DENVR

UNICOM Engineering, Inc.
A Division of UNICOM Global

DEPLOYING IMMERSION-NATIVE ITE IN NEOCLOUDS

Design, Validation, and Deployment at Scale

Amy Short, PhD | Executive Officer, Advanced Data Center Technologies, Denvr
Austin Hipes | Chief Technologist & VP, Engineering, UNICOM Engineering

Abstract

This paper documents the engineering methodology, validation approach, and operational outcomes of a production-scale deployment of immersion-native IT equipment (ITE) within a NeoCloud environment. The work was executed through a joint program between Denvr and UNICOM Engineering, combining facility-level and cluster-level operational data from Denvr with component- and server-level hardware design and validation from UNICOM Engineering. The resulting system is an immersion-native, H200-optimized platform validated for high-density AI deployment.

The deployment addresses a specific and growing engineering constraint: as GPU Thermal Design Power (TDP) exceeds 700 W per component and rack-level power densities grow beyond 100 kW, conventional forced-air cooling is no longer a viable thermal management strategy at scale. Water-based direct to chip present their own challenges related to performance, reliability, and facility requirements. For many use cases, single-phase dielectric immersion cooling resolves these constraints. Further, when implemented with purpose-designed hardware, we were able to achieve a 2x reduction in server form factor, a ~4x server density per rack, power savings of approximately 800 W per server, and an Open Compute Project Open Rack (ORV2.2 and ORV3) power configuration. All of this was supported in a zero-water Denvr modular data center with PUE below 1.04.

Validation methodology across both sites covered component, server, system, and AI production workloads, including NVQual stress testing and ambient thermal stress testing to 60°C. Both teams targeted additional testing related to known or perceived concerns in IT equipment performance in an immersion environment including compatibility, thermal and power throttling, and signal integrity. It proved performance under even the most stressful conditions, including power, thermal and signal integrity function as expected. The division of responsibility between the two organizations – UNICOM Engineering at the hardware and component level, Denvr at the system and production level – reflects a collaborative model that the authors consider as transferable as the technical findings themselves.

Table of Contents

Thermal and Power Constraints in High-Density AI Compute

Modern GPU platforms present thermal management challenges that fall outside the design envelope of conventional forced-air cooling systems. The H100 and H200 SXM platforms, for example, operate at TDPs approaching and exceeding 700 W per GPU, with eight-GPU server configurations generating sustained thermal loads in the range of 10–12 kW per server. At rack densities required for large-scale AI training clusters, these loads cannot be effectively managed through recirculated air without either accepting significant thermal headroom penalties, throttling-induced performance degradation, or impractical physical separation between compute nodes.

The fundamental physics of the constraint is straightforward. The convective heat transfer coefficient of air, typically 10–100 W/m²K under forced convection, is several orders of magnitude lower than that of dielectric immersion fluids operating in single-phase forced convection (roughly 200–1,000 W/m²K). This differential means that a given thermal load requires proportionally less heat exchange area, lower temperature gradient, and less mechanical cooling infrastructure when managed through liquid immersion. The consequence for air-cooled rack design is that thermal capacity, not electrical capacity, is the binding constraint on density.

Five related constraints are converging for AI infrastructure operators:

- **Power density per rack:** Available power per rack exceeds what air cooling can thermally support at the server densities required for large-scale GPU clusters.
- **Thermal variability under dynamic load:** AI training workloads produce highly variable compute utilization across GPUs within a server. In air-cooled environments this creates GPU junction temperature swings of 20–25°C within a single training step, driving thermal fatigue accumulation and warranty risk.
- **Water consumption and sustainability:** Evaporative cooling systems, while effective, carry water consumption liabilities that are increasingly constrained by regional permitting, ESG commitments, and physical availability.
- **Non-traditional deployment environments:** AI workloads must be deployable at stranded power sites, edge locations, and modular facilities where traditional HVAC infrastructure cannot be readily installed.
- **Long-term hardware reliability confidence:** Dielectric immersion introduces material compatibility, signal integrity, and firmware behavior considerations that require rigorous validation before operators can commit to large-scale deployment under standard warranty terms.

