FUJITSU
shaping tomorrow with you

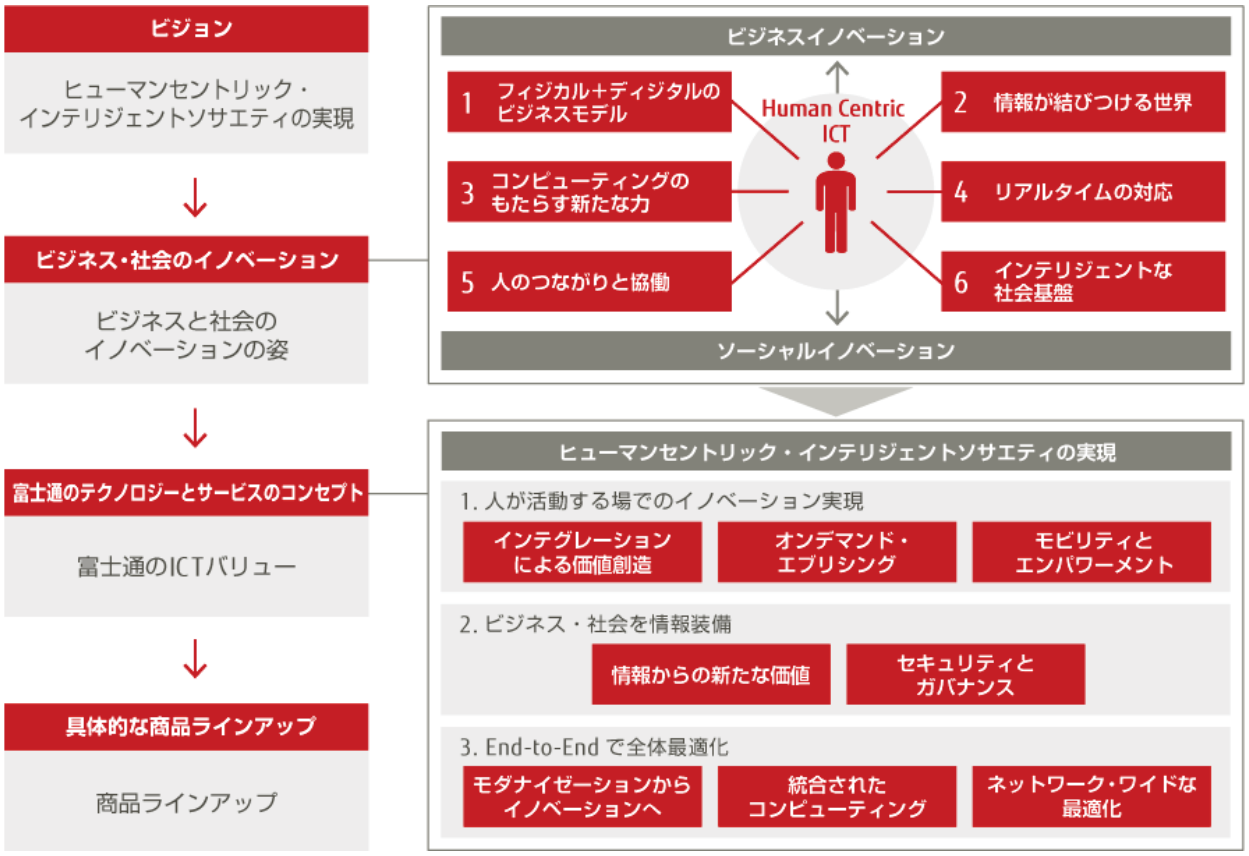
ストレージ，ネットワークを含めた今後のクラウドの方向性

2014年1月27日
株式会社　富士通研究所
システムソフトウェア研究所
湯原 雅信

富士通が考える新たな社会のビジョン

FUJITSU

Fujitsu Technology and Service Vision



http://jp.fujitsu.com/vision/paper/

本日のメッセージ

FUJITSU

今後のIaaSクラウドの方向性

- 分散技術の利用範囲拡大
- 専用装置から汎用ハード＋ソフトへ
- Software Defined xx へ

IaaS: Infrastructure as a Service

目次

FUJITSU

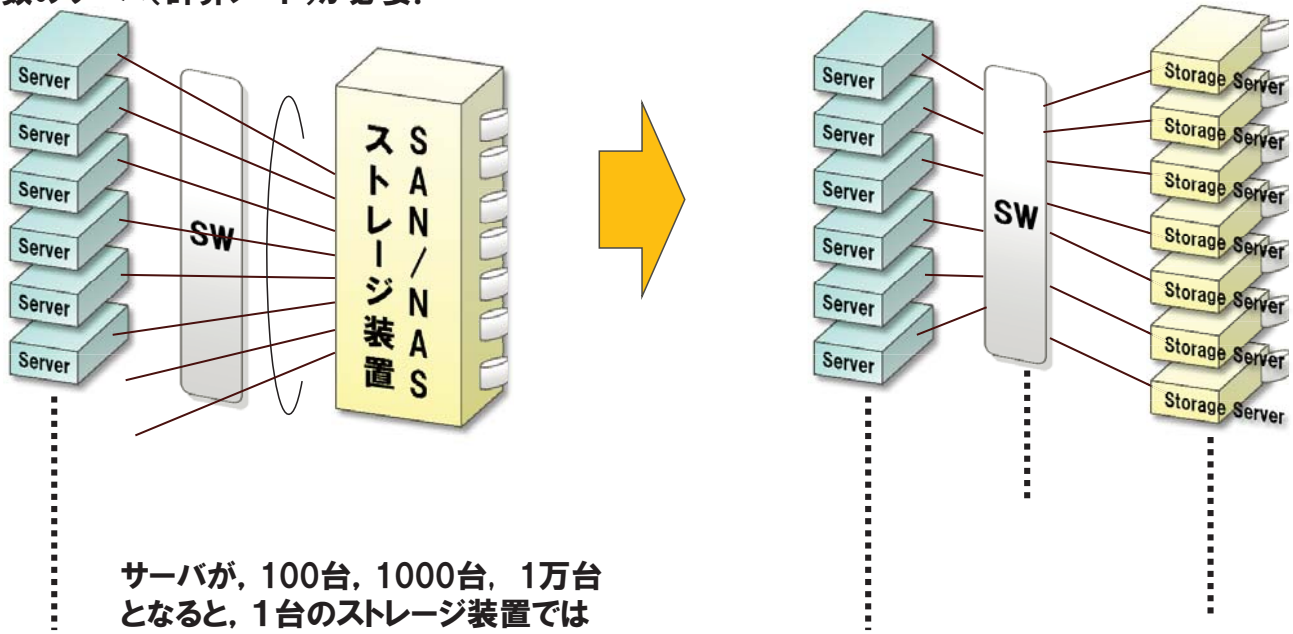
- ストレージ関連のトピック
- ネットワーク関連のトピック
- サーバ関連のトピック
- 今後のクラウドの方向

ストレージ関連のトピック

- 分散ストレージ
- 不揮発性メモリ

ビッグデータから求められるストレージ

ビッグデータの大規模高速処理には，
多数のサーバ（計算ノード）が必要．



サーバが，100台，1000台，1万台
となると，1台のストレージ装置では
対応できない（スケールしない）．

ストレージ側も台数を追加できる構造にする
（スケーラビリティを確保）．

分散ストレージ

分散ストレージ実現上の課題

FUJITSU

① 論理構成管理（ここでは論理＝ストレージ利用者への見せ方）
- ストレージ機能モデル(データの論理的な構成)をどうするか
  - ブロックストレージ: block、ファイルシステム: ディレクトリ/ファイル、オブジェクトストレージ: bucket/object
- それらの論理的な情報を内部でどう管理するか

② 物理構成管理（ここでは物理＝物理サーバや物理デバイスへの格納方法）
- データ本体、論理的な情報(メタデータ)を物理的にどうマッピングするか
  - 集中、単純分散ハッシュ、RING、CRUSH、他

(A)

③ 一貫性制御
- 1つの論理データ(物理的に複製されている)への複数アクセスの協調をどうするか
- キャッシュを分散して持つ場合の一意性をどう保障するか
  - メタサーバ管理型、楽観的制御型、他

④ 構成サーバ管理
- サーバの追加、削除、同期など

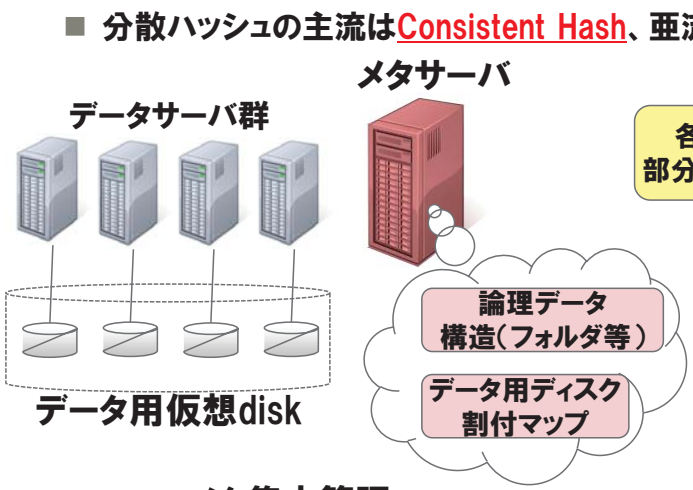
⑤ RAS
- データの信頼性(正しいこと)、保全性: 整合性チェック
- データのavailability(アクセスできること)、durability(失くさないこと):レプリケーション、分散erasure-code
- 保守性(メンテナンスしやすいこと):サーバ故障時の動作
- セキュリティ: マルチテナント対応

