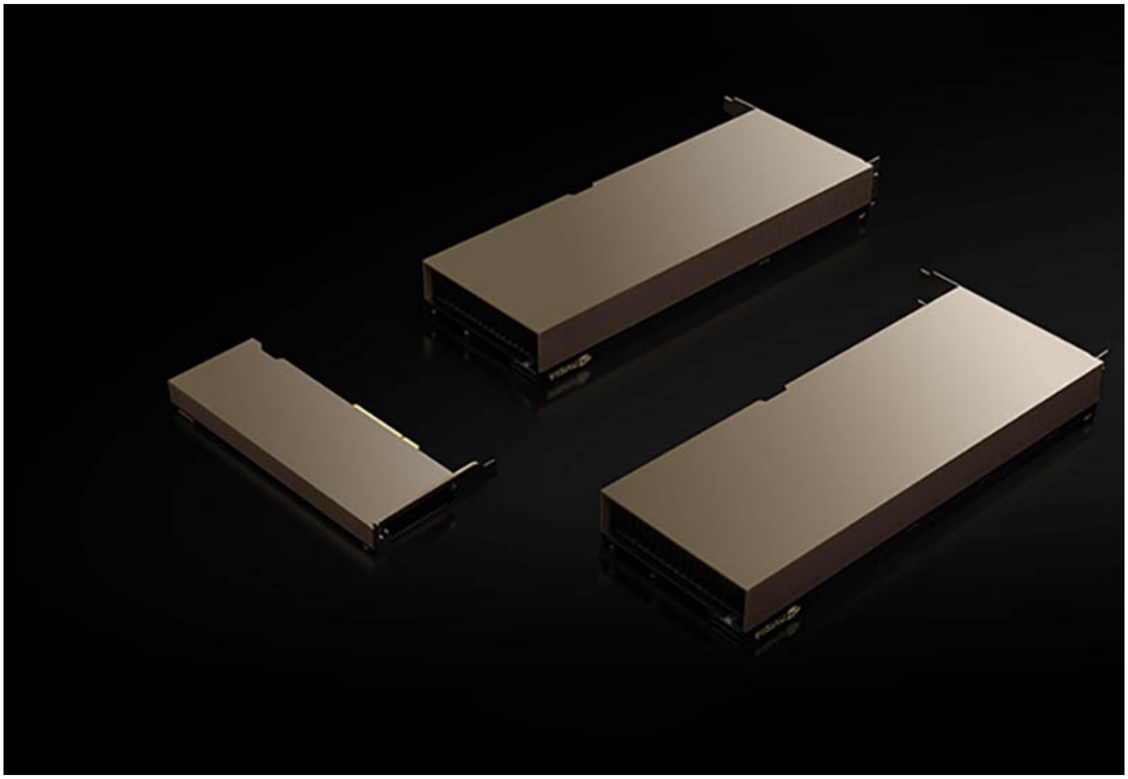
NVIDIA A2 Low-Profile AI Inference Card Replaces the NVIDIA T4

