HOKUSAIシステムの運用報告とShoubuの利用募集

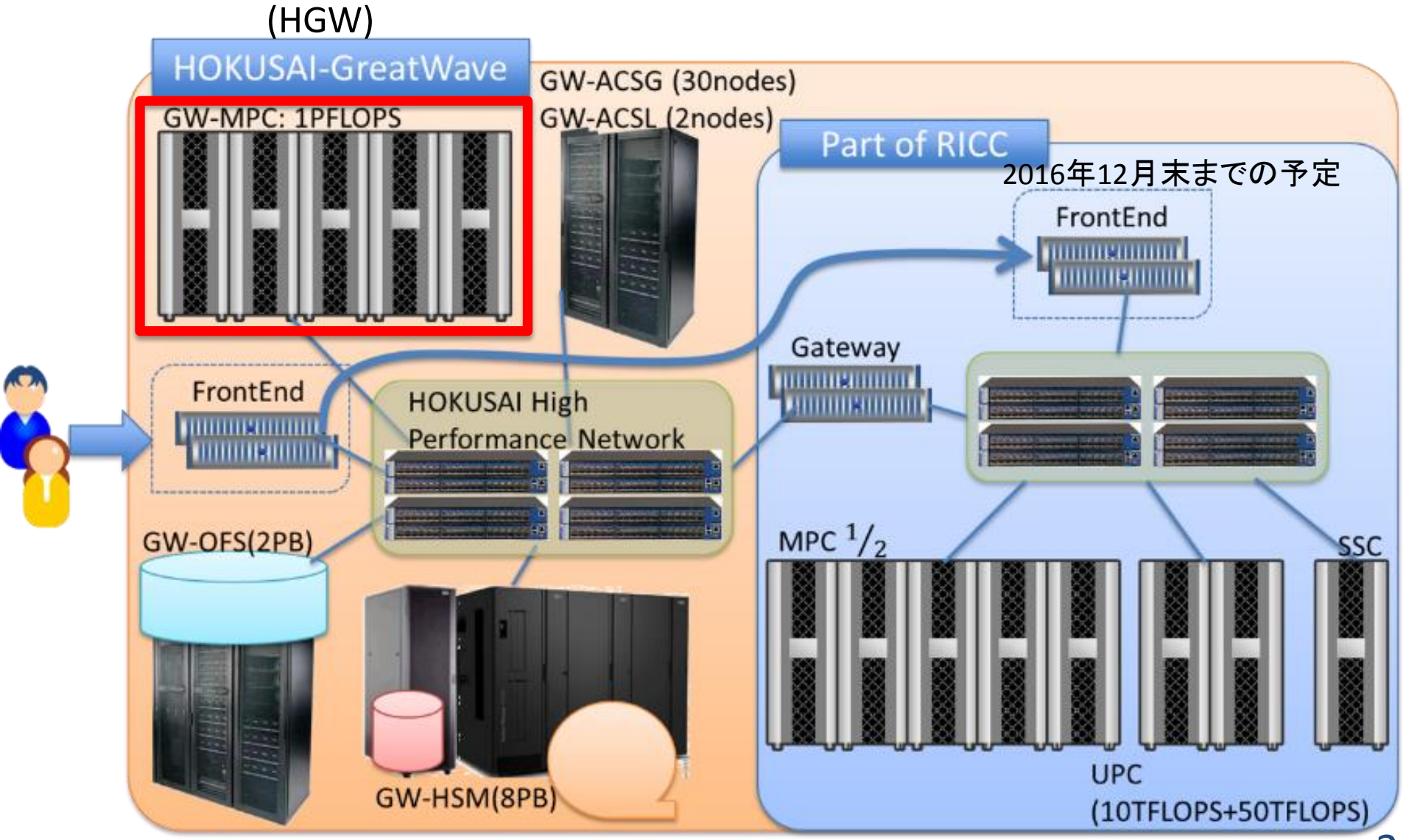
理化学研究所 情報基盤センター
2016/6/8 和光

理研シンポジウム2016
スーパーコンピュータHOKUSAIとShoubu、研究の最前線

Outline

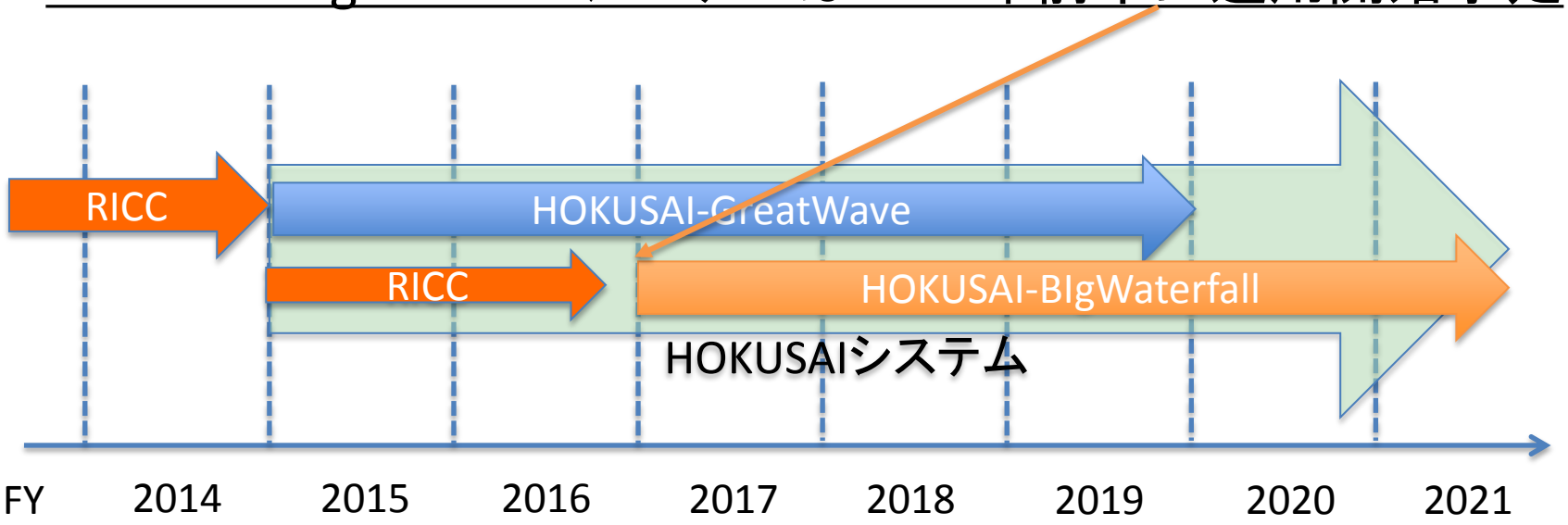
- HOKUSAIシステム2015年度の運用報告
  - HOKUSAIシステムの概要
  - 課題数、アカウント数
  - HOKUSAI利用状況
  - 運用方針の変更
  - ハードウェア障害件数
- Shoubu(菖蒲)の利用募集