(B)

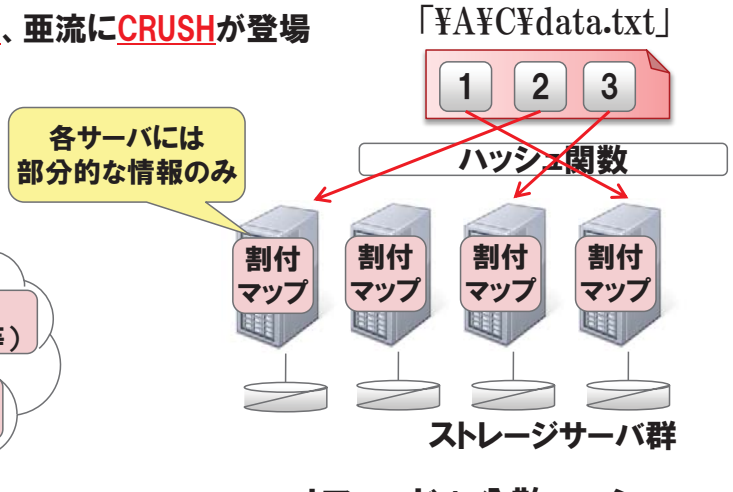
(A) 物理構成管理

FUJITSU

- 更新のアトミック性保証が課題
  - 例:ファイル/オブジェクトの構成ブロックの配置
  - 例:ファイルとブロックの関係と、ブロック割付マップの同時更新
- 近年のデータ量増大で、ストレージ構成変更の性能/負荷が新たな課題に
  - 例:ノード追加時の大量ファイルの再配置の抑制
- 古くはメタ集中管理、近年はオフロード＋分散ハッシュがトレンド
  - ブロック割付マップを複数のストレージサーバに分散させたハッシュテーブルで実現
    - 各ストレージサーバは部分的な情報しか持たないが、全体として、必要なブロックにたどり着ける
  - 分散ハッシュの主流はConsistent Hash、亜流にCRUSHが登場



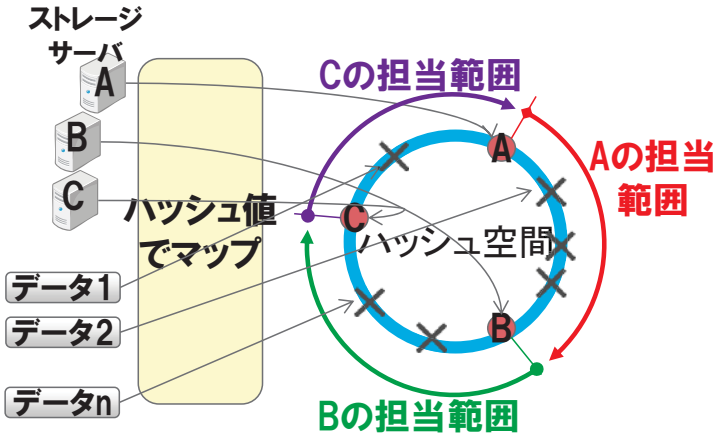
メタ集中管理
(GPFS、OCFS2、GFS2)

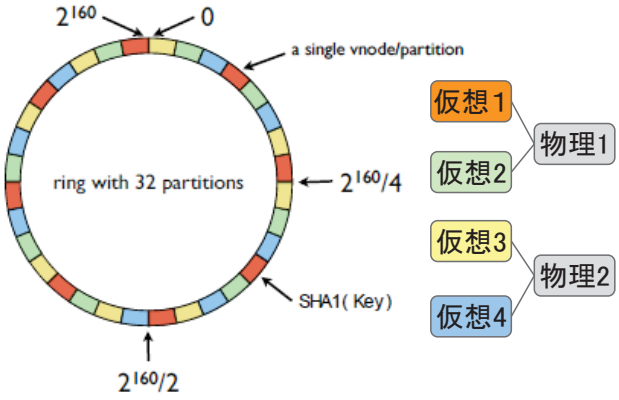


オフロード＋分散ハッシュ
(Ceph、Gluster、Swift、Cassandra)

Consistent Hash

- 基本原理
  - 各ストレージサーバの担当範囲を、ストレージサーバIDのハッシュ値から決定
  - 各データIDのハッシュ値から担当ストレージサーバを決定
- 利点
  - 管理用のデータが小さい（ハッシュ関数と、サーバのリスト）
  - ストレージサーバの追加/削除をハッシュ空間上隣接するサーバのみで対応
    - ➔ リハッシュが不要で、データ移動量が少
- 拡張
  - 物理サーバ毎に、性能/容量に応じた仮想ノードを定義しハッシュ空間に配置
    - ➔ 通常物理サーバ間の負荷の不均衡を低減





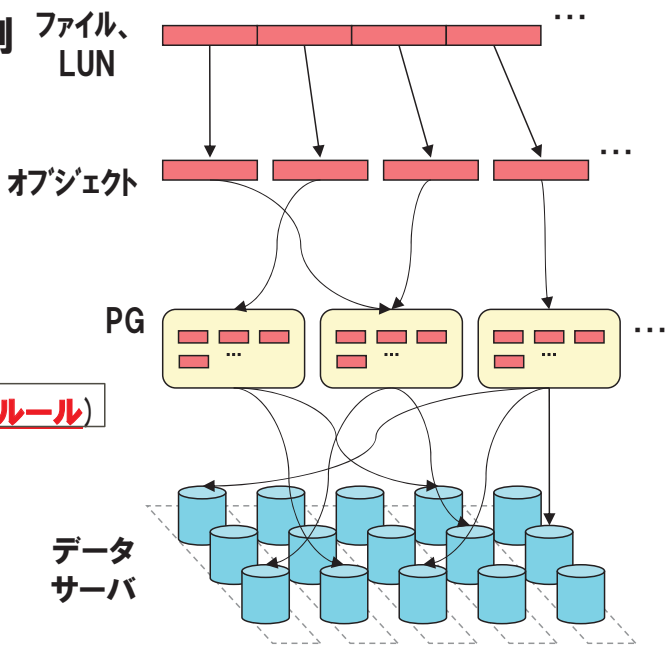
CRUSH

※CRUSH: Controlled Replication Under Scalable Hashing

1. ファイル/ボリュームをオブジェクトに分割
   - 4MB単位がデフォルト
2. オブジェクトをグループ化
   - Placement Group (PG) への単純ハッシュ対応

   PG識別子 = hash(OBJ識別子) & mask
3. PGの担当データサーバを計算
   - 複製するときは、複数データサーバを算出

   {データサーバ} = F(PG識別子, クラスタ構成, ルール)



- ルール(crush-map)は以下の考慮を上記関数F(CRUSH)にインプット
  - ノード間トポロジによる障害発生単位(failure group)のグループ化
  - 各Diskの容量や処理性能の差を考慮した「重み付け」
  - マップ計算速度とデータサーバ増減時のデータ移動量に応じたハッシュ関数 Fの選択

軽量なルール＋クラスタ構成情報(版数付き)を各サーバに分配、各サーバで、オブジェクトの担当を自律的に計算可能

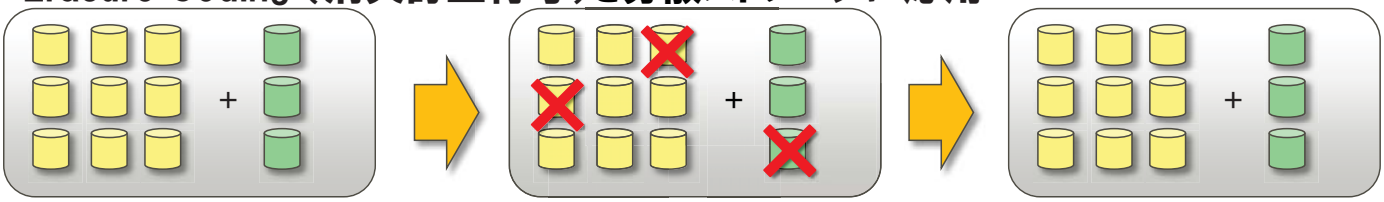
(B) データのdurability、availability

FUJITSU

- 背景
  - Diskやストレージサーバの多重故障時にも、データを失わず(durable)、かつ、データにアクセスできる(available)必要性が高まってきた
  - これまで、処理が単純な3重レプリケーションを利用.
    しかし、データ増で容量効率が問題(コスト増)

