FUJITSU
shaping tomorrow with you

Fujitsu Software Technical Computing Suite ジョブスケジューラーのご紹介

富士通株式会社
次世代テクニカルコンピューティング開発本部
三鴨　利彰

目次

- Technical Computing Suite とは
- ジョブスケジューラーの役割と要件
- TCS ジョブスケジューラーの特徴
- 最新状況

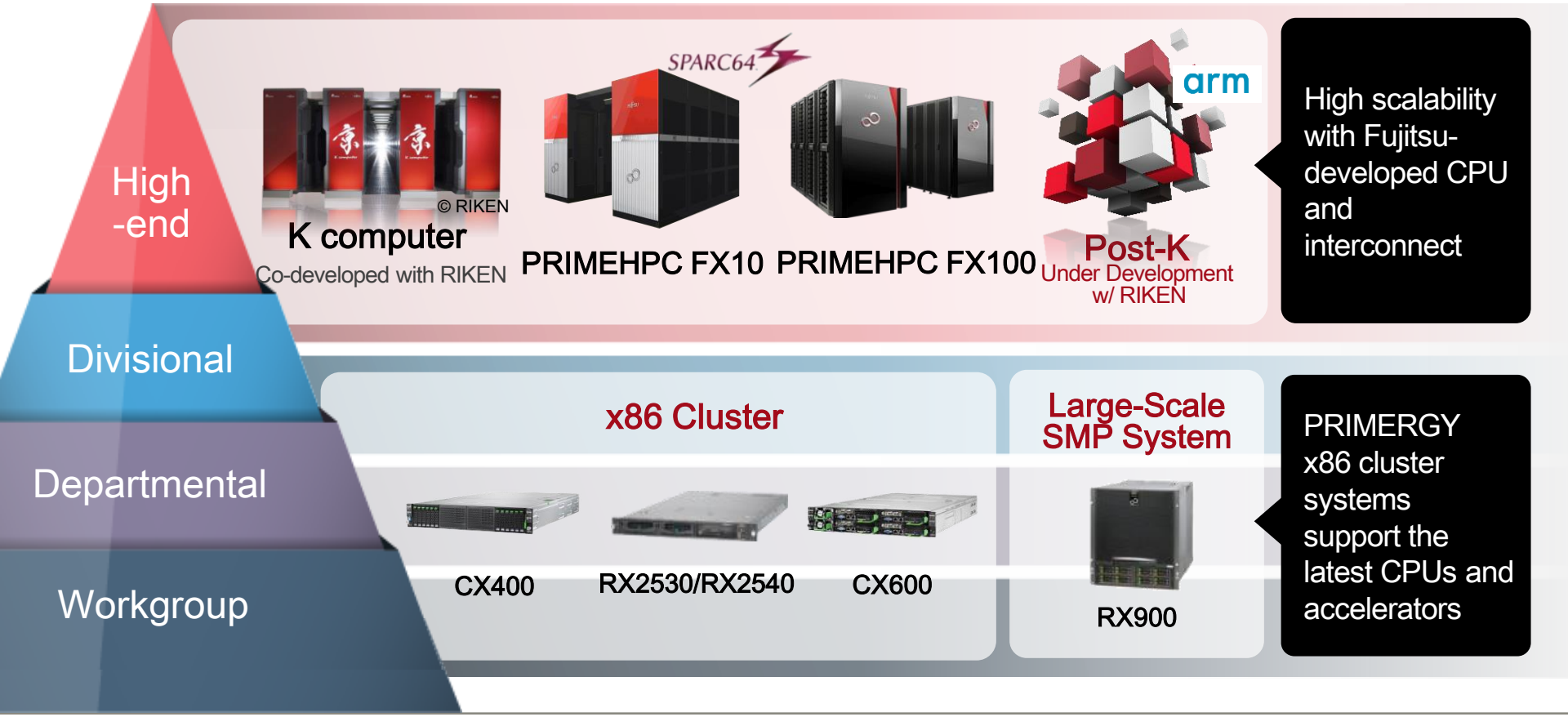
FUJITSU

Technical Computing Suite とは

お客様の要求に応える HPC ソリューション

FUJITSU

- 富士通スパコン(PRIMEHPC)と PCクラスタの両方をサポート
- システムソフトウェア(Technical Computing Suite)でシングルシステムイメージで利用
- 高性能、高可用性、高信頼性の追求



弊社導入システム

- スーパーコンピュータ「京」をはじめ、大規模システムでの確かな運用実績
- スーパーコンピュータ「京」はTOP500(2018.06)で16位になるも、今なお世界第2位のシステム規模(88,128台) ※TOP500リストから算出

Site	Computer	Name
最先端共同HPC基盤施設	PRIMERGY	Oakforest-PACS
理化学研究所 計算科学研究センター	K computer	
九州大学情報基盤センター	PRIMERGY	ITO - Subsystem A
宇宙航空研究開発機構(JAXA)	FX100	SORA-MA
名古屋大学情報基盤センター	FX100	
核融合科学研究所	FX100	Plasma Simulator
理化学研究所 情報システム部	PRIMERGY	HOKUSAI BigWaterfall
理化学研究所 情報システム部	FX100	HOKUSAI GreatWave
気象庁気象研究所	FX100	

