Overview of Next-Generation Green Data Center

● Kouichi Kumon

The dramatic growth in the demand for data centers is being accompanied by increases in energy consumption and operating costs. To deal with this problem, we have begun to research and develop a new system architecture called the "Next-Generation Green Data Center." Our aim through this architecture is to demonstrate a level of performance that exceeds what can be achieved by energy savings and performance enhancement at the equipment level. By reexamining equipment roles, vertically integrating and reallocating functions, and integrating the design of power-feeding and cooling facilities, which has traditionally been handled separately, with the design of information devices, we seek to achieve a data center that removes duplicated functions and eliminates waste caused by functional mismatching, thereby raising energy efficiency and reducing operating costs. We follow five key policies—resource pooling, hardware functions as middleware, commodity hardware, unified facility optimization, and integrated operations—to develop technologies for a new type of data center supporting a Human-Centric Intelligent Society.

1. Introduction

As cloud computing expands, huge amounts of data are coming to be distributed over the Internet to be stored, processed, and used, and much information is being created and shared. The Internet data center (IDC) plays a central role in the processing of these massive amounts of data.

The IDC is a large-scale information processing facility consisting of many types of information and communications technology (ICT) devices, but, as capacity continues to be bolstered to cope with growing demands for data processing, the amount of power consumed by IDCs throughout many industries is expected to increase at an alarming rate. The social need to reduce energy consumption has resulted in the power consumed by IDCs becoming a major issue.

To transform the IDC into a power-saving facility, we are proposing and developing a new data center architecture called the "Next-Generation Green Data Center." In this paper, we describe the five basic policies we are following to achieve the Next-Generation Green Data Center including background to their development (**Figure 1**).

1) Resource pooling
2) Hardware functions as middleware
3) Commodity hardware
4) Unified facility optimization
5) Integrated operations

For details on a prototype system for a Next-Generation Green Data Center, we refer the reader to the article titled "New System Architecture for Next-Generation Green Data Centers: Mangrove" included in this issue.

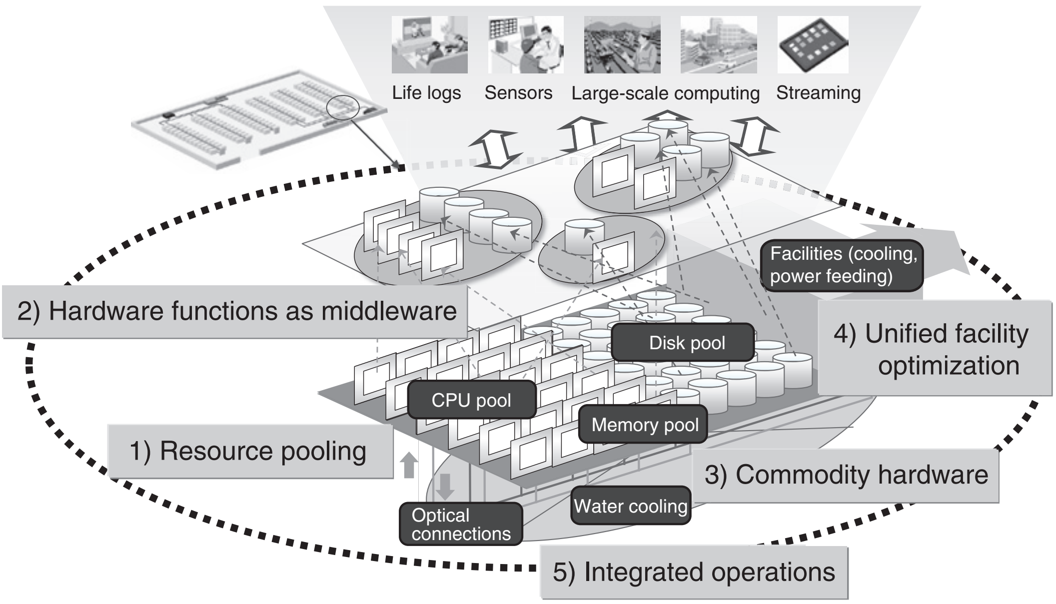

Figure 1
Overview of Next-Generation Green Data Center.

2. Data center issues in the cloud era

Efforts are already being made to develop energy-saving equipment as well as energy-saving data centers as countermeasures to global warming.[1] Looking forward, however, the demand for data centers will continue to grow, and, by 2025, the power consumed by ICT equipment is forecast to reach about 20% of all power generated in Japan and about 15% of all power generated worldwide (**Figure 2**).[2] There is consequently an urgent need to develop technologies that enable energy savings on an even higher level.

