NYRIAD

BeeGFS and UltraIO™ Combine to Create a More Efficient and Simplified Enterprise Storage Option

WHITE PAPER

What is Nyriad UltraIO™?

UltraIO™, Nyriad's flagship product, is fundamentally changing the foundations of storage by simplifying how data is stored, protected, accessed, and managed. Our new erasure-code-based architecture utilizes patented algorithms powered by GPUs and CPUs to deliver extreme performance, resilience, and efficiency, enabling massive amounts of data to be managed in a single storage platform. Achieving 20 GB/s sustained throughput with its intelligent data placement (which uses all of the available drives all of the time) delivers an incredibly dense and efficient solution. Nyriad delivers a POSIX compliant storage solution that empowers businesses to grow, adapt, and stay competitive in a data-driven world.

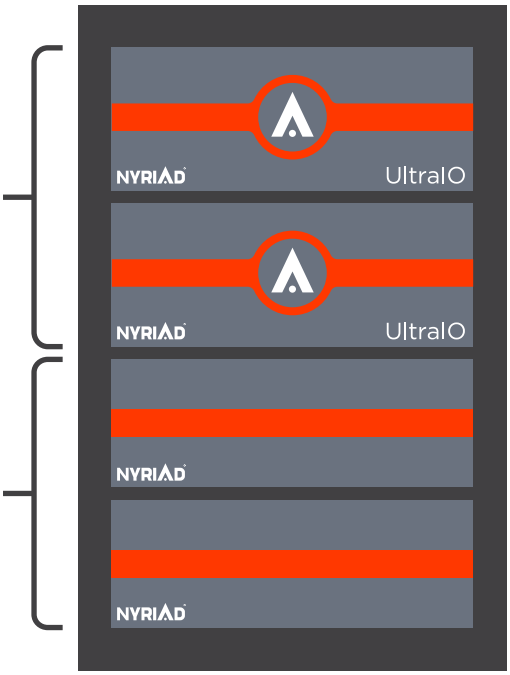
Dual HA Controllers

Up to 20 GB/s throughput

Up to 200 Gbps Network throughput

3.67 PB raw capacity per array

Up to 3.2 PB usable capacity



3.67 raw capacity as shown

RESTful API, GUI CLI

Simplified Storage Management

Unique CPU.GPU Block Erasure Code Architecture

Additional Checksum with real time rebuild

Intelligent Data Placement

Figure 1

In a combined GPU + CPU processor architecture array, an erasure-coded pool of 102 drives can be created with 10 redundant drives. In this case, data would only be lost after the 11th drive failure, essentially having 5x the resilience of a typical RAID 6, 8+2 architecture. It would also have half the capacity tied up in parity of 100 drives as 10 separate 8+2 configurations. Moreover, the Nyriad UltraIO array maintains near optimal performance[1] even after losing all 10 redundant devices (see Fig. 2). In the event of eventually rebuilding the array, we provide an extremely quick rebuild to reach an optimal state.

Nyriad extends this concept to account for as many as 20 drives out of 204 to fail, with no data loss. For those not needing this high level of protection, users can choose as few as 5 parity segments per 102 drive enclosure and save capacity. The large data stripes allow for extremely fast rebuilds, as the data needed to rebuild is pulled from a large number of drives. This limits the rebuild effect on the array performance, in addition to speeding up the rebuild itself.

The UltraIO™ storage system allows the user to configure how much resilience they require. If you want to focus on extreme performance and capacity, you might choose less. If you want to have the array run for many years without touching it, you might configure more. In a 204 drive array, with a 204 drive wide RAID stripe, as an example, the user could choose to use 10 total parity drives. This means only 5% of the total capacity is consumed by parity while still allowing a level of protection for any 10 drives to fail in the array with no data loss.

Nyriad accelerates the UltraIO array by taking advantage of parallel access to many drives at once. Data is placed in intelligent patterns that ensure all the drives are not only kept busy, but are utilized in a streaming pattern to best take advantage of the drives most efficient sequential data access pattern. This ensures a high level of overall performance while enabling the use of cost-efficient high capacity drives. Data written in sequential patterns are usually read in sequential patterns and implementing write placement to account for this leads to both accelerated reads and writes.

[1]Based on Drive-pull IOR bandwidth on UltraIO v1.0 and BeeGFS 7.2.5. Test performed by System Fabric Works 7.12.2021

As an additional benefit, the UltraIO array takes advantage of checksums for each data stripe segment placed on the individual drives. Arrays with no checksum capabilities rely on wasted performance and power from patrol reads, or data scrubbing. In that scenario, the array scrubs the data looking for bad bits to correct. Despite being wasteful with resources, this still doesn't provide 100% protection, as a bit can be corrupted since the last pass of a data scrub routine.

With UltraIO, in the event an alpha particle changes a binary state, bit rot occurs, or some flaw in the media causes an indeterminate data state, checksum identifies the problem and fixes it transparently. The data is decoded with the checksum as each segment is read. If a bit(s) is found corrupted, the full stripe is read, the user receives their requested data and the stripe is rewritten in a new location while the bad segment is marked as unusable, all in a self-healing, transparent, operational path. All of the resources wasted on data scrubs/patrol reads are given back in the form of power savings or additional performance.

