Flash Only от Huawei – презентация новой программы по системам хранения данных

Алексей Июдин
iyudin.alexey@huawei.com

HUAWEI

Скорость работы СХД критически важна для функцинирования сервисов Заказчиков

Data is consolidated by Shanghai Customs every night. **200,000 batches** of data for customs declaration, clearance, and trade supervision

Customs clearance cannot be completed on time at **08:00 a.m.**, **detaining a large shipment of goods.**

Step 2: Risk data delivery from e-Administration (data extract)

Step 3: Results analysis by e-Customs (Data integration)

Step 1: Customs declaration, clearance through e-Port (data input)

Data consolidation

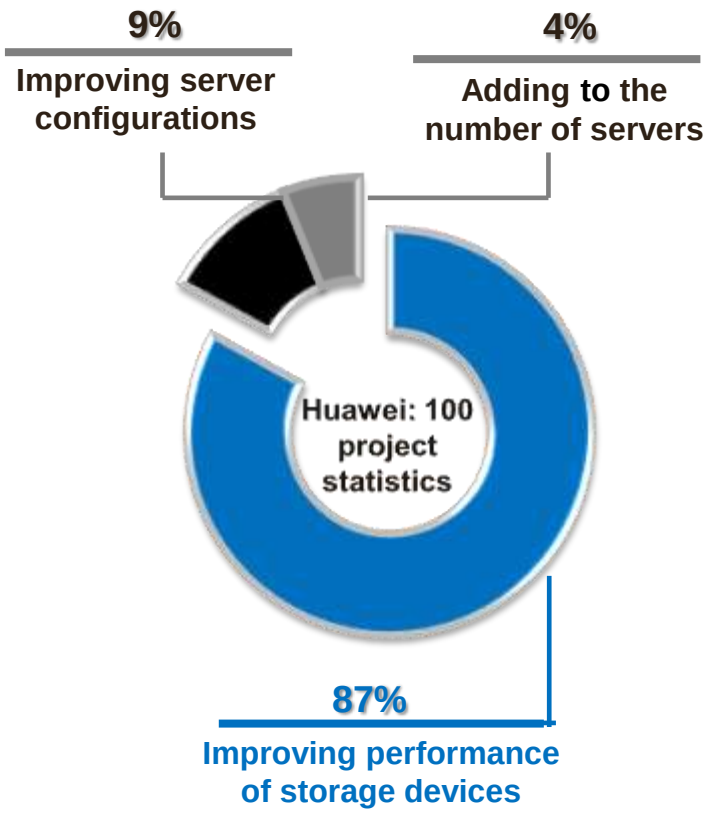
Must be completed between 00:30 a.m. – 06:00 a.m.

HDD storage

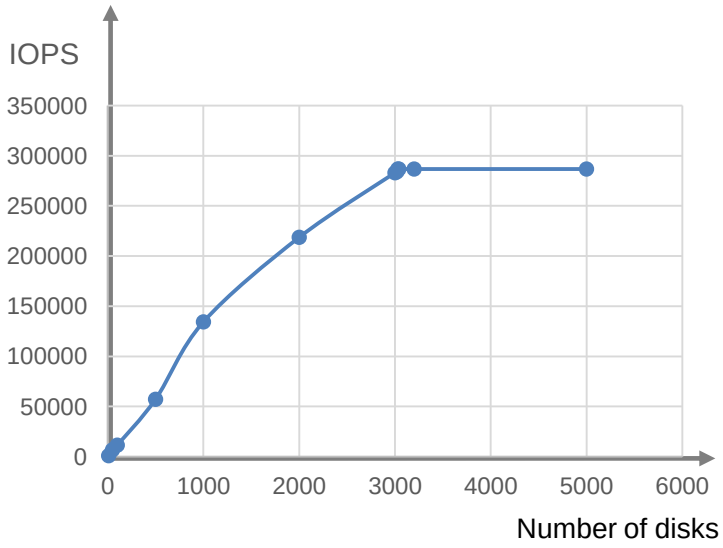
Data consolidation at night takes up to **7.5 hours**

Как можно попытаться улучшить работу инфраструктуры ЦОД

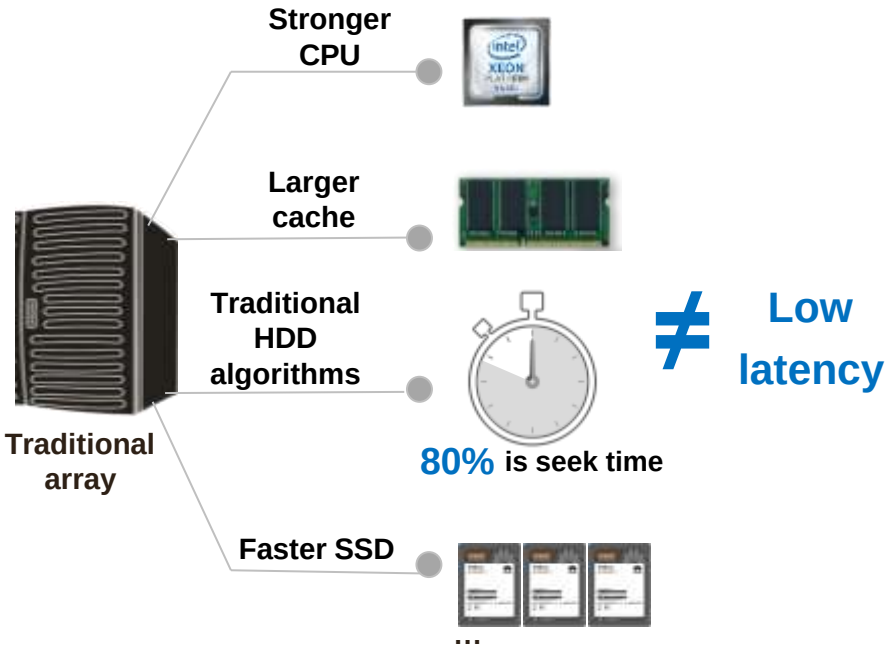
Solution 1
Unable to speed up system response using server-based improvement



Solution 2
Unable to solve performance bottleneck by simply stacking disks



Solution 3
Unable to lower latency using traditional arrays + SSDs



HUAWEI

В чем причины медленной работы БД?

