



**AmpliStor: Unbreakable Object Storage
for petabyte-scale unstructured data**

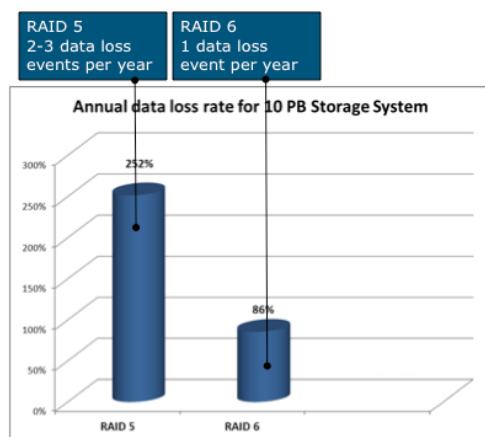
**Amplidata White Paper
April 2011
V 3.0**

Introduction: Online data growth and insufficient reliability of RAID based storage

Most storage industry experts agree that the market is ready for more efficient and reliable solutions for petabyte-scale online digital storage. The result of trends such as cloud computing, the success of online software applications (SaaS), the requirement to retain data for very long time periods and the growth of digitally stored data has exceeded the expectations of many of these experts. One study by IDC ⁽¹⁾ forecasts a thirty-fold growth in digital content to 35 million Petabytes by 2020, with only a 1.4 times increase in the number of people available to manage that storage. At this level of data growth, but a nearly flat talent pool, this also provides a strong indicator of the need for scalable storage systems that are increasingly self-managing and self-healing, to reduce the burden on human administrators.

Many online service providers forecast a healthy growth of their business – and thus in many cases their infrastructure will require 100+TB storage systems for production environments before the end of 2011. The challenge is how to build systems with these high capacities that also provide acceptable reliability and availability levels. Most of today's operational storage systems for online services use traditional RAID and replication based data protection mechanisms. These mechanisms were not designed for the capacities we are addressing today, and are therefore not cost-efficient, do not scale and – critically – suffer from *decreasing reliability and availability as the capacity increases*, making it very difficult to maintain advertised SLA's.

While most see the fast growing capacities of hard disks as benefits in footprint and lower cost, there are also some key negative aspects to the trend. RAID technologies, which were designed to avoid data loss due to failing hard disks, can no longer meet the availability requirements of modern storage systems. Large capacity disk drives will proportionally increase RAID rebuild times. Based on simple statistics, this implies an increasing probability of data loss as drive densities increase. During a RAID rebuild, data becomes unprotected (RAID5) or reduced in protection (RAID6). Statistically, this increases the chance that another one (or more) of the remaining disks will fail as well due to the increased activity.



groups will still nearly guarantee 1 data loss event per year, as demonstrated in the figure.

Studies have shown that RAID sets show increasing probability of a subsequent disk failure after the first failure. This is due to several factors including the fact that these disks are usually of the same age, from the same manufacturing lot, and have been subject to similar read/write patterns as the other disks in the RAID group. More importantly, the probability of a drive encountering an unrecoverable read error (due to bit rot) during a RAID rebuild is the most statistically inevitable event, and would lead to data loss during the rebuild process. In a 10PB data set, data stored in RAID5 ⁽²⁾ groups will statistically experience 2-3 data loss events per year. Storing the data in RAID6 ⁽²⁾

In modern SATA disk drives, drive manufacturers are providing ratings in the range of 1 bit read error in every 10^{14} bits read. This was not an issue a few years ago, but with the current generation of 2TB drives representing 1.6×10^{13} bits, this implies that reading just six drives will be sufficient to encounter such an error, on average. While bit rot or disk rot – the degeneration of stored data – is considered as a problem that rarely occurs and for which various protective

measures exist, the phenomenon is a nightmare for online archive managers. Data archives are very rarely accessed and the data that is kept in archives is extremely static and is hence rarely checked. RAID does not provide any active protection against bit errors or bit rot on disk, block replication or multiple data copies.

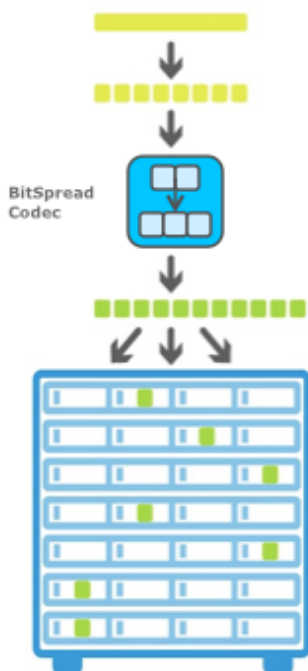
Additional technologies to handle the weaknesses of RAID, such as data replication (mirroring), only mask the issues with RAID and make other aspects worse. Notably, cost inefficiency and complexity of the infrastructure will become the key issues. Mirroring data doubles the raw storage required to protect data from the weaknesses of RAID (each 1 PB becomes 2.5 PB in a mirrored RAID6 configuration), and in some cases three copies of data are maintained (triple mirroring), in an effort to maintain SLA's, as one of the copies may be taken offline for rolling maintenance and upgrade windows. The cost of powering, cooling and housing 2 or 3 full copies of these large data sets becomes exorbitant. In many cases the power requirement alone becomes the main operational expense in a large-scale data center.

