第31回コンピュータシステム・シンポジウム

FUJITSU
shaping tomorrow with you

スーパーコンピュータ「富岳」のシステムソフトウェアについて

富士通株式会社
Linuxソフトウェア事業部　Linux開発部　張　雷
2019.12.10

アウトライン

- 自己紹介
- 富岳システムの紹介
- 富岳のシステムソフトウェアについて

自己紹介

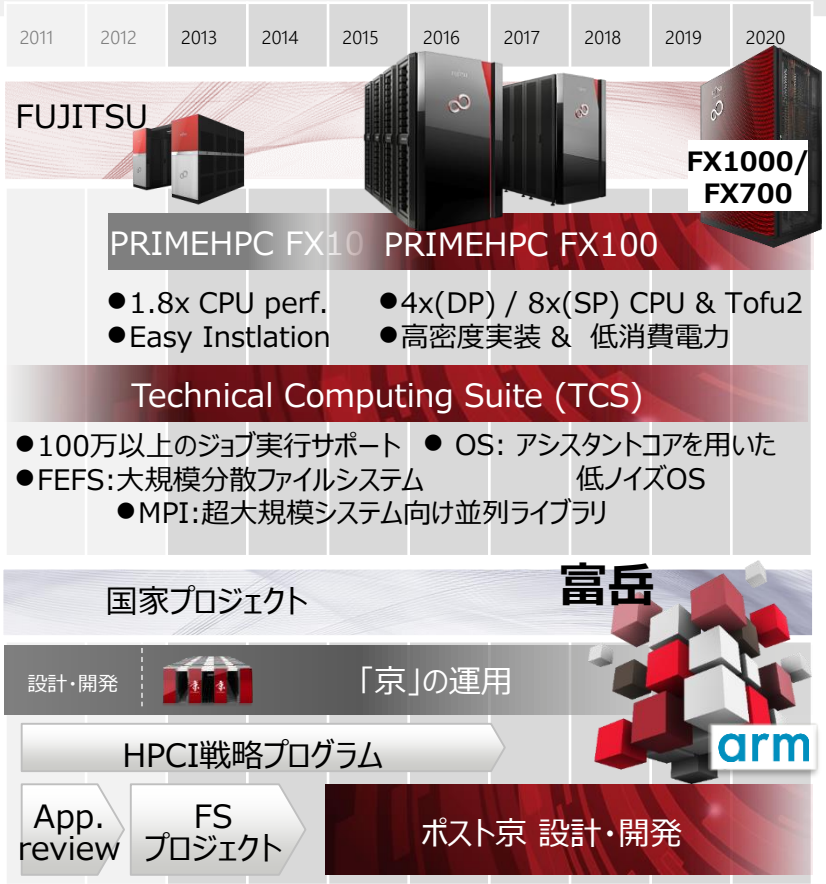
FUJITSU

- 中国生まれ、2005年日本へ
- 2008.3まで筑波大学　システム情報系和田研究室にて並列処理の研究に従事
- 2008.4より富士通株式会社に入社
- 2011.11よりPRIMEHPC FX10/FX100/富岳の開発に参加。HPC拡張機能(カーネルモジュール)の開発に関わり、Linux communityに参加。

FUJITSU

富岳システムの紹介

スーパーコンピュータ「京」とその後



- スーパーコンピュータ「京」
  - 10PFLOPSの実用・汎用スパコンを実現
  - システムソフト「TCS」の開発・提供により、ハードウェアの機能・性能を引き出す
  - 7年間運用され、多くのアプリが開発された
- TCSを進化させ種々の要件に対応
  - FX10, FX100を開発し、TCSで性能を引き出し、提供
  - x86クラスタとのハイブリッドシステム
- スーパーコンピュータ「富岳」を理研と開発

富岳の開発目標とアプローチ

FUJITSU

- 開発目標
  - 高いアプリケーション性能・高い電力効率
  - ユーザに対する利便性の高さ
  - 「京」で動作していたHPCアプリケーションに対する互換性の維持
- アプローチ
  - 高い性能とスケーラビリティ、独自CPUコアの開発

    【性能】　幅広いSIMD、数学演算命令、高いメモリバンド幅

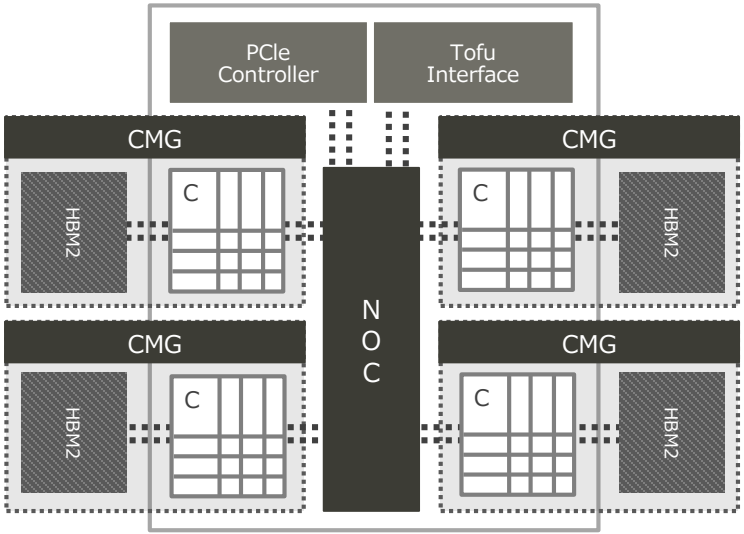
    【スケーラビリティ】　スケーラブルなTofuインターコネクトを強化

    【電力効率】　デバイステクノロジー、電力制御機能、最適な資源利用
  - Arm ISA採用によるArmバイナリ互換

富岳CPU“A64FX”概要

FUJITSU

- Arm SVEを採用した高性能・高効率CPU
  - 倍精度演算性能　>2.7 TFLOPS,　>90%@DGEMM
  - メモリバンド幅　1024 GB/s,　>80%@STREAM Triad



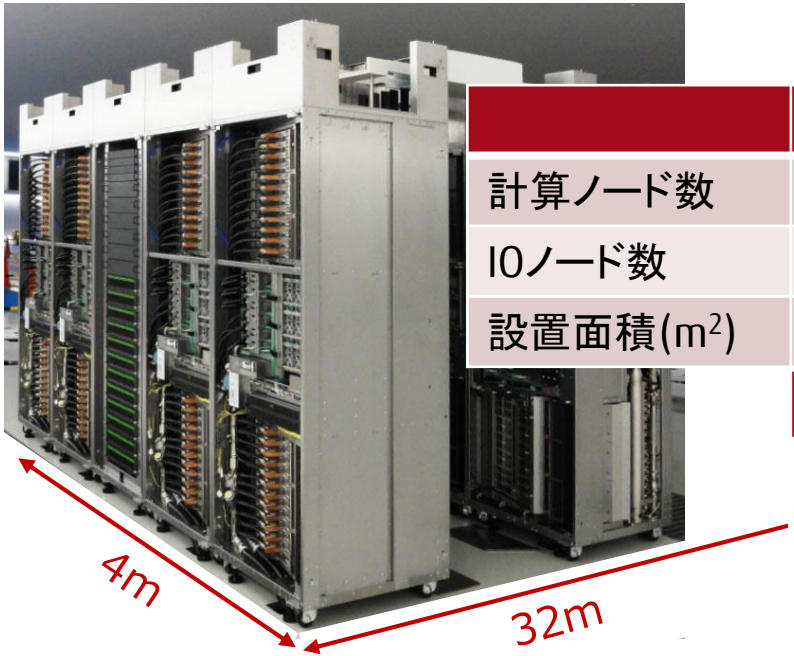
CMG : Core Memory Group　NOC : Network on Chip

