Intel® Solid State Drive Data Center Family for PCIe* in Baidu's Data Center Environment

Case Study

Order Number: 334534-002US

Ordering Information

Contact your local Intel sales representative for ordering information.

Revision History

Revision Number	Description	Revision Date
001	• Initial release	June 2016
002	• Corrections to Select, QPS values in Table 3-2	June 2016

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase.

For more complete information about performance and benchmark results, visit http://www.intel.com/performance.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. **No computer system can be absolutely secure.** Check with your system manufacturer or retailer or learn more at **intel.com.**

Intel, the Intel logo, Intel SDD Data Center Family for PCIe*, Intel SSD DC P3600 Series, P3600, Intel Cache Acceleration Software, Intel CAS, and Intel Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others.

Contents

Tables

Figures

1 Introduction of Intel® SSD

Solid State Drive (SSD) performance has achieved many improvement breakthroughs compared to traditional hard disk drive (HDD); excellent stability, silent operation and because it has no moving parts, low power consumption.

Intel® SDD Data Center Family for PCIe* combines Intel's strengths in technical innovation, design, manufacturing, testing, and certification, giving Intel's PCIe SDDs with NVMe excellent and consistent performance. Intel's 5-year limited warranty and cost benefits make Intel's PCIe SSDs an excellent solution for demanding data center environments and helps to meet the needs of the growing data center market.

Intel SSD Data Center Family for PCIe combines the proven benefits of Intel's data center SSDs and NVMe technology which is designed specifically for non-volatile storage, resulting in revolutionary SSD products.

As one of China's largest Internet companies, Baidu is renowned for its leading internet search service. Baidu's constant pursuit of innovation and new technology has enabled Baidu to maintain its leading position in the industry. Baidu has achieved great success over the years by combining Intel SSDs, software optimization, and storage management solutions along with years of investment and accumulation in Research and Development (R&D). Currently a large portion of Baidu application deployment is performed by SSDs, 90%+ of which are Intel Data Center SSDs; performing all online services, including search, relational database (MySQL*), KV database, content distribution network (CDN), etc.

Based on more than a year of testing and software optimization the PCIe-based Intel® SSD DC P3600 Series with NVMe meets the needs of Baidu performance applications.

2 Management of Intel's PCIe SSDs in Baidu Data Center

2.1 Basic Mode

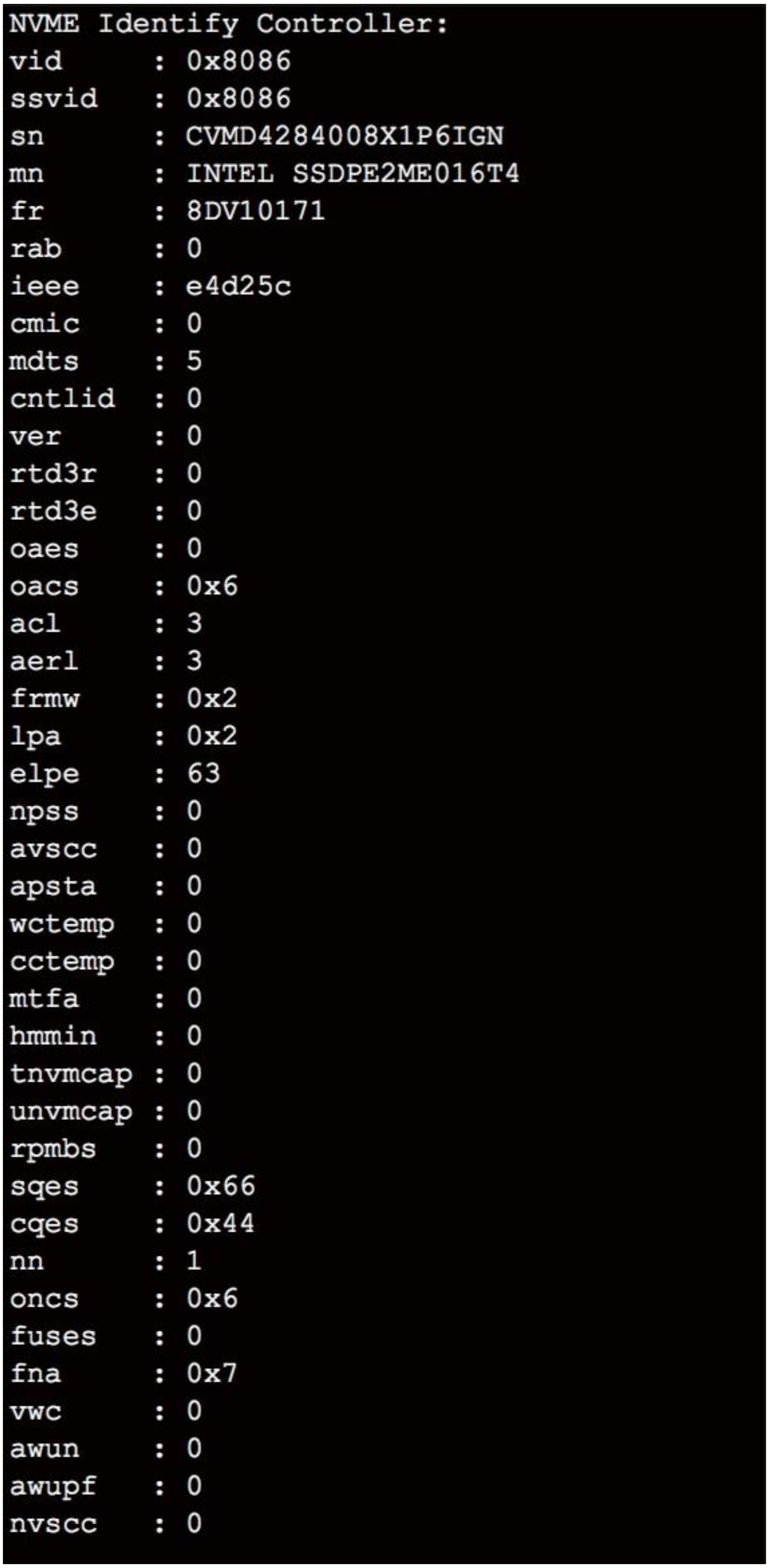
Because NVMe is a new SSD interface protocol that cannot use SCSI management tools, Baidu has developed a new management tool. Introducing command line program

