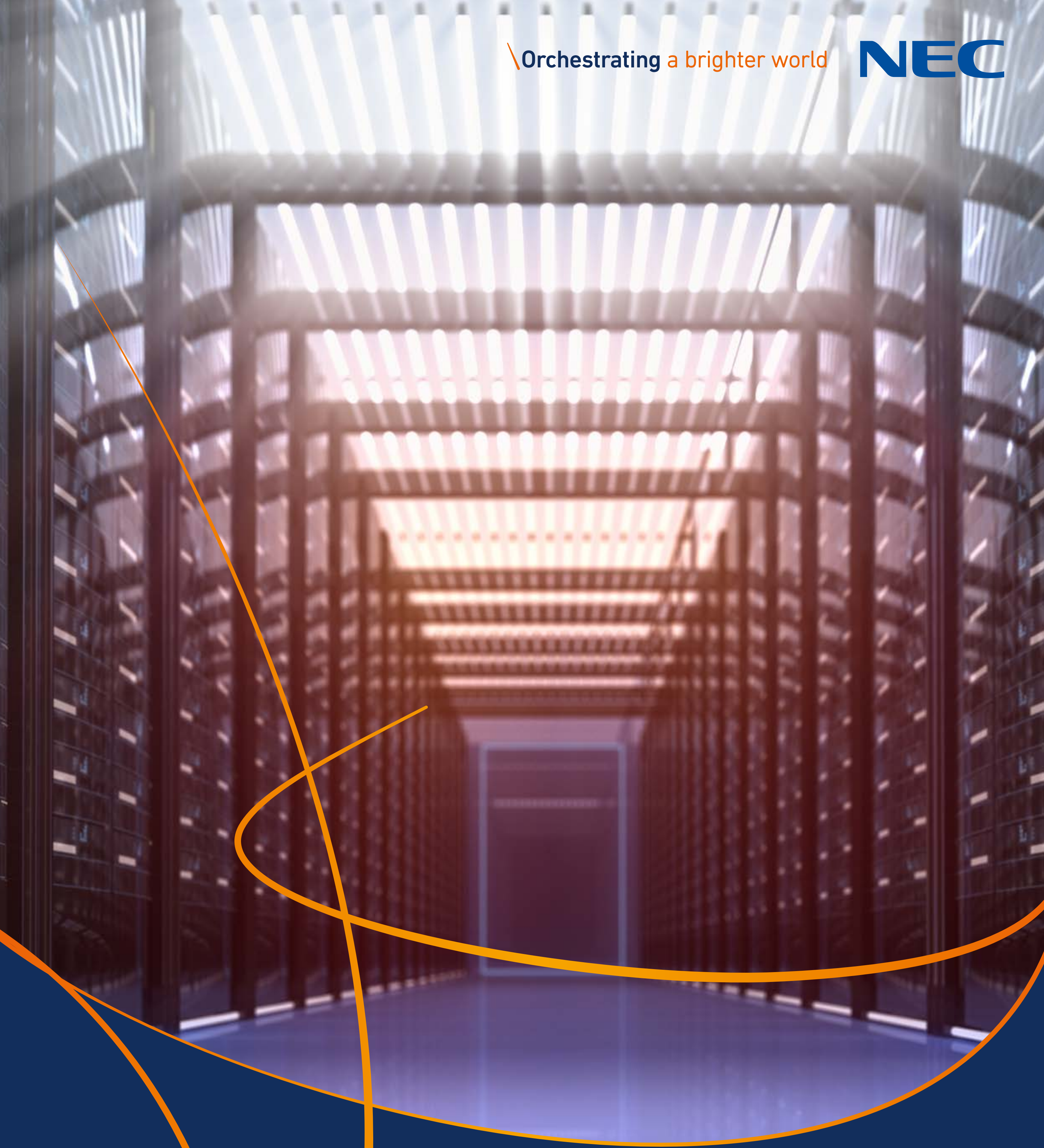
\Orchestrating a brighter world

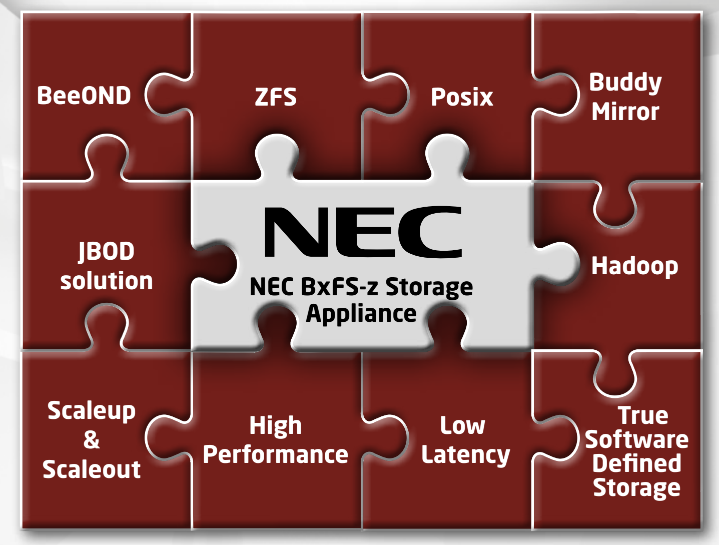


High Performance Computing

# NEC BxFS Storage Appliance

# NEC BxFS-z Storage Appliance

With the ever increasing performance of modern processors and network technologies, the need for collected and processed data is also growing. In order to handle this huge amount of data and to ensure optimal performance in the calculation, the NEC BxFS-z Storage Appliance based on BeeGFS parallel file system has been developed. BeeGFS (also known as Fraunhofer Parallel Filesystem, formerly FhGFS) is a parallel file system, specially optimized for high-performance computing (HPC). In addition to the very good scalability of the system, the NEC BxFS-z Storage Appliance attaches great importance to uncomplicated handling and high-availability combined with a maximum of flexibility and scalability. The NEC BxFS-z Storage Appliance is a true **software defined storage** platform based on open source software combining the scalability and features of BeeGFS with the data protection and RAID features of ZFS as underlying file system for the storage targets.