	A64FX
ISA (Base, extension)	Armv8.2-A, SVE
プロセステクノロジ	7 nm
倍精度ピーク性能	>2.7 TFLOPS
SIMD幅	512-bit
コア数	48 + 4/2
メモリ容量	32 GiB (HBM2 x4)
メモリバンド幅	1024 GB/s
PCIe	Gen3 16 lanes
インターコネクト	TofuD integrated

京コンピュータと富岳の1ペタシステム

FUJITSU

- 「京」
  - 計算ラックx80とディスクラックx20



- 「富岳」
  - 1ラック（SSD含む）
  - 10ラック ＝「京」

	「京」	「富岳」
計算ノード数	7,680(=96x80)	384
IOノード数	4,80(=6x80)	
設置面積(m^2)	128(=4x32)	1.1
	SPARC Linux	Arm Linux

Arm Linuxのオープンソースコミュニティの活動、コラボレーションにより、多くのアプリ、システムソフトが利用可能になることを期待

Green500, Nov. 2019

A64FX prototype –

Fujitsu A64FX 48C 2GHz

ranked #1 on the list

768x general purpose A64FX CPU w/o accelerators

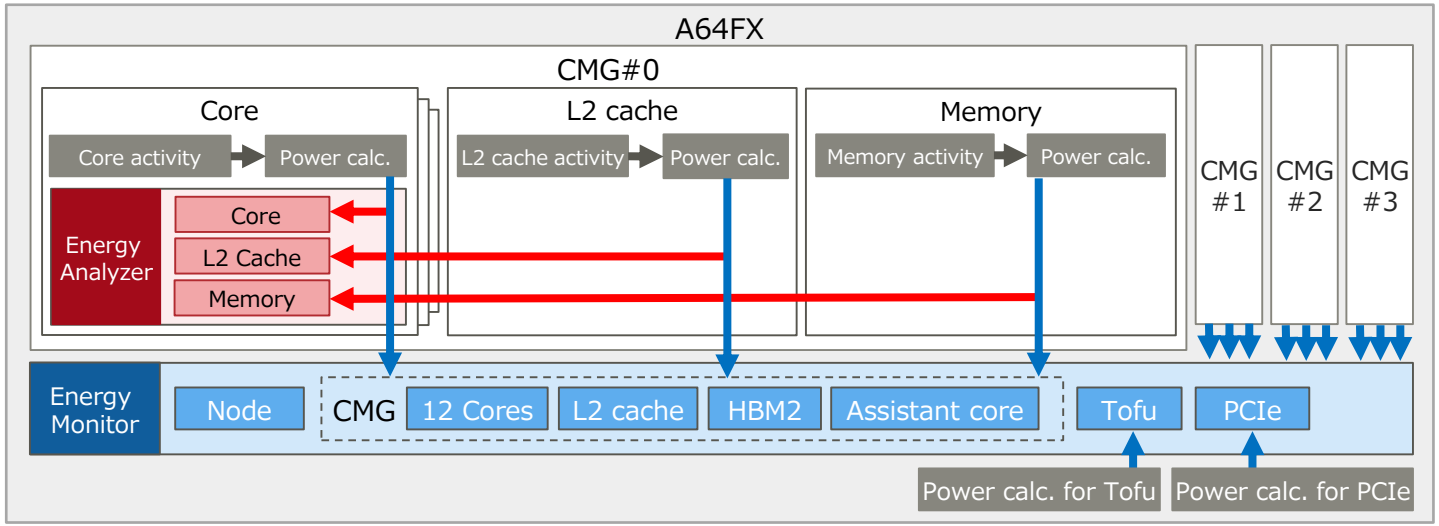
- 1.9995 PFLOPS @ HPL, 84.75%
- 16.876 GF/W
- Power quality level 2

エネルギーモニタ/アナライザ

- エネルギーモニタ(チップ単位)
  - ノード電力をPower API*1 で取得
  - ノード、CMG等の平均電力(~msec).
- エネルギーアナライザ(コア単位)
  - PAPI*2 による論理電力の取得
  - コア、L2キャッシュ、メモリなどの平均電力(~nsec)

*1: Sandia National Laboratory

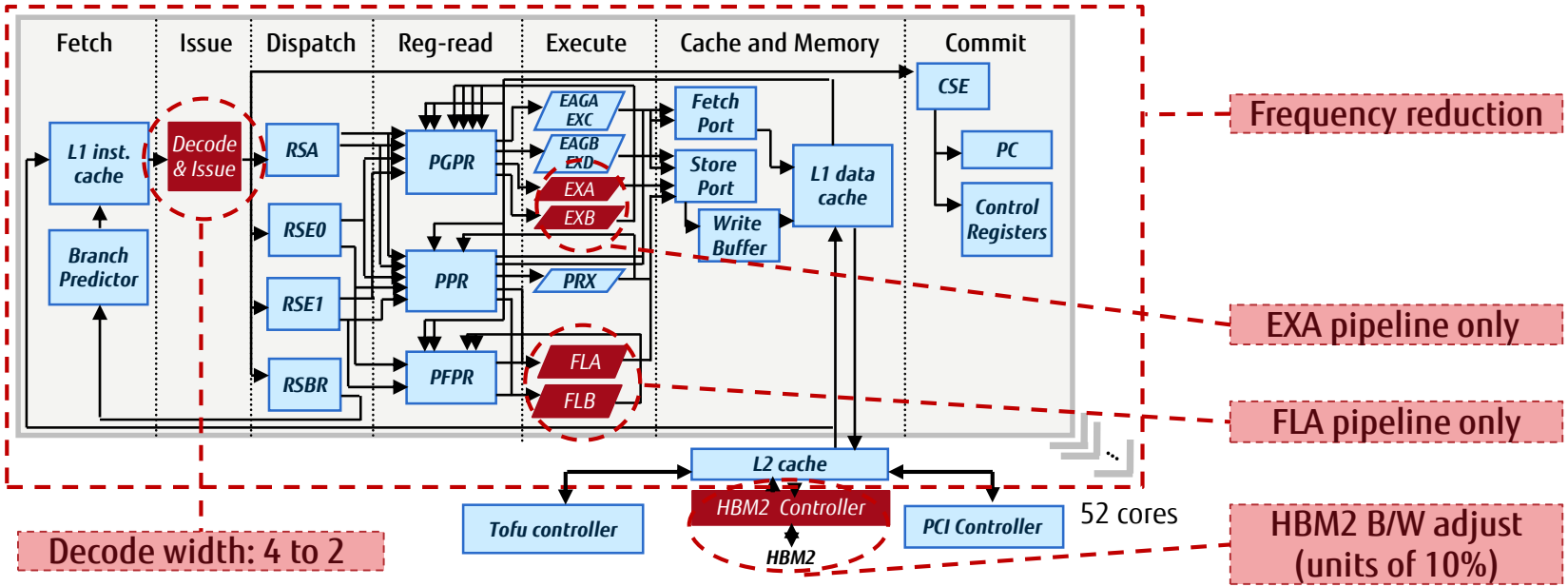
*2: Performance Application Programming Interface



