20. Параллельные и последовательные универсальные системные интерфейсы на примерах PCI и PCI-express.

- https://ru.wikipedia.org/wiki/PCI
- http://dmilvdv.narod.ru/Translate/LDD3/ldd_pci_interface.html
- https://ru.wikipedia.org/wiki/PCI_Express
- http://we-it.net/index.php/zhelezo/materinskie-platy/141-interfejs-pci-express-osnovnye-kharakteristiki-i-obratnaya-sovmestimost

Интерфейс PCI

PCI (англ. Peripheral component interconnect, дословно — взаимосвязь периферийных компонентов) — шина ввода-вывода для подключения периферийных устройств к материнской плате компьютера.

Стандарт на шину PCI определяет:

- физические параметры (например, разъёмы и разводку сигнальных линий);
- электрические параметры (например, напряжения);
- логическую модель (например, типы циклов шины, адресацию на шине).

Развитием стандарта PCI занимается организация PCI Special Interest Group.

Интерфейс широко применялся в бытовых компьютерах в период 1995-2005 год[1][2].

Хотя многие пользователи компьютеров думают о PCI как о способе подключения электрических проводов, на самом деле это полный набор спецификаций, определяющих, как должны взаимодействовать разные части компьютера.

Спецификация PCI охватывает большинство вопросов, связанных с компьютерными интерфейсами. Архитектура PCI был разработана в качестве замены стандарту ISA с тремя основными целями: получить лучшую производительность при передаче данных между компьютером и его периферией, быть независимой от платформы, насколько это возможно, и упростить добавление и удаление периферийных устройств в системе.

Шина PCI достигает лучшей производительности за счёт использования более высокой тактовой частоты, чем ISA; она работает на 25 или 33 МГц (фактическое значение зависит от частоты системы), и недавно были развернуты также 66 МГц и даже 133 МГц реализации. Кроме того, она оснащена 32-х разрядной шиной данных и в спецификацию было

включено 64-х разрядное расширение. Независимость от платформы является частой целью в разработке компьютерной шины и это особенно важная особенность PCI, поскольку в мире ПК всегда доминировали стандарты интерфейсов, зависимые от процессора. В настоящее время PCI широко используется на системах IA-32, Alpha, PowerPC, SPARC64 и IA-64, а также некоторые других платформах.

Однако, наиболее актуальной для автора драйвера является поддержка PCI автоопределения интерфейса плат. PCI устройства бездждамперные (в отличие от большинства старой периферии) и настраивается автоматически во время загрузки. Затем драйвер устройства должен быть в состоянии получить доступ к информации о конфигурации в устройстве с целью завершения инициализации. Это происходит без необходимости совершать какое-либо тестирование.

АРХИТЕКТУРА

Шина децентрализована, нет главного устройства, любое устройство может стать инициатором транзакции. Для выбора инициатора используется арбитраж с отдельно стоящей логикой арбитра. Арбитраж «скрытый», не отбирает времени — выбор нового инициатора происходит во время транзакции, исполняемой предыдущим инициатором.

Транзакция состоит из 1 или 2 циклов адреса (2 цикла адреса используются для передачи 64-битных адресов, поддерживаются не всеми устройствами, дают поддержку DMA на памяти более 4 Гб) и одного или многих циклов данных. Транзакция со многими циклами данных называется «пакетной» (burst), понимается как чтение/запись подряд идущих адресов и даёт более высокую скорость — один цикл адреса на несколько, а не на каждый цикл данных, и отсутствие простоев (на «успокоение» проводников) между транзакциями.

Специальные типы транзакций используются для обращений к конфигурационному пространству устройства.

«Пакетная» транзакция может быть временно приостановлена обоими устройствами из-за отсутствия данных в буфере или его переполнения.

Поддерживаются «расщеплённые» транзакции, когда целевое устройство отвечает состоянием «в процессе» и инициатор должен освободить шину для других устройств, захватить её снова через арбитраж и повторить транзакцию. Это делается, пока целевое устройство не ответит «сделано». Используется для сопряжения шин с разными скоростями (сама PCI и frontside процессора) и для предотвращения тупиковых ситуаций в сценарии со многими межшинными мостами.