HOKUSAIシステムの概要



HOKUSAIシステムの運用スケジュール

- HOKUSAIシステムからシステムを2段階で導入
  - システムを使えない期間の短縮化
  - 最新の計算資源を提供
- HOKUSAI-GreatWaveシステムを2015年4月から運用開始
  - RICCシステムの半分程度も運用を継続(2016年12月末までの予定)
  - HOKUSAI-BigWaterfallシステムは2017年前半に運用開始予定



HOKUSAI システム構成図

超並列演算システム(1FPLOPS)

Fujitsu PRIMEHPC FX100
- ノード数：1080
  - コア数：34,560コア（32コア/ノード）
- メモリ量：34.6TB (32GB/ノード)
- インターコネクト：Tofu2
  - 通信速度：50GB/s × 2/ノード
  - 隣接通信：12.5GB/s × 2
- 外部IO速度：204GB/s

2016年12月末までの予定

RICCシステム
# of nodes: 586 (4,688 cores)
- # of CPUs: 2/node (8cores/node)
  - CPU: intel Xeon X5570 2.93GHz
- Total mem: over 8TB
- Network: Infiniband DR(2GB/s/node)

フロントエンド

高速広帯域ネットワーク

Mellanox SX6036 × 12（InfiniBand FDR） FBB構成

理研
ネットワーク

オンライン・ストレージ(2.1PB)

MDS: PG RX300S8+Eternus DX200S3
OSS: PG RX300S8+NetAppE5600 × 14
ファイルシステム: FEFS
理論IO帯域：190GB/s

階層型ストレージ(7.9PB)

IBM TS4500 + TS1140 × 6
階層構成：GPFS + TSM

管理サーバ群

管理用Ethernet

アプリケーション演算システム（GPU搭載）(13+157TFLOPS)

SGI C2110G-RP5
- ノード数：30（720コア）
- CPU数：2/ノード（24コア/ノード）
  - CPU:Intel Xeon E5-2670 2.3GHz
- メモリ量：1.9TB（64GB/ノード）
- GPU：NVIDIA Tesla K20X(4枚/ノード）
- ネットワーク：InfiniBand FDR (6.8GB/s/ノード）

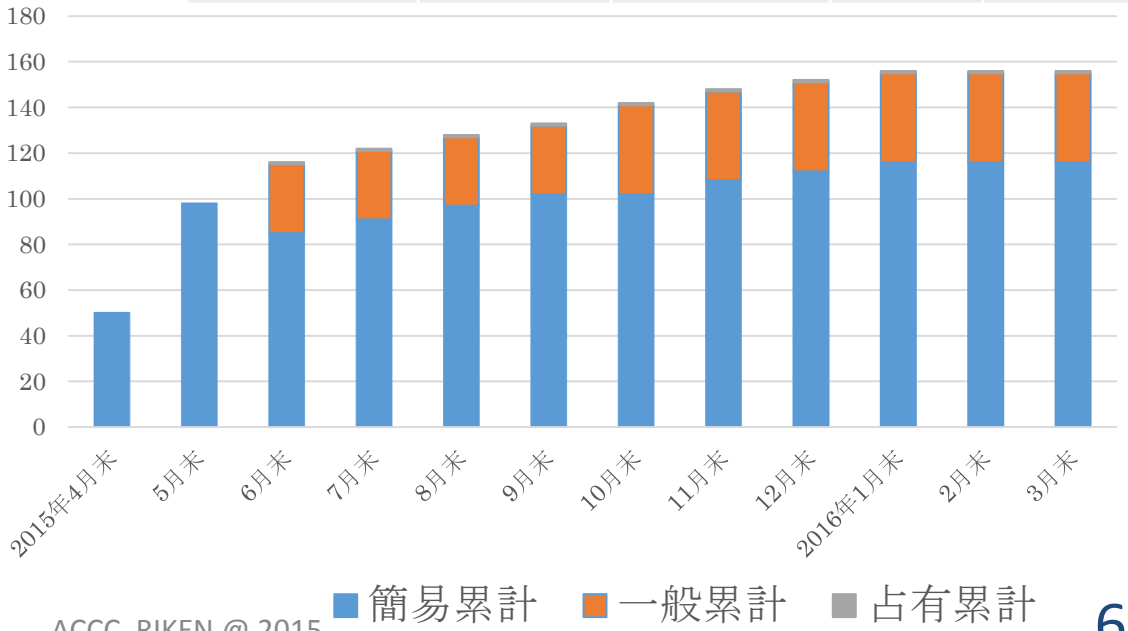
アプリケーション演算システム（大容量メモリ搭載）(2.4TFLOPS)

Fujitsu PRIMERGY RX4770 M1
- ノード数：2（120コア）
- CPU数：4/ノード（60コア/ノード）
  - CPU:Intel Xeon E7-4880v2 2.5GHz
- メモリ量：2TB（1TB/ノード）
- ネットワーク：InfiniBand FDR × 2
  (13.6 GB/s/ノード）

2015年度の課題採択数

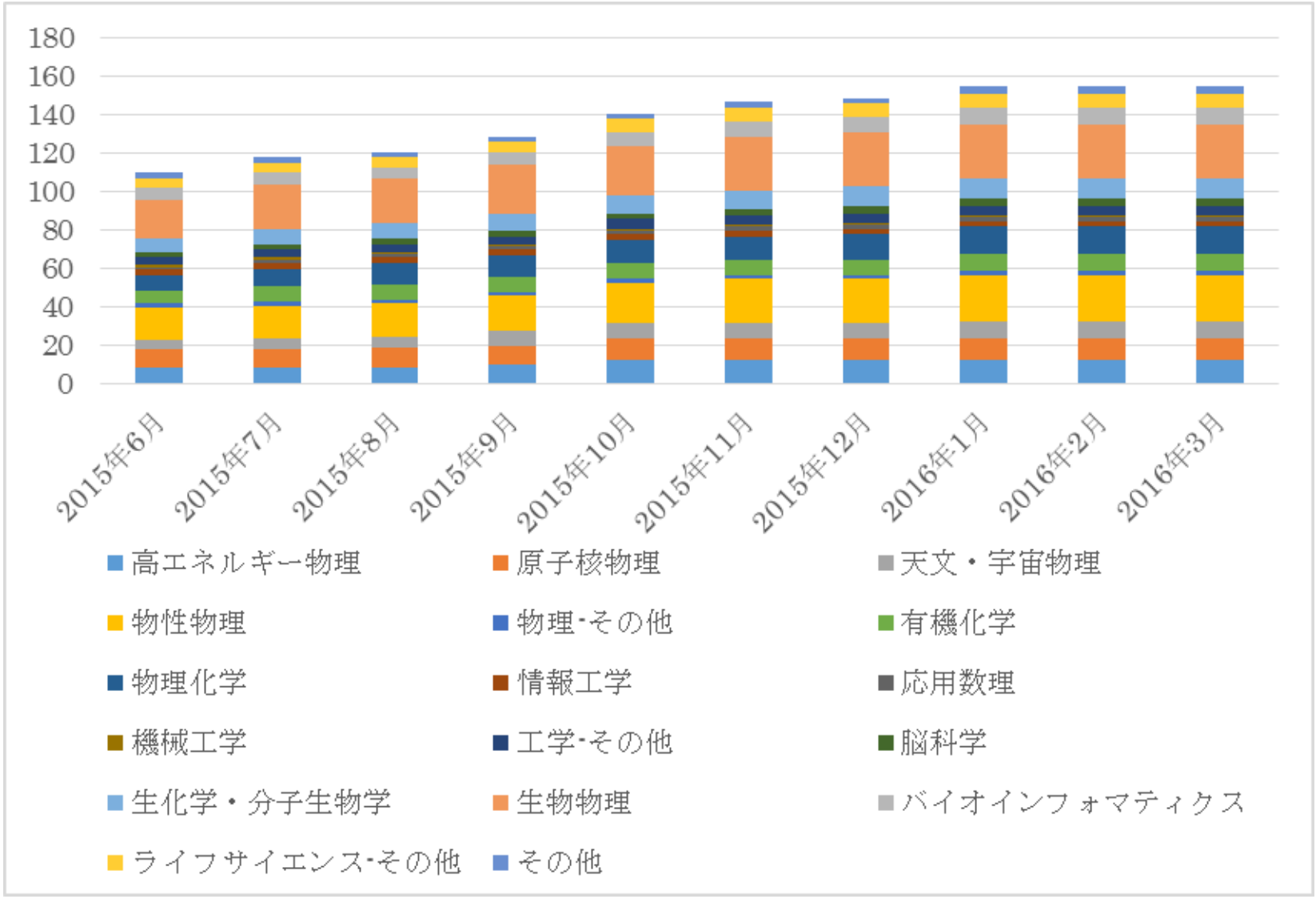
- 4,5月はトライアル運用
  - 98グループが参加
- 6月から本運用開始
  - 一般利用39課題
  - 占有利用1課題
  - 簡易利用116課題

