

1 理研スーパーコンピュータ・システム

1.1 はじめに

理化学研究所情報基盤センター（以下、「当センター」）は、新スーパーコンピュータ・システム HOKUSAI システムを 2015 年 4 月から稼働させた。5 月までの 2 か月間はトライアル期間として利用者による新システムの試行および運用準備・試用を行い、6 月から本運用を開始した。

HOKUSAI システムからは、システムを 2 段階で導入する運用となり、今回は第 1 期システムである HOKUSAI GreatWave (HGW) の稼働開始となった。第 2 期システムである HOKUSAI BigWaterfall (HBW) は 2 年程度後に稼働開始となる予定である。このことにより、システムが全く使えない期間を短くし、日進月歩の勢いで進歩する計算機器やシステムに追随し、最新の計算資源を提供することが可能となる。

また、RICC システムの半分程度の計算資源を、HOKUSAI システムの一部として引き続き稼働させることとした。HGW システムの主計算資源は RICC システムと計算機アーキテクチャが異なるため、アプリケーションによっては HGW の利用が非効率になる場合があるため、利用者がアプリケーションに適したシステムを使い分けることができるようになっている。

1.2 HOKUSAI システムと運用スケジュール

今年度から HOKUSAI システムの運用を始めたが、そのシステム構成を図 1 に示す。今年度の HOKUSAI システムは、新たに導入された HGW システムに、半分の規模になった RICC システムからなる。フロントエンドは HGW と RICC で分かれており、利用者は HGW のフロントエンドには直接アクセスできるが、RICC のフロントエンドにアクセスする際には、一旦 HGW のフロントエンドにアクセスした後、RICC のフロントエンドにアクセスするという手順となる。ファイルシステムについては共通で、HGW システムのファイルサーバに RICC システムからもアクセス可能になっている。

運用スケジュールの概要を図 2 に示す。歴代のスーパーコンピュータ・システムで RICC システムまでは 1 つのスーパーコンピュータ・システムは 5 年間の運用を行ってきたが、HOKUSAI システムからは、全体システムの運用期間を 7 年間とし、2 段階のシステムを 2 年程度ずらして立ち上げ、各段階のシステムは 5 年間の運用を行う。

2 段階の運用にするのは、システム完全停止期間の最小化と新技術への迅速な対応の 2 つの目的があるからである。今までのシステム更新時期においては、ハードウェアの物理的な入れ換えやデータ移行などを行うために 1 か月程度の全停止を行う必要があったが、2 段階の立ち上げを行うことにより完全停止期間を 1 週間以内にできると考えている。また、2、3 年に一度新システムを導入することにより、より最新の技術の計

算機を提供することが可能となる。

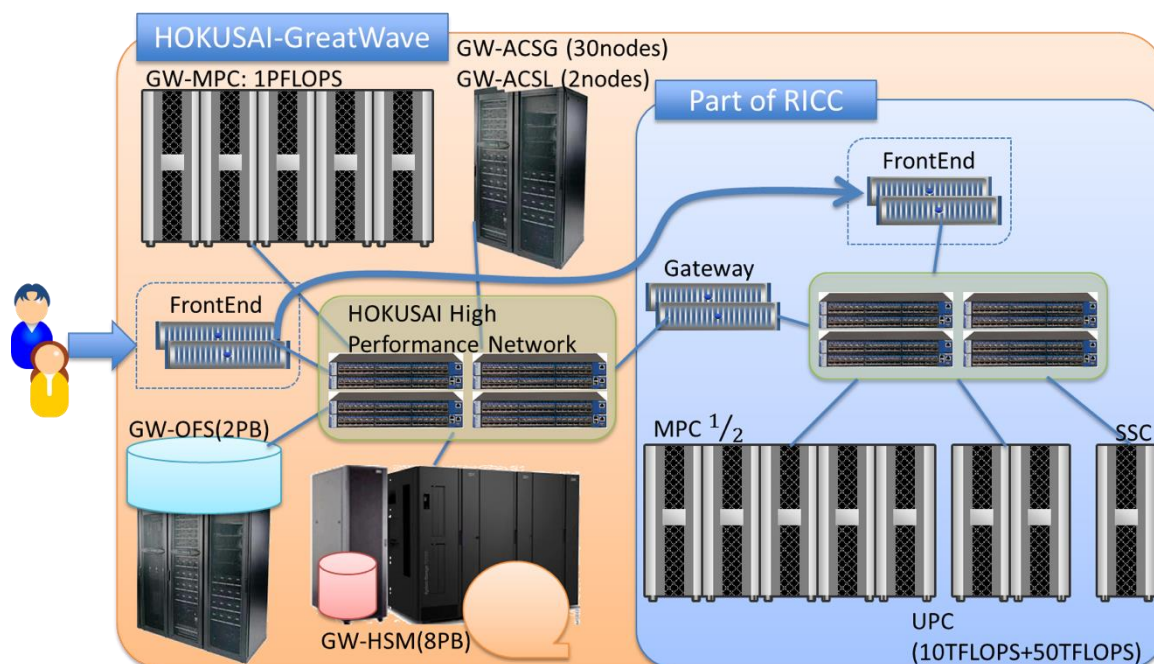


図 1 HOKUSAI システムの概要

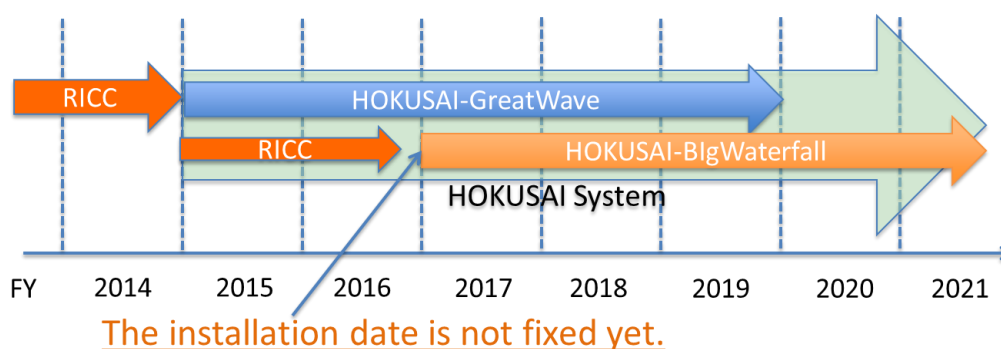


図 2 運用スケジュール概要

1.3 HGW システムの概要

HOKUSAI システムの第 1 期システムとして稼働開始された HGW システムは、以下のコンセプトの元に設計を行った。

1. 強力な汎用 CPU とメモリとインターコネクトを用いた超並列演算環境の提供
2. GPU を用いた強力な単一ノードの計算環境の提供
3. 大規模な共有メモリ空間を持つ計算環境の提供
4. 高速ファイル IO が可能な広帯域ファイルシステムの提供

5. 大容量・低消費電力で利便性の高いアーカイブシステムの提供
6. 並列計算やファイル IO でボトルネックが発生し難い低遅延・広帯域ネットワークの提供

調達の結果として、上記に対応する 3 種類の計算資源と 2 種類のストレージ、高速ネットワークのシステムとなった。

1. GW-MPC : 超並列演算システムとして、スーパーコンピュータ「京」と互換性のある Fujitsu PRIMEHPC FX100 を 1,080 ノード (1PFLOPS)
2. GW-ACSG : 4GPU を搭載可能な SGI C2010G-RP 演算サーバ 30 ノード(別途調達した K20X をノード当たり 4 枚、合計 120 枚搭載した)
3. GW-ACSL : 1TB のメモリを搭載した Fujitsu PRIMERGY RX4770M1 サーバ 2 ノード
4. Online storage : Fujitsu Exabyte File System(FEFS)による並列分散ファイルシステムにより 190GB/s の広帯域を実現した、実容量 2.2PB の NetApp5600 からなるストレージ
5. HSM storage : IBM GPFS(General Parallel File System)+TSM(Tivoli Storage Manager)により HSM 構成し、ディスク容量 300GB、テープ容量 7.9PB のシステム
6. High bandwidth network : Infiniband FDR を FBB(Full Bi-section Bandwidth)で構成した高速ネットワーク

