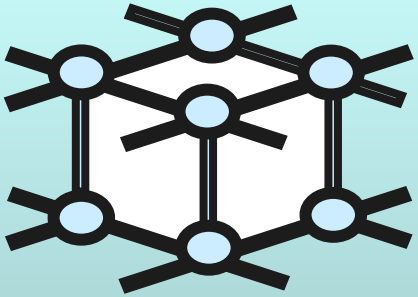


Сибирский Суперкомпьютерный Центр
ИВМ и МГ СО РАН

АРХИТЕКТУРА ТЕХНИЧЕСКИХ СРЕДСТВ НРС

Глинский Б.М.
Кучин Н.В.

ОБЛАСТИ ПРИМЕНЕНИЯ НРС

Проектирование инженерных сооружений, автомобилей, судов и летательных аппаратов, комплексный экологический мониторинг атмосферы и гидросферы, предсказание погоды, астрофизика и космические исследования, нанотехнологии, молекулярные науки, генетика, медицина, разработка лекарственных препаратов, рациональное использование лесных и минеральных ресурсов, разведка нефтегазовых месторождений, управляемого термоядерного синтеза, геоинформационных систем, систем распознавания и синтеза речи, систем распознавания изображений и другие направления деятельности человека просто немыслимы в настоящее время без применения компьютерного моделирования с использованием высокопроизводительных вычислений и параллельных компьютерных технологий.

СОВРЕМЕННЫЕ СУПЕРКОМПЬЮТЕРЫ

Три группы компьютеров для HPC: векторные; высокопроизводительные универсальные; специализированные

Векторные: создатель Сеймур Крей, SX-9 (NEC) имеет 8192 процессора и 840 TFLOPS векторной производительности (970 TFLOPS— с одновременной работы скалярного блока с плавающей запятой). Около 1 PFLOPS! Обмен с памятью 2Тбайт/с на 4 процессора.

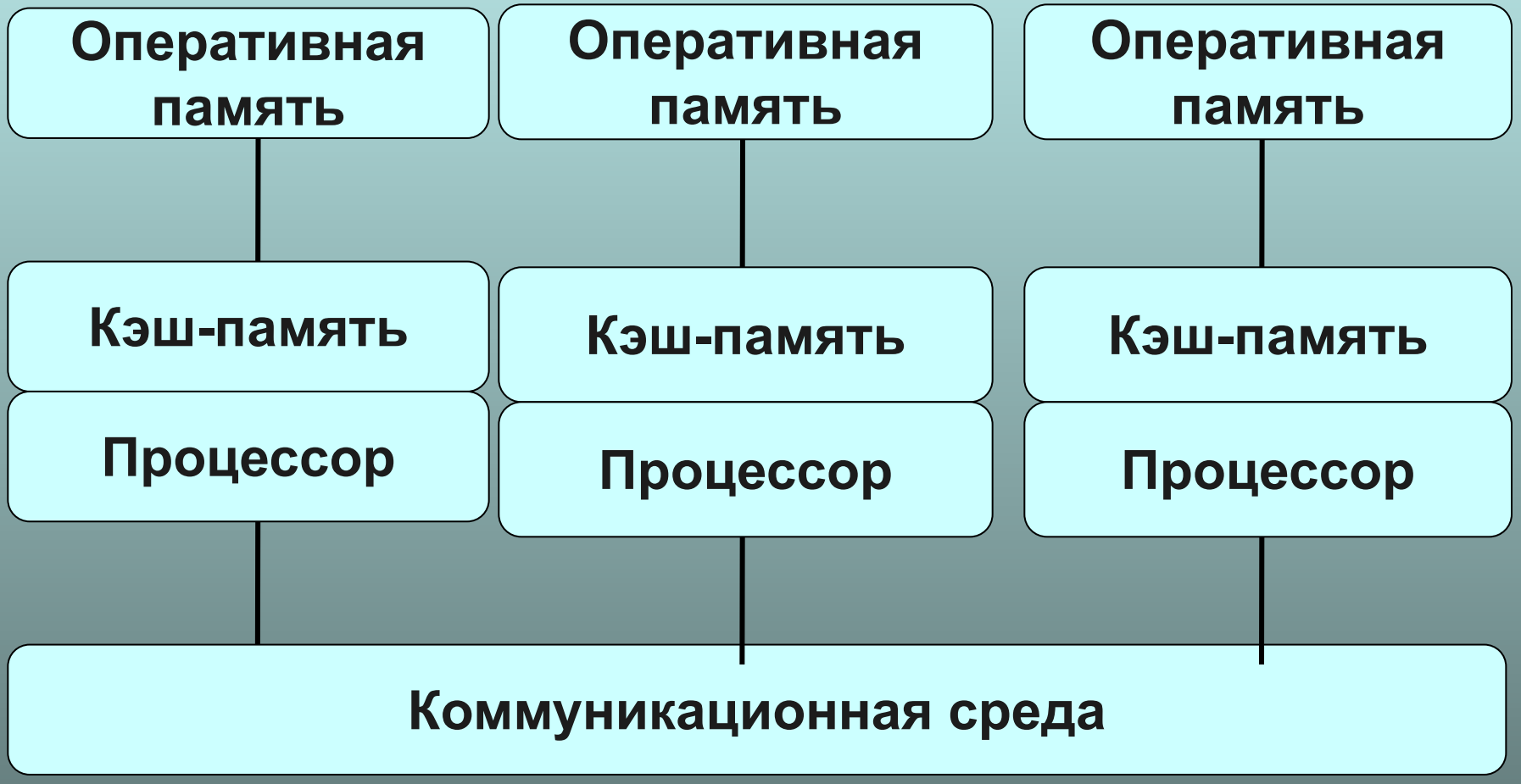
Универсальные: лидер RISC-процессор IBM Power6 с тактовой частотой 4,7 ГГц, 20 GFLOPS, 64 разряда (65 nm). Повышение производительности – многоядерность, копируются ядра на кристалле.

Специализированные: 32 разряд. программируемые графические процессоры (Graphical Processor Unit, GPU), процессоры Cell/B.E разработки IBM, Sony и Toshiba , ускорители вычислений с плавающей запятой типа ClearSpeed и др. Некоторые из этих средств реализованы в форме отдельных плат— «акселераторов» вычислений, другие (например, Cell) интегрируют в одной микросхеме и «спецсредства», и универсальные процессоры.

