HPC on Azure のご紹介

日本マイクロソフト株式会社
Azure ソリューション技術部
佐々木邦暢 (@ksasakims)

本日の内容

- Microsoft HPC Pack
- Azure の計算リソース
- 事例

Microsoft HPC Pack

Windows HPC Server (HPC Pack) とその歴史

ユーザーフレンドリーであるのみならず大規模クラスタにも対応

HPC Pack (旧称: Compute Cluster Pack)

- オンプレミス・クラウドを統合管理できるジョブスケジューラー
- MPICH2 ベースの MPI ライブラリ (MS-MPI)
- 使いやすい GUI 管理ツール
- 効率的なコマンドライン管理ツール
- Excel 高速化機能
- 最新版は HPC Pack 2012 R2 Update 1 (2014 年 11 月リリース)

2006年 Compute Cluster Pack (HPC v1)

- 三菱UFJ証券様のクラスタがTop 500にランクイン。(1760コア、6.52TFlops) http://www.top500.org/system/174885

2008年 HPC Pack 2008 (HPC v2)

- 上海スーパーコンピューティングセンターのDawning 5000AがTop500で11位にランクイン。(30,720コア、180.6TFlops) http://www.top500.org/system/176118

2010年 HPC Pack 2008 R2 (HPC v3)

- 東工大のTSUBAME 2.0で初のペタフロップス越え。1.13PFlops. Top500の5位相当の記録。

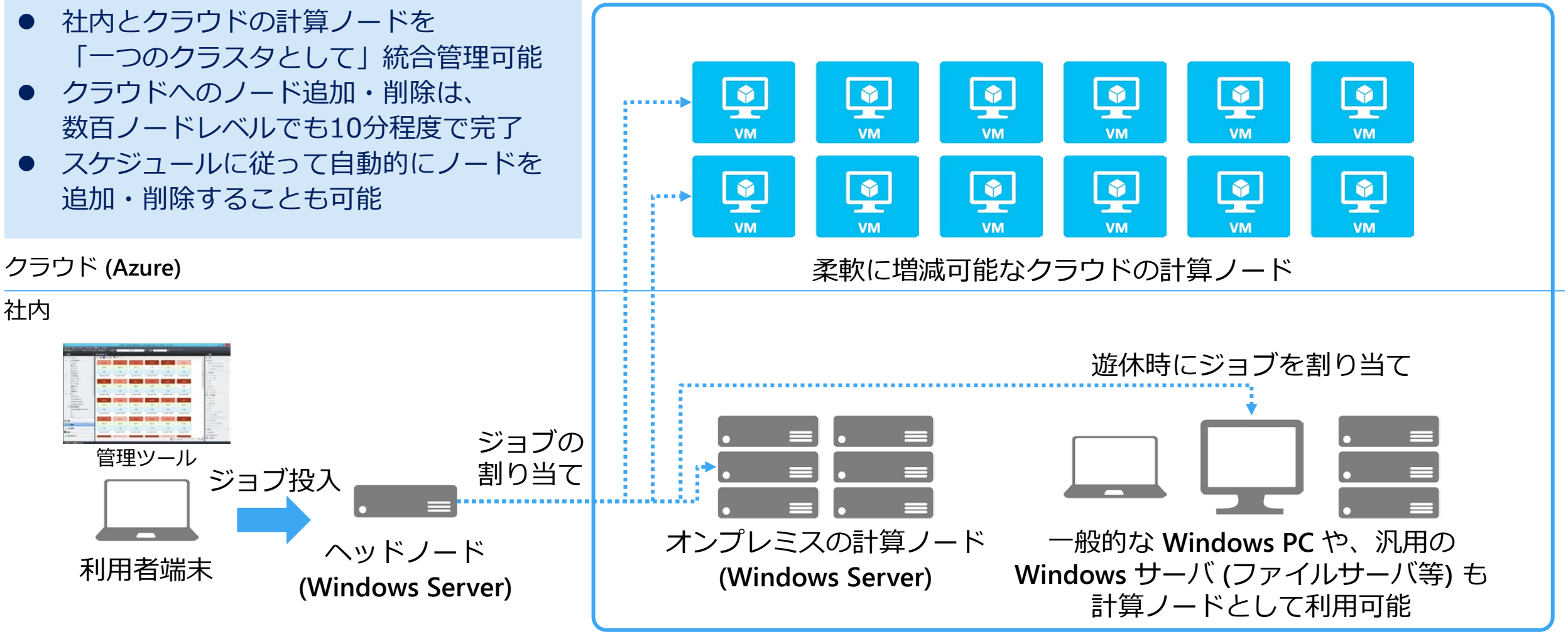
2012年 HPC Pack 2012 (HPC v4)

- HPC用AzureインスタンスでTop500にランクイン (8064コア, 151.3 TFlops) http://www.top500.org/system/177982

