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

● KINOSHITA Atsuhiro

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. Introduction

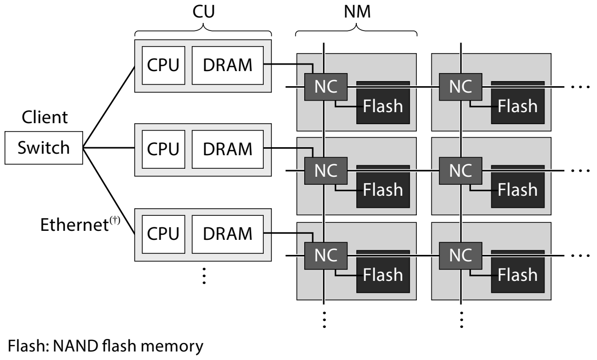
In recent years, the volume of data handled by private enterprises and public organizations has been growing at a tremendous rate, and data processing is also becoming increasingly complex. In order to efficiently process big data or "high-volume, high-velocity, and high-variety (3V) information assets," a novel computer platform is required. To process a huge volume of data that comes in a variety of formats, it is not enough to improve the performance of each computer; a large-scale parallel distributed system such as Hadoop(†) is also necessary to distribute large data sets across a number of computers running in parallel. However, in order to achieve centralized management of distributed data, typical parallel distributed systems require many processes other than actual data processing, including data movement, preprocessing, and metadata processing. This makes it difficult to increase the computing performance of large-scale systems.

Against this background, Toshiba has devised a novel computer platform that has outstanding horizontal scalability and computing performance. The new computer platform uses NAND flash memory and a proprietary data-centric architecture, and is designed to minimize data management processes necessary for parallel distributed processing. Furthermore, this architecture also provides high reliability and high availability required for enterprise use. This report describes its characteristics and theoretical performance.

2. Characteristics of the new architecture

2.1 Data-centric architecture

The new architecture has some noteworthy characteristics. First, all active data are stored in NAND flash memory. NAND flash memory delivers cost-per-bit one-fifth to one-tenth that of dynamic random access memory (DRAM) and is 500 to 1 000 times faster than hard disk drives (HDDs). Because of the cost and performance advantages, NAND flash memory is well suited to the processing of big data. The second characteristic of the new architecture concerns data centricity (**Figure 1**). Whereas the conventional CPU-centric architecture tries to increase computing performance



Flash: NAND flash memory

Figure 1. Configuration of new architecture.
Multiple CUs share a single NAND flash memory space that consists of multiple NMs interconnected in a matrix using the network ports of NCs.

by sending a large amount of data to powerful CPUs at the highest possible speed, data-centric architecture improves the overall system performance by increasing the number of CPUs operating on data sets in parallel. Therefore, data-centric architecture is more suitable for big-data analytics platforms that must process a huge volume of data of great variety at high speed.

The new architecture consists of many node controllers (NCs) that are associated with their own NAND flash memory, and all NCs have a network port to form a two-dimensional (2D) matrix so as to send and receive data to/from one another. An NC and the associated NAND flash memory comprise a unit called a node module (NM). The entire 2D network of NMs serves as one high-speed storage system addressable as a single address space.

Each NC has an interface that directly connects to a control unit (CU) consisting of a CPU and a DRAM, and many CPUs can share and operate in parallel on a huge amount of data sets in the NAND flash memory space. The parallel operation aggregately results in high computing power even if each CPU is not best in class.

The new architecture will greatly benefit medium- to large-scale private and public cloud service providers, high-performance appliance distributors, and so on. These customers tend to have stringent requirements not only for performance and cost but also for system availability and data reliability. To address these requirements, the new architecture provides full component-level redundancy to eliminate single points of failure (SPOF) and thus allow in-service system maintenance, as well as redundant array of independent disks 5 (RAID5) to protect data with high reliability without compromising performance.

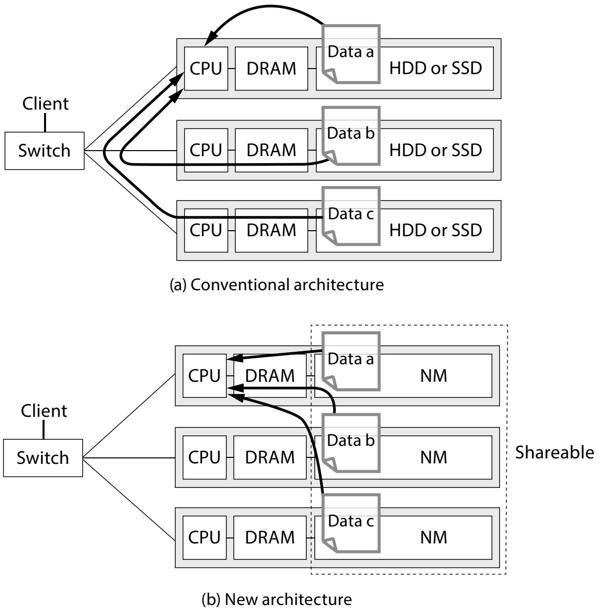
2.2 Horizontal scalability of the new architecture