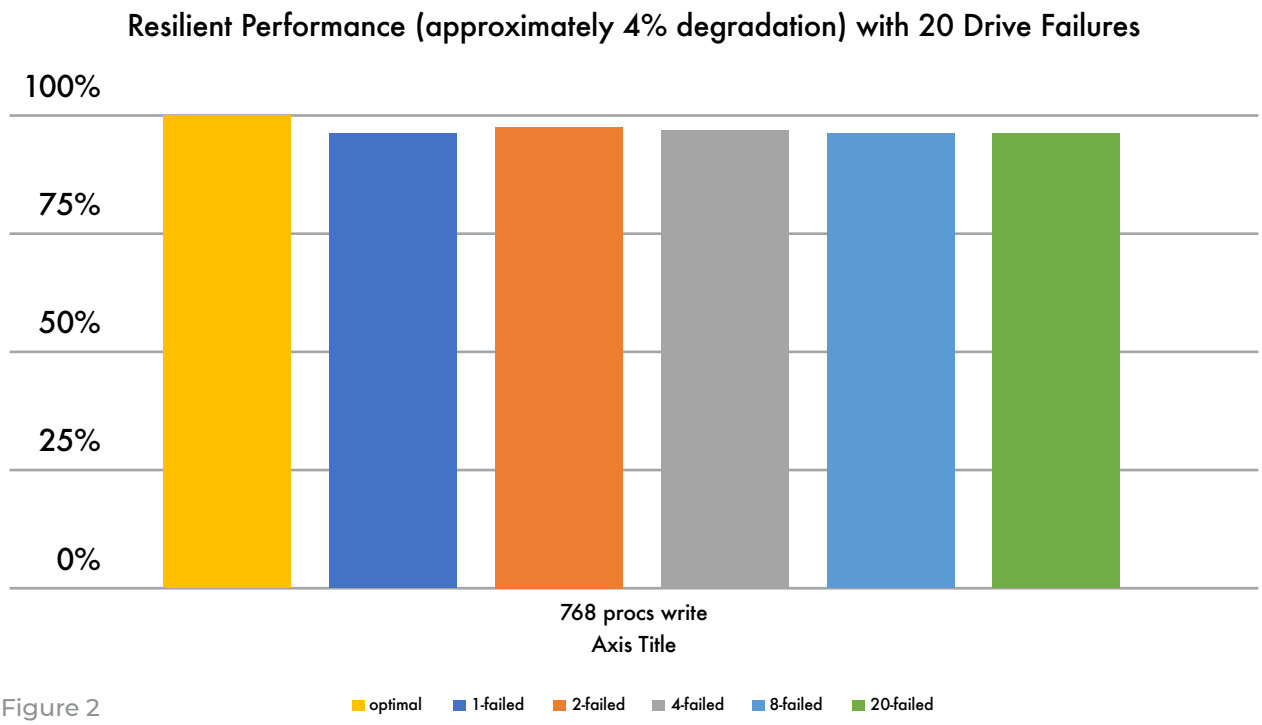


Figure 2

Resilient performance of combined GPU + CPU processor architecture with approximately 4% write performance degradation with up to 20 drives in a failed state. Based on Drive-pull IOR bandwidth on UltraIO v1.0 and BeeGFS 7.2.5.

Nyriad's extremely powerful combined GPU + CPU processor architecture is the next big step in storage, providing performance, resiliency, and efficiency:

- Sustained high performance[2], even with multiple drive failures
- Up to 20 drive fails with no data loss
- Extremely high, scalable performance
- Incredibly dense hardware footprint
- Ability to detect and recover from nearly all data corruption
- Lower power and cooling requirements versus traditional small capacity scale out enclosures
- Storage workload enhanced and stabilized with GPU and CPU involvement
- Built with off-the-shelf hardware
- Compatible with existing POSIX compliant storage targets and file systems

The UltraIO system delivers performance density and efficient drive and data failure protection in the most cost efficient architecture. What does this mean for a BeeGFS environment?

BeeGFS

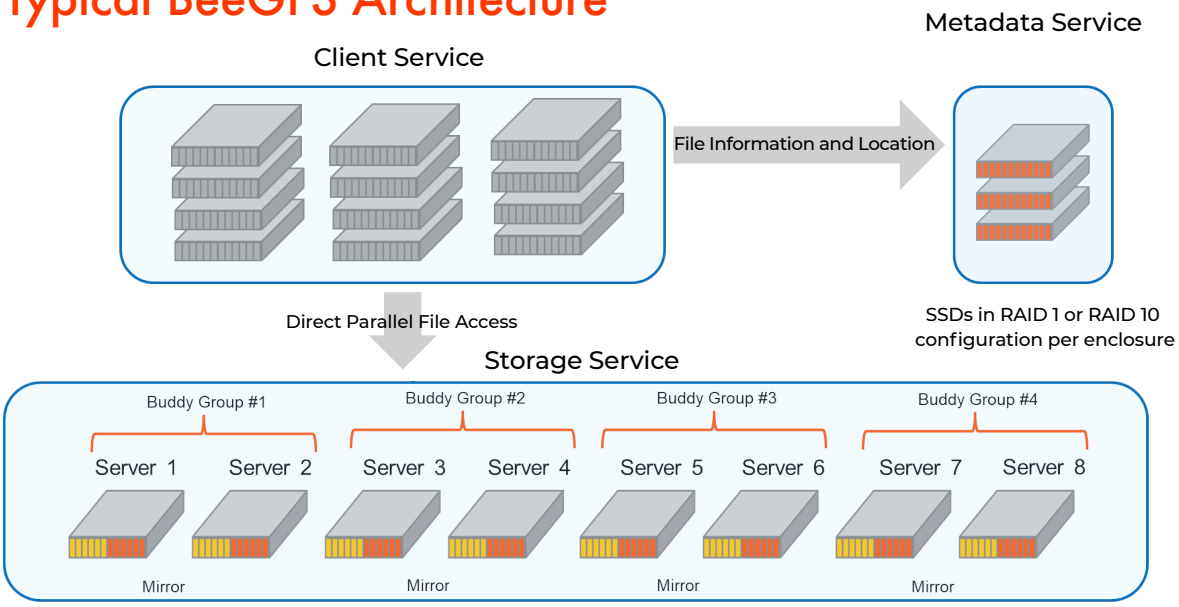
BeeGFS is a parallel clustered file system. The primary focus is on maximum performance and scalability. Performance density is determined by the performance and capacity point per storage target. The scalability comes from the clustering aspect that allows for capacity and overall performance to be scaled out by seamlessly adding additional targets to the namespace. BeeGFS spreads, or stripes, data across multiple storage targets. The storage targets are typically servers running RAID 6. This allows a user to seamlessly aggregate capacity and performance of many servers.

BeeGFS is a scale-out network file system that is POSIX compliant, while POSIX (Portable Operating System Interface for Unix) is a set of IEEE standards. From a storage target point of view, POSIX compliance makes sure there are clear standards dictating how files are accessed, written and read. This allows solutions to become more HW agnostic, for more widespread use of Software Defined Storage offerings.

BeeGFS stores user data and metadata. The user data is stored in a parallel fashion in the Storage Service domain. The metadata is typically stored in a separate location from user data in what is called the Metadata Services domain.

[2]Based on Drive-pull IOR bandwidth on UltraIO v1.0 and BeeGFS 7.2.5.
Test performed by System Fabric Works 7.12.2021

Typical BeeGFS Architecture



- RAID 6 on each enclosure to protext for 2 drive failures per enclosrue
- Every 2 server enclosures are mirrored together to protect against server failure in addition to drive failure

Figure 3

Client Service mounts the file system to access the stored data, reaching out to the Metadata Services to determine where the data resides and whether the request is valid. The Client Service also informs the Metadata Service when a new write is made or if an existing file is modified. Once the file location for a read is determined or a new write has been assigned a location, the Client Service will directly engage the Storage Service and access the server(s) where the read or write is taking place.

Metadata Service stores access permissions for data and information about the striping on the storage services. Think of this as the gatekeeper to protect who accesses what data. For those who have access to a particular file, it provides a map of where the data resides. When a file is written or changed, the metadata is updated to reflect the changes.

Storage Service stores the distributed user file contents and allows file access when requested. The data protection with RAID 6 happens inside of the Storage Domain, but is handled by local data protection methods such as HW RAID HBAs, or head nodes with no interaction with the services that interact with the Client Services. The onboard HW RAID interacts with the Management Service for tracking health and configuration needs. To the Storage Service in general, the HW RAID protection is transparent and autonomous.

Management Service is a 4th service in BeeGFS. This is a registry and watchdog for all other services. You can use it to add or remove components from any of the other

services, as well as monitor the health of their components.