課題審査委員会	利用開始月		応募	採択
第1回	6月	占有利用課題	1	1
		一般利用課題	30	30
第2回	10月	占有利用課題	募集なし	0
		一般利用課題	9	9

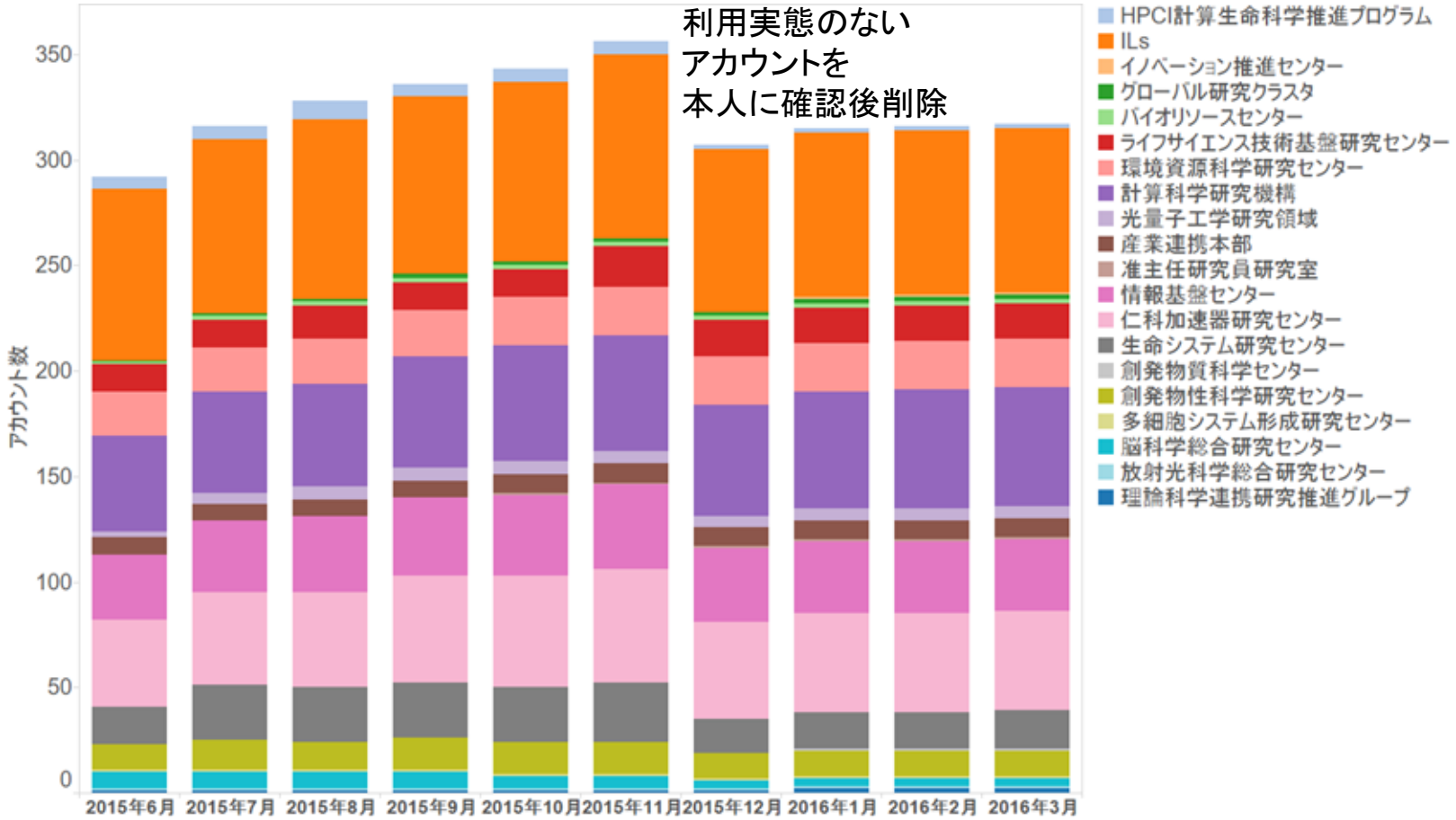


2014年度以前と比べて
一般利用は同程度
簡易利用は少し減少

2015年度の分野別課題数



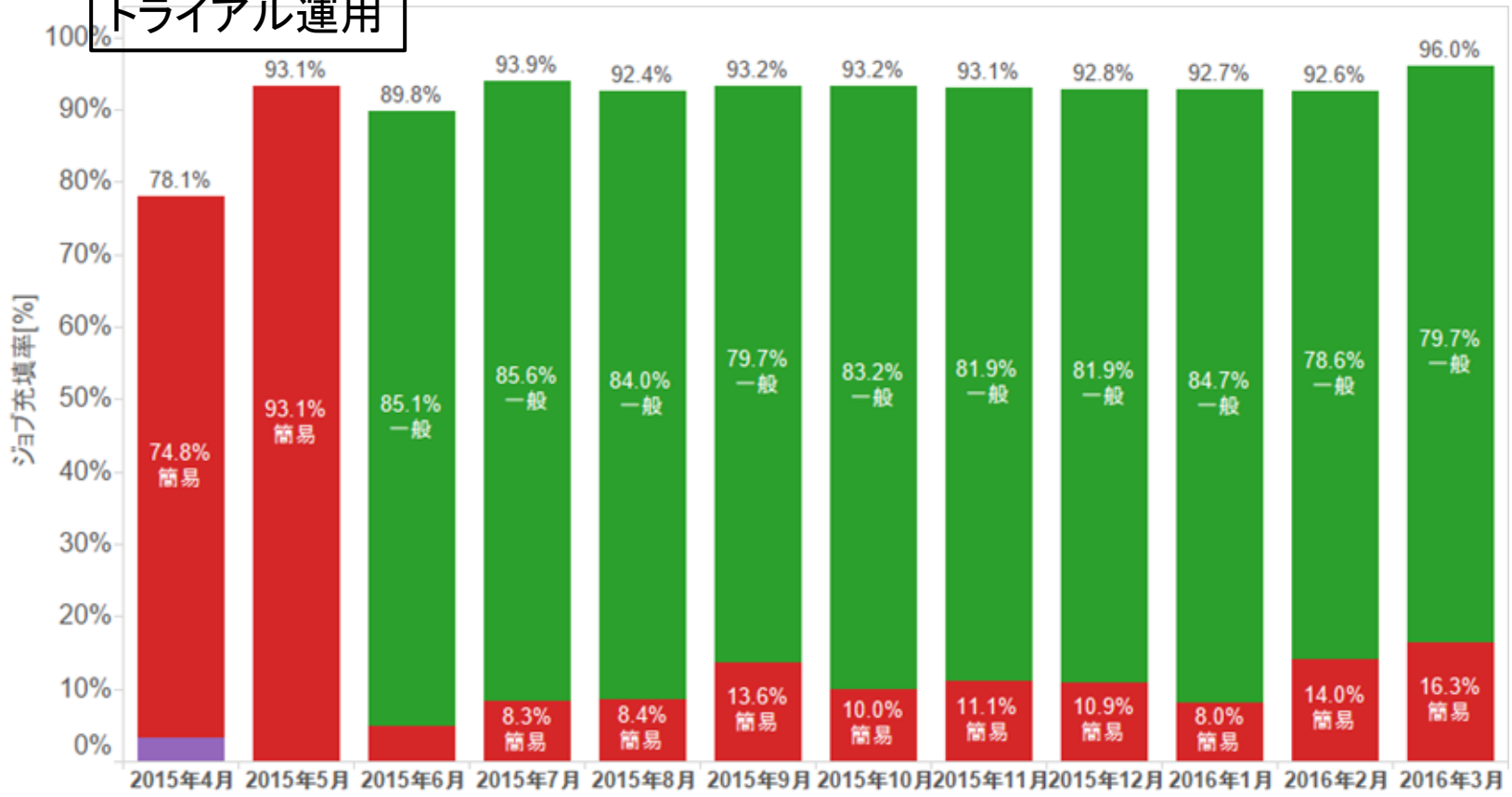
2015年度のセンター別アカウント数



アカウント数は300程度
主任研究員研究室、仁科加速器研究センター、AICSなどが多い

2015年度のHOKUSAI-GreatWave(HGW)全体の利用状況

4,5月は
トライアル運用



RICCの10倍の性能に増えているのにシステム稼働当初から高い稼働率
簡易課題は10-15%程度の利用率で推移

ジョブの稼働率が高いのは。。

