筑波大学
University of Tsukuba

2022 年 12 月 22 日
国立大学法人筑波大学

ビッグメモリスーパコンピュータ Pegasus の試験稼働開始
～第 4 世代 Intel Xeon、NVIDIA H100 PCIe GPU、Intel 不揮発性メモリを搭載～

筑波大学計算科学研究センターは、新たなスーパコンピュータ Pegasus（ペガサス）の試験稼働の開始を発表しました。演算性能、メモリ帯域幅[注 1)]、メモリサイズを大きく向上させ、計算科学のみならずビッグデータ解析、超大規模 AI 分野を強力に推進します。

Pegasus に搭載される演算加速装置（GPU）は、倍精度浮動小数点演算[注 2)]における理論ピーク性能が 51 TFlops[注 3)]（従来よりも 2.7 倍高速）であり、CPU（中央演算処理装置）と高帯域幅の PCIe Gen5[注 4)]により世界で初めて接続されます。DDR5 メモリ[注 5)]（従来よりも約 2 倍高速）、不揮発性メモリ[注 6)]を搭載し、大容量メモリまたは超高速ストレージとしての利用が可能です。また、ネットワークについても最新の 400Gbps ネットワークプラットフォームを利用します。

Pegasus は 120 ノードの計算ノードで構成され、全体の理論ピーク性能は GPU 部分だけでも 6.1 PFlops を超えます。筑波大学計算科学研究センターは、共同利用・共同研究拠点として、学際共同利用、HPCI[注 7)]共同利用、一般利用などの各種利用プログラムにより、Pegasus を全世界のユーザに提供し、さらなる計算科学の発展に寄与します。

背景

2012 年のスーパーコンピュータ「京」から 2020 年の「富岳」までの 8 年間において、演算性能は約 50 倍となりましたが、メモリ容量は 3.8 倍にしかなっていません。演算性能に対しメモリサイズが減少する傾向は富岳に限らず近年のスーパコンピュータで顕著となっています。ところが、データ駆動科学、AI 駆動科学においては計算ノードのメモリサイズは重要です。大規模データを高速に扱うためには、なるべく多くのデータをメモリに載せデータの再利用率を高めるとともに、計算ノード間の通信を減らした方が効率を上げられるためです。また、大規模なデータを高速に扱うためにはストレージ性能も重要となります。

一方、DRAM は大容量のメモリを搭載すると消費電力が高くなり、また高コストとなる問題があります。この問題を解決するため、Pegasus では DRAM と不揮発性メモリを導入することとしました。不揮発性メモリは DRAM に比べ消費電力が低く、容量単価は一桁下です。アクセス遅延については幅があるものの、最低遅延は DRAM に迫ります。不揮発性メモリは拡張メモリとして利用可能なだけではなく、不揮発であるため永続データ構造を用い超高速ストレージとしても利用可能です。計算科学研究センターでは 2020 年より不揮発性メモリを導入して評価を行ってきました。不揮発性メモリによるメモリ拡張によりメモリ容量を超える問題サイズの計算が可能となり、また計算科学アプリケーションの性能はほとんど変わらないことがわかりました。

また、計算科学、データ駆動科学、AI 駆動科学において、演算加速装置（GPU）が有効であることは既に示されています。Pegasus においては、最先端の CPU、GPU、メモリ、不揮発性メモリを搭載し、最先端のネットワークで接続するよう設計を行いました。

設計

Pegasus では、高帯域大容量メモリと超高速ストレージによる計算科学、ビッグデータ解析、超大規模 AI を促進するため、次のような設計を行いました。

GPU メモリ、CPU メモリ、入出力の帯域幅を向上させるために、最新の HBM2E、DDR5 メモリ、PCIe Gen5 を用います。HBM2E は 3 次元積層 SDRAM で、富岳で利用されている HBM2 より高速です。DDR5、PCIe Gen5 は現行の DDR4、PCIe Gen4 に比べ帯域幅が 2 倍となります。GPU については、高い演算性能はもとより、高帯域の PCIe Gen5 で CPU と接続することにより CPU-GPU 間の転送性能を向上させます。不揮発性メモリは、DDR5 メモリの DIMM スロットに搭載し、アクセス性能を向上させます。また、不揮発性メモリは、メモリの拡張としての利用と、不揮発性メモリの直接アクセスの両方を利用可能とし、拡張メモリのサイズは利用者が指定可能なようにします。

大規模データを効率的にアクセスするため、大容量で高性能な並列ファイルシステムを用います。しかしながら、並列ファイルシステムの性能と演算性能のギャップが広くなっているため、各計算ノードには不揮発性メモリに加え、高帯域な NVMe SSD を搭載します。計算ノードの不揮発性メモリ、NVMe SSD は並列ファイルシステムのキャッシュ、および一時的なストレージ領域として利用することにより、性能ギャップを埋めます。