## Highlights

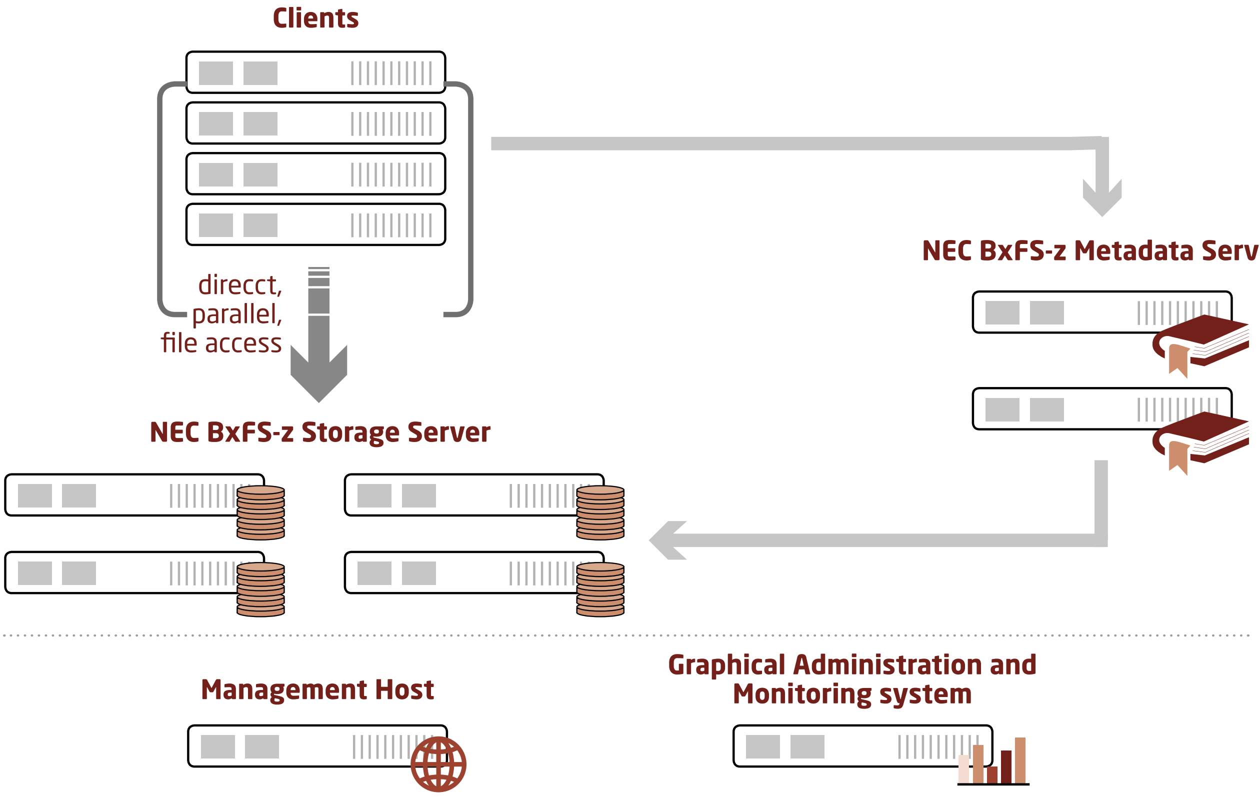
- BeeGFS-based high performance parallel file system appliance
- Software Defined Storage solution relying on ZFS for backend storage
- On demand parallel file system BeeOND included
- Fully redundant high-availability setup
- NEC SNA high density JBOD solution
- Buddy Mirroring – built-in data replication feature for high-availability
- Complete storage solution, delivered with software stack fully installed and configured
- InfiniBand, Omni-Path or Ethernet as high-speed interconnect with dynamic failover capability between different network topologies, with RDMA or RoCE support
- Multiple instances of BeeGFS can run in any combination on the same appliance
- Flexible data striping per directory or file
- NEC support for both soft- and hardware

# BeeGFS Architecture

From the beginning BeeGFS has been developed and optimized for **data throughput with a strong focus on scalability and flexibility.** Conceptually BeeGFS combines multiple storage servers to provide a highly scalable shared network file system with **striped file contents**. This way, it allows to overcome the tight performance limitations of single servers, single network interconnects, a limited number of hard drives etc. In such a system, high throughput demands of large numbers of clients can easily be satisfied, but even a single client or a single stream will benefit from the aggregated performance of all the storage servers in the system.

By design **BeeGFS separates metadata and file contents**. While storage servers are responsible for storing stripes of the actual contents of files, metadata servers do the coordination of file placement and striping among the storage servers and inform the clients about certain file details when necessary. When accessing file contents, BeeGFS clients directly contact the storage servers to perform file I/O and communicate with multiple servers simultaneously, resulting in **truly parallel access** to the file data. To keep the metadata access latency (e.g. directory lookups) at a minimum, BeeGFS can also distribute the metadata across multiple servers, so that each of the metadata servers stores a part of the global file system namespace. The following picture shows the system architecture and roles within an BeeGFS instance.

In the picture below, all services are running on different hosts to show which services generally exist in a BeeGFS storage cluster. However, it is also possible to run any combination of BeeGFS services (client and server components) together on the same machines. Performance and capacity of a NEC BxFS-z environment can easily be scaled by adding NEC BxFS-z Storage Appliance building blocks to the level needed. Adding additional storage or metadata building blocks can be done without interrupting running data services.


