ORACLE

[会員企業技術紹介]

Oracle Cloud Infrastructure (OCI)とHPC

PCクラスタワークショップin 神戸2022「クラウドとHPC」

2022年6月23日
日本オラクル株式会社
クラウド事業統括　公共営業本部　松山 慎

Safe harbor statement

The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

Oracle Cloud Infrastructure (OCI) 概要

Oracle Cloud Infrastructure(OCI)のリージョン

2022年1月現在：37リージョン提供中、さらに7リージョン計画

2012年 Oracle Cloud サービス開始
2018年 アーキテクチャを**Gen2**に全面刷新

デュアル・リージョン：基本的に全ての国/地域で2つ以上のリージョンを提供し、お客様の業務継続要件に対応していく
(日本の場合は東京-大阪)
各リージョンはOracle Backboneで接続

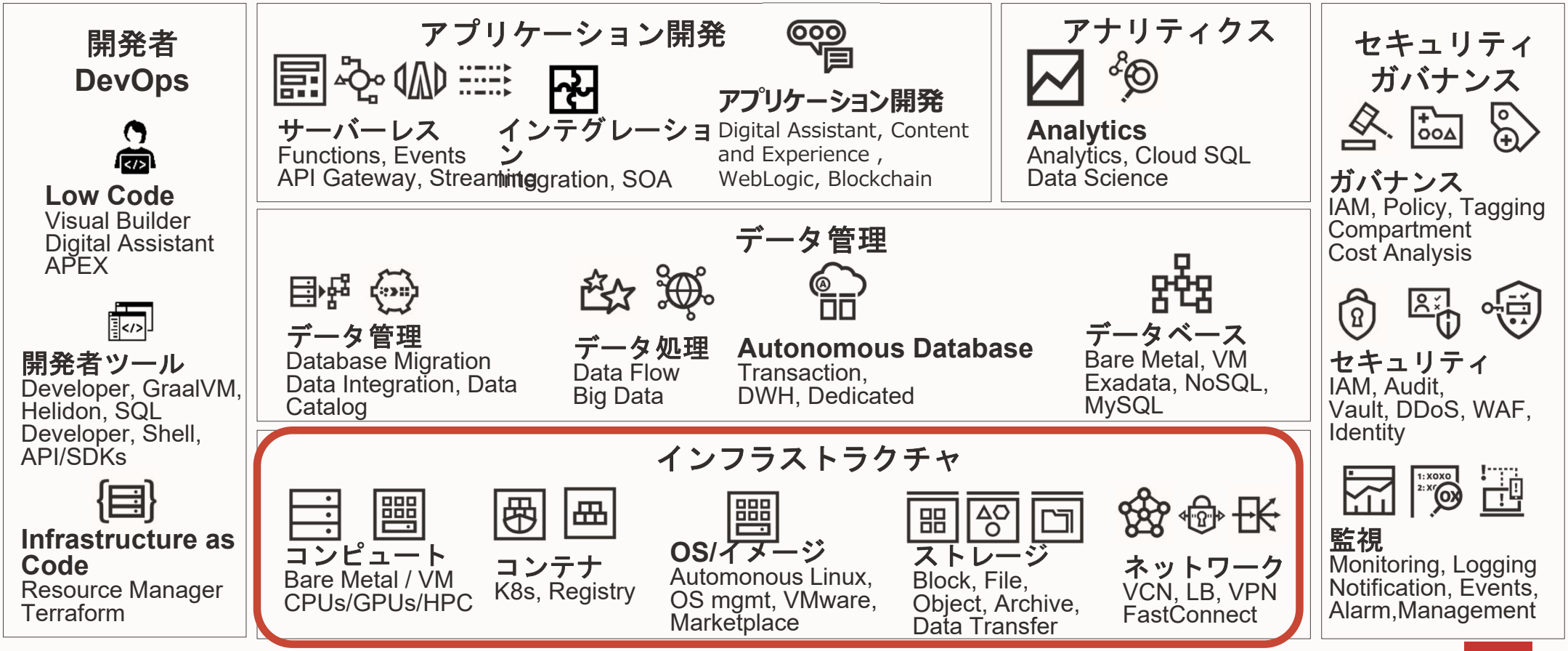
サステナビリティ：2025年までに、すべてのリージョンにおいて、100%再生可能エネルギーを使用することを表明 (欧州リージョンは達成済み)

https://www.oracle.com/cloud/architecture-and-regions.html

Oracle Cloud Infrastructure の日本国内展開

リージョン	大阪（Japan Central）	東京（Japan East）
開業年	2020年2月	2019年5月
SINETクラウド接続	2020年7月	2019年10月
ISMAP	登録済み	
準拠法	日本国の実体法と手続法を適用	
裁判管轄	東京地方裁判所を第一審の専属的合意管轄裁判所とする	
決済方式	現地通貨(日本円) および請求書ベースの支払い	
単価	ワールドワイドで同一の単価 (月10TB超のインターネットアウトバウンドデータ量を除く)	

