\Orchestrating a brighter world

High Performance Computing

NEC LxFS Storage Appliance

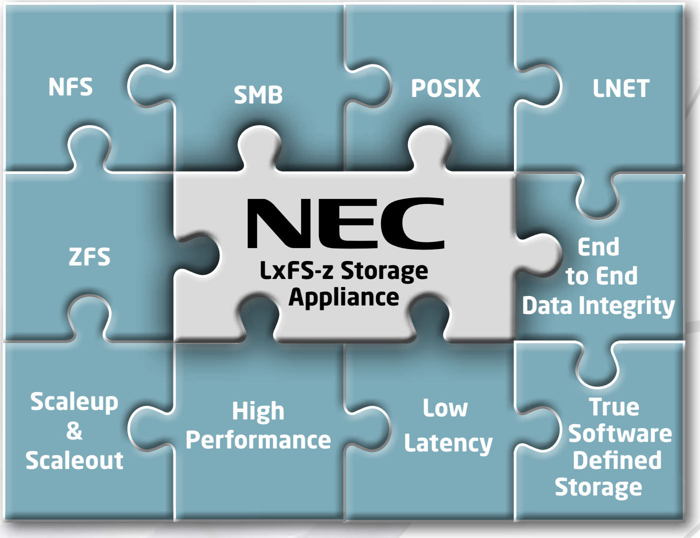
NEC LxFS-z Storage Appliance

In scientific computing the efficient delivery of data to and from the compute nodes is critical and often challenging to execute. Scientific computing nowadays generates and consumes data in high performance computing or Big Data systems at such speed that turns the storage components into a major bottleneck for scientific computing. Getting maximum performance for applications and data requires a high performance **scalable storage solution**. Designed specifically for high performance computing, the open source Lustre parallel file system is one of the most powerful and scalable data storage systems currently available. However, the managing and monitoring of a complex storage system based on various hardware and software components will add to the burden on storage administrators and researchers. **NEC LxFS-z Storage Appliance** based on open source Lustre customized by NEC can deliver on the performance and storage capacity needs without adding complexity to the management and monitoring of the system.

NEC LxFS-z Storage Appliance is a true **software defined storage** platform based on open source software.

NEC LxFS-z Storage Appliance relies on two pillars: Lustre delivering data to the frontend compute nodes and ZFS being used as filesystem for the backend, all running on reliable NEC hardware.

As scientific computing is moving from simulation-driven to **data-centric computing** data integrity and protection is becoming a major requirement for storage systems. Highest possible data integrity can be achieved by combining the RAID and caching mechanisms of ZFS with the features of Lustre.



Highlights

- Lustre based parallel file system appliance for scientific computing
- Software defined storage solution based on ZFS for backend storage
- Complete storage solution, delivered with software stack fully installed and configured
- NEC SNA Storage Systems for best high-density, performance and reliability characteristics
- Fully redundant hardware and high-availability software suite for always-on operation
- NEC LxFS-z Storage Building Block concept for cost-efficient system configuration matching performance and capacity demands
- Support for high-speed access via external protocols including SMB/CIFS and NFS
- Built-in granular monitoring of all system components to ease maintenance and maximise uptime
- Hadoop adapter for Lustre for seamless integration with Hadoop Infrastructures
- MapReduce adapter for HPC available to integrate in Big Data Analytics environments
- NEC support for both hardware and software

