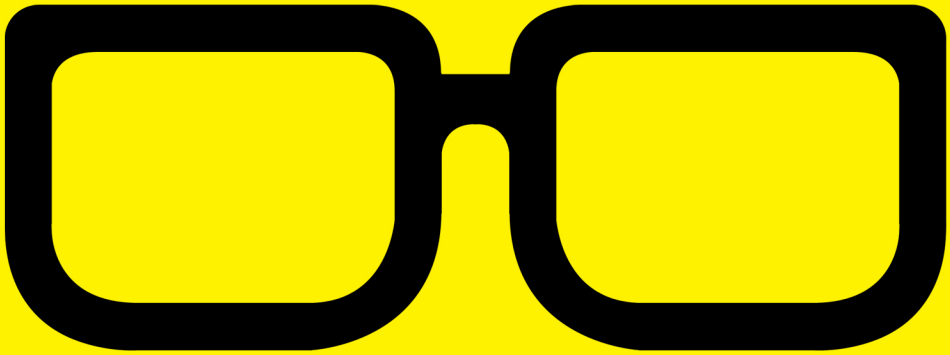
SPONSORED BY

SILICON MECHANICS
SERVERS • STORAGE • HPC

&

intel®

GEEK GUIDE



Get in the Fast Lane with NVMe

LINUX™ JOURNAL

Table of Contents

MIKE DIEHL has been using Linux since the days when Slackware came on 14 5.25" floppy disks and installed kernel version 0.83. He has built and managed several servers configured with either hardware or software RAID storage under Linux, and he has hands-on experience with both the VMware and KVM virtual machine architectures. Mike has written numerous articles for *Linux Journal* on a broad range of subjects, and he has a Bachelor's degree in Mathematics with a minor in Computer Science. He lives in Blythewood, South Carolina, with his wife and four sons.

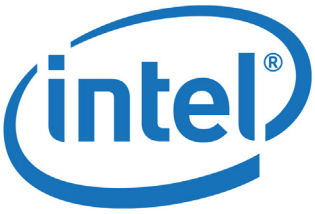
GEEK GUIDES:
Mission-critical information for the most technical people on the planet.

About the Sponsor

Silicon Mechanics, Inc., is a leading provider of servers, storage and high-performance computing technologies to the world's most innovative organizations. Since 2001 Silicon Mechanics has supported customers with its "Expert included." approach, reflecting the company's passion for providing complete customer satisfaction and customer confidence in the return on their technology investments. Recognized as one of the fastest growing companies in the greater Seattle technology corridor, Silicon Mechanics is changing the way systems providers engage with customers.

To learn more about our NVMe-optimized offerings, visit siliconmechanics.com/nvme.

To configure your custom system with one of our experts, call us at 866-352-1173.



Intel Corporation manufactures industry-leading solid-state drives targeting data center, professional and consumer products. Intel SSDs feature leading performance, high levels of reliability, and are manufactured by one of the most trusted brands in the world.

To learn more about the Intel SSD Data Center Family for NVMe, visit intel.com/ssd.

Get in the Fast Lane with NVMe

MIKE DIEHL

Speed Is Everything

The automotive industry spends millions of dollars each year racing cars because it helps build better cars for regular people like you and me. If an auto manufacturer can build a V-8 engine with four valves per cylinder, take it to the race track and have a successful season, the same manufacturer can use that knowledge and experience to build an in-line four-cylinder engine that's powerful, reliable and relatively inexpensive for consumers.

In the IT world, however, speed is everything. People don't buy slow servers, and they don't want their websites

When compared to conventional spinning disks, SSDs are like taking a horse-drawn carriage and strapping an internal combustion engine to it; it's a completely different experience.

to load more slowly than the rest. So in this regard, IT is much like auto racing. You build faster servers and faster networks because it allows you to provide better service to your customers.

Fortunately, you don't have to go to the track to learn to build high-performance servers. Server technology is constantly changing, and it's constantly getting faster and cheaper. The challenge is to keep up with those changes.

