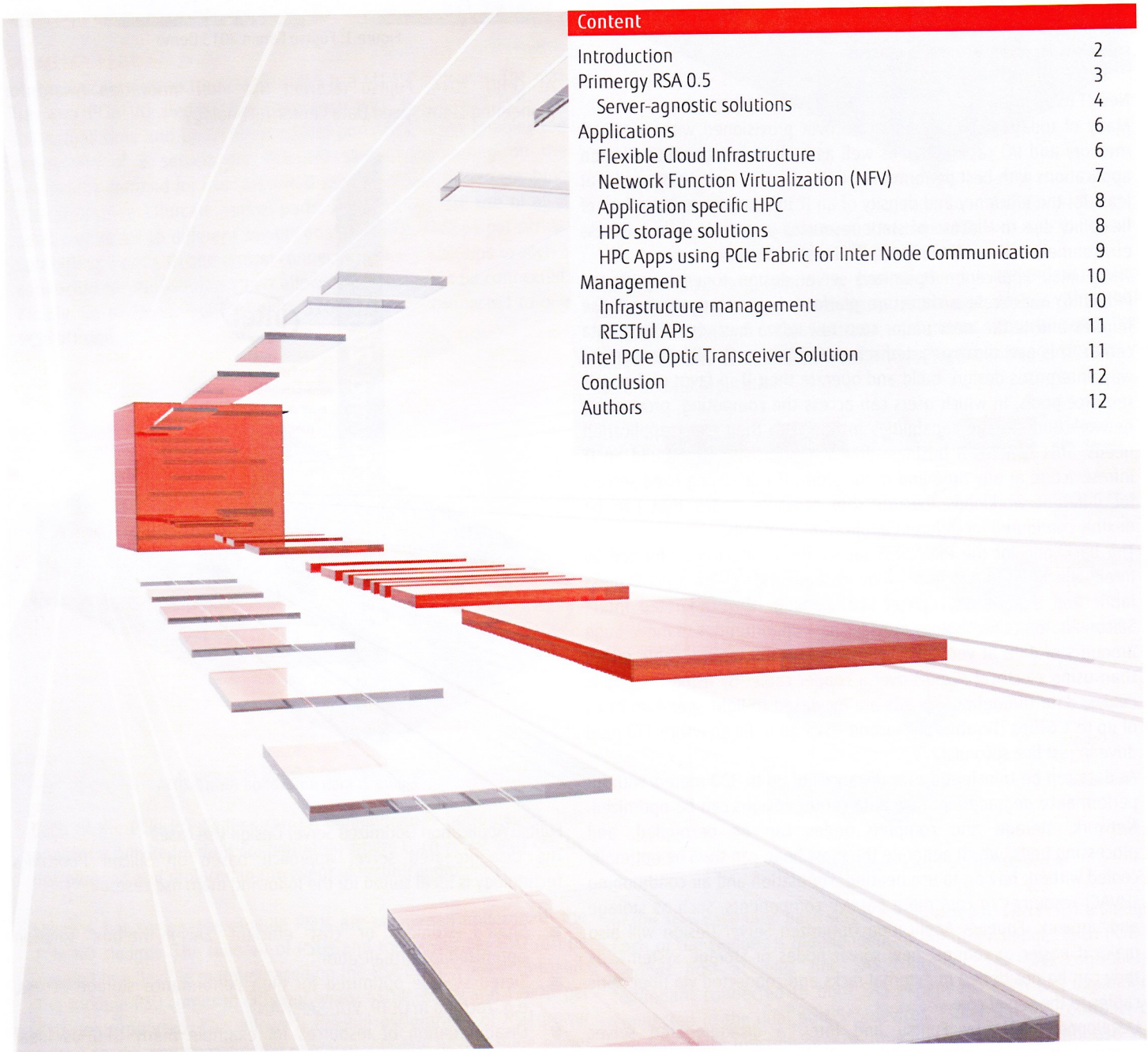
FUJITSU

White Paper
FUJITSU Application-Optimized Server Design

This White Paper provides an overview of Fujitsu's Application Optimized Server Design featuring Intel® Silicon Photonics that will unveil the next major step towards a business-centric data center. This new rack-scale architecture will enable a business-driven approach to (re-) build an IT infrastructure at any time so that users can access the computing, processing, network and storage capabilities in line with their exact application needs.



Content

Introduction

Is your data center keeping pace with modern developments?

The main challenges for many IT organizations that hinder them from making their IT processes more efficient and from focussing on innovations are the speed of the changes as well as the emerging technologies, such as mobile, social and collaboration tools, big data, analytics and cloud. For instance, many companies today have data growth of 60% or more annually – everything from structured databases and text to enormous multimedia files.

However, this unprecedented explosion of data and digital information also results in companies having to deal more intensively with the availability and security of IT services. Almost every business-critical process depends on IT as the failure of important systems in the data center can directly impair the business activity of a company and result in lost sales.

In these times, characterized by rapid changes, new IT infrastructures and architectures are required to fulfill the demands for more simplicity, flexibility and efficiency.

New IT usage models

Many of today's servers tend to be over provisioned with compute, memory and I/O capabilities as well as built-in functionalities to run applications with best performance. Compromises have to be made, at least for the efficiency and density of an IT infrastructure. But a lack of flexibility due to the use of static resources in today's agile business environment can result in missed business opportunities.

The Fujitsu application optimized server design together with the PRIMERGY rack-scale architecture platform helps to overcome these hurdles and is the next major step towards a business-centric data center. This new disaggregated server approach radically changes the way enterprises design, build and operate their IT in favor of dynamic resource pools, in which users can access the computing, processing, network and storage capabilities in line with their exact application needs. This enables a business-driven approach to (re-) build an IT infrastructure at any time and to transform IT from using fixed servers for different workloads towards slim compute nodes that can be flexibly configured for different application demands.

The backbone for the PRIMERGY rack-scale architecture is formed by Intel® Silicon Photonics technology in combination with a PCI-Express fabric and a new MXC optical fiber cabling. The integrated Intel® Silicon Photonics technology provides new opportunities to move huge amounts of data at very high speeds over a thin optical cable rather than using electrical signals over a copper cable. By using fiber-optic cables, data throughput speeds are increased to light speed, at rates of up to 1.6Tbps (Terabits per second, enough to fill an entire 1TB hard drive in just five seconds).

As data can be transferred over distances of up to 300 meters without performance degradation, new data center designs can be optimized. Network, storage and compute nodes can be decoupled, and processing units, which generate the most heat, can then be optimally cooled without having to use heating, ventilation and air conditioning (HVAC) resources to cool more passive components, such as storage and network. Fujitsu's Application Optimized Server Design will also make it easier to connect new server nodes or storage systems, as they can be positioned in external racks and connected via fiber optic cables to the server core.

Developed jointly by Fujitsu and Intel, a disaggregated server approach based on Silicon Photonics Technology was shown for the first time outside the laboratories in November 2013, at the Fujitsu Forum event in Munich, Germany.