> *In air-cooled environments, thermal headroom, not electrical capacity, is the binding constraint on GPU cluster density. The convective heat transfer differential between dielectric fluid and air is three to four orders of magnitude.*

The Production Evidence Gap

Immersion cooling for high-performance compute is not a new concept. Single-phase and two-phase dielectric immersion systems have been characterized in literature and demonstrated in production environments for over a decade. The engineering gap has not been theoretical understanding but operational validation—specifically, the absence of publicly documented production deployments that cover the full system stack from hardware design through sustained production operation under real AI workloads.

This work addresses that gap. It documents a production deployment of single-phase immersion-native ITE supporting H100 and H200 SXM GPU platforms in a live NeoCloud environment. The system has been immersed since April 2024, representing 2 years of continuous exposure to the immersion environment. Validation artifacts span component, server, system, and production levels. The work does not rely on controlled laboratory benchmarks alone; performance and reliability characterization were conducted under operational AI workloads with real utilization profiles.

The engineering decisions documented here—hardware design departures from air-cooled baselines, power architecture choices, signal integrity methodology, and thermal validation approach—are intended to provide a reproducible framework for future deployments rather than to document a single bespoke configuration. Repeatability at the methodological level is the primary deliverable.

Deployment Context: Why Vertical Integration Enables Collaboration

This deployment is hosted within Denvr's NeoCloud infrastructure. Denvr operates as a vertically integrated AI infrastructure provider, with ownership and operational responsibility spanning facility design, power delivery architecture, cooling system, and IT stack. This integration is practically significant for immersion deployment: it enables co-optimization decisions across layers that would otherwise be constrained by organizational and contractual boundaries between a facility operator, a colocation tenant, and an ITE vendor.

Immersion-native deployment requires coordination across all of these layers simultaneously. Server form factor must be compatible with tank geometry; power delivery architecture must match bus specifications at the chassis level; firmware must account for the thermal behavior of the immersion environment rather than forced air; and operational procedures must address fluid handling, component swap methodology, and long-term fluid chemistry. In environments where these layers are owned and operated independently, each interface introduces a constraint that limits what any single party can change. Denvr's vertically integrated model removes those constraints and is part of why this deployment was achievable at this scale and timeline.

Denvr's Modular Data Center (MDC) architecture embodies this integration. The MDC achieves 1 MW per 600 sq ft—a density of approximately 1.67 kW/sq ft—compared to roughly 0.16 kW/sq ft in a conventional air-cooled colocation facility. This 10x density improvement is not achieved through any single technology choice but through systematic co-optimization of the facility thermal loop, power delivery infrastructure, and server hardware design. The MDC is zero-water in operation and is designed for rapid deployment at sites with available power but

without traditional data center HVAC infrastructure, supporting scalability from single-unit deployments to 100 MW campus configurations.

10x	~1.67 kW	0 gal	<1.04
Compute density vs. air-cooled co-lo	per sq ft (MDC)	Water consumption	Measured facility PUE

The target use cases for this infrastructure are AI compute at NeoCloud scale, Cloud Service Provider and Data Center Operator deployments requiring fast and cost-effective AI infrastructure expansion, and sovereign or regulated deployments—including government and public sector environments requiring air-gapped operation, data sovereignty controls, and Canadian security standards compliance—where conventional large-scale colocation facilities are either impractical or unavailable.

Technical Roles and Validation Scope

The technical challenge of deploying immersion-native AI infrastructure spans hardware design, power architecture, cooling engineering, signal integrity, and operational practice. No single organization can validate that full stack in isolation. This program was therefore structured as a bilateral engineering effort, with each organization responsible for the validation layer closest to its operating domain. UNICOM Engineering led the component- and server-level work required to convert conventional air-cooled hardware into an immersion-native platform. Denvr led the system- and production-level validation required to prove that platform under real NeoCloud operating conditions.

