Not All GPUs Are Created Equal: Characterizing Variability in Large-Scale, Accelerator-Rich Systems

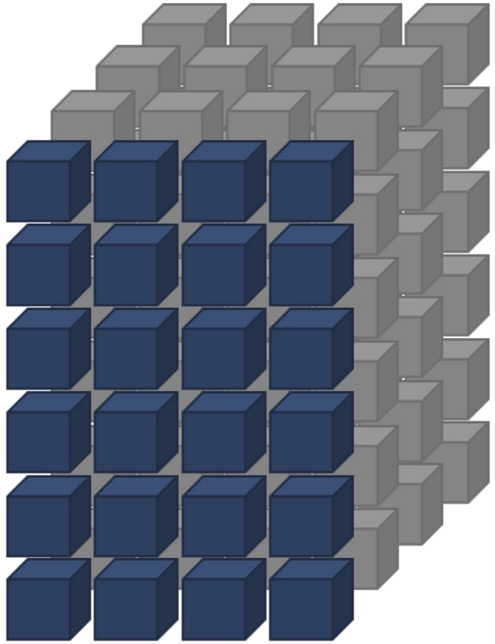
Prasoon Sinha, Akhil Guliani, **Rutwik Jain**, Brandon Tran, Matthew D. Sinclair and Shivaram Venkataraman
Computer Sciences Department, University of Wisconsin-Madison

