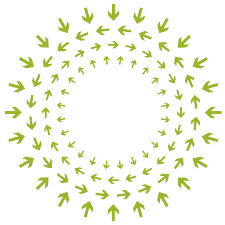


Open Compute Project

オープンコンピュートプロジェクト　ジャパン

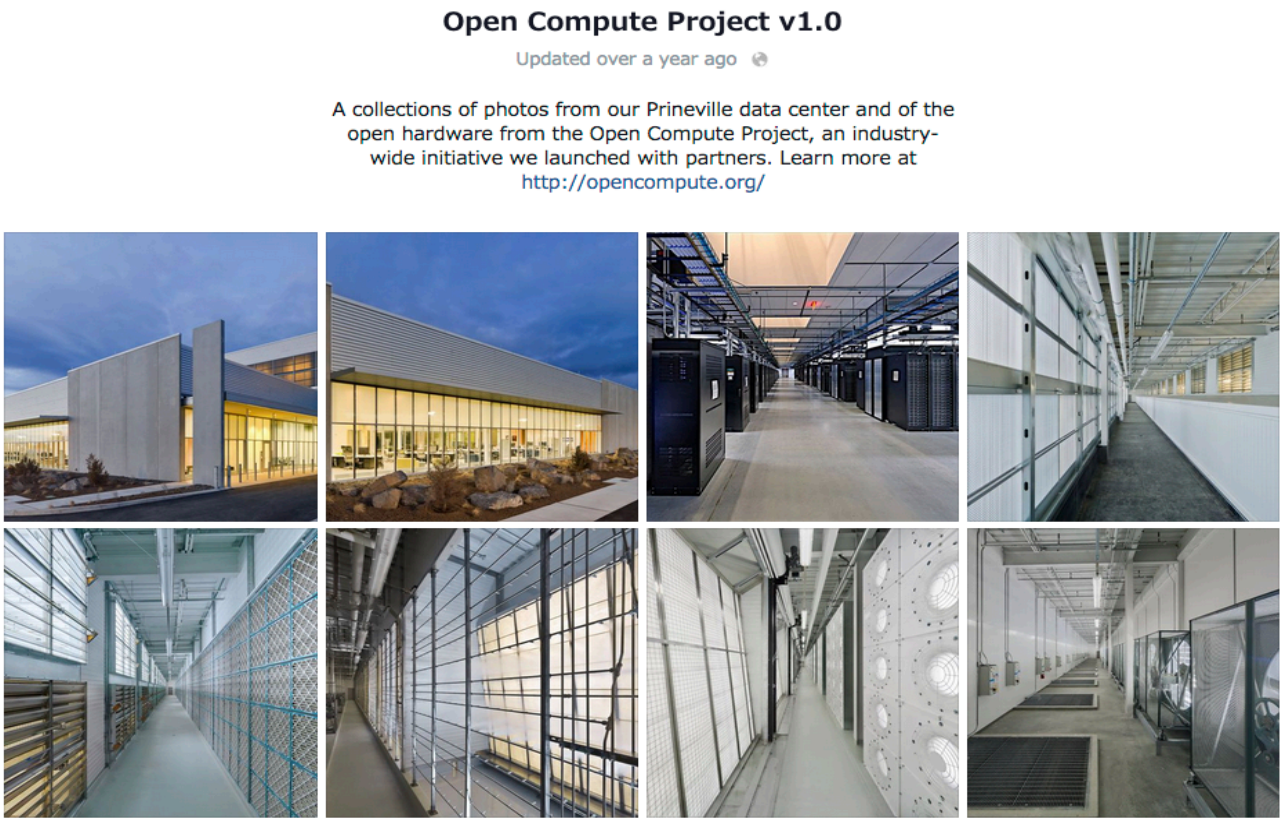
藤田　龍太郎
ryu@netone.co.jp

OCPとは

OCPの目的
OCPプロダクト
OCPの普及
まとめ
OCP-J

2011年

Facebookオレゴンのデータセンタの写真がFBで公開
その仕様書がOCPサイトで公開される

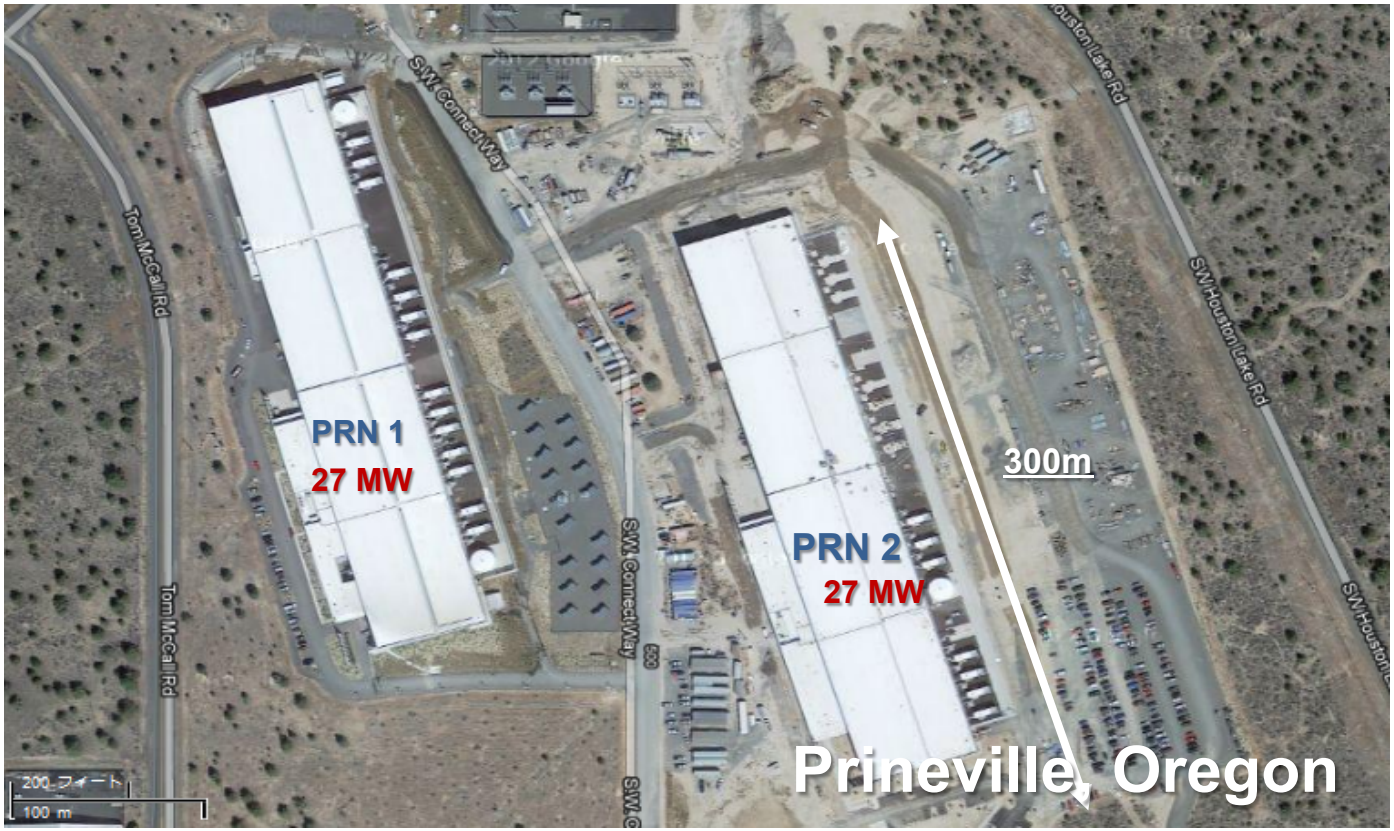


出典：https://www.facebook.com/media/set/?set=a.10150151683427694.289087.193287527693

Facebook の Oregon DC

300m x 60m の床面積と、27MWの電力密度

1棟に10万台のServer



OCP のサーバーとラック

フタもなければ、
フロントパネルもない

ネジを１本も使わずに
キッティング

出典：http://wp.me/pwo1E-2Ku

OCP のサーバーとラック

スイッチとコネクタを
すべてフロントに

ホット・アイルには
ファンが並ぶだけ

出典：http://wp.me/pwo1E-2Ku

FB が 20分間に処理するデータ

- Shared links: 1,000,000
- Wall Posts: 1,587,000
- Status updates: 1,851,000
- Photos uploaded: 2,716,000
- Comments: 10,208,000
- Message: 4,632,000

出典：http://highscalability.com/blog/2010/12/31/facebook-in-20-minutes-27m-photos-102m-comments-46m-messages.html

Open Compute Project

2011年4月にFacebook社が提唱

- オレゴン州の自社DCを公開し、同DCで採用しているエネルギー利用効率の高いサーバーとDCの仕様やベストプラクティスを業界全体で共有するための取り組み
- 最も効率の良いサーバー/ストレージ/データセンターなどのハードウェアを設計提供していくためのエンジニアのコミュニティ
- アイデアやスペックなどの知的財産を共有
- 更なる「Open Compute Project」の加速と個人および組織との間で、知的財産を共有するための構造としてOpen Compute Project Foundationを設立

主な参加企業

Facebook
AMD
Dell
HP
Intel
Goldman Sachs
ARM Holdings
Broadcom

Quanta
wistron
Gigabyte

Vmware
Microsoft
Apple

Cisco
juniper
Schneider Electric

OCP 採用企業

GoldmanSacks

Riot Games

Bloomberg

Facebook

Orange

Fiderity

Microsoft

Rackspace

OCPとは

OCPの目的

OCPプロダクト

OCPの普及

まとめ

OCP-J

price/performance
and
performance/watt

Facebook のコスト削減

Facebook CEO Mark Zuckerberg, at left, discusses the company's infrastructure with Tim O'Reilly of O'Reilly Media yesterday at the Open Compute Summit in San Jose, Calif. (Photo: Colleen Miller)

