INTELLIGENCE UPDATE

GPU breakthroughs bring real-time CFD analysis closer


John O'Brien

5 Aug 2025

Advances in GPU performance are bringing real-time computational fluid dynamics (CFD) analysis closer to reality. Real-time CFD analysis would enable CFD simulations to experience minimal latency, allowing operations teams to study and respond to the live physical system.

For over two decades, CFD has been utilized to design cooling systems, optimize data center layouts and identify inefficiencies. Today, broader applications in operational efficiency and resiliency are supported through integration with physical systems such as power and IT infrastructure (see *Digital twins: the role of simulations*).

However, slow processing and rendering of simulations remain persistent challenges to using CFD in operations. For example, analyzing airflow in a data hall can take anywhere from a few hours to several days.

A recent cooling simulation experiment by supplier Cadence, took 8 hours and 15 minutes to render on 16-core Intel Xeon Gold 5120 CPUs. In contrast, a single Nvidia GB200 GPU completed the same simulation in under 10 minutes — and with higher fidelity. It is likely that Cadence's software was running on Nvidia's own GPU infrastructure.

GPUs may offer a solution for significantly reducing CFD analysis timeframes because they are optimized to run calculations in parallel, across tens of thousands of discrete processing elements. For CFD workloads, this approach has an advantage over using dozens of faster and more complex CPU cores.

No more rear-view mirror analysis

While a 10-minute analysis time is a significant improvement over most current CFD applications, what if GPUs could reduce that to 5 minutes — or even just a single minute?

In that case, attempts to perform system diagnostics would no longer rely on past performance. Instead, analysis and decisions would be driven by real-time feedback from the live operational environment. Since conditions in live environments can change in seconds, reducing the time to

analysis is critical.

Another key benefit would be the ability to test virtual operating scenarios in a safe sandbox, using live system data. Both current and potential future environments could be modeled and tested over time.

CFD could then be used to not only monitor the cooling environment — helping to identify inefficiencies and areas for improvement — but also to compare OEM equipment and alternate configurations for suitability within existing and future hybrid cooling systems (see below).

Currently, the closest alternative applications are less computationally intensive, machine learning or AI-based software-as-a-service platforms. For example, EkkoSense's EkkoSim cooling simulation platform offers an "actual versus expected" system performance analysis, based on live and historical data. While useful for incremental improvement of live environments, these tools may not match the accuracy of CFD predictions, which utilize physical data and physics calculations.

Understanding that CFD utilizes precise physical data, calculations, and live machine data from sensors and equipment may enhance its credibility in future automated control applications.

AI factories and giant data centers

Large-scale, high-density data centers for AI face some of the most complex infrastructure challenges. Their power and cooling requirements raise significant resiliency concerns, as current-generation data center mechanical and electrical (M&E) equipment was not designed to support multi-hundred-kilowatt racks, let alone gigawatt data center capacities.

An immediate concern is managing hybrid air and liquid cooling systems, right down to the server. Emerging CFD applications are now capable of modeling hybrid topologies and flows through chilled water systems, liquid-cooled IT racks, immersion cooling tanks, rear-door heat exchangers and air economizers. While simulations may prove less useful than sensors and alarms for immediate emergency response, real-time CFD feedback based on reliable physics data may offer valuable early warnings of potential problems.

Suppliers such as Cadence, Schneider Electric, ETAP and Vertiv now integrate their software and data — including 3D models of their power and cooling equipment — into Nvidia's Omniverse digital twin platform for GPU-accelerated simulations.

While the stated aim is to create blueprints for the design and optimization of gigawatt-scale AI factories, there is also a critical engineering imperative to reduce the risk and improve operating resiliency at such high densities.

Will most operators need real-time CFD?

Achieving real-time CFD analysis would give operators more than just the time they previously lost waiting for their physics simulations to process. It would increase the long-term value of their CFD investment, moving it beyond what is often seen as a once-and-done exercise at the start of a cooling infrastructure implementation or upgrade.

However, real-time CFD analysis applications will require vast amounts of data and computational processing power to deliver on some of their potential.

Today, most enterprises and operators may consider investing in high-end GPUs impractical or cost-prohibitive. However, this may change if advances in GPU technology continue to drive improvements in price-performance.

GPU servers — such as those offered by Supermicro, Dell and HPE — may provide more enterprises and operators access to the high-performance infrastructure needed for real-time CFD, while also appealing to those that wish to retain their operational data on-premises.

GPU-as-a-service offerings from hyperscale cloud providers and neoclouds, such as CoreWeave, may be utilized in pilot projects and early experiments. However, these options could gain wider acceptance as longer-term solutions — provided they are trusted to deliver reliable, secure and cost-effective high-performance GPU computing on demand.

Other related reports published by Uptime Institute include:
Uptime Institute Global Data Center Survey 2025
Digital twins: the role of simulations
Digital twins: reshaping AI infrastructure planning
Neoclouds: a cost-effective AI infrastructure alternative

ABOUT THE AUTHOR

John O'Brien

John is Uptime Institute’s Senior Research Analyst for Cloud and Software Automation. As a technology industry analyst for over two decades, John has been analyzing the impact of cloud migration, modernization and optimization for the past decade. John covers hybrid and multi-cloud infrastructure, sustainability, and emerging AIOps, DataOps and FinOps practices.

jobrien@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. Its Tier Standard is the IT industry's most trusted and adopted global standard for the proper design, construction, and operation of data centers – the backbone of the digital economy. For over 25 years, the company has served as the standard for data center reliability, sustainability, and efficiency, providing customers assurance that their digital infrastructure can perform at a level that is consistent with their business needs across a wide array of operating conditions.

With its data center Tier Standard & Certifications, Management & Operations reviews, broad range of related risk and performance assessments, and accredited educational curriculum completed by over 10,000 data center professionals, Uptime Institute has helped thousands of companies, in over 100 countries to optimize critical IT assets while managing costs, resources, and efficiency.