新旧システム更新の舞台裏

国立研究開発法人理化学研究所
情報基盤センター
黒川　原佳
<motoyosi@riken.jp>

Outline

- はじめに
  - 理研のスパコンの位置づけ、システムの方向性と利用者の概要。
- 前スパコンRICCのまとめ
  - 利用状況や故障統計など
- 現システムにむけた検討
  - アンケートや利用状況
  - 設備更新とシステム更新の同時進行
- 施設設備工事
  - 電源設備と空調設備の更新作業
- 現システムへの更新
  - HOKUSAI GreatWaveシステムについて
- おわりに

はじめに

- プロダクション・システム
  - 理研のスーパーコンピュータ・システム（≠京）は、理研内の研究者の日常に直結したプロダクション実行を行うシステムです。
- 実験とシミュレーションの連携
  - 実験研究者が自身が利用したり、シミュレーション結果を元に実験を行ったり、実験結果を元にシミュレーションを行っているような研究課題が多い。
  - 実験データ処理にも対応できるように。
- 保守的と進歩的のバランス
  - センターとしては、利用者との意見交換のみではなく、5年後のシステムを見据えたシステム設計を行いたいと思っている。
- 貪欲な利用者とそれ以外の利用者
  - システムが空いていればいくらでも使うという貪欲な利用者と少し計算して時間が空いてまた計算する利用者との利用機会のバランスが取れる運用ができるようにする。

システム構成について

【システム構成】
超並列PCクラスタ＋GPUクラスタ＋専用機クラスタ＋大容量メモリ計算機


【超並列PCクラスタ】1039Nodes（8312core）
ノード性能：93.0GFLOPS, 12GB(mem), 500GB(hdd) ,
DDR IB×1

【多目的PCクラスタ】100Nodes(800core)
9.3テラフロップス
＋【GPGPUアクセラレータ】93.3TFLOPS

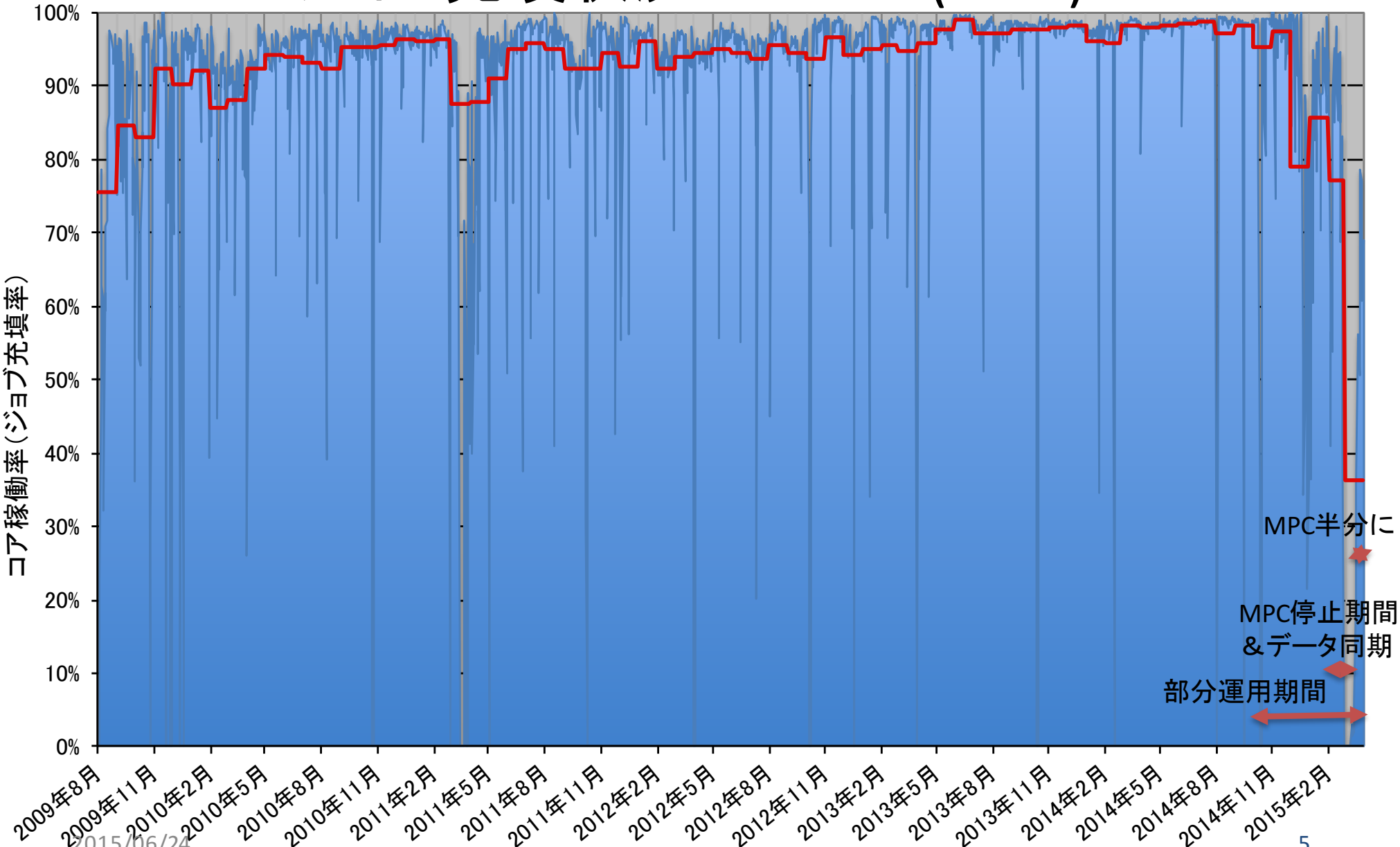
テープアーカイブ装置【HPSS】
(4PB)

