Архитектура вычислительных систем. Лекция 3

- Векторно-конвейерные вычислительные системы
- Реализация параллелизма в Cray C90
- Факторы, влияющие на производительность ВК-систем. Примеры ВК-систем
- Реализация векторных вычислений в современных процессорах
- Массивно-параллельные компьютеры с распределенной памятью
- Характеристика Cray T3D/T3E
- Факторы, влияющие на производительность МП-систем. Примеры МП-систем
- Вычислительные кластеры
- Примеры суперкомпьютерных проектов

Векторно-конвейерные системы (1)

Первый ВК компьютер Cray1 появился в 1976. Архитектура его оказалась столь удачной, что он положил начало целому семейству компьютеров, название которому (PVP) дали 2 принципа в архитектуре процессоров:

- конвейерная организация обработки потока команд (pipe)
- введение в систему команд набора векторных операций, которые позволяют оперировать с целыми массивами данных (vector).

Основной признак PVP-систем – наличие специальных ВК процессоров, в которых предусмотрены команды однотипной обработки векторов независимых данных, эффективно выполняющиеся на конвейерных ФУ.

Длина одновременно обрабатываемых векторов в современных ВК ЭВМ составляет, как правило, 128 или 256 элементов. Основное назначение векторных операций состоит в распараллеливании выполнения операторов цикла, в которых обычно сосредоточена основная часть вычислительной работы. Циклы подвергаются процедуре векторизации, чтобы они могли реализовываться с помощью векторных команд. Это выполняется автоматически компиляторами, поэтому ВК ЭВМ не требовали специальной технологии программирования, что явилось решающим фактором в их успехе на компьютерном рынке.

Требовалось лишь соблюдение некоторых правил при написании циклов, чтобы компилятор мог их эффективно векторизовать.

Внезапное и очень резкое увеличение производительности ЭВМ за счет внедрения векторной обработки послужило причиной внезапного роста популярности ВК ЭВМ. В 80е векторные машины практически стали обыденным явлением и почти синонимом суперкомпьютеров.

Векторно-конвейерные системы (2)

- С середины 90х доля ВК ЭВМ в суперкомпьютерном парке неуклонно снижается ввиду их дороговизны и невысокой масштабируемости.
- Интерес к этому направлению в значительной степени был ослаблен и в связи появлением достаточно мощных суперскалярных микро-процессоров.
- Суперкомпьютеры с ВК архитектурой стали проигрывать системам с массовым параллелизмом.
- Уровень развития микроэлектронных технологий не позволяет в настоящее время производить однокристальные векторные процессоры.
- **Однако!!!** идеи векторизации нашли воплощение в технологиях типа SSE, 3DNow!, поддерживаемых современными микропроцессорами, а также в графических процессорах и процессорах Cell.
- В марте 2002 корпорация NEC представила систему Earth Simulator из 5120 ВК процессоров, которая в 5 раз превысила производительность предыдущего обладателя рекорда - MPP системы ASCI White из 8192 суперскалярных микропроцессоров. Это заставило многих по-новому взглянуть на перспективы ВК систем.
- В редакции списка TOP500 за ноябрь 2011 этот компьютер все еще входил в первую сотню (занимал д94е место).
- В настоящее время ВК процессоры включают в состав высокопроизводительных суперкомпьютеров с гибридной архитектурой.

Векторизация вычислений (1)

Пример – поэлементное сложение двух одномерных массивов B[ndim] и C [ndim] размерности ndim и запись результата в A [ndim]: **A=B+C.**

Данной операции присущ *параллелизм* – сложение ***i***-х элементов массивов **A[i]** и **B[i]** может исполняться параллельно для разных ***i***.

```
for( i = 0; i<ndim; i++)
        A[i] = B[i]+C[i];
```

Скалярный режим: цикл выполняется последовательно, компилятор генерирует инструкции для следующей последовательности действий

1. прочитать элемент B[i],
2. прочитать элемент C[i],
3. выполнить сложение,
4. записать результат в A[i],
5. увеличить параметр цикла,
6. проверить условие цикла, вернуться к п.1, если оно не выполнено

Векторный режим: фрагмент будет выполняться иначе:

- загрузить 128-элементные порции массивов B, C (эти две операции будут выполняться со сдвигом в один такт, т.е. почти одновременно),
- векторное сложение,
- запись порции массива A в память,
- При ndim>128 повторить эту последовательность нужное число раз

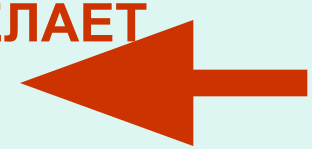
Векторизация вычислений (2)

Таким образом, если размерность векторов NDIM меньше или равна длине вектора процессора NA (в нашем случае 128), то сложение векторов производится за один такт процессора, если же размерность NDIM>NA, то за несколько тактов (пренебрегли накладными затратами).

Это называется ***сегментация (секционирование)***. Условно, при **векторизации вычислений** цикл преобразуется в:

```
Do j = 1, [ndim/128]
  Do i=1,128
        k = (j-1)*128+i
        B1(i) = B(k)
        C1(i) = C(k)
        A1(i) = B1(i)+C1(i)
        A(k) = A1(i)
  EndDo
EndDo
```

ДЕЛАЕТ КОМПИЛЯТОР!



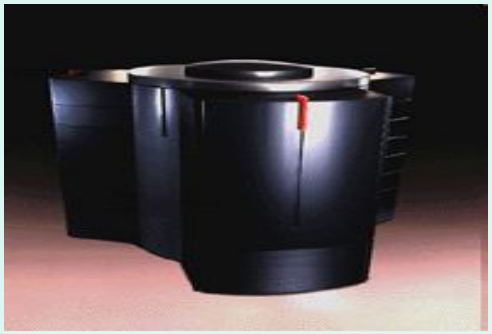
Плюс выполнение операции над остатком ndim-[ndim/128].

