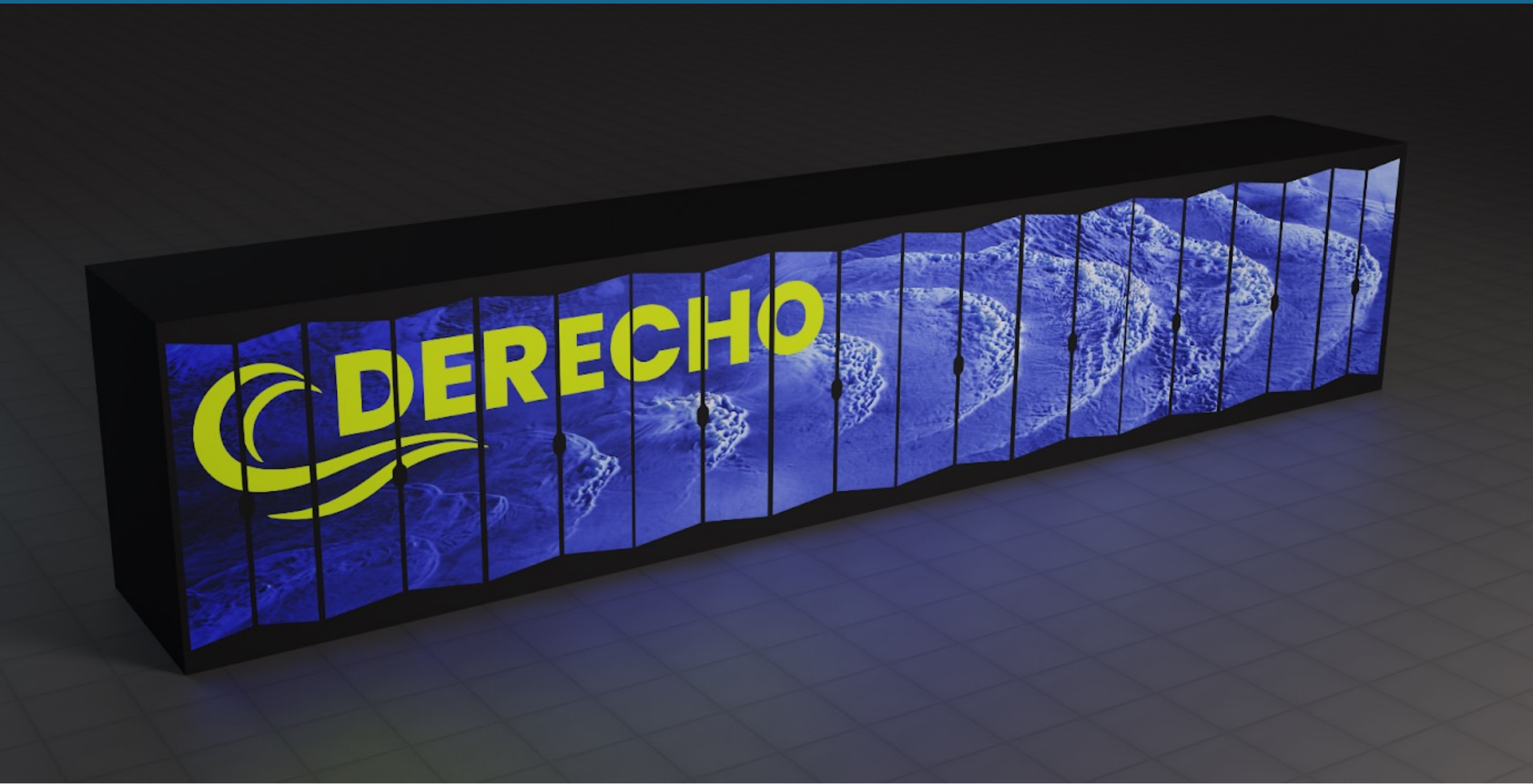
Navigating the evolving path to exascale with NCAR's Derecho

