IT@Intel White Paper

Intel IT
IT Best Practices
Solid-State Drives
July 2012

Accelerating Data Center Workloads with Solid-State Drives

In performance trials, Intel IT determined that some of the latest Intel® Solid-State Drives provide significant cost, reliability, performance, and endurance advantages over traditional enterprise-quality hard disk drives.

Christian Black
Enterprise Architect, Intel IT

Darrin Chen
Senior Systems Engineer, Intel IT

Executive Overview

Interest continues to grow for the use of solid-state drives (SSDs) in data centers, particularly for high random I/O applications where SSDs excel. The greatest obstacles to widespread adoption are cost and concerns about SSD endurance, specifically their ability to withstand large amounts of data writes. In performance trials in our test environment and with our security-compliance application database, Intel IT determined that the latest SSDs provide significant cost, reliability, performance, and endurance advantages over traditional enterprise class hard disk drives (HDDs).

The impetus for our research was the observation that 15K revolutions per minute HDDs presented a serious performance bottleneck to the 100-percent random workload generated by our security-compliance application database. The amount of HDD head travel required for the patching and reporting functions of this database slowed performance to an unacceptable level. Our goal was to find a solution to speed up the database without incurring excessive costs or increasing complexity.

Our evaluation found that switching to the latest SSDs can:

- Eliminate storage performance bottlenecks, increasing disk performance up to 5x on random disk I/O tasks.
- Reduce read latency by up to 10x, write latency by up to 7x, and maximum latency by up to 8x, for faster response to patching and compliance data read/write requests.
- Provide faster transition from idle to active state, plus display no I/O penalties (longer seek times) as drives fill to capacity.
- Justify higher initial costs through reduction of IT time spent dealing with the effects of maximum disk queue depths, drive endurance equivalent to a 25-year lifespan, elimination of compliance issues involving backlogs in recording monitoring data, and reduction of potential losses from delays in patching monitored systems.
- Lower disk power demands by more than 50 percent while producing one-third less heat.

In this paper, we discuss our test methodology, trial results, and actual results when deployed in our data center. We also provide guidance on how to determine whether a particular application is a good fit for running on SSDs.

Based on our results, Intel IT is now researching additional SSD use cases, particularly for known high random I/O workloads.

Contents

Executive Overview 1

Background 2

- Comparing the Causes of Hard Drive Failure 2
- Recent Improvements in SSD Endurance 3

Solution 3

- Test Methodology 3
- Measuring the Existing Workload 4
- Modeling the Workload in Iometer 5

Test Results 6

- Validating Model Workload Against Real Workload 6
- Determining Top Line Performance 7
- Calculating Endurance 7
- Calculating the Uncorrectable Bit Error Rate 8
- Power and Heat Comparisons 9

Production Environment Results 10

Conclusion 11

Acronyms 12

IT@INTEL

The IT@Intel program connects IT professionals around the world with their peers inside our organization – sharing lessons learned, methods and strategies. Our goal is simple: Share Intel IT best practices that create business value and make IT a competitive advantage. Visit us today at www.intel.com/IT or contact your local Intel representative if you'd like to learn more.

BACKGROUND

As interest continues to grow for using solid-state drives (SSDs) in data centers, there is also increasing focus on SSDs' advantages and write endurance—the ability of an SSD to withstand large amounts of data writes. In the past, endurance has been a potential concern, but recent improvement in SSD endurance and other compelling SSD benefits now seem to justify their greater use.

At Intel, as in many other organizations, data in servers is often stored on hard disk drives (HDDs), which write to and read from magnetic disks. SSDs, on the other hand, use semiconductor-based memory to store data. This memory is usually NAND (short for "NOT AND") flash memory, the same storage medium used for USB thumb drives. NAND memory is ideal because unlike RAM, it is non-volatile and data is not lost when the device is powered down.

While the form factor and interface of most SSDs are compatible with HDDs, SSDs have no moving parts. With no moving platters or an actuator arm to read and write data, there is nothing mechanical in an SSD to wear out. This provides a number of advantages.

- **High reliability.** SSDs have a mean time between failure of 2 million hours.
- **Fast access.** With SSDs, there is no waiting for the drive to come up to speed from idle or perform head seek operations.
- **Excellent resistance to impact and vibration.** SSDs can withstand shock and vibration while maintaining data integrity.
- **Low power consumption.** SSDs consume over 50 percent less power compared to an HDD.
- **Lower heat generation.** Systems with SSDS have less heat dissipation.
- **Silent operation.** SSDs have no moving parts to make noise.

A less well-known advantage of SSDs comes in running database applications that rely heavily on I/O operations per second (IOPS) to determine performance. Because SSDs incur no head seeking to read or write data, they deliver higher IOPS and lower access times than enterprise—15K revolutions per minute (RPM)—HDDs.

One disadvantage often cited for SSDs is cost. A serial-attached SCSI (SAS) SSD can cost up to three times more than a traditional 15K RPM SAS HDD of equivalent capacity. SSD prices continue to improve though, especially as the cost of NAND flash memory drops and production volumes increase.

Comparing the Causes of Hard Drive Failure

The main source of failure for HDDs is mechanical. The most frequent failure is a head crash where the actuator arm physically contacts the disk platter causing data loss. Over time, other components simply wear out: platters vibrate due to bearing wear, actuators lose precision, and lubricants evaporate. The result is more retries, more corrupted data requiring error correction code (ECC) recovery, higher drive temperatures, greater power draw, and eventually failure. The magnetic media itself has virtually no limit on the number of writes, but the magnetic bit strength has a half-life of just five to seven years. Nonetheless, a HDD will most likely fail for mechanical reasons long before magnetic bit strength has any appreciable effect on performance or data integrity.

