A KAMINARIO WHITE PAPER

K2 All-Flash Array Architecture White Paper

January 2015

kaminario.

Table of Contents

Executive Summary

Flash and All-Flash Arrays have become synonymous with performance. However, they are also regarded as expensive and lacking the full set of enterprise features, and are to be used as a solution for a specific pain point or a single application.

Flash prices have gone down considerably over the last few years and Flash as a storage media has gained maturity, yet the adoption and deployment gap between HDD-based storage and Flash-based storage is still prevalent in spite of the obvious shortcomings of HDD storage arrays.

In its fifth generation, the Kaminario K2 manages to bridge that gap and even more so, to supersede HDD legacy storage as the right choice for all primary storage purposes.

It comes down to enterprise resiliency of the storage, how the storage infrastructure can facilitate business agility and, most important, what is the total cost of ownership.

K2's Scalable Performance and Resilience Architecture (SPEAR™) is designed to harness the power of Flash with the right storage efficiency features such as global inline selective deduplication, inline compression, thin-provisioning, efficient and robust Kaminario RAID scheme and highly efficient metadata management. The outcome is the most cost-effective All-Flash Array (AFA), with better cost than HDD storage. However, there is no compromise on enterprise resiliency, which is gained via SPEAR's native snapshot and replication features, high availability (HA) and non-disruptive upgrades (NDU).

SPEAR's unique scalability features of scale-out and scale-up drives business agility to the maximum, enabling independent linear growth of capacity and performance according to datacenter needs. Combined with a global adaptive block size algorithm, single management pane and full VMware integration, K2 is able to sustain the performance of multiple environments and mixed workloads while keeping it simple and easy to manage.

This white paper describes the SPEAR architecture in depth, detailing its core features and functionalities and how they come to play in the Kaminario K2 All-Flash Array.

Introduction

Flash as a storage media is a game changer. It finally allows storage to match the evolution, progress and performance of CPU power and networking. However, to use Flash in a storage array requires inventing a whole new storage architecture, since the legacy architectures are tightly designed to match the characteristics of spinning disks.

SPEAR is designed from the ground up to utilize Flash with the insight of preparing for the next generation of Flash, CPU and networking. As an example, the K2 uses SSDs in its array, and SPEAR's metadata design can accommodate any SSD capacity size.

Kaminario K2's SPEAR architecture is a facilitator for driving Flash as primary storage to the next level with the following tangible business benefits:

Cost efficiency – Flash storage is considered expensive, but with the right storage efficiency features such as global inline selective deduplication, inline compression, thin-provisioning and the Kaminario K-RAID™, the CAPEX of the K2 AFA is lower than legacy HDD storage. The K2 AFA is also much more economic on power, cooling and footprint so the OPEX is considerably lower. Native capacity-efficient snapshots can be utilized to create multiple production-like environments for even better ROI.

Enterprise resiliency – SPEAR's high availability combined with non-disruptive upgrades results in continuous availability with no planned down time. However, availability without consistent performance is not worth much. SPEAR also facilitates excellent performance during a failure, so the productivity of the environment is not impacted. Native snapshot and replication deliver data protection with quick restore capabilities. SPEAR's HealthShield™ component monitors the array and offers proactive and preventive serviceability.

Business agility – Being able to adapt to the customer's datacenter growth means the storage array needs to grow in capacity and/or performance. SPEAR has the unique ability to scale out and/or scale up, non-disruptively, to accommodate such growth. The K2 scales, but it is still a single pool of storage, with a single pane of management and with automatic load balancing. Last but not least, the K2 accelerates production with Flash performance that serves mixed workloads via a global variable block size algorithm. It delivers better user-experience in VDI environments, removes the IO-blender effect from virtual servers and allows the customer to receive real-time reports from analytics environments (OLAP) and faster database queries in OLTP environments.

System Overview

GENERAL

SPEAR™ is the secret sauce that binds best-of-breed enterprise hardware components to an All-Flash Array. SPEAR is software-defined, meaning it is agnostic to any hardware or technology. SPEAR can easily and economically adapt to any advancements made in CPU power, networking and Flash media. Scalability combined with non-disruptive upgrades guarantees continuity and progress of both hardware and software within the array.

K-BLOCK

The K-Block is the building block of the K2 All-Flash Array, and encapsulates the following hardware components:

- Two K-Nodes, each a 1U storage controller
- Each K-Node provides full redundancy for its components:
  - Two 8Gbps FC ports and two 10GbE iSCSI ports for external host connectivity
  - Two 40Gbps InfiniBand ports for K-Node backend interconnectivity
  - Two 6Gbps SAS ports for internal connectivity with the K-Block's SSD shelves
  - Two hot-swappable PSUs
  - Two hot-swappable Battery Back Units (BBU)
- Base SSD shelf, 2U in size
  - 24 hot-swappable SSDs; Enterprise grade SSD with MLC Flash
  - Two hot-swappable controllers
  - Two hot-swappable PSUs
- Expansion SSD shelf, 2U in size
  - Same as the base SSD shelf

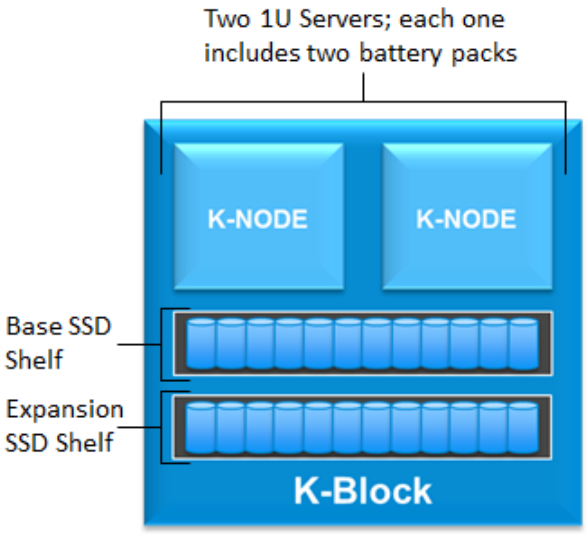


Figure 1: Basic K-Block Components

SCALABLE ARCHITECTURE

SPEAR is designed to facilitate linear growth in both capacity and performance[1] while maintaining consistently low latency, thus implementing a scale-out architecture. In addition, it has the ability to facilitate only growth in capacity with no impact on performance, thus implementing a scale-up architecture. The combination of both scale-out and scale-up architectures in a single storage array is the key feature for building a storage infrastructure that can scale in the most cost-effective way, ensuring that the exact datacenter requirements for new and existing applications are met. Any increase in capacity, results in an automated rebalancing of data within the array, with no intervention of human management.

The starting point for any K2 All-Flash Array configuration is a single K-Block, as shown in Figure 2, below:

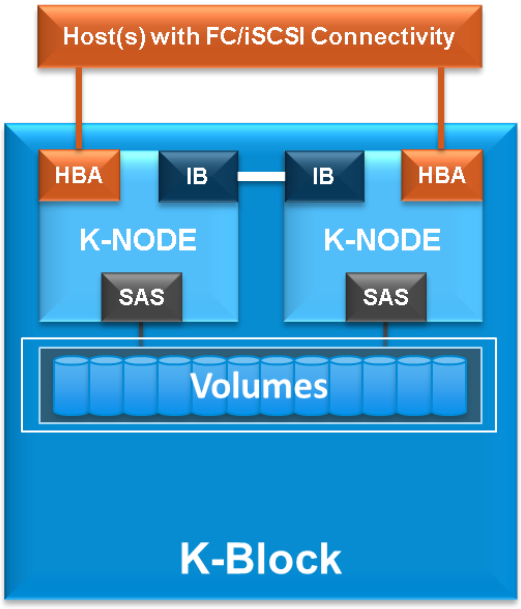


Figure 2: Single K-Block array

- The K-Nodes are connected redundantly using point-to-point IB connectivity
- Each K-Node is connected redundantly to the SSD shelf using SAS connectivity
- The K-Nodes have Active/Active connectivity either directly to the host(s) or through a switch, via FC or iSCSI
- Volumes and metadata are automatically distributed between all SSDs in the array and can be accessed from every K-Node in the array

From this basic building block, the customer has the flexibility to scale the K2 All-Flash Array according to the datacenter needs, using a scale-up and/or a scale-out approach.