磁気ディスク装置
(550TB)

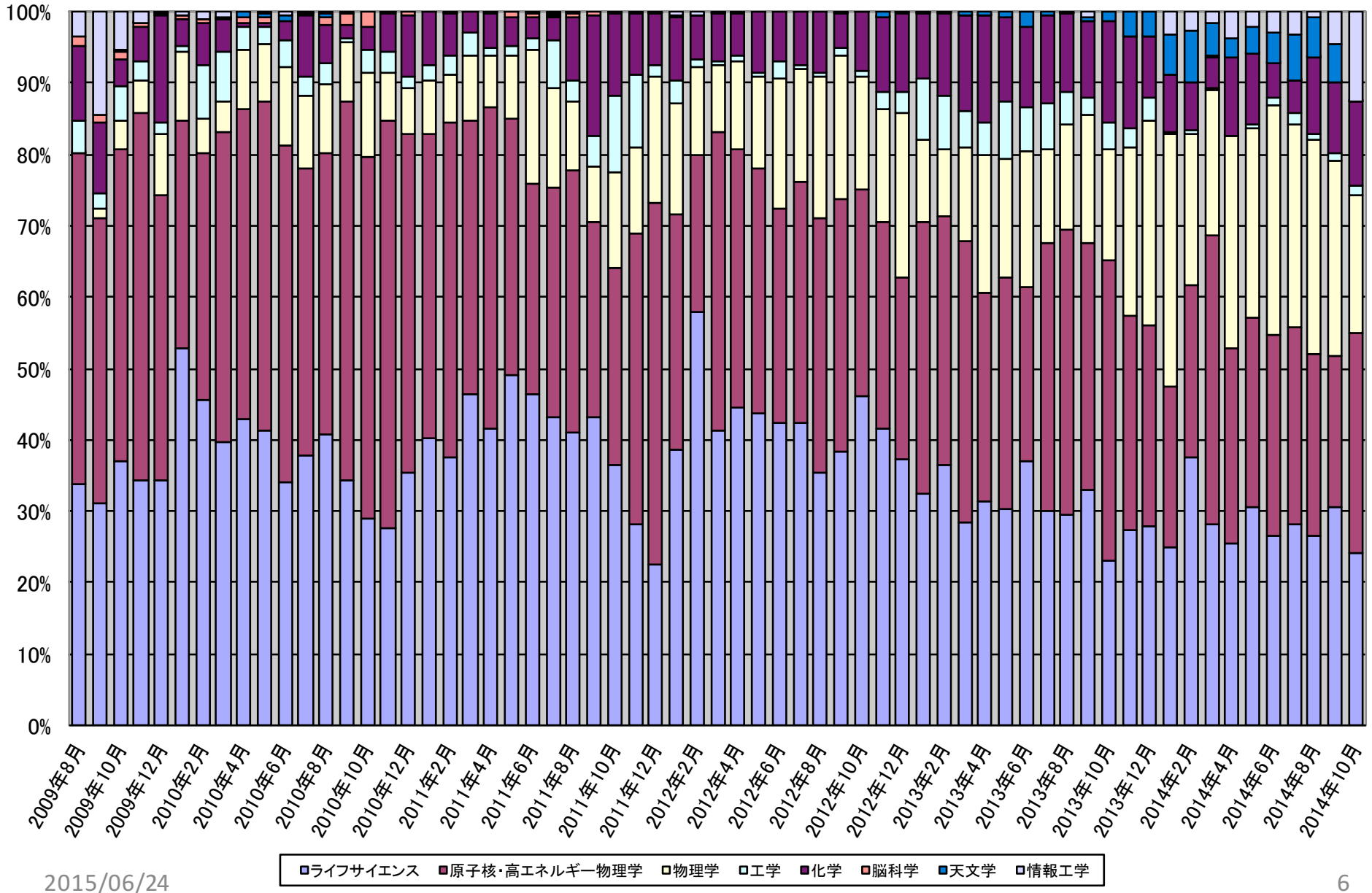
【MDGRAPE3クラスタ】3TFLOPS
＋【MDGRAPE-3】64テラフロップス
＊理研で開発した分子動力学専用計算機を接続

