Facebook намерен использовать сетевое оборудование собственной разработки

Компания Facebook, исходя из своих стратегических целей и планируя иметь в будущем возможность гибкого масштабирования своего бизнеса, пришла к необходимости разработки отдельных устройств для своей сетевой инфраструктуры. Для этого, около трёх лет назад, был запущен Open Compute Project. Через два года были получены первые результаты, которые отличались в лучшую сторону от конкурирующих решений по энергоэффективности и себестоимости.

Поддержкой инициатив компаний, желающих через повышение энергоэффективности сократить затраты на электричество, в США занимаются Department of Energy (DOE) и Environmental Protection Agency (EPA).

"According to the U.S. Environmental Protection Agency (EPA), data centers consumed between 1.7 percent and 2.2 percent of the total electricity used in the United States in 2010. The EPA has reported that U.S. data center electricity use nearly doubled between 2000 and 2005, and it increased by approximately 36 percent between 2005 and 2010. The EPA further claims that data center energy consumption will continue to grow at a rate of more than 9 percent per year through 2020...", — сообщал в прошлом году The Hill.

Эти государственные структуры обеспечивают интересующихся информационной поддержкой, собирая и публикуя сведения о лучших проектах, организовывают встречи, коллективные обсуждения лучших реализаций. Одним из параметров, по которым оценивается степень технологического совершенства ЦОДа — это PUE (power usage effectiveness). Environmental Protection Agency (EPA), готовивший в далёком 2007 году доклад* для конгресса США, ЦОДы с PUE меньше 1,5 относил к категории "лучшие проекты".

Power usage effectiveness (PUE) is a metric developed by The Green Grid that measures data center efficiency. It"s a ratio of the total power consumed by a data center to the energy delivered to the servers. The ideal PUE is 1.0, meaning 100% of the data center energy powers the IT equipment. Best practice PUE for the IT sector is 1.5.

От инженеров Facebook ожидали таких решений, которые имеют возможность масштабирования и при этом позволяют экономить в период эксплуатации за счёт большей энергоэффективности. Не менее важной было решить и другую задачу — оптимизировать себестоимость оборудования собственной разработки с учётом её влияние на конечную стоимость решений, которые будут в дальнейшем наращиваться.

"Если хочешь что-то сделать хорошо, сделай это сам"
Центр обработки данных (ЦОД)

Первым "продуктом", разработанным специалистами Facebook, стал data center (центр обработки данных, ЦОД). Проектирование с "нуля" позволило инженерам изменить всё, что они посчитали необходимым: например, чтобы уменьшить потери, в системе электропитания используется напряжение 480 вольт (3 фазы). Кроме этого, предложена своя система размещения оборудования (Open Rack), серверы собственной разработки, повторное использование воздуха, в том числе для обогрева офисных помещений, отказ от центрального ИБП — всё это позволило ЦОДу в Prineville ("Prineville Data Center"), при той же производительности, тратить энергии на 38% меньше. За счёт этого у компании Facebook появилась возможность в Prineville снизить свои затраты на 24%.

Описание/слайды (плюс более подробные pdf) отдельных элементов инфраструктуры ЦОД — "8 Data Center Lessons From Facebook".

Facebook сумел в Prineville Data Center обеспечить PUE на уровне 1,07-1,08. Ниже есть иллюстрации, где видно, что этот показатель не является постоянным и может меняться в некоторых пределах.

High Level Metric

PUEs: Reported & Calculated	PUE
EPA Energy Star Average	1.91
Intel Jones Farm, Hillsboro	1.41
T-Systems & Intel DC2020 Test Lab, Munich	1.24
Google	1.16
Leibniz Supercomputing Centre (LRZ)	1.15
National Center for Atmospheric Research (NCAR)	1.10
Yahoo, Lockport	1.08
Facebook, Prineville	1.07
National Renewable Energy Laboratory (NREL)	1.06