nvme-cli, Baidu's new open source NVMe SSD user-mode management tool with adaptive compatibility that is very well suited for Baidu's application scenarios.

Baidu has integrated the command line program nvme-cli tool into its Linux* distribution to accomplish the basics of NVMe device management using the following commands; ***id_ctrl*** and ***smart-log/smart-log-add***.

- The ***id_ctrl*** command retrieves device information, including device model, serial number and firmware version; information needed to perform efficient asset management. The resulting information is shown in Figure 2-1.

Figure 2-1: NVMe SSD Profile

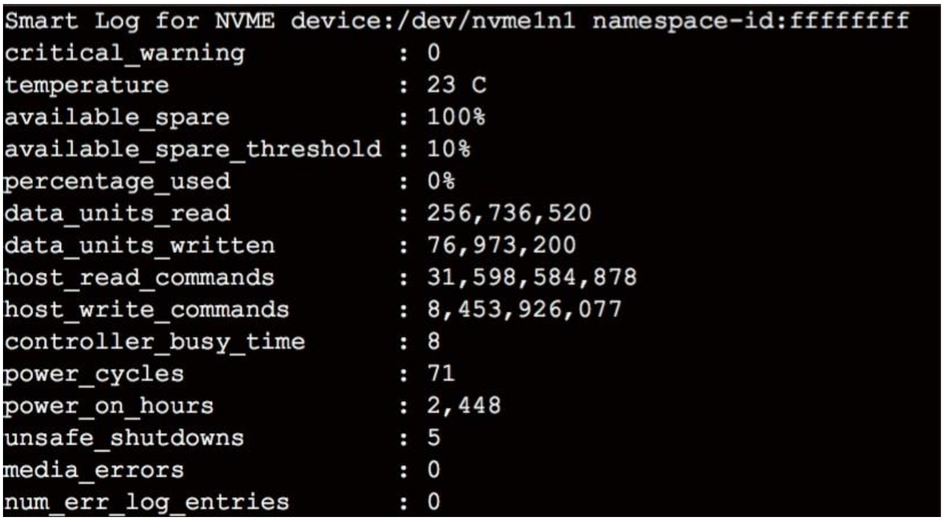


```
NVME Identify Controller:
vid     : 0x8086
ssvid   : 0x8086
sn      : CVMD4284008X1P6IGN
mn      : INTEL SSDPE2ME016T4
fr      : 8DV10171
rab     : 0
ieee    : e4d25c
cmic    : 0
mdts    : 5
cntlid  : 0
ver     : 0
rtd3r   : 0
rtd3e   : 0
oaes    : 0
oacs    : 0x6
acl     : 3
aerl    : 3
frmw    : 0x2
lpa     : 0x2
elpe    : 63
npss    : 0
avscc   : 0
apsta   : 0
wctemp  : 0
cctemp  : 0
mtfa    : 0
hmmin   : 0
tnvmcap : 0
unvmcap : 0
rpmbs   : 0
sqes    : 0x66
cqes    : 0x44
nn      : 1
oncs    : 0x6
fuses   : 0
fna     : 0x7
vwc     : 0
awun    : 0
awupf   : 0
nvscc   : 0
```