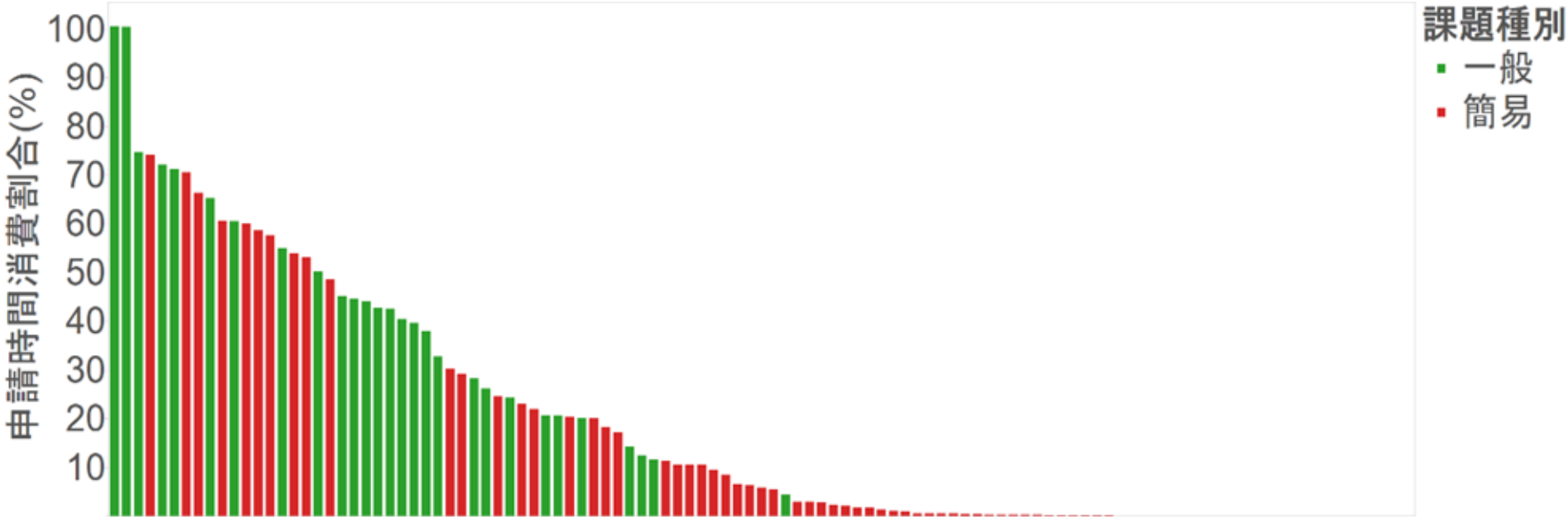
- 運用側にとっては良いことなのだが、、、、
  - 利用者のジョブが極めて流れにくいということ。
  - 最大の問題は、申請を受け付けたコア時間の合計が大き過ぎて、利用者が割り当てられたコア時間を消費できなかった点。

- 原因は、先代RICC時代と同じスキームで課題募集および審査を行ったこと。
  - RICCまでは稼働率は高かったが、これほどの申請コア時間を受け付けることはなかったという経験値を優先したこと。
  - HGWの性能は10倍上がるため、おそらくほどほどに裁ける申請コア時間であろうと高を括って、申請を受け付けたこと。
  - ある程度混み合うのはあるとして、RICCでは審査結果による課題間の実行優先度制御を行ったが、今回対応が遅くなったこと。
    - これらは極めて反省しないと行けない点でした。