ОСНОВНЫЕ КЛАССЫ ПАРАЛЛЕЛЬНЫХ КОМПЬЮТЕРОВ

Основным параметром классификации паралелльных компьютеров является наличие общей (SMPОсновным параметром классификации паралелльных компьютеров является наличие общей (SMP) или распределенной памяти (MPPОсновным параметром классификации паралелльных компьютеров является наличие общей (SMP) или распределенной памяти (MPP). Нечто среднее между SMP и MPP представляют собой NUMAОсновным параметром классификации паралелльных компьютеров является наличие общей (SMP) или распределенной памяти (MPP). Нечто среднее между SMP и MPP представляют собой NUMA-архитектуры, где память физически распределена, но логически общедоступна. КластерныеОсновным параметром классификации паралелльных компьютеров является наличие общей (SMP) или распределенной памяти (MPP). Нечто среднее между SMP и MPP представляют собой NUMA-архитектуры, где память физически распределена, но логически общедоступна. Кластерные системы являются более дешевым вариантом MPP. При поддержке

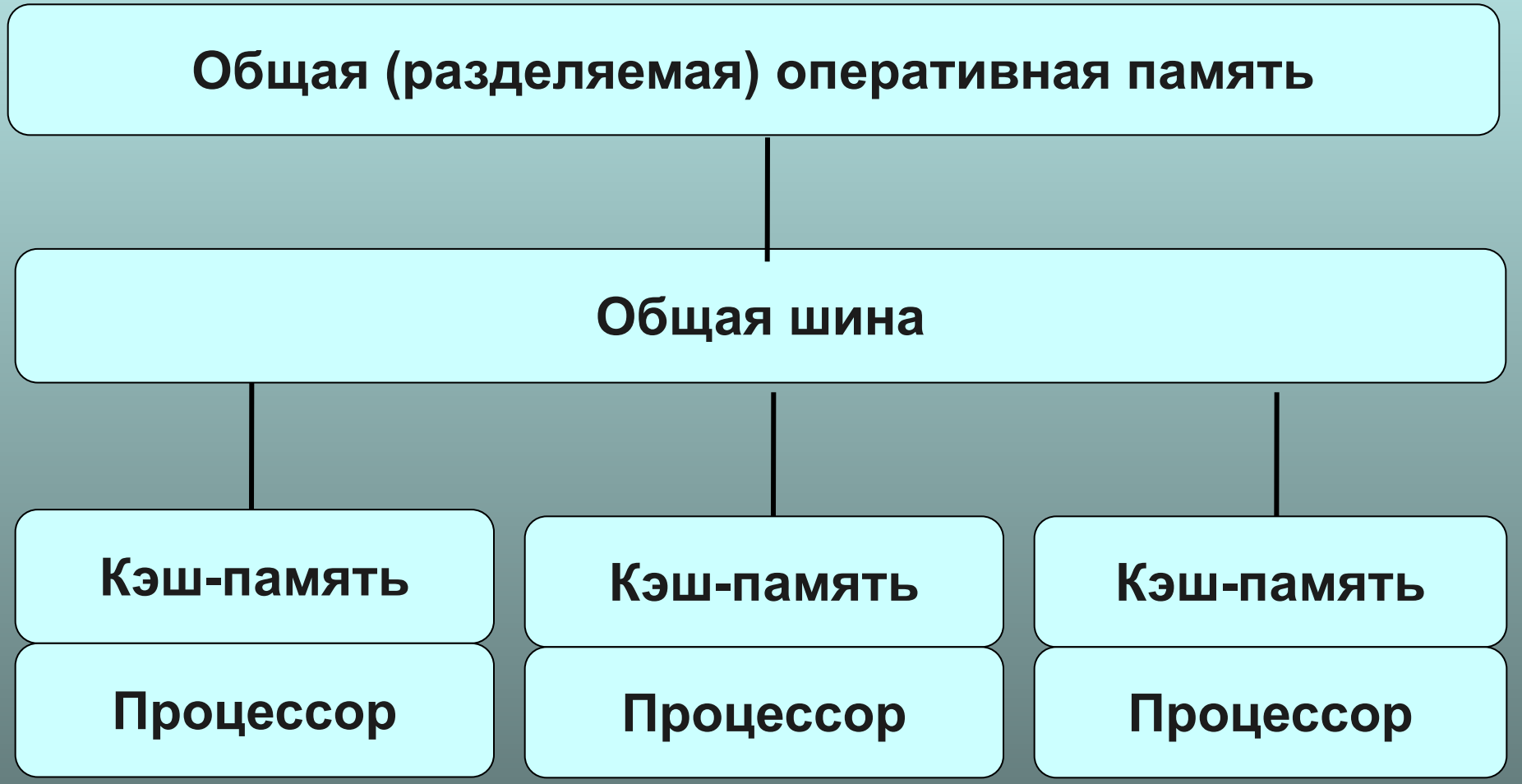
Системы с распределенной памятью (MPP)



Массивно-параллельные системы MPP

Архитектура	Система состоит из однородных *вычислительных узлов*, включающих: . один или несколько центральных процессоров (обычно RISC), . локальную память (прямой доступ к памяти других узлов невозможен), . коммуникационный процессор или сетевой адаптер . иногда - жесткие диски (как в SP) и/или другие устройства В/В К системе могут быть добавлены специальные узлы ввода-вывода и управляющие узлы. Узлы связаны через некоторую коммуникационную среду (высокоскоростная сеть, коммутатор и т.п.)
Примеры	IBM RS/6000 SP2, Intel PARAGON/ASCI Red, CRAY T3E, Intel PARAGON/ASCI Red, CRAY T3E, Hitachi SR8000, Intel PARAGON/ASCI Red, CRAY T3E, Hitachi SR8000, транспьютерные системы Parsytec.
Масштабируемость	Общее число процессоров в реальных системах достигает нескольких тысяч (ASCI Red, Blue Mountain).
Операционная система	Существуют два основных варианта: 1. Полноценная ОС работает только на управляющей машине (front-end), на каждом узле работает сильно урезанный вариант ОС, обеспечивающие только работу расположенной в нем ветви параллельного приложения. Пример: Cray T3E. 2. На каждом узле работает полноценная UNIX-подобная ОС (вариант, близкий к кластерному подходу). Пример: IBM RS/6000 SP + ОС AIX, устанавливаемая отдельно на каждом узле.
Модель программирования	Программирование в рамках модели **передачи сообщений** (MPI (MPI, PVM (MPI, PVM, BSPlib)

