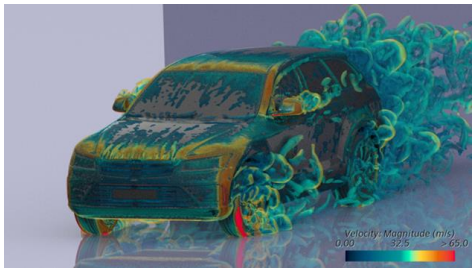
SOLUTION BRIEF

SUPERMICRO AND NVIDIA CREATE SOLUTIONS TO ACCELERATE CFD SIMULATIONS FOR AUTOMOTIVE AND AEROSPACE INDUSTRIES

Supermicro Servers with NVIDIA Datacenter GPUs Delivers Significant Speedup for CFD Simulations, Reducing Time To Market for Manufacturing Enterprises



AeroSUV (image courtesy of Siemens)

TABLE OF CONTENTS

Executive Summary

As automotive companies strive to reduce vehicle development time and bring new vehicles to market faster, there has been a growing need for computer-aided engineering for automotive development. One recent advancement in high performance computing (HPC) has been leveraging GPUs for computational fluid dynamics (CFD) to significantly speed up calculation time while reducing carbon footprint and cost.

Supermicro and NVIDIA have partnered to develop a reference architecture for HPC applications with optimum hardware configurations. In this Solution Brief, we will review the key requirements to get the most out of GPU acceleration for Computational Fluid Dynamics (CFD).

Improving Vehicle Design Through Simulation

The need to improve vehicle efficiency, cost, and comfort are critical targets in the automotive industry. With the move to hybrid and electric vehicles, improving vehicle range has become particularly important. Given that the range is directly proportional to the drag, it's critical to understand and reduce it early in the vehicle design. Computer Aided Engineering (CAE) enables engineers to quickly assess changes without the need to build expensive and time-consuming prototypes. Computational Fluid Dynamics (CFD) tools such as Siemens Simcenter STAR-CCM+ allow engineering professionals to understand aerodynamic performance promptly. A standard automotive CFD benchmark case is used to assess GPU performance

The New Era of Simulation Using GPUs

In a 2021 survey of vehicle engineers, 52% of respondents expressed that their demands to reduce the time required to complete design cycles have been one of the most challenging aspects of their jobs. To address this growing need for acceleration, over the past 15 years, NVIDIA has worked closely with simulation software leaders to bring the benefits of GPUs' massive parallel processing power and the lightning-fast interface between memory and GPU cores to vehicle simulation. By running vehicle simulation applications on GPUs, leading simulation software providers have demonstrated up to 30X acceleration of their applications, thereby improving engineer productivity and helping reduce vehicle time to market.

Doing More with Less - Improving Capital Expenditures for Simulation

Increased computational efficiency means more than workload acceleration, but just as importantly, it must include a smaller capital expenditure footprint. The incredible efficiency of GPUs for computer-aided engineering results in huge infrastructure cost savings for automotive customers. Because modern servers can fit up to 10 dual slot GPUs, offloading computationally heavy workloads to GPUs allows customers to achieve large amounts of server consolidation. Through the joint effort between Supermicro and NVIDIA, we have demonstrated that a single server with 10 NVIDIA H100 GPUs replaces 30 servers, each with 2 x CPUs for CFD workloads such as Siemens Simcenter STAR-CCM+. For equivalent performance, this results in 3x cost savings

Accelerated Computing is Sustainable Computing

Although much attention is focused on reducing vehicle tailpipe emissions and EV power consumption, reducing carbon footprint along the entire vehicle design, testing, and manufacturing workflows is equally important. The third value proposition of Supermicro servers with NVIDIA GPUs for computer-aided engineering is the incredible savings in computing power necessary for these computationally heavy workloads. For equivalent performance, this results in a 5x power saving by running on NVIDIA GPUs, as shown in the performance results section.

Configuring the Optimal Server for your Workload

Supermicro's servers are ideal for CAE because of their versatility in the number of GPUs they support. NVIDIA recommends using 100 and 200 class products (NVIDIA A100, NVIDIA H100, NVIDIA H200, NVIDIA B100) for CAE applications because of their need for double precision (FP64) compute and high memory bandwidth.

