11 | 22 Reprint for Thomas-Krenn.AG

IT Administrator
The magazine for professional system and network administration

Thomas-Krenn Open-E RA1112 Metro Cluster

www.it-administrator.com

Thomas-Krenn Open-E RA1112 Metro Cluster

Mirror, mirror in the data center

by Jürgen Heyer

With the Open-E RA1112 (All-Flash) Metro Cluster, Thomas-Krenn has launched a very compact storage appliance based on the robust ZFS file system and the well-known Open-E JovianDSS operating system. Two identical servers form a metro cluster to ensure highly available data storage through synchronous mirror operation. We took a closer look at this dual server setup in the lab.

Based out of Freyung in Lower Bavaria, Thomas-Krenn has been offering hardware according to the build-to-order principle with a focus on individual server and storage systems for 20 years. The manufacturer relies on renowned hardware and software components to build the systems – and the Metro Cluster we tested is no exception.

Measuring in at just one height unit in a 19-inch rack, the server is characterized by a very compact design, whereby the required usable capacity can be individually configured via different drive configurations.

The Open-E RA1112 can also be operated as a single system, but according to Thomas-Krenn, these appliances are now mainly sold as pairs for operation in failover clusters with synchronous data mirroring. Ideally, these are operated at two sites so that – in the event of a disaster and the loss of one site – data is not lost and can continue to be accessed. In terms of distance, cable lengths of up to 100 meters are standard, and up to 80 kilometers are possible with special and more expensive transceiver cables.

Conveniently, despite current supply problems, Thomas-Krenn was able to provide us with a pair of these cutting-edge servers in cooperation with Open-E and Intel. They should be publicly available by the time this test is published. Adaptations to the respective environment and network infrastructure can be configured individually in the online shop using the modular system, depending on delivery capability.

Proven hardware from ASUS and Intel

The basis of the RA1112 is the ASUS server RS500A-E11-RS12U, which is equipped with full remote management with KVM over LAN and IPMI 2.0 via a dedicated network port. Two additional RJ45 1 Gbit/s network ports are installed as standard, one of which was used to connect to the management network for the Open-E operating system in our configuration.

The test device also had two additional dual-port network cards from Broadcom, a BCM57416 with 10 GBE-Base-T ports, suitable for the connection to our data network, as well as a BCM57414 with 25 GBESFP28 ports as a double direct connection between the two servers for data mirroring. Twelve SAS or SATA hard drives or SSDs (NVMe) can be inserted into the device for data storage from the front in 2.5-inch hot-plug frames.

Twelve Intel Solidigm NVMe D7-P5620 6.4 TB SSDs were installed in our test device. Two M.2 slots are provided internally for the operating system. Here, two NVMe SSDs with a capacity of 480 GB each were installed in the test device and ran in mirror mode. Due to this all-flash configuration, no additional volumes were required for write logs or as read cache, as we knew from previous test settings with ZFS file system and Open-E JovianDSS operating system.

The appliance is switched on via a small button, which can be found on the front in the narrow frame under the hot-plug slots. Two further buttons enable a reset and switching of the locator LED. Some LEDs in the frame indicate disk accesses as well as network traffic. Identical buttons including LEDs can be found on the back. There is no USB port in the front, but there are two in the back. In addition, there is a VGA port for a monitor. This makes it possible to adjust the IP addresses directly on the device, for instance, if needed.

