大規模並列分散システムの超高速データ処理を実現するデータセントリックアーキテクチャ

Data-Centric Architecture to Realize Ultra-High-Speed Data Processing for Large-Scale Parallel and Distributed Systems

木下　敦寛
■KINOSHITA Atsuhiro

企業や公的機関が取り扱うデータ量やデータ処理の複雑さが飛躍的に増加し続けており，大量かつ多様なデータを高速に処理するには，データを多数のコンピュータに分散して処理できる，大規模並列分散システムが必要となる。ところが，一般的な並列分散システムでは，実際のデータ処理以外にも，データの移動や前処理などの数多くのプロセスが必要となり，所望のデータ処理速度を実現することは困難であった。

東芝は，NAND型フラッシュメモリを活用して超高速データ処理を実現するための新しいコンピュータプラットフォームを考案した。データセントリックアーキテクチャを採用し，ノードコントローラにネットワークポートを備えることで，優れたスケールアウト特性と高速なデータ処理性能を兼ね備えるとともに，エンタープライズ用途にも適用可能な高信頼性及び高可用性も実現している。大容量データの高速処理が必要とされるビッグデータ解析プラットフォームとしての利用が期待できる。

With the continuing increase in the volumes of data and complexity of data processing handled by facilities in both the private and public sectors, large-scale parallel and distributed systems, which can process data distributed to a large number of computers in parallel, are necessary to deal with large volumes of diverse data at high speed. However, typical parallel and distributed systems require a number of processes such as moving or preprocessing of the data in addition to the actual data processing, making it difficult to achieve the desired data processing speed.

To resolve this issue, Toshiba has devised a new computer platform for ultra-high-speed data processing utilizing NAND flash memories. By adopting a data-centric architecture incorporating node controllers equipped with a network port, the new platform realizes excellent scale-out characteristics and high-speed data processing performance as well as sufficient reliability and availability to handle enterprise applications. It is expected to be utilized as a platform for big data analysis that requires high-speed processing of large volumes of data.

1　まえがき

近年，企業や公的機関が取り扱うデータ量やデータ処理の複雑さが飛躍的に増加し続けており，容量（Volume），多様性（Variety），及び速度（Velocity）の3Vを兼ね備えた大規模データ処理，いわゆるビッグデータ処理を効率的に行えるコンピュータプラットフォームが求められている。大量かつ多様なデータを高速に処理するには，コンピュータ単体の性能を向上させるだけでは不十分で，データを多数のコンピュータに分散して並列処理ができる，Hadoop(†)などに代表される大規模並列分散システムが必要となる。ところが，一般の並列分散システムでは，実際のデータ処理以外にも，分散されたデータを一元的に管理するためにデータの移動や，前処理，メタデータ処理といった数多くのプロセスが必要になり，大規模なシステムで所望のデータ処理速度を実現することは困難であった。

東芝は，NAND型フラッシュメモリと独自のデータセントリックアーキテクチャを採用し，並列分散システムに必要なデータ管理のプロセスを最小限にすることで，優れたスケールアウト特性と高速データ処理を実現できる新しいコンピュータプラットフォームを考案した。このアーキテクチャは，エンタープライズ用途にも適用可能な高信頼性及び高可用性も実現している。ここでは，その特長と理論性能などについて述べる。

2　新アーキテクチャの特長

2.1　データセントリックアーキテクチャ

新アーキテクチャの第1の特長は，アクティブなデータを全てNAND型フラッシュメモリに格納することである。NAND型フラッシュメモリは，DRAMの1/5 ～ 1/10という安価なビット価格と，HDD（ハードディスクドライブ）の500 ～ 1,000倍という高い応答速度を併せ持っており，ビッグデータを高速処理するのに適している。新アーキテクチャの第2の特長は，データセントリックアーキテクチャとなっていることである（**図1**）。データセントリックアーキテクチャは，強力なCPUになるべく高速で大量のデータを送って高速処理を実現しようとする，従来型のCPUセントリックアーキテクチャとは異なり，データ処理に参加するCPUの並列数を増やすことでシステム全体の処理速度を増大させる。したがってこれは，大量かつ多種なデータの高速処理が必要な，ビッグデータの解析プラットフォームとして適したアーキテクチャと考えられる。

