# DISTRIBUTED DATA PROTECTION

## — *MatrixDDP*

WHITE PAPER | **JUNE 2017**

# SUMMARY

**This whitepaper will show that by designing the data distribution and data protection scheme from scratch leveraging modern SSD technology, WekaIO has created a new category of data protection with integrated intelligence that offers greater protection than RAID, Erasure Coding, or replication without the performance or overhead impact associated with these legacy technologies.**

Data protection is the key to building a scalable storage system.  Without the appropriate protection schema file systems will have to be limited in size to accommodate the rebuild time window and data exposure risk.  Popular data protection schemes that deliver on scalability hurt performance while the performant schemes hurt scalability.

RAID was the defacto data protection standard when the average disk size was 20 megabytes of storage, but with disk drive capacities now at 12 terabytes (TB) RAID has reached its limit as a reliable data protection schema. Erasure Coding (EC) improved upon RAID, but falls short in production environments where EC consumes a significant number of CPU cycles and suffers from long latencies, making the technology unsuitable for high performance applications that are I/O and latency sensitive. Triple Replication (TR) is suitable for low latency, small I/O workloads, but hurts large file throughput because every piece of data needs to be written across the network three times. TR also carries an economic penalty because it requires three times as much purchased capacity for a given amount of usable storage, driving up operational costs (power, cooling, and floor space). This model is untenable at scale when SSD technology is the primary storage medium.

The design philosophy of WekaIO was to create a software solution that has the performance of an all flash array with the scalability and economics of the cloud. This required devising a new coding schema that would solve the problems inherent in RAID, EC, and replication. RAID cannot scale, EC has poor small file IO performance, and replication is both expensive at scale and has poor throughput. The resulting solution from WekaIO is a patented distributed data protection scheme that is faster than replication, has minimal capacity overhead, and can scale to billions of files.

# CONTENTS

1 **Introduction**

1 **The Limitations of RAID, Erasure Coding, and Replication**

1 *RAID Cannot Scale*

1 *Erasure Coding Requires Heavy CPU Cycles and is Low Performing*

2 *Replication Drives Up TCO*

2 **WekaIO Matrix Distributed Data Protection (MatrixDDP)**

2 *Industry First Server-level N+4 Data Protection*

3 *Fast Efficient Data Protection*

3 *Lowest Write Amplification*

4 *Dramatically Reduced Rebuild Times*

4 *End-to-End Data Protection (EEDP)*

5 *Data Protection on Power Failure*

5 *Data Re-balancing*

6 **Conclusion**

# Introduction

Data is the currency of the modern world, and protecting your precious assets is the most important function of any storage system. WekaIO Matrix is not only the fastest distributed file system in production, but it also delivers the highest levels of data protection, durability, and availability. This whitepaper will show that by designing the data distribution and data protection scheme from scratch leveraging modern SSD technology, WekaIO has created a new category of data protection with integrated intelligence that offers greater protection than RAID, Erasure Coding, or replication without the performance or overhead impact associated with these techniques. Matrix Distributed Data Protection (MatrixDDP) is a patented coding scheme that delivers unprecedented levels of data protection at scale.

**Key features include:**

- N+2 or N+4 data protection that can be configured dynamically
- Fastest data encoding scheme for highest I/O performance
- Cluster-wide data rebuilds for fastest return to a more protected level
- Prioritized, intelligent rebuilds to return to a more protected level
- High space efficiency for best utilization of expensive SSD capacity
- Lowest levels of write amplification to ensure longer SSD lifespan
- End-to-end data protection to ensure data integrity
- Smart re-balancing of data for optimal utilization of all resources and linear performance scalability

The end result is a cloud scale storage solution that delivers the highest resiliency, fastest performance and lowest cost of ownership.

# The Limitations of RAID, Erasure Coding, and Replication

## *RAID Cannot Scale*

RAID technology protects data at the disk group level and requires only 20% to 30% more capacity to fully protect a data set. It was developed when the average disk size was 20 megabytes of storage, but with disk drive capacities now at 12 terabytes (TB) RAID has reached its limit as a reliable data protection schema. It is not suitable for server node level data protection and does not allow storage to scale. The higher the drive capacity, the greater the probability of a disk failure. Higher capacity also means longer rebuild times and a greater probability of a second drive failure during a rebuild, resulting in catastrophic data loss. A 12 TB drive will take days to rebuild at full bandwidth at a cost of severely degraded storage system performance. In addition, RAID protects data at the block level, so when a drive fails every block on a protected drive must be rebuilt, even if there is no actual data on the blocks. The result to end users is long periods of running in degraded mode which adversely affects application performance and heightens the chance of data loss from another failure during the extended rebuild window.

