Gfarm ワークショップ2021

Amazon Web Services におけるクラウド HPC ストレージ

～Amazon FSx for Lustre を中心として～

Daisuke Miyamoto, Ph.D.
Specialist Solutions Architect, HPC
Amazon Web Services Japan K.K.

2021/03/05

自己紹介

- 名前
  宮本 大輔

- 所属
  アマゾン ウェブ サービス ジャパン 株式会社
  技術統括本部
  Specialist Solutions Architect, HPC

本日の概要

- HPC 領域におけるクラウド活用の利点と構成例
- HPC on AWS を支えるサービス
- Amazon FSx for Lustre のご紹介とクラウド HPC ストレージの考え方

クラウド HPC とは？（一例）

クラウドの持つ
「必要な時に」「必要なリソース」を「必要な量だけ」
確保できるという特性を HPC に活用

『NIST によるクラウドコンピューティングの 定義』
基本的な特徴
スピーディな拡張性 (Rapid elasticity): コンピューティング能力は、伸縮自在に、場合によっては自動で割当ておよび提供が可能で、需要に応じて即座にスケールアウト／スケールインできる。ユーザにとっては、多くの場合、割当てのために利用可能な能力は無尽蔵で、いつでもどんな量でも調達可能のように見える。

IPA による翻訳: https://www.ipa.go.jp/files/000025366.pdf

クラウド HPC とは？（一例）

クラウドの持つ
「必要な時に」「必要なリソース」を「必要な量だけ」
確保できるという特性を HPC に活用

そのためにはクラウド・ HPC 両方の観点で
考慮すべきポイントがある

今回、特に後半では Amazon FSx for Lustre というサービスを例に
クラウド HPC のどういうところが面白いのか、どういうところが難しいのかを紹介

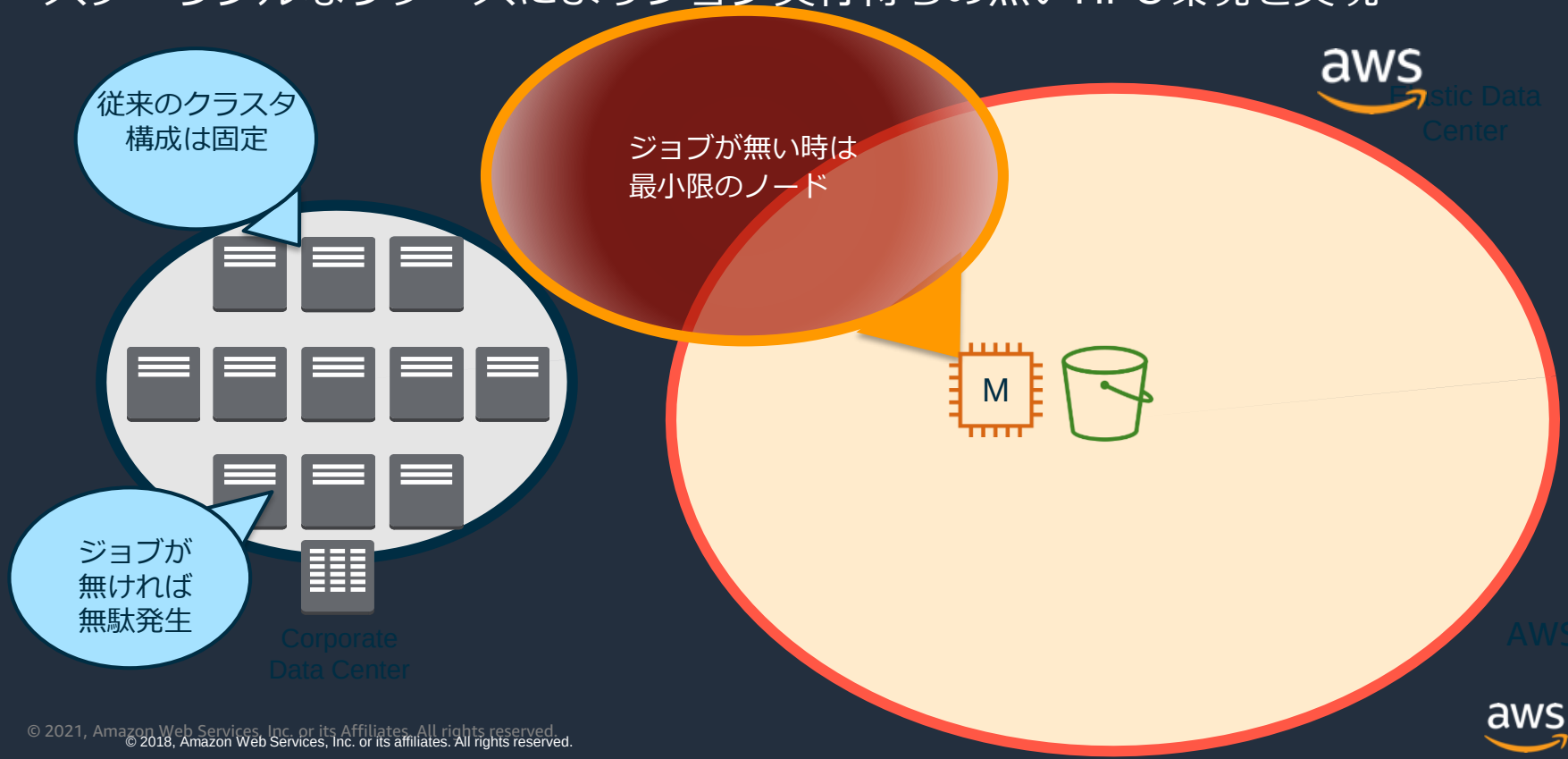
HPC on AWS の特徴・利点

これまでの HPC クラスタの課題

- サーバ台数が限られており、需要が増加する時期には長大なキュー待ち時間が発生する
- 同じ環境を複数メンバーで共有するため、アプリケーションによってはリソースが無駄になることも
- サーバ台数が多く、ハードウェアの保守・管理が煩雑

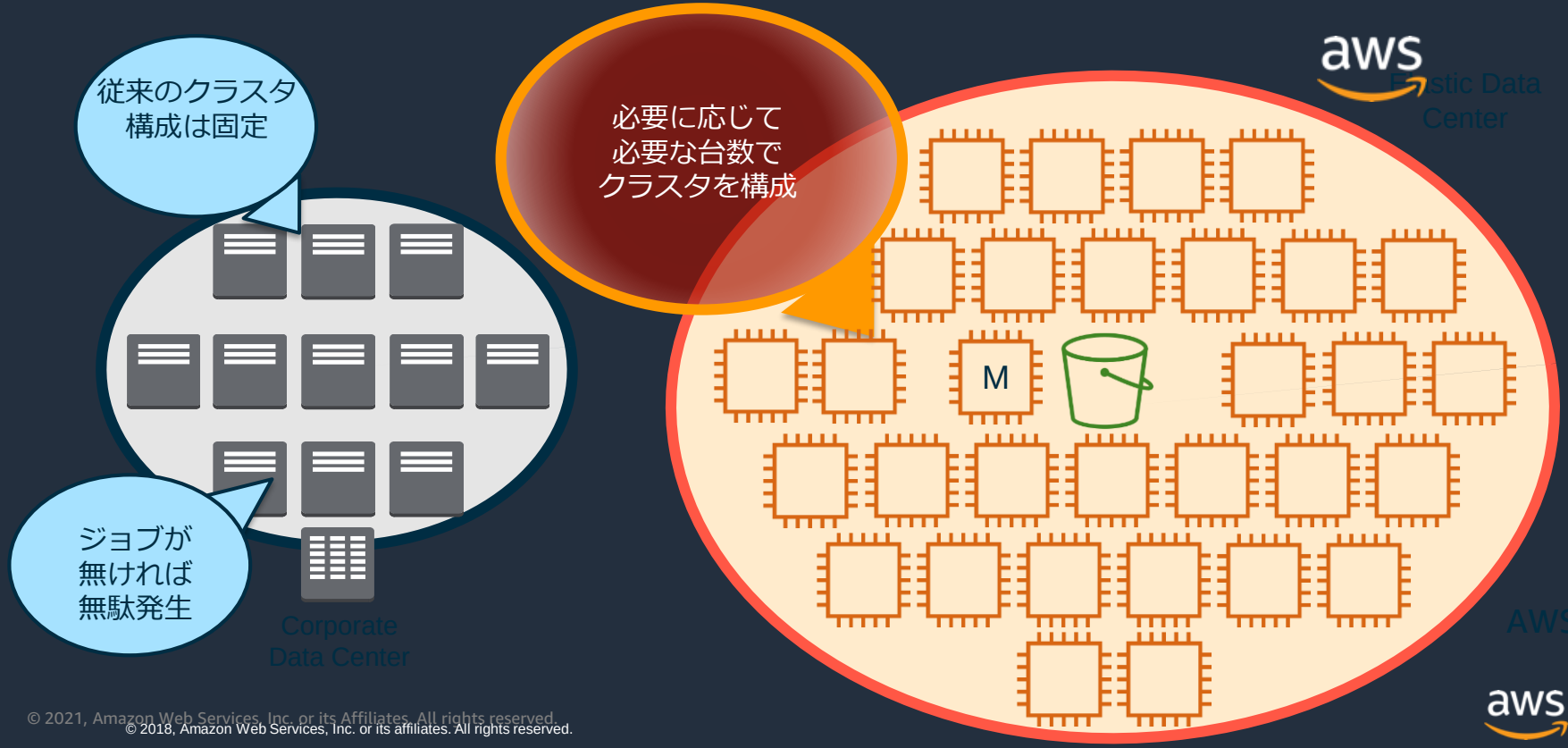
AWSなら、必要な時に必要なだけ利用可能

スケーラブルなリソースによりジョブ実行待ちの無いHPC環境を実現



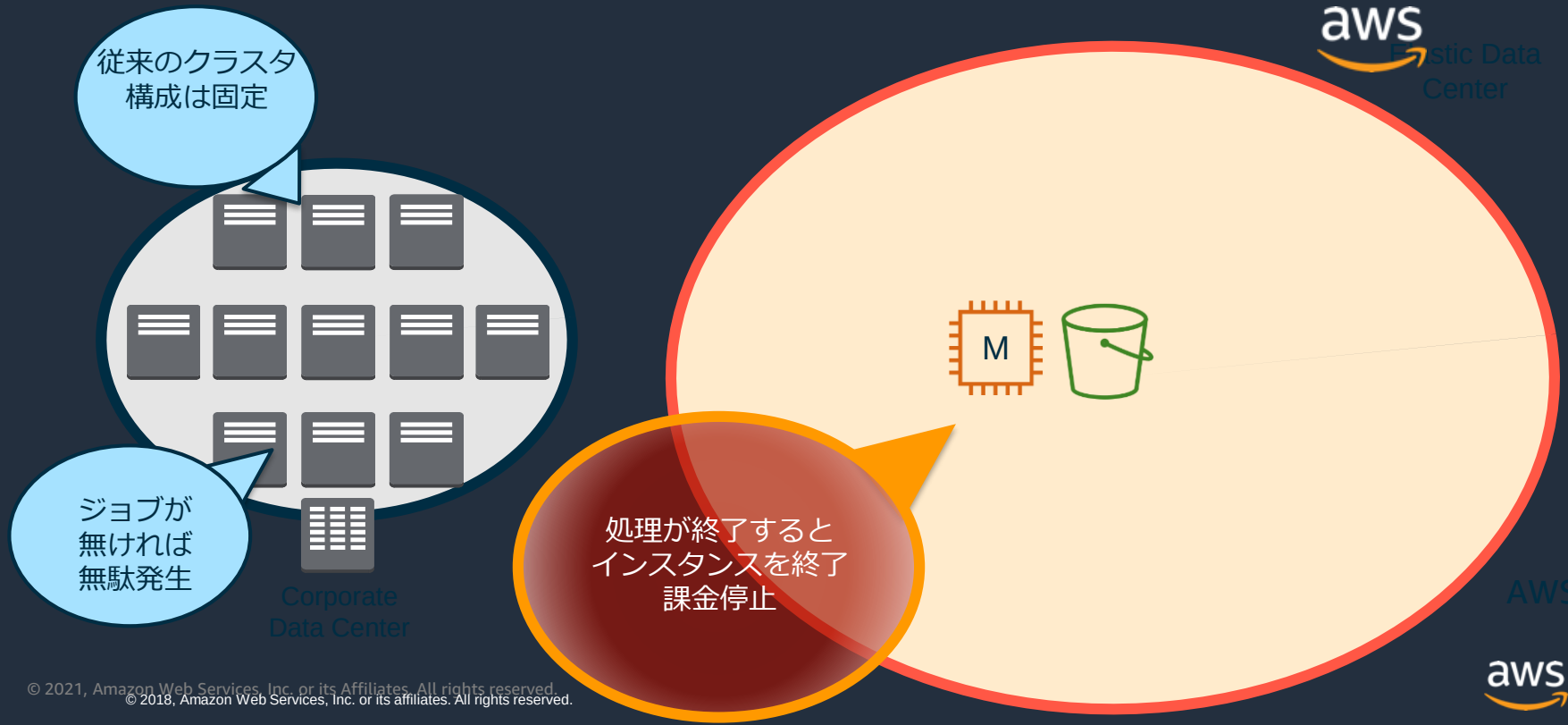
AWSなら、必要な時に必要なだけ利用可能

スケーラブルなリソースによりジョブ実行待ちの無いHPC環境を実現



AWSなら、必要な時に必要なだけ利用可能

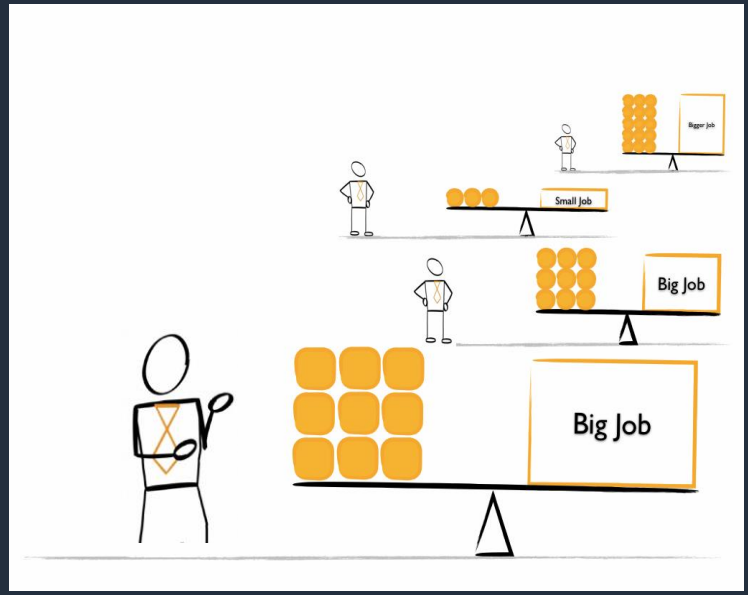
スケーラブルなリソースによりジョブ実行待ちの無いHPC環境を実現



アプリケーションに合わせた構成のクラスタを構築可能

ユーザやタスク単位で専用のクラスタを構築できるため
要件や規模に合わせて、最適構成のクラスタを作成可能

- CPUコア/メモリ
- ストレージ
- アクセラレータ
- ネットワーク
- インストールするソフトウェア



One size does not fit all!

