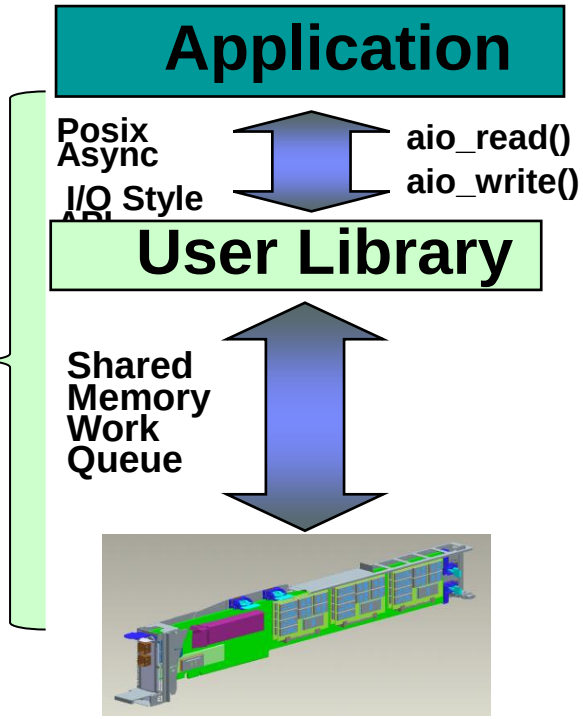
Интерфейс CAPI



20K Instructions

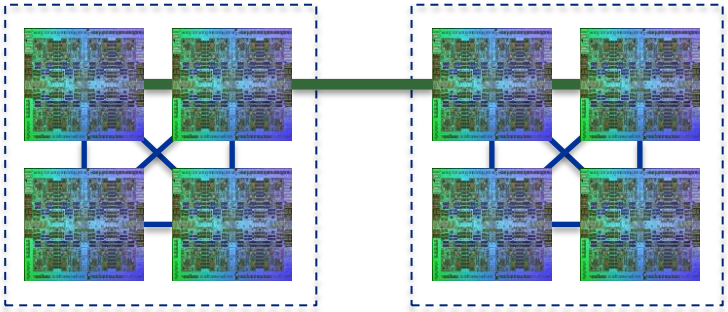
< 500 Instructions

Attach flash memory to POWER8 via CAPI coherent Attach



- **Приложение непосредственно выдает инструкции Read/Write. Уменьшение инструкций до 97%. (CAPI Flash controller Operates in User Space)**
- **Saves 10 Cores per 1M IOPs**