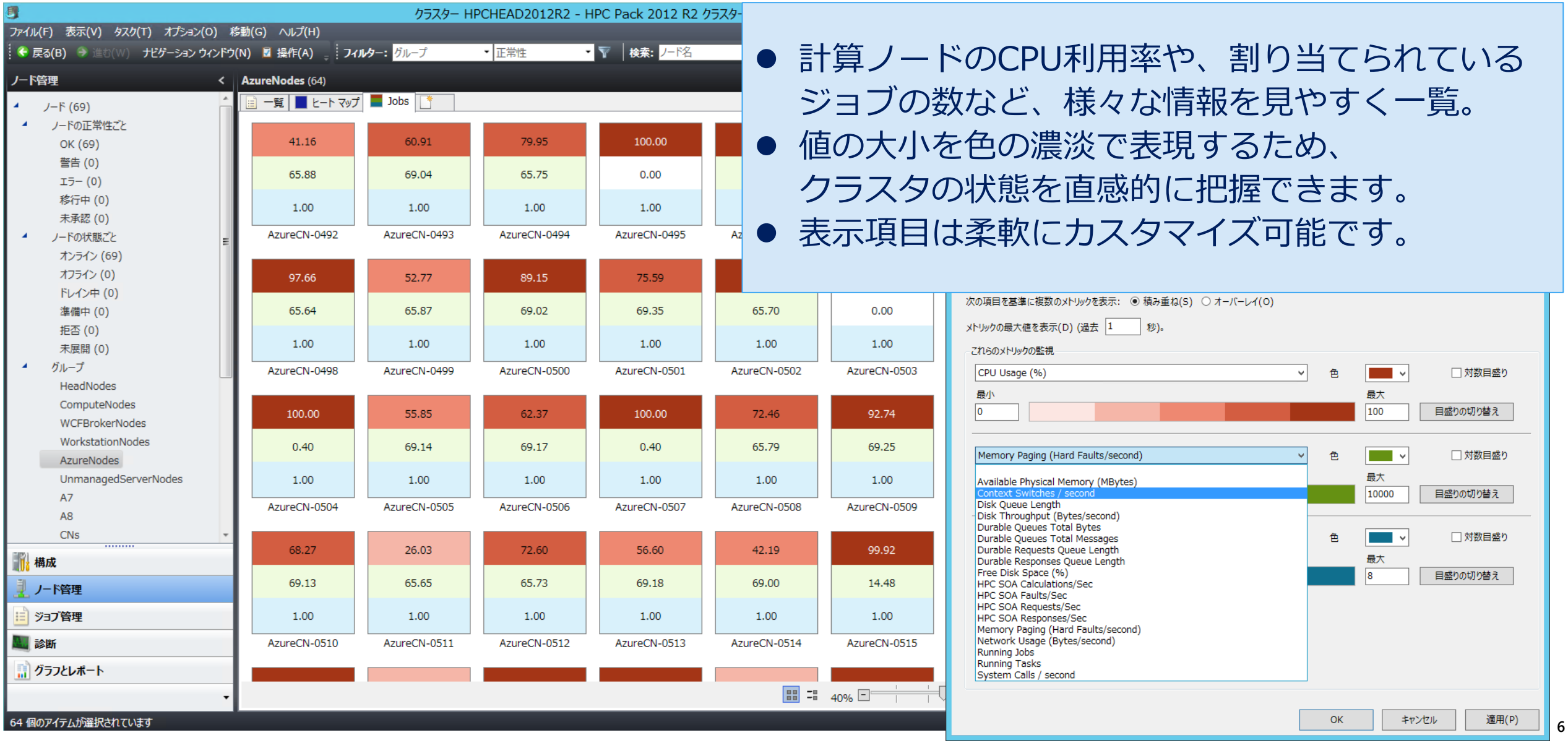
オンプレミス + クラウドの統合クラスタ

PC, サーバー, クラウド. 様々なコンピューターを計算ノードに

- 社内とクラウドの計算ノードを「一つのクラスタとして」統合管理可能
- クラウドへのノード追加・削除は、数百ノードレベルでも10分程度で完了
- スケジュールに従って自動的にノードを追加・削除することも可能