計算機管理の手間を抑える

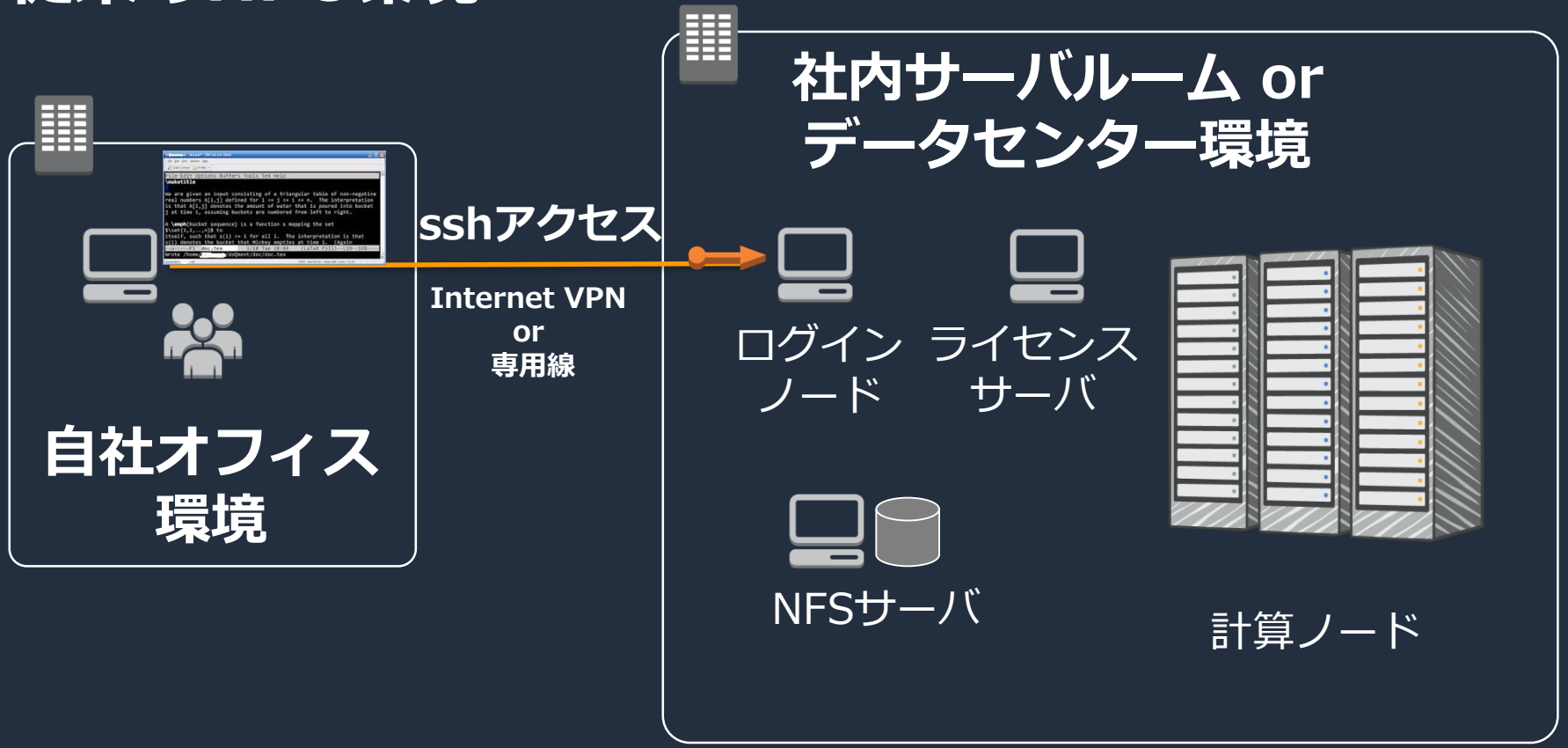
・ハードウェア保守
・ネットワーク管理/保守
・電源管理
・空調管理
・設置場所の費用/運用

計算機の規模が大きくなればなるほど
大変に、、、

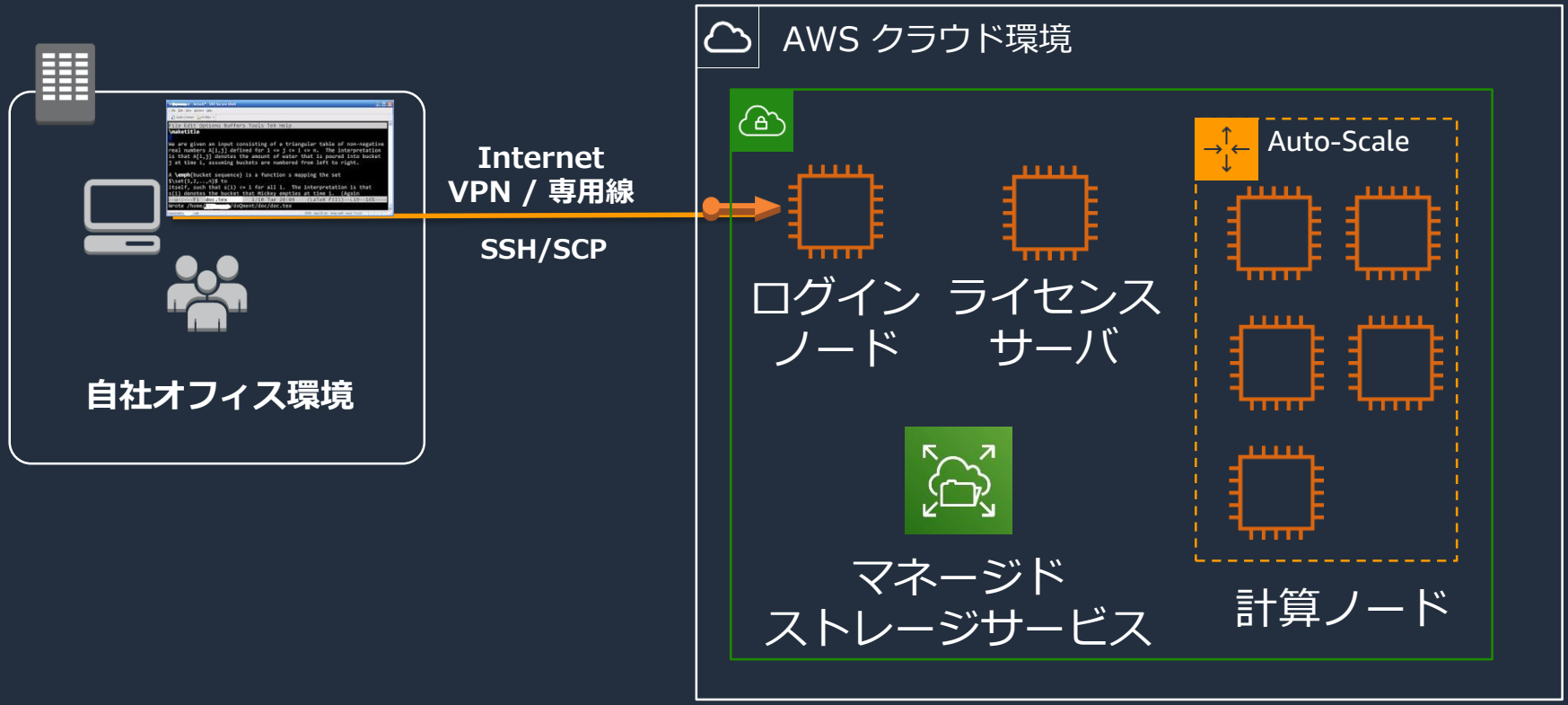
競争優位につながらない物理的管理は全てAWSにお任せ
他社と差別化可能な部分に集中