- The ***smart-log*** command conducts real-time monitoring of NVMe storage and determines the health condition of the device.

When ***smart-log*** detects a problem, a message is automatically generated by combing the registered information of the device and pushing the information to the operations and management (O&M) platform to enable rapid recovery. The obtained NVMe SSD status information is shown in Figure 2-2.

Figure 2-2: PCIe/NVMe SSD Status Information



```
Smart Log for NVME device:/dev/nvme1n1 namespace-id:ffffffff
critical_warning                    : 0
temperature                         : 23 C
available_spare                     : 100%
available_spare_threshold           : 10%
percentage_used                     : 0%
data_units_read                     : 256,736,520
data_units_written                  : 76,973,200
host_read_commands                  : 31,598,584,878
host_write_commands                 : 8,453,926,077
controller_busy_time                : 8
power_cycles                        : 71
power_on_hours                      : 2,448
unsafe_shutdowns                    : 5
media_errors                        : 0
num_err_log_entries                 : 0
```

2.2 Customizable Smart Features

In addition to achieving the basic management mode, Baidu and Intel are co-developing a smarter NVMe storage management mode. Currently, they are making great efforts in the following two areas:

- NVMe log export function – conducts rapid fault analysis and fault category classification.
- Physical slot positioning – rapidly repositions the failed device to improve maintenance efficiency.

In the future, more functions (e.g. dynamically adjusted OP, SSD life expectancy) will be introduced to achieve the requirements of software-defined storage (SDS); further enhancing the performance advantages of storage devices and improving manageability.

2.2.1 Log export:

The test allows Baidu to integrate the command line program **nvme-cli** into the process of extract smart log information using the commands, ***smart-log* / *smart-log-add***. This features then does a before and after comparison, and characterization of those results, allowing the disposition of the NVMe faults found. This process regularly scans NVMe device information and pushes the discovered exception logs to the fault pool for processing by the O&M engineer.

2.2.2 Physical slot positioning:

Confirming the relationship between PCIe bus number and slot on the front panel, allows the O&M team to determine the physical slot positioning of NVMe device. The O&M team can leverage NVMe hot plug functionality to rapidly replace failed SSDs to improve maintenance efficiency.

3 Applying and Optimizing Intel PCIe SSDs in Baidu Data Centers

3.1 Intel PCIe SSD in Baidu Key-Value Database System

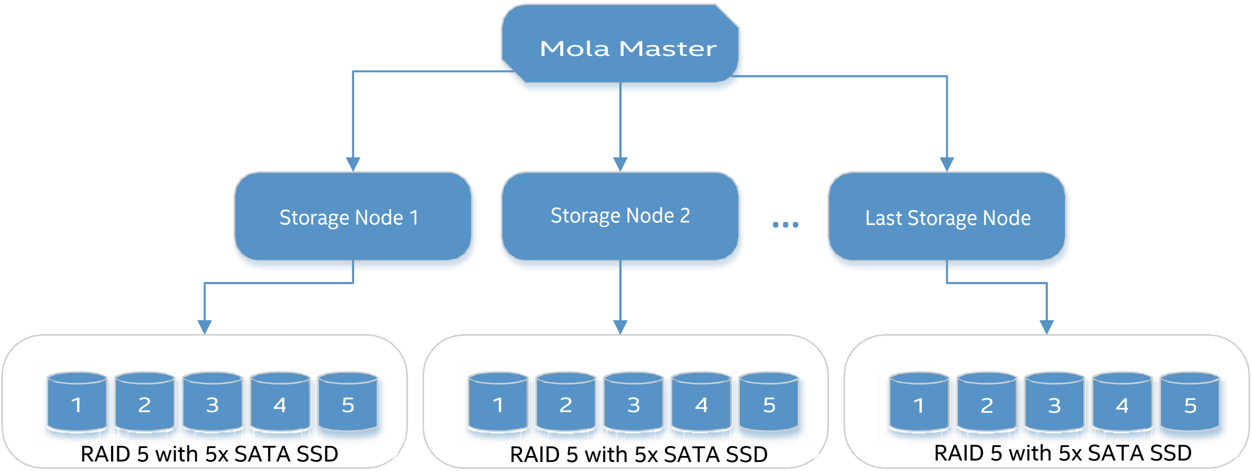
3.1.1 Overview of Baidu KV Database System