2 理研スーパーコンピュータ・システム HOKUSAI システムの運用報告

2.1 HOKUSAI システム稼働開始と初年度の運用概要

HOKUSAI システムは本年度4月から運用を開始し、4,5月はトライアル運用を行い、6月から本運用を開始した。トライアル運用は新しい HGW システムに利用者に慣れてもらうための期間であり、運用に向けたパラメータ調整やシステムの不具合をあぶり出し、修正するための期間でもあった。ただし、後述するように HGW システムのハードウェアの故障が本運用開始後も多発し、早期に修正を計る必要があった。

6月からの本運用では、利用開始当初から想定以上に利用が進み、計算リソースが逼迫することとなった。これは、HGW の主要な計算資源である GW-MPC (FX100) と共通のアーキテクチャであるスーパーコンピュータ「京」やその互換機でアプリケーションの開発が進んでおり、大規模な利用者による利用が殺到したためと考えられる。その結果として、GW-MPC は大変混雑することになり、ジョブ実行までに時間がかかるなど利用者に多大な不便を掛けることになった。そのため、来年度からは課題審査の仕組みを大幅に変え、特に大規模利用については審査を丁寧に行うこととする。

2.2 課題審査と利用者数統計

本年度は4月から新システムである HOKUSAI GreatWave の運用を開始し2ヶ月間のトライアル期間を経て課題募集を行った。トライアルには98グループが参加した。一般・専有利用審査は前システム RICC の課題審査のように複数項目（科学的価値）の採点を行わず総合的に採択すべきかどうか、採択された課題の中で特に優先されるべきかどうかという観点で行った（表1）。簡易利用については6月からの10ヶ月間内で随時申請を受け付けた。月毎の課題採択数を図3に、課題数の推移を図4に示す。4、5月の課題はトライアル課題であり、6月以降本運用が始まり2016年3月末時点で156課題が利用していた。図5に役職別の課題数、図6に分野別の課題数を示す。