Clients
NEC BxFS-z Metadata Server
direcct,
parallel,
file access
NEC BxFS-z Storage Server
Management Host
Graphical Administration and
Monitoring system

BeeGFS architecture

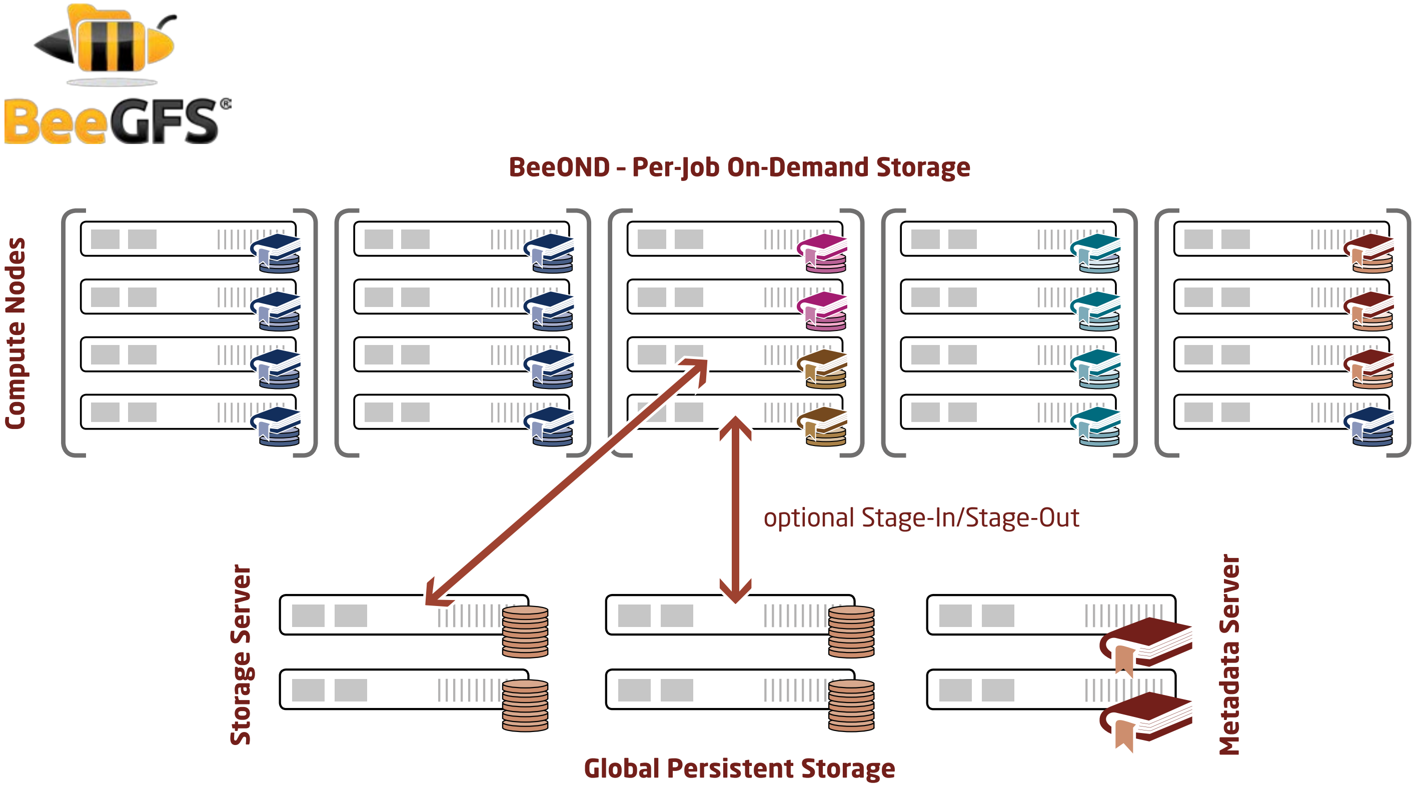
Besides the three basic roles in BeeGFS (metadata service, storage service, client) there are two additional system services that are part of the NEC BxFS-z Storage Appliance. The first one is the management service, which serves as registry and watchdog for clients and servers, but is not directly involved in file operations and thus not critical for system performance. The second one is the optional administration and monitoring service (Admon), which provides a graphical frontend for installation and system status monitoring. Besides detailed storage performance metrics live per-user and per-client statistics are available using Admon. The NEC BxFS-z Storage Appliance will be delivered as a **turnkey solution** with all services preconfigured.

