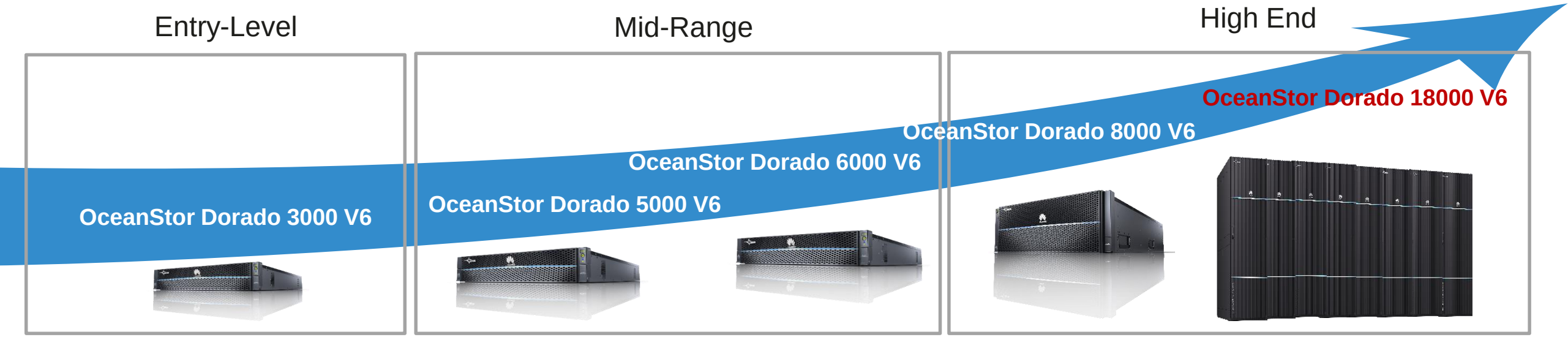
OceanStor Dorado V6 Technical Deep Dive

Security Level: Internal Only

HUAWEI

Overview of Product Portfolio



	Entry-Level	Mid-Range		High End	
Type	OceanStor Dorado 3000 V6	OceanStor Dorado 5000 V6	OceanStor Dorado 6000 V6	OceanStor Dorado 8000 V6	OceanStor Dorado 18000 V6
Height / Controllers of each Engine	2U/2C	2U/2C	2U/2C	4U/4C	4U/4C
Controller Expansion	2-16	2-16	2-16	2-16	2-32
Maximum Disks	1200	1600	2400	3200	6400
Cache/Dual Controller	192G	256G/512G	512G/1024G	512G/1024G/2048G	512G/1024G/2048G
Front-end ports	8/16/32G FC, 10/25/40/100G Ethernet				
Back-end ports	SAS 3.0	SAS 3.0/100G Ethernet			

Overview of OceanStor Dorado V6

20,000,000 IOPS
0.1ms

0 service interrupt
0 impact when upgrade

0 data migration for 10 years

End-to-End Symmetric Architecture

HyperMetro

SmartMatrix

RAID 2.0+

Fixed-length &
variable-length deduplication

FlashEver

End-to-End NVMe

Intelligent Read Cache

Intelligent Front-end Adapter

Intelligent DAE

High density DAE

Self-developed Chipsets

HUAWEI

New Generation Innovative Hardware Platform

Rear Panel

High-end controller enclosure

4U, 28 shared interface slots

Mid-range controller enclosure

2U, 2 controllers per controller enclosure

Entry-level controller enclosure

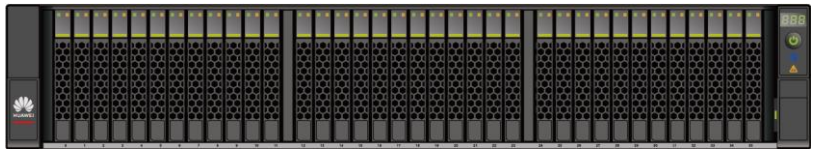
2U, 2 controllers per controller enclosure

Intelligent DAE

2U, 2 controllers per controller enclosure

Front Panel

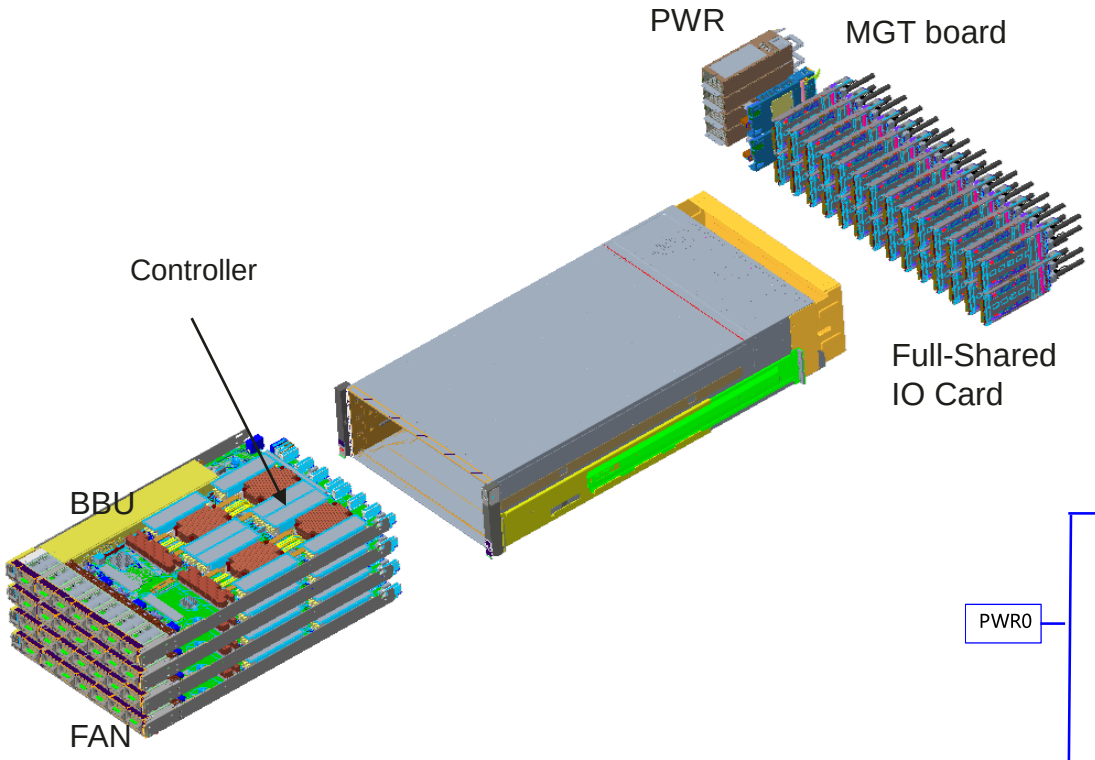
4U, 4 controllers per controller enclosure

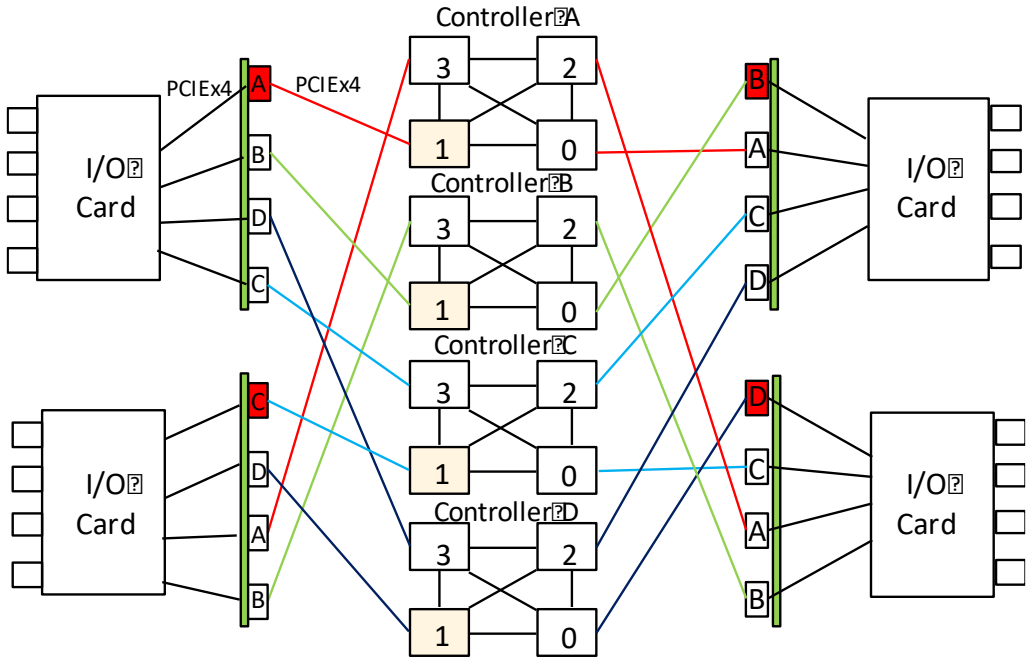


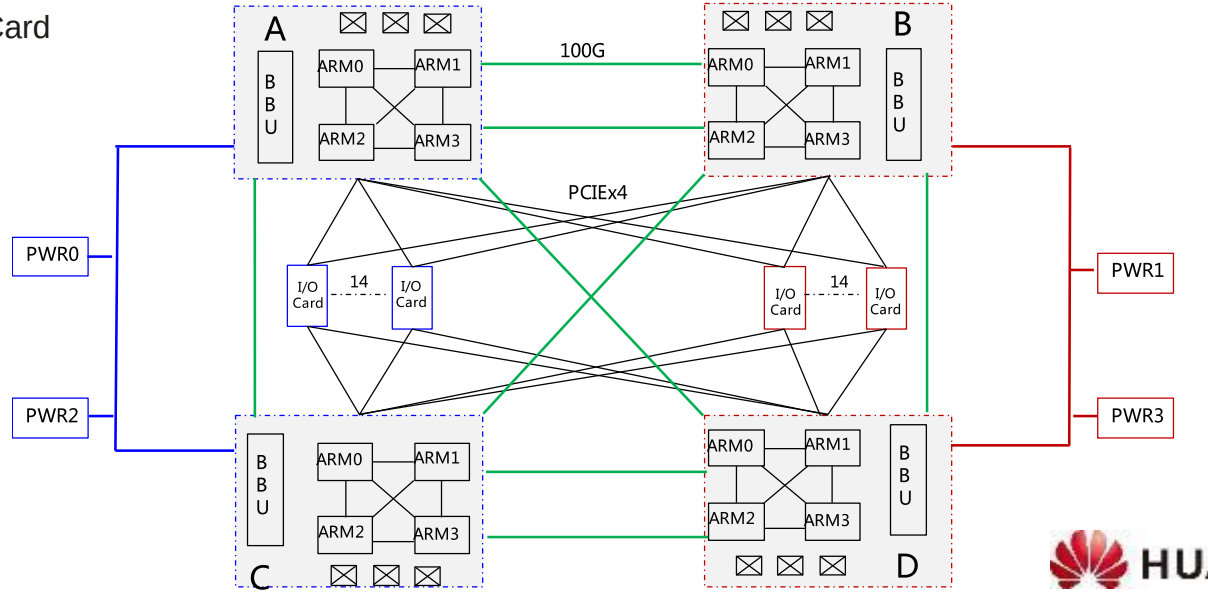
2U, 36 NVMe SSDs(high density)

2U, 25 SAS SSDs

Controller design for high-end series



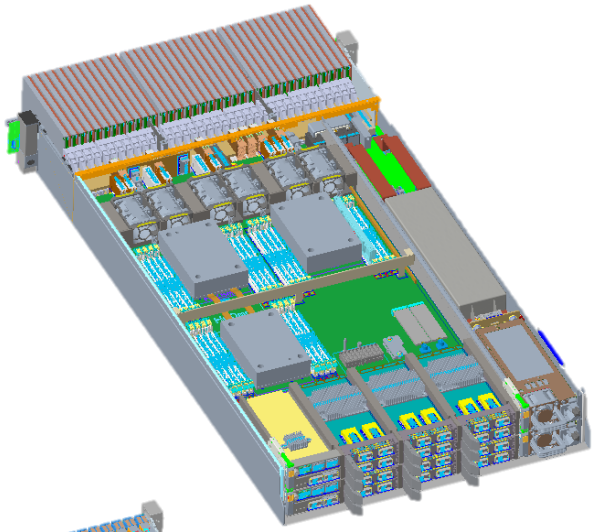




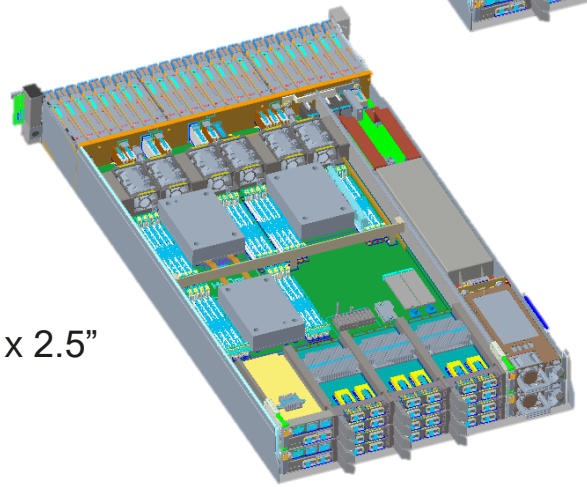
HUAWEI

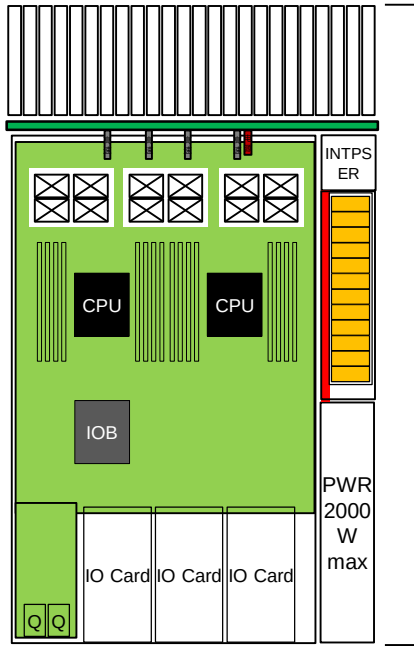
Controller design for Middle-range series

