OpenStack Days Tokyo 2016
7th July 14:40 – 15:20 4F B-2

FUJITSU
shaping tomorrow with you

Newtonでのベアメタルは何が違う？ ～開発内容の紹介とVMとの機能面のギャップについて～

富士通株式会社 プラットフォームソフトウェア事業本部 Linux開発統括部

古川 勇志郎
椎名 宏徳

自己紹介(1/2)

FUJITSU

古川 勇志郎 / Yushiro Furukawa

- 富士通株式会社 所属
  - プライベートクラウドの運用管理ソフトウェアの開発
  - FUJITSU Cloud Service K5の開発
- 現在はOpenStack Neutron/Ironic にて、以下の開発に従事:

1. ベンダープラグイン(ML2)の設計開発
2. ベアメタルサーバのマルチテナント対応
3. ベアメタルサーバのセキュリティグループ対応
4. ネットワークパケットログ採取APIの設計開発
5. FWaaS v2の開発

自己紹介(2/2)

FUJITSU

椎名 宏徳 / Hironori Shiina

- 富士通株式会社 所属
  - ミッションクリティカル分野でのミドルウェア開発を経験
- 現在はOpenStack Nova/Ironic にて、以下の開発に従事:

1. NovaにNMI送信のAPIを追加
2. ベアメタルサーバのマルチテナント対応
3. ベアメタルサーバのコンソール対応

アジェンダ

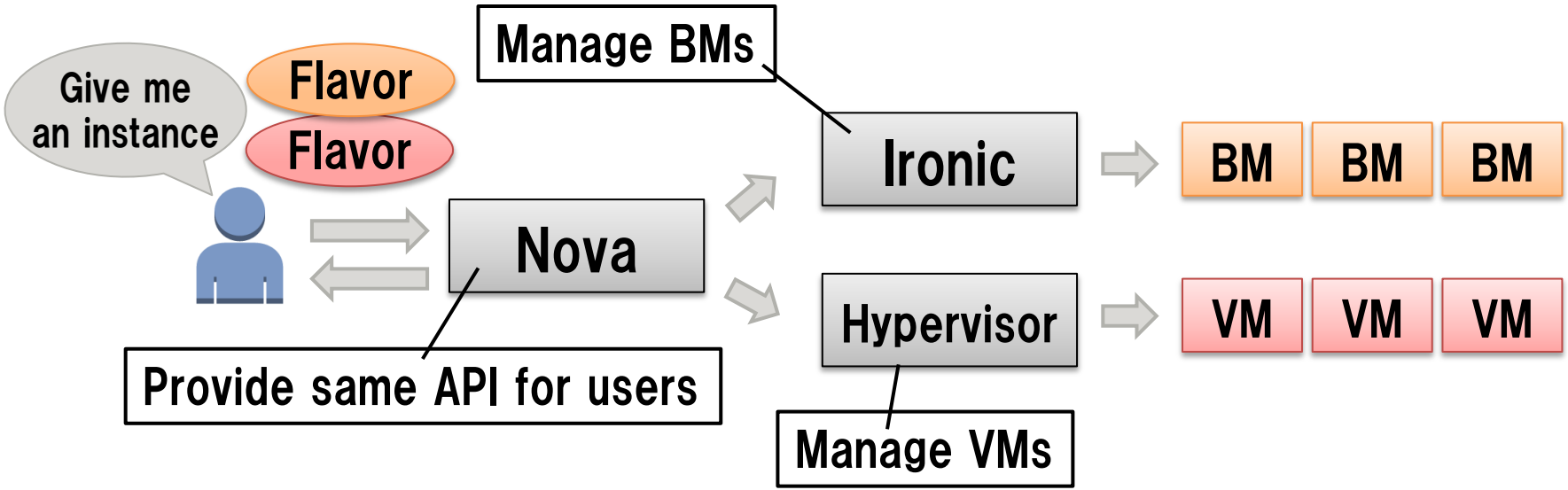
FUJITSU

1. OpenStack Ironicとは
2. 目指すべき世界に向けて
3. 開発内容紹介
   - 導入フェーズ(配備)
   - 正常時運用フェーズ
   - 異常時運用フェーズ
4. コミュニティ状況

OpenStack Ironicとは？

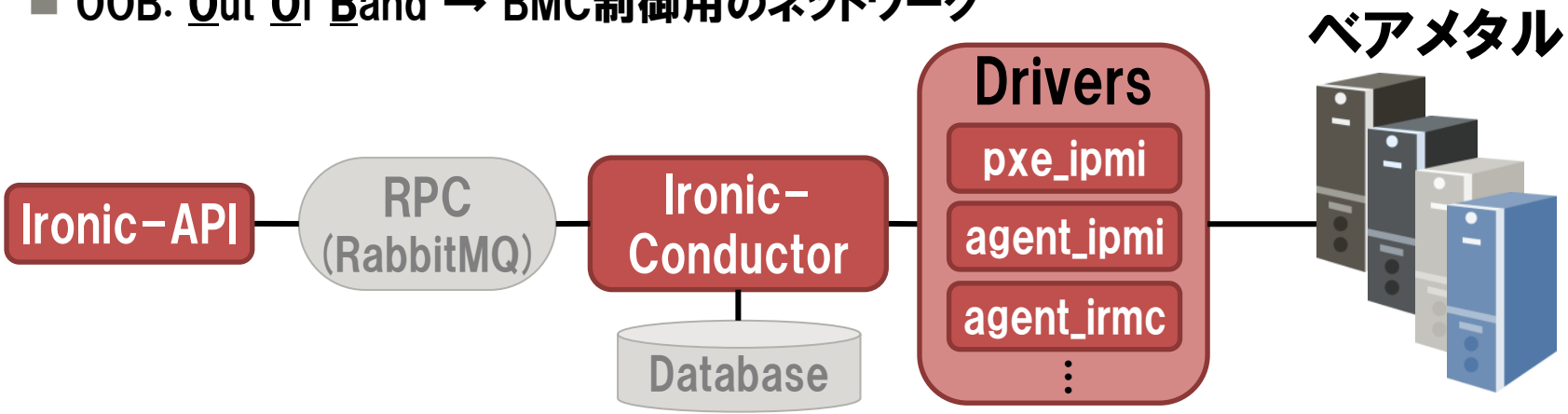
FUJITSU

- **物理マシンの配備を行うコンポーネント**
  ベアメタル（BM: Bare Metal）プロビジョニング
- **テナント利用者はNova経由でベアメタルサーバを配備可能**



