RICCからHOKUSAI GreatWaveへ

情報基盤センター

はじめに

- 2009年8月から運用していたRICC(RIKEN Integrated Cluster of Clusters)は2015年3月末で運用を停止し、4月から現システムが稼働を開始しました。
- 本発表では、
  - RICCの稼働状況や運用状況などを紹介
  - RICCと現システムの入替の合間を縫って実施した施設工事の概要の紹介。
  - 現システムの紹介。

システム構成について

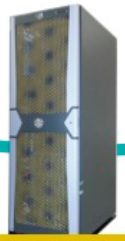
【システム構成】
超並列PCクラスタ＋GPUクラスタ＋専用機クラスタ＋大容量メモリ計算機


【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)

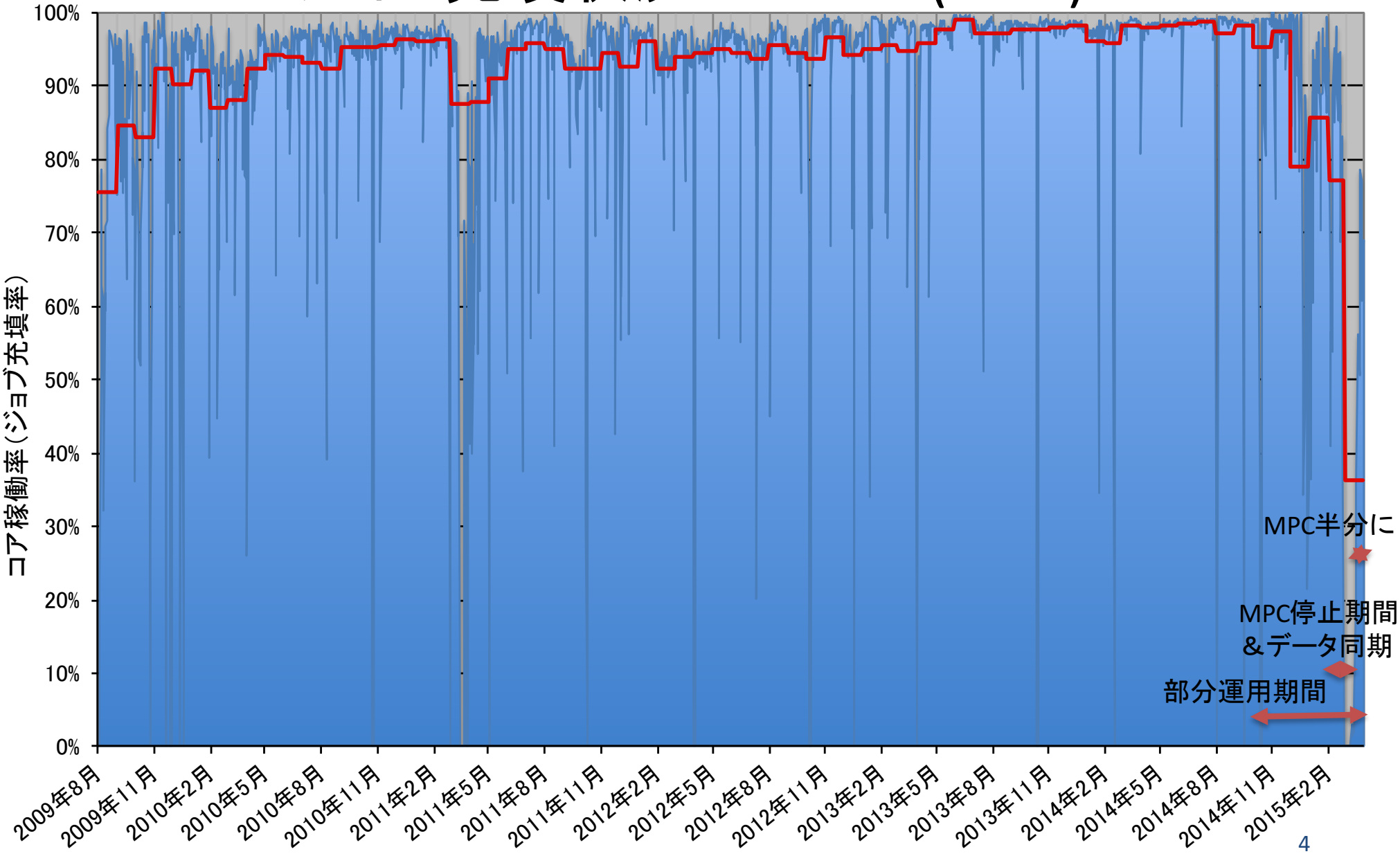
磁気ディスク装置
(550TB)