Through the years, CPUs have accelerated, and when they couldn't speed up any further, manufacturers started putting more of them on a single chip, and then more chips on a single motherboard. Now computers are blindingly fast, fueled by amazingly dense memory systems.

Conventional spinning hard drives are about as quick as they can be; however, they have been getting bigger and cheaper as well. But, then something happened. Solid-state drives (SSDs) started to hit the market. When compared to conventional spinning disks, SSDs are like taking a horse-drawn carriage and strapping an internal combustion engine to it; it's a completely different experience. Many industries gain a substantial market advantage just by having faster servers and mass storage systems.

Who Needs Fast Servers and Mass Storage Systems?

HPC: In this case, the name says it all—High-Performance Computing. HPC includes large-scale simulations, graphics render farms and other research-oriented organizations. These organizations tend to have a huge number of CPUs and GPUs, and they manipulate large quantities of data. HPC organizations understand how to add CPUs to their enterprises, but eventually storage becomes a bottleneck. At that point, the system administrator has a choice: add complexity or add speed. Complexity comes in the form of segmented storage and file synchronization schemes. Speed is easier if you can find faster storage. As you're about to see, you can.

The Cloud: Cloud and hosting providers win big by being able to build efficient, scalable servers. Chassis-count has an intrinsic, recurring cost. Most data centers charge for server rack space, and some charge for power. Then there is the recurring cost of support and maintenance. All of this, of course, is after you buy and configure the server in the first place. If each bare-metal server in the enterprise could be made to operate more efficiently, it could support more virtual machines and, thus, more customers, before additional servers would be needed. In the meantime, the customer experiences better performance from its dedicated servers, virtual servers or shared hosting services.

Virtualization: Many organizations have opted to virtualize their in-house servers because it makes them easier to manage. Virtualization allows the server manager to run multiple virtual servers on a single bare-metal server. This reduction in chassis count brings all of the benefits mentioned earlier for cloud services, but it also makes it economically feasible to maintain

“Big data” isn’t any good if it’s “slow data”.

spare servers, as not as many bare-metal servers are needed for production operation. Consolidating virtual services onto a smaller number of bare-metal servers does place a much larger burden on the memory and disk drives though. SSDs are powerful enough to withstand this extra burden.

Big Data: Financial, health care, insurance and other “big data” or large database-driven sectors realize obvious benefits from increased data throughput. There are insurance underwriters that have data centers with petabytes of data on-line. Being faster, to them, means being able to process more claims each day, which means larger contracts, which means more money. “Big data” isn’t any good if it’s “slow data”.

Smaller Data: Big data isn’t the only data-driven market sector that would benefit from having more powerful servers. Any process or service that uses a database is going to perform better by using a faster back-end storage engine. A blog site, for example, is more database-intensive than you might think. Also, consider that e-commerce and point-of-sale systems almost always have databases behind them. Finally, calendaring and scheduling systems often take a beating from many users at once and certainly would benefit from migrating from spinning disks to SSD.

Almost all server managers buy the fastest servers they can afford and then tune them to get as much out of them as possible. But eventually, they hit a bottleneck. Invariably, that bottleneck is the hard drive spindle.

HDDs vs. SSDs

HDDs: Hard drives have come a long way through the years—from the old ST-506 standard to IDE and SCSI. In the case of IDE, the industry has moved from parallel ATA (PATA) to serial ATA (SATA). SCSI eventually was improved upon with the introduction of serially attached SCSI. While older drives often rotated at 3,600 RPM, newer drives spin as fast as 15,000 RPM. But, conventional hard drive speeds have plateaued; the spindle will turn only so fast.

One of a tape drive's intrinsic limitations is that it can only read and write serially. Sure, you can position the read/write heads to any location on the media and perform I/O, but that positioning operation takes time. Even worse, the time it takes to position the read/write heads is proportional to how far away the data is on the media. And even in the best case, this positioning operation takes eons when compared to the time it takes for a CPU to request and process the data. The hypothetical worst-case scenario is, of course, reading a file backward against the direction in which the platters are spinning, because each read request has to wait for the spindle to make a complete revolution before the next data can be read. Modern spinning disks mitigate this performance bottleneck by having huge onboard caches and by implementing read-ahead algorithms.