Oracle Cloud Infrastructure (OCI)のサービス構成

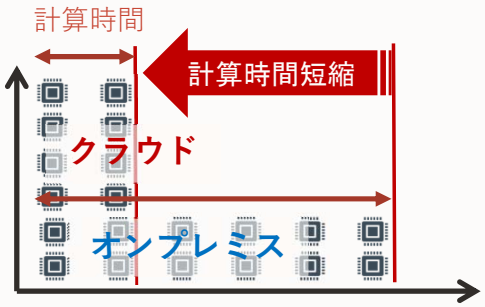


Compute Cloud Service for HPC & AI

HPCクラウド利用の想定されるメリットの例

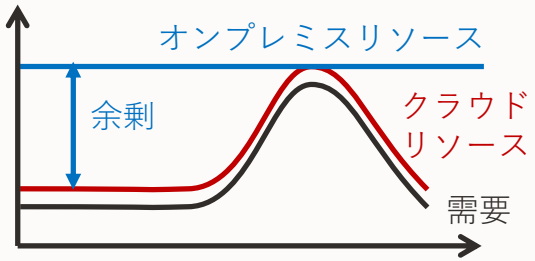
計算時間短縮

必要な計算リソースを迅速に割当て、計算時間を短縮



コスト削減

需要に合せリソースの追加/削除することで、コストを削減
CAPEXからOPEXへの転換



将来技術の利用

最新のCPUやGPUをすぐ利用可能、１秒単位の課金で評価等の短時間利用も低コスト

迅速なリソース導入・追加

設置場所や電源の確保/導入/設置に掛る時間を大幅に短縮

障害から素早い復旧

障害発生リソースを解放し同様構成の再作成のみ、データは常に複数のレプリカを保持

最新のセキュリティ対策 様々な連携サービスの提供

最新のセキュリティ対策の自動適用もしくは実装支援を提供

可視化や各種サーバレスサービスなど様々なサービスと連携可能

HPC用途での黎明期のクラウドサービスに帯する評価

～2015年までのクラウドサービス

オーバーサプスクリプションのある仮想サーバとネットワークのため、ストレージやネットワークのI/Oが不安定で、性能再現性がない。

HPCで実用的なIOPS値を備えるブロックストレージが高価。

アウトバウンド通信データ量に課金され大量のデータを扱うHPC分野では通信費用が高額となる。

日本国内データセンタのサービスが高く設定される。ドル単価で為替影響あり。

HPC用途での黎明期のクラウドサービスに帯する評価

～2015年までのクラウドサービス

- オーバーサプスクリプションのある仮想サーバとネットワークのため、ストレージやネットワークのI/Oが不安定で、性能再現性がない。
- HPCで実用的なIOPS値を備えるブロックストレージが高価。
- アウトバウンド通信データ量に課金され大量のデータを扱うHPC分野では通信費用が高額となる。
- 日本国内データセンタのサービスが高く設定される。ドル単価で為替影響あり。

Oracle Cloud Infrastructure (2016～)

- ✓ 安定して高速な計算資源
- ✓ 高性能で堅牢なストレージ
- ✓ 低コスト

Oracle Cloud Infrastructure (OCI) のHPC Capability

ベアメタル

フラット&ノンブロッキングネットワーク / オーバーサブスクリプションなし

HPC分野における OCI の強み

✓ 安定して高速な
計算資源

ベアメタル ＋
帯域/遅延が均一なRDMA
(オーバーサブスクリプションを排除し、高い性能再現性を提供)

TCP/IPネットワーク、仮想マシンも
オーバーサブスクリプション無し
(リソースアロケーションによらず性能均一)

✓ 高性能で堅牢な
ストレージ

IO性能の高い
各種ストレージサービス
(標準Boot Volume
3kIOPS:24MB/s～
25kIOPS:480MB/s)

高いデータ年間耐久性
99.999999999%(11-9)
(Object Storage、File Storage Service)