2018.06 TOP500 からピックアップ

Technical Computing Suite（TCS）

FUJITSU

■ HPCで必要となるソフトウェア群をパッケージングした統合製品

アプリケーション

FUJITSU Software Technical Computing Suite

運用管理ソフトウェア	ファイルシステム	プログラミング環境
システム運用管理	Lustreベースの 分散ファイルシステム （FEFS）	MPI (Open MPI)
ジョブ運用管理 (ジョブスケジューラー)		OpenMP, COARRAY, Math Libs
		Compilers (C, C++, Fortran)
		Debugging and tuning tools

Linux OS

PRIMEHPC FX10/FX100

PRIMERGY (PCクラスタ)

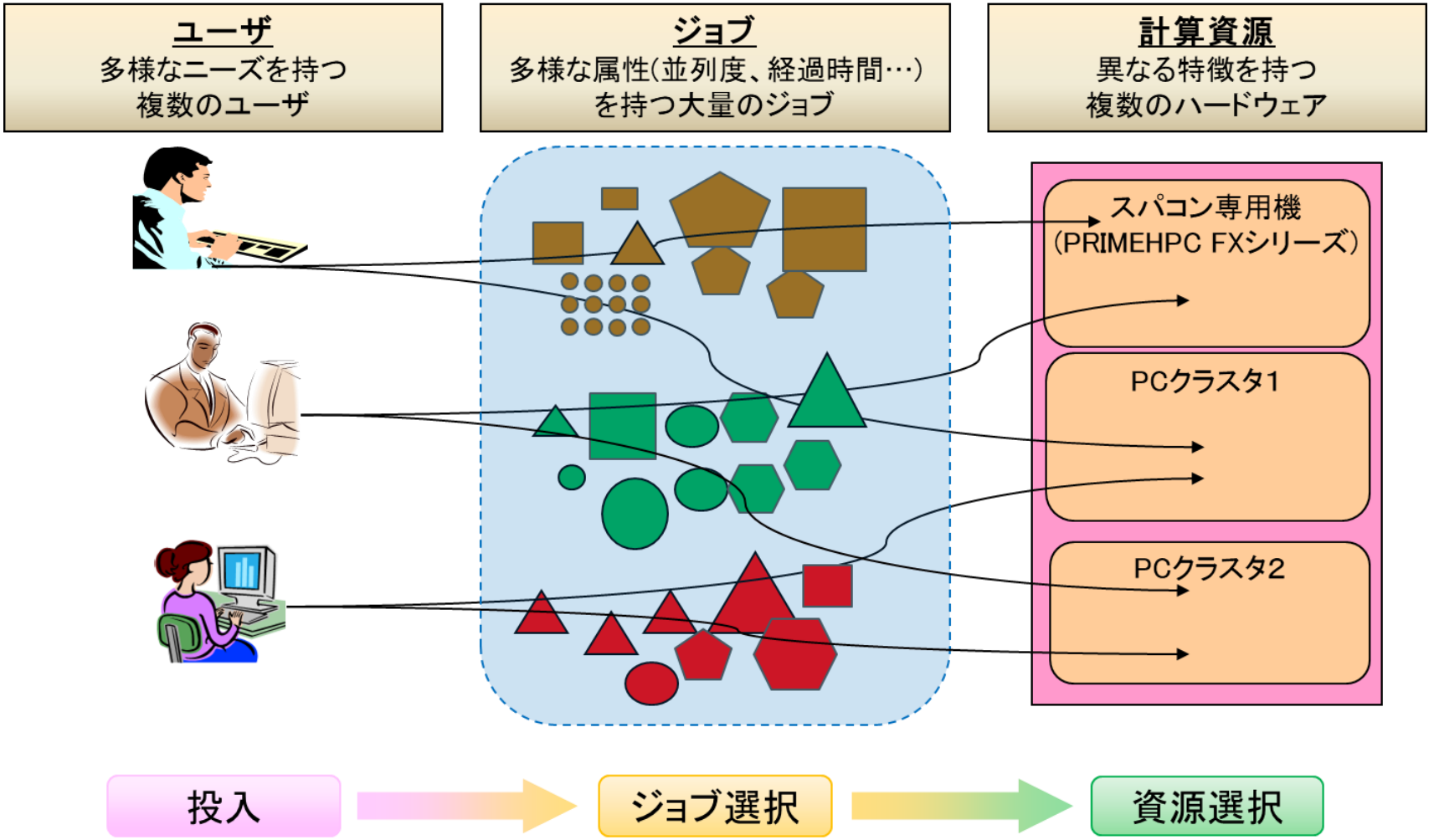
FUJITSU

ジョブスケジューラーの役割と要件