Богатая поддержка межшинных мостов. Богатая поддержка режимов кэширования, таких, как:

- posted write — данные записи немедленно принимаются мостом, и мост сразу отвечает «сделано», уже после этого пытаясь провести операцию записи на ведомой шине.
- write combining — несколько запросов на posted write, идущих подряд по адресам, соединяются в мосте в одну «взрывную» транзакцию на ведомой шине.
- prefetching — используется при транзакциях чтения, означает выборку сразу большого диапазона адресов одной «взрывной» транзакцией в кеш моста, дальнейшие обращения исполняются самим мостом без операций на ведомой шине.

Прерывания поддерживаются либо как Message Signaled Interrupts (новое), либо классическим способом с использованием проводников INTA-D#. Проводники прерываний работают независимо от всей остальной шины, возможно разделение одного проводника многими устройствами.

КОНФИГУРИРОВАНИЕ

PCI-устройства с точки зрения пользователя самонастраиваемы (Plug and Play). После старта компьютера системное программное

обеспечение обследует конфигурационное пространство PCI каждого устройства, подключённого к шине, и распределяет ресурсы.

Каждое устройство может затребовать до шести диапазонов в адресном пространстве памяти PCI или в адресном пространстве ввода-вывода PCI.

Кроме того, устройства могут иметь ПЗУ, содержащее исполняемый код для процессоров x86 или PA-RISC, Open Firmware (системное ПО компьютеров на базе SPARC и PowerPC) или драйвер EFI.

Настройка прерываний осуществляется также системным программным обеспечением (в отличие от шины ISA, где настройка прерываний осуществлялась переключателями на карте). Запрос на прерывание на шине PCI передаётся с помощью изменения уровня сигнала на одной из линий IRQ, поэтому имеется возможность работы нескольких устройств с одной линией запроса прерывания; обычно системное ПО пытается выделить каждому устройству отдельное прерывание для увеличения производительности.

ИНТЕРФЕЙС PCI-E

PCI Express, или PCIe, или PCI-e (также известная как 3GIO for 3rd Generation I/O; не путать с PCI-X и PXI) — компьютерная шина (хотя на физическом уровне шиной не является, будучи соединением типа «точка-точка»), использующая программную модель шины PCI и высокопроизводительный физический протокол, основанный на последовательной передаче данных.

Разработка стандарта PCI Express была начата фирмой Intel после отказа от шины InfiniBand. Официально первая базовая спецификация PCI Express появилась в июле 2002 года.[1][2] Развитием стандарта PCI Express занимается организация PCI Special Interest Group.

ОПИСАНИЕ

В отличие от стандарта PCI, использовавшего для передачи данных общую шину с

подключением параллельно нескольких устройств, PCI Express, в общем случае, является пакетной сетью с топологией типа звезда.

Устройства PCI Express взаимодействуют между собой через среду, образованную коммутаторами, при этом каждое устройство напрямую связано соединением типа точка-точка с коммутатором.

Кроме того, шиной PCI Express поддерживается:[1][2]

- горячая замена карт;
- гарантированная полоса пропускания (QoS);
- управление энергопотреблением;
- контроль целостности передаваемых данных.

Шина PCI Express нацелена на использование только в качестве локальной шины. Так как программная модель PCI Express во многом унаследована от PCI, то существующие системы и контроллеры могут быть доработаны для использования шины PCI Express заменой только физического уровня, без доработки программного обеспечения. Высокая пиковая производительность шины PCI Express позволяет использовать её вместо шин AGP и тем более PCI и PCI-X.[2] Де-факто PCI Express заменила эти шины в персональных компьютерах.

ОПИСАНИЕ ПРОТОКОЛА

Для подключения устройства PCI Express используется двунаправленное последовательное соединение типа точка-точка, называемое линией (англ. lane — полоса, ряд); это резко отличается от PCI, в которой все устройства подключаются к общей 32-разрядной параллельной двунаправленной шине.

Соединение (англ. link — связь, соединение) между двумя устройствами PCI Express состоит из одной (x1) или нескольких (x2, x4, x8, x12, x16 и x32) двунаправленных последовательных линий.[1][2] Каждое устройство должно поддерживать соединение, по крайней мере, с одной линией (x1).

