WHITE PAPER

How SanDisk® Removes DBaaS Adoption Obstacles

SanDisk®
a Western Digital brand

Western Digital Technologies, Inc.
951 SanDisk Drive, Milpitas, CA 95035

www.SanDisk.com

Barriers to Cloud DBaaS (Database-as-a-Service) Adoption

There are numerous reasons DBaaS has yet to reach its full potential. One that is often cited is that current application code may be a bit burdensome to port to the DBaaS underlying SQL or NoSQL database. As every database administrator (DBA) knows, there are assorted incompatible differences between different types of SQL databases. NoSQL databases have their compatibility issues as well. Fixing those issues isn't hard, just time consuming. And because it's a one-time task most DBAs will not typically shy away from it.

Other reasons cited such as lack of database compression or table partitions are relatively easy to rectify by the DBaaS provider.

The most crucial reasons DBAs shy away from DBaaS are their loss of ability to control database:

1. Consistent Performance
2. Availability and Reliability
3. Cost

Consistent Performance

Consistent performance is always an issue for any database. It tends to be more of an issue for transactional applications running on SQL than business intelligence (BI) or analytical applications running on NoSQL; however, it is always a key issue for both. Inconsistent performance rankles users and devastates user productivity. When that performance inconsistency occurs in a private data center or private cloud then the DBA is inundated with complaints. An unhappy DBaaS end user can lead to unchecked out shopping carts (E-Commerce), tarnished reputation, lost customers, lost revenue, and long-term lost business. For the DBaaS provider, inconsistent database performance will ultimately lead to a declining DBaaS business.

DBAs are used to manipulating databases, storage, and storage-related infrastructure to tune and squeeze as much consistent and predictable high performance as possible to meet the needs of the application at all times including escalating loads of writes and reads. DBAs have several database techniques in their arsenal of performance management tools. These tools range from sharding tables, running master-slaves, multi-masters, or clustering to control database read and write performance for the applications. Utilization of a specific technique will depend on the service provider's DBaaS and the underlying SQL or NoSQL database. Most of these database performance-tuning techniques are database dependent. Some providers offer a plethora of DBaaS to enable a broader potential range of SQL and NoSQL applications.

Storage is another DBA manipulation to provide database consistent performance. DBAs will utilize server-side DRAM, server-side flash storage (a.k.a. "software-defined storage"), or shared (SAN or NAS) all flash array (AFA) or hybrid flash storage to accelerate reads and writes. DBAs in the past (although some still do it today) squeezed higher hard disk drive (HDD) performance from short stroking[1].

DBaaS providers, as part of a public cloud offering, cannot economically or securely provide the direct access to their storage that DBAs are used to having. However, they are endeavoring to provide similar DBA storage performance controls through other means. For example, providers may provide different DBaaS tiers of storage performance that can be rented. That performance is based on IOPS or throughput with contractual SLAs. But commonly the performance guarantees are bracketed in a range. That range is quite broad because the storage performance is frequently delivered in a broad range. Overcoming the DBA performance objections and delivering the consistently high performance they require is a steep challenge. This is because delivering DBaaS read and write performance considered to be predictable high performance requires predictable low latency storage. Low latency storage performance today typically is measured in microseconds (noted by the Greek letter µ). Achieving that predictably consistent high performance requires predictably consistent low latency under load. The way that is measured is by the number of nines guaranteed to deliver that performance and latency. The number of nines equates into a percentage. Two nines equal 99% of the time, three nines equal 99.9%, etc. This combination of very low latency

[1] HDDs are electro-mechanical devices with spinning platters and a head that writes and reads the data on the platters. By limiting all writes and reads only to the outer HDD platter performance latencies decrease by up to 50% because the head only has to move a very short distance. Short stroking requires wasting 67-90% of the HDD capacity. Hence, many more HDDs are required to make up for capacity requirements in short-stroking configurations which result in high costs (as much as 10x).

storage at a high level of nines is called a high level of Quality of Services (QoS). Achieving that high QoS level is definitely a non-trivial task.