想定以上の申請コア時間と対策

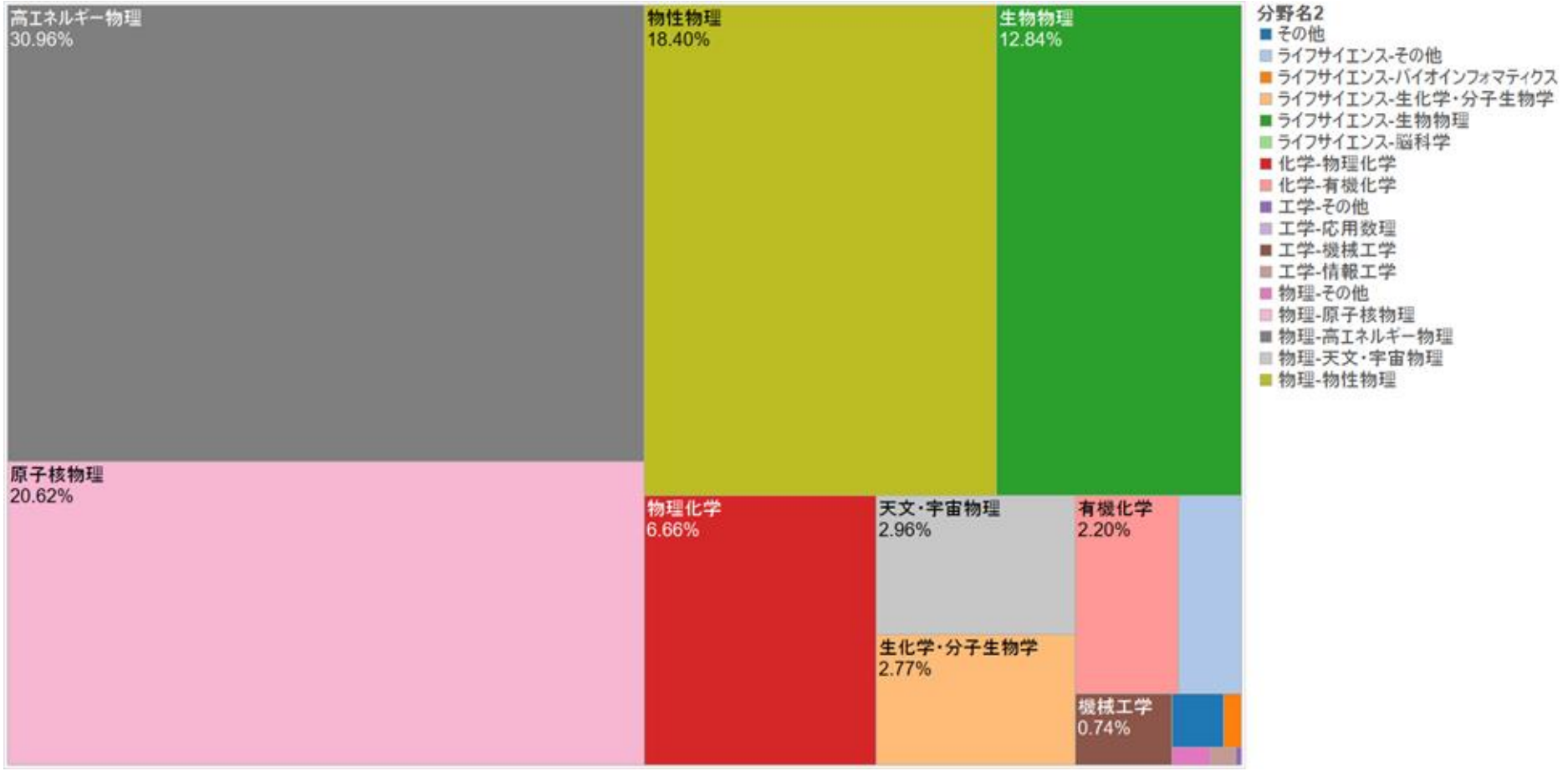
- 第1回の課題審査段階(6月)での割当コア時間の合計
  - GW-MPC:195%
  - GW-ACS(G/L):102%
- 第1+2回の課題審査段階(10月)での割当コア時間の合計
  - GW-MPC:258%
  - GW-ACS(G/L):136%
- 結果として、想定以上のコア時間が申請されGW-MPCは割当コア時間の40%程度の利用しかできないことに
  - ユーザの方々には非常に申し訳ありませんでした
  - 遅くなったが一時的な対策として、2016年1月22日に運用ポリシーを変更し、評価の低い課題の優先度の回復率を半減
  - 2016年度以降は課題審査方法を変更することに

2015年度の課題毎の割当コア消費割合



一部を除いて一般課題は40%程度の消費割合に

分野別でのHGWの総コア時間消費



高エネルギー物理と原子核物理で半分程度と偏りがあり、
大規模利用な課題が多かった

利用申請の方向性の議論

- スパコンの課題審査はスーパーコンピュータ課題審査委員会の事案
  - 事務局などは情報基盤センター
  - 問題点の整理と実施案の検討を行った。
- 基本的な方向性
  - 申請時間の制限、評価を厳格に行い不採択とすることも。
    - 課題審査を始めた際から今までは不採択となった課題はなかった。
- 検討案色々
  - 課題毎に最大利用資源を提供可能資源の10-20%の上限を付ける
  - 提供可能資源の上限を設け、評価の高い順に採択し、枠に入らない評価の低い課題は不採択とする
  - 分野毎に利用可能資源割合の上限を付ける
  - シンポジウムなどで話を聞いて審査する