表 1 課題応募・採択数

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

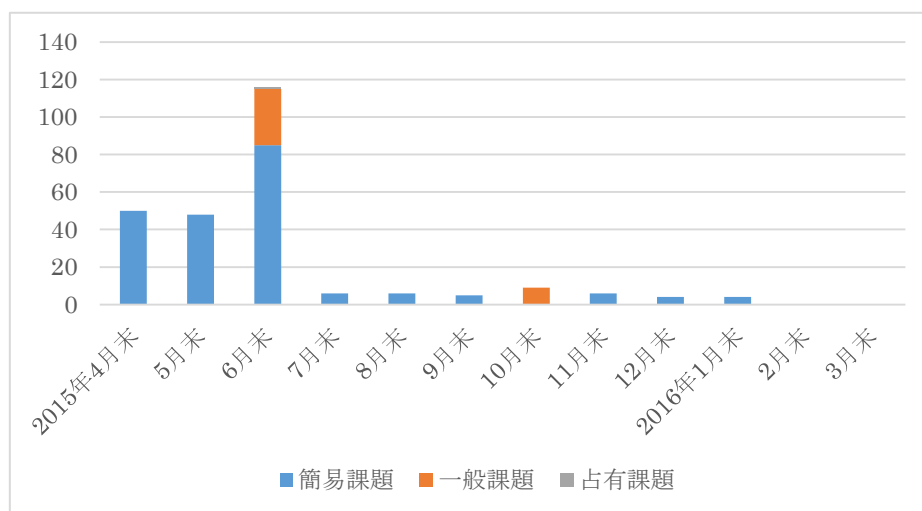


図 3 課題採択数推移

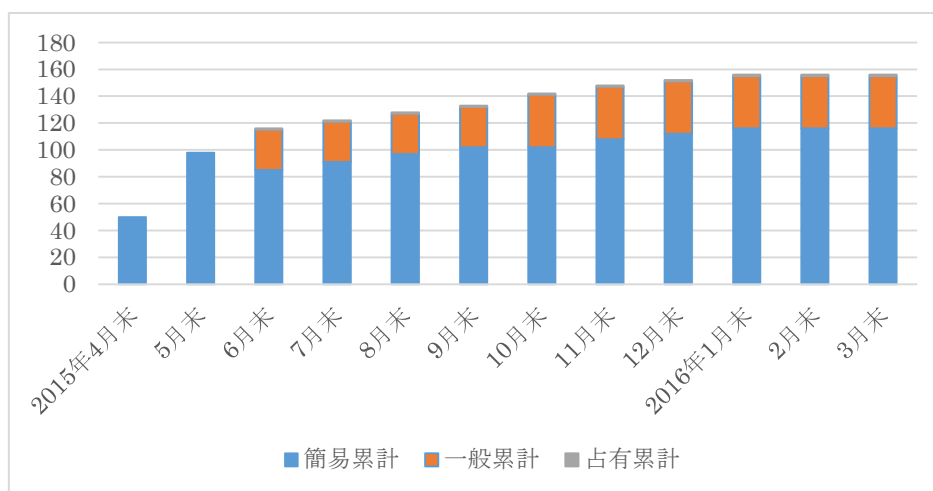


図 4 課題数累積推移

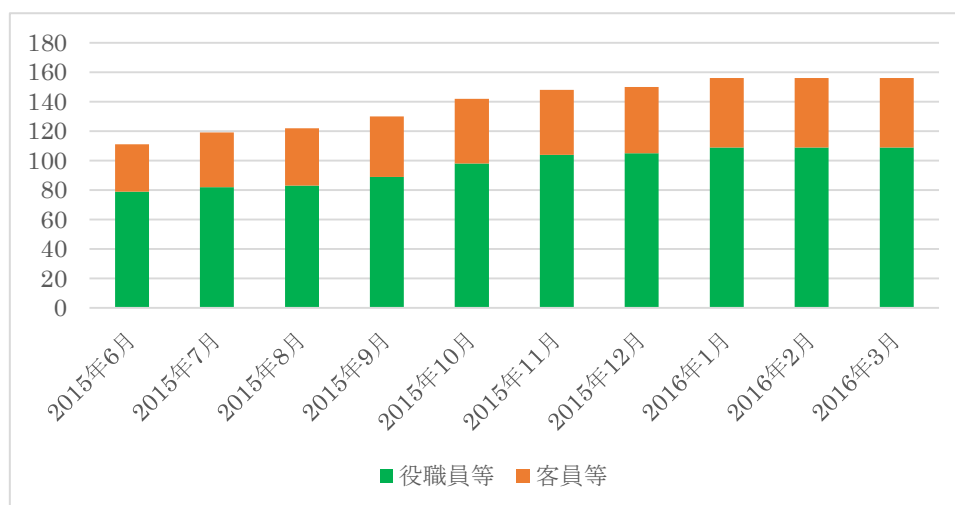


図 5 役職別の課題数累積推移

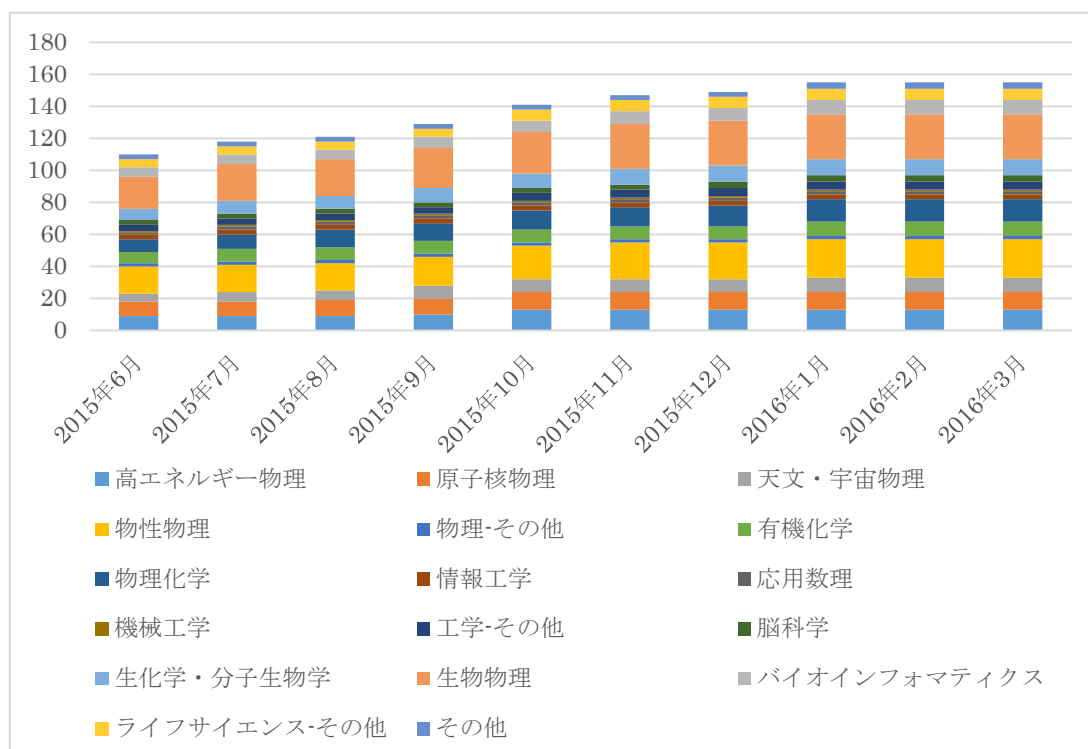


図 6 分野別課題数累積推移

図 7 に役職別のアカウント数を、図 8 にセンター別のアカウント数を示す。また、課題開始後長期間にわたりログインを行わなかった利用者については、資源利用予測が難しくなるため、課題のメンバーから登録削除する措置を行った。対象となったのは 49 アカウントであった。課題募集時に使用予定のないメンバーを含む申請を行わないよう注意を促しているが次年度は代表者に課題の適切な管理について理解を促したい。

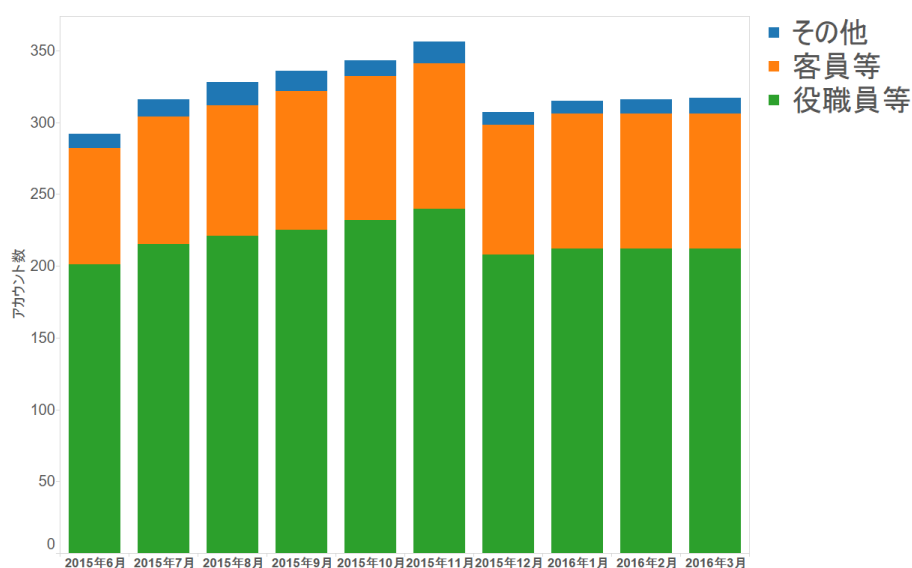


図 5 役職別アカウント数累積推移

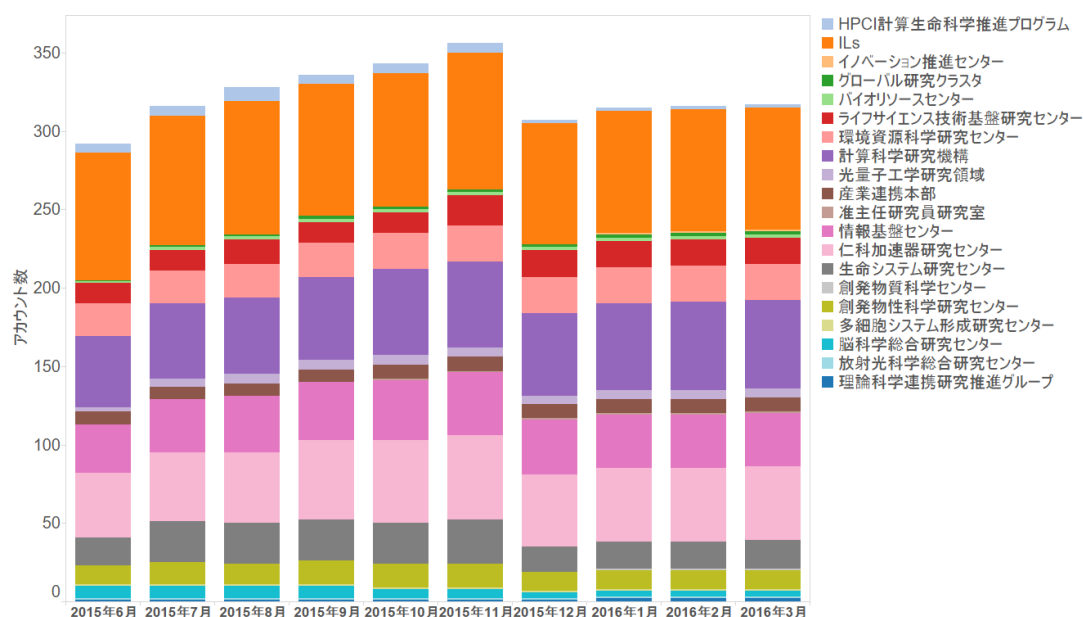


図 6 センター別アカウント数累積推移

2.3 HGW システムの稼働状況

HGW システム全体の利用状況を図 9 に示す。簡易利用は赤で、一般利用は緑、システムの性能測定に使った特殊な課題を紫で色分けしてある。なお、4,5 月はトライアル期間で、トライアル利用の課題は簡易利用として分類してある。6 月の本運用開始当初から 90%程度と高い稼働率で、その後も 90%を超える高い値で推移した。また、簡易利用は 10-20%で収まるような設計を行っているが、本運用開始後はその通りの利用状況となった。

