LATENCY MATTERS

In the era of flash storage, latency is the metric that matters.
Jump in the driver's seat and win the storage performance race.

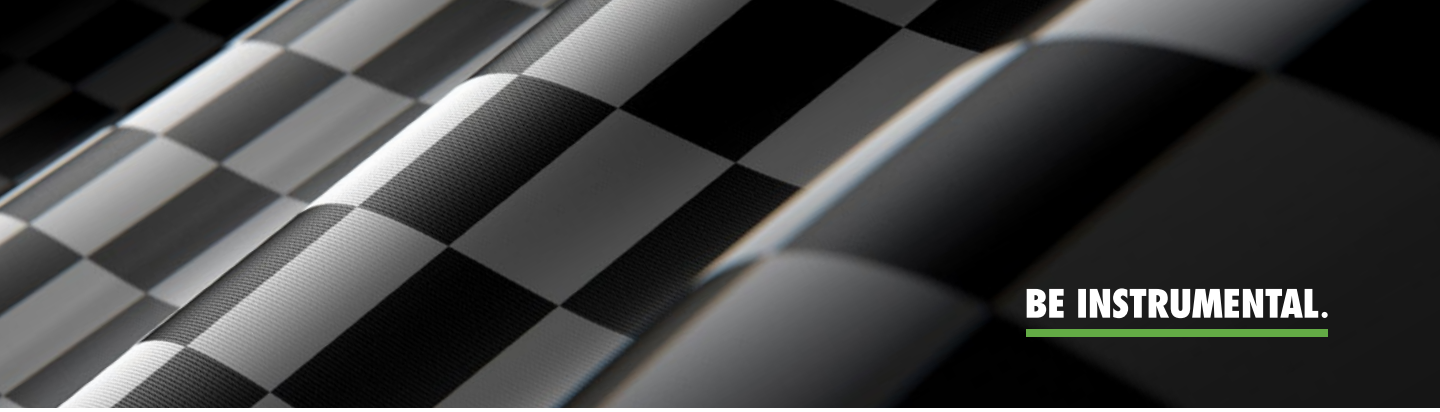
Violin® MEMORY

BE INSTRUMENTAL.


BE INSTRUMENTAL
Violin MEMORY
ULTRA LOW LATENCY
AFFORDABILITY
DENSITY
SCALABILITY
PERFORMANCE
Violin MEMORY
BE INSTRUMENTAL.

TABLE OF CONTENTS

Violin MEMORY



FOREWORD: BE INSTRUMENTAL

As we look at the world of digital data, storage is the focal point for the delivery and long term protection of this information lifeline. We demand seamless access to this information from any device, any time and any place.

There has probably been no other change in the storage industry that has had as profound an impact on our lives as flash memory. Flash has enabled consumer electronics and the portable devices we depend on for access to everything important to us and at price points that are extremely compelling.

As we look at the Enterprise, flash has played a vital role in processing transactions with incredible speed that drive the core applications of business today. Flash has also enabled a whole new set of applications commonly called Big Data. Companies are doubling their investment in all flash solutions to address real-time decision making based on deep data machine learning and fueled by the Internet of Things (IoT).

Formula 1 Racing characterizes the need for speed as well as anything. The highest top speed ever achieved in Formula 1 was by Juan Pablo Montoya in the 2005 Italian Grand Prix hitting a top speed of 231mph. However, in Formula 1 what really matters is winning races, not just executing a single fast lap.This is also true with storage as we see all kinds of published hero numbers from storage vendors. What really matters is what kind of performance is sustainable, consistent, and predictable. Unlike the racetrack where there are a fixed number of laps and then the race is over, enterprise customers need to run the race every hour of every day as the race never ends.

In this collection of essays, we're going to take a look together at several perspectives on the statement: "Latency Matters." As we take this journey together, you will learn some useful tips and hopefully, start thinking about when it makes sense for you and your organization to start moving your storage infrastructure to an all flash storage solution that delivers the lowest latency.

The Violin "Latency Matters" Series of six chapters to follow will address the following topics and answer the following questions:

"LATENCY MATTERS" SERIES

1. THE NEED FOR SPEED: IOPS VS. LATENCY

We start the conversation about performance, IOPS and latency.

Which storage will win the race — the one with a higher top speed (IOPS) or the one with better acceleration (latency)?

2. LAP TIMES: CONSISTENCY DRIVES BETTER OUTCOMES

This article discusses how customer experiences can be threatened by unpredictable performance of the storage environment.

What has a greater impact on databases and applications driving the business – the number of transactions that can be completed at the same time or how long it takes for each individual transaction to be completed?

3. STREET CIRCUIT VS. ROAD CIRCUIT: LOWERING LATENCY DRIVES WORKLOAD MIXING

This article discusses how the unpredictable latency of mixed and multiple workloads can threaten customer experiences.

Can I mix different types of workloads on the same all flash storage platform and still achieve consistent performance?

4. PERFORMANCE IS BOTH SPEED AND LATENCY

This article discusses the compounding effect of high latency on meeting SLAs for mission critical applications.

Are hero numbers of the highest possible IOPS that can be achieved a relevant metric for evaluating storage array performance?