セキュアなストレージサービス
保存データ常時暗号化

✓ 低コスト

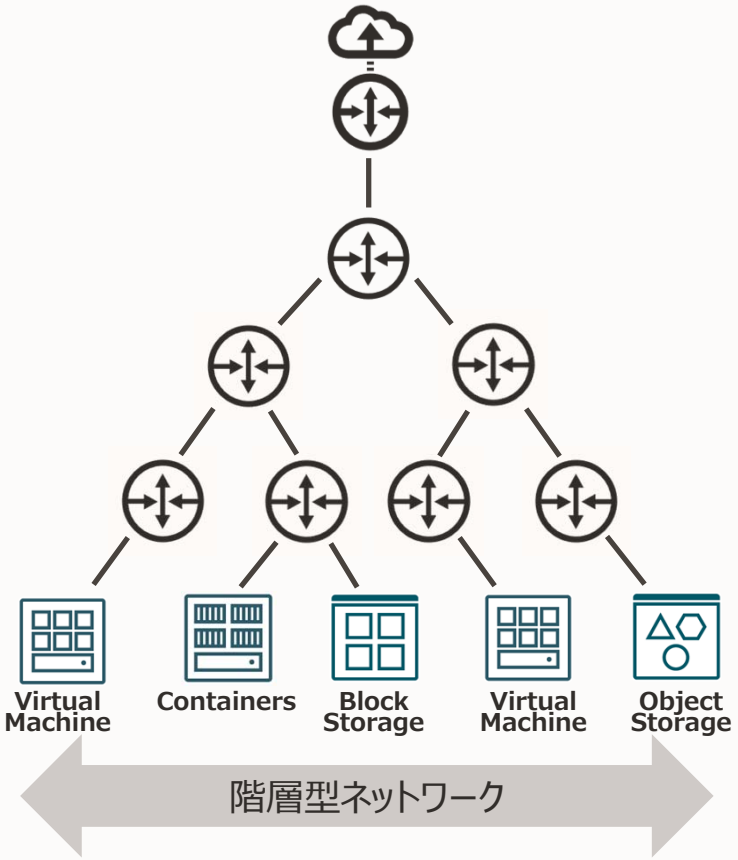
SINETクラウド接続サービス経由
の転送データ量が上り下りとも
無料

計算資源単価他社比較
23～55%程度安価

ストレージ単価他社比較
最大97%安価

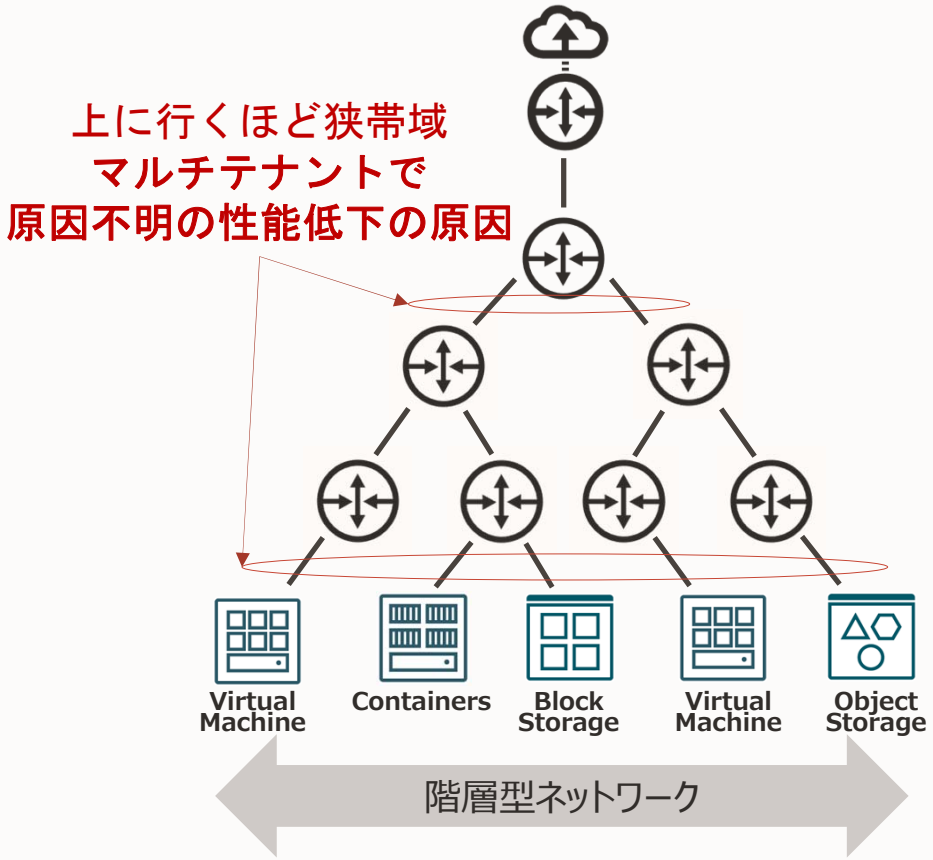
2015年までのクラウドサービスで状況から想定されるネットワーク構成

黎明期のクラウドの想定ネットワークトポロジ



2015年までのクラウドサービスで状況から想定される ネットワーク構成

黎明期のクラウドの想定ネットワークトポロジ



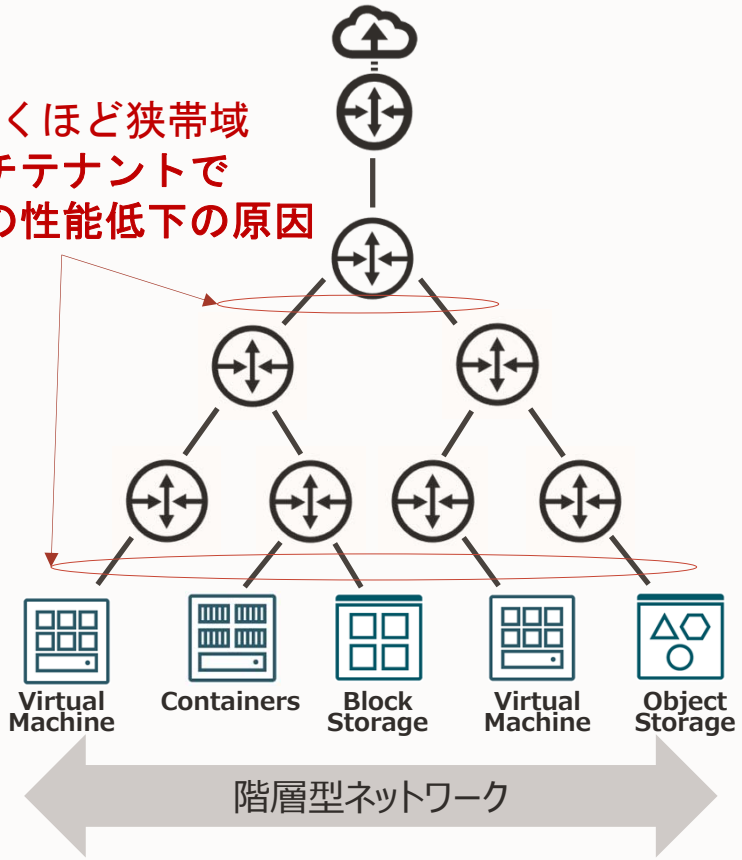
- 階層型ネットワーク基盤をマルチテナントで利用すると、他のテナントの通信状況により帯域が狭まる状況が発生し、システムとして性能/実行時間が安定しない状況(ノイジーネイバー)が多々発生する。

Oracle Cloud Infrastructure (OCI)

高速かつ安全で高コストパフォーマンスの第二世代クラウド (ベアメタル+フラットネットワーク)