DOMAIN
SCIENTISTS

APPLICATIONS

IMAGE RECOGNITION
PHYSICS SIMULATION
MACHINE TRANSLATION
GRAPH ANALYTICS
MOLECULAR DYNAMICS
SPEECH RECOGNITION
GENOMICS

SUPERCOMPUTER



 = 1 node

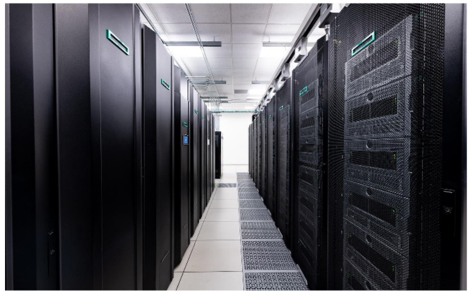
DOMAIN SCIENTISTS

APPLICATIONS

IMAGE RECOGNITION
PHYSICS SIMULATION
MACHINE TRANSLATION
GRAPH ANALYTICS
MOLECULAR DYNAMICS
SPEECH RECOGNITION
GENOMICS

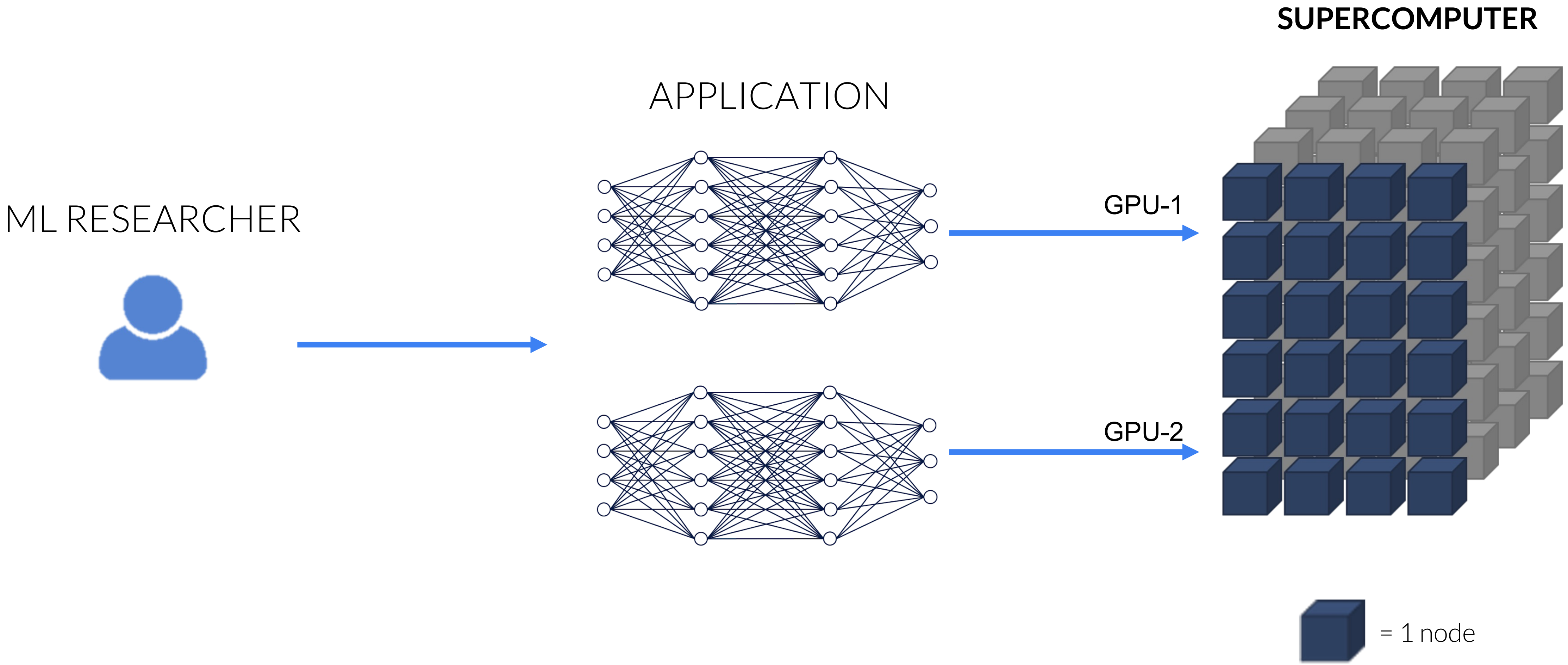
SUPERCOMPUTERS

ORNL SUMMIT
27,000 GPUs



NCSA DELTA
840 GPUs

TACC LONGHORN
416 GPUs

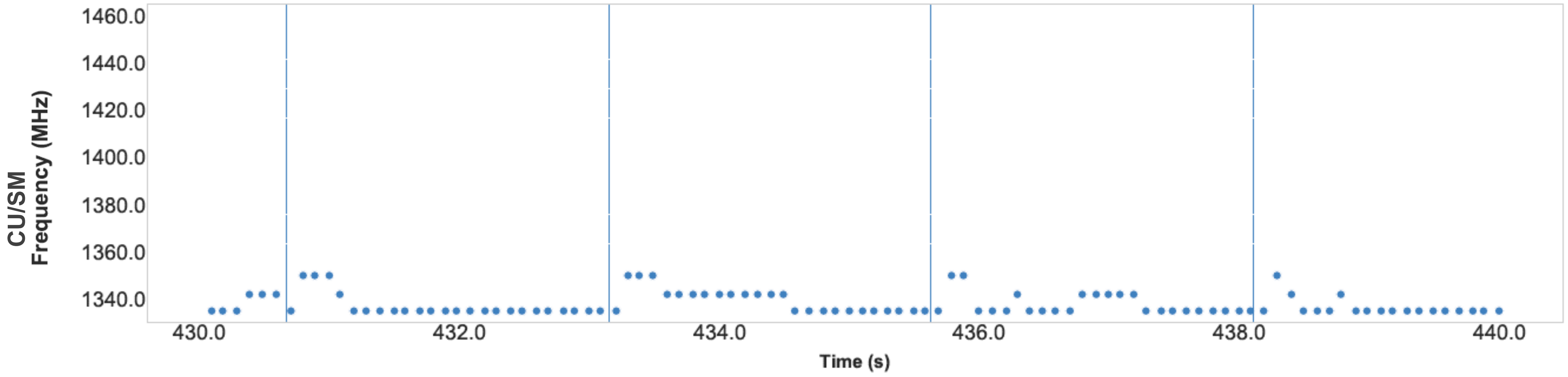

ML RESEARCHER
APPLICATION
SUPERCOMPUTER
GPU-1
GPU-2
= 1 node

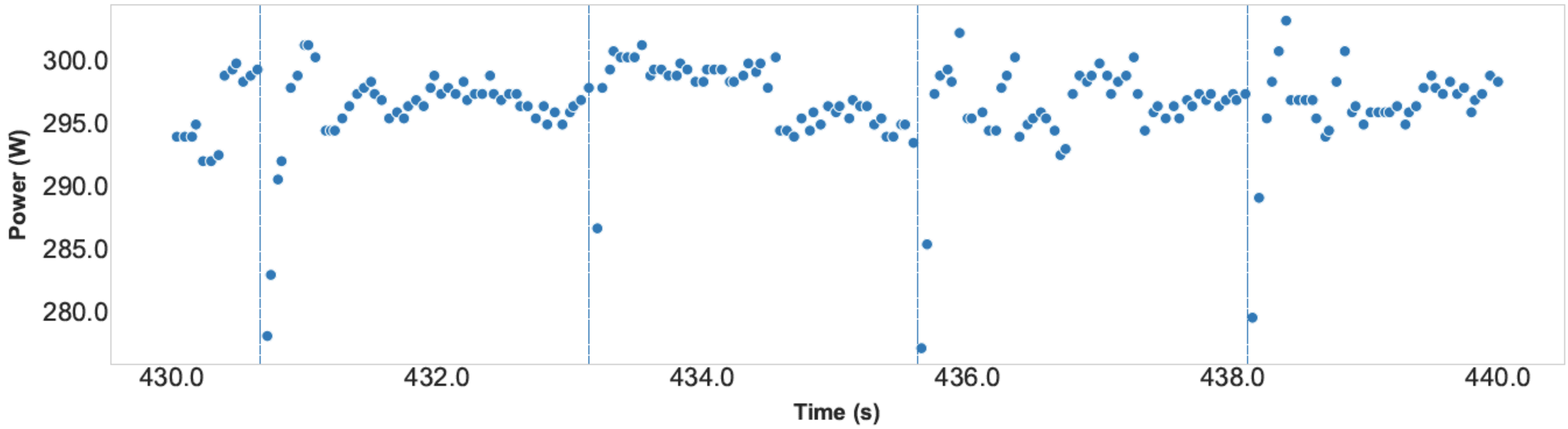
GPU-1

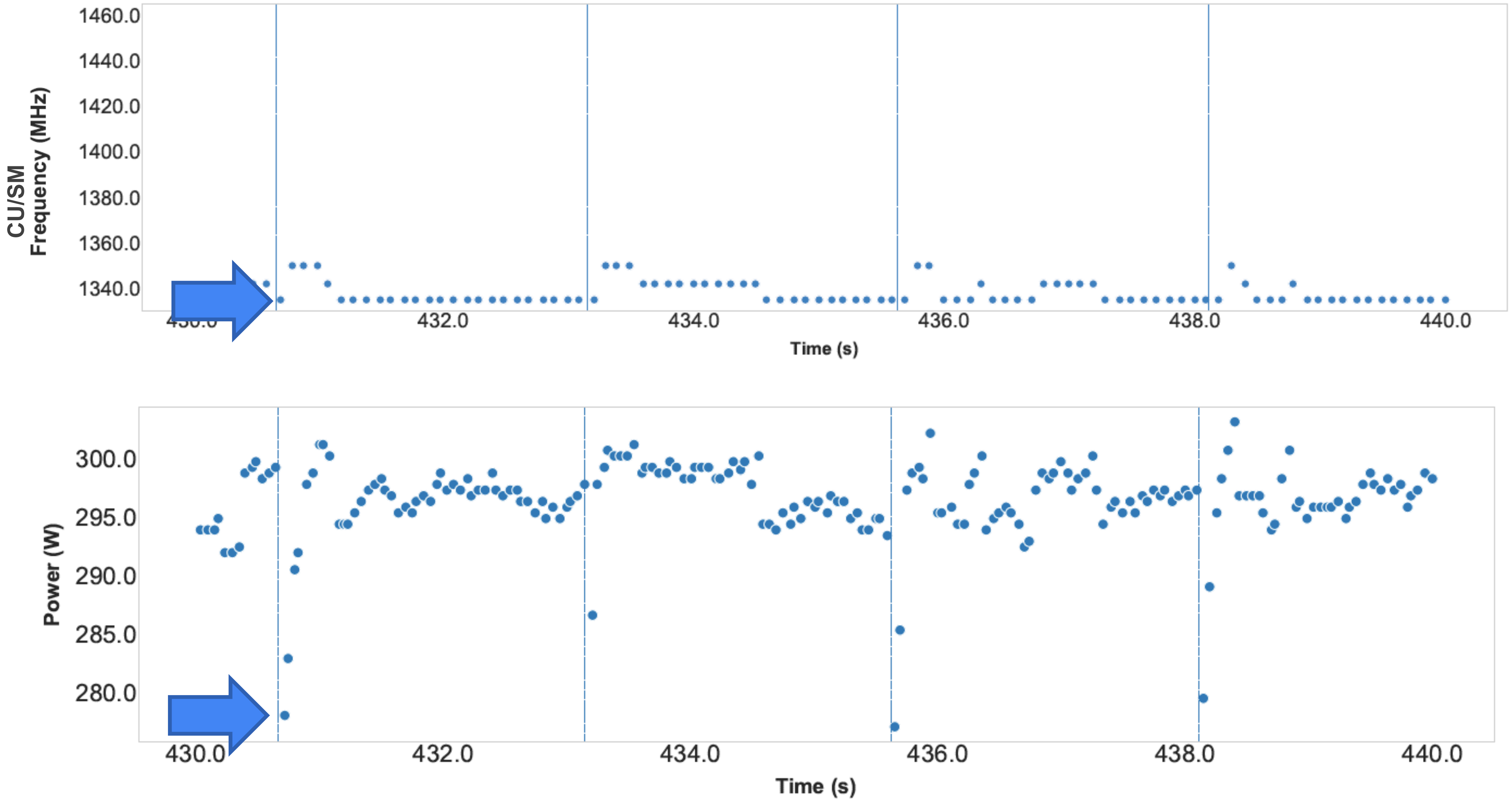
8 to 22% performance difference

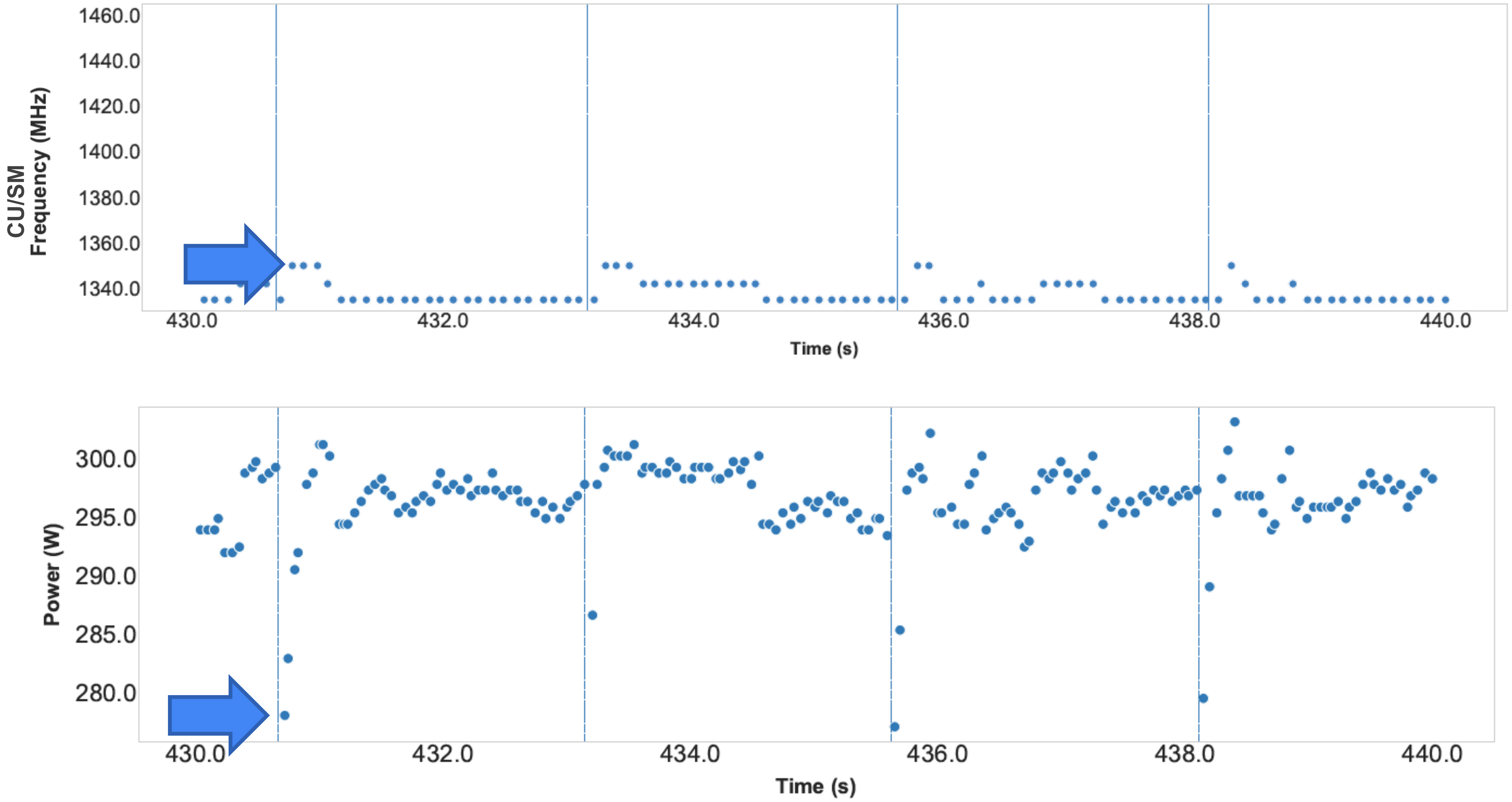
INCONSISTENCY!

