AWSのゲノミクスソリューション

鳥羽 祐輔

アマゾンウェブサービスジャパン合同会社
ソリューションアーキテクト

アジェンダ

1. AWS クラウド活用によって得られる価値
   A. AWS とは
   B. HPC 環境の課題とクラウド活用による解決
   C. 活用事例: 国立がん研究センター様、第一三共株式会社様
2. ゲノム領域における AWS ソリューション
   A. ストレージとデータ転送
   B. AWS HealthOmics

アジェンダ

1. AWS クラウド活用によって得られる価値
   A. AWS とは
   B. HPC 環境の課題とクラウド活用による解決
   C. 活用事例: 国立がん研究センター様、第一三共株式会社様
2. ゲノム領域における AWS ソリューション
   A. ストレージとデータ転送
   B. AWS HealthOmics

Our mission

Amazon は、

地球上で最もお客様を大切にする企業、
そして地球上で最高の雇用主となり、
地球上で最も安全な職場を提供すること

を目指しています。

AWS とは

2006 年より、他社にさきがけてクラウドサービスを提供

245 の国と地域、世界数百万、日本では数十万以上のお客様

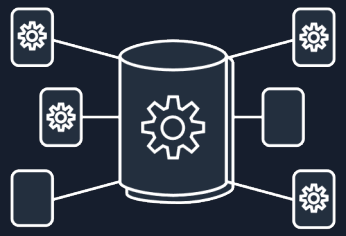
全国をカバーする**パートナーコミュニティ**

累計で **134 回以上の値下げ**をして利益をお客様へ還元

※ お客様とはアクティブカスタマー数を指します。アクティブカスタマーとは、AWS クラウド無料利用枠を含む AWS アカウントの先月の使用状況のあるアマゾン会員でない対象アカウントです。

AWS クラウド活用の真価 = お客様が価値提供に集中できること

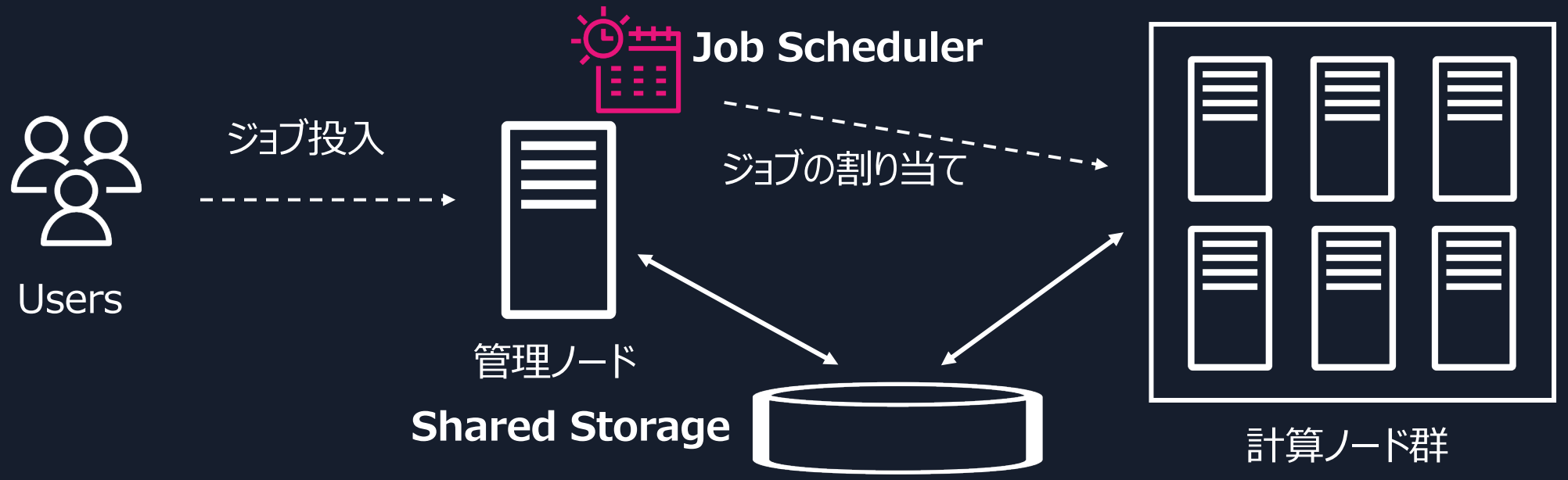
必要なときに
必要なだけ調達可能
な高い柔軟性と可用
性を併せ持つITリソー
ス



マネージド型
サービスの活用で
新たな技術革新を
迅速に適用可能

サービスとして提供される
セキュリティ機能の実装や
厳格なコンプライアンス
要件への対応

HPC環境の典型的な構成要素 & 課題



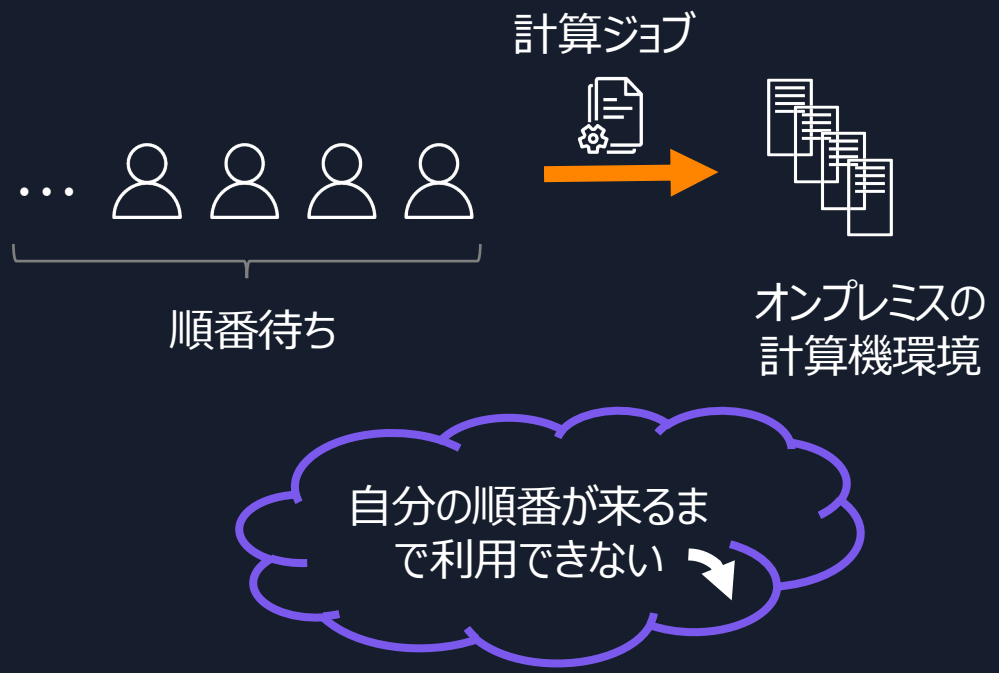
多くの研究機関等で共有のHPC環境（クラスターコンピュータ・スパコン）を整備

共有HPC環境の課題

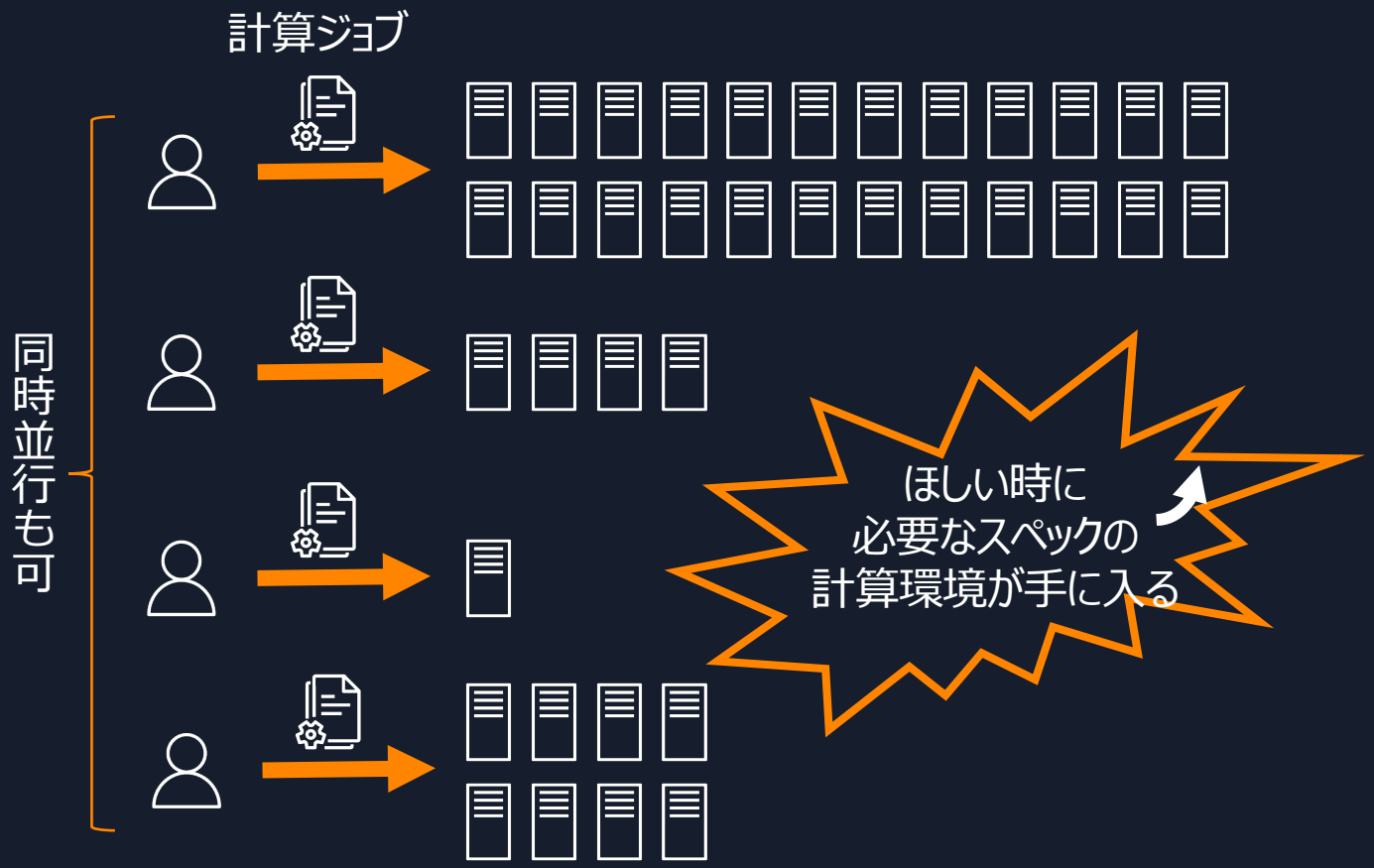
- ジョブ待ちが頻繁に起こる
- OSやライブラリのバージョンが固定。ソフトウェアが動かない場合 対処が困難
- オンプレミスは調達に時間が掛かり、手続きが煩雑。メンテナンスなどの費用負担が大きい

