High-speed Data Processing through Ultra-high-speed Data Management Using InfiniBand

● Shoji Yamamoto ● Toshiaki Yamada ● Daisuke Shimabayashi
● Hideo Sarashina

Today, many enterprises seek to link core business operations with the processing of big data on computers, so the need is growing for accurate processing of data that can be instantaneously generated in huge amounts. Fujitsu is responding to this need by applying its extensive experience in mission-critical systems and advanced technologies and by offering ultra-high-speed in-memory data management software. This software uses InfiniBand, which is widely used in the supercomputer field, as the network communications medium. Incorporation of InfiniBand's remote direct memory access (RDMA) function in the network management layer enables this software to achieve response times less than half those of conventional systems and eliminates the need for modifying interfaces to existing applications, thereby resolving compatibility issues. This paper introduces Fujitsu's approach toward providing a highly reliable system that can provide stable system response times and reduce the risk of accumulating or losing large volumes of data at the time of a system fault.

1. Introduction

In Japan, recent advances in hardware performance and transformations in the business model have been promoting a steady increase in financial and credit transactions and other types of data processing events. Data processing systems must therefore handle not only the data generated by conventional core business operations but also the large quantities of data generated by ubiquitous mobile devices and radio-frequency identification (RFID) devices, data required for auditing such as access logs, and data generated by commuter passes, smart meters, etc. In the past, these data would be processed simply as part of an information system, but today, the need has arisen for processing them as transaction data, since treating such data as reliable and useful information can create new value. There is therefore a need to receive and process these data as transaction data to be linked to core business operations and analyzed appropriately as big data, but data accumulation in the back end of the system must be continuous and ongoing without affecting processing in the front end. It is also necessary to cope with burst traffic in which a large number of transactions such as orders are received within a certain period of time.

Fujitsu provides ultra-high-speed in-memory data management software (hereinafter referred to as "this software") that features large-scale and high-reliability data processing in addition to high-speed response and high throughput. This software was designed to reduce response time and improve reliability in accordance with the characteristics of the transformed business model in which core business operations are being linked with big data processing.

This paper describes Fujitsu's approach to achieving high-speed response and high throughput beyond conventional levels while simultaneously achieving a highly reliable system. This approach features the use of InfiniBand,[1),2)] a network communications technology that was also adopted for the K computer, which was developed jointly by RIKEN and Fujitsu.

2. Features and application of this software

2.1 Features

This software features high-speed response, reliability, and extendibility, which enable extremely large-volume transaction processing in a distributed server configuration. A good balance is achieved

among these three key features through vertical integration and optimization of transaction management, network management, and clustering management. This three-component configuration is summarized below.

1) Transaction management

A completely diskless structure is used, resulting in stable processing performance on the microsecond order. All data, including log data, are managed in-memory and are saved across multiple servers using only the network instead of hard disk drives, resulting in guaranteed data processing.

2) Network management

Communications that require high reliability have generally used the Transmission Control Protocol (TCP), but this software uses the User Datagram Protocol (UDP) in conjunction with proprietary technologies such as delivery confirmation to perform network processing that achieves both low latency and high reliability.

3) Clustering management

"Clustering" refers to the process of connecting multiple servers, and "failover" refers to the process of switching to another server if a server fails. Clustering and failover are managed by using three different LANs, one for control, one for work, and one for synchronization. This enables a faulty server to be detected without fail and server switching to be completed within several seconds. Links with hardware control mechanisms such as a Management Board (MMB) and an integrated Remote Management Controller (iRMC) ensure accurate switching.

2.2 Application

Performing extremely large-volume transaction processing in a distributed server configuration is generally referred to as Extreme Transaction Processing (XTP), and software schemes such as in-memory data grids and key-value store (KVS) are commonly used to execute this type of processing. These software schemes provide data access in the form of "key: identifier" and "value: actual value." In addition to this key/value type of access, Fujitsu's ultra-high-speed in-memory data management software also provides message-queuing access that guarantees first-in first-out (FIFO) transaction processing. This enables construction of a highly reliable, event-driven system. In other words, applying a message-queuing function as a system-wide architecture in addition to simple data access enables the construction of a system with high-speed response and high throughput.

