# NERSC update

## ~~and the Next Procurement - NERSC-10~~

HPC User Forum 2024
Reston VA

Nick Wright
Chief Architect
& Advanced Technologies Group Lead
10th April 2024

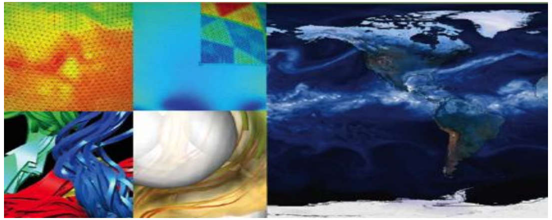
# NERSC update - Power Usage Characteristics of Perlmutter and Implications for Future Procurements

HPC User Forum 2024
Reston VA

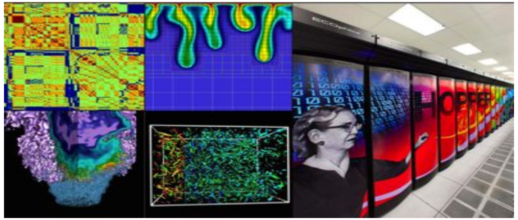
Nick Wright
Chief Architect
& Advanced Technologies Group Lead
10th April 2024

# NERSC: Mission HPC for DOE Office of Science Research

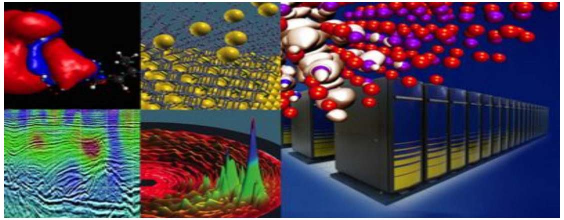
Largest funder of physical science research in the U.S.



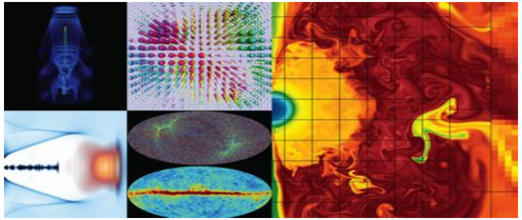
Biological and Environmental Research



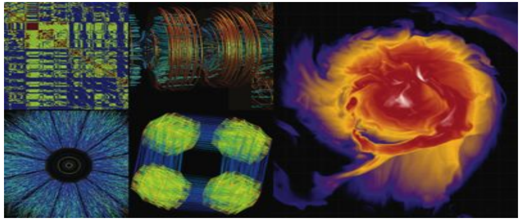
Computing



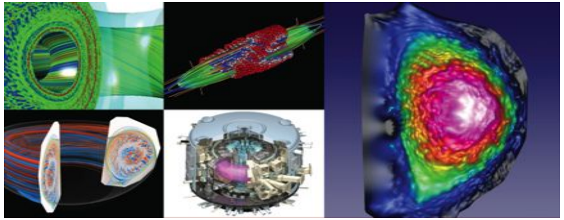
Basic Energy Sciences



High Energy Physics



Nuclear Physics



Fusion Energy, Plasma Physics

# Nobel-Prize Winning Users

*for the development of multiscale models for complex chemical systems*

2013 Chemistry

Martin Karplus

*for the discovery of the accelerating expansion of the Universe through observations of distant supernovae*

2011 Physics

Saul Perlmutter

*for the discovery of the blackbody form and anisotropy of the cosmic microwave background radiation*

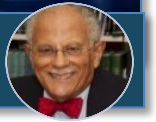
2006 Physics

George Smoot

*for their efforts to build up and disseminate greater knowledge about man-made climate change*