На электрическом уровне каждое соединение использует низковольтную дифференциальную передачу сигнала (LVDS), приём и передача информации производится каждым устройством PCI Express по отдельным двум проводникам, таким образом, в простейшем случае устройство подключается к коммутатору PCI Express всего лишь четырьмя проводниками.

Использование подобного подхода имеет следующие преимущества:

- карта PCI Express помещается и корректно работает в любом слоте той же или большей пропускной способности (например, карта x1 будет работать в слотах x4 и x16);
- слот большего физического размера может использовать не все линии (например, к слоту x16 можно подвести проводники передачи информации, соответствующие x1 или x8, и всё это будет нормально функционировать; однако при этом необходимо подключить все проводники питания и заземления, необходимые для слота x16).

В обоих случаях на шине PCI Express будет использоваться максимальное количество линий, доступных как для карты, так и для слота. Однако это не позволяет устройству работать в слоте, предназначенном для карт с меньшей пропускной способностью шины PCI Express. Например, карта x4 физически не поместится в стандартный слот x1, несмотря на то, что она могла бы работать в слоте x1 с использованием только одной линии. На некоторых материнских платах можно встретить нестандартные слоты x1 и x4, у которых отсутствует крайняя перегородка, таким образом, в них можно устанавливать карты большей длины, чем разъём. При этом не обеспечивается питание и заземление выступающей части карты, что может привести к различным проблемам.

PCI Express пересылает всю управляющую информацию, включая прерывания, через те же

линии, что используются для передачи данных. Последовательный протокол никогда не может быть заблокирован, таким образом задержки шины PCI Express вполне сравнимы с таковыми для шины PCI (заметим, что шина PCI для передачи сигнала о запросе на прерывание использует отдельные физические линии IRQ#A, IRQ#B, IRQ#C, IRQ#D).

Во всех высокоскоростных последовательных протоколах (например, гигабитный Ethernet), информация о синхронизации должна быть встроена в передаваемый сигнал. На физическом уровне PCI Express использует метод канального кодирования 8b/10b (8 бит в десяти, избыточность — 20 %)[1][2] для устранения постоянной составляющей в передаваемом сигнале и для встраивания информации о синхронизации в поток данных. Начиная с версии PCI Express 3.0 используется более экономное кодирование 128b/130b с избыточностью 1,5 %.

Некоторые протоколы (например, SONET/SDH) используют метод, который называется скремблинг (англ. scrambling) для встраивания информации о синхронизации в поток данных и для «размывания» спектра передаваемого сигнала. Спецификация PCI Express также предусматривает функцию скремблинга, но скремблинг PCI Express отличается от такового для SONET.

21. Суперкомпьютер, кластер, мэйнфрэйм. Особенности построения этих систем и области применения.

- https://www.nkj.ru/archive/articles/7365/
- https://ru.wikipedia.org/wiki/%D0%A1%D1%83%D0%BF%D0%B5%D1%80%D0%BA%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80
- https://ru.wikipedia.org/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80_(%D0%B3%D1%80%D1%83%D0%BF%D0%BF%D0%B0_%D0%BA%D0%BE%D0%BC%D0%BF%D1%8C%D1%8E%D1%82%D0%B5%D1%80%D0%BE%D0%B2)
- https://ru.wikipedia.org/wiki/%D0%9C%D0%B5%D0%B9%D0%BD%D1%84%D1%80%D0%B5%D0%B9%D0%BC#%D0%9E%D1%81%D0%BE%D0%B1%D0%B5%D0%BD%D0%BD%D0%BE%D1%81%D1%82%D0%B8_%D0%B8_%D1%85%D0%B0%D1%80%D0%B0%D0%BA%D1%82%D0%B5%D1%80%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B8_%D1%81%D0%BE%D0%B2%D1%80%D0%B5%D0%BC%D0%B5%D0%BD%D0%BD%D1%8B%D1%85_%D0%BC%D0%B5%D0%B9%D0%BD%D1%84%D1%80%D0%B5%D0%B9%D0%BC%D0%BE%D0%B2

СУПЕРКОМПЬЮТЕР

Суперкомпью́тер (англ. Supercomputer, СверхЭВМ, СуперЭВМ, сверхвычисли́тель) — специализированная вычислительная машина, значительно превосходящая по своим техническим параметрам и скорости вычислений большинство существующих в мире компьютеров.