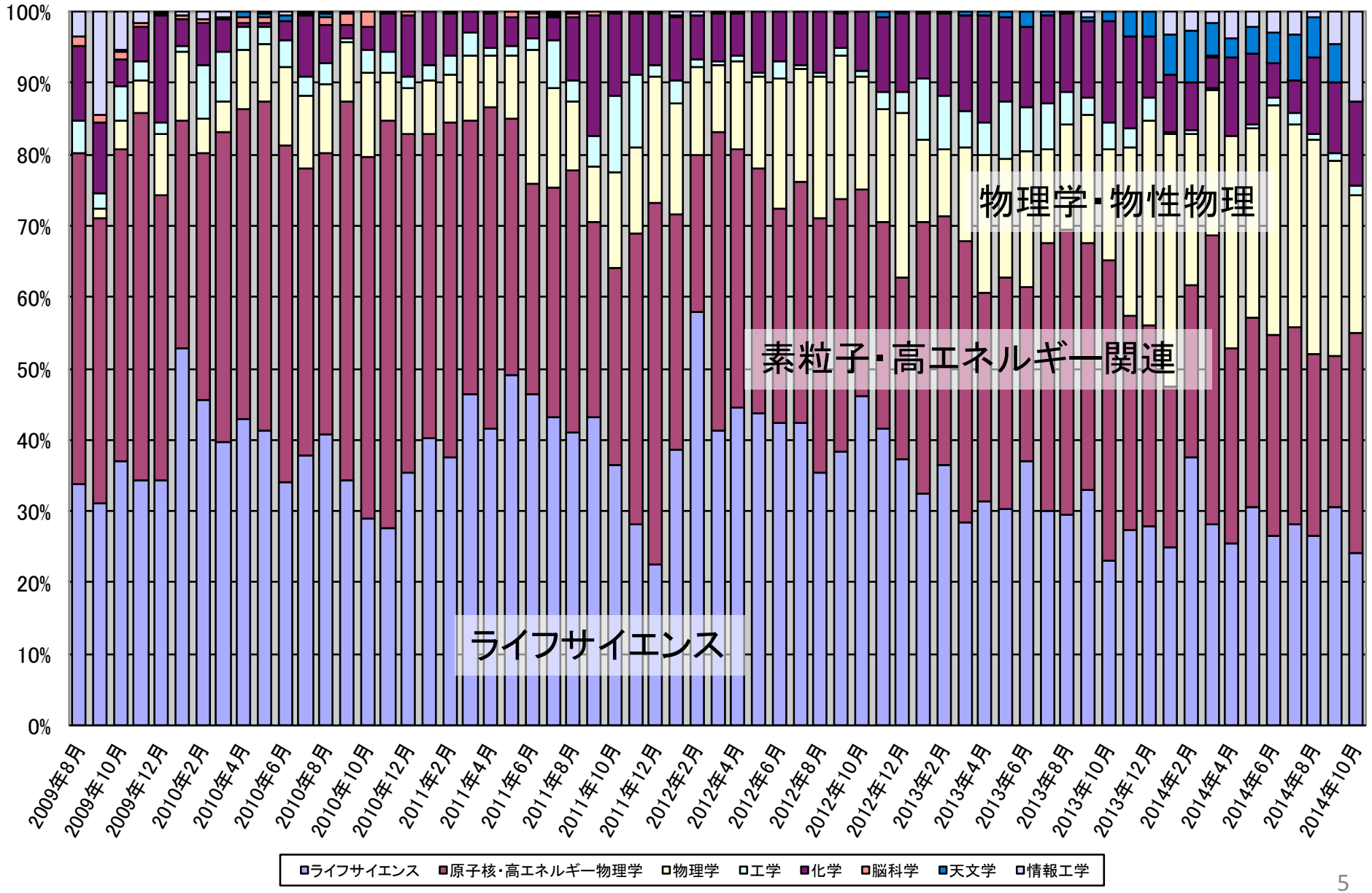
【大容量メモリ計算機】
0.24テラフロップス、512GBメモリ
＊1プロセスで500GB以上のメモリを利用可能

【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

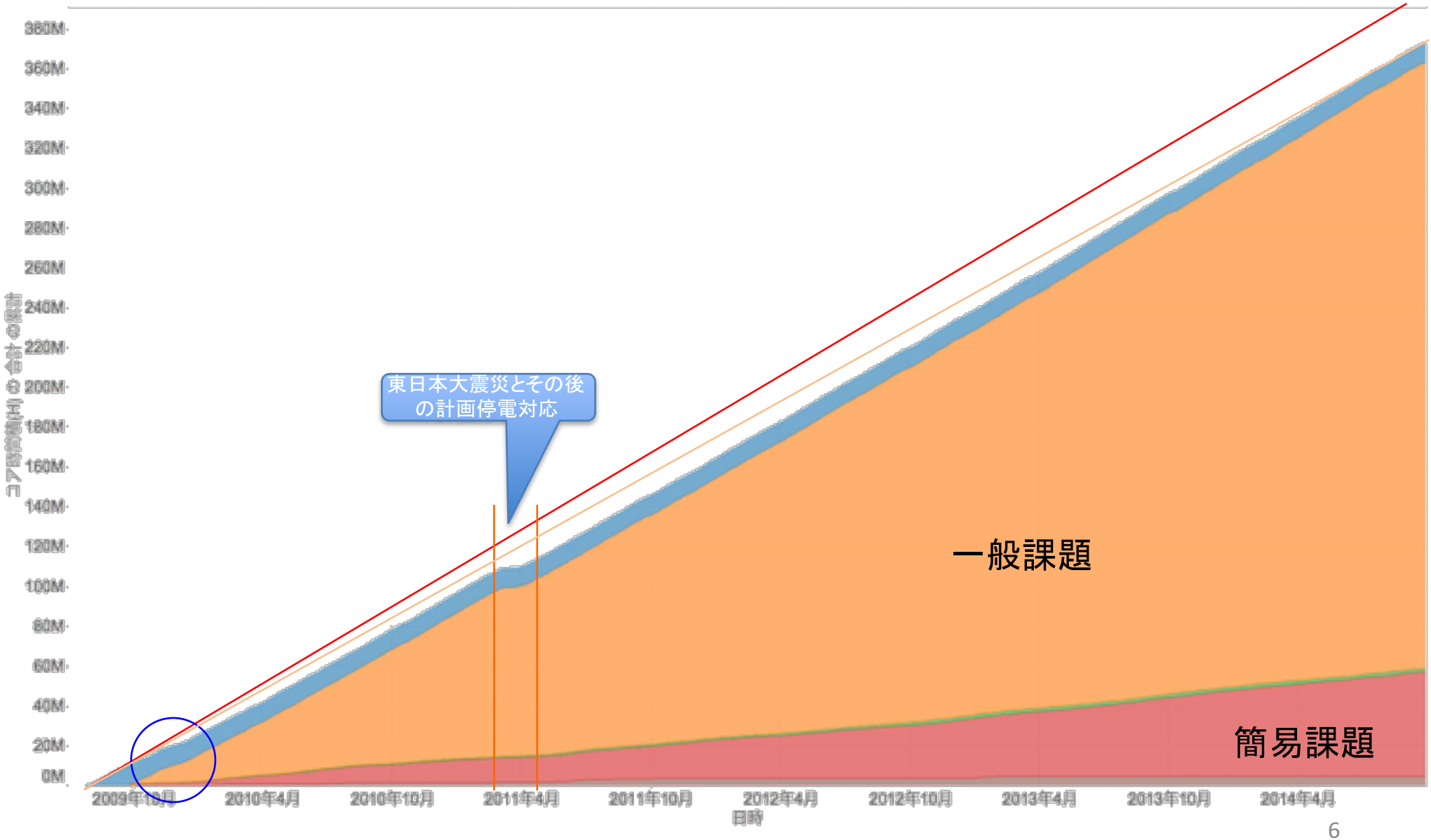
ジョブ充填状況について(MPC)