## *Erasure Coding Requires Heavy CPU Cycles and is Low Performing*

Erasure coding (EC) has been widely implemented in large scale environments such as scale-out NAS and cloud object storage to overcome the challenges of RAID. Data is broken into chunks that are distributed across nodes or even physical locations. EC can protect against considerably more failures than RAID. However, EC implementations that are in production consume a significant amount of CPU cycles and suffer from long latencies, making the technology unsuitable for high performance applications that are I/O and latency sensitive.

### *Replication Drives Up TCO*

Replication is precisely what the name suggests; data is broken into chunks and replicated across multiple locations. The minimum protection level is known as mirroring, which ensures a second copy of data exists. Most enterprise systems that employ replication use triple replication to ensure they can survive up to two failures. Replication is suitable for low latency, small I/O workloads, but triple replication hurts large file throughput because every piece of data needs to be written across the network three times. It also carries an economic penalty because it requires three times as much purchased capacity for a given amount of usable storage, driving up operational costs (power, cooling, and floor space). This model is untenable at scale when SSD technology is the primary storage medium.

## WekaIO Matrix Distributed Data Protection (MatrixDDP)

Data protection is an integral part of a scalable storage system design, however most storage systems that claim scale-out are leveraging technologies designed for dedicated appliances and storage arrays. Legacy data protection schemes such as RAID and EC force limitations on the file system or volume to overcome the challenge of a long rebuild that increases the risk of a catastrophic failure.

The design philosophy of WekaIO was to create a software solution that has the performance of an all flash array with the scalability and economics of the cloud. This required devising a new coding schema that would solve the problems inherent in RAID, EC, and replication because any one of these schemes would violate this design philosophy. RAID cannot scale, EC has poor small file IO performance, and replication is both expensive at scale and has poor throughput. The resulting solution from WekaIO is a patented distributed data protection scheme that is faster than replication, has minimal capacity overhead, and can scale to billions of files.

### *Industry First Server-level N+4 Data Protection*

WekaIO Matrix is designed to run in a cloud environment, and data protection is defined at the node level, which can contain a single or multiple SSDs. Data protection levels are flexible depending on the size and scale of the server cluster—the larger the cluster, the larger the recommended data stripe size for best utilization of SSD capacity. The data protection level can be set to two or four, meaning that the system can survive two or four simultaneous drive or node failures without impacting data availability.

Our N+4 data protection level is unique in cloud storage and delivers exponentially better resiliency than triple replication, without the expensive storage capacity and throughput implications. N+2 level protection is ideal for scaling out clusters with relatively low capacity per server (up to 10TB for example), and it is the most space efficient. We recommend using N+4 protection for customers who want to scale up capacity inside each node in the cluster or for clusters where the compute environment is not considered stable. In such environments, our patented prioritized rebuild scheme will return the cluster to the next best level of protection (from N-4 to N-3) in minutes.

The amount of network bandwidth provided to Matrix during the rebuild process will directly impact the rebuild time. The rebuild window can be shortened by apportioning more network bandwidth to the rebuild process, this can be modified at any time based on customer needs.
The following table (Table 1) demonstrates that more network bandwidth will accelerate rebuild times.

Table 1:  100 Node Cluster, 10TB of storage per node, on a 10GbiE network connection

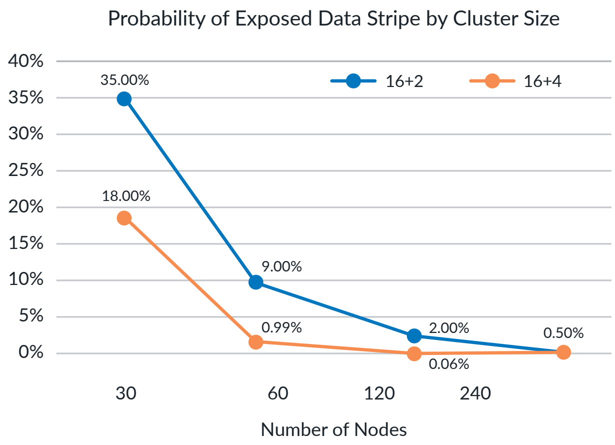
| 100 Node, 10TB/Node, 10 GbiE Network | 25% of 10GbiE | 50% of 10GbiE |
|---|---|---|
| Time to return to full redundancy after a two node failure (h:m:s) | 2:52:44 | 1:26:22 |