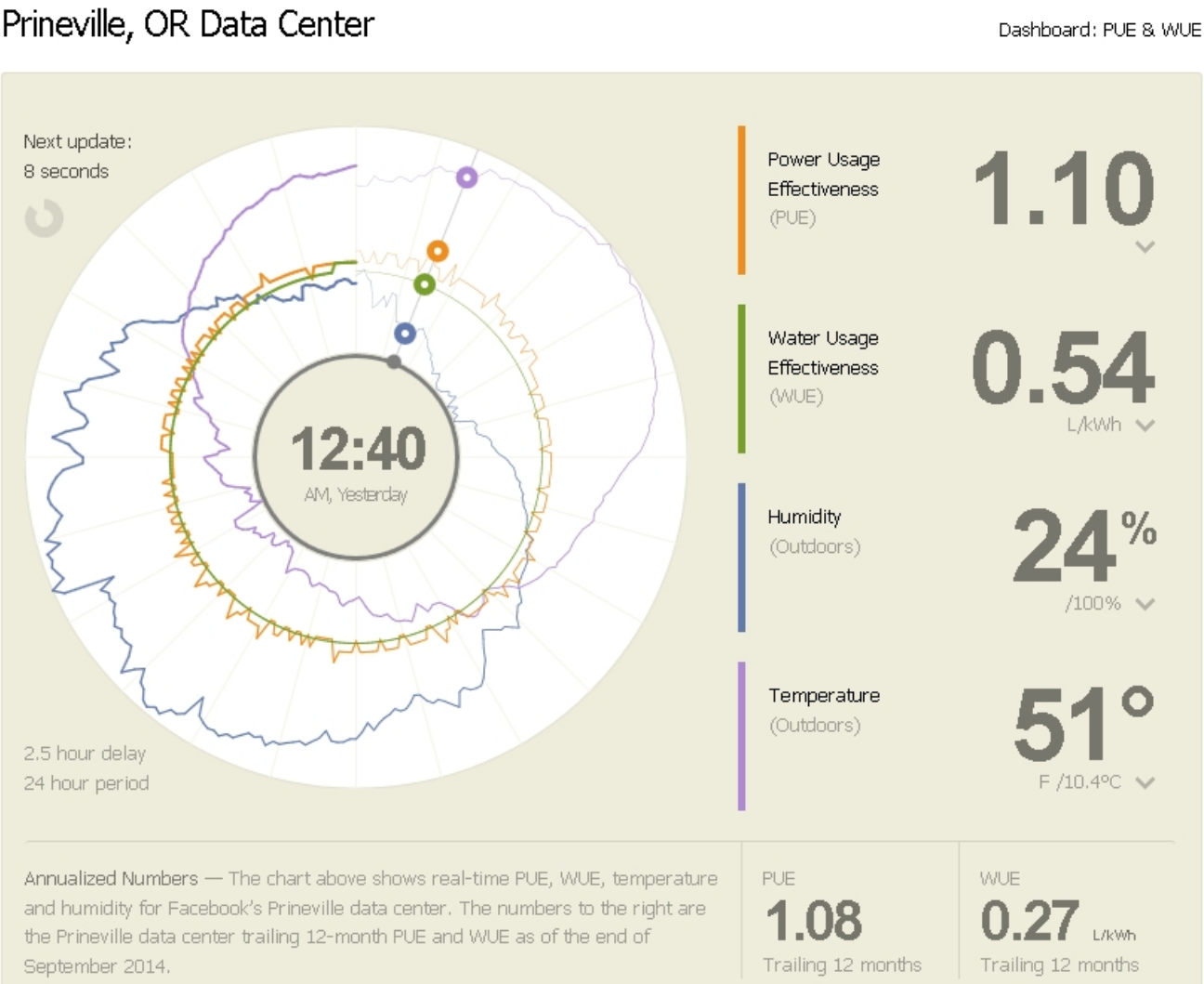
Slide Courtesy Mike Patterson, Intel

Better Buildings

U.S. DEPARTMENT OF ENERGY

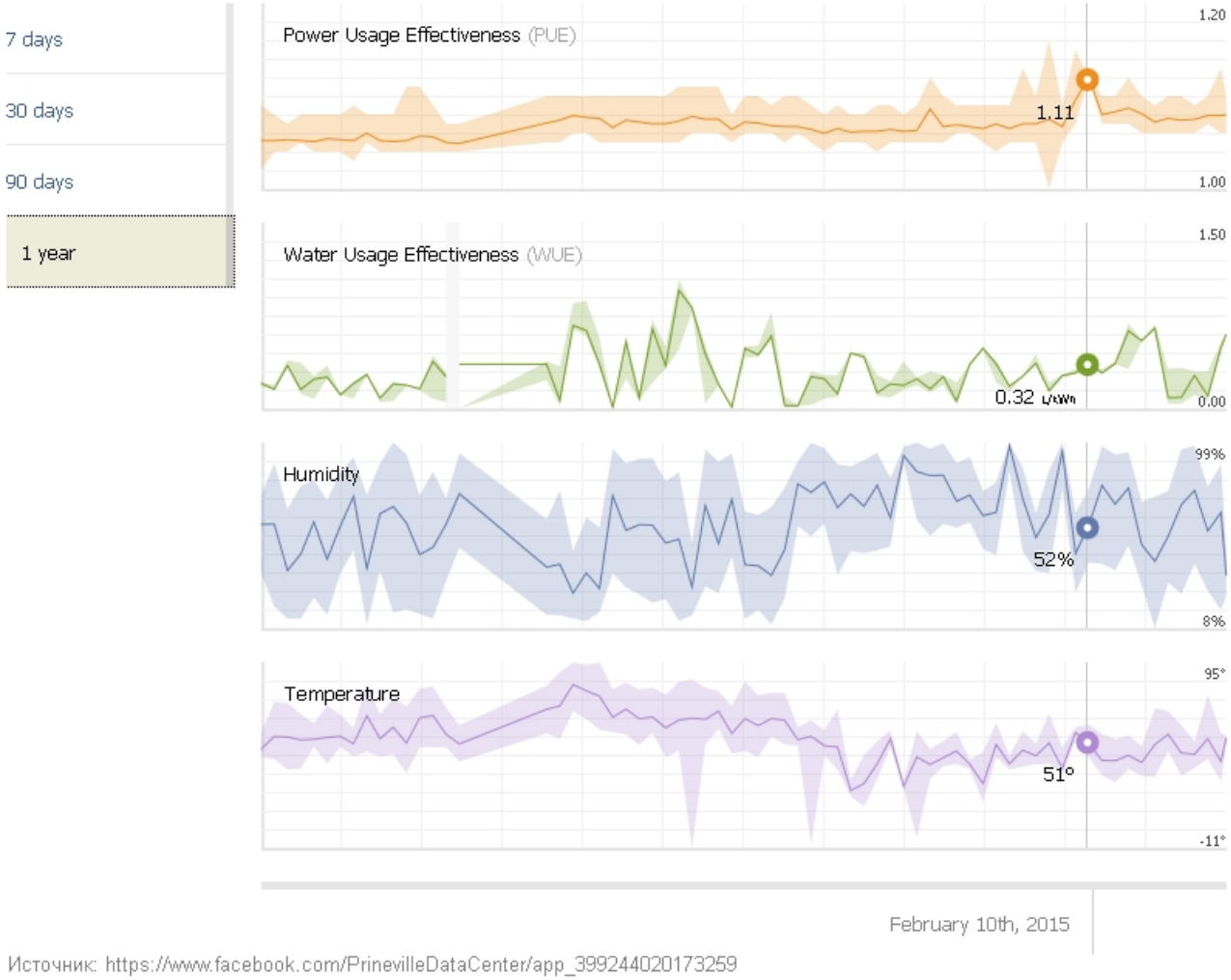
Значения PUE (power usage effectiveness) в ЦОДах разных компаний и организаций