ジョブスケジューラーの役割

FUJITSU

■ 多種多様なニーズを持つ複数のユーザに対して、
計算資源を有効的に共同利用させること



システム管理者からの要件

FUJITSU

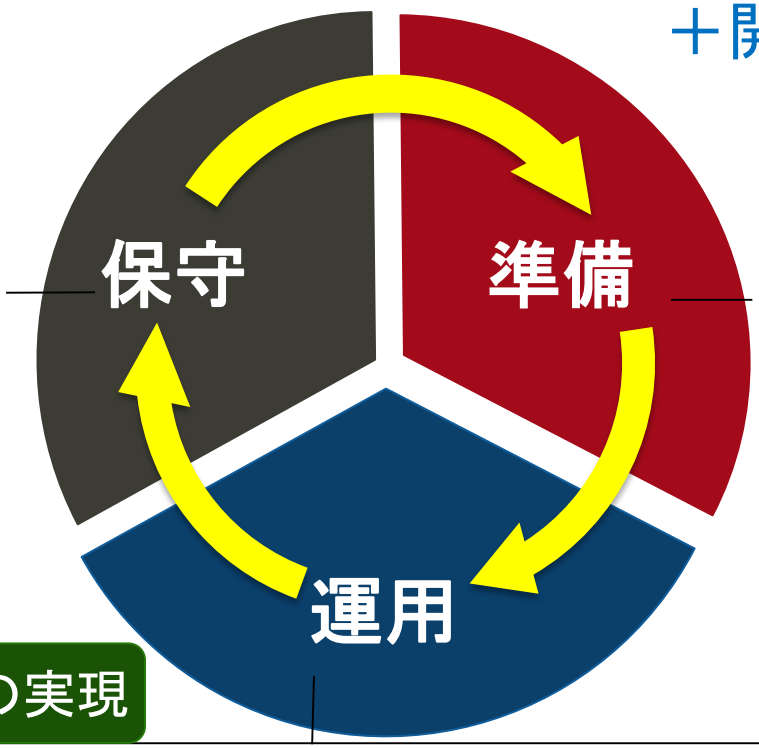
【柔軟な運用】　センター／ユーザからの多種多様な要件への対応

【高スループット】システムリソースを最大限に発揮

【サポート力】　運用分析・改善への支援➔弊社SEの運用サポート＋開発元の技術サポート

安心サポート

＜運用分析＞
・運用状況(利用状況/稼働率etc) の分析
・課題への対策立案



柔軟な運用設定

＜運用設定＞
・スケジューラの設定
(キュー/資源制限 etc.)
・センタ固有の処理の組込
(改札制御 etc)
・・・

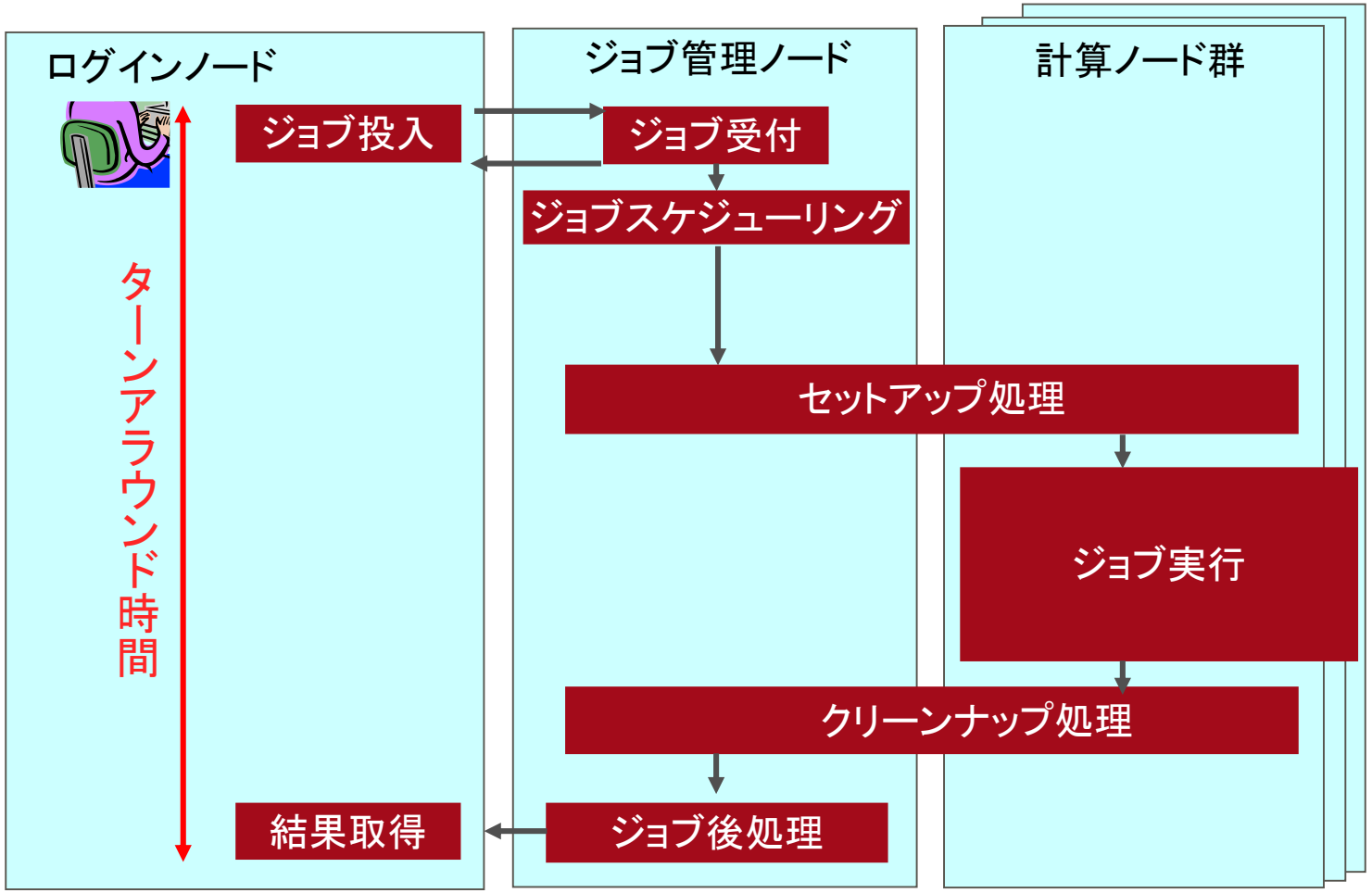
高スループットの実現

＜実行・監視・支援＞
・計画に従い実行(運用切替etc.)　・利用状況/稼働率の監視
・ユーザの利用支援 (問合せ対応/教育)