David L. Hart, NCAR
Computational & Information Systems Lab

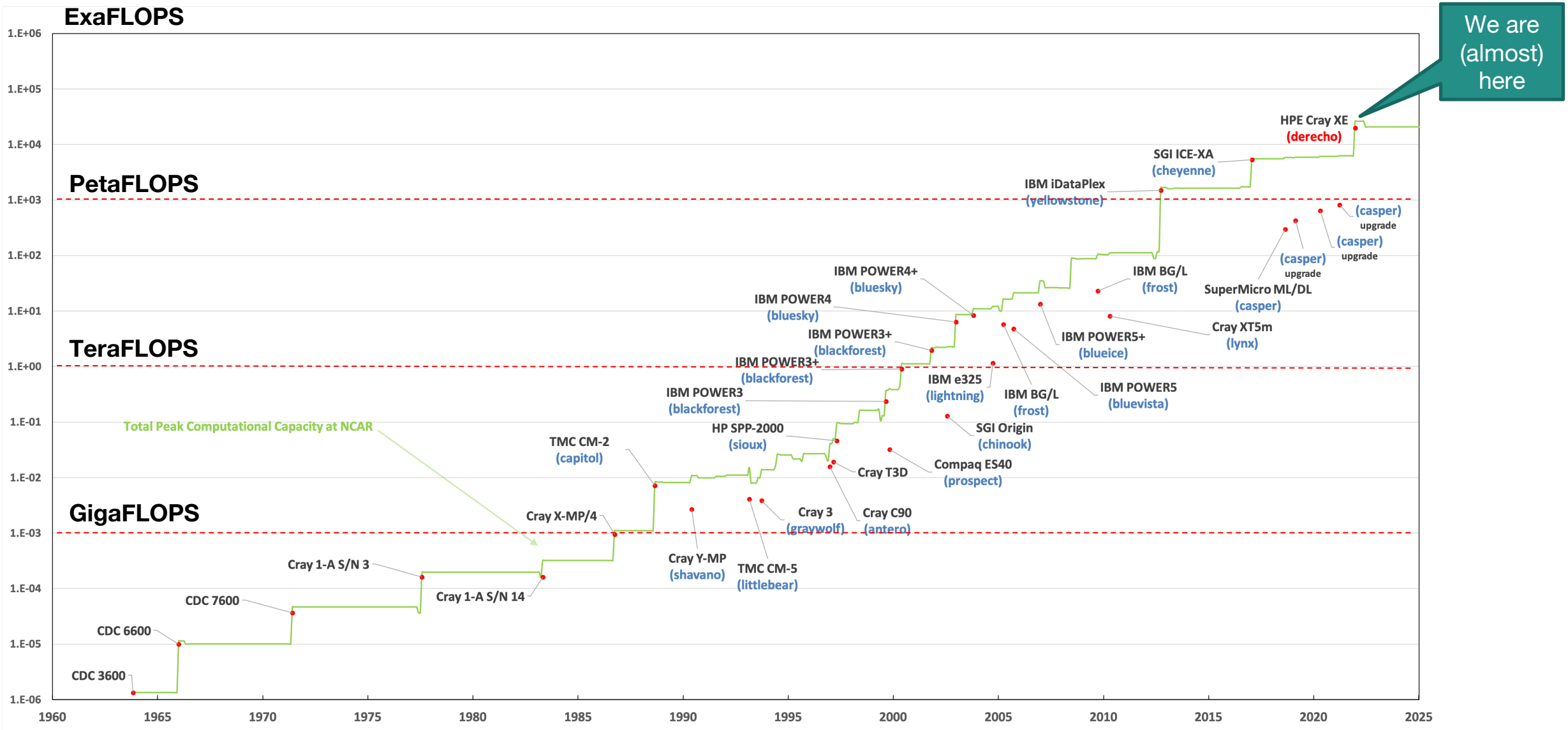
20 September 2021

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Coming in 2022 to NCAR

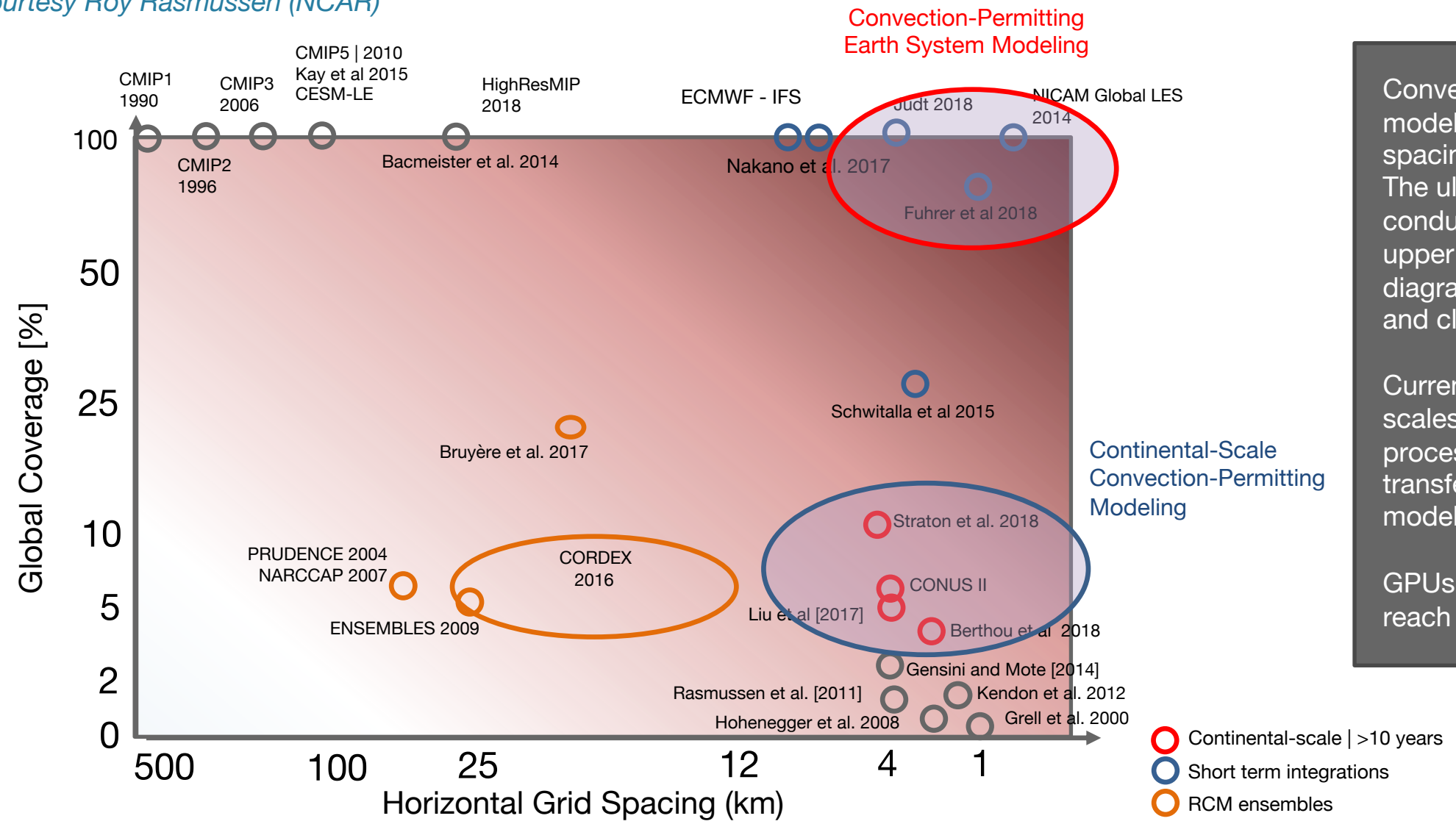


The Path to Exascale HPC at NCAR



Earth Systems Science Drives Us toward Exascale

Courtesy Roy Rasmussen (NCAR)



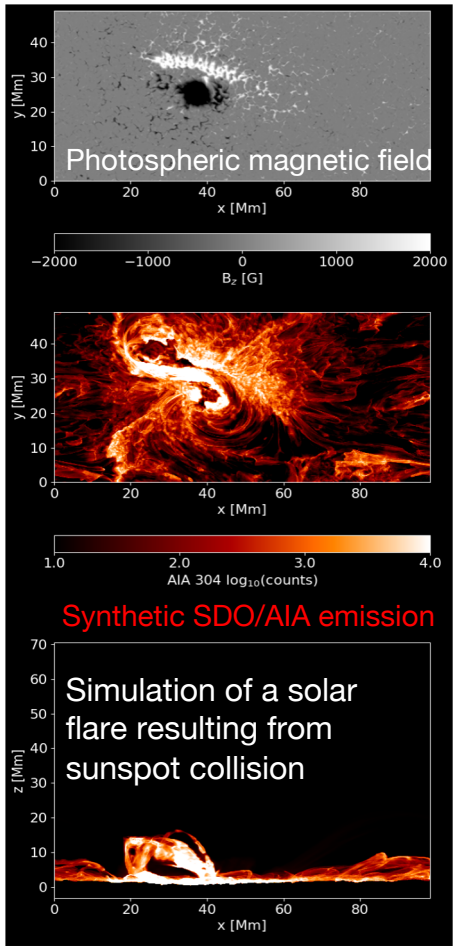
Convection-permitting modeling occurs at grid spacings of less than 4 km. The ultimate goal is to conduct research in the upper right portion of the diagram at both weather and climate time scales.

Current work at continental scales is developing a process that can be transferred to global models.

GPUs will be required to reach this goal.

Advancing Understanding of Solar Phenomena

Solar simulations need faster throughput *and* higher resolution. MURaM OpenACC will help meet these requirements.



Simulation of a solar flare resulting from sunspot collision

- Max Planck University of Chicago Radiative MHD (MURaM) models the solar atmosphere from upper convection zone to lower solar corona
- Goals for MURaM-OpenACC