“必要な時に” “必要な計算機環境を” 手に入れる

限られた環境しかない場合

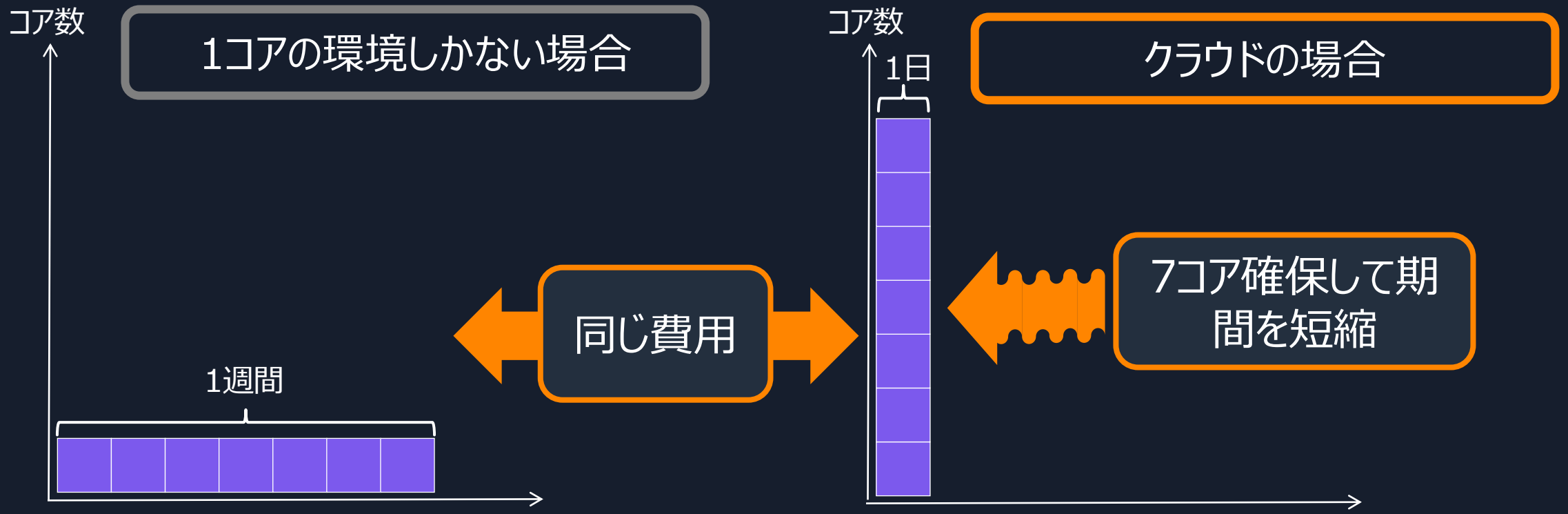


AWS クラウドを利用した場合



クラウドの “拡張性” を活かした計算時間の短縮

例) 1コアで1日かかる計算  を 7 つ実施したい



200 を超えるサービス マネージドサービス

AWS の提供する 90%+ のサービスや機能はお客様からの意見をもとに開発、残りもお客様の潜在的な要望を汲み取って作られています


コンピューティング


モバイル


ARとVR


エンドユーザーコンピューティング


ストレージ


データベース


ネットワークとコンテンツ配信


AWS コスト管理


機械学習


IoT


ロボット工学


ビジネスアプリケーション


メディアサービス


分析


マネジメントとガバナンス


開発者用ツール


アプリケーション統合


Game Tech

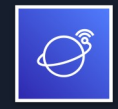

量子テクノロジー


カスタマーエンゲージメント


移行と転送


ブロックチェーン


セキュリティ、ID、コンプライアンス


人工衛星

医療・製薬業界に向けたAWSサービス

ヘルスケアおよびライフサイエンスのお客様に特化したサービス

AWS HealthOmics

ゲノムやトランスクリプトーム、その他のオミックスデータの保存と変換処理により、洞察を得るサービス

AWS HealthLake

医療情報(HL7 FHIR)を蓄積し、機械学習やBIツールからREST APIや使い慣れたSQLでデータ操作できる分析サービス

AWS HealthImaging

医用画像(DICOM)をペタバイト規模で保存、共有、分析できるストレージサービス

AWS HealthScribe

患者と医師の会話から話者を識別し、文字起こしと生成AIを用いた臨床ノートを自動生成するサービス

2024年1月現在、AWS Healthシリーズは
東京リージョン(ap-northeast-1)で未提供

AWS Summit Online 2021　国立がん研究センター

国立がん研究センター

目指すのは自律的な知識の獲得
がんゲノム解析で進むAWS活用

日本人の2人に1人がなるといわれる、がん。より効果的な治療方法を探る上で、不可欠なものとなっているのがクラウドテクノロジーだ。強力なコンピュートリソースや、データ分析機能を駆使することで、がんの発生要因となるゲノム変異を探索する。アマゾン ウェブ サービス（AWS）を活用し、そのための取り組みを展開する国立がん研究センターの活動と、ここまでの成果について紹介する。

先進テクノロジーを駆使し、世界中で進むゲノム解析

がんは日本人の国民病といわれる。これはゲノム変異によって生じることが分かっており、予防や治療の方法を探る上でゲノム解析がカギになるということも広く知られるようになってきた。これについて、国立がん研究センター 研究所の白石 友一氏は次のように語る。

「ゲノムとはDNAの文字列で表される遺伝情報のことで、いわば"細胞の設計図"です。生まれたときは、我々の体のどの細胞も同じゲノムを有していますが、年を重ねるにつれ、放射線やウイルスなどの影響による変異が蓄積してきます。この変異のほとんどは無害なのですが、ゲノムの特別な場所に入ってしまうと、

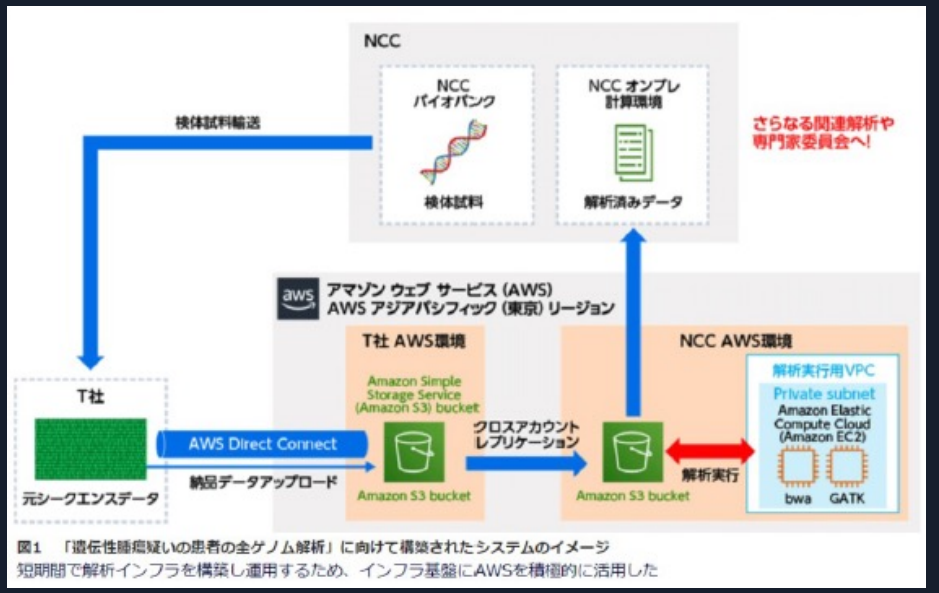


図1　「遺伝性腫瘍疑いの患者の全ゲノム解析」に向けて構築されたシステムのイメージ
短期間で解析インフラを構築し運用するため、インフラ基盤にAWSを積極的に活用した

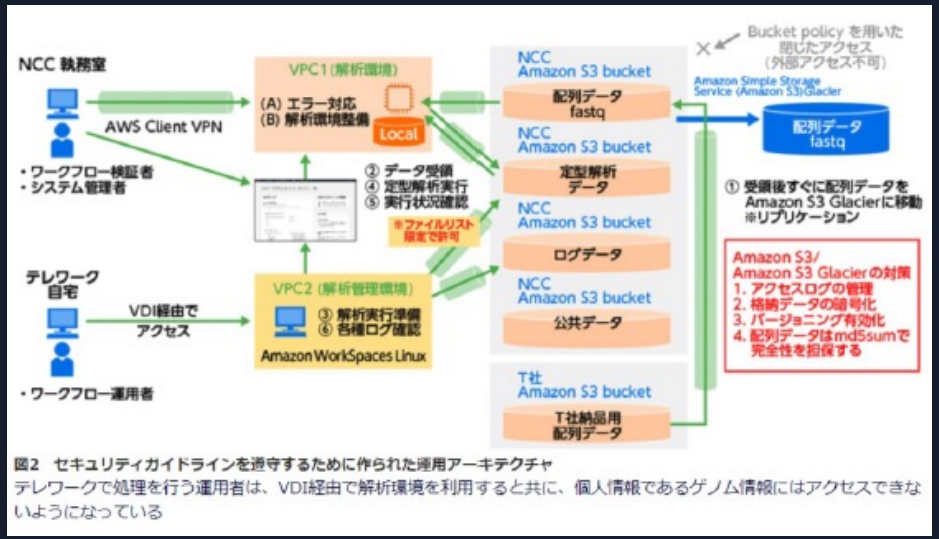


図2　セキュリティガイドラインを遵守するために作られた運用アーキテクチャ
テレワークで処理を行う運用者は、VDI経由で解析環境を利用すると共に、個人情報であるゲノム情報にはアクセスできないようになっている

資料: https://d1.awsstatic.com/events/jp/2021/summit-online/CUS-20_AWS_Summit_Online_2021_NCC.pdf
動画: https://www.youtube.com/watch?v=yOgEcIqv6MM

AWS Summit Online 2021　国立がん研究センター

高度なデータ解析を支える情報基盤を用意する為に掲げた要件

①短期間での構築・運用が可能であること

- ✓ 単年度の研究プロジェクト → 短期間でのインフラ構築が不可欠
- ✓ AWS マネージドサービスをフル活用して解析を実施
- ✓ 検体試料を外部の委託会社へ輸送 → ゲノムの塩基配列を決定するシークエンス →「AWS Direct Connect」経由で委託会社内にある「Amazon S3」へ格納 → クロスアカウントレプリケーションによって国立がん研究センターのAmazon S3へ転送 → ゲノム変異の網羅的な検出などの解析処理を実行

