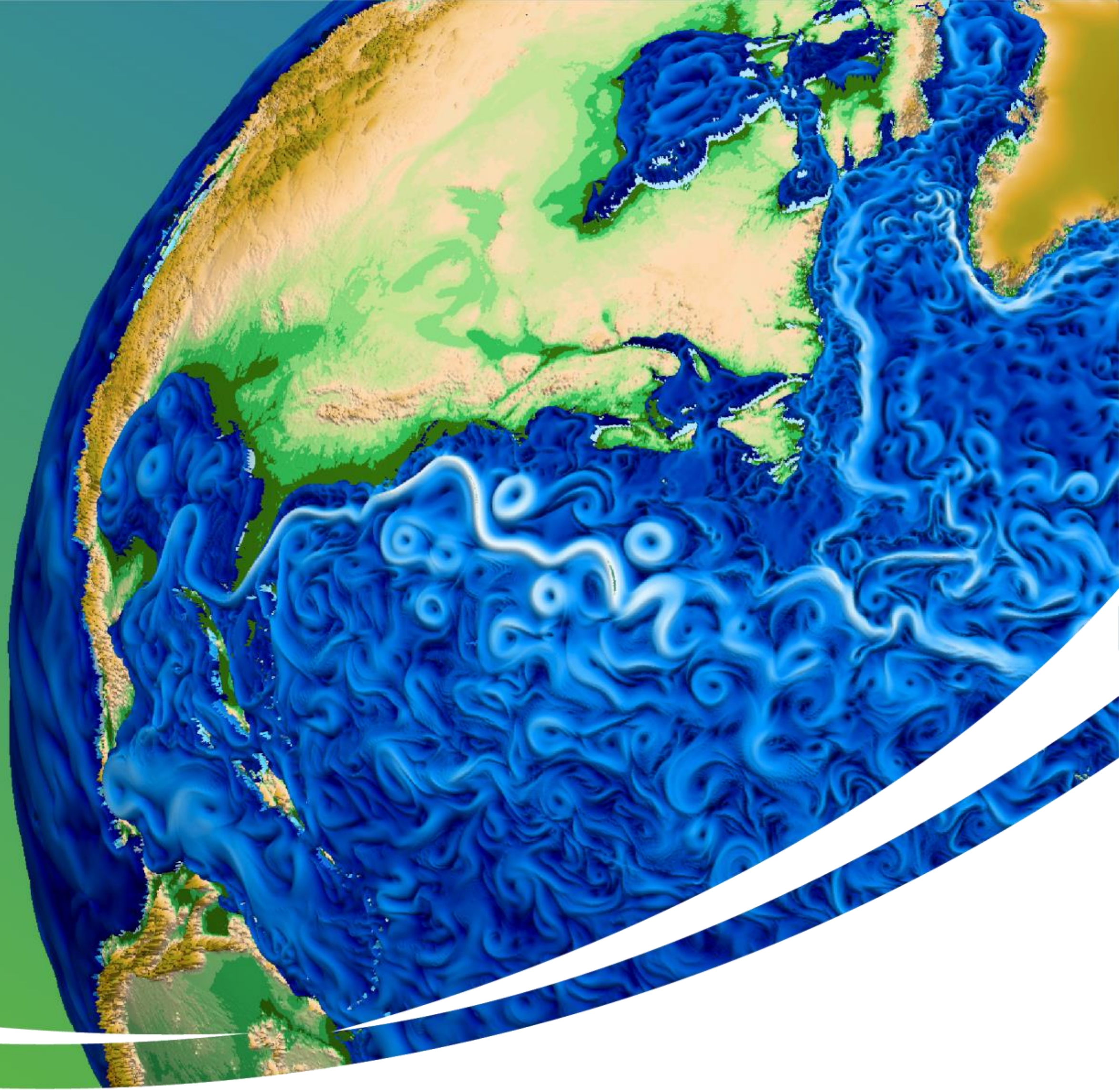
High Performance Computing for Climate

Sarat Sreepathi

Oak Ridge National Laboratory

Performance Analytics for Computational Experiments

PACE

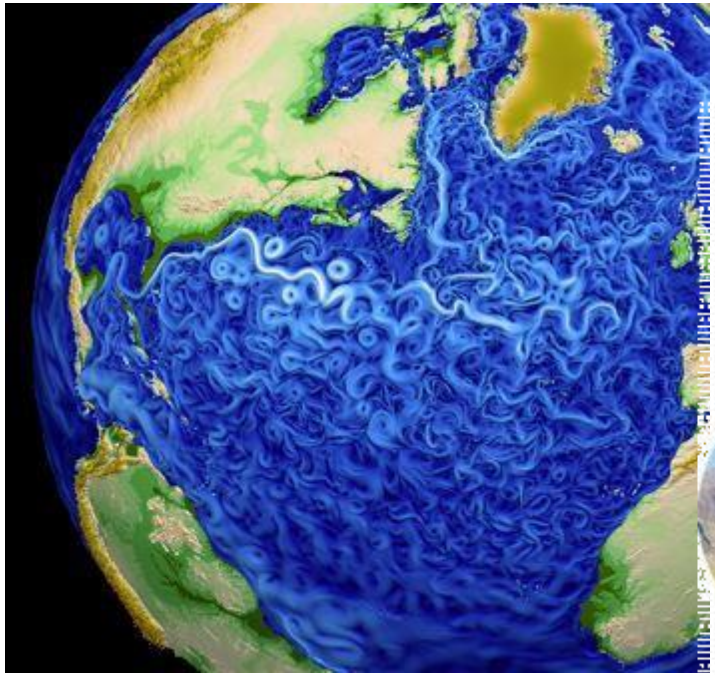


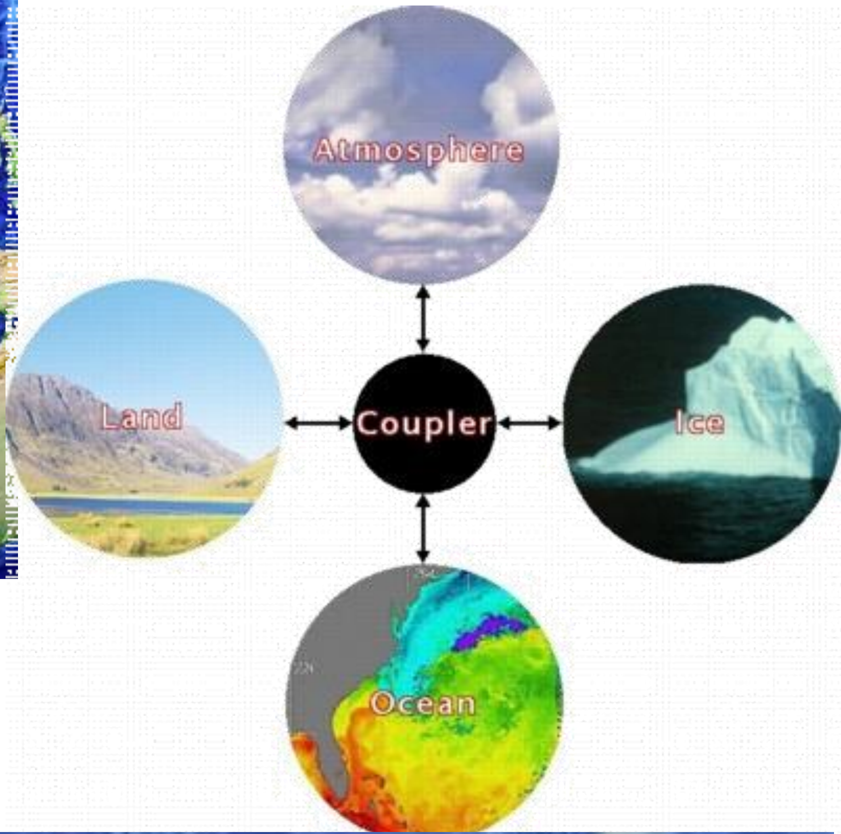
U.S. DEPARTMENT OF ENERGY

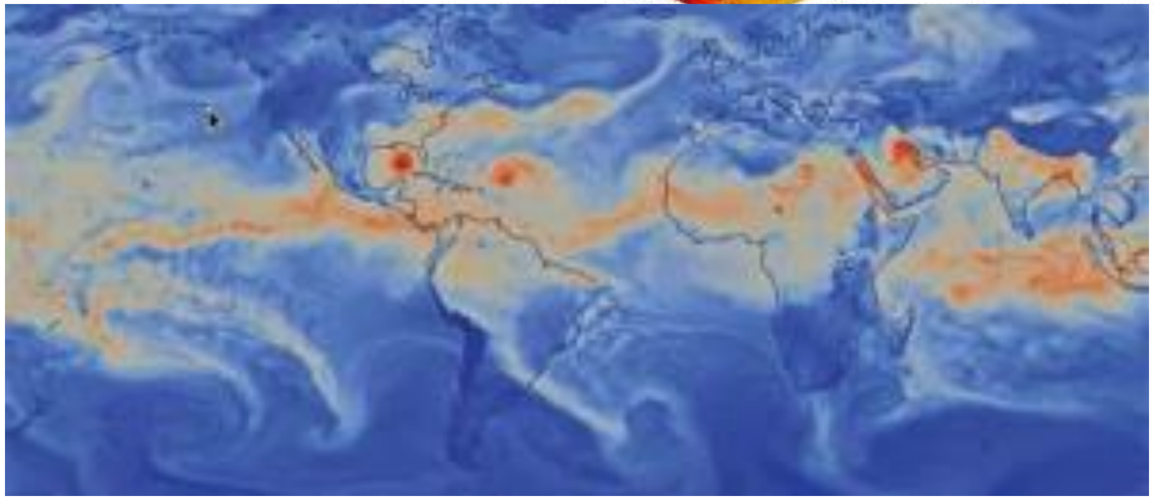
E3SM
Energy Exascale Earth System Model

Energy Exascale Earth System Model (E3SM)

- Global Earth System Model
- Atmosphere, Land, Ocean, Ice, … component models
- 8 DOE labs, 12 university partners,… ~$30+ M/year
- Development driven by DOE mission interests: Energy/water issues looking out 40 years
- **Key computational goal: Ensure E3SM effectively utilizes DOE exascale supercomputers**
- E3SM is open source / open development
  – Website: www.e3sm.org
  – Github: https://github.com/E3SM-Project





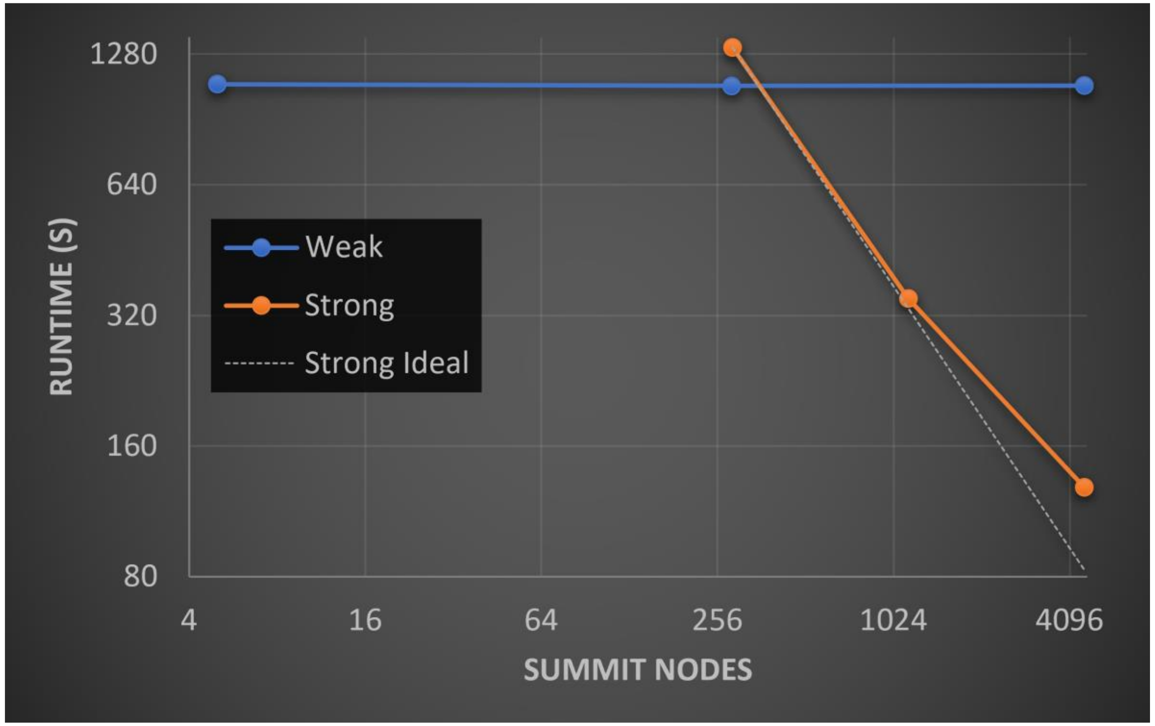


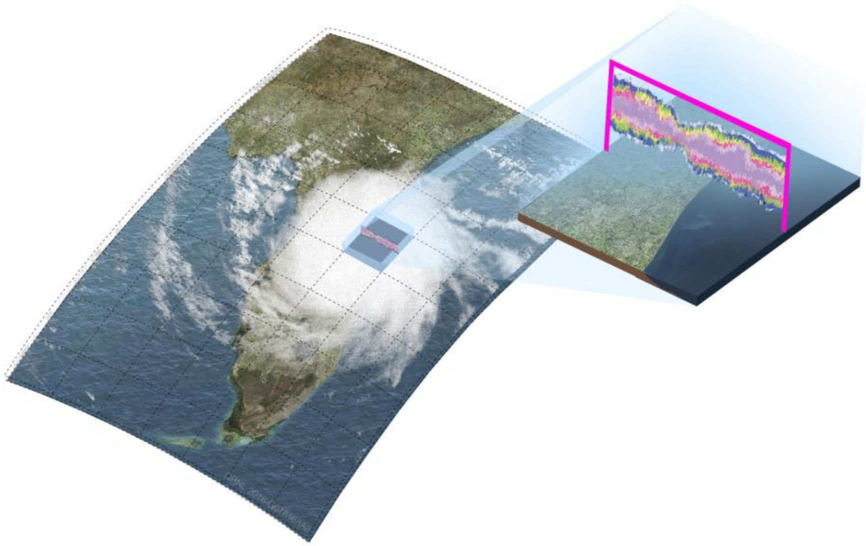
Mission: Use exascale computing to carry out high-resolution Earth system modeling of natural, managed and man-made systems, to answer pressing problems for the DOE.

E3SM-MMF Cloud Resolving Climate Model

EXASCALE COMPUTING PROJECT

Goal: Develop capability to assess regional impacts of climate change on the water cycle that directly affect the US economy such as agriculture and energy production.





