Управляется программно: the Facebook

Существует три вещи, на которые комментаторы Нага реагируют со страстью и напором: SDN, Facebook и "отечественный производитель". Для поддержки градуса обсуждений, решено написать материал про все три составляющие и вот он. Можно прочитать и поучаствовать.

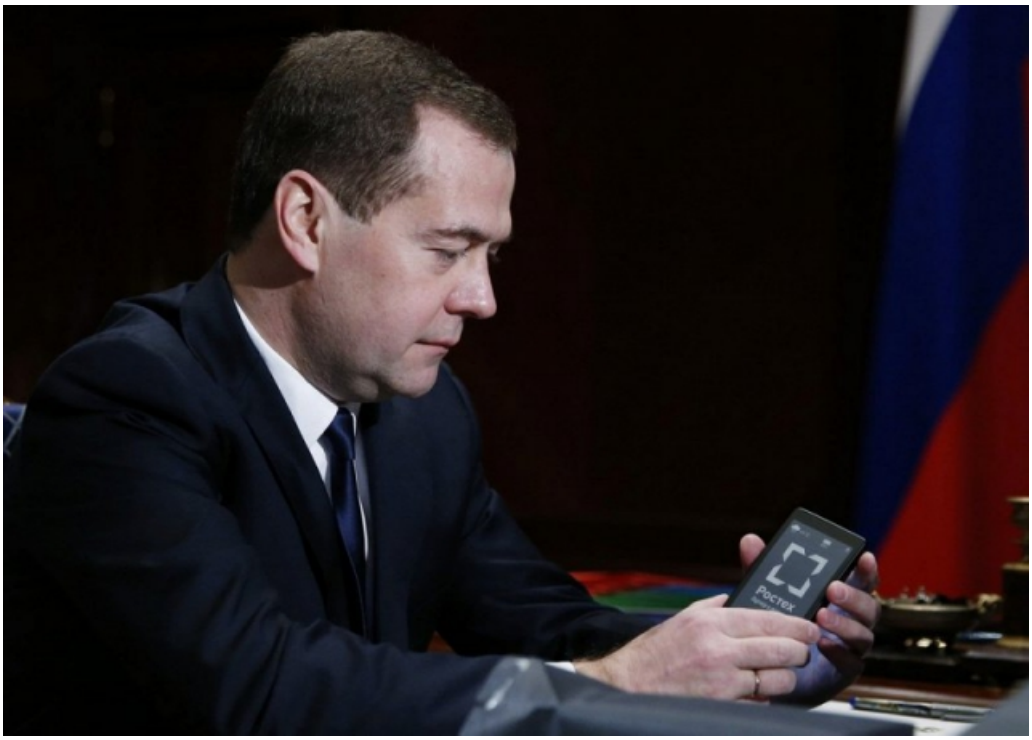
Логика повествования будет соответствующая - с конца. То есть, с "отечественного производителя", ибо тема эта модная, нужная и важная. Есть надежда, что прочитают в министерстве управления отраслью по Твиттеру, ибо, уверен, что ссылку на статью в Твиттер запостят (да вот хотя бы я!).

Итак, пару недель назад, соответствующие уголки рунета взорвали сообщения прессы про "первый честно отечественный маршрутизатор". Вентилятор дул со стороны Финского Залива из ~~шараги~~ НИИ, название которого мы упоминать не станем. Но шила в интернетах утаить невозможно, ибо, как оказалось, "*честный отечественный маршрутизатор*", который может "*производительность - до 2 млн. пакетов в секунду (при условии, что пакет - 64 байта)*", отечественный только отчасти. Ну как - от части? От части аппаратного решения - не получилось. Нехорошие комментаторы воспользовавшись вражеской технологией поиска по картинкам, нашли реального производителя железки в Китае. От части программного получилось только "*глубоко переработанное ядро Linux*". (все фразы заключенные в кавычки италиком являются цитатами официального оправдания).

В общем, как обычно.

Теперь в пословицы, поговорки и сетевой фольклор прочно вошли следующие уровни определяющие "отечественность технологий":

1. Да, отечественные (образец выше);
2. Честно отечественные;
3. Мамой клянусь, отечественные;
4. Век воли не видать, отечественные, Президенту показывали.



Поскольку так никто и не смог внятно сформулировать Пятый, высший уровень отечественности компьютерно-коммуникационных технологий, придется приходить на помощь. В общем, я сейчас вам покажу в Сети место для вдохновения.

