Scalable Resource Disaggregated Platform That Achieves Diverse and Various Computing Services

YOSHIKAWA Takashi, KAN Masaki, TAKAHASHI Masahiko, MIYAKAWA Shinya, HIDAKA Yoichi, ABE Shinji

Abstract

As usage of the cloud expands, cloud data centers will need to be able to accommodate a wide range of services including not only office applications but also on-premises services and in the future, the Internet of Things (IoT). To meet these needs, it requires the ability to simultaneously handle multiple demands for data storage, networks, numerical analysis, and image processing from various users This paper introduces a Resource Disaggregated Platform that will make it possible to perform computation by allocating devices from a resource pool at the device level and to scale up individual performance and functionality. By using standard conventional hardware and software resources to build these disaggregated computer systems, it is possible to deliver faster, more powerful, and more reliable computing solutions effectively that will meet growing customer demand for performance and flexibility.

Keywords
hardware customize, PCI express, disaggregate, Ethernet, distributed system, reconfigurable system, scale up

1. Introduction

Cloud data centers utilize large numbers of low-cost commodity servers as their hardware platform. At the same time, use cases for cloud services are expanding from conventional office applications to various fields which demand more flexible and more powerful computing performance. For example, IoT (Internet of Things) service application requires frequent data access, sensor and image data analysis, in addition to large-scale big data simulations.

To customize a computer hardware platform for such specific purposes, it is necessary to use expensive devices that are not incorporated in ordinary commodity servers, such as high-speed flash memory storage devices, high-performance accelerators for image processing, and low-latency interconnection for clustering.[1), 2)]

Because the capabilities provided by these devices are not always required, incorporating them in all servers would result in increasing costs and power consumption.

To solve this problem, we developed a Resource Disaggregated Platform capable of delivering versatile computer by incorporating special devices only when necessary.[3)] Hardware resources such as CPU/memory (for computing), storage, networks, and accelerators are all modularized, allowing the hardware to be customized and dynamically optimized for specific tasks by changing the combination of the modules dynamically. The configuration can be software-defined to facilitate maximum levels of functionality and performance with high resource usage and availability.

The basic structure and benefits of the Resource Disaggregated Platform will be discussed in Section 2, the interconnections and resource orchestration technologies will be described in Sections 3 and 4, and the high-performance scalable storage configured using the features of this platform will be explained in Section 5.

2. Outline of the Resource Disaggregated Platform

The architecture of the Resource Disaggregated Platform (hereinafter referred to as RD-PF) is shown in **Fig. 1**. The physical layer, which is the lowest layer, is composed of hardware resources modularized in a device level. The virtualization layer, which lies between the physical layer and the upper logical layer, is composed of a fabric interconnect equipped with hardware virtualization functions that connects the modules in the physical layer so that they operate as the nodes of

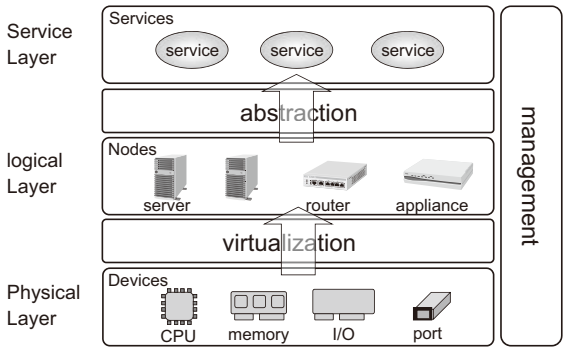

Fig. 1 Architecture of Resource Disaggregated Platform.

servers, routers, and switches. The nodes can be combined in various ways to create a system to execute a wide range of services. At the same time, management software, which interacts across all layers, dynamically allocates resources as required to meet the performance levels demanded by requirements of service applications and policies for reliability management.

The RD-PF offers the following features.