Power is provided via two redundant hot-swap power supplies with 800 watts

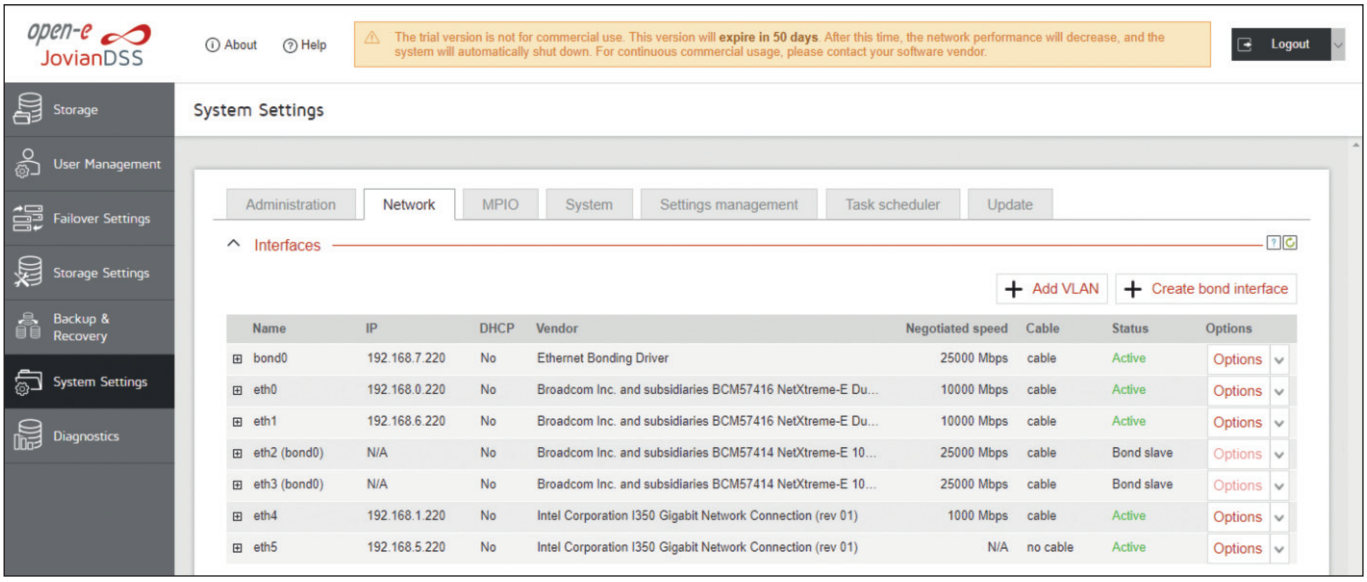

Figure 1: The network configuration of a cluster is complex and requires advance planning.

each, which allow swapping during operation. A fixed thumbscrew is used to open the enclosure lid, so there are no loose parts that can fall off when working on the appliance.

Despite the many network connections, cabling the devices was done in no time, not least thanks to Open-E's good documentation. The two 25 GBE ports of the servers are directly connected via an SFP28 cable, the possible cable length determines the maximum distance between the two servers. The two 10 GBE ports of each server are ideally distributed crosswise on two switches for redundancy reasons, while the 1 GBE port is connected to the management network on one of the two switches. The remaining remote IPMI ports play only a minor role. Some companies use their own VLANs here only for the remote ports, otherwise they can be used for the management network.

IP address assignment requires care

In contrast to the rather simple wiring of the network connections, the subsequent IP configuration requires special attention and care. Very helpful is a step-by-step guide from Open-E, which describes the individual steps in detail with many screenshots and corresponding explanations. There is also a very clear web GUI for the administration of both appliances, which can be accessed via the management addresses. While the initial steps must be performed separately on each system, administration can later be performed for the entire cluster via one node.

Although we were provided with our test device completely pre-configured, we wound up deleting the settings, in part because the entered IP addresses did not exactly match our test network. It should be noted here that several subnet areas (3x bonding, 2x storage access, 1x management) are required for communication between the two server nodes, and duplicate assignment should be avoided at all costs – it is advisable to maintain precise documentation here. The two subnets for storage access must be different from the subnets in which the intended client servers are located. Instead, they are accessed in the same way as other cluster configurations via virtual IP addresses that can be defined in the context of the storage pools.

With our test devices, we also had to take into account that the Ethernet port numbers were not assigned in the same order by the operating system, so we could not orientate ourselves in this way. Rather, when configuring the bonding over the 25-GBE SFP ports, we had to disconnect again to see the mapping. In order to set up the bonding correctly, suitable IP addresses from two subnets must first be entered for the two connected port pairs. Then a so-called bonding interface is to be defined on each

Thomas-Krenn Open-E RA1112 (All-Flash) Metro Cluster

Product

Storage cluster based on Open-E JovianDSS Linux and the ZFS file system.

Manufacturer

Thomas-Krenn
www.thomas-krenn.com/de

Price

The price for the system used in our testing, including twelve months of support, is EUR 69,270. The system can be individually configured.