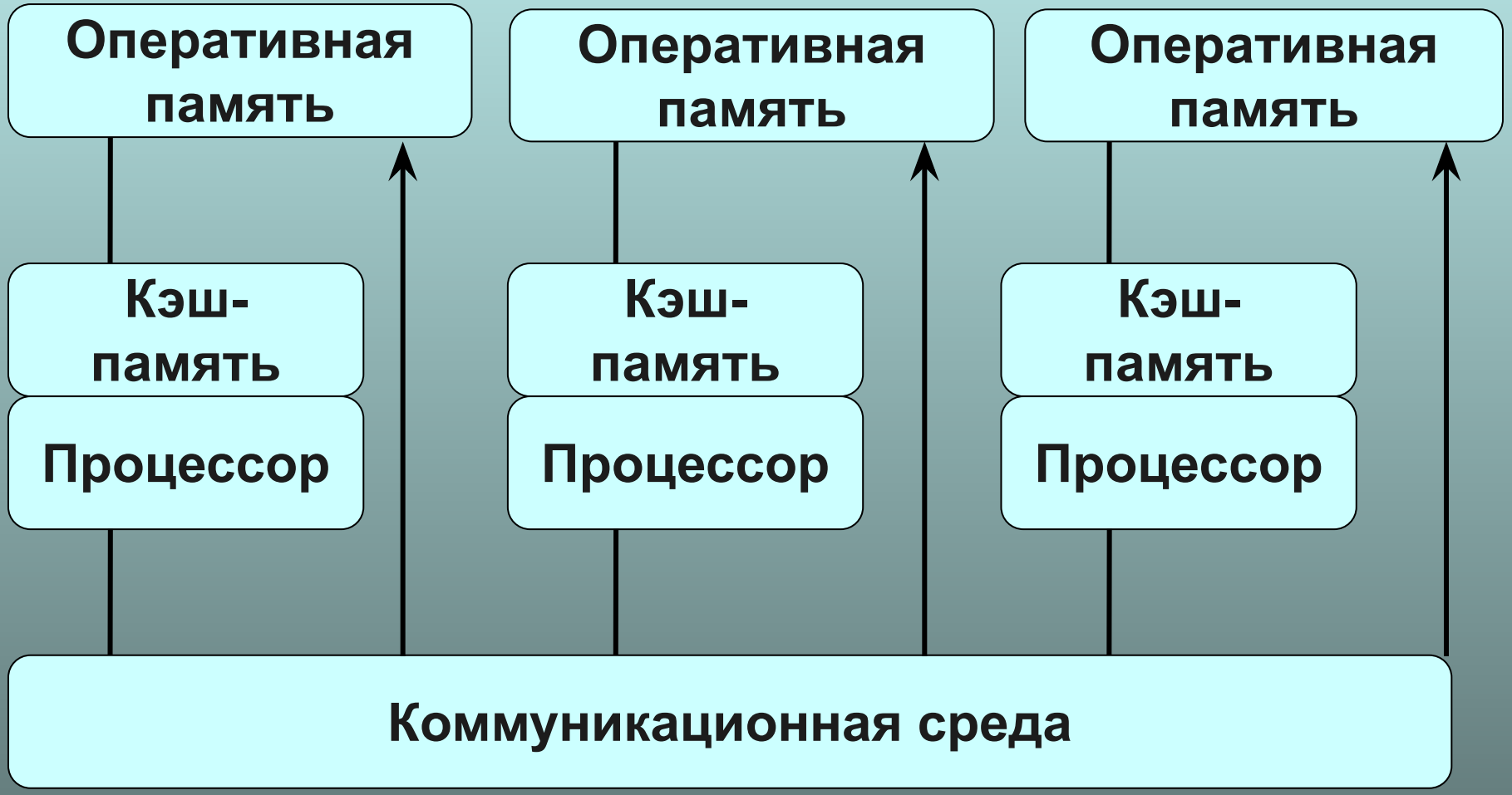
Системы с общей памятью (SMP)



Симметричные мультипроцессорные системы (SMP)

Архитектура	Система состоит из нескольких однородных процессоров и массива общей памяти (обычно из нескольких независимых блоков). Все процессоры имеют доступ к любой точке памяти с одинаковой скоростью. Процессоры подключены к памяти либо с помощью общей шины (базовые 2-4 процессорные SMP-сервера), либо с помощью crossbar-коммутатора (HP 9000). Аппаратно поддерживается когерентность кэшей.
Примеры	HP 9000 V-class, N-class; SMP-сервера и рабочие станции на базе процессоров Intel, IBM, HP, Compaq, Dell, ALR, Unisys, DG, Fujitsu и др.
Масштабируемость	Наличие общей памяти сильно упрощает взаимодействие процессоров между собой, однако накладывает сильные ограничения на их число - не более 32 в реальных системах. Для построения масштабируемых систем на базе SMP используются кластерныеНаличие общей памяти сильно упрощает взаимодействие процессоров между собой, однако накладывает сильные ограничения на их число - не более 32 в реальных системах. Для построения масштабируемых систем на базе SMP используются кластерные или NUMA-архитектуры.
Операционная система	Вся система работает под управлением единой ОС (обычно UNIX-подобной, но для Intel-платформ поддерживается Windows NT). ОС автоматически (в процессе работы) распределяет процессы/нити по процессорам (scheduling), но иногда возможна и явная привязка
Модель	Программирование в модели **общей памяти**. (POSIX threads, OpenMP. (POSIX threads, OpenMP). Для SMP-систем существуют сравнительно

Системы с NUMA-архитектурой



Системы с неоднородным доступом к памяти (NUMA)