- Multiscale Modeling Framework (MMF) / Super-Parameterization
- Replaces traditional parameterizations with cloud resolving model within each grid cell of global climate model

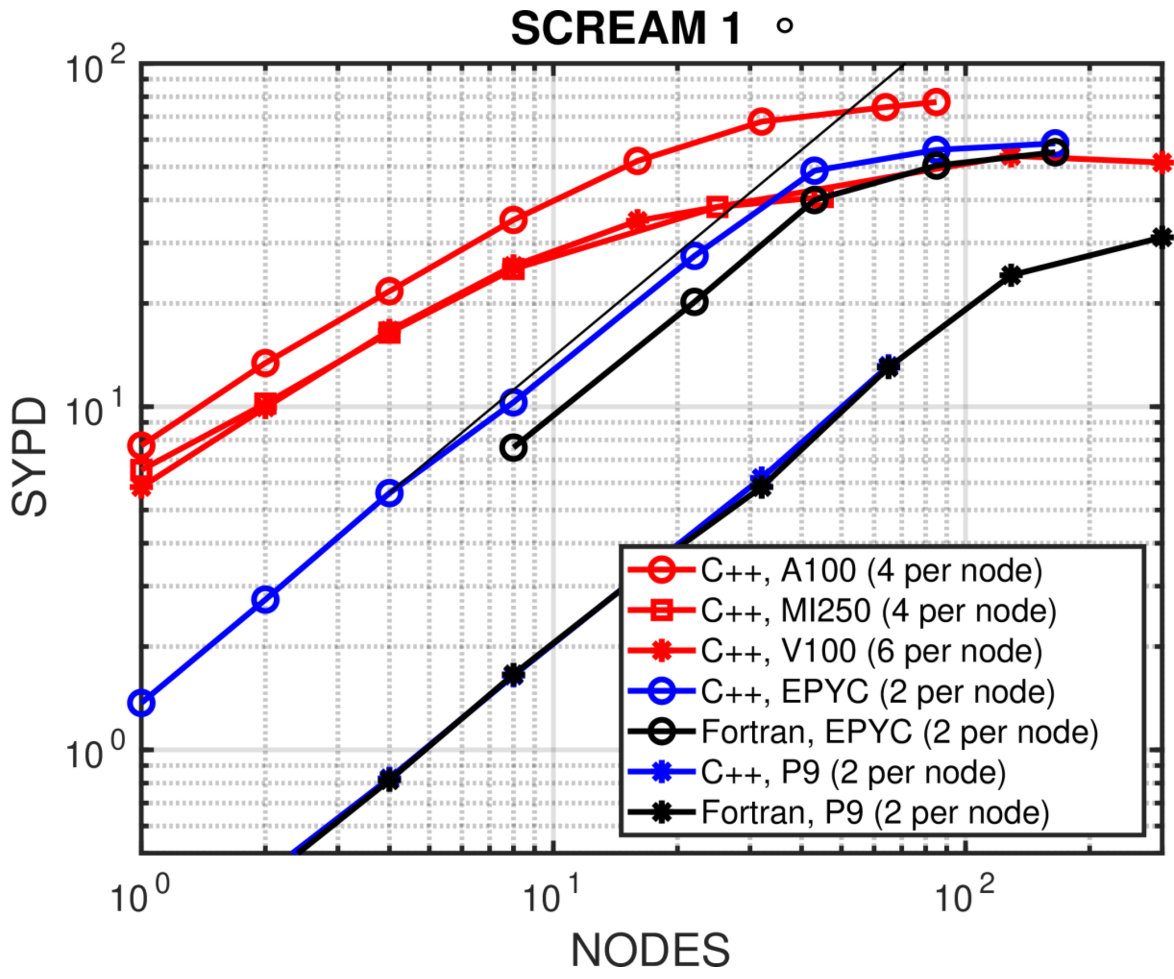
Programming Models

- C++ with templates (Kokkos or YAKL)
  - Robust and well supported solution across most hardware
  - Requires minimal vendor support
- Fortran with OpenACC or OpenMP offload
  - Relies heavily on (lagging) vendor compiler support
  - Remains immature w.r.t. advanced Fortran features
  - Good performance requires major code refactoring
- Domain Specific Languages
  - Promising approach (e.g. GT4Py/GridTools, PSyclone)
  - Need additional investments to support algorithms & meshes in E3SM components
  - Most experience within DOE labs is with C++

C++/Kokkos: Performance Portability

E3SM's Atmosphere model (EAMXX in "SCREAM" configuration)
1 degree resolution: 128 vertical levels, nonhydrostatic (NH) dycore, 10 tracers, P3/SHOC physics with prescribed aerosols, no convective parameterization

- Performance portability
  - IBM P9, AMD EYPC
  - NVIDIA V100, A100
  - AMD MI250
- CPU performance:
  - C++/Kokkos as fast or faster than Fortran
- GPU performance:
  - Large scaling range where GPU nodes are 4-10x faster than CPU nodes



Early Evaluation of Fugaku A64FX Architecture Using Climate Workloads

Sarat Sreepathi
Oak Ridge National Laboratory

Mark Taylor
Sandia National Laboratories

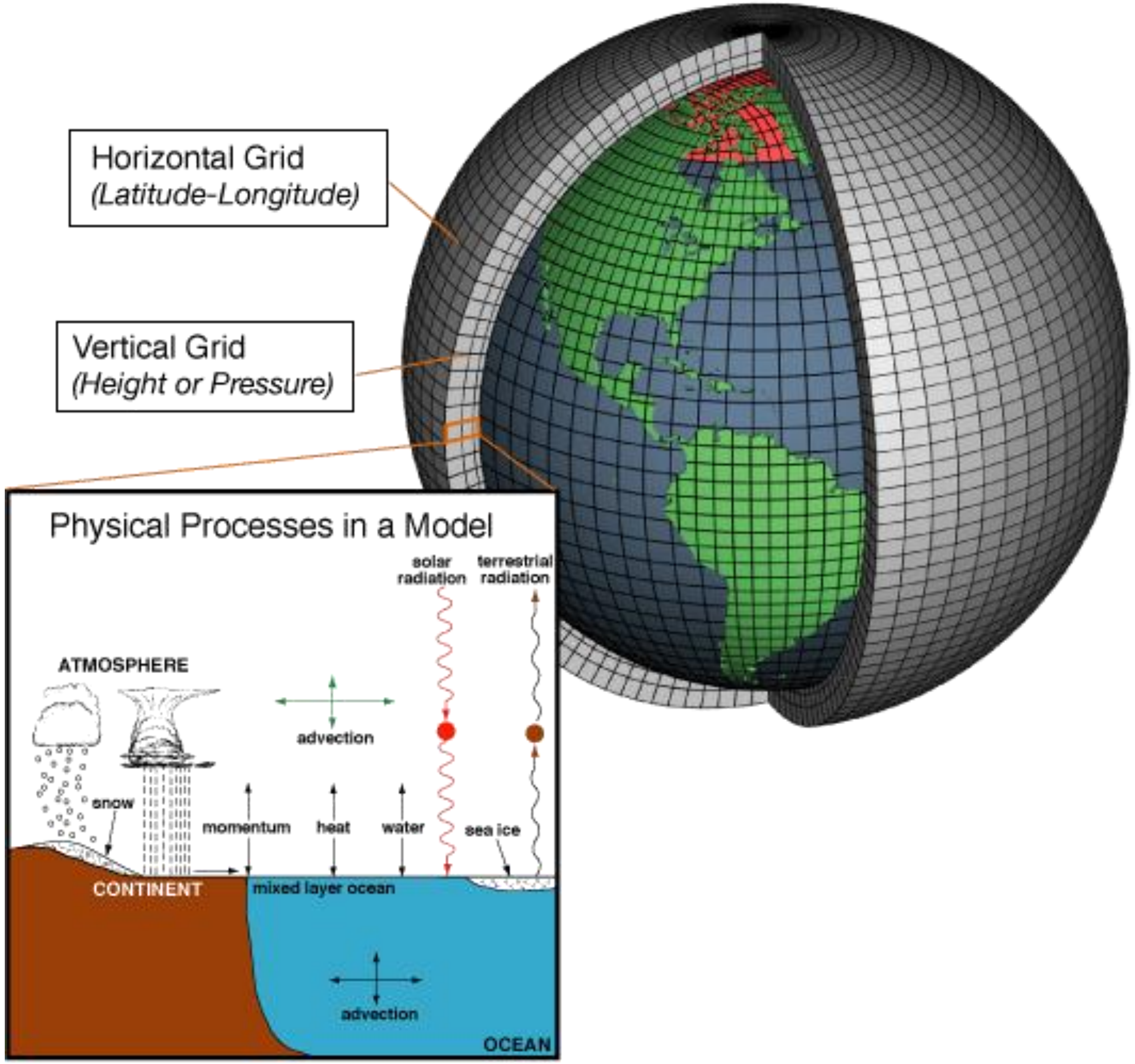
Adapted from talk given at
EAHPC Workshop
IEEE Cluster 2021
September 7, 2021

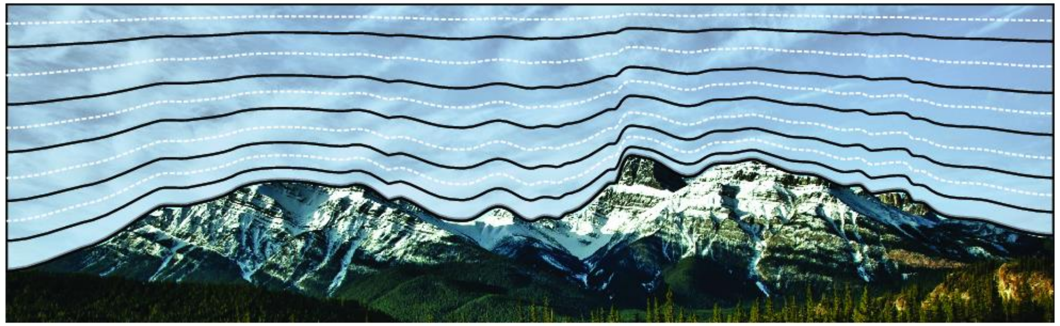
EXASCALE COMPUTING PROJECT

National Nuclear Security Administration

U.S. DEPARTMENT OF ENERGY | Office of Science

Atmosphere Component





hydrostatic-pressure terrain-following coordinates

- Dynamical Core
  - Solves the Atmospheric Primitive Equations
  - Linear transport of 40 atmospheric species
  - 72 vertical levels – 0.8 km avg. spacing
  - Benchmark (two versions): Fortran (preqx) and C++ (preqx_kokkos)

Terrain following figure: D. Hall, CU Boulder
Source: http://celebrating200years.noaa.gov/breakthroughs/climate_model/welcome.html

Fugaku

- #2 on Top500
- RIKEN Center for Computational Science
- Key Characteristics of A64FX*
  - Arm 64-bit with 512-bit SVE (Scalable Vector Extensions)
  - High Bandwidth Memory
  - Low Power

TOP 500 The List.

HOME LISTS STATISTICS RESOURCES ABOUT MEDIA KIT

Home »RIKEN Center for Computational Science »
Supercomputer Fugaku - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu...

SUPERCOMPUTER FUGAKU - SUPERCOMPUTER FUGAKU, A64FX 48C 2.2GHZ, TOFU INTERCONNECT D

Site:	RIKEN Center for Computational Science
System URL:	https://www.r-ccs.riken.jp/en/fugaku/project
Manufacturer:	Fujitsu
Cores:	7,630,848
Memory:	5,087,232 GB
Processor:	A64FX 48C 2.2GHz
Interconnect:	Tofu interconnect D
Performance	
Linpack Performance (Rmax)	442,010 TFlop/s
Theoretical Peak (Rpeak)	537,212 TFlop/s
Nmax	21,288,960
HPCG [TFlop/s]	16,004.5
Power Consumption	
Power:	29,899.23 kW (Optimized: **26248.36** kW)
Power Measurement Level:	2

*https://www.fujitsu.com/downloads/SUPER/a64fx/a64fx_datasheet_en.pdf

https://www.top500.org/system/179807/

EXASCALE COMPUTING PROJECT

Architecture Comparison: Metrics

- Single node workload for understanding h/w trends (ca 2012+)
- Performance Efficiencv metric: number of element remap timesteps **per second**

$$E_{perf} = \frac{N_e * N_t}{(prim\_main\_loop * num\_devices)}$$

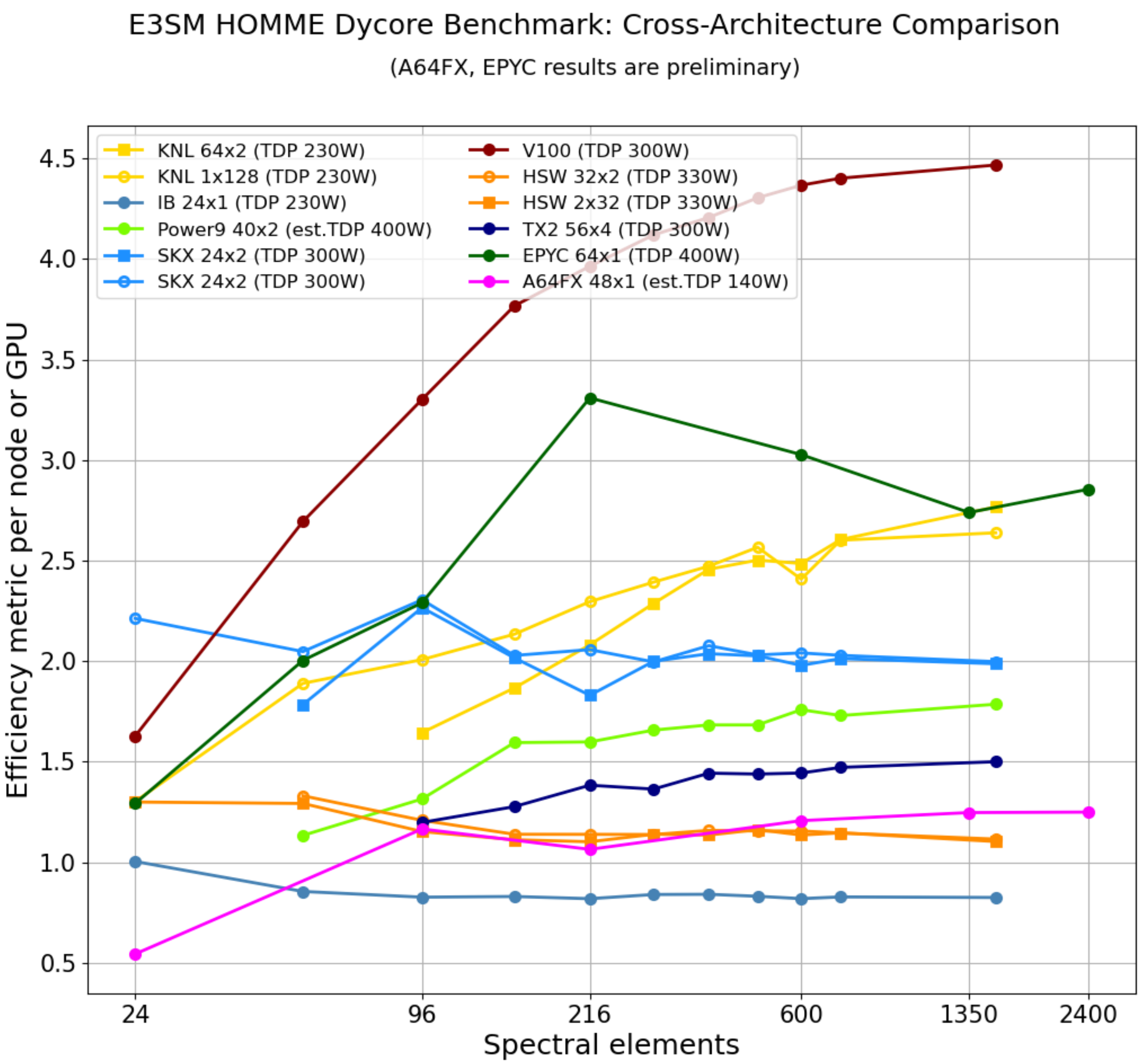
  - N_e is the number of spectral elements
  - N_t is the number of remap timesteps (34 for the Fugaku experiments)
  - *prim_main_loop* is the main computation loop timer
  - *num_devices* is 1 for CPU nodes or the number of GPUs per node for GPU systems
- Power Efficiency metric: number of element remap timesteps **per Watt**

$$E_{power\_tdp} = \frac{E_{perf}}{\text{TDP}}$$

  - Thermal Design Power (TDP)

EXASCALE COMPUTING PROJECT

Architecture Comparison: Performance Efficiency



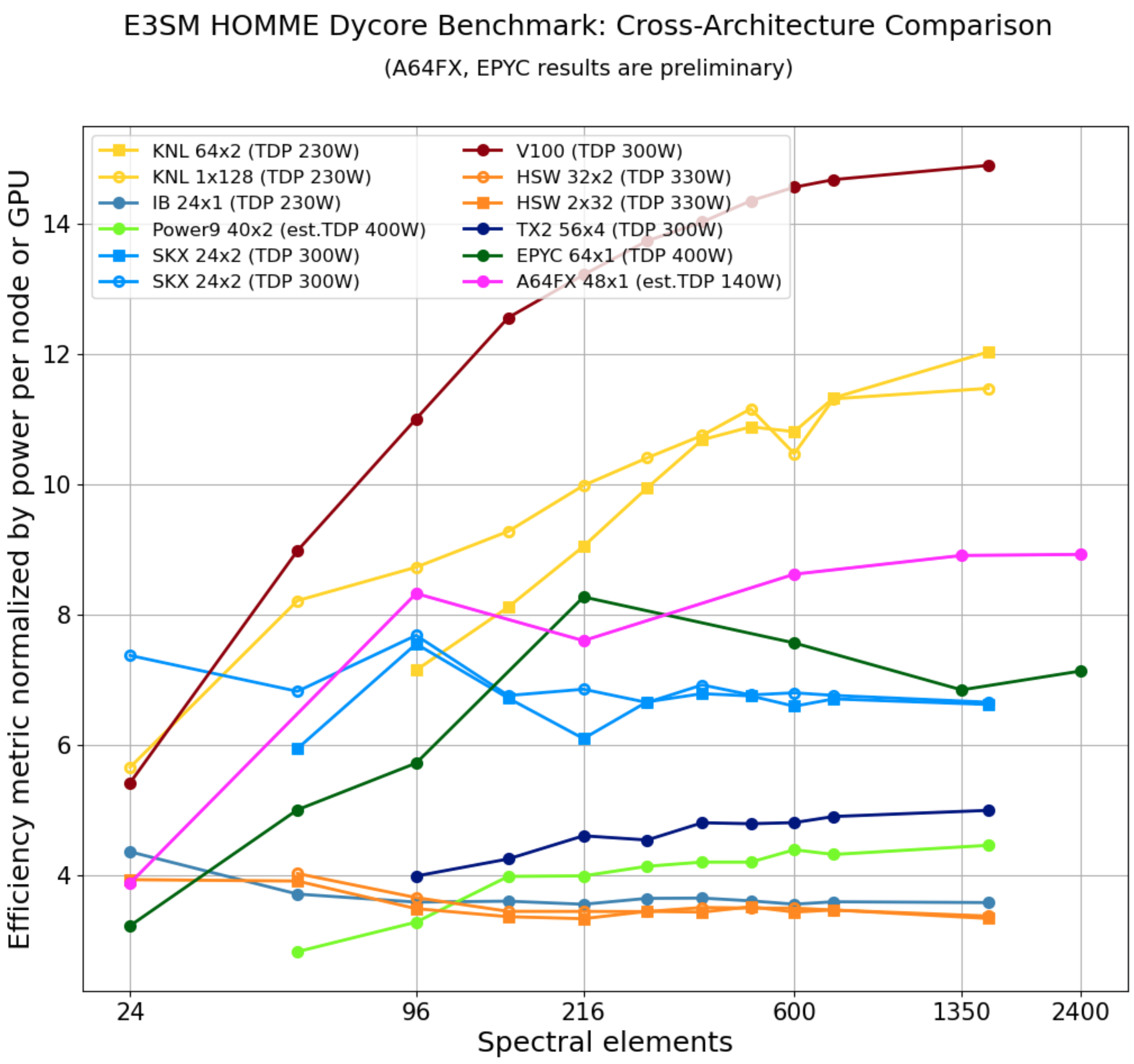
Inform configurations where GPU systems can outperform CPU systems

Fugaku Node: Single A64FX socket
GNU Fortran + MPI (48 ranks)

Note: Top Red (Volta V100), Pink (A64FX), Orange (Dual-socket Haswell)
Higher is better

* Plot of the efficiency metric normalized by power consumption on various hardware architectures. The legend includes a short descriptor for each architecture along with the number of parallel processes times (x) the number of threads and includes TDP in parenthesis. Specifically, the labels map as follows: KNL (Intel® Knights Landing), IB (Intel®Ivy Bridge), SKX (Intel® Skylake), V100 (NVIDIA® Volta), HSW (Intel® Haswell), A64FX (Fujitsu® A64FX), Power9 (IBM® POWER9), TX2 (Marvell®ThunderX2), EPYC (AMD® EPYC).

