Application Performance発表資料

クラウドインフラを支える
高性能ネットワーク、ストレージ技術

さくらインターネット（株）
研究所　大久保 修一
ohkubo@sakura.ad.jp

免責事項

- 「会場のみ」と記載しているスライドについては、一部数値等を伏せさせていただいております。ご了承ください。
- 資料中の性能値は、発表者個人の経験に基づくものであり、弊社の公式見解ではありません。また、製品やリビジョンによって異なる場合がありますので、利用者自身において確認をお願いします。
- この資料は、資料作成時における最新情報をご参考のために提供することを目的として記載されており、弊社は情報の正確性、完全性または有用性について何ら保証するものではありません。また、内容は予告なしに変更または更新されることがあります。
- この資料の情報に基づいて導入・設定・運用した結果について、弊社はいかなる保証も責任も負いかねますので予めご了承ください。

自己紹介

- 所属：さくらインターネット(株) 研究所
- 氏名：大久保 修一
- 数年後のビジネスのネタになりそうな技術の評価、検証など
  - クラウド技術
    ネットワーク仮想化、分散ストレージ
  - IPv4アドレス枯渇対策
    IPv6移行技術
    トランスレータ、トンネル技術(6rd、MAP他)
- さくらのクラウド、ネットワーク、新ストレージ担当

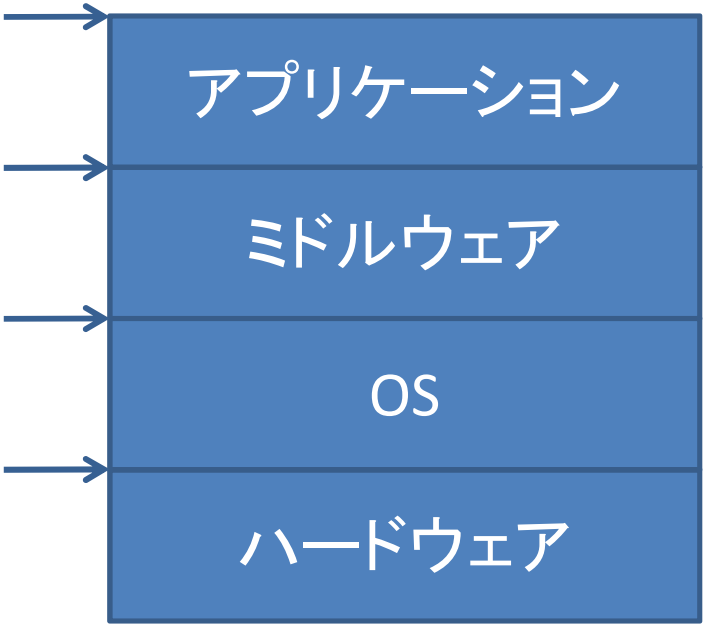
Agenda

- IaaSクラウドの概要
- 弊社クラウドシステムの紹介
- ネットワークのボトルネックと対策について
- ストレージのボトルネックと対策について
- まとめ

IaaSクラウドとは？

- 基本的なコンピューティングリソースを提供するサービス。
- 即時に利用、構成変更、解約可能。

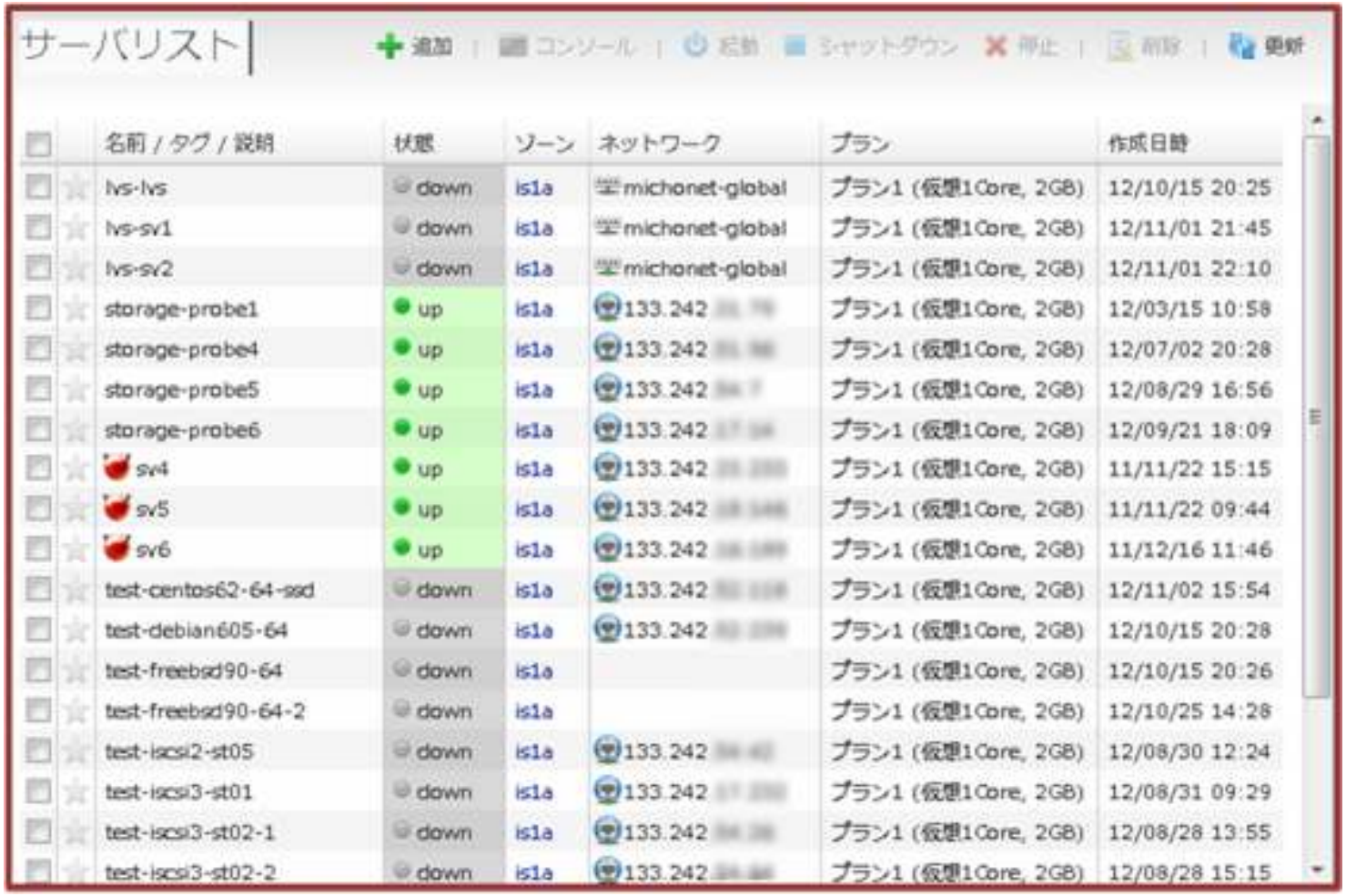
- SaaS (Software as a Service)
- PaaS(Platform as a Service)
- IaaS(Infrastructure as a Service)
- HaaS(Hardware as a Service)



ケーススタディ：さくらのクラウド

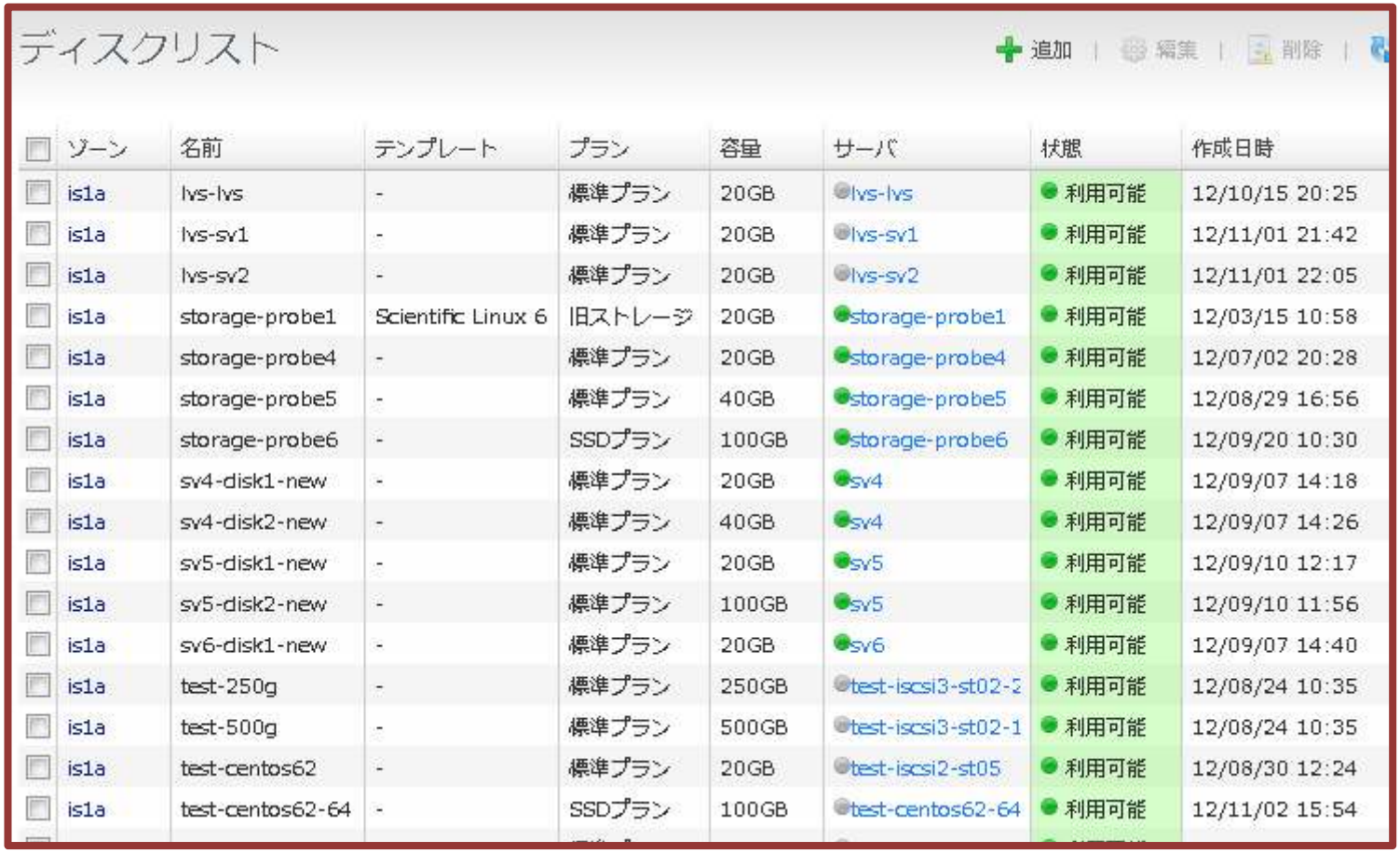
- コンセプト：開発者向けのシンプルクラウド
  - 物理サーバの使い勝手をコンパネ上に再現
- 以下のリソースを提供
  - サーバ（CPU、メモリ）
  - ネットワーク
  - ストレージ
- ネットワークを柔軟に構成可能（専用セグメントあり）
- SSDプラン提供中（2012/11/1～）

例：サーバリソースの一覧



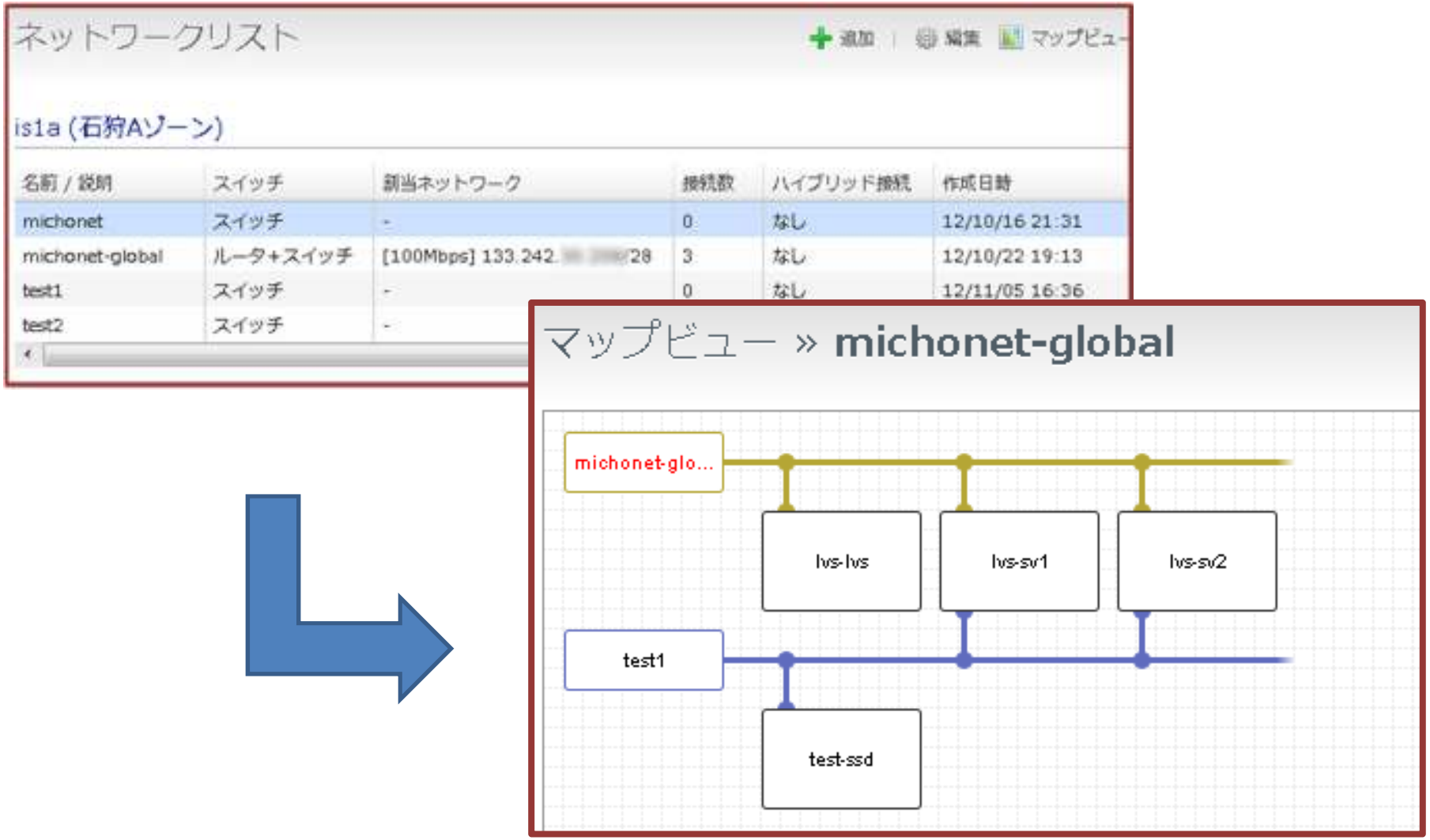
名前 / タグ / 説明	状態	ゾーン	ネットワーク	プラン	作成日時
lvs-lvs	down	is1a	michonet-global	プラン1 (仮想1Core, 2GB)	12/10/15 20:25
lvs-sv1	down	is1a	michonet-global	プラン1 (仮想1Core, 2GB)	12/11/01 21:45
lvs-sv2	down	is1a	michonet-global	プラン1 (仮想1Core, 2GB)	12/11/01 22:10
storage-probe1	up	is1a	133.242	プラン1 (仮想1Core, 2GB)	12/03/15 10:58
storage-probe4	up	is1a	133.242	プラン1 (仮想1Core, 2GB)	12/07/02 20:28
storage-probe5	up	is1a	133.242	プラン1 (仮想1Core, 2GB)	12/08/29 16:56
storage-probe6	up	is1a	133.242	プラン1 (仮想1Core, 2GB)	12/09/21 18:09
sv4	up	is1a	133.242	プラン1 (仮想1Core, 2GB)	11/11/22 15:15
sv5	up	is1a	133.242	プラン1 (仮想1Core, 2GB)	11/11/22 09:44
sv6	up	is1a	133.242	プラン1 (仮想1Core, 2GB)	11/12/16 11:46
test-centos62-64-ssd	down	is1a	133.242	プラン1 (仮想1Core, 2GB)	12/11/02 15:54
test-debian605-64	down	is1a	133.242	プラン1 (仮想1Core, 2GB)	12/10/15 20:28
test-freebsd90-64	down	is1a		プラン1 (仮想1Core, 2GB)	12/10/15 20:26
test-freebsd90-64-2	down	is1a		プラン1 (仮想1Core, 2GB)	12/10/25 14:28
test-iscsi2-st05	down	is1a	133.242	プラン1 (仮想1Core, 2GB)	12/08/30 12:24
test-iscsi3-st01	down	is1a	133.242	プラン1 (仮想1Core, 2GB)	12/08/31 09:29
test-iscsi3-st02-1	down	is1a	133.242	プラン1 (仮想1Core, 2GB)	12/08/28 13:55
test-iscsi3-st02-2	down	is1a	133.242	プラン1 (仮想1Core, 2GB)	12/08/28 15:15