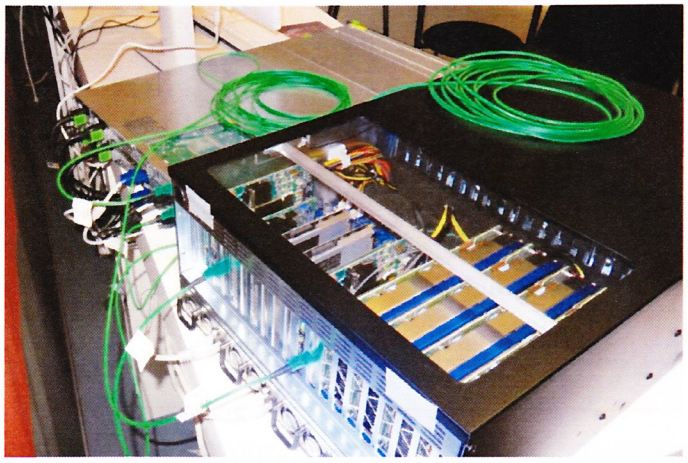


Figure 1: Fujitsu Forum 2013 Demo

At CeBIT 2014, Fujitsu received the Intel Innovation Award for Pioneering Light-Speed Data Center Technology:

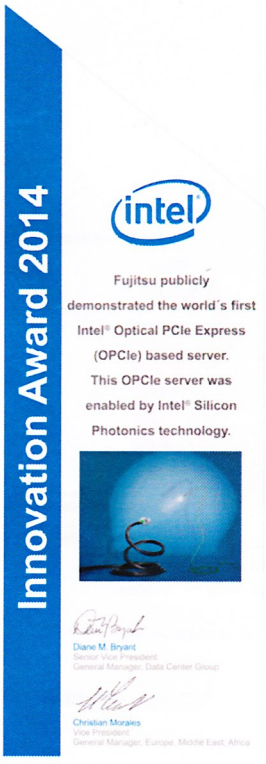


Figure 2: Intel Innovation Award 2014

Fujitsu Application Optimized Server Design Use cases

The disaggregated server approach based on Silicon Photonics Technology is ideal suited for the following main use cases:

- Shared storage, or cost efficient SAN-in-the-box solution optimized for virtualization
- Tiered storage optimized for high performance storage access, e.g. caching in OLTP, VDI, SAP HANA
- Disaggregation of resources for example many GPGPUs ideal suited for HPC environments

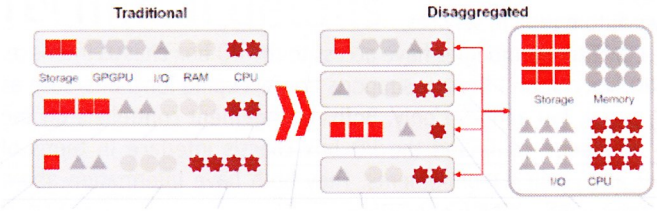


Figure 3: Traditional vs disaggregated server

The above figure shows different traditional rack server systems, each including specific CPU, memory, storage and other IO device resources. The disaggregated approach enables more flexible and cost effective system composition. In the disaggregated approach server systems are built from lean compute nodes with the capability to provision more hardware from disaggregated resource pools.

PRIMERGY RSA 0.5

The rack-scale architecture platform 0.5 is a 4U 19" rack unit. The rear has two PCIe I/O sledges available each with 8 standard PCIe 3.0 full-height slots and up to eight optical ports. The latter enables the connection of 8 servers per PCIe I/O sledge. Depending on the bandwidth demand for the connected server it is possible to combine optical ports or bifurcate optical ports. For example, the use of all 8 ports connected to different servers enables up to 64Gbps per server. Combining 2 ports to one optical connection allows 128Gbps to such a connected server. Splitting a port allows more servers to be connected. Finally, up to 16 servers each with 32Gbps can be connected to one PCIe I/O sledge.

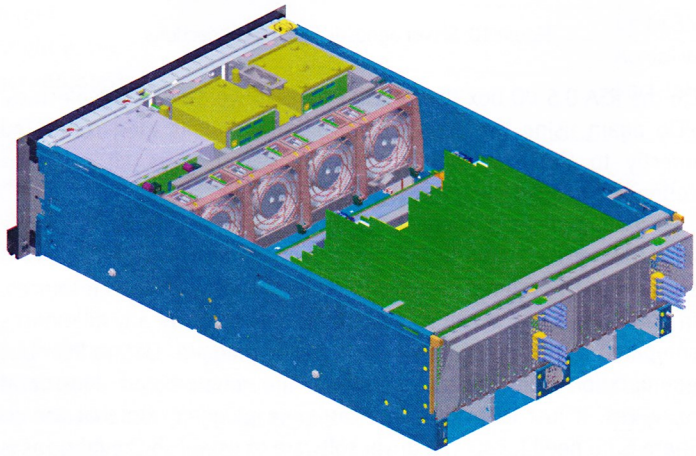


Figure 4: PY RSA 0.5 rear view I/O box

Each of the PCIe I/O sledges can be independently hot-plugged. This enables a fully redundant configuration where each active part can be hot-plugged and replaced in the event of technical problems.

Up to 4 hot swappable module PSUs are positioned underneath the PCIe I/O sledges. The number of PSUs and the power per PSU can be decided in a flexible manner based on the RSA I/O Box power demand. The module PSU comes in different types: 450W, 800W and 1200W per PSU are supported. This allows a flexible assignment of PSU power based on the box demand. For example, GPGPUs have a huge power demand while storage usage cases require much less power. The PSU supports N+1 and N+N redundancy and can be hot-swapped. The fans in the RSA I/O box are redundant. The FANs can cool up to 3600W. The 3600W are derived from 8 x GPGPU + PCIe switch + fan power.

The RSA 0.5 I/O box supports 32 x 2.5" hot-plug storage devices on the front side. This can be HDDs, SSD's with SAS interface (up to 32) or PCIe SFF SSDs (up to 24). Dual port is needed if dual host implementation is required. The PCIe SFF SSDs support dual host connections via 2 x2 PCIe lanes.

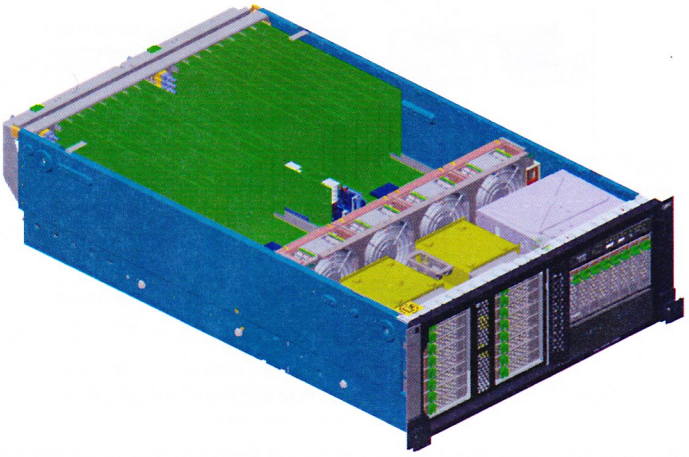


Figure 5: PY RSA 0.5 Front view I/O box

The dual host requirement is part of the full redundant storage connection chain, from hosts down to storage devices. This chain includes redundant PCIe switches, redundant storage controllers, redundant SAS or PCIe Expanders and redundant storage devices via RAID or data replication.