3. Software application results to date

This software provides the infrastructure for the "arrowhead" trading system of the Tokyo Stock Exchange launched in January 2010 and has provided stable system operation even as order volume has continued to increase. It is also being used in the online trading operations of Daiwa Securities to provide stable performance and reliability in session management servers, for which the load can increase greatly during large-scale, simultaneous access. This software is also establishing a good record in supporting the infrastructure of the interest rate futures trading system of the Tokyo Financial Exchange. Whatever the system, this software is gaining a reputation for providing high performance as well as ensuring business continuity.[3] Moreover, the code comprising this software was completely developed in-house by Fujitsu, which takes great pride in its high maintainability. The results obtained in troubleshooting problems through close coordination among personnel within Fujitsu have been highly praised.

4. Supporting transformed business model

At Fujitsu, we came to realize that even faster processing and greater reliability were needed to continuously support changes in the transformed business model that have come to affect even social infrastructure systems. In the system developed to support this model, see **Figure 1**, the huge amounts of data generated by linking with core business operations are handled as transaction data, and conventional back-end processing is linked to big data analysis.

The system is able to process vast amounts of data due to the use of a distributed configuration over multiple servers connected via the network. In this configuration, much of the processing time is occupied by network processing, and one way to reduce this time is to speed up this software's in-memory processing logic, but this alone does not sufficiently reduce it. A significant improvement in network processing is also needed. We achieved high-speed response

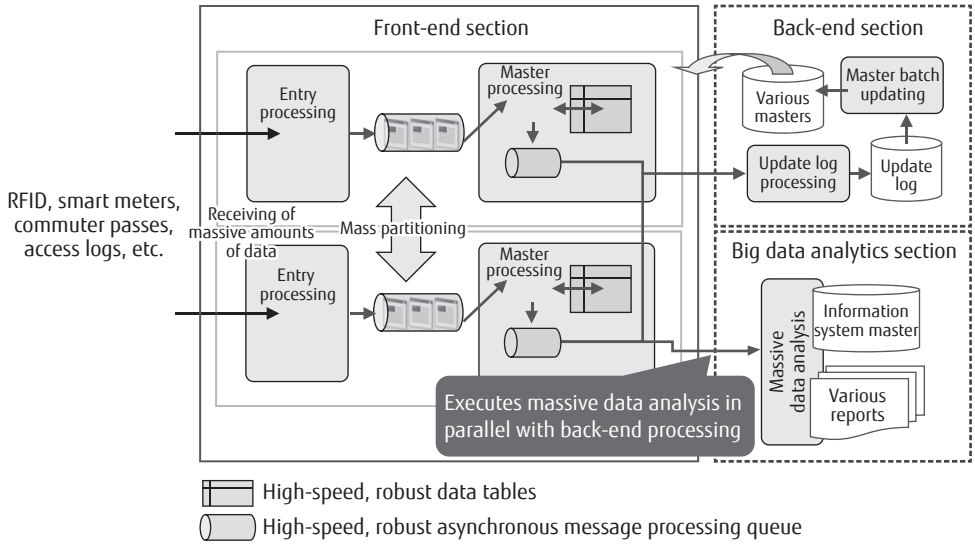


Figure 1
System supporting new business model.

and high reliability by using a proprietary protocol in combination with UDP, despite the fact that UDP has even lower reliability than TCP. We took this approach since network-processing time had to be further reduced to achieve even higher speeds. However, faster processing speeds means an increase in the volume of instantaneously processed transactions, which, in turn, means that a large amount of unprocessed data may be generated even during a very short disruption in business operations, a situation that can lead to system delays. The effect of such a system failure does not stop at the corporate level—it can have a far-reaching impact on society, the economy, and industry. For this reason, it was necessary to ensure business continuity by executing an automatic switchover on detection of an anomaly without the intervention of a human operator; i.e., the system had to be fail-safe.