2007 Peace

Warren Washington



***for developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution***

2017 Chemistry

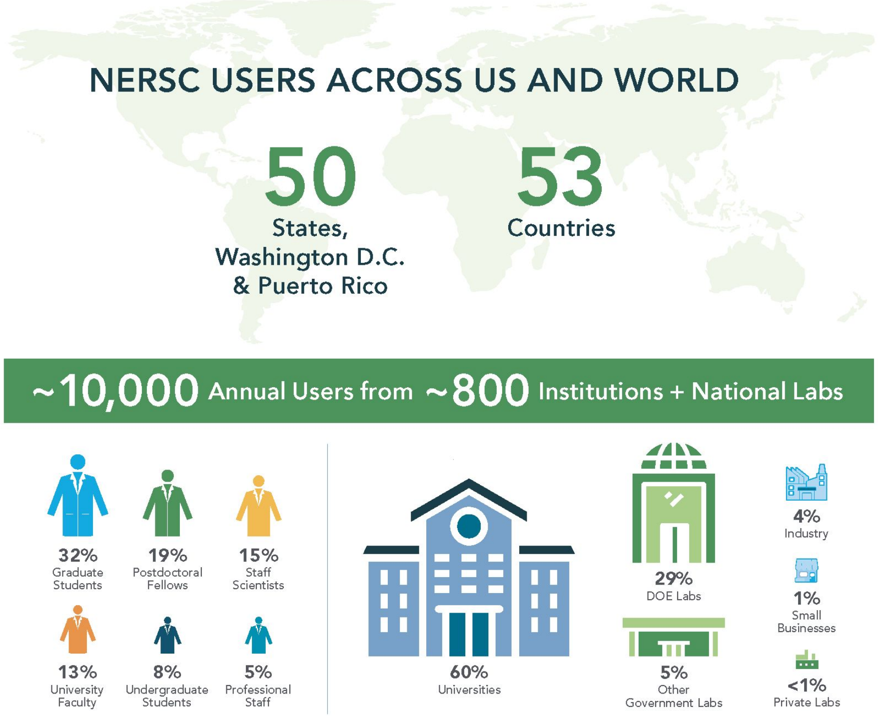
Joachim Frank

***for the discovery of neutrino oscillations, which shows that neutrinos have mass***

2015 Physics

SNO Collaboration

# NERSC by the Numbers

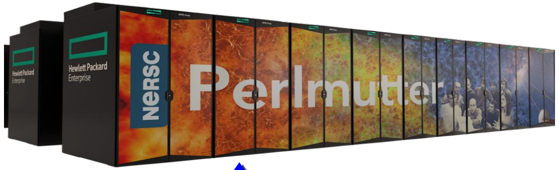


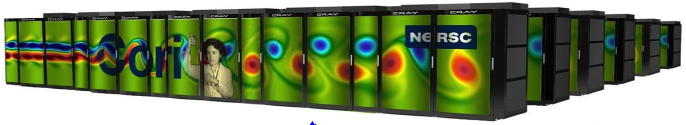


**NERSC has been acknowledged in 5,829 refereed scientific publications & high profile journals since 2020**

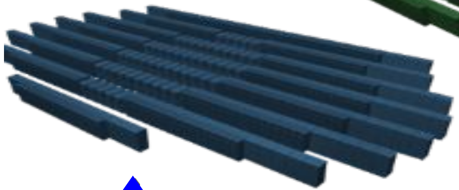
- Nature [32]
- Nature Communications [116]
- Proceedings of the National Academy of Sciences [55]
- Science [21]
- Nature family of journals [232]
- Monthly Notices of the Royal Astronomical Society [248]
- Physical Review B : Condensed Matter and Materials Physics [206]
- Physical Review D : Particles, Fields, Gravitation, and Cosmology [200]

# NERSC Systems Roadmap







**2013**
**NERSC-7: Edison**
Multicore CPU

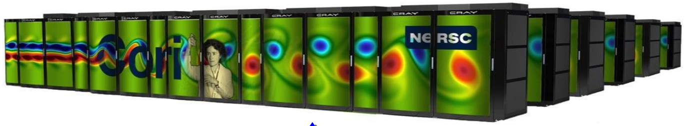
**2016**
**NERSC-8: Cori**
Manycore CPU
NESAP Launched: transition applications to advanced architectures

**2020**
**NERSC-9: Perlmutter**
CPU and GPU nodes
NESAP Expanded
Simulation, Learning & Data: Continued transition of applications and support for complex workflows

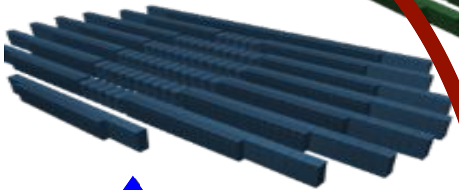
**2026**
**NERSC-10:**
Exa system
NESAP Workflows: Accelerating end-to-end workflows with technology integration

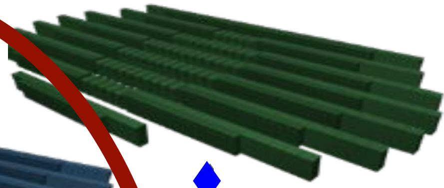
**2030+**
**NERSC-11:**
Beyond Moore

# NERSC Systems Roadmap







**2013**
**NERSC-7: Edison**
Multicore CPU

**2016**
**NERSC-8: Cori**
Manycore CPU
NESAP Launched: transition applications to advanced architectures

**2021**
**NERSC-9: Perlmutter**
CPU and GPU nodes
NESAP Expanded Simulation, Learning & Data: Continued transition of applications and support for complex workflows

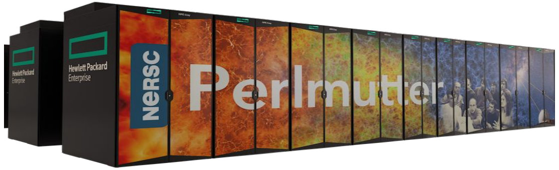
**2026**
**NERSC-10:**
Exa system
NESAP Workflows: Accelerating end-to-end workflows with technology integration

**2030+**
**NERSC-11:**
Beyond Moore

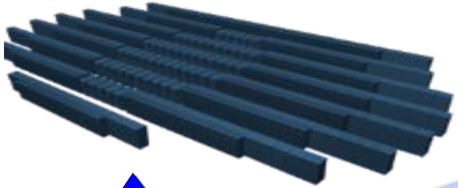
# NERSC-10 RFP is on the Street !

- RFP released 13th March
- Responses due 23rd April
- Delivery
  - 4QCY2026

# NERSC Systems Roadmap







**2016**

**NERSC-8: Cori**
**Manycore CPU**
**NESAP Launched: transition applications to advanced architectures**

**2020**

**NERSC-9: Perlmutter**
**CPU and GPU nodes**
**NESAP Expanded Simulation, Learning & Data: Continued transition of applications and support for complex workflows**

**2026**

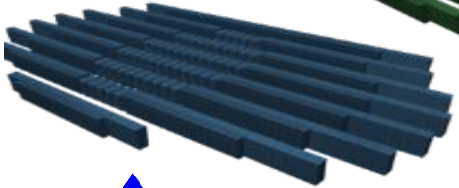
**NERSC-10:**
**Exa system**
**NESAP Workflows: Accelerating end-to-end workflows with technology integration**

