Московский государственный университет имени М.В.Ломоносова
Факультет Вычислительной математики и кибернетики

СУПЕРКОМПЬЮТЕРЫ И ПАРАЛЛЕЛЬНАЯ ОБРАБОТКА ДАННЫХ (лекция 5)

А.С.Антонов

Вед. н.с. НИВЦ МГУ, к.ф.-м.н.

asa@parallel.ru

ВМК МГУ, 2024

Компьютеры с общей памятью

Архитектуры SMP и NUMA

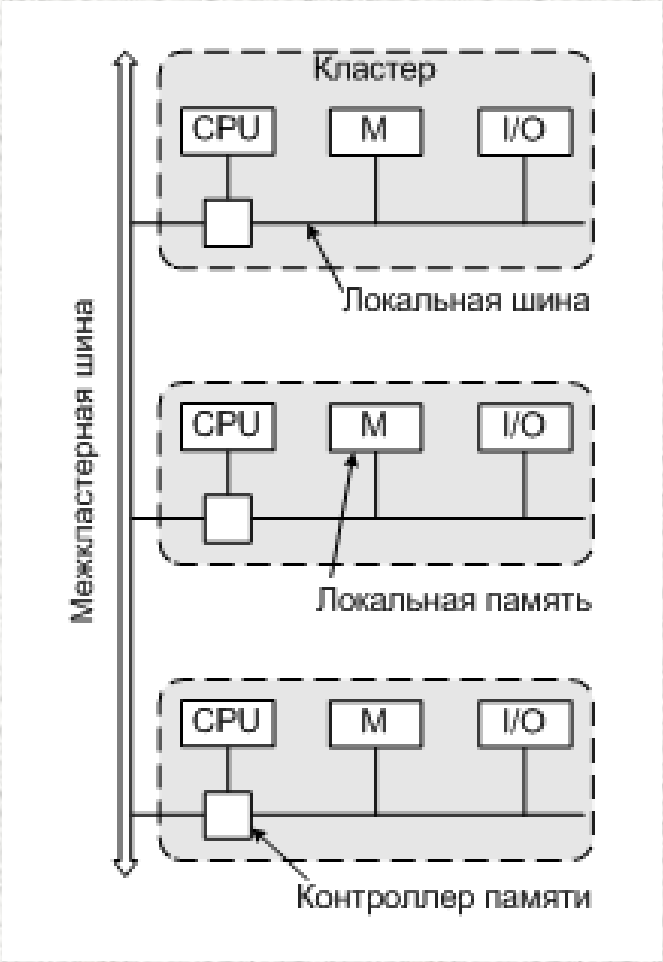
Иногда говорят, что классические SMP-компьютеры обладают архитектурой UMA (Uniform Memory Access), обеспечивая одинаковый доступ любого процессора к любому модулю памяти.

Как сохранить единое адресное пространство и одновременно сохранить возможность для увеличения числа процессоров?

…например, отказаться от однородности доступа к памяти, перейти от архитектуры SMP к архитектуре NUMA:

Non Uniform Memory Access.

Архитектура NUMA
(компьютер Cm*)



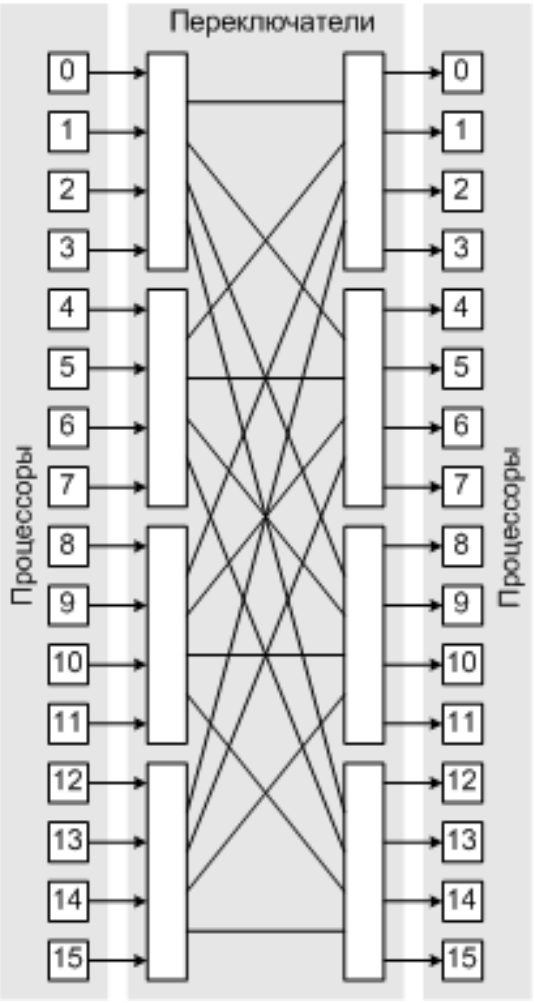
Конец 70-х годов 20-го века.

Контроллер памяти по старшим разрядам адреса определяет, в каком модуле хранятся нужные данные.

Запрос выставляется либо на локальную шину, либо на межкластерную шину.

Межкластерная шина – узкое место данной архитектуры.