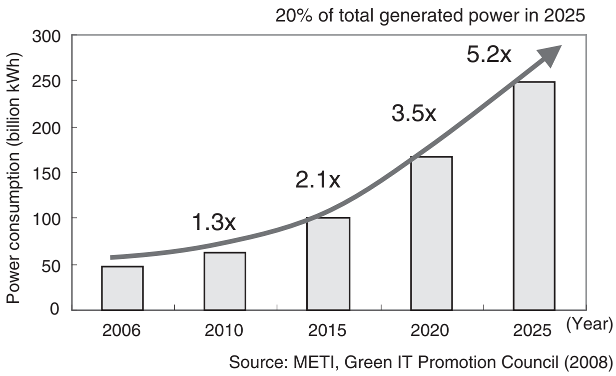

Figure 2
Estimated power consumption of ICT equipment.

In addition to the problem of energy consumption, three types of costs must be considered as data centers are increased in size to expand their processing capacities.

1) Equipment costs
2) Energy costs
3) Management costs

At present, the introduction of high-cost-performance equipment makes it possible to raise performance without increasing cost. Nevertheless, even if equipment costs can be controlled in this way, increasing the number of units of equipment drives up the amount of power consumed overall. Additionally, there is a high probability that the unit price of energy will rise in the years to come as methods of supplying power are reexamined on a global basis. The need for energy-saving measures is consequently expected to become increasingly greater.

Furthermore, in relation to operating and managing costs, increasing the number of units of equipment can be expected to increase the complexity of the system and make management all the more complicated. Against this scenario, the use of existing technologies in their present form may result in an increase in management costs in proportion to the number of equipment units, which would be a major obstacle to improving the processing performance of data centers.

Finding optimal solutions to these data center problems will require the optimization of the entire IDC as a system such as through cooperation between ICT and facility equipment.

In the following section, we reexamine the functional configuration of ICT equipment using a specific example and propose a pooling architecture as a solution to existing problems.

3. Resource pooling

Resource pooling in terms of the Next-Generation Green Data Center means the bringing together of many functional units in an area called a "pool." The idea here is to enable a system that meets certain objectives to be configured by selectively allocating units from the pool as needed and connecting them to other units in the data center via a network. This type of architecture is called "pooling architecture."

In conventional data centers, performance levels and specifications like memory capacity and number of central processing units (CPUs) for each server as well as storage capacity, throughput, network redundancy, and transmission speeds are designed in accordance with the objectives of the target system. Then, on the basis of this design, optimal equipment and devices are procured and the system is constructed.

In contrast to this approach, the designers of an IDC for cloud services seek to minimize variation in equipment type and to use many units of equipment having the same specifications so that services can be easily distributed among multiple servers and storage devices to improve performance and reliability. This type of IDC operates in combination with "scale-out software."

Using many units of the same type of equipment in this way makes it easy to transfer processing to other equipment in the event of equipment failure. However, limiting the system to a relatively narrow range of equipment specifications may prevent an optimal configuration from being achieved for all objectives, and, in some cases, it may prevent some services from being offered or raise the cost of providing them.

With a pooling architecture, however, the unit functions making up the pool are even simpler and smaller than server and storage devices. In the case of servers, for example, their hard disk drives (HDDs) and motherboards made up of CPUs and memory can be selected as constituent elements of the pool. In the following, we use a specific case to show how creating a pool with small units enables a configuration applicable to real applications to be achieved.

We investigate a system architecture for achieving Hadoop, which is a software framework widely used for cloud-oriented processing. Hadoop achieves high performance by enabling many servers to read out and process data from storage in parallel. In general, the HDDs locally connected to a server are used as storage in this scheme, and each server in the system independently reads out and processes data from its HDDs in parallel with other servers. However, as these are locally connected HDDs, the number of HDDs and their capacities in a server are fixed at the time of hardware installation and cannot be freely modified.

In contrast, a configuration that externally connects conventional common-storage equipment to servers makes it possible to overcome the physical capacity limits of each HDD. However, the processing performance limits of the storage equipment make it difficult to provide high data throughput to multiple servers, resulting in a performance bottleneck. Furthermore, as Hadoop processing itself can maintain reliability through multiplexing, the high reliability that can be provided by storage equipment through Redundant Array of Independent Disks (RAID) technology cannot be considered a particular advantage, and the drop in memory capacity caused by data redundancy in this technology is actually a disadvantage. These problems originate in the fact that this division of roles between server equipment and storage equipment is not suitable for Hadoop processing.

For a system running common business applications or a database management system (DBMS), the virtualization functions and enhanced reliability provided by such a storage system are important, but if storage equipment were to be replaced by local HDDs for such applications, both availability and reliability would suffer. At the same time, considering that high CPU performance is not a strong requirement in storage processing, many HDDs can be connected to a single controller, thereby reducing overall costs.