例:Hadoop の分散ファイルシステム HDFSは、デフォルトで3重化

  - ➔ 高容量効率のErasure Codingに期待
- Erasure Coding(消失訂正符号)を分散ストレージに応用

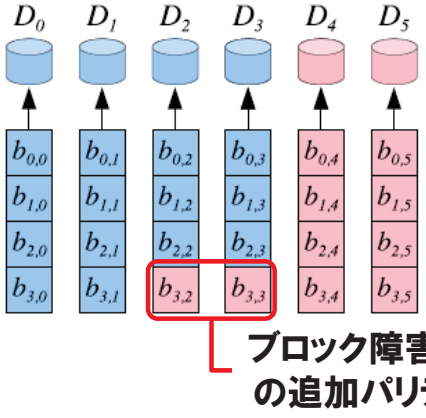


近年の分散 Erasure Coding技術

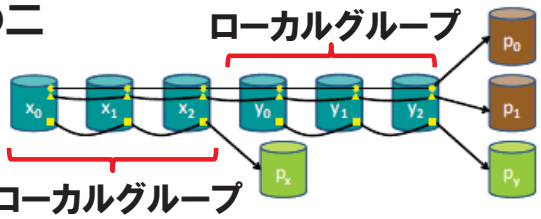
FUJITSU

復旧中の大量データ通信による業務干渉の懸念
➔高容量効率でも、復旧時のデータ転送量の削減が課題

- SD Code (容量効率向上優先)
  - 最小限のパリティ追加で、ディスク障害に加え、特殊ケースの障害を救済
  - 復旧時のデータ通信量は従来RAIDと同等



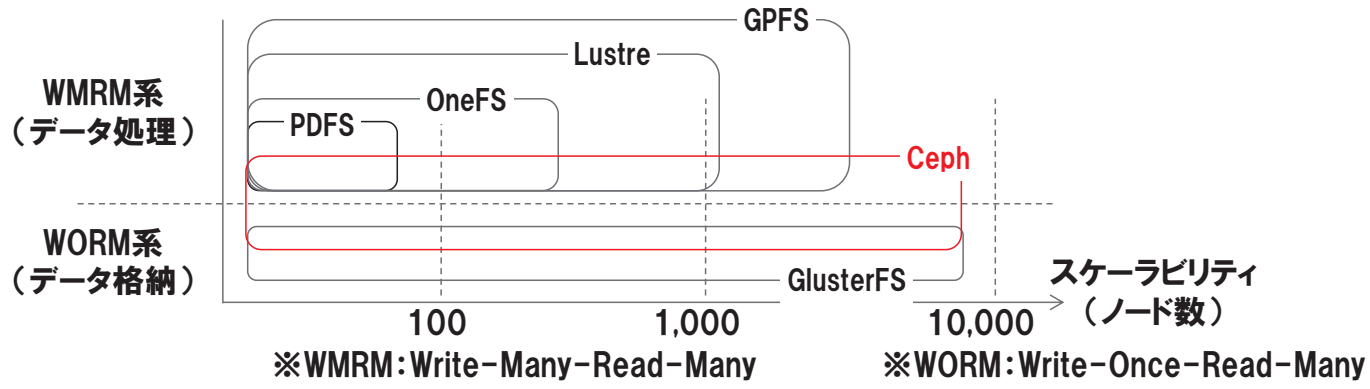
- LRC、Xorbas (データ転送量削減優先)
  - 全ディスクのパリティ(グローバルパリティ)と、ローカルグループのパリティ(ローカルパリティ)の二種のパリティを導入
  - 一重故障復旧でのデータ転送を、関連するローカルグループ内に限定



LRC: Local Reconstruction Code

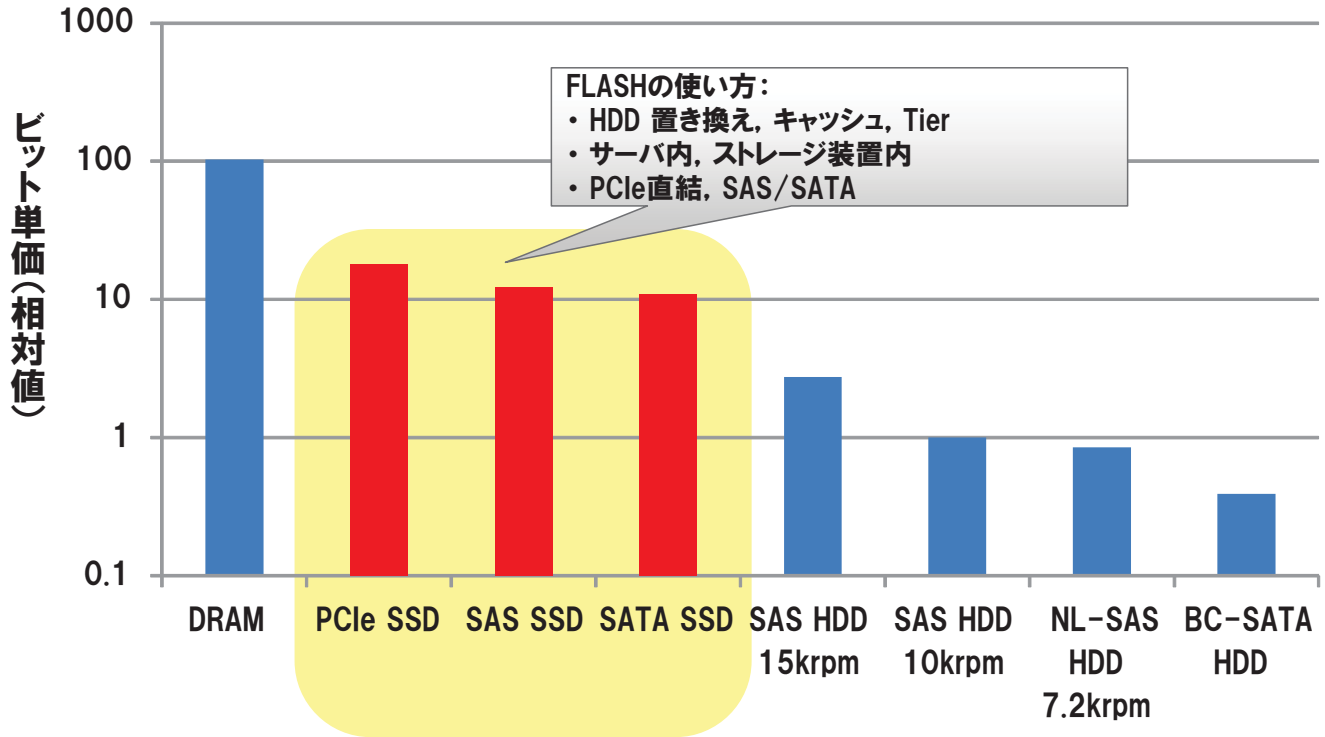
分散ファイルシステムの例

	PDFS	OneFS	Lustre	GPFS	GlusterFS	Ceph
ベンダ	富士通	EMC(Isilon)	Oracle/富士通/OSS	IBM	OSS(Red Hat)	OSS
主要な利用分野	企業向け(基幹/情報系)	大量アーカイブ、大量分析、情報系	HPC	大量アーカイブ、大量分析、HPC、情報系	大量アーカイブ、大量分析	大量アーカイブ、大量分析、情報系
容量	128TB	72TB(Sシリーズ)～10PB(Xシリーズ)	64PB、16PB/file(理論)	2^{99}B、FS数256個、ファイル数2G個、16EB/file	2^{64}B(理論)	2^{64}B(理論)
ノード数	32台	144台	データノードは1,000台	3,794台(HPC) 30台(SoNAS)	2^{64}台(理論)	2^{64}台(理論)
備考	Hadoop対応		富士通はFEFS	用途毎にチューニング	緩い一貫性制御	



DRAMとHDDのギャップを埋めるFLASH/SSD

富士通のIAサーバ(PRIMERGY)のある型の標準価格をもとに試算



レイテンシ: 100nsクラス　　10msクラス

FLASH/DRAMを置き換えるかもしれない ストレージクラスメモリ

FUJITSU

- MRAM, PRAM, ReRAMなどの新しい不揮発性メモリの実用化が近づいている.
- 特徴
  - 不揮発性
  - FLASHに対するアクセス高速性(DRAM並みまで)
  - DRAMに対する低消費電力性

  の他,
  - FLASHに対する書き込み回数(MRAMは事実上制限なし)
  - SRAMに対する電源オフ→オン移行時間の高速性
- 想定される用途
  - FLASH置き換え(PRAM, ReRAM)
  - DRAMの置き換え (MRAM)

ストレージ/データ処理が大きく変わる可能性あり

FUJITSU

ネットワーク関連のトピック

- SDN
- NFV

SDN

FUJITSU

- SDN（Software Defined Networking）とは
  - ソフトウェアによりコンピュータネットワーク全体を集中制御・管理すること．また，そのようなネットワークのこと．
- もう少し言うと…
  - 遅れてやってきた仮想化
    - 多くの場合，物理的なネットワーク（wire once）の上に，仮想的なネットワークをオンデマンドで構成することを可能にする．
  - ハードウェア機器に括り付けられてきたネットワーク上位機能を，機器の外部のソフトウェアで実現可能にする
    - オープン化による競争原理のメリット
    - 実現までの時間を短縮