Архитектура NUMA
(компьютер BBN Butterfly)



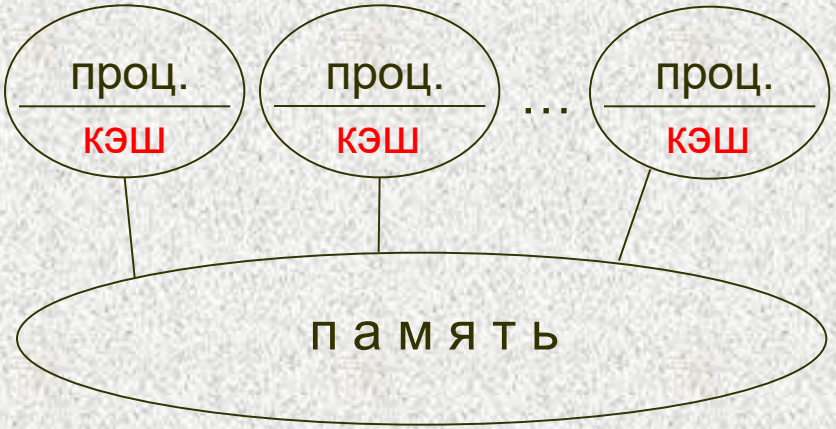
Максимальная конфигурация – 512 процессоров.

Связь процессоров с удалёнными блоками данных – через коммутатор Butterfly.

Локальные ссылки – около 2 мкс, удалённые – около 6 мкс.

Архитектура NUMA и ccNUMA

(все бы хорошо, да кэш-память появилась…)



В процессорах первых NUMA-компьютеров кэш-памяти не было.

Если процессор P1 сохранил значение x в ячейке q, а затем процессор P2 читает значение ячейки q, что он получит? Если x не вытеснено из кэш-памяти в оперативную память, то получит старое значение.

Проблема согласования содержимого кэш-памяти (cache coherence problem, проблема когерентности кэшей)

Архитектура NUMA и ccNUMA

(все бы хорошо, да кэш-память появилась…)

Специальная модификация NUMA-архитектуры:

ccNUMA:
cache coherent Non Uniform Memory Access.

Проблема когерентности кэшей решается на аппаратном уровне.

Важный вопрос – насколько неоднородна архитектура NUMA?

Если обращение к памяти другого узла требует на 5-10% больше времени, чем обращение к своей памяти, то к такой системе в большинстве случаев можно относиться, как к UMA (SMP).

Но чаще всего разница составляет 200-700%, и это существенно влияет на производительность.

Компьютеры с общей памятью
(Hewlett-Packard Superdome)

Выпускается с 2000 г. Крайне успешный – в 2001 году в Top500 занимал 147 позиций.

В стандартной комплектации объединяет от 2 до 64 процессоров с возможностью расширения системы.

Архитектура ccNUMA.

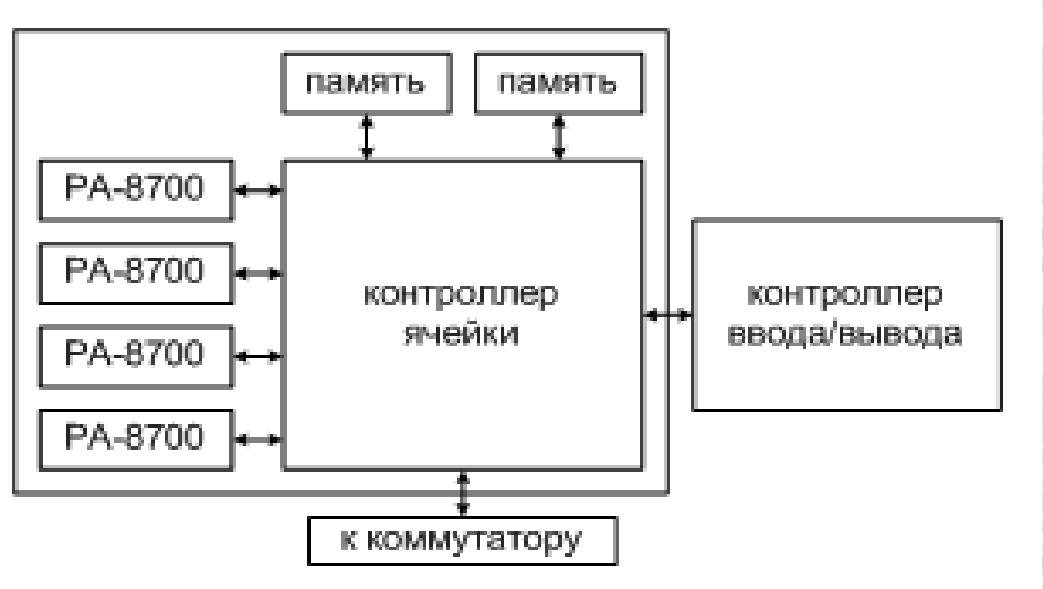
До 256 Гбайт оперативной памяти.

Используемые процессоры:
- *PA-8600, PA-8700, PA-8900*
- *Intel IA/64: Itanium, Itanium 2*

Компьютеры с общей памятью

(Hewlett-Packard Superdome)

