ALTAIR

SUSTAINABLE COMPUTING FOR HPC AND AI

ENVIRONMENTALLY FRIENDLY COMPUTING AT ANY SCALE

INTRODUCTION

Concerns about global warming and unsustainable levels of carbon dioxide (CO2) emissions loom large for organizations of all sizes. Manufacturers are under more pressure than ever to design energy-efficient products and move towards net zero emissions to meet corporate environmental, sustainability, and governance (ESG) goals. Virtually all organizations that operate large-scale data centers have similar concerns, from financial services firms to government agencies to pharmaceutical companies to online service providers. While high-performance computing (HPC) and artificial intelligence (AI) play key roles in designing and delivering more energy-efficient products, ironically, the widespread adoption of AI – with its enormous appetite for electricity – is driving dramatic increases in energy demand across all industries.

According to research from the International Energy Agency (IEA), global energy consumption due to data centers, AI, and the cryptocurrency sector is poised to double from an estimated 460 terawatt-hours (TWh) in 2022 to more than 1,000 TWh by 2026 — an amount roughly equal to Japan's total annual electricity consumption.[1]

Fortunately, technologies pioneered in HPC can play a critical role in improving the efficiency of energy-intensive AI model training and inference (prediction). By leveraging these technologies and taking other measures, such as deploying more energy-efficient servers or improving data center PUEs[2], organizations can dramatically reduce their carbon footprint and help realize a greener, more sustainable future. In addition to providing environmental benefits, more energy-efficient data centers help organizations reduce costs and increase profitability.

In this guide, we discuss the challenge of global warming, the promise of HPC, and the considerable challenges of curbing AI's enormous appetite for energy. We also suggest some strategies and best practices that can help organizations increase efficiency and reduce emissions while simultaneously boosting productivity and profitability.

Global electricity consumption from data centers, AI, and the cryptocurrency sector is poised to double by 2026.

IEA, Electricity 2024, IEA, Paris — https://www.iea.org/reports/electricity-2024

Table of Contents

SUSTAINABILITY

A Critical Concern

Sustainability has emerged as a key requirement in modern design and manufacturing environments. Faced with longer heatwaves, more wildfires, increasingly intense storms, and growing threats to agriculture and the global food supply, 193 countries and the European Union signed the Paris Agreement, pledging to pursue efforts to limit global temperature rise to 1.5 degrees C above those recorded in pre-industrial times. The goal is to avoid the catastrophic projected impacts of climate change.

Driven by consumer demand and government regulation, most organizations have made public commitments to meet carbon reduction goals as part of corporate ESG programs, and they're incorporating these requirements into their supply chains. Suppliers are increasingly required to disclose greenhouse gas emissions and climate risks in public and private procurements. Common strategies to meet sustainability goals include reducing scope 1 and scope 2 carbon emissions from operations, investing in renewable energy, responsible sourcing and recycling, and purchasing carbon credits to offset emissions.

A Two-Pronged Approach to Reducing Greenhouse Gas Emissions

Manufacturers play an outsized role in enabling a low-carbon future because the lifetime environmental impact of their products tends to dwarf the impact of product design and manufacturing over the product life cycle. Automobiles are a good example. According to the IEA, the total carbon impact of manufacturing a typical internal combustion engine vehicle is approximately six tons of CO2 equivalent (tCO2e). By contrast, operating the car over its lifetime results in approximately 35.9 tCO2e. In other words, roughly 86% of life cycle emissions are due to the fuel cycle and tailpipe emissions.

To meet sustainability goals, manufacturers face two critical challenges:

1. Designing more sustainable, energy-efficient products that exhibit a lower lifetime carbon footprint.
2. Improving the energy efficiency and emissions associated with internal operations, including design, manufacturing, and data center resources.

Over its lifetime, a gasoline-powered automobile has a carbon footprint of approximately 41.9 tCO_2e.