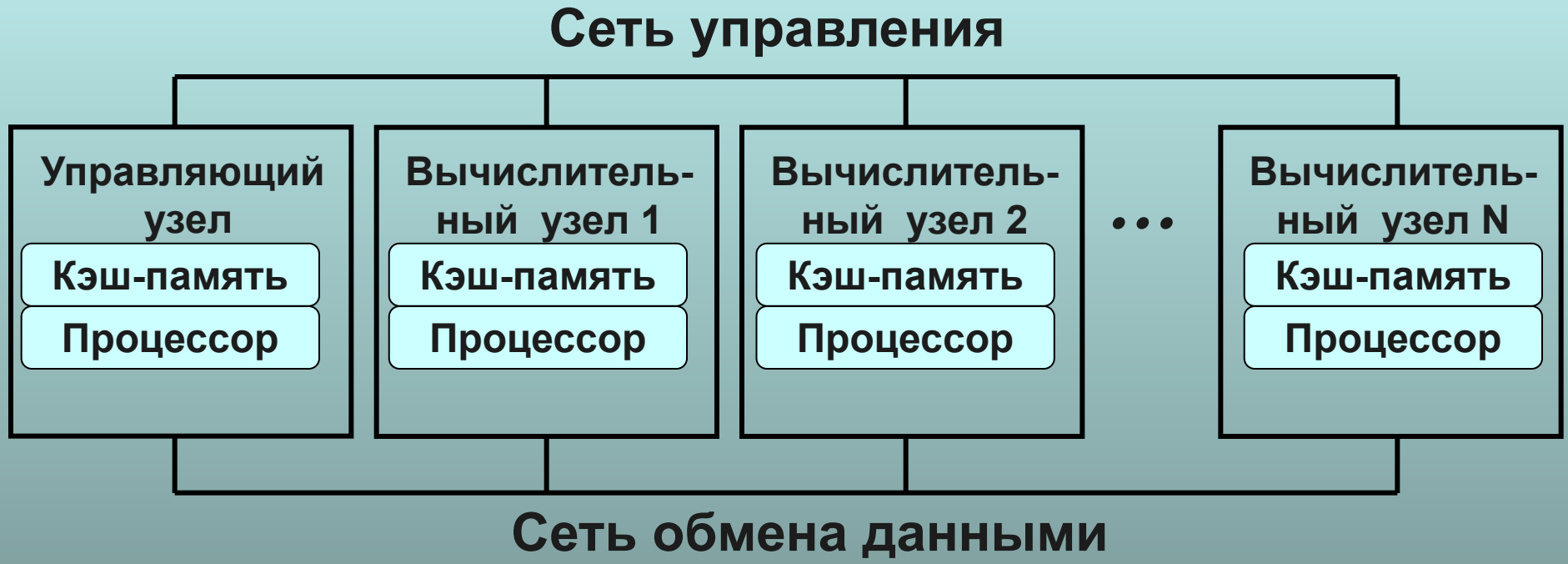
Архитектура	Система состоит из однородных базовых модулей (плат), состоящих из небольшого числа процессоров и блоков памяти. Модули объединены с помощью высокоскоростного коммутатора. Поддерживается единое адресное пространство, аппаратно поддерживается доступ к удаленной памяти, т.е. к памяти других модулей. При этом доступ к локальной памяти в несколько раз быстрее, чем к удаленной. В случае, если аппаратно поддерживается когерентность кэшей во всей системе (обычно это так), говорят об архитектуре **cc-NUMA (cache-coherent NUMA)**
Примеры	HP HP 9000 V-classHP HP 9000 V-class в SCA-конфигурациях, SGI Origin2000HP HP 9000 V-class в SCA-конфигурациях, SGI Origin2000, Sun HPC 10000HP HP 9000 V-class в SCA-конфигурациях, SGI Origin2000, Sun HPC 10000, IBM/Sequent NUMA-Q 2000HP HP 9000 V-class в SCA-конфигурациях, SGI Origin2000, Sun HPC 10000, IBM/Sequent NUMA-Q 2000, SNI RM600.
Масштабируемость	Масштабируемость NUMA-систем ограничивается объемом адресного пространства, возможностями аппаратуры поддержки когерентности кэшей и возможностями операционной системы по управлению большим числом процессоров. На настоящий момент, максимальное число процессоров в NUMA-системах составляет 256 (Origin2000).
Операционная система	Обычно вся система работает под управлением единой ОС, как в SMP. Но возможны также варианты динамического "подразделения" системы, когда отдельные "разделы" системы работают под управлением разных ОС (например, Windows NT и UNIX в NUMA-Q 2000).

КЛАСТЕРННЫЕ СУПЕРКОМПЬЮТЕРЫ

Появление высокопроизводительных кластеров не явилось большой неожиданностью. Вопрос об объединении сетевых ресурсов в единый вычислительный пул «висел в воздухе». Соответствующее решение было найдено с использованием технологий, предназначенных для локальных сетей (прежде всего Ethernet), достигших к тому времени нужного уровня развития. Будущее показало, что локальные сети — не единственный способ превратить множество вычислительных узлов в единый компьютер: сети такого рода могут быть как глобальными, так и сетями на одном кристалле.

В 1994 году сотрудники NASA Дональд Беккер и Томас Стерлинг создали кластер Beowulf. Он состоял из 16 процессоров Intel 486DX4, соединенных 10-мегабитной сетью Ethernet. Кластеры Beowulf сегодня доминируют во всех списках самых производительных вычислительных систем. Главной особенностью таких компьютеров было и остается то, что их можно собрать из имеющихся на рынке продуктов, раньше это были системные блоки, а сегодня — серверы-лезвия.

Общая структура кластерного суперкомпьютера



Один управляющий узел, остальные вычислительные, связанные в локальную сеть.

Управляющий узел: подготовка параллельных программ и данных, взаимодействие с вычислительными узлами через управляющую сеть.

Вычислительные узлы: выполнение параллельной программы, обмен данными через коммуникационную сеть.