直近の 3年間で$1.2 Billion 以上のコストを削減

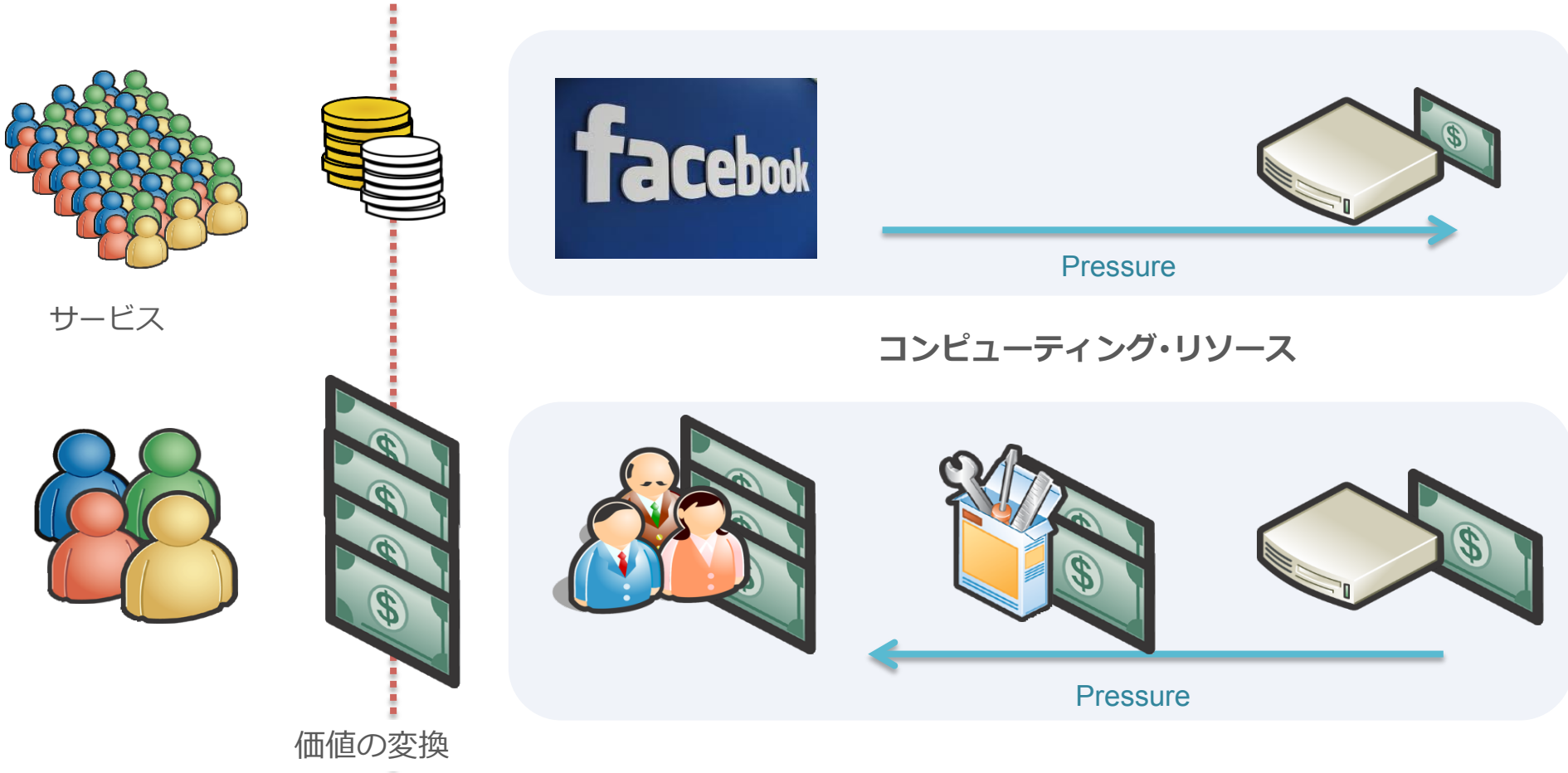
・データセンターやサーバーを効率化する Open Compute のデザインを使用

・デザイン／アーキテクチャ／プロセスにおける小さな改善の集大成

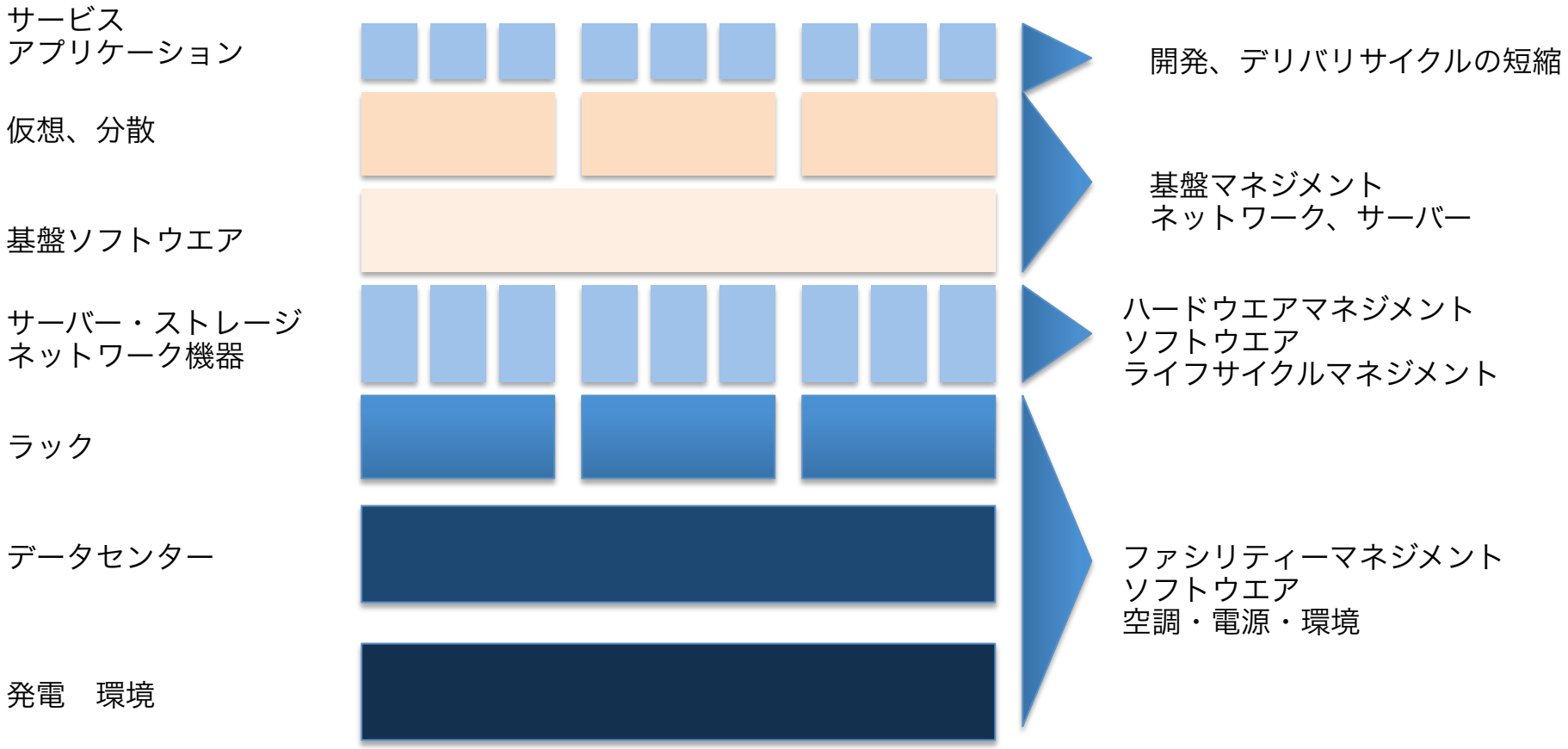
・何十万台ものサーバー群
※20-60万台/DC

@2014 Open Compute サミット

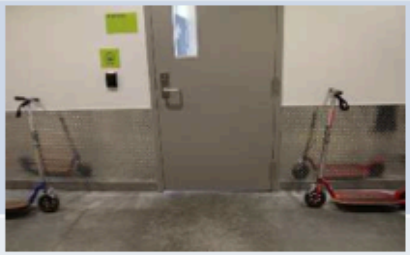
OCP エコシステム と　従来からのサプライチェーン



サービス中心の基盤



Facebookのデータセンター

場所	稼働時期	概要
プラインヴィル オレゴン州	2011年4月	 OCPのベース 効率アップ38%、 コスト減24% PUE1.07
フォレストシティ NC州	2012年4月	外気冷却　高温 高湿度対応 
ルレオ スウェーデン	2013年6月	 水力発電　100%再生可能 PUE1.07
アルトゥーナ アイオワ州	2015年	100%風力発電 