This division of responsibility also shaped the validation sequence. Early A100 HPL testing established the failure modes associated with non-optimized immersion implementation. Subsequent H200 HPL testing validated the optimized immersion-native platform under controlled high-stress conditions. Production Llama-3.1-70B inferencing then demonstrated stable behavior under real AI workload conditions.

UNICOM Engineering: Component and Server-Level Engineering

UNICOM Engineering's role encompassed the hardware design, validation, and production-hardening work required to produce an immersion-native server platform from a conventional air-cooled baseline. UNICOM's liquid cooling practice is built on the principle that at GPU TDPs above 350 W per component, liquid cooling transitions from an optimization to a thermal engineering requirement. The company maintains validated immersion-ready platform configurations across Dell Technologies, Intel, and HPE hardware families, with a fluid partner ecosystem that includes ExxonMobil, Lubrizol, and GRC for material compatibility and long-term fluid chemistry validation.

- Physical redesign of the Dell PowerEdge XE9680 platform into the UNICOM Engineering H200 HGX Node immersion-ready configuration, reducing server form factor from 6U to 3OU and shrinking the physical volume by half.
- Power architecture migration from AC distribution to 48V-54V DCbus, ORV 2.2-inspired (see power section for technical detail on the departure from standard ORV bus bar voltage specification).
- Thermal hardware redesign: removal of fans; installation of immersion-optimized copper heat sinks with geometry re-envisioned for single phase dielectric fluid flow dynamics.
- Firmware co-development: updated thermal management, telemetry, and fan-control logic appropriate for an immersion environment where there are no fans and thermal behavior differs substantially from air-cooled operation, including ambient thermal stress testing to 60°C fluid temperature.
- Long-duration compatibility and reliability confirmation.
- Component-level and server-level signal integrity validation via NVQual testing.
- Safety and regulatory certification continuity through the conversion process—a requirement for sovereign and regulated deployment contexts.

Denvr: System and Production-Level Validation

Denvr's role was to validate single server- and cluster-level behavior under real operational conditions in a modular data center environment. Denvr's proprietary server validation test program has been proven in support of R&D innovation activities and practically applied during installation and commissioning of new deployments. Drawing on a background that includes over 15 GW of global hyperscale cloud infrastructure deployment and more than $1 B in advanced AI compute architectures, Denvr applied production NeoCloud operational context—including sustained AI training workloads, non-traditional deployment constraints, and sovereign use case requirements—to validate that the hardware design choices made at the component and server level translated to predictable, reliable cluster behavior.

- Server-level and cluster-level characterization of thermal profiles, power profiles, and performance under controlled and AI production workload scenarios.
- Server-level and cluster-level performance benchmarking in high stress scenarios. This includes controlled and AI production workloads known to apply stress across the server as well as to components of interest (e.g., GPU, CPU, RAM, storage).
- Evaluation of fluid performance and quality through continuous in-line monitoring and physical sampling for high-volume and specialized characterization (e.g., elemental analysis, dielectric properties, breakdown voltage, oxidation, water content, etc.).
- Continuous monitoring and operational impact assessment of the Denvr MDC infrastructure.
- Continuous monitoring of system behavior and performance under long-term production operating conditions.
- Sovereign and regulated use case validation in non-traditional, modular deployment environments.

This division of responsibility allowed each organization to apply its deepest operational expertise at the appropriate validation layer, while maintaining alignment across the full system stack. The combined validation

artifacts—from NVQual test data at the component level through production AI workload benchmarking—provide the multi-level confidence required before committing to large-scale deployment.

Immersion-Native Hardware Design: Departures from Air-Cooled Architecture

Adapting an air-cooled server platform to an immersion environment without systematic hardware redesign introduces several categories of engineering risk: suboptimal thermal performance from heat sink geometry designed for forced-air flow rather than natural or forced convective flow in dielectric fluid; power architecture incompatibilities between AC fan control and DC immersion environments; signal integrity degradation from impedance changes introduced by the dielectric medium; and material compatibility failures from fluid-polymer, fluid-elastomer, or fluid-PCB laminate interactions over extended operational periods.