例：ストレージリソースの一覧

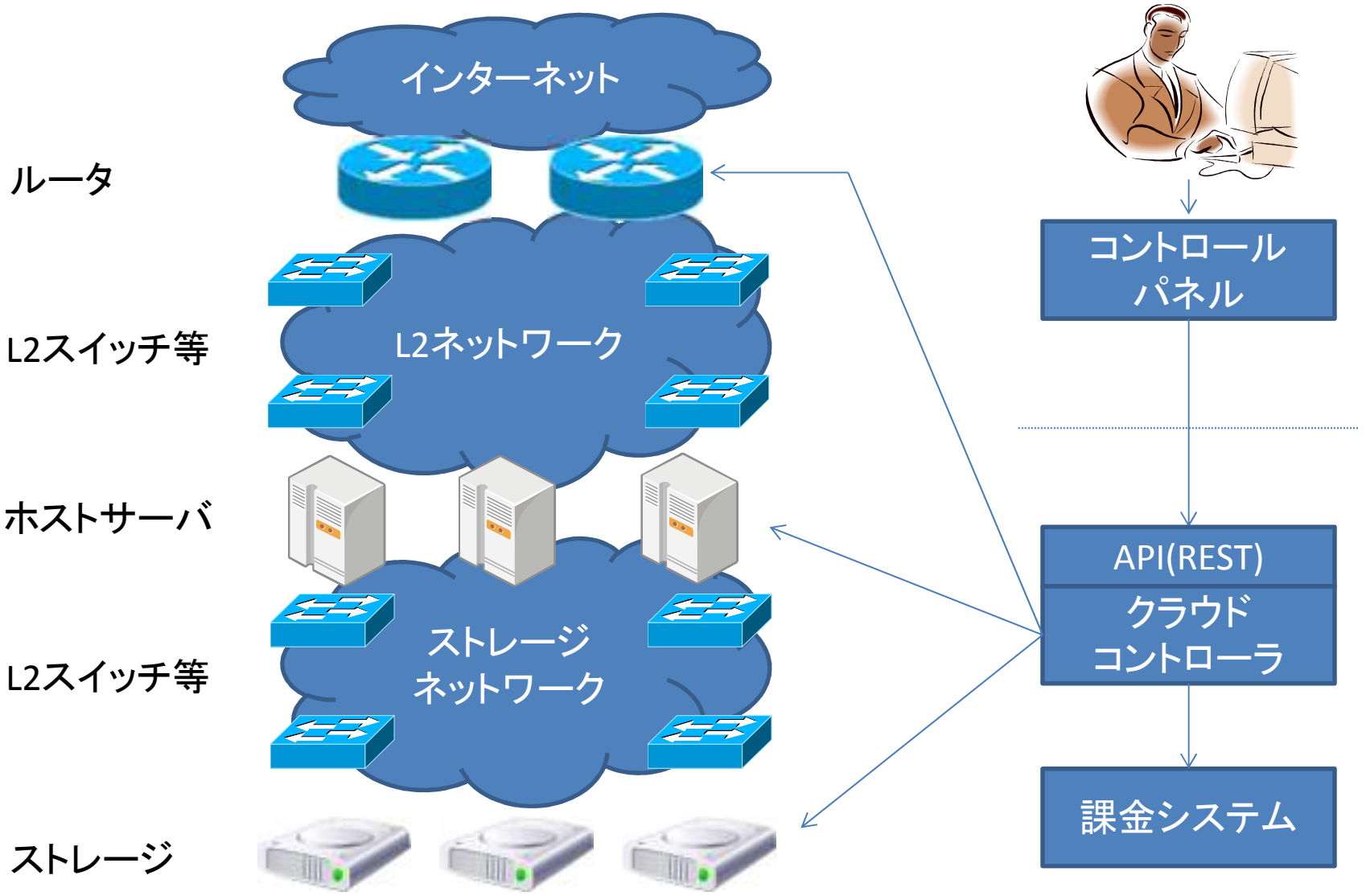


ゾーン	名前	テンプレート	プラン	容量	サーバ	状態	作成日時
is1a	lvs-lvs	-	標準プラン	20GB	lvs-lvs	利用可能	12/10/15 20:25
is1a	lvs-sv1	-	標準プラン	20GB	lvs-sv1	利用可能	12/11/01 21:42
is1a	lvs-sv2	-	標準プラン	20GB	lvs-sv2	利用可能	12/11/01 22:05
is1a	storage-probe1	Scientific Linux 6	旧ストレージ	20GB	storage-probe1	利用可能	12/03/15 10:58
is1a	storage-probe4	-	標準プラン	20GB	storage-probe4	利用可能	12/07/02 20:28
is1a	storage-probe5	-	標準プラン	40GB	storage-probe5	利用可能	12/08/29 16:56
is1a	storage-probe6	-	SSDプラン	100GB	storage-probe6	利用可能	12/09/20 10:30
is1a	sv4-disk1-new	-	標準プラン	20GB	sv4	利用可能	12/09/07 14:18
is1a	sv4-disk2-new	-	標準プラン	40GB	sv4	利用可能	12/09/07 14:26
is1a	sv5-disk1-new	-	標準プラン	20GB	sv5	利用可能	12/09/10 12:17
is1a	sv5-disk2-new	-	標準プラン	100GB	sv5	利用可能	12/09/10 11:56
is1a	sv6-disk1-new	-	標準プラン	20GB	sv6	利用可能	12/09/07 14:40
is1a	test-250g	-	標準プラン	250GB	test-iscsi3-st02-2	利用可能	12/08/24 10:35
is1a	test-500g	-	標準プラン	500GB	test-iscsi3-st02-1	利用可能	12/08/24 10:35
is1a	test-centos62	-	標準プラン	20GB	test-iscsi2-st05	利用可能	12/08/30 12:24
is1a	test-centos62-64	-	SSDプラン	100GB	test-centos62-64	利用可能	12/11/02 15:54

