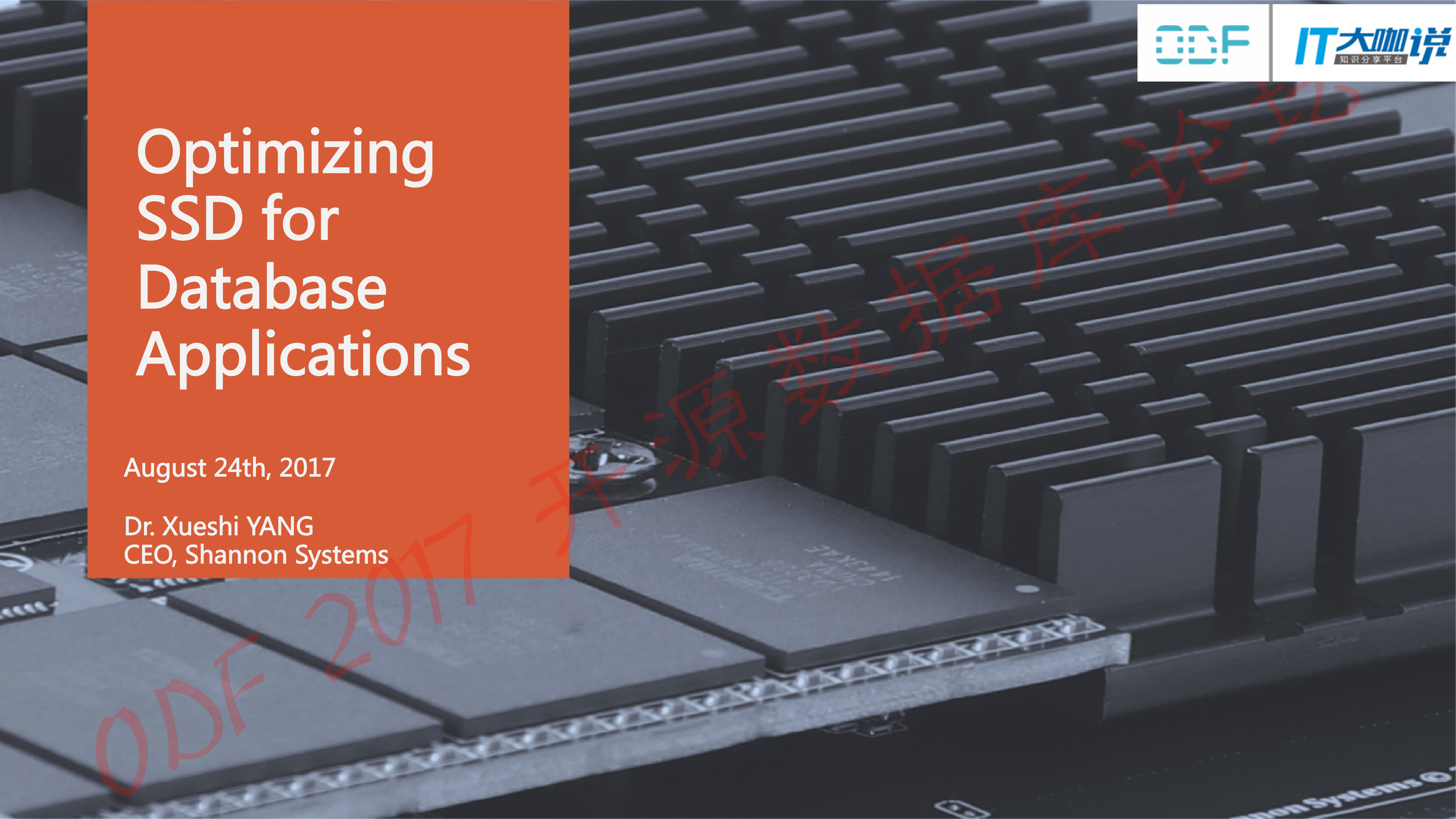


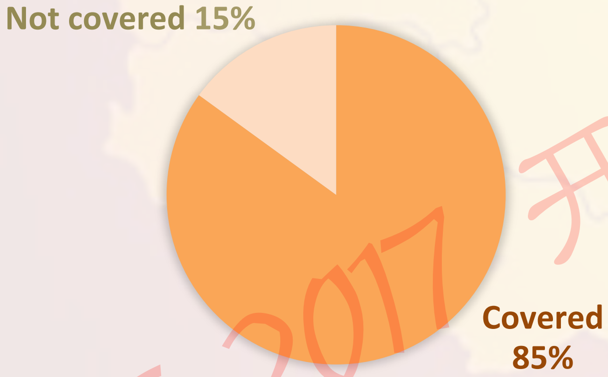
Optimizing SSD for Database Applications

