新スーパーコンピュータシステムの紹介

島袋 友里 [1)], 當山 達也 [1)], 疋田 淳一 [1)]

1) 京都大学 情報部

shimabukuro.yuri.8e@kyoto-u.ac.jp

Introduction of New Supercomputer System

Yuri Shimabukuro[1)], Tatsuya Tohyama[1)], Junichi Hikita[1)]

1) Information Management Department, Kyoto Univ.

概要

京都大学学術情報メディアセンターでは、全国の学術研究者に対してスーパーコンピュータシステムを提供しており、2023 年 4 月より新スーパーコンピュータシステムの正式サービスを開始した。本稿では、システムの構成およびサービス内容について紹介し、新システムの性能について報告する。

1 はじめに

京都大学学術情報メディアセンター（以下、「本センター」と言う）では、全国共同利用サービスとして、学術研究者に対してスーパーコンピュータシステムを提供している。高性能な演算環境を提供するために 4〜6 年間隔でシステムを更新しており、2023 年 4 月より新スーパーコンピュータシステム（以下、「新システム」と言う）の正式サービスを開始している [1]。

新システムは、Camphor3（System A）、Laurel3（System B）、Cinnamon3（System C）、Gardenia（System G）、クラウドシステムからなる 5 種の演算システムと、大容量ストレージ、高速ストレージからなる 2 種のストレージシステムによって構成された、総演算性能 11 PFLOPS、総メモリ容量 370 TiB、総ストレージ容量 44 PB の高性能かつ大規模なシステムである。

本稿では、新システムの構成およびサービス内容について紹介し、新システムの性能について報告する。

2 システム構成

2.1 システム概要

新システムの構成図を図 1 に示す。演算システムの内、Camphor3、Laurel3、Cinnamon3 については前システムの設計理念を継承しており、Camphor3 は演算性能を、Laurel3 は汎用性を、Cinnamon3 はメモリ容量を重視した構成となっている。Gardenia は今回追加したシステムであり、近年著しく発展・拡大している機械学習・深層学習用途に対応するために、GPU を搭載した構成である。クラウドシステムは、クラウド事業者が提供するリソースを活用することで、柔軟に運用可能なシステムとして位置付けている。ストレージシステムは、HDD で構成した大容量ストレージと、SSD で構成した高速ストレージの 2 種である。

2.2 Camphor3（System A）

Camphor3 は、第 4 世代の Intel Xeon CPU Max 9480（Sapphire Rapids）プロセッサ 2 基を搭載した演算ノードを 1120 ノード接続した大規模クラスタ構成のシステムである。各ノードには 112 個の CPU コアを搭載し、6.80 TFLOPS の演算性能を有している。メモリは高いメモリバンド幅を有する HBM2e を 128 GiB 搭載しており、理論上のメモリバンド幅は 3.2 TB/sec である。ノード間は 50 GB/sec の転送性能を持つ Infiniband NDR による Fat-Tree 構成で接続している。総演算性能は 7.63 PFLOPS、総メモリ容量は 140 TiB である。

HBM2e と Infiniband NDR による高速なデータ転送能力により、CPU の演算性能を発揮しやすい大規模な並列演算が可能なシステムとなっている。

2.3 Laurel3（System B）

Laurel3 は、第 4 世代の Intel Xeon Platinum 8480+ プロセッサ（Sapphire Rapids）を 2 基搭載した演算ノードを 370 ノード接続したクラスタ構成のシステムである。各ノードには 112 個の CPU コアを搭載し、7.17 TFLOPS の演算性能を有しており、メモリは最新の規格である DDR5-4800 を 512 GiB 搭載し、

