BRIEFING B1 UNCLASSIFIED

Direct-to-Chip Cooling

How AI hardware is cooled today, where cooling is heading, and where the CF-JET cold plate and SOL's AI framework fit

John Kolb
SOL Engineering Services

PUBLISHED
13 November 2026

READING TIME
15 min

Overview

CF-JET is an additively manufactured, liquid-cooled jet-impingement cold plate with an optimized carbon fiber target plate, designed with AI and machine learning. The cold plate is the product, but the work that produces it spans a much wider landscape: artificial intelligence, advanced manufacturing, materials development, mechanical engineering, modeling and simulation, thermal sciences, electronics cooling and experimental testing, all carried out with a re-usable AI framework.

> Remember that the AI revolution is about hardware development more than the software development that it enables.

This briefing shows how liquid cooling for AI hardware works today, where it is heading, and where CF-JET and SOL's AI framework fit. Each figure is a simplified schematic drawn for this briefing; the text beside it explains what it shows and why it matters to the project.

1. What the Proposal Really Delivers

The CF-JET proposal is an AI-enabled, modular open systems approach (MOSA) to designing and developing a product that meets the requirements of three ARL broad agency announcement (BAA) topics at once. Its purpose is to create seminal innovation through AI-enabled generative design that advances cold plate design, or any of the 60 topics identified in the separate Topic Development research paper that supports the proposal.

The approach answers ARL-BAA-0038, *Artificial Intelligence and Machine Learning with Extremely Sparse Data*, through the neuro-symbolic method SOL has developed: the SOL Aquifer™ protocol and the Forest-to-Desert Framework™. The lasting contribution of the proposal is this AI approach and the application being built with it, which reach well beyond the cold plate used to develop and prove them.

1.1 A “Code-to-Architecture” framework that is reused, not rebuilt

Software development today is more than writing code that stands on its own. SOL Aquifer takes a “Code-to-Architecture” approach: every effort has a product target, but the application-specific content sits on the standard layers of the Forest-to-Desert Framework, which stay the same whatever solution is sought. The framework is therefore re-usable rather than a one-off, and that is what lets AI work with the sparse data that real hardware programs produce. The MOSA approach mirrors the research “sandboxes” that AI laboratories such as Anthropic and OpenAI use to make continuous, incremental improvements.

1.2 Ownership of the algorithmic baseline

For modernization priorities that include advanced computing, advanced infrastructure and manufacturing, space technology, hypersonics, sustainment and logistics, and trusted AI and autonomy, the Department of the Air Force (DAF) states:

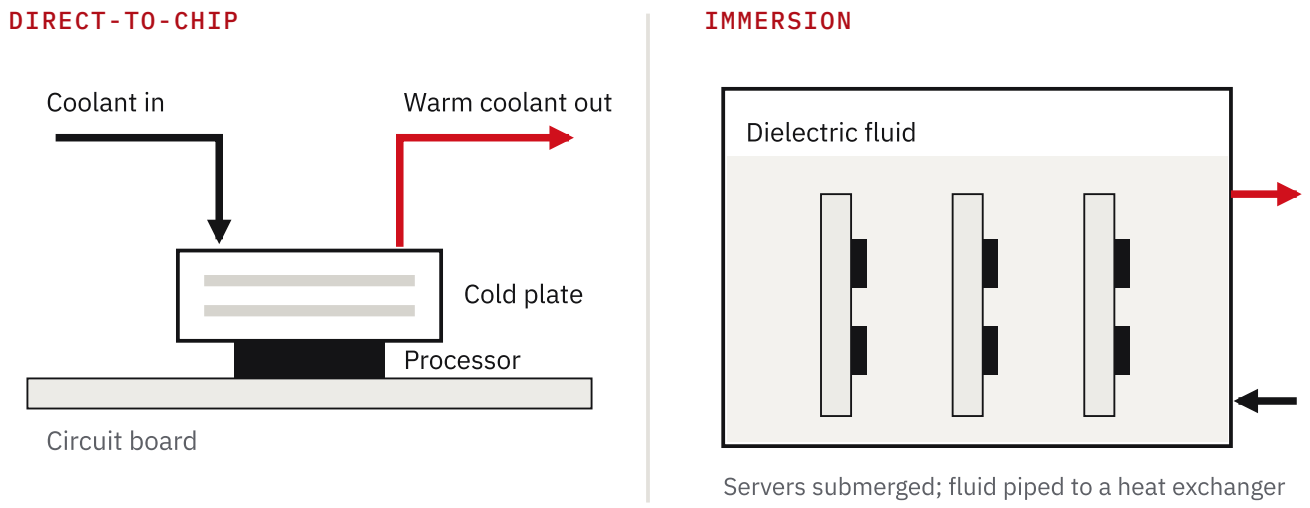
> “Offerors must explicitly demonstrate that they possess the intellectual property rights or native ownership of the algorithmic baselines being brought forward. This foundational proof cannot be purely conceptual or marketing-driven: it must contain raw technical and analytical validation.”