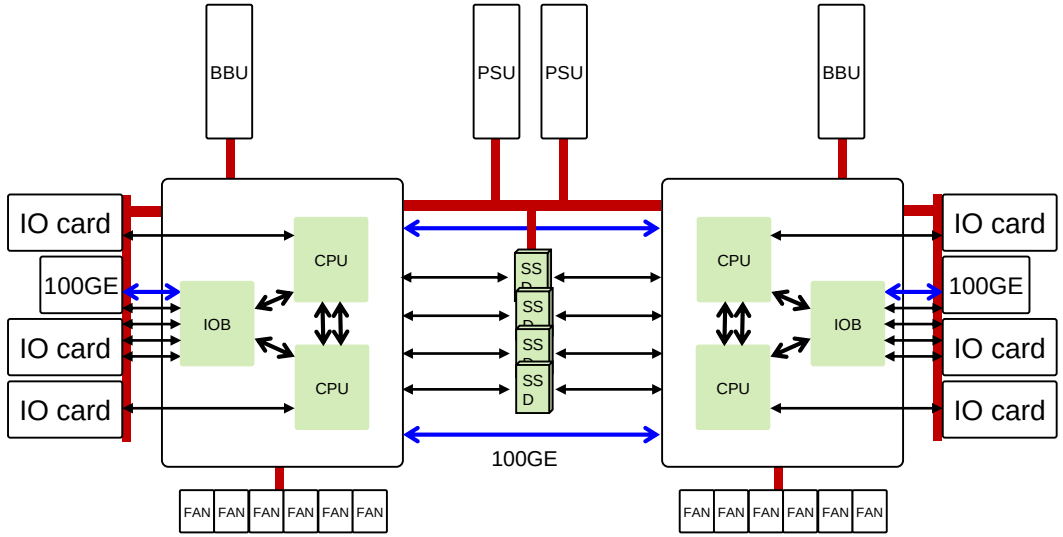
36 x Palm



25 x 2.5”



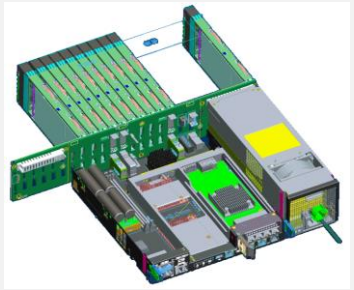




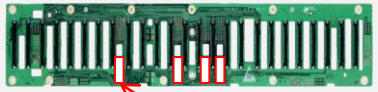
HUAWEI

2U, 36 disks, high capacity density

Traditional architecture design



Single backplane

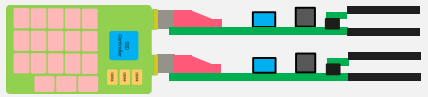


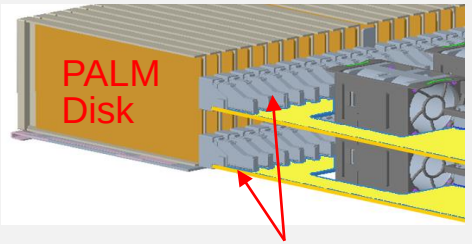
1. The heat dissipation window is small and the wind resistance is large.
2. Double-sided connector, interfering with each other. The number of hard disks is limited.

> 25-disk double-sided connectors cannot be staggered

Dual Horizontal Orthogonal Architecture Design

System side view





Horizontal backplane and orthogonal connection

1. The window area increases by 50%, and the heat dissipation capability increases by 25%.
2. Orthogonal connection without dual-side interference, increasing the number of hard disks by 44%

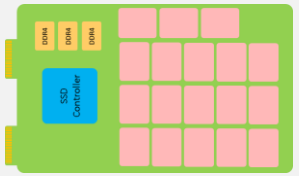
2U integrated equipment, 36 Palm SSDs, 44% SSDs increasing in industry

Traditional

Size:
100.6*14.8*70
Volume:
103cm³

U.2 NVMe SSD

User-defined Palm form



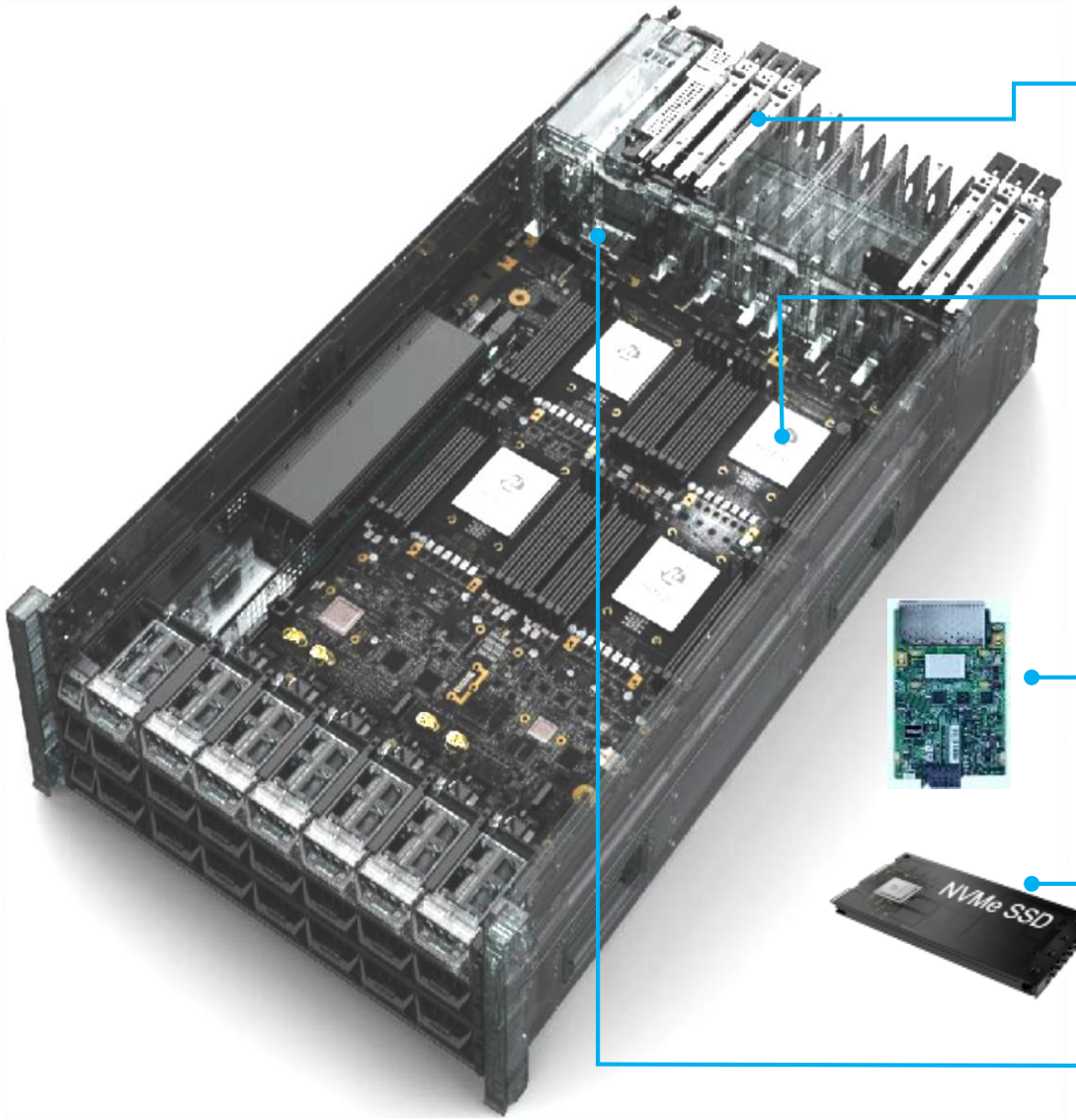
Dual-port Palm SSD

Size:
160*9.5*79.8
Volume:
121cm³

The number of 44% disk slots is added to the width of the 19inch cabinet.

Same capacity, width reduced by 36%

Innovative Hardware Platform Overview: with self-developed chipsets



Network Chip Hi1822
- **lower** Network latency 160μs→80μs

CPU Chip Kunpeng 920
- **NO.1** ARM CPU, **930+** SPECint
- Intelligent enclosure, CPU integrated.

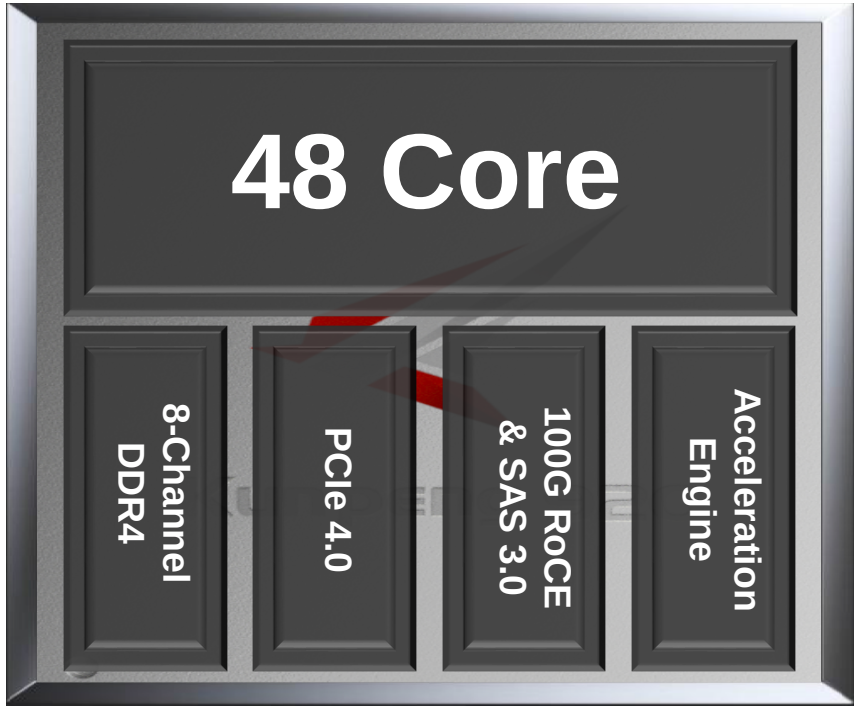
AI Chip Ascend 310
- **AI SoC** for mini-scale training

SSD Chip Hi1812e
- **lower** SSD Latency 40μs→20μs (write)

