Суперкомпьютерные системы

Судаков А.А.

“Параллельные и распределенные вычисления” Лекция 10

План

- История
- Типы суперкомпьютерных систем
- Векторно-конвейерные
- Массивно-параллельные
- NUMA – с неоднородным доступом к памяти
- Высокопроизводительные кластеры
- Метакомпьютеры

Литература

- Суперкомпьютерные системы http://www.top500.org/ORSC/2004/
- Транспьютеры http://maven.smith.edu/~thiebaut/transputer/descript.html
- Метакомпьютеры http://setiathome.ssl.berkeley.edu/
- GRID системы

http://www.grid.org

Исторические сведения

- Первые векторные, конвейерные и суперскалярные процессоры CDC 1960-года
- CRAY - CDC 1976
- В СССР – БЭСМ 6 1967 г
- Сейчас – SGI, HP, NEC
- Массовое распространение - кластеры на базе широкодоступных компонентов

Типы суперкомпьютерных систем

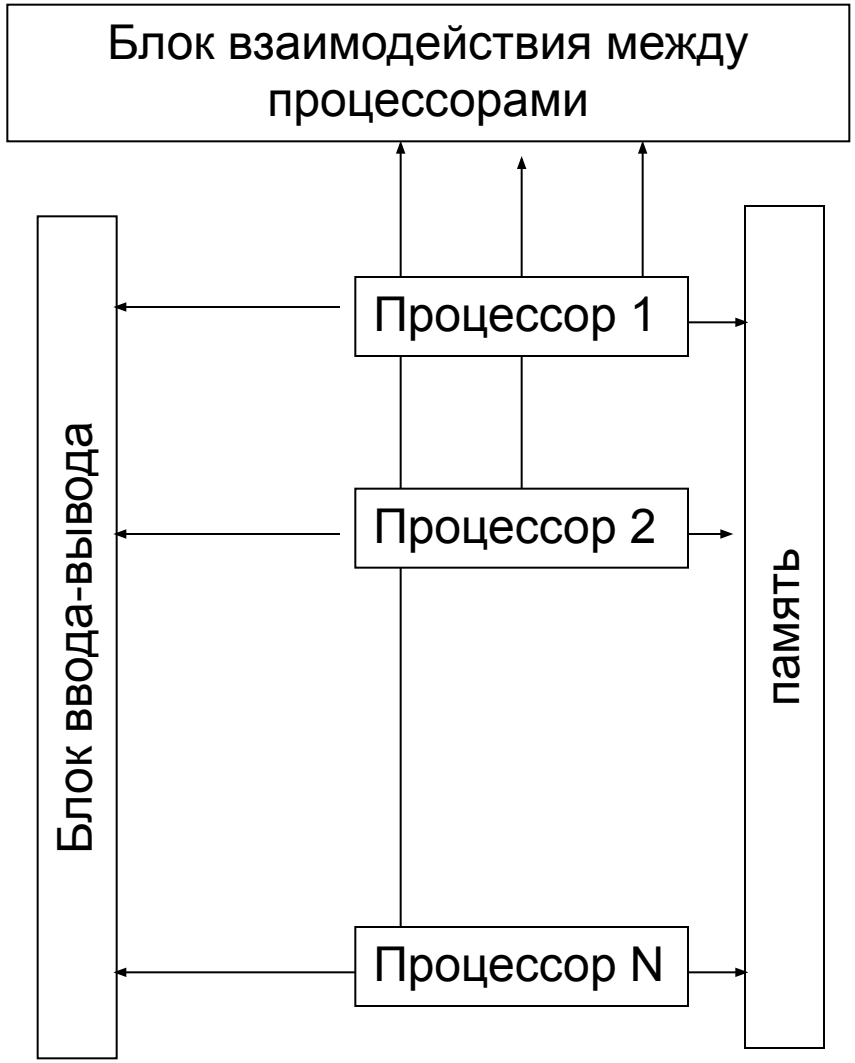
- Векторные (PVP, array, matrix, векторно-конвейерные vector-pipeline) компьютеры
- Массивно-параллельные суперкомпьютеры
- NUMA системы

Векторно-конвейерные суперкомпьютеры Cray Y-MP C90

- Конец 1980-х
- До 16 процессоров с тактовой частотой до 250 МГц
- Памяти до 1 ТВ
- ОС UNICOS
- Производительность 1 процессора 1 GFlops

Структура

- Векторно-конвейерные процессоры подключены к общей памяти как SMP
- Отсутствует кэш
- Каждый процессор может взаимодействовать с
  - памятью
  - устройствами ввода-вывода
  - С другими процессорами