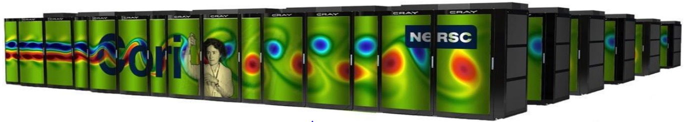
**2030+**

**NERSC-11:**
**Beyond Moore**

**Increased Energy Efficiency**

# NERSC Systems + Facilities Roadmap





**2013**
**NERSC-7: Edison**
Multicore CPU

**2016**
**NERSC-8: Cori**
Manycore CPU
NESAP Launched: transition applications to advanced architectures

**2020**
**NERSC-9: Perlmutter**
CPU and GPU nodes
NESAP Expanded Simulation, Learning & Data: Continued transition of applications and support for complex workflows

**2026**
**NERSC-10:**
Exa system
NESAP Workflows: Accelerating end-to-end workflows with technology integration (20 MW)

**2030+**
**NERSC-11:**
Beyond Moore

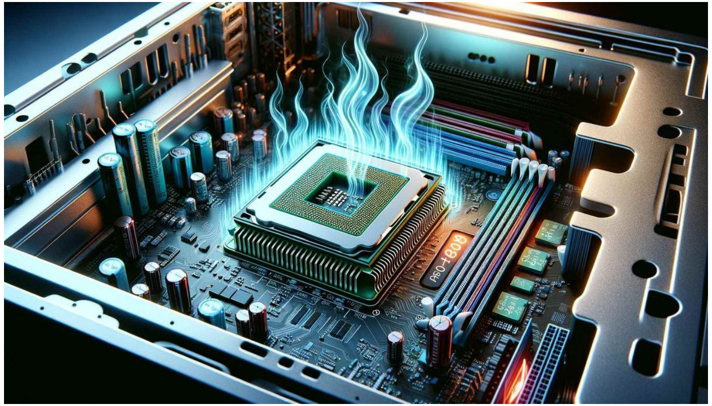
♦ NERSC Relocation from Oakland to Wang Hall

♦ Major Facility Power Upgrade (12.5 MVA)

♦ Facility Upgrade 2 (+10 MVA) & water-efficient cooling

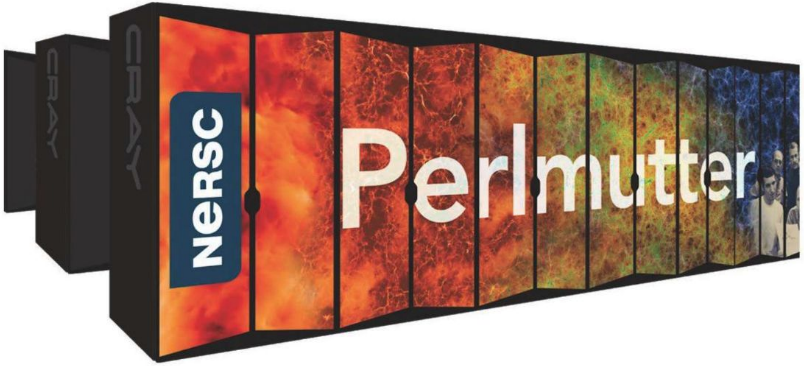
# What is Thermal Design Power (TDP)?

- Maximum power a computer chip, such as a CPU or GPU, can use in Watts
  - 100 - 1000 W range
- Can also have
  - Node TDP - sum of the max power of each component in a node
    - KW
  - Machine TDP - sum of the max power of each component in a machine
    - MW



# Perlmutter - HPE Cray EX System Based AMD Milan CPUs and NVIDIA A100 GPUs

- 1792 GPU accelerated nodes with 1x Milan CPU and 4x NVIDIA A100 GPUs;
- 3072 CPU nodes with 2x AMD Milan CPUs;
- Slingshot 11 interconnect
- 35 PB all Flash Lustre file system




| Thermal Design Power (TDP) | | | | |
|---|---|---|---|---|
| **CPU Socket** | **GPU Socket** | **CPU Node** | **GPU Node** | **System** |
| 280 W | 400 W | 823 W | 2,340 W | 6.9 MW |

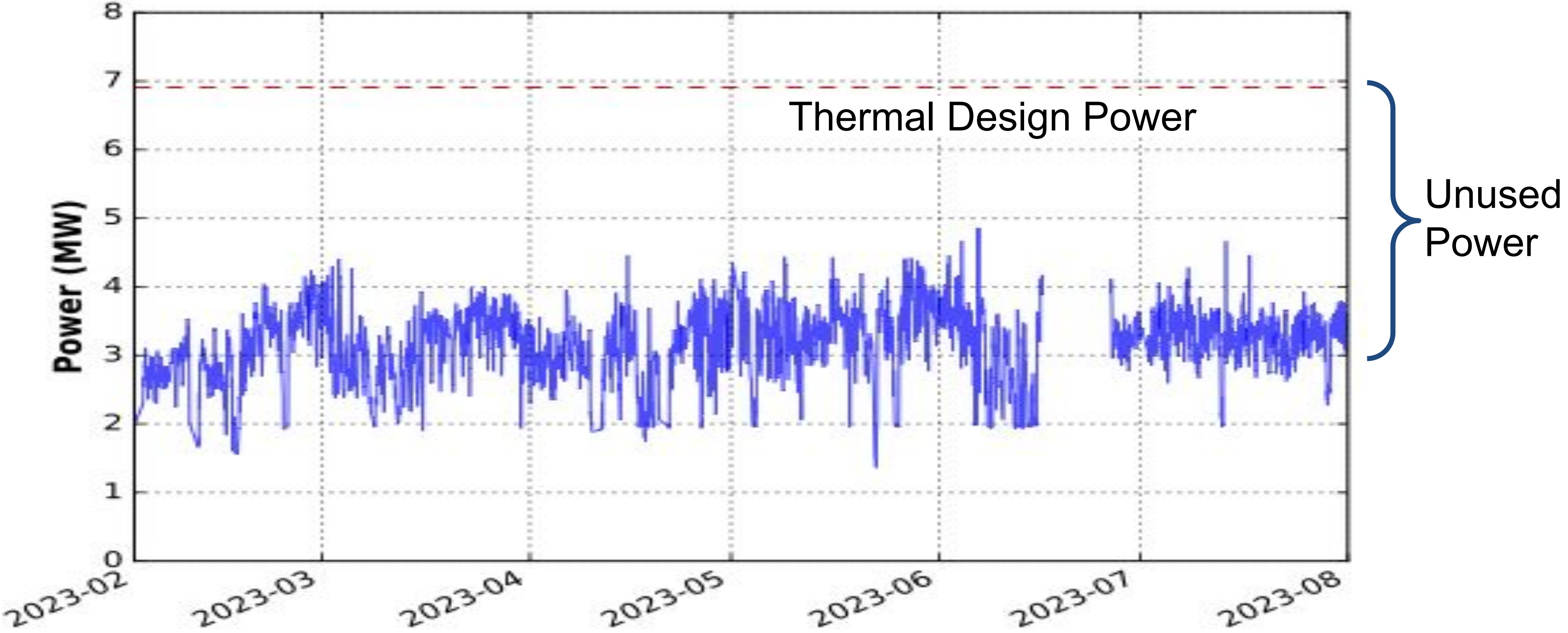
Power measurement sources:

- Cray power monitoring (PM) counters and NVIDIA DCGM - nodes
- Modbus - cabinets, substations

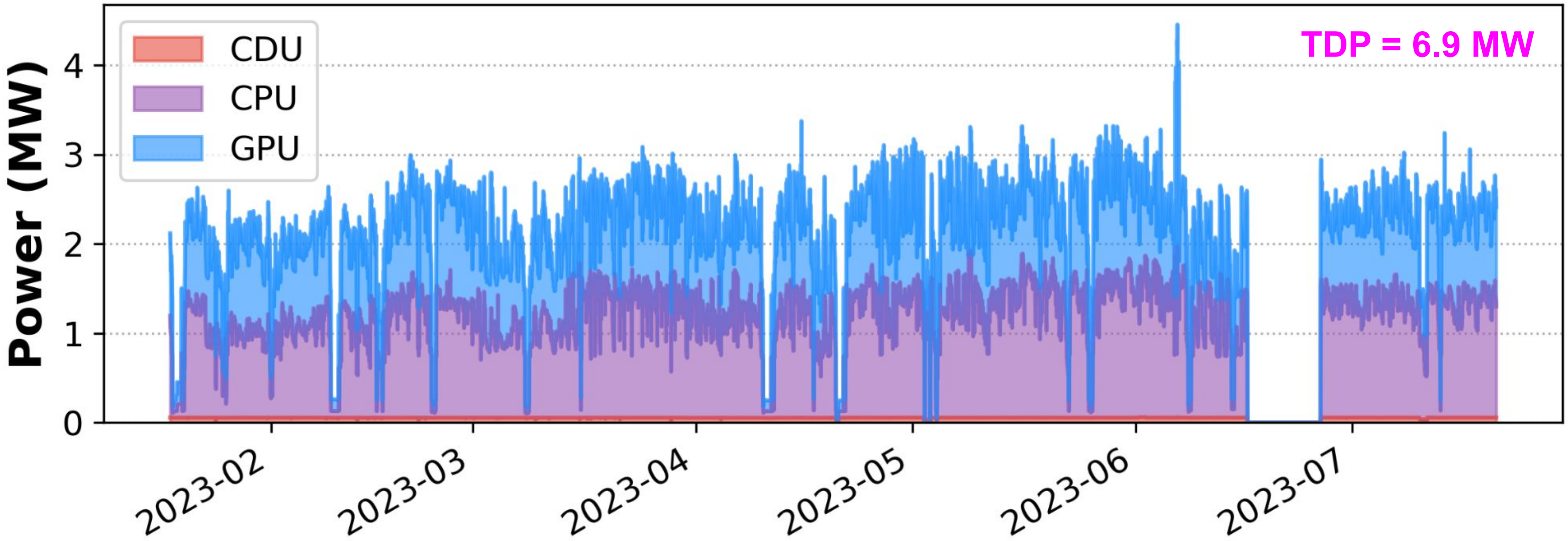
# Question 1

How much power does Perlmutter use?