HPC on AWS の利用イメージ

従来のHPC環境

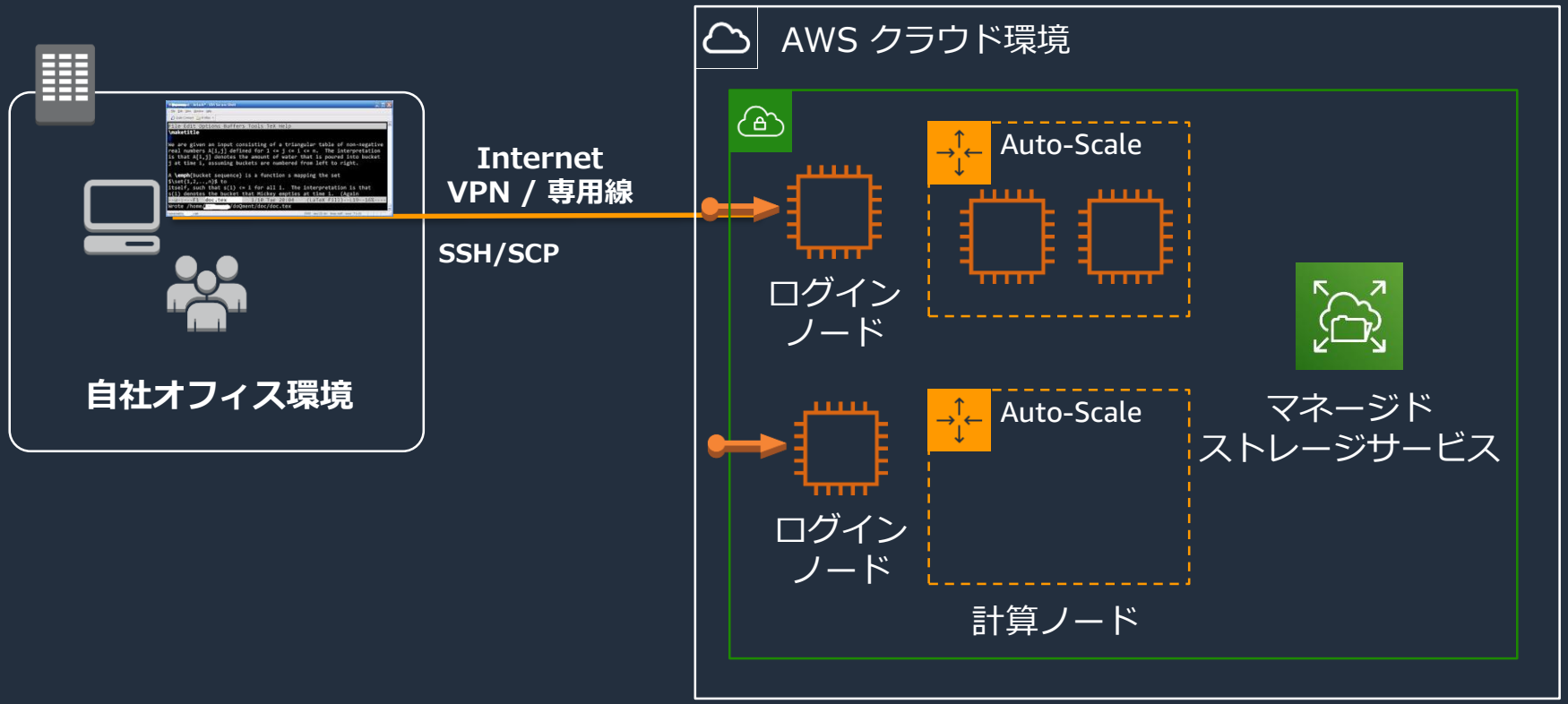


AWSでも基本的なシステム構成は同じ



需要に応じて伸縮する計算環境・マネージドサービスの活用

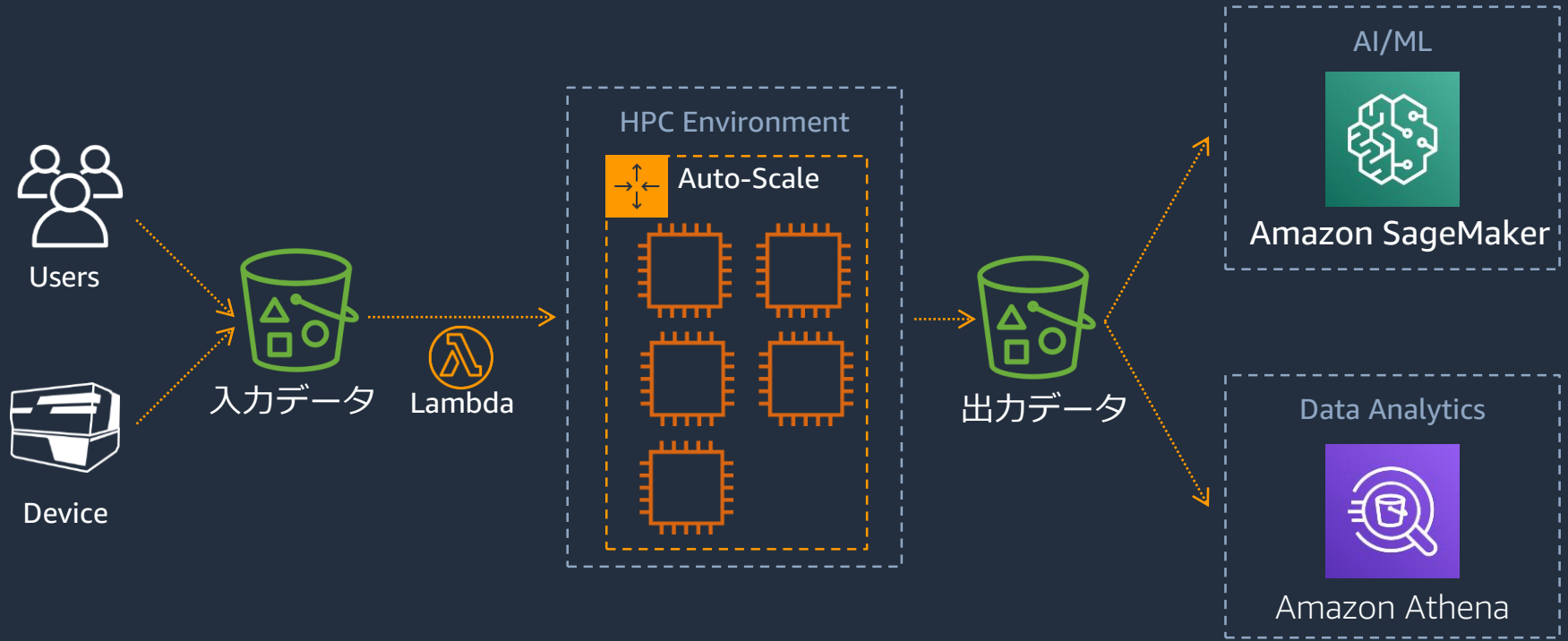
AWSでは、1人1クラスタなどより柔軟性の高い構成も可能



計算ノードは自動でスケールするため、複数クラスタを作成しても維持コストが低い

データドリブンなHPC環境とデータ活用

データのアップロードをトリガーにHPC環境を展開し自動で処理を行う
更にS3のデータレイク化により大規模シミュレーション結果を機械学習環境で活用



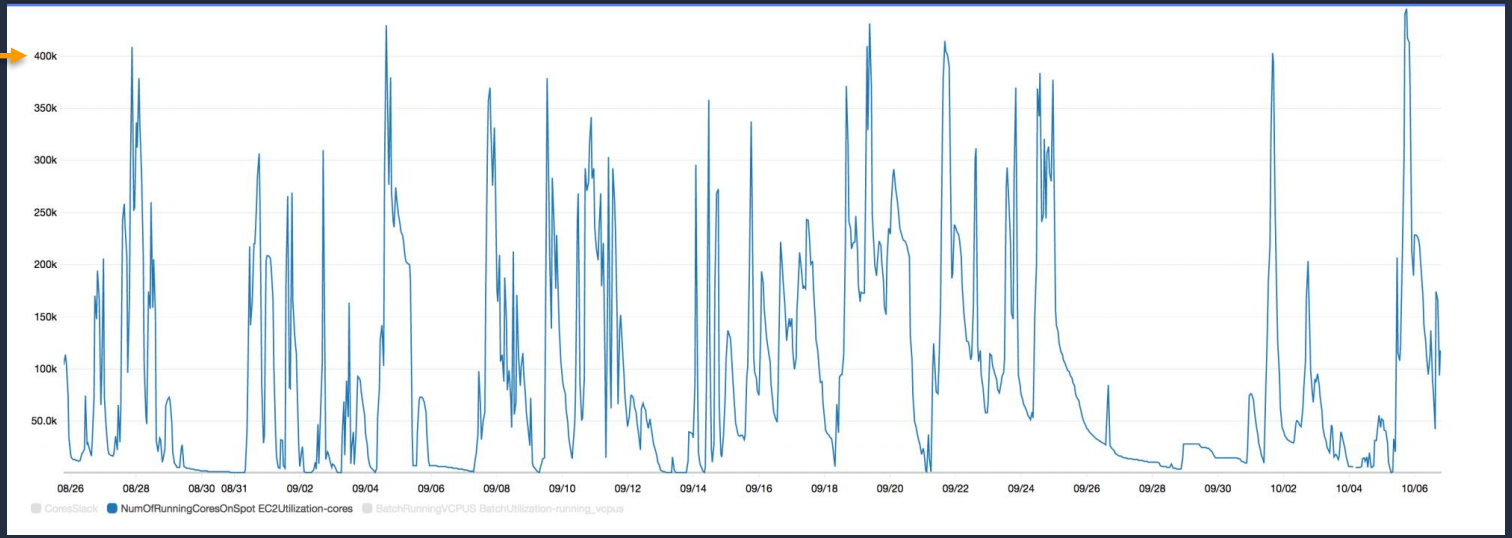
HPC on AWS 活用事例

Mobileye

大規模な自動運転のシミュレーション環境を AWS Batch を用いて構成

最大同時 500,000 CPUコアを利用して、1時間あたり70年分のコンピューティングを実行、S3上のデータを1ヶ月あたり100PB処理

40万 vCPU



https://d1.awsstatic.com/events/reinvent/2019/Navigating_the_winding_road_toward_driverless_mobility_AUT307.pdf

Descartes Labs

Amazon EC2 C5インスタンスを使用したクラスタで約2PFLOPSのLINPACK性能を達成

- Top500(2019/06)で136位にランクイン
- 41,472コア(Xeon Skylake 3.0GHz)利用
- 2.6h稼働、コストは約$5,000で実現

> “One of the more interesting aspects of the story is that **we didn’t ask Amazon to give our engineers any special dispensation, discount, or custom planning or setup**. We wanted to see if we could do this on our own, which if completed successfully, would also be a testament to the self-service model of AWS. ”

Amazon EC2 C5 Instance cluster us-east-1a - Amazon EC2 Instance Cluster C5, Xeon Platinum 8124M 18C 3GHz, 25G Ethernet

Site:	Descartes Labs
System URL:	https://aws.amazon.com/ec2/instance-types/c5/
Manufacturer:	Amazon Web Services
Cores:	41,472
Memory:	157,824 GB
Processor:	Xeon Platinum 8124M 18C 3GHz
Interconnect:	25G Ethernet
Performance	
Linpack Performance (Rmax)	1,926.4 TFlop/s
Theoretical Peak (Rpeak)	3,981.31 TFlop/s
Nmax	2,985,984
Power Consumption	
Power:	
Software	
Operating System:	Amazon Linux 2
Compiler:	gcc 7.3.1
Math Library:	Intel MKL
MPI:	OpenMPI 4.1.0

https://medium.com/descarteslabs-team/thunder-from-the-cloud-40-000-cores-running-in-concert-on-aws-bf1610679978

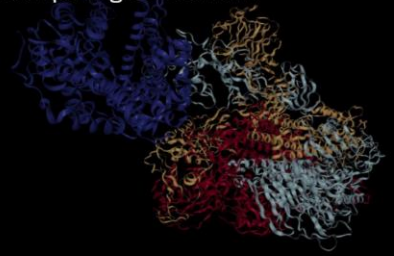
re:Invent で紹介された HPC の取り組み

The COVID-19 High Performance Computing Consortium に参加し、計算リソースを提供

moderna も COVID-19 ワクチン開発のために AWS を活用

How has HPC impacted the worldwide response to COVID-19?

The COVID-19 High Performance Computing Consortium



AWS provides no-cost access to HPC capacity to select research projects

- Over **$1.5M in AWS credits** across **17 projects** since March 2020
- Offering free technical support

novel techsciences

Covid19Tree

Moderna uses AWS to accelerate its COVID-19 vaccine development

moderna™
messenger therapeutics

"AWS's breadth and depth of services are supporting our mission to create a new generation of medicines for patients and are instrumental in our quest to develop a vaccine for COVID-19 and other life-threatening diseases."

Stéphane Bancel
Moderna CEO

https://virtual.awsevents.com/media/1_v2c1s59g

日本国内でも多様な分野で利用が広がる HPC on AWS

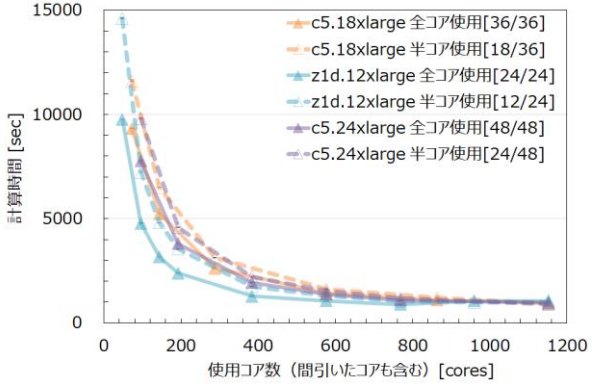
- CAE・製造
- 半導体デバイス設計
- 創薬
- ゲノミクス
- 金融
- 気象シミュレーション

三菱電機株式会社 先端技術総合研究所 様

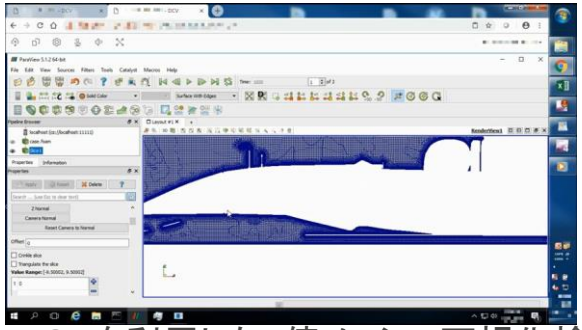
解析システムのクラウド利用推進を目的として、汎用熱流体解析ソフトウェアの検証を実施

- オープンソースベースの汎用熱流体解析ソフトウェア(iconCFD)を用いてAWS上で並列性能ベンチマークテストを実施
- Amazon EC2 の通常のイーサネットでも600コア程度まで性能がスケールすることを確認(今後100Gbpsのネットワークや Elastic Fabric Adapter(EFA)を使用することで更なるスケーラビリティー向上も期待できる)
- 計算結果の可視化処理もAWS上で実行、GPU搭載インスタンスを利用することで1億メッシュ規模のデータも問題なく扱えた
- クラウドの利用により、初期投資を小さくし、大規模な解析をすぐに実行することができる

・iconCFDは，Icon Technology & Process Consulting Ltd.とIDAJによって開発されました.
・iconCFDは，GPLに準じたオープンソースベースの汎用熱流体解析プログラムです.



Amazon EC2上での並列処理性能検証結果



NICE-DCVを利用した1億メッシュ可視化検証

塩野義製薬株式会社 様

SHIONOGI

「一日でも早く薬を届ける」ためにクラウドHPCを活用

多様な創薬ワークロードに対して、適したAWSのサービスを選択

ゲノム分析

- **AWS Batch** を活用し、ゲノム分析に必要な解析パイプラインを構築
- TB級のデータを必要な時間内での処理を実現
- **スポットインスタンス**の利用でEC2コストを約50%～節約

タンパク質立体構造解析

- **AWS ParallelCluster** を使用し既存のジョブスケジューラから変えずに移行
- P3インスタンスによるGPU活用
- **NICE-DCV**により計算結果をクラウド上で即座に確認

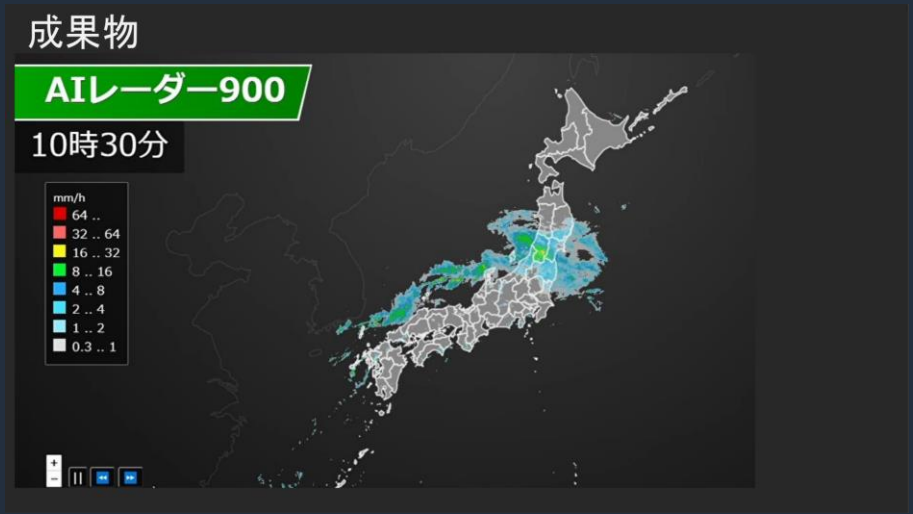
AWS Batch

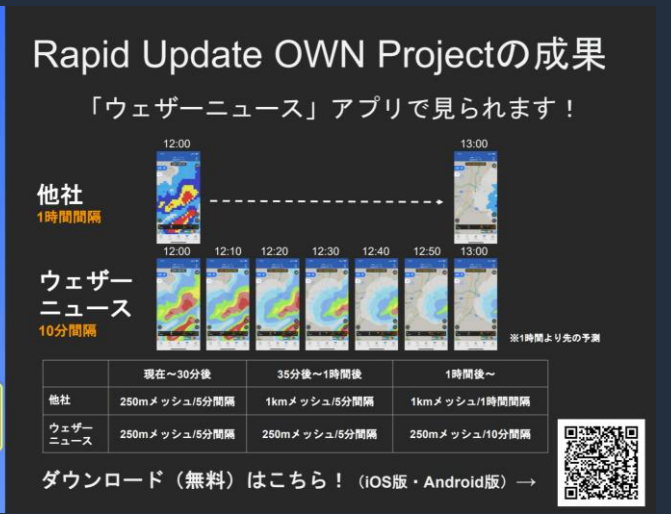
AWS ParallelCluster

aws

株式会社ウェザーニューズ様

台風やゲリラ豪雨など気象リスクに対して
短い予報間隔・高い更新頻度の気象予報をユーザーに配信するためクラウドを活用





AWS Summit Online Japan 2020
新たな気象リスクへの挑戦を可能にした HPC on AWS
https://resources.awscloud.com/aws-summit-online-japan-2020-on-demand-industry-3-12034/cus-98-aws-summit-online-2020-weathernews

株式会社ウェザーニューズ様

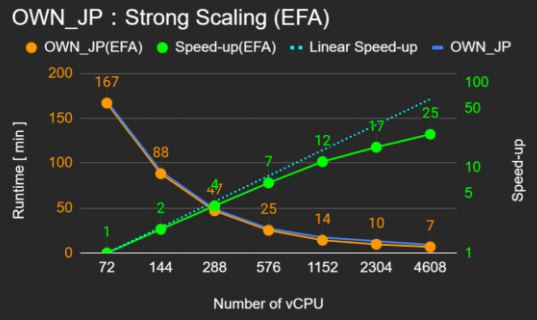
AWS ParallelCluster + Elastic Fabric Adapter によるスケーラビリティ
一定期間内に計算を完了させるための Multi Region Fail Over クラスタ構成
スポットインスタンス活用によるコスト削減

