lenovo FOR THOSE WHO DO.™

Jülich Supercomputing Centre

Leveraging a breakthrough Technical Computing storage solution

Overview

The need

Jülich Supercomputing Centre supports complex research projects and needed a reliable storage solution that could meet the supercomputer's enormous bandwidth demands.

The solution

The center deployed an innovative System x® GPFS™ Storage Server solution, which provides 7 PB of usable capacity, I/O bandwidth of up to 200 GB/s, and unique GPFS Native RAID capabilities.

The benefit

Delivers extreme data integrity and reduced latency with faster rebuild times.

Jülich Supercomputing Centre (JSC) is part of the publicly funded Forschungszentrum Jülich GmbH—the largest research center in Germany. JSC supports research into a wide range of fields including fundamental physics, life sciences, climate change, and energy by providing the powerful infrastructure and technical expertise needed to run large and complex simulations quickly and effectively. Jülich Supercomputing Centre employs around 200 people, including approximately 140 experts for all aspects of supercomputing and simulation sciences.

Modeling complex systems requires huge amounts of data—especially the so-called “scratch data” that is used during simulations, but not stored for the longer term. JSC must provide the supercomputer with reliable, rapid access to scratch data.

Jülich Supercomputing Centre found that a new storage infrastructure was required to match the capabilities of the supercomputer. The most important criteria for this new storage infrastructure were reliability together with high bandwidth and capacity.

Lothar Wollschläger, Storage Specialist at Jülich Supercomputing Centre, explains, “In a large storage environment like ours, a small but significant proportion of hard drives will fail every week. Even using 2 TB disks, compared to the 1 TB disks in our previous storage solution, we need more than 4,600 disks to meet our capacity demands. In our previous conventional RAID storage array, it took around 12 hours to rebuild just a 1 TB disk, and the I/O and processing power required for the rebuild process reduced performance and bandwidth by up to 30 percent. Our new System x GPFS Storage Server (GSS) solution enables us to rebuild a 2 TB disk to a non-critical state in just under one hour, ensuring consistent, high-speed access to large volumes of data for our simulations.”

JSC had to put major effort into managing its previous storage infrastructure efficiently. All the external storage controllers required management by the technical team, increasing costs and adding complexity. GSS manages all its hardware components from a central location.

JÜLICH
FORSCHUNGSZENTRUM

Solution components

Hardware

- System x® GPFS™ Storage Server
- System x3650 M4

Software

- IBM® General Parallel File System (GPFS)

Innovative storage solution

Jülich Supercomputing Centre decided to implement a System x General Parallel File System (GPFS) Native RAID solution that eliminates the need for separate physical storage controllers. GPFS software acts as the basis for the System x GPFS Storage Server, which uses the "declustered" RAID technology to deliver not only outstanding throughput, but also extreme data integrity, faster rebuild times, and enhanced data protection.

The center deployed 20 System x GPFS Storage Server Model 24 building blocks, each of which includes two System x3650 M4 servers and four disk enclosures. These enclosures come with five disk drawers each and a total of 232 NL-SAS disks and six solid state drives (SSDs). The solution has a total of 4,640 NL-SAS disks, each with 2 TB raw capacity, providing a total usable capacity of 7 PB. The SSDs are used for declustered buffering small write I/Os and logging the RAID metadata. The IBM GPFS Native RAID, which runs on the servers, offers sophisticated data integrity, using end-to-end checksums for both read and write operations, and dropped-write detection.

Extreme data integrity

Studies have shown that disks do not report some read faults and occasionally fail to write data, while actually claiming to have written the data. These errors are often referred to as silent errors, phantom-writes, dropped-writes, and off-track writes. To cover for these shortcomings, GPFS Native RAID implements an end-to-end checksum calculated and appended to the data by the client that can detect silent data corruption caused by either disks or other system components that transport or manipulate the data.

If the checksum or version numbers are invalid on read, GPFS Native RAID reconstructs the data using parity or replication and returns the reconstructed data and a newly generated checksum to the client. Thus, both silent disk read errors and lost or missing disk writes are detected and corrected.

Faster rebuild times

Compared to conventional RAID, GPFS Native RAID implements a sophisticated data and spare space disk layout scheme that uniformly spreads (or "declusters") user data, redundancy information, and spare space across all the 58 disks of a declustered array.

A declustered array can significantly shorten the time required to recover from a disk failure, which reduces the rebuild overhead for client applications. When a disk fails, erased data is rebuilt using all 57 operational disks in the declustered array, the bandwidth of which is 6 times greater than that of the 10 disks of a conventional RAID6 group.

> "In our previous conventional RAID storage array, it took around 12 hours to rebuild just a 1 TB disk. Our new System x GPFS Storage Server solution enables us to rebuild to a non-critical state in just under one hour, ensuring consistent, high-speed access to large volumes of data for our simulations."
>
> —Lothar Wollschläger, Storage Specialist at Jülich Supercomputing Centre

Furthermore, GSS distinguishes the critical rebuild of RAID stripes, which have already lost their parity protection (so another disk failure could cause data loss), from normal rebuild of RAID stripes, which are degraded but still have some level of parity protection. When a classical RAID6 array experiences two disk failures, all of its 8+2P RAID stripes become critical and need to be urgently rebuilt. When two disks fail in a 58-disk declustered array, on average only 2.7 percent of the 8+2P RAID stripes become critical. Rebuilding just those critical stripes to get them into a non-critical state takes only a few minutes. GSS performs only critical rebuilds at the highest priority (which may impact application I/O performance for a few minutes). The remaining non-critical rebuild will only take place when there is no user I/O activity on the system, avoiding any performance impact on user applications. This is an essential feature to maintain high performance in multi-petabyte file systems, in which disk failures are the norm rather than rare and exceptional cases.

Enhanced data protection

GPFS Native RAID automatically corrects for disk failures and other storage faults by reconstructing the unreadable data using the available data redundancy of either 2- and 3-fault-tolerant Reed-Solomon codes or 3-way and 4-way replication, which respectively detect and correct up to two or three concurrent faults. Reed-Solomon offers much higher performance compared to conventional RAID by requiring much fewer load/store operations.

The disk hospital is a key feature of GPFS Native RAID that asynchronously diagnoses errors and faults in the storage subsystem. GPFS Native RAID times out an individual disk I/O operation after about ten seconds, thereby limiting the impact from a faulty disk on a client I/O operation. The suspect disk is immediately admitted into the disk hospital where it is determined whether the error was caused by the disk itself or by the paths to it. While the hospital diagnoses the error, GPFS Native RAID uses the redundancy codes to reconstruct lost or erased strips for I/O operations that would otherwise have used the suspect disk.

High performance supports complex research studies

The outstanding performance of the solution supports researchers as they conduct studies in a wide range of fields. For example, Prof. Dr. med. Katrin Amunts has been working on a project called The BigBrain at Jülich Supercomputing Centre. The study aims to create the first ultra-high resolution 3D model of the human brain to redefine traditional neuroanatomy maps and support advancements in neurosurgery.

"We are highly satisfied with their offering and delighted to have such a reliable partner to advance our research work," says Klaus Wolkersdorfer, Head of High-Performance Computing Systems at Jülich Supercomputing Centre.

For more information

To learn more about System x contact your Business Partner or visit: **lenovo.com**/servers

For more information on Jülich Supercomputing Centre, please visit: www.fz-juelich.de/ias/jsc/EN

Please Recycle