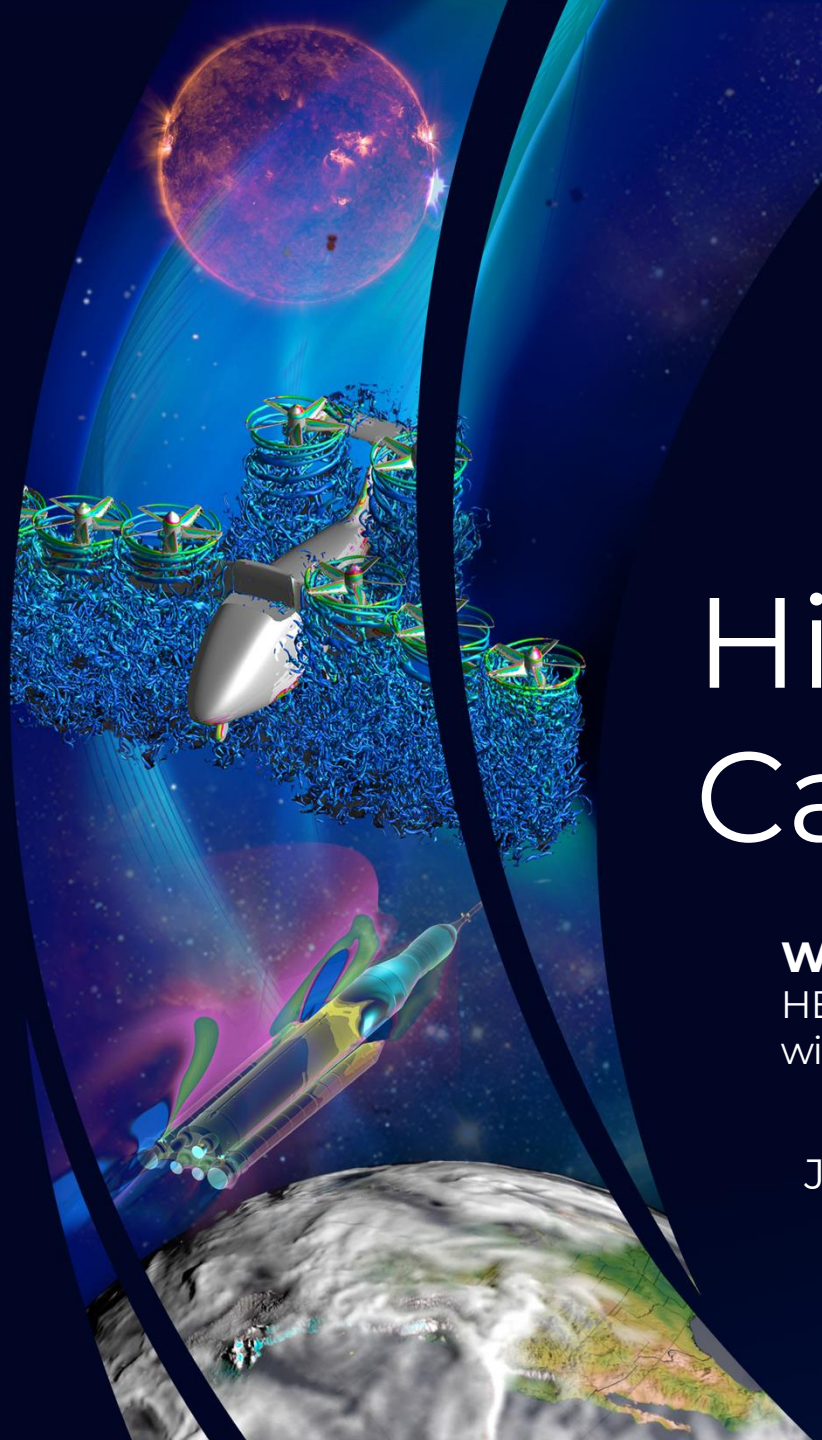


National Aeronautics and Space Administration

High-End Computing Capability

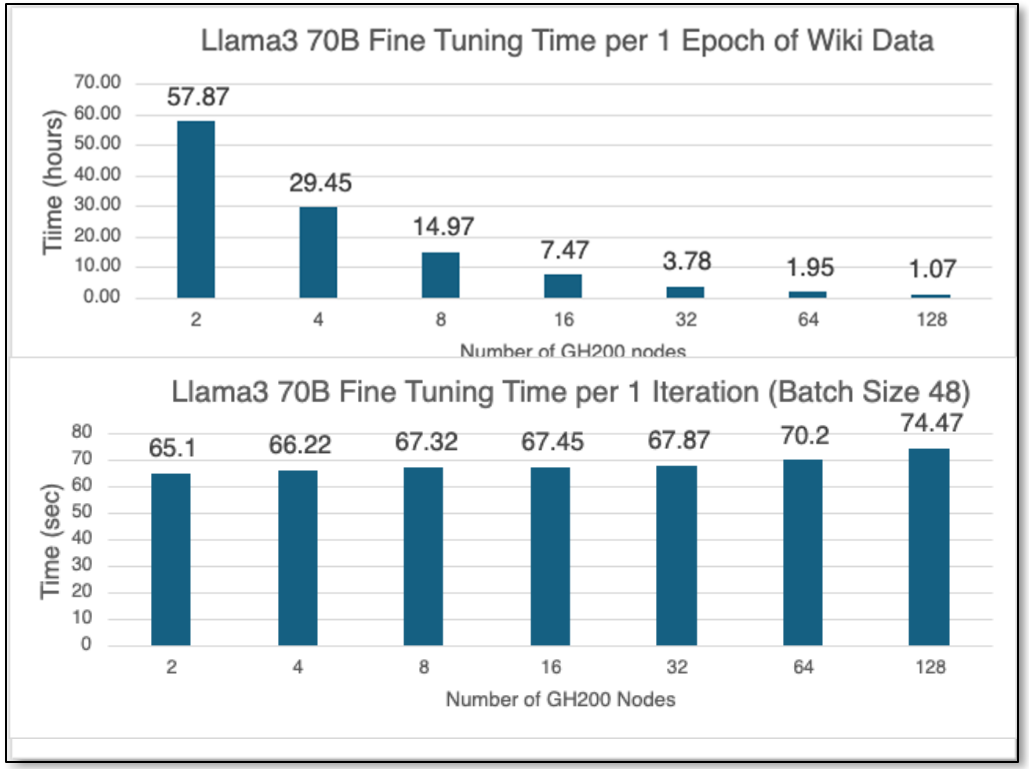
William Thigpen
HECC Component Manager
william.w.thigpen@nasa.gov

June 10, 2025

Fine Tuning LLMs with Grace Hopper Nodes

- Using HECC's Grace Hopper nodes allows for training large language models (LLMs) and foundation models at scale. Benefits include:
  - More efficient than training on A100 nodes due to larger GPU memory and unified memory architecture using NVIDIA NVLink Chip-2-Chip interconnect in a single Superchip, allowing for better data loading from CPU.
  - Fine-tuning using Quantized Low-Rank Adaptation (QLoRA) reduces memory usage and computational cost.
- Distributing the model using Fully Sharded Data Parallel (FSDP) allows for very large models like Llama3 70B (a state-of-the-art large language model designed for enhanced reasoning, coding, and broad application across multiple languages and tasks) to fit in memory by splitting the model across nodes for training.
  - The FSDP model allows Llama3 70B (~140 GB) to fit in GPU memory (96 GB) and use large batch sizes (48) while saturating the GPU memory.
- A near-linear time reduction was achieved by increasing the number of nodes for fine-tuning Llama3 70B, demonstrating efficient utilization of Grace Hopper nodes.
- The model is stored in the HECC data analytics team's Hugging Face repository and can be accessed and fine tuned by other users by saving the tuned weights, minimizing memory storage costs for LLMs.

IMPACT: Near linear scaling for training and fine-tuning of LLMs using HECC's Grace Hopper nodes allows for efficient development of NASA foundation models.