August 24th, 2017

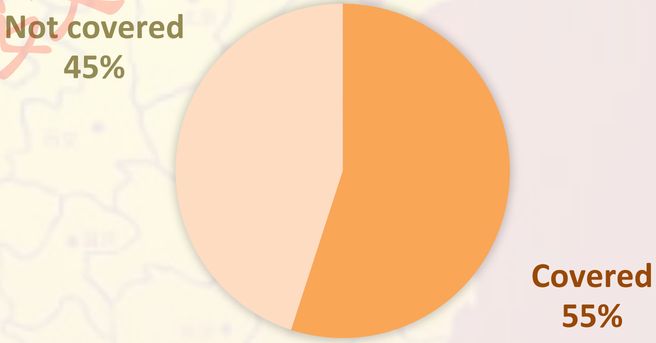
Dr. Xueshi YANG
CEO, Shannon Systems

Shannon Systems

宝存科技



Shannon SSDs inside ~85% of internet companies



We power more than 50% of e-commerce databases

喫 喝 玩 樂

大众点评 dianping.com

口碑网

五粮液

酒仙网 www.Jiuxian.com

去哪儿?
Qunar.Com
聪明你的旅行

途牛旅游网 Tuniu.com

大麦 damai.cn

唱 唱吧 最时尚的手机KTV

衣 食 住 行

唯品会 vip.com

聚美优品 JUMEI.COM

饿了么

美团 meituan.com

lianjia 链家 .com

房天下 Fang.com

中国铁路

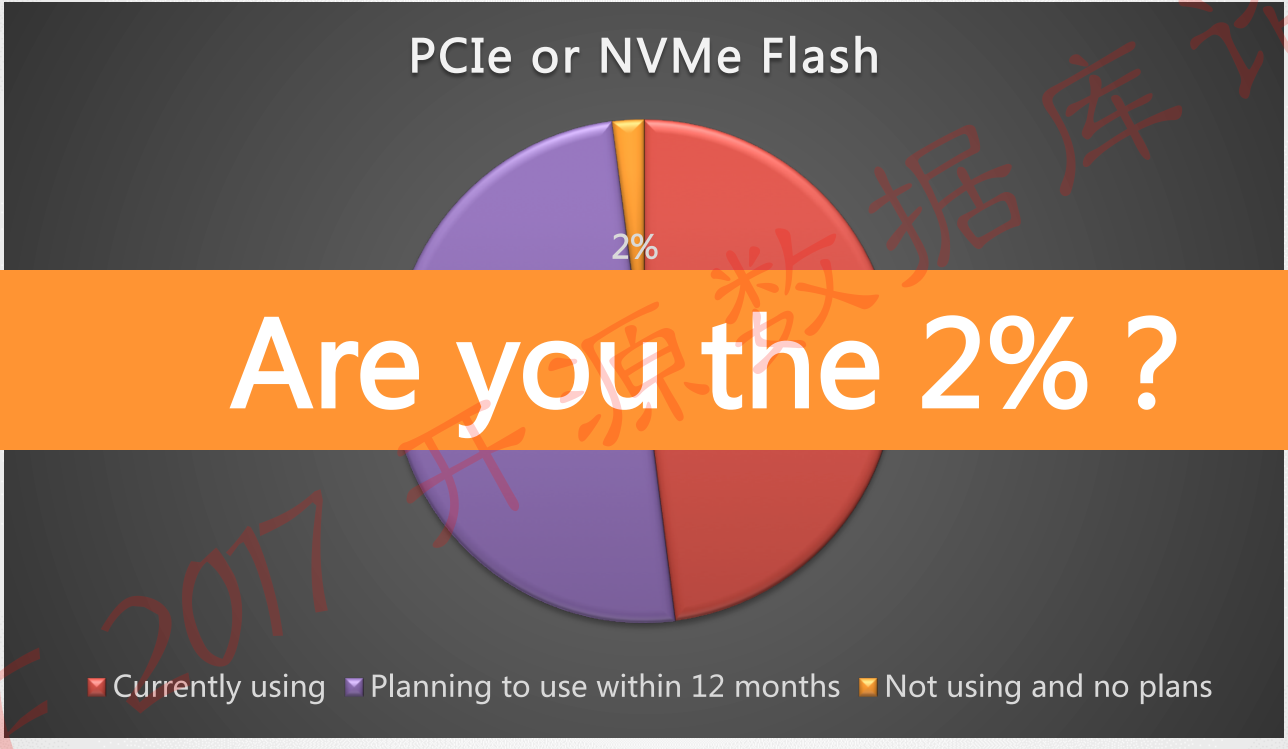
Alibaba Group 阿里巴巴集团

JD.COM 京东

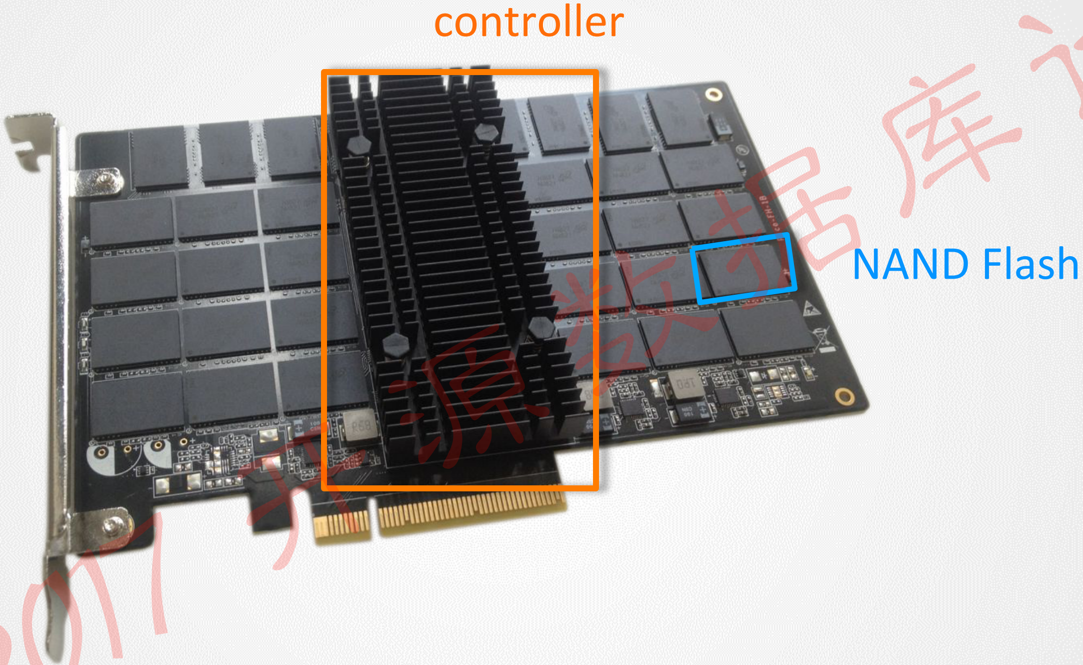
or find your significant other...