パワーノブ（電力削減機能）

FUJITSU

- “パワーノブ”はユーザAPIで機能ユニットの動作を抑制することで、性能/電力を改善可能

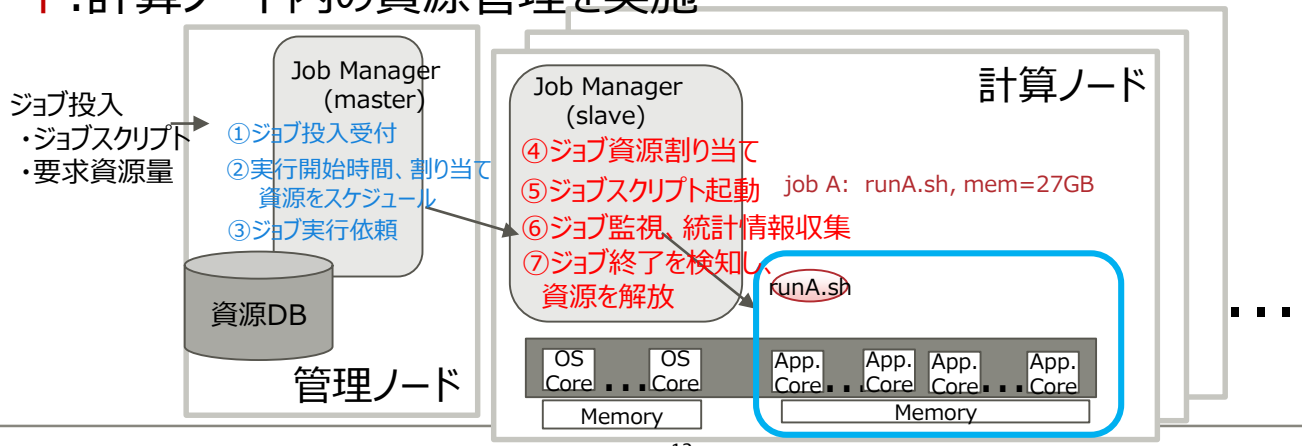


FUJITSU

富岳のシステムソフトウェアについて

前提知識

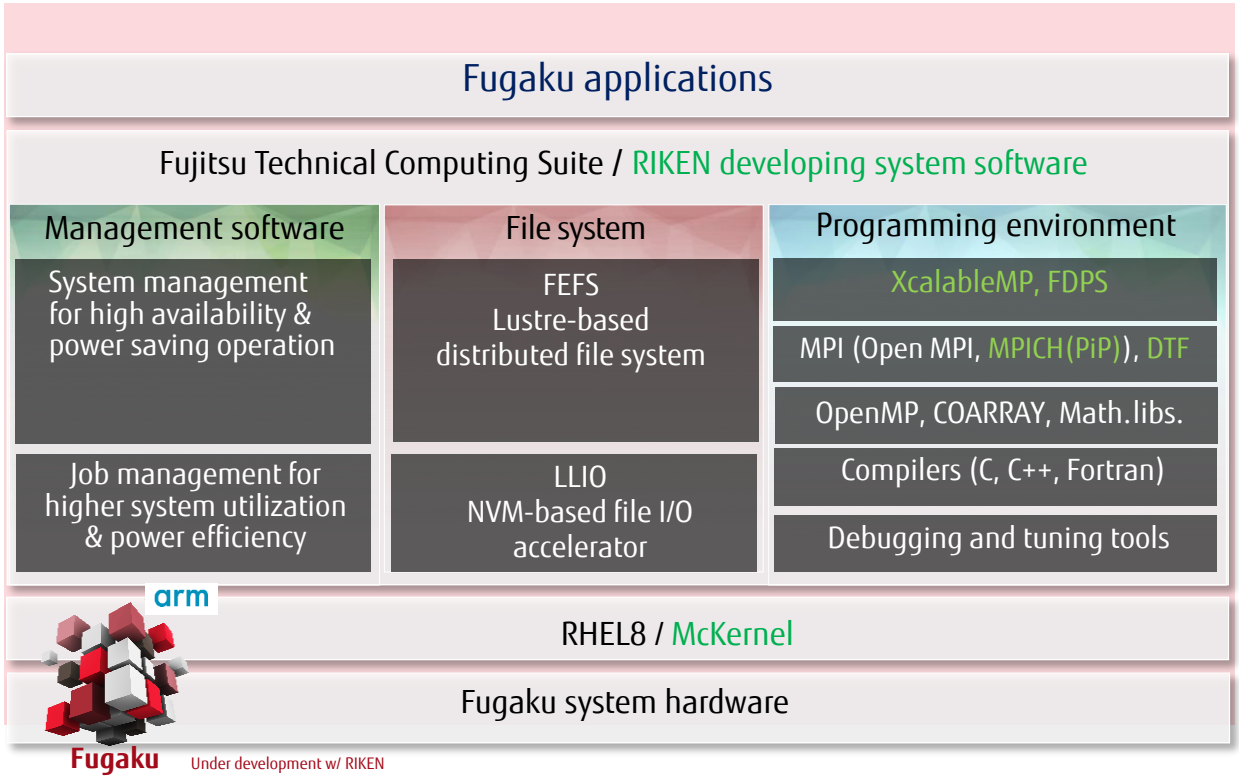
- スーパーコンピュータのセンター運用はバッチジョブシステム中心
- 共用環境で多数のユーザ要求を効率的に運用することが目的
  - ジョブ実行要求キューイングし順次バッチ実行
- 運用ソフトウェアの構成
  - 管理ノード:ジョブ要求の受付、ジョブスケジューリングを実施
  - 計算ノード:計算ノード内の資源管理を実施



富岳向けシステムソフトウェア

FUJITSU

■ 理研と富士通で開発中の富岳向けソフトウェアスタック



FS2020プロジェクトではAIプラットフォームとしてTensorlflow, Chainer Pytorchなども整備中

富岳のシステムソフトウェアの特徴

FUJITSU

- **多様なユーザニーズに応える**
  - 様々なOSSを活用できる汎用OSディストリビューション
  - LWK/コンテナなどの多様な実行環境のサポート
- **柔軟なセンター運用の実現に向けて**
  - 最大電力を抑えるための電力制限ジョブスケジューリング
  - ジョブスケジューラ・コマンドをセンター毎にカスタマイズできるAPI
  - 運用中のパッチ適用を可能とするローリングアップデート
- **アプリケーションの高速な実行のために**
  - 安定した高速な共有ファイルシステム
  - システム性能を十分に引き出すHPC拡張機能
  - A64FXの性能を最大限引き出すコンパイラ
  - 大規模並列でもスケールするMPI実行環境
  - アプリケーションの高速化の手助けをするプロファイラ

