WHITEPAPER | ATTO

DISAGGREGATED STORAGE OVER ETHERNET USING THE ATTO XSTREAMCORE ET 8200 WITH A VIKING ENTERPRISE SOLUTIONS NDS-41020 JBOD

Historically, compute nodes and storage arrays have been closely associated with each other, by keeping the storage inside the server box. Extending storage has been achieved by aggregating a second enclosure with the server, typically using the SCSI protocol which has evolved from parallel SCSI to serially attached SCSI (SAS). The SCSI protocol has evolved to be a robust workhorse for data storage, allowing, for example, cascading storage or daisy chaining of several enclosures with up to 16256 individual devices and a cable length of up to 6.0 m.

Any of these iterations are still burdened with the functional aggregation of the storage enclosure with a server, adding cost for hardware acquisition, per CPU licenses and direct and indirect power consumption to the data center budget. On the other hand, the advantage of this kind of aggregate storage is that from a server perspective it doesn't matter whether the drives are in the server enclosure or whether they are attached through a JBOD enclosure since it is not the physical location of the drive but only the SAS controller that defines the feature set, like for example RAID or Hierarchical Storage management.

A recent trend in data centers is moving towards disaggregation of hardware as a first step towards software-defined hardware. In short, compute nodes are stripped off their direct attached storage arrays, which are moved into a separate physical space for storage only using a dedicated low power storage server as interface. The storage server in turn uses Ethernet or Fibre Channel to interface with the outside world or data-center internal compute nodes.

Disaggregating storage from the compute nodes in a data center does not eliminate the need for a conventional storage server with all aforementioned cost adders but it adds flexibility for the overall hardware deployment. Instead of having multiple instances of servers and storage arrays that at times are underutilized, the individual components (storage, servers) can be allocated on demand if a specific workload requires more resources.

STORAGE SERVER VS. PROTOCOL CONVERSION APPLIANCE

A conventional x86 based server running a Linux or Windows-based operating system and using a storage controller is certainly well suited as a hub for the attached storage. However, servers are not optimized for just providing a platform to tie together the SAS and Ethernet protocols to interface the hard disk or solid state drives with the outside world. In other words, depending on the use case, servers may be overkill and inflate the total cost of ownership. An alternative solution could be to customize the platform to completely eliminate the x86 environment and make it OS agnostic. One possible solution for this is to use an FPGA that embeds all the drivers for third party components such as IOCs and NICs and have a simple and streamlined protocol

conversion appliance. Custom features to add intelligence to the JBOD can be used to make the appliance suitable for advanced features such as hierarchical storage management.

In this context, it may be worth mentioning that it is not always necessary to support all SAS features in a single appliance. For example, the SAS specifications support up to 16256 target devices but for the purpose of an appliance, supporting the full number of drives according to the specifications does not add a tangible value.

ATTO Technology Inc.'s XstreamCORE® ET 8200 fits into this niche. In contrast to a standard storage server, the XstreamCORE ET 8200 is a protocol conversion appliance to bridge Ethernet to SAS at direct-attached speed while being compatible with any SAS-based storage regardless of whether the underlying hardware is using hard disk, solid state drives or tape in a JBOD, JBOF, RAID or Tape format. An added value feature is the Extended Copy (XCopy) data mover that allows it to internally copy data from one target device to another without involving the host. This feature can be used to enable hierarchical storage management (HSM) by shadowing data to an SSD within the JBOD and then deleting the SSD entry if the data are not requested within a specified period and cool off below a predefined threshold temperature.

HIERARCHICAL STORAGE MANAGEMENT

At the core of the XstreamCORE HSM capability is the Data Mover which acts without burdening the host interface with data traffic. Data are transferred from one drive to another using the Extended Copy SCSI command which is executed by the XstreamCORE. The host initiates the XCOPY transfer between drives by issuing the Extended Copy command which identifies the source and destination of the data within the array. The XstreamCORE processes the Extended Copy command, moves the data and reports back with the status. No additional host data traffic is required in this case, even though the host server still manages the file system.