Ironicのコンポーネント

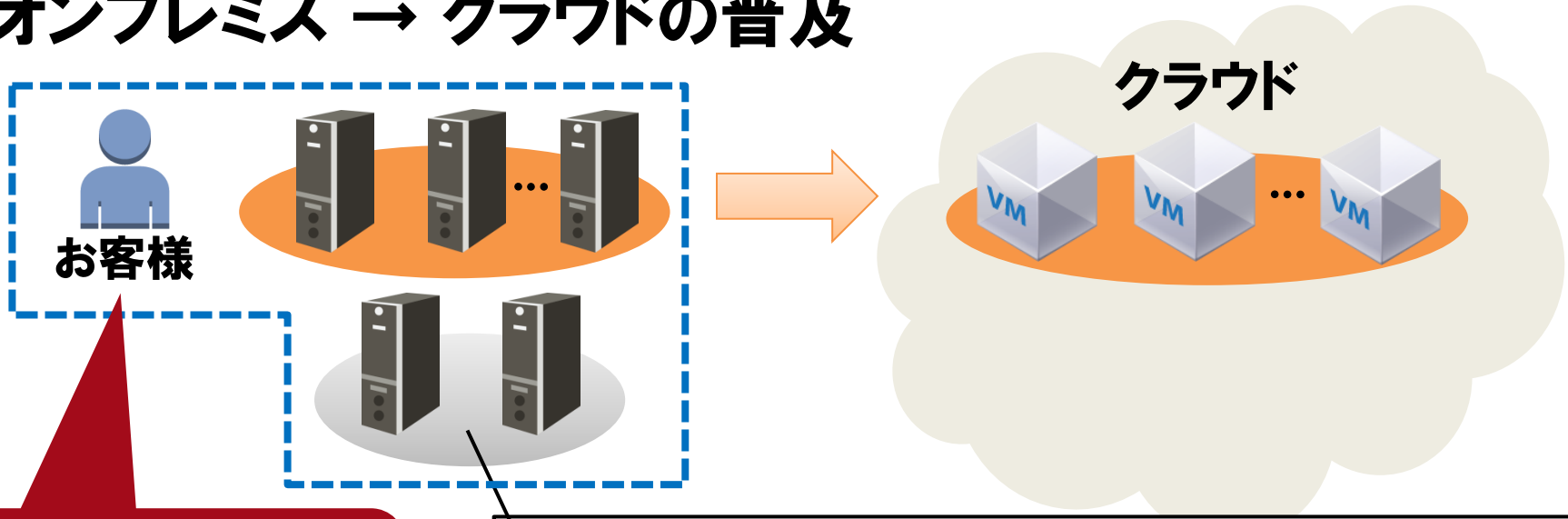
- Ironic-API ：REST APIの受け口。管理者のみ利用可能
- Ironic-Conductor: Ironicの中枢。ドライバ呼び出しやDB操作を行う
- Drivers: サーバ毎の差異を吸収する仕組み。OOB経由でBMCを制御
  - BMC: Baseboard Management Controller →サーバ管理用コントローラ
  - OOB: Out Of Band → BMC制御用のネットワーク