黎明期のクラウドの想定ネットワークトポロジ

上に行くほど狭帯域
マルチテナントで
原因不明の性能低下の原因

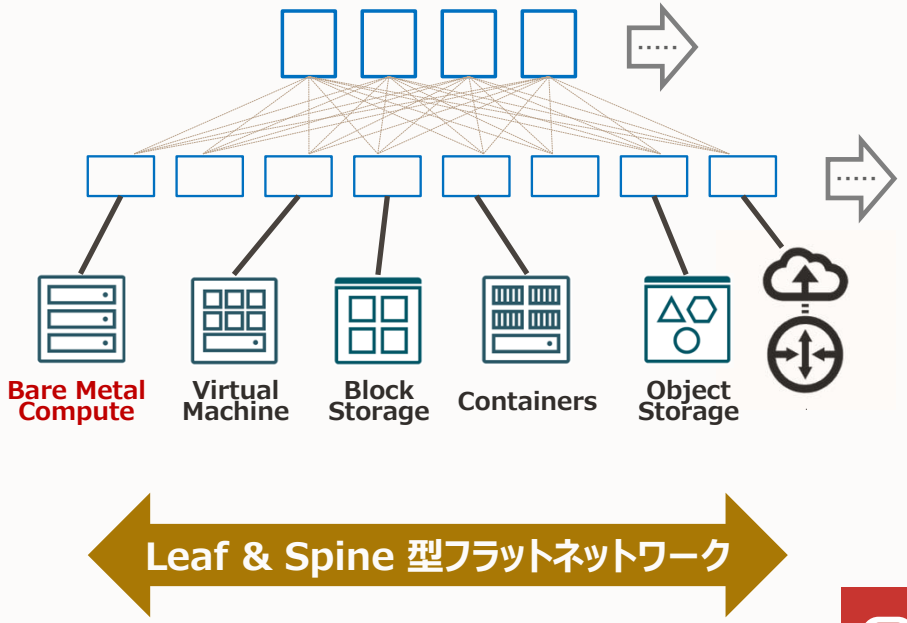


VS

ORACLE
Cloud Infrastructure
(Gen.2 Cloud)

2ホップ以内で関連リソースにアクセス

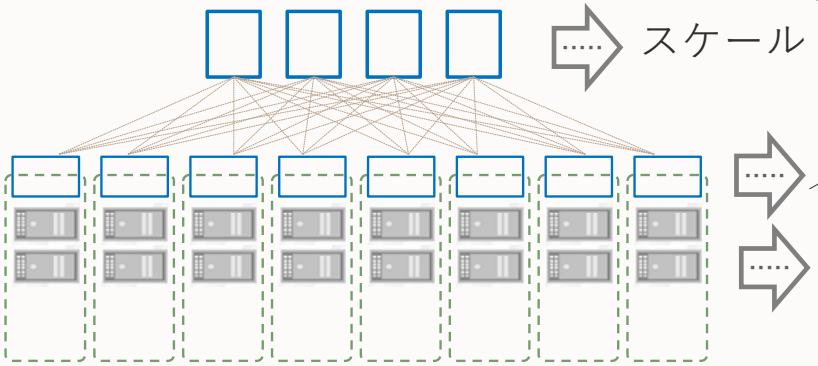
Performance SLA定義
・ネットワーク帯域
・ブロックボリュームIOPS
- <90% 99.9%/月



Oracle Cloud Infrastructure
安定して高速・低遅延のネットワーク基盤

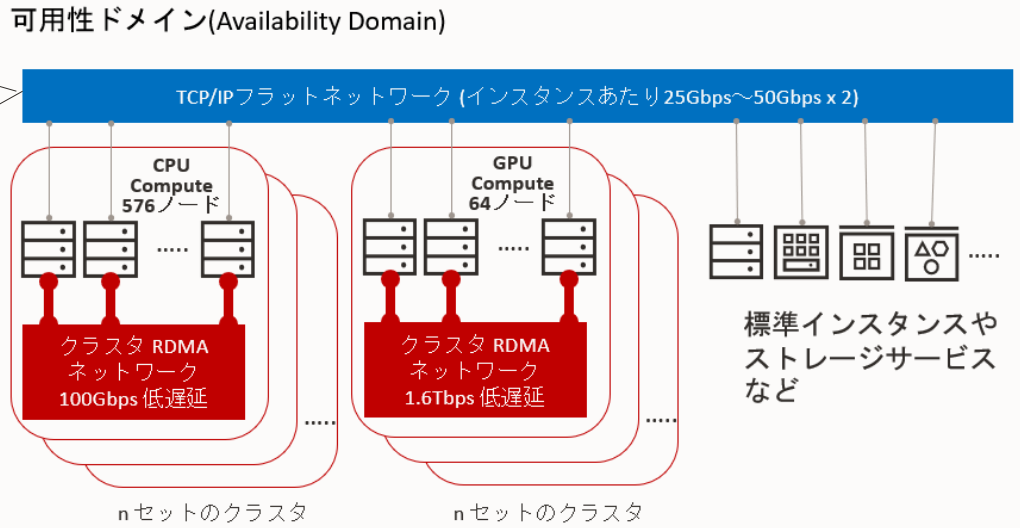
TCP/IPネットワーク基盤

- ✓ 各サービスをフラットなSpine-Leafネットワークで接続
  あらゆるサービスに低遅延・広帯域でアクセス
- ✓ **リソースアロケーションの制約無く均一な性能を提供**
  （コンピュート、ストレージやデータベースなど各種サービス）



HPC向けRDMAネットワーク

- ✓ HPCクラスタ向け、**広帯域・低遅延**ネットワーク
- ✓ RoCEv2 (RDMA over Converged Ethernet) による
  安全なマルチテナント環境を提供
- ✓ **帯域と遅延が均一なRDMA網**で接続されたクラス
  タ(アイランド)をリージョン内に複数配備



Bare Metal RDMAクラスタ 構成概要
マルチテナントで安全に利用可能な複数セットの性能再現性のあるHPC環境

CPUクラスタ

CPUクラスタインスタンス (BM.Optimized3.36) 1ノード	
CPU	Intel Xeon Ice lake 3.0-3.6GHz 18core x 2CPU 理論ピーク3.456TFLOPS/ノード
Memory	512GB (DDR4-3200 x 8チャネル/ソケット)
Storage	3.2TB NVMe SSD + Block Storage (up to 1TB)
Network	100Gbps RDMA (RoCEv2、CX-5) x 1 50Gbps TCP/IP x 1