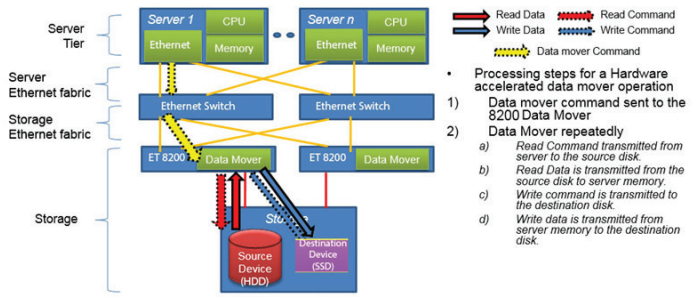

Figure 1: Local XCOPY of data between devices without server data traffic

The distribution, that is, the drive slot assignment of cache SSDs within the JBOD requires a modicum of consideration. Large JBODs often use a cascading expander architecture with the primary expander bridging the connections to the outside world with those to the secondary expanders. For optimal performance, the cache SSDs must be distributed across the secondary expanders in a ratio reflecting the number of SAS lanes between the primary and secondary expanders. Further, the cache SSDs must shadow copy the data to HDDs mapped to the secondary expanders in a mutually exclusive pattern. In other words, the shadow copy must not be assigned to an SSD mapped to the same secondary expander as the HDDs that are shadowed.

The total number of SSDs used for HSM caching depends on several factors:

- System-wide throughput bottlenecks
  - SAS connectivity
  - Ethernet connectivity
- Total capacity vs. cache capacity
- Cooling off of data vs. time

The XstreamCORE ET 8200 features two 40GbE Host Ports which add up to approximately 10 GB/s combined host bandwidth (line speed) or 6-8 GB/s usable payload bandwidth and 4 x 4 12Gb SAS 3.0 ports supporting an aggregate bandwidth of close to 19.6 GB/sec. Based on these data we estimate that 10 SAS SSDs with a throughput of approximately 1GB/s each may saturate the Ethernet pipeline of a single XstreamCORE ET 8200.

An important factor for HSM is the locality and temperature of data which will depend on the type of data. For example, web traffic may concentrate on 5% of available content, in

the case of IOT data, the estimated "signal to noise" ratio, that is, the relevant portion of data within repetitive sensor data is estimated to be less than 1%. This aligns well with a total number of 5-10% of cache devices within a hybrid HDD/SSD storage array to achieve a cache size of around 1% of the total array capacity. One example is the Viking Enterprise Solutions NDS-41020 SAS 12G JBOD. The NDS-41020 supports up to 102 3.5" HDDs in a 1m 19" rack. HDD capacity currently may be up to 14 TB/drive, resulting in a total storage capacity of over 1.4 PetaBytes if large capacity HDDs are used exclusively. Even if several drive slots are repurposed for SSD-based caching, it will still leave upwards of one PetaByte of raw capacity.

Of particular relevance for caching or tiering performance is the distribution of fast drives across the slot topology in order to avoid blocking of traffic. The architecture of the NDS-41020 uses three secondary expanders feeding into a single primary expander as the interface to the outside world. Consequently, the fast drives should be evenly distributed across the three secondary expanders, which lends itself to multiples of 3 SAS SSDs for optimal utilization of the internal SAS bandwidth. Depending on the usage model, therefore, we propose that the optimal number of SAS SSDs used for HSM in an NDS-41020 should be 6 or 9 at a capacity of 1.6 – 3.2 TB. Assuming 9 SSDs and 93 HDDs at a storage capacity of 14TB for each HDD and 1.6 TB for each SSD, this provides 1.3 PetaBytes of HDD storage cacheable in 14.4 TeraBytes of SSD media. Reducing the number of SSDs to 6 changes the ratio to 13.2 PB and 9.6 TB for HDDs and SSDs, respectively.

