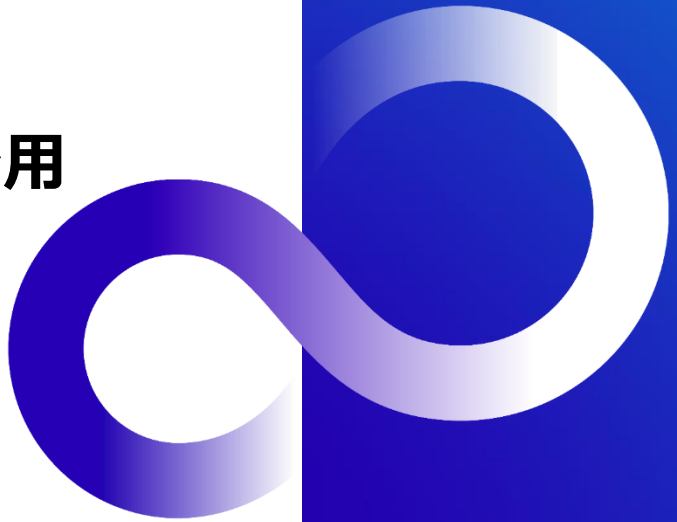
2024/02/05
PCCC AI/HPC OSS活用ワークショップ

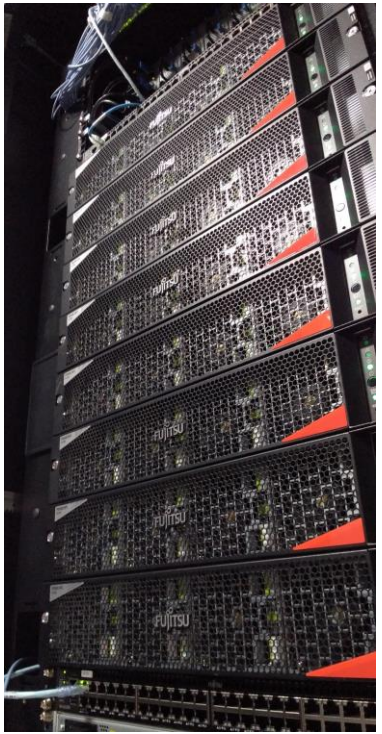
Arm CPU搭載サーバを用いた量子コンピュータシミュレーション用クラスタの構築と運用

富士通株式会社
コンピューティング研究所
大辻 弘貴



本発表の概要

- 富士通の量子コンピュータシミュレーション用クラスタについて、OSS活用の観点でご紹介します
  - 量子コンピュータシミュレーション
  - 構築・運用の流れ
  - ハードウェアおよびシステムの構成・OSSの活用
  - 大規模量子シミュレーション向けのスケジューリング技術
  - 運用実績



FUJITSU

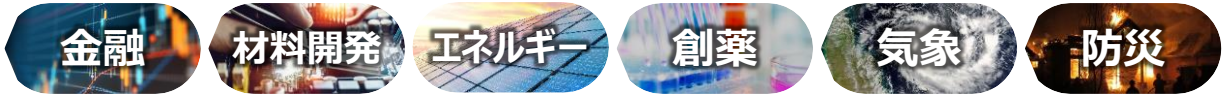
量子コンピュータシミュレーション

富士通がComputing as a Serviceで目指す世界

Provide the top-class Computing Technologies “as a Service”

FUJITSU

Application



Platform

AI | Simulation | Data Analytics

Computing as a Service

Low power consumption | Great computing power | Trust platform

Available on Public Cloud

Middleware

OS

Hardware

High Performance Computing (HPC)