Память

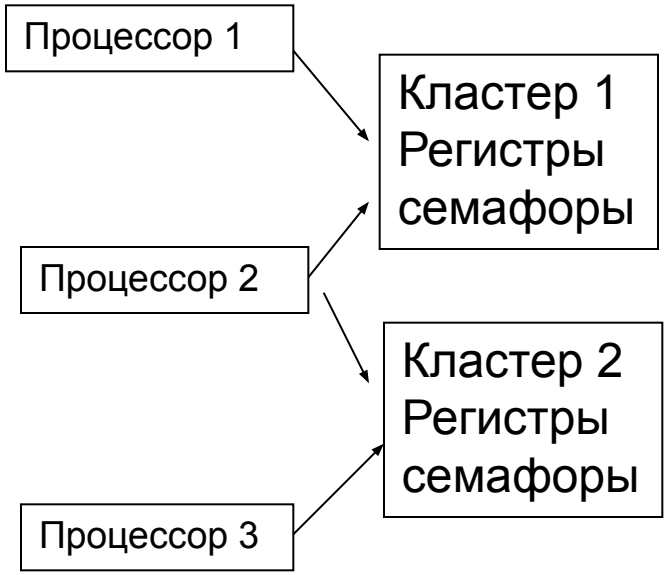
- До 1024 банков
  - К разным банкам можно обращаться одновременно
  - При обращении к одном банку задержка до 6 тактов
- 16 банков – одна подсекция
- 8 подсекций – 1 секция
- Процессоры обращаются к памяти через 4 порта ввода-вывода
  - 1 порт всегда на запись
  - 1 порт всегда секция ввода-вывода
  - Остальные по ситуации

Секция ввода-вывода

- Для связи с внешними устройствами и обмена информацией
  - Low-speed (LOSP) channels - 6 Mbytes/s
  - High-speed (HISP) channels - 200 Mbytes/s
  - Very high-speed (VHISP) channels - 1800 Mbytes/s