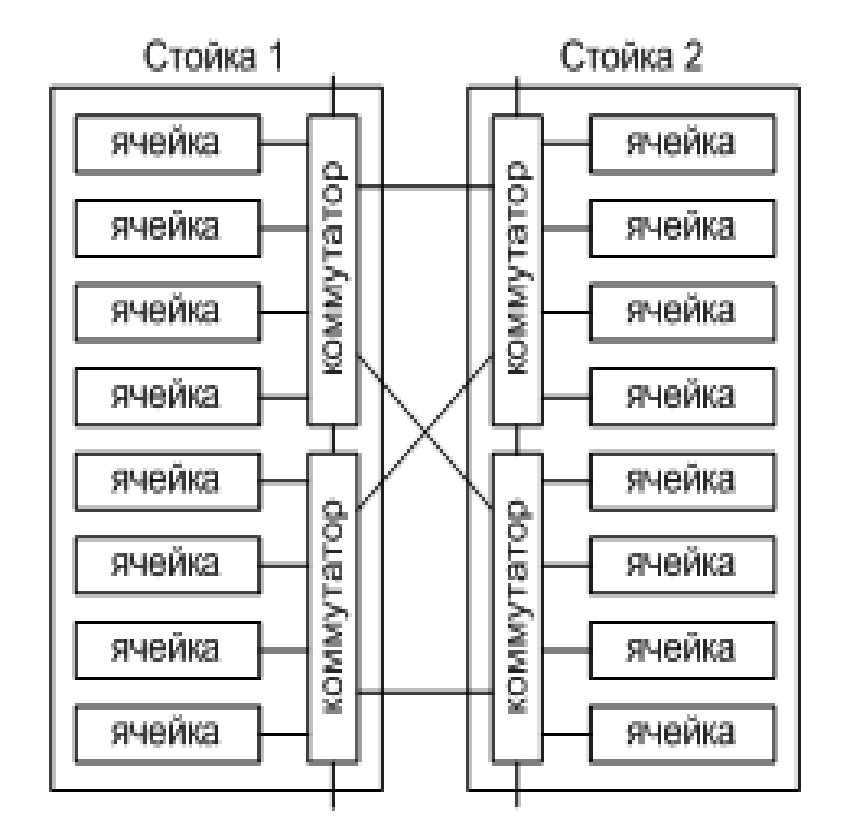
Вычислительные ячейки (cells):



До 4 процессоров.
До 16 Гбайт оперативной памяти, разделённой на 2 банка.
Контроллер ячейки (24 миллиона транзисторов) – интерфейсные функции, когерентность кэшей.
Канал процессор-контроллер – 2 Гбайт/сек.
Канал банк памяти-контроллер – 2 Гбайт/сек.
Канал контроллер-внешний коммутатор – 8 Гбайт/сек.

Компьютеры с общей памятью
(Hewlett-Packard Superdome)

Базовая конфигурация:



2 стойки.

В каждой стойке – 2 восьмипортовых коммутатора.

Все порты коммутаторов работают со скоростью 8 Гбайт/сек.

К каждому коммутатору подключено 4 ячейки.

3 других порта для связи с коммутаторами.

8-ой порт – для формирования многоузловой конфигурации.

Компьютеры с общей памятью

(Hewlett-Packard Superdome)

Виды задержек при обращении процессора к памяти, являющихся своего рода платой за высокую масштабируемость системы в целом:

- *процессор и память располагаются в одной ячейке – в этом случае задержка минимальна;*
- *процессор и память располагаются в разных ячейках, но обе эти ячейки подсоединены к одному и тому же коммутатору;*
- *процессор и память располагаются в разных ячейках, причём обе эти ячейки подсоединены к разным коммутаторам – в этом случае запрос должен пройти через два коммутатора, и задержки будут максимальными.*

В HP Superdome средняя задержка при переходе от 4 до 64 процессоров возрастает только в 1.6 раза.

Компьютеры с общей памятью

(Hewlett-Packard Superdome)

Компьютер HP Superdome может быть классическим единым компьютером с общей памятью. Однако его можно сконфигурировать и таким образом, что он будет являться совокупностью независимых разделов (nPartitions), работающих под различными операционными системами, в частности, под HP UX, Linux и Windows. Также предусмотрены организация эффективной работы с большим числом внешних устройств, возможности горячей замены всех основных компонент аппаратуры, резервирование, мониторинг базовых параметров.

Процессор PA-8700 имеет тактовую частоту 750 МГц, может выполнять 4 операции за такт, что даёт пиковую производительность 3 Гфлопс. Следовательно, пиковая производительность базовой 64-процессорной конфигурации – 192 Гфлопс.

Причины снижения производительности компьютеров с общей памятью

1. *Закон Амдала.*
2. *ccNUMA (неоднородность доступа к памяти).*
3. *ccNUMA (необходимость согласования кэш-памяти).*
4. *Конфликты при обращении в память.*
5. *Сбалансированность вычислительной нагрузки процессоров.*
6. *Производительность отдельных процессоров.*
7. *…*

Причин много, все они в той или иной мере проявляются в каждой программе. Их нужно ясно себе представлять, чтобы выбрать правильный способ решения.

Закон Амдала

$$S \le \frac{1}{\alpha + \frac{(1-\alpha)}{p}} \qquad S \approx \frac{1}{\alpha}$$

где:

α *– доля последовательных операций,*

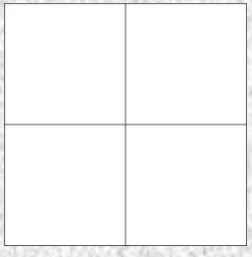
p – число процессоров в системе.

Если в программе 20% всех операций должны выполняться строго последовательно, то ускорения больше 5 получить нельзя вне зависимости от числа использованных процессоров.

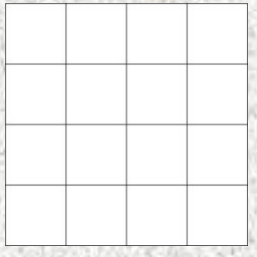
При использовании моделей параллельных программ с общей памятью возникают дополнительные участки последовательного кода, связанные с синхронизацией доступа к общим данным, например, критические секции. Реально эти фрагменты будут последовательными участками кода.