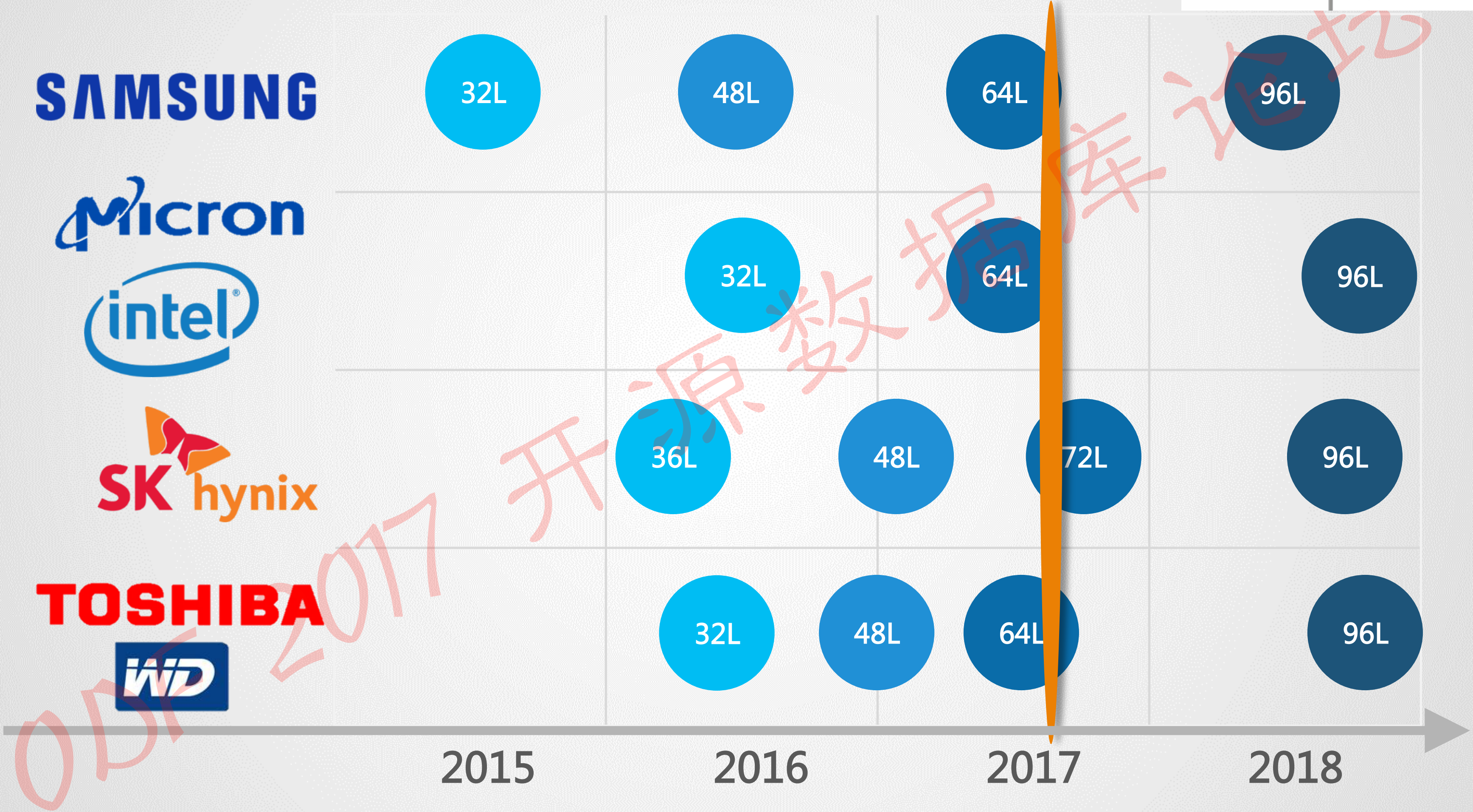
珍爱网
zhenai.com

百合 baihe.com
实名婚恋网开创者


PCIe or NVMe Flash
2%
🟥 Currently using
🟪 Planning to use within 12 months
🟧 Not using and no plans

Are you the 2% ?


controller
NAND Flash

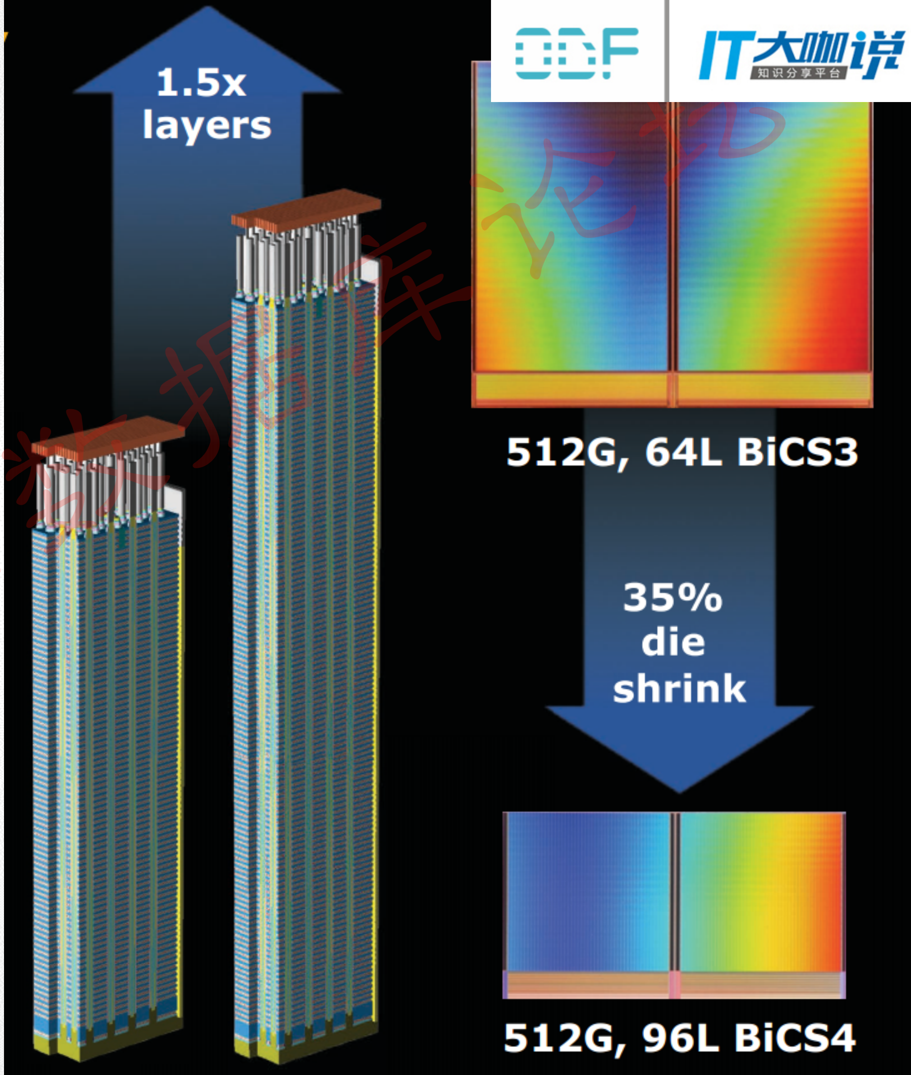

SAMSUNG
32L
48L
64L
96L
Micron
intel
32L
64L
96L
SK hynix
36L
48L
72L
96L
TOSHIBA
WD
32L
48L
64L
96L
2015
2016
2017
2018