新アーキテクチャでは，NAND型フラッシュメモリを読み書きするノードコントローラ（NC）が，互いにデータを送受信する

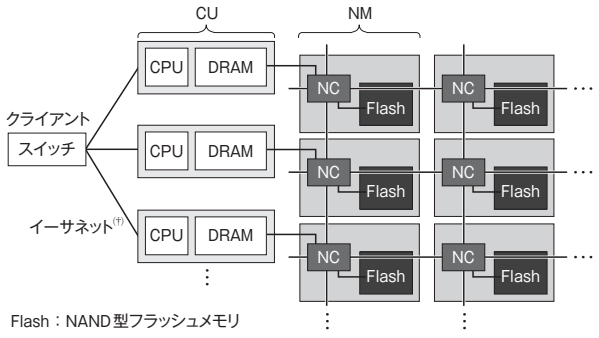

図1．新アーキテクチャの構成 — NCのネットワークポートを用いて複数のNMをマトリクス状に接続して構成されたNAND型フラッシュメモリ空間を，複数のCUが共有している。
Configuration of new architecture

ためのネットワークポートを持っており，このポートを相互に接続することでマトリクス状の2次元ネットワークが構成されている。このNCとNAND型フラッシュメモリをまとめてノードモジュール（NM）と呼ぶ。NMの2次元ネットワークは，全体が単一のアドレス空間でアクセス可能な高速ストレージとなっている。

NCは，CPUやDRAMから成るコネクションユニット（CU）と直接接続可能なインタフェースを持っており，NAND型フラッシュメモリ空間にある巨大なデータセットを多数のCPUで共有し，並列に処理できる。これにより，個々のCPUが多少非力でも，多数接続することで，システム全体として高性能にできる。

この新アーキテクチャは，中規模から大規模なプライベート及びパブリッククラウドサービスなどのサービス事業者や，高性能アプライアンスのディストリビュータなどにとって非常に有用と考えている。これらの顧客は，性能やコストに加え，データの信頼性及び可用性に対する要求が非常に高い。そこで新アーキテクチャは，コンポーネントレベルでの単一障害点（SPOF：Single Point of Failure）をなくすことで，部品交換などの保守作業をオンサービスで行える高可用性を持たせたり，RAID5（Redundant Array of Independent（Inexpensive） Disks 5）によるデータ保護機能を実装して性能の劣化なくデータの高信頼性を確保したりすることも考慮されている。

2.2　新アーキテクチャのスケールアウト特性

従来型の並列分散システムと比較した場合の新アーキテクチャの利点は大きく分けて二つある。一つ目はスケールアウト特性に優れていることである。ビッグデータ処理を指向したシステムの規模を拡大する場合，コンピュータ単体の性能を向上させる“スケールアップ”で対処すると，処理対象となるデータの規模が大きすぎて追いつかないことが多い。そこで，並列分散処理を活用して，全体の性能を規模に比例して向上させる“スケールアウト”と呼ばれる手法がとられるが，後述するように，大規模システムになるほど，データの管理プロセスが複雑になり，実現するのが困難になる。

図1で示したように，新アーキテクチャでは，任意のCU上のCPUから，任意のNAND型フラッシュメモリに対してアクセスが発生する。システム全体の性能を決めているのはこの際のアクセス時間である。アクセス時間は，①CPU内部の処理時間，②2次元ネットワーク内のデータの往復転送時間，及び③NAND型フラッシュメモリのアクセス時間とコントローラの処理時間，の三つの要素から成るが，システム規模を大きくした際に増加するのは②である。新アーキテクチャでは，2次元ネットワーク内のデータのやり取りを専用ハードウェアで高速化して，②の時間が③と比べて十分小さい数十μsになるように設計されており，システム規模を拡大してもアクセス時間が大きく変わらないようになっている。システム規模を大きくすれば，CPU数すなわち並列処理能力は規模に比例して増加するため，結果として全体の性能が規模に比例して向上する。したがって，新アーキテクチャは，優れたスケールアウト特性を持ったシステムを実現できるアーキテクチャと言える。