立場によりさまざまな思惑がありそう

FUJITSU

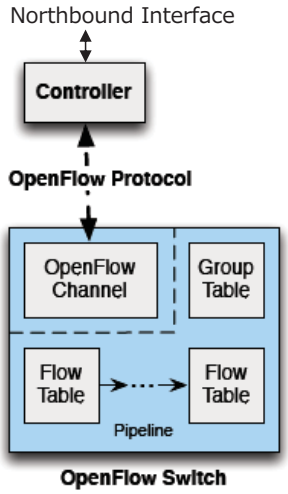
- クラウド提供者，キャリア
  - 特定ベンダの高価な機器(専用ハード)に縛られたくない
    - RootがネックになるTreeトポロジーからの脱却（⇒ Leaf & Spineトポロジー）
    - ネットワークサービスをIAサーバ上で実現
  - 新サービス提供までの時間を短縮したい
  - VLAN ID の4094個制限をきれいに解消したい
  - 広域ネットワークを効率よく利用したい
- 新興ネットワークベンダ
  - 新しい機器APIを普及させることで，土俵を変えてビジネスしたい
- 老舗ネットワークベンダ
  - 世の中の流れ(SDN)に乗り遅れたくない，むしろ先頭を走りたいが，これまで成功してきたビジネス構造・技術を変えたくない
- サーバベンダ
  - ネットワークも含めて仮想化し，ICT機器を一括提供したい
- サーバ仮想化ベンダ
  - ネットワークの仮想化も配下に入れ，ICTの仮想化を一括提供したい

SDN関連の標準化・OSS (1)

FUJITSU

ONF (Open Networking Foundation)

- 新しい機器の機能とAPI (OpenFlowプロトコル) を中心とした標準化団体
- 2011年設立
  (Deutsche Telekom, Facebook, Google, Microsoft, Verizon, Yahoo!).
- 1/3 現在, 117社が参加.
- OpenFlow 1.4まで公開.
- Northbound I/F (Controllerへの指示)
  の制定に遅れ.



フローテーブルエントリの主要項目

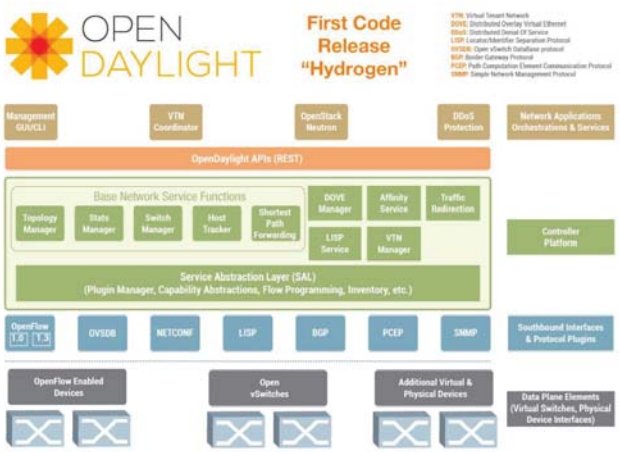
Match Fields	Priority	Counters	Instructions	Timeouts	Cookie

SDN関連の標準化・OSS (2)

FUJITSU

OpenDaylight

- SDN対応機器を制御・管理するソフトウェアをオープンソースとして開発するコミュニティ.
- 2013年4月設立 (Arista Networks, Big Switch Networks, Brocade, Cisco, Citrix, Dell, Ericsson, Fujitsu, HP, IBM, Intel, Juniper Networks, Microsoft, NEC, Nuage Networks, PLUMgrid, Red Hat, VMware)
- Linux Foundation の Collaborative Project の1つ
- OpenFlowだけでなく, BGP-LSのような標準や, ベンダ依存のAPIにも対応可能な抽象化層を提供
  (SAL: Service Abstraction Layer)
- コードとして Northboud IF が提供されるはず



SDN関連の標準化・OSS (3)

FUJITSU

OpenStack Neutron

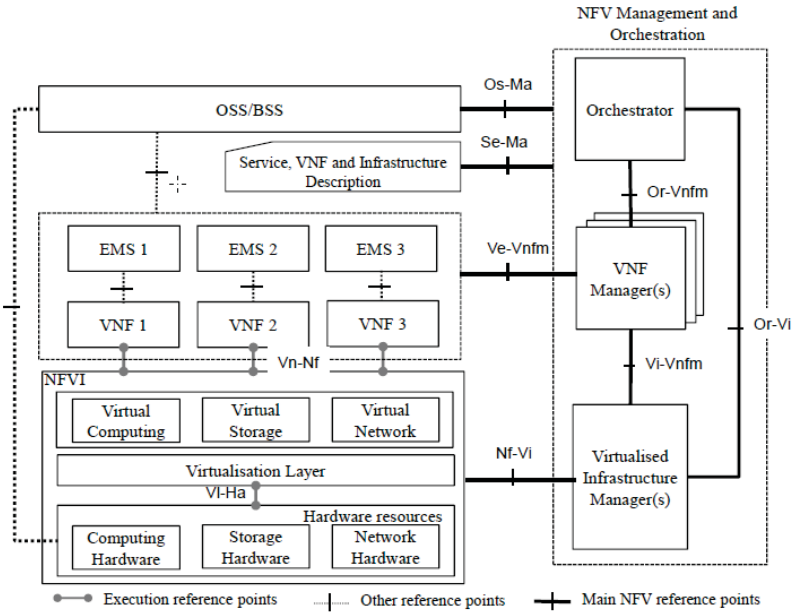
- OpenStack(オープンソースのクラウド管理ソフト)のネットワーク管理コンポーネント(旧称 Quantum)
  - テナント毎に分離された仮想ネットワークを提供
  - プラグインにより複数の実装と接続可能
- 2011年9月の Diablo(D=第4版) で開始, 2012年4月 の Essex (E=第5版) からincubationプロジェクト. Neutron 中心メンバはCisco, Citrix, HP, Nebula, Nicira, Rackspace, Red Hat, Stackopsなど.
- 2012年9月のFolsom (F=第6版)から, 正式な「コア」コンポーネントになった.
- OpenDaylightとつなぐための プラグインあり.

SDN関連の標準化・OSS (4)

FUJITSU

ETSI NFV (Network Functions Virtualization)

- 専用ハードに括り付けだったネットワーク機能を, 汎用サーバ上の仮想マシン (VM) で実現
  - 価格低減
  - サービス提供までの時間短縮
  - ハード利用の効率化
- キャリア主導
  - 仮想化／クラウド技術を キャリア内に導入
- VNF (Virtualized Network Function) の例
  - Firewall
  - SIP server
  - 携帯網のコアネットワーク (LTEのEPCなど)
  - CDN (Content Delivery Network)
- キャリア以外に一般化



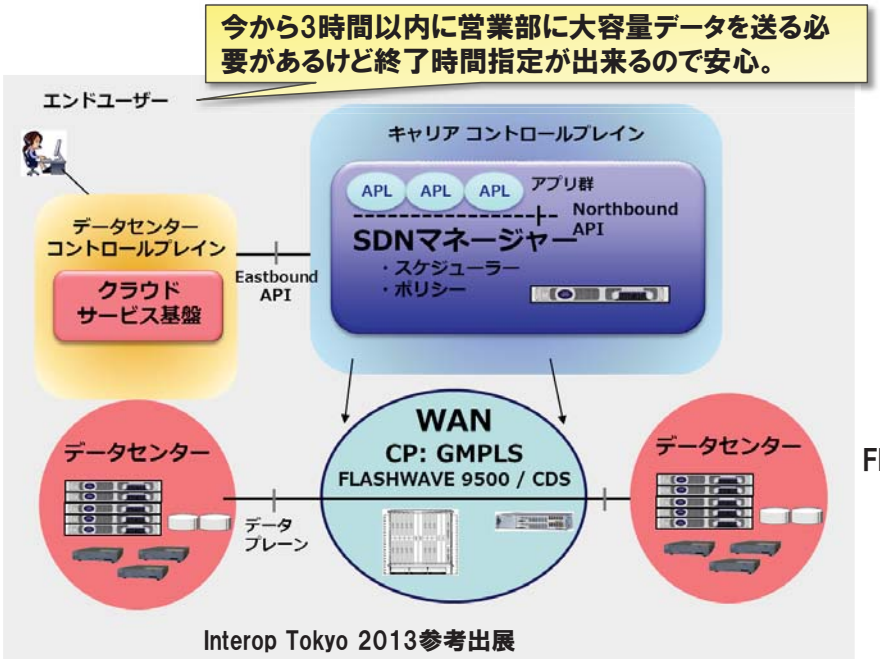
ETSI: European Telecommunications Standards Institute

各社の対応

- Cisco eXtensible Network Controller（XNC）
  - OpenFlow標準と Cisco独自の onePK に対応
  - OpenDaylightのアーキテクチャに基づく
    - 公開可能な部分をOpenDaylightへ提供
- VMware NSX
  - VMwareの Software Defined Datacenter構想の一環
  - ネットワークと（ネットワーク）セキュリティの仮想化基盤
  - VMwareの VXLAN と，買収した Nicira の STT の両方を扱う（どちらもトンネリングプロトコル）.
- Microsoft
  - トンネリングプロトコルとして，NVGREを提案.