5. Solution for achieving even higher speeds

It was decided to adopt InfiniBand, which has been used worldwide for network communications owing to its proven track record with the K computer and its support for high-speed processing.

5.1 Reasons for high-speed communications with InfiniBand

While Ethernet is throughput-oriented, InfiniBand is latency-oriented (**Figure 2**).

1) Remote direct memory access (RDMA)

The purpose of RDMA is to reduce processing time by bypassing kernel processing and accessing memory directly in the target node and by reducing the number of memory copies through direct writing by hardware.

2) Packet notification system

Whereas Ethernet allows packets to accumulate for a certain amount of time before processing them, InfiniBand continuously checks for arrived packet by polling and immediately processes them, resulting in low latency and high throughput.

5.2 Features of this software supporting InfiniBand

Using InfiniBand as the network communications medium can reduce latency by about 50% compared with Ethernet. Since burst traffic is bound to occur in a system that handles large volumes of transactions, it is essential to make response times not only shorter but also stable so that system performance and service access are stable. This software has the following structure to maximize InfiniBand performance and stabilize response times (**Figure 3**).

1) Architecture supporting non-uniform memory access (NUMA)[4)]

Intel CPUs have become mainstream in Intel Architecture (IA) servers, and the memory is arranged

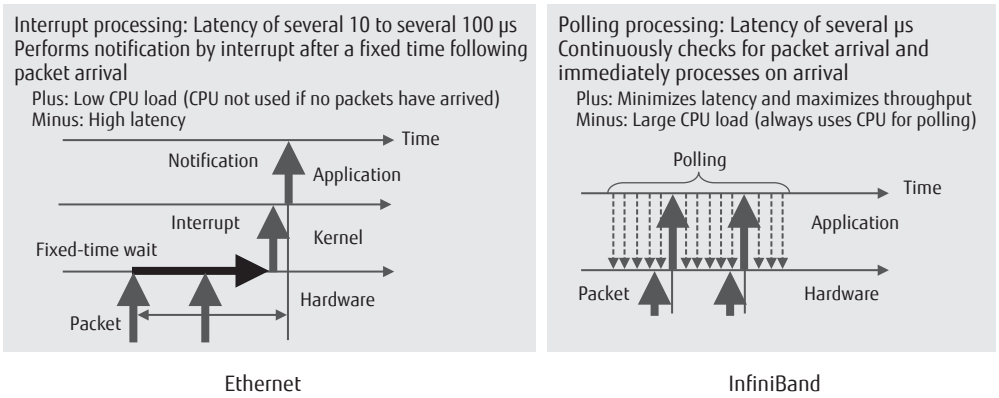


Figure 2
High-speed communications with InfiniBand.