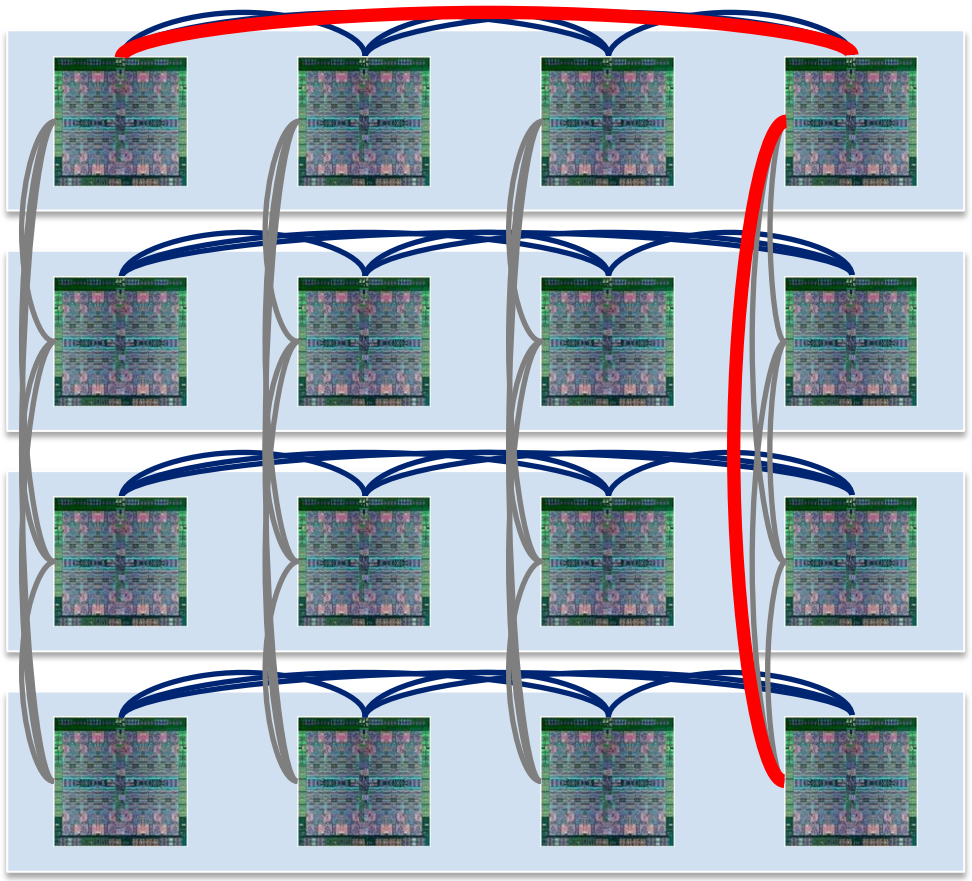
Соединения SMP



On-Module
1 Hop

Off-Module
1 to 3 Hops

POWER7

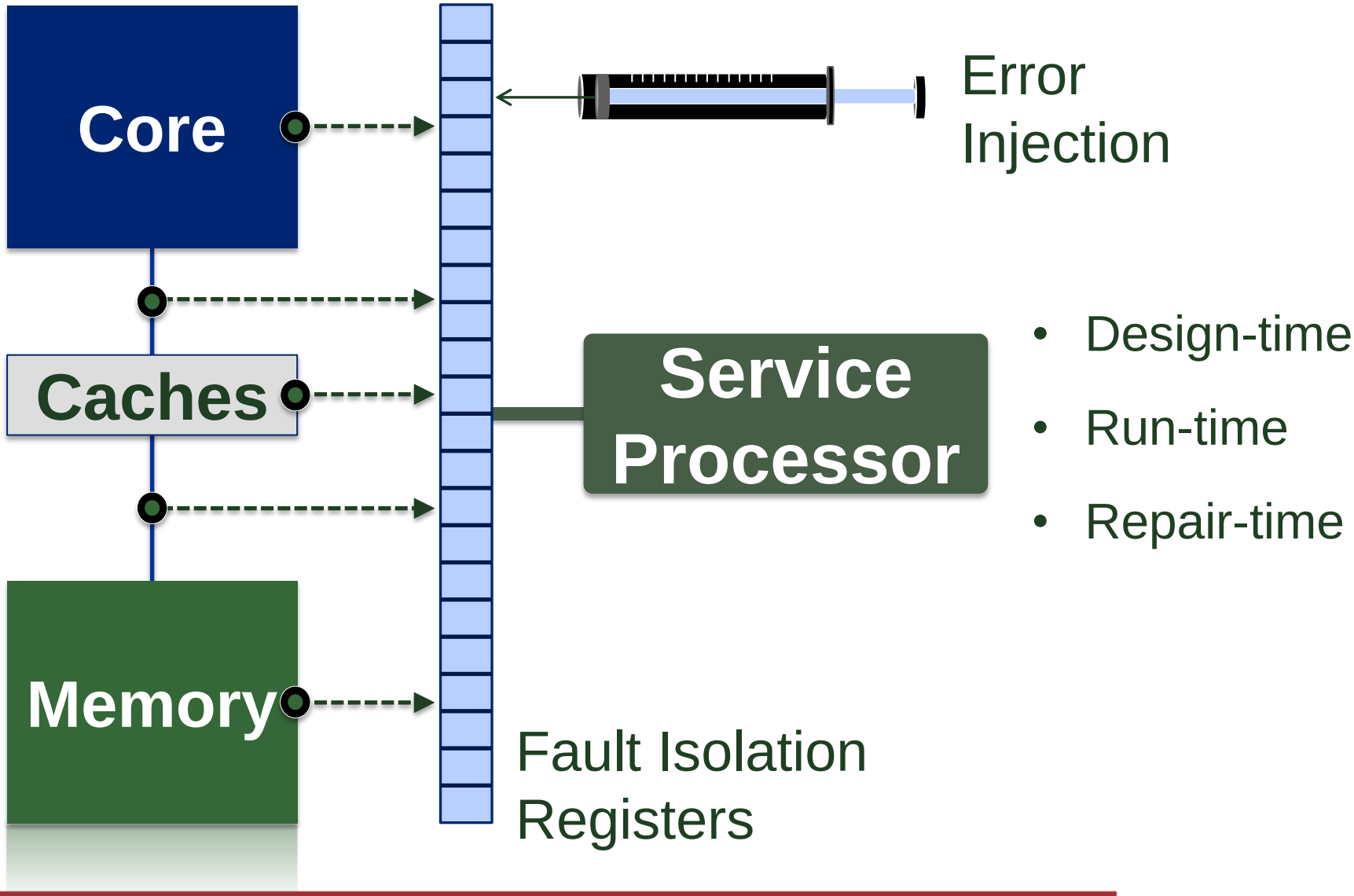


Off-Module 2 Hops

POWER8

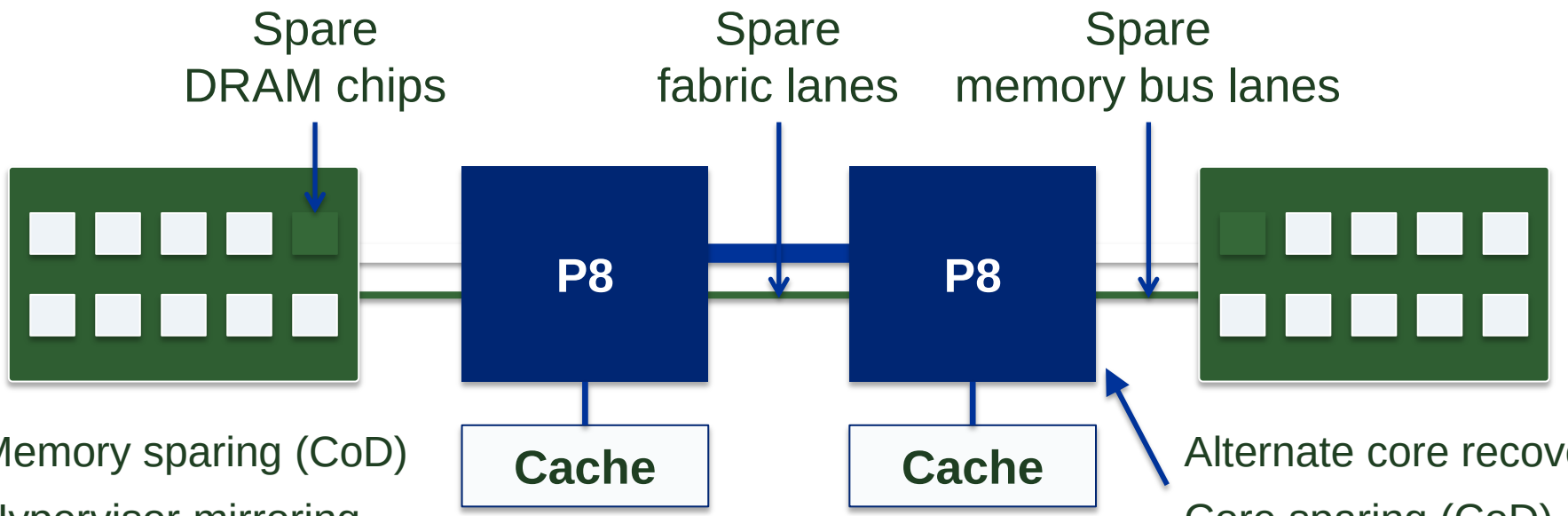
Надежность

First Failure Data Capture



- Design-time
- Run-time
- Repair-time

Авто-мониторинг и восстановление



Memory sparing (CoD)