BMC Chip Hi1710
- Trouble shooting accuracy rate **93%**

HUAWEI

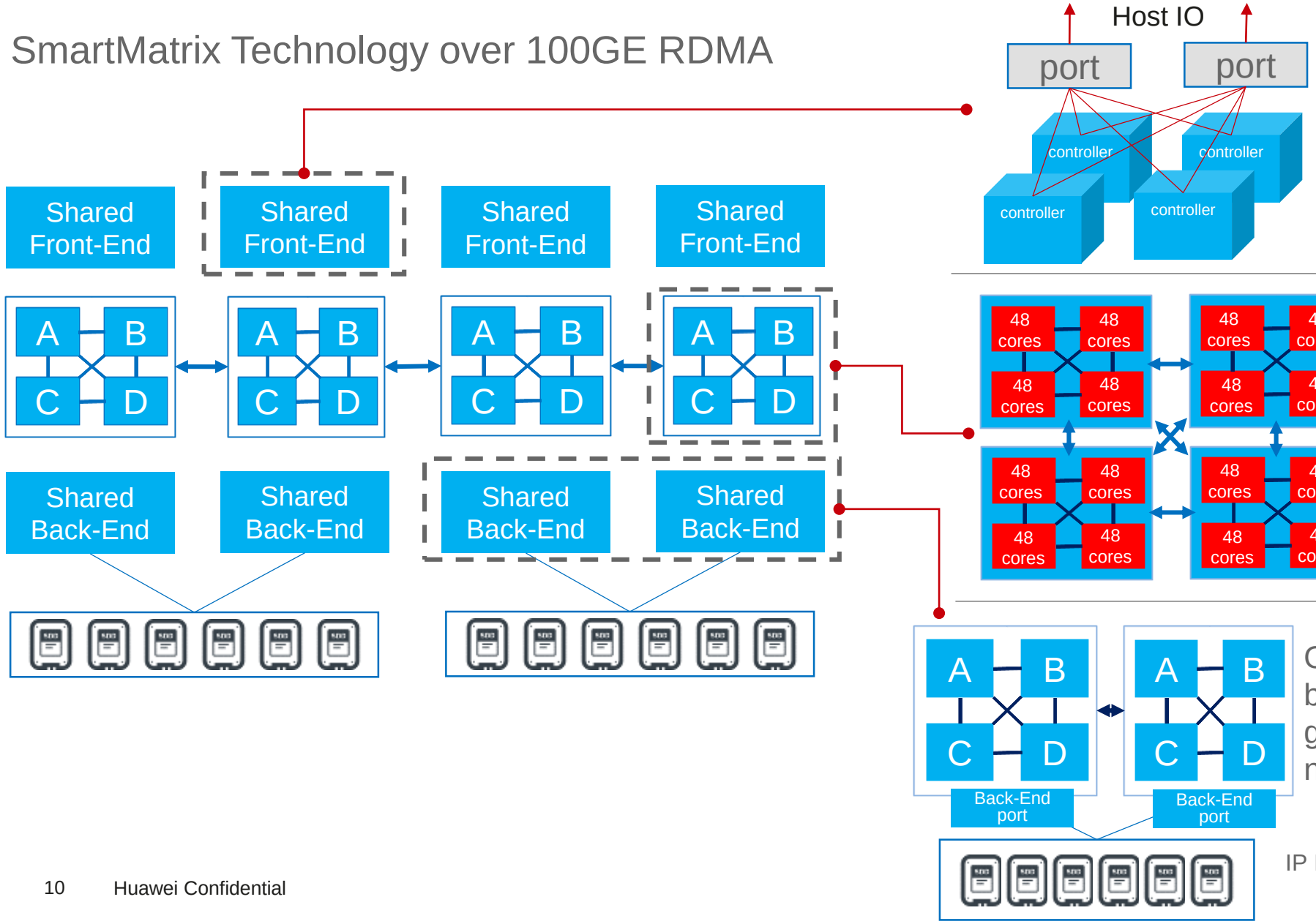
Kunpeng 920, the best processor for storage



High concurrency	**Up to 48 cores in one CPU**
High integration	**Not only computing**
High performance	**Hard acceleration engine to release the computing power**

Introduction of Connectivity: More reliable and More balanced

SmartMatrix Technology over 100GE RDMA



Every physical port connects all the four controllers in one engine

Full Mesh interconnection between all controllers in each engine

Shared interconnection module for connecting between the engines

One intelligent disk enclosure can be accessed by 8 controllers(2 engines) through the shared back-end module

The best Active-Active design

Vendor1: 2 Controllers Scale-out

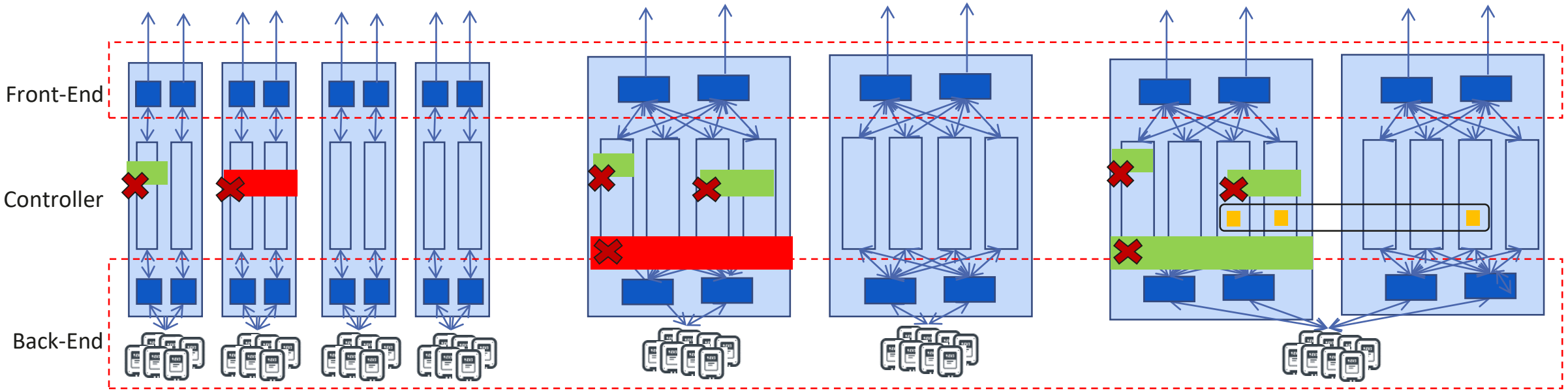
IO interface belongs to one controller, controller fails causes link switch over.

Vendor2: 4~8 Controllers Scale-out

Shared front-end, no switch over when any controllers failed.

OceanStor Dorado V6

Shared front-end, no switch over when any controllers failed. Controller failure is transparent to the host



Disk enclosure shared with dual-controller, dual-controller(one engine) failure causes service interruption.

Disk enclosure shared with four-controller, 4 controller failure(one engine) causes service interruption.

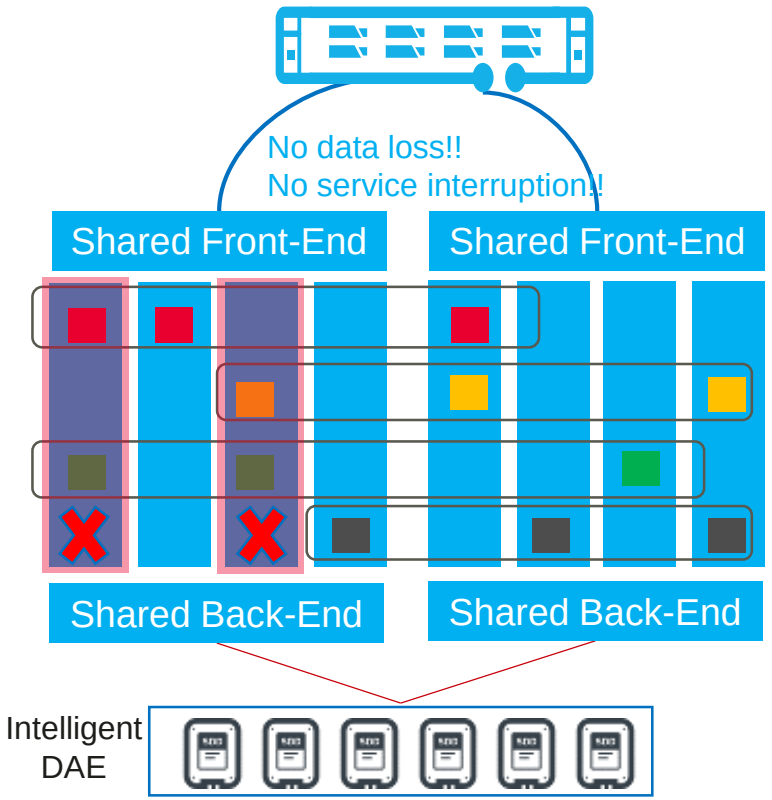
Global cache provides continuous mirroring technology and 3 copies across 2 engines

Disk enclosure shared with 8 controllers(2 engines)

No service interruption: any 2 controllers failed at the same time; 1 engine failed; 7 controllers failure one by one of 8 controllers(2 engines)

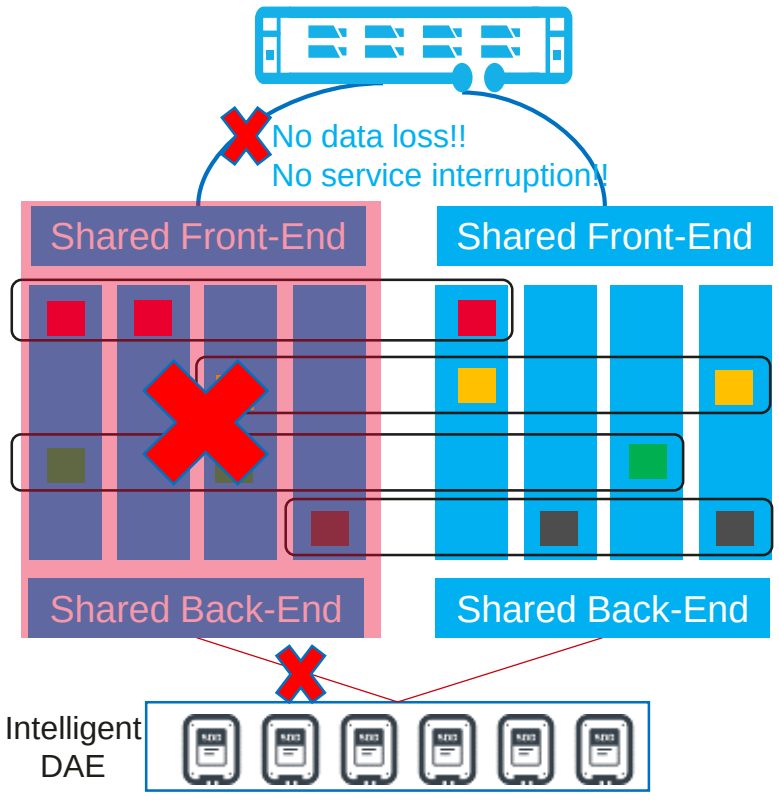
High availability Architecture(HyperMetro-inner for High-end series)

Tolerance of 2 controllers failure simultaneously



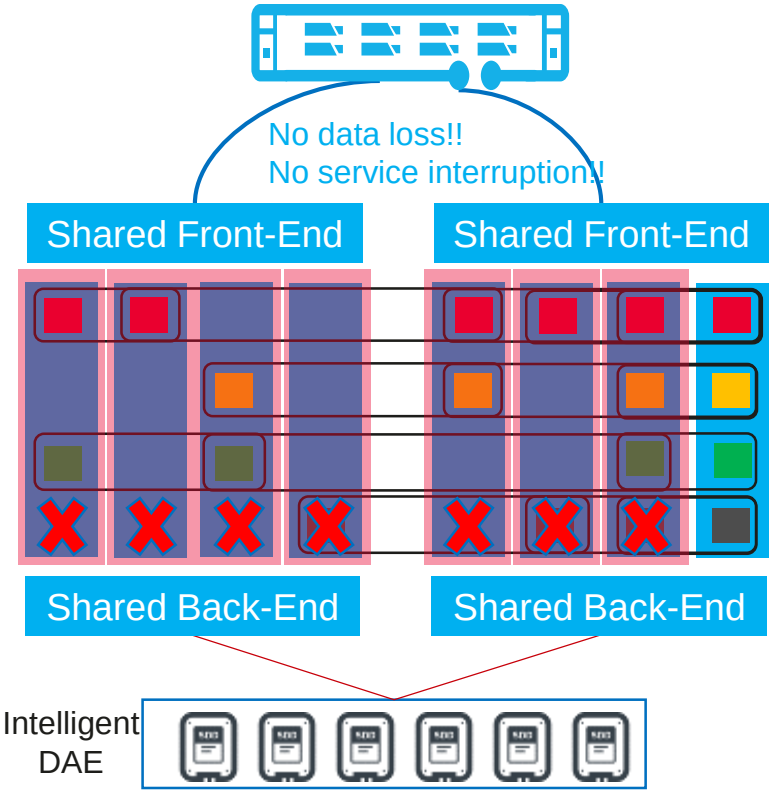
- Global Cache supports 3 copies across two engines.
- Guarantee at least 1 cache copy available if 2 controllers failed simultaneously.
- Only one engine can also tolerate 2 controllers failure at the same time with 3 copies Global Cache

Tolerance of 1 engine failure



- Global Cache supports 3 copies across two engines.
- One disk enclosure can be accessed by 8 controllers(2 engines) through the shared back-end module
- Guarantee at least 1 cache copy available if one engine failed.