Нашёлся любопытный источник и свежей, и исторической информации о ЦОДе в Prineville, там можно посмотреть, какими были и в каких пределах изменялись такие параметры как PUE, WUE (water usage effectiveness), Humidity (Outdoors) и Temperature (Outdoors). Можно посмотреть подробные изменения этих параметров в течение последних 24 часов, ниже скриншоты максимальных значений одного из параметров, зафиксированных в течение суток 3 апреля 2015 года.

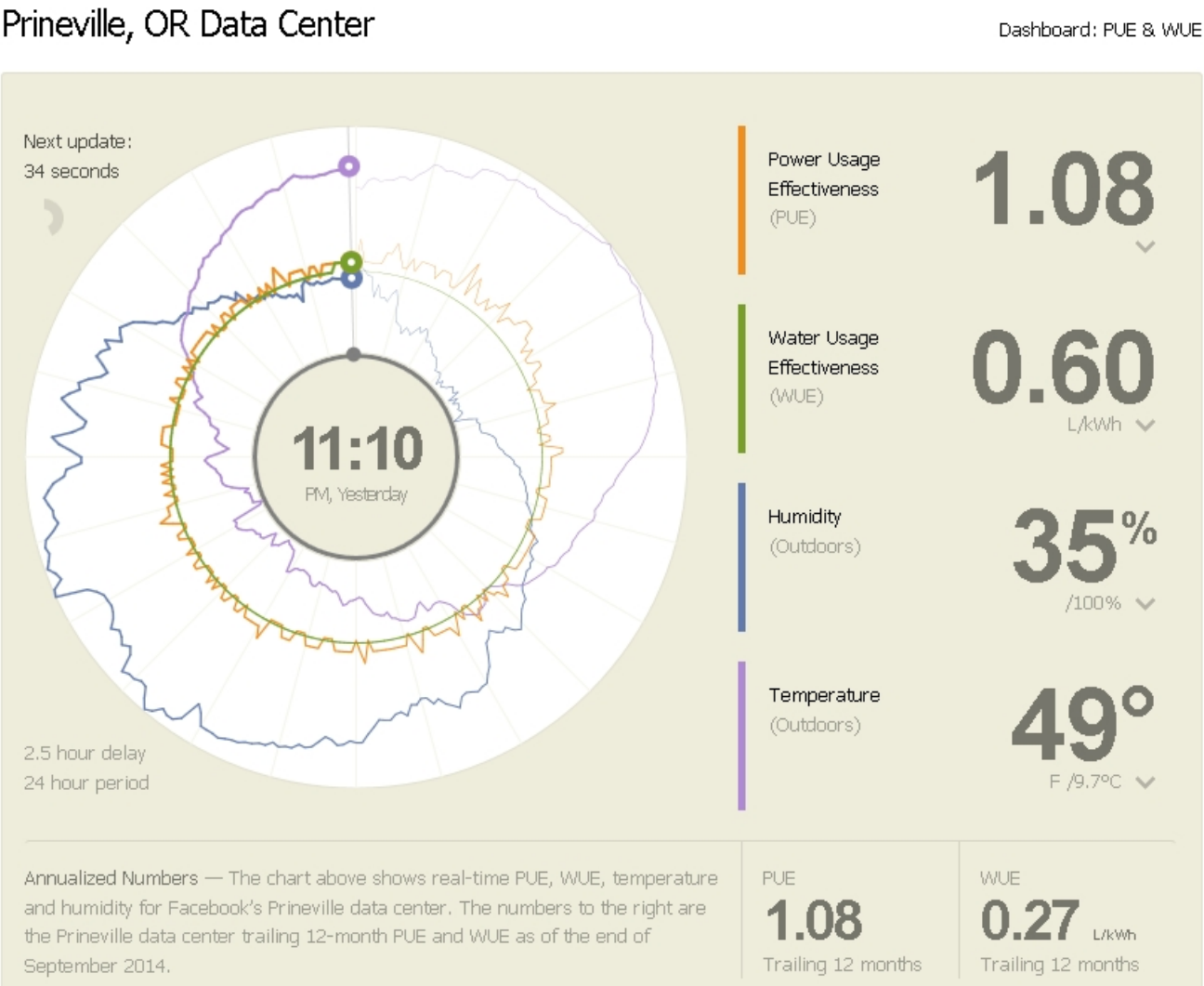


Худшее значение PUE в течение суток