System specifications

- One height unit in a 19-inch rack
- Mainboard ASUS KMPA-U16
- 1x AMD EPYC 72F3 (3.7 GHz, 8-core)
- 512 GB ECC-DDR4-3200-RAM (8x 64 GB)
- 2x 480 GB ATP-N600Sc M.2 NVMe SSDs with PLP for the operating system
- 12x 6.4 TB Solidigm-NVMe-D7-P5620 in hot plug frame for data
- 2x 1-GBit/s-onboard-LAN (Intel I350-AM2)
- Remote management (KVM over LAN, IPMI 2)
- 1x Broadcom BCM57414 dual-port 10/25 Gbit SFP28
- 1x Broadcom BCM57416 dual-port 10 Gbit BASE-T
- 2x 800 watt hot-swap power supply unit

Technical data

www.it-administrator.de/downloads/datenblaetter

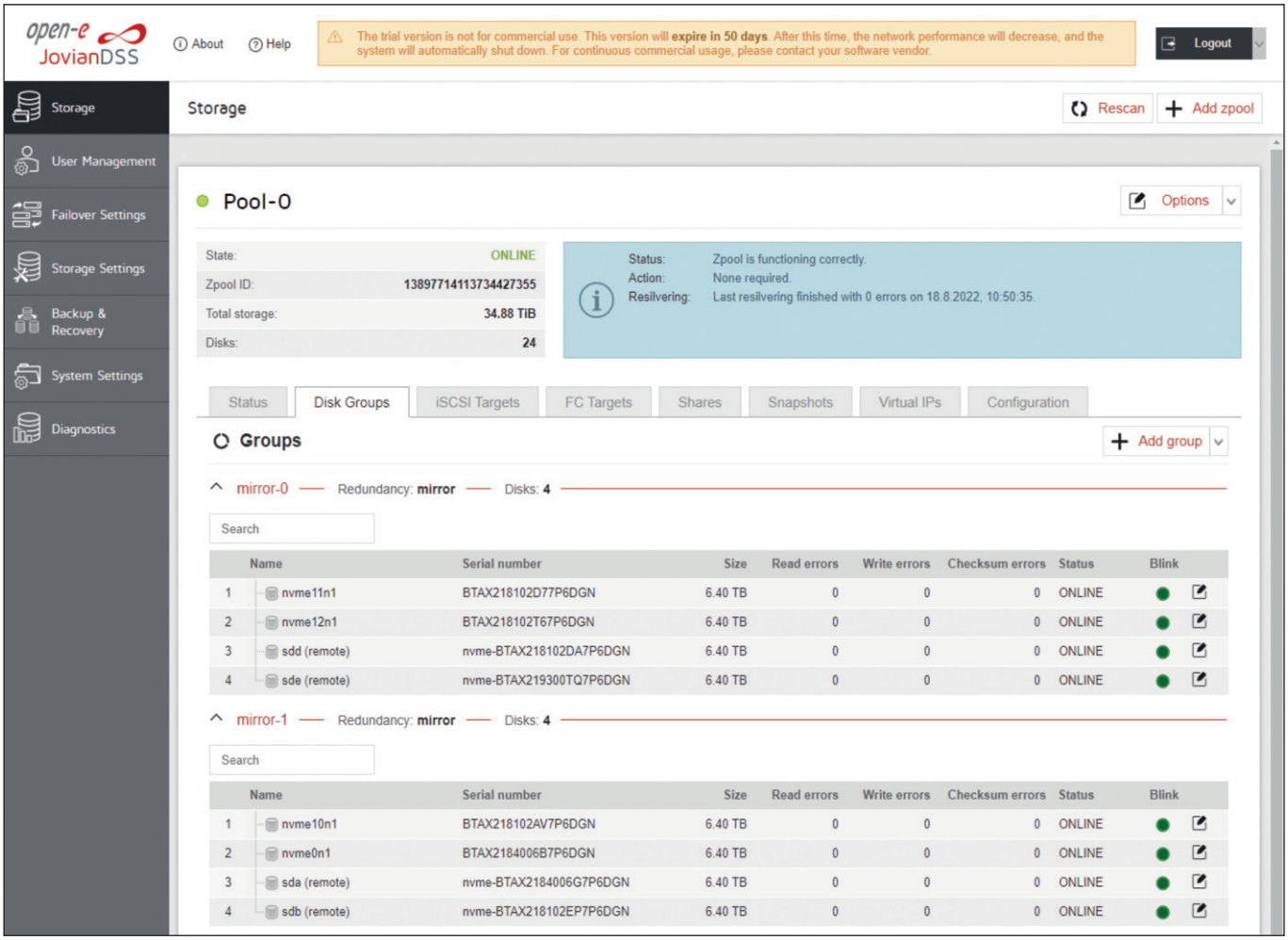

Figure 2: In this view, it is easy to see how the system combines local and remote drives into disk groups for cluster operation with mirroring.

server via the two bonding ports with a corresponding IP address from a further subnet in order to set up a mirroring path via it in a later step.