Как правило, современные суперкомпьютеры представляют собой большое число высокопроизводительных серверных компьютеров, соединённых друг с другом локальной высокоскоростной магистралью для достижения максимальной производительности в рамках подхода распараллеливания вычислительной задачи.

ПРИМЕНЕНИЕ

Суперкомпьютеры используются во всех сферах, где для решения задачи применяется численное моделирование; там, где требуется огромный объём сложных вычислений, обработка большого количества данных в реальном времени, или решение задачи может быть найдено простым перебором множества значений множества исходных параметров (см. Метод Монте-Карло).

Совершенствование методов численного моделирования происходило одновременно с совершенствованием вычислительных машин: чем сложнее были задачи, тем выше были требования к создаваемым машинам; чем быстрее были машины, тем сложнее были задачи, которые на них можно было решать. Поначалу суперкомпьютеры применялись почти исключительно для оборонных задач: расчёты по ядерному и термоядерному оружию, ядерным реакторам. Потом, по мере совершенствования математического аппарата численного моделирования, развития знаний в других сферах науки — суперкомпьютеры стали применяться и в «мирных» расчётах, создавая новые научные дисциплины, как то: численный прогноз погоды, вычислительная биология и медицина, вычислительная химия, вычислительная гидродинамика, вычислительная лингвистика и проч., — где достижения информатики сливались с достижениями прикладной науки.

4 СОВРЕМЕННЫХ НАПРАВЛЕНИЯ В СУПЕРКОМПЬЮТЕРАХ

Векторно-конвейерные компьютеры

Две главные особенности таких машин: наличие конвейерных функциональных устройств и набора векторных команд. В отличие от обычных команд векторные оперируют целыми массивами независимых данных, то есть команда вида A=B+C может означать сложение двух массивов, а не двух чисел. Характерный представитель данного направления - семейство векторно-конвейерных компьютеров CRAY, куда входят, например, CRAY EL, CRAY J90, CRAY T90 (в марте этого года американская компания TERA перекупила подразделение CRAY у компании Silicon Graphics, Inc.).

Массивно-параллельные компьютеры с распределенной памятью

Идея построения компьютеров этого класса тривиальна: серийные микропроцессоры соединяются с помощью сетевого оборудования - вот и все. Достоинств у такой архитектуры масса: если нужна высокая производительность, то можно добавить процессоры, а если ограничены финансы или заранее известна требуемая вычислительная мощность, то легко подобрать оптимальную конфигурацию. К этому же классу можно отнести и простые сети компьютеров, которые сегодня все чаще рассматриваются как дешевая альтернатива крайне дорогим суперкомпьютерам. (Правда, написать эффективную параллельную программу для таких сетей довольно сложно, а в некоторых случаях просто невозможно). К массивно-параллельным можно отнести компьютеры Intel Paragon, ASCI RED, IBM SP1, Parsytec, в какой-то степени IBM SP2 и CRAY T3D/T3E.

Параллельные компьютеры с общей памятью

Вся оперативная память в таких компьютерах разделяется несколькими одинаковыми процессорами, обращающимися к общей дисковой памяти. Проблем с обменом данными между процессорами и синхронизацией их работы практически не возникает. Вместе с тем главный недостаток такой архитектуры состоит в том, что по чисто техническим причинам число процессоров, имеющих доступ к общей памяти, нельзя сделать большим. В данное направление суперкомпьютеров входят многие современные SMP-компьютеры (Symmetric Multi Processing), например сервер HP 9000 N-class или Sun Ultra Enterprise 5000.

Кластерные компьютеры

Этот класс суперкомпьютеров, строго говоря, нельзя назвать самостоятельным, скорее, он представляет собой комбинации предыдущих трех. Из нескольких процессоров, традиционных или векторно-конвейерных, и общей для них памяти формируется вычислительный узел. Если мощности одного узла недостаточно, создается кластер из нескольких узлов, объединенных высокоскоростными каналами. По такому принципу построены CRAY SV1, HP Exemplar, Sun StarFire, NEC SX-5, последние модели IBM SP2 и другие. В настоящее время именно это направление считается наиболее перспективным.