# System Power Timeline For Perlmutter



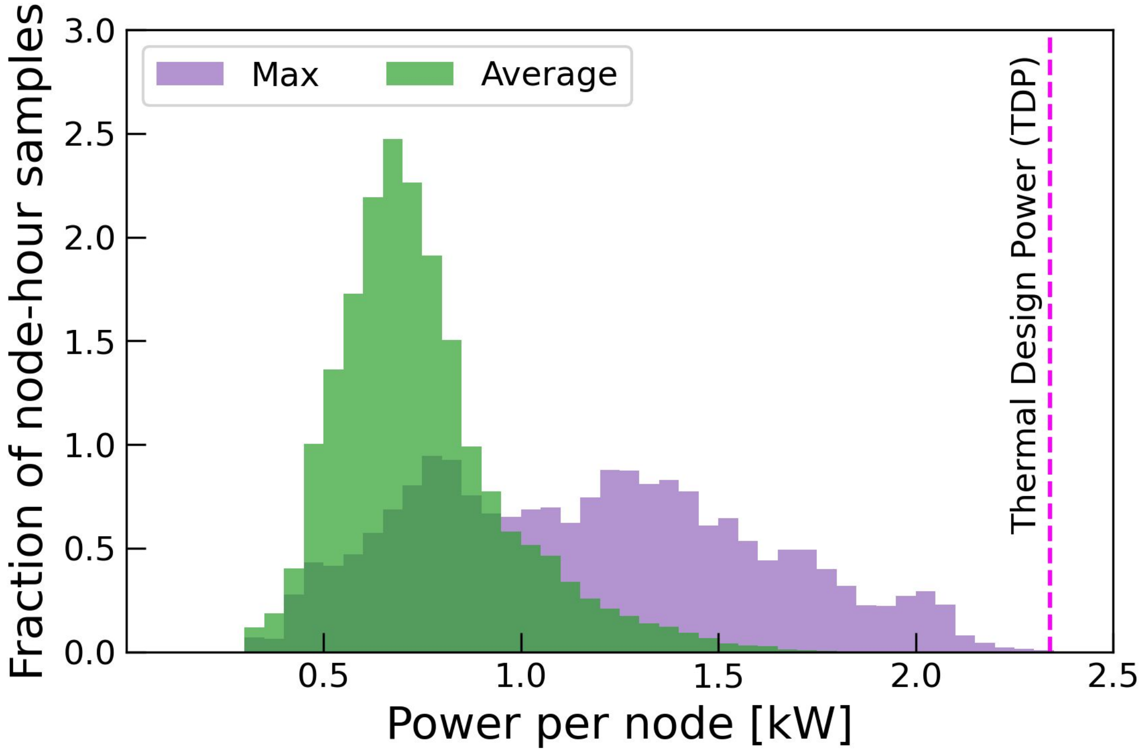
# Perlmutter System Power Usage Fluctuates Significantly and is Much Lower than TDP, Particularly for the GPU Partition




**This work focuses on the GPU partition.**


15

# The Difference Between the Average and Peak Power Distribution Indicates Significant Power Fluctuations During Job Runs



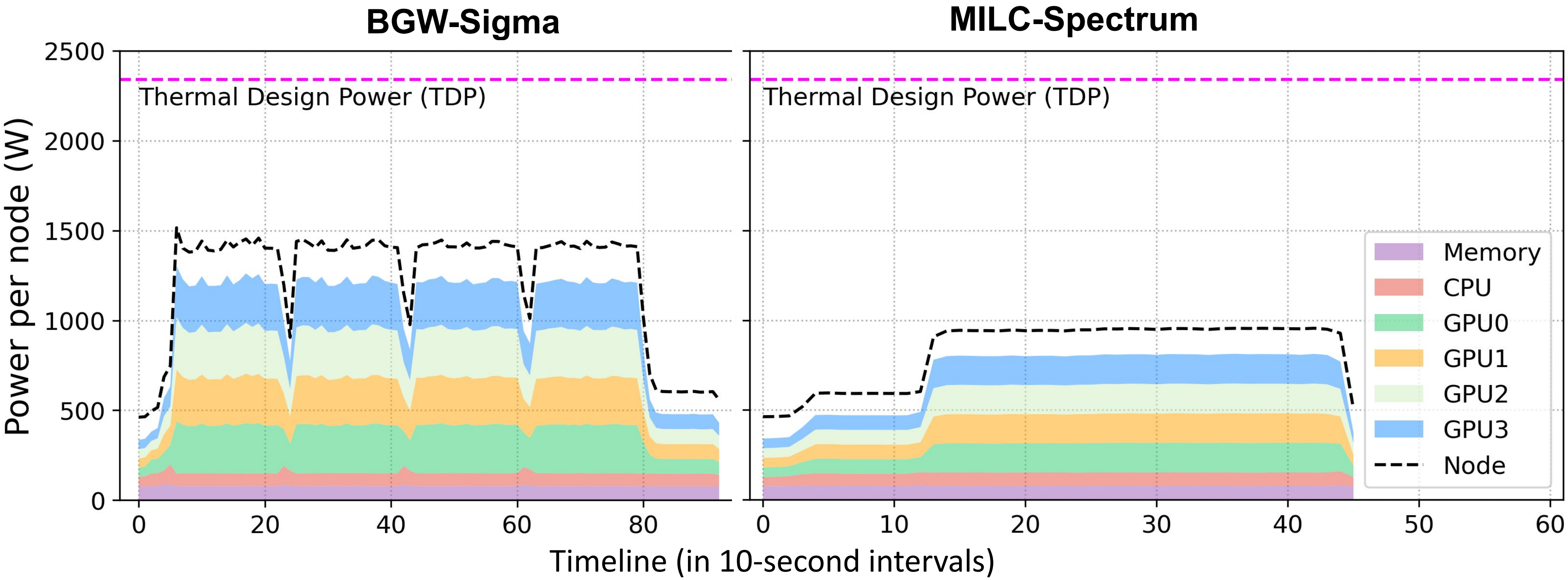

- Data for GPU partition only
- Average way below 1KW/node
- Peak broad, and extends beyond 2 KW

# Question 2

Why is the system power significantly below the TDP?

Not because the utilization is low !