SSDs, in comparison, can occasionally fail because of data retention or read-error issues on individual NAND flash cells. Each block of a flash-based SSD provides one million to five million write cycles—granted, a very large number—before it fails. This is known as *write endurance*. These errors occur only after reaching the maximum number of program/erase (P/E) cycles for any individual block on an SSD. For many use cases, this is hardly an issue.

Recent Improvements in SSD Endurance

The endurance of an SSD is dependent on its overall capacity and the amount of P/E cycles its NAND flash cells can support. When a host issues a write command to an SSD, the SSD data management scheme may consume multiple P/E cycles simply performing this command on individual NAND flash cells. The ratio of NAND writes to host writes during this operation is known as *write amplification*. When writing 100 gigabytes (GB) to an SSD, for example, NAND flash cells may be written two times, resulting in 200 GB of NAND flash cell writes. This amounts to a write amplification of 2 (200 GB/100 GB = 2).

For a number of years, SSD capacity increased through a technology known as *multi-level cell* (MLC) *NAND*. MLC NAND uses multiple charge levels per cell to allow more bits to be stored using the same number of cells. While suitable for client applications, MLC NAND P/E cycle limits may be insufficient to meet the endurance needs of data center applications. An SSD targeted for less rigorous applications may try to overcome these issues by using an onboard ECC engine. However, once beyond a certain level of P/E cycles, as these SSDs reach the end of their functional life they can fail to recover data.

High Endurance Technology (HET), a new Intel solution designed for data center environments and other endurance-focused applications, extends the endurance of SSDs by using endurance-validated MLC NAND. HET's silicon-level and system-level optimizations enable it to use beyond-ECC error recovery steps to extend MLC NAND capability to a higher P/E cycle count and employ special programming sequences to mitigate program state disturb issues that may occur. Though a scheme called *background data refresh*, an SSD with HET moves data around during periods of inactivity to re-allocate areas that have incurred heavy reads. Additionally, an SSD with HET comes with a spare area that lowers write amplification. The combined effect of these items enables an SSD with HET to deliver the endurance and data retention necessary for many data center applications.

There is a trade-off though. A standard MLC NAND SSD can retain data for 12 months without power. An MLC NAND SSD with HET can retain data for only three months without power. However, because an MLC NAND SSD with HET provides up to 30x the number of write cycles compared to a standard MLC NAND SSD, this trade-off actually is quite favorable for enterprise workloads where the power is always on.

SOLUTION

Intel IT needed a new drive solution for a security-compliance database that had reached the point where incoming data consistently exceeded the write capacity of the HDD array. The unwieldy I/O queue depths were forcing staff to spend precious time manually throttling the security patching and compliance data, which created recording backlogs. To handle the high random I/O demands of this database, we decided to take advantage of the faster read/write speeds of SSDs for this kind of data.

For our tests, we selected high endurance HET-based SSDs—the Intel® Solid-State Drive (Intel® SSD) 710 series. Our goal was to first test the performance of these SSDs in a controlled environment, and then, based on the data, measure their performance while running the production security-compliance database.

Test Methodology

To determine the viability and advantages of using SSDs for a high I/O application such as our security-compliance database, we chose the following methodology:

1. Measure the existing database server workload using Perfmon, which is a performance monitoring tool.
2. Model the workload in Iometer, which is an open source I/O load generation, measurement, and characterization tool.
3. Using production-identical lab hardware, validate that the model workload accurately represents the real workload.
4. Test multiple configurations and controller settings of both 15K HDDs and SSDs in an eight-drive redundant array of independent disks (RAID) sets using both RAID 5 and RAID 10 settings.
5. Determine top-line performance for both the 15K HDDs and the SSDs.
6. Calculate endurance—the projected useful lifetime of an SSD—with the modeled workload.
7. Calculate the uncorrectable bit error rate (UBER) by determining the probability of encountering an uncorrectable error.
8. If the solution looks promising, conduct a test on an actual security-compliance database.

Manufacturer Support Agreements

The solid-state drives (SSDs) we used in our testing and then brought online in our data center are not currently supported by the server manufacturer supplying our systems. For most IT departments, including Intel IT, manufacturer support agreements covering the entire system are a requirement for production data center servers. To retain support for every part of the server except the SSD drives, we alerted the manufacturer of our intention to use SSDs that were validated internally for some of the disk arrays in the server. We also said we would continue using the manufacturer-supplied and validated HDDs for the system's boot/OS drive. This resulted in a support agreement that excluded the SSDs and provided coverage of all the other usual system components. Leaving the boot/OS drives as manufacturer supplied also allows support staff to run OS-based diagnostics on the system if necessary.

Measuring the Existing Workload

In measuring the existing workload with Perfmon, our primary interest was queue depth, which is the number of outstanding I/Os queued in the disk controller. Figure 1 shows the queue depth (in grey) for the eight-disk RAID 10 F:\ array on which the security-compliance database runs. A standard rule of thumb is that for every disk in an array the queue depth should be no more than one or two. With an eight-disk F:\ array, the queue depth then should be no more than 16 for any extended period. Figure 1 shows that the F:\ array is exceeding this limit; the queue is stacking up at times to 255 outstanding I/Os. This means that large amounts of data for read or write are piling up in the queue—the maximum the SAS controller is designed by industry specification to handle.

Both the read and write throughputs in Figure 1 are very low for an array of 15K HDDs. The write workloads in particular are averaging under 25 megabytes per second (MBps), with the exception of backup jobs happening from around 11:30 p.m. to midnight. This workload amount indicates an inability to handle the write transactions, forcing them to pile up in the queue throughout the day. A single 15K 300-GB SAS drive, for instance, with a perfectly sequential write workload should be roughly capable of writing up to 200 MBps to disk.[1] An eight-disk RAID10 array with four disks writing and four disk mirroring should be capable of writing up to 800 MBps. These estimates do not take into account the write cache on the controller, which increases performance, or the mirroring activity, which depending on the controller can decrease performance.