Сбалансированность вычислительной нагрузки процессоров

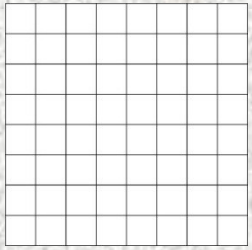
4 процессора

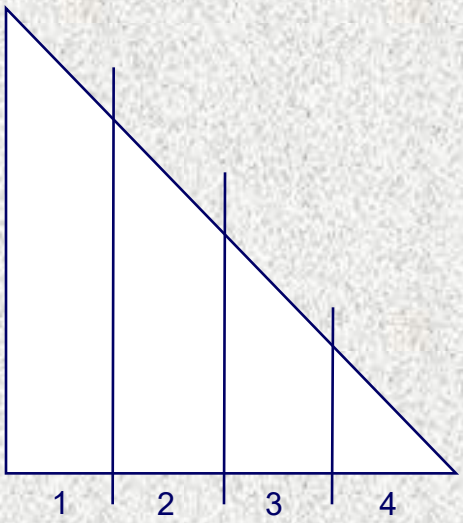


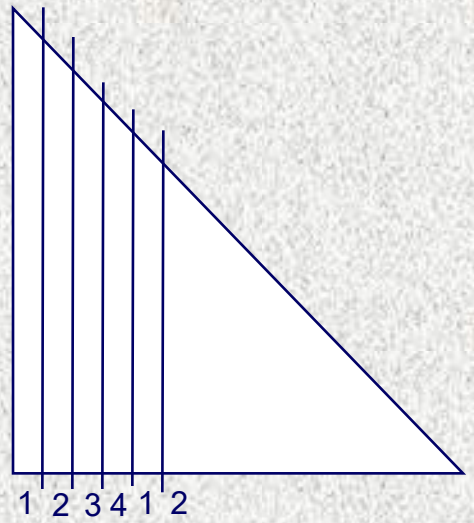
16 процессоров

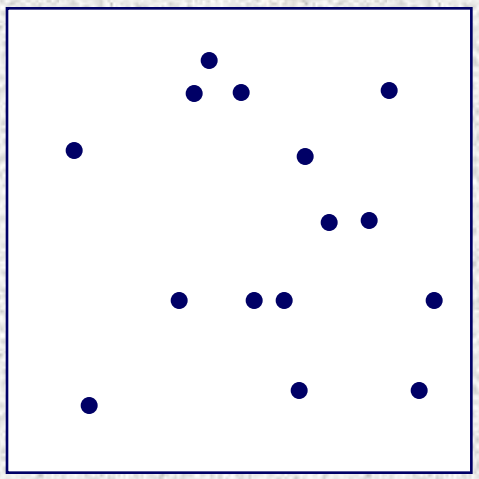


64 процессора









В случае систем с общей памятью ситуация упрощается тем, что практически всегда системы являются однородными, поэтому о сложной стратегии распределения работы речь, как правило, не идёт.

Компьютеры с распределенной памятью

Компьютеры с распределенной памятью

- *Используемые процессоры,*
- *Коммуникационная сеть.*

Примеры:

- *Intel Paragon: Intel i860, двумерная прямоугольная решетка.*
- *IBM SP1/SP2: IBM Power, высокоскоростной коммутатор каждый-с-каждым.*
- *«Ломоносов»: Intel Xeon + NVIDIA Tesla,QDR Infiniband, топология «толстое дерево»*

Компьютеры с распределенной памятью

Топология - конфигурация графа сети. Определяет взаимное расположение узлов и каналов связи.

В архитектуру сети помимо топологии входят также алгоритмы маршрутизации (routing) и управления потоками (flow control).

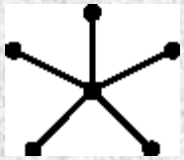
Компьютеры с распределенной памятью

Примеры коммуникационных топологий:

- *линейка – для системы из p узлов нужно p-1 соединение, средняя длина пути между двумя узлами равна p/3;*

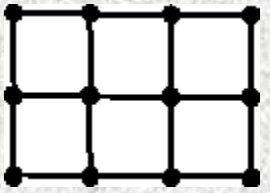
- *кольцо – для построения системы из p узлов нужно p соединений, средняя длина пути между двумя узлами равна p/6; увеличивается отказоустойчивость за счёт того, что передача сообщений может идти по двум направлениям;*

- *звезда – общение только через центральный узел, p-1 соединение, длина пути между концевыми узлами равна 2; большая нагрузка на центральный узел;*

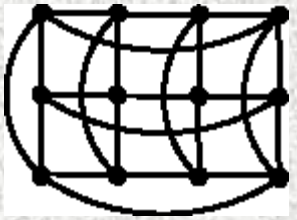


Компьютеры с распределенной памятью

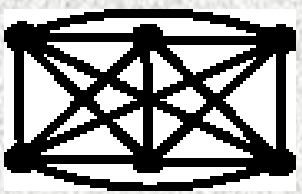
Примеры коммуникационных топологий:



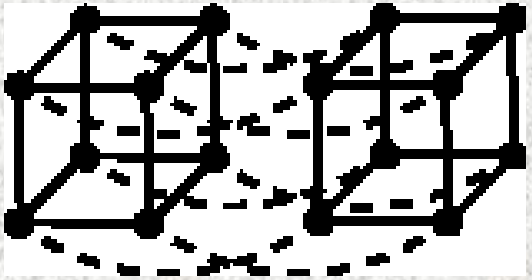
- *двумерная решётка, двумерный тор, … многомерные аналоги;*



- *полносвязная топология – каждый узел имеет связь с каждым другим узлом; для соединения p узлов требуется p(p-1)/2 связей;*



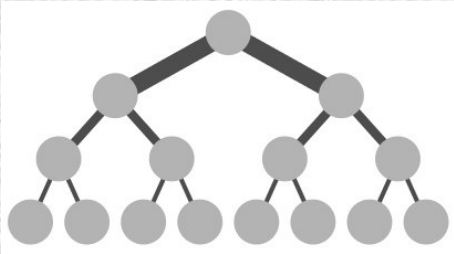
- *двоичный гиперкуб – в n-мерном пространстве поместим $p=2^n$ узлов системы в вершины единичного n-мерного куба; каждый узел соединим с соседом вдоль каждого из n измерений – всего $\log_2 p$ соединений; максимальное расстояние между вершинами n-мерного гиперкуба равно n.*