Architecture Comparison: Power Efficiency



A64FX: Promising performance/watt

Fugaku Node: Single A64FX socket
GNU Fortran + MPI (48 ranks)

Note: Top Red (Volta V100), Pink (A64FX), Yellow (KNL)
Higher is better

* Plot of the efficiency metric normalized by power consumption on various hardware architectures. The legend includes a short descriptor for each architecture along with the number of parallel processes times (x) the number of threads and includes TDP in parenthesis. Specifically, the labels map as follows: KNL (Intel® Knights Landing), IB (Intel®Ivy Bridge), SKX (Intel® Skylake), V100 (NVIDIA® Volta), HSW (Intel® Haswell), A64FX (Fujitsu® A64FX), Power9 (IBM® POWER9), TX2 (Marvell®ThunderX2), EPYC (AMD® EPYC).

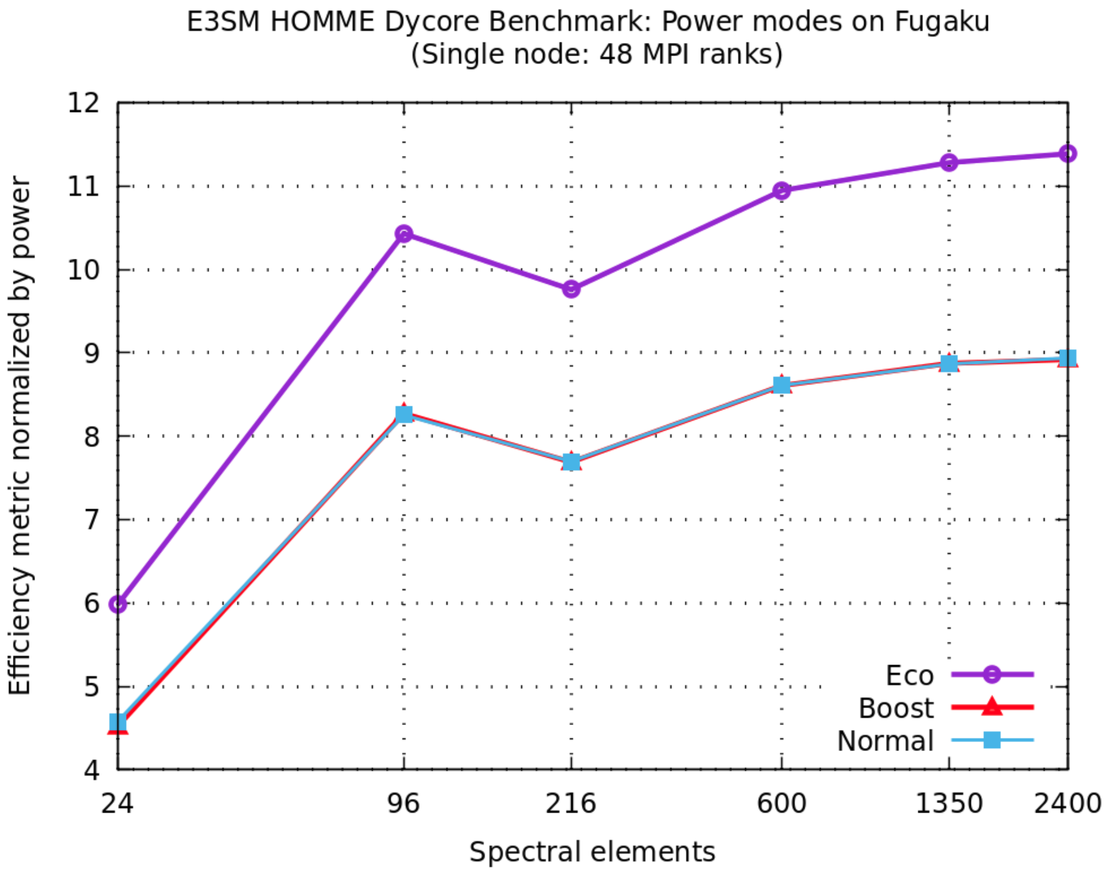
Power and Performance tradeoffs

- Power Efficiency metric normalized by the measured power on the compute node

$$E_{power\_measured} = \frac{E_{perf}}{measured\_power}$$

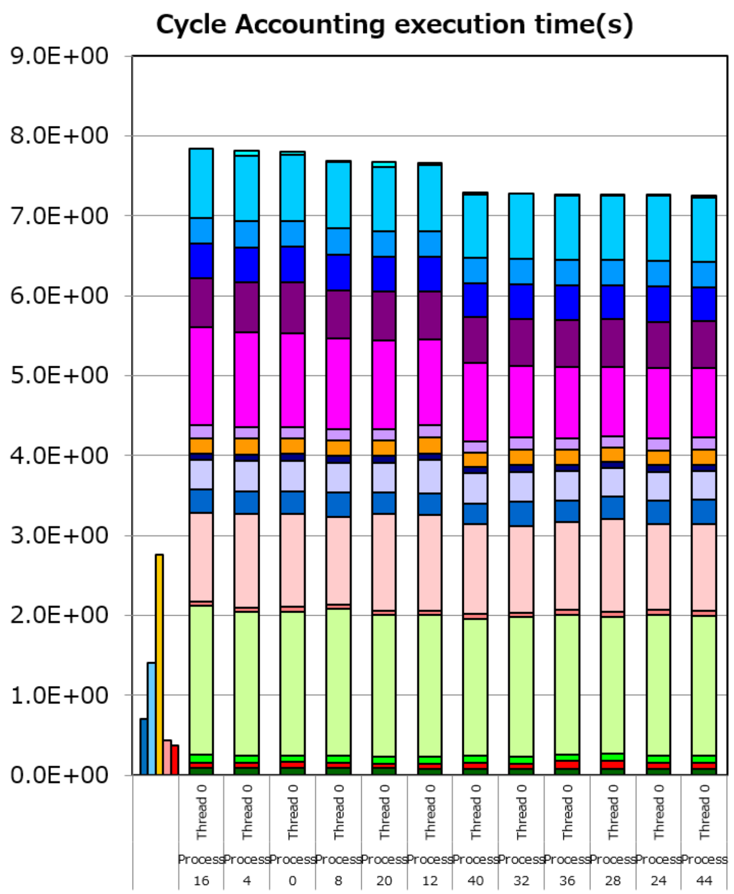
- PowerAPI
- Three modes
  - Normal (2 GHz)
  - Boost (2.2 GHz)
  - Eco (2 GHz/eco_state=2)
- Fortran version with GNU

Higher is better



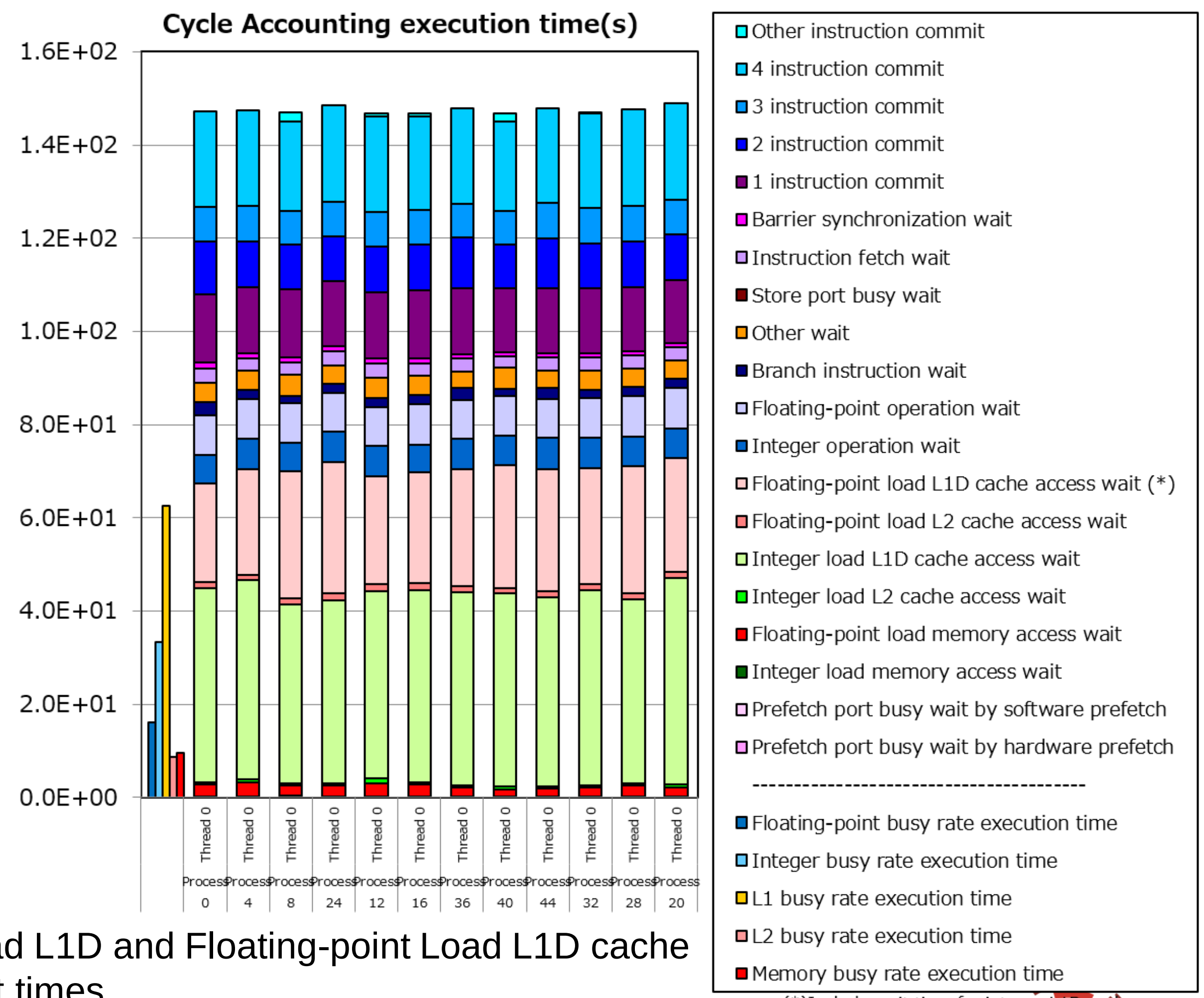
Performance Characterization: Instruction mix

Spectral Elements: 96



Spectral Elements: 2400

20 Categories

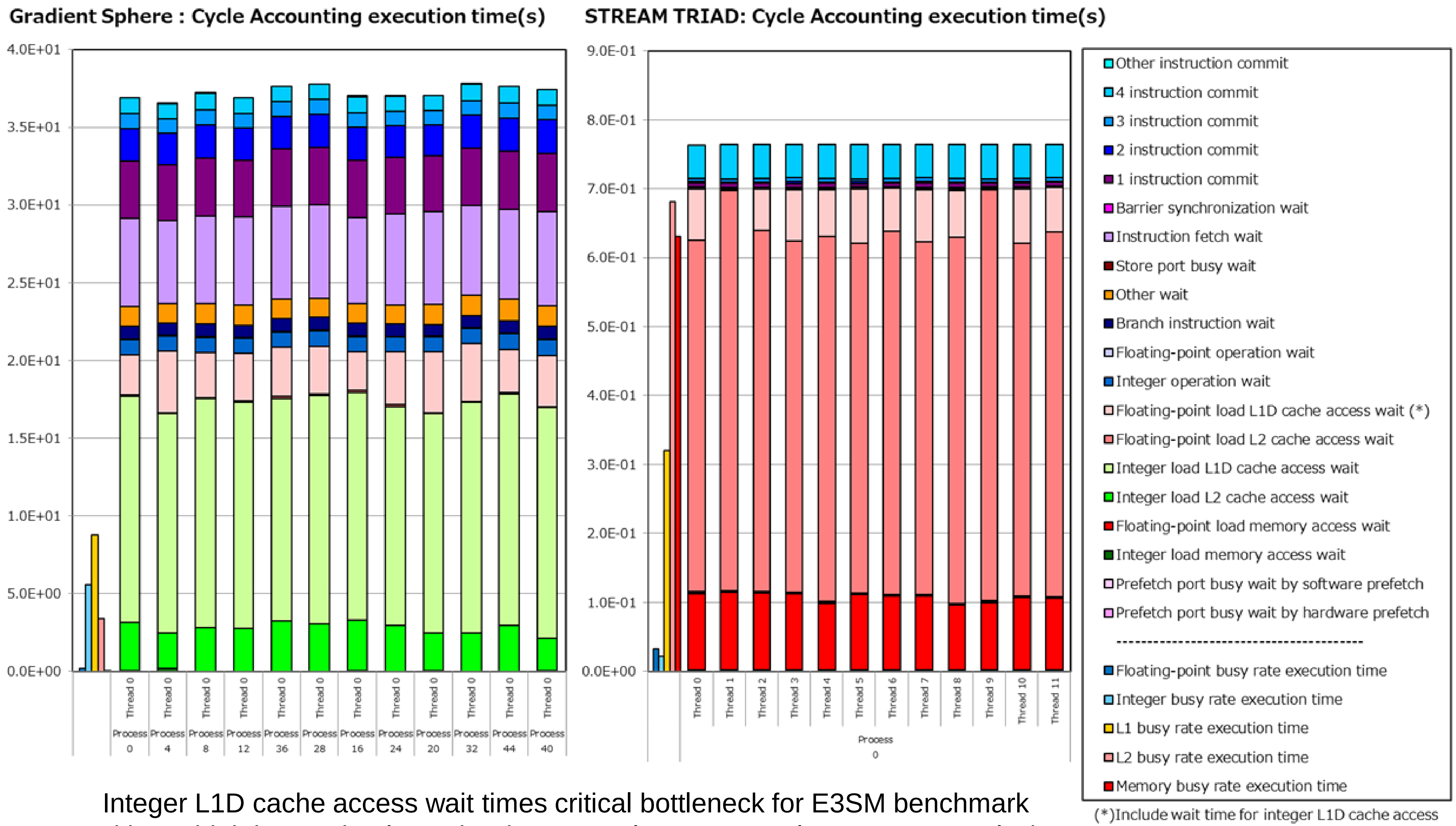


(*)Include wait time for integer L1D cache access

Significant fraction of runtime in the Integer Load L1D and Floating-point Load L1D cache access wait times