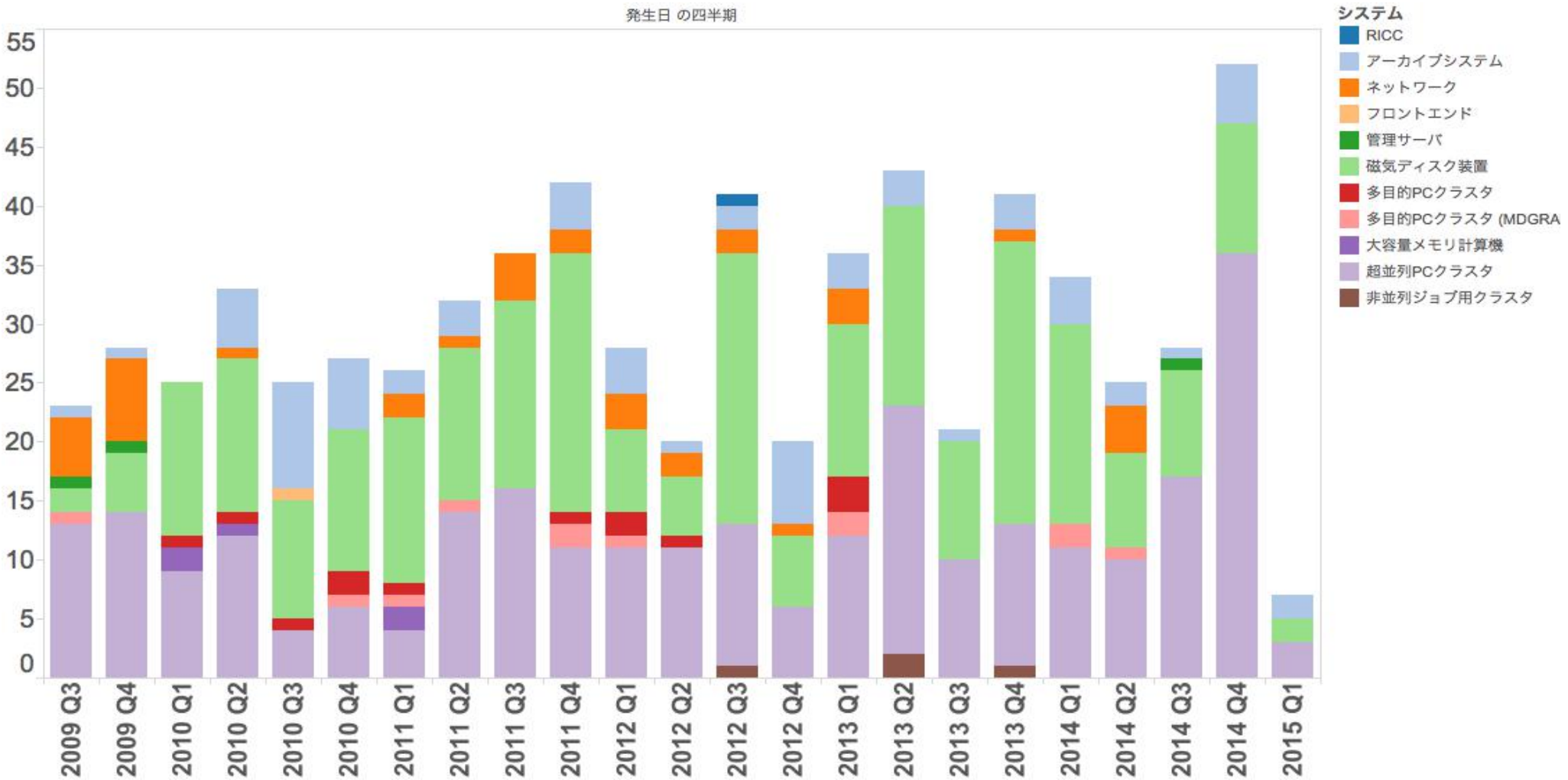
分野毎のコア時間利用比率



利用区分別コア時間累積



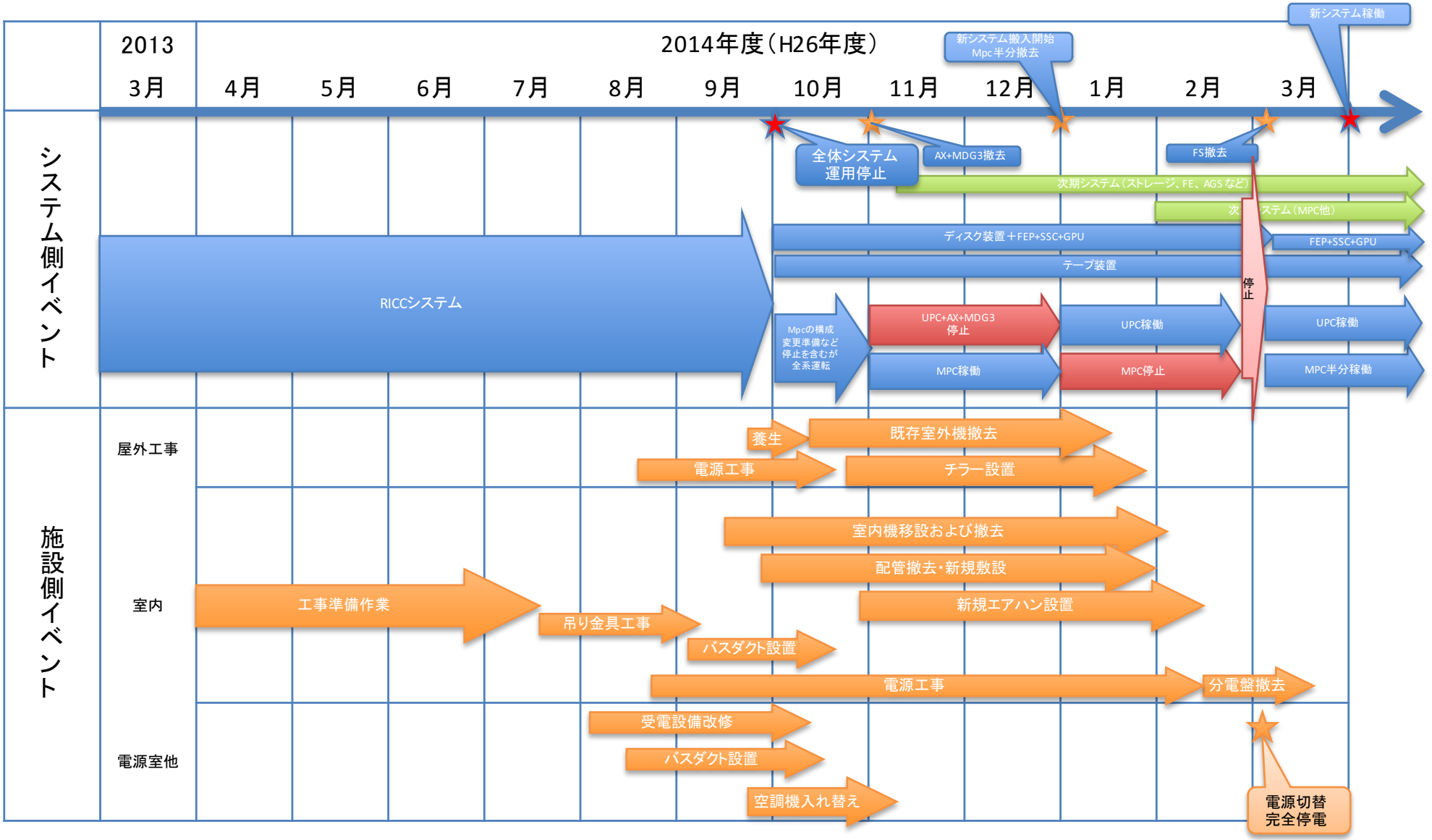
故障状況


発生日 の四半期
55
50
45
40
35
30
25
20
15
10
5
0
2009 Q3
2009 Q4
2010 Q1
2010 Q2
2010 Q3
2010 Q4
2011 Q1
2011 Q2
2011 Q3
2011 Q4
2012 Q1
2012 Q2
2012 Q3
2012 Q4
2013 Q1
2013 Q2
2013 Q3
2013 Q4
2014 Q1
2014 Q2
2014 Q3
2014 Q4
2015 Q1
システム
RICC
アーカイブシステム
ネットワーク
フロントエンド
管理サーバ
磁気ディスク装置
多目的PCクラスタ
多目的PCクラスタ (MDGRA
大容量メモリ計算機
超並列PCクラスタ
非並列ジョブ用クラスタ

現システムに向けたスケジュール

- H25年度
  - 受電設備改修工事
  - 空調と電源の改修設計
  - 現システムに向けたWGやアンケートや調達作業開始
- H26年度
  - 空調と電源の改修工事
    - 改修工事は4月から開始
  - システム調達
    - 8月に現システムの導入業者が決定
    - 工事の佳境は10月から
      - 旧空調機の撤去が始まる。
    - システム導入作業は11月から
      - データ移行用のストレージを導入。
  - どちらも終了は3月末