システム開発

計算ノードのストレージは並列ファイルシステムと演算性能のギャップを埋めるために用いますが、効率的に活用するためのソフトウェアがまだ開発されていないため、システム開発を行っています。

計算ノードはジョブが割当てられ、実行されてから初めて利用可能となります。そのため、ジョブ実行中にしか利用することはできません。これまで主に計算ノードのストレージは、ジョブ開始時に並列ファ

イルシステムから指定したデータをコピーし、ジョブ終了時に指定したデータを並列ファイルシステムに書き戻すことで用いられてきました。この処理は、データの指定が煩雑で、また間違えるとジョブが正しく実行できない、あるいは実行結果が失われる問題がありました。また、ジョブは複数の計算ノードで実行されますが、他の計算ノードのストレージをアクセスすることができず、効率的な利用ができていませんでした。

そこで、ジョブが実行される複数の計算ノードのストレージを用いた、並列キャッシュファイルシステム CHFS/Cache の開発を行っています。この並列キャッシュファイルシステムは、ジョブの実行中に一時的に構築されます。キャッシュするデータは指定することもできますし、自動的にキャッシュすることもできます。更新されたデータは自動的に並列ファイルシステムに書き出されます。これにより、データ指定が間違っていても正しくジョブが実行されます。また並列アクセスが可能であり、高い性能も実現できます。これまでの評価では、並列ファイルシステムより高い帯域幅、メタデータ性能、スケーラビリティを示し、並列ファイルシステム性能と演算性能のギャップを埋めることに成功しています。

運用

筑波大学計算科学研究センターは、共同利用・共同研究拠点として、学際共同利用、HPCI [注 7)] 共同利用、一般利用などの各種利用プログラムにより、Pegasus を全世界のユーザに提供し、さらなる計算科学の発展に寄与します。一般のユーザーの利用は 2023 年 4 月からを予定しています。

仕様

システム名称	Pegasus
製造	NEC
全体性能	> 6.1 PFlops
ノード数	120
ネットワーク	NVIDIA Quantum-2 InfiniBand プラットフォームによるフルバイセクションファットツリーネットワーク
並列ファイルシステム	7.1PB DDN EXAScaler (帯域幅 40 GB/s)

計算ノード

CPU	第 4 世代 Intel Xeon スケーラブルプロセッサ（旧コードネーム Sapphire Rapids）（48 コア）
GPU	NVIDIA H100 PCIe GPU (FP64 テンソルコア演算 51 TFlops、80GB HBM2E、2 TB/s)
メモリ	128GiB DDR5 (282 GB/s)
不揮発性メモリ	Intel Optane 不揮発性メモリ（コードネーム Crow Pass）
SSD	2 x 3.2TB NVMe SSD (7 GB/s)
ネットワーク	NVIDIA Quantum-2 InfiniBand プラットフォーム (200 Gbps)

用語解説

注１） 帯域幅

メモリ、GPU、ストレージ等の転送性能。毎秒の転送バイト数で表現される。帯域幅が高いとデータ転送が高速に行われる。

注２） 倍精度浮動小数点演算

64 ビットで表現される浮動小数点数の演算。IEEE 754 で定められる。計算科学で通常利用される。

注３） TFlops, PFlops

倍精度浮動小数点演算性能を表す単位。1 TFlops は 1 秒間に 1 テラ（= 10^{12}）回演算を行う性能であり、1 PFlops は 1 秒間に 1 ペタ（=10^{15}）回演算を行う性能である。

注４） PCIe Gen5

入出力シリアルインタフェースである PCI Express の第 5 世代の規格。第 4 世代の PCIe Gen4 の 2 倍の速度を持つ。

注５） DDR5 メモリ

DRAM の規格で、DDR4 メモリに比べ帯域幅が 2 倍になる。

注６） 不揮発性メモリ

メモリと同様にアクセス可能な不揮発性デバイス。メモリと異なり電源を切ってもデータは保持される。一般的にフラッシュ等の NAND デバイスより高速で、メモリに近い性能をもつ。

注７） HPCI

革新的ハイパフォーマンス・コンピューティング・インフラ（HPCI）。筑波大学を含む国内の大学や研究機関の計算機システムやストレージを高速ネットワークで結んだ環境であり、その環境を使うために HPCI 共同利用プログラムがある。

問合わせ先

【スーパコンピュータに関すること】
建部 修見（たてべ　おさむ）
筑波大学計算科学研究センター　教授
URL: https://ccs.tsukuba.ac.jp/

【取材・報道に関すること】
筑波大学 計算科学研究センター 広報・戦略室
TEL: 029-853-6260
E-mail: pr@ccs.tsukuba.ac.jp