Documenting the IP addresses is important insofar as after creating the bonding interface, one is no longer able to determine which IP addresses the underlying bonding ports have received in the GUI. After creating the two bonding interfaces, the default gateway must be assigned to a port and time synchronization must be set up, preferably via NTP.

Perfectly suited for VMware environments

Next, we connected the two nodes using the previously created bonding interface and then created two so-called rings. The first ring consists of the two IP addresses of the bonding interfaces. For the second, the administrator can select another port pair, such as the two matching ports for client data access.

After that, a minimum of two and a maximum of six ping nodes must be specified with suitable IP addresses in the various subnets that the cluster can use later to detect interruptions in the network and avoid a split-brain situation. Afterwards the previously mentioned mirroring path has to be set up, consisting of the two bonding interfaces.

It is helpful that these final steps are all to be entered on one page in the web GUI, called "Failover Settings", and the two nodes always check the communication, which is confirmed with a green indication "active" or "reachable" if successful. If everything is green, the administrator can start the failover operation at the top of this page.

Switching to the "Storage" page, we could now see both the local disks and the remote SSDs of the second node on two tabs. This is the prerequisite to create a mirrored data pool, here called a Zpool.

Overall, it can be said that the IP configuration of the cluster may sound more complicated than it is in practice – especially if the administrator has the step-by-step guide at their disposal and has already gained some experience with other cluster configurations. It's worth mentioning, however, that we focused our configuration on deployment via iSCSI to a VMware environment. For use cases as a NAS with NFS or SMB, there are slightly different recommendations in the guide. Furthermore, there are separate configuration notes if, for example, it is not possible to perform the bonding via a direct connection.

Failover operation with quadruple mirroring

If all available volumes are not displayed for the following step to create a Zpool, there is the option for a rescan. We now had a total of 24 data SSDs in our two nodes, which we configured as six mirror pairs for optimal performance. To do this,

we selected two NVMEs from each of the two nodes, for a total of four SSDs, and formed a mirror group from them.

We repeated this six times and ended up with a Zpool of about 34.7 TiBytes. Hot spare drives were not available in this case, as this would have meant doing without a mirror group with just under 6 TiByte capacity. In other configurations with even more drives, additional hot spare disks might make sense, but in our test configuration, the data was identically available four times anyway due to the mirror groups of four disks each, which meant that the usable capacity was also only 25 percent of the gross value. Even if individual SSDs fail, mirror redundancy continues in this configuration, as well as if one of the two nodes fails completely or only restarts.

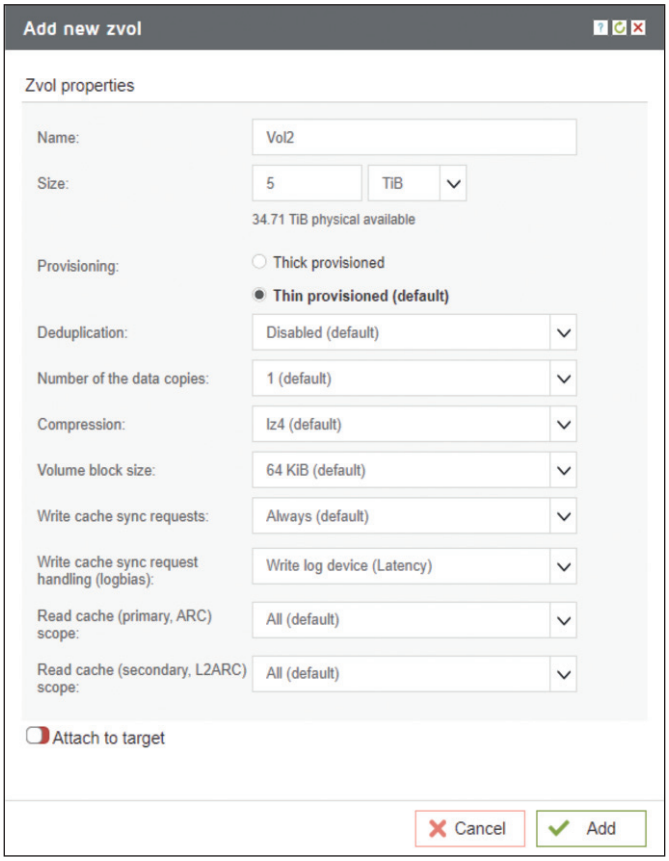

Figure 3: When creating a new volume, options such as deduplication and compression can be selected.