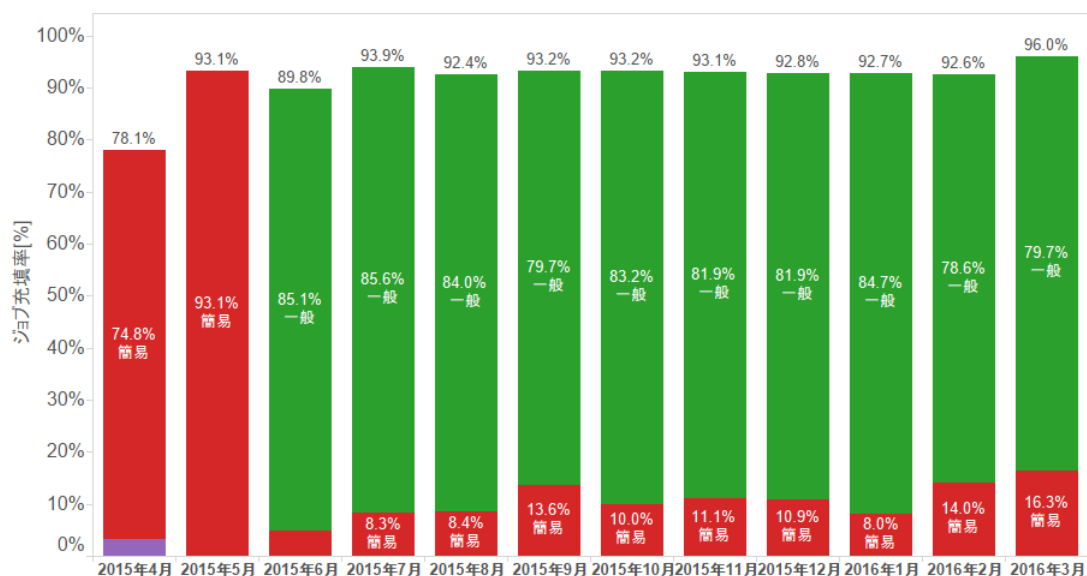


図 9 HGW 利用状況

HGW のコア時間消費状況を分野別で示したものを図 10 に示す。高エネルギー物理と原子核物理で総コア時間の半分を消費しており、物性物理、ライフサイエンス、物理化学の分野の利用が次に多くなっている。

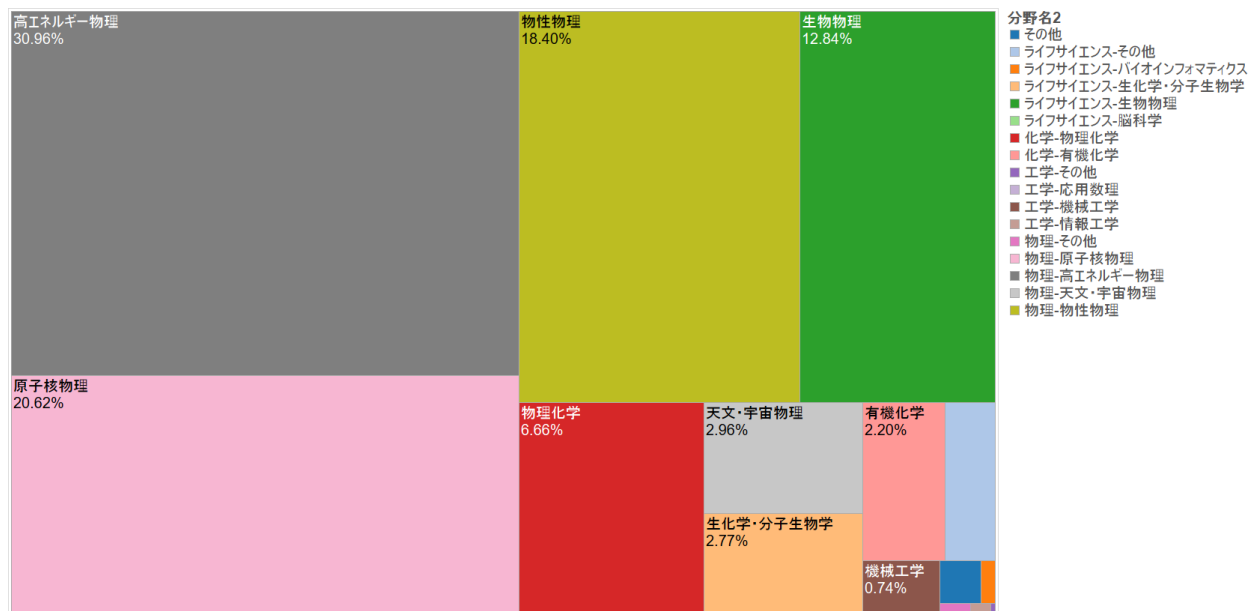


図 10 分野別での HGW の総コア時間消費の概要

課題毎の HGW の割当コア時間に対する消費割合を図 11 に示す。緑が一般利用で、赤が簡易利用である。一般利用については一部で高い消費率の課題もあったが、GW-GW-MPC が混んでいたこともあり 40%程度の消費率となった課題が多数あった。割り当てたコア時間の半分以下しか使えない課題が数多く、利用者には大変な不便を掛けることになってしまった。このようなことを繰り返さないために、来年度からはスーパーコンピュータの申請方法や審査を大きく変えることとなった。簡易課題については、10 課題程度が 50%程度の消費率となり、次の 20 課題程度が 10-20%程度の消費率で、残りの課題はかなり低い消費率となっていた。

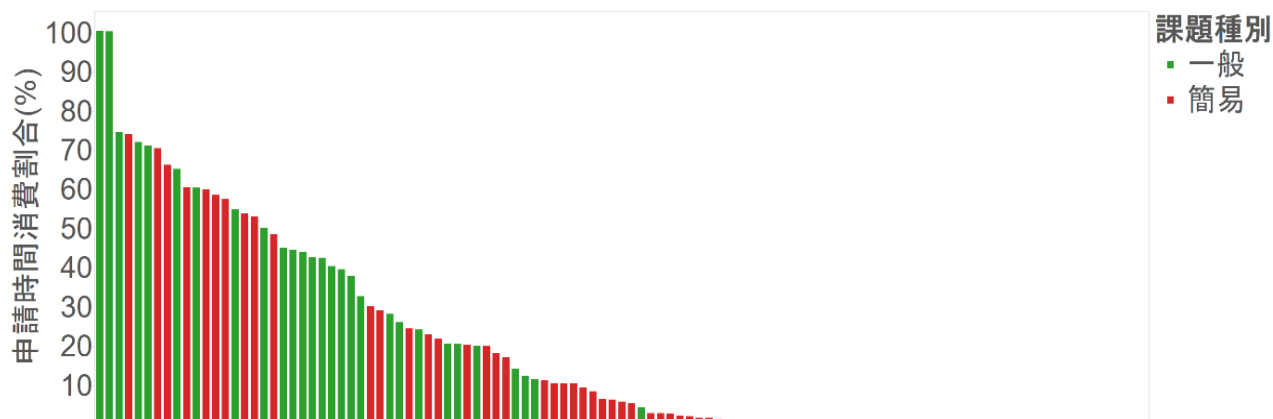


図 11 課題毎の割当コア時間消費割合

2.4 個別システム稼働（ジョブ稼働）状況

HGW（図 12-14）と RICC（図 15-16）の個別のシステムの稼働率を以下の図に示す。主要な計算資源である GW-MPC も RICC-MPC は、本運用開始直後から概ね 90%を超える高い稼働率を示している。その他の GW-ACSG、GW-ACSL、RICC-UPC は月によっては、80%以下の稼働率となり比較的空いていることもあったが、基本的には高い稼働率で推移した。

