

Revolutionary Methods to Handle Data Durability Challenges for Big Data

Intel and Amplidata address the storage challenges presented by the coming tsunami of unstructured data that will stress data center scalability

“Big Data will earn its place as the next ‘must have’ competency in 2012 as the volume of digital content grows to 2.7 zettabytes, up 48 percent from 2011. Over 90 percent of this information will be unstructured...full of rich information, but challenging to understand and analyze.”

– IDC

Dr. Sam Siewert, PhD
Founder and Senior Partner,
Trellis Logic, LLC

EXECUTIVE SUMMARY

The growth in unstructured data is pushing the limits of data center scalability at the same time that disk drive vendors are pushing the limits of data density at tolerable device level bit error rates (BER).¹ For organizations delivering Cloud-hosted services involving images, videos, MP3 files, social media and other applications, data reliability will be a primary concern. The traditional RAID (Redundant Array of Inexpensive Disks) approach in wide use today simply will not provide the levels of data durability and performance required by enterprises dealing with the escalating volume of data. New approaches that go beyond traditional RAID promise to improve rebuild times on high-density disk drives, and reduce susceptibility to disk-error induced corruption,² which otherwise would result in crisis if traditional RAID is simply scaled up using current algorithms.

In this paper, we will discuss why RAID doesn't scale for Big Data, why erasure code is a better option, and how various erasure code alternatives compare.

We will use the long-standing mean-time-to-data-loss (MTTDL) model to compute the risk of data loss over time and show how the Amplidata computationally intense BitSpread algorithm deployed on Intel® Xeon® processor-based platforms deliver high levels of storage durability with a significant reduction in raw disk capacity overhead. BitSpread is Amplidata's rateless erasure coding software which is delivered commercially in the AmpliStor Optimized Object Storage system, a Petabyte-scale storage system purpose built for storing massive amounts of big unstructured data.

Cloud Data Scaling

Total data produced per year surpassed one zettabyte in 2010 and continues to more than double every two years.³ To put this in perspective, about two zettabytes (ZB) of digital universe will be created (per IDC) in 2011 is two million terabytes, or over one million new 3.5" SATA disks with two terabyte (TB) capacities. At the same time, not only is total data stored growing rapidly, but so is storage density. For example, Seagate demonstrated HAMR (Heat Assisted Magnetic Recording) technology with a terabit per square inch hard disk drive that is expected to lead

to 60 TB 3.5" hard disk drives within a decade.⁴ High capacity drives will challenge traditional RAID with long rebuild times when a single 60 TB drive fails and has to be rebuilt from parity, or mirrored data restored. Huge drives mean RAID systems will have to operate longer in degraded mode with risk of double or triple faults, and data loss will increase.

The marketing research community generally agrees that Cloud-hosted data will be somewhere between 10 to 20 percent of all annually generated data, so when total data hits 10 ZB somewhere around 2016, then over one ZB will likely be

Table of Contents

Cloud Data Scaling 1
Scaling the Cloud with RAID..... 2
Why RAID MTTDL Models are Unrealistic 3
Are There Better Alternatives to Traditional RAID? 4
Amplidata BitSpread Algorithm ... 5
Comparing Amplidata BitSpread to Traditional RAID 6
A Reference Architecture Proposal 7
Summary..... 7

stored in the public Cloud. This does not include private and personal Cloud storage (where we'll likely find the other nine ZB). Projections from IDC are even more bullish, stating that "Finally, Big Data will earn its place as the next 'must have' competency in 2012 as the volume of digital content grows to 2.7 zettabytes, up 48 percent from 2011.⁵ Over 90 percent of this information will be unstructured (e.g., images, videos, MP3 files, and files based on social media and web-enabled workloads)—full of rich information, but challenging to understand and analyze." Needless to mention, it will be challenging to store all of that data reliably.

Users of public Cloud services will expect their data to be safe. But if the same hard disk drive technology is used in the Cloud as is used at home and in small businesses, the only way this will be true is if it is based on enterprise data systems design. When a component storing data fails in an enterprise system, we call that a **data erasure**. Cloud systems will be expected to not only recover that data, but to do so such that the data remains available with minimal to no service interruption.