A64FX Technology

Quantum-Inspired Technology

Digital Annealer

Quantum Simulator

Quantum Technology

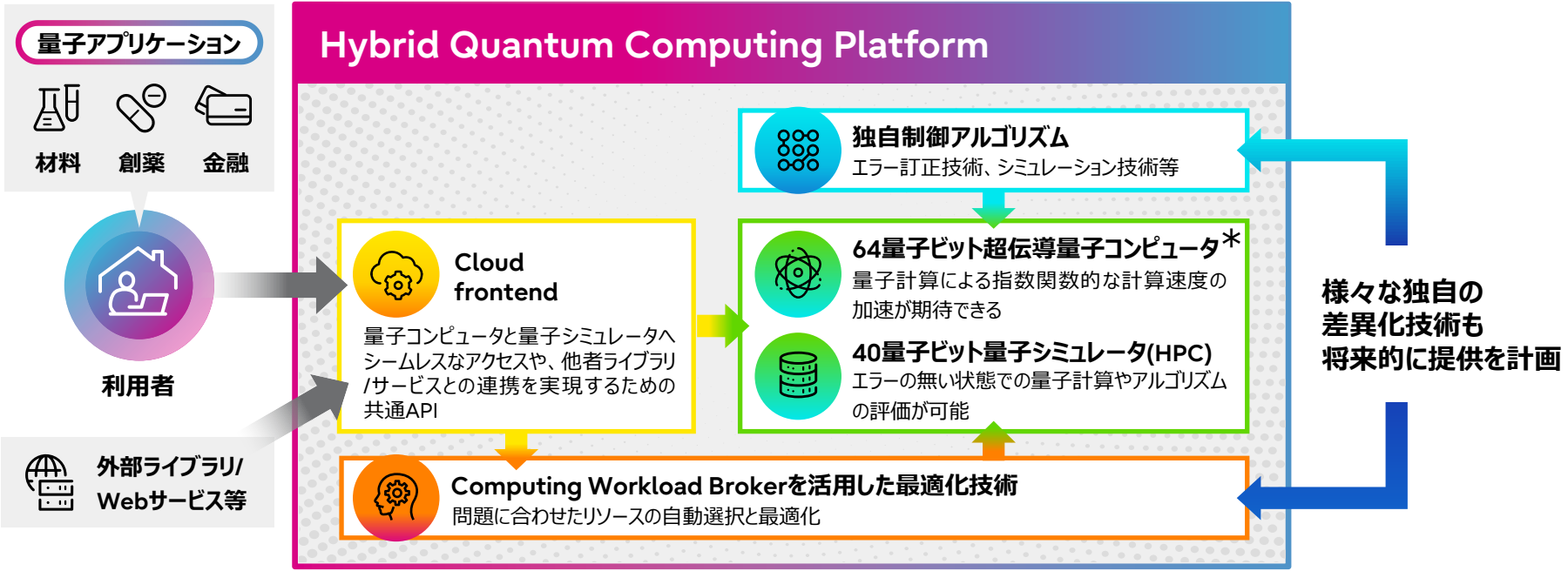
©RQC

Superconducting Qubit
Diamond Spin Qubit

ハイブリッド量子コンピューティングプラットフォーム

FUJITSU

- 量子コンピュータ（実機）と量子シミュレータのシームレスな操作を実現
- 量子コンピュータと量子シミュレータ双方のメリットを活かした計算手法の開発にも活用

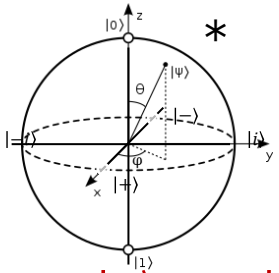


*理研RQC-富士通連携センターで開発

量子コンピュータシミュレーション

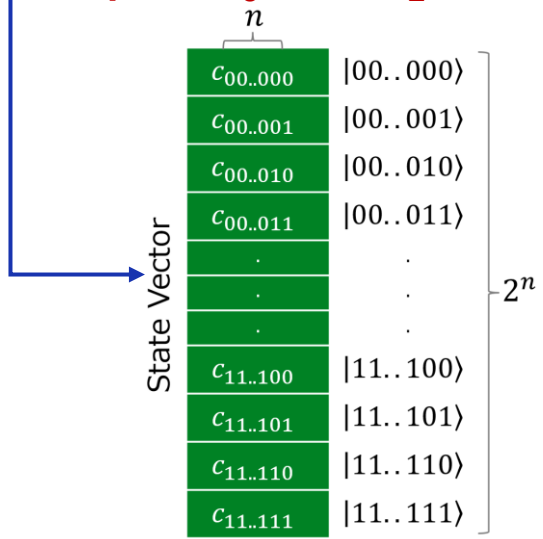
FUJITSU

- 量子ビットの状態変化を通常のコンピュータ上でシミュレーション
  - 量子ビットの状態をメモリ上で数値として表現
- 量子ビット数(n)の場合において、表現には2^nの空間が必要
  - →量子ビット（Qubit）数を増やすためには膨大なメモリおよびそれらに対する演算が必要
  - →40 Qubitで16 TiB
- **<u>クラスタ型計算機が不可欠</u>**
  - →高性能計算向けArm CPUを採用したサーバを用いてクラスタを構築