②低コストであること

- ✓ 大量のコンピューティングリソースが必要になる解析作業を安価に行う必要があった
- ✓ 各モジュールをDockerコンテナ上のバッチ処理として実装
- ✓ 1つのモジュールにおける処理が終わった段階で計算リソースを解放し モジュールの処理内容に応じて最適なインスタンスを利用 コスト削減を図った
- ✓ 各処理を短時間で集中的に実行することで スポットインスタンスを積極的に利用

③セキュリティガイドラインを遵守できること

- ✓ ゲノムデータは個人情報に当たり 解析を行うシステムは セキュリティに十分な配慮をすることが必要
- ✓ 所属機関のセキュリティ運用規定に加え 厚生労働省が策定している「医療情報システムの安全管理に関するガイドライン」など 各種ガイドラインに適合することを重視
- ✓ このプロジェクトはコロナ禍の真っ只中で進められ テレワークを前提とした運用環境の整備も必要であった

AWS Summit Japan 2024
第一三共株式会社における創薬研究クラウドプラットフォーム

Amazon Web Services ブログ

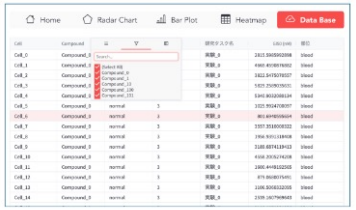
AWS Summit Japan 2024 第一三共株式会社における創薬研究クラウドプラットフォーム

by Takehiro Nakajima | on 03 10月 2024 | in Amazon Elastic Container Service, Amazon FSx for Lustre, AWS Batch, AWS Fargate, AWS ParallelCluster, AWS Step Functions, General, Healthcare, Life Sciences | Permalink | Share | 原文なし

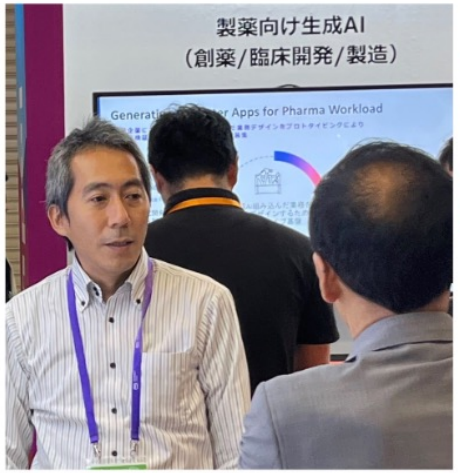
このブログは、第一三共株式会社 研究統括部 研究イノベーション企画部と、アマゾン ウェブ サービス ジャパン合同会社 ソリューション アーキテクト 中島丈博による共著です。

アジャイルアプリケーション開発に取り組んだ事例の紹介

アプリケーション開発にあたって、開発者とユーザーのワーキンググループを結成し、ユーザーのニーズに基づいて研究データ解析のためのアプリケーション開発に取り組んでおります。内製でアプリケーション開発を行う場合、ユーザーと開発者との距離が近いため、コミュニケーションが密に取りやすいと実感しています。そのため、要件を適宜確認し、フィードバックを頻繁にもらい、アプリケーションの改善を複数回にわたって行ってきました。実際にこのアプローチにて医薬品候補化合物の活性評価試験結果を可視化するビューワーを作成しました。



ユーザーの意見を反映した可視化方法を採用する等により、データサイエンティストと研究者が共同で効果的なアプリケーションの開発を実現しました。



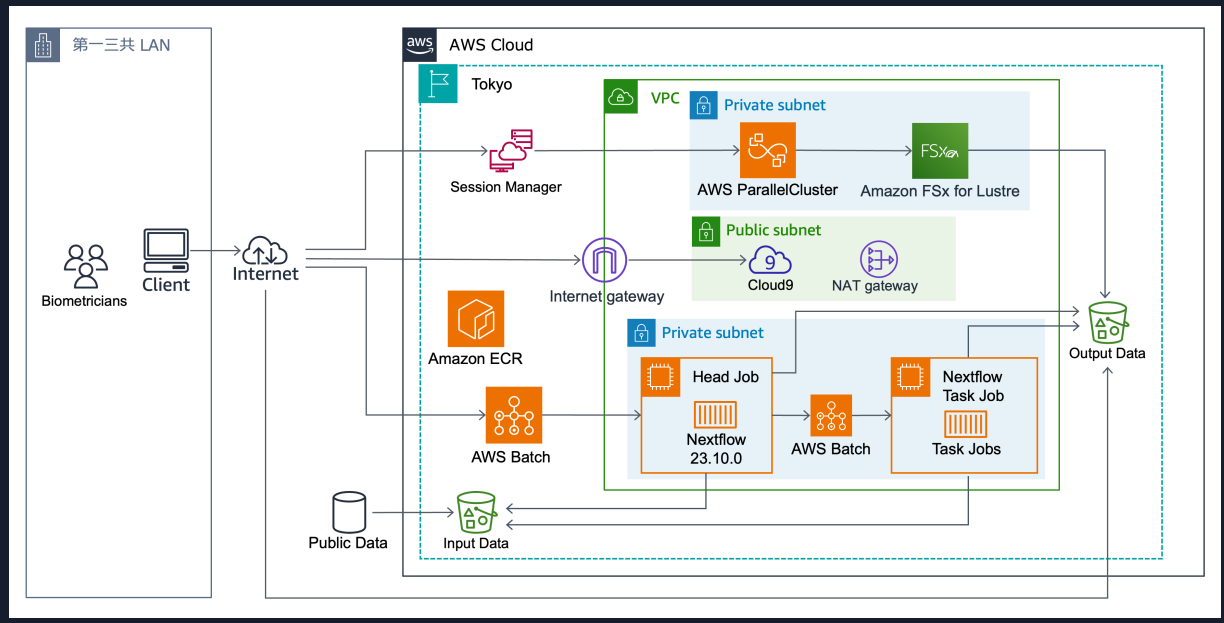
https://aws.amazon.com/jp/blogs/news/aws-summit-japan-2024-daiichisankyo-drug-discovery-research-cloud-platform/

AWS Summit Japan 2024

第一三共株式会社における創薬研究クラウドプラットフォーム

1. 実験データ転送の自動化
   - NGS のラン終了の 1–2 時間後には出力されたデータの解析を開始できるようになり、研究サイクルが迅速化

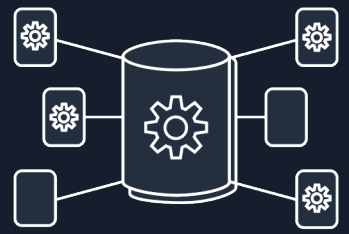
2. 柔軟に利用可能な HPC クラスター環境
   - 必要な計算リソースの見積もりが困難なオープンソースの解析ツールの検証を、柔軟性のある AWS クラウド環境で実現
   - ファイル入出力の負荷が大きい NGS のデータ解析も、分散ファイルシステムである Lustre のマネージドサービスを活用することで高速に処理



https://aws.amazon.com/jp/blogs/news/aws-summit-japan-2024-daiichisankyo-drug-discovery-research-cloud-platform/

AWS クラウド活用の真価 = お客様が価値提供に集中できること (再掲)

必要なときに
必要なだけ調達可能
な高い柔軟性と可用
性を併せ持つITリソー
ス



マネージド型
サービスの活用で
新たな技術革新を
迅速に適用可能

サービスとして提供される
セキュリティ機能の実装や
厳格なコンプライアンス
要件への対応

アジェンダ

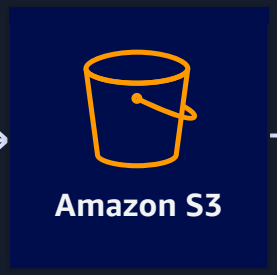
1. AWS クラウド活用によって得られる価値
   A. AWS とは
   B. HPC 環境の課題とクラウド活用による解決
   C. 活用事例: 国立がん研究センター様、第一三共株式会社様
2. ゲノム領域における AWS ソリューション
   A. ストレージとデータ転送
   B. AWS HealthOmics

高精度な解析のための AWSコンピューティングフレームワーク

データ転送とストレージ

サンプル

DNA シーケンサー



データ&分析

ワークフローと解釈のためのカタログサービス

長期保管

監査と災害復旧

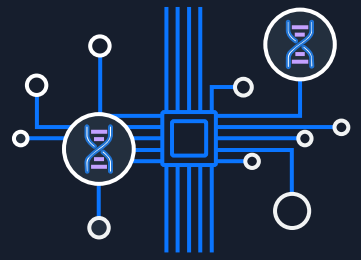
HIPAA 対応サービスを活用しセキュリティとプライバシーを高める

AWS HealthOmics

プロダクションレディーなオミクス解析環境をフルマネージドで迅速に提供

マルチオミックスと
マルチモーダル分析

集団ゲノム解析レベル
の規模に対応



フルマネージドな
バイオインフォマティクス
計算環境

組み込みのセキュリティ、
プライバシー、
コンプライアンス
(HIPAA 適格)

- 過去10年間にAWSがGenomics England, Stanford, Philips, AstraZeneca, Illumina, DNA nexus などのお客様と取り組んだゲノミクス関係の活動の知見に基づいて設計されている
- 一般利用開始。米国東部（バージニア北部）、米国西部（オレゴン）、欧州（アイルランド）、欧州（ロンドン）、欧州（フランクフルト）、アジアパシフィック（シンガポール）、アジアパシフィック（メルボルン）、イスラエル（テルアビブ）にて

AWS HealthOmicsでオミクス解析をスケールさせる

数百
同時実行回数

数千
同時実行タスク数

80,000+
同時実行 vCPU
(オンデマンド)

500+
同時実行 GPU 数
(オンデマンド)