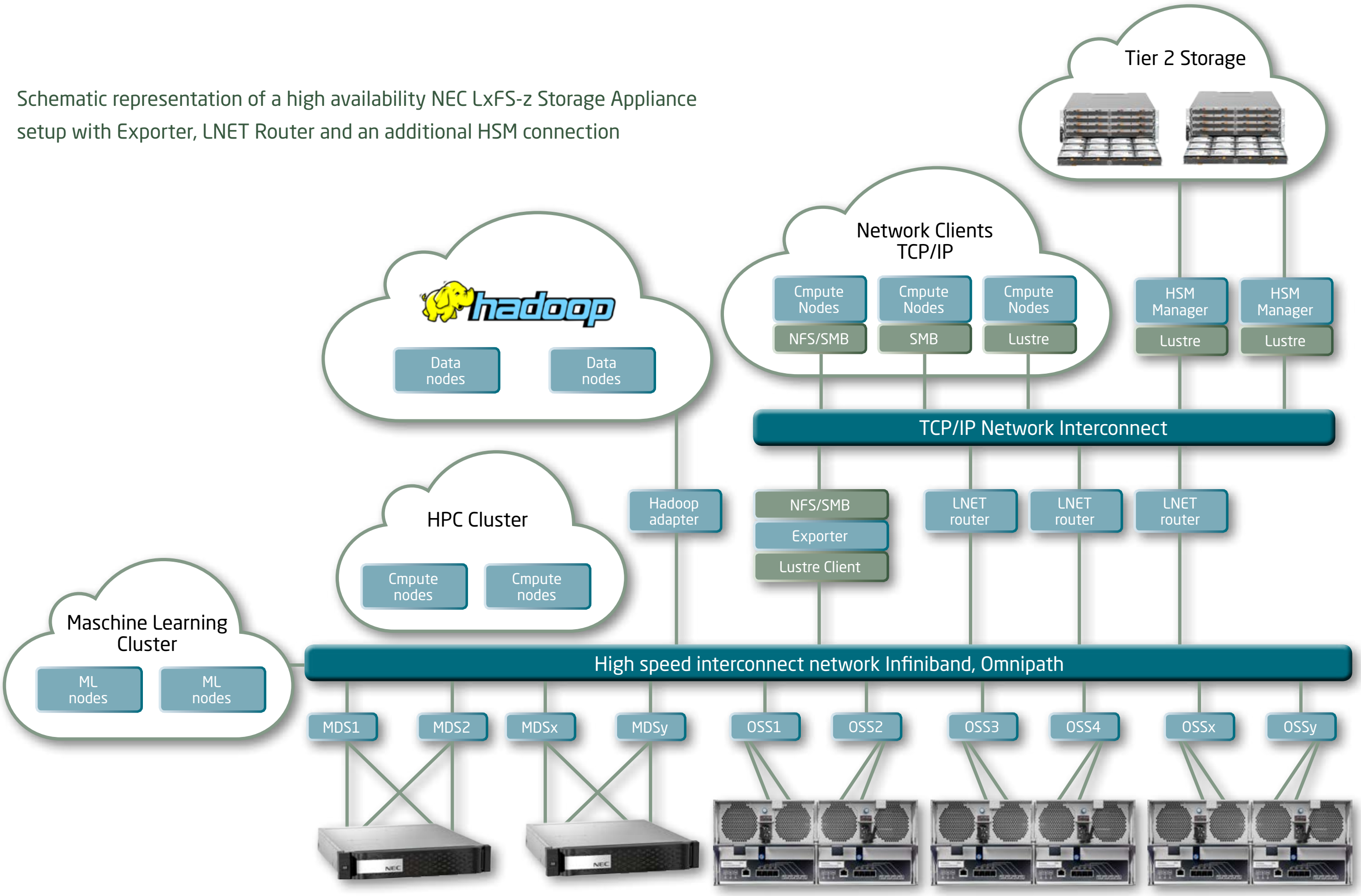
Lustre Architecture

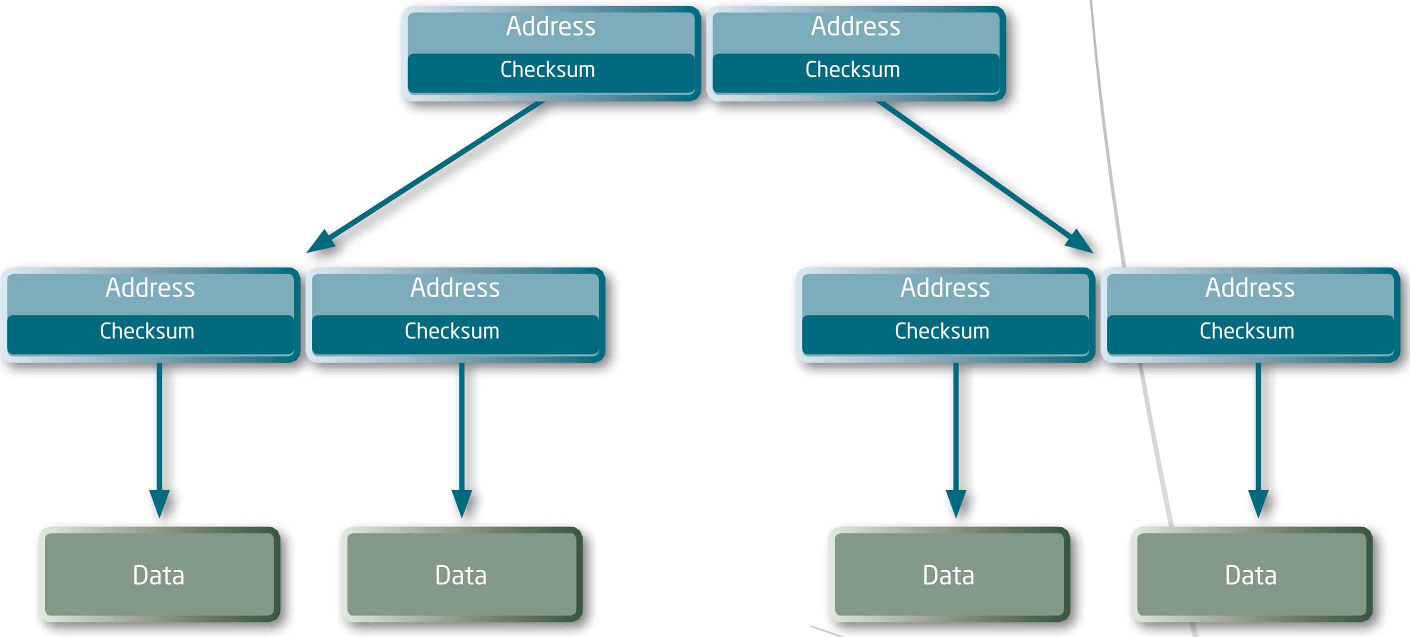
Lustre is an object based high performance parallel file system which separates file metadata management and actual file data handling and stores them on different targets. File metadata like name, permissions, layout, is stored on **Metadata Targets** (MDT) and processed by Metadata Servers (MDS) while file data is split into multiple objects and stored on **Object Storage Targets** (OST) which are mounted on Object Store Servers (OSS). The file system capacity is the sum of the OST capacities. On client side Lustre presents a file system using standard **POSIX** semantics that allows concurrent and coherent read and write access. When a client accesses a file, it completes a filename lookup on the MDS. The MDS returns the layout of the file to the client. After locking the file on the OST, the client will then run one or more read or write operations on the file by delegating tasks to the OSS. Direct access to the file is prohibited for the client. This approach enhances scalability, ensures **security and reliability**, while decreasing the risk of file system corruption from misbehaving or defective clients.

Servers and clients communicate through the Lustre Network **(LNET)** protocol, a network stack that is able to use TCP/IP and Infiniband networks at performances close to their native speed. Lustre Clients can tolerate failures of servers as long as the targets can be mounted from other servers. **LNET Routers** can connect multiple InfiniBand or Ethernet networks and route LNET traffic between them. Lustre clients can re-export the file system through **NFS** or **SMB/CIFS** or act as data movers to/from a second tier HSM file system. The architecture enables scaling to thousands of OSTs, tens of thousands of client nodes, hundreds of Petabytes of storage capacity and over a Terabyte per second aggregated I/O bandwidth .Features like an **online file system check** allow to repair inconsistencies of the file system while the file system is up and running.