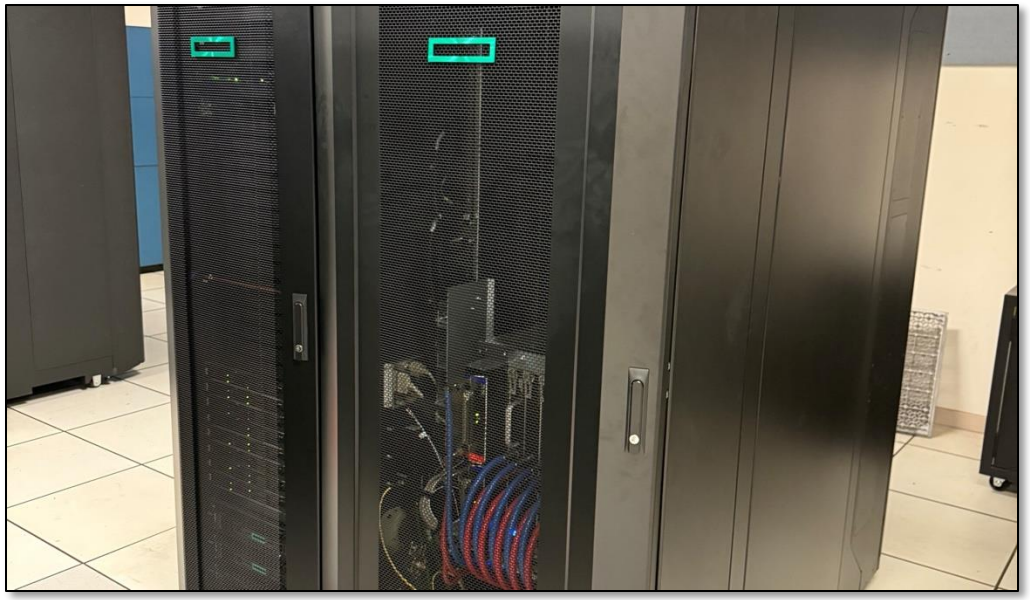


Llama3 70B tuning timings. *Jonathan Gee, NASA/Ames*

Athena Test and Development System Delivered

- Hewlett Packard Enterprise delivered and physically installed the Athena Test and Development System (TDS) on May 13. The TDS will be used to create and test the configuration, establish early benchmarks, and define operational processes for the production system scheduled to arrive in July.
- The system is configured with:
  - Three cluster management nodes in a high availability configuration.
  - Two fabric managers.
  - A PBS server.
  - One front-end node.
  - Four compute nodes (one blade).
    - Dual AMD Turin EPYC 9745 CPUS (128 cores, 2.4 gigahertz).
    - 768 gigabytes (GB) of Double Data Rate 5 (DDR5) 6400 memory.
    - Dual port 200 gigabit HPE Slingshot system.
- After Athena is deployed, the TDS system will provide a scaled-down representative system to test and validate changes prior to implementation on the production system.

IMPACT: The Athena Test and Development System will allow for pre-production development work to start, reducing the time required to deploy the Athena production system.

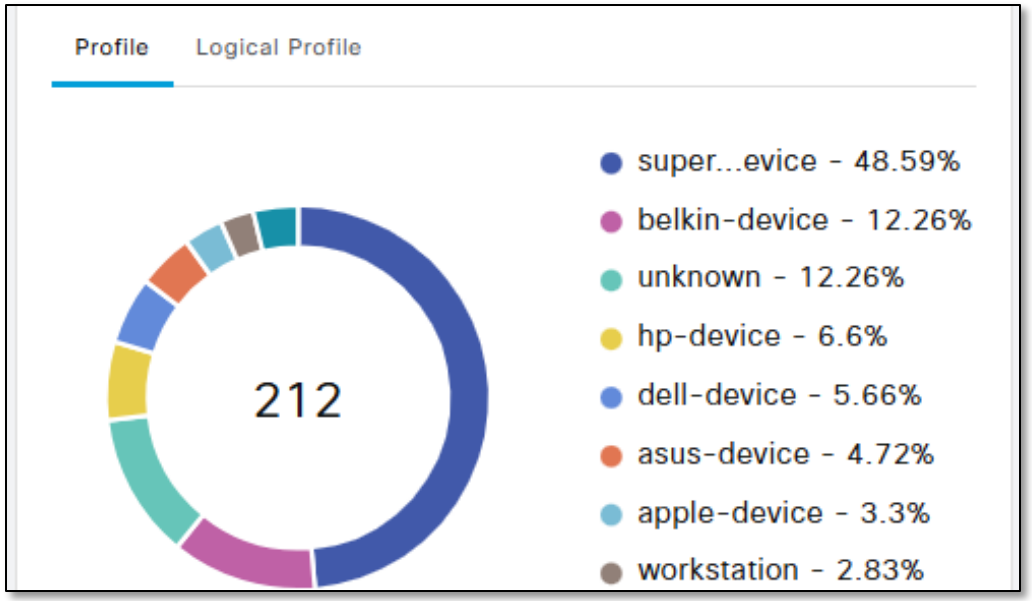


The Athena Test and Development System is housed in two cabinets. including service nodes and water-cooled compute nodes.
Davin Chan, NASA/Ames