  - **Short-term:** Solar models capable of running models at the resolution of DKIST telescope observations
  - **Long-term: Better prediction of space weather events** using data-driven models of solar eruptions
- Status Refactoring of MURaM for GPU using OpenACC
  - Refactoring, optimization focused on radiation transport solver (RTS)
  - Have achieved about 69x a Skylake core on a V100 (1.76x a node) on RTS
  - Have scaled to ~100 GPUs.

Slide courtesy of Rich Loft (NCAR)

Daniel K. Inouye Solar Telescope

MURaM OpenACC project is an HAO/CISL collaboration with the University of Delaware and the Max Planck Institute for Solar System Research & Lockheed

Expanding Machine Learning Activities at NCAR

Artificial Intelligence for Earth System Science (AI4ESS) Summer School — 2020

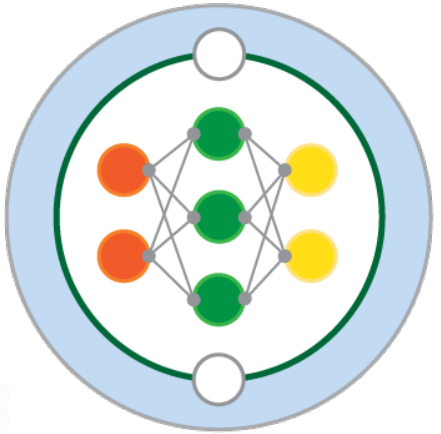
- Averaged 1,500 attendees daily
- 150 hackathon participants throughout the week

Trustworthy Artificial Intelligence for Earth System Science (TAI4ESS) Summer School — 2021

- In partnership with AI2ES center
- About 500 attendees

NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES) — *ai2es.org*

Multiscale Machine Learning In Coupled Earth System Modeling (M²LInES) — *m2lines.github.io*