富岳のシステムソフトウェアの特徴

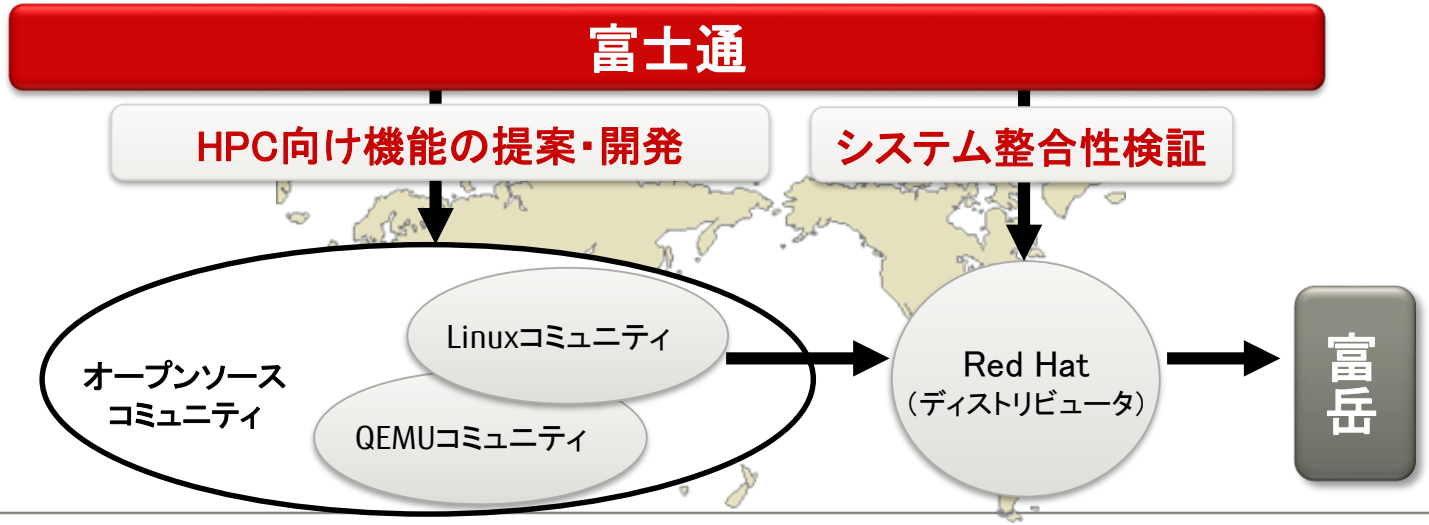
FUJITSU

- **多様なユーザニーズに応える**
  - **様々なOSSを活用できる汎用OSディストリビューション**
  - LWK/コンテナなどの多様な実行環境のサポート
- **柔軟なセンター運用の実現に向けて**
  - 最大電力を抑えるための電力制限ジョブスケジューリング
  - ジョブスケジューラ・コマンドをセンター毎にカスタマイズできるAPI
  - 運用中のパッチ適用を可能とするローリングアップデート
- **アプリケーションの高速な実行のために**
  - 安定した高速な共有ファイルシステム
  - システム性能を十分に引き出すHPC拡張機能
  - A64FXの性能を最大限引き出すコンパイラ
  - 大規模並列でもスケールするMPI実行環境
  - アプリケーションの高速化の手助けをするプロファイラ

標準OSディストリビューションの採用(「富岳」新規)

FUJITSU

- 「京」では、独自OSを採用し、Linuxカーネルやライブラリ、ツールまで独自に開発
  - SPARCアーキのCPUではOSSのエコシステムが活発ではなく、独自開発せざるを得なかった
- 「富岳」ではRHEL8を採用
  - Armv8 Linux標準とバイナリ互換： OpenHPC, SPACKなどArm HPCエコシステム充実
    - PCクラスタと同じ使い勝手で「富岳」を利用でき、Python/Ruby等のスクリプト言語で運用性向上
  - OSSコミュニティ（e.g. Linux, QEMU）上で機能開発し、RHELに取り込む➔エコシステムにも貢献



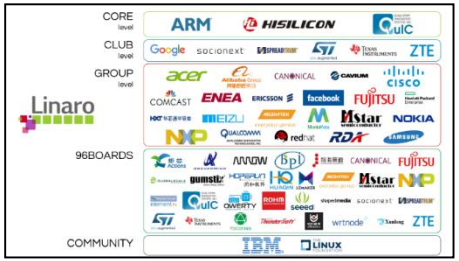
ARM HPCエコシステム形成に向けた連携

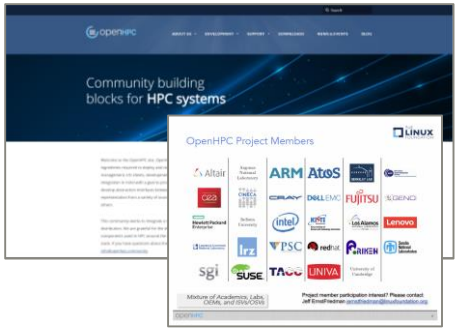
FUJITSU

- ARM
  - Linux, GCCのSVE対応他、OpenHPCのARM対応など
    積極的にARM HPCエコシステム形成に貢献
    https://developer.arm.com/hpc
- Linaro
  - ARM基本ソフト(Linux Kernel, glibc, GCC他)の標準化とUpstream
  - ARM HPCにおけるバイナリレベルでのポータビリティ確保
  - SVE対応ソフトのOSSへのUpstream、普及推進
    https://www.linaro.org/sig/hpc/
- OpenHPC
  - PCクラスタソフトの標準化
    （IAとARM間のソフトウェアポータビリティ確保による相互補完）

  http://www.openhpc.community/
- 富士通
  - HPCの経験・技術をコミュニティに展開

    →ARM HPCエコシステムの環境整備
  - 最新で強力なコンパイラの早期整備・展開と利用環境の整備





富岳のシステムソフトウェアの特徴

FUJITSU

- **多様なユーザニーズに応える**
  - 様々なOSSを活用できる汎用OSディストリビューション
  - **LWK/コンテナなどの多様な実行環境のサポート**
- **柔軟なセンター運用の実現に向けて**
  - 最大電力を抑えるための電力制限ジョブスケジューリング
  - ジョブスケジューラ・コマンドをセンター毎にカスタマイズできるAPI
  - 運用中のパッチ適用を可能とするローリングアップデート
- **アプリケーションの高速な実行のために**
  - 安定した高速な共有ファイルシステム
  - システム性能を十分に引き出すHPC拡張機能
  - A64FXの性能を最大限引き出すコンパイラ
  - 大規模並列でもスケールするMPI実行環境
  - アプリケーションの高速化の手助けをするプロファイラ