*

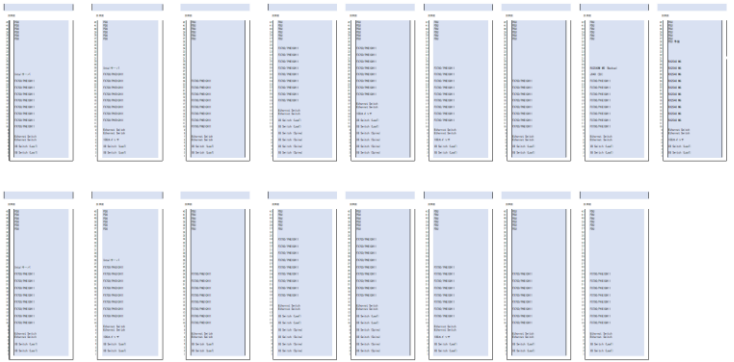
$$|\psi\rangle = c_0|0\rangle + c_1|1\rangle$$



* https://ja.wikipedia.org/wiki/%E3%83%96%E3%83%AD%E3%83%83%E3%83%9B%E7%90%83

システムの概要

- システム規模
  - FUJITSU Supercomputer PRIMEHPC FX700 1,056ノード
    - 富士通製A64FXプロセッサ搭載(Arm v8.2-A SVE)
  - FUJITSU Server PRIMERGY RX2540 M6 約10ノード
    - 管理系やログインノード、共有ファイルシステム、統計収集用
  - Intel CPU搭載ノード（数台）
    - Intelプロセッサを必要とするコード用計算ノード
  - ネットワーク
    - EthernetおよびInfiniBand HDR100/HDR200
- 設置場所
  - 自社データセンター
  - 17ラックを使用



ラック搭載配置図

構築・運用の流れ

スケジュール

構築スケジュール

FUJITSU

- 2022年4月～8月中旬*
  - システム構成検討・ソフトウェア選定
  - ラック配置・結線の設計資料作成・テスト環境における構築シミュレーション
- 2022年8月中旬～9月中旬
  - 工事・ケーブル敷設・ラック搭載（物理的な作業は委託）
  - 設置に合わせて実地検証をスタート、9月中旬に全機器ハード確認完了
- 2022年9月中旬～9月末
  - 管理系サーバおよび計算ノード520ノードのシステム設定を完了、**稼働開始**
- 2023年7月
  - 計算ノードを1056ノードに増設

*2022年3月に発表した36量子ビットシミュレータとは別システムとして開発

事前検証・設定自動化により迅速な構築を実現

構築体制

FUJITSU

- 研究レベル技術の短期適用のため研究者が主導して構築・運用
  - 量子コンピュータシミュレーションソフトウェアの開発・展開
  - ユーザとの調整や利用に関わるインタフェース開発、外部ネットワークの構築
  - コンピューティングクラスタとしてのシステム構成検討・構築
  - 計算ノード・管理系のソフトウェア設定・環境構築