The requirement makes sense. Government regulation of AI is coming, and owning a validated baseline puts SOL on the defining end of the standards that follow. An open-source stack removes structural cost and the rigidity of a proprietary platform, but it leaves the customer without control; a single, agile foundation such as SOL Aquifer gives that control back. The requirement also matches how the Industry 4.0 market sells distributed software solutions and royalty-based usage contracts.

2. Two Ways to Cool AI Hardware with Liquid

Air can no longer carry away the heat that modern processors produce, so high-density computing is moving to liquid. Two architectures dominate (Figure 1). **Direct-to-chip** cooling mounts a cold plate on each hot component and pumps coolant through it. **Immersion** cooling submerges whole servers in a non-conducting (dielectric) fluid.

Figure 1. Direct-to-chip and immersion cooling



Schematic, not to scale. Dark arrows show cool coolant; red arrows show coolant warmed by the chip.

Each approach has its own strengths, summarized in Table 1. Direct-to-chip cooling fits into existing racks and servers and is the more mature of the two. Immersion removes heat from every component at once and suits the densest systems.

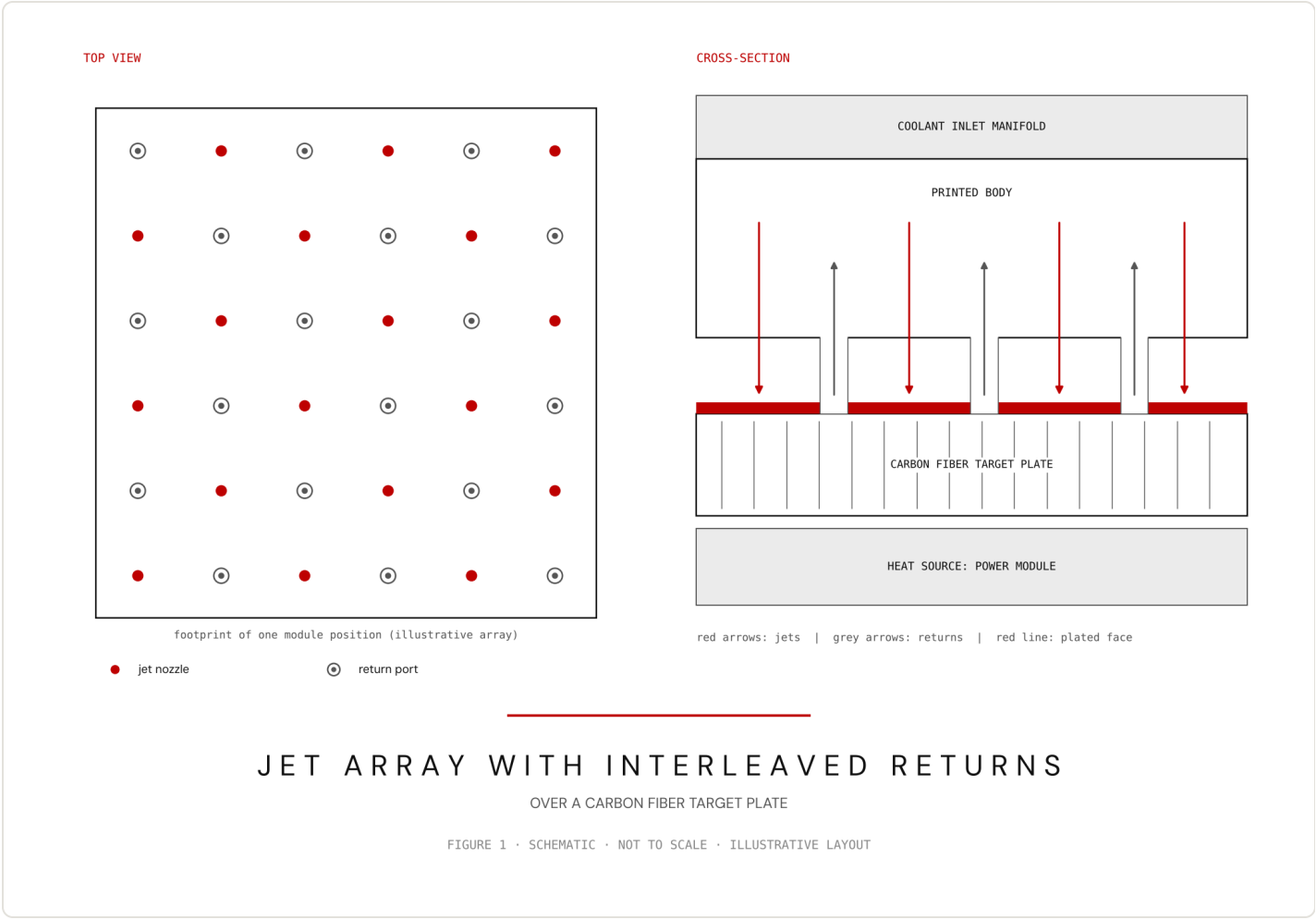
Table 1. Strengths of the two approaches

Direct-to-chip cooling	Immersion cooling
Precise coolant delivery to the hottest components	Removes heat from every component in the server
High thermal performance per processor	Energy efficiency and sustainability
Mature technology and easy maintenance	Quiet operation with no server fans
Flexible, scalable deployment in standard racks	Suited to ultra-high-density systems

Summarized from industry descriptions of the two architectures.