[1] Qualifications for specific configurations are on the roadmap (targeted for Q1 2015)

SCALE-UP

Scaling up means adding more capacity by adding Expansion Shelves without adding K-Blocks.

Scaling up a single K-Block with an Expansion Shelf is shown in Figure 3, below:

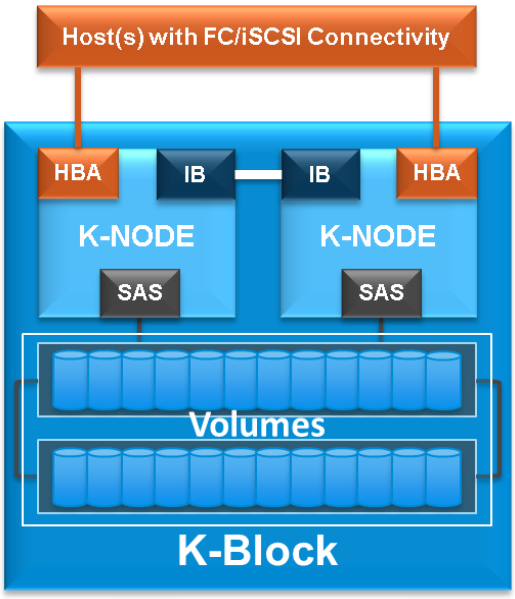


Figure 3: Scaling up a Single K-Block array with an Expansion Shelf

- The expansion increases the capacity density and reduces the cost per GB
- The Expansion Shelf is connected redundantly using SAS connectivity
- The expansion is done online with no downtime or decrease in performance
- The new configuration has the same performance as before
- Existing volumes are automatically redistributed between all the SSDs in the array
- No need to change any host connections or definitions

SPEAR is optimized for cost and performance as well as for the right balance between performance and capacity. To achieve the best optimization of cost, capacity and performance, current configurations support each K-Block to scale-up with a single Expansion Shelf.

SCALE-OUT

Scaling out means to increase the number of K-Blocks, thus adding more capacity and compute power.

Scaling out from a single K-Block to a dual K-Block array is shown in Figure 4:

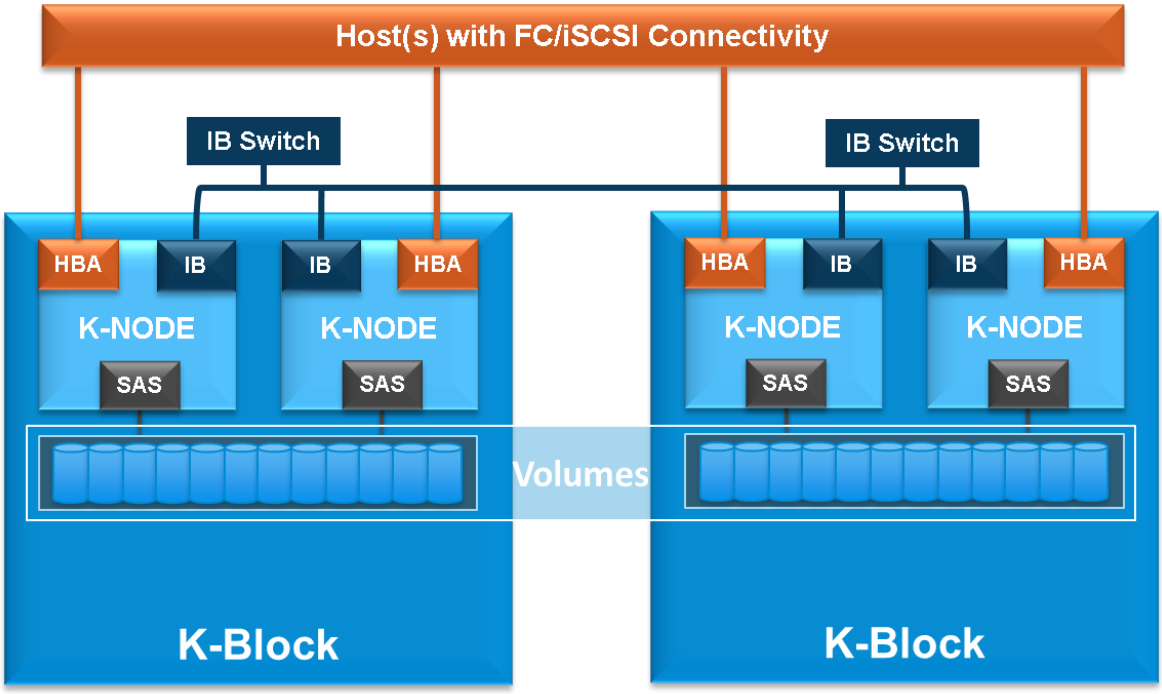


Figure 4: Scaling Out from a Single K-Block array to a Dual K-Block array

- The expansion linearly increases the capacity, IOPS and throughput. The latency is kept consistently low and is indifferent to the expansion.
- Two InfiniBand switches are required to support the interconnect between all the K-Nodes in the array. Scaling beyond two K-Blocks will not require any additional networking hardware.
- The expansion is performed online with no downtime or decrease in performance
- Existing volumes are automatically redistributed between all the SSDs in the array and can be accessed from every K-Node in the array.
- New hosts can be connected to the new K-Block, and the new and existing hosts can access all new and existing volumes.

SCALE-OUT AND SCALE-UP

There is no particular order in which the K2 scales. The K2 can first scale up, then scale out and vice versa. This flexibility removes any compromises of using a storage array that is not tailored to the customer's needs and datacenter requirements.

For example, scaling-up in a dual K-Block array with two Expansion Shelves is shown in Figure 5, below:

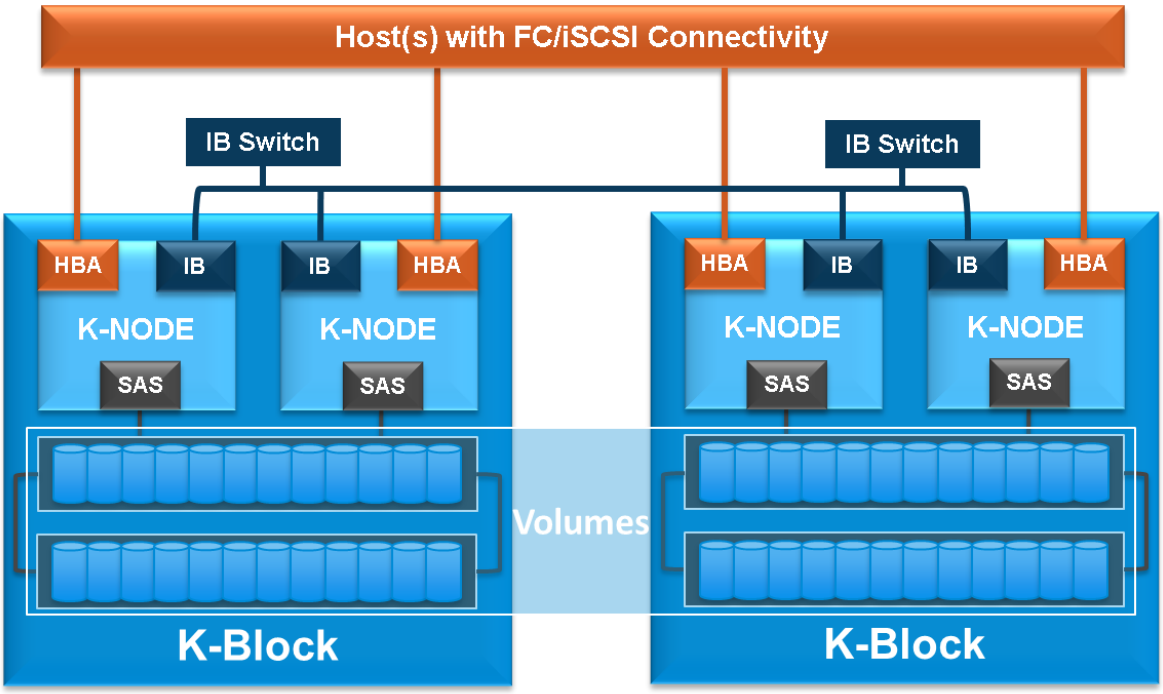


Figure 5: Scaling Up in a Dual K-Block array

Similar to the scale-up of a single K-Block, also in this case:

- The expansion increases the capacity density and reduces the $/GB factor
- The Expansion Shelves are connected redundantly using SAS connectivity
- The expansion is performed online with no downtime or decrease in performance
- Existing volumes are automatically redistributed between all the SSDs in the array
- No need to change any host connections or definitions

The system can continue to scale out with K-Blocks and their Expansion Shelves. Continuing the example above, scaling from a dual K-Block array with Expansion Shelves to a quad K-Block array, thus adding two more K-Blocks and two more Expansion Shelves, is shown in Figure 6, below:

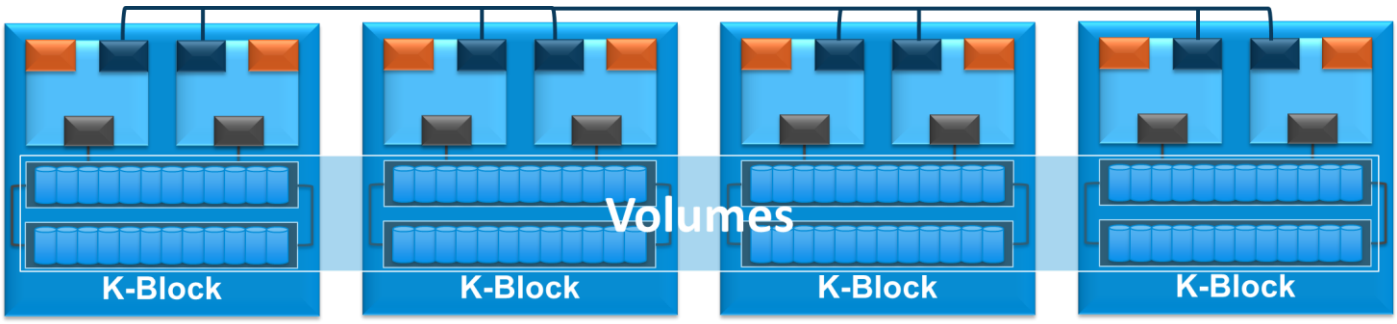


Figure 6: Scaling from a Dual K-Block array with Expansion Shelves to a Quad K-Block array

SUMMARY TABLE

The ability to accommodate various capacity sizes of SSDs and the flexibility of scalability allows a tailored-fit configuration to meet the requirements of the individual customer. The table below captures the most common configurations and capacities.

# of K-Blocks	# of Expansion Shelves	# of IB Switches	Rack Footprint	Usable Capacity (*)
1	0	0	4U	7TB – 180+ TB
	1	0	6U	
2	0	2	10U	14TB – 360+ TB
	2	2	14U	
3	0	2	14U	21TB – 540+ TB
	3	2	20U	
4	0	2	18U	28TB – 720+ TB
	4	2	26U	

* The Usable Capacity range covers net physical capacity of different SSD sizes, as well as common data reduction ratios gained by inline deduplication and compression. For some datasets such as VDI, the range can be higher.

Storage Efficiency

The ability to build a cost-effective AFA relies almost entirely on the efficiency of the architecture or, in other words, how much effective capacity can be generated from raw physical SSD capacity.

SPEAR is focused on being highly efficient, but without compromising on other features such as enterprise resiliency and consistent performance.

It is therefore a combination of features that allow the efficient usage of Flash media and also complement the attributes of Flash. These features play a major role in the IO processing, as described in the IO Flow.

DEDUPLICATION

SPEAR's global inline selective deduplication meets the demanding requirements of eliminating redundant data so that it is stored on the array only once. The deduplication is performed globally and its processing is distributed across all the array's K-Nodes, enabling higher deduplication ratios, high performance and consistent low latency. As the array scales out, so does the deduplication. SPEAR offers the unique option, amongst AFAs, of selective deduplication. It allows storing data without deduplication for applications whose data redundancy is negligible and additional performance is preferred (such as database applications like Oracle or SQL Server), as well as for security-sensitive applications where deduplication is prohibited.

COMPRESSION

SPEAR uses inline real-time data compression that is optimized for low latency performance. The data reduction is gained regardless of whether the data is dedupable or not, which makes compression the de-facto method of data reduction for non-dedupable data sets that are common in database environments such as Oracle and SQL Server.

SPEAR uses the LZ4 compression algorithm with the ability to compress data in the granularity of bytes. This byte-aligned compression prevents internal fragmentation and facilitates better compression ratios. The compression is performed in a 4KB granularity rather than on bigger data segments, ensuring that small reads do not result in decompression of unnecessary data.

K-RAID™

Kaminario developed the K-RAID - a RAID scheme that is highly efficient with an 87.5% utilization rate. This high rate is gained without compromising on either the protection level, which is an optimized version of RAID6 protection, or on performance. This rate is achieved by deploying efficient erasure coding on each 24 SSD shelf. This erasure coding consists of two logical RAID groups, each one with a parity (P1 and P2), and an additional parity for the two groups (Q), as show in Figure 7, below:

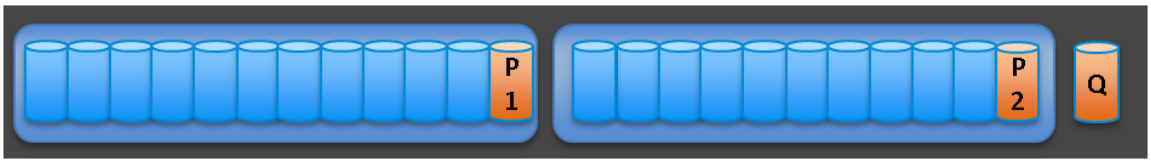


Figure 7: K-RAID logical layout of two RAID groups and a Q parity for both groups

The K-RAID protection and recovery abilities are detailed in the Enterprise Resiliency section.

THIN-PROVISIONING

Thin-provisioning allows maximum utilization of the storage array with the ability to plan storage provisioning for the long term. However, thin-provisioning can be truly utilized only with a scalable architecture that can facilitate capacity growth within the array, with no limitations. All the volumes in the K2 are thinly-provisioned, with a fine granular on-demand growth of 4KB. Un-map operations are also supported in the same granularity. SPEAR delivers the required management tools that bring the thin-provisioning feature to the next level, where the capacity management of volume provisioning is easy and hassle-free.

Enterprise Resiliency

The Kaminario K2 is architected and built to meet the requirements of the most sensitive enterprise applications. The enterprise resiliency starts by deploying only enterprise grade hardware components and continues with High Availability (HA) throughout SPEAR's design, scalability of fault domains and providing the right features for building an enterprise product.

K-RAID™

Aside from being highly efficient, the K-RAID is extremely robust. It can sustain two concurrent SSD failures and up to three SSD failures within each separate SSD shelf, without loss of data. As the K2 scales capacity, so does the number of system-wide SSD failures that the system can sustain. The K-RAID has a dual parity protection that adapts according to the failure at hand. An SSD failure is quickly recovered thanks to efficient metadata and real-time system health monitoring. It has minimal performance impact during the rebuild and no performance impact on the array's performance once the rebuild is completed.

K-RAID is built from two logical RAID groups, each one with a parity (P1 and P2), and an additional parity for the two groups (Q), as show in Figure 8, below:

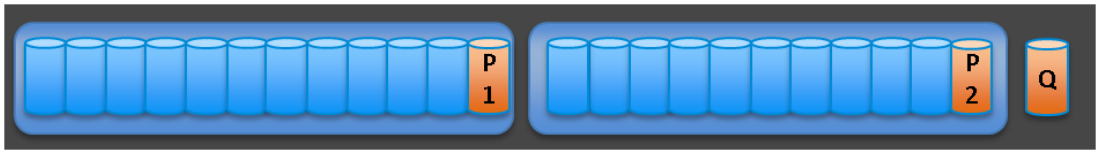


Figure 8: K-RAID logical layout of two RAID groups and a Q parity for both groups

The K-RAID layout is pseudo-randomly distributed throughout each SSD shelf, meaning that there is no fixed Parity or Data role per SSD and there is no hot spare, meaning that all the SSDs are utilized for protection and performance at all times.