Top 5 wait events in OLTP model

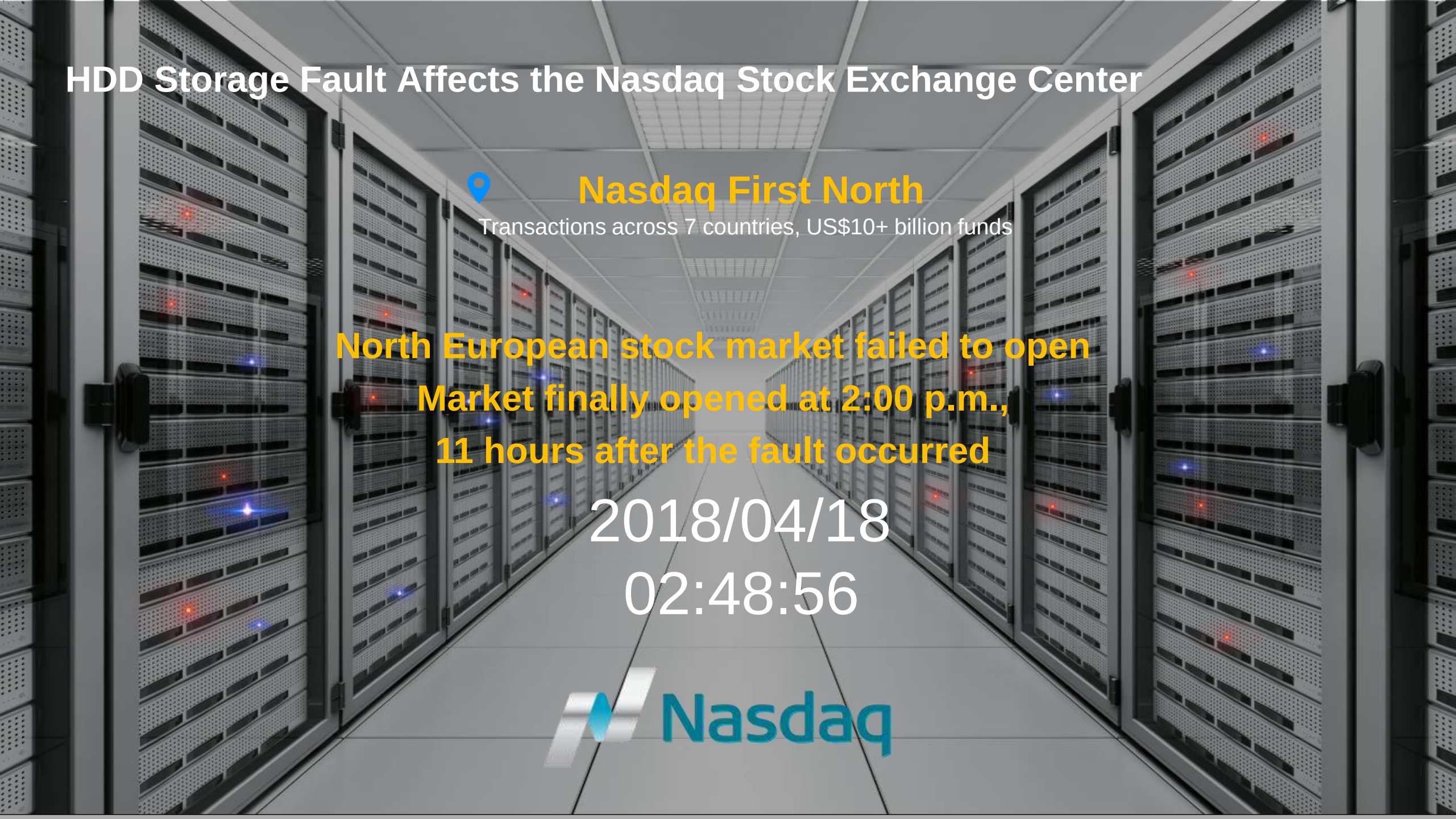
Event	Waits	Time(s)	Avg wait (ms)	% DB time	Wait Class
db file sequential read	426,694	6,492	15	96.87	User I/O
DB CPU		355		5.30	
db file parallel read	729	58	79	0.86	User I/O
log file sync	39,938	30	1	0.45	Commit
gc cr grant 2-way	56,407	18	0	0.27	Cluster

Top 5 wait events in OLAP model

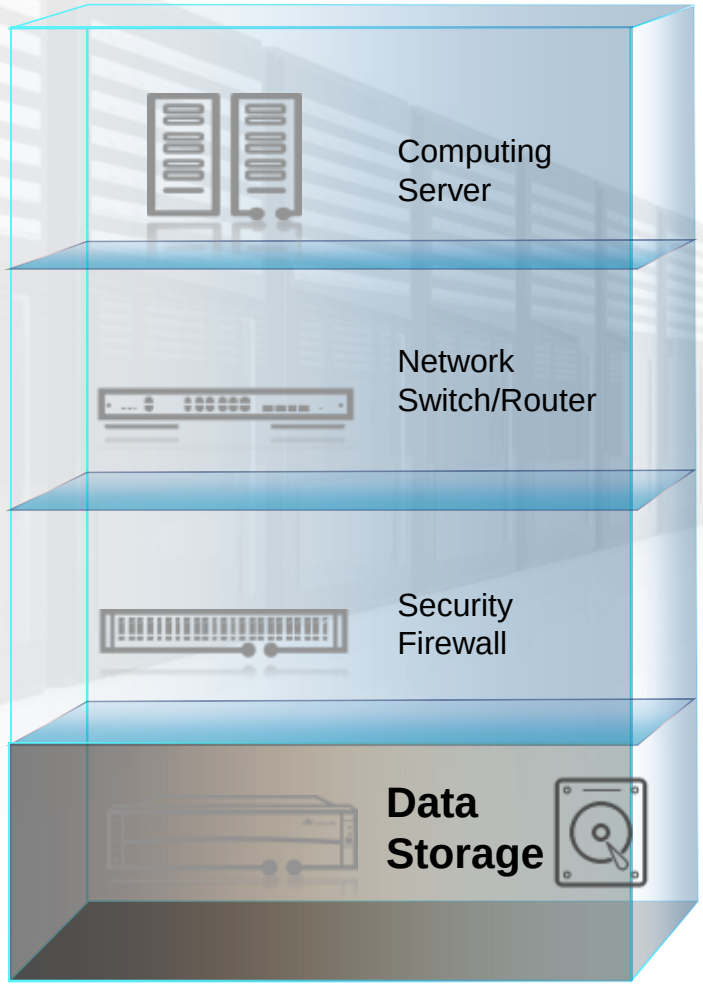
Event	Waits	Time(s)	Avg wait (ms)	% DB time	Wait Class
direct path read	4,604,339	567,141	123	63.67	User I/O
direct path read temp	1,955,162	147,298	75	16.54	User I/O
DB CPU		38,874		4.36	
db file sequential read	117,944	16,399	139	1.84	User I/O
direct path write temp	597,138	13,507	23	1.52	User I/O

Source: Oracle AWR performance analysis report on real projects

- **More than 80% of time is wasted on I/O delays, causing idle server resources**
- **Reducing I/O latency is the best solution for database efficiency improvement**


HDD Storage Fault Affects the Nasdaq Stock Exchange Center
Nasdaq First North
Transactions across 7 countries, US$10+ billion funds
North European stock market failed to open
Market finally opened at 2:00 p.m.,
11 hours after the fault occurred
2018/04/18
02:48:56
Nasdaq

HDD – это последний «механический» компонент ЦОД



Enterprise DC

HDD

1. Performance bottleneck

Limit rates of mechanical rotation cause disk performance far behind the improvement of CPU computing. The ultimate performance of a single disk is 400 IOPS.

2. Business availability blackhole

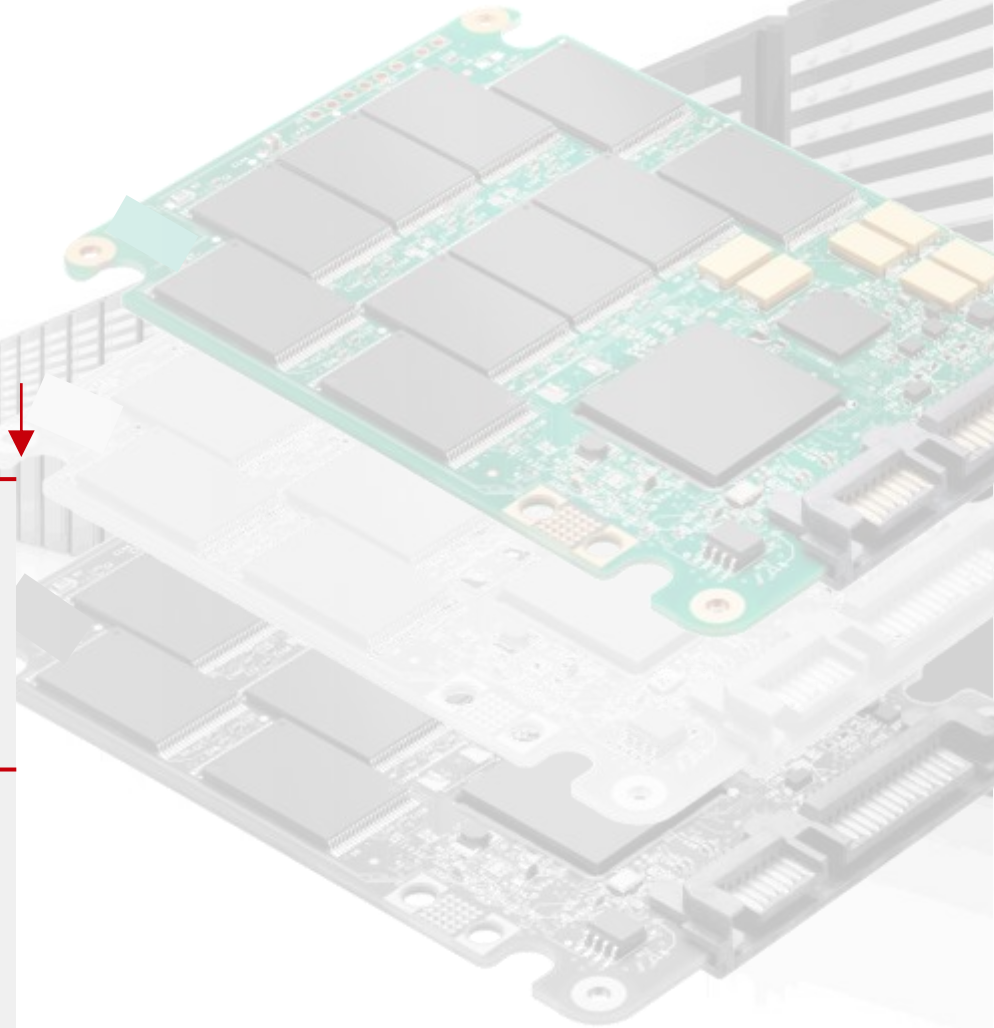
Inherent disk failure rates are high (2% to 10% per year), resulting in plenty of business breakdowns, data losses, and maintenance difficulties.

