

*Gfarm*ファイルシステム on Azure

*Gfarm*ワークショップ2016

2016年10月21日
株式会社H P Cソリューションズ
河野 証



Gfarmファイルシステム on Azure

内容

1. HPCソリューションズ会社概要
2. Microsoft Azure概要
3. クラウド環境におけるGfarmの利用イメージ
4. 性能テスト
5. Azure概算費用

H P Cソリューションズ会社概要



会社概要

- 株式会社H P Cソリューションズ

- 本社： 東京都中央区日本橋大伝馬町3-2
- 大阪： 大阪府大阪市淀川区宮原4-5-36

- PCクラスタを中心としたシステム販売

- 主な取り扱い製品

- | | |
|------------------------|--------------|
| • DELL, HP, Supermicro | ～ PCクラスタシステム |
| • Mellanox Infiniband | ～ 超高速ネットワーク |
| • DataDirect Networks | ～ 大規模ストレージ |
| • Intelソフトウェア | ～ プログラム開発環境 |

- コンサルティング

- PCクラスタシステム構築
- Peta Scaleストレージシステム構築
- 大規模システムの運用支援
- 解析プログラムの並列化、高速化、GPGPUプログラミング、コンサル
- Eco & Site Planning

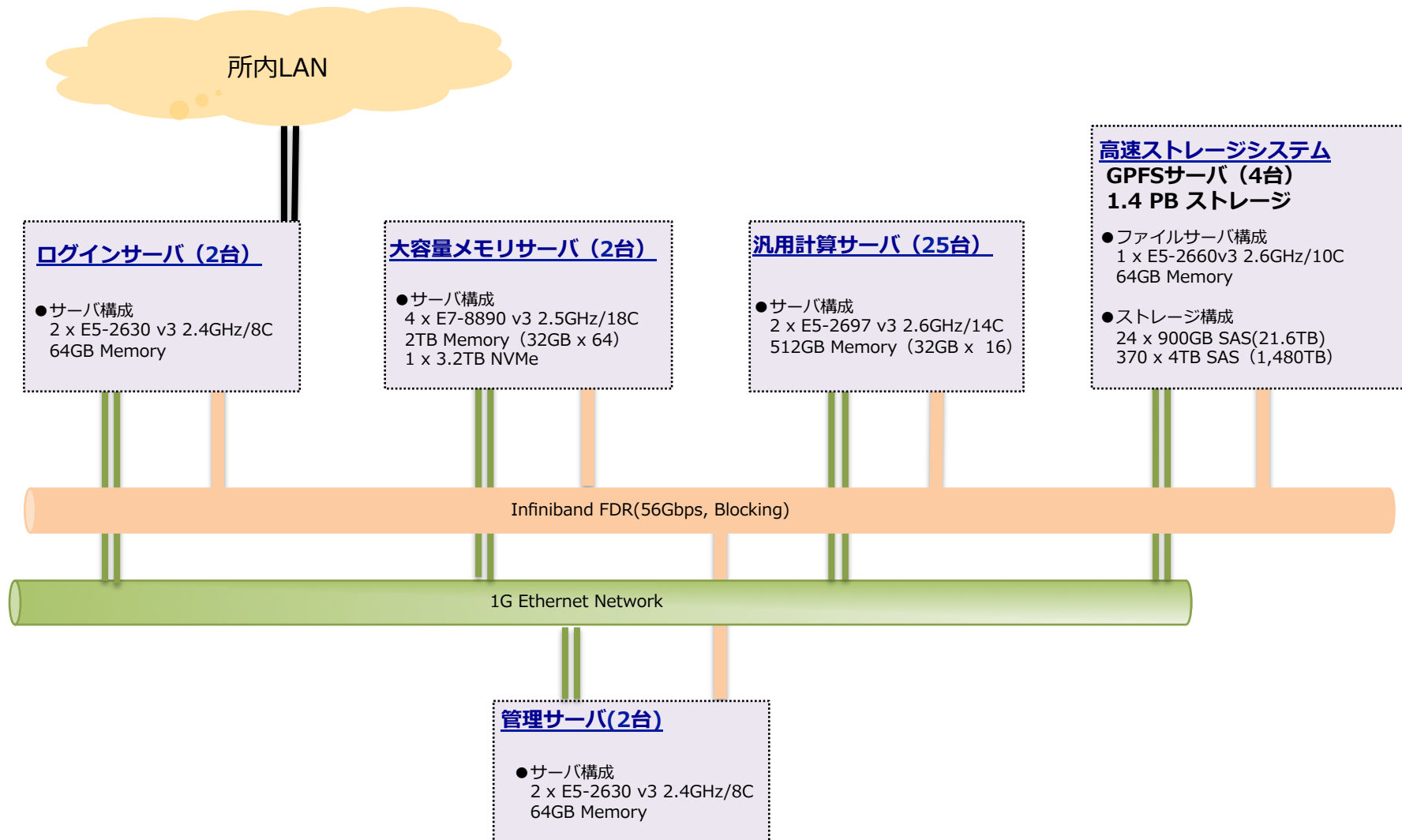


PCクラスタシステム トータルソリューション