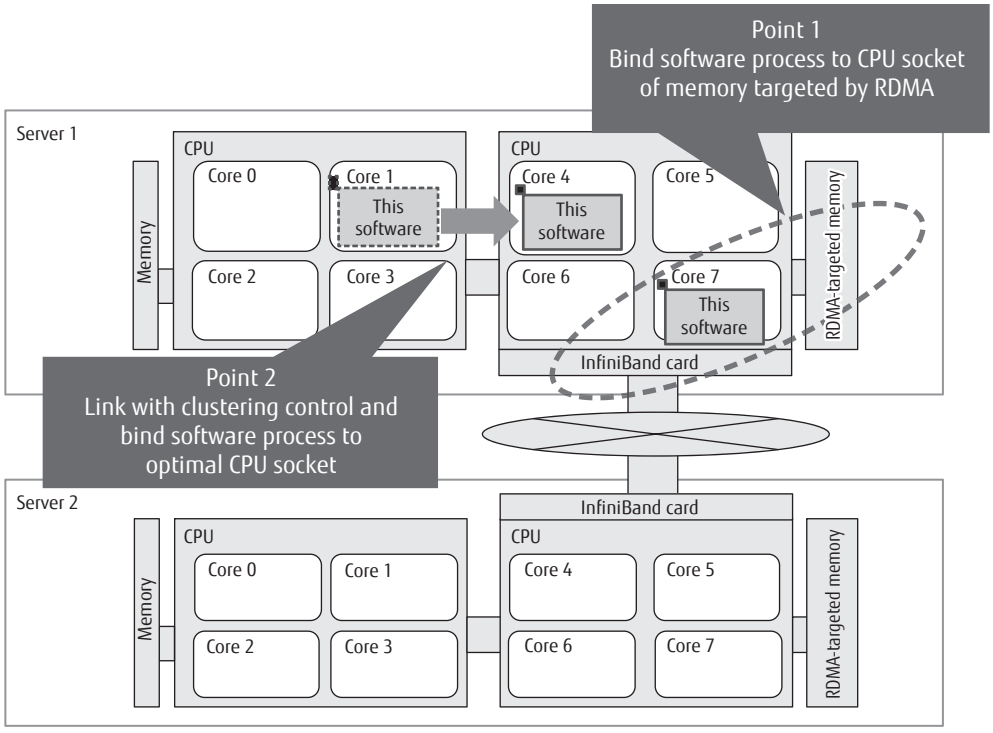
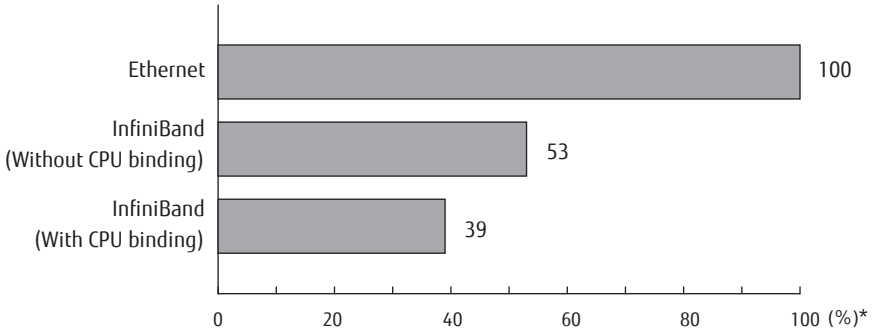


Figure 3
Software architecture supporting NUMA.

in NUMA format. In NUMA, CPU sockets and memory are managed in terms of sets, and, as a result, a fluctuation of several tens of microseconds in response time will occur when accessing memory depending on the location of the access destination (memory) and the access source (process). To suppress this fluctuation in response time, this software provides a function that binds the software process for processing data to the CPU incorporating the memory targeted for RDMA with the aim of maximizing the NUMA features. This software also provides a function for linking with clustering control to dynamically change the process to be bound with the aim of making effective use of the limited number of cores. Using clustering control in this way makes it possible to maintain an optimal state of CPU-process binding not only under normal operating conditions but also in the event of a server malfunction. This results in a level of performance



* Percentages indicate relationship between measured values for Fujitsu-specified model

Figure 4
Relative function response times for this software by type of communications medium.

that takes full advantage of the features of NUMA and the high-speed characteristics of InfiniBand. Relative function response times when accessing memory by this software for different combinations of network communications media and tuning system are shown in **Figure 4**. The 14% difference in function response time between the cases with and without CPU binding shows how CPU binding can contribute to the high stability required by social infrastructure systems.

2) Provision of software-dedicated subnet manager

The weak point in maintaining network reliability and continuity is path switching at the time of a device or cable failure. The mainstream approach is to switch communication paths through network devices. InfiniBand uses network management software, a "subnet manager," to carry out this network management. Subnet managers including the commonly used OpenSM have a function for automatically switching communication paths at the time of a path anomaly. However, this form of automatic control may not match the requirements of a mission-critical system. This is because automatic control, though convenient, tends to generate a considerable amount of device-heartbeat processing at the time of a switching event, which will likely affect the throughput of network processing for normal business operations. This problem has been solved through the provision of a subnet manager dedicated to this software and oriented to mission-critical systems, which have a strong requirement for process transparency. In this software, data is made redundant and managed in memory even for the subnet manager so that stable performance can be provided at the time of switching without having to process large amounts of management data for communications. Additionally, to guarantee stable operation, proprietary technology is provided for selecting two paths and transmitting data along both of them at the same time. This has the effect of completely eliminating delays and interruptions in business operations in the event that one of the two paths fails.