SSDs: SSDs support completely random access. You can access data on an SSD almost instantly, no matter where it is on the media. The head positioning bottleneck is completely gone with SSD. An SSD can find data as fast as the CPU can ask for it. But then you discover another bottleneck, which I discuss shortly.

As I'm sure you know, there are many forms of racing. There's NASCAR, drag racing, Formula-1 and off-road racing,

In the quest to build a faster server, your first task is to replace spinning disks with SSDs. Then you should consider whether the old-style disk controller is well suited to controlling this new style of drive.

just to name a few. It turns out that the things that make a car go fast on one type of track, often aren't very effective on another type of track.

For example, a dragster goes only in a straight line on a smooth surface. And obviously, dragsters have adapted to this particular style of racing. A Formula-1 car, on the other hand, has a light, aerodynamic body that helps keep it on the ground while maneuvering around neck-wrenching corners.

These two completely different forms of racing require completely different cars. A dragster, while being able to muster a 300 mph top speed, isn't be able to make the sharp turns that a Formula-1 car is expected to make. And, the Formula-1 car would be about 100 mph too slow to be competitive at the drag strip.

Race car manufacturers take into consideration the specific requirements of their particular style of racing when they build their cars. It only makes sense that a hard drive controller would be tailor-made for the type of drive that it controls. An IDE controller that was designed to control a spinning disk just isn't well adapted to controlling an SSD. In the quest to build a faster server,

your first task is to replace spinning disks with SSDs. Then you should consider whether the old-style disk controller is well suited to controlling this new style of drive. Let's think about what that means.

A Look at SATA and SAS

The old PATA/SATA drive controllers were designed with specific requirements in mind. First, they were made with the assumption that they would be controlling a single spindle with a stack of platters on it. In order to boost performance, these ATA controllers come equipped with onboard cache. The controller is managing a single spindle full of platters, so only one command queue is needed. But, because the CPU is putting commands on the queue and the controller is taking commands off of it, a locking mechanism is designed into the system and this adds overhead.

Many of these controllers were built around the ISA and PCI buses, and they inherit some of their design limitations from those buses. Sure, PCIe SATA disk controllers exist, but they certainly don't take advantage of all that the PCIe bus can provide, as you'll see.

The first-generation SSDs didn't use SATA because it was particularly well suited to the job. SATA was used because it was available. SATA is well understood and has name recognition with server managers. And most importantly, SATA drivers were readily available for almost every operating system and system BIOS out there. None of these reasons are good reasons to stick with an older technology.

Just like you wouldn't put a Volkswagen engine in a Ferrari, it doesn't make sense to strap the latest SSD hard drive to an older SATA drive controller if there are better alternatives.

Serial-attached SCSI (SAS) was designed to replace SATA and to meet the speed demands of the modern server. SAS has a richer command set than SATA, and with a 12Gb/s throughput capability, it has twice the throughput as SATA. But once again, it's geared toward a spinning, magnetic platter, and NVMe is still 3x faster.

NVMe Performs Better

An NVMe controller is a different beast altogether. First and foremost, an NVMe controller can have up to six times as much throughput as a SATA controller—six times is fast, indeed!

Just like you wouldn't put a Volkswagen engine in a Ferrari, it doesn't make sense to strap the latest SSD hard drive to an older SATA drive controller if there are better alternatives. So, in the quest to improve server performance, the next step is to forgo the older drive controller technology and step up to the next-generation NVMe controller designed specifically for the PCIx bus and SSD drives.

So, why do NVMe controllers perform so much better? Well, let's kick the tires a bit and find out.

NVMe Highlights

- Designed for NV RAM from the start.
- Open standard.
- Six times the throughput as SATA.
- Half the latency as SATA.
- Ready for next-generation SSDs.
- Half the CPU overhead per IOP than SATA.

