Breakthrough in Nearline Disk Storage: Spectra ArcticBlue

October 12, 2015

V1.0

800-833-1132 • 303-449-6400 • (Fax) 303-939-8844 • 6285 Lookout Road • Boulder, CO 80301 USA • SpectraLogic.com

Contents

Abstract

As data grows beyond normal human comprehension and the value of that data also continues to increase, the demand to store it all for longer and longer periods of time is exploding. Traditional methods of data storage continue to slowly evolve, but not at the pace required to keep up with data growth. Predictions range up to the tens of Zetabytes over the next 20 years and new revolutionary technologies are required to keep up with that demand.

Companies are faced with not only huge costs of storage but the prospect of building larger data centers, adding more and more power, and staffing adds both complexity and burden to businesses. Cloud options exist and are gaining popularity, but businesses are also faced with threats from every angle seeking to hack, destroy, access, and exploit data and most companies are choosing to maintain some or all of their critical data in-house with some copies completely off-line for further protection.

At the same time, data needs to be more accessible across nearly every market segment with the speed of access to data frequently rivaling the importance of protection. Traditional offline storage methods clearly have both the cost and safety advantage, but suffer from the online accessibility required for many applications.

Digital Preservation is an all-encompassing concept that covers every aspect of protecting and preserving data including, but not limited to:

- Fingerprinting
- CRC/error checking
- End-to-end CRC
- System replication-redundancy
- Media redundancy/RAID
- No single points of failure
- Genetic Diversity (no single media types, locations)
- Migrating to next storage technology (or higher density systems)

Of course, the critical demand is to do all this cost effectively with simple integration into existing workflows, with reasonable retrieval performance, and with monitoring that guarantees the data is truly safe long term.

Introduction

Traditionally, data is stored in tiers; sometimes in as many as 5 or 6 with complicated data movers responsible for categorization, deduplication, organizing, and storing on the most appropriate tier to meet both performance and cost goals. At the same time, the entry of cloud storage as a viable alternative has caused organizations to also consider where these services play while also diving into the value of RESTful interfaces, like S3, with object storage behind it. While every top level software package already knows how to save to a file system, that is not always the most efficient methodology for the deeper storage tiers.

Logically, disk, due to the higher costs involved, is at the top of the storage tier classes bested only by flash based SSD, and tape is used exclusively for long term storage and disaster recovery.

The question at hand, then, is what happens if disk can be inserted into the storage system at nearly the cost and longevity of tape and at the same time be tightly linked to deeper tiers of storage with a low cost, no middleman, system-level approach to true Digital Preservation. As storage architectures simplify, it becomes cost effective to have multiple copies on different types of media for Genetic Diversity. This is a storage concept that mimics what happens in nature in that diversity provides guaranteed protection. The simplest example might be a hacking attack. While a wildly successful hacker might be able to corrupt any online data copy on disk or even tape, an eject copy of the data in an un-connected cold storage location would not initially be affected.

Incremental storage density and cost improvements have been interesting, but an inflection point has been reached by applying new storage technologies to this problem and the direct result is ArcticBlue. The use of Shingled Magnetic Recording [SMR] drives, applying Data Lifecycle Management with intelligent idling and power down technology, and front ended by an industry standard S3 interface bundled with Spectra's 36 year focus on Digital Preservation all come together in a single product that meets all these diverse needs.

At a glance, ArcticBlue:

- Provides Object storage for managing both structured and unstructured data on nearline fast access disk
- Simple scalability, over 6.1PB per rack with fully integrated tape copy as an even more affordable tier - in the zettabyte scale
- Cost effective Power Down SMR-SATA disk storage, allowing lowest initial cost of $.10/GB USD.
- Appliance based BlackPearl controller with fast cache allows use of SMR drives efficiently – unlike other software only solutions
- 1GB/s or more throughput plus if data is stored on a powered down drive, then less than 30 seconds to first byte with random access thereafter from all bands
- Power down capability uses up to 75% less power and more than doubles disk life
- Tightly controlled drive monitoring and alerts combined with simple self-service options to secure data and keep costs down
- 1 in 2 million years probability of data loss on disk combined with an optional tape copy results in a nearly unmeasurable "nines" of reliability
- Industry standard S3 front end interface for simple integration
- Specifically designed to be part of a storage ecosystem – BlackPearl offers seamless storage tiers that are policy controlled and temporal in nature – completely eliminating the need for third party data movers while maintaining support for diverse and multi-copy media across multiple sites
- When added to a tape system, upgrades that system to be able to use unmodified S3 commands to archive to tape – a truly new industry offering

Spectra ArcticBlue Snapshot

- **4U Master Node,** includes system controller with BlackPearl Software: a large cache of 4TB SAS HDDs standard (8TB optional) and an object storage database on mirrord SSD
- **4U Expansion Node**, holds up to 96 8TB SMR drives each, expansion node connects to the master node using dual external SAS cabling
  - The 4U master accepts up to 8 expansion nodes, for over 6.1PB raw uncompressed configurations
  - Minimum of two bands per master node system (allowing intelligent power down of bands)

- **Drives**
  - Shingled Magnetic Recording [SMR]
  - 8 TB raw each in 20+3 (triple parity) bands

Physical Characteristics

- Dimensions:
  - 4U Master node: 7” H x 19” W x 29.5” D
    178 mm H x 437 mm W x 699 mm D
  - 4U Expansion node: 6.9” H x 19”W x 40” D
    175 mm H x 483 mm W x 1016 mm D
- Weights (approximate):
  - 4U master node with 12 drives: 120.2 lb. (54.5 kg);
  - 4U expansion node with 96 drives: 248.8 lb. (112.9 kg)

Power

- 4U Master node: 1280W Redundant Power Supplies
  - 750W typical R/W
- 4U Expansion node: 1100W Redundant Power Supplies
  - 775W typical R/W and 140W idle (with DLM)

Simple Pricing

- 4U Master Node lists for $31K USD
- 2U Master Node lists for $19K USD
- Expansion nodes (full) list for 10¢/GB ($77K USD)

Spectra ArcticBlue Snapshot

High-Availability Features

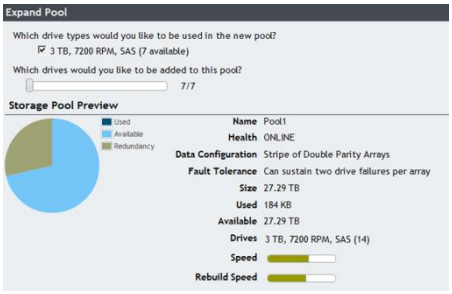
- Redundant, hot-pluggable cooling system
- Redundant power supplies, each with its own AC power connector and an LED to indicate status
- Hot-swap/spare drives
- Redundant boot drives for the operating system, object storage system, logical volume manager, S3 interface, and SNMP servers, and the Spectra StrataView web interface
- Mirrored SSD, Object Storage Database

Simple to Use

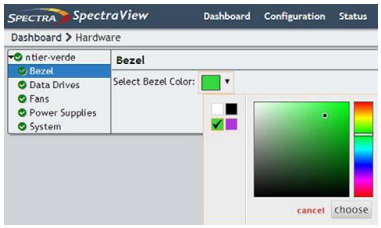
- Buttons: Power on/off, system reset
- LEDs: power, hard drive activity, network activity (x2), system overheat, power fail
- Direct integration and configuration with Spectra Tape Libraries thorugh integrated BlackPearl software, where default and most common data policies are auto configured and ready for production

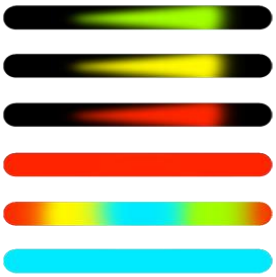
Interface

- Graphical interface for ease of use. You can use the web interface to manage ArcticBlue through an Internet browser.
- Command line interface for flexibility



Bezel with Status Light Bar: The Visual Status Beacon light bar in the front bezel provides an at-a-glance status of the array. The light bar changes color to indicate the status of the array. If a BlackPearl system requires attention, the lit bar, or beacon, helps administrators identify the unit quickly. Customize these colors using the web interface.





Access Protocols

- Spectra S3 API
- Many AWS S3 API calls

Spectra ArcticBlue Snapshot

Network Monitoring and Configuration Support

- DHCP
- SNMP
- SMTP
- NTP

Data Transfer Specifications

Full network bandwidth use in high-performance setup, using one of the following network data transfer connections from hosts to:

4U BlackPearl master node:

- One 10 Gig-Base-t copper port
- One 10 GigE optical/SFP+ ports
  - Aggregate both 10 GigE ports for 20GigE network connection
- Optional upgrade to a 40GigE network card instead of the included 10GigE
  - Aggregate both 40 GigE ports for 80GigE network conection
- Optional updade to 10Gbase-T network card instead of included 10GigE (SFP+)
  - Aggregate only two 10Gbase-T ports on network card for 20Gb connection

Additional 4U Master Node Ports

- 1/10 Gbase-T (RJ45) port: BlackPearl-ArcticBlue management connection to the browser-based web interface (not used for data traffic)
- Adding expansion nodes requires an HBA card to be installed, which has SAS ports used to connect to one 4U ArcticBlue expansion node
- Adding tape attach HBA card to the master node to connect a tape library to the BlackPearl ecosystem (either FC or SAS HBA)

Monitoring

- Hardware status
- System messages
- Remote access
- SNMP client support
- Thermal monitoring
- Email notification when issues arise
  - Log collection

Support

- 1-year warranty for hardware and software
- AutoSupport feature
- Recovery from failed drives with efficient automatic rebuilds
- On-site professional services available
- Support service options:
  - Basic phone support with standard shipping for replacement parts
  - Upgrade to 24x7 phone support
  - Next Business Day On-Site
  - Next Business Day On-Site with 24x7 phone support
  - Four-Hour On-Site with 24x7 phone support and price lock

Spectra ArcticBlue Snapshot

Additional features supported in ArcticBlue disk systems:

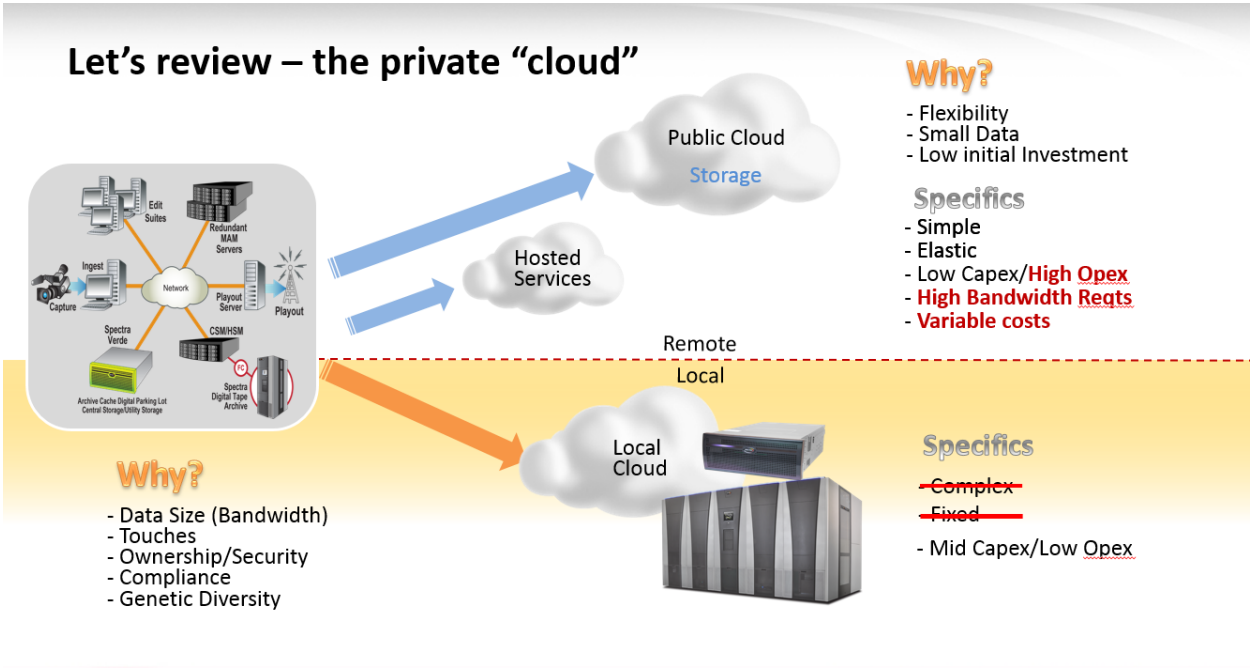
- **Data integrity** – Advanced checksums protect against undetected errors and result in a much better bit error rate than traditional disk. Checksums can be enabled all the way from the original object down to the final saved media.
- **Triple parity** – Allowing up to 3 drive failures per Very Wide Band group without data loss. This enhances reliability when working with large data sets.
- **Compression** – Compression is always turned on, and we are using LZ4 compression which can compress to roughly 2:1 without performance impact - stats
- **On-demand integrity check** – ArcticBlue features an on-demand and scheduled data integrity check for data drives configured per storage pools. The check scans the drives for data corruption and corrects any errors found. Since each pool is managed independently just one band could be powered on for a integrity check.
- **Intelligent rebuilds** – Instead of rebuilding an entire failed drive, ArcticBlue rebuilds only the portion of the drive that held data, potentially saving hours on rebuilds. Additionally Drive Lifecycle Management can have other band(s) powered on for data I/O as the pools is rebuilding – reducing drive use and further increasing rebuild times.
- **Redundant power** – Each ArcticBlue node ships with two high-efficiency redundant power supplies.
- **Performance monitoring –** The web interface lets you view the performance of pools, hard drives, tape drives, CPUs, and the network.
- **Global spare drives** – A drive that is not configured in a storage pool acts as a global spare drive. Each band of 24 drives (192TB) includes 1 global spare. If a drive failure occurs on the ArcticBlue, the array immediately activates a global spare. When the failed drive is replaced, the replacement drive now acts as the global spare. Spectra Logic recommends having four hot spare drives per ArcticBlue node in your array to permit immediate rebuilds of a pool in the case of data drive failure.
- **Hot-swappable data drives** – ArcticBlue's drives are easily inserted into the expansion unit, without tools, by on-site data center staff.
- **AutoSupport** – Automatically contact mail recipients upon generation of messages. Also auto generate logs for Spectra Logic Technical Support.