Meeting that task has previously meant a very high cost in storage and storage infrastructure. That might make sense in a private data center or even a private cloud. But in the exceedingly competitive public cloud based DBaaS market excessive cost is a non-starter. Cloud services in general and DBaaS cannot be high cost or they will not be rented. As previously discussed, cost is one of the primary objections to DBA DBaaS adoption. Adding a lot of cost to overcome the performance objection is simply counter-productive.

DBaaS providers as a result are demanding new storage standards in predictable consistent low latency, high IOPS, high throughput, and lower cost from the industry. When those demands come from cloud industry giants Amazon, Microsoft, Google, IBM, and others, they cannot be ignored.

The obvious answer is the flash SSD. However, not every flash SSD delivers the same performance or predictable consistency. In fact, contrary to conventional wisdom, flash SSDs hugely vary by vendor and model. DBaaS workloads require flash SSDs optimized for database reads and writes (generally 4K blocks with 70% reads and 30% writes) with predictably consistent extremely low latencies (ranging from dozens to hundreds of microseconds depending on queue depth). There are significant differences in write IOPS, read IOPS, throughput, latency, and most important that predictable consistency under load. The QoS for every SSD is not the same. Selecting the flash SSDs best optimized for DBaaS will eliminate the DBA consistent performance adoption objection. Selecting flash SSDs not optimized well for DBaaS without an acceptable QoS will not eliminate the DBA adoption objection.

Something to always keep in mind is the difference in bits per cell utilized in the flash NAND of every flash SSD. SLC (single level cell) is one bit per cell. MLC (multi-level cell) is 2 bits per cell. TLC (triple level cell) is 3 bits per cell. As the number of bits increases so do the number of states that can be captured and represented in a NAND cell. One bit per cell has two states (0,1); two bits per cell has four states (00, 01, 10, 11); and three bits per cell has eight states (000, 001, 010, 011, 111, 100, 110, 101). Fewer bits per cell require less electricity to write and read providing lower latency, higher performance, and longer wear life. More bits per cell are just the opposite. Cost is inversed to the number of bits per cell with fewer bits per cell costing more.

SLC is the lowest latency, highest performing, longest wear life, and highest cost. TLC is by far the worst latency, slowest performing, worst wear life, and lowest cost. That may change in the future when performance, availability, and reliability are improved. But for today, TLC flash SSDs are not an optimum choice for DBaaS. MLC is in-between and has been consistently improved to the point where it is considered Enterprise grade and an excellent choice for DBaaS because of the balance between performance, reliability, availability, and cost. More details on this will be discussed in the next two sections.

Availability and Reliability

No matter how consistent and predictable the performance, if it is not available and reliable it is useless. Database reliability in DBaaS is table stakes. That requires the DBaaS storage to be equally reliable. From a hardware perspective, that means being up and always available. Outages hurt cloud service reputations, revenues, and customer adoption of those services, even though they occur at a lower frequency than outages in private data centers or private clouds.

Just as DBaaS optimized flash SSDs should eliminate the DBA consistent performance objection, so can they mitigate the availability and reliability objections. Flash SSDs have a field record that’s shown them to be far more reliable than HDDs as described in table below.