3. High power consumption

Power consumption of a single disk is up to 15 W. Considering factors such as heat dissipation and cooling, OPEX exceeds 50% of purchase costs.

SSD: революция в технологиях хранения данных

	10x higher performance	3-5x higher reliability	70% lower energy consumption
HDD	• IOPS:100 on average • bandwidth: Reads: 100 MB/s Writes: 50 MB/s-70 MB/s • Delay: 3+ ms	• Failure rate: 1%-5% • Vulnerable, sensitive to vibration and collision	• Power consumption: 5 W-13 W • Operating temperature: 5 ºC to 55 ºC
SSD	• IOPS: 10,000 on average • Broadband: 500+ MB/s reads and writes • Delay: below 0.1 ms	• Failure rate: below 0.5% • Collision, shock, or vibration cannot cause any mechanical fault because there is no movable mechanical component inside	• Energy consumption: down to 4 W–7 W • Operating temperature: 0 ºC to +70 ºC



SSD | All electronic component design

Что сдерживает распространения SSD

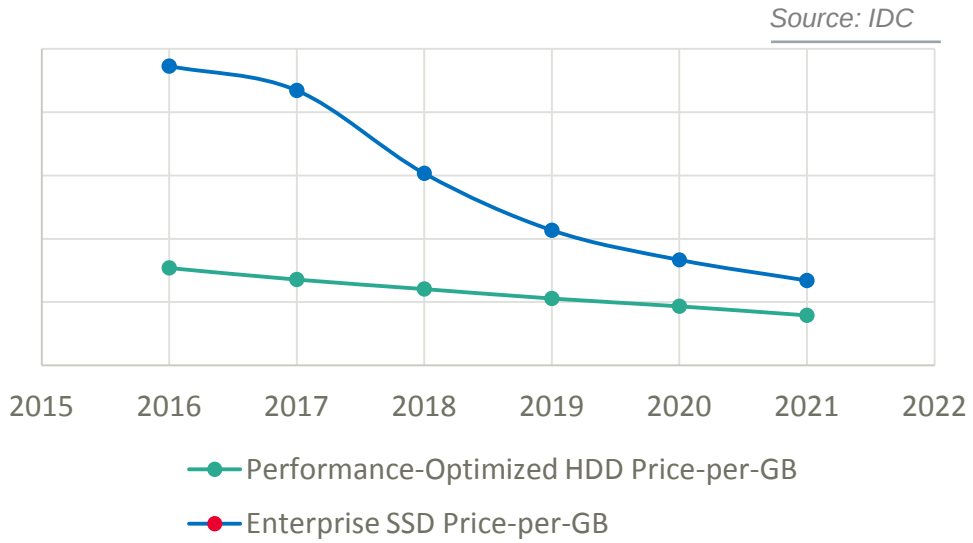
Price

Really high?

Tested data: SSD price is close to that of 10K SAS disks

SSD price is reduced from **3x** higher than that of SAS disks in 2016 to **1.6x** higher at the end of 2018

(If data reduction and TCO in the whole lifecycle are further considered, SSDs have great advantages.)



1

Tested data: SSD reliability is much higher than that of HDDs

Google

After the tracking of 10 types of SSDs for over 6 years, it can be found: The average annual failure rate of SSDs is **only about 1/3 that of HDDs.**

HUAWEI

Shipment data in the past decade shows that the return rate of SSDs is lower than 0.5%, **3x to 5x better** than that of HDDs.

2

Reliability

Really enough?

DWPD (one of SSD reliability indicators):

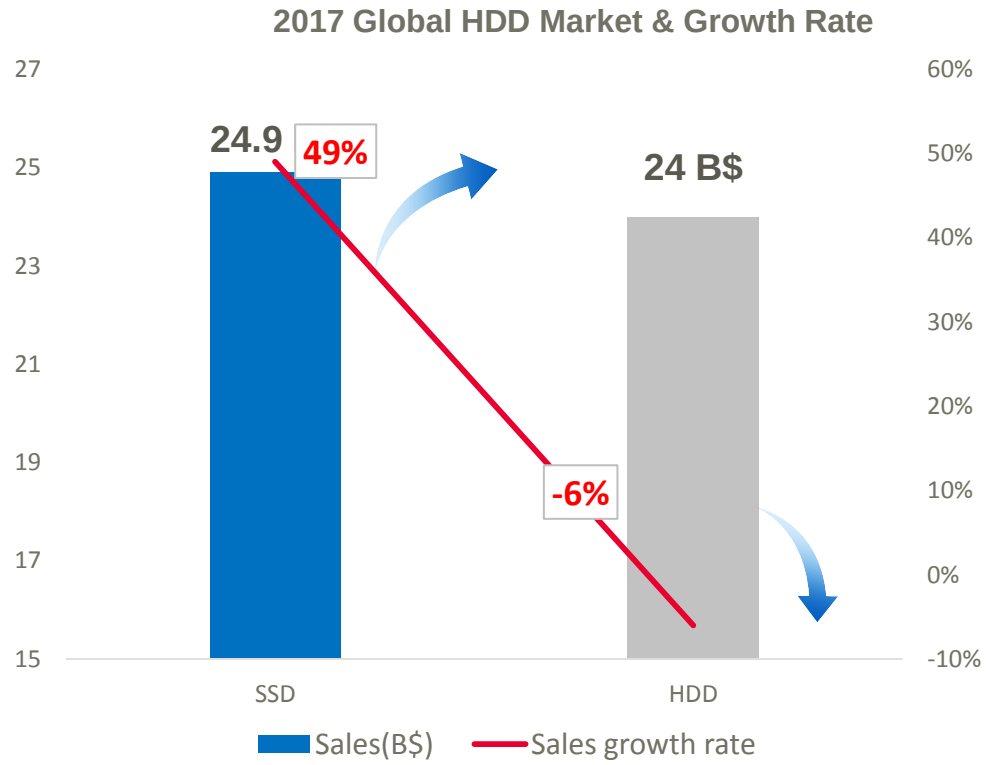
Number of drive overwrites each day in the expected service life

In the industry, if DWPD is 1, the SSD can ensure 5 to 7 years of service life in intensified write scenarios.

HUAWEI

Продажи SSD растут

SSDs are in and HDDs are on the way out. Overriding HDDs, SSDs are soaring into the stage center.



High-performance HDDs reach the stage end

Top HDD vendors steer their directions

Western Digital:

2018 Q1 saw the formal reform of mechanical disk services. **10K rpm and 15K rpm SAS** disks have been stopped releasing, declaring the withering of once super HDDs.

Seagate:

In Oct. 2016, **the last-gen 15K rpm HDDs** were released.

So far, Seagate's high-performance disks have all turned to SSDs.