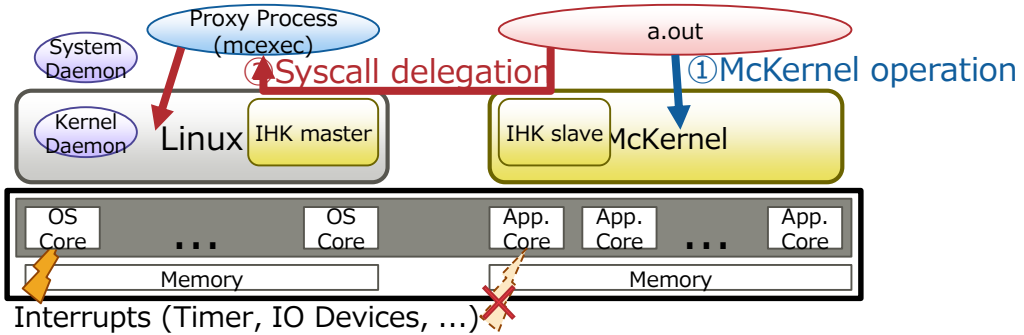
多様な実行環境のサポート

FUJITSU

- 複数のジョブ実行モードの中から最適な環境をユーザ自身が選択可能
- 通常モード
  - デフォルトのジョブ実行環境モード
  - システムにインストールされたRHEL上でジョブプログラム※が実行可能
    ※Singularityも実行可能
- McKernelモード
  - McKernel上でアプリケーションを実行するモード
  - ノイズレス実行環境でジョブプログラムが実行可能
- Docker/KVMモード
  - 仮想環境でアプリケーションを実行するモード
  - クラウド上に流通するDocker/KVMイメージを実行可能

富岳で採用するLWK:IHK/McKernel

- McKernelは理研が開発するLWKであり、Linuxと互換性を持つ
  - LWK(Light Weight Kernel)とは通常のOS機能から最低限の機能を抽出した軽量OS
- McKernelの特徴
  - Linuxとは独立したCPUとメモリで動作
  - アプリケーションでよく使われるAPIのみ実装し、軽量化
  - 割り込みや大部分のシステムコールなどの重い処理はLinux側で代行処理



富岳のシステムソフトウェアの特徴

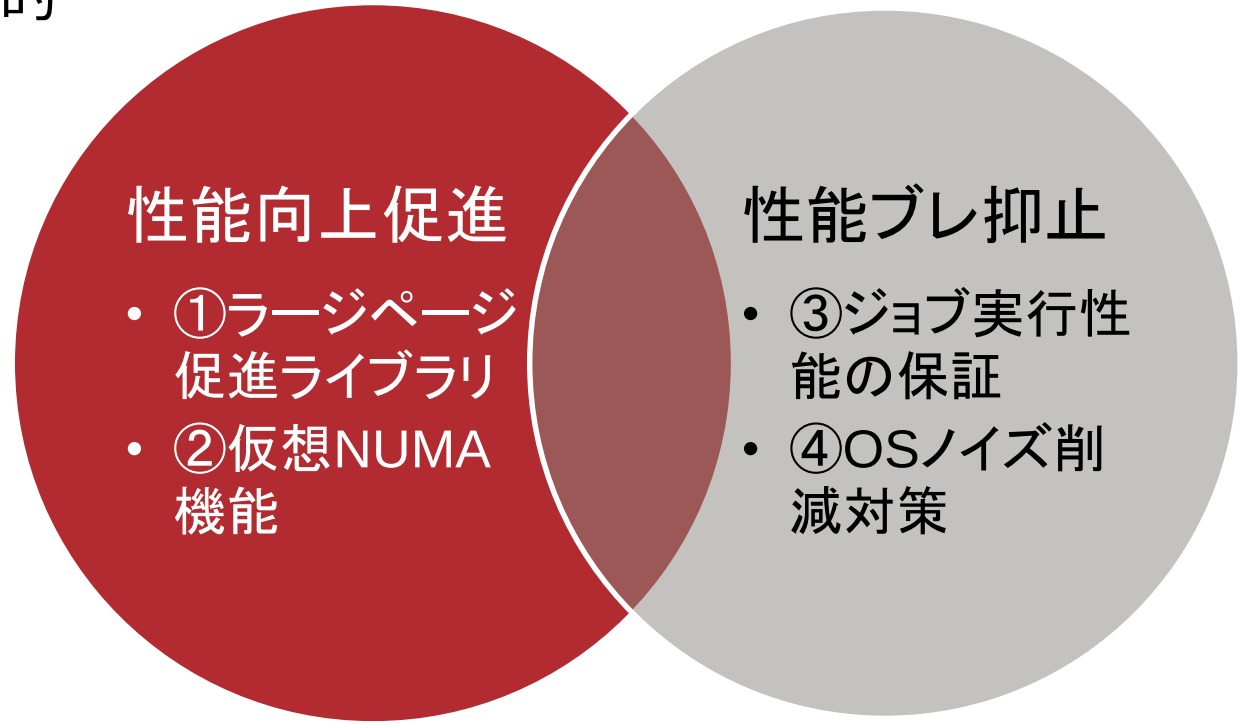
FUJITSU

- **多様なユーザニーズに応える**
  - 様々なOSSを活用できる汎用OSディストリビューション
  - LWK/コンテナなどの多様な実行環境のサポート
- **柔軟なセンター運用の実現に向けて**
  - 最大電力を抑えるための電力制限ジョブスケジューリング
  - ジョブスケジューラ・コマンドをセンター毎にカスタマイズできるAPI
  - 運用中のパッチ適用を可能とするローリングアップデート
- **アプリケーションの高速な実行のために**
  - 安定した高速な共有ファイルシステム
  - **システム性能を十分に引き出すHPC拡張機能**
  - A64FXの性能を最大限引き出すコンパイラ
  - 大規模並列でもスケールするMPI実行環境
  - アプリケーションの高速化の手助けをするプロファイラ

システム性能を引き出すHPC拡張機能

FUJITSU

■ 開発目的



①ラージページ促進ライブラリ

FUJITSU

- HPCではメモリアクセスを高速化するため、ページサイズを大きくして、使うことが一般的
  - Linuxのラージページの実装としてはTHPとHugeTLBfsがある
- ラージページを確実に獲得かつ使い勝手が良くするため、ラージページ促進ライブラリを開発
  - ラージページ促進ライブラリでglibcが提供しているメモリ獲得・解放に関わるAPIの機能を置換し、ユーザに意識させることがなく、ラージページを積極的に獲得する