A key-value (KV) database system is generally used in a Baidu software stack for Key-Value storage. In the search service, it is used to store index data and its inverted index in the search engine. It is also extensively applied in other cloud services, mainly hot storage and IO intensive applications that rely on SSDs.

3.1.2 Introduction of Original Solution

The original hardware and software architecture of KV database is shown in Figure 3-1. The master of the KV database is responsible for responding to the query request and sending this request to different storage nodes, then the storage engine on the node will, according to the Key of the request, query whether the data exists in the memory and decide whether to read the data from the disk to return the request result. Different engines have different operation modes here. In the meanwhile, to improve single-server IOPS, Baidu applies RAID 5 with five or six SATA-based Intel® SSD DC S3500 Series to maximize the performance; and the reliability is addressed by replication mechanism. Baidu has done a lot in software and hardware to improve the single-server IOPS of KV database, and P3600 is a high-performance Intel PCIe SSD with NVMe and provides extremely high IOPS and extremely low latency, and its application in KV database can be expected to yield good results.

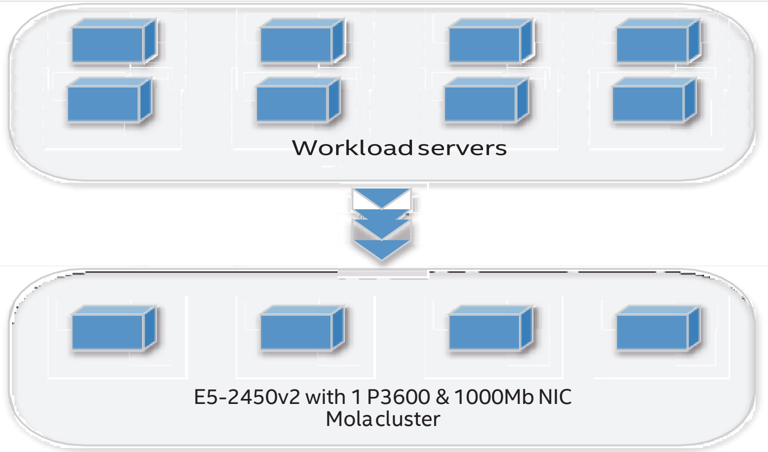
Figure 3-1: Original Hardware Architecture of Baidu KV Database



3.1.3 NVMe SSD Solution and Performance Optimization

To verify that Intel's PCIe SSD with NVMe was able to improve the performance of KV database, a 12-node test environment was constructed, with 8 set up as the workload servers and 4 as a KV database cluster. To improve the burst capability of single-node IOPS, each node of KV database was equipped with one Intel® SSD DC P3600 Series, as shown in Figure 3-2.

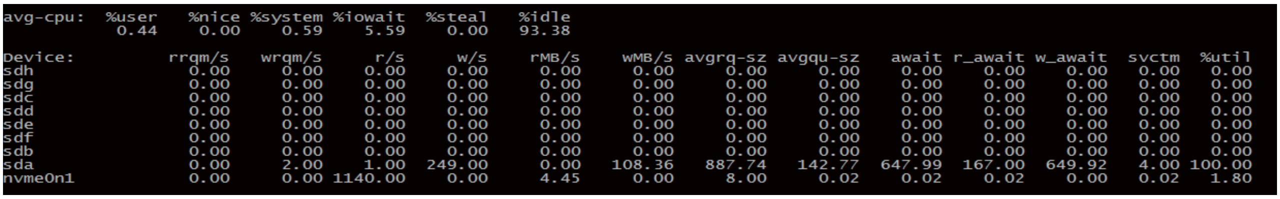
Figure 3-2: Test Configuration for Baidu KV Database System based on Intel PCIe SSDs with NVMe



During the performance test, the first issue encountered was that the performance did not improved after installing P3600. The main reasons shown in Figure 3-3, and sda's %util reaches 100%. The primary cause is that the single engine of KV database writes the log into sda and, to improve the problem positioning, the logging level is DEBUG mode, which is not an obvious issue when SCSI SSD rather than P3600 is installed. The temporary solution is to raise the log level while the long-term solution is to record the logs on the PCIe SSD with NVMe. Here, we first raise the log level to INFO.