In some cases a Cloud service provider might use mirrored data, even triple or N-way mirrored, but this is the most costly method to protect against data loss when drives fail.

Scaling the Cloud with RAID

An enterprise measures the durability of data in terms of how long a system is expected to operate with no data loss. Standard methods include mirroring of data or RAID level 1, parity RAID or RAID level 5, and more advanced Reed-Solomon (RS) erasure codes which allow for multiple storage component failures before data is lost or becomes inaccessible. Mirroring data is by far the simplest, but also the most costly because it requires full data duplication and twice the resources for RAID1, and three times for triple mirroring. RAID level 5 uses simple exclusive OR logic to reduce that duplication to N/N+1, or for example, with five 2 TB hard drives a RAID5 volume would provide 8 TB of usable capacity. This results in an 80 percent storage efficiency ratio (usable capacity/raw capacity). RAID level 6 extends this parity approach to include double parity, using several different competing encoding algorithms, with the most general being Galois field mathematics for RS encoding of P (XOR parity) and Q (Galois field parity).

In context of the standard MTTDL model (Equation 1), RAID breaks down at petascale and beyond (the likely scale for Cloud data centers of the future hosting zettabytes of total data) and lacks the durability needed for safe hosting of user data in the Cloud.

$$MTTDL \propto \frac{MTTF^{(X+1)}}{(\text{Number of Data Loss Combinations}) * (MTTR)^X}$$

Where, X = simultaneous erasure tolerance

Equation 1.

Simply put:

- MTTDL is directly proportional to Mean Time to Multiple Failures (MTTF) (raised to X+1 power). This means, the bigger the MTTF, the bigger is the MTTDL—leading to better data durability.
- MTTDL is inversely proportional to the Number of Data Loss Combinations. This means that the lower the number of combinations we have, the bigger the MTTDL leading to better MTTDL durability.
- MTTDL is inversely proportional to the Mean Time To Recovery (MTTR) (raised to the Xth power). This means, the smaller the MTTR, the bigger is the MTTDL—leading to better data durability.

In summary, any increase in drive durability has a huge payoff because MTTF is increased, and for the more fault tolerant RAID levels this has a power law, so the durability increase is significant. Likewise, any reduction in MTTR improves durability and furthermore, for the more fault tolerant levels, this MTTR term is a power law, and therefore significantly improves durability for RAID6 and beyond compared

to single fault tolerant methods. The ideal system maximizes MTTF and minimizes MTTR, but also, perhaps more important, provides fault tolerance to a higher number of drive failures in a protection set. Or, restated, it vastly increases the protection by minimizing the potential for concurrent failures and by shrinking down the exposure window.

Figure 1 summarizes the annual probabilities of data loss for RAID5, RAID6, mirroring (RAID1), three-way mirroring (RAID1 3-way), and triple-parity erasure coding. In this figure, any scheme that results in data loss probability of one or higher affords weaker data durability, and any scheme that results in lower than one probability affords stronger data durability. As is clear from the figure, we expect to lose data unless we resort to triple mirroring or compute intensive and complex triple-erasure coding. The example in Figure 1 shows what might be a common full-rack of storage with several hundred disk drives (at 3 to 4 TB each) and the probability of data loss as a function of useful capacity (not including parity or mirrored data, just unique data).

In context of the standard MTTDL model, RAID breaks down at petascale and beyond.

Knowing that real systems can suffer greater loss potential than this standard MTTDL model predicts and because even RAID6 approaches a data loss probability of one at just 350 TB, it is clear that petascale systems demand better protection. With RAID, the only viable work-arounds for better protection than RAID6 are either 3-way mirroring or RAID6+1 (mirrored RAID6).

Why RAID MTTDL Models Are Unrealistic

The RAID MTTDL formulations are overly optimistic because they do not consider failure modes such as infant mortality, end-of-life, stress modes—like overheating in a data center, or anything beyond what is expected from the manufacturers highly accelerated (perhaps simulated) lifetime testing of a population of the devices. These models do not consider any sort of partial failure of a device (it is a two state Markov model), so devices are considered to either be in a probable working state or less probable total failure state, and not some sort of partially working state.⁶ Nor do these formulations estimate the magnitude of data loss which is expected to be at least one more drives-worth of data per RAID set, but loss is considered to be an all or none type of calculation. Perhaps more convincing that the model is optimistic are the large statistical studies such as the one million disk drive study summarized in Figure 2 (see next page).