ユーザからの要件

【高いターンアラウンド性能】ジョブ投入から結果取得までの時間短縮

【使いやすさ】 使い勝手の良い操作性



FUJITSU

TCS ジョブスケジューラーの特徴

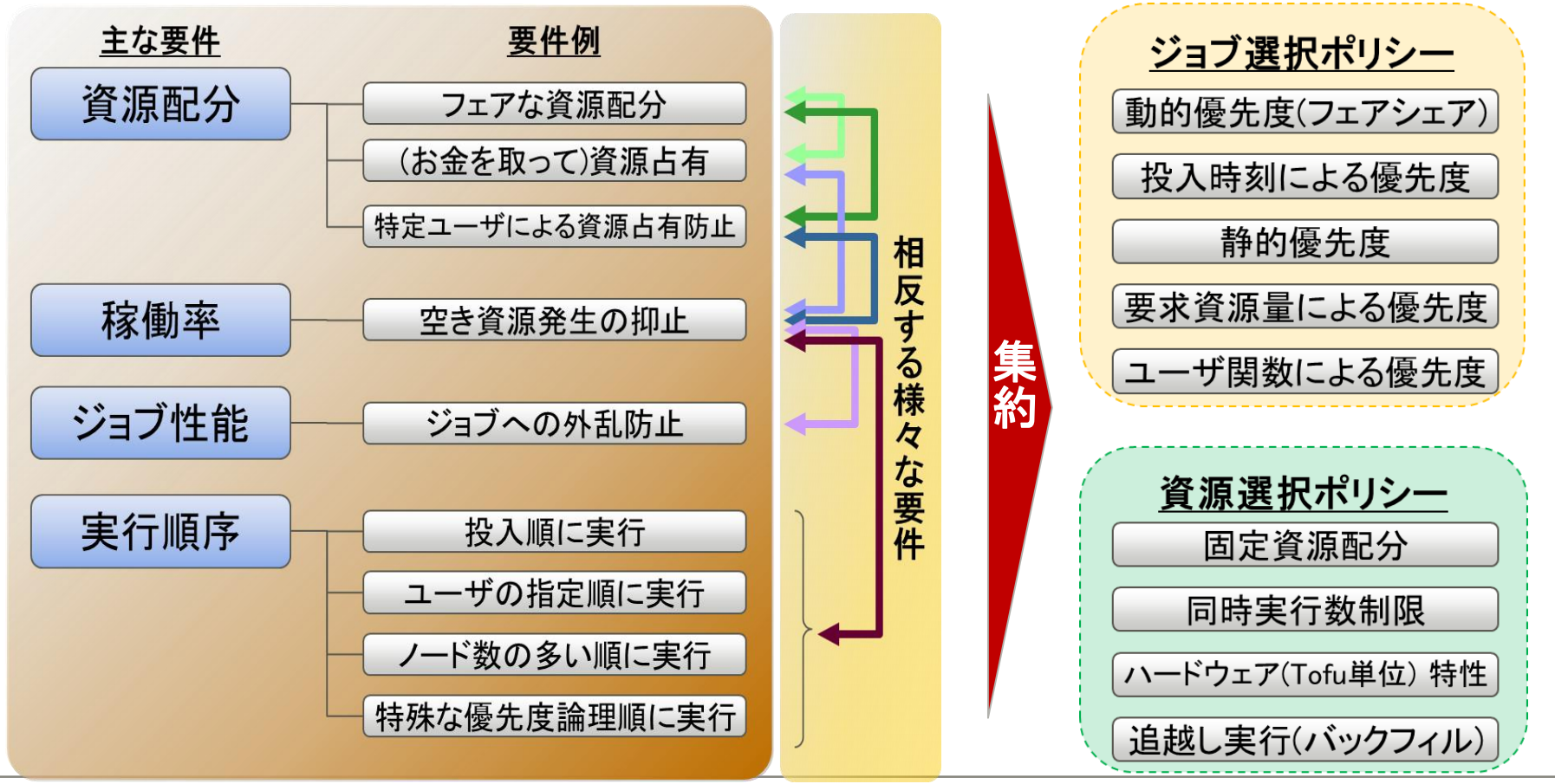
柔軟な運用 （1／2）

FUJITSU

豊富なスケジューリングポリシー

■ 相反する要件を「ジョブ選択ポリシー」「資源選択ポリシー」に集約

運用方針に合わせて自由に選択



柔軟な運用 （2／2）

FUJITSU

きめ細かな制御を可能とするジョブACL

- ユーザ・グループ・リソースグループ(キュー)の利用資源量の制限（下限、上限、デフォルト値）

  76種類の制限値を用意

  - 利用制限したい資源をカスタマイズも可能(ISVライセンス等)
- ユーザ・グループごとにスケジューラ機能の利用許可を定義

  運用の安全確保

センター固有の要件を実現する豊富なAPI

- ジョブスケジューラー内部の動作ポイント(ジョブ状態遷移等)でセンター固有の処理を組み込めるフック機能

  予算チェック・改札制御
- 外部のプログラムにジョブ状態遷移やスケジューリング結果を通知するイベント通知機能

  ジョブや空き資源の外部監視

高スループット ＆ 高いターンアラウンド性能（1／3）

FUJITSU

大規模システムの厳しいワークロードに耐える

- ジョブの受付から終了までのあらゆる処理をマルチプロセス、マルチスレッド化して並列処理

徹底した分散並列処理

- ジョブ管理ノードの階層構造を利用してジョブを高速起動

システムサイズ	10万ノード / 100万コア規模
ジョブ投入性能	3ミリ秒
スケジューリング性能	2,500ジョブ/秒　※1ノードジョブ
大規模MPI起動	数秒 (万オーダーのプロセス生成)