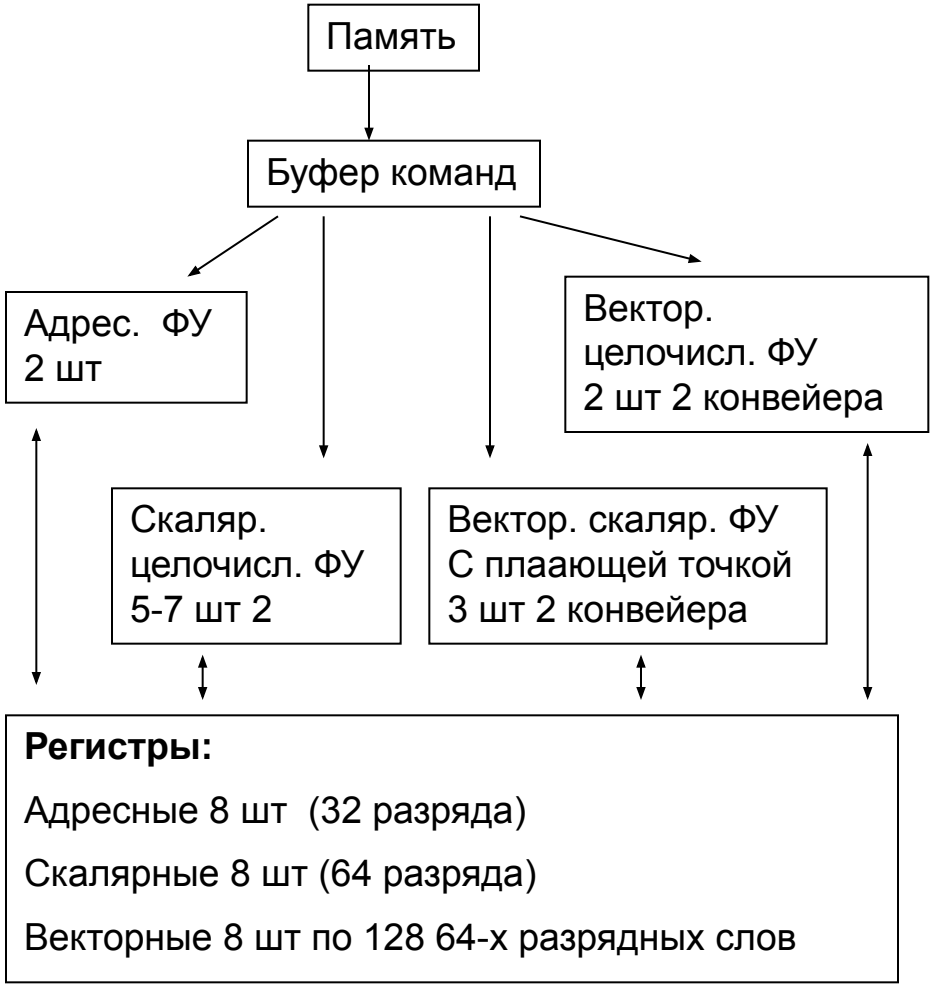
Блок взаимодействия между процессорами

- Для быстрой передачи данных между процессорами (задержка 1 такт)
- Несколько коммуникационных кластеров
- Каждый кластер содержит
  - Информационые регистры
  - Битовые семафоры
- Каждый процессор может обращаться в каждый момент времени только к одному кластеру
- К одному кластеру могут обращаться несколько процессоров
- Используются для «зацепления» процессоров
  - Данные с одного процессора конвейером могут передаваться на другой



Векторно-конвейерный процессор

- Команды считываются блоками
- Все операции являются конвейерными
- Все функциональные устройства могут работать параллельно
- Векторные операции могут выполняться двумя параллельными конвейерами
- Максимум – 4 операции за такт



Особенности использования PVP

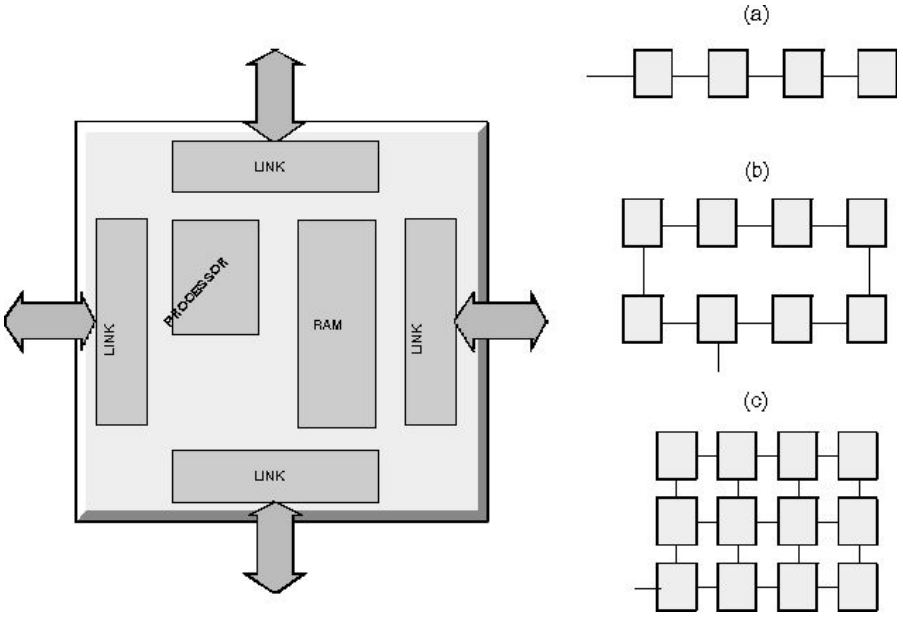
- Эффективны для выполнения большого количества однотипных (векторных, матричных) вычислений
- Не эффективны, если операции не векторизуются
- При большом количестве процессоров становятся очень дорогими и эффективность снижается по причине обращений к общей памяти
- Существуют специальные распараллеливающие/векторизующие компиляторы и распараллеленные библиотеки
  - MPI
  - SHM

Транспьютеры

- Микропроцессоры, специально разработанные для параллельных вычислений
- 1980-е года компания INMOS
- Основная идея - возможность непосредственного соединения процессоров

Особенности

- Транспьютер:
  - Процессор
  - Память
  - Соединения
- До 4-х соединений с соседними процессорами



Массивно-параллельные компьютеры

- Набор блоков с общей памятью (UP, SMP, PVP) соединенных с помощью коммуникационной подсистемы
- Массивно-параллельные компьютеры – системы с распределенной памятью
  - Каждый блок (узел) может обращаться только к своей локальной памяти
  - Данные из памяти других блоков могут передаваться только по сети
- Обычно существует один или несколько центральных блоков и большое количество рабочих узлов
- Важный элемент - коммутатор