【大容量メモリ計算機】
0.24テラフロップス、512GBメモリ
＊1プロセスで500GB以上のメモリを利用可能

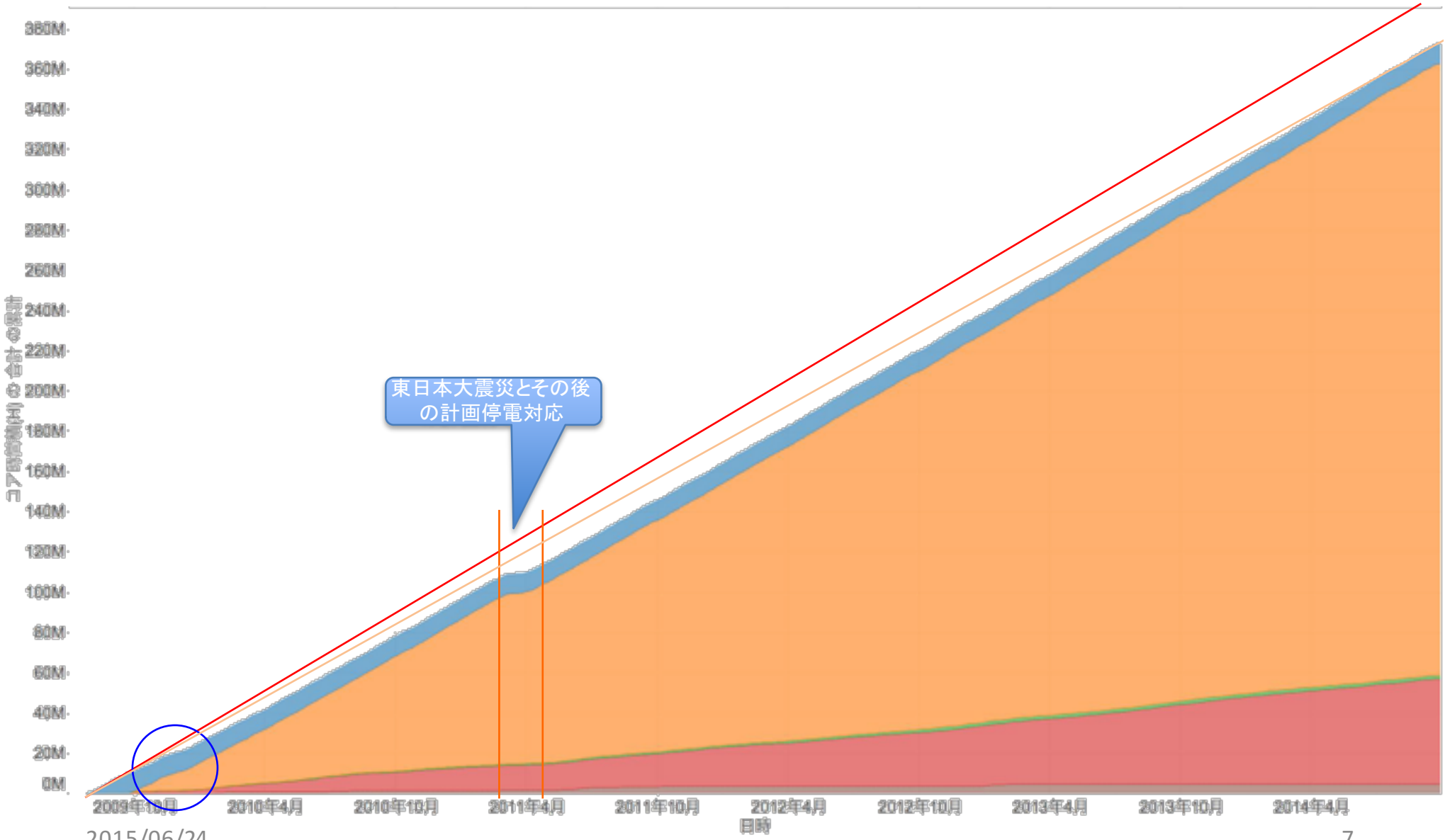
ジョブ充填状況について(MPC)