The K-RAID is fully automatic and does not require any configuration or human intervention, which means another IT task is offloaded to the Kaminario K2.

K-RAID FLOW

1. The flow of how K-RAID works in the scenario of an SSD failure is shown in Figure 9, below:

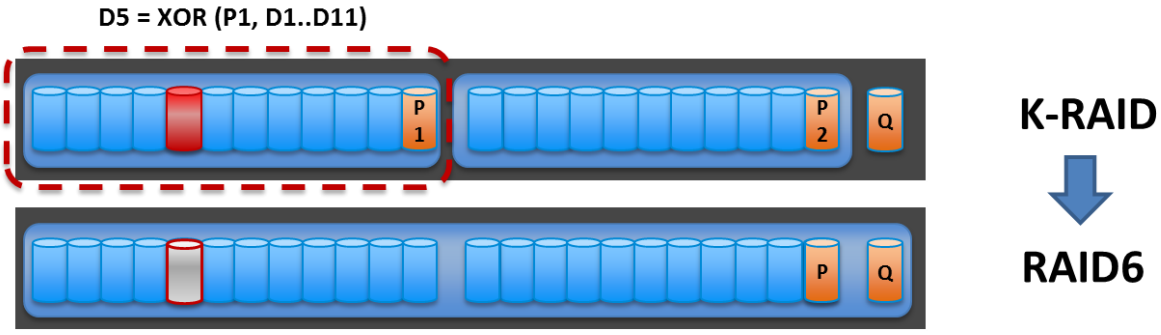


Figure 9: K-RAID Flow during SSD Failure

- The SSD failure is shown in the left RAID group. However, since the K-RAID layout changes within the SSD Shelf every 64MB, all the different K-RAID layouts experience a different failure, i.e., D5 is not necessarily affected in each layout. To be precise, it will only be D5 in 1/24 of all permutations of the K-RAID layouts.
- Due to the different K-RAID layouts, the rebuild process of the failed SSD is performed from all the 23 healthy SSDs, which results in a faster rebuild time.
- When reading from a RAID group that lost a Data segment, the XOR calculations are performed only from that affected RAID group, which is built roughly from half the SSDs in the SSD shelf, meaning superior read performance during rebuild. In that RAID group, the Parity segment will take over the role of the lost Data segment and the Parity of the unaffected RAID group (P2 in the figure) will become P, a product of XOR(P1,P2).
- When reading from a RAID group that lost a Parity segment (P1, P2 or Q), there is no performance hit at all.
- Once the rebuild is completed, the K-RAID falls back to a RAID6 scheme, meaning it is still capable of handling two more concurrent SSD failures without losing data.
- When the failed SSD is restored, the K-RAID will be restored to its original layout.

2. The flow of how K-RAID works in the scenario of a second SSD failure is shown in Figure 10, below:

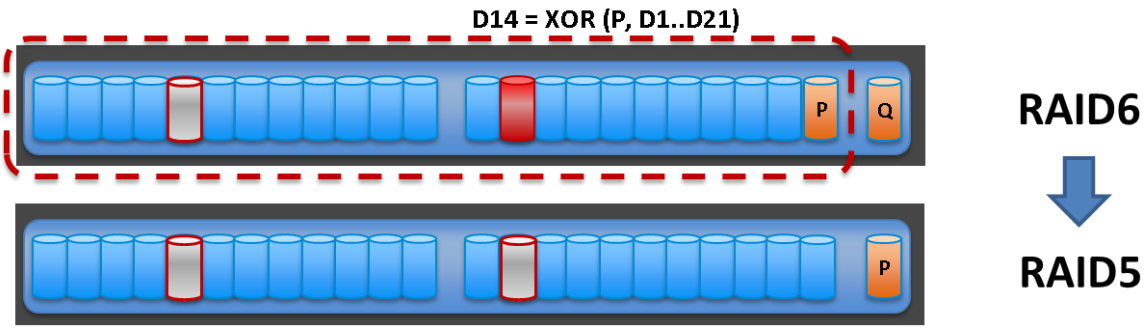


Figure 10: K-RAID Flow During Second SSD Failure

   - This is a standard RAID6 recovery.
   - When reading from a K-RAID layout that lost a Parity segment (P or Q), there is no performance hit at all.
   - The Parity segment (P) will take over the role of the lost Data segment and the second parity Q will become the new (P).
   - Once the rebuild is completed, the K-RAID falls back to a RAID5 scheme, meaning it is still capable of handling an SSD failure without losing data.
   - When the failed SSDs are restored, the K-RAID will be restored to its original layout.

3. The Q parity – handling concurrent failures. When two SSDs fail concurrently, the Q parity comes into play for those K-RAID layouts that lost two Data segments or a single Data segment and its Parity, as show in Figure 11, below:

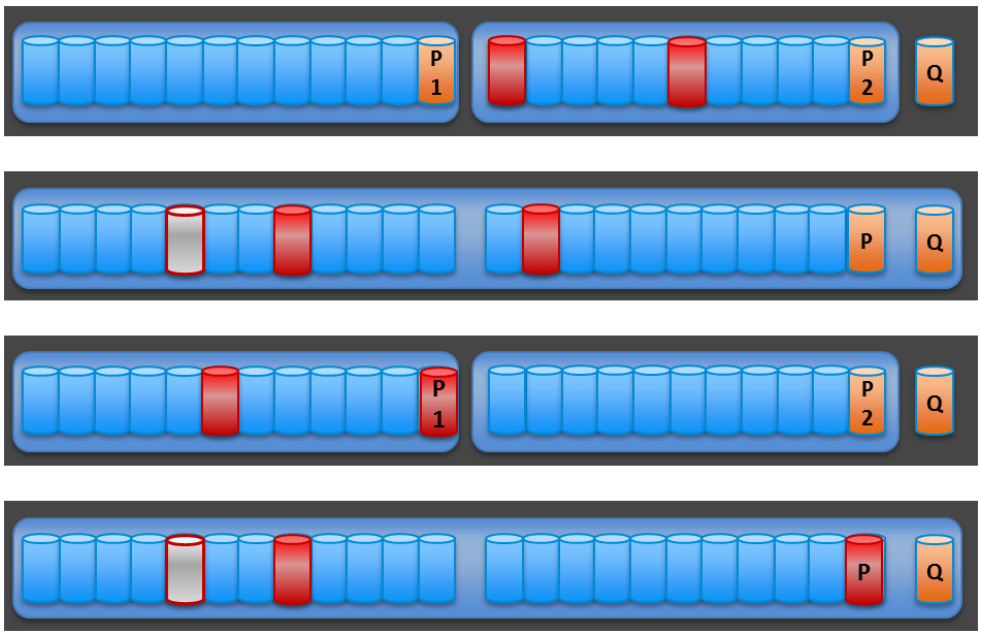


Figure 11: Using the Q Parity for two concurrent SSD failures

- SPEAR uses a set of highly efficient mathematical procedures for rebuilding the data out of the healthy Data segments, P and Q. In some of the K-RAID layouts, two concurrent SSD failures will be treated as two single, unrelated SSD failures. In other K-RAID layouts, one or two of the lost segments can be P1, P2 or Q.

NO SINGLE POINT OF FAILURE (SPOF)

SPEAR maintains a double-everything approach for K2's hardware components and all data, and metadata at rest is protected by the dual-parity K-RAID™. However, the K2 does not have passive or idle components in the array; all of its resources are being utilized at all times. There is full redundancy of every component in the system and there is not a single component that can fail and cause unplanned down time or data loss. Each K-Block is a standalone failure domain, which means that the entire array can sustain more failures as the array scales and the MTBF of the array stays constant.

POWER LOSS

A storage array with enterprise capabilities must have the ability to sustain a power outage in the datacenter and still keep the data intact and available for when power returns. Any metadata and/or data that were already acknowledged by the storage controllers (K-Nodes) before being stored to the K-RAID™ are saved in two distinct K-Nodes for redundancy. Each K-Node in the K2 is equipped with an internal battery that will provide sufficient power for the controller to destage any inflight data that has not been stored to its K-RAID. These batteries do not occupy any space within the rack since they are internal to the K-Nodes, which are 1U in size. Data that is already stored to the K-RAID is kept persistent and is sustainable through power cycles.

