

# HPC



# «Пара слов» о себе

- ФИО: Окунев Дмитрий Юрьевич
- Должность: начальник отдела UNIX-технологий, НИЯУ МИФИ
- Контакты:
  - ✉ dyokunev@mephi.ru (PGP: 8E30679C)
  - ☎ +7 (495) 788 56 99, доб. 8255
  - IRC: irc.campus.mephi.ru#mephi

В лицее создан  
высокопроизводительный  
вычислительный кластер «lambda»



В лицее создан  
высокопроизводительный  
вычислительный кластер «lambda»

And then what?



# And then what?

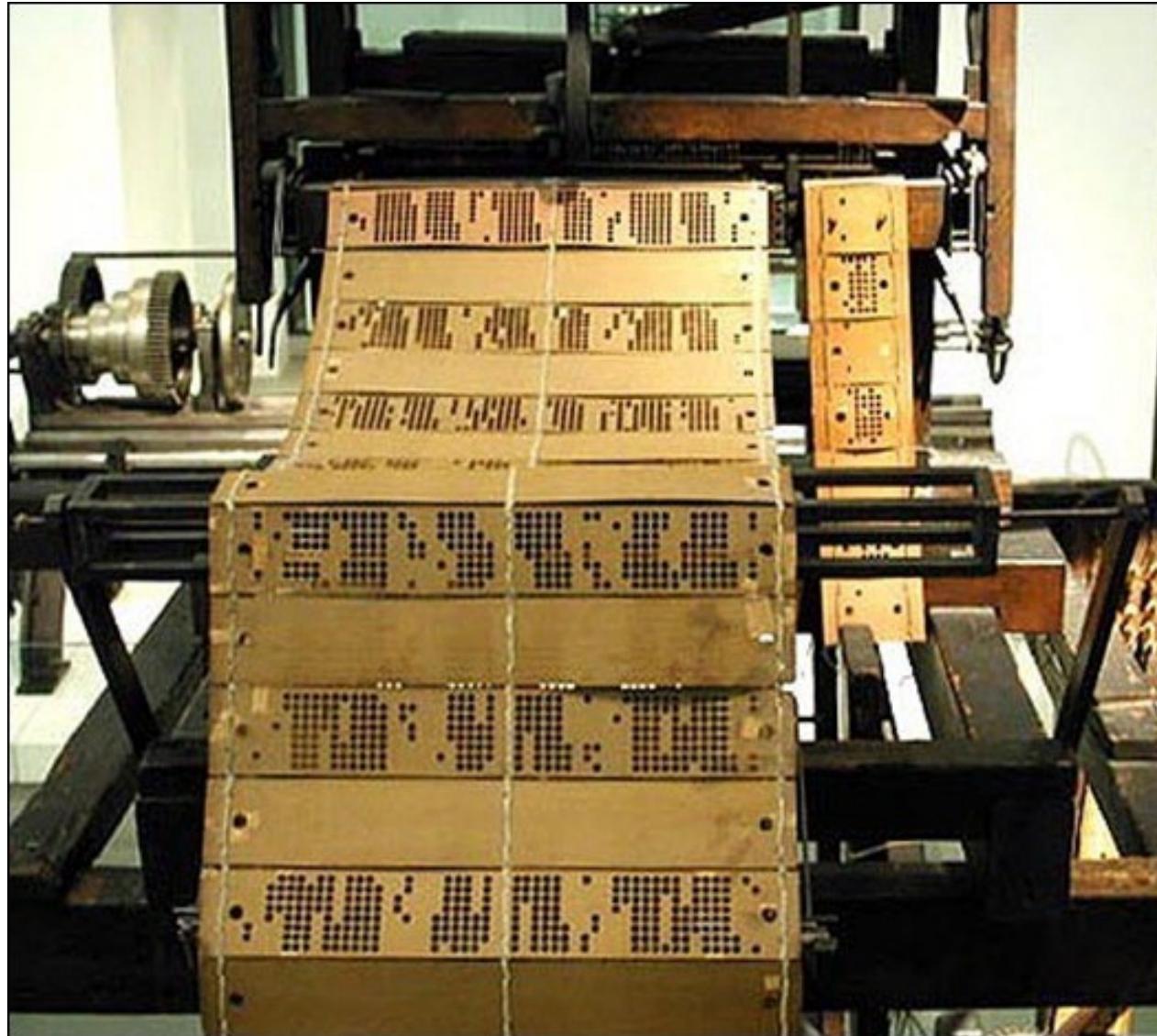
Постараемся разобраться, что это за кластер, и  
как вы можете его использовать.

Но в первую очередь вспомним о том, что такое  
современный компьютер.

# Что это за устройство?

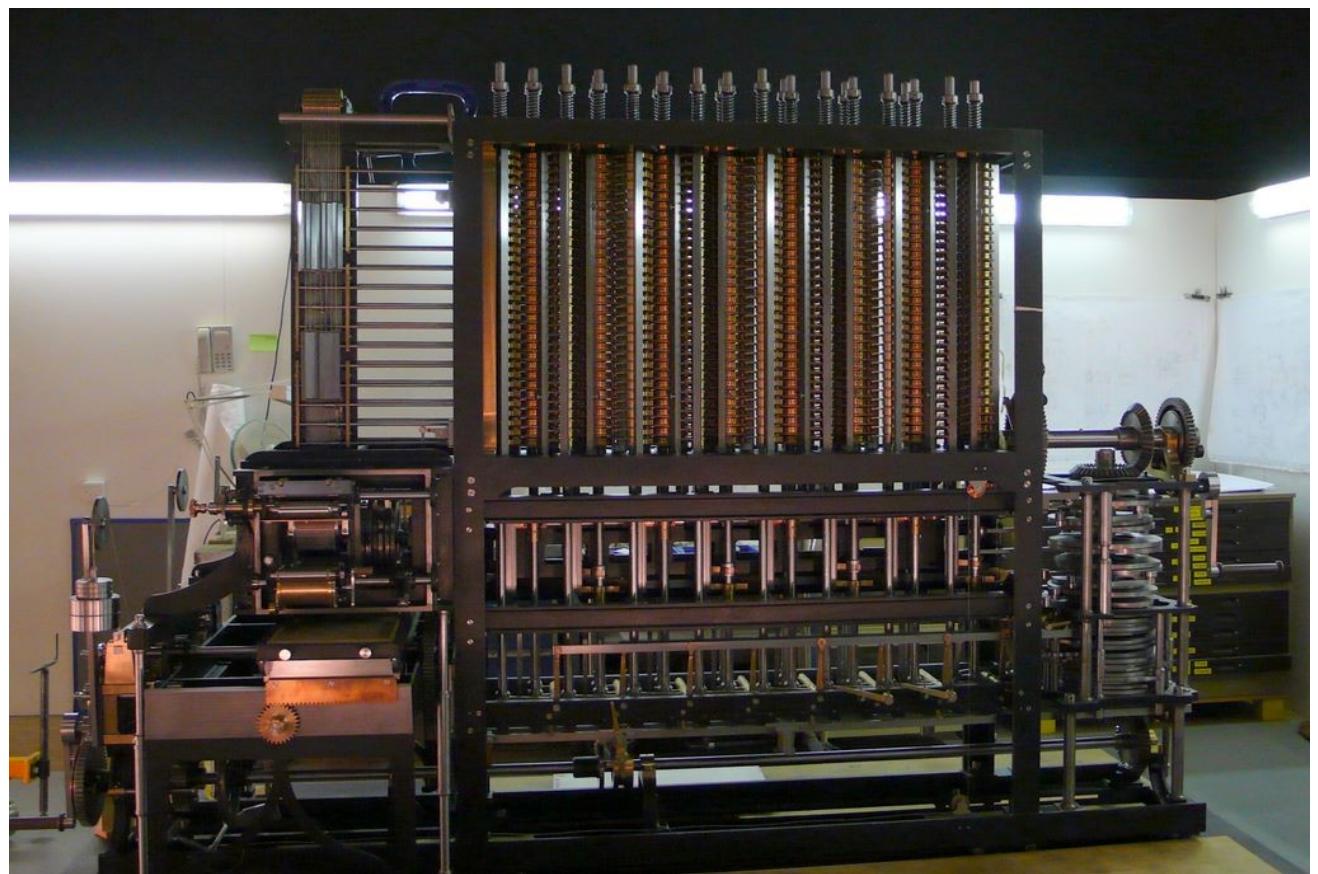


# А это?



# Первый в мире программист

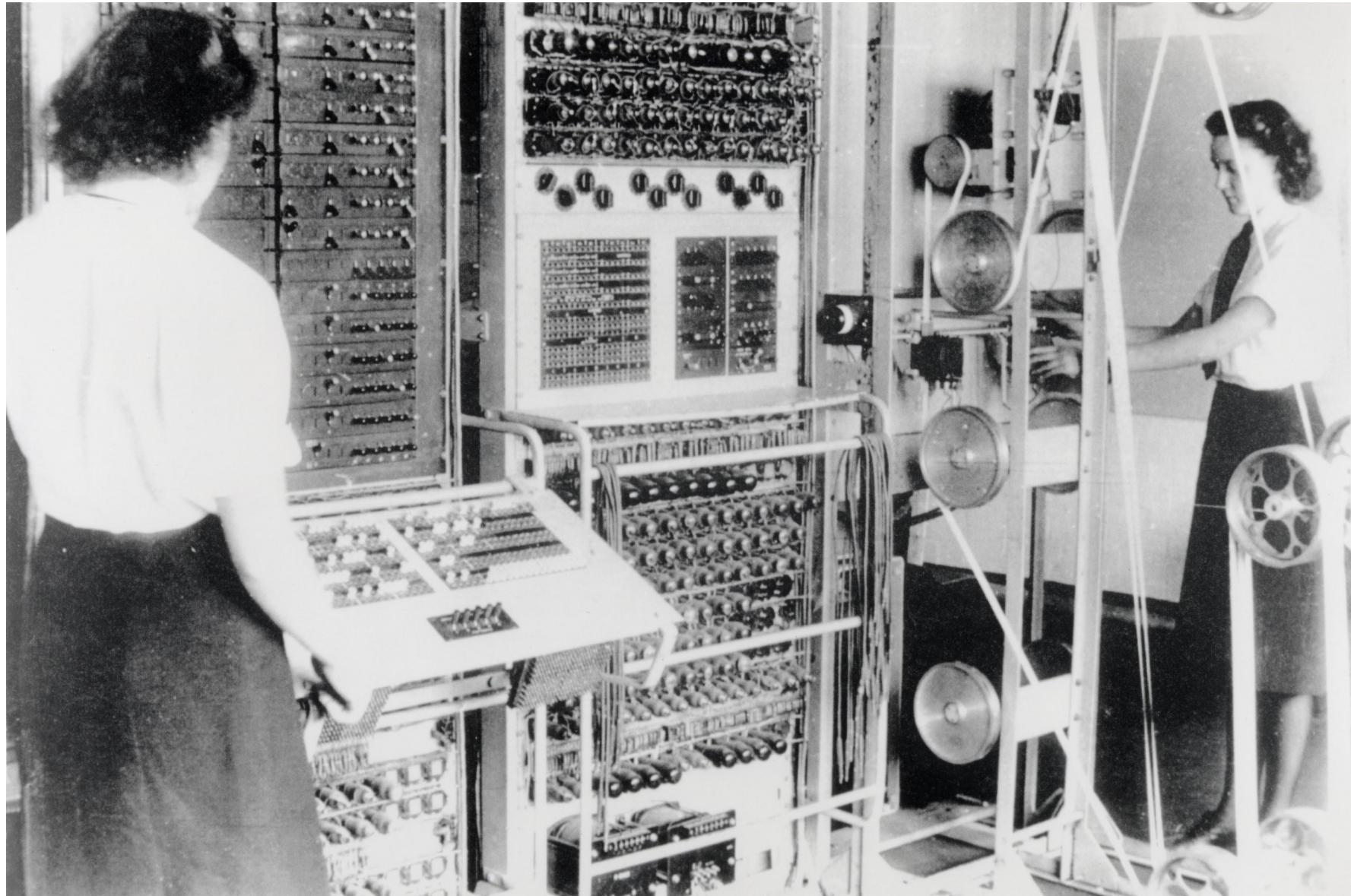
- Августа Ада Кинг Лавлейс



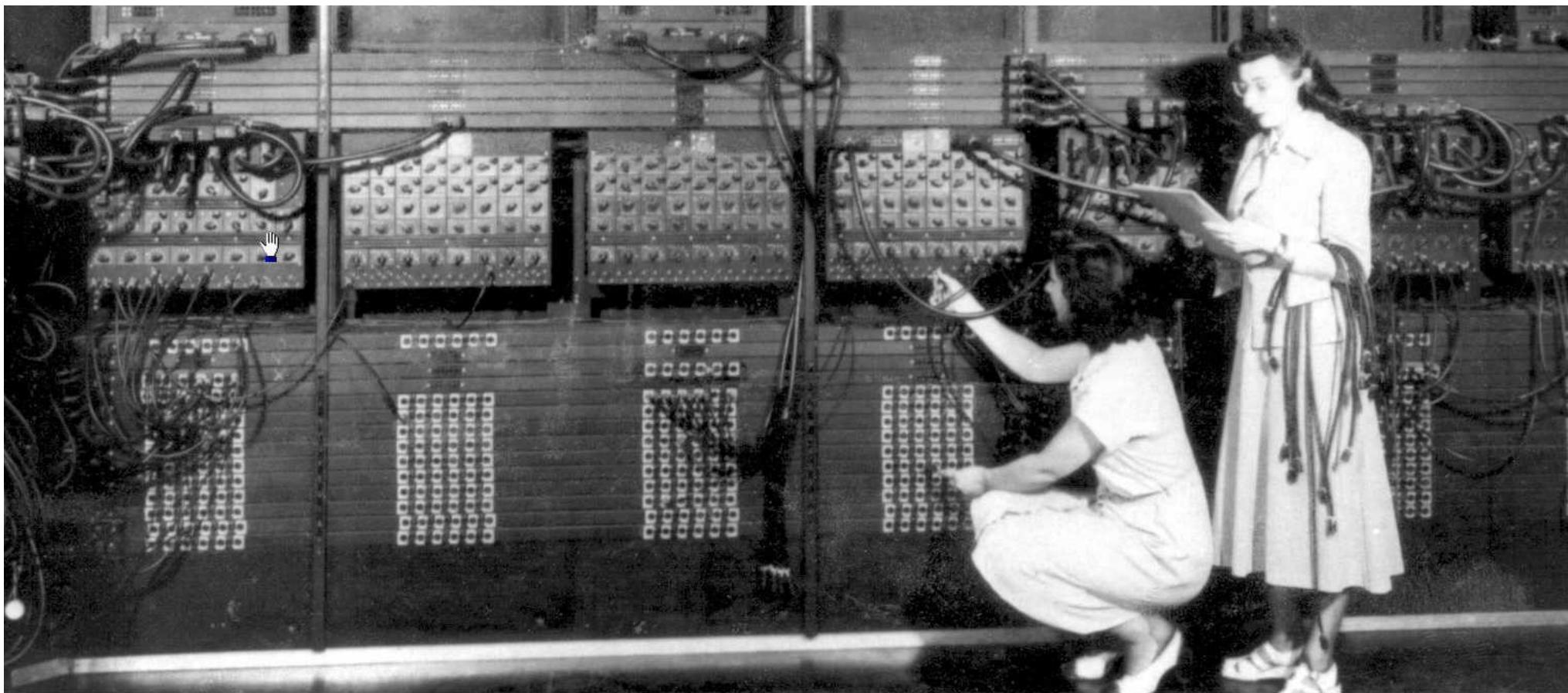
# Что такое компьютер и зачем он нужен?

?

# Что такое компьютер и зачем он нужен?



# Что такое компьютер и зачем он нужен?



# Зачем нужен компьютер?

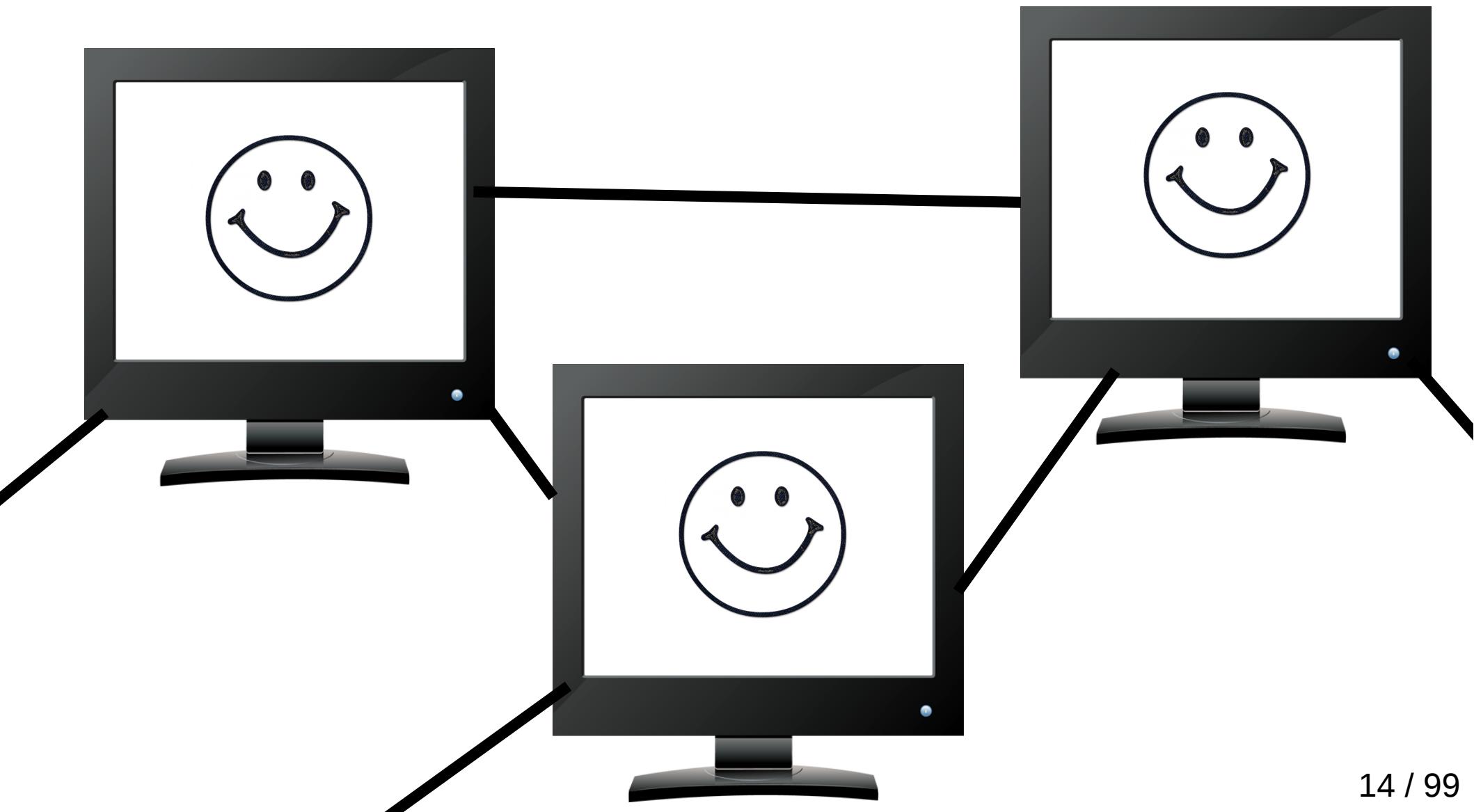


Поручи задачу компьютеру, если сможешь

# Что делать, если возможностей компьютера не хватает?



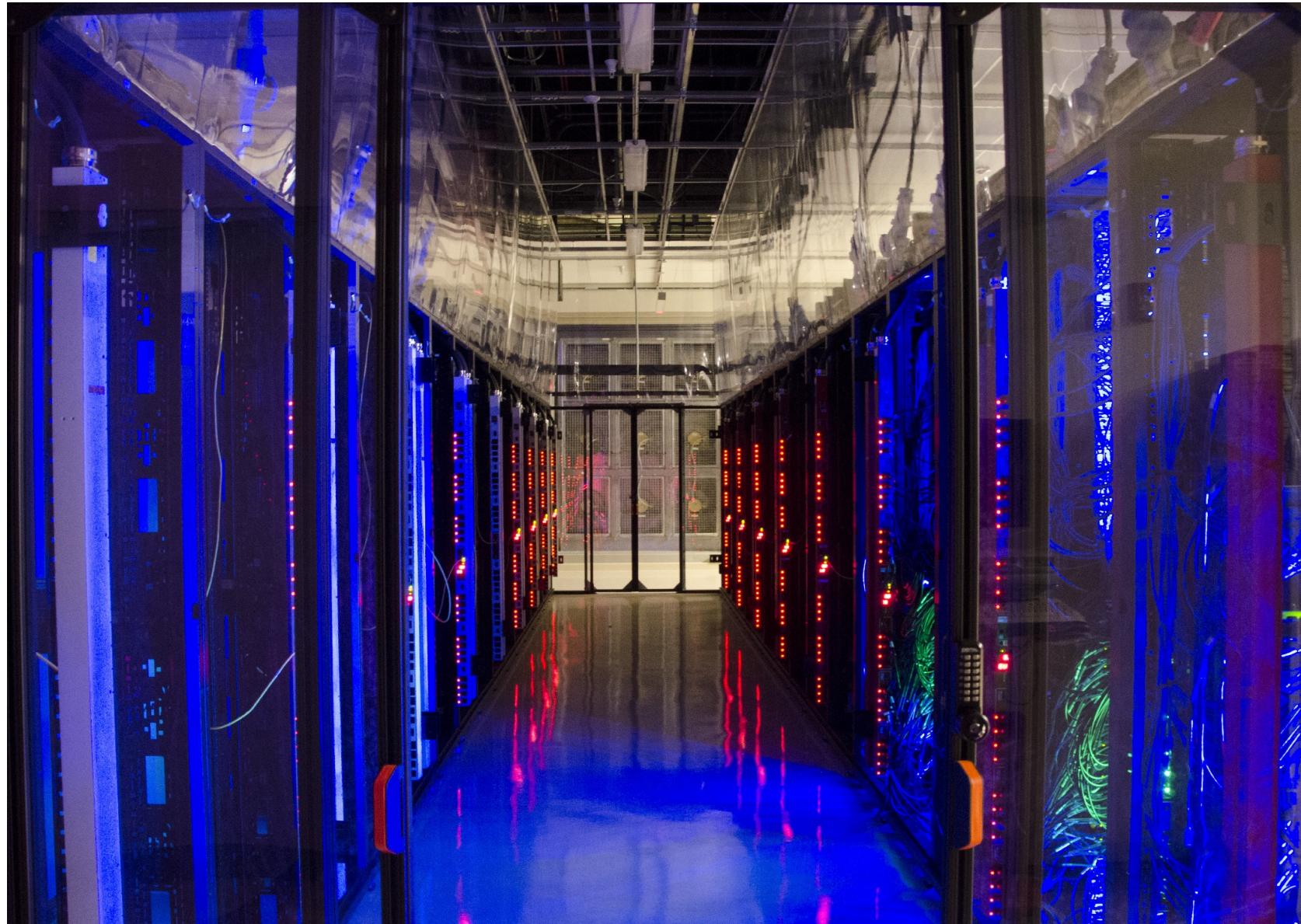
# Что делать, если возможностей компьютера не хватает?



# Что такое НРС?

- НРС – это о больших вычислительных комплексах, возможности которых существенно превосходят возможности обычных компьютеров.
- НРС == High Performance Computing (высокопроизводительные вычисления)
- Вы придумали для компьютера задачу, а он не справляется → идём за помощью в НРС

# Что же такое HPC?



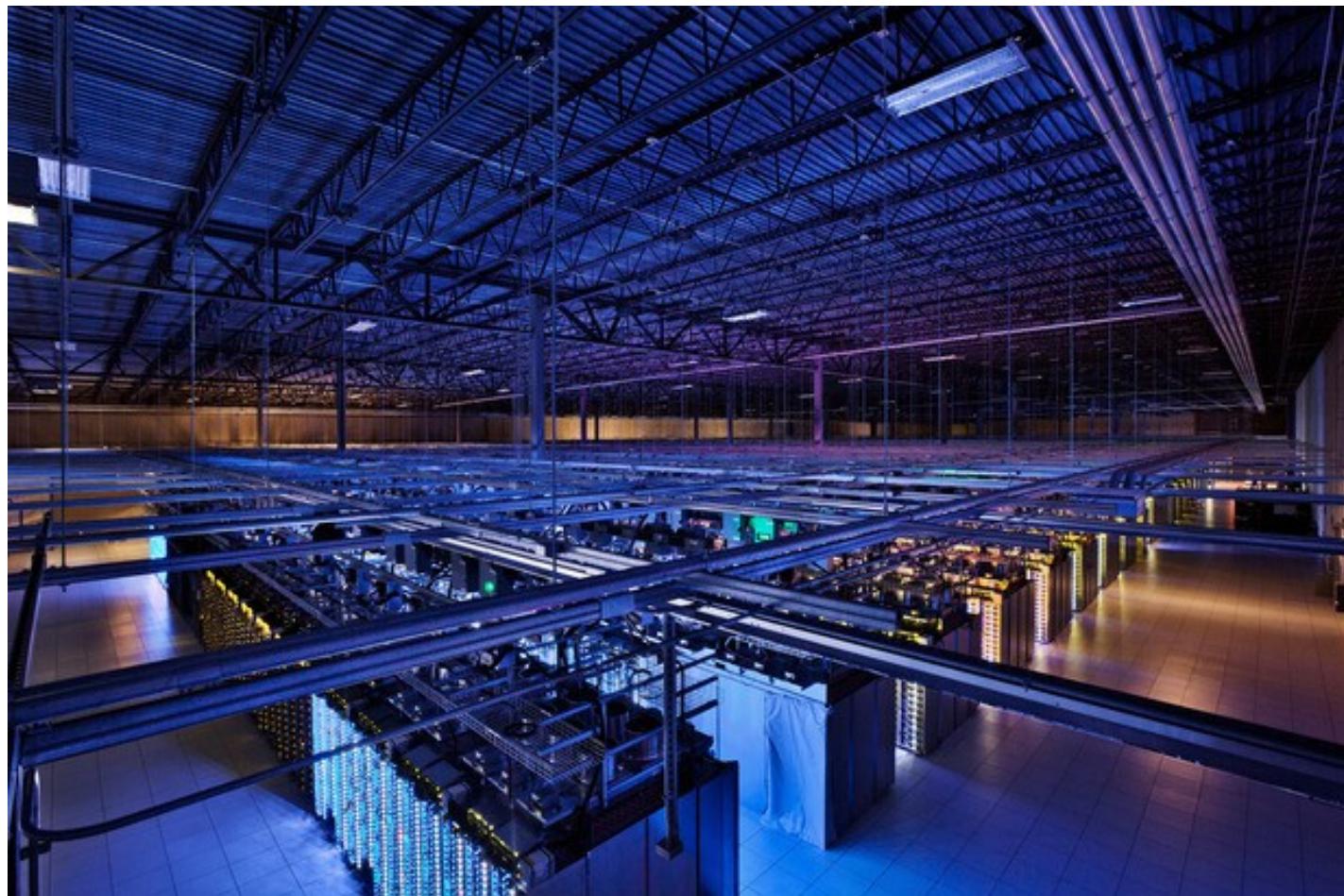
# Основные характеристики вычислительного кластера

- Пиковая (теоретическая) производительность (Tflops)
- Linpack производительность (Tflops)
- БогоMIPS производительность?
- Пропускная способность Interconnect (Gbps)
- Задержки Interconnect (ns)

# Основные характеристики вычислительного кластера

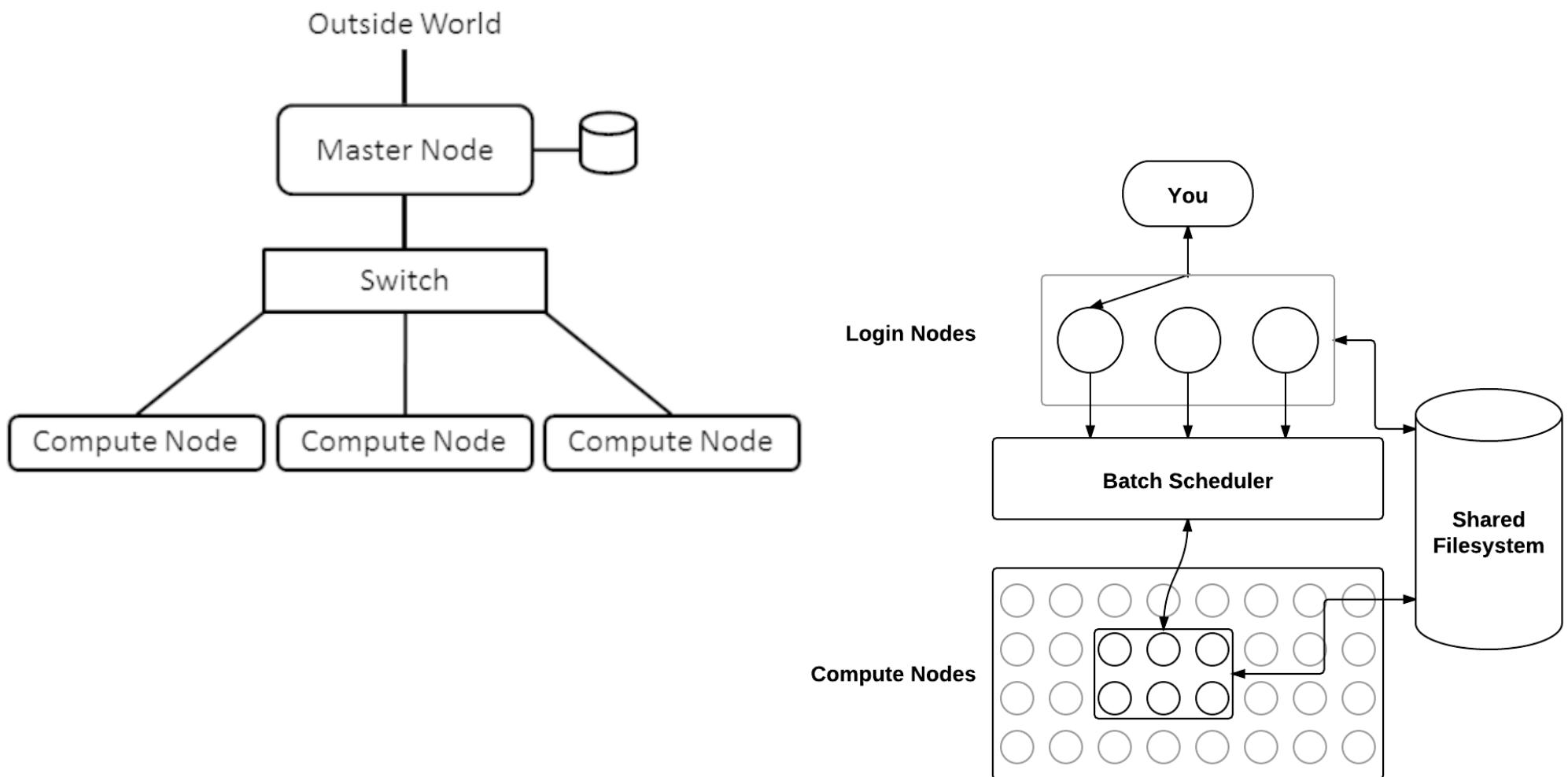
- Пиковая (теоретическая) производительность (Tflops)
- Linpack производительность (Tflops)
- БогоMIPS производительность?
- Пропускная способность Interconnect (Gbps)
- Задержки Interconnect (ns)