クラウド対応SDN　WANソリューション例

- 大容量データ転送用業務アプリ利用者が、サービス利用画面に転送するデータ容量，送付先，時間等を入力するだけでストレスのないデータセンター間のデータ転送サービスの利用が可能です。
- クラウド側からWAN側に必要帯域をオンデマンドでリクエストするだけで迅速なネットワーク設定変更を可能とします。



Interop Tokyo 2013参考出展

FLASHWAVE9500：パケット統合光伝送システム
システム性能
- 480G SONET/SDH
- 44波長多重 x 10G (per ROADM degree)
- 最大 8 ROADM degree

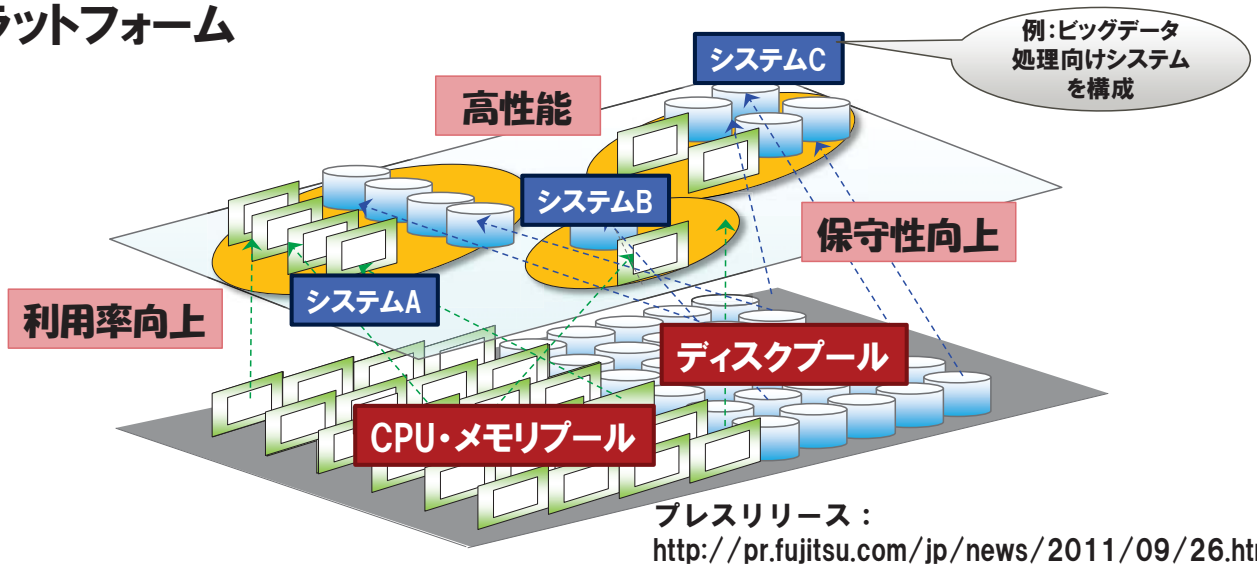
サービスインタフェース：
- 10GbE の場合，最大48ポート

サーバ関連のトピック

- Software Defined Server

Software Defined Server （資源プール化アーキテクチャ）

■ 高性能かつ柔軟な構成の物理マシンをオンデマンドに提供するプラットフォーム



プレスリリース：
http://pr.fujitsu.com/jp/news/2011/09/26.html

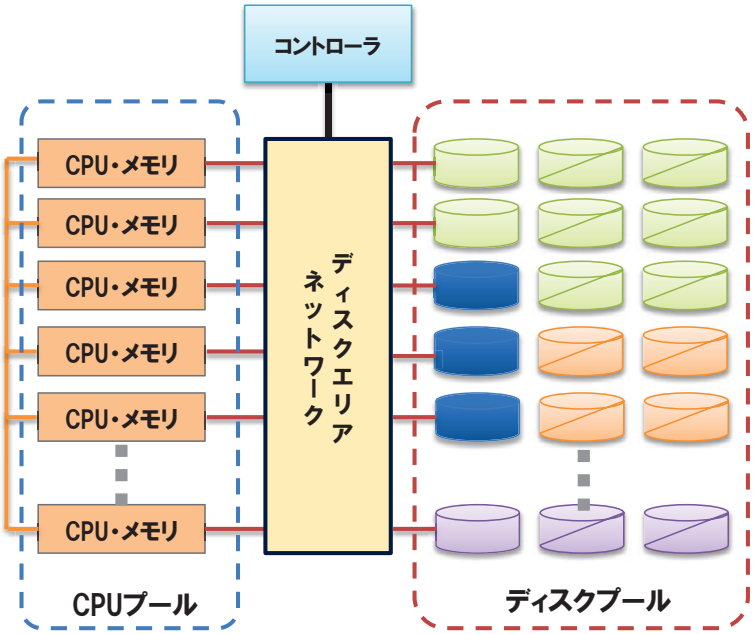
■ ハードウェアリソースをコンポーネントごとにプール化

■ 各リソース間を高速なインターコネクトで接続

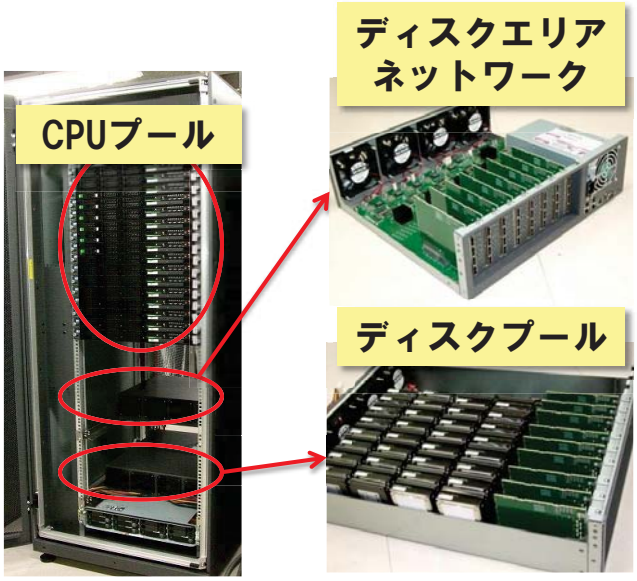
■ リソースプールからユーザの要求に応じて切り出したリソースを提供

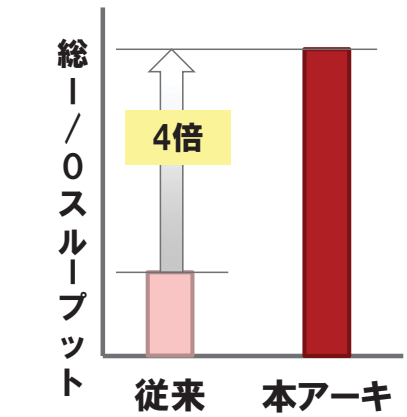
Software Defined Server プロトタイプ

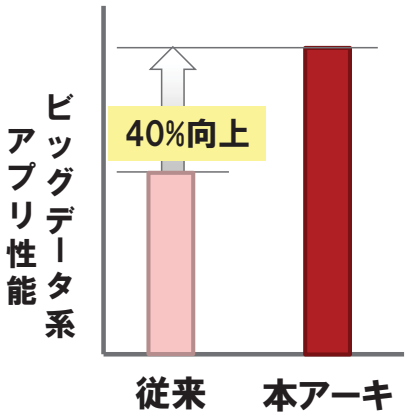
- CPUプール(CPU・メモリ)をディスクレスサーバにより構成
- ディスクプールをディスクドライブ群で構成
  - HDDプール，SSDプール
- ディスクエリアネットワーク(DAN)により，サーバに接続したディスクドライブは，機能的にも，性能的にも，通常のローカルディスクとして見える