In the case of warm storage, a single hybrid storage array can be put as a head node in front of several additional HDD storage arrays as shown in Figure 2 below:

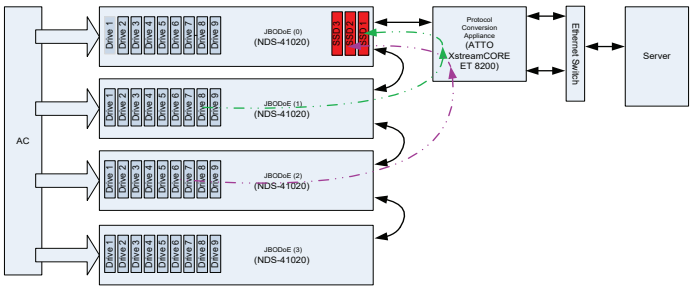

Figure 2: Hybrid Storage Array in front of daisy-chained HDD JBODs

Note that the movement of data (dotted lines) still needs to go through the SAS cabling from the JBOD arrays to the ET 8200 and then back to the SSDs in the hybrid appliance.

A different topology is shown in Figure 3 where multiple hybrid storage arrays are shown in a fan-out topology. It is understood that the daisy-chaining as shown in Figure 2 can be combined with the fan-out topology of Figure 3.

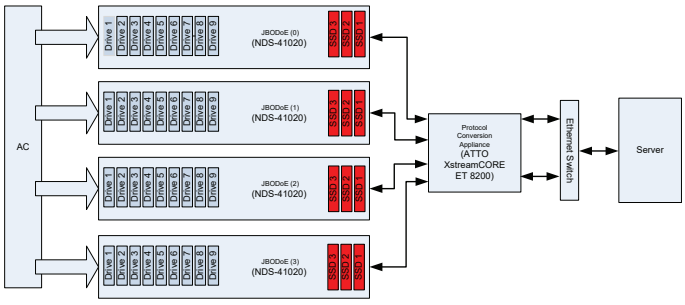

Figure 3: Hybrid Storage Arrays in a fan-out configuration

The two examples shown above are just two possible configurations out of a plethora of flexible options, which could also include dedicating a single NDS-41020 as an all flash appliance (JBOF) as a tier 0 array and using the XCopy command to move frequently requested data to the SSDs for accelerated access. This particular usage model may be especially suited for streaming media in front of a large audio-visual library.

XSTREAMCORE ET 8200 HARDWARE

The ATTO XstreamCORE ET 8200 storage controller is a high performance appliance designed to bridge 40Gb Ethernet to 12Gb SAS with consistent <2 microseconds of added latency when using iSER. Ethernet protocols supported are iSER (iSCSI Extensions for RDMA which uses RoCE v2, RDMA over Converged Ethernet) and iSCSI over TCP/IP. The XstreamCORE ET 8200 allows concurrent iSER and iSCSI connections to up to 64 hosts from up to 240 SAS/SATA flash SSD or hard disk drives.

Built upon the ATTO intelligent Bridging Architecture™ design philosophy that efficiently converts from SAS, or

SATA to other protocols. Designed with powerful hardware and advanced acceleration algorithms for negligible added latency, primary features that propel the 8200, performance wise, are hardware acceleration (xCORE™) and offload (eCORE™) processors.

The xCORE processor features multiple, parallel ATTO accelerator engines to efficiently route commands and data through the controller to provide unparalleled bridging performance. xCORE also features end to end I/O processing, hardware buffer allocation management and real-time performance and latency analytic capabilities.

The eCORE offload processor adds common, open storage services, handles reservations, storage routing and host and LUN mapping functions. eCORE also manages traffic for data mover offload functions with added error handling and diagnostic tools. eCORE handles various run time conditions, board health monitoring and various peripherals such as Ethernet and serial management ports.