3D NAND Flash

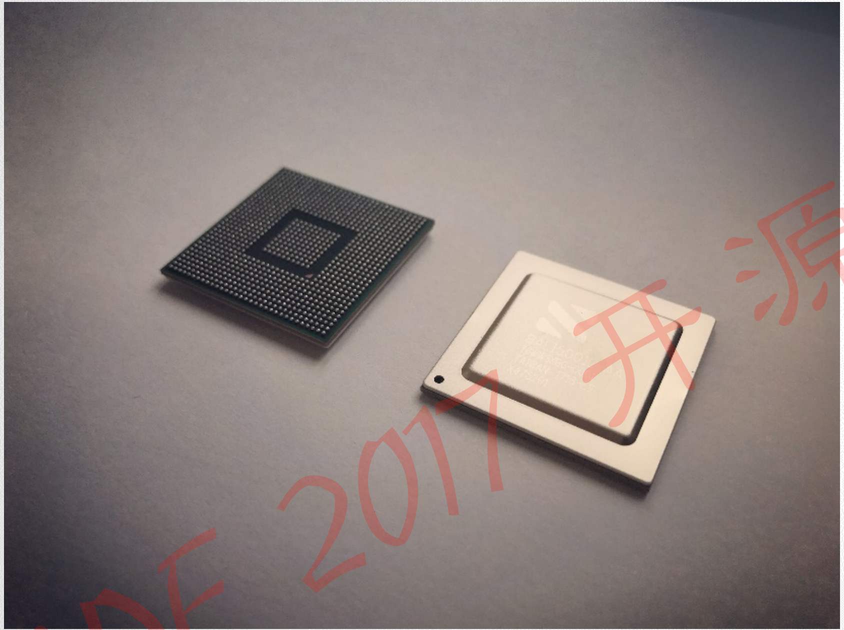
- 96 layer NAND
- 4 bits/cell

512Gb per die
1TB per single BGA package

Source: WDC



5Th generation Shannon Direct-IO controller

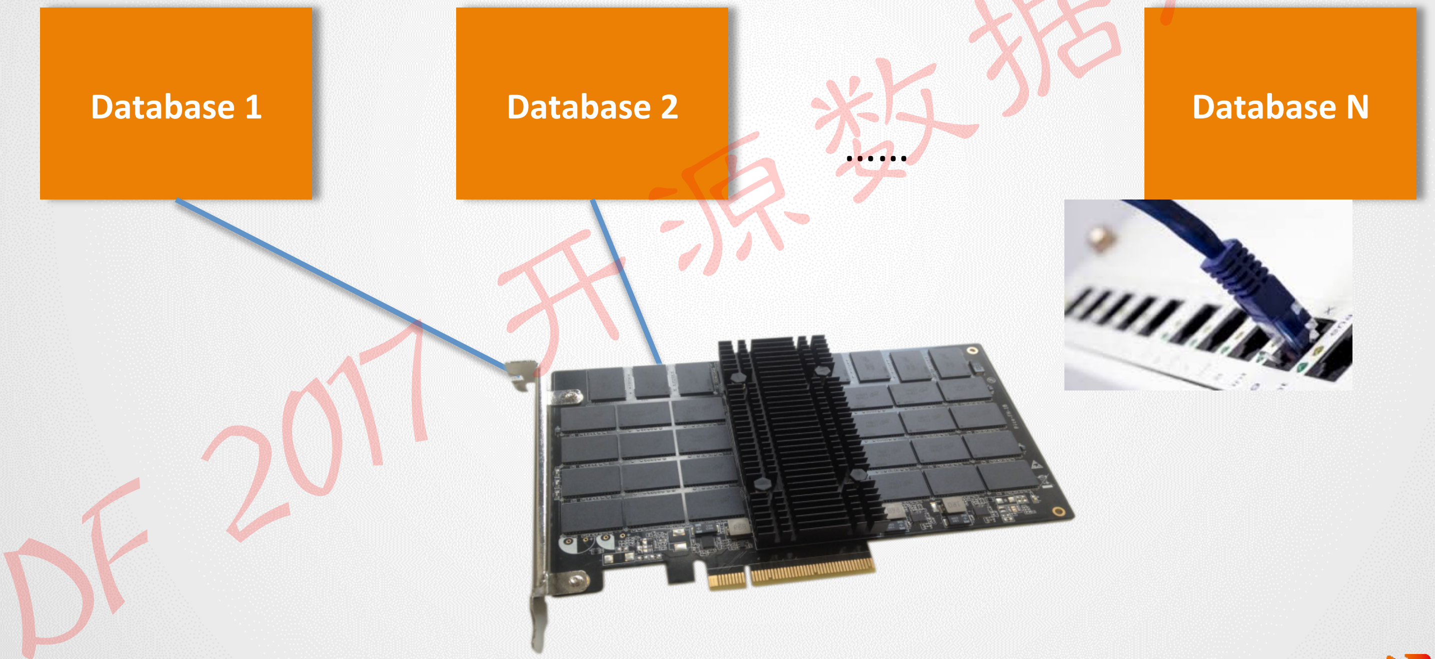


Tailor-made FFSA controller

- 1 Million IOPS
- 36TB capacity
- micro-second access latency

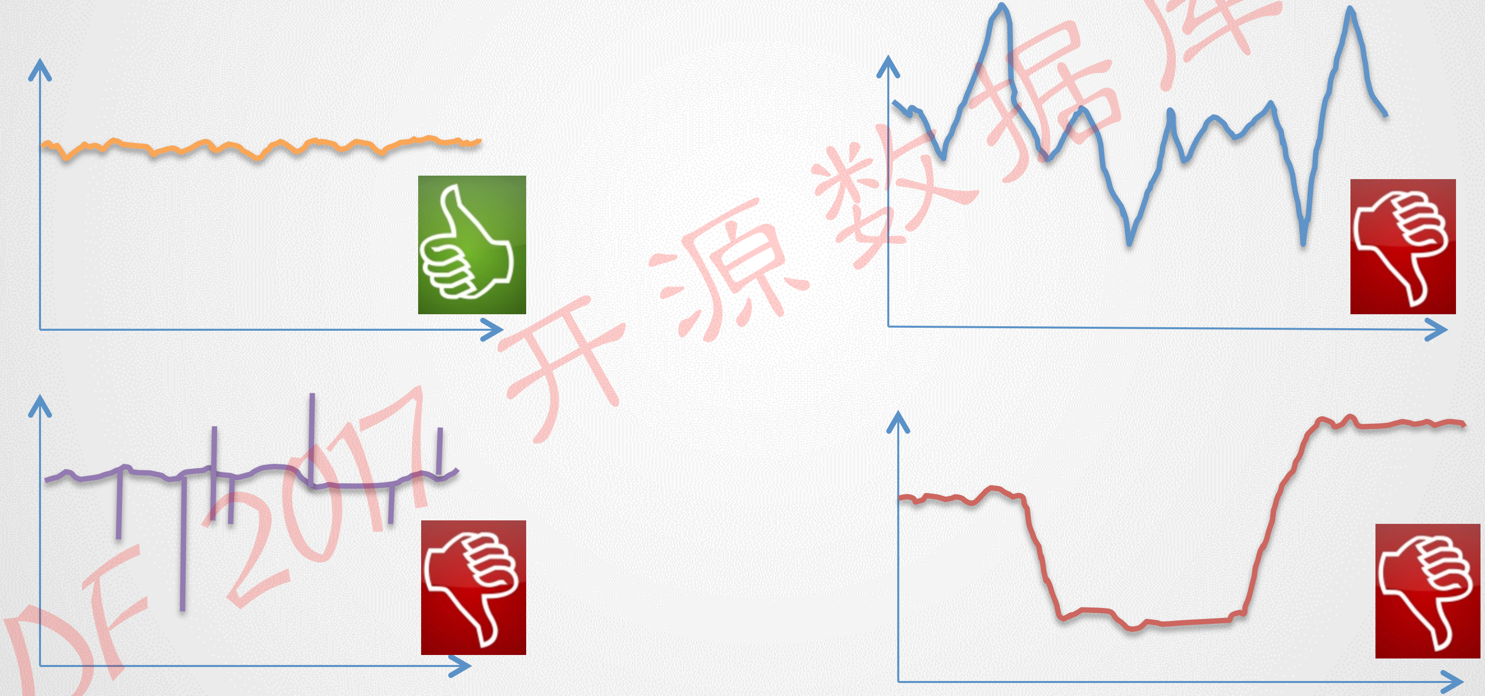
Sharing by multiple Databases

- Capacity and performance sharing



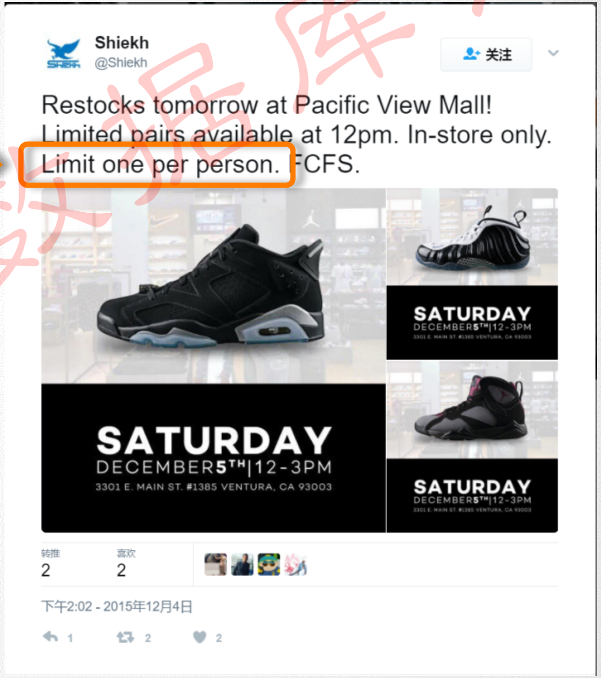
SERVICE
LEVEL
MAX
NO
LOW
MED
HIGH
MAX
SATISFACTION

IO consistency



IO separation

Approach: IOPS limitation per instance



Shiekh
@Shiekh
关注

Restocks tomorrow at Pacific View Mall! Limited pairs available at 12pm. In-store only. Limit one per person. FCFS.