なぜベアメタルサービスが必要なのか？

FUJITSU

■オンプレミス → クラウドの普及



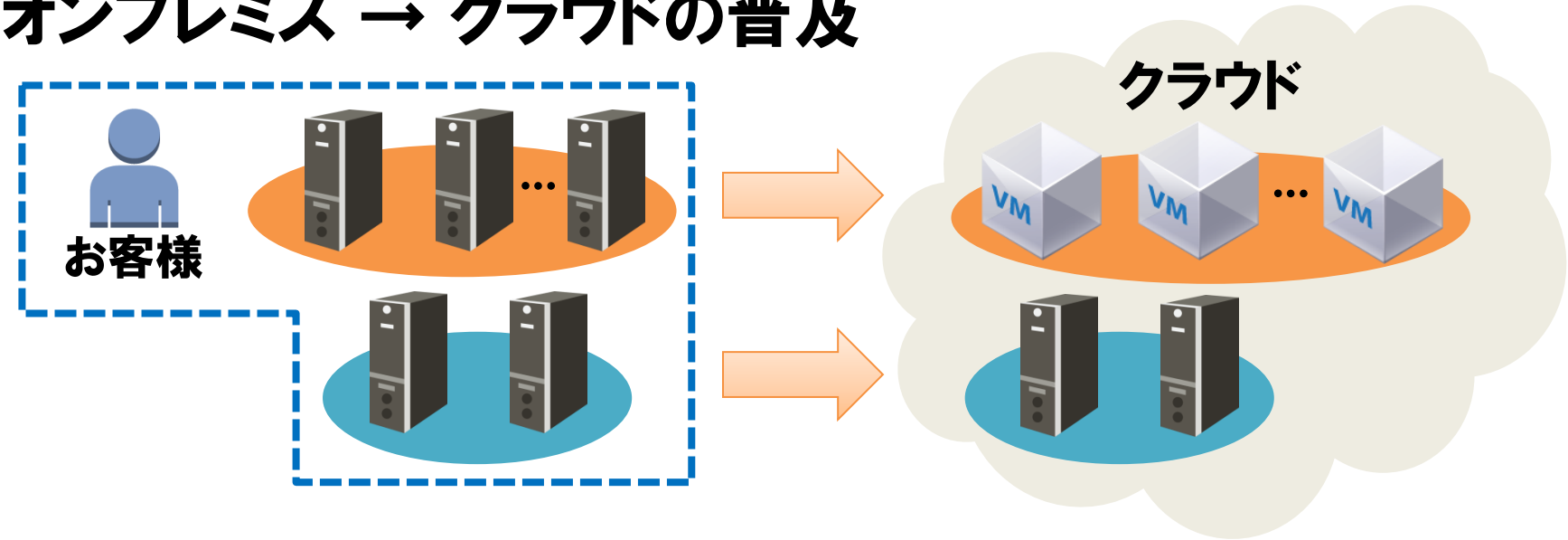
二つの管理を
一本化したい

仮想マシン上では運用に適さない業務
（CPU負荷 or ディスクI/O性能が重視される
ワークロードなど）

なぜベアメタルサービスが必要なのか？

FUJITSU

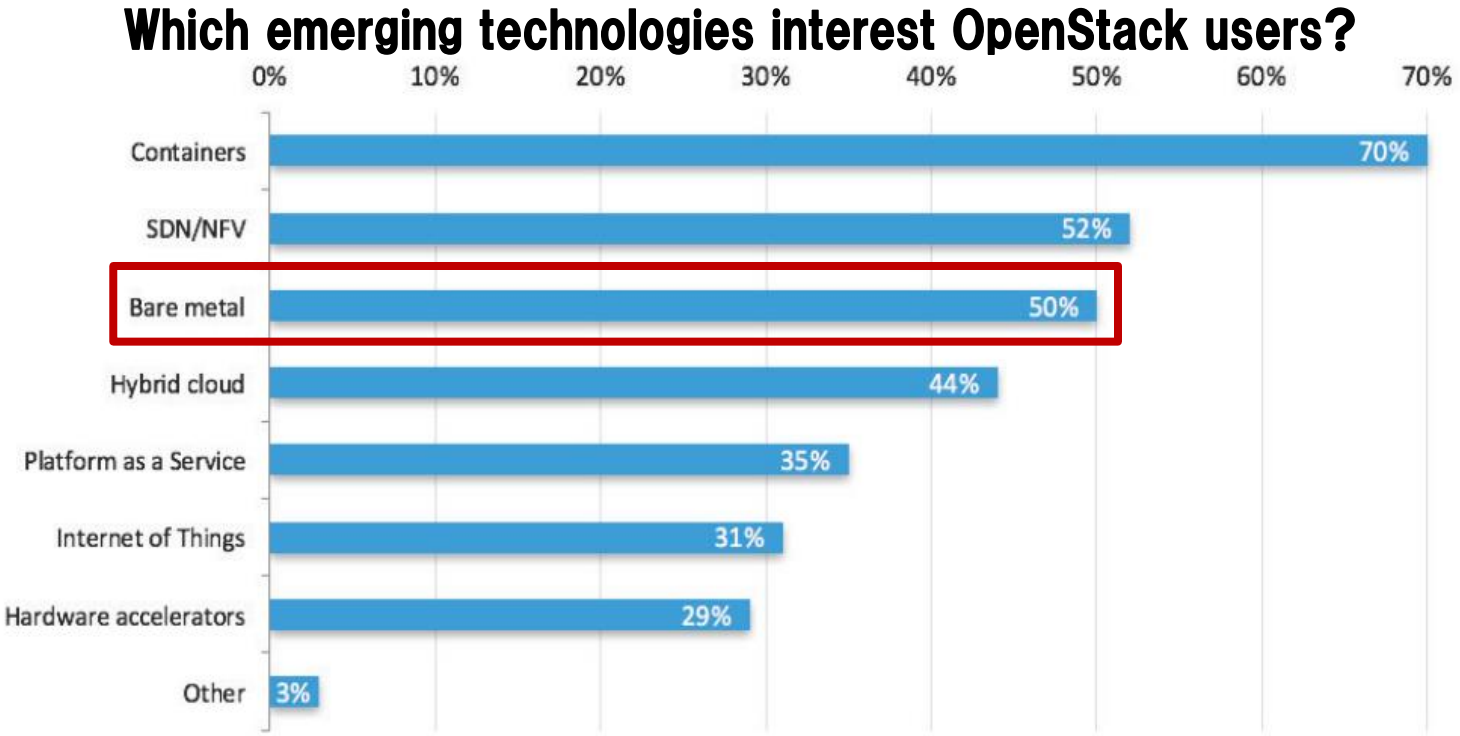
■オンプレミス → クラウドの普及



管理コストの削減・クラウドが持つ便利な機能の享受

【補足】利用者の関心

公式の User Surveyより［1］

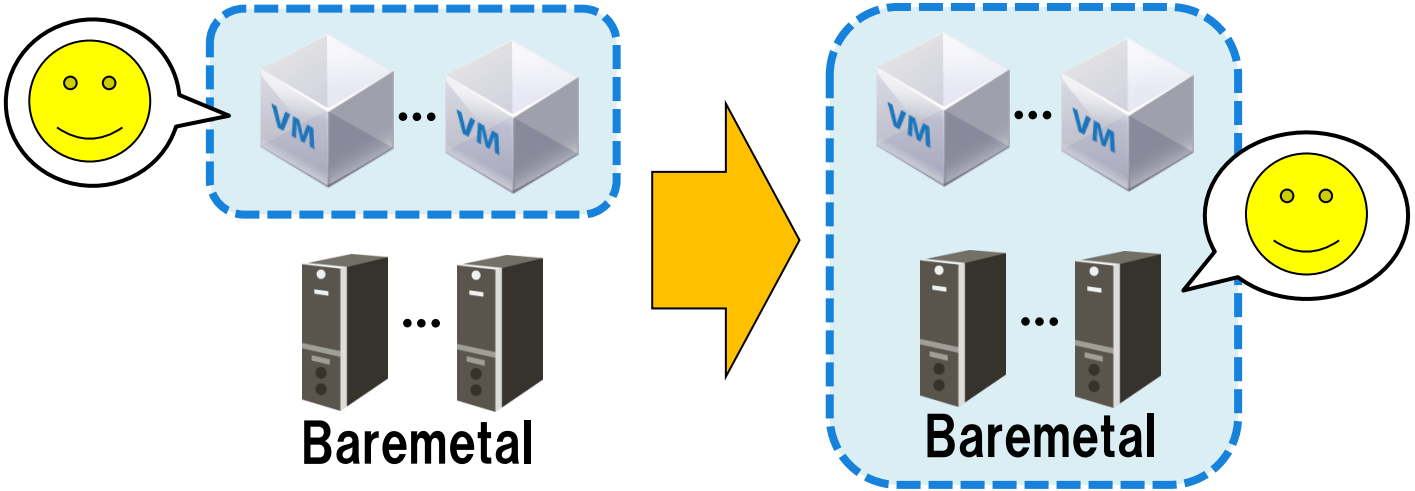


[1] https://www.openstack.org/assets/survey/April-2016-User-Survey-Report.pdf#page=23 Figure 2.5 n=1131

目指すべき世界に向けて

我々が目指すべき世界

FUJITSU



マルチテナント | SecurityGroup | 仮想Volume | Graceful Shutdown | Console

仮想/クラウドから出た技術をベアメタルでも実現する

FUJITSU

開発内容紹介

- ☑ 導入フェーズ(配備)
  - マルチテナント、LAG(冗長化)
  - SAN対応(Cinder連携)
- ☐ 正常時運用フェーズ
- ☐ 異常時運用フェーズ

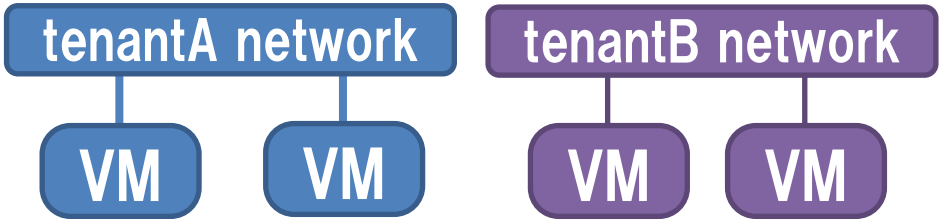
マルチテナント対応状況

FUJITSU

仮想（VM）

- テナントごとに分離された任意のネットワーク上に配備可能
- 分離方式として以下：

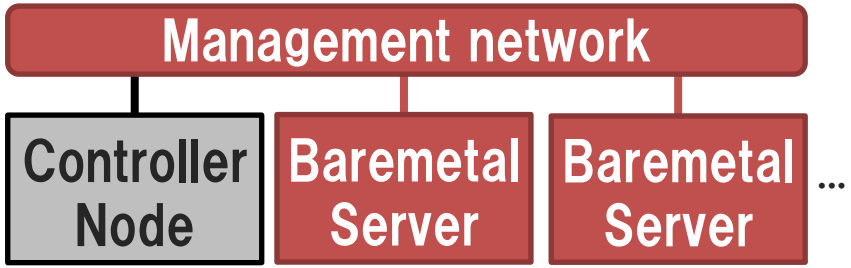
  VLAN, VXLAN, gre, geneve



【Mitaka時点】

物理（BM）

- Controllerと同じネットワーク上にのみ配備可能
- Flatネットワークのみのサポート（テナント間分離なし）

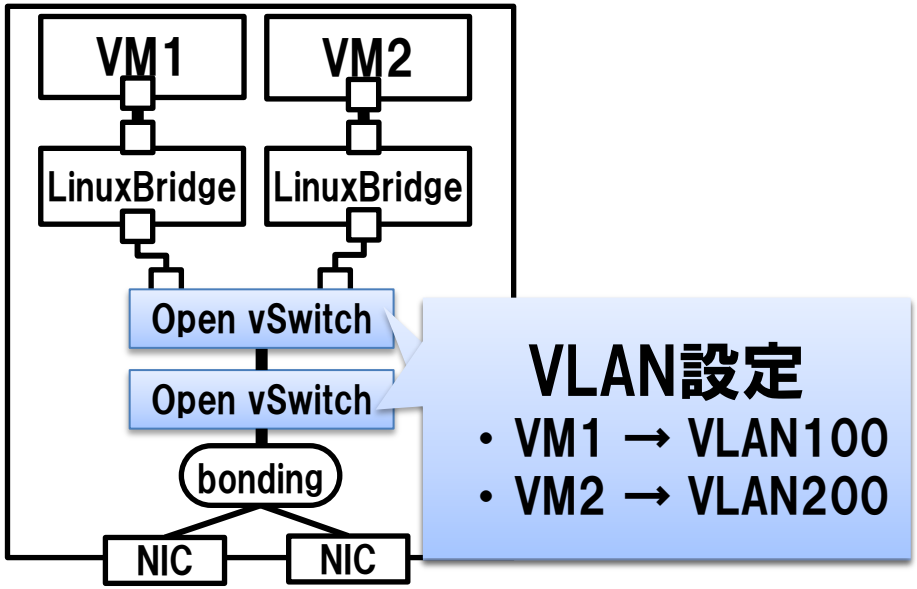


開発内容: マルチテナント対応

FUJITSU

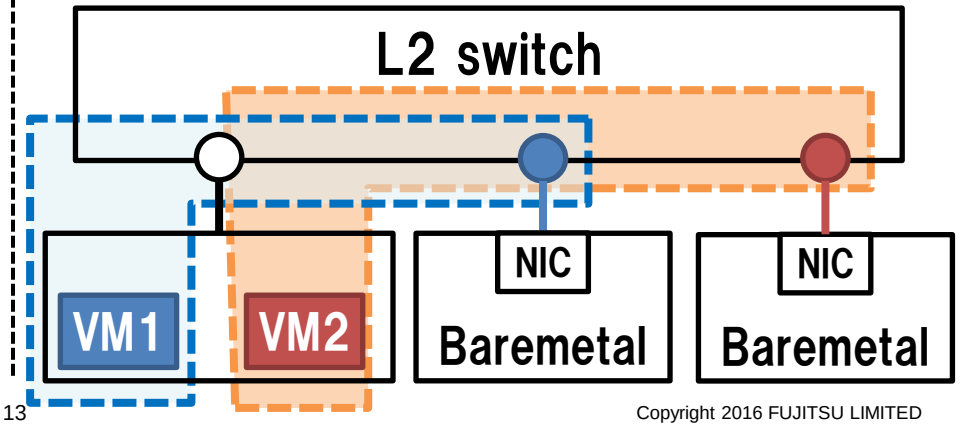
仮想(VM)