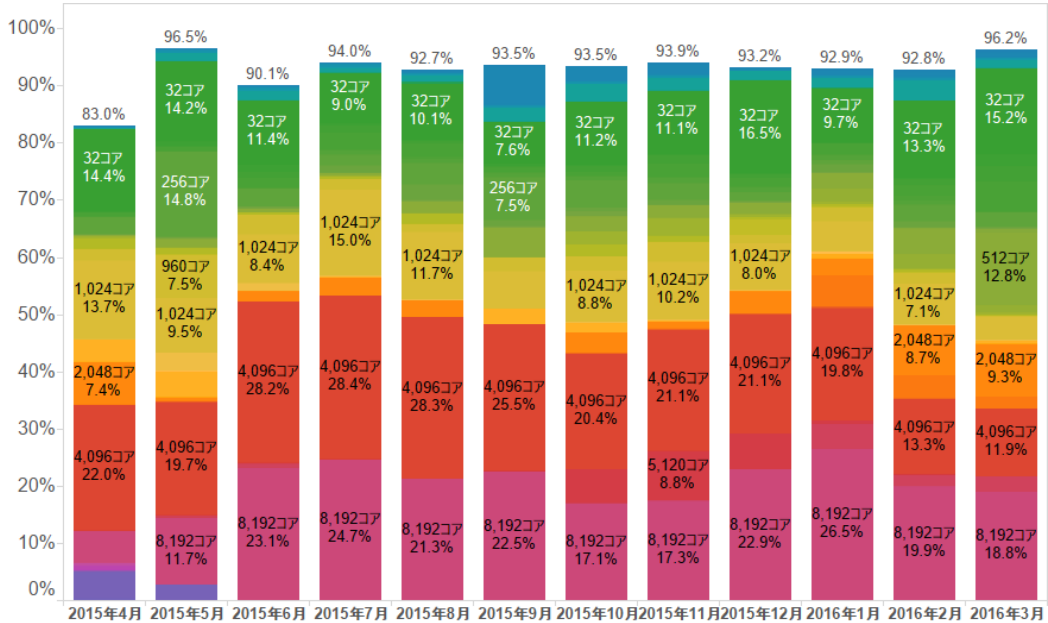


図 12 GW-MPC 稼働率（ジョブ充填率）

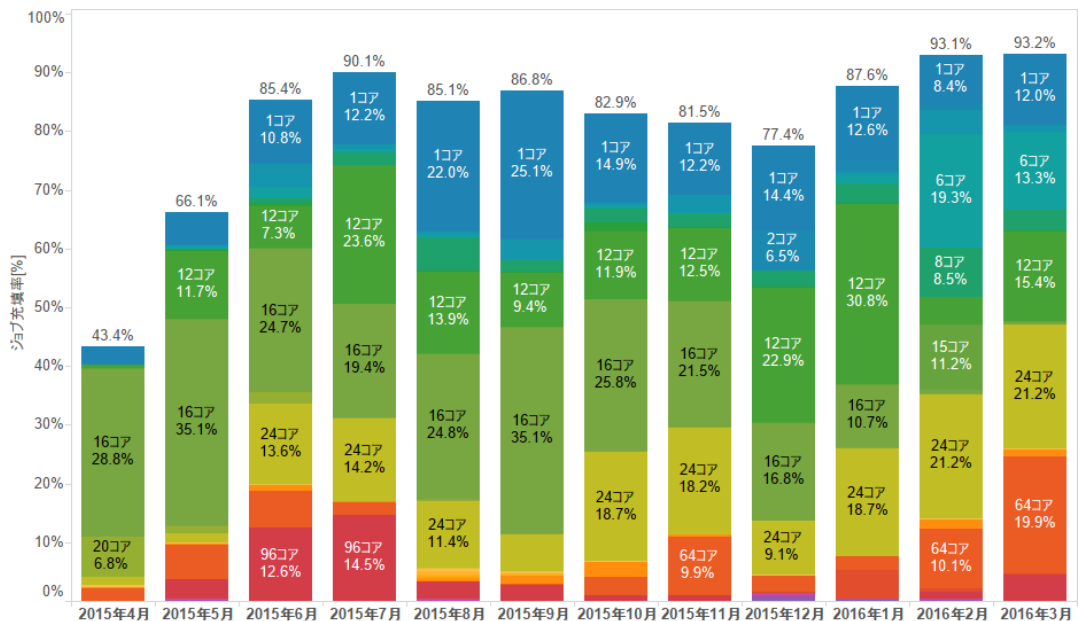


図 13 GW-ACSG 稼働率（ジョブ充填率）

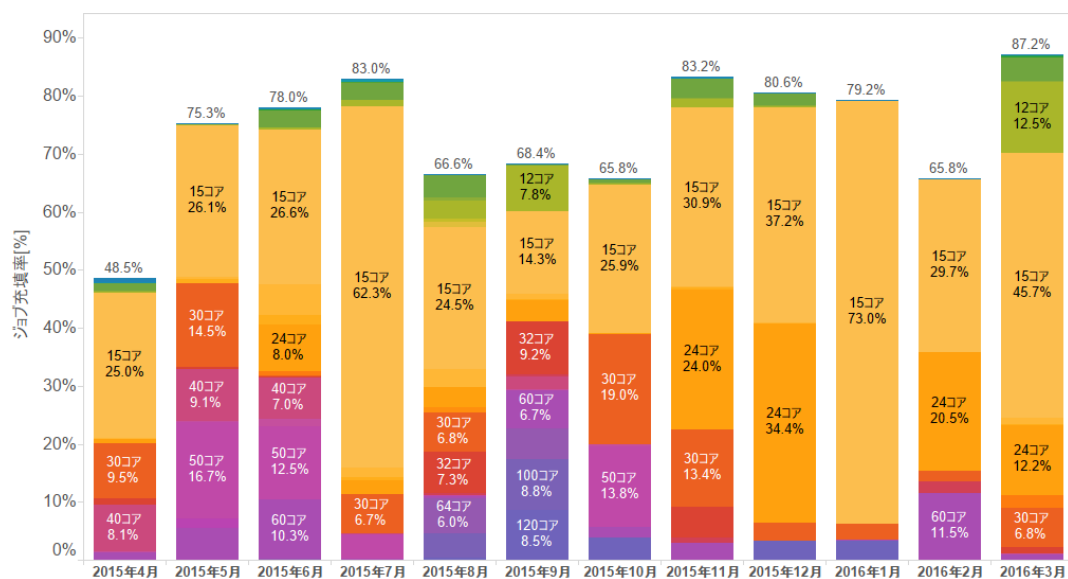


図 14 GW-ACSL 稼働率（ジョブ充填率）

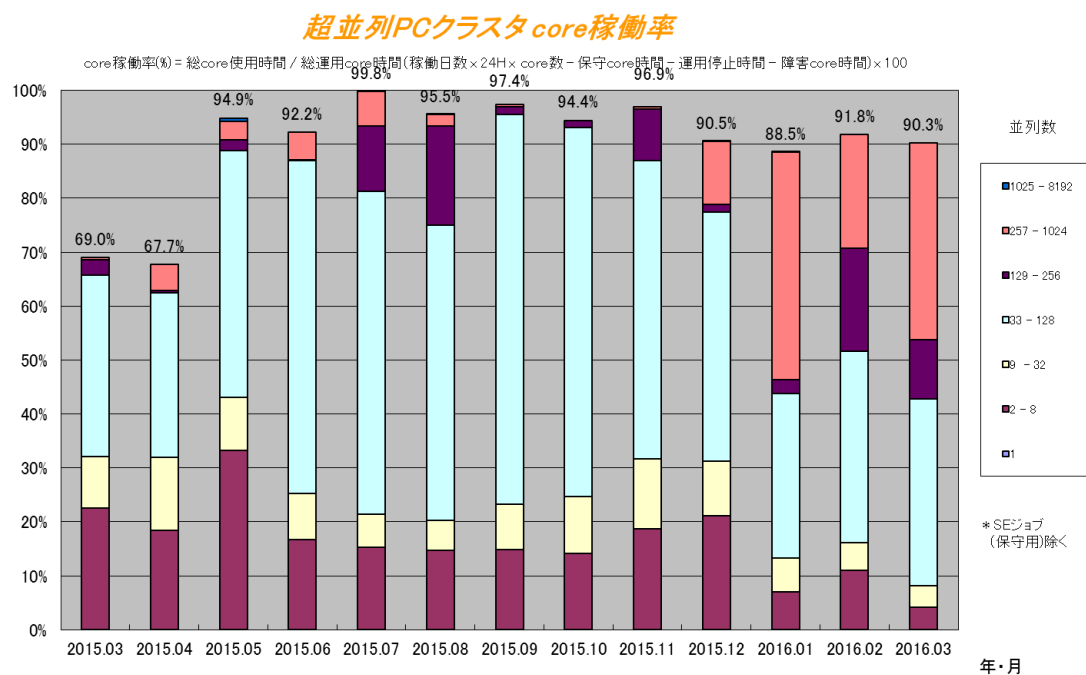


図 15 RICC-MPC 稼働率（ジョブ充填率）

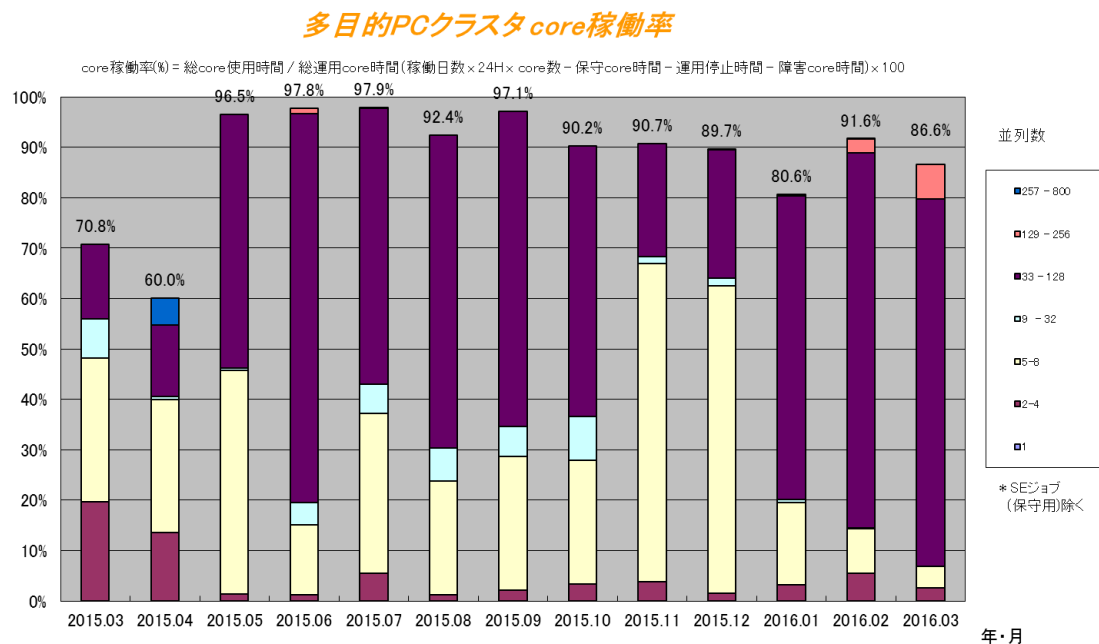


図 16 RICC-UPC 稼働率 (ジョブ充填率)

2.5 HOKUSAI データ預かりサービス

HOKUSAI データ預かりサービスは、スーパーコンピュータ・システム HOKUSAI を構成する。テープ・ライブラリ装置を利用してスーパーコンピュータ利用者以外の理研の方にも大容量データのバックアップを提供するサービスである。

RICC では、同等のサービスとして D2S (Data Depository Service) を提供していたものを、新システムでは HOKUSAI データ預かりサービスとして利用要件とインターフェースを変更して 2015 年 12 月にサービスを開始した。旧 D2S の利用者でデータ移行を希望する利用者のデータは新サービスへと移行した。

新サービスにおける具体的な変更点は、容量の管理を個人単位から研究室単位へと変更したこと、データ通信には高速データ転送が可能で再送機能を有し、堅牢な通信プロトコルを利用する専用ツールだけへと変更したことである。新サービスでは、最大容量が研究室あたり 52TB と拡大され、さらにファイルの送受信や削除等の履歴を保存することができるようになった。またデータ通信に特化した高速で堅牢なプロトコルを使用することで、大容量データを短時間で転送することが実現でき、通信が途中で止まった際にも途中から再送信を行えるようになった。

図 17 に利用者数の累積，図 18 にデータ量とファイル数の履歴を示す。2015 年 12 月に旧 D2S から移行した利用者 14 名でスタートし，2016 年 3 月の時点では利用者数は 23 名，総データ量は 107TB，総ファイル数は 9 万ファイルである。