Performance Considerations for Metadata Services

The overall speed of the File System is directly tied to how fast the client can be authorized to access a file and given a pathway to reach that file. In addition, after a write or modification of a file, the write isn't complete until the metadata has been updated. As a result, if the access to stored metadata is slow, the entire read/write path to data will be slow. This happens no matter how much performance the underlying data storage has. If the user has to wait to get a clear path to the data, or the write can't be fully committed until the metadata completes the update, then the CPU must wait on storage to complete operations. In this case, the Client Service CPU as well as the Storage Service are waiting on the Metadata service to complete a task. As a result of this performance sensitivity, Metadata Services are typically built with SSDs in a performance rich, but capacity inefficient, mirror or RAID 10 configuration. As the user scales out more Storage Servers in the Storage Service domain, or adds more clients in the Client Service domain, the Metadata Service has to scale as well. This is to allow for more capacity to service more file mapping and provide increased performance to accommodate a growing number of file accesses.

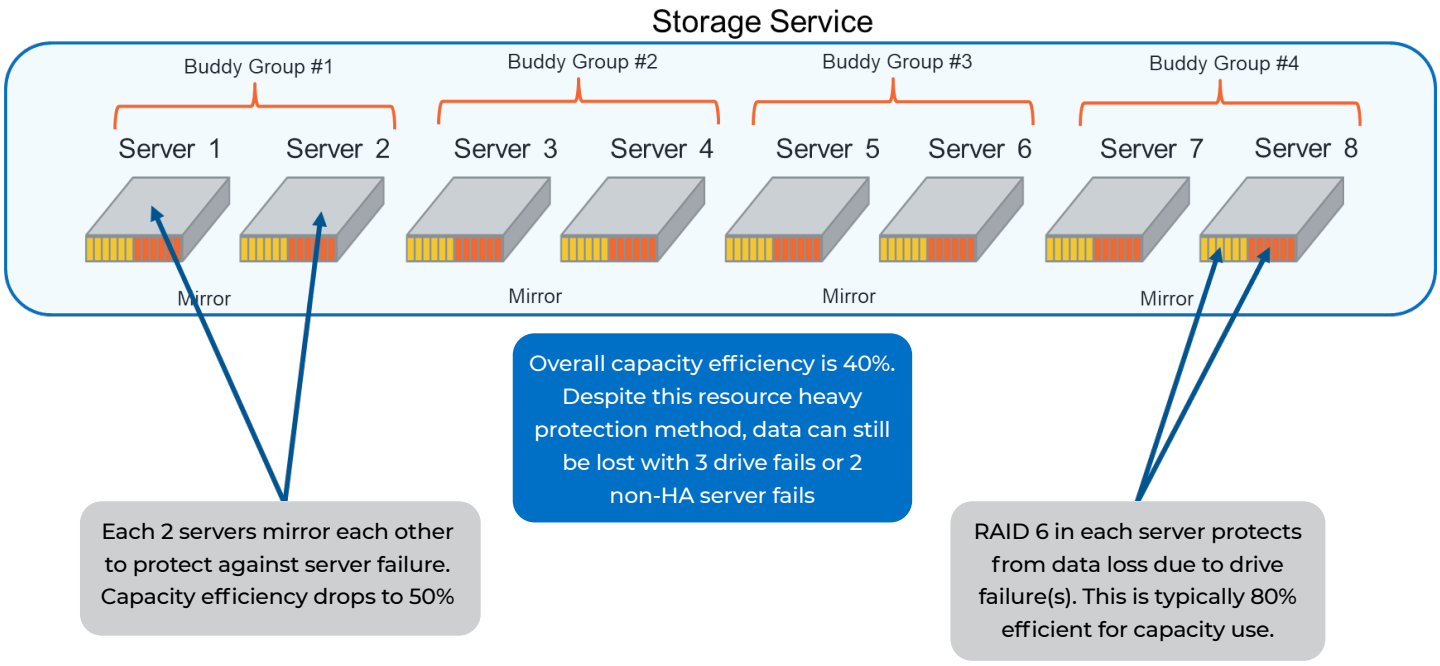


Figure 4

The trend in the industry had been to move away from centralized storage and rely on individual non-HA HW elements for additional redundancy to make up for the HW reliability. This is clearly present in Hadoop with three racks holding three copies of data. Ceph was originally built around a similar model of at least three copies to overcome the failure rate of non-HA HW. BeeGFS does something similar: 60% of the BeeGFS storage capacity in the reference design above is used to overcome the reliability weaknesses of the non-HA HW.

As most of those reading this know, this design proved a better solution than average SAN solutions. Large scale capacities of storage tied into a single SAN was a massive failure domain and had to be bulletproof to ensure a companies' livelihood wouldn't be lost by a storage failure. Every component and every data path was in duplicate, all resulting in an expensive cost structure that was complicated to manage. Scale out designs with multiple copies utilized by BeeGFS were just as reliable and less expensive. This was true even with the triple copy approach seen in Ceph, Hadoop and others.

Typical RAID 5 or RAID 6 designs in traditional SAN aren't convenient to use large capacity drives. The width of the RAID stripe is limited in the width most would deploy, because there's only 1 or 2 drive failures worth of protection. Using large capacity drives in traditional RAID SAN is painful because if you limit the width of the drive stripe, you have a limited number of drives in the rebuild. With only 2 drives of failure protection, long rebuild times slowed by the small number of drives in the rebuild limit drive capacity points. This leads many to have hot spares built into the design, so the rebuild can start immediately, but this wastes expensive capacity and doesn't help with actual rebuild time. Disaggregated or Distributed RAID (DRAID) speeds up rebuilds by laying stripes out over larger drive counts, allowing more drives in the rebuild. Yet DRAID leads to complicated and expensive designs that require multiple enclosures to ensure a failure doesn't take the array down.

The UltraIO system takes a new approach. Utilizing the combined power of the CPU and GPU, UltraIO storage provides an erasure code based solution that provides the best aspects of performance, efficiency and resiliency. GPU driven erasure codes allow for performance that is protected by up to 20 drive failures. This level of erasure code protection delivers up to 204 drive wide stripes in a single erasure code group. The large erasure code group means an extremely high level of parallelism, and thus performance. In addition to high levels of resiliency and performance, intelligent

data placement enables all write workloads to be aggregated and sent to the array in a sequential pattern. This utilizes the strengths of hard disk drives while allowing for the use of large capacity, cost efficient drives.

This design's resilient, large capacity arrays provide the simplicity of a single large array and enjoy the cost reduction seen in scale out non-HA designs, all the while proving to be better than both of them.

In BeeGFS when a client wants to retrieve data a 2-step process is followed.

Step one

The client reaches out to the Metadata Service requesting the location of the data. The metadata service ensures that the requesting client has security access to the data. If the Client has permission to access the data, the metadata provides the location(s) of the data.

Step Two

The client reaches out to the servers and locations and retrieves the data.

Potential Step Three

If the client chooses to alter the data, the metadata is made aware of the change and any change to location as the result of shrinking or growing the file.

An UltraIO storage design fits into this infrastructure seamlessly and has an extremely positive effect. Instead of writing to data stripes on individual servers, BeeGFS can simply write to stripes that all reside on an individual UltraIO system (See Figure 5). An UltraIO array provides a more efficient level of underlying protection than the RAID 6 and Buddy Copy methods. RAID 6 is removed from the design and Buddy Copies are not needed since the UltraIO system provides the GPU-based Erasure Code protection.