By **Cliff Robinson** - November 9, 2021



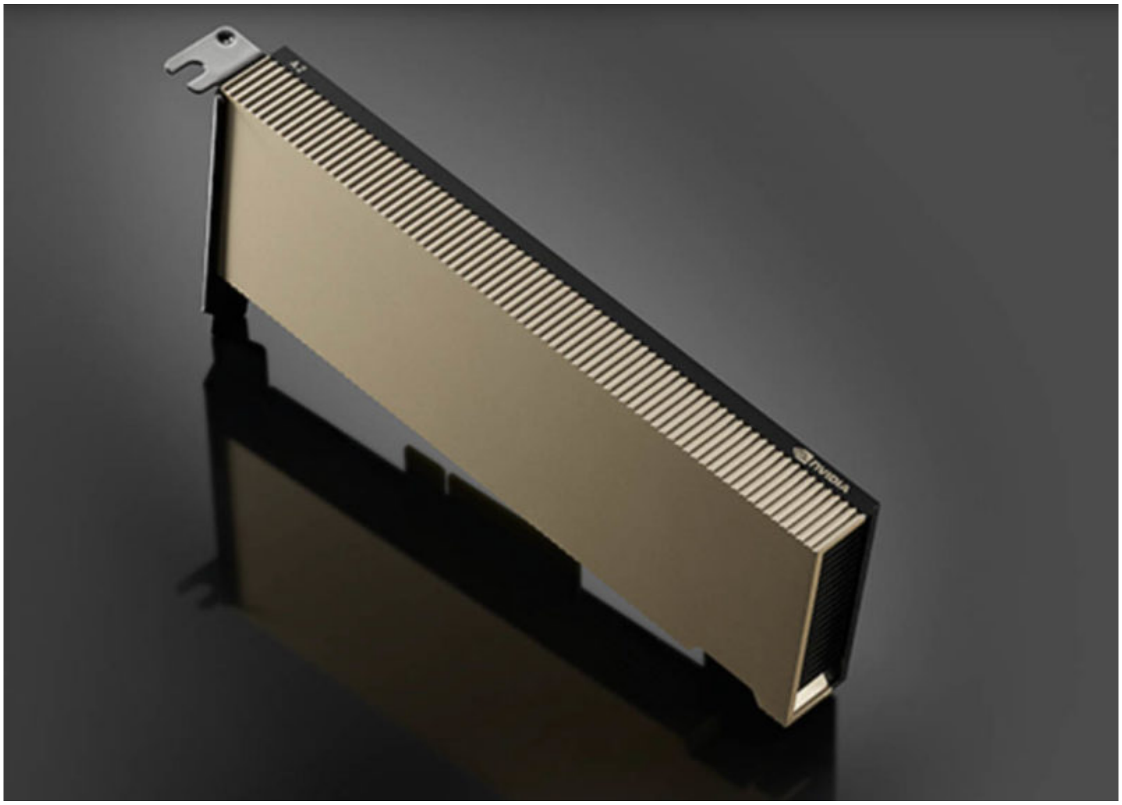
NVIDIA A2 With A100 And A30 Cover

One of the big announcements at today's GTC Fall 2021 is on the GPU side, but not where many may have expected is with the NVIDIA A2. While at the high-end we have seen replacements for the data center GPUs, along with the workstation, and even consumer markets, the one area that we have not seen get an update is the AI inference space. More specifically, the low-profile AI inference card space. The NVIDIA A2 finally updates the segment held by the NVIDIA T4.

NVIDIA A2 and Market Perspective

The NVIDIA A2 is a low power and low profile PCIe card. Specifically the TDP is only 40-60W. The interface is also PCIe Gen4 x8. While one may immediately question why anyone would want a GPU without an x16 connector and such low TDP in the data center, the reason is simply that these are easy to put into severs and both to power and cool.

One of the biggest breakthroughs of the NVIDIA T4 that we reviewed was that it was a low-profile entry into the NVIDIA line-up. It could therefore be physically slotted into places normally reserved for NICs and other typically low-profile devices. The NVIDIA T4 also had a double-digit TDP. Like the NVIDIA A2, it is designed to be powered off of the PCIe bus instead of requiring external power connections. This helps to improve airflow in a chassis and reduces the system requirements for being able to integrate the A2. With lower power, it also means that the A2 can be placed into servers at the edge where there may be tighter power envelopes to adhere to.



NVIDIA A2

We are going to put a full set of key specs below, but the card also has 16GB of GDDR6 memory. One feature we did not see listed is we did not see MIG (many-instance GPU) support listed. Previously, NVIDIA's position was that one would be more likely to use a larger GPU with MIG to get AI inference in edge servers. While that may work for density, it does not work for those cases where there is only 40-120W available and one needs 1-3 inference cards. That seems to be the market that NVIDIA is targeting here.

Final Words

Overall, this is a good step for NVIDIA. It also feels very late in the cycle. PCIe Gen4 is going to be replaced by PCIe Gen5 in the data center starting in Q2 2022. NVIDIA says the cards are available today, but there is only a short window until we see Gen5 devices at this point. Indeed, we already have been looking at PCIe Gen5 on the desktop. In the meantime, the market has had to use the NVIDIA T4 or use MIG with larger GPUs. This is Ampere coming to a popular and established market segment around six quarters after the NVIDIA A100 launched.

Still, we are excited to see more NVIDIA A2 servers, as the T4 has been extremely popular.

NVIDIA A2 Key Specs

Here are the NVIDIA A2 key specs from NVIDIA’s website as of launch:

Peak FP32	4.5 TF
TF32 Tensor Core	9 TF | 18 TF[1]
BFLOAT16 Tensor Core	18 TF | 36 TF[1]
Peak FP16 Tensor Core	18 TF | 36 TF[1]
Peak INT8 Tensor Core	36 TOPS | 72 TOPS[1]
Peak INT4 Tensor Core	72 TOPS | 144 TOPS[1]
RT Cores	10
Media engines	1 video encoder 2 video decoders (includes AV1 decode)
GPU memory	16GB GDDR6
GPU memory bandwidth	200GB/s
Interconnect	PCIe Gen4 x8
Form factor	1-slot, low-profile PCIe
Max thermal design power (TDP)	40–60W (configurable)
Virtual GPU (vGPU) software support[2]	NVIDIA Virtual PC (vPC), NVIDIA Virtual Applications (vApps), NVIDIA RTX Virtual Workstation (vWS), NVIDIA AI Enterprise, NVIDIA Virtual Compute Server (vCS)

Cliff Robinson

They call me "the STH news guy" for a reason.