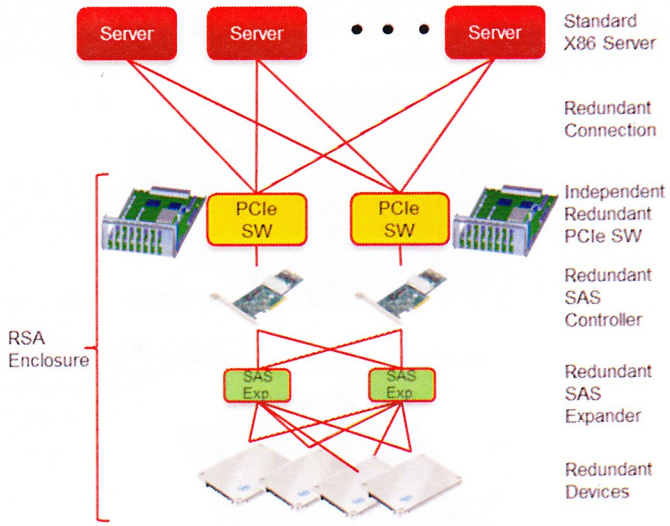


Figure 6: Built-In redundancy with SAS infrastructure

In the above scenario each server is connected to two PCIE I/O sledges. Each of the sledges is independent of the other sledge. Each PCIE I/O sledges host at least one SAS controller. The SAS controller is redundant connected to both SAS expanders. Each SAS expander is connected to the dual port SAS devices. Either the SAS controller can provide RAID functionality or upper layer software can use data replication among devices to achieve redundancy. As a result we have no single point of failure in the storage connection chain.

If more performance e.g. more IOPS and lower latency, is required it is also possible to redundantly connect PCIe SSD SFF devices to the PCIe switch. Up to 24 PCIe SSD SFF devices are supported. Each connected by PCIe 3.0 x2 with 16Gbps.

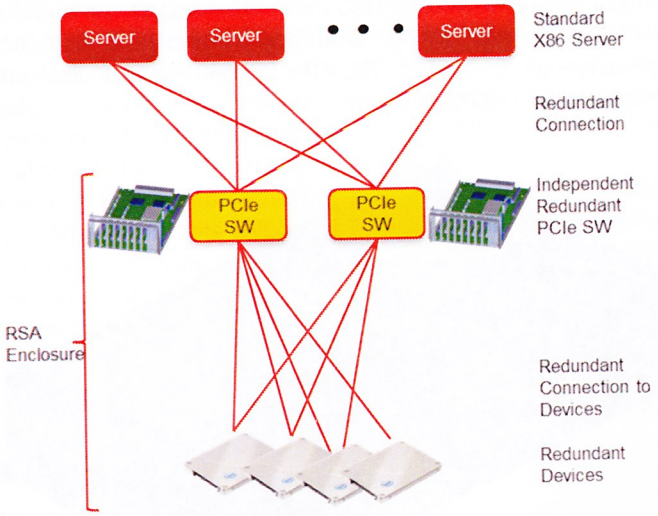


Figure 7: Built-in redundancy with PCIe end device infrastructure

PSU and fans are shared inside the RSA 0.5 I/O box but redundant as well. All midplanes and backplanes are implemented in such a way that failures are not single point of failures (SPF).

The next picture shows a populated motherboard, so-called PCIe I/O sledges. Two of these motherboards can be plugged into the RSA 0.5 I/O Box. Up to eight standard full heights full length PCIe 3.0 cards can be hosted. In addition, four low profile high density PCIe cards can be plugged into standard x16 PCIe 3.0 connectors. The I/O bracket for the optical transceiver is not shown in this picture. Depending on the topology and connectivity requirements either 16 x4 or 8 x8 or 4 x16 MXC connectors can be provided.

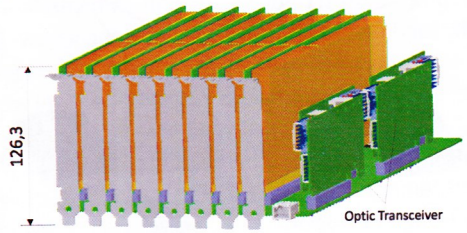


Figure 8: Fully packed PCIe I/O sledge

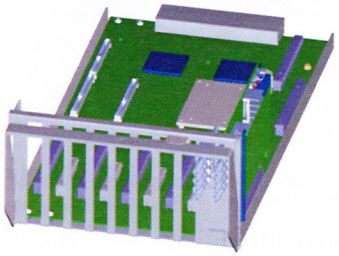


Figure 9: Rack-scale architecture platform sledge

For redundancy reasons each of the two PCIe I/O sledges has its own COMExpress module which runs the management software for the I/O sledge. Assignment and composition of I/O parts, such as PCIe controller or storage devices to connected servers are managed here. A RESTful API is the interface to upper management software and also allows automation of services.
Both COMExpress modules have access to the I/O box chassis interface. The I/O box chassis interface supports housekeeping functions, such as front panel, fan and PSU only. Both management controllers on the PCIe I/O sledge command the I/O box chassis interface in terms of power and cooling requirements. They both have information about PSU, front panel and fan operations.

Server-agnostic solutions

The rack-scale architecture platform 0.5 as part of the "Application Optimized Sever Design" is designed to optimally support any type of PRIMERGY server but also includes support for non-PRIMGERY x86 servers. The concept is that standard PCIe x8 or x16 slots can be used to connect to the RSA 0.5 I/O box. The oPCIe card, a standard form factor low-profile PCIe card, is used for the optical interconnect. This card converts the physical layer from electrical PCIe to optical PCIe.

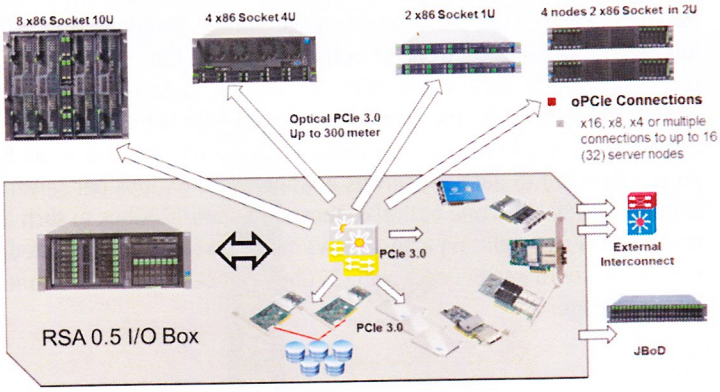


Figure 10: Server agnostic RSA 0.5 connections

On the RSA 0.5 I/O box the optical PCIe is converted back to electrical PCIe again using the oPCIe card. This PCIe link is then connected directly to the PCIe switch. The PCIe switch hosts all resources connected to build the resource pool.
The management station is used to assign resources from the available resource pool to the connected PCIe link. From a connected server view this remote resource looks like directly attached resources. In other words, the server software does not recognize any differences compared to the PCIe card which is locally plugged. During PCIe bus enumeration the remote resources are enumerated like local resources. It just looks like new cards are plugged into that server. There is no need for special driver software to establish the fabric as is needed for Fibre-Channel or Ethernet. Of course, the standard drivers for the remote cards are required e.g. Ethernet driver for Ethernet NIC or MegaRaid driver for storage, as they are required for locally plugged cards. As each server link functions independent of the other, server links from different server types can be connected to the RSA 0.5 I/O box.

To summarize, RSA 0.5 I/O box is agnostic to connected servers because there are no special requirements - other than having a standard PCIe slot available. That is also basically true for 3rd vendor servers but some limitation with respect to higher management support has to be accepted.