Tolerance of 7 controllers failure



- Global cache provides continuous mirroring technology
- Tolerates 7 controllers failure one by one of 8 controllers(2 engines)

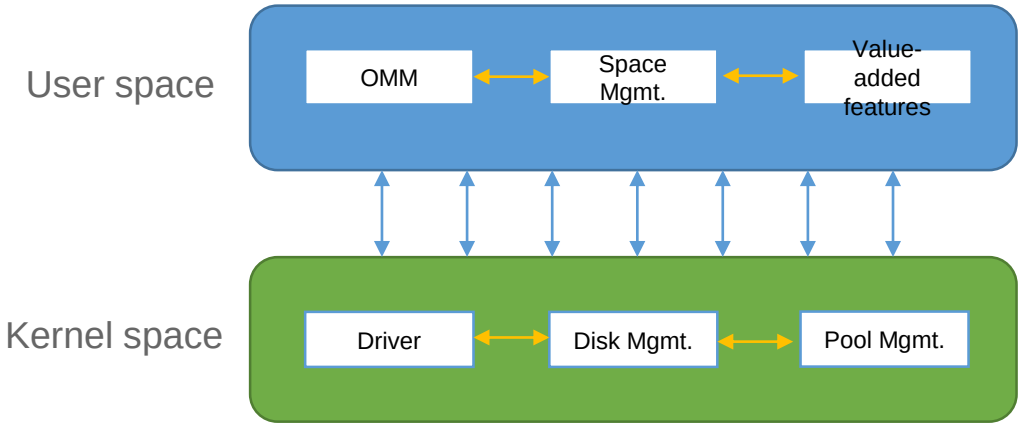
HUAWEI

Nearly all software components are in user-mode

Software components are in the user mode, Components can be quickly upgraded

Traditional design

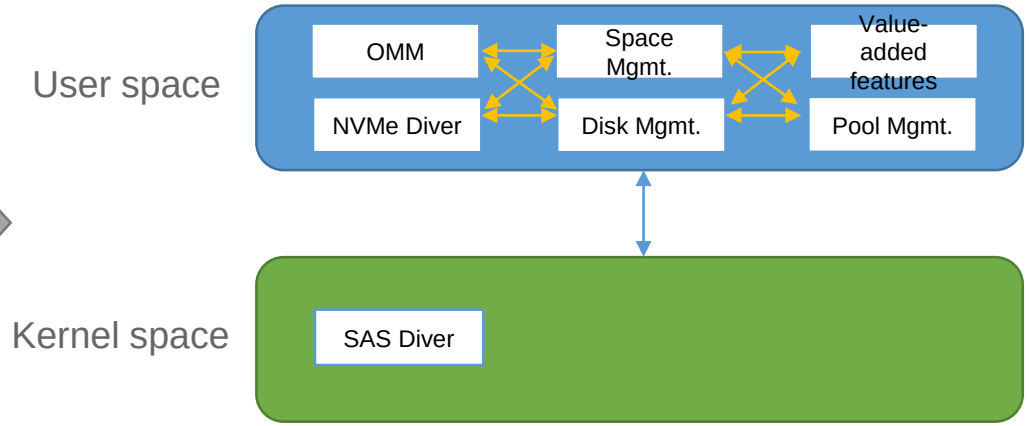
less user-level components, more interaction



Call each other between user mode and kernel mode
High Latency

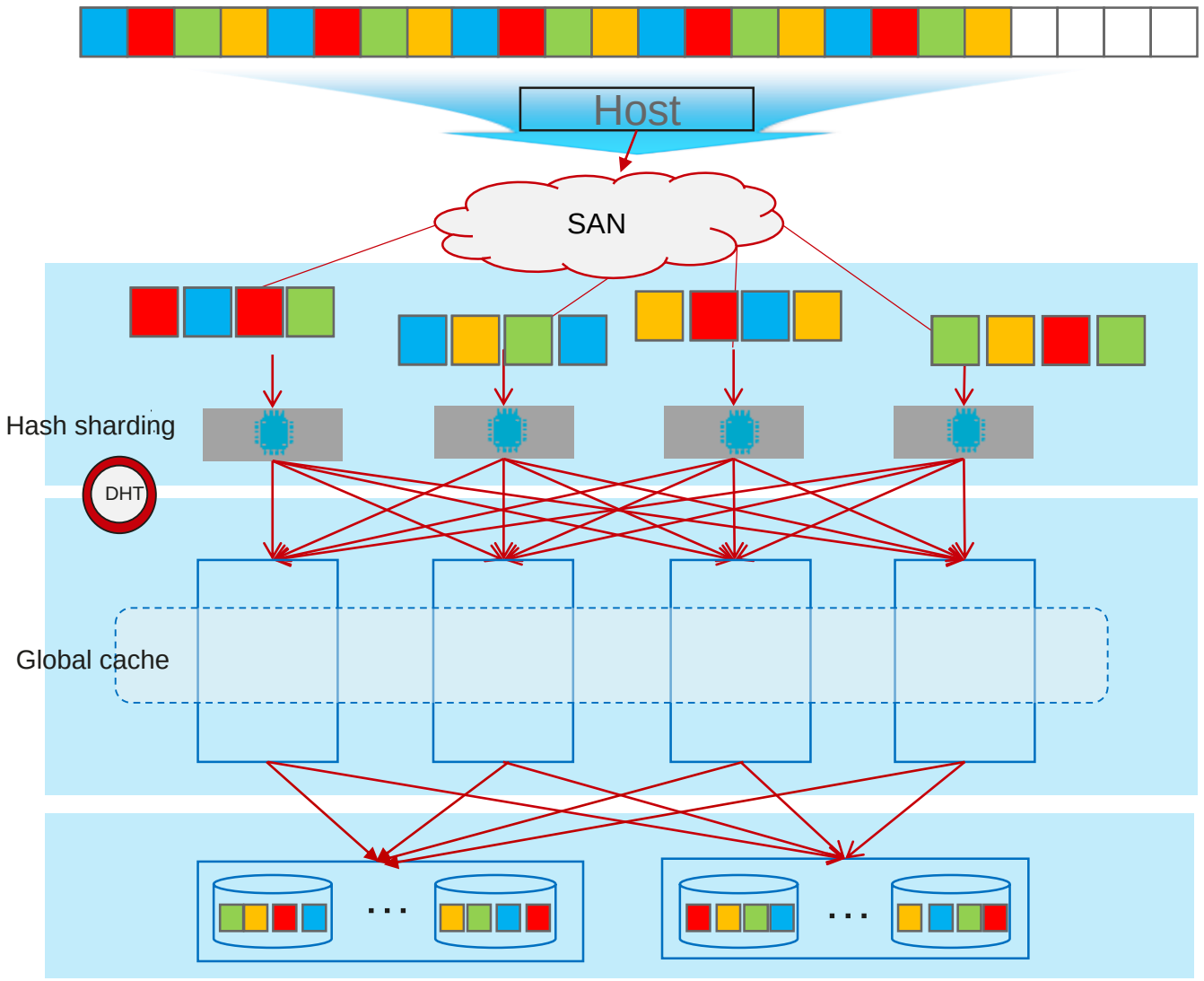
OceanStor Dorado V6

full user mode, less interaction



Reduce interactions between two modes
Low Latency

End-to-End Symmetric Architecture



Symmetric interface

- All series Support Active-Active access mode of the hosts, requests can evenly distribute on every frontend link
- LUNs of all series have no ownership controller, easy for use and load balance(LUNs are divided into slices and slices are distributed evenly on all the alive controllers by using DHT algorithm)
- High-end series provide shared and intelligent frontend IO module which can divide LUNs into slices and send the requests to their target controller for reducing latency

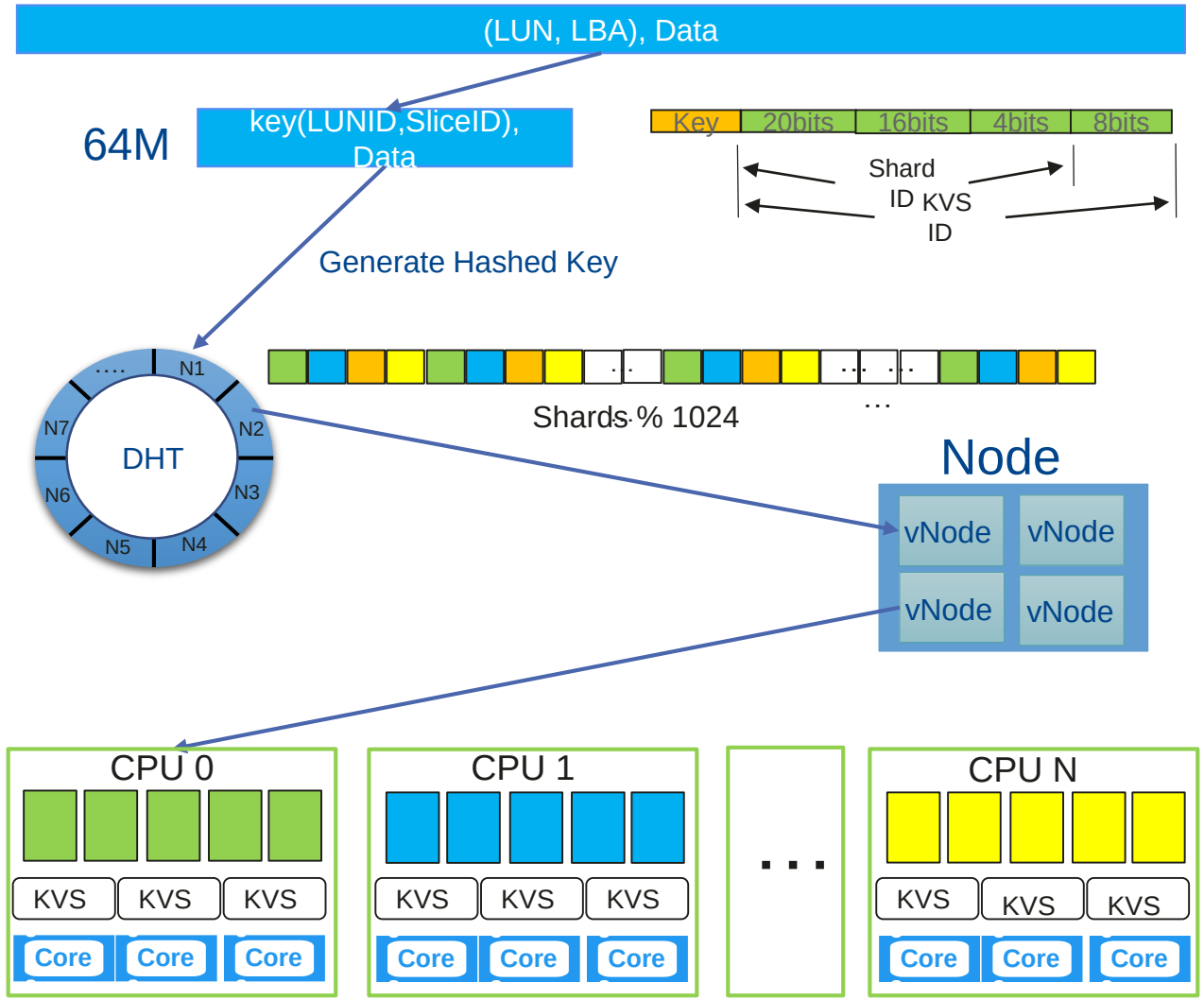
Global Cache

- IOs(located in one or more slices) of LUNs can be written to the cache of all the controllers and then be responded to the host
- The intelligent read cache of all the controllers can pre-fetch all the LUNs' data and meta data for cache hitting

Global Pool

- Storage pool can spread across all the controllers and use all the SSDs connected to the controllers to store all the LUNs' data and meta data by RAID2.0+

vNode End-to-End Service Equalization Scheduling Algorithm

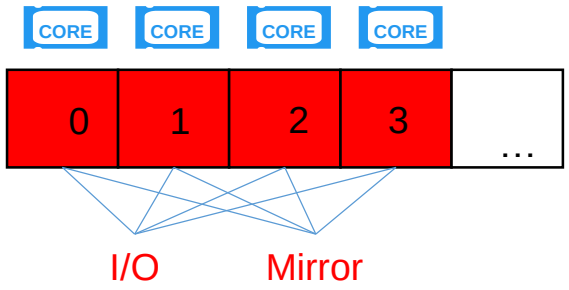


- Each LUN is divided into several Shards based on a fixed granularity (64M). Each Shard calculates the hash value based on LUN ID + Slice ID. Each Shard falls on the DHT ring. The 64M belongs to the same shard for sequential flow identification. Different 256Ks are distributed to different KVSs for load balancing.
- The DHT hash algorithm uses the 48Bits key. The first 40 bits are used to calculate a global Shard ID, and all 48 bits are used to calculate the KVS ID. Use the Shard ID to locate the vNode and the KVS ID to locate the Core.
- Each vNode is processed by only one physical CPU or Controller. The service performs CPU and memory affinity design based on vNode to reduce forwarding between multiple physical CPUs. vNode performs inter-core balancing and scheduling-free scheduling based on KVS.

CPU multi-core load balancing optimization:

No grouping -> Grouping -> Grouping + Intelligent scheduling

Traditional

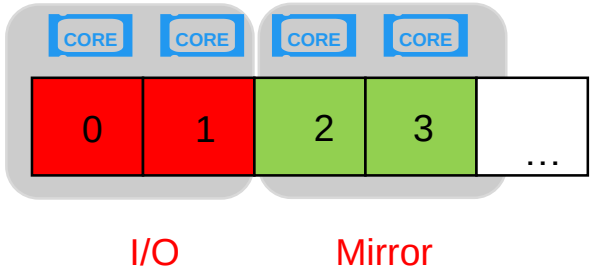


Challenge:

Different tasks compete for time slices of different cores of the CPU, resulting in frequent copying of data IO between different cores, resulting in high latency.