IEA — Comparative life-cycle greenhouse gas emissions of a mid-size BEV and ICE vehicle

HPC AND SIMULATION

Reducing Life Cycle Carbon Emissions

HPC and simulation are essential in designing more sustainable, environmentally friendly products. By employing advanced computer simulation and optimization, organizations can:

- Design lighter, more fuel-efficient products
- Reduce material usage and waste while meeting product performance goals
- Use more sustainable processes and recycled materials
- Employ software-based crash simulations to avoid wasteful physical testing
- Improve product durability and life span to reduce waste
- Reduce transport and packaging costs
- Leverage generative AI to create more energy-efficient designs

The Altair Enlighten Award honors the greatest sustainability and lightweighting advancements that successfully reduce carbon footprint, mitigate water and energy consumption, and leverage material reuse and recycling efforts.

The Enlighten Award attracts interest from industry, engineering, policymakers, educators, students, and the public. It showcases the latest and greatest technology innovations dedicated to sustainability and highlights the critical role played by advanced design tools, software simulation, and HPC.

> "Advanced simulations run on HPC infrastructure have helped avoid millions of tons of CO2 emissions."

Dr. Rosemary Francis, Chief Scientist, HPC, Altair. HPC's Sustainability Challenge: The Elephant in the Room. https://altair.com/resource/hpc-sustainability-challenge

IMPROVING HPC EFFICIENCY

HPC Is Energy-Intensive

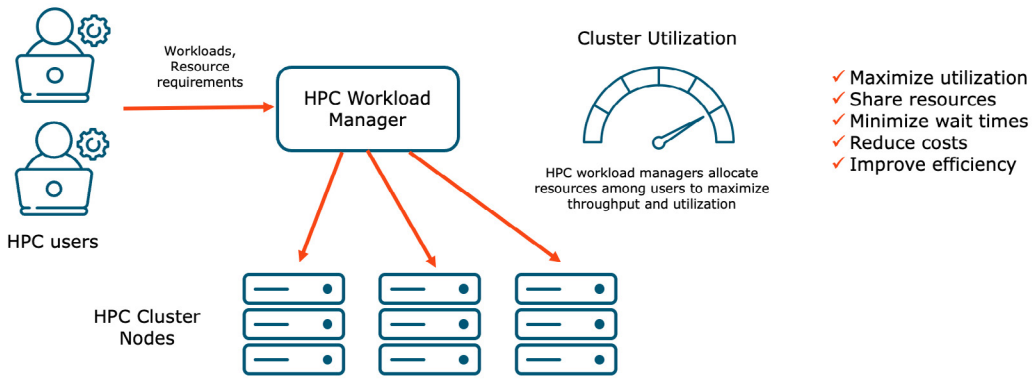
While large-scale simulation and data analytics are critical to designing and manufacturing sustainable products, HPC simulation also has a considerable carbon footprint.

Consider a single large-scale computational fluid dynamics (CFD) simulation conducted on a cluster with 524K CPU cores lasting 500 hours. While energy costs and carbon footprint will vary depending on the duration of the simulation, number of cluster nodes executing in parallel, resolution, energy efficiency of each node, and proportion of energy coming from renewable sources, the **carbon footprint of a large simulation can be as high as 106 kg (10,000 tons) of CO_2**[4]. To put this in perspective, this is roughly equivalent to flying a Boeing 777 from New York to Beijing and back. Such a simulation is also equivalent to the lifetime carbon emissions of roughly 240 gas-powered passenger vehicles. Other types of HPC simulation, such as vehicle crash tests, are similarly energy- and carbon-intensive.

Reducing HPC's Carbon Footprint

Today, most organizations use HPC workload management software to help improve efficiency and reduce the cost and carbon footprint associated with simulations. To reduce HPC's carbon footprint, organizations can:

- Use more efficient algorithms to reduce the cost and energy requirements for each simulation
- Employ energy-efficient graphics processing units (GPUs) where feasible to increase throughput per watt
- Leverage generative AI and AI-guided simulation to optimize parameter selection and reduce the total simulation space for faster, cheaper, less energy-intensive simulations
- Employ advanced workload managers to maximize resource utilization and shift workloads to the most energy-efficient resources for particular jobs



AI'S INSATIABLE APPETITE FOR ELECTRICITY

New Challenges for Sustainability

Advances in AI are transforming a broad variety of industries, from manufacturing to healthcare to retail to financial services. Even with powerful GPUs that deliver ~6x the model training performance per socket compared to CPUs alone, the power required to train large models is enormous.