- Open vSwitch上にVLAN設定



物理(BM)

- 物理スイッチの特定のポートに対して VLAN (Untagged) 設定
- NeutronのML2プラグインを利用
- ポイント:ネットワークフリップ

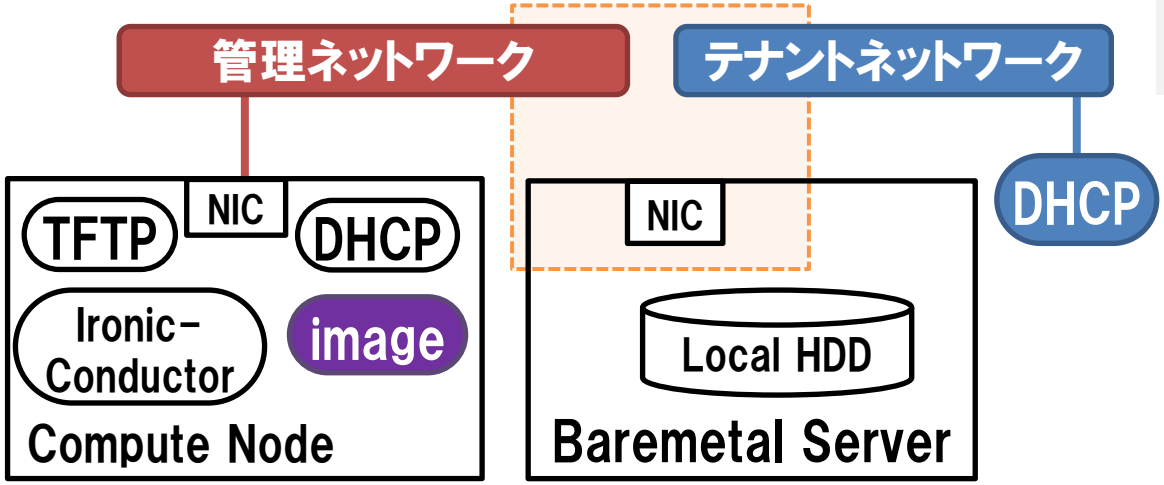


ネットワークフリップ（1/3）

FUJITSU

配備中にネットワークを切り替える：

1. **管理ネットワークへ接続**
2. **接続解除**
3. **テナントネットワークへ接続**



管理ネットワーク上にあるもの

- **TFTPサーバ**
- **ironic-conductor**
- **Deploy image（配備用）**
- **Boot image（お客様用）**

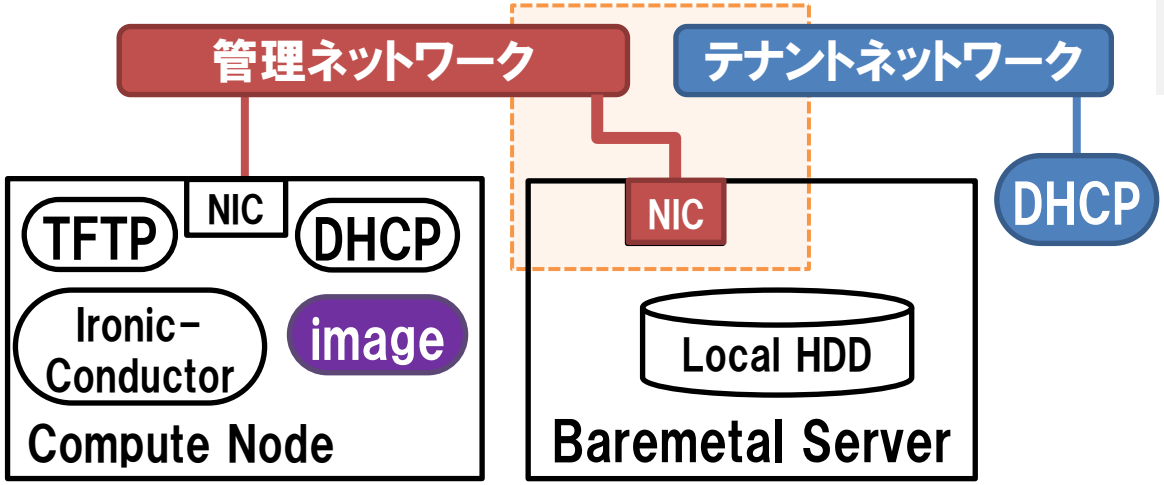
配備中に一旦
接続する必要がある

ネットワークフリップ（1/3）

FUJITSU

配備中にネットワークを切り替える：

1. **管理ネットワークへ接続**
2. **接続解除**
3. **テナントネットワークへ接続**



管理ネットワーク上にあるもの

- **TFTPサーバ**
- **ironic-conductor**
- **Deploy image（配備用）**
- **Boot image（お客様用）**

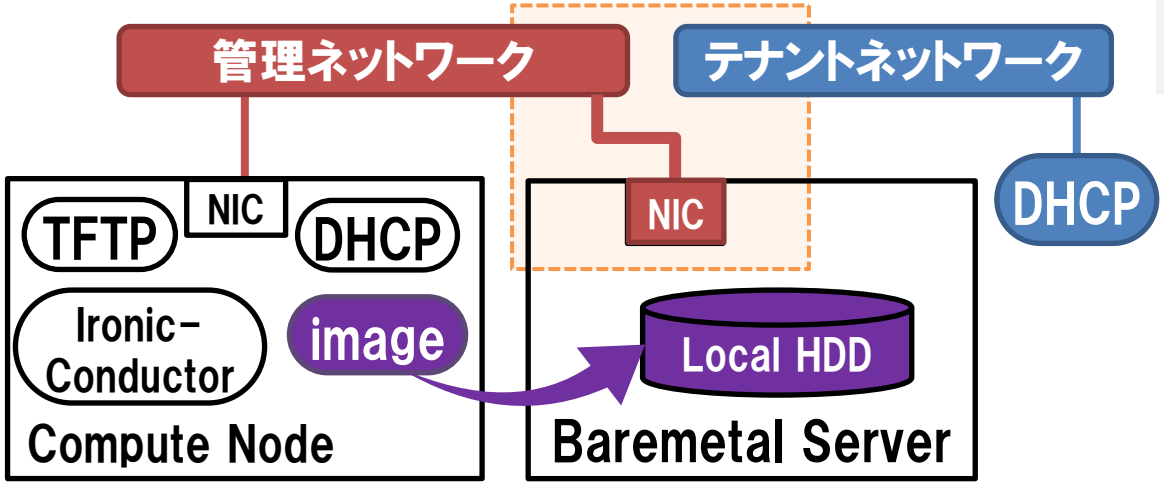
配備中に一旦
接続する必要がある

ネットワークフリップ（1/3）

FUJITSU

配備中にネットワークを切り替える：

1. **管理ネットワークへ接続**
2. **接続解除**
3. **テナントネットワークへ接続**



管理ネットワーク上にあるもの

- **TFTPサーバ**
- **ironic-conductor**
- **Deploy image（配備用）**
- **Boot image（お客様用）**