Hypervisor mirroring (E870/E880/E850)

Alternate core recovery

Core sparing (CoD)

Spare L2/L3 cache columns

L2/L3 cache line delete

L4 persistent fault handling

Redundant I/O adapters, I/O drawers, I/O drawer links, Virtual I/O Servers

Серверы

Семейство IBM Power Systems

BigData & Analytics

S812LC

8348-21C

S822L

8001-22C

High Performance Computing

S822LC for HPC

8335-GTB

Commercial Computing & Cloud

S822LC for Commercial

8335-GCA

S821LC

8001-12C

„LinuxCloud&Cluster-Systems“

8284-2

S812

AIX: 4-core/128GB
IBM i: 1core/64GB
No virtulaization !

8284-22A

S822

1 or 2 socket, 2U
4 (AIX) / 6 - 20 cores

8286-41A

S814

1 socket, 4U
4 - 8 cores

8286-42A

S824

2 socket, 4U
6 - 24 cores

8247-21L

S812L

1 socket, 2U, Linux
10 - 12 cores

8247-22L

S822L

2 socket, 2U, Linux
16 - 24 cores

8247-42L

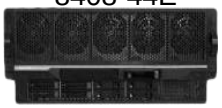
S824L

2 socket, 4U, Linux
8 - 24 cores

„Linux-Systems“

„Cloud-Systems“

8408-44E



9080-MME

9080-MHE

E880C

8 - 192 Cores
256 GB – 32 TB Memor
8 - 192 PCI Adapters

E870C

8 - 64 Cores
256 GB – 16 TB Memory
8 - 96 PCI Adapters

E850C

16 - 48 Cores
128 GB – 4 TB Memory
7 - 51 PCI Adapters

E870 и E880

Power E870C:
8 to 64 cores @ 4.19 GHz
256 to 16TB Memory
1 or 2 nodes per system

Power E880C:
8 to 128 cores @ 4.35 GHz
8 to 160 cores @ 4.19 GHz
8 to 192 cores @ 4.00 GHz
256 to 32TB Memory
1 to 4 nodes per system

E870 и E880

19-inch Rack
(E870 Shown)

EXP24SX SAS Drawers (2U)

- 24 SFF SSD/HDD
- Connects via 2 SAS adapters

PCIe I/O Drawers (4U)

- 12 PCIe expansion slots
- Connects via 2 PCIe adapter slots

System Control Unit (2U)

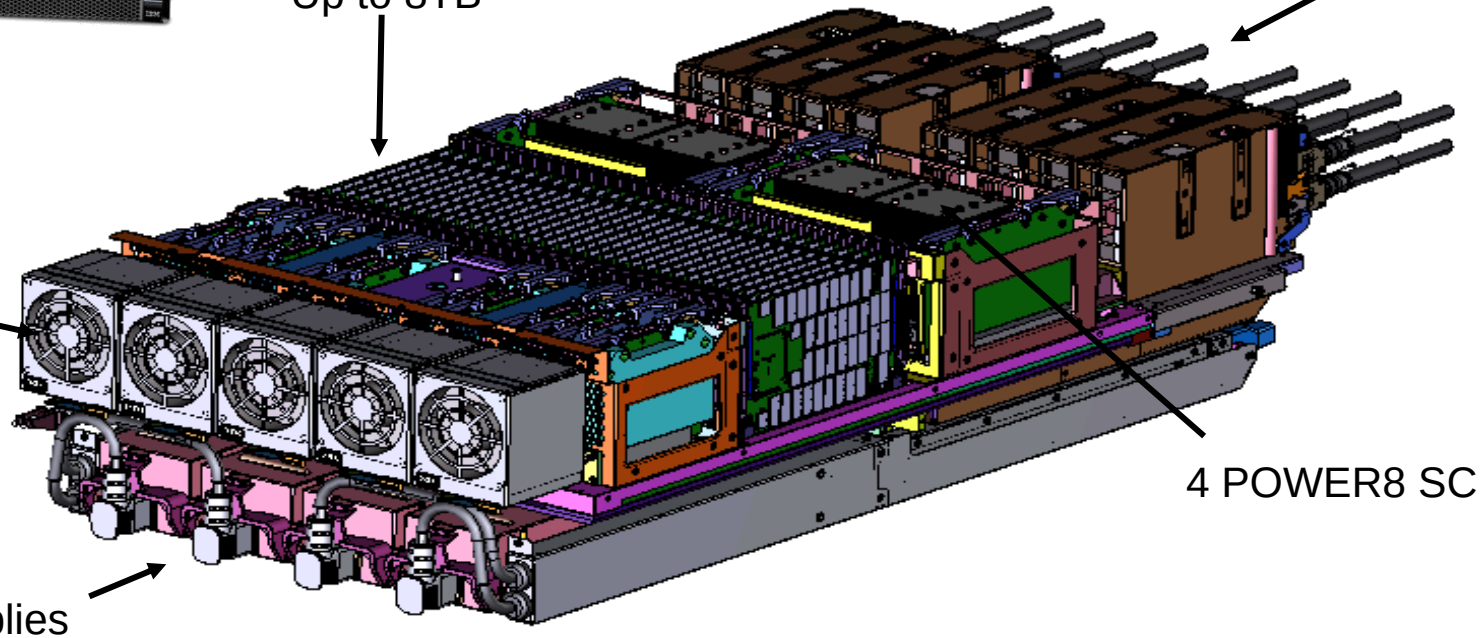
System Nodes (5U)