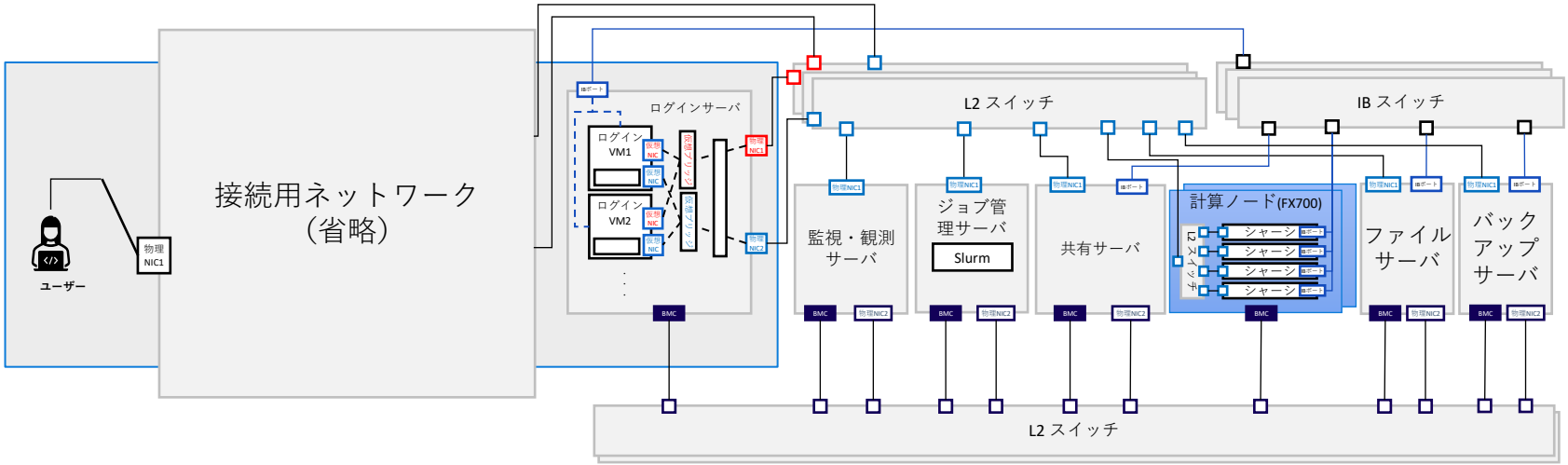
- 構築後の通常運用は委託
  - 運用マニュアルの作成やエスカレーション先は研究部門
  - サポート契約なしにOSSを利用

FUJITSU

ハードウェアおよび
システムの構成・OSSの活用

システムの全体構成

- 主な計算資源およびインターコネクト
  - FUJITSU Supercomputer PRIMEHPC FX700 1056ノード
  - InfiniBand HDR100 Full-bisection FatTree接続(48ポートHDRスイッチx50)
- 管理系やストレージは汎用サーバで構築
  - ストレージサーバの記憶デバイスはNVMe SSDおよびOptane不揮発メモリを採用



OSSの活用

FUJITSU

- 全ての要素をOSSで構成
  - 計算ノードOS・デプロイ
    - Rocky Linux 8系、Kickstartによるインストール、Ansibleによる設定
  - ジョブスケジューラ
    - OSSであり機能拡張が可能であることから、Slurmを選択
      - インタラクティブ性を高めるための拡張については後述
  - 共有ファイルシステム
    - GlusterFS
      - ストレージノードレベル・デバイスレベルの冗長性をソフトウェアで確保
    - CHFS (Consistent Hashing File System)
      - Optane不揮発メモリとInfiniBandを活用し、高速データアクセス領域を提供
  - 監視・観測
    - Prometheus, Grafana, Alertmanager, node-exporter
      - 状態監視、アラート通知

