New System Architecture for Next-Generation Green Data Centers: Mangrove

● Takashi Miyoshi ● Kazuichi Oe ● Jun Tanaka
● Tsuyoshi Yamamoto ● Hiroyuki Yamashima

With the aim of configuring next-generation green data centers, we introduce a new system architecture called "Mangrove" that achieves total optimization by vertically integrating servers, storage drives, networks, middleware, and facilities. Mangrove embodies resource pooling and the offloading of control functions to middleware, which are key concepts of the Next-Generation Green Data Center. An information technology (IT) platform based on Mangrove can use resources flexibly and efficiently, perform quick reconfigurations, and improve availability and reliability, thereby lowering costs and saving energy. Mangrove consists of a server/storage architecture that pools hardware resources, storage functions running on the pooled resources via middleware, a scalable data center network, high-speed interfaces supported by low-cost, highly consolidated optical interconnects, and operations and management technology for optimizing virtual machine (VM) placement. This article describes the purpose and features of each of these Mangrove elements.

1. Introduction

The expansion of cloud computing is bringing about a major change in the role of data centers. In addition to supporting existing information technology (IT) systems geared to enterprise needs, data centers are being asked to serve as a computing platform that can accommodate new services now and into the future.

Fujitsu Laboratories is working to optimize the computing platform through vertical integration of IT equipment, operation and management methods, and building facilities with the aim of achieving a Next-Generation Green Data Center (Green IDC). This article introduces an architecture that we call "Mangrove" for vertically integrating the data center based on the Green IDC concepts of resource pooling and middleware-based functions. An IT platform based on Mangrove consists of servers, storage equipment, and network devices as basic constituent elements plus interconnects for connecting those elements and operation and management methods for operating those elements in an integrated manner. The aim with Mangrove is to vertically integrate the data center through unified construction and operation, including the building facilities. In the following sections, we describe the purpose and features of each of the above constituent elements.

2. Server and storage architecture

The server and storage architecture of Mangrove is aimed at achieving data center flexibility based on resource pooling. In a conventional data center, a large number of server and storage enclosures are arranged in rows, and a resource pool is configured by software means.[1] As a consequence, the addition and removal of resources can only be performed

by adding or removing pieces of equipment, which makes it difficult to make changes once an initial configuration has been set. With Mangrove, however, the granularity of the resource pool is made even finer to configure a hardware resource pool at the component level, that is, at the level of central processing units (CPUs), memory devices, and hard disk drives (HDDs). This type of architecture improves the resource utilization rate and enables prompt configuration changes.

A conventional server consists of a motherboard and storage drives, which may be HDDs or solid state drives (SSDs). To construct a resource pool at the component level, we begin by separating motherboards from storage drives to form a motherboard pool and a disk pool, as shown in **Figure 1**. The formation of a disk pool provides a beneficial effect from the viewpoint of hardware implementation. Separating high-heat-generating CPUs from low-heat-generating HDD/SSD devices optimizes the cooling structure, and by enclosing and rack-mounting the motherboard pool and disk pool using optimal methods for each, the layout of a data center with presumed resource pooling can be optimally designed. The formation of a flexible disk pool also has the effect of optimizing storage configurations in accordance with user system requirements. For example, an ordinary general-purpose server can be configured by combining one motherboard and four HDDs, while a storage server can be configured by connecting one motherboard to many disks.

An important factor in realizing a practical disk pool as described above is incorporating a means of interconnecting the motherboard pool with the disk pool. We call the interconnection method used here a disk area network (DAN) to distinguish it from a storage area network (SAN) based on conventional Fibre Channel[2] and to emphasize the connections with disks. Since the connections are to be made to components corresponding to conventional local disks, DAN must have both high-performance and low-cost