Compared to conventional parallel distributed systems, the new architecture has two major advantages. The first advantage is its excellent horizontal scalability. In the case of an architectural model called vertical scaling (scale-up), resources are added to an existing single computer when there is a need to scale a computer system. However, for computer systems oriented for big-data applications, a scale-up model is unsuitable, in general, because of difficulty in handling data volume growth. An alternative architectural model is horizontal scaling (scale-out), which increases the overall system performance by adding more nodes to a system. However, as described above, a larger system means increased management complexity and thus increased difficulty in implementation.

In the new architecture, a CPU in any CU can access NAND flash memory in any NM, as shown in Figure 1. The overall system performance is governed by this access time, which consists of three terms: (1) the CPU run-time, (2) the time taken for data to make a round trip within the 2D network, and (3) the access time of NAND flash memory plus the controller run-time. An increase in the system size causes only the second term to increase. The new architecture is designed using dedicated hardware to speed up the data traffic within the 2D network, reducing the second term to a few tens of microseconds, which is considerably smaller than the third term. Consequently, an increase in the system size does not affect the access time significantly. As the system size is scaled, the number of CPUs increases, resulting in a greater parallel processing capability and thus higher overall system performance. Therefore, the new architecture makes it possible to build a computer system with excellent horizontal scalability.

2.3 Parallel distributed processing in the new architecture

The second advantage of the new architecture is the ease of data management, which leads to faster computing. Typical parallel distributed systems split and store a given data set across multiple computer nodes. When a given computer node works on a collection of data sets, each node needs to communicate with other nodes where the required data are stored, and then move them to itself via Ethernet[(†)] (**Figure 2**(a)). In order to access these data items, the computer node must trace a set of



(a) Conventional architecture

(b) New architecture

Figure 2. Differences between conventional and new architectures.

(a) depicts the conventional architecture. In order for the top CPU to operate on data items a, b and c in HDDs or SSDs, the data items stored in other computer nodes (b and c) must be brought to the CPU via Ethernet(†) . (b) illustrates the new architecture that eliminates the need for this data movement and thus simplifies data management.

data (called metadata) that describe or give information about the target data items many times. Thus, finding and moving data in a distributed system causes significant performance degradation.

In contrast, all the CPUs in the new architecture share the entire NAND flash memory space as described above, eliminating the need to move data via Ethernet(†) (Figure 2(b)). Such a model is called a shared-everything architecture. In a conventional parallel distributed system, computer nodes must exchange data with one another. Consequently, the amount of Ethernet(†) network traffic increases according to the number of nodes and can be a limiting factor for horizontal scaling. The new architecture is free from this network bottleneck.

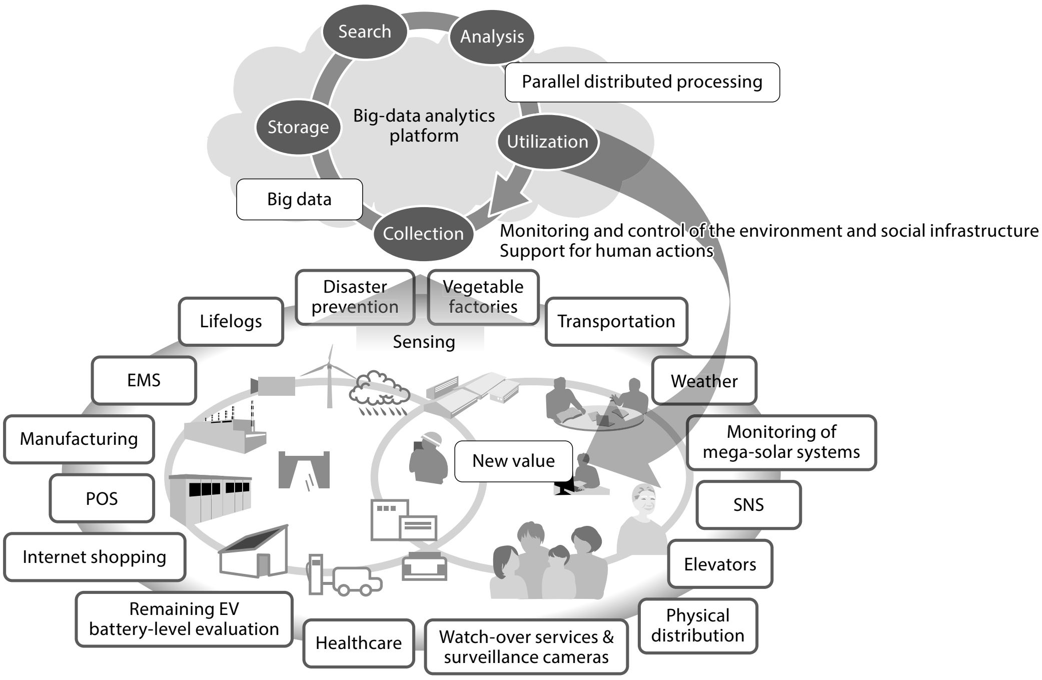
Nevertheless, the shared-everything architecture has its own drawbacks. Sharing all data across multiple CPUs means that while a given CPU is working on a certain data set, other CPUs can also access the same data set. Therefore, fatal errors could occur in processes during which data coherency must be maintained. To avoid this problem, a shared-everything architecture generally requires exclusive control over data.