With the increase of pressure, KV database soon reached the bottleneck point. The performance of single server was about 25,000 IOPS, but it still did not reach the expected result. At this time, the performance bottleneck mainly occurs in the Gigabit NIC, which can be addressed by upgrading to a 10 Gigabit NIC. A common issue for the 10 Gigabit NIC is unbalanced soft interrupt, as shown in Figure 3-4. The hard interrupts are addressed by different hardware queues while the soft interrupts are balanced manually.

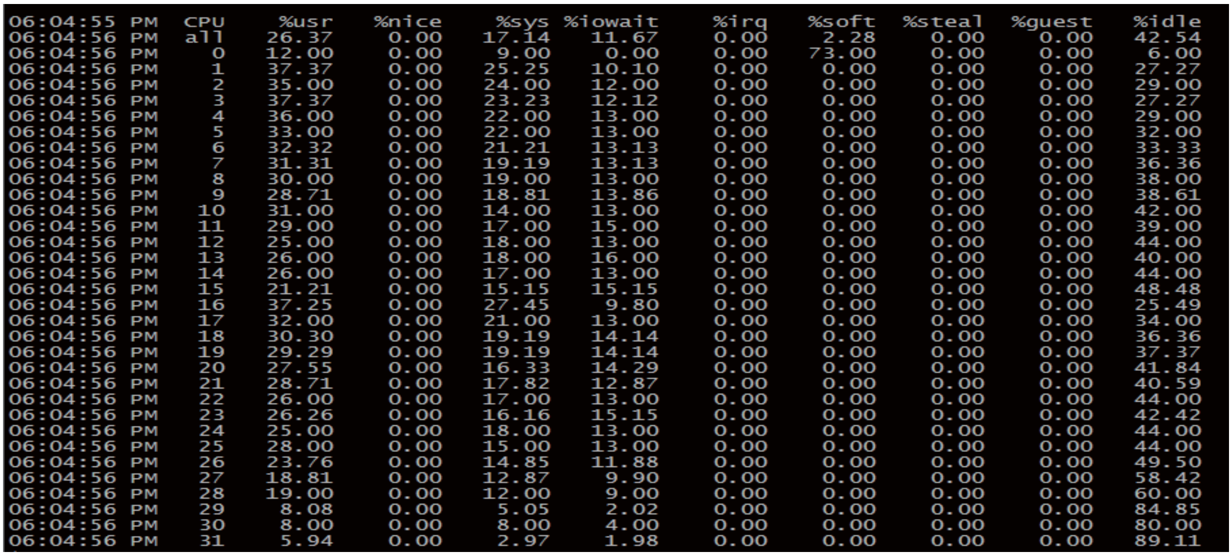
Figure 3-3: Disk IO Performance Data



```
avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           0.44    0.00    0.59    5.59    0.00   93.38

Device:         rrqm/s   wrqm/s     r/s     w/s    rMB/s    wMB/s avgrq-sz avgqu-sz   await r_await w_await  svctm  %util
sdh               0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00    0.00    0.00   0.00   0.00
sdg               0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00    0.00    0.00   0.00   0.00
sdc               0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00    0.00    0.00   0.00   0.00
sdd               0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00    0.00    0.00   0.00   0.00
sde               0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00    0.00    0.00   0.00   0.00
sdf               0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00    0.00    0.00   0.00   0.00
sdb               0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00    0.00    0.00   0.00   0.00
sda               0.00     2.00    1.00  249.00     0.00   108.36   887.74   142.77  647.99  167.00  649.92   4.00 100.00
nvme0n1           0.00     0.00 1140.00    0.00     4.45     0.00     8.00     0.02    0.02    0.02    0.00   0.02   1.80
```