ハードウェアプロトタイプ（第1世代）

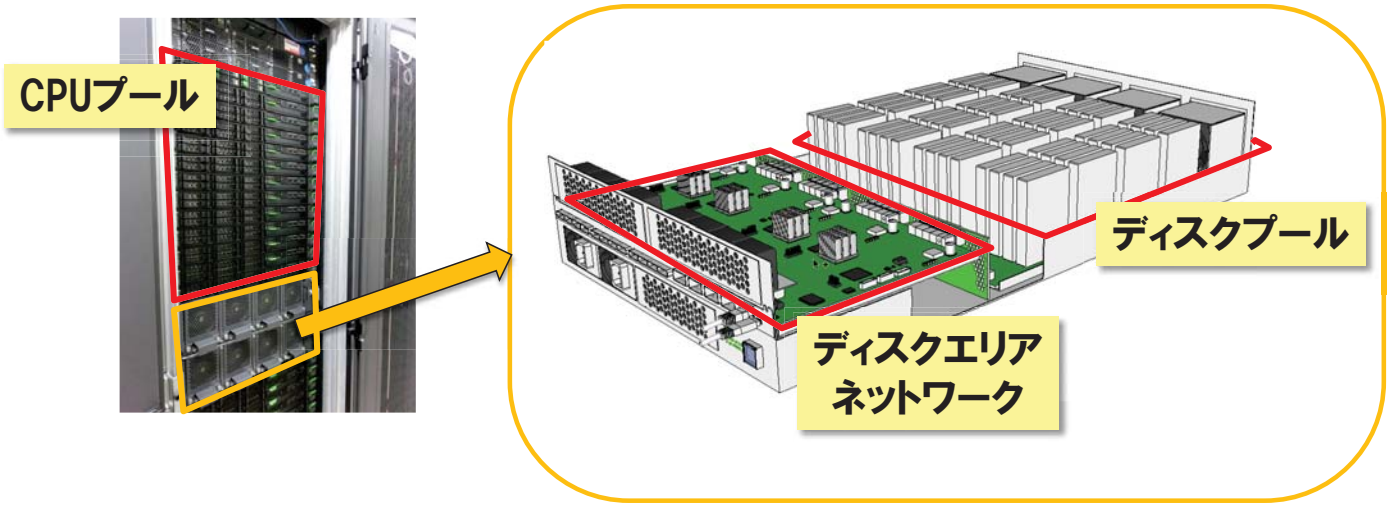






ハードウェアプロトタイプ（第2世代）

FUJITSU



- 配線数減
- カスケード機能によるスケーラビリティ向上
- メンテナンス性向上

Software Defined Serverを物理IaaSに利用

FUJITSU

従来のホスティングサービスと，仮想IaaSの良いとこどり

IaaS: Infrastructure as a Service

	従来のホスティングサービス	仮想IaaS	今回の物理IaaS
物理サーバの占有・設定変更	○	×	○
物理構成の配備時変更	○	×	○（※）
物理構成の稼働中変更	×	×	○（※）
サーバ配備時間	数日	すぐ	すぐ（10分）
サーバ配備・撤収指示の自動化対応	×	○	○

物理サーバを占有したい理由の例：
- 安定した性能（他VMの影響なし）
- Hypervisorによる性能劣化なし，メモリオーバヘッドなし
- Hypervisorバグによるセキュリティリスクなし
- VM未対応のハード利用
- ハード設定（BIOS等）の変更
- Hypervisorの開発・テスト

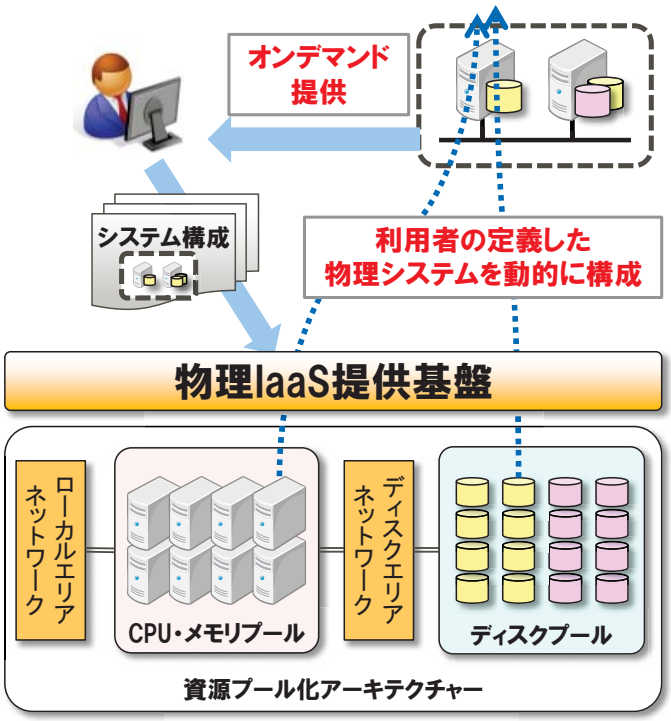
※: 今回はローカルディスクの追加・削除に対応

プレスリリース
http://pr.fujitsu.com/jp/news/2013/07/4.html

物理IaaS基盤

FUJITSU

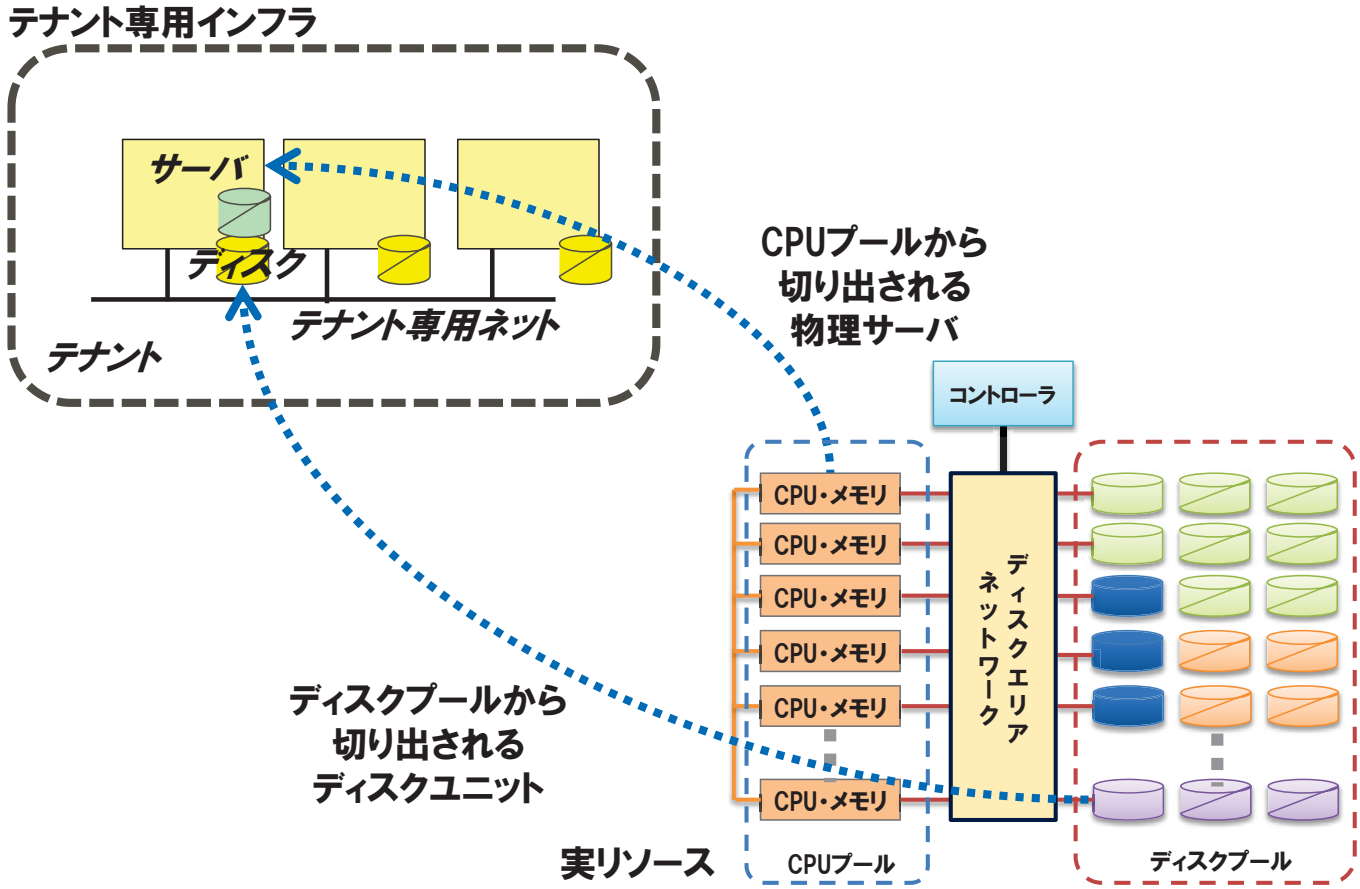
- 資源プールから物理サーバを高速に構成・配備
  - 利用者のニーズに合わせた**物理サーバをオンデマンド提供**可能な物理IaaS基盤を開発
  - 「FUJITSU Software ServerView Resource Orchestrator」を拡張
  - 稼働前，稼働中の**物理サーバに対しローカルディスクの追加・削除が可能**
  - 物理サーバをテナント専用ネットワーク（LAN）に接続



「FUJITSU Software ServerView Resource Orchestrator」:
ダイナミックリソース管理ソフトウェア。ICTリソースの有効活用と運用・管理の効率化を実現するプライベートクラウドの基盤ソフトウェア。
http://software.fujitsu.com/jp/ror/

