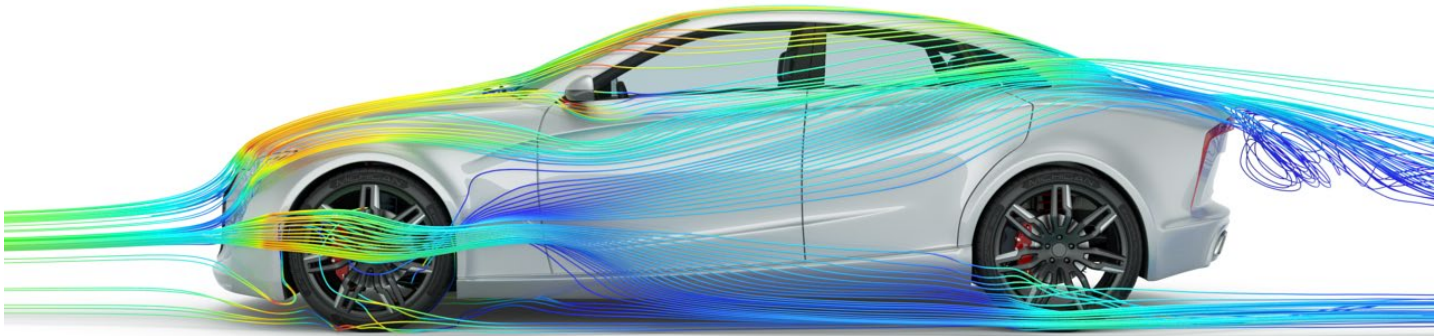


ACCELERATING INNOVATION: RUNNING GPU-POWERED SIMULATIONS WITH ALTAIR ONE ON MICROSOFT AZURE

Altair | Microsoft | NVIDIA



Introduction

In the race to bring better products to market faster, design and engineering teams rely on multi-physics simulations. Increasingly, these simulations involve state-of-the-art solvers and AI tools that require high-performance GPUs. For computer-aided engineering (CAE) simulations, traditional on-premises infrastructure often falls short—it is slow, expensive, and frequently too hard to deploy and scale. Fortunately, the cloud is changing that.

With [Altair One™](#) and [Microsoft Azure®](#), designers and engineers can run simulations using industry-leading tools and the latest cloud infrastructure on demand. Whether optimizing the aerodynamics of a vehicle, modeling structures, or predicting flow behavior around complex geometries, design teams can leverage the latest NVIDIA® GPUs in the Azure cloud to accelerate simulations, explore design alternatives, and make informed decisions—without bottlenecks.

This paper explains how engineers can leverage GPU-accelerated solvers on Azure cloud infrastructure and provides real-world benchmarks that demonstrate the benefits of running Altair One CAE simulations on Microsoft Azure.

Evolving requirements for design and simulation

Computer-aided engineering (CAE) is a cornerstone of modern manufacturing, enabling faster product development, reduced costs, and the delivery of higher-quality, more reliable products. As product designs become more complex, engineers must simulate assemblies at higher resolutions and account for intricate multi-physics interactions, including fluid dynamics, structural mechanics, heat transfer, and electromagnetics.

In this environment, manufacturers compete based on the strength of their high-performance computing (HPC) capabilities. Advanced simulation and AI workloads require powerful CPUs and GPUs, large memory footprints, and high-speed interconnects.

However, building and managing HPC infrastructure presents significant challenges, including high capital expenditures (CAPEX), physical limitations such as space, power, and cooling, as well as the need to recruit skilled personnel and keep pace with new technologies. As a result, many organizations are turning to the cloud to overcome these barriers and gain the agility required to innovate at scale.

Operating in the cloud brings clear benefits

By shifting HPC and AI workloads to the cloud, organizations can avoid many of the pitfalls of on-premises deployments while realizing multiple benefits:

- On-demand access to the latest CPUs and GPUs
- Faster deployment times for compute and storage
- Seamless scaling to address changing business needs
- Predictable pricing through flexible, consumption-based models
- Reduced capital expenditures by avoiding costly infrastructure investments

While the cloud addresses many infrastructure-related challenges, access to scalable HPC infrastructure is only part of the problem. Manufacturers also need to deploy and manage the full range of design and simulation software critical to engineering productivity. They also need tools that can facilitate collaboration between multidisciplinary design and engineering teams working in different locations.

Altair One – the cloud innovation gateway

Altair One is a revolutionary cloud service for collaborative engineering, data engineering, and the development of analytical applications. Built on a robust HPC backbone and decades of simulation, HPC, and AI expertise, Altair One provides engineers access to the tools, data, and computing resources they need across every stage of the product development life cycle.

Through the Altair One cloud portal, users can choose from over **200** Altair and third-party design, simulation, analytics, and AI-based tools. These tools can be downloaded for local execution, run directly within Altair One (hosted on Microsoft Azure), or launched to other supported public clouds—all from a unified interface.

As shown in Figure 1, applications can be launched onto the Azure Cloud by selecting custom or pre-packaged application appliances via the Altair One Marketplace.

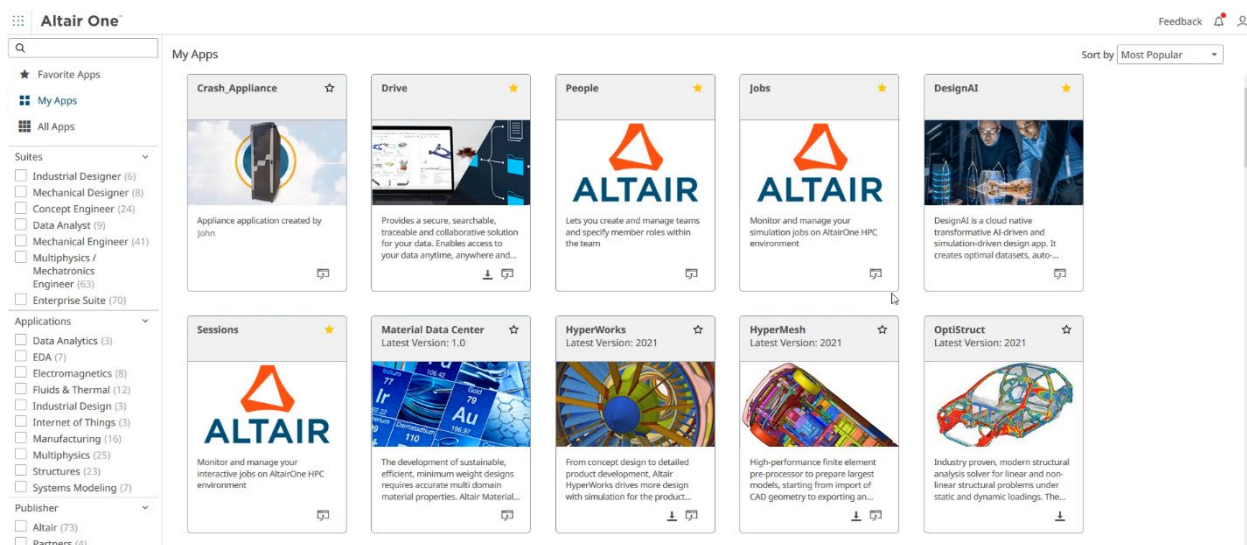


Figure 1 – Easily access CAE applications via the Altair One Marketplace

Just as operating in the cloud helps organizations avoid the cost and complexity of managing infrastructure, Altair One brings the same benefits to software. With a simple appliance-based model, applications and essential components, such as workload managers, remote access solutions, license servers, and prerequisite middleware and software libraries, are all deployed and managed automatically.