6. Achieving even greater reliability

From a system reliability standpoint, disruption to business operations must be reduced to the utmost limit. To achieve automatic switching even for cases that have conventionally given the operator no other choice but to perform manual switching, Fujitsu has developed a solution to the problem described below.

6.1 Problem[5)]

One problem with conventional cluster systems is called the "split-brain syndrome." It occurs when a network error or other difficulty affects the servers making up a cluster system,[note)] resulting in a state in which the cluster cannot be controlled. In this state, multiple servers may become activated, making it impossible to maintain data consistency. This state is fatal to a cluster system and must be prevented at all costs. A typical cluster system makes use of hardware control

note) A cluster system consists of an active system and standby system. If the active system enters an inoperable state due to a hardware problem or other difficulty, the standby system is made the active system.

mechanisms and shared disks and commences server activation upon verification that a server in the cluster has shut down. If verification cannot be obtained, the system suspends cluster-switching control and enables manual switching by a human operator to avoid the split-brain syndrome. Such human intervention, however, means that someone has to decide which server to keep alive, which could result at the least in several minutes to several tens of minutes of disruption to business operations. Today, with the trend toward high-speed, large-capacity computer systems, a difference of just a few seconds in business downtime can mean a major loss in a customer's business opportunities. For this software, which uses InfiniBand to enable a much larger volume of transactions to be processed compared with existing systems, even a short downtime in business operations can have a devastating effect. For this reason, finding a means of keeping business downtime to an absolute minimum is the key to differentiating not only Fujitsu products but also customers that provides systems using Fujitsu products.

6.2 Solution

This problem is overcome through reliable detection of faulty servers by performing comprehensive heartbeat diagnosis using communication paths with business applications (work LAN) and communication paths to synchronize data among the servers comprising the cluster system (synchronization LAN) in addition to conventional cluster interconnects (control LAN). This software also interacts with hardware control mechanisms built into Fujitsu servers to control power supplies and to precisely and quickly terminate a degenerating server so that business downtime can be kept to an absolute minimum. Nevertheless, there are still cases in which a failure in the communication path to such hardware control mechanisms or other unexpected failures could prevent another server from being reliably shut down, thereby requiring a human operator to intervene and perform manual switching.

To obviate the need for such human intervention, we have developed a means of suppressing the split-brain syndrome and logically isolating a server from the network (**Figure 5**). This is accomplished by taking advantage of the vertical integration of transaction management, network management, and clustering management and by blocking network access on the basis of clustering information. In this way, we have enabled automatic switching and improved business continuity even for situations in which automatic switching was not possible in the past.