# Что же такое HPC?

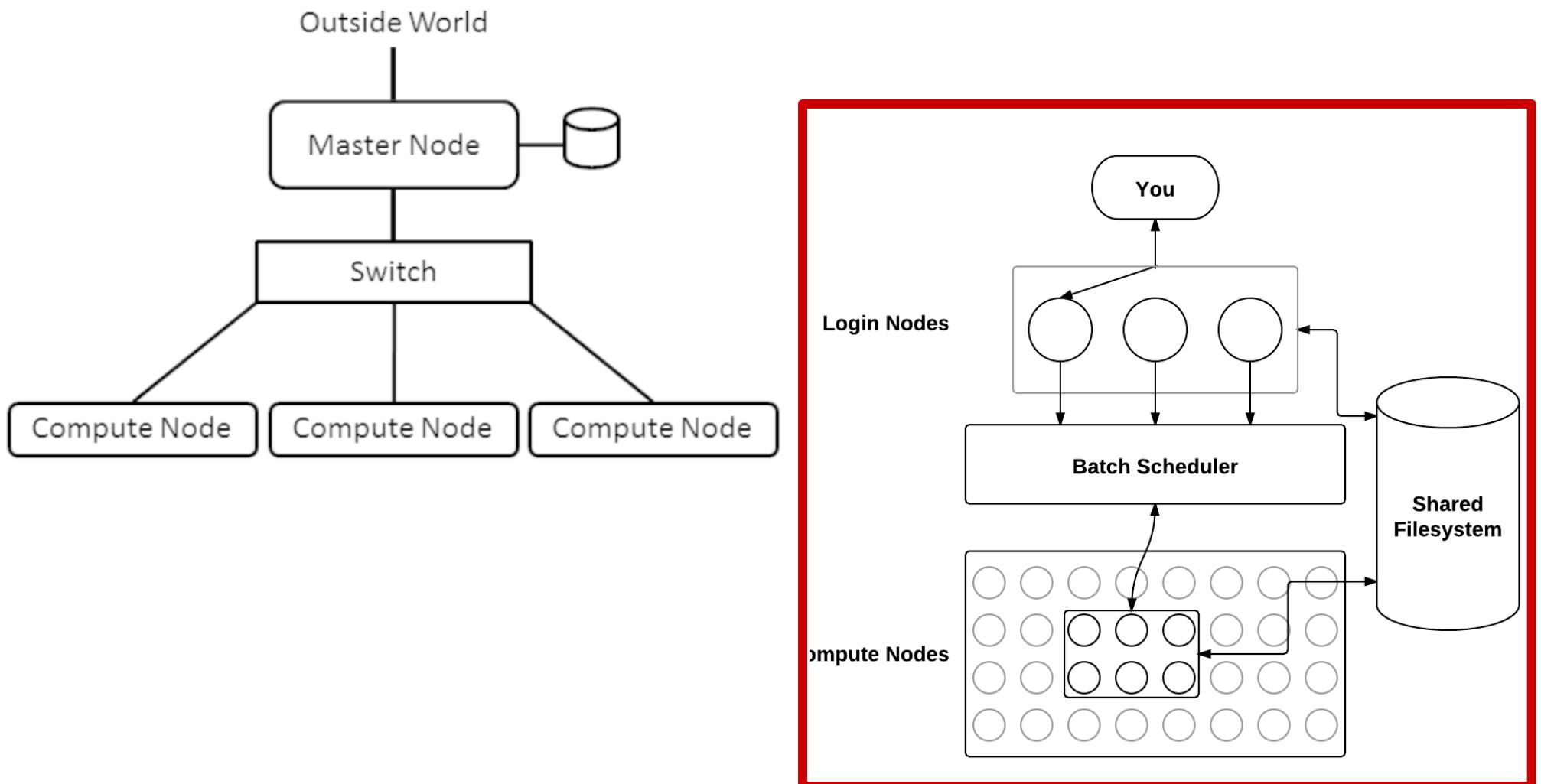


Вычислительный ресурс	Мощность, TFlop/s, peak
Top1 «Sunway TaihuLight»	125000
Top52 «Lomonosov 2»	2962
Суммарно в НРС-Центре МИФИ	21

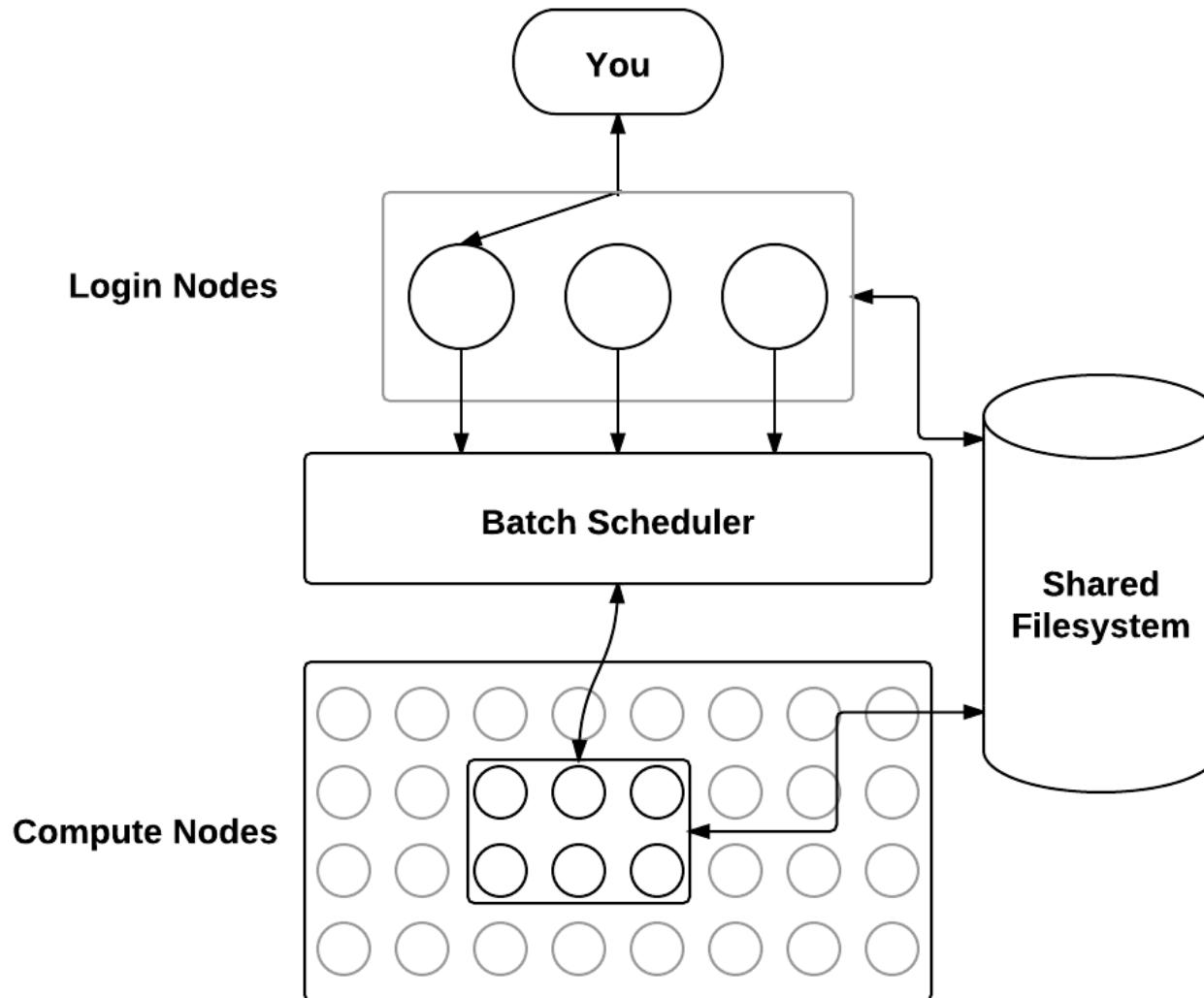
# Архитектура вычислительного кластера



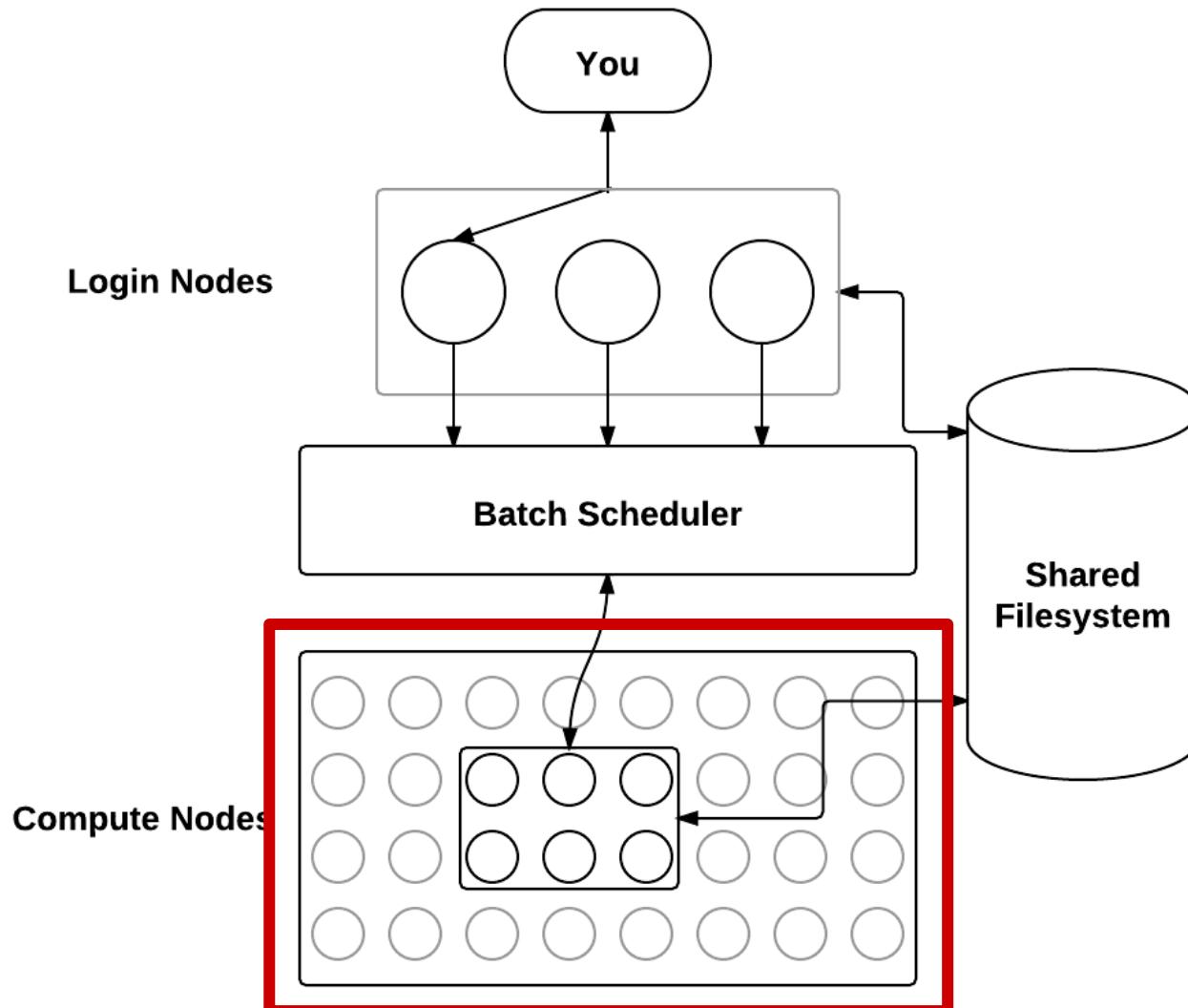
# Архитектура вычислительного кластера



# Архитектура вычислительного кластера

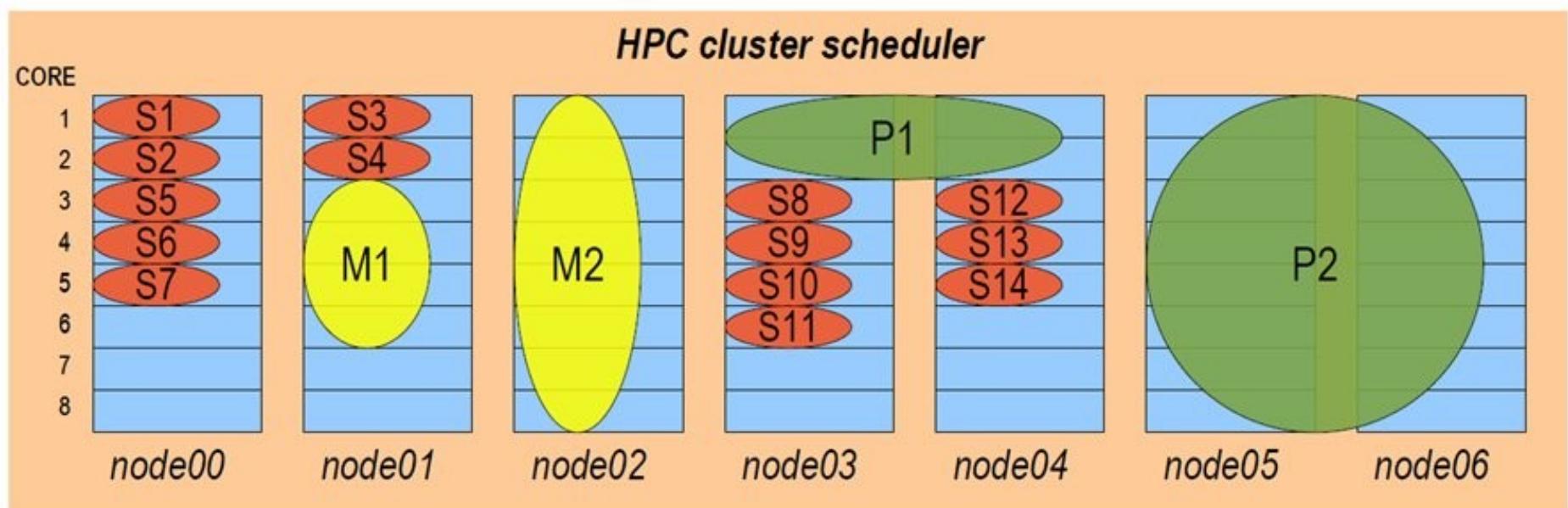


# Архитектура вычислительного кластера

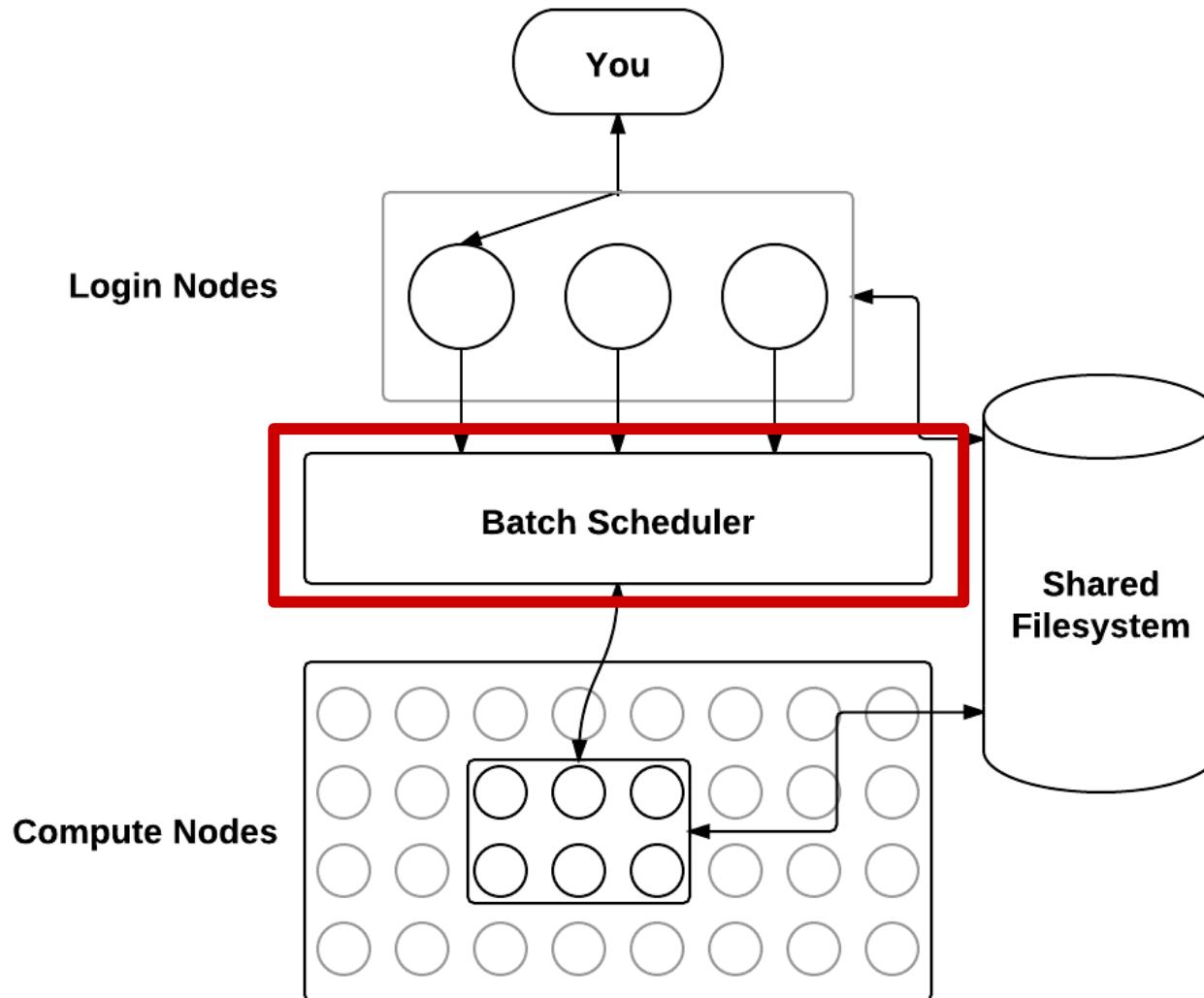


# Архитектура вычислительного кластера: вычислительное поле

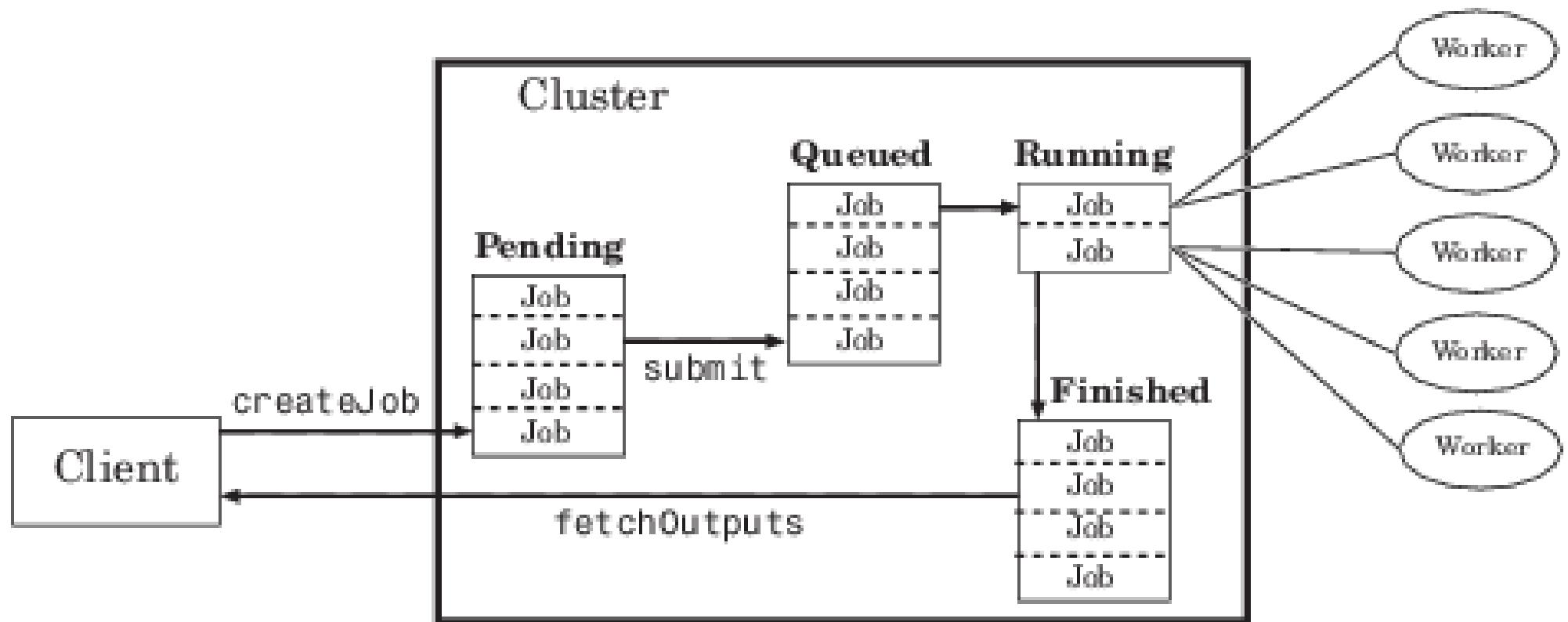
- S Serial jobs (single core, single node)
- M Multi-threaded jobs (many cores, single node)
- P Parallel jobs (many cores, many nodes)



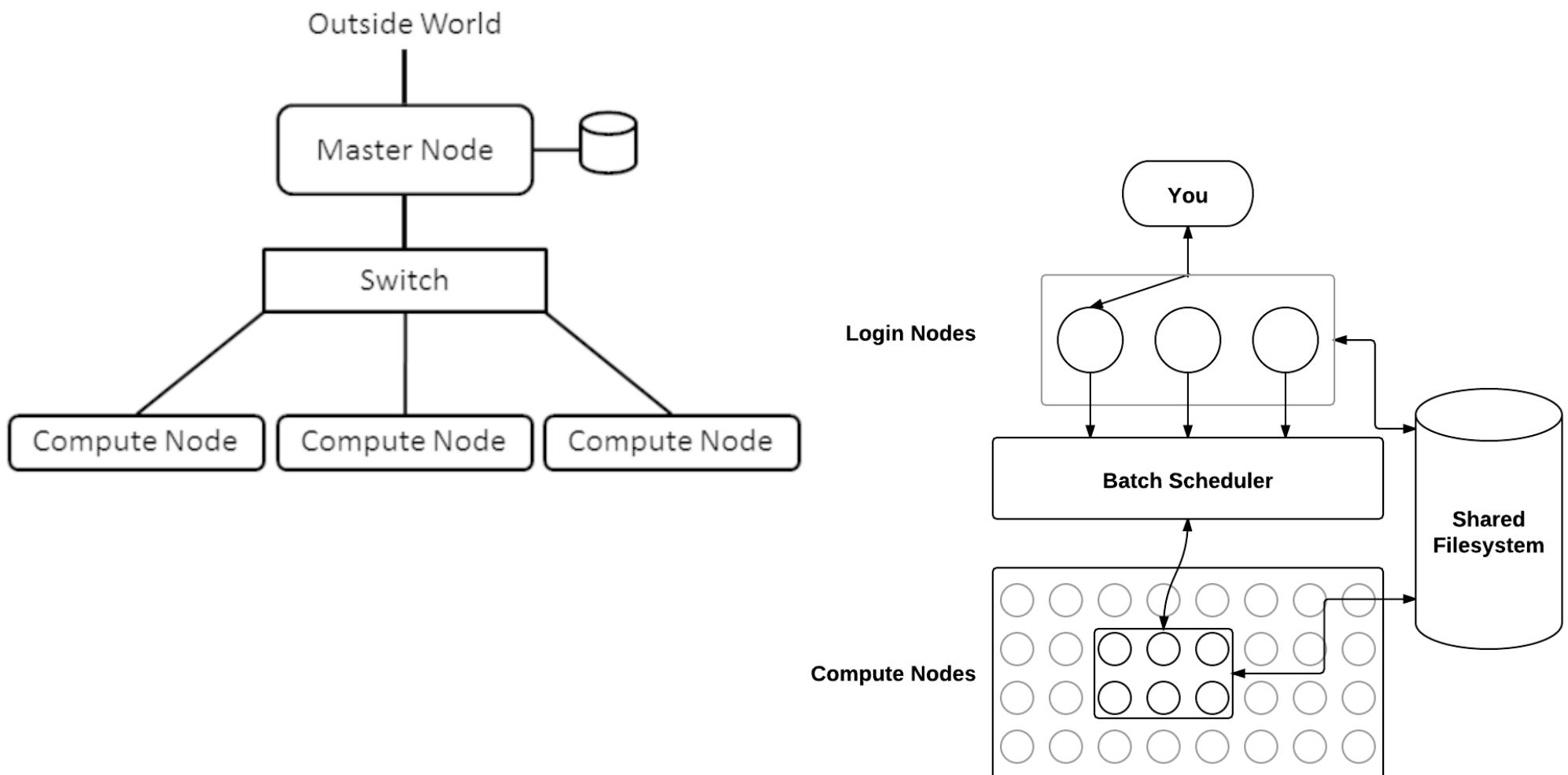
# Архитектура вычислительного кластера



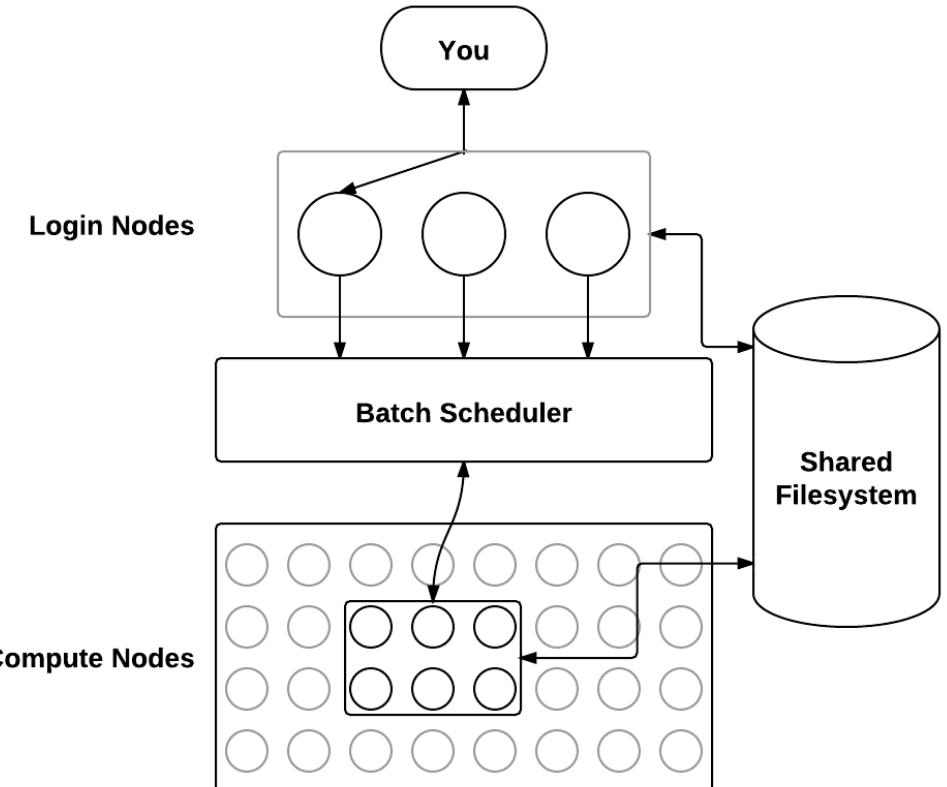
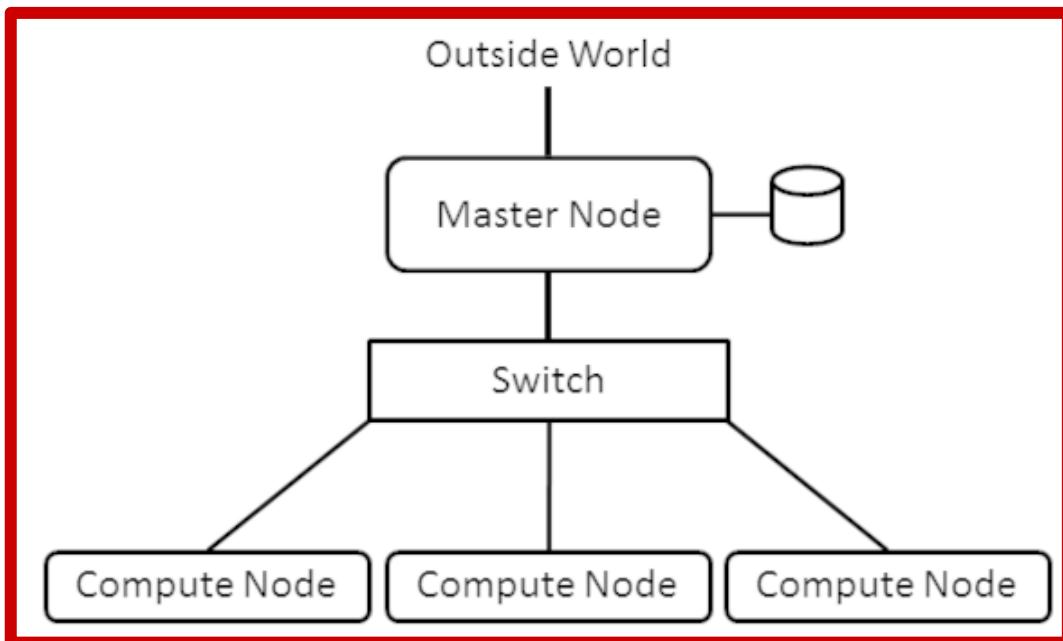
# Архитектура вычислительного кластера: вычислительное поле