## Fast Efficient Data Protection

WekaIO has invented a data protection scheme that delivers the scalability and durability of EC, but without the performance penalty. In addition, our data protection provides significantly faster rebuild times. Data is distributed across logical entities that span several hosts and several SSDs. Matrix never stripes data inside the same server (or failure domain), so in the event of a single server outage there can never be the possibility of a double degrade, even if the node has multiple SSDs. When an entire server goes down, the system considers it a single failure.

As clusters grow in node count, the probability of a hardware or drive failure goes up proportionally. Matrix overcomes this issue by distributing stripes across nodes and also by utilizing a highly randomized data placement methodology. The resulting cluster actually becomes more resilient than a smaller cluster because the probability of two nodes having the same date set goes down dramatically.
Example: for a stripe size of 18 (16+2) and a cluster size of 50, the number of possible stripe combinations is 1.8 to the power of 13 based on the following formula: C!/(S!*(C-S)!). If the cluster size is increased to 100, then the number of possible combinations is 3 to the power of 19.

The following graph illustrates how the probability of data exposure decreases as the cluster size increases due to the distributed data placement.



## Highest Performance Writes

MatrixDDP overcomes the significant write penalty that occurs with erasure coding. Write Amplification is the ratio between the number of writes that a user does vs. the number of writes that the filesystem does to acknowledge a user write. Write amplification affects system performance and reduces the useful life of the underlying SSD devices. As Table 2 indicates, MatrixDDP has the lowest write amplification impact of any data protection scheme in production. Our N+2 protection level has a 1.5x write amplification factor compared to 3x for a comparable RAID or EC protection level.

Table 2: Weka IO write amplification compared to legacy data protection schemes

| Scheme | Space Efficiency | Resiliency Level | Write Overhead | Write Amplification |
|---|---|---|---|---|
| 1+1 Mirroring | 50% | 1 | 2W | 200% |
| Triple Mirroring | 33% | 2 | 3W | 300% |
| RAID-5 8+1 | 89% | 1 | 2R+2W | 200% |
| RAID-6 (10+2) | 83% | 2 | 3R+3W | 300% |
| WekaIO 16+2 | 89% | 2 | 3R+1.5W | 150% |
| WekaIO 16+4 | 80% | 4 | 3R+2W | 200% |

Note: WekaIO calculation with tiered object store. We reserve an additional 20% of capacity if there is no object

## *Dramatically Reduced Rebuild Times*

Protecting against data loss is the single most important function that a storage system must perform, but storage is physical, so it is not a case of if a failure will occur but when. MatrixDDP was designed to ensure that the system is restored to a more protected level as fast as possible. The solution deploys several innovative strategies to achieve this.

Firstly, all nodes in the system participate in the rebuild process. This means that the larger the cluster size, the faster the rebuild occurs because more processors are available for the data reconstruction. By contrast, in a replicated system only the mirrored servers participate in the rebuild process, resulting in significant impact to application performance.

In the event of multiple failures, the system prioritizes data rebuilds starting with data that is least protected. MatrixDDP looks for data stripes that are common to the failed nodes and rebuilds these data stripes first so the system can be returned to the next level higher of resiliency as fast as possible. This prioritized rebuild process continues until the system is returned to full redundancy.

MatrixDDP rebuild rates are unprecedented in the industry, Table 3 shows that a 100 node cluster with 30TB capacity per node can be returned from 4 failures to 3 failure resiliency in under 1 minute, while a return to full redundancy would take 2 hours and 25 minutes, assuming the entire 120TB must be rebuilt.

Table 3: Node rebuild times after 16+2 and 16+4 failure

| Stripe Size(data+parity) | 16+2 | 16+2 | 16+4 |
|---|---|---|---|
| Number of Nodes | 50 | 100 | 100 |
| Capacity per Node | 10TB | 10TB | 30TB |
| Rebuild Bandwidth (GB/Second) | 0.5 | 0.5 | 1.0 |
| Time to Full Protection (h:m:s) | 2:54:34 | 1:26:22 | 2:25:50 |
| Time to 1st failure resiliency (h:m:s) | 1:00:37 | 0:14:52 | 0:28:00 |
| Time to 2nd failure resiliency (h:m:s) | | | 0:05:11 |
| Time to 3rd failure resiliency (h:m:s) | | | 0:00:55 |

Finally, because WekaIO Matrix protects data at the file level and not at a block level like RAID systems and all-flash arrays, the software only needs to rebuild data that is actually stored on the failed server or SSD. This means that the rebuild times shown in the previous example would be significantly lower in practice.