XstreamCORE is available in a 1U 19" rackmount enclosure; complete with dual power supplies and cooling. The storage controller is designed to provide market leading performance while integrating Ethernet and Fibre Channel capabilities into midrange and enterprise-level SAS/SATA JBOD/JBOF/ RAID and tape and optical storage systems.

DISAGGREGATED RAID:

Overcoming the limitations of physical locality by moving from SAS to Ethernet allows for much greater levels of protection of data than any deployment of drives within the same physical location. For example, RAID arrays can now be created using drives in different enclosures that may or may not be located within the same data center. In more detail, as long as every drive can be exposed as an iSCSI target, it can be picked by any server and assigned to a RAID array. In other words, every enclosure can be part of as many RAID arrays as it contains iSCSI targets. Figure 4 below shows an example of such a configuration for additional protection. Eight JBODs have two redundant IO Modules and each IOM is connected to an accelerated protocol conversion appliance (XstreamCORE ET 8200). The XstreamCORE ET 8200 are connected to an Ethernet switch fabric (for simplicity purposes, only a single switch is shown), which provides the interface to a number of servers (Server A-C).

BASE CONFIGURATION (COLD STORAGE)

In the base configuration, the SAS ports of two XstreamCORE ET 8200 storage controller appliances are split with 8 SAS ports each between two JBOD enclosures in a cross-bar configuration to allow establish a fully redundant data path. Additional JBOD enclosure can be daisy-chained to the remaining 8 SAS ports of the first JBOD.

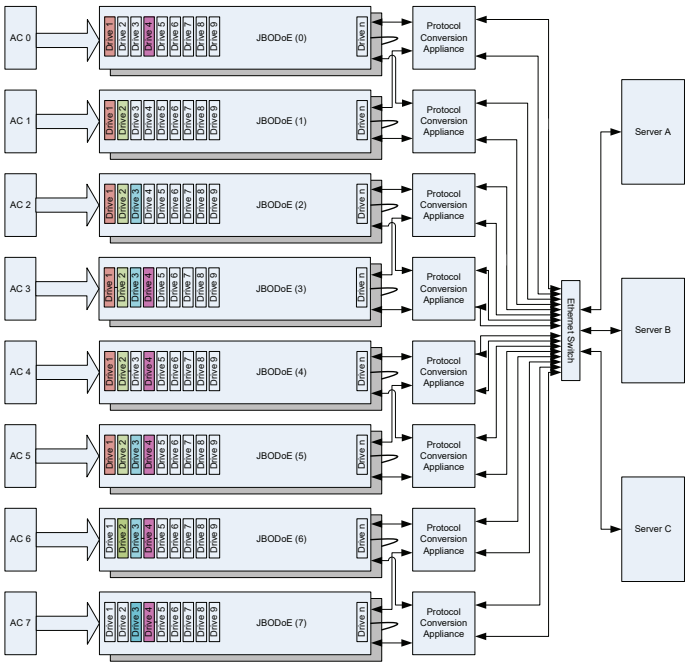

Figure 4: Disaggregated RAID over Ethernet (Base Configuration)

Each IOM uses 8 SAS lanes to connect to the XstreamCORE ET 8200 and an additional 8 lanes to connect to a second daisy-chained JBOD (Cold Storage). Up to 10 JBODs with 102 drives each can be supported by each XstreamCORE ET 8200 in a 42U rack

HIGH PERFORMANCE CONFIGURATION

In a high-performance configuration, each JBOD is connected to two XstreamCORE ET 8200. The SAS ports of both XstreamCORE ET 8200 storage controller appliances can be connected to one IOM each or split between the two IOMs. The second option may provide slightly better performance by alleviating the potential bottleneck between the primary and the secondary SAS expanders. Depending on the throughput required, some SAS lanes may be allocated for daisy-chaining of additional enclosures.