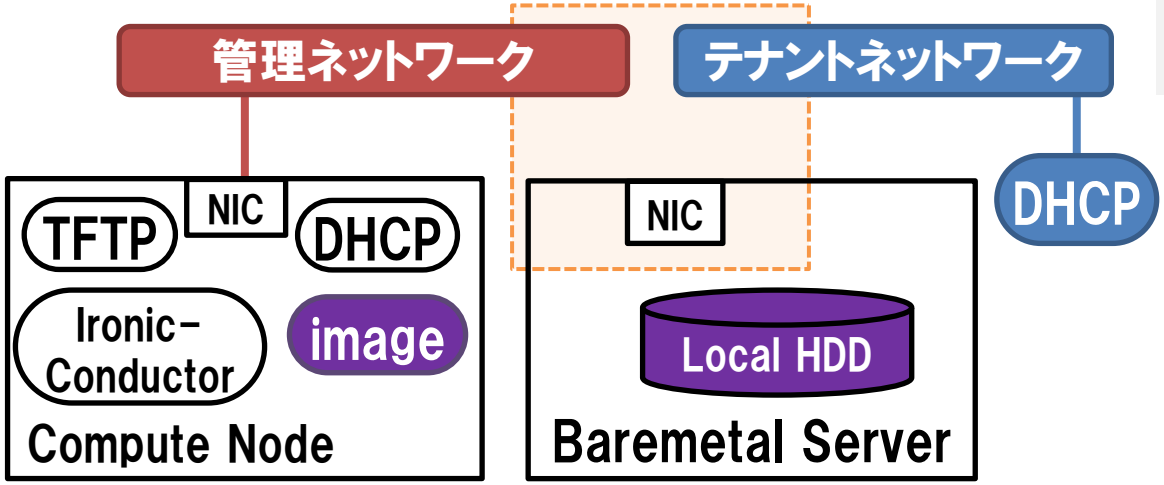
配備中に一旦
接続する必要がある

ネットワークフリップ (2/3)

FUJITSU

配備中にネットワークを切り替える:

1. **管理ネットワークへ接続**
2. **接続解除**
3. **テナントネットワークへ接続**



管理ネットワーク上にあるもの

- TFTPサーバ
- ironic-conductor
- Deploy image(配備用)
- Boot image(お客様用)

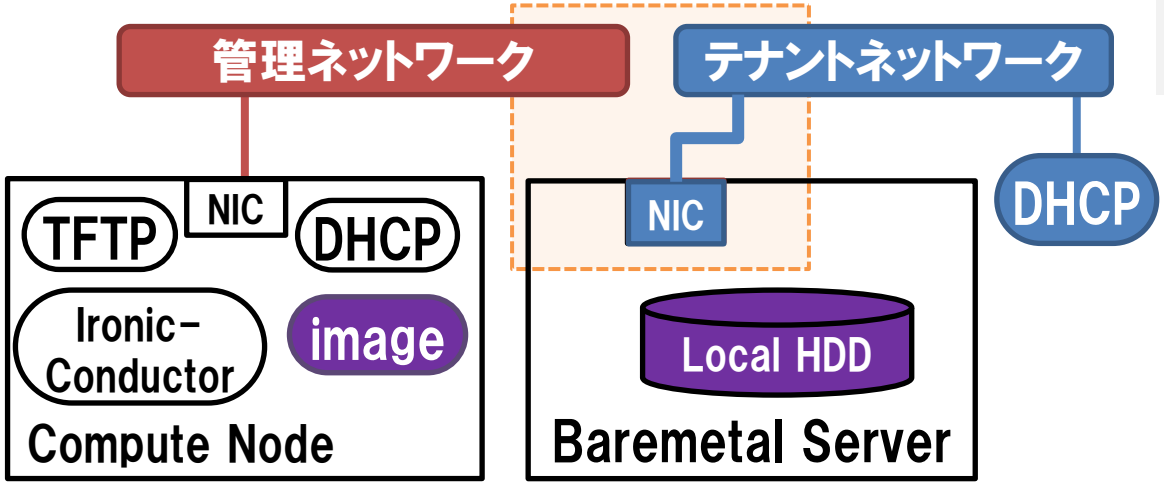
配備中に一旦
接続する必要がある

ネットワークフリップ (3/3)

FUJITSU

配備中にネットワークを切り替える:

1. **管理ネットワークへ接続**
2. **接続解除**
3. **テナントネットワークへ接続**



管理ネットワーク上にあるもの

- **TFTPサーバ**
- **ironic-conductor**
- **Deploy image(配備用)**
- **Boot image(お客様用)**