What makes AI workloads particularly challenging is that, depending on the model, inference — using an AI model to analyze new data and make predictions — can be even more energy-intensive than model training. Google estimates that 60% of the cost and environmental impact of AI is from inference, due to the sheer number of times predictive models are invoked.

Some examples that highlight the enormous energy requirements and environmental impact of AI model training and inference:

Training a single generative AI model can take 150,000 GPU hours, equivalent to 11,250 kg CO2e. This is just for images — new AI-powered video generators are even more energy-intensive.

Generating 1,000 images using the Stable Diffusion XL text-to-image model uses the energy equivalent of driving 4.1 km in a gas-powered car. Generating a single image takes as much energy as charging a mobile phone.

Training a large language model (LLM) such as GPT-4 requires approximately 25,000 GPUs. For GPT-4's 100-day training period (and factoring in overhead), this adds up to ~60,000,000 GPU hours and consumes ~28,800,000 kWh. This translates into 6,912 tCO2e, or the equivalent of powering 1,300 homes for one year.

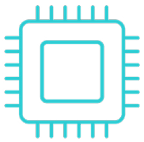
According to the IEA, a single ChatGPT query requires 2.9 watt-hours of electricity, compared with 0.3 watt-hours for a Google search — a ~10x increase in energy intensity.

If all the world's ~1 billion vehicles were suddenly self-driving, the cumulative inference workload would be approximately 21.6 million billion inferences per day — roughly 20,000X that of Facebook.

As AI features make their way into more products and services, the demand for data-center power is poised to explode. Data centers already account for 4% of U.S. electricity demand, a figure expected to grow to 10% by 2030.[14]

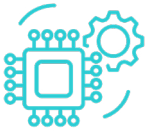
REDUCING YOUR HPC AND AI ENERGY FOOTPRINT

Fortunately, many of the same techniques pioneered in HPC can help organizations reduce the cost and environmental impact of their AI workloads.



GPU-aware scheduling
Given the enormous energy requirements of GPUs and accelerated processing units (APUs), maximizing their utilization with GPU-aware scheduling is critical to containing costs and reducing total power, cooling, and associated emissions.

Run mixed workloads
Consolidate AI, HPC, and cloud-native Kubernetes workloads onto a shared infrastructure with robust GPU and container support to improve resource usage and minimize energy consumption.



Optimize hardware selection
For HPC and AI, scheduling on-premises or cloud resources optimized for particular workloads is essential for reducing power requirements. For example, using specialty tensor processing units (TPUs) can boost performance per watt by 2-5x compared to general-purpose GPUs.[15]

Employ purpose-built models
Many predictive models are over-parameterized. For narrow tasks such as sentiment analysis or classification, purpose-built models can be far more energy-efficient than general-purpose LLMs, dramatically reducing training and inference costs.[16]

Cloud bursting
With policy-based cloud bursting, organizations can reduce scope 2 emissions by taking advantage of the energy efficiency and lower power usage effectiveness (PUE) ratios of carbon-neutral cloud data centers.

