新スーパーコンピュータ「ITO システム」の紹介

上田　将嗣 [1)], 小野　真 [1)], 平島　智将 [1)], 原田　浩睦 [1)], 南里　豪志 [2)]

1) 九州大学 情報システム部
2) 九州大学 情報基盤研究開発センター

ueda@cc.kyushu-u.ac.jp

Introduction of New Supercomputer System ITO

Masatsugu UEDA [3)], Makoto ONO [3)], Tomoyuki HIRASHIMA [3)], Hiroyoshi Harada [3)], Takeshi NANRI [4)]

3) Information System Department, Kyushu University.
4) Research Institute for Information Technology, Kyushu University.

概要

九州大学情報基盤研究開発センター（以下、「センター」という。）では、2017 年 10 月より新スーパーコンピュータシステム ITO の運用を開始する。本稿では、ITO の概要について紹介する。

1　はじめに

センターでは、新スーパーコンピュータシステム ITO の運用を 2017 年 10 月から開始する[1]。

ITO は九州大学伊都キャンパスに導入される最初のスーパーコンピュータであり、第 5 期科学技術基本計画に示された AI（人工知能・機械学習）・ビッグデータ、さらにデータサイエンスの研究およびこれらを活用した研究に対応した研究基盤の提供を目指して仕様策定したもので、日本国内に設置されるスーパーコンピュータシステムとしては初めて、詳細な電力モニタリング機構や本格的なクラウド連携の仕組みを導入し、従来にはない新しいスーパーコンピューティングの方向性や利用者層・課題の拡大に向けたインフラの提供を目指している。

2　ITO の狙い

2.1 対話的利用環境の強化

近年、計算規模の拡大に伴い、計算結果の入出力に係るデータが飛躍的に増加しており、手元の端末でプリポスト処理を行うための、データ転送に要する時間が大きな問題となっている。

そこで、平成 24 年度に導入した高性能演算サーバシステムでは、大規模メモリを有した可視化サーバ 5 ノードを導入し、可視化サーバを時間単位で予約して利用できる可視化予約システムを導入した[2]。

ITO では、これをさらに発展させ、フロントエンドを大幅に増強し仮想環境も取り入れることで、プリポスト処理の強化を図っている。

2.2 データサイエンス分野の利用支援

センターでは、従来の科学技術シミュレーション向けの利用に加えて、データサイエンス研究の支援を重大なミッションと考えている。ITO では、最新 GPU（NVIDIA Tesla P100）と、TensorFlow をはじめとする関連ソフトウェア群を整備している。また、先に述べたフロントエンドもデータサイエンスに活用できると考えている。

2.3 パブリッククラウド連携

これまで、特に論文の締切前に短期間で計算結果が必要な場合等、突発的に発生する緊急性の高い計算需要への対応が困難になる場合があった。また、近年みられる電力の需給逼迫及び電気代の高騰により、計算機の縮退運転の頻度が今後増加することが予想される。これらの、不足する計算資源の補填は、計算機センターの重要な課題であり、ITO では、その対策としてパブリッククラウド連携機能を組み込んでいる。