SCHMIDT FUTURES

Using the Cloud for Specialized HPC Use Cases

Today: Cheyenne + Cloud

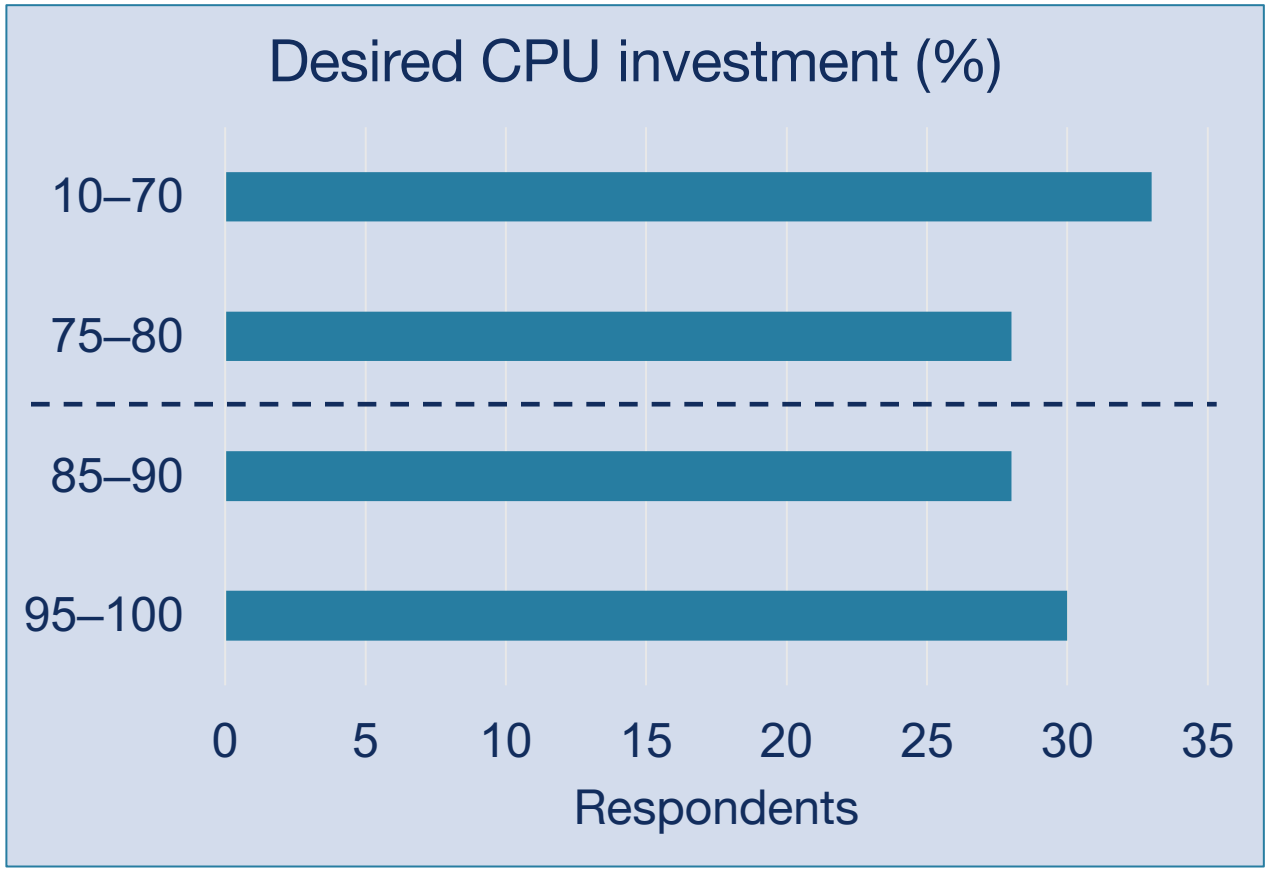
- NCAR-operated Antarctic Mesoscale Prediction System (AMPS) produces twice-daily weather forecasts covering Antarctica
- During system maintenance, the AMPS forecast workflow shifts to the Penguin On Demand HPC cloud or Amazon Web Services

Tomorrow: Derecho and integrated cloud bursting

- Derecho scheduler to include PBS Cloud Connector
  - cloud credit needed to use commercial cloud
- Enhanced support for expanded use cases
  - seamless support for **high-availability** needs
  - **on-demand** support for urgent computing
  - extensible **high-throughput** computing capacity
- Sample images for NCAR models being developed

Community Feedback Contributed to Derecho's Design

- Science Requirements Advisory Panel convened (SRAP)
  - 44 members from NCAR and university community
- Provided application drivers
- Considered Cheyenne workload analysis
- Reviewed Community Survey input
- Considered likely technology options
  - Processors, memory, storage
- Made key recommendations to Derecho design
- NCAR also conducted a co-design process with potential vendors



Among other results, our community survey found that roughly half of respondents would invest 80%+ of funds in conventional CPUs (rest in GPUs), while the other half wanted to spend less on CPUs (more in GPUs). Such results helped guide NCAR's plans for Derecho.

Derecho architecture

- **HPE Cray EX cluster**
- **19.87 PetaFLOPS (peak)**
- **60 PB usable storage capacity (Cray Lustre)**
- **Slingshot v11 interconnect**
- **3.51x performance of Cheyenne**