GPU-2

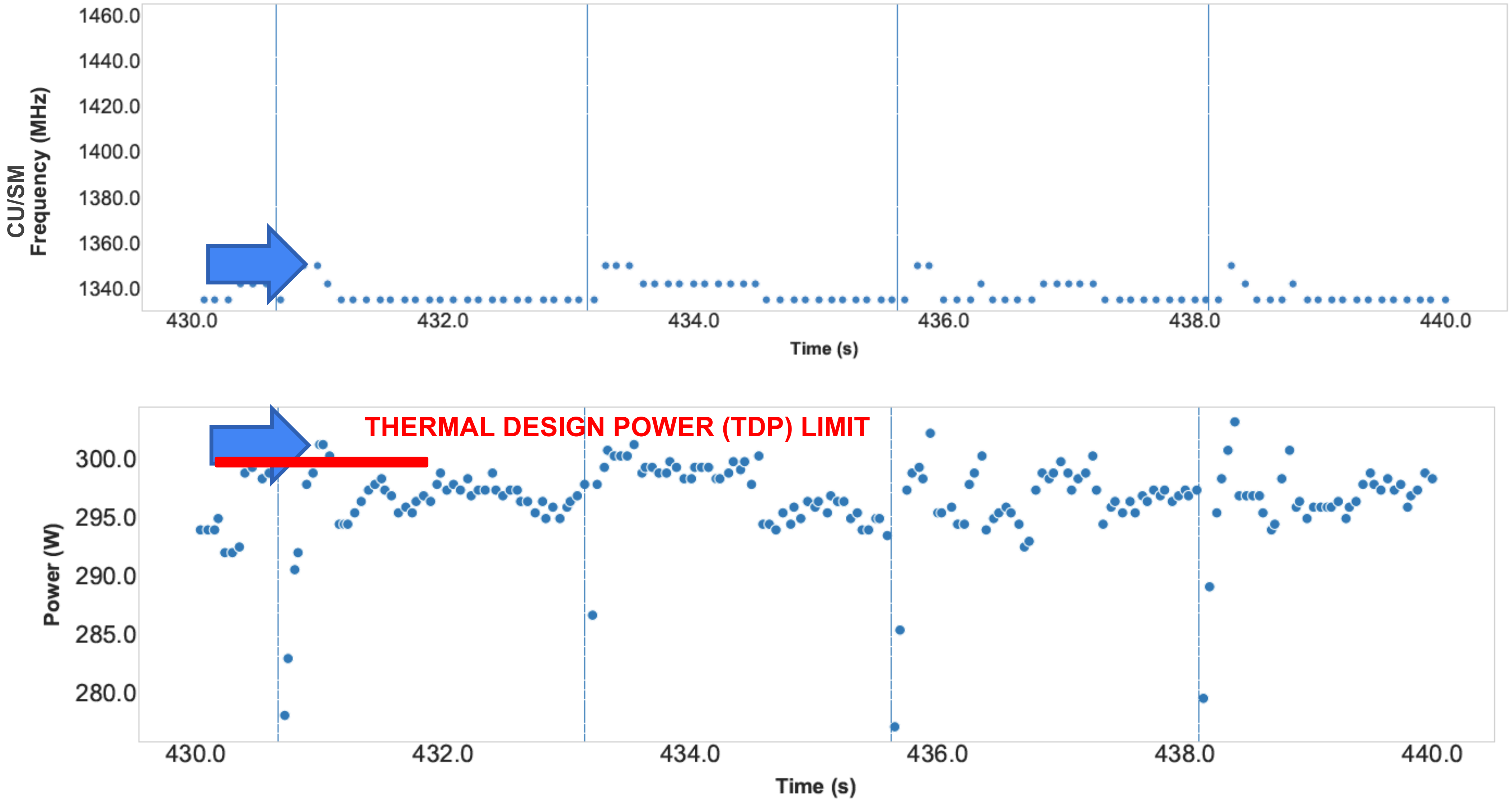

1460.0
1440.0
1420.0
1400.0
1380.0
1360.0
1340.0
CU/SM Frequency (MHz)
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)

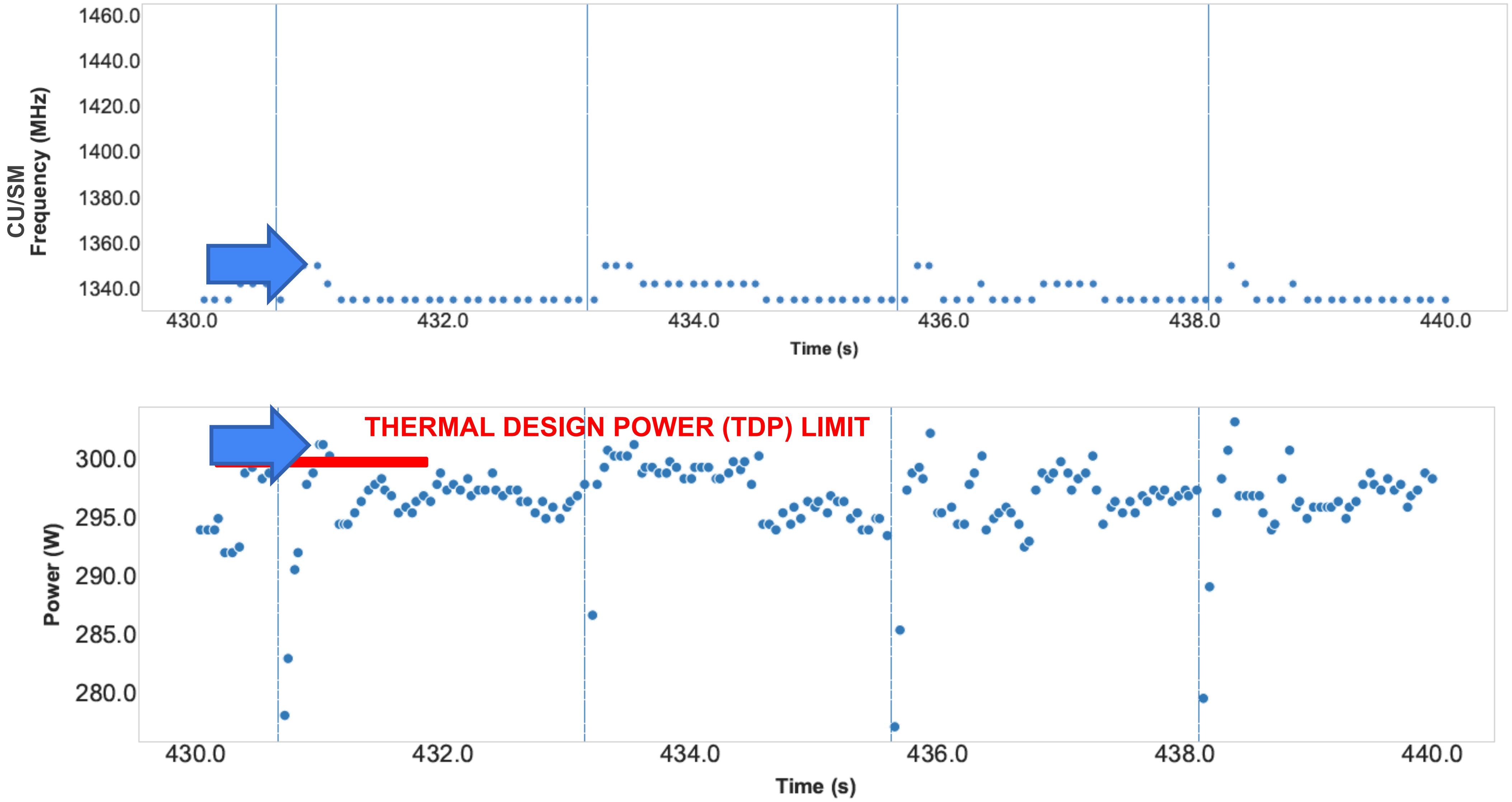

300.0
295.0
290.0
285.0
280.0
Power (W)
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)

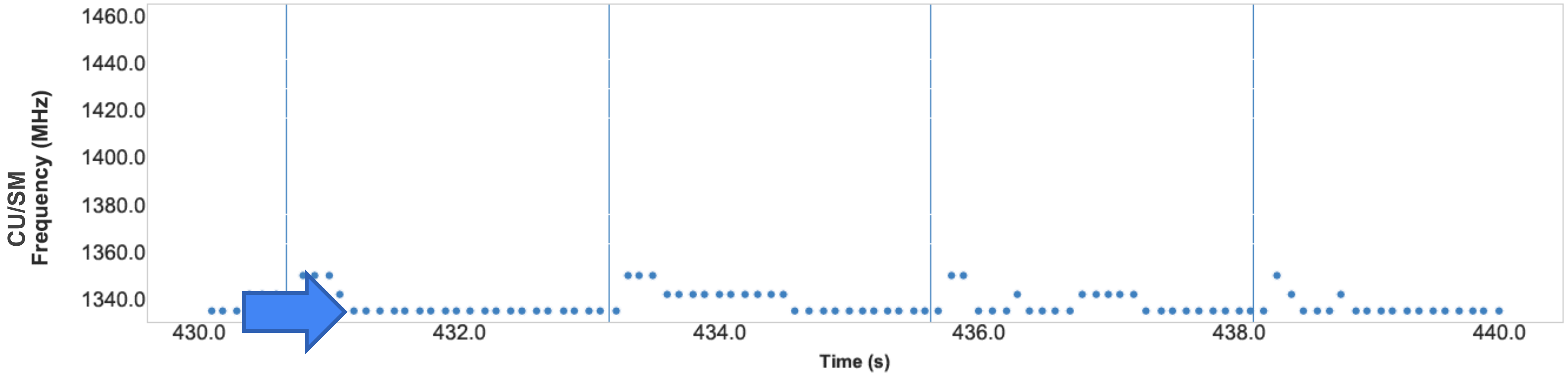

1460.0
1440.0
1420.0
1400.0
1380.0
1360.0
1340.0
CU/SM Frequency (MHz)
432.0
434.0
436.0
438.0
440.0
Time (s)
300.0
295.0
290.0
285.0
280.0
Power (W)
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)