EFAの利用

Parallel Clusterの設定
でEFAを用いた場合

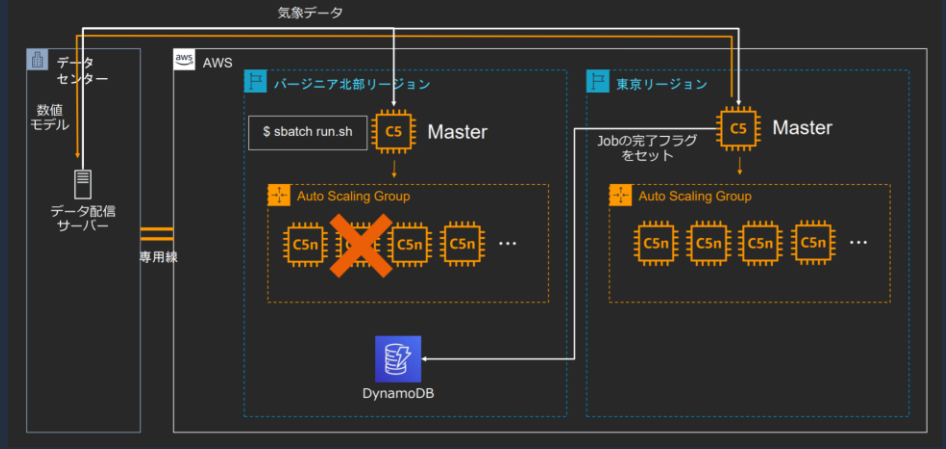
特に並列数を増やした
場合に高速化された

vCPU数が少ない場合は
1~3%の改善だが
4068まで並列化させた場合は
25%の改善が見られた



改善率(%)							
	1	3	4	7	18		25

副系での計算処理の流れ



AWS Summit Online Japan 2020
新たな気象リスクへの挑戦を可能にした HPC on AWS
https://resources.awscloud.com/aws-summit-online-japan-2020-on-demand-industry-3-12034/cus-98-aws-summit-online-2020-weathernews

HPC on AWS を支えるサービス

AWS における HPC 関連サービス

多様な HPC ワークロードに対応するための数多くのサービス

コンピュート

Amazon EC2

用途に応じて多様なインスタンスを利用可能な仮想サーバサービス

- P4d: NVIDIA A100 GPU 搭載
- F1: Xilinx Virtex UltraScale+ 搭載
- C5n: 100 Gbps のネットワーク帯域

スポットインスタンスの活用で大幅なコスト減も可能

ネットワーク

Placement Group

EC2インスタンスの基盤上の配置を制御してネットワークを高速化

Elastic Fabric Adapter

MPI/NCCL 専用の低レイテンシネットワークアダプタ

ストレージ

FSx for Lustre

S3連携可能な高速な分散ファイルシステムをフルマネージドで提供

管理自動化

AWS ParallelCluster

AWS上に HPC クラスタを自動で構築。**Slurm** / SGE / Torque といったジョブスケジューラに対応しており既存HPC環境からの移行が容易

AWS Batch

コンテナベースの大規模バッチジョブコンピューティング環境をフルマネージドで提供

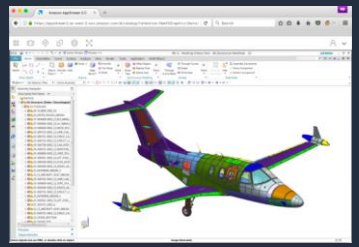
可視化

NICE-DCV

GPUアクセラレーションに対応し、インタラクティブなアプリケーションに適したデスクトップ仮想化

Amazon AppStream 2.0

フルマネージドのアプリケーションストリーミングサービス



仮想サーバサービス Amazon EC2 (Elastic Compute Cloud)

- 必要なときに必要な計算リソースを確保可能な仮想サーバサービス
- 数分で起動し、秒単位の従量課金（一部タイプについては1時間単位）
- 独自の仮想化基盤 Nitro System により、仮想化オーバーヘッドを極小化
- ワークロードに応じて様々なインスタンスタイプを選択可能

高性能計算向けインスタンスタイプの例

高性能CPUの選択肢

Intel Xeon processor
(x86_64 arch)

C5 インスタンス
最大3.9GHz駆動
Cascade Lake or Skylake
M5zn インスタンス
最大全コア4.5 GHz駆動
Cascade Lake

AMD EPYC processor
(x86_64 arch)

C5a インスタンス
最大3.3GHz駆動
Rome

AWS Graviton Processor
(64-bit Arm arch)

C6g インスタンス
64bit Arm Neoverse N1ベース
Graviton2 CPU搭載

アクセラレータの選択肢

NVIDIA
P3: V100 GPU搭載
P4d: A100 GPU搭載
G4: T4 GPU搭載

Xilinx
UltraScale+ FPGA
F1インスタンス

EC2 購入オプション

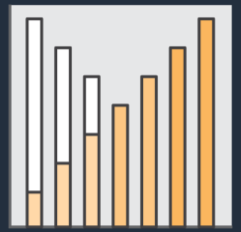
オンデマンドインスタンス

長期コミット無し、使用分への支払い(**秒単位/時間単位)**。Amazon EC2の定価

スパイクするようなワークロード

リザーブドインスタンス (Savings Plans)

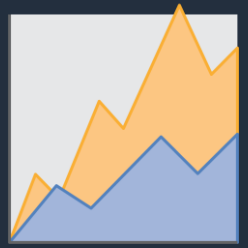
1年/3年の長期コミットをする代わりに**大幅なディスカウント**価格



一定の負荷の見通しがあるワークロード

スポットインスタンス

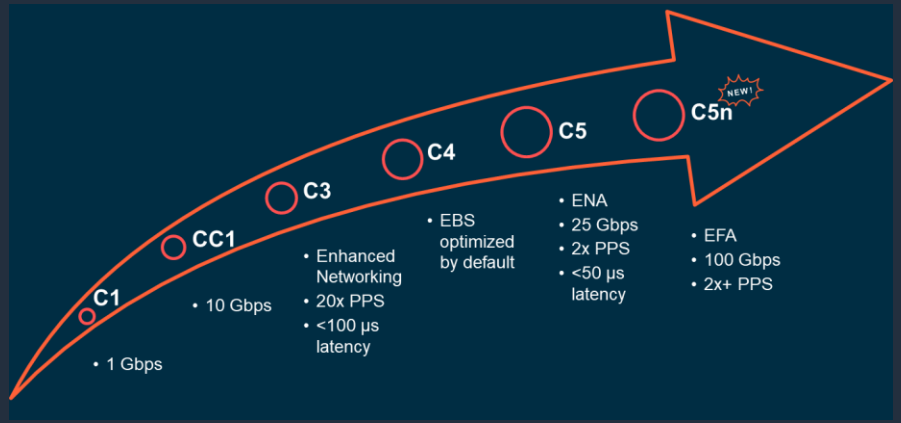
Amazon EC2の空きキャパシティを活用し、**最大90%値引き。中断が発生することがある**



中断に強く、かつ様々なインスタンスタイプを活用できるワークロード

HPC 等では特にスポットインスタンスを活用することで
コストパフォーマンスの良い計算が可能

EC2 の高性能ネットワーク技術



EC2のネットワークも進化を続け
現在はEC2インスタンスあたり
最大 400Gbps (P4dインスタンス) まで
サポート※

※インスタンスタイプとサイズによって通信帯域は異なります

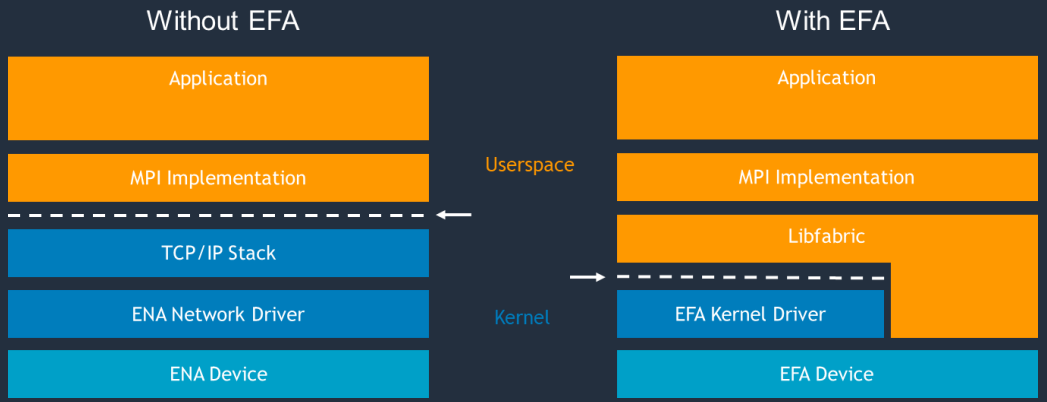
- **拡張ネットワーキング**
  - SR-IOVに対応し、仮想化オーバーヘッドを低減することで低レイテンシでの通信が可能
- **Cluster Placement Group**
  - **インスタンスの配置を最適化すること**で広帯域/低レイテンシ/フルバイセクション通信を実現
- **Elastic Fabric Adapter**
  - **HPC向けに、MPI (Message Passing Interface) やNCCL (NVIDIA Collective Communications Library) など**の libfabric 対応のアプリケーションでの通信を**より低レイテンシ化**

http://docs.aws.amazon.com/ja_jp/AWSEC2/latest/UserGuide/enhanced-networking.html

EFA (Elastic Fabric Adapter)

MPI/NCCL専用のネットワークアダプタ Elastic Fabric Adapter により
低レイテンシでのノード間通信を実現

- 利用には対応したMPI環境（Intel MPI/OpenMPI）が必要だが、プログラムの変更は原則不要
- EFA対応インスタンス：c5n.18xlarge, c6gn.16xlarge, p4d.24xlarge , p3dn.24xlarge etc.



L. Shalev, H. Ayoub, N. Bshara and E. Sabbag, "Supercomputing on Nitro in AWS Cloud," in IEEE Micro, doi: 10.1109/MM.2020.3016891.
https://ieeexplore.ieee.org/document/9167399

EFAで使用されているSRD (Scalable Reliable Datagram)

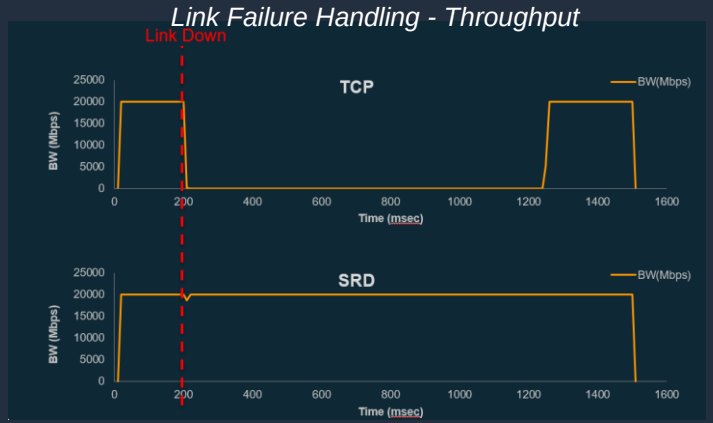
AWS のデータセンターネットワーク向けに
新たに開発されたトランスポートプロトコル

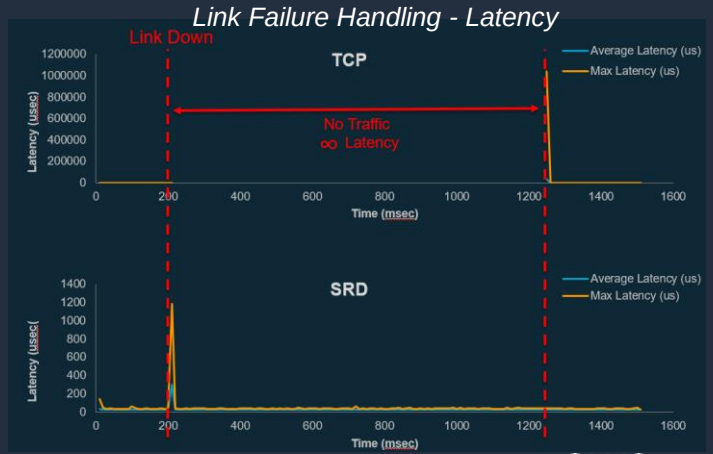
配信保証：EC2のリソースを使用せずに保証を行う

マルチパスルーティング：データセンターの複数のネットワーク経路を活用

レイテンシ・ジッターの低減：独自の link/switch ダウン検出、輻輳制御

アウトオブオーダーでの転送：ブロックを抑制

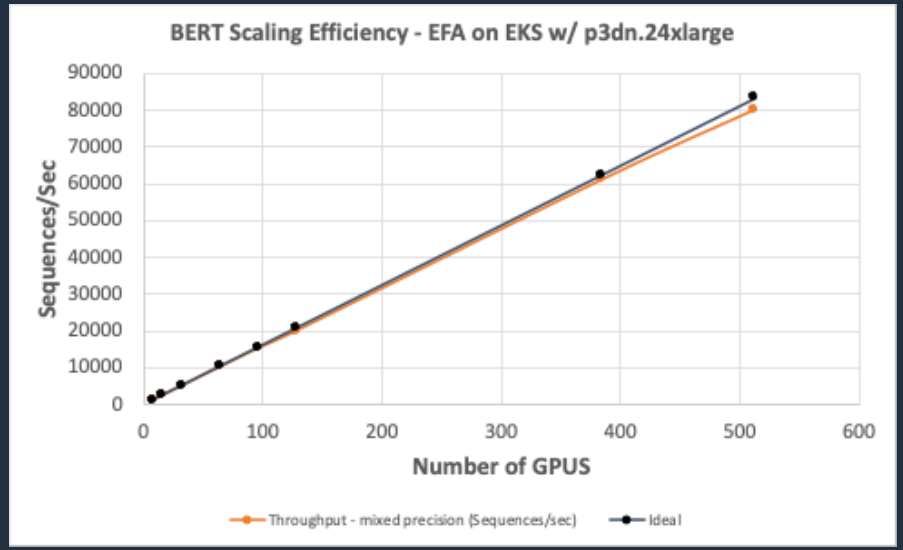




L. Shalev, H. Ayoub, N. Bshara and E. Sabbag, "Supercomputing on Nitro in AWS Cloud," in IEEE Micro, doi: 10.1109/MM.2020.3016891.
https://ieeexplore.ieee.org/document/9167399

参考：Elastic Kubernetes Service (EKS) & EFA

EFA は EKS による Kubernetes 環境上でも利用可能



Number of Nodes	Number of GPUS	Batch Size / GPU	Max Accumulation Steps	Sequence length	Throughput - mixed precision (Sequences/sec) - EFA
1	8	64	128	128	**1303**
2	16	64	128	128	**2561**
4	32	64	128	128	**5114**
8	64	64	128	128	**10254**
12	96	64	128	128	**15315**
16	128	64	128	128	**20053**
48	384	64	256	128	**61161**
64	512	64	256	128	**80190**