The NEC LxFS-z Storage Appliance scales not only in terms of bandwidth and performance but also in terms of connectivity. Using the Hadoop Adapter for Lustre a shared data repository for all compute resources can be established. The usage of data in place without import and export of data combined with a high performance low latency Lustre filesystem provided by the NEC LxFS-z Storage Appliance allows the creation of data centric workflows. You can have dedicated compute and data nodes or run a single compute cluster that can be used for variety of workloads. To support machine learning technologies, storage systems have to deliver throughput and performance at scale. I/O bottlenecks and massive data storage can be considered as a major challenge for machine learning and AI environments The NEC LxFS-z Storage Appliance effectively reduces the problem of I/O bottlenecks and with its focus on high performance access to large data sets, NEC LxFS-z Storage Appliance combined with burst buffers and flash devices can be considered a valid solution to full scale machine learning systems.

Schematic representation of a high availability NEC LxFS-z Storage Appliance setup with Exporter, LNET Router and an additional HSM connection





ZFS Solution for Data Integrity with Lustre

ZFS uses a **copy on write** transactional model for writing data. Blocks on disk with active data will never be overwritten and data will be always consistent on disk and a snapshot of the data can happen at no cost and without any performance impact anytime.

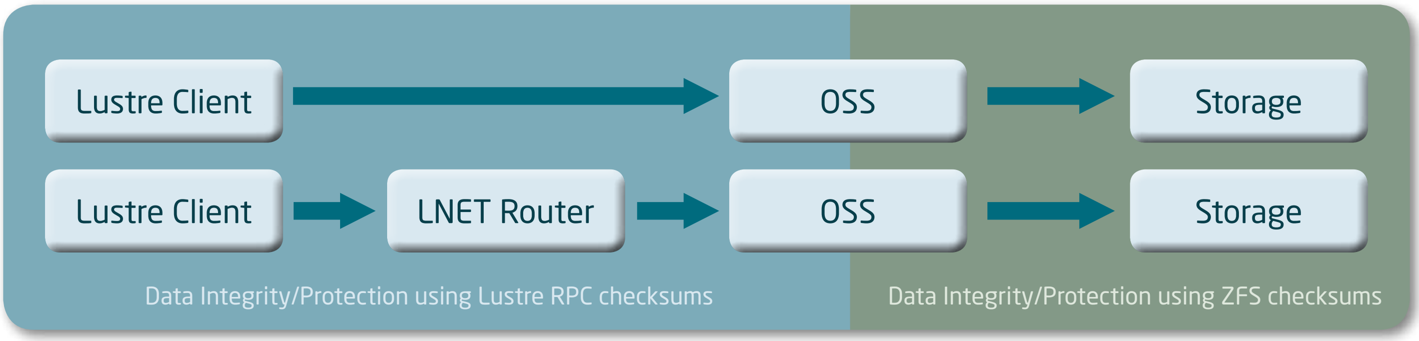
One of the key features of ZFS is the extremely powerful **software RAID engine** that allows single, dual, and even triple parity raid configurations. An important objective in the development of ZFS was to eliminate expensive hardware RAID controllers for building enterprise class storage solutions. The so-called RAID-Z software RAID implementation of ZFS has several outstanding features. To prevent the so-called "RAID-5 write hole" which can also happen when using RAID 6, RAID-Z uses variable-width RAID stripes resulting in all writes being full-stripe writes. Full stripe writes guarantee not only data protection, they also greatly improve write performance. In combination with a highly tuneable and reliable I/O scheduler ZFS outperforms most of the hardware RAID-controller based storage systems