Особенности систем

- Масштабируемость
  - Легко расширяются установкой новых блоков
  - Может быть большое количество процессоров (несколько тысяч в отличие от ранее рассмотренных систем)
- Операционная система
  - С одной копией ОС – устанавливается только на центральный узел
  - С распределенными копиями ОС – устанавливаются отдельно на каждую машину
- Библиотеки
  - MPI
  - PVM

Особенности использования

- Передача данных между блоками требует значительно большего времени, чем внутри блока
- Передача данных между блоками может выполняться параллельно с обработкой данных
  - В общей памяти – только конвейерно
- Увеличение количества процессоров не приводит к уменьшению эффективности за счет обращения к общим ресурсам
  - Эффективность уменьшается за счет возрастания времени передачи данных

Реализации

- CRAY T3E
- IBM SP2
- Классические большие суперкомпьютеры

IBM SP2

- Процессоры Power 2
- Узел Узлы IBM RS/6000
  - До 16 CPU
- До 16 узлов + коммутатор
- Коммутатор
  - Набор плат
  - Каждая плата 4 внешних выхода + 4 внутренних для связи с 4 другими платами
  - для сложных систем – промежуточные коммутаторы
  - Скорость обмена 300 МБайт/с

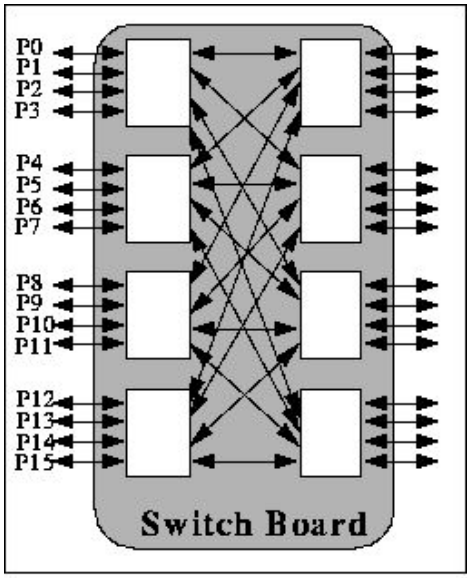
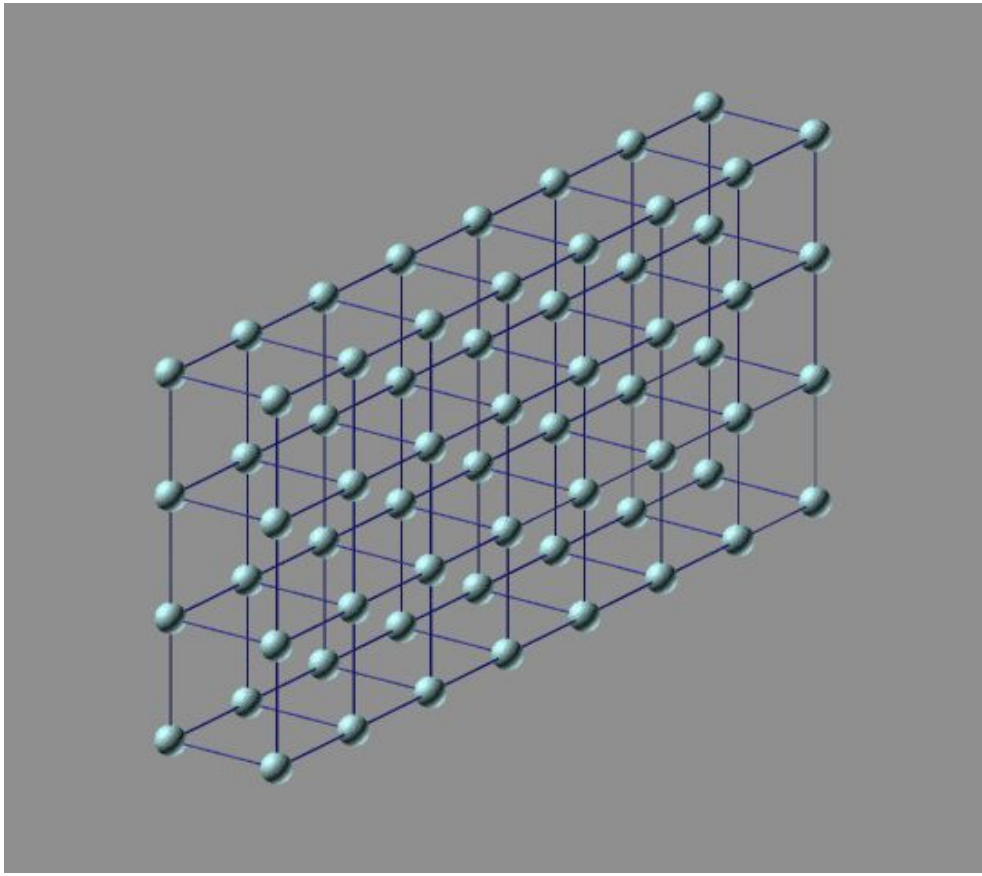