Petabases
保存されているデータ

100,000+
バリアントストアに
保存されたサンプル数

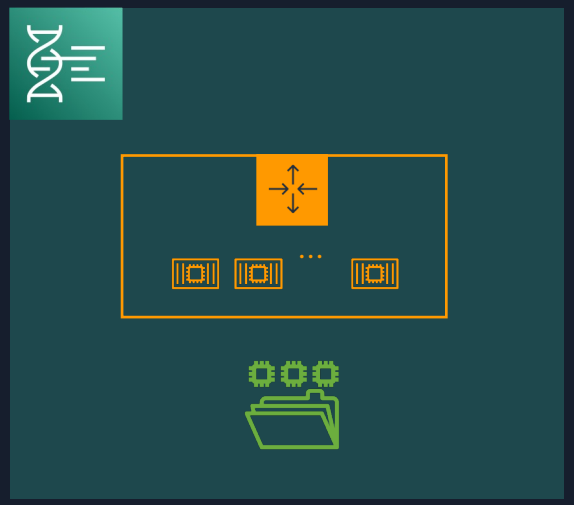
AWS HealthOmics 全体像

3 つの主要コンポーネントから構成

HealthOmics Storage

オミクスデータの保存

HealthOmics Workflow



オミクスデータの 2 次解析

HealthOmics Analytics

オミクスデータの 3 次解析と
マルチモーダル解析

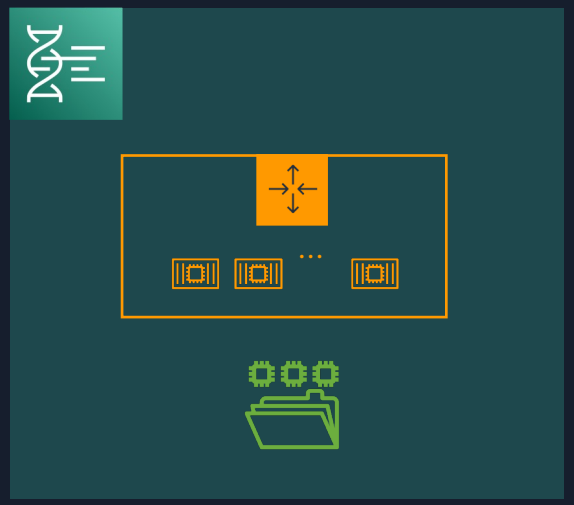
AWS HealthOmics 全体像

3 つの主要コンポーネントから構成

HealthOmics Storage

オミクスデータの保存

HealthOmics Workflow



オミクスデータの 2 次解析

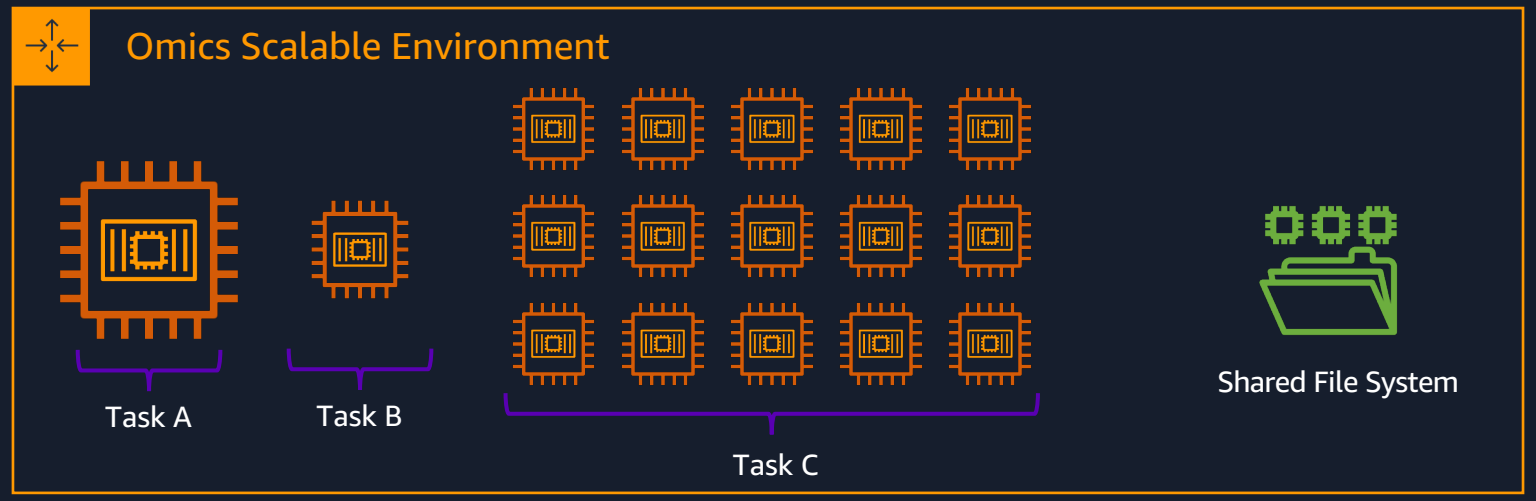
HealthOmics Analytics

オミクスデータの 3 次解析と
マルチモーダル解析

本日のスコープ

HealthOmics Workflows

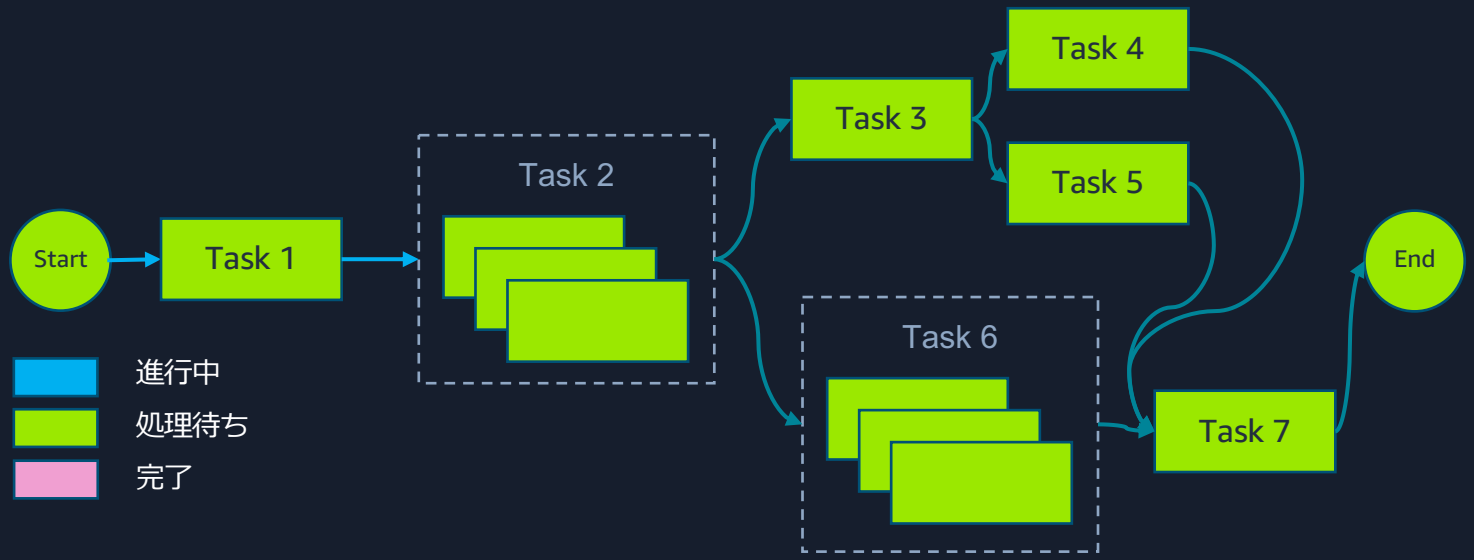
フルマネージドの伸縮自在なワークフロー実行環境により大規模解析を効率的に実施

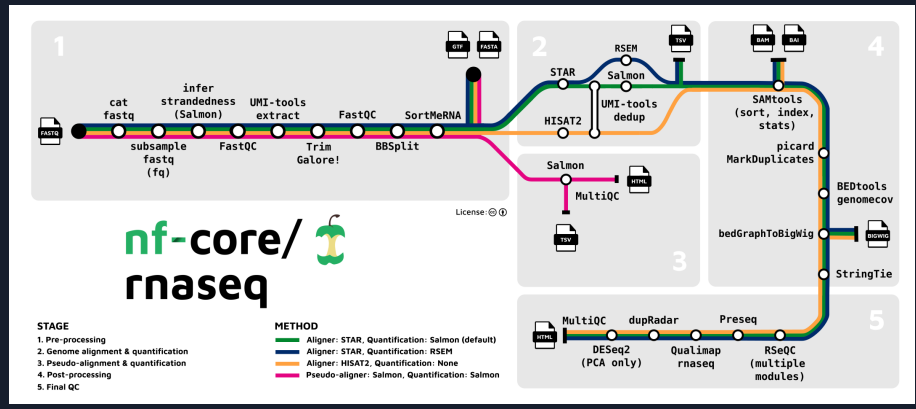


- WDL/Nextflow/CWL でバイオインフォマティクスワークフローを記述し実行
- ワークフローは進行状況に応じて Omics Workflow フルマネージドスケジューラが必要リソースに応じてコスト最適なインスタンスを自動で起動・停止し、複雑な依存関係を解決
- ワークフロー各ステップはタスクと呼ばれ、ワークフロー実行中存在するフルマネージド共有ファイルシステムを介してやりとり
- ワークフローは、最大 vCPU 数・計算時間に制限を設けることが可能な Run Group 内で実行可能