## On-Demand Storage BeeOND

The NEC BxFS-z Storage Appliance provides a persistent global storage solution at the same time, it also offers the possibility of creating a temporary on-demand parallel file system on the nodes in the compute cluster.

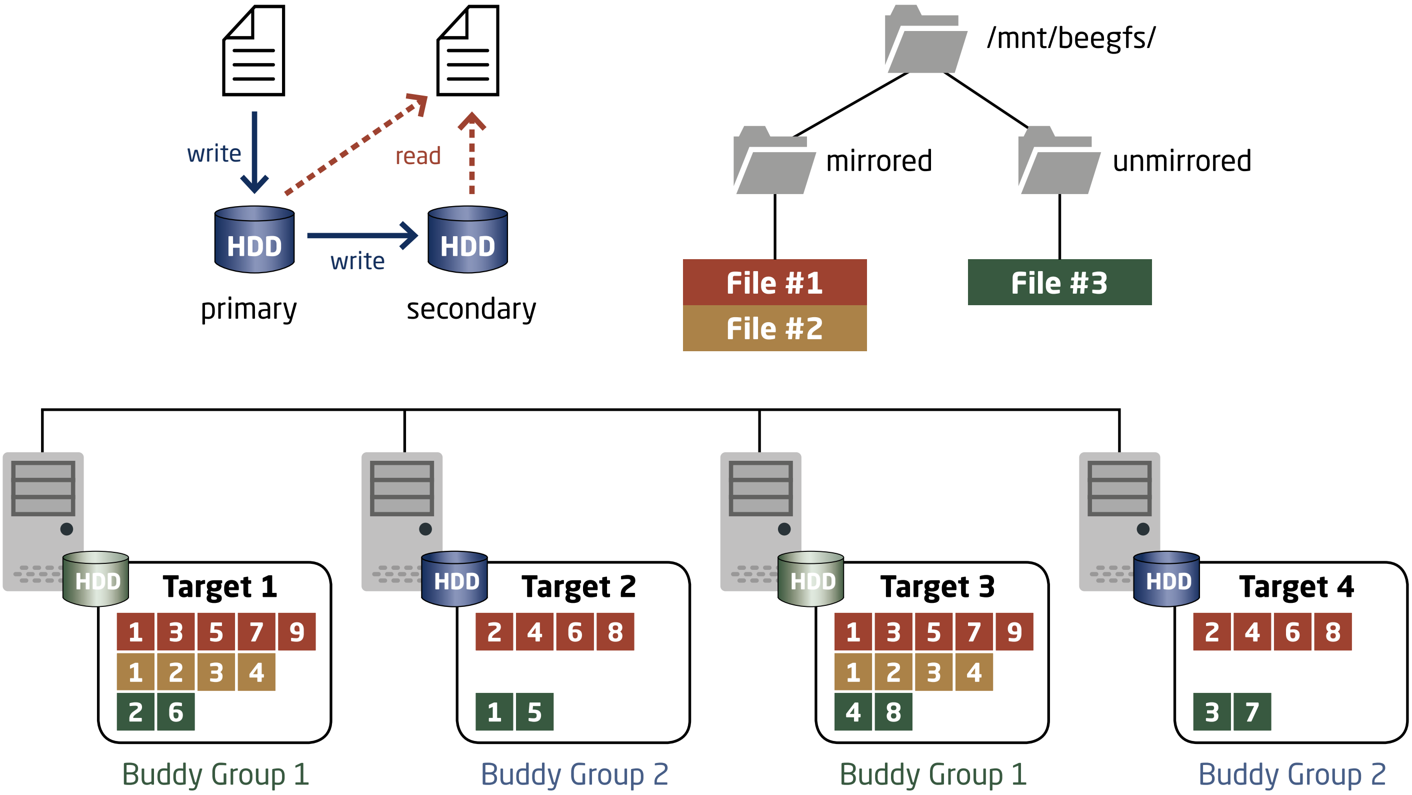
The efficient mapping from application data model to storage hardware is increasingly more complex. Therefore **BeeGFS on Demand (BeeOND)** was developed, to bring storage I/O closer to the computation layer. **BeeOND** allows on-the-fly creation of a complete, parallel file system instance on a given set of compute nodes with just one command. BeeOND is designed to integrate with cluster batch systems to create temporary, parallel file system instances on a per-job basis on internal spinning or flash devices on the compute nodes, which are part of the compute job. This provides a very fast buffer while keeping much of the I/O load for temporary, random access files away from the global cluster storage. At the beginning of a job data, can be staged from global persistent storage to the BeeOND parallel filesystem.