The poor energy efficiency of RAID systems has become another key issue. RAID arrays are extremely power hungry, mainly due to the fact that all disks in a RAID group need to keep spinning all the time. There is no possibility of spinning down one or more drives in a RAID group, even if they are infrequently accessed. This impacts power consumption even in data that is regularly dormant, such as archival data.

Finally, RAID systems do not retain a history of the stored data, making it impossible to go back in time to recover from possible data corruption. For all of these reasons, distributed storage architectures where data blocks are spread over low-cost appliances are gaining popularity and are making a strong case to replace RAID.

AmpliStor: Unbreakable and efficient distributed storage

Amplidata is creating next-generation distributed storage based on erasure-coding technology. Amplidata solutions use smart codecs to provide much higher availability and reliability than achievable with RAID 6. This strategy typically requires 50-70% less storage capacity than replication-based systems while consuming a fraction of the power. The key Amplidata technology is the BitSpread codec that splits and encodes data objects into multiple blocks that encode redundancy within the data blocks. These blocks are spread widely over the entire storage infrastructure. The codec requires only a subset of these blocks to restore the original data object, as determined by the user-specified availability policy. In this way, multiple devices can fail simultaneously without data loss. In large scaled out systems, full racks or even data centers can fail and data can be reconstructed from the remaining blocks. The BitSpread approach creates a system that increases availability and reliability as it scales, as opposed to RAID systems that suffer the opposite effect. In a multi-petabyte data set, an AmpliStor system is 100,000's of times more reliable than a RAID5 or RAID6 system.



The Amplidata distributed storage system is termed AmpliStor. The AmpliStor solution is broadly applicable to a range of storage use cases, and was purposely designed to comply with the reliability, availability and efficiency requirements of petabyte-scale online data storage, cloud computing, SaaS and archival

systems. AmpliStor employs the BitSpread mechanism to eliminate the problems incurred by RAID on high-density (multi-terabyte) disk drives at large scale. The system distributes data across storage nodes, by splitting and encoding stored data objects into many smaller blocks. The system only needs a subset of these blocks to restore the original data object. This allows for one or several devices to fail or become unavailable without data loss or affecting data availability.

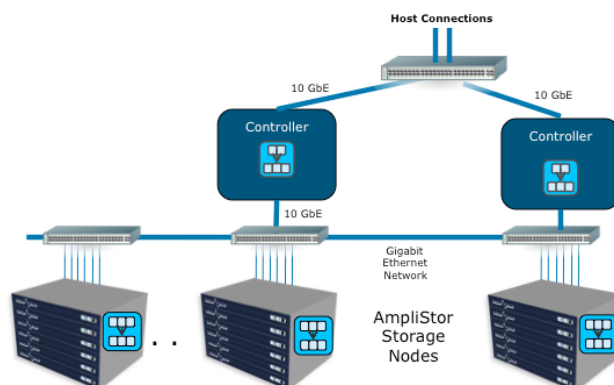
Key Advantages of AmpliStor

The AmpliStor approach provides solutions to several critical problems incurred by systems based on traditional data protection schemes. By eliminating the use of RAID and replication, AmpliStor provides the following benefits:

- **AmpliStor provides the highest levels of data reliability and availability:** AmpliStor provides policy-based reliability to protect data against any number of simultaneous failures, plus provides active data integrity assurance. RAID technologies become less reliable as disk densities and RAID groups grow. In contrast, AmpliStor actually becomes more reliable and available as data is spread over more disks and nodes as the system scales.
- **AmpliStor provides the lowest storage overhead:** RAID and replication requires doubling or tripling (2-3x higher) the raw capacity over the usable capacity to protect data from loss. AmpliStor provides better protection from data loss without doubling the storage capacity. This means that AmpliStor will require much less data center rack and floor space, as well as lowering Capex and maintenance costs.
- **AmpliStor provides low power consumption:** AmpliStor has the lowest storage overhead, can leverage low-power disk drives, plus power efficient storage enclosures to provide storage with the lowest Watts per TB.
- **Scale-out capacity and performance:** Additional performance or capacity can be incrementally added to the system on-the-fly. The system also auto-detects and automatically utilizes new capacity, to auto-scale the system with minimal need for human intervention.
- **AmpliStor provides low Total Cost of Ownership:** lower Capex, lower data center floor space, lowest power and cooling costs, and lowest management overhead leads to an incredibly low TCO.