lustre™

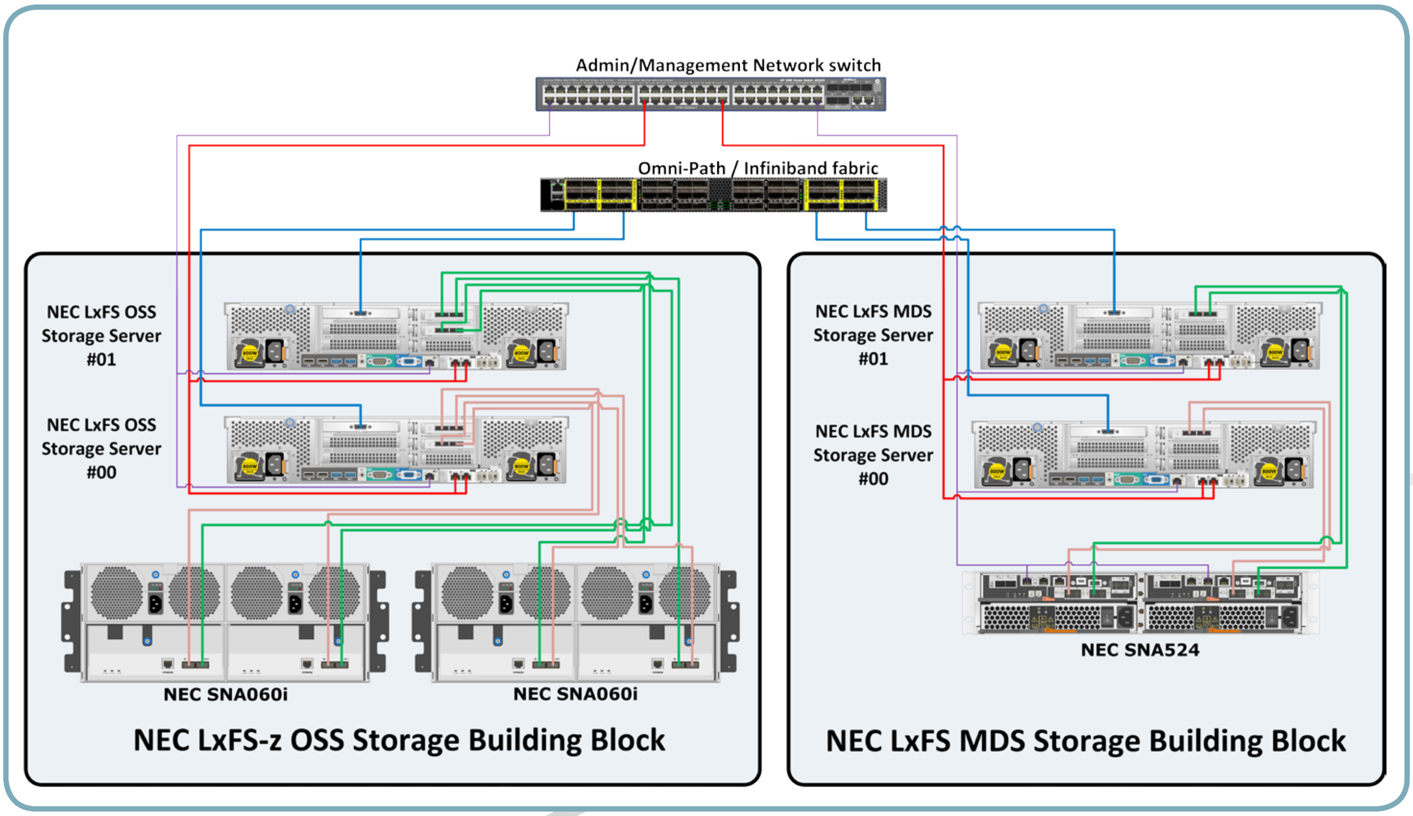
Intelligent caching algorithms greatly improve performance of a ZFS based system. Conceptually ZFS differentiates three caching methods. The **Adaptive Replacement Cache** (ARC) is the first destination of all data written to a ZFS pool, and as it is the DRAM of the server it is the fastest and lowest-latency source for data read from a ZFS pool. When the data is in the ARC. The contents of the ARC are balanced between the most recently used and most frequently used data. Level Two ARC (**L2ARC**) is an extension of the ARC based on SSD. The L2ARC Cache is a read cache to take the pressure from the ARC cache. The algorithms that manage ARC to L2ARC Migration work automatically and intelligently. The ZFS Intent Log (ZIL) is used to handle synchronous writes-write operations that are required by protocol to be stored in a non-volatile location on the storage device before they can be T10-PI data protection, which protects only against silent data protection ensure data stability. ZFS can do this by using placing the ZIL on a mirror of enterprise grade write-optimized SSD. All writes (whether synchronous or asynchronous) are written into the DRAM based ARC, and synchronous writes are also written to the ZIL before being acknowledged. This is comparable to the concept of NVRAM used in a hardware RAID-controller. Under normal conditions, when ARC is flashed to drives, the data in the ZIL is no longer relevant Especially the way ARC and hard disks work together is one of the keys to performance for ZFS backed systems.