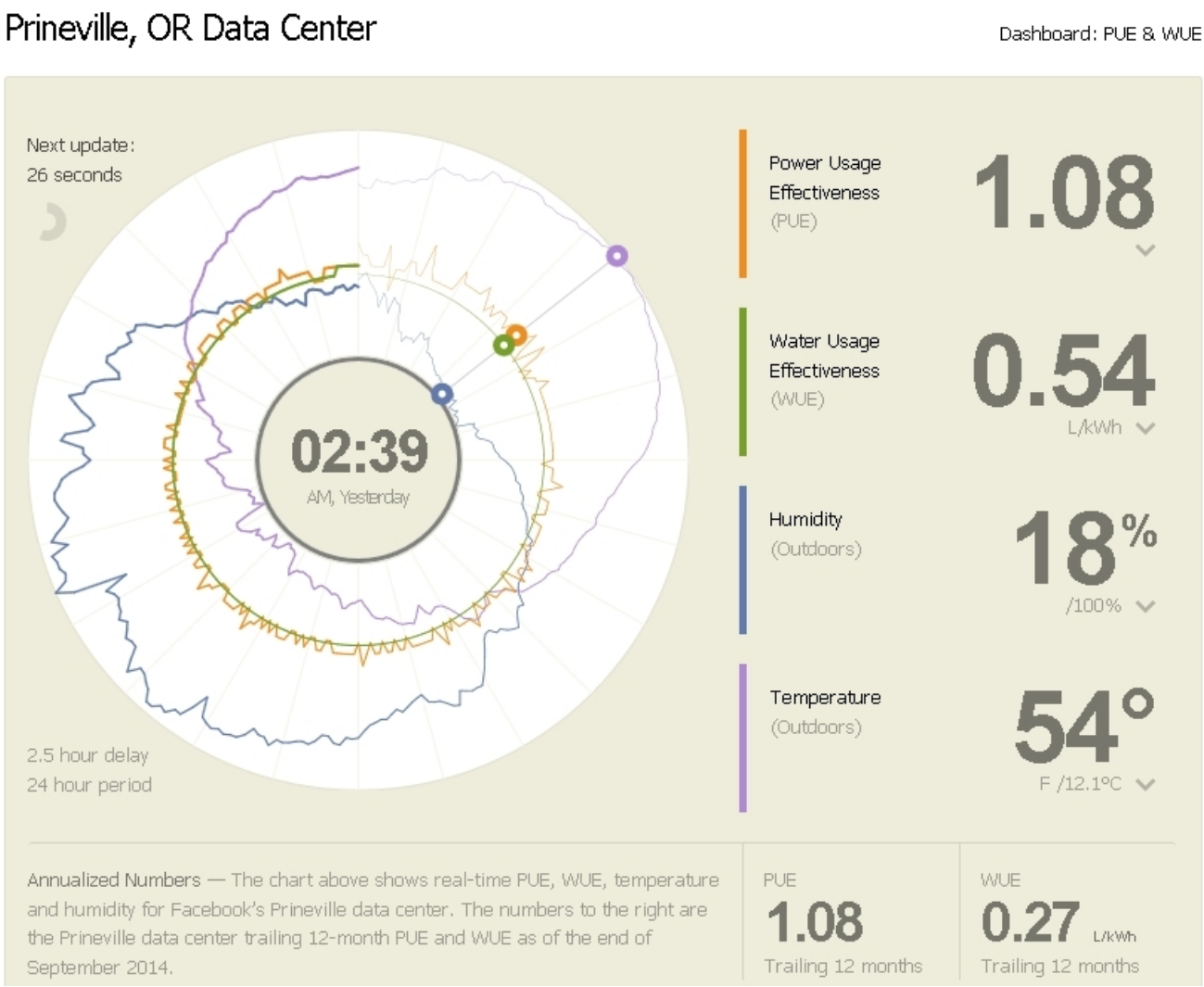
Есть данные и за другой временной промежуток (можно выбрать неделю, месяц, последние три месяца или год). И тогда можно обнаружить ещё более худшее значение PUE.



В феврале 2015 года в Prineville Data Center значение PUE было даже 1,11



Худшее значение WUE (water usage effectiveness) в течение суток



Максимальная температура в течение суток

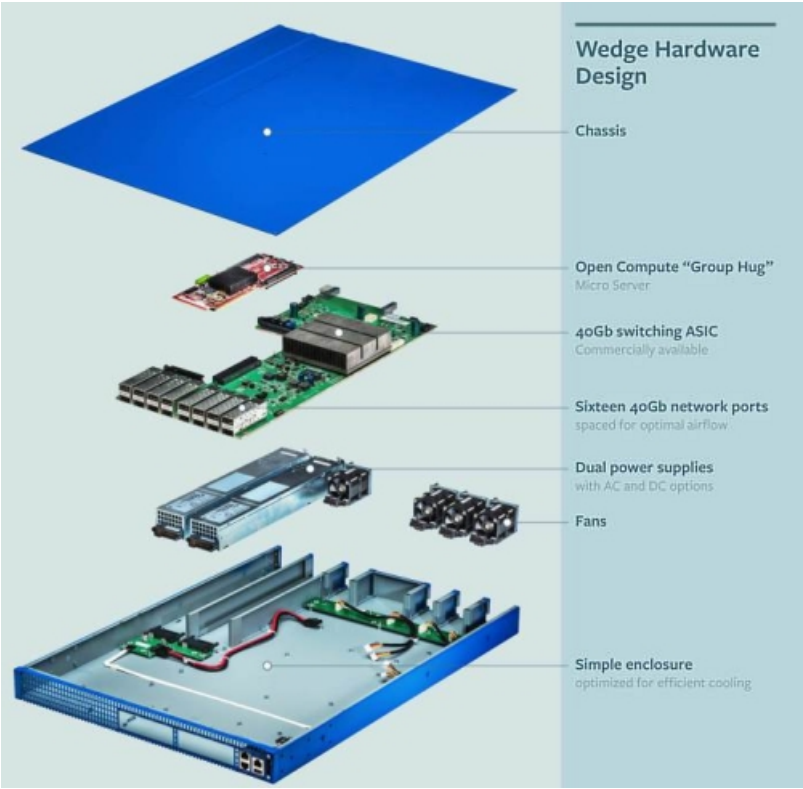
Есть ли подобные "живые" и публичные источники о работе какого-либо ЦОДа в России?

