UKAI Virtual Disk Storage

This article introduces our research into the UKAI*1 storage system, which enables flexible control over the actual data locations of virtual disk images of virtual machines for which location management is difficult.

3.1 Infrastructure with Virtualization as a Prerequisite

The word virtualization no longer has a fresh ring very often these days. Virtualization technology now occupies an important position as a component technology for supporting services. Of course, virtualization technology itself is not new at all, having been used in a range of situations since computers first came into practical use. Even the multitasking technology implemented in current mobile phone OSes was first conceived for the parallel use of expensive physical CPUs among multiple programs via virtualization. The Java language widely used in enterprise services could also be called a type of virtualization technology that utilizes Java virtual machines. Despite this, virtualization has attracted a great deal of attention in recent years, because the prospect of executing the server environments we use every day as virtual machines at an acceptable speed is within sight.

Some network-based services have already been composed of virtual equipment thanks to the maturity of the operation technology of virtual machines. The use of virtualization technology allows for the swift deployment of services, while also making it comparatively simple to expand resources urgently to cope with high loads, and reduce resources when usage drops off, which would be difficult if physical machines were used. Performance is of course inferior to using physical machines directly, but we are now at a point where this gap is evened out by the ease of operation. Virtual machines will never catch up to the performance of physical machines, but based on advances in basic performance, it is not hard to imagine that virtual environments will one day meet the level of quality required by many services. In the same way that applications that monopolize CPU resources have all but disappeared, it is likely that installing service OSes directly on a single piece of hardware will one day be a thing of the past. We will soon be in an age where virtual machines are a fundamental component in service infrastructure, other than a few special applications.

The first step in operating virtual infrastructure efficiently is to construct data centers dedicated to virtual machines. IIJ's Matsue Data Center Park*2 is one of these data centers, and this is used as the infrastructure for the IIJ GIO cloud service. One way to raise the efficiency of a data center is to increase its scale. However, in a country with limited space such as Japan, it

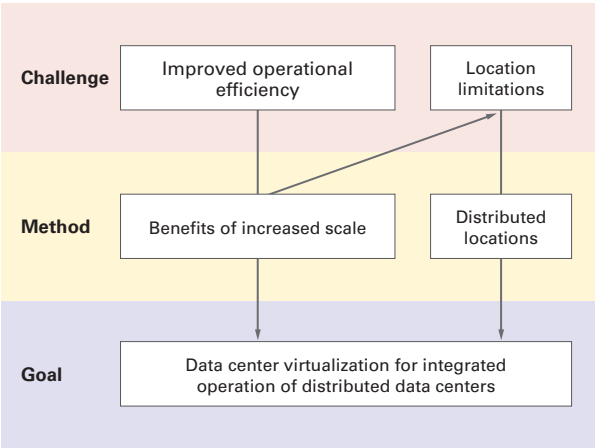

Figure 1: Virtual Data Center Demand and Challenges

*1 Keiichi Shima, "UKAI: Centrally Controllable Distributed Local Storage for Virtual Machine Disk Images", In Proceedings IEEE Globecom 2013 Workshop on Cloud Computing Systems, Networks, and Applications (CCSNA), December 9, 2013.

*2 IIJ's Matsue Data Center Park (http://www.iij.ad.jp/DC/dcpark/) (in Japanese).

is hard to make a significant investment and construct massive data centers when commensurate demand is not immediately anticipated. This leads to the concept of virtual data centers, in which geographically separate medium-scale data centers are operated as a virtual large-scale data center. When configuring infrastructure services across distributed data centers, virtual resources must be arranged and operated efficiently, taking into consideration geographical conditions and network latency (Figure 1). Systems for the flexible location and relocation of virtual resources will be increasingly important in the future.

3.2 The Necessity of Disk Storage for Virtual Machines

Broadly speaking, virtual machines are comprised of three resources. The first is the CPU and memory that form the core of a computer, the second is the network for interconnection between computers, and the last is storage for retaining systems and data. Technology that enables the flexible location and relocation of these resources will contribute greatly to the operation of adaptable virtual infrastructure in the future.