2.3　新アーキテクチャにおける並列分散処理

新アーキテクチャの二つ目の利点は，データ管理が容易で高速なデータ処理が可能なことである。従来型の並列分散システムでは，一連のデータでも異なるコンピュータノードに分割されて格納されている。したがって，必要なデータを集めて処理をする場合，CPUはデータを格納している自分以外のコンピュータノードと通信し，イーサネット(†)を介して必要なデータを自ノードに移動させてから所望の処理を行う必要がある（**図2**(a)）。しかも，目的のデータにたどり着くためには，何らかの目印となるデータ（メタデータ）を何度もたどらなければならないことも多く，必要なデータを見つけて移動するという処理は，パフォーマンスを低下させる大きな要因となっている。

新アーキテクチャでは，前述のようにNAND型フラッシュメモリ空間が全てのCPUで共有されているため，必要なデータをイーサネット(†)を介して移動させる必要がない（図2(b)）。このようなアーキテクチャをシェアードエブリシングアーキテクチャと言う。各コンピュータノードがデータのやり取りをしなければならない従来型の並列分散システムでは，ノード数が増えると行き交うデータ量もそれに伴って増え，いずれイーサネット(†)を介したデータ移動がスケールアウトの阻害要因となってしまう。新アーキテクチャではこのようなことはない。

シェアードエブリシングアーキテクチャは利点ばかりではない。全てのデータが複数のCPUで共有されるということは，あるCPUがデータを処理している間に他のCPUがそのデータにアクセスできてしまうことを意味する。これは，データの一貫性などが求められる処理では致命的な不具合を引き起こすことがある。したがって，シェアードエブリシングアーキテクチャは一般にデータに対する排他制御の仕組みを持つ必要がある。

従来のシェアードエブリシングアーキテクチャを持った並列

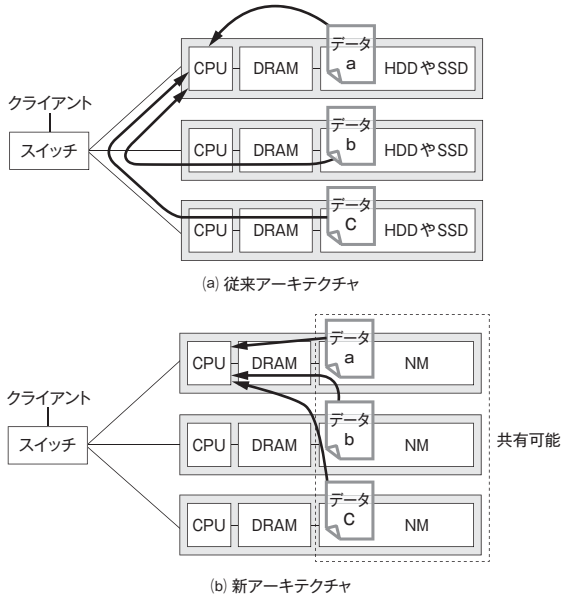

図2. 従来アーキテクチャとの違い ― 従来アーキテクチャでは，HDDやSSD（ソリッドステートドライブ）にある一連のデータ（a，b，及びc）を一番上のCPUで処理する場合，他のコンピュータノードにあるデータ（b及びc）をイーサネット(†)を介して移動させる必要がある。新アーキテクチャではこの必要がないので，データ管理が容易になっている。

Differences between conventional and new architectures

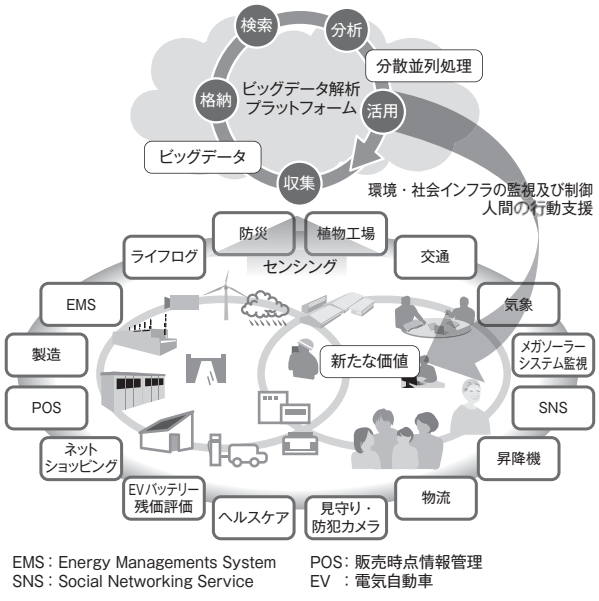

図3. 新アーキテクチャに基づくプラットフォームの利用イメージ ― ビッグデータ解析プラットフォームを中心に，ビッグデータを収集して解析し，活用することで新たな価値を生み出すといったサイクルを構築できれば，環境・社会インフラの監視及び制御や，人間の行動支援など，様々な応用が考えられる。