Data stripes on the Matrix file system are not exclusive to a single file, therefore when a file has been tiered off to an object store, there is no need to rebuild the tiered data within an exposed stripe. For example, if only 50% of the system capacity is stored on the hot tier, with the remainder on object store, then the actual rebuild time will be half of what it would take for a full stripe. This is because only half of the exposed stripes contain exposed data, as a copy of the data is already protected on the object store. This is true even in the case of tiered data that continues to be used by the hot tier as a local cache.

## *End-to-End Data Protection (EEDP)*

Protecting written data is critical, but it is also important to write the correct data to begin with. A bit flip on the network can inject silent corruption to data being written, or a drive fault can cause bad data to be read. WekaIO Matrix provides end-to-end data protection to ensure corrupt data can never be committed to SSD or returned to the application on read.

All written data gets a checksum on write that is calculated by the network adapter so it does not consume CPU cycles and cause performance degradation. When the data block arrives at the file system the checksum is saved in the file system data structures. When data is written to the SSD the software verifies the checksum again. If it does not match the original checksum, the file system will not commit the block write. The process is repeated on read, ensuring that you will never have data corruption.

To guarantee data integrity, the EEDP checksum must be stored independent of the data block.  If a block of data self-verifies, then the system cannot detect a scenario where the SSD did not actually commit the write. If a storage system stores the EEDP with the data block, and a previous write was successfully executed, but the following write was not committed, then the next read will verify the old data based on the EEDP checksum, returning incorrect data.

## Data Protection on Power Failure

Power loss during a write commit from the application is problematic and can lead to data loss, so most storage systems rely on non-volatile memory to protect against a power failure during a write commit. This solution works well for dedicated appliances where the hardware architecture is tightly controlled; however in a cloud environment where servers are virtual machines, you cannot assume that systems will have non-volatile memory available for this task.

WekaIO has implemented a journaling system that can fully return the consistency of the file system without the reliance on non-volatile memory. WekaIO Matrix is unique in that the file system algorithm chooses data placement rather than the data protection algorithm. As a result, when a write is acknowledged at the OS level, it is safely protected from server failure or even a data center wide power failure. Furthermore, thanks to Matrix's innovative data layout and algorithms, recovering from a data-center wide power failure takes minutes because there is no need to do a complete file system consistency check (FSCK). For most other file systems, the FSCK process recovery time is proportional to the size of the recovered file system. In large scale deployments, this recovery can take days.

## Data Re-balancing

MatrixDDP plays an important role in managing and maintaining consistent overall performance levels for the entire server cluster. The system calculates the average capacity of cluster nodes (referred to as the fill level); if it detects a heavily filled node (relative to the overall cluster average), the file system redistributes the data so that the system returns to a more balanced state. The result is that the overall performance of the cluster is maintained (by adding more SSDs) or improved (by adding more nodes with SSDs) as the cluster is expanded.

# Conclusion

Historically, enterprises have relied on storage arrays to protect their valuable data from disk failures and silent data corruption, and vendors have relied on age-old protection schemes to provide this protection. However, with today's multi-terabyte disk capacities and multi-petabyte storage systems these protection schemes are no longer viable, or severely limit overall storage system performance. Enterprises are left with a false sense of security that their data will be accessible when they need it.

As the value and volume of data grows, enterprises should look beyond traditional data protection schemes to innovative solutions that are designed to protect data at massive scale while delivering on performance and economics. WekaIO's patented distributed data protection scheme (MatrixDDP) offers N+4 protection that is faster and more cost efficient than legacy protection schemes, has integrated load balancing and wear leveling technology to extend the life of your storage resources, and scales to billions of files to protect your data as it grows. With a massively parallel rebuild process, tens of terabytes of data can be rebuilt in hours instead of days.

To find out more, please contact us at Info@Weka.IO

WEKA.IO
2001 Gateway Place, Suite 400W, San Jose, CA 95110 USA T 408.335.0085 E info@weka.io www.weka.io