BERT (128 sequence length). Dataset is the Wikipedia/Books Corpus prepared from NVIDIA Deep Learning examples

At 16 nodes of p3dn.24xl (128 V100 GPUs), we achieve ~96% scaling efficiency

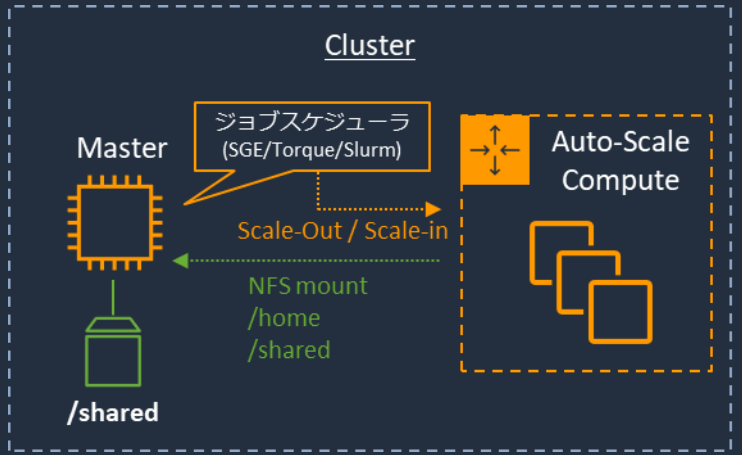
For more, visit https://github.com/aws-samples/eks-efa-examples

AWS ParallelCluster とは

ジョブ投入に応じて自動でスケールするクラスタを
AWS 上に構築可能な AWS 公式のオープンソースソフトウェア

AWS ParallelCluster の特徴

- 既存のHPC向けジョブスケジューラと Auto-Scaling を連携した環境を作成
  Slurm / SGE / Torque ※に対応
- 少しのコマンド操作でクラスタ作成可能
- MPI/NCCL 環境がセットアップ済みで、すぐに利用可能
- 使用するOSやネットワーク環境、ストレージ構成などを柔軟にカスタマイズ可能
- オープンソースソースのプロジェクトであり、誰でもソースコードを入手可能
  https://github.com/aws/aws-parallelcluster

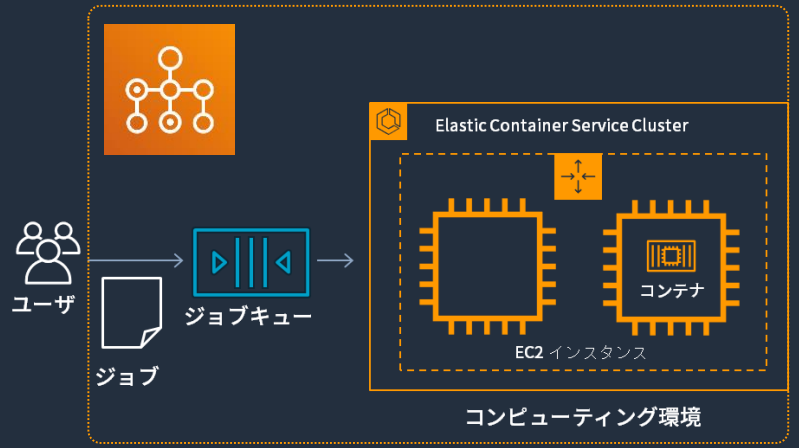


※ 将来的にSGE/Torqueについてはサポートの終了がアナウンスされており、Slurmの利用を推奨

AWS Batch とは

大規模バッチ処理のため環境をフルマネージドで提供

- AWS Batch がインスタンスの起動や停止を行うため、スケジューラや計算ノードなどの **管理が不要**
- ジョブは **Docker コンテナイメージ** を元に作成し、自動でスケールするコンピューティング環境で実行する
- コンピューティング環境ではインスタンスタイプや vCPU 数、スポットインスタンス利用有無などを任意に指定可能



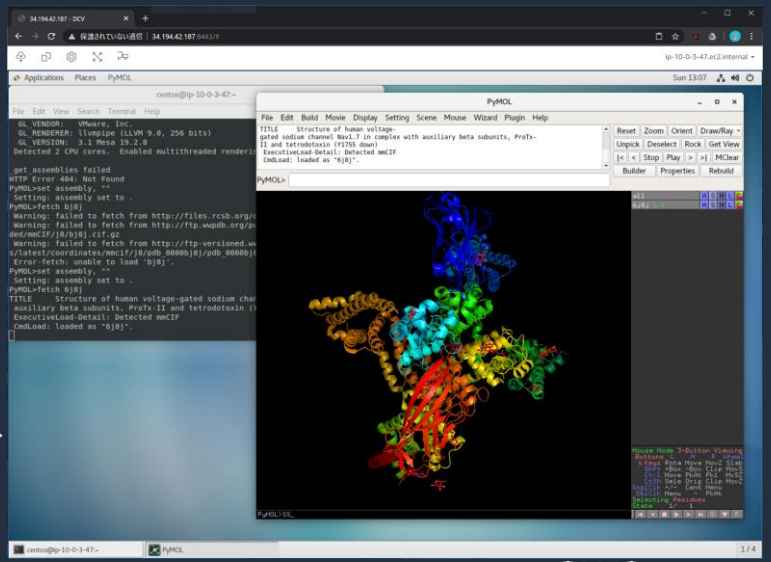
コンテナイメージを用意するだけでスケーラブルな大規模バッチ処理環境が得られる

NICE-DCV とは

クラウド上のデスクトップ画面をストリーミングするためのソフトウェア

特徴

- 専用のプロトコルによる高速かつスムーズなストリーミング
- サーバはWindows、Linuxの両方に対応
- GPUにも対応し、G4dn/G4adインスタンス等を利用することでより高速な描画が可能
- ネイティブクライアントの他、HTML5対応ブラウザからも利用可能



コスト

- Amazon EC2で利用する場合はライセンスコスト無しで利用可能 (EC2以外での利用は有償)

[AWS Black Belt Online Seminar] HPC on AWS

AWS の HPC 関連サービスについて、より詳しく知りたい方は：

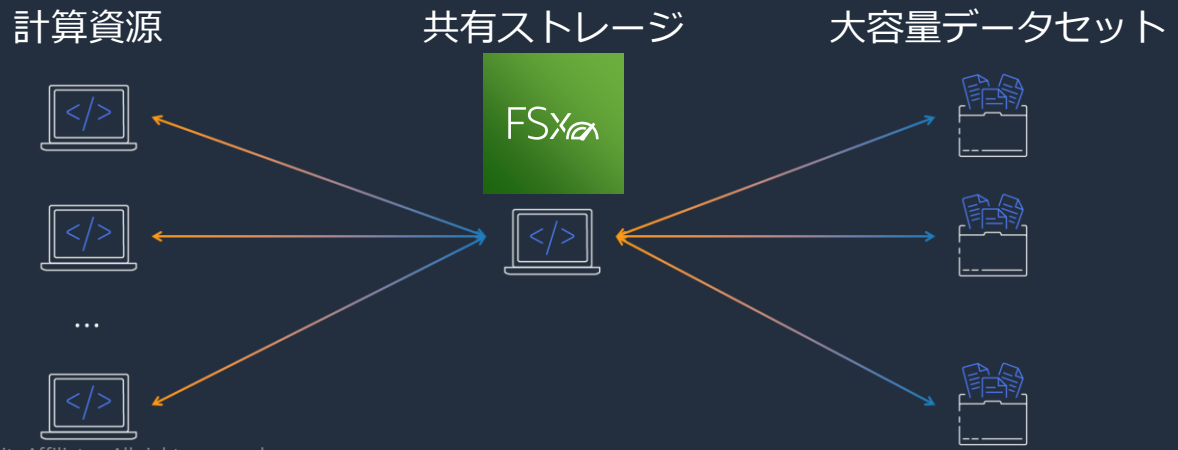
https://aws.amazon.com/jp/blogs/news/webinar-bb-hpconaws-2020/

Amazon FSx for Lustre のご紹介と クラウド HPC ストレージの考え方

Amazon FSx for Lustre とは

高速な分散ファイルシステムである Lustre をフルマネージドで提供

- 数クリックでファイルシステムを作成でき、サーバを直接管理する必要がない
- 容量に応じて高いパフォーマンスを提供
- 階層化ストレージ機能により Amazon S3 と連携可能