КЛАСТЕР

Кластер — группа компьютеров, объединённых высокоскоростными каналами связи, представляющая с точки зрения пользователя единый аппаратный ресурс. Кластер - слабо связанная совокупность нескольких вычислительных систем, работающих совместно для выполнения общих приложений, и представляющихся пользователю единой системой. Один из первых архитекторов кластерной технологии Грегори Пфистер дал кластеру следующее определение: «Кластер — это разновидность параллельной или распределённой системы, которая:

- состоит из нескольких связанных между собой компьютеров;
- используется как единый, унифицированный компьютерный ресурс».

Обычно различают следующие основные виды кластеров:

- отказоустойчивые кластеры (High-availability clusters, HA, кластеры высокой доступности)
- кластеры с балансировкой нагрузки (Load balancing clusters)
- вычислительные кластеры (High performance computing clusters, HPC)
- системы распределенных вычислений

Вычислительный кластер – это совокупность компьютеров, объединенных в рамках некоторой сети для решения крупной вычислительной задачи. В качестве узлов обычно используются доступные однопроцессорные компьютеры, двух- или четырехпроцессорные SMP-серверы (Symmetric Multi Processor). Каждый узел работает под управлением своей копии операционной системы, в качестве которой

чаще всего используются стандартные операционные системы: Linux, NT, Solaris и т.п. С учетом полярных точек зрения кластером можно считать как пару персональных компьютеров, связанных локальной 10-мегабитной сетью Ethernet, так и обширную вычислительную систему, создаваемую в рамках крупного проекта. Такой проект объединяет тысячи рабочих станций на базе процессоров Alpha, связанных высокоскоростной сетью Myrinet, которая используется для поддержки параллельных приложений, а также сетями Gigabit Ethernet и Fast Ethernet для управляющих и служебных целей.

Состав и вычислительная мощность узлов может меняться даже в рамках одного кластера, давая возможность создавать обширные гетерогенные (неоднородные) системы с задаваемой вычислительной мощностью. Выбор конкретной коммуникационной среды определяется многими факторами: особенностями класса решаемых задач, финансированием, необходимостью последующего расширения кластера и т.п. Возможно включение в конфигурацию специализированных компьютеров, например файл-сервера, и, как правило, предоставлена возможность удаленного доступа к кластеру через Интернет.

В большинстве случаев, кластеры серверов функционируют на раздельных компьютерах. Это позволяет повышать производительность за счёт распределения нагрузки на аппаратные ресурсы и обеспечивает отказоустойчивость на аппаратном уровне.

В то время как обычный суперкомпьютер содержит множество процессоров, подключенных к локальной высокоскоростной шине, распределенные, или GRID-вычисления, в целом являются разновидностью параллельных вычислений, которое основывается на обычных компьютерах (со стандартными процессорами, устройствами хранения данных, блоками питания и т.д.), подключенных к сети (локальной или глобальной) при помощи обычных протоколов, например Ethernet.

МЕЙНФРЕЙМ

Мейнфре́йм (также мэйнфрейм, от англ. mainframe) — большой универсальный высокопроизводительный отказоустойчивый сервер со значительными ресурсами ввода-вывода, большим объёмом

оперативной и внешней памяти, предназначенный для использования в критически важных системах (англ. mission-critical) с интенсивной пакетной и оперативной транзакционной обработкой.

Основной разработчик мейнфреймов — корпорация IBM, самые известные мейнфреймы были ею выпущены в рамках продуктовых линеек System/360, 370, 390, zSeries. В разное время мейнфреймы производили Hitachi, Bull, Unisys, DEC, Honeywell, Burroughs, Siemens, Amdahl, Fujitsu, в странах СЭВ выпускались мейнфреймы ЕС ЭВМ.

Особенности и характеристики современных мейнфреймов

Среднее время наработки на отказ. Время наработки на отказ современных мейнфреймов оценивается в 12-15 лет. Надёжность мейнфреймов — это результат их почти 60-летнего совершенствования. Группа разработки операционной системы VM/ESA затратила 20 лет на удаление ошибок, и в результате была создана система, которую можно использовать в самых ответственных случаях.