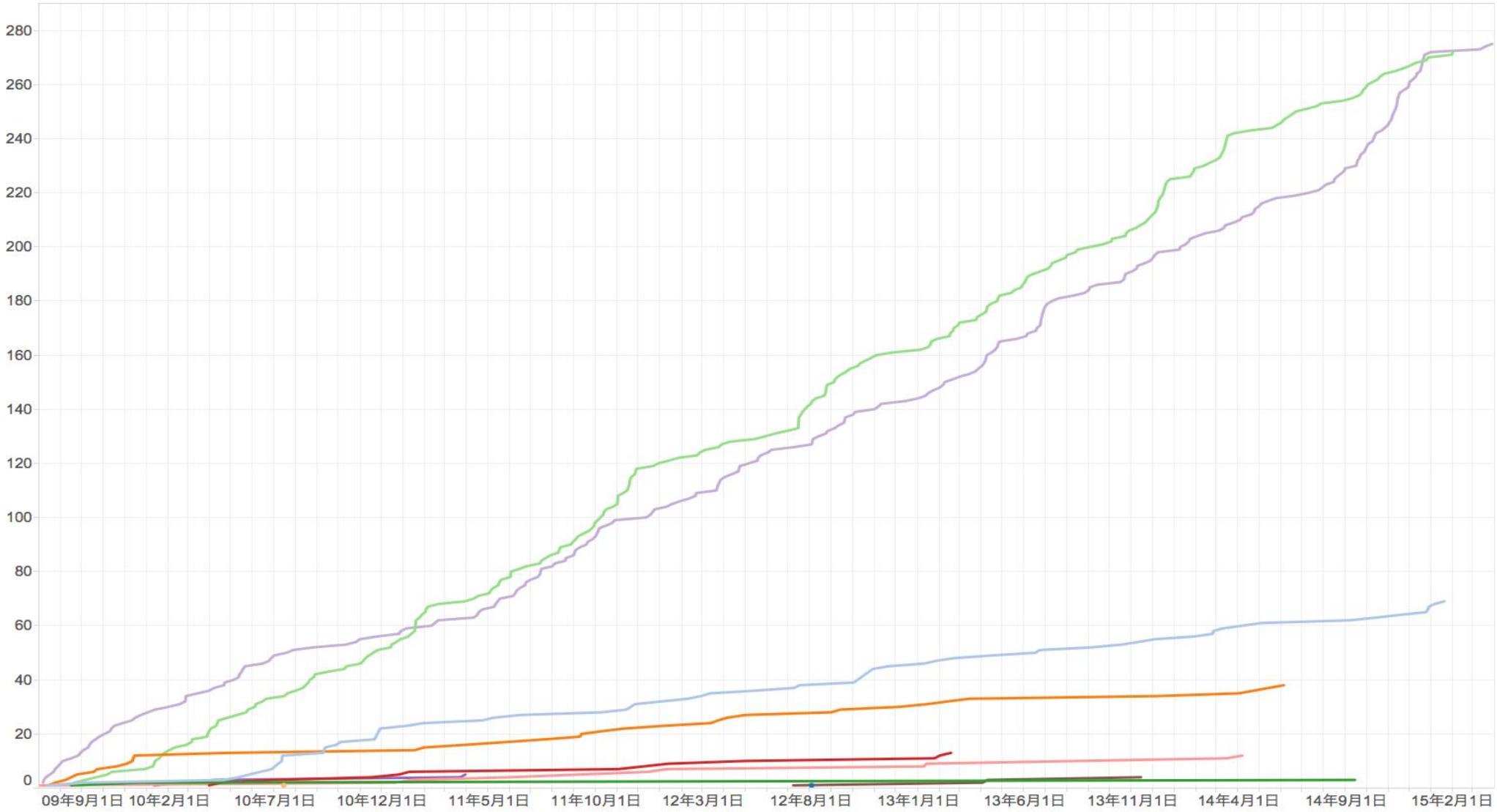
分野毎のコア時間利用比率



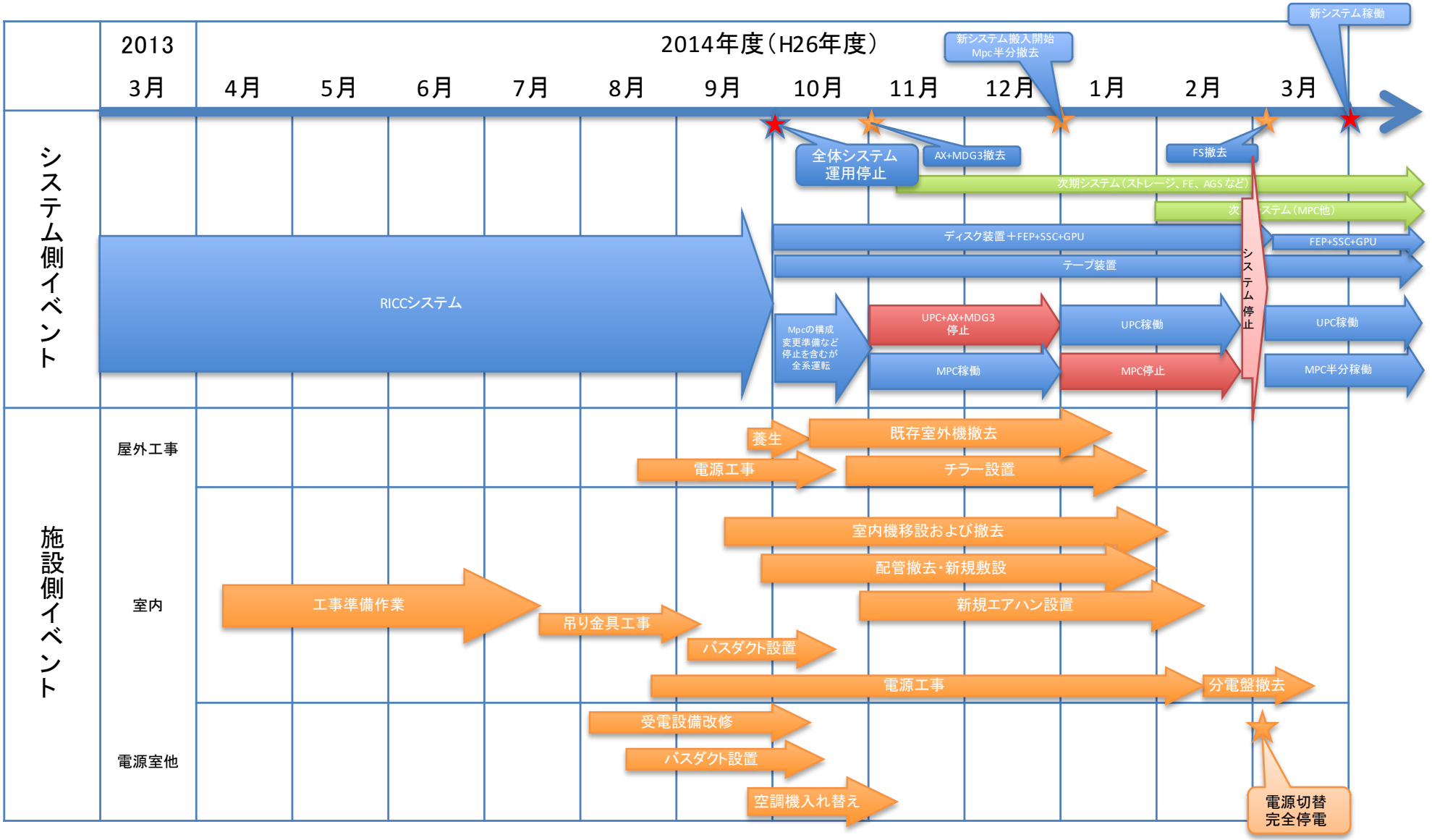
利用区分別コア時間累積



故障状況

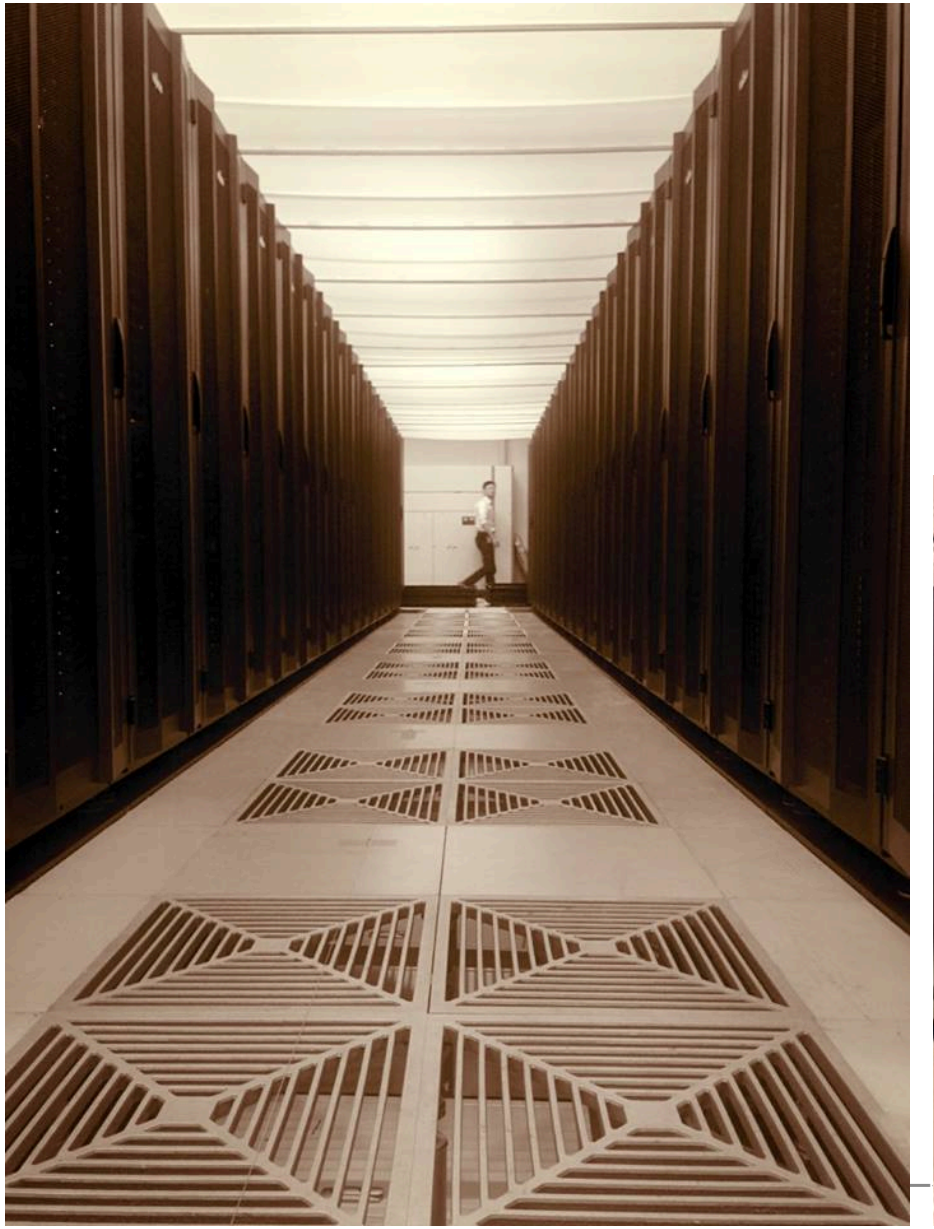

280
260
240
220
200
180
160
140
120
100
80
60
40
20
0
09年9月1日
10年2月1日
10年7月1日
10年12月1日
11年5月1日
11年10月1日
12年3月1日
12年8月1日
13年1月1日
13年6月1日
13年11月1日
14年4月1日
14年9月1日