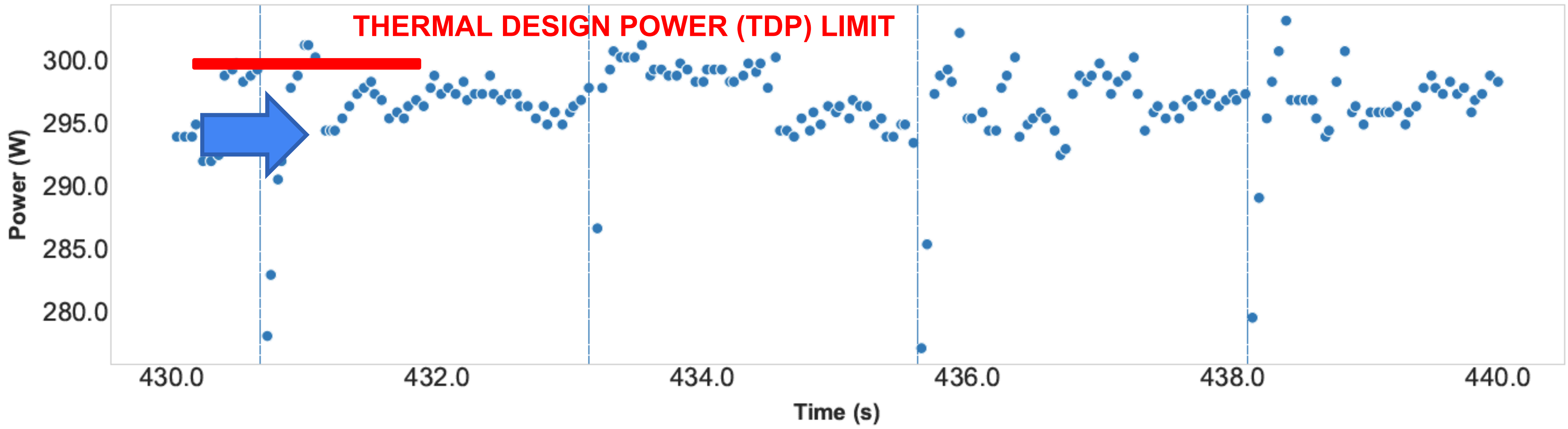


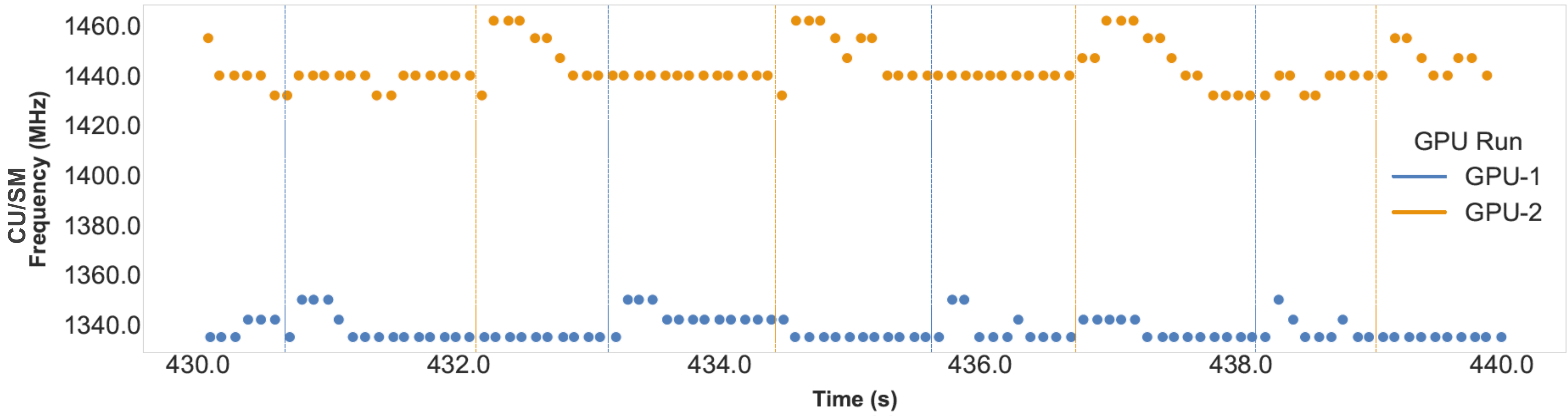
CU/SM Frequency (MHz)
1460.0
1440.0
1420.0
1400.0
1380.0
1360.0
1340.0
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)
Power (W)
300.0
295.0
290.0
285.0
280.0
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)


CU/SM Frequency (MHz)
1460.0
1440.0
1420.0
1400.0
1380.0
1360.0
1340.0
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)
THERMAL DESIGN POWER (TDP) LIMIT
Power (W)
300.0
295.0
290.0
285.0
280.0
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)

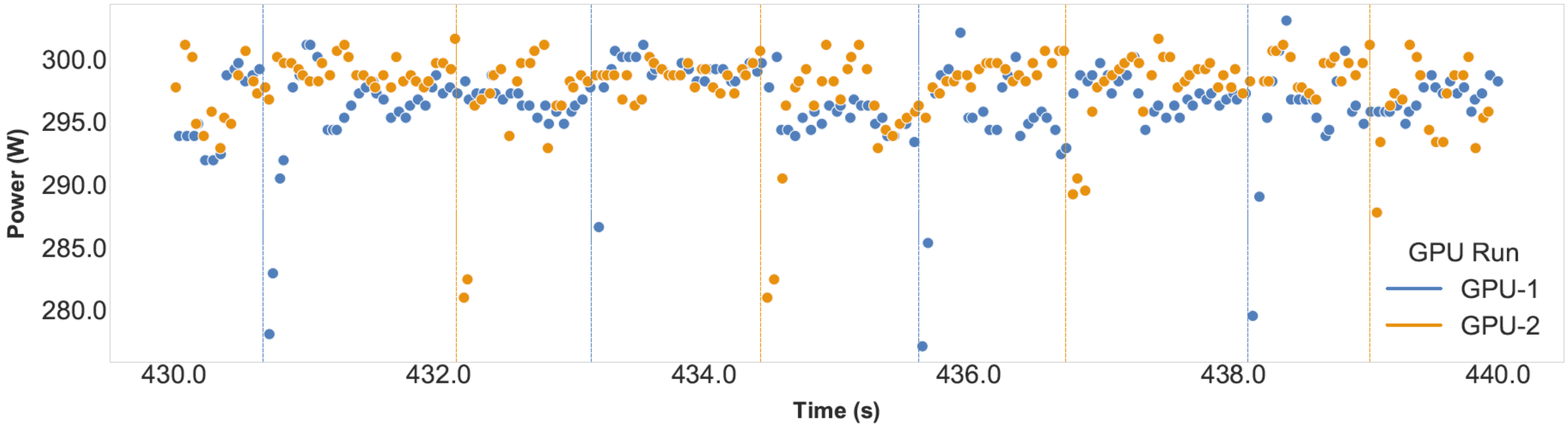

1460.0
1440.0
1420.0
1400.0
1380.0
1360.0
1340.0
CU/SM Frequency (MHz)
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)
THERMAL DESIGN POWER (TDP) LIMIT
300.0
295.0
290.0
285.0
280.0
Power (W)
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)


1460.0
1440.0
1420.0
1400.0
1380.0
1360.0
1340.0
CU/SM Frequency (MHz)
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)

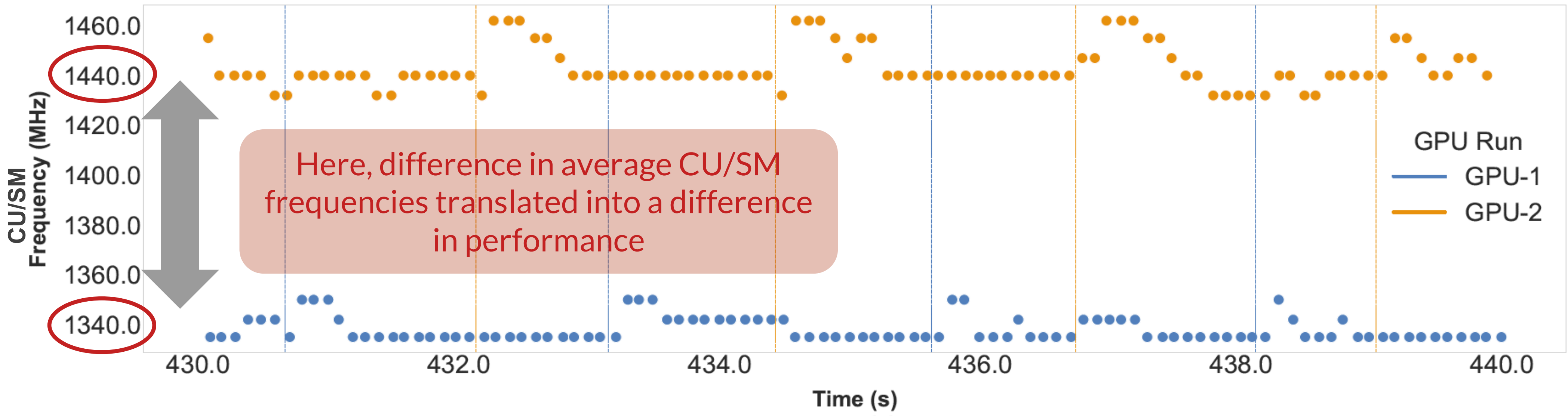

THERMAL DESIGN POWER (TDP) LIMIT
300.0
295.0
290.0
285.0
280.0
Power (W)
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)