5. WINNING: INCREASING BUSINESS VALUE WITH LOWER LATENCY ALL FLASH STORAGE

This article discusses the real-world experiences of enterprise businesses and reveals how lower latency has dramatically improved the return on investment (ROI) of all flash storage.

Can lower latency have an impact on greater people productivity, higher company morale, increased work quality, greater workplace efficiencies, and lower operating expenses?

6. MEET THE WINNER IN THE ALL FLASH STORAGE WINNER'S CIRCLE: IOPS ARE FUN, BUT LATENCY WINS

We announce the winner in our Need for Speed Sweepstakes and provide more answers to why lower latency matters.

Can something as small as 500 microseconds of latency difference actually be directly monetized into anything of significance?

KEY QUESTIONS

Enterprise applications clearly have a Need for Speed. The essential element for achieving application speed starts with all flash arrays. However, success depends on having the optimal speed to meet the services level objectives of the business. This raises the following technology and business questions:

TECHNOLOGY QUESTIONS:

How do we measure performance and what metrics should I use when evaluating the speed of my storage?

Which performance metrics have the greatest impact on customer application speed?

Are the storage vendors publishing performance specifications that are reflective of what customers really need to drive their application workloads?

BUSINESS QUESTIONS:

What impact does the response time of my storage have on the customer experience?

What does it cost my company when applications experience negative spikes in performance?

Are the storage vendors all the same in their ability to meet my specific performance requirements?

The answers to these questions really matter. Keep in mind that most customer environments have mixed workloads that are all vying for performance, so how do I meet the Service Level Agreements (SLAs) required for the company's success?

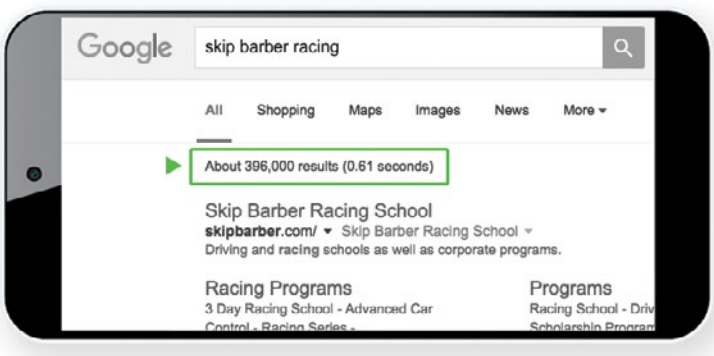
WHAT LIES AHEAD

The chapters that follow reveal various reasons latency matters in your decision about a storage strategy and your investments in a primary storage platform. Chapters will look at the impacts of latency to your business technically, financially, and strategically. They will also explore how latency impacts your customers and their online experience. Speed is important and we look forward to taking this journey with you as we explore your future considerations for a high performance all flash data center. The race is on!

1

THE NEED FOR SPEED – IOPS VS. LATENCY

Many in IT strive for the Google experience – the ability to deliver SLAs that allow for instant log-on, instant access to information, and instant results. Do a search on Google – and the expectation is the result will be served up in under a second, as shown in the below example:



We are now conditioned for this rapid response and so are our respective internal and external customers. The question to ask yourself is: "Can my storage environment support this expectation?"

To help you **Be Instrumental** in delivering this need for speed, we are going to explore why latency matters over the course of the next five chapters, sharing leading third-party points of views as well as our own so as to equip you with the right insight so that you can be successful.

SO LET'S START DRIVING

Many will profess a need for speed. Maybe that's why all flash storage is becoming more common within enterprise data centers. Companies' fascination with their own dashboards, in particular their speedometer quite perplexing.

Across storage brands and models, the constant metric used compares I/Os per second (IOPS) that their array can deliver to that of their competitors. Given its prevalence, you would think that this is the one metric that separates the best arrays from the rest.

We get the human mindset and need for speed, but for those who choose to test drive and ultimately buy high-performance machines (whether they are race cars or storage arrays), what really matters is how fast can you go from zero to 60 (or 90 or 120) and can you sustain that performance over time. What doesn't matter is if your speedometer reads 120, 180 or even 260 (420 km/h) for merely a split second.

Storage expert Scott Lowe's article on Enterprise Storage Guide states that while IOPS is certainly a useful number to have, it is only valuable when taken in context of how it relates to latency.

IOPS VS. LATENCY

IT professionals should think of IOPS as the number of transactions that an array can pass through an aggregate of all its ports in a given second, while latency is the amount of time it takes to process the transaction within the system.