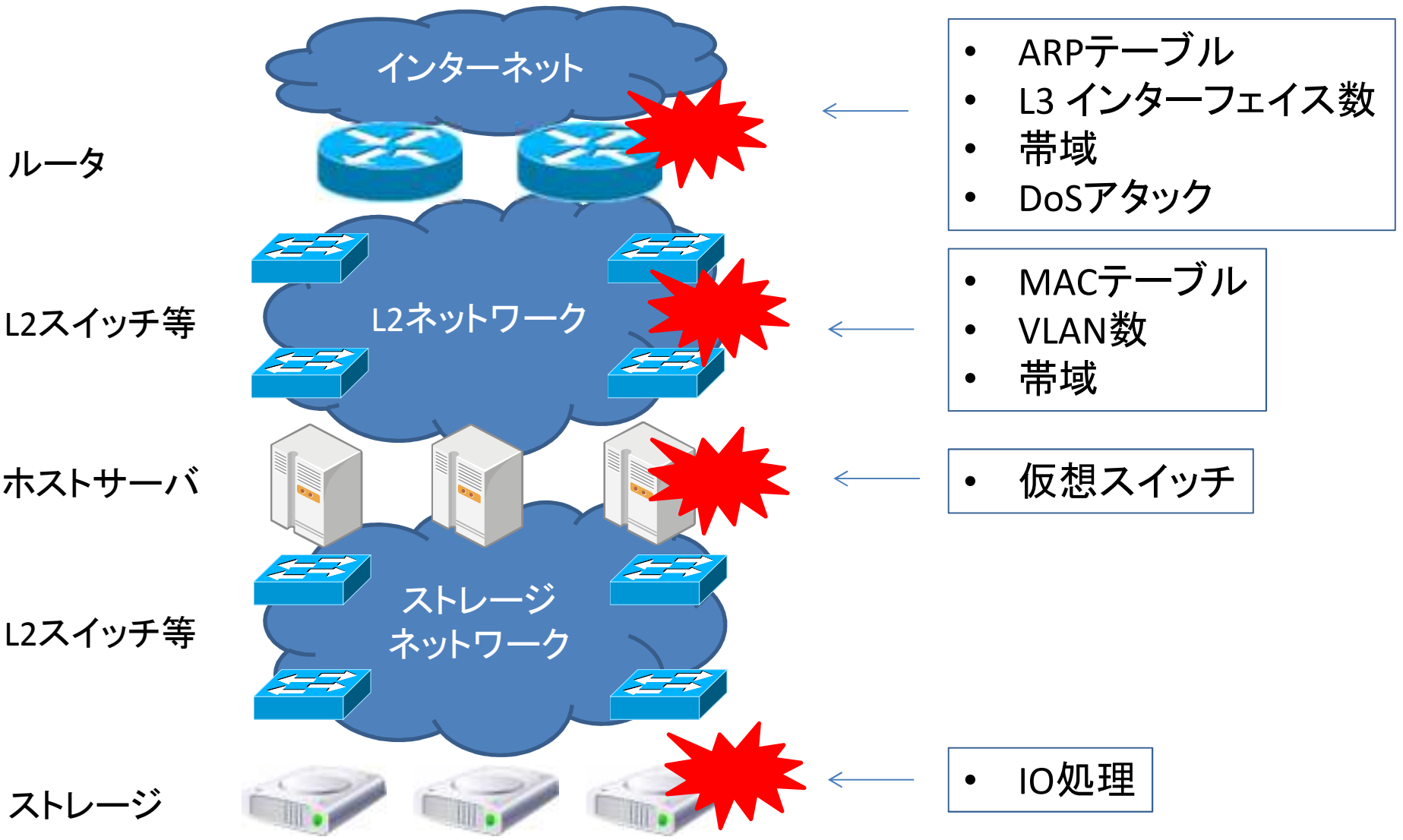
例：ネットワークリソースの一覧



クラウドインフラの構成



クラウドインフラボトルネックマップ

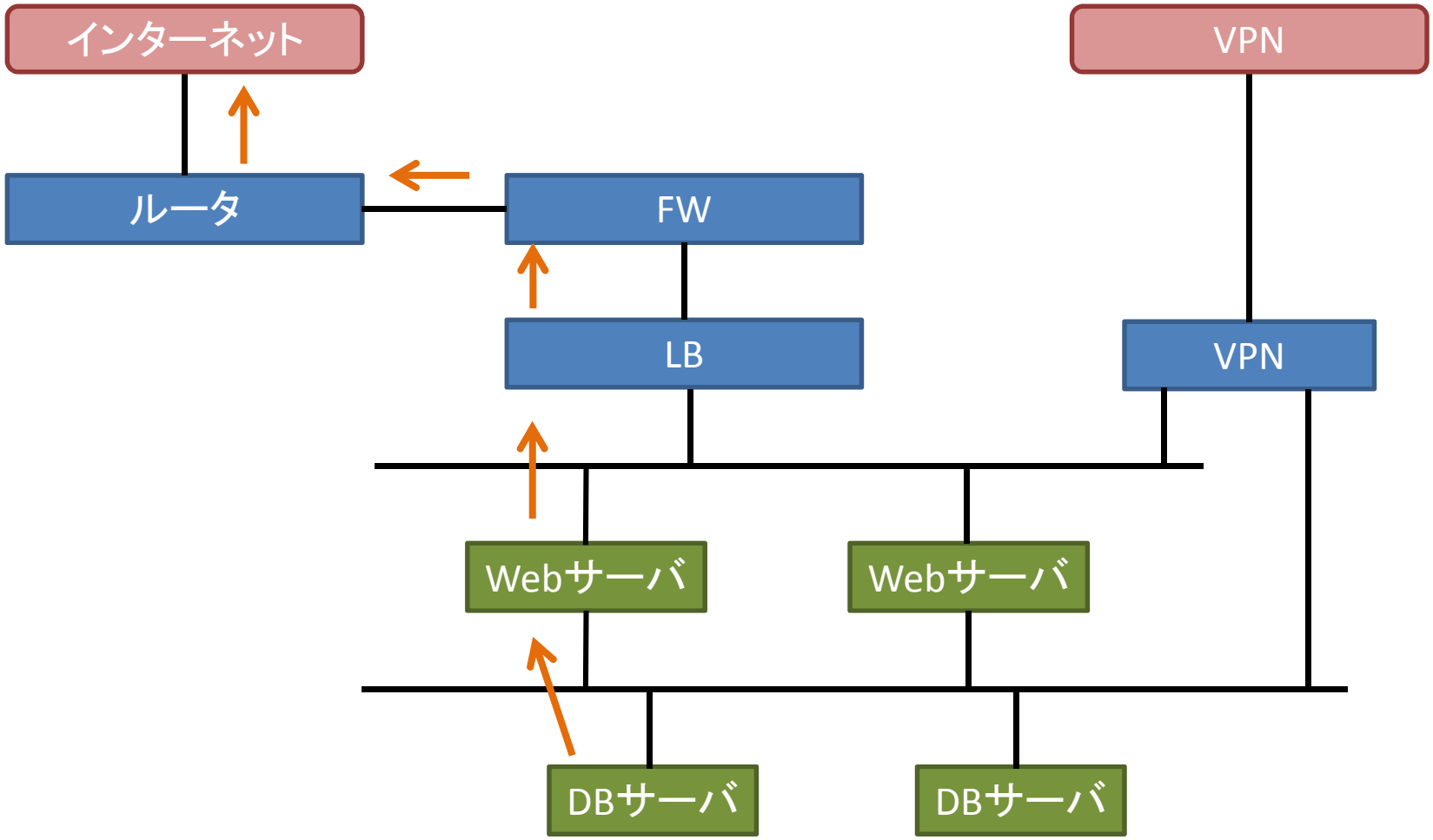


ネットワーク編

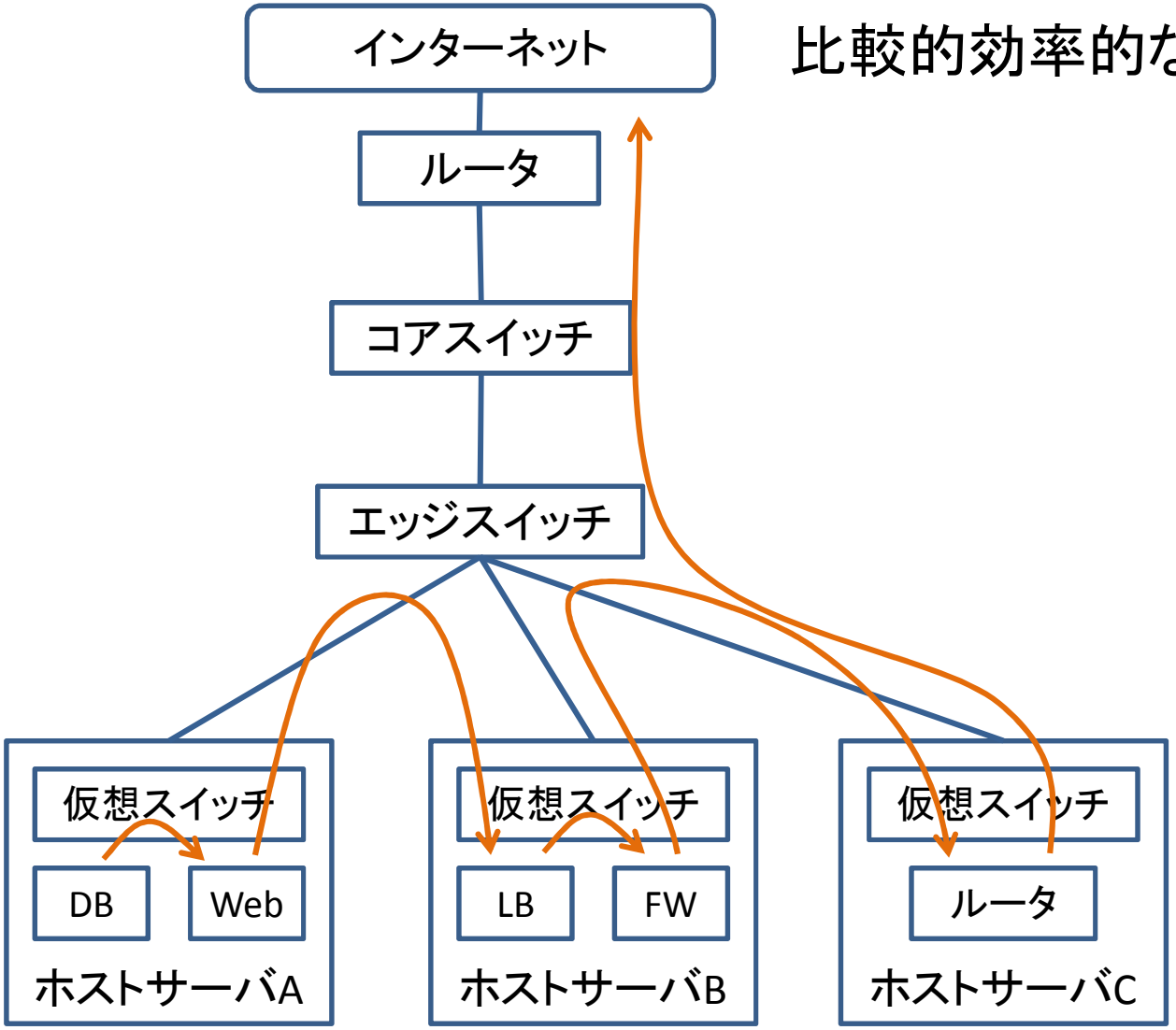
L2NWの帯域圧迫

- サーバ間トラフィックの増大
- 10Gbps以上の接続が必須
- 弊社ではホストサーバをInfiniBand(40Gbps)にて接続
- 10GbEスイッチ、EoIB変換、IBスイッチ群にて構成

例：仮想ネットワークトラフィックフロー



物理トラフィックフロー(パターン1)

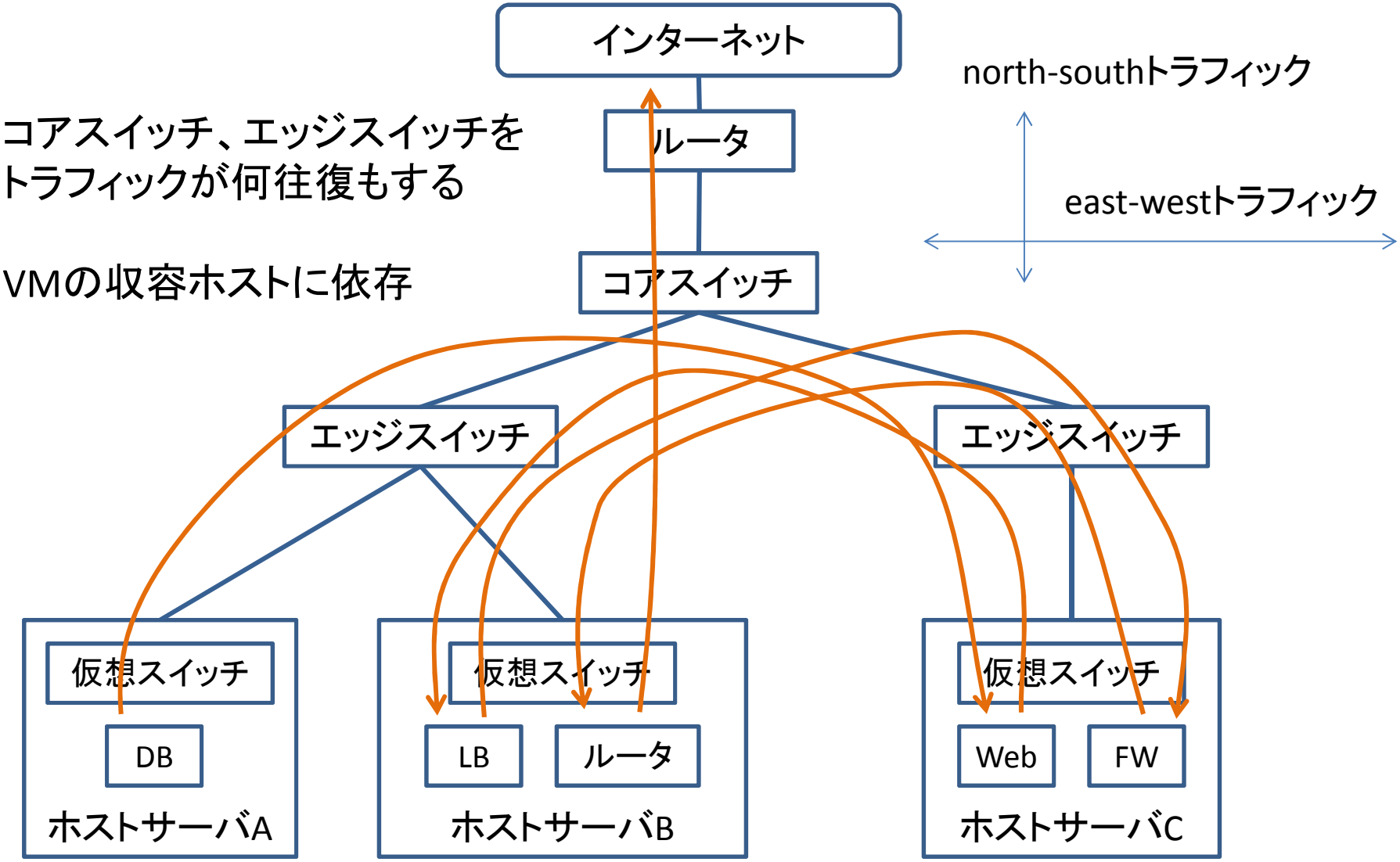


比較的効率的なケース

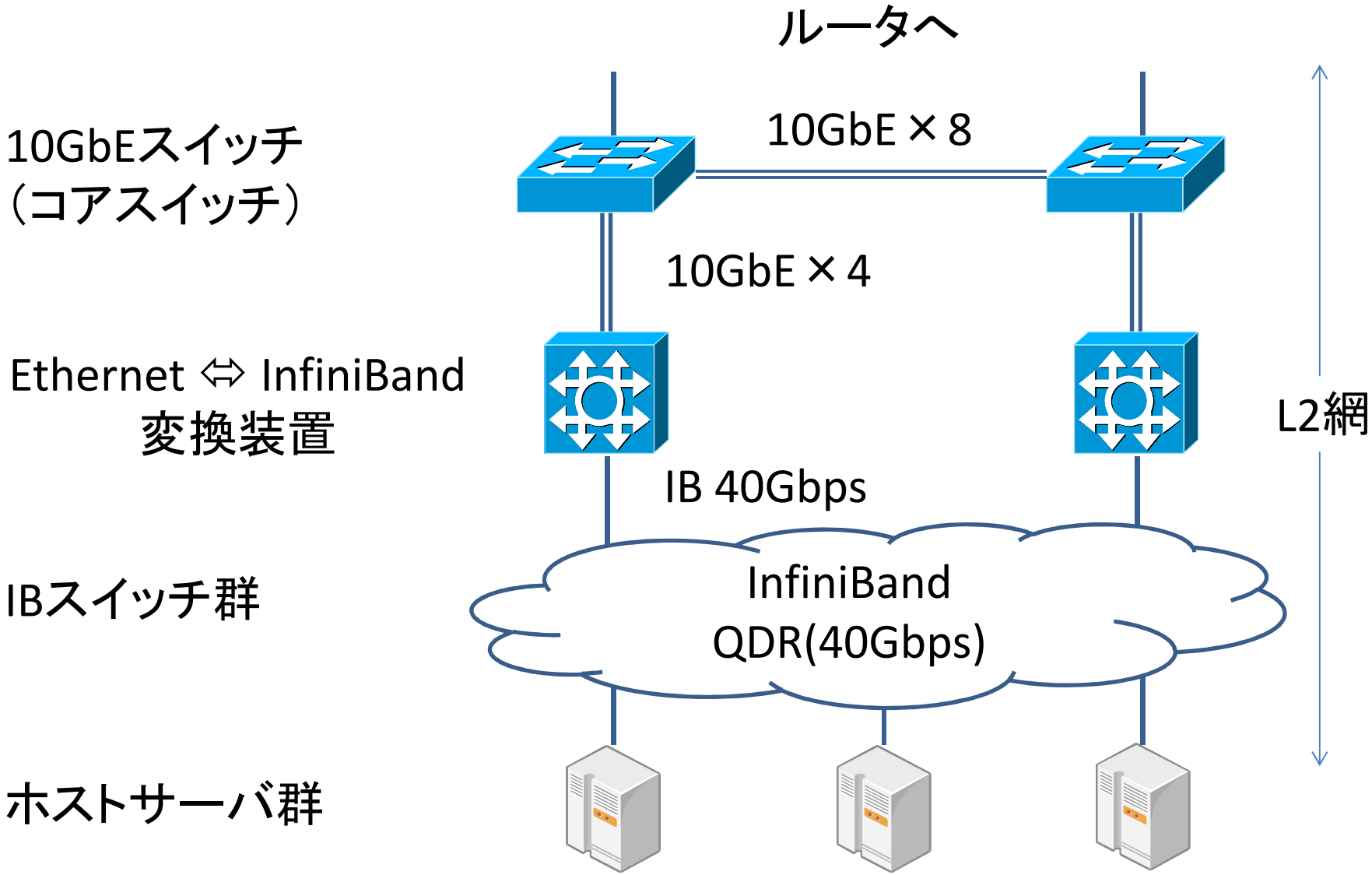
物理トラフィックフロー(パターン2)

コアスイッチ、エッジスイッチを
トラフィックが何往復もする

VMの収容ホストに依存



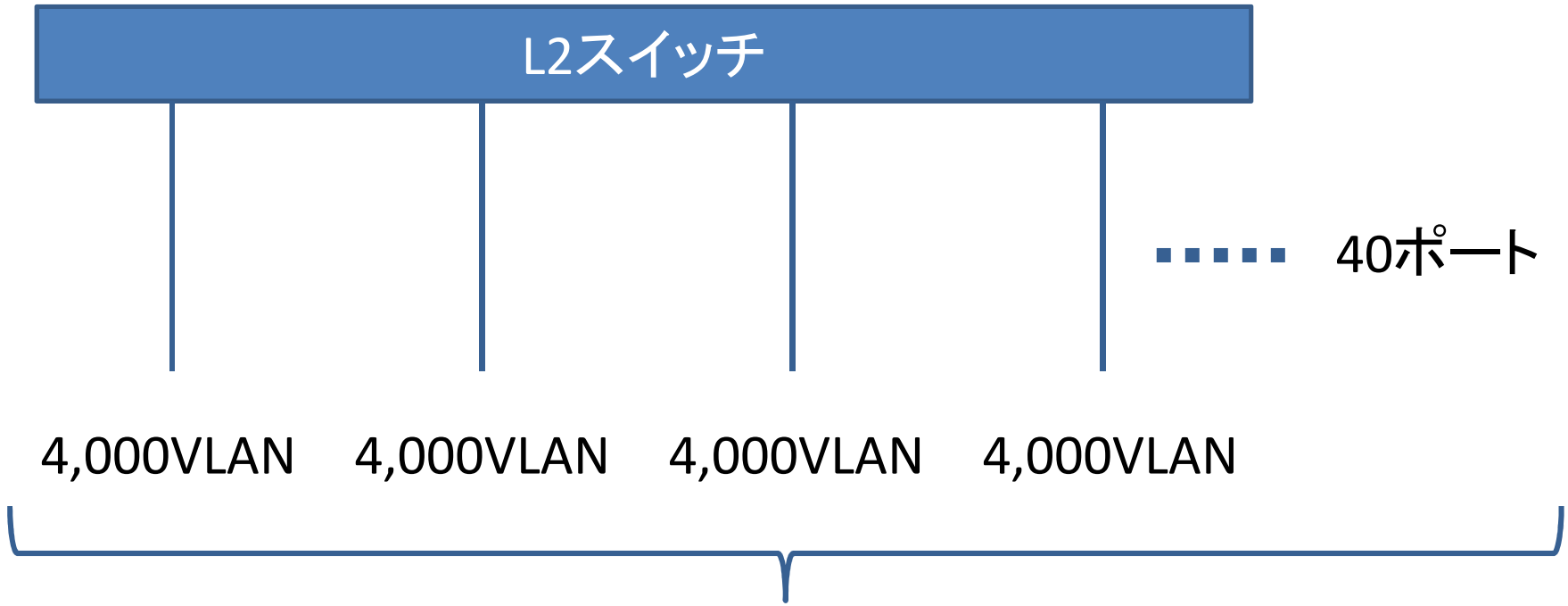
現在のL2ネットワークの構成



VLAN数の制限

- VLAN ID数の不足(12bit≒約4,000VLAN)
  - 1ユーザあたり10VLANとして、400ユーザが上限
- ロジカルポート数の不足(次ページで説明)
- 装置によって扱えるVLAN数に制限があるケース
- 弊社では当初2,000VLANをデプロイ