Details such as data management, access control, and launching and scaling cloud infrastructure and shared storage are handled automatically within the Altair One environment.

The rise of GPU-based simulation

In CAE, GPU-based solvers have been making steady inroads. GPUs are now used to accelerate physics-based simulations in multiple domains, including dynamic explicit analysis, computational fluid dynamics (CFD), and electromagnetics (EM).

Traditionally, high-fidelity simulations required hours or even days to run on large clusters of high-performance servers. Today, engineers can significantly reduce the time and cost needed for large simulations by leveraging the latest NVIDIA® GPUs, which often enable the execution of the same workloads on a single server or cloud-based virtual machine (VM). These improvements also bring significant gains in energy efficiency and lower carbon emissions compared to traditional CPU clusters.

In the early 2010s, most CFD codes were CPU-centric, and GPUs were considered niche and experimental. The use of GPUs was limited to small or medium-sized models due to memory limitations and a mismatch between solver algorithms and GPU architectures.

This began to change around 2018 with the introduction of more capable GPUs featuring higher memory bandwidth, larger memory footprints, and improved interconnects. These advances enabled support for larger simulation models. At the same time, new solvers emerged with GPU-native algorithms, delivering faster and more cost-efficient simulations.

As examples, Altair® ultraFluidX®, part of Altair CFD™, uses the Lattice-Boltzmann Method (LBM)—a highly parallel, GPU-optimized approach to CFD. Similarly, Altair® nanoFluidX®, employs Smoothed Particle Hydrodynamics (SPH), a meshless, Lagrangian method that maps naturally to GPU architectures.

Today, GPU simulation is a mainstream technology. Depending on the model size, solver, and simulation parameters, large-scale CFD workloads can run 10 to 18x faster on GPUs than on CPUs—fundamentally transforming engineering workflows and enabling faster, more frequent design iterations.¹ Figure 2 shows key milestones in this evolution.

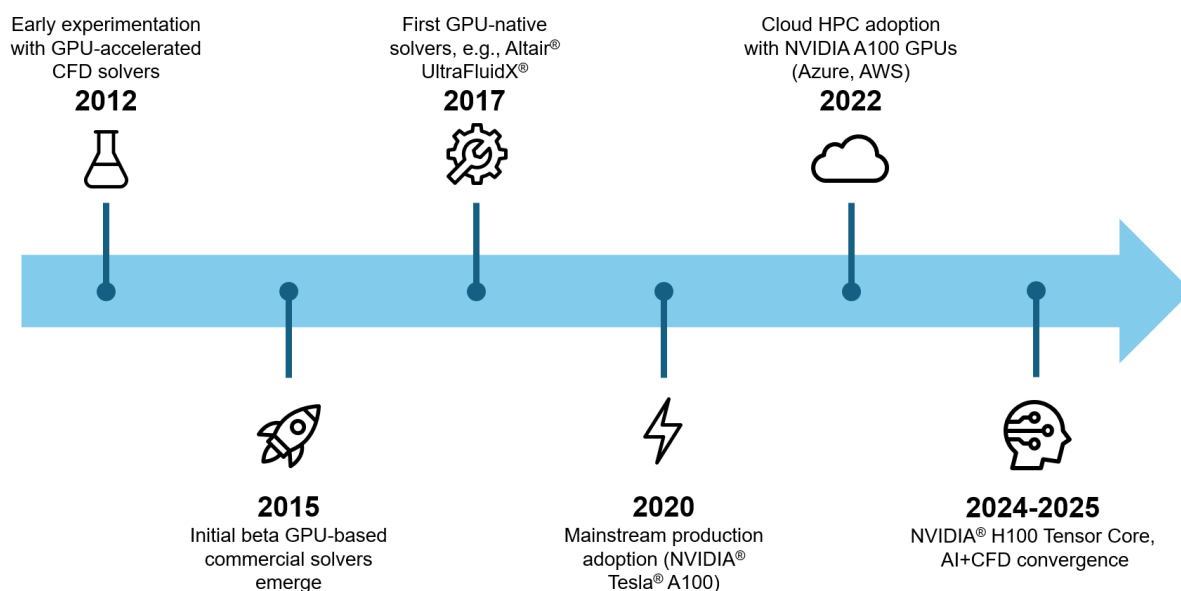


Figure 2 - A timeline showing the evolution of GPU-accelerated CFD

GPU accelerators are also critical for new AI-powered workloads. CAE workflows increasingly leverage AI-enhanced tools for concept generation, meshing, and pre-processing. In addition, machine learning (ML) models trained on prior simulation data can accurately predict results in near real time without requiring exhaustive simulation, providing a significant productivity boost.

GPU-enabled simulations in the cloud

Altair One dramatically simplifies running GPU-based simulations. Users simply choose from pre-defined or custom appliances (VM images) and launch them in the cloud. Engineers can access a comprehensive suite of solvers spanning multiple physics domains. The solvers tested in this paper are listed in Table 1.

Table 1 - GPU-accelerated Altair solvers tested.

Solver	Use cases	Solver type
Altair® ultraFluidX®	Aerodynamics, thermal simulation, blood flow, and flows in complex geometries	Lattice Boltzmann Method (LBM)
Altair® nanoFluidX®	highly dynamic fluid-structure interactions, such as splashing, sloshing, and lubrication in rotating machinery	Smoothed-particle Hydrodynamic (SPH), Lagrangian solver

¹ Results will vary depending on the GPUs and CPUs compared. This comparison is based on an Altair nanoFluidX simulation comparing two different 32-core Intel® Xeon® processors to a configuration with four NVIDIA® Tesla V100 Tensor Core GPUs. See the [Altair nanoFluidX documentation](#) for details.

Altair® EDEM™	Bulk material handling, granular flow, Hoppers, conveyors, mixers, mills, crushers, etc.	Discrete Element Method (DEM), particle-based Lagrangian solver
---------------	--	---

When users run applications in Altair One, VMs are automatically deployed in the Azure cloud to support functions such as Altair Desktop, the cluster head node, and various CPU- and GPU-enabled compute nodes.² Simulations that benefit from acceleration are automatically dispatched to GPU-capable VMs.

Most production users will prefer to create and run custom appliances using [Altair® Navops®](#), a hybrid cloud scaling and cost management facility included in Altair One. With custom appliances, customers have full control over the applications installed on the appliance, the VMs used to run their workloads, the cloud regions where simulations run, and the accounts used for billing purposes.

In Azure, users can utilize Microsoft Cost Management in conjunction with Altair Navops to monitor and manage cloud spending across various CAE teams, leveraging Altair and third-party tools.

NVIDIA GPUs for HPC and AI

NVIDIA's latest H100 and H200 Tensor Core GPUs represent the cutting edge of accelerated computing, offering exceptional performance for deep learning, generative AI, and computer-aided engineering (CAE) workloads.

