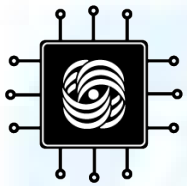
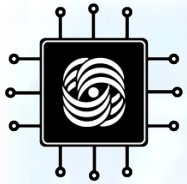


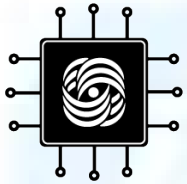
АРХИТЕКТУРА КОМПЬЮТЕРНЫХ СИСТЕМ

Лекция 10:
Параллельные вычисления (2)



План лекции

- Классификация многопроцессорных вычислительных систем
  - Мультипроцессоры – системы с общей памятью
  - Мультикомпьютеры – системы с распределенной памятью
- Типовые схемы коммуникации процессоров
- Примеры параллельных систем



Классификация вычислительных систем

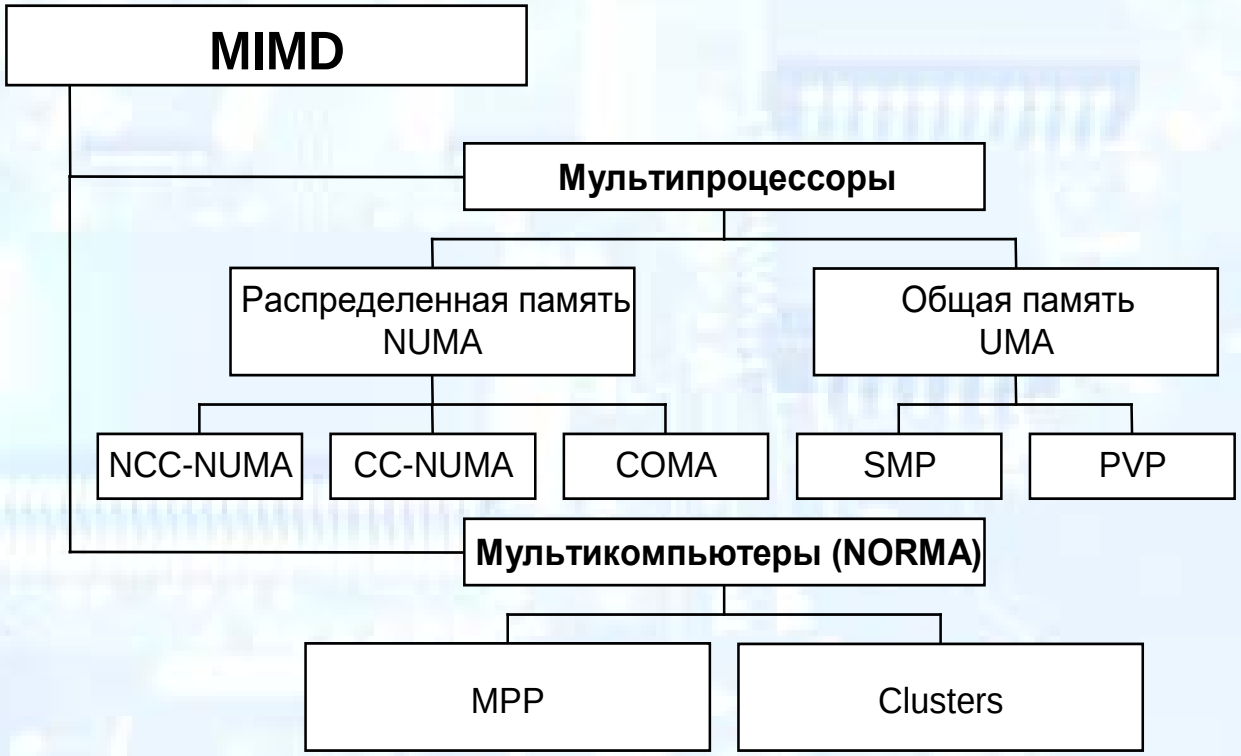
- **Систематика Флинна** (Flynn)
  - Классификация по способам взаимодействия последовательностей (*потоков*) выполняемых команд и обрабатываемых данных:
    - **SISD** (Single Instruction, Single Data)
    - **SIMD** (Single Instruction, Multiple Data)
    - **MISD** (Multiple Instruction, Single Data)
    - **MIMD** (Multiple Instruction, Multiple Data)

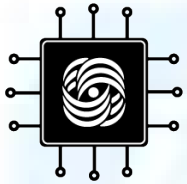
Практически все виды параллельных систем, несмотря на их существенную разнородность, относятся к одной группе ***MIMD***

Классификация вычислительных систем

- **Детализация систематики Флинна…**
  - Дальнейшее разделение типов многопроцессорных систем основывается на используемых способах организации оперативной памяти,
  - Позволяет различать два важных типа многопроцессорных систем:
    - ***multiprocessors*** (***мультипроцессоры*** или системы с общей разделяемой памятью),
    - ***multicomputers*** (***мультикомпьютеры*** или системы с распределенной памятью).