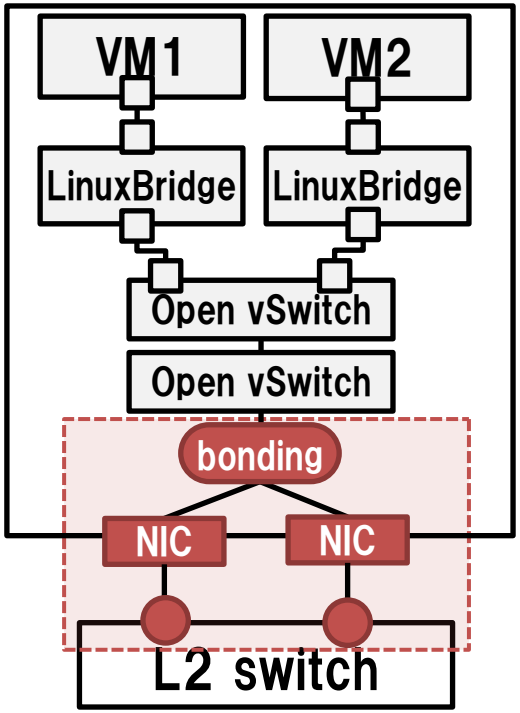
配備中に一旦
接続する必要がある

開発内容：ネットワーク経路冗長化（LAG）

FUJITSU

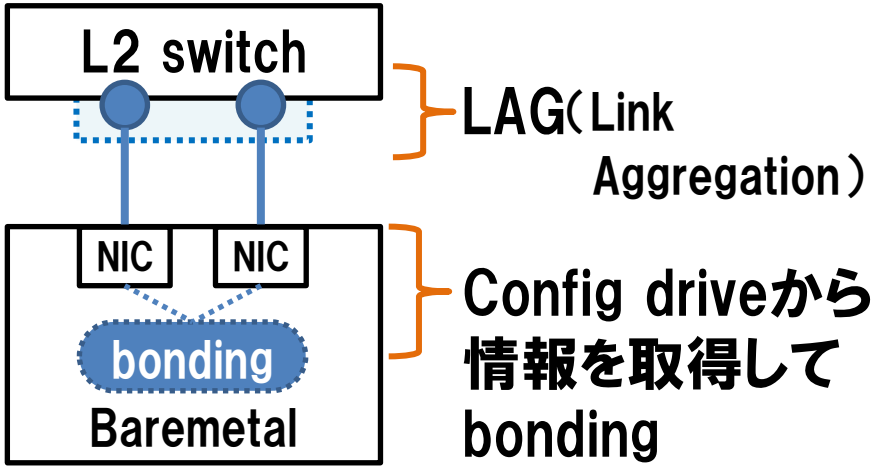
仮想（VM）

- ComputeNodeの物理NICにて実現



物理（BM）

- 物理スイッチ上でLAGを設定
- ベアメタルのOS上でNICのbonding
  - Nova configdrive を利用

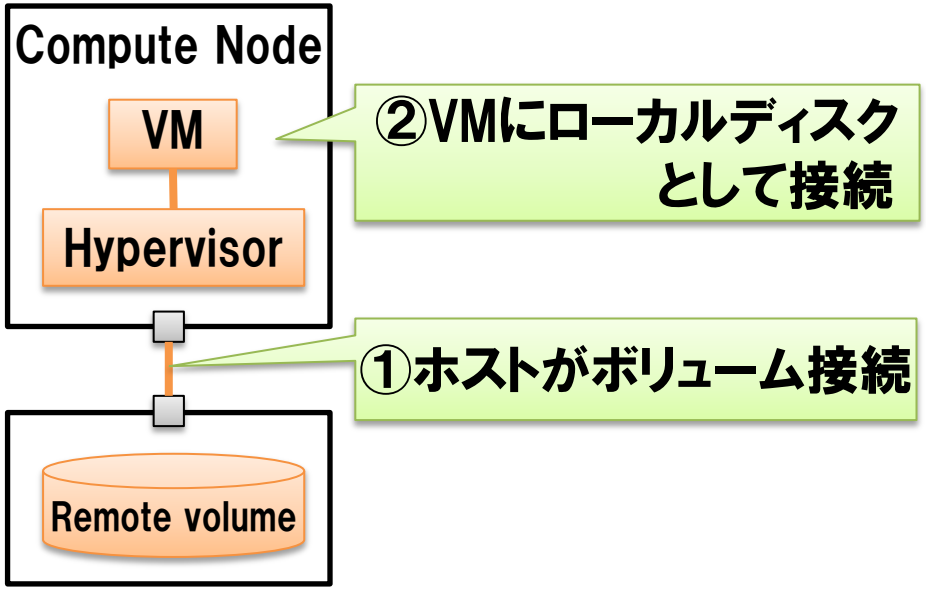


開発内容：SAN対応（Cinder連携）

FUJITSU

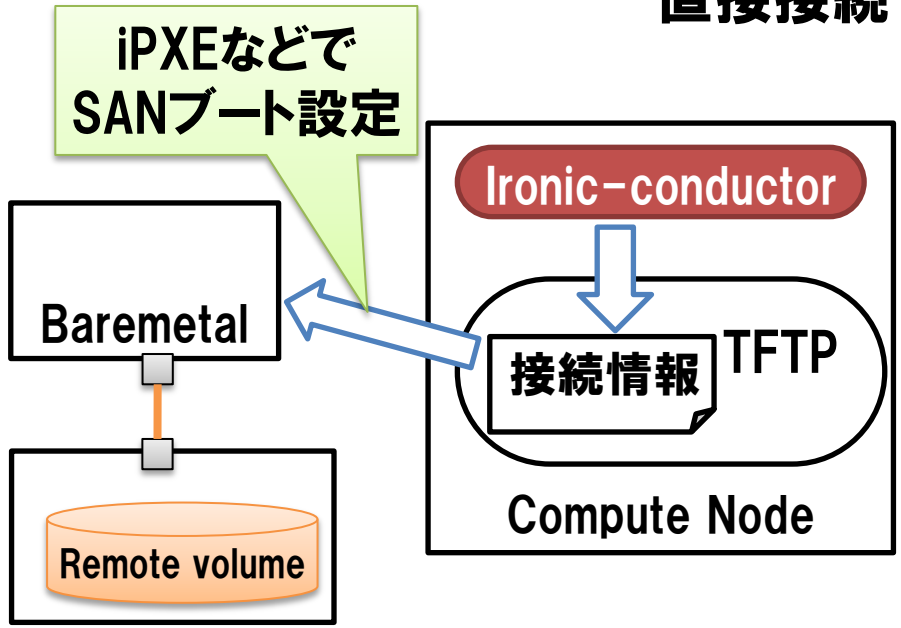
仮想（VM）

- ホストがボリュームに直接接続
- VMはボリュームの実体を意識不要



物理（BM）

- ベアメタルサーバをボリュームに直接接続

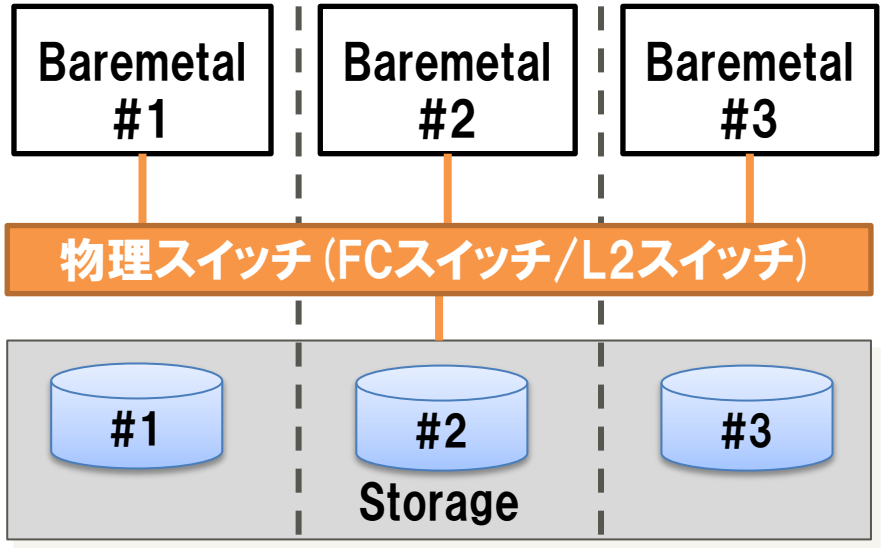


Cinder連携の課題

FUJITSU

■ストレージのマルチテナント対応

各テナント間のデータが参照できないように
ネットワーク、ストレージ装置での分離が必要



FUJITSU

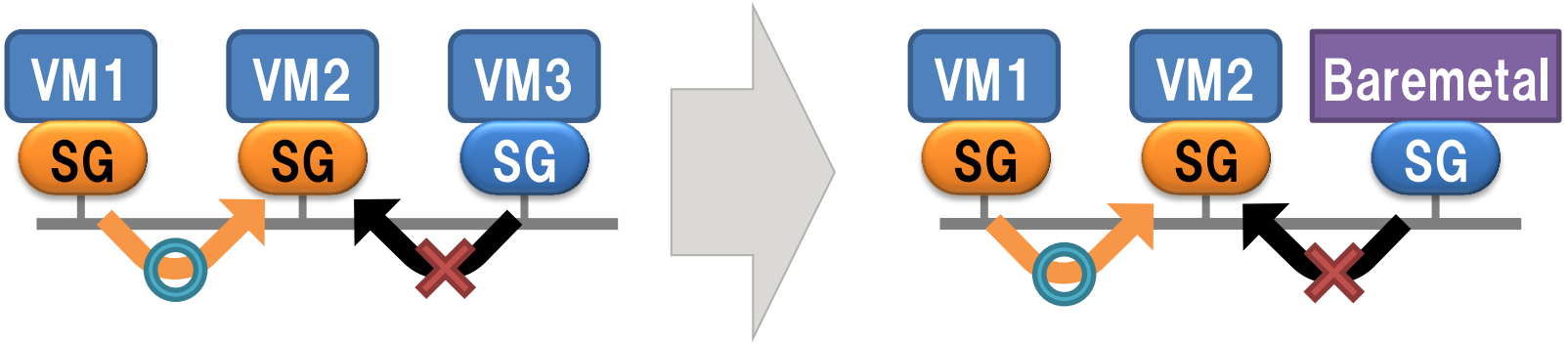
開発内容紹介

- ☐ 導入フェーズ（配備）
- ☑ 正常時運用フェーズ
  - SecurityGroup適用
  - 停止（シャットダウン）
- ☐ 異常時運用フェーズ

SecurityGroupとは