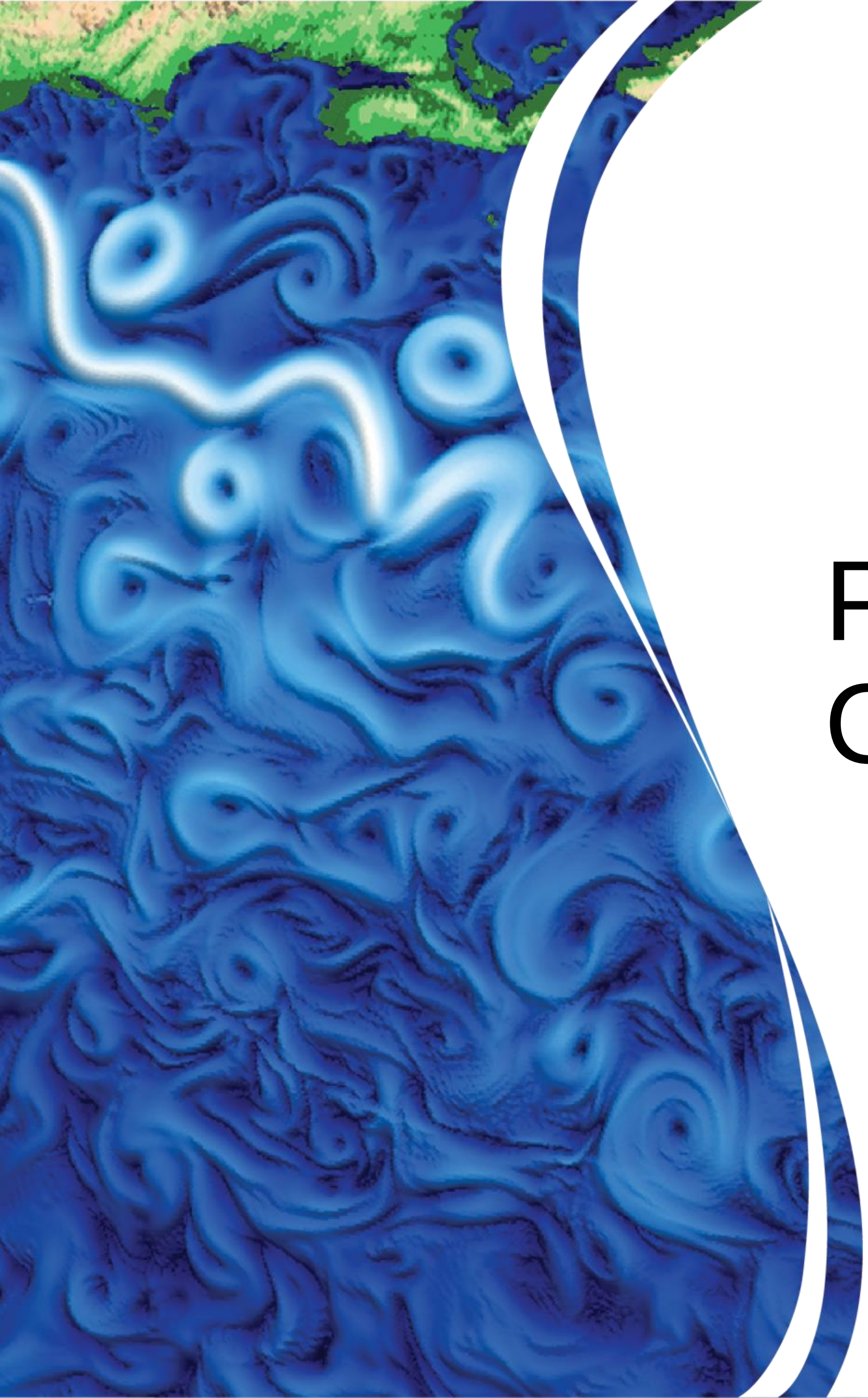
Left: pink section is Barrier synchronization wait

Instruction mix Comparison: Gradient sphere kernel vs. STREAM TRIAD



Integer L1D cache access wait times critical bottleneck for E3SM benchmark

Mitigate high instruction latencies (INT: 5 cycles, FP: 8 cycles, SVE: 11 cycles)



Energy Exascale Earth System Model

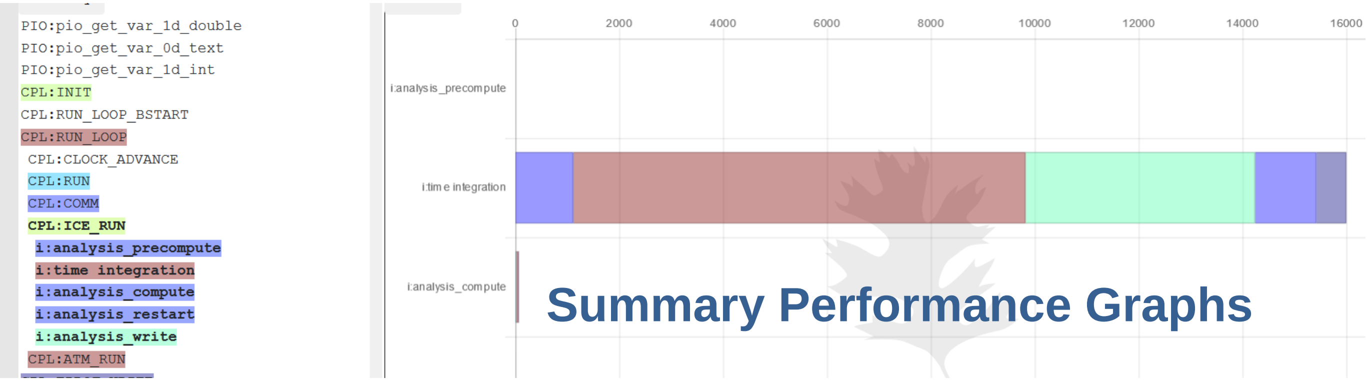
Performance Analytics for Computational Experiments

PACE

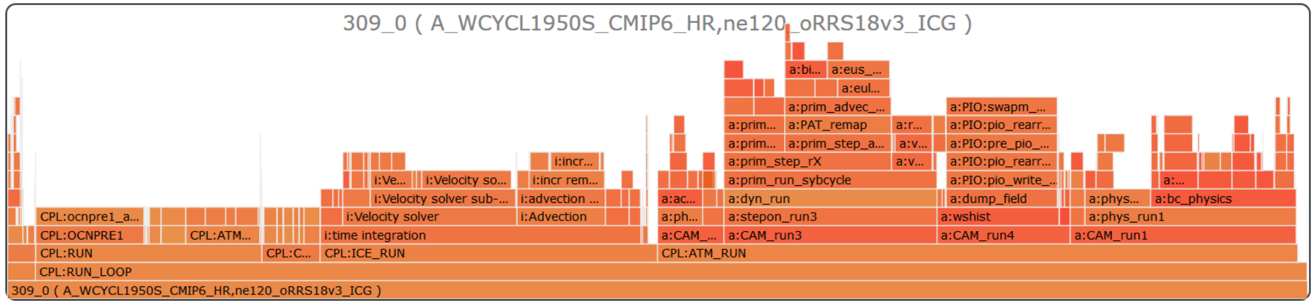
U.S. DEPARTMENT OF ENERGY

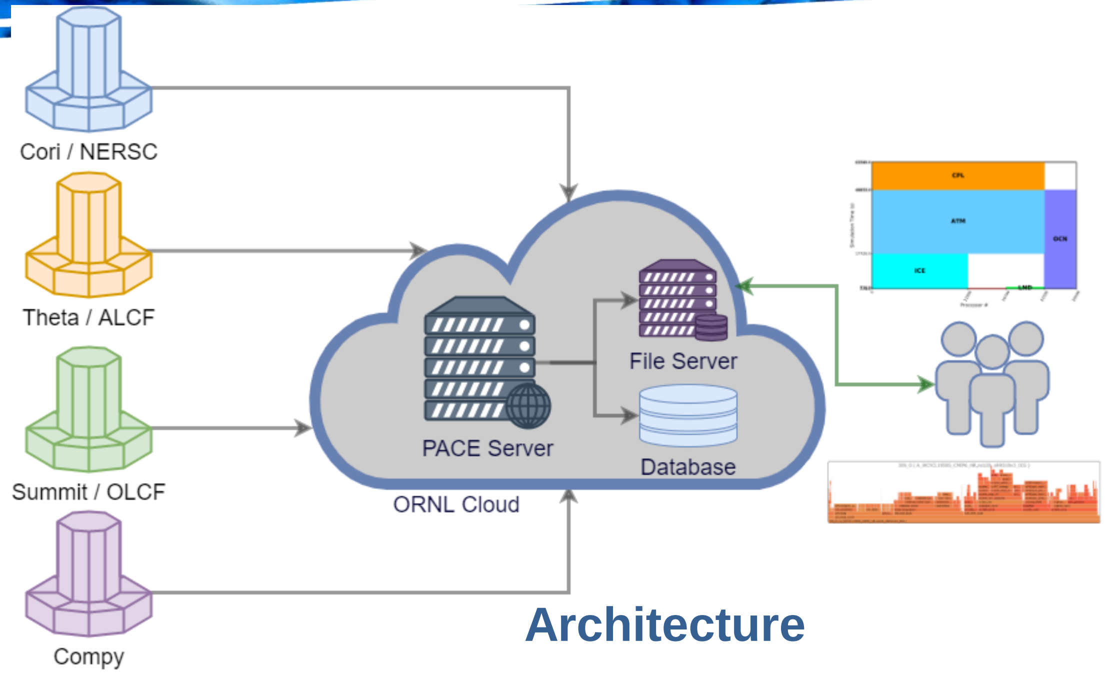
Summary

- Captures every E3SM experiment run on DOE supercomputers *automatically*
  - Performance Summary & Provenance
  - Facilitate performance research



Summary Performance Graphs



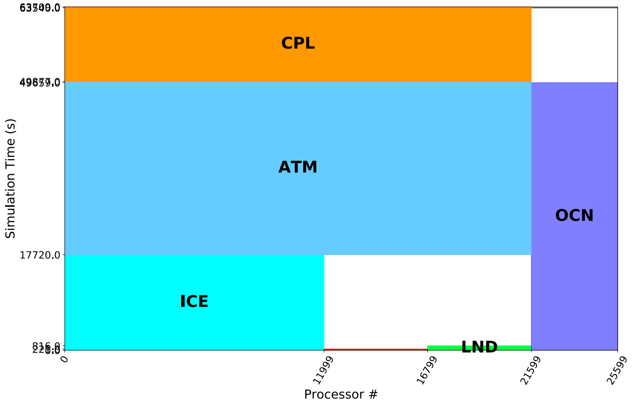


Architecture

Stats

- **130k experiments**
- **3+ million** input files
- 200+ users
- 14 platforms

Load Balancing

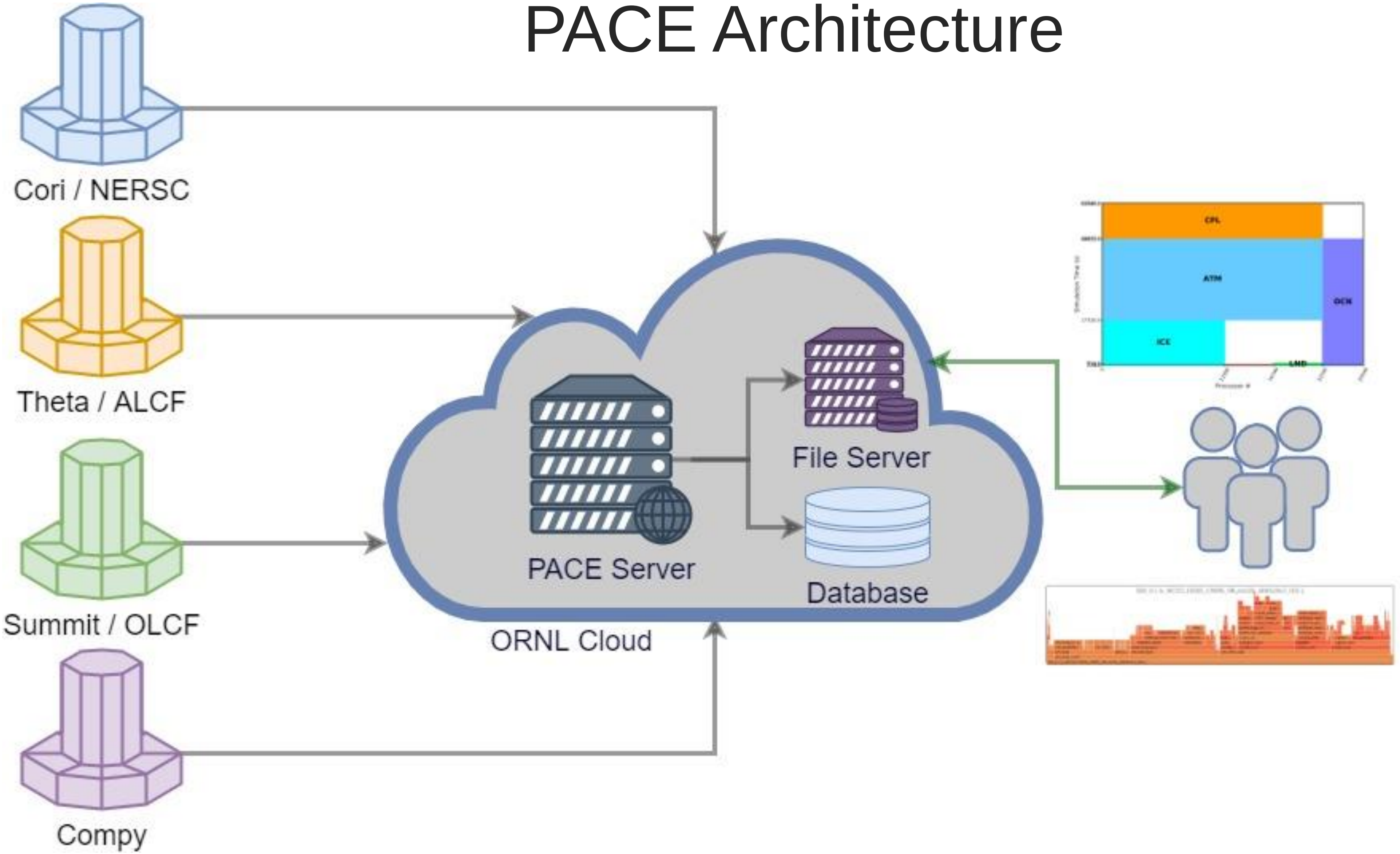


https://pace.ornl.gov

E3SM Performance Data

- Lightweight performance profiling by default
  - Utilizes General Purpose Timing Library (GPTL) timers
  - Mark start/stop at defined application phases
  - Aggregate statistics for parallel processes
  - Collect computation, communication and I/O performance data
  - Support for PAPI hardware counters
- Performance Archiving
  - Enabled on supported platforms at OLCF, ALCF, NERSC etc.
  - Archive performance data in project wide locations
  - Provenance data for context and reproducibility
  - System state and various logs

PACE Architecture

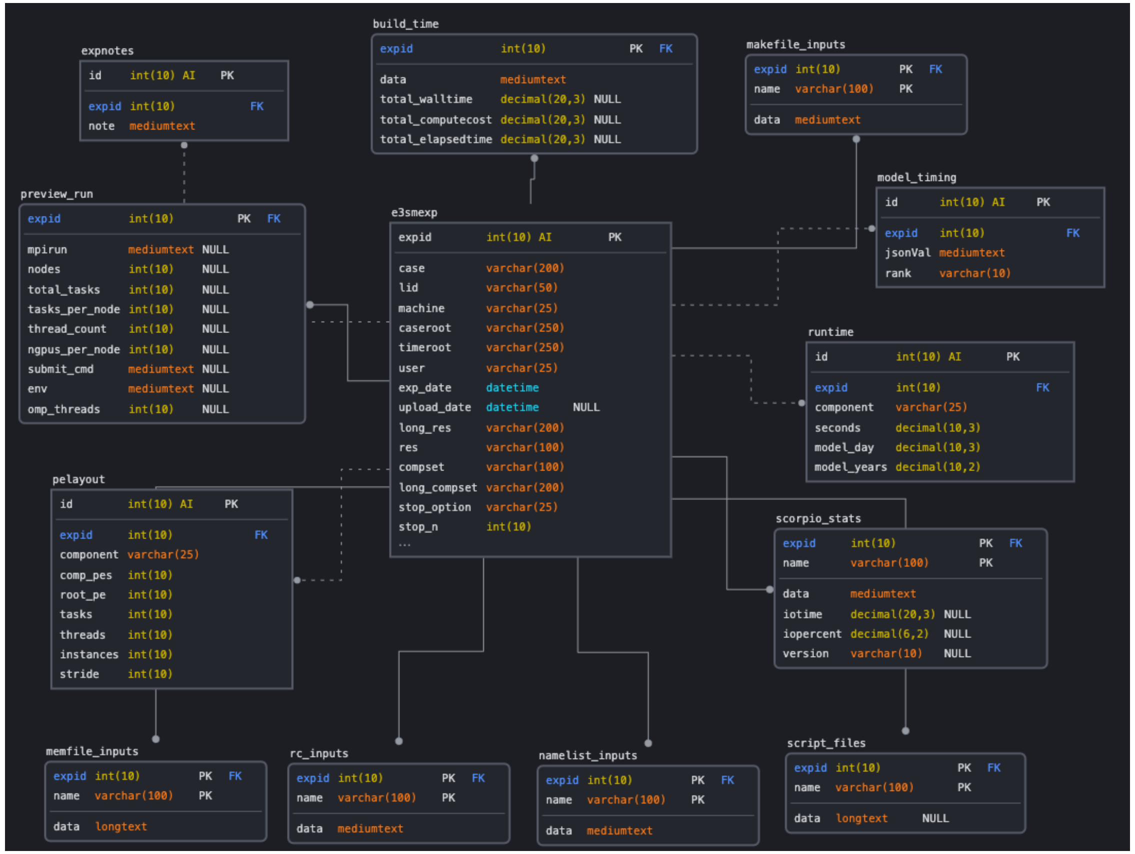


E3SM Energy Exascale Earth System Model

Technology Stack

- Infrastructure
  - ORNL Cloud (CADES)
  - OpenStack VM
- Nginx Web Server + Reverse Proxy
- Python-Flask middleware
  - Application Server
  - Process model inputs/timings
- Minio File Server
  - Object based storage for raw data
- MariaDB database
  - Structured and semi-structured data
  - Flexible Schema
- JavaScript
  - Frontend and visualization