The XE9680-IR addresses these risks through deliberate, layer-by-layer hardware redesign rather than incremental modification. The H100 and H200 SXM platform was selected as the target compute node given its relevance to large-scale AI training and the severity of its thermal requirements (TDP ~ 700 W per GPU). Validation work was carried out on both H100 and H200 platforms over the course of the project. The redesign covered four principal domains:

Air-Cooled Baseline (6U)	Immersion-Native XE9680-IR (3OU)
6U rack-mount form factor	3OU open-rack form factor; 50% height reduction
AC power input; fan-speed-modulated thermal control	48V-54V DC bus (ORV 2.2- and ORV3-inspired); fan and blower elimination
Aluminum heat sinks; forced-air convection design	Immersion-optimized copper heat sinks; fluid-convection design
Air-cooled firmware; fan telemetry and control	Updated firmware: immersion telemetry, thermal management, no-fan logic
3–4 H200 SXM servers per rack at practical density	13–14 immersion servers per tank; 4x effective rack density

Form Factor and Mechanical Design

The reduction from 6U to 3OU is enabled by the elimination of airflow plenum space required in conventional rack-mount servers. In an air-cooled design, a significant fraction of server height is allocated to air inlet and exhaust geometry, fan assemblies, and the clearance required for effective airflow through heat sink fin arrays. In an immersion environment, fluid contacts component surfaces directly (or via heat sinks), eliminating the need for plenum geometry and enabling a substantially more compact chassis form factor. The result is tank population of 13–14 servers in the same footprint that would accommodate 3–4 in an air-cooled open rack.

Power Architecture

The power architecture was redesigned around a 48V-54V DC bus using three bus bars—an ORV 2.2-inspired configuration with a deliberate departure from the ORV 2.2 specification, which defines bus bars at 12V DC when 3 busbars are utilized, or a single busbar at 50V-54V DC. The 48 V selection was driven by the current-carrying requirements at the power levels required for high-density H200 deployment: at 48 V, the same power delivery requires one-quarter the current of a 12 V bus, substantially reducing I^2R losses and bus bar sizing requirements. The design is extensible toward future ORV 3 specifications as the OCP standards community converges on higher-voltage bus architectures for next-generation AI compute platforms. Over 2000W of fans utilized for air cooling are eliminated, recovering approximately 800 W per server under load in normal usage conditions and blower power circuits were eliminated, recovering approximately 800 W per server under load.

Thermal Hardware: Heat Sink Design for Immersion

Copper was selected over aluminum for heat sink construction based on its superior thermal conductivity (approximately 400 W/m·K for copper versus 205 W/m·K for aluminum) and its compatibility with the engineered dielectric fluids used in this deployment. The heat sink geometry was redesigned for convective heat transfer in dielectric fluid rather than forced-air flow: fin pitch, fin height, and base contact geometry were selected to optimize fluid contact area and promote effective natural or forced convection within the tank, rather than to minimize airflow resistance. The innovations in thermal coupling are primarily internal to the heat sink structure and are not visible from external inspection.

Firmware and Software

Fan control firmware, thermal throttling logic, and operational telemetry all required modification for the immersion environment. In an air-cooled server, thermal management is largely mediated through fan speed adjustment in response to temperature sensor feedback; in an immersion environment, fans are absent and thermal management relies on fluid flow rate and heat exchanger setpoints external to the server chassis. Server firmware was co-developed to report accurate thermal telemetry, remove fan fault conditions that would otherwise trigger shutdown, and adjust throttling thresholds appropriate to the more stable thermal environment of dielectric immersion.

Thermal Performance: Stability as the Operative Metric

Thermal characterization of immersion cooling is commonly framed around absolute temperature reduction—the difference in steady-state GPU junction temperature between air-cooled and immersion-cooled operation at equivalent power. While this metric is relevant, it is not the primary driver of performance or reliability outcomes in AI compute deployments. The operative metric is thermal stability: the magnitude of GPU junction temperature variation under real, dynamic AI workloads.