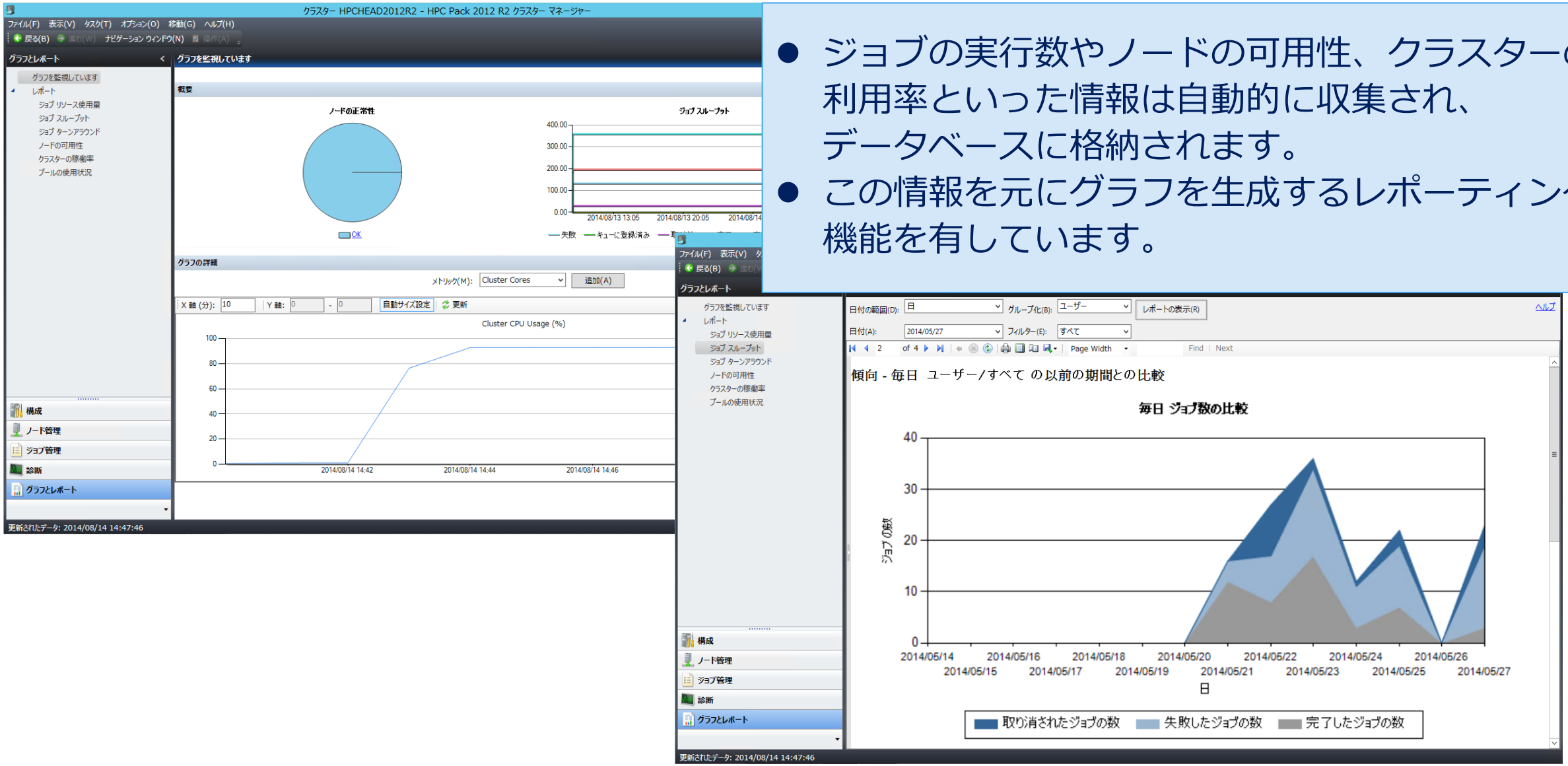


ヒートマップでクラスタの状態を可視化



- 計算ノードのCPU利用率や、割り当てられているジョブの数など、様々な情報を見やすく一覧。
- 値の大小を色の濃淡で表現するため、クラスタの状態を直感的に把握できます。
- 表示項目は柔軟にカスタマイズ可能です。