DERECHO
NCAR Supercomputer

Derecho CPU

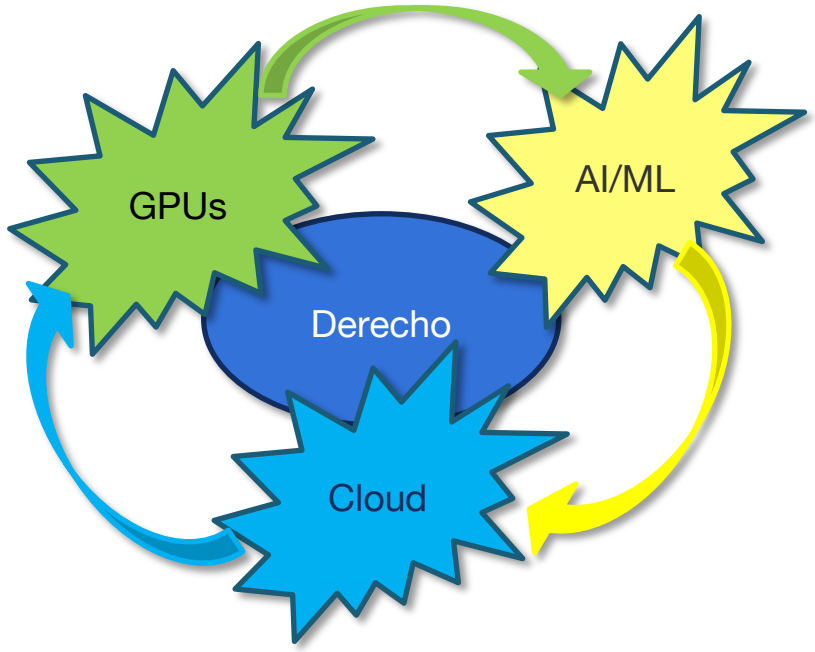
- 323,712 processor cores
- 637 GB total memory
- 2,488 dual-socket nodes
  - 64-core AMD EPYC 7763 “Milan” processors
  - 256 GB memory per node
  - 1 Slingshot injection port
- 2.84 Cheyenne Sustained Equivalent Performance (CSEP)
- 80% of expected Derecho performance
- 13.47 Petaflops (peak)

Derecho GPU

- 328 total Nvidia A100 GPUs
- 40 GB HBM2 memory per GPU
- 82 GPU nodes
  - 4 Nvidia 1.41-GHz A100 Tensor Core GPUs
  - 600 GB/s NVLink
  - 512 GB DDR4 memory
  - 4 Slingshot injection ports

- 0.67 CSEP
- 20% of expected Derecho performance
- 6.4 Petaflops (peak)

Major increase in GPU Capability at NCAR



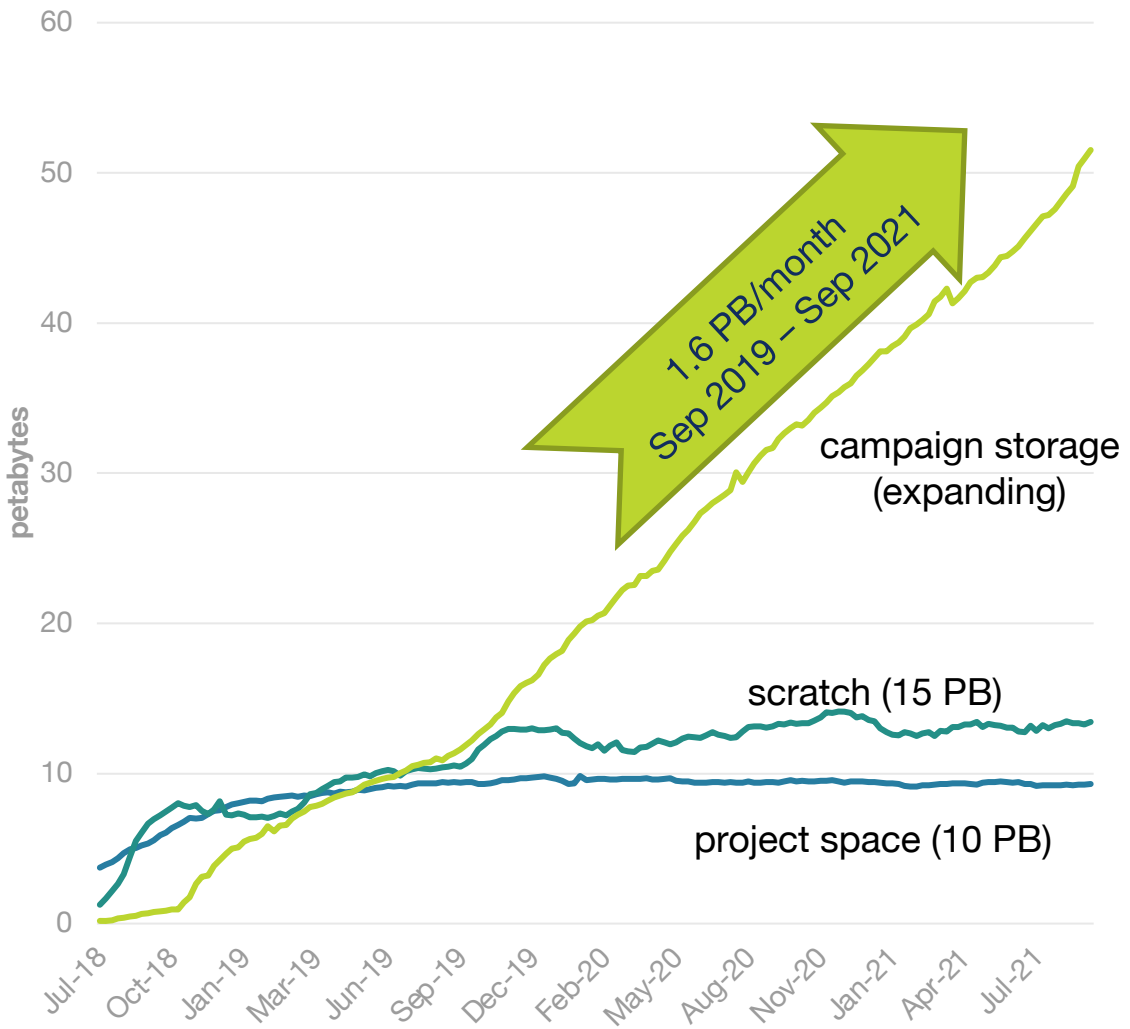
- Derecho will bring together GPU architectures, machine learning, cloud computing, and related software technologies
- **Major training, outreach, and support effort required**
  - Optimizing software-stack configuration
  - Managing the hardware, software ecosystem, and user environment
  - Exploring new capabilities (e.g., GPU Direct Storage)
  - Developing GPU expertise within user community
  - Preparing GPU porting guides (particularly for Fortran)
  - Maintaining knowledge base/best practices
  - Offering regular training on GPUs and AI/ML for Earth science problems