自社サービスインフラを開発環境として提供

Disaggregate

モジュール化

ベンダ、ハードウエア種類、ラック単位で構成されていた要素技術を分解、構成部品単位にモジュール化

接続ポイントを高速化

スケールアウト、スケールアップ

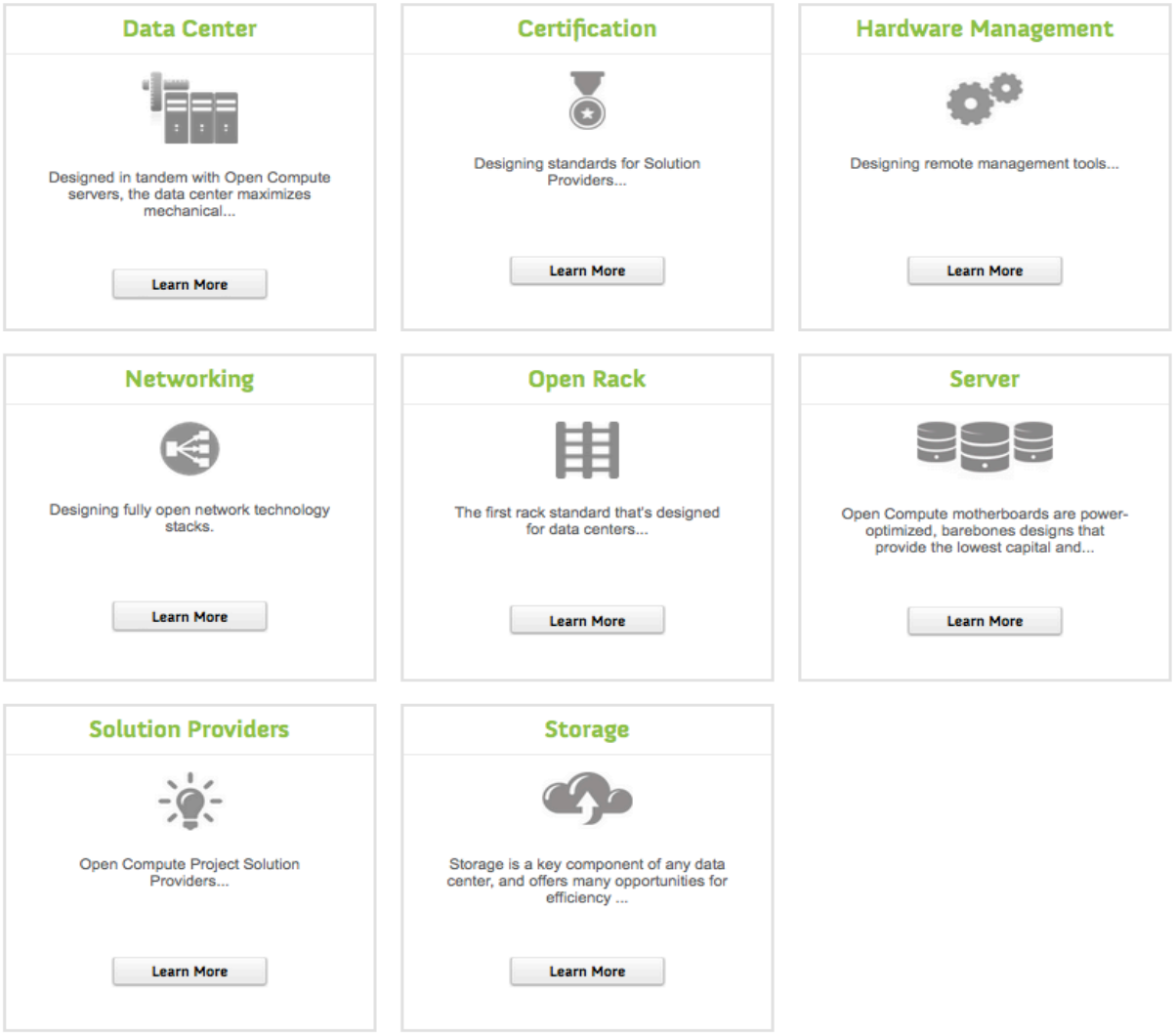
モジュールの組み合わせ

集中管理、集中運用

ハードウエアマネジメント、プロビジョニングソフトウエアを共通化

OCPとは
OCPの目的
OCPプロダクト
OCPの普及
まとめ
OCP-J

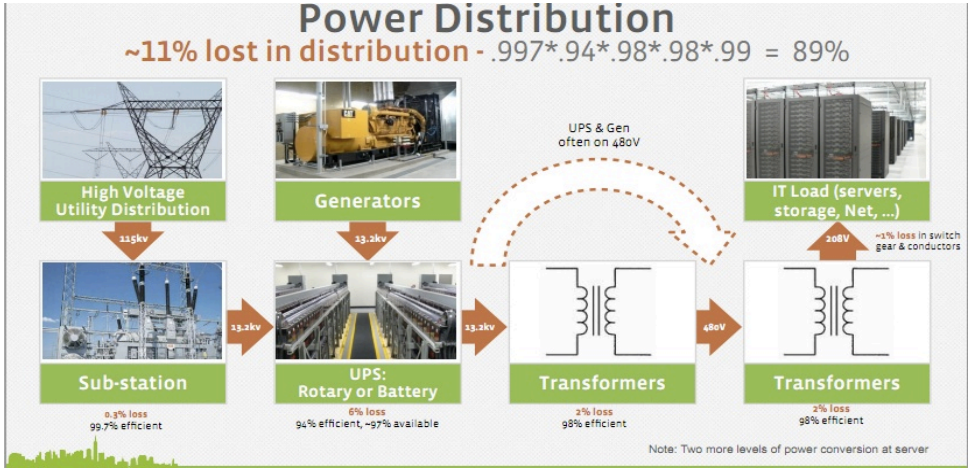
Projectで定義し　仕様を公開 共有

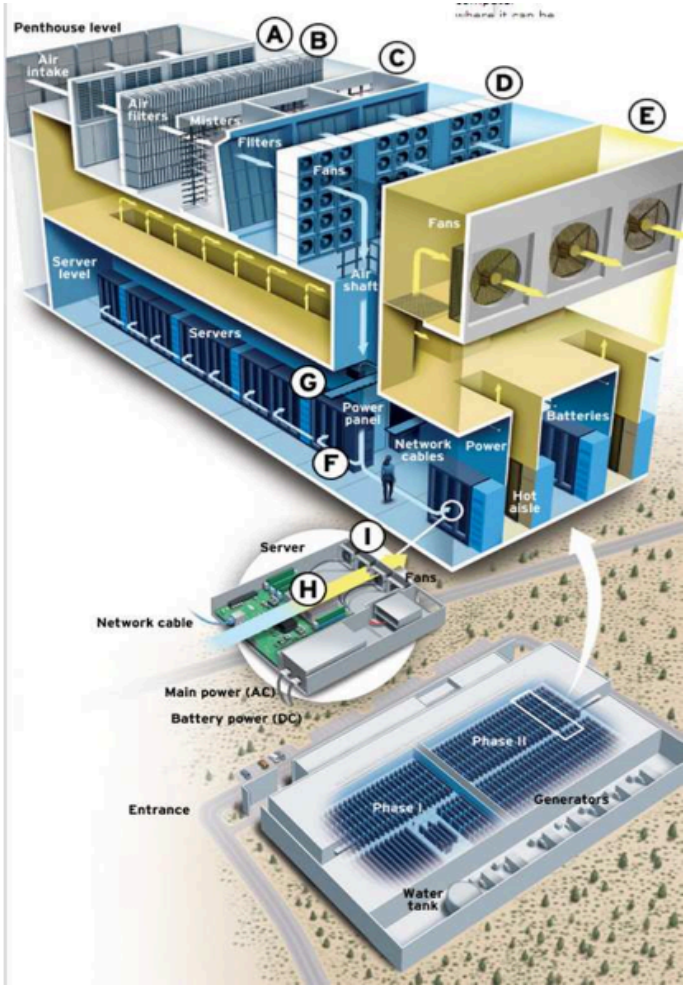


Data Center Design

省電力性能の向上

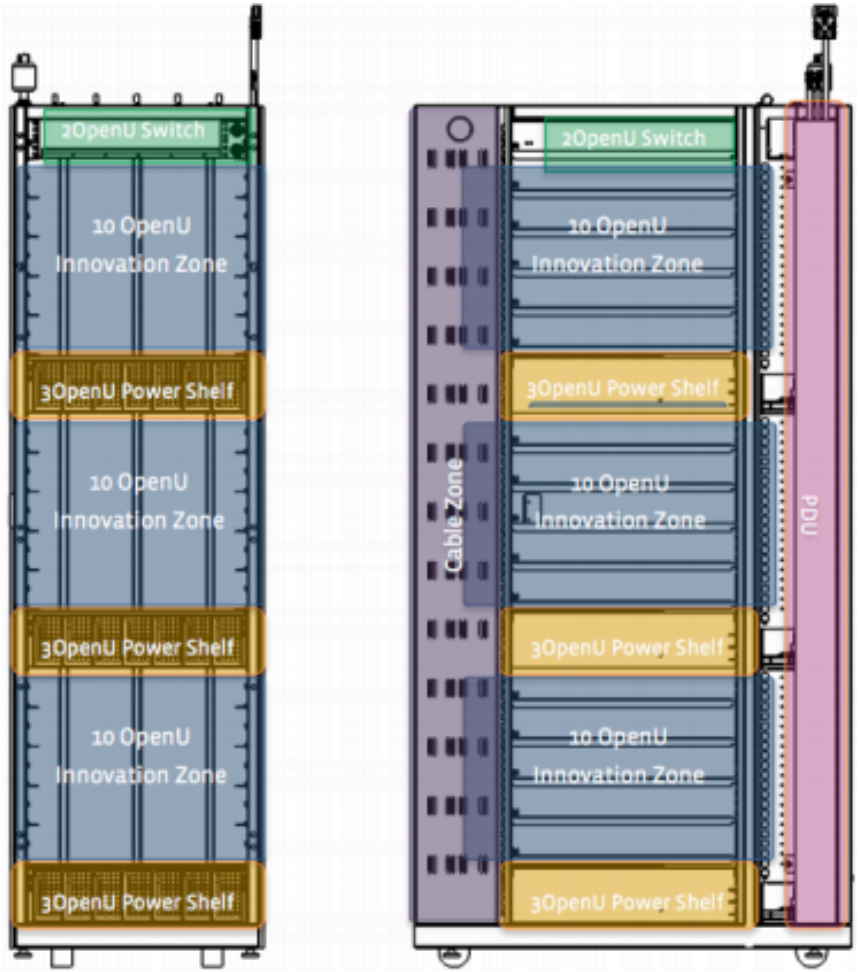
- 環境性能
- 冷却方式
- 発電～給電～ラック配電





Open Rack

19インチ、21インチ
電源・UPSを包括
バスバー給電
接続コネクタ
ファンをラックに搭載
空調、電源等、ラック単位のマネジメントシステム
工具なしにServerが交換可能

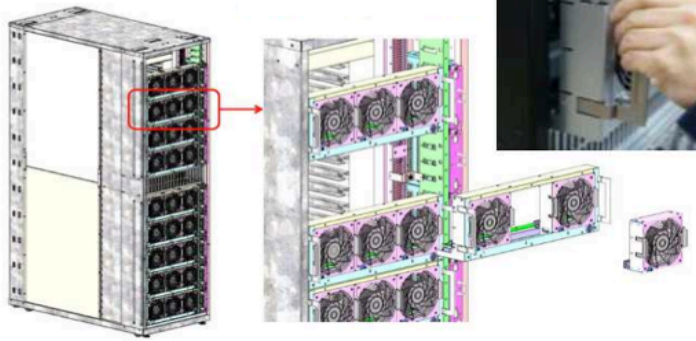


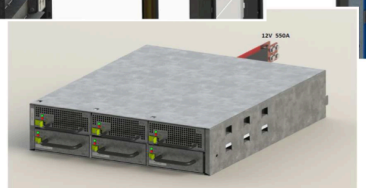
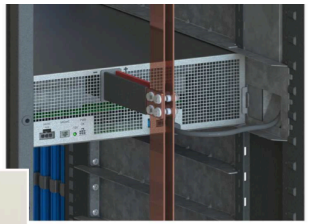
Open Rack

