KAMINARIO FLASH ECONOMICS: ACHIEVING A $1/GB PRICE POINT

Author(s): Chris M Evans

A REVIEW OF THE COST EFFICIENT ARCHITECTURAL FEATURES OF THE K2 PLATFORM

First Published: September 2015
Latest Update: September 2015
Document ID - LB1CD0080
Release 4

This White Paper was commissioned by Kaminario Inc and written and distributed by Langton Blue Ltd.

Table of Contents

Figures

Executive Summary

Choosing the right all-flash architecture is becoming increasingly difficult in an expanding and mature market. At the outset of the availability of all-flash systems, purchases were justified on application performance requirements that couldn't be achieved using traditional HDD-based systems. The market has moved on and reached a point where choosing all-flash for all production applications is a viable proposition and that means being able to reach a price per GB ($/GB) figure that makes wide-scale adoption of the technology financially acceptable.

Over the last decade, flash SSDs (solid state disks) have increased in capacity, performance and endurance while dramatically reducing in cost. These achievements have been realised through a number of incremental steps, including an evolution in the way data is stored (SLC, MLC and now TLC), improvements in the manufacturing process (shrinking the size of each cell) and moving from planar to 3D technology. SSDs are cheaper than ever before, with capacities pushing 4TB per drive and the promise of 16TB drives in the near future.

All-flash systems need to be built upon the right architecture in order to exploit the benefits of new flash technology and make the effective $/GB ratio more acceptable. This means minimising the amount of data physically written to disk using space reduction features such as data de-duplication and compression. In tandem, flash updates need to be as low as possible, as flash lifetime (endurance) is finite and lowest with high-capacity, low cost SSDs.

Finally, cost reductions can be fully realised through efficient scaling of all-flash architectures. This means offering both scale-out and scale-up to obtain the most benefit from all components in an all-flash system.

K2 from Kaminario delivers all-flash storage at an effective cost of less than $1/GB with a 7-year guarantee. This is achieved through architectural features including a scale-up and scale-out design that provides a usable capacity of over 2PB per rack. K2 is able to use the latest highest capacity TLC flash drives available, through extending the endurance of drives by between 5-10 times and providing a cost saving of between 60-90% compared to traditional MLC. This combines with space efficiency features including data de-duplication and compression, with a typical total saving ratio of around 6:1. Data protection is implemented with K-RAID, a proprietary version of RAID-6 that minimises writes to flash and delivers a highly efficient 87.5% usable capacity after protection.

The combination of the right architecture and ability to safely use the latest NAND flash products means Kaminario's K2 system is able to deliver one of the most competitive effective capacity price points available in the market today. It is important to note that the $1/GB mark is achieved without needing to deploy the maximum configuration and inclusive of all software licences, meaning there are no hidden surprises when features like snapshots and replication are enabled. Customers can make the jump straight to all-flash, bypassing more complex hybrid solutions based on tiering and caching, as K2 is competitive on price, even compared to the latest hybrid platforms.

In summary this makes K2 the logical choice for all future production storage needs.

Introduction

Objective

This report looks at the architectural features of Kaminario's K2 platform that are used to achieve a usable price of less than $1/GB with the release of version 5.5 of K2. The contents highlight the key design strategies and their benefits compared to other all-flash storage solutions.

Audience

Decision makers in organisations looking to evaluate the financial impact of moving to all-flash configurations will find this report provides in-depth information on the features and functionality that should be considered when implementing flash-based solutions. The report provides a basis for decision makers to develop a comparison methodology and selection criteria.

Contents of This Report

- **Executive Summary** – a summary of the background and conclusions reached in this report.
- **Evolution of the all-flash Array Market** – a discussion on the growth of all-flash systems as a viable solution for all primary data requirements.
- **Assessing the Cost Factors** – a discussion of the factors impacting the cost of all-flash storage, including the evolution of NAND, space reduction techniques and architecture design.

Evolution of The All-Flash Array Market

Over the last 5-10 years we have seen issues arise with the ability of traditional hard-disk drive (HDD) external storage arrays to deliver to the performance requirements of modern applications. Some of this has been driven by the inevitable increase in processor performance and some through the now de-facto use of server virtualisation as the main deployment vehicle for new applications.