Data growth rates…

The Digital Universe is expected to double every two years growing from 4.4 ZB in 2013 to 44 ZB in 2020. With nearly as many digital bits as there are stars in the universe, data growth is expanding as fast as the cosmos. By 2020, it is expected that nearly 60% of the data created will come from emerging markets countries, 40% of all data will be touched by the cloud in some way (stored or processed), and 10% of the data will be generated by the Internet of Things.

Data retention requirements can come from many sources including legal requirements or corporate governance. Outside of regulatory reasons for storing data for long term there is another reason, the unknown value of the data being stored.

Public vs. Private Cloud – what do I really need?



It's logical to begin with the question of whether local storage is needed or not. If we consider cloud storage in a generic sense, there are clearly some advantages including flexibility, low initial costs, simple wide area access, and ease of setup for small usage models. For a small business with very little data, backing up only a few computers or sharing personal files with multiple computers, there really is no better solution than some of the free or very low cost online storage. The cloud is also an excellent place for a third copy of data for disaster recovery even for larger data sets, as long as the time it takes to download from the cloud is reasonable cost, given available bandwidth.

The obvious question, then, is why not use cloud storage for everything. Organizations have different driving priorities but the two most prevalent are cost and ownership.

It seems counterintuitive that cost would be a driving factor in considering local vs. cloud storage since most entry level packages are free or very low cost. New entrants to this market have publicly advertised 1c/GB per month and some even a fraction thereof, but there is always more to the story. As data grows, that seemingly small cost adds up quickly, and the added variable costs for retrieval or even moving data grow exponentially. On top of all that is the bandwidth required to move data. For data that needs to be accessed at all, it has to be local for performance, but even moderate amounts of backup completely swamp typical WAN speeds and require pipe upgrades. Even a moderate 50TB backup that must complete within a weekend window quickly drives bandwidth costs well over $100K/month. A specific whitepaper comparing cloud bandwidth usage costs to private cloud total system costs is available on the Spectra Logic website.

Rather than diving into the pros and cons of physical ownership, let's just pose some questions and carefully consider the responses for your particular organization.

- Who really has access to your data
- Perhaps more importantly, who is really in control of it
- If you ever stop paying your bill what happens to your data
- Is it possible for access to your data to be denied due to court ruling or government policy

Clearly local access has its advantages. The question then shifts to what is the most economical and secure local storage available. For long term Deep Storage, currently nothing competes with object storage on tape libraries, but for data that needs to be accessed and used as part of an active workflow, a copy on ArcticBlue nearline storage is unquestionably the right choice.

Bulk Nearline storage to Meet Data's Mid-Life *(& long-aged)* Demands

In data's mid-life phase, the data needs to remain accessible but is too infrequently accessed to leave on high-performance, expensive disk. This longer term storage phase warrants its own specialized storage platform. This platform needs to protect data against the challenges of longer term storage, including accessibility, affordability, scalability and concerns surrounding the risk of data corruption.

Spectra Logic is now exceeding this challenge with the introduction of ArcticBlue, a type of storage that is designed specifically to address these requirements at ground breaking costs. This storage is affordable and easily used and scaled. With ArcticBlue, sites can handle data that continues to expand at dizzying rates while remaining confident in the data's integrity.

Purpose-Built nearline storage: ArcticBlue

Spectra Logic designed ArcticBlue to address the specific requirements of digital preservation, which are:

Spectra T680 library and ArcticBlue System

- Storing data affordably - ArcticBlue does this through a modular design with parts that can be easily swapped and upgraded in place to preserve initial investment. The disk also reduces data center overhead with its high density capacity, power density and small footprint requirements. Power down capabilities extend system life reducing overall cost of ownership.
- Easy to install, maintain, scale, and protect - ArcticBlue provides an unparalleled benefits to the BlackPearl ecosystem with ease of use in maintaining, expanding, and upgrading disk systems.
- Preserving data integrity - ArcticBlue systems preserve data integrity by providing multiple levels of integrity checking beyond that found in typical disk systems.
- ArcticBlue leverages BlackPearl hybrid storage architecture providing a high speed, large cache, and native tape out creating a seamless redundant second copy on tape automatically all with a simple, efficient and high speed S3 (REST) interface for increased performance.

Disk Technologies

It is important to split our descriptors of disk technology into three distinct categories: interface (SAS or SATA), drive type (consumer or enterprise), and recording technology (PMR, GMR, SMR, etc). Although we commonly use just one of these to refer to the drive, they are each becoming independent items that no longer necessarily imply the other two. In the past, SAS (Serial Attached SCSI) drives were used exclusively in the enterprise and were built to last with extra mechanical support of the drive platters and long wearing components. Likewise, SATA (Serial ATA) drives have traditionally been consumer drives with only single point spindle attachments and lower end components. The SAS interface actually has other advantages such as dual port capabilities, the ability to put a semi-fabric like system together where many drives are on the same controller via switches, and other speed enhancements.

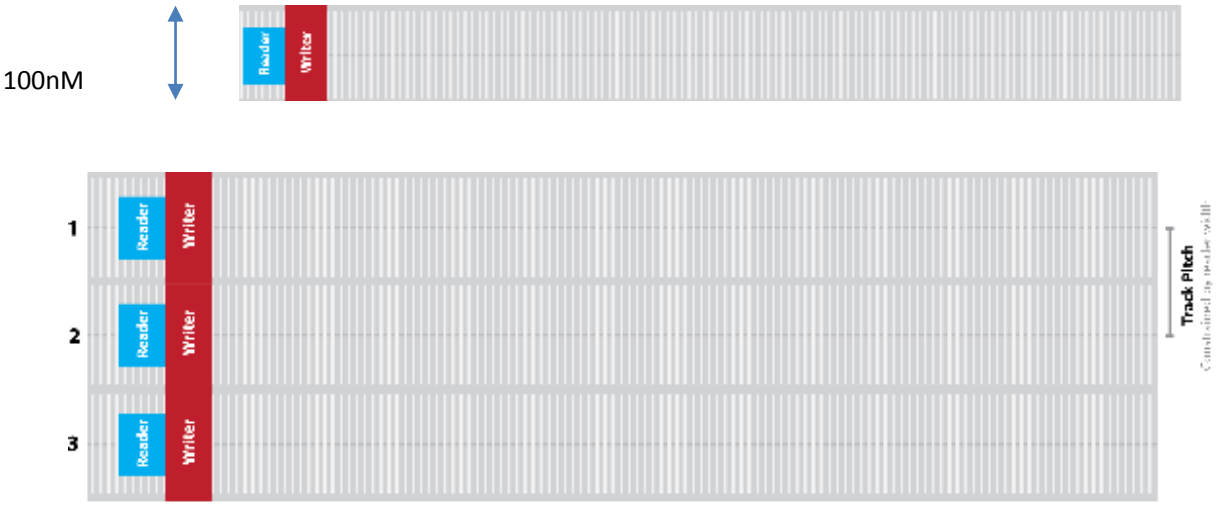
SATA, however, no longer directly implies a consumer grade drive as manufacturers have begun blurring the lines between the different technologies. At the same time, new recording technologies, such as how a magnetic “head” actually records bits onto the surface of a metallic platter, are also changing rapidly. The race for “areal density”, or how many bits can be fit into an area on a platter, is becoming aggressive but has recently not been able to achieve the same magnitude of advancement as the processor companies have. Disks have always had a lot of levers to pull to increase capacity from smaller read/write heads, to adding more platters, to using different recording technologies, but each new technology has only so much to offer and when it maxes out, a technology change must occur.

Pioneering new SMR technology

Shingled Magnetic Recording (SMR) is one of the latest technologies to come to the disk drive market. To properly understand this new technology jump, we need to first clarify how current, let's call them traditional, hard drives record onto a platter.

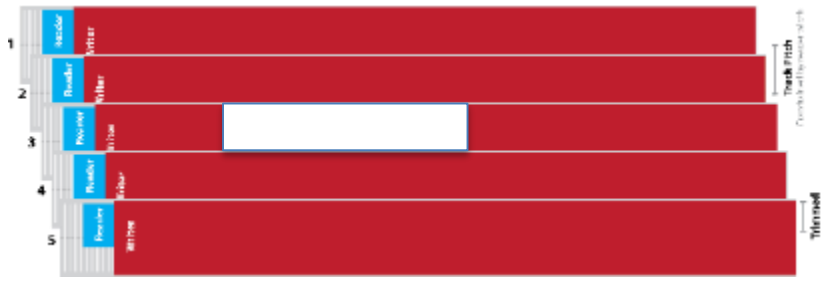
If you remember from elementary school, making an electromagnet by wrapping wire around a nail and then attaching a battery, you have the basic structure of a disk "head". Energizing a miniature coil on the end of an arm that "flies" over a metallic platter, much like the arm of a record player, creates a magnetic field that, when pulsed, can flip the orientation of little magnetic particles on the platter to record data on that platter. Just like a record player, data is recorded in concentric rings on the platter. This is on an incredibly small scale with each "track" traditionally being around 100-200nM wide; a human hair is typically around 100,000nM to give you an idea of the scale.

Traditional hard drives typically record data in concentric rings with small empty spaces in between to make sure that data from one track does not overwrite the next.



An interesting law of physics is that it takes more space to write a track than to read one mostly because the larger magnetic force required to flip magnetic particles make a larger magnetic field than it takes to just sense the orientation of particles and read them. So, write heads are "larger" than read heads. If there were some way to make written tracks smaller, you could still read them and you could fit more data on a platter. Shingled Magnetic Recording is one way to make this happen. The write heads are still the same size but now, with each successive track we partially over-write the previous track. The analogy is like shingles on the roof of a house where each layer partially overlaps the layer under it. You can still see each individual shingle but not all of it. That is really where the analogy stops since a write head actually overwrites (destroys) part of the track above.

Notice that the read head simply reads the non-overlapped part of each track. This recovers as much as 25% of space. Unfortunately, there is more to the story. Since each track is partially overwritten by the next track, it becomes more difficult to fill empty spaces.



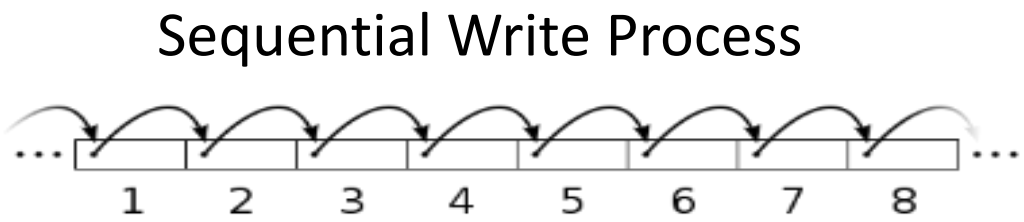
If we simply put the head in that gap and overwrite, it will destroy part of the next track, already full of data, as well.