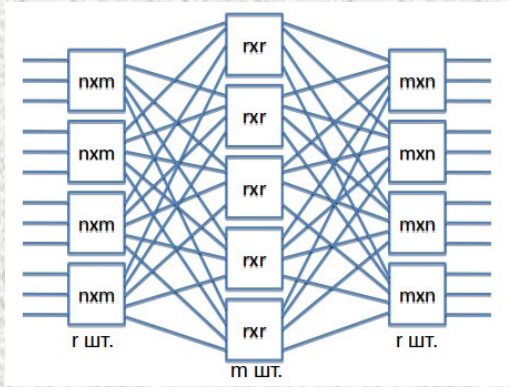
Компьютеры с распределенной памятью

Примеры коммуникационных топологий:

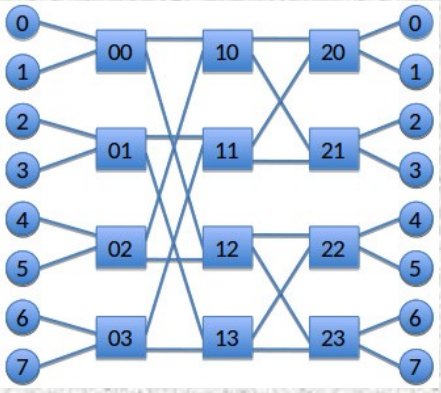
- *дерево, толстое дерево;*



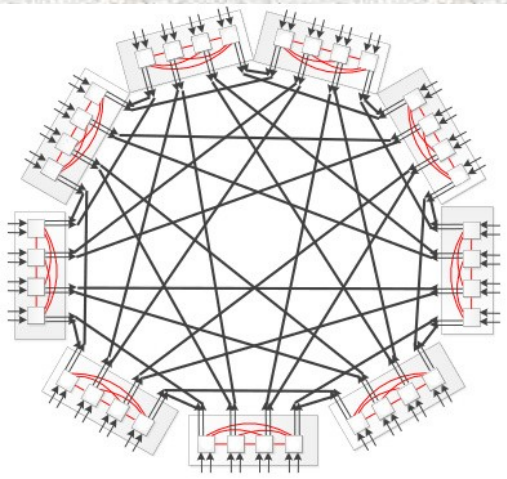
- *сети Клоза;*



- *бабочки (k-ary n-flies), flattened butterfly;*



- *dragonfly.*



Компьютеры с распределенной памятью

Параметры коммуникационных сетей:

- *Длина критического пути (диаметр) – минимальное количество элементарных связей, которые нужно пройти для коммутации двух самых удаленных процессоров.*
- *Связность – количество элементарных связей, которые нужно удалить, чтобы схема распалась на две несвязанные части.*
- *Сложность – общее количество необходимых элементарных связей.*

Компьютеры с распределенной памятью

Схема коммутации	**Длина критического пути**	**Связность**	**Сложность**
линейка	p-1	1	p-1
кольцо	$\lfloor p/2 \rfloor$	2	p
звезда	2	1	p-1
двумерная решётка	$2(\sqrt{p}-1)$	2	$2(p-\sqrt{p})$
двумерный тор	$2\lfloor \sqrt{p}/2 \rfloor$	4	2p
полносвязная топология	1	p-1	p(p-1)/2
двоичный гиперкуб	$\log_2 p$	$\log_2 p$	$p\log_2 p/2$

Компьютеры с распределенной памятью

Другие параметры коммуникационных сетей:

- *Количество портов маршрутизаторов (radix): low-radix/high-radix.*
- *Масштабируемость.*
- *Расширяемость.*
- *Простота эксплуатации.*
- *Устойчивость к отказам.*
- *Многообразие путей.*
- *…*

Компьютеры с распределенной памятью

(семейство Cray XT)

Семейство: T3D, T3E, XT3, XT4, XT5, XT6, XE6
Cray T3D – начало 90-х годов
Cray XT5 – на ноябрь 2010 г. второе место в Top500

Используемые процессоры:
T3D/T3E – DEC ALPHA,
XT3/XT4/XT5/XT6/XE6 – AMD Opteron,
XE6 + NVIDIA Tesla

Топология коммуникационной сети: трёхмерный тор.

Компьютеры с распределенной памятью

(семейство Cray XT)

Узлы суперкомпьютера:

- ***пользовательские**: компиляция, командные файлы, однопроцессорные задачи;*
- ***операционной системы**: выполнение многих системных сервисных функций ОС, в частности, работа с файловой системой;*
- ***вычислительные**: выполнение программ пользователя в монопольном режиме.*

Реальные конфигурации Cray T3E: 24/16/576 или 7/5/260 (пользовательские узлы/узлы ОС/вычислительные узлы)

Вычислительные узлы:
процессор, локальная память, …
+ сетевой интерфейс (он же – контроллер прямого асинхронного доступа в память).
Любой процессор через свой сетевой интерфейс может обращаться к памяти любого другого процессора, не прерывая его работы.