Классификация вычислительных систем



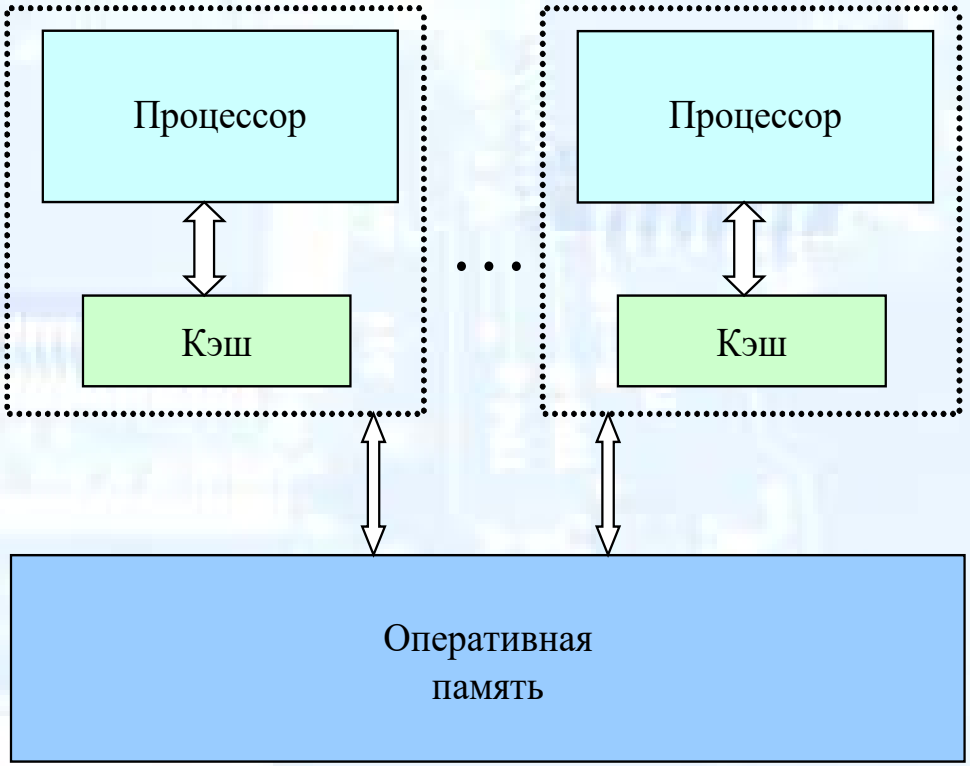


Классификация вычислительных систем

- **Мультипроцессоры** с использованием единой *общей памяти* (*shared memory*)…
  - Обеспечивается *однородный доступ к памяти* (*uniform memory access or UMA*),
  - Являются основой для построения:
    - *векторных параллельных процессоров* (*parallel vector processor or PVP*). Примеры: Cray T90,
    - симметричных *мультипроцессоров* (*symmetric multiprocessor or SMP*). Примеры: IBM eServer, Sun StarFire, HP Superdome, SGI Origin.

Классификация вычислительных систем

- **Мультипроцессоры** с использованием единой *общей памяти…*



Классификация вычислительных систем

- **Мультипроцессоры** с использованием единой *общей памяти…*

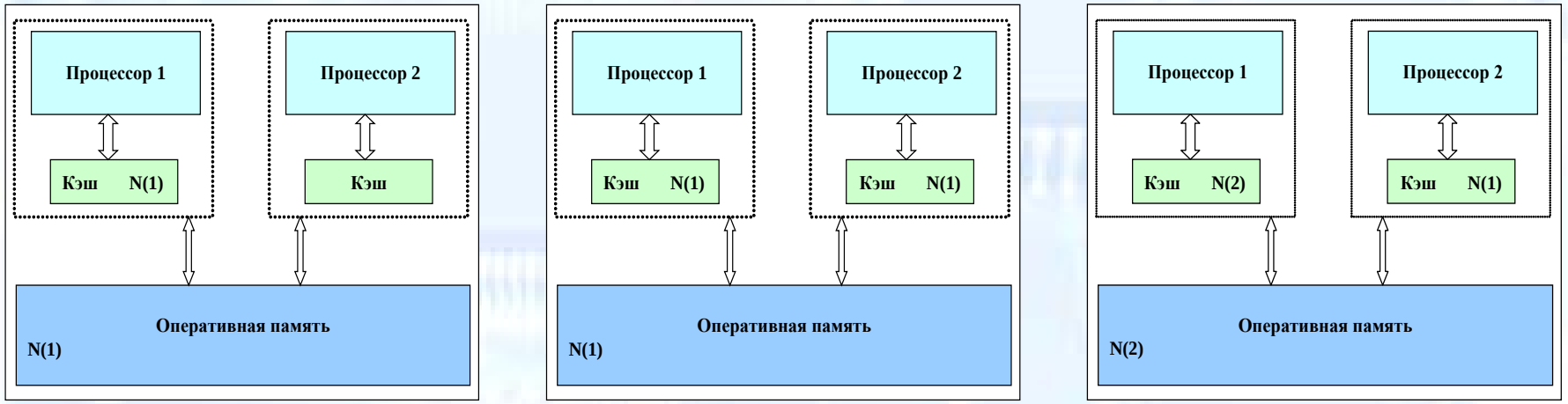
Проблемы:

  - Доступ с разных процессоров к общим данным и обеспечение, в этой связи, *однозначности (когерентности) содержимого разных кэшей* (*cache coherence problem*),
  - Необходимость *синхронизации взаимодействия* одновременно выполняемых потоков команд