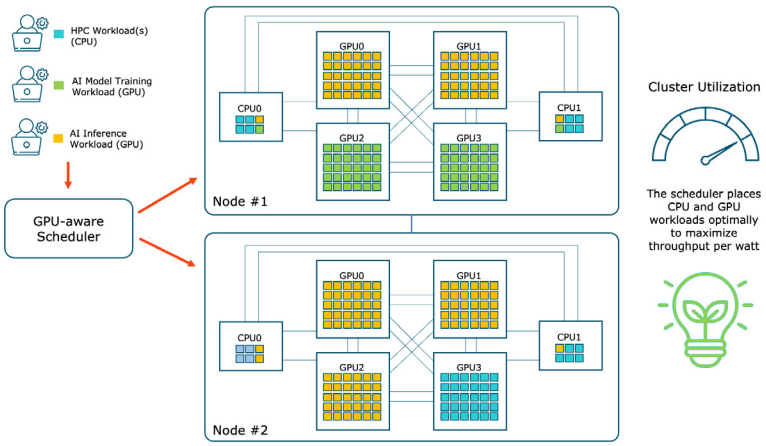
Energy-aware scheduling
Leverage schedulers that consider power requirements in scheduling decisions with power capping, per-job power profiles, and energy accounting to reduce total power consumption.

Allocate resources based on need
By prioritizing AI model training jobs and HPC simulations based on business needs, organizations can avoid redundant or unnecessary computations and reduce their overall costs and carbon footprint.

Deploy workload and resource monitoring and reporting
Use workload management together with monitoring solutions that can report on resource usage and power consumption to optimize efficiency and reduce energy usage over time.

GPU-Aware Scheduling Dramatically Improves Sustainability for AI Workloads

GPU and topology-aware scheduling optimally place components of distributed GPU workloads, considering resource requirements and underlying CPU, GPU, and memory architectures to optimize performance, avoid conflict, and make unused CPU cores available for other workloads. The result is that expensive and power-hungry GPU and CPU cores are fully utilized, and the CPU portion of workloads are pinned to cores in proximity to GPUs, resulting in lower latency, higher throughput, and higher throughput per watt.



In addition, for partitioned models that span multiple GPU-capable nodes, topology-aware scheduling considers underlying network architectures (where power considerations are also important), including intra-server GPU interconnects and InfiniBand or Ethernet links between servers, minimizing communication overhead and reducing the time and energy required to train a model.

Sharing expensive resources among HPC, model training, and inference workloads results in additional financial and energy savings. Mixing CPU—and GPU-intensive workloads improves utilization on large compute nodes, and time-critical inference jobs can be prioritized to minimize wait times.

With cloud bursting and energy-aware scheduling, workloads can automatically be directed to the most cost- and energy-efficient computing resources, on-premises or in the cloud.

Regardless of industry, GPU and topology-aware scheduling can help organizations dramatically reduce the carbon footprint of their AI workloads.

HOW ALTAIR CAN HELP

Whether HPC and AI workloads run on-premises or in public, private, or hybrid clouds, organizations need to optimize productivity and maximize resource utilization. Altair's HPC solutions can help organizations manage large-scale workloads effectively while minimizing energy usage and carbon emissions.

Augment AI-Powered Innovation with a Powerful HPC Backbone

The Altair® HPCWorks® HPC and cloud platform provides a rich set of tools to access, control, and optimize computing resources. It enables users to move seamlessly between on-premises and cloud environments and make better decisions with detailed monitoring and reporting data. Organizations can take advantage of convenient Jupyter Notebook integration, GPU acceleration, and rapid scaling to enable the latest analytics and AI workloads with flexible scheduling and workflow design.

- Workload managers, including Altair® PBS Professional® and Altair® Grid Engine®, provide extensive support for containerized GPU workloads and rich, topology-aware scheduling and cloud-bursting features to help organizations simplify administration and maximize infrastructure usage for a wide variety of HPC, AI, and analytics workloads.
- Altair® Access™ offers a simple, powerful, consistent interface for submitting and monitoring jobs on remote clusters, enabling data scientists and analysts to focus on their work and access the most energy-efficient resources for their workload requirements.
- For organizations that need an easy-to-use application for monitoring cluster configuration and reporting in HPC and AI environments, Altair® Control™ supplies a control center for managing, optimizing, and forecasting resources with advanced analytics to support data-driven decision-making.
- Altair® NavOps® lets you define intelligent, business- and workload-aware scaling automations to maximize resource utilization. NavOps reduces costs and cuts down on common problems such as the energy wasted bringing machines up and down too often, or the waste resulting from the wrong machines being scaled. NavOps works with Altair schedulers and cloud providers to dynamically scale on-demand cloud resources while providing detailed visibility and control over cloud spending.
- Altair Mistral™ provides live system telemetry and I/O monitoring for data-intensive distributed model training workloads, quickly pinpointing compute and storage-related bottlenecks to maximize on-premises and cloud resources in HPC and AI environments. Altair Breeze™ profiles application file I/O to optimize data handling and ensure data-hungry model training workloads run at peak efficiency and utilization.