Figure 3. Switch Board Connections (16-way)

CRAY T3D

- топология 3D тор
- До 512 векторно-конвейерных процессорных блоков



NUMA системы

- NUMA – неоднородный доступ к памяти
- Набор SMP плат, связанных коммутатором
  - Доступ процессоров к «своей» памяти выполняется быстро
  - Доступ процессоров к «чужой» памяти выполняется в несколько раз медленнее (NUMA factor)
  - Вся память составляет одно общее адресное пространство

Особенности

- Те же, что у SMP
  - Когерентность кэшей
    - ccNUMA –аппаратное обеспечение когерентности
    - Программное обеспечение когерентности
- Обеспечение эффективного использования памяти
  - Алгоритмы консистентности памяти
- Операционная система
  - Одна копия ОС для всей системы (как SMP)
- Модель программирования – общая память
- Размеры системы ограничены размером адресного пространства
- Все современные суперкомпьютеры с общей памятью строятся по такой схеме

SGI Altix 3000

- Несколько блоков
  - Связь NUMALink 3 – 3.2 Гбайт/с
  - В сумме до 512 процессоров
- 1 Блок
  - 2 узла
  - Связь NUMALink 4 – 6.4 Гбайт/с
- 1 узел 2 процессора Itanium2
- ОС Linux

Кластеры

- Кластер – набор вычислительных систем, которые могут работать независимо, связаны между собой и используются как одна логическая система
- Все машины кластера работают как один большой компьютер для решения некоторых задач

Особенности кластеров

- Все узлы кластера являются вычислительными системами, которые выполняют свою копию ядра операционной системы
  - Кластер SMP систем
  - Кластер NUMA систем
- Кластер является одной системой лишь в контексте тех задач, для которых он предназначен, при рассмотрении с других точек зрения машины кластера могут оказаться несвязанными
- Все современные суперкомпьютеры являются кластерами
- Кластеры легко строить на базе широкодоступных компонент
  - Компьютеры общего назначения
  - Средства коммуникации общего назначения
- Основные функции в кластерных системах выполняет программное обеспечение

Использование кластеров

- High Performance Clusters, HPC
- high availability cluster, HAC
- load balancing cluster, virtual server
- Storage cluster, storage area network
- database cluster
- management clusters

Вопросы стоимости

- Закон Гроша (Grosch)
  - *стоимость суперкомпьютера пропорциональна квадрату его производительности*
  - Для микропроцессорных систем перестал действовать, но стоимость суперкомпьютеров очень высока (больше сотен тысяч долларов)
- Стоимость кластера = сумме стоимостей компонент и достаточно низка
- Кластер – дешевый вариант MPP компьютера
- Кластер хорошо использовать для обеспечения надежности за счет избыточности

Другие классификации кластеров

- Гомогенный
  - Все машины кластера одинаковы (в определенном контексте)
- Гетерогенный
  - Машины кластера – различны
- С одной копией операционной системы
  - Все ресурсы всех машин кластера видятся как ресурсы общей операционной системы
  - Для программ пользователя создается полная иллюзия того, что они работают на одно большой системе
- С распределенными копиями операционной системы
  - Каждый узел выполняет свою копию операционной системы, которая обслуживает ресурсы только своего узла

Исторические сведения

- Мультикомпьютерные системы
  - Конец 1970-х годов
- Первый промышленный кластер
  - 1983 г *VAX кластер*, DEC
- Промышленные кластеры
  - SUN, HP, IBM
- Массовые высокопроизводительные кластеры
  - 1996 г проект Beowulf

Метакомпьютеры

- Метакомпьютеры – использование существующих (простаивающих) компьютерных ресурсов для решения задач
  - Компьютерный класс
  - Компьютеры в пределах Интернет

Использование мощности существующих компьютеров

- В ночное время компьютеры часто простаивают
  - Потенциальная мощность простаивающих компьютеров может быть очень большой
  - Очень дешевые ресурсы
  - Возможность обеспечить избыточность ресурсов
- Недостатки
  - Надежность каналов передачи небольшая
  - Из-за малой скорости передачи данных в пределах Интернет невозможно выполнять параллельные вычисления