参考：ロジカルポート数の不足



ロジカルポート数が12,000や24,000の制限があるスイッチがある

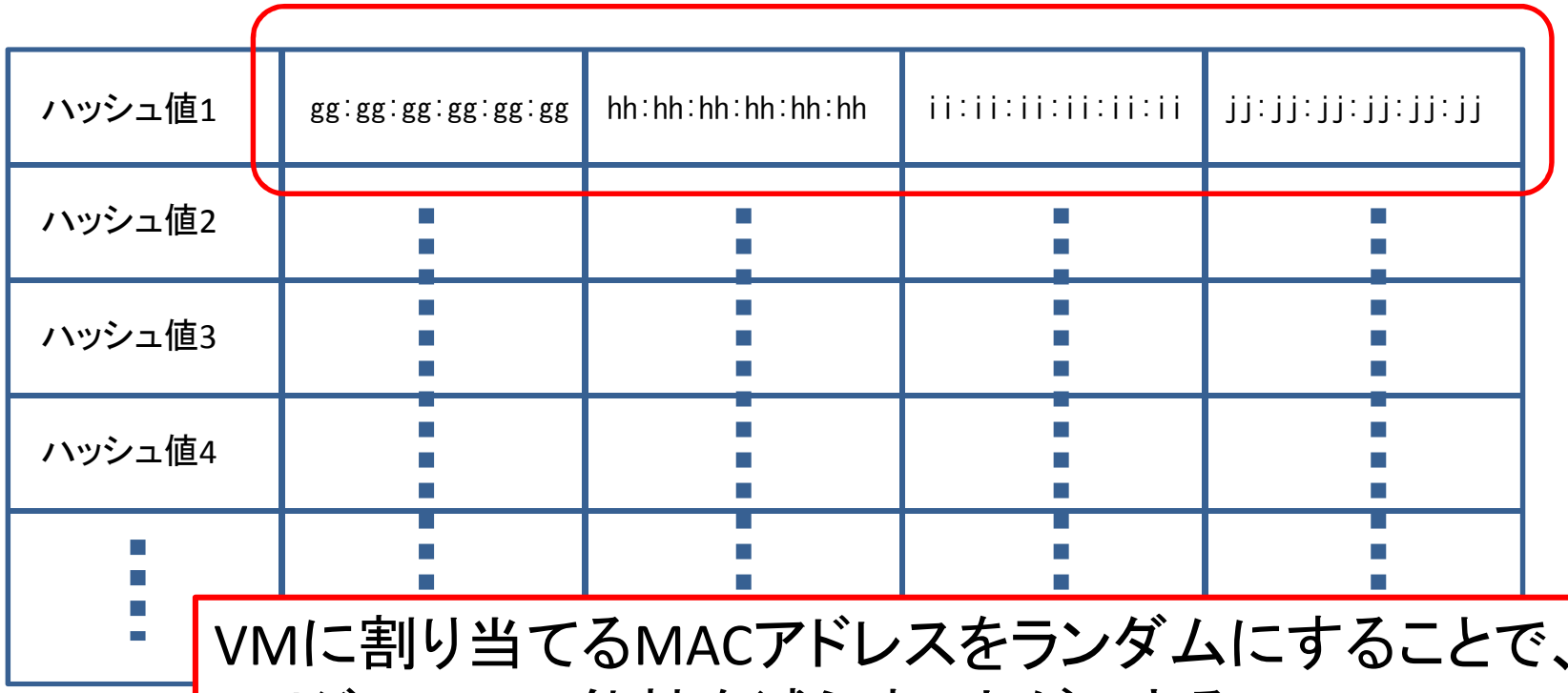
MACアドレス数の制限

- MACテーブルエントリ数
  - ハッシュコリジョン問題(次ページで説明)
  - Internalで消費されるMACアドレス
  - 実質、カタログスペックの半分くらいしか使えない
  - 弊社での要件は16,000エントリ

参考：ハッシュコリジョン問題

同じハッシュ値をとる5つ目のMACアドレスが
来ると学習できない

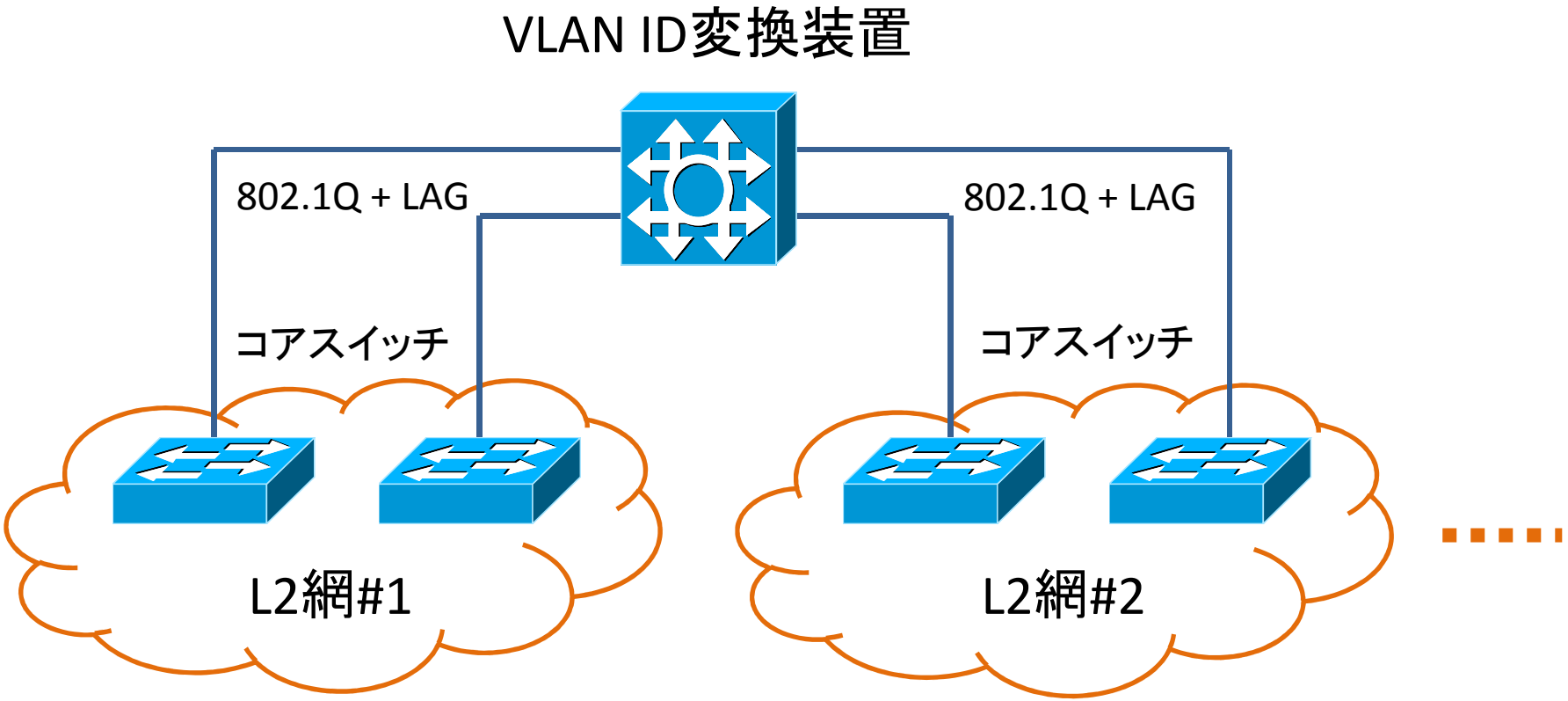
4段の例



VMに割り当てるMACアドレスをランダムにすることで、
コリジョンの可能性を減らすことができる

L2網のスケーラビリティ確保

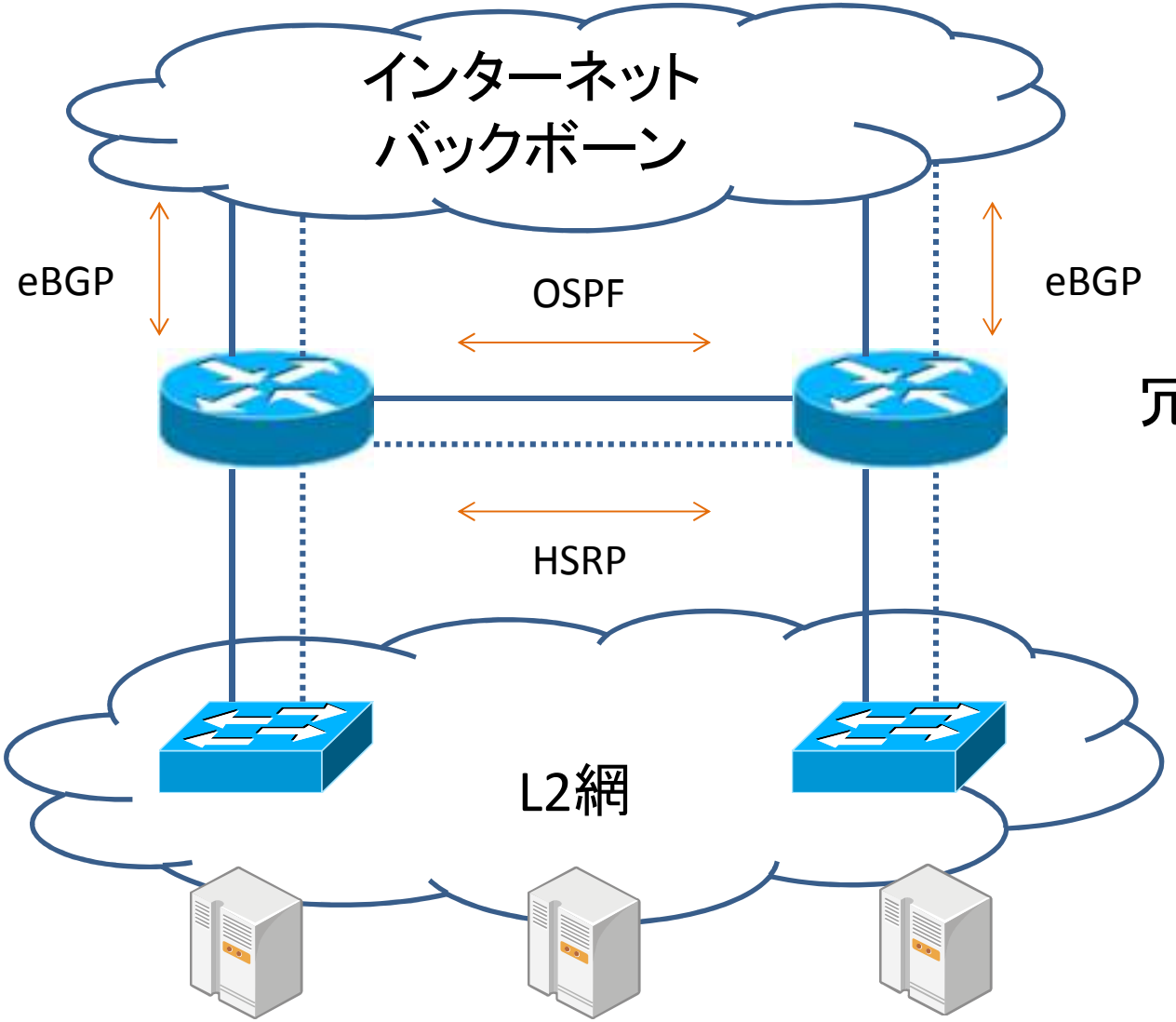
VLAN数制限、MACアドレス数制限により、
基本的に網を分割していくしかない。
分割されたL2網間を接続する仕組みが別途必要



ルータのボトルネック

- 帯域
  - 基本的に10Gbpsで接続、足りなければN本に増速
- ARPテーブル
  - インターネットに接続するVMのIPアドレス、MACアドレスを学習
  - 弊社の要件は12,000エントリ程度
- L3仮想インターフェイス数
  - インターネットに接続するVLANの収容
  - 冗長プロトコルのインスタンス数の制限にあたりやすい
  - 弊社の要件は2,000インスタンス程度
- ルータ宛てのDoSアタックへの耐性

ルータの構成

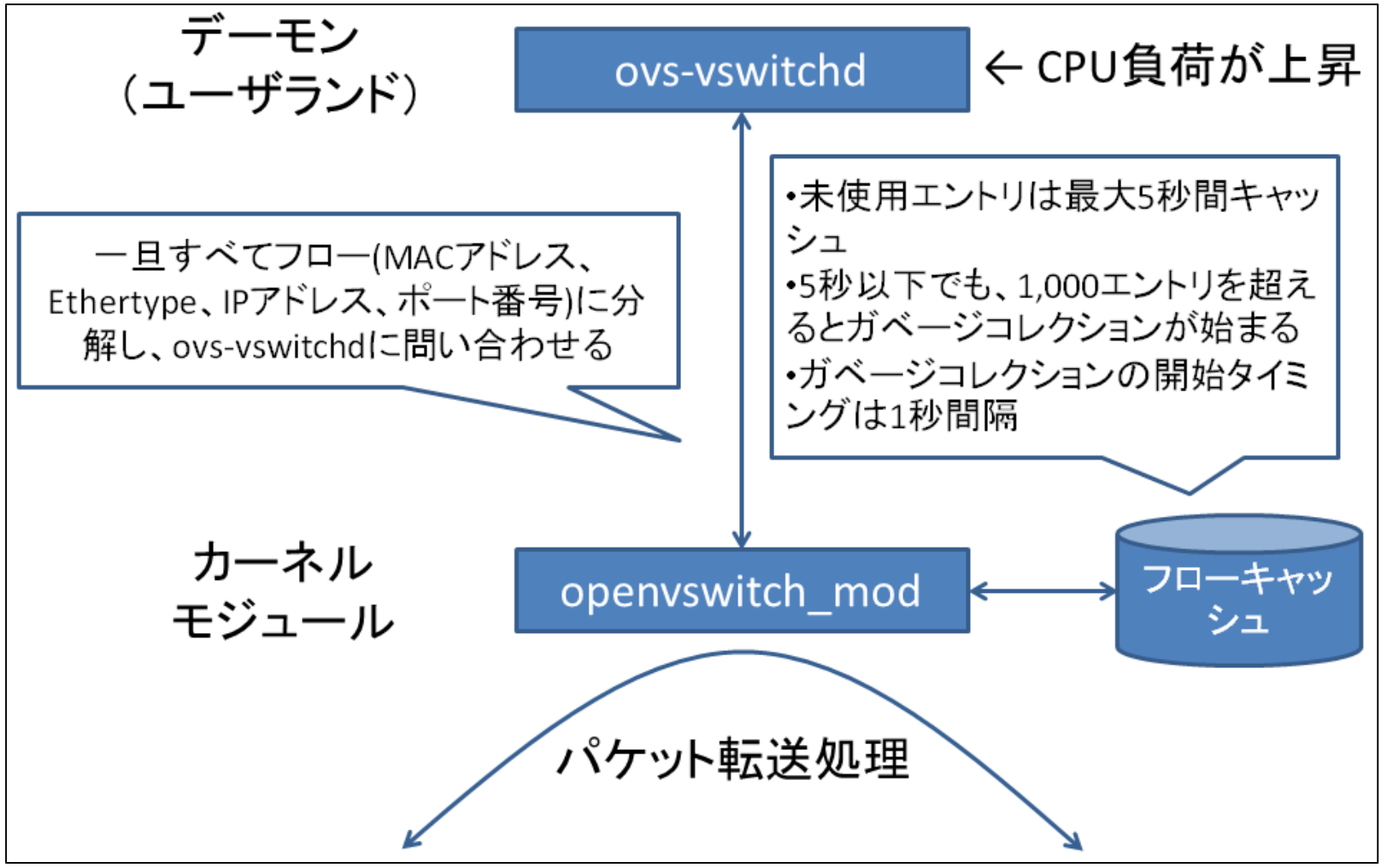


冗長構成

仮想スイッチのボトルネック

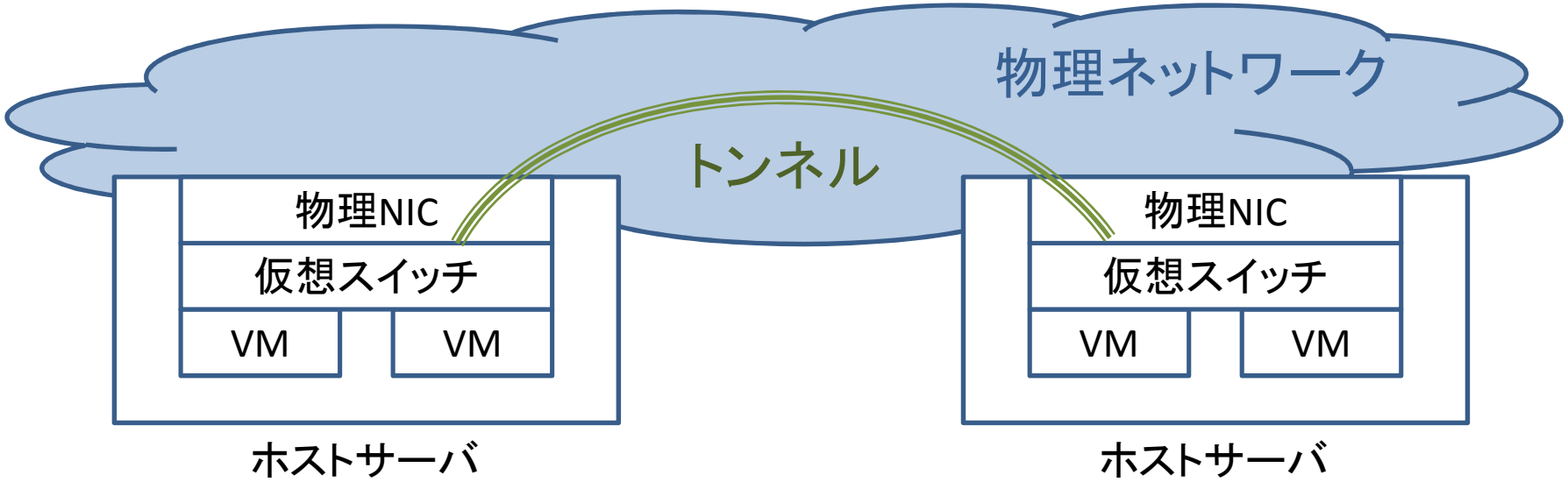
- 弊社ではOpen vSwitchを使用
- DoSアタックなど、大量のフローが発生すると、転送性能が極端に低下する
- 同一ホストの別のお客様のVMに影響が及ぶ
- 弊社で実施した対策
  - 新しいバージョンを使用する
  - ガベージコレクション開始の閾値を変更 (1,000⇒120,000)
  - フロー数監視⇒異常な通信の検出