Storage I/O performance can be measured using a metric known as I/O density, or the number of IOPS that can be delivered per GB or TB of storage capacity. I/O density has increased steadily over time, as improvements in processor and memory performance have been delivered to the market. This has put pressure on HDD-based systems, where drive capacity has increased exponentially but performance has remained relatively static. In many cases, the performance required for applications was achieved by over-provisioning HDD capacity simply to gain the required spindle count (number of drives) necessary to deliver the throughput needed. Whilst this temporarily addressed the throughput issue, it didn't solve the issue of I/O latency.

Until recently, all-flash systems have been used to store data for only the most I/O intensive applications, where the cost of storage can be justified against the needs of the application or where HDD systems simply couldn't deliver the I/O density or required latency. As the all-flash market has matured, newer more flexible solutions such as K2 from Kaminario have pushed the cost boundaries to levels that can justify moving more and more data to all-flash.

All-flash storage is moving to be the mainstream de-facto choice for many IT organisations, and as this happens the traditional metrics around value for money based on $/GB once again become relevant to the total cost of ownership of storage.

Assessing the Cost Factors

The ability to achieve a low price point is based on a number of factors:

- **NAND Flash platform** – the specific choice of NAND chip architecture.
- **Data Reduction** – techniques such as de-duplication and compression.
- **Data Protection** – efficient use of RAID technology.
- **Efficient architecture** – scale-out in conjunction with scale-up and being "flash-friendly".

NAND Flash Evolution

It has been less than a decade since NAND flash technology was introduced into traditional shared external storage arrays. The availability of NAND technology was famously driven by the demand created with the release of the iPod music device by Apple Inc. Initial flash devices were based on a technology known as SLC (Single Level Cell), in which each "cell" (a group of transistors able to store state information) recorded a single bit of data, either "0" or "1". SLC technology is relatively expensive and was quickly superseded as the primary choice for SSDs by MLC (Multi-Level-Cell). MLC devices, despite their name record two bits per cell by storing four voltage states, representing the values "00", "01", "10" and "11".

The ability to store two bits per cell reduces the cost of MLC compared to SLC, however the impact of the MLC design is to also reduce the endurance of the NAND itself by a factor of ten compared to SLC. SSD vendors have worked hard to develop flash management techniques such as wear levelling and intelligent garbage collection that improve endurance as much as possible. The result of this has been to increase the reliability of SSDs to at least that of traditional hard drives, when measured over a five-year lifetime. A by-product of this work has resulted in a range of MLC SSDs with varying levels of endurance based on cost; once again $/GB is back into the storage equation.

The next step in NAND flash evolution is the move to TLC or triple-level-cell, where each cell stores three bits of data across eight voltage states (from "000" to "111"). In the same way as the transition from SLC to MLC traded cost for endurance, so TLC offers cheaper devices with a lower level of endurance compared to both MLC and SLC (but with continually improving algorithms to mitigate this). It's worth noting however that endurance of flash devices is based on data written and reading data from flash has no impact on NAND lifetime. This makes it possible to introduce tiering for flash, based on both performance and workload profiles.

All NAND technology is reaching limits of miniaturisation, as the size of the "production process" reduces towards the 10nm (nanometre) range. As a result, NAND manufacturers have looked to new techniques in order to increase flash density, one of which is known as 3D or V-NAND. 3D-NAND stacks (or more accurately digs out) multiple cells vertically in the silicon substrate, creating a 3D structure compared to the traditional 2D or planar NAND. Vertical stacking promises to dramatically increase flash densities, with existing products already using 32 and 48 layers.

All of the evolutions in flash point to a similar capacity growth trajectory to that of hard disk drives, without the restriction of a low I/O density (IOPS/GB). Flash will evolve to deliver multi-petabyte devices, with a range of price, performance and endurance specifications to suit all workload types. As flash has been adopted by the IT industry, it has become clear that the level of write IOPS needed in an all-flash array to sustain typical workloads doesn't require high endurance flash, when combined with "flash-friendly" technology and so greater business value will be achieved by matching endurance to workload profiles.

Kaminario K2 currently supports 480GB, 960GB (3D) and 1.92TB 3D TLC flash SSDs.[1]

Data Reduction Techniques

The rate of data growth continues to be a major issue for all IT organisations from both a cost and manageability perspective. In the early days of flash storage, the relative expense of flash was justified for only a subset of high performing applications. In order to reduce the effective $/GB cost, all-flash vendors looked to solutions that would reduce the physical amount of data stored compared to that logically written to the system. Reducing the number of actual writes to flash had a dual effect; firstly, it reduced the impact on endurance and second it allowed prices to be quoted using "effective" rather than "raw" $/GB figures. This made flash storage more attractive compared to traditional HDD-based arrays. Two techniques used for reducing the physical amount of data stored are data de-duplication and compression. Both techniques look for and eliminate repeated or redundant data within the storage infrastructure.