レポーティング機能

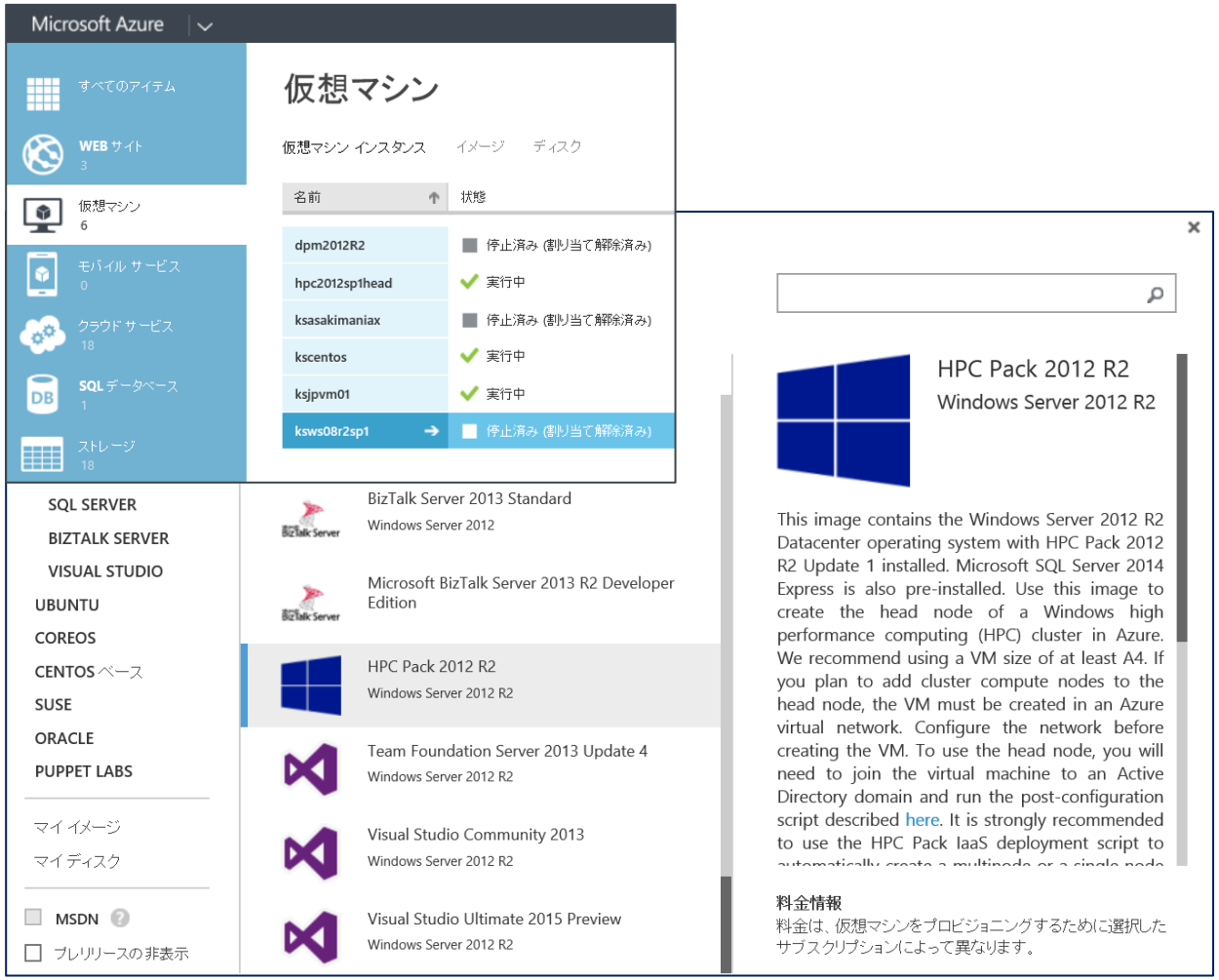


- ジョブの実行数やノードの可用性、クラスターの利用率といった情報は自動的に収集され、データベースに格納されます。
- この情報を元にグラフを生成するレポーティング機能を有しています。

Azure の計算リソース

Microsoft Azure『仮想マシン』サービス

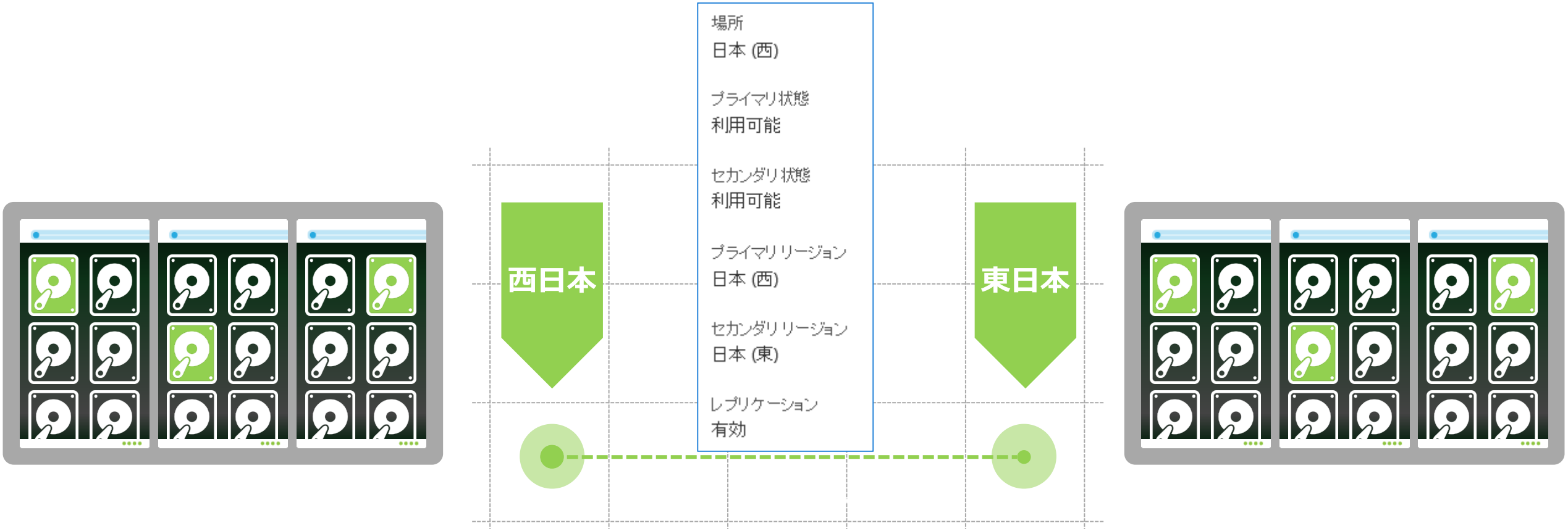
仮想マシンを素早く構築できる「サービスとしてのインフラ」(IaaS)



- 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。

リージョン間複製機能を持つ堅牢なストレージ

リージョン内3多重 + リージョン間複製 = 6多重複製



Azure は日本国内に、400㎞ 以上離れた2リージョンを有します。
データを国外へ持ち出すことなく、DR 構成が可能です。

仮想マシンのサイズとスペック

サイズ	コア数	メモリ容量 (GB)	作業用ディスク容量 (GB)	データディスク数	最大 IOPS
A0 (XS)	共有	0.768	20	1	1 x 500
A1 (S)	1	1.75	70	2	2 x 500
A2 (M)	2	3.5	135	4	4 x 500
A3 (L)	4	7	285	8	8 x 500
A4 (XL)	8	14	605	16	16 x 500
A5 (メモリ集中型)	2	14	135	4	4 x 500
A6 (メモリ集中型)	4	28	285	8	8 x 500
A7 (メモリ集中型)	8	56	605	16	16 x 500

http://azure.microsoft.com/ja-jp/pricing/details/virtual-machines/

SSD 搭載インスタンス: Dシリーズ

サイズ	コア数	メモリ容量 (GB)	作業用SSD 容量 (GB)	作業用SSD 最大IOPS (8KBブロック)	作業用SSD R/W性能 (MB/s)	データ ディスク数 最大接続数	データ ディスク 最大 IOPS
D1	1	3.5	50	3,000	48/24	1	2 x 500
D2	2	7	100	6,000	96/48	2	4 x 500
D3	4	14	250	12,000	192/96	4	8 x 500
D4	8	28	500	24,000	384/192	8	16 x 500
D11	2	14	100	6,000	96/48	2	4 x 500
D12	4	28	200	12,000	192/96	4	8 x 500
D13	8	56	400	24,000	384/192	8	16 x 500
D14	16	112	800	48,000	768/384	16	16 x 500