AmpliStor System Architecture

The AmpliStor system is physically comprised of multiple AmpliStor Storage Nodes, which provide the back-end storage (disk enclosures and processing) capacity. Application data access is through the AmpliStor controllers, which integrate the patent-pending BitSpread encoder. The controllers are connected to the AmpliStor nodes through 10 and 1 Gigabit Ethernet networks, as described further below.



The controller provides native object interfaces for object storage services. The system provides an http/REST style API (with PUT, GET, DELETE style access semantics), as well as a C language library and a Python based command line interface (CLI). Controllers provide high-throughput access for large object & file storage and retrieval. Throughput is scalable with multiple controllers, with fully shared access to the back-end storage pool, even for concurrent writers.

AmpliStor integrates the following key Amplidata technology innovations:

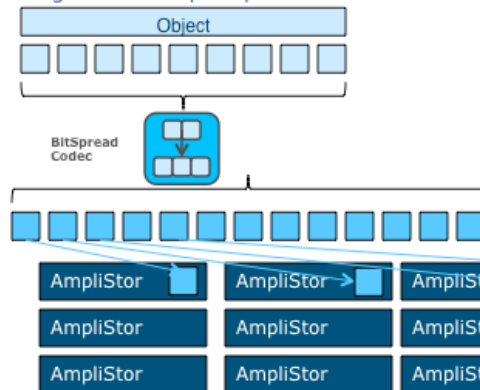
BitSpread - “Data encoding for high availability, reliability and lowest overhead”

The BitSpread technology performs the actual data distribution, and is hosted on the controller host. The system manages data objects that are first divided into multiple data blocks, which are then encoded into a larger number of check blocks. The BitSpread codec can reconstruct the original data object from any subset of check blocks as soon as a sufficient number of them can be retrieved. The order in which the check blocks are retrieved and the disks they are retrieved from is irrelevant to the algorithm. Conceptually, this is similar to solving a Sudoku puzzle: as soon as a sufficient number of fields are filled, the remaining fields can be recalculated without having to read all data. The way data is distributed across the available AmpliStor nodes and disks is policy controlled. Users specify availability requirements as simple policies on each AmpliStor Namespace, with the following parameters:

- The minimum # of disks that need to be included in the data spread
- The # of simultaneous failures in this spread that the system needs to be able to survive
- The geographical spread rules

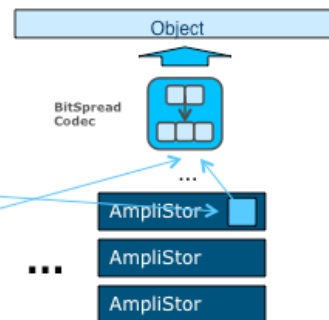
• **Storing Objects: BitSpread Encoding**

- Object is divided chunks
- Chunks are encoded with redundancy information as per reliability policy
- New check-blocks written to distributed storage nodes as per spread width



• **Fetching Objects: BitSpread Decoding**

- Check-blocks are fetched from distributed storage nodes
- Object is reconstructed as soon as sufficient number of check-blocks are returned
- Location and order of check-blocks does not matter!



11

As an example, the user may specify a “16/4” policy which instructs the system to spread across a minimum of 16 disks, but to tolerate the failure of any 4 disks within the Namespace. The data can then be reconstructed from any 12 disks, independent of the location or order in which it receives the blocks. In the case of a “16/4” policy, BitSpread only requires 60% overhead (1PB usable requires 1.6PB raw) in disk capacity to provide this level of data protection. This compares favorably against the alternate schemes, as demonstrated in the table below.

Technology	6+2 RAID6 + Replica	3 Copy Cloud Store	16/4 BitSpread
Source	1.00	1.00	1.00
Raw	2.67	3.00	1.6
Overhead	167%	200%	60%

Based on the selected policy parameters, the system selects on which disks data is stored. The data spread will be maximized across the available disks and their locations. This approach minimizes the impact of a component failure and ensures that data availability is never jeopardized. In the event of a failure, the system automatically reconfigures itself to store new incoming data according to an alternative spread of available disks, as defined in the policy.

BitDynamics - “Out-of-band storage verification and optimization, scalability and optimization”

The BitDynamics technology executes as an out-of-band agent on the AmpliStor storage nodes to provide automated storage management functions such as data integrity verification, self-monitoring, automatic data healing, scrubbing and garbage collection. BitDynamics keeps the storage system healthy and optimized without the need for manual human intervention.

