ASIC, как много в этом звуке

ASIC, как много в этом звуке.

Не все моменты удалось прояснить, поэтому в логике изложения есть существенные упрощения. Так что возможно по итогам обсуждения на форуме можно будет выстроить уже правильную "картину мира".

Основа любого современного коммутатора – ASIC (application specific integrated circuits), они же заказные специализированные микросхемы. Говоря упрощенно, они умеют очень быстро делать ограниченное число операций по перекладыванию пакета из одного буфера в другой в соответствии с заданным алгоритмом.

Основные интерфейсы этого "черного ящика" можно представить как:

1. Ethernet MII (Media Independent Interface) в определенном типе и количестве (сейчас чаще SMII, это фактически "свосьмиренный" MII). На сегодня типично – 8 портов для "маленьких" и 24 для "больших", иногда в комбинациях типа: 24 "сотки" и 4 "гигабита"
2. Интерфейс процессора управления – шина типа Интел или Моторола (в тех, что попроще - последовательный управляющий интерфейс I2C SPI), сюда подключается управляющий процессор, с точки зрения которого этот ASIC является набором регистров, которые можно читать и записывать.
3. Высокоскоростной интерфейс – как правило, частный интерфейс производителя ASIC, с очень большой производительностью, который используется для объединения микросхем между собой. Именно таким образом получаются стековые коммутаторы – стековая шина реализуется отдельным чипом и соединяется с интерфейсом, а так же модульные – этим интерфейсом идет соединение через бэкплэйн с модулем свич-фабрики, где стоит уже другой ASIC, обеспечивающий коммутацию между этими интерфейсами. Иногда на подобный интерфейс можно установить специальный процессор (Network Processor) для сложной программируемой обработки пакетов на очень высоких скоростях.

Кстати, сейчас никто не дает гарантии что внутри ASIC нет (к примеру) микропрограммного ядра со своим набором команд и навечно вшитой программой. Поэтому можно сказать, что ASIC представляет собой логический автомат, ориентированный на максимальную производительность с возможностью задания заранее установленных производителем настроек. И нет возможности напрямую управлять обработкой пакетов в реальном масштабе времени.

Чтобы компенсировать функции, отсутствующие в ASICе, достаточно у него задействовать порт процессора (или до более умного ASICa). Обрабатываемый пакет дожидается результата обработки (для этого процессор берет копию заголовка из той же памяти) и двигается дальше по установленным правилам. Ожидание пакетом своей участи для ASICa, владельца этой очереди, - не зависящий от него форсмажор, и существенное снижение скорости обработки.

Более простые ASIC в случае затруднений вообще не мудрствуя перенаправляют отфильтрованную часть трафика в отдельный порт CPU, давая последнему всю полноту решения (и задержки времени).

Хотя для устройства в целом ситуация "особой" обработки вполне штатная - так реализуются большинство недостаточно устоявшихся протоколов, а так же некоторые особо сложные. Но управляющий процессор относительно слаб, и это приводит к ряду вполне объяснимых, но очень неприятных эффектов. Иногда достаточно нескольких настроек, что бы отправить пакеты на программную обработку, а коммутатор - в глухой коматоз.

Ранее производители оборудования самостоятельно проектировали ASIC, это было овеянное романтикой поиска время технологической борьбы. Но прогресс повлек усложнение и специализацию, на первое место выдвинулись разработчики чипов – типа Broadcom и Marvell. Поэтому производители коммутаторов вынуждены использовать "что есть", и разница между "одноклассниками" разных вендоров уменьшилась до неприличия.

В настоящее время функциональность и производительность управляемого коммутатора примерно на 80% определяется используемым ASIC, и на 20% - программным кодом, работающим на процессоре, который управляет ASICом. Если, конечно, в софте нет клинических ошибок. :-) Соответственно, для самых простых ASIC, которые программируются считыванием EPROM, или вообще запаиваемыми на землю перемычками, эта разница вообще трудноуловима.

Надо отметить, что прогресс не стоит на месте, сейчас есть и программируемые ASIC, да и архитектура коммутаторов далеко ушла от "классической" (чего только стоит широкое применение TCAM, Ternary Content Addressable Memories, для хранения CAM таблиц, ACL и маршрутов FIB). Но пока оставим это для дальнейших выпусков.