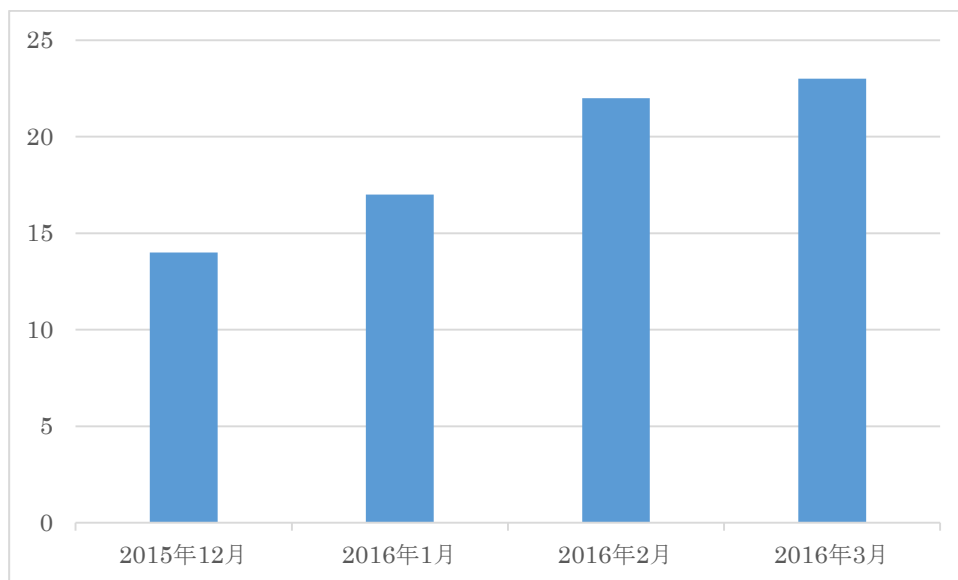


図 17 データ預かりサービス利用者数累計推移

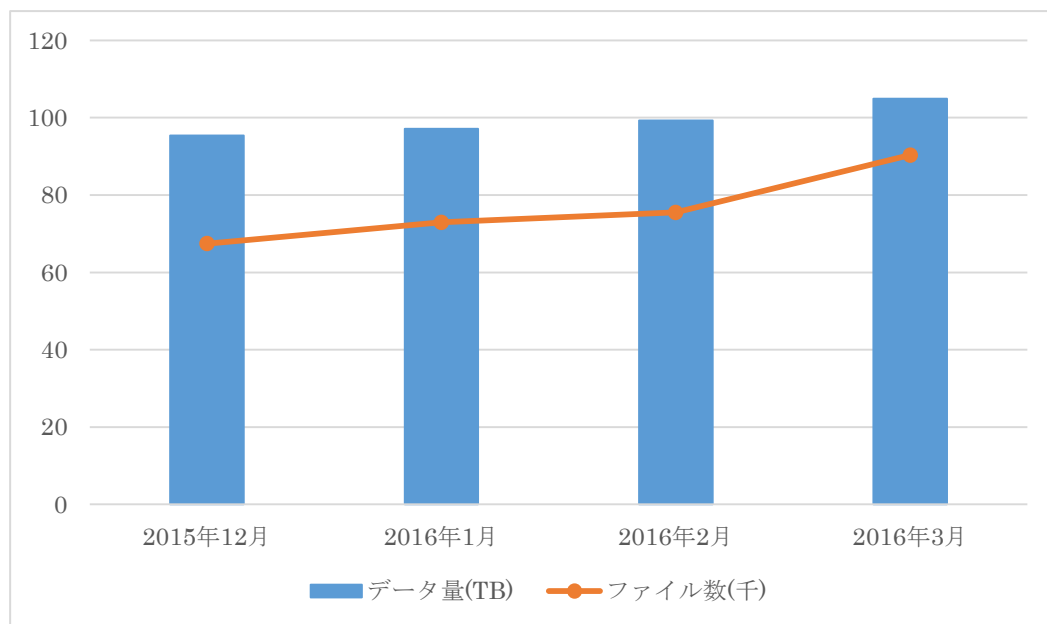


図 18 データ量とファイル数推移

2.6 障害情報

HGW システムのハードウェアの障害件数を図 19 に示すが、GW-MPC が故障件数のほどんどを占めていた。GW-MPC は稼働当初より障害件数が多く、徐々に頻度は下がって来てはいるが、今年度前半は高い発生率で推移した。GW-MPC の障害の内訳を図 20 に示すが、メモリ障害(HMC)とネットワーク関連(Tofu)の障害件数が多くなっているが、年度後半にかけてかなり減少した。しかし、GW-MPC の障害対応は、引き続きハードウェアベンダーと協議し、さらに障害発生を減らしたいと考えている。また、RICC システムのハードウェアの障害件数を図 21 に示す。