Database Schema



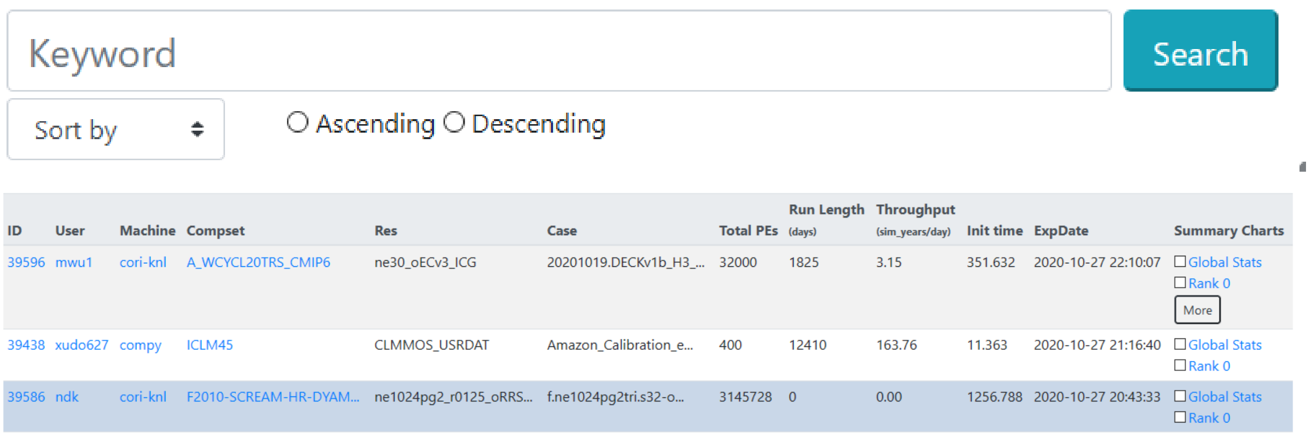
Last but definitely not least:
Cybersecurity compliance at a DOE lab

U.S. DEPARTMENT OF ENERGY

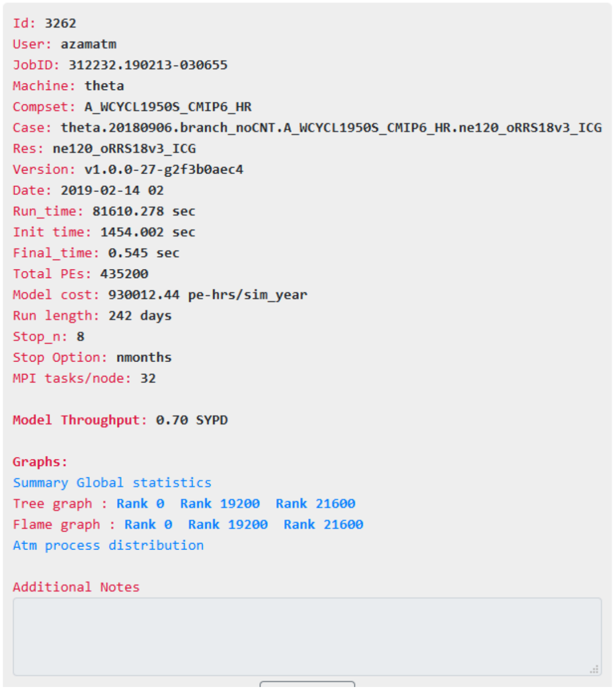
https://pace.ornl.gov

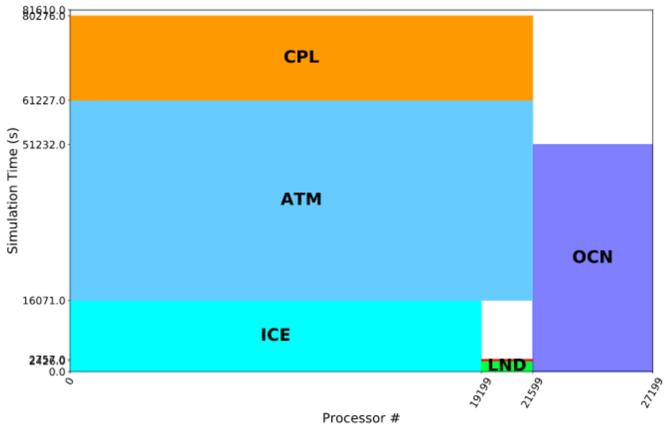
Usage

- Search for existing experiment using case, compset, grid, user etc. (Autocomplete supported)
  - Sort by desired criterion
- Click on a row from search results to dive into specific experiment
- Experiment details page contains
  - Metadata: user, machine, date etc.
  - Provenance: Browse model inputs
  - Performance overview
    - Model, Component throughputs
  - Process layout diagram
  - Links to detailed performance graphs



Experiment Details

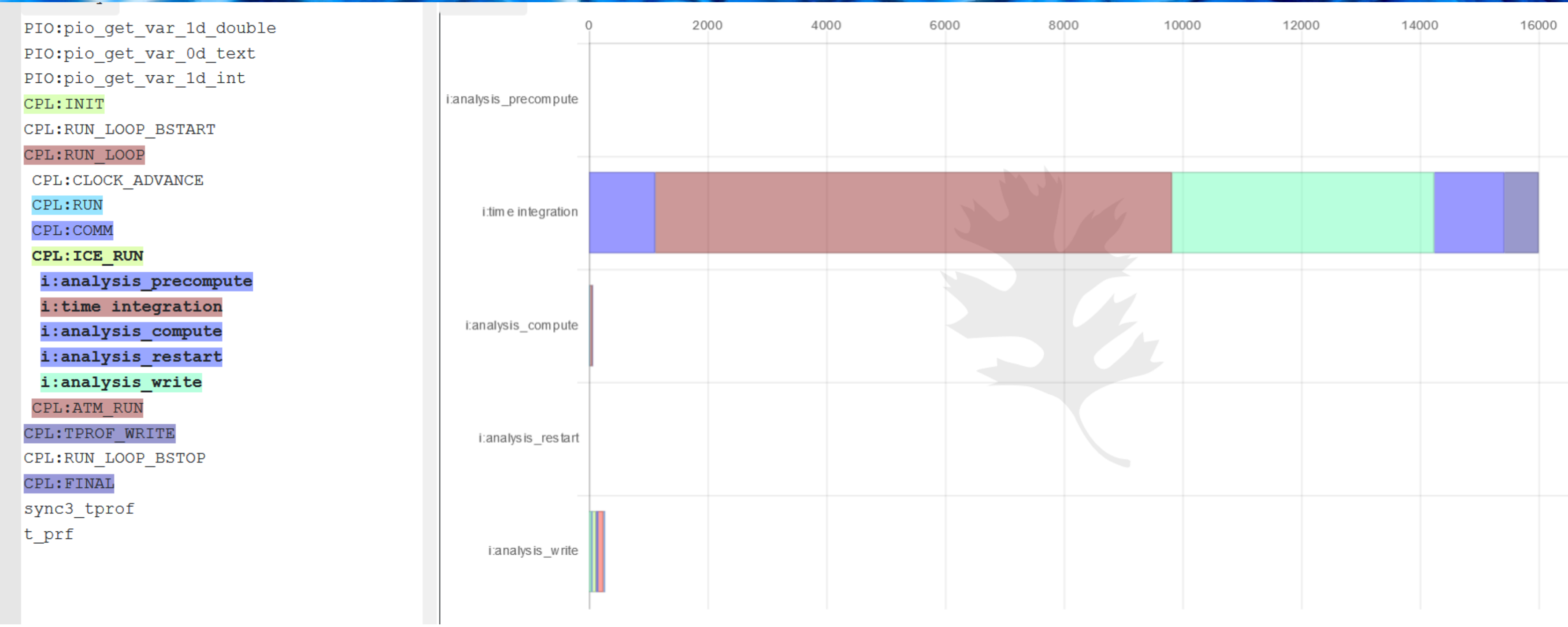




Tree Graph

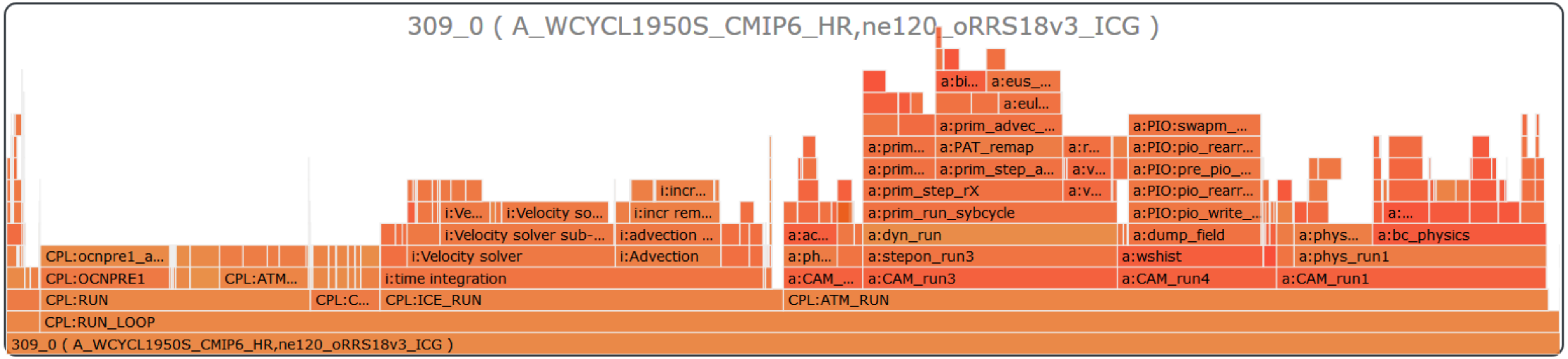
Summarize time taken by model components

Recursively explore time taken by model sub-regions



Flame Graph

High-level overview of a parallel process execution time

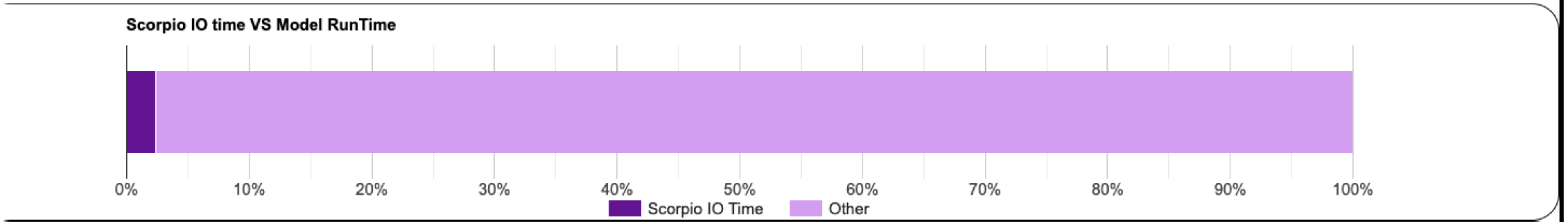


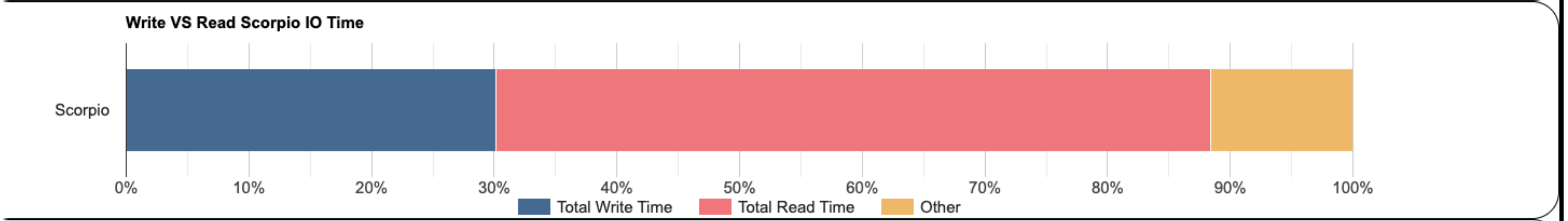
I/O Performance

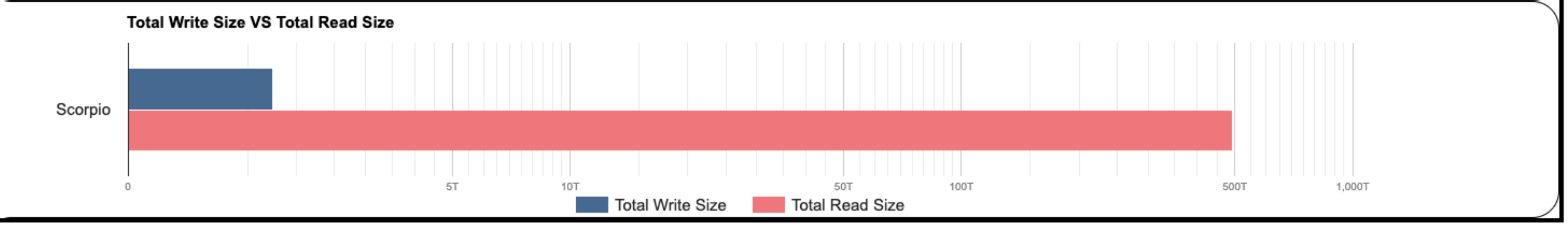
Overall I/O Summary Statistics

▶ Collapse

Total Run Time(s)	Total IO Time(s)	Total Write Time(s)	Total Read Time(s)	Avg Write Throughput(MB/s)	Avg Read Throughput(MB/s)	Total Write Size(GB)	Total Read Size(GB)	Na
140,791.72	3,285.97	991.81	1,911.73	1,673.52	246,476.79	1,740.44	494,084.65	Sc







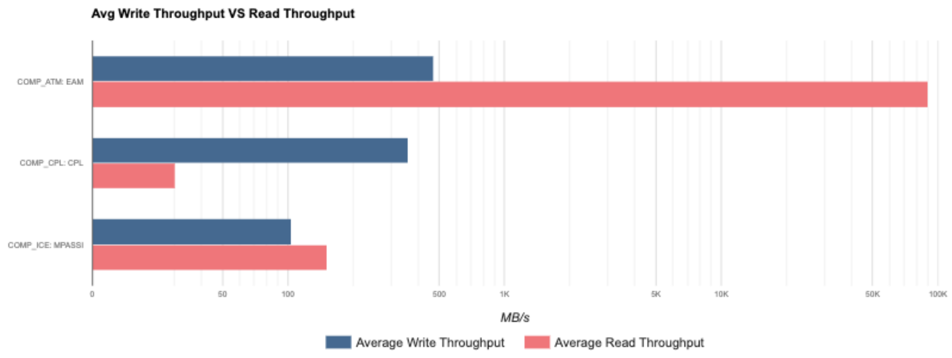
U.S. DEPARTMENT OF ENERGY

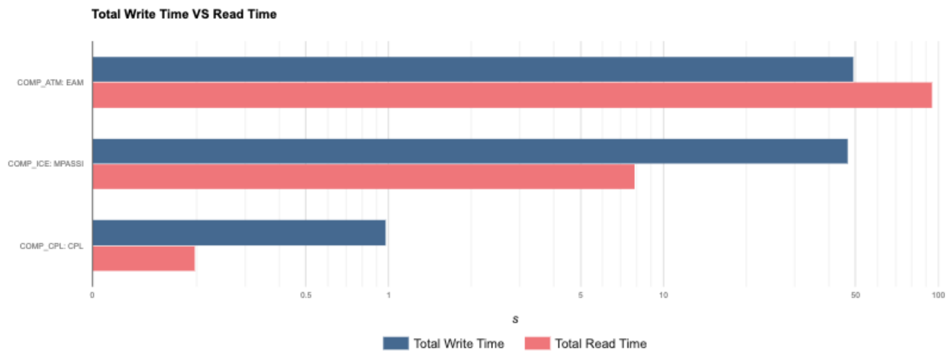
I/O Performance Details

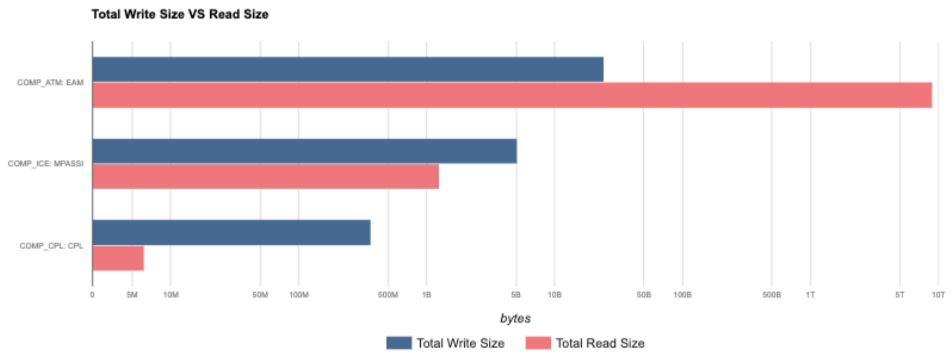
Model Component I/O Statistics

▸ Collapse

Total IO Time(s)	Total Write Time(s)	Total Read Time(s)	Avg Write Throughput(MB/s)	Avg Read Throughput(MB/s)	Total Write Size(GB)	Total Read Size(GB)	Name
147.65	49.34	95.06	470.36	90,044.44	24.33	8,975.11	COMP_ATM: EAM
56.87	46.79	7.85	103.50	152.12	5.08	1.25	COMP_ICE: MPASSI
1.29	0.98	0.20	359.36	30.09	0.37	0.01	COMP_CPL: CPL

