ACTIONABLE INSIGHTS FROM INDUSTRY EXPERTS

# GREENER DATA | VOLUME THREE

FOREWORD BY DEAN NELSON
INTRODUCTION BY JAYMIE SCOTTO CUTAIA


MARY ALLEN | DENITZA ARGUIROVA | DR. ATIF ANSAR | JACK BACKES | ALLAN BEDWELL | JAMES BEER
PIERRE-ADRIEN BEL | FERNANDA BELCHIOR | BRUNO BERTI | ROBERT BIANCO | FABIOLA BORDINO
MICHAEL BRAND | JIM BUIE | TATE CANTRELL | STEVEN CARLINI | MANFREID CHUA | NANCY COBLENZ
WES CUMMINS | JOHN DAY | ALISON DEANE | JAY DIETRICH | MICHAEL DONOHUE | ANNA DOWSON
USAMA EL SHAMY | MATTHEW ENGLERT | SEAN FARNEY | LUKE FOXLEY | MATHIEU FRANCOIS
ELIF GAMZE KAYA OK | MIRANDA GARDINER | JENNIFER HOLMES | JONATHAN JÜRGENS | MÜGE KARASAHIN
SUSANNA KASS | BILL KLEYMAN | PETER LANTRY | ANDY LAWRENCE | LINDA LESCUYER | PATRICIA LEYVA
ANDREW LINDSEY | LOUIS LIU | YURY LUI | NABEEL MAHMOOD | CARA MASCINI | CHRIS MILLER | EHSAN NASR
DINA NASSAR | PETER PANFIL | OLIVIER PASCAL MALANDA | KAREN PETERSBURG | PAUL QUIGLEY
SAMUEL RABINOWITZ | JAKE RASWEILER | M. REALI-ELLIOTT | MARTIN REED | ELENA REHMAN | SHAOLEI REN
MAXIE REYNOLDS | AUGUSTIN ROCA | JAMES ROGERS JONES | MUHAMMAD SARWAR | KATHERINE SCHWIND
CHARLIE SELLARS | BILL SEVERN | MELISA SIMIC | NICOLE STAROSIELSKI, PhD | FRANCOIS STERIN
DAVE STERLACE | JIM SUMMERS | WES SWENSON | CJ TANG | MELINA TISOPULOS | JIM TYLER
ALICK WAN | MARGOT WEISTROFFER | VICKI WORDEN

**ISBN:** 979-8257586545

TWENTY-SEVEN

# Accelerating AI Cooling While Managing Risk and Sustainability

CJ Tang, OptiCool Technologies

The space race was not only a contest of technological ambition, but a large-scale exercise in risk management. Putting humans on the moon required unprecedented capital investment, rapid innovation, and long-term planning under conditions of extreme uncertainty. Engineers were forced to make architectural decisions years in advance, knowing that technology, mission requirements, and constraints would continue to evolve.

The Apollo program ultimately succeeded by favoring flexibility and modularity over rigid, monolithic design. Systems were engineered to adapt, provide redundancy, and remain viable even when assumptions changed — a philosophy that proved critical during missions such as Apollo 13. The program inspired generations, delivered transformative technological advancements, and did so at a cost of approximately $280 billion when adjusted for inflation.

Today's race to build AI infrastructure shares many of the same characteristics. Massive capital deployment, accelerated technology cycles, and uncertain long-term outcomes are forcing data center operators to make high-stakes infrastructure decisions before requirements are fully known.

In 2025 alone, venture capital funding for AI reached $202 billion, a 75% increase over the prior year, excluding hyperscaler and enter-

prise spending. On the demand side, strong interest does not always translate into binding, long-term commitments, and announced projects may be delayed, renegotiated, or canceled. On the supply side, intense competition for power, equipment, and skilled labor has extended delivery timelines. By the time facilities come online, technology roadmaps and market conditions have often shifted.

In this environment, adaptability is no longer optional. Infrastructure decisions must limit downside risk while remaining flexible enough to evolve. Nowhere is this more evident than in data center cooling, where choices made today can lock operators into long-term capital commitments, operational constraints, and environmental consequences.

## Large Investments, Long-Term Reward, and Near-Term Risk

Large language models and generative AI represent major technological achievements, yet the industry remains in its early stages. Building the IT, cloud, and AI infrastructure required to support sustained economic value is a generational challenge — one marked by experimentation, false starts, and rapid iteration.

The early automotive industry offers a useful parallel. When Henry Ford introduced the Model T in 1908, several hundred automobile manufacturers existed. By 1929, only 44 remained. Whether today's leading AI platforms will dominate long-term or be displaced by emerging entrants remains an open question.

Industry analysts largely agree on this trajectory. Gartner estimates that more than 40% of agentic AI projects will be canceled by the end of 2027 due to escalating costs, unclear business value, or inadequate risk controls,[1] while an MIT study places that figure as high as 95%.[2] At the same time, Gartner predicts that by 2028, at least 15% of day-to-day work decisions will be made autonomously through agentic AI, and one-third of enterprise software applications will include agentic AI functionality.[3]

For data center operators, this combination of long-term potential and short-term volatility makes risk mitigation a central infrastructure strategy.