Technical details

PRIMERGY RSA05		
Base unit	PY RSA05	
Housing types	Rack	
Storage drive architecture	32x 2.5-inch SAS/SATA or 24x PCIe SFF SSD	
Power supply	Hot-plug	
Mainboard		
Mainboard type	PCIe 3.0 switch mainboard	
High availability	Redundant mainboard design	
Interfaces		
Optical connector	16x MXC connector (128 PCIe 3.0 lanes ~2Tbps send+ receive)	
Management LAN (RJ45)	2 x dedicated management LAN port for chassis management boards	
Slots		
PCI-Express 3.0 x16	8x full height full length slots (75 W)	(4 on each mainboard)
PCI-Express 3.0 x8	8x full height full length slots (225W)	(4 on each mainboard)
Dimensions / weight		
Rack (W x D x H)	482.6 mm (Bezel) / 448 mm (Body) x 736 x 177 mm	
Mounting depth rack	700 mm	
Height unit rack	4 U	
19" rack mount	Yes	
Weight	up to 35 kg	
Weight notes	Actual weight may vary depending on configuration	
Rack integration kit	Rack integration kit as option	
Environmental		
Operating ambient temperature	10 - 35 °C	
Operating relative humidity	10 - 85 % (non-condensing)	
Operating environment	FTS 04230 – Guideline for Data Center (installation specification)	
Electrical values		
Power supply configuration	1-4x 1200 W hot-plug power supply	
Max. output of single power supply	1200 W (94 % efficiency)	
Power supply efficiency	94 % (80 PLUS platinum)	
Hot-plug power supply output	1200 W (94 % efficiency)	
Hot-plug power supply redundancy	Yes	
Rated voltage range	100 V - 240 V	
Rated frequency range	47 Hz - 63 Hz	
Rated current in basic configuration	100 V - 240 V	
Active power (max. configuration)	2400 W	
Active power note	To estimate the power consumption of different configurations use the Power Calculator of the System Architect: http://configurator.ts.fujitsu.com/public/	

Applications

Flexible Cloud Infrastructure

No two clouds look the same.

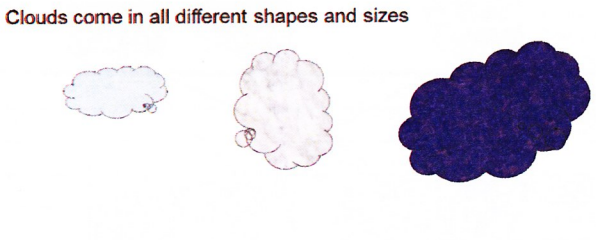


Figure 11: Clouds

What's true in nature is also true in the IT cloud environment.

The trend to bring more types of application into the cloud is accelerating. With more application types living in the cloud, the requirements for the cloud infrastructure diverge from the early homogenous cloud infrastructure.

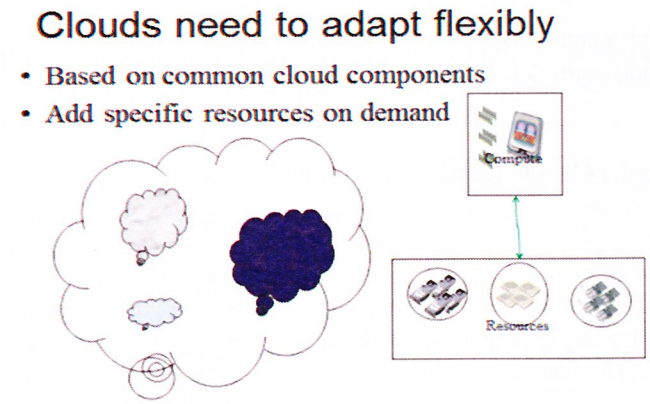


Figure 12: Flexible cloud infrastructure

It is increasingly necessary to adapt the HW to the changing cloud workloads. The time that workloads remain on a specific cloud platform is decreasing. Taking these trends into account it is necessary to provide flexibility in cloud hosting environments. At the same time the core values of the cloud platform need to be preserved. This requires a cloud-aware flexible solution to adapt the HW infrastructure to the requirements of the cloud workload.

A flexible infrastructure for the cloud is needed as follows:

- preserves common compute, memory and networking
- adds HW accelerators such as GPGPU in a flexible way
- attaches the right networking resources

With the transition of application usage models towards the cloud approach more and more applications are available in the cloud. That is true even for enterprise type of applications, either on premise as well as off premise. From the IT infrastructure perspective more flexibility is required to fully support the cloud approach and to support a software-defined infrastructure and the provision of a software-defined data center. The software defined methodology together with well-designed management software provides more flexibility and keeps the management effort low. OpenStack, VMware vCenter or Microsoft System Center are good examples for such management software on top of a flexible infrastructure. As there is no fixed application to IT infrastructure assignment in the cloud, as was common in traditional data centers, the IT infrastructure must become flexible in order to support different application demands. However, there is no one IT infrastructure that fits all applications demands. Each application needs different compute power, memory space, network access and storage. In addition, the demand for such resources depends on the customer usage model for each application. More users connecting to the application or more replicas of applications will change the resource needs and utilization. Lifecycle changes of applications, such as updates or upgrades, change the application resource footprint.

Figure 13: Cloud application examples

A flexible infrastructure solves the "One fits all" dilemma of the traditional basic cloud infrastructure by adapting to the applications needs.

Virtualisation or containers are one way to gain such a flexible infrastructure but this suffers from the fixed hardware platform that is typically provided. This is where the Application Optimized Server Design as part of the Fujitsu Rack-Scale Architecture 0.5 steps in, providing capabilities to deploy a new IT infrastructure from a resource pool on demand. From the customer view, this is quite similar to deploying virtual servers, but with more flexibility in terms of the hardware components that can be composed to build such an IT infrastructure.

There are some basic steps to follow:

1. Collect sever parts from the resource pool and compose your server based on the application demand
2. Set up your virtualization environment on top of the new hardware
    - Install your application on the virtual machine environment.

Alternative

3. Install your application directly on the new composed server environment

    - For applications which will not directly benefit from virtualization, for example they are already designed as scale-out cluster solutions or using simple load balancers to steer loads for replicated applications

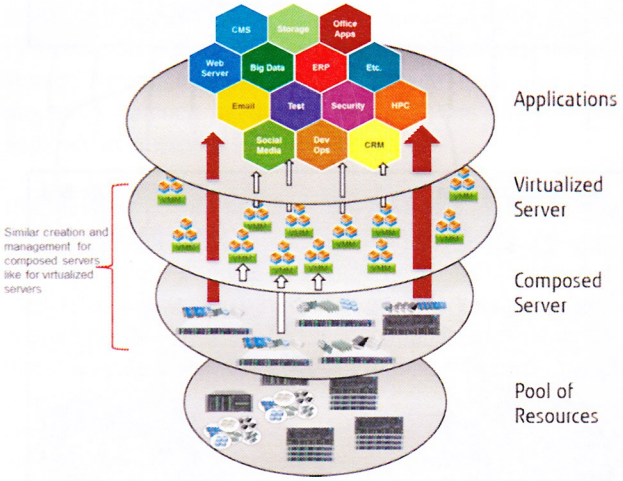


Figure 14: Resource pooling

The server instances are composed from resource pools according to the application profile. These composed servers can be used to support a software virtualization layer and enable further manageability and flexibility.