Throughput: In the racing world, there's an old saying, "there's no substitute for cubes". What this means is that, everything else being equal, the car with the biggest engine (in cubic inches) is going to be the fastest. In the server world, there's no substitute for throughput. As I already mentioned, NVMe has up to six times as much throughput as SATA. In fact, NVMe can take advantage of the inherent parallel nature of both SSD and PCIe. The PCIe bus has four data pathways, and because SSD is completely solid state, it's more than happy to travel down all four highways at the same time.

Latency: As if being capable of six times as much throughput as SATA isn't enough, NVMe also has half the latency. NVMe doesn't have as many disk control instructions as SATA, and it certainly doesn't have the rich instruction set that SAS implements. The instructions that NVMe has have been optimized for controlling SSDs and only SSDs. This goes

back to my analogy of adapting a race car for a particular type of racing. In this case, the disk control instruction set has been adapted to control a particular type of disk and do it as efficiently as possible. The results are a much more responsive disk and controller pair.

Command Queues: Obviously, it doesn't do any good to have the fastest car on the track if you don't have enough fuel to finish the race. Any high-throughput device is going to place at least some burden on the host CPU. Data transfers have to be set up and torn down by the CPU after all. Commands have to be put on the command queue by the CPU to be processed by the device as well. In the case of a SATA disk controller, there is only one command queue, and it has to be locked before additional commands can be put on it by the CPU or taken off of it by the controller. This locking requirement introduces the possibility of lock contention, where the CPU has to wait to put new instructions on the command queue while the controller takes commands off of it and vise versa. All of this takes time and slows things down.

An NVMe controller has 64K command queues, each of which can queue up 64K commands and doesn't require any locking. So, the CPU is free to put commands on any queue at any time without any kind of delay. Having multiple command queues allows the operating system to spread commands across multiple queues in order to prevent any one process from consuming all of the available I/O. Because of NVMe's highly tuned command set, it actually requires fewer CPU cycles per I/O operation (IOP) than any of the other controller technologies I've

discussed. So, not only is the controller quicker and more responsive, it's more CPU-efficient as well. This means a given server will have a longer usable lifespan before it needs to be upgraded because the CPU can't keep up with the demands placed on it.

A Word on PCIe

Native PCIe SSD storage systems are available that are not based on NVMe. And sure, they perform well, because they can take advantage of the PCIe bus just like NVMe does. But, they are proprietary, vendor-specific devices. In an age of open standards, proprietary systems tend to die out. NVMe, on the other hand, is an open standard backed by a consortium of about 80 different companies. You can find additional information on NVMe, such as benchmarks and technical specifications, at http://www.nvmexpress.org.

Future-Proof Your Investment

Whereas a race car only races for a single season before it's replaced by a faster car, you want your servers to provide high-performance service for as long as possible. Because you want to future-proof your server investment, it doesn't make sense to shackle yourself to closed-standards technology. And since speed is the name of the game, it is making less and less sense to stay in the slow lane with SATA or SAS.

NVMe is the standard of the future. It's been designed from the ground up to be able to keep up with the SSDs of today and tomorrow. And since NVMe is an

open standard, it's only going to become more available as time goes by. As is almost always the case, the proprietary offerings will soon be forgotten in favor of more standardized products. You can expect the SATA-based SSDs to become less popular as server manufacturers and integrators move to NVMe, because no vendor wants second-place servers.

NVMe isn't just some future standard; it's here now. Intel has NVMe-based SSDs on the market at the time of this writing. Its DC P3500 SSD drives come in 400GB, 1.2TB and 2TB capacities. The P3500 is rated for .3 drive writes per day (DWPD) and a Mean Time Before Failure (MTBF) of 2 million hours. The P3600 adds 800GB and 1.6TB capacities and is rated for three DWPD. Finally, the P3700 is available in 400GB & 800GB (10 DWPD) and 1.6TB & 2TB (17 DWPD) capacities. For more information on these drives, visit http://www.intel.com/content/www/us/en/solid-state-drives/intel-ssd-dc-family-for-pcie.html.

