HUAWEI

Huawei New-Gen OceanStor A600 High-Performance AI Storage



Huawei New-Gen OceanStor A600 High-Performance AI Storage

Data Sheet



- Ultra Performance
- Flexible Expansion
- AI Native

Huawei OceanStor A600 is the new-gen high-performance distributed storage designed for AI, offering ultra performance to accelerate large AI model training and inference. It provides 160 GB/s bandwidth and 3.6 million IOPS per controller enclosure.

OceanStor A600 is built on an innovative data and control plane separation architecture. A single cluster can scale out to 32 controllers needed for the smooth evolution of multimodal models with over 1 trillion or even more than 10 trillion parameters. OceanStor A600 supports new data paradigms, such as vectors, tensors, and KV cache, as well as a built-in high-performance vector retrieval engine that accelerates retrieval from large-capacity vector knowledge bases and reduces AI hallucinations. It also provides multi-level KV cache to reduce inference latency while improving cost-effectiveness. This enables a single storage system to meet the data processing needs of the entire AI training and inference process.



OceanStor A600 high-performance AI storage

Product Introduction

- OceanStor A600 is an all-flash high-performance AI storage product that delivers excellent performance, high capacity, and elastic scalability based on brand-new hardware. It integrates disks and controllers, with each controller enclosure supporting 36 high-speed NVMe SSDs and providing 1105.92 TB of raw capacity.
- Based on an innovative data and control plane separation architecture, OceanStor A600 provides 160 GB/s bandwidth and 3.6 million IOPS. A single storage system accelerates the entire AI training and inference process. OceanStor A600 is widely applicable to financial AI, healthcare AI, regional intelligent computing centers, AI computing centers, scientific research platforms, autonomous driving platforms, and more.

Product Features

Ultra Performance for Higher AI Cluster Computing Power Utilization

New-gen OceanStor A600 AI storage uses the DataTurbo high-performance file acceleration engine to directly read data from and write data to the storage system in kernel mode. Kernel-mode caching enables memory data mobility and halves memory usage. OceanStor A600 offers 160 GB/s bandwidth per controller enclosure and TB/s-level aggregate bandwidth per cluster.

OceanStor A600 uses NPU Direct Storage (NDS) to build an optimal performance path from NPUs to the storage system, bypassing complex protocol stacks and eliminating the need for data copy, improving performance by 30%.

More Flexible Expansion

Based on the brand-new hardware architecture, OceanStor A600 requires at least two controllers. A single cluster can scale out to 32 controllers and 10 PB-level storage capacity, ensuring high performance and capacity density.

A single new-gen OceanStor A600 AI storage system can meet diverse ecosystem requirements. It is compatible with file, object*, and among other protocols, and supports the entire large AI model development process, such as data ingestion, preprocessing, model training, and deployment and inference. OceanStor A600 eliminates the need for data copy and accelerates the iteration and development of large AI models.

OceanStor A600 provides SmartQuota. With this file system quota technology, system administrators can manage and control the storage resources used by users, and user groups. By limiting the disk space that can be used by specified subjects, SmartQuota prevents excessive resource occupation and maximizes the value of storage space.

More Efficient Inference with AI-Native Data Paradigms

The data paradigms supported by new-gen OceanStor A600 AI storage range from NAS and objects to new vectors and graph-based adaptive retrieval engines. OceanStor A600 enables high-performance retrieval from massive amounts of multimodal knowledge. The solution's competitive advantage lies in its inference acceleration based on RAG knowledge bases and Unified Cache. OceanStor A600 provides a built-in multimodal AI knowledge base* for industry-leading QPS. This enables faster knowledge base creation and more accurate queries.

OceanStor A600 significantly accelerates inference thanks to its Unified Cache feature, which offers multi-level KV cache from on-chip memory to SSDs. This establishes a PB-level global shared resource pool for KV cache, enabling flexible allocation across multiple compute nodes and xPUs and eliminating repeated computing. This slashes the Time to First Token (TTFT) by 78% while improving the throughput by 60%. It also uses sparse attention algorithms for retrieval to compress tokens to enable the inference of ultra-long sequences beyond the context window, ultimately making inference smoother, faster, and more cost-effective.

Key Specifications

Model	OceanStor A600
Hardware Architecture	Disk and controller integration
Maximum Raw Capacity per Controller Enclosure	1105.92 TB
Height of Each Controller Enclosure	2 U
Controllers per Controller Enclosure	2
Disks per Controller Enclosure	36
Processors per Controller	Two Kunpeng processors
Maximum Memory per Controller	512 GB
Data Disk Type	Palm-sized NVMe SSD
Network Type	• 25/100 Gbit/s TCP/IP • 25/100 Gbit/s RoCE
Key Features	Quota (SmartQuota), Quality of Service (SmartQoS), inference acceleration (Unified Cache), end-to-end data integrity check (DIF)
Enclosure Dimensions (H x W x D)	86.1 mm × 447 mm × 950 mm (including the cover)
Operating Temperature	–60 m to +1800 m altitude: 5°C to 30°C (cabinet)/35°C (enclosure) 1800 m to 4000 m altitude: The max. temperature threshold decreases by 1°C for every altitude increase of 220 m
Operating Humidity	10% to 90% RH

* If this feature is required, contact the sales personnel.

To learn more about Huawei storage, please contact your local Huawei office or visit the Huawei Enterprise website: https://e.huawei.com.

Huawei IT

Copyright © Huawei Technologies Co., Ltd. 2025. All rights reserved.

No part of this document may be reproduced or transmitted in any form or by any means without the prior written consent of Huawei Technologies Co., Ltd.

Trademarks and Permissions

, **HUAWEI,** and are trademarks or registered trademarks of Huawei Technologies Co., Ltd.
Other trademarks, product, service and company names mentioned are the property of their respective holders.

Disclaimer

The content of this manual is provided "as is". Except as required by applicable laws, no warranties of any kind, either express or implied, including but not limited to, the implied warranties of merchantability and fitness for a particular purpose, are made in relation to the accuracy, reliability or contents of this manual.

To the maximum extent permitted by applicable law, in no case shall Huawei Technologies Co., Ltd be liable for any special, incidental, indirect, or consequential damages, or lost profits, business, revenue, data, goodwill or anticipated savings arising out of, or in connection with, the use of this manual.

HUAWEI TECHNOLOGIES CO., LTD.
Bantian Longgang District
Shenzhen 518129, P.R. China
Tel: +86-755-28780808

www.huawei.com