An Aside: How Derecho Got its Name

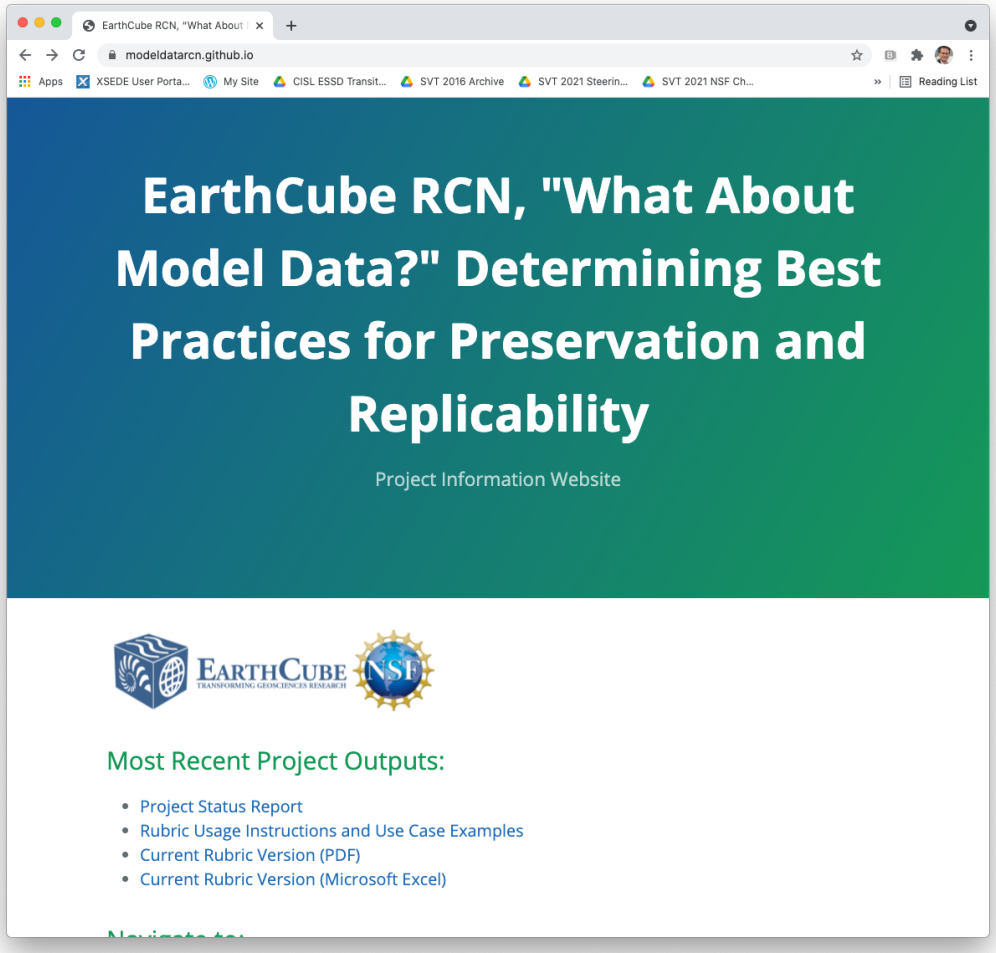
- NCAR held a contest open to students in the state of Wyoming to name our next supercomputer
  - In partnership with Wyoming Dept of Education and Wyoming Governor's Office
  - More than 200 entries received
- Winning entry submitted by Cael Arbogast, a student at Riverton Middle School in Riverton, Wyoming
  - Town of 10,750, surrounded by the Wind River Indian Reservation
  - 560 km from NCAR's Mesa Lab
  - 430 km from NWSC data center
- Cael won an iPad and his class won a pizza party

Storage challenges on the horizon

- With great (computing) power comes great (data) responsibility!
- With Derecho, 60 PB scratch file system will have (at least) six month retention
  - potentially up to one year
- Campaign storage expands each year
  - designed to hold data for the duration of "projects" (e.g., research funding awards)
  - net expansion will slow as disks are retired
- Growth rate will **increase** with Derecho
  - *potentially 5 PB / month or more!*
- Data management becoming more and more critical to ensuring science impact