Вот оно:

[Управляется программно: the Facebook]

Для представителей ~~шаражек~~ НИИ импортозаместительных технологий и Министерства даю подсказку - если нажать на картинку, то попадете на воистину клондайк для инноваторов. Поскольку проект делается исключительно на английском языке, то перевод документации на русский, уже будет значительным прорывом. Примерно как "Маршрутизатор мультипротокольный специального назначения". А может быть и более сложный нейминг, но я сразу предупреждаю, что бренд "Управляется программно" я ~~продам за очень большие деньги~~ оставляю за собой.

Проект же инициировал, поддерживает и финансирует Facebook. Использует решения в своих целях, и строит на базе опубликованных документов уже десятый (кто ж их считал-то) ЦОД в городке Лулео, что на северо-востоке Швеции, на берегу Ботнического залива:

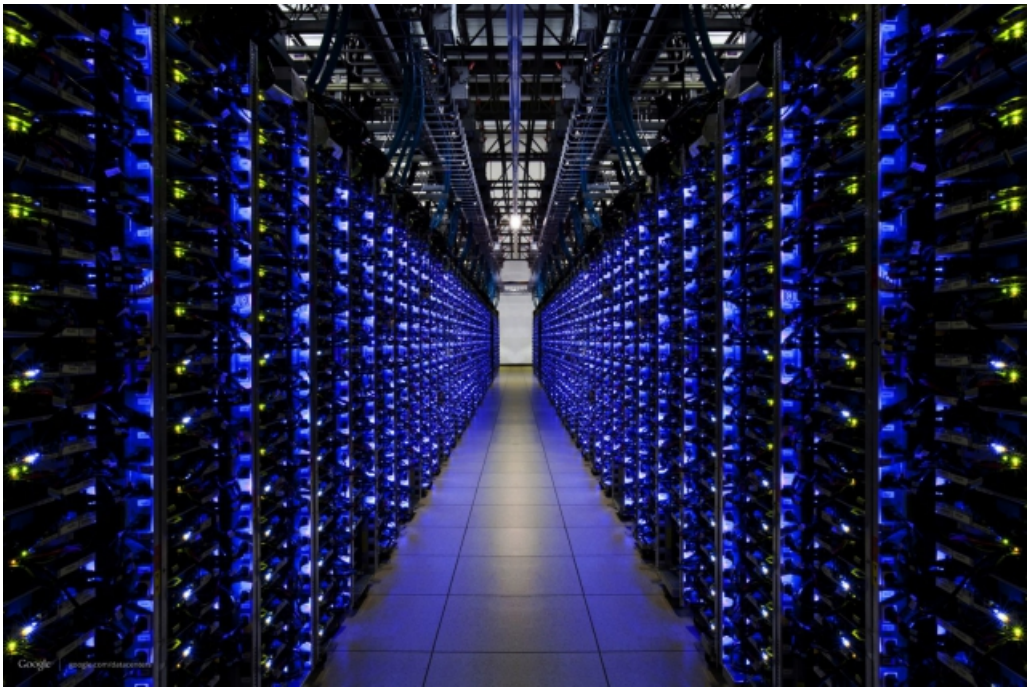
Вот этот ДЦ с высоты полета коптера:

(есть еще очень детальное фото для изучения и удовлетворения любопытства - осторожно, трафик)

Вот проходная с Гугл.Стрит (выше, с карты можно "побродить" в окрестностях):

Вот очень милый логотип изо льда и снега - как-никак, ДЦ находится на широте Полярного круга:

А вот - внутри:



Но вернемся к проекту Open Computer.

Весь проект состоит из трех крупных разделов с подразделами. Это "серверные технологии", "Системы хранения данных" и "Технологии дата-центра". Немного рассмотрим по каждому пункту.

Server Technology

Раздел описывает инженерно-технические решения серверов, стоек, и прочей нужной начинки. Самый, на мой взгляд, интересный раздел, где можно серфить ночами. Имеет множество подразделов, а именно:

Networking

Логично, что это сетевые составляющие Дата Центра: коммутаторы и маршрутизаторы. О них будет подробнее.

Server Design

Устройство серверов и дизайн материнских плат. Причем можно найти решения для Intel, AMD и ARM. Есть и решение класса "устройство на чипе" - SoC. Блоки питания тоже описываются.