Open vSwitchのアーキテクチャ



将来的に

- SDN/OpenFlowの導入を検討
- エッジオーバレイ方式により、L2網のスケーラビリティ向上



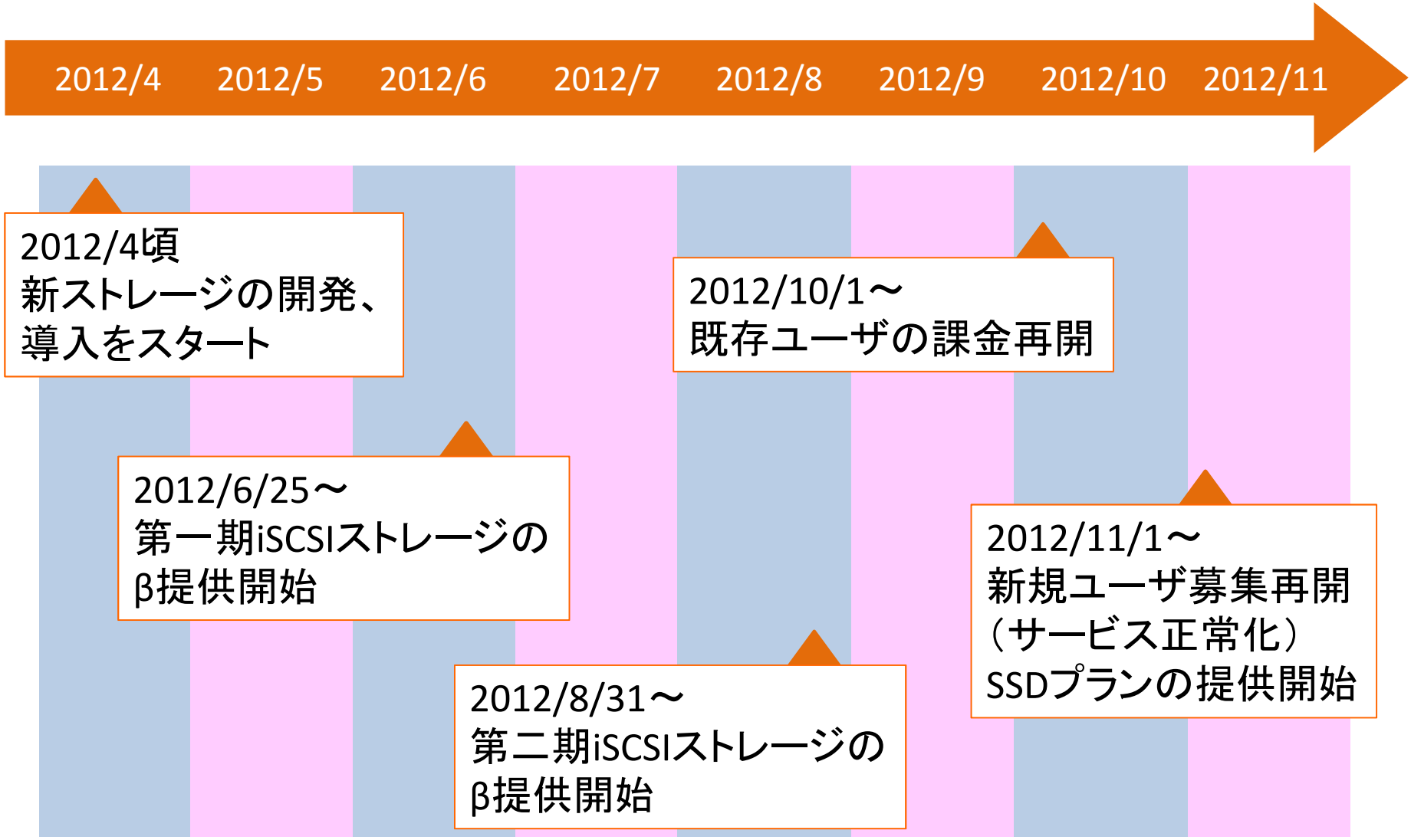
続きは SDN Japan 2012 にて！
2012/12/6〜12/7開催（於:ベルサール九段）
http://www.sdnjapan.org/

ストレージ編

弊社ストレージ障害について

- サービス開始当初、導入したストレージの不調
- 2011/12頃から2012/3まで障害が多発
  - 負荷上昇により、ホストサーバとの接続が切断
  - 冗長プロトコル(IPMP)のフラップ
  - IBインターフェイスの問題
  - シェア数増大による制御遅延
  - IOパフォーマンス劣化
  - アップデートメンテによる障害
- 2012/3よりサービスの課金、および新規申し込みを停止させていただく
- 大変ご迷惑をおかけしました m(__)m

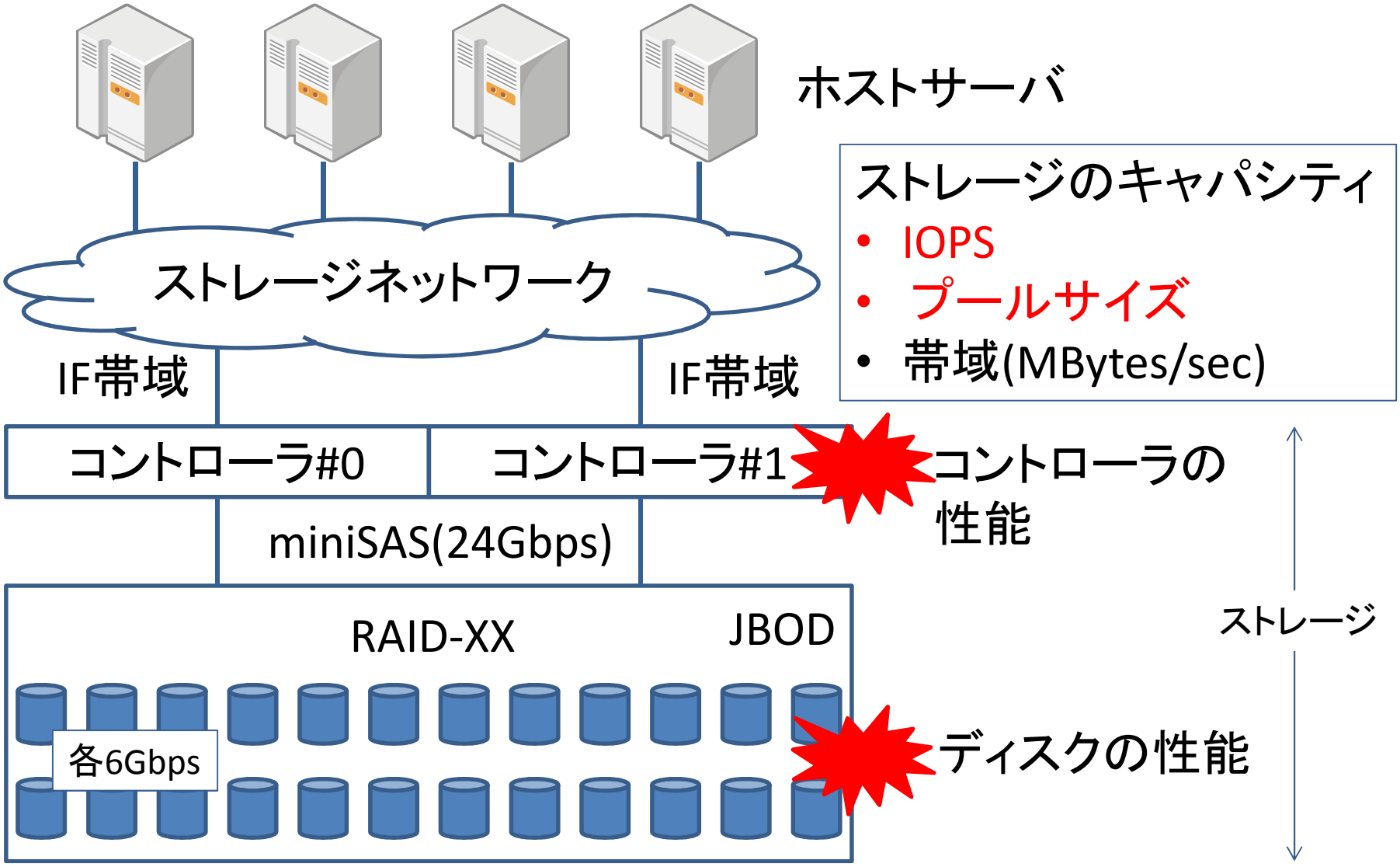
新ストレージ導入の経緯



新ストレージの導入にあたって

- 0からの見直し（旧ストレージを捨てる）
- ボトルネックの洗い出し
- 収容設計の手法を確立
- ユーザからのIO負荷制限
- 自社で運用に責任をもつ
  - 旧ストレージはベンダ任せの部分があった
- 高負荷時の挙動を把握できるように
- デグレ時の性能が読めるように

ストレージのボトルネック

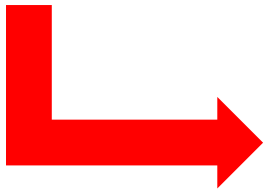


ストレージのボトルネック

- HDDはIOPSが絶対的に不足
  - 1本のディスクを数十ユーザで共有
  - 共有ストレージは難しい
  - 専用サーバ等では1本のディスクを1ユーザが専有（DAS最強説？）
- その他の制限
  - IF、miniSASの帯域（あまり気にする必要ない？）
  - iSCSIセッション数上限（1ポートあたり256セッションなど）

キャパシティを左右するパラメータ

- ディスクのタイプ（SAS ~~or NL-SAS or SATA~~）
- HDDのサイズ（300GB or 600GB or 900GB）
- 回転数（10Krpm or 15Krpm~~ or 7.2Krpm~~）
- RAID構成（RAID10 or RAID60~~ or RAID50~~）
- ホットスペア（２～４本？）
- HDDの本数（JBODにささるきりのいい数字に）
- スナップショット領域



IO性能（IOPS）、プールサイズ(GB)

注意・検討が必要な事項

- IaaS用途ならSASのディスクが望ましい？
- HDDのサイズはGiBではなく、GBである。
- ストレージ製品によってパリティなどの予備領域が確保されるため、使用できる容量はかなり減る。
- RAID5(50)はuncorrectable errorによりリビルドが失敗するリスクが高くこわい。
- RAID6(60)はRAID1(10)よりも安心。

RAID構成による違い

性能1/3

	RAID10	RAID60(8+PQ)	RAID50(4+P)
random write IOPS	N/2	N/6	N/4
random read IOPS	N	N×4/5	N×4/5
容量	M/2	M×4/5	M×4/5
リビルドの安全性	○	◎	△

容量1.6倍

- N=HDD1本あたりのIOPS × 本数
- M=HDD1本あたりの容量 × 本数

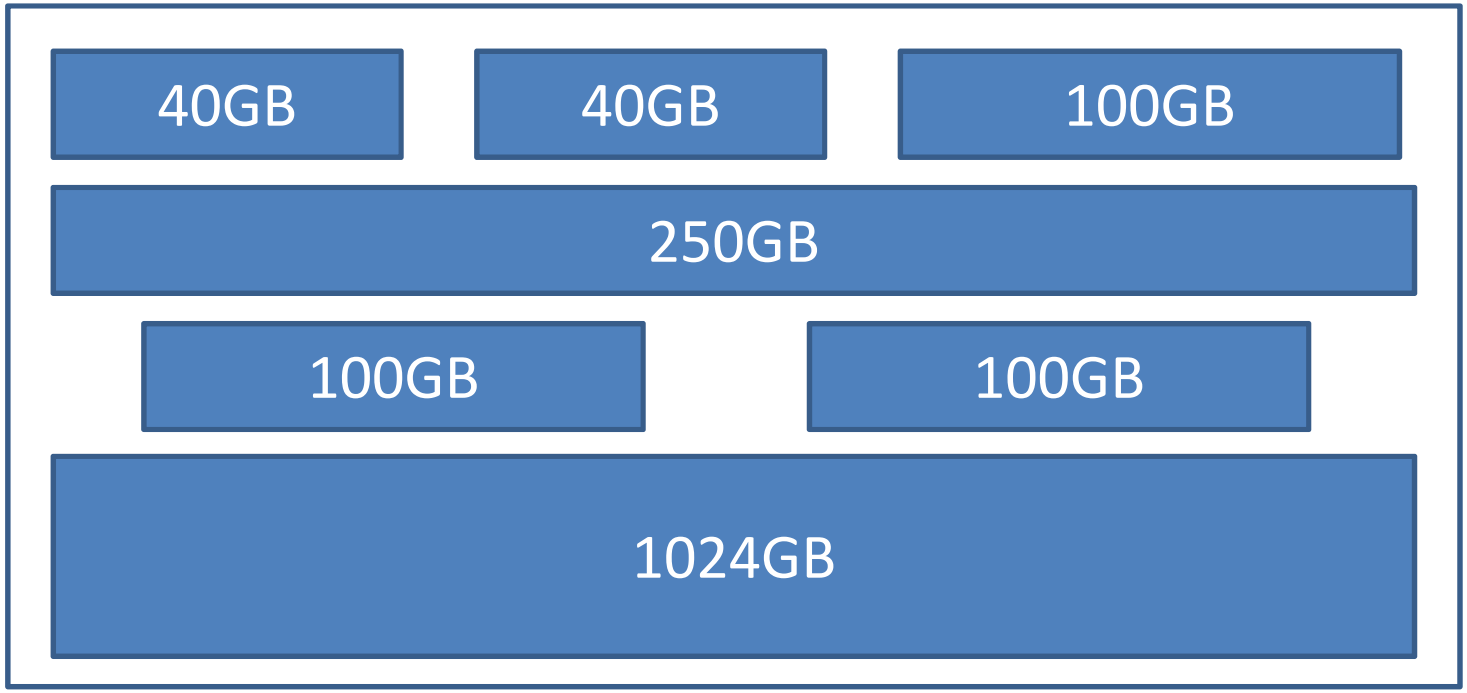
収容設計の考え方（弊社の例）

- 4KiB random writeにてストレージのIOPS性能を測定する。
- 1ユーザあたり平均XXIOPSとして、収容ユーザ数の上限を決定する。
- 容量が無駄にならないように、もしくは容量が先に頭打ちしないように各種パラメータを調整する。