Even with these approximate figures for maximum write speed, it is readily apparent from the 25 MBps write speed of our F:\ disk array and its constant high queue depth that a random pattern of I/O is slowing down performance. The randomness is forcing so much back-and-forth movement of the drive's read/write head—head seeks—to find spaces

[1] Drives are rated using synchronous transfer rate. The synchronous transfer rate for HP 300-gigabtye 6G serial-attached SCSI 15K RPM dual-port enterprise hard disk drives is 6 gigabits per second (Gbps) – see: http://h18000.www1.hp.com/products/quickspecs/12244_na/12244_na.html. To compute our estimated write speed, we divided the 6 Gbps by 8 (8 bits per byte). That equals 750 gigabytes per second or about 750 megabytes per second (MBps). A single physical disk will never be able to achieve 6 Gbps; based on Intel IT experience, it's normally about 25 percent of this speed. In a redundant array of independent disks (RAID) array, it normally takes at least 4 disks RAID 0 (set for no parity/redundancy, just performance) to saturate the interface under perfect conditions. So for one disk, given the most ideal conditions and the benefit of the doubt, we generously divided the 750 MBps by 4 and rounded up to get ~200 MBps per 15K RPM drive.

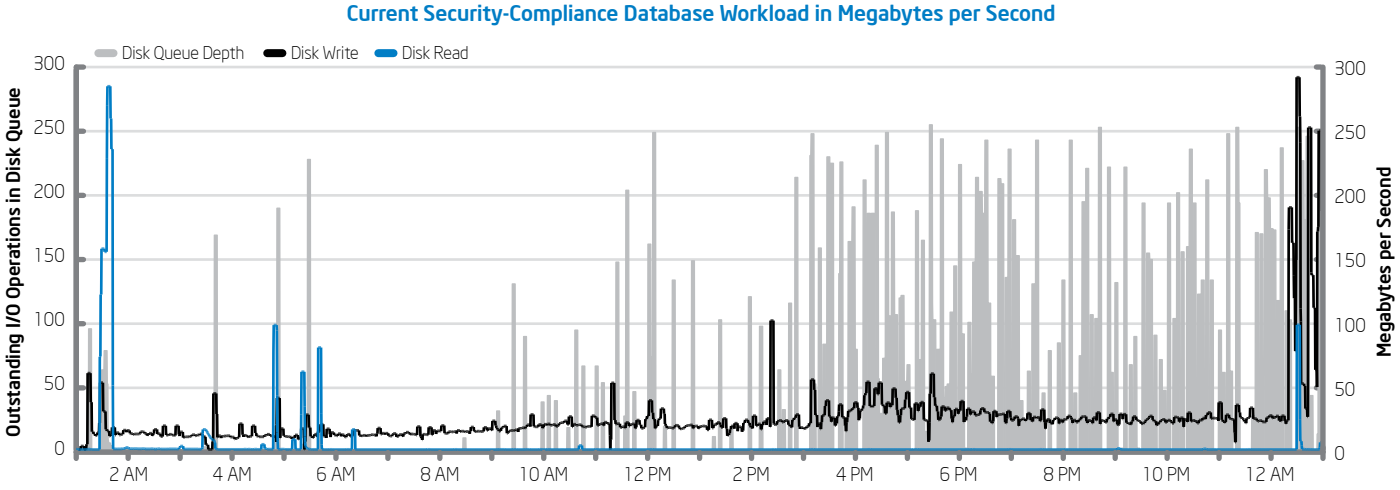

Figure 1. Measurement of a current security-compliance database workload in megabytes per second. Note the disk queue depth.

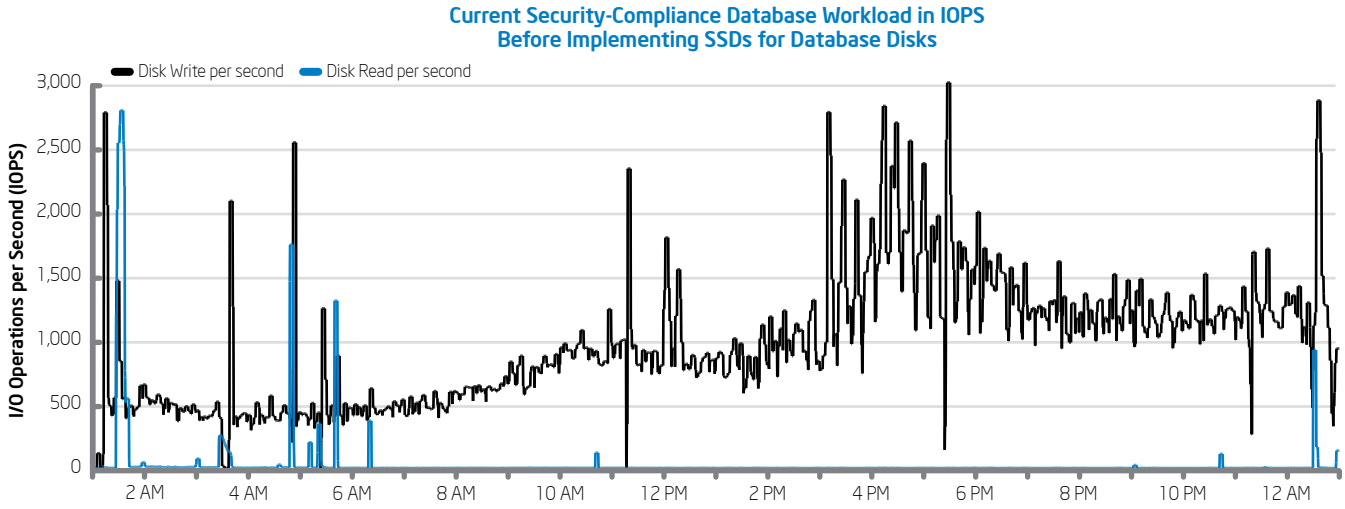

Figure 2. Measurement of current security-compliance database workload in I/O operations per second. Note the write-intensive workload.

to write data that the drives are unable to keep up with the incoming data.