(1) Pin-point scale-up of functionality and performance by adding resources

In a conventional computer system, if there is a performance bottleneck for a single item only - CPU processing, storage, or networks -, there is no way to boost the performance of that item alone; instead, the entire system has to be upgraded, with changes being made to the full set of computers as a unit.

Not only is this an uneconomical way of improving performance, it also results in wasteful power consumption by running resources that are not used. Moreover, when performance is upgraded by scaling out, as is the case in conventional data centers, the addition of new servers does not immediately improve performance; instead, there is some time lag because of the consistency control necessary for the scale-out software.

With the RD-PF, on the other hand, performance can be boosted for a specific item when required simply by adding the necessary resources.

(2) Isolation of compute (CPU/memory) problems from input/output (I/O)

With a conventional computer system, if a server that is a key component of the system configuration fails, the I/O device provided with the server will also be out of service. The opposite is also true: when a network interface card (NIC) installed in a server is damaged, no access is possible from other servers, with the same results as a total server failure.

With the RD-PF, on the other hand, because resources are disaggregated, even when one device starts malfunctioning, the system can continue to operate without interruption by reconnecting a new compute and I/O devices. In other words, hardware issues can be isolated within the relevant resources.

(3) Devices are shared via interconnect

Conventionally, since high-performance devices (GP-GPU and high-speed SSD card, for example) are installed inside the server housing, they cannot be used by other servers. However, the RD-PF shares those resources via interconnect. The devices are connected to a single compute only when necessary, and when no longer required can be disconnected and made available for use by another compute.

In the next section, we will look more closely at the interconnect technology used to connect the resources and the orchestration software that stitches together the hardware and software to deliver the services required by the customer.

3. RD-PF Technology: Interconnect

In order to provide the flexibility to accommodate a wide range of user requirements, while keeping costs to a minimum, it should be possible to utilize commercially available hardware and software resources based on the PCI Express (PCIe) - the de facto standard for computer systems.[4] To achieve this, we developed ExpEther (PCI Express switch over Ethernet) technology. A fabric interconnection for RD-PF, ExpEther combines two technologies, a virtualization of the PCIe switch and highly-reliable transport over Ethernet.

With conventional PCIe switches, the connection distance and the port counts are limited to a few meters and a few dozen ports respectively, and only one compute device (root) can be connected. With ExpEther technology, on the other hand, more than a thousand PCIe resources, regardless of the distinction between compute and device, can be connected over a distance of a few kilometers (**Fig. 2**).

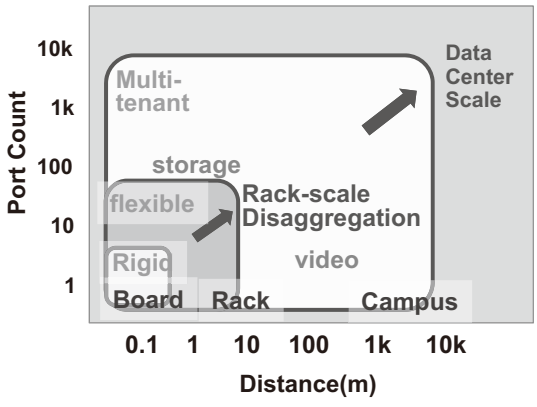

Fig. 2 Increase of connection distance and number of ports by ExpEther.

ExpEther expands the PCIe switch chip virtually over the Ethernet by extracting the chip's internal bus and exchanging it over the Ethernet. In a virtualization, the Ethernet is transparent to the OS and software. Therefore, the ExpEther chips connected via Ethernet are recognized as a standard PCIe switches. As a result commercially available hardware, driver software, and Ethernet switches can be utilized as is without any modification (**Fig. 3**).