Классификация вычислительных систем

- **Мультипроцессоры** с использованием единой *общей памяти…*

Проблема: *Обеспечение однозначности (когерентности) содержимого разных кэшей* (*cache coherence problem*)



1. Процессор 1 читает значение переменной N

2. Процессор 2 читает значение переменной N

3. Процессор 1 записывает новое значение переменной N

При изменении данных необходимо проверять наличие "старых" значений в кэш-памяти всех процессоров (обеспечивается на аппаратном уровне, но становится сложным при большом количестве процессоров)

Классификация вычислительных систем

- **Мультипроцессоры** с использованием единой *общей памяти…*

Проблема*: Необходимость синхронизации взаимодействия одновременно выполняемых потоков команд…*

Классификация вычислительных систем

Пример: Пусть процессоры выполняют последовательность команд

```
N = N + 1
Печать N
```

над общей переменной *N* (в скобках указывается значение этой переменной)

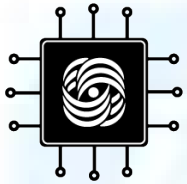
Вариант исполнения 1

Время	Процессор 1	Процессор 2
1	Чтение N (1)	
2		Чтение N (1)
3		Прибавление 1 (2)
4	Прибавление 1 (2)	
5	Запись N (2)	
6	Печать N (2)	
7		Запись N (2)
8		Печать N (2)

Вариант исполнения 2

Время	Процессор 1	Процессор 2
1	Чтение N (1)	
2	Прибавление 1 (2)	
3	Запись N (2)	
4	Печать N (2)	
5		Чтение N (2)
6		Прибавление 1 (3)
7		Запись N (3)
8		Печать N (3)

Временная последовательность команд может быть различной – необходима синхронизация при использовании общих переменных !



Классификация вычислительных систем

- **Мультипроцессоры** с использованием единой *общей памяти…*

Проблема*: Необходимость синхронизации взаимодействия одновременно выполняемых потоков команд…*

- Рассмотренный пример может рассматриваться как проявление общей *проблемы использования разделяемых ресурсов* (общих данных, файлов, устройств и т.п.)

Классификация вычислительных систем

- Для **организации разделения ресурсов** между несколькими потоками команд необходимо иметь возможность:

  - *определения доступности* запрашиваемых ресурсов (ресурс свободен и может быть выделен для использования, ресурс уже занят одним из потоков и не может использоваться дополнительно каким-либо другим потоком);

  - *выделения свободного ресурса* одному из процессов, запросивших ресурс для использования;

  - *приостановки (блокировки) потоков*, выдавших запросы на ресурсы, занятые другими потоками.

Классификация вычислительных систем

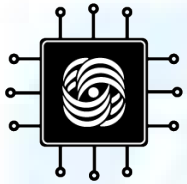
- **Мультипроцессоры** с использованием единой *общей памяти*

Проблема: *Необходимость синхронизации взаимодействия одновременно выполняемых потоков команд*

- Доступ к общей переменной в рассмотренном примере в самом общем виде должен быть организован следующим образом:

```
<Получить доступ>
  N = N + 1
  Печать N
<Завершить доступ>
```

- Полное рассмотрение проблемы синхронизации будет выполнено позднее при изучении вопросов параллельного программирования для вычислительных систем с общей памятью

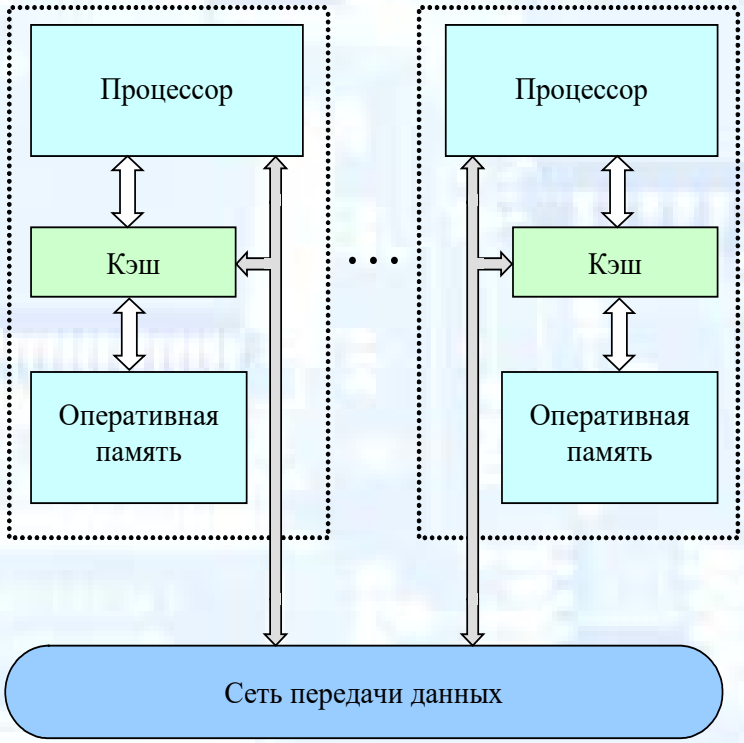


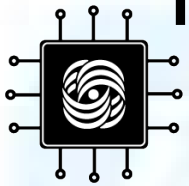
Классификация вычислительных систем

- **Мультипроцессоры** с использованием физически распределенной памяти (*distributed shared memory* or *DSM*):
  - *Неоднородный доступ к памяти* (*non-uniform memory access or NUMA*),
  - Среди систем такого типа выделяют:
    - *cache-only memory architecture or COMA* (системы KSR-1 и DDM),
    - *cache-coherent NUMA or CC-NUMA* (системы SGI Origin 2000, Sun HPC 10000, IBM/Sequent NUMA-Q 2000),
    - *non-cache coherent NUMA or NCC-NUMA* (система Cray T3E).