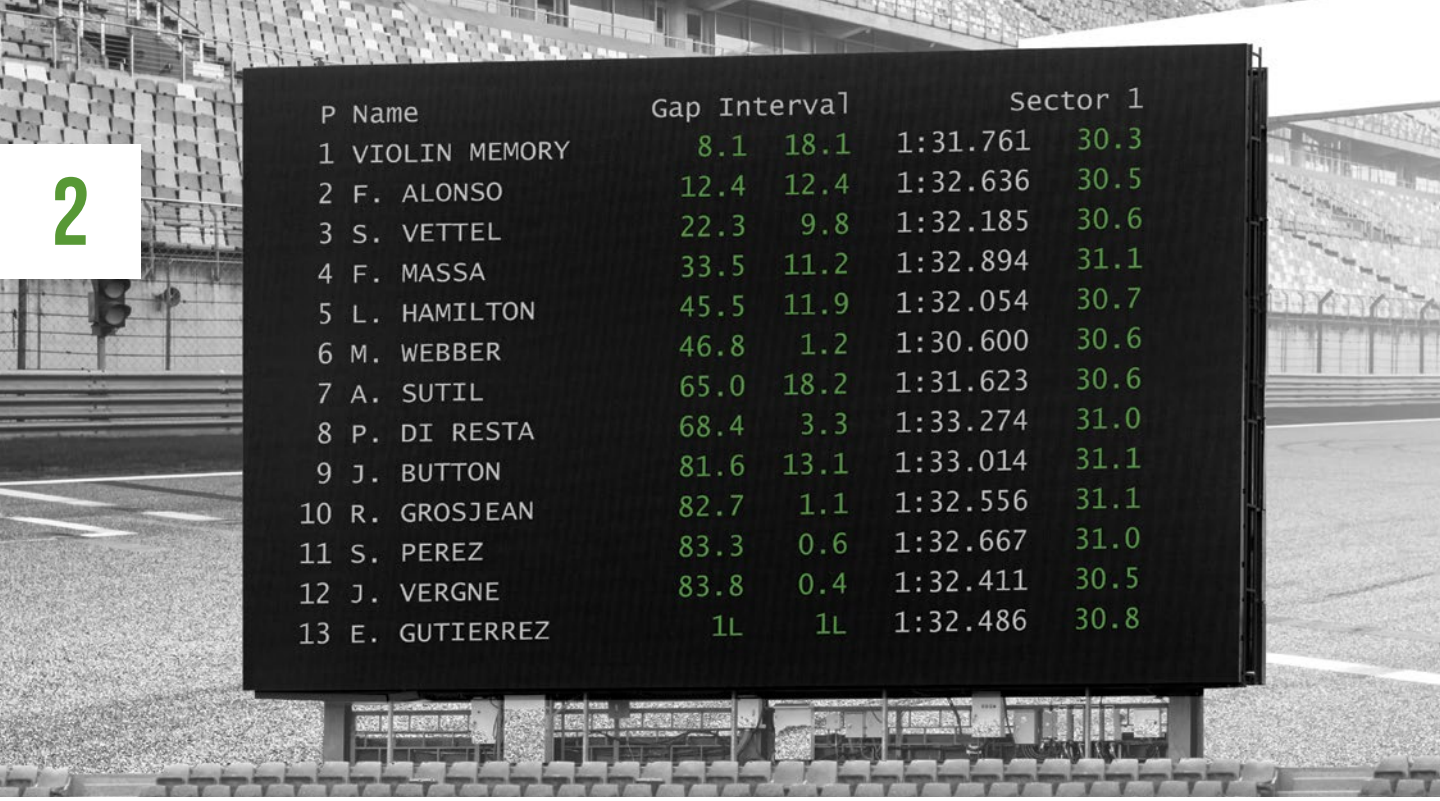
You can think of IOPS as a measurement of the top speed a particular race car can reach. If one car has a top speed of 130 miles per hour and another has a top speed of 200 miles per hour, which one will win the race? Well, it depends on the context of their latency.

...the difference between [an array] that can do 50,000 IOPS at 500 microseconds compared to one that does the same 50,000 IOPS but at 2 milliseconds can fundamentally alter your data center. It may sound like we are talking about the difference of a few thousandths of a second, but in this example it represents a real-world performance difference of 4x.

How long does it take each car to reach their top speed? If the turns on the track limit the cars to about 80 mph and the 200 mph car accelerates to 80 in 12 seconds, but the 130 mph car can do it in only 8 seconds, which car is going to win? The one with a higher top speed (IOPS) or the one with better acceleration (latency)?

IOPS are an important measurement of storage performance when taken in the context of latency. The difference between an array that can do one million IOPS and one that can do 300,000 IOPS is not relevant to you if your application only does 50,000 IOPS. However, the difference between one that can do 50,000 IOPS at 500 microseconds compared to one that does the same 50,000 IOPS but at 2 milliseconds can fundamentally alter your data center. It may sound like we are talking about the difference of a few thousandths of a second, but in this example it represents a real-world performance difference of 4x.

The less latency that your storage has, the more that it can do. Your storage can act on requests faster and deliver more data to more processors in less time. This means your applications can run faster, you may need fewer servers, and the servers you already have can do more. That's why latency matters.



2

LAP TIMES – CONSISTENCY DRIVES BETTER OUTCOMES

Companies rely on Information Technology (IT) infrastructure to support their business objectives. This means databases and applications need to run fast enough to deliver information to users in a timely fashion, but determining what it takes to achieve performance is not as straightforward as many expect. The challenge is that performance is determined by how latency and IOPS (Input/Output Operations Per Second) interact, especially in all flash storage.

CONSISTENT LOW LAP TIMES WIN CHAMPIONSHIPS

There is an excellent analogy between Formula 1 (F1) racing and data center storage. Just as technology is key to winning in F1 racing (technology is used to deliver the fastest road course racing available), technology providing low latency delivered through all flash arrays results in the fastest data center storage.