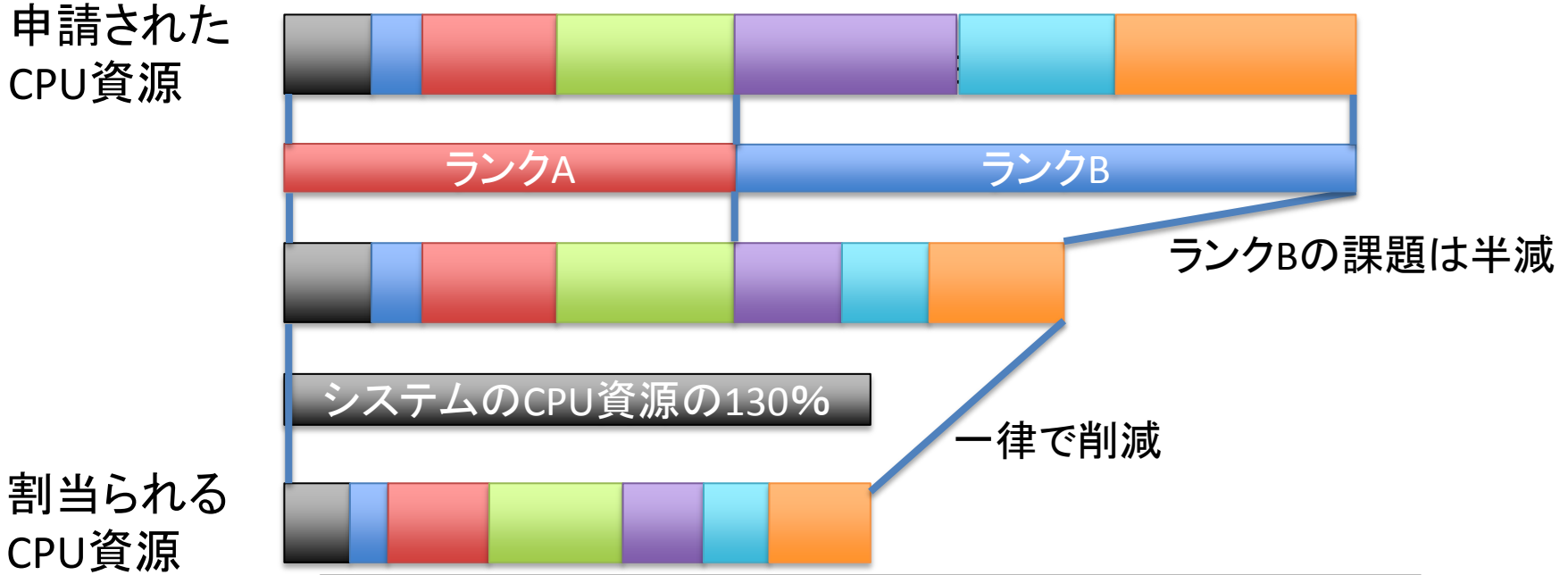
スパコン利用の新たな方針

- 一般利用課題全体で割当てる計算資源を提供可能資源の130%以下にする
  - 審査委員の評価によって削減率を変える
- 1つの課題が申請できるCPU資源の上限を各システム20%以下に
  - 1人のユーザが(複数の課題に渡って)申請できるCPU資源の上限も20%以下とする
- 全システムの10%程度以上の申請を大規模利用として、より厳しい審査を行う
  - 外部の専門家に審査に加わっていただく
  - 大規模利用は不採択の可能性も高くなる

- まだまだ、これで問題無いとはせず、常に問題点の改善を考えながらシステム運用を行いたい。

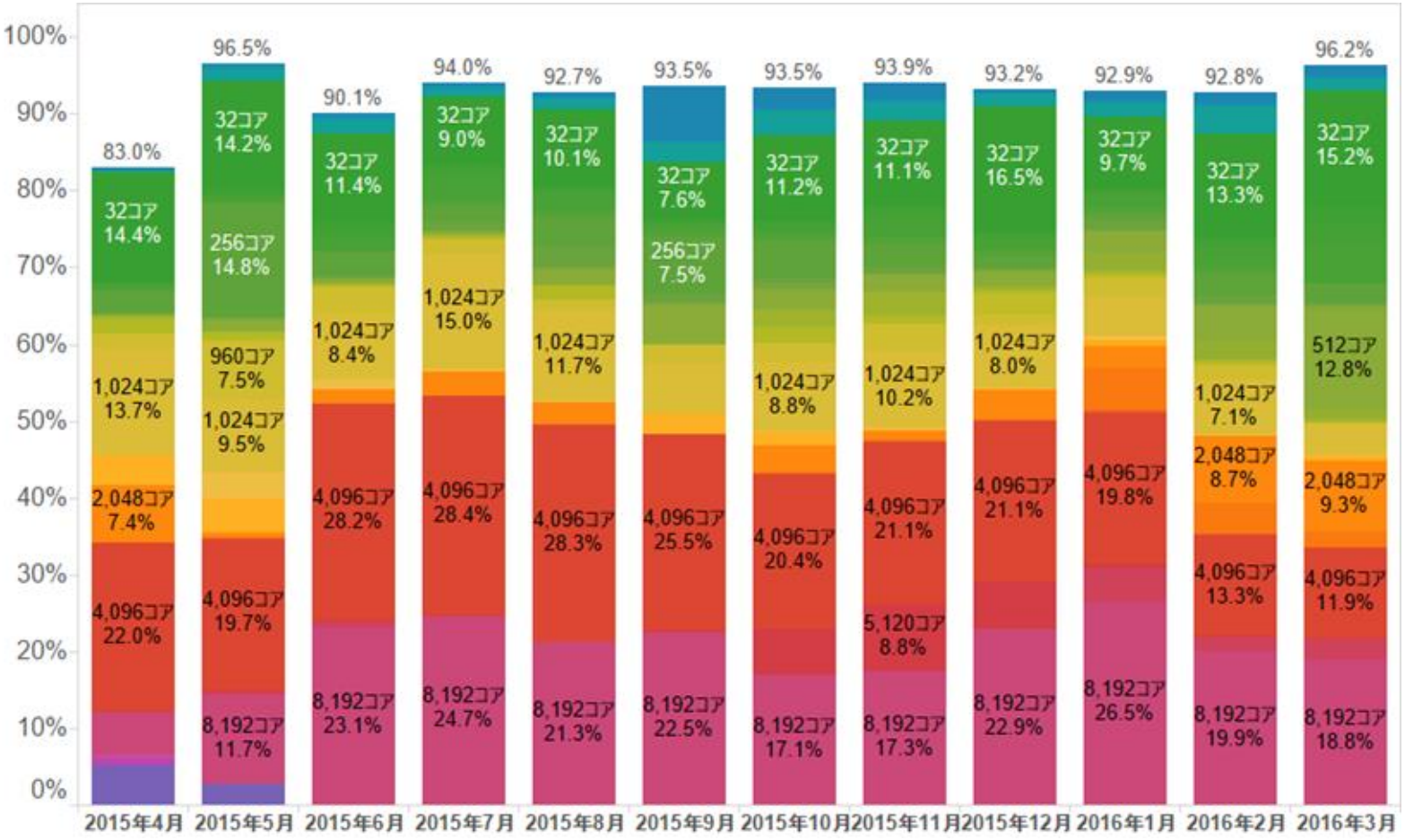
CPU資源の割当方法

- CPU資源の合計が130%以下になるまで以下を実行する
  1. 審査委員の評価によって不採択となれば、簡易利用に変更
  2. 評価点によりランクAとランクBに分け、ランクBの課題はコア時間を半減
  3. 一律でコア時間を削減する



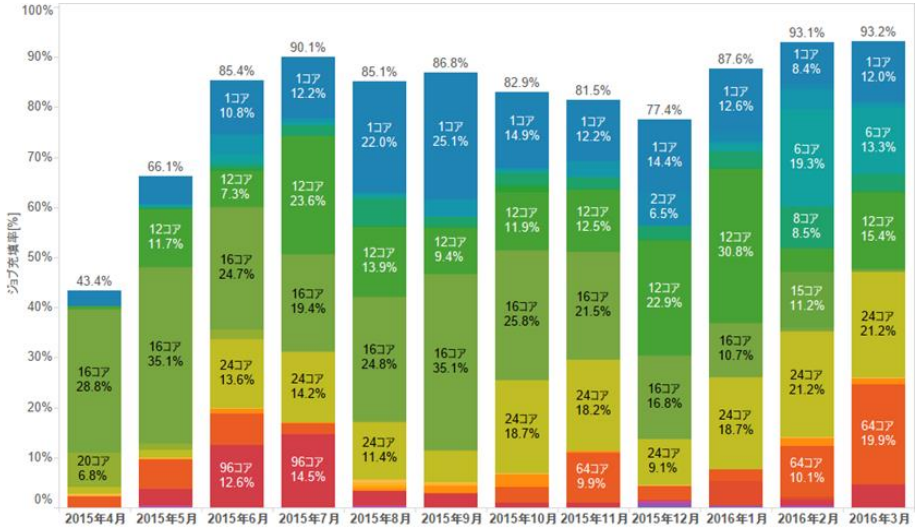
2016年度については2課題が不採択とされ、
2課題が半減され、130%以下になった