Кластерные системы

Архитектура	Набор рабочих станций (или даже ПК) общего назначения, серверов используется в качестве дешевого варианта массивно-параллельного компьютера. Для связи узлов используется одна из стандартных сетевых технологий (Fast/Gigabit Ethernet, Myrinet, InfiniBand) на базе шинной архитектуры или коммутатора. При объединении в кластер компьютеров разной мощности или разной архитектуры, говорят о **гетерогенных (неоднородных) кластерах.** Узлы кластера могут одновременно использоваться в качестве пользовательских рабочих станций. В случае, когда это не нужно, узлы могут быть существенно облегчены и/или установлены в стойку.
Примеры	NT-кластерNT-кластер в NCSA, Beowulf-кластеры.
Операционная система	Используются стандартные для рабочих станций ОС, чаще всего, свободно распространяемые - Linux/FreeBSD, вместе со специальными средствами поддержки параллельного программирования и распределения нагрузки.
Модель программирования	Программирование, как правило, в рамках модели передачи сообщений (чаще всего - MPI). Дешевизна подобных систем оборачивается большими накладными расходами на взаимодействие параллельных процессов между собой, что сильно сужает потенциальный класс решаемых задач.

Суперкомпьютер СКИФ МГУ

Общая характеристика

Пиковая производительность	60 TFlop/s
Производительность на Linpack	47.04 TFlop/s (78.4% от пиковой)
Число процессоров/ядер в системе	1250 / 5000
Модель процессора	Intel Xeon E5472 3.0 ГГц
Объём оперативной памяти	5.5 Тбайт
Дисковая память узлов	15 Тбайт
Число стоек всего/вычислительных	42 / 14
Число блэйд-шасси/вычислительных узлов	63 / 625
Производитель	Т -Платформы

Blade-шасси, СКИФ МГУ
10 модулей T-Blade, 960 GFlop/s

Скиф МГУ Площадь зала 98 кв. метров

Суперкомпьютер Roadrunner, IBM

(10 июня 2008, Источник: BBC News)

- Пиковая производительность 1,5 Pflop/s
- 3456 оригинальных серверов tri-blade
- Производительность сервера tri-blade – 0,4 Tflops
- Гибридная архитектура
- 20 тысяч высокопроизводительных двухъядерных процессоров – 6948 AMD Opteron20 тысяч высокопроизводительных двухъядерных процессоров – 6948 AMD Opteron и 12 960 Cell Broadband Engine производства самой IBM
- Системная память - 80 Терабайт
- Занимаемая площадь около 560 кв. метров
- Общий вес оборудования - 226 тонн
- Энергопотребление - 3,9 мегаватта (376 миллионов операций на один ватт)

Платформы HPC

На формирование образа суперкомпьютеров близкого будущего повлияют несколько ключевых технологических факторов:

- технологии миниатюризации серверов; это, прежде всего, широкое принятие лезвий, хотя одновременно появляются и новые конструкции тонких «стоечных» серверов, перспективность которых также нельзя исключать;
- широкое распространение технологий организации межсоединений на основе Infiniband;
- новая система тестирования, расширяющая «классические» тесты Linpack Benchmarks.

ВЫЧИСЛИТЕЛЬНЫЕ УЗЛЫ (СЕРВЕРА)
История создания блейд - серверов

- Лезвия изобрел Крис Хипп во время Internet-бума конца 90-х. В основе банальная идея — заменить монолитные серверы простыми платами с процессорами архитектуры x86, работающими под управлением ОС Linux. В декабре 1999 года эта идея окончательно оформилась, и 1 января 2000 года была создана компания RocketLogix.
- Первым суперкомпьтером на блейд-серверах стал кластер Green Destiny.
- Компании производители: в первую очередь, IBM и HP (купила RLX Technologies); в меньших масштабах Dell, Sun Microsystems и другие.
- Сегодня крупнейшим производителем лезвий является компания HP. В HP выдвигается лозунг о «всеобщей блейдизации» (Blade Everything).

Одно из первых лезвий

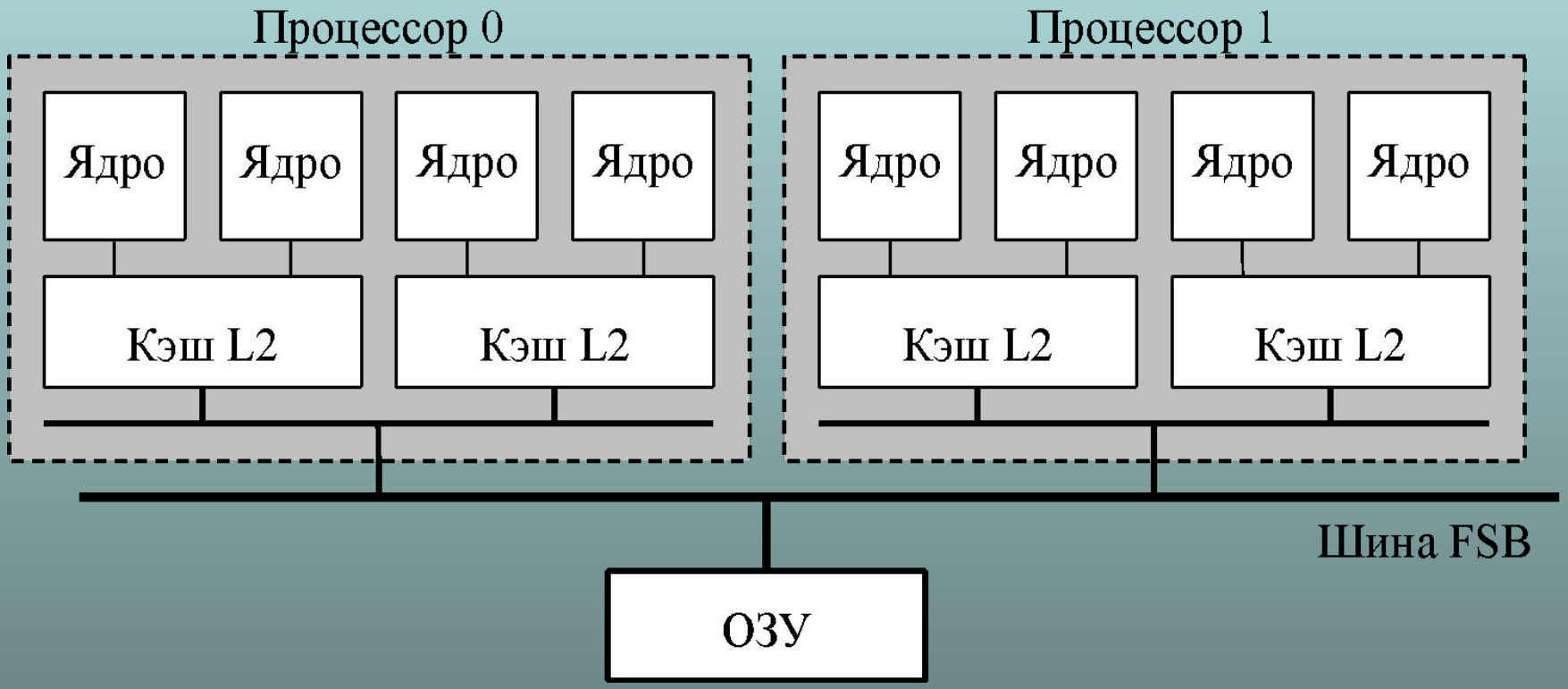
Суперкомпьютер Green Destiny — кластер Beowulf (на лезвиях)