Network Function Virtualization (NFV)

Internet Service providers are the main drivers of the industrial initiative NFV (Network Function Virtualization). The main objectives of the initiative are:

- meet telecom customers' need for agile, flexible services, for scale-up/-down of services; to support services based on software requiring industry-standard server hardware
- accelerate time-to-market deployment time of network services by more dynamic adaption capability to changing business requirements
- reduce OpEx by simplifying deployment and administration of network services.
- reduce CapEx by using lower cost standard HW/SW instead of special vendor HW/SW implementations for network functions and by eliminating infrastructure overprovisioning supported by pay-as-you-grow deployment models.

NFV offers a new way to design, deploy and manage networking services. It decouples network functions, such as network address translation (NAT), firewalling, intrusion detection, domain name service (DNS), storage services, switching/routing etc. from proprietary hardware appliances. It is designed to consolidate and deliver networking services based on a fully virtualized server, storage and network function infrastructure. It uses standard IT virtualization technologies running on high-volume service, compute, switch and storage hardware to virtualize network functions.

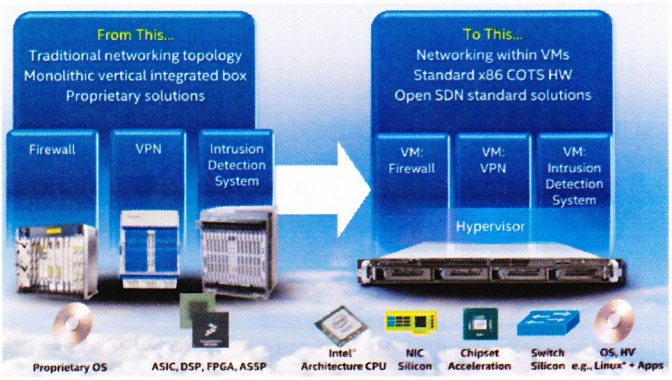


Figure 15: From proprietary to standard IT (Source: IDF2014)

The requirements for the service provider's compute, network and storage HW infrastructure (capacity, bandwidth and latency) can vary depending on the service function requirements to be built. Many service functions, such as video or network switching functions, can significantly benefit from special hardware such as high-performance network controllers, accelerators for encryption and packet switching/forwarding, new storage tiers, such as SSDs, storage class memory etc..

The disaggregation model of RSA0.5 enables hardware separation from compute nodes. The modularity of NFV hardware resource functions is to identify replaceable units. A separate upgrade, scale, lifecycle management of critical IO components becomes possible. The sharing capability from pools of resources across multiple IT system compositions provide improved amortization of expensive components.

ETSI is working on NFV architecture standards with special consideration of server disaggregation. Standard management and orchestration interfaces are under definition to enable service, customer and business-driven orchestration and composition of standard hardware compute, storage and network function capacities.

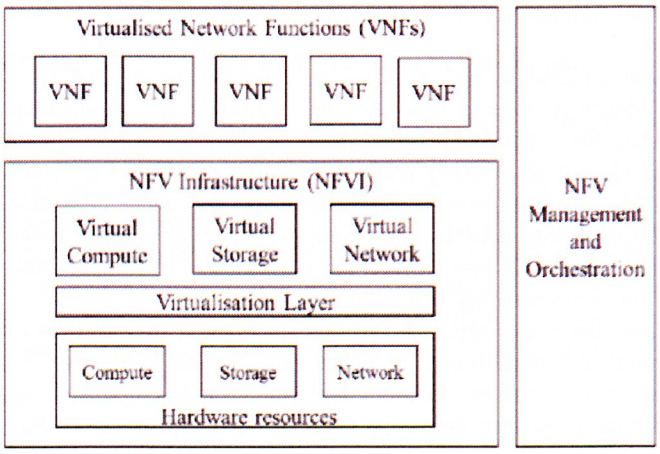


Figure 16: High-level NFV framework (Source: ETSI.org)

The aim of the PRIMERGY RSA approach is to support NFV reference architecture requirements as under definition by ETSI and ONF (https://www.opennetworking.org,http://www.etsi.org). All of this supports an NFV-driven optimized IT infrastructure build at scalable capacity, performance and costs.

Application specific HPC

For a wide range of scientific-engineering, secure-network and secure-storage applications, hardware accelerator technology can help to get highest IT solution performance with low power consumption, low space requirements and low costs. The market requirement for hardware accelerators has thus grown significantly.

Hardware accelerator devices, such as GPGPUs, typically come with highly parallel compute and memory architectures, which can result in excellent application-specific performance. Efficient mapping of processes to a high number of CPU cores is supported by specific software stacks for such devices. Successful application areas for GPGPUs are, for example 3D image reconstruction, molecular dynamics (MD) simulation, high-speed geometrics, geometric algebra models and complex network processing at wire speeds.

Most accelerator devices come with PCIe interface and are packed in typical PCIe device form factors. This makes them basically usable in standard server systems. However, GPGPUs such as the Intel® XEON Phi™, occupy double-width PCIe slots. GPGPUs mostly require longer than standard slot space and multiple times the power of standard PCIe cards. Some applications scale well performance-wise with the number of accelerator devices connected to a single compute node, others apps do not. The way in which accelerators are connected to server systems is also important for the performance/cost ratio of application specific HPC installations. Some workloads can significantly benefit from multiple high bandwidth connections between the accelerator devices and the CPU/memory complex; other, more compute-sensitive workloads, do not need high bandwidth connectivity between CPU/memory and accelerator devices.
Many costly preventive arrangements have to be implemented in a server system in order to afford GPGPU mounting options. It is increasingly difficult for IT vendors to build server system enclosures, which can support the increasing diversity of I/O system configuration requirements at reasonable costs.

The increase in HPC as a service offerings means that the IT industry is required to evolve HPC system designs, from mainly statically configured systems to more flexible (re-)configurable server systems. Making GPGPU, such as HW accelerators assignable from resource pools to compute-nodes will allow on-demand business, administrator and customer-driven system configuration. This avoids the over-provisioning of expensive, costly and power hungry infrastructure operation.
The Fujitsu RSA approach can host up to eight GPGPUs. RSA can support a broad range of uplink options with in-between PCIe switching capability. Different deployment models, from sharing a single accelerator device by multiple hosts, up to eight GPGPU assignments to one compute node, will be supported.

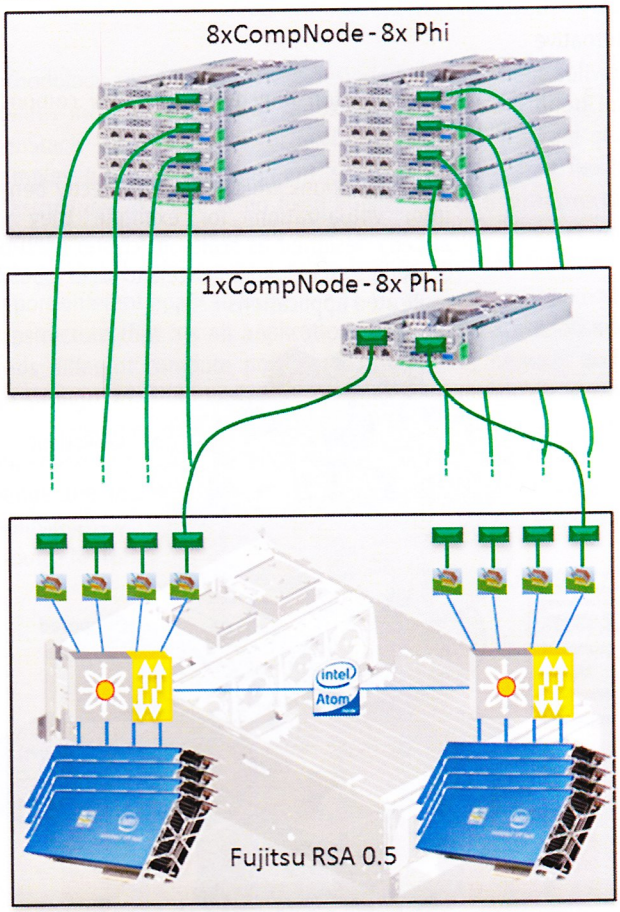


Figure 17: Some RSA 0.5 configuration options with Intel® XEON Phi™

HPC storage solutions

The picture below shows a state-of-the-art reference architecture for HPC applications.