ユーザーに応じ様々な
組み合わせが存在
バスバー、パワーシェルフが異なる

- Open Rack V1/V2
- Rackspace Open Rack
- Fidelity Open Rack
- China Scorpio 2.0

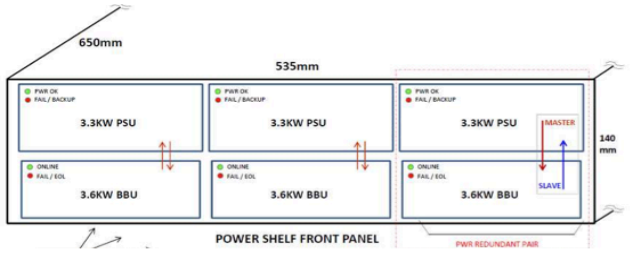


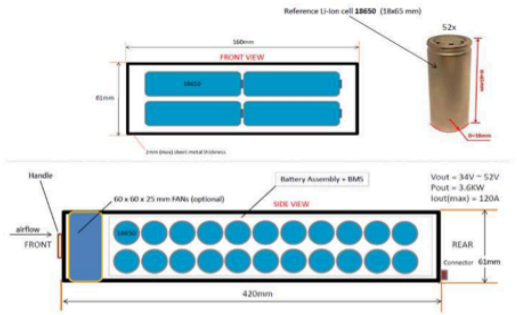


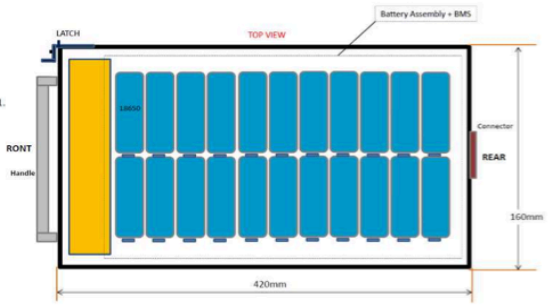




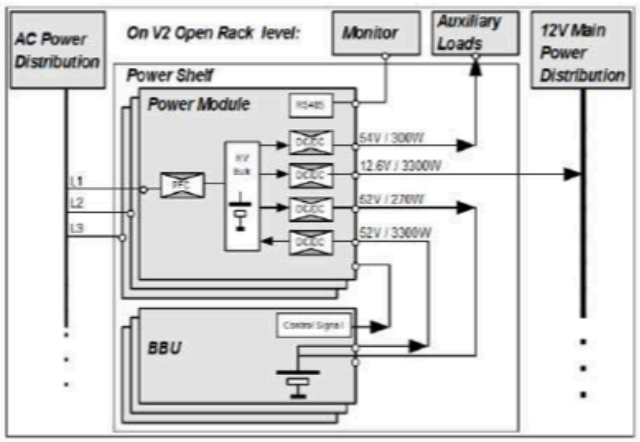
Power shelf

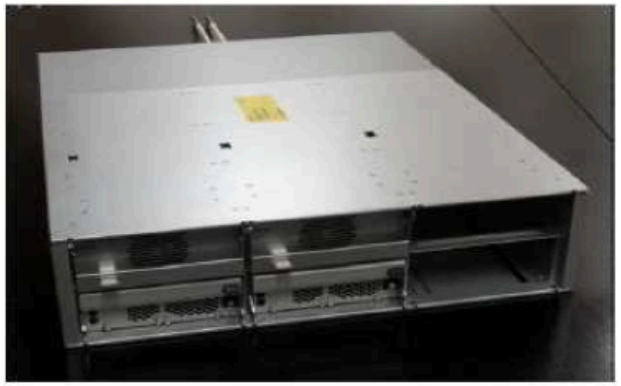
Power modules and Li-ON batteries in the same shelf
Single 12V Busbar output(535A)
Three-phase input
2+1 Redundancy + batteries
534mm x 612mm x 19mm

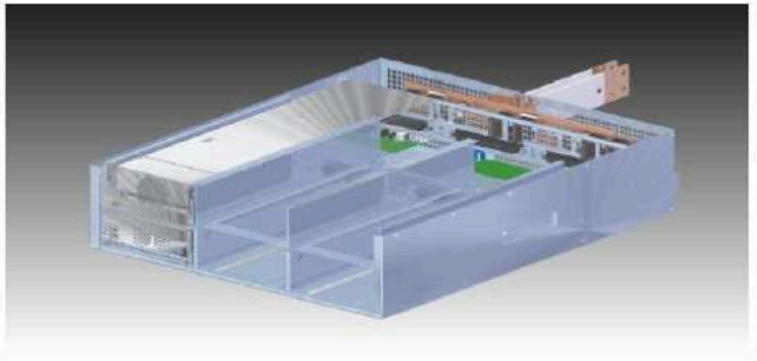












Facebook Server /Storage types

Standard Systems	Type I	Type II	Type III	Type IV	Type V	Type VI
CPU	High 2 x EN2670	Low 1 x 6128HE (AMD)	Medium 2 x X5650	Medium 2 x X5650	Low 1 x L5630	High 2 x EN2660
Memory	Low 16GB	High 144GB	High 144GB	Medium 48GB	Low 18GB	High 144GB
Disk	Low 250GB	Low 250GB	High IOP 6 x 600GB SAS +2x1.3TB Flash	High 12 x 3TB SATA	High 12 x 3TB SATA	Medium 1TB SATA
Services	Web, Chat, Ads	Memcache, Ads	Database	Hadoop	Photos, Video	Multifeed, Search

Standard Systems	Type I	Type II	Type III	Type IV	Type V	Type VI
CPU	High	Low	Medium	Medium	Low	High
Memory	Low	High	High	Medium	Low	High
Disk	Low	Low	High IOPs	High	High	Medium
Services	Web, Chat, Ads	Memcache, Ads	Database	Hadoop	Photos, Video	Multifeed, Search

省電力

電源、ファンをラックに搭載
21inc 幅とフロントパネルの排除で冷却効率をアップ

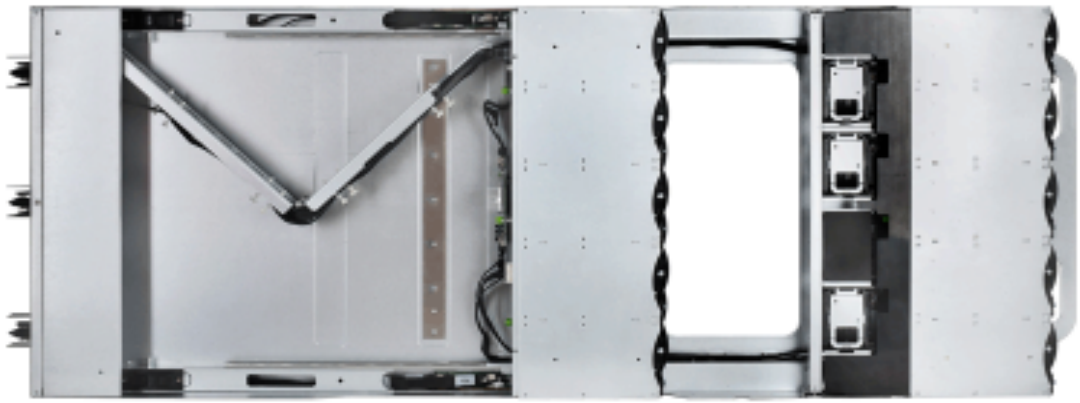
高集積

2CPU/16メモリスロットを横に3枚
3.5inc　HDDを14本

運用性

工具なしで部品交換が可能

Quanta JBOD for OCP



JBR

High Density 2U JBOD with Tool-less Tray Design

- Front load screw-less HDD trays
- Lock-in mini-SAS module
- 2OU JBOD with QCT Patented
- support up to 28 x 3.5" hot-swappable SATA/SAS HDDs

出典：http://www.quantaqct.com/Product/

Microsoft OCP & Cloud Server

Microsoft　OCPにコントリビュート

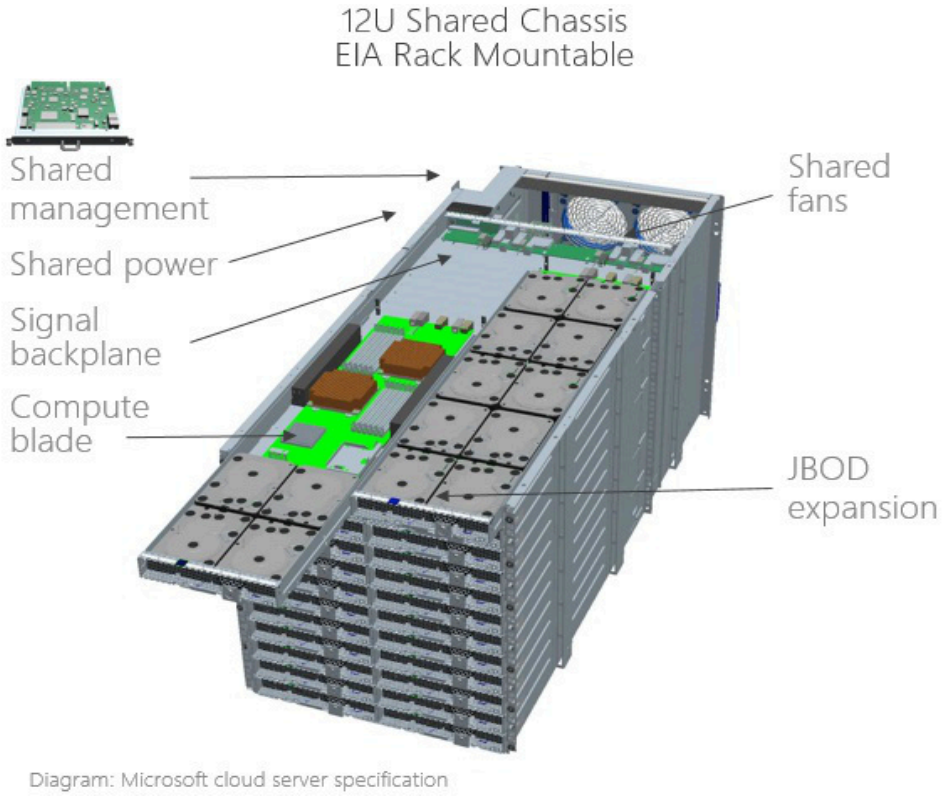


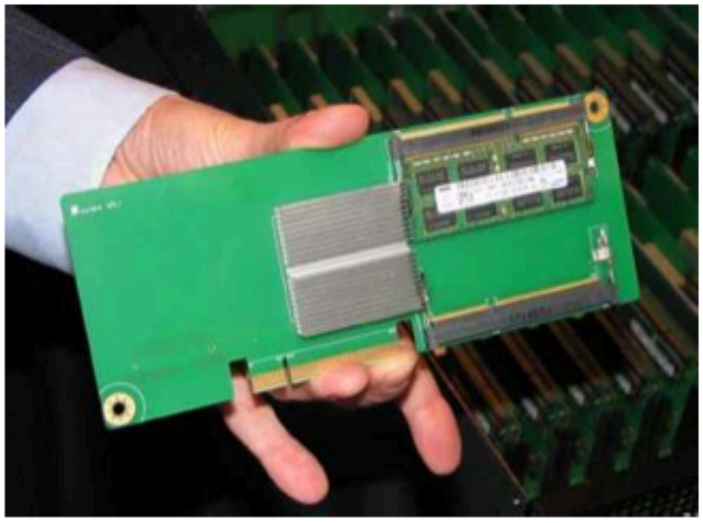
Diagram: Microsoft cloud server specification

chassis v1.0
Blade v1.0
JBOD v1.0
Chassis Management v1.0
Network Mezzanine v1.0
SAS Mezzanine v1.0
Chassis Management Software source code