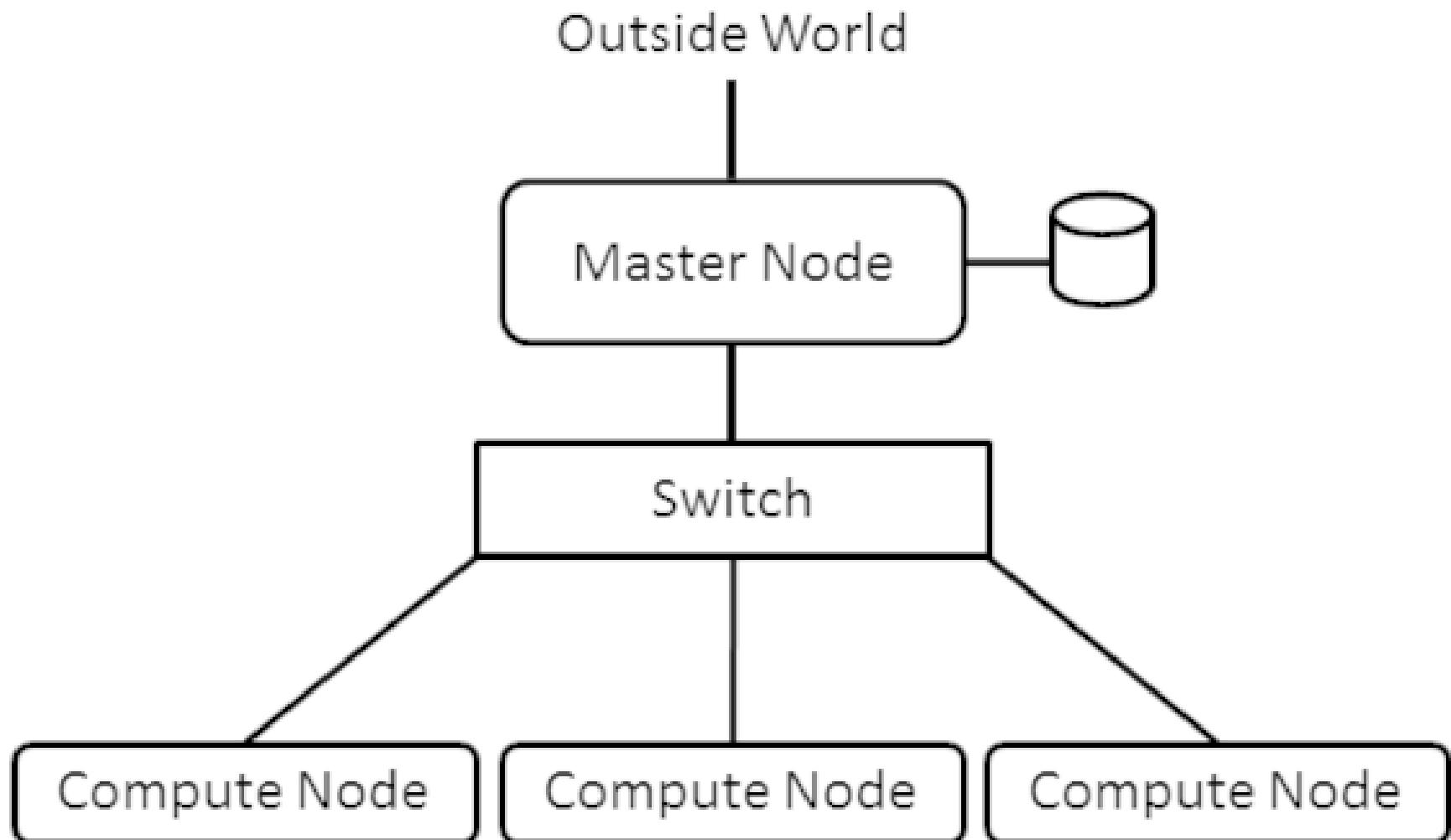
# Архитектура вычислительного кластера



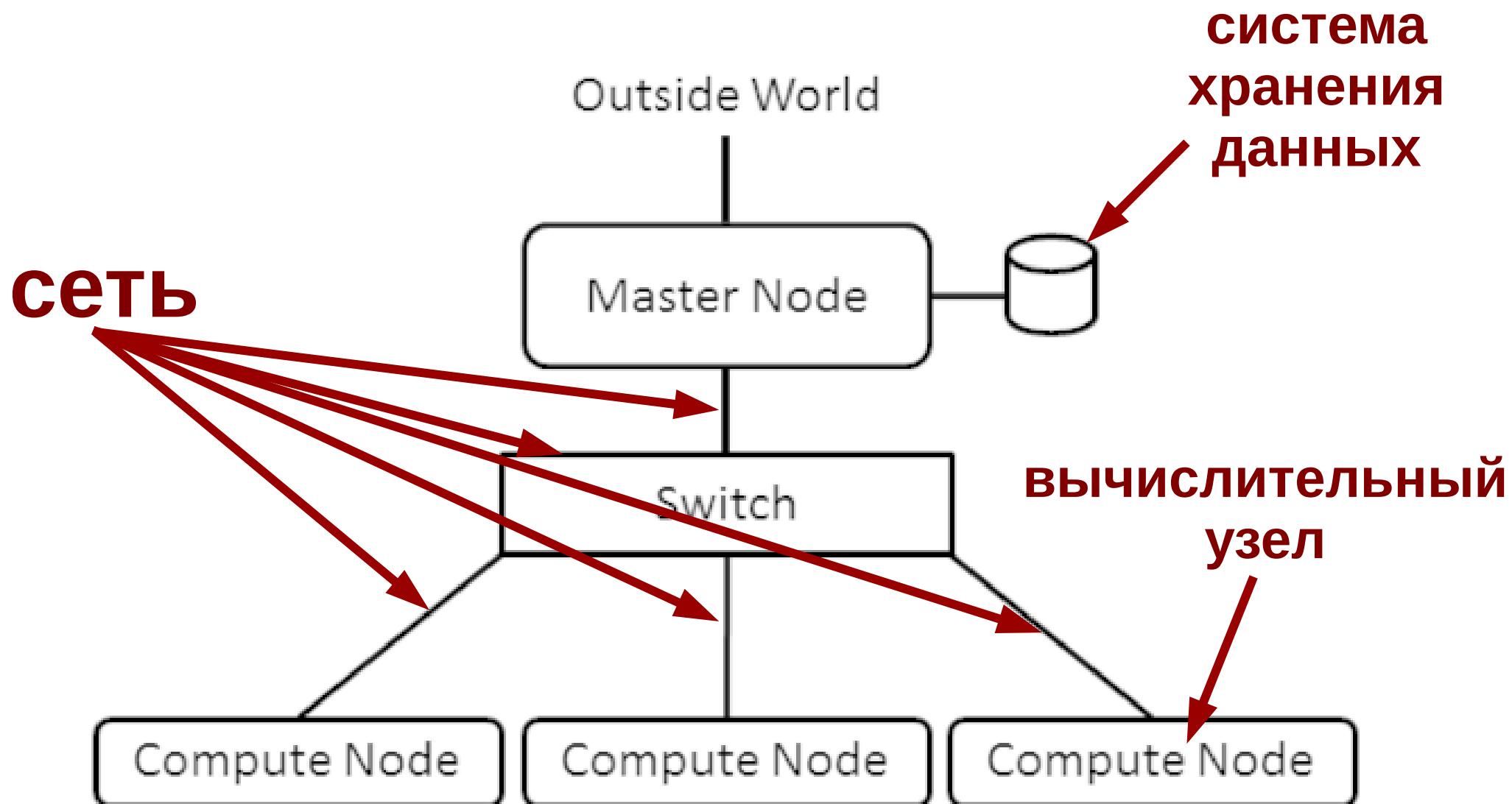
# Архитектура вычислительного кластера



# С аппаратной точки зрения: основные моменты

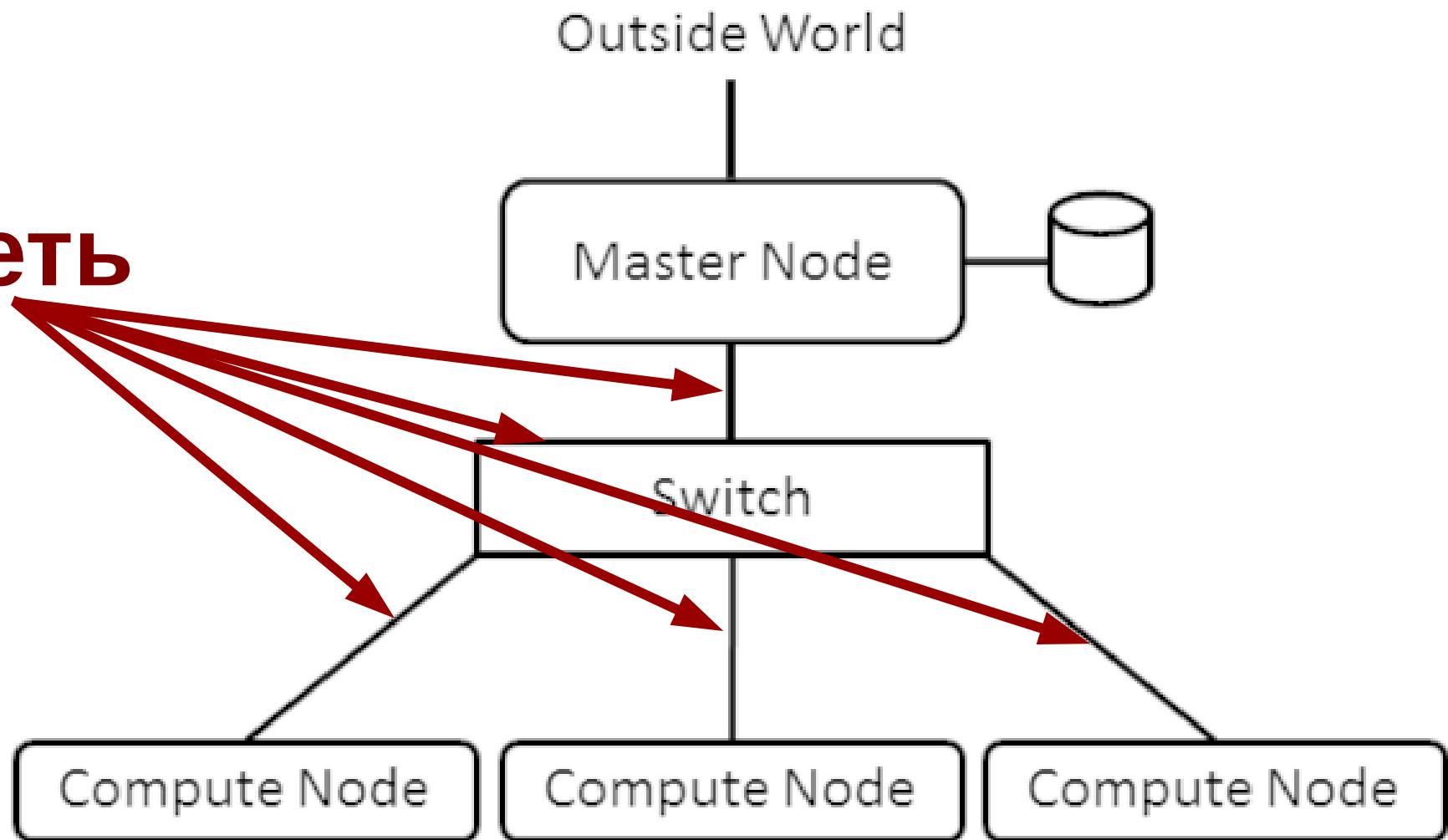


# С аппаратной точки зрения: основные моменты

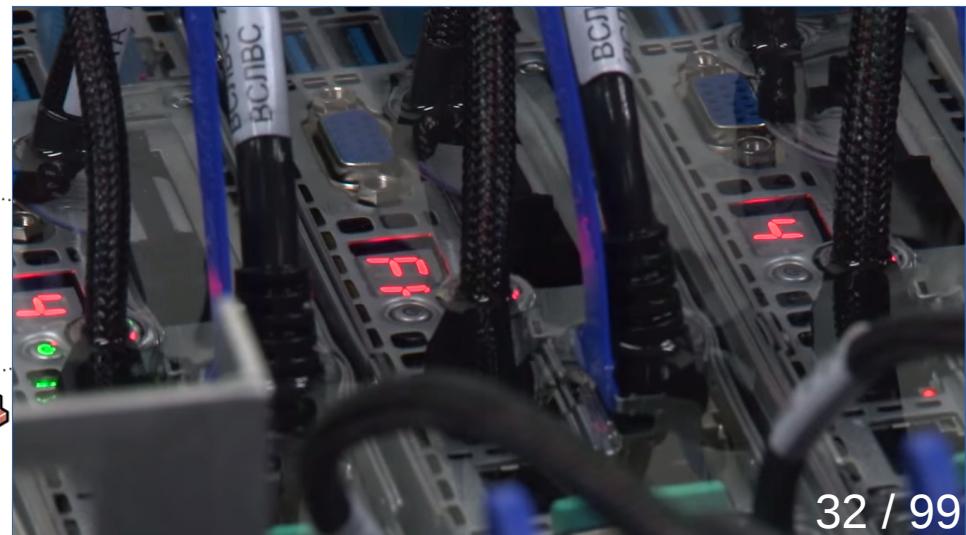
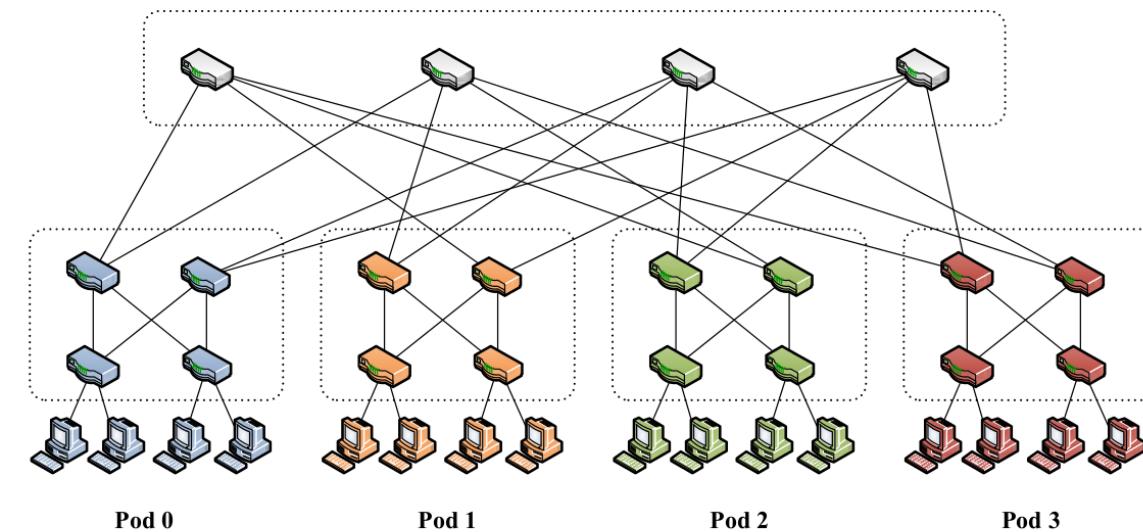
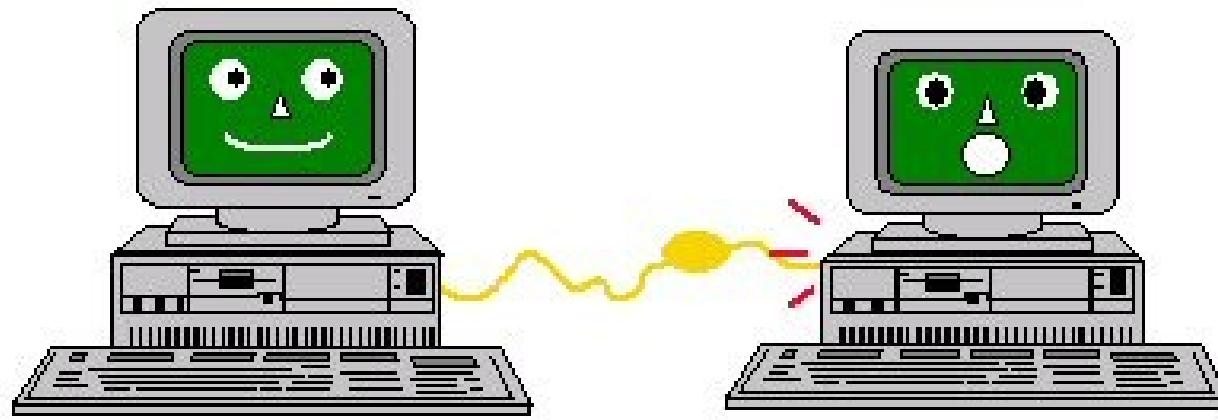


# С аппаратной точки зрения: основные моменты

**сеть**



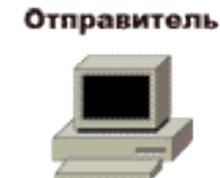
# С аппаратной точки зрения: сеть



# С аппаратной точки зрения: сеть

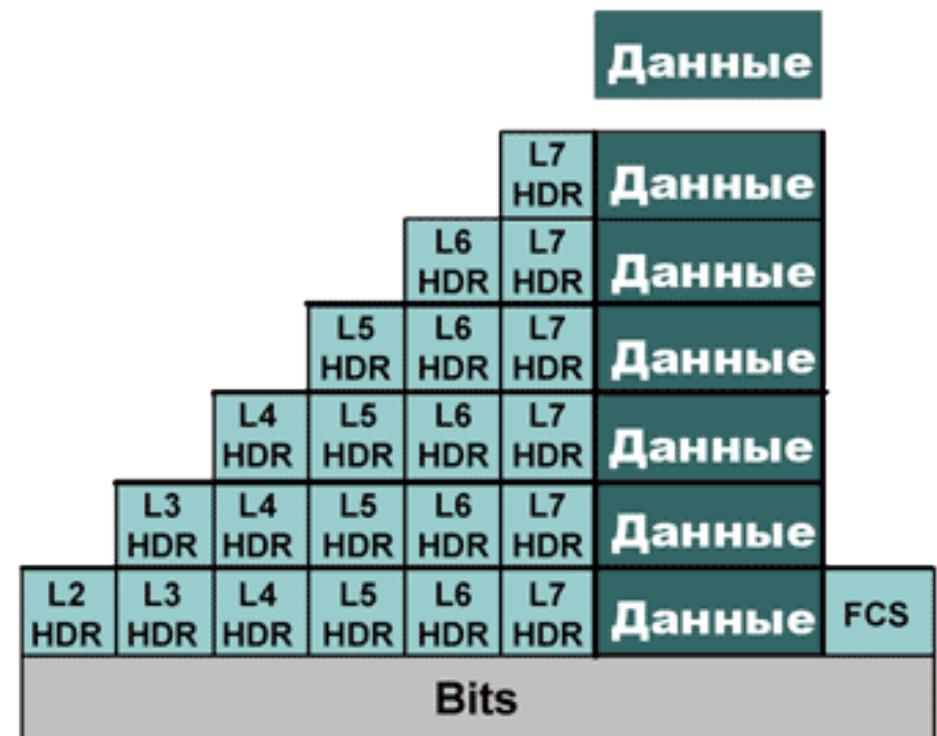
TCP → IP → Ethernet:

- IP
- TCP
- network syscalls



## InfiniBand:

- Простой протокол, быстрый
- RDMA
- IPoIB



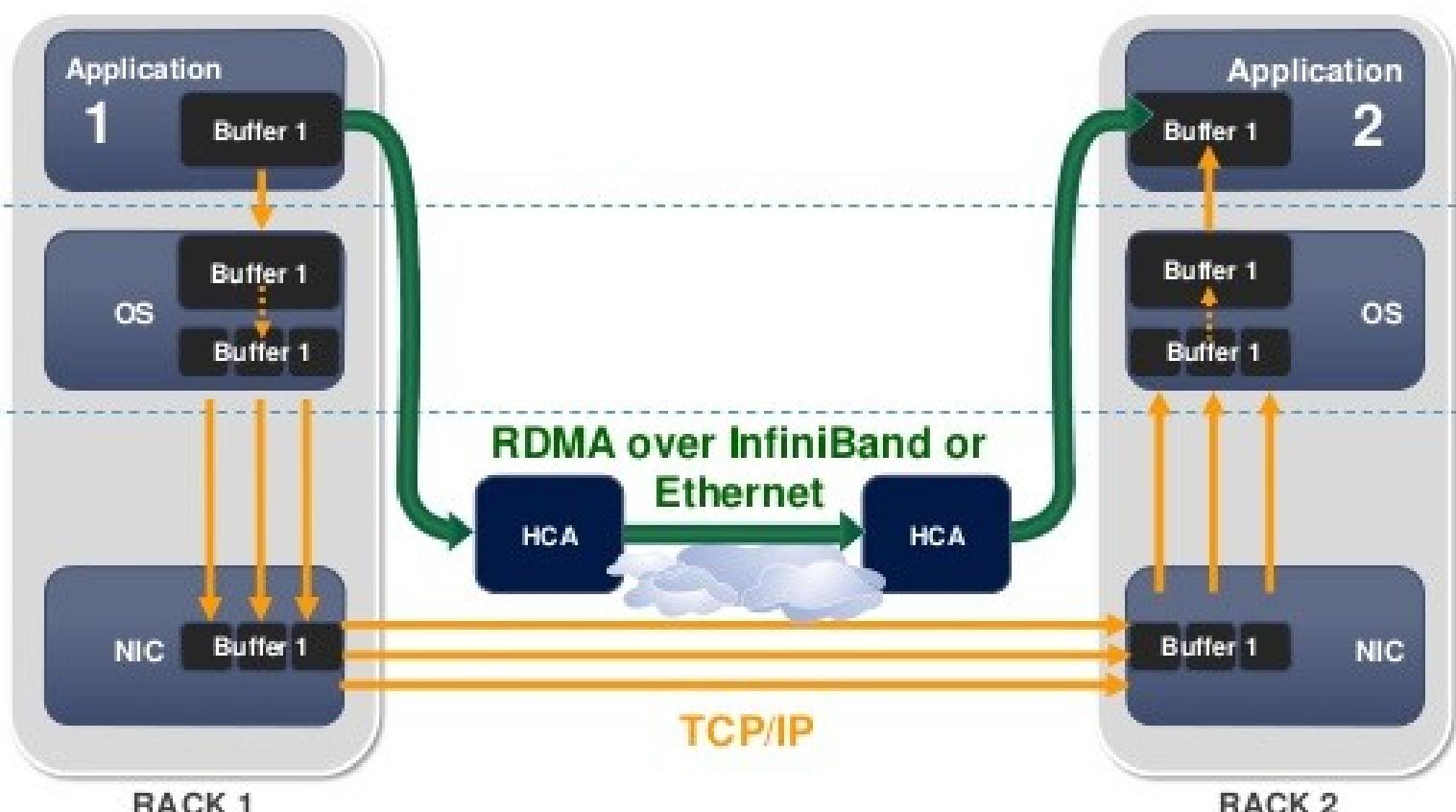
HDR = Заголовок

# RDMA

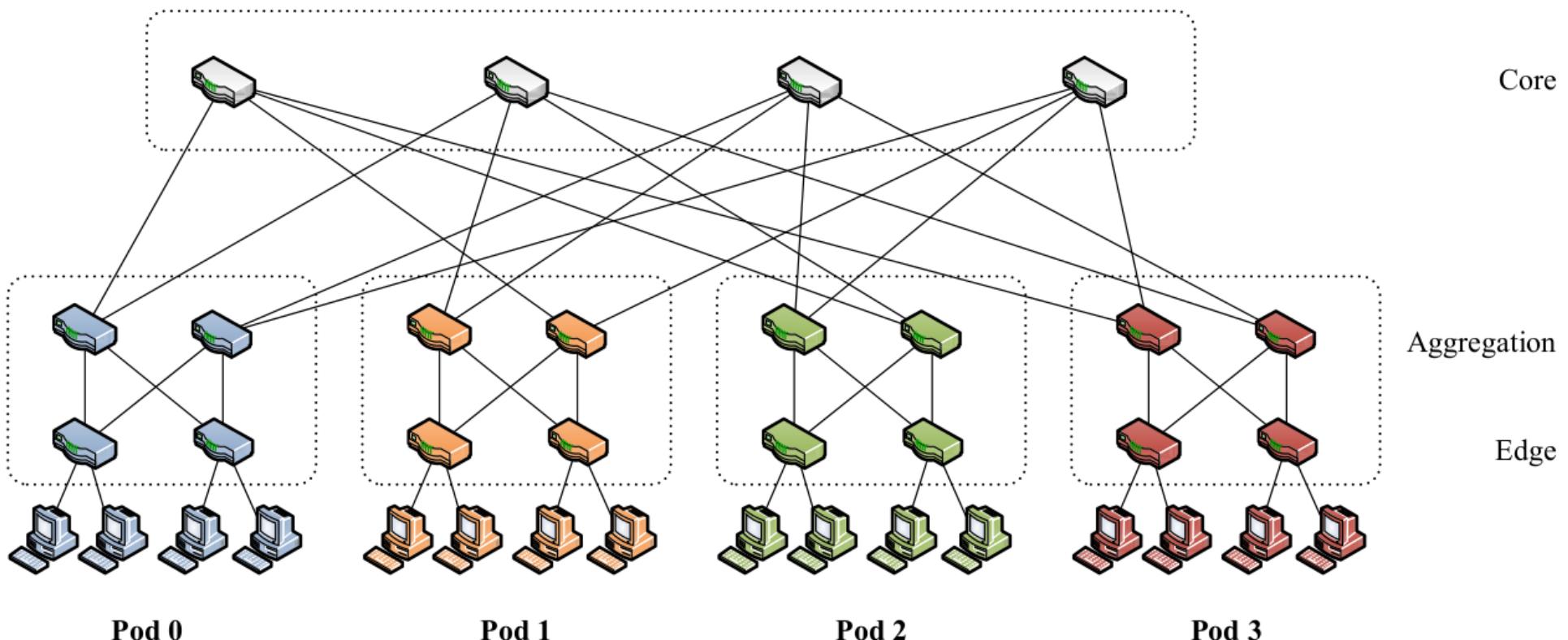
## RDMA/RoCE I/O Offload



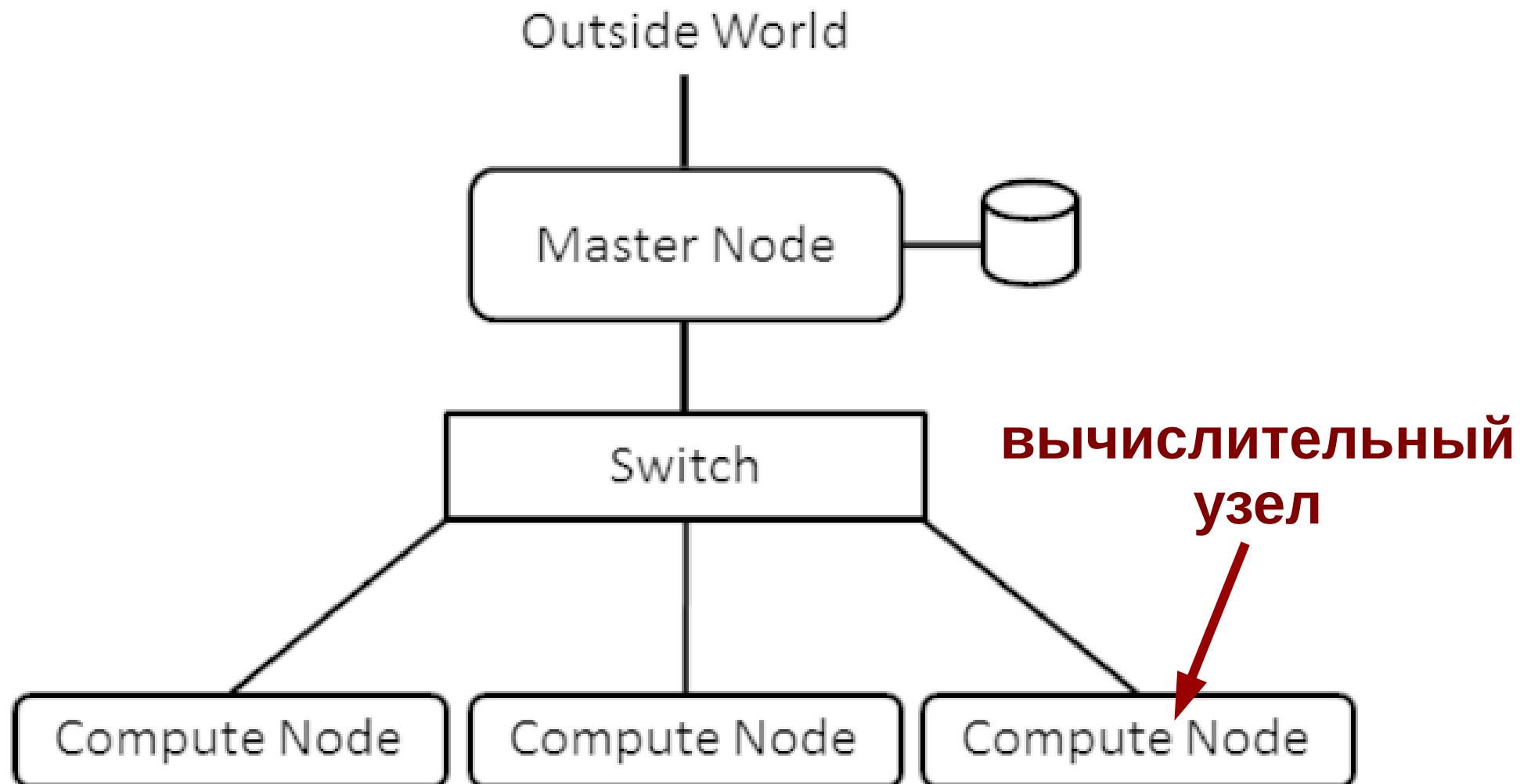
USER  
KERNEL  
HARDWARE



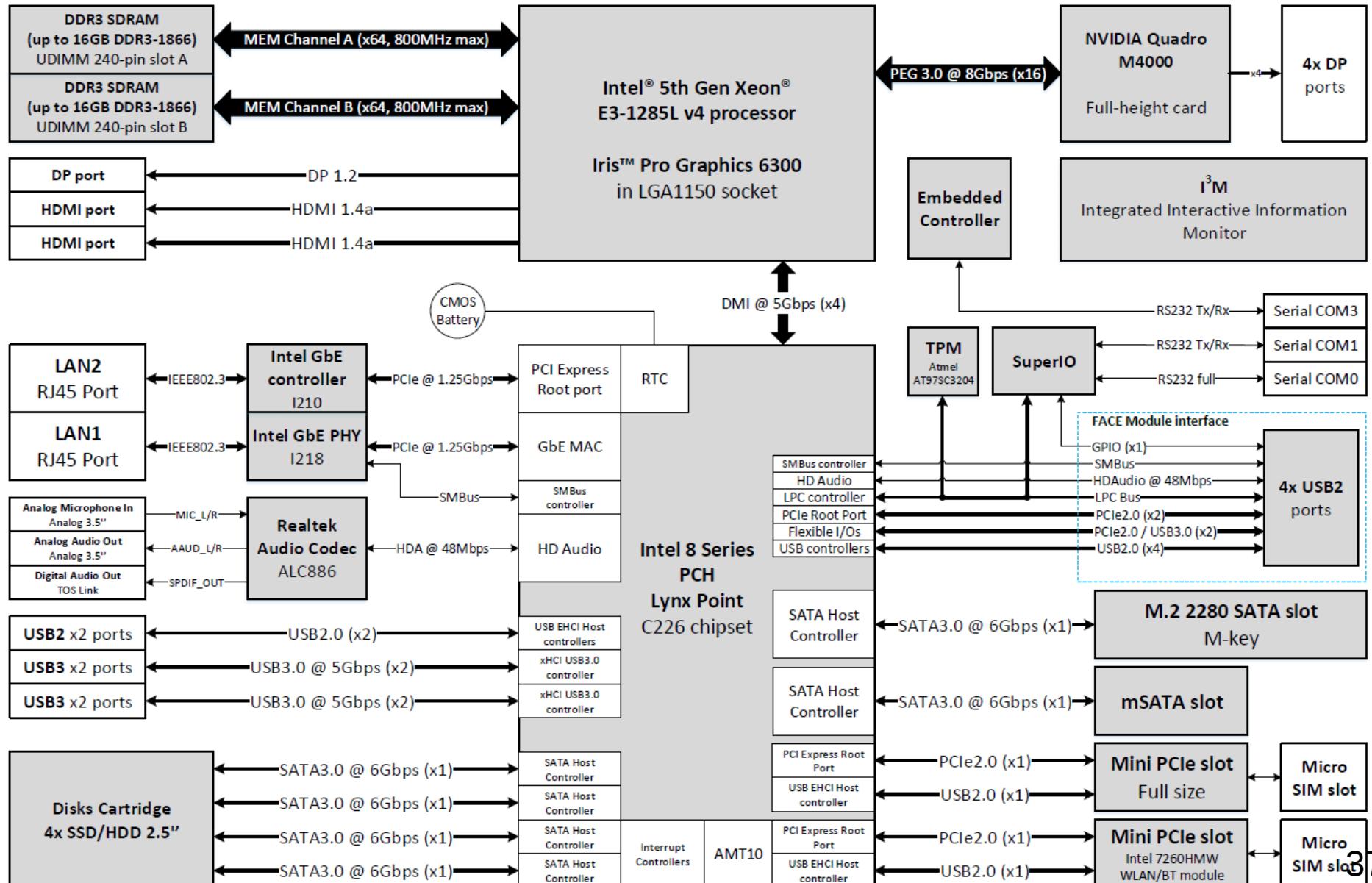
# С аппаратной точки зрения: InfiniBand



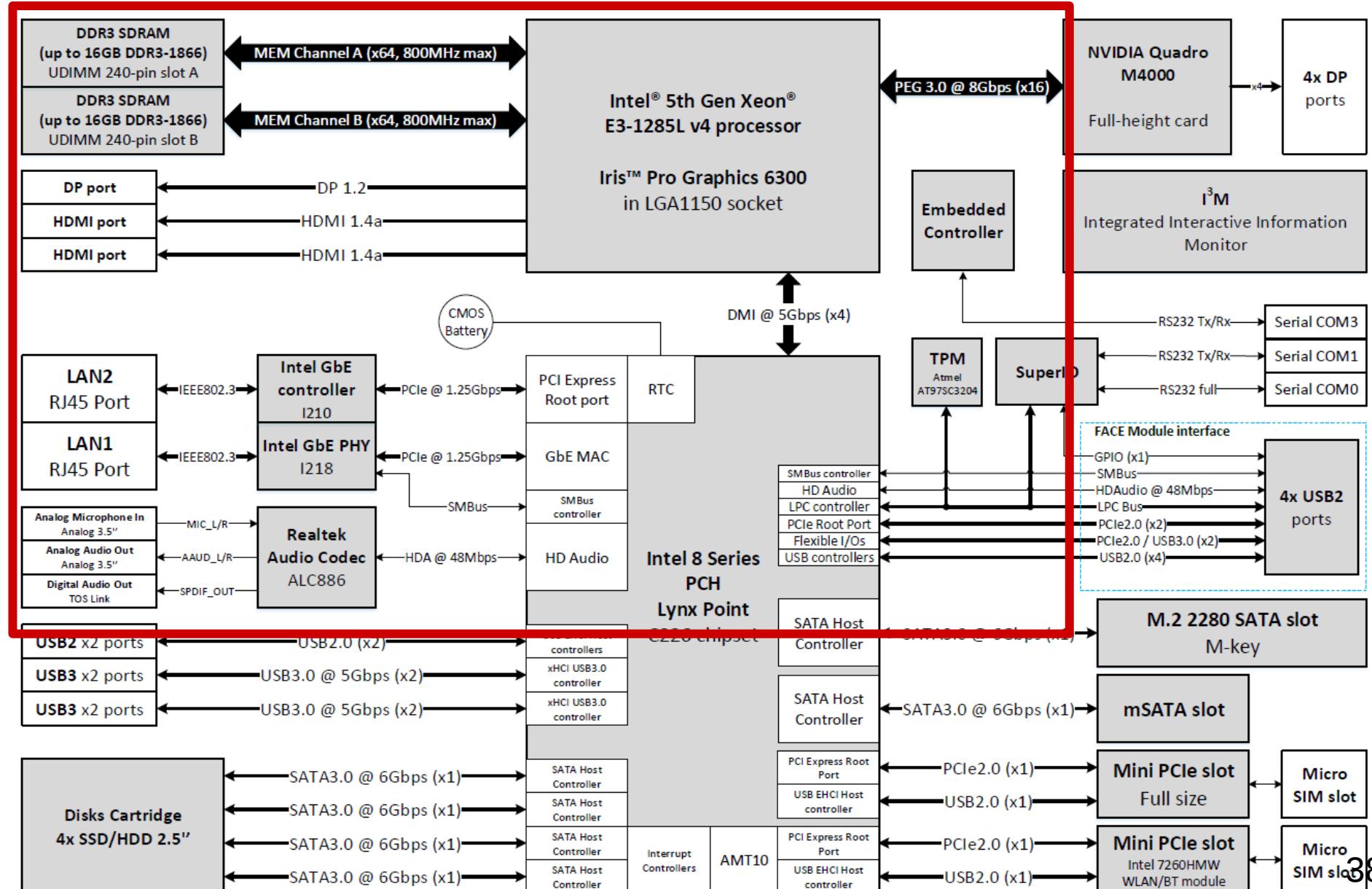
# С аппаратной точки зрения: основные моменты