Open Vault (Storage)

Системы хранения данных, как юниты, так и целые стойки. Для _очень_ больших хранилищ, пишут нам разработчики. Фото юнита хранения:

Hardware Management

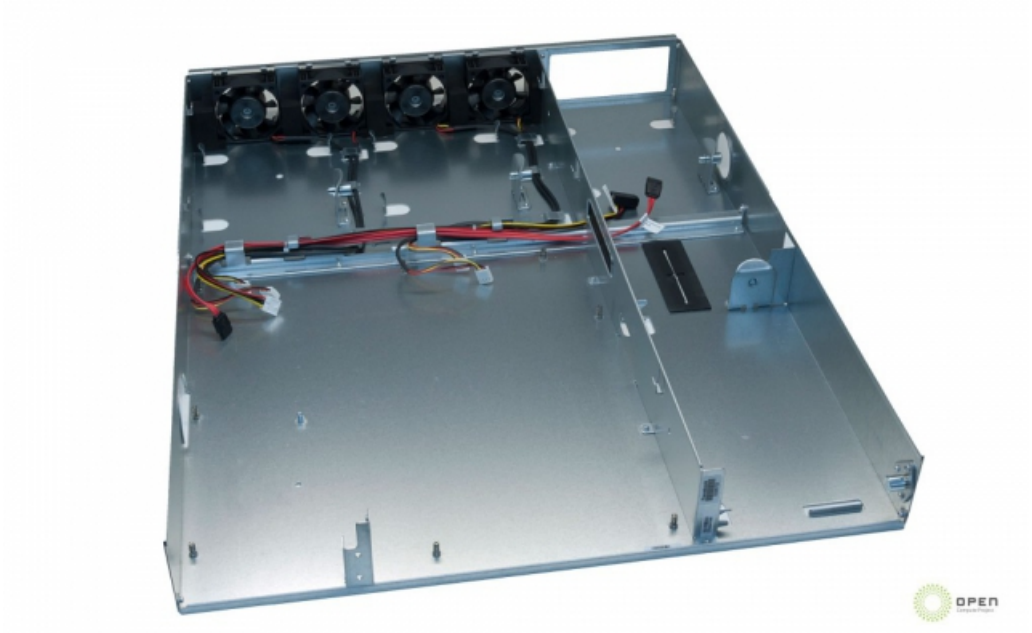
Описание и спецификации инструментария управления оборудованием. Считаю, что это маст рид каждому системному администратору по определению.

Power Supply

Думаю, что понятно из названия - это энергоснабжение. Описание, CAD-файлы и все остальное нужное для того, чтобы серверы правильно энергоснабжались. Блоки питания, например.

Chassis

Это механическая составляющая. То куда крепятся блоки питания и материнские платы. Стандартная конструкция, под которую разработаны предыдущие составляющие, выглядит так:



Дешево и стандартно - главный девиз. Как мне кажется, мы наблюдаем создание нового отраслевого стандарта, ибо все, что требуется - взять чертежи и согнуть пластины. Края обработать напильником. Очень технологично.

Storage Technology

Может показаться, что разделы дублируются. ~~Так оно и есть.~~ Но на самом деле, в предыдущем серверном разделе речь шла о модулях хранения, а здесь - о системах. Раздел состоит из всего одного подраздела:

Storage

Но внутри имеется несколько проектов:

Cold Storage - системы "длительного хранения". Например, резервные копии и/или данные, необходимые для "юридической поддержки". Идеологам "хранить логи полгода" должно понравиться.

Fusion-io - проект системной PCIE-платы для хранилищ высокой плотности. Один адаптер обслуживает несколько (15) дисковых массивов с плотностью хранения "до 3,2TB".

Hyve - дизайн двухюнитового модуля, куда упаковываются 15 дисков под управлением Fusio-io.

OpenNVM - концепция модуля хранения на "нестандартных носителях". Например - SSD. Это в большей степени программный проект - подробности пока опустим.

И уже упомянутый выше **Open Vault** - так же входит в раздел Storage.

Data Center Technology

В данном разделе собраны технологии и спецификации по общему управлению ДЦ, включая менеджмент, инженерное обеспечение, энергоснабжение, механику и безопасность. Просто перечислим разделы, а пытливый читатель самостоятельно разберется при наличии навыка читать веб-странички на английском языке. Пусть даже и с гугл-транслейтед,~~ как я~~.