クラウドの伸縮性で必要な時に必要な計算リソースを活用

ワークフローの各ステップが要する計算スペックごとにコスト最適なリソースを自動で割り当て、ワークフロー言語で記述されたジョブの依存関係や並列計算を管理

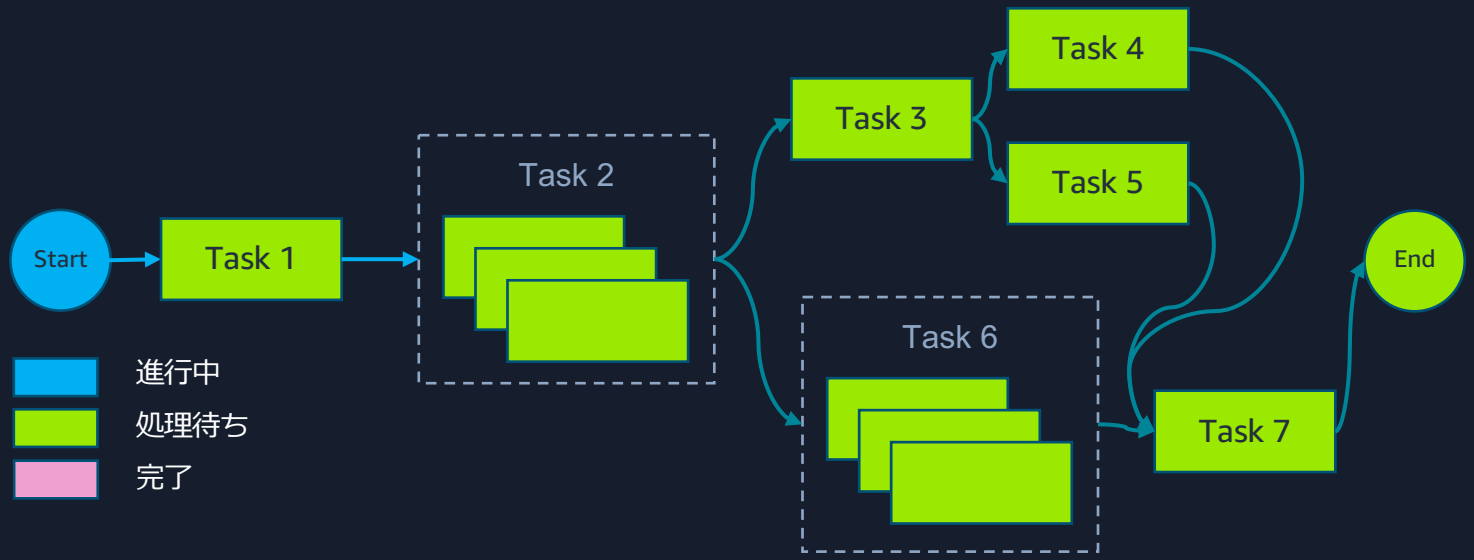


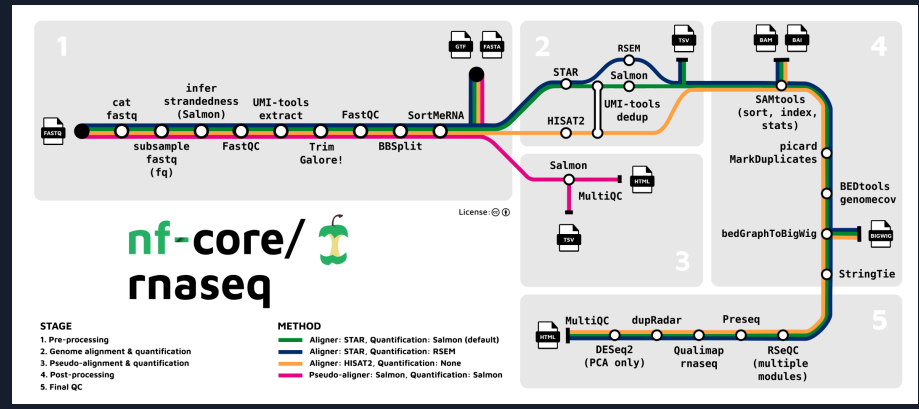


https://nf-co.re/rnaseq

クラウドの伸縮性で必要な時に必要な計算リソースを活用

ワークフローの各ステップが要する計算スペックごとにコスト最適なリソースを自動で割り当て、ワークフロー言語で記述されたジョブの依存関係や並列計算を管理





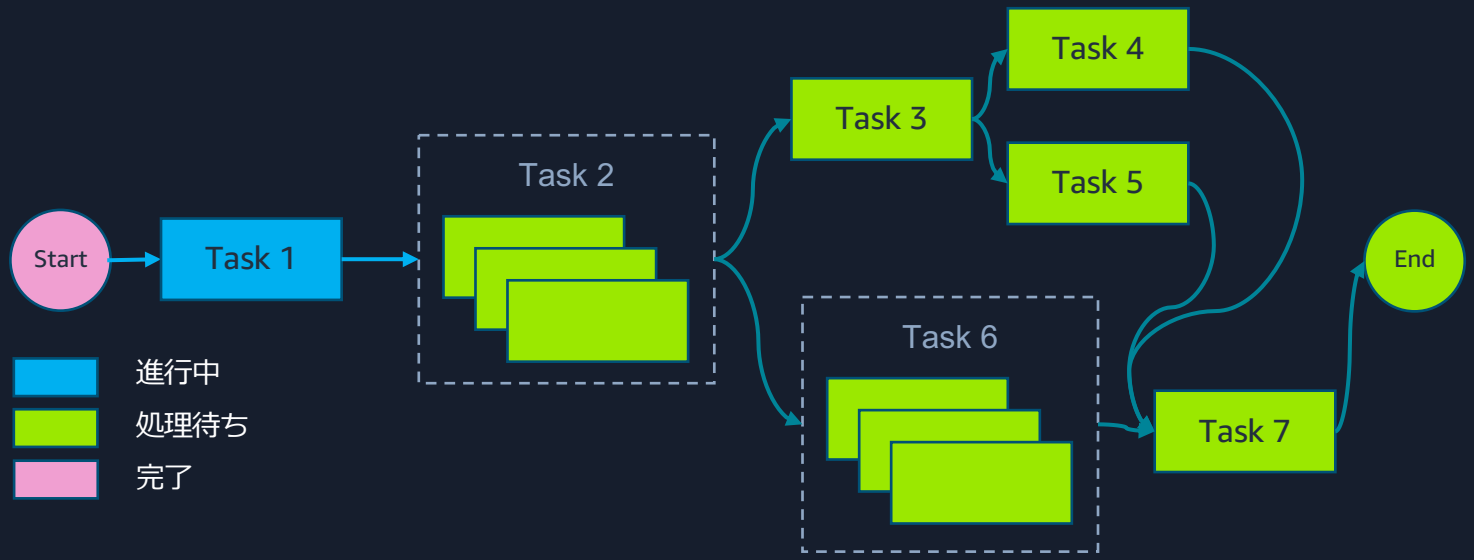
https://nf-co.re/rnaseq

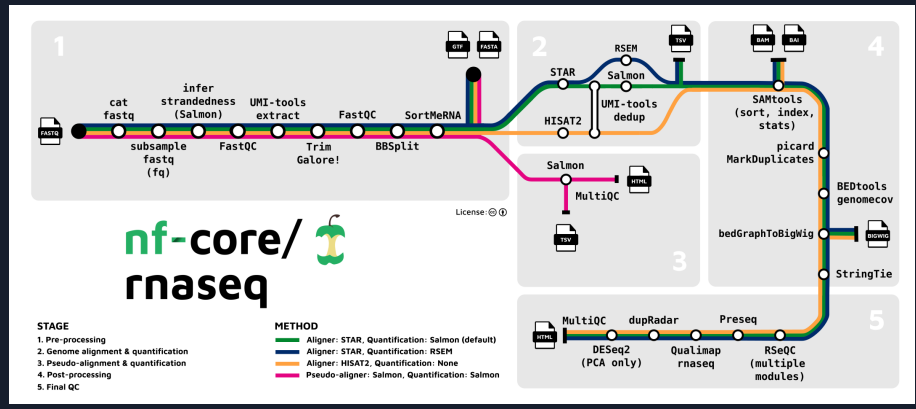
Omics Scalable Environment

開始前、リソースは**全く 0** で課金はない
(常駐ジョブは不要)