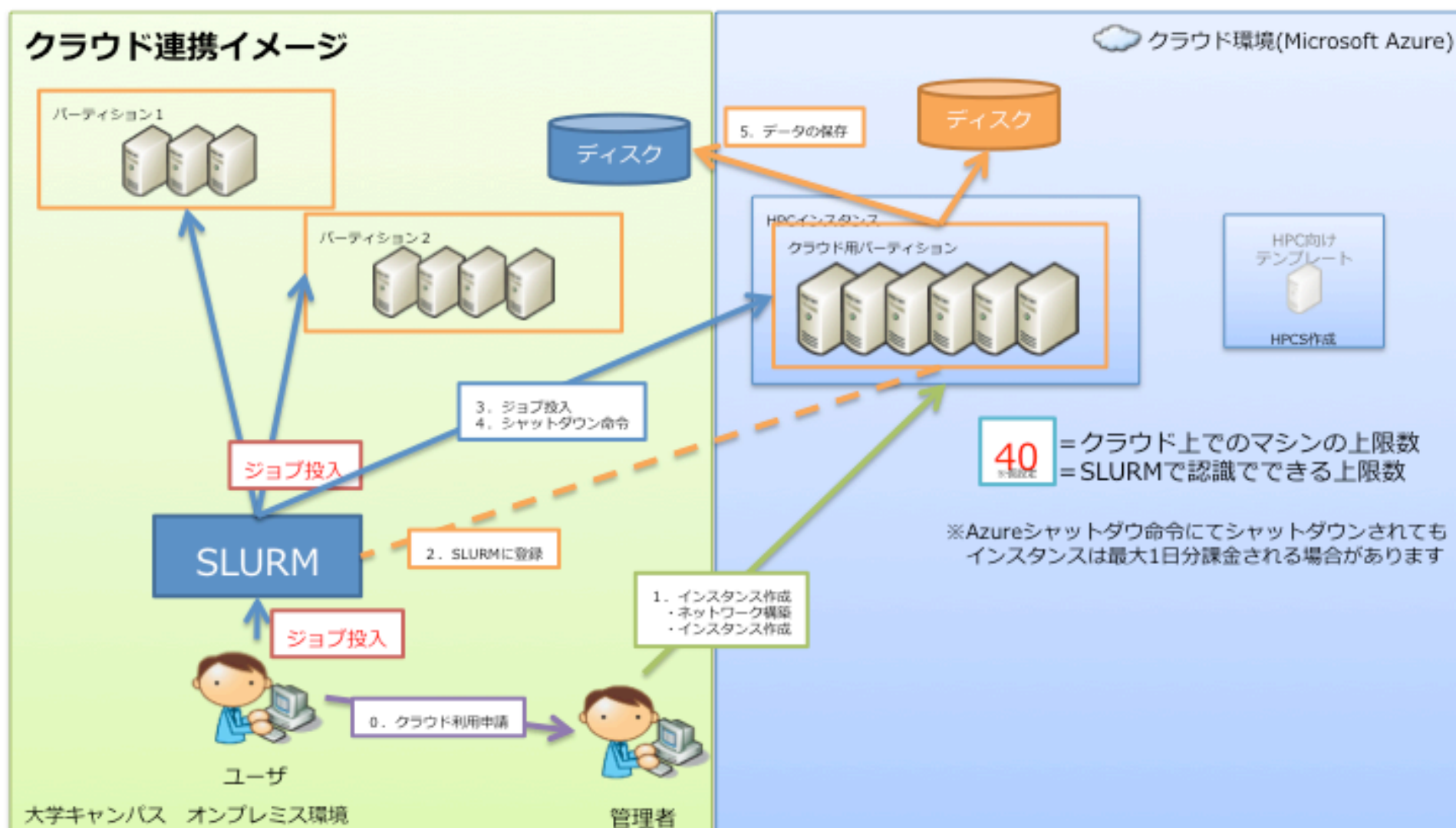


※ 機器設定、システム構築など、一部のサービスのみでもご提供可能です。まずはご相談ください。

国立研究所 様（大規模ゲノムシーケンス解析システム）

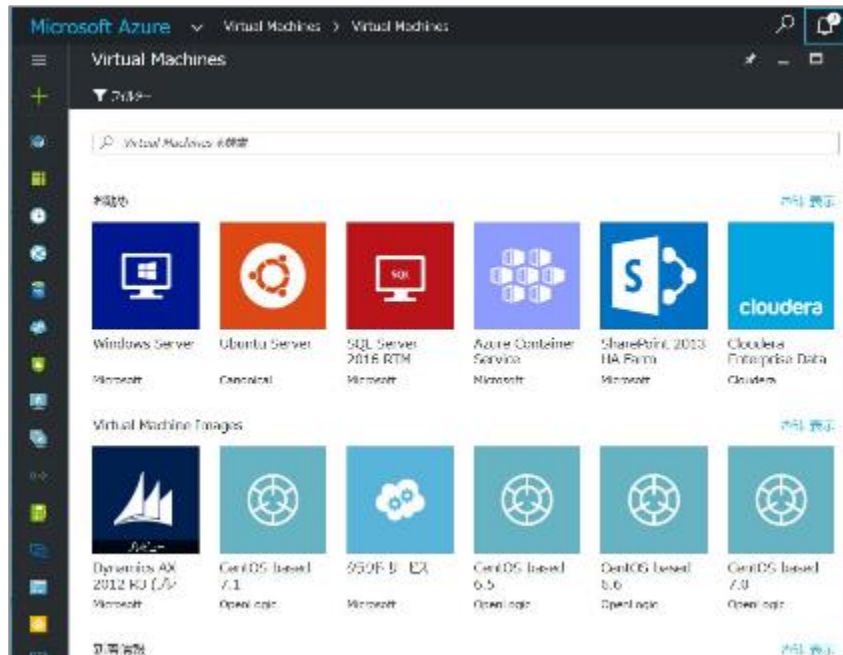


国立大学 様（オンプレミス+クラウド連携システム）



Microsoft Azure（アジュール）概要

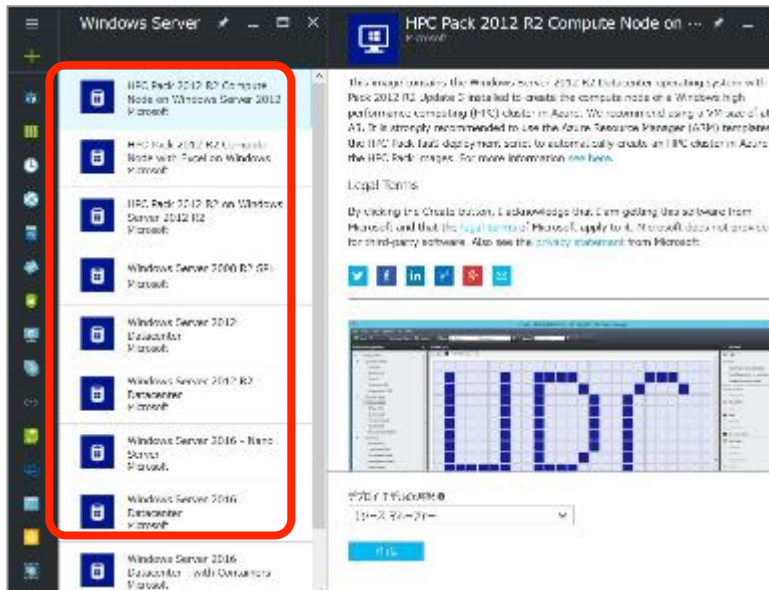
Microsoft Azure『仮想マシン』サービス



- Windows Server あるいは 各種 Linux の「OS 導入済み仮想マシンイメージ」が多数用意されており、それらを選択することでサーバー環境を容易に構築することが可能です。
- また、必要なアプリケーションを導入するなどカスタマイズを施した仮想マシンを「自分用のカスタムイメージ」として登録できます。
- HPC Pack を導入済みの仮想マシンイメージも用意してあります。

様々な仮想マシンイメージをご提供

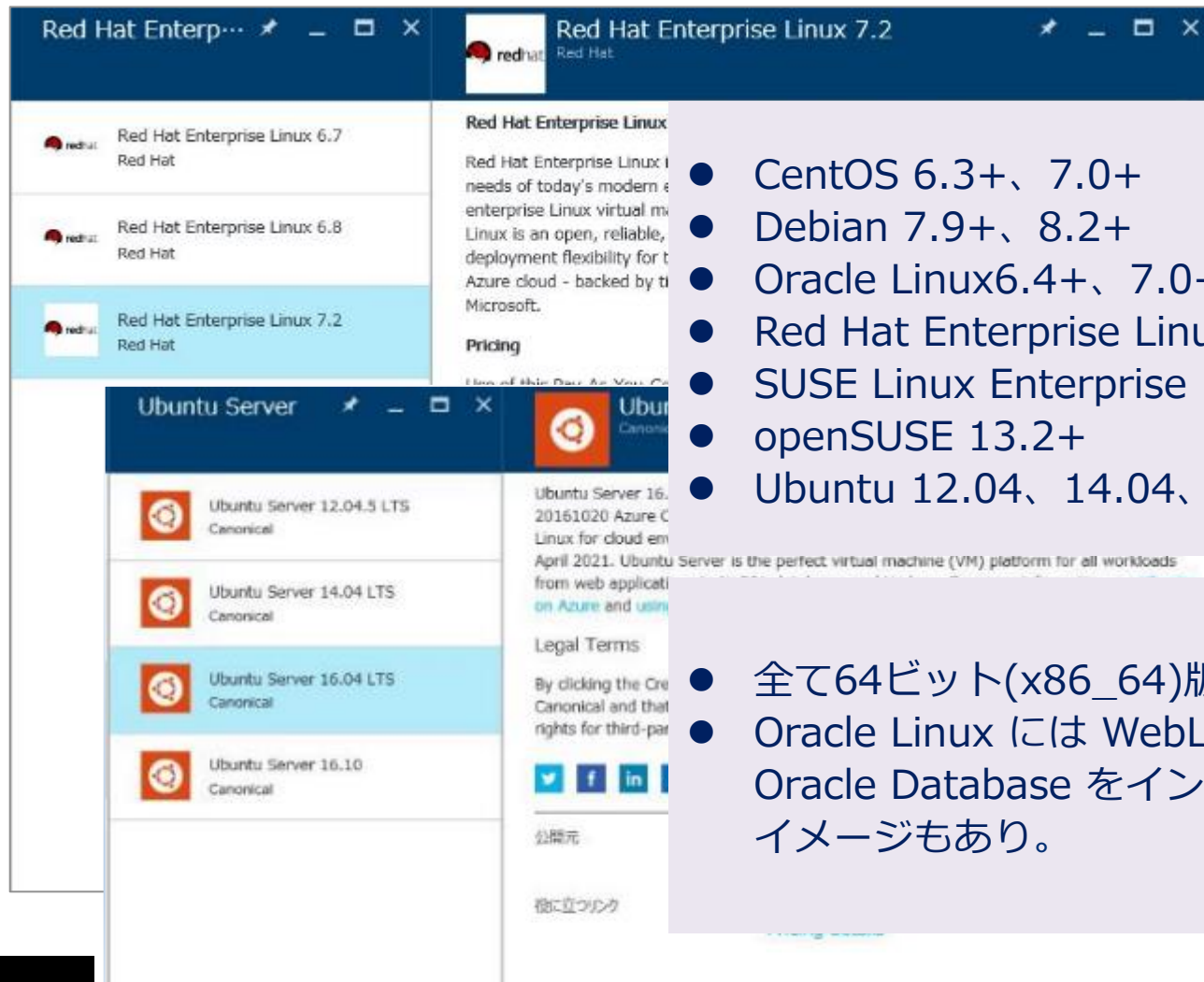
基本的な Windows Server



- Windows Server 2016
- Windows Server 2012 R2
- Windows Server 2012
- Windows Server 2008 R2 SP1
- Windows Server Essentials Experience
- HPC Pack