A direct-to-chip cold plate can be single-phase, where the liquid only warms up, or two-phase, where the cold plate acts as an evaporator and the coolant boils as it absorbs heat. Within the cold plate, coolant can either flow straight through channels or be fired at the hot surface as jets. CF-JET is a jet-impingement cold plate (Figure 2), and additive manufacturing is what allows its complex internal pathways: jets and return ports interleaved over the full footprint of the module.

Figure 2. Where CF-JET fits: a jet array with interleaved returns over a carbon fiber target plate

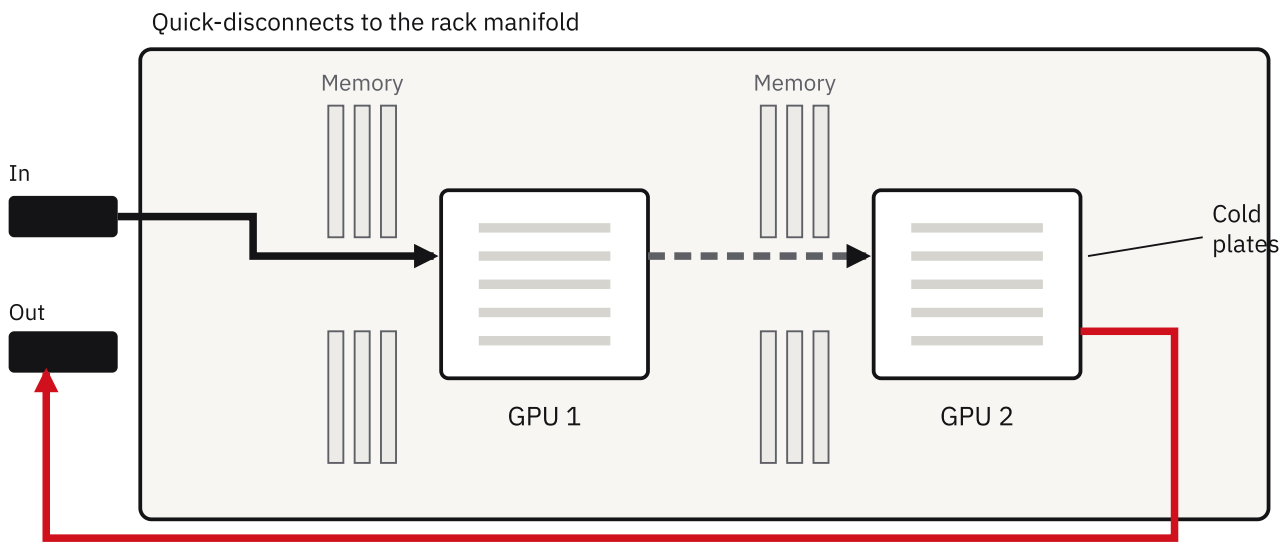


Schematic, not to scale; illustrative layout. Also shown as Figure 1 of the CF-JET research briefing.

3. Inside the Server

In a direct-to-chip server, each processor (GPU or CPU) carries its own cold plate (Figure 3). Flexible tubing links the cold plates to quick-disconnect fittings at the back of the server, so a server can be removed from the rack without draining the system. Coolant usually passes through the cold plates in a planned order so that each chip sees coolant cool enough for its heat load. Memory modules and voltage regulators either share the airflow that remains or get cold plates of their own.

Figure 3. A liquid-cooled server board, seen from above



Schematic, not to scale. Coolant enters cool (dark), passes the first cold plate (grey, partly warmed) and returns warm (red).