Table 1: Why MLC flash SSDs are more reliable and available than HDDs

Why	What it Means

Overprovisioning	MLC flash SSD failures tend to be limited to program erase (P/E) blocks. When a P/E block fails, it is removed from the blocks available for writes. No capacity is lost because of MLC flash SSD overprovisioning. Overprovisioning is the amount of capacity not included in the rated usable capacity. MLC SSD overprovisioning is utilized to ensure write blocks are always available and do not have to wait for blocks with outdated data to be erased before they can write. Overprovisioning also enables the flash SSD to replace failed P/E blocks. Write optimized flash SSDs typically have ~20% or more overprovisioning, whereas read optimized flash SSDs are overprovisioned at less than 10%. (Reads create no wear on P/E blocks.)
Superior UBER Measurement	Error correcting code (ECC) in MLC flash SSDs has become quite extensive and keeps getting better. The best MLC flash SSDs today have an UBER range (depending on vendor and model) of 10^{-17} to 10^{-18}; UBER is the abbreviation for **U**ncorrectable **B**it-**E**rror **R**ate, a metric for the data corruption rate equal to the number of data errors per bit read after applying any specified error correction (ECC) method. The best HDD UBER available today is 10^{-16}; meaning MLC flash SSDs have a 100x (2 orders of magnitude) better UBER than the best HDDs. The way MLC flash SSDs treat an unrecoverable bit error (UBER) is very different as well. HDDs treat an UBER as a drive failure. MLC flash SSDs treat an UBER as a P/E block failure. Failed HDDs must be rebuilt. Failed P/E blocks are simply replaced from the overprovisioned pool.
No Parity RAID Required	RAID parity is a standard requirement on HDDs. This is to protect the data. As HDD capacity has grown, the time to rebuild drives has lengthened increasing the risk of additional HDD failures and data loss. Today standard RAID is now RAID 6 (dual parity) to protect against two concurrent HDD failures in a RAID group. RAID 6 consumes approximately 25% of a drive's raw capacity. If MLC SSDs are placed in the parity RAID group, when an UBER does occur or a write block fails, the RAID controller will take the SSD out of the group, require the drive be replaced, and start an unnecessary rebuild. It is simpler, faster, and just as reliable to copy any lost data, which would be minimal, from a zero capacity snapshot or mirror. Parity RAID is necessity for HDDs, unnecessary for MLC flash SSDs and a hindrance to today's cloud architectures.
Extensive Wear Life	Trade press has made much noise about MLC flash SSDs wear life and when they "wear out". It is called P/E cycles or the number of times a flash write block can be written and erased before it fails. MLC flash SSD manufacturers have gotten quite good at spreading that wear over every P/E block to make them last for quite a long time. Flash SSD wear life is rated as the number of drive writes per day (DWPD) over a guaranteed period of time. For example: a 1.6TB MLC flash drive rated at 1.8 DWPD warrantied or guaranteed for 5 years would have to write 3.24TBs per day, every day, to that SSD for 5 years straight before it wears out. That's 5.9 petabytes written to that 1.6TB drive. And even then it probably would not be worn out. Additionally, the SSD performance will maintain a steady state until it is worn out. That is not the case for HDDs. Ongoing HDD studies from BackBlaze and others, clearly demonstrates that HDD failure rates increase significantly after 3 years and performance declines.

Most flash SSDs optimized for DBaaS reads and write are more reliable than HDDs. There are exceptions. If the flash SSD is not optimized for DBaaS reads and writes or if the flash SSD is TLC (triple level cell.) TLC flash SSDs, even 3D TLC flash SSDs have a much shorter wear life when it comes to writes and tend to be read optimized. TLC flash SSDs also have a much higher UBER (Uncorrectable Bit-Error Rate), typically up to 1,000 times higher than the best MLC flash SSDs.

Cost

There is a pervasive perception that flash SSD costs are much higher than HDDs when compared on a price-per-gigabyte basis. The first flash SSDs were SLC based and were in fact pretty expensive. Times have changed. Today, MLC flash SSDs have a total cost per gigabyte that is equivalent or better than HDDs and their cost per IOPS is approximately two orders of magnitude (~100 times) lower than HDDs. Perceptions have not kept up. DBaaS optimized MLC flash SSDs are far more cost effective than ever before.

Take the example of a 1.6TB DBaaS-optimized MLC flash SSD. In raw capacity alone, it has 267% more than a 15K or 10K RPM HDD. When parity RAID 6 is used on the HDDs (mandatory requirement today) and no parity RAID is utilized on the SSDs, when deemed unnecessary in a RAID 0 configuration, SSDs have 356% more raw capacity. In some public cloud instances, service providers elect to offer RAID 0 to use all storage capacity in conjunction with snapshots and Availability Zone replication to address costs savings along with industry practices to protect customer data.