The database application we use actually performs a lot of cleanup in the background. Tables are cleaned up and placed in a sequential order. This means the security database file itself is not fragmented. Instead, the data flowing into the database tables is highly random and thus makes the database behave like an extremely fragmented file.

Figure 2 shows read-and-write disk throughput in IOPS. A comparison of the read-and-write IOPS indicates a write-intensive workload. We can also surmise from the low amount of disk reads that these reads are neither impeding write activities nor causing the excessive queue depths. Except for a few spikes throughout the day from backup activities, disk reads on the F:\array are consistently low.

Looking at the disk writes in terms of IOPS provides important data for helping model the workload for testing SSDs. By dividing throughput (MBps) by IOPS, we can see that not only is the write activity random, but it is also approximately 16 kilobytes (KB) in block size, which is consistent with a Structured Query Language (SQL) database such as our security-compliance database solution.

From the workload data we collected (Figures 1 and 2) and what we know about the database and the HDD disk array, we came to the following conclusions:

- The database, which is 80 GB, uses approximately the first 27 percent of the eight 73-GB disks making up the RAID 10 F:\ array.
- The fact that only 27 percent of the disk space is being used, the queue depth is extremely high, and the throughput rates are very low, indicates a workload that corresponds to a 100-percent random I/O at approximately a 16-KB block size with 30-percent read rate.

Modeling the Workload in Iometer

Taking what we know from the data collected in production, the next step was to model the workload with a tool for testing. Iometer is an open source I/O subsystem measurement and characterization tool (www.iometer.org) for single and clustered disk systems. It is used as a benchmark tool, for troubleshooting, and for modeling I/O workloads. It can be configured with disk parameters, such as maximum disk size, number of outstanding I/Os, number of I/O threads, percent random or sequential distribution, read/write distribution, and block size, to replicate the behavior of applications such as our security-compliance database.

For our SDD test, our goal with Iometer was to model as closely as possible the security-compliance database workload that we measured and profiled with Perfmon.

We first configured Iometer with the data we collected in examining the existing workload: 16-KB block sizes, 100-percent random distribution, and a 30-percent read rate—our worst case scenario. (Simultaneous reads and writes are the worst case for every drive.) At the same time, we didn't want to artificially load the system up beyond real-life usage, so we set it up to run four individual worker processes with four queues each to keep the total queuing on the system under test to 16 queues, with two queues per drive. This is consistent with the workloads our system currently handles—SQL in the background will launch between only four and eight different processes that are going to write to a database at any given time. Also, a general engineering rule of thumb is to have no more than two I/O queues for each drive in a RAID set.

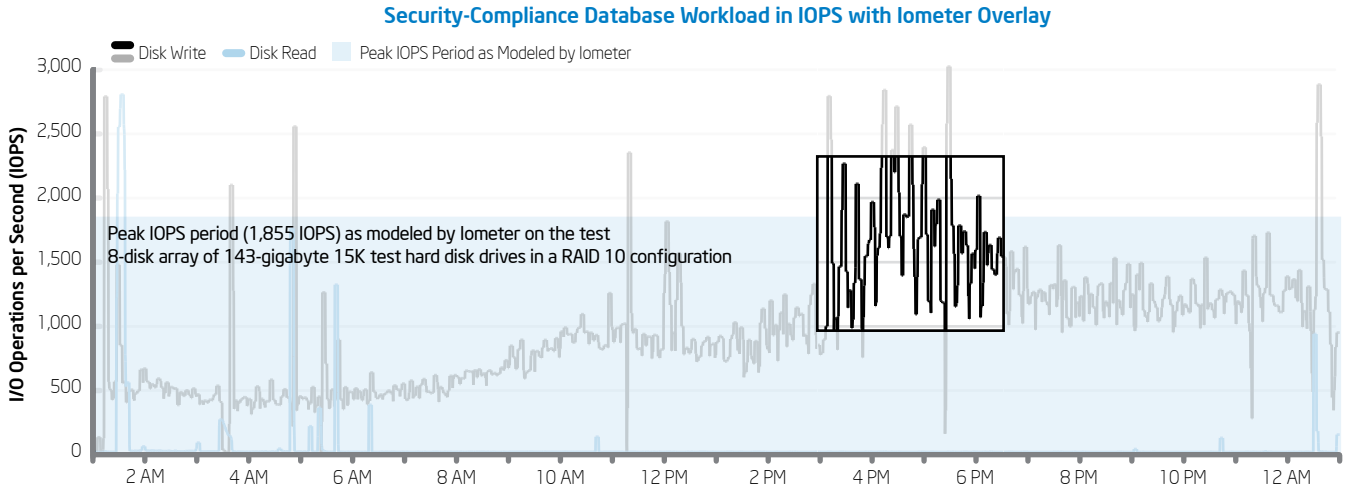

Figure 3. Peak period, shown in the black box, of write I/O operations per second of security-compliance database workload on the day measured.

One advantage of SDDs having no moving parts is that they do not require time to come out of idle. To ensure a fair test, we included a two-minute ramp time to allow for any necessary spin-up time in the 15K HDDs and to negate the effects of cache bursts. This ramp time was followed by a 10-minute runtime to measure performance between the two drive types. Each test was run three times. The run variance was less than plus-or-minus 5 percent. Results from the three runs were averaged to provide the end result displayed in our performance graphs.