Because all the functions of the ExpEther are implemented in a chip (ASIC, FPGA), the access delay of PCIe devices including the ExpEther chip and the Ethernet switch, is suppressed to the microsecond order. This makes it possible to obtain almost the same performance as when a PCIe switch is used. Since data transfer is performed on the Ethernet by using simple memory transfer (DMA), performance significantly exceeds that available from a system (iSCSI for example) that uses TCP/IP, which requires complex transaction-layer-protocol processing. **Fig. 4** shows performance measurement results of random read for high-speed PCI-SSD. "Direct Insertion"

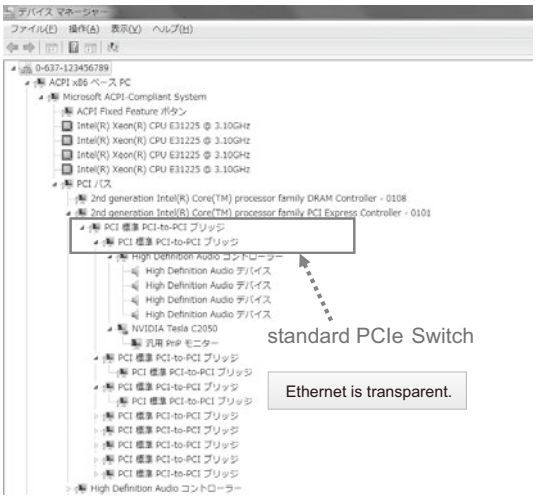

Fig. 3 OS view of PCIe resources connected by ExpEther.

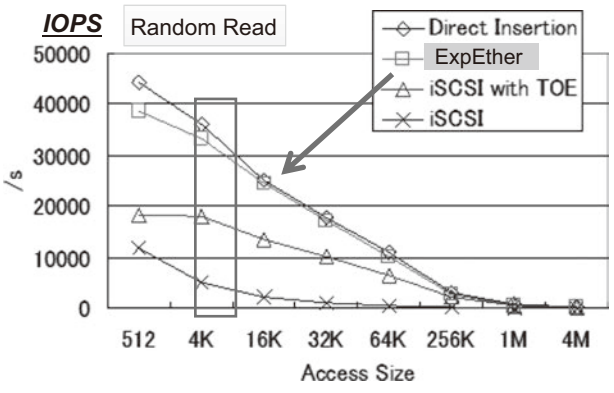

Fig. 4 Performance comparison between ExpEther and various systems.

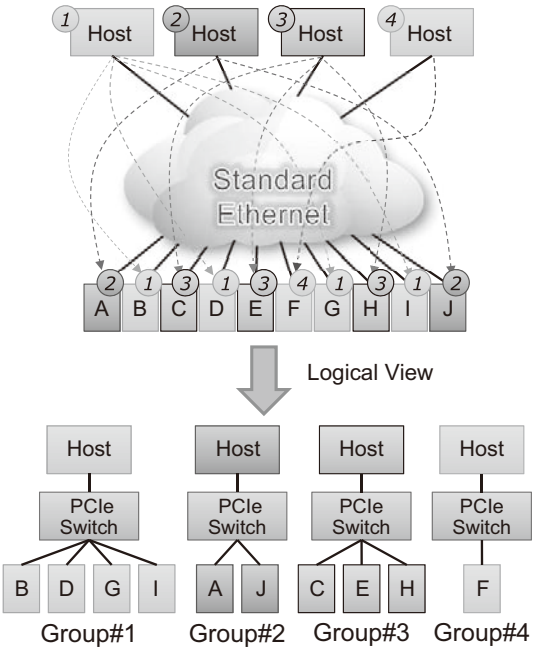

Fig. 5 Automatic formation of PCIe trees by setting group ID.

means that the device is directly inserted in the PCIe slot in the PC's motherboard. Performance degradation in the ExpEther connection for random read access in the 4K size, for example, is insignificant as shown in Fig. 4.

Another advantage of ExpEther is that the computer system configuration can be changed via management software by setting the group ID to the ExpEther chip remotely. This means PCIe logical connections can automatically be formed among the resources connected to ExpEther chips with the same group ID (**Fig. 5**). In other words, as long as resources have been connected to the ExpEther network in advance, hardware configurations can be modified by the resource management software described in the next section without physically inserting devices into or removing them from PCIe slots.