Source: IDC 2017

Более 10 лет на рынке SSD

Internet

Telecom

Financial

Education

Government

Media

Enterprise

2005 — Start

2007 — 1st Gen
ES2000
PCIe SSD
SLC
128 GB/256 GB

...

2014 — 5th Gen
ES3000 V2
PCIe SSD
19/20nm MLC
600 GB - 3.2 TB

2016 — 6th Gen
ES3000 V3
NVMe SSD
15/16nm/3D
800 GB - 6.4 TB

2017 — 6th Gen
ES3000 V3
SAS SSD
3D NAND
800 GB - 3.84 TB

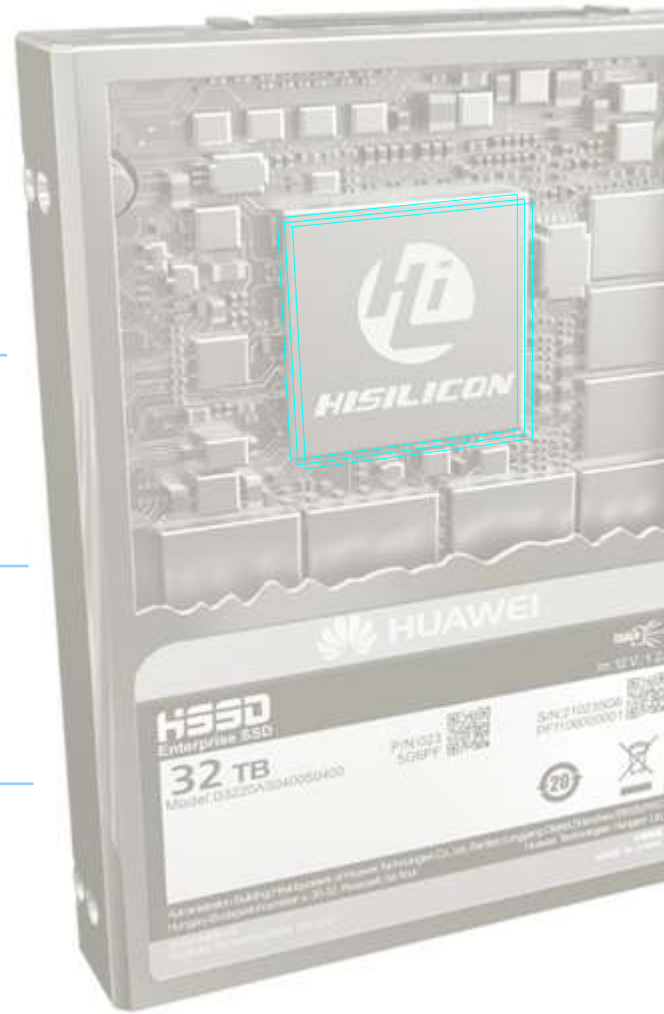
2018 — 7th Gen (Developing)
ES3000 V5
NVMe SSD/SAS SSD
3D NAND
800 GB - 32 TB

SSD массив начального уровня: OceanStor Dorado3000 V3

OceanStor Dorado3000 V3

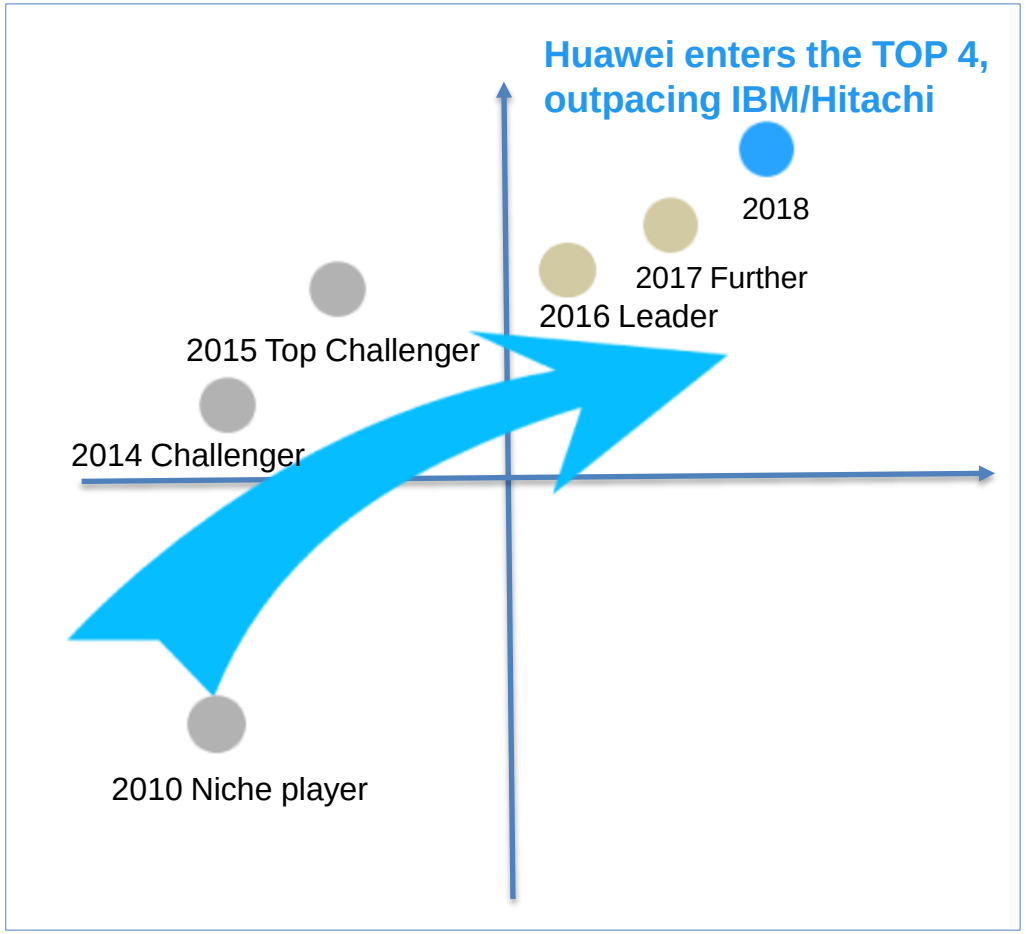
Lightning fast and rock solid

- 3x improvement in application performance
- 99.9999% business availability
- 75% savings in OPEX
- Non-disruptive migration