Let's explain. Winning F1 championships requires achieving the lowest lap times in a series of races on different courses and roads (workloads). Achieving better results (performance) requires improving consistency since championships are awarded to the driver maintaining low lap times across an entire season.

Participating in a lot of races (IOPS) is important but completing races before others (latency) matters most. What doesn't help at all is winning one or a few races (low latency) and then losing all others by a large margin (high latency). Latency is the key storage performance metric and low latency is vital for all flash storage environments.

"There is an overwhelming emphasis on IOPS (Input/Output Operations per Second) and the impact of latency is conspicuously ignored."

— *"Why Low Latency Matters" by George Crump, President and Founder at Storage Switzerland*

LOW PIT TIMES MATTER TOO

Most of us have not driven a F1 racecar. But let's give you another example. You can perform your own highly personalized, real-world, IOPS versus latency demonstration, at your favorite retail store. All you need is a new chipped credit or debit card. Chances are you will experience the problems that can happen when all flash storage decisions are based on maximum IOPS rather than consistent low latency.

Racing through your favorite store doesn't feel like winning when you have to make a long pit stop during check out. All of us are experiencing poor performance due to high latency when using chipped cards. It does not matter if chipped cards

facilitated more transactions (higher IOPS). Every transaction is now slower (higher latency) and transaction times differ at stores (consistency). Chipped cards are eroding customer experiences, negatively impacting consumers and retailers.

CONSISTENCY DRIVES BETTER OUTCOMES

If your "day job" involves ensuring companies have positive and productive customer experiences, one of the best ways to achieve this outcome is to surpass expectations—consistently. No one wants unpredictable or mixed results, whether we're talking about F1 racing, storage latency, or customer experiences.

Understanding all flash performance used to be straightforward when the comparison was against storage with hard disk drives. All flash storage alternatives outperformed traditional storage systems in obvious ways. A focus on IOPS performance—an inherent limitation of hard disk drives—amplified differences. Even the largest and most slow-to-change storage providers now agree the all flash storage era has arrived, and comparing all flash storage against all flash storage is more complicated.

The primary driver of business value from all flash storage is not how many transactions can be completed at the same time (IOPS), it's how consistently each and every individual transaction can be rapidly completed (latency). In other words, databases, applications, and users want the next piece of data they need ASAP (low latency), and they want this to happen often (consistency), rather than different pieces of data (high IOPS) they need later.

The next time you have some quality time to invest online, check out Storage Switzerland and see what George and his colleagues have to say about flash storage. In the meantime, here are a few more quotes from George's "Why Low Latency Matters" article to help you win more races.

"In a world that demands 'instant gratification,' forcing a customer, prospect or employee to wait for a response is the kiss of death."

"For most data centers the number one cause of these "waits" is the data storage infrastructure, and improving storage performance is a top priority for many CIOs."

"The traditional three-tier infrastructure of servers, network, and compute benefits by having storage systems that directly respond and service existing I/O requests faster and thus have the capability of supporting significantly more applications and workloads on the same platforms."

— "Why Low Latency Matters" by George Crump, President and Founder at Storage Switzerland

3

STREET CIRCUIT VS. ROAD CIRCUIT – LOWERING LATENCY DRIVES WORKLOAD MIXING

The diverse teams of businesses can resemble the challenges businesses face when managing multiple and mixed workloads within their data centers. Just as a Formula 1 driver must race multiple circuits with different mixes of straights and turns using the same car, a company's IT infrastructure must enable mixed departments with multiple projects to complete tasks. Substitute workloads for departments, IOPS for projects, and latency for tasks and the connection becomes clear.

The key reason why companies are deploying all flash enterprise storage is the expectation that their preferred solution will run multiple and mixed workloads simultaneously. Examples include big data analytics, online transaction processing, databases, applications, server virtualization, and private cloud.

Choosing an all flash storage solution that favors higher IOPS at the expense of lower latency will lead to unpleasant surprises (I can't mix different workloads?) and deliver disappointing outcomes (I can't consolidate multiple workloads?). Unfortunately, this happens far too often.

LOW LATENCY WINS ON RACE TRACKS AND IN DATA CENTERS

The previous chapter used the analogy of auto racing to describe why consistent performance wins in Formula 1 racing and all flash storage. Let's build on the analogy to include the importance of lower latency when mixing many workloads of different types on the same all flash storage system.

Formula 1 racing is a mixed workloads environment where drivers race on different courses under diverse racing conditions during a season. The races occur on different types of road courses including close city streets, combinations of public roads and permanent track, and permanent racing facilities.

Formula 1 racing is also a multiple workloads environment with many drivers competing to cross the finish line before everyone else during each race. Each driver is one of many workloads on the course with individual drivers needing to run their race with minimal interference from other drivers.

The bottom line is this: low latency wins championships in environments with mixed and multiple workloads, whether they are Formula 1 races or enterprise data centers.

REDUCING LATENCY BEATS INCREASING IOPS

The confusion between the benefits of lower latency versus higher IOPS is understandable, especially with all flash storage. Databases, applications, and users all need their tasks to complete sooner (low latency), regardless of how many other things are happening at the same time (high IOPS), but this is seldom the experience in enterprise data centers. Ultimately, doubling performance involves reducing latency by half rather than doubling IOPS.