## Cooling Infrastructure Decisions as Risk and Sustainability Strategy

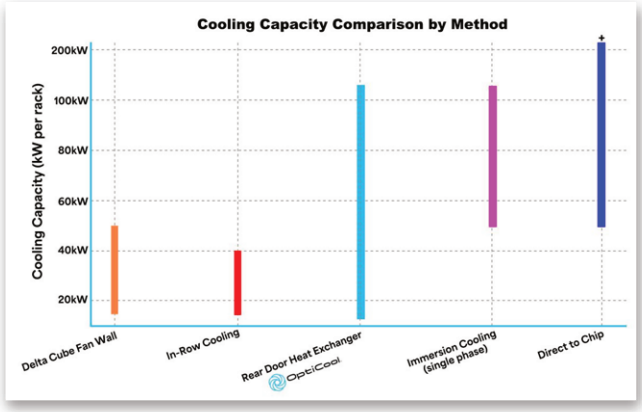
Mitigating risk requires limiting downside exposure in the face of uncertainty—financial, operational, and increasingly environmental. So how do operators manage risk in an ever-evolving environment? One of the clearest examples can be seen through the lens of data center cooling strategies, where decisions made today can lock operators into long-term capital commitments with limited flexibility.

Much of the industry focus has centered on direct-to-chip (DTC) liquid cooling to support escalating power densities. While DTC solutions are well suited for extreme hyperscale deployments, they often require large upfront investments and are tightly coupled to specific processor designs. As silicon roadmaps continue to evolve at an accelerated pace, the risk of premature obsolescence—and stranded infrastructure assets—becomes increasingly real.

Beyond financial exposure, stranded assets also carry environmental consequences. Cooling strategies that extend infrastructure lifespans, enable reuse, and delay large-scale retrofits reduce embodied carbon and material waste. In this context, flexibility is not only a risk management tool, but a sustainability strategy.

## Operational Flexibility with Liquid-Cooled Rear Door Heat Exchangers

Liquid-cooled rear door heat exchangers (RDHx) occupy an important middle ground between traditional air cooling and DTC liquid cooling. While hyperscale AI clusters may operate at 200kW or more per rack, a wide range of enterprise AI and HPC workloads fall below that threshold and encompass diverse server architectures and heat densities.




It is within this mixed-density environment that RDHx solutions excel. Cooling is applied at the rack level — more targeted than room-based air cooling, yet non-invasive to server hardware and independent of processor-specific cold plates. Often referred to as liquid-assisted air cooling, RDHx combines high-capacity liquid heat removal at the rear door with traditional air cooling within the rack, maximizing compatibility across server types.

Maximum heat removal capacity for liquid-cooled rear door heat exchangers reaches approximately 120kW per rack, making them well suited for hybrid cooling architectures in high-density AI environments. As rack power densities continue to scale beyond the limits of air cooling alone, many next-generation systems are expected to rely on direct-to-chip liquid cooling for the majority of heat removal, supplemented by rack-level solutions to manage residual thermal load and airflow efficiency.

For example, Nvidia's Rubin Ultra NVL576 platform, targeted for release in 2027,[4] is designed to support power densities of up to 600kW per rack. Industry estimates suggest that approximately 80% of this heat load will be removed via direct-to-chip liquid cooling, leaving roughly 120kW of residual heat to be managed at the rack and room level. In these environments, RDHx systems can be integrated alongside direct-to-chip cooling as part of a hybrid strategy, removing remaining heat without redesigning server hardware or data hall infrastructure.

By upgrading cooling capacity at the rack level rather than

redesigning entire data halls, operators can scale AI deployments while minimizing new construction, material use, and the environmental impact associated with overbuilding infrastructure.

## Reducing the Risk of Cooling Infrastructure Decisions

DTC liquid cooling requires significant upfront capital investment and multi-year amortization. As GPU power consumption continues to rise, new cooling architectures are emerging. Nvidia has announced future AI platforms targeting rack power densities in the 600kW range7, utilizing newly designed microchannel cover plates. Microsoft has also introduced microfluidic cooling concepts where liquid channels are etched directly into silicon.

In such a fast-moving landscape, the risk of today's cooling investment becoming obsolete tomorrow is substantial.

One effective de-risking strategy is to deploy liquid-cooled RDHx solutions to capture near-term AI demand while deferring capex-intensive DTC investments until technology roadmaps and market requirements stabilize. RDHx systems offer lower entry costs and leverage existing cooling infrastructure:

- Rear doors can be retrofitted onto existing racks with minimal disruption
- They operate alongside room-level air cooling, sharing thermal load
- They integrate seamlessly with hot- and cold-aisle containment

This approach avoids early lock-in to processor-specific cooling architectures while enabling immediate revenue generation.