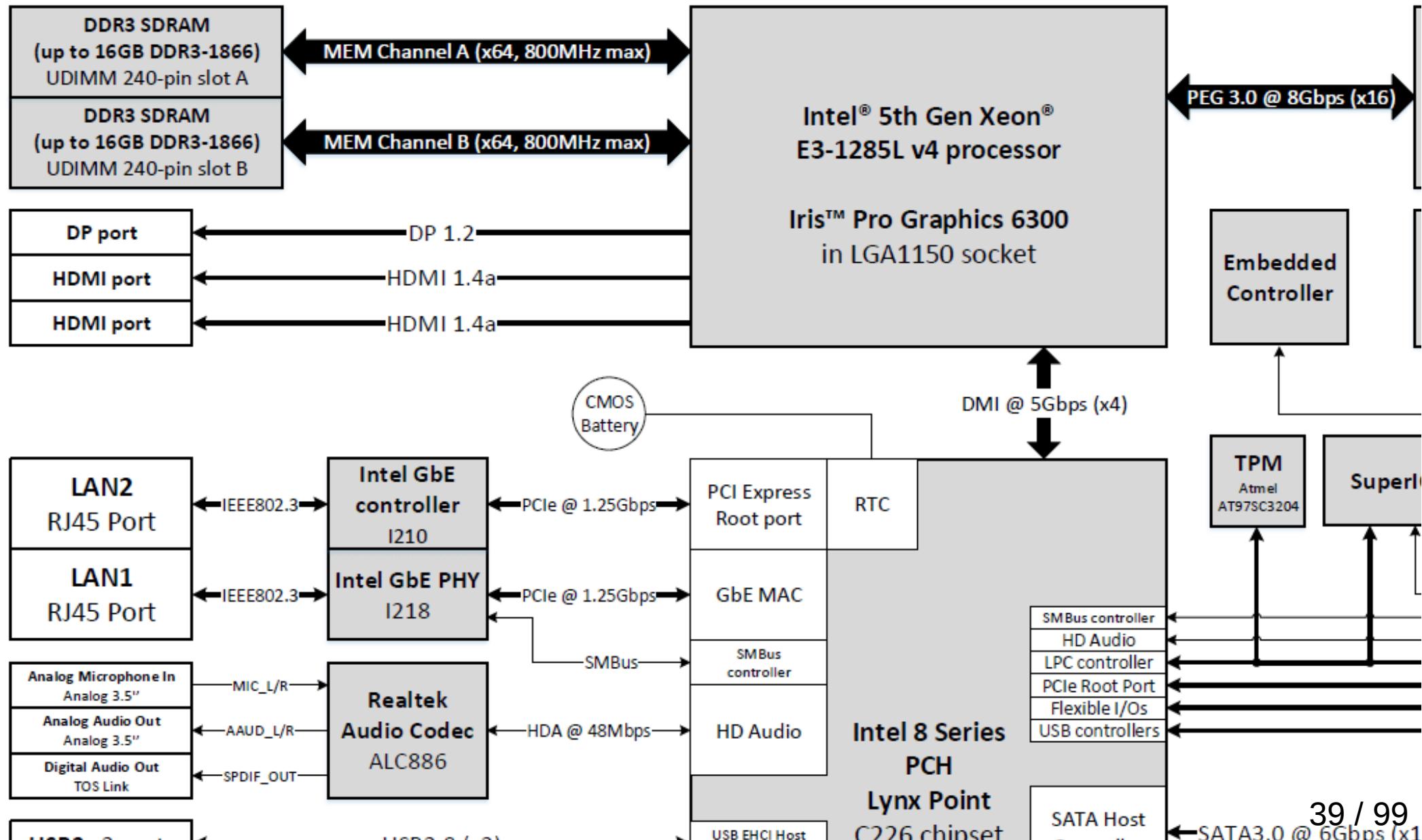
# С аппаратной точки зрения: вычислительный узел



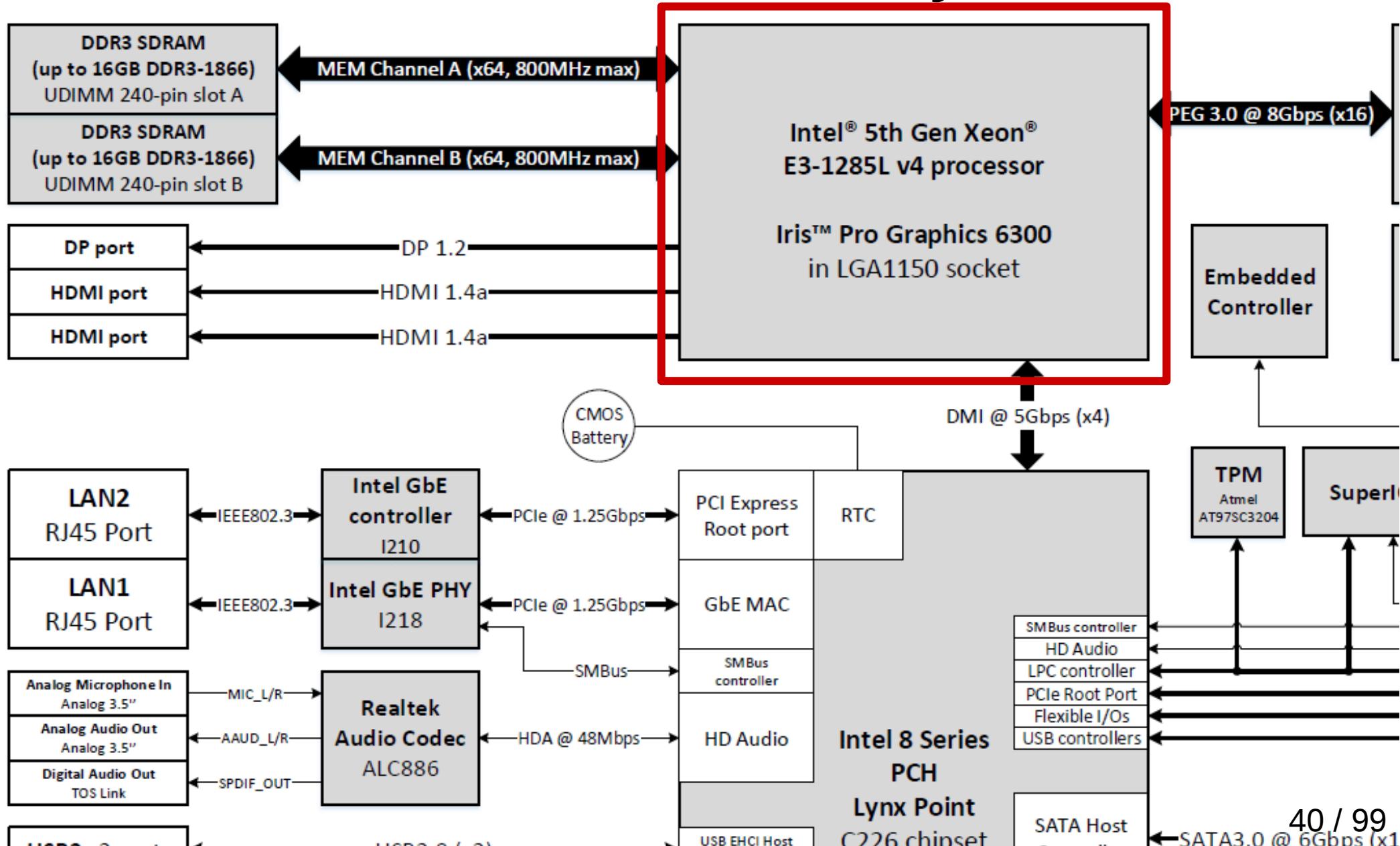
# С аппаратной точки зрения: вычислительный узел



# С аппаратной точки зрения: вычислительный узел

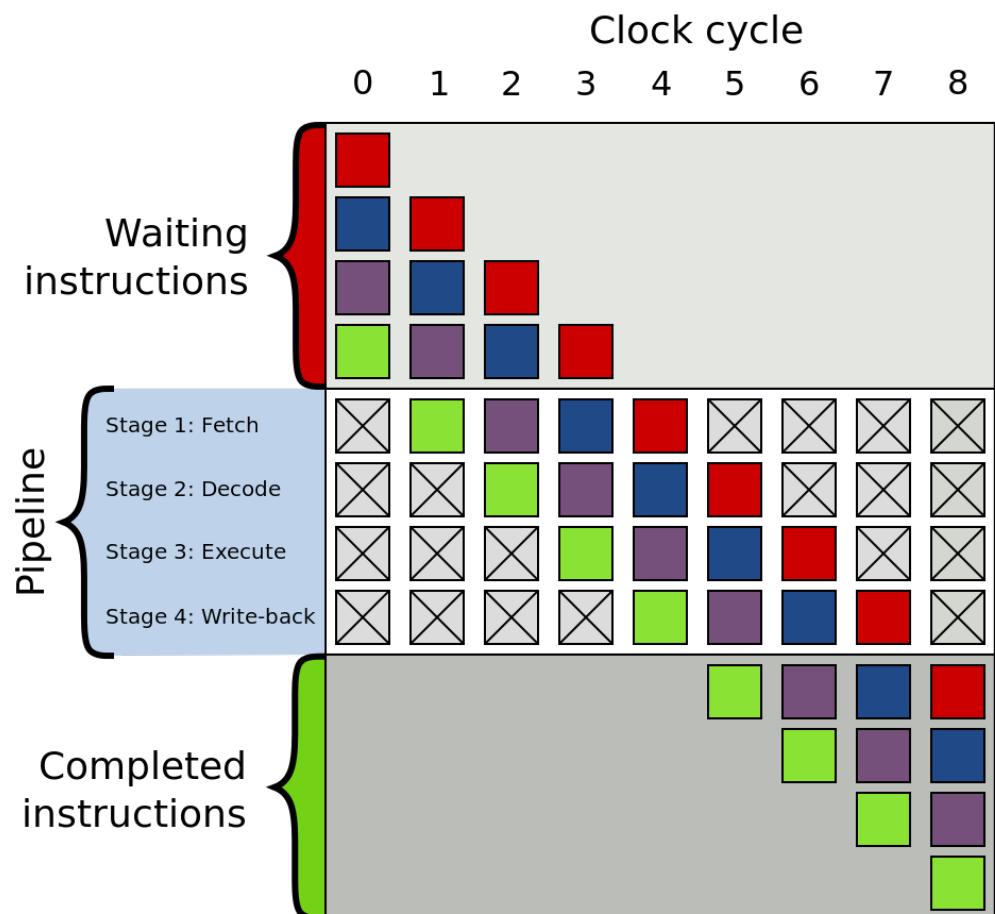
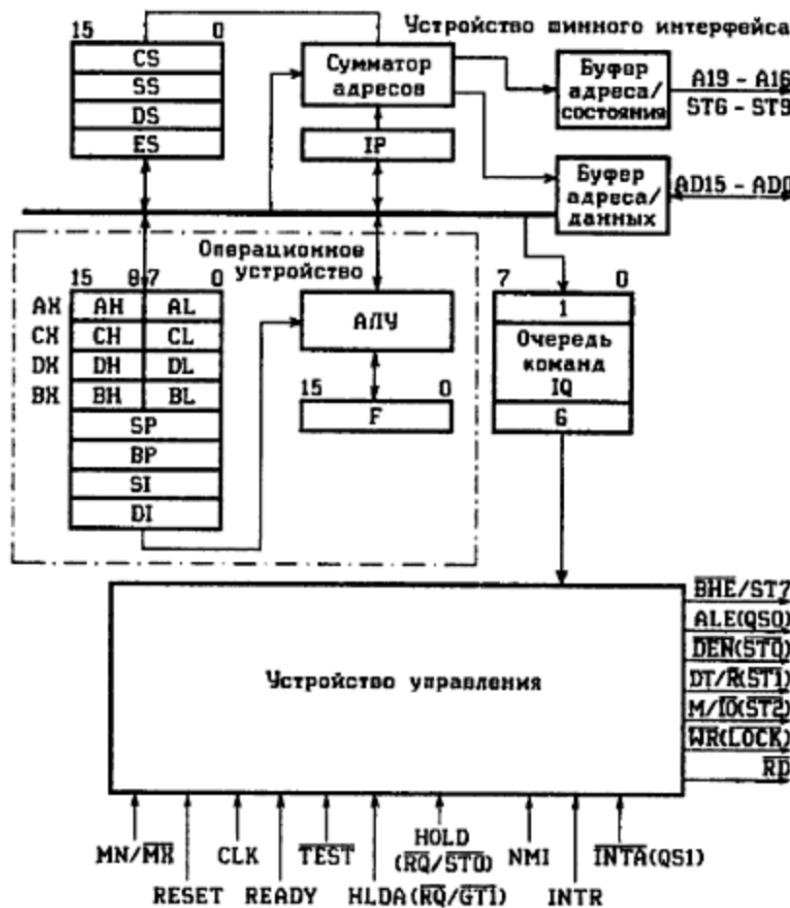


# С аппаратной точки зрения: вычислительный узел

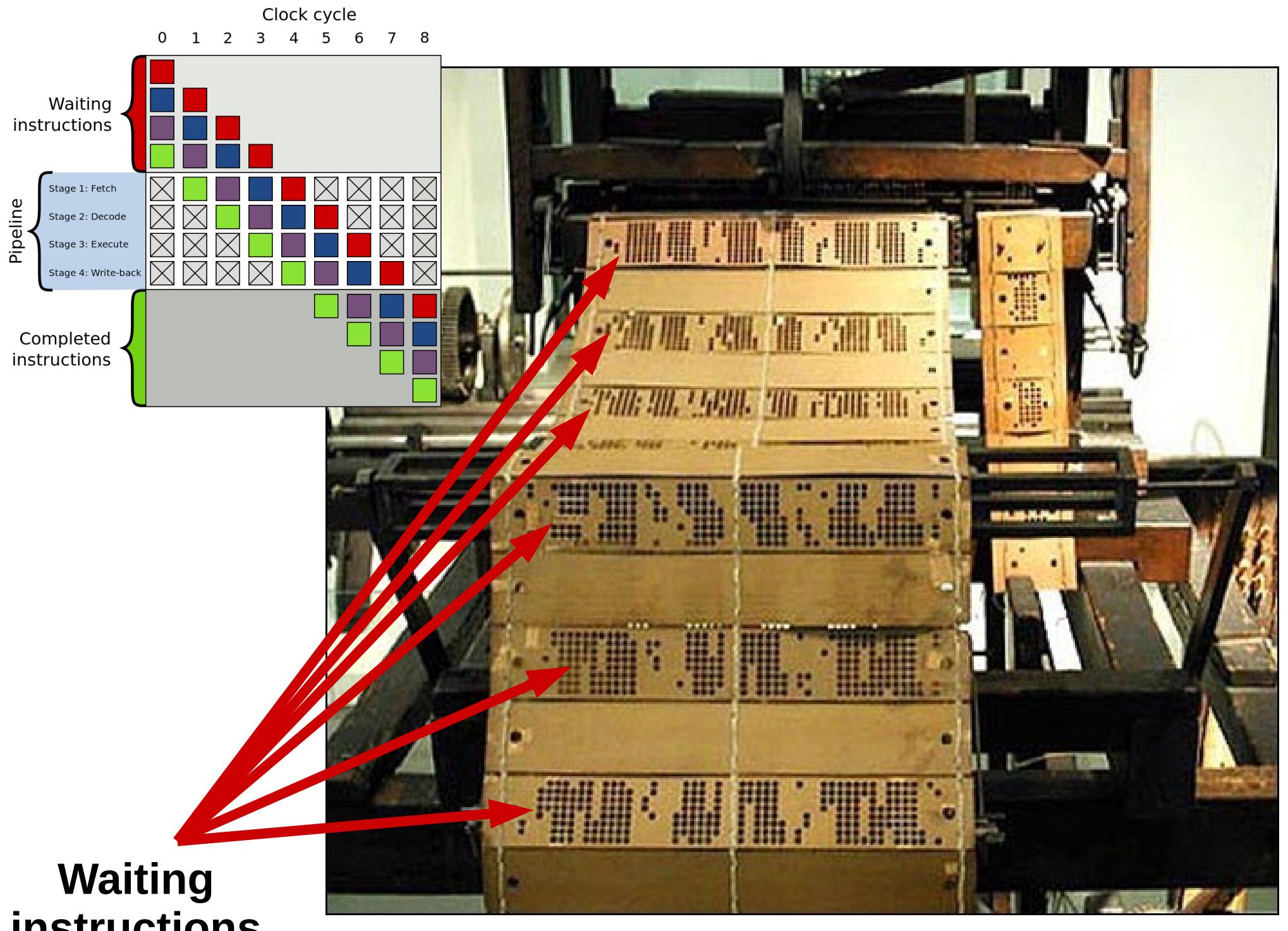


# Архитектура процессоров x86

- 8086, 8088, 80186: real mode

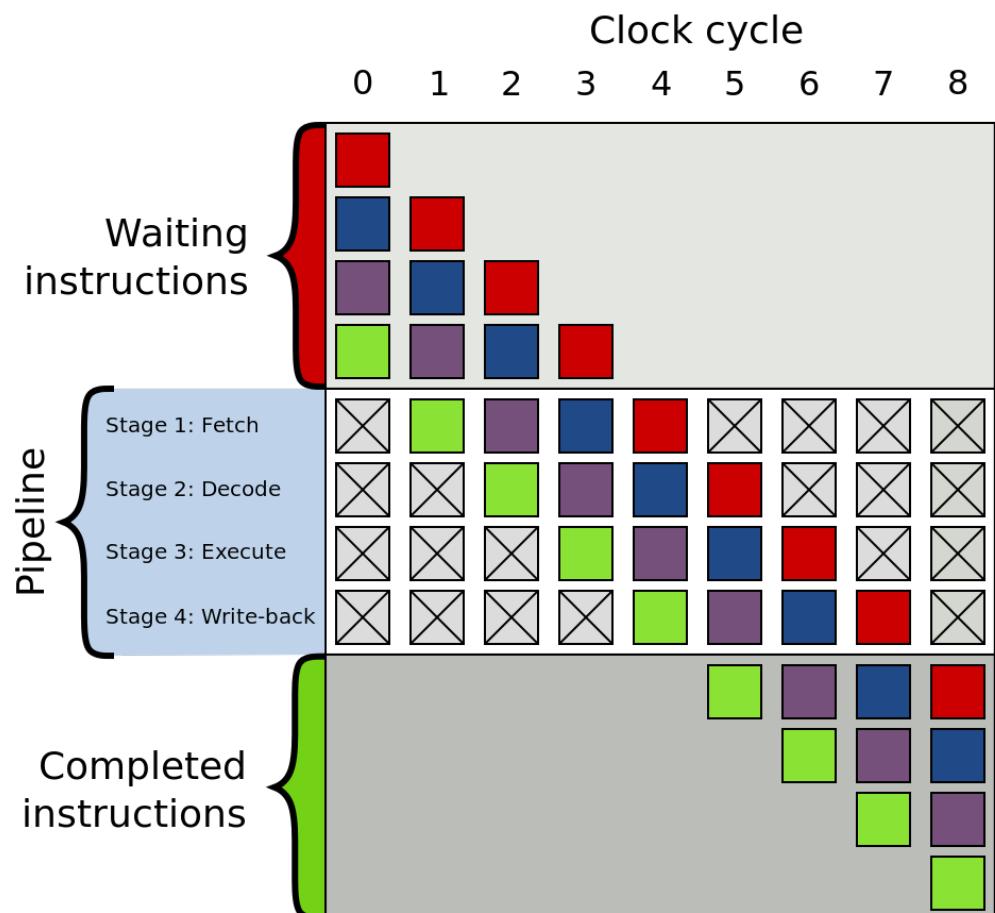
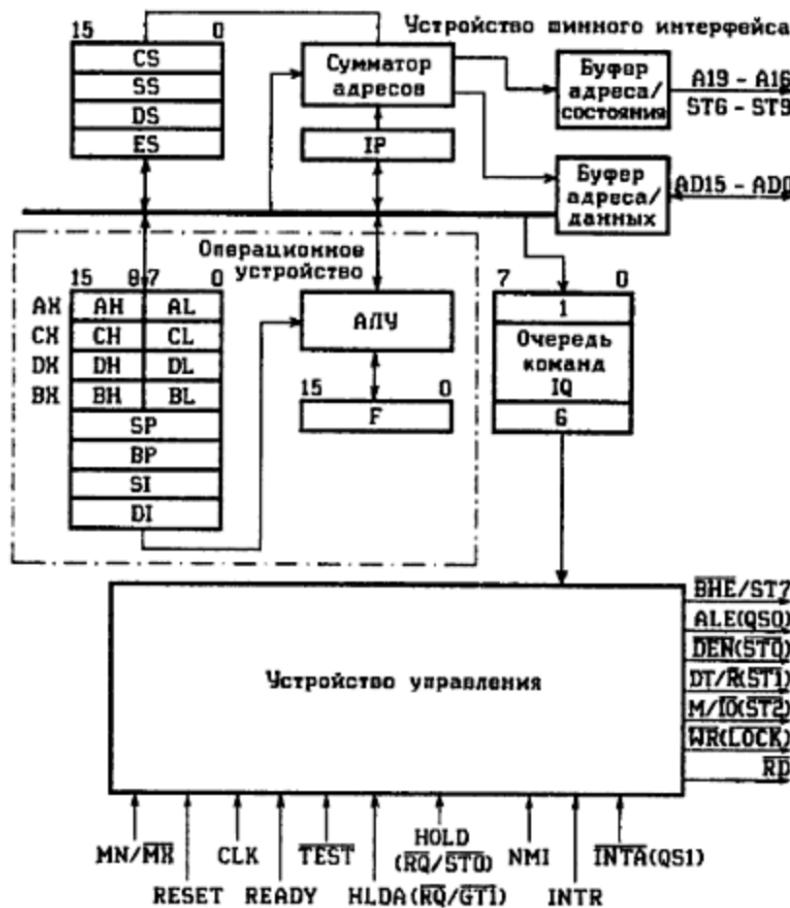






# Архитектура процессоров x86

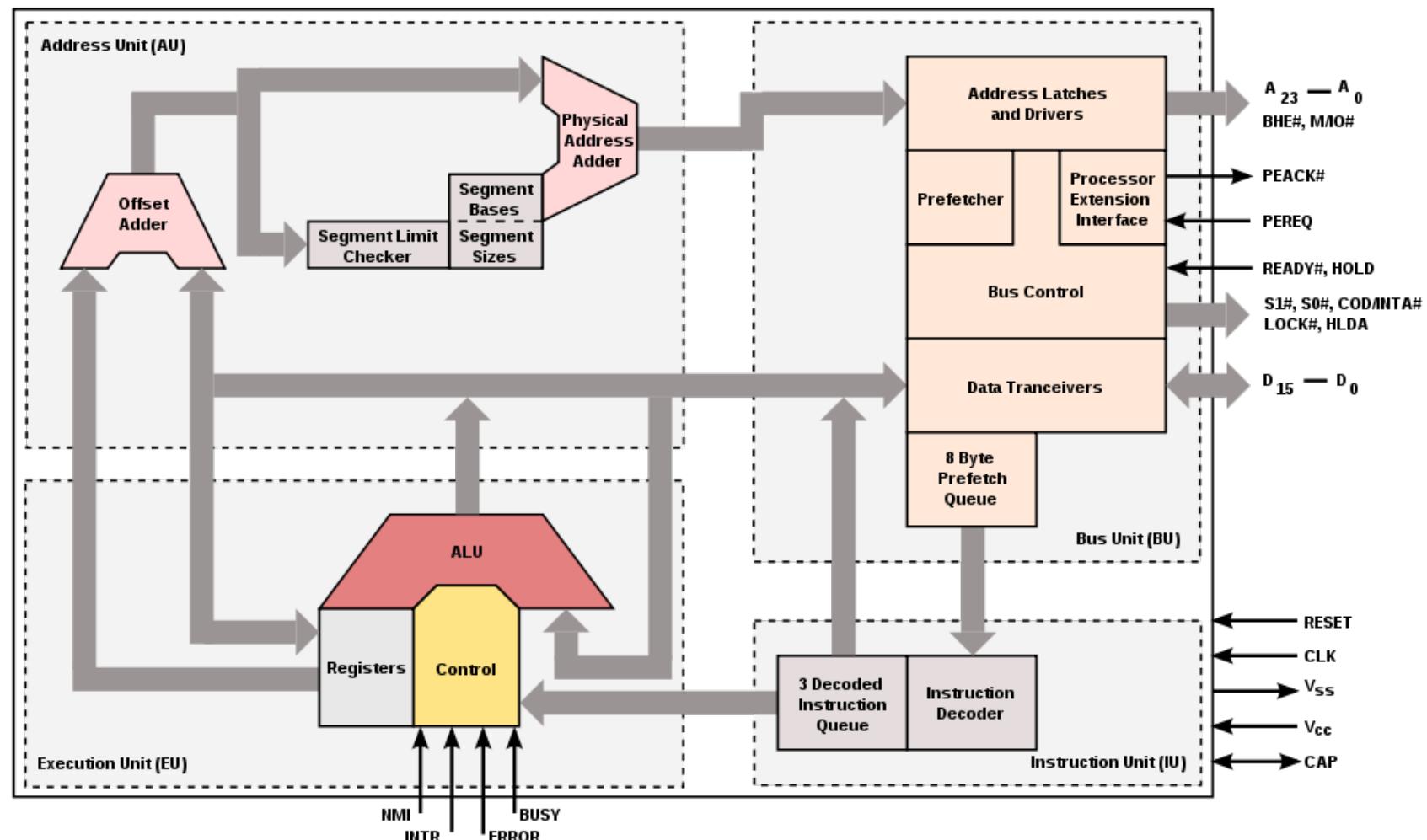
- 8086, 8088, 80186: real mode



# Архитектура процессоров x86

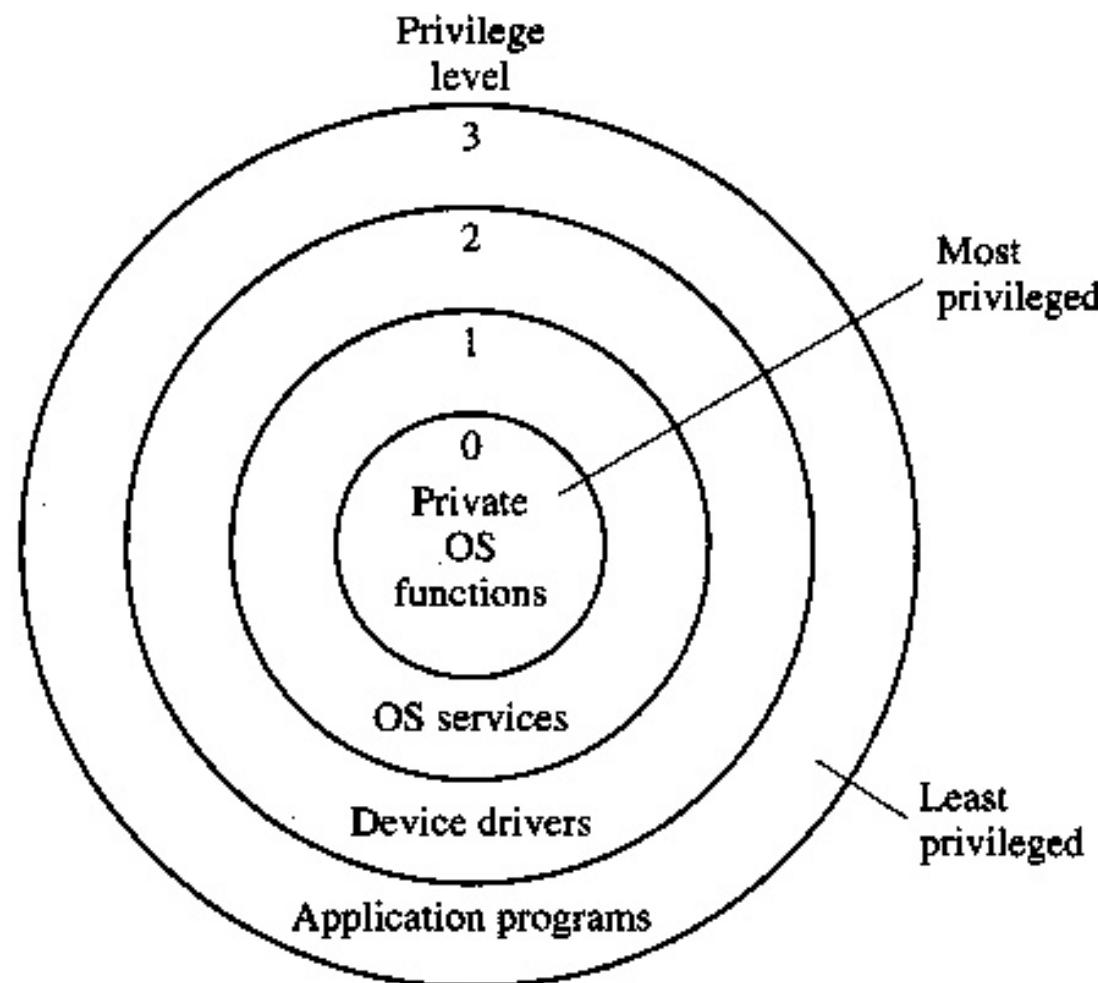
- 80286: protected mode, virtual memory, 20MHz