物理IaaSサービス

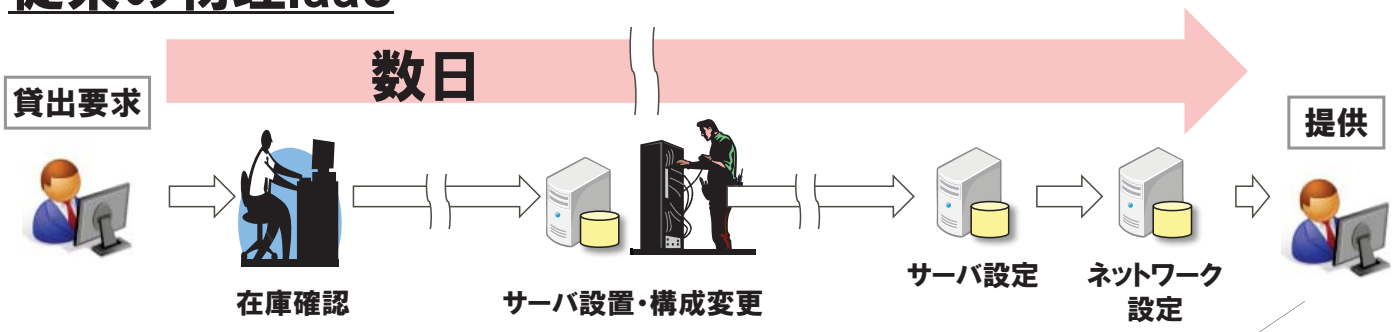
FUJITSU



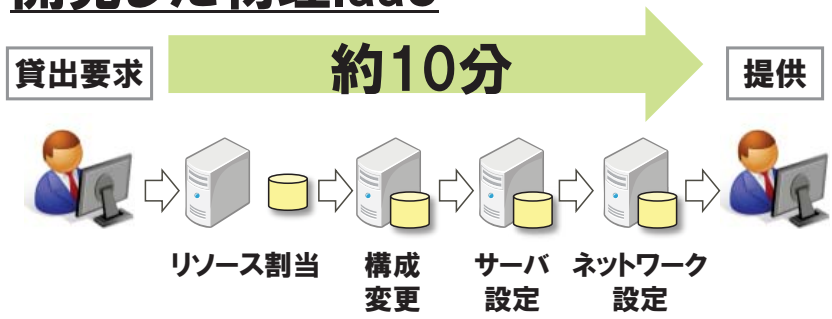
効果(１)

FUJITSU

従来の物理IaaS



開発した物理IaaS

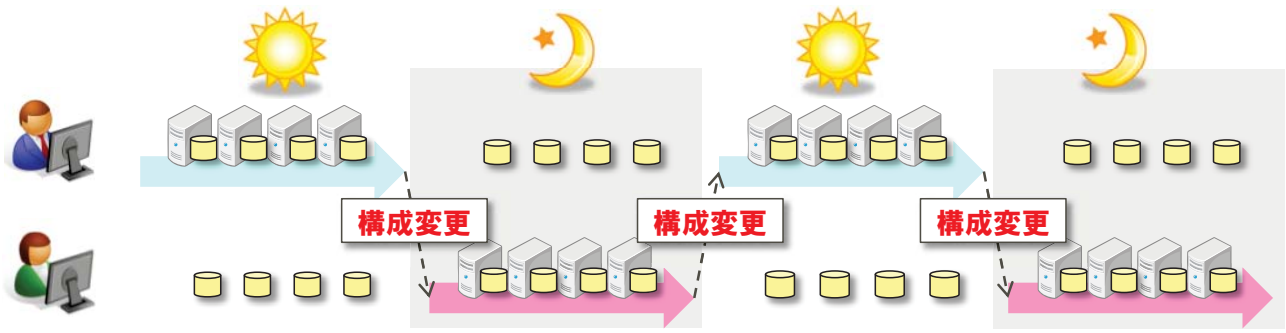


提供時間を大幅短縮

効果(２)

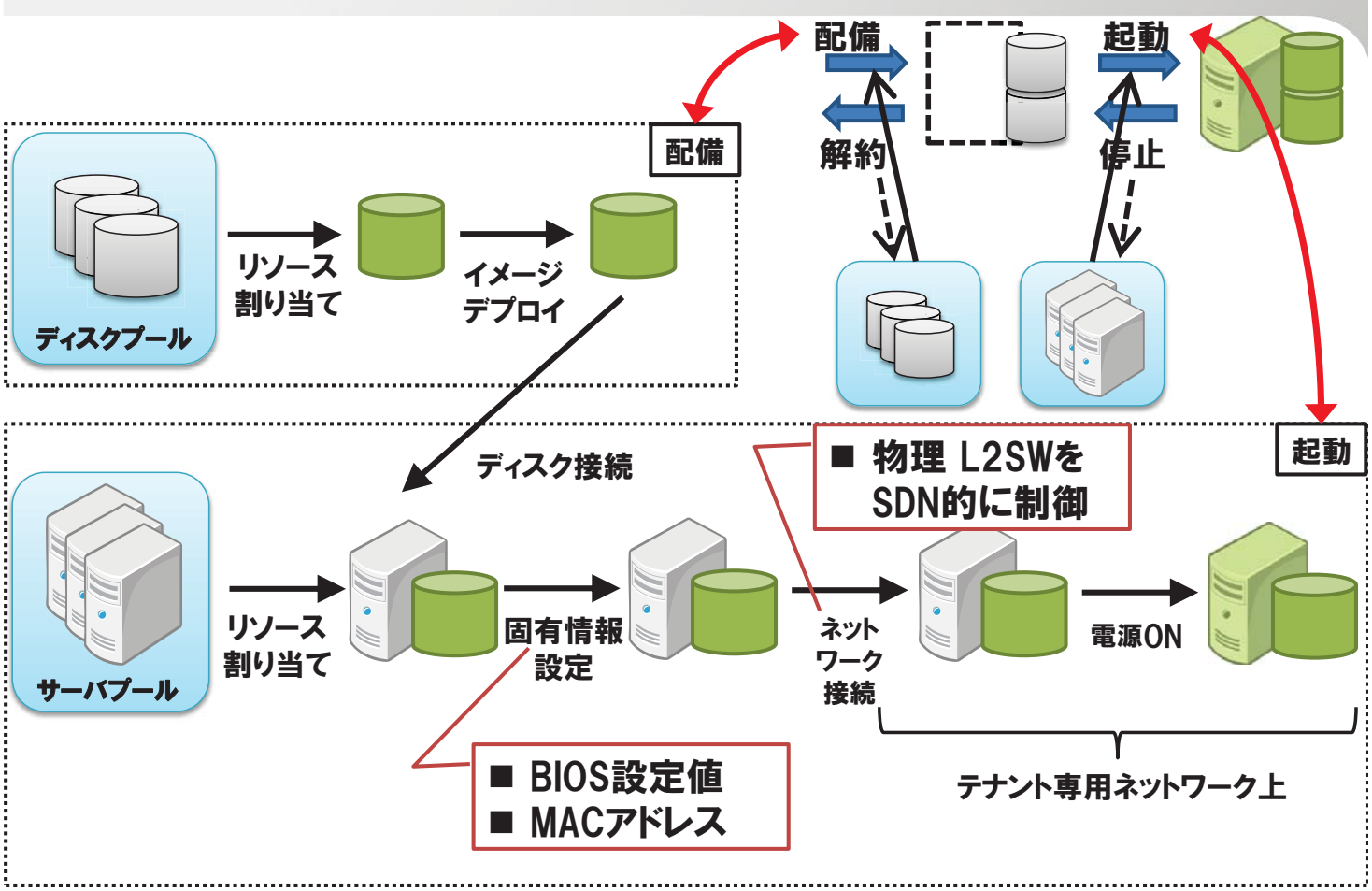
FUJITSU

- ハードウェアリソースの利用効率向上
  - サーバの電源オフを検出すると、利用者のディスクを物理サーバから切断．物理サーバリソースだけ回収し、再利用．
  - ユースケース
    - 昼と夜に別々の利用者がバッチ処理を行う（下図）
    - 開発済のシステム(テスト用，旧版など)をすぐ動かせる状態で保存しておく



ホスティングサービスでは，物理サーバを占有したまま．
構成変更時間が大きく短縮されたため可能になった．
この例では，サーバ利用効率が2倍になる．

ライフサイクル管理(起動シーケンス)