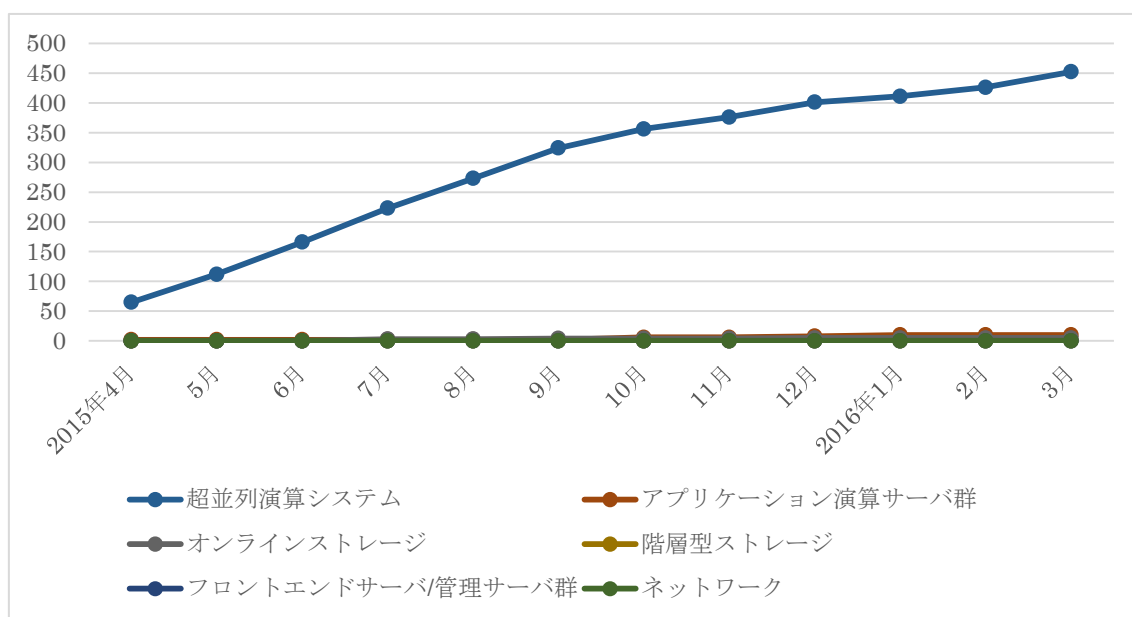


図 19 2015 年度 HGW ハードウェア障害件数累計推移

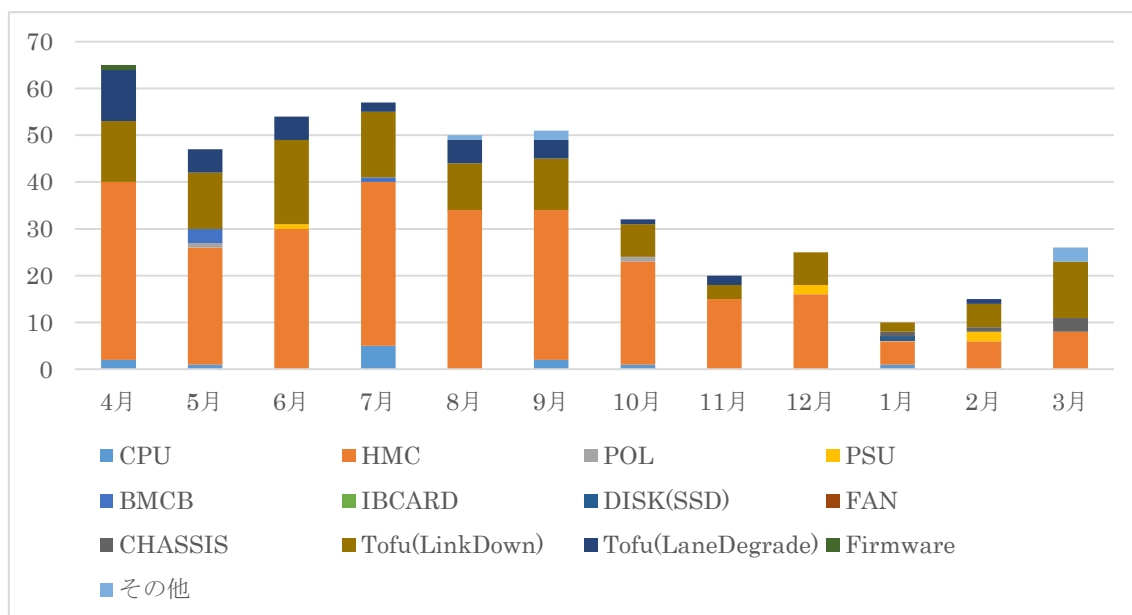


図 20 2015 年度 GW-MPC ハードウェア故障詳細推移

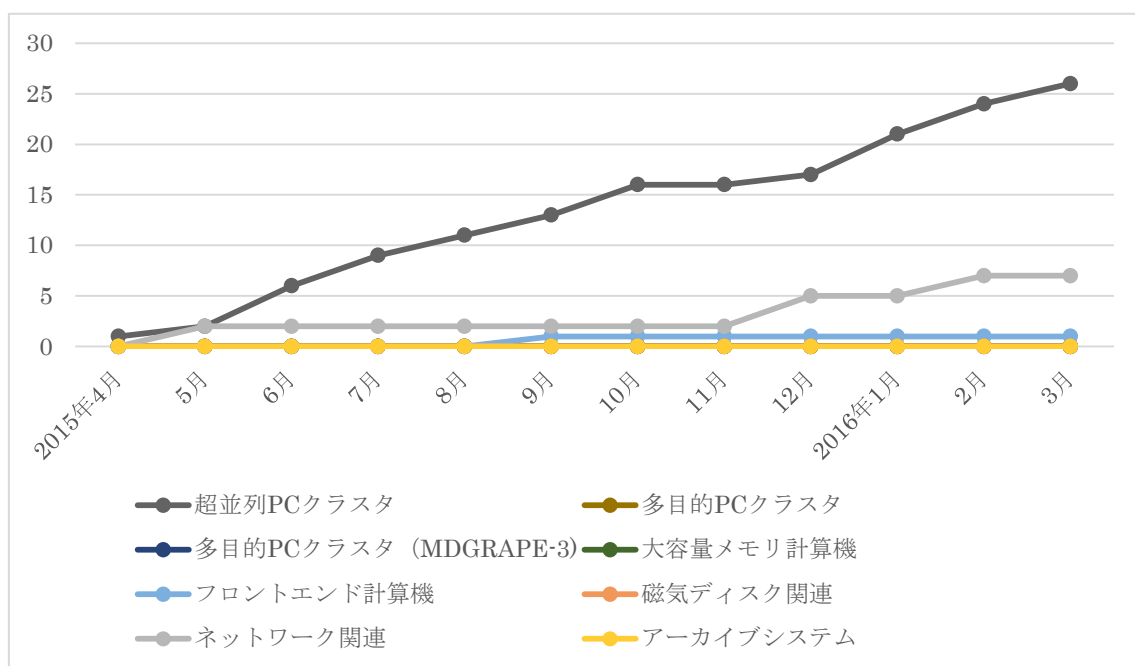


図 21 2015 年度 RICC ハードウェア障害件数累計推移

3 講習会の実施

本年度は従来行っていた少人数で行う座学によるプログラミングなどの基本的な内容の講習会の開催を行わなかった。利用者からの要望もあり FX100(超並列 PC クラスタ)のチューニングに関する講習会のみを実施した。今後も利用者の要望に従って必要な講習会や情報提供を行う方針である。

表 2 講習会の実施状況

講習会名	講座内容	言語	開催月	参加数
いますぐ役立つ FX100 チューニング	RICC、京コンピュータ、FX10 等のスーパーコンピュータの利用経験があり、HOKUSAI で初めて FX100 を利用する方を対象に、FX100 におけるチューニングのポイント、および RICC、京コンピュータ、FX10 から FX100 へプログラム移行する際の注意点について説明する。一般的なスカラ計算機でのプログラム高速化手法の基本知識があることを前提とした内容。	日	6 月 29 日	16

4 アウトリーチ・成果公開

4.1 見学

本年度のスーパーコンピュータ視察・見学は28回（954名）であった。今年度は昨年度に引き続き一般公募による見学会を所内各研究部門と連携し開始し、情報基盤センターでも4度公募による見学者を受け入れたが、そのうち1回は高校生を対象とし見学会を実施した。見学会の告知方法などに工夫し次年度以降も実施したいと考えている。

表 3 視察・見学者数

日付	訪問者	人数
4月18日	一般公開日	280
5月12日	多賀城市立多賀城中学校	6
5月28日	埼玉県立羽生第一高等学校	20
6月4日	高浜市立南中学校	4
6月9日	トプコン US	4
6月18日	山梨県立吉田高校	45
7月7日	東北経済連合会	40
8月3日	上海交通大学	30
8月4日	子ども大学和光(和光市主催)	30
8月6日	秋田県大仙市内中学生首都圏研修	20
8月10日	公募:高校生見学ツアー	30
8月26日	トプコン USA	20
9月4日	近畿大学	20
9月9日	群馬県立富岡高等学校見学会	20
9月11日	公募:見学ツアー	16