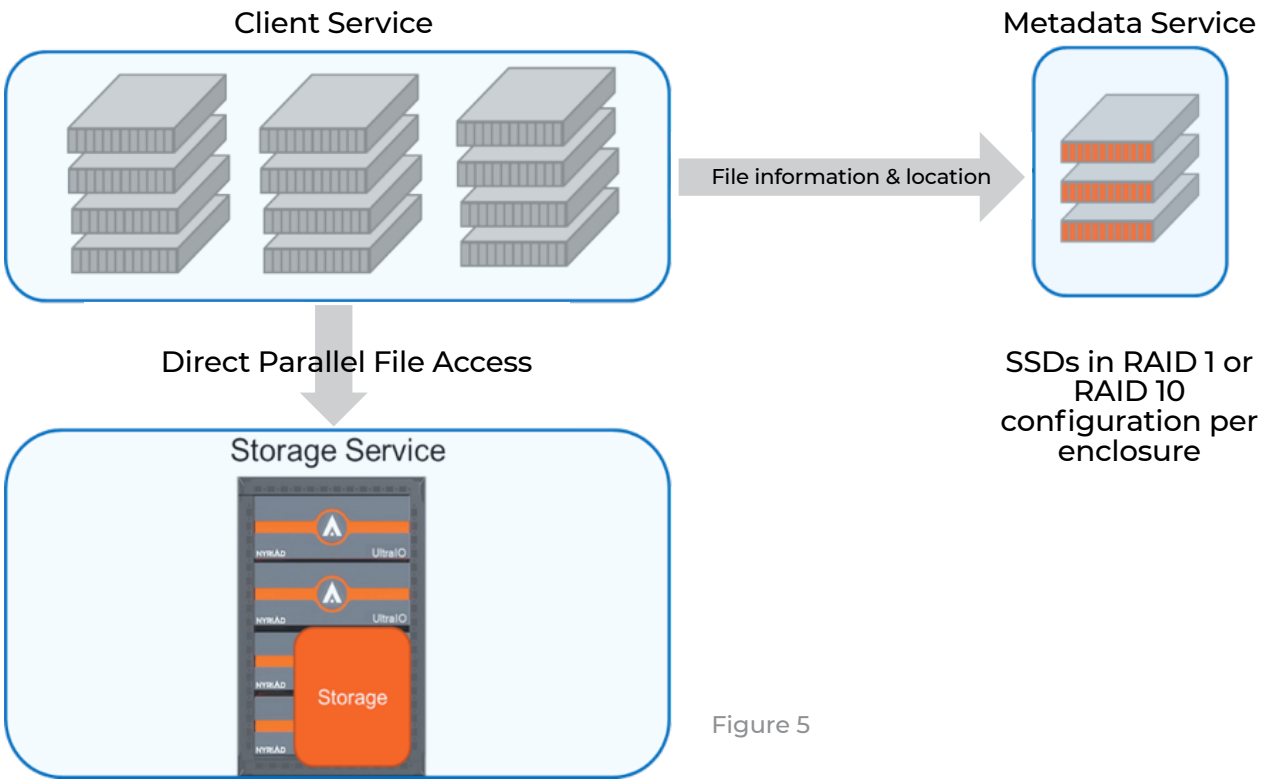


Figure 5

Of course, the design can be scaled in the same manner that would be used by individual server scale out and increased client Service Nodes (See Figure 6).

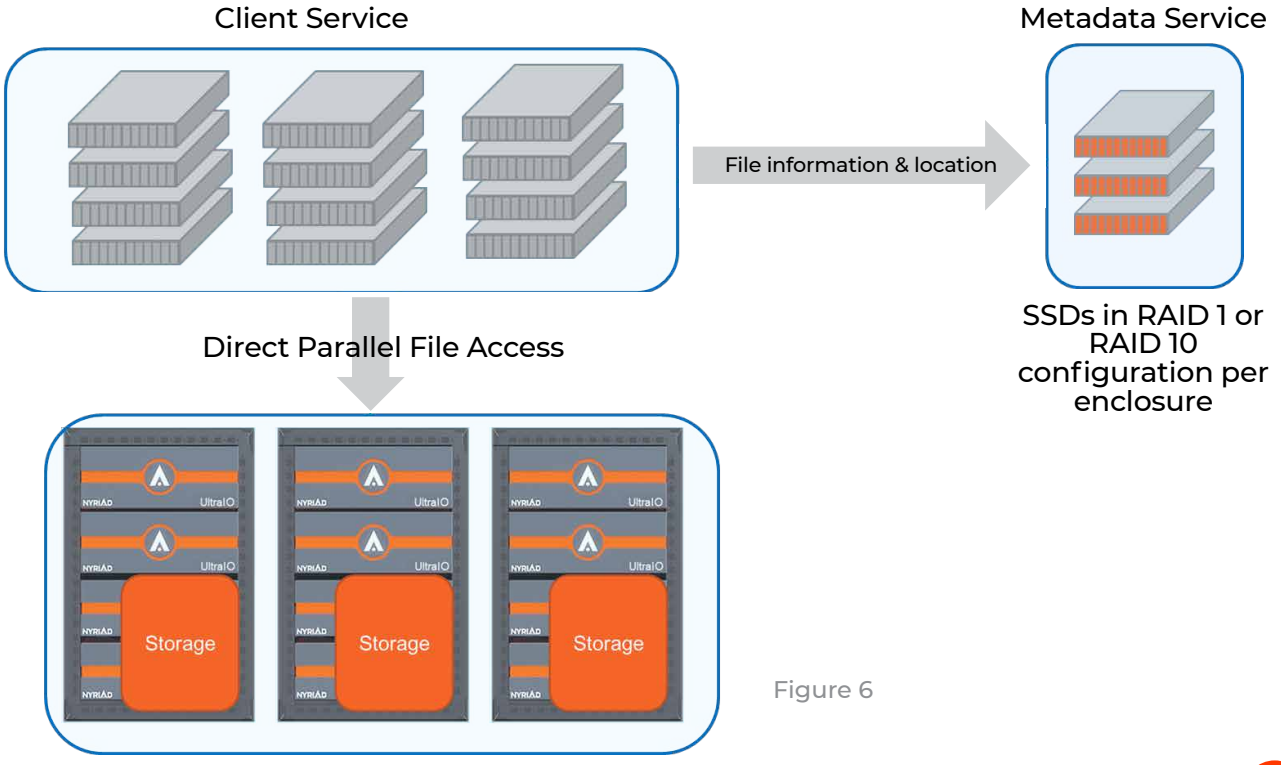


Figure 6

These illustrations show how to build a BeeGFS design that utilizes UltraIO storage. The benefits are many.

- Hardware footprint consolidation
- Lower power costs
- Lower cooling costs
- Sustainability advantage (Smaller CO_2 footprint versus original)
- Ability to fail 20 drives per array with no data loss, while existing RAID 6 and buddy mirror can lose data if the right 5 drives fail, even with 60% of the capacity consumed by redundancy
- Provides 20 drives of protection but only consumes 10% of the storage capacity
- Checksum corrects bit error, bit rot, and bits residing on damaged drive surfaces, eliminating data scrubbing
- Checksums protect against data corruption at a higher level than data scrubbing. Bits can become corrupted since the last data scrub, but checksums catch them all.

