全国共同利用大規模並列計算システム調達の背景

伊達 進[1] 木越信一郎[2]

大阪大学サイバーメディアセンター[1]　大阪大学情報推進部情報基盤課[2]

{date, kigoshi}@cmc.osaka-u.ac.jp

概要：大阪大学サイバーメディアセンターでは、2017 年 12 月より全国共同利用大規模並列計算システム(通称：OCTOPUS)による計算サービスを開始する。本システムの調達は、大阪大学サイバーメディアセンターの教職員が中心となり、2015 年後半期より検討を開始し、当該センターの利用状況、利用者へのアンケート、調達予算額を総合的に鑑み、当該システムの仕様策定がなされた。本稿では、大阪大学に新規に導入される全国共同利用並列計算システムの調達背景についてまとめる。

1　まえがき

現在、大阪大学サイバーメディアセンターでは、NEC 製ベクトル型スーパーコンピュータ SX-ACE、スカラ型スーパーコンピュータ大規模可視化対応 PC クラスタ(通称: VCC, 31.1TFlops)および汎用コンクラスタシステム(通称: HCC)による計算機サービスを提供している。しかし、SX-ACE を主に利用するユーザは、その性能、使いやすさから、高い評価がある一方、スカラ型スーパーコンピュータを利用するユーザあるいは利用したいと考えるユーザからは、性能、ノード数の少なさ（実行ジョブの並列度数)、待ち時間の長さ、仮想化資源への反発等、現状システムへの不満が寄せられている現状がある。

本稿で取り扱う全国共同利用大規模並列計算システム(通称：OCTOPUS)[1]の調達は、本センターの有する HCC の後継システムをターゲットとして行われた。本稿では、本センターのスカラ型スーパーコンピュータによる問題点、および、後継システムへの要望・期待等を始めとする調達背景とともに、どのようなシステムが導入されたのかについてを記す。これにより、本センターの計算サービスを近い将来担当することとなる教職員、あるいは、他計算機センターの教職員が計算機システムの調達を行う際の一助となれば幸いである。

本稿の構成は以下の通りである。2 節では、本センターの有するスカラ型スーパーコンピュータの問題点について説明する。3 節では、本センターが実施した利用者へのアンケートから、後継システムへの要望・期待をまとめる。4 節では、本稿執筆時点で構築を進めている全国共同利用並列計算システムの概要について記し、5 節で本稿をまとめる。

2　サイバーメディアセンターのスカラ型スーパーコンピュータの問題点

2.1　スカラ型スーパーコンピュータ

大阪大学サイバーメディアセンターでは、1 節で述べたように、スカラ型スーパーコンピュータとして、VCC および HCC の計算サービスが提供されている。前者は、2013 年の補正予算「高性能汎用計算機高度利用事業「京」を中核とする HPCI の産業利用支援・裾野拡大のための設備拡充」の一環として、大阪大学サイバーメディアセンターが実施した調達「HPCI と連動するネットワーク共有型可視化システム」の一環として導入された計算クラスタシステムである。後者は、2012 年に実施された調達「汎用コンピュータシステム」で導入されたクラスタシステムである。

VCC は、本稿執筆時点では、Intel Xeon E5-2670v2 プロセッサ 2 基、64GB の主記憶が搭載された計算ノードが 66 ノード、Intel Xeon E5-2690v4 プロセッサ 2 基、64GB の主記憶が搭載された計算ノードが 3 ノードを InfiniBand FDR によるインターコネクトで接続したクラスタシステムとなっている。また、当該システムは、59 基の NVIDIA 製 GPU Tesla

K20 がリソースプール化されている構成となっており、システムハードウェア仮想化技術 ExpEther により、任意の枚数の GPU を計算ノードに比較的高い自由度をもって割り付けることのできる再構成可能システムとしての特徴を有している。そのため、例えば、1 計算ノードに対し GPU4 枚、1 計算ノードに対し GPU2 枚といった構成を、利用者の計算要求に応じた設定が可能である。

一方、HCC は、2012 年 10 月より稼働する NEC 製 Express5800/53h 575 台からなるクラスタシステムである。本システムのそれぞれの計算ノードは、Intel Xeon E3-1225v2(1CPU4 コア)、主記憶として 8GB あるいは 16GB のメモリを有するものの、本システムの計算サービスは仮想計算機としてのサービスとなっている。