Cluster Network 想定最大構成	
Latency	1.5マイクロ秒程度
Max nodes	576
Max total cores	**20,736**
Max total memory	288TiB
Max total NVMe storage	1.8PiB
理論ピーク性能/クラスタ	1.99 PFLOPS

GPU搭載クラスタ

GPUクラスタインスタンス (BM.GPU4.8) 1ノード	
CPU	AMD EPYC Rome 2.9-3.4GHz 32core x 2CPU
Memory	2TiB (DDR4-3200 x 8チャネル/ソケット)
GPU	NVIDIA A100/40GB mem x 8 (NVLINK)
Storage	24TiB NVMe SSD + Block Storage (up to 1TiB)
Network	100Gbps RDMA (RoCEv2、CX-5)x16 (1.6Tbps) 50Gbps TCP/IP x 1

Cluster Network 想定最大構成	
Latency	1.5マイクロ秒程度
Max nodes	64
Max total CPU cores	4,096
Max total Host memory	128TiB
Max total NVMe storage	1.5PiB
MAX total GPU (A100)	**512GPUs**

OCI HPC CLIによるHPCクラスタ作成/削除

クラスタ作成: $ ocihpc deploy --stack ＜Cluster Name> --node-count <Node数> --region ap-tokyo-1 --compartment-id <compartment-id>

クラスタ削除: $ ocihpc delete --stack ＜Cluster Name>

最新のハードウェアを低価格で提供

Intel Ice lake
HPCクラスタ

CPU ¥6.48 /コア時間
メモリ ¥0.18 /GB時間

- Intel Xeon Ice lake 搭載 (3.0GHz - 3.6GHz)
- ベアメタル (¥325.44/ノード時間)
  - 36コア 512GiBメモリ
  - NVMe SSD 3.2 TB搭載
  - RDMAネットワーク 100Gbps、低遅延
  - 仮想マシン:Flexible VMs

AMD EPYC Milan
汎用ワークロード

CPU ¥3.00 /コア時間
メモリ ¥0.18 /GB時間

- AMD EPYC Milan 搭載 (2.55GHz - 3.5GHz)
- ベアメタル (¥752,64/ノード時間)
  - 128コア 2TiBメモリ
  - ネットワーク 50Gbps x 2
- 仮想マシン: Flexible VMs
  - 1コア、1GB単位で組合せ

Arm Ampere Altra
コストパフォーマンス最高

CPU ¥1.20 /コア時間
メモリ ¥0.18 /GB時間

- Ampere Altra 搭載 Neoverse N1、3.0GHz
- ベアメタル (¥376.32/ノード時間)
  - 160コア 1TiBメモリ
  - ネットワーク 50Gbps x 2
- 仮想マシン:Flexible VMs
  - 1コア、1GB単位で組合せ

NVIDIA A100
高性能 GPU

A100
¥366 / GPU時間

- 最新のNVIDIA GPUを搭載 (¥2,928/ノード時間)
- 40GBメモリ/GPU
- 8GPU - NVLINK
- ホスト
  - EPYC Rome 最大64コア
  - 2TiBメモリ
  - NVMe SSD 24TB搭載
  - ホスト間RDMA
    - 1.6Tbps、低遅延

他社に比べ、圧倒的なコストパフォーマンスを実現

	Oracle Cloud の強み	Oracle Cloud	某クラウドベンダ 同等サービス(公開情報)	
Compute 	✓ ベアメタル ✓ FlexibleVM * ✓ 遅延1.5μsのRDMA環境も提供	¥80.82/時 Compute (VM.Standard2.8; 8コア, 120GiB, Windows)	¥180.48/時	¥ 55%低価格
GPU 	✓ A100 NVLINK 8GPU ✓ NVIDIA GPU Cloudや NVIDIA GRIDも利用可能	¥354/時 Compute - GPU (VM.GPU3.1: NVIDIA Tesla V100/16GB x 1GPU)	¥461.34/時	¥ 23%低価格
Storage 	✓ 1PB年間374万円の Archive Storageも提供	¥5,222/月 Block Volume (1TB, 20K IOPS)	¥171,000/月	¥ 97%低価格
Network 	✓ AD間無償 ✓ **10TB/月まで無償** ✓ **専用線接続時はデータ転送無償**	¥18,972/月 FastConnect (1Gbps, 100TB) *専用線接続	¥614,645/月	¥ 97%低価格

*FlexibleVM : 1CPU物理コア単位、1GBメモリ単位で設定してデプロイ出来る仮想マシン

HPC関連ソリューション

Job Scheduler
- Altair PBS Professional (Market Place)
- Altair Grid Engine
- Slurm (Market Place)、など
  - Cloudbursting スクリプトサンプル提供

Container
- Docker, Singularity

Cluster Management
- Xtreme-D AXXE-L
- 学認クラウドオンデマンド構築サービス、など

HPC SaaS
- Altair HyperWorks Unlimited
- Rescale
- NIMBIX

Parallel File System
- Lustre (Market Place)
- BeeGFS (Market Place)
- Gfarmなど

Object Storage NFS Gateway
- OCI Storage Gateway (無償提供)

