NVIDIA ADA CRAFT

The engineering marvel of the RTX 4090.

V1.0

Table of Contents

List of Figures

An Engineering Marvel

When the NVIDIA Ampere architecture was launched in 2020, excitement surrounding the technical achievements and workload accelerations was accompanied by another, surprise topic: aesthetics. While the Turing architecture marked a full departure from the enclosed blower-driven coolers, Ampere architecture was something different—beautiful, memorable, cool, and quiet. A sleek ribbon of metal bristling with blackened heatsinks that encircled deep, opposing fans. It was elegance achieved through meticulous engineering—and thoroughly redeveloped for the new NVIDIA Ada architecture.

The NVIDIA Ampere generation introduced a number of design rules, patterns, and principles that would serve multiple GPU generations. The symmetrical “back is the front” motif, where every side of the GPU was attractively designed, was also part of this vocabulary. This evergreen design language cultivates an iconic design based on simple forms, real metals, and enduring finishes. It is also about celebrating what was previously hidden. Heatsink fins, for example, used to be hidden behind shrouds or fans. Instead, for Ada architecture, the guiding mantra was: “Less, but better.”

Materials

Form is driven by function across the Ada family. The material selection process for every component needs to consider a variety of mechanical, thermal, and physical requirements ranging from electrical conductivity to thermal expansion to harmonic resonance. The guiding principle is that quality should never be compromised in the name of aesthetics. If a material looks good but affects quality, then it is a poor material choice. If you want something to look like metal, use an actual metal.

Figure 1. RTX 4090 Thermal Cooling

The minimalist frame wraps everything in a single seamless structure and is built from a strong, lightweight, and durable extruded aluminum alloy that can receive a variety of treatments. Beyond creating an enduring and beautiful finish, anodizing treatments for the aluminum frame allowed us to create better-distinguished product tiers, with different frame shapes as well as different colors, finishes, and accents. Nestled within the frame are heatsink fins, which are emphasized (rather than hidden away) and are spaced slightly further apart as the cooler height increases to maintain consistent airflow and pressure. They are made from a 99% pure aluminum alloy that is lightweight and rigid with great thermal conductivity.

Attention to detail can be found in subtler forms, like the magnetically latched extender cover and the aerodynamics-inspired curve on the 3-slot coolers, which lends both a sleek appearance and reduces weight. Product logos used to be manufactured as separate parts and then fused into the frame by a CNC-machined gap, but manufacturing processes and tolerances meant that visible or tactile separations would be left over. With Ada architecture board design, we can injection-mold the logo right into the metal frame. This eliminates the panel gaps to better create that seamless, premium look.

Mechanics

Consisting of 18432 CUDA Cores and over 77 billion transistors, the AD102 GPU used in our flagship Ada-based graphics card—the GeForce RTX 4090—provides incredible performance for graphics and compute workloads. Next to our own Hopper H100 data center GPU, it is the most

powerful GPU architecture the world has ever seen. NVIDIA thermal engineers pushed even harder to maximize the performance of the new cooler, to deliver the most efficient thermals, acoustics, and power.