"Если хочешь что-то сделать хорошо, сделай это сам"
Маршрутизатор

Кроме перепроектирования ЦОДов в Facebook занимались и конструкциями других устройств (["How Facebook aims to reinvent hardware"]). Летом прошлого года представитель компании в интервью изданию Gigaom рассказал о новом подходе к тому, каким должен быть сетевой маршрутизатор. Была информация и в Network World (["Facebook-Led Open Compute Project Tackles Network Switches"]).

Достижение Facebook в 2015 году — "свой" маршрутизатор

Предложенная модульная конструкция сетевого устройства, в котором ещё и разделены "железо" и программная "начинка" (об ONIE информация ниже), была названа Wedge. Конструктивное решение дополнено операционной системой FBOSS на базе Linux.



Wedge

Основой для построения и развития сетевых ресурсов в Facebook станет 6-уровневая платформа, которая создана в соответствии с идеями Wedge.

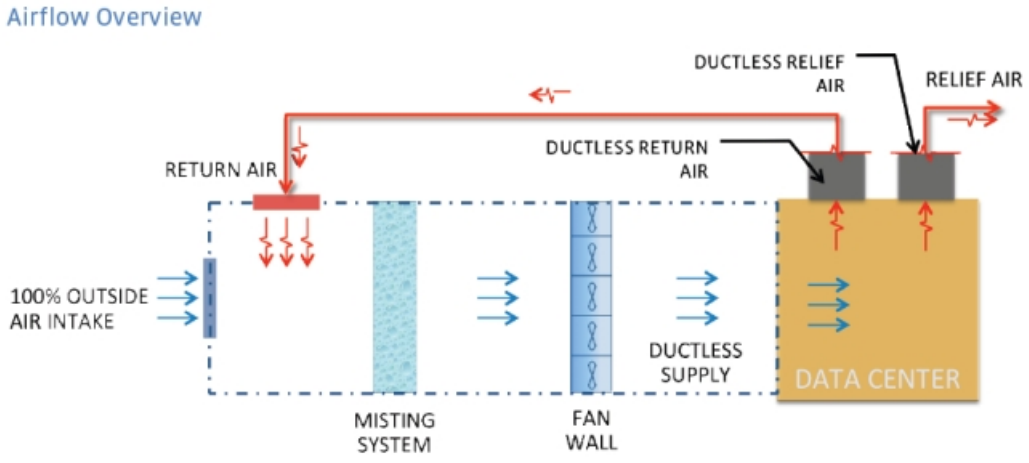
В этой платформе есть 12 независимых коммутирующих элементов, каждый из которых может управлять трафиком данных 1,28 Тб/с. Есть две конфигурации: для первой реализовали на фронтальной стороне 16x40 GE, а на тыльной — возможность подключения канала до 640 Гб/с,

вторая конфигурация отличается от первой отсутствием 16 фронтальных коннекторов, а сзади есть — 640 Гб/с.

Таким образом, за относительно небольшой период времени инженеры Facebook собственными силами переработали ключевые элементы, используемые в инфраструктуре компании: серверы, маршрутизаторы, конструктив (Open Rack) для размещения сетевого, вычислительного оборудования и систем хранения данных и реализовали своё комплексное решение на базе перечисленных решений и подходов в виде масштабируемого энергоэффективного ЦОД.

Результаты Facebook — следствие нового подхода

Проект не просто так называется Open Compute Project: на сайте доступны спецификации, а разработчики открыты для предложений, которые помогут им что-то улучшить. Там есть описание системы хранения (Storage); описание конструкций для размещения оборудования (Open Rack) с расчётом на определённую сейсмическую устойчивость; дизайн-проект ЦОДа (Data Center Design project). Понятно, что разработки и рекомендации на Open Compute Project подготовлены с учётом американских стандартов, поэтому есть смысл обращать внимание только на интересные идеи и удачные решения.



Из описания решений Facebook для ЦОДа

Непосредственно в разделе** Open Compute Networking Project нет спецификаций или подобных материалов, но есть контакты руководителя этого подпроекта, есть предложение подписаться на рассылку, есть IRC Channel и Webchat Link for IRC.

Угроза гегемонии поставщиков проприетарных решений

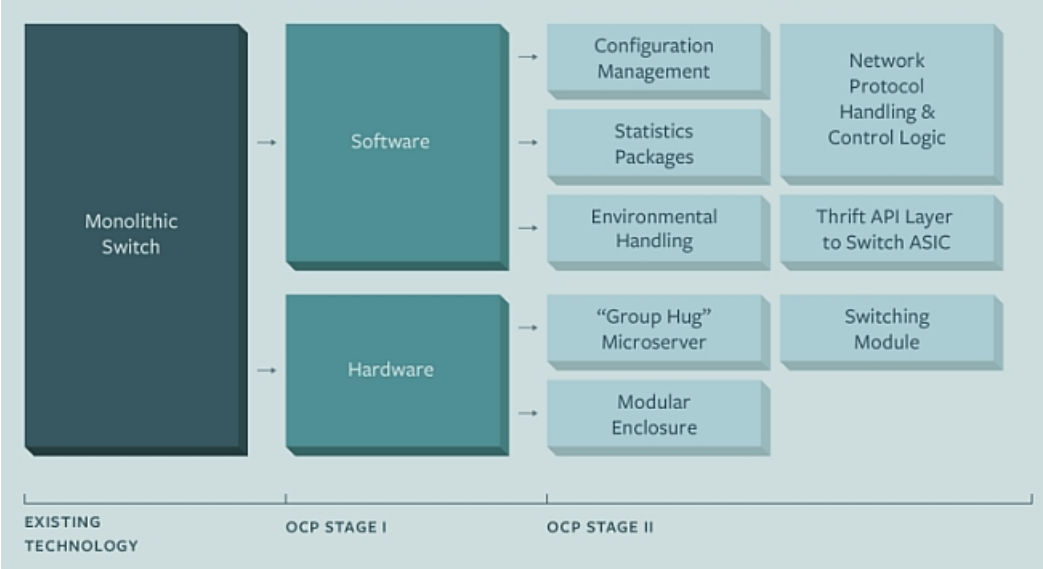
Поиск своих вариантов оптимизации, эффективных решений для нынешних и будущих задач привели к тому, что специалисты Facebook минимизировали использование проприетарных решений, а весной 2011 года объявили о запуске Open Compute. В 2013 году этот проект охватил новые направления, связанные с сетями и сетевым оборудованием.