Drive controllers were set at their recommended settings.

- SSD: Drive cache on, adaptive read-ahead and write-back caching on
- HDD: Drive cache off, adaptive read-ahead and write-back caching on

TEST RESULTS

For our test in the controlled environment, we used 100-GB SSDs and 143-GB 15K HDDs in eight-drive arrays in both RAID 5 and RAID 10 configurations. The logical disk capacity of the eight-drive array setups tested was set at 320 GB, our database growth limit target. Test servers were identical in configuration, including their array controllers.

Validating Model Workload against Real Workload

Once the workload was modeled in Iometer, we wanted to validate that the Iometer load would replicate the security-compliance database workload satisfactorily in the control environment. Figure 3 shows that the peak I/O measurement during the day—shown in the black box—averages about 1,850 IOPS on the actual 80-GB database workload on our data center running on an eight-disk array of 73-GB HDDs in a RAID 10 configuration. Note that this is with about a 27-percent disk capacity utilization, which is the amount we want to replicate.

For the Iometer model test, we set up a 160-GB logical disk just for this particular test on an eight-disk array of 143-GB 15K test HDDs in a RAID 10 configuration. We selected 143-GB HDDs because they were the smallest we could find; 73-GB HDDs are no longer available. The blue-shaded area of Figure 3 tops out at about 1,855 IOPS with about a 28-percent disk capacity utilization. Since these results are nearly the same percentage of disk capacity utilization and peak write IOPS as our production environment, they validate our model in Iometer and clear the way for testing using this model.

Determining Top Line Performance

Figure 4 shows the average results during the 10-minute runtime with a logical disk capacity setting of 320 GB, which is our goal for handling future growth of the database. The SSD array in a RAID 5 configuration handles 5x the number of IOPS of the HDD arrays. In the RAID 10 configuration, the advantage is also nearly 5x. This performance advantage clearly demonstrates the advantages of SSD arrays in random data accesses. Being able to read directly from any location with no mechanical motion, the SSD arrays show little or no I/O penalty as Iometer delivers the workload to the disks and they fill with data. Facing the same workload, the HDD disk arrays bog down.

For average latency, shown in Figure 5, again there is a large discrepancy between the SSD arrays and the HDD arrays. The SSD array in the RAID 5 configuration shows a 10x lower read latency and the SSD array in the RAID 10 configuration records an 8x lower read latency. In write latency, both the SSD arrays in the RAID 5 and RAID 10 configurations show a 7x lower score. Particularly revealing is maximum latency, shown in Figure 6, which represents the worse performance the drives will deliver in responding to a request for I/O. The difference between the 15K HDD arrays score of 1.2 seconds compared to the SSD array score of 154 milliseconds in the RAID 5 configurations is dramatic. In the RAID 10 configurations, the SSD array records an 8x lower maximum latency in comparison to the HDD array. It's clear that both SSD arrays should provide much faster response to database requests.

Calculating Endurance

An important part of our evaluation was calculating endurance. While the performance tests show that SSD arrays are clearly superior in both IOPS throughput and access time (read and write), most IT departments want to know if SSDs will last long enough to justify their higher cost.

As previously noted, MLC NAND has a finite number of block-erase cycles. The finite number of block-erase cycles is further reduced by write amplification, which is the amount of data an SSD controller has to write in relation to the amount of data the host controller wants it to write. A write amplification of 1 is ideal. It means, for instance, that 1 MB was the desired amount to be written, and 1 MB was written. Because NAND must be erased before it can be rewritten, the process to perform these operations involves moving, or rewriting, data more than once. This multiplying effect increases the number of writes required over the life of the SSD, shortening the time it can reliably operate.

HET in the Intel SSD 710 Series extends SSD endurance by a variety of methods beyond what can be achieved using standard MLC NAND. These methods include providing a spare write area to reduce the effects of write amplification and to increase the total write capacity of the drive. For instance, a standard MLC drive might have a total write capacity of 550 terabytes (TB), whereas the same drive replaced with HET NAND and the Intel SSD 710 Series logic might have a 1 petabyte (PB) total write capacity. The method of determining how long the drive will last is the same, with differing start values for total write capacity.

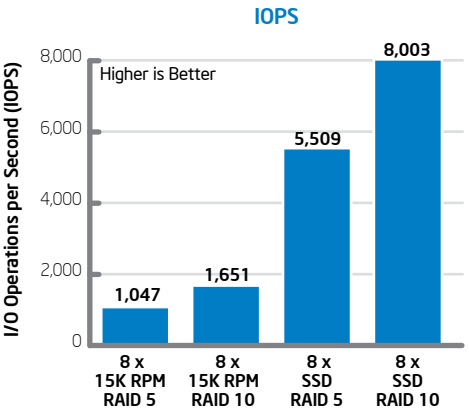

Figure 4. Comparison of I/O operations per second. 100-percent random write and 30-percent read running a 16-kilobyte block workload on a 320-gigabyte logical drive.

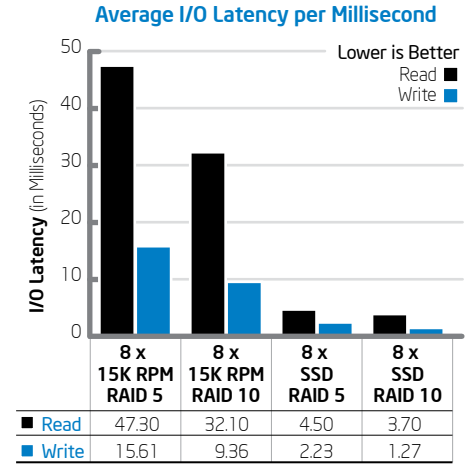

Figure 5. The average I/O latency. 100-percent random write and a 30-percent read running a 16-kilobyte block workload on a 320-gigabyte logical drive.