Intel 80286 architecture



# Архитектура процессоров x86

- 80286: protected mode, virtual memory, 20MHz



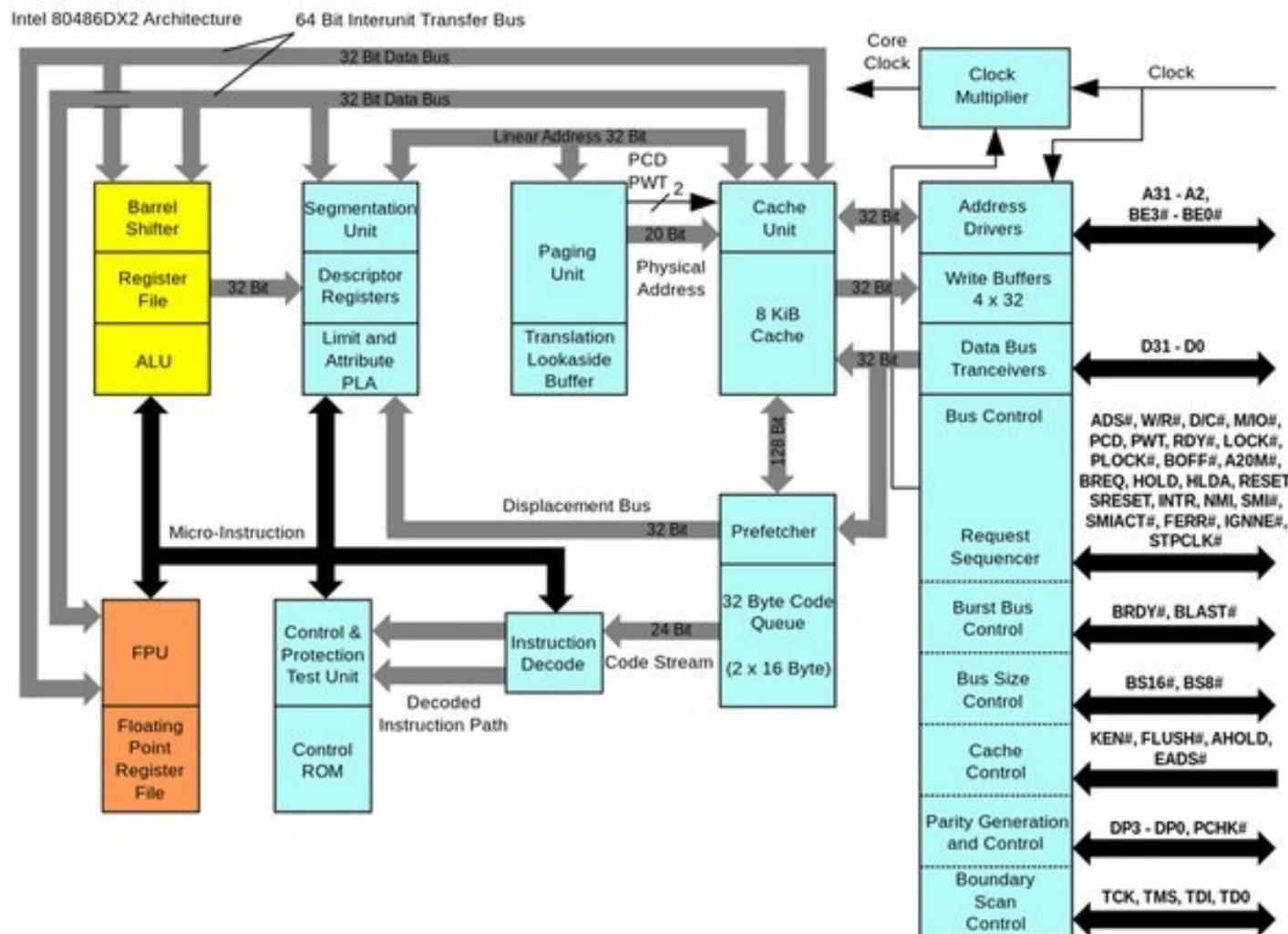
# Архитектура процессоров x86

- 80386 (i386): 32bit, ext. cache, 40MHz



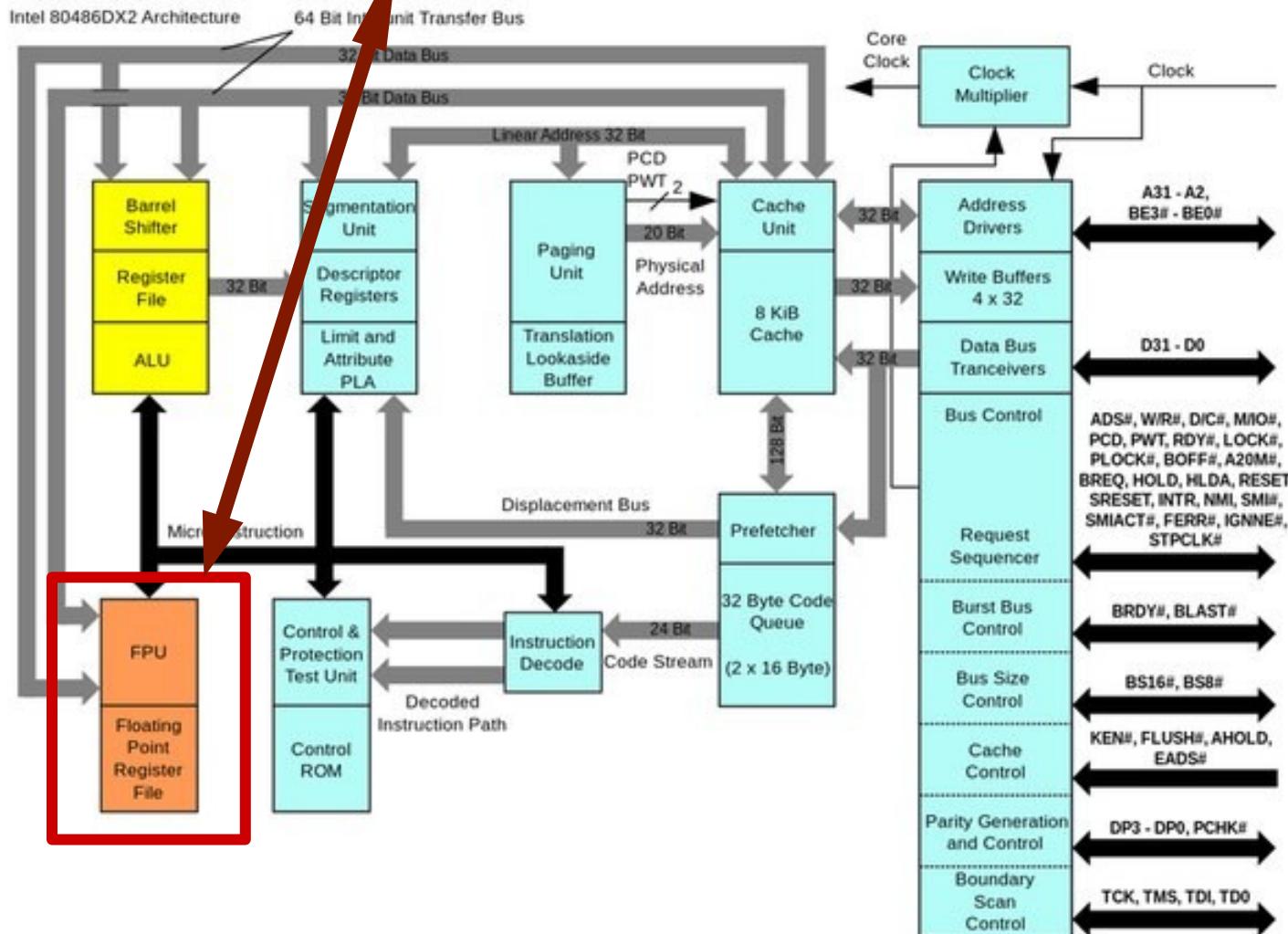
# Архитектура процессоров x86

- 80486 (i486): FPU, int. cache, FSB multiplier



# Архитектура процессоров x86

- 80486 (i486): **FPU** int. cache, FSB multiplier



# Floating Point Unit

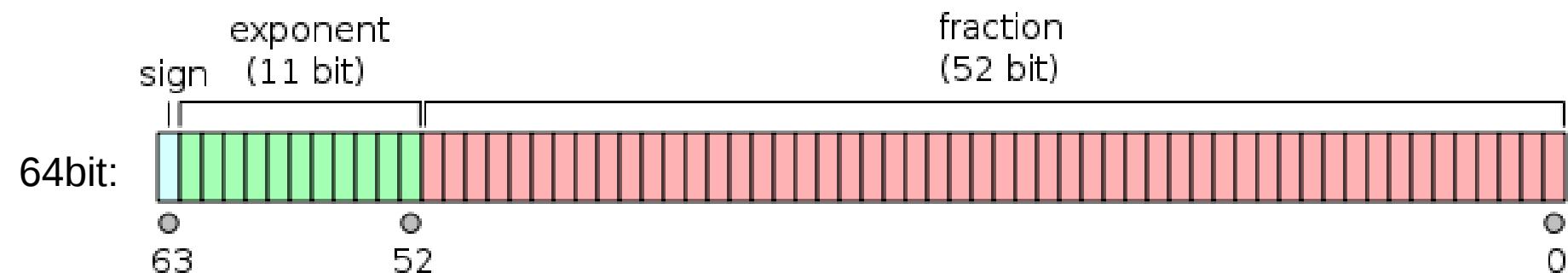
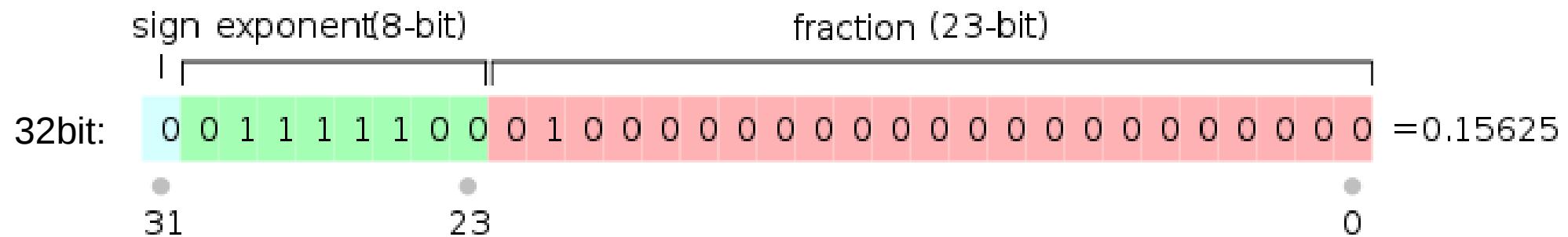
- Реальные физические значения (скорость, масса, давление и т. п.) не ограничиваются целыми числами ( $2\frac{m}{c}$ ,  $3\frac{m}{c}$ ; но не  $2.5\frac{m}{c}$ ).  
Требуется возможность работы с действительными числами.
- В качестве аналога действительных чисел в компьютере используются floating point numbers (числа с плавающей точкой).

# Floating Point Unit

- В распространённых компьютерах процессор работает с двоичными значениями. Каким образом представить нецелое число в двоичной форме?
- Требуется:
  - Большой диапазон значений
  - Простота реализации процессора (и ПО)
  - Высокая скорость выполнения арифм. операций
- В стандарте IEEE754 закреплено следующее представление: ...[след. слайд]...

# Floating Point Unit

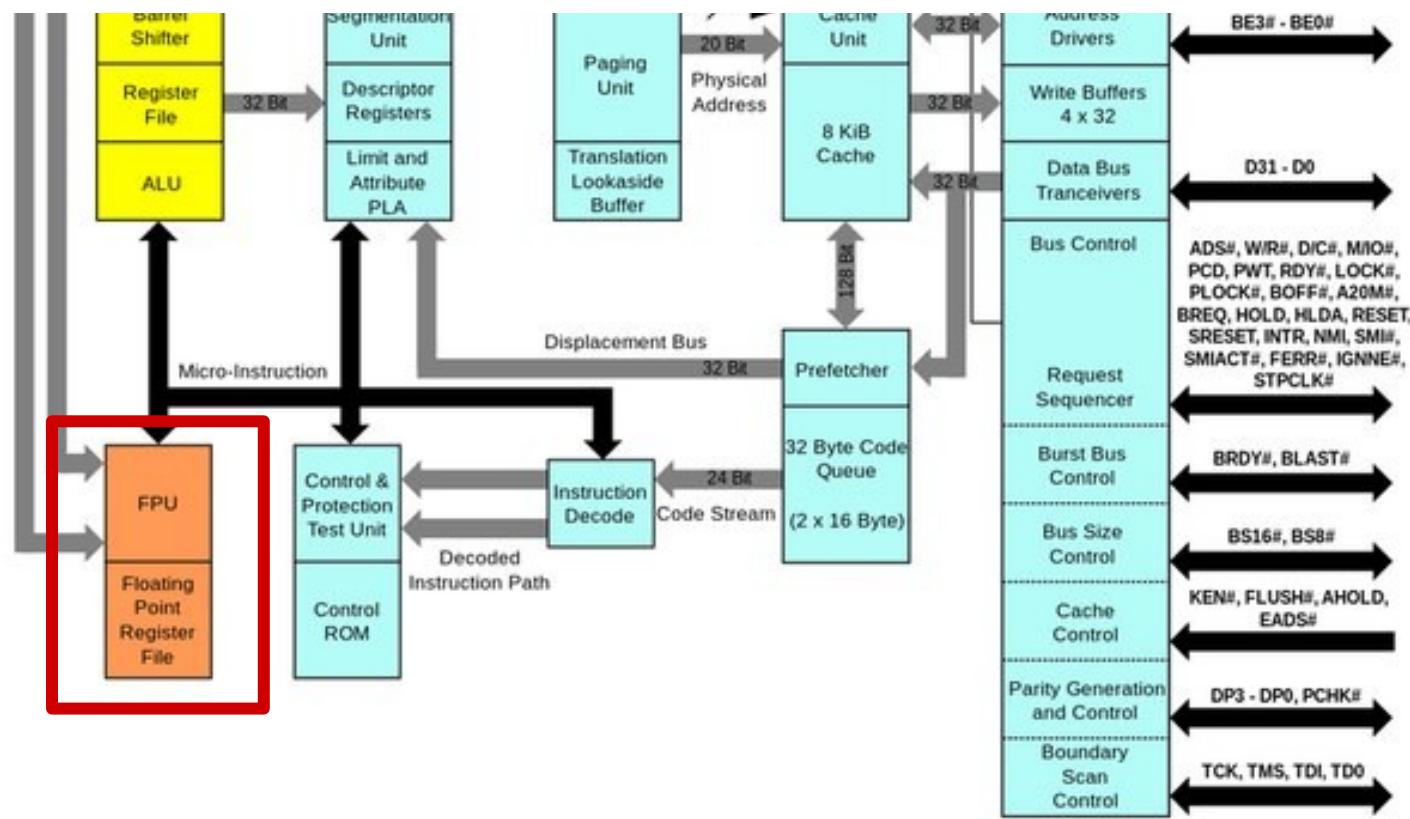
- В стандарте IEEE754 закреплено следующее представление:



significand  $\times$  base<sup>exponent</sup>

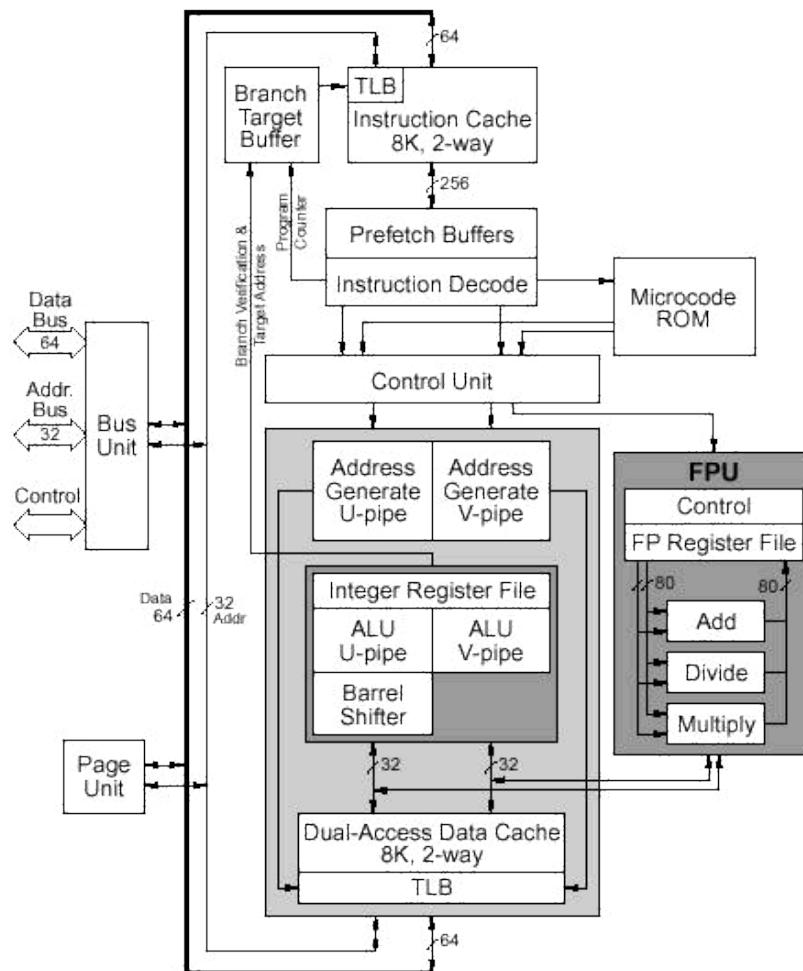
# Floating Point Unit

- FPU — это специальный модуль, выполняющий операции над числами, представленными в соответствии с IEEE754



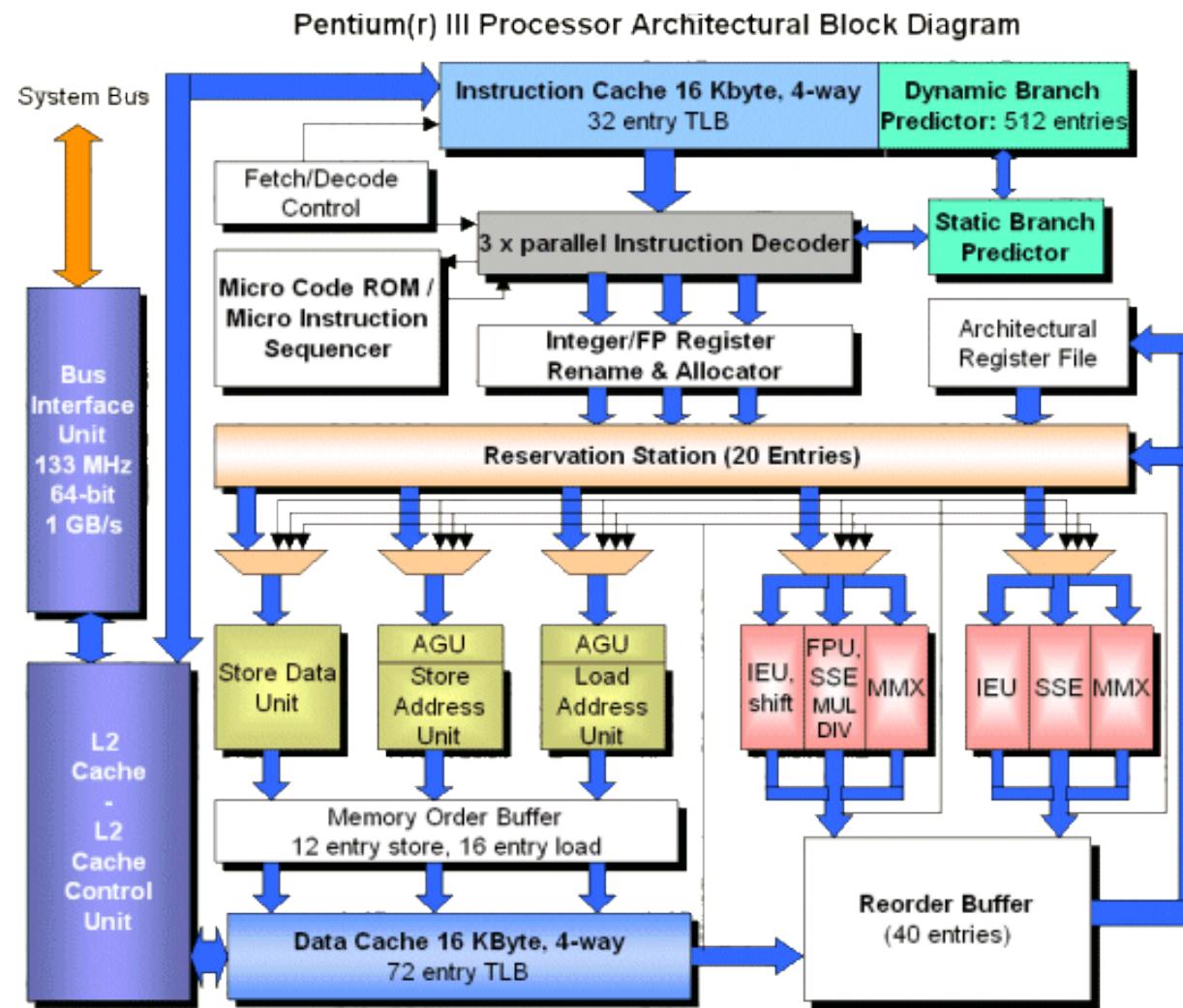
# Архитектура процессоров x86

- i586 (Pentium MMX): MMX, superpipelining



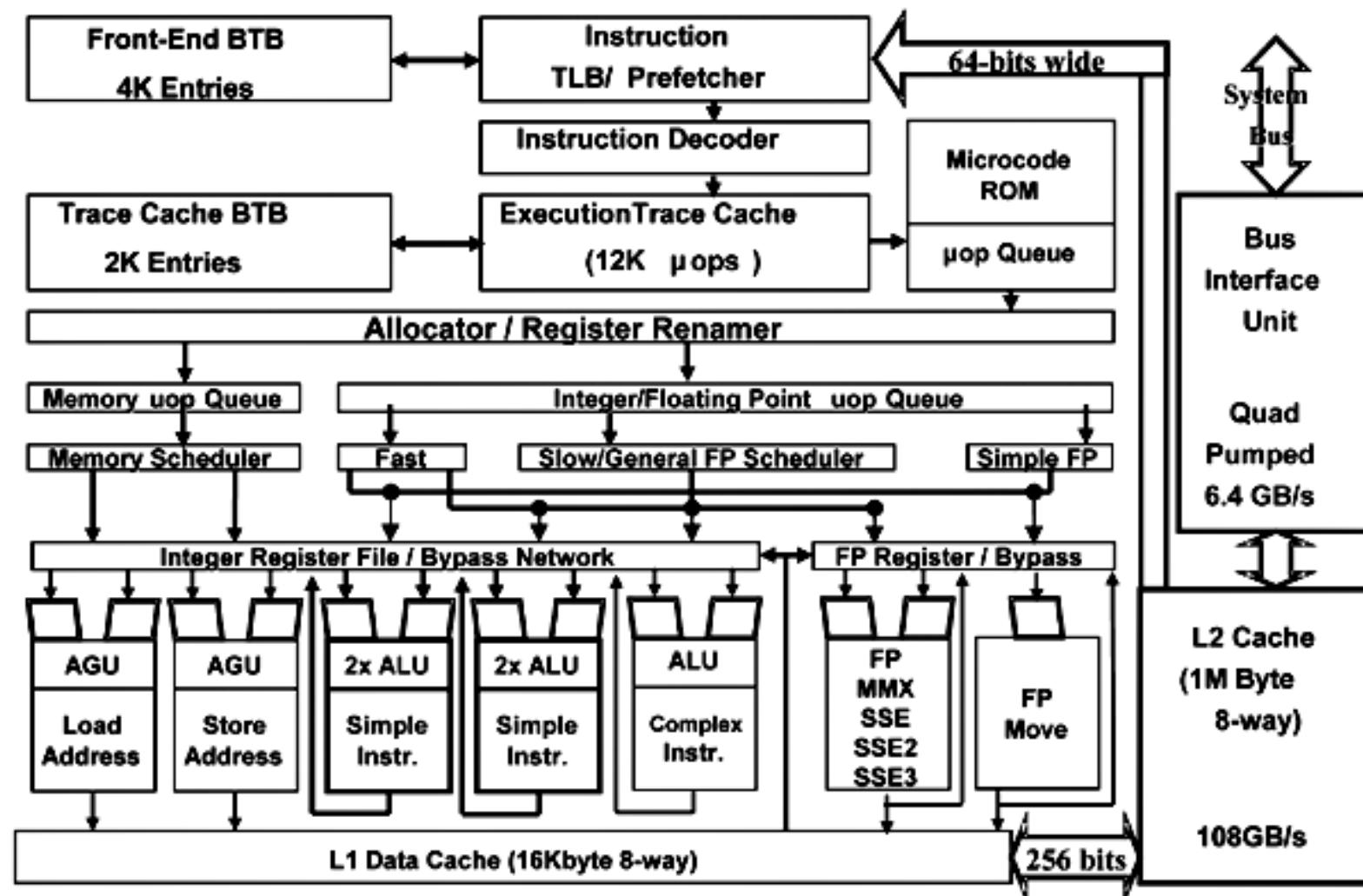
# Архитектура процессоров x86

- Pentium 3 (i686): RISC, int L2 cache, more SIMD



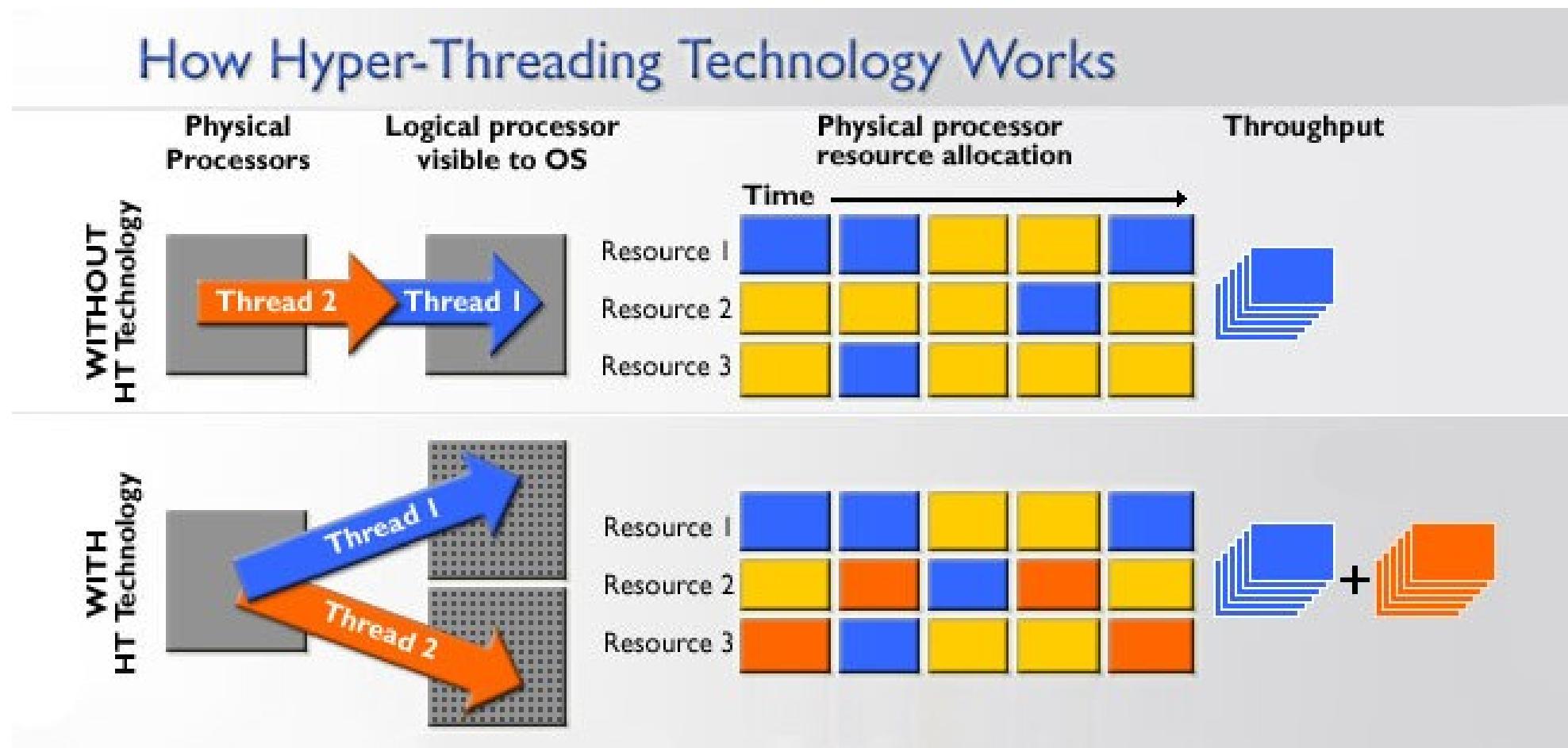
# Архитектура процессоров x86

- Pentium 4: Hyper-threading, hyperpipelining

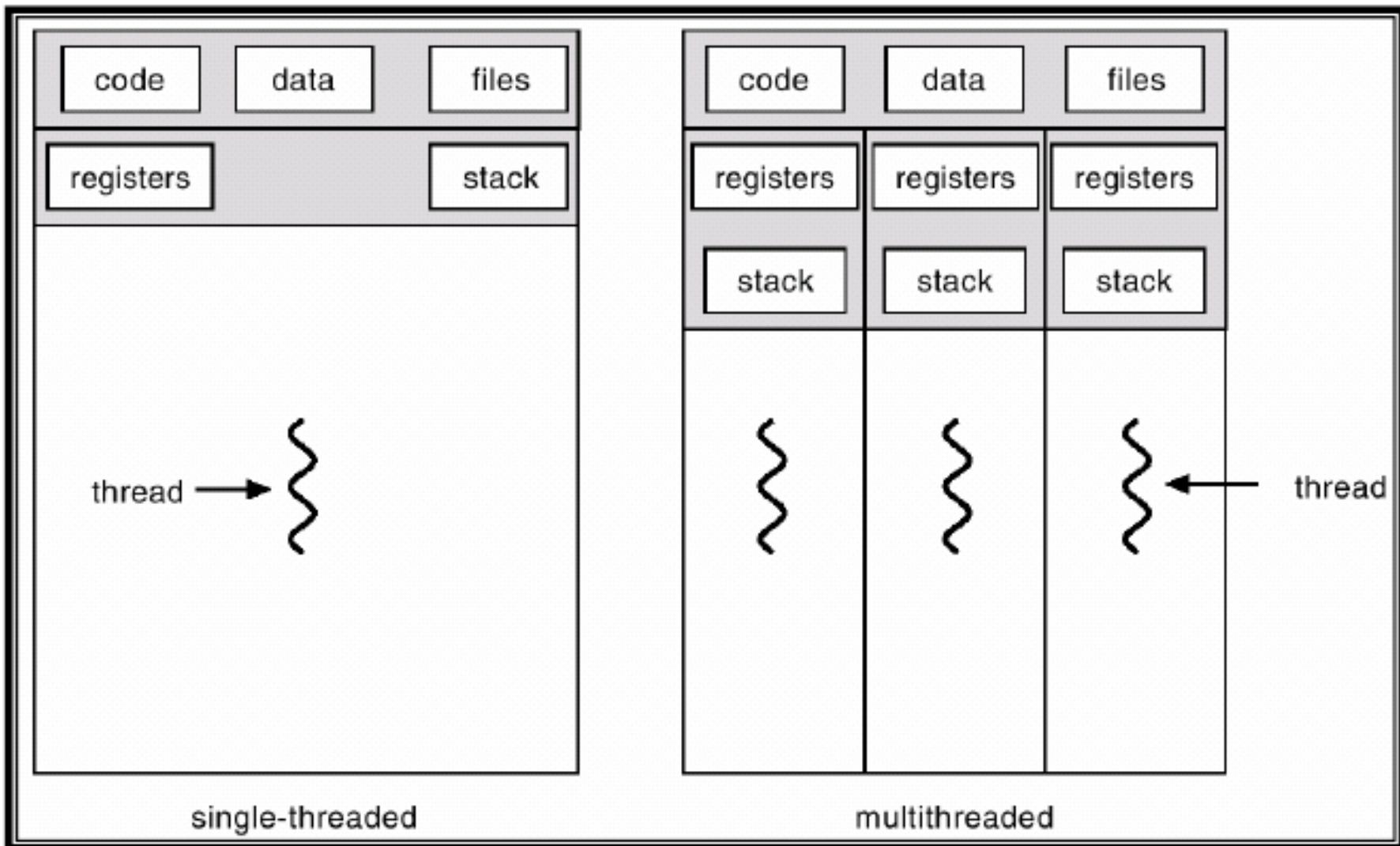


# Архитектура процессоров x86

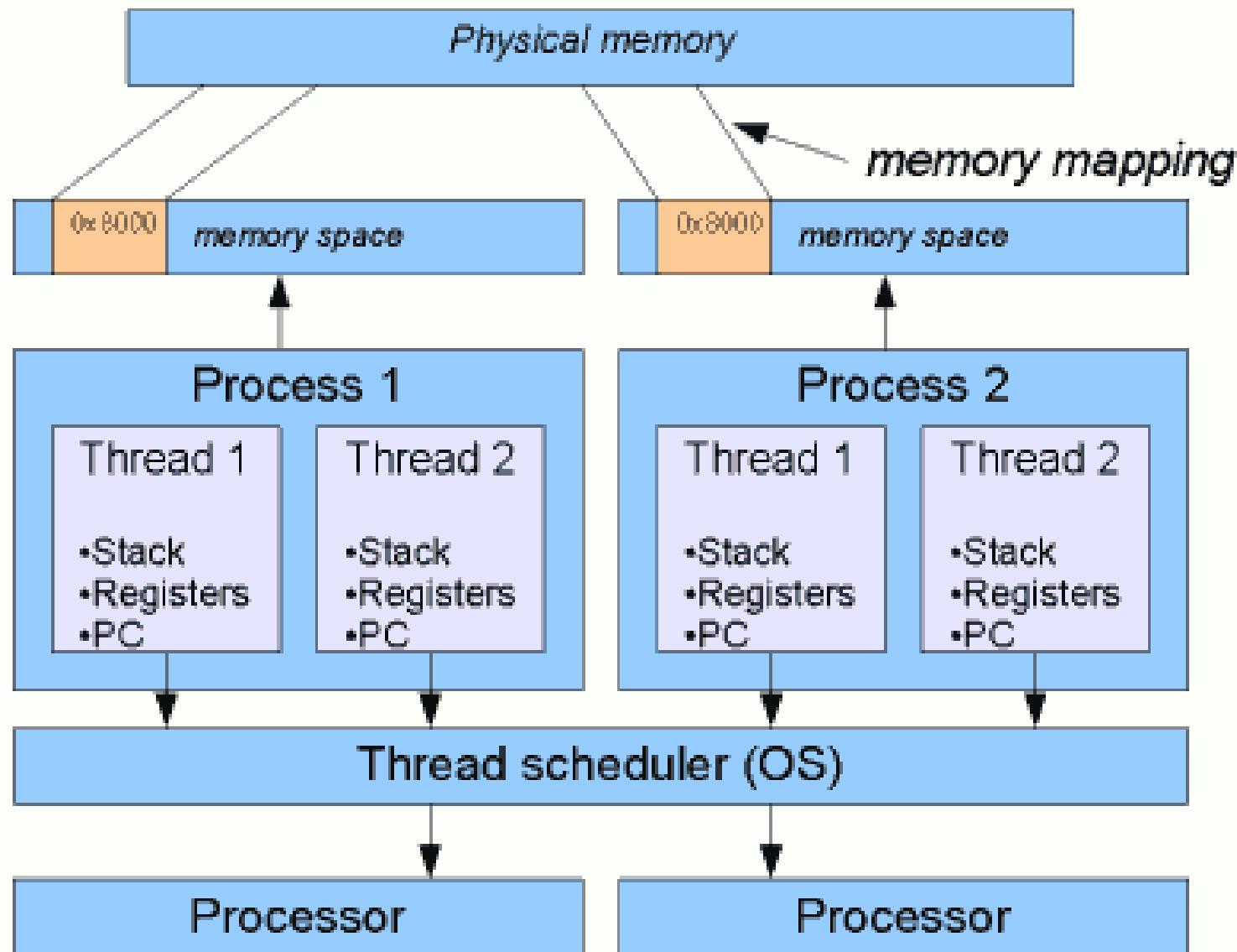
- Pentium 4: Hyper-threading, hyperpipelining