- OS インストール済み。すぐに使い始められます。
- Windows Server のライセンスは仮想マシンの料金に含まれており、別途調達は不要です。
(Windows Server CALも含まれています)
- OS は英語版ですが、日本語化可能です。
- 日本語化などのカスタマイズを施した上で「マイ イメージ」として登録しておけば、2台目以降の展開が楽になります。

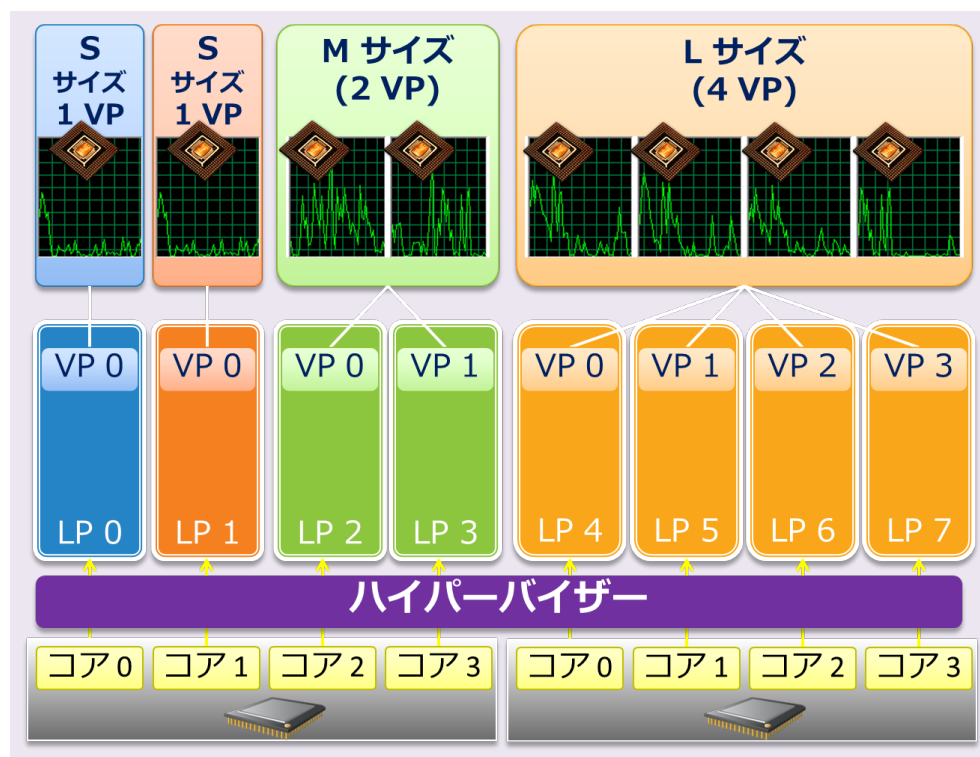
様々な仮想マシンイメージをご提供 複数の Linux ディストリビューションを用意



- CentOS 6.3+、7.0+
- Debian 7.9+、8.2+
- Oracle Linux 6.4+、7.0+
- Red Hat Enterprise Linux 6.7+、7.1+
- SUSE Linux Enterprise 11 SP4、SLES 12+
- openSUSE 13.2+
- Ubuntu 12.04、14.04、および 16.04

- 全て64ビット(x86_64)版
- Oracle Linux には WebLogic Server や Oracle Database をインストール済みのイメージもあり。

Azure VMの仮想CPUは物理コアを占有



- オンプレミスの Hyper-V と異なり、Azure 上の仮想マシンは、LP と VP の比が 1:1 です。
- つまり、物理的なコアが一つ丸ごと、仮想マシン上の CPU コアになります。
- 「他の仮想マシン」とコアを共有することがないため、物理コアの性能をすべて享受できます。

※ ただし、A0 (XS)サイズだけは例外で、複数の仮想マシンでコアを共有します。
(LP:VP が 1:Nになります)

Azureリージョン



ワールドワイドに30リージョン
日本国内に2つのリージョン
(リージョンは増設進行中)

場所

アメリカ

リージョン	場所
※国東部	バージニア州
※国東部 2	バージニア州
※国中部	アイオワ州
※国中北部	イリノイ州
※国中部	テキサス州
※国中部	米田中部
※国西部	カリフォルニア州
※国西部 2	米田西部 2
※国政府バージニア	バージニア州
※国政府アイオワ	アイオワ州
カナダ東部	ケベックシティ
カナダ中部	トロント
ブラジル東部	サンパウロ州

新規発表

※国防総務東部	後日発表予定
※国防総務西部	後日発表予定

ヨーロッパ

リージョン	場所
北ヨーロッパ	アイルランド
西ヨーロッパ	オランダ
ドイツ中部	フランクフルト
ドイツ北東部	マクデブルク
英田西部	ハーディフ
英田南部	ロンドン

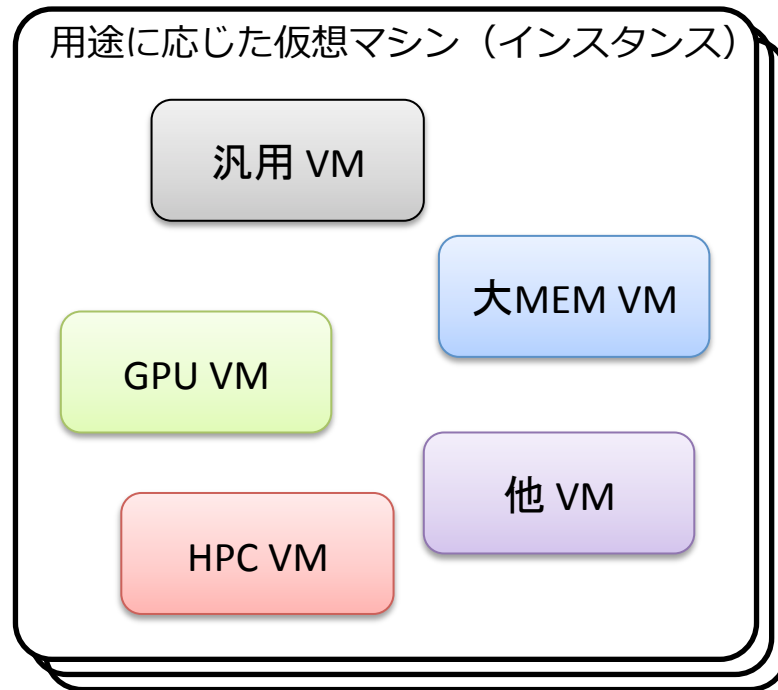
アジア

リージョン	場所
東南アジア	シンガポール
東アジア	香港
オーストラリア東部	ニュー・サウス・ウェールズ州
オーストラリア東部 2	ビクトリア州
インド中部	ブネー
インド西部	ムンバイ
インド南部	チェンナイ
東日本	東京、埼玉
西日本	大阪
中国東部	上海
中国北部	北京