# GPU Applications Have Distinct Power Profiles and the Four GPUs Consume Most Power on the Node

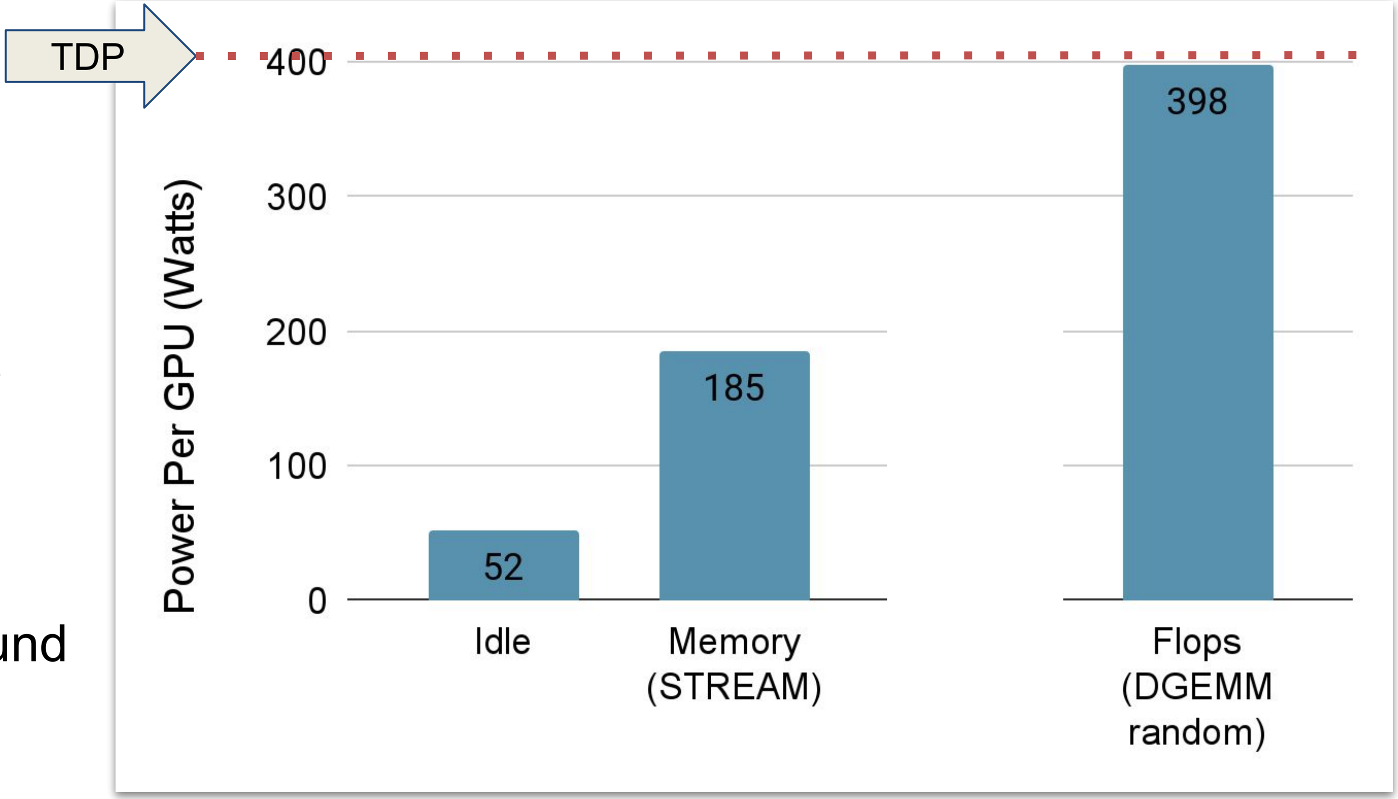




18

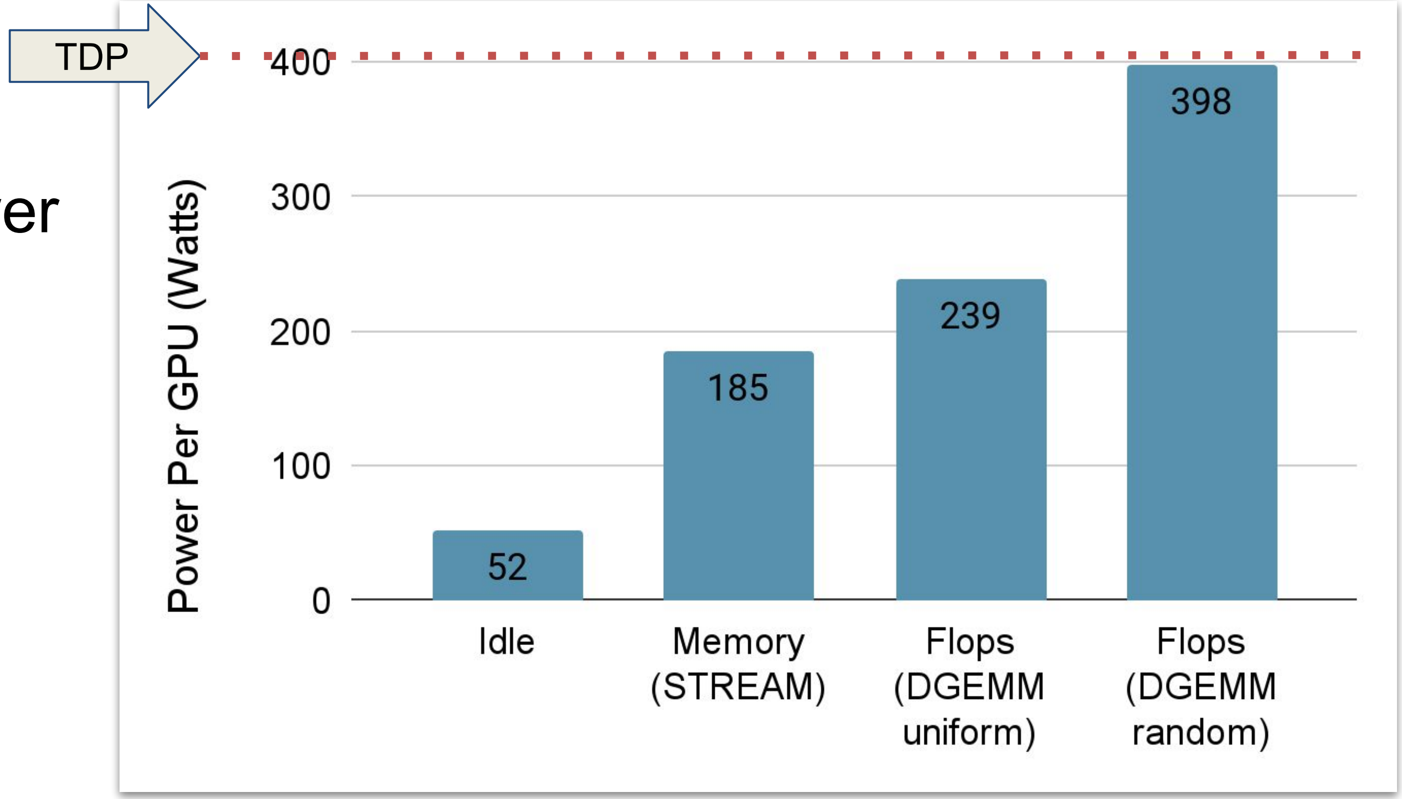
# Per GPU Microbenchmark Power Usage

- Floating point