В списке первых участников Open Compute Networking Project (OCP) нет таких известных производителей разнообразного сетевого оборудования, как Cisco Systems и Hewlett-Packard*. Понятно, что идея OCP — отказ от жёсткой "связки" фирменного "железа" и ПО одного и того же производителя — подрывает их бизнес.**

Интерес Facebook и других крупных потребителей к решениям Open Compute Networking Project подталкивает тот же Cisco к тому, что ценовые предложения для этих покупателей (а это хорошие, растущие покупатели) должны учитывать наличие альтернативных вариантов. Одновременно Cisco критикует Open Compute Networking Project.

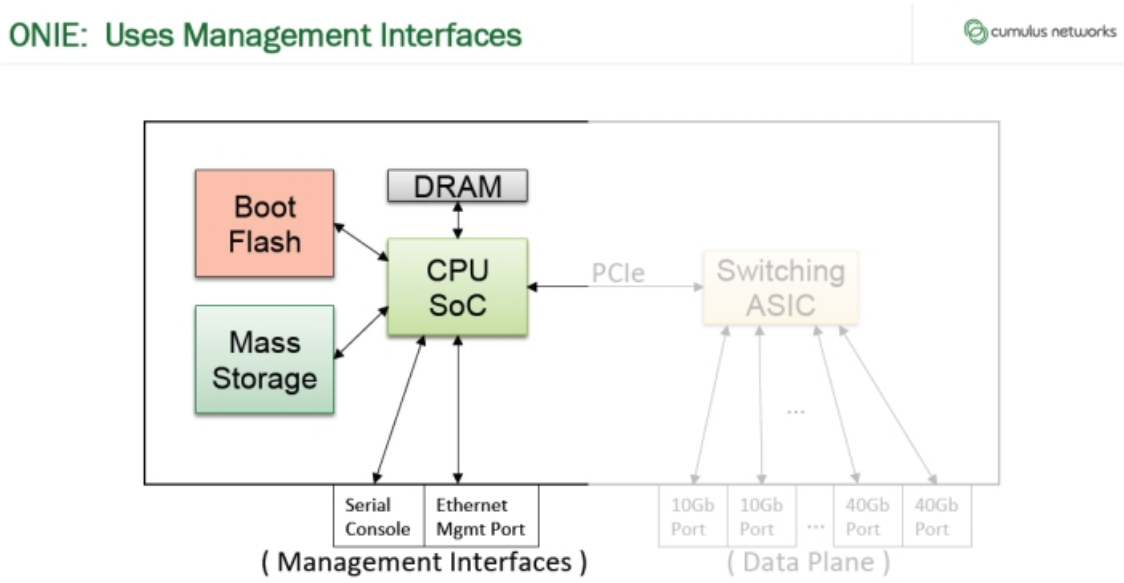
Позже появлялась информация, что "Cisco and HP are starting to sell their computing hardware without their proprietary software". Например, одно из сообщений по этому поводу было в марте этого года — см. "Cisco might ONIE up".

Информация о том, чем в настоящее время заняты специалисты есть в Wiki, они работают над программным обеспечением для Open Network Install Environment (ONIE) — направления, которое поддерживают компании Accton, Agema, Big Switch Networks, Broadcom, Cumulus Networks, Dell, Penguin Computing, Quanta и другие.