Things become clear when considering how latency and IOPS actually affect all flash storage:

- **Lowering latency allows each storage operation to complete sooner.**
- **Increasing IOPS allows more storage operations to work at once.**

Since many all flash storage systems can be configured to provide more IOPS than enterprise data centers need, IOPS specifications are no longer an effective predictor of real world performance. Let's revisit the Formula 1 analogy to explore why.

Formula 1 teams focus on the amount of time their drivers require to complete races because finishing sooner wins championships. In other words, the entire team strives to minimize their driver's race times (lower latency). Racing more cars (higher IOPS) doesn't help any individual driver win more races, but it does increase everyone's race times (higher latency) due to more traffic.

"The goal is to remove performance off the table as an issue by creating architectures with enough head room so that the storage system will never be the bottleneck.

We will get to a point where you won't have to worry about storage performance testing because the storage systems will be such speed demons that it will handle anything you throw at it."

— *"Storage Performance: Important Things to Consider" by Tony Asaro, The INI Group, LLC*

4

PERFORMANCE IS BOTH SPEED AND LATENCY

When you are looking for enterprise storage, performance is what you really care about. Reliability is important, redundancy is important, compatibility is important. A lot of things are important, but nothing is as important as performance. The only reason enterprise storage exists is to house data and then to deliver that data to applications as quickly as possible.

We can all agree that performance is the most significant criteria when customers are evaluating and selecting all flash storage, and it is, in fact, the main reason companies are moving from hard disks to flash. So you would think that if performance is that important in selecting storage, then at the very least we should all be able to agree on how to measure performance.

So how do you measure speed? Well, if you are in a racecar you can measure speed with a speedometer as either miles per hour (mph) or kilometers per hour (kph). Still, that does not give you a real measure of how a car will perform in a race. What you need is a true definition of performance, such as how speed is measured which is a function of distance over time.

In business, we are all in a race to win. It is a race to win customers, to deliver products first to market, to process transactions faster, or whatever it is that drives your company's success. To win in this race, success is not just about achieving great results at a given moment, it is about delivering those great results consistently, over time.

"High sustained latency in a mission-critical app can have a nasty compounding effect. A delay in the DB...and the company could well lose thousands of customers and millions of dollars while the delay is happening. Some companies could also face penalties if they cannot meet certain SLAs."

— *"An Explanation of IOPS and Latency"*
by Dimitris Krekoukias, RecoveryMonkey.org

In racing, whatever top speed a given car hits at a particular moment is not really relevant. What matters for that car is how it performed over the entire distance of the race. The same is true for storage. Often we see arrays that list their IOPS, which may just be a peak number that was achieved at a point in time, as the performance metric that they want to be judged by. Like the speed at a given moment for our racecar that is not really a true measure. For the measurement of IOPS to be truly useful, we need to look at it in the context of latency.

INTO THE PIT

Let's go back to our race analogy. In races there are times when every car has to make a pit stop. In Formula 1 racing there is a speed limit of 80 kph in the pit area. So while a car may be going 80 kph when it enters and leaves the pit, the car's average speed over time will actually be much lower. When the car is being refueled, getting new tires, and whatever else needs to be done, then it is not moving. That impacts its total time. If the car is in the pit too long, it can lose the race regardless of its top speed.

The speed of the car going into and out of the pit is the same kind of measurement as IOPS. The time in the pit, that is latency.

It's the same for enterprise storage. The measurement of IOPS, the speed of data going into and out of the interface ports, is good to know but it is only relevant in terms of overall system latency. How much time does it take the array to process the data once the read/ write request is received?

That is why the important metric is not simply IOPS, but IOPS at latency. The more I/Os an array must manage, the bigger the impact on latency. It makes sense that the more I/Os that we ask an array to perform at a given time (that is, the more I/Os per second), the longer it will take that array to process those tasks. So the question becomes, if an application requires 50,000 IOPS, what will the latency be from a given array? What will the latency be at 100,000 IOPS? At 1 million IOPS?

WHY WE AGREE WITH DIMITRIS KREKOUKIAS OF RECOVERYMONKEY.ORG

This is why in an explanation on IOPS and latency, Dimitris Krekoukias writes, "IOPS numbers by themselves are meaningless and should be treated as such. Without additional metrics such as latency, read vs write % and I/O size (to name a few), an IOPS number is useless."

"High sustained latency in a mission-critical app can have a nasty compounding effect – A delay in the DB…and the company could well lose thousands of customers and millions of dollars while the delay is happening. Some companies could also face penalties if they cannot meet certain SLAs."

WILL YOU WIN THE STORAGE PERFORMANCE RACE?

When you look at many all flash arrays, you will often see IOPS listed, but not latency. Why is that? Perhaps, they can reach a high I/O number, though often under ideal conditions that will never happen in the real world. You must ask the real-world question of what is the latency required to achieve this high I/O number? Maybe, they state what the minimum latency is (the lowest latency they could achieve under unrealistic tests), but at what overall throughput? We have seen some vendors quote 500,000 or even a million IOPS, but then find out that this is at 2-5 ms of latency, and sometimes longer.