Benefits of Combining BeeGFS and UltraIO

To illustrate the benefits, let's look at an example.

- An UltraIO array with 20 drives of redundancy and a total of 204 drives (All hard disk drives) in a Storage Service for a BeeGFS design
  - 2.88 PB of usable capacity
  - 4,100 watts per array
  - 16u hardware footprint
- A traditional BeeGFS design utilizing individual servers.
  - 2.88 PB usable capacity
  - 60 servers needed to provide 2.88 PB of usable capacity
    - 3u server with 16 drives each and 700 watts consumed per server
    - 8 TB HDDs
    - 75% efficiency per RAID array plus buddy mirroring at 50% additional overhead
    - 16 x 8TB = 128TB raw capacity
    - (2) 6 + 2 RAID 6 = 75% efficiency = 96 TB usable
    - After buddy mirroring, 96 TB x 50% = 48 TB usable per server
    - 2,880 TB/48 TB = 60 servers needed
    - 42,000 watts per array (60 servers at 700 watts each)
- 180u of hardware footprint

UltraIO System Replaces This:

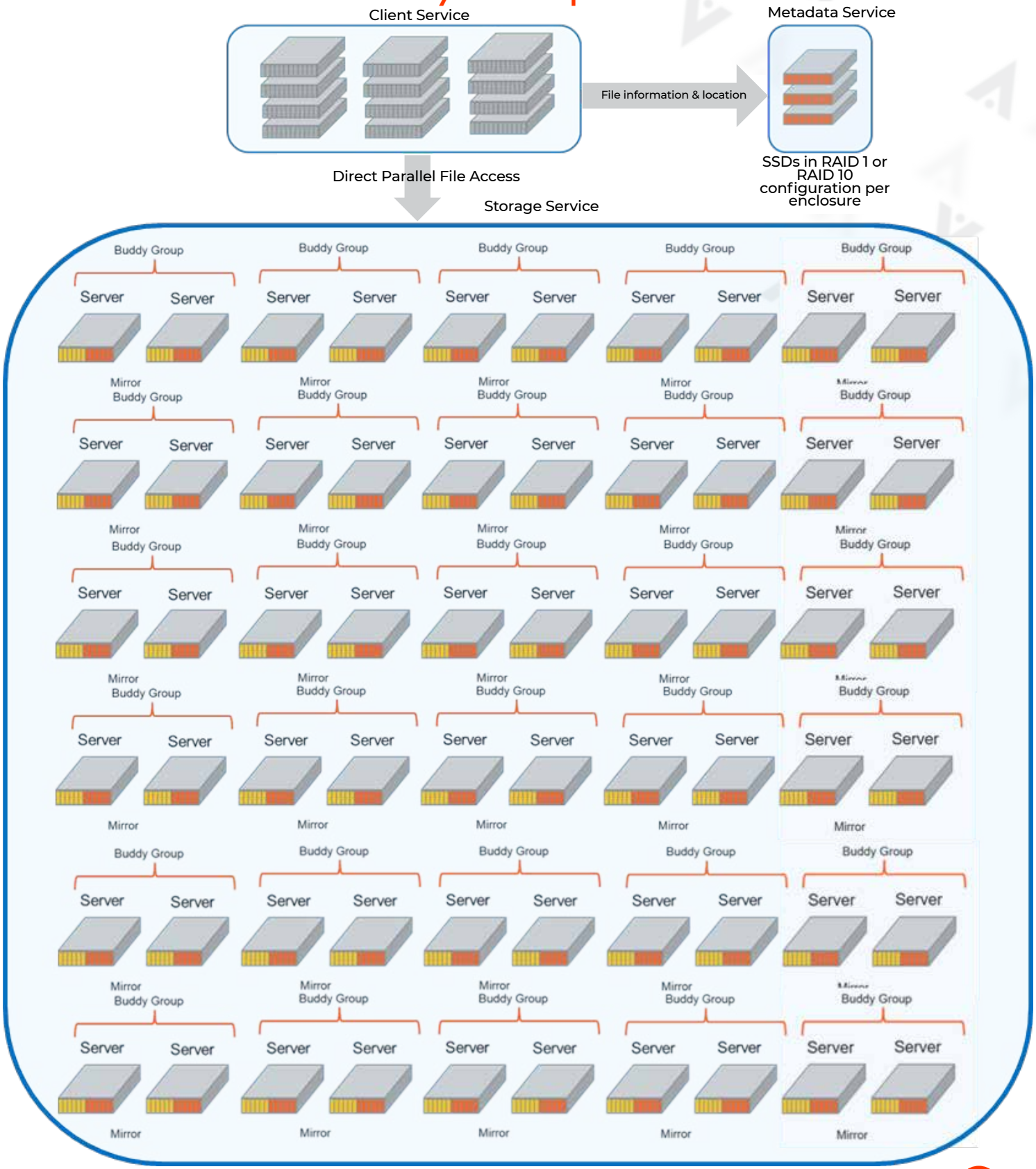


Figure 7

With This:

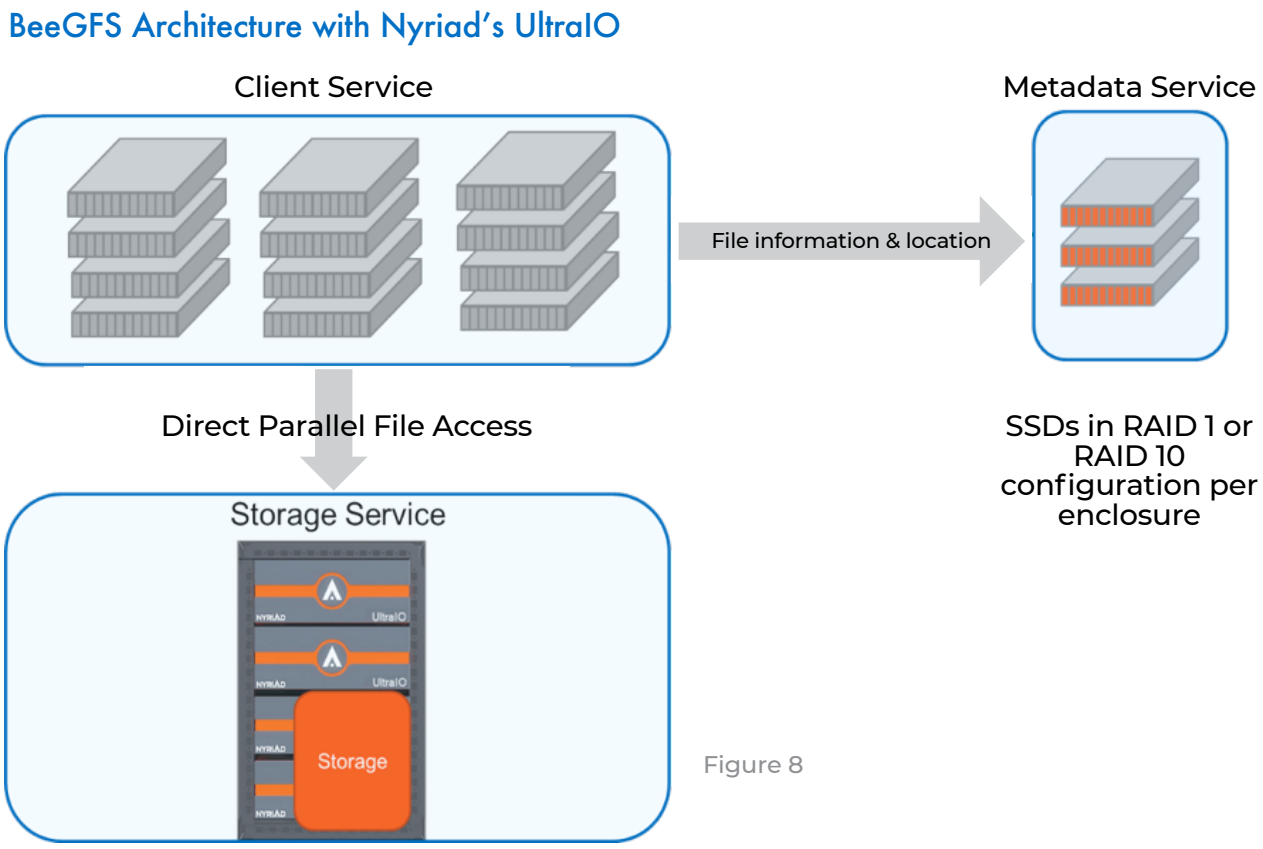
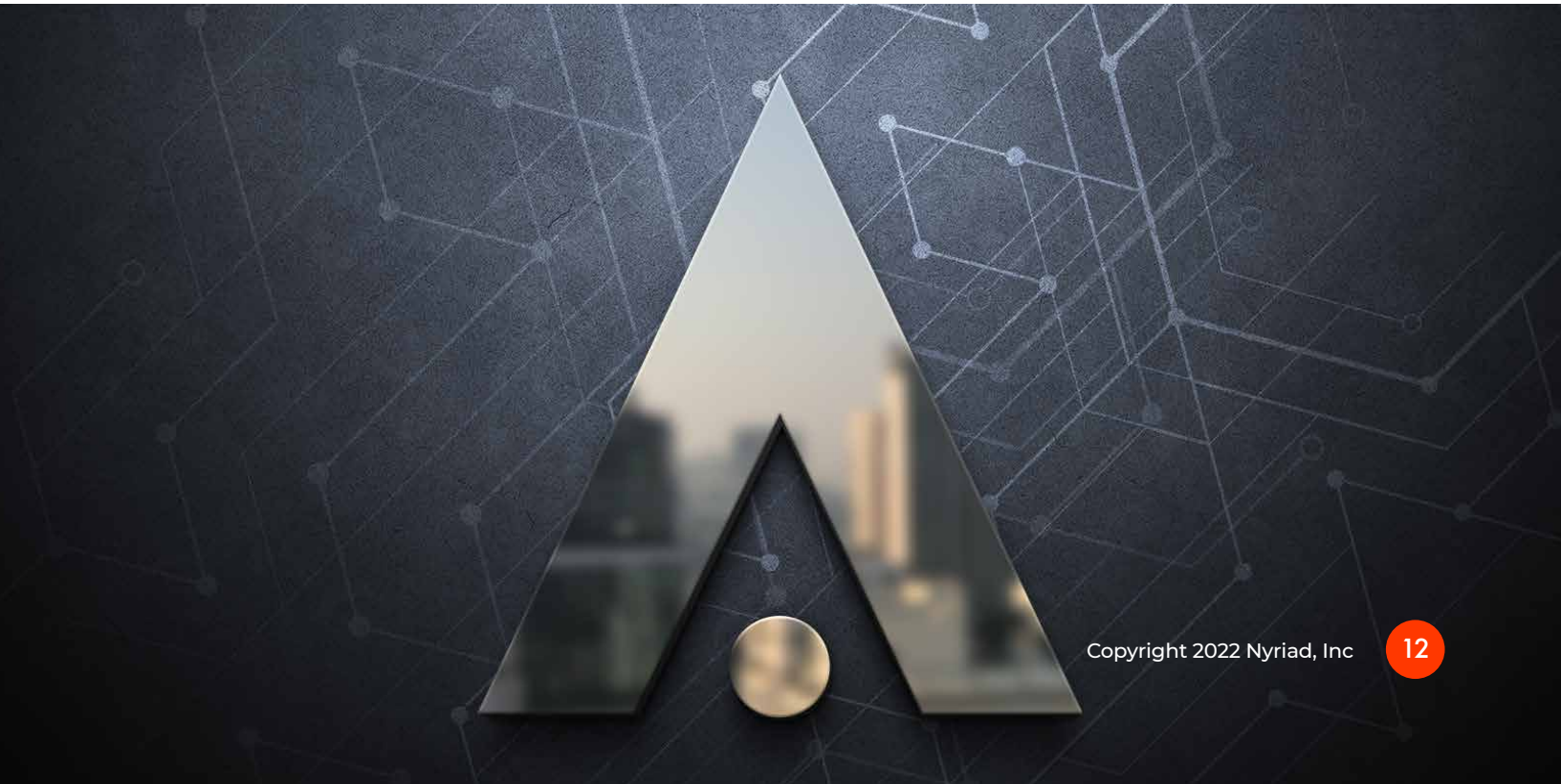


Figure 8



UltraIO system and Sustainability

The UltraIO storage system delivers 2.88 PB of usable capacity in a 16u form factor and consumes 4,100 watts. Traditional BeeGFS delivers 2.88 PB of usable capacity in a 180u form factor and consumes 42,000 watts (See Figure 9)