Obviously, these drives need to be written to sequentially, in large data groups. They need to be carefully controlled, with deletes and defragmenting handled very carefully. Because of this, these are not normal hard drives that you can just substitute for traditional hard drives in any application. Controllers, including hardware RAID controllers, which write to drives randomly can very quickly bring

these drives to their knees, but there are huge benefits as well. Since you can fit more data in a given area you can store more data at a lower cost, and since they are harder to use, the drive manufacturers are incentivizing the industry, through lower costs, to implement controllers for them specifically.

So two important points:

- Hard drives that like to be written to sequentially in large data groups and changed infrequently, is exactly how you would describe a tape library system. It took a tape company to make these disk drives work to their fullest.
- Hardware RAID systems are incompatible with this technology and it takes special software control to make them work effectively



Value-Feature

Together, this explains why Spectra's ArcticBlue software RAID controller is the perfect system to use this new technology. The use of Spectra's implementation of the ZFS software file system allows ArcticBlue to automatically force data to be written sequentially by using a large cache and Copy on Write technology. Additionally with all ArcticBlue data being cached in BlackPearl SAS HDD cache pool first, the cache actually aggregates more data (S3 PUTs) together before writing it all out to the software raid system.

Digital Preservation:

For most systems, the story would end there. 10¢/GB is compelling enough on its own that the product is already revolutionary, but Spectra has a long history of protecting and preserving data for our customers in the tape library market. To that end, ArcticBlue employs a 5 tier Digital Preservation Strategy to ensure your data is safe – forever.

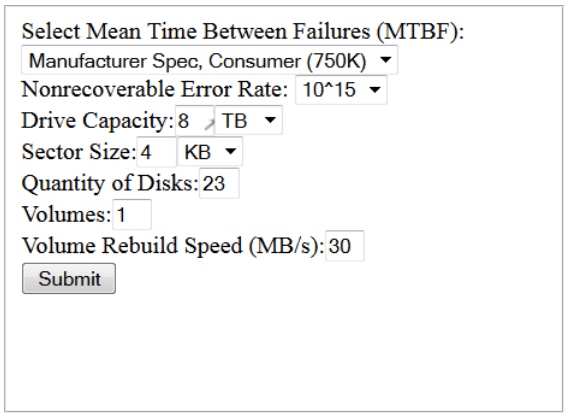
Local Erasure Coding – Probability of data loss less than 1 in 2 Million years

Local Erasure Coding uses technology similar to RAID with calculated parity bits saved as part of each band, providing redundancy without the overhead of mirroring. Unlike pooled RAID stripes in Verde DPE, data is striped only within a band (a single stripe band) in order to allow power down without disrupting data availability. Even with this modification, ArcticBlue's Local Erasure Coding provides the same level of protection by allowing up to three drive failures while providing continuous access to data.

A triple parity system, statistically, will only lose data once in over 2 Million years for the drive type and band size used in ArcticBlue when properly monitored and maintained.

This kind of probability can be easily calculated on any one of a number of online calculators used if we know the MTBF (800K), non-recoverable error rate (10^15), 8TB, 23 band, and a rebuild speed of

between 30 and 50MB/s. Basically the question is, if a drive fails, what is the probability that a second drive fails during that short rebuild window, and in fact a third and fourth drive also fail in that specific time. The answer is, a VERY long time.



RAID Level	Formatted Capacity (GB)	Mean Time To Data Failure (MTTDF) in hours	Bit Error Rate MTTDL	Mean Time To Data Loss (MTTDL) in hours	MTTDL (Years)
RAID 0	173,342.72	32,608.70	< .01	16,304.35	1.86
RAID 1	86,671.36	684,495,684.33	< .01	342,247,842.16	39,069.39
RAID 5	165,806.08	707,123.64	42,313.84	374,718.74	42.78
RAID 6	158,269.44	17,630,614.31	981,495.80	9,306,055.06	1,062.34
RAID-Z3	150,732.80	510,027,205.98	37,652,950,557.88	19,081,488,881.93	2,178,252.16

Achieving this level of protection requires many individual features working together to protect your data, all of which are present and standard in ArcticBlue.

Definition

Triple Parity: A system used which creates a mathematical checksum of every block of data that can be used to recover any lost data up to and including three full hard drive failures in a single band. A system would have to have three full hard drives fail at the same time before any data is lost.

In ArcticBlue in particular, the combination of a wide 20+3 (20 data and 3 parity) drives and automatic rebuild to global spares provides excellent protection as long as global spares are available and the system is properly maintained. Each increment of 192TB purchased includes the 20+3 band as well as 1 Global hot spare drives. See rebuilds in the Maintenance section below for detailed information.

Global Spares

ArcticBlue is constantly monitoring individual drives for failure indications. When the system detects that a drive is no longer reliable, it immediately begins a rebuild of that drive to an available Global Spare. Since ArcticBlue uses triple parity, up to three drives can fail at the same time in each 23 drive band with no loss of data. It is critical that enough Global spares are available so that rebuilding can always start instantly. Spectra recommends keeping four global spares per ArcticBlue expansion node. Rebuilds can take place to any available spare in the system. Once complete, the failed drive is marked as bad and in need of replacement.

Rebuild time

If a drive fails for any reason and a rebuild is required, ArcticBlue immediately begins rebuilding the failed drive to an available Global Spare. That process is math and processor intensive as it requires reading back block by block from each of other drives in that band, performing the math functions, and rewriting the result to the Global Spare. Only the blocks that have data are read back for recalculation, saving time. Current rebuild performance on a 23 wide band is approximately 30MB/s. This currently translates to roughly 70 hours for an 8TB drive. As soon as a drive begins rebuild, a service call can be initiated for basic levels allowing shipment of a replacement drive to take place and be on hand when the rebuild completes.

Value-Feature

Note that in this system, only actual data is rebuilt, unlike in traditional hardware RAID systems where an entire disk is rebuilt sector by sector. Assuming a drive is only half full, it only takes half the time to rebuild. These systems are designed to provide alerts when 80% of capacity is reached so even with a heavily loaded system, a 70 hour theoretical rebuild will typically only take 50 or so hours.

During rebuild, BlackPearl will power-up other bands when possible first for writes, otherwise any external data movement takes priority over the rebuild. The affected pool will still be fully available, and performance impact will be negligible since the pool primarily would only be used for GETs.

With on-site service level options a technician will not be dispatched until the last global spare has begun a rebuild and when they arrive, all suspect drives are replaced in all chassis.

Definition

Checksum

A cyclic redundancy check (CRC) is an error-detecting code commonly used in digital networks and storage devices to detect accidental changes to raw data. Blocks of data entering these systems get a short check value attached, based on the remainder of a polynomial division of their contents.

BlackPearl Checksums

A file level or part of file (chunk/blob) level checksum is provided by the Object Storage system, each file or chunk gets a checksum calculated and stored with the object. An End-to-End checksum can be requested by the host or set at the bucket level, and the host is required to provide a checksum for the file or blob, and BlackPearl verifies this and maintains it providing high Data Integrity through the objects life. Many checksums are supported in the system: MD5, CRC32, SHA256, and SHA512.

ZFS Continuous Checksums

Once the file is sent to the ArcticBlue storage pools ZFS calculates and stores independently a checksum for each ZFS record of data in the system. This checksum is verified on every read access to the media guaranteeing that data corruption is detected before erroneous data is presented to the user.

How Data Can Become Corrupted:

- Data is transmitted through a chain of components each imposing a certain probability that it may corrupt a given bit of the data. This includes but is not limited to components such as software, CPU, RAM, NIC, I/O bus, HBA, cables, and the disks themselves.
- Studies have shown that the error rate in a large data warehousing company can be high as every 15 minutes[1]. This only serves to illustrate the need for data integrity protection when selling systems designed to store hundreds or thousands of terabytes.
- To address the problem for the entire I/O stack, ZFS performs checksums on every block, and critical blocks are replicated.
- ZFS also has the ability to correct errors before they're no longer correctable, on both a passive and active basis.

The combination of ZFS checksums and parity allows ZFS to self-heal in situations that other RAID solutions cannot. For example, hardware RAID filers can be configured to use RAID parity to detect silently corrupted data in much the same way that ZFS uses checksums. However, there is nothing these systems can do to safely reconstruct the data or to inform the user which drives are causing the problem. In this situation, a ZFS solution will try all of the possible combinations of parity reconstruction until the record checksum is correct, detecting which member or members of the RAID have bad data and recovering the original information.

Multi-Level Error Correction Codes (ECC): ArcticBlue performs checksums on every 256K block of data. The block's checksum is stored in a pointer to the data block rather than with the data block, adding a layer of protection between the corrective mechanism and the data itself. If the checksums don't match, ArcticBlue identifies an accurate copy of that data, or rebuilds another copy through smart RAID (see below).

Memory with ECC and Interleaving: An often-overlooked cause of data corruption is the memory component of the storage system. The ArcticBlue system addresses this by incorporating memory that incorporates integrated ECC and interleaving. ECC checksums ensure that individual errors in memory are corrected, and interleaving permits recovery from an error that affects more than a single bit within the memory of the system.

Using these mechanisms, even though raw disk systems are prone to periodic bit errors "bit-flip", "bit-rot", ArcticBlue has the ability to detect those bit anomalies and then use the RAID Z3 system to correct those errors using the triple parity redundancy. With the specific coding scheme used in ArcticBlue, errors are rarely missed. The undetected bit-error rate internally is on the order of 10^{67}. To illustrate just how rare this is, that is roughly the number of atoms in our solar system. So if we were to store the value for every atom in the solar system, one would possibly slip by us wrong and undetected.

BlackPearl Hybrid Storage Architecture Ecosystem

ArcticBlue is a key part of the BlackPearl ecosystem, right beside a deep storage tape library; creating a Hybrid Storage Architecture. The simple S3 interface allows ArcticBlue and BlackPearl to integrate with any S3 "out" application with little to no modification. Within a single system, and by a single vendor, we have both a nearline disk storage and a deep storage library. Actually BlackPearl also supports a third type of storage in its ecosystem, Enterprise SAS HDDs as a storage pool as well; the same drives used in Spectra's NAS disk product – Verde.

BlackPearl has data policies at the bucket level that enable redundant copies of data, in that bucket, on multiple media targets including disk, tape and even an automatically ejected copy for offsite storage.

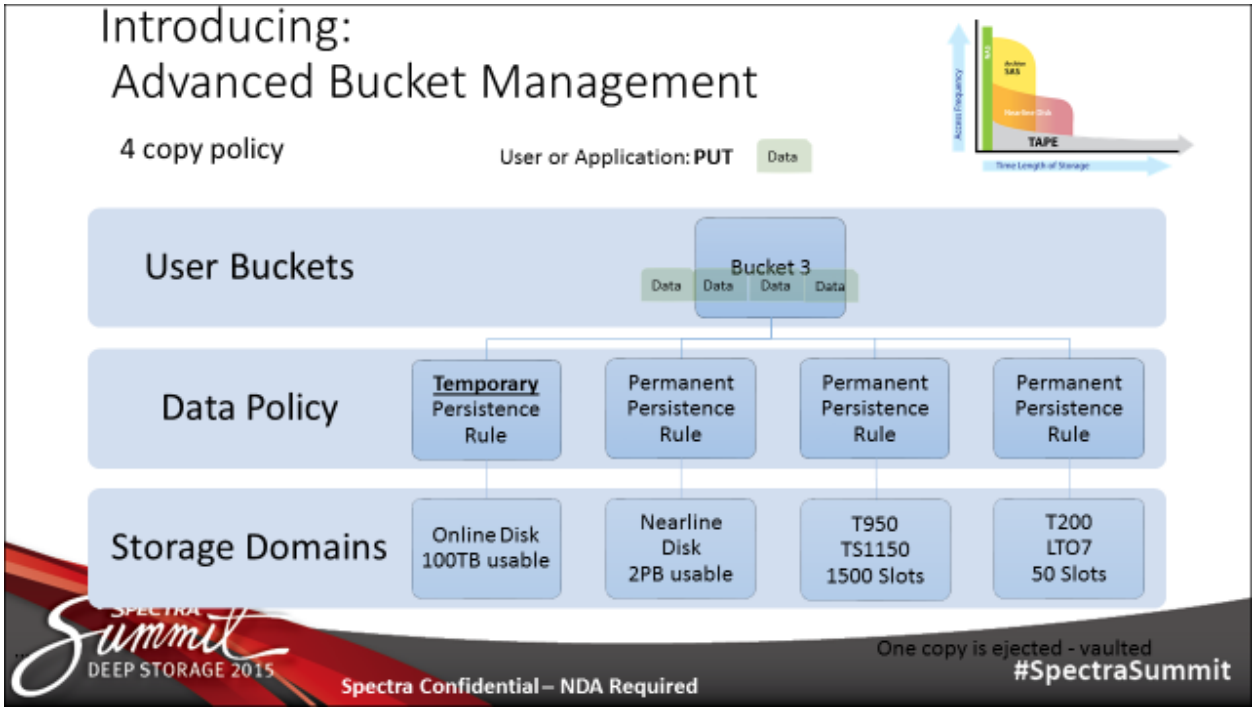
Data Redundancy

Beyond the RAID Z3 triple parity system, the easiest way to add redundancy to a storage system is multiple copies. Making a seamless second copy on a different media target, preferably on one that gets physically separated from the first, prevents data loss due to natural and man-made disasters.