Повышенная устойчивость систем. Мейнфреймы могут изолировать и исправлять большинство аппаратных и программных ошибок за счёт использования следующих принципов:

- Дублирование: два резервных процессора, резервные модули памяти, альтернативные пути доступа к периферийным устройствам.
- Горячая замена всех элементов вплоть до каналов, плат памяти и центральных процессоров.

Целостность данных. В мейнфреймах используется память с коррекцией ошибок. Ошибки не приводят к разрушению данных в памяти или данных, ожидающих вывода на внешние устройства. Дисковые подсистемы, построенные на основе RAID-массивов с горячей заменой и встроенных средств резервного копирования, защищают от потерь данных.

Рабочая нагрузка. Рабочая нагрузка мейнфреймов может составлять 80-95 % от их пиковой производительности. Операционная система мейнфрейма будет обрабатывать всё сразу, причём все приложения будут тесно сотрудничать и использовать общие компоненты ПО.

Пропускная способность. Подсистемы ввода-вывода мейнфреймов разработаны так, чтобы работать в среде с высочайшей рабочей нагрузкой на ввод-вывод данных.

Масштабирование. Масштабирование мейнфреймов может быть как вертикальным, так и горизонтальным. Вертикальное масштабирование обеспечивается линейкой процессоров с производительностью от 5 до 200 MIPS и наращиванием до 12 центральных процессоров в одном компьютере. Горизонтальное масштабирование реализуется объединением ЭВМ в Sysplex (System Complex) — многомашинный кластер, выглядящий с точки зрения пользователя единым компьютером. Всего в Sysplex можно объединить до 32 машин. Географически распределённый Sysplex называют GDPS. В случае использования операционной системы VM для совместной работы можно объединить любое количество компьютеров. Программное масштабирование — на одном мейнфрейме может быть сконфигурировано фактически бесконечное число различных серверов. Причём все серверы могут быть изолированы друг от друга так, как будто они выполняются на отдельных выделенных компьютерах и в то же время совместно использовать аппаратные и программные ресурсы и данные.

Доступ к данным. Поскольку данные хранятся на одном сервере, прикладные программы не нуждаются в сборе исходной информации из множества источников, не требуется дополнительное дисковое пространство для их временного хранения, не возникает сомнений в их актуальности. Требуется небольшое количество физических серверов и значительно более простое программное обеспечение. Всё это, в совокупности, ведёт к повышению скорости и эффективности обработки.

Защита. Встроенные в аппаратуру возможности защиты, такие как криптографические устройства и Logical Partition, и средства защиты операционных систем, дополненные программными продуктами RACF или VM:SECURE, обеспечивают надёжную защиту.

Пользовательский интерфейс. Пользовательский интерфейс у мейнфреймов всегда оставался наиболее слабым местом. Сейчас же стало возможно для прикладных программ мейнфреймов в

кратчайшие сроки и при минимальных затратах обеспечить современный веб-интерфейс.

Сохранение инвестиций — использование данных и существующих прикладных программ не влечёт дополнительных расходов по приобретению нового программного обеспечения для другой платформы, переучиванию персонала, переноса данных и т.д.

Мейнфреймы и суперкомпьютеры

Суперкомпьютеры — это машины, находящиеся на пике доступных сегодня вычислительных мощностей, особенно в области операций с числами. Суперкомпьютеры используются для научных и инженерных задач (высокопроизводительные вычисления, например, в области метеорологии или моделирования ядерных процессов), где ограничительными факторами являются мощность процессора и объём оперативной памяти, тогда как мейнфреймы применяются для целочисленных операций, требовательных к скорости обмена данными, к надёжности и к способности одновременной обработки транзакций (ERP, системы онлайн-бронирования, автоматизированные банковские системы). Производительность мейнфреймов, как правило, вычисляется в миллионах операций в секунду (MIPS), а суперкомпьютеров — в операциях с плавающей запятой (точкой) в секунду (FLOPS).

В контексте общей вычислительной мощности мейнфреймы, как правило, проигрывают суперкомпьютерам.

Эффективность суперкомпьютера на многих прикладных задачах в значительной мере определяется профилем работы с памятью и сетью. Профиль работы с памятью обычно описывается пространственно-временной локализацией обращений — размерами обращений и разбросами их адресов, а профиль работы с сетью описывается распределением узлов, с которыми происходит обмен сообщениями, интенсивностью обмена и размерами сообщений.