Figure 3-4: CPU Performance Data



```
06:04:55 PM  CPU    %usr   %nice    %sys %iowait    %irq   %soft  %steal  %guest   %idle
06:04:56 PM  all   26.37    0.00   17.14   11.67    0.00    2.28    0.00    0.00   42.54
06:04:56 PM    0   12.00    0.00    9.00    0.00    0.00   73.00    0.00    0.00    6.00
06:04:56 PM    1   37.37    0.00   25.25   10.10    0.00    0.00    0.00    0.00   27.27
06:04:56 PM    2   35.00    0.00   24.00   12.00    0.00    0.00    0.00    0.00   29.00
06:04:56 PM    3   37.37    0.00   23.23   12.12    0.00    0.00    0.00    0.00   27.27
06:04:56 PM    4   36.00    0.00   22.00   13.00    0.00    0.00    0.00    0.00   29.00
06:04:56 PM    5   33.00    0.00   22.00   13.00    0.00    0.00    0.00    0.00   32.00
06:04:56 PM    6   32.32    0.00   21.21   13.13    0.00    0.00    0.00    0.00   33.33
06:04:56 PM    7   31.31    0.00   19.19   13.13    0.00    0.00    0.00    0.00   36.36
06:04:56 PM    8   30.00    0.00   19.00   13.00    0.00    0.00    0.00    0.00   38.00
06:04:56 PM    9   28.71    0.00   18.81   13.86    0.00    0.00    0.00    0.00   38.61
06:04:56 PM   10   31.00    0.00   14.00   13.00    0.00    0.00    0.00    0.00   42.00
06:04:56 PM   11   29.00    0.00   17.00   15.00    0.00    0.00    0.00    0.00   39.00
06:04:56 PM   12   25.00    0.00   18.00   13.00    0.00    0.00    0.00    0.00   44.00
06:04:56 PM   13   26.00    0.00   18.00   16.00    0.00    0.00    0.00    0.00   40.00
06:04:56 PM   14   26.00    0.00   17.00   13.00    0.00    0.00    0.00    0.00   44.00
06:04:56 PM   15   21.21    0.00   15.15   15.15    0.00    0.00    0.00    0.00   48.48
06:04:56 PM   16   37.25    0.00   27.45    9.80    0.00    0.00    0.00    0.00   25.49
06:04:56 PM   17   32.00    0.00   21.00   13.00    0.00    0.00    0.00    0.00   34.00
06:04:56 PM   18   30.30    0.00   19.19   14.14    0.00    0.00    0.00    0.00   36.36
06:04:56 PM   19   29.29    0.00   19.19   14.14    0.00    0.00    0.00    0.00   37.37
06:04:56 PM   20   27.55    0.00   16.33   14.29    0.00    0.00    0.00    0.00   41.84
06:04:56 PM   21   28.71    0.00   17.82   12.87    0.00    0.00    0.00    0.00   40.59
06:04:56 PM   22   26.00    0.00   17.00   13.00    0.00    0.00    0.00    0.00   44.00
06:04:56 PM   23   26.26    0.00   16.16   15.15    0.00    0.00    0.00    0.00   42.42
06:04:56 PM   24   25.00    0.00   18.00   13.00    0.00    0.00    0.00    0.00   44.00
06:04:56 PM   25   28.00    0.00   15.00   13.00    0.00    0.00    0.00    0.00   44.00
06:04:56 PM   26   23.76    0.00   14.85   11.88    0.00    0.00    0.00    0.00   49.50
06:04:56 PM   27   18.81    0.00   12.87    9.90    0.00    0.00    0.00    0.00   58.42
06:04:56 PM   28   19.00    0.00   12.00    9.00    0.00    0.00    0.00    0.00   60.00
06:04:56 PM   29    8.08    0.00    5.05    2.02    0.00    0.00    0.00    0.00   84.85
06:04:56 PM   30    8.00    0.00    8.00    4.00    0.00    0.00    0.00    0.00   80.00
06:04:56 PM   31    5.94    0.00    2.97    1.98    0.00    0.00    0.00    0.00   89.11
```

Ultimately, we discovered that PCIe/NVMe based SSDs can achieve better performance than SATA based SSDs, as shown in Figure 3-5. Although P3600 still did not reach its performance limits in this test, the workload servers did reach the bottleneck point, with more than 2x improvement in performance IOPS and 10x improvement in latency, thus meeting our expectations.