To decide on the right number of GPUs for the system, GPU memory size is a critical factor. For CFD, the mesh size is often used as a measure of the size of a simulation. The total GPU memory used is also determined by the complexity of the physics used. As a rule of thumb, we can say that the memory required can be between 1.2-5x the size of the mesh, depending on the complexity of the physics. For example, for a relatively simple case in terms of physics, we'd need 150 GB of memory for a 100

million cell mesh (100x1.5). Given that A100 and H100 have 80GB of memory, they need at least two GPUs, but 3 GPUs would be recommended. Software vendors can give additional guidance on the memory requirements for their software.

Application Support

One of the key questions is whether your CFD tool of choice supports GPU acceleration. The last few years have seen a proliferation of GPU solver support from both mainstream CAE software and new start-ups. Check with your software vendor to see if the physics you intend to simulate is GPU accelerated. Much of the solver runs on the GPU. By running the solver on highly parallel GPUs is the reason for very good acceleration compared to a few years ago when CFD codes were only accelerating part of the solver.

Supermicro Solution - The Supermicro SYS-521GE-TNRT

Basics of the SYS-521GE-TNRT

GPU	10 x NVIDIA H100 GPUs 80 GB (PCIe)
CPU	2x 4th Gen Intel® Xeon® CPU Max Processors (9468)
Memory Installed	1 TB
Operating System	Rocky Linux 8.6
NVIDIA Driver	535.104.12
CUDA Version	12.2

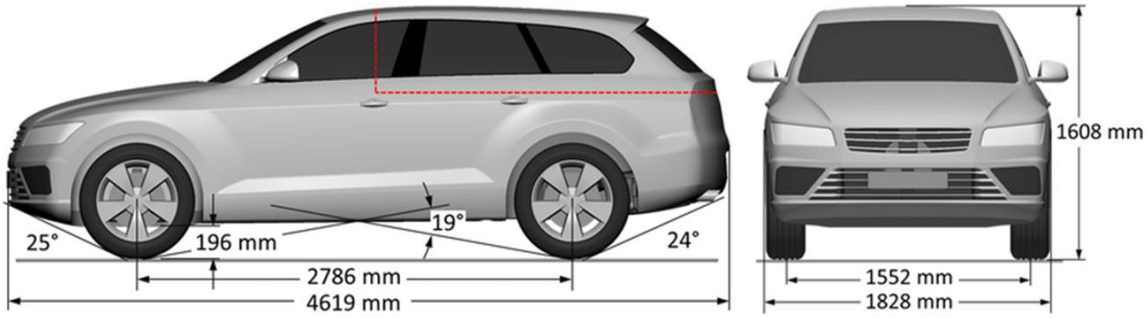
Supermicro SYS-521GE-TNRT Server (2 CPUs, 10 x NVIDIA H100 PCIe GPUs)

The Supermicro SYS-521GE-TNRT server is designed to accommodate up to 10 PCIe NVIDIA H100 GPUs. The CPUs can be either 4th or 5th Gen Intel Xeon processors and can hold up to 8 TB of memory.

Test Case Details

- Standard AeroSUV external aerodynamics test case (AeroSUV by FKFS, Stuttgart, Germany https://www.sae.org/publications/technical-papers/content/2019-01-0646/)
- Mesh size - Server: 57, 106, 320 million cells

- Mixed precision
- Siemens Simcenter STAR-CCM+ : Version 23.06
- Coupled solver



Tanneberger, Max & Zhang, Chenyi & Kuthada, Timo & Wittmeier, Felix & Wiedemann, Jochen & Nies, Juliane. (2019). Development of an SUV reference model for aerodynamic research. 10.1007/978-3-658-25939-6_37.