The conventional parallel distributed systems with a shared-everything architecture achieve exclusive control through communications among computing nodes, but as described above, network traffic becomes a limiting factor for horizontal scaling. In contrast, in our new architecture, NCs provide exclusive control, minimizing network communications between computer nodes.

In essence, complex data management delegated to computer nodes is a major cause of performance degradation or a bottleneck in a parallel distributed system. Thus far, various techniques have been used to prevent this bottleneck from occurring, including the way of data storage and processing, and system configuration. However, with big data, it is difficult to predict the nature of processing or the types of incoming data, causing the bottleneck to show up frequently. The new architecture reduces the performance degradation by minimizing the data management processes and thus achieves high-speed computing.

2.4 Usage scenarios for the new architecture

Big-data analytics platforms with a conventional parallel distributed architecture have a tradeoff between data volume and computing latency. When the data volume is in the order of petabytes (peta: 10^{15}), an HDD-based system is commonly utilized for batch processing. When real-time processing with a latency of less than one millisecond is required, data volume must be kept in the order of terabytes (tera: 10^{12}) to use a faster DRAM-based in-memory system. In this situation, the new architecture can serve as a new big-data analytics platform. Since the new architecture is based on NAND flash memory that has lower per-bit cost than DRAM and faster response time than HDDs, it delivers both


Search
Analysis
Parallel distributed processing
Storage
Big-data analytics platform
Utilization
Big data
Collection
Monitoring and control of the environment and social infrastructure
Support for human actions
Lifelogs
Disaster prevention
Vegetable factories
Transportation
Sensing
EMS
Weather
Manufacturing
Monitoring of mega-solar systems
New value
POS
SNS
Internet shopping
Elevators
Remaining EV battery-level evaluation
Healthcare
Watch-over services & surveillance cameras
Physical distribution

EMS: energy management system POS: point-of-sale SNS: social networking service EV: electric vehicle

Figure 3. Conceptual diagram of applications for platform based on new architecture.

The key is the use of big-data analytics platforms to collect and analyze big data and use the results to create new value. This cycle will make various big-data applications possible, including monitoring and control in the fields of the environment and social infrastructure, and support for human actions.

excellent horizontal scalability and high-speed computing. The new architecture can solve the tradeoff between data volume and computing speed of the conventional parallel distributed systems. Therefore, the expected application area is general big-data analytics, which includes all kinds of applications that collect a huge amount of data that come in a variety of formats from numerous sensors at high speed, analyze the data at very high speed to produce useful information, and provide feedback to the physical world (**Figure 3**).

Target applications cover a broad spectrum of fields, including monitoring and control in the fields of the environment and social infrastructure and support for human actions. Our next step is to select target applications and use cases that will benefit from the new architecture.

3. Theoretical performance

This section describes the results of theoretical calculations under certain conditions regarding a parallel distributed system with the new architecture.

3.1 Data access latency

In the new architecture, an access from a CU to a NAND flash memory occurs when packets including read/write commands or data successively pass through multiple NMs in the 2D network (**Figure 4**(a)). Therefore, data access latency is a function of the number of transfer steps that exist between the CU and the destination NM. It is represented as a straight line with a certain slope and a y-intercept (Figure 4(b)).