Considering power consumption and performance, we recommend using multiple PCIe/NVMe SSDs and distributing the workloads over them, as shown in Figure 3-6.We find that this solution dramatically improves CPU efficiency at the back-end micro-architecture layer, and that using several PCIe/NVMe SSDs greatly improves the latency of a single PCIe/NVMe SSD.

Figure 3-5: Performance Data of KV Database System based on Single PCIe/NVMe SSD on One Server

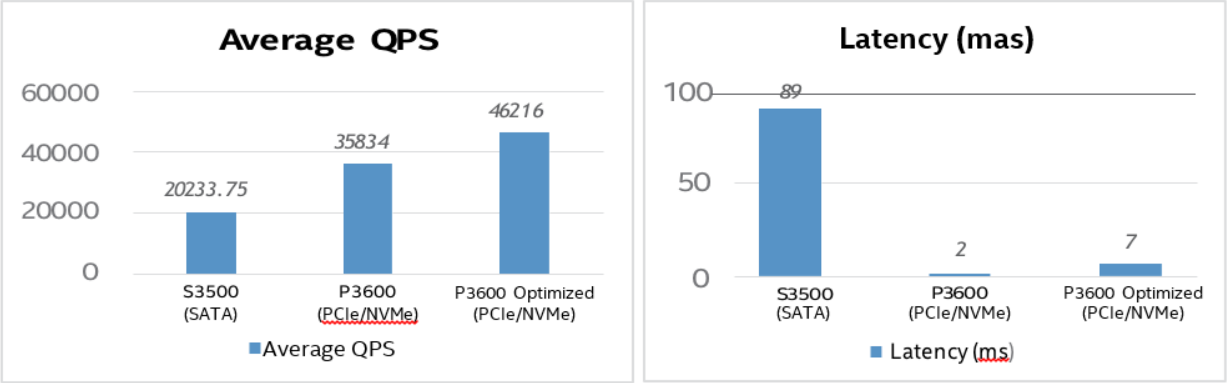
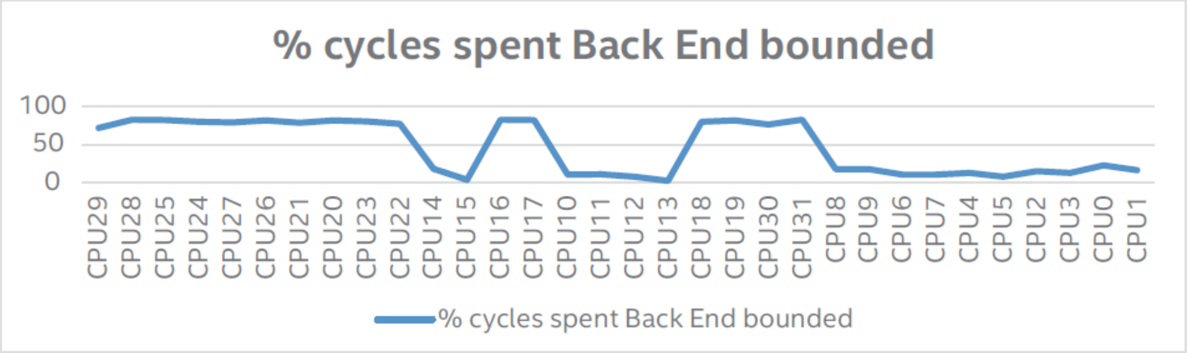


Figure 3-6: Performance Data of KV Database System based on Several NVMe SSDs on One Server



3.2 Application and Optimization of NVMe SSD in Baidu Relational Database (MySQL) System

3.2.1 Overview of Baidu Relational Database (MySQL) System

MySQL is an open source database and is extensively applied in the IPDC field. Baidu's O&M team has its self-defined MySQL system for use by all product lines. MySQL is mainly used to store structured data such as user information. MySQL is also provided externally as a service that demands higher performance in the emerging services-based public cloud.

3.2.2 Introduction of Original Solution

Baidu MySQL cluster has a similar hardware configuration to a KV database in that it adopts RAID 5 with several SATA SSDs; but it behaves differently in that the system disk adopts RAID 1 with 2 SATA HDDs, so it can provide reliability while maintaining performance. In the front end, there is a Master cluster, which is mainly responsible for encapsulating MySQL and providing a cache of the structured data. It corresponds to the data on the background MySQL nodes. Compression allows us to save disk space. The architecture is as shown in Figure 3-7.