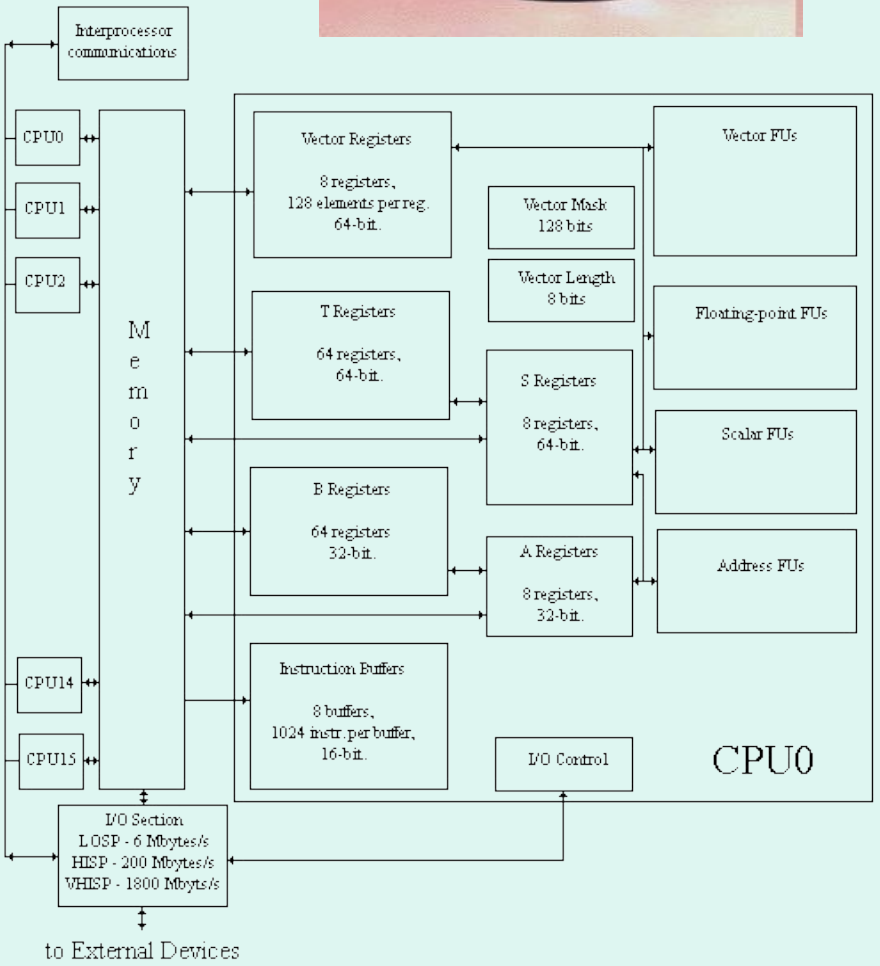
Применение средств векторной обработки позволяет получить выигрыш в скорости вычислений по сравнению с последовательной обработкой.

- НО появились дополнительные операции
- НО требуется время (***инициализация, startup***), связанное с заполнением конвейера и подкачкой аргументов.
- Чем больше длина векторов, тем эффективнее векторизация

Компьютер Cray C90 (1)

- CRAY Y-MP C90 – пример классического ВК-компьютера.
- Объединет в максимальной конфигурации 16 процессоров, работающих над общей памятью.
- Время такта 4.1нс, что соответствует тактовой частоте почти 250MHz.
- Пиковая производительность одного процессора 1Гфлопс.
- Более поздние конфигурации – Cray SV1 Cray T90, Cray X1.
- Системы серии T90 базируются на ВК-процессоре Cray Research с пиковой производительностью 2GFlop/s.
- В Cray X1 используются 16-конвейерные векторные процессоры, пиковая производительность 12.8GFlop/s. Максимальная конфигурация - 4096 процессоров.





Компьютер Cray C90 (2)

Все процессоры имеют одинаковую вычислительную секцию, состоящую из регистров, ФУ и сети коммуникаций.

Регистры и ФУ могут обрабатывать и хранить три типа данных

– адреса (*A*-регистры, *B*-регистры),

– скаляры (*S*-регистры, *T*-регистры),

– вектора (*V*-регистры).

РЕГИСТРЫ. Каждый процессор имеет три набора основных регистров (*A*, *S*, *V*), которые имеют связь как с памятью, так и с ФУ. Для регистров *A* и *S* существуют промежуточные наборы регистров *B* и *T*, играющие роль буферов для основных регистров.

- Адресные регистры: *A*-регистры, 8 штук по 32 разряда, для хранения и вычисления адресов, индексации, указания величины сдвигов, числа итераций циклов и т.д. *B*-регистры, 64 штуки по 32 разряда.
- Скалярные регистры: *S*-регистры, 8 штук по 64 разряда, для хранения аргументов и результатов скалярной арифметики, иногда содержат операнд для векторных команд. *T*-регистры, 64 штуки по 64 разряда.

  Скалярные регистры используются для выполнения как скалярных, так и векторных команд.
- Векторные регистры:

  V-регистры, 8 штук на 128 64-разрядных слова каждый.

  Векторные регистры хранят аргументы и результаты векторных команд, используются только для выполнения векторных команд.

Компьютер Cray C90 (3)

ФУНЦИОНАЛЬНЫЕ УСТРОЙСТВА исполняют свой набор команд и могут работать одновременно друг с другом. Все ФУ конвейерные и делятся на ***четыре группы:***

1. ***Адресные*** (2шт): целочисленное сложение/вычитание, целочисленное умножение
2. ***Скалярные*** (4шт): целочисленное сложение/вычитание, логические поразрядные операции, сдвиг и т.п.
3. ***Векторные*** (5-7шт): целочисленное сложение/вычитание, сдвиг, логические поразрядные операции, умножение битовых матриц; предназначены только (!) для выполнения векторных команд
4. ***Для работы с плавающей точкой*** (3шт): сложение/вычитание, умножение, обратная величина; предназначены для выполнения и векторных, и скалярных команд

Векторные ФУ и ФУ с плавающей точкой продублированы: векторные команды разбивают 128 элементов векторных регистров на четные и нечетные, обрабатываемые одновременно двумя конвейерами (pipe 0, pipe 1). В скалярных операциях, использующих ФУ с плавающей точкой, работает только один конвейер.