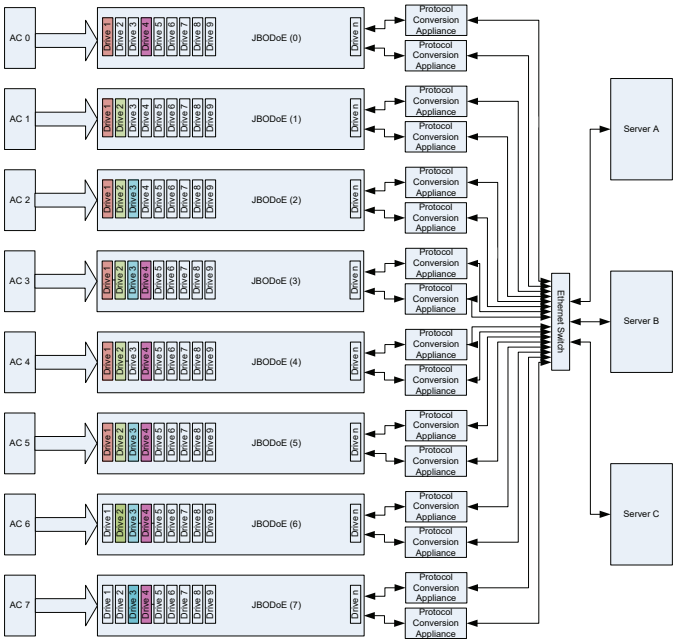

Figure 5: Disaggregated RAID over Ethernet
(High Performance Configuration)
Full BW is achieved by attaching one XstreamCORE ET 8200
to each IOM for a total of 32 SAS lanes and 4 x 40GbE
Ethernet connectivity to the Ethernet switch fabric

Power Failure Resilience

Each JBOD is hooked up to its own AC power node and the drives are configured to be individual iSCSI targets. The majority of drives is allocated to a number of RAID groups which are distributed across the enclosures so that no enclosure has more than one drive of any RAID group. Different RAID groups are not confined to the same enclosures but can span over different groups of enclosures to provide additional resilience against for example AC power loss at any given enclosure.

Overprovisioning

Each JBOD has a small number of spare drives that can be allocated on the fly as a replacement for any failing drive without physically servicing the enclosure. Service can be scheduled if a drive is failing but there is no immediate requirement for service. Any replacement drive is added to the overprovisioning pool and can be cycled into any RAID group as replacement drive if needed.

Enclosure Requirements:

An absolute condition sine qua non for this type of enclosure is high availability of all critical components in the data path. That is, all drives have to be accessible through two different paths, which requires two physically different IO modules (IOMs), redundant power supply units – attached to separate AC sources as well as redundant 5V regulator units.

Access Path Requirements:

In addition, the access path needs to be fully redundant, that is, two independent accelerated protocol translation appliances are required that are connected to one of the IOMs each. For full redundancy also the upstream path needs to have full redundancy, including duplicated Ethernet switches with separate connections to the data center backbone.

On-the-fly Addition of Spares in a RAID Array after Drive Failure:

In a distributed RAID array, that is the array is spanned across multiple physical enclosures, two different failure scenarios are possible:

- Power failure to the enclosure
- Failure of a single drive within an enclosure

In the first scenario, the entire enclosure goes off line while the array still continues to operate but eventually it comes back on-line. If the operating system kernel (MDRAID) does not understand this type of error condition, then the existing data on the drive are essentially useless and the entire drive needs to be rebuilt. In contrast, if for example the ZFS file system is used, it can start a resilvering operation to restore only the blocks that were written after operation went down.

In the second scenario, one of the drives needs to be replaced and the array needs to be rebuilt. Since each drive in the enclosure is an iSCSI target, any overprovisioned or spare drive in the enclosure can be immediately substituted for the failed drive and the rebuilding of the array can be initiated. Depending on the specific environments and needs, different strategies may be employed:

- Create an immediate back-up of the remaining drives and use one of the copies to rebuild the array while keeping the rest of the array on-line. Optionally, all drives can be marked as Read-Only until the rebuilding of the array is complete after which the drives are erased

and recycled into the overprovisioning/spare pool.