4. RD-PF Technology: Management Software

Resource management software requires orchestration functionality - that is, the ability to dynamically allocate resource on a device basis as necessary to meet the performance requirements of service applications and also maintain high reliability/availability. Orchestration software in data centers must be standards-compliant because this assures a consistent uniform management interface not only for RD-PF but for all the other resource disaggregated systems. To ensure standards compliance, the resource management software implemented in the RD-PF is based on OpenStack, which is the de facto standard in open source cloud management software. OpenStack is modularized according to function. The resource man-

agement function is prototyped on Ironic, which is the module for bare metal machine management functionality (**Fig. 6**).

The orchestration function is needed to facilitate dynamic scale-up and scale-down as well as high availability through fault handling - one of the benefits of the cloud. In a conventional cloud, availability is optimized on a server-by-server basis; however, with CPU-I/O disaggregated architecture, availability is optimized on a device basis. Thus, we have created resource monitoring functionality on a device basis.

When requesting a virtual machine (VM) to use in OpenStack in conventional cloud infrastructure, it is not possible to request a VM with specifications surpassing the performance of the physical servers. With the RD-PF, on the other hand, a user could conceivably request a machine with any specifications, and the devices required to meet that request would be allocated and connected from the device pool. The operations that correspond to the requested specifications are executed automatically. The control interface and functions are implemented as an extended driver of Ironic. This implementation way makes it possible to cope with other disaggregated platform by creating each driver respectively.

As an example of system orchestration, we conducted functional validation using an application that dynamically increases and decreases network interface cards (NICs) while tracking the network load (**Fig. 7**).

Shown in the upper left of Fig. 7 is the packet receiving rate of a computer configured on the RD-PF. With packet receiving handled by the LAN on motherboard (LOM) only, saturation was reached at around 900 Mbps, which is the performance limit of LOM. With the RD-PF, on the other hand, NICs were automatically added when the load approached 900 Mbps, scaling up performance to 1.3 Gbps. The orchestration software kept everything running smoothly by automatically adding NIC devices when packet arrival exceeded the maximum receiving rate of the LOM device.

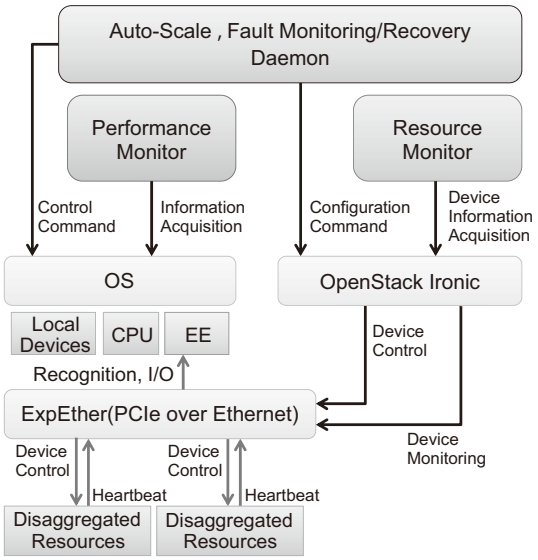

Fig. 6 Architecture of management software.

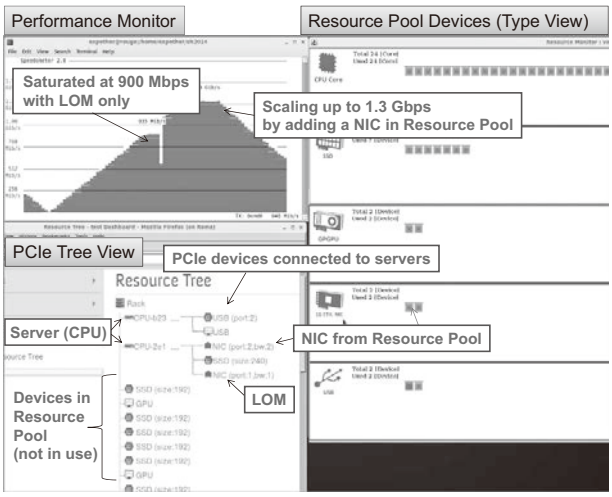

Fig. 7 Example of system orchestration.