実施計画スケジュール

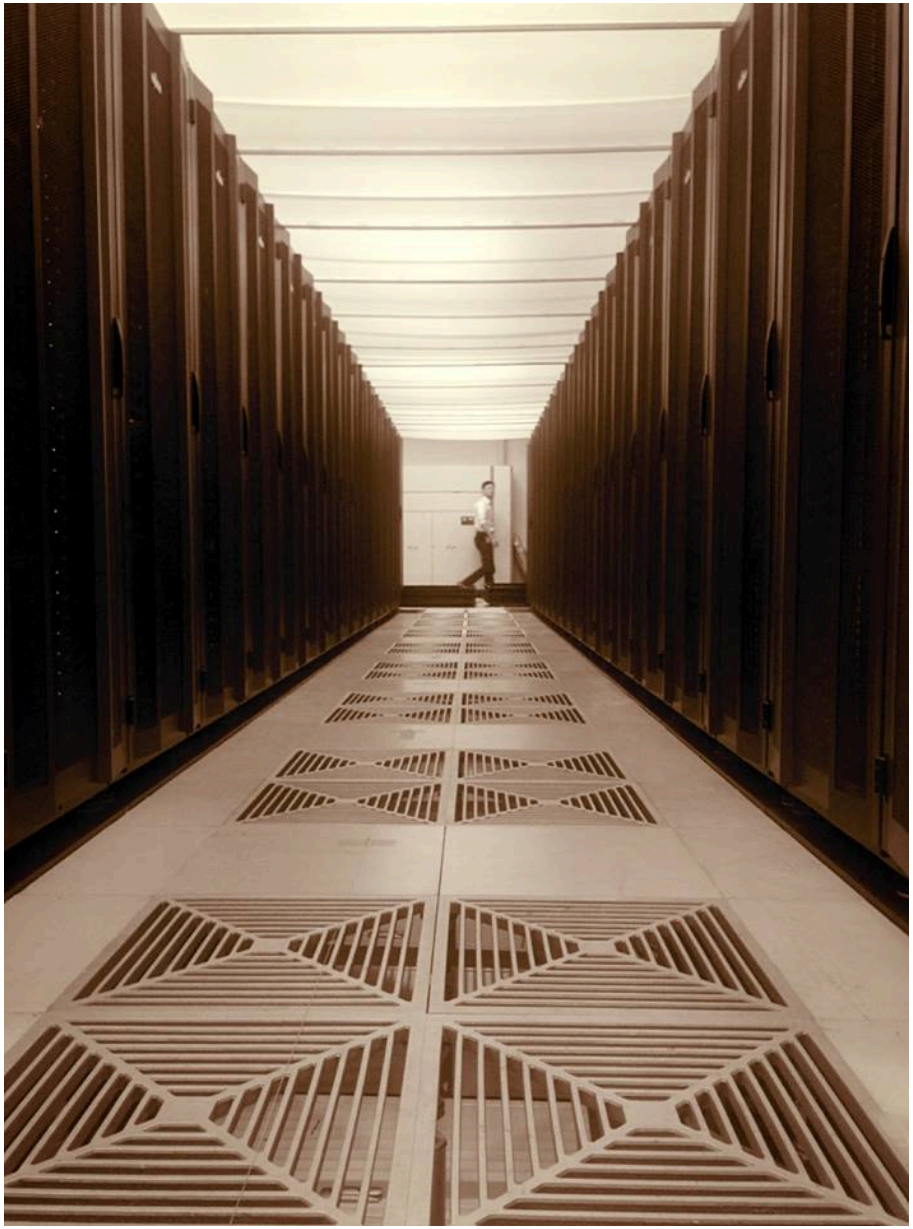

2013
3月
2014年度（H26年度）
4月
5月
6月
7月
8月
9月
10月
11月
12月
1月
2月
3月
新システム稼働
新システム搬入開始
Mpc半分撤去
全体システム
運用停止
AX+MDG3撤去
FS撤去
次期システム（ストレージ、FE、AGSなど）
次システム（MPC他）
ディスク装置＋FEP+SSC+GPU
FEP+SSC+GPU
テープ装置
停止
RICCシステム
Mpcの構成
変更準備など
停止を含むが
全系運転
UPC+AX+MDG3
停止
UPC稼働
UPC稼働
MPC稼働
MPC停止
MPC半分稼働
システム側イベント
施設側イベント
屋外工事
養生
既存室外機撤去
電源工事
チラー設置
室内
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
電源室他
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

今だから笑える話

- 室内空調機をだんだん止めていくと、吹き出しの流量が低下して、廃熱が前面に巻き込まれてきた。
  - ラック上部をプラ段で覆って対応。
- MPC側室内空調機が最小時5台（普段は8台でカバー）体制。
  - 250KW程度の負荷が裁けるはずが。。。
    - 全然吹き出し風量が足りない。
  - 室内温度は35度超えて推移。
    - 廃熱を空調機側に攪拌するために大型ファンを導入。
  - コールドアイル側は覆いを作成して、25度前後を維持していたが、ラックや筐体自体がホットアイル側の熱を吸収して熱くなる。
  - PSUが想定外に壊れた。
    - おそらくコンデンサが駄目になっていったのでしょう。（SEさん談：破裂音が聞こえたこともあったとか）

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。

施設整備の具体的な項目

- 電源設備更新計画
  - 2MW分の電力供給が可能
    - 750kVAキュービクル3系統(@2000A)
    - 分電盤方式ではなく、バスダクトによる給電を採用。
    - 天井下にバスダクトを設置。

- 冷却設備更新計画
  - 空冷型チラー(160kW×8台, 来年度8台追加予定)を新設
    - 冷却は水冷部分1.6MW、空冷部分400KW(現状で全て設置)
  - 冷水で空気を冷やすAHU(72.4KW分)×6台を設置
    - 現状:50馬力パッケージ型空調機(約55KW)×20台
  - 外気冷却および調湿のために外調機(DHU:9.8KW)を設置
  - 空冷型チラーにフリークーリングチラーを設置して、外気および井戸水による冷却を行えることにした。
    - フリークーリングとは、電力を使わずに冷たい空気などで水を冷やすこと。
    - 一部AHUにも井戸水を用いて空気を冷やせるものも。
  - 夏場の空調電力効率はPUEで1.5まで, 冬場は1.1～1.2の予定。
    - PUEとは、IT機器の消費電力の何%を冷却電力に利用しているかという指標。

現システムに向けて

現システム導入に向けての議論

- H24頃から検討WGを発足して議論
  - システム構成
    - H26年頃の想定されるシステムや消費電力など。
      - 冷却は水冷にならざるを得ないか。受電能力をどうにかできるのか。
    - システムの方向性について
      - PCクラスタで行くのか、商用MPPで行くのか？
      - RICCの10倍以上の性能は必要だろう。
    - メモリを多めに搭載したIAサーバ(UPCみたいな)は必要でしょう。
      - GPUの搭載は？
    - 大容量メモリのサーバは必要でしょう。
    - ステージングとかVPNは止めたいとか。
  - システム運用の方向性
    - 現状5年単一システムリース
    - 全く利用出来ない期間が5年に1回1ヶ月強の期間存在する。
      - どうにかしたい。
    - 5年間、全くシステム増強ができなかった。
    - システムの入替が5年毎では新しいデバイスが使えない。
    - WG発足後、施設設備の更新の必要性も出てきた。。