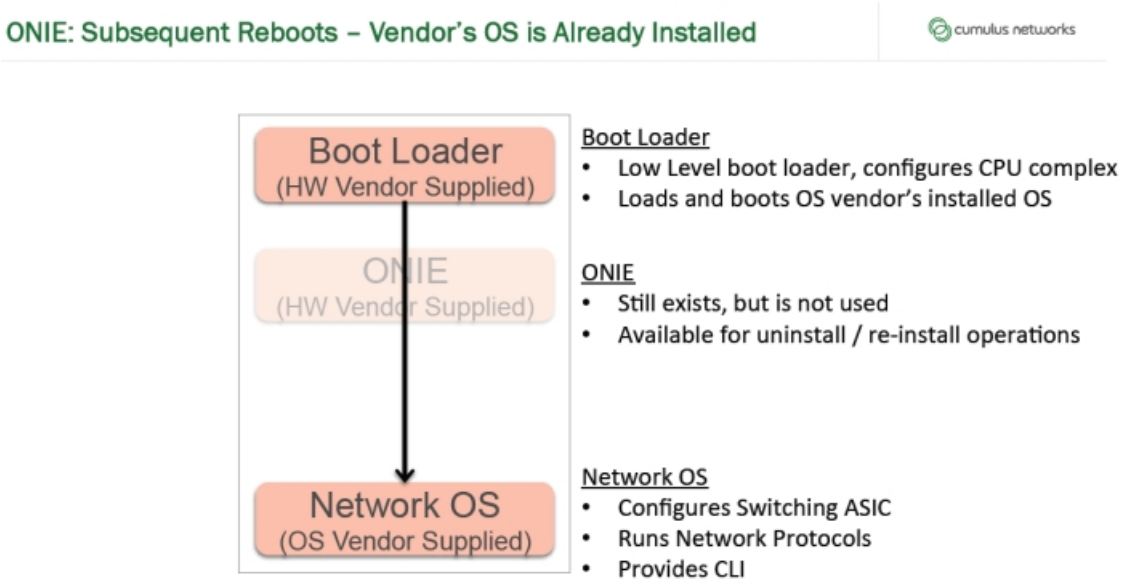


Новый тренд в разработке маршрутизаторов

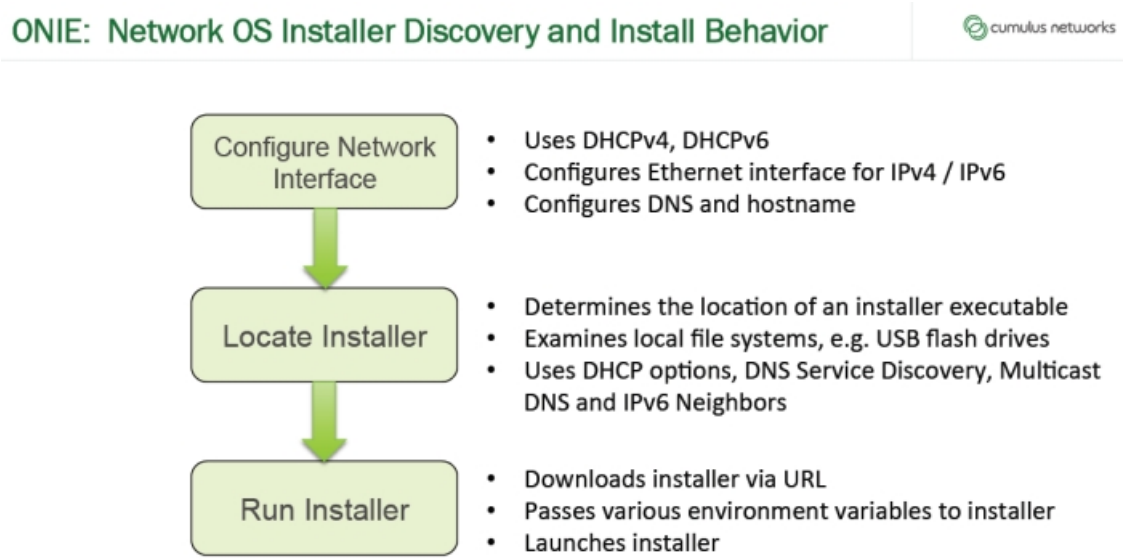
Основная идея — дать провайдеру универсальное "железо" и простейший загрузчик, которые позволят при желании "залить" другую сетевую операционную систему (network OS) и гибко, под себя, всё настроить.



Разделены управляющая часть и управляемая (коммутационная)

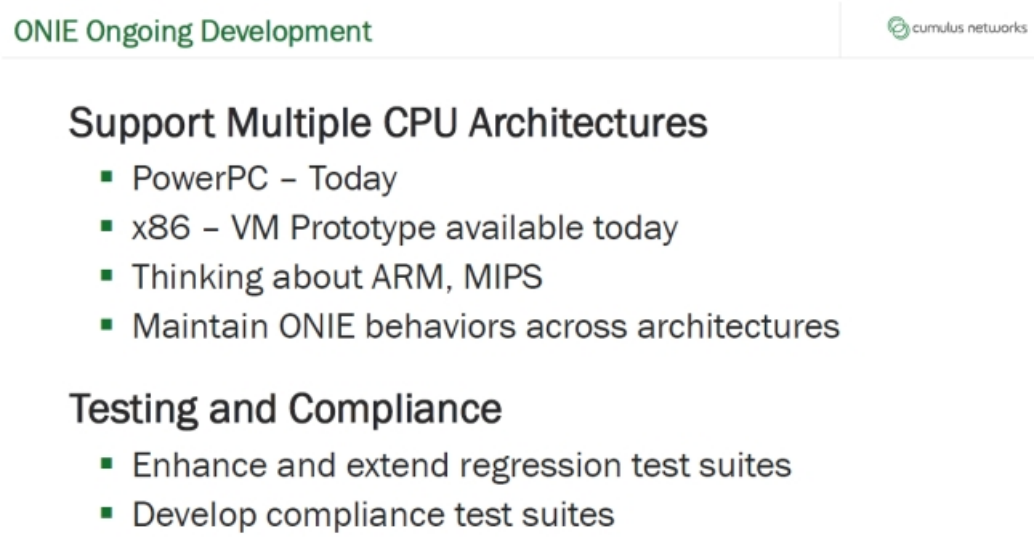


Можно загрузить другую Network OS (Operating System Vendors: Big Switch Networks, Cumulus Networks, Mellanox Technologies)



Другие возможности, реализованные в ONIE

Участники Open Compute Networking Project намерены освоить "железо", работающее на самых разных процессорах



Для "обкатки" и работы ONIE используют "виртуализацию"

