Router-switch.com

NVIDIA A100 GPU Datasheet

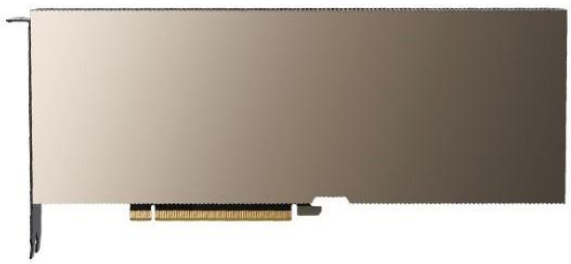
Product Overview

- The NVIDIA A100 is part of the NVIDIA A100 Tensor Core GPU series, designed to deliver extraordinary acceleration for high-performance computing (HPC) and data analytics applications.
- This series is known for enabling the highest-performing elastic data centers, providing versatility for a wide range of workloads.
- The A100 supports a broad array of precision formats, including FP64 (double precision) and INT8 (integer), making it suitable for diverse computational tasks.
- As part of the broader NVIDIA data center platform, the A100 integrates seamlessly with hardware, networking, software, and AI models from the NVIDIA NGC™ catalog, providing a complete solution for researchers and IT managers to drive real-world results and deploy large-scale production systems.
- This makes the A100 a powerful choice for delivering the highest performance across AI and HPC applications.

Product Highlights

- **NVIDIA Ampere Architecture:** Versatile A100 supports MIG and NVLink for efficient GPU utilization across workloads of all sizes.
- **Third-Generation Tensor Cores:** 20X faster deep learning performance compared to Volta GPUs.
- **Next-Generation NVLink:** Enabling up to 16 A100 GPUs with 600 GB/sec bandwidth.
- **High-Bandwidth Memory (HBM2e):** 80GB memory delivering over 2TB/s bandwidth, 1.7X faster than previous generations.
- **Multi-Instance GPU (MIG):** Partition A100 into up to seven instances, maximizing resource utilization.

Detailed Features



A100 GPU

Component	Specification
Turing Tensor Cores	312
NVIDIA CUDA® cores	6,912
Interconnect	Gen4 x16 PCIe, NVLink
Bandwidth	2 TB/s
Power	400 watts

NVIDIA Ampere Architecture

The NVIDIA A100 leverages the groundbreaking NVIDIA Ampere architecture, delivering an unparalleled combination of performance, efficiency, and versatility. Designed for AI, HPC, and data analytics, it enables cutting-edge advancements in computing.

CUDA Cores and Tensor Cores

The A100 integrates 6,912 CUDA cores and 432 Tensor Cores, achieving up to 19.5 teraflops of FP32 performance and 312 teraflops for Tensor operations. These cores optimize AI training, inference, and large-scale parallel computing workloads with superior efficiency.

High-Bandwidth Memory (HBM2)

Equipped with 40GB or 80GB of HBM2 memory and a bandwidth of 1.6 TB/s, the A100 ensures seamless data transfers and high memory throughput. This capability is vital for handling extensive datasets, AI model training, and high-resolution simulations without bottlenecks.

Multiple Precision Formats

The A100 supports multiple precision formats, including FP64, FP32, FP16, and INT8, making it adaptable for a wide range of computational tasks, from scientific computing to machine learning.

Tensor Core Technology

Optimized for deep learning tasks, the A100's Tensor Cores excel in matrix operations, boosting the efficiency of training and inference for AI models.

Energy Efficiency

The A100 is engineered for high performance while maintaining power efficiency. With its multi-instance GPU (MIG) capability, organizations can maximize GPU utilization, optimizing power consumption while delivering high computational output across multiple workloads.

Scalability

The A100 supports multi-GPU configurations, allowing for scalable deployments. This flexibility ensures the A100 can meet the growing demands of AI, ML, and HPC applications.

Accelerated Workloads

The A100 accelerates a variety of workloads, including training large AI models, running simulations, and processing massive datasets. With its ability to deliver unmatched performance in a wide range of tasks, the A100 is a go-to solution for industries such as healthcare, automotive, and finance that rely on complex computing.

Data Center Integration

Designed for seamless integration, the A100 is ideal for modern data centers. It provides the computational power needed for high-performance computing tasks and AI applications at scale. With its flexibility and efficiency, the A100 can easily integrate into existing infrastructures, helping organizations achieve their goals in AI and HPC.