ФУ имеют различное число ступеней конвейера, но каждая ступень срабатывает за один такт, поэтому при полной загрузке все ФУ могут выдавать результат каждый такт.

Архитектура CRAY C90 позволяет использовать регистр результатов векторной операции как входной регистр следующей векторной операции, т.е. выход сразу подается на вход (***зацепление вект. оп-й)***

Компьютер Cray C90 (4)

Параллельное выполнение программ

1. Конвейеризация выполнения команд. Все основные операции, выполняемые процессором – обращение в память, обработка команд, выполнение инструкций – являются ***конвейерными***.

2. Независимость функциональных устройств. Большинство ФУ в CRAY C90 являются независимыми, поэтому несколько операций могут выполняться ***одновременно.*** В частности, векторные операции, использующие различные ФУ и регистры, могут выполняться параллельно.

3. Векторная обработка. Векторная обработка увеличивает скорость и эффективность обработки за счет того, что обработка целого набора данных выполняется одной командой. Скорость выполнения операций в векторном режиме приблизительно в 10 раз выше скорости скалярной обработки.

4. Зацепление ФУ. Возможность выполнения нескольких векторных операций в режиме «макроконвейера» дает дополнительный выигрыш в скорости. Глубина зацепления может быть любой (чтение векторов, сложение, умножение, запись).

5. Дублирование ФУ. Векторные данные обрабатываются одновременно двумя конвейерами для четных и нечетных индексов

6. Многопроцессорная обработка. В максимальной конфигурации может содержать 16 независимых процессоров, которые могут выполнять одновременно несколько независимых программ, либо могут быть назначены для выполнения одной программы в параллельном режиме.

Факторы, влияющие на производительность ВК систем

Факторы, уменьшающие реальную производительность.

- Конфликты в памяти (обращение к одному эл-ту памяти из разных у-в)
- Ограниченная пропускная способность каналов связи
- Затраты на секционирование векторных операций (сегментация)
- Затраты времени на разгон конвейера (инициализация)
- Неэффективность векторизации конкретного кода:
  - зависимость по данным,
  - вложенные циклы,
  - обращение внутри цикла к неизвестной функции

Для успешной векторизации циклов необходимы следующие условия:

(а) цикл не должен быть внутренним;

(б) не допустимы ветвления (условные операторы) внутри тела цикла;

(в) недопустимы вызовы внешних функций и процедур из тела цикла;

(г) не должно быть рекурсии элементов векторов или массивов в теле цикла.

- Ограниченный набор векторных регистров
- Несбалансированность использования векторных ФУ из-за их специализации
- Отсутствие операции деления (Cray C90 и ряд других машин)
- Наличие скалярных операций (закон Амдала)
- Качество используемого компилятора

Примеры систем с векторными процессорами

Earth Simulator – создан в 2002, установлен в Японском центре морских наук и технологий (Japan Marine Science and Technology Center). Занимает здание размером 50x65x17м.

- 640 процессорных узлов, соединенных через высокоскоростной переключатель. В составе узла 8 векст. арифм. процессоров, работающих над общей памятью.
- Пиковая производительность 40 Тфлопс.
- В 2002-2003 – 1е место TOP500, в 2012 – 94е, в 2013 – 342е, т.е. >10 лет в продержался в TOP500.

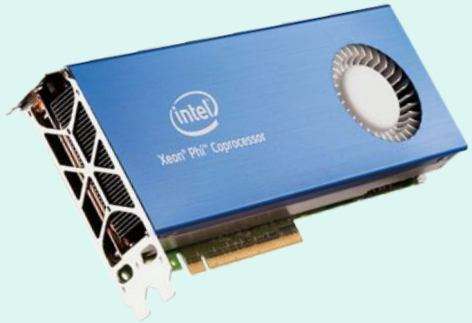
Cray XT5h – один из первых суперкомпьютеров на основе гибридной архитектуры.

- Объединяет скалярные, векторные процессоры, а также процессоры на основе ПЛИС.
- Векторные вычисления: блэйд-сервер Cray X2, состоит из 2 выч. узлов, каждый из которых имеет 4 1-ядерных векторных процессора, работающих над общей памятью.
- Система масштабируется до 32000 процессоров, при этом работа обеспечивается в едином адресном пространстве.

Реализация векторных вычислений в современных процессорах

- В настоящее время векторные расширения присутствуют в наборах команд процессоров различных производителей и архитектур.
- Технология основана на присутствии специализированных регистров с большим числом разрядов и наборе инструкций, обеспечивающих одновременное выполнение арифметических действий над всеми числами, размещенными в таких регистрах.
- ЦПУ Intel: наборы команд MMX – SSE – SSE2 – SSE3 – SSE4 – AVX эволюционировали от векторной целочисленной арифметики (MMX), организованной на специализированных 64-разрядных регистрах и ориентированной преимущественно на обработку звука и видео, до широкого спектра арифметических операций с плавающей запятой, работающей на 128—разрядных (SSE) и 256-разрядных (AVX) регистрах.
- ЦПУ AMD: дополнение 3DNow!; ЦПУ архитектуры ARM: расширение NEON; синергетические ПЭ в процессорах Cell и др.
- Векторные расширения, реализованные в ядрах ускорителей Intel Xeon Phi

Реализация векторных вычислений в Intel Xeon Phi



- Intel Xeon Phi имеют 60 и более ядер x86 с 512-разрядными векторными модулями, работающими на частоте от 1 ГГц, что обеспечивает скорость вычислений двойной точности более 1 TFLOPS.
- В каждом ядре сконструирован специальный модуль векторной обработки (VPU, vector processor unit), содержащий 32 512-разрядных регистра (zmm-регистры) и открывающий новые возможности по обработке данных с использованием векторных инструкций вида $a=a+b*c$, а также некоторых других возможностей.
- Учитывая размер векторного регистра, на каждом ядре Intel Xeon Phi мы можем одновременно производить действия над 16 32-битными целыми числами или 8 64-битными целыми числами или 16 числами с плавающей запятой одинарной точности или 8 числами с плавающей запятой двойной точности.
- Расширенная поддержка некоторых математических функций. Так, появились команды для вычисления в одинарной точности функций 1/x, 1/sqrt(x), $\log_2(x)$, 2^x.