and SAN platform with thin provisioning, compression and deduplication. Despite the Linux base, the system supports VMware, Citrix, Linux, RHEL, MacOS, XEN and OpenStack as well as Hyper-V. However, a Windows-based NAS system might integrate better into pure Microsoft environments, but this will depend on the individual situation.

ZFS is a 128-bit copy-on-write file system that can be used to create snapshots very efficiently. It works transactionally and was developed specifically for server operation. The maximum file size is 16 EiByte (=16*2^{60} bytes), the number of manageable files is 2^{48}. It has built-in RAID functionality, volume management and checksum-based protection against file transfer errors with little impact on performance.

In our setup with one Zpool across the entire capacity, the two nodes operated in active-passive mode because a Zpool can only be completely assigned to one node. If an administrator creates two or more Zpools, they can distribute them between the two nodes and thus run an Active-Active operation.

If you are looking for higher storage efficiency instead of optimal performance, you can also realize this with Open-E JovianDSS, but you have to order a hardware RAID controller to configure a RAID 6 array per node, for example, which is then mirrored between the nodes. The operating system contains all drivers and tools for the RAID controllers from Broadcom (LSI), Microsemi (Adaptec) as well as Areca. Once the Zpool is set up, the admin has to deal with the IP address assignment again and, as already mentioned, assign a virtual IP address from the storage networks to each Zpool. These addresses are always served by the active node and can migrate accordingly in the event of a failover.

ZFS file system – a robust foundation

Thomas-Krenn chose to rely on Open-E's JovianDSS for the operating system when building the RA1112. The OS runs on the 128-bit ZFS file system. This has been used for many years in the Linux environment for storage systems and is proven in practice. It not only ensures the integrity of the stored data, but also supports almost unlimited storage capacity and very large files.

Open-E utilizes all of these advantages and supplements them with some useful enterprise features. An asynchronous replication (on & off-site data protection) locally or remotely enables a quick recovery of critical data in the event of a total failure, providing reliable protection against loss. Furthermore, a cluster configuration with synchronous mirroring is possible, as in our test.

JovianDSS is based on Linux and can therefore be seamlessly integrated into most existing IT structures. It is especially suited for users looking for a unified NAS

Furthermore, ZFS ensures that the file system is consistent at all times and thus no verification is necessary after a power failure, for instance. ZFS supports deduplication, compression and thin provisioning, which can be enabled and disabled in JovianDSS as needed.

Clean web GUI without dashboard

Though a keyboard and monitor can be connected to the appliance, it is most likely to be operated via the web GUI. The interface is very clearly structured with a seven-point menu on the left, but we would have liked to have seen a dashboard for the most essential performance and status data.

However, administrators can at least take a look at individual values like CPU, system and memory load under the last menu item "Diagnostics" and read out the network traffic as well as the disk usage separately for reading and writing. The latter can only be viewed individually for each drive and we searched in vain for an output of the current IOPS and latency.