- Dシリーズは SSD を搭載した新ハードウェアで稼働。作業用ディスクが大幅に高速になっています。
- CPU 性能は同じコア数のAシリーズ比で 60% 程度向上しています

http://azure.microsoft.com/ja-jp/pricing/details/virtual-machines/

さらなる大型インスタンス: Gシリーズ

サイズ	コア数	メモリ容量 (GB)	作業用ディスク(SSD) 容量 (GB)
G1	2	28	406
G2	4	56	812
G3	8	112	1,630
G4	16	224	3,250
G5	32	448	6,500

- スケールアップが必要な用途に最適の大型VM
- Intel Xeon E5-2600 v3 を最大 32 コア搭載
- ローカルディスクはすべて SSD

2015年1月9日
正式リリース済み
(まずは West US リージョンのみ)

http://azure.microsoft.com/en-us/updates/general-availability-g-series-for-virtual-machines/

HPCインスタンス (A8,A9)

高速 CPU, 大容量メモリ, 高速インターコネクト

サイズ	コア数	メモリ容量	プロセッサ	ネットワーク 1	ネットワーク 2
A8	8	56 GB	Xeon E5-2670 2.6 GHz	10 Gbps イーサネット	QDR InfiniBand (w/ RDMA)
A9	16	112 GB			

TOP500 にランクインしました (2012年11月)

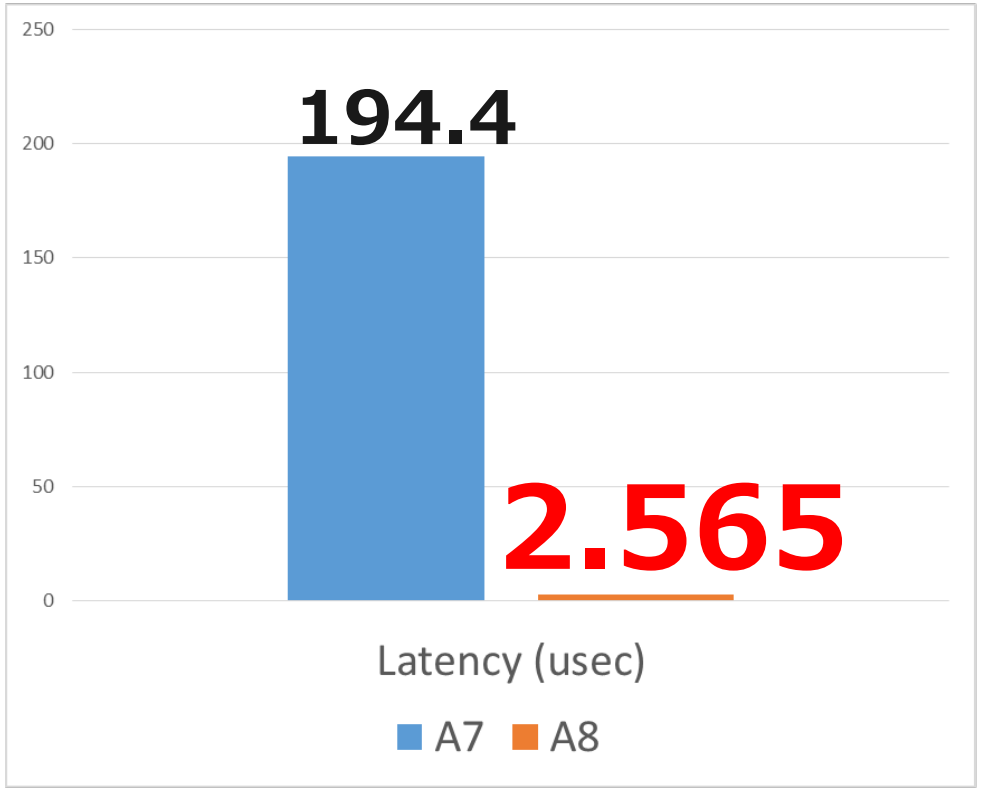
- 151.3 TFLOPS (効率 **90.2**%) で 165 位
- 504 ノード, 8064 コアで実施
- http://www.top500.org/site/50454