bound with random inputs runs at TDP
- Memory bandwidth bound considerably lower

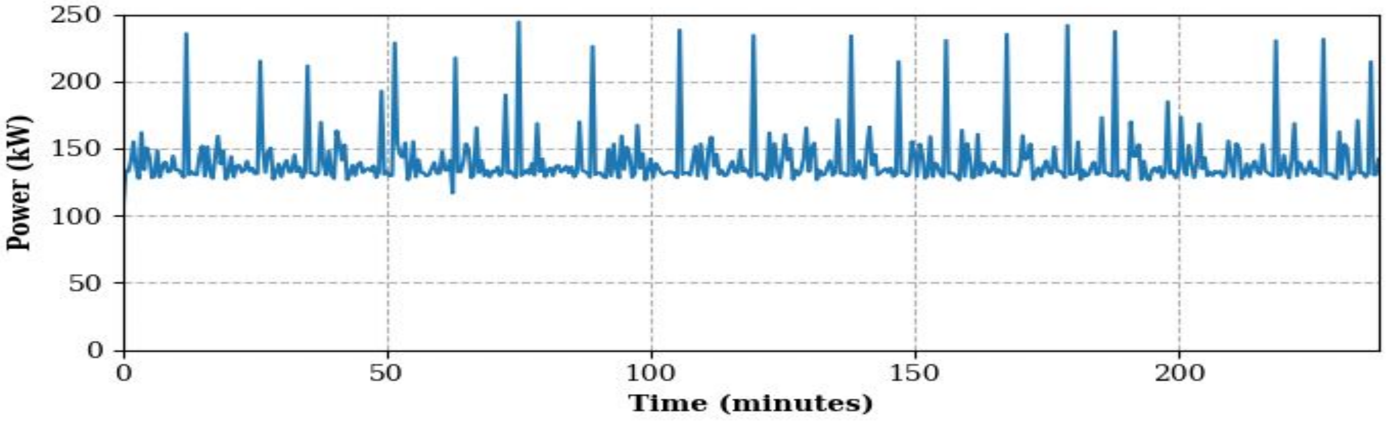


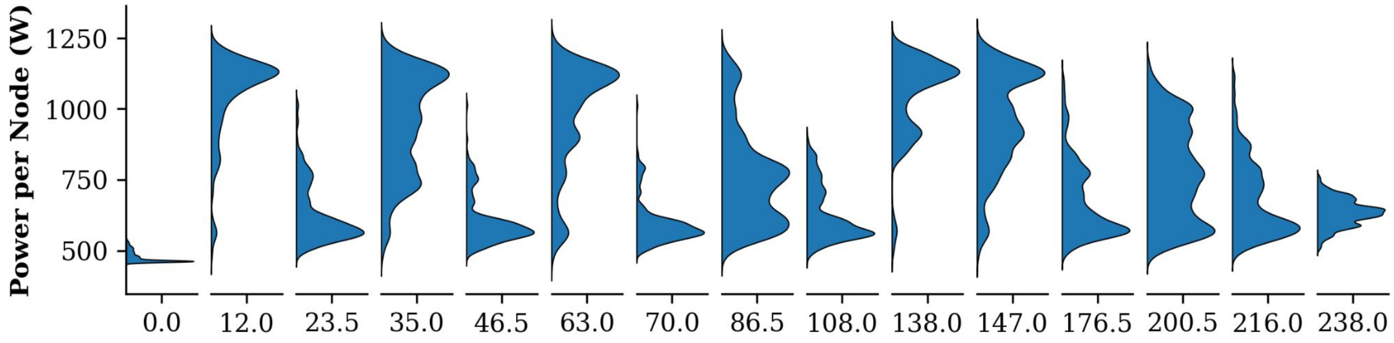
# Per GPU Microbenchmark Power Usage

- Input data effects power usage !