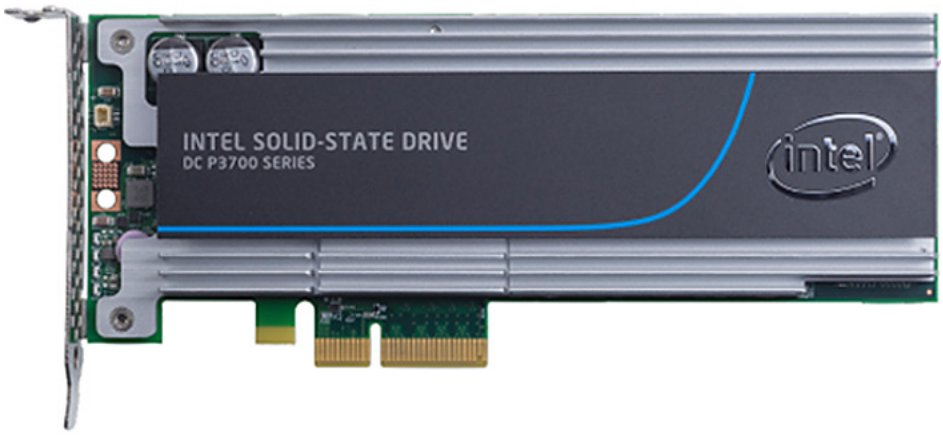


FIGURE 1. NVMe SSDs, such as the Intel DC P3700, are available now.

TABLE 1. Intel's Data-Center NVMe SSD Family

	DC P3500	DC P3600	DC P3700
Disk writes per day	0.3	3	10, 17
400GB	Yes	Yes	Yes
800GB	—	Yes	Yes
1.2TB	Yes	Yes	—
1.6TB	—	Yes	Yes
2TB	Yes	Yes	Yes

Systems integrators are making NVMe convenient to deploy by making it available as a factory-installed option on new servers, such as the Rackform R335v5 from Silicon Mechanics. For more information about this and other NVMe-ready servers, visit http://siliconmechanics.com/nvme.

FIGURE 2. The Rackform R335v5 from Silicon Mechanics

There are already out-of-the box drivers for Linux, Windows and Mac, so you're not going to find yourself downloading drivers just to get your mass storage system recognized by your operating system of choice. These drivers are mature and vendor-supported.

The Finish Line

Throughout this ebook, I've argued that speed is the primary

motivation for upgrading to NVMe-based SSDs, but that's only part of the picture. Earlier, I said that race car drivers buy the fastest cars they can, while server managers buy the fastest servers they can afford. Although speed is important, it doesn't do any good if you can't afford it. There is a 40–50% cost premium over SATA SSDs, and there is a 25x $/GB between NVMe and a spinning disk. But, when you consider the fact that the raw throughput of NVMe-based SSDs is six times that of SATA-based SSDs, you begin to see that NVMe has the potential to bring down the total cost of ownership (TCO) four-fold. And of course, that's only evaluating throughput. The NVMe drives won't tax the host CPU as much as the SATA drives would, so there are serious cost savings to realize over time there as well.

So sure, your next server may be a bit more expensive than your last server was in up-front dollars. But, because of NVMe's higher performance and better CPU efficiency, you'll be able to own that server for a much longer time before your needs outgrow it. You'll have more time to amortize that cost premium, but you also will have put off the costs of the next server upgrade in terms of dollars, labor and downtime. Nobody likes to be on the upgrade treadmill, and anything that a system administrator can do to slow that cycle down is a good thing.

It's pretty clear that almost every server manager feels the pressure to build ever-faster servers while staying under budget. The low-hanging fruit in this effort is to migrate from spinning disks to SSD if you haven't already. While migrating to SSD, it's important to invest in hardware standards that future-proof your investment and provide you with as much bang for your buck as possible. Migrating to NVMe-based SSDs is a cost-effective way to get more horsepower out of your next upgrade.■