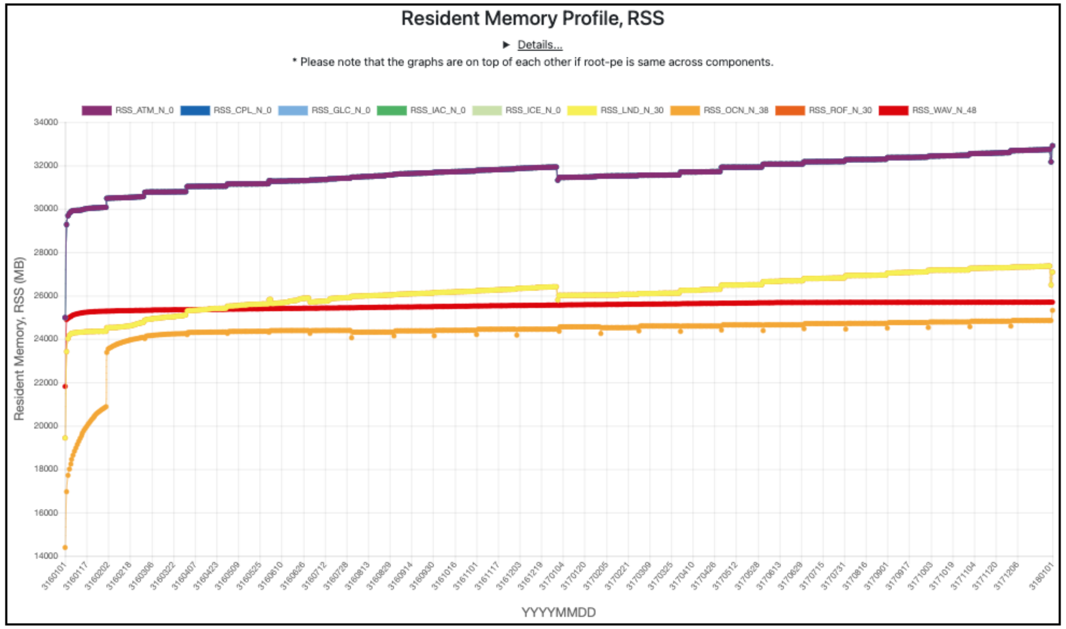


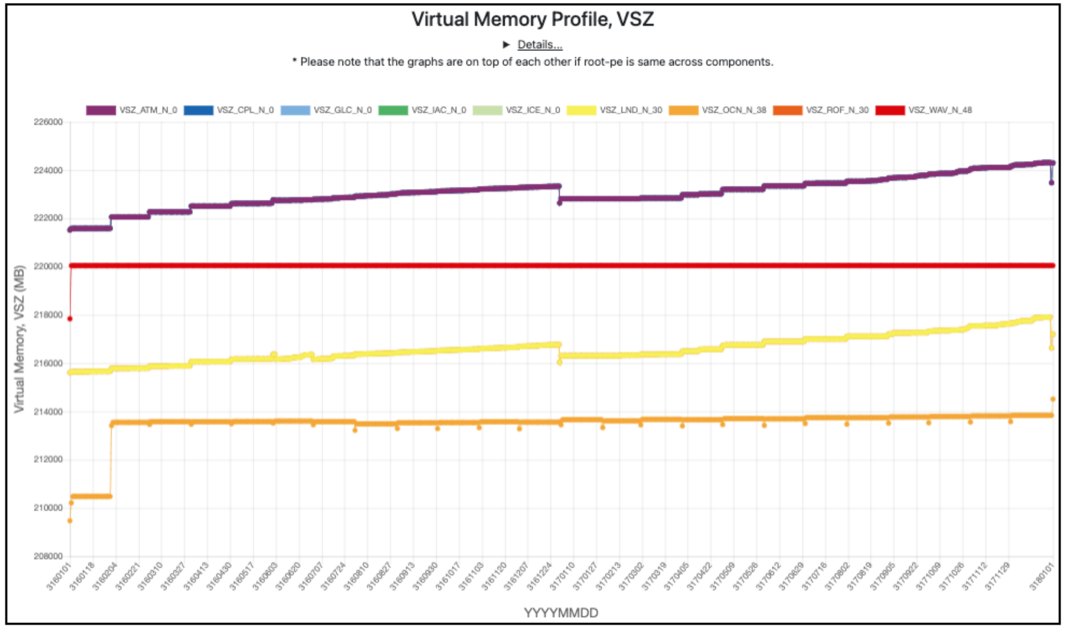




Memory Profiles

Performance Analytics for Computational Experiments (PACE)





E3SM Energy Exascale Earth System Model

Build Profiles

Build Times

Wall Time for build: 2039.520 sec.
Total Compute Cost of build: 2423.561 sec.
Note: This is the total cost associated with compilation across components.
Typically, the wall time is lower due to parallel builds.

▶ Details

Time (s)	File Name
2,039.520000	Total_Build
139.131729	CMakeFiles/lnd.dir/__/__/elm/src/data_types/VegetationDataType.F90.o
63.014300	CMakeFiles/ocn.dir/__/__/core_ocean/driver/mpas_ocn_core_interface.f90.o
60.025588	CMakeFiles/ice.dir/__/__/core_seaice/model_forward/mpas_seaice_core_interface.f90.o
52.475535	CMakeFiles/atm.dir/__/__/eam/src/physics/cosp2/local/cosp.F90.o
51.702540	CMakeFiles/lnd.dir/__/__/elm/src/data_types/ColumnDataType.F90.o
51.515716	CMakeFiles/lnd.dir/__/__/elm/src/biogeochem/CNCarbonFluxType.F90.o
35.808562	CMakeFiles/atm.dir/__/__/eam/src/physics/clubb/mt95.f90.o
27.578397	CMakeFiles/lnd.dir/__/__/elm/src/external_models/fates/main/FatesHistoryInterfaceMod.F90.o
25.040277	CMakeFiles/rof.dir/__/__/mosart/src/riverroute/RtmMod.F90.o

U.S. DEPARTMENT OF ENERGY

Ongoing and Future

Assistant

- Simulation planning
- Process layouts
- Data analytics
- Anomaly detection
- Allocation reports

Steve The Minion – from Pixabay
https://pixabay.com/photos/minions-banana-steve-the-minion-2552584/

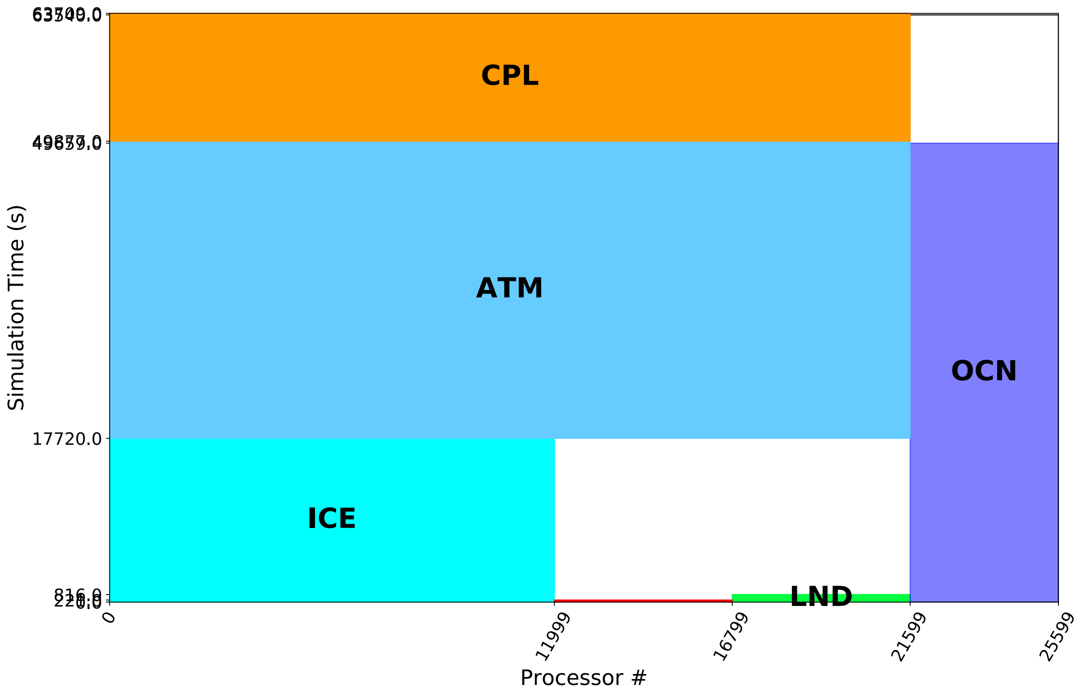
Wizard

- Recommend optimizations
- Optimal resource allocations
- Machine Learning
- Communication optimization
- Active monitoring and reporting

Dennis Jarvis from Halifax, Canada / CC BY-SA (https://creativecommons.org/licenses/by-sa/2.0)

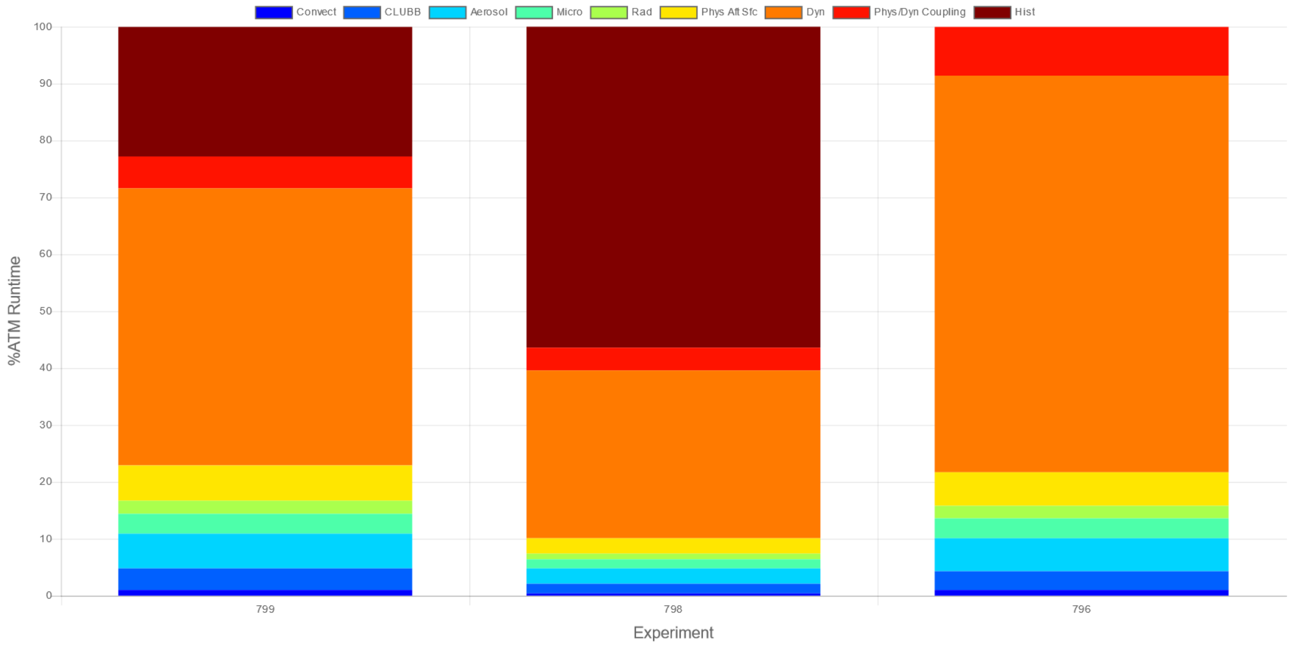
Performance Research Directions

Resource Allocation and Load Balancing



MPI Task Mapping

Targeted Optimization



Atmosphere model time distribution

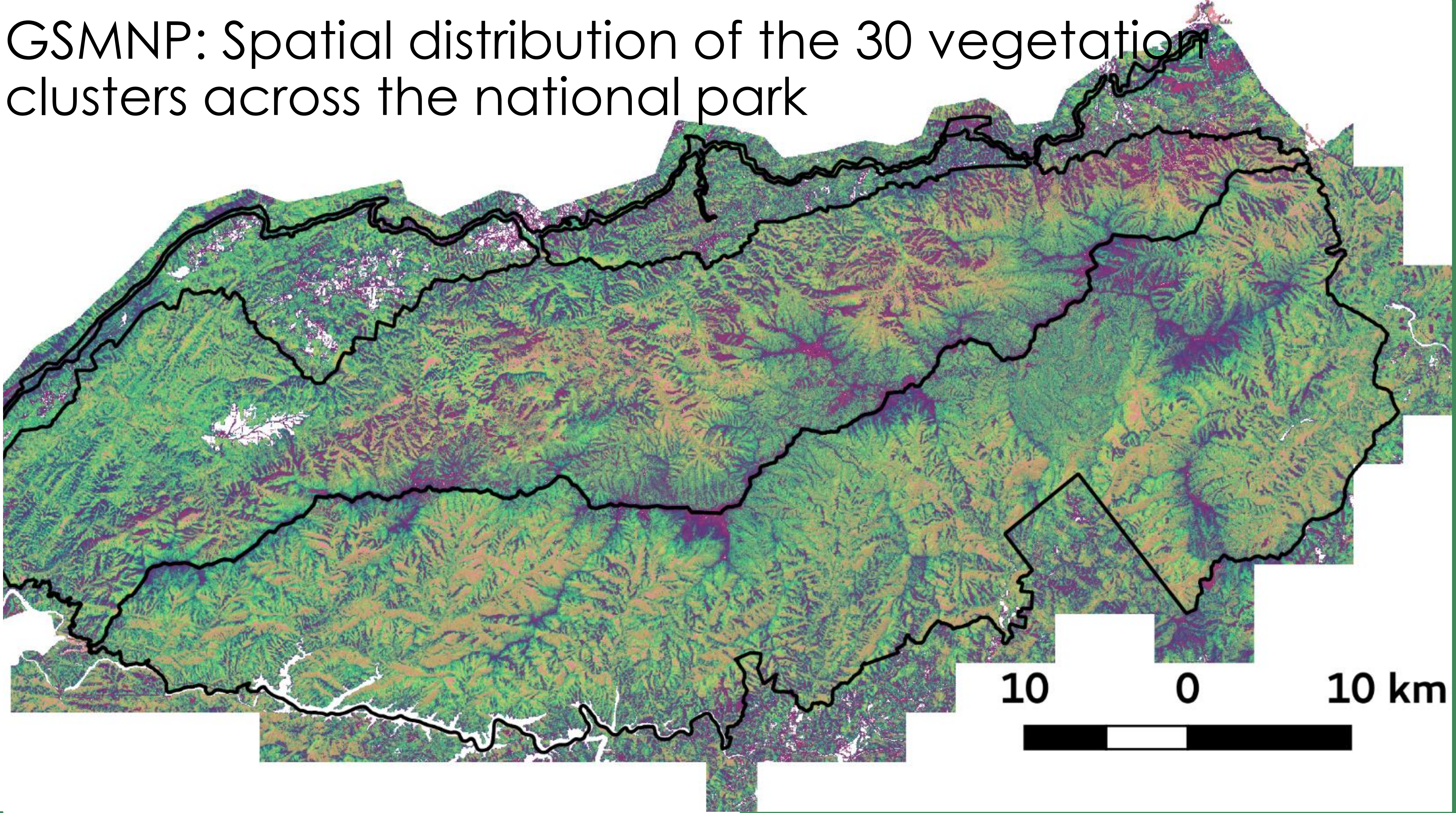
EarthInsights: Parallel Clustering of Large Earth Science Datasets on the Summit Supercomputer

Sarat Sreepathi[1], Jitendra Kumar[1], Forrest M. Hoffman[1], Richard T. Mills[2], Vamsi Sripathi[3], William W. Hargrove[4]

[1]Oak Ridge National Laboratory
[2]Argonne National Laboratory
[3]Intel Corporation
[4]USDA Forest Service