ジョブスケジューラ構成

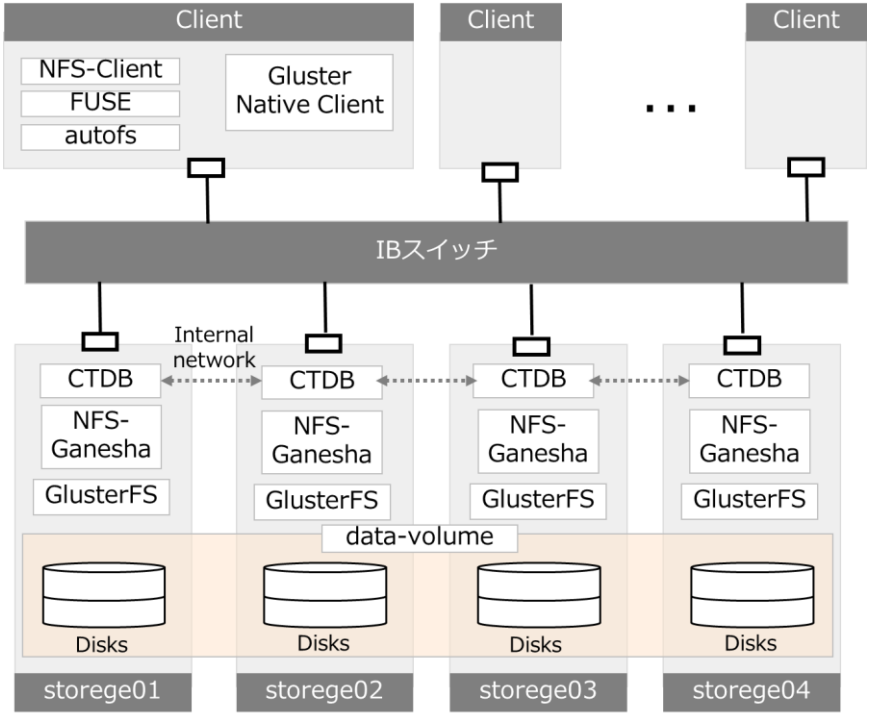
FUJITSU

- 全計算資源はジョブスケジューラを介して利用
- 採用OSS：
  - Slurm (現在は23.xを使用)
  - SlurmとOpenMPIをPMIx経由で連携
- 構成
  - ノード単位のリソース管理
  - 3種類のキュー：
    - バッチキュー、占有インタラクティブ、タイムスライスインタラクティブ
    - タイムスライスインタラクティブについては後述

様々な需要に応えるための複数キューを運用

共有ファイルシステム構成

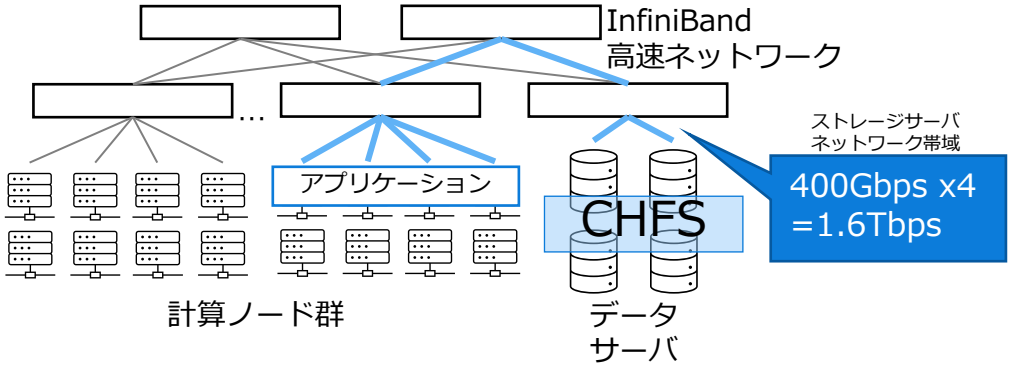
- 全計算ノードに対して単一空間の共有ファイルシステムを提供
- 採用OSS:
  - GlusterFS – RedHatにより開発されている分散ファイルシステム
- 構築のポイント
  - デバイスレベル・ノードレベルのデータ冗長化による耐故障性
    - Distributed Dispersedボリューム
  - フェイルオーバーによる高可用性
  - 小サイズ・多数ファイルアクセス向けにイメージマウント機能を提供

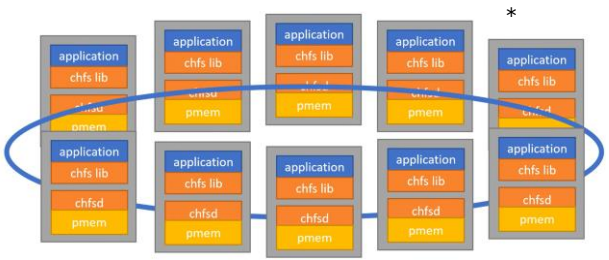


汎用サーバとOSSのみで並列・高可用性・耐故障性を実現

高速ファイルシステム構成

- CHFS (Consistent Hashing File System)を展開中
  - Optane不揮発メモリとInfiniBandの高速通信を活用した共有データストア
- メモリ上データの退避先としても活用予定