2.2 問題点

上述したように、VCC は 2014 年に導入された比較的新しい計算クラスタシステムである。本システムを実際に利用している利用者の多くは、(老朽化の影響はあるものの)本システムの性能に不満があるわけではない一方、(1)高い並列度数でジョブを実行しようとする際の待ち時間の長さ、(2)本システムで投入できる並列度数のジョブの大きさ、に不満を持っているユーザが数多く存在している。これらは、本システムを構成する計算ノード数が 69 と小さいために、本センターのシステム管理上、ノード数の少ない本システムが並列度数の高いジョブをもつユーザによって占有されることを防止する視点から、ジョブスケジューラのキュー構成の最大並列度数を小さくしていること、本センターの採用するジョブスケジューラの特性上並列度の小さいジョブが並列度の高いジョブを先行してジョブが実施される傾向にあることに起因する問題である。また、本システムは、上述したように、該当年度でのみ有効である補正予算での導入であるため、通常、本センターが利用者に計算機利用負担金として課す電気代相当分の費用に加え、導入業者による保守運用費用が利用負担金として課されている。そのため、利用者からは(3)利用負担金が高い、という問題も発生している。

次に、HCC についての問題点をまとめる。HCC の問題点は、上述した調達「汎用コンピュータシステム」は、本センターが全学支援業務として行う、教育、CALL、高性能計算、図書館のすべての要望を取り込んだシステムとして導入されていたことに起因する。調達当時は、仮想化技術への期待も高かったことも関係しているのであろうと思われるが、「汎用コンピュータシステム」の高性能計算に係る機能は、計算機 Express5800/53h 上でオペレーティングシステムとして Windows を稼働させ、仮想化技術 Parallels を導入した上で、当該計算機のもつ 4 コアのうち 2 コアを使う Cent OS 6.1 を稼働させ、それらをクラスタ化することで実現されている。このような構成のため、利用者からは、(1) 高性能計算用途に仮想計算機を利用していること、(2)システムの不安定な挙動(仮想化技術 Parallels と InfiniBand へのパススルー技術の不安定な挙動)などの不満・問題が数多く生じていた。

上述のように、大阪大学サイバーメディアセンターのスカラ型スーパーコンピュータに対する問題点や不満から、HCC の後継システムである全国共同並列計算システムの調達は、スカラ型計算サービスに期待する利用者にとっては必要不可欠かつ重要な調達案件となっていた。

3 新システム導入に向けたアンケート調査

3.1 調査内容

前節で述べたように、本センターの現状有するスカラ型スーパーコンピュータに対する不満や問題点から、HCC の後継システムの導入は非常に重要な調達である。そのため、本センターでは、利用者を対象とし、新システムに対する要望調査アンケートを実施した。当該アンケートでは、より多くの回答数を得られるよう、できるだけ簡易的な下記の質問項目に限定し、本センターのウェブ

ページで実施した。

(1) 下記で最も/2 番目/3 番目に/優先する利用基準を教えてください

A: 並列利用できる最大ノード数

B: ノードあたりのコア数

C: ノード(コア)あたりのメモリ

D: 利用できるアクセラレータ

(2) どのようなアクセラレータを利用したいですか？

XeonPhi

GPU

FPGA

その他

(3) どのようなアプリケーションを利用しますか？

希望しません

Adams

AMBER

ANSYS

Dytran

Gaussian

GROMACS

LAMMPS

Marc

Mentat

Nastran

OpenFOAM

VASP

その他

(4) その他　次期システムに希望すること(希望理由を必ずご記入ください)

3.2 アンケート結果

アンケートの結果、(1)の設問に際しては、図 1 に示されるような結果となり、次期システムに対しては、アンケート回答者のうち約 40%が並列利用できる最大ノード数を最も希望し、約 50%がノードあたりのコア数を 2 番目に優先する結果となった。

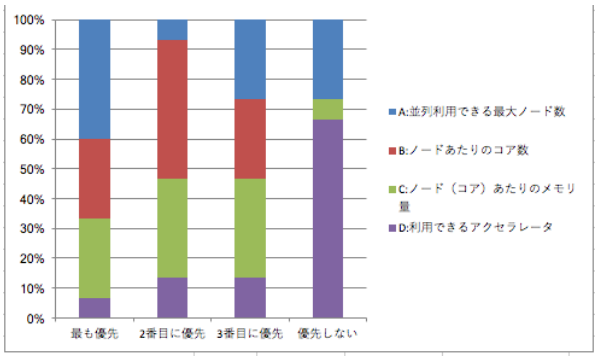

図 1: 計算ノードに対する優先基準

また、(2)の設問に際しては、FPGA については「希望しない」が大多数を占めたが、XeonPhi および GPU については、「希望する」、「希望しない」がおおよそ半数ずつ程度の結果となった。