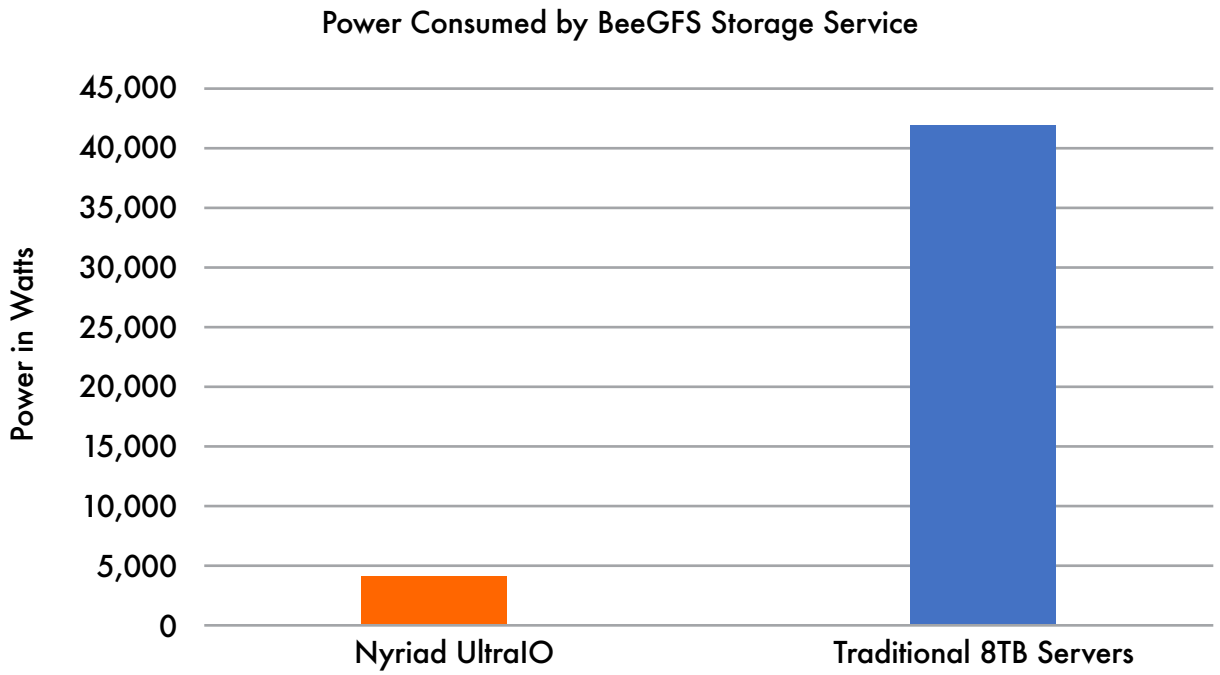


Figure 9

Of course, for all power consumed, there is power used for cooling. The ratio of power consumed versus cooling is expressed as Power Usage Efficiency (PUE). An international average of PUE is 1.7. This means that for every dollar of power spent on power consumed on running the equipment, there is an additional 70 cents spent cooling the area to keep the equipment operational. This cooling includes power consumed to run compressed gas HVAC or pump cool water and run ventilation fans etc. Assuming an average PUE of 1.7, the power numbers increase considerably (See Figure 10).

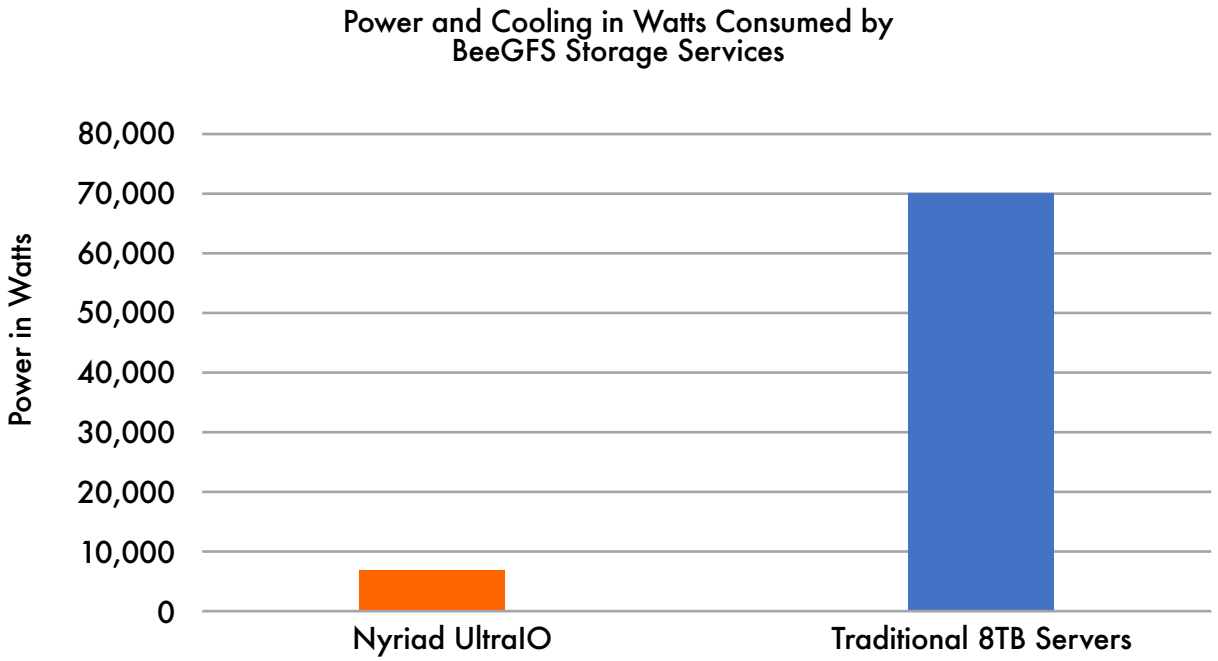


Figure 10

Power costs vary from state to state, so let's pick an acceptable average and use that in this example. We'll use 7 cents per every kWh of power consumed. The numbers above are an instantaneous power draw. There are 8,760 hours in a year, so this means that every kilowatt of power consumed produces 8,760 kWh per year. In this UltraIO storage example, the UltraIO array produces 61,057 kWh per year. The traditional 8TB Storage Service example produces 621,960 kWhours per year. This is more than a 10x delta in power usage. What does this look like with respect to cost to operate each solution? In this example, the UltraIO system costs $4,274 per year to operate. The traditional 8TB Storage Service example costs $43,537 per year to operate (See Figure 11). The move to UltraIO system saves $39,264 per year in power and cooling savings.

Keeping the system for five years results in $196,320 in power and cooling savings.