Компьютеры с распределенной памятью

(семейство Cray XT)

Коммуникационная сеть

Трёхмерный тор (в узлах тора сетевые маршрутизаторы):

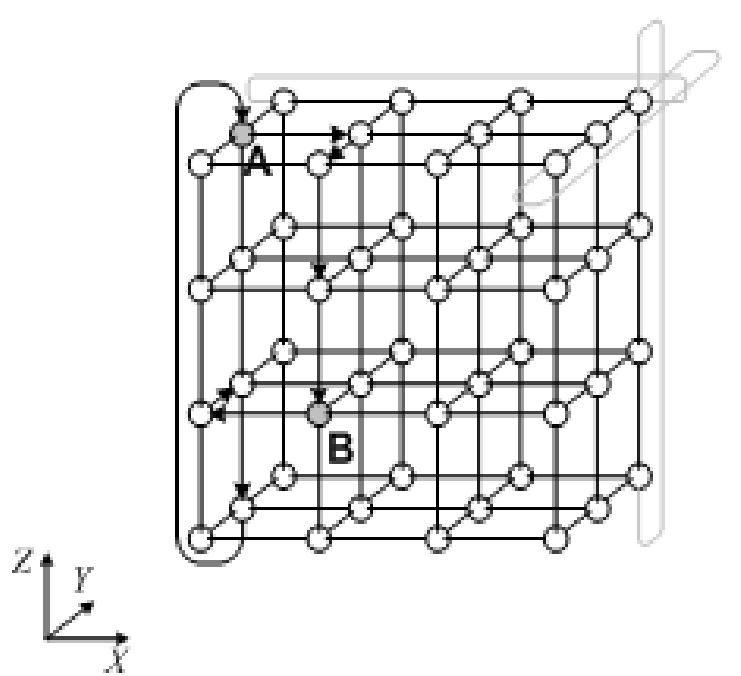
- *отказоустойчивость, возможность выбора альтернативного маршрута для обхода поврежденных связей;*
- *сокращение расстояния между узлами.*

У каждого узла всегда шесть непосредственных соседей.

Между узлами – два однонаправленных канала передачи данных, что допускает одновременный обмен данными в противоположных направлениях.

Компьютеры с распределенной памятью
(семейство Cray XT)

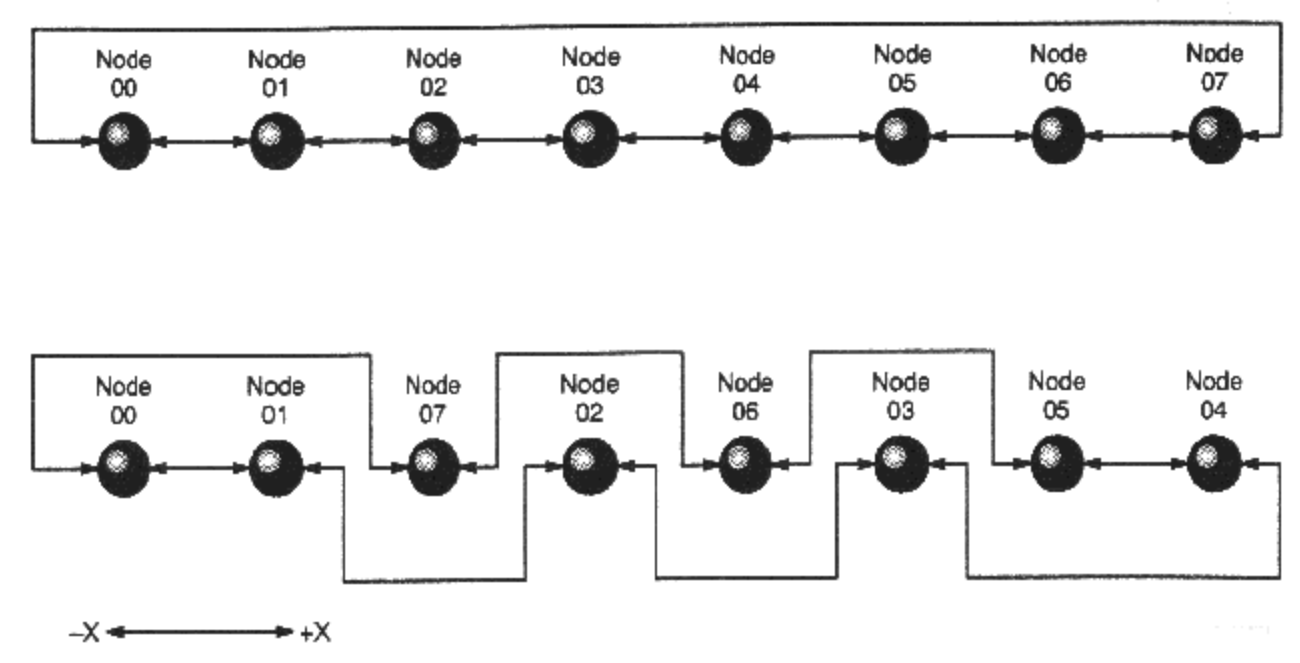
Коммуникационная сеть



Маршрутизация между вершинами A и B:
Сначала смещение по измерению X до тех пор, пока координата очередного транзитного узла и вершины B по измерению X не станут равными. Затем аналогичная процедура по Y, а в конце по Z.
Смещение может быть как положительным, так и отрицательным, минимизируя число перемещений по сети и обходя повреждённые связи.
Параллельный транзит данных по каждому из измерений X, Y и Z может выполняться одновременно.

Компьютеры с распределенной памятью
(семейство Cray XT)