Классификация вычислительных систем

- **Мультипроцессоры** с использованием физически распределенной памяти…



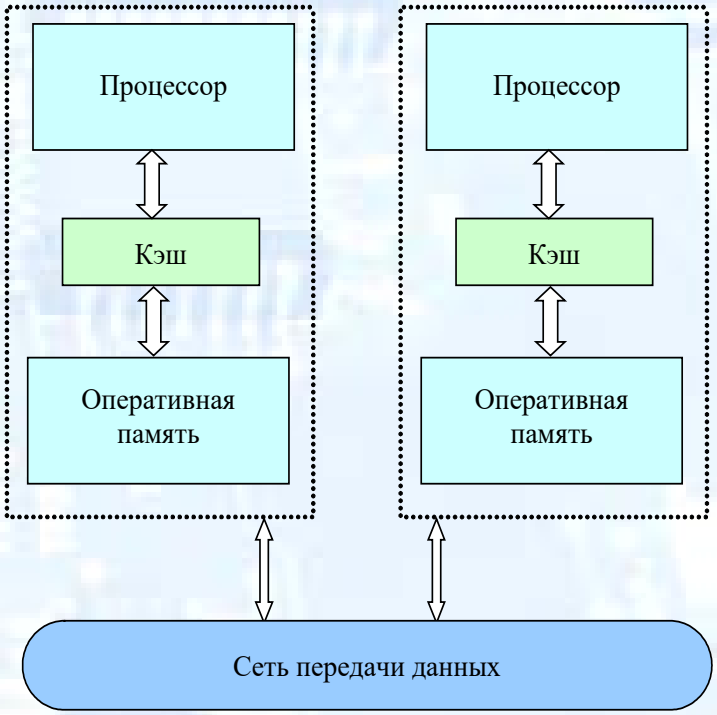


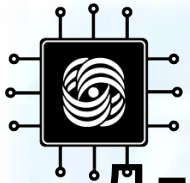
Классификация вычислительных систем

- **Мультипроцессоры** с использованием физически распределенной памяти:
  - Упрощаются проблемы создания мультипроцессоров (известны примеры систем с несколькими тысячами процессоров)
  - Возникают проблемы эффективного использования распределенной памяти (время доступа к локальной и удаленной памяти может различаться на несколько порядков).

Классификация вычислительных систем

- **Мультикомпьютеры**
  …
  - Не обеспечивают общий доступ ко всей имеющейся в системах памяти (*no-remote memory access or NORMA*),
  - Каждый процессор системы может использовать только свою локальную память





Мультикомпьютеры

- Для доступа к данным, располагаемых на других процессорах, необходимо явно выполнить *операции передачи сообщений* (*message passing operations*),
- Основные операции передачи данных:
  - Отправить сообщение (*send*),
  - Получить сообщение (*receive*)

Пример:

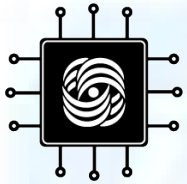
Процессор 1
<Отправить сообщение>
<Продолжение вычислений>

Процессор 2
<Получить сообщение>
<Продолжение вычислений
с использованием данных
полученного сообщения>

Мультикомпьютеры

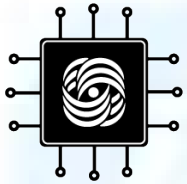
Данный подход используется при построении двух важных типов многопроцессорных вычислительных систем:

- *массивно-параллельных систем* (*massively parallel processor or MPP*), например: IBM RS/6000 SP2, Intel PARAGON, ASCI Red, транспьютерные системы Parsytec,
- *кластеров* (*clusters*), например: AC3 Velocity и NCSA NT Supercluster.



Кластеры

- **Преимущества**:
  - Могут быть образованы на базе уже существующих у потребителей отдельных компьютеров, либо же сконструированы из типовых компьютерных элементов;
  - Повышение вычислительной мощности отдельных процессоров позволяет строить кластеры из сравнительно небольшого количества отдельных компьютеров (*lowly parallel processing*),
  - Для параллельного выполнения в алгоритмах достаточно выделять только крупные независимые части расчетов (*coarse granularity*).