ジョブのアイソレーションで最高性能を発揮

- 計算ノードの資源(CPU、メモリ、ページキャッシュ、GPGPU等)をNUMA構成を考慮してジョブに専用割り当て

計算資源を占有利用

- ネットワークトポロジーを意識した計算ノード割り当て

高スループット （2／3）

FUJITSU

高稼働率下での緊急ジョブ実行

- 実行中ジョブの計算資源を一時的に解放(スワップアウト)して他のジョブ（業務ジョブ等）を緊急実行

緊急ジョブの実行待ち時間短縮

- メモリ資源の3つの解放方式を自動選択

① 論理スワップ	メモリを解放しない
② パーシャルスワップ	メモリを部分的に解放
③ 物理スワップ	メモリをすべて解放

✓ *JAXA様ではデバッグジョブを緊急ジョブとして利用*
✓ *論理スワップのみ利用することでデバッグジョブのレスポンスも良好で満足して頂いています*

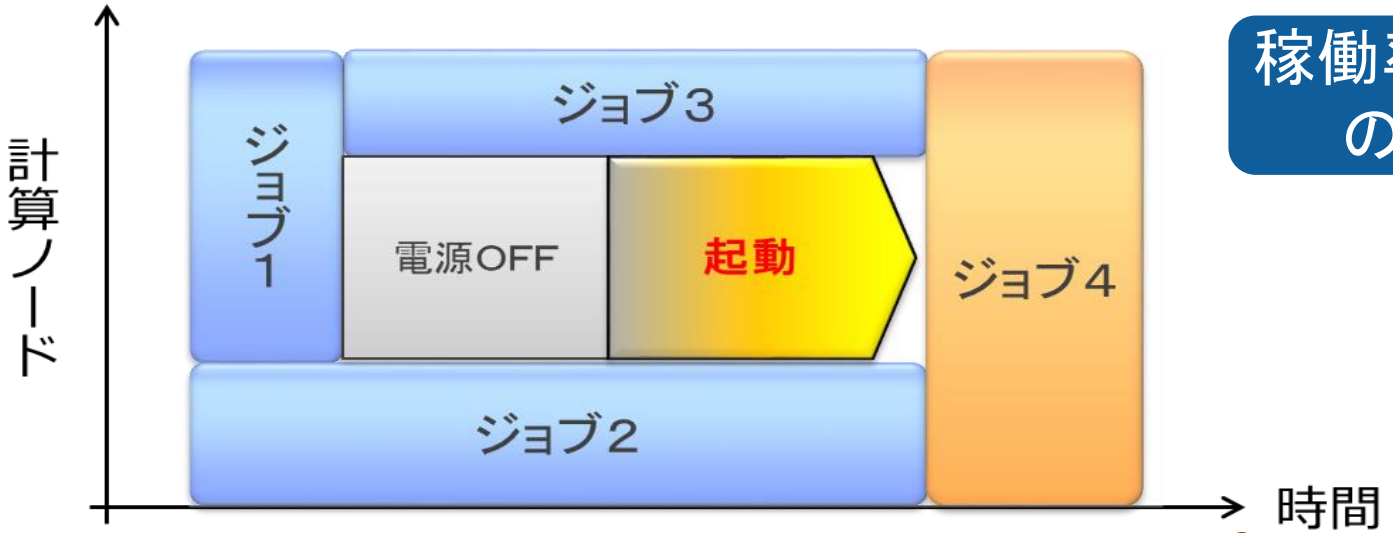
高スループット （3／3）

FUJITSU

JSCAPS: スループットへの影響の少ない節電運用

JSCAPS: *Job SCheduling Aware Power Save*　　情報処理学会研究報告, Vol.2017-HPC-160 No.2 (2017)

- ジョブのスケジューリング情報から計算ノードの予約状況を確認
- ジョブの予約が一定時間以上ない計算ノードを電源停止
- ジョブ実行開始予定時刻の前に起動が完了するように電源投入



稼働率と節電の両立

JAXA様と実装方式の検討から行い、現在、共同で効果検証中

柔軟な運用 & 使いやすさ（1／3）

FUJITSU

コンテナ型仮想化技術による実行環境の配備

- システム管理者、ユーザが作成した任意のDockerイメージをジョブ専用のコンテナ上にディプロイ
- 利用例
  - システム管理者
    - ① ユーザから要望のあるミドルウェアを動的に配備したい
    - ② ISV・OSSの版数に依存関係があるパッケージ・ライブラリ群を同一ノードで配備したい