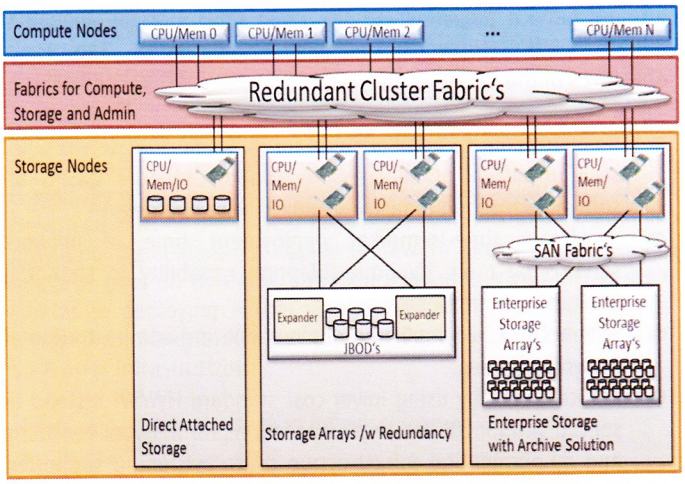


Figure 18: State-of-the-art HPC reference architecture

A low latency, high bandwidth, well-scalable cluster fabric provides inter-CPU node connectivity for high performance inter-compute-node communication. Infiniband is today the most chosen fabric type in HPC applications for low latency intercompute-node communication. The compute nodes may include storage devices not shown in the picture. The data content in compute node local devices can get lost with a failing node, which is why compute node local storage devices are typically used for caching/buffering of remote storage data, only.

In HPC enterprise installations a second independent redundant cluster fabric (Ethernet, Infiniband or Fiberchannel) is used for storage access. Multiple different storage nodes with different properties (latency, bandwidth, reliability) are often concurrently deployed in enterprise HPC installations.

HPC filesystems, such as Lustre, can have very much leverage from having multiple different tier storage nodes deployed at one time. The deployment of different storage node solutions for hot (very frequently accessed) warm, cold or even dark data (Archive) with appropriate hierarchical storage tier management enables data placement on most appropriate storage media. Such different storage pools help to optimize performance while containing costs. Lustre supports an object-based storage model including several abstraction layers for performance and scalability improvements. Files are treated as objects that are locatable through metadata servers. Metadata servers support name space operations, such as file lookups, as creation and file/directory attribute manipulation. The figure below shows a reference architecture of a Lustre multi-tier storage implementation with hierarchical storage management.

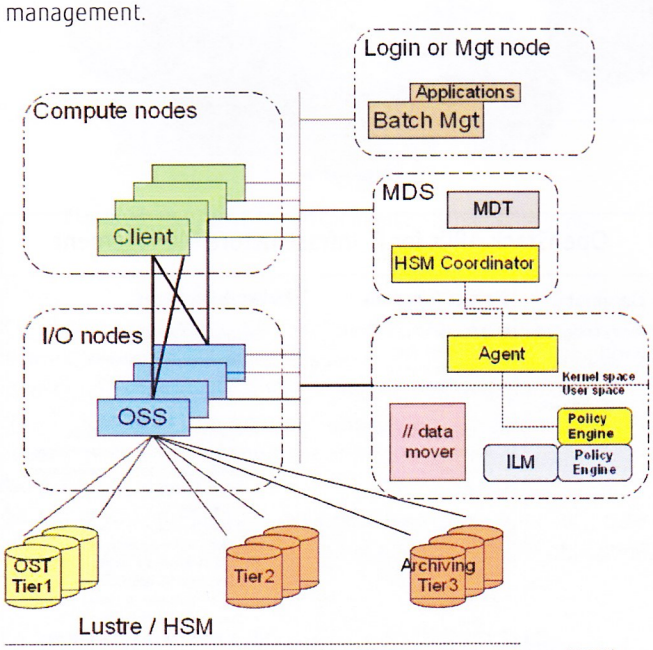


Figure 19: Lustre with Hierarchical Storage Management (HSM)

Upcoming new storage device technologies like storage class memory (SCM), like non NOR/NAND flash based devices in different performance classes, devices with file-size beyond block-size granularity accessibility, will drive the demand for more differentiated Storage TIER support in future HPC deployments.

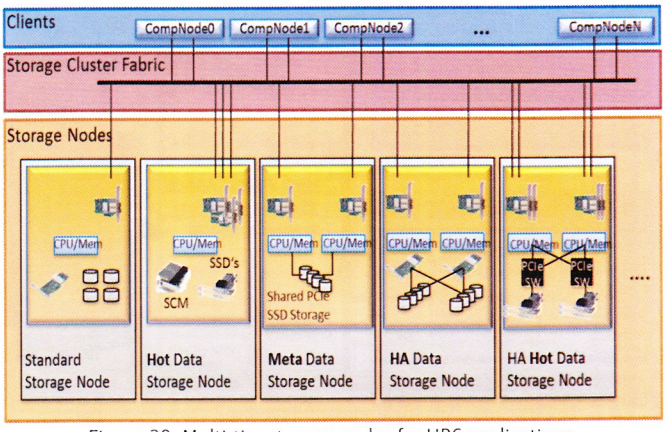


Figure 20: Multi-tier storage nodes for HPC applications

The PRIMERGY RSA 0.5 I/O Box is designed as a storage node building set, supporting different storage tier subsystem build. Different performance/capacity class storage subsystem solutions, from global sharable Storage Class Memory (SCM) to High Available (HA) SSD resource pools can be made accessible to one or multiple compute nodes. Exclusive assignment of storage nodes to single compute-nodes or even subsystem build for sharing from multiple compute nodes with multi path access can be supported. PCIe Endpoint Sharing e.g. for cost saving reasons but also high available (HA) storage subsystems built for HPC applications, will be supported.

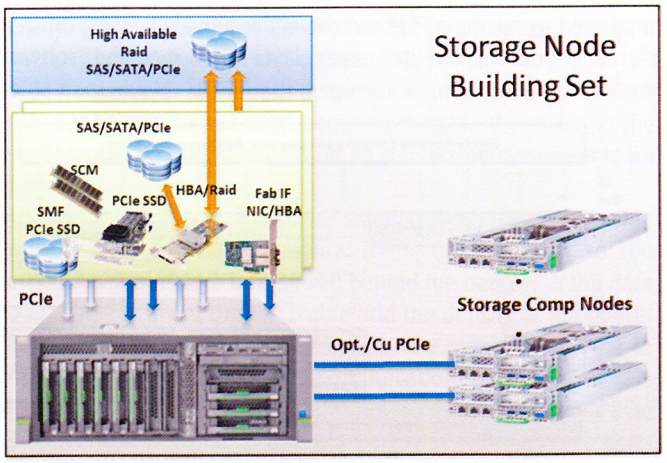


Figure 21: Building set for different tier HPC storage nodes

HPC Apps using PCIe Fabric for Inter Node Communication

The PCIe switch fabric implementation of the PRIMERGY RSA 0.5 I/O box will support standard network protocol based interhost port communication. Up to rack level, lowest latency, high bandwidth inter node communication is possible by having PCIe Fabric built-in support for standard API based network communication. The high bandwidth and low latency IP traffic tunnelling capability over PCIe physical layer enables significant CAPEX, OPEX and space and power savings , which is of special interest in HPC applications.