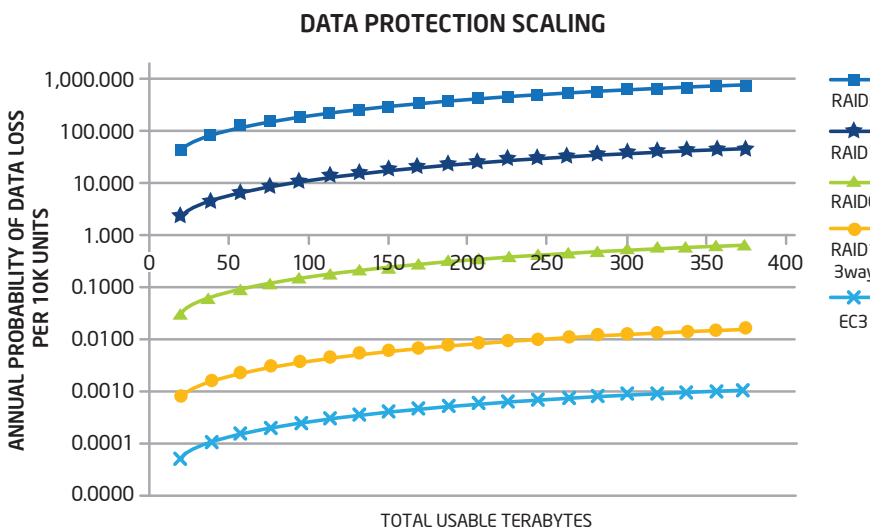


Figure 1. Standard MTTDL model comparison of a variety of RAID protection levels along with triple protection erasure coding (EC3).

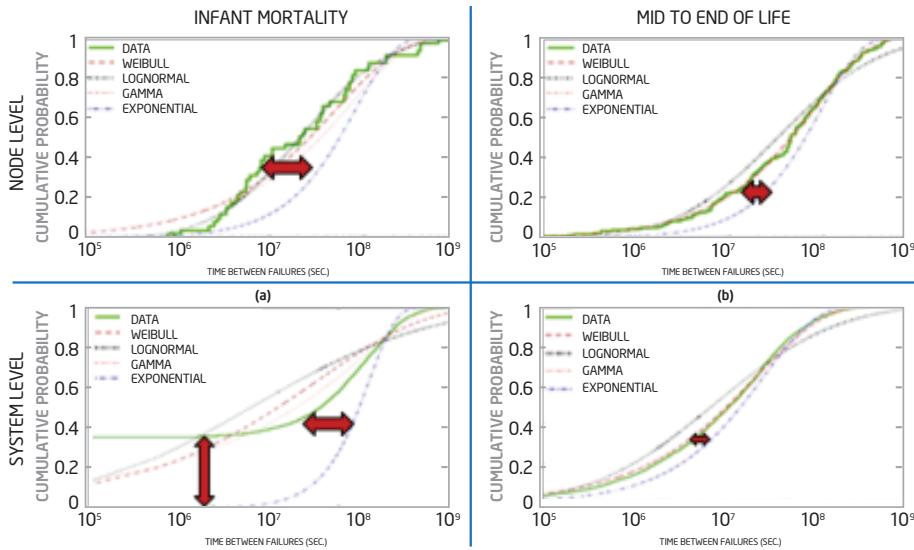


Figure 2. Standard model compared to statistics^{7,8}

The standard MTDL exponential model is the dashed blue line and the alternative distributions are super-imposed along with the green line, real data. For anyone building a Cloud data center, there are four alarming observations. Viewed one chart at a time we can see:

- **Node Level - Infant Mortality (Upper Left):** shows that the number of early failures for a single RAID controller can be up to an order of magnitude more frequent than predicted by the standard model.
- **Node Level - Mid to End of Life (Upper Right):** shows that node failures rates are higher than the standard model.
- **System Level - Infant Mortality (Lower Left):** shows an even wider rate of disk failures in a large set of drives as compared to node failures.
- **System Level - Mid to End-of-Life (Lower Right):** is where the standard model predicts best, but is still optimistic.