The BitDynamics agents are responsible for automatically healing the system in the event of a component (disk or node) failure. In the event of a disk failure, multiple agents will coordinate and work in parallel to reconstruct and then write the failed data blocks to multiple disk drives. The rebuild process is completely automated, to restore the original level of resiliency without any need for human intervention. For example, a system with a 16/4 policy that experiences one disk failure, will still be protected against 3 simultaneous other disk failures. BitDynamics agents then reconstruct the original level of protection to protect against 4 simultaneous failures. By working in parallel across multiple storage nodes and disks, this self-healing process is expedited to the best extent possible. BitSpread provides multiple advantages in protecting data versus a RAID based approach, as shown in the table below.

	RAID 5/6	BitSpread
Maximum Disk Failure Protection	2 or 3	Any number
Data Protection against URE	No	Yes
Degraded mode operation?	Yes	Always protected
Flexible disk allocation	No	Yes, new spread for each object
Impact of rebuilds on performance	Rebuilds in the IO path	Redundancy regenerated out-of-band
Data protection during rebuilds	Writes to degraded disks	Writes only to fully protected disks
Encoding performance	N/A	Optimized
Dynamic reliability policies	No	Yes
Need to hot-swap drives?	Yes	No

Given this level of self-healing and multi-way protection, with a distributed storage architecture it is no longer necessary to urgently replace disks in the case of a failure: the entire system is permanently monitored and it reconfigures itself when needed. Also, data integrity is constantly checked and corrupted data is healed pro-actively. This enables the system to grow to very large levels, preserve data reliability and availability, without requiring additional administrative burden to maintain and manage the system.

AmpliStor Features & Benefits

Feature	Benefit	User Impact
Petabyte scalability	As new AmpliStor nodes are added to the system the BitSpread agent will automatically generate new spreads that include disks with available capacity. As this happens transparently the AmpliStor system can scale effectively to Petabytes without requiring virtual disk reconfiguration.	Seamless growth of capacity to meet business needs without any concerns for hitting limits.
Add storage nodes on-the-fly	New storage nodes can be seamlessly added at any time, without system disruption. BitDynamics agents inform the BitSpread encoders to use the added capacity to store data, without requiring reconfiguration of existing data.	Virtually no management effort to scale the system to Petabytes and beyond.
User-driven availability polices with any level of reliability	Users may specify availability policies directly in the system. An availability policy instructs the system how many disks to spread the data across, and how many simultaneous failures to tolerate. A system may have multiple Namespaces, each with its own unique availability policy.	Directly address business requirements for data reliability and availability.
Self-monitoring and healing	BitDynamics agents monitor and performance maintenance tasks on every storage node. In case of a disk or node failure, BitDynamics agents across a series of storage nodes will generate additional data blocks to substitute the lost data. This processing is happening out-of-band and does not cause a performance impact on system. As multiple BitDynamics agents are working in parallel to heal the	Lowers the time that system administrators spend to address component failures. Administrators can now spend more time on value-added tasks.

	data, the heal time is a multitude shorter than a comparable RAID based system.	
Continuous integrity checking	BitDynamics agents perform frequent background integrity checks on the stored data. This includes data scrubbing, data integrity verification and assurance tasks. Data that may have been corrupted due to write errors, bit rot or tampering, will be detected and proactively corrected by the BitDynamics agent.	The system automatically ensures data integrity, monitoring and performs ongoing maintenance tasks.
Out-of-band Optimizations	BitDynamics agents on the storage nodes perform maintenance tasks in an out-of-band manner, to not impact the IO path for maintenance operations.	The system minimizes the impact of system maintenance and data integrity assurance tasks.
Command Line Interface	The system provides a scriptable command line interface for administrative tasks, management tasks and system monitoring.	Easy to use interface that enables scripting of commands.
Graphical User Interface	The system provides a web-based user interface for administrative tasks, management tasks and system monitoring.	Intuitive, graphical interface for management-at-a-glance.

Sources

1- IDC: By 2020, the digital universe will contain 35 trillion gigabytes of media. 15% of that will be information related to cloud services, requiring cloud-optimized storage. (IDC)

Also according to IDC: although the amount of information in the digital universe will grow by a factor 44, the number of IT professionals will only grow by a factor 1.4. This means that data center administrators will have to manage 20 times more storage.

2 - 10 Petabytes of data, stored on 1.5TB disk drives, 3% annual disk failure rates. RAID5 groups stored as 6+1 disk drives (data/parity) have a 252% chance of data loss per year (2-3 events per year on average). RAID6 groups stored as 6+2 disk drives, have an 86% chance of data loss per year.