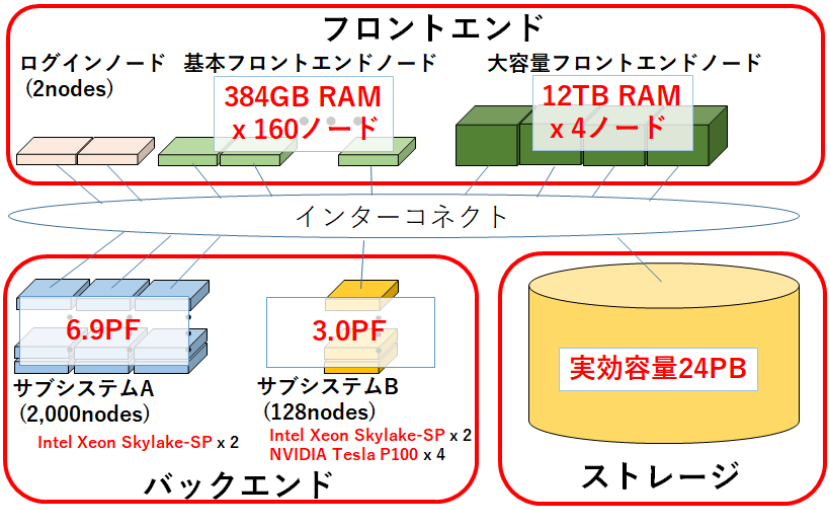

図 1　ITO システム構成図

2.4 省電力コンピューティング

スーパーコンピュータの消費電力は増加の一途を辿っており、今後、省電力技術が、計算機センターの運用において重要になる。ITO では、将来の省電力運用に向けた基礎技術機能を導入している。

3　システム構成

3.1 システム概要

ITO はバックエンド（サブシステム A、サブシステム B）、フロントエンド（ログインノード、基本フロントエンド、大容量フロントエンド）、ストレージから構成されている。システム構成図を図 1 に、旧システムと ITO の比較を表 1 に示す。なお、表中の総理論演算性能は浮動小数点

ITO は 2 段階での導入を予定しており、2017 年 10 月に運用を開始したのはサブシステム B、フロントエンドおよびストレージである。一方、サブシステム A は、2018 年 1 月から運用開始よ予定している。

表 1　旧システムと ITO の比較

	旧システム（合計）	ITO
総理論演算性能	1.958PF CPU 1.290PF Acc 0.668PF	**10.43PF** **CPU 7.72PF** **Acc 2.71PF**
ディスク容量	8.28PB	**24.6PB**
フロントエンド	11 台	**166 台**
アクセラレータ	1.31TF×32 1.17TF×354 1.0TF×210	**5.3TF×512**
ノード当たり主記憶容量	32GB～16TB	**192GB～12TB**

3.2 サブシステム A

サブシステム A は、Intel 社の最新の Xeon プロセッサ（Skylake-SP）を 2 基搭載したノード、2,000 ノードからなるシステムである。各ノードは、倍精度浮動小数点演算で 3.456TFLOPS の理論演算性能、192GiB の主記憶容量を有しており、サブシステム A 全体では、総理論演算性能 6.912PFLOPS、総主記憶容量 384TiB である。

なお、サブシステム A のうち、256 ノードは、ローカルの作業領域として、0.8TB の SSD を内蔵している。

3.3 サブシステム B

サブシステム B は、Intel 社の最新の Xeon プロセッサ（Skylake-SP）を 2 基と、NVIDIA 社の最新 GPU（Tesla P100）を 4 基搭載したノード、128 ノードからなるシステムである。各ノードは、倍精度浮動小数点演算で CPU：2.64TFLOPS、GPU：21.2TFLOPS の理論演算性能、384GiB の主記憶容量を有しており、サブシステム B 全体では、総理論演算性能 3.052PFLOPS、総主記憶容量 49TiB である。

なお、サブシステム B の全ノードは、ローカルの作業領域として、0.8TB の SSD を内蔵している。

3.4 基本フロントエンドノード

基本フロントエンドは、Intel 社の最新の Xeon プロセッサ（Skylake-SP）を 2 基と、NVIDIA 社の GPU（Quadro M4000）を 1 基搭載したノード、160 ノードからなるシステムである。各ノードは、倍精度浮動小数点演算で 2.64TFLOPS の理論演算性

能、384GiB の主記憶容量を有しており、基本フロントエンド全体では、総理論演算性能 422.4TFLOPS、総主記憶容量は 61TiB である。

3.5 大容量フロントエンドノード

大容量フロントエンドは、Intel 社の Xeon プロセッサ（Broadwell-EP）を 16 基と、NVIDIA 社の GPU（Quadro P4000）を 1 基搭載したノード、4 ノードからなるシステムである。各ノードは、倍精度浮動小数点演算で 12.39TFLOPS の理論演算性能、12TiB の主記憶容量を有しており、大容量フロントエンド全体では、総理論演算性能 49.56TFLOPS、総主記憶容量は 48TiB である。