3.2.3 PCIe SSD with NMVe Solution and Performance Optimization

Intel's PCIe SSD, P3600 Series, is installed to replace the existing solution, with one PCIe SSD instead of five SATA SDDs as RAID 5. The test environment is shown in Table 3-1, and the test covers the comparison between Intel® Xeon® E5 v2 and Intel® Xeon® E5 v3 platforms and environment comparison of PCIe SSD on Intel® Xeon® E5 v3 processor. Three MySQL operations, namely—insert, query, and update—are tested; the performance result are shown in Table 3-2.

The objective data indicates that as a result of upgrading the entire software stack, overall performance improved by13%, thereby meeting our guidelines for replacement. Intel PCIe SSDs have great performance potential and, in this environment, are far from reaching their limitations. Currently the performance bottleneck occurs in the CPU for many reasons. One example; when entire computing resources are jammed at data compression, the performance of the PCIe SSD cannot be fully realized. PCIe SSD's superior performance can impact the software by enabling the software architecture to adapt to the high IOPS performance of the PCIe SSD. Intel is currently looking for opportunities to improve the software architecture to gain further advantages.

Figure 3-7: Architecture of Baidu Relational Database (MySQL)

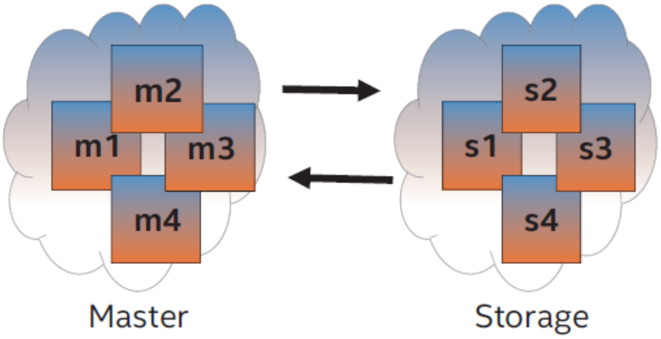


Table 3-1: Test Configuration

Configuration #	CPU	Memory	SSDs	IOPS Specified (W/R)
Congifuration1	E5-2620v2	96G	5 x 480GB SATA SSDs	56,000 / 80,000
Configuration 2	E5-2620v3	96G	5 x 480GB SATA SSDs	56,000 / 80,000
Configuration 3	E5-2620v3	96G	1 x 1.2TB PCIe SSD	56,000 / 80,000

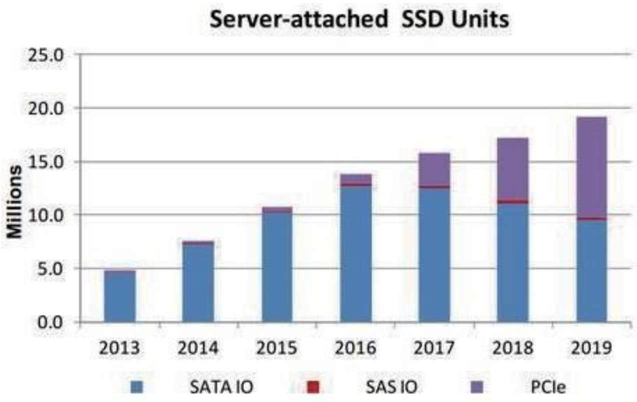
Table 3-2: Test Result

Operation	Test	Configuration 1	Configuration 2	Configuration 3
Insert	QPS	7,100	7,300	8,000
	Total IO_WRITE_KB	280M	240M	312M
	Total IO_WRITE_REQ	54,800	51,920	64,000
Select	QPS	6,000	7,000	8,000
	Total IO_WRITE_KB	152M	200M	210M
	Total IO_WRITE_REQ	8,400	12,000	11,500
Update	QPS	7,000	6,900	11,500
	Total IO_WRITE_KB	280M	235M	300M
	Total IO_WRITE_REQ	52,000	48,120	63,500

4 Summary

Intel and Baidu work closely to apply new technologies and develop innovative products that drive efficiency gains; this beneficial relationship is instrumental in supporting the rapid growth of data center markets. With the continuous technical improvement in SSD technology and products, their decreasing cost, and the industry's positive projected growth of SSD, we expect the complete conversion from traditional HDD storage to SSDs in the near future. (See Figure 4-1)

Figure 4-1: Baidu's Projected Growth in the Data Center SSD Market


Server-attached SSD Units
Millions
SATA IO
SAS IO
PCIe