- 32 – 48 cores / node
- 32 DIMM slots / node
- 8 PCIe Gen3 I/O slots / node

E870 и E880

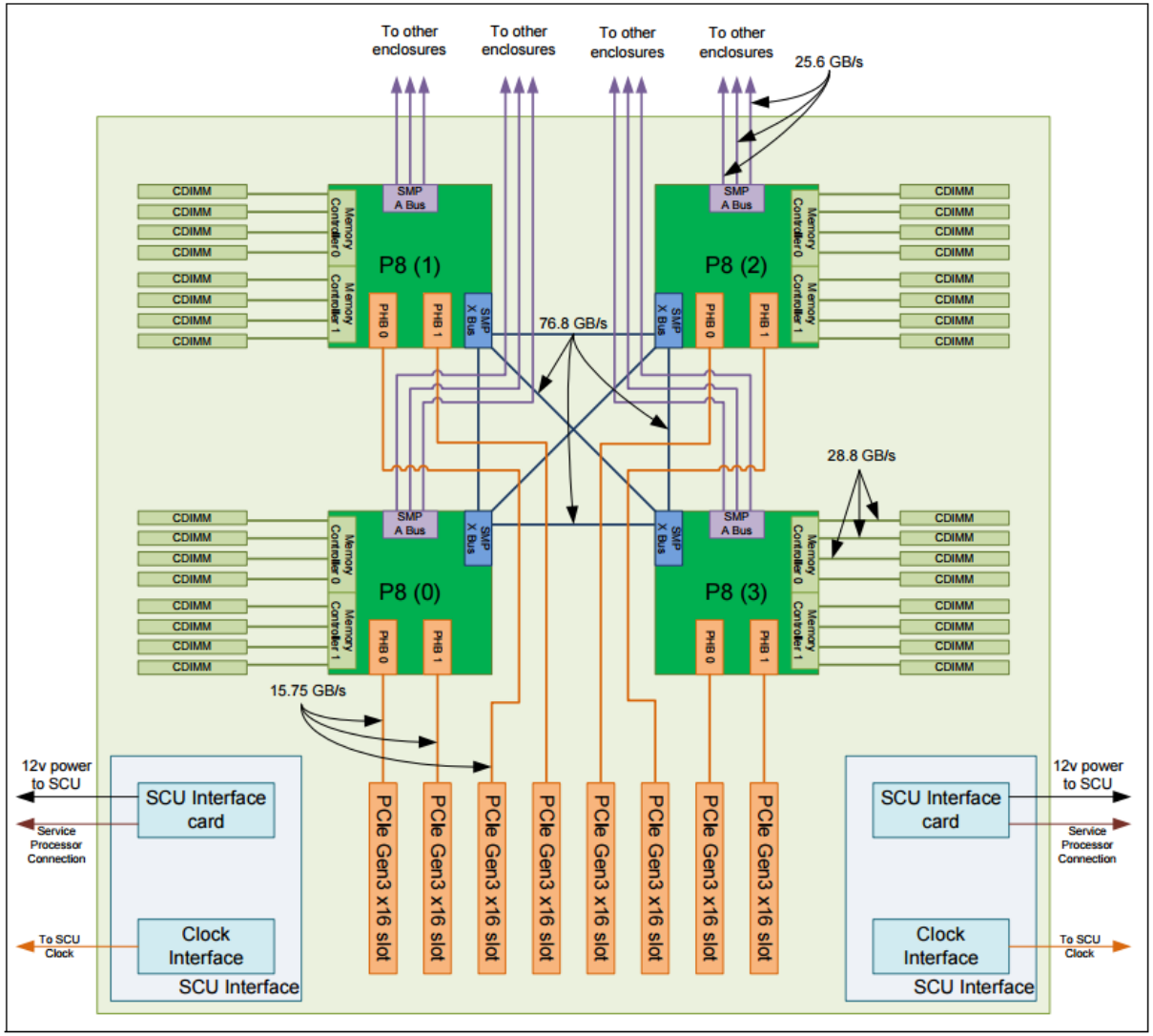
	E870C	E880C		
Node	32-core 4.02 GHz	32-core 4.35 GHz	40-core 4.19 GHz	48-core 4.02 GHz
1 Node	32 cores	32 cores	40 cores	48 cores
2 Nodes	64 cores	64 cores	80 cores	96 cores
3 Nodes		96 cores	120 cores	144 cores
4 Nodes		128 cores	160 cores	192 cores

E870 и E880 System Node (CEC Drawer)



5U Enclosure

Логическая схема



System Control Unit (Midplane)

Один system control unit на каждый сервер

размер - 2U

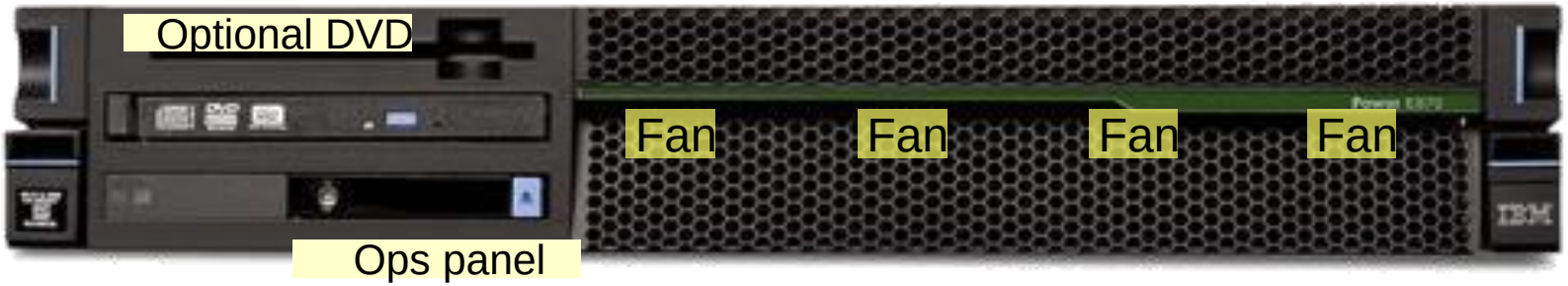
Необходим для работы сервера

System control unit содержит:

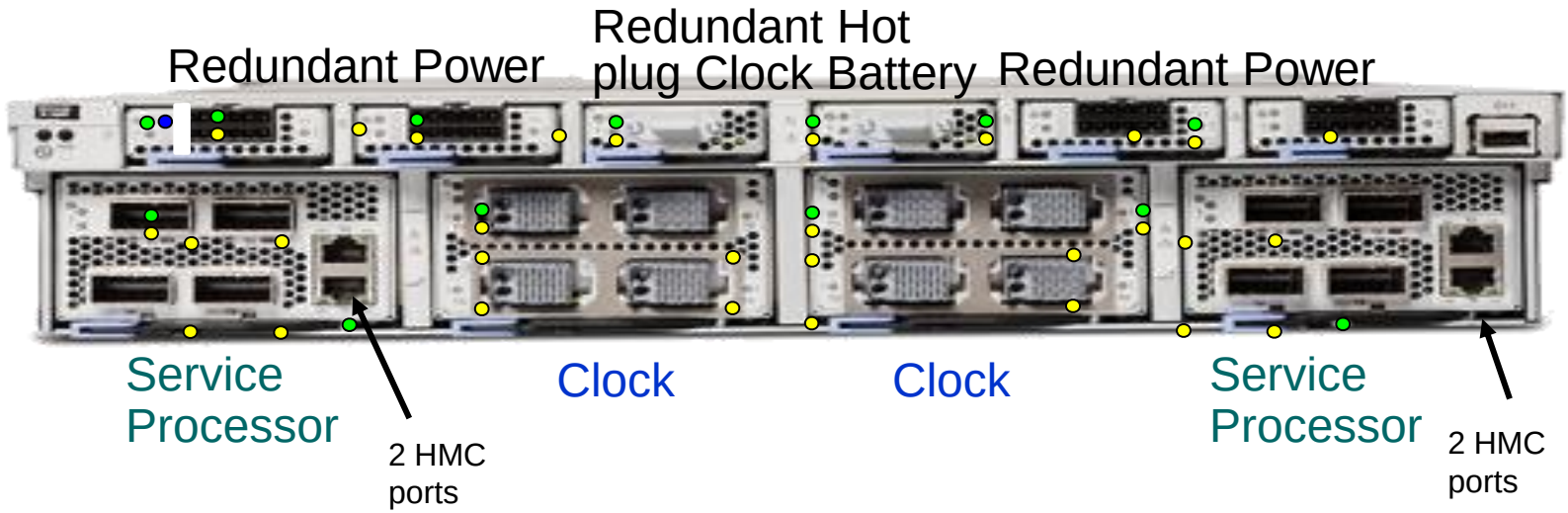
- Сервисные процессора (FSPs)
- HMC порты
- Системные часы
- Операторскую панель
- VPD (vital products data)
- Опционально может иметь DVD

System Control Unit (Midplane)

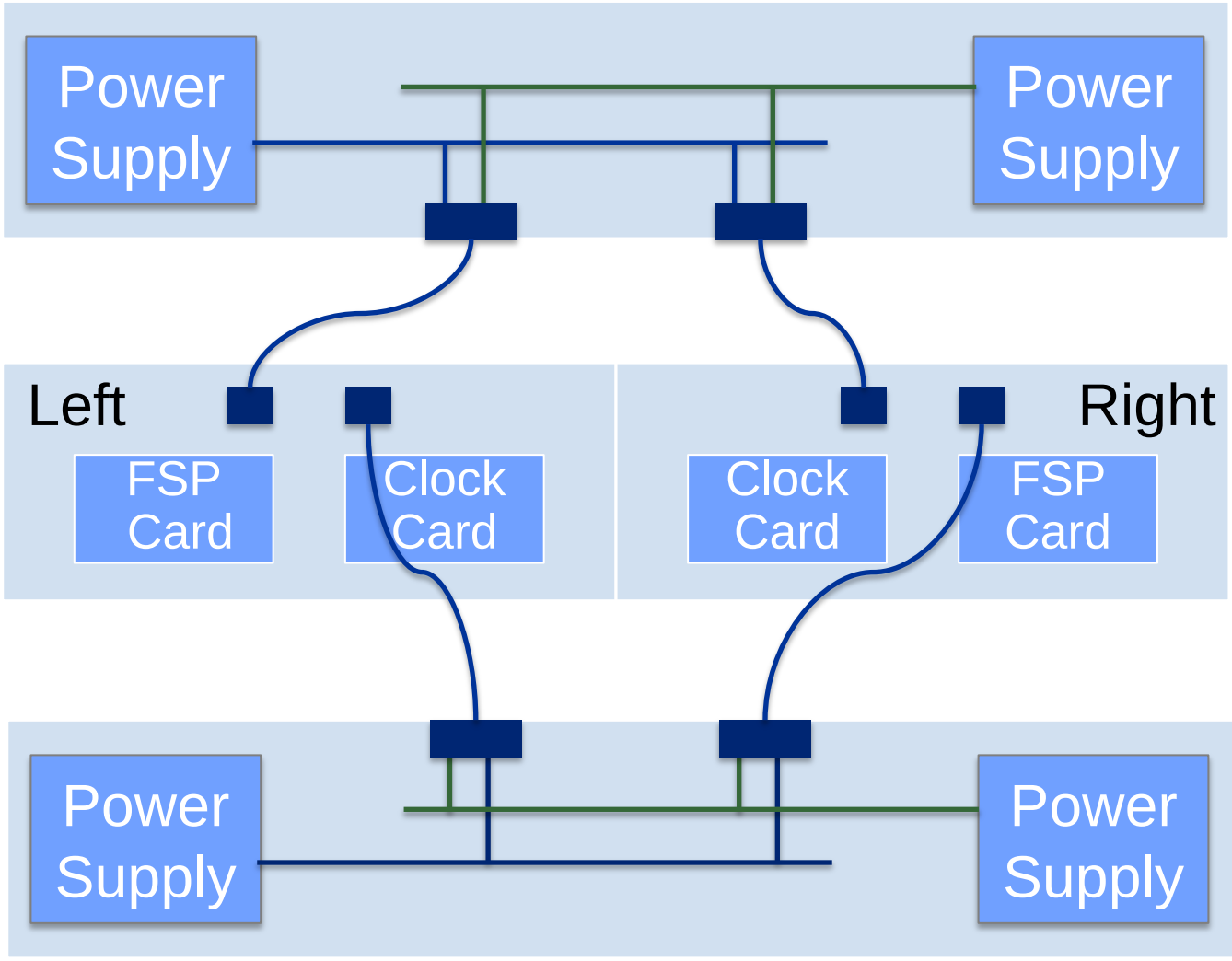
Front View



Rear View



System Control Unit (Midplane)