今後のクラウドの方向

- Software Defined System

Software Defined System

FUJITSU

リソース毎にばらばらに提供

- サーバ仮想化
- ストレージ仮想化
- ネットワーク仮想化

システム全体をセルフサービス・オンデマンドで提供

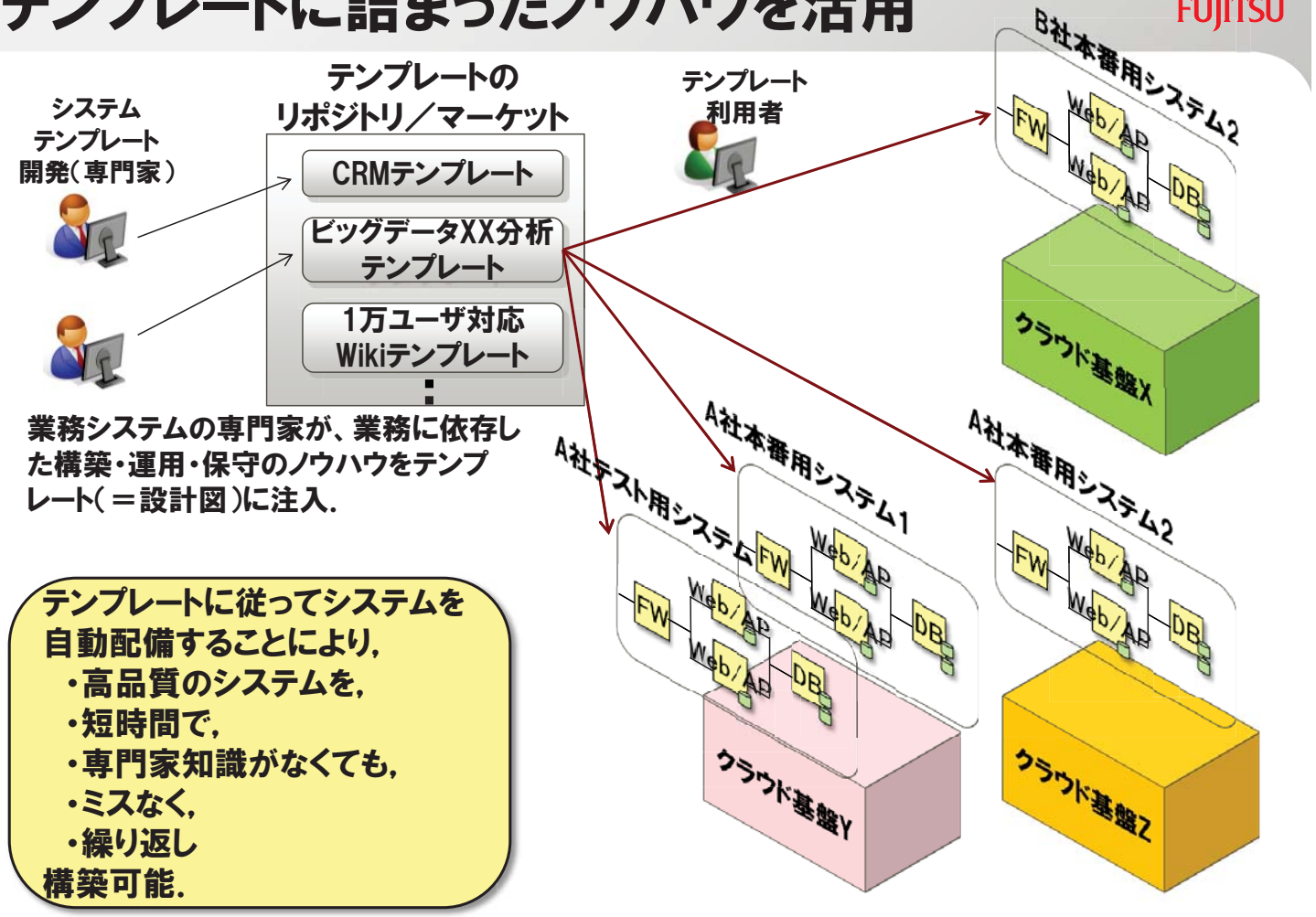
Software Defined System

- サーバ仮想化
- ストレージ仮想化
- ネットワーク仮想化

参考: DMTF CIMI 標準の System
(CIMI: Cloud Infrastructure Management Interface)
FUJITSU Cloud IaaS Trusted Public S5の仮想システム

テンプレートに詰まったノウハウを活用

FUJITSU



業務システムの専門家が、業務に依存した構築・運用・保守のノウハウをテンプレート(＝設計図)に注入.

テンプレートに従ってシステムを自動配備することにより,
- 高品質のシステムを,
- 短時間で,
- 専門家知識がなくても,
- ミスなく,
- 繰り返し

構築可能.

Software Defined System関連技術

FUJITSU

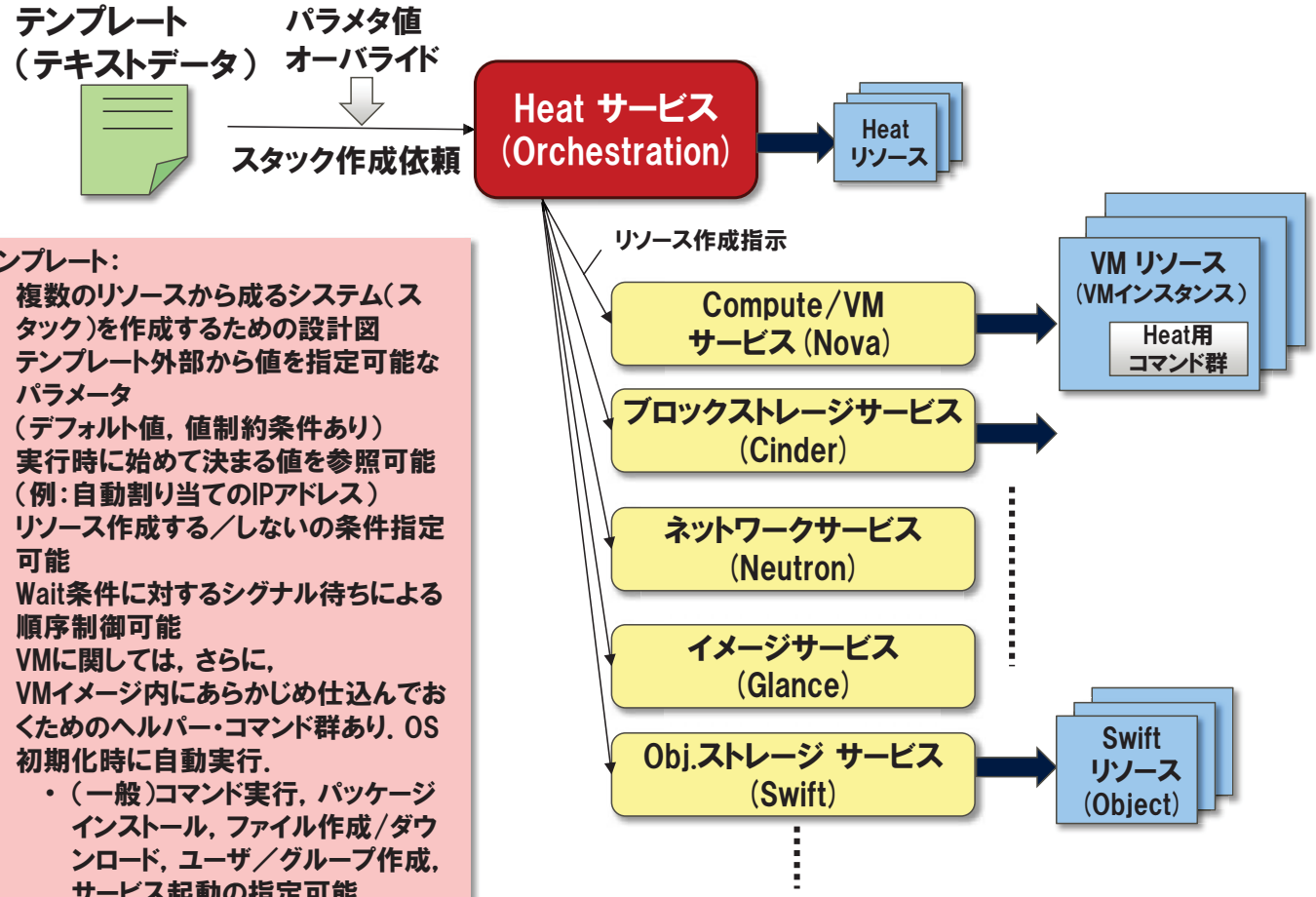
	IaaSリソース群の自動構築支援	VM内部のソフトスタックの自動構築支援
CloudFormation (AWS)	✓	
OpenStack Heat (OSS)	✓	
Chef (OSS/Chef社)		✓
Puppet (OSS/Pupet Labs)		✓
OASIS TOSCA（標準）	✓	✓
PureApplication System /Workload Manager(IBM)	✓	✓

- チェックマークがないところでも，自分で書けば対応可能．
- Heatのテンプレートは，CloudFormation互換と，独自のHoT．
- Cloudformationや Heatは，ChefやPuppetと組み合わせて利用可能．
- OASIS TOSCAが，OpenStack Heatに提案されたが今のところ受け入れられていない模様．
- 現在は，構築に重点があるが，ライフサイクル全体の管理がターゲット．

IaaS: Infrastructure as a Service

OpenStack Heat / AWS CloudFormation

FUJITSU



テンプレート:
- 複数のリソースから成るシステム(スタック)を作成するための設計図
- テンプレート外部から値を指定可能なパラメータ
  (デフォルト値，値制約条件あり)
- 実行時に始めて決まる値を参照可能
  (例:自動割り当てのIPアドレス)
- リソース作成する／しないの条件指定可能
- Wait条件に対するシグナル待ちによる順序制御可能
- VMに関しては，さらに，VMイメージ内にあらかじめ仕込んでおくためのヘルパー・コマンド群あり．OS初期化時に自動実行．
  - (一般)コマンド実行，パッケージインストール，ファイル作成/ダウンロード，ユーザ／グループ作成，サービス起動の指定可能

まとめ

FUJITSU

- ■（並列）分散技術の利用範囲拡大
  - ≪単体性能向上の鈍化とシステムの大規模化≫
    - ■ 分散ストレージ
- ■ 専用装置から汎用ハード＋ソフトへ
  - ≪コスト削減，ソフトのアジャイル開発≫
    - ■ 分散ストレージ
    - ■ OpenFlow
    - ■ NFV（Network Function Virtualization）
- ■ Software Defined XX ＝ サービス化 ⇒クラウド
  - ≪変化へのアジャイル対応≫
    - ■ SDN（Software Defined Networking）
    - ■ Software Defined Server
    - ■ Software Defined System

FUJITSU

shaping tomorrow with you