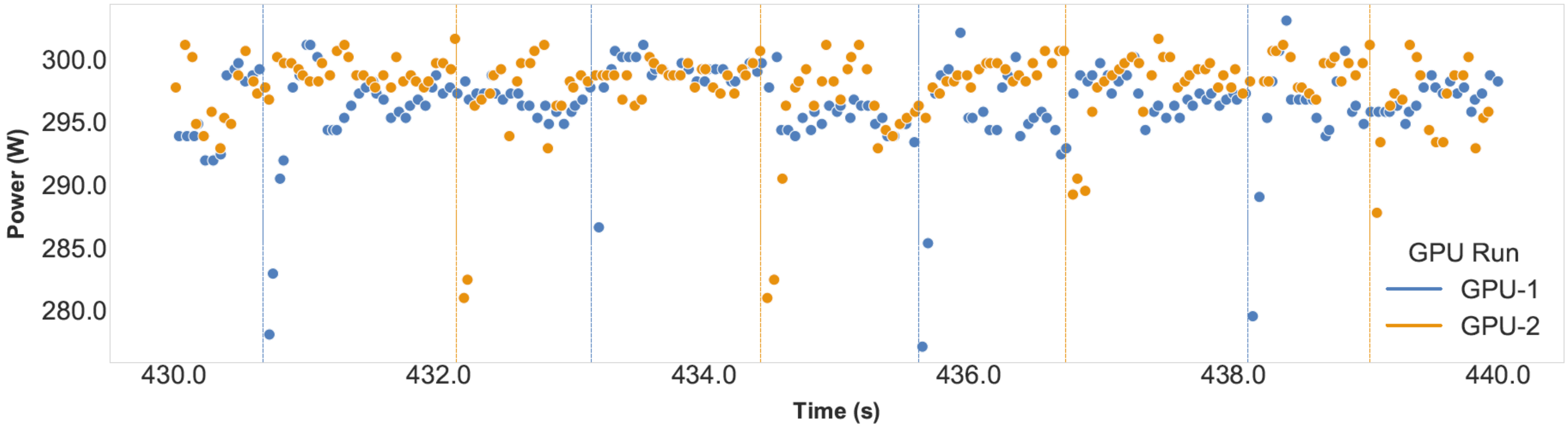


CU/SM Frequency (MHz)
1460.0
1440.0
1420.0
1400.0
1380.0
1360.0
1340.0
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)
GPU Run
GPU-1
GPU-2



Power (W)
300.0
295.0
290.0
285.0
280.0
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)
GPU Run
GPU-1
GPU-2

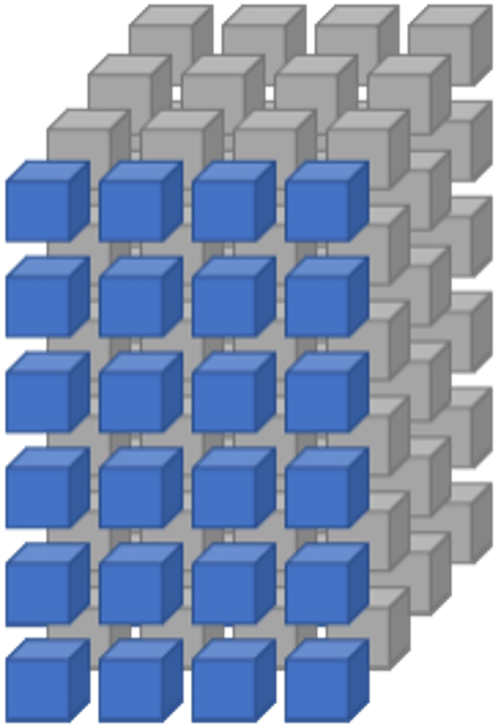

1460.0
1440.0
1420.0
1400.0
1380.0
1360.0
1340.0
CU/SM Frequency (MHz)
Here, difference in average CU/SM frequencies translated into a difference in performance
GPU Run
GPU-1
GPU-2
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)


300.0
295.0
290.0
285.0
280.0
Power (W)
GPU Run
GPU-1
GPU-2
430.0
432.0
434.0
436.0
438.0
440.0
Time (s)

The Problem: Performance Predictability

- Hard to get **repeatable, consistent performance**!
- In multi-GPU experiments, faster nodes keep waiting for slower resources – can lead to **resource underutilization**!

SUPERCOMPUTER

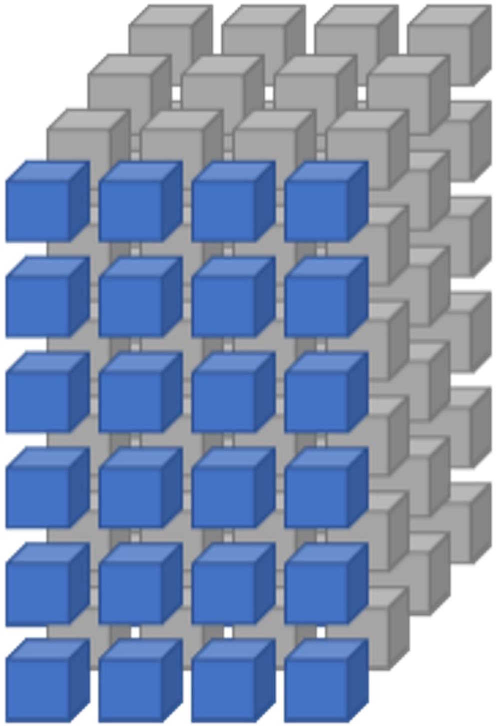


Objective

Understand and characterize **GPU variability** in large scale, accelerator-rich computing clusters

Examine the effects of **scale**, **application type**, **cooling** and **GPU vendors** on variability

SUPERCOMPUTER



Related Work

- **Acun, et al. [ICS'16, IPDPSW'16, HiPC'17], Zhang, et al. [IPDPS'15]**
  - Performance variability in CPU-based HPC systems
  - Solutions: dynamic load balancing, adaptive runtimes, temperature-aware job placement

Related Work

- **Acun, et al. [ICS’16, IPDPSW’16, HiPC’17], Zhang, et al. [IPDPS’15]**
  - Performance variability in CPU-based HPC systems
  - Solutions: dynamic load balancing, adaptive runtimes, temperature-aware job placement
- **Coplin, et al. [IPDPSW’16], Jiao, et al. [GreenCom&ICCPS’10]**
  - Energy, power and performance characterization in GPGPU benchmarks
  - Used older generation GPUs (Kepler/Fermi) and focus on single-GPU workstations

Related Work

- **Acun, et al. [ICS’16, IPDPSW’16, HiPC’17], Zhang, et al. [IPDPS’15]**
  - Performance variability in CPU-based HPC systems
  - Solutions: dynamic load balancing, adaptive runtimes, temperature-aware job placement
- **Coplin, et al. [IPDPSW’16], Jiao, et al. [GreenCom&ICCPS’10]**
  - Energy, power and performance characterization in GPGPU benchmarks
  - Used older generation GPUs (Kepler/Fermi) and focus on single-GPU workstations
- **Scogland, et al. [SC’15]**
  - Studied LINPACK’s performance variation for CPU clusters and AMD GPUs
  - Motivated the need for a more in-depth study

Outline

- ❑ Motivation
- ❑ Related Work
- ❑ **Methodology**
- ❑ Experiments and Results
- ❑ Conclusion