Массивно-параллельные компьютеры с распределенной памятью (1)

Довольно долго (с начала 70х) ВК ЭВМ составляли преобладающее большинство среди высокопроизводительных систем. Однако прогресс в развитии аппаратных и программных средств организации межпроцессорного взаимодействия в системах с большим количеством процессоров к началу 90х сделал реальностью появление на рынке систем с массовым параллелизмом (MIMD-компьютеров по Флинну).

MIMD-компьютер представляет собой набор процессорных узлов (processor node), состоящих как минимум из одного процессора и приданного этому процессору банка памяти. Процессорные узлы соединены между собой в единую систему с помощью высокоскоростной шины или коммутатора. По сути, компьютер представляет собой сеть (матрицу) из ПЭ.

По типу организации доступа к памяти МП компьютеры делятся на машины с распределенной памятью (distributed memory) и машины с общей или разделяемой памятью (shared memory).

Основные причины появления массово-параллельных компьютеров

(1) необходимость построения ЭВМ с гигантской производительностью,

(2) необходимость производства компьютеров в большом диапазоне как производительности, так и стоимости.

Для массивно-параллельного компьютера, в котором число процессоров может сильно меняться, всегда можно подобрать конфигурацию с заранее заданной производительностью и/или стоимостью.

Массивно-параллельные компьютеры с распределенной памятью (2)

- В компьютерах с распределенной памятью каждый процессор непосредственно может адресовать только собственную локальную память. Как правило, объем памяти в процессорных узлах невелик и для больших программ необходим специальный инструментарий для организации глобального доступа к памяти.
- Широкое распространение компьютеры данного класса получили с начала 90х.
- По большому счету различие между разными системами состоит в типе используемых процессоров, организации коммуникационной среды и в наличии\отсутствии хост-компьютера.
- К классу систем с распределенной памятью помимо массивно-параллельных (матричных) суперкомпьютеров можно отнести также кластеры и сети компьютеров.

Архитектура CRAY T3D/T3E (1)

Компьютеры CRAY T3D\T3E – классические МП-компьютеры с распределенной памятью, объединяющие от 32 до 2048 процессоров.

Как и любые другие компьютеры данного класса, содержат два основных компонента – узлы и коммуникационную среду. Каждый процессор имеет непосредственный доступ только к своей локальной памяти, а доступ к данным, расположенным в памяти других процессоров, выполняется более сложными способами.

CRAY T3D подключается к хост-компьютеру, роль которого, в частности, может исполнять CRAY Y-MP C90. Вся предварительная обработка и подготовка программ проходит на хосте (напр., компиляция).

CRAY T3D имеет в своем составе:

- сеть межпроцессорного взаимодействия (или коммуникационную сеть),
- вычислительные узлы
- узлы ввода/вывода.

Вычислительные узлы состоят из двух ПЭ, сетевого интерфейса и контроллера блочных передач. Оба ПЭ, входящие в состав ВУ, идентичны и могут работать независимо друг от друга.

Процессорный элемент содержит микропроцессор, локальную память и некоторые вспомогательные схемы.

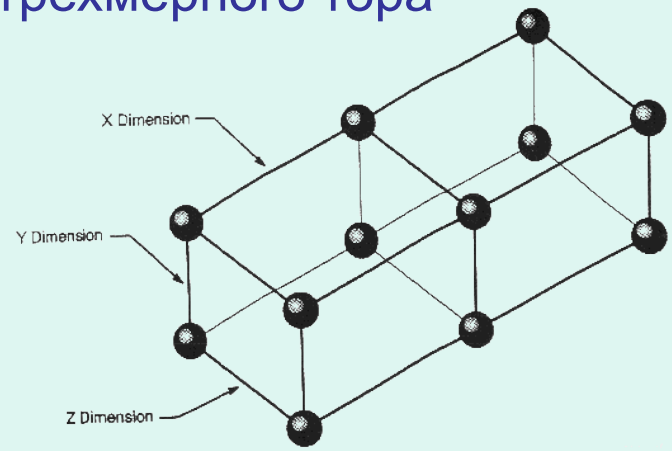
Микропроцессор в T3D – ALPHA21164 или 21164A DEC, работающий на тактовой частоте 150 MHz. В Cray T3E-1200 – ALPHA21164 с частотой 600 МГц. В Cray T3E-1350 ALPHA21164A с частотой 675 МГц.

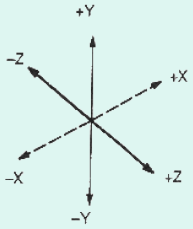
Архитектура CRAY T3D/T3E (2)

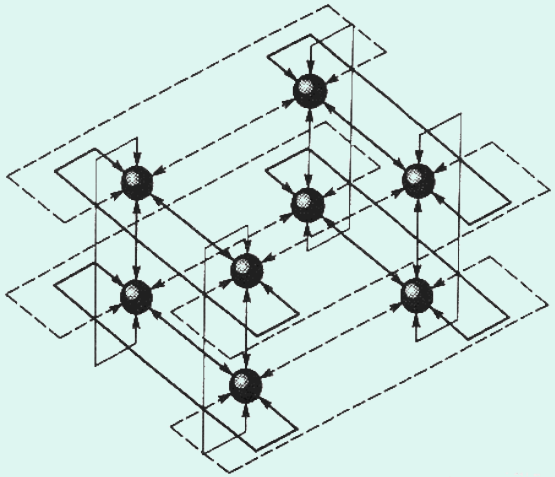
Локальная память каждого ПЭ является частью физически распределенной, но логически разделяемой (или общей), памяти всего компьютера. Память физически распределена, т.к. каждый ПЭ содержит свою локальную память. В то же время, память разделяется всеми ПЭ, т.к. каждый ПЭ может обращаться к памяти любого другого ПЭ, не прерывая его работы. Обращение к памяти других ПЭ в 6 раз медленнее, чем обращение к своей собственной локальной памяти.