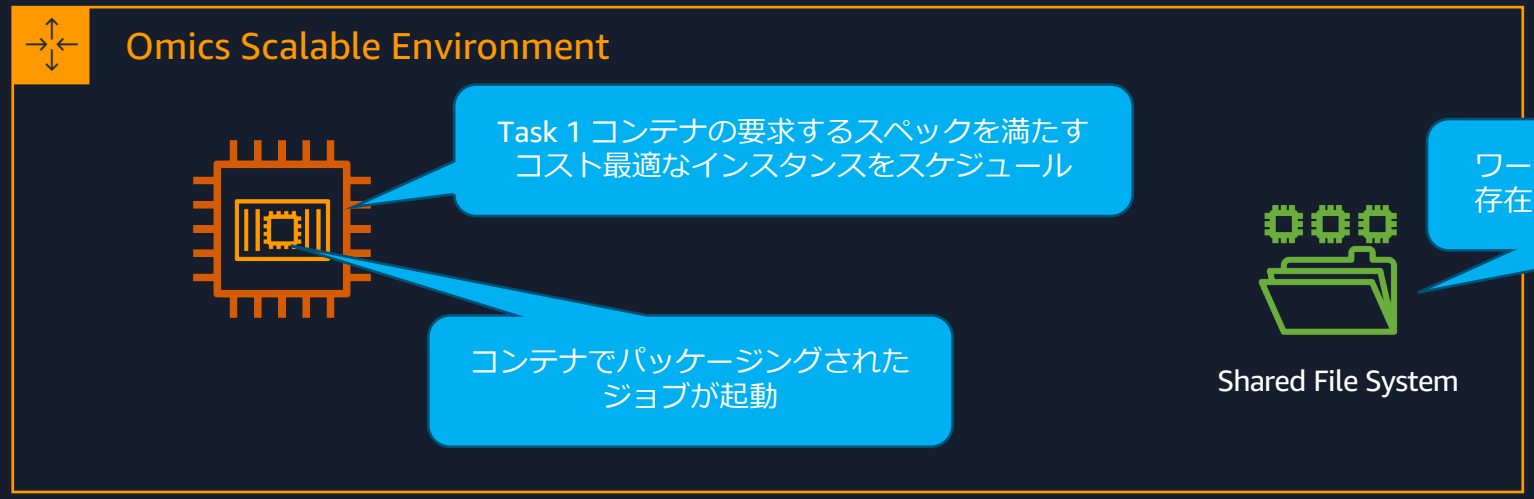
クラウドの伸縮性で必要な時に必要な計算リソースを活用

ワークフローの各ステップが要する計算スペックごとにコスト最適なリソースを自動で割り当て、ワークフロー言語で記述されたジョブの依存関係や並列計算を管理



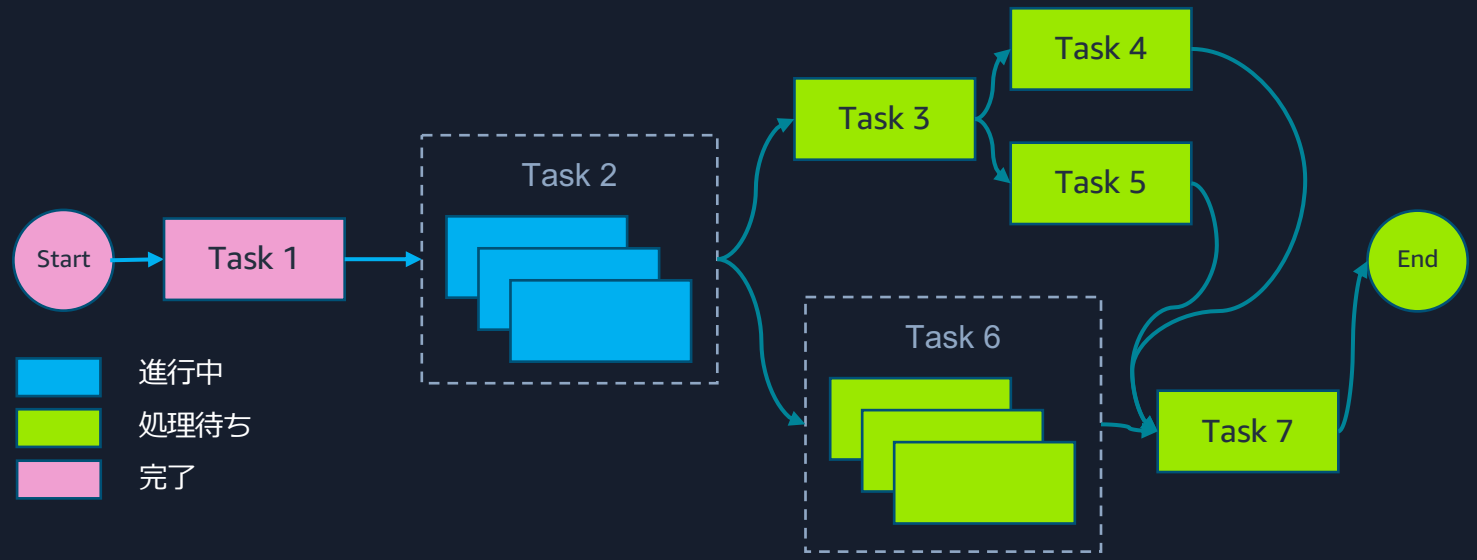


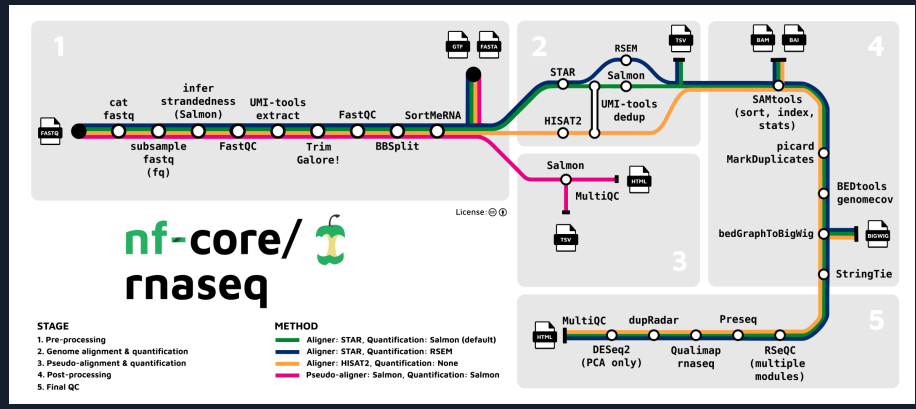
https://nf-co.re/rnaseq



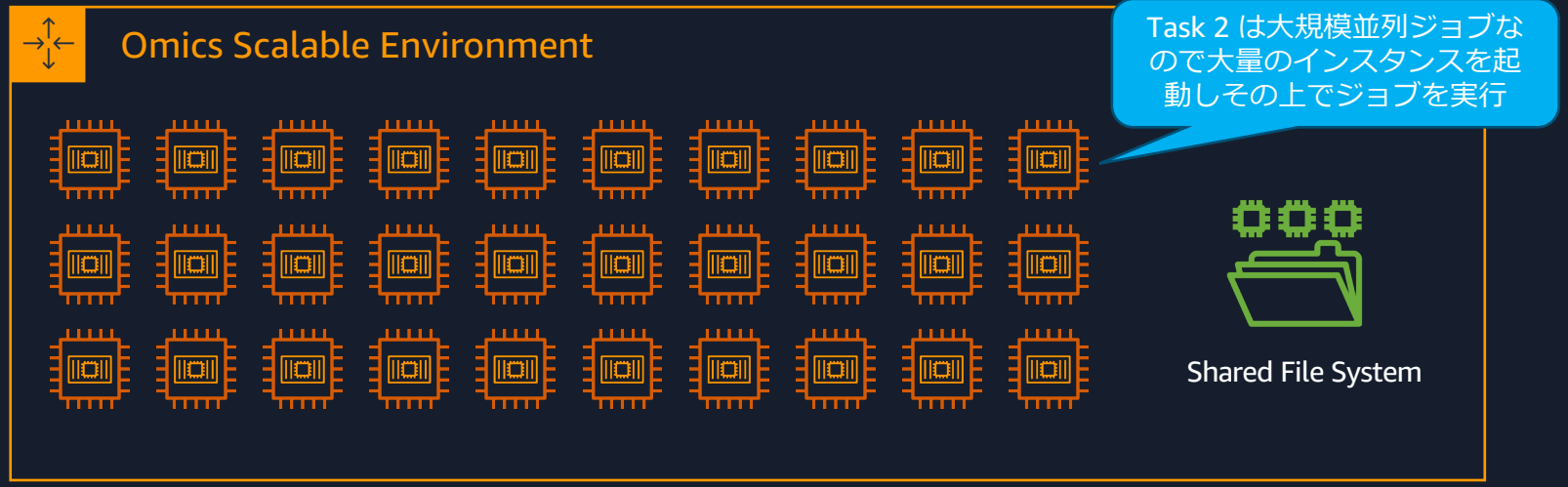
クラウドの伸縮性で必要な時に必要な計算リソースを活用

ワークフローの各ステップが要する計算スペックごとにコスト最適なリソースを自動で割り当て、ワークフロー言語で記述されたジョブの依存関係や並列計算を管理



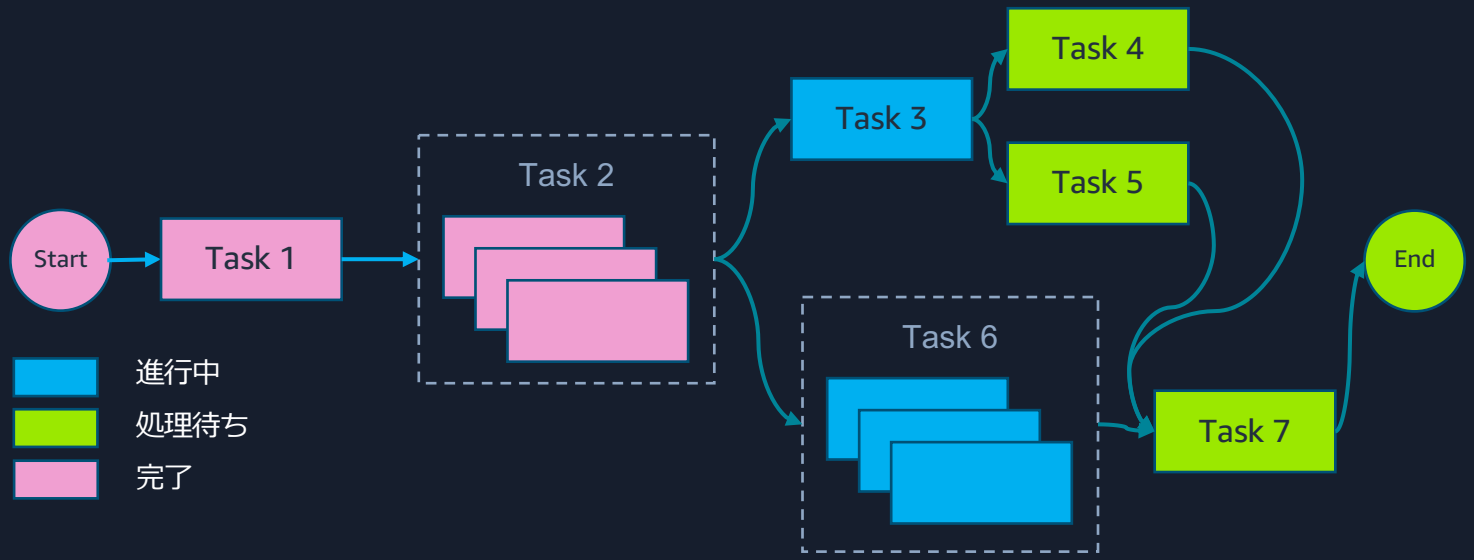


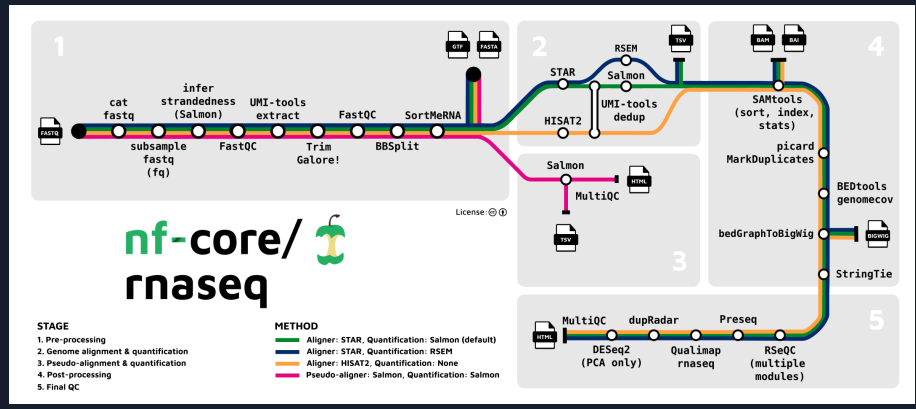
https://nf-co.re/rnaseq



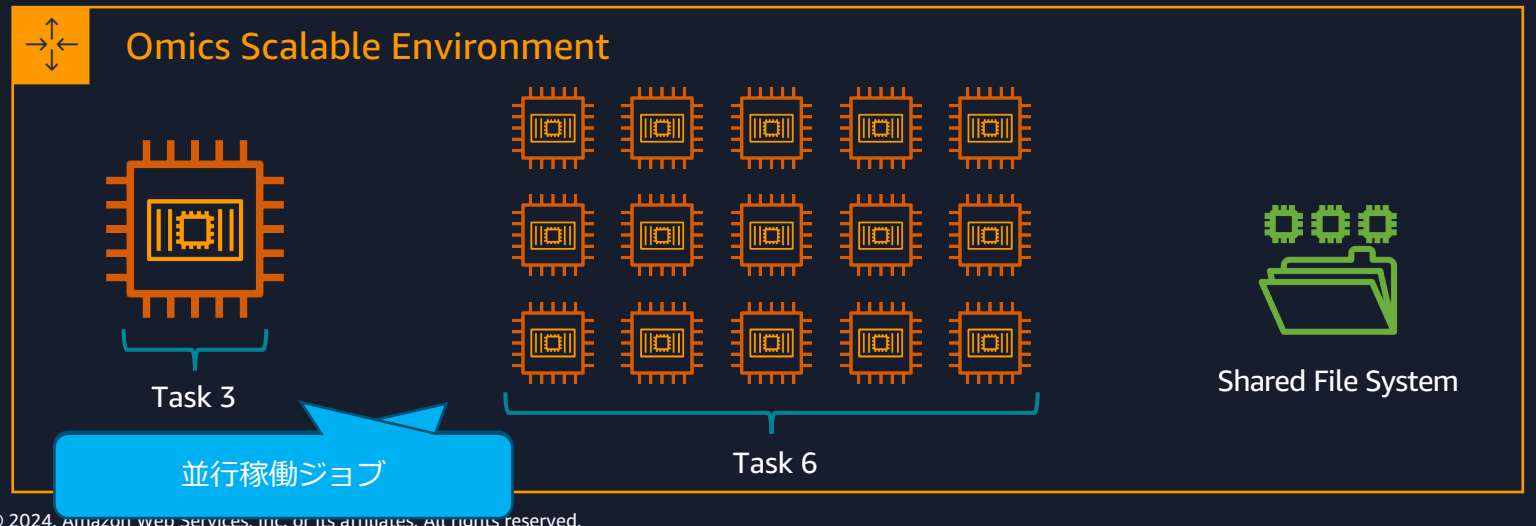
クラウドの伸縮性で必要な時に必要な計算リソースを活用

ワークフローの各ステップが要する計算スペックごとにコスト最適なリソースを自動で割り当て、ワークフロー言語で記述されたジョブの依存関係や並列計算を管理



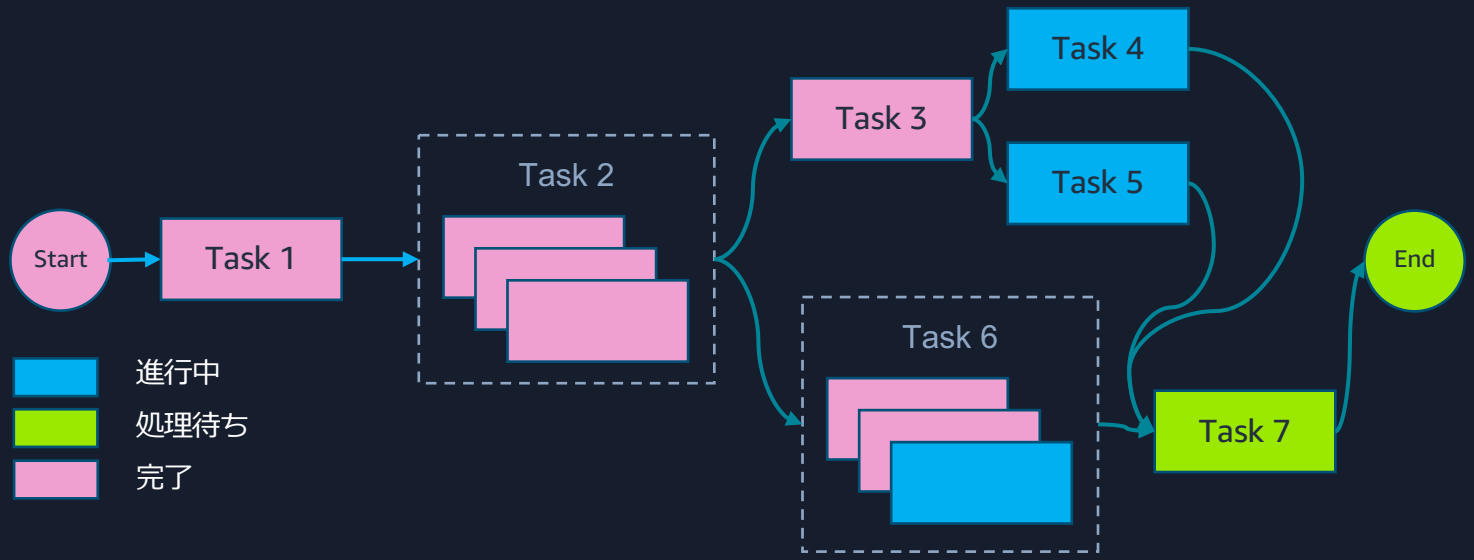


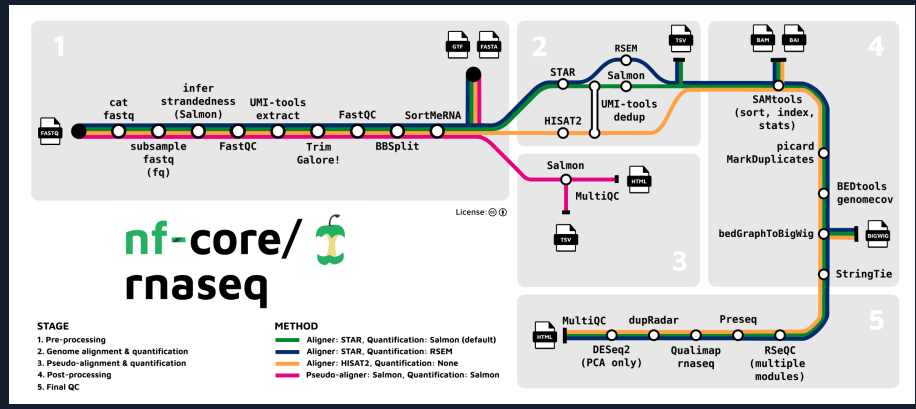
https://nf-co.re/rnaseq



クラウドの伸縮性で必要な時に必要な計算リソースを活用

ワークフローの各ステップが要する計算スペックごとにコスト最適なリソースを自動で割り当て、ワークフロー言語で記述されたジョブの依存関係や並列計算を管理

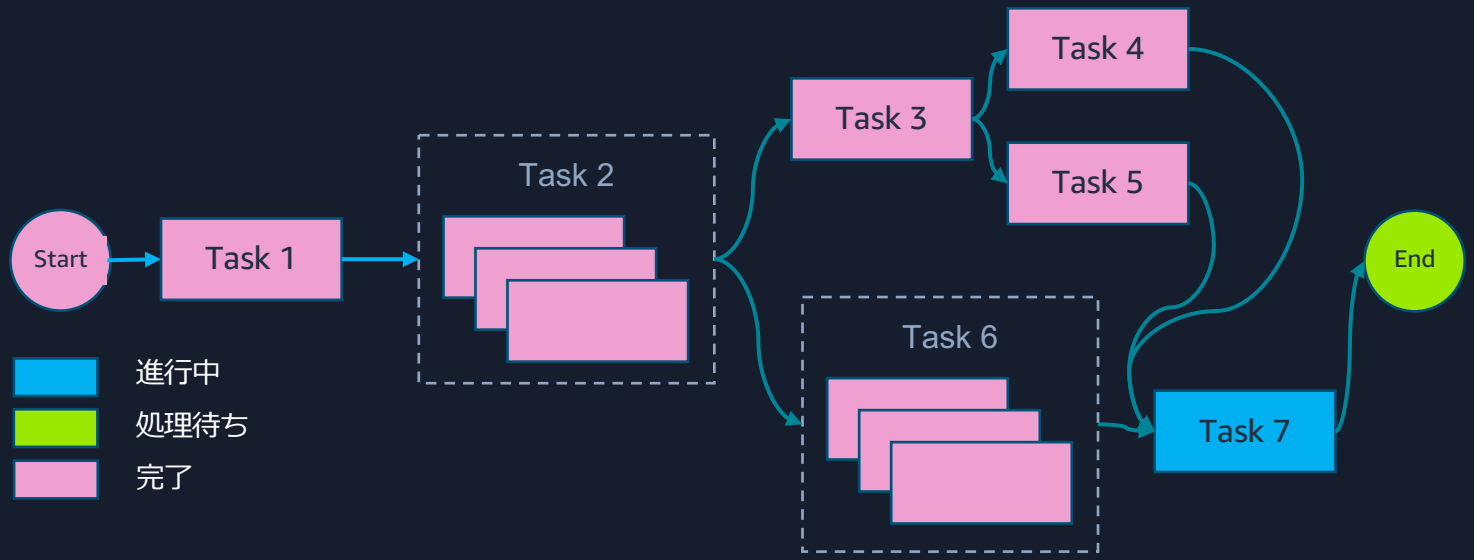


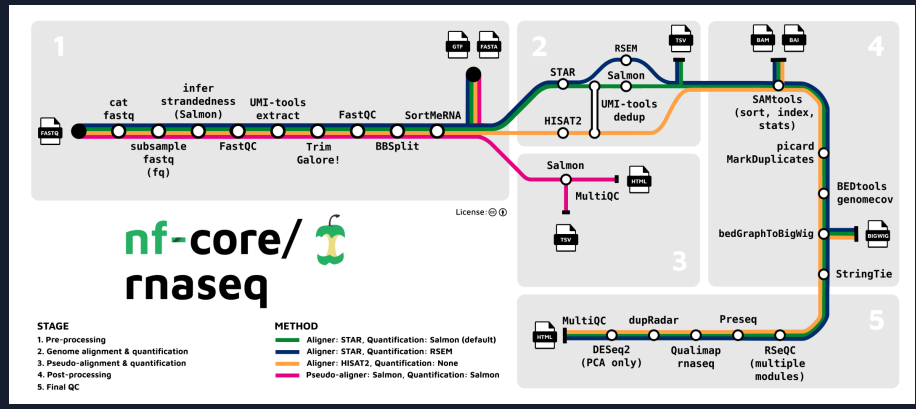


https://nf-co.re/rnaseq

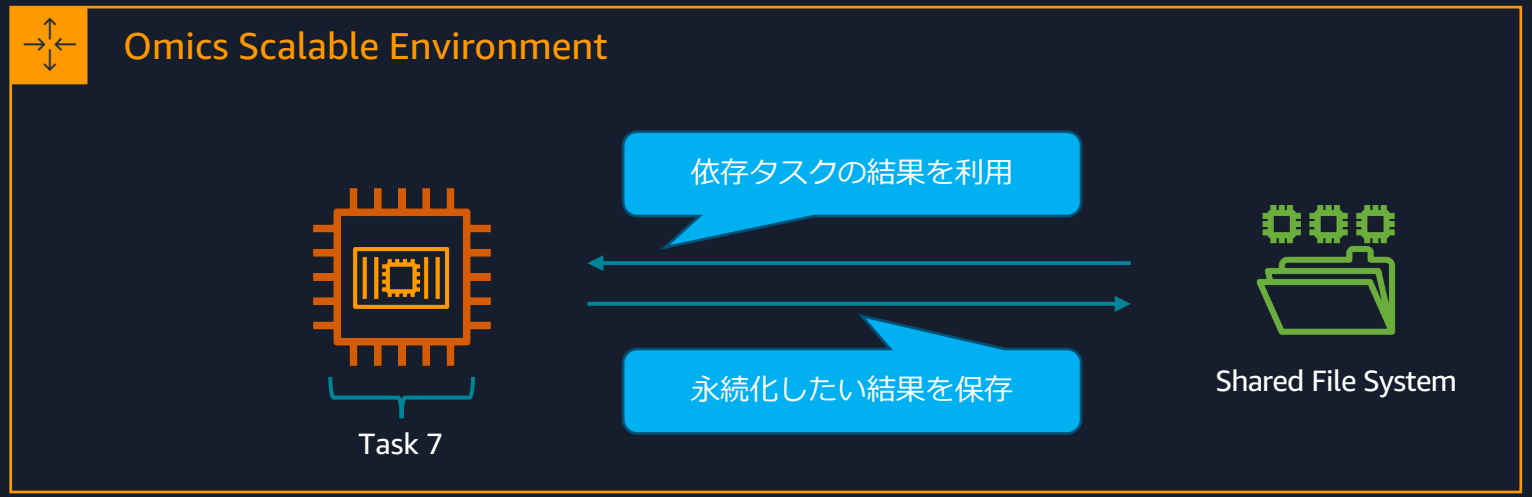
クラウドの伸縮性で必要な時に必要な計算リソースを活用