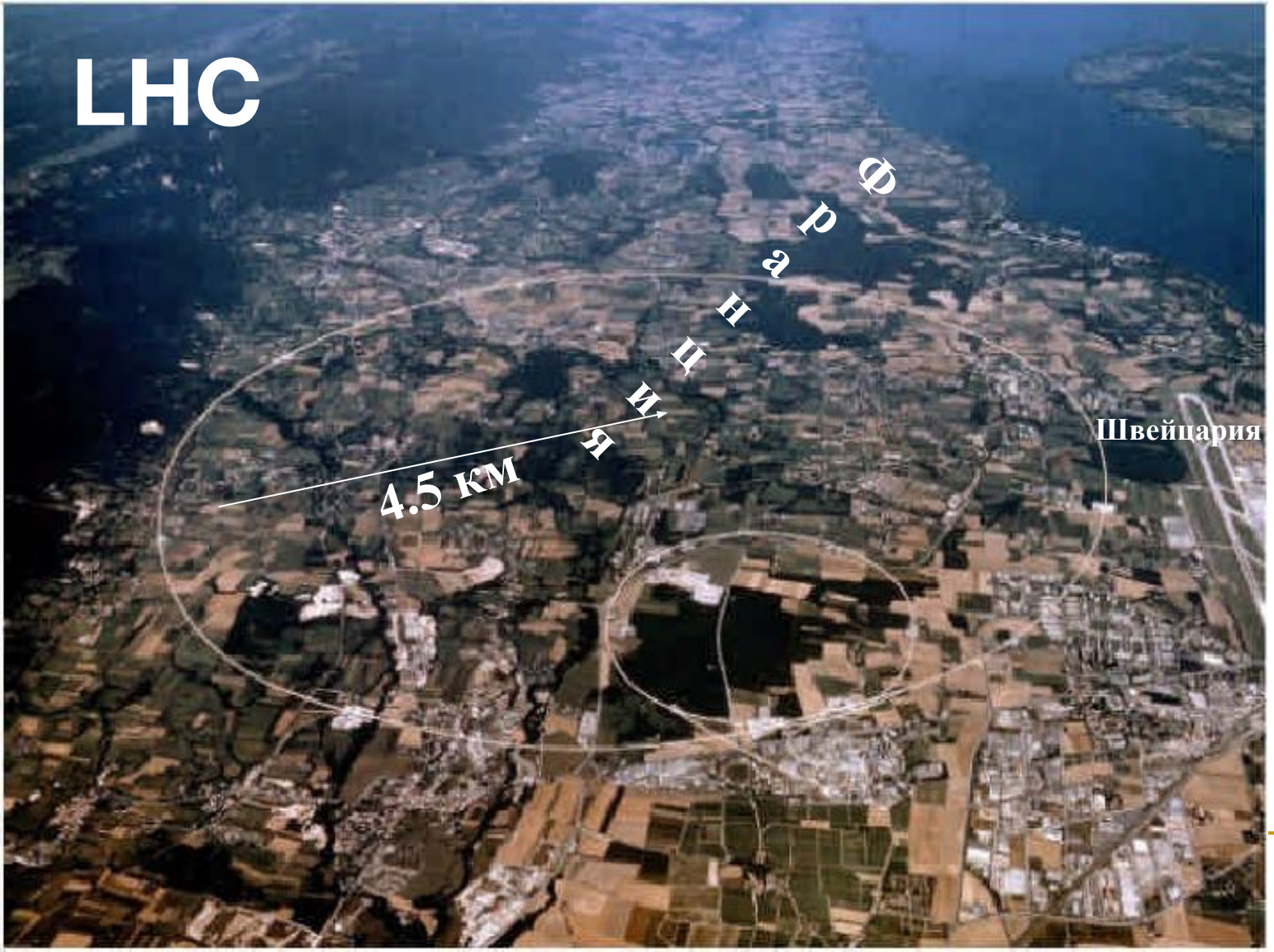
Существующие проекты

- Обработка данных с радиотелескопов по поиску внеземных цивилизаций
  - http://setiathome.ssl.berkeley.edu/
- Взломы алгоритмов шифрования
  - http://www.distributed.net/