Common hardware-raid based storage solutions offer only a small subset of possible methods to assure data integrity. Most common used is T10-PI data protection, which protects only against **silent data corruption** but for example can't protect against phantom writes or misdirected reads or writes. To counter data degradation ZFS uses **checksums** throughout the complete file system tree. Each block of data is checksummed and the checksum value is then saved in the pointer to that block rather than at the actual block itself. Next, the block pointer is checksummed, with the value being saved at its pointer, Thus creating a Merkle tree resulting in a self-validating pool. Each time data are accessed the whole tree will be validated. A background or on demand process called **disk scrubbing** is scans and verifies all data blocks against the checksums and automatically repairs damaged blocks.

The I/O path of the data is protected by two types of checksums form the client for data sent over the network to stable storage. Within Lustre a 32-bit checksum of the data read or written on both the client and OSS is computed, to ensure that the data has not been corrupted in transit over the network. In combination with the checksums generated by ZFS NEC LxFS Storage Appliance provides enterprise data integrity all along the data path from client to disk.


Lustre Client
OSS
Storage
Lustre Client
LNET Router
OSS
Storage
Data Integrity/Protection using Lustre RPC checksums
Data Integrity/Protection using ZFS checksums

NEC LxFS-z Storage Appliance Data integrity providing always consistent data from client to disk.


Admin/Management Network switch
Omni-Path / Infiniband fabric
NEC LxFS OSS Storage Server #01
NEC LxFS OSS Storage Server #00
NEC SNA060i
NEC SNA060i
NEC LxFS-z OSS Storage Building Block
NEC LxFS MDS Storage Server #01
NEC LxFS MDS Storage Server #00
NEC SNA524
NEC LxFS MDS Storage Building Block

NEC LxFS-z Storage Building Blocks

Being a **software defined storage** appliance, the choice of components and the software configuration is crucial for usage in a production environment. Configuration of Lustre and ZFS can get complicated and error-prone, especially when hardware incompatibility causes issues. Therefore well-defined building blocks are the basic units of a NEC LxFS-z Storage Appliance. By design NEC LxFS-z Storage Appliance consists of two types of building blocks one for metadata and one for object storage.

The metadata building block relies on two **NEC HPC128Rh-2** server systems and a dual controller NEC SNA524 raid system using high performance SAS drives in a RAID-10 setup.

connecting the disks redundantly to the server systems. From the disks to the servers, all connections are realized using state-of-the-art SAS-3 technology. Each NEC SNA060i Storage Enclosure is equipped with 60 high capacity NL-SAS disk drives. For each NEC SNA Storage System, four RAIDZ2 sets are configured with 14 NL-SAS disks each. The remaining four disk drives are configured as hot spare disks.