AI training workloads do not present a constant thermal load. Matrix multiplication operations during forward and backward passes drive GPU utilization transiently to near-100%, while data loading, gradient synchronization, and

checkpointing reduce utilization substantially within the same training step. In an air-cooled environment, these utilization transients translate directly to GPU junction temperature swings of 20–25°C within timescales of seconds to tens of seconds. Thermal cycling at this frequency and amplitude drives fatigue accumulation in solder joints, PCB laminates, and die-attach structures—the primary mechanisms of long-term GPU reliability degradation.

The significantly higher thermal mass and convective heat transfer coefficient of the dielectric fluid environment suppresses these transient temperature excursions. Under equivalent AI workload profiles, GPU junction temperature variation in the immersion environment is substantially attenuated—thermal response is dominated by the fluid's specific heat capacity rather than the server's internal air mass. The operational consequences are:

- Elimination of thermal throttling under sustained AI training loads—GPU clock speeds remain at operational NVIDIA GPU Boost throughout the training step.
- Suppression of thermal fatigue accumulation through reduced junction temperature cycling amplitude.
- Improved warranty confidence, since steady-state and transient junction temperatures remain well within manufacturer operational limits.
- Higher sustainable server density per rack, since thermal headroom is no longer consumed by variability margin.

> *The operative thermal metric for AI compute reliability is not steady-state temperature—it is the amplitude of junction temperature cycling under dynamic workloads. Dielectric immersion attenuates this cycling through superior thermal mass and heat transfer coefficient.*

Validation for this deployment extended beyond controlled characterization. H100 GPUs have been in continuous immersion operation since April 2024; NVQual system-level qualification was completed in September 2024. The deployment was subsequently tested for operational conditions (throttled) with ambient fluid temperatures up to 60°C under ambient fluid temperatures up to 60°C, confirming adequate thermal margin across a range of operating conditions relevant to non-traditional deployment environments. Critically, this thermal margin confirms design headroom for future GPU generations with higher TDPs—the immersion architecture is not thermally saturated at current platform power levels.

Power Architecture and Efficiency

Cooling infrastructure accounts for approximately 40% of total power consumption in a conventional air-cooled data center—power that performs no useful compute work. Immersion cooling achieves nearly-complete heat capture from server hardware, eliminating the recirculated-air thermal management chain and the HVAC infrastructure it requires. The efficiency gains are distributed across the facility, power delivery, and server levels simultaneously.

Facility Level: Single-Loop Cooling Architecture

The MDC's single-loop cooling architecture routes dielectric fluid directly from server tanks through an external heat exchanger (dry cooler or chiller depending on ambient conditions), without the intermediate air handling, chilled water distribution, and computer room air conditioning loops that characterize conventional data center cooling. This elimination of intermediate thermal transfer steps reduces auxiliary power consumption and the associated PUE overhead. Measured PUE in deployed MDC configurations is below 1.04, compared to 1.2–1.5 typical of air-cooled facilities at equivalent compute density.

Power Delivery: 48V-54V DC Distribution Efficiency

Migration from AC distribution with intermediate rectification to 48 V DC bus distribution reduces power conversion losses that accumulate through the conventional AC power chain (utility transformer → UPS → PDU → server PSU → internal DC-DC conversion). High-efficiency 48 V DC distribution eliminates at least one conversion stage and improves overall power delivery efficiency. The specific I^2R loss reduction associated with the transition from 12 V to 48 V bus architecture (at constant power delivery) is approximately 16x, substantially reducing thermal load on power delivery infrastructure.

Server Level: Power Savings

Fan and blower assemblies in an H200 SXM server configuration consume approximately 800 W at full operational speed under sustained load. In an immersion environment, thermal management is transferred entirely to the external fluid loop, and fans are removed from the server chassis. The resulting 800 W per-server power recovery is available for compute—effectively increasing the usable compute capacity of a fixed power envelope without additional infrastructure. At the 128-server MDC scale, fan elimination recovers over 100 kW of power—equivalent to the full compute power budget of more than 10 additional H200 servers.