Experience by IT professionals running on-line deep archives is that partial failures and software bugs are a real phenomenon and that RAID level 6 is the minimum protection required but really requires mirrored RAID6 (RAID6+1). What necessitates this in deep archives? The need comes from petabytes of data and the need for this data on demand—common requirements with public Cloud data services. Triple mirroring is rarely used due to the high cost of 33 percent storage efficiency. Even RAID6+1 is likely to be only slightly better since each RAID6 volume has $N/N+2$ storage efficiency (e.g., 80 percent in $8+2$) and with mirroring becomes 40 percent, not much better than triple mirrors, which are simpler to recover when a disk member is lost. RAID level 6 is ok for on-demand but occasional access where the cost of RS encoding and recovery on the fly can be tolerated, but for faster access, this will either require hardware acceleration or a more efficient algorithm with the same or better durability than RAID6+1.

Are There Better Alternatives to Traditional RAID?

The lack of scalability of RAID to petabytes with safe levels of durability has led to re-visiting the fundamentals of data protection using a family of algorithms broadly known as erasure codes. To be accurate, RAID1, 5, and 6 are erasure codes, but they are a specific type known as Maximal Distance Separation (MDS) codes. Table 1 summarizes the RAID algorithms in use today. All of these algorithms meet the MDS criteria for encoding, whereby the data protection level matches the number of code or mirror segments in the RAID set.

As shown in Table 2, sticking with traditional hierarchical RAID algorithms (not meeting MDS criteria), which combine parity RAID with mirroring, are possible. But these approaches have low storage efficiency (resulting in higher acquisition and operation cost) and hence undesirable at Petabyte and higher scale storage systems.

Beyond simple XOR parity and mirroring, the mathematics to compute the code segments for MDS codes become high. So, the sheer complexity of these calculations can limit MDS solutions from achieving line rates when this data is accessed or updated. Some proprietary work-arounds to RAID6 specifically have been developed, such as RAID DP (Double Parity) and the EVEN/ODD algorithm, but general extension of MDS codes beyond double fault protection requires Reed Solomon erasure code.

To go beyond double fault protection with better efficiency than mirroring, triple-protection Reed-Solomon erasure code is an option, but innovators have proposed new erasure codes which are simpler, perhaps not meeting MDS criteria in a small set, yet having all of the protection advantages and are designed for scaling huge.

Algorithm	Data Protection	Code/Mirror Segments	Unique Data Segments	Storage Efficiency
RAID1 2-way (mirroring)	Single fault	1	1	50%
RAID5	Single fault	1	4	80%
RAID6	Double fault	2	4	66%
RAID1 3-way (triple mirroring)	Double fault	2	1	33%
Reed-Solomon Extensions of RAID	N faults (e.g., 3 faults)	k=N (3)	m (configurable) (4)	m/(k+m) (57%)

Table 1.

Hierarchical Algorithm	Data Protection	Code/Mirror Segments	Unique Data Segments	Storage Efficiency
RAID5+1	Double fault	6	4	40%
RAID6+1	Quadruple fault	8	4	33%

Table 2.

Amplidata BitSpread Algorithm

One very interesting example of an innovative algorithm like this is the Amplidata BitSpread algorithm, which is implemented in Amplidata’s AmpliStor Optimized Object Storage system. AmpliStor provides storage services over an http/REST interface that enables applications to store virtually unbounded numbers of objects, of any type and any size. BitSpread is hosted on AmpliStor controller nodes, which encodes objects using its rateless (referring to the ability to generate an unlimited number of equations from the input source) erasure-coding algorithm and distributes the encoded data across the AmpliStor storage nodes. Only a subset of the encoded data elements (actually equations, as demonstrated below), are required to retrieve the original data objects, thereby protecting the data against erasures (due to component failures) or bit error induced corruption.

The BitSpread rateless erasure code is distinct from other erasure code implementations based on RS style MDS codes or its variants. While RS codes have the ability to protect data against higher numbers of simultaneous failures than RAID5 or RAID6 and with storage overhead directly proportional to their protection or “safety” level, they pay a price in several areas:

- Encoding time is typically higher (non-linear with respect to the number of symbols being encoded). This impacts overall system performance.
- Lack of flexibility in reorganizing data for new spread policies. RS codes must recalculate for new spread and safety levels, forcing data to be unloaded and reloaded to make any changes in policy especially when adding or removing new data center sites.
- These algorithms often limit the amount of bit error protection that can be corrected to a few bit errors per object.