ワークフローの各ステップが要する計算スペックごとにコスト最適なリソースを自動で割り当て、ワークフロー言語で記述されたジョブの依存関係や並列計算を管理



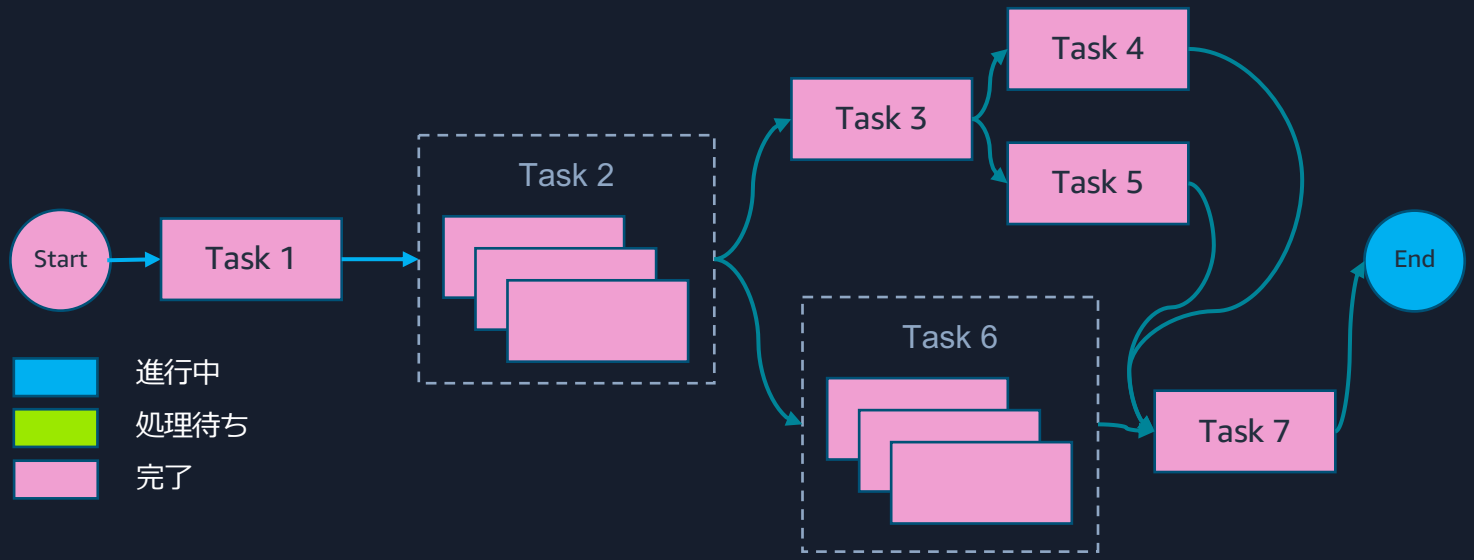


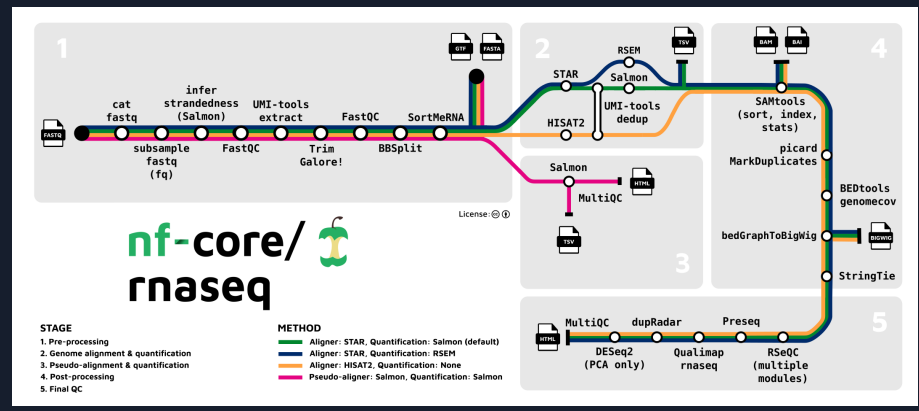
https://nf-co.re/rnaseq



クラウドの伸縮性で必要な時に必要な計算リソースを活用

ワークフローの各ステップが要する計算スペックごとにコスト最適なリソースを自動で割り当て、ワークフロー言語で記述されたジョブの依存関係や並列計算を管理





https://nf-co.re/rnaseq

Omics Scalable Environment

利用中リソースは**全く 0** (共有ファイルシステム含む)になり課金停止

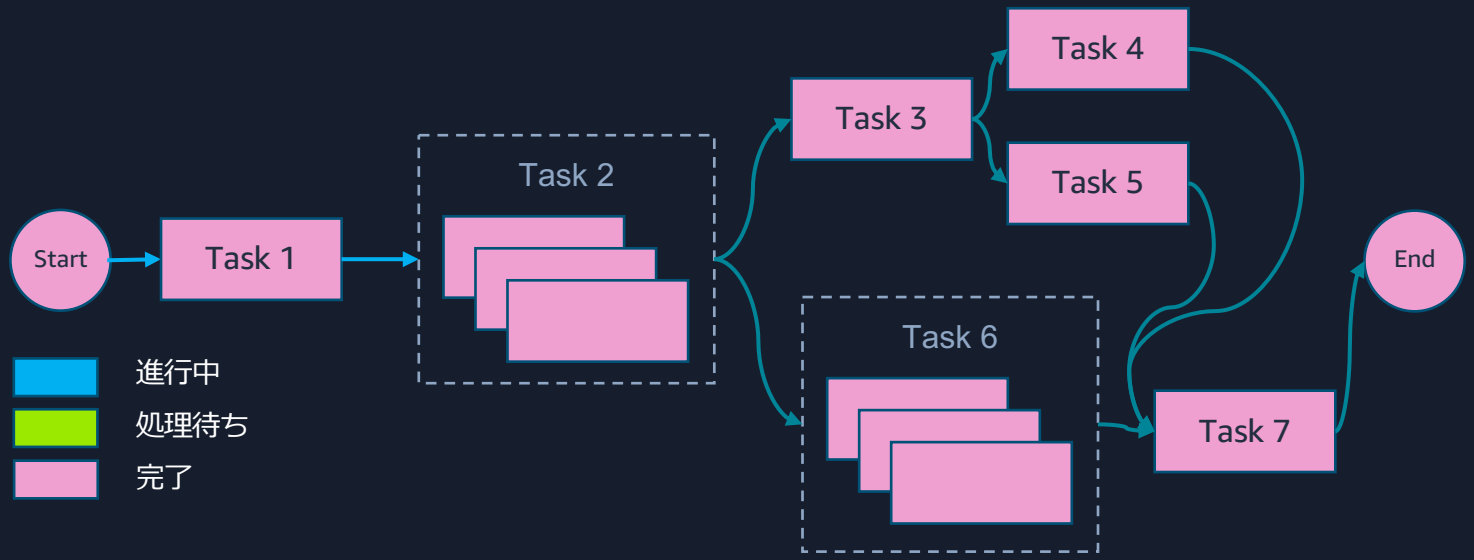
Shared File System

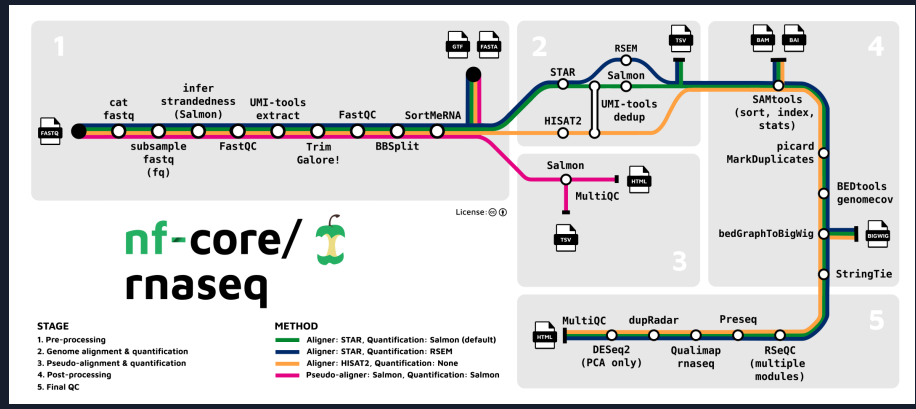
結果は S3 バケットに保存し永続化

Amazon S3

クラウドの伸縮性で必要な時に必要な計算リソースを活用

ワークフローの各ステップが要する計算スペックごとにコスト最適なリソースを自動で割り当て、ワークフロー言語で記述されたジョブの依存関係や並列計算を管理





https://nf-co.re/rnaseq

HealthOmics Workflow の利点 まとめ

- ワークフロー実行のためのインスタンススペックやデータのやり取りの方法は全て HealthOmics Workflow にお任せ
- ワークフロー言語と入出力データのパスの考慮をするだけでワークフローのクラウドスケールでの大規模実行が可能

AWS Healthomics: Ready2Run ワークフロー

1-API Call

Inputs from S3 or Omics storage

AlphaFold

NVIDIA

Element Biosciences

ESMFold

gatk

Sentieon

nf-core

1 回の実行ごとの固定料金

36 Ready2Run マルチオミクスワークフロー

GATK best practice pipelines