Requirements

In the administration GUI data policies are set within the BlackPearl ecosystem.

- The ArcticBlue storage pools are put into a Disk Partition(s)
- The Disk Partition is put into a Storage Domain
- A data policy is created: Ea. Storage Domain gets a persistence rule (temporary or permanent), Default job priorities are set, default (file level) checksum, and if End-to-End CRC is required
    - Each Storage Domain (different of the same one) added to a Data Policy creates another redundant copy of any data PUT into that policy
- Optionally for true data redundancy with Genetic Diversity a Tape Partition is attached to BlackPearl, its put into a Storage Domain and added to any Data Policy

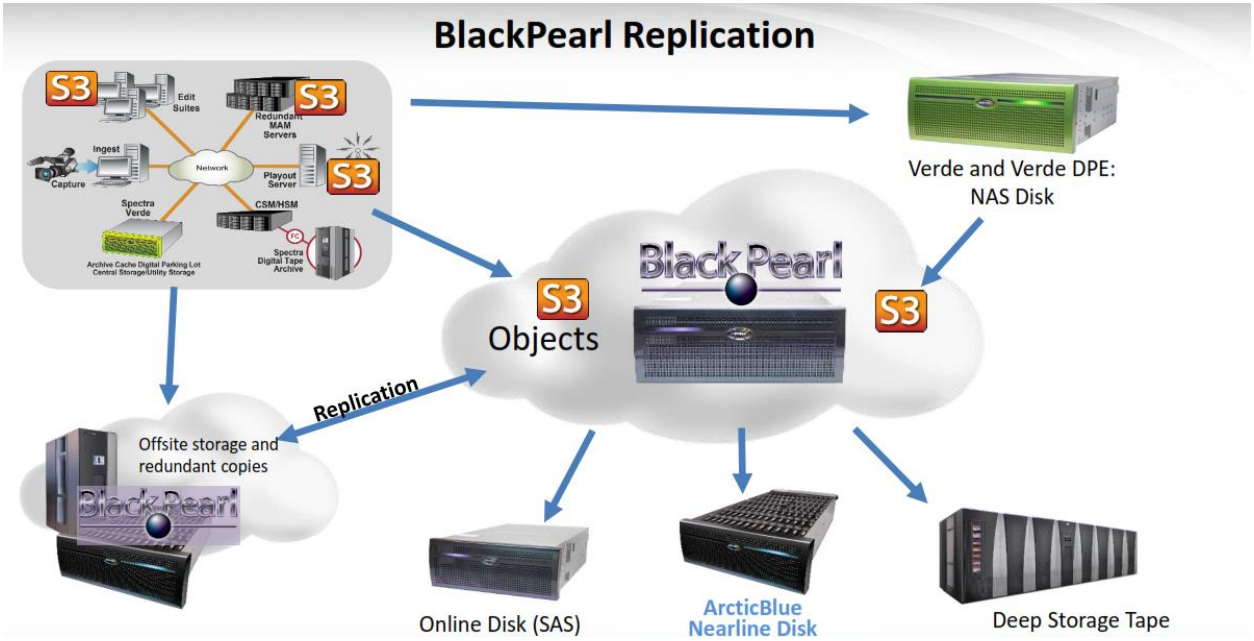


Roadmap Item

BlackPearl Replication (mid 2016 feature enhancement)

BlackPearl affords a degree of site replication that was previously only available to disk and tape users deploying expensive software applications or expensive primary storage that harbored the intelligence to replicate data to multiple sites. Not only was this middle-ware expensive, but it introduced a layer of complexity into the storage environment that was not necessary.

The IP-based, object storage front-end of BlackPearl also delivers data replication capabilities. As the gatekeeper to enormous amounts of storage behind it, BlackPearl can make multiple copies of your data and dispense those copies across different storage domains, and devices behind it. This functionality includes replicating objects from one BlackPearl to another BlackPearl unit or units at alternate locations (sites) in an Active-Active configuration. Should an object not be available from one BlackPearl because it was off-line, the application can request that object from another BlackPearl in the configuration, even at another location. The other BlackPearl receiving the request will go ahead and provide that object back to the application in a seamless manner with no service interruption and very little latency.



A unique advantage of the Active-Active replication architecture, it basically allows a Bucket to Span across multiple BlackPearls, where each BlackPearl uses the respective storage for that location, be-it nearline disk or deep storage tape. In essence allowing data to be replicated while simultaneously sharing access to that data, and providing many endpoints to upload (or PUT) data to the BlackPearl Private Cloud. There is a delay, depending on bandwidth and job priorities before data put into one BlackPearl endpoint can be replicated to the other BlackPearl systems. It is critical that clients always wait until job complete is received from all storage domains, local or remote, before deleting the original data to ensure preservation of that data. [1][2]

[2] Roadmap item CY2016.

The Spectra Ecosystem

In the larger data center ecosystem the ArcticBlue is just one part of a larger storage architecture.

The Verde family provides a file based storage system accessible either through CIFS (Windows primarily) or NFS (a file interface used by Linux and other similar systems) that can be mounted much like any normal network drive. Verde SAS provides fast access to files that can be modified and deleted at will with little effect since the system is based on enterprise grade fast SAS hard drives. Verde DPE typically is used for bulk storage of large data sets that tend to remain on the same volume for long periods of time but can be accessed instantly.

Spectra has been in the business of building digital tape libraries for over 35 years and has a large selection of tape options from a 50 tape slot capacity robotic system to one that holds tens of thousands of tapes and can store nearly an Exabyte of compressed data. Typically, however, complicated and expensive software is required to interface between the “file system” world and storing data on tape libraries. It takes a lot of code to cache data, stream it to tape drives, control the robots that move tapes to and from those tape drives, and catalog the entire process. In an effort to streamline this process and provide an alternative to middleware-software, Spectra launched the BlackPearl Deep Storage gateway. BlackPearl is an appliance that completely automates and controls the process of long term data storage.

BlackPearl has a frontend that accepts data via an industry standard S3 http RESTful interface. This interface was originally pioneered by cloud storage providers as a means to move large amounts of data over the internet, which has uncertain delays, where bandwidth cannot be guaranteed at any given time, but is simple to integrate into software so a direct connection can be made from those applications to various types of storage. BlackPearl essentially provides a large Deep Storage pool on-site using that same interface saving the customer money and providing positive control of data.

Since a normal file system is not directly capable of outputting data using S3, a “client” must be created so an application can “talk S3”. Spectra has developed an extensive Software Development Kit [SDK] to make creating clients for BlackPearl simple and is directly involved in creation of those clients for customers who need this capability to enable a BlackPearl and therefore a tape library.

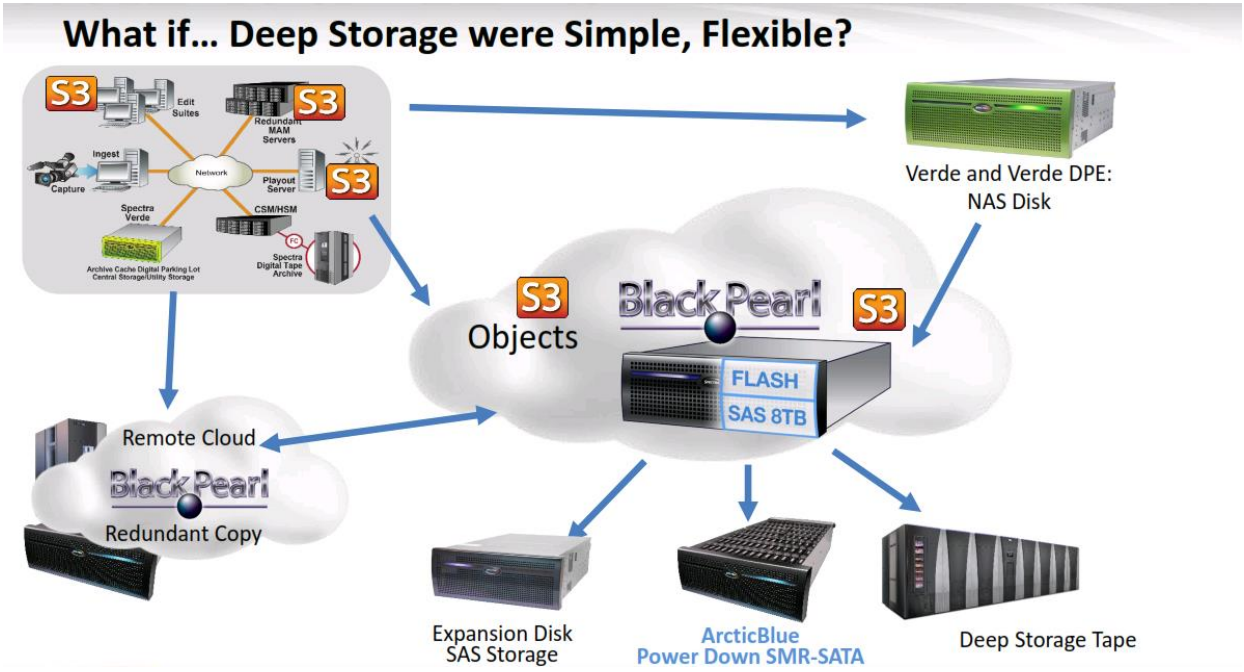
BlackPearl has a very large fast access cache to accept incoming data from the S3 interface, catalogs all the metadata from those objects for later retrieval in internal SSD memory, and directly controls both the robots as well as the tape drives via a fibre channel interface. So issuing a single “put” command via S3 is all that has to be done to send that data off to Deep Storage; BlackPearl handles the rest.

Inside BlackPearl is a technology called Advanced Bucket Management which includes the ability to set data policies for buckets. Buckets are logical containers in cloud devices and object storage systems to

store data in a flat architecture, can't nest a bucket in another bucket. Data policies are defined as groups of both storage media (disk and/or tape) on storage domains where partitions are assigned, as well as persistence rules that govern how long the data is stored on storage domains. As an example, a bucket could be created that spans both an LTO7 partition in a library as well as an expansion SAS disk pool under BlackPearl. To do so, a data policy can then be created such that any data written (PUT) to that bucket automatically creates a copy on disk, a copy on LTO7 tape, and a second copy on LTO7 tape that is automatically ejected from the library for sending to an off-site location for safety.

The question, however, is what is available if a customer has either a manual process for archiving data or does not have an S3 client built into their application that can send data directly to BlackPearl. Verde family products includes their own S3 client that is called NFI – Network File Interface purpose built for BlackPearl and all types of storage behind it.

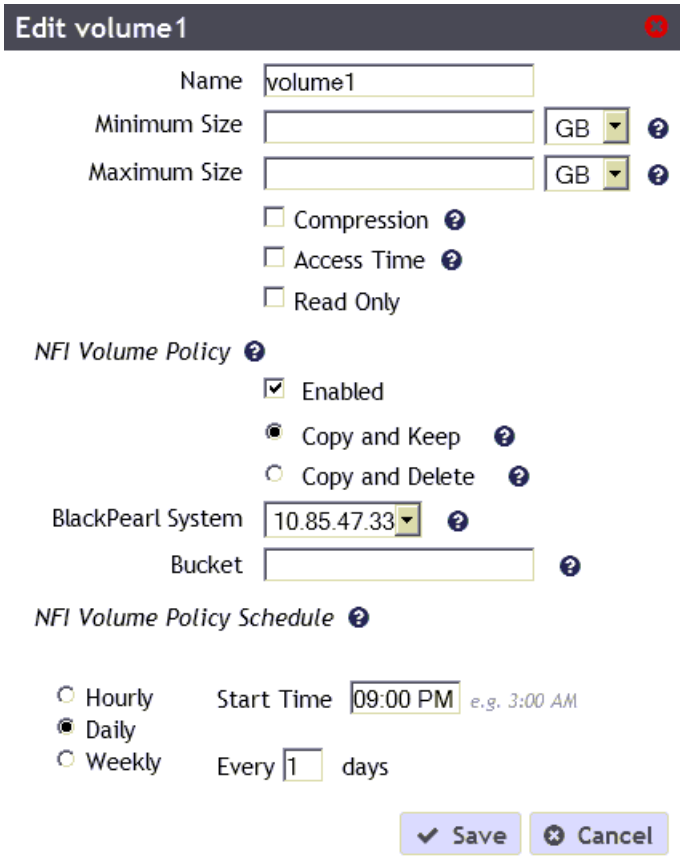
Obviously the goal of NFI is to provide a method to directly archive the contents of the file based bulk-storage system directly to tape through BlackPearl. If that copy is intended to be an eject copy, it can normally be accomplished for around 2.4¢/GB. A system with versioning that can provide basic self-archive capability is under development but for the initial release, a straightforward NFI portal has been provided.