4. From Chip to Atmosphere

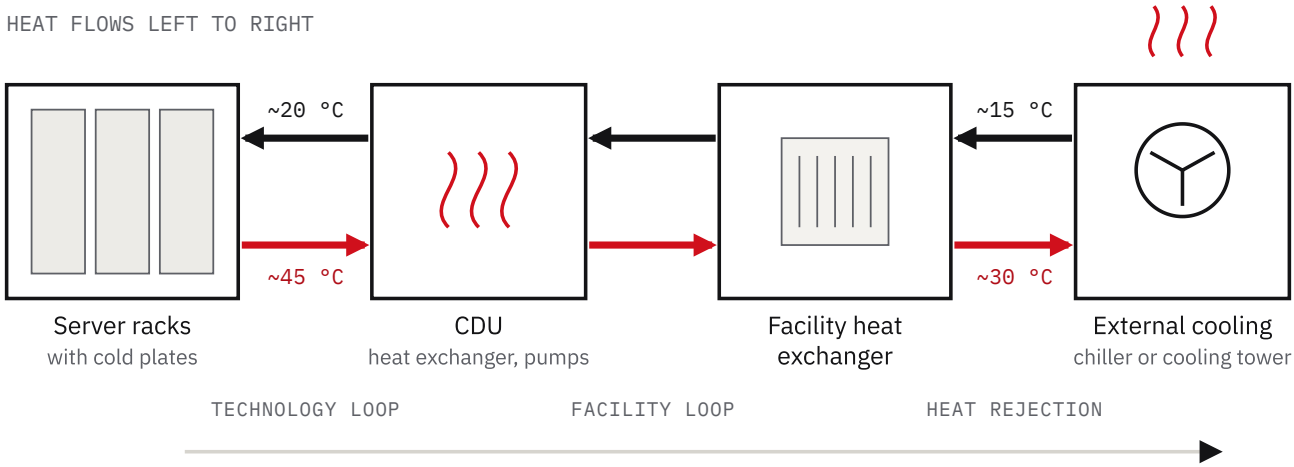
A cold plate is only the first link in a chain that carries heat from the chip to the outside air. The chain has five stages:

01	**Heat generation.** The AI accelerator (GPU or CPU) produces intense heat on its silicon die during computing workloads.
02	**Absorption.** A cold plate mounted directly on the chip absorbs that heat as coolant flows through its internal channels or jets.
03	**Collection.** Flexible hoses and quick disconnects carry the warm fluid to a vertical rack manifold that combines the flow from every server.
04	**Exchange.** The coolant distribution unit (CDU) uses a heat exchanger and internal pumps to pass the heat to the building's water loop.

05 **Rejection.** Dry coolers, chillers or cooling towers, often on the roof, receive the warm water and release the energy to the atmosphere.

Figure 4 follows the same path through the three separate liquid loops of a data center, with typical temperatures. The loop through the servers is kept small, clean and closed; it never mixes with the building water. Many new designs use a sealed, automotive-style coolant loop that consumes no water at all.

Figure 4. Three loops carry heat from the servers to the outside air



Schematic; temperatures are typical values shown for illustration and vary by facility. Dark lines carry cool liquid toward the servers; red lines carry warm liquid away.

4.1 Where SOL already works in this chain

SOL has experience in the design and build of coolant distribution units, and its architecture, engineering and civil past performance covers the rooftop chillers and HVAC equipment at the end of the chain. A CDU can stand alone or be integrated into the rack. Either way, its control can be driven by the SOL Aquifer digital flywheel, which learns from the operating data the unit already collects. SOL has validated the Aquifer neural pipeline on HVAC work, the same kind of equipment that rejects heat at the end of the chain.

5. At Facility Scale

Hyperscale AI data centers repeat this chain thousands of times. Rows of liquid-cooled racks share overhead or under-floor manifolds, banks of CDUs serve each row, and large heat-rejection plants sit

outside. Published imagery of major operators' AI facilities and of new "AI factory" designs shows the same pattern: liquid cooling moving from a specialty option to the default for high-density compute.

~100,000	~150 MW
GPUS IN ONE LARGE AI CLUSTER IN MEMPHIS	REPORTED POWER DRAW OF THAT CLUSTER

Power is becoming the limit. One large AI training cluster in Memphis reportedly runs about 100,000 GPUs at the same time and draws about 150 MW, more than the local grid was built to supply. Concentrating that much load in one place may not be the future; as political and environmental pressure prompts a fresh look at how data centers operate, more efficient cooling and smaller, distributed and on-premises computing become more attractive. SOL intends to be part of that rethink.

6. Not Just Data Centers

Every use of an advanced cold plate such as CF-JET, SOL's developing agile design, follows the same physics: concentrated heat, limited space and a need for long life. The technology developed in the proposal is much more than the product; it is the AI-enabled generative design approach, embodied in SOL Aquifer and the Forest-to-Desert Framework, that can be applied to each of these markets. Tables 2 and 3 list the uses and what CF-JET offers each one.

Table 2. Military uses

Military use	What it cools	Why CF-JET helps
Ground vehicle power electronics	SiC inverters, converters and motor drives in hybrid-electric combat and tactical vehicles	Lower mass; chip-matched expansion for longer module life; no corrosion in glycol loops
Onboard power and silent watch	Generators, power distribution units and battery packs	Lighter cooling hardware frees weight for armor, payload and range
Directed energy	Laser and high-power microwave power supplies and laser diode arrays	High heat flux from jet impingement; low-expansion target plate suits diode and SiC devices
Radar and electronic warfare	GaN transmit and receive modules and high-power amplifiers	Expansion close to GaN-on-SiC; light weight for masts, vehicles and aircraft
Aviation and unmanned systems	Avionics and hybrid-electric propulsion power electronics on rotorcraft and UAS	Weight is critical; one-piece printed part with no brazed joints
Expeditionary power and microgrids	Tactical microgrid inverters and energy storage converters	Corrosion resistance and thermal-cycling life in field conditions