GRID системы

- GRID – сеть
  - Метакомпьютеры промышленного уровня
  - Объединение компьютерных ресурсов в одну систему через Интернет
    - Кластеры
    - Базы данных
    - Средства хранения информации
    - Установки, которые выдают информацию
      - Медицинское оборудование
      - Телескопы
      - Микроскопы
      - Детекторы
  - По аналогии с едиными энергосистемами
    - Если одна электростанция перегружена, а другая нет, то часть клиентов переключаются на свободную электростанцию

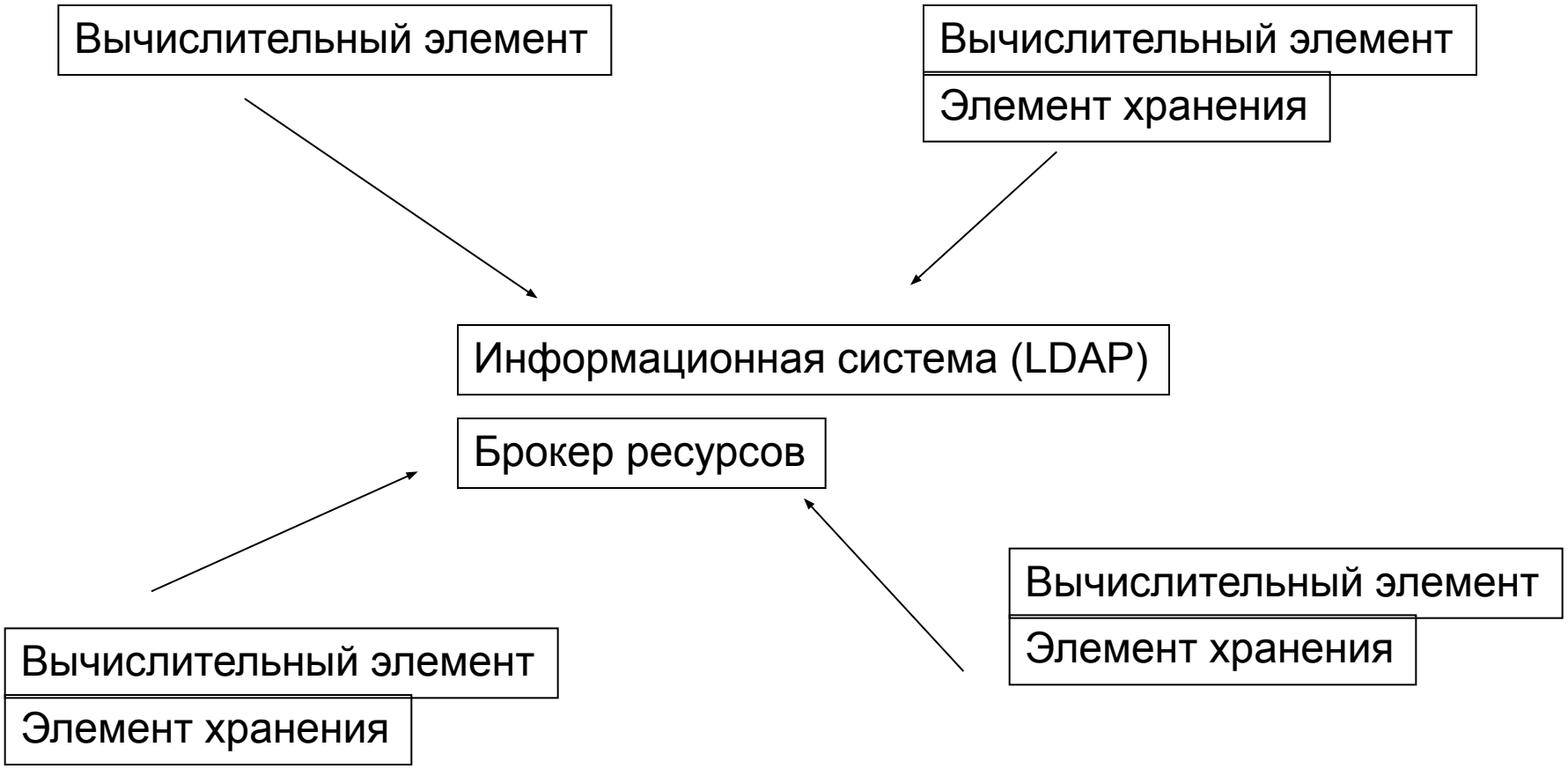
Lagre Hadron Colider - ускоритель



Объемы вычислений

- **Объём LHC данных соответствует**
- **примерно 20 миллионам CD ежегодно!**
- **требует компьютерной мощности эквивалентной ~ 100 000 персональных компьютеров с современными быстрыми процессорами!**