HEALTHSHIELD™

HealthShield is SPEAR's cloud-based Call Home monitoring, analytics and reporting module that facilitates preventive, proactive and automated enterprise level support for the Kaminario K2 AFA. HealthShield continuously monitors all of the array's components and is able, through a unique decision-making algorithm, to detect in real time any error that might occur. Any error or change in the system state triggers an automatic real-time event that is sent to Kaminario Cloud Support. It is also possible for the storage admin to subscribe to the events of interest by category such as LUN configuration, FC/iSCSI connectivity, hardware errors and so on.

On an hourly basis, HealthShield collects system-wide information that is sent to the Kaminario service center data warehouse, where it is analyzed and processed to recognize issues of performance, configuration, capacity management and more. This information is reported back to Kaminario Support, which uses a sophisticated BI system to provide tailored support for each customer.

NON-DISRUPTIVE UPGRADES (NDU)

The K2 can upgrade any of its hardware and software components with no impact on the availability or performance of the array. SPEAR is software-defined, so any new feature/enhancement/bug-fix/firmware can be deployed with no dependencies on maintenance windows or running workloads. In addition, hardware can be replaced/upgraded/added in the same manner. Having NDU combined with a scalable architecture, the K2 provides the best TCO of storage: No fork-lift upgrades, no need to plan down-time, new technologies can be deployed in the existing array and growing datacenter needs can be met by adding more capacity and/or performance. All of these operations are performed non-disruptively and automatically, with no human intervention.

SNAPSHOTS

K2's patented snapshot architecture follows SPEAR's guidelines of storage efficiency, performance and scalability. Snapshots are created instantly, with no performance impact and they do not take up any capacity. They track only the deltas from the volume in a 4KB granularity, using a redirect-on-write (RoW) approach. This storage-efficient design also keeps the impact on SSD endurance to a minimum. The snapshots can be mounted for read/write purposes, which serve to create additional working environments such as QA, Test&Dev, analytics, backup and more, all at a very low cost of storage capacity. Read/write snapshots deliver the same performance of the production volumes, without any impact on the production volumes.

The snapshots are created instantly, with no dependencies on the number or size of the volumes being snapped or how big the array is. Using the snapshots' restore functionality for recovery purposes is done without losing any of the snapshot history and is allowed at any time. The snapshots can be accessed from any of the storage controllers of the K2, without bottlenecks or load balancing of affinity to a specific controller.

Operational Management

SPEAR's principle of HA is also followed in the Kaminario Management Service (KMS). The KMS does not require additional hardware resources, since it is designed to natively use the K2's K-Nodes and is accessed using redundant secured 1GbE connections. There is no need to configure RAID groups, tune the array to a specific application or create affinities between volumes and controllers.

With the KMS, the storage admin has a myriad of options to manage and control the K2. All of these options adhere to two fundamental guidelines: Simple and easy.

GUI

The K2 GUI is slick and highly usable, keeping all the necessary information in a tiled dashboard. Each tile presents the essential details of a K2 management component: Health, Volumes, Hosts, Performance, Capacity and Events.

Figure 12: K2 GUI Dashboard

Clicking a tile will automatically open a more detailed and actionable view of that management component: Create volumes, snapshots, replication policies, capacity policies, assign hosts, manage FC/iSCSI connections and so on.

The GUI is accessible via any web browser and does not require any software installation.

CLI

The CLI is fully scriptable and allows full control of the features and functionality of the K2. The CLI is accessible via an SSH connection using a Linux shell or Putty-like utility.

RESTFUL API

The K2 exposes a full set of RESTful APIs that allows it to be controlled, managed and monitored via automatic third-party software and in-house management platforms.

VMWARE INTEGRATION

SPEAR incorporates full support for storage offloading APIs such as VMware's VAAI, including Thin-Provisioning support. This seamless integration is required especially when combined with K2's global inline deduplication, which facilitates fast cloning and no extra use of physical capacity.

The K2 introduces a vCenter plug-in, delivering the complete user experience of managing the K2 also from VMware's vCenter environment.

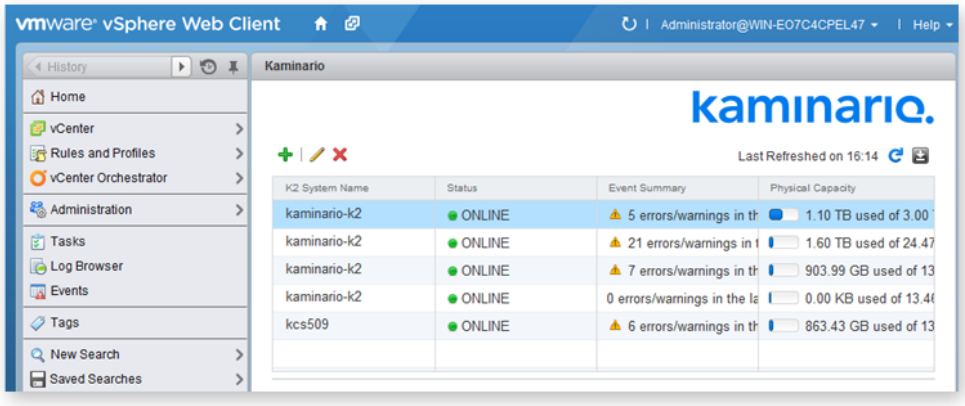


Figure 13: VMware vSphere Web Client, displaying the Kaminario vCenter Plug-in

IO Flow

SPEAR's architecture core functionality is the IO flow. In simple words, it is how data is written to the K2 AFA and how it is read from it. From the user's point of view, the K2 accomplishes these tasks efficiently, reliably and fast.

When taking a closer look at the IO flow, it can be observed that it encapsulates advanced technologies such as global inline selective deduplication, global adaptive block size, inline compression, distributed metadata, K-RAID and more – all developed in a true scale-out architecture. This section details these technologies.

GLOBAL ADAPTIVE BLOCK SIZE

Workloads generated by real applications vary in their block size. Storage arrays – specifically arrays that deploy deduplication – tend to use a fixed block size (usually 4KB), thus fragmenting the application's data blocks into small chunks. This method results in limited bandwidth for the application and multiple IOs for each IO that is bigger than that fixed block size. SPEAR's novelty is to adapt to the application's block size, which in return generates the best performance for the application's real workload without compromising latency, IOPS or bandwidth. The global adaptive block size algorithm allows the K2 to support the real performance requirements of a multitude of application types all running concurrently, which is the core essence of a primary storage array. This patented algorithm is crucial for deploying a true scale-out AFA.

METADATA (MD) MANAGEMENT

Metadata is essential and critical in any storage array, however, its significance grows tenfold when deploying a storage array that supports scalability on one hand and features such as snapshots, deduplication, compression and thin-provisioning on the other. The way SPEAR manages K2's metadata results in better performing deduplication and compression and allows real scale-out Active/Active access to all volumes and snapshots from all the K-Nodes (storage controllers). It also facilitates fast recovery in the event of a failure scenario and an optimized garbage collection process. Metadata is kept both on DRAM and SSD media, using a unique cache algorithm for consistent performance. This advanced choice of media for metadata placement eliminates any restriction on the capacity size of the solid-state drives (SSD) that are used in the array, any restriction on the maximum data reduction ratios and any restriction on the overall capacity within a single array.