As you are no doubt aware, relocation of the first resource (CPU and memory) is already possible at a practical level. The virtual hardware environment that is a prerequisite for virtual machines can be unified to a certain extent, so as long as virtual machine specifications are decided, the hypervisor that executes them can be operated anywhere. Some virtualization technologies also provide a live migration function that migrates virtual machines to another hypervisor while they are running.

The research and development of technology for relocating the second resource (networks) has been underway for some time, with Software Defined Networking (SDN) technology being one of the strongest candidates. VLAN was a form of network virtualization used as standard up until now, but due to the increase in the scale of data centers, it is nearing the bounds of its capacity. Use of SDN technology enables networks to be sectioned out in virtual machine units, allowing for flexible network configurations that were difficult to implement using VLAN in the past.

This leaves the last resource, storage, but at this point in time there are no good options available. NFS and iSCSI are often used as technologies for providing virtual disks to virtual machines. Considering factors such as the granularity of disk volume management and redundancy of the system in case of disk failures, in most cases it is likely that iSCSI products are selected for large-scale infrastructure. This storage is connected via a network, so as long as the network resources of a virtual machine are correctly relocated, it should still be possible to access its storage in theory. However, the data kept in this storage remains fixed to a physical storage server. Considering the prospect of distributed operation based on the concept of virtual data centers in the future, it is possible that when a suspended virtual machine is restarted, it may restart at a different data center depending on the status of hypervisor resource allocation. When storage resources are allocated to a fixed device somewhere, it affects the performance of virtual machines launched in remote locations. We need to be able to flexibly relocate storage resources in the same way as CPU and network resources, but NFS and iSCSI cannot fulfill this requirement.

At the IIJ Research Laboratory, we are working on the research and development of the UKAI storage system that enables flexible control over the actual data location of virtual disks used with virtual machines, taking into consideration environments for the integrated operation of distributed data centers as a virtual data center.

3.3 UKAI: Location Aware Virtual Disk Storage

UKAI is designed to achieve the following three goals.

1. Administrators are able to control the location of actual data for virtual disks based on operational decisions
2. Functions practically in environments where network latency is not uniform
3. Has redundancy in the event of failures

The actual data of virtual disks is stored in storage nodes distributed across various data centers. When creating a virtual machine, the actual data of the virtual disk allocated to that virtual machine will ideally be located close to the place where that virtual machine is running, whenever possible. It is necessary to reduce latency between the virtual machine itself and the actual data of the virtual disk, and prevent performance degradation in a distributed environment, by situating storage nodes at each data center and specifying the location of the actual data for a virtual disk operationally. However, as mentioned earlier, virtual machines are not always located on the same hypervisor on a permanent basis. Virtual infrastructure is accessed and shared by multiple users. Depending on service usage, there may be cases in which you have no choice but to launch a virtual machine on another hypervisor, or even a hypervisor at another data center at times. The minimum requirement is being able to provide access to virtual disks transparently even in these situations. However, accessing storage resources at another data center dramatically degrades virtual machine performance. UKAI separates the management of virtual disks and actual data, enabling the actual data to be migrated while the virtual disk information remains fixed. This makes it possible to freely change the location of the actual data comprising a virtual disk, while the virtual machine is running, based on operational decisions.

Dealing with network latency is an important issue for distributed environments. It is particularly crucial to estimate the impact of latency when providing services across locations distributed over a wide area, as is the case with virtual data centers. In general, there are two types of latency when operating over a wide area. The first is communication latency within a data center. This applies when multiple storage nodes are in operation, and these nodes are located within the same data center. In this case, latency is minimal (around 1 ms or less), and we can expect it to be uniform. The other type is communication latency between data centers. This varies depending on the relative positions of the data centers, and we can expect latency to have greater fluctuations compared to communications within a data center. UKAI leaves the selection of the location for actual data up to the administrator, and operation is carried out after predicting the latency that will occur. Even when launching a virtual machine in a remote location, the degree of performance degradation can be predicted as long as the latency between that virtual disk and the actual data is quantitatively known. A migration plan is then formulated for the actual data based on this.