Table 3. Commercial uses

Commercial use	What it cools	Why CF-JET helps
Electric vehicles	Traction inverters, onboard chargers and battery modules	Mass and module life drive range and warranty cost
Data centers and AI computing	High-power processors and accelerators on liquid-cooled boards	Very high heat flux; even temperature across the chip
Renewable energy and grid storage	Solar and wind inverters and battery storage converters	Long life under daily thermal cycling; no corrosion
Rail, marine and aerospace electrification	Traction converters and electric aircraft power electronics	Weight savings and reliability in harsh environments
Industrial power and lasers	Industrial laser diodes, induction heating and welding power supplies	High heat flux with chip-matched expansion

6.1 Three clusters of demand

The uses group into three clusters, each with a different reason to need a better cold plate.

CLUSTER I

High compute and directed energy

AI and data centers, directed energy weapons, and radar and electronic warfare. Very high heat flux on small devices; even temperatures prevent throttling, and a low-expansion target plate suits laser diodes, SiC and GaN-on-SiC.

CLUSTER II

High mobility and electrification

Ground vehicle power electronics, aviation and unmanned systems, rail and marine, and commercial EVs. Every kilogram counts, so lower mass, chip-matched expansion and one-piece parts with no brazed joints extend range and module life.

CLUSTER III

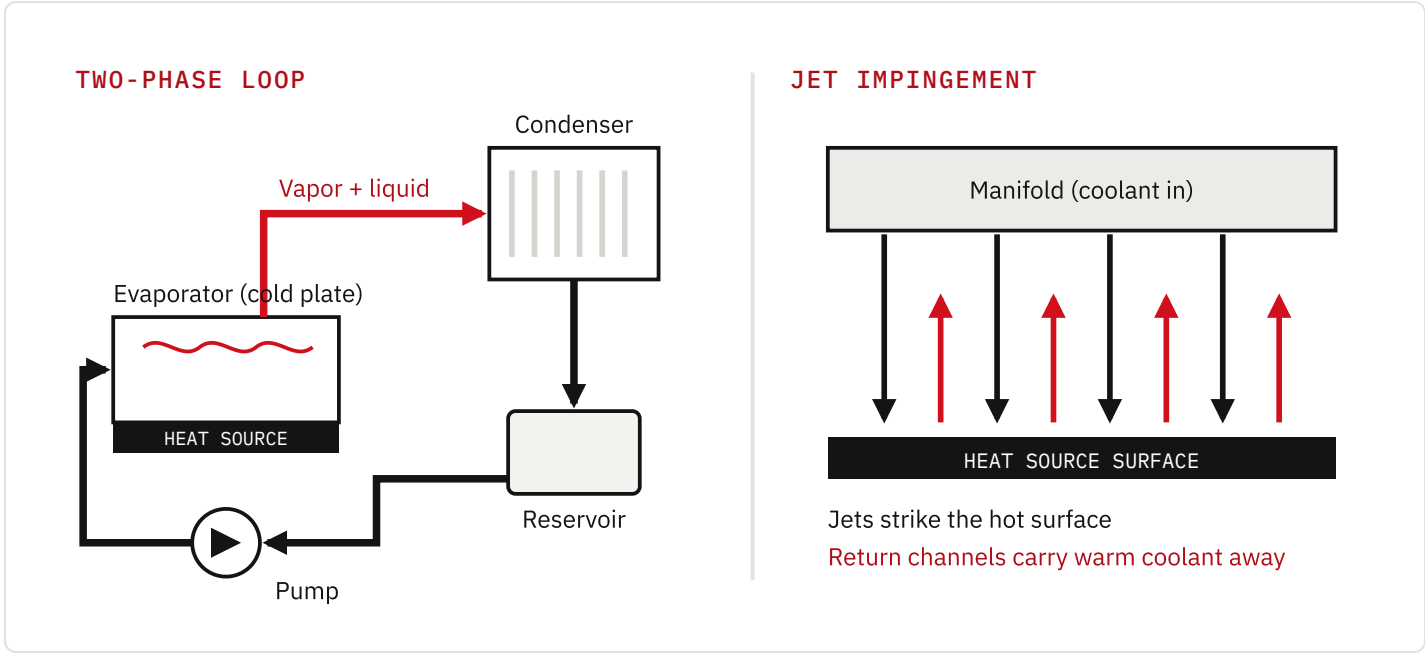
Extreme environments and rugged power

Expeditionary power and microgrids, renewable energy and grid storage, and industrial power and lasers. Daily thermal cycling and harsh conditions demand corrosion resistance and decades of fatigue life.

