# Warm-Water Ready: Benchmarking Two-Phase Vs. Single-Phase Direct-to-Chip Cooling On An NVIDIA B200 Server

An apples-to-apples comparison on a Dell PowerEdge XE9680L shows NeuCool® two-phase cooling running GPU cold plates up to 19°C cooler than the factory single-phase system—and 9-14°C cooler at the full-system level.

ACCELSIUS®

info@accelsius.com

www.accelsius.com

# Executive Summary

As AI racks climb toward and beyond 100 kW+ and operators pursue warm-water free cooling to cut energy costs, the cooling technology inside the server increasingly determines both compute performance and facility economics. To quantify the difference between cooling technologies on identical hardware, an independent third-party systems integrator benchmarked a commercially available Dell PowerEdge XE9680L server with eight NVIDIA B200 GPUs (~10 kW TDP), first with its factory-installed single-phase, water-based direct-to-chip cooling system, then retrofitted with the Accelsius NeuCool® two-phase direct-to-chip system.

Across approximately 40,000 operating points, two-phase cooling ran the thermally limiting GPU cold plate up to 19°C cooler than the stock single-phase solution, and delivered 9–14°C lower junction temperatures at the full-system level depending on facility water flow rate — while using roughly one-third the coolant flow at the cold plate. These margins translate directly into higher allowable facility water temperatures, year-round free cooling in most climates, lower total cost of ownership, and thermal headroom for next-generation chips and rack architectures.

# Why This Test Matters

Cooling comparisons are often muddied by differences in servers, workloads, and instrumentation. This program eliminated those variables: the same server, the same GPUs, and the same simulated AI workloads were used for both cooling technologies. The only change was the cooling system itself. The stock single-phase water loop was removed and replaced with NeuCool two-phase cold plates and manifolds — a true apples-to-apples comparison.

The stakes are rising. NVIDIA has pointed the industry toward warmer facility water temperatures as a key lever for AI factory efficiency, since every 1°C increase in facility water temperature yields roughly 4% in annual cooling energy savings and expands the geography where free cooling (no chillers) is possible. A cooling technology that preserves GPU performance at 50°C+ facility water is therefore worth more than the sum of its degrees: it lowers PUE, defers chiller capex, and lets every rack convert more of its power budget into tokens rather than heat management.

**A cooling technology that preserves GPU performance at 50°C+ facility water is worth more than the sum of its degrees.**

# Test Methodology

An independent third-party systems integrator conducted the benchmark on a Dell PowerEdge XE9680L eight-way NVIDIA B200 server with a total thermal design power of approximately 10 kW. The evaluation swept approximately 40,000 operating points by independently varying GPU core power, high-bandwidth memory (HBM) power, and CPU power under simulated AI workloads, producing a comprehensive thermal map of realistic operating conditions.

Single-phase data were captured using the factory cooling loop served by a CDU operating at a 4°C approach temperature. Two-phase data were captured with the same server retrofitted with NeuCool cold plates, served by an Accelsius NeuCool IR150 CDU. Both configurations were evaluated against common facility water conditions, including a 50°C facility water temperature representative of warm-water, chiller-less operation.

**Figure 1.** Before and after. Left: the tested Dell PowerEdge XE9680L GPU tray with its factory-installed single-phase cold plates. Right: the same server retrofitted with Accelsius NeuCool two-phase cold plates. All other server hardware was unchanged.

# Results

## Cold Plate Performance

At the cold plate level, the two technologies are separated by the physics of vaporization. In the XE9680L, coolant flows through two GPU cold plates in series, making the second (downstream) cold plate the thermally limiting position. On that second cold plate, NeuCool two-phase cooling held the GPU 19°C cooler than the stock single-phase system. Because two-phase cooling absorbs heat through latent heat of vaporization rather than sensible heating alone, the downstream cold plate sees little of the temperature penalty that a series water loop imposes.

Averaged across positions, the NeuCool B200 cold plate achieved a thermal resistance of 0.0171 K/W versus 0.0328 K/W for the stock single-phase plate — a 48% reduction — while operating at 0.55 LPM/kW of coolant flow versus 1.5 LPM/kW for single-phase, roughly one-third the flow. Lower flow means smaller pumps, less pumping energy, and less erosion-related wear on fittings and cold plates over the life of the system.

Table 1 summarizes the measured cold plate metrics:

| Metric | Single-Phase (Stock) | Two-Phase Universal* | Two-Phase (NeuCool B200) |
|---|---|---|---|
| Average thermal resistance, $R_{th}$ (k/W) | 0.0328 | 0.0258 | 0.0171 |
| Cold plate volumetric flow rate (LPM/kW) | 1.5 | 0.55 | 0.55 |
| ΔT, chip junction—cold plate inlet (°C) | 34.8 | 26.1 | 16.7 |

**Table 1.** Measured cold plate thermal performance, stock single-phase vs. NeuCool two-phase, on the Dell PowerEdge XE9680L B200 server. *Two-phase universal cold plate values extrapolated from B200 universal cold plate thermal test vehicle (TTV) testing.

## System-Level Performance

Cold plate performance only matters if it survives the full thermal stackup from chip junction to facility water. When all thermal resistances from the cold plate through the CDU are included, the two-phase advantage narrows but remains decisive. At 50°C facility water and a facility flow rate of 1.5 LPM/kW, the two-phase system held GPU junction temperatures 9°C below the single-phase system — enough that single-phase operation exceeded the B200's 84°C throttle point while NeuCool maintained 9°C of thermal headroom below it.