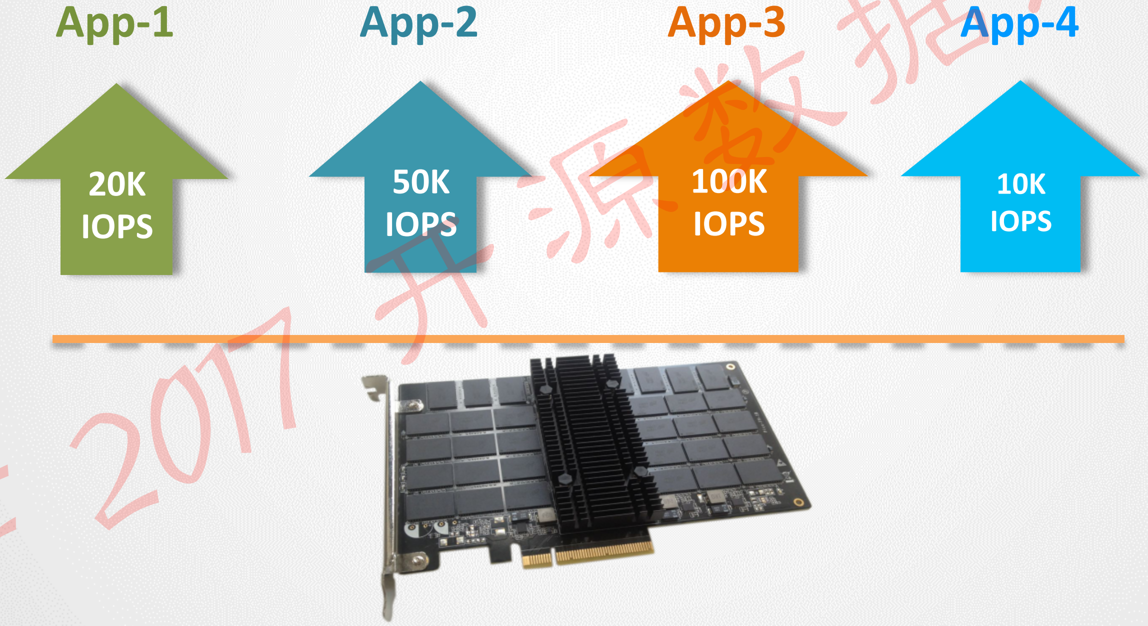
转推 2 喜欢 2

下午2:02 - 2015年12月4日

1 2 2

IO separation

- Limit the maximum available resources for each applications



App-1

20K IOPS

App-2

50K IOPS

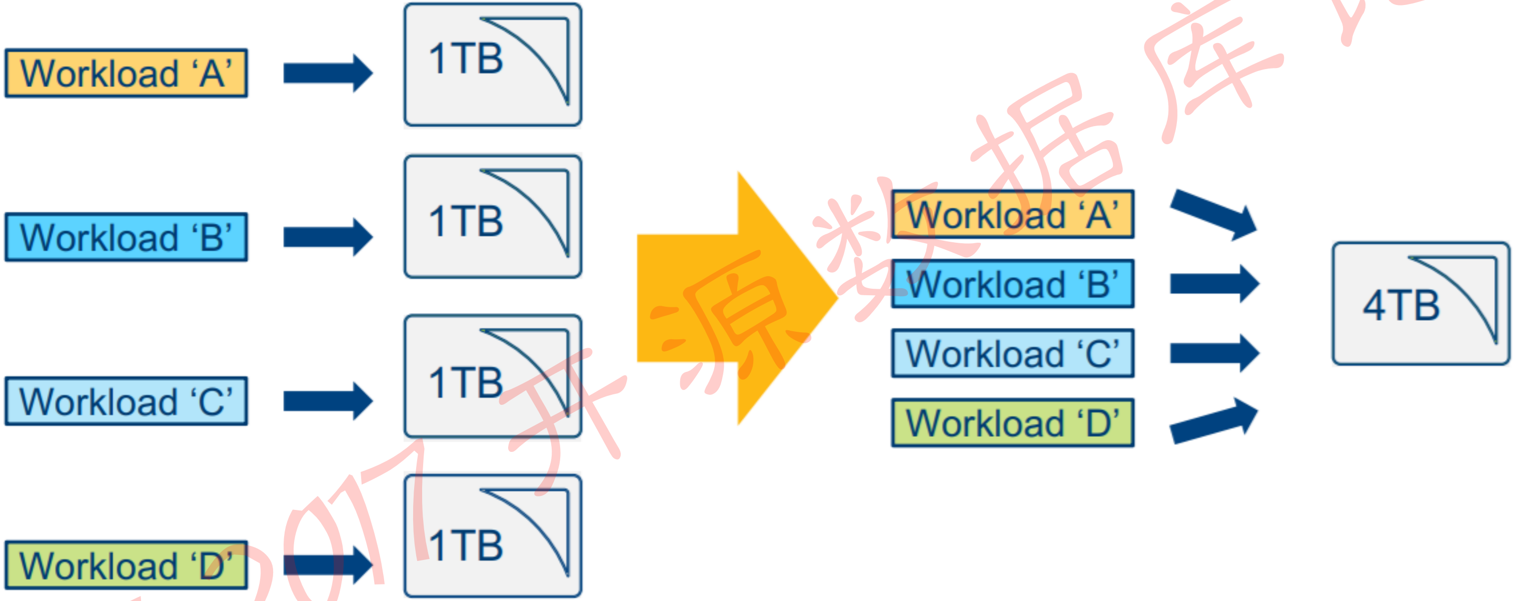
App-3

100K IOPS

App-4

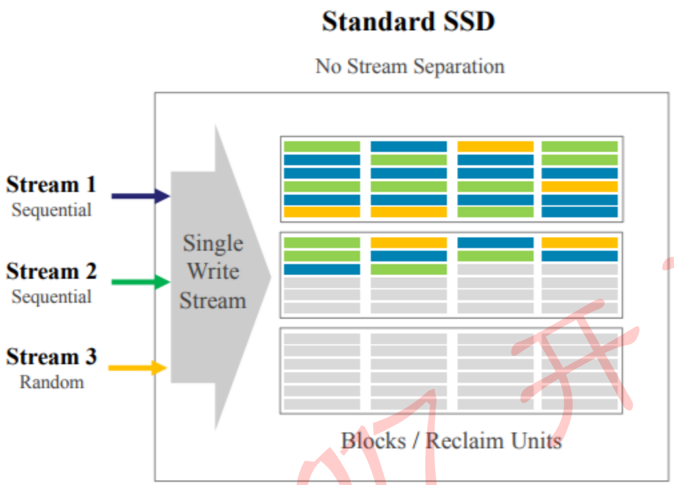
10K IOPS

IO separation via streams



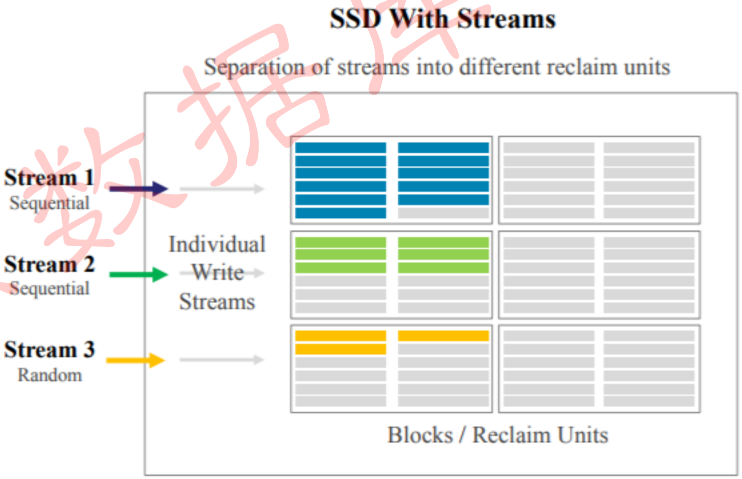
Source: Landsman,FMS2017

IO separation via streams



Mixed data needs garbage collection to reclaim blocks

Higher Write Amplification



Separated data can be trimmed or self-invalidated to reclaim blocks

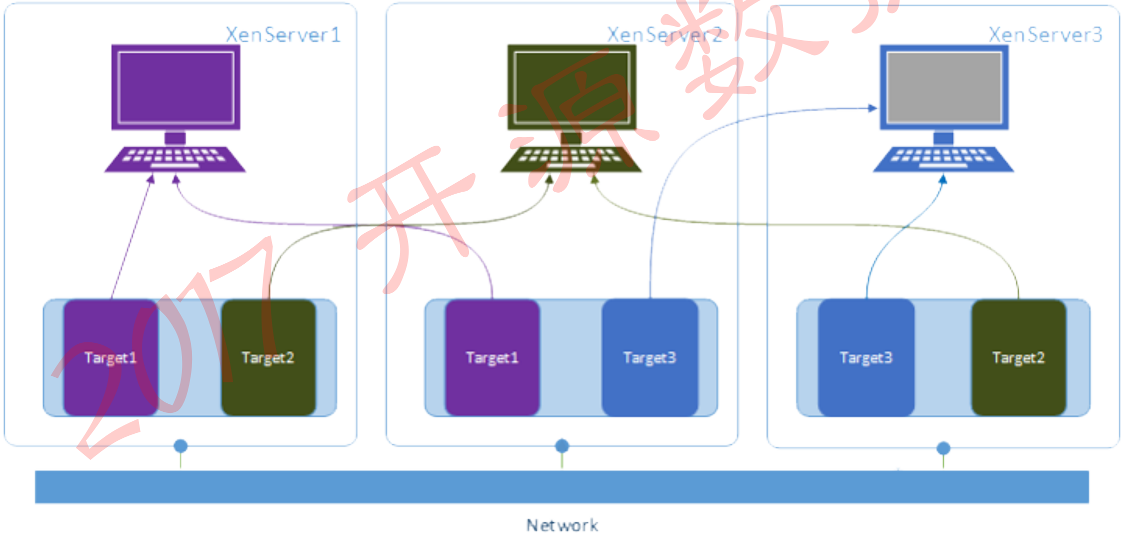
Lower Write Amplification

Source: Landsman,FMS2017