The last goal of redundancy goes without saying. You cannot afford to have the whole service shut down due to a node failure somewhere when carrying out the distributed operation of storage nodes. The same applies to situations in which access to service nodes is lost due to network failure. It is of course impossible to protect against all failures, so the best approach is to anticipate a certain number of failures, and add redundancy to cope with them. For UKAI we anticipate disk and network hardware failures on storage nodes. With regard to device failures within data centers (switch failures, etc.), we assume that redundancy for data center equipment is implemented separately. Because virtual disks and actual data are managed separately in UKAI, it is possible for a certain virtual disk to have multiple copies of a piece of actual data. Placing multiple copies within the same data center enables them to be run as network-based mirror disks. Storing copies in other locations also makes it possible to operate them as a disaster recovery measure, although storage access performance would drop.

3.4 Implementing UKAI

We have released a prototype implementation for validation of the UKAI design*3. This implementation takes into consideration the following points.

a. Compatibility with existing hypervisors
b. Elimination of single points of failure through distributed systems
c. Interfacing with cloud controllers

There are two conceivable methods for providing virtual machines with virtual disks. The direct method involves extended implementation of a new virtual disk format on the hypervisor itself. Directly embedding this is likely to lead to better performance, but it is necessary to develop separate implementations for each hypervisor. Another method involves providing virtual disk images as files. The simplest method many hypervisors have is a system in which a file is treated as a single virtual disk image and provided to the virtual machine. UKAI adopts this latter method, constructing virtual disk storage independent of the types of hypervisors.

When designing a distributed system, the greatest concern is how much redundancy should be implemented for when failures occur. As mentioned earlier, UKAI is able to operate multiple storage nodes, so storage node failures can be coped with to a point. Another issue is how to save the information that constitutes a virtual disk (virtual disk metadata) as a distributed system. In our prototype implementation, the Apache ZooKeeper system for distributed coordination support is used.

Figure 2 shows the module configuration of the prototype system. The hypervisor uses files created on a file system that is mounted on a local system as virtual disks. From the hypervisor's perspective, virtual disk images are seen merely as files, so