Performance like that should be totally unacceptable for flash storage. Enterprise customers require both high throughput and consistent low latency. That is why Violin publishes numbers with both IOPS and latency, because if you are buying a racecar, you want to know you are getting one that will put you in the winner's circle.

5

WINNING – INCREASING BUSINESS VALUE WITH LOWER LATENCY ALL FLASH STORAGE

If you've been following our series on the importance of low latency in all flash enterprise storage, you will have read about consistency, multiple and mixed workloads, and performance. The short story is: lower latency–rather than higher IOPS–accelerates databases and applications when using all flash storage; hence, latency has an incredibly relevant story from a business perspective.

The focus of this chapter is the real-world experiences of enterprise businesses that reveal how lower latency has dramatically improved the return on investment (ROI) of all flash storage. But first, let's keep our racing theme going with another Formula 1 (F1) racing analogy.

Formula 1 is the most technologically advanced car racing, just as all flash storage is the most technologically advanced storage. While F1 budgets are closely guarded secrets, Autoweek once estimated that F1 cars cost USD $2.6 Million (CAPEX) each (some all flash storage solutions cost more) and many F1 teams spend more than USD $100 Million (OPEX) a season. This makes achieving a positive ROI vital—and challenging.

POSITIVE ROI WINS RACES WHILE BUSINESS VALUE CREATION WINS CHAMPIONSHIPS

Achieving positive ROI from any storage system is straightforward since winning with ROI involves a one-time event (calculation), just as winning an F1 race involves finishing in first place. Creating business value with all flash storage is as challenging as earning an F1 championship because both involve numerous and complex factors.

Ultimately, combining positive ROI with business value creation is an ideal outcome that businesses can only experience from all flash storage with lower latency. You see, lower latency means the work can be done faster which translates into greater people productivity, higher company morale, increased work quality, greater workplace efficiencies, and lower operating expenses. Lowering latency accelerates response times and we all know that servicing customers faster and enhancing customer experiences creates customer loyalty and increases revenues.

Below are real-world examples of positive ROI with business value creation experienced by enterprise businesses—using Violin Memory all flash storage solutions, of course—and more are available online here.

LOWER LATENCY FACILITATES BUSINESS AQUISITION

Shortly after the Ferrellgas IT team brought their Violin Flash Storage Platform (FSP) solution online, their leadership finalized a deal to buy another company. The acquired company's entire IT operations had been outsourced. By working in concert, Ferrellgas and Violin were able to ensure sufficient capacity to migrate applications and data to the in-house data center and wind down the outsourcing commitment.

"We're an energy company, growing organically and through acquisitions. Operating savings achieved are estimated by Ferrellgas at almost $1 million annually."

— Bill Evans, VP, Information Technology, Ferrellgas

LOWER LATENCY IMPROVES CUSTOMER SATISFACTION

The largest hearing aid company in Germany, KIND, has more than 500 retail outlets across Germany and an additional 100 outlets across the rest of Europe. Tasks, such as stock checking and accounting, had slowed down significantly, directly impacting employee productivity. KIND's decision to choose Violin FSP as their all flash storage solution improved customer satisfaction because of the speed with which their staff can now provide service to consumers at their retail outlets.

"Violin Memory has helped us to become more efficient as a company with flash storage infrastructure that we can rely on. The benefits we have received more than outweigh the cost of flash storage."

— Christian Emmrich, IT Administrator, KIND Hearing

LOWER LATENCY POSITIONS BUSINESS FOR FUTURE

Valley Health System's patient care and staff support goals required the most robust technology solution possible to both ensure optimal performance and meet or exceed their Disaster Recovery/Business Continuity (DR/BC) requirements. Valley chose to work with Violin Memory to design an enterprise-wide, all flash storage environment around the Violin FSP. The overall system now provides the resiliency demanded by the paperless hospital and the recovery stance inherent in HIPAA, HITECH, and ARRA legislation.

"Using Violin All Flash Storage solutions, we are able to reach our growth targets while continuing to provide outstanding service to our community. We are well positioned for the future."

— Eric Carey , Chief Information Officer, Valley Health System

Valley Health System

INDUSTRY EXPERT REVEALS WHY LOWER LATENCY WINS

Wikibon's online article, titled, "The Potential Business Value of Low-Latency Flash" includes detailed economic models showing all flash storage with lower latency reduces the total cost of ownership (TCO). The lowest TCO is delivered by all flash storage providing a 0.3-millisecond latency that matches the lower latency offered by Violin FSP solutions.

"The starting point should be key business database systems, with an emphasis on providing the lowest storage latency."

— "The Potential Business Value of Low-Latency Flash" by David Floyer, CTO at Wikibon

The Expected Business Value of All Flash Storage = Productivity Savings + Net Revenue Gains + Hardware and Infrastructure Savings – Flash Storage TCO. The bottom line is that fast response times is a metric that truly matters in driving success to your business and lower latency is a powerful enabler of this competitive edge and winning the race.