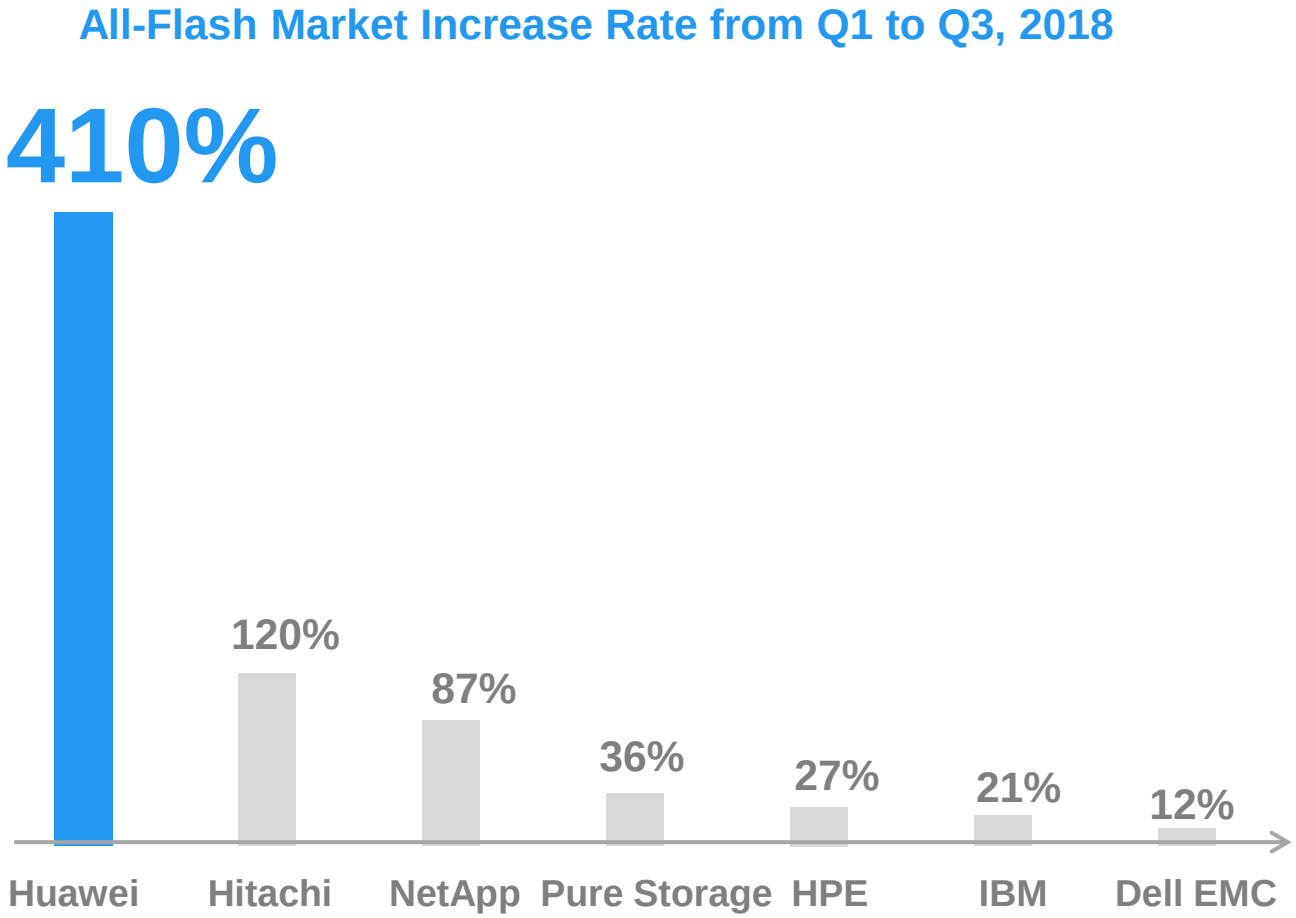


HUAWEI

Международное признание



Source: Gartner Storage Magic Quadrant for General-Purpose Disk Arrays 2018



Source: Gartner market share, 2018 Q3

HUAWEI

Что получают Заказчики после перехода на Dorado?

3x improvement in application performance

Top mining group in Australia

3x

After replacing NetApp FAS2000 in the production system, the VM scale is improved threefold.

Largest ISP in eastern France

30 mins -> 10 mins

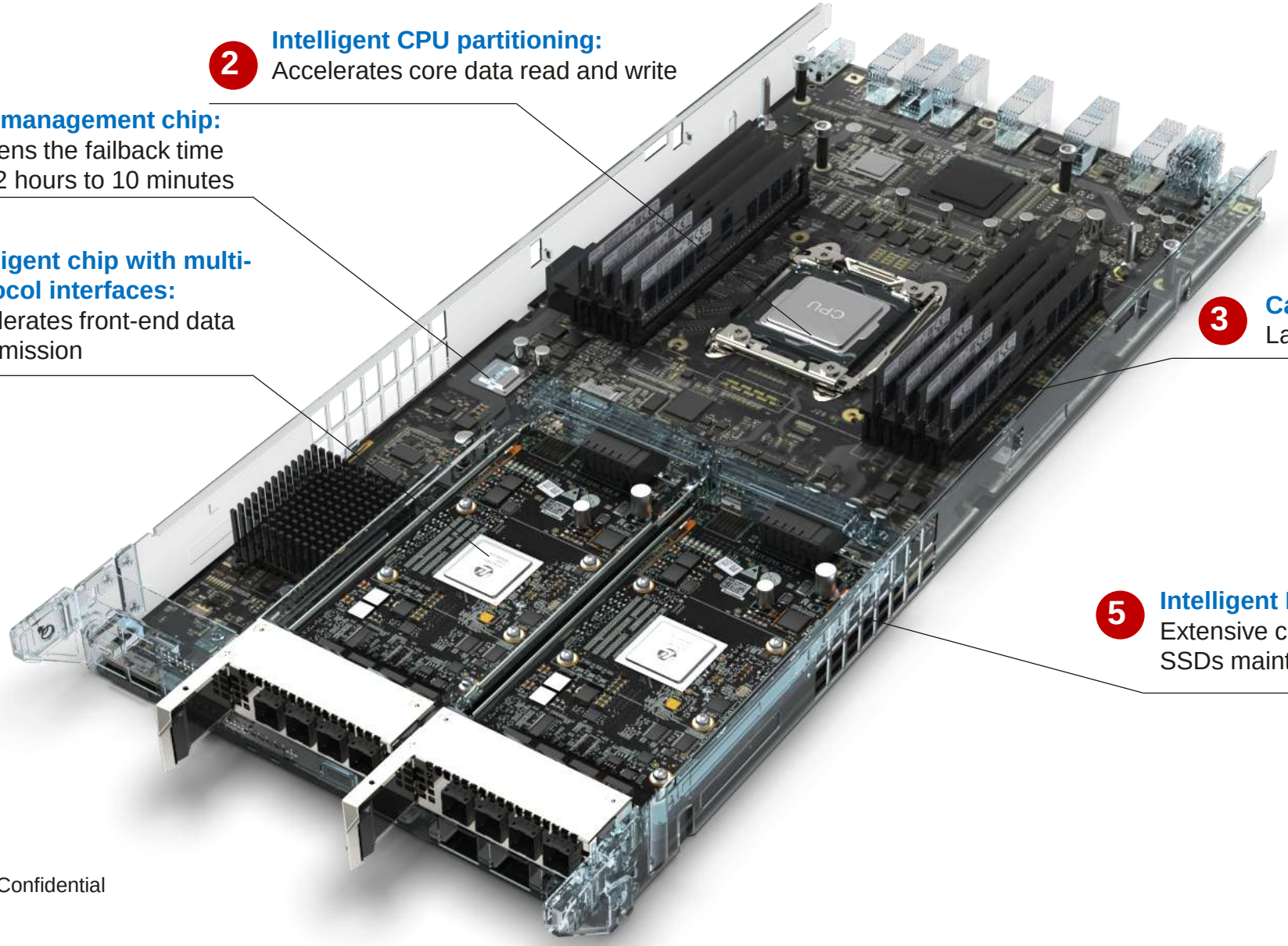
After replacing the EMC VNX virtualization platform, the deployment time of 100 VMs is shortened from 30 mins to 10 mins.

Largest airport in Spain

1/3

After replacing the HP XP9500 in the operations system, the decision-making time is shortened by one-third.