特殊環境の容易な切替

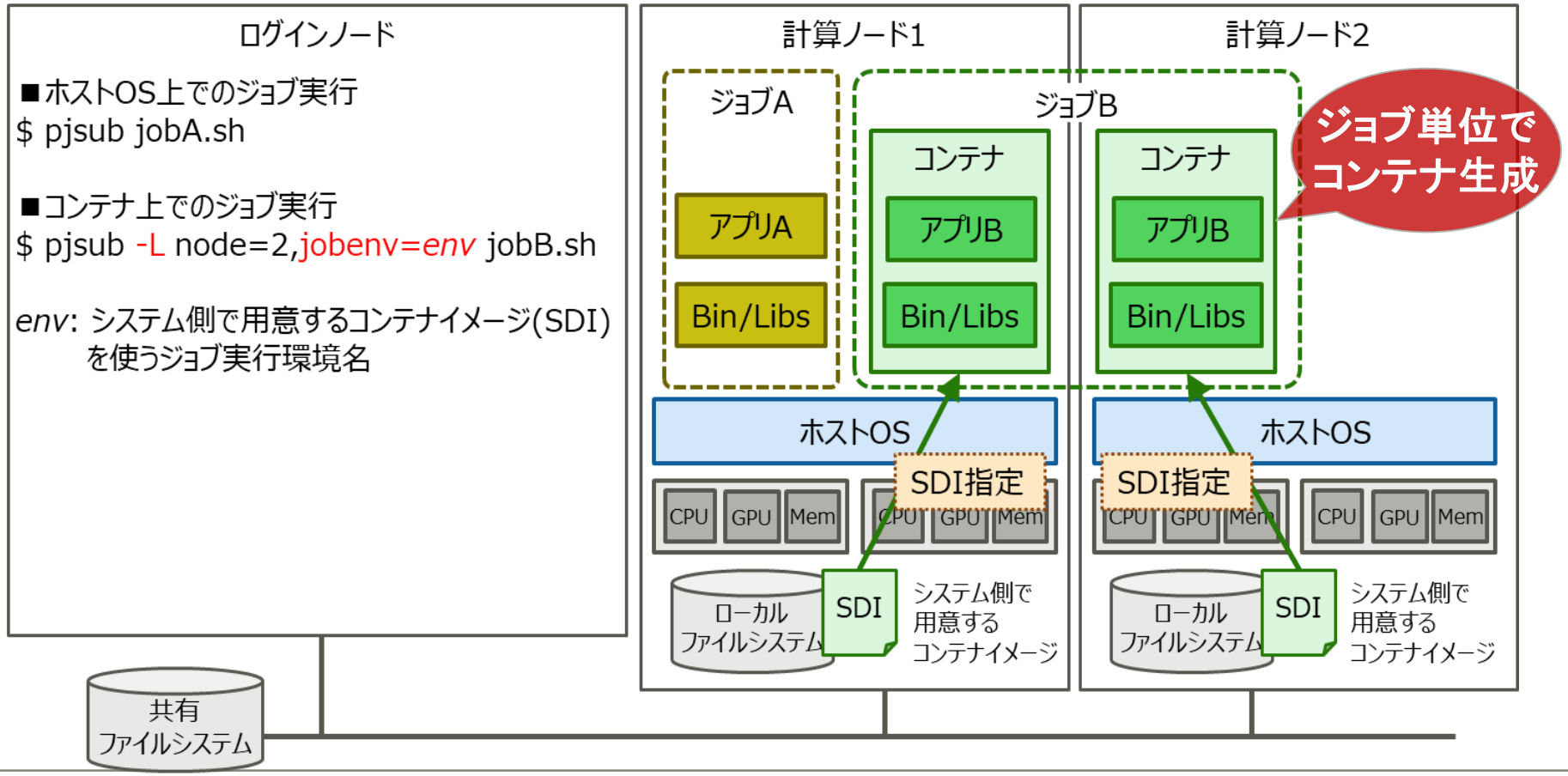
  - ユーザ
    - ① ジョブの実行結果を再現するパッケージ群を配備したい
    - ② WSで使い慣れた開発環境でスパコンを利用したい
    - ③ 世の中にあるDokcerイメージをそのまま利用したい (ディープラーニングフレームワーク etc.)