システム導入の考え方の変更

- 既存システムまでは5年間の単一システムの運用だった。
- 次期システムは約7年間の運用期間とする。
- 1システムを2年程度の時間差で2回に分けて立ち上げる。
  - システム完全停止期間の縮減。
  - システムの需要増加トレンドへの対応。
  - 空調システムのリプレースへの対応。
- 次のシステムは追加とし、一体運用する。

HOKUSAI GreatWave

システム名称

- HOKUSAI (北斎) is the most famous “Ukiyo-e(浮世絵)” artist.
  - First system name is GreatWave (浪裏).
  - Second system name is BigWaterfall (瀑布).

HOKUSAI GreatWaveシステム

- スケジュール
  - 2014年9月末
    - RICCの正式運用停止。
    - その後はできる限り運用し、部分的にシステム停止の部分運用へ。
  - 2014年11月
    - オンライン・ストレージ・テープ・アーカイブ・ACSL搬入
    - ファイルシステム構築・テスト
  - 2014年12月
    - 12月後半からRICCのストレージと同期を開始。
  - 2015年1月
    - テープ上のデータの移行を開始。
  - 2015年2月
    - MPC・ACSG搬入。
    - 2月末ストレージの最終同期のため1.5週間停止。
  - 2015年3月
    - テスト利用を実施。
    - 3月末、検収。

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

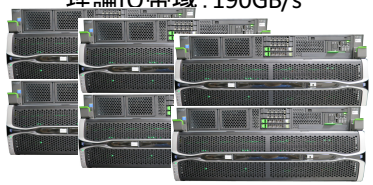
RICCシステム

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)

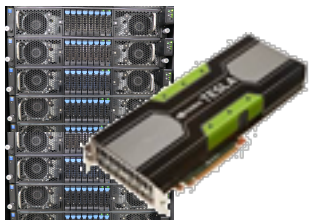
IBM TS4500＋TS1140×6
階層構成：GPFS＋TSM

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)

システム概要

- 演算システム
  - 1PFLOPS級超並列演算システム（GWMPC）
    - 高性能汎用CPUと高性能メモリによる計算ノードを広帯域低遅延インターコネクトで接続した超並列演算システム。
  - 6TFLOPS/nodeの高性能計算ノードのクラスタ（GWACSG）
    - 4×GPUを搭載した高性能IAノードのGPUクラスタ。
  - 大容量メモリサーバ（GWACSL）
    - 1TB/ノードを搭載した大容量メモリサーバ。
- 広帯域ファイルシステム
  - 容量2PB
  - 190GB/sの理論帯域
  - FEFSベース
- 大容量・低消費電力のHSMシステム
  - 容量8PB（テープ）、300TB（キャッシュ）
  - 消費電力10KW以下
- RICC
  - MPC（RICC時代の半分）, UPCは残ります。

cation

ACCC, RIKEN

Advanced Center for Computing and Communication

ACCC, RIKEN

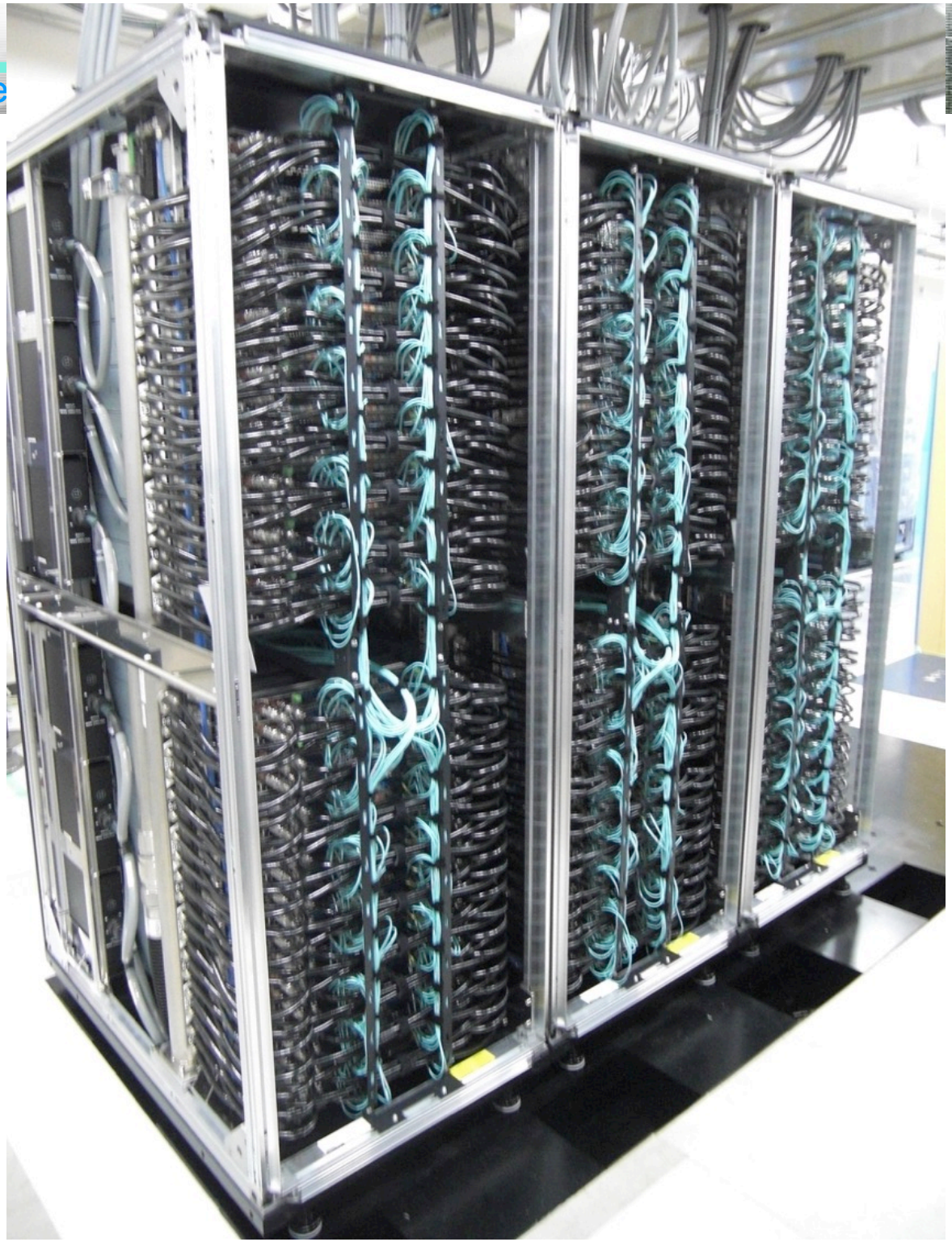
Advanced Center for Computing and Communication