Altair offers a rich portfolio of software and services to help optimize all facets of the AI and machine learning operations pipeline, from data collection to data preparation to feature engineering to model training.

CONCLUSION

Sustainability is a critical concern for organizations of all sizes across all industries. With the widespread adoption of energy-intensive AI applications in the enterprise, energy efficiency has become more important than ever.

Operating an efficient, sustainable computing environment isn't just good for the planet — it's good for the bottom line. Organizations that operate more efficiently can significantly reduce costs related to data center operations, power, and cooling and curb costs in the cloud.

By leveraging HPC management software in compute and data-intensive HPC and AI environments, organizations can:

- Optimize how workloads are deployed on expensive GPUs, APUs, and TPUs to maximize throughput and productivity while reducing total energy requirements
- Run mixed HPC and AI workloads (both CPU- and GPU-intensive) on shared infrastructure to increase utilization and reduce costs
- Shift workloads to energy-efficient cloud data centers, selecting optimal instance types based on considerations such as cost, performance, and throughput per watt
- Realize an infrastructure for more cost-efficient model training, leading to more energy-efficient, purpose-built models that help reduce the life cycle costs of AI in the enterprise

To learn more about HPC and cloud solutions that can help improve the efficiency and sustainability of HPC and AI environments, visit altair.com/altair-hpcworks.

To learn more about Altair software solutions for AI, visit altair.com/ai.

Altair is a global leader in computational intelligence that provides software and cloud solutions in simulation, high-performance computing (HPC), data analytics and AI. Altair enables organizations across all industries to compete more effectively and drive smarter decisions in an increasingly connected world – all while creating a greener, more sustainable future.

To learn more, please visit www.altair.com

REFERENCES

1. IEA (2024), Electricity 2024, Paris, Licence: CC BY 4.0
2. In data centers, PUE refers to Power Usage Effectiveness. PUE is determined by dividing the total power entering a data center by the power used to run IT equipment, with overall efficiency improving as the quotient decreases toward 1.0.
3. Credit: United Nations Framework Convention on Climate Change, The Paris Agreement
4. IEA, Comparative life-cycle greenhouse gas emissions of a mid-sized BEV and ICE vehicle.
5. See Computational Fluid Dynamics: its Carbon Footprint and Role in Carbon Emission Reduction, February 2024. License CC BY 4.0.
6. Assuming a passenger vehicle's lifetime carbon footprint is 41.9 tCO2e (based on IEA data), 10,000t / 41.9t per vehicle = 238.7 vehicles.
7. Comparative Analysis of CPU and GPU Profiling in Deep Learning Models.
8. Energy and Emissions of Machine Learning on Smartphones vs. the Cloud, Jan 2024.
9. HuggingFace – Environmental Impact of training Stable Diffusionv1.3.
10. Making an image with generative AI uses as much energy as charging your phone.
11. The Carbon Impact of Large Language Models: AI's Growing Environmental Cost.
12. Google's greenhouse gas emissions are soaring thanks to AI.
13. HPC's Sustainability Challenge: The Elephant in the Room. Dr. Rosemary Francis, Altair.
14. Power grab: Energy needed to power AI set to double worldwide by 2028.
15. See /dev/sustainability, Influencing the carbon emissions of AI, February 2023.
16. See State of the Planet, Columbia Climate School, AI's Growing Carbon Footprint, June 2023.