OceanStor Dorado V3

Core grouping



Advantage:

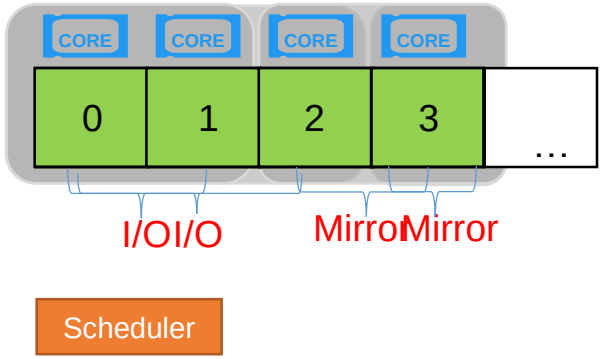
Avoid interference and frequent resource switching

Challenge:

Different services, partial nuclear overload, high latency

OceanStor Dorado V6

Grouping + intelligent scheduling

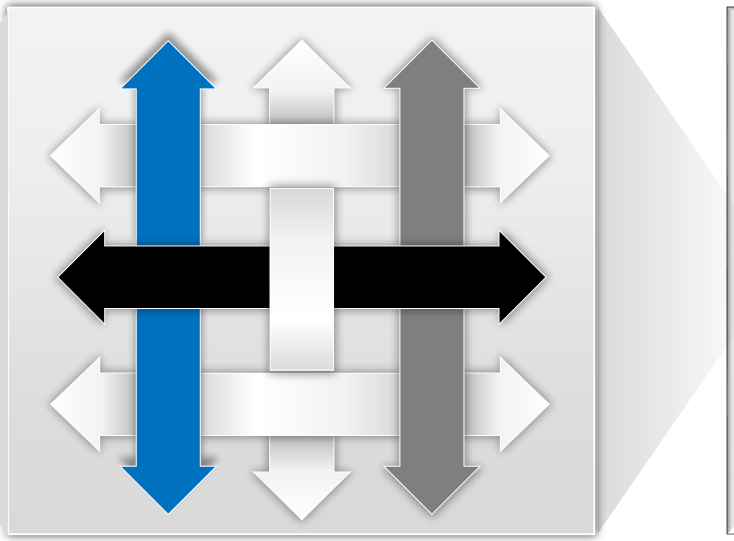


Advantage:

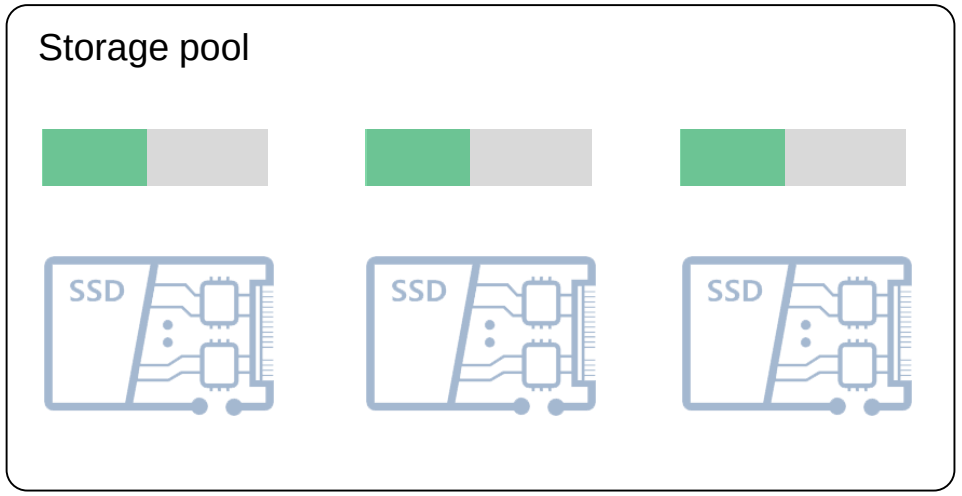
According to the load status, the intelligent scheduler dispatches tasks to other cores to achieve load balancing.

Self-developed SSD disk

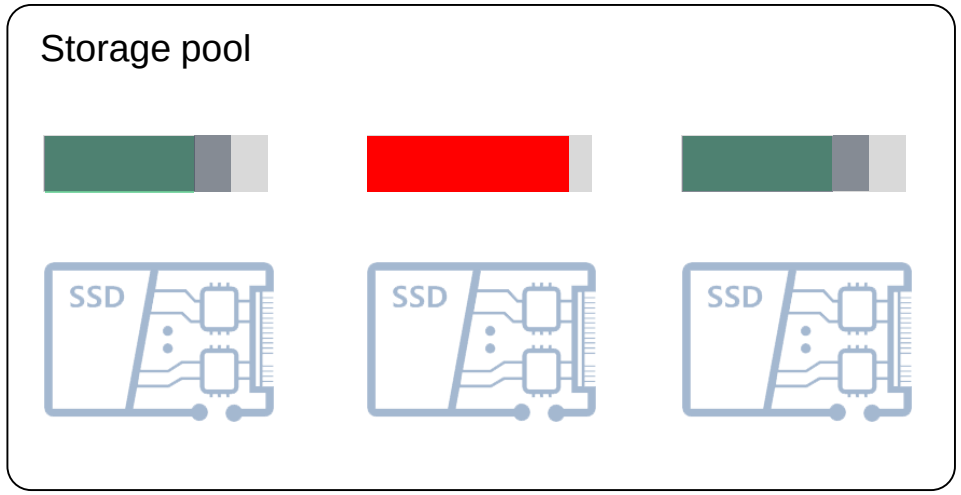
RAID 4 is supported in SSDs to ensure data reliability



OceanStor Dorado V6 supports RAID 5/6/TP, tolerating simultaneous failures of up to three disks

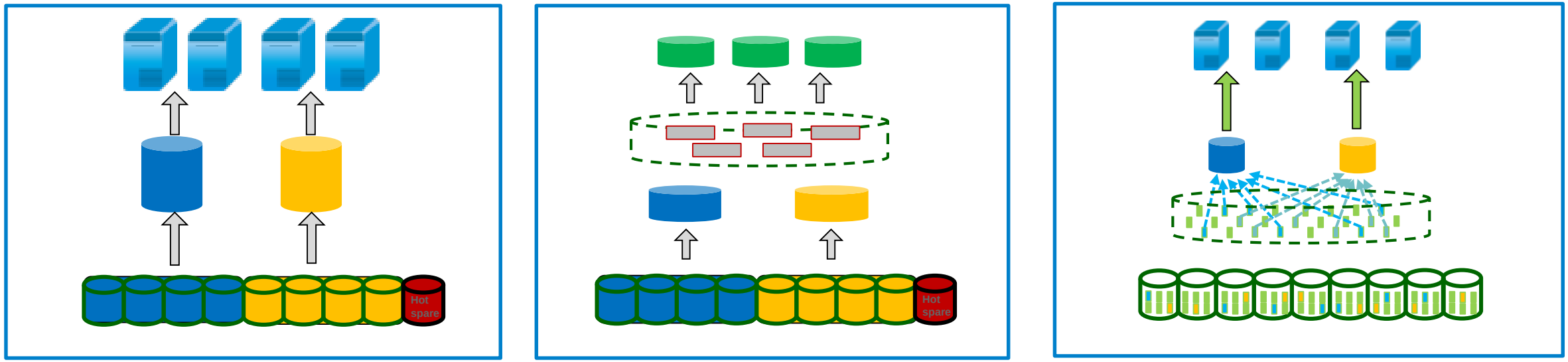


Global wear leveling



Huawei's patent: global anti-wear leveling

RAID 2.0+

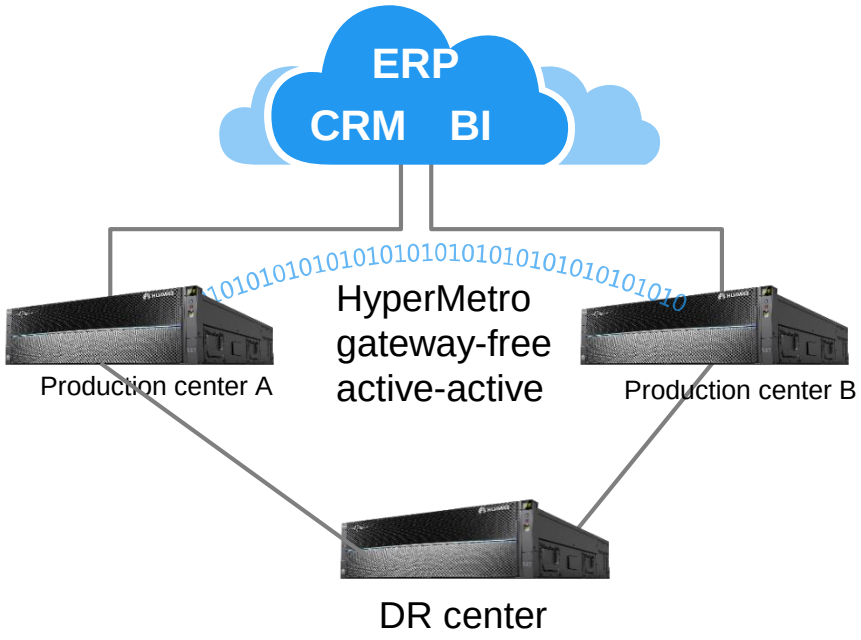


Traditional RAID → **LUN virtualization** → **RAID2.0+ Block virtualization**

Data reconstruction speed is improved 20-fold

- Huawei RAID2.0+: bottom-layer media virtualization + upper-layer resource virtualization for fast data reconstruction and smart resource allocation
- Fast data reconstruction: Data reconstruction time is shortened from 5 hours to only 15 minutes. The data reconstruction speed is improved 20-fold. Adverse service impacts and disk failure rates are reduced.
- All disks in a storage pool participate in reconstruction, and only service data is reconstructed. The traditional many-to-one reconstruction mode is transformed to the many-to-many fast reconstruction mode.

Gateway-Free Active-Active Solution



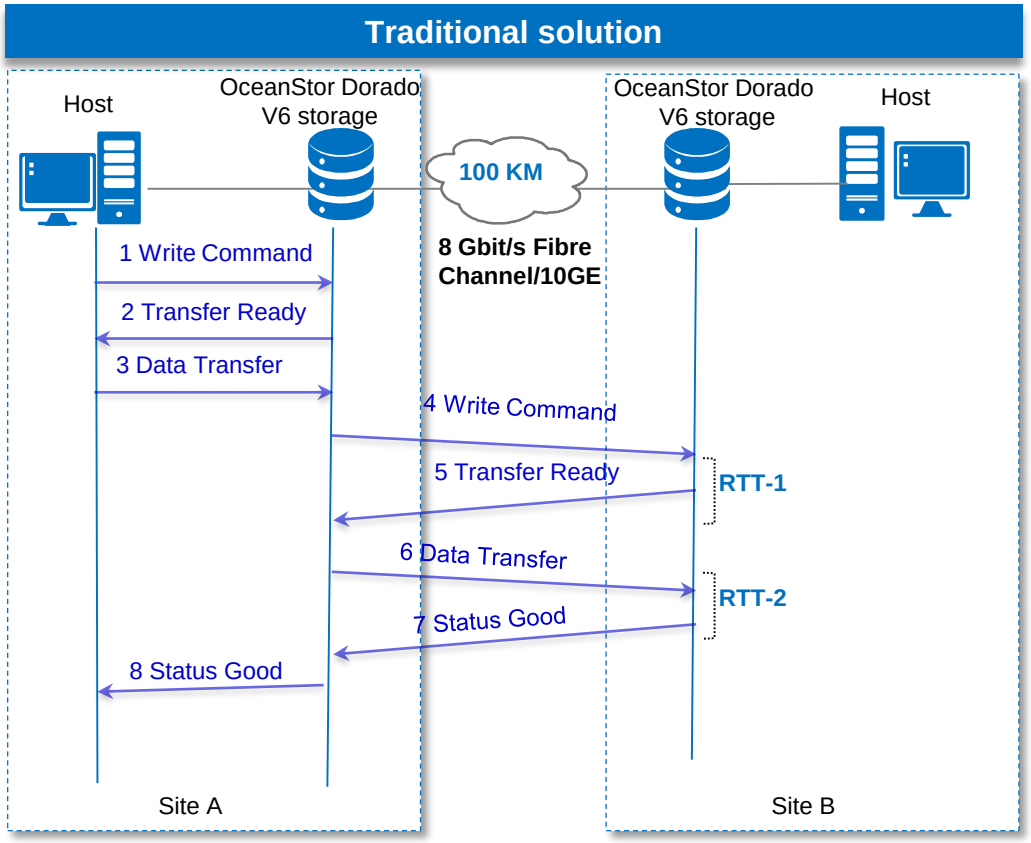
Lightning Fast, Rock Solid

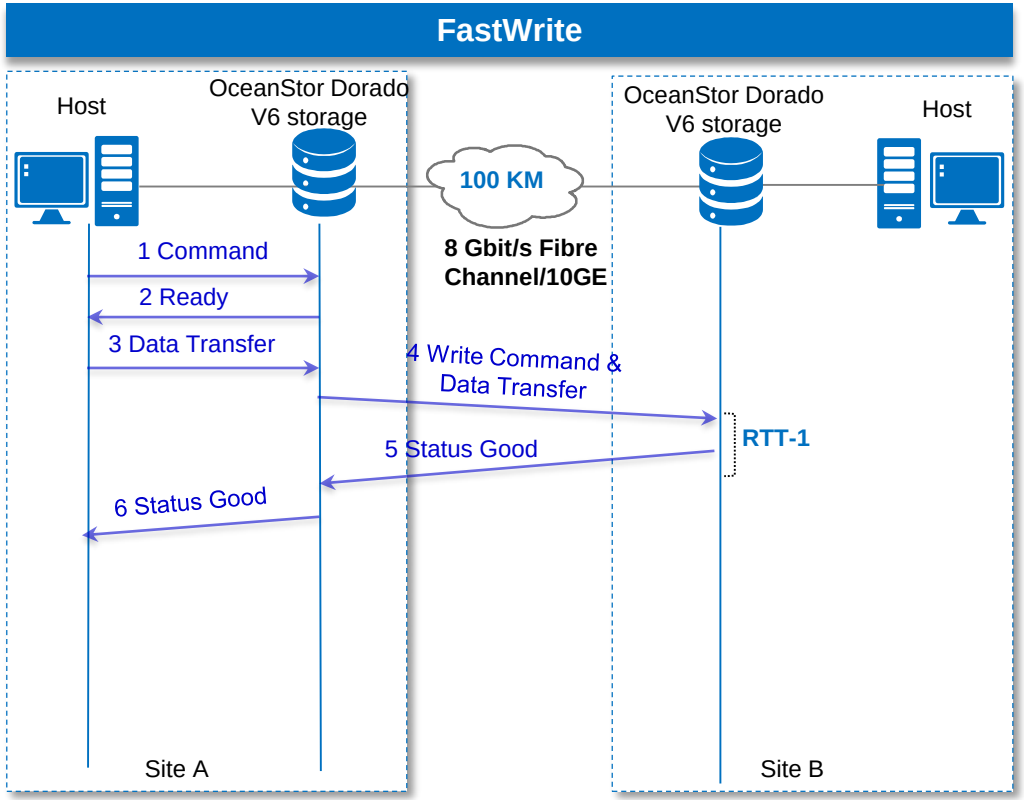
- **Gateway-free:** fewer nodes, simplified management
- **Active-Active:** load balancing between sites, RPO = 0 and RTO ≈ 0