日本リージョンでも利用可能に

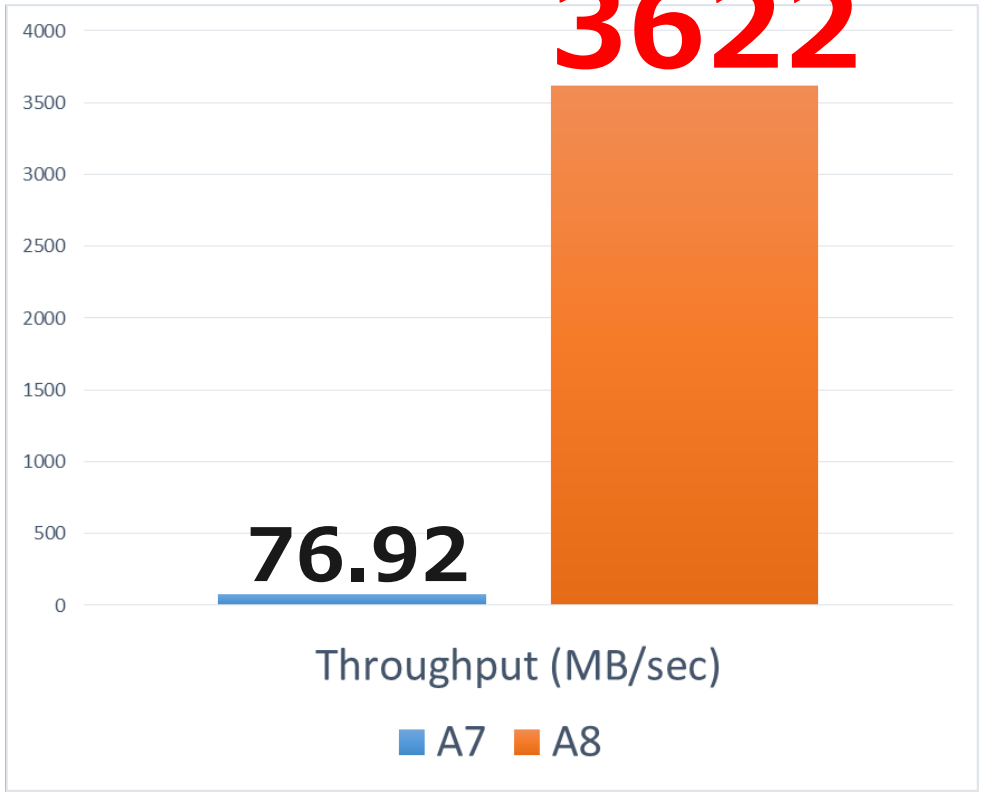
- **5 月から、日本(東)でも利用可能になりました。**
- **MPI を利用するアプリケーションでは、同じコア数の A7 と比べて 5 倍の性能を発揮したケースもあります。**