Таким образом, полностью управляемый коммутатор состоит из:

1. ASIC и необходимая "обвязка" типа осцилляторов и, иногда, дополнительной памяти;
2. Управляющий процессор, как правило, достаточно слабый (в c3750 - 700МГц), и необходимая ему обвязка – flash, RAM, и кнопка reset;
3. Микросхемы PHY – реализующие требуемый физический уровень Ethernet, в большинстве случаев это сборки по 4-8 PHY в одном корпусе, а так же подключаемая к ним аналоговая часть (трансформаторы и разъемы RJ45.);
4. По мере возможности – стековый интерфейс, гнездо под модули и т.д.
5. Горсть резисторов, конденсаторов (желательно танталовых!), светодиодов, и блок питания.

Вот фото наиболее мне симпатичного коммутатора 3com 4400 (не слишком навороченного, для простоты понимания).



Наиболее интересны чипы с нашлепкой Broadcom. Три небольшие микросхемы обрабатывают PHY уровень Ethernet, по 8 физических портов каждая. Обычные чипы массового производства.
PHY - преобразует поток данных из принятой для Ethernet системы кодирования в простой поток бинарных данных. Он никак не обрабатывает поток данных, а выполняет только проверку контрольных сумм и убивает пакеты с неверными контрольными суммами.

Большая микросхема собственно и есть ASIC коммутатора (часто это называют уровнем MAC, хотя сейчас это большая условность). Даташит на этот чип (BCM5615) можно посмотреть тут, там прописаны все возможности устройства. И можно легко понять, почему 3com 4400 не делают в 48-ми портовом исполнении - поддержка связи между ASICами появляется только в следующей их версии, BCM5645, там есть симпатичная диаграмка, показывающая способ объединения.

К сожалению, последнее время производители коммутаторов (например Cisco) взяли моду перемаркировывать ASIC в свои "цвета", и точный тип чипа можно определять только приблизительно. Не лучше и наклеивание радиаторов - отрывать жалко, и мой краткий поход за информацией на склад закончился "ничем" (придется подойти к вопросу более основательно).

Между PHY и MAC есть стандартный цифровой интерфейс SII (в других устройствах могут быть всякие его разновидности, RMII, MII, S3II и т.д...), в котором четко прописаны тактовые частоты и количество линий. Так, даже в самых простейших 8-ми портовых PHY RTL8208 используется SMII в котором под каждый порт выделена отдельная линия и тактовая частота передачи данных составляет 125 Мгц.

Пакеты, приходящие с порта, попадают в PHY, раскодируются из Ethernet и отправляются в MAC. Никакой существенной задержки тут не происходит. Каждый PHY связан с MAC своей физической линией в соответствии со стандартом, поэтому общей шины нет и никаких коллизий или ожиданий быть не может.

Другое дело внутренняя шина MAC, в которой все эти 24 порта должны быть обработаны на wire speed, но все это уже внутри "черного ящика" ASIC. Вероятно, чипмейкер планировал, что свои штатные операции ASIC должен выполнять на "полной скорости" - по крайней мере хочется в это верить.

Однако коммутаторы разных производителей на тестах упорно демонстрируют большие отличия в производительности. Причем часто такие, которые никак не укладываются в описанные выше 20% зависимости от управляющего софта. Самая вероятная причина этого - использование заведомо слабого для заявленных функций ASIC, и отработка недостающего функционала на процессоре.

Да так, что отрыв между реальными и заявленными значениями вообще доходит до неприличных

размеров. Просто они не видны так отчетливо после обычного объяснения "ну дешевая железка, что с нее взять, надо купить помощнее". Грешат этим многие, и это еще один повод для будущего обзора, когда дело дойдет до сравнения параметров ASIC и заявлений производителя коммутаторов.

Иначе складывается неправильная ситуация, когда при выборе оборудования совершенно невозможно ориентироваться на техпараметры, виден только "бренд". Который, увы, не всегда адекватен ситуации в этой сверхдинамичной отрасли.

Но вернемся к технике, рассмотрим путь пакета в ASICe самого простого управляемого коммутатора (надеюсь, в следующих выпусках дело дойдет и до более сложных устройств).