BitSpread is a non-MDS code, and hence it trades a small amount of extra space overhead compared to MDS codes in exchange for much higher-levels of throughput, faster self-healing, dynamic policy changes, and bit-perfect data integrity assurances as follows:

- The BitSpread algorithm achieves linear encode/decode times with respect to the number of symbols being encoded. This maximizes throughput, speeds rebuilds and broadens the set of use-cases. BitSpread furthermore builds upon the Intel Xeon SSE 4.2 instruction sets to achieve further performance gains and drive full 10 Gb Ethernet throughput rates on standard Intel Xeon multi-core servers.
- BitSpread enables flexible policies: e.g., easy-to-shift spreads from local to distributed multi-site policies “on-the-fly” and vice-versa.
- BitSpread protects every object with super-granular bit error protection—over 1000 bit errors can occur per object while preserving full object protection, availability and integrity.

The easiest way to understand the Amplidata BitSpread algorithm is to consider a simple case where the encoding is a set of equations as shown in Figure 3 below. In this figure, the data is encoded with three simple equations and data chunks X and Y derived from a stored object, which become two unknowns in two equations if there is an erasure. This simple approach is not strictly proportional to protection level in terms of bit overhead compared to RS encoding. But it has huge pay-off from the simplicity, which allows the encoding and recovery to run with minimal complexity and achieving better throughput compared to RS.

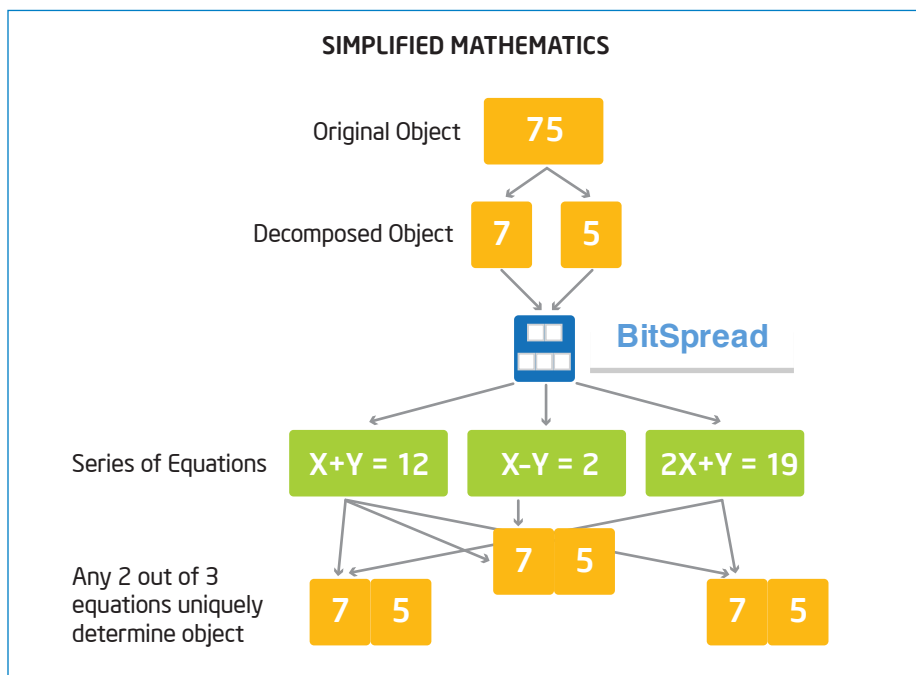


Figure 3⁹

Value	Description	Number nodes=32, 10 drives/node, Cap/Node=30 TB				
		BitSpread m=10, k=6, 1 nodes	RAID0+1 m=10, k=10, 2 nodes	RAID5+1 m=9, k=11, 2 nodes	RAID6+1 m=8, k=12, 2 nodes	RAID1 3-way m=10, k=20, 3 nodes
Efficiency	Raw/usable efficiency	63% ¹⁰	50%	45%	40%	33%
Durability	Relative data loss risk	10 ⁻⁸	2288	1.6	10 ⁻⁶	1

Table 3. Summary of Performance for RAID strategies compared to Amplidata BitSpread [Hardware configuration Large Object Reference Architecture].¹¹

BitSpread trades a small amount of extra space overhead compared to MDS codes in exchange for much higher-levels of throughput, faster self-healing, dynamic policy changes, and bit-perfect data integrity assurances.