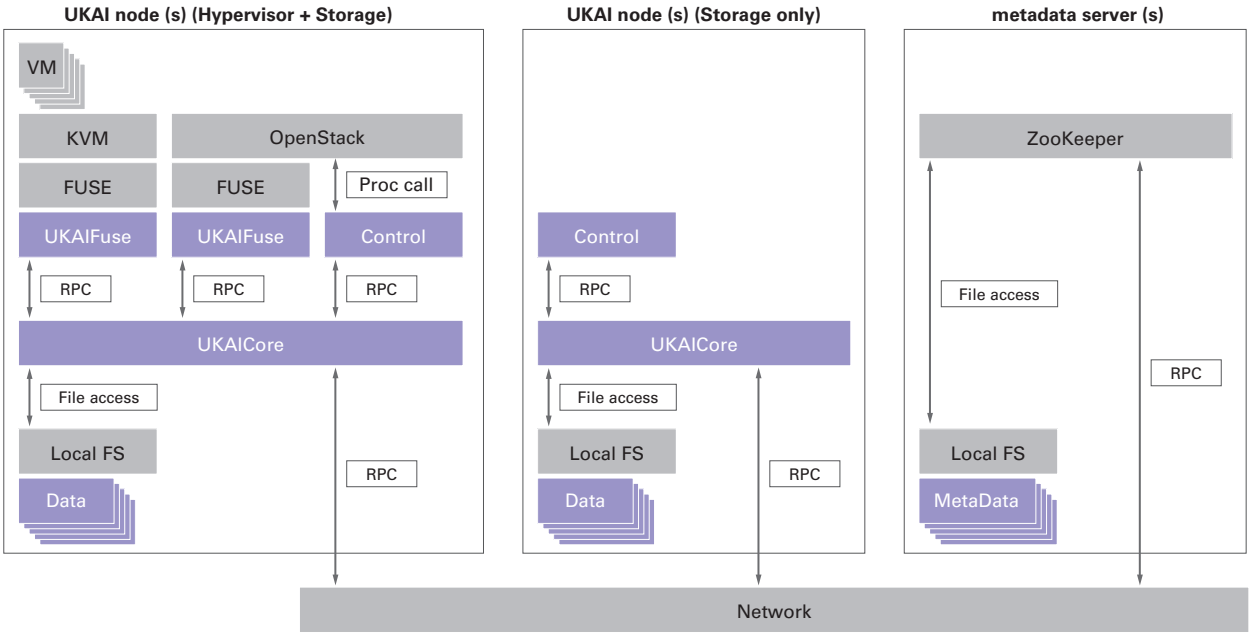

Figure 2: UKAI Prototype System Module Configuration

*3 UKAI: A Location-Aware Distributed Storage Software for Virtual Machine Disk Images (https://github.com/keiichishima/ukai).

they can be operated in the same way as when using local files as virtual disks. UKAI is actually present at the mount point, so file input-output is all intercepted by the UKAI subsystem. In the prototype, FUSE*4 and the fusepy*5 Python binding for it are used to implement the UKAI subsystem. UKAI itself is also implemented using Python.

As their name suggests, virtual disks don't actually exist, and are merely a collection of pointers to the actual data. Figure 3 shows the relationship between virtual disks and the actual data and storage nodes. As this figure indicates, virtual disks are segmented into multiple data blocks, with each data block having pointers to the actual data. Each data block can have pointers to multiple pieces of actual data, to cope with failures on storage nodes.

The relocation of virtual disks is carried out as shown in Figure 4. Let us assume that when a virtual machine is migrated to or restarted on a different hypervisor, a large amount of communication latency is anticipated between it and the storage node containing the actual data. In this case, a copy of the actual data is first created on a storage node close to the hypervisor to