Adapted from talk given at
1st Workshop on Leveraging Artificial Intelligence (AI) in the Exploitation of Satellite Earth Observations & Numerical Weather Prediction(NWP)
NOAA, College Station, MD
April 23-25, 2019

GSMNP: Spatial distribution of the 30 vegetation clusters across the national park

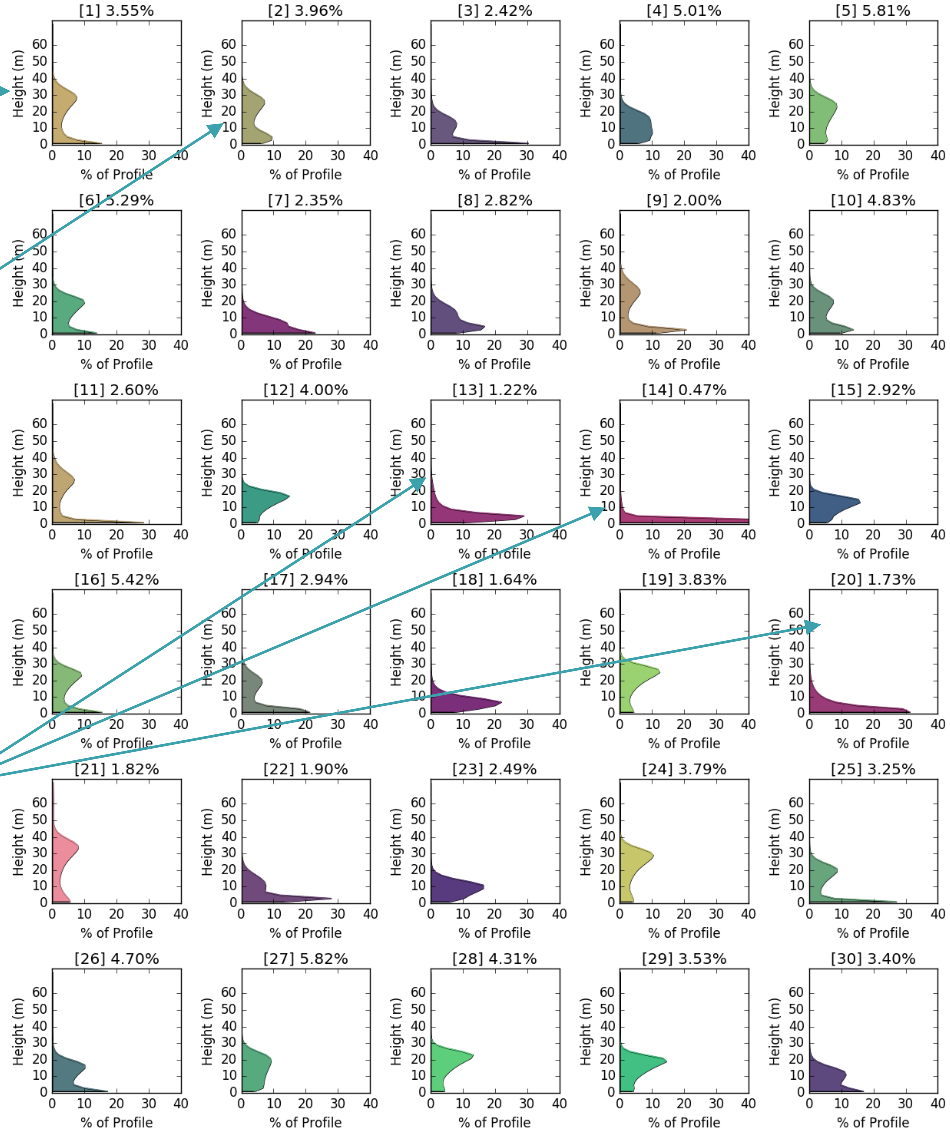


GSMNP: 30 representative vertical structures (cluster centroids) identified

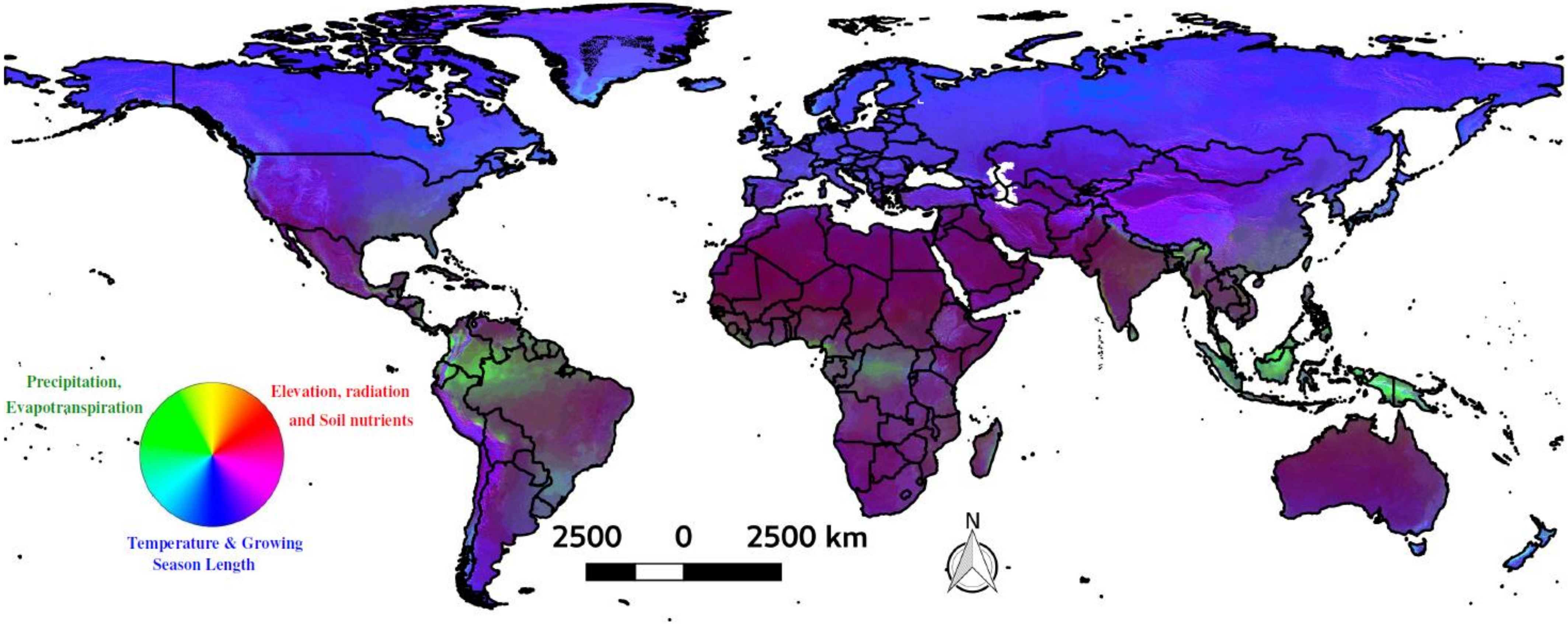
tall forests with low understory vegetation

forests with slightly lower mean height with dense understory vegetation

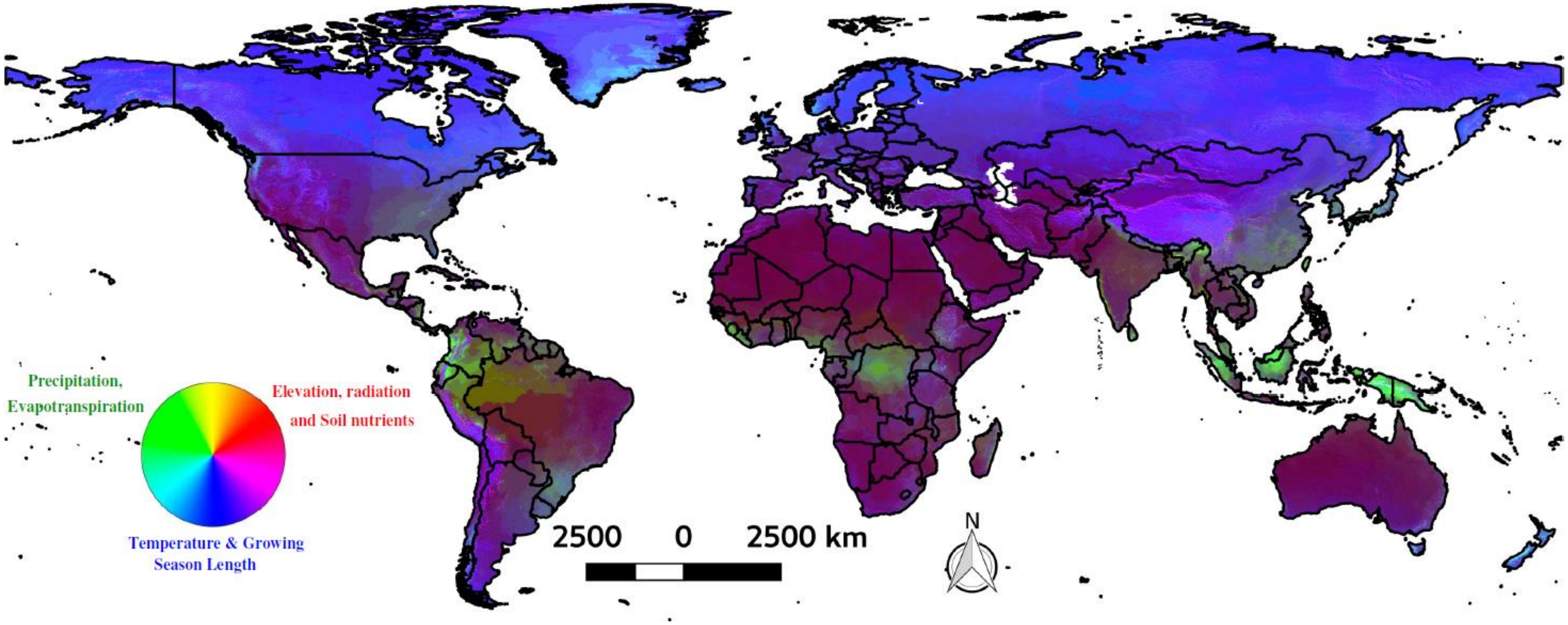
low height grasslands and heath balds that are small in area but distinct landscape type



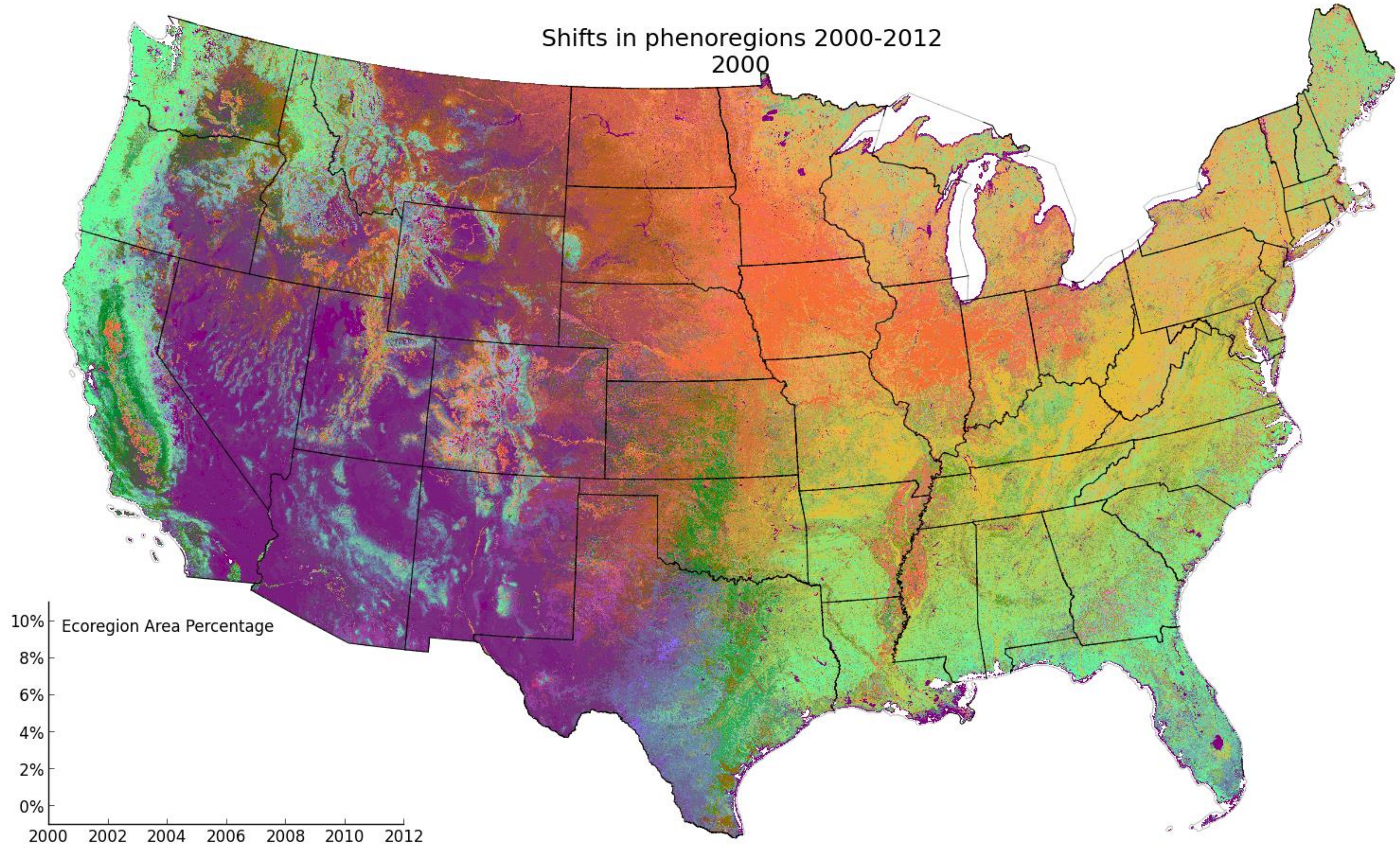
Global Climate Regimes: 1000 clusters Contemporary using Similarity color scheme



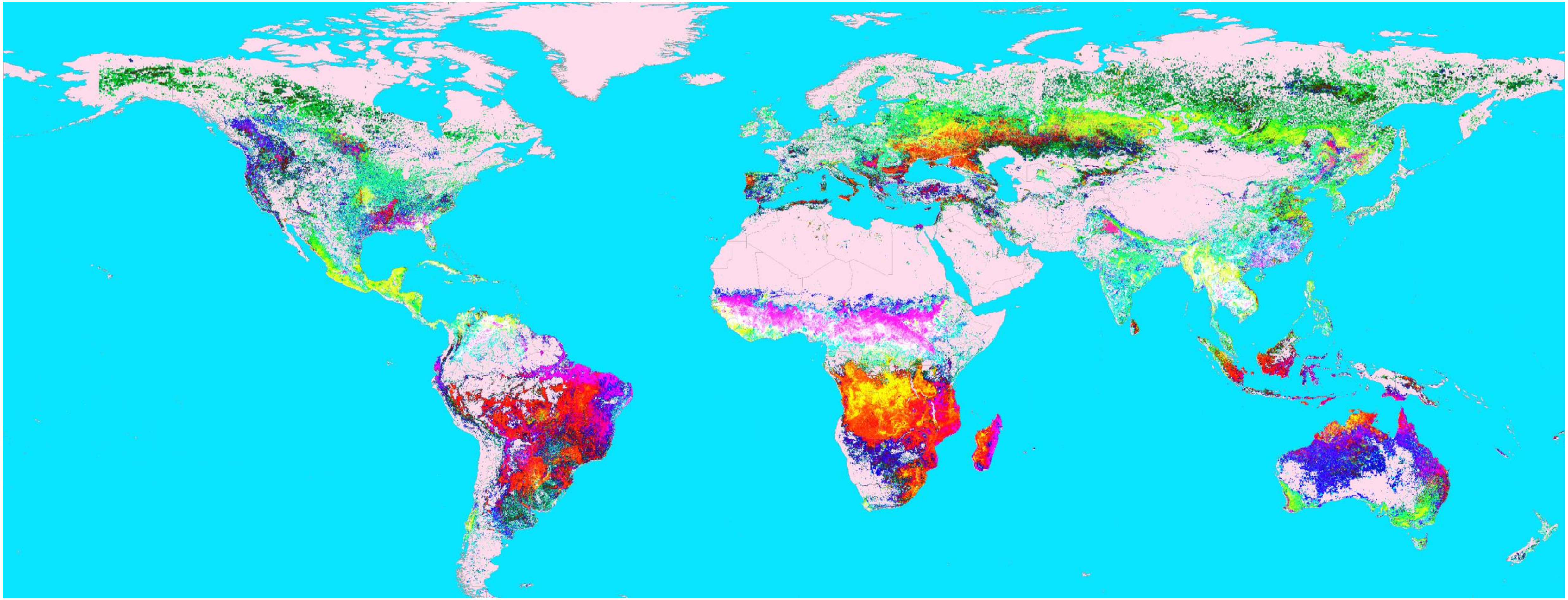
Global Climate Regimes: 1000 clusters 2100 using Similarity color scheme