Hybrid cooling architectures that integrate DTC liquid cooling with RDHx provide a complete and efficient heat removal solution for both current and future data centers. DTC systems remove the majority of heat at the source, while RDHx solutions capture

remaining thermal load at the rack level, stabilize exhaust temperatures, and reduce dependence on large-scale facility redesigns.

Together, these approaches enable operators to support escalating rack densities without overbuilding infrastructure, preserve flexibility, and reduce the risk of stranded cooling assets as AI workloads and roadmaps change.

## Investment Protection and Sustainable Infrastructure Design

Liquid-cooled RDHx systems are highly adaptable across multiple generations of IT equipment. Because heat is removed without cold plates, compatibility with future CPUs and GPUs is preserved. Rear doors can be redeployed, upgraded, or relocated as customer requirements evolve, providing a high degree of investment protection.

This adaptability also supports sustainable infrastructure practices. Reuse, redeployment, and extended asset life reduce the need for frequent replacement and the embodied carbon associated with new equipment. By contrast, DTC and immersion cooling systems are often tightly coupled to specific hardware designs, making repurposing more difficult.

RDHx solutions are particularly well suited to mixed-density colocation and multi-tenant data centers. Rear doors of varying capacities can be deployed rack by rack, eliminating hotspots without segregating AI workloads into dedicated zones. High-density AI racks can operate alongside lower-density workloads, simplifying planning and deployment while maximizing utilization of existing space.

Hybrid cooling strategies further enhance efficiency. Traditional air cooling manages ambient conditions while RDHx handles high-density racks. Heat is rejected via chilled water or DX systems without adding to room-level thermal load. High-efficiency two-phase RDHx systems can achieve PUE values as low as 1.02 (excluding chilled water or DX energy), contributing to reduced overall energy consumption and emissions.

## Conclusion

Liquid cooling — whether direct-to-chip, immersion, or rear doors — is essential to support the escalating thermal demands of AI infrastructure. However, the accelerated pace of silicon innovation introduces both market and technology risk for data center operators, as today's cutting-edge solutions can quickly become tomorrow's constraints.

In this environment, liquid-cooled rear door heat exchangers offer a uniquely low-risk path forward. Their hardware-agnostic design, lower upfront investment, and ability to be redeployed across multiple generations of IT equipment provide both financial resilience and long-term sustainability benefits. By extending infrastructure lifespans and reducing the need for large-scale retrofits, RDHx supports lower embodied carbon and more responsible use of capital and materials.

Much like the Lunar Orbit Rendezvous strategy chosen for the Apollo missions, RDHx provides a modular, adaptable architecture designed for uncertainty. Redundancy, flexibility, and reuse were central to mission success in space — and those same principles are increasingly critical for AI-era data centers. As workloads, power densities, and technology roadmaps continue to evolve, RDHx enables operators to remain efficient, scalable, and sustainable without locking into premature design decisions.

The question is no longer whether liquid cooling is required — it is how to deploy it in a way that balances performance, risk, and sustainability. What's your risk mitigation strategy for AI cooling?

---

1. Gartner. "Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027." June 25, 2025. https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027
2. MIT NANDA. "The GenAI Divide: State Of AI In Business 2025." July 2025. https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf
3. Coshow, Tom. "Intelligent Agents in AI Really Can Work Alone. Here's How." October 1, 2024. https://www.gartner.com/en/articles/intelligent-agent-in-ai
4. Moss, Sebastian. "Nvidia's Rubin Ultra NVL576 rack expected to be 600kW, coming second half of 2027." March 18, 2025. https://www.datacenterdynamics.com/en/news/nvidias-rubin-ultra-nvl576-rack-expected-to-be-600kw-coming-second-half-of-2027/

# About the Author

## CJ TANG

CJ Tang is Chief R&D Scientist at OptiCool Technologies and a seasoned thermal systems engineer with more than 15 years of experience developing and optimizing advanced energy and thermal technologies. His expertise spans gas turbines, heat pumps, refrigeration, thermal management systems, and fuel cells, with a strong focus on translating fundamental thermodynamics into scalable, real-world solutions. CJ holds a PhD in Mechanical Engineering from Penn State University.

At OptiCool, CJ leads research and development efforts focused on advanced cooling technologies. He brings a deep proficiency in multi-domain steady-state and dynamic system modeling, supporting every phase of technology development—from concept definition and proposal preparation through testing, validation, and publication.

Previously, CJ served as modeling and simulation task lead for the U.S. Department of Energy R&D program to develop a 10 MW supercritical $CO_2$ pilot plant test facility. He also served as Principal Investigator for an ARPA-E–funded heat pump program, leading a cross-functional team to develop an advanced absorption heat pump technology.

Greener Data: Volume Three is a timely, industry-driven guide to advancing sustainability across digital infrastructure—from data centers to global networks.

Featuring insights from leaders across the ecosystem, this volume explores how we scale responsibly in the age of AI.

Visit GreenerData.net to learn more and purchase the full book on Amazon.