FUJITSU

- フラットなネットワーク間でも適用可能なパケットフィルタリング機能
- 実体は iptables

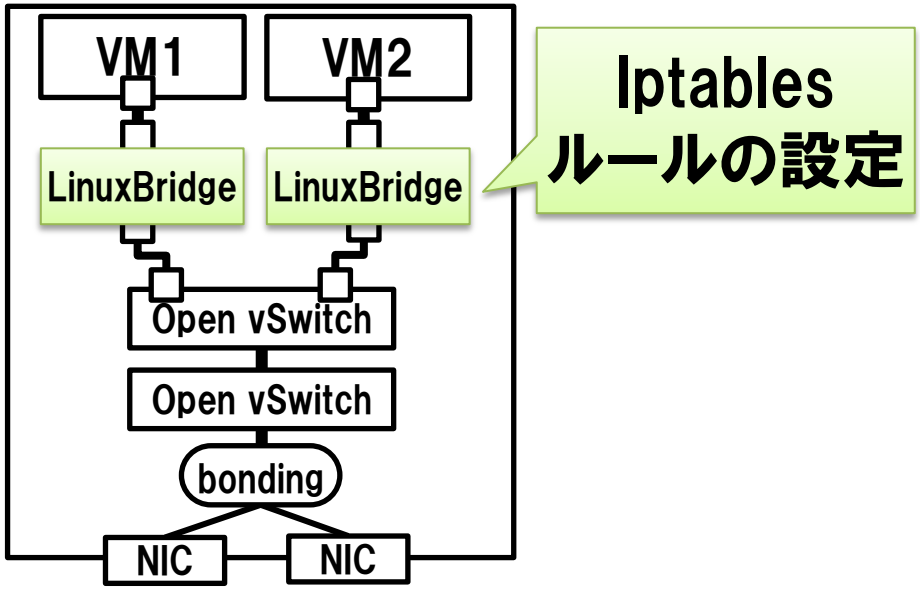


開発内容：SecurityGroup適用

FUJITSU

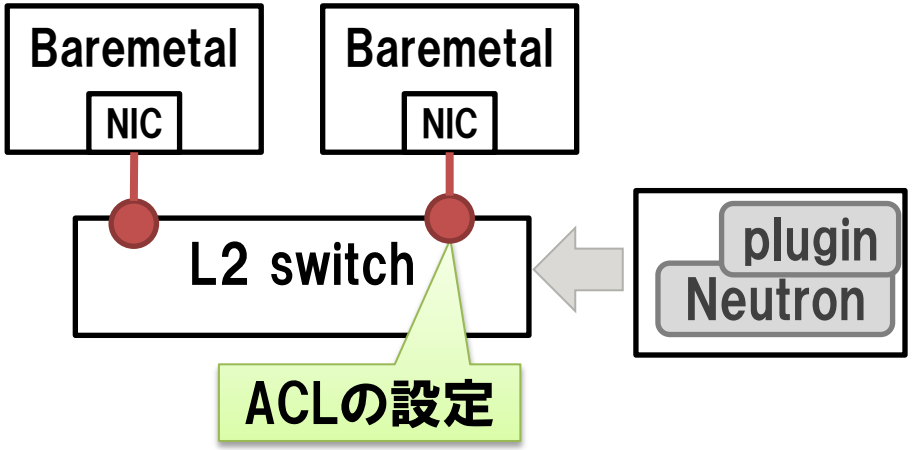
仮想（VM）

- Linux bridge上にiptablesを設定することでSecurityGroupを実現



物理（BM）

- Neutronプラグインを拡張して物理スイッチ上のACL（Access Control List）を制御

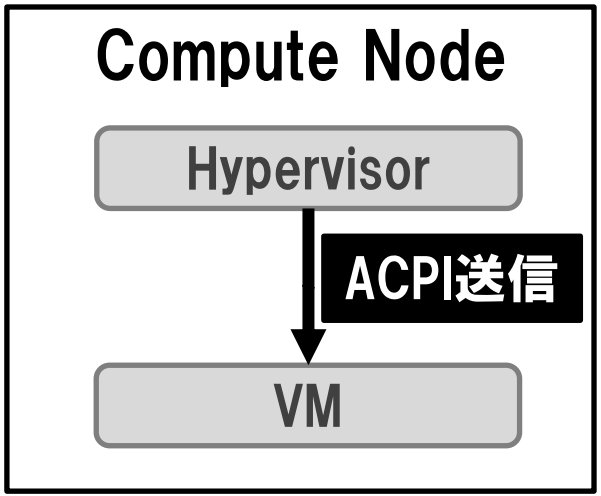


開発内容: Graceful shutdown

FUJITSU

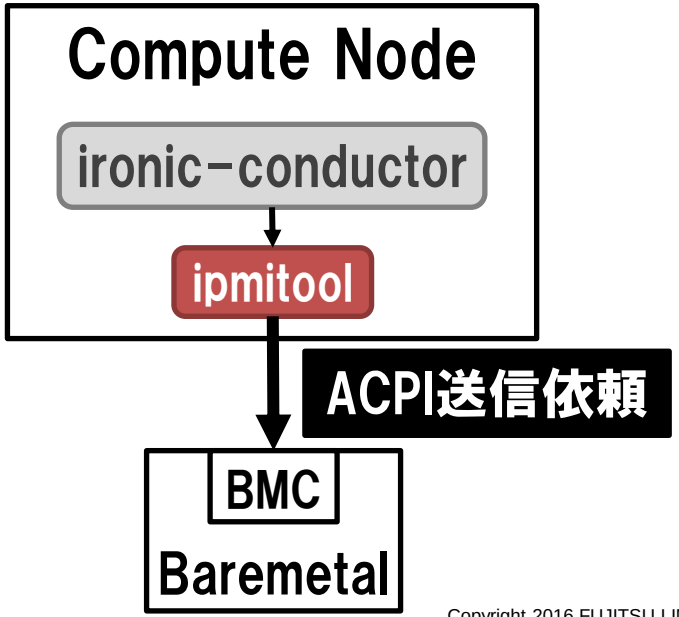
仮想(VM)

- HypervisorがACPIを送信する



物理(BM)

- Ironic-CondurcorがipmitoolBMCからACPIを送信させる



開発内容紹介

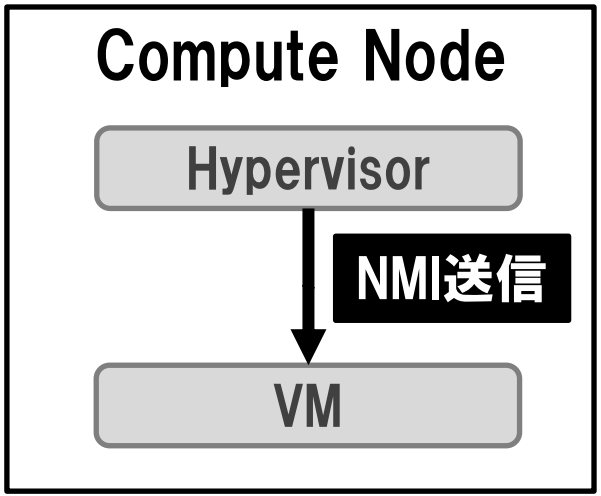
- ☐ 導入フェーズ（配備）
- ☐ 正常時運用フェーズ
- ☑ **異常時運用フェーズ**
  - **カーネルダンプ採取（NMI）**
  - **コンソール接続**

開発内容: NMI送信

FUJITSU

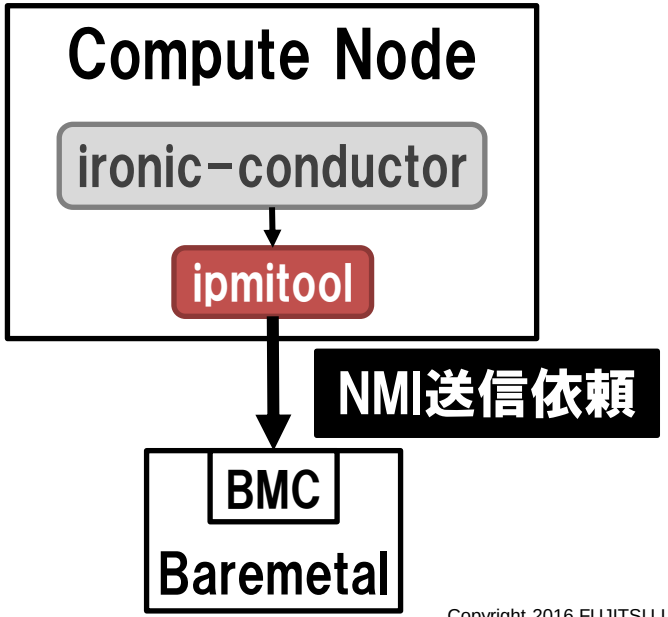
仮想(VM)