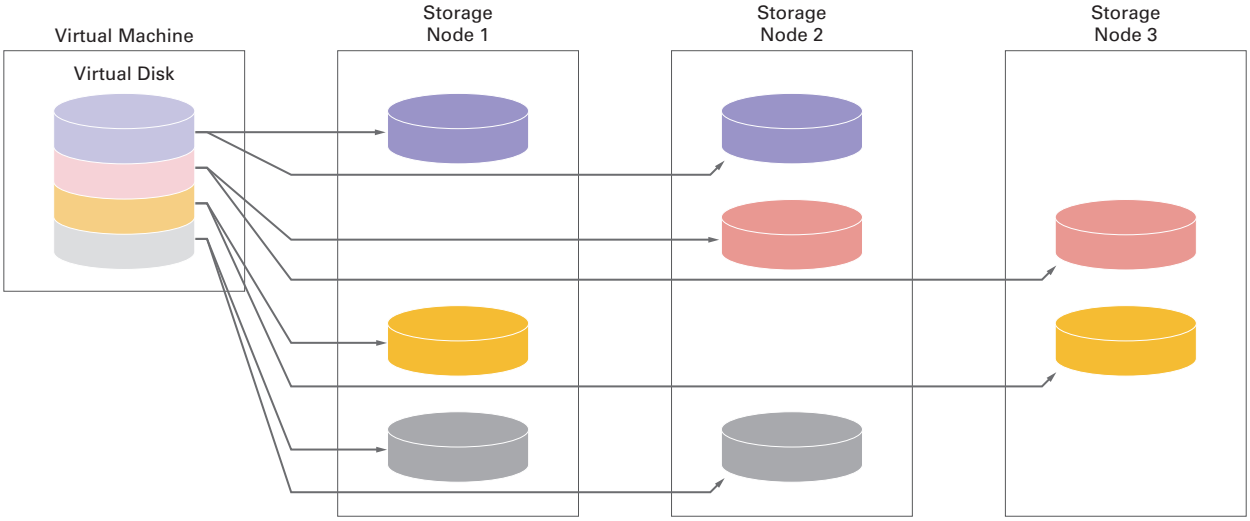

Figure 3: Conceptual Diagram of UKAI Virtual Disk and Actual Data Storage

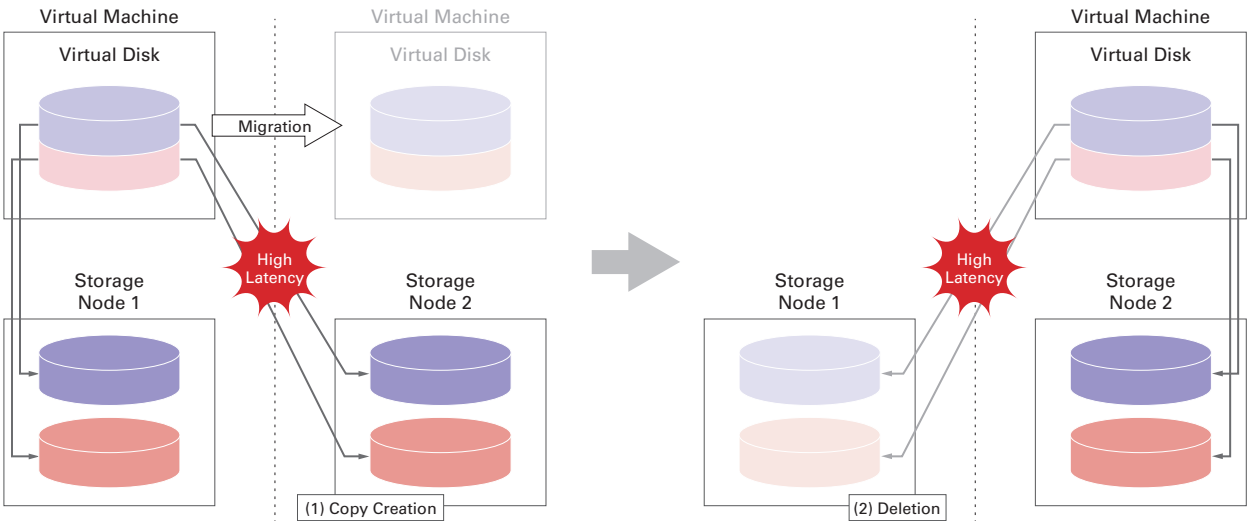

Figure 4: Virtual Disk Relocation

*4 FUSE (http://fuse.sourceforge.net).
*5 fusepy (https://github.com/terencehonles/fusepy/).

which the virtual machine is being migrated. After the virtual machine is migrated, the actual data on the storage node with the higher latency is deleted. Because virtual disks are simply collections of pointers, there is no need for a virtual machine in operation to be aware that a copy of the actual data was created, or that the actual data further away was deleted. Although disk access performance is degraded until relocation is complete, it is possible to optimize the location of the actual data for a virtual disk without stopping the virtual machine.

3.5 Conclusion

As virtual environment performance increases, we believe that many physical machines will be replaced by virtual machines. This will of course have a great impact on services that have continued since the age of physical machines, including server rental, but its true value will be demonstrated when virtual environments that enable the software-based addition or deletion of machines come to be used as components in service delivery models such as SaaS and PaaS. With service infrastructure moving into the cloud, technology for the flexible location and operation of component resources will be essential. Here we have introduced our research on a technology for the relocation of storage, a resource for which location management is difficult. IIJ will continue to pursue technological innovations that serve as infrastructure for stable Internet services.

Keiichi Shima
Senior Researcher, Research Laboratory, IIJ Innovation Institute Inc. Mr. Shima is engaged in the research and development of virtual machine architecture in wide-area distributed computing environments, and technology for the flexible location of virtual resources.