Data de-duplication (commonly called dedupe) looks to highlight blocks of repeated data by storing only a single instance of that data and logically pointing all references to the single physical copy. This technique is highly beneficial where there is a large amount of similar data stored in the array, such as operating system files and email attachments. The technology derived from disk-based data protection, where savings of 95% were typical when storing multiple full-image system backups.

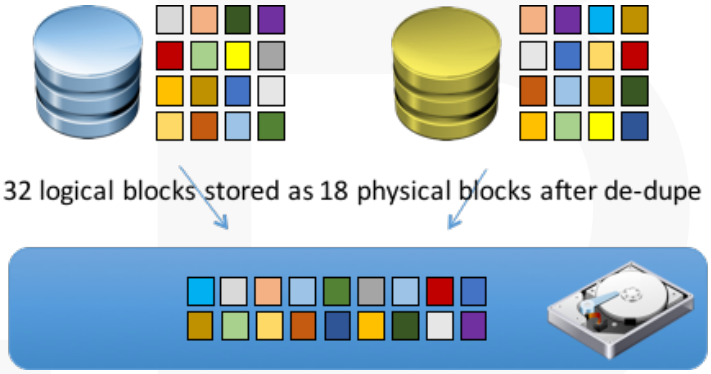

Figure 1 - effect on physical capacity of deduplication

Implementing dedupe requires the right metadata structures in order to track the logical to physical relationships, plus the ability to quickly identify duplicate data "inline" as an application writes to the storage array. Storage array vendors have until recently had to rely on custom ASIC processors to perform the identification of duplicate data, however today's processors are fast enough to perform this process directly. It is worth noting that de-duplication can be done as a post-processing task in order to mitigate any write I/O latency penalty, however this technique provides no benefit in reducing flash writes as the data is written to flash before being optimised at a later time.

[1] More details can be found in the K2 Data Sheet.

Compression is a process that highlights repeated patterns in data and replaces it with an encoded copy using less physical space than the original data. This technique is beneficial when used on data sets that contain lots of repeated data or white space, such as databases. Implementing compression has traditionally been a computationally expensive task, requiring dedicated FPGAs or ASICs, however today's processors can be used to compress data directly.

The savings made from data reduction are highly dependent on the content of the data itself, however as a storage system grows in capacity, then the savings become more predictable. Typical values range from around 4:1 to 10:1 depending on the data type and quality of the reduction algorithms (for example database data sees lower reductions).

K2 systems use both de-duplication and compression. De-duplication is global and distributed across all nodes (K-Blocks) within a K2 system, ensuring savings continue to be achieved as a K2 system is scaled out. The de-duplication feature can be applied selectively, ensuring savings are made on data that is most appropriate for the technology and allowing sensitive data (such as that under compliance) to be excluded from the de-duplication process.[2] It is important to emphasise the savings that are made when de-duplication is globally applied in a scale-out system. When multiple scale-up systems are deployed, de-duplication occurs only within the individual system, resulting in multiple copies of data being retained across the separate arrays. Some scale-out systems are also not capable of de-duplication across nodes. K2 stores one copy of unique data per system.

Compression in K2 systems uses the LZ4 algorithm and is based on 4KB block granularity. The process is byte (rather than block) aligned, ensuring that the highest level of compression is achieved and fragmentation is avoided.[2]

Data Protection

No shared storage systems would be implemented today without some form of local data protection such as RAID (redundant array of independent disks). RAID provides automated protection to recover from device failure, without the need to resort to backups. Implementation schemes include simple data mirroring (RAID-1, RAID-10) and parity-based protection (RAID-5, RAID-6). Parity schemes provide a lower overhead in terms of the additional capacity needed to implement protection (RAID-1 has a 100% overhead or provides only 50% usable capacity, whereas RAID-5 7+1 represents an overhead of only 14% or 87.5% usable), but the updating of data in place within a RAID stripe results in many physical writes for each logical block of data updated.