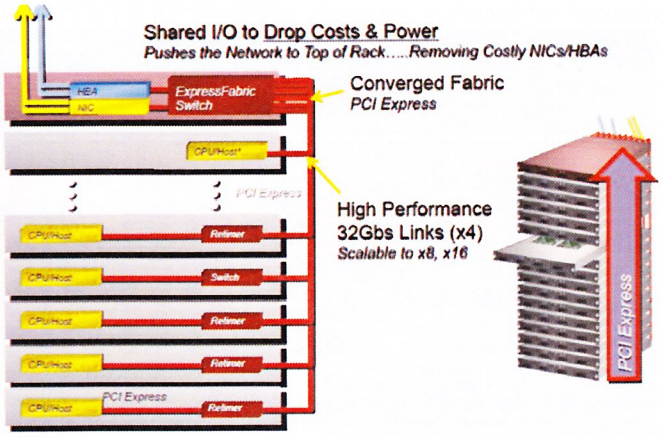


Figure 22: PCIe converged up to rack-scale fabric (Source: plxtech.com)

The OFED™ (OpenFabrics Enterprise Distribution) software stack APIs will be supported for Inter Node communication over PCIe Fabric. A broad range particularly of HPC applications can be supported and will run agnostic to the PCIe type of inter node fabric used as physical transport layer below.

The OFA (Open Fabrics Alliance) which provides the OFED software stack is an open source software alliance that focuses on developing, testing and licensing high performance networking software for servers and storage systems. The figure below shows an overview of OFED software stack functionality, which can be supported by having a PCIe switch vendor provided device driver in place, interfacing to the upper layer OFA provided software modules.

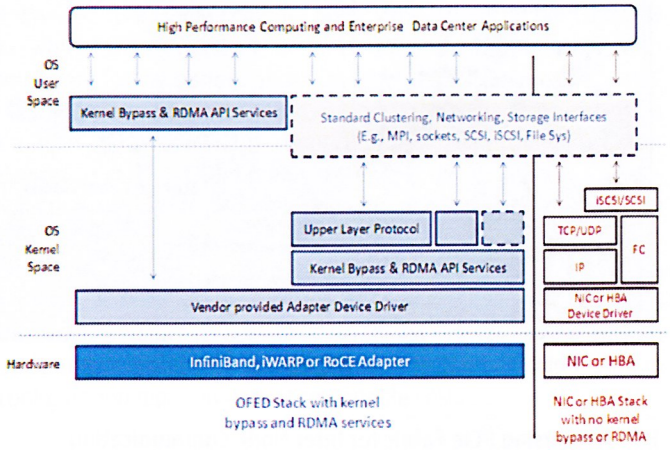


Figure 23: OFED software stack overview (Source: openfabrics.org)

These protocols are supported by the OFED software stack:

- SRP - SCSI RDMA
- iSER – iSCSI over RDMA
- RDS – Reliable Datagram Service, used for lowest latency, high performance IPC
- Internet Protocols - TCP/UDP/IP
- SDP – Socket Direct Protocol
- PureScale/GPFS, RDMA FileIO
- UDAPL – User Direct Access. Programming Library set of user APIs for RDMA
- MPI – Message Passing Interface for parallel programming Computers

Management

Infrastructure management

The DMTF (Distributed Management Task Force, www.dmtf.org) is currently working with the IT industry and other organizations on APIs supporting the "Software Defined Data Center" (SDDC) approach.

Today available cloud computing software can already build and manage public and private cloud installations. However, most of today's available cloud management software can manage virtual compute resources very well but only include basic capabilities for physical resource monitoring and management.

The Software Defined Data Center approach means this is going to change. OpenStack, as one of the most popular software stacks for up to enterprise scale cloud installations, is evolving to obtain better awareness and control of platform hardware capabilities. The OpenStack management software will support more detailed and different compute and storage capabilities by filter schedulers. Provisioning and monitoring of different accelerator devices will also be supported in the near future.

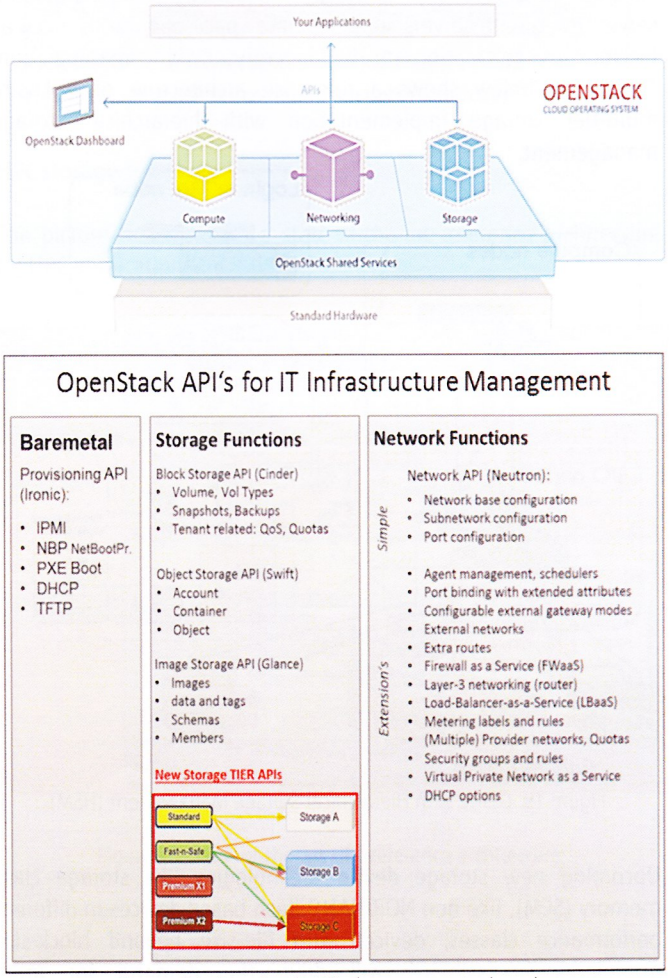


Figure 24: OpenStack APIs (www.openstack.org)

Vendor agnostic APIs are required by the IT industry to make pooled hardware resources manageable from cloud software stacks. The DMTF with supporting working groups has recently made great progress in defining platform management APIs with the support of scalable platform build from resource pools. Fujitsu will implement these standard APIs under definition.

RESTful APIs

Regarding infrastructure management Fujitsu will cooperate with Intel on their rack-scale architecture in order to provide a common API for partners, such as VMware, Microsoft, OpenStack and the Linux community. This will allow third-parties to make use of the rack-scale architecture platform for infrastructure management purposes.