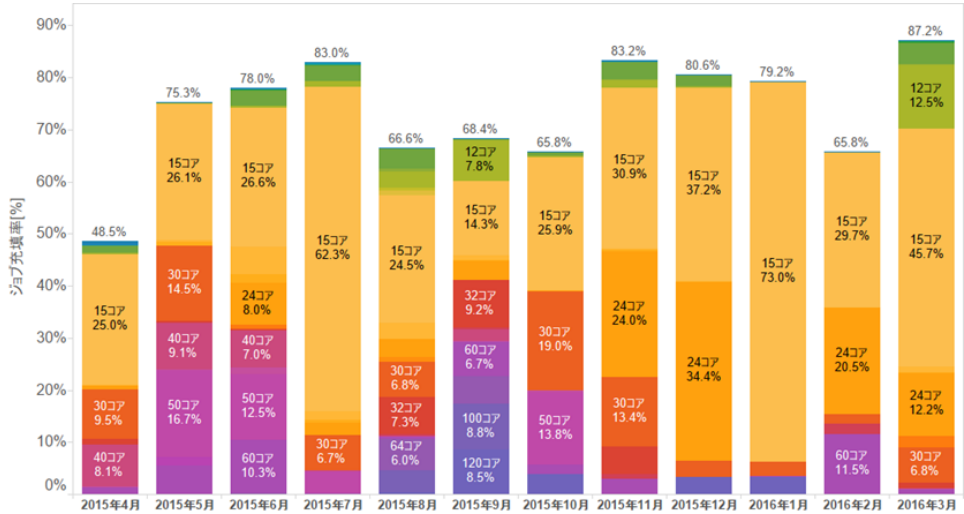
GW-MPC(超並列演算システム)稼働率



システム稼働当初から高いジョブ充填率
1000コア以上の高並列のジョブが半分以上を占めていた

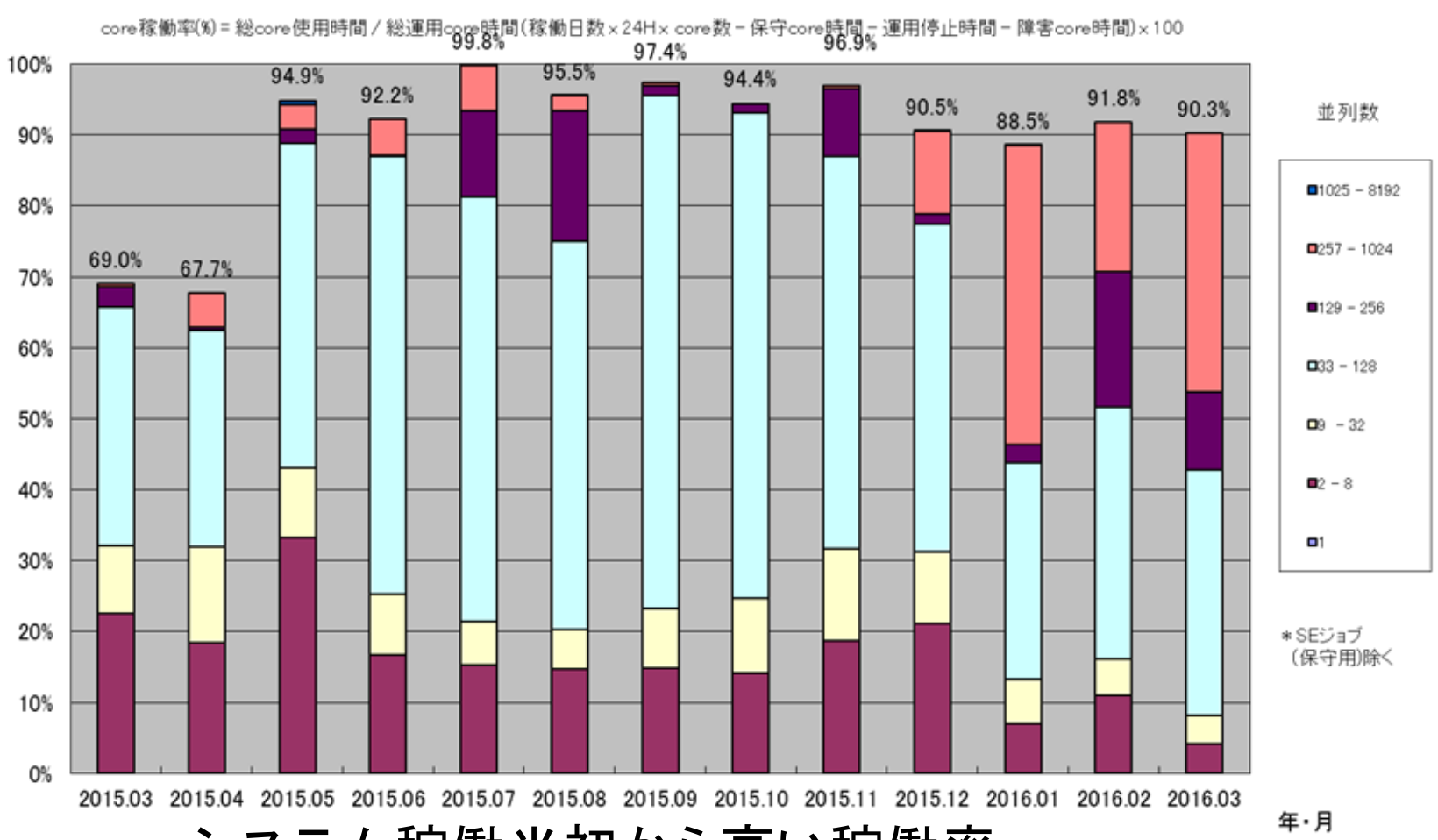
（左)GW-ACSG(GPU)稼働率と (右)GW-ACSL(Large memory)稼働率





