CheckMark

White Paper

Direct-to-Chip Cooling in Data Centers

Executive Summary

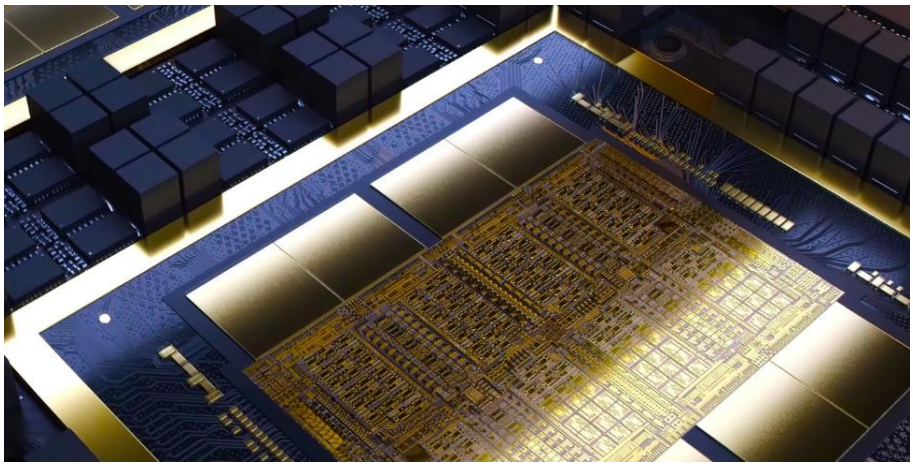
As Data Centers face escalating demands from High-Performance Computing (HPC), Artificial Intelligence (AI), and dense server configurations, traditional air-based cooling systems are increasingly inadequate. Direct-to-Chip (D2C) Cooling, also known as Direct Liquid Cooling or Cold Plate Cooling, emerges as a highly efficient alternative by delivering coolant directly to heat-generating components like CPUs (Central Processing Units) and GPUs (Graphics Processing Units). This method enhances thermal management, reduces energy consumption, and supports sustainability goals. Drawing from industry insights, D2C cooling can lower power usage effectiveness (PUE) metrics, enable higher rack densities, and mitigate overheating risks.

However, it involves challenges such as higher initial costs, potential leakage risks, and the need for customized implementations. This White Paper explores the technology, benefits, challenges, and future implications of D2C cooling, providing Data Center operators with a comprehensive guide to adoption.

Ted J. Pappas, Director of Operations

Introduction

The rapid evolution of Data Center infrastructure is driven by the exponential growth in data processing needs. Modern chips, projected to incorporate up to **one trillion transistors** by 2030, generate unprecedented heat levels that air cooling struggles to dissipate efficiently.



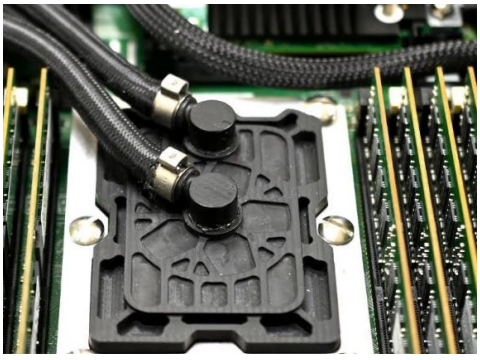
Liquid cooling technologies, particularly D2C, offer a solution by leveraging fluids with superior heat absorption capacities—up to 3,000 times that of air.

Industry projections indicate that nearly 50% of Data Centers will incorporate liquid cooling within five years, up from 22% today, as AI and HPC workloads dominate.

This White Paper examines D2C cooling's role in revolutionizing Data Center efficiency, sustainability, and performance.

What is Direct-to-Chip Cooling?

Direct-to-chip cooling is an advanced thermal management technique where coolant is delivered directly to the hottest components of servers, such as processors, via cold plates. Unlike traditional air cooling, which relies on fans and ambient air circulation, D2C uses a liquid medium, often water or dielectric fluids, to absorb and transfer heat away from the chip.



This method is a subset of in-rack liquid cooling and is particularly suited for high-density environments where power densities exceed 100 kW per rack.

Key characteristics include:

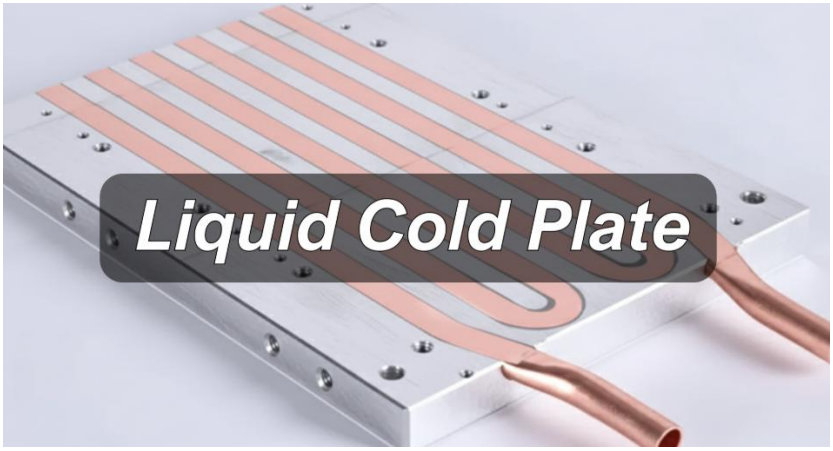
✓ Targeted Heat Removal: Focuses on major heat sources like CPUs and GPUs, leaving other components (e.g., hard drives) potentially reliant on supplementary air cooling.