First, let's define variability

Interquartile Range

$$\mathrm{IQR} = \mathrm{Q}_3 - \mathrm{Q}_2$$

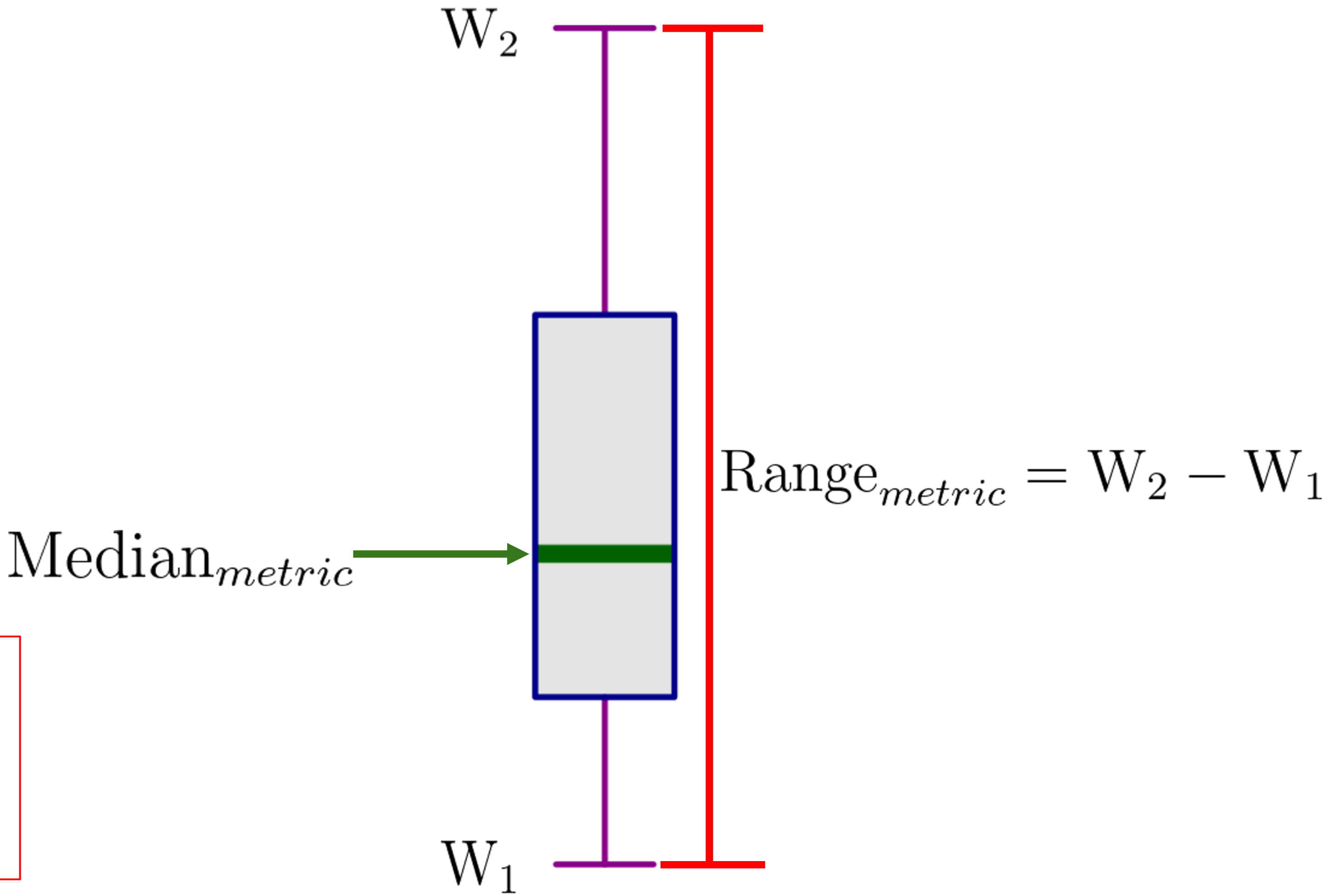
Whiskers and Range

$$\mathrm{W}_2 = \mathrm{Q}_3 + 1.5 \times \mathrm{IQR}$$

$$\mathrm{W}_1 = \mathrm{Q}_1 - 1.5 \times \mathrm{IQR}$$

$$\mathrm{Range}_{metric} = \mathrm{W}_2 - \mathrm{W}_1$$

$$\boxed{\mathrm{Variability}_{metric} = \frac{\mathrm{Range}_{metric}}{\mathrm{Median}_{metric}}}$$



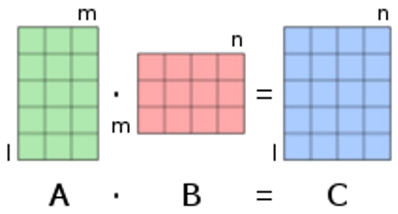
Key Questions

1. How much performance variation is there across GPUs?
2. Do GPU physical metrics (frequency, power and temperature) vary too?
3. How is variability affected by cluster parameters (size, cooling, GPU vendor)?
4. Are these variability observations consistent over time?
5. Is variability application-dependent?

Methodology

- **Benchmark**

  SGEMM



- **Metrics**

  Kernel duration (ms)

  GPU CU/SM **temperature** (°C)

  GPU CU/SM **power consumption** (W)

  GPU CU/SM **frequency** (MHz)