Conceptual diagram of applications for platform based on new architecture

分散システムでは，コンピュータノードどうしが通信を行って排他制御を実現しているが，これは先に述べたのと同じ理由でスケールアウトの阻害要因となっている。これに対して新アーキテクチャでは，NCにこの排他制御の仕組みを持たせることで，排他制御のためのコンピュータノードどうしの通信を最小限に抑えている。

並列分散システムにおけるパフォーマンスの低下要因，すなわちボトルネックをひと言で言えば，“コンピュータノードが複雑なデータ管理を行っている”ことである。これまでは，このボトルネックが顕在化しないように，データの処理や，格納方法，システムの組み方などが様々に工夫され，最適化されてきたが，新アーキテクチャの適用先と考えているビッグデータ処理では，処理の内容やデータの性質などを予測することが難しく，ボトルネックが顕在化してしまうことが多い。新アーキテクチャは，データ管理プロセスを最小限にしてパフォーマンス低下を抑えることで，高速なデータ処理ができる。

2.4 新アーキテクチャの利用イメージ

従来型の並列分散システムによって構築されたビッグデータ解析プラットフォームでは，P（ペタ：10^{15}）バイト級の大規模データ処理は，HDDベースのシステムを用いたバッチ処理で行い，1 msを切るようなリアルタイム処理は，データ量をT（テラ：10^{12}）バイト級に抑えてDRAMベースのインメモリシステムで処理する，というように，データ容量と処理速度の間には明確なトレードオフが存在した。新アーキテクチャは，NAND型フラッシュメモリの“DRAMより安価なビット価格”と“HDDより高い応答速度”を活用しつつ，独自アーキテクチャに基づく優れたスケールアウト特性と高速なデータ処理を実現できる，新しいビッグデータ解析プラットフォームとなる。これは，従来型の並列分散システムにおけるデータ容量と処理速度のトレードオフを打破する可能性がある。したがって適用先として期待されるシステムのイメージは，ビッグデータ解析一般，すなわち実世界に置かれた多数のセンサから多種多様で大量のデータを高速に収集して格納し，それらを超高速に解析することで価値ある新しいデータを生み出し，それを実世界上に作用させて活用する，というものになる（**図3**）。

ターゲットアプリケーションとしては，ビッグデータを生かした，環境・社会インフラの監視及び制御や，人間の行動支援など多岐にわたる。今後，新アーキテクチャに適するターゲットアプリケーションやユースケースの選定を進めていく。

3 理論性能

ここでは，新アーキテクチャに基づいた並列分散システムに期待される基本性能について，いくつかの仮定を置いた理論計算の結果とともに述べる。

3.1 データアクセス速度

新アーキテクチャにおける，CUからNAND型フラッシュメモリに対するアクセスは，読み書きコマンドやデータを含んだパ

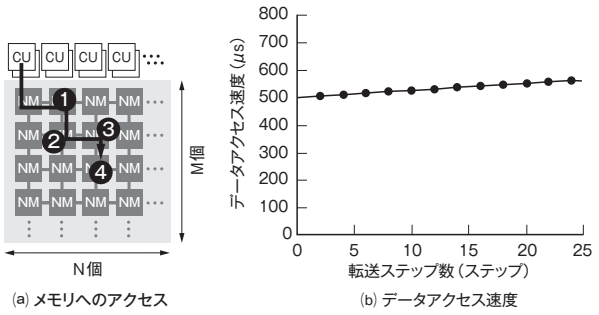

図4．データアクセス速度 — 新アーキテクチャでは命令やデータをNMごとに1ステップずつ転送することで所望のNAND型フラッシュメモリにアクセスする。したがって，データアクセスに掛かる時間は転送ステップ数に比例する。ここでは，1ステップの転送に2.5μs，転送以外の処理に500μs掛かると仮定して計算した。