Коммуникационная сеть

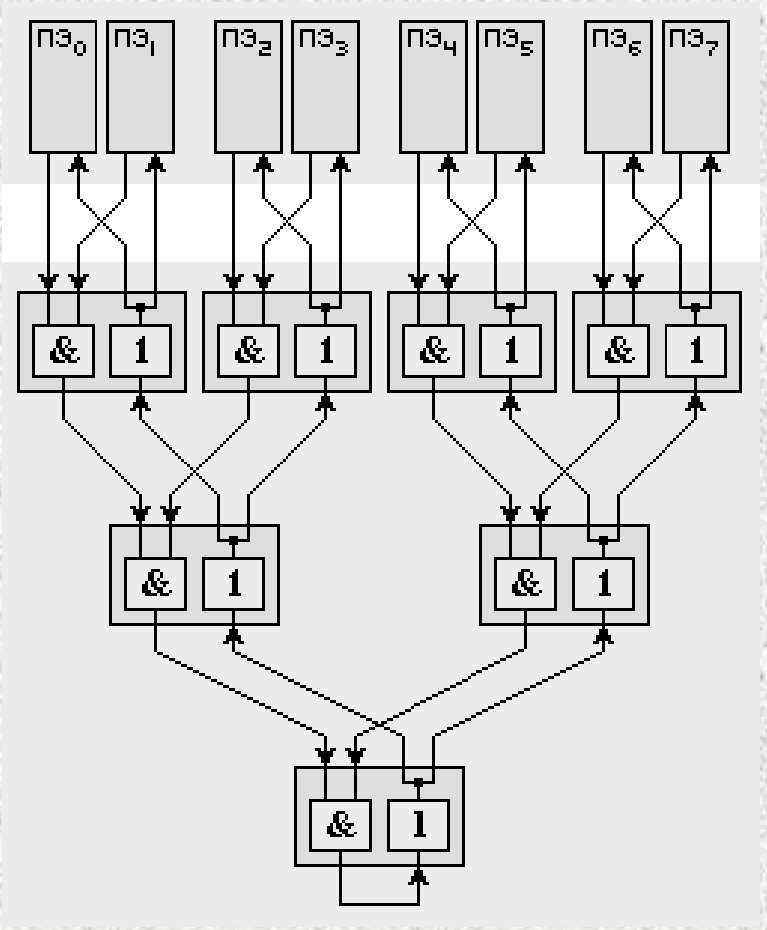


Все узлы в коммуникационной сети в размерностях X и Z расположены с чередованием, что позволяет минимизировать длину максимального физического соединения между узлами.

Компьютеры с распределенной памятью

(семейство Cray XT)

Аппаратная поддержка барьерной синхронизации:



Барьер - точка в программе, при достижении которой каждый процесс должен ждать, пока остальные процессы также не дойдут до барьера. После этого все процессы могут продолжать работу дальше.

Несколько входных и выходных регистров синхронизации. Каждый разряд этих регистров соединён со своей независимой цепью реализации барьера.
Двоичное дерево на основе двух типов устройств: логическое умножение (&) и дублирование входа на два своих выхода (1).

Кластеры, кластеры, кластеры…

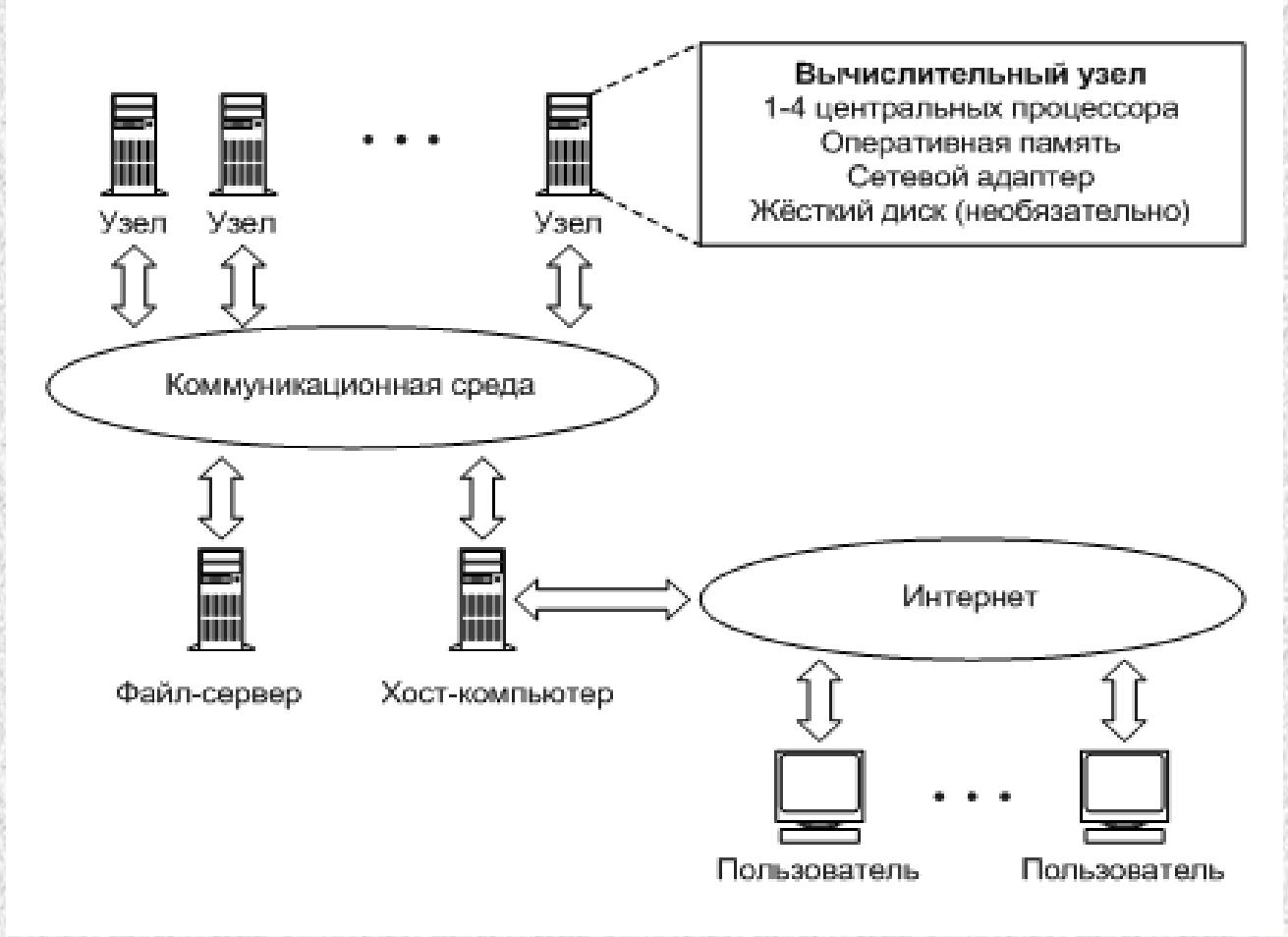
Вычислительный кластер - это совокупность компьютеров, объединённых в рамках некоторой коммуникационной сети для решения одной задачи.