Based on the [NVIDIA Hopper™ architecture](#), the [NVIDIA H100 Tensor Core GPU](#) delivers massive parallelism and features such as fourth-generation Tensor Cores and a Transformer Engine optimized for deep learning, making it ideal for large-scale AI training and inference. The [NVIDIA H200 Tensor Core GPU](#) extends these capabilities with additional high-bandwidth memory (HBM3e), delivering greater memory capacity and bandwidth.

The [NVIDIA Blackwell architecture](#) further advances these capabilities with increased memory bandwidth, higher floating-point throughput, and improved efficiency for AI applications. The specifications of selected NVIDIA GPUs are provided in Table 2.³

Table 2 - Comparison of NVIDIA GPUs

Feature	NVIDIA A100 SXM	NVIDIA H100 SXM	NVIDIA H200 SXM	NVIDIA GB200 (2xB100) ⁴
Architecture	Ampere	Hopper (HBM3)	Hopper (HBM3e)	Blackwell (HBM3e)
FP64 Performance (TFLOPS)	9.7	34	34	80
FP32 Performance (TFLOPS)	19.5	67	67	160
Tensor Performance (TF32)	312	989	989	3,962
Memory Capacity (GB)	80	80	141	384
Memory Bandwidth (TB/s)	2.039	3.35	4.8	8.0
NVLink Bandwidth (GB/s)	600	900	900	1,800

² See the [Altair One documentation](#) for details on default cloud instances used by various Altair One standard appliances.

³ See [NVIDIA A100 SXM specifications](#), [NVIDIA H100 SXM specifications](#), [NVIDIA H200 SXM specifications](#). The NVIDIA GB200 specifications are published in the [NVIDIA Blackwell Architecture Technical Brief](#) (pages 13 and 14). A100 PCIe cards and H100/H200 NVL cards are not shown in this comparison since cloud hyper-scalers don't typically support them.

⁴ The NVIDIA GB200 (Grace Blackwell) is a multi-chip module that combines 2 x NVIDIA B100 GPUs based on the Blackwell architecture, and 1 x NVIDIA Grace (ARM-based) CPU. NVIDIA does not offer the NVIDIA B100 as a standalone product.

High-performance GPU instances on Microsoft Azure

Microsoft offers a wide selection of accelerated VMs, including the [ND-H100-v5](#) and [ND-H200-v5](#) series, designed for large-scale HPC and AI workloads. Ideal for demanding CAE workloads, these VMs are based on NVIDIA's Hopper architecture and feature the H100 and H200 GPUs, respectively.

The Standard_ND96isr_H100_v5 VM provides 96 vCPUs, 1,900 GB of system memory, and 8 x NVIDIA H100 Tensor Core GPUs (SXM form factor), each with 80 GB of HBM3 memory. Each GPU on the VM is equipped with a dedicated 400 Gb/s NVIDIA Quantum-2 CX7 InfiniBand interface, providing 3.2 Tb/s of aggregate interconnect bandwidth per VM and enabling scalability to thousands of GPUs across multiple nodes.

The Standard_ND96isr_H200_v5 offers similar VM-level specs but is powered by eight NVIDIA H200 GPUs, each with 141 GB of HBM3e memory and features 4.8 TB/s of memory bandwidth for optimal performance.

Microsoft also offers the [ND-GB200-v6](#) series VMs based on the NVIDIA Grace Blackwell GB200 platform. These VMs include 2 x NVIDIA Grace CPUs (Arm v9 architecture, 128 vCPUs total), 900 GB of system memory, and 4 x NVIDIA B100 GPUs, each with 192 GB of HBM3e memory. These instances utilize NVIDIA's Grace-Blackwell architecture to deliver unified CPU-GPU memory coherence along with exceptional performance and energy efficiency for AI workloads. Altair One users can choose to run solvers and AI tools optimized for the Arm-based NVIDIA Grace-Blackwell architecture on these powerful VMs.⁵

For engineering simulation, the ND-H100-v5 and ND-H200-v5 series VMs represent a "sweet spot". FP64 and FP32 performance are crucial for both computational efficiency and accuracy. While the superior tensor performance (TF32) of the Grace-Blackwell instances is important for AI, it is less relevant for CFD and other numerical simulation workloads.

The Azure Standard_ND128isr_NDR_GB200_v6 VM offers more memory and ~19% better FP32 performance per GPU compared to the Standard_ND96isr_H100_v5; however, this is more than offset by the fact that the Standard_ND96isr_H100_v5 VM offers more total CPU memory and double the number of GPUs, delivering ~68% better FP32 throughput per VM.⁶

Putting Azure GPU instances to the test

In June 2025, engineers from Altair and Microsoft conducted a comprehensive set of benchmarks to evaluate the performance and scalability of Altair's GPU-accelerated solvers on Microsoft Azure. The benchmarks covered three applications: ultraFluidX and nanoFluidX for computational fluid dynamics (CFD), and EDEM for discrete element modelling (DEM). Benchmarks were run on Microsoft Azure Standard_ND96isr_H100_v5 and Standard_ND96isr_H200_v5 VMs, each equipped with eight NVIDIA H100 or H200 Tensor Core GPUs, respectively.

A total of 20 benchmark tests were run involving 14 different models, as illustrated in Figure 3.⁷ For each test, elapsed simulation time and throughput were recorded on each VM as the number of GPUs was varied.

⁵ Altair announced support for NVIDIA Blackwell and the NVIDIA Grace architecture in Altair One in March of 2025. See [Altair One Cloud Innovation Gateway Achieves Seamless Integration with NVIDIA Omniverse Blueprint for Real-Time Digital Twins](#).

⁶ The NVIDIA H100 SXM delivers 67 TFLOPS of FP32 performance. The NVIDIA B100 GPU (the GPU used in the multi-chip GB200) delivers 80 TFLOPS of FP32 performance, a ~19% improvement (80/67). Given that the ND-H100-v5 has 8 x Hopper GPUs (8 x 67 = 536 FP32 TFLOPS) and the ND-B200-V6 has 4 x Blackwell GPUs (4 x 80 = 320 FP32 TFLOPS), in theory, the ND-H100-v5 offers ~68% better FP32 performance per VM (536/320).

⁷ Additional Altair nanoFluidX test cases were also run on specific GPU configurations for comparison purposes. These additional tests are included in the appendix.



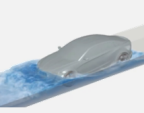
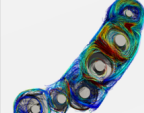



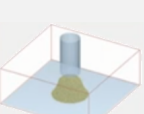
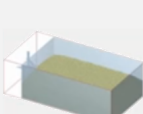


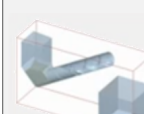


Altair® ultraFluidX®		Altair® nanoFluidX®				
						
Altair ARC SUV model 1 test case	Open Cooling DrivAer (OCDA) 1 test case	Water Wading (WW) model 1 test case	Aerospace Gearbox model (areo_gbx) 2 test cases	Altair e-Gearbox altair_ebgx model 2 test cases	Cuboid model static particles at rest 2 test cases	Dambreak model collapsing water column 4 test cases
Altair® EDEM™						
						
Angle of Repose The steepest angle a pile of granular material can maintain 1 test case	Bed of Material Simulate a bed of materials with a tillage tool 1 test case	Hopper Discharge Model the discharge of powders, including velocity and flow rate 1 test case	Powder Mixer Mixing of bulk solids 1 test case	Screw Auger transporting materials up an incline 1 test case	Mill Bulk material behaviour inside a mill 1 test case	Transfer Chute Simulate the behaviour of a transfer chute 1 test case

Figure 3 - Summary of Altair CFD and Altair EDEM test cases

Details of the results for each test, including configurations tested, elapsed simulation time, and measured throughput where applicable, are included in the Appendix.