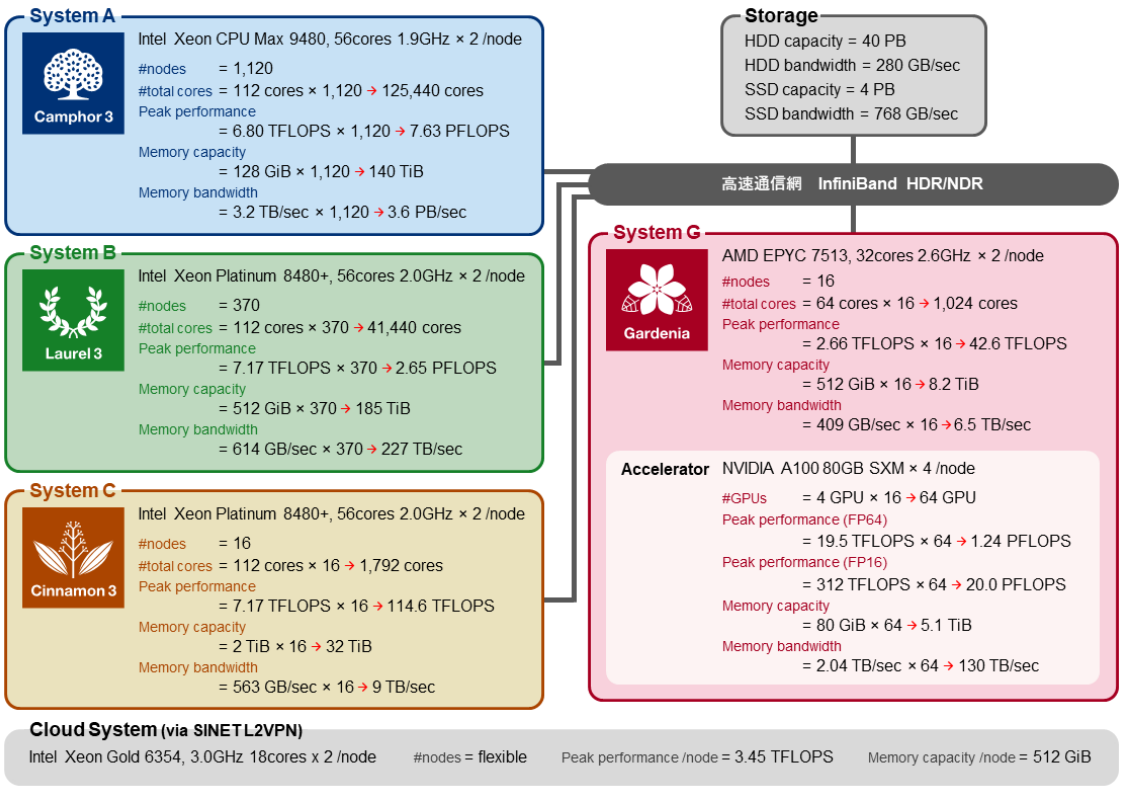

図 1　システム構成図

614 GB/sec のメモリバンド幅を備えている。ノード間は 25 GB/sec の転送性能を持つ Infiniband NDR 200Gbps による Fat-Tree 構成で接続している。総演算性能は 2.65 PFLOPS、総メモリ容量は 185 TiB である。

Camphor3 と比較し、メモリバンド幅、ネットワーク転送性能には劣るが、メモリ容量が大きいことから多様な計算を実行しやすい中規模なシステムとなっている。

2.4　Cinnamon3（System C）

Cinnamon3 は、メモリ以外は Laurel3 と共通のシステムで、1 ノードあたり 563 GB/sec のメモリバンド幅を有する DDR5-4400 を 2 TiB のを搭載した大容量メモリのシステムである。演算ノードは 16 ノードで構成している。総演算性能は 114.6 TFLOPS、総メモリ容量は 32 TiB である。

大規模な可視化やデータ分析を共有メモリ上で実行しやすいように、容量だけでなく、メモリバンド幅も重視したシステムである。

2.5　Gardenia（System G）

Gardenia は、AMD EPYC 7513 プロセッサを 2 基、NVIDIA A100 を 4 基搭載した演算ノードを 16 ノード接続したシステムである。各ノードの CPU 演算性能は 2.66 TFLOPS、GPU 演算性能は倍精度 78 TFLOPS、半精度 1248 TFLOPS である。CPU 総演算性能は 42.59 TFLOPS、GPU 総演算性能は倍精度 1.24 PFLOPS、半精度 20.2 PFLOPS である。TensorFlow、Pytorch 等のツールも備えることで、機械学習・深層学習用途に活用しやすいように整備したシステムである。

2.6　ストレージ

ストレージシステムは、DDN 社の EXAScaler を用いた Lustre ファイルシステムで構成している。HDD で構成される大容量ストレージの物理容量は 40.32 PB、データ転送速度は 280 GB/sec、SSD で構成される高速ストレージの物理容量は 4.06 PB、データ転送速度は 768 GB/sec である。

2.7　ソフトウェア

利用可能なソフトウェアスタックを図 2 に示す。OS は全システム共通で Red Hat Enterprise Linux 8 である。

ジョブスケジューラはオープンソースソフトウェアである Slurm Workload Manager を採用した。本センターの固有機能の一部はプラグインによる機能拡張が行われており、キューごとの資源量保障の仕組みや、要求リソース量を指定する機能、追い越し許容ポ

リシーなどが該当する。

Camphor3、Laurel3、Cinnamon3 向けの開発ツールとして Intel oneAPI を、Gardenia 向けの開発ツールとして NVIDIA HPC SDK を導入した。この他にも、図 2 に示した ISV アプリケーションを導入している。