FLASH ENDURANCE

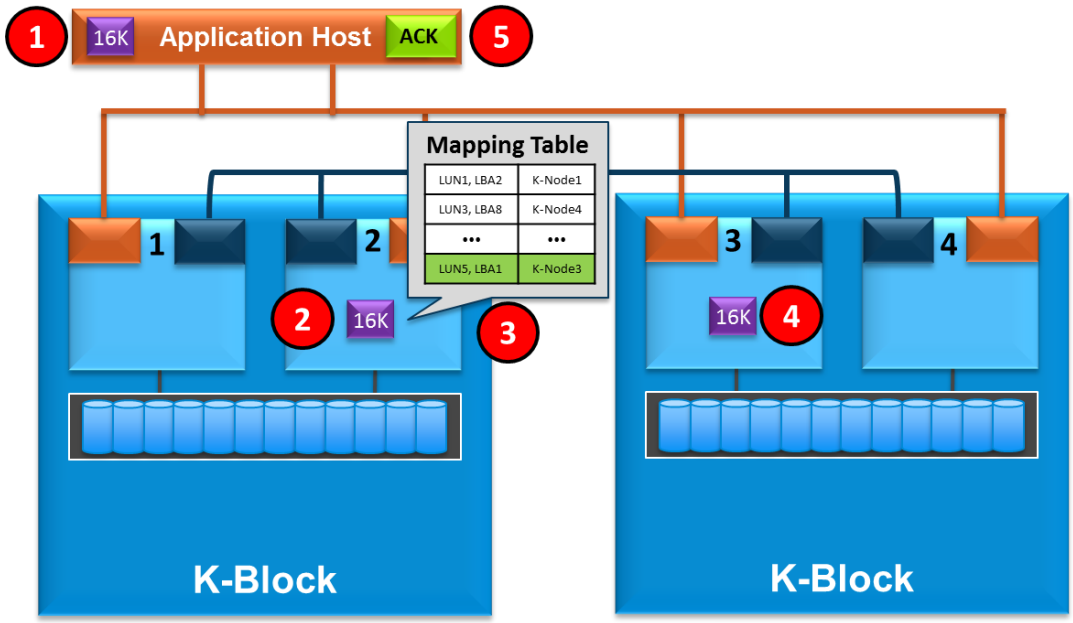
The Kaminario K2 utilizes Flash media. Therefore, SPEAR is designed not only to take full advantage of this media's characteristics but also to limit its shortcomings such as write endurance and write amplification. SPEAR's inline deduplication and compression reduce the actual data that is written to the SSDs; data that is written to the K-RAID is written using log-structured full stripes so that the number of updates per stripe is minimized. Writes are fully distributed across the entire array and a scalable distributed write cache eliminates hotspots.

FLOW OF OPERATIONS

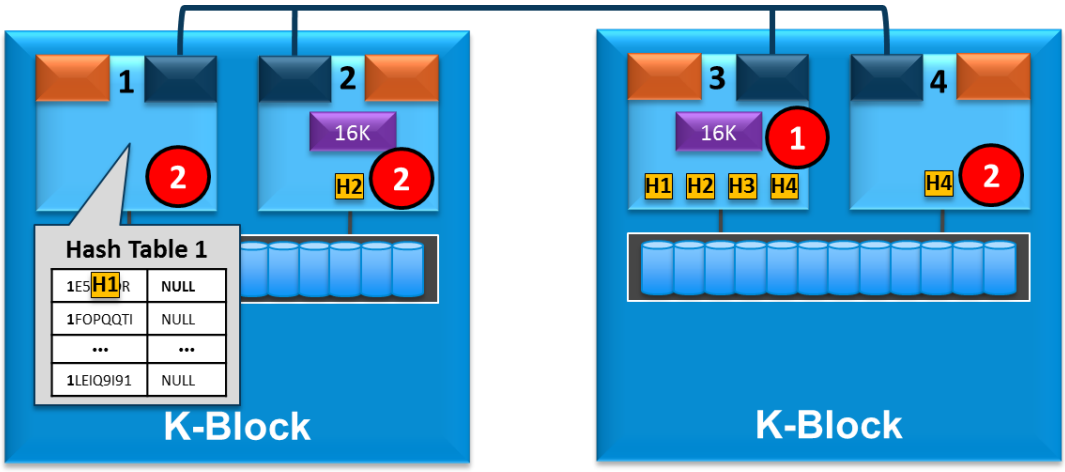
The flow below demonstrates a simplified IO flow of write and read. In the flow, a 16KB block is used. It shows how the global adaptive block size algorithm works in a scale-out array with a real application load. However, this is true also for larger block sizes as well, and of course when hundreds of thousands of IO operations are running concurrently through the array, additional mechanisms such as batching and queuing take place to provide further optimizations.

Unique Write

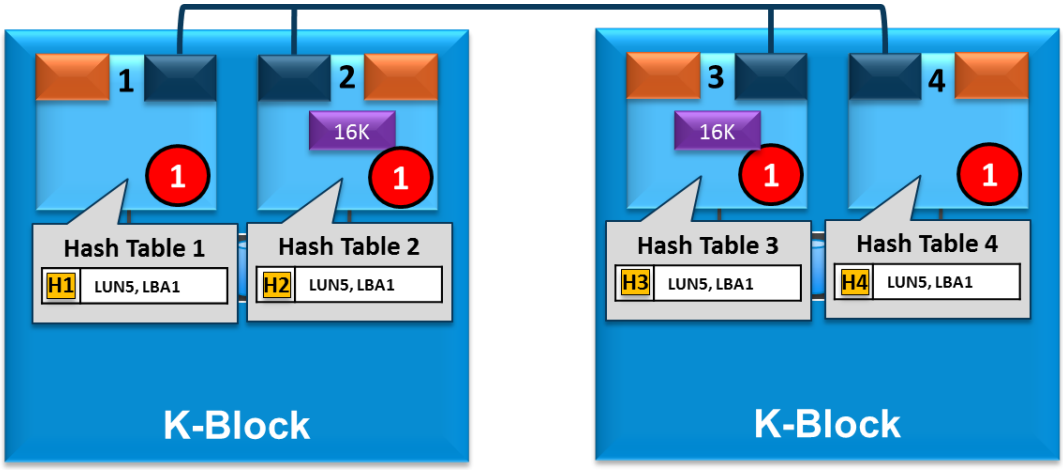
I. An application writes a typical block of 16KB to a logical block address (LBA1) within a block device (LUN5) on the K2 array (1). This block can arrive at any of the K-Nodes, since the K2 uses a scale-out Active/Active scheme. In this case, the block arrived at K-Node2. Once the block is stored to the K-Node's DRAM (2), it must be stored in another K-Node before returning an ACK to the application. The second K-Node is selected according to a mapping table that maps the (LUN, LBA) of the incoming write to a specific K-Node. In this case, K-Node3 is selected (3). This mapping table is identical in all the K-Nodes. The block is then mirrored over the IB fabric to the second K-Node, K-Node3, which also stores the block in its DRAM (4). At this point, the block is stored in two different physical servers, which have battery backups – the first K-Node can now return an ACK to the application's host (5). This means that any subsequent work performed while storing the new block is asynchronous and allows for low host-side latencies.



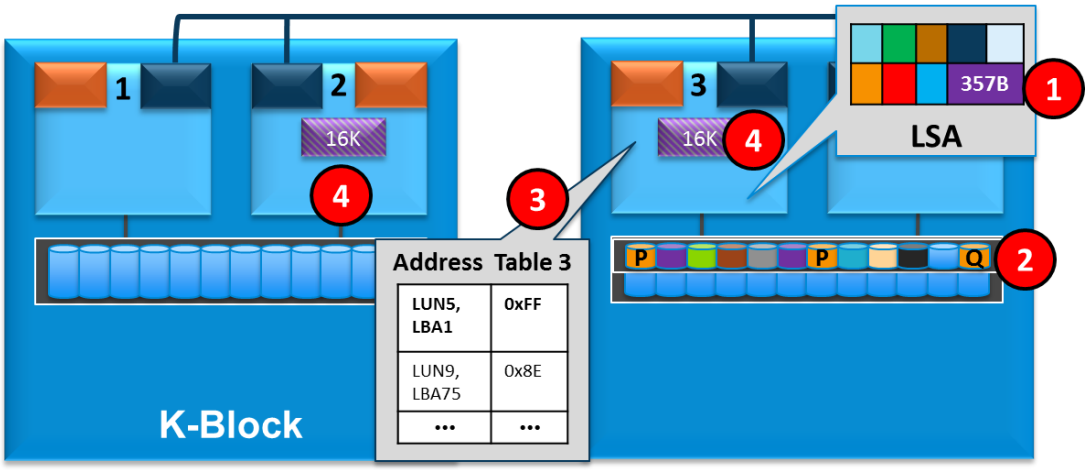
II. On K-Node3, this 16KB block is scanned in a 4KB granularity. For each 4KB, a hash is created, with four hashes all together (1). All possible hash values are divided into several hash tables, according to the number of K-Nodes in the array. Each hash value is sent over the IB fabric to the appropriate K-Node, and a lookup is done for each hash value (2). This will give a first indication of whether the respective 4KB was already written to the system. Since this is a unique write to the K2, all hash lookups will return negative.