6

MEET THE WINNER IN THE ALL FLASH STORAGE WINNER'S CIRCLE – IOPS ARE FUN, BUT LATENCY WINS

You can **Be Instrumental** as a performance champion to your organization by revealing how lower latency, rather than higher IOPS, wins races in all flash storage for enterprise data centers. Like a driver behind the wheel of a Formula 1 (F1) racecar, you want to win that first race. Once you do that, your goal extends to winning other races and your aspiration becomes how do you become champion of the series.

Winning the series signifies that you and your team—pit crew, ownership, etc. —are the best at competing in tough and complex environments, which in F1 means a variety of races on various streets and roads in varying conditions.

In the data center, IT professionals translate the traits of a F1 driver into meeting performance service level objectives every minute of every day for the entire year.

RECOUNTING PIVOTAL OUTCOMES IN STORAGE PERFORMANCE

The Latency Matters series featured five articles and covered a lot of ground on the importance of lower latency in all flash enterprise storage. Here's a quick recap you and your team can race through:

1. The performance of all flash storage solutions is counterintuitive as lower latency rather than higher IOPS drives database and application performance.
2. The primary driver of business value from all flash storage is not how many transactions can be completed at the same time (IOPS), it's the consistency with which each and every individual transaction can be rapidly completed (latency).
3. Low latency wins championships in storage environments with mixed and multiple workloads.
4. Performance is the most significant criteria when evaluating and selecting all flash storage and IOPS at a specific latency is the key metric.
5. Lower latency translates to higher productivity, revenue gains, and compelling business value.

"Most enterprises have at least several applications where extremely low storage latency can directly translate to increased revenues, better customer experience, or differentiated competitive advantage."

— Eric Burgener, Research Director in Enterprise Storage, IDC

IDC ANALYST AS F1 INDUSTRY'S COLOR COMMENTATOR EQUIVALENT

"Most enterprises have at least several applications where extremely low storage latency can directly translate to increased revenues, better customer experience, or differentiated competitive advantage."

"These applications tend to be transactional, database-driven or real-time analytics workloads where a 500 microsecond latency difference can be directly monetized."

"For these environments, AFA architecture can matter a great deal in an ability to consistently deliver sub 300 microsecond performance that companies can base their reputation on, regardless of widely varying workloads and operational workflows."

– Eric Burgener, Research Director in Enterprise Storage, IDC

THREE KEY TAKEAWAYS ON LATENCY IN ENTERPRISE STORAGE

Increasing IOPS theoretically allows more databases and applications to run on all flash storage, but all can become equally unresponsive due to higher latency.

Doubling database and application performance with all flash storage involves reducing latency by half rather than doubling IOPS.

It's lower latency that matters most with all flash storage, especially in environments with mixed and multiple workloads.

AFTERWORD

There are multiple takeaways from the Violin "Latency Matters" Series.

We now know that IOPS are only one part of the performance story, and not nearly as significant as latency. Low latency is required to deliver fast access to information and paramount to meeting the expectations of the customer experience. IT organizations are tasked with delivering the performance required to run mission critical applications and consistent, low latency is the essential ingredient to meeting the SLAs that drive the business because it has a vital impact on the real-time, instantaneous response time that users demand.

THE STATE OF THE STORAGE INDUSTRY

As primary storage rapidly transitions from hard drives to flash, it's clear "Disk Is Dead". It died because it could not keep up with the performance demands of the modern data center. Flash has alleviated the storage bottleneck and empowered customers to not only run their existing applications faster but flash has also enabled the shift to virtualized infrastructure and cloud delivery.

The state of the storage industry was until now comparing all flash storage to hard disk drive arrays. Any of the all flash storage solutions on the market will provide exponential levels of performance advantage over spinning media but we are now at an inflection point to evaluate whether there is a significant performance advantage of one all flash storage array versus another one. This is where latency becomes a real differentiator between all flash array solutions in the marketplace.

The all flash array solution that will be in the Winner's Circle is the one that can deliver lower latency while providing a robust set of data services that meet the selection criteria of enterprise IT customer environments.

Now more than ever, the real world applications that drive the business need a heightened level of performance with the shift to virtualized infrastructure and cloud delivery. From database transaction systems to big data modeling with cognitive analytics, performance to store data, process it and access it, demands low and consistent latency. In some cases lower latency is about faster decision making while in other cases it is about increasing productivity and getting the job done faster. Let us not forget about the software developers of these applications and the test dev environment where speed translates into cutting the development time for applications, or the speed to bring a new rev of software online, adding value to the business.

WHAT DOES THE FUTURE HOLD FOR LOWER LATENCY

NVMe Over Fabric (NVMf): Flash storage goes a long way toward addressing the insatiable appetite for high performance. NVMe over Fabric is an emerging technology and will play a role in powering the next generation of flash storage. This new shared-storage model will further reduce latency.

The emergence of NVMe over Fabric raises new questions:

- **Will applications need to be modified to take full advantage of this new storage network and the associated new storage arrays?**
- **Will NVMe over Fabric storage arrays provide a full set of data services to meet my Enterprise storage requirements?**
- **Can my current storage vendor protect investments in current Fibre Channel and iSCSI SAN solutions and seamlessly integrate new workloads to the NVMe over Fabric storage arrays?**

Customers should seek solutions that can seamlessly integrate application workloads under NVMe Over Fabric while protecting existing Fibre Channel and iSCSI SAN investments.