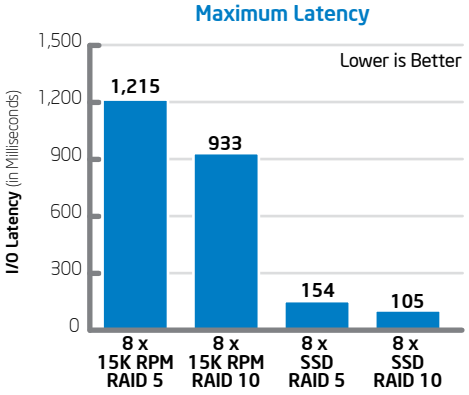

Figure 6. Maximum I/O latency. 100-percent random write and 30-percent read running a 16-kilobyte block workload on a 320-gigabyte logical drive.

Cost Benefits

A solid-state disk (SSD) with 300-gigabyte capacity costs approximately three times as much as a 15K hard disk drive (HDD) of similar capacity. For the eight-disk array used for the data volume of our security-compliance database and its replication partner in another datacenter, this is a considerable expense.

For our use model, this additional expense is more than offset by the savings involved in having a system that:

- Eliminates the considerable staff hours formerly spent manually throttling I/O to handle the unwieldy I/O queue depths.
- Eliminates backlogs in recording the monitoring data, helping to avoid a potential compliance issue.
- Reduces the potential for losses associated with delays in patching monitored systems.
- Provides the performance and capacity to handle the projected growth of the workload over the next three to five years.
- Simplifies the solution to a local disk setup instead of a more complicated solution such as a storage area network or network-attached storage.

Using our database I/O model and a five-day burn-in test with Iometer, we calculated that we write approximately 151 GB per day to the SSD RAID 10 array, which is 54 TB a year. We know that the total write capacity or life of an Intel SSD 710 Series is approximately 1 PB. If we multiply the writes per year (54 TB) times the write amplification (3)—measured using the Intel® Solid-State Drive Toolbox—and then divide 1 PB by that number, we find that the life of an individual SSD drive in our database use case is 6.4 years.

Now we must factor in the effect of the RAID settings for the eight-drive arrays. In a RAID 10 configuration, this spreads the total writes over four drives, so the life of the array becomes four times 6.4 or 25.6 years. A RAID 5 configuration spreads the total writes over seven drives, so the life of the array becomes seven times 6.4 or 44.8 years.

We also need to consider Intel IT's direct experience. In our IT lab, even using beta samples of Intel SSDs on heavy workloads over the last four years, we've experienced only two drive failures in approximately 500 samples. This is a failure rate of just 0.4 percent.

Calculating the Uncorrectable Bit Error Rate

Drives have a variety of methods to ensure data integrity. SSDs typically use parity checking or ECC to correct bit errors and, along with other methods, to avoid data integrity issues. Before using an SSD array for our security-compliance database, we wanted to know the UBER of the drives. The lower the UBER, the better the SSD is at ensuring error-free data.

UBER scales with the read rate of drives. To determine UBER for our intended drives, we first determined the average read bandwidth in MBps, which came to 9.38 MBps. We then determined the total reads (bits) per year by multiplying 31,536,000 (the number of seconds in a year) × 1024 KB/MB × 1024 bytes/KB × 8 bits/byte. This equaled 2.48141×10^{17}.

According to Intel's NAND Solution Group and Intel SSD 710 Series specifications, the NAND cell bit error rate—the statistical probability of an uncorrectable error—based on a 1,000 drive sample is 1×10^{-17}. If we multiply the NAND cell bit error rate by our total reads in bits per year ($1 \times 10^{-17} \times 2.48141 \times 10^{17}$), we get the number of bits that can be expected to fail in a year, which is 2.481 in a deployment of 1,000 drives. This corresponds to a 0.248 percent failure rate. If we then divide this number by three to account for an eight-hour duty cycle per day (the approximate amount of time the server runs the database each day), we get a 0.083 percent probability of one failure over the course of a year in a 1,000-drive deployment.

This extremely low failure rate can be virtually eliminated or masked by the following:

- RAID controller scrubbing, also known as *patrol read,* which is a process of sequentially reading all data and their corresponding parity information, and rebuilding parity whenever needed
- Database transaction log shipping and data replication, which ensures all writes to a database are replicated on another system
- Any file system, including NTFS, that performs sector sparing, which marks bad or inconsistent sectors and remaps them

Power and Heat Comparisons

While power and thermal considerations were not the primary focus of our research, both these factors are a high priority in the selection of components for today's data centers.

For a comparison in power consumption, we used specifications provided by a leading manufacturer of storage systems. These specifications compare the operating and idle power consumption of 100-GB SSDs with 300-GB 15K HDDs (see Table 1). The data shows that power savings of well over 50 percent are possible with SSDs.

How much heat a drive emits in operation affects the overall cooling requirements for individual servers and the data center at large. This is significant because cooling is one of the major power consumers in a data center. To demonstrate the difference in heat output under load between the tested HDDs and SSDs, we used a thermal imaging camera. The photo shown in Figure 7 shows that under load, the SSD produces approximately one third less heat.

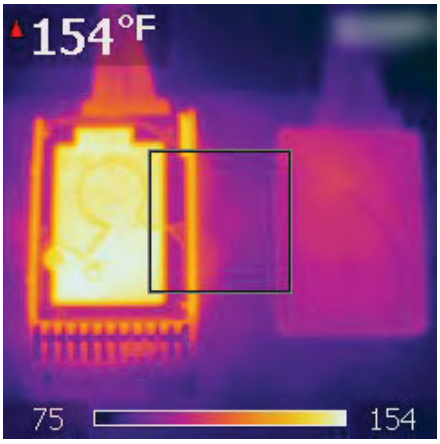

Figure 7. This thermal imaging photo compares the heat output under load of one of the 15K 146-gigabyte (GB) hard disk drives (left) and one of the 100-GB solid-state drives (right) used in our testing.