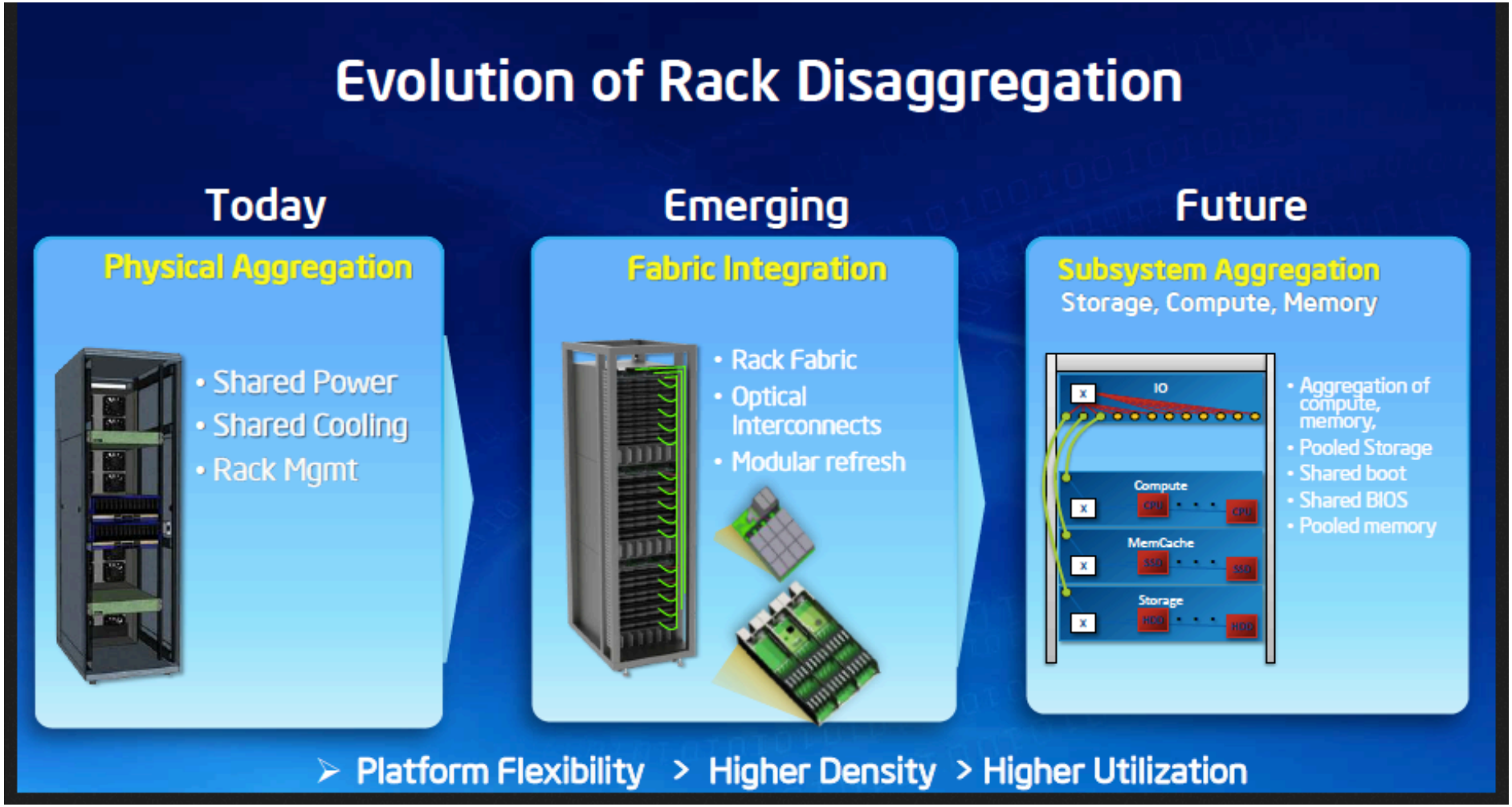
出典：https://gigaom.com/2014/01/27/microsoft-the-software-king-wants-to-tell-the-world-how-to-build-servers/

Rack Scale Architecture

- プロセッサやメモリ、ストレージ等の集合体を、それぞれ1つのモジュールとして扱う
- プロセッサの集合体を単一のプロセッサのように扱い、メモリストレージも同様に管理
- 一般的なラックマウントサーバーでは、CPUトマザーボード、メモリの組み合わせで成り立っているが、RSAではこれらの差異をモジュールで吸収

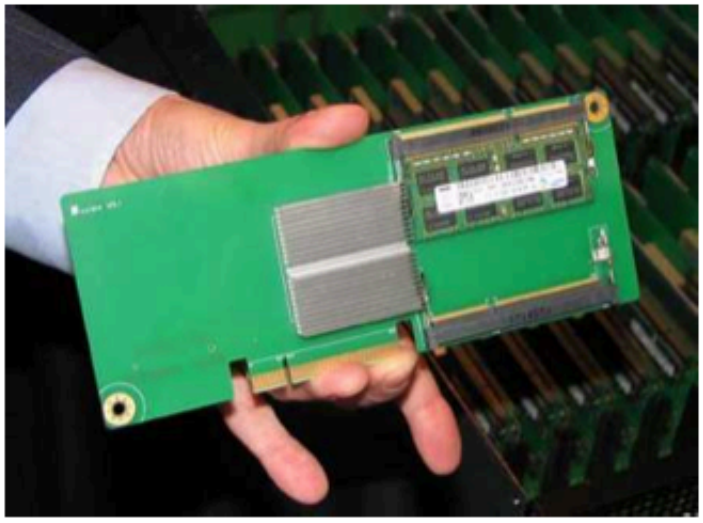


Rack Disaggregation



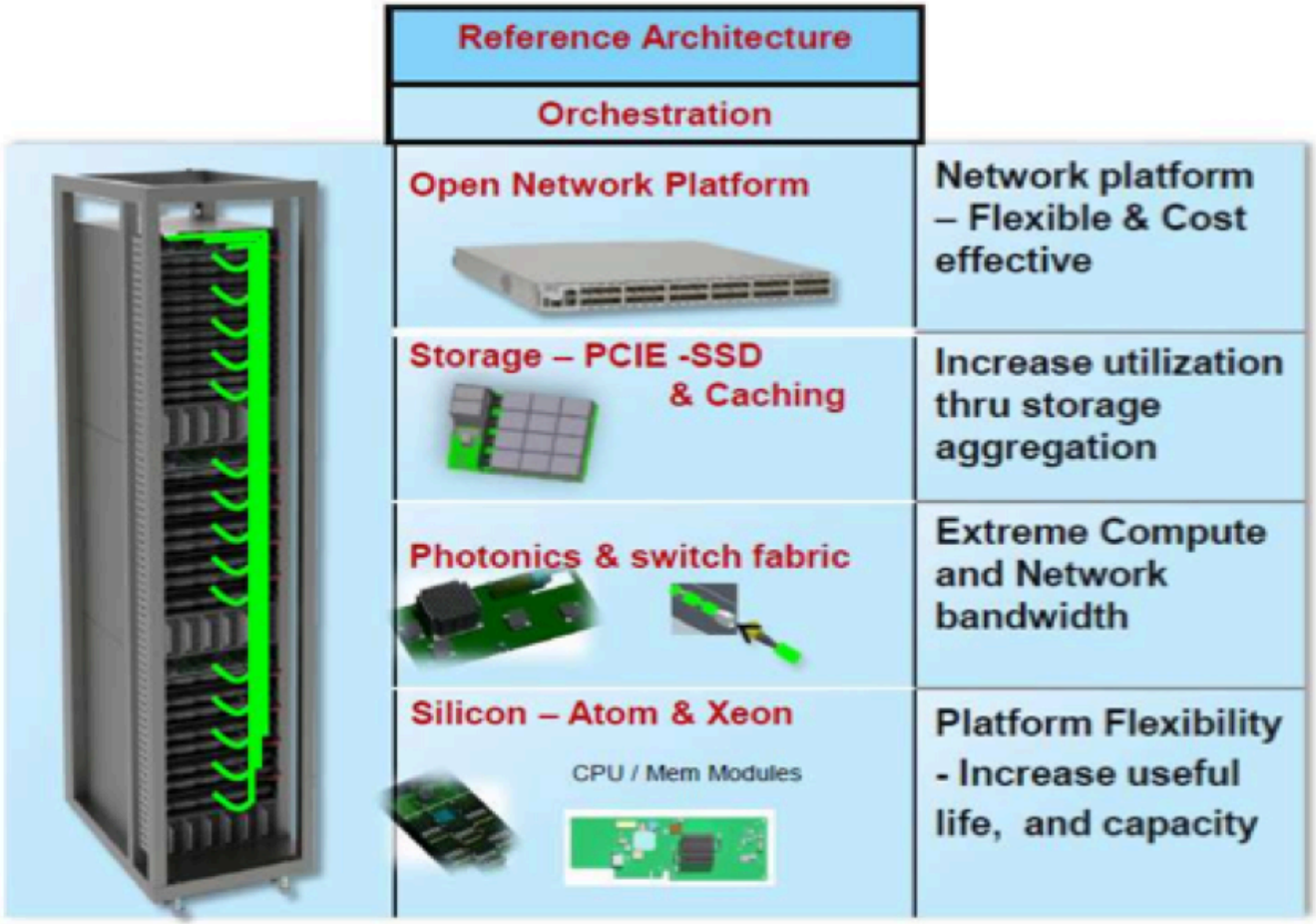
Rack Scale Architecture

- プロセッサやメモリ、ストレージ等の集合体を、それぞれ1つのモジュールとして扱う
- プロセッサの集合体を単一のプロセッサのように扱い、メモリストレージも同様に管理
- 一般的なラックマウントサーバーでは、CPUトマザーボード、メモリの組み合わせで成り立っているが、RSAではこれらの差異をモジュールで吸収

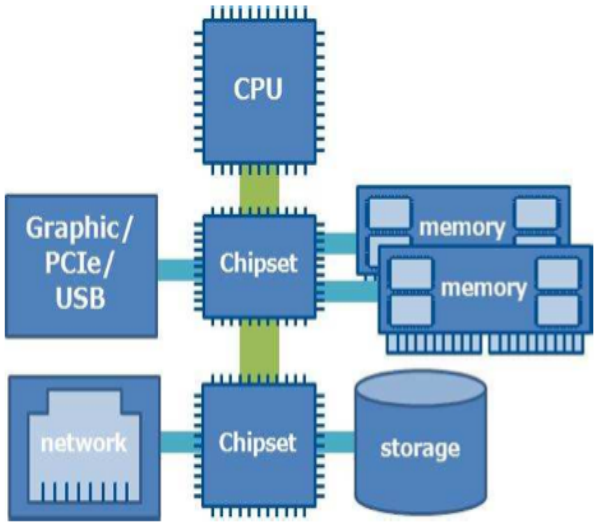


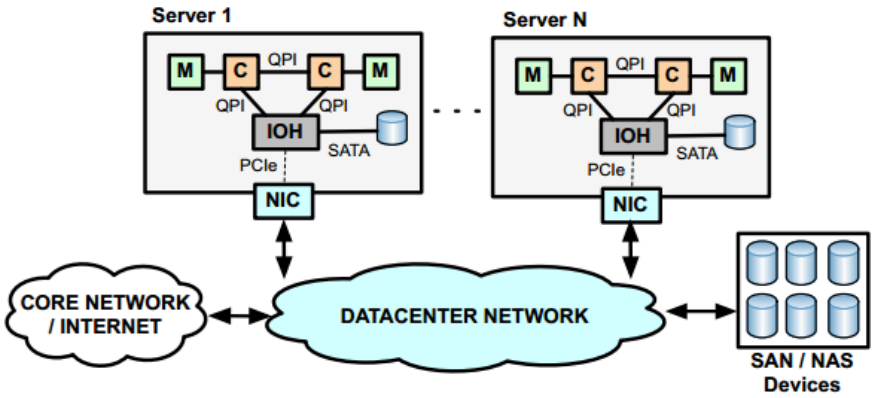
Intel Rack Scale Architecture

- Intel technologies optimized for flexibility, performance & cost
- Open rack scale reference architecture to simplify adoption
- Driving alignment on common standards with broad range of uses (end users, Scorpio and OCP) and OEM implementations