Technical Specifications

Product Specifications

Feature	Value
Product SKU	P1001 SKU 230
NVPN	699-21001-0230-xxx
Thermal Solution	Passive
Mechanical Form Factor	Full-height, full-length (FHFL) 10.5”, dual-slot
Total Board Power	300 W (default) / 300 W (maximum) / 150 W (minimum)
GPU Clocks	Base: 1065 MHz / Boost: 1410 MHz

Feature	Value
PCI Express Interface	PCI Express 4.0 ×16 / Lane and polarity reversal supported
Weight	Board: 1170 g (excluding bracket, extenders, and bridges) / NVLink Bridge: 20.5 g per bridge (×3 bridges) / Bracket with screws: 20 g / Long offset extender: 48 g / Straight extender: 32 g
SR-IOV Support	Supported -- 20 VF (virtual functions)
NVIDIA CUDA® Support	CUDA 11.4 or later
NVFlash	Version 5.695 or later

Memory Specifications

Feature	Value
Memory Clock	1512 MHz
Memory Type	HBM2e
Memory Size	80 GB
Memory Bus Width	5120 bits
Total Board Power	300 W (default) / 300 W (maximum) / 150 W (minimum)
Peak Memory Bandwidth	Up to 1.94 TB/s

Product Comparison

Feature	NVIDIA Tesla P100	NVIDIA Tesla V100	NVIDIA A100
GPU Codename	GP100	GV100	GA100
GPU Architecture	NVIDIA Pascal	NVIDIA Volta	NVIDIA Ampere
GPU Board Form Factor	SXM	SXM2	SXM4
SMs	56	80	108
FP32 Cores / SM	64	64	64
FP32 Cores / GPU	3584	5120	6912
GPU Boost Clock	1480 MHz	1530 MHz	1410 MHz
Texture Units	224	320	432
Memory Interface	4096-bit HBM2	4096-bit HBM2	5120-bit HBM2

Feature	NVIDIA Tesla P100	NVIDIA Tesla V100	NVIDIA A100
Memory Size	16 GB	32 GB / 16 GB	40 GB
Memory Data Rate	703 MHz DDR	877.5 MHz DDR	1215 MHz DDR
Memory Bandwidth	720 GB/sec	900 GB/sec	1555 GB/sec
TDP	300 Watts	300 Watts	400 Watts
Transistors	15.3 billion	21.1 billion	54.2 billion
GPU Die Size	610 mm^2	815 mm^2	826 mm^2
TSMC Manufacturing Process	16 nm FinFET+	12 nm FFN	7 nm N7

Accessories

Category	Accessories
Included Accessories	- Passive Heatsink (for air-cooled configurations)
	- PCIe Power Cables
	- Quick Installation Guide
Optional Accessories	- NVLink Bridge: NVLink-3 4-Slot Bridge (P4381)
	- Active Heatsink: A100-SXM4-HS (for specific cooling solutions)
	- Liquid Cooling Kit: A100-LC-KIT (for data center deployments)
	- PCIe Extender Kit: A100-PCIe-EXT
	- Rack Mount Kit: A100-RMK (for enterprise configurations)
	- High-Speed Interconnect Cables: QSFP-DD 400Gb/s DAC & AOC

Support & Warranty

24/7 Online Service

3-Year Premium Warranty

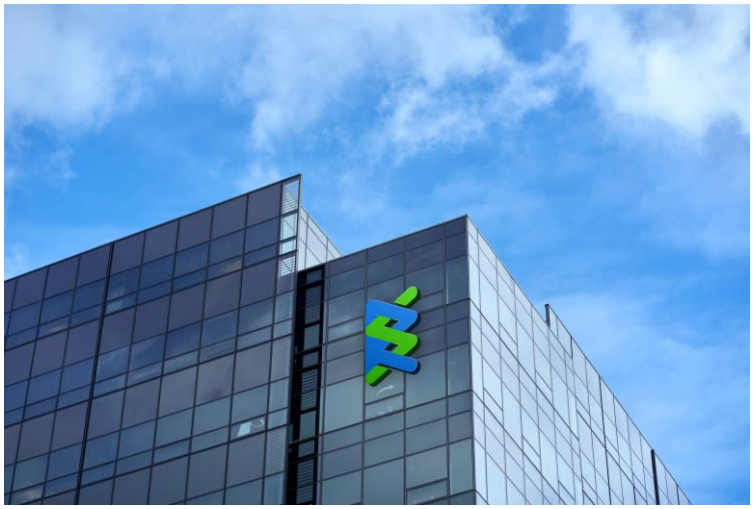
Professional Technical Support

100% Low Price Guarantee

100% Quality Assurance

100% Money Back Guarantee

About Us



Router-switch.com, headquartered in Hong Kong since 2002, has been a trusted global leader in ICT distribution for 23 years. We provide cutting-edge networking, cybersecurity, data center, and AI solutions to meet evolving business needs. Our wide range includes products from top brands like Cisco, Arista, Aruba, Fortinet, Mellanox, and Huawei, ensuring access to the latest technology and innovations.