III. K-Node3, The K-Node that originated the hash lookups is going to take ownership of the 16KB. All the K-Nodes that were queried before, updated their hash tables (at the same time of the query that returned NULL) with the owner of the 4KB data: LUN5, LBA1 (1). K-Node3 has the internal know-how to retrieve any of the 4KB out of the original 16KB when requested.

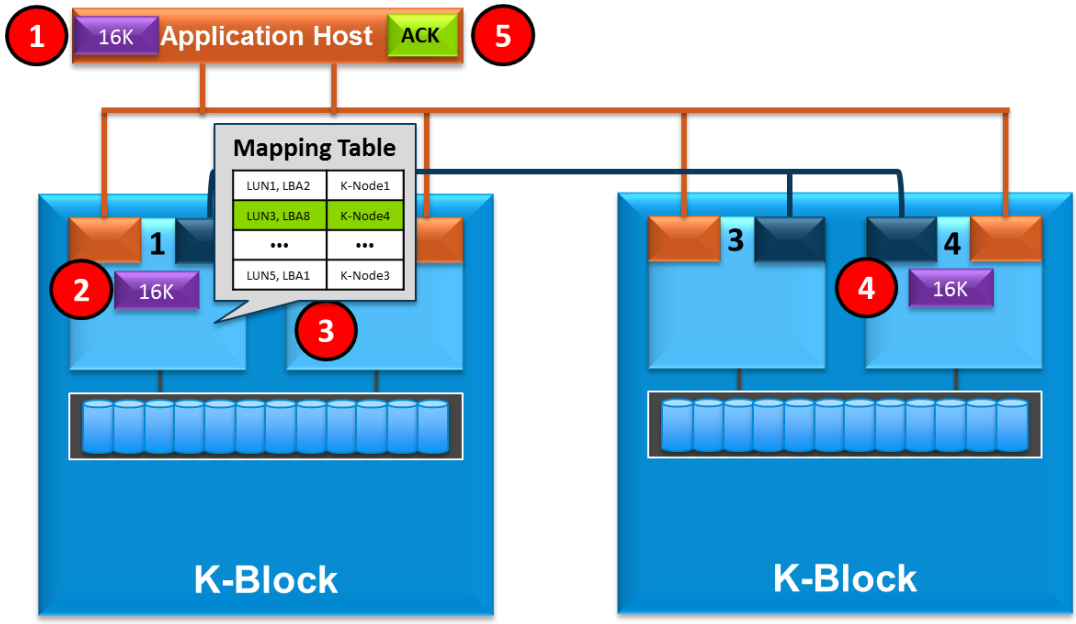


IV. The 16KB is now compressed - each 4KB is compressed separately - to the nearest byte (byte-aligned compression). The compressed 16KB is now placed contiguously in a log-structured array (LSA) page. This page is several MB in size (1). Once the LSA page has been filled, K-Node3 prepares a full RAID stripe calculating the parity in the DRAM and writes it to the K-RAID (2). K-Node3 keeps the location of the compressed 16KB in its Address table, which translates (LUN, LBA) pairs to physical addresses on the K-RAID (3). Once the 16KB is stored to the K-RAID, both K-Node2 and K-Node3 free the 16KB block from their DRAM (4).

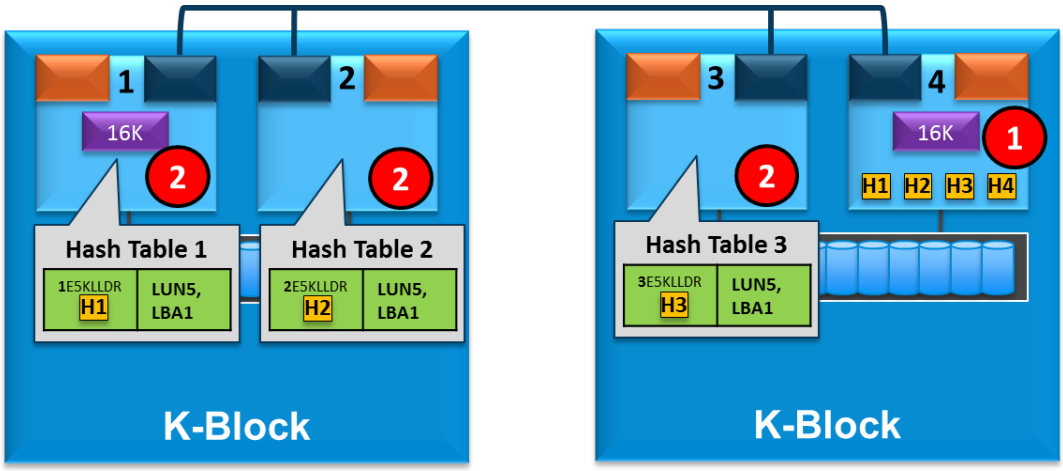


Dedupable Write

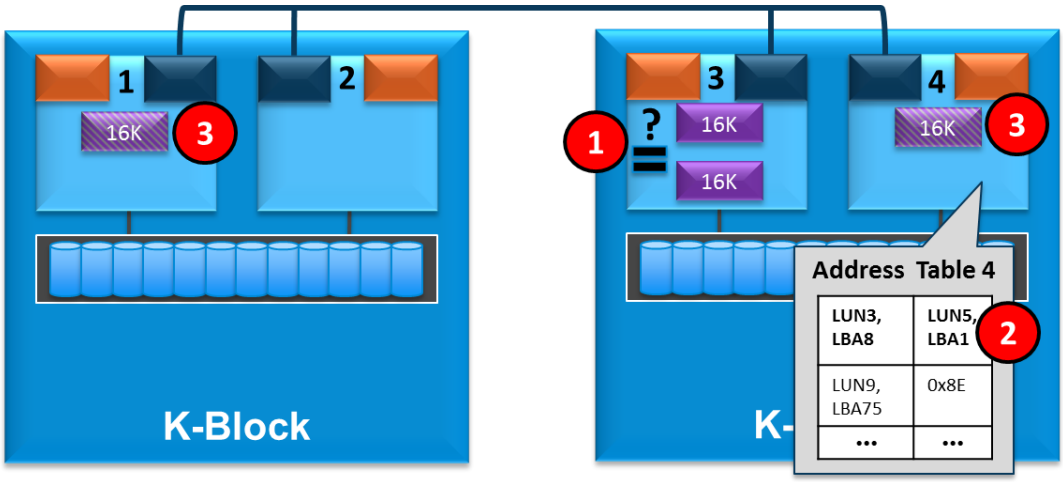
I. The same 16KB block is written again to the K2, with the following differences: The write is designated for LUN3, LBA8 (1). It arrives at K-Node1 and is stored in its DRAM (2). The mapping table indicates that the write should be mirrored to K-Node4 (3). The block is mirrored over the IB fabric to K-Node4 (4). At this point, the block is stored in two different physical servers which have battery backups – K-Node1 can now return an ACK to the application's host (5).



II. On K-Node4, this 16KB is scanned in a 4KB granularity, and for each 4KB, a hash is created, with four hashes all together (1). Each hash value is sent to the appropriate K-Node, and a lookup is performed for each hash value. All the K-Nodes report back that for the hash value they were queried, the address that appears is LUN5, LBA1 (2). Recall that the 16KB block that was written to LUN5, LBA1 was stored to the K-RAID by K-Node3.