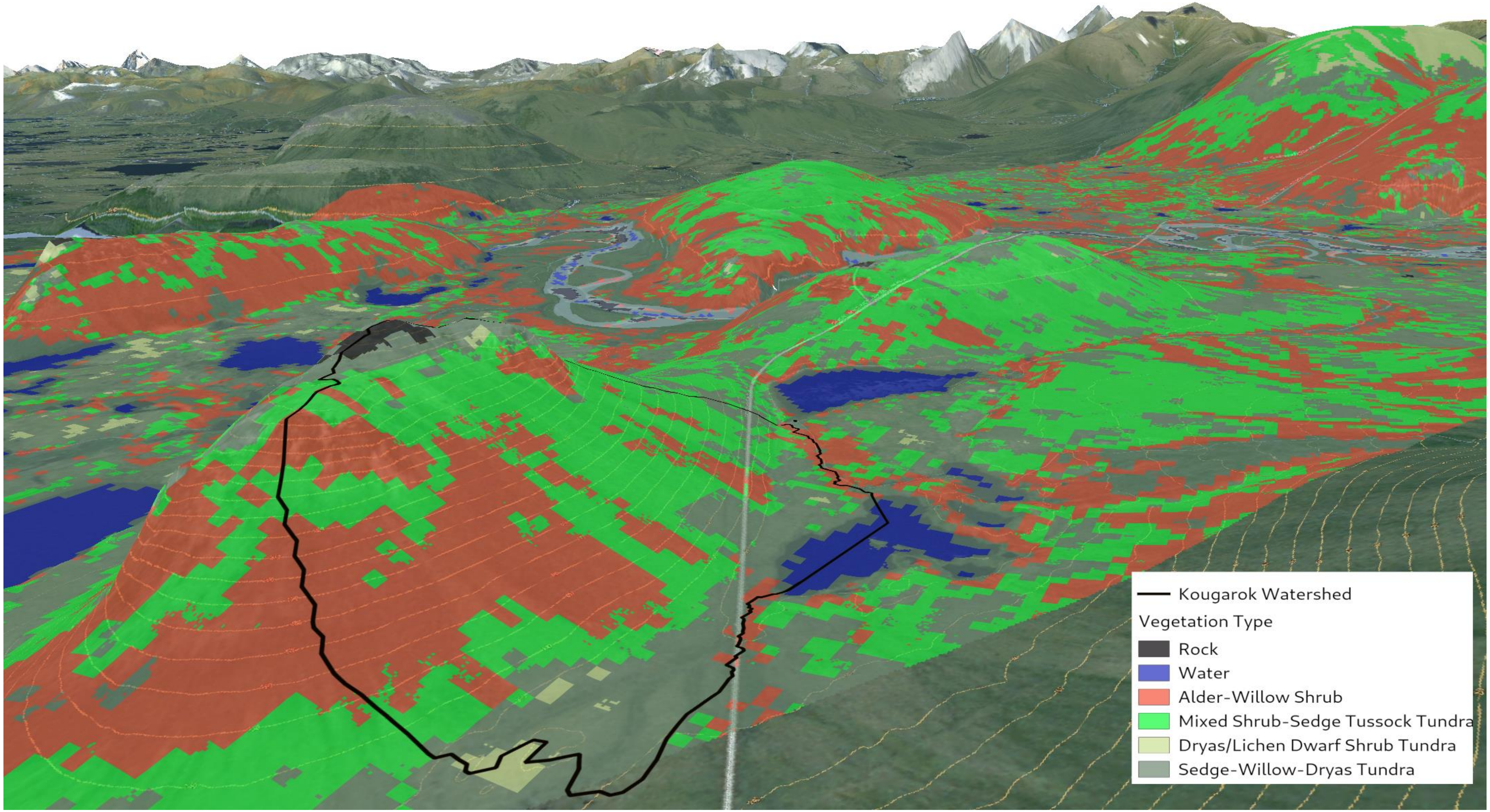
CONUS dynamic phenoregions



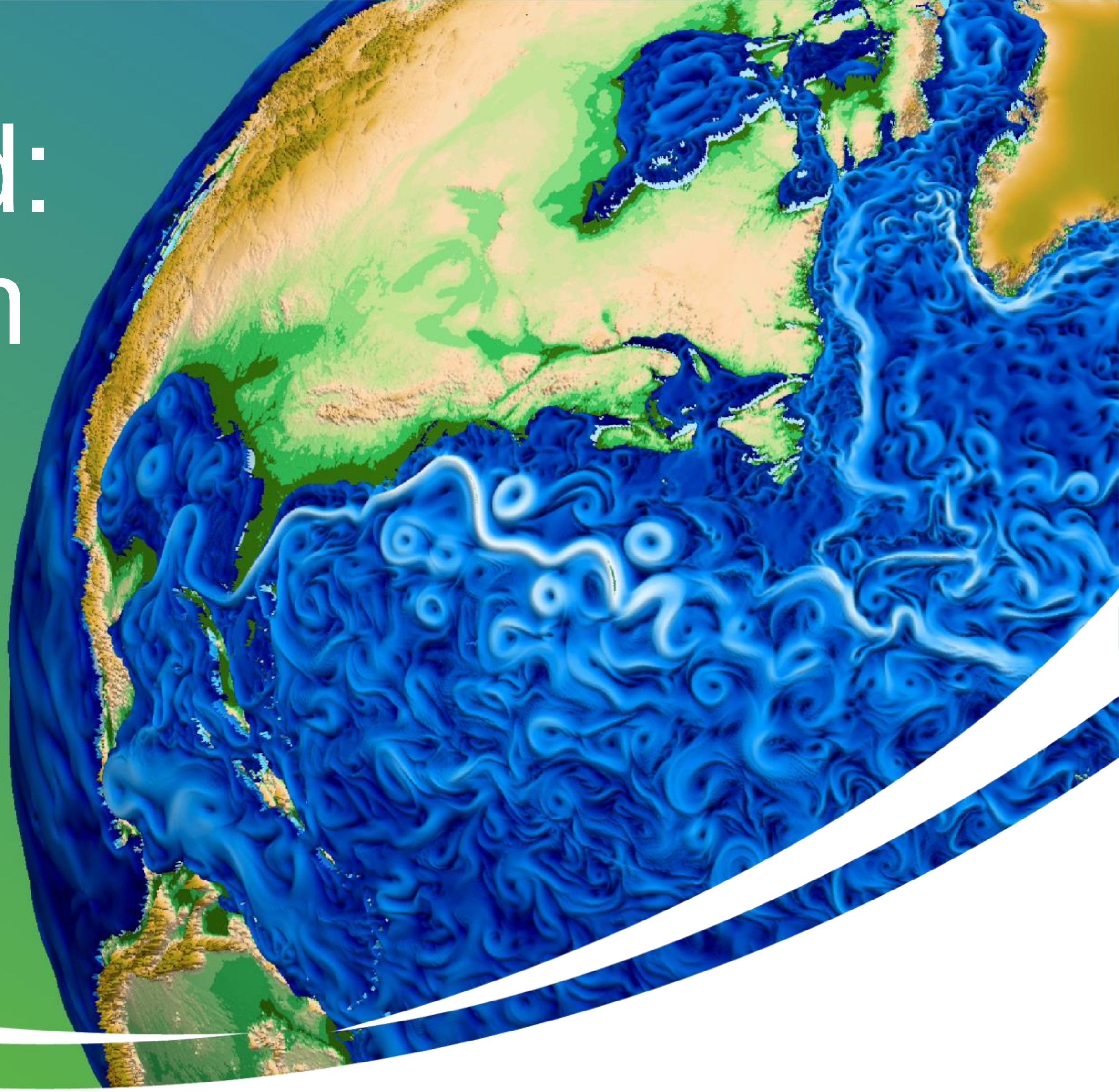
Global Fire Regimes



Arctic: High-resolution vegetation mapping



Exascale and Beyond: Application Co-design

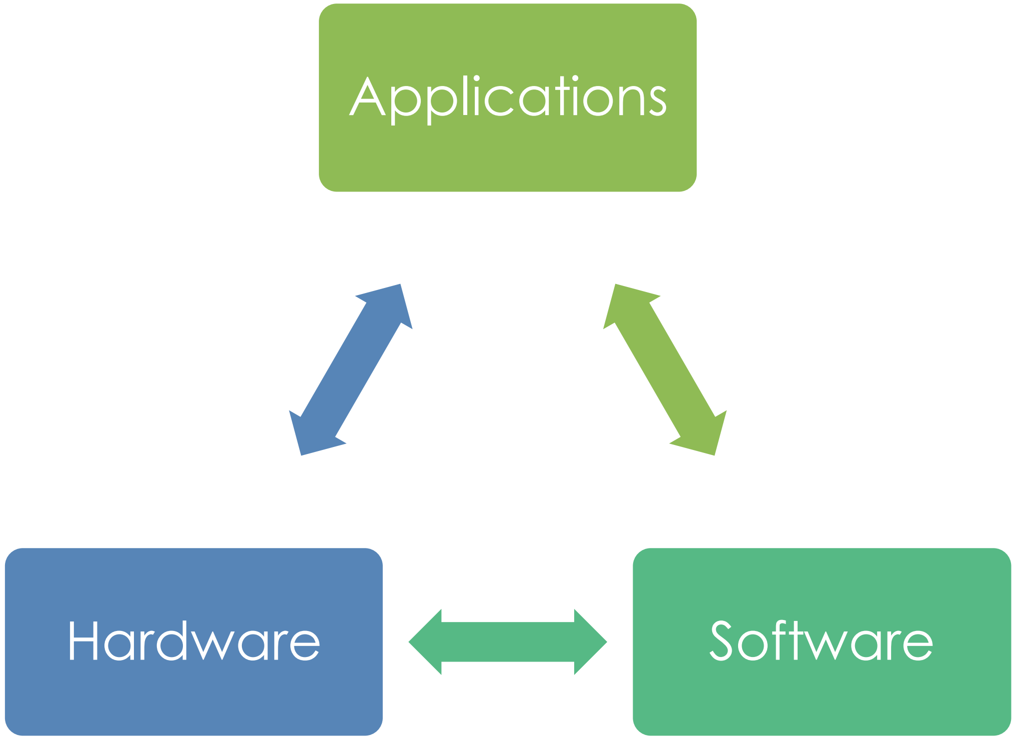


U.S. DEPARTMENT OF ENERGY

E3SM
Energy Exascale Earth System Model

Co-design

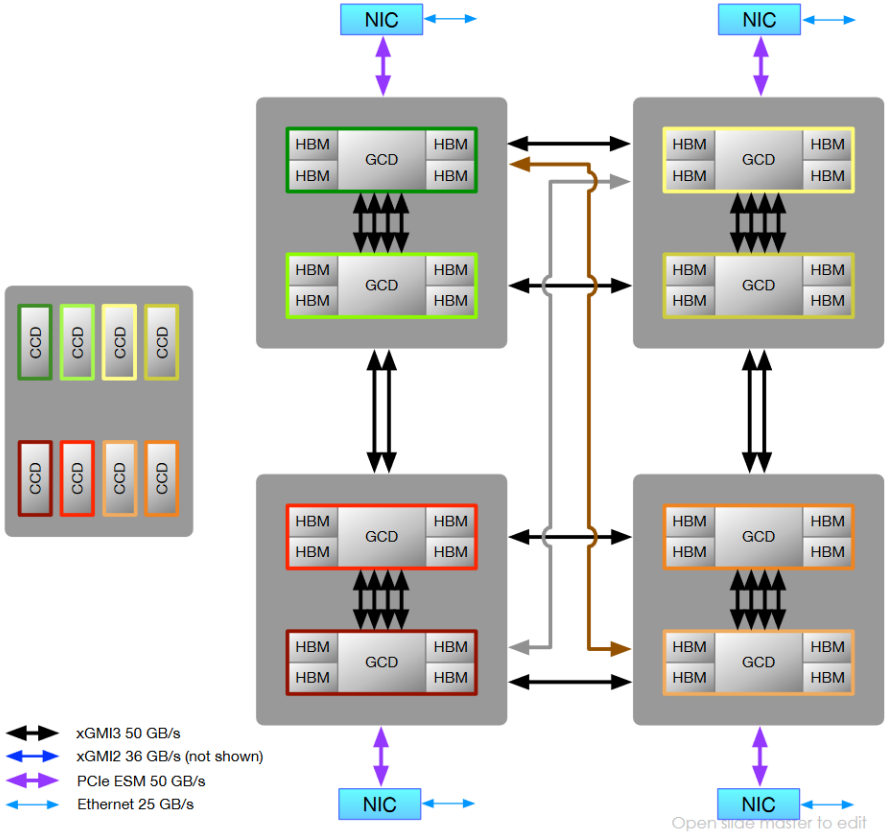
- Feedback loop between applications, system s/w and computer architecture
- Application requirements inform (influence?) hardware design
- Technology choices and constraints guide problem formulation and design of algorithms.



Today's Exascale: Frontier

- 1.1 Exaflops (FP64 – HPL Benchmark)
- 29 MW
- 4,000 ft^2
- 9,408 nodes
- Node
  - 4 AMD MI250X GPUs/node
  - Equiv. **8 logical GPU*s/node**
  - 512 GiB DDR4 (CPU) + **512 GiB HBM2e** (GPU)
  - **GPU Mem B/W**: 8x 1,635 GB/s (**13,080 GB/s Total**)
  - 1 AMD Trento CPU (64 cores)
- GPUs directly connected to high-speed interconnect

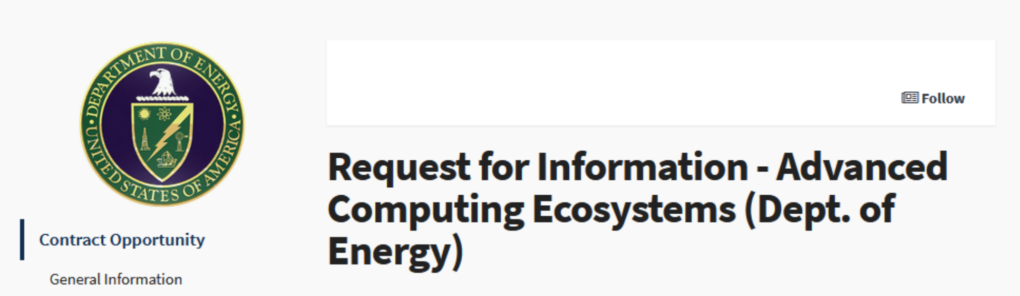
- **Aurora: Still under NDA**



Frontier Compute Node Architecture
1 CPU, 8 GPU*s

One cabinet of Frontier (24 ft^2) has higher HPL than all of Titan (4,500 ft^2) while using lower power (309 kW vs. 7 MW)

DOE Thinking



- DOE RFI – Summer 2022
  - Computing vendors and system integrators
  - Next generation supercomputers for 2025-2030 timeframe
- 10-20 FP64 exaflops in 2025 (8x from 2022)
- 100+ FP64 exaflops in 2030 (64x from 2022)
- 20-60 MW
- 4000 ft^2 (+ 50% more option)

Optional

- Upgradability: Every 1-2 years
- Emerging accelerators (Quantum,…) if feasible
- Hybrid: On-prem + Cloud

New Golden Age for Computer Architecture

- Increasing heterogeneity
- Hybrid chips (APU/XPUs)
- Divergence of AI and HPC

- Chiplet-based System-on-Chips
- Widespread HBM
- 3D stacking
- Low-power ARM (A64FX, Grace,…)
- RISC-V
- Processing in memory
- Silicon Photonics, Optical Interconnects

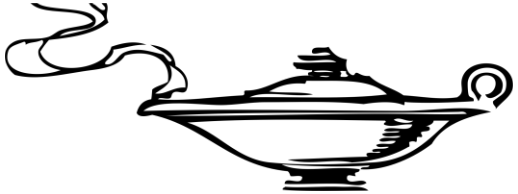
- Open Source Hardware
  - DARPA Electronic Resurgence Initiative (ERI)
- Numerous semiconductor startups

- Moore’s Law, Dennard Scaling
- Quantum: Optimization problems
  - Noisy Intermediate-Scale Quantum (NISQ): Practically useful?
- Neuromorphic: No clear fit
  - Spiking Neural Networks: Perhaps wavefront computations

Planning under uncertainty: A perspective

- Compute Architectures and Science: Friends or Frenemies?
  - Creativity for effective science
- High-end scientific computing: Leading vs. following
  - Cultivate and nurture vendor relationships
  - *Strategize ahead and influence vs. starting after general availability of an architecture*
- Co-design: Key application kernels and mini-apps
  - Impact on hardware: Skepticism warranted
  - Ray of hope (software/compilers)
- Changing Economics of Hardware Design
  - Fugaku ($1B incl. R&D), Frontier $600M procurement
- Wish: Imagine *relatively affordable* custom chips
  - Opinion: Better bet than fusion-powered quantum computers

(Domain-specific architectures)



Acknowledgments

U.S. Department of Agriculture, U.S. Forest Service,
Eastern Forest Environmental Threat Assessment Center.

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Contact:
sarat@ornl.gov