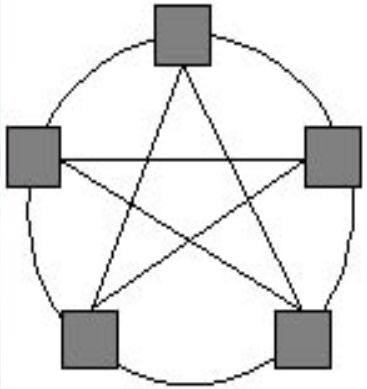


Кластеры

- **Недостатки**:
  - Организация взаимодействия вычислительных узлов кластера при помощи передачи сообщений обычно приводит к значительным временным задержкам
  - Дополнительные ограничения на тип разрабатываемых параллельных алгоритмов и программ (*низкая интенсивность потоков передачи данных*)

Топология сети передачи данных

– **полный граф** (*completely-connected graph* or *clique*) – система, в которой между любой парой процессоров существует прямая линия связи



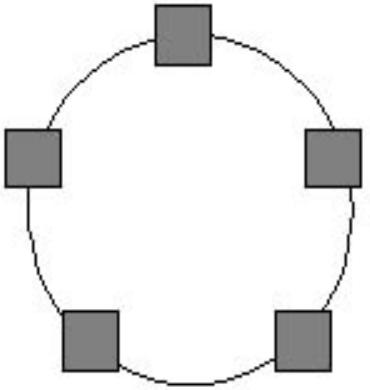
Полный граф (*completely-connected graph* or *clique*)

– **линейка** (*linear array* or *farm*) – система, в которой все процессоры перенумерованы по порядку и каждый процессор, кроме первого и последнего, имеет линии связи только с двумя соседними

Линейка *(linear array or farm)*

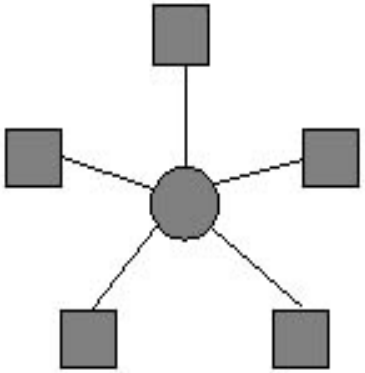
Топология сети передачи данных

– **кольцо** (*ring*) – данная топология получается из линейки процессоров соединением первого и последнего процессоров линейки



Кольцо (*ring*

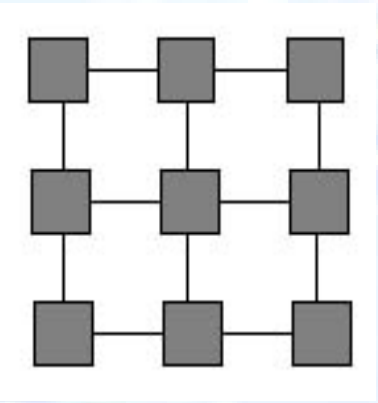
– **звезда** (*star*) – система, в которой все процессоры имеют линии связи с некоторым управляющим процессором



Звезда (*star*)

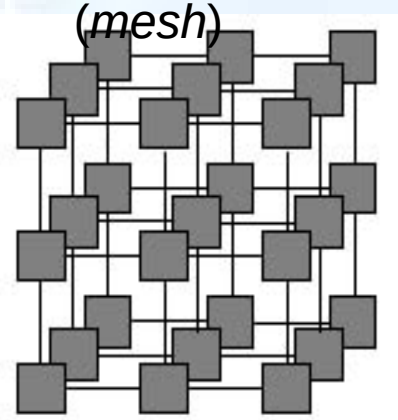
Топологии сети передачи данных

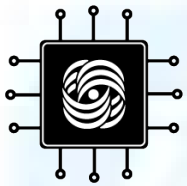
– **решетка** (*mesh*) – система, в которой граф линий связи образует прямоугольную сетку



Решетка (*mesh*)

– **гиперкуб** (*hypercube*) – данная топология представляет частный случай структуры решетки, когда по каждой размерности сетки имеется только два процессора.





Характеристики топологии сети

- ***диаметр*** – максимальное расстояние между двумя процессорами сети; характеризует максимально-необходимое время для передачи данных между процессорами,
- ***связность*** (*connectivity*) – минимальное количество дуг, которое надо удалить для разделения сети передачи данных на две несвязные области,
- ***ширина бинарного деления*** (*bisection width*) – минимальное количество дуг, которое надо удалить для разделения сети передачи данных на две несвязные области одинакового размера,
- ***стоимость*** – общее количество линий передачи данных в многопроцессорной вычислительной системе.

Характеристики топологии сети

Топология	Диаметр	Ширина бисекции	Связность	Стоимость
Полный граф	1	$p^2/4$	(p-1)	p(p-1)/2
Звезда	2	1	1	(p-1)
Линейка	p-1	1	1	(p-1)
Кольцо	$\lfloor p/2 \rfloor$	2	2	p
Гиперкуб	$\log_2 p$	p/2	$\log_2 p$	$p \log_2 p/2$
Решетка (N=2)	$2\lfloor \sqrt{p}/2 \rfloor$	$2\sqrt{p}$	4	2p

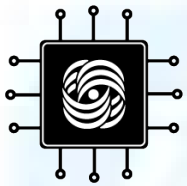
Суперкомпьютеры. Программа ASCI

(*Accelerated Strategic Computing Initiative*)