- Since the array is "software-defined over Ethernet" it may be disassociated from the original storage server and a dedicated "rebuild server" can be associated on-demand with the copy used for rebuilding of the array.
- If the failing drive is partially recoverable, copy the recoverable data to a spare drive (Drive s+1) in Figure 2 and use a ZFS resilvering operation to determine what part of the data need to be rebuilt using the drives in JBODs 1-n

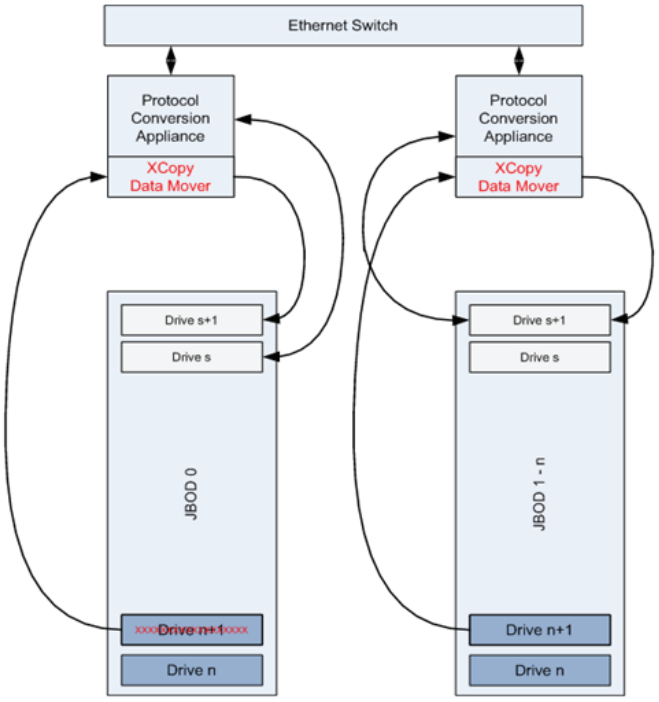

Figure 6: Rebuilding a failed drive
after partial recovery using ZFS resilvering

Dedicated On-demand Rebuild Servers

After failing of a single drive, for additional protection, the contents of the remaining drives in the RAID group may be backed up to a spare drive within the same enclosure and either copy may be used to rebuild the array while the second copy is used to resume normal operation. Because the array is software defined to a server, a second software defined server may be tasked with rebuilding the array while the original server maintains provides access to the data stored in the RAID group. This scenario may cause data coherence problems, that may be resolved for example with a ZFS resilvering operation. Alternatively, the RAID group used for data access may be flagged as Read-Only until the second copy of the array has been completely rebuilt, after which the drives may be erased and recycled into the overprovisioning pool.