Traditional erasure coding schemes implemented by competitive storage solutions have limited device-level BER protection (e.g., 4 four bit errors per device). But Amplidata’s solution can tolerate up to 1500-1600 BER per object stored without loss of data integrity.

Comparing Amplidata BitSpread to Traditional RAID

Using the standard MTTDL model as an optimistic proxy for reality, which we believe correctly allows for comparison of data durability based on MTTF and MTTR along with the protection set design, we can make a first level comparison of the Amplidata BitSpread protection sets to traditional RAID, including 3-way mirrors and RAID6+1 mirrored parity sets. The following table summarizes this numerical comparison of methods derived from the two-state Markov model for risk of data loss, which if anything, understates the need for new methods of data protection beyond simple RAID.

These numbers indicate that the storage efficiency is almost twice as efficient with Amplidata’s BitSpread than RAID6+1 and RAID1 3-way, providing Cloud applications with high data durability storage configurations that offer twice the usable capacity. We see that there is less than one chance in 100 million of data loss for 10 thousand units shipped using the BitSpread approach. By comparison, there is high likelihood that data will be lost with RAID1 3-way. While RAID6+1 is durable, it is still 100 times more likely that we will see data loss with this method compared to Amplidata BitSpread.

Initial Capex is driven by the cost of the controllers and the storage nodes and disks, so it is directly proportional to the reduction in storage overhead. Furthermore, Opex is predominately due to power use and heat generation of the drives along with IT monitoring costs. The reduction in storage overhead provided

Component	Quantity Per Rack	Configuration
Controller Node	3	<ul style="list-style-type: none"> ▪ Dual Intel® Xeon® Processor E5-2600 Family ▪ 64 GB DDR3 ECC memory ▪ 2 Intel® Ethernet Controller X540 (Twinville) Dual 10Gbase-T (4 ports) ▪ 8 Port SAS on board ▪ 2 x 10K rpm 900GB SAS HDD ▪ 2 Intel® SSD 710 Series RAID 0 Cache ▪ BMC (DCMI Support) ▪ 2 - PCIe x8 Gen3 Expansion
Storage Node	Up to 39 (in 44U rack)	<ul style="list-style-type: none"> ▪ Intel® Xeon® Processor E3-1200 Family (4 Core) ▪ 8 GB DDR3 ECC memory ▪ Intel® GbE Dual Ethernet Controller ▪ 8 Port SAS/SATA IOC ▪ 12 High Capacity (3 TB) Enterprise SATA Drives
Network Switch	2	<ul style="list-style-type: none"> ▪ 2 x 48 port Ethernet switches ▪ 44 x 1 GbE ports and 4 x 10 GbE SFP+ ports ▪ 4 x 10 GbE SFP+ ports and 44 x 1 GbE ports

Table 4. Reference architecture and component configuration.

by BitSpread means fewer spindles are needed for data protection, and therefore less power consumed per gigabyte stored, along with reductions in management costs due to self-healing.

A Reference Architecture Proposal

Intel and Amplidata have proposed a reference architecture (Table 4) to stand-up a robust storage system for enterprises, Internet portals and service providers.

Deploying Amplidata BitSpread on Intel Xeon processor based platforms affords high performance scalability, optimal power consumption and very high levels of durability for large-scale data storage.

For controller nodes, Amplidata BitSpread uniquely builds upon the Intel Xeon instruction sets (SSE 4.2) to provide scalable high-throughput capabilities. With low-power Intel processors (such as the Xeon Processor E3 1220L) used in building the storage nodes, Amplidata enables multi-petabyte scale storage with both high-density and high-performance but

also the lowest power consumption of any large scale storage system, requiring under 3 Watts per TB idle.