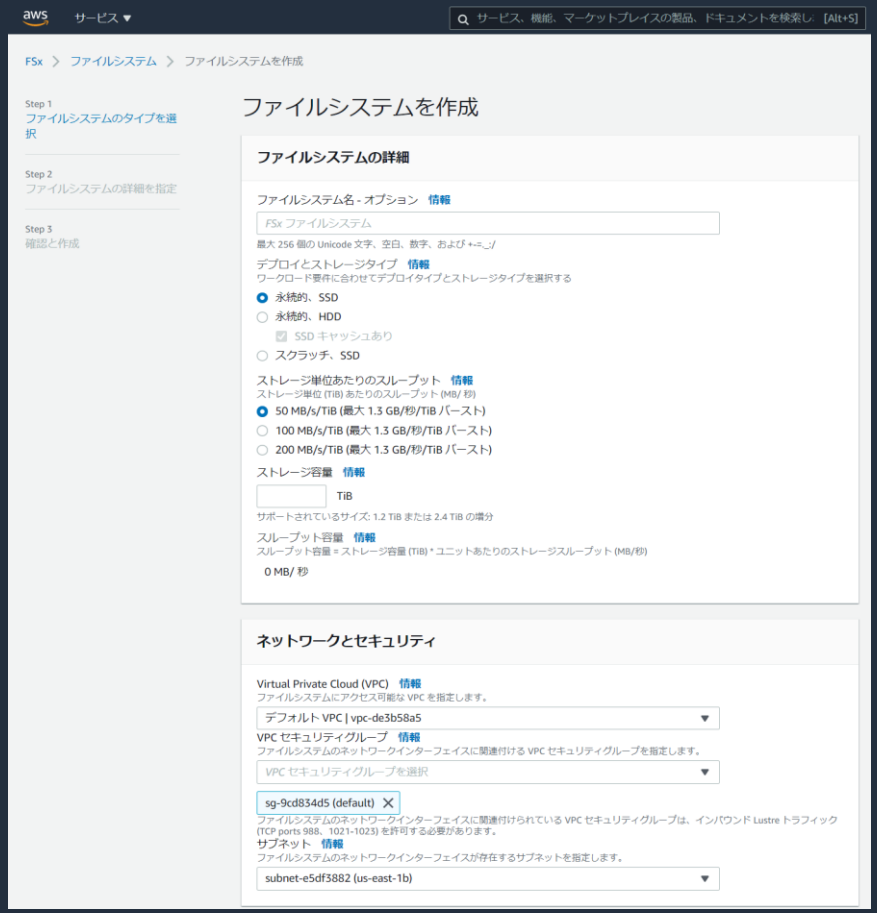


FSx for Lustre の作成方法

- 容量やストレージタイプを指定
- ファイルシステムを作成するVPC/サブネットを指定

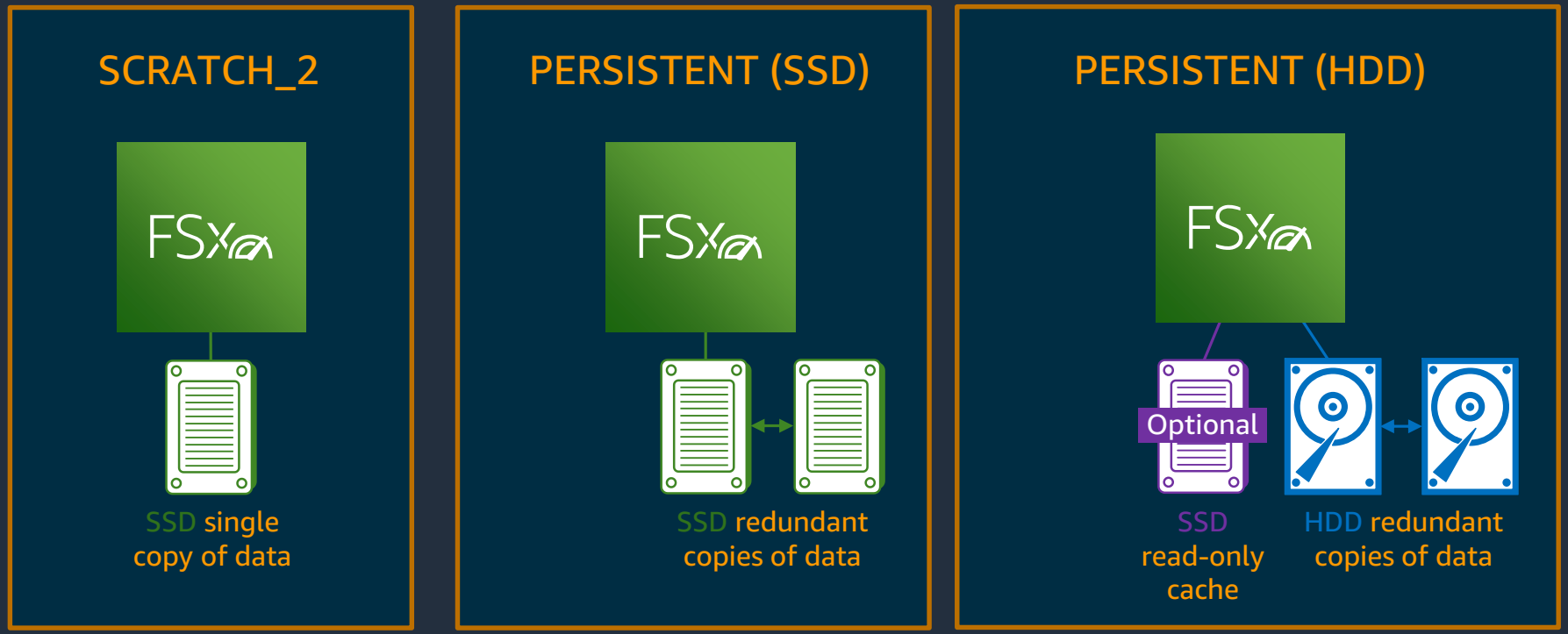
ファイルシステムを作成

数分で Lustre ファイルシステムが作成完了



ストレージタイプ

一時領域向けの SCRATCH_2、永続領域として利用可能な
PERSISTENT (SSD or HDD) といった多様なバリエーションから選択可能

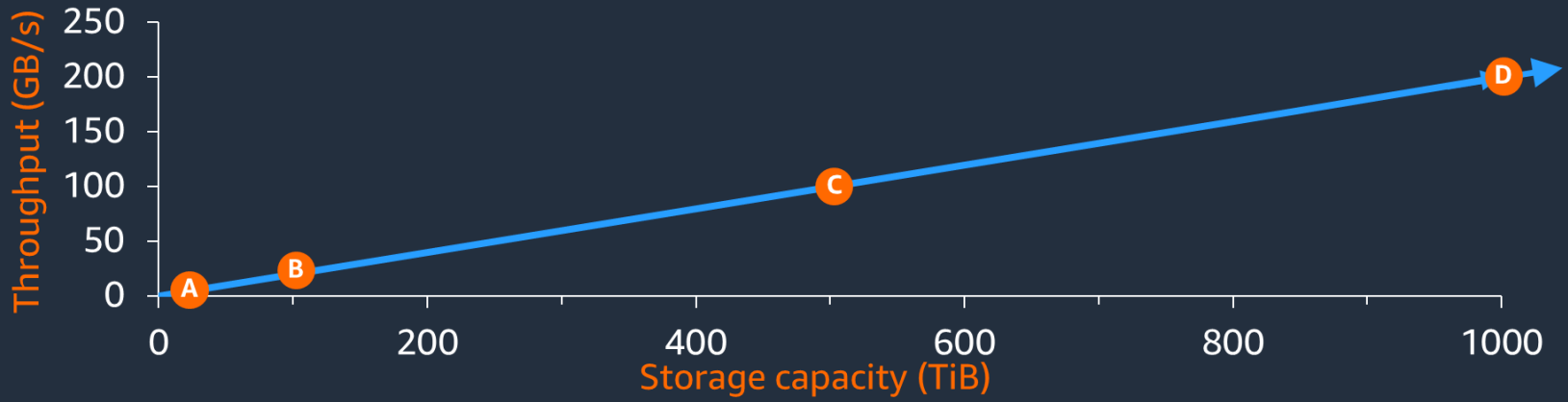


aws

パフォーマンス特性

プロビジョン（使用する量の設定）した容量に応じてパフォーマンスが変化

SCRATCH_2 の場合、1 TiB あたり、 200 MB/s のパフォーマンスを提供

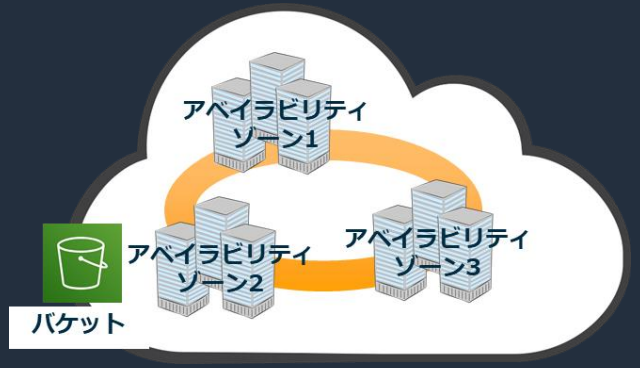


	Storage capacity	Baseline throughput	Burst throughput	IOPS	Latencies
A	10 TiB	2 GB/s	13 GB/s	Tens of thousands	Submillisecond
B	100 TiB	20 GB/s	130 GB/s	Hundreds of thousands	Submillisecond
C	500 TiB	100 GB/s	650 GB/s	>1 million	Submillisecond
D	1,000 TiB	200 GB/s	1,300 GB/s	Millions	Submillisecond

参考: Amazon S3

AWS の提供するオブジェクトストレージサービス

- **容量無制限**
  (1オブジェクトは最大5TBまで)
- データを3つ以上の AZ (データセンタ群) に保管し 99.999999999% という**高い耐久性**
- **低コスト**
  Standard: 0.023 USD/GB※
  ～ S3 Glacier Deep Archive: 0.00099 USD/GB ※
- **スケーラブルで安定した性能**
  データ容量に性能が依存しない
  （ユーザが、サーバ台数、媒体本数やRAID、RAIDコントローラを考える必要がない）



※2021年3月現在のus-east-1での価格
価格はストレージクラスによって異なる
https://aws.amazon.com/jp/s3/pricing/

大規模なデータを扱う時の要望

データ保管時は S3 に格納してコストを抑えつつ
処理を行う時だけ高速なファイルストレージとしてアクセスしたい

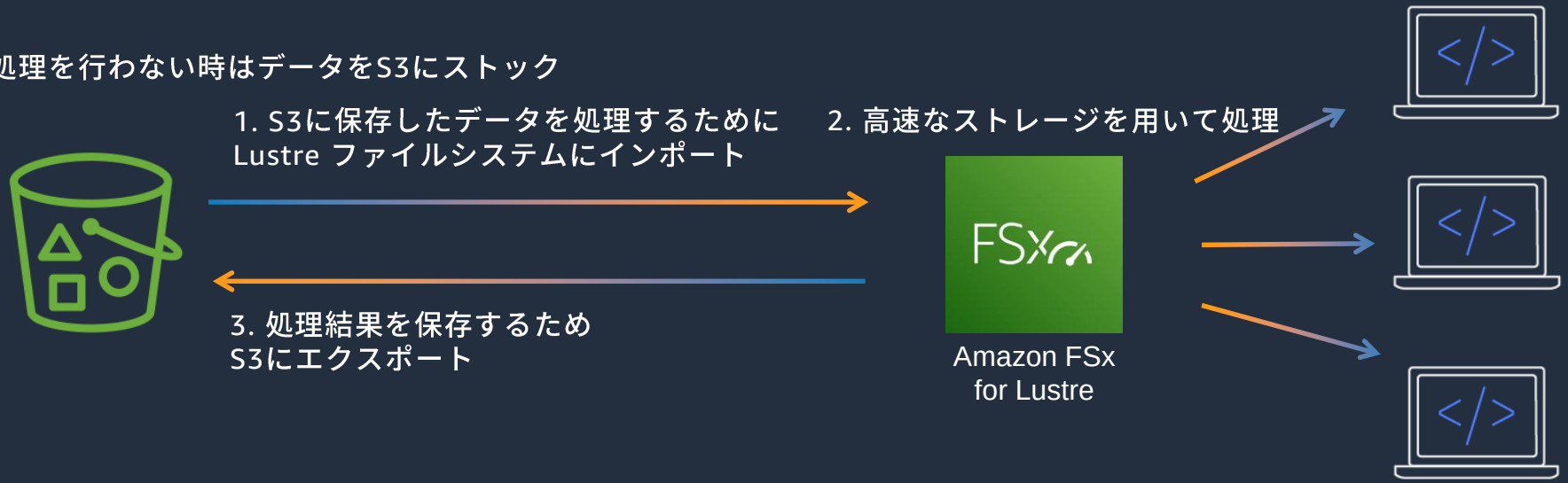
↓

FSx for Lustre の S3 連携機能を活用

S3 とのシームレスな統合

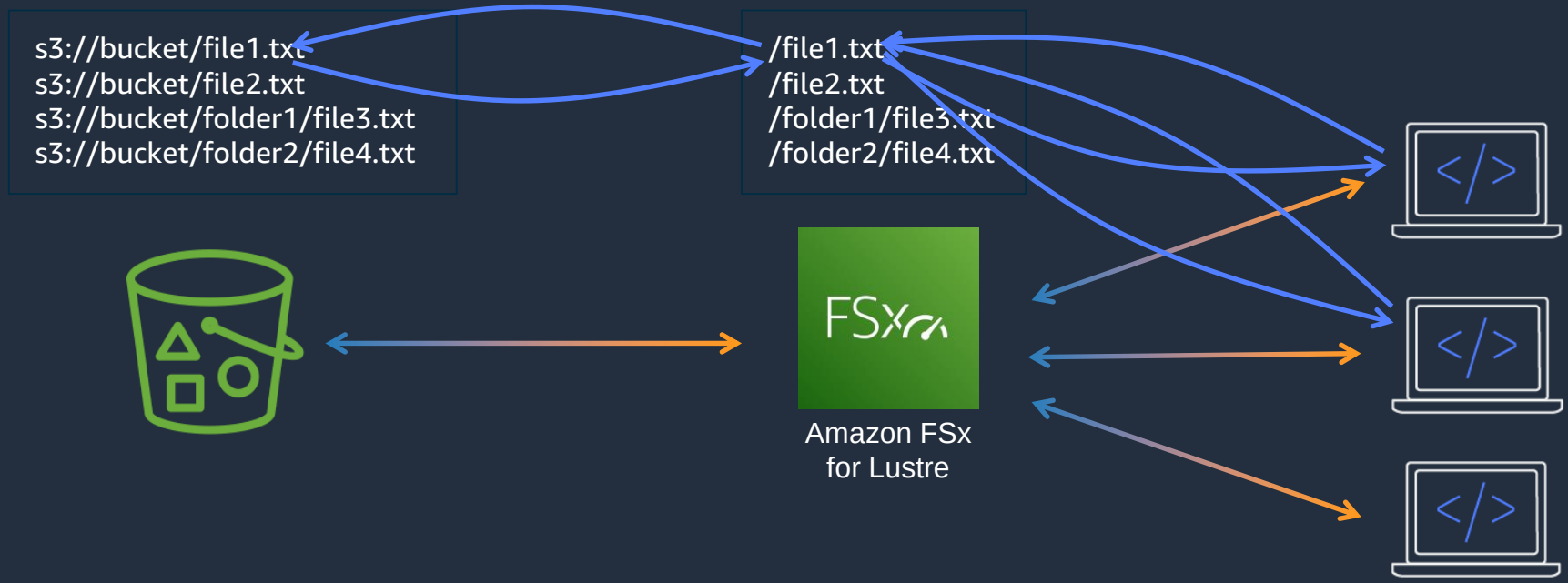
Amazon S3のデータセットと Amazon FSx for Lustre ファイルシステムを関連付け
実際に処理を行う時にのみ FSx for Lustreを使用する

0. 処理を行わない時はデータをS3にストック



従量課金のため、処理が終了したらファイルシステムを削除することで
FSx for Lustre の課金は停止する

Amazon S3 への透過的アクセス



- ファイルシステム作成時にS3 bucketを関連付け、ファイルのインデックスが作成される
- 各ファイルに初回アクセスがあった時点でデータがS3からファイルシステムに自動でインポートされる（インポートのための若干のレイテンシが発生）
- 二度目のアクセスからは高速なレスポンスが可能

S3 への透過的アクセスの確認 (import)

lfs hsm_state コマンドによりファイルの状態を確認可能
（HSM: Hierarchical Storage Management）

```
$ ls
README.txt
$ lfs hsm_state README.txt
README.txt: (0x0000000d) released exists archived, archive_id:1
$ cat README.txt
Hello from FSx
$ lfs hsm_state README.txt
README.txt: (0x00000009) exists archived, archive_id:1
```

この例ではファイルにアクセスすることで Lustre にファイル本体が格納される
（ released フラグが消える）

https://docs.aws.amazon.com/fsx/latest/LustreGuide/fsx-data-repositories.html

S3 への透過的アクセスの確認 (export)

自動的に S3 への export を行う設定（非同期）に加え、
明示的に export を行うことも可能

```
$ echo Hello from EC2 >> README.txt
$ lfs hsm_state README.txt
README.txt: (0x0000000b) exists dirty archived, archive_id:1
$ lfs hsm_archive README.txt
$ lfs hsm_state README.txt
README.txt: (0x00000009) exists archived, archive_id:1
```

S3に格納されていたファイルに追記を行うことで dirty フラグが立つ
lfs hsm_archive を実行することで、ファイルはS3に export され、dirty フラグは消える

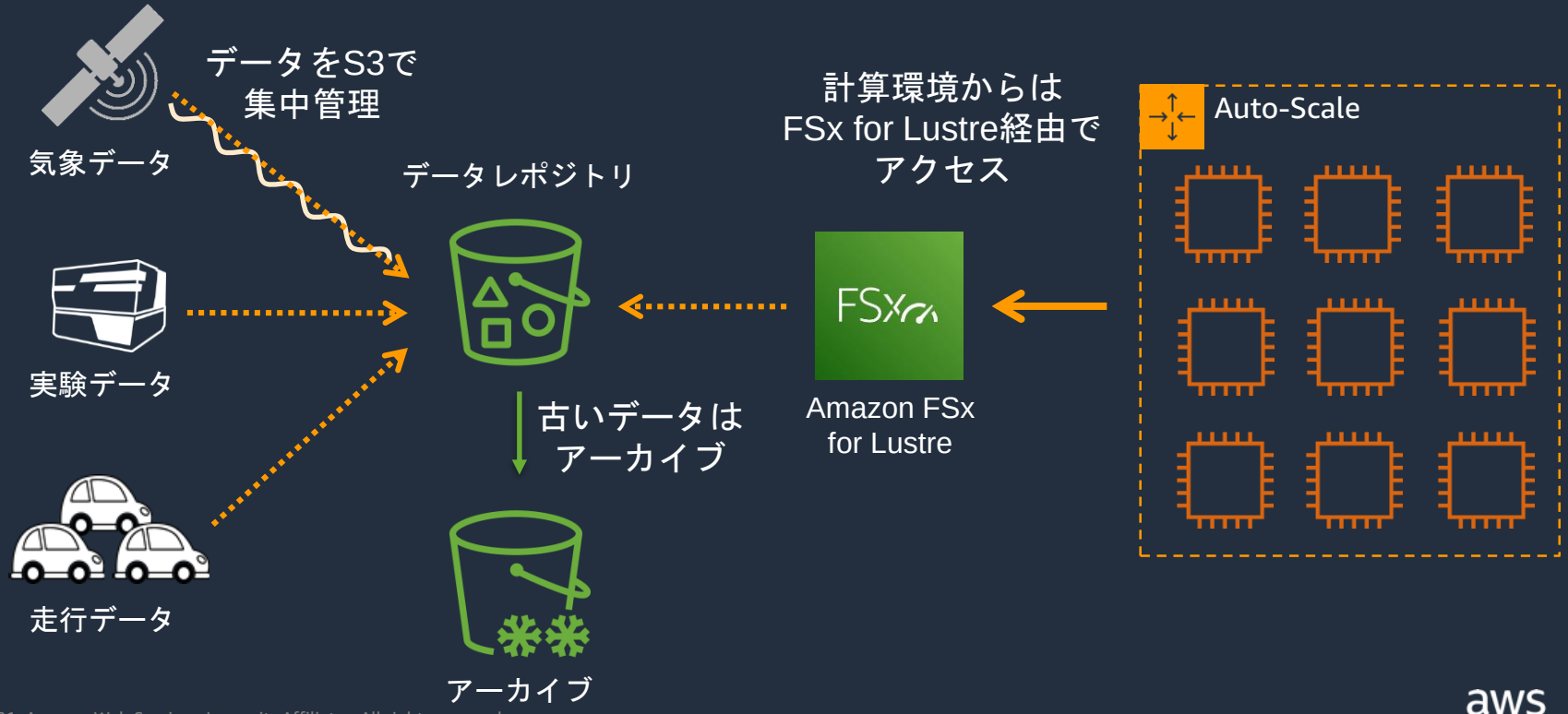
https://docs.aws.amazon.com/fsx/latest/LustreGuide/fsx-data-repositories.html