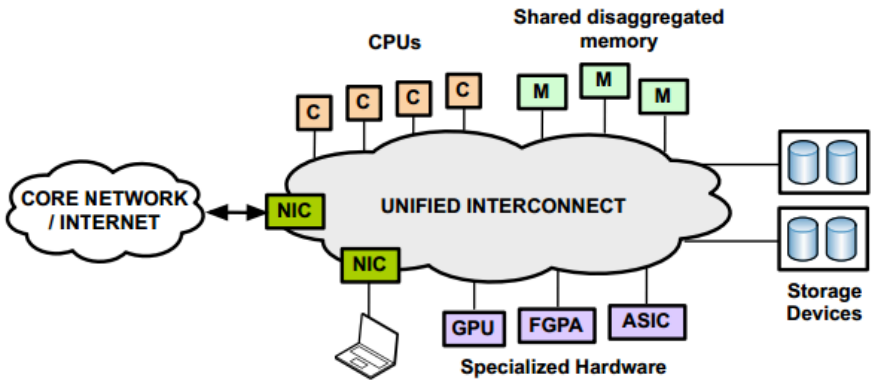


Disaggregated Datacenter





(a) Current datacenter

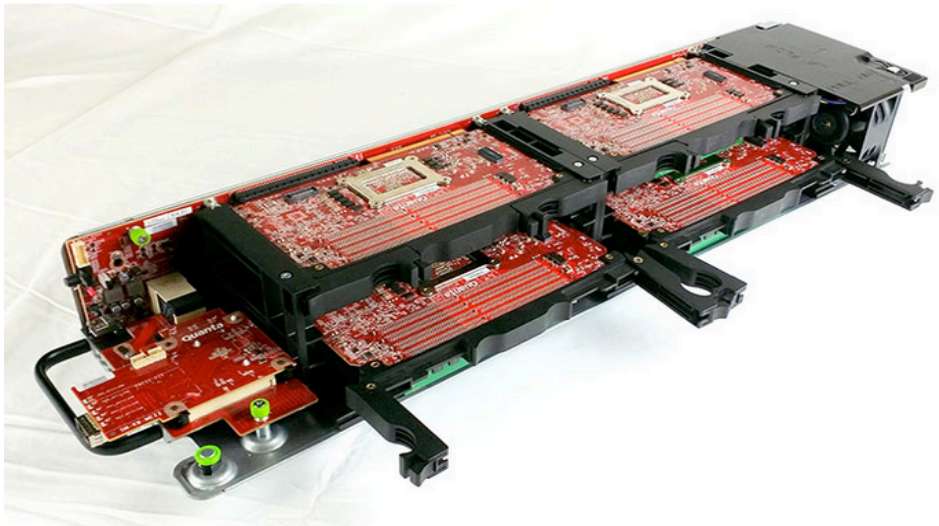


(b) Disaggregated datacenter

http://conferences.sigcomm.org/hotnets/2013/papers/hotnets-final40.pdf

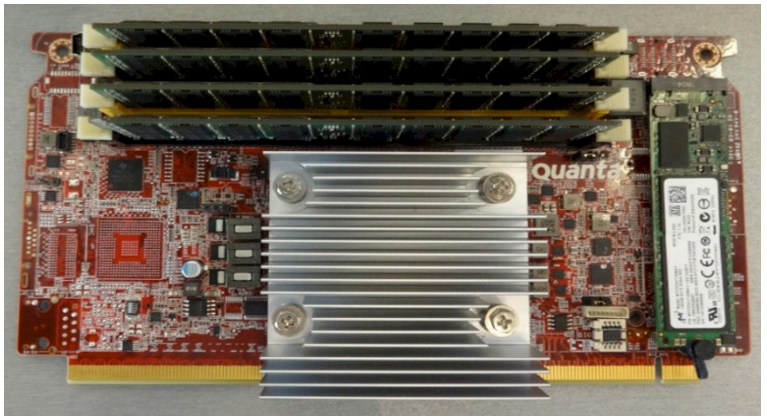
https://lazure2.wordpress.com/2013/12/10/disaggregation-in-the-next-generation-datacenter-and-hps-moonshot-approach/

Yosemite / 1S Server



OpenRackV2 192 SoC servers
PCI-Express x16 mechanical slots
X86 （ARM, Power）
40GbE Mellanox C-4 hybrid mezzanine card
400W

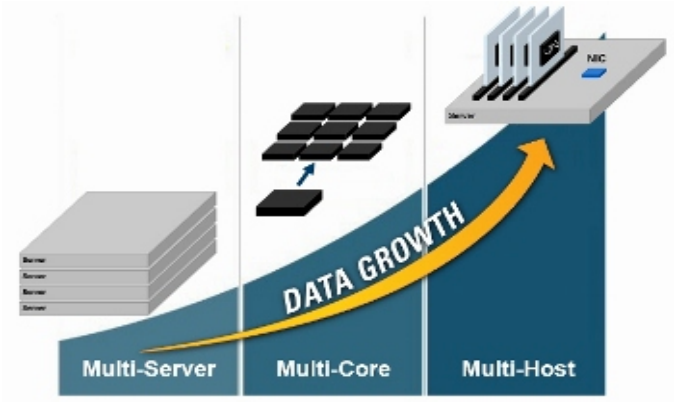
Intel Xeon D-1500 SoC
HighPowered-SoC Micro Server
210×110mm
M.2 SSD
10GbE
ローカル管理コントローラ
65W

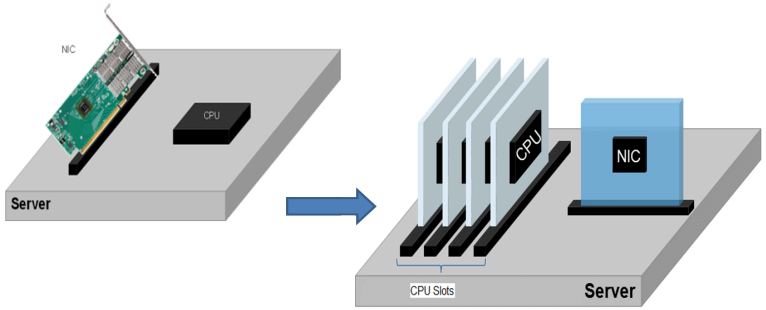


intel created with Xeon D processor and worked with Quanta to design the board
and to get the microserver manufactured.
Facebook and Quanta designed the sideplane and the hybrid mezzanine card along with Mellanox.

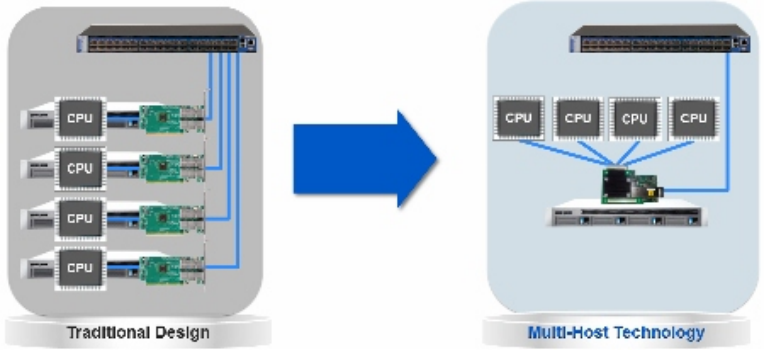
出典：https://code.facebook.com/posts/1616052405274961/introducing-yosemite-the-first-open-source-modular-chassis-for-high-powered-microservers-/

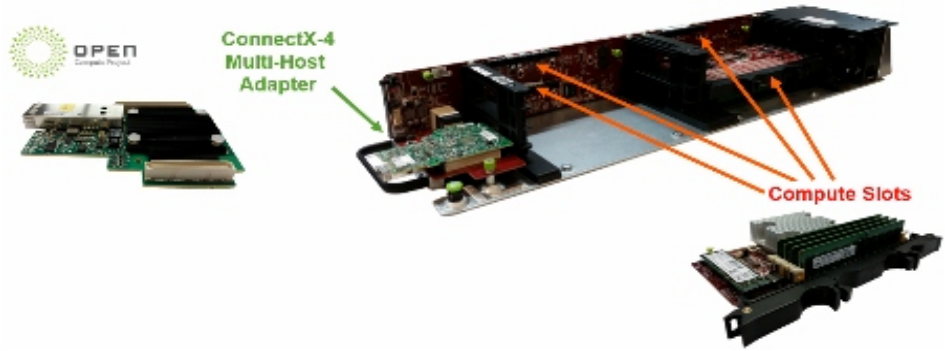
ConnectX-4 and Multi-Host





10/25/50/100 Gigabit Ethernetアダプタ用IC
4つの完全に独立したPCIeバス
ホスト間で独立したQoS
単一のネットワーク·コントローラに複数の
異種ホスト（x86、ARM、GPUなど）の直接接続





出典：http://www.mellanox.com/page/products_dyn?product_family=210&mtag=multihost

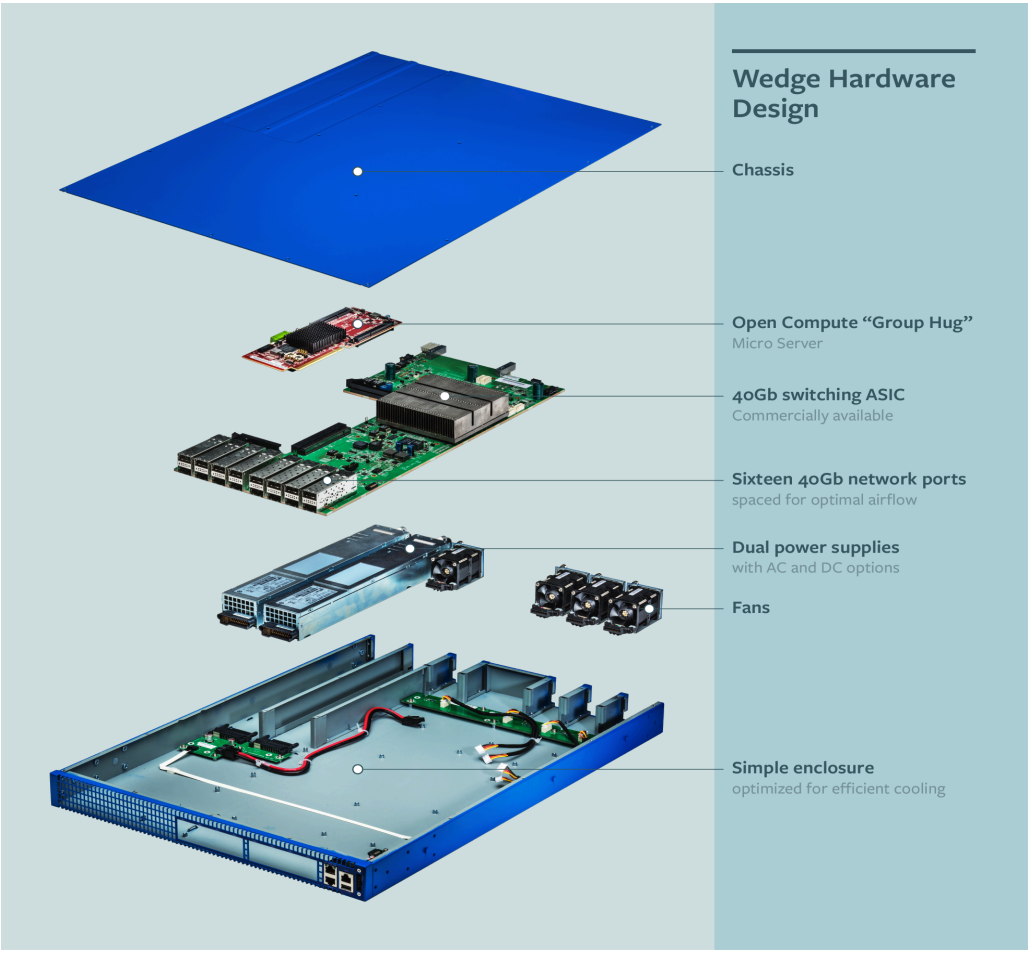
Switch

We wanted to make switches feel like servers.