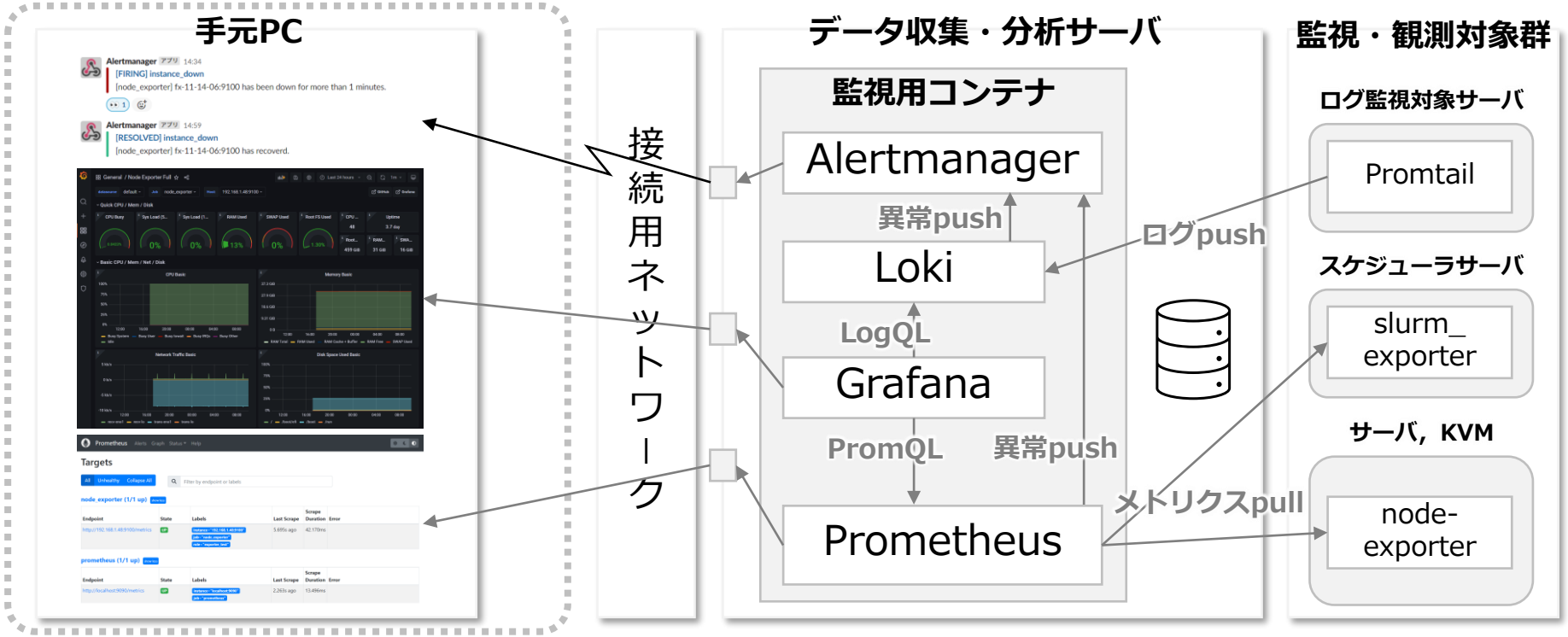




* O. Tatebe et al., CHFS: Parallel Consistent Hashing File System for Node-local Persistent Memory, HPC Asia 2022