Производительность суперкомпьютера на задачах с интенсивным обменом данными между

узлами (задачи моделирования, задачи на графах и нерегулярных сетках, вычисления с использованием разреженных матриц) в основном определяется производительностью сети, поэтому применение обычных коммерческих решений (например, Gigabit Ethernet) крайне неэффективно. Однако реальная сеть – это всегда компромиссное решение, при разработке которого расставляются приоритеты между ценой, производительностью, энергопотерблением и другими требованиями, во многом конфликтующими между собой: попытки улучшения одной характеристики могут приводить к ухудшению другой.

Коммуникационная сеть состоит из узлов, в каждом из которых есть сетевой адаптер, соединенный с одним или несколькими маршрутизаторами, которые в свою очередь соединяются между собой высокоскоростными каналами связи (линками).

22. Топологии коммуникационных сетей суперкомпьютеров. Подробнее о топологиях «толстое дерево» и «многомерный тор». Характеристики топологий.

- https://sites.google.com/site/kompsetisofochka/home/topologia-setej
- https://ru.wikipedia.org/wiki/Fat_Tree
- https://sites.google.com/site/exemsenko/6-topologia-kommunikacionnyh-setej-multiprocessornyh-sistem-sovremennye-superkomputery-i-vs-primenenie-ih-v-socialno-sfere

Коммуникационные сети

Эффективность суперкомпьютера на многих прикладных задачах в значительной мере определяется профилем работы с памятью и сетью. Профиль работы с памятью обычно описывается пространственно-временной локализацией обращений — размерами обращений и разбросами их адресов, а профиль работы с сетью описывается распределением узлов, с которыми происходит обмен сообщениями, интенсивностью обмена и размерами сообщений.

Производительность суперкомпьютера на задачах с интенсивным обменом данными между узлами (задачи моделирования, задачи на графах и нерегулярных сетках, вычисления с использованием разреженных матриц) в основном определяется производительностью сети,

поэтому применение обычных коммерческих решений (например, Gigabit Ethernet) крайне неэффективно. Однако реальная сеть – это всегда компромиссное решение, при разработке которого расставляются приоритеты между ценой, производительностью, энергопотерблением и другими требованиями, во многом конфликтующими между собой: попытки улучшения одной характеристики могут приводить к ухудшению другой.

Коммуникационная сеть состоит из узлов, в каждом из которых есть сетевой адаптер, соединенный с одним или несколькими маршрутизаторами, которые в свою очередь соединяются между собой высокоскоростными каналами связи (линками).

Структура сети, определяющая, как именно связаны между собой узлы системы, задается топологией сети (обычно решетка, тор или толстое дерево) и набором структурных параметров: количество измерений, количество уровней дерева, размеры сторон тора, число коммутаторов на уровнях дерева, число узлов сети, портов у маршрутизаторов и т. д. На рис. 1 приведен пример топологии четырехмерный тор 3x3x3x3.

Архитектура маршрутизатора определяет структуру и функциональность блоков, отвечающих за передачу данных между узлами сети, а также необходимые свойства протоколов канального, сетевого и транспортного уровней, включая алгоритмы маршрутизации, арбитража и управления потоком данных. Архитектура сетевого адаптера определяет структуру и функциональность блоков, отвечающих за взаимодействие между процессором, памятью и сетью; в частности, на этом уровне осуществляется поддержка MPI-операций, RDMA (Remote Direct Memory Access — прямой доступ к памяти другого узла без участия его процессора), подтверждений получения другим узлом пакета, обработки исключительных ситуаций, агрегации пакетов.

Для оценки производительности коммуникационной сети чаще всего используются три характеристики: *пропускная способность* (количество данных, передаваемых за единицу времени); *коммуникационная задержка* (время передачи данных по сети); *темп выдачи сообщений* (обычно отдельно рассматривают темп выдачи при

посылке, приеме и передаче пакетов между внутренними блоками маршрутизатора).