新規発表

韓国中部	ソウル
韓国南部	後日発表予定

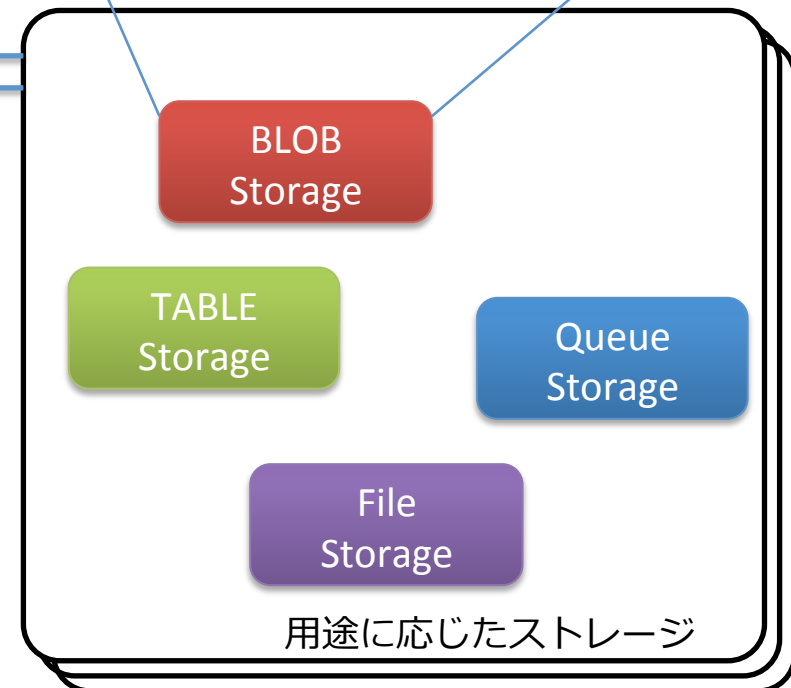
仮想マシンと仮想ストレージ(1)



- 汎用 VM
E5プロセッサシリーズ
- 大MEM VM
～ 448GB RAM
- GPU VM
GPGPU搭載
- HPC VM
高クロックCPU
Infiniband対応

- 性能オプション
Standard: HDD
Premire: SSD
- 冗長オプション
LRS : 3xコピー(同一施設内) デフォルト
ZRS : 3xコピー(1-2リージョン、2-3施設内)
GRS : 6xコピー(プライマリリージョンに3
セカンダリリージョンに3)
RA-GRS : GRS+セカンダリリージョンへの
Read-only Access

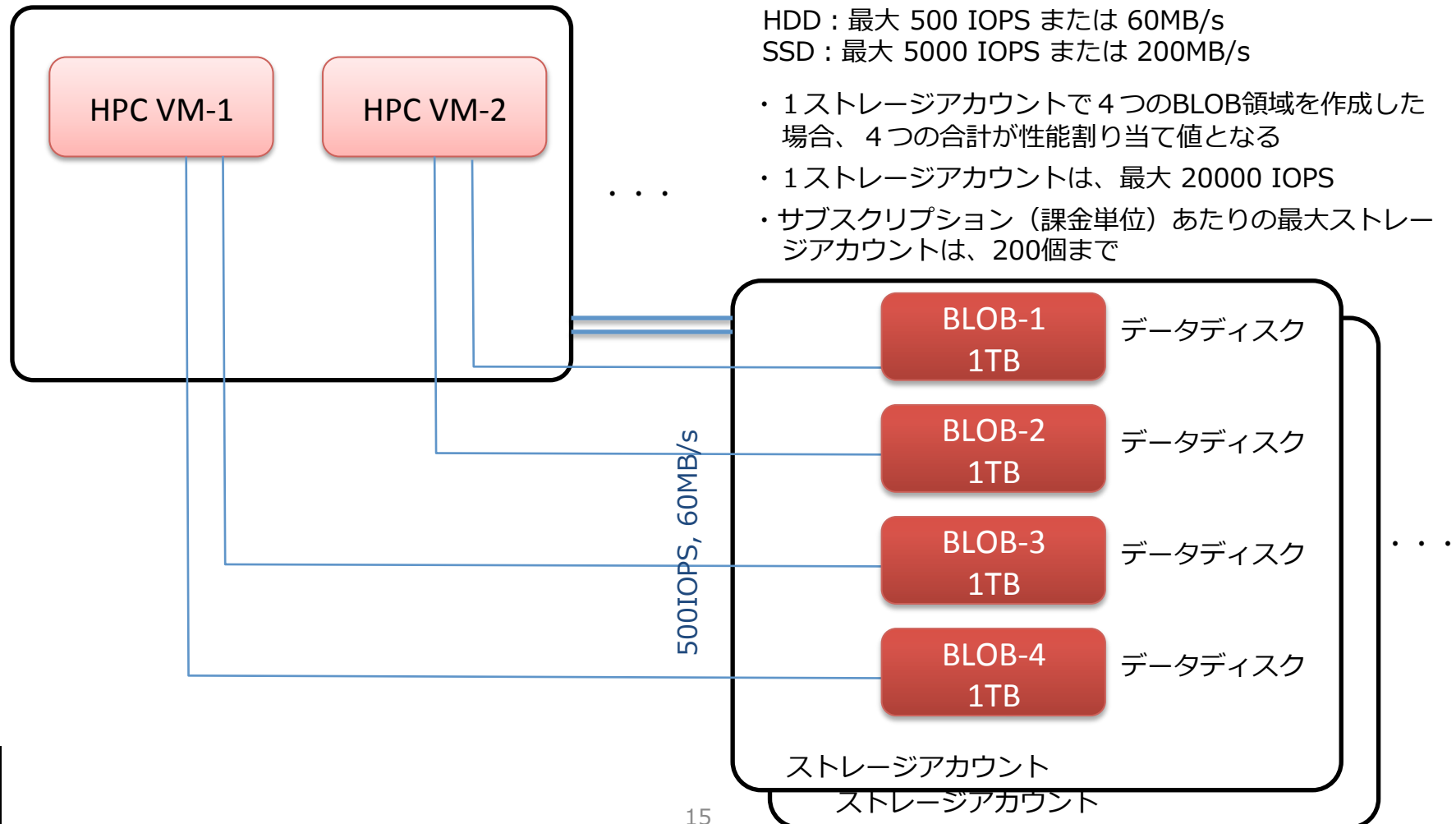
...



...

仮想マシンと仮想ストレージ(2)

- データディスクサイズ
HDD, SSDいずれも最大約1TB
- マウント数
VM (インスタンス) 毎に異なる (1~64ディスク)
- ストレージアカウント
HDD : 最大 500 IOPS または 60MB/s
SSD : 最大 5000 IOPS または 200MB/s
 - ・ 1ストレージアカウントで4つのBLOB領域を作成した場合、4つの合計が性能割り当て値となる
 - ・ 1ストレージアカウントは、最大 20000 IOPS
 - ・ サブスクリプション (課金単位) あたりの最大ストレージアカウントは、200個まで