HECC Enforces Network Access Control

- The HECC Networks team deployed Network Access Control (NAC) with MAC Address Bypassing (MAB) to meet a Department of Homeland Security requirement and improve network security.
- This shift from port security enables session logging for visibility and control of endpoint devices such as laptops, workstations, and printers.
- NAC enforcement benefits include:
  - **Centralized policy control:** Seamless access for devices like laptops, Linux workstations, and printers by authenticating using their MAC addresses.
  - **Provides dynamic authorization and profiling:** Ability to automatically detect device type via profiling and assign tailored access, enhancing security and minimizing manual configuration.
  - **Better security than port security:** Unlike static port security, NAC enforcement with MAC Address Bypass supports dynamic policies, device mobility, and real-time monitoring, reducing overhead and increasing network agility.
  - **Reduced administrative overhead:** Ability to view endpoint device connection status in real time and simplified onboarding of endpoint devices.

IMPACT: Network Access Control enforcement provides centralized policy control, improving network security and reducing administrative overhead for managing endpoint devices on the network.



This diagram depicts the devices that authenticated utilizing Network Access Control Enforcement.

Congressman Ro Khanna Visits NAS Facility

- Congressman Ro Khanna, the U.S. representative for California's 17th congressional district, visited the NASA Advanced Supercomputing facility's hyperwall Visualization Lab on May 12.
  - The HECC Visualization team showcased work from multiple NASA mission directorates, including visualizations of the Estimating Circulation and Climate of the Ocean (ECCO) model and computational fluid dynamics (CFD) simulations of the High-Lift Common Research Model (CRM).
  - The team also presented visualizations of experimental and CFD data from a recent Space Launch System (SLS) test, which was conducted in the Unitary Plan Wind Tunnel (UPWT) using the unsteady Pressure-Sensitive Paint (uPSP) technology and the NAS-UPWT Red Rover system.
  - The presentation concluded with a 5-gigapixel edge-on view of the Milky Way Galaxy, as imaged by the Spitzer Space Telescope.
- Following the presentation, the Congressman expressed strong support for the agency and for Ames Research Center.
- The 128-node, billion-pixel NASA Ames hyperwall, which is tightly coupled to the high-performance computing systems, enables researchers to interactively visualize, analyze, and explore their large high-dimensional datasets in new and impactful ways.

IMPACT: Communicating NASA's national impact to key stakeholders and representatives is critical for securing continued support for the agency's important work facilitated by HECC.



California Congressman Ro Khanna viewing a 5-gigapixel edge-on view of the Milky Way Galaxy, as imaged by the Spitzer Space Telescope. *Donald Richey/NASA*

HECC Participates in Ames Partnership Days

- The HECC Publications and Media team organized the cross-divisional Advanced Computing exhibit during Ames Partnership Days, held April 29 – May 1, to showcase our resources, services, and facilities as an Ames core competency.
- More than 90 representatives from industry, U.S. federal labs, government agencies, and academia gathered at Ames to learn about the center's groundbreaking research and development capabilities, and to explore possible partnerships.
- Seven researchers spoke during the event's first technical session, chaired by NAS Division Chief Donovan Mathias. The researchers described NASA's High End Computing Capability, Modular Supercomputing Facility, science data pipeline processing, visualization and data analysis, computational fluid dynamics, quantum computing, and neuromorphic computing. Another NAS researcher spoke during a separate session about data science, machine learning, and artificial intelligence.
- The team recruited staff and subject matter experts to participate in the exhibit throughout the event to show visualizations, answer questions, and provide information to many of the registered participants.
- Visitors to the exhibit who wanted to talk further about exploring possible partnerships included representatives of the University of Alabama, Agscent, Rancher Government Solutions, and Agilent.

IMPACT: As part of an Ames core competency, HECC and NAS participate in key center events to communicate the role of supercomputing and modeling and simulation at NASA projects and missions.