Table 1. Power Consumption Comparison between Solid-State Drives and Hard Disk Drives

State	100-gigabyte (GB) Solid-State Drive	300 GB Hard Disk Drive	Power Savings
Idle	1.38 watts	8.74 watts	7.36 watts 84 percent
Operating	4.97 watts	12.92 watts	7.95 watts 61 percent

Note: Figures were derived from page 4 of this data sheet: http://www.emc.com/collateral/software/specification-sheet/h8514-vnx-series-ss.pdf

Endurance Calculation Methodology

Here are the steps to calculate the expected life of a solid-state drive (SSD) and an SSD array for a particular workload.

1. Measure the existing workload.
2. Model the workload in Iometer.
3. Validate the workload model by running the workload in a controlled environment using Iometer to read and write to a system and drive array that is similar to those used to measure the existing workload.
4. Adjust the workload as necessary to reproduce the production workload accurately.
5. Install the test SSD drive array in the test system.
6. Run the test for a set number of days.
7. Take one drive at a time out of the array and put it in another server. Use Intel® Solid-State Drive Toolbox (www.intel.com/go/ssdtoolbox) to determine the write amplification and to collect the necessary data to calculate the amount of data written per day (write/day).
8. Determine from the write/day the amount of data that would be written in a year on the drive and multiply by the amplification.
9. Divide the expected life (in petabytes) of the disk by the write/year. This is the expected life of the disk.
10. Use the average of the expected life of each disk in the array to determine the expected life of the array, taking into consideration the RAID configuration used.

The Intel® Xeon® Processor E5 Product Family

Our testing was completed before the release of the Intel® Xeon® processor E5 product family. This processor integrates the I/O controller and hub onto the processor die, reducing I/O latency by up to 30 percent. An evaluation revealed that on this platform, a single Intel® Solid-State Drive 710 Series 300-gigabyte using the motherboard non-RAID controller can perform with our production security database workload as well as a RAID controller running an eight-disk RAID 10 set. This performance opens the doorway to novel software-based RAID configurations in the future that use the power of the Intel Xeon processor E5 product family to break the bottleneck for local storage, which now squarely rests at the RAID-enabled storage controller.

PRODUCTION ENVIRONMENT RESULTS

Based on the favorable results in our test environment, we replaced the 15K HDD RAID 10 eight-disk array used for the data volume of our security-compliance database with a RAID 10 eight-disk array of 300-GB SSDs. We also updated the six-drive 15K HDD RAID 10 six-disk array used for the log volume with a RAID 10 six-disk array of 300-GB SSDs. We then monitored our results for 15 days.

In the results shown in Figures 8 and 9 we found that we have:

- Eliminated the performance bottlenecks (backlogs in monitoring remediation) in the security-compliance database
- Eliminated the need to manually throttle data collection
- Created headroom to decrease the polling interval for security objects

Figure 8 shows HDD read-and-write disk activity in IOPS before implementing SSDs and the same workload after SSD implementation. Of particular interest in this graph of a complete day's activity is the ability of the disk workload to spike 25 percent higher than the HDD implementation, which demonstrates the greater responsiveness of the SSDs. Also, the afternoon IOPS workload appears more stable without frequent changes in amplitude, which demonstrates the smoother operation of the SSDs. Note that the total IOPS performed over the course of the day before and after the SSD upgrade are identical within five percent.

Figure 9 shows read and write disk throughput in MBps and disk queuing for the HDD implementation. It also shows the production data collected after SSD implementation. Some activities, such as backups, still run queues up, but for the most part the sustained queue depths are completely eliminated. It is worth noting that queuing in sequential disk operations, such as backups, actually improves the performance of the operation.

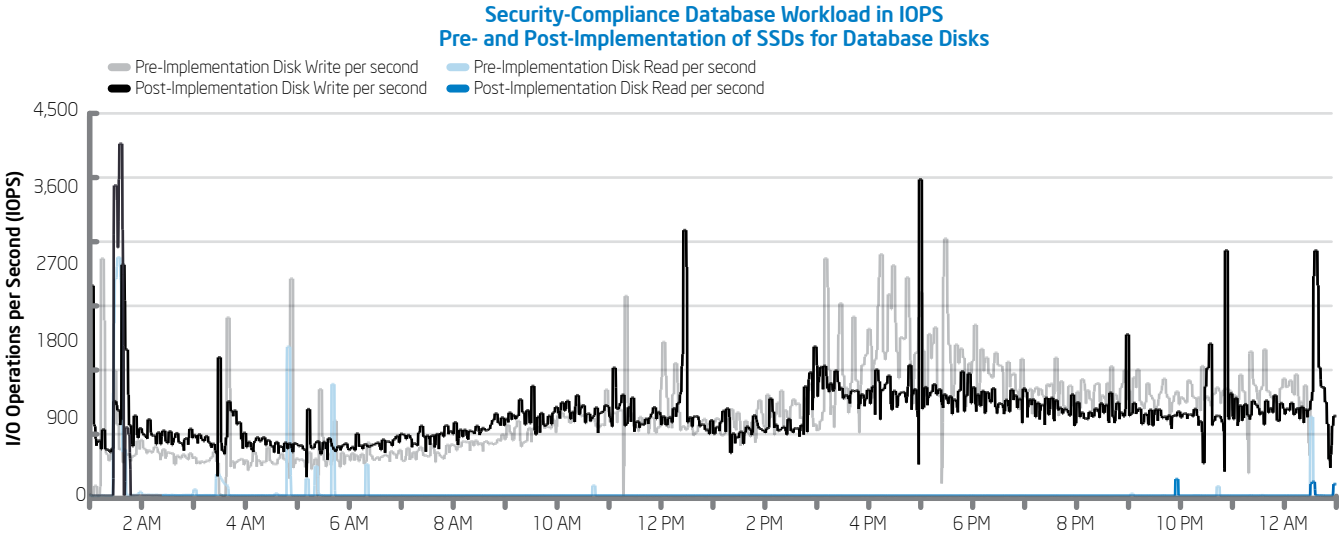

Figure 8. Measurement of security-compliance database workload in I/O operations per second, before and after implementing SSDs for the database disks.