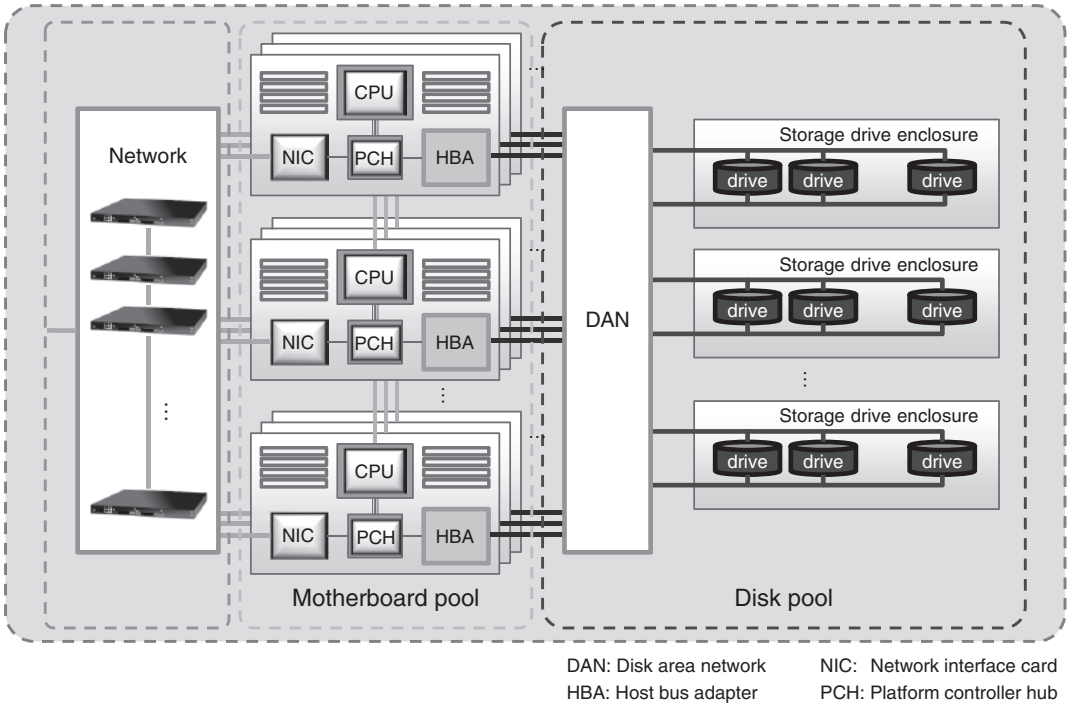

Figure 1
Disk pool system configuration.

features.

The DAN format has the following features in contrast to SAN.

- Topology: There is no need for one disk (target) to be shared by multiple servers (nodes).
- Routing: Simple circuit switches are sufficient since configuration changes are rare.

In the above way, DAN simplifies topology and routing, which makes it easy to achieve a low-cost implementation. A disk pool configured by DAN also enables high-speed and flexible connections of physical disk resources. This makes it possible to propose a disk configuration in accordance with system requirements and to improve the utilization rate of storage-drive resources.

3. Storage system configuration

The conventional approach to constructing a storage system is to decide on the type and number of required units of storage equipment through capacity planning on a case-by-case basis. With Mangrove, however, the user can freely combine any number of CPUs and storage drives and can therefore configure the storage equipment that best fits the target application. What this means is that there is no need to prepare beforehand storage equipment with a configuration that differs from case to case. To achieve this objective, the following issues must be addressed.

1) A mechanism must be developed for extracting in an on-demand manner the resources needed from the motherboard and disk pools to construct the target system.
2) A function must be developed for placing middleware on the constructed system to provide storage functions.

For issue 1), we have developed "MangroveManager," a general-purpose mechanism for constructing not just storage functions but systems as well. For issue 2), we have developed "Akashoubin," a mechanism for achieving Redundant Array of Independent Disks (RAID) functions.

MangroveManager operates as follows.

1) Server construction

MangroveManager extracts resources from each pool in accordance with the type and number of CPUs and storage drives specified by the user and controls DAN to make the necessary connections.

2) OS/middleware installation

MangroveManager installs the OS and middleware software specified by the user on the constructed server.[3)]

Akashoubin is configured as follows (**Figure 2**).

- AsbM (Akashoubin Manager)

Manages multiple AsbRCs ("alive monitoring")

- AsbRC (Akashoubin RAID Controller)

Connects to multiple RAID groups and performs RAID functions

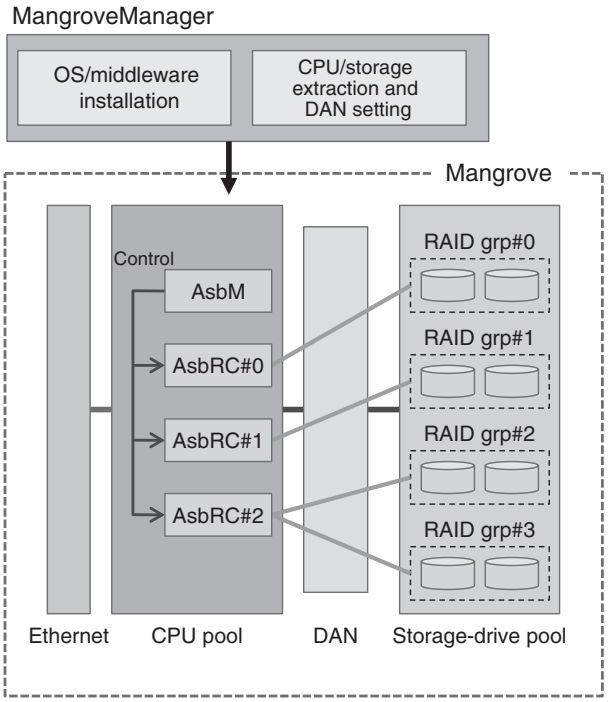

Figure 2
Akashoubin configuration.