6.2 Two cold plate families for these uses

Two families of cold plate serve these markets (Figure 5). In a **two-phase** loop, the cold plate is an evaporator: the coolant boils as it absorbs heat, a condenser turns the vapor back to liquid, and a pump returns it through a reservoir. Boiling absorbs a great deal of heat at a nearly constant temperature, which suits high-flux devices such as ground vehicle SiC inverters and directed energy lasers. In a **jet-impingement** plate, a manifold fires coolant jets directly at the hot surface and return channels carry it away, giving very high local heat transfer with lower thermal resistance; radar transmit-receive modules and UAS avionics are typical uses. Commercial designs also use micro-channel cold plates, micro-channel matrices and vapor (evaporation) chambers.

Figure 5. Two-phase and jet-impingement cold plates



Schematic, not to scale. Dark arrows show cool liquid; red arrows show warmed liquid or vapor.

7. Why AI Chips Are Pushing Cold Plates Further

NVIDIA's Hopper generation of AI processors, introduced in 2022 with the H100, set the pattern: an 80 GB stack of HBM3 memory beside the processor and a heavily loaded voltage regulator module, cooled by air or liquid. The industry is now moving to the Blackwell family, where heat output is higher still and liquid cooling is standard for high-end systems. Liquid-cooled Blackwell servers route coolant tubing to every GPU. Significant improvement in cooling technology is needed, and it is already in motion.

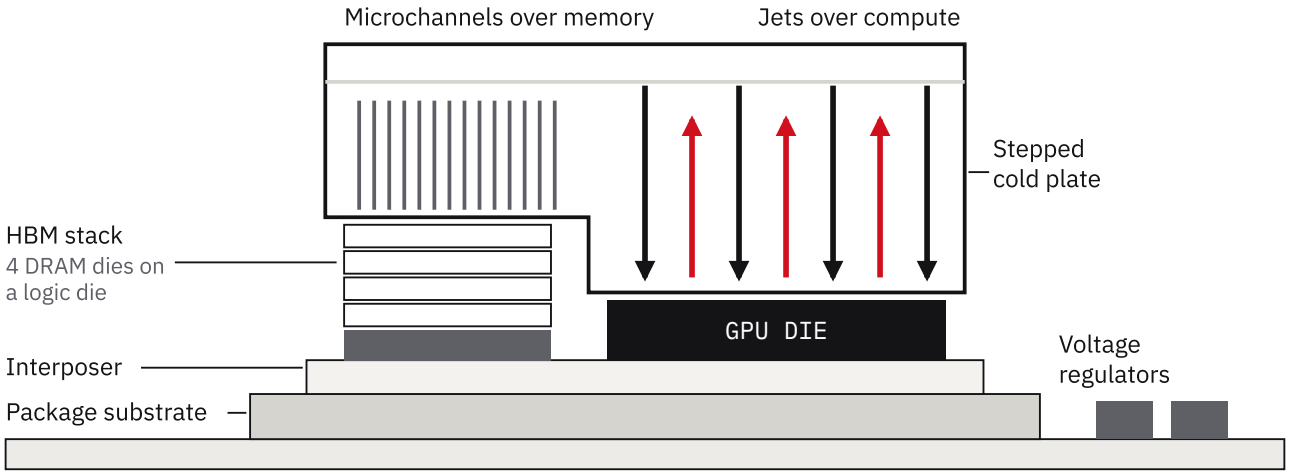
7.1 Voltage regulators carry hundreds of amps

A voltage regulator module (VRM) takes the incoming power and steps it down to the low voltage a GPU needs. It therefore carries hundreds to more than 1,000 amps and gets hot. The draw changes within microseconds, so the VRM must respond instantly. If fans cannot cool the VRM, a cold plate does, either one plate that extends across everything with separately controlled zones, which is the SOL approach, or a separate plate. VRMs and their MOSFET switches are familiar territory for electrical engineering and engineering technology students, which makes them a good teaching example.

7.2 High Bandwidth Memory is the next cooling challenge

High Bandwidth Memory (HBM) is what feeds data to an AI GPU. Training and inference move enormous amounts of data; without enough bandwidth the GPU sits idle, waiting. HBM stacks several DRAM memory chips vertically on a logic die, and the stack sits on the same silicon base, the interposer, right next to the GPU die (Figure 6). Speed comes from the width of the connection and the short distance, at 2 TB/s and above, and so does a great deal of heat. The HBM stack must stay cooler than the GPU die, or it slows down and loses reliability.

Figure 6. An AI processor package and a stepped, hybrid cold plate



Schematic, not to scale. Dark arrows show cool coolant entering; red arrows show warmed coolant returning.

Two cold plate topologies address these packages:

- **Parallel microchannels.** Ultra-thin, closely spaced fins, often 50 to 200 µm wide, maximize surface area for convective heat transfer. Because the fluid travels long paths through narrow channels, this topology cools strongly but has a high pressure drop and needs strict filtration to prevent clogging.
- **Microconvective jet impingement.** Coolant is forced through an array of microscopic nozzles perpendicular to the target plate, breaking up the thermal boundary layer directly over hot spots. Nozzle density can be tuned locally to send colder, faster fluid straight onto the compute cores while dedicating separate jet zones to the HBM stacks.