Структурная схема



Реализации

- Инструментарий
  - Globus
  - Condor
- Грид системы
  - EDG
  - ALIEN
  - LCG

Средства коммуникации для кластерных систем

- Кластер в основном управляется программно
- Средства коммуникации – наиболее критичная аппаратная часть для кластеров

Технологии коммуникации

- Ethernet
  - Fast Ethernet
  - Gigabit Ethernet
  - 10Gigabit Ethernet
- Myrinet
- SCI
- Infiniband
- QSNet
- cLan
- GigaNet

Характеристики средств коммуникации

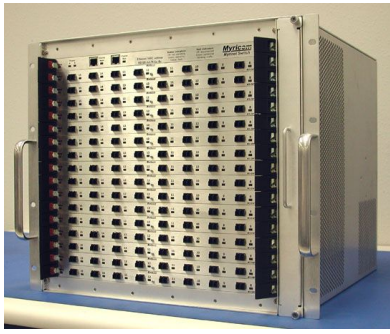
- Скорость передачи данных
  - Сколько времени занимает передача единицы информации
- Латентность (начальная задержка)
  - Сколько времени приходит от начала передачи данных до прихода информации на приемник
- Топология
  - Структура соединений между узлами
- Тип передачи
  - Коммутация каналов
  - Коммутация пакетов

Ethernet

- Среда
  - Медь
  - оптика
- Скорость передачи
  - Fast 100 Mbit/c (125 Мбайт/с)
  - Gigabit 1000 Мбит/с
- Латентность
  - Fast 125 мкс
  - Gigabit 33 мкс
- Топология
  - Звезда
- Тип коммутации
  - Передача пакетов
  - Каждый два узла параллельно
  - Широковещательная передача
- MTU
  - Fast 1500 байт
  - Gigabit 9000 байт
  - Минимальное 64 байт

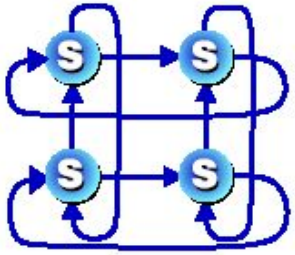
Myrinet (www.myri.com)

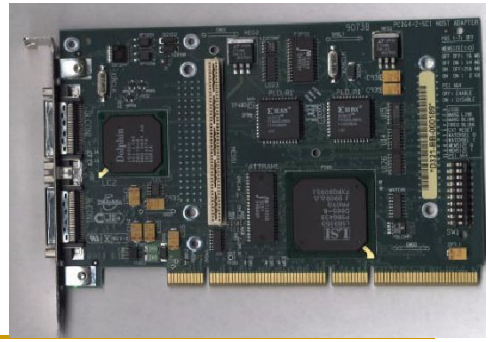
- Среда
  - Медь
  - Оптика
- Скорость передачи
  - до 10 Gbit/c
- Латентность
  - От 5 мкс
- Топология
  - Звезда
  - Гипердерево
- Тип коммутации
  - Передача пакетов



SCI (dolphinics.com)

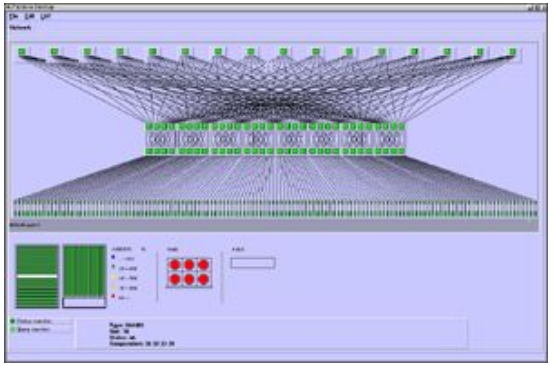
- Среда
  - Медь (шлейф)
- Скорость передачи
  - До 8008 Мбайт/с
- Задержка
  - 1.2 мкс !
- Топология
  - Линейка
  - Тор
  - Звезда
- Тип коммутации
  - Передача пакетов
- Общая память, перехват шины





QSNet (quadrix.com)

- Среда
  - Медь (параллельная шина)
- Скорость
  - До 1064 MBytes/sec в одном направлении
- Латентность
  - около 3 мкс
- Топология
  - Гипердерево
- Тип коммутациии
  - Коммутация пакетов
- Общая память



Другие технологии

- Infiniband
  - 1 GB/sec (7 мкс)
- Memory channel
  - 100 MB/sec, латентность - 3 мкс.