監視・観測システムの構成

- OSSを活用してメトリクス収集や可視化、アラート通知を実施



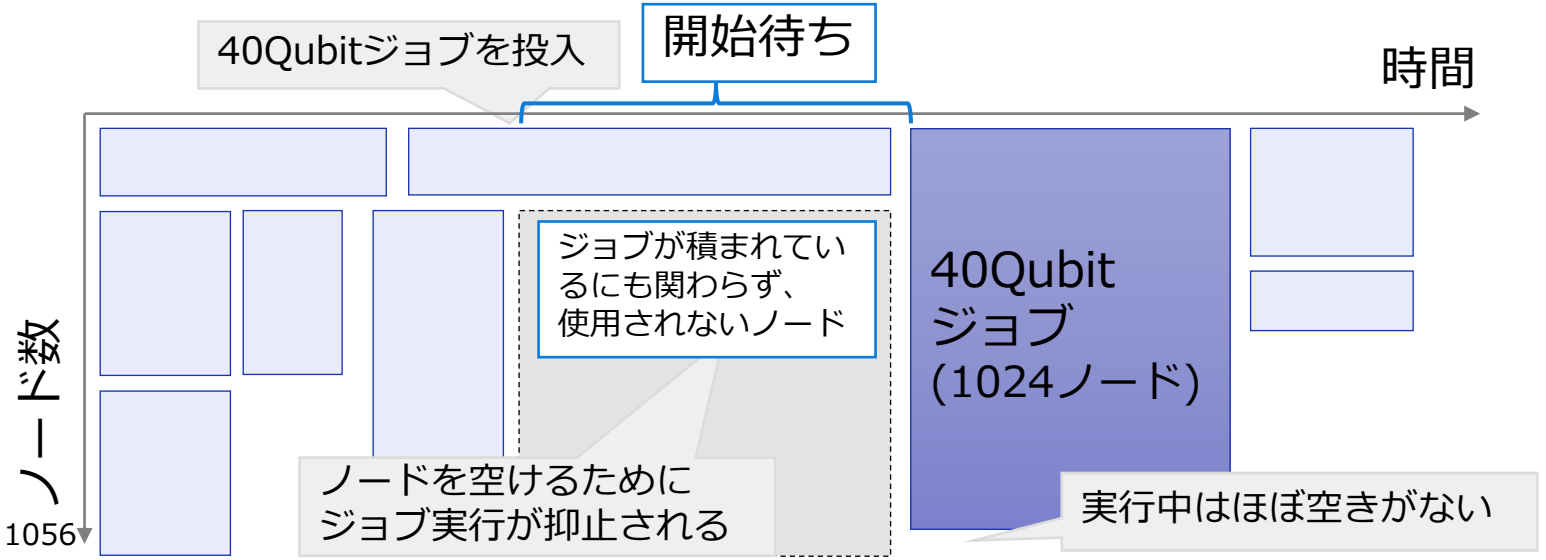
FUJITSU

大規模シミュレーション向けのスケジューリング技術

大規模シミュレーションジョブ投入時の課題

FUJITSU

- 40Qubitシミュレーションジョブは1024ノードを使用
  - 投入すると実行前・実行中に他のジョブ実行が難しくなる

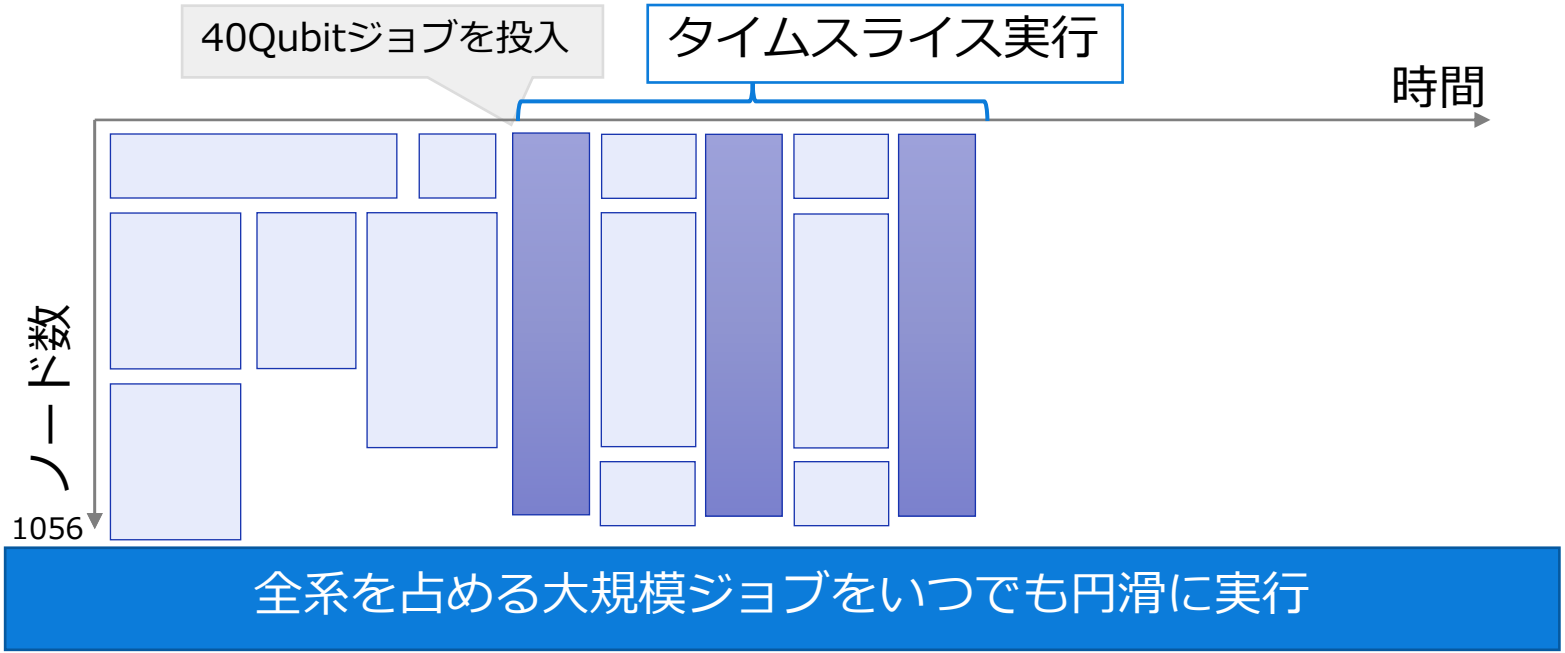


大規模ジョブの実行前・実行中に他のジョブが滞る

タイムスライス実行による大規模ジョブ効率化

FUJITSU

- 大規模ジョブを他ジョブと細粒度タイムスライス実行
  - 40Qubitジョブを即時実行可能になり、計算ノード利用率も向上

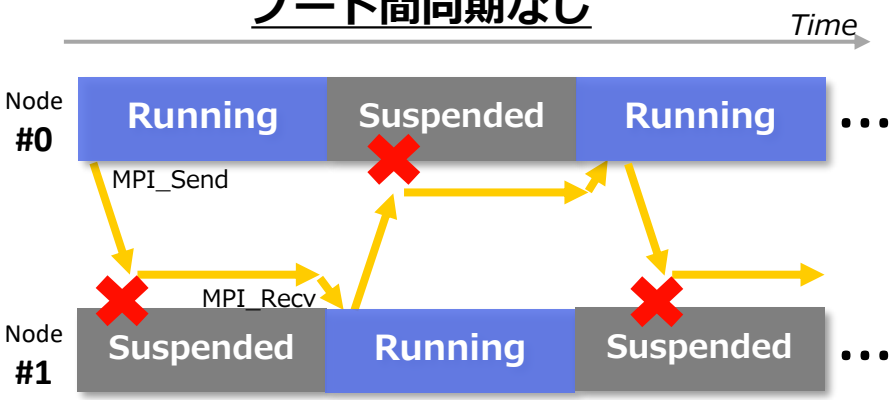


タイムスライス実行のための全系高精度同期

FUJITSU

- 並列プログラムの並行動作のためにはノード間同期が不可欠
  - いわゆるGang scheduling

ノード間同期なし



性能低下が発生

ワーストケースでは1スライスにつき1メッセージしか送受信できない

ノード間同期あり

通常バッチ実行並みの性能を実現

一度のタイムスライスで大量のメッセージを通常通り送受信可能

全てのノードで高精度なジョブ状態の同期が必要

細粒度タイムスライス実行機能のSlurm統合

FUJITSU

- Slurmの標準機能は5秒未満のタイムスライス実行に非対応
  - 短いタイムスライスでは性能低下が発生する可能性がある

- 大規模環境においても高精度にジョブ切替を行う機構を開発(*1)