Коммуникационная сеть обеспечивает передачу информации между вычислительными узлами и узлами ввода/вывода. Сеть образует трехмерную решетку, соединяя сетевые маршрутизаторы узлов в направлениях *X*, *Y*, *Z*. Каждая элементарная связь между двумя узлами – это два однонаправленных канала передачи данных, что допускает одновременный обмен данными в противоположных направлениях.

Коммуникационная сеть организована в виде двунаправленного трехмерного тора







Архитектура CRAY T3D/T3E (3)

Топология в виде двунаправленного трехмерного тора имеет следующие преимущества перед другими способами организации связи:

- быстрая связь граничных узлов и небольшое среднее число перемещений по тору при взаимодействии разных ПЭ: максимальное расстояние в сети для конфигурации из 128 ПЭ равно 6, а для 2048 ПЭ равно 12;
- возможен выбор другого маршрута для обхода поврежденных связей.
- длина максимального физического соединения между ПЭ минимизирована за счет чередования узлов в размерностях X и Z.

Маршрутизация в сети и сетевые маршрутизаторы.

- При выборе маршрута для обмена данными между двумя узлами сетевые маршрутизаторы всегда сначала выполняют смещение по размерности X, затем по Y, затем по Z. Смещение может быть как положительным, так и отрицательным, Этот механизм помогает минимизировать число перемещений по сети и обойти повреждения.
- Сетевые маршрутизаторы каждого ВУ определяют путь перемещения каждого пакета и могут осуществлять параллельный транзит данных по каждому из трех измерений X, Y, Z.

Нумерация вычислительных узлов.

Каждому ПЭ в системе присвоен уникальный физический номер, определяющий его физическое расположение, который и используется непосредственно аппаратурой.

Факторы, влияющие на производительность МР-систем

Пример поэлементного сложения двух векторов

```
for(i = 0;i<NDIM; i++)
    A[i] = B[i]+C[i];
```

Организация расчета в параллельном режиме на МП компьютере: каждый процессорный узел в системе из **N** процессоров выполняет поэлементное сложение для своего диапазона индексов массива A, т.е. по **NDIM/N** операций (если **N** кратно **NDIM**). Для этого необходимо

- разослать данные по процессорным узлам (каждому нужные элементы)
- выполнить сложение
- собрать результат в главный (root) процесс

ПОЛЬЗОВАТЕЛЬ САМ МЕНЯЕТ КОД!

Теоретическое ускорение параллельного кода по сравнению с последовательным составит **~N**, однако из-за накладных расходов (рассылка и сбор данных, синхронизация процессов и др.) реальное ускорение будет существенно меньше.

Факторы, влияющие на производительность

- Реальная производительность одного процессора и их количество
- Закон Амдала (особенности алгоритма)
- Скорость передачи информации между процессорами
- Взаимодействие процессоров – **латентность** (время инициализации посылки) и **скорость** (время самой передачи сообщения)
- Ожидание завершения взаимодействия процессоров
- Ожидание прихода сообщения
- Неравномерность загрузки (простой части процессоров)

Суперкомпьютеры CRAY XT3,XT4, XT5, XT6

Это представители MPP-систем с распределенной памятью. Отличаются в основном элементной базой и вариантами компоновки. Как и в предшествующих моделях Cray, базовый элемент – 1-процессорный узел (ПЭ), состоящий из одного микропроцессора AMD (1-,2- и более-ядерного), имеющего свою память и средства связи с другими ПЭ.

Cray XT3 комплектуется 1-ядерными процессорами AMD Opteron 2.4 ГГц.

В **Cray XT4** используются 2-ядерные процессоры AMD Opteron 2.6ГГ.

В моделях **XT5-HE** использованы 6-ядерные процессоры Opteron 2.6 GHz.

В **XT6-HE** – процессоры Opteron 12-core 2.1 GHz.

ВУ компонуются в стойки (до 96 ВУ на стойку). Возможны различные конфигурации по числу стоек и числу процессоров в стойке.

Максимальная конфигурация **Cray XT3\4** до 30502 ВУ (320 стоек по 96 ВУ).

Число ПЭ в моделях **XT5**-HE – 224162, в моделях **XT6**-HE – 30720.

ВУ объединены на основе топологии «трехмерный тор» с использованием высокоскоростных коммуникационных технологий.

Пиковая производительность Cray XT4 – 318TFlop/s (38е место в Top500),

Cray XT3: 147Tflop/s (50е место в 2011, модифицированная модель XT3\XT4 с 38208 ПЭ)

XT5-HE: 3е место в 2011, пиковая производительность 2331Тфлоп\с (499е в 2015)

XT6-HE: 49е место в TOP500 в 2011, пиковая производительность 258Тфлоп\с

Вычислительные кластеры (1)

Помимо традиционных MPP-суперкомпьютеров класс компьютеров с распределенной памятью активно расширяется за счет вычислительных кластеров.

Единственным способом взаимодействия процессоров в рамках MPP-систем (MPP — massive parallel processing) было их общение через некоторую коммуникационную среду, объединяющую процессоры в единую вычислительную установку. Однако все уникальные решения, да еще и с рекордными характеристиками обычно недешевы, поэтому и стоимость подобных систем никак не могла быть сравнима со стоимостью систем, находящихся в массовом производстве.

Прогресс в области сетевых технологий привел к появлению на рынке недорогих, но эффективных коммуникационных решений. Это, а также распространение серийных однокристальных микропроцессоров высокой мощности, предопределило появление кластерных вычислительных систем, фактически являющихся одним из направлений развития компьютеров с массовым параллелизмом

Кластер представляет собой два или более компьютеров (узлов), объединяемых при помощи сетевых технологий на базе шинной архитектуры или коммутатора и предстающих перед пользователем как единый информационно-вычислительный ресурс.

В качестве **узлов** кластера могут быть выбраны рабочие станции, серверы и даже обычные ПК. На каждом узле работает собственная копия операционной системы. При сбое одного узла другой может взять на себя нагрузку и пользователи не заметят прерывания в доступе.

Вычислительные кластеры (2)

Возможности масштабируемости (увеличение числа узлов) делают такие системы наиболее дешевыми, т.к. они собираются на базе стандартных процессоров, коммутаторов, дисков и т.д.