When the job is finished, the temporary parallel filesystem will automatically be shut down, data can be staged out using a parallel copy to persistent storage before the file system will be stopped. Typical BeeOND use cases would be for jobs that produce lot of temporary data, read input x-times or read and modify small chunks of data in-place. Not only for the use cases described BeeOND, offers a kind of **smart burst buffer** solution which can be easily implemented. In addition to this temporary parallel filesystem, the NEC BxFS-z Storage Appliance has built-in high-availability features.


BeeGFS®
BeeOND - Per-Job On-Demand Storage
Compute Nodes
optional Stage-In/Stage-Out
Storage Server
Metadata Server
Global Persistent Storage

BeeOND principle of operation

**High-Availability** for data and metadata plays a key role in scientific computing. Typically a shared storage architecture is used to keep availability of data and services high. NEC BxFS-z Storage Appliance follows this approach to handle storage server failures. If even the failure of a complete NEC BxFS-z Storage Appliance building block or a metadata server must be covered the built-in **BeeGFS Buddy Mirroring** can be used. With this feature enabled, data chunks are mirrored within primary and secondary targets the so-called Buddy Mirror Group. While reading is possible from both targets, modifying operations are sent to the primary target and forwarded to the secondary target. BeeGFS Buddy Mirroring automatically replicates data, handles storage server failures transparently for running applications, and provides automatic self-healing when a server comes back online, efficiently resyncing only the files that have changed while the machine was offline. BeeGFS Buddy Mirroring can be enabled on a per-directory base, therefore targets with Buddy Mirroring enabled can also store non-mirrored chunks.

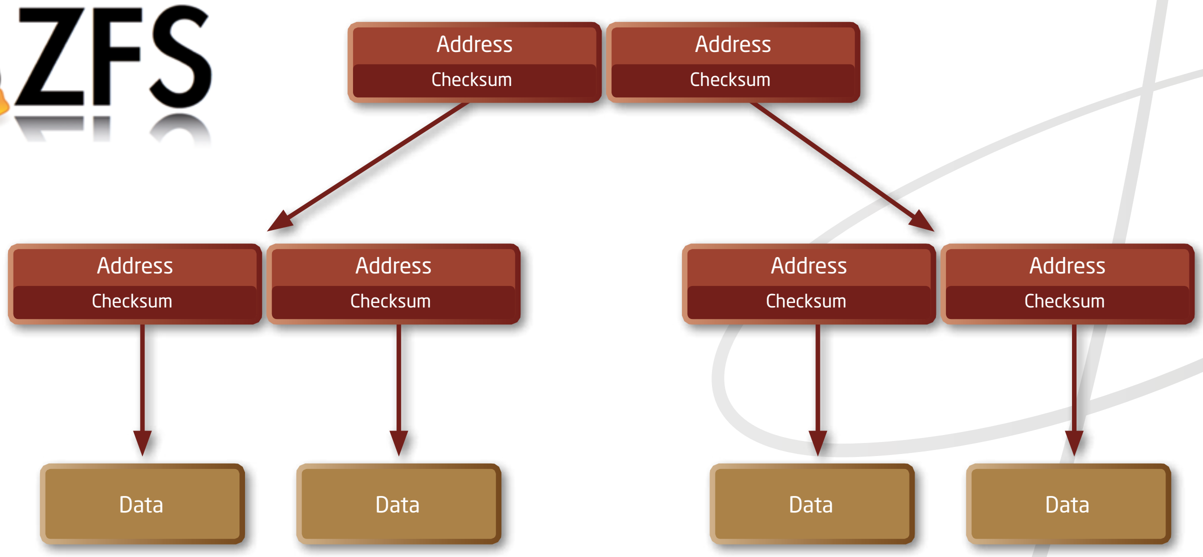


Combined with the sophisticated high-availability architecture of the NEC BxFS-z Storage Appliance, Buddy Mirroring for metadata in particular increases the availability of data without adding complexity. Buddy Mirroring offers a flexible solution to increase availability of data or metadata, as it can be enabled on a per-directory base.