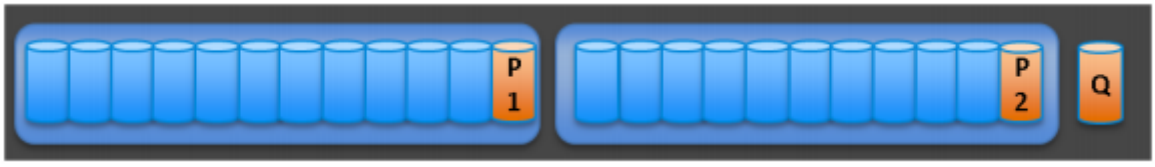

Figure 2 - K-RAID dual parity protection scheme

[2] Further details can be found in the K2 Architecture White Paper

This “write amplification” effect seen in RAID is something to be avoided in all-flash systems where the amount of write I/O needs to be kept to a minimum. It is therefore imperative that RAID implementations in all-flash systems are both space and I/O efficient.

K2 systems implement a proprietary RAID system known as K-RAID^2. In terms of resiliency, K-RAID is able to sustain a failure level of two concurrent SSD failures within any single shelf, without data loss. This is comparable to a dual-parity RAID-6 protection scheme, yet offers a highly efficient 87.5% utilisation rate. K-RAID also ensures that flash endurance isn’t compromised, by writing data to disk using a sequential (log structured file) scheme.

Architecture

It’s clear from the points presented so far that the architecture of all-flash systems is important in driving down cost and improving efficiency. In addition to the issues already raised, the ability to scale represents a key benefit in reducing the overall $/GB cost of flash storage.

Legacy storage arrays were originally based on scale-up architecture. This treated the array as a single monolithic device, where capacity was increased by adding extra drives to the system. Performance mapped directly to the capability of the system controllers, placing a limit on the throughput and bandwidth a single system could provide.

Scale-out systems are based on multiple nodes, each of which has controllers and storage capacity. Some implementations such as K2, build resiliency into each node, whereas some solutions cater for and expect node failure. Both performance and capacity are increased through the deployment of additional nodes as a cluster in a single system; clusters can be tightly or loosely coupled, depending on the design.

Comparing the two architectures, scale-up systems are easier to implement at a lower cost but have more restrictive expansion capabilities. Scale-out systems are more complex to implement (at a higher cost) but provide better scalability, performance and resiliency. The combination of both scale-out and scale-up in a single architecture provides the ability to gain the best features of both architectures, scaling in either direction as capacity or performance requires.

In Kaminario K2 systems, capacity can be increased on each K-Block (the unit of scale-out expansion) by adding an expansion flash SSD shelf. This allows the performance capacity of the K-Block controllers to be fully exploited and is a key advantage in delivering mixed SSD capacity configurations. Performance and capacity can be increased by adding additional K-Blocks to a K2 configuration, bringing in controllers and disk shelves.

K2 supports mixed configurations, allowing the customer to install the latest K-Block hardware within an existing system. Scale-out therefore allows K2 systems to take advantage of K-Blocks with increased controller performance and larger capacity drives. Scale-out is a significant benefit when looking to replace old storage systems, by avoiding the “forklift upgrade”.