NAS and HECC staff and researchers participated in the Advanced Computing exhibit to speak about the role of supercomputing in NASA's projects and missions, and to explore possible partnerships. From left: Michelle Moyer, HECC Publications and Media; Blaise Hartman, HECC User Services. *Ted Manning, NASA/Ames*

NASA-GPT: Searching NASA Reports Using AI*

- Researchers at NASA Ames used artificial intelligence (AI) tools to create an internally hosted chatbot and AI-enhanced search tool with access to several agency report servers and data repositories, such as the NASA Technical Reports Server and the JPL Technical Reports Server.
- Using NASA-GPT, users can ask specific questions about agency programs and quickly access key data from published reports, presentations, and logs. For example: "What was the size of the inlet bleed holes on the XB-70?"—related to the image at right.
- Users can also get answers to technical and procedural questions about how to use the computational resources provided by HECC.
- A key feature of NASA-GPT is that it provides links to relevant sources so that users can verify that the language model is not just hallucinating an answer, a phenomenon other chatbots have recently come under fire for.
- Running large language models like NASA-GPT is computationally demanding; HECC provided many (6,000-10,000) GPU hours to convert three million PDFs to text to make them readable by the model. The full NASA-GPT model requires at least four NVIDIA A100 GPUs on the Cabeus supercomputer to run, although multiple users can use it simultaneously.

IMPACT: With HECC supercomputing resources, NASA-GPT can re-create much of the capability of commercial tools for an internal userbase. Developed in just a few weeks, this tool saved hundreds of thousands of dollars compared to procuring a tool from a commercial provider.



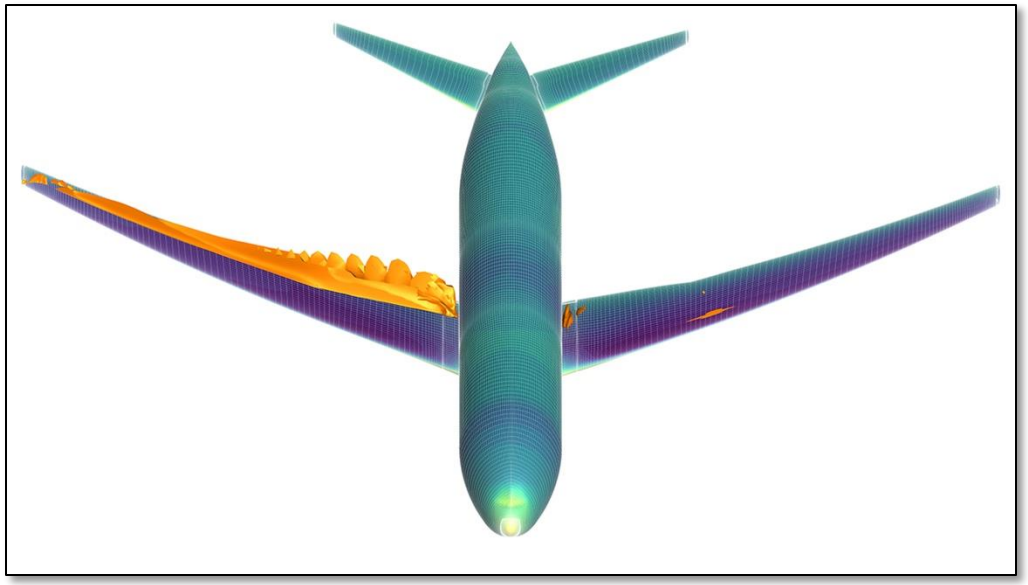
Cropped image from a demonstration of NASA-GPT, answering a specific design question about a retired experimental aircraft from a paper written in 1966. The answer was helpful for a current NASA civilian hypersonic aircraft project. *Derek Dalle, NASA/Ames*

** HECC provided supercomputing resources and services in support of this work*

Optimizing Aircraft Performance through CFD*

- Researchers in the Launch, Ascent, and Vehicle Aerodynamics (LAVA) group in the NASA Advanced Supercomputing Division are utilizing their in-house-developed computational fluid dynamics (CFD) software to support agency efforts to optimize aircraft designs for improved efficiency, lower fuel costs, and reduced noise.
- By fine-tuning the shapes of wings, fuselages, and other components, aircraft designers can reduce air resistance, which reduces the fuel required for flight in addition to reducing noise levels. This approach makes air travel more cost-effective for airlines and passengers.
- Aerodynamic design optimization requires running hundreds of simulations to explore a variety of design possibilities. The team's optimization of a full aircraft configuration using 1,280 Electra Skylake cores took 138 hours to complete.
- By applying aerodynamic design optimization to an existing configuration, the LAVA team was able to reduce the drag by over 4%, translating to significant fuel savings in real-world applications.
- The team's next goal is to make the simulations more computationally efficient, allowing them to perform these high-fidelity optimizations faster and using fewer computational resources.