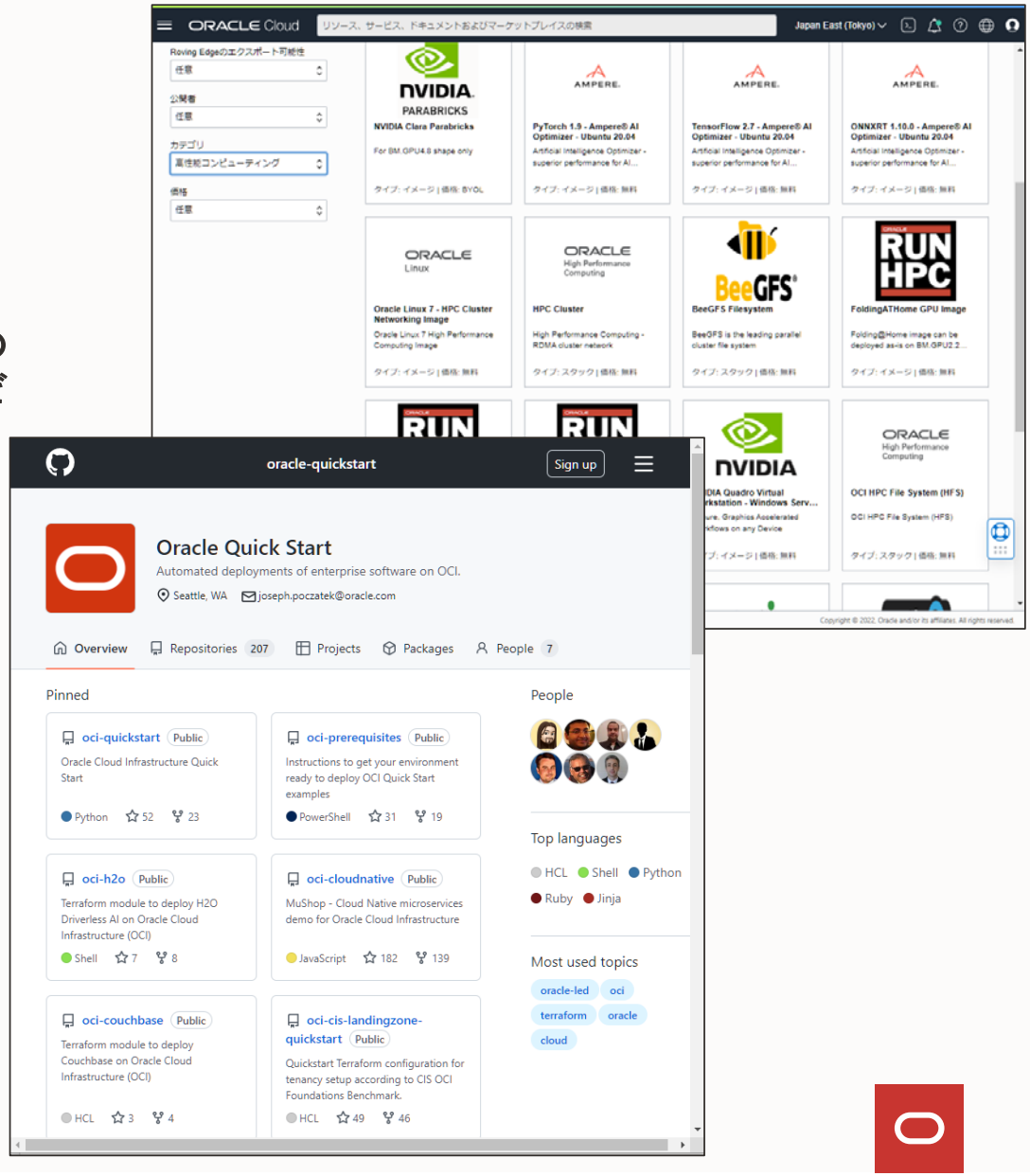
Storage Tiering
- iRODS
- クラウディアン HyperStore 階層化
  - カスタムエンドポイント
- Gakunin RDM拡張ストレージ、など

経験豊富なHPCサポートSIパートナ

HPCソリューション短期デプロイ支援
OCIマーケットプレイス &
github Oracle Quick Start

CAE、グラフィックス・ワークステーション、高性能ストレージ、AI、ゲノム解析など、様々なHPCソリューションの短期デプロイ可能なテンプレートスタック、リポジトリなどを提供