Networked SSDs

- IO access across nodes or space is the key enabling component

Multiple-Nodes Hyper-Converged Infrastructure for XenServer

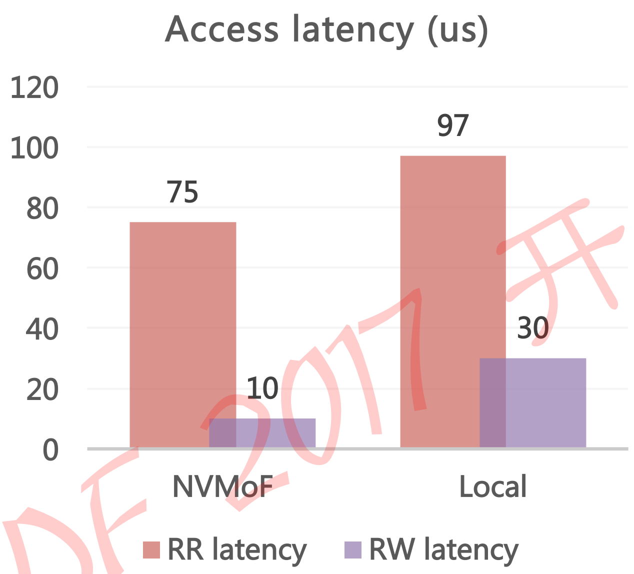


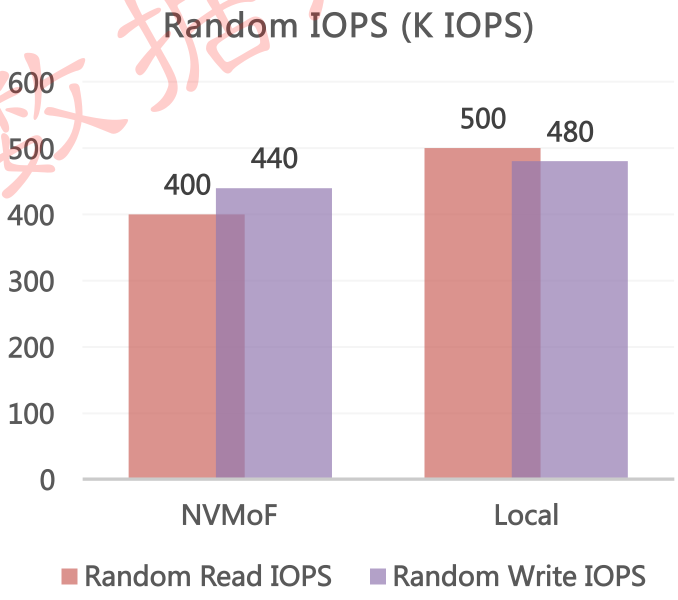
Networked IO

- RDMA technology enables SSD sharing across nodes with minimum latency overhead
  - e.g. NVMeoF
- Proprietary software layer ensures IO determinism across space
  - Access latency becomes node agnostic
  - Storage pool is constructed across multiple nodes
  - SSD pool becomes HA (highly available)

Networked IO

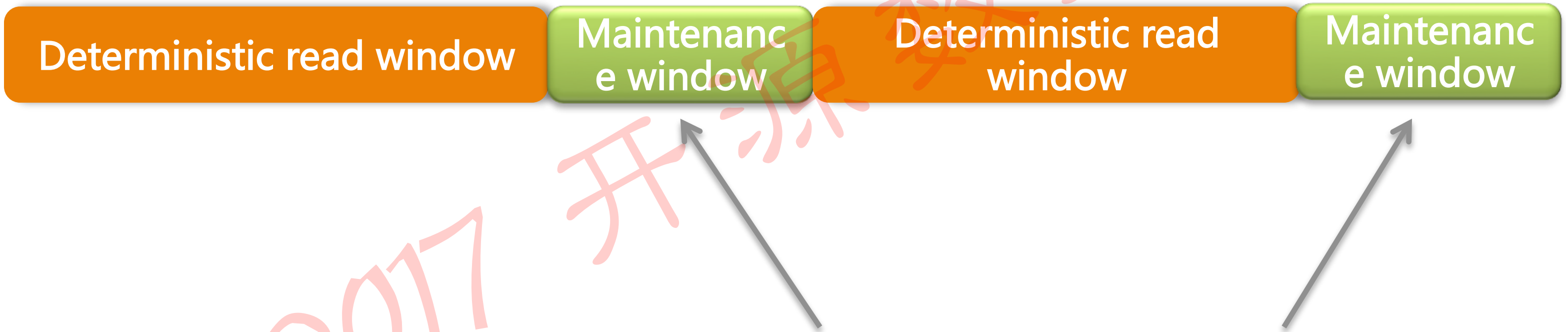
- IO access through high speed network via RDMA





IO consistency –RW separation