The NFI policy on Verde DPE is a bit different than Verde SAS, as an example we'll show how configuring the NFI portal is a simple process through the Verde DPE GUI at the volume level. The setup steps include:

- Setup the link to an existing BlackPearl System using Configuration>Services>NFI>Action>Configure new BlackPearl system
    - On this screen the BlackPearl IP address and credentials are entered.

- On the NAS>Volume screen, either create a new volume or edit an existing one. Note that the NFI function is restricted to only be available on volumes of 4TB or less. This prevents overlapping NFI sessions as the minimum schedule is once/day.
- Choose the BlackPearl previously set up, a schedule for the NFI job start, and a bucket name
    - Note that a generic name can be chosen in which case a new bucket will be created on BlackPearl if that bucket does not yet exist.
    - The date and time will be appended to the bucket name for each job. This prevents duplicate files from causing an error in BlackPearl since a new job will be in a new bucket. If any duplicates were to exist, BlackPearl would report and error and reject the entire job, this method prevents this.
    - Each job, since it is a new bucket name, will take at minimum a single tape since each bucket is put on a unique tape or bank of tapes. Care should be taken to schedule such that tapes are not wasted.

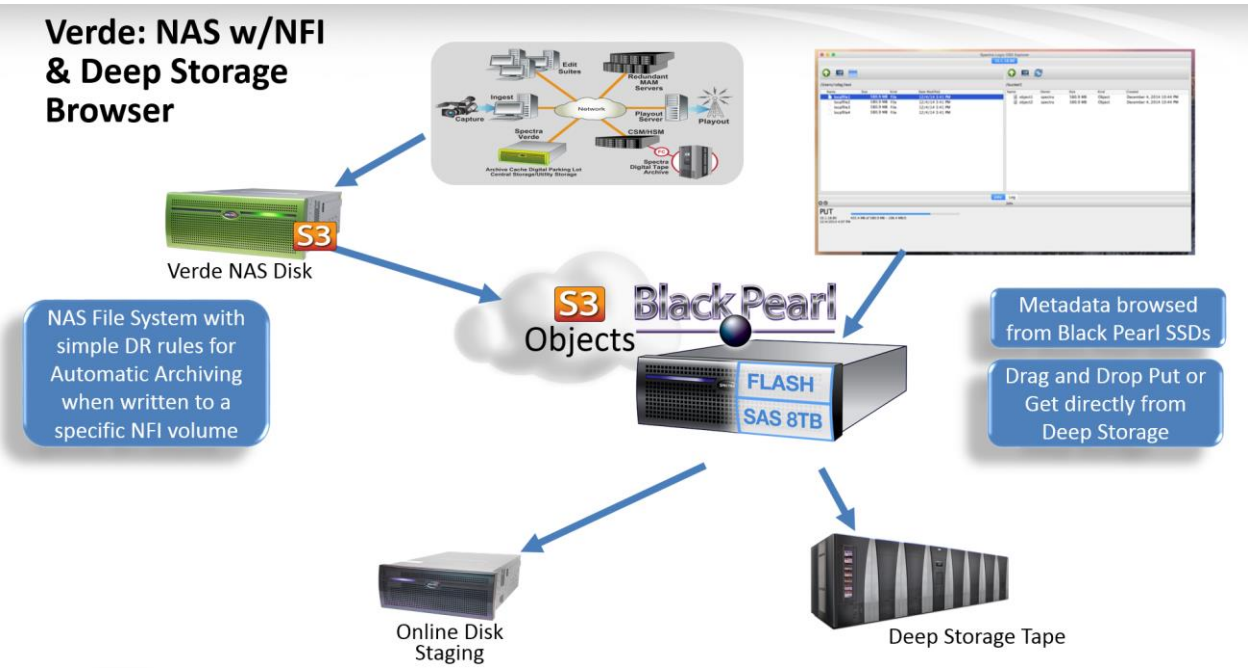


Once the volume is shared to either NFS or CIFS, then any data written to that share/volume will be output to BlackPearl on the set schedule.

With a tape backend, the tape cartridges are allocated as more data is added to a bucket. This allows for the BlackPearl to share empty cartridges across many buckets. When disk is on the backend, the buckets are created in a ZFS disk pool and data is stored there. There is also the ability to combine both disk and tape behind BlackPearl, for more on that see the BlackPearl web page. The tape system behind BlackPearl can be any Spectra Logic tape library, ranging from as small as the T50e up to and including the TFinity. The library does not need to be dedicated to BlackPearl, BlackPearl works with a partition in

a library. The minimum partition in a library is 1 drive and 10 data slots, the maximum partition size is the number of slots you have licensed in the library.

Once any data is stored on tape or disk through BlackPearl, in order to view or retrieve it the Deep Storage Browser [DSB] is used.



This is a simple multi view (panes) system that allows search as well as drag and drop between file systems and deep storage object based systems. DSB is loaded onto any client computer with network access to BlackPearl and the same credentials are used as in the initial setup of NFI. The interface provides the local file system on one side and a representation of the object side on the other. The object database can quickly be traversed or searched as it resides in SSD on BlackPearl but once an object or objects are chosen for retrieval, if they exist only on tape (they have expired from online disk and cache) it takes some time to retrieve as physical tapes have to be loaded by the robot for retrieval.

File System or Object Storage

File Systems are different than Object Storage. Main difference is a file system protocol allows for files to be edited with open, modify, and close commands; among many other commands. And Object Storage is a much easier protocol, with PUT, GET, Delete, List, and Copy. Object storage systems typically treat any new file/object as an entirely new thing and if versioning is enabled keeps the previous version. And if versioning isn't enabled would delete the old object before "Putting" the new one. This type of protocol is very advantageous to sequential media, like tape and SMR drives. Additionally, the simplified interface is more efficient, meaning easier for a single client to saturate a network connection to an object storage system than a CIFs or NFS file system where many clients are typically required.

Traditional File Based SMR workflows (where to use it, and where not to…)

Given what we discussed above on how SMR drives work, it is important to lay out exactly how a file system (Verde DPE) should be used. It might actually be easier to start with how not to use an SMR product with file system access. Considering that SMR drives take more effort to fill empty gaps, workflows that include a very high rate of data modification and/or a lot of frequent deletes that leaves random holes are non-optimal. A good example would be running a large airline reservations database on this or attempting to do real time video editing. Just say no.

Using the same logic to consider ArcticBlue use cases, a significant advantage is added because of the RESTful S3 interface. Sitting behind this layer, BlackPearl can intelligently cache and sequence data with the large disk cache, prevent random data modifications, intelligently control power down and still not have the higher level system time out, and apply system policies to data movement all out of the visibility of the end user.

While writes are carefully controlled, reads are completely unrestricted with full random access reads available across any and all bands at the same time once they are powered up.

Key Information

So, the ideal data set would be information that is stored in large chunks, can be written, or cached to be written, sequentially, rarely deleted or changed but can be read in any order, at any time, randomly, and as often as needed. Aside from the corner cases mentioned above that encompasses almost every other kind of data: Backups (with deduplication), archive, scientific data, general IT (shared drives etc.), video surveillance, video and media parking during editing, and many other types of data that include unstructured data.

Performance – what to expect

Very Wide Bands for performance and capacity

A unique feature of ArcticBlue systems is the use of Very Wide Bands [VWB]. Most RAID systems use smaller stripes of 5-10 disks to balance system performance and RAID 5 or 6 single or double parity. While some other object storage systems use a 20+6 erasure coding with higher overhead. Spectra's implementation of ZFS, as previously described, uses a triple parity Local Erasure Coding system to maximize data integrity. While this provides excellent data protection, it also requires extra hard drives. To counter this, a 20+3 band size is used, giving up to 83% usage vs raw capacity.

Compared to other object storage systems: a common 20+6 erasure coding has 77% usage, or a 3 copy redundancy is only 33% usage. The BlackPearl ecosystem provides the flexibility to provide the best of both worlds, a zero overhead copy on tape and redundant ArcticBlue copy with minimal overhead.

As another benefit, a VWB also helps throughput. As previously mentioned, SMR drives can achieve up to 100MB/s, so a 20 wide usable band provides up to 2GB/s of raw throughput. Of course with system overhead, only half of that around 1GB/s is to be expected.

Typical performance and how to measure

A user setting up a VWB pool of 20+3 drives with an expectation of 1GB/s typical performance. To compare lets first look at file system. A windows/CIFS mount to do a drag and drop of large 100MB files can be monitored on the performance screen on the Verde GUI; it is likely with this setup a user would only see only 100MB/s. There are two unique advantages of REST and S3 protocol to maximize performance and the question is what effect different types of data has on that performance.

Key Information

File size implications

Object Storage has an overhead per file, creating a UUID and DB record, uploading any metadata into the DB, then actually caching the data, calculating a checksum, and persisting it into storage. This overhead starts to show with smaller files, in the Megabyte or less range. As each object storage system, like BlackPearl and ArcticBlue improve, the file size where throughput degrades keeps decreasing.

Any files larger than tens of megabytes receive the maximum throughput to the system.

BlackPearl cache

All PUTs into ArcticBlue storage go through the cache first. The cache aggregates single file PUTs or accepts Spectra S3 SDK/API bulk PUTs where the data is already batched together. From the cache it's then written out to storage.

The cache needs to be large and fast, and affordable. Flash doesn't strike a good balance of all three of these, and they wear out quicker. The cache in BlackPearl is made up of 10 or 20 HDDs, with 4TB capacity drives providing a good cache balance while staying affordable.

The cache can sustain over 1GBps in the S3 interface, while simultaneously writing out around 800MBps to the backend (ArcticBlue or Tape).

S3 Put Single File

ArcticBlue will work with start/native S3 single file PUTs following the AWS spec. To increase performance to sequential media the BlackPearl cache will aggregate these files together then write them out to backend storage in larger 'jobs'. The sequential nature of SMR drives will enjoy this additional batching process, and its key for writing data to LTFS tape. If the BlackPearl ecosystem includes tape, the data gets persisted from the cache to LTFS tape. The way data is persisted to a self-describing LTFS tape is after each job the index partition has to be updated which equates to a tape stop/restart. The larger the batch, the more efficient the tape throughput will be and therefore the overall BlackPearl ecosystem throughput.

The throughput to ArcticBlue will be sustained in the 1GBps range.

Spectra S3 Bulk PUT

The Spectra API was designed for sequential media, specifically tape. It added additional BULK commands that are like a priming and batch command combined. If the application workflow is project based, or capable of staging many files in front of BlackPearl the Bulk PUT with an SDK instantly saturates the BlackPearl cache performance with a single job (given the file size requirement above), while sustaining 800MBps to six tapes drives on the backend, or a 1GBps if the backend is only writing to ArcticBlue. And if the BlackPearl backend is writing to both tape and disk storage, the overall backend throughput for two copies should be in that same range (or a bit more than half their respective throughput – i.e. 400+ MBps to tape and 500+ MBps to ArcticBlue).

System CPU and DRAM

Unlike traditional hardware based RAID systems in the past, the only way to achieve reasonable performance, particularly for rebuilds, was to use custom silicon with dedicated math engines to perform the XOR and multiply functions quickly. By tying system performance to the processor speed improvement curve, however, software RAID systems and local erasure coding have been able to provide exceptional performance gains. In BlackPearl in particular, a dual CPU (Xeon 6 core) plus 64 or 128GB of DRAM is used to support the object storage and ZFS file system for ArcticBlue. The system memory acts as a virtual cache and also supports fast rebuilds of parity blocks as necessary.

Network setup considerations

Most throughput and connectivity issues, and hence the vast majority of support calls, are a direct result of network setup inconsistencies.

The management port is separated from the data ports. The management port and data ports have their own default routes. This does not prevent a user from having management and data utilizing the same network if desired.