GATK BP germline fq2vcf for 30x
GATK BP FASTQ to BAM
GATK BP Germline BAM to VCF for 30x
GATK BP Somatic WES
GATK BP Somatic WGS

Single cell transcriptomic analysis
scRNAseq with STARsolo
scRNAseq with Kallisto + BUStools
scRNAseq with Salmon Alevin-fry + AlevinQC

Protein folding prediction
ESMFold for up to 800 residues
AlphaFold + MSA for up to 600 residues
AlphaFold + MSA for 601–1200 residues

Element Biosciences

Element Biosciences
ElementBio Bases2Fastq for 2x75
ElementBio Bases2Fastq for 2x150
ElementBio Bases2Fastq for 2x300

Sentieon

Sentieon Inc.
Sentieon Germline DNAseq FASTQ WGS& WES
Sentieon Somatic WGS & WES
Sentieon LongRead for PacBio HiFi
Sentieon LongRead for ONT
Sentieon Germline DNAseq BAM WES & WGS

NVIDIA PARABRICKS

NVIDIA
NVIDIA Germline (DeepVariant)
NVIDIA Germline (Haplotype Caller)
NVIDIA BAM2FQ2BAM
NVIDIA FQ2BAM
NVIDIA Somatic WGS & WES

まとめ

- AWS クラウドを活用することで、差別化に繋がらないような作業をオフロードし、新たな価値を創造する活動に注力することができる
- AWS はゲノミクス業界の多様なニーズに応える幅広いソリューションを提供
  - ゲノムデータの転送、および高コスト効率なストレージへの保管
  - 計算環境構築の手間を大幅に削減

Thank you!