Two-phase cooling also capitalizes on facility-side improvements in a way single-phase cannot. Raising facility water flow from 1.5 to 3.0 LPM/kW widened the two-phase advantage from 9°C to 14°C; single-phase realized no comparable gain. Put in facility terms: the performance single-phase delivers at 45°C facility water, NeuCool matches at 54°C — and at the higher facility flow rate, at up to 59°C. Those are facility water temperatures at which free cooling is available in most locations around the world for most of the year.

**Two-phase cooling capitalizes on facility-side improvements in a way single-phase cannot.**

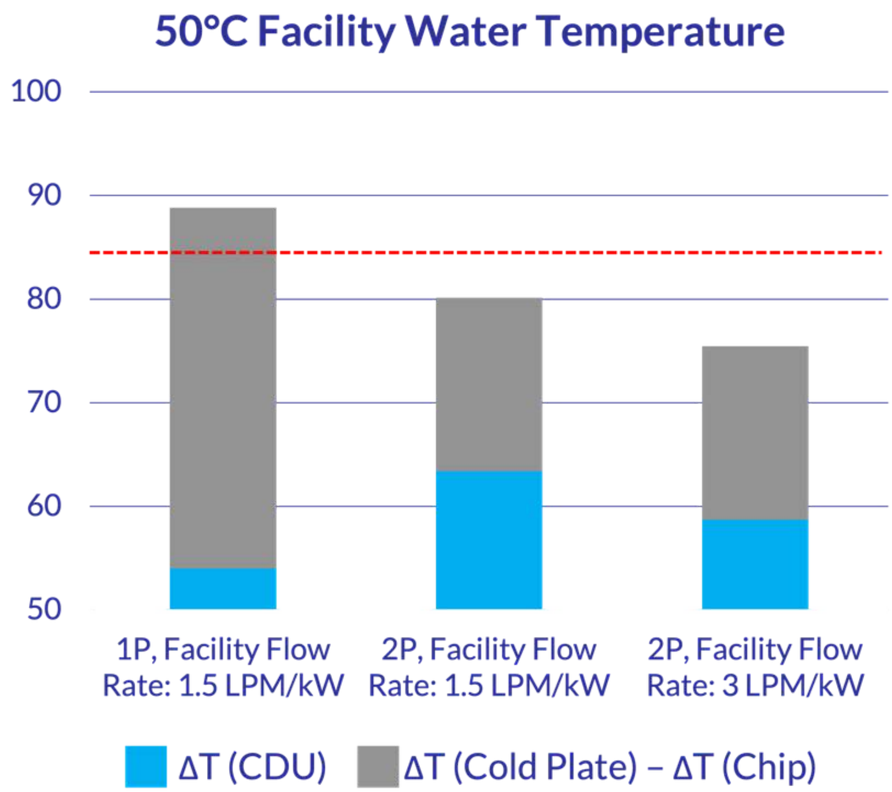


**Figure 2.** GPU junction temperature stackup at 50°C facility water, showing the contribution of the CDU approach and the cold-plate-to-chip resistance for three cases: single-phase at 1.5 LPM/kW facility flow, two-phase at 1.5 LPM/kW, and two-phase at 3.0 LPM/kW. The dashed red line marks the B200's 84°C throttle point; only the single-phase configuration exceeds it.

# What This Means For AI Factories

These results reinforce what Accelsius has long maintained: as facility water temperatures rise, two-phase cooling pulls away from single-phase, because the physics of vaporization deliver disproportionately greater benefit under warm-water conditions. For operators, the 9–14°C system-level margin can be spent in several ways. Taken as warmer facility water, it is worth roughly 36–56% in annual cooling energy savings at ~4% per °C, and it converts chillers from a necessity into a contingency. Taken as thermal headroom, it keeps GPUs safely below throttle limits at sustained boost clocks, protecting training throughput and inference token economics. And taken as design margin, it provides a path to next-generation chips and denser rack architectures whose TDPs will overwhelm single-phase water loops — without re-plumbing the facility.

Because these gains were demonstrated on unmodified, commercially available server hardware by an independent third party, they represent a retrofit-ready path: the same server, running cooler, at warmer water, with one-third the coolant flow at the chip.

# Conclusion

In a controlled, apples-to-apples benchmark on a Dell PowerEdge XE9680L NVIDIA B200 server, Accelsius NeuCool two-phase direct-to-chip cooling outperformed the factory single-phase system by up to 19°C at the thermally limiting cold plate and by 9–14°C at the system level, cutting cold plate thermal resistance nearly in half at one-third the coolant flow. Two-phase cooling is not an incremental improvement on single-phase — it is the enabling technology for warm-water AI factories, delivering lower total cost of ownership, material energy savings, and the heat removal capacity that next-generation chip and rack architectures will demand.

**Two-phase cooling is not an incremental improvement on single-phase — it is the enabling technology for warm-water AI factories.**

# About Accelsius

Accelsius empowers data center operators and AI neoclouds to achieve their business, financial and sustainability goals through advanced cooling solutions. The proprietary NeuCool® platform provides best-in-class thermal efficiencies through a safe, two-phase, direct-to-chip liquid cooling system that scales from single racks to entire data centers. For more information, visit accelsius.com or follow us on LinkedIn.

ACCELSIUS®

info@accelsius.com

www.accelsius.com