~800 W	<1.04	>100 kW	40%
Power savings per server (vs. air cooled)	Measured facility PUE	Power recovered in 128-server MDC	Typical cooling share of DC power eliminated

Signal Integrity in Dielectric Immersion Environments

Signal integrity in high-speed digital systems is governed by the electromagnetic properties of the transmission medium and surrounding dielectric materials. Transitioning from an air-cooled to a dielectric immersion environment alters the dielectric constant of the medium surrounding PCB traces, connectors, and cable assemblies—with direct consequences for characteristic impedance, propagation velocity, and inter-symbol interference on high-speed interfaces including NVLink, PCIe Gen 5, and high-speed memory buses.

The dielectric constant (εr) of engineered single-phase immersion fluids is typically in the range of 1.8–2.1, compared to εr ≈ 1.0 for air. For a transmission line whose impedance is determined in part by the surrounding medium, this change in dielectric constant will reduce characteristic impedance and alter signal propagation velocity. Uncontrolled, these effects can cause impedance mismatches at connectors and PCB layer transitions, increase return loss, and degrade eye margin on high-speed links—potentially violating signal integrity specifications developed and validated for air-cooled operation.

The correct engineering response is to treat signal integrity as a first-class design deliverable in immersion hardware—not as a property that can be assumed to transfer from air-cooled validation. This requires:

- PCB trace geometry and impedance design that accounts for the dielectric constant of the immersion fluid at the design stage, not post-hoc.
- Connector selection validated for dielectric fluid compatibility and electrical performance when immersed.
- Cable and interconnect routing that maintains specified impedance continuity in the immersion environment.
- System-level qualification via NVQual testing, which exercises the full NVLink and PCIe interconnect fabric under operational conditions and provides objective pass/fail data independent of design-phase assumptions.

In this deployment, signal integrity was designed in from the hardware redesign stage, with PCB layout, connector selection, and interconnect geometry reviewed and adjusted for the specific dielectric properties of the fluid in use. NVQual qualification testing and sustained production benchmarking confirmed that the H200 HGX Node, in immersion meets or exceeds signal integrity specifications for H100 and H200 SXM operation. This outcome is not self-evident and should not be assumed for other hardware configurations without equivalent validation.

SIGNAL INTEGRITY: DESIGN AND VALIDATION REQUIREMENTS FOR IMMERSION

- Design PCB trace impedance for the dielectric constant of the immersion fluid (εr ≈ 1.8–2.1), not for air (εr ≈ 1.0).
- Select connectors qualified for both dielectric fluid material compatibility and electrical performance when immersed.
- Validate at the full system level via NVQual—component-level measurements do not capture system-level impedance interactions.
- Do not transfer signal integrity assumptions from air-cooled hardware validation to immersion deployments without re-characterization.
- Treat signal integrity qualification as a gate condition for production deployment, not a post-deployment verification step.

Production Validation: Performance Under Sustained AI Workloads

Laboratory and controlled characterization provide necessary but insufficient evidence for production deployment confidence. The relevant performance questions—GPU throughput stability under realistic workload profiles, thermal behavior under sustained multi-hour training runs, cluster-level interconnect performance at scale—can only be answered in a production environment with real workloads.

Technical Validation Sequence: From Failure Mode to Production Proof

Denvr's NeoCloud environment provided the production context for validating immersion-native AI infrastructure under real operating conditions. The validation sequence was structured to compare three engineering states:

- Performance characteristics of an unoptimized immersion system under a variety of controlled high-load conditions .
- Performance characteristics of immersion-native H100 and H200 systems under a variety of controlled high-load conditions.
- Operational behavior of immersion-native H100 and H200 systems under a variety production AI workloads.

The results presented below are representative examples drawn from a broader set of validation activities across each of these conditions. For the unoptimized immersion system, an R&D dataset exploring heat sink fin optimization for A100 SXM server was used. In early High Performance Linpack (HPL) testing at Denvr, this A100 configuration exposed the failure modes associated with a "worst case scenario" server design where heat sinks were least effective at facilitating thermal management. Servers experienced increasing thermal slowdown counters, heterogeneous GPU behavior, inconsistent power draw, and clock-rate reduction under sustained thermal stress. This result demonstrates that without optimized heat sink design and system-level optimization, GPUs can still exceed thermal limits and invoke firmware-level protective throttling. In this scenario, servers are still functional and can support customer workloads, but with sub-optimal throughput.