5. Storage System Using the RD-PF

By taking advantage of the ability to dynamically add resources, we built a scalable storage system called RDStore (**Fig. 8**).

When the RD-PF is used as a storage system, two benefits can be enjoyed: sharing of storage devices among multiple CPUs, and CPU/device based system expansion. Moreover, because data communication with storage devices is handled by DMA, which has low latency and high bandwidth, higher I/O performance can be obtained that is possible with distributed storage using TCP/IP for data communication.

To extract even higher performance, we turned our attention

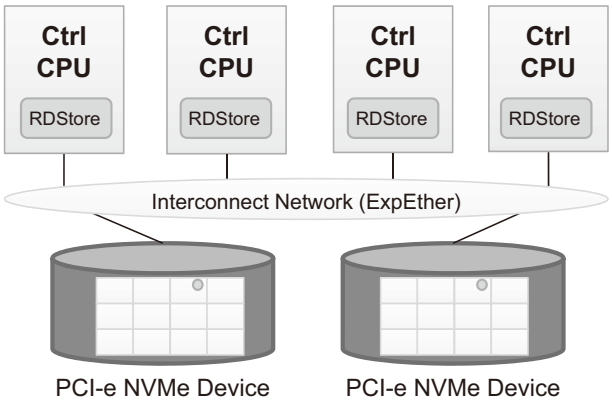

Fig. 8 Implementation example of RDStore.

to the fact that storage devices are placed remotely and shared. Therefore, RDStore has to deal with access delays resulting from the interconnection and access collision of shared resources.

To mask this communication delay, RDStore uses an address resolution method that reduces the occurrence of communications. In addition, exclusive access to shared memory devices implemented in the device can also decrease inter-server communications and improve the load-balance algorithm. We have implemented "fused operation" functionality on the memory devices to achieve exclusive access, without significantly impacting performance. Evaluation of the prototype system showed that response time of "memory device write" operations with shared memory devices was decreased by 40 percent.

RDStore also makes it possible to individually increase hardware resources, including control CPUs and storage devices, in order to scale up performance to meet specific requirements. We confirmed that the I/O performance linearly improved as the number of controlling CPUs was increased under conditions in which a bottleneck existed in the performance of the control CPUs as shown in **Fig. 9**. Conventional distributed storage cannot afford individual performance expansion like this since storage devices (memory) and CPUs are integrated. Therefore, performance is not increased linearly.

Additionally, RDStore can scale up immediately with the addition of servers independent of the amount of data. This is because there is no need to reconfigure the data allocation, and the resource can be used as soon as it is added. This feature is especially effective when data analysis performance needs to be immediately improved in accordance with the changes in the real world, such as in IoT.

In conventional distributed systems (such as Hadoop and distributed NoSQL based systems), which scale out performance by adding servers, it takes time to rebalance the data, so it is difficult to immediately obtain the benefit of the added servers. For example, it takes about 10 minutes with 10 TB of data, more than 1 hour with 50 TB, and more than 7 hours with 300 TB (the same is true when control CPUs (servers) are substituted if a fault occurs in one or more servers). In contrast, RDStore is able to scale up performance the instant the resources are added.

6. Resource Pool System

Photo shows a resource pool system that is in-service at Osaka University. Servers and devices are pooled across six racks, all of which are connected by ExpEther.

Various computers are configured arbitrarily according to user requirements, and they are offered to universities all over Japan through SINET as high-performance computer resources. This is the world's first system that uses commercially available devices and OSs as is while operates resource distributed architecture across multiple racks.