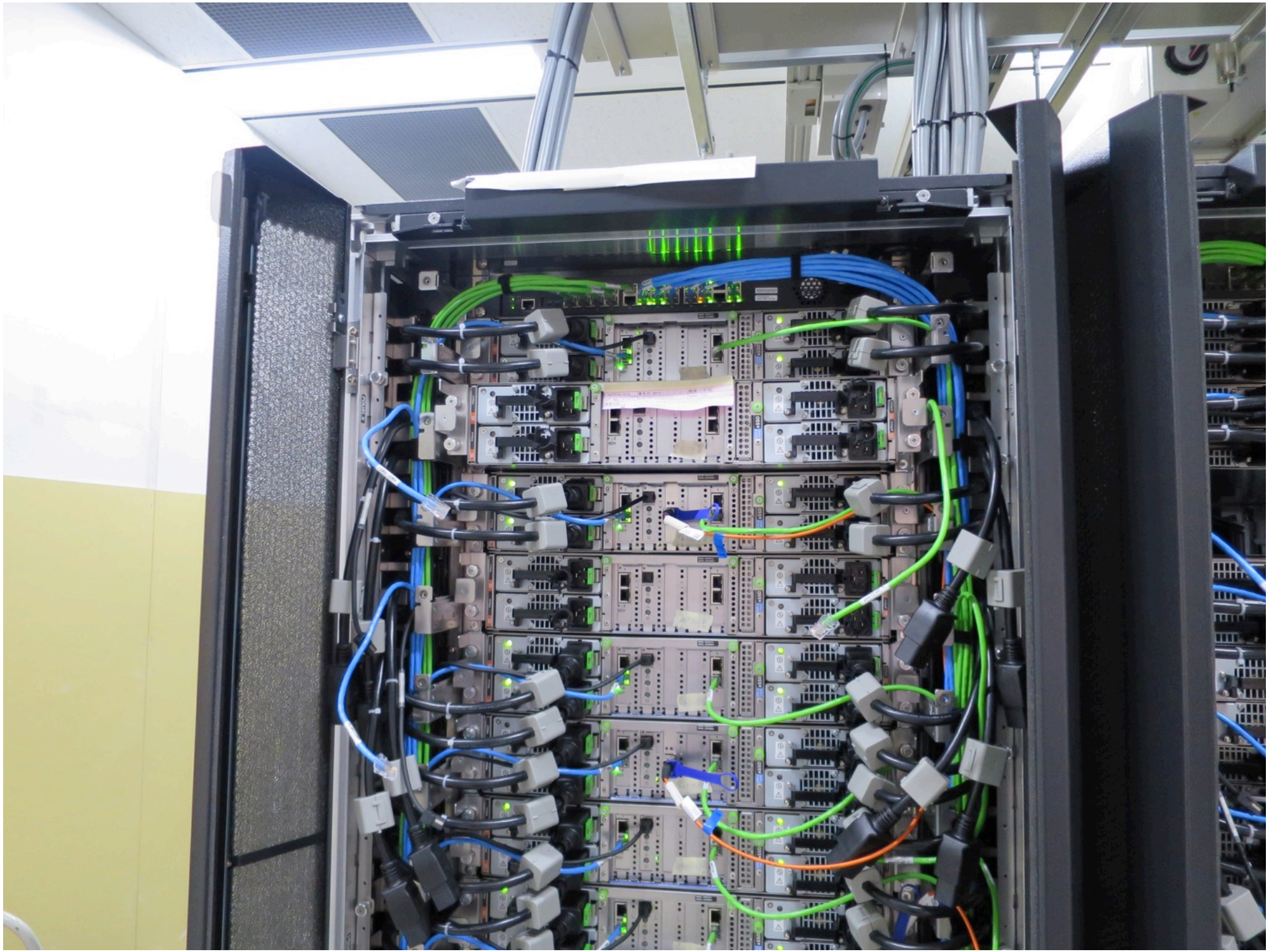

KAWASAKI
KAWASAKI


TOYOTA
COMPACT
GENEO 25


KAWASAKI


#3
PRIMERGY FX100





ACCC, RIKEN

Advanced Center for Computing and Communication

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

システム不具合の状況（特にFX100）

- 導入当初からかなりあった。
  - ハードウェア
    - メモリ（信号エラー）
    - ネットワーク・光リンク（リンクダウン、リンク縮退）
  - ソフトウェア
    - 低レベル通信ライブラリ（大規模並列時に通信フリーズする場合がある）
      - 発生源は上記のネットワーク障害
      - これは京やFX10で取れてても良いバグのように思われる。
      - 現状、ハードウェアバリア機能をOFFとして運用回避。
    - OSのメモリ管理周り
- 7月中には、顕在化している不具合を解決する予定。