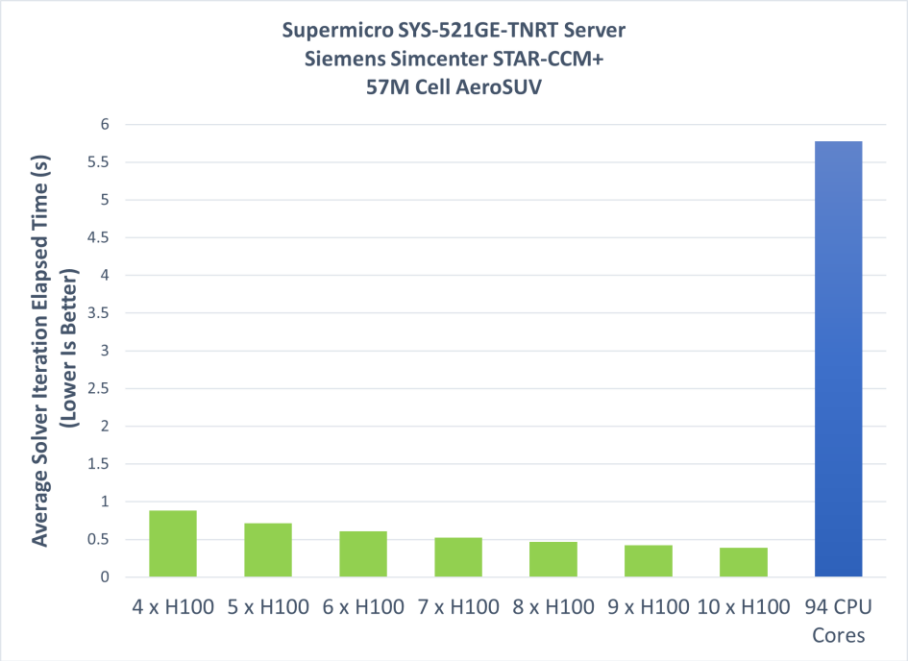
Performance Results

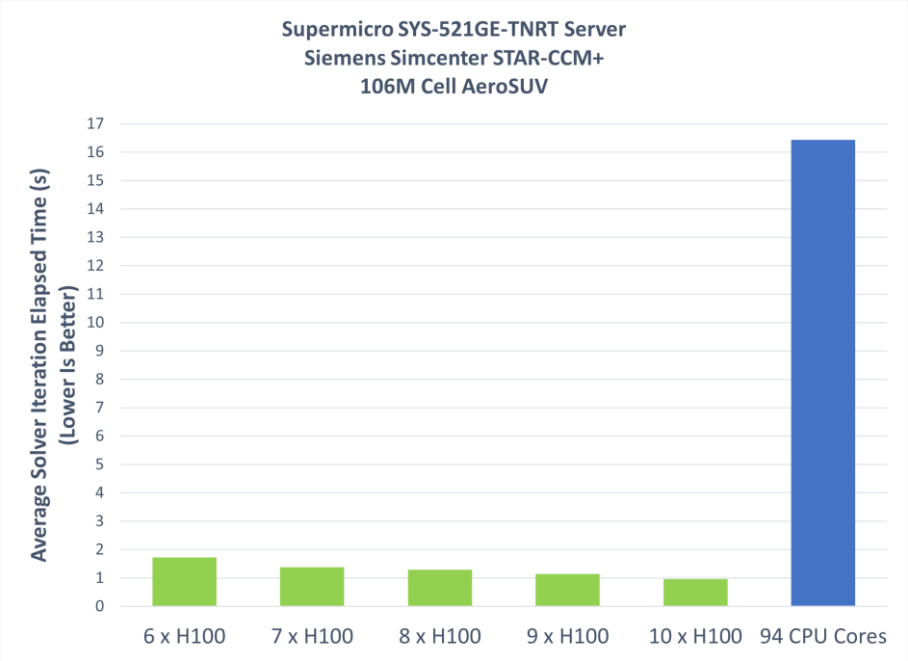
The chart below shows results for running the same AeroSUV test case of various mesh sizes on different numbers of H100 GPU on the server configuration. Note the 57 million cell case can be run on less than 4 GPUs if required.

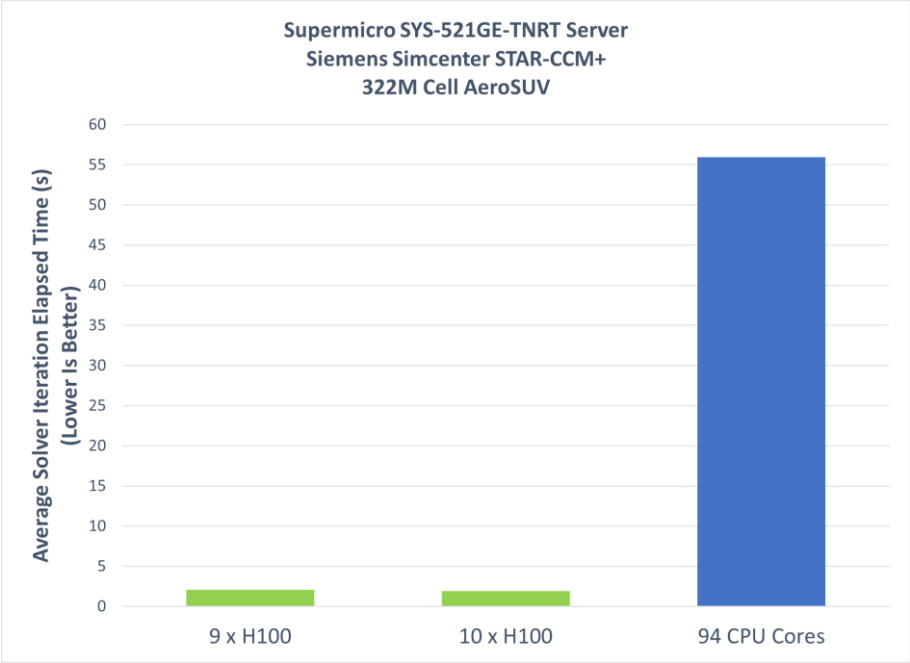
Key Results:

- For the smallest mesh size (57 million cells) with 4 NVIDIA H100 PCIe GPUs are 6.6x faster than 94 CPU cores
- For the medium mesh size (106 million cells) with 6 NVDIDIA H100 PCIe GPUs are 9.5x faster than 94 CPU cores
- For the largest mesh size (322 million cells) with 10 NVIDIA H100 PCIe GPUs are 29.3x faster than 94 CPU cores

The results above demonstrate the significant improvements in turnaround time that can be achieved by running CFD simulations on NVIDIA GPUs leveraging Supermicro hardware. Using GPUs allows more designs to be considered in less time, resulting in a more optimized product.







Price Performance and Sustainability

As well as comparing pure performance across one server, it is essential to consider the overall price performance. Assuming a GPU server (8-10 H100 GPUs) and a speed-up of ~30x compared to a CPU server, customers would have to invest in 30 CPU servers to match the performance of one GPU server. These results assume linear scaling. Therefore, running on GPU, we could estimate it would be 3x cheaper than CPU for equivalent performance. Alternatively, spending the same budget as the CPU on the GPU provides a 3x improvement in throughput.

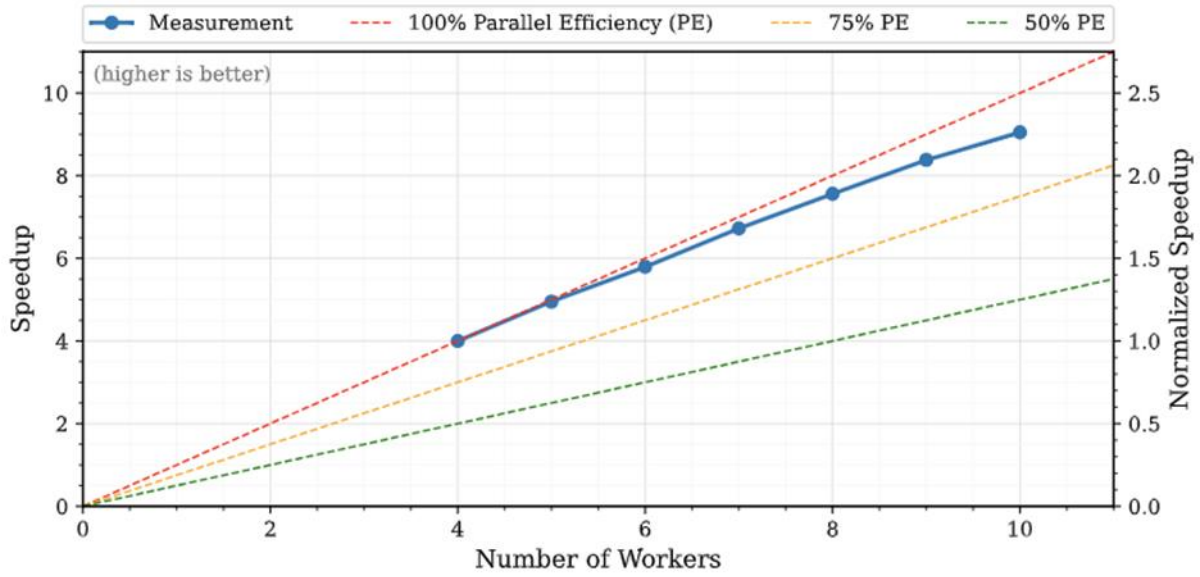
Similar cost savings can also be seen in the power required for the same performance. Estimating from the thermal design power, a GPU node would require approximately 3000W compared to a CPU node ~500W per CPU node, given the need for 30 servers to match GPU performance. This represents a 5x power saving.