# Процессы, потоки

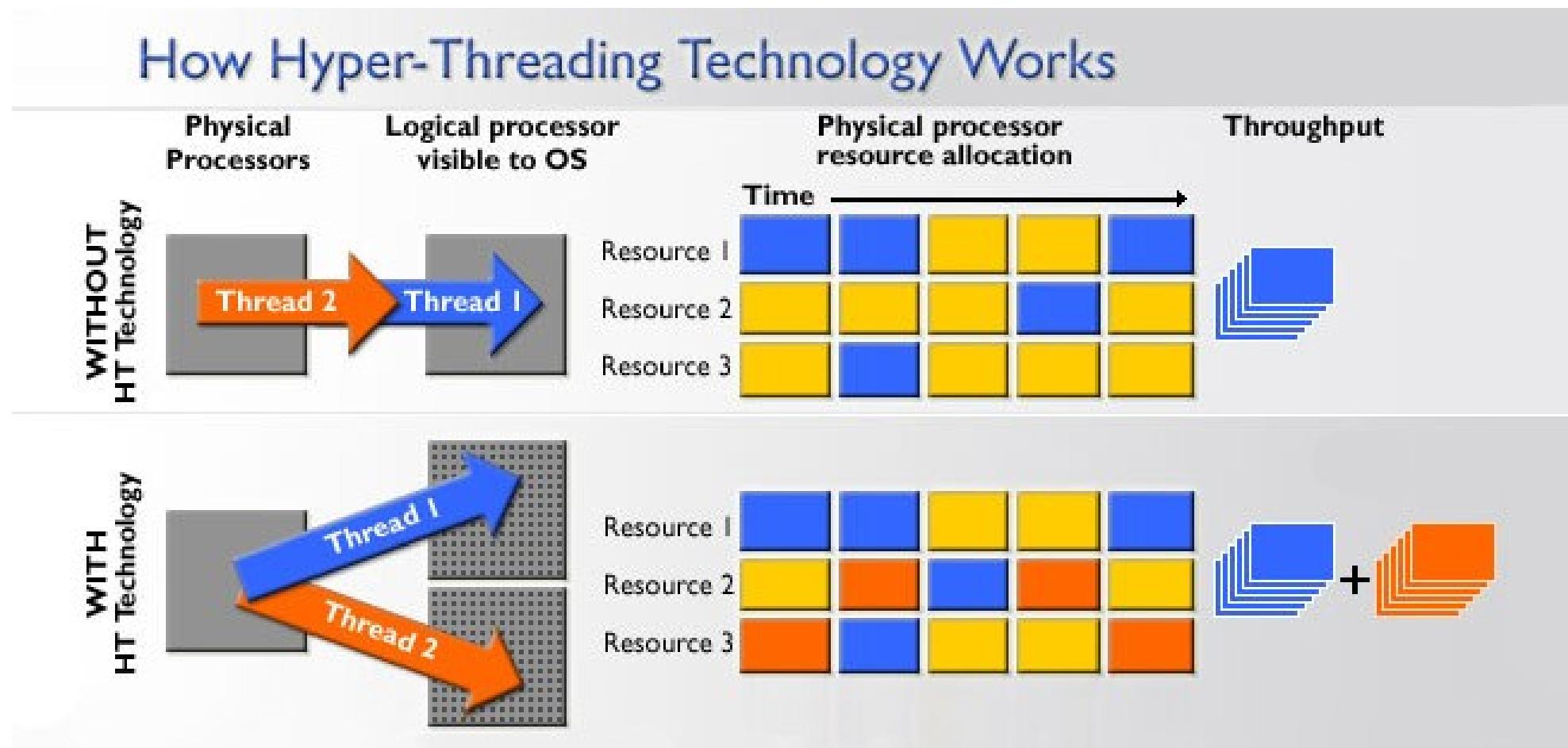


# Процессы, потоки



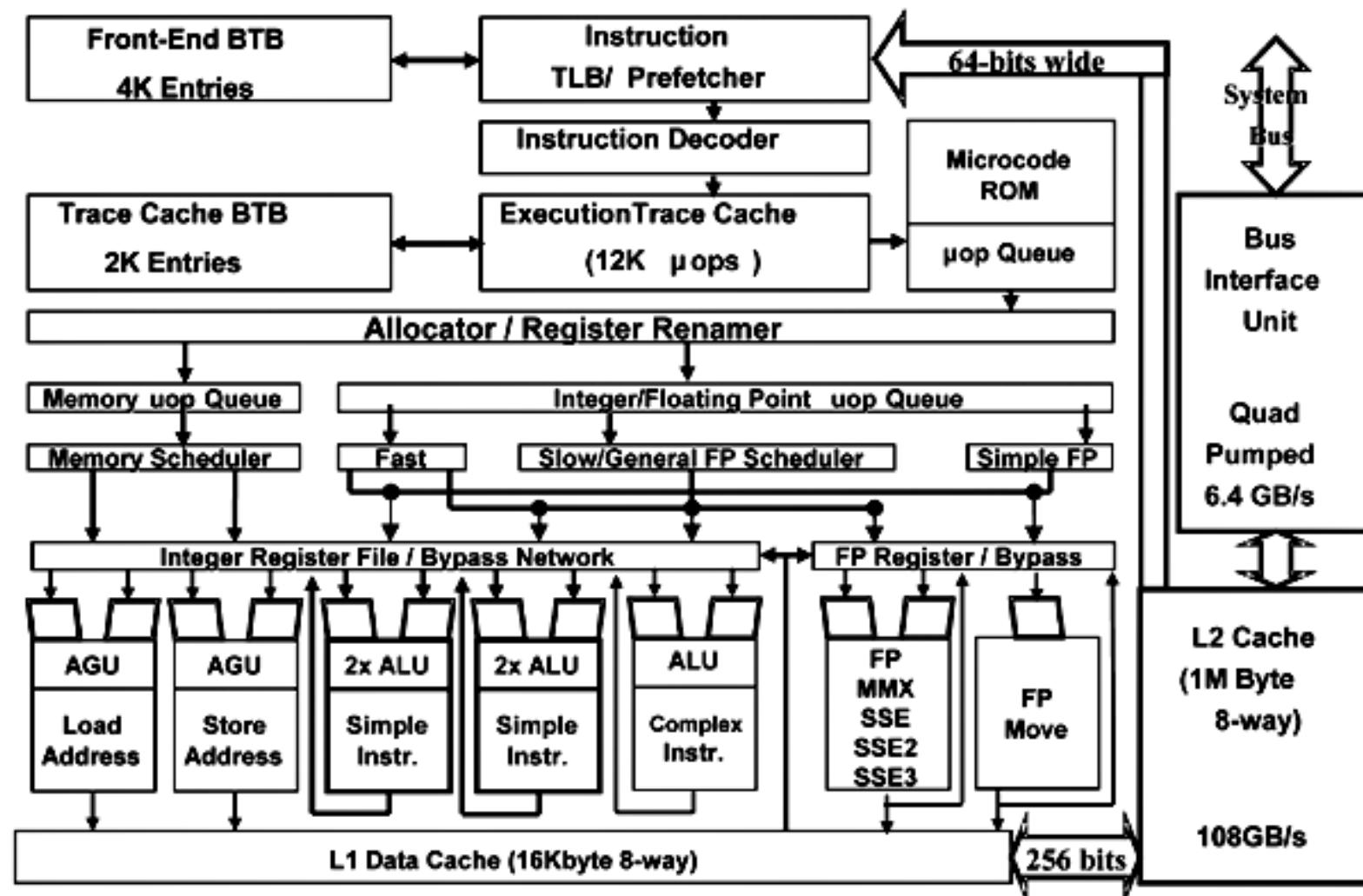
# Архитектура процессоров x86

- Pentium 4: Hyper-threading, hyperpipelining



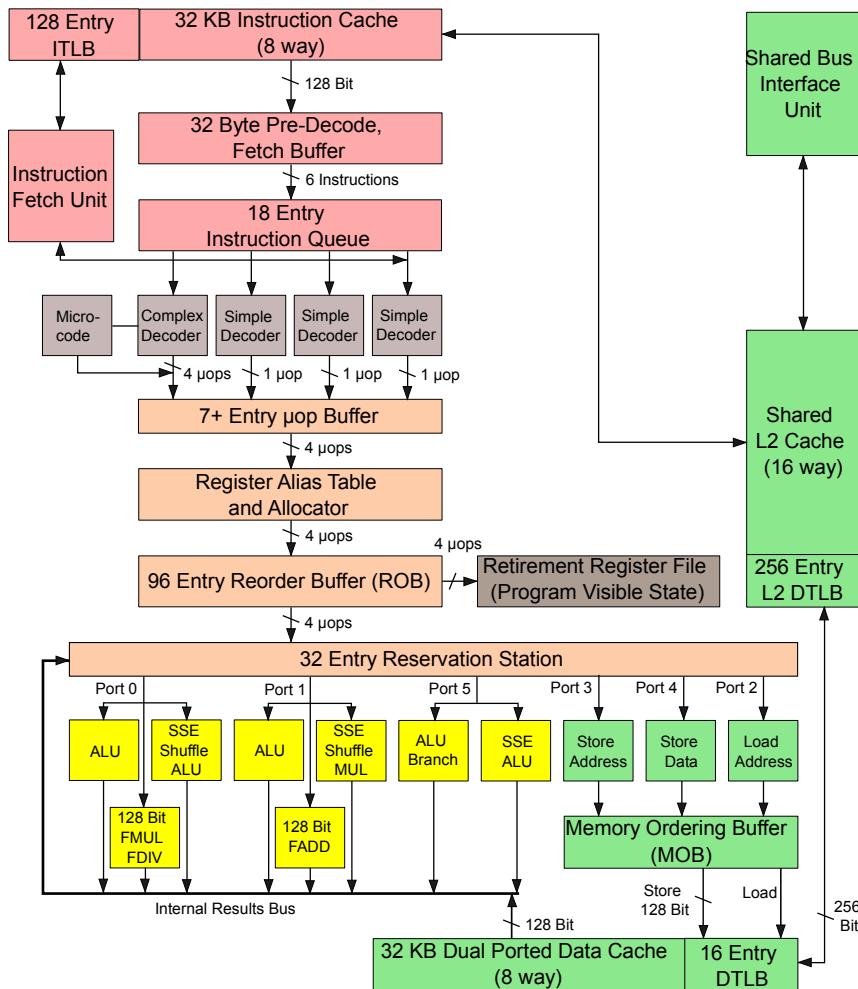
# Архитектура процессоров x86

- Pentium 4: Hyper-threading, hyperpipelining

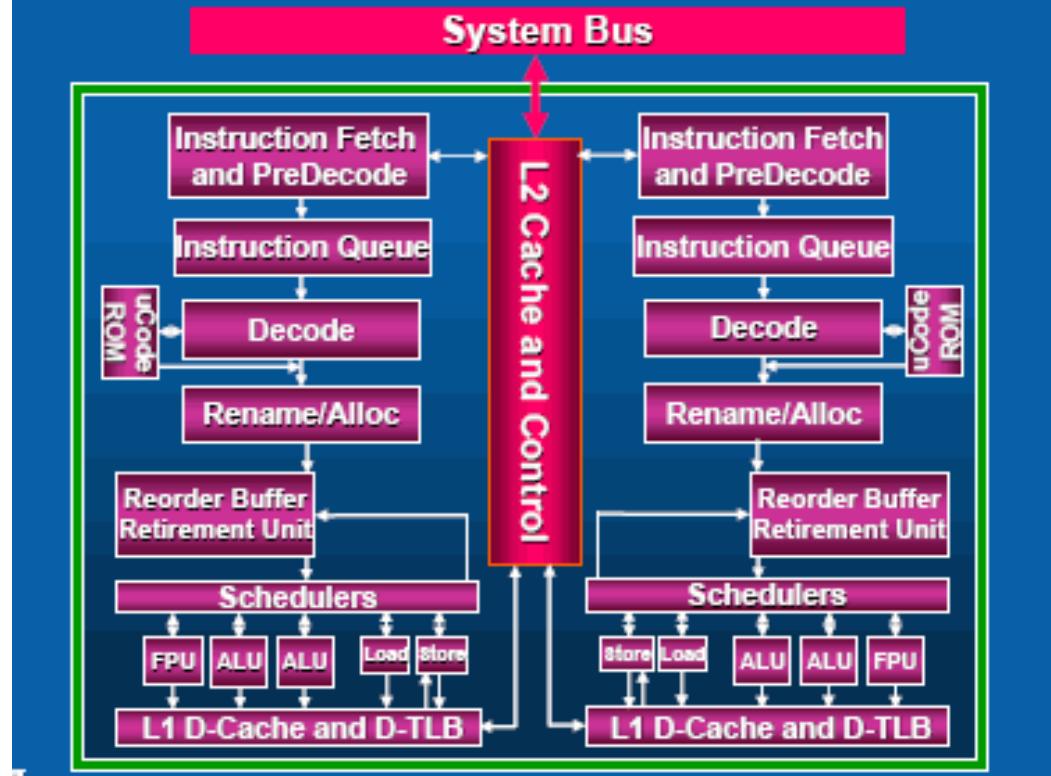


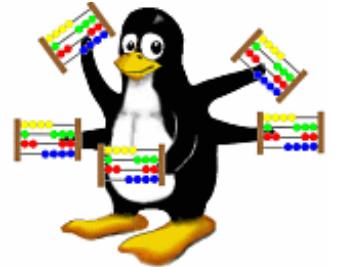
# Архитектура процессоров x86

- Intel Core/Core 2: multicore

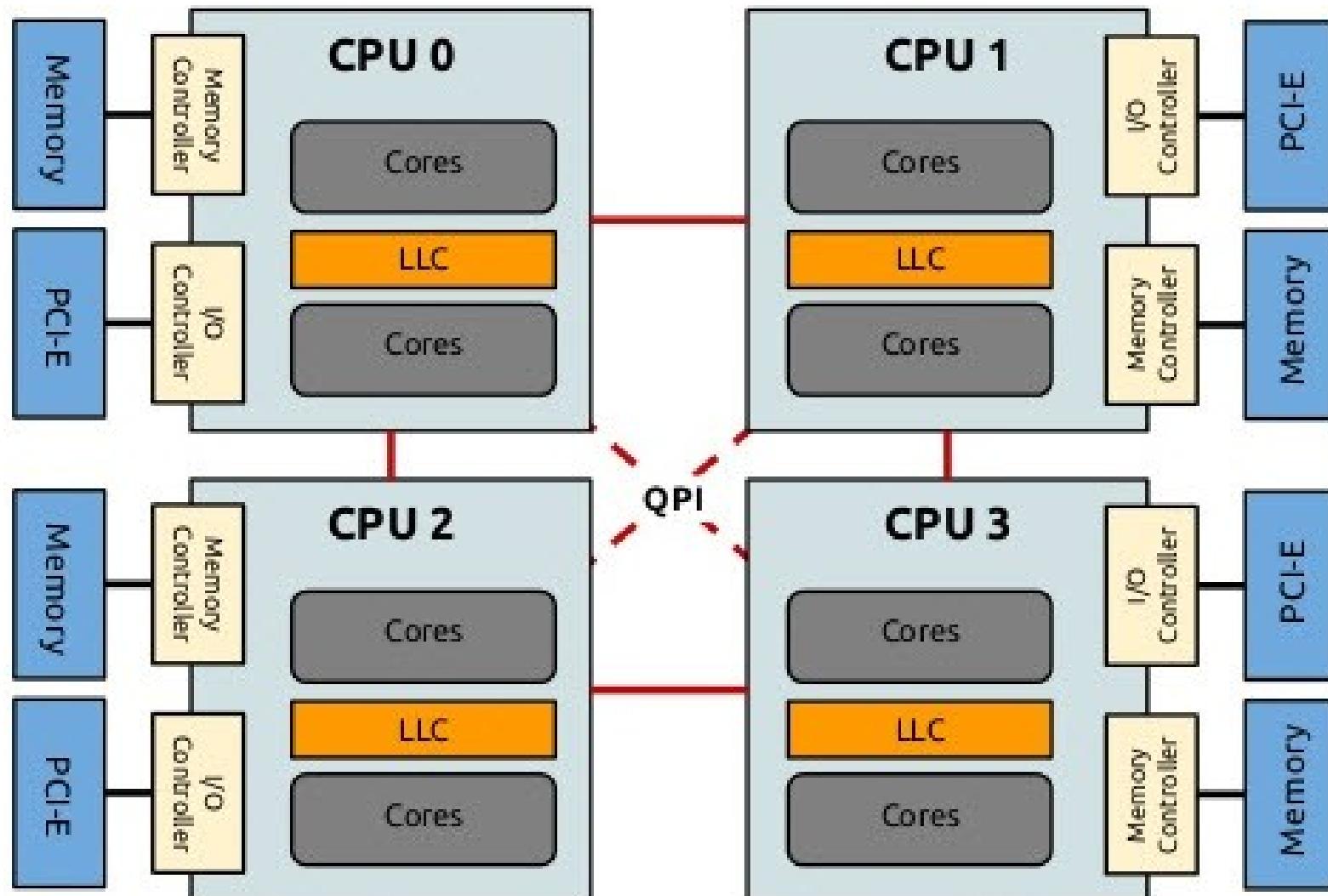


**Merom, Conroe and Woodcrest Block Diagram**



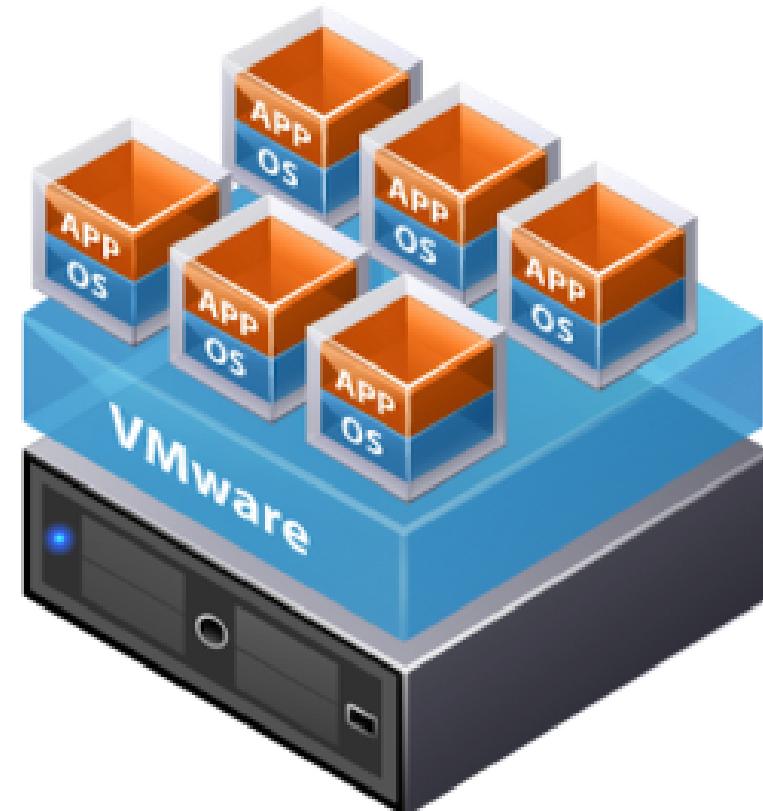
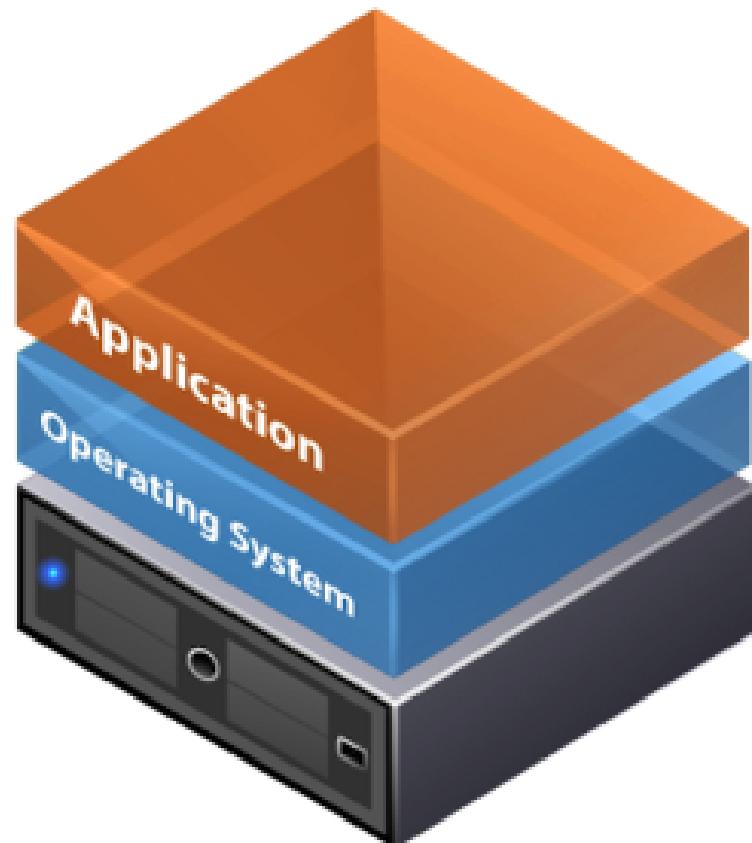


## CPU architecture (Intel Sandy Bridge)

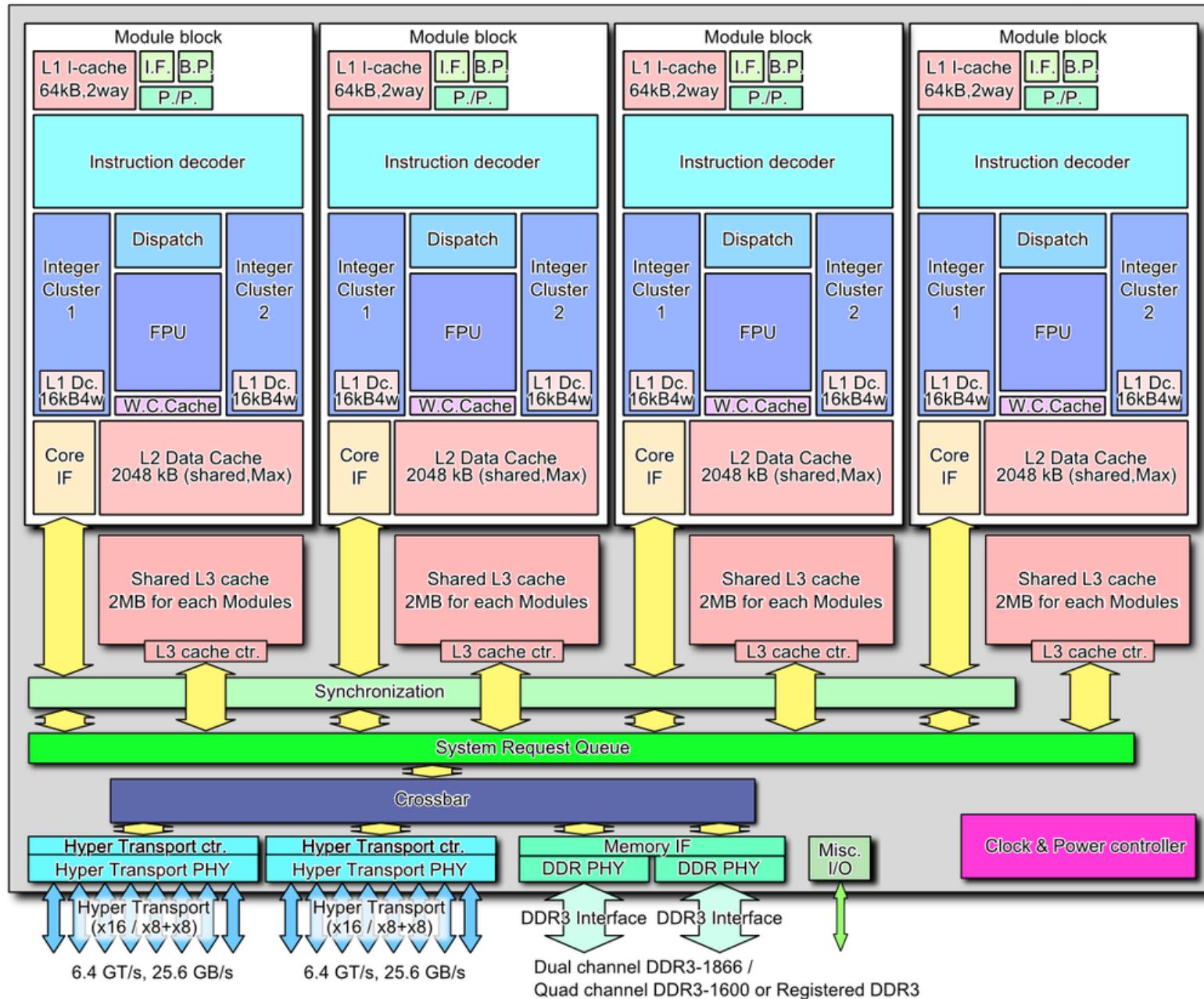


# Архитектура процессоров x86

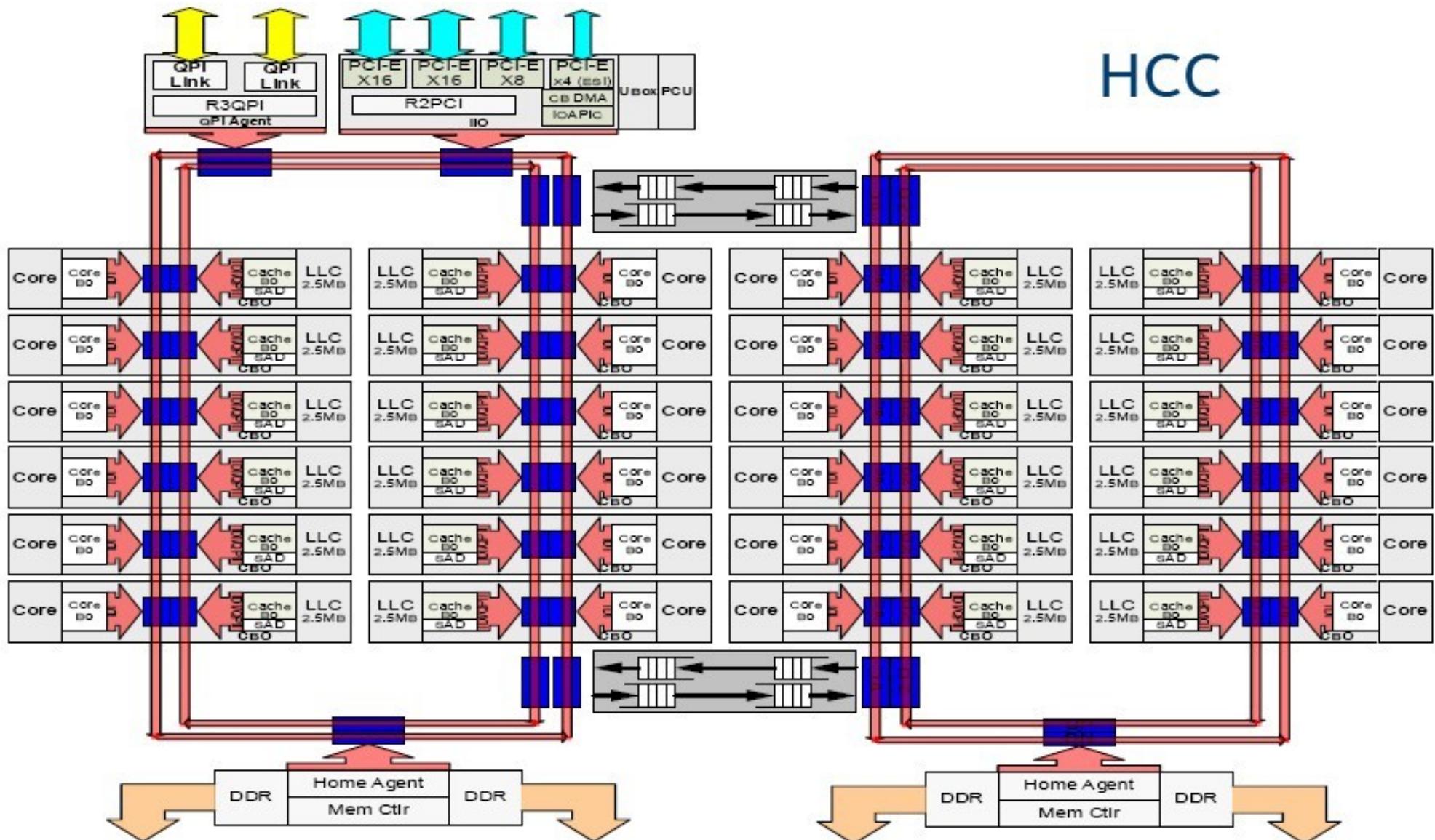
- Intel Core i3/i5/i7: vmx, 3-channel MMU