Как это устроено



2 Intelligent CPU partitioning:
Accelerates core data read and write

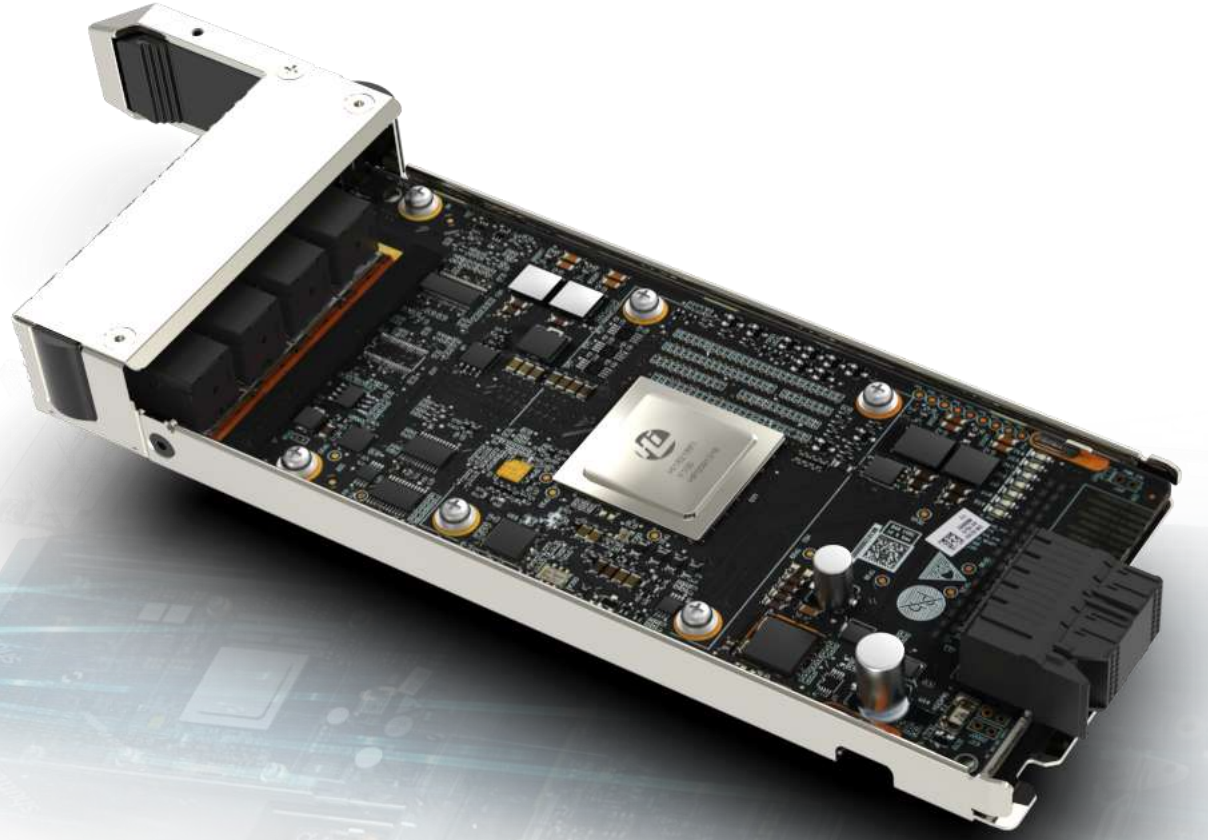
4 BMC management chip:
Shortens the failback time from 2 hours to 10 minutes

1 Intelligent chip with multi-protocol interfaces:
Accelerates front-end data transmission

3 Cache:
Latency fluctuation rate < 2%

5 Intelligent FlashLink®
Extensive coordination between controllers and SSDs maintains a stable latency of 0.5 ms

Как это устроено



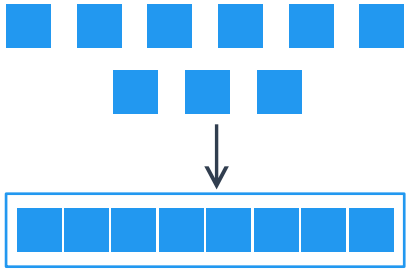
- Industry-leading interfaces support **32G, 16G, and 8G Fiber Channel as well as 100, 40, 25, and 10 Gb Ethernet** (on-demand)
- The chips handle protocol parsing that would otherwise require CPU processing, thus reducing data access latency by 10%

Как это устроено

- Comprehensive and intelligent fault management technology
- Fault locating accuracy rate is improved to 93%.
- Failback time is shrunk **from 2 hours to 10 minutes**.

FlashLink® - технология работы с SSD

Sequential writes of large blocks

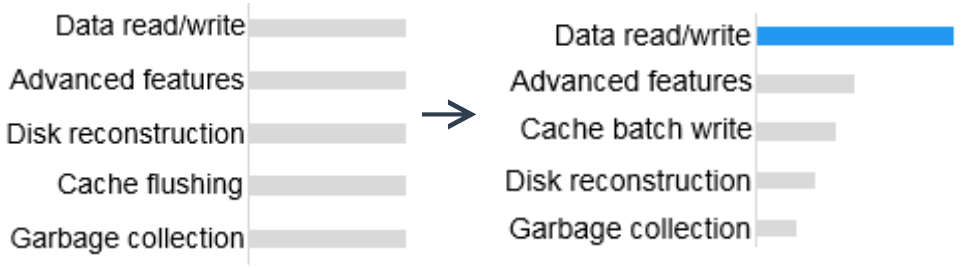


Discrete writes of multiple small blocks

↓

Sequential write of one large block
Less write amplification

I/O priority adjustment



All I/Os handled chronologically with the same level of priority

↓

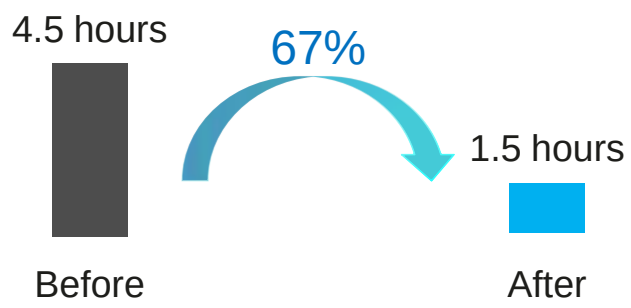
The system always gives data read/write I/Os top priority, which ensures the lowest latency for data read/write.

Controllers automatically sense and synchronously adjust data layouts in SSDs to reduce performance loss and guarantee a stable latency.

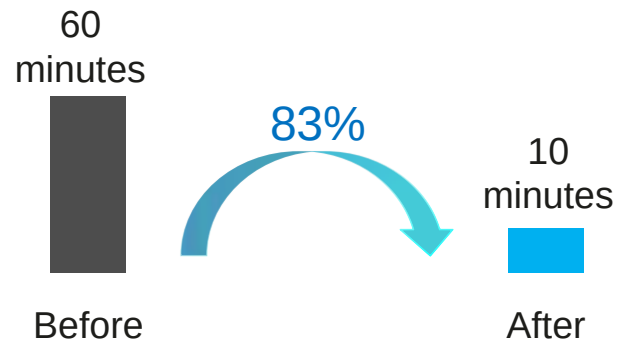
Пример перехода на AFA массив в производстве

BYD: China's largest private carmaker and a leader in new-energy vehicles

Time taken to extract BW data



Inventory transfer time of thousand spare parts



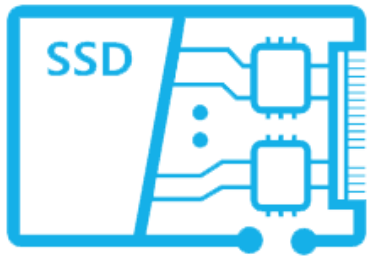
ROCK SOLID

24/7 always-on

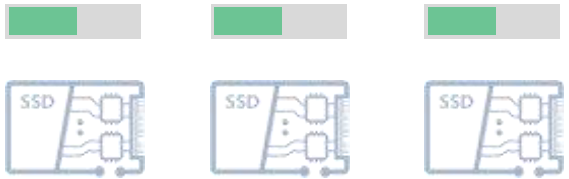
2,000+ инсталляций

Используемы технологии: SSD Core Technologies

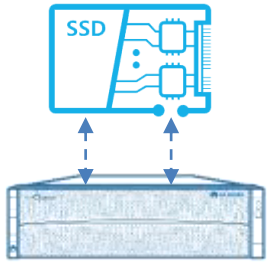
Industry-leading Reliability



SSD Core Tech



Wear leveling/Anti–wear leveling, providing longer SSD service life



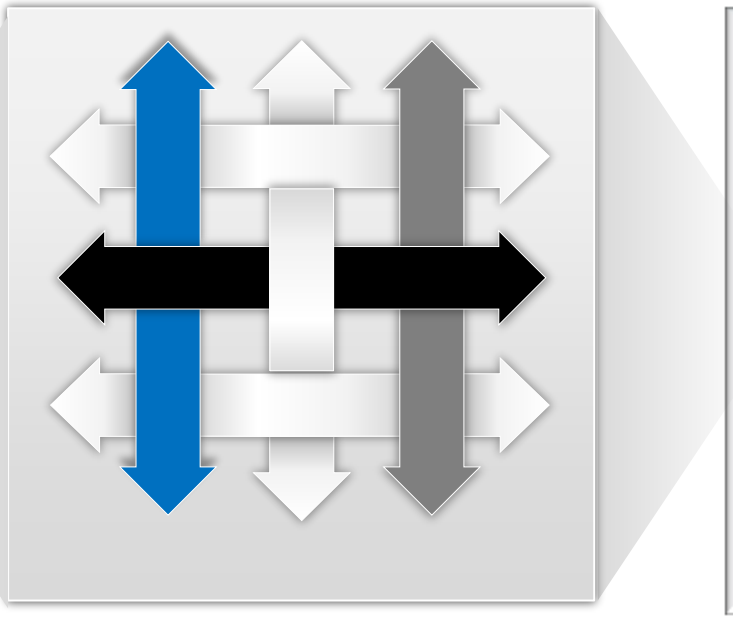
SSDs and system-level RAID technology, ensuring solid reliability