Возможно также создание кластеров с более высокими параметрами на основе эксклюзивных и более дорогих комплектующих.

- системы высокой надежности
- высокопроизводительные вычисления
- многопоточные системы

Согласно Aberdeen Group, кластер – многомашинный комплекс, который

- выглядит с точки зрения пользователя как единая система
- обеспечивает высокую надежность (готовность)
- имеет общую файловую систему
- обладает свойством роста производительности при добавлении ресурсов (масштабируемость)
- гибко перестраивается
- управляется/администрируется как единая система

Компания **DEC** первой анонсировала ***концепцию кластерной системы*** в 1983, определив ее как группу объединенных между собой ЭВМ, представляющих собой единый узел обработки информации.

VAX-кластер представлял собой слабосвязанную многомашинную систему с общей внешней памятью, обеспечивающую единый механизм управления и администрирования.

Вычислительные кластеры (3)

VAX-кластер обладал свойствами, определившими его эффективность и ставшими неотъемлемой частью современных кластерных систем:

Разделение ресурсов. Компьютеры VAX в кластере могут разделять доступ к внешней памяти. Все компьютеры VAX в кластере могут обращаться к отдельным файлам данных как к локальным.

Высокая готовность. Если происходит отказ одного из VAX-компьютеров, задания его пользователей автоматически могут быть перенесены на другой компьютер кластера.

Высокая пропускная способность. Ряд прикладных систем могут пользоваться возможностью параллельного выполнения заданий на нескольких компьютерах кластера.

Удобство обслуживания системы. Общие базы данных могут обслуживаться с единственного места. Прикладные программы могут инсталлироваться только однажды на общих дисках кластера и разделяться между всеми компьютерами кластера.

Расширяемость. Увеличение вычислительной мощности кластера достигается подключением к нему дополнительных VAX-компьютеров.

Эффективность VAX-кластера определялась двумя главными факторами:

- ✔ высокоскоростной механизм связи,
- ✔ системное программное обеспечение, которое обеспечивает клиентам прозрачный доступ к системному сервису. Физически связи внутри кластера реализуются с помощью трех различных шинных технологий с различными характеристиками производительности.

Вычислительные кластеры (4)

Один из известных кластерных проектов, также оказавших оромное влияние на развитие кластерных технологий – **beowulf** (1994)

В центре NASA 16 процессоров Intel 100МГц по 16Mb ОП на каждом, объединены с использованием обычной сети Ethernet (10 мбит\сек).

Проект стал основой общего подхода «a la Beowulf» к построению параллельных кластерных компьютеров, состоящих из одного серверного узла (головного) и нескольких подчиненных узлов, соединенных посредством стандартной компьютерной сети. Кластер строится на основе стандартных компонент (ПК под LINUX), стандартных сетевых адаптеров. Нет специального программного пакета **beowulf**. Есть несколько пакетов программного обеспечения, которые пользователи нашли пригодными в рамках кластеров данной концепции. LINUX, MPI, системы управления ресурсами и другие стандартные продукты.

Итак, кластеры строятся на основе стандартных процессорных узлов, стандартного сетевого оборудования, стандартных ОС и систем управления ресурсами.

Критерии выбора комплектующих: надежность, распространенность, предсказуемость производительности, проверка нагрузочными тестами.

Развитие кластеров идет

- по наращиванию числа процессоров,
- использованию все более мощных процессоров (Intel, Alpha, AMD),
- развитию коммуникационных технологий – Gigabit Ethernet, Myrinet et al

Вычислительные кластеры (5)

Умеренная цена оборачивается существенными накладными расходами на взаимодействие узлов.

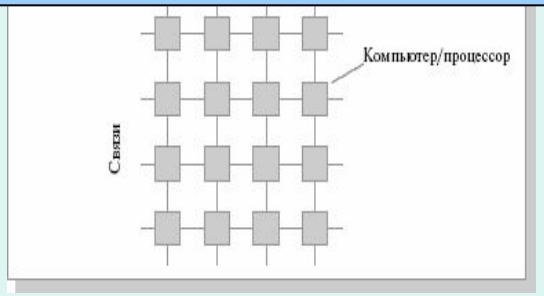
- ***Латентность*** – время начальной задержки при передаче сообщений;
- ***Пропускная способность сети*** – скорость передачи информации.

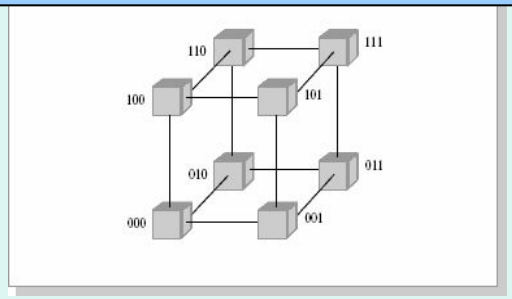
Много коротких сообщений – латентность играет главную роль.

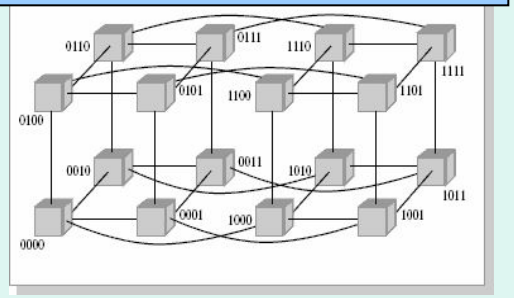
Длинные сообщения – основным фактором становится скорость.

Топология соединения процессоров друг с другом определяет производительность кластера в не меньшей степени, чем тип используемых в ней процессоров.