Easy-to-Scale

- **Smooth upgrade to 3DC** provides a higher level of reliability.
- **Serial, parallel, and ring 3DC networking** meets the most demanding enterprise reliability requirements.
- **Interconnection with traditional storage** reduces the costs of building disaster recovery systems.

FastWrite: Dual-Write Performance Tuning

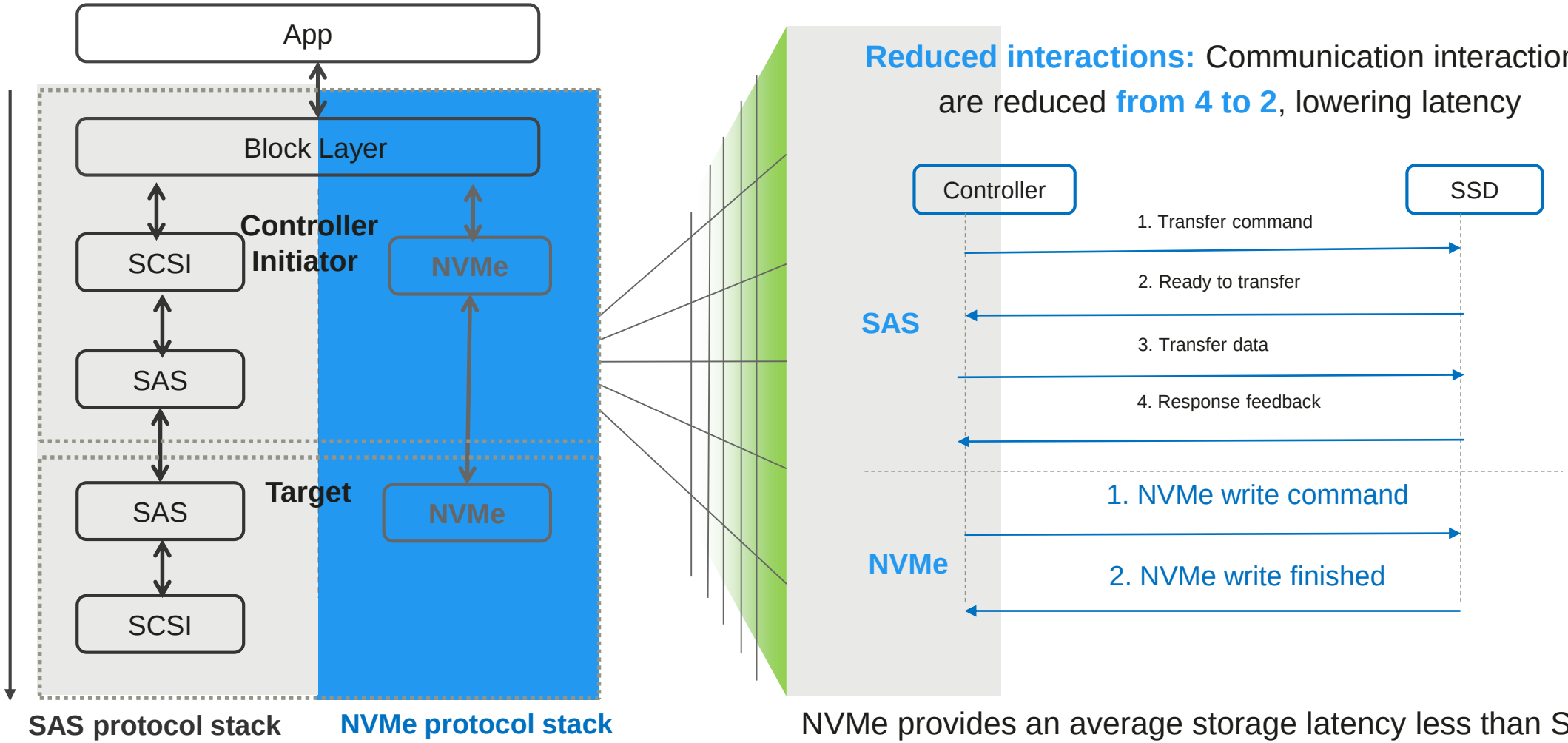




- Traditional solution: Write I/Os experience two interactions at two sites (write command and data transfer).
- 100 km transfer link: RTT (≈1.3ms) x 2

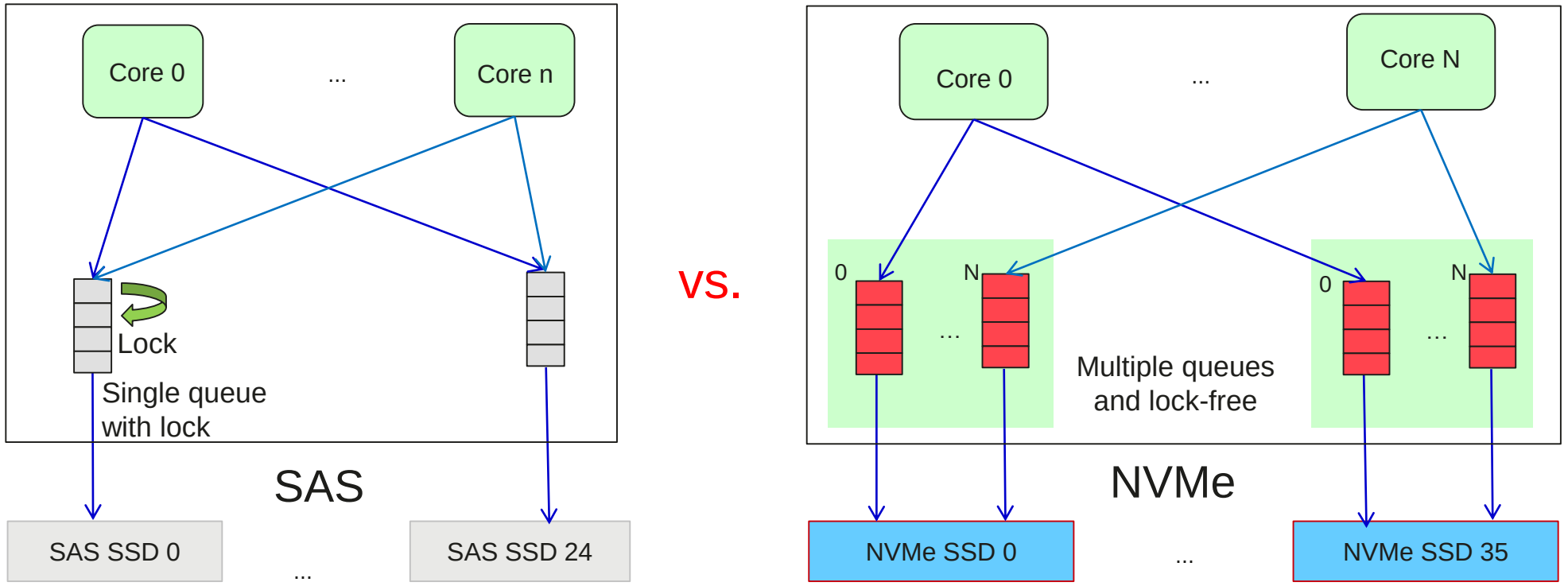
- FastWrite: A private protocol is used to combine the two interactions (write command and data transfer). The cross-site write I/O interactions are reduced by 50%.
- 100 km transfer link: RTT for only once, improving service performance by 25%

NVMe Reduces Protocol Processing Latency



NVMe provides an average storage latency less than SAS 3.0.

NVMe Concurrent Queue and Lock-Free Processing

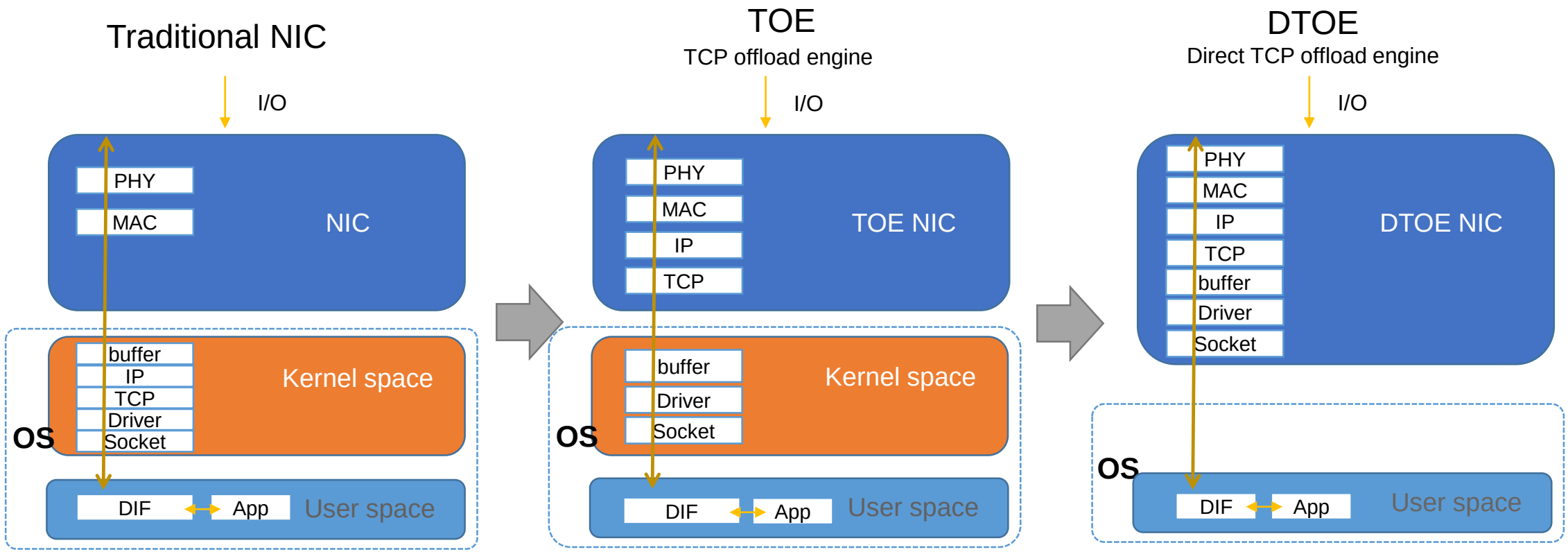


Number of queues = 25 (**OceanStor Dorado 5000 SAS with 25 SSDs**)

Number of queues = 288 (**OceanStor Dorado 5000 NVMe with 36 SSDs, N = 7**)

- NVMe: Every CPU core has an exclusive queue on each SSD, which is lock-free.
- Count of queues for each controller = Count of disks * Count of CPU cores for processing back-end I/O.
- SAS: Each controller has a queue to each SSD, which is shared by all CPU cores. Locks are added to ensure exclusive access of multiple cores. The number of queues for a single controller equals to the number of disks.