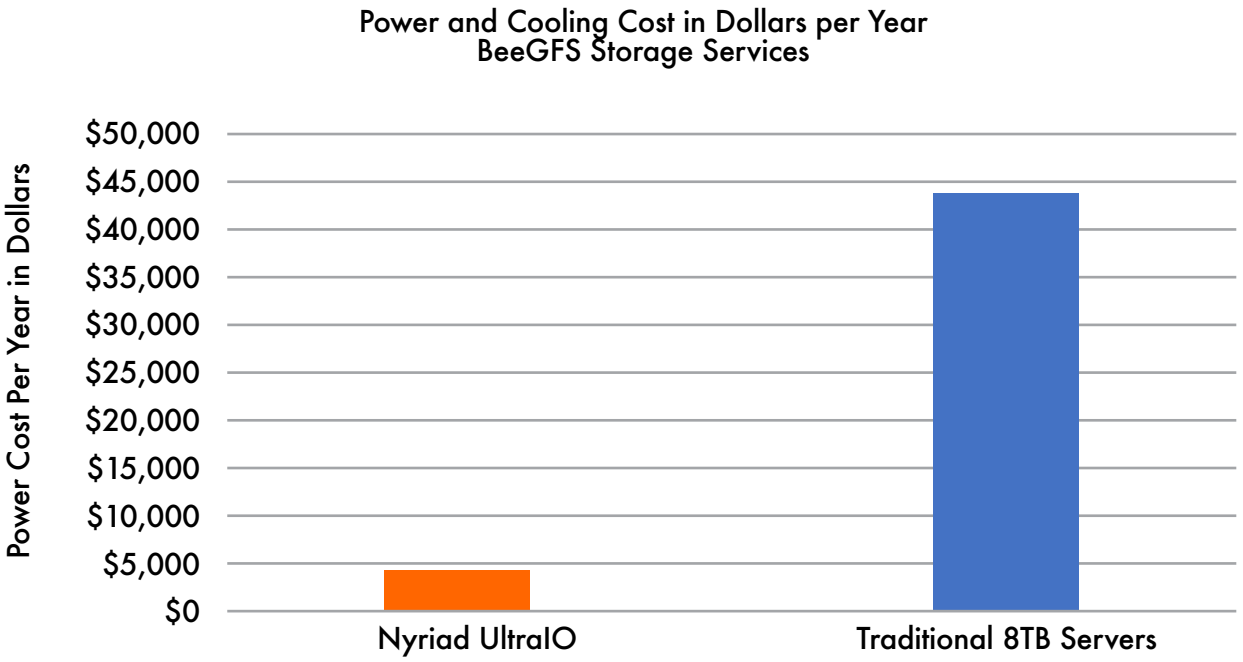


Figure 11

More power consumed means a larger CO_2 footprint. According to eia.gov[3], for 2021, only 12% of the power produced in the United States is renewable energy, with another 8% produced by nuclear sources. The remaining 80 percent is produced by fossil fuels. Unfortunately, 1/3 of the 12% that is listed as renewable energy is Biomass, which is essentially biofuels and wood burning, which also produces CO_2 like fossil fuels. Needing to achieve close to 100% uptime for compute and storage limits the ability to run these data centers on solar or wind. And less than 3% of the United States power comes from hydroelectric power. The reality is that most of the power driving today's IT infrastructure for both compute and storage is generated from Carbon emitting fuels, and result in CO_2 emissions.

Consider how much power and resultant CO_2 we generate with ordinary tasks. According to the University of Michigan's Center for Sustainable Systems[4], per kWh, Coal produces 2.2 pounds, petroleum, 2.0 pounds, and natural gas, 0.9 pounds of CO_2. Many sources will list CO_2 produced per kWh as an average of .954 pounds of CO_2 per kWHour or something similar. Our industry is not generally using wind or solar due to a need for consistent power and is rarely located near power plants, so this average is actually low for data center and IT calculations. Removing solar and wind power from

[3]U.S. energy facts explained - consumption and production - U.S. Energy Information Administration (EIA)

[4]http://www.css.umich.edu

the equation and coupling that with less efficient power usage means the results are impacted negatively. If we focus on an average assumption of 1.021 pounds of CO_2 per KWh, this assumes the bulk of power usage comes from fossil fuels with some rare occurrences of non CO_2 emitting sources.

We can calculate the amount of CO_2 produced per year for each of the examples. With 1.021 pounds of CO_2 produced for every kWh and the UltraIO system consuming 61,057 kWh of power per year, the UltraIO solution produces 62,339.2 pounds of CO_2 per year. That is 31.169 Tons of CO_2 produced per year.

Next, we can calculate the amount of CO_2 produced per year for each of the examples. With 1.021 pounds of CO_2 produced for every kWh and the traditional 8TB example consuming 621,960 kWh of power per year, the solution produces 635,021 pounds of CO_2 per year. That is a massive 317.51 Tons of CO_2 produced per year (See Figure 12). As expected, the ratio of CO_2 production is roughly 10x more for the 8 TB example.

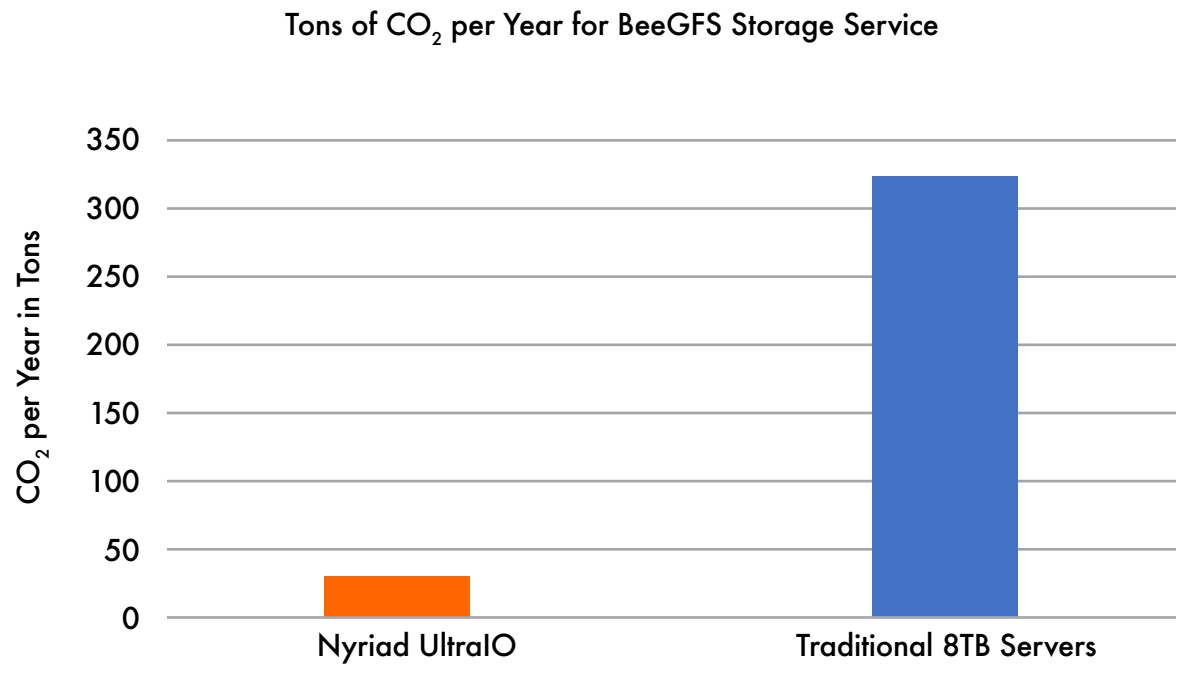


Figure 12

Summary

The UltraIO solution provides:

- Extremely resilient 20 drive fails (versus “the wrong five” failing in a RAID 6 and Buddy Mirror pair)
- Sustained 20 GB/s of performance per usable 2.88 PB array
- Usable capacity efficiency (increases from 40% to 92% depending on configuration)
- 10X reduction in power consumption
- 10x reduction in CO_2 produced
- Over 11x hardware footprint reduction (16u versus 180u)
- Simplified design that works seamlessly and transparently in BeeGFS environment
- Completely scalable in the same manner as traditional BeeGFS
- $196,320 in power and cooling savings over five years, compared to a traditional 8TB solution

About Nyriad

Nyriad, Inc. is the developer of the UltraIO storage system, an all-new system that combines the processing speed of GPUs and advanced algorithms to deliver unprecedented performance, resiliency, and efficiency. The ground-breaking design enables UltraIO systems to support block storage media and block, file, and object data types in a single system for maximum flexibility. UltraIO systems run on industry-standard hardware, use the highest capacity, most efficient storage media, and simplify storage management to achieve low total cost of ownership. Headquartered in Austin, Texas, Nyriad empowers businesses to grow and adapt their storage to stay competitive in a data-driven world. For more information, visit us on the web at www.nyriad.io.

NYRIAD®

www.nyriad.io