- RAID groups

A RAID group can be configured from multiple storage devices extracted from the disk pool.

Akashoubin simplifies operations since it enables faulty CPUs and storage devices to be replaced by other units in the pools.

4. Data center network

In this section, we describe a Layer 2 network (meaning conventional LAN/SAN) as a data center network between systems consisting of Mangrove-based servers and storage drives.

A basic concept of the Green IDC is "optimization of cost performance by vertical integration." To provide a next-generation cloud platform that embodies this concept, the following network-related points take on importance:

- Scalability (on a scale, for example, of about 2000 servers)
- Flat characteristics (uniform delay and throughput across all servers)
- Virtual-network security and reliability
- Low-cost, power-saving features

The network that we are now researching and developing broadly consists of the following three elements. The idea here is to synergistically integrate these elements to achieve a flat and scalable data center network (**Figure 3**).

1) High-density, large-capacity switches

To connect a high-density Mangrove system, we envision a 10GE many-port, high-density switch. This type of switch can offload local CPU functions for performing routing and other tasks to the Mangrove CPU pool. Relegating heavy control plane processing to the CPU pool and light management processing to the switch's local CPU in this way makes for flexible operations in which costs are proportionate to the amount of processing performed.

2) Server-side networking

The cost of switches will ordinarily rise as network functions become more complicated. Here, however, our aim is to raise the functionality and lower the cost of the entire system by requiring the high-density, large-capacity switch described above to have

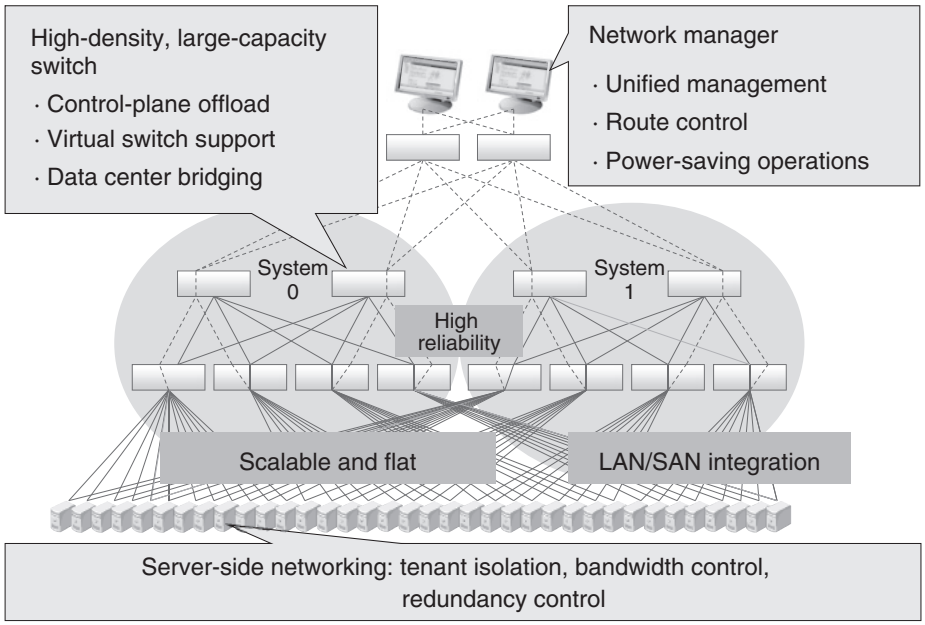

Figure 3
Data center network.

essential transfer functions and a basic level of performance while diverting complicated network functions to the server side. This is called "server-side networking." Specifically, functions like isolation of scalable, virtual networks[4], user-by-user fine-bandwidth control, and redundancy control are to be achieved by the host OS on a server.

3) Network manager

To manage this high-density, large-capacity switch and achieve server-side networking, total layer-2 network management including resource management, topology management, route control, circuit-accommodation design, and visualization will have to be performed. The role of route control will be to reduce the power consumption of the network by putting the switches with a light load into sleep mode while maintaining a certain level of data-transfer performance. In contrast to conventional learning-based Ethernet processing, the network manager will perform total route control to achieve optimal control across the entire system.