- **1996**, система **ASCI Red**, построенная Intel, производительность 1 TFlops,
- **1999**, **ASCI Blue Pacific** от IBM и **ASCI Blue Mountain** от SGI, производительность 3 TFlops,
- **2000**, ASCI White с пиковой производительностью свыше 12 TFlops (реально показанная производительность на тесте LINPACK составила на тот момент 4938 GFlops)

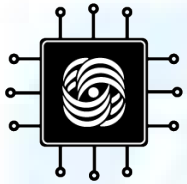
Суперкомпьютеры. ASCI White

- Система с 512-ю симметричными мультипроцессорными (SMP) узлами, каждый узел имеет 16 процессоров
- Процессоры IBM RS/6000 POWER3 с 64-х разрядной архитектурой и конвейерной организацией с 2 устройствами по обработке команд с плавающей запятой и 3 устройствами по обработке целочисленных команд, они способны выполнять до 8 команд за тактовый цикл и до 4 операций с плавающей запятой за такт, тактовая частота 375 MHz
- Оперативная память системы – 4 TB,
- Емкость дискового пространства 180 TB



Суперкомпьютеры. ASCI White

- Операционная система представляет собой версию UNIX – IBM AIX,
- Программное обеспечение ASCI White поддерживает смешанную модель программирования – передача сообщений между узлами и многопотоковость внутри SMP-узла,
- Поддерживаются библиотеки MPI, OpenMP, потоки POSIX и транслятор директив IBM, имеется параллельный отладчик IBM.



Суперкомпьютеры. Система BlueGene

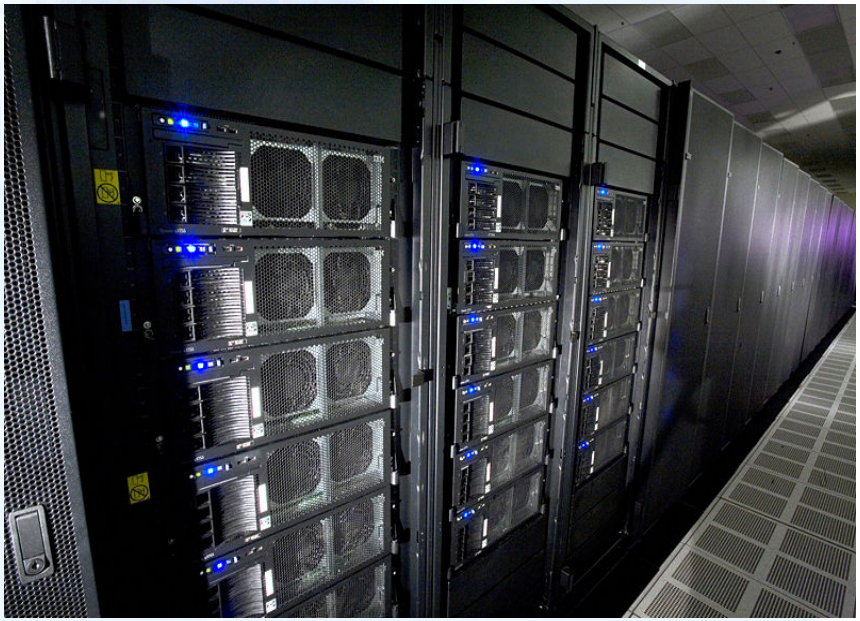
- Первый вариант системы представлен в 2004 г. и сразу занял 1 позицию в списке Top500
- Расширенный вариант суперкомпьютера (ноябрь 2007 г.) по прежнему на 1 месте в перечне наиболее быстродействующих вычислительных систем:
  - 212992 двухядерных 32-битных процессоров PowerPC 440 0.7 GHz,
  - пиковая производительность около 600 TFlops, производительность на тесте LINPACK – 478 TFlops

ASCI White & Blue Gene

Система RoadRunner

– **RoadRunner** является наиболее быстродействующей вычислительной системой (2008) и первым в мире суперкомпьютером, производительность которого превысила рубеж **1 PFlops** (1000 TFlops):