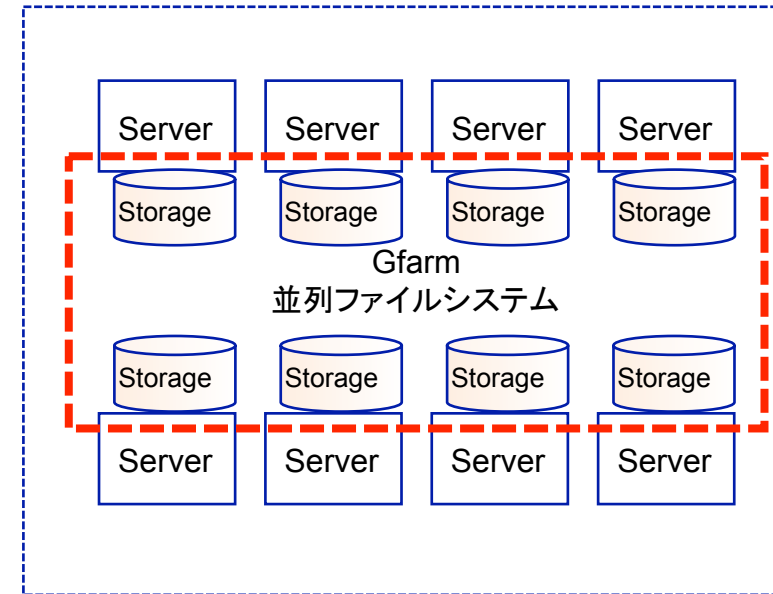
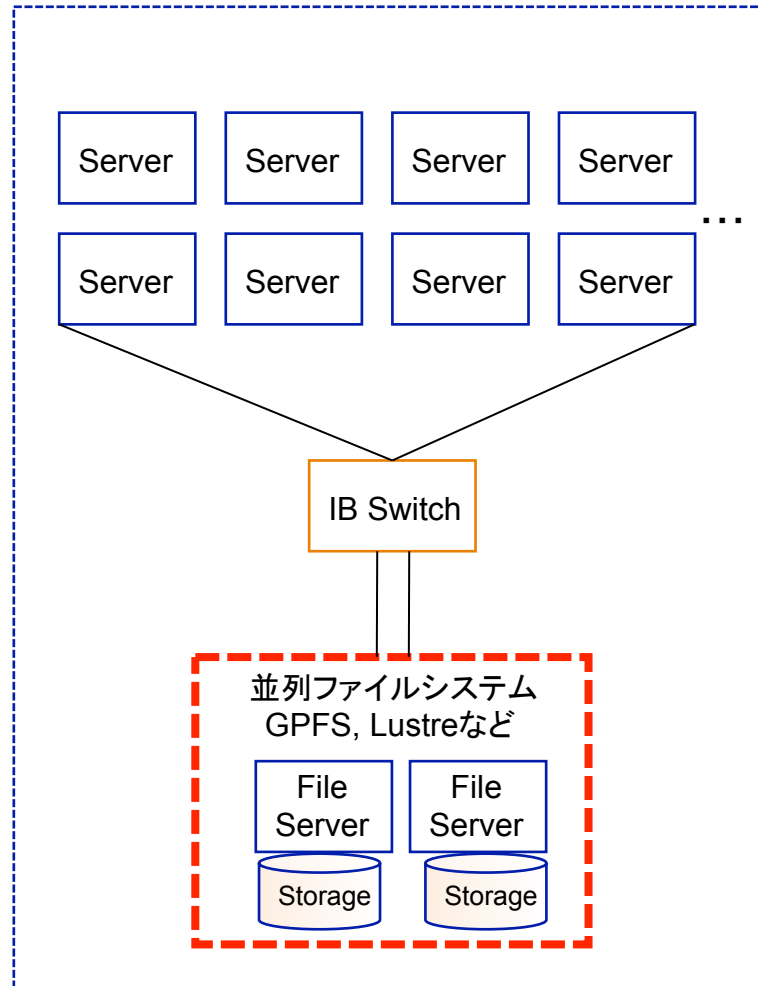


クラウド環境におけるGfarmの利用イメージ

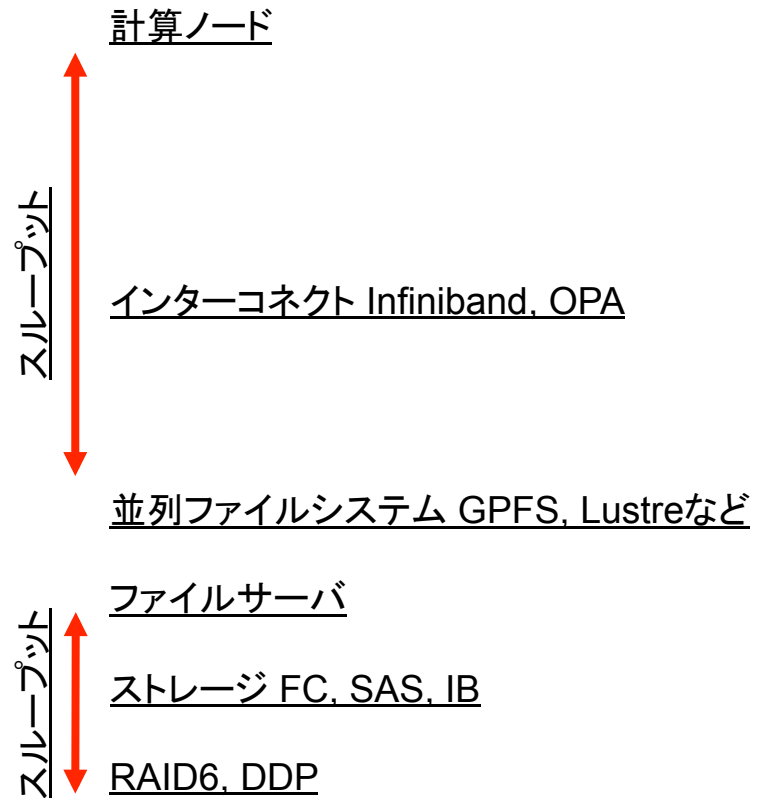
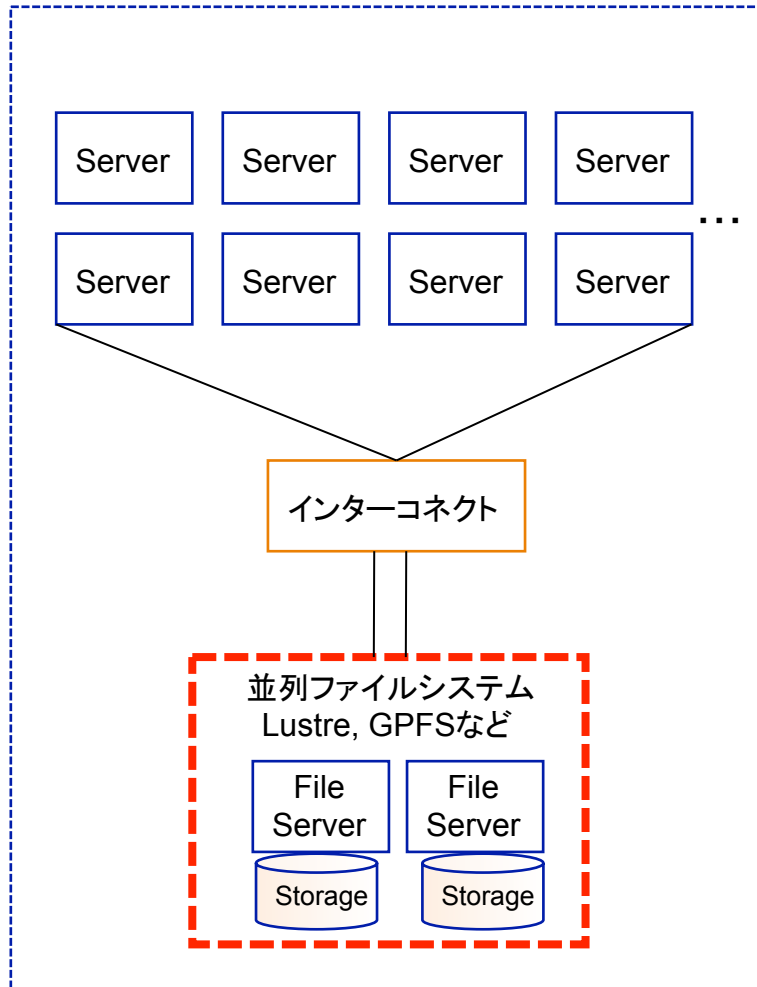
Gfarm利用イメージ1