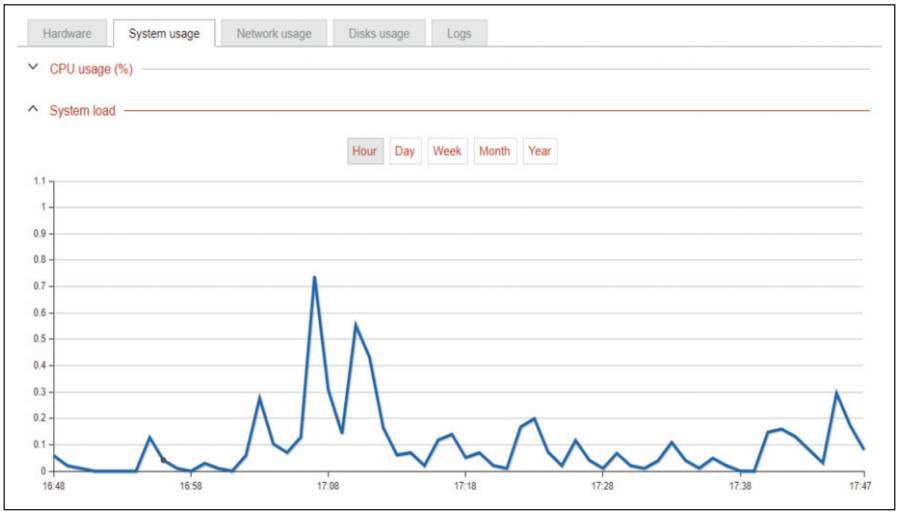

Figure 4: Die Angaben zur aktuellen Lastsituation sind bei JovianDSS recht unübersichtlich auf diverse Menüs verteilt und teils wenig aussagekräftig.

We would like to see a bit more of an overview here as it is difficult to determine the actual load.

Once at least one Zpool has been created as described above, volumes or datasets and shares for the actual external access must now be created, depending on the requirements and intended use. The RA11112 is multi-protocol capable, supporting iSCSI and Fibre Channel (FC) block protocols, as well as file-based NFS and SMB shares. FC usage requires that there is at least one FC host bus adapter in the device, which was not the case for us. We therefore opted for iSCSI volumes and NFS shares in the test to mount to the ESXi servers in our VMware vSphere environment.

When creating an iSCSI target, the wizard first suggests a typical default name. Then the administrator must create a volume (zvol), by default with thin provisioning, so that the space is not reserved. Such a volume can then be defined larger than the available space in the pool. Deduplication is possible but disabled by default. To control iSCSI access, the appliance supports CHAP, Mutual CHAP, and restriction based on IP addresses. In testing, the setup and subsequent connection to our three ESXi servers worked without a hitch. However, we could not configure a complete multipathing (MPIO) up to the servers due to the lack of a second 10-GBE switch.

File-based sharing possible

To define a file-based share instead of an iSCSI volume, a so-called dataset must be created. To control the size, JovianDSS supports both quotas and reservations. A quota describes a hard limit on how large a dataset can become and includes any snapshots. The reservation ensures that the specified space is guaranteed to be available. Therefore, if part of an active data pool is already occupied, a reservation cannot be created larger than the remaining free space. Snapshots and clones are also included in the reservation. Quotas and reservations can also be combined. Within the dataset, the administrator can now create a share, allowing access for NFS and/or SMB.

In case of SMB access, the NAS system allows authentication either via an Active Directory or LDAP as well as via local users and groups. Open guest access without authentication is also possible, as well as a restriction to read-only access. For an NFS share, access can be restricted to IP addresses with optionally read/write or read only access. As part of our test, we integrated the cluster into our Active Directory and defined some shares with different access rights on a trial basis, which worked smoothly.

Granular remote access

Administrative access to the appliance can be set in a pleasingly granular manner. For instance, HTTPS access to the GUI can be restricted to specific IP addresses. We were also able to password protect the local console and enable/disable REST access as well as SNMP and remote console (SSH) access to the command line (CLI). Logs can be forwarded to a log server and messages can be sent by e-mail.

The system supports manual snapshots for volumes and datasets in order to retain individual data statuses. To automate this, snapshot creation can be set up as a backup task. Here, administrators can define their own retention rules, i.e. at what intervals a snapshot should be created and how long it should then be retained.

Impressive mirror performance

To put our cluster through its paces performance-wise, we used a special Linux VM as a load generator with various operations. This VM under CentOS runs the Unix command dperf with various options. We primarily used the IOPS test here, which generates continuous random IO. To set up our test, we rolled out this VM twice each to three ESXi hosts, to which we assigned an iSCSI volume of the cluster, so that a total of six VMs accessed the nodes in parallel.

With this arrangement, we achieved 116,000 IOPS, which is a pleasingly good result. It should be noted that this performance was achieved with synchronous mirroring.

In the case of operating a single RA1112, the result should be slightly better. According to Open-E, the bonding with "only" 25 GBE is the bottleneck and Thomas-Krenn is considering possibly offering a 100-GBE variant at a slightly higher price. In our opinion, however, this is already a very high performance level and every prospective customer should determine their own requirements for availability and performance in advance to achieve the optimal price-performance ratio.