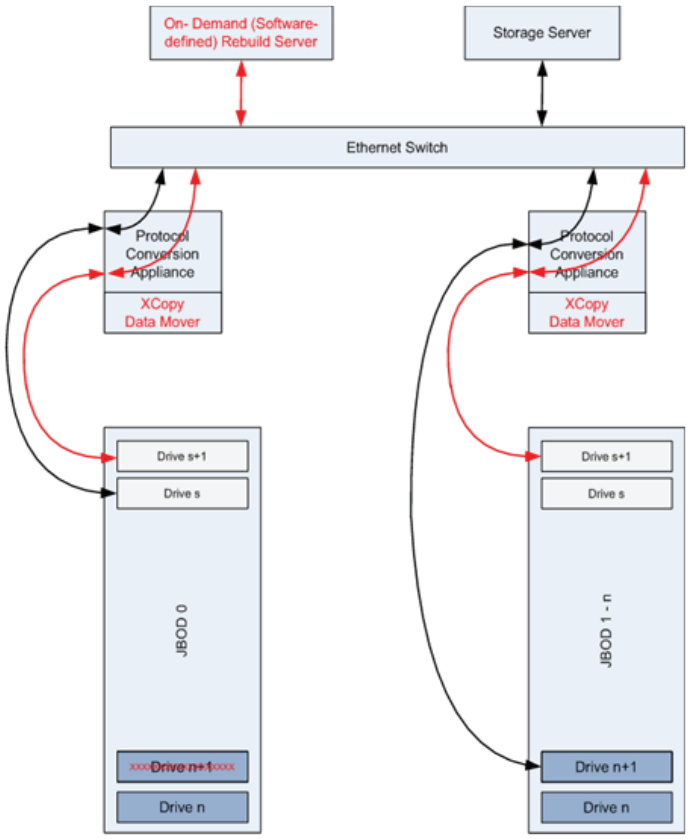

Figure 7: Back-up and restore
using a software-defined "Rebuild Server"

CONCLUSION:

Accelerated protocol conversion appliances connected to large and space-efficient JBOD enclosures such as the Viking Enterprise Solutions NDS-41020 with 102 large form factor drives in a standard 1m rack 4U enclosure may emerge as an interesting tool for software-defined storage over fabric. Certain limitations such as added latency may slightly attenuate performance but features like XCopy / Datamover / Redfish management support and planned integration of Swordfish extensions add some very interesting perspectives to HSM. Specifically the host-independent copying of data from the main array to a lower tier as a local cache for hot data can significantly speed up access to frequently used data and at the same time allow for power saving by spinning down drives that are not accessed over a certain period of time. Additional advantages could be the increased

resiliency against power failure distributing RAID or Erasure Coding across multiple enclosures that may be separated into different racks, different parts of the same data center or even different geographical locations. The advantage of reclaiming partial data from failing drives by using a resilvering operation to speed up the rebuilding of the array will foster the more widespread adaptation of advanced file systems like ZFS with integrated logical volume manager.

Defining each drive as iSCSI target also allows for overprovisioning of the enclosures and cycling any spare drive into an array on demand, either for intermediate back-up or rebuilding of a damaged array.

In addition to having software-defined storage, disaggregation also allows for software-definition of dedicated rebuild virtualized servers.

Taking advantage of the capabilities will require some amount of rethinking of storage, including the migration to newer, more intelligent file systems that are capable of recovering fragments of data from a failing drive and use those as basis for an accelerated rebuild of the array using spares. Also, component failure needs to become a "feature" with the appropriate fail-safe mechanisms in place to avoid unnecessary service calls, or eliminate them altogether because of the availability of overprovisioned hardware – at least for the scheduled life span of the "zero maintenance" storage appliance.

NETWORK FUNCTION VIRTUALIZATION

A growing trend in the IT industry is the move towards Network Function Virtualization (NFV), taking software defined networking (SDN) to the next level towards a lower cost network infrastructure with more agility. SAS over Ethernet greatly facilitates the deployment of already existing hardware, including storage and servers by defining each virtual device as a network function that can be allocated on demand. This is where Ethernet-attached storage appliances like the above described arrays of Viking Enterprise Solutions NDS-41020 behind the ATTO XstreamCORE ET 8200 with their built-in flexibility epitomize the vantage for an agile and flexible infrastructure with very little slack and best resource utilization resulting in best TCO possible.

VIKING™
Enterprise Solutions

Global Locations

US Headquarters	*Colorado Research Ctr*	*European Sales*	*Software Development Ctr*
2700 N. First Street San Jose, CA 95134 U.S.A. Toll Free: +1 855 639 7838 Main: +1 408 964 3730	6385 Mark Dabling Blvd Colorado Springs, CO 80918 Main: +1 719 266 5398	Sanmina-SCI Holding GmbH & Co.KG Lerchenstr. 1 91710 Gunzenhausen, Germany +49 89 14010707 (UK)	University Technology Centre, Building 2, Curraheen Rd., Cork, T12 NY5T.

For sales information, email us at sales@vikingenterprise.com, or visit our website for all global locations and contact information.

www.vikingenterprisesolutions.com

R 1.01