会場のみ

収容設計のイメージ

40～1024GBのユーザを最大N個詰めて、容量が余らないプールを作りたい ⇒ ナップサック問題を解く！



SAS XXXGB 2.5” 10Krpm XX本 RAIDXX
X,XXXIOPS（最大XXXユーザ）、XX,XXXGiB

会場のみ

第一期iSCSIストレージ

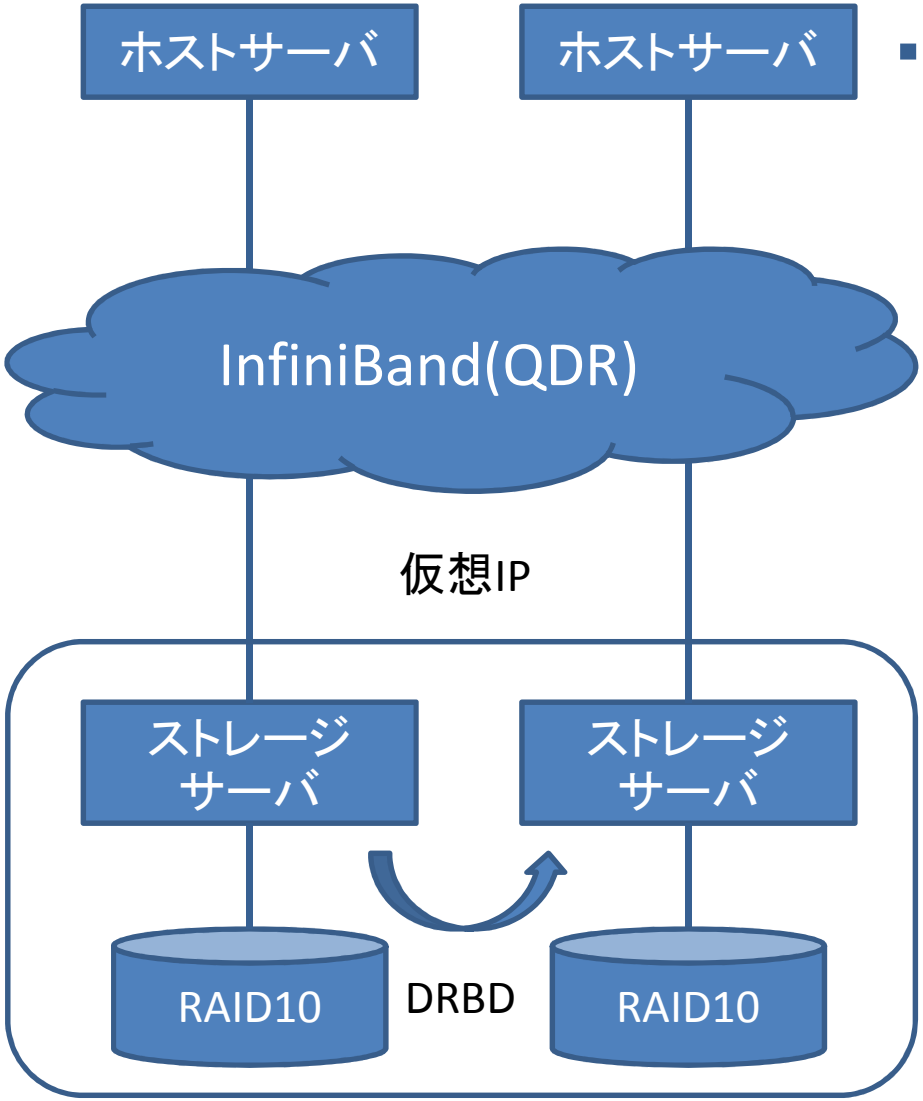
- 2012/6/25～提供中
- 市販サーバにHDD搭載
- IB接続
- OS: CentOS 6.2
- クラスタ制御: Pacemaker + DRBD
- ボリューム制御: LVM
- iSCSI Target: tgtd (scsi-target-utils)

RAID構成、収容ユーザ数

- RAID: RAIDXX
- HDD: XXXGB 2.5” SAS 10Krpm × XX本
- IOPS性能: X,XXXIOPS
- 容量: X,XXXGiB（スナップショット領域500GiB別途）
- 収容ユーザ数: 20GiB × XXXユーザ
- IOPS収容率: XX%
- 容量収容率: XX%

会場のみ

接続構成



150台

- IPoIBによる接続
- ターゲットは仮想IP
- IB経由でデータのミラー

5セット

開発、運用中に生じた課題

- パフォーマンスがでない
  - DRBDのバージョンを8.4から8.3にダウンすることで解消
- DRBDの不具合
  - RHEL6系からbarrier命令に対するカーネルの挙動が変化したことによるデータ不整合
  - no-barrierに変更、Verify+再同期によって解消
  - 参考情報:

    http://www.3ware.co.jp/aboutus/news/news-20120911.html

運用してみた結果

- メリット
  - ストレージサーバの内部まで作りこめるので、クラウドコントローラとの接続を好きにできる。
  - ノード冗長なので、バックプレーンの故障などにも対応できる。
- デメリット
  - ノード冗長により、HDDが2倍必要になる
  - （RAID101） => 4冗長
  - HDDのコストがかかる

第二期iSCSIストレージ

- 2012/8/31～提供中
- コスト削減のため、大容量ユーザ向けに商用のストレージ製品を導入。
- 異なる国内メーカの二機種を評価し、某N社さんのストレージを導入
- 一般的なミッドレンジストレージ
- 10GbE接続、マルチパスによる冗長化
- 大容量ユーザを収容
  - 40,60,80,100,250,500,750,1024GiB

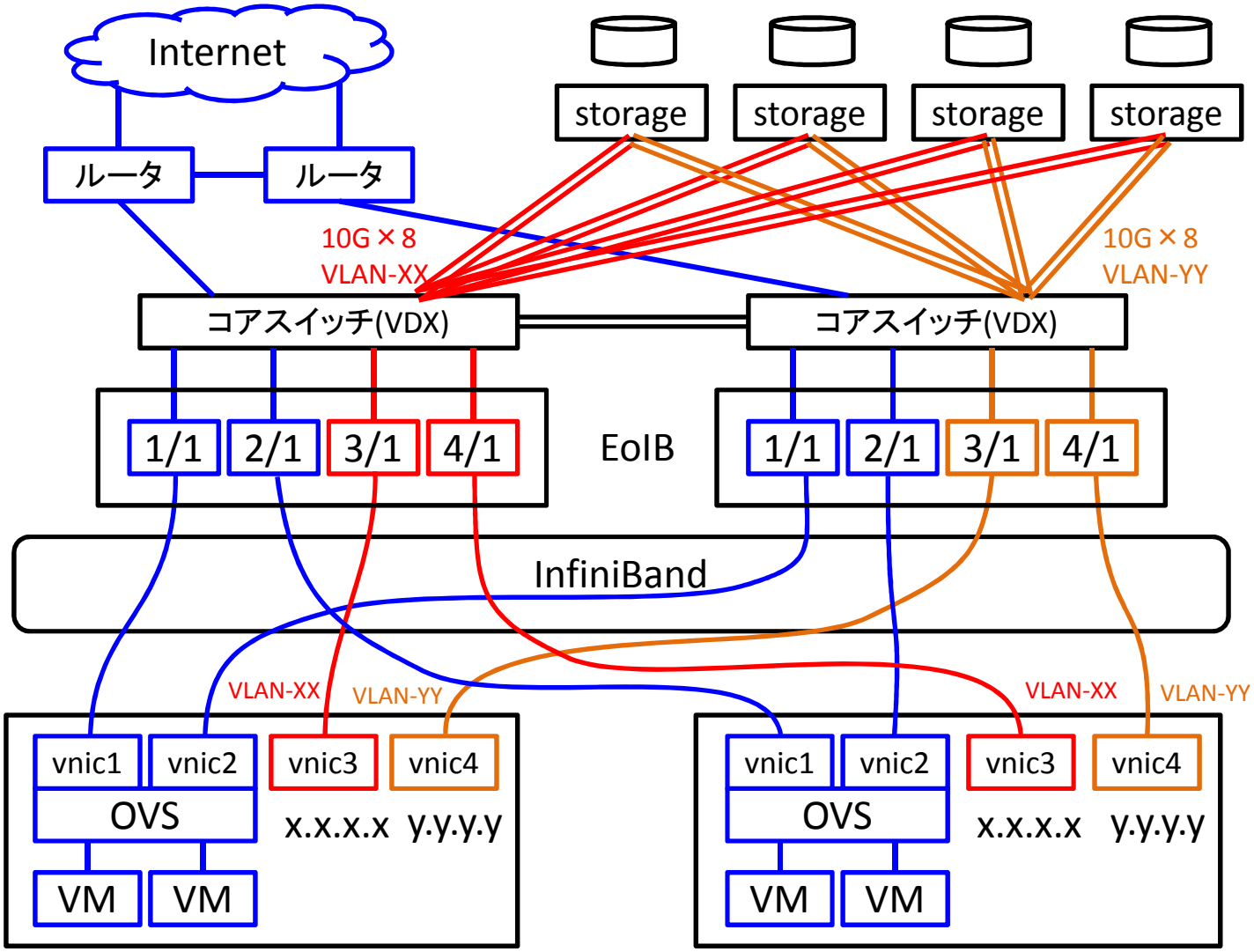
RAID構成、収容ユーザ数

- RAID: RAID60（8+PQ）
- HDD: XXXGB 2.5” SAS 10Krpm × XX本
- ホットスペア: 別途4本
- IOPS性能: X,XXXIOPS
- 容量: XX,XXXGiB（スナップショット領域1,024GiB別途）
- 収容ユーザ数: 最大XXXユーザ
- IOPS収容率: XX%
- 容量収容率: XX%

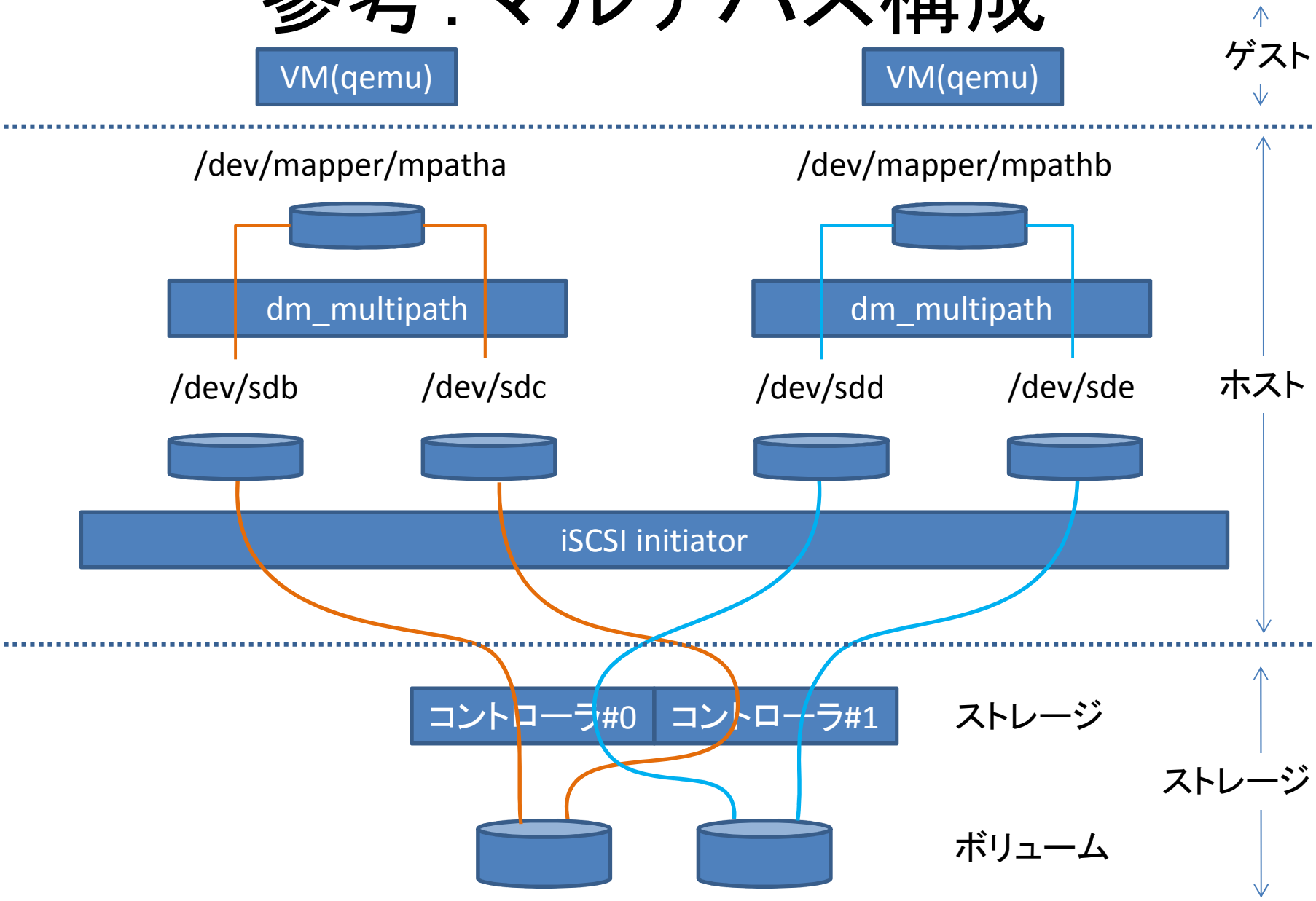
（平均XXXGiBで試算）

会場のみ

接続構成


Internet
ルータ
ルータ
storage
storage
storage
storage
10G × 8
VLAN-XX
10G × 8
VLAN-YY
コアスイッチ(VDX)
コアスイッチ(VDX)
1/1
2/1
3/1
4/1
EoIB
1/1
2/1
3/1
4/1
InfiniBand
VLAN-XX
VLAN-YY
VLAN-XX
VLAN-YY
vnic1
vnic2
vnic3
vnic4
OVS
x.x.x.x
y.y.y.y
VM
VM
vnic1
vnic2
vnic3
vnic4
OVS
x.x.x.x
y.y.y.y
VM
VM

参考：マルチパス構成


VM(qemu)
VM(qemu)
ゲスト
/dev/mapper/mpatha
/dev/mapper/mpathb
dm_multipath
dm_multipath
/dev/sdb
/dev/sdc
/dev/sdd
/dev/sde
ホスト
iSCSI initiator
コントローラ#0
コントローラ#1
ストレージ
ストレージ
ボリューム

参考：マルチパスと仮想IPの違い

- iSCSIイニシエータのタイマー値が全く異なる。
- 仮想IPアドレスによる冗長
  - ストレージの切替りの間、ユーザのIOを保留して切り替わりを待つ。

```
node.conn[0].timeo.noop_out_interval = 0
node.conn[0].timeo.noop_out_timeout = 0
node.session.timeo.replacement_timeout = 86400
```

- マルチパスによる冗長
  - 即座にdm_multipathにIOエラーを返し、切替えを行う。

```
node.conn[0].timeo.noop_out_interval = 5
node.conn[0].timeo.noop_out_timeout = 5
node.session.timeo.replacement_timeout = 5
```