9月17日	宝仙学園中学校	22
10月13日	石川県立七尾高校	40
10月19日	茨城県立日立第一高校	30
10月20日	大使館科学技術担当者見学会	40
11月5日	蒲田火災予防協会	35
11月5日	埼玉県副市長会	40
11月10日	裁判官研修	2
12月1日	栃木県立宇都宮高校	30
12月15日	海城中学高等学校	30

1 月 8 日	公募:見学ツアー	20
1 月 22 日	機械学会関東支部公募	30
3 月 10 日	海城中学高等学校	30
3 月 11 日	公募:見学ツアー	20
	合計	954

4.2 学会での展示

2015 年 11 月 16 日から 20 日まで米国テキサス州オースチンで開催されたハイパフォーマンスコンピューティングに関する国際会議である SC15 (The International Conference for High Performance Computing 2015) 展示会場にて計算科学研究機構および GRAPE プロジェクトと共同で理研ブースを設け、新システム HOKUSAI GreatWave の説明および流体シミュレーションに関するデモを行い各国から集まった多くの研究者、技術者と情報交換を行うことができた。

4.3 利用者向け Web サイトの刷新

スーパーコンピュータ利用者および利用予定者向けに情報公開を行っている Web サーバの刷新を行いコンテンツも作りなおしを行った。大きな変更点を以下に示す。

- ・ 情報が古く今後更新する見込みがなく一般的に有用と思われるコンテンツを削除した。
- ・ メニューに表示する項目を精査し大幅に削減した。
- ・ 外部の検索エンジンをトップページから利用できるようにし、目的の情報を探しやすくした。

今後も効率よく必要な情報が得られる Web サイトになるよう心がけたい。

5 問い合わせ対応件数

4 月から新システムの運用が始まり 2 ヶ月のトライアルを経て 6 月から正式な運用が始まったため 4, 5, 6 月は多くの質問、相談が寄せられた。2 月は利用報告書や自年度の課題募集が始まり受付システムからの年度末報告書の受領確認メールで受け取ったメールの数が増加しているが多くはオンライン受付システムからの通知メールであったため年度末に過度にメールに依存した業務を行う必要は無く受付システムの導入効果が得られている。

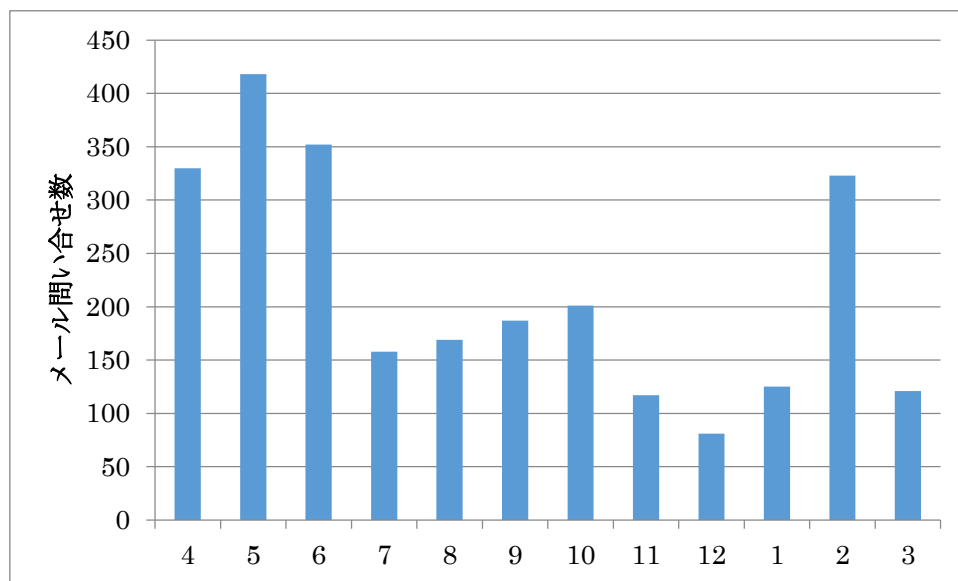


図 22 月別のスーパーコンピュータに関するメール問い合わせ数