Defining More Consistent Practices for Data Retention



https://modeldatarcn.github.io/

- NCAR is co-leading an EarthCube project to define best practices for preservation and replicability of Earth system model data
- Two workshops in 2020 have led to the creation of a quantitative rubric and supporting instructions for researchers
  - Formal publication likely in 2022
- Rubric asks researcher to assess value and need for data across 8 “themes”
  - Community commitment
  - Repository access capabilities
  - Simulation workflow repeatability
  - Post-processing workflow repeatability
  - Research workflow output accessibility
  - Research feature reproducibility
  - Cost of running simulation workflow
  - Repository services cost

Latest Derecho timeline

Phase / Milestone	Timeline
Procurement kickoff	June 2018
Technical evaluation	Aug. 2018 – Aug. 2020
Science requirements advisory panel (SRAP)	Nov. 2018 –July 2019
NWSC facility upgrades	Dec. 2019 – Jan. 2022
NWSC HPC fit-up	May 2020 – June 2021
Assembly, Delivery, Installation, Acceptance	Jan. 2021 – Mid-2022
Accelerated Scientific Discovery (ASD)	Mid-2022 (two months)
General user access	After ASD projects

- NCAR completing Module A at NCAR-Wyoming Supercomputing Center (NWSC) for Derecho's installation
  - 3 MW electrical capacity added
- Currently awaiting hardware availability
  - delays due to global chip shortage

Accelerated Scientific Discovery (ASD) on Derecho

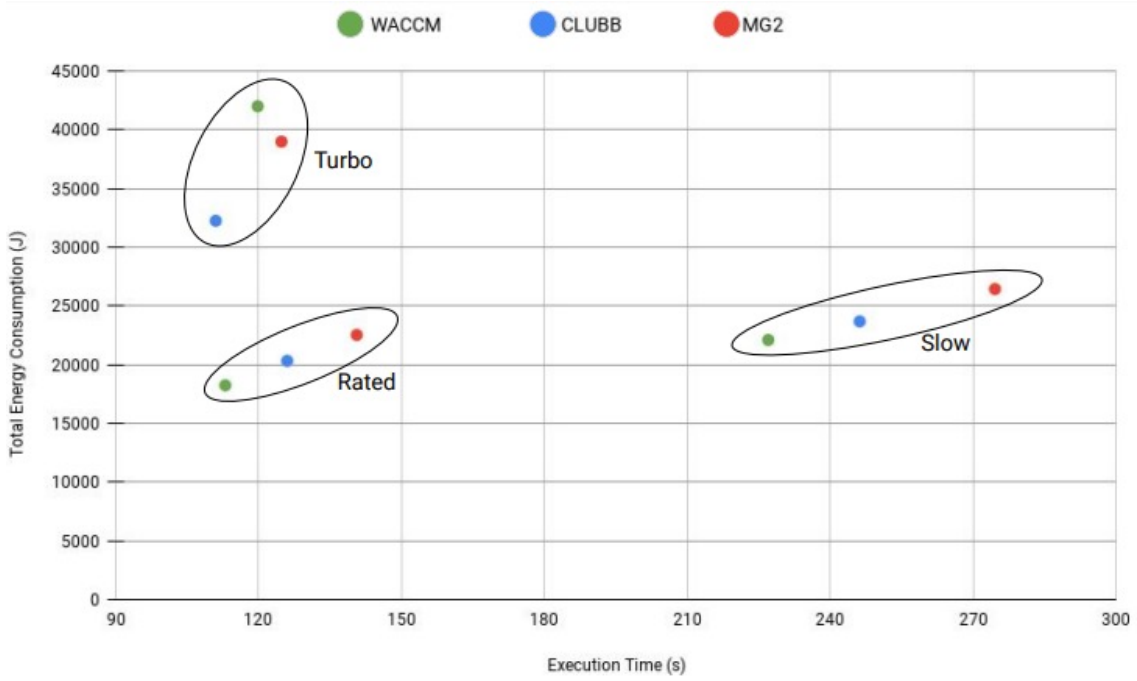
- Small number of projects given access to Derecho system for its first two months
  - Half university-led, half NCAR-led (roughly)
- **450 million core-hours on the AMD EPYC nodes**
  - 10-12 projects
  - 30M core-hours minimum
- **450,000 GPU-hours on the Nvidia A100 nodes**
  - approx. 6 projects
  - 50k GPU-hours minimum

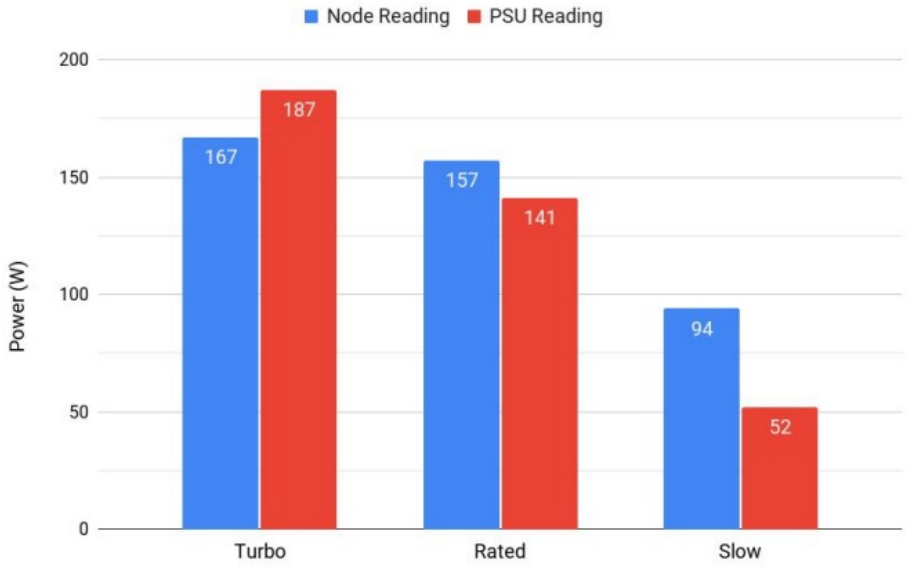
- 10 University-led project proposals now under review
  - Final selections to be made by November
- NCAR strategic selection process underway
  - 18 pre-proposals submitted with 10 moving to full proposal phase
  - Full proposal review completed by March

Panorama of the 2020 Midwest derecho with shelf cloud.
Photo courtesy of Maddie Murphy from National Weather Service.