Summary

The RA1112 Metro Cluster from Thomas-Krenn is suitable as a high-availability storage system for SMEs that want to keep their data synchronized at two lo-

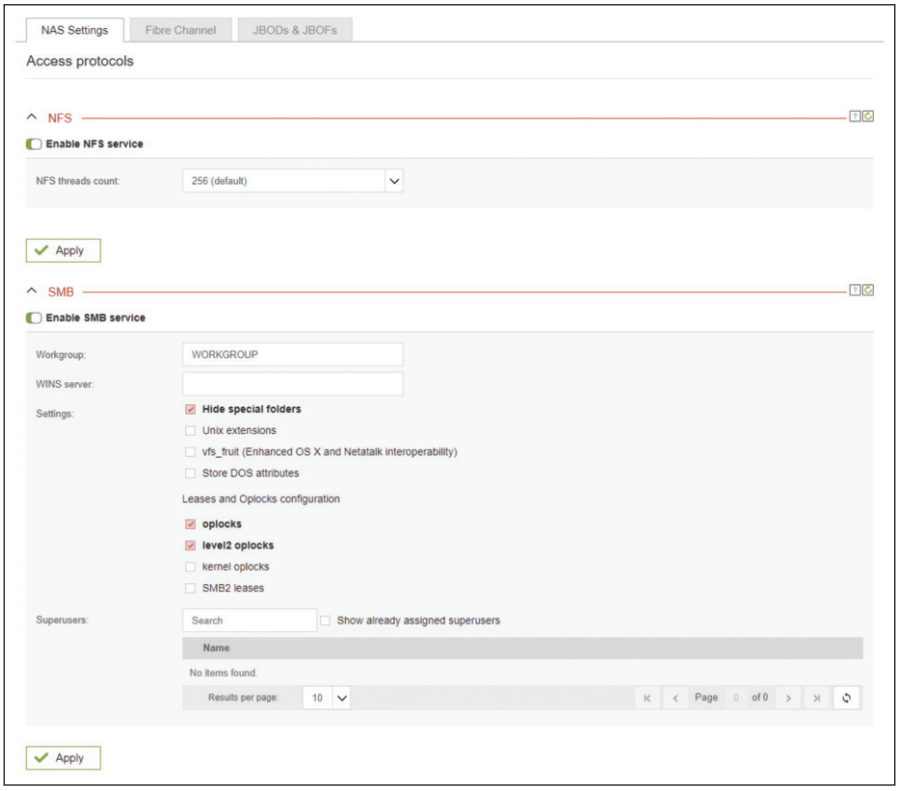

Figure 5: Instead of iSCSI, access via NFS or SMB can also be set up very easily.

cations. We see the main application primarily within virtualization environments under VMware vSphere, Citrix or KVM. Use in conjunction with Microsoft Hyper-V is also possible and certified, though Microsoft offers Windows-based alternatives that will likely integrate better into certain environments.

In the test, the system performed extremely well with up to 116,000 IOPS in synchronous mirroring. Open-E's Linux-based JovianDSS operating system is easy and intuitive to configure. However, we would like to see more concentrated information about the current workload within the WebGUI. In contrast, the cluster and failover status display is pleasantly tidy. With support for iSCSI, Fibre Channel, NFS and SMB, the system is versatile – both when working block-based or file-based.

The network connection can be configured individually at Thomas-Krenn, two 1-GBE ports as well as a management port for IPMI are on board ex works, while 10, 25 or even 100-GBE cards can be added individually. When sizing such clusters, we generally recommend working together with the manufacturer to ensure that the performance is adequate, any necessary expandability is provided and, ultimately, that the price is right. *(ln)*

The IT-Administrator verdict

Hardware design flexibility	8
Commissioning	8
Redundancy	9
Performance	9
Scalability	7

This product is

Optimally suited for use in SMEs as a highly available storage cluster for virtualization environments, especially under VMware vSphere.

Conditionally suited for use in combination with Hyper-V. The combination has been tested and confirmed, but a Windows-based storage platform would enable better integration in certain constellations

Not suited for low availability requirements that do not require cluster operation and can be covered with a simpler NAS system, possibly even with only one RA1112 node.