As training and inference loads rise, the answer is likely a hybrid: impingement on the compute side and microchannels over the DRAM dies, on a stepped cold plate that follows the height of each component (Figure 6). Each zone would be developed and optimized with SOL's AI protocol, which optimizes a design for its performance target and, in parallel, drives the advanced, additive and mechatronic manufacturing processes that make it possible to build. SOL's code is intended to become part of the control sequence that keeps the HBM stack cooler than the GPU die, and that kind of stepped cold plate is best made through AI-enabled additive manufacturing.

7.3 From Hopper to Blackwell

Table 4 compares the two generations. Power per GPU rises from about 700 W to 1,000–1,200 W, memory bandwidth more than doubles, and liquid cooling becomes standard for the rack-scale systems.

Table 4. Hopper and Blackwell compared

	Hopper (H100 / H200)	Blackwell (B200 / GB200)
Chip design	One large die	Two dies joined by a very fast link (~10 TB/s) that act as one GPU
Transistors	~80 billion	~208 billion
Tensor Cores	4th generation	5th generation
Lowest precision for AI	FP8 (Transformer Engine)	FP4 and FP6 (2nd-generation Transformer Engine), roughly doubling throughput and fitting larger models
Memory	H100: 80 GB HBM3, ~3.35 TB/s; H200: 141 GB HBM3E, ~4.8 TB/s	B200: 192 GB HBM3E, ~8 TB/s; Blackwell Ultra (B300): up to 288 GB
GPU-to-GPU link (NVLink)	4th generation, ~900 GB/s per GPU	5th generation, ~1.8 TB/s per GPU
System scale	Typically 8 GPUs per server (HGX)	Up to 72 GPUs in one rack acting as one large GPU (GB200 NVL72), with Grace CPUs paired directly to GPUs
Power per GPU	Up to ~700 W	~1,000–1,200 W
Cooling	Air or liquid	Liquid cooling standard for high-end and rack-scale systems
Other additions	Confidential computing introduced	Dedicated reliability engine (RAS), data decompression engine, improved confidential computing

Values as published for each generation; approximate.

8. SOL Memphis AI Center

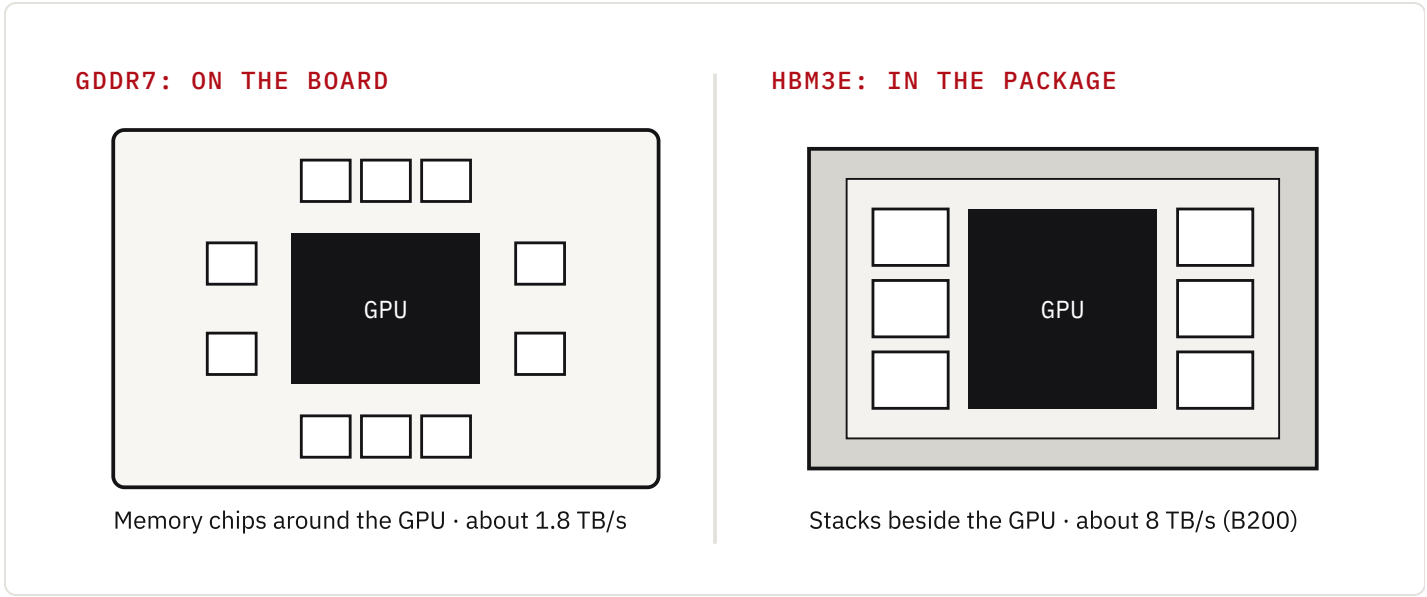
SOL will use its on-premises AI server for this work: an NVIDIA RTX PRO 6000 Blackwell system with 96 GB of GDDR7 memory with error correction (ECC), housed at SOL's Memphis center. On-premises AI and simulation can train SOL Aquifer surrogates, run neural operators and ensembles, and run physics post-processing at the engineer's desk. On-premises compute, storage and networks that meet CMMC requirements are likely to become a larger part of the overall data center picture.