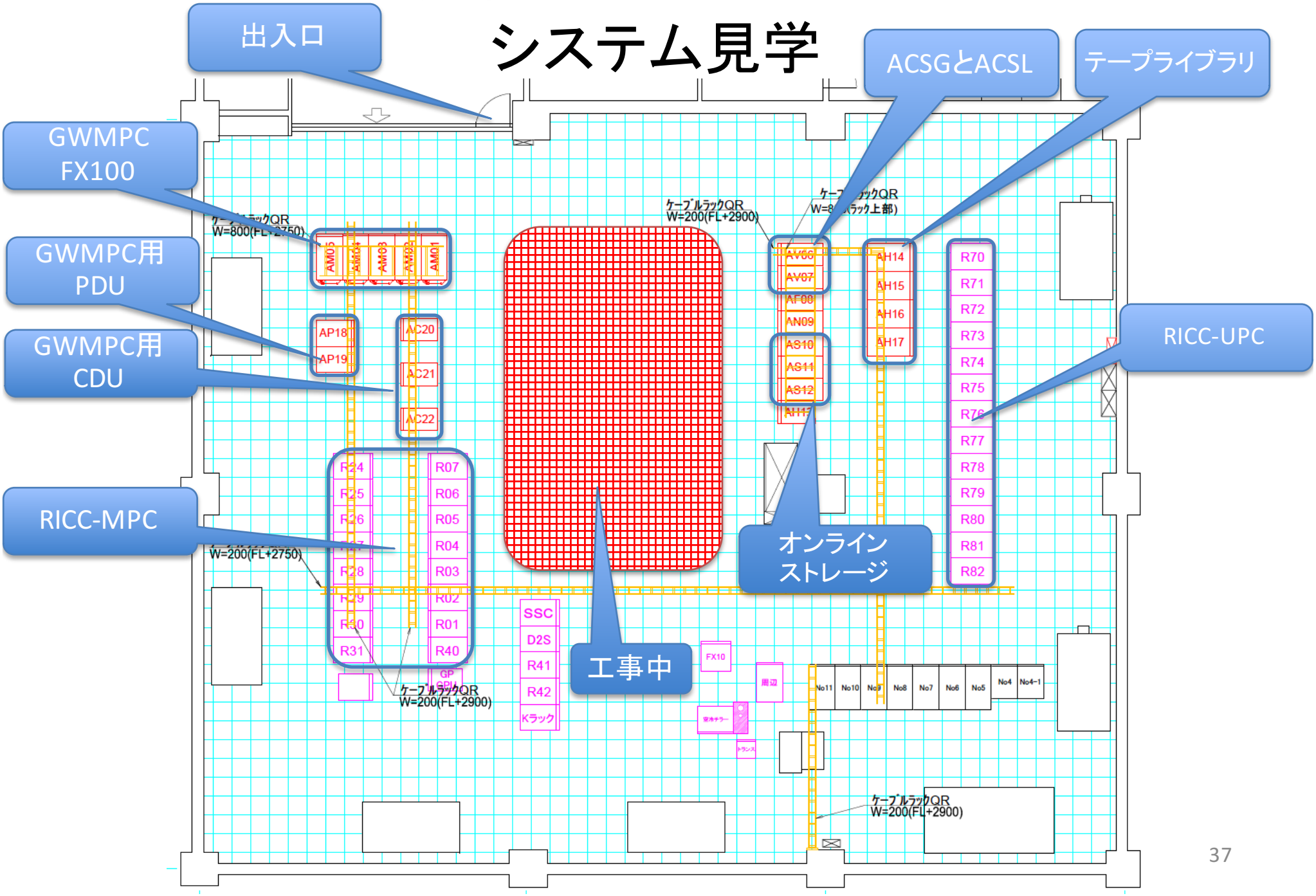
おわりに

- 6月1日から本運用を開始
  - 一般課題30、簡易課題約90、全体で120課題程度
  - 登録されている利用者は約280名
  - 5月のGWMPCのジョブ充填率は80%超
    - すでにかなり利用率は高いです。
- 29日にFX100向けの実践チューニング講習会を開催。
  - 原則として理研の方対象です。

- RICCよりも効率的で柔軟でかつ公正な運用を心がけて運用を続けていきたいと考えています。
  - ご意見やコメントなどあれば以下のアドレスまで。

hpc@riken.jp

システム見学


出入口
ACSGとACSL
テープライブラリ
GWMPC
FX100
GWMPC用
PDU
GWMPC用
CDU
RICC-UPC
RICC-MPC
オンライン
ストレージ
工事中
ケーブルラックQR
W=800(FL+2750)
ケーブルラックQR
W=200(FL+2900)
ケーブルラックQR
W=200(FL+2750)
ケーブルラックQR
W=200(FL+2900)
ケーブルラックQR
W=200(FL+2900)
AM05
AM04
AM03
AM02
AM01
AP18
AP19
AC20
AC21
AC22
AV06
AV07
AF08
AN09
AS10
AS11
AS12
AH14
AH15
AH16
AH17
R70
R71
R72
R73
R74
R75
R76
R77
R78
R79
R80
R81
R82
R24
R25
R26
R28
R29
R30
R31
R07
R06
R05
R04
R03
R02
R01
R40
SSC
D2S
R41
R42
Kラック
GP
GPU
FX10
周辺
No11
No10
No9
No8
No7
No6
No5
No4
No4-1