Networking

Open Rack

Battery Cabinet

Data Center Electrical

Data Center Mechanical

Data Center

Шесть частей коммутатора

Собственно, подходим к главному, ради чего и затевалась эта статья:

Этот агрегатЪ называется "6-pack" и представляет собой модульный программный коммутатор для "очень высоких нагрузок".

Полное описание (на английском, конечно) можно найти вот здесь, а кое-что о модулях, на которых построен сей девайс - вот здесь.

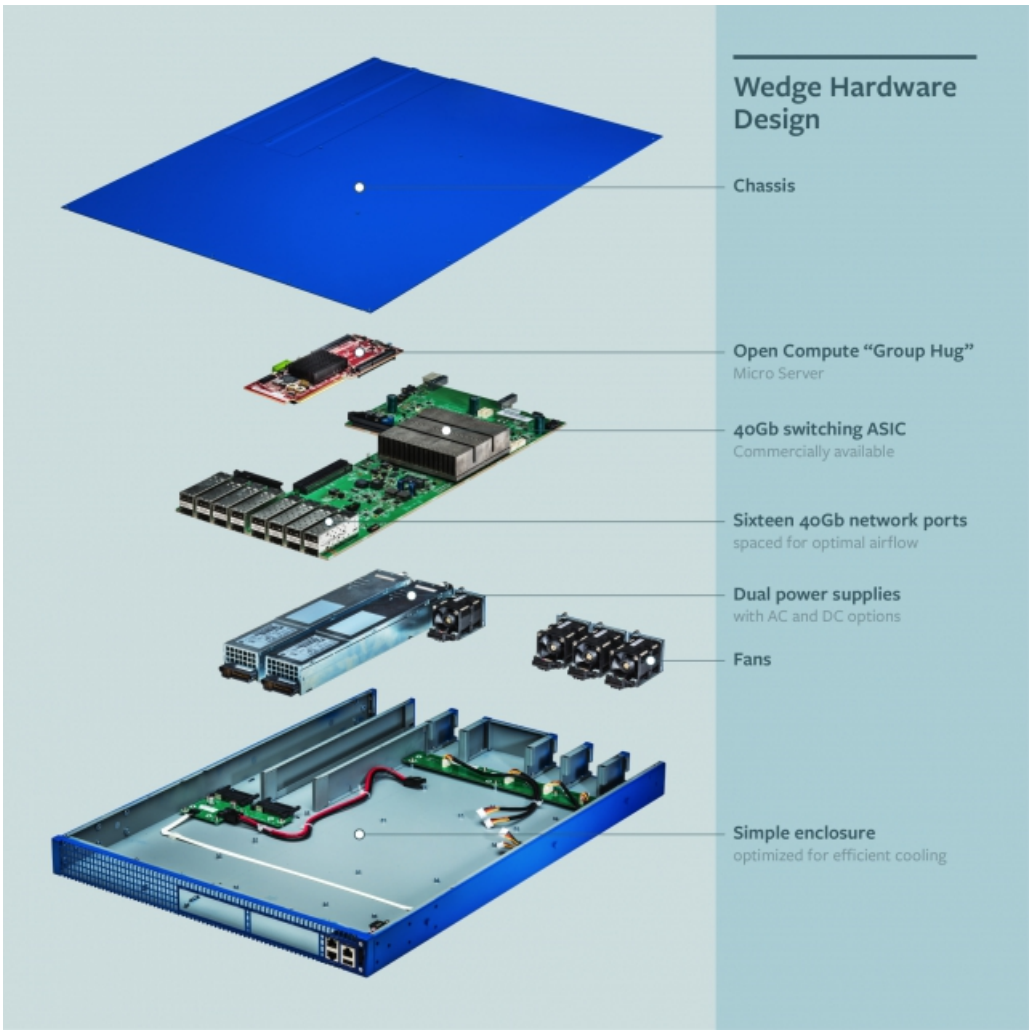
Система работает под управлением специального ПО под Like-Linux-OS под названием FBOSS - Facebook Open Switching System. Собственно FBOSS можно даже скачать в исходных кодах на GitHub.

Через какое-то непродолжительное время Facebook обещали вывести 6-Pack в открытую продажу, а пока аппарат находится в боевом тестировании на собственных задачах FB, о котором сообщили 11 февраля сего года. Ждем-с, мониторим Ebay.