5. Optical interconnects

The Green IDC consolidates CPUs and storage drives into functionally separate pools with the aim of achieving a high-density arrangement of servers. High-density servers require a large number of signal connections, underscoring the need for high-bandwidth data connection technology. Bandwidth has been increasing in recent years thanks to the use of optical technology having high transmission characteristics, but the need for very-high-capacity connections in a Green IDC has prompted us to investigate the use of optical technologies that are even further advanced. We here describe optical-interconnect technologies for this purpose (**Figure 4**).

1) Inter-server optical interconnects

With Mangrove, the core technology of the Green IDC, data connections between many enclosures are needed to establish DAN connections, and the key to achieving this is low-cost, small-size interconnects. Although our aim is to achieve a drastic reduction in cost by applying low-priced consumer optical technology to servers, the reliability of this type of technology is not sufficiently high, so we investigated a redundant configuration to satisfy server requirements. We surveyed optical modules having a 10 Gb/s transmission capacity because of their high cost performance and performed a system test on an optical transceiver for PC use as a primary candidate. In this test, we evaluated the error rate at the transmission speed required for SAS (serial attached SCSI) signals and found that this optical transceiver could indeed be applied to servers. We selected

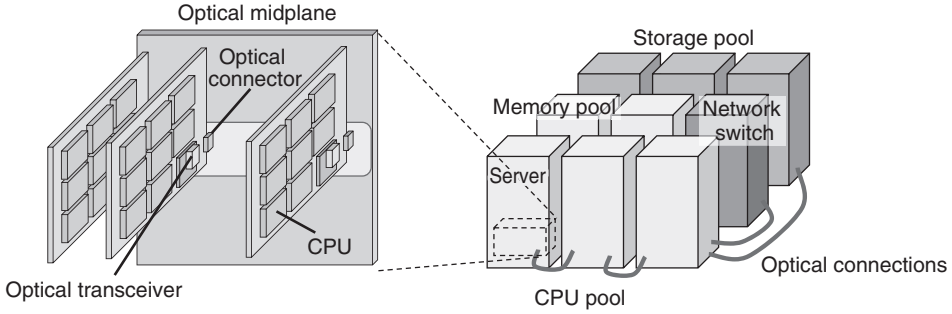

Figure 4
Optical interconnects' server applications.

two ports per module in order to strike a balance between the need for redundancy and the need for compactness. We plan to test the module on a small-scale prototype system.

2) Intra-server optical interconnects

Improving server performance requires high-bandwidth signal connections not just between servers but also within a server (between system boards). There is also a need for extendibility of the input/output interface and memory. This requires extending the internal bus to the outside. These requirements call for high-bandwidth connections in an extremely limited area within the server, which calls for high-speed, high-density (and, of course, low-cost) optical-interconnect technology. Fujitsu Laboratories has developed a high-speed, high-density optical transceiver and a high-density optical midplane (arranging approximately 2000 optical fibers in a compact configuration) as elemental technologies. We constructed a pseudo server for evaluation purposes based on these technologies and tested the transmission of 10 Gb/s optical signals via the midplane and the conversion of peripheral component interconnect express (PCI-e) signals to optical signals. The results of this test demonstrated the feasibility of achieving high-bandwidth optical interconnects within a server.

6. Optimization of VM placement

We are also building and evaluating a framework for virtual machine (VM) placement design, either as specified by a system manager or in a fully automated manner, both of which are considered to be necessary within usage scenarios envisioned for Mangrove. This framework has three advantages over conventional implementations of control.

1) Adoption of a plug-in structure for optimization functions, such as power savings and fault tolerance, enabling each function to be independently enhanced, added, or deleted (**Figure 5**).

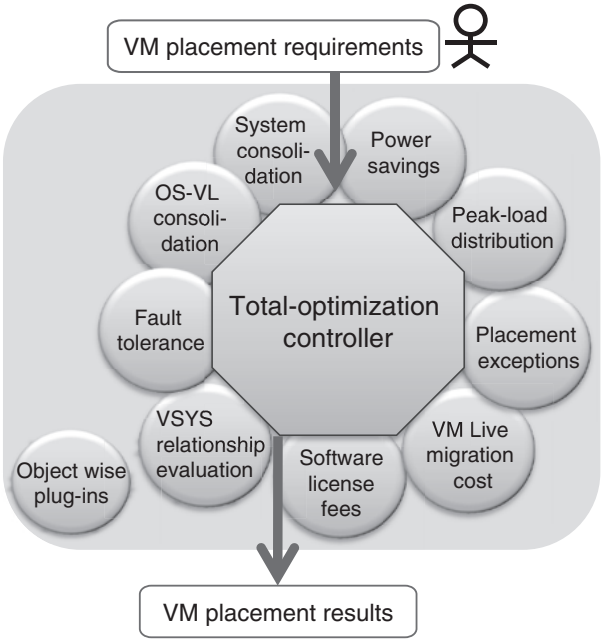

Figure 5
Basic concept of evaluation plug-in.