Our goal is to help make networking hardware that is open, and to foster a wide variety of open source software that can run on top of it.

出典：https://code.facebook.com/posts/681382905244727/introducing-wedge-and-fboss-the-next-steps-toward-a-disaggregated-network/

TOR SW Wedge



Merchant Silicon
Trident II
1.28TbpsASIC
40Gbps×16

X86 Micro Server
OCP Group Hug

Software
FBOSS(Facebook)
ONIE
Open Network Linux

Baseboard Management Controller
OpenBMC

標準的なLinuxベースのOSで
スイッチをプロビジョニング

With "FBOSS," all our infrastructure
software engineers instantly become network engineers.

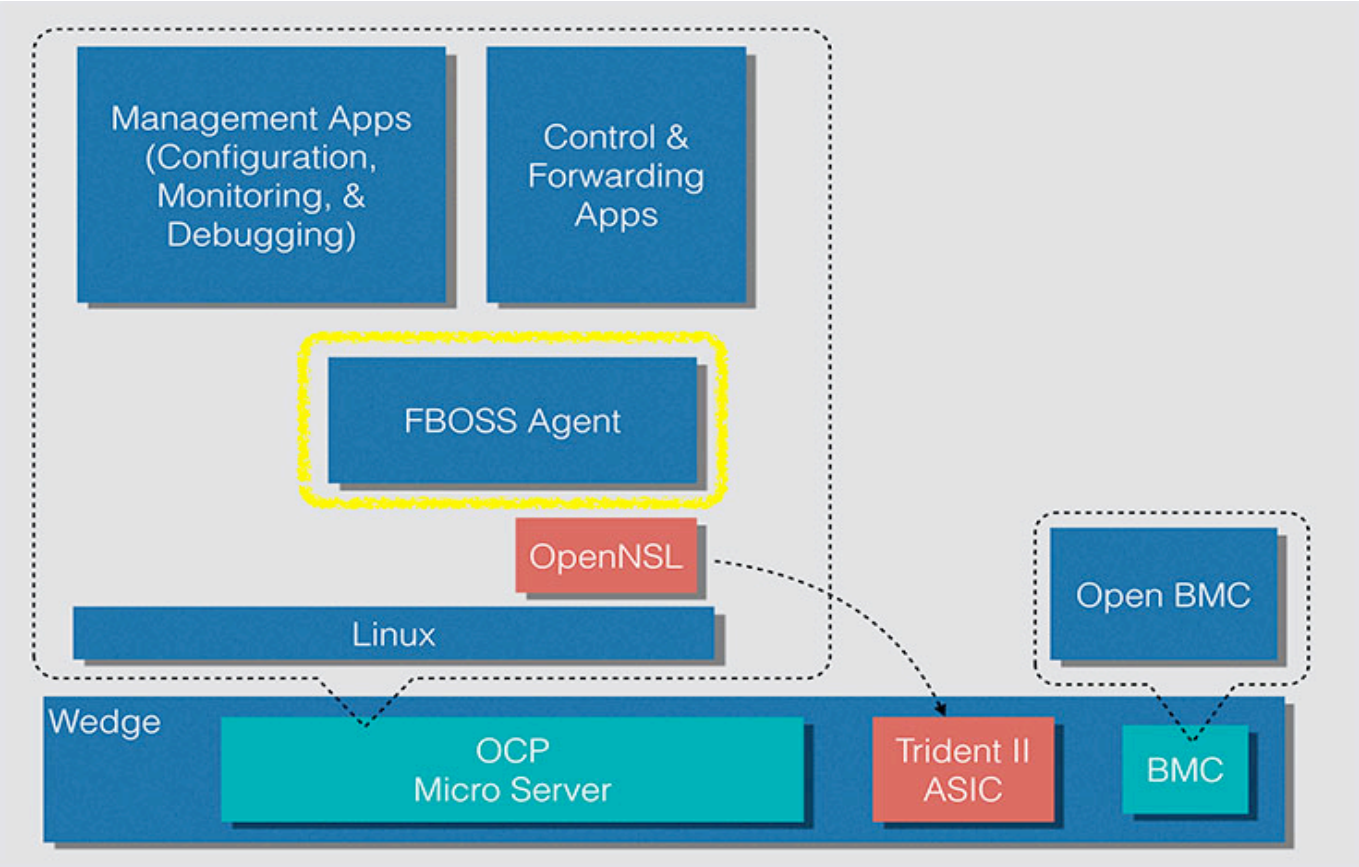
出典：https://code.facebook.com/posts/681382905244727/introducing-wedge-and-fboss-the-next-steps-toward-a-disaggregated-network/

Open Network Linux

SWのOSを共同開発するプロジェクト

Facebook	FBOSS
NTT	L3 Routing
Big Switch Networks	Open Flow

FBOSS / OpenNSL / OpenBMC



FBOSS
標準Linux上で実行可能な
SWアプリケーションの集合

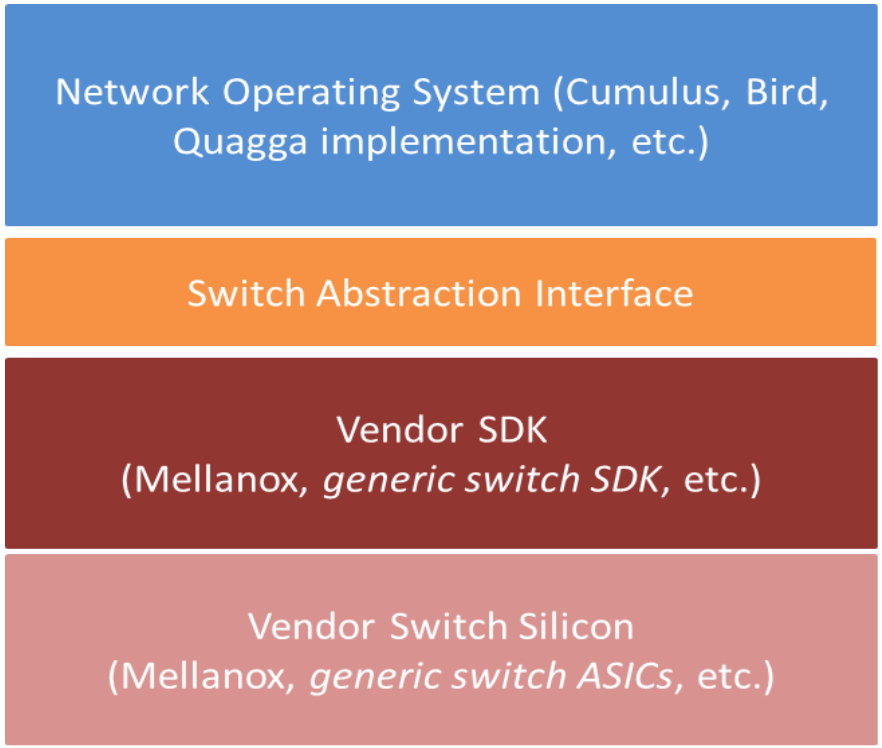
Open NSL
SW　ASICのAPI
ASICのプログラミングが可能

Open BMC
リモート電源、環境制御　監視
サーバーのホストCPUとメモリのエラー
ログ等のシステム管理

Up until now, building open source switching software has been difficult, because there are only a handful of companies that build switching ASICs. Aided in part by the efforts of the OCP, several ASIC vendors are now beginning to open up some of their APIs and SDKs.

出典：https://code.facebook.com/posts/843620439027582/facebook-open-switching-system-fboss-and-wedge-in-the-open/

Switch Abstraction Interface (SAI)



スイッチ抽象インタフェース

Microsoft,Mellanox, OCP

アプリケーション、プロトコルを異なるベンダーのASIC上でシームレスに動作させる

ハードウェアのSDKと接続

スイッチング、ルーティング

ポート管理、データ転送

ACL、QoS などの機能を統合

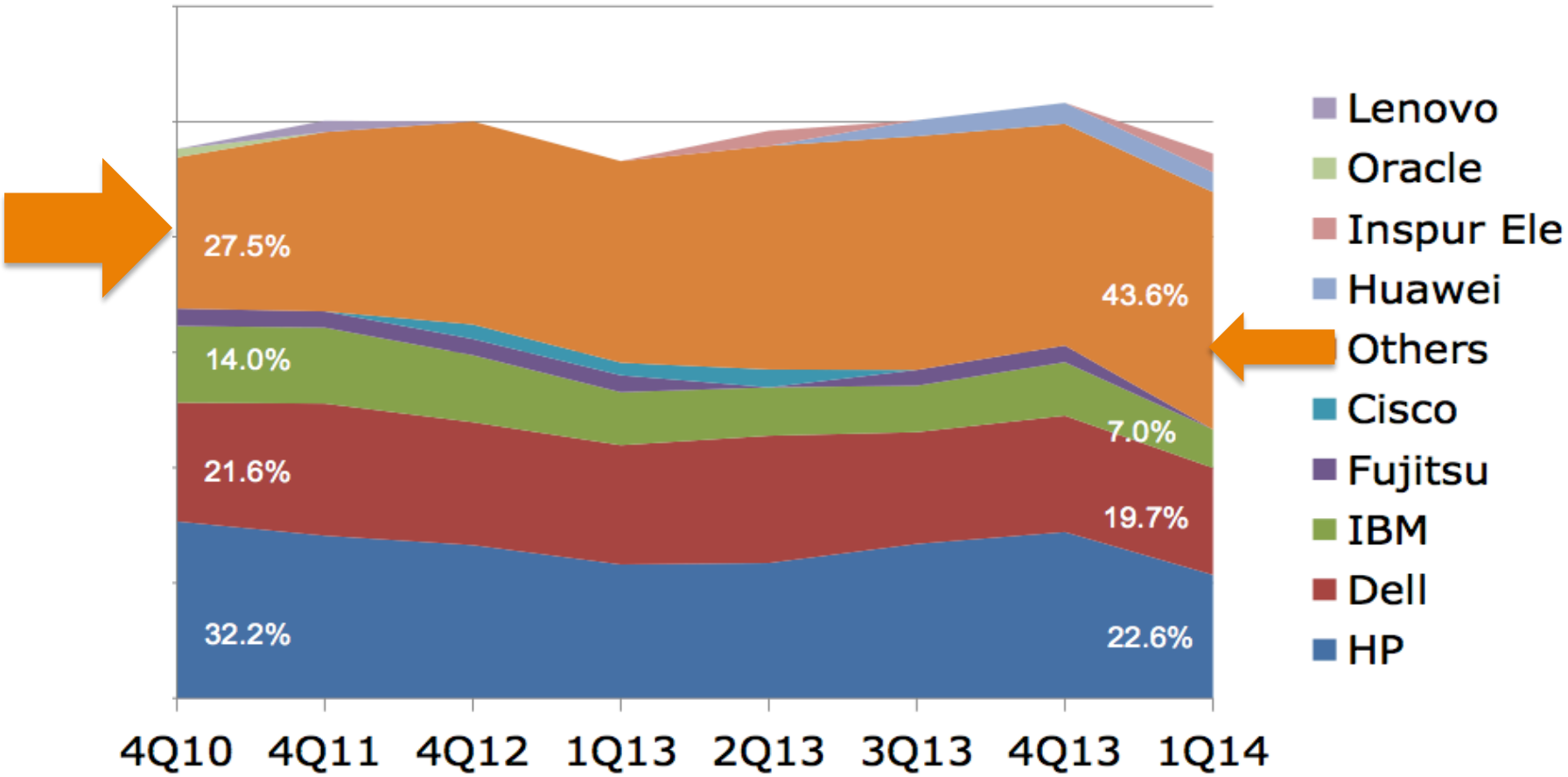
出典：http://www.mellanox.com/page/products_dyn?product_family=210&mtag=multihost