- ✓ Closed-Loop Systems: Coolant circulates in a sealed loop, minimizing evaporation and contamination risks.
- ✓ Compatibility: Can be retrofitted into existing data centers, though it requires server modifications.

How It Works

D2C Cooling operates through a series of integrated components:

1. **Cold Plates**: Conductive plates (typically copper) are mounted directly onto heat-generating chips. Coolant flows through microchannels within the plate, absorbing heat via conduction.



2. **Coolant Circulation**: A pump propels the fluid (e.g., water-glycol mixtures or dielectric fluids) through the system. Heated coolant is directed to a heat exchanger.
3. **Heat Rejection**: Coolant Distribution Units (CDUs) transfer heat from the D2C loop to a facility's cooling water loop or external rejection systems like dry coolers.
4. **Thermal Interface**: A material between the chip and cold plate ensures efficient heat transfer.

The process maintains chip temperatures below **Throttling Thresholds** (the maximum temperature at which a processor (CPU or GPU) can operate before it automatically reduces its performance to prevent overheating and potential damage) allowing sustained high performance. For instance, in a closed-loop setup, coolant enters the cold plate at around 20-30°C, absorbs heat, and exits at 40-60°C before cooling.

Benefits

D2C Cooling provides significant advantages over air-based systems:

- ✓ **Energy Efficiency**: Reduces energy use by up to 25% by eliminating the need for energy-intensive fans and chillers, leading to lower PUE (often below 1.2).
- ✓ **Higher Density and Performance**: Supports rack densities beyond 100 kW, enabling more compute power per square foot and preventing thermal throttling in HPC/AI workloads.
- ✓ **Sustainability**: Reduced electricity consumption and potential heat reuse (e.g., for district heating).
- ✓ **Reliability and Longevity**: Maintains lower operating temperatures, reducing equipment failure rates and extending hardware lifespan.
- ✓ **Cost Savings Over Time**: While initial costs are higher, operational savings from energy efficiency can yield a positive ROI within 2-3 years.

Challenges and Considerations

Despite its advantages, D2C cooling presents hurdles:

- ✓ **High Implementation Costs**: Custom designs and retrofits can be expensive, often requiring server modifications and specialized vendors.
- ✓ **Leakage Risks**: Although dielectric fluids are non-conductive, leaks can disrupt operations; robust detection systems are essential.
- ✓ **Maintenance Complexity**: Servicing involves handling fluids, and not all components (e.g., memory) are directly cooled, necessitating hybrid systems.
- ✓ **Environmental Concerns**: Some coolants may pose pollution risks if not managed properly.
- ✓ **Lack of Standardization**: Material incompatibilities and varying vendor approaches can complicate adoption.

Mitigation strategies include rigorous testing, staff training, and adherence to emerging standards from organizations like IEEE SA (Institute of Electrical and Electronics Engineers (IEEE) Standards Association)

Key Role of IEEE SA in Data Centers

IEEE SA facilitates collaboration among industry experts, scholars, and policymakers to develop technical guidelines that mitigate issues such as excessive energy consumption, cooling inefficiencies, and equipment reliability. For instance, standards from IEEE SA ensure that Data Center components, from servers and networks to cooling systems, operate seamlessly and reduce environmental impact.

Implementation Strategies

Successful D2C deployment requires careful planning:

1. **Assessment**: Evaluate current infrastructure for compatibility, focusing on power density and available cooling loops.
2. **Vendor Collaboration**: Partner with specialists to design tailored systems, incorporating CDUs and heat rejection methods suited to the facility.
3. **Hybrid Approaches**: Combine D2C with air cooling for comprehensive coverage.
4. **Retrofitting**: Minimize downtime by phasing implementation, starting with high-density racks.
5. **Monitoring**: Integrate leak detection and performance analytics for ongoing optimization.

For existing chilled water-cooled Data Centers, adding CDUs and manifolds can enable D2C without full overhauls. D2C excels in targeted applications but may require supplementation, while immersion offers comprehensive cooling at a potentially higher cost.

Future Trends

Looking ahead to 2030 and beyond, D2C cooling will integrate with AI-driven optimizations and two-phase systems for enhanced efficiency.

Industry standards will accelerate adoption, and heat reuse (make the Heat do work) will become standard for net-zero Data Centers.

As chip power transients increase, responsive D2C designs will address challenges like material incompatibility.

Ted J. Pappas, Director of Operations

Projections suggest liquid cooling will dominate, with D2C playing a pivotal role in sustainable, high-performance infrastructure.

Conclusion

Direct-to-Chip Cooling represents a paradigm shift in Data Center Thermal Management, offering unparalleled efficiency and performance for modern workloads. By addressing its challenges through strategic implementation and innovation, operators can future-proof their facilities while controlling energy consumption. As the industry evolves, D2C will be integral to meeting the demands of an increasingly Data-Driven world.

Ted J. Pappas, Director of Operations