Data inspection algorithm, avoiding data distortion

LDPC algorithm, with 3x better error correction capability

Используемы технологии : RAID на уровне диска

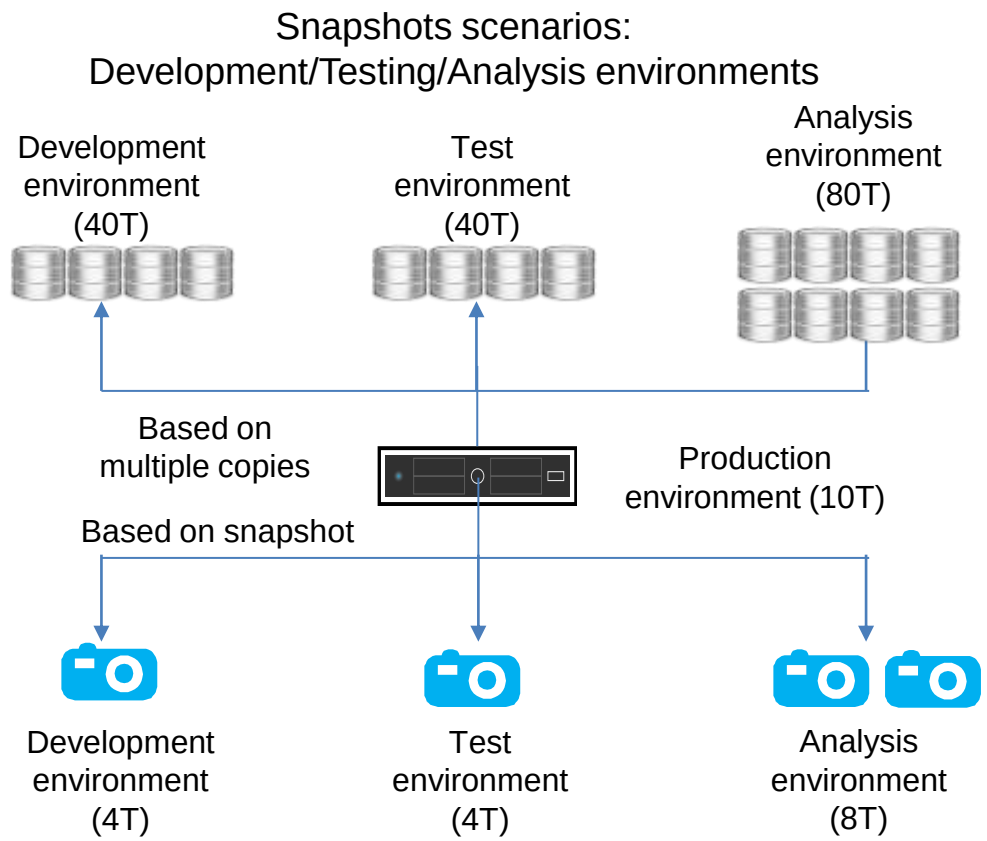
RAID 4 is supported in SSDs to ensure data reliability

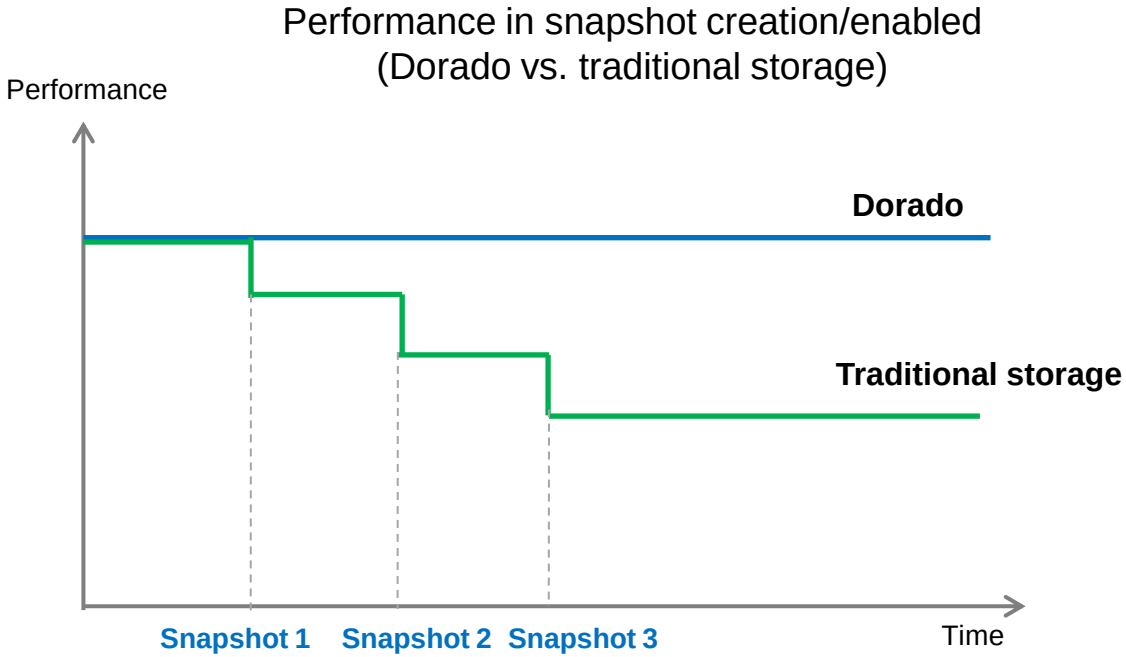


Dorado supports RAID 5/6/TP, tolerating simultaneous failures of up to three disks

- At the end of service life, SSDs cannot support single disk recovery using RAID protection technology (such as two channel failures). System-level RAID groups are invoked to recover data to the disk.
- Data in the faulty disk is moved to an operating SSD, ensuring no change to intra-disk OP, performance, and reliability.

Используемые технологии: Snapshots без снижения производительности





- No need of several copies, improving capacity utilization
- Less copy data volume, mitigating the performance loss of the production system

- Stable during snapshot creation
- Stable while snapshot is enabled
- Stable if multiple snapshots are continuously created with the interval of several seconds

HUAWEI

Резервные копии в облако

Data center 1

- Active-active data protection ensures uninterrupted data services.
- Continuous data protection within seconds ensures zero data loss

Data center 2

- No backup software or gateway, saving investment
- Immediate availability of snapshot-based copies and minute-level service recovery

Cloud DR center

- Backup data on the cloud and cloud-based data restoration
- Unified on-premises and cloud-based management, simplifying O&M

Service recovery duration

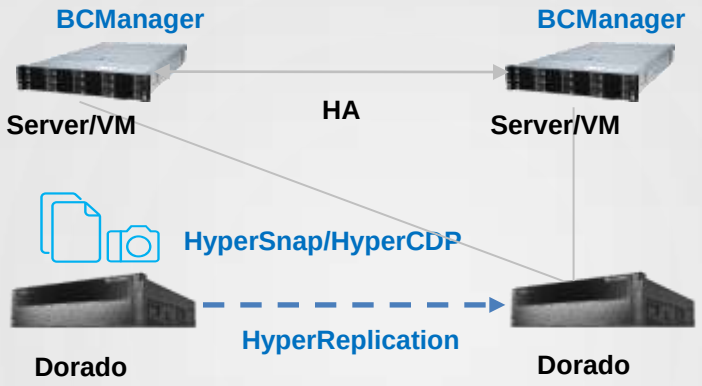
Second-level → Minute-level → Hour-level