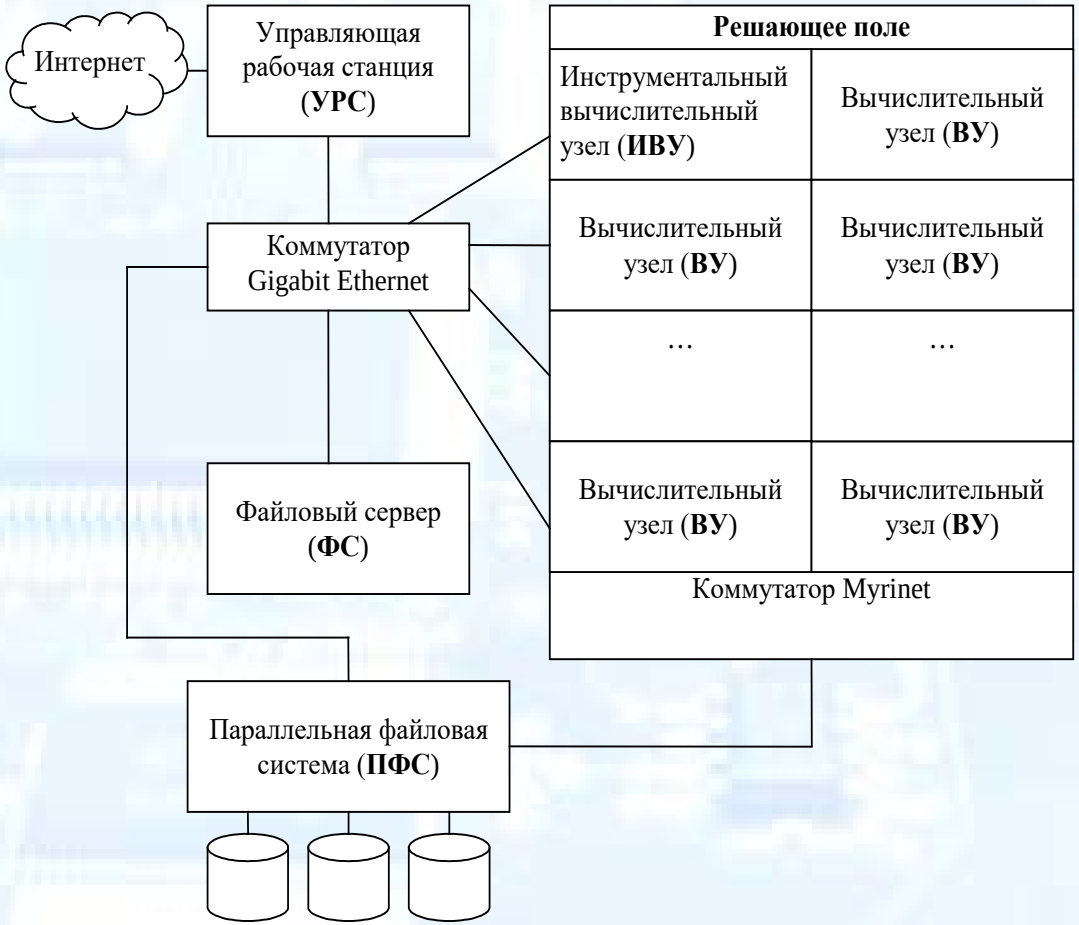
- **12960** процессоров IBM PowerXCell 8i и **6480** двухядерных процессоров AMD Opteron,
- пиковая производительность около **1700** TFlops, производительность на тесте LINPACK – **1026** TFlops



Суперкомпьютеры. МВС-15000

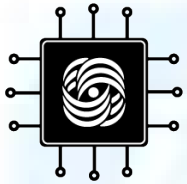
- Общее количество узлов 276 (552 процессора). Каждый узел представляет собой:
  - 2 процессора IBM PowerPC 970 с тактовой частотой 2.2 GHz, кэш L1 96 Kb и кэш L2 512 Kb,
  - 4 Gb оперативной памяти на узел,
  - 40 Gb жесткий диск IDE,
  - Операционная система SuSe Linux Enterprise Server версии 8 для платформ x86 и PowerPC,
  - Пиковая производительность 4857.6 GFlops и максимально показанная на тесте LINPACK 3052 GFlops.

Суперкомпьютеры. МВС-15000



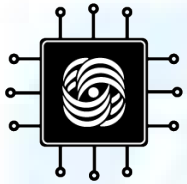
Суперкомпьютеры. СКИФ МГУ

- Общее количество двухпроцессорных узлов 625
  (1250 четырехядерных процессоров Intel Xeon E5472 3.0 ГГц),
  - Общий объем оперативной памяти – 5,5 Тбайт,
  - Объем дисковой памяти узлов – 15 Тбайт,
  - Операционная система Linux,
  - Пиковая производительность 60 TFlops, быстродействие на тесте LINPACK 47 TFlops.



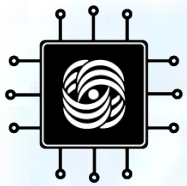
Кластер типа “Beowulf”

В настоящее время под кластером типа “*Beowulf*” понимается вычислительная система, состоящая из одного серверного узла и одного или более клиентских узлов, соединенных при помощи сети Ethernet или некоторой другой сети передачи данных. Это система, построенная из готовых серийно выпускающихся промышленных компонент, на которых может работать ОС Linux/Windows, стандартных адаптеров Ethernet и коммутаторов.



Кластер типа “Beowulf”

- **1994**, научно-космический центр NASA Goddard Space Flight Center, руководители проекта - Томас Стерлинг и Дон Бекер:
  - 16 компьютеров на базе процессоров 486DX4, тактовая частота 100 MHz,
  - 16 Mb оперативной памяти на каждом узле,
  - три параллельно работающих 10Mbit/s сетевых адаптера,
  - операционная система Linux, компилятор GNU, поддержка параллельных программ на основе MPI.



Кластеры. Beowulf

– **1998**, Система **Avalon**, Лос-Аламосская национальная лаборатория (США) , руководители проекта - астрофизик Майкл Уоррен:

- 68 процессоров (позднее расширен до 140) Alpha 21164A с тактовой частотой 533 MHz,
- 256 Mb RAM, 3 Gb HDD, Fast Ethernet card на каждом узле,
- операционная система Linux,
- пиковая производительность в 149 GFlops, производительность на тесте LINPACK 48.6 GFlops.