Конструктивно "6-частей" выполнен в отдельном шасси и состоит из четырех модулей Wedge (совершенно стандартных сетевых коммутатора, над которым ФБ работает с 2012 года) и двух модулей "сетевой фабрики", которые представляют собой стандартный же ФБ-сервер.

Каждый модуль независим и имеет свою собственную операционную систему. Оркестрируется все тем же FBOSS.

Wedge (переводится как "клин") в свою очередь тоже модульный:



На шасси "накидывается" по мере необходимости блоки питания и охлаждение, "микросервер", и, разумеется, собственно коммутационные модули (один или два на юнит) с QSFP-посадочными местами. Получается, что систему можно собирать как из конструктора Лего™ в той конфигурации, которая требуется. И наращивать по мере необходимости.

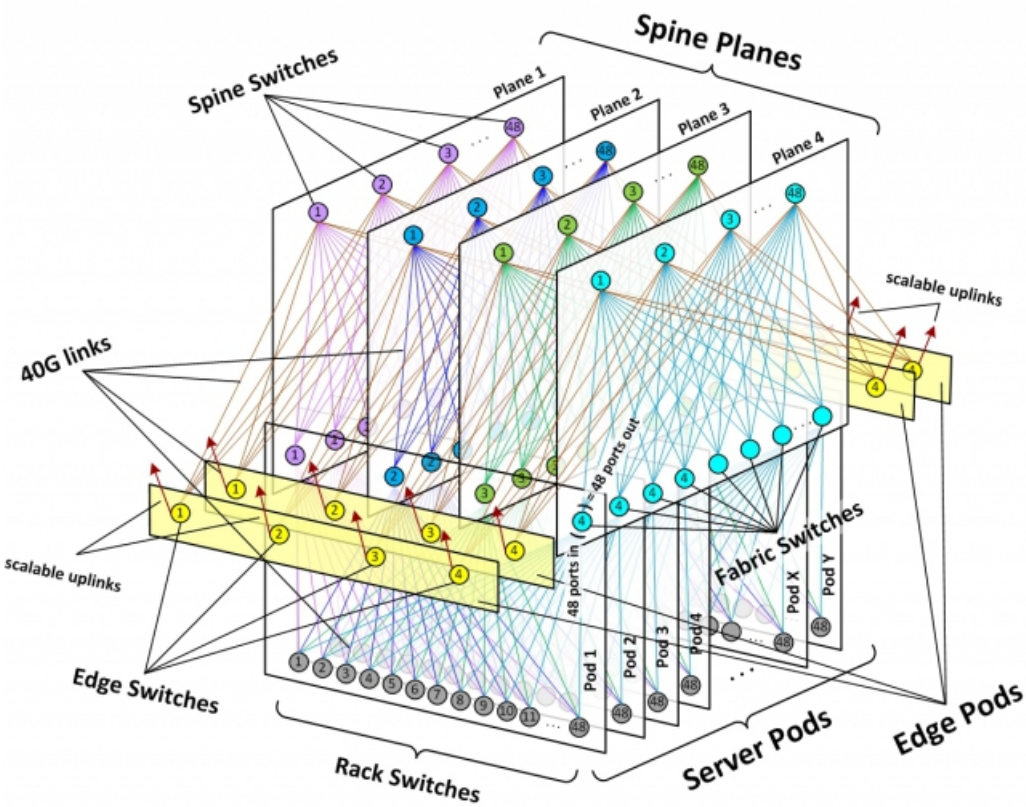
"Коммутационный модуль" выглядит так:

Все вместе это хозяйство представляет собой программный SDN-коммутатор производительностью до 2,5 терабит в секунду.