# AMD Bulldozer



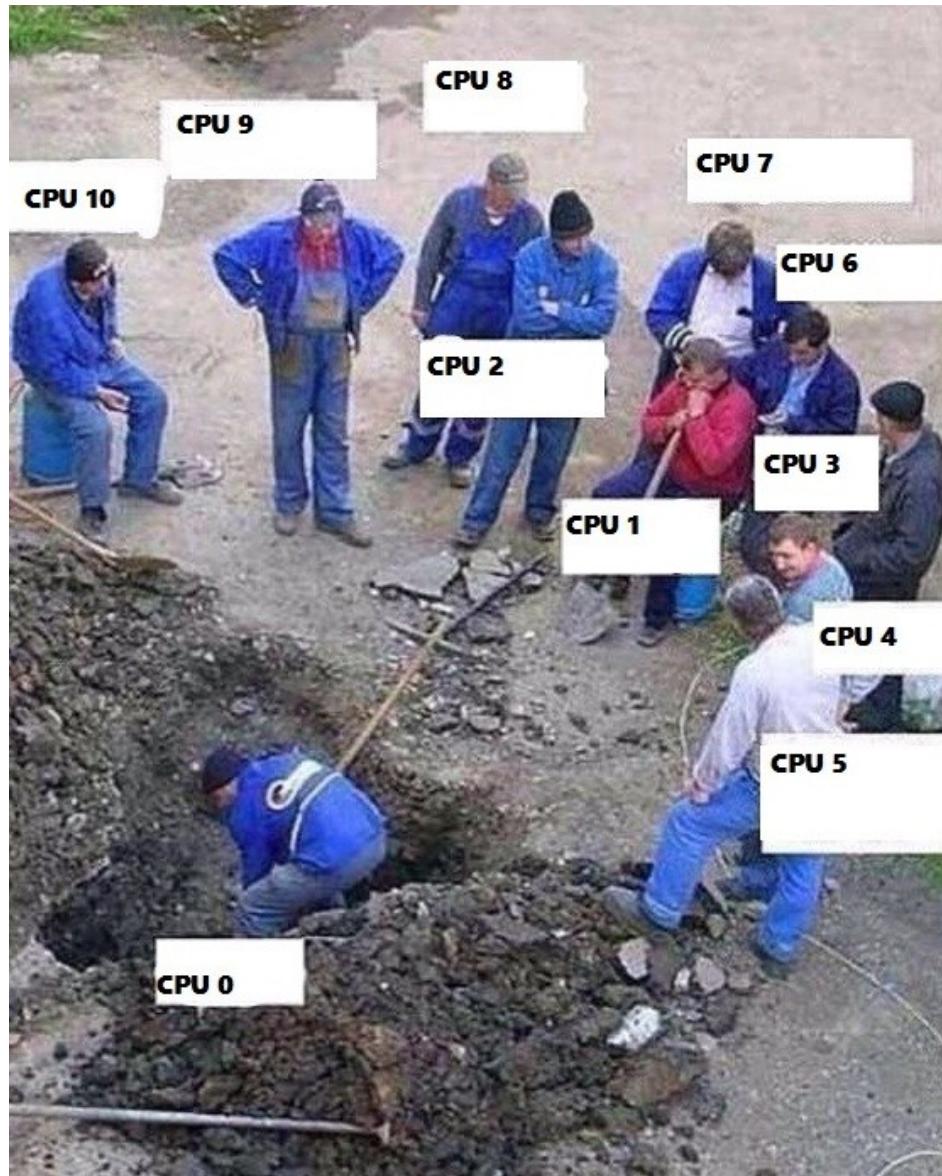
# Intel Xeon E5



# Multicore, hyperthreading



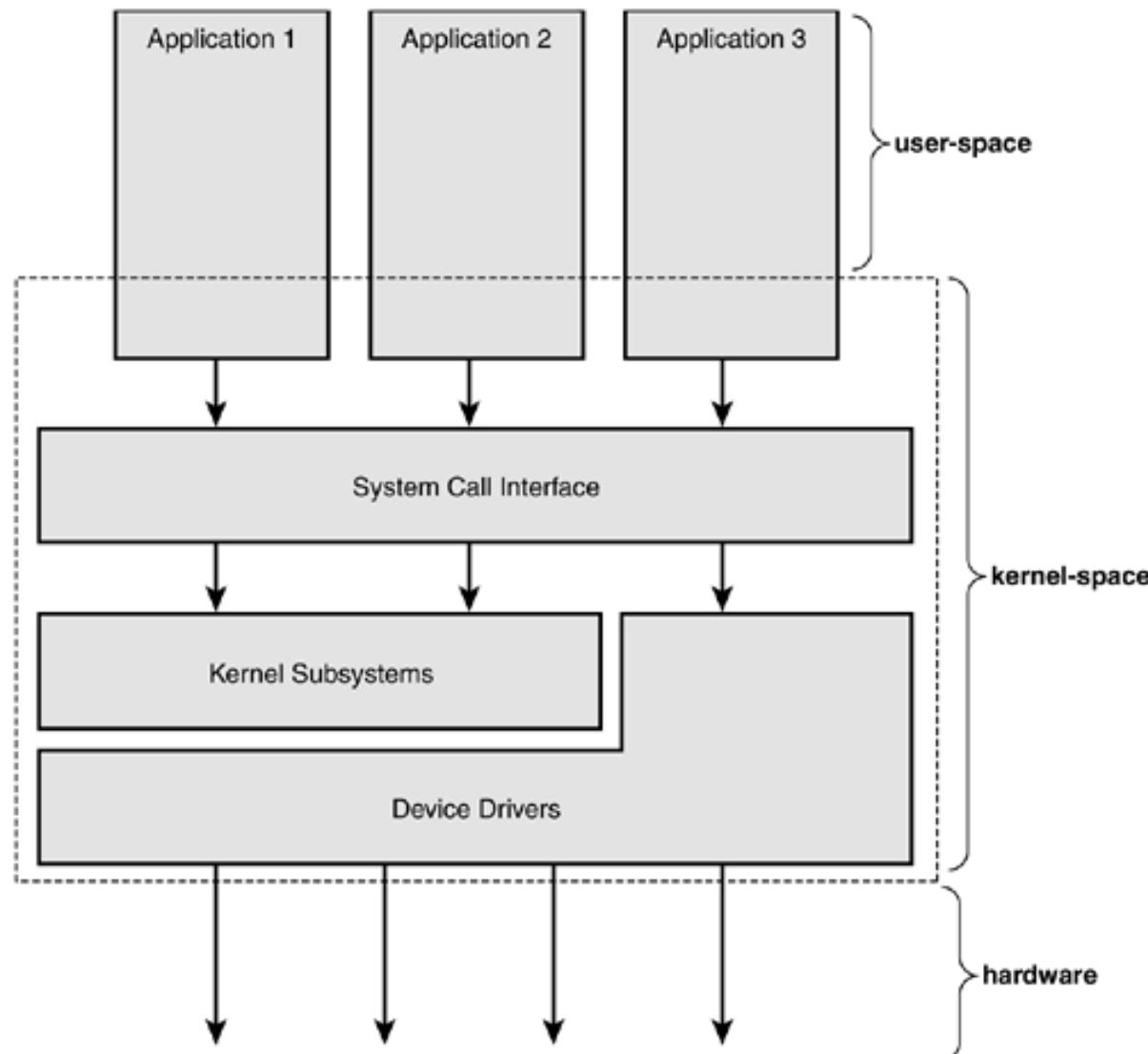
# Multicore, hyperthreading



# Современный процессор (с точки зрения HPC)

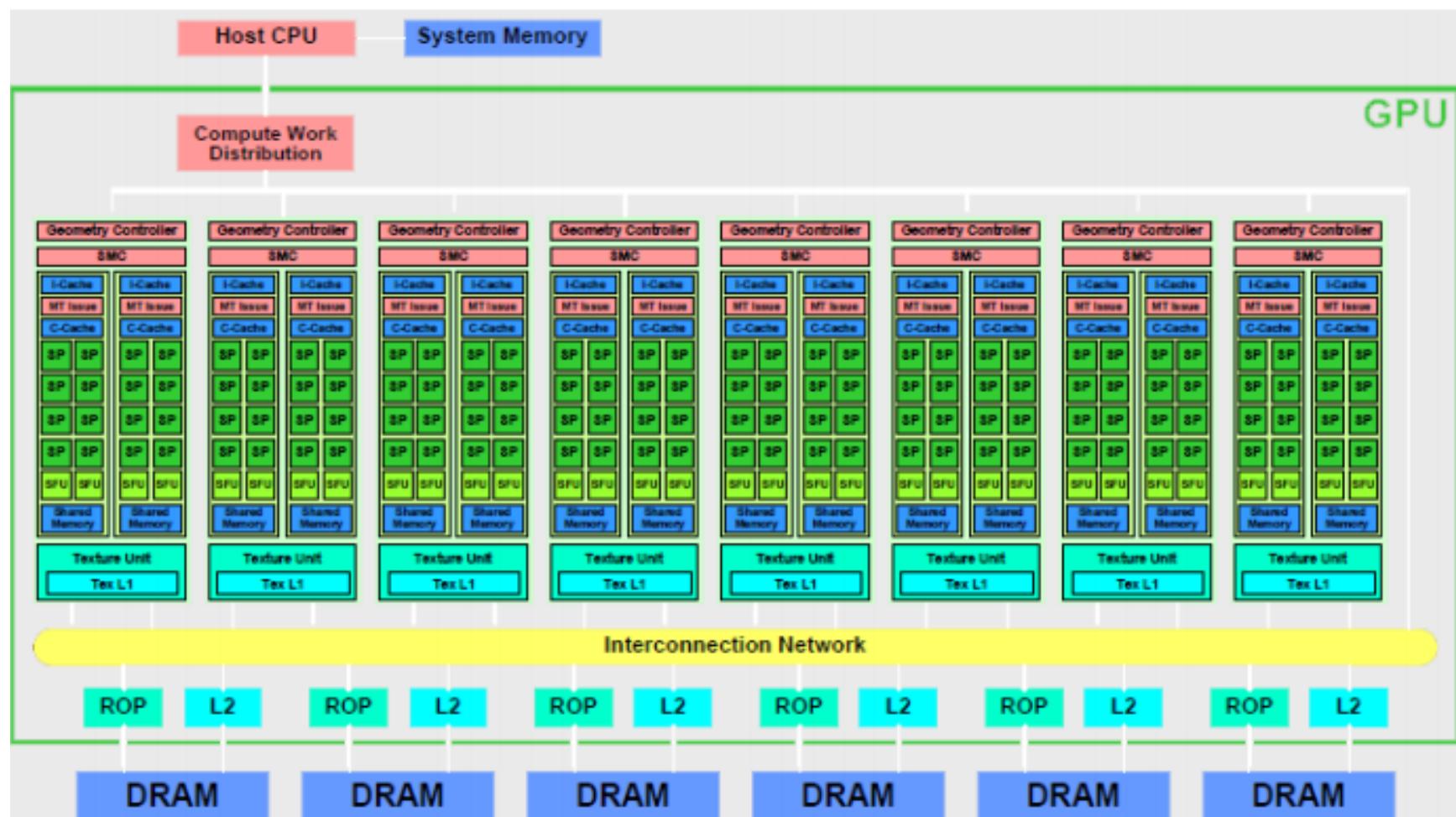
- Многоядерность
- Малопредсказуемое поведение cache и других систем
- SIMD (AVX, FMA)
- Сложное взаимодействие с ОЗУ и с периферией
- Другие архитектуры: VLIW
- Кроме этого необходимо помнить: cache (L1/L2/L3, TLB, ...); branch prediction; context switch.

# kernel-space ↔ user-space



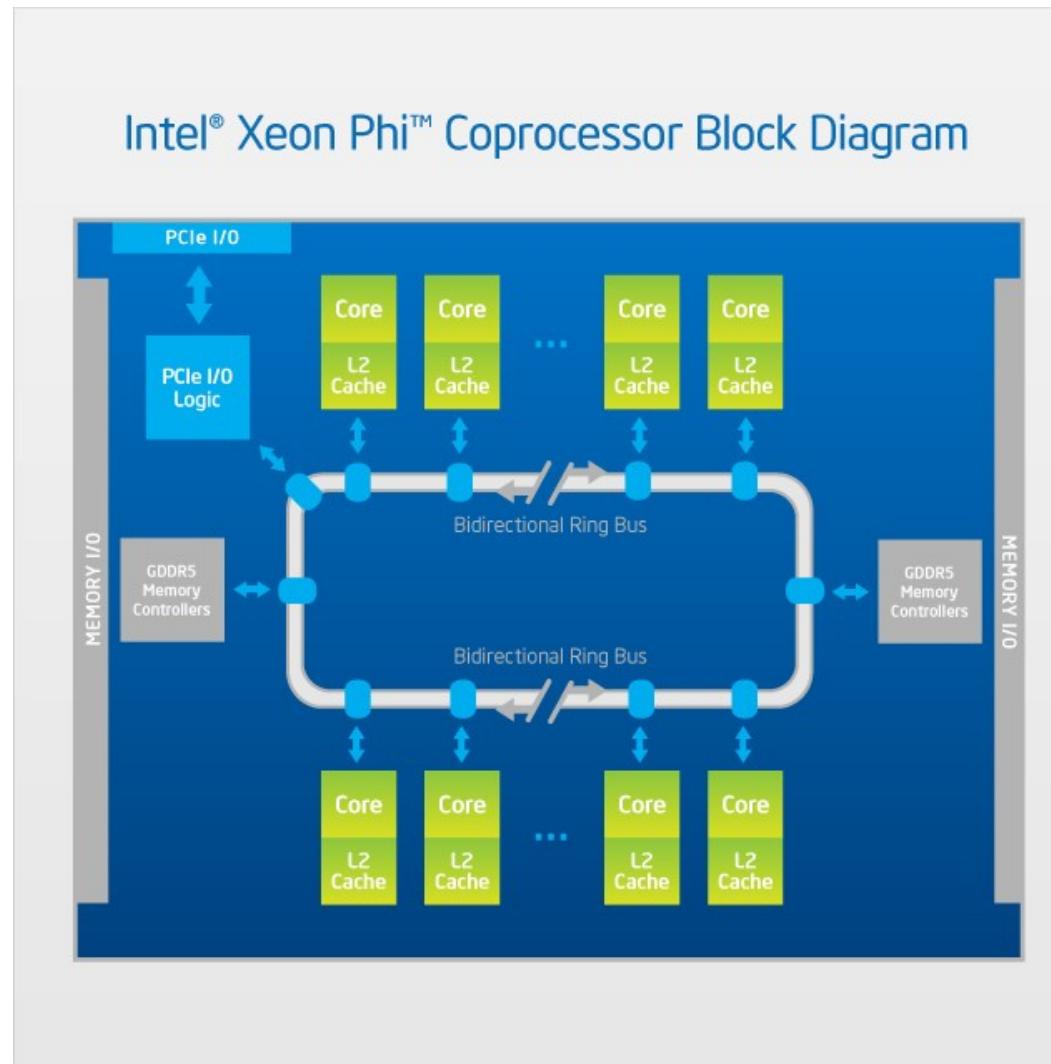
# Аппаратные ускорители

- GPGPU:

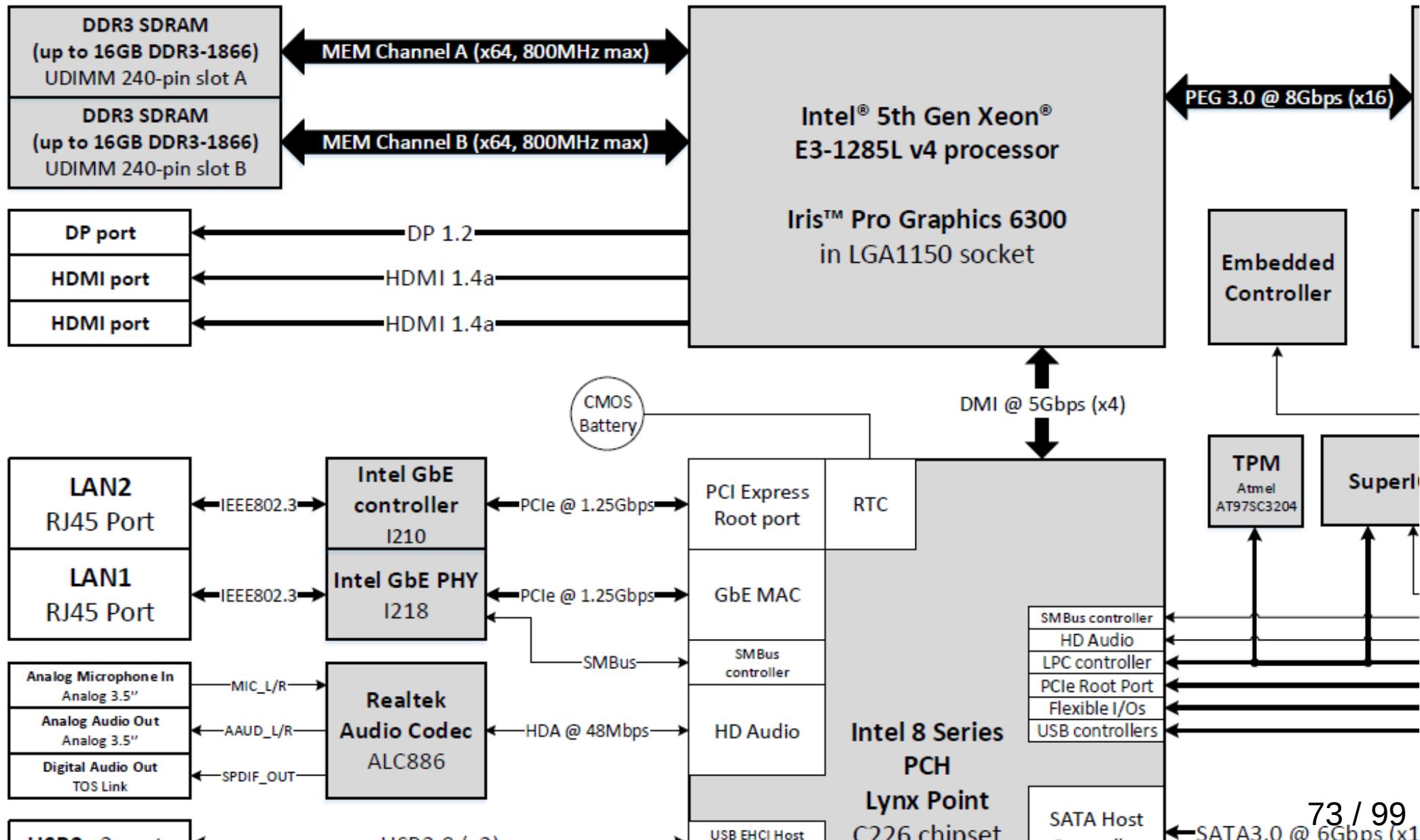


# Аппаратные ускорители

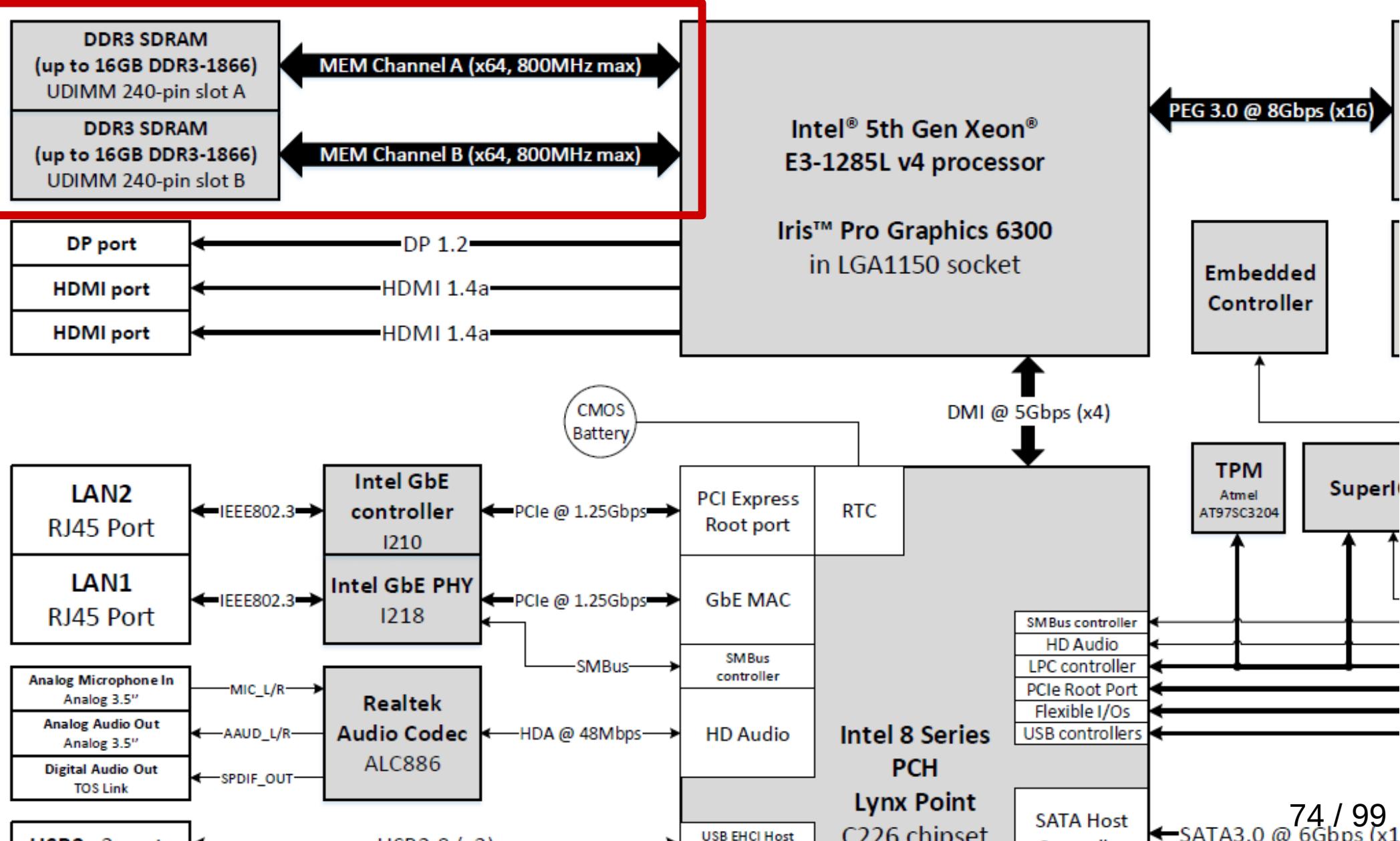
- Xeon Phi:



# С аппаратной точки зрения: вычислительный узел



# С аппаратной точки зрения: вычислительный узел

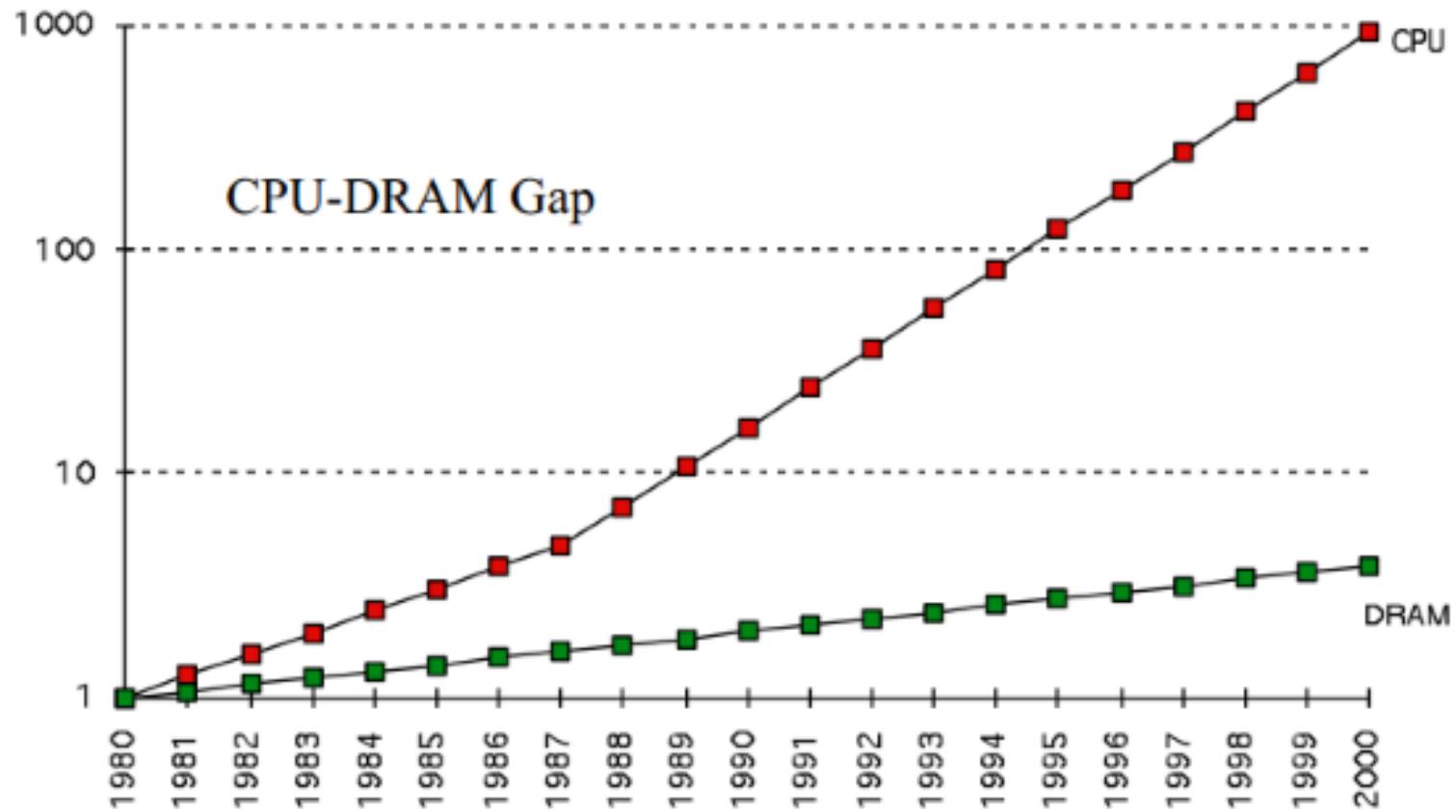


# RAM

- CPU-RAM performance GAP
- NUMA

# RAM

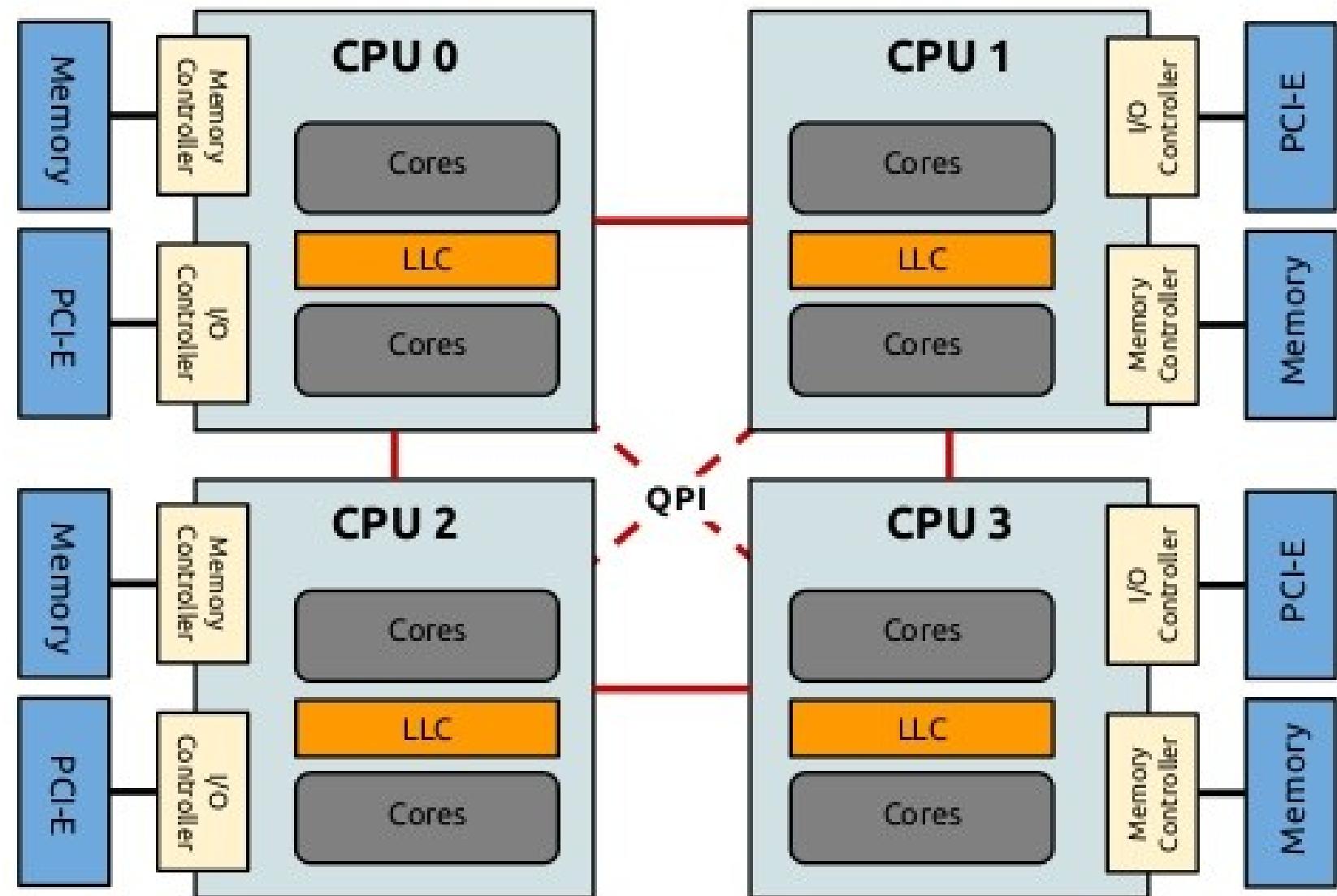
- Processor vs Memory Performance



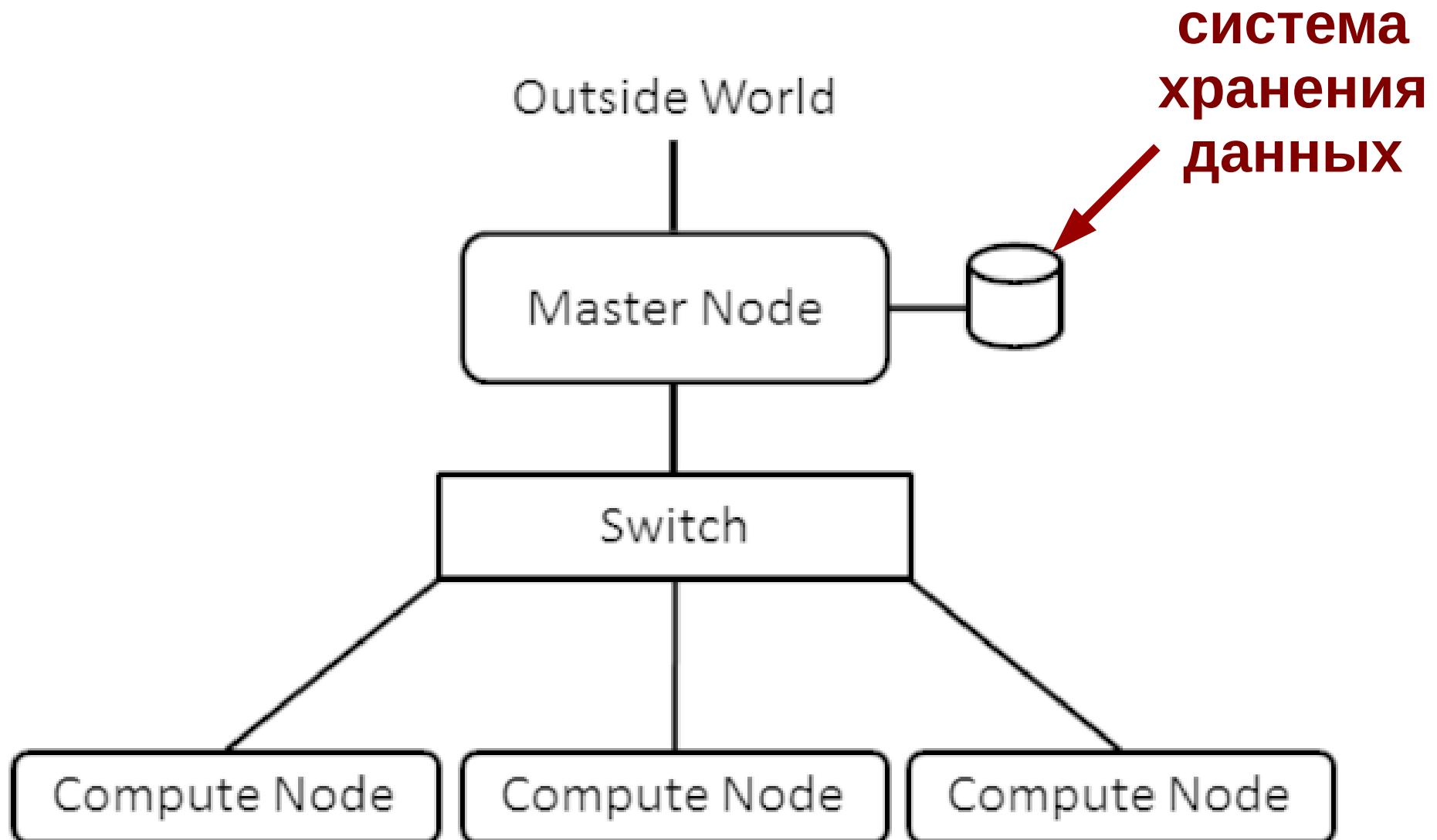
1980: no cache in microprocessor;

1995 2-level cache

# RAM



# С аппаратной точки зрения: основные моменты



# Система хранения данных

Конечной задачей является создание ФС, доступной с любого вычислительного и управляющего узла

Используемые решения:

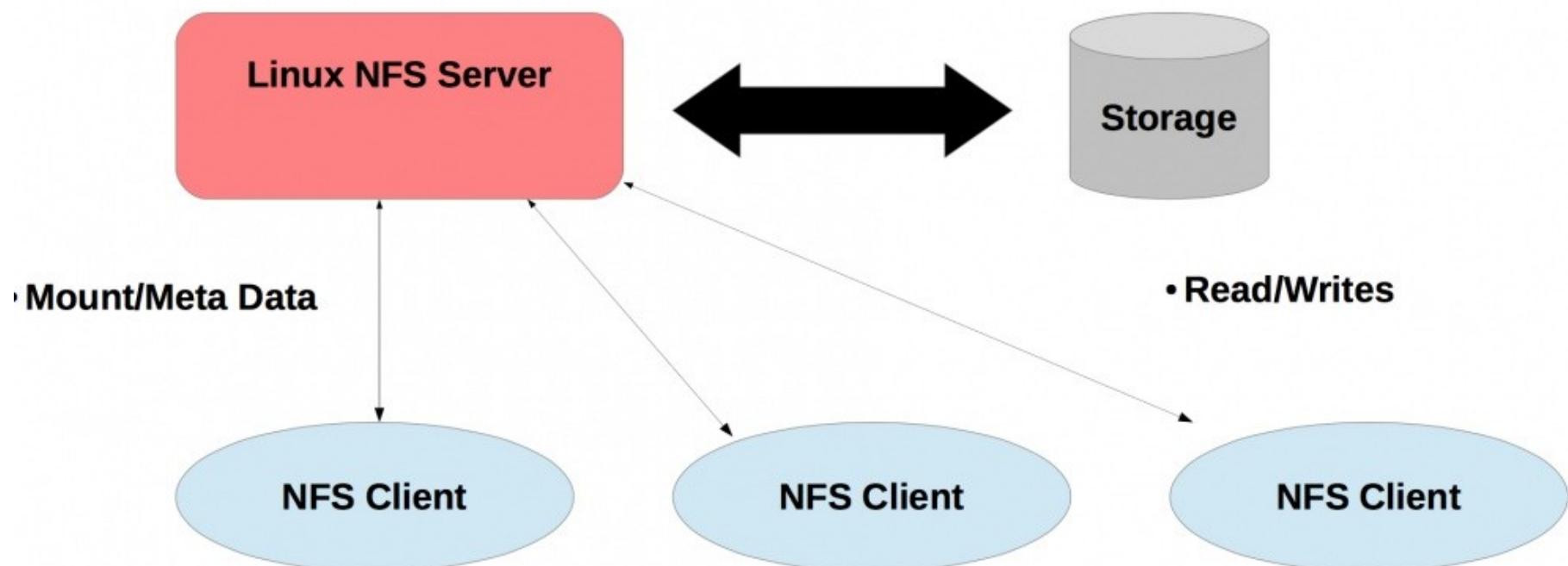
- Дорогие enterprise-решения (некомпетентно для HPC)
- Собственные решения:
  - Экспорт ФС:
    - NFS/pNFS
    - Ceph
    - Lustre/OrageFS
    - AUFS
    - другое...
  - Экспорт блочных устройств:
    - FC
    - iSCSI
    - iSER
    - SRP
    - другое...