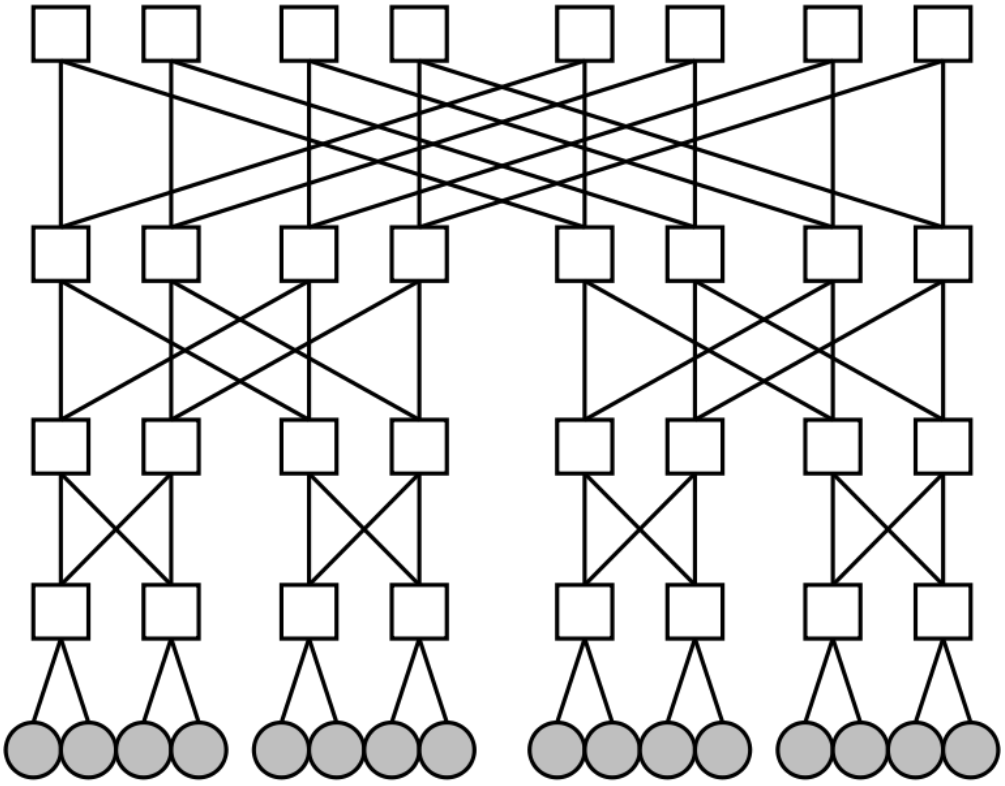
Для полноты картины данные характеристики измеряются на разных видах трафика, например, когда один узел рассылает данные всем остальным, либо, наоборот, все узлы шлют данные одному, либо когда все узлы посылают данные случайным адресатам. К современным сетям предъявляются требования по функциональности:

- эффективная реализация библиотеки Shmem, как варианта поддержки модели односторонних коммуникаций, и GASNet, на которой основаны реализации многих PGAS-языков;
- эффективная реализация MPI (обычно для этого требуется эффективная поддержка механизма кольцевых буферов и подтверждений для принятых пакетов);
- эффективная поддержка коллективных операций: широковещательной рассылки (посылки одинаковых данных одновременно многим узлам), редукции (применение бинарной операции, например, сложения, ко множеству значений, получаемых от различных узлов), распределения элементов массива по множеству узлов (scatter), сборки массива из элементов, находящихся на разных узлах (gather);
- эффективная поддержка операций межузловой синхронизации (как минимум барьерной), эффективное взаимодействие с сетью большого количества процессов на узле, обеспечение надежной доставки пакетов.

Также важна эффективная поддержка работы адаптера с памятью узла напрямую без участия процессора.

ТОЛСТОЕ ДЕРЕВО

Сеть fat tree (утолщенное дерево) — топология компьютерной сети, изобретенная Charles E. Leiserson из MIT, является дешевой и эффективной для суперкомпьютеров. В отличие от классической топологии дерево, в которой все связи между узлами одинаковы, связи в утолщенном дереве становятся более широкими (толстыми, производительными по пропускной способности) с каждым уровнем по мере приближения к корню дерева. Часто используют удвоение пропускной способности на каждом уровне.



Сети с топологией fat tree являются предпочтительными для построения кластерных межсоединений на основе технологии Infiniband.