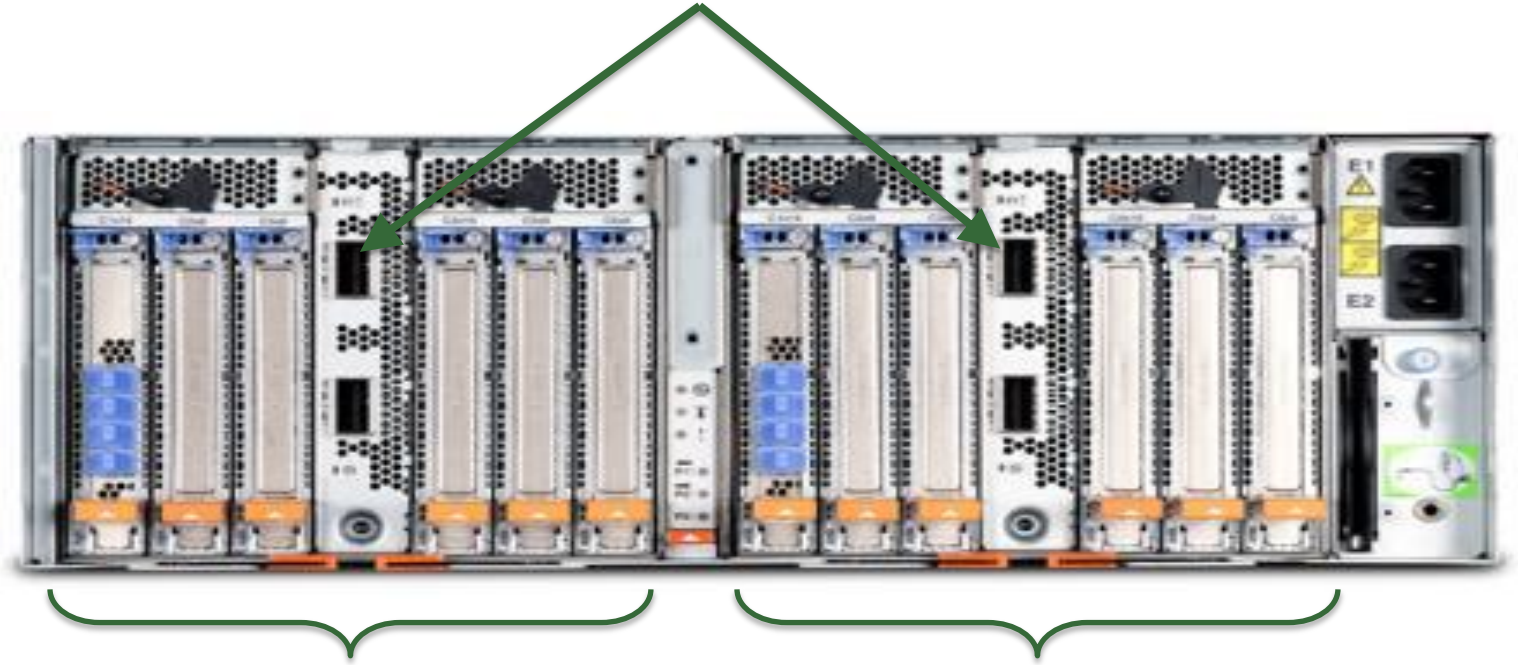
System Node 1

System Control Unit

System Node 2

POWER8 I/O Drawer

PCIe Optical Interface to System Node

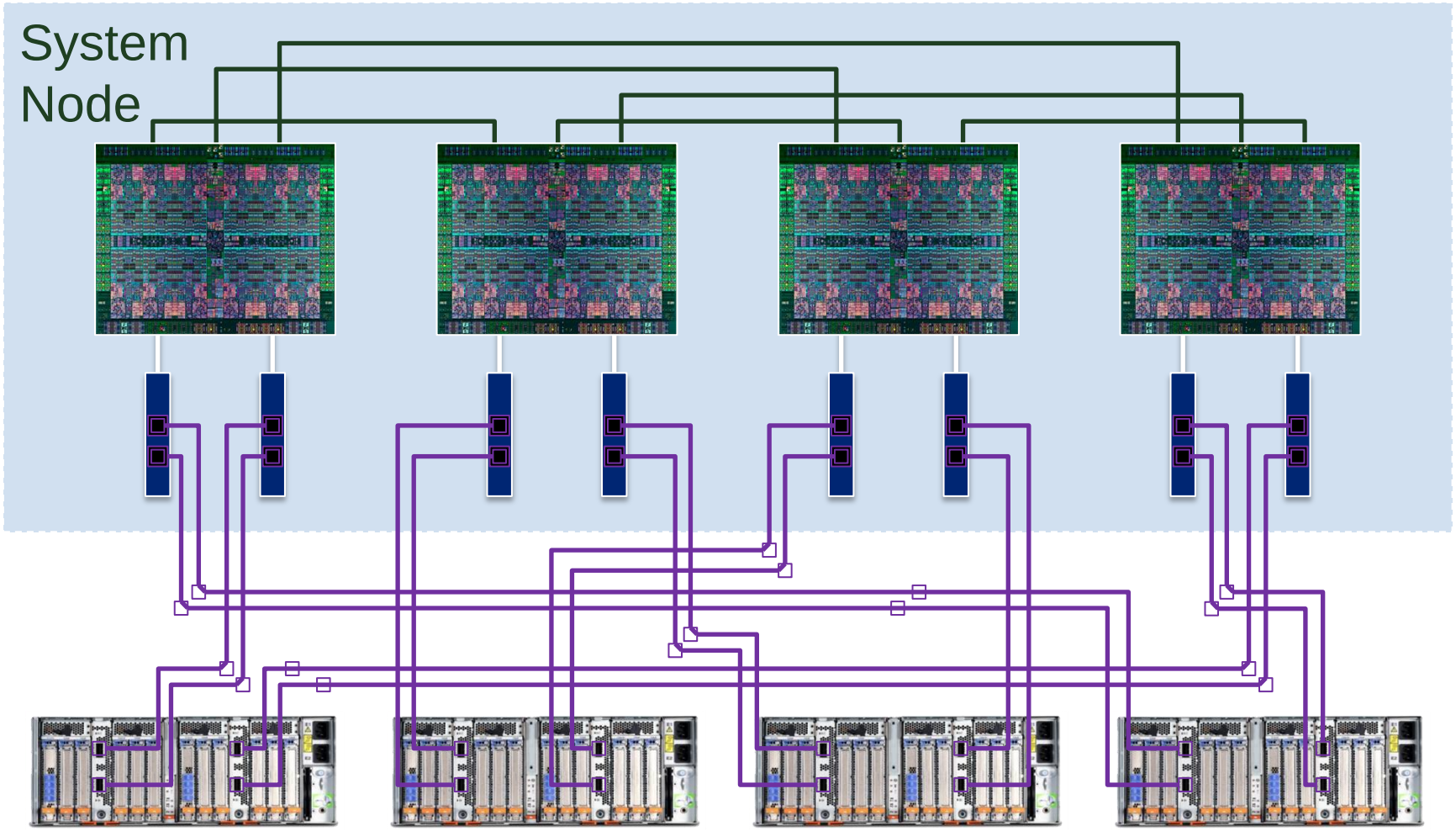


Dual Power Cords

Fan-out Module
6 PCIe Gen3 slots
4 x8 and 2 x16

Fan-out Module
6 PCIe Gen3 slots
4 x8 and 2 x16

POWER8 I/O Drawer



SAS Expansions

- EXP12SX

- 12 3.5-inch SAS bays (LFF-1)

- Low cost per TB, big, 7200 rpm

- 3.86TB or 7.72TB disk drives