That difference in raw capacity translates into fewer drawers, rack space, and floor space. Space in a cloud data center is always at a premium and has a well-defined cost. SATA MLC flash SSDs deployed in a standard 2-socket x86 server, for example, requires less server or storage system infrastructure than SAS HDDs (SAS is a mandatory DBaaS requirement of HDDs). DBaaS-optimized MLC flash SSDs utilize much less power per gigabyte or IOPS than HDDs at less than 10% on average. These are real costs especially for cloud data centers.

Finally, flash SSDs dedupe and compress two to four times more than HDDs increasing the usable capacity at a much greater extent than HDDs and with far less additional latencies. Lower cost, higher performance, and greater capacities equal a better solution. Once again, not all flash SSDs are equivalent. Some cost more, overprovision less, deliver less performance, have a lower UBER, and don't offer SATA interconnect, etc.

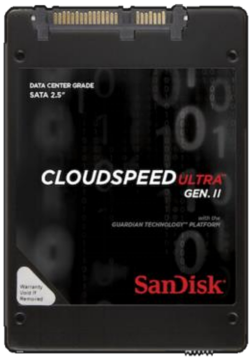
SanDisk® DBaaS Solution: CloudSpeed Ultra™ Gen. II SATA SSD

CloudSpeed Ultra Gen. II SATA SSD is the first MLC flash SSD highly optimized for DBaaS requirements. It delivers incredibly predictable high performance (up to 32,000 4K random write IOPS), at very low latency (see table 2 below) with a very high level of predictable consistency QoS (Table 2), in an extremely dense write optimized (20% overprovisioned) high capacity package specifically aimed at the DBaaS very common 30% random writes and 70% random reads (400GB, 800GB or 1.6TB raw[2]), for a very inexpensive cost (check with your local SanDisk representative).

Table 2: CloudSpeed Ultra™ Gen. II SATA SSD QoS

QoS*	< µ = Lower latency
99.9% QoS Random Write Latency 4K Queue depth 1	**80µ (microseconds)**
99.9% QoS Random Write Latency 4K Queue depth 32	**3850µ (microseconds)**
99.9% QoS Random Read Latency 4K Queue depth 1	**130µ (microseconds)**
99.9% QoS Random Read Latency 4K Queue depth 32	**319µ (microseconds)**

**DBaaS architects prefer 99.9% QoS to average latency for predictability and SLAs*



CloudSpeed Ultra Gen. II SATA SSD QoS is not found in consumer-grade SSDs and is 6 times better than its next closest competitor. Cost per IOPS comes in at approximately 1.2% of SATA or SAS HDDs. In other words the cost per IOPS is approximately 99% less than HDDs.

Reliability and availability are second to none. CloudSpeed Ultra Gen. II SATA SSD UBER rate is 1 in 10^{-18} according to the JEDEC 64.8 specification. It is also warrantied for 1.8 DWPD for five years. That fits the example previously discussed. It allows 3.24TB written to the drive every single day for five years meaning the drive has approximately 5.9 PBs of writes in its wear life warranty.

CloudSpeed Ultra Gen. II SATA SSD costs are highly competitive, but for DBaaS providers there are other cost considerations. And SanDisk CloudSpeed Ultra Gen. II SATA SSD is designed to reduce those costs as well. Rackspace density is a good example. High IOPS in a small package means a 2-socket, 24 bay server can deliver up to 768,000 random write IOPS in 2U.

This means fewer servers are required to deliver the performance required by the DBaaS databases; reducing rack and floor space; reducing the number of database instances required to meet customer application performance requirements resulting in fewer licenses and/or support agreements; which in turn reduces cloud provider administrative costs; and much lower power, heat, and cooling costs.

Conclusions

DBaaS adoption by DBAs faces several barriers including consistent and predictable performance, availability and reliability, and cost. CloudSpeed Ultra Gen. II SATA SSD is specifically optimized to eliminate or at least mitigate each of these barriers making DBaaS far more palatable to the reluctant DBA.

For More Information

Contact SanDisk at: SanDisk CloudSpeed Ultra Gen II SATA SSD or www.sandisk.com/about/contact.

[2] The raw capacities increase by 20% (480GB, 960GB, and 1.92TB) when CloudSpeed Ultra Gen. II SATA SSDs are optimized for reads with 10% random writes and 90% random reads.