2) Shared framework among individual optimization targets from different viewpoints, enabling the scores to be maximized or minimized in a uniform manner.
3) Isolation of VM placement design from VM placement execution, enabling any type of VM technology to be supported.

There are a variety of requirements that must be met in infrastructure operations, including reducing power consumption, minimizing software-license fees, reducing network traffic load on core routers, and enhancing fault tolerance. The priorities among requirements, however, can change depending on the business environment and social demands, meaning that an implementation may soon become obsolete. At the same time, the framework should enable temporary suspension of a service for operational reasons, such as scheduled maintenance, patch application, and faulty device replacement.

To address these various requirements in a consistent manner, we have adopted an

evaluation system that calculates a score for each VM placement for each plug-in to be evaluated. For a power consumption plug-in, the system calculates a smaller score for a physical server whose power is ON than for a physical server whose power is OFF. As a result, the physical server whose power is ON would have a lower score and be selected for VM placement.

Thus, the system calculates evaluation scores for individual plug-ins corresponding to different optimization targets and optimizes the grand total of those scores. To avoid ambiguity when calculating base scores for plug-ins and to clarify the practical meaning of the scores, we associate the scores with actual costs. In this way, a variety of optimization goals with different evaluation axes can be considered in terms of minimizing cost, and they can be defined using objective values such as electric-power charges and software-license fees. There is still arbitrariness in calculating scores for imaginary costs, such as for durability against a partial physical server fault and for consumption of infrastructure capacity. We assume that this is a problem in the modeling of the optimization targets and that it can be overcome by refining the model, in other words, by replacing an evaluation plug-in.

Another technical problem is the computational load imposed by VM placement calculation. With an increase in the scale and/or power of the resources, the number of VM placement candidates increases, so the number of calculations becomes overwhelming. Known methods for reducing the amount of computation include the exclusion of duplicate patterns, the use of a cache for calculated values, and heuristic techniques. After examining these methods, we introduced the "equivalent set evaluation" method, which uses the fact that most candidate physical servers for VM placement receive the same evaluation score.

With the approach described above, the calculations for a typical VM placement can be completed within one second in most cases. It has therefore become possible to execute VM-placement design in a real-time, on-demand manner in an infrastructure as a service (IaaS) cloud, instead of statically calculating VM placements in advance.

7. Conclusion

This paper described the constituent elements of the Mangrove architecture, which is aimed at configuring a Next-Generation Green Data Center. We discussed a method for flexibly configuring user systems through a server architecture and storage system that pools hardware resources and proposed a data center network that optimizes cost performance. We also introduced high-speed interfaces made possible through low-cost, highly consolidated optical interconnects and described operations and management technologies for optimizing VM placement.

As a next step, we plan to perform a "proof of concept" study by integrating these constituent elements into a single prototype system. We will also investigate new technologies such as memory pools and object storage.

References

1) H. Yoshida et al.: Service Oriented Platform. *Fujitsu Sci. Tech. J.*, Vol. 46, No. 4, pp. 410–419 (October 2010).
2) T11 Home Page. *http://www.t11.org/index.html*
3) Cobbler Website. *https://fedorahosted.org/cobbler/*
4) K. Onoue et al.: Host-based Logical Isolation Technology for Scalable Cloud Networks. SACSIS, 2011. (in Japanese).

Takashi Miyoshi
Fujitsu Laboratories Ltd.
Mr. Miyoshi is engaged in server architecture research.

Kazuichi Oe
Fujitsu Laboratories Ltd.
Mr. Oe is engaged in storage-related research.

Jun Tanaka
Fujitsu Laboratories Ltd.
Mr. Tanaka is engaged in the research of data center networks.

Tsuyoshi Yamamoto
Fujitsu Laboratories Ltd.
Mr. Yamamoto is engaged in the research of optical interconnects for server use.

Hiroyuki Yamashima
Fujitsu Laboratories Ltd.
Mr. Yamashima is engaged in the research of operations and management methods for cloud data centers.