3.6 ストレージ

ストレージのファイルシステムは FEFS で、実効容量が 24.6PB である。ストレージへのデータ転送速度は、サブシステム A から合計 100GB/秒以上、サブシステム B から合計 30GB/秒以上、フロントエンドから合計 30GB/秒以上である。

3.7 インターコネクト

サブシステムやフロントエンドの各ノード、及びストレージを相互接続しているインターコネクトは片方向理論転送性能がポート当たり 12.5GB/s の Mellanox 社 InfiniBand EDR で構成している。特にサブシステム A 及び B では、それぞれの内部のノード間接続を Full Bisection Bandwidth の Fat Tree トポロジーとすることにより、通信同士の衝突の影響を受けにくくなっている。

さらにこのインターコネクトには、並列計算で多用される集団通信について、制御を CPU ではなくネットワークカードやスイッチで行わせる機能があるため、CPU を計算に専念させることができ、特に大規模並列計算で計算効率の向上が図れる。

各システムのノード当たりのポート数は、以下の通りである。

- サブシステム A　1 ポート
- サブシステム B　2 ポート
- 基本フロントエンド　2 ポート
- 大容量フロントエンド　4 ポート

3.8 ソフトウェア

ITO で提供するソフトウェアを表 2 に示す。これまでと比べて、「機械学習、データ解析」関連ソフトウェアを新たに導入した。今後も利用者からの要望に応じて、ソフトウェアの拡充する予定である。

表 2　ソフトウェア一覧

分野	ソフトウェア
コンパイラ	富士通コンパイラ、 インテルコンパイラ、 PGI コンパイラ、CUDA
数値計算ライブラリ	SSL Ⅱ、C-SSL Ⅱ、LAPACK/BLAS ScaLAPACK、NAG ライブラリ、 FFTW、PETSc
その他ライブラリ	HDF5、NetCDF
計算化学	Gaussian、GaussView、CHARMM、 VASP、Molpro、SCIGRESS、AMBER、 GAMESS、GROMACS
流体・構造解析	Marc、Marc Mentat、 MSC Nastran、MSC Patran、 ANSYS CFX、OpenFOAM
機械学習 データ解析	TensorFlow、SAS、ENVI/IDL、R
科学技術計算	Mathematica、Matlab
画像処理	FieldView、AVS
その他	Exceed onDemand

4　ITO で新たに導入する運用制度及び機能

4.1 ノード固定利用

旧システムでは、利用者に対して契約したノード数を常に確保する「占有タイプ」を提供しており、実行待ち時間を考慮する必要がないことから好評を得てきた。一方でこのタイプは、年間を通じて必ずしも利用率が高いわけではなく、稼働率の面で課題があった。

ITO では、「占有タイプ」に代わって新たに「ノード固定タイプ」を提供することとした。これは、利用者に対して契約したノード数を固定的に割り当て、それらのノードにおいて、どのようにシステムが混雑していても必ず 1 時間以内にジョブを実行できる権利を提供する制度である。これにより、ノード固定タイプで提供したノード上で、他の利用者の 1 時間以内のジョブを実行することが可能となる。そこで、実行時間を 1 時間以内に制限した「デバッグ用ジョブクラス」を用意し、このジョブクラスのジョブの優先度をノード固定タイプのジョブよりも低く設定することで、ノード固定タイプの利用者の権利を保証しつつ、システムの稼働率向上を図る。

4.2 フロントエンド予約システム

フロントエンドには、基本フロントエンドノード 160 台、大容量フロントエンド 4 台が用意されている。これらのノードは、富士通株式会社のフロントエンド予約システム「UNCAI」により、指定した時刻に、指定した資源量で対話的に利用できる。また、これらのノードの利用形態は、物理ノード単位で利用する「ベアメタル」、もしくはノード内に複数起動した仮想システム単位で利用する「仮想マシン」を選択することが出来る。