The NEC BxFS-z Storage Appliance was designed with easy administration in mind. The graphical administration and monitoring system enables dealing with typical management tasks in a simple and intuitive way, while everything is of course also available from a command line interface. The monitoring system includes live load statistics even for individual users, storage service management and health monitoring.

# ZFS Solution for Data Integrity with BeeGFS

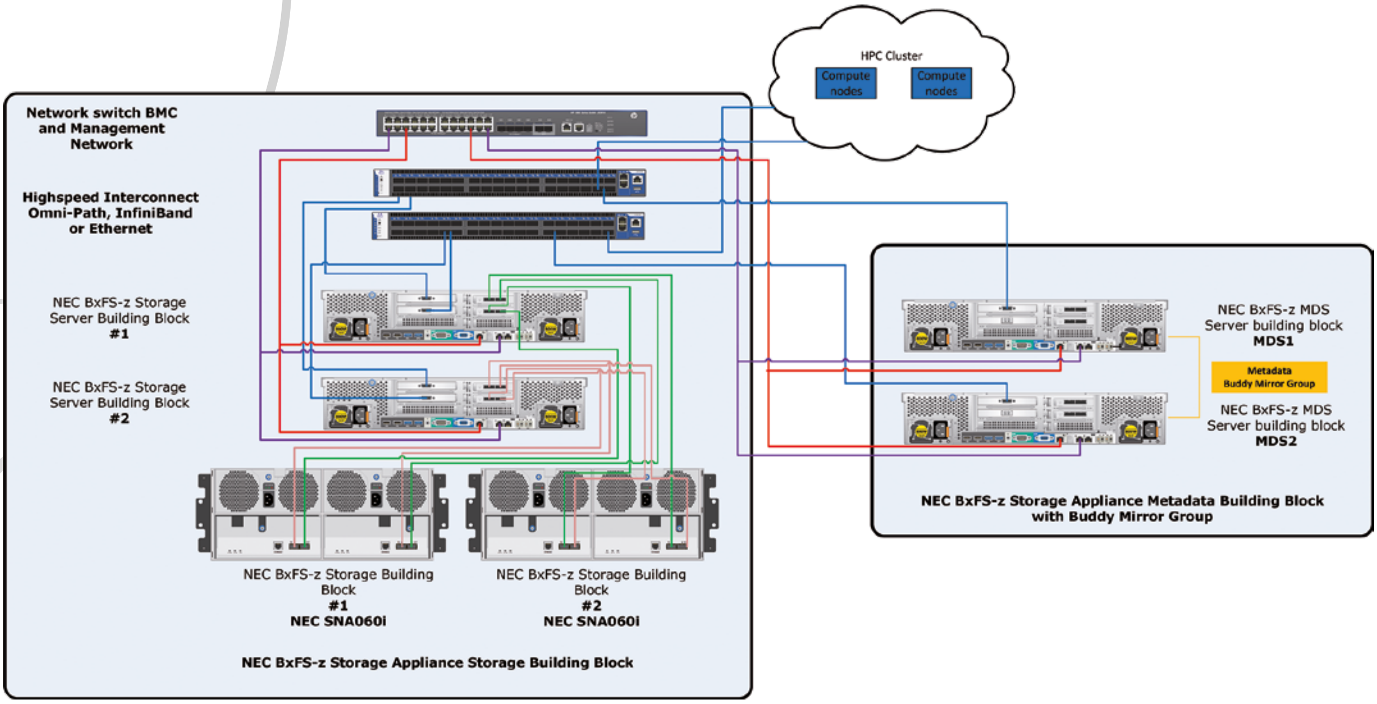
The NEC BxFS-z Storage Appliance uses ZFS as filesystem for the storage backend. ZFS applies  a **copy on write** transactional model for writing data. Blocks on disk with active data will never be overwritten and data will be always consistent on disk and a snapshot of the data can happen at no cost 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 tunable and reliable I/O scheduler, ZFS outperforms most of the hardware RAID-controller based storage systems Intelligent caching algorithms greatly improve the 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 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 write – write operations that are required by protocol to be stored in a non-volatile location on the storage device before they can be acknowledged to the host. ZFS can do this by 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 commonly used is T10-PI data protection, which protects only against **silent data corruption** but cannot 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 in 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, thus ensuring that only validated data will be read from stable storage. 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 underlying ZFS architecture ensures data integrity and protection at a high level making the NEC BxFS-z Storage Appliance best choice even for mission critical data. As NEC has a long-lasting experience with ZFS on Linux (ZOL) the selection and configuration of the ZOL version is optimally adapted to the hardware.