Testing of H100 and H200 servers validated the optimized immersion-native platform under a more demanding condition. The system operated as expected at ~700 W GPU TDP, in a significantly higher-density 3OU chassis, using the same high-stress HPL methodology as described above. Unlike the unoptimized configuration, the servers showed no detected thermal slowdown events, stable and consistent GPU power draw across accelerators, homogeneous GPU behavior, and no evidence of overheating or thermal instability. This confirms that server's capacity for heat dissipation exceeded system load and that the platform sustained full GPU performance without inducing protective throttling.

Production workload validation then extended the test beyond synthetic benchmarks. Continuously batched Llama-3.1-Instruct 70B inferencing was used to evaluate system behavior under sustained, production-relevant AI demand. This workload introduced dynamic utilization patterns, memory-compute interaction, and continuous request handling that are not fully represented by standardized stress tests alone. Under these conditions, the system maintained stable GPU power profiles, consistent performance across accelerators, and no evidence of thermal throttling or progressive degradation over time.

The entirety of the test program results establish the full validation case. The A100 R&D baseline shows the worst case scenario: thermal throttling, inconsistent GPU behavior, and reduced performance under load. The

immersion native H200 HPL result shows the engineered correction: stable operation under higher thermal and power density. The Llama-3.1-70B inferencing result confirms production behavior: predictable performance under real AI workload conditions.

Two deployment contexts provided distinct validation scenarios:

- **AI NeoCloud training clusters:** Sustained large-model training workloads with dynamic GPU utilization profiles, validating thermal stability and interconnect performance under realistic AI compute demand.
- **Sovereign and regulated deployments:** The modular MDC architecture enables deployment in environments with data sovereignty requirements, security boundary controls, and operational constraints—including air-gapped configurations and compliance with Canadian security standards—where conventional hyperscale colocation infrastructure is not available or permitted.

A relevant finding from production benchmarking: the thermal margin observed in this deployment is not exhausted by current H100 and H200 TDP levels. The immersion architecture retains meaningful headroom for next-generation GPU platforms with higher TDPs, confirming that the engineering approach is not platform-specific but applicable to the trajectory of AI compute hardware development.

Engineering Observations from Production Deployment

The following observations reflect direct operational and engineering experience from this deployment and are intended to inform the methodology of future immersion deployments:

KEY ENGINEERING OBSERVATIONS

1. Immersion-native hardware design is a prerequisite, not an optimization. Adapting air-cooled hardware to an immersion environment without systematic redesign of thermal geometry, power architecture, and firmware introduces failure modes that are not evident in short-duration characterization but manifest under sustained production operation.

2. Thermal stability—not absolute temperature—is the correct reliability metric. For AI GPU workloads, the amplitude of junction temperature cycling under dynamic load governs fatigue accumulation and long-term reliability. Immersion's primary reliability benefit is suppression of thermal cycling and dramatic changes in temperature, not necessarily a reduction of peak temperature.

3. Signal integrity requires dedicated design and validation for the immersion dielectric environment. The dielectric constant of immersion fluids alters transmission line characteristics relative to air-cooled designs. NVQual system-level qualification is required; component-level measurements are not sufficient.

4. Power architecture co-design is required. The 50V-54V DC bus selection in this deployment was driven by current-carrying and loss considerations specific to the server density achieved—a choice that could not have been made independently of the tank population and thermal design.

5. Fluid chemistry and material compatibility require deliberate validation. Long-term dielectric fluid stability, fluid-polymer compatibility, fluid-elastomer compatibility, and fluid-PCB laminate interactions may be verified in short-duration testing and confirmed in long-duration installation. A validated fluid partner ecosystem and established material compatibility protocol are necessary program inputs.

6. Multi-disciplinary validation is not optional. Component, server, system, and production-level validation address qualitatively different risk categories. Compressing this validation hierarchy accelerates deployment timelines at the cost of discovering failure modes after production commitment.