- Доступные процессорные платформы
- Стандартные сетевые технологии
- Свободное ПО: Linux, MPI, GNU...
- Невысокая стоимость

438 компьютеров из Top500 мира – кластеры

(по данным 63-ей редакции списка Top500, июнь 2024 г.)

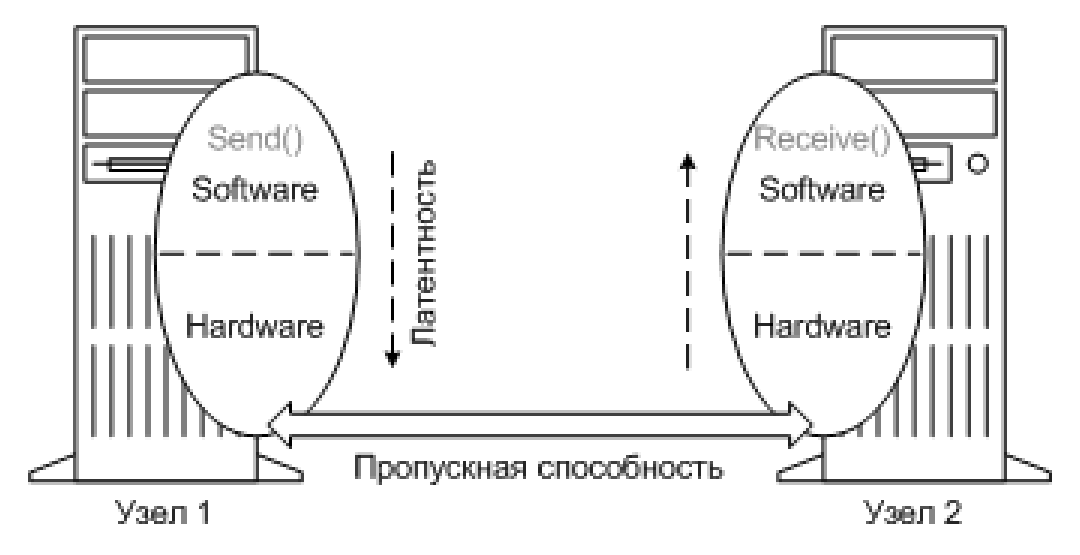
Компьютеры с распределенной памятью
(вычислительные кластеры)



Возможность формирования архитектуры кластера с учетом характерных особенностей решаемых задач

Компьютеры с распределенной памятью

(вычислительные кластеры)



Латентность и пропускная способность сети – основные параметры коммуникационной сети кластеров.

***Латентность** - время начальной задержки при посылке сообщений.*

***Пропускная способность** сети определяется скоростью передачи информации по каналам связи и измеряется объёмом передаваемой информации в единицу времени.*

Компьютеры с распределенной памятью

(вычислительные кластеры)

Время на передачу сообщения по коммуникационной сети вычисляется по следующей формуле:

$t_N = t_0 + N/S$,

Где t_0 - латентность, N - объём передаваемых данных, S - пропускная способность сети.

Если в программе много маленьких сообщений, то сильно скажется латентность.

Если сообщения передаются большими порциями, то важна высокая пропускная способность каналов связи.

Наличие латентности определяет и тот факт, что максимальная скорость коммуникаций не может быть достигнута на сообщениях с небольшой длиной.

Компьютеры с распределенной памятью

(вычислительные кластеры)

На практике пользователям не столько важны заявляемые производителем пиковые характеристики, сколько реальные показатели, достигаемые на уровне приложений, например, из программ, использующих технологию MPI.

Измерение “латентности” на практике:

- *процесс p_1 замеряет время t_1;*
- *процесс p_1 посылает сообщение длины 0 процессу p_2;*
- *процесс p_2 принимает сообщение от процесса p_1 и сразу посылает сообщение длины 0 обратно процессу p_1;*
- *процесс p_1 принимает сообщение от процесса p_2;*
- *процесс p_1 замеряет время t_2;*
- *под латентностью понимают величину $t_0 = (t_2 - t_1)/2$.*

Компьютеры с распределенной памятью

(вычислительные кластеры)

Пропускная способность и латентность сильно зависят от алгоритмов маршрутизации и топологии коммуникационной сети.

Важна также стоимость коммуникационной сети, которая складывается из:

- *Количества и стоимости маршрутизаторов.*
- *Длины и типа кабелей.*

Что снижает производительность компьютеров с распределенной памятью?

1. *Закон Амдала*
2. *Латентность передачи по сети*
3. *Пропускная способность каналов передачи данных*
4. *Особенности использования SMP-узлов*
5. *Балансировка вычислительной нагрузки*
6. *Возможность асинхронного счета и передачи данных*
7. *Особенности топологии коммуникационной сети*
8. *Производительность отдельных процессоров*
9. *...*

Московский государственный университет имени М.В.Ломоносова
Факультет Вычислительной математики и кибернетики

СУПЕРКОМПЬЮТЕРЫ И ПАРАЛЛЕЛЬНАЯ ОБРАБОТКА ДАННЫХ (лекция 5)

А.С.Антонов

Вед. н.с. НИВЦ МГУ, к.ф.-м.н.

asa@parallel.ru

ВМК МГУ, 2024