Основные требования к блейд – серверам (HP)

- Гибкость
- Снижение энергопотребления и ресурсов охлаждения
- Средства консолидированного управления
- Активные средства обеспечения безопасности
- Прозрачный механизм виртуализации
- Автоматизация выполнения рутинных и трудоемких процедур и задач

Сервер BladeCenter HS21 для кластера IBM 1350

Схема сервера



- Кластер состоит из 6 «блейд-серверов» IBM BladeCenter HS21xx, один из которых управляющий, один запасной и четыре вычислительных.
- Вычислительный модуль состоит из двух 4-х ядерных процессор Intel Xeon 5320 “Clovertown”. Кэш второго уровня динамически разделяется между двумя ядрами.
- Два ядра, в зависимости от взаимного расположения могут обмениваться информацией либо через кэш 2 уровня, либо через шину FSB. Все 8 ядер для чтения /записи данных в ОЗУ, а так же для синхронизации одних и тех же данных в разных кэшах используют общую шину (FSB).

Характеристики BladeCenter HS21

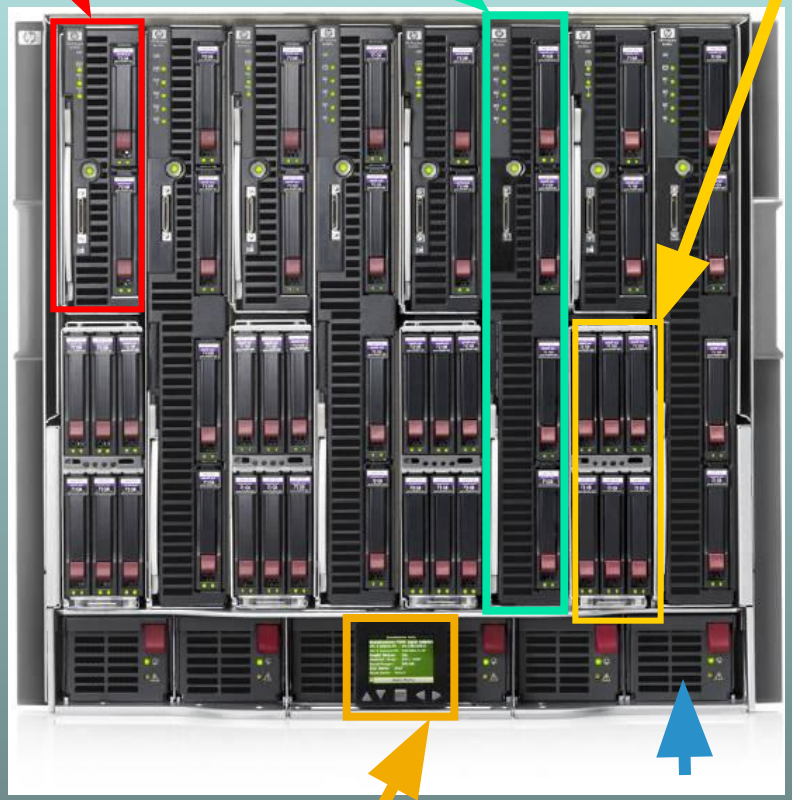
Характеристика	Вычислительный модуль	Управляющий модуль
Сервер	BladeCenter HS21	BladeCenter HS21XM
Процессор	2xIntel Xeon 1.86GHz	2xIntel Xeon 2.33GHz
Ядер в узле	4 (2 x 8)	4 (2 x 8)
ОЗУ	4x2GB	8x2GB
Кэш-память	L1: 32 + 32 KB L2: 4 × 4 MB*	
Межсоединение	InfiniBand	
Производительность ядра (пиковая)	7.44 GFlops	9.28 GFlops
Производительность кластера (пиковая)	238 GFlops	-

Программное обеспечение

- Операционная система - Red Hat Enterprise Linux
- Параллельная файловая система - IBM General Parallel File System.
- Система управления пакетными заданиями – IBM Load Leveler.
- Компиляторы - Intel C++ и Intel Fortran версии 10.1,
- Библиотеки - Intel MKL версии 10.0 и Intel MPI 3.0.