利用者は、使用するコア数と利用開始及び利用終了時刻を指定してベアメタル及び仮想マシンの予約を行うことができる。「UNCAI」のリソース予約画面を図 2 に示す。

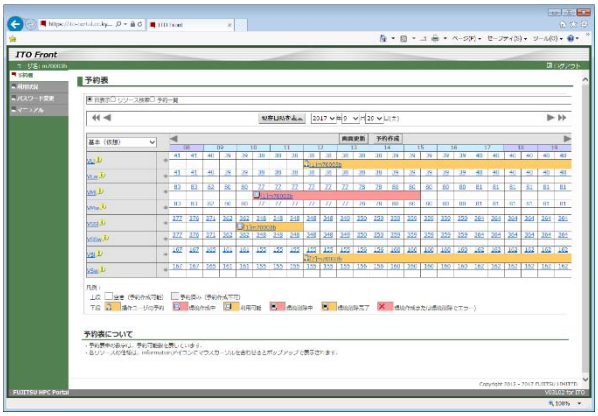

図 2　「UNCAI」リソース予約画面

4.3 パブリッククラウド連携

ITO は、ストレージ上の利用者プログラムやデータ等のファイルを、パブリッククラウド環境にアップロードして実行し、結果をストレージにダウンロードしたうえで、クラウド上のインスタンスを自動的に削除するインターフェースを有している。

また、基本フロントエンドノード、大容量フロントエンドノード、およびパブリッククラウドに、共通のユーザインタフェースでジョブを投入可能な機能も有している。

これらのインタフェースの運用には、特にパブリッククラウド側の利用料金やアカウントの管理について、従来と異なる制度設計が求められる。そこで、当面は、利用希望者と協力して試験運用を進め、将来のパブリッククラウドを活用した運用に繋げる予定である。

4.4 電力制御・モニタリング

ITO では、将来の省電力運用に向けた基礎技術として以下の機能を有している。

- ジョブ単位で消費電力を収集、蓄積する機能
- CPU コアの周波数変更等による電力キャッピング機能
- 使用電力を考慮したジョブ割り当て機能
- 使用状況に応じた演算ノードの自動 On/Off 機能

また、上記の機能を支援するために、PDU によりラック毎に 1 分単位で消費電力を測定・蓄積できる仕組みも導入している。電力モニタリングシステムの画面を図 3 に示す。

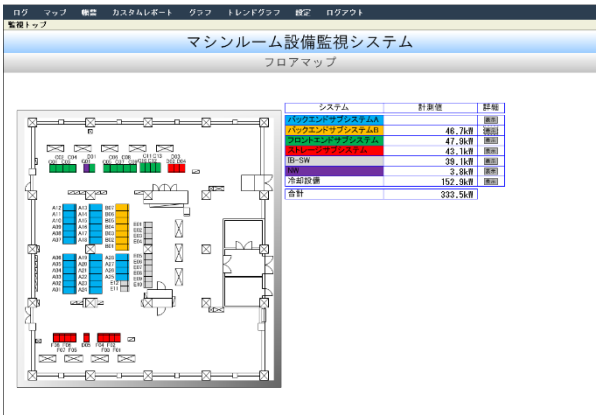

図 3　電力モニタリングシステム

なお、これらの省電力機能を運用するうえで、従来の計算機台数と理論演算性能に基づいた課金体系では、対応が難しいと考えている。そのため、学内外の研究者や、ITO の開発ベンダーである富士通株式会社と共同で省電力運用に関する研究に取り組む予定である。

5　おわりに

本稿では、今年度より運用を開始した新システム ITO の概要を紹介した。執筆段階では、システム構築中のため、運用実績等を報告することはできていないが、旧システムの 5 倍以上となる理論演算性能や、大幅に増強したフロントエンドによる対話型環境の強化等、利用者の研究の質の向上及び利用者の裾野拡大に貢献できるシステムを導入できたと考えている。

また、パブリッククラウド連携や電力制御・モニタリング等の新しい取り組みに関しては、試験運

用や共同研究を通して知見を深めていき、今後のITO 運用、ITO の次のスーパーコンピュータ仕様策定、さらに将来のスーパーコンピュータセンター運用に役立てたいと考えている。

参考文献

[1] 新スーパーコンピュータシステム “ITO” https://www.cc.kyushu-u.ac.jp/scp/new-system.html

[2] 平島智将 原田浩睦 小野真 上田将嗣 南里豪志、可視化サーバ予約システムの導入と運用、大学 ICT 推進協議会 2015 年度年次大会