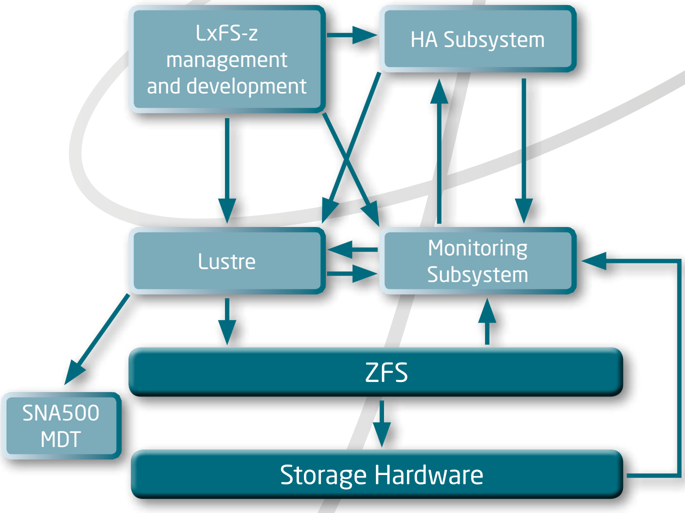
NEC LxFS-z Storage Appliance comes with a fully configured software stack **including high-availability**. The NEC LxFS-z Storage Appliance modular building block concept allows easy sizing and scaling of any I/O setup and data workflow. The building block concept allows to grow in capacity or bandwidth according to your demands. NEC LxFS-z Storage Building Blocks are delivering high performance and are due to the fully redundant configuration **without single points of failure** designed for always-on operation.

The NEC LxFS-z Storage Appliance comes with a completely configured software stack and an integrated monitoring solution. Based on the profound knowledge and a long-time operating experience with Lustre and ZFS NEC LxFS-z Storage Appliance is designed and configured for high bandwidth and reliable operation.

NEC LxFS-z Software Stack

Core of the software stack is the hardened version of the open source Lustre and ZFS on Linux. It is embedded into a framework of other software components providing reliability, management and monitoring features. Each building block runs an instance of the Pacemaker High Availability System which handles server failures gracefully without interruption of service.

In detail monitoring of important hardware and software components is provided by the health monitoring framework consisting of Nagios NEC-provided extensions. Performance monitoring data are collected by Ganglia and related NEC tools, that provide a large number of metrics about server loads, I/O, Network and Lustre activity. Parameters and thresholds of the monitoring system are based on the long-time experience of NEC in building, running and supporting LxFS-z Storage Appliances. Monitoring data and system state can be conveniently visualised and accessed from anywhere by accessing the built-in webserver which provides highly configurable user-customisable monitoring views. Based on alert rules alarm notifications will be sent in case of a failure using the preferred notification methods. The system comes with a set of CLI tools for easy management of the high-availability HA system and other administrative tasks, and for debugging and problem reporting. An automated setup deploys the NEC LxFS-z Storage Appliance in a reproducible way. This sophisticated software stack in combination with performance optimised building blocks makes the NEC LxFS-z Storage Appliance best of breed in software defined storage.



Block diagram of NEC LxFS-z Storage Appliance

NEC as a provider of Storage Appliances

The building blocks of the LxFS-z Storage Appliance are architected, integrated, tested, and optimized to work flawlessly together, thus cutting complexity and eliminating risks. This results in easier deployment and upgrades, and more efficient data and systems management. NEC not only provides hardware, but also optimal storage solutions based on know-how and experience of our employees. Consulting, benchmarking, implementation and support during all stages of a project from first design to 3rd level support are covered by NEC experts.

NEC has successfully implemented and is supporting LxFS-z Storage Appliances up to petabyte scale with a proven performance of more than 100 Gigabytes per second.

NetApp®

NEC Deutschland GmbH
HPC EMEA Headquarter
Fritz-Vomfelde-Straße 14-16
D-40547 Düsseldorf
Tel.: +49 (0) 211 5369 0

www.nec.com

HPC Division
Raiffeisenstraße 14
D-70771 Leinfelden-Echterdingen
Tel.: +49 (0) 711 78 055 0

HPC Division
3 Parc Ariane
F-78284 Guyancourt
Tel.: +33 (0) 139 30 66 00