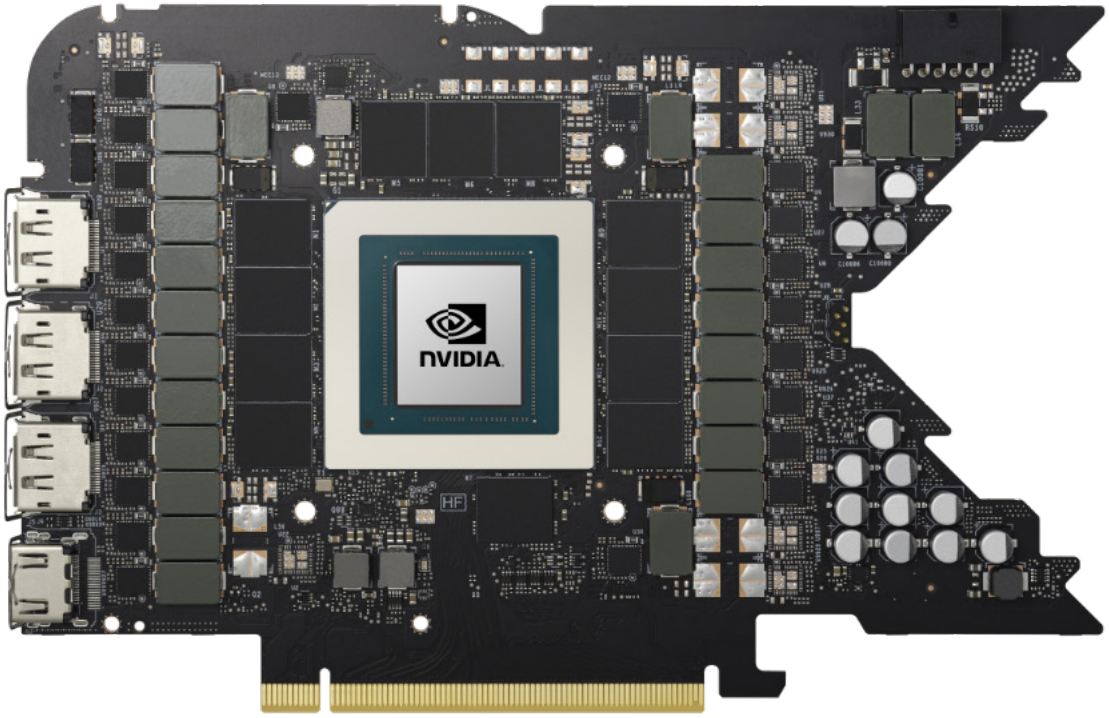

Figure 2. RTX 4090 AD102 GPU

For the previous Ampere architecture flagship GPU, the GeForce RTX 3090, NVIDIA engineers implemented a dual-axial flowthrough design, which used airflow to cool the board fins. The new RTX 4090 coolers may look similar to the RTX 3090 at first glance, but NVIDIA engineers have created an all-new design that has been tailored to increase airflow as much as possible and make the heatsink more effective.

The design of the RTX 4090 delivers the highest airflow ever measured in a discrete NVIDIA GPU, clocking in at 80 cubic feet per minute—enough to inflate about 3.5 regulation basketballs every second. This design achieves 20% more airflow than an RTX 3090, all thanks to the purposeful designs of the heatsink and fans.

Fan affinity laws show that airflow increases cubically with fan diameter, so by carefully designing the unibody frame and its cross section, fan sizes have grown throughout the Ada generation for incredible cooling performance. On their own, larger fans would provide higher airflow without changing acoustics, but swapping to fluid dynamic bearings and using counter-rotating fans results in cooler and quieter operation. The blades of the fans are formed from glass fiber-reinforced plastic, which is resilient and easy to manufacture, while the fan hub uses fluid dynamic bearings to improve acoustic performance, durability, and reliability.

Across the Ada family, the distance between heatsink fins is slightly different depending on the height (or “thickness”) of the cooler, as taller fin stacks need to have more distance between the fins to maintain consistent airflow and pressure. The distances between the fins are approximately 1.7mm for 2-slot coolers and 2.0mm for 3-slot coolers.

Generational Changes

Compared to the GeForce RTX 3090, the RTX 4090 has the following electrical differences:

- PCIe Gen 5 (ATX 3.0) powered with smart dongle which configures power limits based on the number of 8-pin connectors used.
- Shifted the GPU chip north to improve power layout (impedance balance between phases)
- Added two PCB layers to enhance power efficiency
- Improved G6x signal integrity with better ground isolation from power supply noise (2 additional PCB layers) and upgraded the dielectric materials to also improve G6x signaling (IT150GS to NPG-170D)
- Added four additional core power phases (from 16 to 20) for greater performance and efficiency

The RTX 4090 has the following thermal/mechanical design changes from RTX 3090:

- Larger 7-blade fans (from 110mm to 116mm) selected after studying over 50 different 9-blade and 7-blade fans with different blade shapes to optimize for airflow and acoustics. Fin pitch relaxed from 1.7mm to 2.0mm. Airflow increased by 20% (20% reduction in exhaust temperature to ambient delta).
- Increased height (2.7 slot to 3 slot), reduced length (12.3 inch to 12 inch).
- Notched memory pedestal provides more even GPU to heatsink contact which improves overall cooling efficiency (reduces unbalanced pressure from memories). 2mm to 1.5mm memory TIM thickness to improve memory cooling.
- Optimized vapor chamber and heat pipes:
    - Increased the number of heat pipes extending to the east fan from 4 to 6 to increase temperature uniformity on fins and also increase Qmax capability.
    - Changed to a bigger vapor chamber to cover entire west fin to increase temperature uniformity, and removed 2 additional heat pipes on west side; amount of total heat pipes the same as RTX 3090
    - Optimized vapor chamber internal design (Reduce evaporator wick thickness to 0.4mm to reduce thermal resistance) to support >650W Qmax (10% higher peak cooling capacity before vapor chamber dry-out)
    - Increase powder pillars density in die area to increase Qmax and reduce thermal resistance, thicker vapor chamber base wall (to 1.0mm). Total increase vapor chamber strength by 43%
    - More water in vapor chamber to increase Qmax
- Changed from ball bearing fans to fluid dynamic bearing fans for improved acoustics
- TIM visual blocking plate and moved capacitors away from the edge of the board for a cleaner appearance.