- **Profiler**

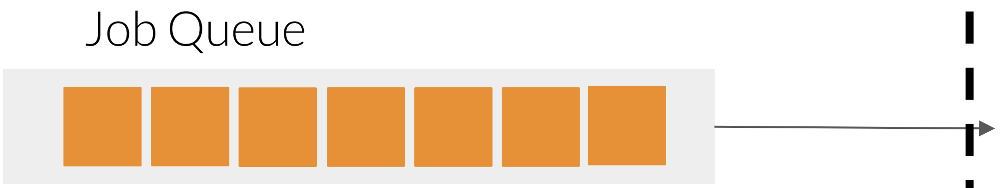
  nvprof
  NVIDIA

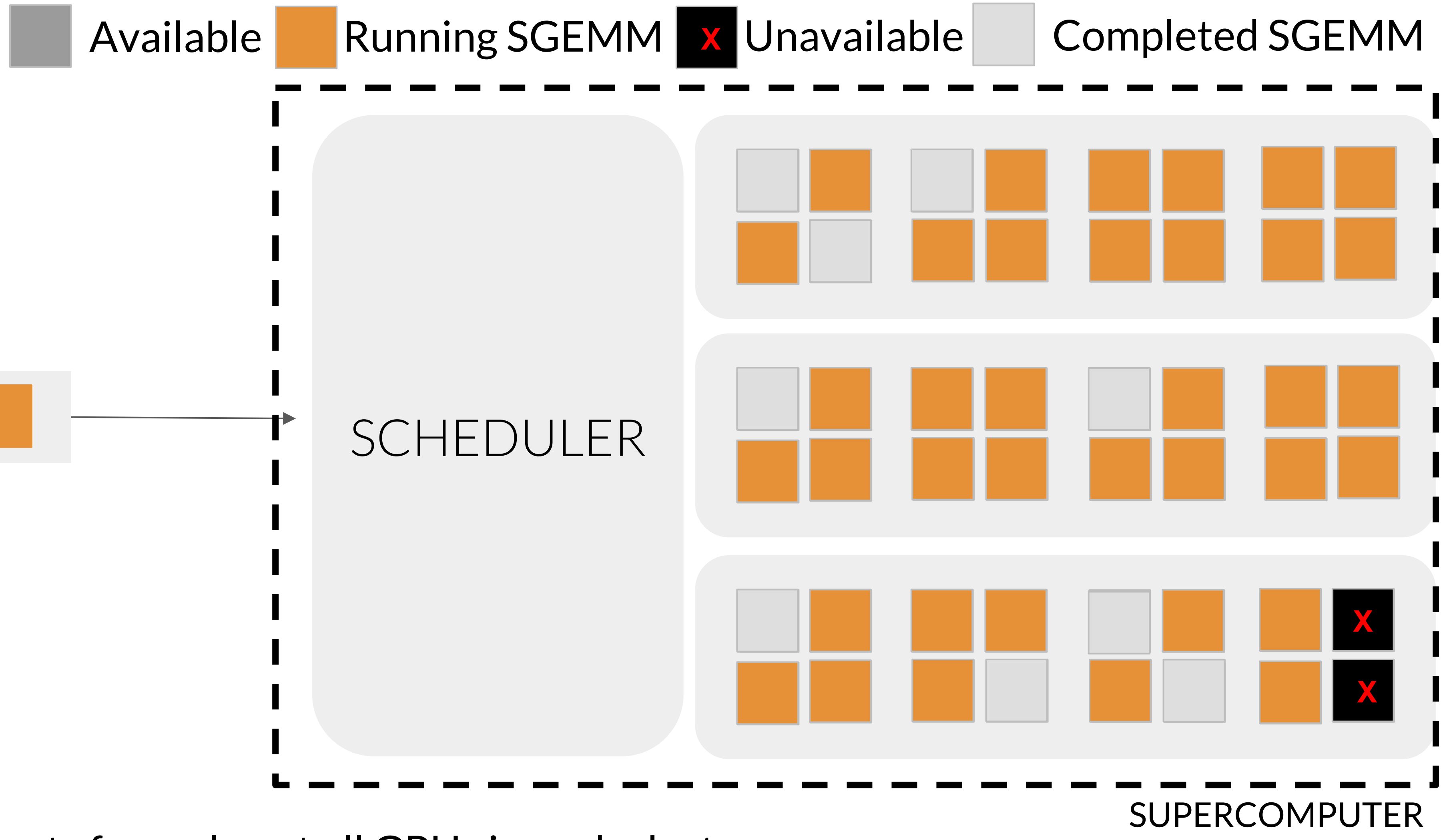
- **Cluster specifications**

TACC TEXAS ADVANCED COMPUTING CENTER **Longhorn**

  - **416** GPUs
  - NVIDIA **Volta V100**
  - air-cooled

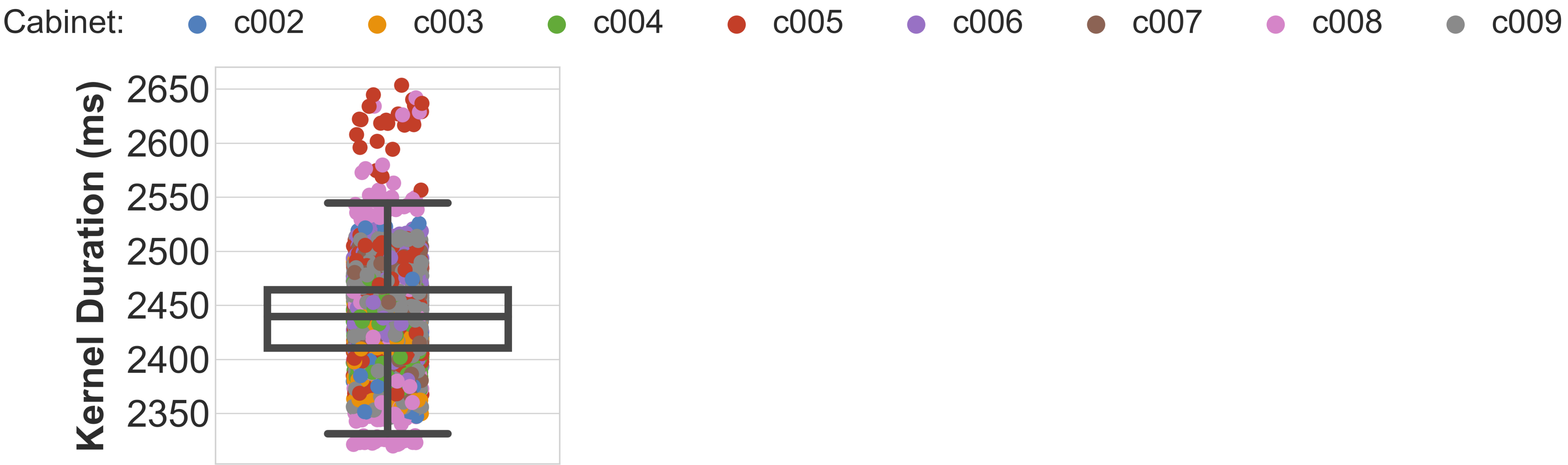
Methodology





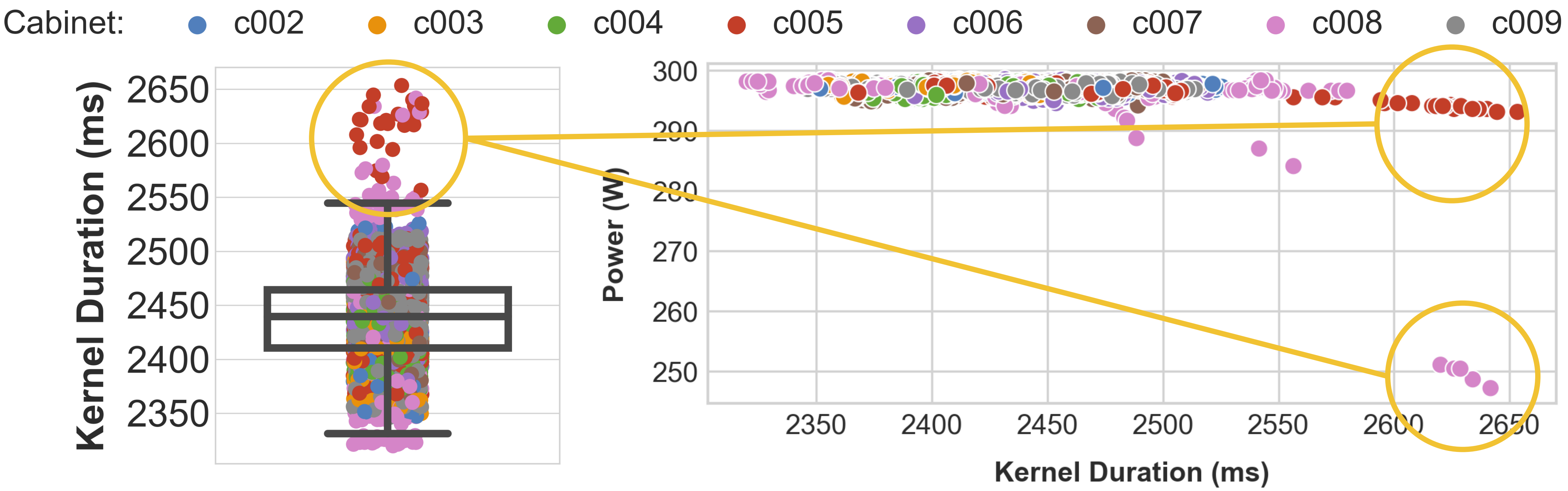
- **Sample Size**
  - Sample measurements from almost all GPUs in each cluster
  - Profiled 2.9x more GPUs than worst-case recommendations for statistical significance [Scogland, et al. SC'15]

SGEMM on TACC Longhorn: Performance



Performance variability
9%

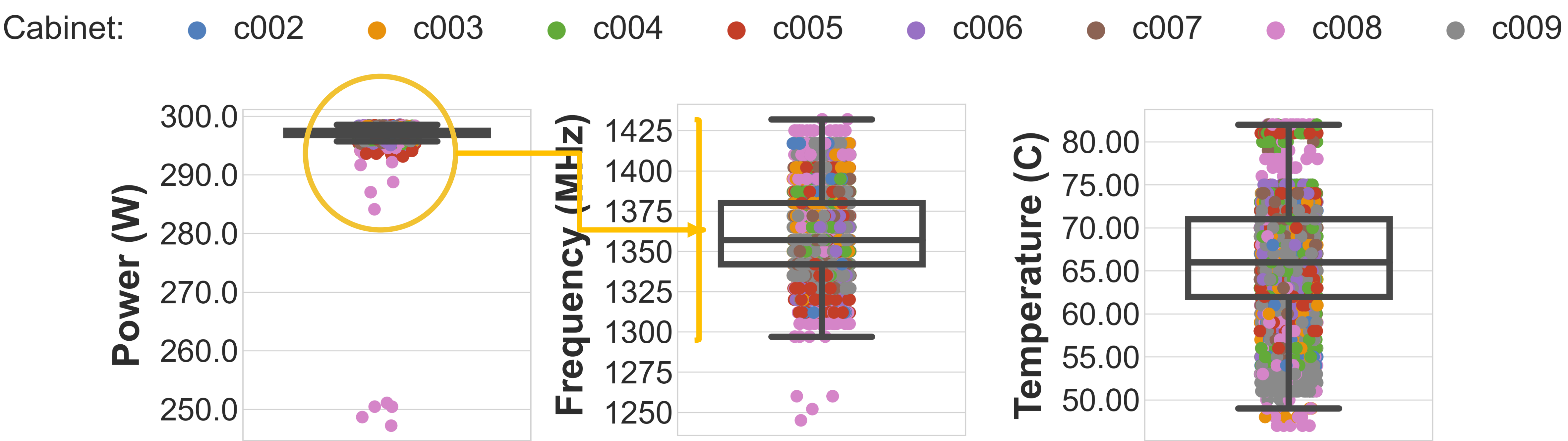
SGEMM on TACC Longhorn: Scatterplot



Performance variability
9%

power and performance
weakly correlated
$\rho = -0.35$

SGEMM on TACC Longhorn: Other Metrics



Power variability
1%

Frequency variability
11%

Temperature variability
33°C b/n Q1 and Q3

SGEMM on TACC Longhorn: Other Metrics

Cabinet: c002 c003 c004 c005 c006 c007 c008 c009

SC'20

GPU Lifetimes on Titan Supercomputer: Survival Analysis and Reliability

George Ostrouchov
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN, USA
ostrouchovg@ornl.gov

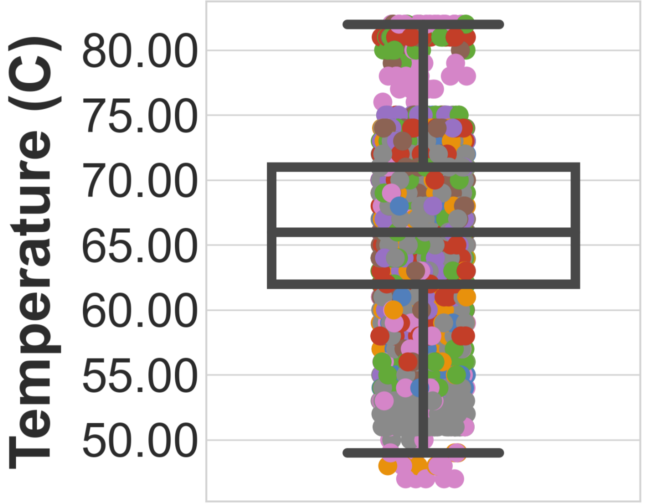
Don Maxwell
National Center for Computational Sciences
Oak Ridge National Laboratory
Oak Ridge, TN, USA
maxwellde@ornl.gov

Rizwan A. Ashraf
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN, USA
ashrafra@ornl.gov

Christian Engelmann
Computer Science and Mathematics Division
Oak Ridge National Laboratory
Oak Ridge, TN, USA
engelmannc@ornl.gov

Mallikarjun Shankar
National Center for Computational Sciences
Oak Ridge National Laboratory
Oak Ridge, TN, USA
shankarm@ornl.gov

James H. Rogers
National Center for Computational Sciences
Oak Ridge National Laboratory
Oak Ridge, TN, USA
jrogers@ornl.gov