7. Conclusion

Demands on data centers are growing rapidly, and the ability to deliver a wide range of services and handle many and diverse processing requirements is critical. NEC's Resource Disaggregated Platform provides an elegant solution that is enormously flexible and highly cost-effective, making it possible to add required hardware resources as required, while also dynamically scaling performance and functionality with targeted precision.

8. Acknowledgment

We would like to express our gratitude to Professor Shinji

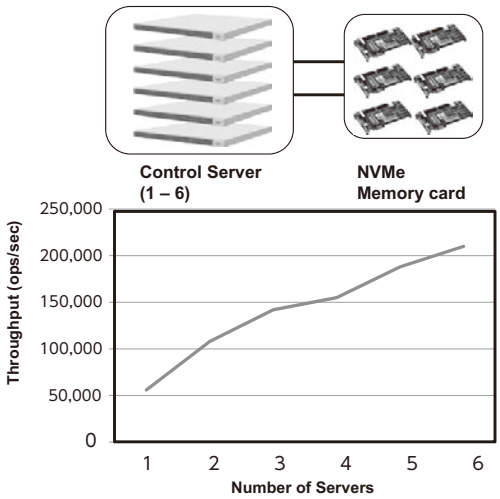

Fig. 9 Scaling up performance by adding control CPUs.

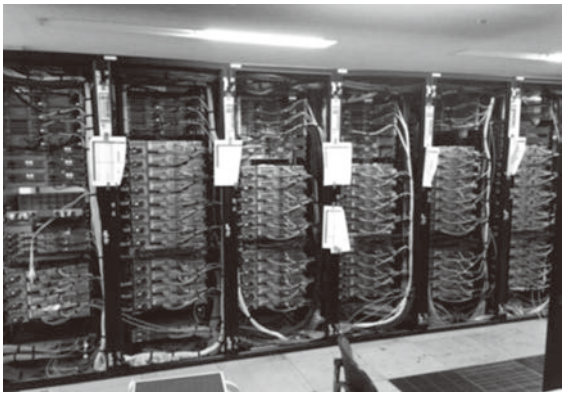

Photo Resource Pool System in service at Osaka University.

Shimojo and Associate Professor Susumu Date of the Cybermedia Center of Osaka University, who provided us with the photo of the Resource Pool System.

* Ethernet is a registered trademark of Fuji Xerox Co., Ltd.
* OpenStack is a registered trademark or trademark of OpenStack Foundation.
* Hadoop is a registered trademark or trademark of The Apache Software Foundation.

Reference

1) Andrew Putnam, et al., "Large Scale Reconfigurable Computing in a Microsoft Datacenter," HotChips26, 2014
2) Jason Cong , et al., "Customizable Domain – Specific Computing," IEEE Design & Test of Computers, March/April, pp.6-14, 2011
3) YOSHIKAWA T. et al., "'Generation Free Platform,' Technology for the Dependable Network Platform," NEC Technical Journal, Vol.1 No.3, July 2006
4) Jun Suzuki, et al., "Express Ether - Ethernet- Based Virtualization Technology for Reconfigurable Hardware Platform," HotInterconnect14, 2006

Authors' Profiles

YOSHIKAWA Takashi
Senior Principal Researcher
Green Platform Research Laboratories

KAN Masaki
Assistant Manager
Green Platform Research Laboratories

TAKAHASHI Masahiko
Principal Researcher
Green Platform Research Laboratories

MIYAKAWA Shinya
Principal Researcher
Green Platform Research Laboratories

HIDAKA Yoichi
Senior Expert
System Devices Division

ABE Shinji
Senior Manager
IT Platform Division

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese

English

Vol.9 No.2 Special Issue on Future Cloud Platforms for ICT Systems

Vol.9 No.2
June, 2015

Special Issue TOP

NEC Information