Storage Class Memory (SCM): SCM refers to the many types of non-volatile memory, such as 3D XPoint and ReRAM, which will make up the future of solid state enterprise storage. Each of these will offer unique levels of density, high performance, and low latency.

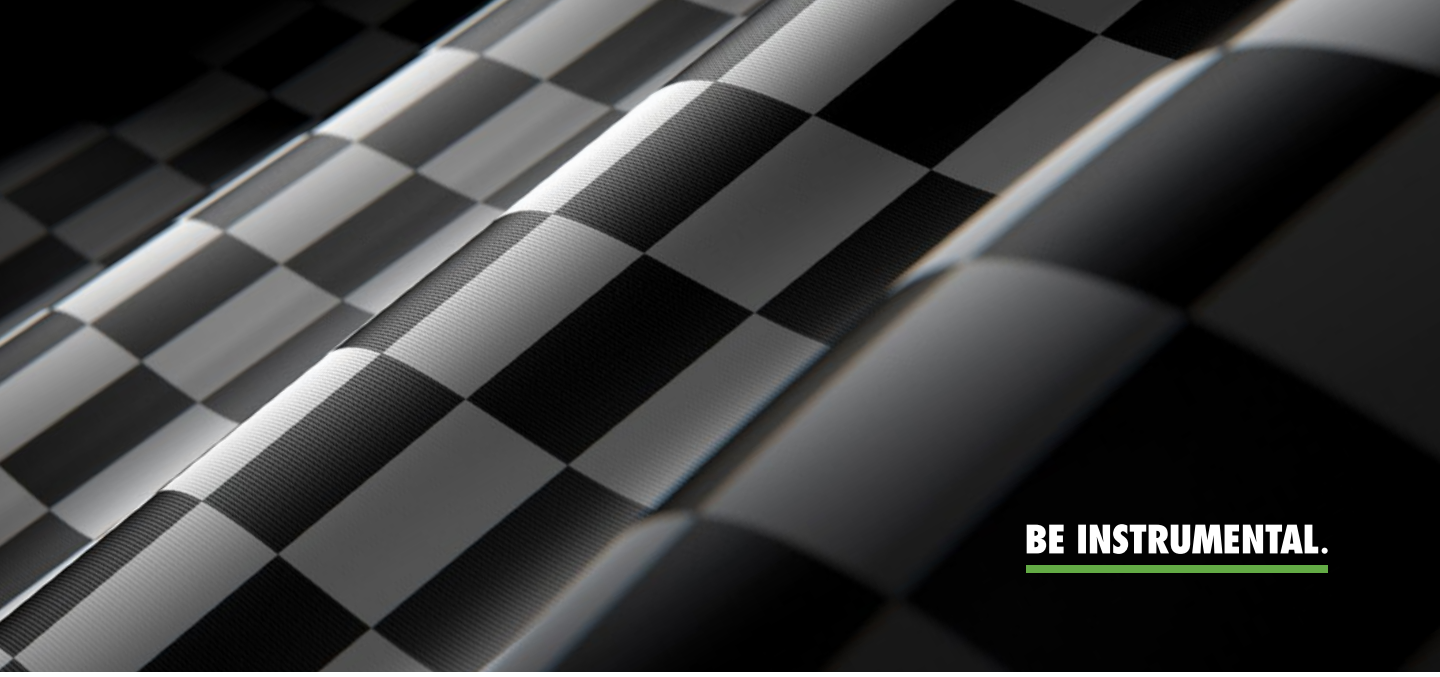
The future of SCM raises new questions:

- **Will my current storage solution support the transition to SCM or will I have to buy a whole new set of storage hardware and software?**
- **Can my storage provider offer a single software stack of data services and a single pain of glass (SPOG) to manage SCM storage arrays along with current investments with all flash storage arrays?**
- **Will it be affordable to move to SCM storage arrays and can I justify the CapEx and OpEx investment?**

CONCLUSION

It is reasonable to expect that in a few years we will move on to ubiquitous storage that is always at lightning speed and never even think about performance again. But for now, customer environments will continue to wrestle with how to deliver IT Services cost-effectively and at the highest levels of performance. While this challenge exists, "latency matters" and will be the key criteria for evaluating and selecting the all flash storage solution of choice.


BE INSTRUMENTAL.



APPENDIX B
ADDITIONAL RESOURCES

George Crump
"Why Low Latency Matters"
Article on *Storage Switzerland*
https://storageswiss.com/2016/02/02/why-low-latency-matters

Dimitris Krekoukias
"An Explanation of IOPS and Latency"
Article on *Recovery Monkey*
http://recoverymonkey.org/2012/07/26/an-explanation-of-iops-and-latency/

Tony Asaro
"Storage Performance: Important Things To Consider"
Article on *Contemplating IT*
http://www.contemplatingit.com/blog1.php/2014/03/12/storage-performance-important-things-to

David Floyer
"The Potential Business Value of Low-latency Flash"
Article on Wikibon
http://wikibon.com/the-potential-business-value-of-low-latency-flash/

BE INSTRUMENTAL.

Violin® MEMORY

www.violin-memory.com/latency-matters