In this way, an optimal system configuration for Hadoop differs from that for common business applications, but, if we compare the internal configurations of server equipment and storage equipment, we find that there is no major difference. Both server equipment and storage equipment consist of two types of units as main constituent elements: a CPU board made up of CPUs and memory, and many HDDs for storing data. The difference here from a hardware perspective is that storage equipment can connect more HDDs than servers, and the difference from a functional perspective is that storage equipment is equipped with firmware for performing storage functions and storage-related operation and management tasks.

Given these differences in the number of connectable HDDs and the use of firmware, the pooling architecture now under development provides for two types of pools: a CPU-board pool consisting of CPUs and memory and an HDD pool consisting of HDDs. We have also prepared an intra-system interconnection capability called disk area network (DAN) to enable flexible connections between these pools to be made and server and storage functions to be dynamically configured.

For pooled hardware, the role of providing specific functions is assigned to middleware running on pooled equipment. We call this mechanism “hardware functions as middleware,” which we will explain next.

4. Hardware functions as middleware

Most present-day ICT equipment is equipped with embedded software called firmware that operates within the equipment to improve operability and functionality. In contrast to ordinary software, firmware is not intended to be installed or executed by users, and its development environment and internal specifications are not publically released. In many cases, an OS designed for embedded devices is used in the firmware operating environment, which means that software development using typical open source software (OSS) is difficult owing to different application programming interfaces (APIs) and other factors. This kind of closed environment limits what developers can do.

In general, the development of middleware on a general-purpose OS enables OSS to be used and contributions from a wide range of software developers to be made while also shortening the technology development time and reducing development costs.

Taking storage as an example, there is a technology called “object storage” that is attracting attention as a means of achieving flexible storage management. In this technology, an upper-layer application conveys to storage the characteristics of the data to be stored so that data placement can be optimized. OSS for achieving object storage is available, and it is already being used to construct large-scale, scalable file systems.

Achieving equipment functions by converting hardware functions to a software library and using a pooling architecture in the above way is called “hardware functions as middleware.”

There is a similar concept called “virtual appliance” in which functions formerly provided as appliance hardware are achieved by software running on commodity hardware.

In the Next-Generation Green Data Center, this concept is extended to configurations of basic equipment like servers and storage units with the aim of generating various types of server and storage functions from the same pool.

5. Commodity hardware

Servers in a cloud-oriented data center are not special equipment—they are treated as just one kind of system component. This is because they have come to be produced as low-cost, high-performance commodities.

There are many cases in which technologies developed for personal computers (PCs) as commodities have been appropriated for use in servers, and the use of Intel CPUs in servers is a good example, but the use of components that are exactly the same is rare. The reason for this is that the reliability requirements of a server differ from those of commodity PCs.

The technologies behind commodity products are progressing rapidly, however, and, in comparisons limited to performance and functionality, many commodity products surpass non-commodity ones. For example, solid state drives (SSDs) based on flash read-only memory (ROM) first came to be used in notebook PCs owing to their shock-resistance properties, but, in terms of random-access performance, they are far superior to high-performance disks designed for server use. Yet, if such high-performance commodity products were to be applied to data center applications, the need for reliability would still have to be addressed.

In the case of commodity products that are not designed for 24-hour continuous operation, there are many that do not have a sufficient lifetime and that vary in quality. Since the conventional approach to securing reliability is to build it up layer-by-layer in the order of components, boards, equipment, and system, it would not be possible to secure system reliability if commodity components with inferior reliability were to be used without taking appropriate measures. Conversely, if efforts were made to raise reliability at the component level, there is always the possibility that the cost of components would dramatically jump. Thus, if the objective is to configure a low-cost, high-reliability system using commodity products, there is a need for separate technology that can handle commodity equipment effectively.

A good example of such equipment-handling technology is RAID, which has been used widely for servers. The essence of RAID is to use redundancy to raise reliability and to improve performance through scale out by using an array of identical products. This basic concept can be widely applied.

When talking about equipment-handling technology, it is important to understand what failure mode means. In the case of RAID, it is assumed that HDD data can either be read normally or cannot be read at all and that failure mode means that data cannot be read. Consequently, if we were to suppose that incorrect reading of data constitutes failure mode, it would not be possible to guarantee correct operation by using RAID schemes like RAID1 and RAID5. Thus, it is also important to consider how the people using equipment functions would understand failure mode. In this case, factors like the conditions of fault occurrence, the frequency of occurrence, whether the failure is accidental or due to wear and tear, whether it is temporary or permanent, and whether it can be completely concealed would come into play. It is important that cost reductions be pursued only after clarifying necessary and sufficient reliability.

In the Next-Generation Green Data Center, we plan to use optical interconnect technology that has already been commoditized to achieve high-bandwidth interconnects spanning intermediate distances (approximately 30 m) as required for implementing a DAN. We also envision the use of low-cost, high-capacity Serial Advanced Technology Attachment (SATA) disks for HDDs, and, to this end, we will study means of ensuring their reliability.