...

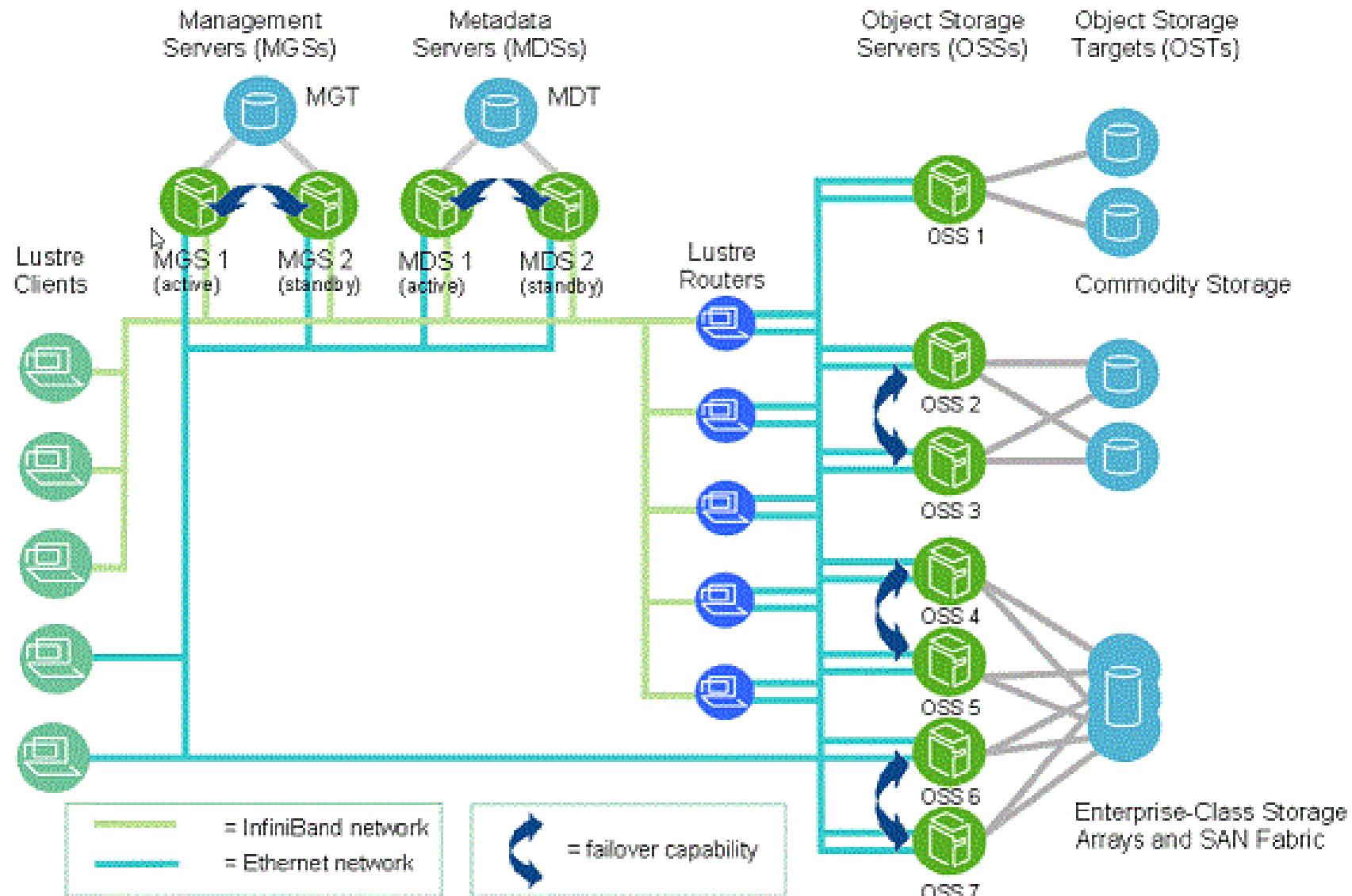
# Простая система хранения

## Traditional NFS

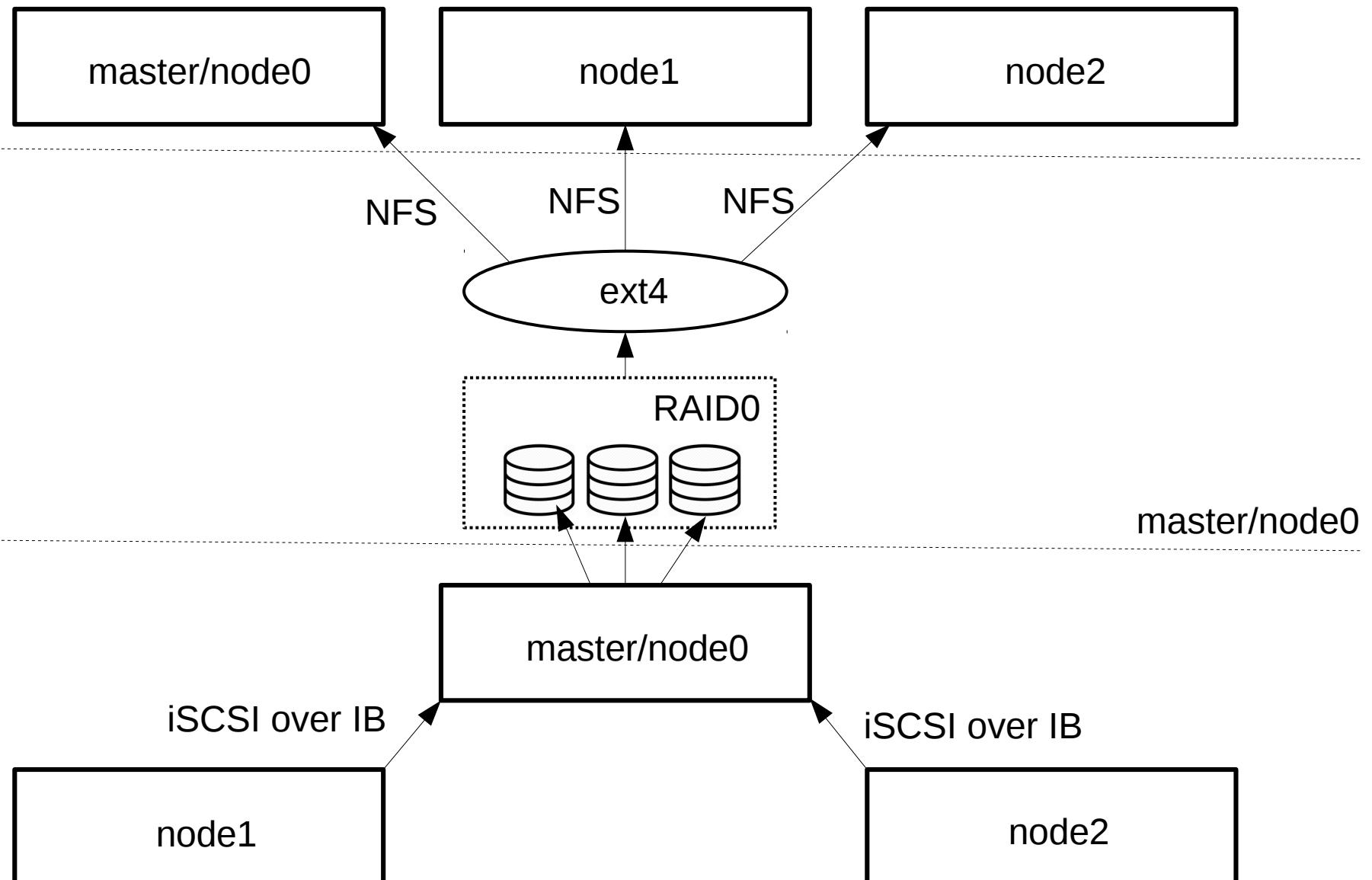
One Server for Multiple Clients  
= Limited Scalability



# Lustre FS



# Система хранения кластера «lambda»

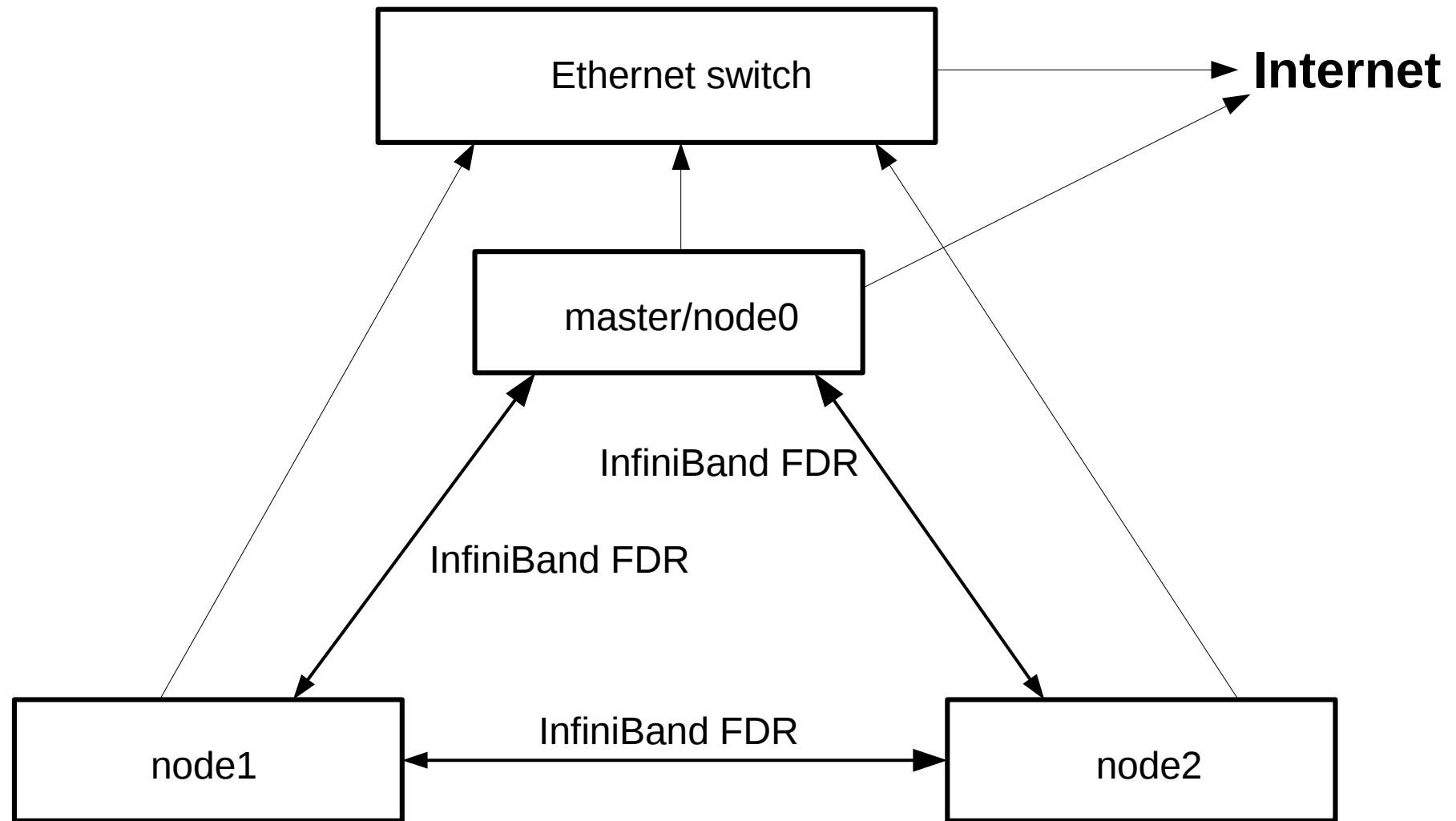


# Кластер «lambda»

Современное HPC-оборудование



# Кластер «lambda»



# Имеется оборудование, что дальше?

Области деятельности специалистов:

- Охлаждение, электрика, пожарная безопасность и т.п.
- Аппаратное обеспечение и системное администрирование
- Программирование (формулировка вычислительных задач)

# Как этим пользоваться?

Области деятельности специалистов:

- Охлаждение, электрика, пожарная безопасность и т.п.
- Аппаратное обеспечение и системное администрирование
- Программирование (формулировка вычислительных задач)

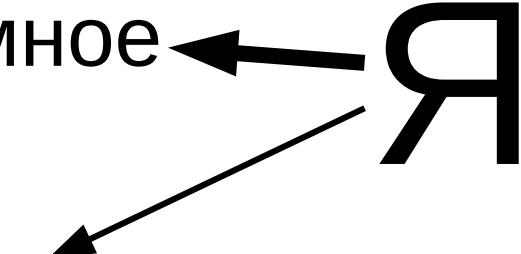
пользователь

# Как этим пользоваться?

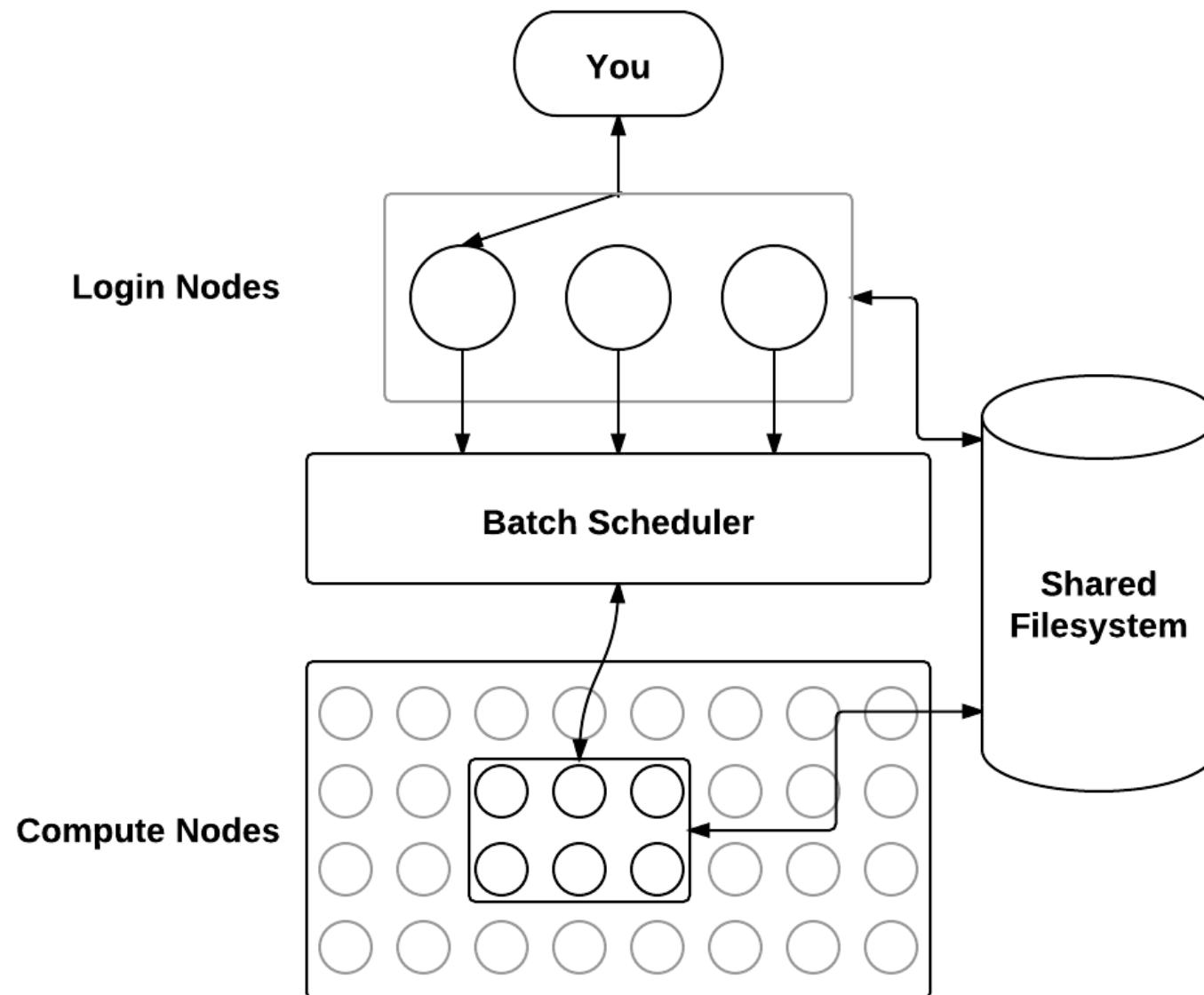
Области деятельности специалистов:

- Охлаждение, электрика, пожарная безопасность и т.п.
- Аппаратное обеспечение и системное администрирование
- Программирование (формулировка вычислительных задач)

Я



# Как этим пользоваться-то?



# Терминал

```
000 d[21:46:15] [xaionaro@shadow ~]$ ssh hpc
Password:
Available HPC clusters:
  basov
  cherenkov
  unicluster [current]

To log on other cluster, type:
  ssh cluster_name

For example:
  ssh basov

xaionaro@master.unicluster ~ $ █
```

# Примеры задач

- Вычисление числа Пи
- Газодинамика
- Физика высоких энергий
- Криптография
- И мн. др.

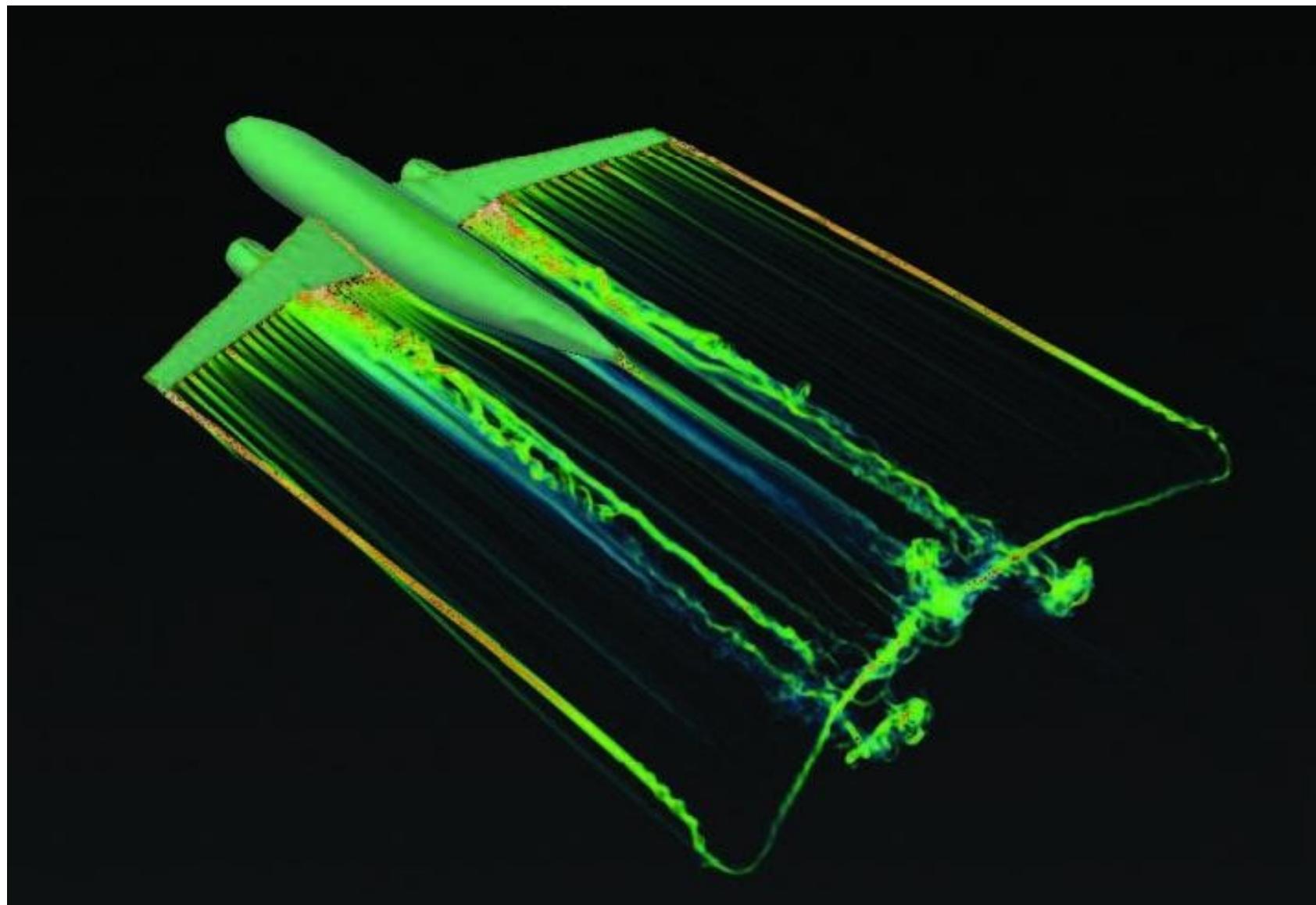
# Пользовательские инструменты HPC-систем

- Стандартные утилиты GNU/Linux
- Готовые научные приложения
- Компиляторы, библиотеки, средства отладки
- Менеджер ресурсов

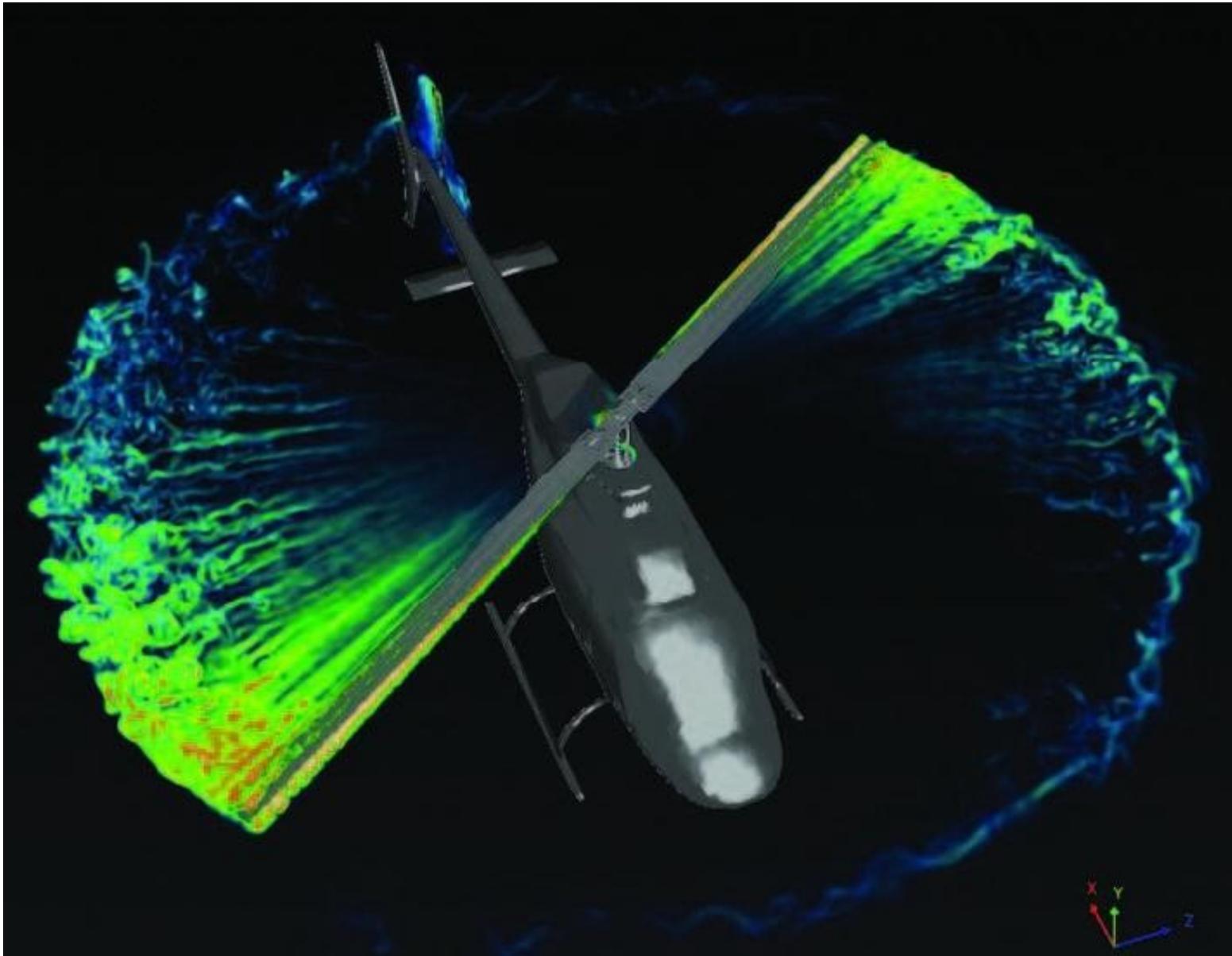
# Задача от начала до конца

- Мат. аппарат; численный метод (с оглядкой на параллельность)
- Написание приложения (использование существующего)
- Отладка на базе известных данных
- Запуск необходимого расчёта
- Сбор данных
- Визуализация

# Визуализация



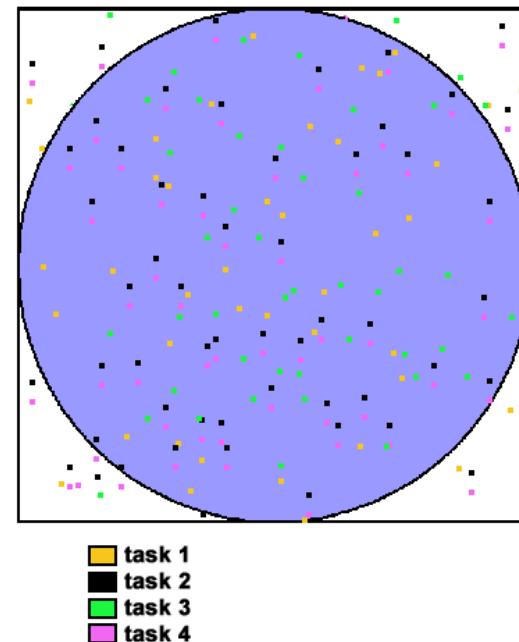
# Визуализация



# Демонстрация

- Поиск числа Пи. Популярные варианты:

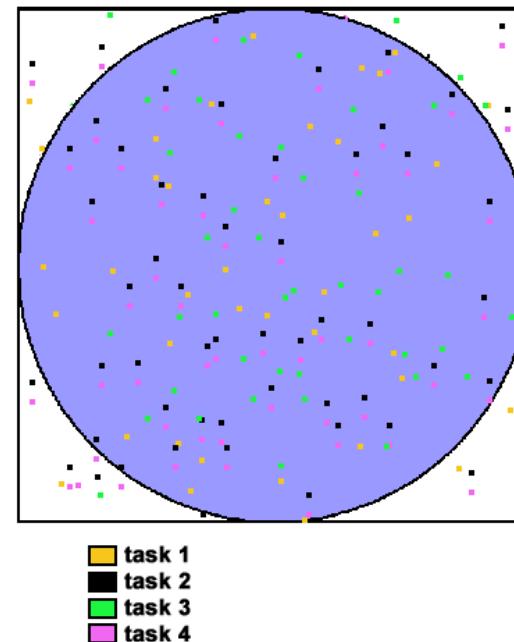
$$\pi = \int_0^1 \frac{4}{1+x^2} dx$$



# Демонстрация

- Поиск числа Пи. Популярные варианты:

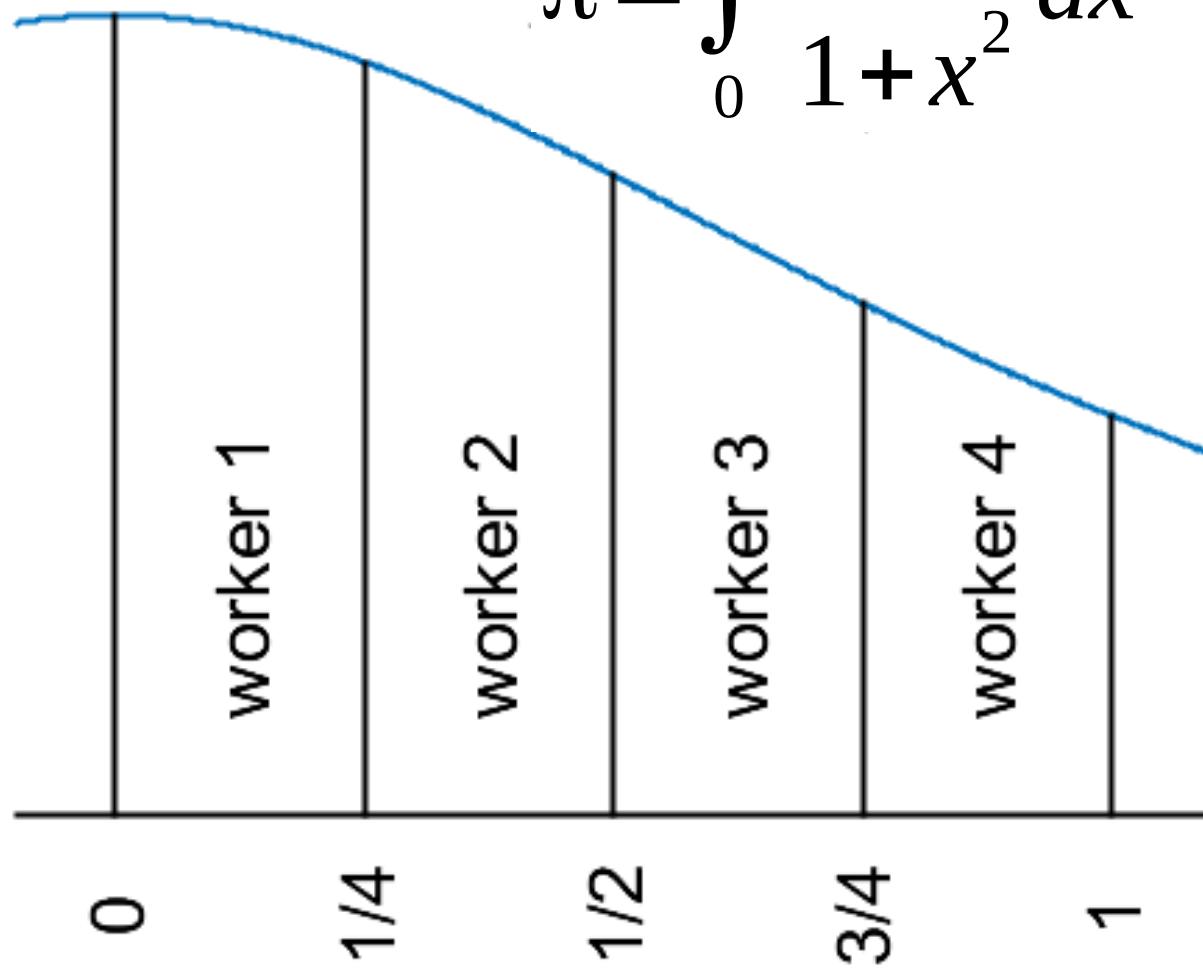
$$\pi = \int_0^1 \frac{4}{1+x^2} dx$$



# Демонстрация

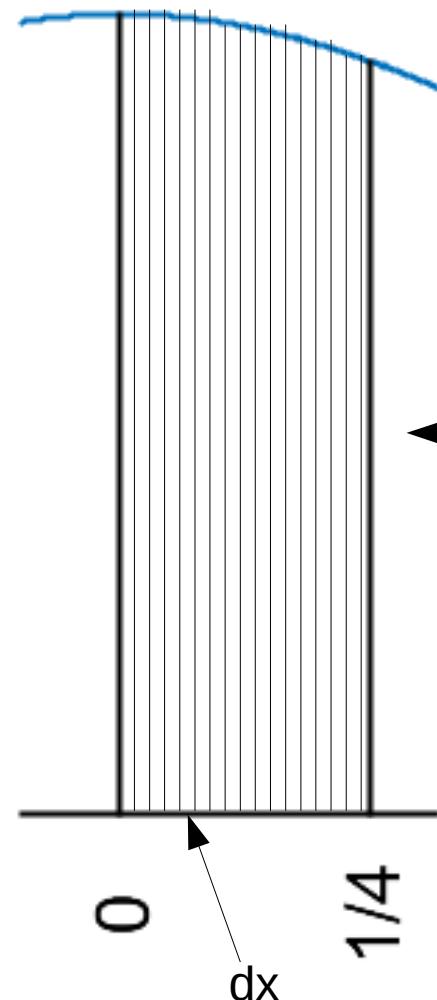
- Поиск числа Пи

$$\pi = \int_0^1 \frac{4}{1+x^2} dx$$



# Демонстрация

- Поиск числа Пи



$$\pi = \int_0^1 \frac{4}{1+x^2} dx$$

# Q&A