Optimum Performance - Scaling

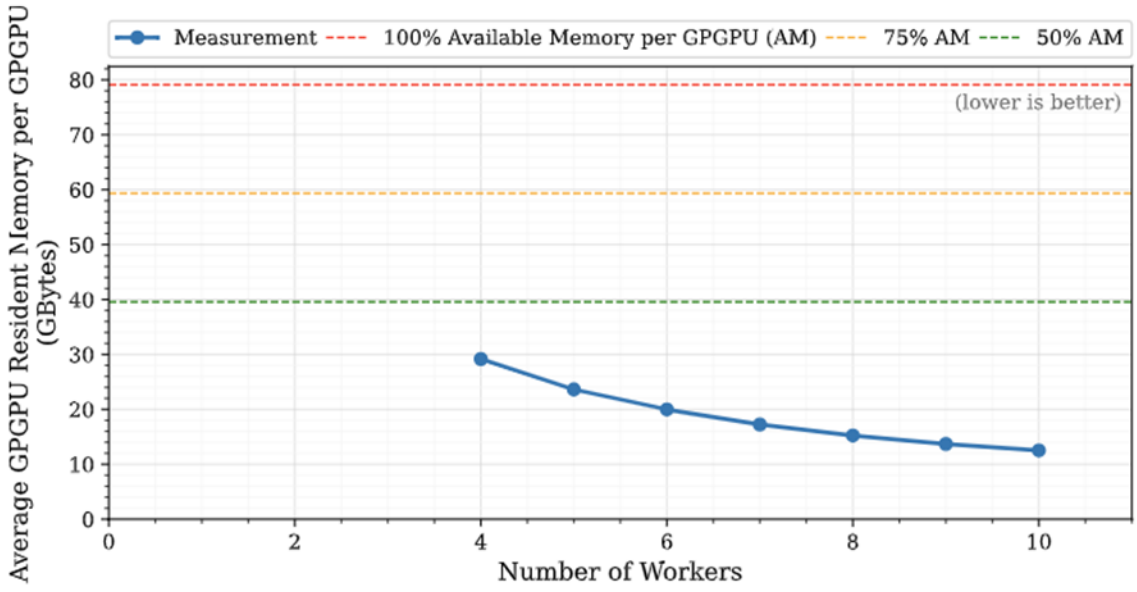
In the previous sections, we discussed the minimum number of GPUs required for running the simulation, but what is the optimum number? Below are charts that show the speed-up and the memory usage for the 57 million cell aerodynamics case. We can see that performance scales up to around 8 GPUs. Suppose we look at the memory usage per GPU. Below 6 GPUs that memory usage is below 20GB per GPU, so less than 25% of the GPUs' available memory. The conclusion is that scaling will suffer if running a case only occupies a small percentage of GPU memory. Therefore, having at least 25% of the GPU memory occupied is good for maintaining reasonable scaling.

Speedup

Speedup



Average GPGPU Resident Memory per GPGPU



Summary

The results presented above show that GPU acceleration offers a significant opportunity to reduce CFD simulations' turnaround time and save on hardware costs. Power savings are also considerable and becoming more critical as companies look to implement more sustainable solutions. Supermicro provides a range of efficient solutions ranging from deskside workstations to 10 GPU systems for maximum performance. To find out more about the benefits of CAE for your automotive workflows, please contact Supermicro. NVIDIA and Supermicro jointly work together to help our customers develop the most optimal solution for their CAE workloads.

For More Information

Supermicro NVIDIA Accelerated Systems – https://www.supermicro.com/accelerators/nvidia

Supermicro 10x GPU PCIe Server - https://www.supermicro.com/en/products/system/gpu/5u/sys-521ge-tnrt

SUPERMICRO

As a global leader in high performance, high efficiency server technology and innovation, we develop and provide end-to-end green computing solutions to the data center, cloud computing, enterprise IT, big data, HPC, and embedded markets. Our Building Block Solutions® approach allows us to provide a broad range of SKUs, and enables us to build and deliver application-optimized solutions based upon your requirements. See www.supermicro.com

NVIDIA

Since its founding in 1993, NVIDIA (NASDAQ: NVDA) has been a pioneer in accelerated computing. The company's invention of the GPU in 1999 sparked the growth of the PC gaming market, redefined computer graphics, ignited the era of modern AI and is fueling industrial digitalization across markets. NVIDIA is now a full-stack computing company with data-center-scale offerings that are reshaping industry. More information at https://nvidianews.nvidia.com.