- Maintenance window to guarantee read access latency



Deterministic read window

Maintenance window

Deterministic read window

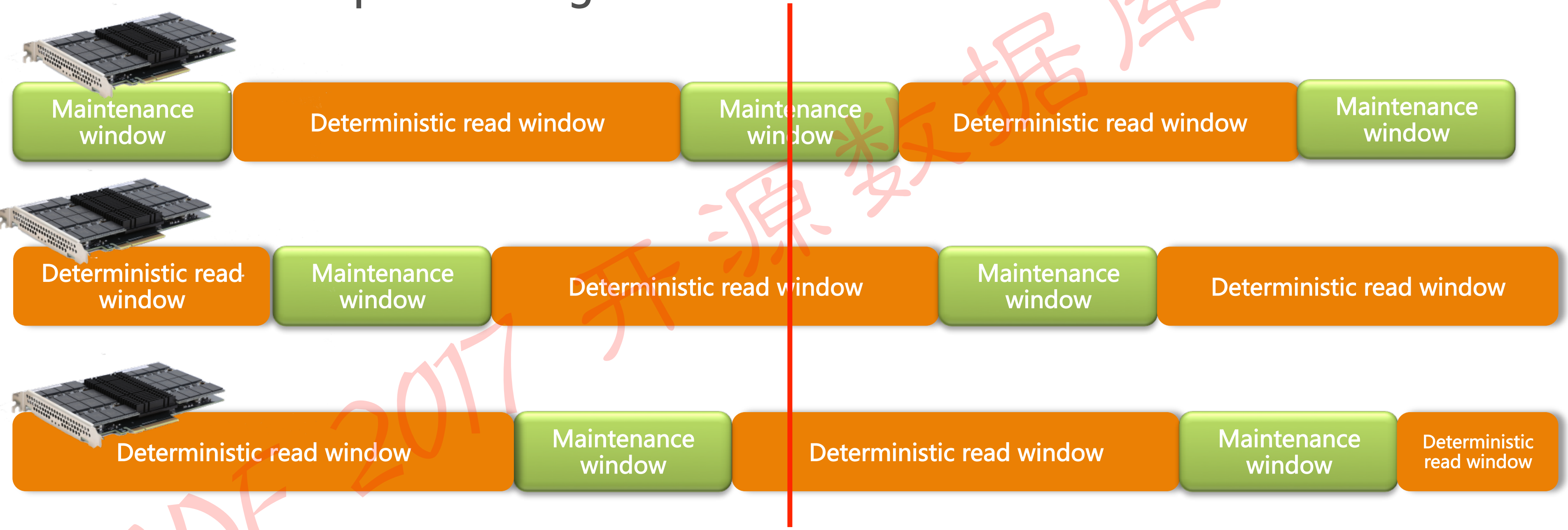
Maintenance window

GC/WL, background scrub etc

Source: Amber Huffman/intel, UCSD NVM workshop 2017

IO consistency –RW separation

- Pool multiple SSDs together via RDMA network


Maintenance window
Deterministic read window
Maintenance window
Deterministic read window
Maintenance window
Deterministic read window
Maintenance window
Deterministic read window
Maintenance window
Deterministic read window
Deterministic read window
Maintenance window
Deterministic read window
Maintenance window
Deterministic read window