Intelligent NIC optimization: Traditional NIC -> TOE -> DTOE



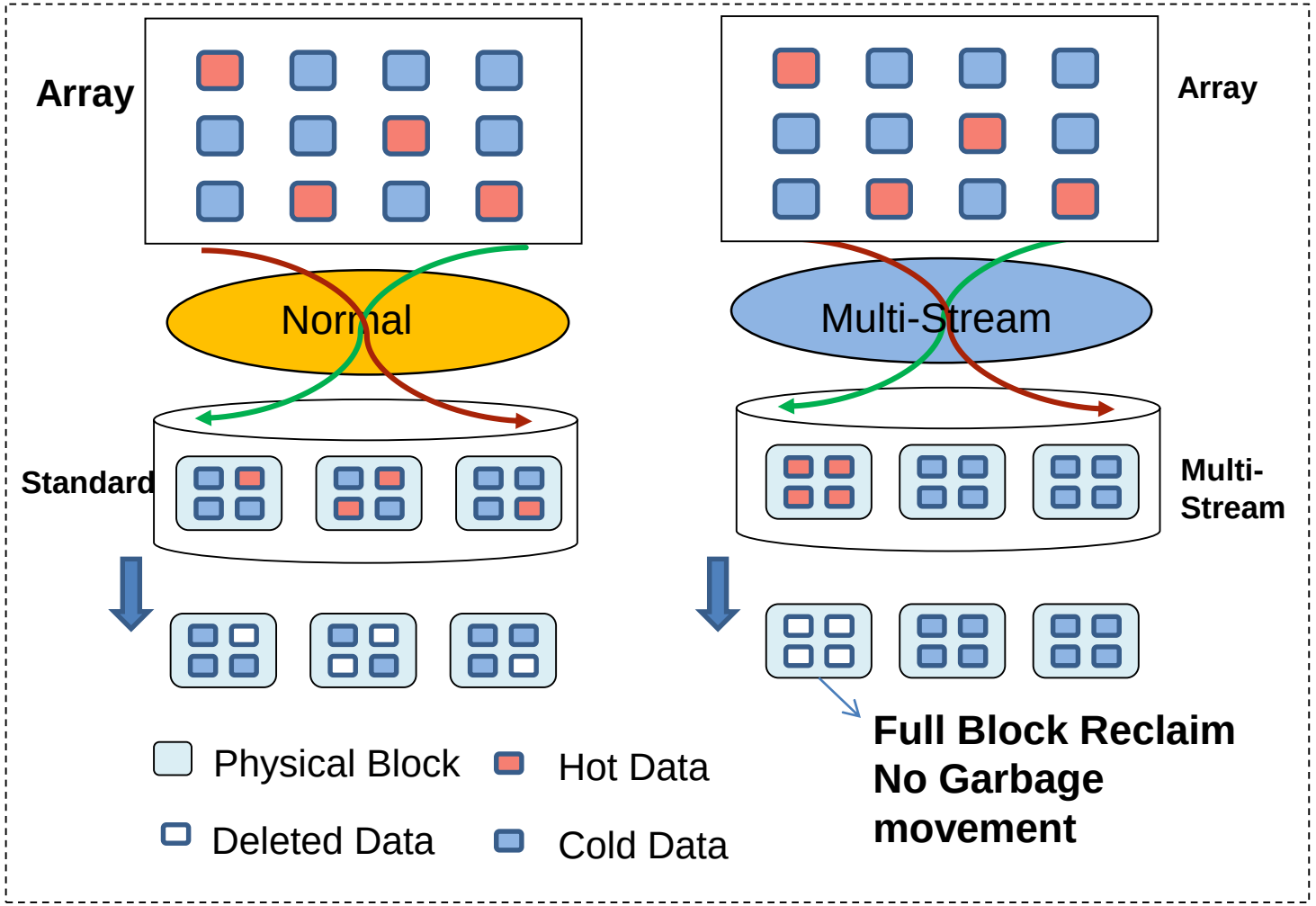
Challenge:
A traditional network card needs to trigger an interruption for processing each data packet, and CPU resource consumption is severe.

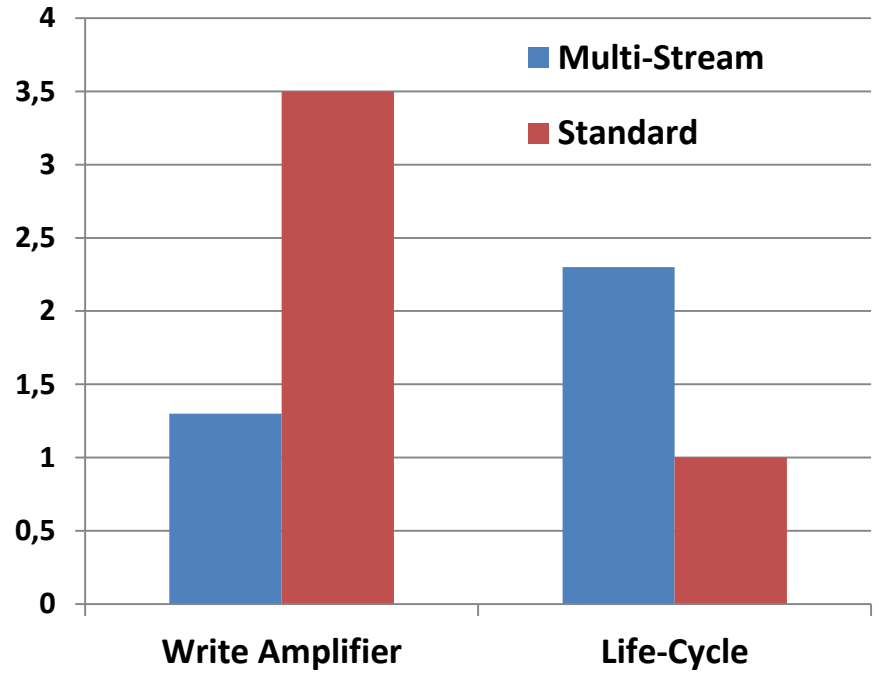
Advantage:
Each application can finish a complete data processing process before triggering an interrupt, significantly reducing the server's response to the interruption.

Challenge:
There are still high latency overheads such as kernel mode interrupts, locks, system calls, and thread switching.

Advantage:
1. Move processing of the transport layer to the Huawei customized network card 1822's microcode
2. Optimize storage application software to adapt the new architecture
3. Implement data (from the link layer) directly to the application memory
4. Bypassing the kernel state, significantly reducing the latency

FlashLink: Smooth GC with Multi-stream to reduce WA by 60%

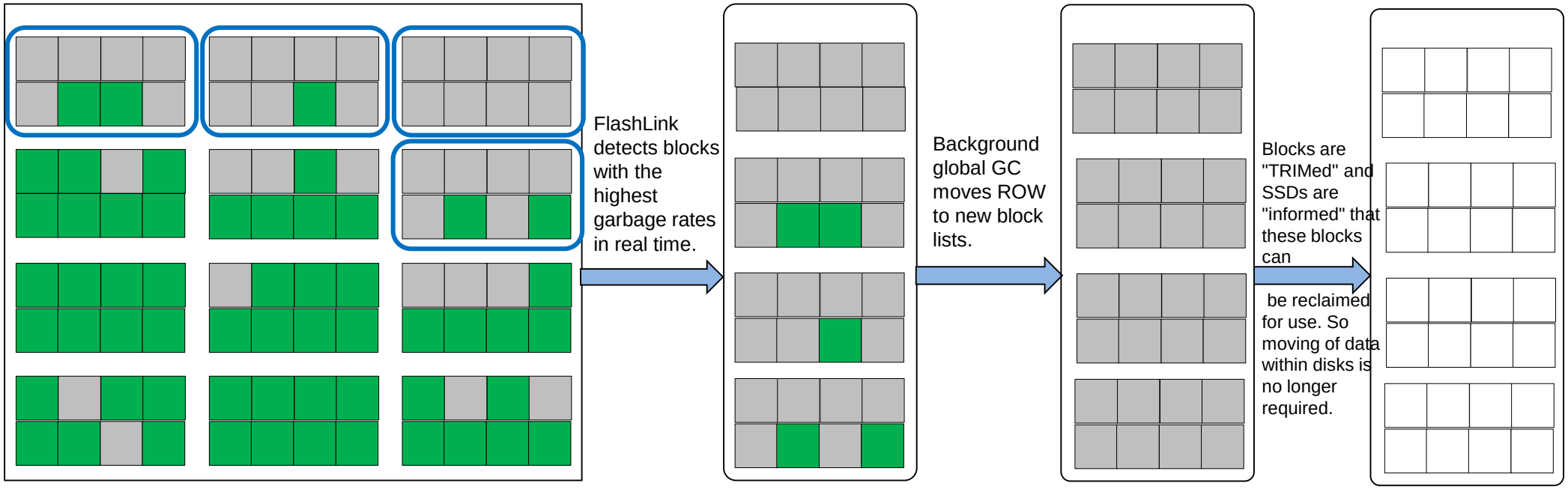




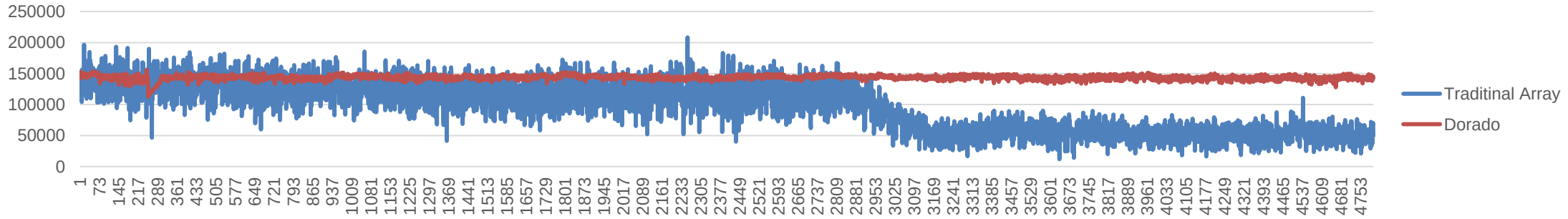
Write Amplification reduces over 60%, life cycle expands 2 times

HUAWEI

FlashLink: Global Garbage Collection

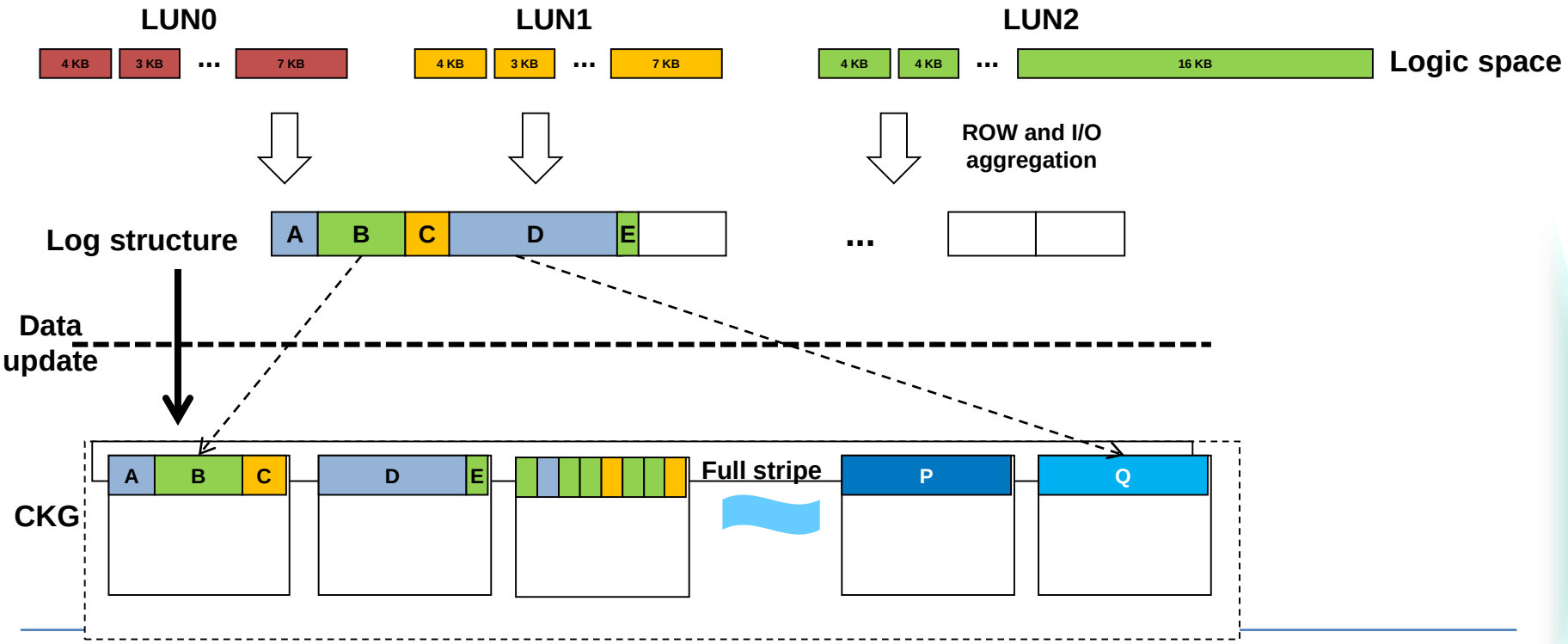


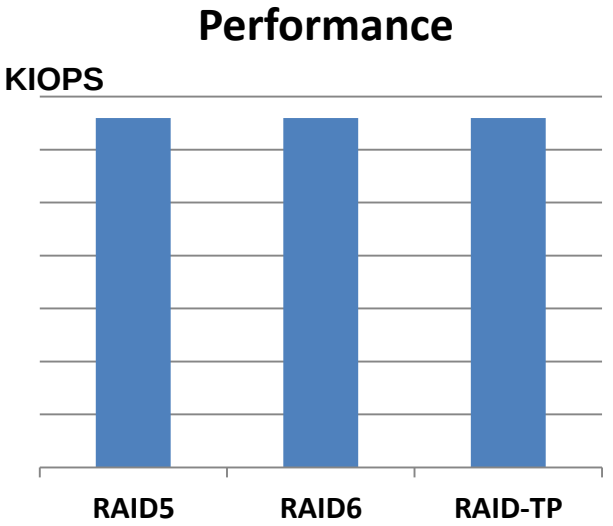
SSD FTL Only VS. OceanStor Dorado Global FTL — Performance Comparison