# NEC BxFS-z Storage Appliance Building Blocks

With the NEC BxFS-z Storage Appliance being a **software defined storage** appliance, the choice of components and the software configuration is crucial for use in a production environment. Therefore well-defined building blocks are the basic units of the NEC BxFS-z Storage Appliance. The idea of building blocks is to combine proven components needed to build a fully functional BeeGFS environment. By design, the NEC BxFS-z Storage Appliance consists of two types of building blocks one for metadata storage and one for file data storage. The BxFS-z MDS Storage Server building block for metadata relies on **NEC HPC128Rh-1** server systems with internal high endurance SSD in a RAID 10 setup. To assure availability of metadata, two NEC BxFS-z MDS Storage Server can be configured as a buddy mirror group, thus replicating metadata. The NEC BxFS-z Storage Building Block has two components the storage server and the storage target. The performance-optimized NEC BxFS-z storage building block consists of two redundant **NEC HPC128Rh-2** server systems acting as storage server. Each storage server is connected redundantly to at least two high-density 60-bay **NEC SNA60i** JBODs for the storage target. The NEC SNA060i JBOD has redundant SAS extenders connecting the disks redundantly to the server systems From the disk to the server, all connections are realized using state-of-the-art SAS-3 technology. Each JBOD is equipped with 60 high-capacity NL-SAS drives. For each JBOD, four RAIDZ2 sets are configured with 14 NL-SAS hard drives. The remaining four hard disks are configured as hot spare disks. The NEC BxFS-z Storage Appliance comes with a fully configured software stack **including high-availability**. The NEC BxFS-z Storage Appliance has built-in by design high-availability for BeeGFS. The default configuration of the NEC-BxFS-z Storage Appliance allows the failure of a storage or a metadata server without subsequent failure of the filesystem. The NEC BxFS-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.


HPC Cluster
Compute nodes
Compute nodes
Network switch BMC and Management Network
Highspeed Interconnect Omni-Path, InfiniBand or Ethernet
NEC BxFS-z Storage Server Building Block #1
NEC BxFS-z Storage Server Building Block #2
NEC BxFS-z Storage Building Block #1 NEC SNA060i
NEC BxFS-z Storage Building Block #2 NEC SNA060i
NEC BxFS-z Storage Appliance Storage Building Block
NEC BxFS-z MDS Server building block MDS1
Metadata Buddy Mirror Group
NEC BxFS-z MDS Server building block MDS2
NEC BxFS-z Storage Appliance Metadata Building Block with Buddy Mirror Group

NEC BxFS-z Storage Building Blocks deliver high performance and are due to the fully redundant configuration **without single points of failure** designed for always-on operation. Based on the profound knowledge and the long-time operating experience with ZFS and BeeGFS, the NEC BxFS-z Storage Appliance is designed and configured for high-bandwidth and reliable operation.

# NEC as a provider of Storage Appliances

The building blocks of the NEC BxFS-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 staff. 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 supports NEC BxFS-z Storage Appliances up to petabyte scale.

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