- AIX, Linux, VIOS

EXP24SX

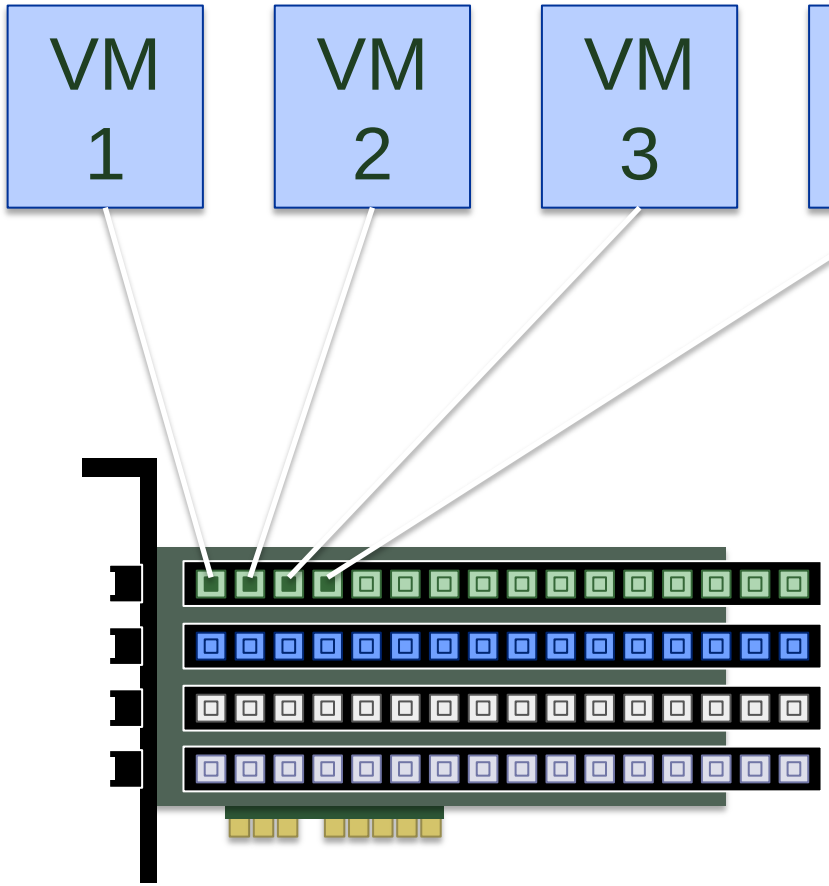
- 24 2.5-inch SAS bays (SFF-2)

- SSD & 10k & 15k drives

- AIX, IBM i, Linux, VIOS

- Подключается через PCIe3 SAS адаптеры
- Mode 1, 2 or 4 --- Can change mode in field carefully using specific procedure

Single Root I/O Virtualization (SR-IOV)



- Прямая Ethernet виртуализация
- Низкая загрузка CPU
- Лучше пропускная способность
- QoS совместимость