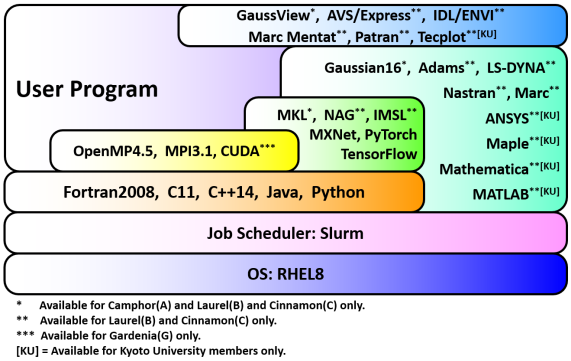

図 2　ソフトウェアスタック

3　提供サービス

本センターでは、全国の学術研究者に対し、定額の利用負担金によるサービスを提供している。サービスコースごとに利用可能な資源量やジョブの実行可能時間が異なっており、ユーザは利用目的に応じたコースを選択できるようになっている。

3.1　サービスコースの種類

本センターでは、エントリコース、パーソナルコース、グループコース、専用クラスタコースの 4 種類からなるサービスコースを用意している。

パーソナルコースは個人で利用するユーザのためのコースであり、Camphor3/Laurel3 は最大 2 ノード相当、Cinnamon3 は最大 1 ノード相当、Gardenia は最大 1GPU 相当の資源を提供している。

グループコースは研究室や共同研究グループ等の複数名での利用を想定したコースである。申請ノード数が常に確保されることを保証する「占有」タイプ、申請ノード数の 1/2 の確保を保証し、残りはシステムの空き状況に応じて提供する「優先」タイプ、申請ノード数の 1/4 の確保を保証し、残りはシステムの空き状況に応じて提供する「準優先」タイプ、常に確保するノードは無く、ベストエフォートで提供する「準々優先」タイプの 4 種類のタイプから選択が可能である。なお、「準々優先」は新システムから開始したサービスであり、電気代が高騰した状況においても、負担金を抑えたタイプを用意することで計算需要に応えつつ、ノードの確保を保証しないことで、節電のための縮退運転等、柔軟に運用できるように設計したタイプである。

専用クラスタコースは計算ノードの一部をスーパーコンピュータから切り離し、クラスタの先頭ノードにグローバル IP アドレスを割り当てることで、研究室のクラスタに近い環境を提供するコースである。

エントリコースは他のサービスコースに属さないユーザのためのコースである。最大 0.5 ノード相当の資源を提供している。

3.2　ストレージの提供

ホームストレージ、大容量ストレージ、高速ストレージの 3 種類のストレージ領域を提供している。

ホームストレージは、ユーザのホームディレクトリとして全ユーザに対して 100GB を割り当てている。大容量ストレージはパーソナルコース向けに 8TB、グループコース向けにはタイプ・申請ノード数に応じて標準提供している。

大容量ストレージは 2 つのファイルシステムに分割して構成しており、そのうち 1 つをバックアップ領域とし、週 1 回を目安にバックアップを行っている。大容量ストレージのバックアップはユーザ自身で無効にすることも可能であり、バックアップが有効の場合、使用可能な容量は提供している容量の 1/2 となるが、無効とすることで、提供している容量を全て使用することも可能である。標準提供の大容量ストレージでは不足する場合は、別途申請を行うことで容量を 10TB 単位で追加することも可能である。

高速ストレージについては、2023 年度を試用期間とし、パーソナルコース向けに 500TB、グループコース向けには申請ノード数に応じた容量を無償で提供している。

3.3　その他のサービス

前述のサービスの他に、短期間だけ高並列ジョブを実行したいユーザのための、Camphor3 または Laurel3 を 1 週間（7 日）単位で利用可能な大規模ジョブコース、研究室の PC にアプリケーションをインストールして利用可能な ENVI/IDL のライセンスサービス、学会や研究会のポスターセッションなどへの投稿支援のための大判プリンタのサービスを提供している。

4　利用支援制度

本センターで実施している、スーパーコンピュータシステムのユーザに向けた利用支援制度について紹介

する。いずれの制度も、課題募集後に審査を行い、採択課題を決定する方式である。

4.1　プログラム高度化共同研究

スーパーコンピュータをグループコースまたは専用クラスタコースで利用中の研究グループを対象に、大規模計算プログラムの高度化・高性能化を支援する制度である。本制度に採択された課題は、本センターとの共同研究として、プログラムの性能評価やより良いアルゴリズムの検討を実施し、プログラムを改良する取り組みである。本制度のプログラム開発等に要する費用は本センターが負担することで、ユーザの費用負担なしで実施している。