FacebookとOCP
OCPとは
OCPの目的
OCPのアーキテクチャー
OCPプロダクト

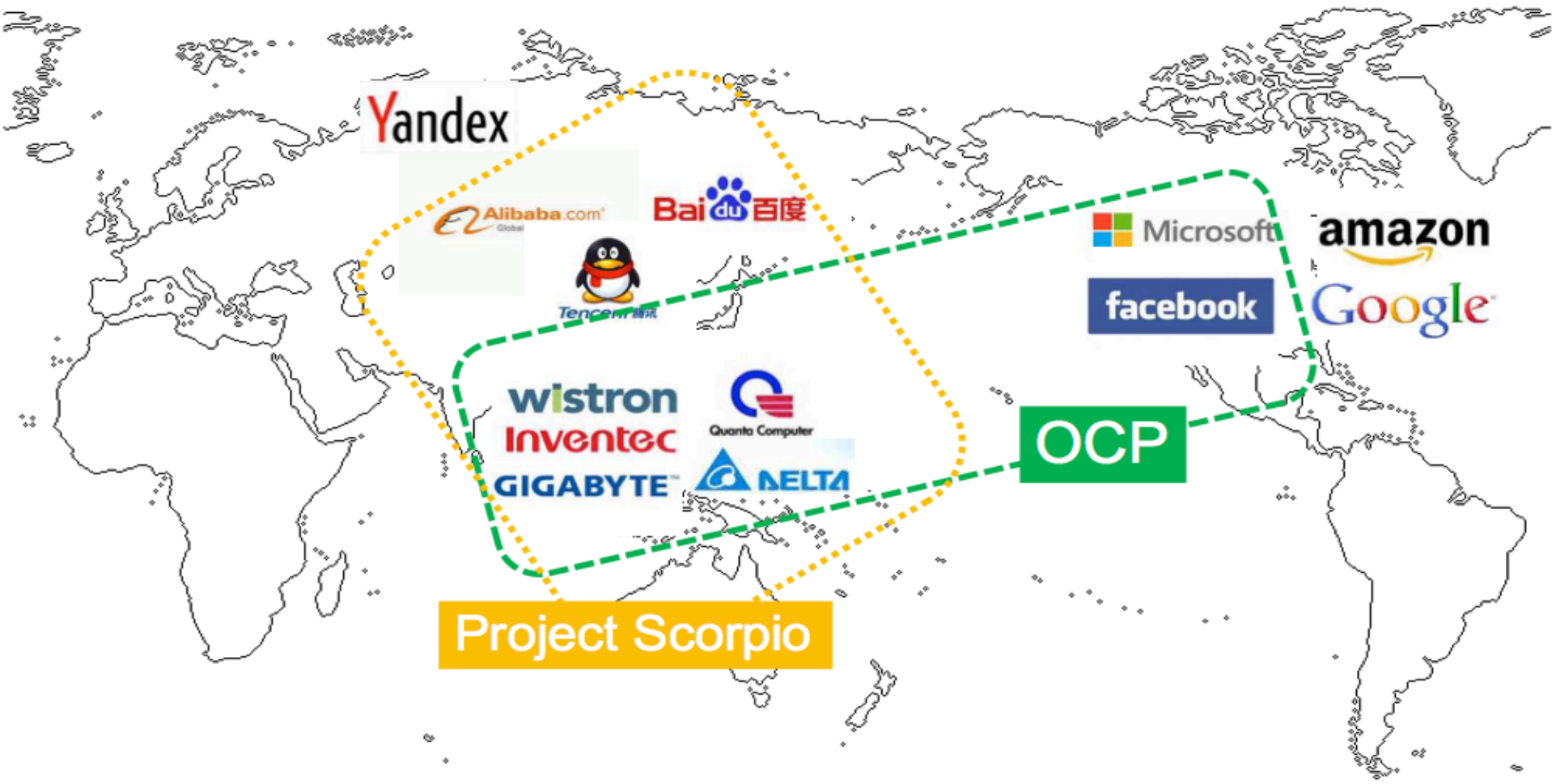
OCPの普及

まとめ

Gartner Worldwide: Server Vendor Shipment Estimates



OCPとProject Scorpio



OCP

安価で高性能な データセンター

- アプリケーション、ソフトウエア指向
- ベンダーレイヤーに捉われない、ライフサイクルマネジメント

Disaggregate

- ハードウエアをモジュール単位で分解 再設計
- プログラマブルな運用IF
- ソフトウエアとハードウエアの分離

インフラ･コミュニティの育成

- サプライチェーンのショートカット

Open Compute Project Japan

主な活動内容

- Open Compute Project 成果のシェア、日本からの提言
- 国内ファシリティー開発者との技術情報のシェア
- 日本国内の関連技術のオープンソース化
- 先端データセンターによる実証実験(実証値測定)
- 省電力化及び全体最適化、運転手法の実証実験
- 海外情報、仕様のトランスレーション
- 情報公開、共有手法の検討、
- ナレッジサイトの運営

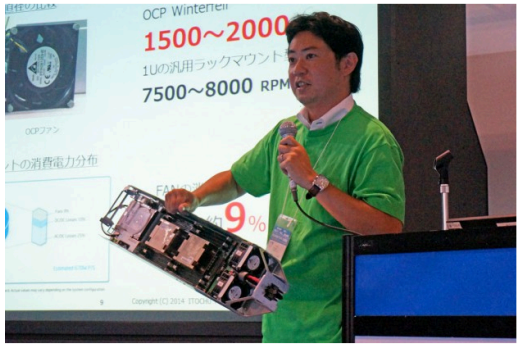
OPEN
Compute Project
Japan

Open Compute Japan WG

Proof of Concept WG

概要及び目的

OCP CERTIFIED/READYのサーバ、ストレージ、オープンラックなどの検証からPoC(Proof of Concept)を実施し、システムアーキテクチャから日本市場の技術条件に一致した各種仕様の検討/策定を目的とします。



INTEROP Tokyo 2014 (2014.06.11-13)

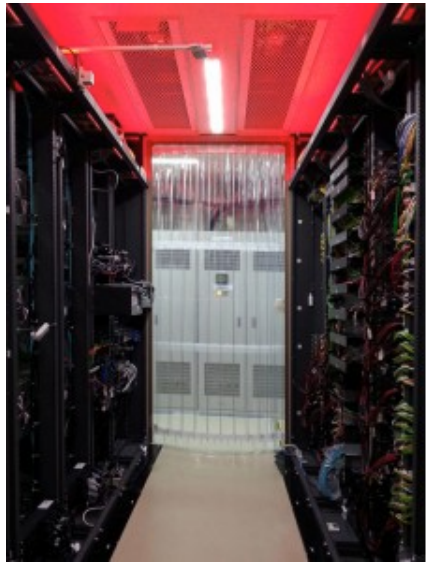
HVDC WG

概要及び目的

消費電力量の削減を図り、給電信頼度を高め、省スペースを実現するHVDCシステムをデータセンタ向けの日本発の技術としてOCPに提言する。

内容

(1)HVDCシステムの広報活動
(2) GUTP(グリーン東大)DCIM-WG HVDC-SWGと連携した標準化



Open Compute Japan WG

Public Relations WG

概要及び目的

OCP及びOCPJ(またはOCPT)の活動状況を外部に発信する、及び、OCPの活動の推進。

Translation WG

概要及び目的

OCPが発行する各種ホワイトペーパーなどを日本語に翻訳する。また、その成果物である日本語ドキュメントを広報WGから配布することで、Open Computeに関する情報共有を促進する。

Compliance & Interoperability WG

概要及び目的

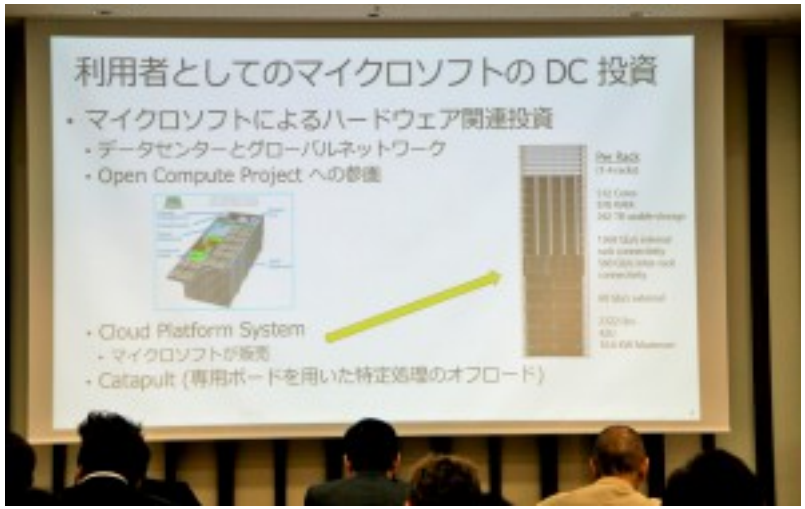
OCPJ C&I WGはOCPが定義したガイドラインにとどまらず、日本の市場が要求する日本独自のスペックやS/Wレイヤー、OSSコンパチビリティまで含めたOCPJ推奨仕様を検討する。

OCP Taiwan and OCP Japan – 2014 Cloud Computing Day Tokyo(2014/09/19)

OCPJ Meet up 2014/05/19

2015.10.26.OCPT & OCPJ 2015 Cloud Computing Day Tokyo

OCPJ Meet-up 2015 夏(2015/06/26)



http://www.opencomputejapan.org