Breakthrough performance and scalability

Altair ultraFluidX test results

The ultraFluidX benchmarks simulated external aerodynamics for two reference geometries:

1. Altair's ARC SUV model simulated over 0.05 seconds.
2. The Ford Open Cooling DrivAer (OCDA) model simulated over 0.19 seconds.⁸

These benchmarks highlight the performance benefits of scaling ultraFluidX from a single GPU to an eight-GPU configuration on Azure ND v5 VMs. Figure 4 summarizes simulation runtimes and GPU scaling results for both models.

For the ARC model, elapsed time for the computational phase decreased from 2,744 seconds (45m 44s) on a single H200 GPU to just 470 seconds (7m 50s) on eight H200 GPUs. This represents a 5.83x speedup, corresponding to 73% strong scaling efficiency, and underscores the excellent multi-GPU scalability of the ultraFluidX solver.⁹

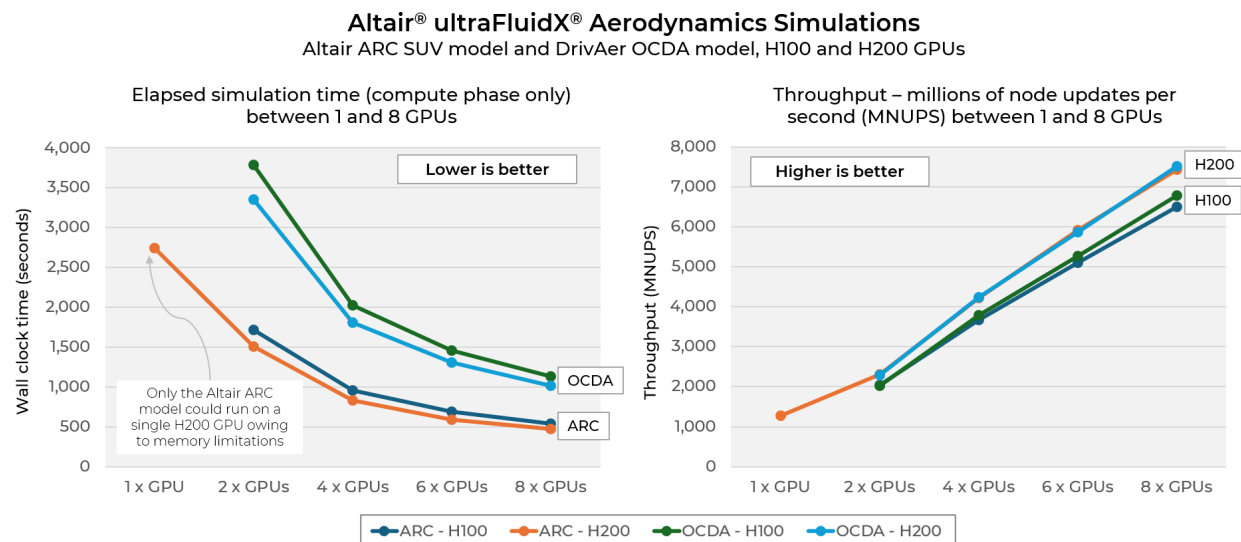


Figure 4 – Summary of Altair ultraFluidX benchmarks

It was not possible to run the 118 GB ARC model on a single H100 GPU, as the model exceeded the H100's 80 GB memory capacity. Similarly, the OCDA model could not be run on a single H100 or H200 GPU due to its memory footprint. As a result, the OCDA model was benchmarked using configurations with two, four, six, and eight GPUs.

As expected, the OCDA model required longer compute time than the ARC model, due to its larger size and longer simulated time (0.19 vs. 0.05 seconds).

Although the H100 and H200 GPUs have the same theoretical compute performance (in TFLOPS), for the ARC model, the Standard_ND96isr_H200_v5 VM achieved a 1.14x speedup compared to the Standard_ND96isr_H100_v5 VM when using all eight GPUs. This performance difference is likely attributable to the higher memory bandwidth of the H200's HBM3e memory (4.8 TB/s vs. 3.35 TB/s).

The OCDA model running across eight GPUs on the Standard_ND96isr_H200_v5 VM delivered 7,515 million node updates per second (MNUPS), a record result for this benchmark. Details of all results, including wall clock times and throughput metrics, are provided in the Appendix.

Altair nanoFluidX test results

A set of ten nanoFluidX benchmark cases was executed on Microsoft Azure Standard_ND96isr_H100_v5 and Standard_ND96isr_H200_v5 virtual machines, scaling from one to eight GPUs. Additionally, a production-scale "water wading"

⁸ Note: The aerodynamics simulations were run just long enough to provide a reliable average throughput. In a production simulation, runtimes would typically be 20x longer or more.

⁹ The total simulation time for the ARC model, including pre-processing, output file initialization time, and total computational time, was reduced from 8,484 seconds (2h 21m 24s) on a single H200 GPU to 4,088 seconds (1h 8m 8s) on 8 x H200 GPUs, a 52% reduction in wall clock time. During the pre-processing and file initialization stages, the 8 x H200 configuration performed 1.49x and 3.14x better than a single H100, respectively, suggesting that CPU and I/O were bottlenecks during these stages of the simulation.

simulation was conducted on both instances. This test simulated an Altair CX-1 vehicle moving at 10 km/h through a 24-meter-long water channel over a 15-second physical time interval.

Figure 5 summarizes the results of the ten standard benchmarks for both VMs. Full details, including wall-clock times and throughput metrics for all test cases, are provided in the Appendix.

- **Runtime range** — Elapsed simulation times for the standard nanoFluidX tests ranged from 2.56 seconds to 12,278 seconds (3h 22m 58s), reflecting the diversity in model sizes and simulation parameters.¹⁰
- **Scaling efficiency** — Across ten test cases, moving from one H100 GPU to eight H200 GPUs on the Azure ND v5 instances delivered an average 5.74x speedup, equivalent to 72% strong scaling efficiency. This substantial reduction in turnaround time enables engineering teams to explore more design alternatives within tight development timelines.

In the compute-bound nanoFluidX simulations, the Standard_ND96isr_H200_v5 VM consistently delivered slightly higher throughput than the H100-based instance, with an average runtime improvement of 1.52% across the ten test cases tabulated.