4.2　大規模計算支援枠

スーパーコンピュータをパーソナルコース、グループコースまたは専用クラスタコースで利用中の研究グループを対象に、「大規模ジョブコース」の利用負担金を一定範囲で本センターが負担する制度である。既に大規模なプログラムを持っているユーザや、プログラム高度化共同研究を利用したユーザが、その成果を元に、大規模計算を実行するための制度である。

4.3　若手・女性研究者奨励枠

40 才未満の若手研究者または女性研究者を対象とした奨励研究制度である。パーソナルコースまたはグループコースを利用可能であり、それらのコースを利用するために必要な利用負担金のうち、最大 10 万円を本センターで負担する制度である。また、採択された課題は「学際大規模情報基盤共同利用・共同研究拠点（JHPCN）」の萌芽型共同研究へ推薦する候補としている。

5　性能向上の達成状況

新システムを含めて、本センターで稼働していた直近 3 世代それぞれの総理論演算性能を図 3 に示す。2012 年～2016 年に稼働していたシステムから 2016 年～2022 年に稼働していた（以下、「旧システム」と言う）システムへの更新の際は、約 6.7 倍の性能向上を達成していたが、今回の更新では、約 1.8 倍の向上に留まっている。システム更新の計画を立てた当初は、CPU の性能向上のスピードが鈍化している状況や予算上の制約をふまえて、旧システムの 2～2.5 倍程度の達成を目指していたところであった。しかし、調達のタイミングでの円安や物価上昇等の影響もあり、当初想定以上にシステム規模を抑えざるを得なかった。

理論演算性能の見かけ上の伸びは想定より小さくなったが、高いメモリバンド幅を持つシステムを導入

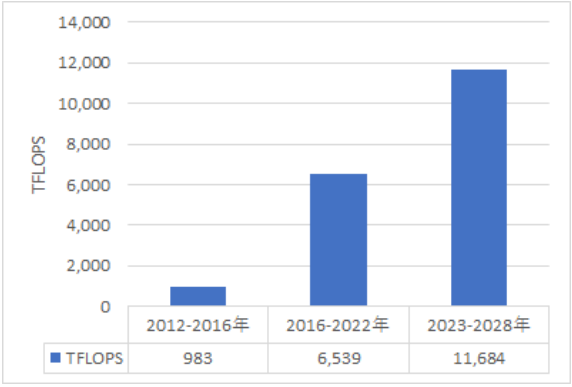

図 3　過去 3 世代性能比較

したことで、実際の性能の面では旧システムよりも性能を発揮しやすい環境になったと考えている。表 1 に示す新/旧システムの Camphor の Top500[2] のデータからは、実効効率が大幅に上昇し、実効性能を示す Rmax の値では、2.19 倍の向上を達成している状況を確認できている。ユーザの体感的にも性能向上のメリットを感じられるものと考えている。本稿での紹介は省略するが、いくつかのベンチマークプログラムでは、2～5 倍の性能向上を得られていることを確認できている状況である。

表 1　Camphor 性能比較

	旧システム	新システム	新旧比
Rmax [TFLOPS]	3,057	6,708	2.19 倍
Rpeak [TFLOPS]	5,483	7,627	1.39 倍
実行効率 [%]	55.76	87.96	1.58 倍

6　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の概要およびサービスについて紹介し、以前導入していたシステムとの性能比較を行った。

Camphor3 の総演算性能は 7.63 PFLOPS、Laurel3 の総演算性能は 2.65 PFLOPS、Cinnamon3 の総演算性能は 114.6 TFLOPS となっている。また、今回新しく導入した Gardenia では、ノードあたり GPU を 4 基搭載しており、機械学習・深層学習方面での利用が見込まれる。

サービスコースは新システムからグループコースに「準々優先」タイプを新設したことで、システムの柔軟な運用が可能になった。

システムの性能としては総理論演算性能として約1.8 倍のシステムを導入した。更新計画当初の目標には届かなかったものの、高いメモリバンド幅を持つシステムを導入したことで旧システムより性能を発揮することができるようになったと考えている。

サービス面、システム面ともに、今後もより良い計算機環境を実現するための取り組みを続けていく所存である。

参考文献

[1] 深沢圭一郎. 新スーパーコンピュータシステムのご紹介. 京都大学情報環境機構広報誌「Info!」. 2023, vol 28, p.10-12. https://www.iimc.kyoto-u.ac.jp/info28.pdf, (参照 2023-09-22).

[2] “Top500”. TOP500. https://www.top500.org/, (参照 2023-09-22).