HPCクラスタシステムの並列ファイルシステム

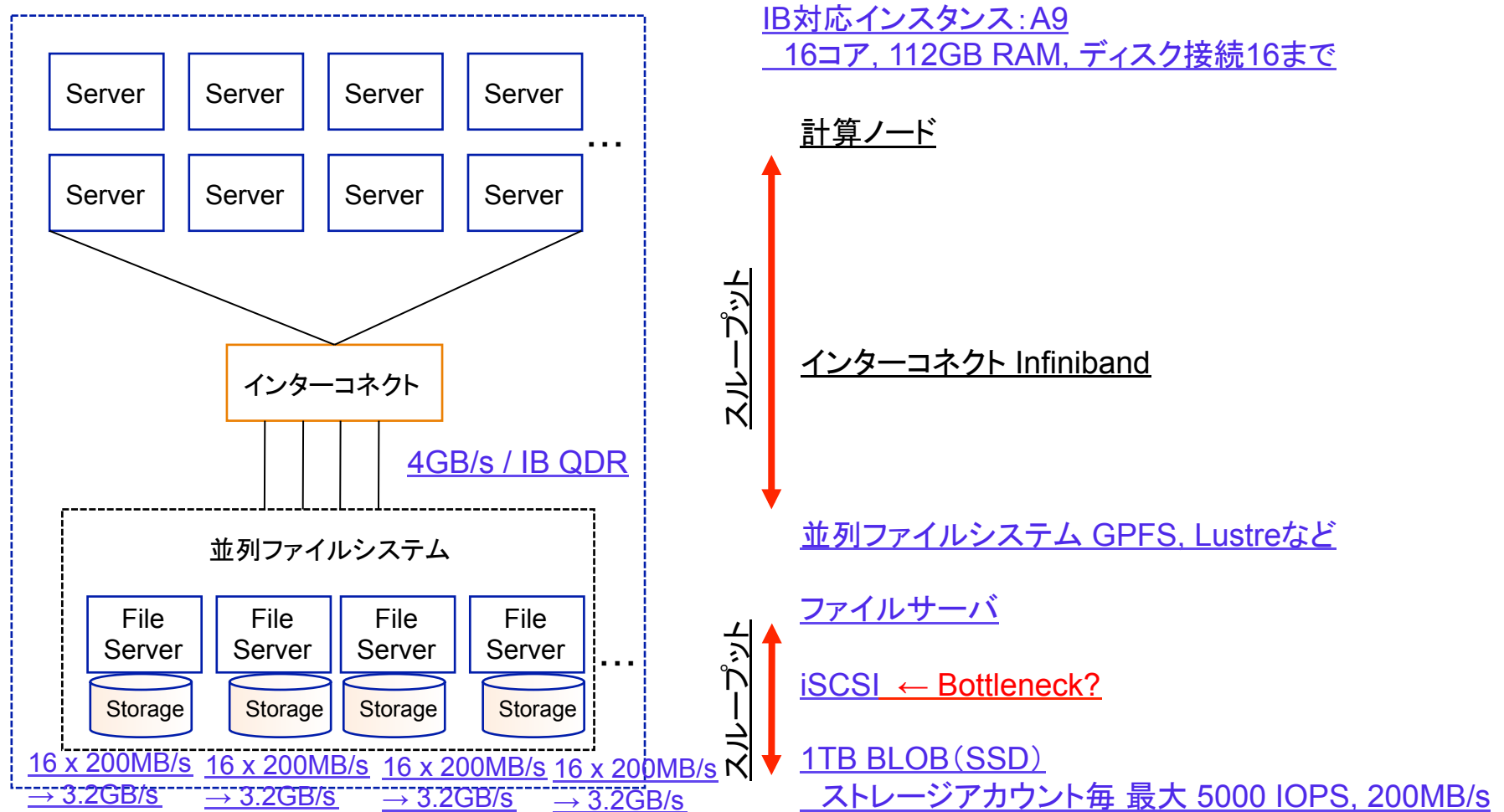
Gfarm利用イメージ1：HPCクラスタシステムの並列ファイルシステム



オンプレミス : Lustre, GPFS スループットを考慮した設計例



Azure : Lusre, GPFS スループットを考慮した設計例



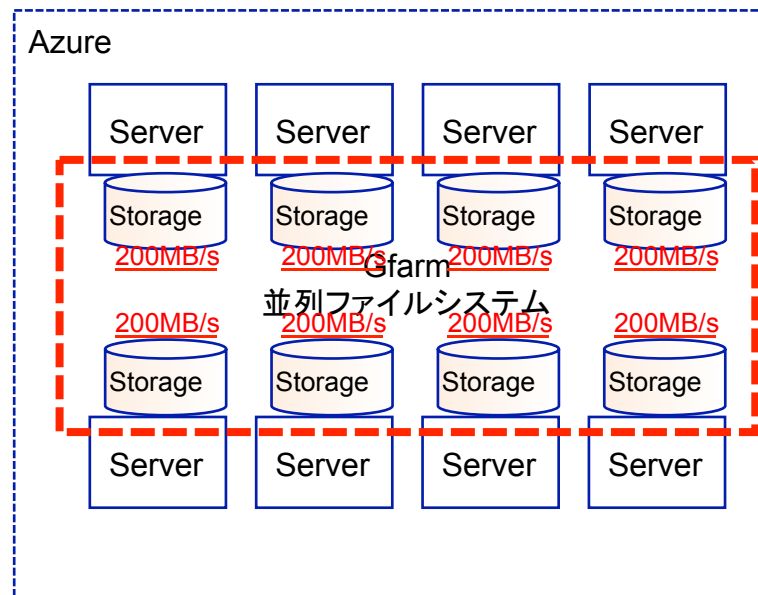
※ 最大200 x ストレージアカウント／サブスクリプション → 最大200TB, 40GB/s
 但し、ファイルサーバ～Storage間の帯域次第

Azure : Gfarm

スループットを考慮した設計例

IB対応インスタンス:A9

16コア, 112GB RAM, ディスク接続16まで



- スループット
 $200 \text{ MB/s} \times \text{ストレージアカウント数} \times \text{Server数}$

- 最大スループット
ストレージアカウント上限 : 200

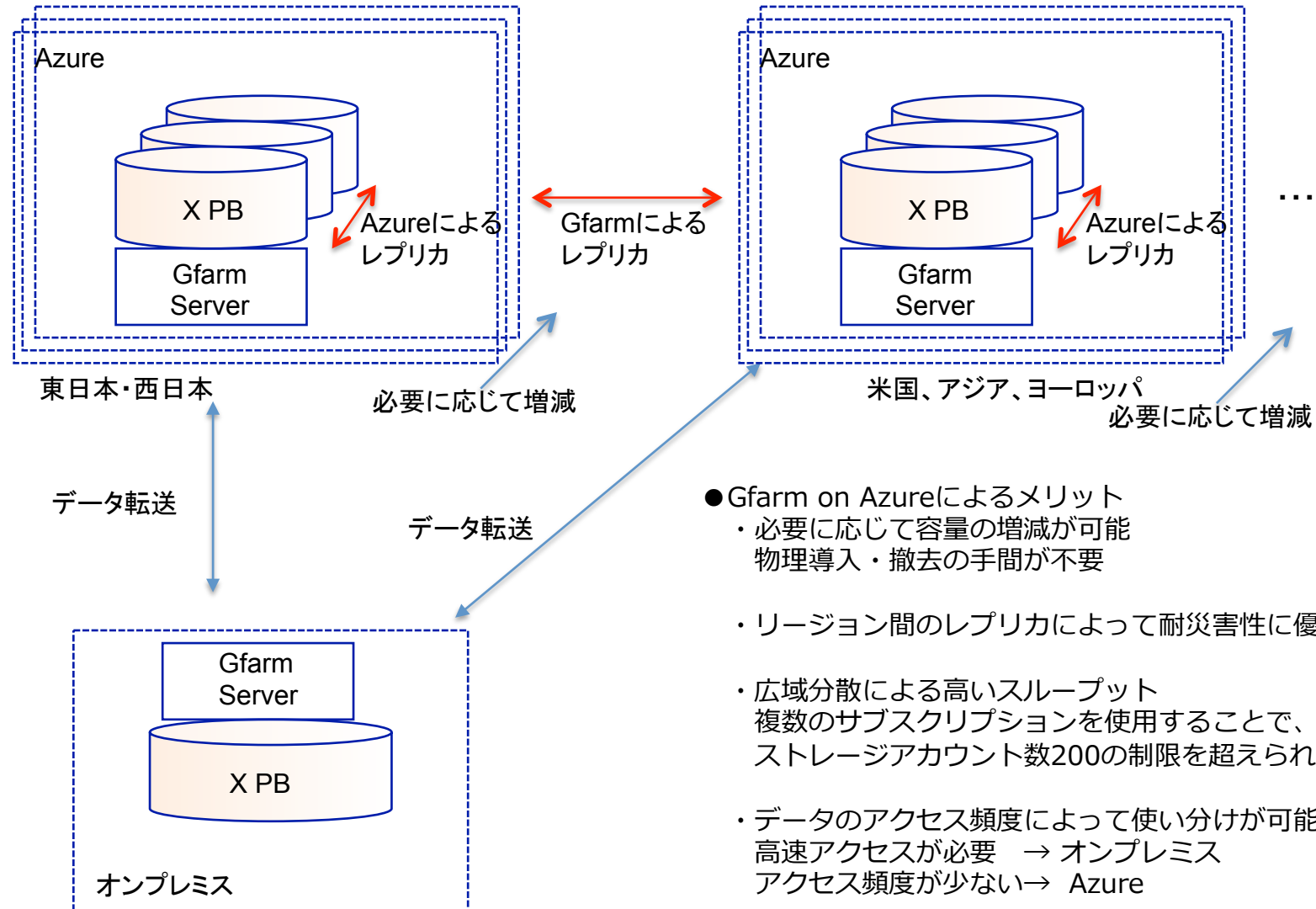
理論最大スループット
 $10 \text{ ノード} \times 20 \text{ データディスク} \times 200 \text{ MB/s}$
 $= 40 \text{ GB/s}$ は理論上可能
但し、Server~Storage間の帯域次第