18,000+ global customers	**600,000+** end-users	**200+** countries & regions	**23** years experience
500+ global vendors	**100,000+** SKUs available	**700+** local sales experts	**50-98%** off global list prices

Contact Us

Email

Sales Inquiries: sales@router-switch.com
Expert Technical Support: ccie-support@router-switch.com
Cooperative Partnerships: partner@router-switch.com

Phone

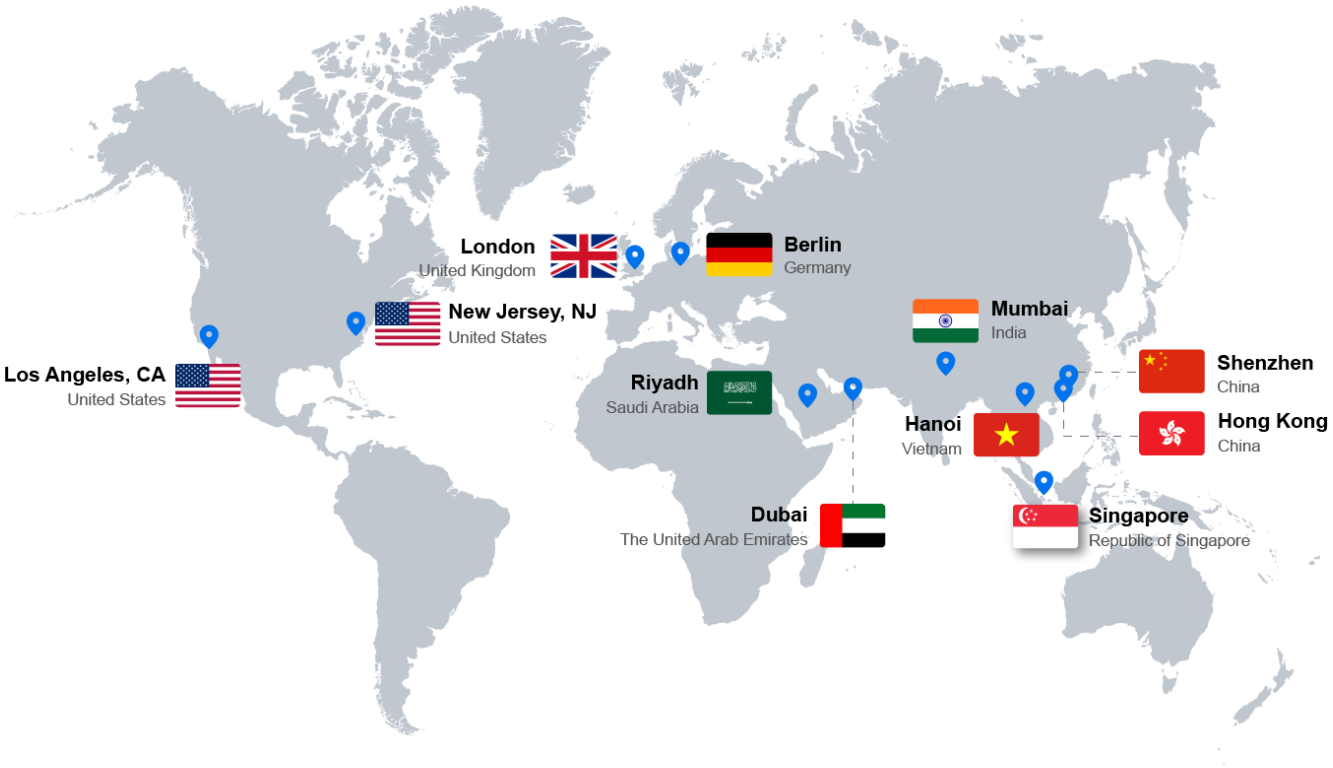
USA: +1-626-655-0998
Hong Kong: +852-25925389 / +852-25925411

Follow Us

Facebook: @Routerswitchdotcom
LinkedIn: Router-switch.com
X: @routerswitchcom
Instagram: @routerswitchdotcom

Global Footprint

Global Warehouses & Service Centers Across Continents.



Global Branches

Headquarters (Hong Kong)

Rm 605, 6/F, Fa Yuen Comm Bldg, 75-77 Fa Yuen St, Mongkok, Kowloon, Hong Kong, China

Shenzhen Branch

Jingfeng Building, 1001 Shangbu South Road, Futian District, Shenzhen, China

USA Branch

3592 Rosemead Blvd, B #220, Rosemead, CA 91770, USA

UK Branch

Third Floor, 207 Regent Street, London W1B 3HH, UK

References

[1] NVIDIA. (2020). NVIDIA Ampere Architecture Whitepaper. Retrieved from https://images.nvidia.com/aem-dam/en-zz/Solutions/data-center/nvidia-ampere-architecture-whitepaper.pdf
[2] NVIDIA. (2020). NVIDIA A100 Tensor Core GPU Datasheet. Retrieved from https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-nvidia-us-2188504-web.pdf