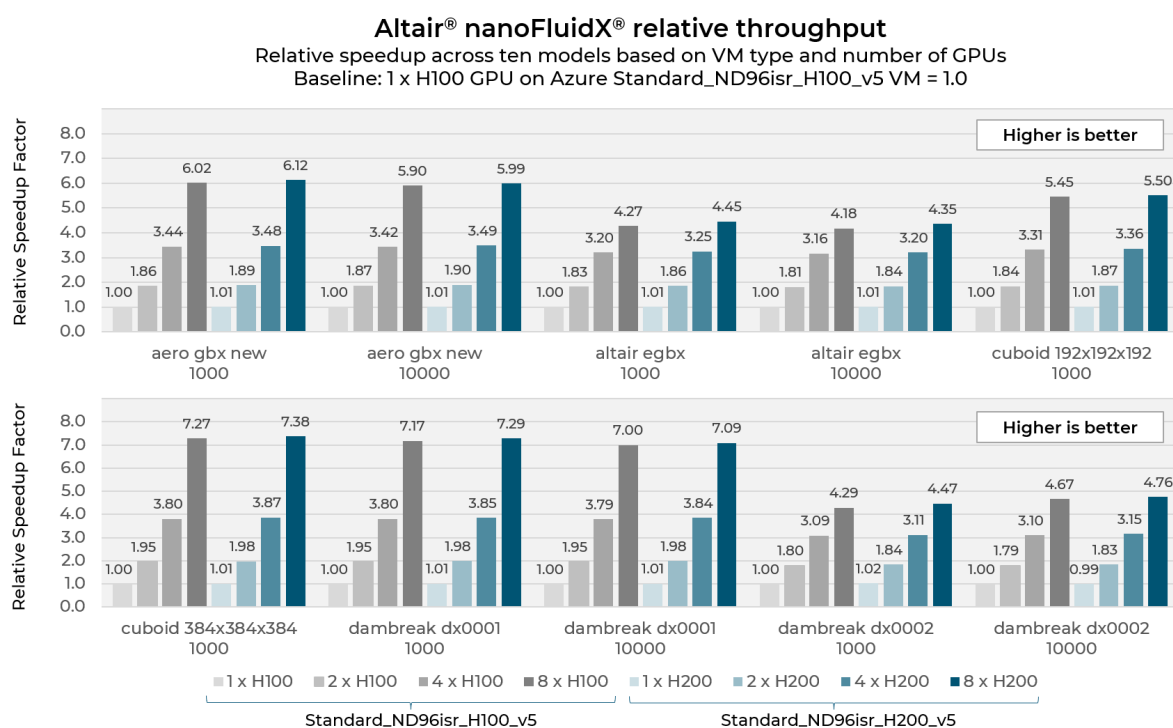


Figure 5 - Altair nanoFluidX benchmarks

Due to the memory demands of the water wading model, the minimum viable configuration was two H200 GPUs. Figure 6 presents the results of this large-scale test. Running the model on eight H200 GPUs reduced the total runtime from 12,869 seconds (3h 34m 29s) to 3,605 seconds (1h 0m 5s), representing a 3.57x throughput gain.

¹⁰ The cuboid_192x192x192_1000 test ran for 2.56 on eight H200 GPUs. The full-scale altair_egbx test ran for 12,278 seconds on four H100 GPUs. All test runtimes are included in the Appendix.

Altair® nanoFluidX® Water Wading benchmark

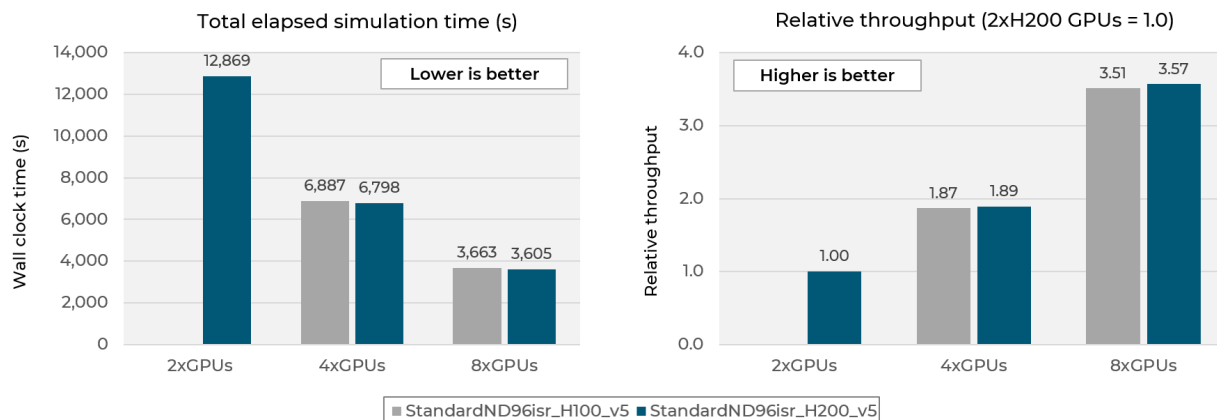


Figure 6 - Altair nanoFluidX water wading benchmark result

Altair EDEM test results

Altair EDEM benefits enormously from GPU-based acceleration. In prior Altair benchmarks involving 2 x NVIDIA A100 GPUs and the same models tested here, performance improvements ranged from 26.8x faster to 187.1x faster compared to a 32-core CPU.¹¹

The EDEM benchmarks were run on both the Standard_ND96isr_H100_v5 and Standard_ND96isr_H200_v5 VMs for each model with one, two, four, and eight GPUs, and the results are presented in Figure 6. Detailed results, including model parameters, elapsed wall-clock time, and relative speedup for each test, are provided in the Appendix.

Scalability varied depending on the model. Running with all eight GPUs active on the Standard_ND96isr_H100_v5 VM resulted in an average 3.90x speedup relative to a single H100 GPU across the tested models. A slightly higher average uplift of 3.95x speedup was observed when using the Standard_ND96isr_H200_v5 VM. The best result was achieved with the Powder Mixer model, where the 8 x H200 GPU configuration ran 5.31x faster than a single H100 GPU, reducing the runtime from approximately 45 minutes to 8 minutes, demonstrating a significant productivity improvement.

Like the other solvers tested, with EDEM, the larger memory and higher memory bandwidth of the Standard_ND96isr_H200_v5 VM delivered slightly better throughput for all test cases compared to the H100 VM. The best choice will depend on whether models fit in memory, instance pricing, and instance availability in a selected cloud region.

¹¹ Altair EDEM supports both CPU and GPU acceleration, making direct comparisons possible. In tests conducted by Altair in 2022, the Screw Auger model ran 26.8x faster on two NVIDIA A100 GPUs compared to a single 32-core CPU. The Bed of Material model ran 187.1x faster. Details are available at community.altair.com.