Continued Pursuit of Power Efficiency

- Student intern in summer 2020 performed experiments comparing performance and power consumption of running three application kernels at turbo, rated, and slow processor speeds.
- At rated speed, benchmarks saw an average 45% decrease in energy consumption versus an average 7% increase in execution time compared to Turbo

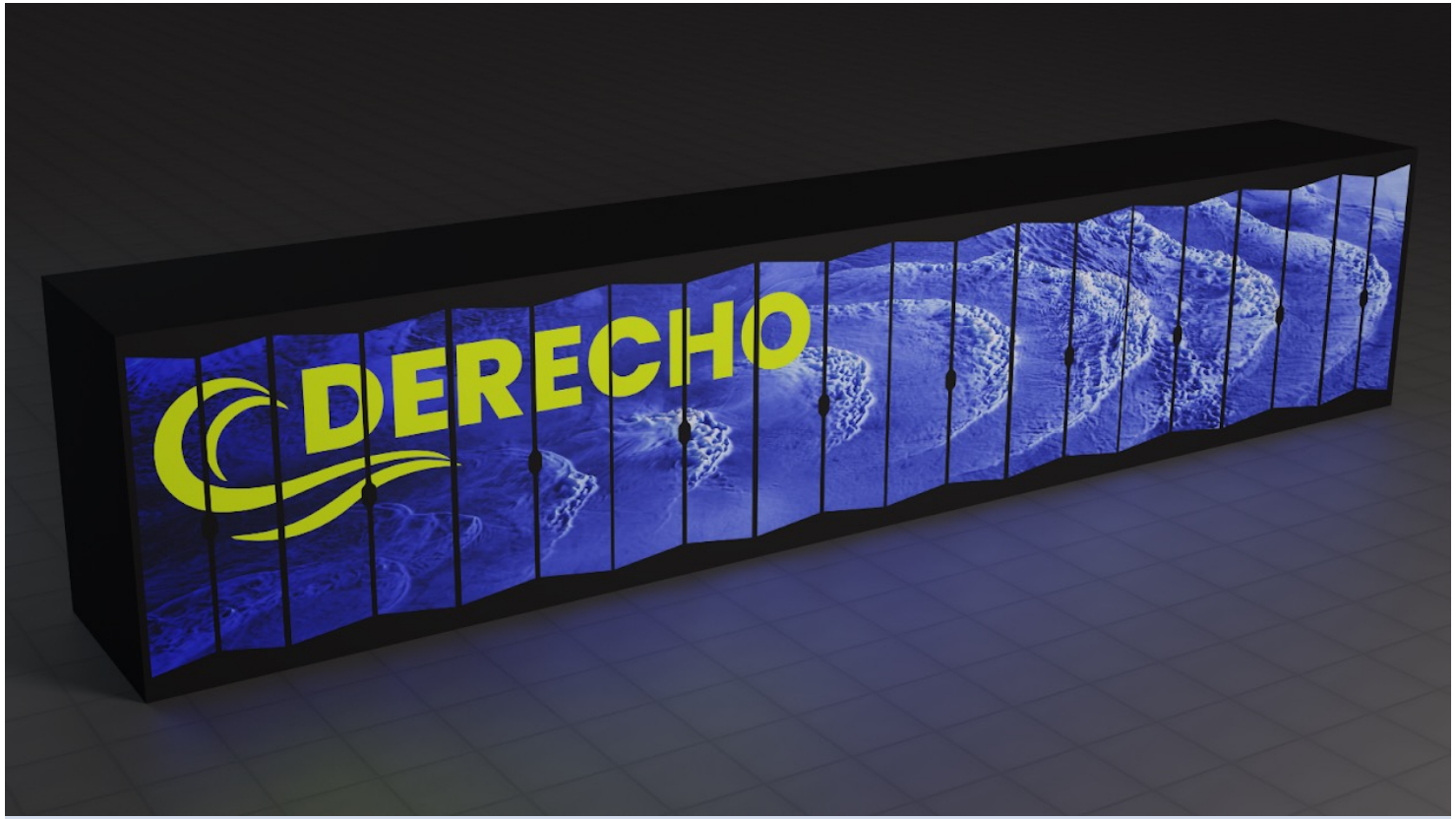




- Student also showed that downclocking idle/sleeping nodes could also have non-trivial power impacts
- Idling at Slow uses 43.5% less power than idling at Turbo
- Experiments will continue on Cheyenne and Derecho

Spencer Diamond, 2021, *"Lowering the Cost of Climate Research: Energy Consumption vs Clock Speed for Various Application Profiles"*

Coming in 2022 to NCAR



User communities will transition to Derecho in second half of 2022
Cheyenne to remain in operation until end of 2022