ONIE Ongoing Development

cumulus networks

New Features

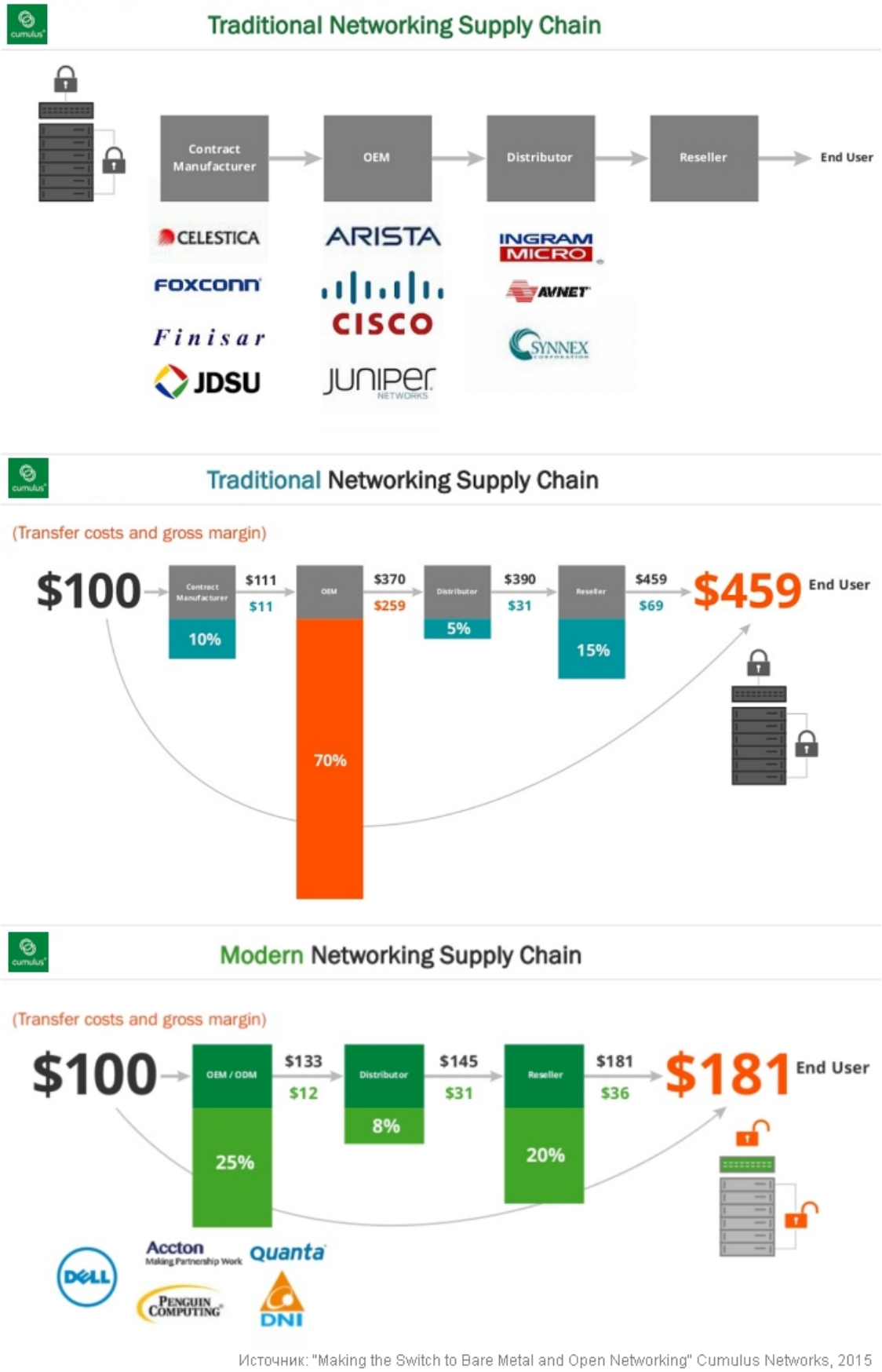
- DHCPv6
- DNS Service Discovery / Multicast DNS discovery
- HW Vendor diagnostic

Releases

- Quarterly release cadence
- Ongoing maintenance
- Enhancements

ONIE не только воспроизводит, повторяет имеющиеся возможности, но и совершенствует и добавляет новые

Такой подход позволяет быстрее подстраиваться под требования провайдеров: пусть производители "железа" занимаются свои делом, поставляя "bare metal network switches", а другие компании, сотрудничая и соперничая, займутся совершенствованием и развитием программной "начинки". Преимущество этого подхода не только в возможности вносить изменения в ПО, есть и финансовый выигрыш.



Источник: "Making the Switch to Bare Metal and Open Networking" Cumulus Networks, 2015

Новый подход позволяет заметно уменьшить финальную стоимость оборудования, за счёт изъятия того "навара", который получали поставщики проприетерных решений

Изменения в отношениях между поставщиками проприетарного оборудования и крупными покупателями-потребителями (Amazon, Google, Facebook) отметили в Gartner. Там считают, что инициативы, подобные Open Compute Networking Project, возникли не случайно и будут развиваться**** в ближайшее время: цель понятна — организовать бизнеса наиболее эффективно с учётом того, что для глобального охвата потребуется значительное масштабирование инфраструктуры. Достижения на этом пути не скрываются, проект открытый, поэтому информация доступна и другим, более мелким компаниям (см. "Overview" и "Code repository for ONIE" на GitHub).

---
* "Report to Congress on Server and Data Center Energy Efficiency" — ссылка на pdf.

** Open Compute Networking Project — www.opencompute.org/projects/networking/

*** Hewlett-Packard (HP) присоединилась к проекту весной 2012 года. Новые компании-члены, кроме HP, это AMD, Fidelity, Quanta, Tencent, Salesforce.com, VMware, Canonical, DDN, Vantage, ZT Systems, Avnet, Alibaba, Supermicro и Cloudscaling. Чем занялись новые участники — см. здесь.

*** "Web-scale IT is a pattern of global-class computing that delivers the capabilities of large cloud service providers within an enterprise IT setting by rethinking positions across several dimensions." - см. "Gartner Says By 2017 Web-Scale IT Will Be an Architectural Approach Found Operating in 50 Percent of Global Enterprises".