導入の過程で生じた問題

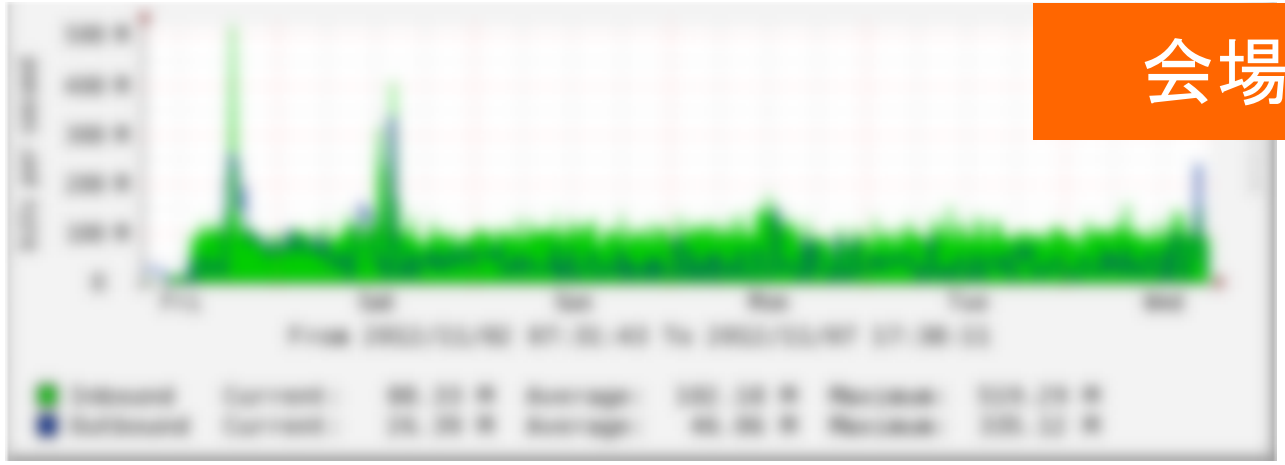
- 検証中にiSCSI回りで不具合が見つかった
  - iSCSIセッションが落ちる
  - コントローラがダウンする
  - エラーiSCSIセッションでIFダウンが発生
  - セッション数が増えるとパフォーマンス低下
- メーカさんのスペシャル対応で即改善 ☺

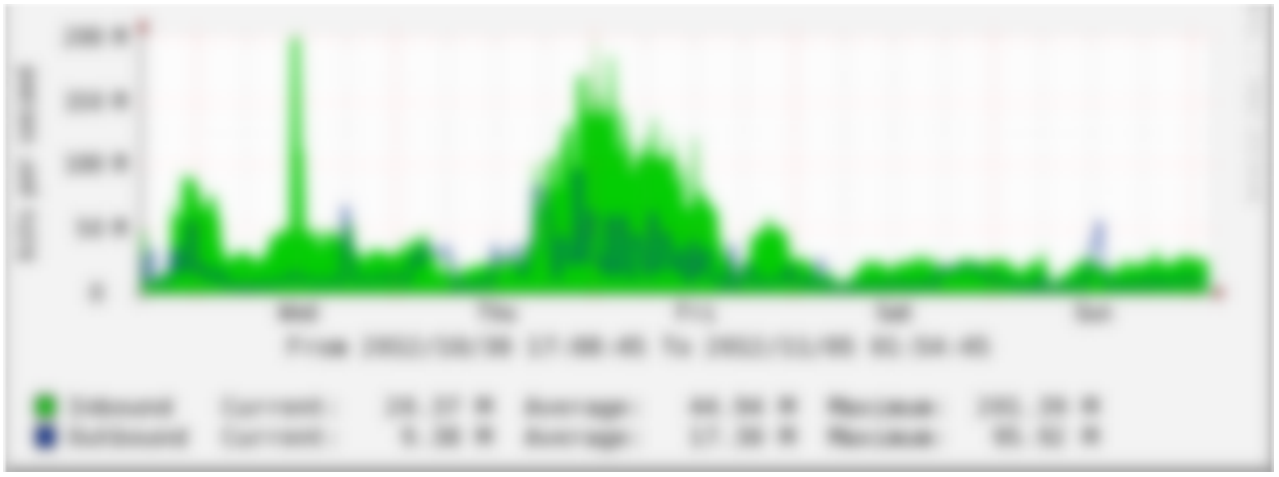
運用してみて

- 検証中にバグを踏みまくったおかげで、現在安定して動作 ☺
- 性能監視を好きなように作れない（メーカ提供の専用アプリが必要）
- ホストサーバにEtherがないので、接続のためにEoIBのリソースを消費
- 帯域はそれほど食わない（次ページ資料）

参考：ストレージの帯域

会場のみ





10Gbpsあれば全然問題ないくらい

SSDストレージ

- 2012/11/1～提供中
- 第一期iSCSIストレージと同じアーキテクチャ
  - 市販サーバ＋DRBD＋IPoIB
- HDDをSSDに換装
- Intel SSDXXX XXXGB XX本
- ホットスペア: 2本（SSD Guard、後ほど詳細）
- 1セットあたり100GiB × XXユーザ収容

会場のみ

SSDとHDD、単体の性能比較

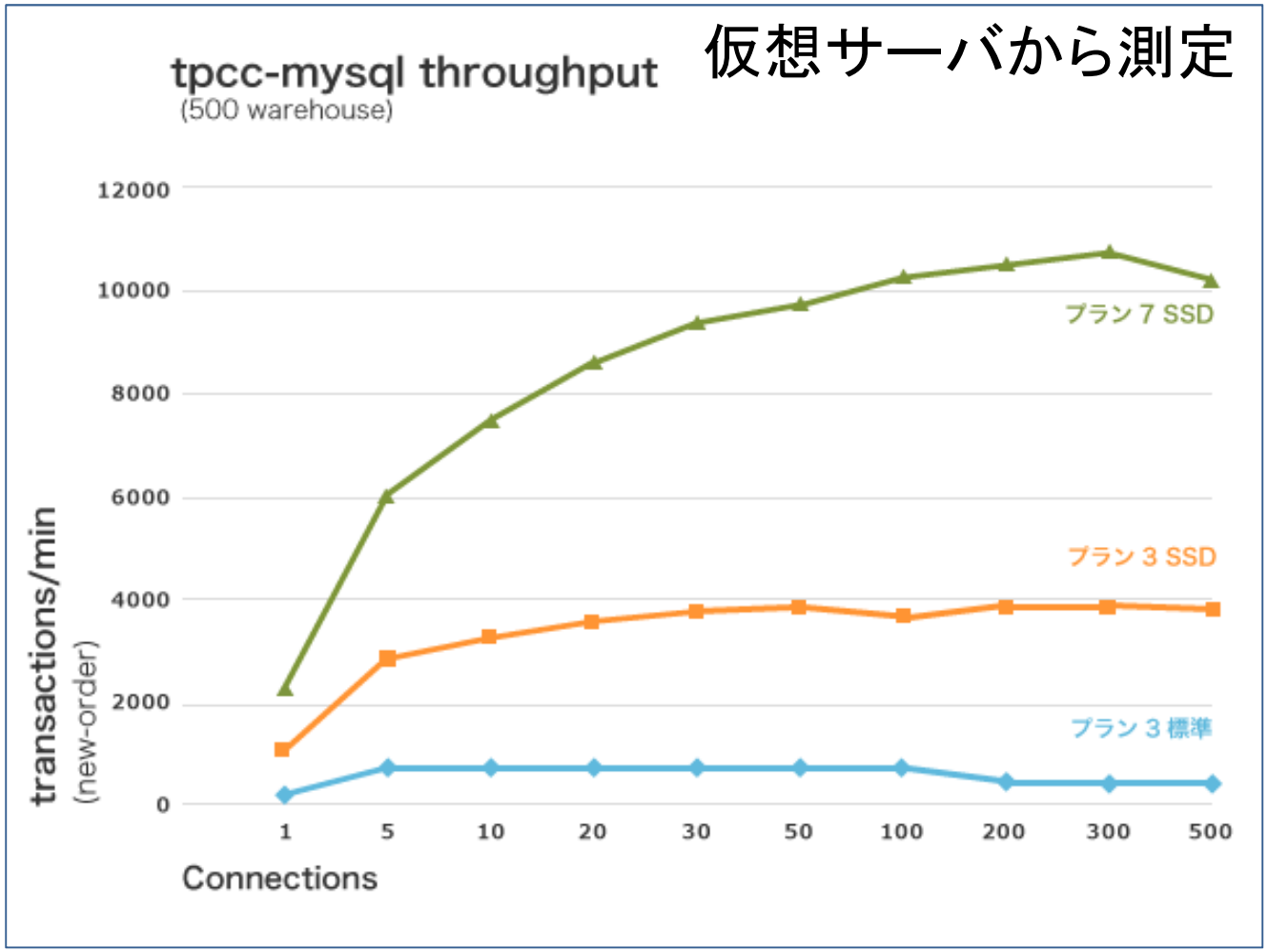
HDD

転送帯域: 150MB/sec
IO性能: 300IOPS
latency: 5ms

SSD

転送帯域: 550MB/sec
IO性能: 50,000IOPS
latency: 0.1ms

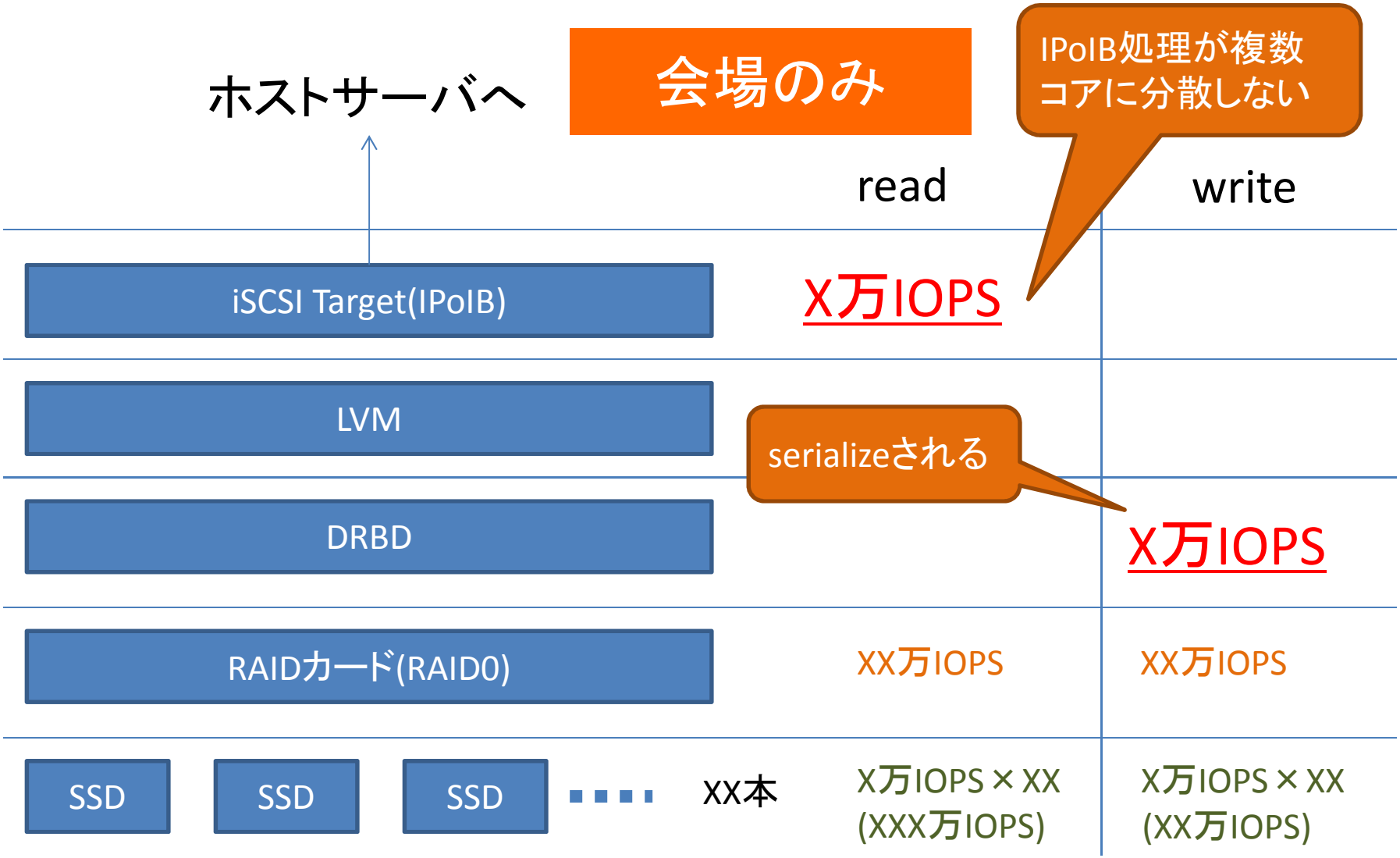
弊社で公開しているベンチ結果



仮想化ドライバの比較

- qemuには、IDEとvirtioの2種類が存在
- IDE
  - 1並列でしかIOを出せない（NCQのような仕組みを提供していない）
  - 実質QD=1
- virtio block
  - 複数IOを並列で出せる
  - デフォルトではQD=32（iSCSIイニシエータに依存）
- SSDはvirtio利用を前提

SSDストレージサーバボトルネック



SSDのEndurance管理

- SMART情報の監視

```
@sac-is1a-iscsi4-st01a
Enc Slot ID DG RAID  E8  E9 Firmware state
 11    0  8  0    1 100 100 Online, Spun Up
 11    1  9  0    1 100 100 Online, Spun Up
 11    2 10  1    0 100 100 Online, Spun Up
 11    3 13  1    0 100 100 Online, Spun Up
 11    4 14  1    0 100 100 Online, Spun Up
<snip>
```