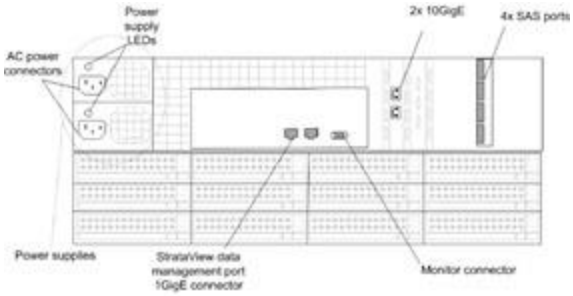
The basic steps on configuring the management and data ports for access to the network are simple and straight-forward. However, each customer network environment is unique and may require some additional troubleshooting in order to properly connect to Verde and utilize the 10 Gb interfaces properly.

Configuration

The first step is to configure the management and data ports accordingly using the user interface, i.e. GUI. Do not attempt to access the system directly via the root console and modify interfaces directly. The management code is tightly integrated with the base operating system and other actions occur based on network changes.

Connectivity to the Network

The data path is supported in the following manner:



- Single 10 Gb logical connection utilizing the onboard 10 Gb (10Base-t copper) physical port

  OR

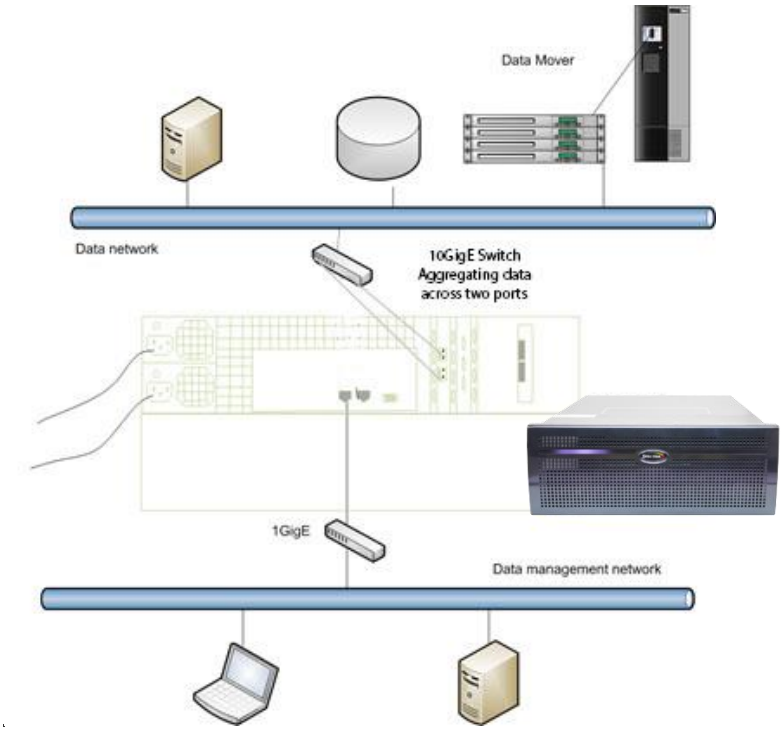
- Single 10 Gb logical connection utilizing one of the 10 GbE optical physical ports with SFP in the PCI expansion card

  OR

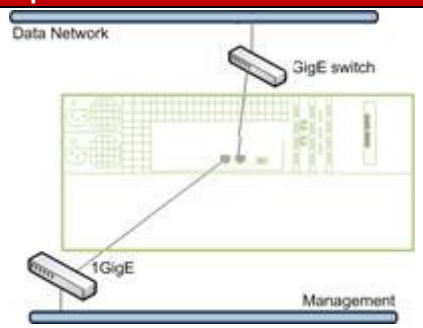
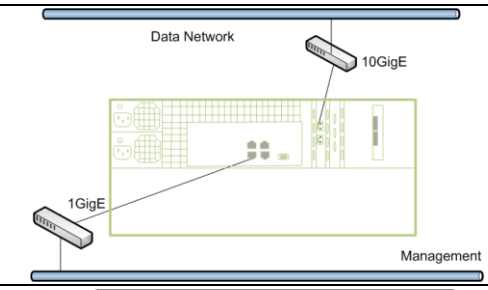
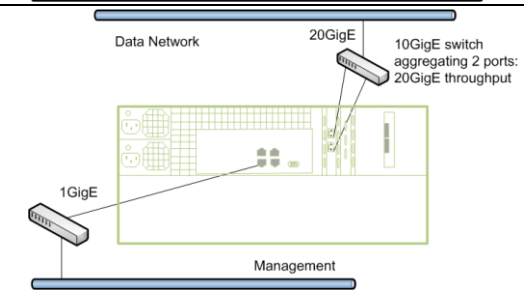
- Single 20 Gb logical connection utilizing two 10 GbE optical physical ports with SFP in the PCI expansion card (Link Aggregation)

4U BlackPearl: Example Network Diagram

The following diagram shows an example environment with the Verde system used in an architecture supporting data transfer and data management networks.



Additional network configurations are possible; a few of these are illustrated in the following table.

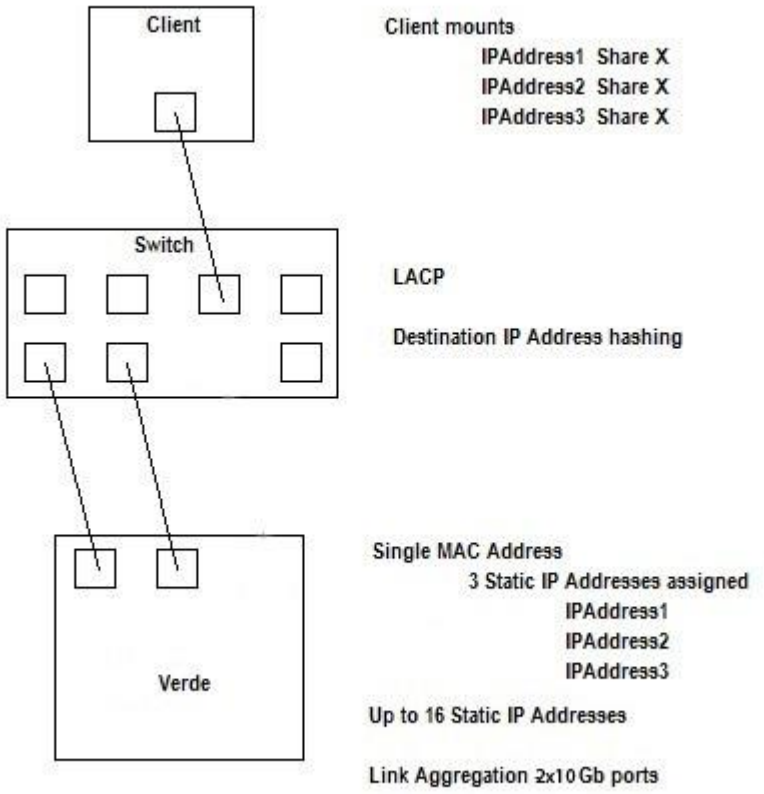
Connectivity	Options
10 Gb data transfer port 1/10 GigE management port	
10/40 GigE data transfer port 1/10 GigE management port	
Two 10/40 GigE data transfer ports aggregated to one port with 20/80 GigE throughput 1/10 GigE management port	

The user should assign the appropriate IP address either statically or via DHCP to the management and data ports. If setting the MTU to something other than 1500, the user should ensure that their switch configuration supports larger MTU settings as well as all hosts on the network. The Verde can support Jumbo frames (MTU=9000), but all switches and hosts on the same network must be configured to support Jumbo frames if this is chosen or performance may be degraded. Additionally, switches must also be able to support Link Aggregation if the user specifies in the Verde network configuration to aggregate or "trunk" the data ports together to provide higher bandwidth into the Verde storage unit. Switches must support LACP and hash the destination IP addresses as there are multiple methods for link aggregation. The user must manually configure LACP on those switch ports. LACP does not get enabled automatically.

Link Aggregation Notes

Various switches use different methods of routing traffic from clients to NAS servers. There are also many different network configurations to move data from clients to NAS servers. For example, some Cisco switches route traffic based on the MAC address and the IP address. The unit presents only one

MAC and IP address when the data ports are aggregated via DHCP. If static link aggregation is chosen, the unit presents only one MAC address, but can have up to 16 IP addresses aliased to the MAC address. It is up to the switch to rotate data transfers amongst the ports being used to physically connect to the Verde DPE unit in order to achieve the highest throughput possible. This is an issue when a customer has only a single client connected to Verde and is measuring performance. A customer may only see 100 MB/s performance over two aggregated data ports since the other ports are not being utilized by the switch. If one were to connect multiple clients to the Verde unit or mount a share multiple times using different IP addresses and start transfers from all three clients, one would see up to 1GB/s typical. A user may have to configure more than three IP addresses on the Verde to get the switch hashing algorithm to utilize all physical ports and maximize performance.



Once configured properly and attached to the network, the status in the UI should indicate the speed of the connection and whether the port is active. The link lights on the network ports should be on and active at both the BlackPearl and network switch.

The user should be able to “ping” the assigned IP for the given management or data port that was set during the configuration of the ports from a client external to Verde on the customer network. If not, please follow the troubleshooting tips below to ascertain if the problem is a network setup issue.

Troubleshooting Network issues
Port link LED does not light

Solution 1: Check the port configuration on the network switch. The Verde DPE system supports only auto-negotiation. Make sure the switch is configured to match speeds on both ends of the connection. In the past many environments did not properly handle auto-negotiation.

Solution 2: Check the cables that connect the port to the other device. Make sure that they are connected. Verify that you are using the correct cable type and connectors. This is especially true for 10 Gb connections utilizing SFPs.

Solution 3: Verify that the switch ports are not administratively disabled.

Port link LED is lit, but I cannot ping the Verde

Solution 1: Check the LACP settings on the switch. If you are using link aggregation on the Verde, the switch MUST be configured to use LACP on those ports. If you are not using link aggregation, the switch must NOT be configured to use LACP on those ports.

Solution 2: Check the VLAN settings on the switch. Ensure that those ports are assigned to the correct VLAN.

Ping

The **ping** command is a simple tool, based on a request-response mechanism, to verify connectivity to a remote network node. The **ping** command is based on ICMP. The request is an ICMP Echo request packet and the reply is an ICMP Echo Reply. Like a regular IP packet, an ICMP packet is forwarded based on the intermediate routers’ routing table until it reaches the destination. After it reaches the destination, the ICMP Echo Reply packet is generated and forwarded back to the originating node.

For example, to verify the connectivity from the switch to the IP address, run the command shown below from the switch command line or client:

Example:

ping 192.168.2.10

All ICMP Echo requests should receive replies. There is also additional information about the round trip time it took to receive the response. 0 msec means that the time was less than 1 ms. If the request times out then check the settings on the switch to which the Verde is connected.

Traceroute

You can use the **traceroute** command or something similar if it is available to not only verify connectivity to a remote network node, but to track the responses from intermediate nodes as well. The **traceroute** command sends a UDP packet to a port that is likely to not be used on a remote node with a TTL of 1. After the packet reaches the intermediate router, the TTL is decremented, and the ICMP time-exceeded message is sent back to the originating node, which increments the TTL to 2, and the process repeats. After the UDP packet reaches a destination host, an ICMP port-unreachable message is sent back to the sender. This action provides the sender with information about all intermediate routers on the way to the destination. The command shown below

Example:

traceroute 192.168.2.10

In the output, you will see an enumerated numbered list indicating how many hops on the way from the switch to the Verde are encountered when tracing the packet.

ArcticBlue implementation

Theory of Operation

Software Components

Two mirrored boot disks in the BlackPearl system are dedicated to the software necessary to run the product including the operating system, and ZFS.

The software has a 'Front End' and a 'Back End' to the product.

The front end is the main software of BlackPearl where the Data Planner controls most aspects of the system. The front end also has the HTTP server (tomcat) and object storage system. The back end has the storage domain manager and the LTFS tape manager.

Master node operating and file system: The BlackPearl master node has internal specialized dual boot disks (separate from the data disks) that support the unit's operating system, integrating a logical volume manager and file system, used for data stored on ArcticBlue storage. The boot disks are mirrored, and control the structure and management of data storage. The BlackPearl operating system provides data verification to protect against corruption.

SNMP Server: The system accepts SNMP queries used by some network management applications, making it easy to track Verde DPE status in the context of the entire network.

HTTP Server: The system is running a tomcat HTTP server to host the Spectra S3 API, provide version 2 authentication, and IP network system access.

Web Interface: The web interface provides browser-based configuration, management, and monitoring of the ArcticBlue node. See the *System Ease of Use* section for more information.

Storage Domain: The system uses Storage Domain as a logical location for a physical partition (disk or tape) to place a copy of data. E.g. a Storage Domain is created for a ArcticBlue pool disk partition.