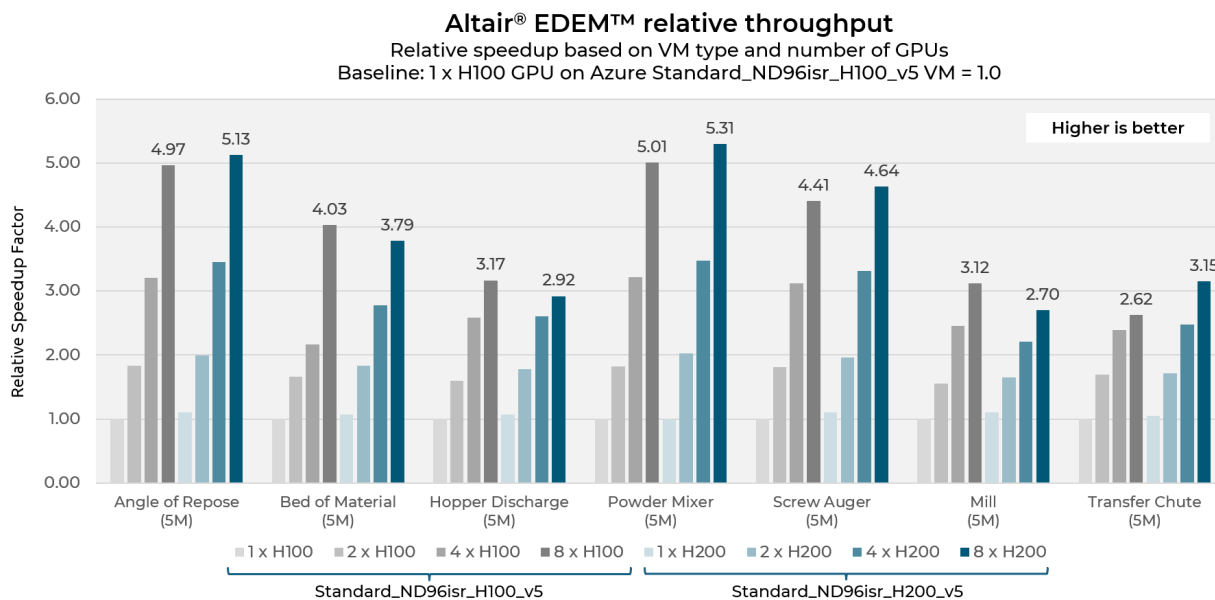


Figure 7 – Summary of Altair EDEM benchmarks

AI-powered CAE in the cloud

Increasingly, manufacturers rely on machine learning (ML) and artificial intelligence (AI) to enhance productivity and efficiency throughout all stages of the product lifecycle. Altair embeds AI capabilities in many of the [Altair HyperWorks](#) tools offered via the Altair One portal. The advantage of this approach is that designers and engineers can benefit from AI-assisted design and simulation, and even train deep learning models, without having specialized knowledge of ML tools, frameworks, or coding.

For example, [Altair® PhysicsAI™](#) utilizes geometric deep learning to learn from prior simulations, thereby creating fast, reduced-order models (ROMs) or surrogate models that provide near real-time predictions of complex physics without requiring exhaustive computer simulations.

AI-based prediction can dramatically impact the design and engineering process. PhysicsAI models can deliver predictions up to 1,000x faster than conventional solver simulations, enabling teams to evaluate more concepts and make better design decisions.¹² Engineers can easily access PhysicsAI through the Altair One portal and train predictive models using Azure-based GPU instances from prior simulations. GPUs are also recommended for other AI tools accessible through Altair One, including [Altair® DesignAI™](#) and [Altair® romAI™](#).

For business and process-level problems, Altair One users can leverage [Altair® RapidMiner®](#), a suite of enterprise data analytics and GPU-accelerated AI tools accessible via Altair One that seamlessly integrates AI into the business fabric.

A Generative Models extension for Altair RapidMiner enables organizations to support new use cases and out-innovate competitors by deploying state-of-the-art Large Language Models (LLMs). Use cases include automated code reviews, natural language interfaces, knowledge extraction from documentation, and design reviews. Users can build or refine new models using their own data, leveraging the 450,000+ publicly available HuggingFace.co and OpenAI models.¹³

Engineers can augment the rich AI-based capabilities available in Altair One, including Altair RapidMiner, with [Azure Machine Learning](#), a fully managed service that enables building, training, and deployment of ML models at scale.

¹² For details, see [Altair® PhysicsAI™ Geometric Deep Learning](#).

¹³ For additional information on Altair RapidMiner generative AI extensions, visit <https://altair.com/generative-ai>.

Getting started with Altair One on Microsoft Azure

To get started running CAE simulations and AI tools via the Altair One portal on Azure ND-H100-v5 and ND-H200-v5 series instances, users can follow these simple steps:

- Register for an Altair One account by visiting <https://admin.altairone.com/register>.
- If you don't have a Microsoft Azure account, you can obtain one by visiting <https://azure.microsoft.com/en-us/pricing/purchase-options/azure-account/>.

Once logged into Altair One, you can access a rich set of resources, including Altair Marketplace applications, documentation, and various educational and community support resources.

By default, when users run appliances and launch simulations in Altair One, workloads are automatically deployed to Microsoft Azure. To deploy simulation environments that use ND-H100-v5 or ND-H200-v5 series VMs, Altair One users with “appliance manager” credentials can easily create custom appliances and deploy them by providing their Azure account details. Step-by-step directions are provided in the [Altair One Documentation](#).

- From the Altair Appliances menu, select Altair Navops and [Create an Appliance](#).
- When creating appliances on Azure for the first time, you will be asked to [add your Microsoft Azure Cloud Account to Altair One](#). Provide your Azure Subscription ID, your Azure tenant ID (Active Directory tenant ID), as well as your Client ID and Secret Key.
- Using these steps, you can easily create custom appliances (i.e., My_CFD_Appliance), and select Azure ND-H100-v5 or ND-H200-v5 VMs (or other instance types) to run your GPU-accelerated Altair workloads.

By running simulations on Azure Cloud infrastructure via the Altair One portal, users avoid the hassle of installing application software, workload managers, remote access solutions, and infrastructure prerequisites such as MPI and CUDA, and CuDNN.

Conclusion

Running GPU-accelerated Altair One simulations and AI workloads on Microsoft Azure helps deliver both breakthrough results and outstanding business agility. Together, Altair and Microsoft can help manufacturers realize multiple benefits:

- **Superior time-to-solution** — Across 20 Altair CFD and Altair EDEM test cases, GPU-accelerated Azure VMs slashed runtimes by up to 5-6x versus a single GPU, achieving CFD simulation throughputs of over 7.5 billion node updates per second with ultraFluidX. Highly granular EDEM particle simulations achieved a speedup of up to 5.31x on eight H200 GPUs, compressing simulation runtimes from hours to minutes.
- **Right-sized economics** — Microsoft Azure ND-H100-v5 series VMs provide outstanding throughput and value for most simulation workloads, while the additional memory and bandwidth offered by the ND-H200-v5 series VMs provides a performance boost and headroom for oversized meshes and AI workloads.
- **Effortless scale, management, and governance** — Altair One and Microsoft Azure dramatically simplify CAE simulation in the cloud. With appliance templates and Navops scheduling, engineers can launch jobs with automated infrastructure provisioning in minutes, while IT retains full visibility into spending and resource governance.
- **A single gateway for simulation and AI** — From GPU-native solvers to PhysicsAI surrogate models, Altair One lets teams efficiently share the same Azure compute environments, data stores, and security controls across physics-based and AI-driven workflows, reducing cost and avoiding duplicated infrastructure.

By consolidating eight top-bin NVIDIA GPUs in ND v5 series VMs, Microsoft helps manufacturers achieve the same simulation throughput as dozens of CPU-based nodes, dramatically reducing costs and energy usage while freeing up valuable floor space in the data center.