Example: 4-port PCIe3 10Gb Ethernet Adapter

Enterprise Pool

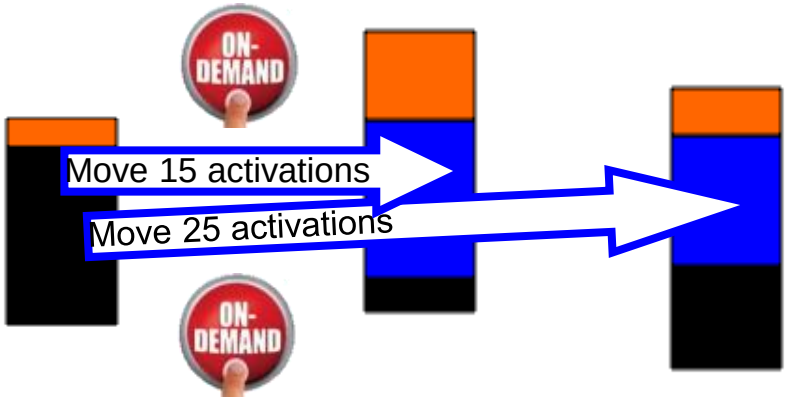
Sys A
64-core E880
4.35 GHz
Activations:
10 static
0 mobile
54 “dark”

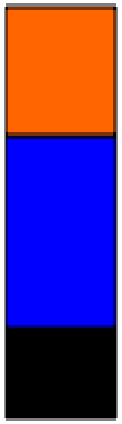
Sys B
96-core 795
3.7 GHz
Activations:
30 static
55 mobile
11 “dark”

Sys C
96-core 780
3.7 GHz
Activations:
16 static
45 mobile
35 “dark”

Sys D
128-core 795
4.0 GHz
Activations:
40 static
60 mobile
28 “dark”

Going back to initial starting point and moving the activations differently





Pool Totals

Activations:
96 static
160 mobile
128 “dark”

Производители платформ

中太服务器 ZOOM SERVER

System

Redpower C210 is the world's first OpenPOWER technology based 2-socket server, which have been GA released in October 2015. By selecting different processing capacity, memory capacity and ability to meet the various application requirements such as IO, database, virtual memory, calculation oriented application. In dealing with similar work load, C210 server can achieve more excellent throughput with higher economic benefits.

TYAN

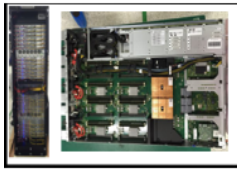
System

Tyan GT75-BP012 platform is a 1U POWER8-based server solution that reveals the spirit of the OpenPOWER Foundation. the using the ppc64 architecture in the 1U single-socket system, the TYAN GT75-BP012 provides huge memory footprint as well as outstanding performance for HPC and server virtualization applications

Mark III Systems

System

Mark III Systems will commercialize the first OpenPOWER-enabled system built to the Open Compute Project design specification, which will follow the Barreleye server design created by Rackspace, Broadcom, IBM, Ingrasys, Mellanox, Micron, and Samsung. As a long-time IBM Premier Partner, Mark III will also offer value-added services and expertise to clients to help them plan for, implement, and maintain Barreleye within the context of their existing and future technology strategies.



rackspace HOSTING

Server

Rackspace partnered with Broadcom, IBM, Ingrasys, Mellanox, Micron, and Samsung to create "Barreleye" an OpenPOWER / Open Compute Platform designed for Cloud 2.0. The first Barreleye servers came online in 3Q 2015. With our first data center shipment in process, we look forward to bringing customers aboard within the next few months.

inspur 浪潮

System

INSPUR Allure - 1S or 2S 4U POWER8

4U2S Rack Server Optimized for New Data Center and Big Data Application: Supports 2 IBM OpenPOWER Processors, 64 DDR3—2TB, 12 PCIE Gen3

wistron

System

E4 COMPUTER ENGINEERING

Wistron Polaris – 2 Sockets POWER8 Server with Nvidia GPUs—Polaris is a ready-to-ship 2U 2-socket system built on POWER8 with CAPI (Coherent Accelerator Processor Interface) to deliver high performance, reliability, scalability in high-performance computing (HPC) environments.

STACK VELOCITY

System

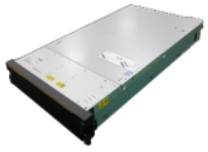
Saba - 2U High-Performance Data Analytics Engine.

StackVelocity has designed a high-performance platform, Saba, featuring OpenPOWER™ POWER8™ and CAPI. Saba boasts leadership performance, enhances cloud efficiency and enables open innovation on POWER. An extremely smart platform with high number-crunching capabilities, Saba is perfect for big data analytics and hyper performance computing applications.

ITL 英特力

System

ITL CP8128 on-board server is a kind of server which is designed and developed independently based on IBM open OpenPOWER chip architecture. ITL server supports Redhat operating system and virtual technology. The structure of server uses module design and can satisfy customer's customized demand of on-board server.



Neu Cloud Oriental

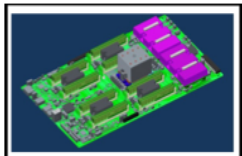
System

The NL2200 is the first jointly developed product by Beijing Neu Cloud Oriental System Technology Co., Ltd. and partners from OpenPOWER Foundation. It can support 16/20/24 cores CPU, up to 1TB memory, 2 HDDs, 2 GPUs, 100GB IB adapter & PCIe SSD with NCO self-developed BMC."

PENGUIN COMPUTING

System

Penguin Magna 2001, powered by a single OpenPOWER Power8 generation CPU, is targeted at software development and testing on OpenPOWER architecture. Magna 2001 supports rich I/O options, including CAPI, to support variety of I/O driven workloads.



Inventec

Board

Inventec's First OpenPOWER project targets to provide high-performance, scalability and flexibility for large Data Center deployments and service applications. Using IBM Centaur Memory buffer and PLX PEX8725 switch supported 1 IBM P8NVL CPU, 16 DDR4 DIMMs and 2 nVIDIA P100 SXM2 GPUs. Through NVLINK connects CPU and GPUs, and GPUs to bring high computing efficiency.

IBM

SUPERMICRO

System

IBM 1U/2S Power8 Prototype It supports up 512GB DDR4 memory, 1 NVIDIA K80 or up to 2 Alpha-Data KU3 CAPI adapters and 4x LFF/SFF hot swap bays.

Спасибо
за внимание!