Data Policy: Data placement rules are created, whereby each storage domain assigned gets a persistence rule (temporary or permanent) for each copy needed. The same storage domain can be used more than once to in the same data policy to create multiple copies on the partition in that storage domain.

Example Case of Pools, Protection, and Arrays

The following figure illustrates the physical disks used in this example.

Note: *Throughout this example, physical hard drives are shown for illustrative purposes only.*

Virtualizing Nearline Disk Pools

A storage pool is a virtualization of multiple disks; in other words, it's a logical grouping of a set of physical drives. The pool is the location where disk storage partitions reside.

ArcticBlue always uses 23 disks per pool with the others available for use as global spares. If, at a later time, more capacity is added by adding at least 23 new drives to create more pools. When more storage for bucket(s) is needed the BlackPearl data planner will use the additional pools.

ArcticBlue is only sold in 192TB increments

The reasoning behind ArcticBlue only being available in 192TB increments (bands) is a result of very careful analysis:

- Performance – Very Wide Bands [VWB] – specifically 20+3 provides optimal throughput performance for SMR drives with balanced (low) overhead
- Included Global Spare Drives, with space for them in the chassis, is critical for system reliability and Digital Preservation (chassis has 96 slots: 4 x 23 SMR band plus four global spares)
- That specific number (20+3) statistically provides the best storage utilization that maximizes usable space while still providing high data protection and performance
- Sales model is simplified and allows for consistent pricing as well as reaching the 10¢/GB sooner

The minimum configuration is two bands

In order to provide the Drive Lifecycle Management – powering down bands when idle, two bands are required. If there was only one band the system can never guarantee to put the band into a Drive Lifecycle Management state and power it down to extend the drives life.

Raw vs Usable space

For each 192TB band, which is its' own pool, 24 drives (8TB*24=192TB) are included. Of those 24 drives, 23 are in the triple parity local erasure code band and one is a global spare. While parity is actually written across the band, triple parity takes up the equivalent space of three drives in the band, so 20 are usable. This is 20*8T=160TB or 83% usable.

As is common knowledge, there is also a difference between GB and GiB. The Seagate datasheet for this specific drive states that it is an 8TB or 1 trillion bytes of storage which is actually 7450.6 GiB (2^{30}). This corresponds well to windows which reports one of these drives as 7452.04GB (you always get just a little more than the label). This has implications, however, as the customer needs to be very specific as to which "scale" they are referring when asking for a particular capacity.

10¢/GB – how much capacity is required to get there

The BlackPearl controller only has cache drives and database flash drives so the cost of ArcticBlue storage (per GB) has to be amortized over additional cost in a total price.

As soon as the order includes a full ArcticBlue node, a full node holds four bands: 768TB, that ArcticBlue node's cost with 768TB is priced at 10¢/GB. The same is true for each/any full node ordered whether only one is purchased or a rack full of ArcticBlue nodes is purchased. That way a customer reaches the 10¢/GB target with only the first full ArcticBlue node.

Expansion of Capacity

Any ArcticBlue system can be expanded just past 6.1PB in a single rack with a single BlackPearl controller. Each capacity increment is individually priced, meaning a ½ full node has a higher price per GB than a ¾ full node; where the best $/GB achieved when buying full ArcticBlue node. If an ArcticBlue node is purchased only partly full, customer upgradeable stipes (192TB increments) can be purchased

separately and self-installed. BlackPearl system is keyed for a specific number of slots (hard drives) so upon expansion, with the drive pack will come a new key. Each key will be an add-on key to provide an easy order experience and installation process. Note that slot keys come only with initial purchase or disk expansion packs but are not time sensitive, they never expire.

Since ArcticBlue nodes are large and require the addition of a PCI-SAS card in the head controller in order to add more SAS cabling from the controller to the new ArcticBlue expansion node, all ArcticBlue expansion nodes require Spectra on-site installation support.

Since all increments are 192TB, total system capacity can easily be expanded dynamically by the system up to 32 total 192TB bands or just over 6.1PB.

Physical Buildout considerations

Weight and rack considerations

Each full ArcticBlue node weighs 249lbs and comes with heavy duty slide-out rails as well as a cable management system. In order to account for safety concerns, Spectra recommends best practices that include:

- Build systems from the bottom up with heaviest components on the bottom of a rack
- Ensure any rack used meets weight and structural requirements
- Carefully evaluate any raised floor weight specifications to make sure that current as well as potential expansion capacities do not overload the floor
- Spectra highly recommends bolting the rack to the floor to prevent tipping
    - Note that Spectra racks for this system come with bolt down connectors that should be used only on reinforced floor systems. Simply bolting the rack to a raised unstable floor can create a significant safety hazard.
- Consider building large systems horizontally across multiple racks instead of vertically in a single rack
- See users guide and installation guide with best practices for a full list of requirements

Power Considerations

The power usage of an expansion frame depends on the number of drives powered on and active in the unit. The table below lists the power usage in WATTS for various configurations.

Number Of Drives Installed	On & Idle	On & Actively Writing	Idle & Powered Off
0	118	118	118
23	245	275	118
46	372	430	118
69	503	590	118
92	620	750	118
96	640	775[2]	140[3]

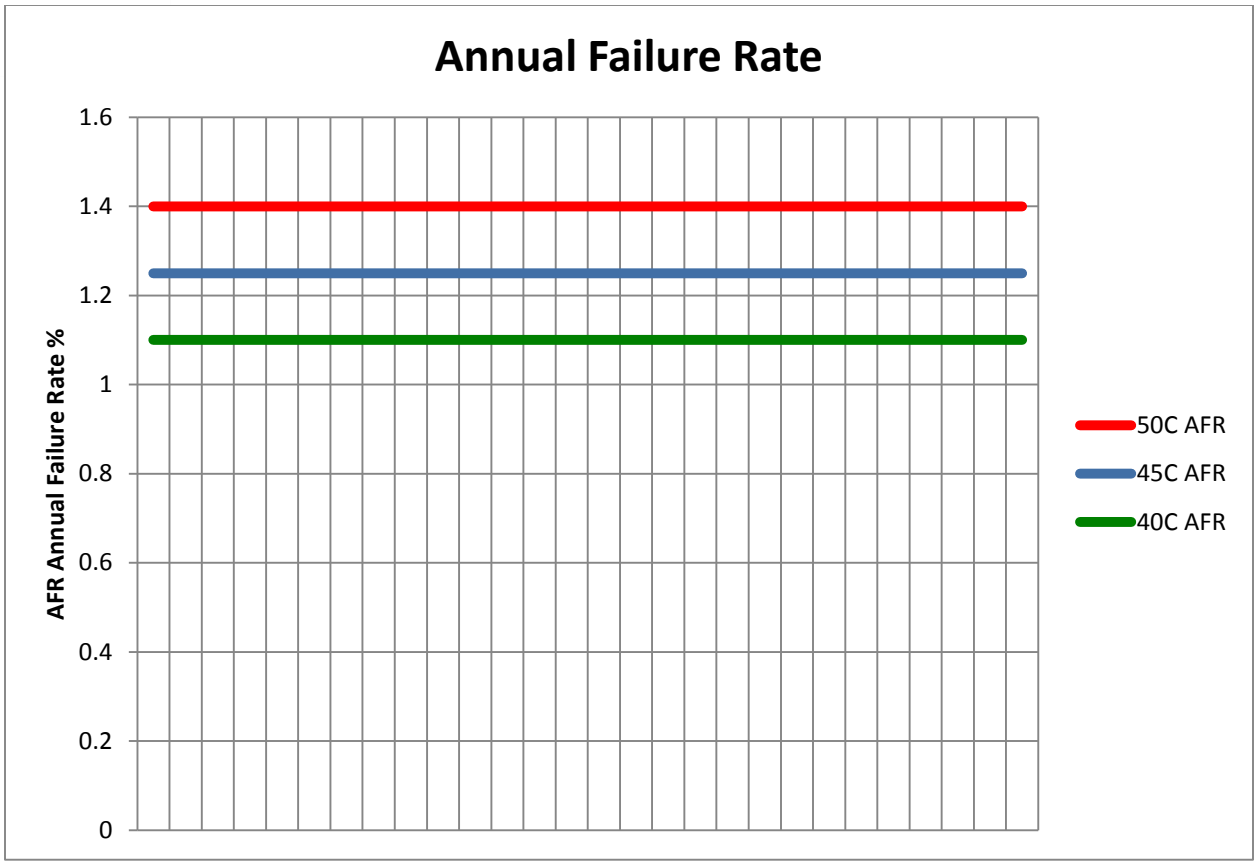
Cooling considerations

Proper temperature control is perhaps the single most important element to support a stable long term storage system. Drive maximum temperature is, in fact, the most critical aspect to monitor in your ArcticBlue system. If expressed as annual failure rate, the nominal operating temperature including all contributors including self and case heating, is 40 degrees C. Even a temperature increase of 10 degrees results in a major increase of the annual failure rate of drives as shown in the following graph.

In order to maintain a maximum 40C drive temperature, a data center ambient temperature of maximum 21°C is specified. This should be the inlet air temperature at the front of the ArcticBlue chassis. If temperature rises above this for a period of time, a warning is sent to the administrator via email. Significant violation of the 40C drive temperature will void the ArcticBlue warranty. The amount of drives powered on in ArcticBlue is highly dependent on system activity, a lot of restores across all data stored in the nearline disk system will cause more bands to stay powered on. On a very inactive or low restore system, more if not almost all of the bands will be powered off most of the time, lowering the risk of high temperature warnings.

[2] System with 96 drives powered on, and 92 of these actively writing

[3] System with 92 drives powered off, and 4 spares powered on. Note most configurations should have 4 spares on all the time.



For this reason it is critical that both ambient temperature be carefully controlled and adequate airflow be provided to cool ArcticBlue nodes. Cooling is from front to back and requires some special considerations. The HDDs are top loaded into the ArcticBlue nodes and orientated front to back as well.

- See users and installation manual for details on cooling recommendations as well as HDD placement and buildout
- In Expansion units:
    - Looking at the front of the unit the buildout is on a row basis (left to right) like stacking the HDDs on top of each other
    - Begin installing the first band(s) in the middle of the unit in the rows with the I/O module
    - Using “dummy HDD blanks” to fill any partially filled horizontal rows, the two rows with the SAS I/O module must be full
    - And each/any row with a HDD must be full (using blanks) to create an even air flow
- System will automatically monitor temperatures and alert the user via the front panel LED light bar, email alerts, and in extreme cases audible tones and even self-shut down
- BlackPearl powers down idle ArcticBlue bands, it would take a lot of GET request across the entire storage system (and the system must have a lot of data – more full) to keep all bands powered on 100% of the time

Note that the system records temperature over time and excessive violations of the recommendations in the user's manual will void warranty and support.

Add-in card options:
BlackPearl 4U controller comes with a dual 10GigE optical network card - standard.

Units can be upgraded with other network cards, a 10Base-T copper network card, or a 40GigE network card.

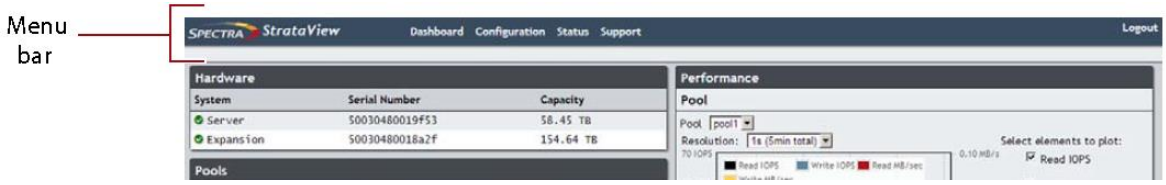
All network cards are dual port.

Each ArcticBlue node requires dual SAS cables from the controller unit and an expansion PCI-E SAS HBA card is added to the controller.

Note that initially, quad port 6Gb SAS HBA cards will be used providing support for two ArcticBlue expansion nodes per HBA card. Once a quad port 12Gb SAS HBA card is available from the same vendor, they will be qualified, resulting in higher throughput in some workflow cases.

Monitoring and Maintenance

The BlackPearl system has a simple interface that is extremely easy to use in configuring, managing, and monitoring system status. The system interface is password-protected, and lets you remotely monitor and manage the system. The initial screen, dashboard view, shows system and storage information. Navigate easily between this screen and the menu bar options that include Configuration, Status, and Support screens.