実行環境の可搬性

理研和光様の専用クラスタに試供し、評価を頂いています

柔軟な運用 & 使いやすさ（2／3）

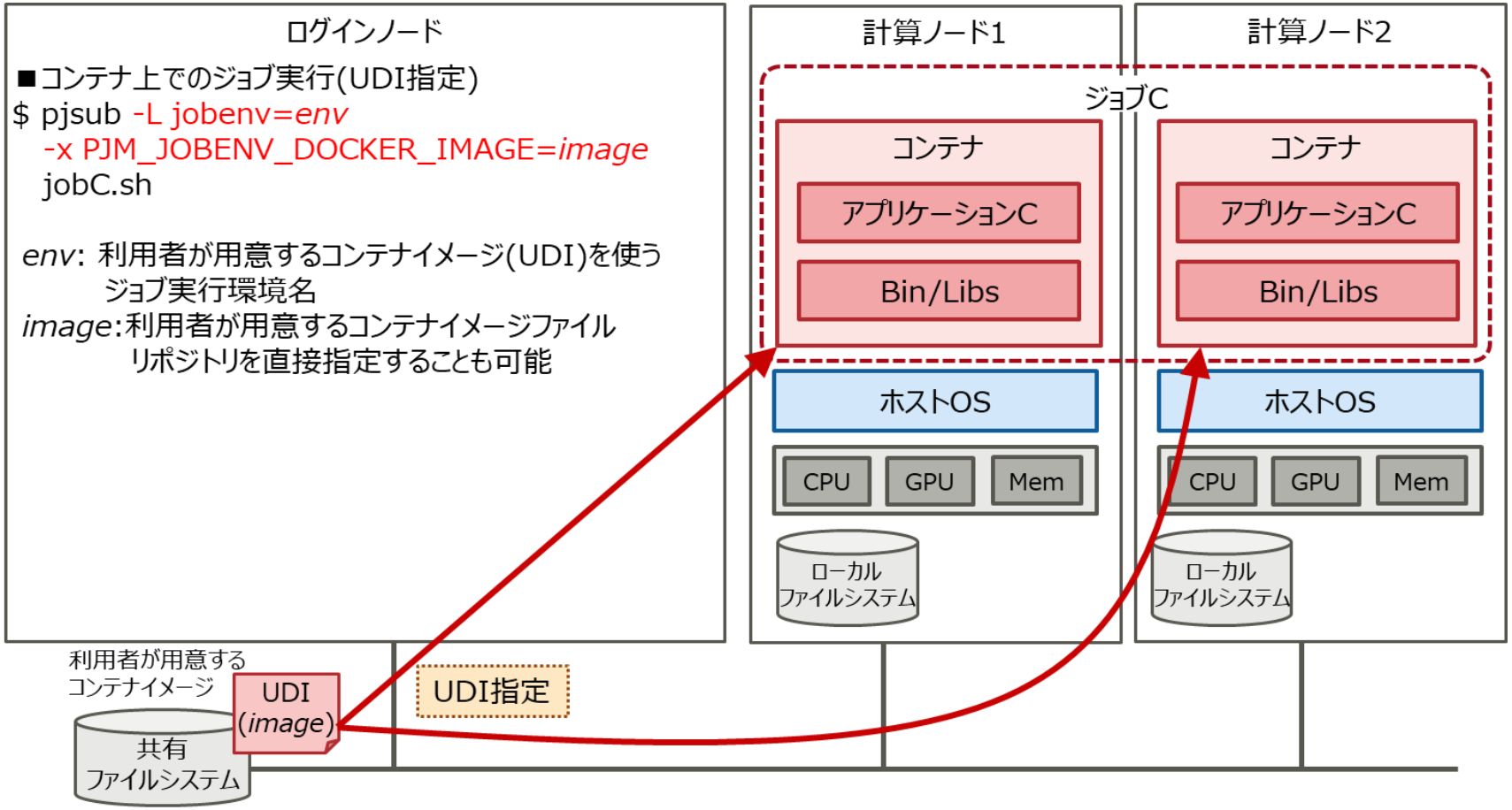
- システム管理者がDockerイメージ(SDI)を配備 して、ジョブ実行環境の名前を定義
  SDI: *System Deployed Image*
- ユーザはジョブ投入時にジョブ実行環境名を指定するだけ
- ジョブスクリプトの修正は不要。マルチノードのMPIにも対応



柔軟な運用 & 使いやすさ（3／3）

- ユーザが用意したDockerイメージ(UDI)をジョブ投入時に指定
- ジョブスクリプトの修正は不要。マルチノードのMPIにも対応
- ユーザにDockerコマンド権限を与える必要はなくセキュリティ確保