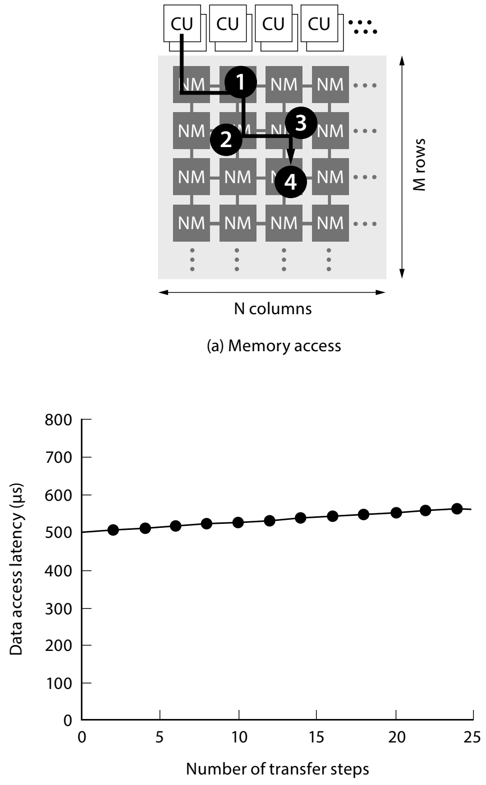
The y-intercept is a time that has nothing to do with the number of transfer steps. Major components of the y-intercept are the access time of NAND flash memory and the controller run-time. The slope of the line represents the time taken to transfer a packet once, which is equal to the packet size divided by the throughput of the 2D network.

The latency of a given access depends on where the target NAND flash memory is located within the 2D network. The average number of transfer steps can be calculated by assuming that the location of the destination NM will be random. The average number of transfer steps increases as a system is scaled. In order to minimize the slope, the 2D network can be interconnected in a torus (ring-shaped) topology. The use of a torus network topology reduces the average number of transfer steps by half.

3.2 Horizontal scalability

Figure 5 shows the relationship between data access latency and the number of computer units. The assumption is that each computer unit consists of 24 rows by 16 columns of NMs. Under the assumption, the average data access latency is hardly affected by the number of computer units. As described in Section 2.2, the new architecture is designed to maintain a roughly constant latency regardless of the system size. Therefore, the new architecture is expected to deliver excellent horizontal scalability.

Data throughput, or system performance, is given



(a) Memory access

(b) Data access latency

Figure 4. Data access latency.
In the new architecture, a target NAND flash memory is accessed by sending instructions or data successively through NMs one step at a time. Therefore, the time taken to access the target NM is a function of the number of transfer steps. The above graph shows the results of calculation assuming that it takes 2.5 μs to transfer data to the step and 500 μs for non-transfer operations.

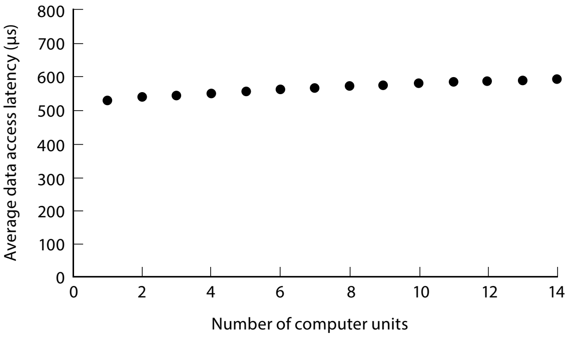


Figure 5. Relationship between data access latency and number of computer units.
The assumption is that a unit consists of 24 rows by 16 columns of NMs. The number of computer units does not affect data access latency significantly.

as a reciprocal of latency. **Figure 6** shows the relationship between system performance and system size. The assumption is that each NM has 64 Gbytes of NAND flash memory. As shown in Figure 6, the performance of a system increases with its size, demonstrating the excellent horizontal scalability of the new architecture.

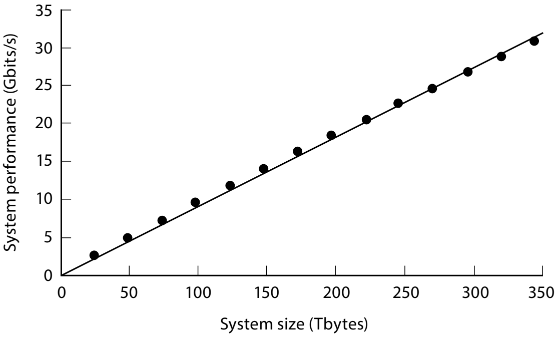


Figure 6. Relationship between system performance and system size.

The above graph shows the results of system throughput calculation assuming that each NM contains 64 Gbytes of NAND flash memory and that each unit incorporates a total of 24 CPUs, each of which has 16 threads, to access data. System performance increases according to the system size, showing good horizontal scalability.

4. Conclusion

We have developed a parallel distributed processing system that delivers ultra-high-speed computing power by combining NAND flash memory and our proprietary new architecture. We will continue research and development of this system to realize an efficient big-data analytics platform that will support the Human Smart Community envisioned by Toshiba.

- *Hadoop is a registered trademark of the Apache Software Foundation in the U.S.A. and other countries.*
- *Ethernet is a registered trademark of Fuji Xerox Co., Ltd.*

KINOSHITA Atsuhiro, Ph.D., Eng.

Senior Specialist. Storage Solution Promotion Department, Storage Products Division, Semiconductor & Storage Products Company. He is an architect involved in the development of new storage solutions.

Note:

This publication may include Toshiba's original estimation, and/or data based on Toshiba's evaluation.