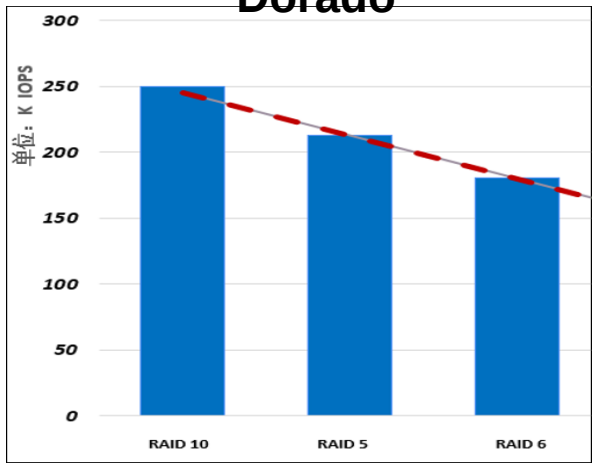


HUAWEI

FlashLink: Full stripe writing in RoW, Same performance across different RAID levels







Traditional Way

Configuration	Extra Reads	Extra Writes	Total IOs (extra IO)
RAID-5	2	1	4 (3)
RAID-6	3	2	6 (5)
RAID-TP	4	3	8 (7)

OceanStor Dorado way

Configuration	Extra Reads	Extra Writes	Total IOs
RAID-5	0	0	1
RAID-6	0	0	1
RAID-TP	0	0	1

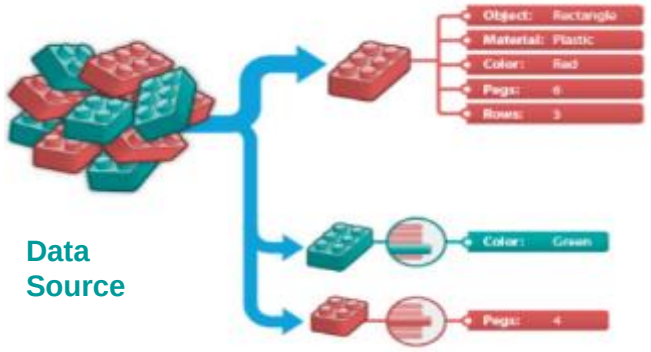
HUAWEI

Background deduplication by Variable length blocks

Adaptive deduplication between inline and background

- Feature Data Reduction
- Mandatory Inline Deduplication
- Software and hardware Adaptive Compression
- Adaptive Fixed length Deduplication
- Adaptive variable-length Deduplication

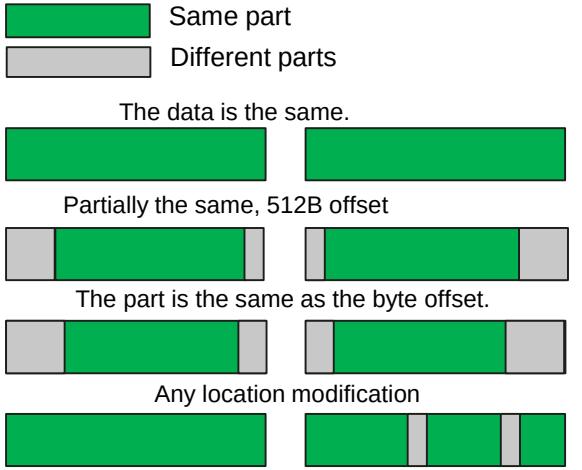
Fixed-length + variable-length deduplication



Reference block: Save the complete content.

Similar block: Differences in stored content

Fixed-length block: Count of saved references.

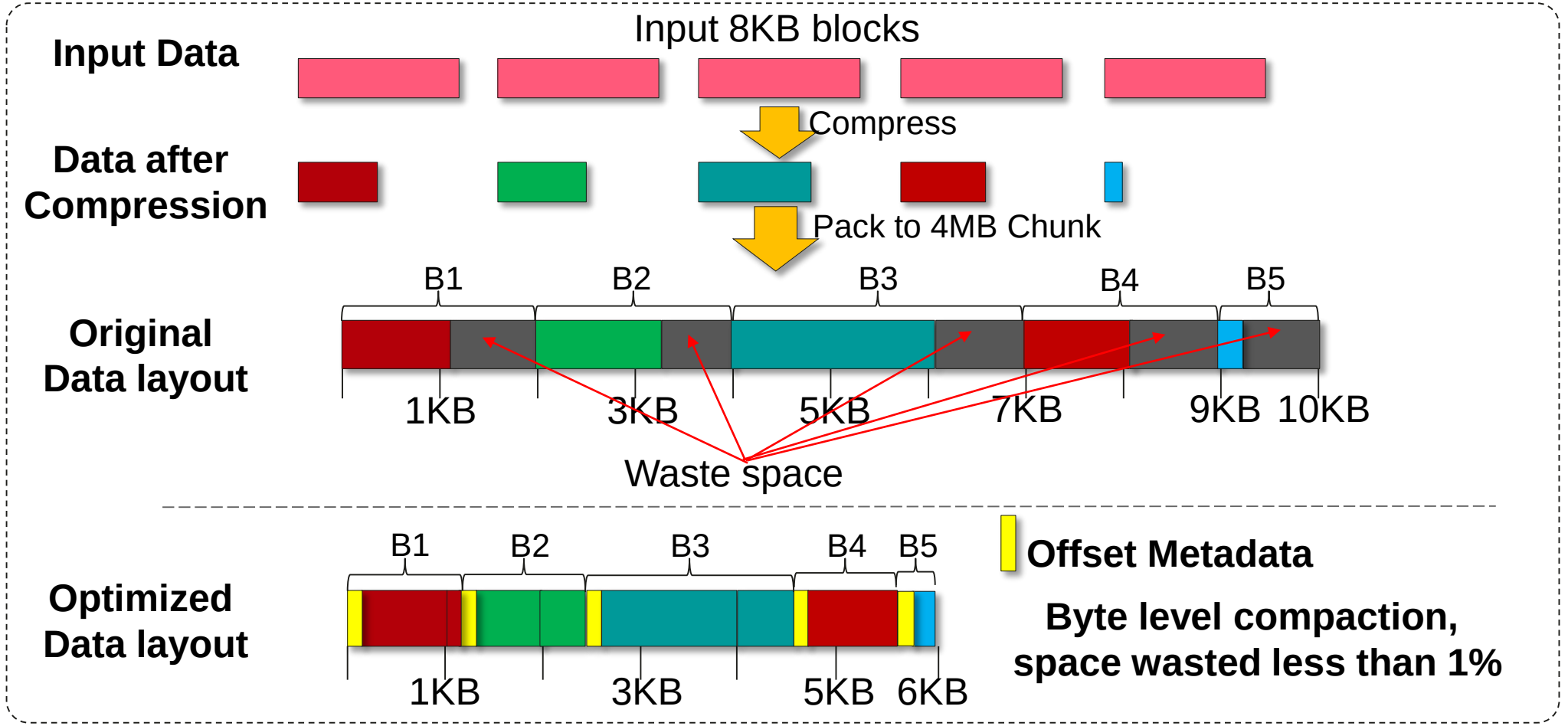


Deduplication is supported.

Supports hybrid deduplication of fixed-length and variable-length modes, achieving ultimate data reduction rate.

HUAWEI

Inline Compaction: Data Compaction By Byte



Self-Learning on workloads awareness:

Makes real-time intelligent analytics possible

Intelligence analytics

Intelligence Cache | Intelligence QoS | Intelligence dedup | ……

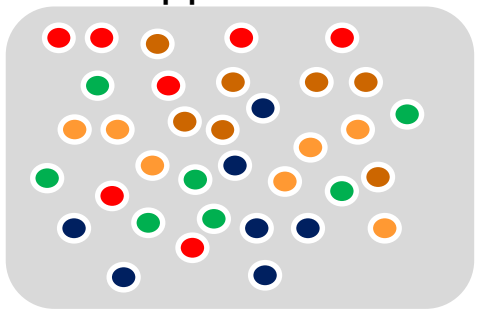
Data correlations | Data similarity

Health analytics | Adaptive optimization

Data temperature | Failure perdiction

ML/DL Engines

Application



Tens of million IOs

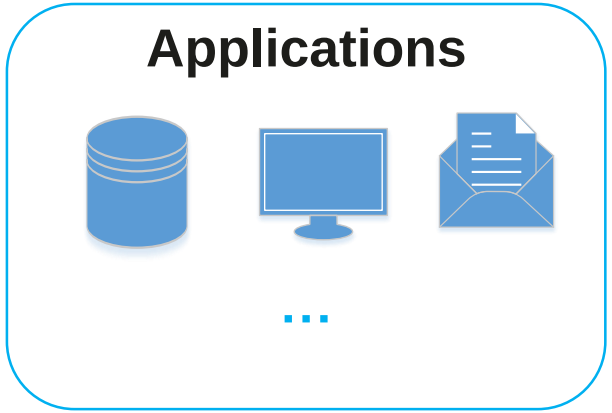
Thousands of sensors

A large amount of computing power is required.
(**>1Tera FLOPS** for real-time analytics)

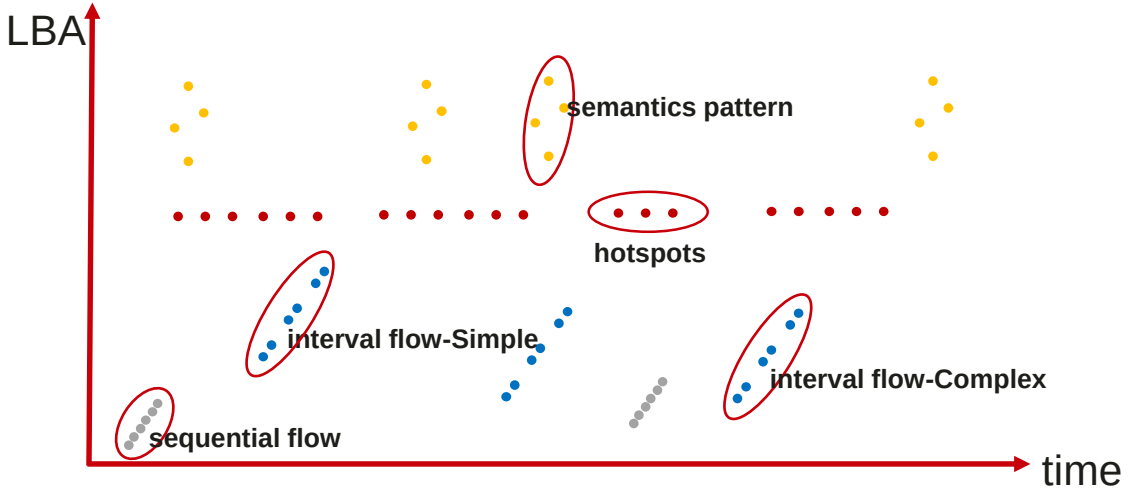
Ascend 310:
Architecture: Da Vinci
FP16: 8Tera FLOPS
INT8: 16TeraOPS
Max power: 8W

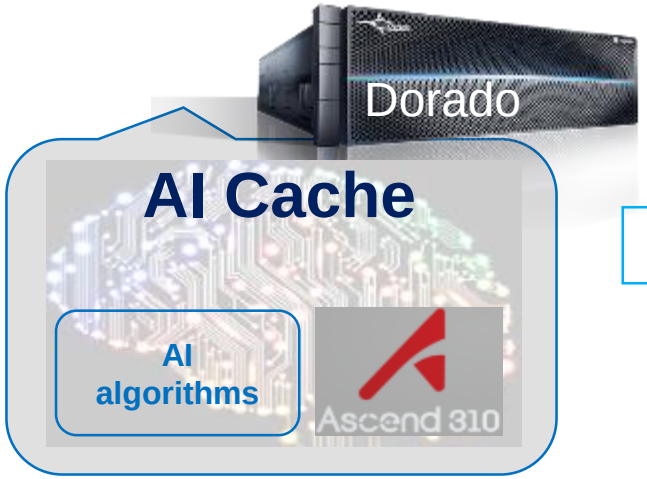
HUAWEI

AI Cache Effect: GPU for Deep Learning



IO sequence





Benefit on cost

Pattern	resource	model	accurate
sequential flow	CPU	Statistical learning	100%
hotspots	CPU	Statistical learning(MQ)	100%
interval flow	CPU	Statistical learning	100%
complex interval flow	CPU/ GPU	Machine learning(Bayesian network, EM and LZ77…)	>98%
semantics pattern	GPU	Deep learning(CNN and RNN)	>95%

Furthermore, we use deep reinforcement learning to realize model self-optimization.

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.