- **Air-gapping.** Levels of air-gapping are a likely place for future government controls. SOL is anticipating this trend in the way it brings the project forward.
- **Controlled access for university researchers.** Giving university partners access to SOL's own AI center keeps the work inside a controlled environment, which simplifies U.S.-person and compliance questions.
- **Separating controlled work.** Ordinary commercial cloud and public AI tools can still be used, since most of the work is unrestricted, but the setup lets controlled unclassified information (CUI) be kept separate, as future contract and sponsor work will require.
- **Working with partners.** The center makes it easier to work interactively with the project's research partners. The alternatives, an AWS GovCloud account or access through the DoD High Performance Computing Modernization Program (HPCMP), are obtainable for large sweeps but are not needed for this work.

8.1 The server

The RTX PRO 6000 uses the same Blackwell GPU generation as the B200 data center part, with GDDR7 memory instead of an HBM stack. It is the full-power Blackwell workstation and server card: a PCIe accelerator that slides into a standard slot, unlike the B200's data center modules, and it can be air-cooled. The machine can be upgraded in future.

Figure 7. Board-mounted GDDR memory and in-package HBM



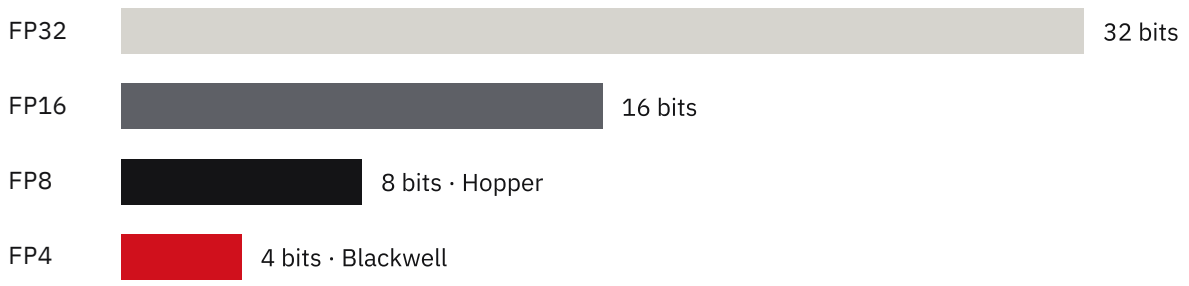
Schematic, not to scale; chip counts are illustrative.

GDDR and HBM. GDDR7 chips sit around the GPU on the circuit board rather than stacked in the same package (Figure 7). GDDR is cheaper and easier to cool, but it has less bandwidth than HBM: about 1.8 TB/s

on this card against about 8 TB/s on a B200. For training surrogates and running simulations at the desk, that trade is a good one.

FP4 precision. AI models store and process numbers at a chosen precision. Each halving of the number of bits roughly doubles speed and halves memory, so larger models fit and run faster (Figure 8). Neural networks tolerate the rounding well, and scaling techniques in the Transformer Engine keep FP8 and FP4 accuracy acceptable. That is why Blackwell's FP4 support is a large gain for running large language models.

Figure 8. Fewer bits per number: less memory and more speed



Bar length is proportional to the bits used for each number.

Error-correcting code (ECC). ECC memory corrects single-bit errors on the fly, so the program never notices, and detects larger multi-bit errors and reports them instead of silently passing on bad data. Without ECC, a flipped bit might crash a program or, worse, quietly corrupt a result: a wrong value in a CFD run or a corrupted weight in an AI model, with no warning.

9. Key Takeaways

- Liquid cooling is becoming the default for high-density computing, and the cold plate is where the heat first leaves the chip.
- CF-JET's jet-impingement design with interleaved returns, made possible by additive manufacturing, targets the high heat flux and even temperatures that AI processors, SiC power electronics and directed energy systems need.
- The same need appears across military and commercial markets: ground vehicles, radar, directed energy, EVs, renewable energy, rail, marine and industrial lasers.

- Rising GPU power and HBM memory point toward hybrid, stepped cold plates whose zones are designed by AI and built by additive manufacturing.
- The lasting contribution is the re-usable AI framework, SOL Aquifer and the Forest-to-Desert Framework, run on SOL's own controlled AI center.