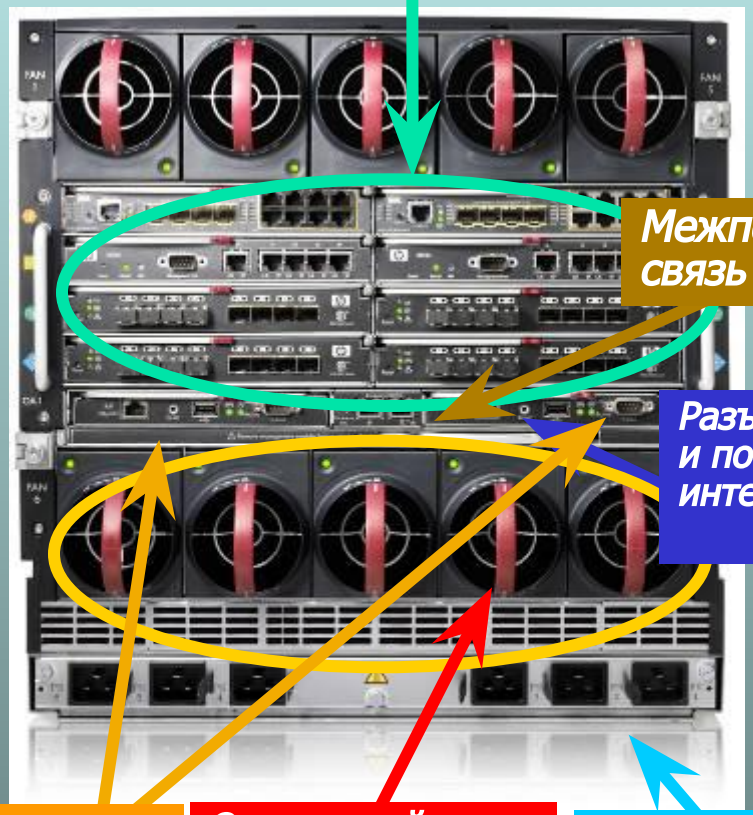
HP Полка c7000 — вид спереди

- Два форм-фактора блейдов
  - Полноразмерный блейд-сервер (до 8 в одной полке)
  - Блейд-сервер половинной высоты (до 16 в одной полке)
  - Storage-блейд половинной высоты (до 90 дисков в одной полке)
- Блоки питания
  - До 6 блоков питания с горячей заменой при мощности 2250 Вт
- Управление
  - BladeSystem Insight Display
  - Модуль управления «Onboard Administrator»

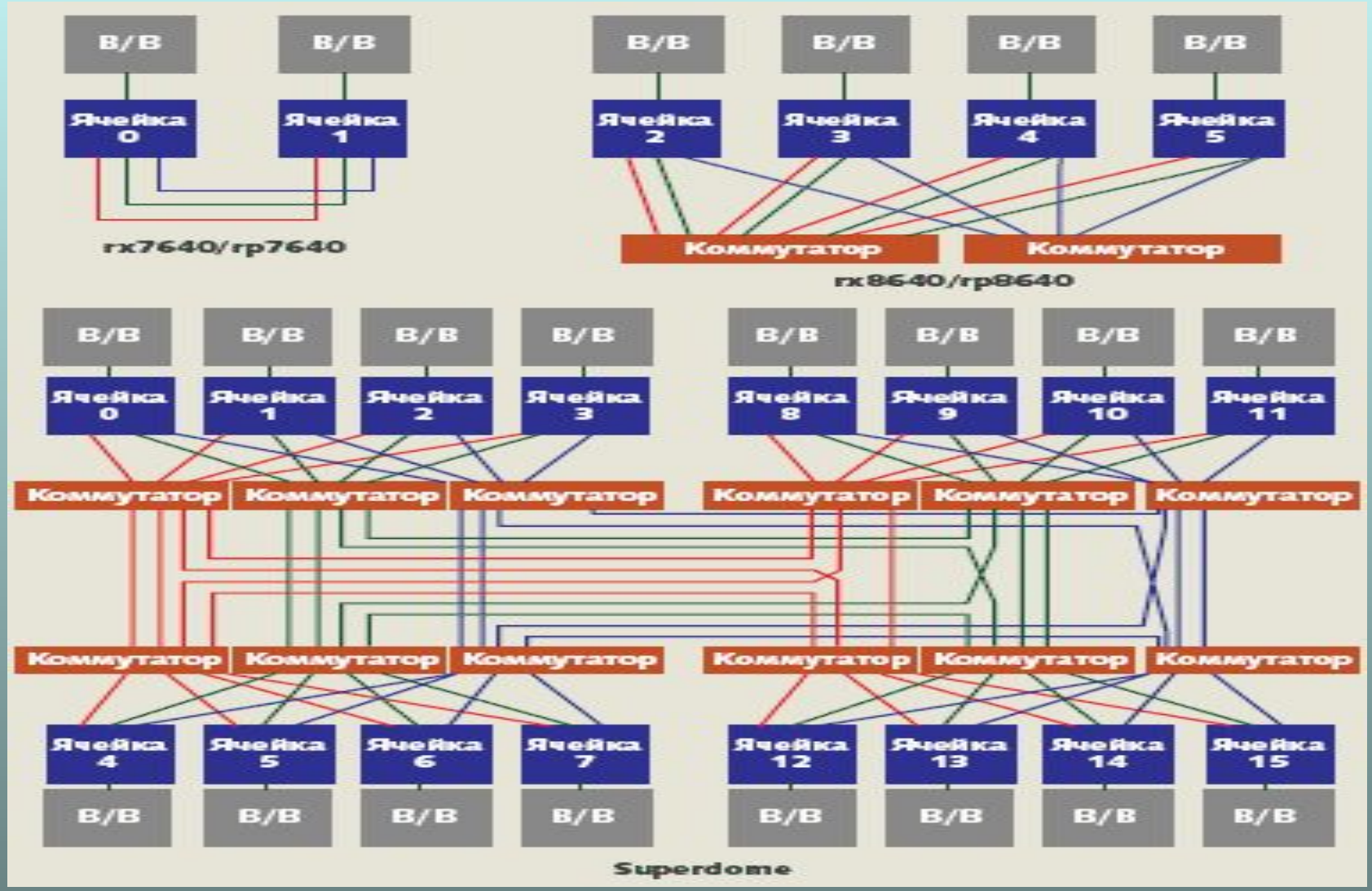


Полка c7000 — вид сзади

- Восемь коммутационных отсеков
  - До 4 резервированных фабрик ввода/вывода
  - Ethernet, Fibre Channel, iSCSI и InfiniBand
  - До 94% уменьшение количества кабелей
- 2 модуля управления «Onboard Administrator» (один стандартно)