Whether you're optimizing an EV cooling channel, predicting slurry flow, or training a PhysicsAI model, Altair One and Microsoft Azure give you the speed, scale, and flexibility to iterate faster and design better products—today and into the future.

To learn more about Altair One, visit altairone.com.

To learn more about Microsoft Azure, visit azure.microsoft.com.

Appendix – Benchmark Details

Altair ultraFluidX benchmarks

The runtimes shown below for the Altair ARC SUV and the Open Cooling DrivAer (OCDA) model are expressed in seconds. Single-GPU ARC results for the NVIDIA H100 and single-GPU OCDA results are not available because the models would not fit in GPU memory. Results were obtained with Altair ultraFluidX 2025.0.

Arc model	1 x GPU		2 x GPUs		4 x GPUs		6 x GPUs		8 x GPUs	
GPU type	H100	H200	H100	H200	H100	H200	H100	H200	H100	H200
pre-processing time (s)	-	5,000	4,633	4,482	3,762	3,486	3,190	3,084	3,412	3,353
output file initialization (s)	-	740	404	405	280	267	231	245	138	236
total computation time (s)	-	2,744	1,719	1,511	953	828	686	590	538	470
total elapsed time (s)	-	8,484	6,756	6,398	4,995	4,581	4,107	3,919	4,088	4,088
MNUPS (compute phase)	-	1,273.4	2,032.3	2,312.4	3,663.6	4,218.2	5,094.0	5,915.1	5,094.0	7,421.0
Speedup vs. 1 x H200	-	9	0	8	8	1	7	2	7	8
OCDA (DrivAer) model	1 x GPU		2 x GPUs		4 x GPUs		6 x GPUs		8 x GPUs	
GPU type	H100	H200	H100	H200	H100	H200	H100	H200	H100	H200
pre-processing time (s)	-	-	5,005	4,876	3,384	3,272	2,618	2,424	2,539	2,434
output file initialization (s)	-	-	432	443	307	299	256	243	203	229
total computation time (s)	-	-	3,783	3,352	2,026	1,806	1,454	1,306	1,128	1,019
total elapsed time (s)	-	-	9,220	8,671	5,717	5,377	4,328	3,973	3,870	3,682
MNUPS (compute phase)	-	-	2,023.8	2,284.4	3,778.7	4,240.0	5,264.6	5,860.4	6,786.0	7,514.7
	-	-	2	8	1	6	9	3	3	7

Altair nanoFluidX benchmarks

The nanoFluidX results were obtained using standard Altair benchmarks involving four different models. The numbers 1,000 and 10,000 refer to the number of timesteps in the simulation. The "dx" figures refer to the spatial resolution — dx0001 refers to 0.0001 meters (0.1 mm) and dx0002 refers to 0.0002 meters (0.2 mm). For some models, additional comparisons were made using different simulation parameters for specific GPU counts to provide additional points of comparison. These further results are shown in *italics* below and are not included in the summary benchmark tabulations.

Model / GPU(s)	ND-H100-v5				1 x H200	ND-H200-v5		
	1 x H100	2 x H100s	4 x H100s	8 x H100s		2 x H200s	4 x H200s	8 x H200s
Wall clock time (seconds)								
aero gbx new 1000	70.63	37.97	20.51	11.72	70.23	37.43	20.31	11.53
aero gbx new 10000	620.23	332.14	181.09	105.19	616.66	326.61	177.79	103.62
aero gbx new full				8,529.98				8,376.78
altair egbx 1000	24.03	13.12	7.49	5.62	23.73	12.93	7.39	5.39
altair egbx 10000	258.46	142.43	81.80	61.90	255.61	140.13	80.81	59.41
aero egbx full			12,278.31				12,034.99	
cuboid 192x192x192 1000	14.13	7.69	4.26	2.58	13.99	7.55	4.20	2.56
cuboid 198x198x198 1000	15.43				15.25			
cuboid 384x384x384 1000	108.28	55.52	28.46	14.88	107.37	54.76	28.01	14.67
dambreak dx0001 1000	106.41	54.43	28.00	14.82	105.57	53.67	27.62	14.59
dambreak dx0001 10000	1,131.46	579.23	298.58	161.63	1122.95	571.66	294.27	159.68
dambreak dx0001 full				1,649.70				1,628.92
dambreak dx0001 tol5 full				1,640.06				1,614.66

Model / GPU(s)	1 x H100	ND-H100-v5			1 x H200	ND-H200-v5		
		2 x H100s	4 x H100s	8 x H100s		2 x H200s	4 x H200s	8 x H200s
dambreak dx0002 1000	16.00	8.90	5.18	3.73	15.72	8.70	5.14	3.58
dambreak dx0002 10000	169.56	94.52	54.75	36.31	171.27	92.58	53.77	35.62
<i>dambreak dx0002 full</i>			<i>254.04</i>				<i>250.25</i>	
<i>dambreak dx0002 tol5 full</i>			<i>251.42</i>				<i>246.51</i>	

Performance — The table below shows the results for each of the ten nanoFluidX benchmarks expressed as average compute time in seconds per particle per timestep (s/prtl/it).

Model / GPU(s)	ND-H100-v5				ND-H200-v5			
	1 x H100	2 x H100	4 x H100	8 x H100	1 x H200	2 x H200	4 x H200	8 x H200
<i>Average time in seconds per particle per timestep (s/prtl/it)</i>								
aero gbx new 1000	2.64E-09	1.42E-09	7.68E-10	4.39E-10	2.63E-09	1.40E-09	7.61E-10	4.32E-10
aero gbx new 10000	2.32E-09	1.24E-09	6.78E-10	3.94E-10	2.31E-09	1.22E-09	6.66E-10	3.88E-10
<i>aero gbx new full</i>				<i>4.21E-10</i>				<i>4.14E-10</i>
altair egbx 1000	1.99E-09	1.08E-09	6.19E-10	4.65E-10	1.96E-09	1.07E-09	6.11E-10	4.46E-10
altair egbx 10000	2.13E-09	1.18E-09	6.76E-10	5.11E-10	2.11E-09	1.16E-09	6.67E-10	4.91E-10
<i>aero egbx full</i>			<i>1.06E-09</i>				<i>1.04E-09</i>	
cuboid 192x192x192 1000	2.00E-09	1.09E-09	6.03E-10	3.66E-10	1.98E-09	1.07E-09	5.94E-10	3.63E-10
<i>cuboid 198x198x198 1000</i>	<i>1.99E-09</i>				<i>1.97E-09</i>			
cuboid 384x384x384 1000	1.91E-09	9.81E-10	5.03E-10	2.63E-10	1.90E-09	9.67E-10	4.95E-10	2.59E-10
dambreak dx0001 1000	1.66E-09	8.47E-10	4.36E-10	2.31E-10	1.64E-09	8.35E-10	4.30E-10	2.27E-10
dambreak dx0001 10000	1.76E-09	9.01E-10	4.65E-10	2.52E-10	1.75E-09	8.90E-10	4.58E-10	2.49E-10
<i>dambreak dx0001 full</i>				<i>2.92E-10</i>				<i>2.88E-10</i>
<i>dambreak dx0001 tol5 full</i>				<i>2.90E-10</i>				<i>2.86E-10</i>
dambreak dx0002 1000	1.72E-09	9.54E-10	5.56E-10	4.00E-10	1.69E-09	9.33E-10	5.51E-10	3.84E-10
dambreak dx0002 10000	1.82E-09	1.01E-09	5.87E-10	3.89E-10	1.83E-09	9.92E-10	5.76E-10	3.82E-10
<i>dambreak dx0002 full</i>			<i>6.19E-10</i>				<i>6.09E-10</i>	
<i>dambreak dx0002 tol5 full</i>			<i>6.12E-10</i>				<i>6.00E-10</i>	