15年2月1日

実施計画スケジュール


2013
3月
2014年度（H26年度）
4月
5月
6月
7月
8月
9月
10月
11月
12月
1月
2月
3月
新システム稼働
新システム搬入開始
Mpc半分撤去
全体システム
運用停止
AX+MDG3撤去
FS撤去
次期システム（ストレージ、FE、AGSなど）
次システム（MPC他）
ディスク装置＋FEP+SSC+GPU
FEP+SSC+GPU
テープ装置
システム停止
RICCシステム
Mpcの構成
変更準備など
停止を含むが
全系運転
UPC+AX+MDG3
停止
UPC稼働
UPC稼働
MPC稼働
MPC停止
MPC半分稼働
システム側イベント
施設側イベント
屋外工事
室内
電源室他
養生
既存室外機撤去
電源工事
チラー設置
室内機移設および撤去
配管撤去・新規敷設
工事準備作業
新規エアハン設置
吊り金具工事
バスダクト設置
電源工事
分電盤撤去
受電設備改修
バスダクト設置
空調機入れ替え
電源切替
完全停電

ACCC, RIKEN

Advanced Center for Computing and Communication



新空調・電源設備

- パッケージ型空調機から空冷式チラーで冷水を供給するシステムに変更。
  - フリークーリング、外気導入や井戸水などを利用して空調電力削減。
- 分電盤方式からバスダクトによる給電方式に変更。
  - レイアウトの自由度と電源工事の工期短縮に貢献。