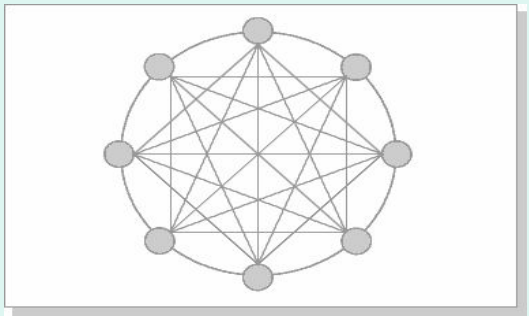
плоская решетка (матрица) куб (8) гиперкуб (> 8)

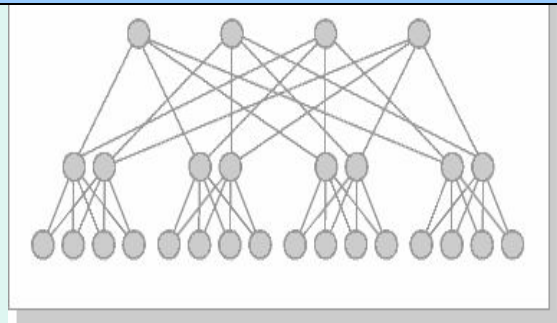


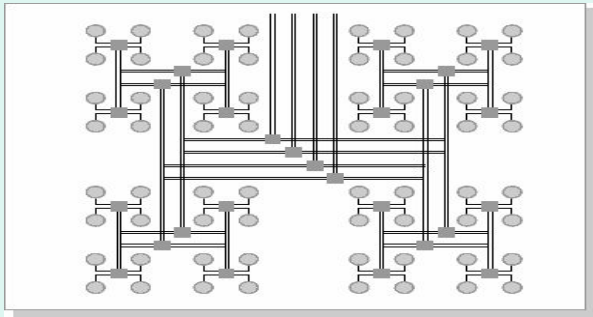




кольцо с полной связью по хордам фэт-три «толстое дерево»







Вычислительные кластеры (6)

- Итак – вычислительные кластеры, как и MPP-компьютеры, имеют вычислительные узлы (ВУ), взаимодействующие через коммуникационную среду. В качестве ВУ выступают ПК или рабочие станции, подключаемые к сети через обычные сетевые платы.
- На начальных этапах развития кластерных технологий взаимодействие между ВУ в кластерах происходило гораздо медленнее, чем в MP-компьютерах. Главными преимуществами кластеров были тогда простота и гибкость конструкции, а таже цена.
- С развитием коммуникационных технологий различие в скорости взаимодействия сглаживается, так что и в настоящее время кластеры вытесняют MP-компьютеры и занимают лидирующие позиции среди высокопроизводительных вычислительных систем.
- Централизованные кластеры – ВУ компактно смонтированы в единую структуру в пределах одной комнаты или зала.
- Децентрализованные кластеры – ВУ разбросаны в пределах здания или организации ⇒ идеи метакомпьютинга

Суперкомпьютер «Ломоносов»

Поставлен компанией «Т-Платформы» для НИВЦ МГУ в рамках реализации российско-белорусского проекта СКИФ по разработке семейства высокопроизводительных систем.

В июне 2011 занимал 13е место в TOP500. В 2014г - 42е, в 2017 – 63е (Lomonosov-2) и 227е (Lomonosov).

Первый гибридный суперкомпьютер такого масштаба в РФ и Восточной Европе. 3 вида вычислительных узлов и процессоры с различной архитектурой.

В качестве основных узлов, обеспечивающих >90% производительности системы, используются модули T-Blade2 на базе 4- и 6-ядерных микропроцессоров Intel Xeon 2.93 GHz + ускорители Nvidia 2070 GPU.

Сеть: Infiniband QDR.

Заявка на производительность 1.3Пфлопс

Отношение реальной и пиковой производительности (эффективность): **78%.**

В состав суперкомпьютерного центра НИВЦ МГУ входят также кластеры «Чебышев» и «Менделеев», также построенные в рамках проекта СКИФ.

Третья российская система в TOP500-2017: RSC Tornado, 412е место, Санкт-Петербургский политехнический университет

Другие российские системы в списках TOP500:

- Курчатовский институт,182е место (2011);
- Южно-Уральский университет, 185е место (2011) проект СКИФ-АВРОРА;
- RSC Tornado SUSU, 154е место (2014)

РФ в TOP500-2020:

36е место (Сбер); 131е (Ломоносов)

Fujitsu K computer, SPARC64 VIIIfx 2.0GHz – пример удачного соотношения реальной и пиковой производительности

- Самый быстрый кластер согласно Top500-2012.
- В 2014 – 4е, в 2016 – 5е, в 2017 – 10е
- Установлен в Японии, RIKEN Advanced Institute for computational science (AICS)
- Число ядер:705024.
- Реальная производительность на тесте LINPACK 10.51 petaflops (~10.51×10^{15} операций с плавающей точкой в секунду);
- пиковая производительность – 11.28 petaflops.

Удачное соотношение реальной и пиковой производительности – **93% (!!)**

Суперкомпьютеры IBM BlueGene (1)

Проект IBM, направленный на достижение производительности 1Pflop/s, стартовал в 1999. Предусмотрены этапы: Blue Gene/L, Blue Gene/P и Blue Gene/Q, 100 млн долларов вложений.

Цель разработчиков – создание не только самой быстрой машины, но и самой эффективной с точки зрения показателей «терафлоп\доллар», «терафлоп\куб.м», «терафлоп\ватт». Отказ от стратегии выбора самых быстродействующих компонент независимо от цены. Выпущен собственный однокристальный процессорный элемент, работающий с умеренной скоростью и низким энергопотреблением, чтобы на этой основе построить машину с компактным и эффективным расположением элементов.

BlueGene/L (2006) — 1е поколение: первый заказчик – Ливерморская лаборатория. В максимальной конфигурации 131072 процессора, пик. произв. 360Tflop/s. В Top500 2006 вошли 27 Blue Gene/L (включая 1е, 2е места).