通常のインスタンス(A7)との比較

レイテンシの比較
パケットサイズ: 4バイト



スループットの比較
パケットサイズ: 4MB



検証事例: Particleworks on Azure

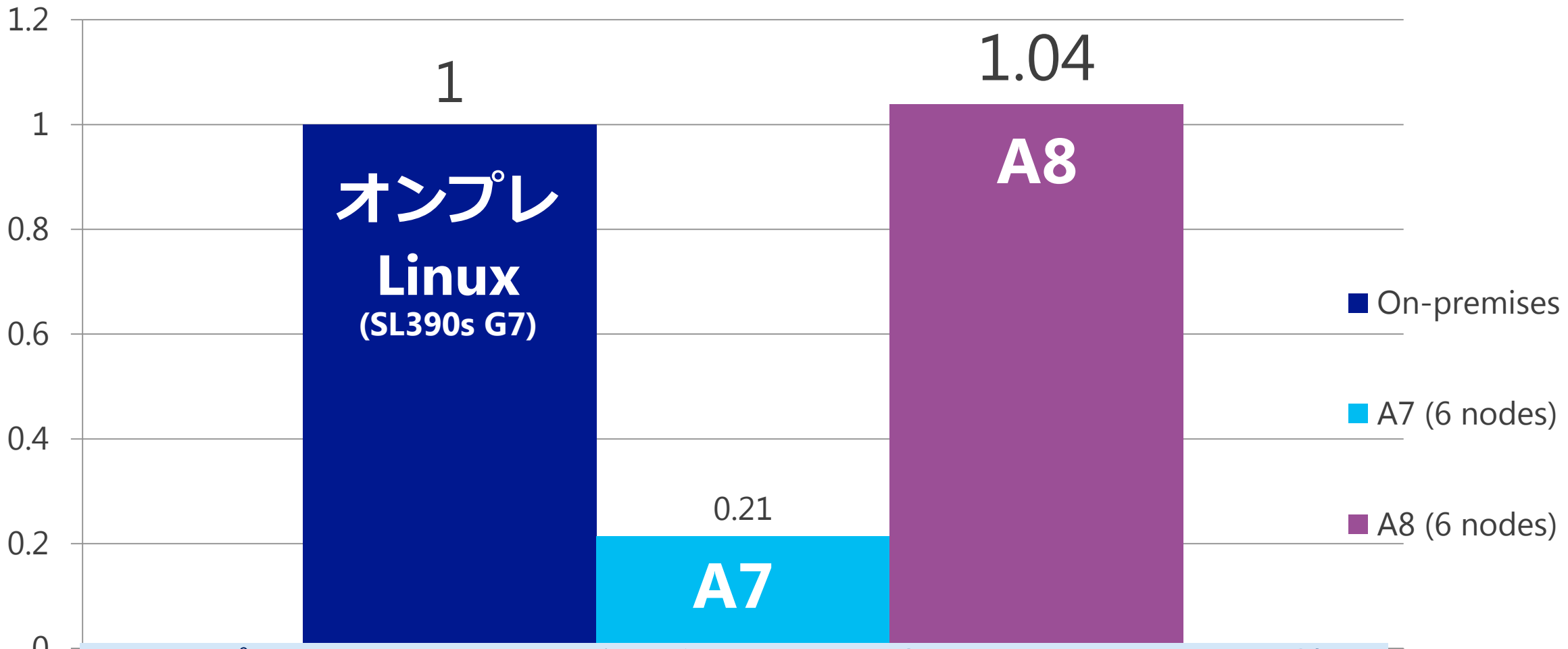
実施内容

- 4000万粒子規模の解析
- 内容: 自動車の水はね

オンプレミスのLinuxマシンとの比較を実施

- オンプレミスの **Linux** クラスタと、**Azure**の**A7,A8,A9**インスタンスで同じ解析を実施し、実行時間を比較。
  - 機種: **ProLiant SL 390s G7 x 4 ノード (計48コア)**
  - **CPU：Intel Xeon X5675 3.06GHz 6 cores ×2**
  - **RAM：4GB×12 = 48 GB**
  - **QDR InfiniBand 40Gbps×2**

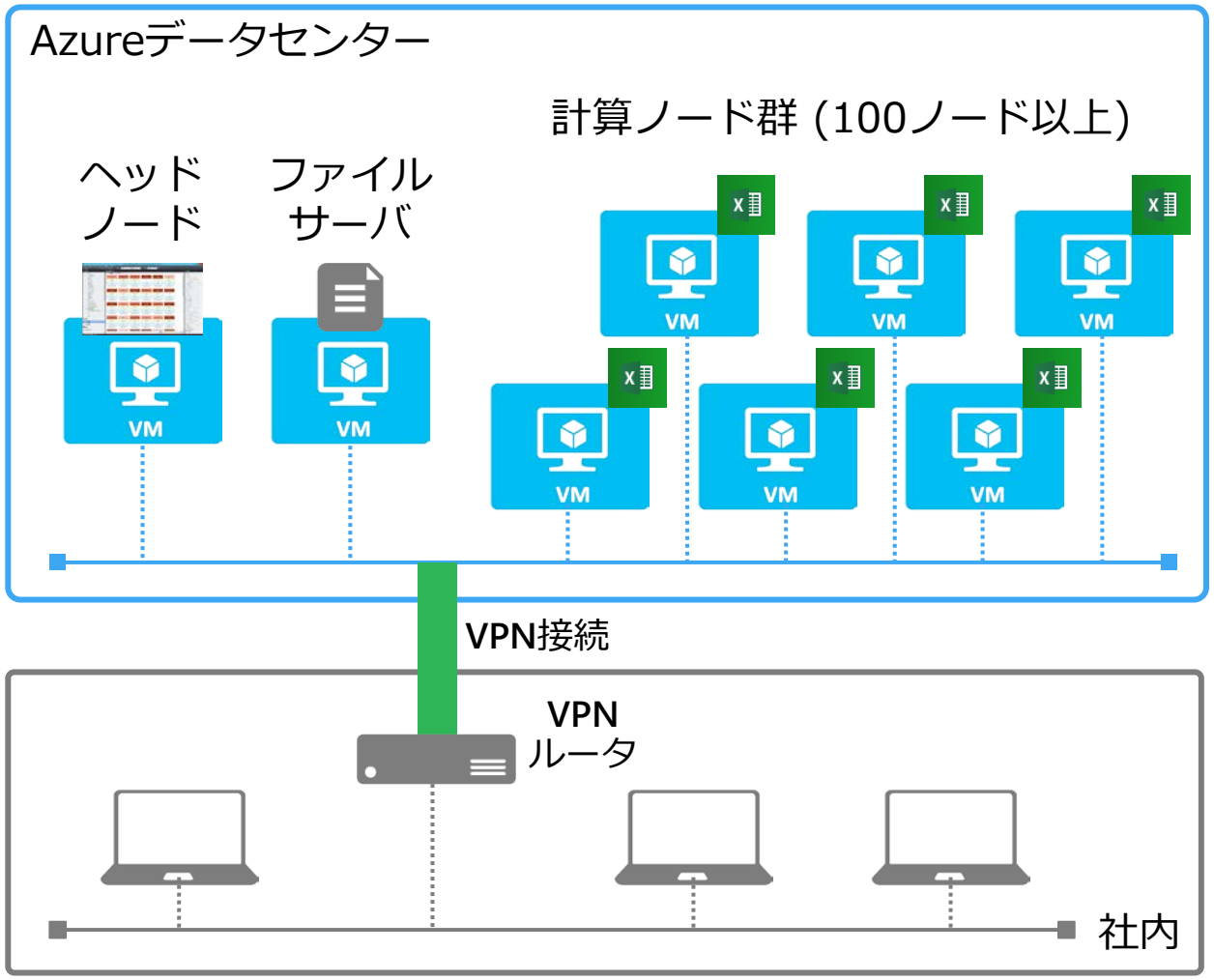
オンプレミスの物理マシンと遜色ない性能



オンプレミスの InfiniBand 付き物理マシンの性能を 1 とした場合の比較
A8/A9 は物理マシンと同等、A7 との比較では 5 倍の性能を記録
(流体計算アプリケーションでの性能検証結果)