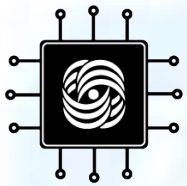
Кластеры.
AC3 Velocity Cluster

- **2000**, Корнельский университет (США), результат совместной работы университета и Advanced Cluster Computing Consortium, образованного компаниями Dell, Intel, Microsoft, Giganet:
  - 64 четырехпроцессорных сервера Dell PowerEdge 6350 на базе Intel Pentium III Xeon 500 MHz, 4 GB RAM, 54 GB HDD, 100 Mbit Ethernet card,
  - 1 восьмипроцессорный сервер Dell PowerEdge 6350 на базе Intel Pentium III Xeon 550 MHz, 8 GB RAM, 36 GB HDD, 100 Mbit Ethernet card,
  - операционная система Microsoft Windows NT 4.0 Server Enterprise Edition,
  - пиковая производительность AC3 Velocity 122 GFlops, производительность на тесте LINPACK 47 GFlops.

Кластеры. NCSA

NT Supercluster

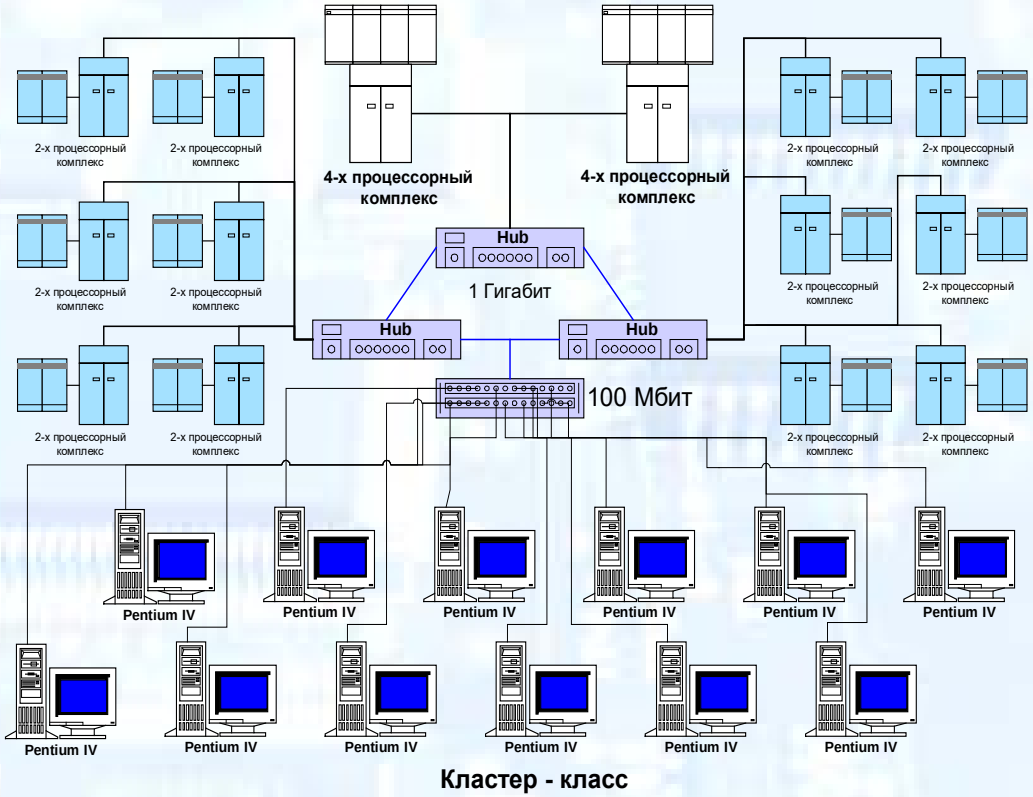
- **2000**, Национальный центр суперкомпьютерных технологий (National Center for Supercomputing Applications):
  - 38 двухпроцессорных систем Hewlett-Packard Kayak XU PC workstation на базе Intel Pentium III Xeon 550 MHz,
    1 Gb RAM, 7.5 Gb HDD, 100 Mbit Ethernet card,
  - операционная система ОС Microsoft Windows,
  - пиковая производительностью в 140 GFlops и производительность на тесте LINPACK 62 GFlops.



Кластеры. Thunder

- **2004**, Ливерморская Национальная Лаборатория (США):
  - 1024 сервера, в каждом по 4 процессора Intel Itanium 1.4 GHz,
  - 8 Gb оперативной памяти на сервер,
  - общая емкость дисковой системы 150 Tb,
  - операционная система CHAOS 2.0,
  - пиковая производительность 22938 GFlops и максимально показанная на тесте LINPACK 19940 GFlops (5-ая позиция списка Top500).

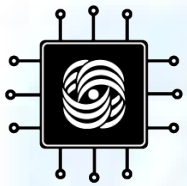
Кластеры. Вычислительный кластер ННГУ



Кластеры. Вычислительный кластер ННГУ

– **2007**,

- 64 вычислительных сервера, каждый из которых имеет
  2 двухядерных процессора Intel Core Duo 2,66 GHz, 4 GB RAM,
  100 GB HDD, 1 Gbit Ethernet card,
- пиковая производительность
- ~3 Tflops
- операционная система
- Microsoft Windows.



Спасибо за внимание!