6. Unified facility optimization

In conventional data centers, power-feeding equipment for supplying energy to ICT equipment and cooling equipment for expelling thermal energy discharged from those devices to the outside are treated simply as building facilities. They are not, as a result, required to have functions other than those for performing those tasks. However, given the need to create energy-saving IDCs, reducing power-feeding and cooling losses—which account for 20–50% of the energy consumed by these facilities—has become an important issue.

To reduce the amount of energy consumed by these facilities in the Next-Generation Green Data Center, the methods for connecting both power-feeding equipment and cooling equipment to ICT equipment must be revised, and the number of times that energy is converted must be reduced. Targets of revision in power-feeding interfaces would include AC/DC power feeds, power-feeding voltages, and methods for ensuring reliability during power outages. Making appropriate revisions in these areas should reduce waste in energy consumption. As for cooling interfaces, we can consider modifying the heat carrier for transporting heat generated by chips and enabling high-temperature operation at the component level in ICT devices to reduce the amount of power required for cooling. Here, the high heat-transport efficiency of water makes it a good candidate for use as a heat carrier, and, to make the use of water as a heat carrier possible, we are studying heat-exhaust interfaces different from those currently used in ICT devices and technologies for using generated heat as thermal energy instead of expelling it to the outside as waste heat.

Integrating and optimizing ICT equipment and power-feeding/cooling facilities in this way will bring about a major change in the conventional relationship between equipment and facilities. While the implementation of facilities faces tight constraints since facility operation is based on physical phenomena, the physical positions of ICT processing on equipment may be centralized or distributed as desired through the use of virtualization technology. The environment-related information possessed by facilities can be used to give instructions on the positioning of ICT processing and thereby minimize the amount of energy consumed. In the following section, we describe integrated operations for performing total management including the above control processing.

7. Integrated operations

The basic idea behind this integrated approach is that breaking down functions into elements and reintegrating those elements will serve to eliminate waste and promote flexibility. But to make these effects a reality, we need an operations and management system for optimally positioning a large number of resources and for monitoring and controlling the operation of both data center and facility equipment.

From the perspective of system operators, however, there is a desire to maintain continuity with the present system. Thus, as a first step, we can begin by concealing the pooling architecture and making it look as if a large number of ordinary ICT devices are arranged in arrays. However, there will still be a need for functions like operation monitoring, fault inspecting, and maintenance support.

To achieve facility control, we are preparing a tool to make it easy for managers to investigate configurations that can achieve optimal processing performance. This tool will combine real-time multipoint temperature-sensing technology, IDC-wide high-speed thermal simulation technology, and IDC power-consumption simulation technology.

While research has just begun on achieving optimal operations, we have already obtained results showing that optimally arranging virtual machines on existing hardware on the basis of cost can indeed reduce costs compared to using a standard arrangement of virtual machines. The pooling architecture, in addition to providing flexibility in hardware, can also provide a greater diversity of services at low cost with energy savings.

To perform future processing through optimal arrangements and controls based on policies established under given constraints, we plan to use Fujitsu ICT to the fullest and to develop technologies for implementing a green operating environment.

8. Conclusion

The Next-Generation Green Data Center architecture is aimed at achieving energy and cost savings and flexibility by rethinking the division of roles among existing types of equipment and eliminating waste. The architecture introduced is focused on revising the roles of servers and storage equipment as well as ICT devices and facilities. The end result is a system consisting of two types of pools: a CPU-board pool consisting of CPUs and memory and an HDD pool consisting of HDDs. However, the suitability of breaking down functions to a certain level and reconfiguring them depends on the benefits that can be obtained by such reorganization and on the technologies available, especially those for interconnections. For example, an I/O pool that makes peripheral component interconnect express (PCI-e) connections common among multiple servers can be created using existing technologies if the need is great enough. On the other hand, separating CPU and memory to create separate pools would require a CPU-memory interconnect with a delay under several 10 ns and a throughput of several 100 Gb/s, which would be difficult to achieve at present from the viewpoints of component cost and size.

Looking to the future, we can expect the development of elemental technologies for creating resource pools to continue evolving. By reexamining targets of pooling as needed, we should be able to put these technologies to good use in achieving efficient architectures and creating platforms in support of future ICT.

References

1) M. Otagiri et al.: Trends of Energy Saving in Data Centers and Fujitsu Group's Approach. *FUJITSU Sci. Tech. J.*, Vol. 46, No. 4, pp. 352–358 (2010).
2) Ministry of Economy, Trade and Industry (METI), Commerce and Information Policy Bureau: Green IT Initiative. May 2008. (in Japanese). *http://www.meti.go.jp/committee/materials/downloadfiles/g80520c03j.pdf*

Kouichi Kumon
Fujitsu Laboratories Ltd.
Mr. Kumon is engaged in the research and development of Next-Generation Green Data Centers and high-performance computer systems.