HOKUSAI GreatWave

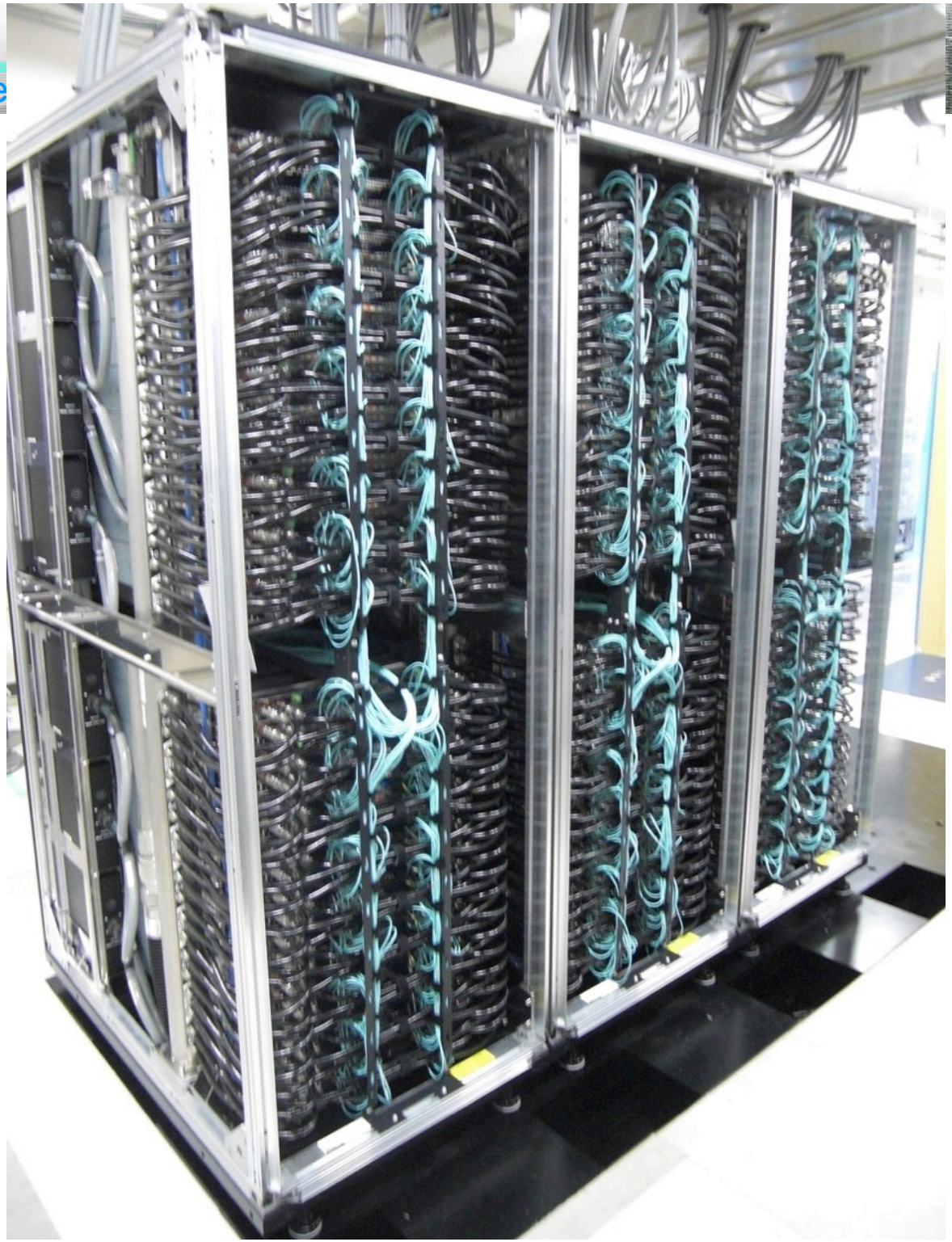
cation

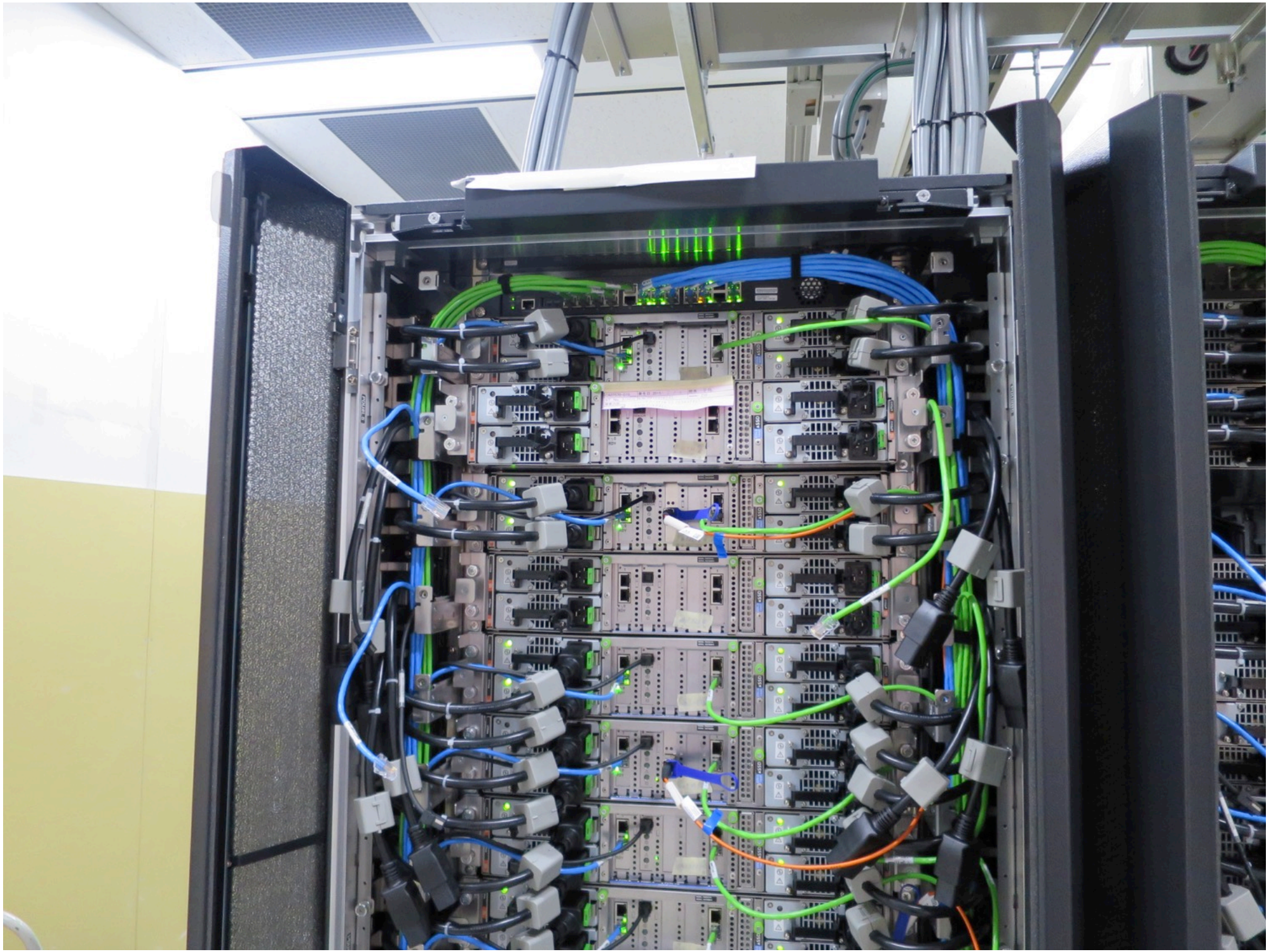

KAWASAKI
KAWASAKI


TOYOTA
COMPACT
GENEO 25


KAWASAKI


#3
PRIMERGY FX100





ACCC, RIKEN

Advanced Center for Computing and Communication

HOKUSAI GreatWave システム構成図

超並列演算システム
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

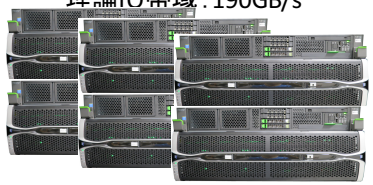
フロントエンド

高速広帯域ネットワーク
Mellanox SX6036×12（InfiniBand FDR） FBB構成

RICCシステム
of nodes: 589 (4712 cores)
- # of CPUs: 2/node (8cores/node)
  - CPU: intel Xeon X5570 2.93GHz
- Total mem: over 7TB
- Network: Infiniband QDR(4GB/s/node)

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

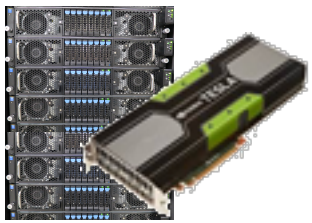


階層型ストレージ(7.9PB)
IBM TS4500＋TS1140×6
階層構成：GPFS＋TSM

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）
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）
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)

RICCとHOKUSAI GreatWaveのMPCの性能能力比較

	RICC-MPC/UPC (FY2009-)	GW-MPC(FY2015-)	
Performance	96TFLOPS	1PFLOPS	About 10 times
Total # of nodes / cores	1,024 /8,192	1080/ 345,560	About 4.3 times (Cores)
# of core/ node	8	32	4 times
Memory / node	12GB	32GB	About 2.5 times

MPC(RICC)
Air cooling
Over 30 racks

MPC(HOKUSAI GreatWave)
Water cooling
5racks

おわりに

- システムと施設設備の同時入替作業も無事完了。
- 4月1日からトライアル運用を開始した。
- 6月1日から本運用も開始されている。
- MPC（FX100）のジョブ充填率はすでに高い。
  - 5月期実績で85%以上。
- ただ、FX100は問題点もまだまだ存在する。
  - システムの細かな不具合がまだ取れ切れていない。
  - 不可解なソフトウェア仕様やバグも見受けられる。
  - 現時点判明している不具合は7月中には解決予定。
    - ただし、不具合の解消とまで行くかどうか。。