100% without service interruption or data loss

Сценарии отказоустойчивости

Integration of DR and backup

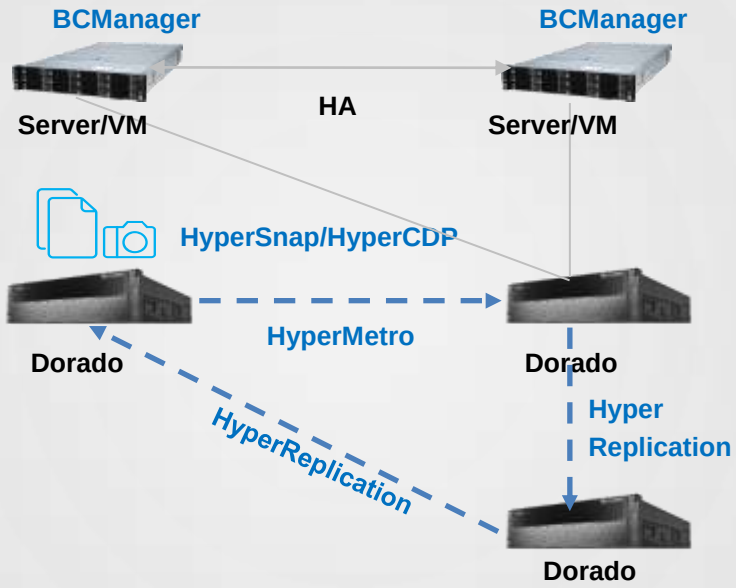
Scenario: Healthcare/Manufacturing



- HyperSnap/HyperCDP: Fast backup within seconds
- HyperReplication: Asynchronous remote replication for DR
- BCManager: DR/Backup policy configurations

3DC: Active-active + DR and backup integration

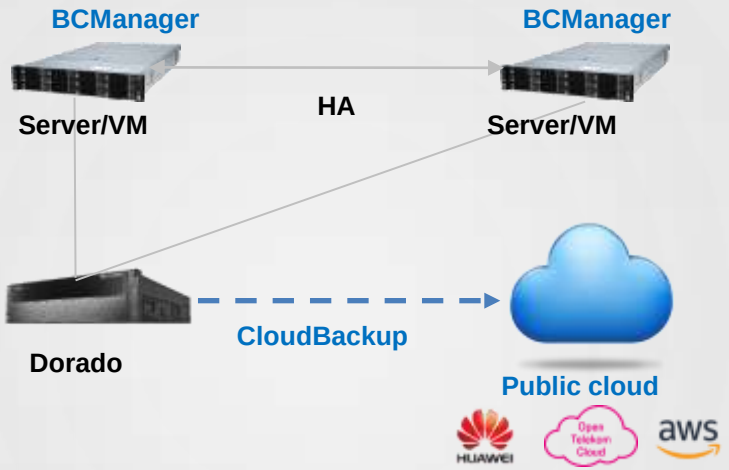
Scenario: Healthcare/Manufacturing



- The active-active function ensures high availability and seamless expansion to the 3DC layout
- HyperSnap/HyperCDP: Fast backup within seconds
- BCManager: DR/Backup policy configurations

CloudBackup

Scenario: Low-cost backup for enterprises

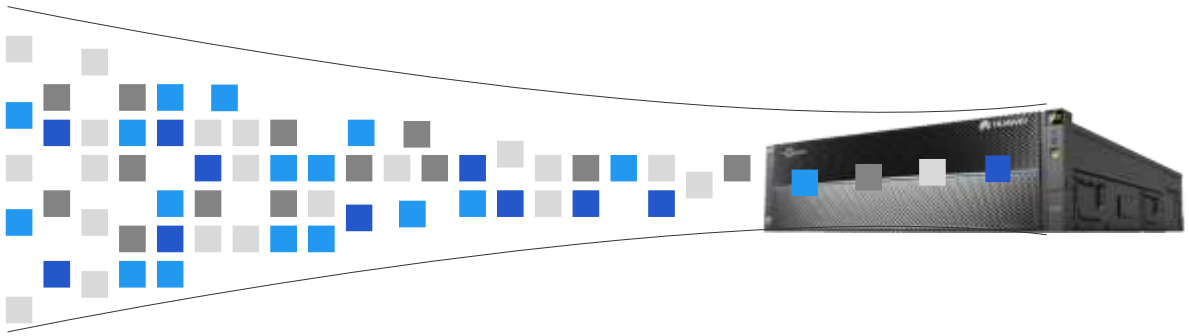


- CloudBackup: Gateway-free cloud-based backup enables services to be started on the cloud within minutes*
- BCManager: DR/Backup policy configurations

Note: CloudBackup will be supported in the first half of 2019 (services started on the cloud: HUAWEI CLOUD - GA on March 30, 2019; Huawei jointly-operated cloud - GA on June 30, 2019)

HUAWEI

Дедупликация и компрессия

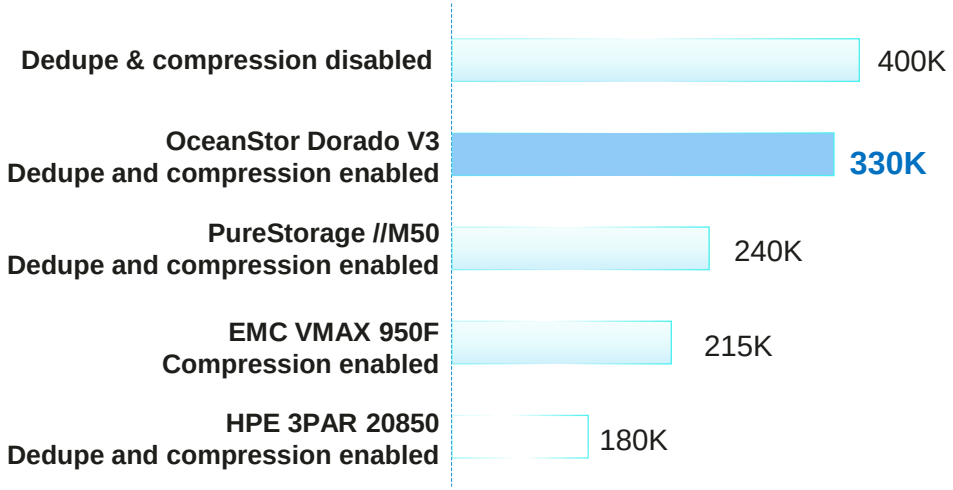


Dorado: Dedupe and compression deliver 75% OPEX savings.

Space 85% ↓ Power 74% ↓

Field-test performance of main AFAs

Test conditions: Dual-controllers, RAID6, 80% capacity occupation, and 1 ms latency



Others: Dedupe and compression compromise the performance a lot.

The impact on Dorado IOPS is the smallest after enabling the dedupe and compression. With the industry-leading performance, Dorado deals with peak workloads with ease.

HUAWEI

Следование стандартам

Compatible with 300+ mainstream storage and mainstream IT infrastructures in the industry, 6,000+ pages of interoperability matrixes

Note: For details, see Huawei Storage Interoperability Navigator

HUAWEI

OceanStor Dorado3000 V3 All-Flash Storage

Lightning-Fast

3x higher service performance

- ✓ Huawei-developed SSDs and chips
- ✓ FlashLink intelligent algorithm

Rock-Solid

99.9999% service availability

- ✓ Multi-level reliability technology
- ✓ RAID-TP technology, tolerating three disk failures
- ✓ Gateway-free active-active design and CDM solution

Cost-Efficient

75% OPEX reduction

- ✓ Inline deduplication/compression

Lightning Fast Rock Solid

HUAWEI

Flash Only

1 : 1 SSD : HDD*

Same capacity, same price

Till December 31, 2019

*10K SAS HDD

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and organization for a fully connected, intelligent world.