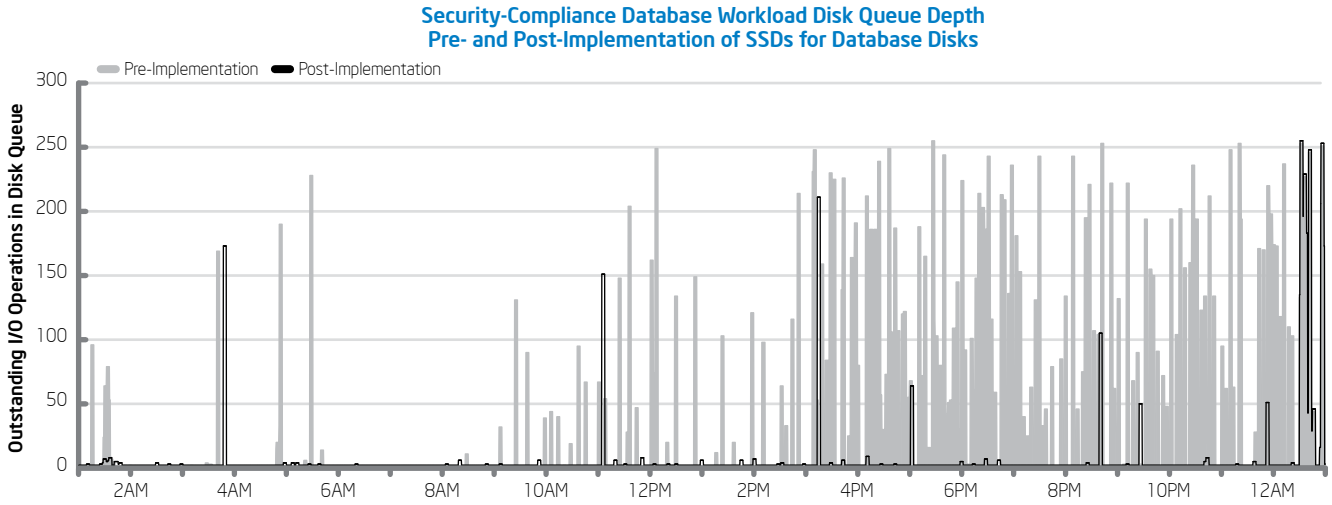

Figure 9. Measurement of security-compliance database workload showing queue length before and after implementing solid-state drives as the database disks.

The SSDs resulted in a 40x improvement in reducing the number of queued I/Os while maintaining read and write MBps. This particular improvement will allow the team to decrease the security-polling interval and move compliance reporting 4x closer to real time than the current setting.

CONCLUSION

Based on our testing, we have implemented SSD arrays for handling the data and log volume operations of our security-compliance database, and we are already seeing results.

Our testing in a controlled environment and then in our production data center demonstrated that for workloads generating random disk I/O, SSD arrays significantly increase performance. For such applications, their performance, high reliability, functional lifespan, and lower power and cooling requirements offset their higher initial cost.

Our testing revealed that switching to the tested SSDs can achieve the following:

- Reduce performance bottlenecks by increasing disk performance up to 5x on random disk I/O tasks
- Deliver up to 10x lower read latency, up to 7x lower write latency, and up to 8x lower maximum latency for faster response to patching and compliance data read-and-write requests
- Provide faster performance when spinning up from idle and incur no penalties as drives fill or fragmentation increases
- Offer significant cost benefits in everything from reducing staff hours spent dealing with long I/O queue depths to improving compliance by reducing the time it takes to respond to patching requests and record them.

We found that what is important is the methodology of our measurement, replication, testing, implementation, and final production measurement using available tools and hardware in the lab. This is what demonstrated the actual improvement for disk I/O in the security database workload, and gave us the confidence to make the leap to SSDs. At the end of the day, we improved the user experience for the security database team and the 120 console users who now think their application is "snappy." Based on these results, we plan to look for other applications in our data center with other workloads that could benefit from a switch to SSD arrays.

For more information on Intel IT best practices, visit www.intel.com/it.

CONTRIBUTORS

Ed Giguere

Mark Jackson

Terry Yoshii

ACRONYMS

ECC	error correction code
GB	gigabyte
Gbps	gigabits per second
HDD	hard disk drive
HET	High Endurance Technology
IOPS	I/O operations per second
KB	kilobyte
MB	megabyte
MBps	megabytes per second
MLC	multi-level cell
NAND	not and (electronic logic gate)
NAS	Network-Attached Storage
P/E	program/erase
PB	petabyte
RAID	redundant array of independent disks
RPM	revolutions per minute
SAN	Storage Area Network
SAS	serial-attached SCSI
SCSI	Small Computer System Interface
SQL	Structured Query Language
SSD	solid-state drive
TB	terabyte
UBER	uncorrectable bit error rate

Printed in USA ♻ Please Recycle 0712/ACHA/KC/PDF 327463-001US