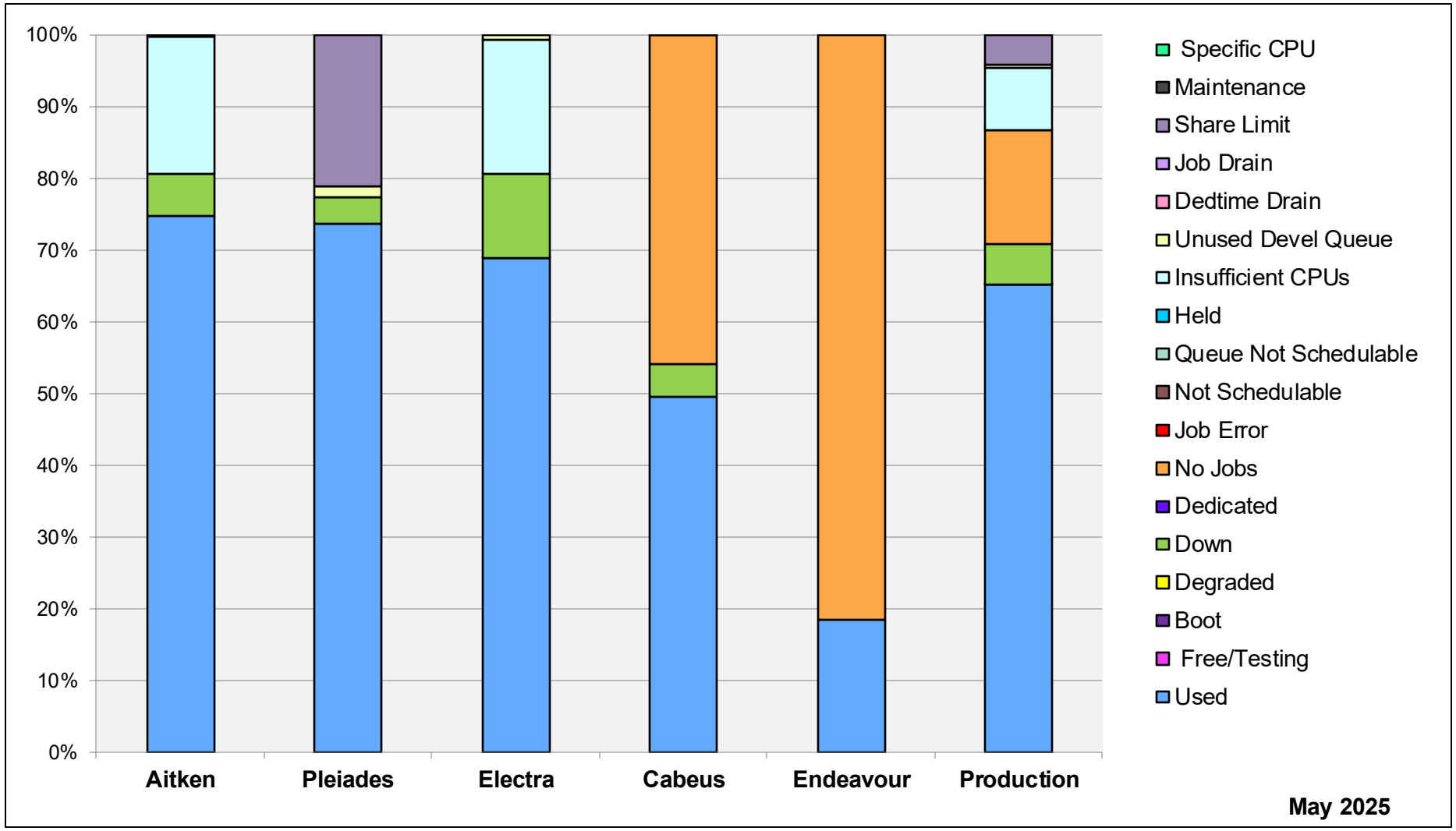
IMPACT: One of NASA's primary aviation goals has been to continually address the challenge of cutting aircraft fuel consumption without compromising safety or performance.



Comparison of aircraft designs: The initial geometry (left) and the optimized shape (right). The surface is colored by the air pressure on the aircraft, with orange surfaces showing shock waves. The optimized design reduces drag by 4% compared to the original, leading to improved fuel efficiency. *Brandon Lowe, NASA/Ames*

** HECC provided supercomputing resources and services in support of this work*

HECC Utilization



HECC Utilization Normalized to 30-Day Month