	THP	HugeTLBfs	ラージページ促進ライブラリ
ラージページ獲得保証	△	○	○
ユーザの使い勝手	○	△	○

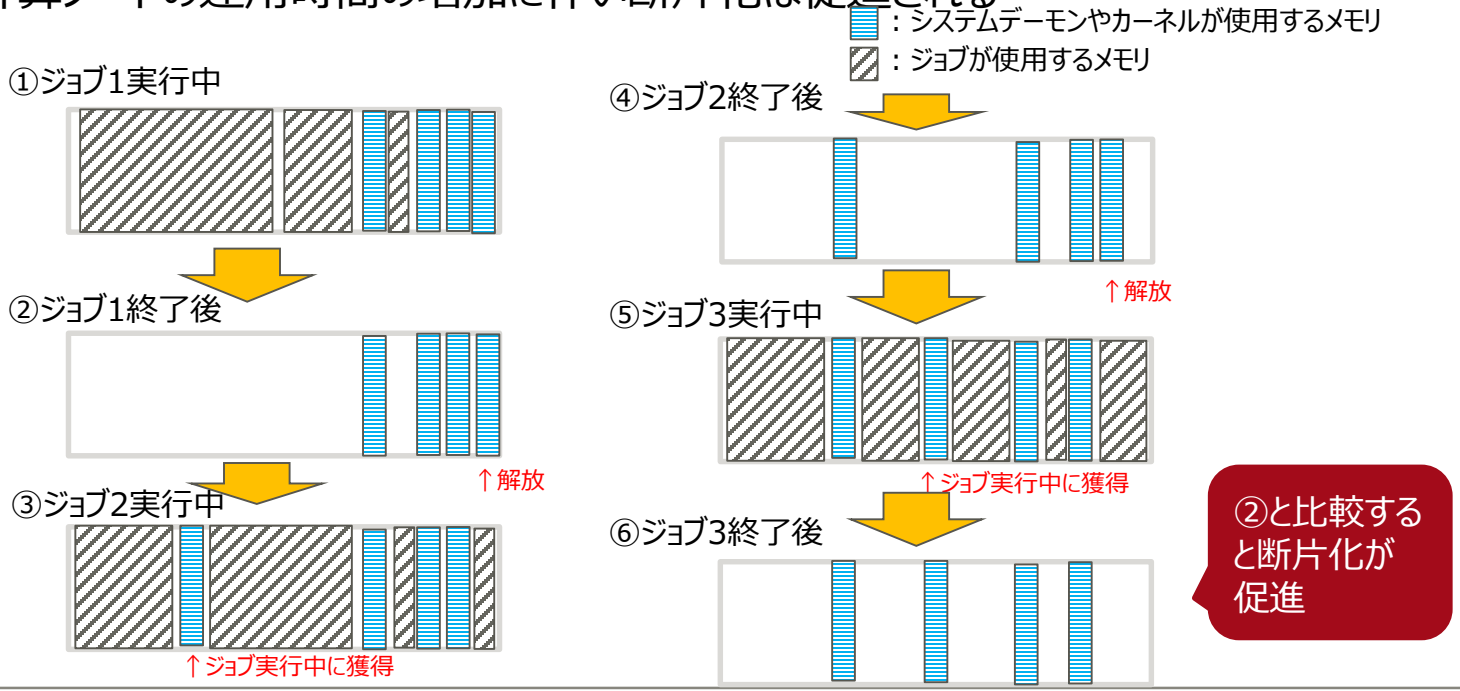
- ラージページ促進ライブラリの効果
  - 20GiBのメモリを獲得するアプリケーションでmallocによるラージページの獲得枚数を測定

→結果10245枚が獲得でき、20GiBのメモリはすべて2MiBのラージページであることを確認できた

②仮想NUMA機能 1/5

FUJITSU

- 課題：メモリの断片化により、ラージページを獲得にくい
  - ジョブ実行中に獲得されたシステムデーモンやカーネル処理のためのメモリは、ジョブ終了と連動して開放されない
  - 一般に計算ノードの運用時間の増加に伴い断片化は促進される