GW-MPCに比べると余裕があった
ただし、月によっては90%程度のジョブ充填率を占めた

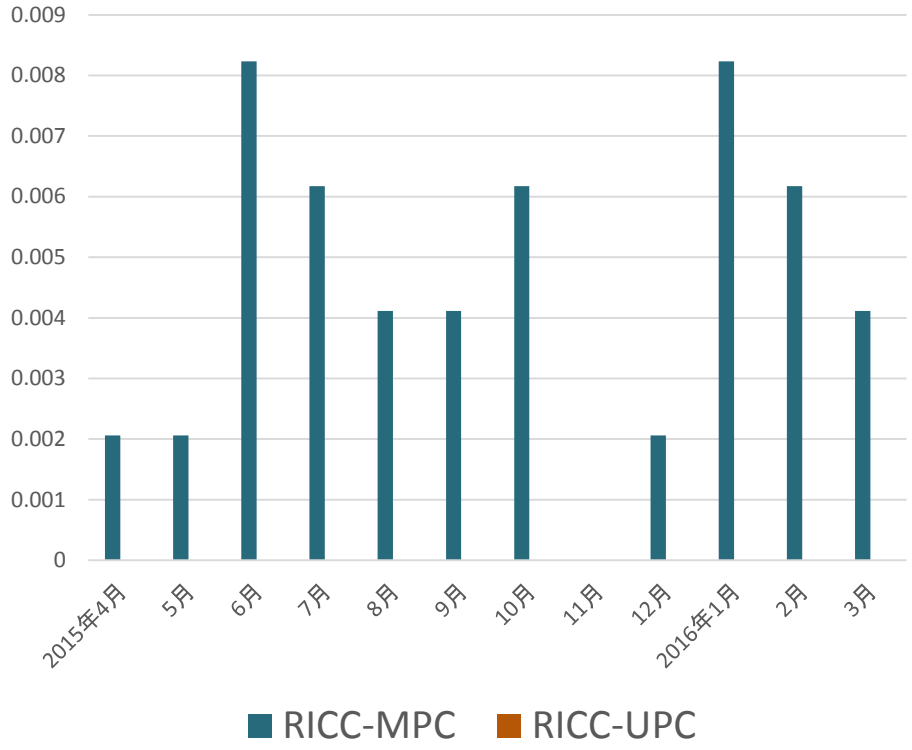
RICC-MPC(超並列演算システム)稼働率



システム稼働当初から高い稼働率

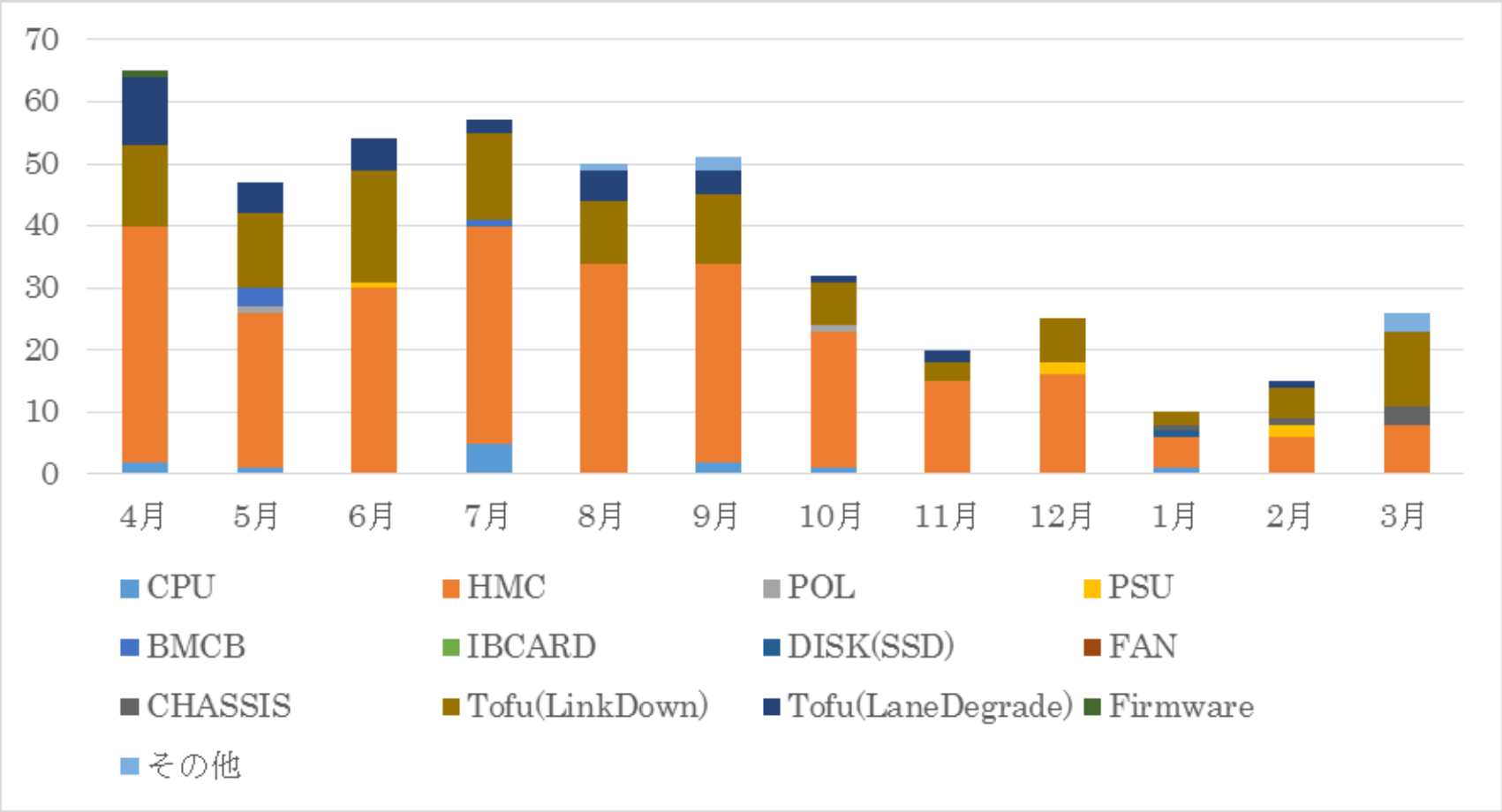
本運用開始後はほぼ90%を超えるジョブ充填率

(左)HGWハードウェア障害率(障害件数/ノード数)
(右)RICCハードウェア障害率(障害件数/ノード数)



GW-MPCとGW-ACSG/Lは障害率が大きい
RICC-MPCとRICC-UPCは障害率が非常に小さい

GW-MPCハードウェア障害詳細推移



メモリ関連(HMC)とネットワーク関連(Tofu)の障害件数が多かった
障害を減らす対応(8月頃)の後も、まだ十分低いとは言えない

Shoubuの概要

- Shoubu(菖蒲)の名前の由来
  - 古来より縁起の良い植物
  - 水辺の植物で、葉が扁平で積層になっている
  - 「勝負」にも通じる
- 理研の役割
  - 実アプリケーションの性能評価とシステムのDC設置・運用の評価
- Shoubuシステムの経緯
  - 2015年6月に2 PetaFlops級のExaScaler-1.4の5台構成で、情報基盤センターに設置される
  - 2015年7月と11月のGreen500で1位に
  - 2016年5月に換装されて、ZettaScaler-1.6に

Name Acorus calamus Family Acoraceae Original book source: Prof. Dr. Otto Wilhelm Thomé Flora von Deutschland, Österreich und der Schweiz 1885) Gera, Germany

Shoubu(菖蒲)利用募集

- 利用目的
  - Shoubu(菖蒲)を使った、実アプリケーションの開発や性能測定
  - 大きなリソースを占有して使う計算を行う場合は要相談
- 利用資格
  - 情報基盤センターの運用系スパコン等と異なり、広く利用可能
    - 日本の居住者であること
  - 共同研究利用条件に対する同意(署名）
  - 利用内容に基づき、可否を情報基盤センターで判断
  - 利用後に、情報基盤センターに実行状況や利用結果について、公開可能なレポートを提出すること
- 利用募集
  - 7月から情報基盤センターのweb siteにて公募 http://accc.riken.jp/
  - 第1次利用として20件程度を想定
- *特記*事項
  - 年に2回ぐらい（5月、10月頃）は利用出来ない期間あり
  - まずは、GPUで動作していることが望ましい
  - 利用サポートは基本的にはありません。ベストエフォート対応。