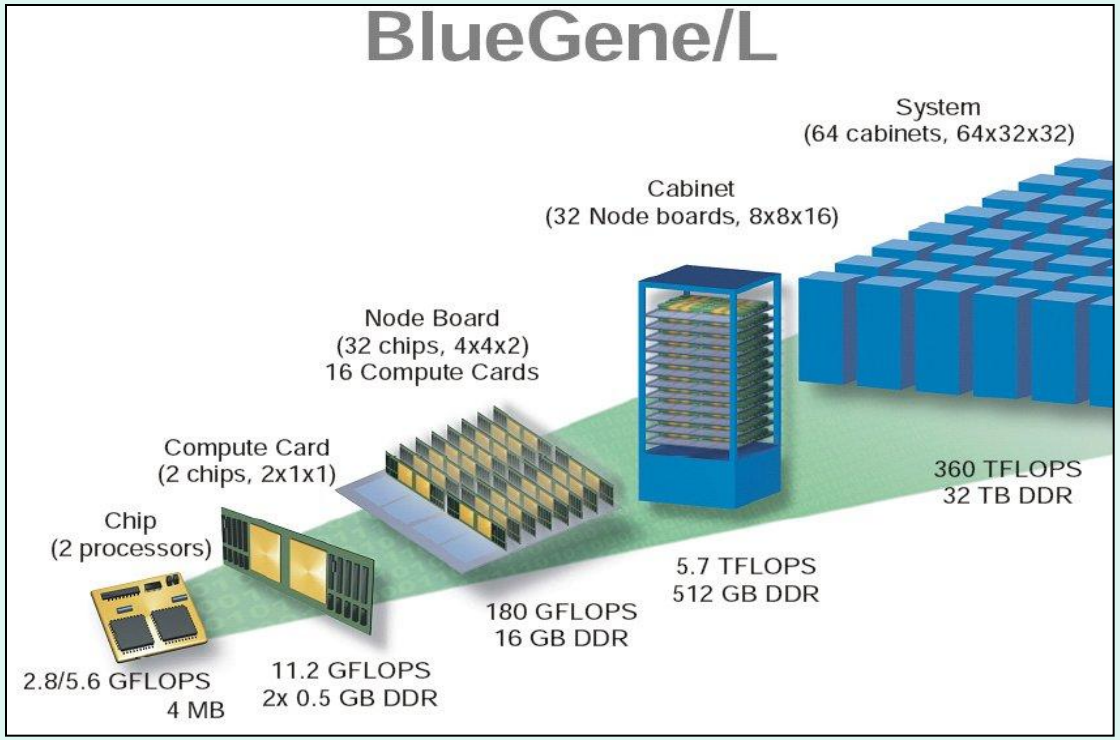
BlueGene/P (2007) — 2е поколение: 294912 процессора, достигнута пиковая производительность 1Pflop/s, 13е место Top500 в ноябре 2011.

BlueGene/Q (2011) — 3е поколение. Цель разработчиков – 20 Pflop/s в 2011-2012. Эволюционное продолжение /L, /P, работающее на более высокой частоте и потребляющее меньше энергии. Примерно то же число процессоров при большем количестве ядер. 1я система /Q – **Sequoia** в Ливерморской лаборатории (2011): 98304 ВУ, 1.6млн ядер. 1е место в 2012, 3е в 2014, 4е в 2016, 6е в 2017. **Mira** – 5е место в 2016, 11е в 2017. Всего в TOP500-2020 – 4 Blue Gene/Q (включая 16е место)

Суперкомпьютеры IBM BlueGene (2)

Особенности архитектуры:

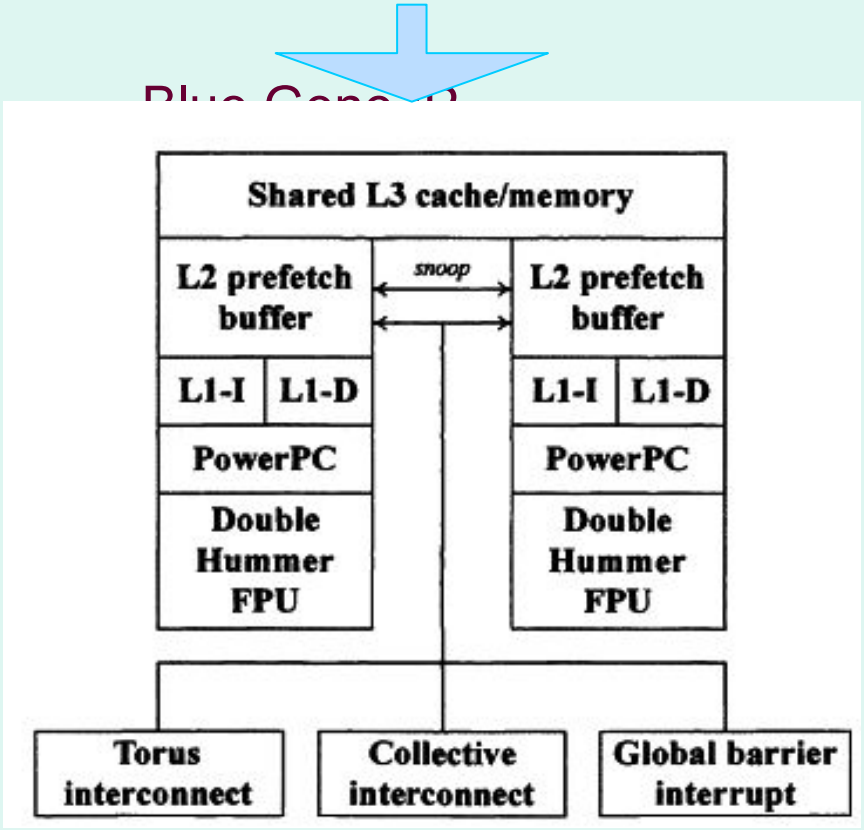
-MPP-компьютер;
-масштабируемая сотовая архитектура;
-сверхплотная компоновка;
-нужная конфигурация собирается из однотипных стоек, стойка занимает меньше 1кв.м, объединяет до 1024 ВУ
-масштабируемая и высокопроизводительная схема соединений
-Базовый элемент имеет 2,4,18 процессоров соответственно для /L, /P, /Q



Суперкомпьютеры IBM BlueGene (3)

SMP-узел Blue Gene /L:

- Два 770МГц процессора PowerPC.
- У каждого – кэши 1го и 2го уровней и специализированный процессор Double Hummer FPU для арифметики с плавающей точкой.
- Общий кэш 3го уровня + общий блок разделяемой памяти.
- Связь кэша L2 через порт с коллективной сетью



4-процессорная микросхема