- 最大容量
 $1 \text{ TB} \times 200 = 200 \text{ TB}$

※ BLOBを使用せず、IB非対応のローカルSSD搭載インスタンスを使用する選択肢もある。
ローカルSSDを使用した場合は、インスタンス停止と共にデータはクリアされるため、
スクラッチ領域向き

Gfarm利用イメージ2 広域ストレージ領域

Gfarm利用イメージ2：広域ストレージ領域



● Gfarm on Azureによるメリット

- ・ 必要に応じて容量の増減が可能
物理導入・撤去の手間が不要
- ・ リージョン間のレプリカによって耐災害性に優れる
- ・ 広域分散による高いスループット
複数のサブスクリプションを使用することで、
ストレージアカウント数200の制限を超えられる
- ・ データのアクセス頻度によって使い分けが可能
高速アクセスが必要 → オンプレミス
アクセス頻度が少ない → Azure
- ・ オンプレミスシステム入れ替え時のデータ移行の軽減
- ・ 1 オンプレミスとの連携以外に、公共データの公開にも向く

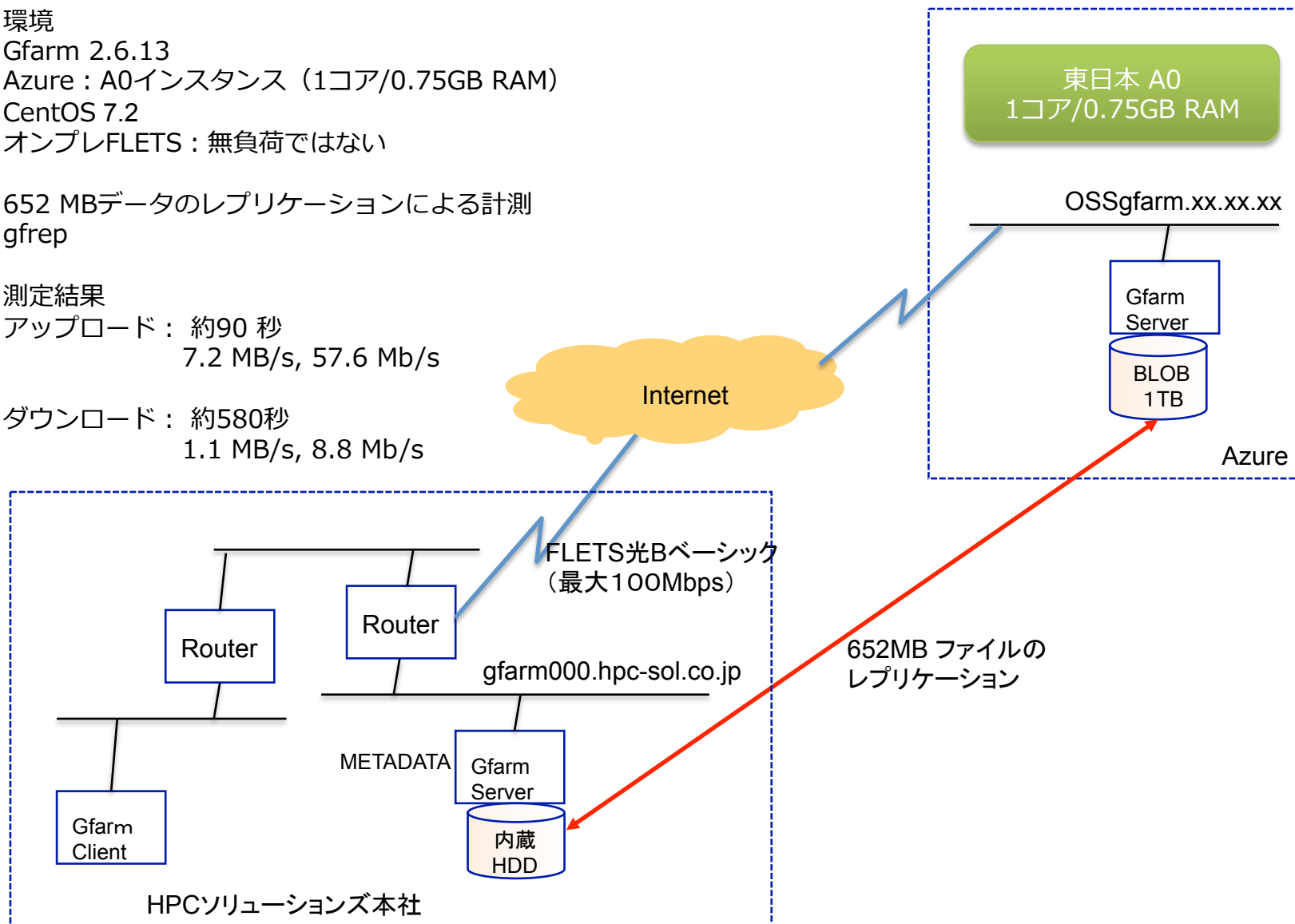
性能テスト

性能テスト 1

- 環境
Gfarm 2.6.13
Azure : A0インスタンス (1コア/0.75GB RAM)
CentOS 7.2
オンプレFLETS : 無負荷ではない