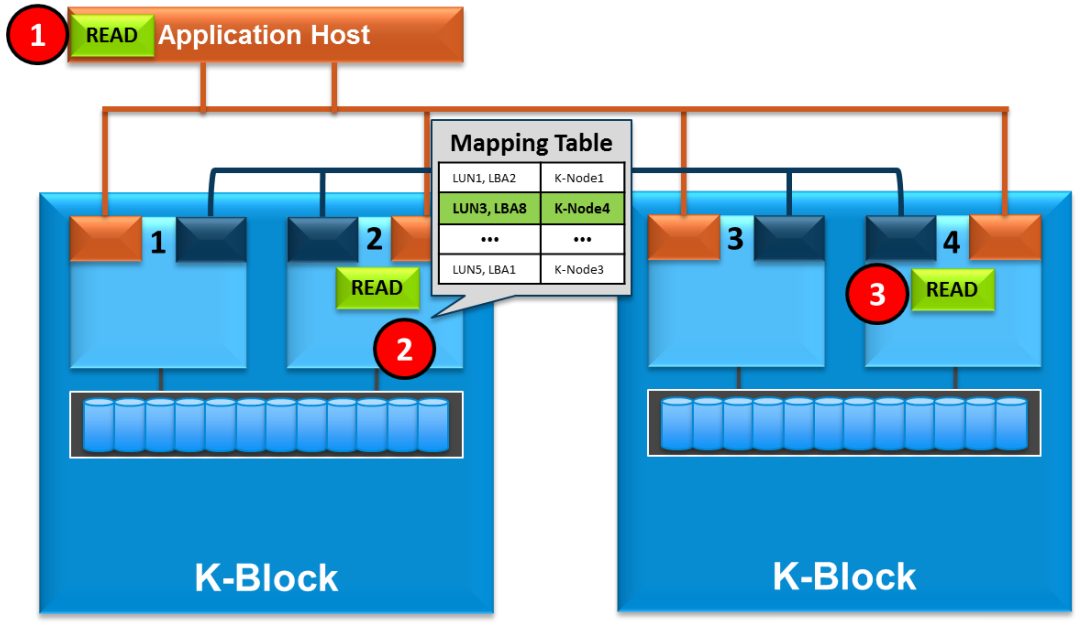


III. To avoid any probability of hash collision, it is necessary to compare the actual data. K-Node4 sends the 16KB block to K-Node3 for a full compare (1). This requires a single read operation by K-Node3 from the K-RAID, since it was stored as a single 16KB block. Once it checks out, K-Node4 updates its Address table at LUN3, LBA8 to point to LUN5, LBA1 (2). These metadata updates are mirrored between K-Nodes and eventually destaged to the K-RAID, so HA of metadata is kept at all times. It is now possible for K-Node1 and K-Node4 to free the 16KB block from their DRAM (3).

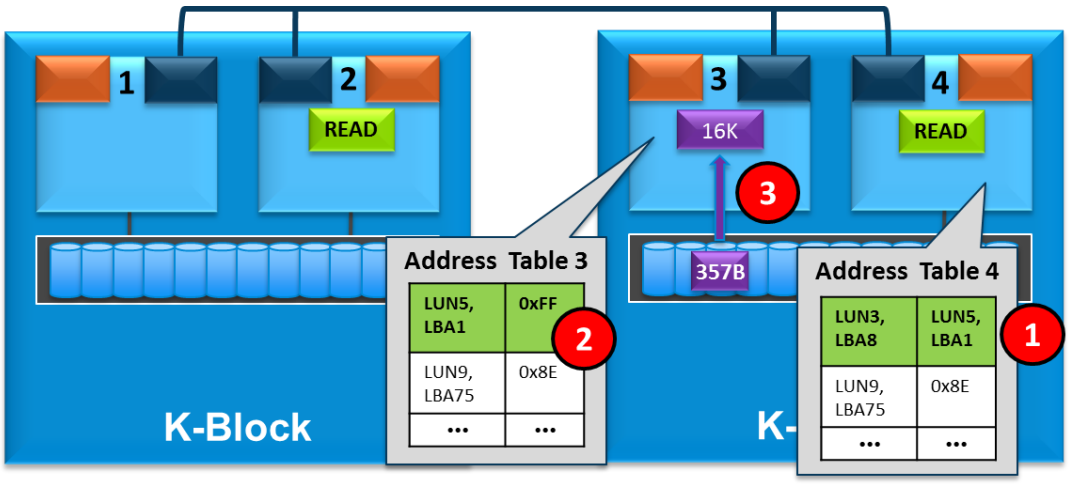


Read

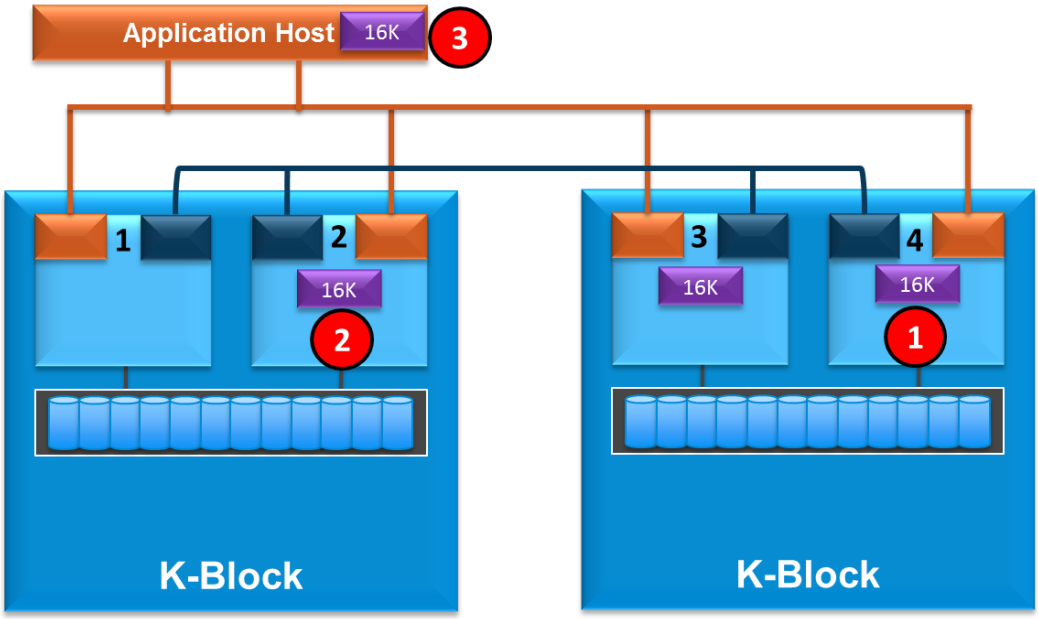
I. The application now reads 16KB from LUN3, LBA8 (1). This read request can arrive at any of the K-Nodes. In this case, it is received by K-Node2, which performs a lookup in its Mapping table for the owner of LUN3, LBA8 (2). According to the Mapping table lookup, K-Node2 relays the read request to K-Node4 (3).



II. K-Node4 looks up LUN3, LBA8 in its Address table and finds reference to a logical address, LUN5, LBA1 rather than a physical address (1). Using the Mapping table, it requests the data from K-Node3, the owner of LUN5, LBA1. K-Node3 looks up the physical address (2), retrieves the data and decompresses it (3).



III. K-Node3 sends the requested 16KB over the IB fabric to K-Node4 (1). K-Node4 sends the requested 16KB over the IB fabric to K-Node2 (2). K-Node2 sends the requested 16KB back to the application's host (3).



Summary

The Kaminario K2 All-Flash Array succeeds in positioning Flash as the right choice of storage media for primary storage arrays.

With a comprehensive set of storage efficiency features such as global inline selective deduplication, inline compression, thin-provisioning and the K-RAID™, the Kaminario K2 is challenging the cost of HDD-based and hybrid storage arrays.

SPEAR provides the Kaminario K2 AFA with a complete software stack of enterprise resiliency features such as high availability (HA), non-disruptive upgrades (NDU), snapshots, replication and cloud-based HealthShield™ for enterprise serviceability. In addition, the Kaminario Management Service delivers ease of use and various, flexible management abilities. With scale-out and scale-up capabilities, the same ease of use and management simplicity is gained for any array size, with performance and capacity to suit the datacenter needs - driving business agility to the top.

The Kaminario K2 All-Flash Array answers all the requirements of a primary storage array with no compromises.

For more information, please visit us at www.kaminario.com.