- HypervisorがNMIを送信する



物理(BM)

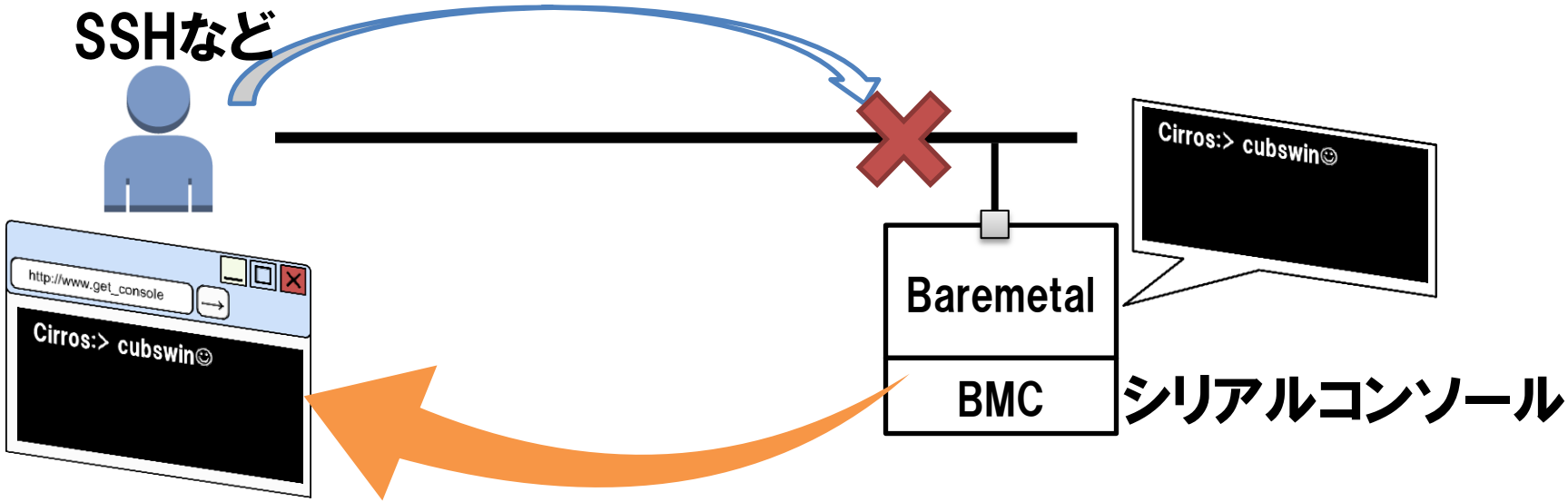
- Ironic-CondurcorがipmitoolBMCからNMIを送信させる



シリアルコンソール

FUJITSU

■OSブート時の異常や、ネットワーク異常時の調査手段としてシリアルコンソールアクセスが必要

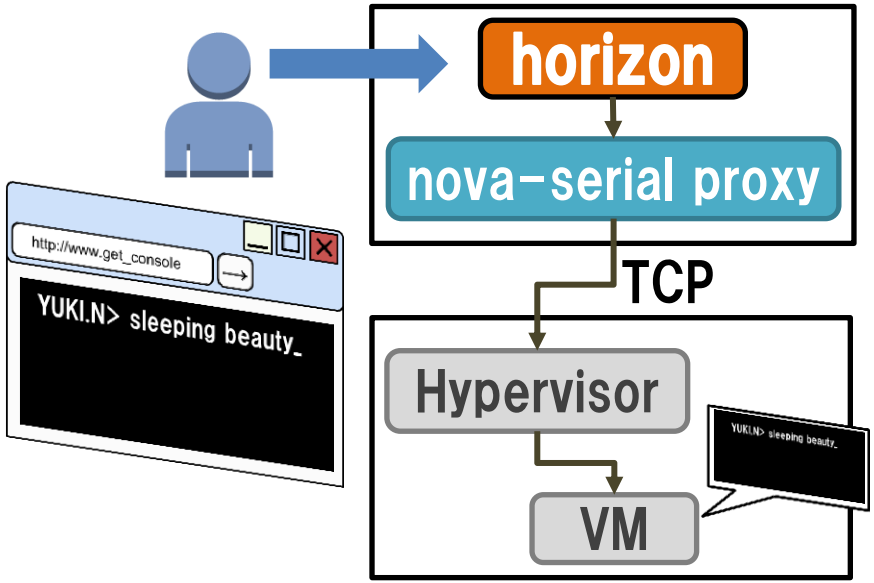


開発内容: シリアルコンソール対応

FUJITSU

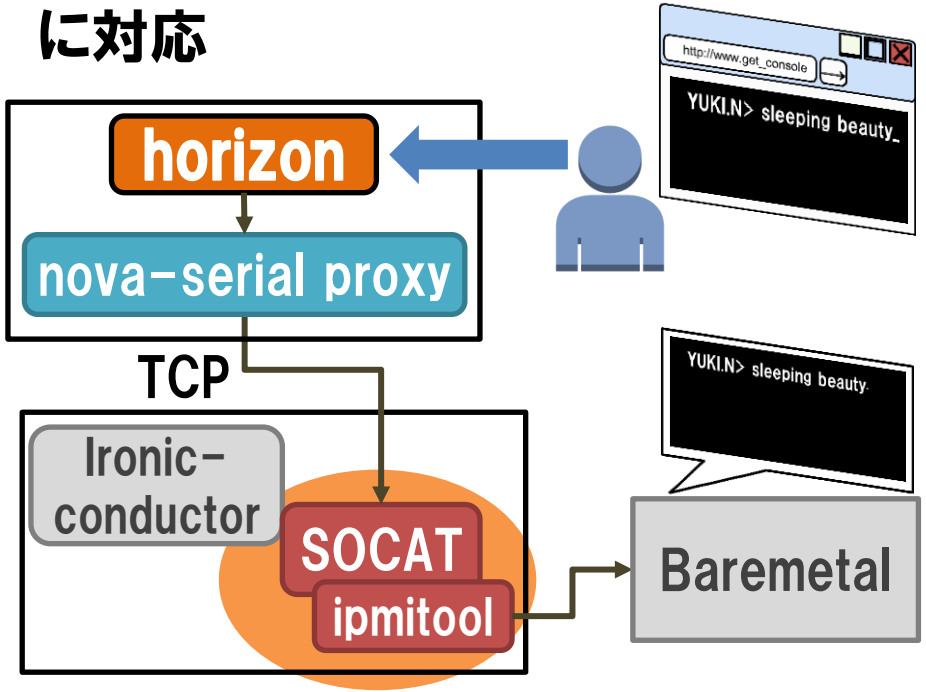
仮想(VM)

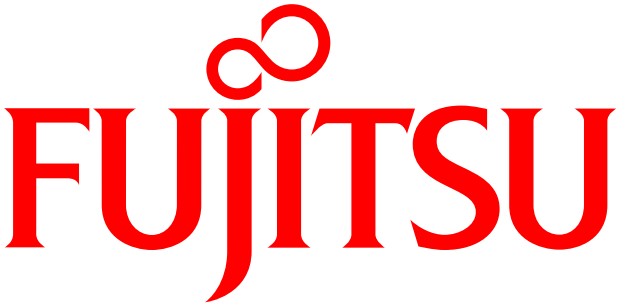
- Nova-serialproxyがHypervisor側の待ち受けポートにプロキシ



物理(BM)

- SOCATを使ってNova-serialproxyに対応





shaping tomorrow with you