Status Bar

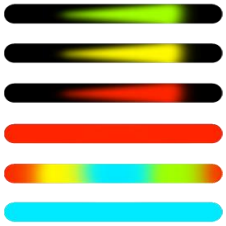
The web interface provides the status of the system at a glance, providing component status and information about any messages that require attention. The status bar, at the bottom of every screen, provides the following:

Icon	Meaning
	Component OK The component is functioning correctly.
	Information An informational message about a system component is available. Check messages to determine the component.
	Warning A system component requires attention. Check messages to determine the component.
	Error A system component has experienced an error condition. Check messages to determine the component and its error condition.
	Unknown The status of a system component cannot be determined. Check messages to determine the component.

- Icons that indicate hardware status at a glance
- Severity, date, and time of the most recent warning or error message
- Controls for rebooting and shutting down an array

Visual Status Beacon

The Visual Status Beacon light bar in the front bezel provides an at-a-glance status of the system. The light bar changes color to indicate the status. If a BlackPearl system requires attention, the beaconing blue bar helps to administrators identify the unit quickly.



Scalability: The Modular Design of Spectra Storage Systems

Spectra Logic's storage systems, including BlackPearl and ArcticBlue expansion nodes, are designed with modular media and components that let users add or swap drives and replace components stored on-site as needed, most with no downtime.

Modular Expansion: Scaling the Spectra System

To scale the BlackPearl system to meet your site's evolving storage requirements, you can easily add capacity and performance by adding ArcticBlue expansion node. The expansion node connects to the master node using external SAS cabling. ArcticBlue supports up to 8 expansion nodes.

Drive storage bays:

	Total Drive Bays	Front Drive Bay Access	Back Drive Bay Access	Top Load Access
ArcticBlue node*	96			96

** All expansion nodes are 4U*

Modular Design: On-Site, Swappable Components

Spectra Logic designs products so that major components are optionally customer-replaceable, so that customers who would prefer to maintain their system can do so with user-installable on-site spare components.

Spectra disk systems are the only ones in the industry that are available with on-site, user-replaceable components that include:

- System drive
- Power supplies
- Fans
- 10/40 GigE Ethernet cards
- Boot Drives

Management and Reporting Features

Spectra BlackPearl's many built-in management and reporting features simplify its management and monitoring. The graphical interface, shown in an earlier section and in following sections, increases system ease of use.

Command Line Interface

The command line interface provides an alternate method of accessing the BlackPearl system, duplicating the functions available through the graphical interface. System administrators may use SSH to remotely access and manage the BlackPearl system using commands and associated parameters.

For example, use the following command:

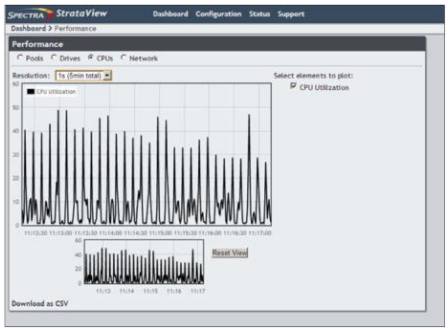
```
config mailrecipient delete --id=BlackPearl DPEEmail2
```

to delete the mail recipient with the id of BlackPearl DPEEmail2.

SNMP Management Protocol

The BlackPearl system supports SNMP (Simple Network Management Protocol), with a MIB (Management Information Base) available through the web interface that can be used to communicate between the system and other servers on the network.

Performance Monitoring

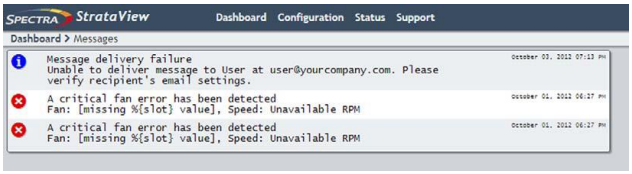


The performance pane displays the performance of

- Disk Storage Pools
- CPUs
- Network
- Tape Drives

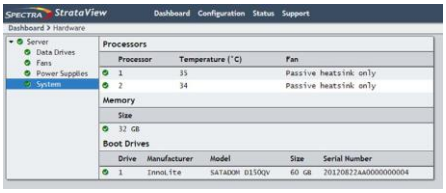
System Messages

The BlackPearl system provides ready access to error messages, including flagged messages that may require attention.



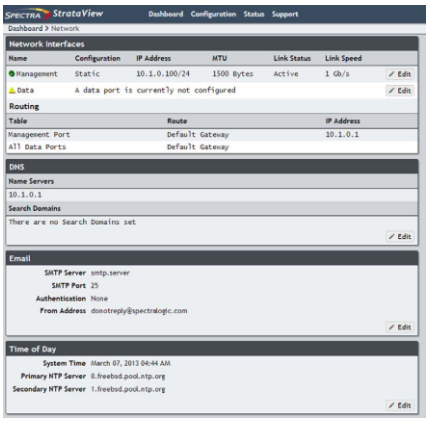
Hardware Status

Through a web browser from any location, you can use the web interface to check hardware status. From the interface's main dashboard, select hardware then select the component you are checking on. You can check status of data drives, fans, power supplies, and the system. The system screen provides information about processors, memory, and the boot drives.



Network Configuration

The system displays information about the configuration of the network, including all network connections and status of each, DNS servers, and email configuration. This greatly simplifies remote management of the BlackPearl unit.



Support and Continuity Features

A set of system features help expedite issue resolution. These features help identify possible issues and let administrators address them before they interfere with ongoing system operations. BlackPearl also provides on-site repair options, and is backed by Spectra Logic's around-the-clock support.

The following are some of the BlackPearl built-in support features.

AutoSupport Phone Home Feature

Spectra storage systems have an AutoSupport feature that can be configured to automatically create and send e-mail messages about issues or problems to designated e-mail users, including Spectra Logic Technical Support.

Paired Recovery

BlackPearl is designed with a modern, highly reliable single controller design. Typical deployments of the system do not require the complexity or expense of traditional high-availability disk storage solutions. In cases where users need a method that helps sites quickly recover from a system failure, BlackPearl offers ColdPair recovery options.

The ColdPair option for BlackPearl lets users quickly and economically recover from a master node failure. This involves storing on-site a spare master node without any hard drives or SSD. In the case of a master node failure, the customer can rapidly bring the system back up using the spare master node by following these steps: make sure the original BlackPearl system is powered off, move cables, network and HBAs, and data drives to the ColdPair chassis, then power up the ColdPair node. The system reads all configuration data from the migrated disks, as the disks store Replicated System Configuration (RSC). All pools, buckets, and data access typically comes on-line in less than 30 minutes.

Hot-Swappable Hard Drives

Hard drives in the ArcticBlue are on drive sleds that can be pulled out easily. With this, a failed drive can be replaced with a new one without requiring tools, and without requiring downtime.

Intelligent Rebuilds

When a drive fails, instead of rebuilding the entire drive, BlackPearl rebuilds only the portion of the drive that held data, potentially saving hours on rebuilds.

Redundant Power

Each BlackPearl master node and ArcticBlue expansion node ships with two redundant, high-efficiency power supplies.

SpectraGuard Support and Warranty Overview

The Spectra BlackPearl and ArcticBlue disk system has a warranty that extends one year from the date of installation. This warranty includes a SpectraGuard local business hours phone support and standard shipping replacement of parts after phone diagnosis. The following service options are also available to customers purchasing a Spectra data storage system. Refer to our website for more details: https://support.spectralogic.com/services-and-contracts/support-offerings/.

SpectraGuard Basic Warranty Extension: only available on a disk *only* system
Basic warranty support can be extended for years 2 through 5 at time of purchase or with annual contract renewals. Basic level of support includes SpectraGuard from 8:00 a.m. to 5:00 p.m. (customer local time) and standard shipping replacement of parts after phone diagnosis. Basic warranty extension is required for all support plans starting in year 2.

SpectraGuard 24x7 Phone support
An option to add 24x7 phone support to any service contract is available.

SpectraGuard Next Business Day On-Site Service: base level of support for tape libraries and for BlackPearl (with or without ArcticBlue nodes) when a tape library is attached

- Access to a SpectraGuard technical support representative on any business day (not including evenings, weekends, or holidays) from 8:00 a.m. to 5:00 p.m. (customer local time).
- A service visit from a field service representative, upon verification that the unit has malfunctioned. If Spectra Logic is notified by 4:00 p.m. (customer local time), a field service representative will be dispatched that day for arrival the following business day before 12 noon.
- 24x7 phone support can also be added to this plan as an uplifted option for
- Note that this is in addition to Basic Warranty support
- Technicians will not be dispatched for drive failures until the last available global spare has begun a rebuild.

SpectraGuard Four-Hour On-Site Service

- 4-hour on-site service, 5 days a week, from 8:00 a.m. to 5:00 p.m. (customer local time).

- 24 x 7 x 365 telephone access to a SpectraGuard technical representative (includes evenings, weekends, and holidays).
- Upon verification that the unit has malfunctioned, a field service representative is dispatched to the site within four hours of the dispatch request.
- Technicians will not be dispatched for drive failures until the last available global spare has begun a rebuild.

Professional Services

Spectra Logic's professional services group offers additional on-site services for prevention, maintenance, and site-specific consulting. These services include:

- Preventive maintenance
- Customized training
- Optimization services
- Media migration support
- Security assessment and consulting
- System consolidation
- Backup and disaster-recovery consulting
- System relocation and reintegration services

Assisted Self-Maintenance Support Options

The Assisted Self-Maintenance option is available to customers purchasing a Spectra BlackPearl and ArcticBlue disk system. This option can be purchased to supplement the on-site service options above. Refer to the *Modular Design: On-Site, Swappable Components* section for details.

Conclusion

The growth of data has driven the accretion of storage in the data center, with most storage built on general-purpose platforms: SATA/JBOD disk and tape. The data growth has led to new uses of data, along with increased demands for data protection and long-term data retention. The current model of data use and storage, starting with creation on primary disk and then storage on SATA and tape for backup and disaster recovery, no longer sufficiently support how data is accessed and used. These uses for and demands on data require a new class of data storage that integrates specific features that ensure high data integrity through longer periods of time, all on systems that are affordable and easy to use.

This new model is implemented in ArcticBlue. With a focus on data integrity and long-term storage, along with flexibility, scalability, and affordability, storage can now tackle the serious issues of long-term data storage and protection while supporting easier management of steadily increasing data growth.

Specifications

Environmental Specifications

Description	Values	
Temperature range - operating environment	10° C to 21°C**	50° F to 70° F**
Temperature range – environment when storing & shipping	-40° C to 70° C	-40° F to 158° F
Relative humidity - operating environment	8%-90% (non-condensing)	
Relative humidity –environment when storing & shipping	5%- 95% (non-condensing)	
Altitude - operating environment	Sea level to 3,048 meters	Sea level to 10,000 feet
Altitude – environment when storing & shipping	Sea level to 12,000 meters	Sea level to 39,370 feet
Maximum wet bulb temperature - operating environment	29° C	84° F
Maximum wet bulb temperature– environment when storing & shipping	35° C	95° F

Power

Unit	Specification
Current	4.2 amps (4U master) 4 amps (2U master) 4.5 amps (expansion)
Input frequency	50-60 Hz
BlackPearl 4U Master Input voltage	100-140 VAC, 12-8 A, 1,000 W maximum 180-240 VAC, 8-6 A, 1,280 W maximum
ArcticBlue Expansion Node Input voltage	100-140 VAC, 13.5-9.5 A, 1,100 W maximum 180-240 VAC, 9.5-7 A, 1,400 W maximum

Data Storage

Parameter	Specification
Drive type	5600 RPM SMR SATA
Single drive capacity, native	8TB
System capacity options	**BlackPearl Master Node** • No SMR data storage **ArcticBlue Node** • Minimum – 384TB (first node only, 192TB on second+ node) • Maximum – 768TB

BlackPearl System (4U)

Parameter	Specification
CPU	2 multi-core processors
System boot drives	Two boot HDDs
Memory	128 GB (8 x 16 GB DIMMs)
Interface connections	• 2 integrated 10Gigabit 10GBase-t Ethernet ports • 1 dual-port 10 Gigabit Ethernet NIC • 1 dual-port 10 Gbase-T Ethernet NIC (option) • 1 dual-port 40 Gigabit Ethernet NIC (option)

** Ambient temperature maximum is a critical installation requirement. Disk drive operation temperature is carefully monitored and if exceeded will void the warranty.

All Specifications, features, functions, performance metrics and any future plans are subject to change without notice at the sole discretion of Spectra Logic. Nothing in this document constitutes a guarantee of performance or any commitment on the part of Spectra Logic.

www.SpectraLogic.com

Spectra Logic Corporation

6285 Lookout Road
Boulder Colorado 80301 USA

800.833.1132
303.449.6400

For more information, please visit www.SpectraLogic.com