  - 複数のSlurmジョブを1秒以下のタイムスライスで切り替え
    - ジョブ情報やプロセス情報を収集し、ギャングスケジューリングを実施
  - 切り替え信号をブロードキャスト通信により伝達
    - 高い同期性を実現

(*1) Ohtsuji et al., Scalable Fine-Grained Gang Scheduling for HPC Systems with Unreliable Broadcast Synchronization Mechanisms, SC23 Poster, 2023

キューを指定してジョブを投入するだけで利用可能

FUJITSU

運用実績

富士通 Quantum challenge

FUJITSU

日程

昨年 2月 21日に、Webサイトおよびシリコンバレー開催イベントにおいて募集を開始

多くの国や地域から 計 43 チームの応募

選出された 20 チームがコンテストに参加

$100,000 Prize

First Prize	**$50,000**
Second Prize	**$30,000**
Third Prize	**$20,000**

授賞式

今年1月に、オランダのデルフトで開催予定のFujitsu Quantum Day eventにて受賞者を発表

コンテストの実施など多くの方々にご利用いただいています

- Obtain feedback from Participants on performance and scalability

FUJITSU

まとめ

まとめ

- 量子コンピュータシミュレーション向けクラスタの構築
- クラスタ構築におけるOSS活用
- 量子シミュレーション向けスケジューリング技術
- 運用実績

Thank you