事例: 国内生命保険会社様

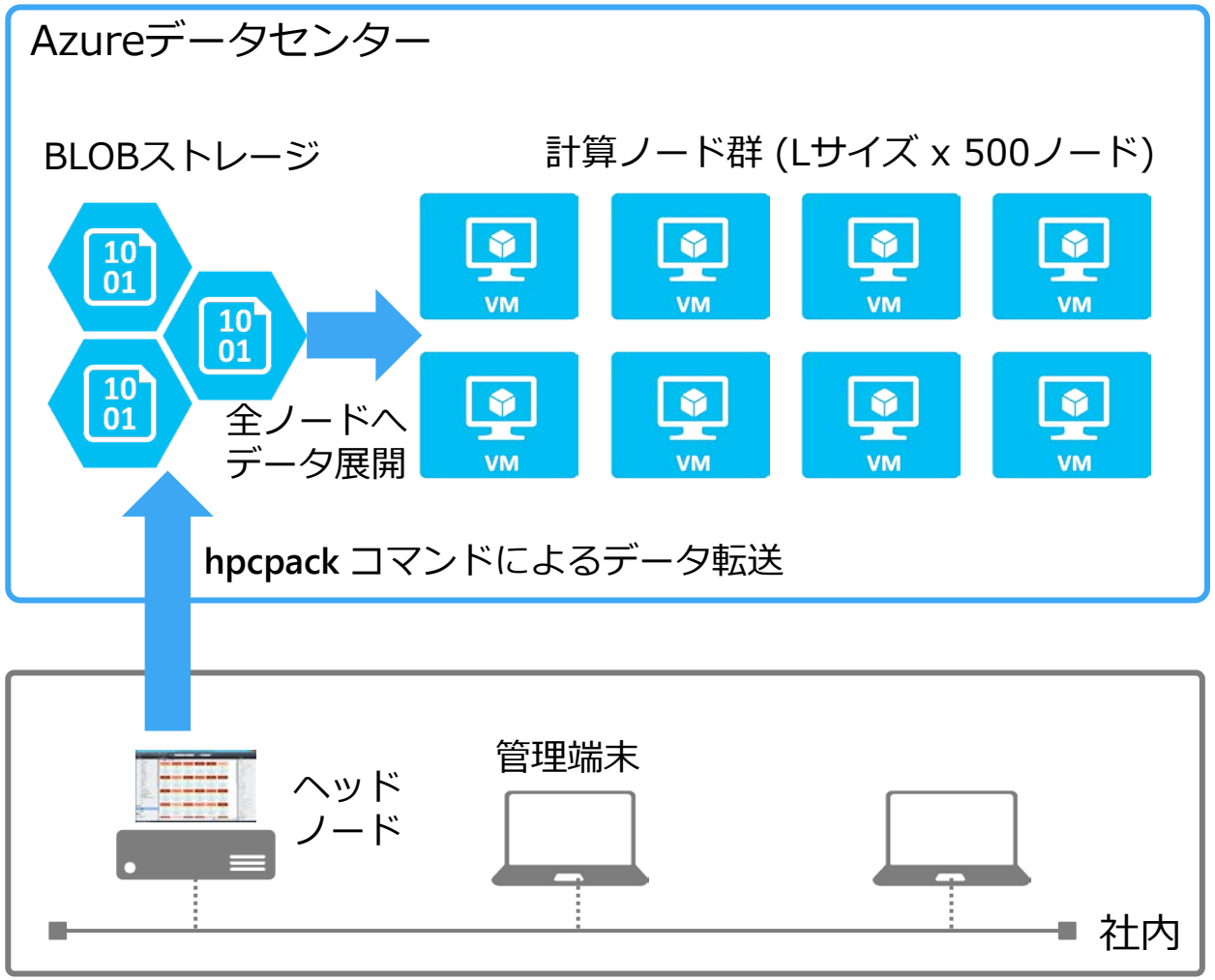
Excel 並列実行による保険数理計算を Azure 上で



- 期末の繁忙期に必要となる大量の計算処理をクラウドにオフロードした事例
- ヘッドノードと計算ノードをすべて Azure 上に配置した完全クラウド構成
- 通常時はすべてのノードをシャットダウンしておくことで課金を抑制
- 必要になった時点でヘッドノードを起動し、計算ノードを展開。100ノード以上でも15分程度で利用可能に
- 社内システムとは Azure 仮想ネットワークの VPN 機能で接続し、入力データのクラウドへのアップロード、計算結果ファイルのクラウドからのダウンロードはこの VPN 経由で実施

事例: 海外証券会社様

2000コアのクラスタを夜間バッチの時間帯に実行



- 日中のオンライン処理で約定した大量の取引データを夜間バッチ処理で処理
- オンライン時間終了後、夜間バッチの時間帯のみAzure上に計算ノードを展開。
  - Lサイズ(4コア) x 500ノードの合計 2000 コア
- 計算元データは hpcpack コマンドを利用して HTTPS 接続で Azure 上の BLOB ストレージへ。
- 全ノードへのデータ展開はAzureのデータセンタ内で行うことで、オンプレ → クラウド間の転送量を最小化


Microsoft