Throughput — The metric in the table below is calculated as $1 / (\text{the average time per particle per timestep})$ from the table above, yielding particle timestep updates per second. Only the tests with results for one, two, four, and eight GPUs are shown here:

	ND-H100-v5				ND-H200-v5			
	1 x H100	2 x H100	4 x H100	8 x H100	1 x H200	2 x H200	4 x H200	8 x H200
Throughput – particle-timestep updates per second								
aero gbx new 1000	378.11	703.33	1,301.53	2,277.26	380.26	713.39	1,314.25	2,314.70
aero gbx new 10000	430.64	804.14	1,474.86	2,538.84	433.13	817.76	1,502.23	2,577.37
altair egbx 1000	503.75	922.06	1,614.44	2,150.77	510.11	935.59	1,636.24	2,243.24
altair egbx 10000	468.50	850.12	1,480.16	1,956.07	473.72	864.07	1,498.20	2,037.74
cuboid 192x192x192 1000	500.59	919.44	1,659.15	2,729.88	505.42	935.73	1,682.49	2,753.48
cuboid 384x384x384 1000	522.86	1,019.65	1,988.66	3,803.73	527.30	1,033.89	2,020.86	3,856.49
dambreak dx0001 1000	603.81	1,180.34	2,293.73	4,331.74	608.64	1,196.94	2,325.91	4,401.08
dambreak dx0001 10000	567.91	1,109.33	2,152.03	3,975.31	572.22	1,124.03	2,183.54	4,023.93
dambreak dx0002 1000	582.89	1,048.05	1,799.04	2,497.83	593.27	1,071.56	1,814.30	2,602.84
dambreak dx0002 10000	550.51	987.55	1,704.68	2,570.40	545.00	1,008.21	1,735.82	2,620.24

	ND-H100-v5				ND-H200-v5			
	1 x H100	2 x H100	4 x H100	8 x H100	1 x H200	2 x H200	4 x H200	8 x H200
Average relative throughput (compared to 1 x H100)	1.00	1.87	3.41	5.62	1.01	1.90	3.46	5.74

Altair nanoFluidX water-wading benchmarks

The water wading benchmark represents an Altair CX-1 car model traveling at 10 km/h through a 24-meter wading channel simulated for 15 seconds. Results were obtained with 2, 4 and 8 GPUs on each cloud VM. The dual H100 configuration was not run, because the model was too large to fit in 160 GB (2 x 80 GB).

VM / Wall clock time (s)	2x GPUs	4x GPUs	8x GPUs
Standard_ND96isr_H100_v5	-	6,887.35	3,663.46
Standard_ND96isr_H200_v5	12,868.51	6,797.58	3,604.63

Altair EDEM benchmarks

A standard set of Altair EDEM benchmarks was run on both the Standard_ND96isr_H100_v5 and Standard_ND96isr_H200_v5 VMs with 1, 2, 4, and 8 GPUs. Parameters for each simulation are shown below:

Name	Angle of Repose	Bed of Material	Hopper Discharge	Powder Mixer	Screw Auger	Mill	Transfer Chute
Description	Cylinder angle of repose	A bed of material with a tillage tool	Hopper emptying into a container	Mixing of bulk solids	Transporting materials	Bulk material inside a mill	Simulate the behavior of a transfer chute
Particle Radii (m)	0.0005 - 0.001	0.002 - 0.004	0.003	0.0005	0.001	0.005	0.0045 - 0.009
nSpheres	3	3	3	1	1	1	3
Size Distribution	Random	Random	Fixed	Fixed	Fixed	Fixed	Random
nParticles	5,000,000	5,000,000	5,000,000	5,000,000	5,000,000	5,000,000	5,000,000
Physics	Hertz-Mindlin	Hertz-Mindlin +JKR	Hertz-Mindlin	Hertz-Mindlin	Hertz-Mindlin	Hertz-Mindlin	Hertz-Mindlin +JKR
Timestep (s)	5.73E -06	6.22E -05	4.00E -05	9.20E -06	1.40E -05	0.00016	5.97E -05
Total Time (s)	0.5	1	1	1	1	1	1
Save Interval (s)	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Grid cell size (x Rmin)	3	3	3	3	3	3	5

The results show elapsed wall clock time in seconds for each cloud instance with varying numbers of GPUs.

Benchmark / wall clock time (s)	1 x H100	2 x H100	4 x H100	8 x H100	1 x H200	2 x H200	4 x H200	8 x H200
Angle of Repose (5M)	2,558.01	1,398.87	798.52	515.00	2,319.58	1,280.87	740.89	498.76
Bed of Material (5M)	513.83	309.56	237.35	127.46	479.36	279.94	185.00	135.68
Hopper Discharge (5M)	338.52	211.67	131.01	106.81	315.87	189.87	130.00	115.88
Powder Mixer (5M)	2,676.90	1,470.72	832.55	534.26	2,676.90	1,319.94	770.08	504.49
Screw Auger (5M)	1,793.89	990.29	574.10	406.98	1,628.07	914.29	540.41	386.94
Mill (5M)	247.39	159.03	100.84	79.28	224.78	149.43	112.25	91.52
Transfer Chute (5M)	443.82	262.67	185.80	169.17	421.66	258.84	179.08	140.69

The table below shows the same results expressed as relative throughput compared to a single NVIDIA H100 GPU on the Standard_ND96isr_H100_v5 instance (throughput = 1 / wall clock time).

Benchmark / relative speed-up vs. single NVIDIA H100 GPU	1 x H100	2 x H100	4 x H100	8 x H100	1 x H200	2 x H200	4 xH200	8 x H200
Angle of Repose (5M)	1.00	1.83	3.20	4.97	1.10	2.00	3.45	5.13
Bed of Material (5M)	1.00	1.66	2.16	4.03	1.07	1.84	2.78	3.79
Hopper Discharge (5M)	1.00	1.60	2.58	3.17	1.07	1.78	2.60	2.92
Powder Mixer (5M)	1.00	1.82	3.22	5.01	1.00	2.03	3.48	5.31
Screw Auger (5M)	1.00	1.81	3.12	4.41	1.10	1.96	3.32	4.64
Mill (5M)	1.00	1.56	2.45	3.12	1.10	1.66	2.20	2.70
Transfer Chute (5M)	1.00	1.69	2.39	2.62	1.05	1.71	2.48	3.15
Average relative throughput (compared to 1 x H100)	1.00	1.71	2.73	3.90	1.07	1.85	2.90	3.95