- SSD Guard
  - LSI社の技術
  - RAID0ボリュームで、SMART情報より壊れそうなSSDのデータをホットスペアにコピー

ストレージ以外のチューニング

- ユーザからのIO負荷制限
- IOスケジューラ
- ページキャッシュ
- アライメントの整合性

ユーザからのIO負荷制限

- iSCSIで見えるブロックデバイスに対して cgroupを掛けている
- 一部QDを制限

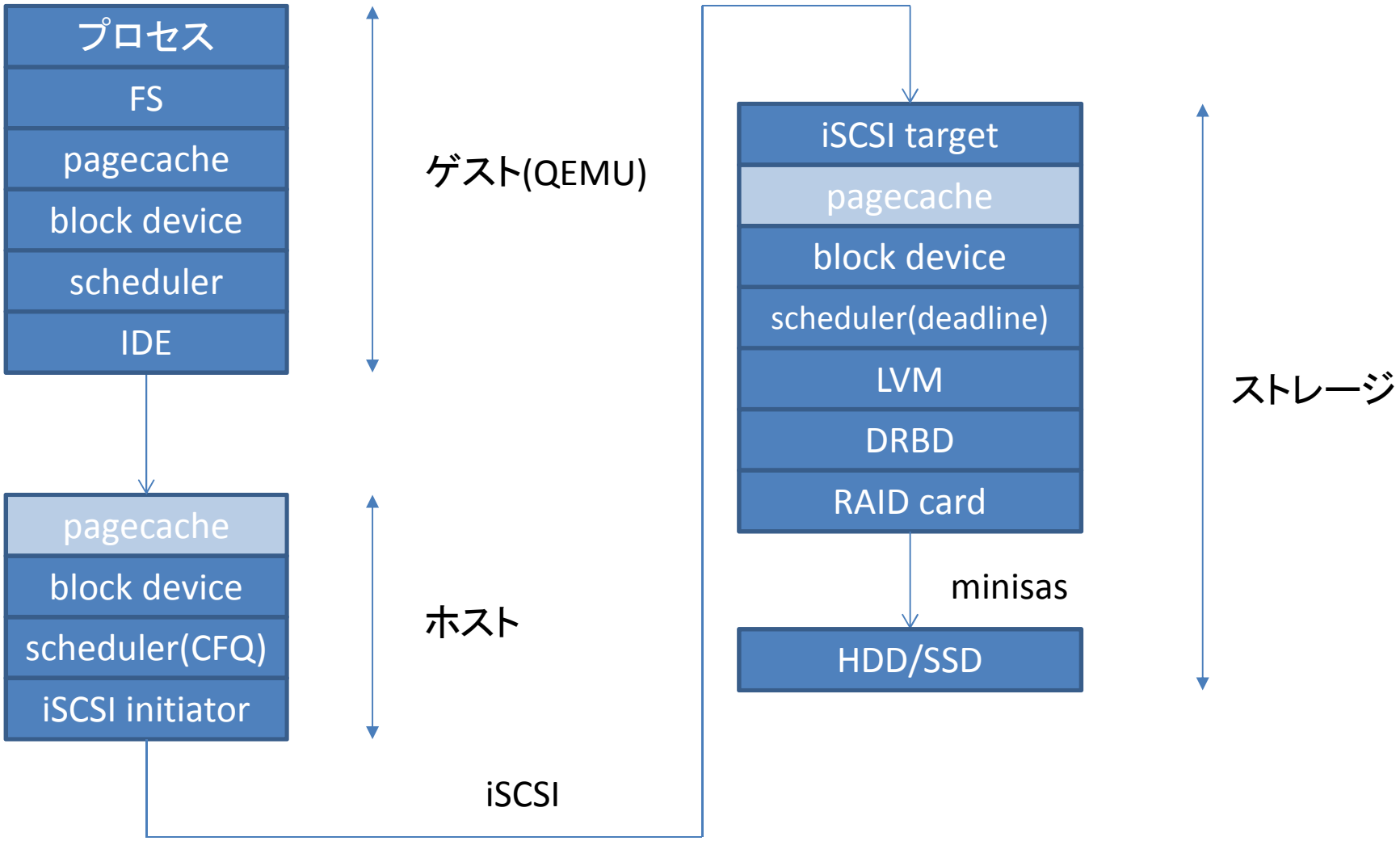
会場のみ

ストレージ	IOPS		BW		QD
	read	write	read	write	
HDD IDE	XXXX	XXXX	XXXMB/s	XXXMB/s	実質1
HDD Virtio	XXXX	XXXX	XXXMB/s	XXXMB/s	XX
SSD	XXXX	XXXX	XXXMB/s	XXXMB/s	実質4

負荷制限の悩みどころ

- OSの起動に時間がかかる
  - ブートストラップが512Bでseq readする
  - Linuxのイメージが約20MBなので約40,000IO程度発生
- readはIOPS制限なしでQDで調整するのがいいのか？

多段IOスケジューラの苦悩

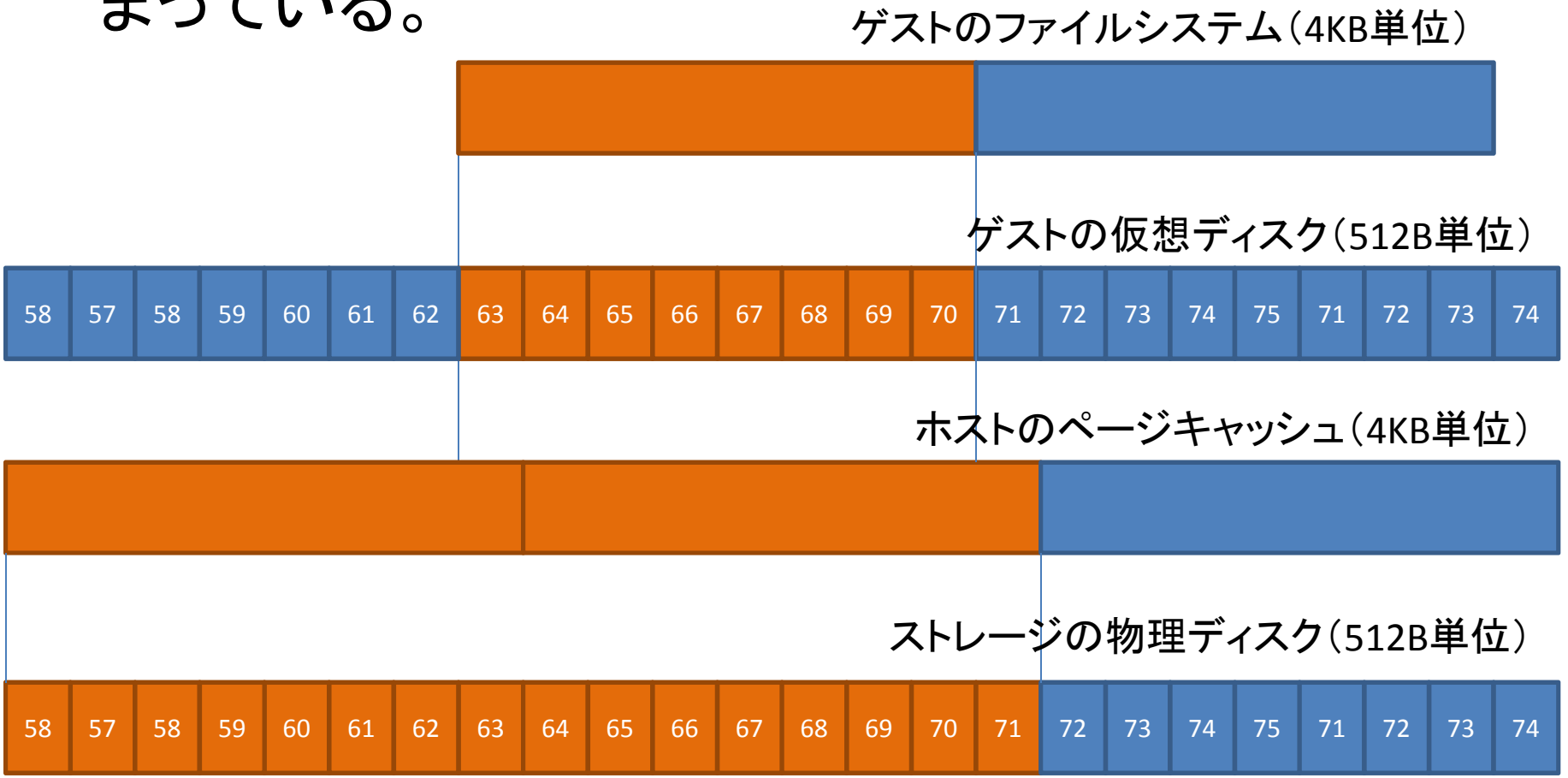


ホストサーバのページキャッシュ

- Linuxはブロックデバイスに対して4KiB単位でメモリにキャッシュする⇒一見よさそうだが？
- cgroup(IO制限)が意図通りに動かない
  - cgroupはページキャッシュの下で動いている
- 性能が安定しない
  - ホストサーバのメモリが減ると遅くなる
  - メモリが少ないVMプランでもIOが爆速
  - ユーザの不公平感がある
- 4KiBアライメントずれによる性能劣化
- ページキャッシュは使わない（cache=none）

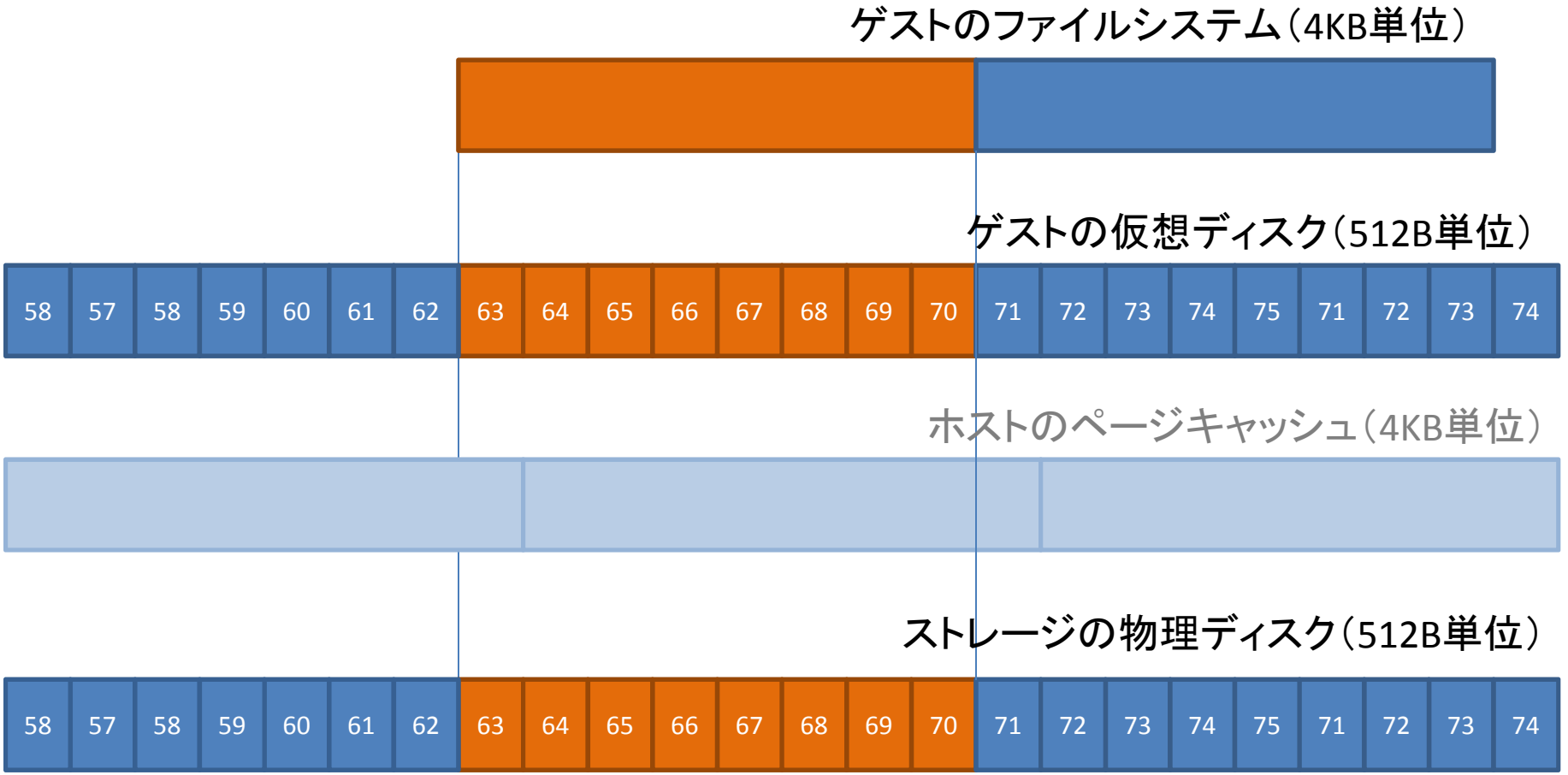
アライメントのミスマッチ

- 古いOSでは、パーティションが63セクタ目から始まっている。

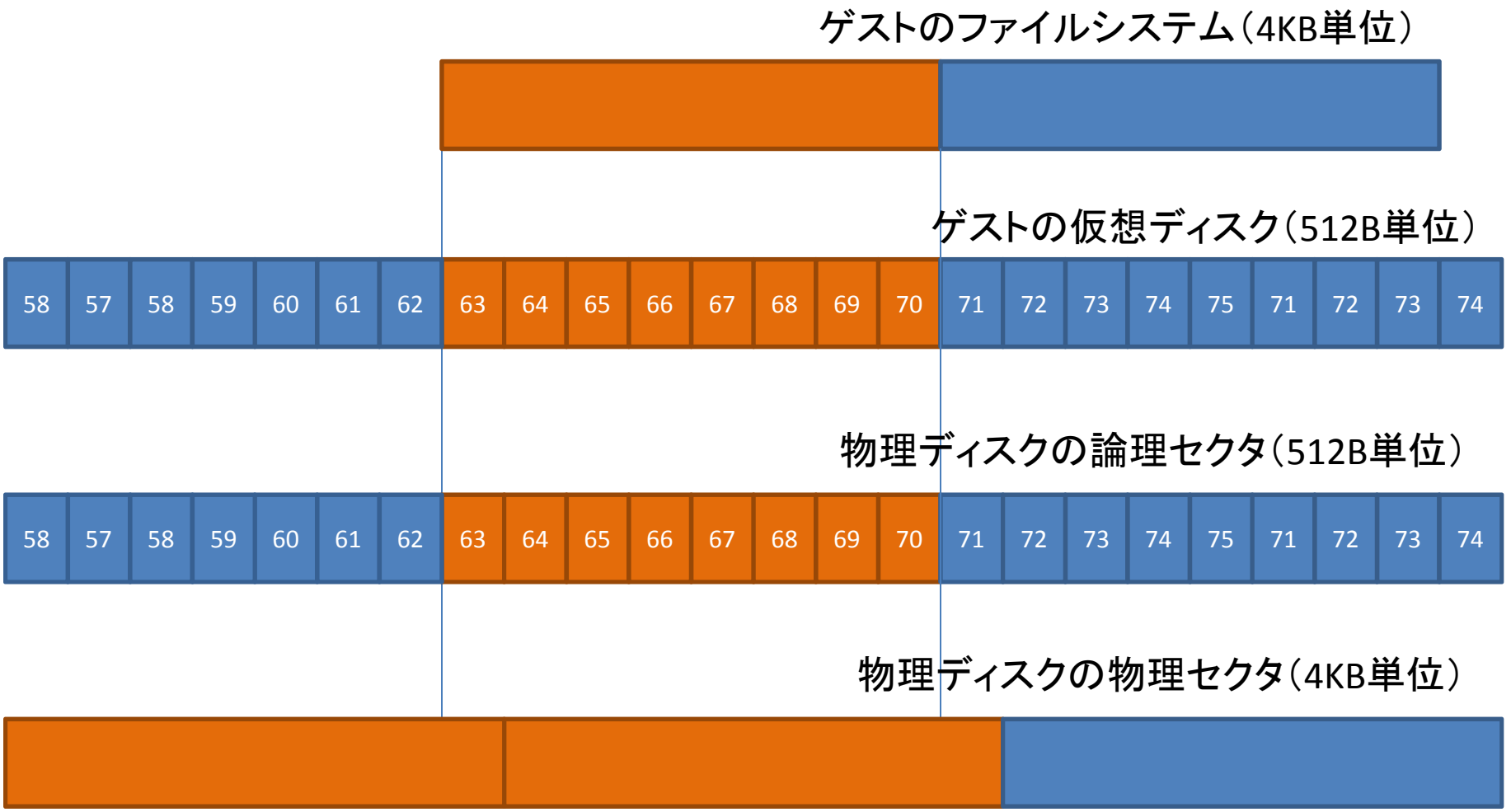


アライメントのミスマッチ

- ページキャッシュをバイパスすれば問題ない



AFT、SSDでも同様の問題が再現



個AFT、SSDの場合（ディスクの内部で処理）

今後

- 引き続き、安定稼働、安定供給に注力する。
- しばらくは、２種類のアーキテクチャを導入
- 今後取り組みたい課題
  - HDDへのIO負荷低減対策
  - 小容量のHDDストレージをSSDで実装
  - 階層化ストレージ、シンプロビジョニング、その他
  - データバックアップ処理の改善
  - iSCSIセッション制限の克服

まとめ

- ネットワークの使用帯域はどんどん増える。増速によって対応。L2ネットワークのボトルネックは、将来的にSDN的なソリューションによって解消を図りたい。
- ストレージIOは、HDDを使う限り最初から「詰んでいる」状態。ゲストとストレージ間の中間層の最適化、SSDの導入などで、負荷低減、オフロードを図る必要がある。