- 652 MBデータのレプリケーションによる計測
gfrep

- 測定結果
アップロード : 約90 秒
7.2 MB/s, 57.6 Mb/s
ダウンロード : 約580秒
1.1 MB/s, 8.8 Mb/s



※ 西日本、米国、ヨーロッパとの計測結果も概ね同じであった

性能テスト2

- 環境

Gfarm 2.6.13

Azure : 8 x A1インスタンス (1コア/1.75GB RAM) → (恐らく) 物理的には1ノード

8 x 1TB BLOB ストレージアカウント

CentOS 7.2

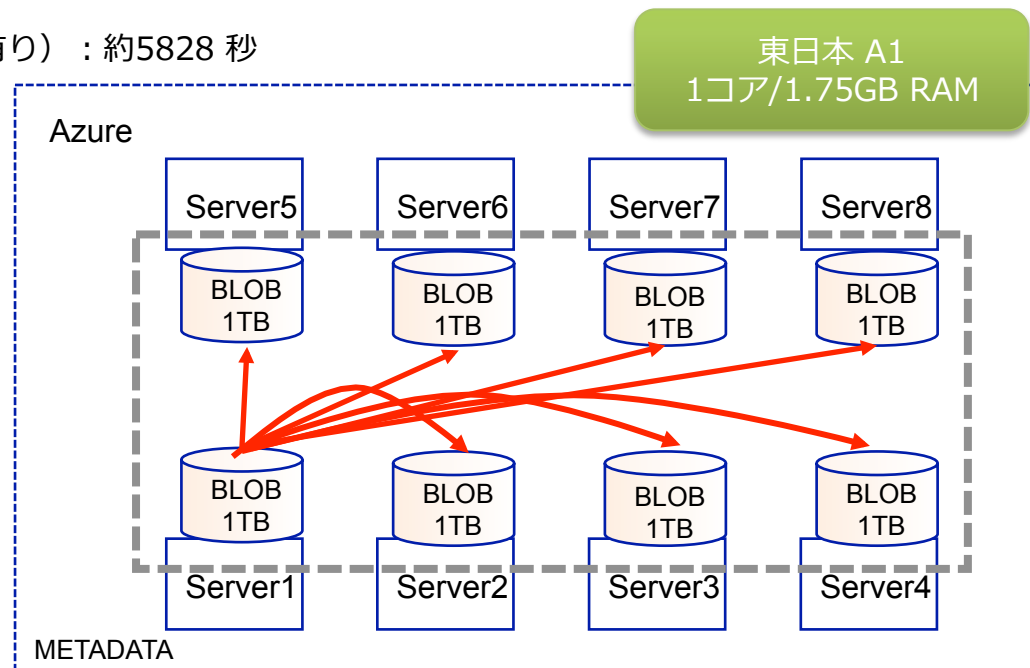
- 7.7GBデータのレプリケーションによる計測
gfrep

- 測定結果

[1] Server1 → Server2 : 約645 秒
11.9 MB/s

[2] Server1 → Server2-8 : 約5281 秒
10.2 MB/s

[3] Server1 → Server2-8(checksum有り) : 約5828 秒
9.2 MB/s



性能テスト3

- 環境

Gfarm 2.6.13

Azure : 8 x A1インスタンス (1コア/1.75GB RAM) → (恐らく) 物理的には1ノード

8 x 1TB BLOB ストレージアカウント

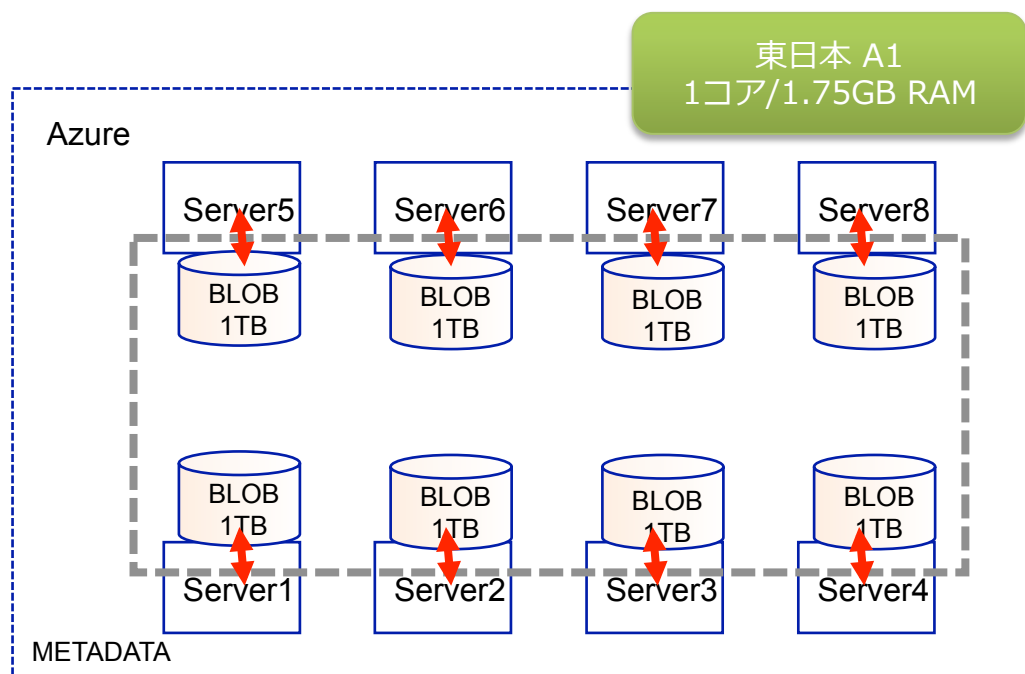
CentOS 7.2

- IOR 2.10.3

mpirun --allow-run-as-root -v -np プロセス数 -hostfile hostfile ./IOR -w -r -F -t 1mb -b 5gb -e -o out -i 4

- 測定結果

プロセス数	Write[MiB/s]	Read[MiB/s]
1	22.89	22.38
2	41.82	38.95
4	39.34	23.33
8	31.01	23.62



性能テスト4

- 環境

Gfarm 2.6.13

Azure : 4 x A11インスタンス (16コア/112GB RAM) → 物理的にも4ノード

4 x 1TB BLOB ストレージアカウント

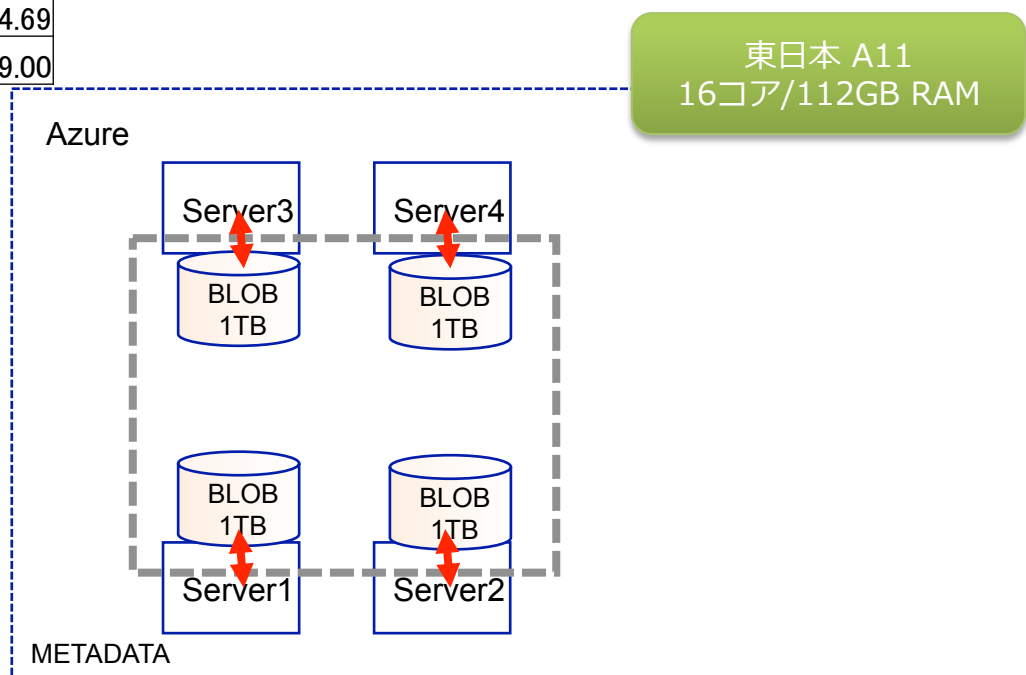
CentOS 7.2

- IOR 2.10.3

mpirun --allow-run-as-root -v -np プロセス数 -hostfile hostfile ./IOR -w -r -F -t 1mb -b 5gb -e -o out -i 4

- 測定結果

プロセス数	Write[MiB/s]	Read[MiB/s]
1	54.06	1,415.46
2	100.10	2,784.69
4	213.89	5,389.00



Azure概算費用

Azure概算費用（月額, 税別）

2016年10月時点

仮想サーバ

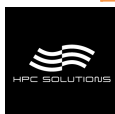
インスタンス	東日本	西日本	米国西部
A9 16コア, 112GB RAM, IB	¥183,269.52	-	¥147,981.60
D14 16コア, 112GB RAM, 800GB SSD	¥146,615.62	¥127,491.84	¥117,019.30

ダウンロード容量

容量	東日本	西日本	米国西部
1TB	¥14,343.44	←	¥9,042.61
10TB	¥144,067.86	←	¥90,825.39
100TB	¥1,386,999.06	←	¥803,160.75
500TB	¥6,452,727.06	←	¥2,996,568.75

ストレージ(HDD保存容量, LRS=ローカルレプリカ)

容量	東日本	西日本	米国西部
1TB	¥2,506.75	←	¥2,506.75
10TB	¥25,067.52	←	¥24,691.51
100TB	¥250,675.20	←	¥244,450.10
500TB	¥1,253,376.00	←	¥1,213,727.54



まとめ

1. Gfarm on Azure

オンプレ + Azureおよび、Azure内で容易に構築できることが確認できた。

Gfarmの基本機能も問題無く動作確認できた。

2. HPCクラスタシステムの並列ファイルシステムとしてのGfarm

データディスクの刻みが1TB, HDD:60MB/s、SSD:200MB/sであるため、これを束ねた時のピーク性能の計測は今後の課題

3. 広域ストレージ領域としてのGfarm

本来Gfarmの特徴である広域分散共有ファイルシステムとして用いることで以下が期待される。

- ・大容量・高スループットストレージの提供
- ・データの保管、移動のに伴う運用上の負担軽減

今後、実運用を想定した実験が必要

4. 性能

今回は計測が主体

今後、I/Oスループットの設計ができるレベルの計測結果と構成情報の整理が必要

5. コスト

オンプレとの単純比較ではなく、実運用を想定したランニングコスト、管理コストも含めた比較検討が必要