(3)の設問に際しては、「希望しません」が全体の 30%となり、他のアプリケーションについての希望は分散の大きいものとなった。

(4)の設問に際しては、自由記述であったが、数多くの回答が得られた。以下にいくつかを抜粋する(一部、個人、組織名を特定できる箇所は修正をいれています。)。

- 計算待ち時間が長く混雑しているように思われますので、ノード数を増やして頂ければと思います。
- 京につながるような、中規模(とはいえ並列度数千)の並列計算が可能なシステムを望みます。
- ベクトルコンピュータを希望します。
- 大規模メモリ型 x86 サーバを希望いたします。Pascal-GPGPU の搭載も希望いたします。
- NVIDIA の GPU が速くて個人的には好みです。(中略) GPU を採用いただける場合は、CUDA や OpenACC のユーザー講習会をぜひともよろしくお願いいたします。
- 自前の並列コード(■□■□シミュレーション) を使いますので、アプリについては特に希望なしです。通常の MPI、OpenMP 可能なクラスタが最優先で、そして将来的に GPU、FPGA などのアクセラレータの使用も考慮し

ていく予定です。ベクトルマシンは、汎用性がおちるのでやめた方がいいと思います。

- 1 ノードあたりのメモリ容量は、64GByte 以上(SX と同レベル以上)できれば、128GByte とかが望ましいです。最近の■□■□のサイズが極端に大きくなってきているので、それなりのメモリが欲しい所です。それなりのメモリのアクセス速度も重要な要件です。
- 通信速度にも重点を置いてください。(■□大学の話をきいたのですが、通信速度をケチったために、■□■□コードの実行性能が 0.1%にまで落ちて使えなかったそうです。)ピーク性能を高くするために、通信をケチってしまったために性能を出せなくなってしまったという話を他でも聞きます。通信はケチらないでください。

本アンケートの結果から、本センターの全国共同利用大規模並列計算システムでは、ノード数を最優先項目と設定するとともに、コア数、メモリ容量、アクセラレータ、インターコネクトに関する多様なユーザの計算要求を柔軟に収容できるハイブリッド型クラスタシステムの導入を目指した。

4 全国共同利用大規模並列計算システム

本節では、全国共同利用大規模並列計算システムとして導入が決定したシステムの概要を図 2 に記す。本システムの調達では、上述したようにシステムの有する計算ノード数を優先項目と定め、調達を推進した。調達当時、プロセッサのリリース時期・価格等の詳細が明確ではなかったため、本調達に際しては、各提案候補ベンダの担当者とは何度も何度も話をし、プロセッサ、アクセラレータ、相互結合網の価格感を想定し、総合評価方式による加点を調整することで、プロセッサ性能のよいプロセッサを搭載したノード数を多く含むシステム提案が高得点となるよう設定した。その結果、本センターの利用できる予算内で、提案候補ベンダ間の競争も働き、本センターの利用者の望むノード数を多く搭載するとともに、多様な計算ニーズを収容できるハイブリッド型クラスタシステムの搭載が実現できたと考えている。

総演算性能	1.463 PFLOPS	
ノード構成	汎用CPUノード 236ノード (471.24 TFLOPS)	プロセッサ：Intel Xeon Gold 6126 (Skylake / 2.6 GHz 12コア) 2基 主記憶容量：192GB
	GPUノード 37ノード (858.28 TFLOPS)	プロセッサ：Intel Xeon Gold 6126 (Skylake / 2.6 GHz 12コア) 2基 GPU：NVIDIA Tesla P100 (NV-Link) 4基 主記憶容量：192GB
	Xeon Phiノード 44ノード (117.14 TFLOPS)	プロセッサ：Intel Xeon Phi 7210 (Knights Landing / 1.3 GHz 64コア) 1基 主記憶容量：192GB
	大容量主記憶搭載ノード 2ノード (16.38 TFLOPS)	プロセッサ：Intel Xeon Platinum 8153 (Skylake / 2.0 GHz 16コア) 8基 主記憶容量：6TB
ノード間接続	InfiniBand EDR (100 Gbps)	
ストレージ	DDN EXAScaler (Lustre / 3.1 PB)	

図 2: 全国共同利用大規模並列計算システム概要.

5 まとめ

本稿では、本センターが 2017 年 12 月より計算サービスを開始する全国共同利用大規模並列計算システム調達の背景についてまとめた。本システムの調達に際しては、システムコードネームを octopus として、本センターの教職員が団結し、利用者のニーズを満たすことのできるシステムを目指した。その結果、汎用 CPU ノード群(236 ノード)、GPU ノード群(37 ノード)、XeonPhi ノード(44 ノード)、大容量主記憶搭載ノード 2 ノード、大容量ストレージ 3.1PB から構成される 1.46PFlops のハイブリッド型クラスタシステムの導入が確定した。本稿執筆時点では、いまだ構築作業中であるが、今後、多様なユーザの計算ニーズを収容する目的を達成できる OCTOPUS (Osaka university Cybermedia center Over-Petascale Universal Supercomputer)へと完成させていきたいと考えている。

参考文献

[1] OCTOPUS, http://www.hpc.cmc.osaka-u.ac.jp/octopus/.