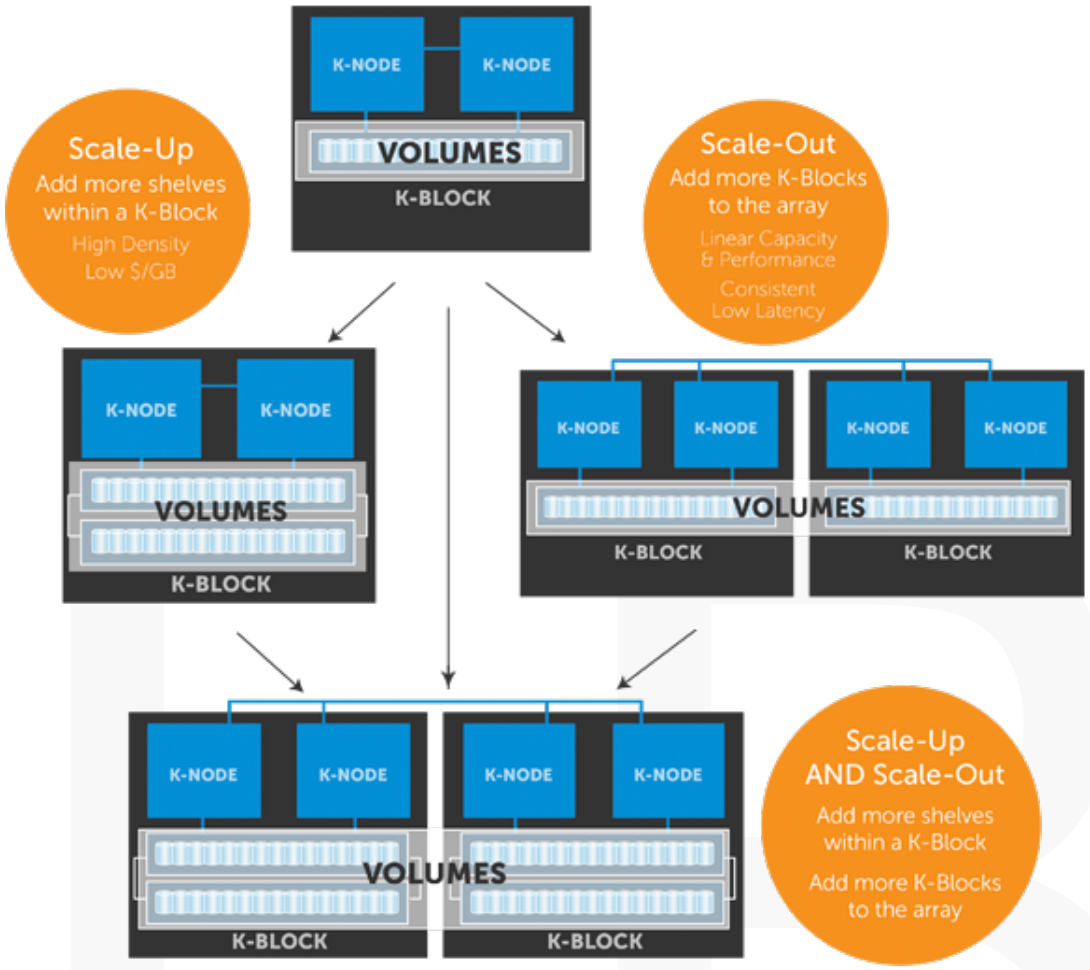

Figure 3 - scale-up and scale-out options with K2

Modern all-flash architectures write to flash in a way that improves flash endurance. These features include (as already discussed), inline data reduction, SSD "friendly" writing (log structured files), automated load balancing of data and metadata across all nodes and optimised RAID (full stripe writes).

Having a flash-friendly architecture reduces the level of over-provisioning needed on SSDs, where a portion of space on the drive remains unused in order to mitigate the impact on performance of garbage collection (a technique analogous to "short stroking" HDDs). K2 systems are capable working with much lower levels of over-provisioning, while improving the endurance of flash by 5-10 times and is one of the reasons Kaminario can offer a 7-year guarantee, even with the latest TLC technology.

Finally, we should highlight the effect of licensing on $/GB pricing. Some vendors licence snapshot and replication features separately from the hardware. K2 systems however are delivered fully inclusive of all software licences and therefore provide a predictable cost to the customer. This makes delivering storage services to end users more transparent and easier to calculate.

More Information

Full details of the Kaminario K2 architecture can be found in the white paper "K2 Architecture White Paper" available on the Kaminario website.

For additional technical background or other advice on the use of flash in the enterprise, contact enquiries@langtonblue.com for more information.

Langton Blue Ltd is hardware and software independent, working for the business value to the end customer. Contact us to discuss how we can help you transform your business through effective use of technology.

Website: www.langtonblue.com
Email: enquiries@langtonblue.com
Twitter: @langtonblue
Phone: (0) 330 220 0128

Post:

Langton Blue Ltd
133 Houndsditch
London
EC3A 7BX
United Kingdom

The Author

Chris M Evans has worked in the technology industry since 1987, starting as a systems programmer on the IBM mainframe platform, while retaining an interest in storage. After working abroad, he co-founded an Internet-based music distribution company during the .com era, returning to consultancy in the new millennium. In 2009 he co-founded Langton Blue Ltd (www.langtonblue.com), a boutique consultancy firm focused on delivering business benefit through efficient technology deployments. Chris writes a popular blog at http://blog.architecting.it, attends many conferences and invitation-only events and can be found providing regular industry contributions through Twitter (@chrismevans) and other social media outlets.