その他の便利な機能

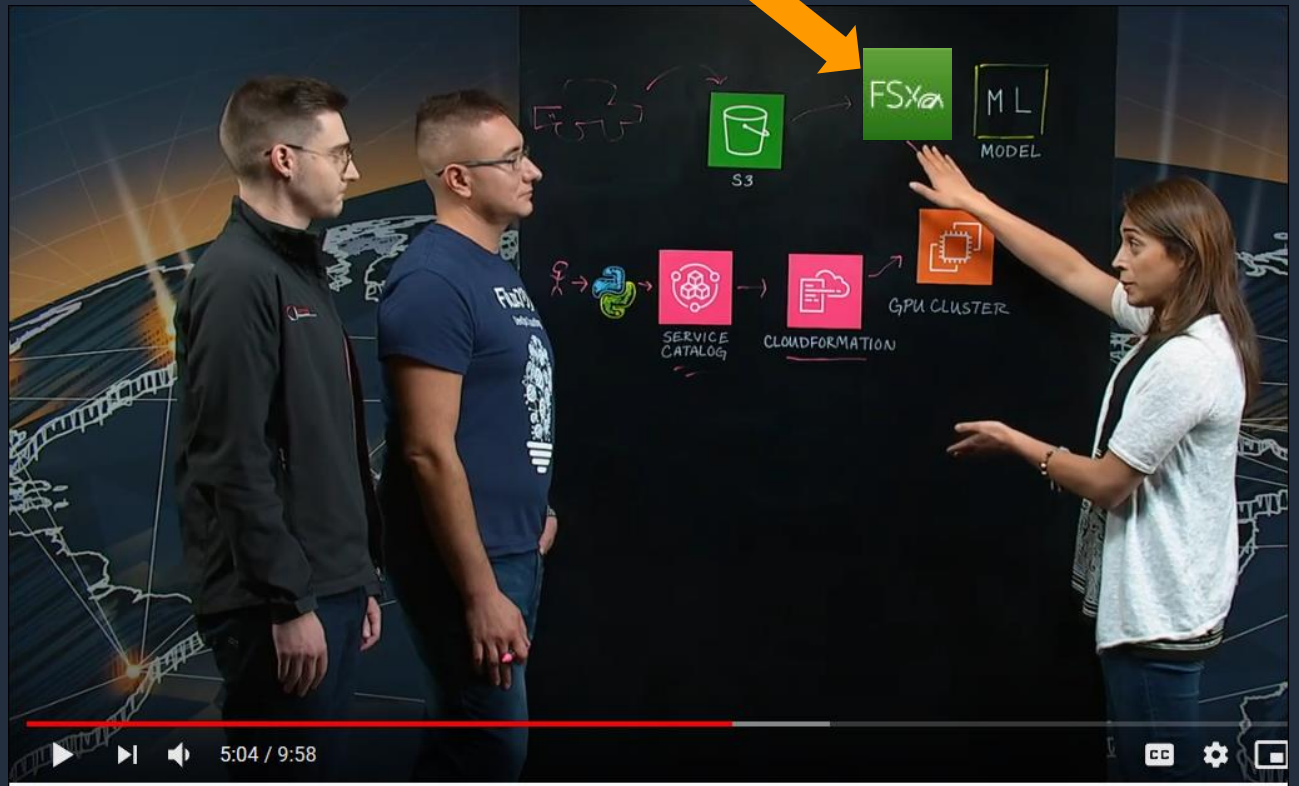
- **オンラインでの容量拡張**
- 1 分単位のモニタリング（スループット・IOPS）
- スナップショット作成（PERSISTENT タイプ）
- ユーザ・グループごとの Quota 設定 （ `lfs setquota` ）
- CSI ドライバの提供により Kubernetes 環境からの制御（作成・削除等）が可能

FSx for Lustre + S3 によるデザインパターン

S3を中心としつつ、FSx for Lustreの階層化ストレージ機能を活用することで
コスト効率よく大規模のデータを保管・解析が可能となる



Toyota Research Institute: Autonomous Vehicles



Toyota Research Institute: On-Demand Self-Service Portal for Data Scientists to Process Data Sets

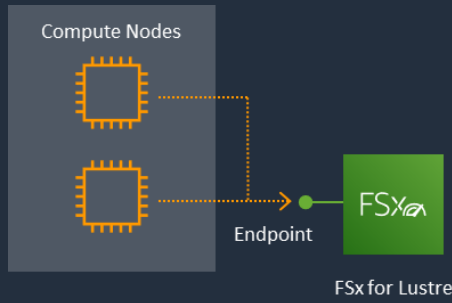
[YouTube Link]

用途に応じたストレージ構成パターン

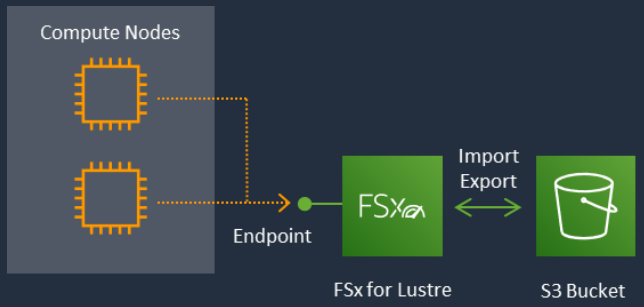
計算ノードでのデータの利用パターン等によって選択

- **FSx for Lustre**: 高速かつスケーラブルな共有ディスクが必要な場合に選択
- **FSx for Lustre + S3**: 大容量データセットのうちの一部を使用し、同一のデータが複数のノードで使用される場合に選択
- **S3 to local disk**: 大容量データセットのうちの一部を使用し、計算ノード間でデータを共有する必要がない場合に選択

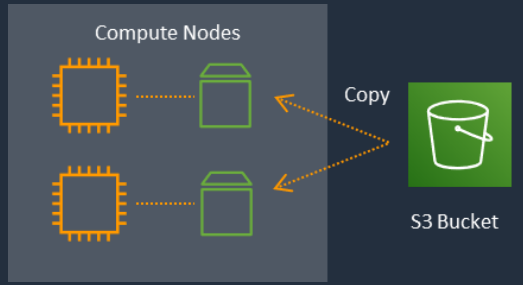
FSx for Lustre



FSx for Lustre + S3



S3 to local disk



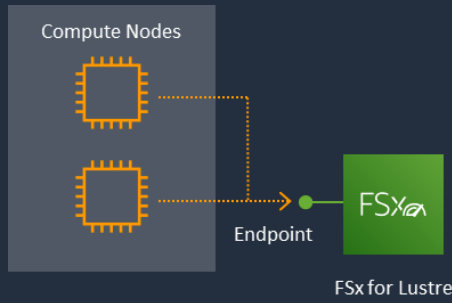
用途に応じたストレージ構成パターン

計算ノードでのデータの利用パターン等によって選択

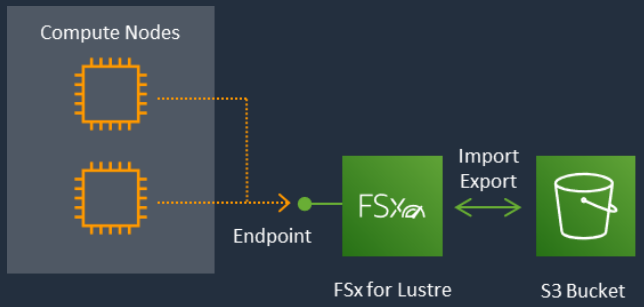
- FS
- FS
- S3

どのパターンが適しているかは
行いたい処理のワークフローやデータが決まらないと
判断できない

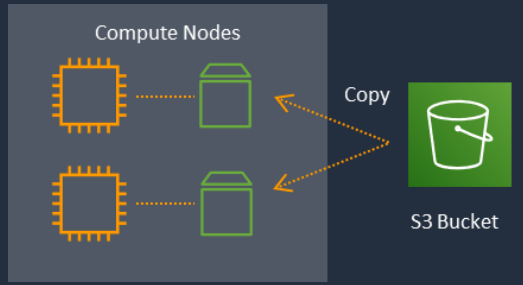
FSx for Lustre



FSx for Lustre + S3



S3 to local disk



クラウド HPC を活用する際の考慮事項

クラウドの持つ「必要な時に」「必要なリソース」を「必要な量だけ」確保できるという特性を HPC に活用したい

- ワークロードに合わせて様々な構成を設計でき、自由度が高いぶん、ワークロードへの理解も重要
  （そもそもワークロードが決まっていないと判断ができない）
- クラウド・HPC 両方の観点で考慮が必要
  - **そのワークロードはどういうデータフロー・計算特性なのか**
  - **どのようなクラウドインフラ・サービスが適しているか**

現実には、、

- 最高ではなく、現実的なコスト・パフォーマンスで妥協することも重要
- 実際のワークロードで**すぐに試せる**のはクラウドのメリット

まとめ

クラウド HPC の考え方

- クラウドの持つ 「必要な時に」「必要なリソース」を「必要な量だけ」確保できるという特性を HPC に活用
- 構成の自由度が高いため、クラウド・HPC両方の観点で考慮が必要
    - **そのワークロードはどういうデータフロー・計算特性なのか**
    - **どのようなクラウドインフラ・サービスが適しているか**

HPC on AWS の利点・特徴・関連サービス

- 多様な EC2 の種類や EFA による高速・低レイテンシなMPI通信
- FSx for Lustre による高速な分散ストレージ + S3 との階層化機能による大容量データセットの取り扱いが可能

クラウドを活用した HPC の形について一緒に考えていければ幸いです！

AWS Educate

教育機関向けにAWSの学習環境を無償で提供

- **AWS 利用環境**
    - 機関校加盟済みの場合：教員 $200/年、学生 $100/年
    - 非加盟校の場合：教員 $75/年、学生 $30/年
    - 別途授業やゼミ向けの追加クレジットを教員がリクエスト可能
- クレジットカード登録が不要で、AWSが使用できるスターターアカウントも提供
    - 一部サービスの利用に制限あり
- 20以上の自習用オンラインコースも提供
- 100万人以上が加盟（2020年10月現在）

Webより申し込み可能
是非、機関校加盟もご検討ください
http://www.awseducate.com

Appendix

各サービスについてより詳しく知りたい方は

Black Belt オンラインセミナー

HPC on AWS

- https://aws.amazon.com/jp/blogs/news/webinar-bb-hpconaws-2020/

AWS ParallelCluster ではじめるクラウドHPC

- https://aws.amazon.com/jp/blogs/news/webinar-bb-aws-parallelcluster-cloudhpc-2020/