BitSpread is highly optimized for Intel Xeon multi-core processors, multi-threading capabilities and high memory-bandwidth through advanced software algorithms that take advantage of the processing power. The configuration also provides scalable, full line rate 10 GB Ethernet network throughput per controller.

Summary

The standard MTTDL model for data durability is an imperfect and optimistic predictor, but good enough for comparing the efficiency, cost/scaling, and protection afforded between erasure codes including traditional RAID and more advanced erasure codes like Amplidata BitSpread.

The model indicates the same negative trend for traditional RAID strategies even with this optimistic model (it only gets worse with more accurate models). We have shown here using the optimistic

standard model that Cloud data centers require data protection better than RAID6 and with better efficiency than RAID1 3-way or RAID6+1.

Amplidata’s BitSpread is an efficient, scalable, and practical alternative to the stop-gap of combined RAID levels like 6+1 and N-way mirroring. Furthermore, Amplidata’s BitSpread is more practical to implement and affords better operational efficiencies than Reed-Solomon erasure codes based storage systems.

Deploying Amplidata’s BitSpread technology based storage systems using Intel platforms can result in significant acquisition and operational savings for the data center implementers while achieving very high levels of data durability.

Revolutionary Methods to Handle Data Durability Challenges for Big Data

For more information on Amplidata,
AmpliStor, and BitSpread please visit
www.amplidata.com

¹ "An Analysis of Latent Sector Errors in Disk Drives," Lakshmi Bairavasundaram, Garth R. Goodson, et al.

² "An Analysis of Data Corruption in the Storage Stack," Lakshmi Bairavasundaram, Garth R. Goodson, et al.

³ "World's data will grow by 50x in next decade, IDC study predicts," by Lucas Mearian, Computerworld, June 2011.7

⁴ "Seagate hits 1 terabit per square inch, 60TB hard drives on their way," by Sebastian Anthony, March 2012.

⁵ The research firm IDC predicts that the volume of digital content will grow to 2.7 zettabytes in 2012, from <http://www.idc.com/getdoc.jsp?containerId=prUS23177411>

⁶ "Mean time to meaningless: MTTF, Markov models, and storage system reliability," Kevin Greenan, James Plank, Jay Wylie, Hot Topics in Storage and File Systems, June 2010.

⁷ Schroeder, Bianca and Gibson, Garth, "A large-scale study of failures in high-performance computing systems (CMU-PDL-05-112)" (2005). Parallel Data Library. Paper 46. <http://repository.cmuedu/pdl/46>

⁸ "Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you?," Bianca Schroeder, Garth A. Gibson.

⁹ "Unbreakable Object Storage for Exabyte-scale Unstructured Data – Amplidata Technology Paper," V5.3, Amplidata Inc., 2011.

¹⁰ BitSpread overhead has been further reduced in the current release.

¹¹ Next Generation Scalable and Efficient Data Protection," Sam Siewert, Greg Scott, Intel Developer's Forum, San Francisco, June, 2011.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. UNLESS OTHERWISE AGREED IN WRITING BY INTEL, THE INTEL PRODUCTS ARE NOT DESIGNED NOR INTENDED FOR ANY APPLICATION IN WHICH THE FAILURE OF THE INTEL PRODUCT COULD CREATE A SITUATION WHERE PERSONAL INJURY OR DEATH MAY OCCUR.

Intel may make changes to specifications and product descriptions at any time, without notice. Designers must not rely on the absence or characteristics of any features or instructions marked "reserved" or "undefined." Intel reserves these for future definition and shall have no responsibility whatsoever for conflicts or incompatibilities arising from future changes to them. The information here is subject to change without notice. Do not finalize a design with this information.


The products described in this document may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. Contact your local Intel sales office or your distributor to obtain the latest specifications and before placing your product order. Copies of documents which have an order number and are referenced in this document, or other Intel literature, may be obtained by calling 1-800-548-4725, or by visiting Intel's Web site at www.intel.com.

Copyright © 2012 Intel Corporation. All rights reserved. Intel, the Intel logo, and Xeon are trademarks of Intel Corporation in the U.S. and other countries.

*Other names and brands may be claimed as the property of others.

Printed in USA

0912/CSP/HBD/PDF

 Please Recycle

327961-001US