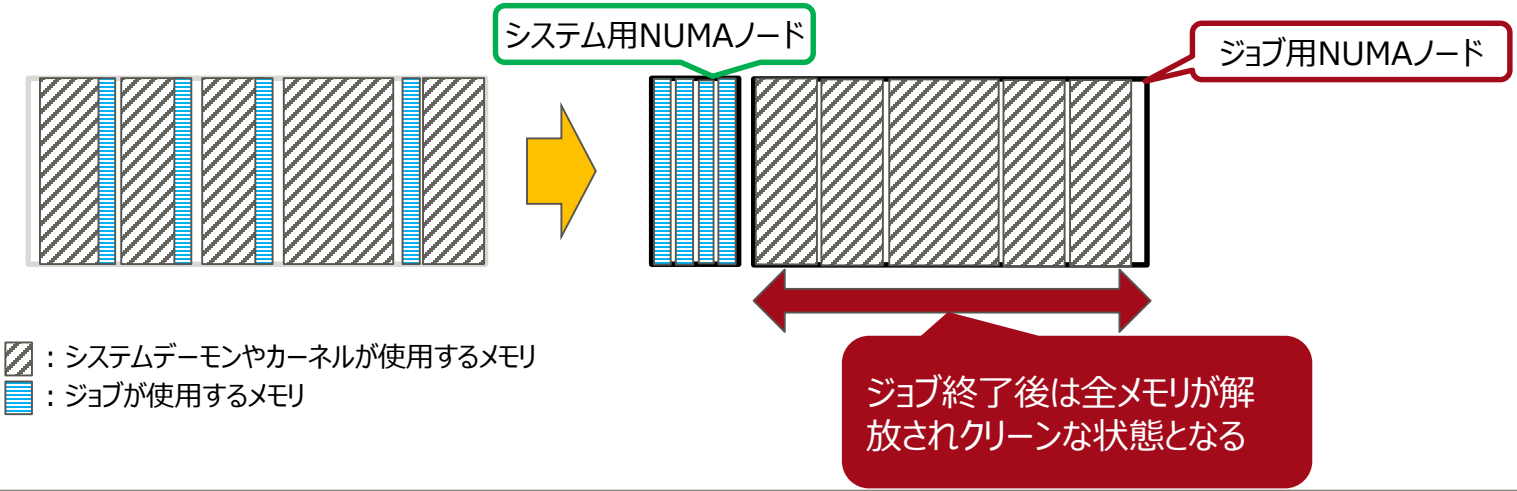
②と比較すると断片化が促進

②仮想NUMA機能 2/5

FUJITSU

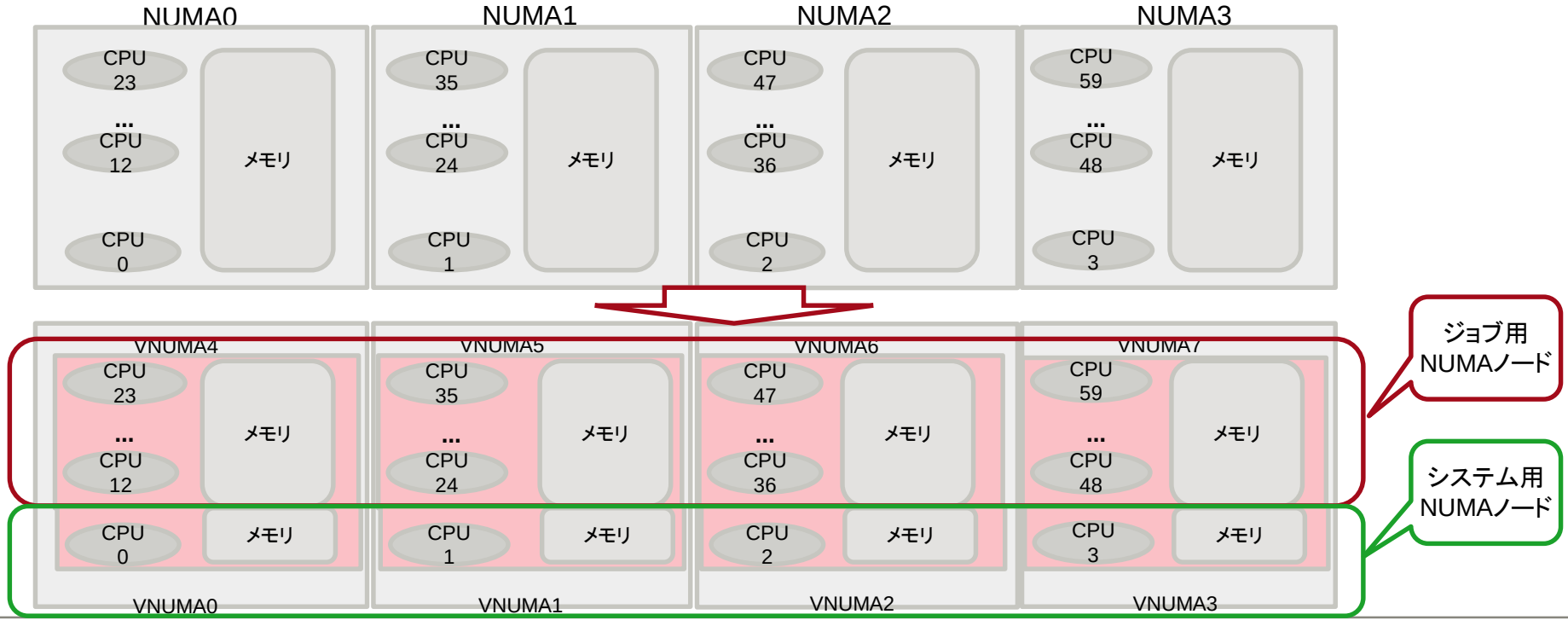
- 対策：ジョブに割り当てるメモリを独立したNUMAノードとして分割管理することで断片化を防止
  - NUMA構成情報を書き換え仮想的なNUMAノード構成を作成
    - LinuxはNUMAノード単位でメモリを管理するため、分割した各NUMA毎に割り当てポリシを変更可能
  - システムデーモンやカーネル処理用のメモリはシステム用NUMAノードから割り当て

⇒ ジョブ用NUMAのメモリをジョブに占有利用させることが可能



②仮想NUMA機能 3/5

- 実現イメージ：物理NUMAは4つ。仮想的に8つのNUMAを見せかける
  - ACPI TableのSRAT(各NUMA Nodeの構成要素を表現)/SLIT(NUMA Node間のメモリアクセスレイテンシを表現)を修正することで実現



②仮想NUMA機能 4/5

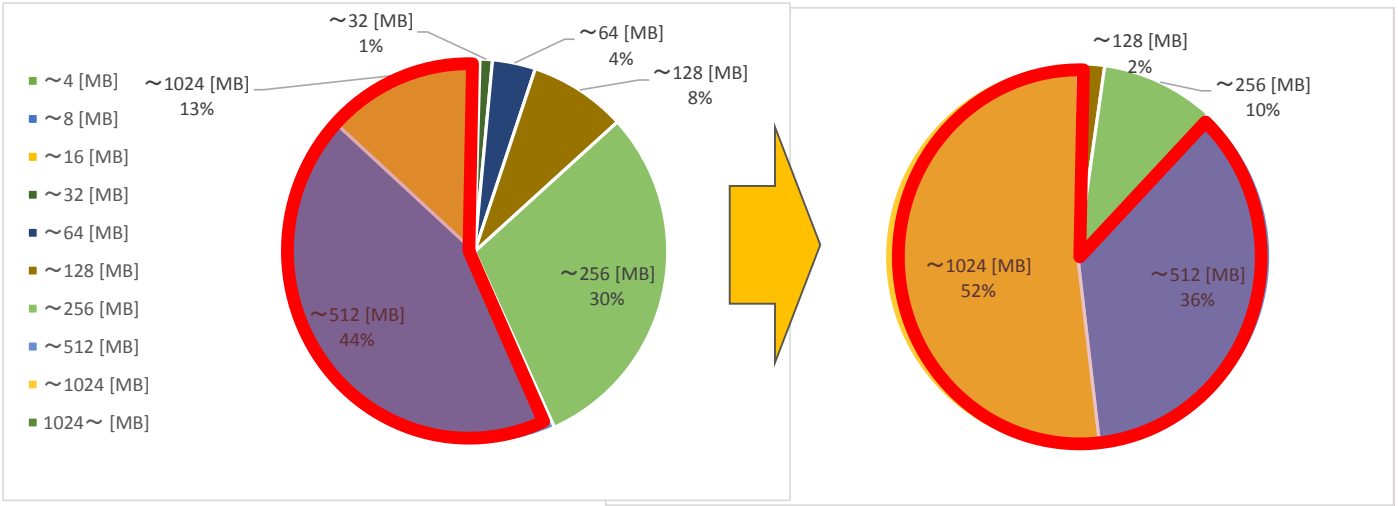
FUJITSU

- 断片化検証テスト: メモリ高負荷状態での運用を擬似的に再現
  - 擬似システムデーモンを動作させ、システム用メモリの最大95%を消費
    - malloc/freeを繰り返すプロセス×2: アノニマスメモリ、カーネルメモリを消費
    - ファイルIOを繰り返すプロセス×1: ファイルキャッシュ、カーネルメモリを消費

    → ジョブと連動しないシステムデーモン、カーネル処理が頻繁に動作し、断片化が促進される状態となる
  - ジョブ実行を繰り返し、ジョブ用メモリも最大95%消費し高負荷状態とする
    - 開始時にmallocで大規模な領域を獲得し終了時に開放するプロセス×4
    - ファイルIOを繰り返すプロセス×1
  - 12時間実行し、終了時の物理メモリの状態を確認

②仮想NUMA機能 5/5

- 断片化検証テスト結果：256MB以上の連続メモリ領域の割合が、57%から88%に改善
  - 断片化の抑止に効果があることを確認できた
  - 測定環境：Intel(R) Xeon(R) CPU E5520 @ 2.27GHz