AWS Batch

- https://aws.amazon.com/jp/blogs/news/webinar-bb-aws-batch-2019/

Amazon FSx for Lustre

- https://aws.amazon.com/jp/blogs/news/webinar-fsx-title-2019/

Amazon EC2

- https://aws.amazon.com/jp/blogs/news/webinar-bb-amazon-ec2-2019/

Amazon EC2 Deep Dive: AWS Graviton2 Arm CPU搭載インスタンス

- https://aws.amazon.com/jp/blogs/news/webinar-bb-aws-graviton2-2020/

HPC on AWS ベンチマーク例

密結合ワークロードでのEFA活用例

計算流体力学

- Solvers - OpenFOAM, Fluent, Star-CCM+, LS-DYNA, OVERFLOW, FUN3D

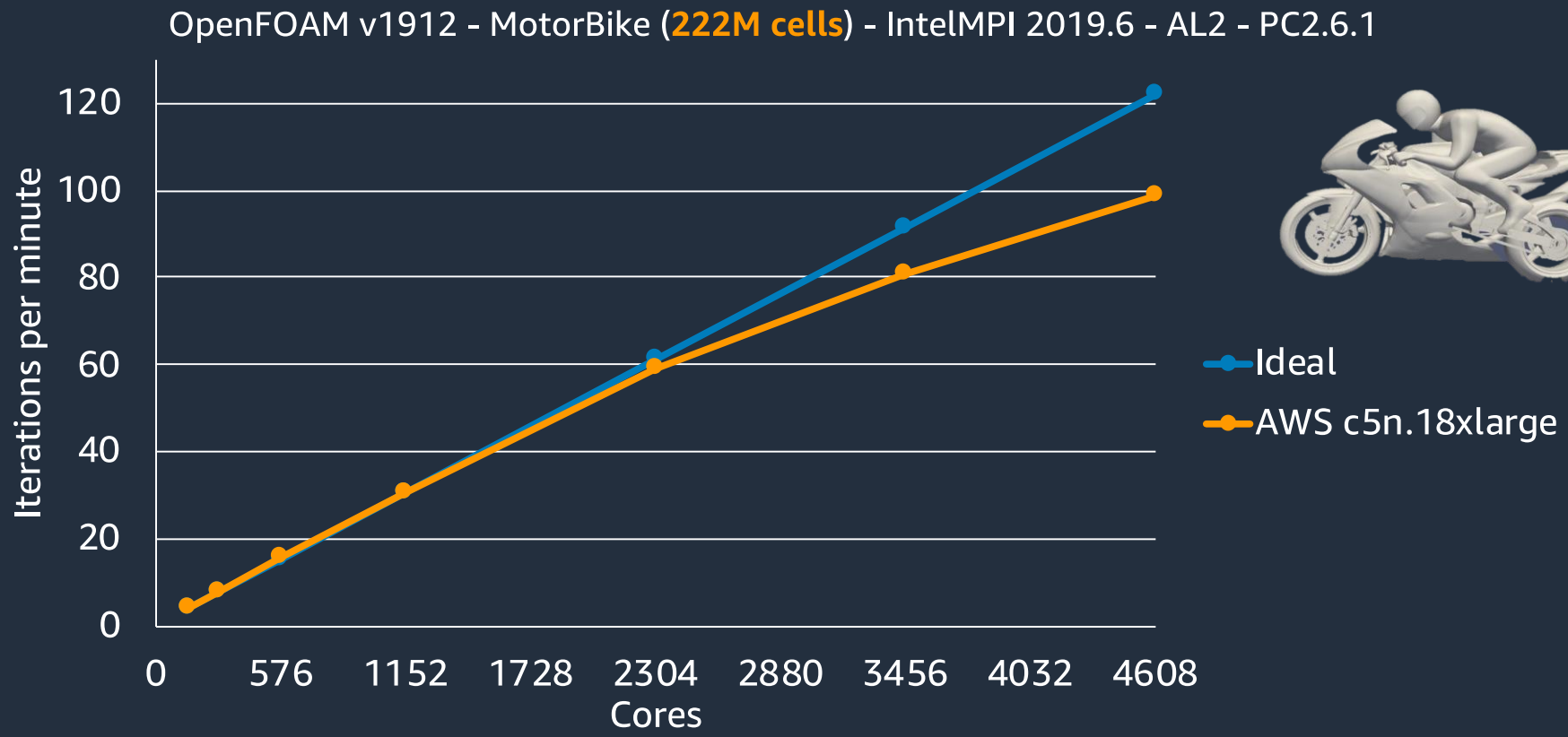
数値気象予測

- Weather Models - WRF, FV3, MPAAS

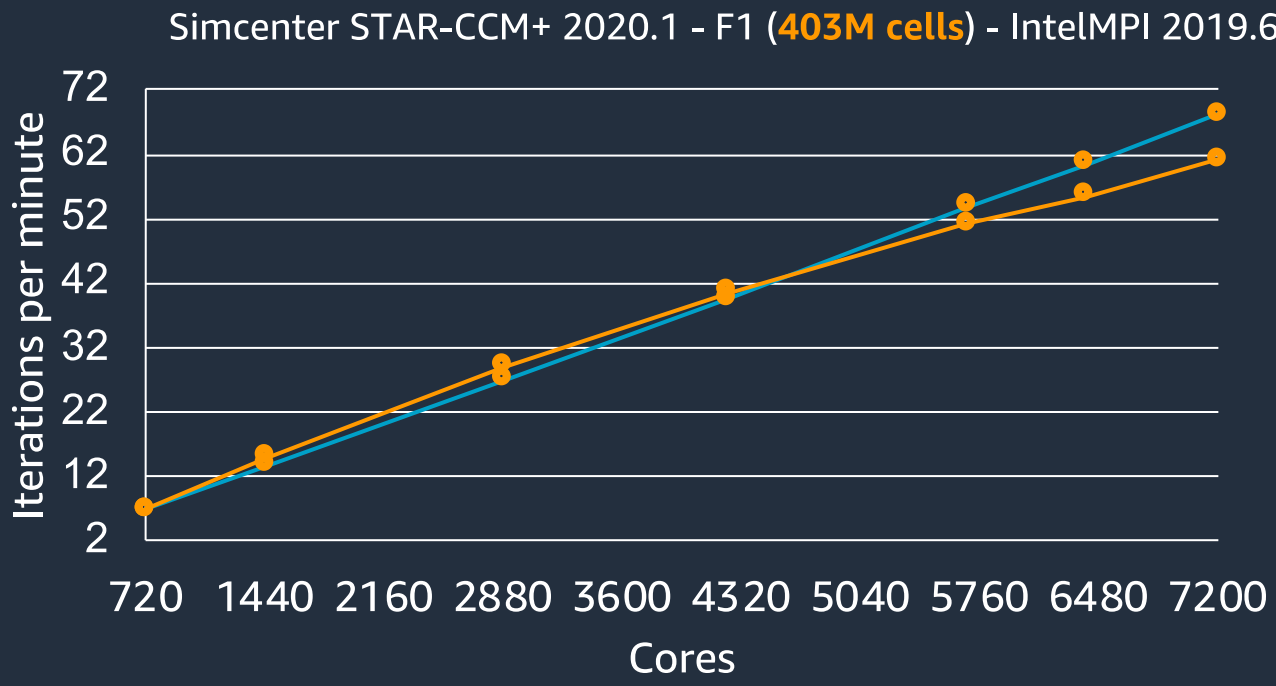
分散機械学習

- ML Models - BERT, Seq-2-seq

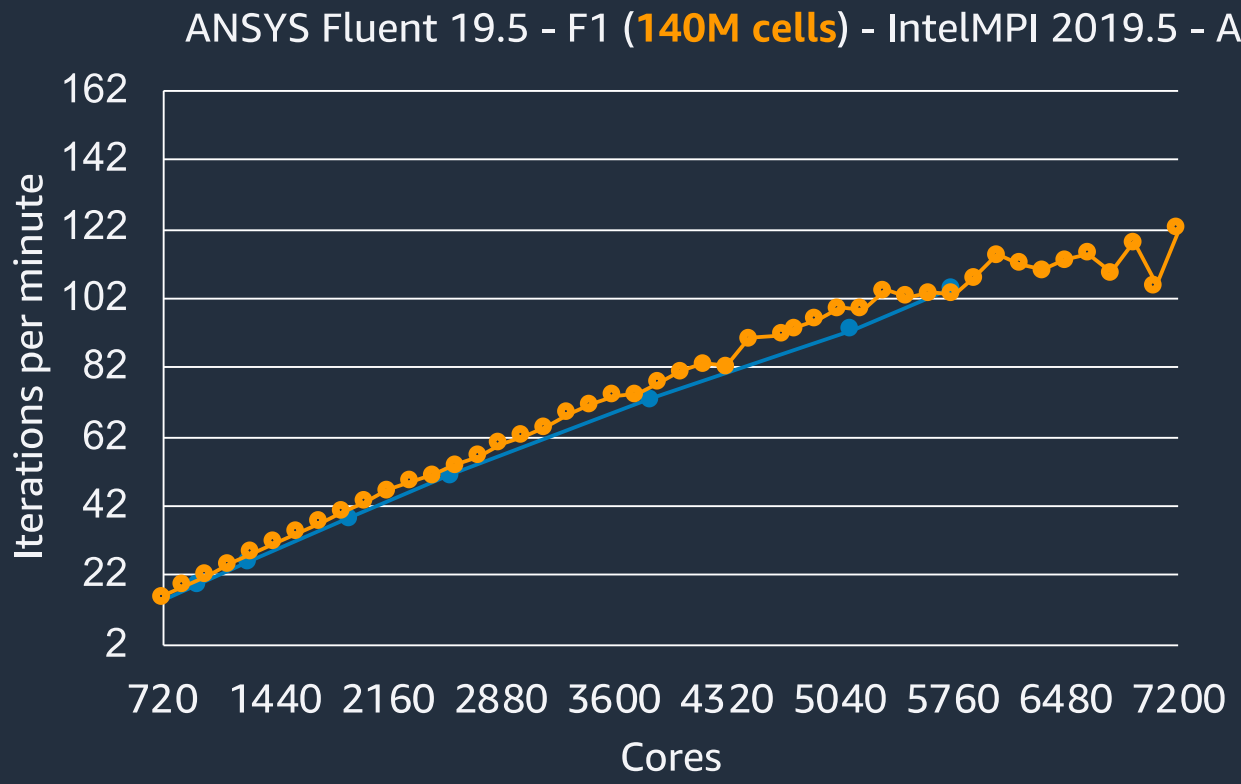
Scaling on AWS - OpenFOAM



Scaling on AWS – STAR-CCM+

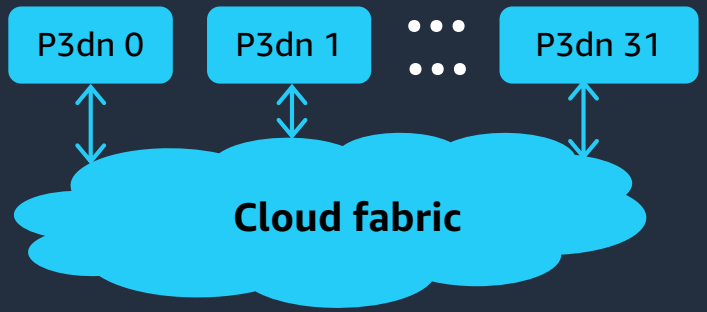


Scaling on AWS – Fluent



Facebook AI Research (FAIR) & EFA

32 * P3dn nodes, NVLINK, all-reduce via NCCL		
256 V100 GPUs	**AWS ENA**	**AWS EFA**
BERT (NLP)	1.0x	1.5x
TDS-Seq2Seq (ASR)	1.0x	1.3x



ML Training is sensitive to interconnect

- All-reduce is most common collective
- Large messages O(1-10MB) are common

P3dn with EFA provides up to 100 Gbps Model parallelism

- Different communication patterns
- Needs high bisection bandwidth

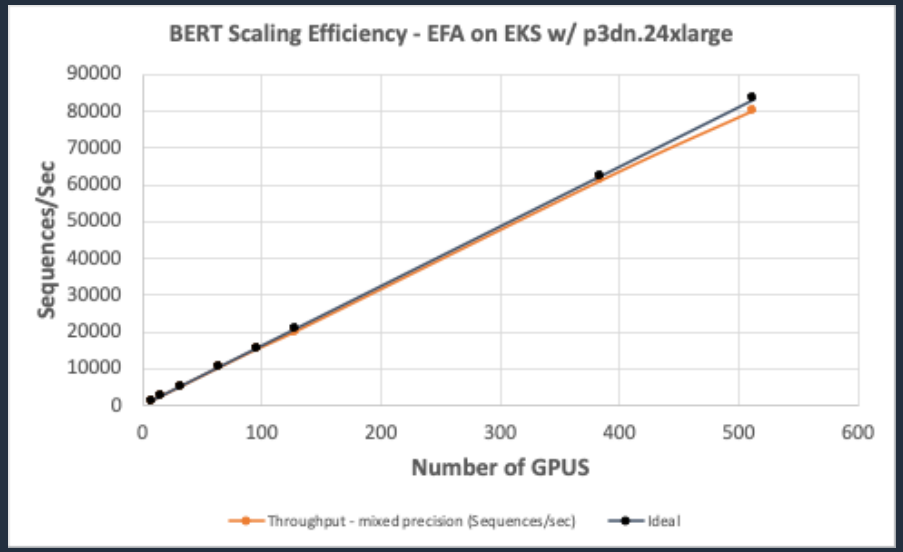
New emerging workloads: graph learning

ENA - Elastic Network Adapter

EFA - Elastic Fabric Adapter

aws

参考：Elastic Kubernetes Service (EKS) & EFA



Number of Nodes	Number of GPUS	Batch Size / GPU	Max Accumulation Steps	Sequence length	Throughput - mixed precision (Sequences/sec) - EFA
1	8	64	128	128	**1303**
2	16	64	128	128	**2561**
4	32	64	128	128	**5114**
8	64	64	128	128	**10254**
12	96	64	128	128	**15315**
16	128	64	128	128	**20053**
48	384	64	256	128	**61161**
64	512	64	256	128	**80190**

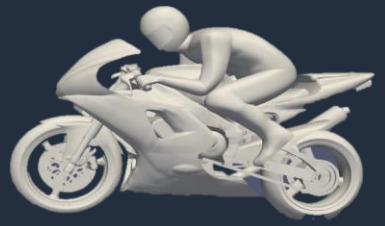
BERT (128 sequence length). Dataset is the Wikipedia/Books Corpus prepared from NVIDIA Deep Learning examples

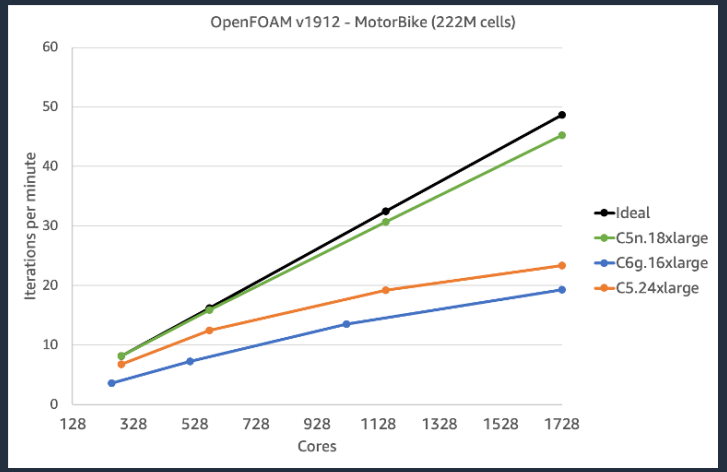
At 16 nodes of p3dn.24xl (128 V100 GPUs), we achieve ~96% scaling efficiency

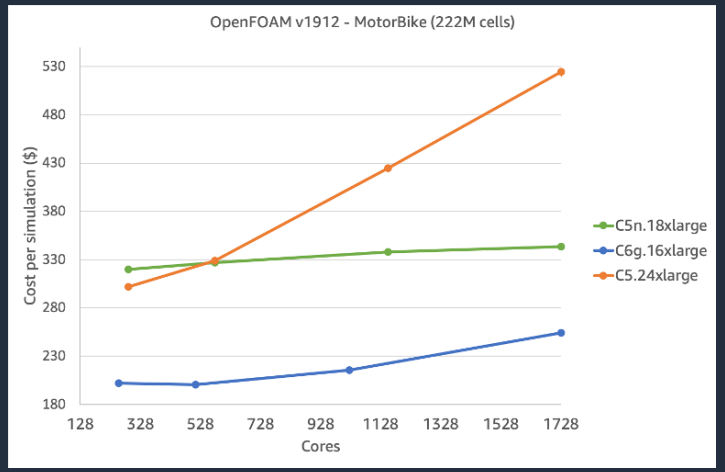
For more, visit https://github.com/aws-samples/eks-efa-examples

C6g による OpenFOAM Benchmarks

- 222M cell Motorbike モデルを用いて評価
- 計算性能としては、EFAを搭載したC5nが最も高い
- シミュレーションあたりのコストでは、C6g が最も安価







https://aws.amazon.com/blogs/compute/c6g-openfoam-better-price-performance/
https://gitlab.com/arm-hpc/packages/-/wikis/packages/openfoam

Arm 公式 Blog での HPC 利用例

- Evaluation of the NEMO Ocean Model on Arm Neoverse-based AWS Graviton2
    - https://community.arm.com/developer/tools-software/hpc/b/hpc-blog/posts/evaluation-of-the-nemo-ocean-model-on-aws-graviton2
- Demonstration of low mach-number CFD modeling with Nalu on AWS Graviton2 M6g instances
    - https://community.arm.com/developer/tools-software/hpc/b/hpc-blog/posts/low-mach-number-cfd-modeling-with-nalu-on-graviton2-aws-m6g
- Seismic Modeling with Arm Neoverse N1 and AWS Graviton2
    - https://community.arm.com/developer/tools-software/hpc/b/hpc-blog/posts/seismic-modeling-with-arm-neoverse-n1-and-aws-graviton2

HPC on AWS

aws