7. Conclusion

Fujitsu's ultra-high-speed in-memory data

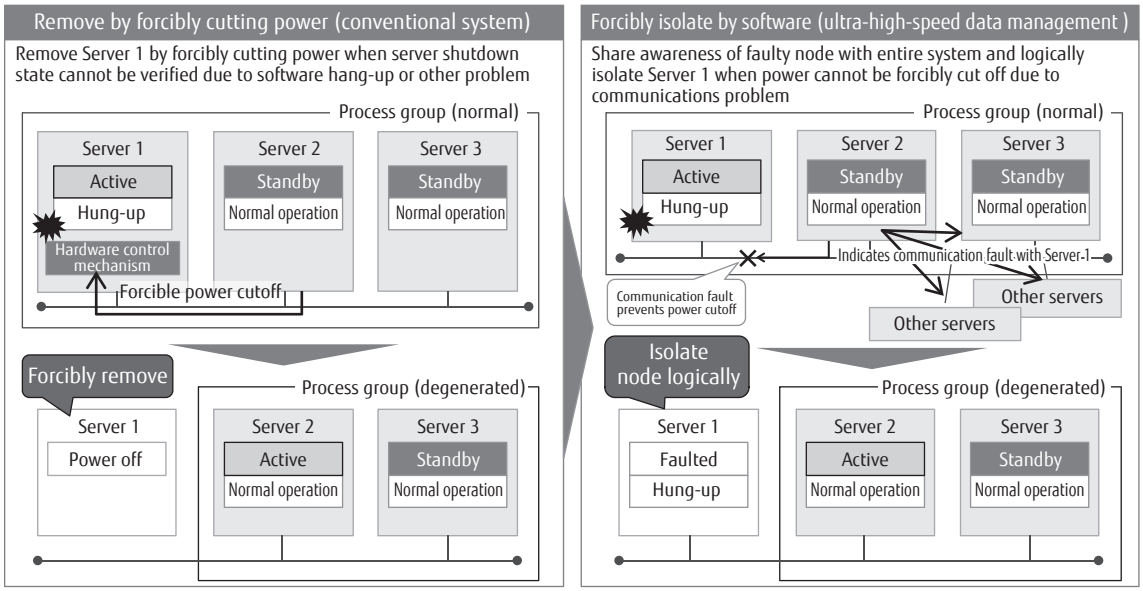


Figure 5
Improving business continuity through node-isolation function.

management software is the first middleware product of this type to be used by social infrastructure systems, which demand high-speed response and high reliability. In this paper, we described how this software can be used in conjunction with InfiniBand technology. Going forward, we plan to aggressively pursue high-speed response and high reliability in general by assessing ICT trends and constructing a software architecture that uses commodity components and maximizes performance for given hardware specifications. At the same time, we intend to pursue new technologies on the product layer to enhance performance without having to modify application interfaces. To date, the successful application of this software has been centered about the finance and securities markets. From here on, we aim to provide this software as a solution tailored to the new business model of "ultra-high-speed processing of massive amounts of data beyond common sense," as has come to be required by new markets in energy provision and telecommunications, and we plan to support our customers' system platforms in this way. We also plan to feed back the technologies that we cultivate with this software to other Fujitsu middleware products so that we can provide our customers with full support in their business operations.

References

1) N. Imamura et al.: New Technologies for Supporting Large Scale e-Business Sites. *FUJITSU*, Vol. 52, No. 4, pp. 338–344 (2001) (in Japanese). *http://img.jp.fujitsu.com/downloads/jp/jmag/vol52-4/paper13.pdf*
2) ALTIMA: Graphic Explanation! What is InfiniBand? (in Japanese). *http://www.altima.co.jp/products/mellanoxtechnologies/whats_infiniband.html*
3) Fujitsu: Case Studies of Introducing Fujitsu Middleware (in Japanese). *http://software.fujitsu.com/jp/middleware/casestudies/*
4) Y. Kimura et al.: Architecture Evaluation Tool for Server Machines: MUSCAT.*FUJITSU*, Vol. 50, No. 4, pp. 202–209 (1999) (in Japanese). *http://img.jp.fujitsu.com/downloads/jp/jmag/vol50-4/paper07.pdf*
5) Invitation to Linux Clustering (2): Structure and issues of failover (in Japanese). *http://www.atmarkit.co.jp/ait/articles/0104/14/news003.html*

Shoji Yamamoto
Fujitsu Ltd.
Mr. Yamamoto is engaged in the development of ultra-high-speed in-memory data management software.

Daisuke Shimabayashi
Fujitsu Ltd.
Mr. Shimabayashi is engaged in the development of ultra-high-speed in-memory data management software.

Toshiaki Yamada
Fujitsu Ltd.
Mr. Yamada is engaged in the development of ultra-high-speed in-memory data management software.

Hideo Sarashina
Fujitsu Ltd.
Mr. Sarashina is engaged in the development of ultra-high-speed in-memory data management software.