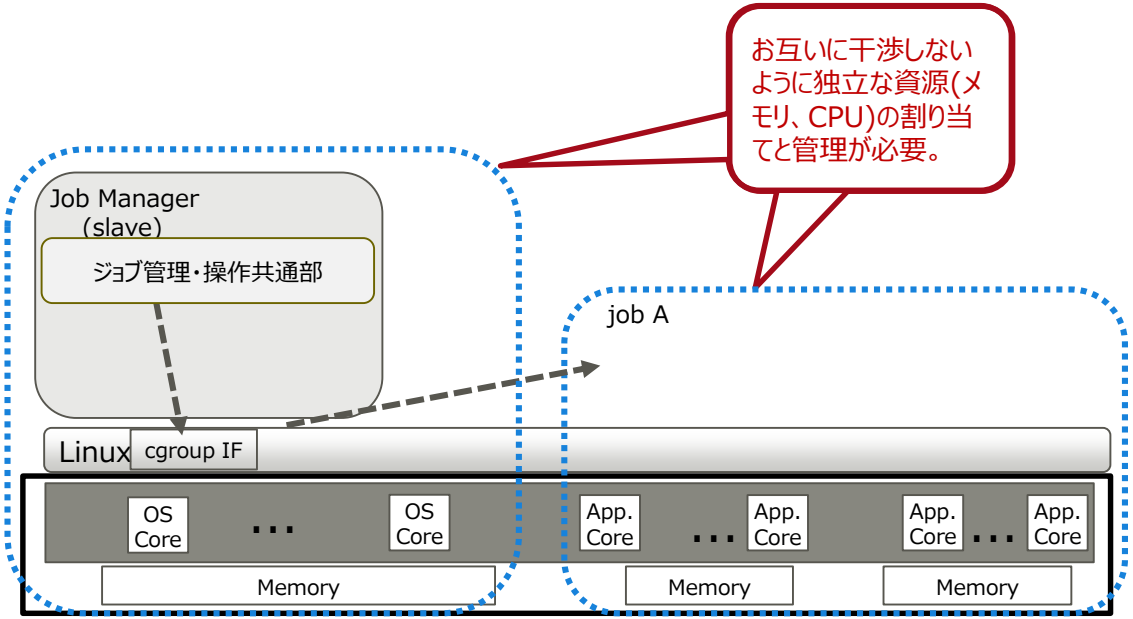


対策実施前　　対策実施後

③ジョブ実行性能の保証　1/2

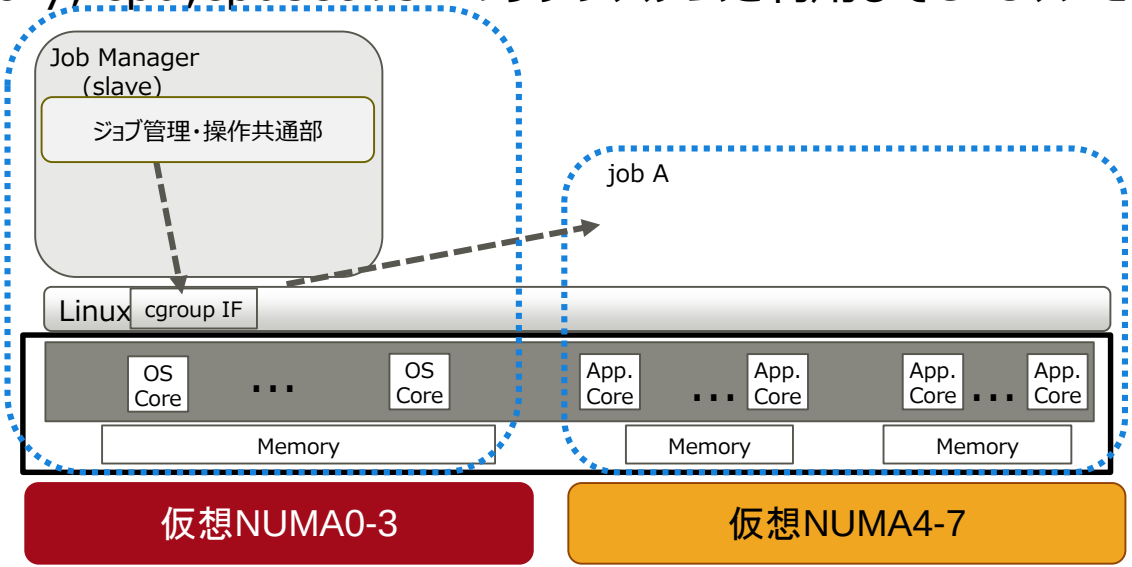
FUJITSU

- 課題：システム運用するための資源とジョブ資源を干渉しないように管理する必要がある



③ジョブ実行性能の保証　2/2

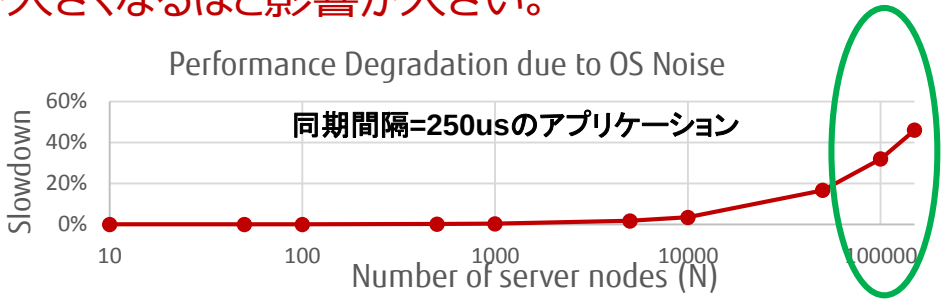
- 対策：仮想NUMA機能+Linux標準の資源管理機能cgroupsで制限が可能
  - プロセス、スレッド群をグループ化し、資源制限、統計情報の取得、優先度制御などの資源管理が可能
  - memory, cpu,cpusetの3つのサブシステムを利用してCPU、メモリ資源を管理



④OSノイズ削減対策

FUJITSU

- OSノイズとは
  - 本来ジョブの実行とは関係なしシステムデーモン、カーネルデーモン、割込み処理によるジョブの性能劣化/ブレ。システム規模が大きくなるほど影響が大きい。



- OSノイズ低減の方針
  - 同期間隔1ms の16万並列アプリケーションの性能低下率が 1～5%以下となること
  - 更に細粒度で同期するアプリケーションにはMcKernelによるノイズレス環境で対応
- OSノイズの低減対策
  - アシスタントコアの活用： OS動作(割込/デーモン等)をアシスタントコアにオフロード
  - Ticklessカーネルの採用：Linuxのノイズ削減機能であるnohz_fullの利用

New PRIMEHPC Lineup

FUJITSU

PRIMEHPC FX1000

Supercomputer optimized for large scale computing

High Performance | High Scalability | High Density

A64FX processor
384 nodes/Rack
Tofu Interconnect D
Water Cooling
Fujitsu Software Stack
for Supercomputing

PRIMEHPC FX700

Supercomputer based on standard technologies

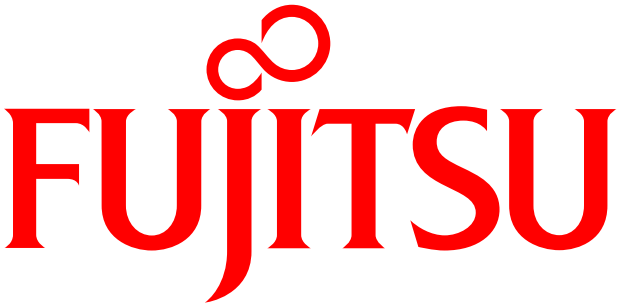
Ease to use | Installation

A64FX Processor
8 nodes/2U Rackmount
InfiniBand
Air Cooling
Utilize ISV and Open Source Software Stack

まとめ

FUJITSU

- 今日紹介したこと
  - 富岳システムの紹介
    - Green500で2019.11に世界一達成
  - 富岳のシステムソフトウェアについて
    - **多様なユーザニーズに応える**
    - **柔軟なセンター運用の実現に向けて**
    - **アプリケーションの高速な実行のために**
- 最新情報
  - 12/3から富岳の搬入開始



shaping tomorrow with you