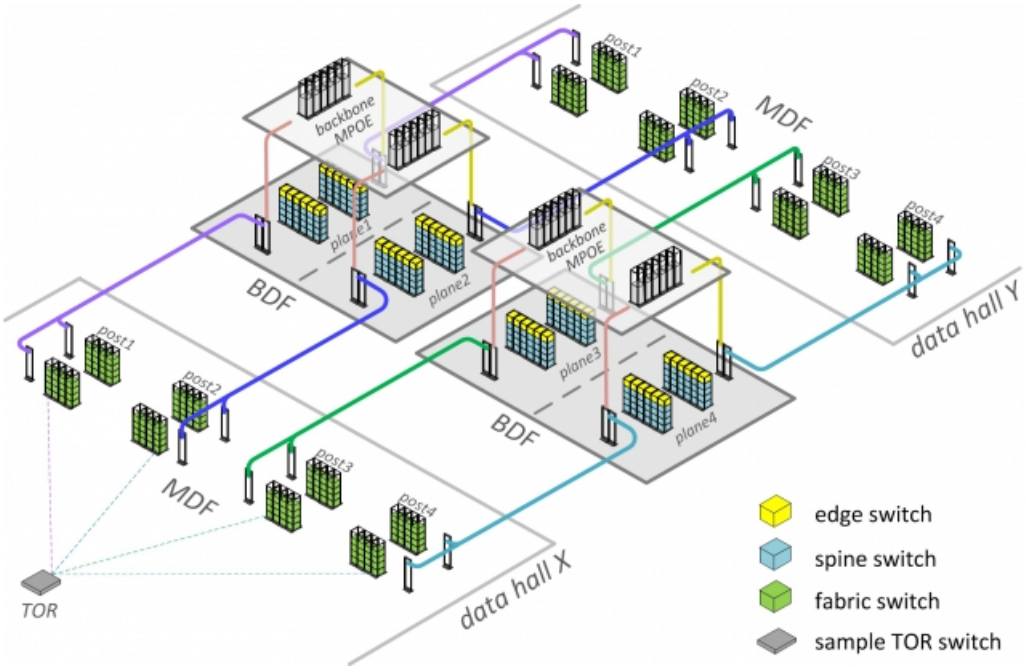
Я не самый продвинутый в техническом плане специалист, посему, углубляться в технические детали не стану - все равно найдется комментатор, который "поймает ошибку" и начнет спорить о деталях. Но поскольку я в большей степени телеком-маркетолог, то вот об этом и поговорим.

Разумеется, данное решение нужно собственно Фейсбуку - будучи самым посещаемым интернет-ресурсом в мире (20 миллиардов уникальных пользователей в месяц!), компании просто необходимо иметь такие молотилки, которые еще и кластеризуются и параллелятся. В "реальной жизни" такое решение, наверное, может применяться где-нибудь в ДЦ или очень больших телекомах.

Собственно ФБ рисует такую картину масштабирования:



Которая, не смотря на кажущуюся сложность, на самом деле, выглядит куда более понятной, если измерения масштаба развернуть на физическую инфраструктуру:



И тут даже ~~министр связи~~ такой гуманитарий, как я, сообразит, в чем преимущество программно-конфигурируемых сетей: абсолютная масштабируемость и гибкость. Мощность сети мы можем повышать практически бесконечно, добавляя новые, в том числе и перспективные, модули. А гибкость обеспечивается функционалом SDN.

Кстати, во время демонстрации 6-pack присутствовали функционеры Cisco, которые "высоко оценили решение". И тут же заявили, что "компания имеет собственные SDN-решения - Nexus 9000, например". Но понятно, что у Facebook имеются свои уникальные требования, и Cisco не видит рынка для подобных решений вне этого.

Однако, к инициативе OCP (Open Computing Project) все же присоединились и "увлеченно работают в команде". Впрочем, как и Juniper и HP. Однако, ~~вторая~~ третья по масштабам интернет-компания Yahoo! все же приняли решение развиваться на идеях OCP, закрыв контракты с Cisco. Очень приглядывается к решению и Google, который вкладывает немалые средства в собственные разработки и вообще, "корпорация добра" очень любит Open Source. Прочие крупные интернет-проекты находятся в глубокой задумчивости, но когда дело дойдет до подсчетов бюджетов развития на сетевую инфраструктуру, результат можно спрогнозировать.

Другими словами, вот прямо сейчас мы наблюдаем очередной тектонический сдвиг в поведении крупных интернет-компаний и провайдеров услуг по всему миру. И кажущееся спокойствие Cisco и прочих крупных телеком-вендоров вполне может хлопнуть, когда дело дойдет до массового исхода на SDN, которая очевидно является угрозой для высокой рентабельности телеком-вендоров. А рентабельность от продажи сетевого оборудования, как правило, составляет около 60%. И вот-вот эти деньги перейдут к производителям софта.

А в это время, где-то в России, режиссер перманентно убыточного кино получал деньги на сеть быстрого питания, правоохранительные органы закрывали спектакли, а один грустный разработчик глубоко перерабатывал ядро Линукс с целью построить мультипротокольный маршрутизатор специального назначения...