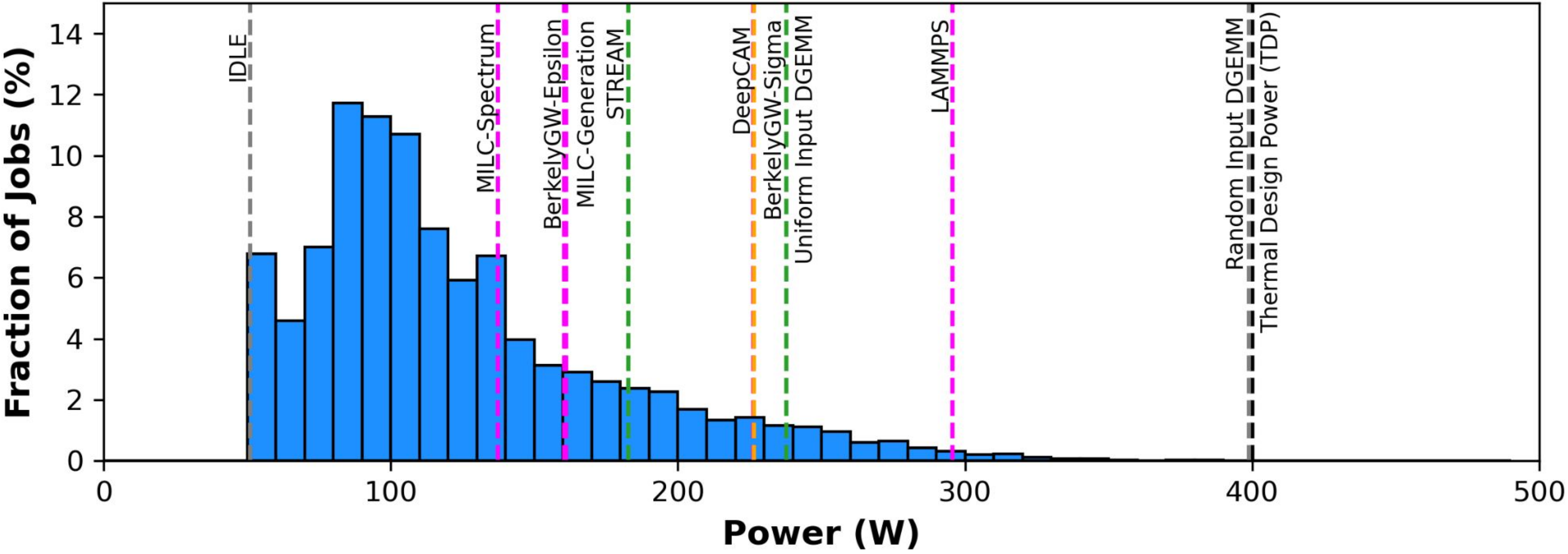
# Power Fluctuations within XGC Application on Perlmutter




- The application ran on 224 nodes for over 2 hours.
- Power timeline characterized by fast power spikes up to 125 kW.
- Power distribution among nodes shows load Imbalance.

# Majority of Applications Use Less Power than STREAM






BERKELEY LAB
Bringing Science Solutions to the World
U.S. DEPARTMENT OF ENERGY
Office of Science

# Summary

- System power draw is consistently below TDP
  - TDP is increasingly a useless metric
- Average Application power on Perlmutter
  - is consistently below what would be expected if it was purely compute or memory bound
  - can vary significantly during the course of a run.
    - Many reasons why - Control flow from GPU<->CPU, MPI, disk I/O,
- Application power is not only dependent on the algorithm also depends in numerical inputs
  - Implications for iterative solvers - and for power predictions !

# What does this Mean for the HPC Community?

- Need to develop power projection methodologies
  - Highly likely NERSC-11 RFP will ask for this!
- Research into power management is abundant and increasingly urgent.  Potential strategies include:
  - power-aware scheduling,
  - coarse- and fine-grained power capping,
  - frequency throttling
- HPC center operators should provide vendors with power usage data - Need system designs that reflect production mode average power in addition to TDP
  - Lets get rid of TDP as a metric !

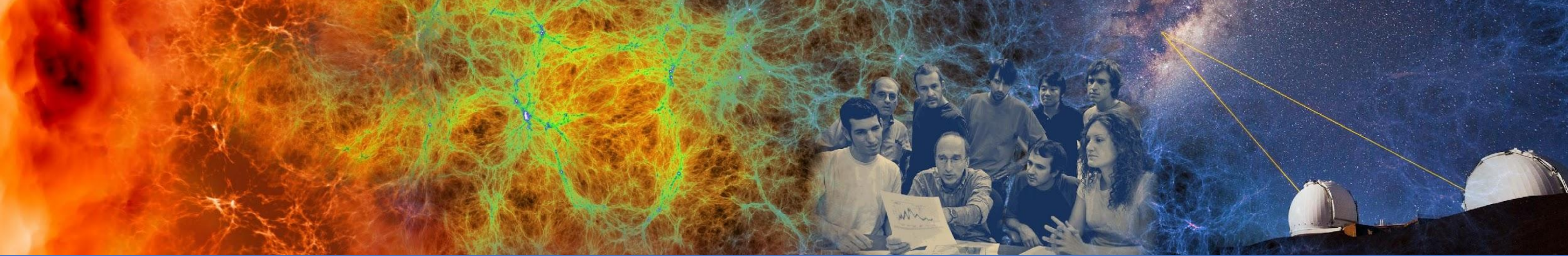
# Thanks!

**Ermal Rrapaj** **Sridutt Bhalachandra** **Zhengji Zhao**

**Brian Austin** **Hai Ah Nam**

**References**


- *Power Consumption Trends in Supercomputers: A Study of NERSC's Cori and Perlmutter Machines.* Ermal Rrapaj, Sridutt Bhalachandra, Zhengji Zhao, Brian Austin, Hai Ah Nam, Nicholas Wright. ISC 2024
- *Power Analysis of NERSC Production Workloads.* Zhengji Zhao, Ermal Rrapaj, Sridutt Bhalachandra, Brian Austin, Hai Ah Nam, Nicholas Wright. Proceedings of the SC'23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis.
- *Understanding the Impact of Input Entropy on FPU, CPU, and GPU Power.* Sridutt Bhalachandra, Brian Austin, Samuel Williams, Nicholas J Wright. _ arXiv preprint arXiv:2212.08805
- *Understanding power variation and its implications on performance optimization on the Cori supercomputer.* Sridutt Bhalachandra, Brian Austin, Nicholas J Wright. 2021 International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS).



# Questions?