UDI: *User Deployed Image*

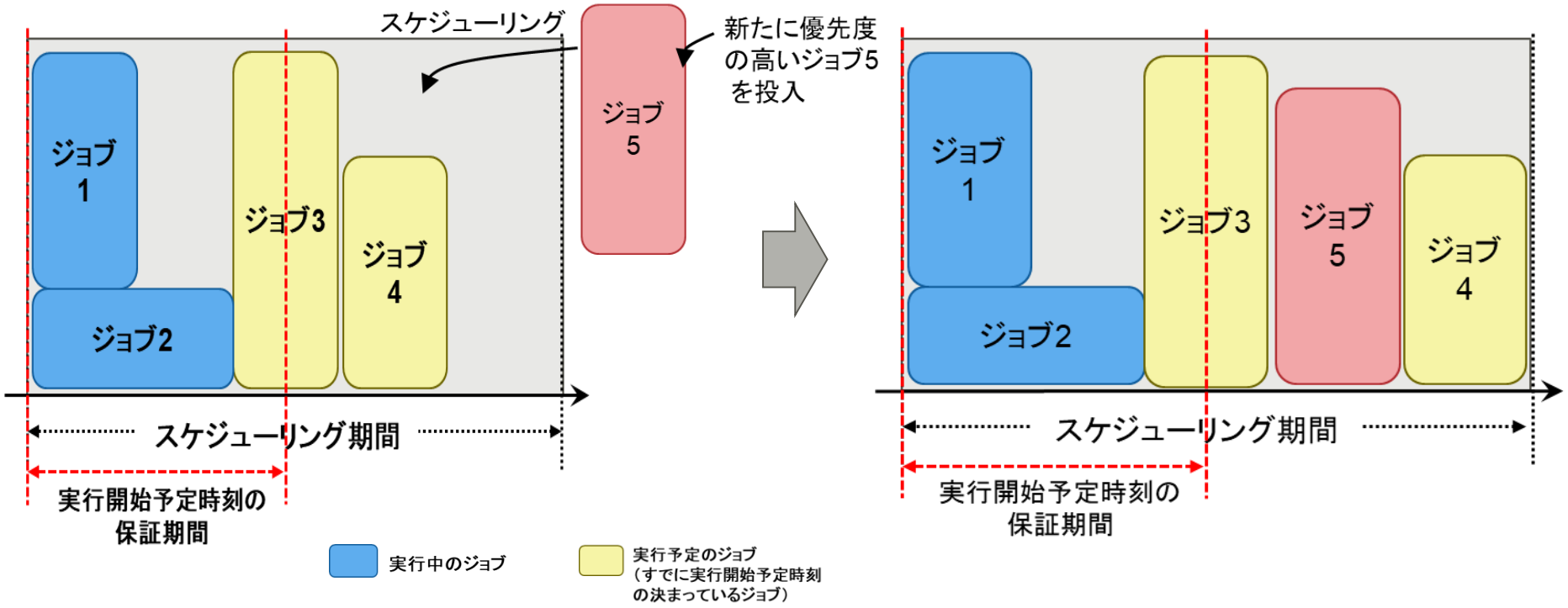


使いやすさ

ジョブ実行開始予定時刻の後退を抑止

- ジョブスケジューラーが一度決定した実行開始予定時刻を保証
- 実行中ジョブが早く終わり、計算ノードが空いたら、実行開始予定時刻を前倒し

ジョブがいつ完了するか予測可能



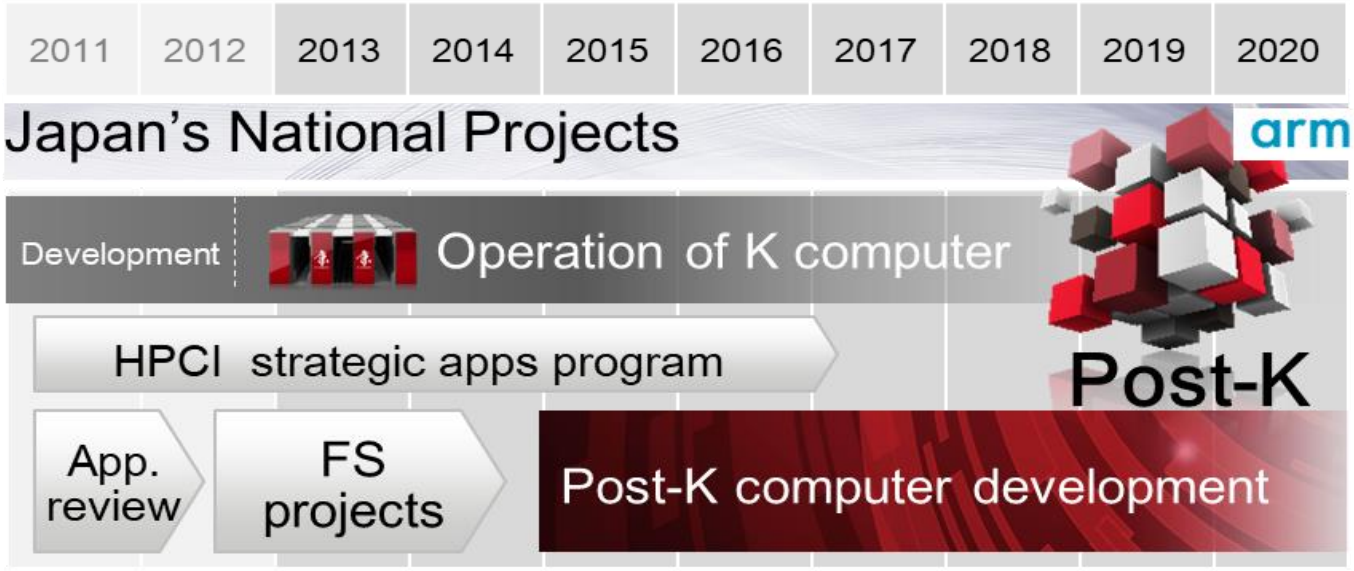
FUJITSU

最新状況

エクサスケール時代に向けて

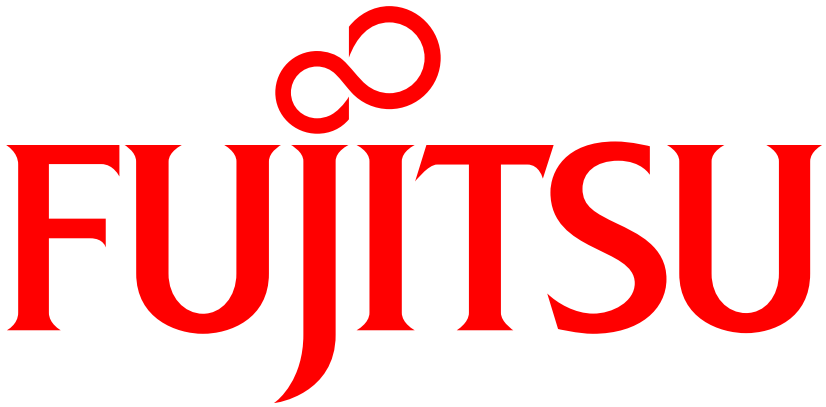
FUJITSU

■ スーパーコンピュータ「京」の後継機：ポスト「京」を開発中



- 2018.5.17-18 富士通フォーラム2018
  ポスト「京」試作機を展示
- 2018.8.22 Hot Chips 30
  ポスト「京」に搭載するCPU
  「A64FX™」の仕様を公表
  http://pr.fujitsu.com/jp/news/2018/08/22-1.html

ポスト「京」で開発した成果をPCクラスタへ展開していきます



shaping tomorrow with you