Data access latency

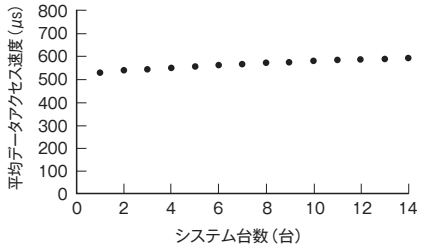

図5．システム台数とデータアクセス速度の関係 — 1システムが24×16 NMで構成されていると仮定して計算した。システム台数を変化させても，アクセス速度には大きな影響がない。

Relationship between data access latency and number of systems

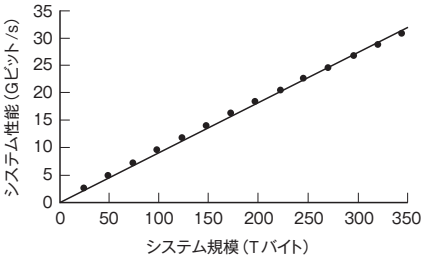

図6．システム規模とシステム性能の関係 — 1 NMが64 GバイトのNAND型フラッシュメモリを搭載するとした場合のシステム規模（ストレージ容量）と，1システムが24個のCPUを搭載し，各CPUが16スレッドでアクセスを行うとした場合のシステム性能（システム全体のデータスループット）を計算した。システム規模の拡張に伴ってシステム性能が線形に増加しており，良好なスケールアウト特性を持っていることがわかる。

Relationship between system performance and system scale

ケットが2次元ネットワークによってバケツリレーのように転送されることで実現されている（**図4**(a)）。したがって，データアクセスに掛かる時間（レイテンシ）は，CUからNMに至るまでの転送ステップ数を変数として，切片を持った直線となる（図4(b)）。

切片は，転送ステップ数とは無関係な時間で，CPU内部の処理時間やNAND型フラッシュメモリのアクセス時間とコントローラの処理時間などが主な要素である。また，直線の傾きはパケットを1回転送するのに掛かる時間，すなわちパケットサイズを2次元ネットワークのスループットで除したものとなる。

個々のデータアクセスを見た場合のレイテンシは，目的のNAND型フラッシュメモリの場所によって変わるが，この場所がランダムに決まると近似できれば，平均ステップ数を求めることができる。この平均ステップ数はシステムを拡張するにつれて大きくなる。新アーキテクチャでは，このレイテンシの増分を最小限にするため，ドーナツ状の2次元トーラスネットワーク化することもでき，その場合の平均ステップ数はトーラス化しない場合の1/2になる。

3.2　スケールアウト特性

新アーキテクチャにおけるシステム台数と平均データアクセス速度の関係を**図5**に示す。ここでは，1システムが24×16個のNMから構成されると仮定している。この計算の範囲では，システム台数を変えても，平均データアクセス速度は大きくは変わらない。2.2節で述べたように，新アーキテクチャは，システム規模を拡大してもレイテンシが大きく変わらないようになっていて，優れたスケールアウト特性が期待できる。

レイテンシの逆数を取ることで，データのスループット，すなわちシステム性能を計算できる。新アーキテクチャのシステム規模（ストレージ容量）とシステム性能の関係を**図6**に示す。ストレージ容量は，1 NM当たり64 GバイトのNAND型フラッシュメモリが搭載されていると仮定して算出した。システム規模の拡張に伴ってシステム性能が線形に増加しており，新アーキテクチャが良好なスケールアウト特性を持っていることがわかる。

4　あとがき

NAND型フラッシュメモリと当社独自の新アーキテクチャを組み合わせ，超高速データ処理を実現する並列分散処理システムを考案した。このシステムが，当社の目指すHuman Smart Communityを支えるビッグデータ処理を効率的に行うためのビッグデータ解析プラットフォームとして貢献できるように，今後も研究開発を進めていく。

木下　敦寛　KINOSHITA Atsuhiro, Ph.D.
セミコンダクター＆ストレージ社　ストレージプロダクツ事業部　ストレージソリューション推進部主査，博士（工学）。アーキテクトとして新規ストレージソリューションの開発に従事。
Storage Products Div.