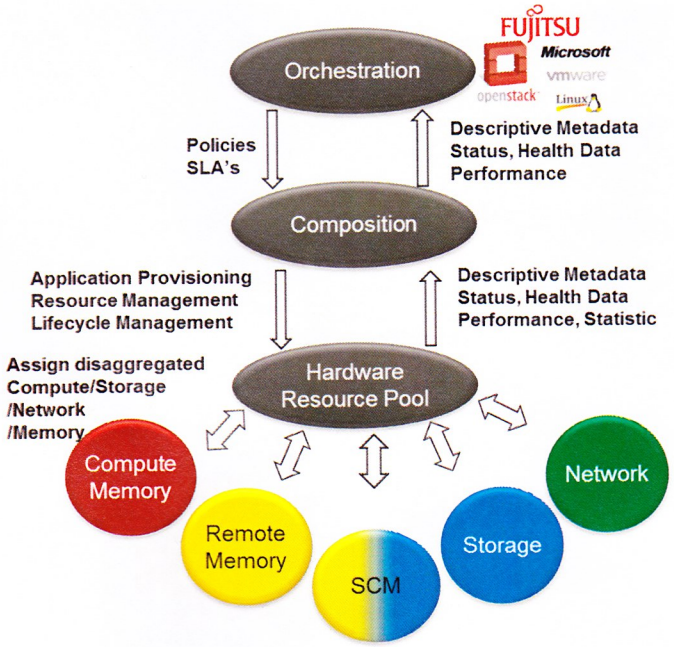


Figure 25: Orchestration and composition of hardware pools

The objective is to provide a management of pooled compute, network, storage and other resources. This includes support for discovering resource capabilities and automated provisioning of logical resources based on application requirements. The measurement and management of resources already consumed is also maintained.
The target is support for a policy-driven orchestration of pooled resources to meet application demand and service level obligations for customer workloads.

The server side of the management should be untouched. The Board Management Controller (BMC) is the server management interface to the upper layer management frameworks. The composition of servers and pooled parts out of the RSA 0.5 I/O box are typically carried out pre-boot via the independent RESTful API and the RSA service Layer. If the OS or Hypervisor is running, PCIe Hot-Plug operation may be used to change the composition of server and pooled parts.

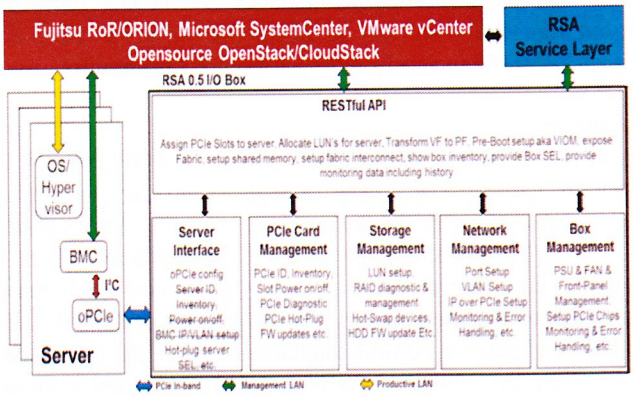


Figure 26: Rack Scale Architecture Platform Management

The management controller on the RSA 0.5 I/O box is a standard Type 10 COMExpress module with a powerful Celeron CPU running a Linux OS. This COMExpress module provides all the necessary interfaces to the box components. The chassis manager software runs on that COMExpress module. As interface to the higher management layers the I/O box chassis manager will provide a RESTful API. The chassis manager includes RSA 0.5 I/O Box Management, Network Management, Storage Management, PCIe card Management and management of the server interface.

Intel PCIe Optic Transceiver Solution

To make a PCIe fabric happen server nodes and I/O must be connected. With copper cable only solutions restrictions in terms of cable length and connector density have to be accepted. Just up to 5 meter is supported and the connector is as huge as a PCIe low profile front shield. The length restriction limited the reach of such a PCIe connection to just neighbour nodes inside the rack and the density problem made it impossible to connect many servers to an I/O box.
With the Intel Silicon Photonics solution both restrictions are overcome. The reach of such optical connection is now up to 300 meters allowing remote boxes even behind fire barriers in the data center. The density of the MXC cable and the density of the optical engine allow high port count switches. The capability of PCIe to combine lanes or to split lanes (bifurcation) allows a flexible optical port configuration. For example an individual x16 optic PCIe transceiver can be used as one big port with 128Gbps or 2 ports with 64Gbps or 4 ports with 32Gbps.

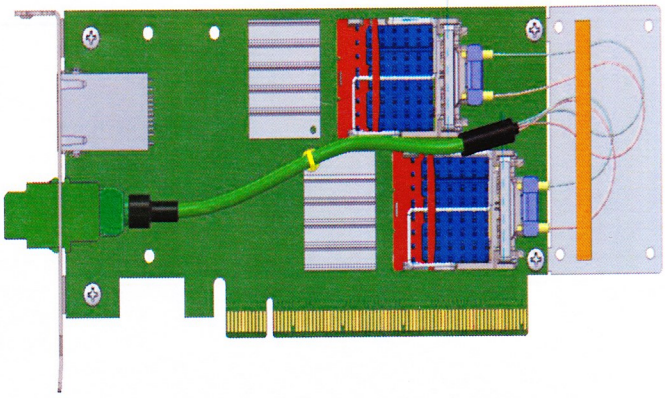


Figure 27: Intel x16 PCIe Optical PCIe Board - Preliminary

The image above shows a preliminary x16 optic PCIe transceiver design from Intel, which can be used in the Fujitsu RSA box and in Fujitsu PRIMERGY servers for electrical to optic conversion.

Conclusion

Using Intel's Silicon Photonics Technology, the PRIMERGY RSA 0.5 platform provides a highly modular, flexible and robust architecture – future-proof, fault resilient, versatile and designed to fit various use cases. A huge variety of I/O components can be plugged into the platform - from the highest performance non-volatile memories, such as PCIe add-in cards or 2.5" SFF SSDs, to high performance network interfaces or high-end co-processor cards. They are connected via Silicon Photonics based links with multiple servers over distances up to 300m. The implementation ensures that devices inside the PRIMERGY RSA platform can be dedicated to single remote servers or can be shared between multiple remote servers. Sharing PCIe based storage components inside a rack between servers will dramatically optimize a shared storage infrastructure, eliminating the need for a SAN and for SAS connectivity. It simplifies the setup and optimizes OPEX and CAPEX. In parallel the PRIMERGY RSA platform avoids single point of failures and provides excellent RAS features. The architecture is server agnostic – PCIe over Silicon Photonics based links provides the single connection between a host server and the PRIMERGY RSA platform. Slot connectivity is ensured as long as the server has a free PCIe slot.

There is a huge range of use case scenarios for the RSA platform 0.5. The basic principle is to provide pools of resources to compute entities, maintaining a high level of flexibility and the ability to reconfigure at any time. In this sense the configuration can be adapted to the varying needs of the application which the end-user wants to start on his server – a truly application-optimized architecture.

Authors

Georg Müller	FTS PSO PM&D SV E HW	georg.gm.mueller@ts.fujitsu.com
Bernhard Homölle	FTS PSO PM&D SV E HW	bernhard.homoelle@ts.fujitsu.com
Ewald Harms	FTS PSO PM&D SV PM DCS	ewald.harms@ts.fujitsu.com
Timo Lampe	FTS MKT P Server	timo.lampe@ts.fujitsu.com
Bernhard Schräder	FTS PSO PM&D IN FL	bernhard.schraeder@ts.fujitsu.com

Contact

FUJITSU
Fujitsu Technology Solutions GmbH
Website: www.fujitsu.com