OCP Technical Specifications: Foundation for Standardized Design

Several Open Compute Project technical workstreams provided specifications that directly informed the hardware design decisions in this program. The OCP framework was not the organizing basis for the collaboration between Denvr and UNICOM Engineering, but it provided a common technical language and a set of validated baseline specifications that grounded design choices and reduced the solution space that required original engineering work.

The OCP contributions specifically leveraged in this program were:

- **Open Rack power distribution specifications:** The ORV 2.2 and ORV3 open rack specifications provided the reference architecture for the 50V-54V DC bus design. As noted in the power architecture section, the implementation departs from ORV 2.2's specified 12 V bus bar voltage in favor of 48 V to meet the current-carrying requirements of high-density GPU deployment—a gap that the OCP community is addressing through ongoing ORV 3 specification development.
- **Immersion Cooling Environments specifications:** OCP immersion requirements documentation provided baseline guidance on fluid properties, material compatibility testing methodology, and thermal design parameters. These specifications informed the fluid selection criteria and material compatibility validation approach used in this program.
- **Signal integrity guidance:** OCP technical documentation on signal integrity considerations for immersion environments informed the design review process for PCB layout and connector selection.
- **Safe handling, cleaning, and operational procedures:** OCP operational guidance for immersion ITE provided a starting framework for fluid handling, component swap procedures, and maintenance protocols.

- **FMEA documentation:** OCP failure mode and effects analysis for immersion systems provided a structured reference for identifying and mitigating operational risk categories.

The gaps identified during this program—particularly in power architecture specifications for GPU-density deployments and in signal integrity validation methodology for high-speed interconnects in immersion—represent areas where practitioner experience from production deployments can most directly improve OCP technical documentation.

Conclusion

Single-phase dielectric immersion cooling, implemented with purpose-designed hardware and validated across the full system stack, resolves the thermal and power density constraints that limit conventional air-cooled AI infrastructure. This work documents a production deployment that demonstrates the engineering methodology and operational outcomes achievable when facility design, power architecture, hardware engineering, and validation methodology are co-developed rather than independently optimized.

The primary technical contributions of this program are: (1) a validated immersion-native server platform (H200 HGX Node) covering H100 and H200 SXM GPU configurations, with documented departures from air-cooled design in form factor, power architecture, thermal hardware, and firmware; (2) a multi-level validation methodology spanning single- and multi-component NVQual qualification through sustained production AI workload operation; and (3) engineering observations from production deployment that address the failure modes and design considerations not captured by laboratory characterization alone.

The design and validation methodology documented here is intended to be reproducible. The specific hardware configuration will evolve with GPU generations and power architecture standards; the multi-disciplinary validation approach, the signal integrity design requirements for dielectric environments, and the thermal performance metrics that matter for AI workload reliability are stable engineering principles that apply across platform generations.

> *The hardware configuration is generation-specific. The engineering methodology is not.*

About the Authors and Organizations

Amy Short, PhD is Executive Officer, Advanced Data Center Technologies at Denvr, with expertise spanning dielectric fluid system design, thermal architecture, and the operational engineering of immersion-native AI infrastructure.

Austin Hipes is Chief Technologist and VP of Engineering at UNICOM Engineering, with deep expertise in server architecture, signal integrity, power delivery systems, and immersion-ready hardware platform design and validation.

Denvr operates NeoCloud infrastructure for next-generation AI workloads, with a vertically integrated model spanning facility, power, cooling, and IT deployment. The organization draws on 15+ GW of global hyperscale infrastructure experience and over $1B in AI compute architectures deployed, including Canada's first commercial AI Cloud.

UNICOM Engineering is a leading provider of purpose-built application platforms and life-cycle deployment services, with an immersion-ready platform portfolio spanning Dell Technologies, HPE hardware families, and a validated fluid partner ecosystem including Lubrizol, Shell, ExxonMobil, HF Sinclair, and SK enmove.

Open Compute Project technical workstreams: https://www.opencompute.org/projects