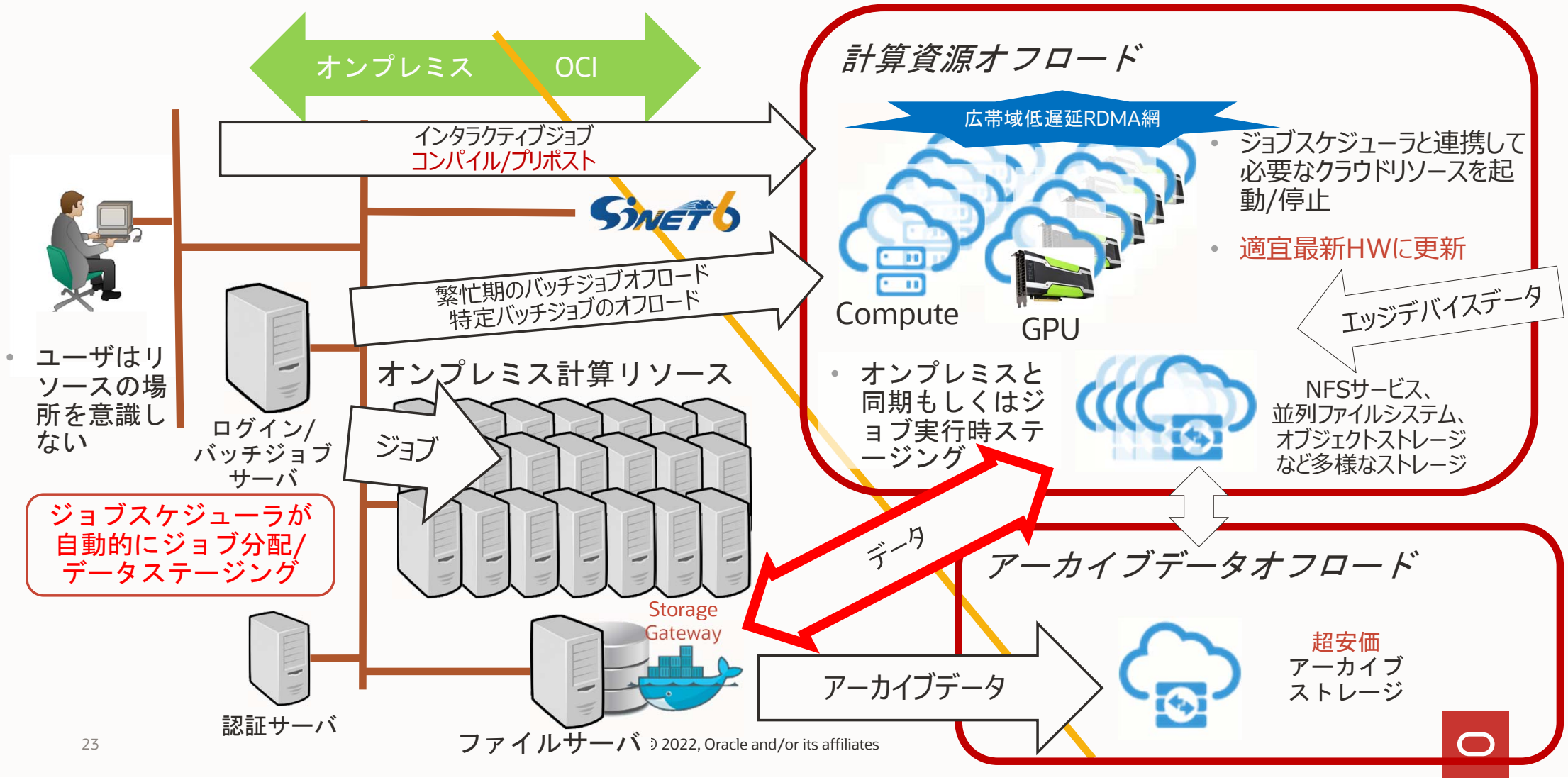
- 低遅延RDMAクラスタ
- 並列ファイル・システム
- AI/機械学習用イメージ
- ゲノム解析用イメージ
- 各種標準的HPCアプリケーション
- GPU 仮想ワークステーション・イメージ
- HPC コマンド (ocihpcコマンド)
- HPC GUI



HPC用途における黎明期クラウドサービスの課題とOCI

～2015年までのクラウドサービス	Oracle Cloud Infrastructure (2016～)
オーバーサブスクリプションのある仮想サーバとネットワークのため、ストレージやネットワークのI/Oが不安定で、性能再現性がない。	物理ネットワークはオーバーサブスクリプションの無いフラットネットワークでバンド幅とレーテンシを確保。広帯域/低遅延のRDMAネットワークも提供。性能再現性を提供。
HPCで実用的なIOPS値を備えるブロックストレージが高価。	最新ハードウェアを用いたベアメタルサーバを時間単位で提供。
アウトバウンド通信データ量に課金され大量のデータを扱うHPC分野では通信費用が高額となる。	仮想マシンも物理コア分割でCPUコア、メモリ、NIC帯域にオーバサブスクリプション無し。
日本国内データセンタのサービスが高く設定される。ドル単価で為替影響あり。	利用サービス切換自由、契約変更無しにテクノロジリフレッシュ可能。
	Outbound Traffic 毎月10TBまで無料。SINETなど閉域接続では通信データ量無制限に無料。
	低価格 IaaSをWW同一円定価で提供。

Oracle Cloud Infrastructureを用いたHybrid HPC システムイメージ



OCI for HPC & AI ユースケース1
- 計算資源オフロード –

中・小規模ジョブを性能再現性あるクラウドにオフロードし、イノベーションサイクルを加速

[解決出来る課題やメリット]

- オンプレミス大規模計算資源のより有効な活用
- 繁忙期のリソース待ち時間軽減

- シーズナリティ/利用量に応じたコスト最適化
- 契約変更不要で常に最新技術を利用可能
- オンプレミス設置スペース/消費電力軽減
- 障害からの復旧短縮
- 計算資源利用経費のCAPEXからOPEXへの移行

[メリットの大きいワークロードタイプ]

- シーズナリティが大きい、散発、突発的もしくは短期間
- オンプレミスとワークロードの行き来がない/少ない
- 標準的アプリケーション利用

[対象となるワークロード種類・サイズ → 適用サービス]

- 2万CPUコア/288TiBメモリ程度、もしくは512GPU+4096CPUコア/128TiBまでのジョブ → ベアメタル+RDMAネットワーク
- 128コア/2TiB、160コア/1TiB、もしくは8GPU+64CPUコアまでのジョブ → ベアメタル
- パラメータサーチ、アンサンブルなどEPなジョブ → ベアメタル または FlexibleVM*注

*FlexibleVM : 1CPU物理コア単位、1GBメモリ単位で設定してデプロイ出来る仮想マシン

OCI for HPC & AI ユースケース2
- アーカイブデータオフロード -

データ転送量課金のないメンテナンス不要で高耐久性を提供する安価なストレージ基盤

[解決出来る課題やメリット]

- オンプレミス高速ファイルサーバ空き容量確保
- オンプレミス設置スペース/消費電力軽減
- 他機関からの容易なアクセス

- 99.999999999%(11-9)の年間データ耐久性
- ストレージメディアメンテナンス不要
- ストレージデバイス陳腐化対策不要
- SINET経由ダウンロードデータ量無課金

[メリットの大きい対象データ]

- オンプレミスからアクセスの少ないデータ
- ソースデータ
- 複数機関から参照されるデータ

[関連サービス]