Входящий пакет попадает в ASIC. Из MAC адреса получателя пакета выбирается определенное количество бит в соответствии с размером Look-up таблицы коммутатора. К примеру, если размер таблицы составляет 1024 записи, то выбирается 10 бит. Которые являются адресом для чтения информации из Look-up таблицы.
Прочитанное из таблицы соответствий значение будет списком портов, куда ASICу нужно отправить пакет.

Т.е. на недорогих свитчах для коммутации используется не весь MAC адрес. Причина - используемая в буферах сверхбыстрая многопортовая память достаточно дорогая, и увеличение разрядности адресных шин резко повышает стоимость устройства в целом.

Но чем меньше "окно" выборки, тем больше вероятность коллизии (совпадение части адресов, и соответственно рассылка по нескольким адресам). Поэтому иногда используют специальный буфер CAM, который хранит только коллизии, что существенно уменьшает вероятность переполнения ими таблицы.

Хуже того, если злоумышленник специально или случайно изменит MAC так, что это не попадет в "контролируемую" область - MAC-секьюрити окажется бессилен. Конечно, есть методы контроля этого процесса через CAM (буфер, где хранятся коллизии), но не всегда и не везде они доступны.

Так что нельзя забывать, что недорогой коммутатор отделяет от состояния хаба только не слишком высокая вероятность совпадения, которую может перекрыть "умелый" китайский производитель, сэкономивший на правильной прошивке MAC-адресов в сетевые карты.

Еще один пример. Если попробовать реализовать по описанной схеме MAC-секьюрити (блокировку всех MAC, кроме избранных), то пропускать любой пакет коммутатор все равно будет, в механизме секьюрити проверка на MAC источника не ведется. Но вот ответ можно будет уже отследить по look-up таблице, и... Закрыть порт совсем (или на время). Блокировать иначе однонаправленный поток невозможно.

Именно поэтому в большинстве простых коммутаторов нет функции "не пропускать MAC-адреса кроме указанных". С точки зрения чипмейкеров это не устраняет атаку, а только ухудшает ситуацию. :-) А желания российских домашних сетей (ведь по идее можно и "заморозить" "обучение" (learning) на порту) им сугубо параллельны...

Подводя итог, можно сказать, что в простейшем случае для рассылки пакета нужно сделать всего несколько действий:

1. Взять адрес из MAC адреса получателя;
2. Прочитать по этому адресу значение в Look-Up;
3. Отправить пакет в те порты, которые указанны в этом значении.

Аналогично делается и заполнение таблицы look-up, только адрес берется из MAC отправителя и делается не чтение, а запись. При этом обе операции могут проходить параллельно.

Для работы с Vlan 802.1q пакеты нужно обрабатывать в ASICах по MAC+VID (VLAN ID). Фактически, это единственная сложная функция коммутаторов L2, требующая быстрой обработки в реальном времени. Но реализовать ее можно несколькими путями.

... продолжение следует (не все Vlan одинаково полезны).

Разное.

Для начала, забавное:



Прислал **Smile**

Интересные вещи выясняются, если внимательно читать иные новости. Вот, например, заметка о активно распиариваемой скупке сетей NetByNet'ом. Вот две цитаты:
... пояснил "Бизнесу" советник гендиректора NetByNet Александр Милицкий... ... Весной 2006 года разрозненные сети, обслуживающие 25 тыс. пользователей в разных районах Москвы, после приобретения NetByNet за $10 млн западным инвестфондом были объединены в три компании... Но все равно NetByNet, занимая на рынке широкополосного доступа в Москве всего лишь 1,1%, не может конкурировать с грандами",- резюмирует Овчинников.

Интересные оценки, если принять весь рынок москвы в 1,1 млн, то фактически аналитик и представитель владельца разошлись в 2-2,5 раза... Так кто же прав? Или просто is fecit cui prodest?

Оказывается, не перевелись еще энтузиасты домашних сетей на Руси. Вот, например, проектик из Воронежа - www.net.vrn.ru. Так сказать, визуализация сетей города.

И еще один любительский проект - замена знаменитого сетевого киллера, чата vypress, на более корректную софтину . Нужны бетатестеры. ;-)

Интересная аналитика по МРК. Либерализация рынка связи ударила по "Ростелекому". "*Ростелеком" опубликовал отчетность по РСБУ за I полугодие 2006 года, согласно которой чистая прибыль снизилась по сравнению с аналогичным периодом 2005 года в 2 раза - до 2,735 млрд руб.* И далее все подробно и по пунктам. Похоже, Ростелеком ждут не легкие времена. Впрочем, это все слабо пересекается с передачей данных.