Топология серверов HP Integrity



- Integrity rx7640 имеют две связанные напрямую ячейки; координатные коммутаторы не используются (восемь процессорных разъемов, в стойке высотой 10U, мах ОЗУ 64 Гбайта).
- В rx8640 четыре ячейки соединены двумя коммутаторами (16 процессорных разъемов в стойке высотой 17U, мах ОЗУ 64 Гбайта).
- В старших моделях, Superdome, имеется восемь ячеек и шесть коммутаторов, расположенных в двух стойках (32 процессорных разъема в 2-х стойках высотой 17U, мах ОЗУ 128 Гбайт).
- Производительность серверов. На тестах TPC-C (задержка при обращении к ОЗУ): 2-х процессорные , 4-х ядерные серверы Integrity rx4640 с Itanium 2/1,6 ГГц достигли показателя 200829 tpmC со стоимостью $2,75/ tpmC; 8-ми ядерный 4-х процессорный сервер HP rx4640-8 с процессорами Itanium 2/1,6 ГГц (Montecito) имеет 290644 tpmC при стоимости $2,71 /tpmC

КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ

- Основные: Fast EthertnetОсновные: Fast Ethertnet, Gigabit EthernetОсновные: Fast Ethertnet, Gigabit Ethernet, MyrinetОсновные: Fast Ethertnet, Gigabit Ethernet, Myrinet, cLANОсновные: Fast Ethertnet, Gigabit Ethernet, Myrinet, cLAN (Giganet), SCIОсновные: Fast Ethertnet, Gigabit Ethernet, Myrinet, cLAN (Giganet), SCI, QsNetIIОсновные: Fast Ethertnet, Gigabit Ethernet, Myrinet, cLAN (Giganet), SCI, QsNetII (QSW), MEMORY CHANNELОсновные: Fast Ethertnet, Gigabit Ethernet, Myrinet, cLAN (Giganet), SCI, QsNetII (QSW), MEMORY CHANNEL, ServerNet IIОсновные: Fast Ethertnet, Gigabit Ethernet, Myrinet, cLAN (Giganet), SCI, QsNetII (QSW), MEMORY CHANNEL, ServerNet II, InfiniBandОсновные: Fast Ethertnet, Gigabit Ethernet, Myrinet, cLAN (Giganet), SCI, QsNetII (QSW), MEMORY CHANNEL, ServerNet II, InfiniBand, Flat Neighborhood.
- *Fast Ethernet*
- **Производители оборудования:** Intel, CISCO, 3Com и др.
- **Показатели производительности:** Пиковая пропускная

Gigabit Ethernet

- **Производители оборудования:** Intel, 3COM и др.
- **Показатели производительности:** Пиковая пропускная способность - 1 Gbit/sec (125 MB/sec), полный дуплекс. В рамках TCP/IP достигаются скорости порядка 500 Mbit/sec (60 MB/sec), в рамках MPI - до 45 MB/sec
- **Программная поддержка:** Драйверы для многих версий UNIX и Windows NT, протоколы TCP/IP.
- **Комментарии:** Преимуществом данной технологии является совместимость и возможность плавного перехода с технологий Ethernet/Fast Ethernet.

Myrinet 2000

- **Производители оборудования:** Myricom
- **Показатели производительности:** Пиковая пропускная способность - 2 Gbit/sec, полный дуплекс. Аппаратная латентность порядка 5 мксек. В рамках TCP/IP достигаются скорости порядка 1.7-1.9 Gbit/sec (240 MB/sec). Латентность - порядка 30 мксек.
- **Программная поддержка:** Драйвера для Linux (Alpha, x86, PowerPC, UltraSPARC), Windows NT (x86), Solaris (x86, UltraSPARC) и Tru64 UNIX. Пакеты HPVM (включает MPI-FM, реализацию MPI для Myrinet), BIP-MPI и др.
- **Комментарии:** Myrinet является открытым стандартом. На физическом уровне поддерживаются сетевые среды SAN (System Area Network), LAN (CL-2) и оптоволокно. Технология Myrinet дает высокие возможности масштабирования сети и в настоящее время очень широко используется при построении высокопроизводительных кластеров.

InfiniBand

- **Производители оборудования:** InfiniBand Trade Association
- **Показатели производительности:** Пиковая пропускная способность каналов 10 GB/sec, латентность - 7 мксек.
- **Программная поддержка:** MPICHMPICH - бесплатная переносимая реализация MPI,MPI/Pro - реализация MPI для Linux RedHat 7.3, 7.3.
- **Комментарии:** InfiniBand предлагает удалённый прямой доступ в память (remote direct memory access - RDMA), позволяющий доставлять данные непосредственно в память процесса, не вовлекая системные вызовы. Данные могут передаваться 1-о,4-х и 12-ти кратной скоростью. Латентность на свиче InfiniBand составляет 160 наносекунд.

Архитектура InfiniBand

- **Адаптер канала хоста (Host Channel Adapter, HCA)**. Инициация и организация обмена.

  Взаимодействие: с аналогичными адаптерами HCA,; с целевыми адаптерами канала; с коммутатором InfiniBand.
- **Менеджер подсети (Subnet Manager, SM).**

  Управление и мониторинг «матрицей InfiniBand» (InfiniBand fabric). Активный менеджер SM может размещаться на одном из узлов или непосредственно на одном из коммутаторов,

  Снабжение необходимой коммутационной и конфигурационной информацией всех коммутаторов, входящих в InfiniBand fabric. Согласованная работа инфраструктуры поддерживается тем, что все остальные узлы структуры включают в себя специальные агенты, обеспечивающие обработку данных, относящихся к обмену. Менеджеры и агенты взаимодействуют по алгоритмам, заложенным в датаграммы Management Datagram.

Архитектура InfiniBand

- **Целевой адаптер канала (Target Channel Adapter, TCA).**

  Используется для подключения не серверов, а внешних устройств, в том числе устройств хранения и интерфейсов ввода/вывода, к инфраструктуре InfiniBand. Обычно адаптер TCA включает контроллер ввода/вывода, соответствующий специфическому протоколу устройства (SCSI, Fibre Channel, Ethernet и т.д.), и он же обеспечивает взаимодействие с адаптерами HCA и коммутатором InfiniBand.

- **Коммутатор InfiniBand.**

  Масштабируемость инфраструктуры InfiniBand. Он позволяет подключать необходимое число адаптеров HCA и TCA, а также дополнительные коммутаторы InfiniBand fabric.

  Организация сетевого трафика, проверка заголовоков пакетов данных и направление их по месту назначения.

  InfiniBand fabric может состоять несколько коммутаторов.

Модуль InfiniBand на 24 канала

InfiniBand Architecture PotentialInfiniBand