- クラウド上の分析/可視化サービスを利用した簡便なプリポスト処理

RIKEN 富岳 Fugaku

Oracle Cloudで「富岳」の高度な計算資源の有効活用と研究成果創出を促進

- パブリック・クラウドと連携した「富岳」の柔軟な利活用の支援を目的に、学術情報ネットワークSINET を介した、「富岳」 とOracle Cloud Infrastructureとの接続
- 膨大な研究データの転送コストを気にすることなく、「富岳」のネットワークから、Oracle Cloud Infrastructureの高い性能のコンピュートやストレージ・リソースなどを予測可能なコストで利用可能
- 安全かつ低コストでの接続により、セキュリティ、パフォーマンス、伸縮性に優れたコンピュートやストレージのリソースを「富岳」のネットワークから低コストで利用可能に

大阪大学、日本電気共同、日本オラクル共同研究事例

～ 新型コロナウィルス感染症などに向けクラウドバースティング を通じたベアメタル計算資源の提供～

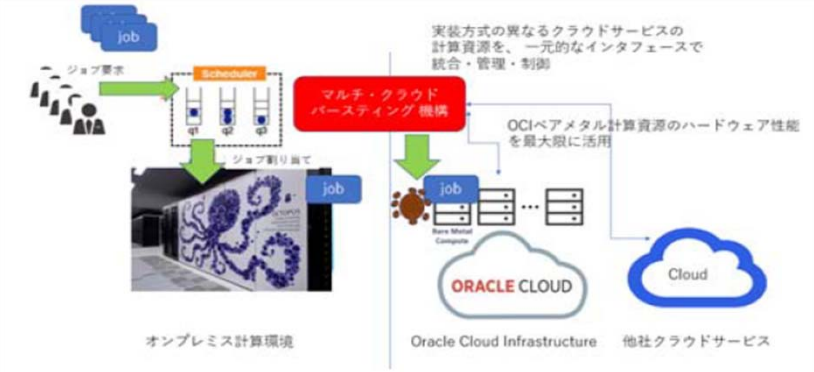
大阪大学 OSAKA UNIVERSITY

研究背景

- CMC のスーパーコンピュータOCTOPUS は、利用者からのスカラ型スーパーコンピュータに対する様々な計算ニーズ・需要を収容可能であり、非常に高い利用率で利用される状況になっています。しかし、その一方で、利用者の計算要求から計算完了までの待ち時間が定常的に長時間になるという問題が深刻になりつつあり、利用者からの問い合わせ・相談の声も大きくなっていた。

ソリューション

- Oracle Cloud Infrastructure　bare metal



図：マルチクラウドバースティング機構を通じた OCI ベアメタルクラウド計算資源.

Why Oracle Cloud

- ベアメタルマシンも仮想マシンと同様にオンデマンドに必要な時に起動し、不要な場合は停止するHPC 環境を構築することが可能
- 最新のCPU、GPU、高性能なノード間通信等のHPC 関連技術がリリースされた際には迅速に提供さる。
- 一般的にHPC用途では多くの分野でデータが大きくなる傾向があるが専用ネットワーク接続サービスを利用することで、転送データ量が無制限に無料
- OCI のベアメタル計算資源であれば、オンプレミスのOCTOPUS 計算ノードと同様にユーザの計算要求を実行出来ることが確認された
- 多くの計算ノードを利用する並列計算においてクラウド計算資源を利用した場合でも高いスケーラビリティが得られる

【 研究成果のポイント】

大型計算機におけるクラウドバースティングが実現可能であることが実証されると利用者が計算結果を得られるまでの時間の削減ができるだけでなく、COVID-19 対策のような急な計算需要拡大への対応含め、計算機リソースの問題で解決できなかった事象を解析できるようになる。

これにより、大型計算機を用いた研究分野において、学術的・教育的に大きな成果が出ることが期待されます。

また、低コストでのスケールアウトを示すことにより、学術機関と企業との連携による産業利用・産学連携の加速、企業・社会課題の解決が加速されることが期待されます。

OCIのカーボンニュートラルへの対応

電力使用量の大きなHPC環境をエネルギー効率に優れたクラウド環境へ移行

Oracle Cloudの取り組み

59%

世界中のOracle Cloudのデータ・センターにおける使用電力のうち、59%が再生可能電力として2019年に認証済み

100%

ヨーロッパの10のリージョンで、100% 再生可能エネルギーを使用

100%

2025年までに、Oracle Cloudの電力を100%再生可能エネルギーとするゴールを設定

Oracle Cloud Infrastructureの導入により、IT運用コストの大幅な削減、エネルギー消費量の20%削減、管理とコンプライアンスの簡素化、そして持続的な成長計画に必要なスケーラビリティを実現しました。

Vlad Moca, Deputy Group IT Director, KMG Rompetrol SRL

HPCでクラウドとオンプレミスを連携させるにあたっての課題

- ✓ オンプレミスとクラウドサービスの間のデータ移動/同期に掛る時間と費用 (データグラビティ対策)
  - 利用形態など踏まえデータ転送時間の最小化/隠蔽する工夫が必要なケースあり
  - OCIはSINETクラウド接続サービス経由の転送データ量費用は無料
- ✓ オンプレミスとクラウドの認証連携
- ✓ オンプレミスとクラウドの適用区分
  - 採択課題/研究室毎、アプリケーション毎、CPU/GPU/メモリサイズ/並列数など
  - Xヶ月アクセスのないファイルなど
- ✓ クラウド利用の費用負担モデル
  - HPCIリソースとの価格差
  - センター or ユーザ、センター経由ユーザ
- ✓ クラウドサービス(役務、従量性)の複数年継続利用可能な契約手続き

【お問い合わせ先】

日本オラクル株式会社
クラウド事業統括 公共営業本部
松山 慎 (まつやま まこと)
E-Mail: Makoto.Matsuyama@oracle.com
電話: 080-1289-8315

お気軽にご連絡ください!


ORACLE