Очередной технологический труп - I-mode. Хоть и смежная отрасль, но весьма показательно, как модное и внешне перспективное направление развития превращается в большие и безвозвратные финансовые потери.

И вот, кстати, обратный пример. Movielink лицензировал технологию продажи DVD в онлайне. *Онлайновый видеосервис Movielink LLC лицензировал у Sonic Solutions технологию продажи содержимого DVD через интернет для последующего прожига на диск и просмотра на DVD-плеерах.* Никак не идут продажи потокового видео через сеть. Зато способы обхода растут и множатся... Впрочем, можно предположить, что создатели сервиса и не думают о зарабатывании на предлагаемой услуге. Более реальная цель - захват ниши для последующих продаж файлов для прямого просмотра в ТВ, причем без каких либо значительных инвестиций. Чем не **FoD**?

В завершение - радио-фонари от Голден Телекома. Оригинально, и более-менее эффективно.

Обновление в разделах:

- Новые иллюстрации в Фотографиях сетевого оборудования, Planet WAP 4033.
- Пополнение в Сертификатах и лицензиях, Серверы SUN: Netra, SunFire, Sun Enterprise (CCC), Приложение.
- Свежие Инструкций к оборудованию на русском языке, Руководство по установке и эксплуатации SkyStar 2.
- В разделе **shop.nag.ru** две тотальные распродажи. Во-первых, **3com 610 (компактная версия знаменитого 3com 1100)** за **$95** и во-вторых, **Nortel BPS2000 (до L4)** за **$200**.

Фотогалерея

Стойка на одном из узлов...



Снизу вверх:

- Блок розеток;
- 3 аккумулятора Delta 105 А/ч;
- ИБП Инелт 1кВт;
- Оптический кросс;
- Корзина Dlink с конвертерами;
- Ну и Nortel BPS2000.

Пока работает на 100Мбит/с, но в планах гигабит, SFP модуль уже стоит.

Прислал **Alexander Popov**, iWAN Company

Красота монтажного дела...



- Ниточки из оптики;
- Верхолаз;
- Оптика без оболочки.

Прислал **Cowboy**

Картинки Мультиплексора Ericsson (бывший Axxessit) Axx Connect 9100, который стоит примерно 5000$...



- Крупный план;
- Вид сбоку.

Прислал **Vasiliy**

Фотографии WiFi сетей Владивостока.



Линк - 31 км, без усилителей. BlueBox прошивка на стандартном Dlink AP2100, POE. Ходит VoIP 40 телефонов, и немного интернета.

- Антенны
- Антенны-2
- Активное оборудование

Прислал **Иван**

Традиционно для NATa в eth-сеть провайдера, и как FTP-сервер на остеклённой лоджии стоял старенький П-1, который немного не дожил до десятилетнего возраста из-за почти полного апгрейда основного компьютера.



Кроме замены шлюза было решено прикупить ИБП, переставить в него имеющиеся диски и облагородить окружающее пространство. Вот что из этого получилось...

- ИБП на 500 и 750 ВА для шлюза и основной машины (15 минут)
- Мелкий свичик на 5 портов и провода
- Athlon 2800 и 512Мб оперативки
- 2 SATA- и 2 IDE-диска общей ёмкостью примерно 700Гб
- Ради охлаждения дисков пришлось поработать ножовкой

Прислал **Dmitry V. Davidov**

Анонс

- Заметка про мониторинг и построение графиков с помощью rrdtool в юниксах;
- Выгодно ли покупать небольшие сети? (сбор статистики, жду писем);
- Послегрозовой ремонт управляемых коммутаторов;
- Утилита для управления сетью через SNMP;
- Традиционный пункт - ссылки на интересные места Сети. Присылайте письма - они очень нужны для обзоров. Обязательно сообщайте, нужна ли Ваша подпись, ссылка, или лучше обойтись без нее;
- В "ужастиках" - Субботняя гроза в середине лета.

Долгострой:

- Пример настройки сервера для терминирования PPPoE (перенесен в долгострой);