At any time, one copy exists
for deterministic read IO

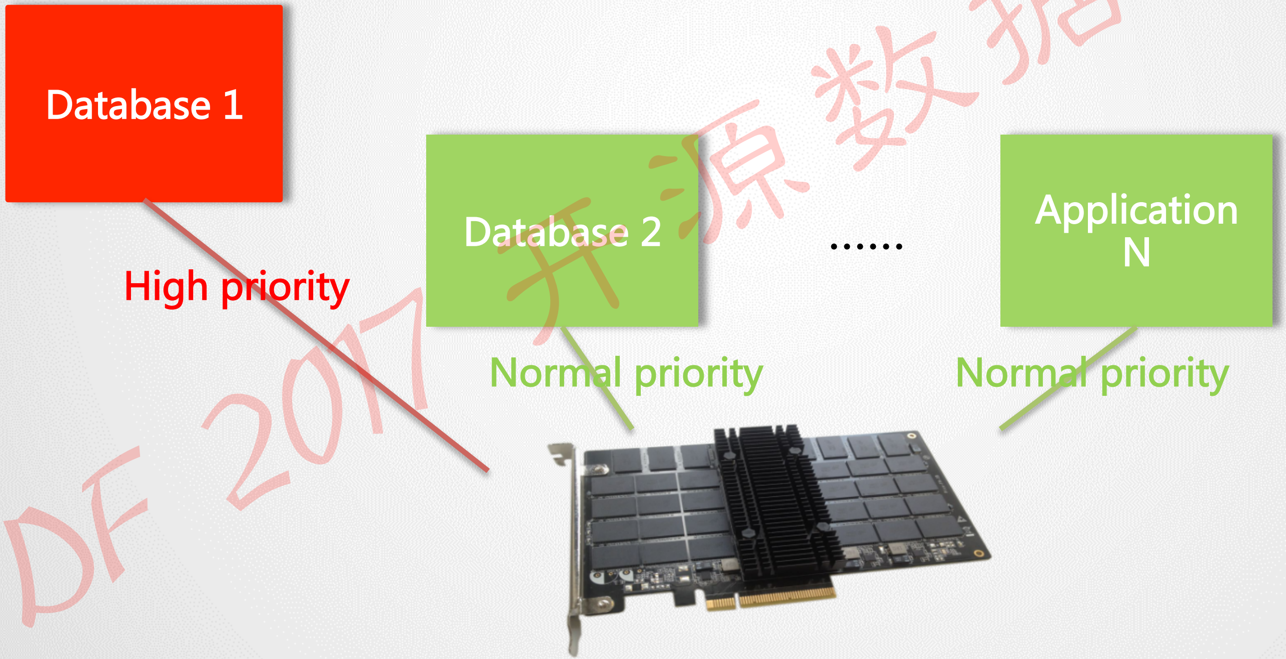
Source: Amber Huffman/intel, UCSD NVM workshop 2017

IO prioritization

- Provide priority service for certain traffic

IO prioritization

- Higher priority processing for IO from a particular application
- Priority Read/write operations



Database 1

High priority

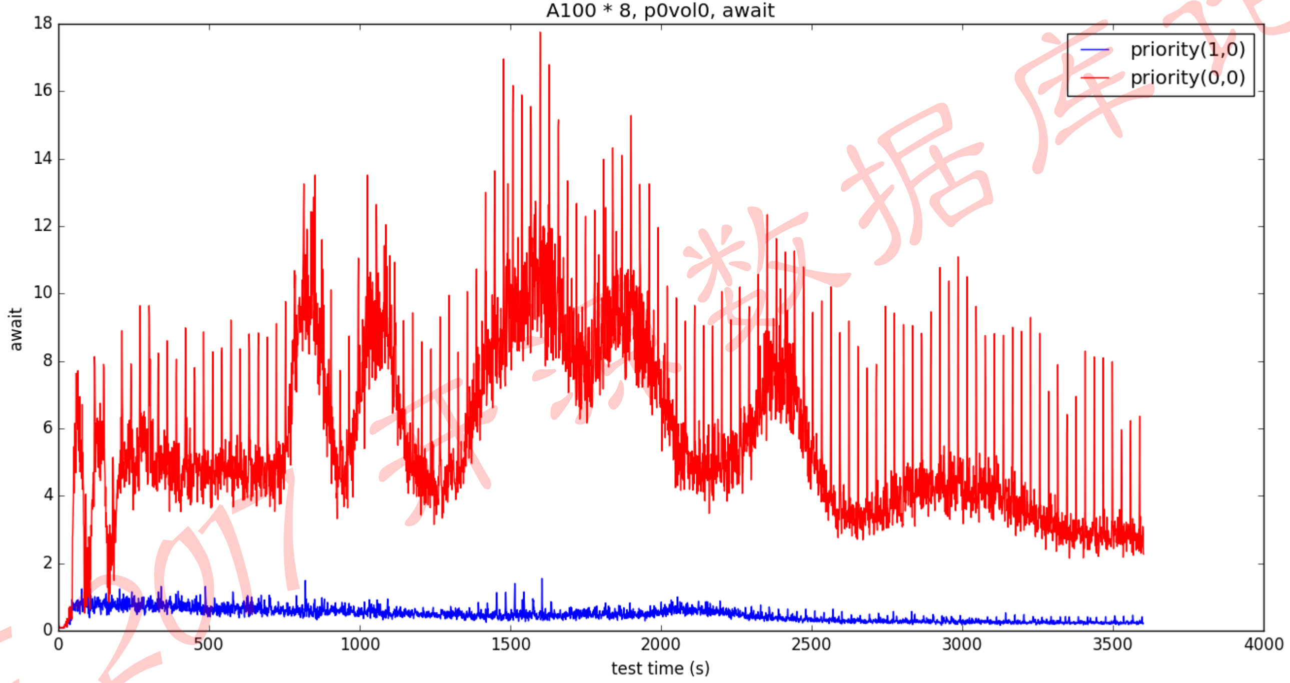
Database 2

Normal priority

......

Application N

Normal priority


A100 * 8, p0vol0, await
await
test time (s)
priority(1,0)
priority(0,0)
0
2
4
6
8
10
12
14
16
18
0
500
1000
1500
2000
2500
3000
3500
4000

IO atomicity

- IO (minimum size) sent to storage system
  - Success
  - Complete fail
- No intermediate state is allowed
- For example, Shannon Direct-IO PCIe Flash guarantees atomicity for any IO less than 32KB
- MariaDB supports atomic writes without using double write for improved performance.

Summary

1

SSD capacity/ performance admits resource sharing among applications

2

IO determinism elevates QoS

- separation
- prioritization
- atomicity

3

IO determinism is enabled via software-defined SSD

- Intelligence is offered in host SSD driver

4

Database applications can benefit substantially from better IO determinism.

Shannon Direct-IO™ PCIe Flash G4i

User Capacity

12.8TB

Access Latency

90/9us

IOPS
Random 4KB read/write

495/650K

上海宝存信息科技有限公司
Shannon Systems