Temperature variability
33°C b/n Q1 and Q3

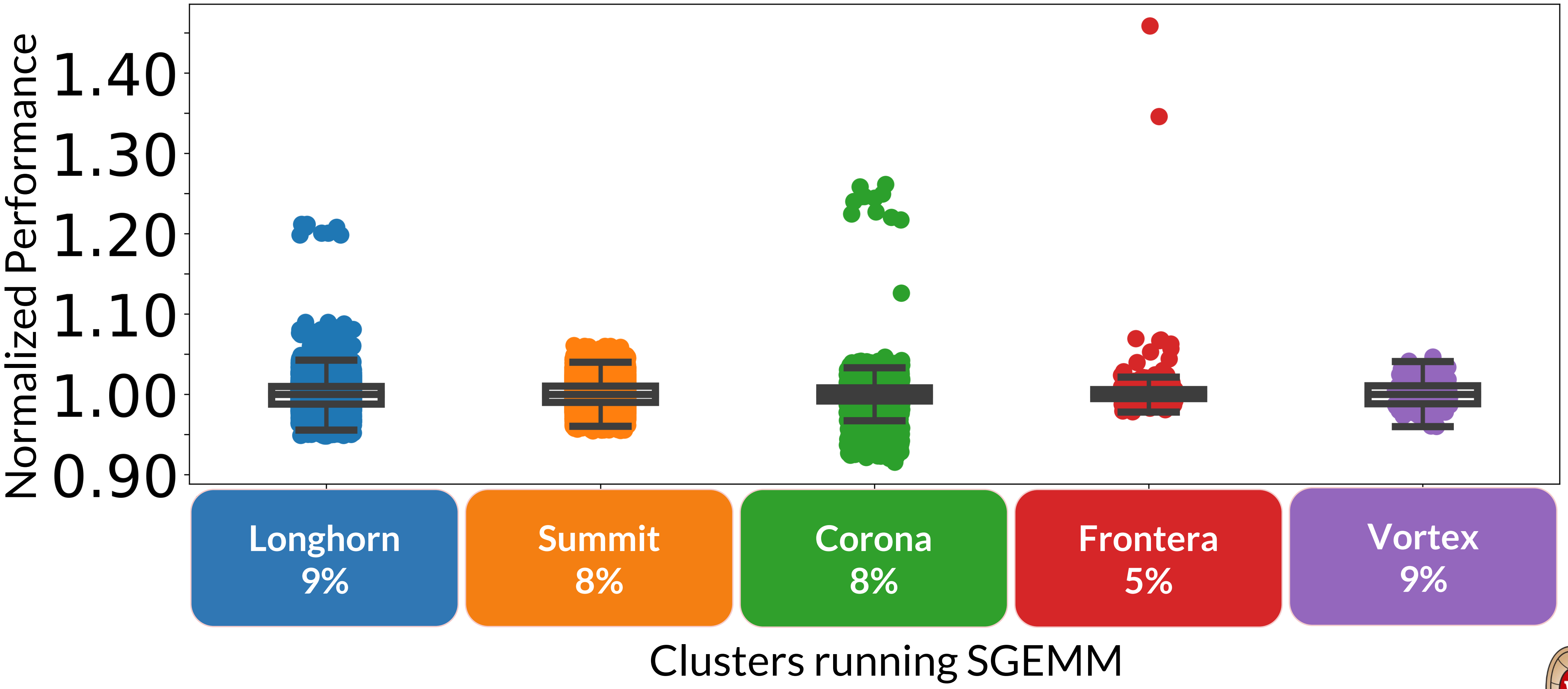
Key Questions

1. How much performance variation is there across GPUs?	**9% for SGEMM**
2. Do GPU physical metrics (frequency, power and temperature) vary too?	**Yes**
3. How is variability affected by cluster parameters (size, cooling, GPU vendor)	
4. Are these variability observations consistent over time?	
5. Is variability application-dependent?	

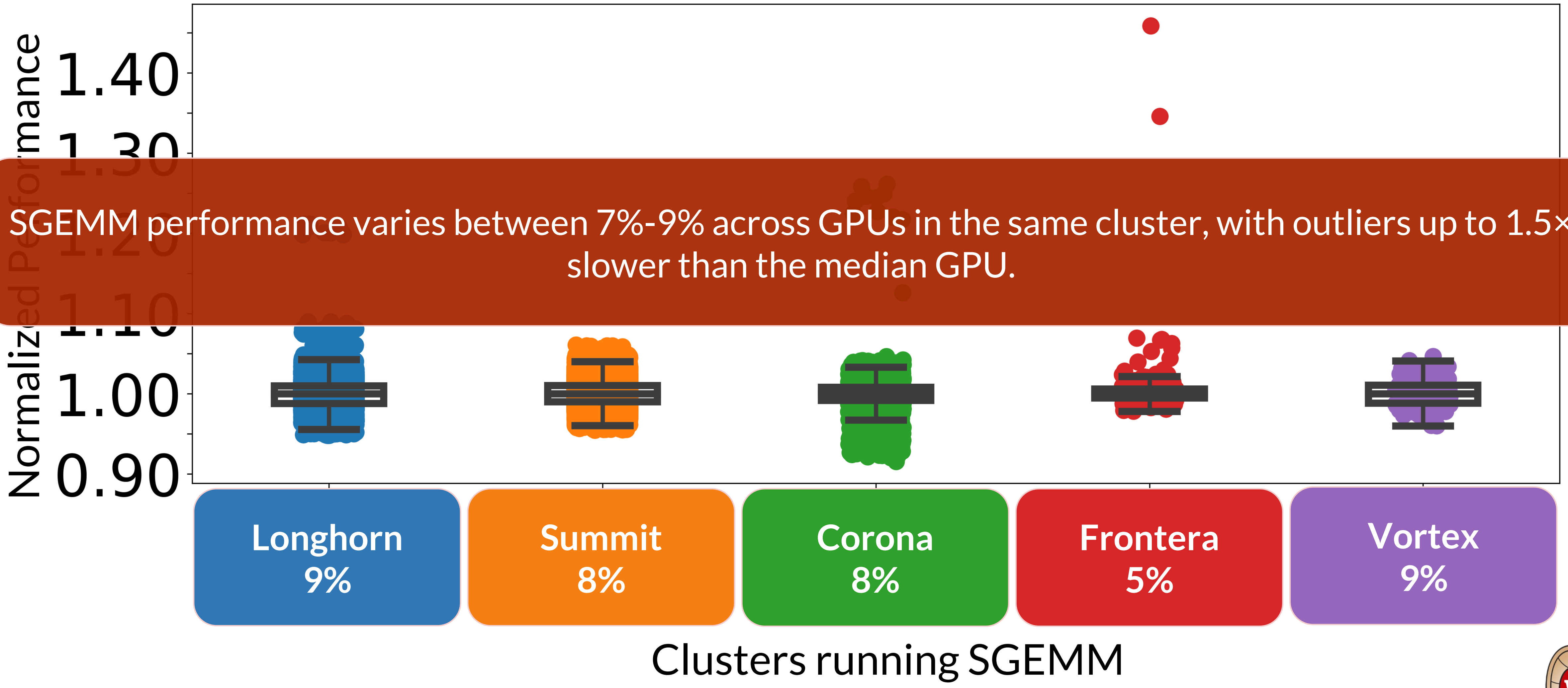
Variability & Cluster parameters

TACC TEXAS ADVANCED COMPUTING CENTER	TACC **Longhorn**	NVIDIA V100 (**Volta**)	**416** GPUs	**air** cooled	
TACC TEXAS ADVANCED COMPUTING CENTER	TACC **Frontera**	NVIDIA RTX5000 (**Turing**)	**360** GPUs	**mineral oil** cooled	cooling
Sandia National Laboratories	SNL **Vortex**	NVIDIA V100 (**Volta**)	**216** GPUs	**water** cooled	
OAK RIDGE National Laboratory	ORNL **Summit**	NVIDIA V100 (**Volta**)	**27648** GPUs	**water** cooled	scale
Lawrence Livermore National Laboratory	LLNL **Corona**	AMD **MI60**	**328** GPUs	**air** cooled	GPU vendor

SGEMM across clusters



SGEMM across clusters



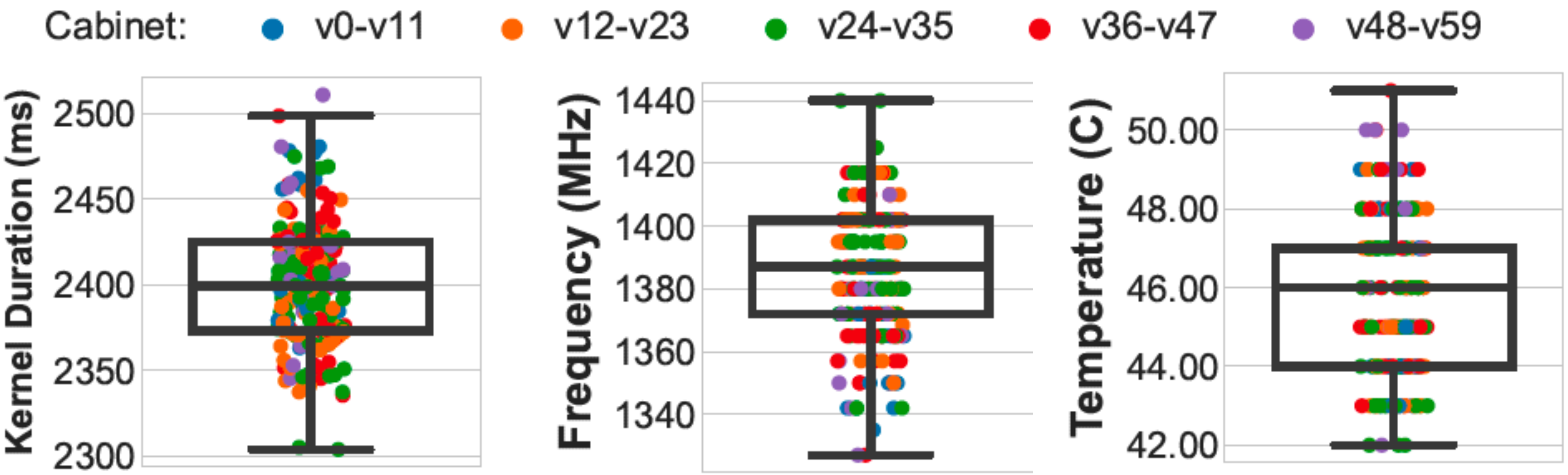
SGEMM performance varies between 7%-9% across GPUs in the same cluster, with outliers up to 1.5× slower than the median GPU.

Comparing cooling methods

TACC	TACC **Longhorn**	NVIDIA V100 (**Volta**)	**416** GPUs	**air** cooled
TACC	TACC **Frontera**	NVIDIA RTX5000 (**Turing**)	**360** GPUs	**mineral oil** cooled
Sandia National Laboratories	SNL **Vortex**	NVIDIA V100 (**Volta**)	**216** GPUs	**water** cooled
Oak Ridge National Laboratory	ORNL **Summit**	NVIDIA V100 (**Volta**)	**27648** GPUs	**water** cooled
Lawrence Livermore National Laboratory	LLNL **Corona**	AMD **MI60**	**328** GPUs	**air** cooled

cooling

Water-cooling: SGEMM on SNL Vortex

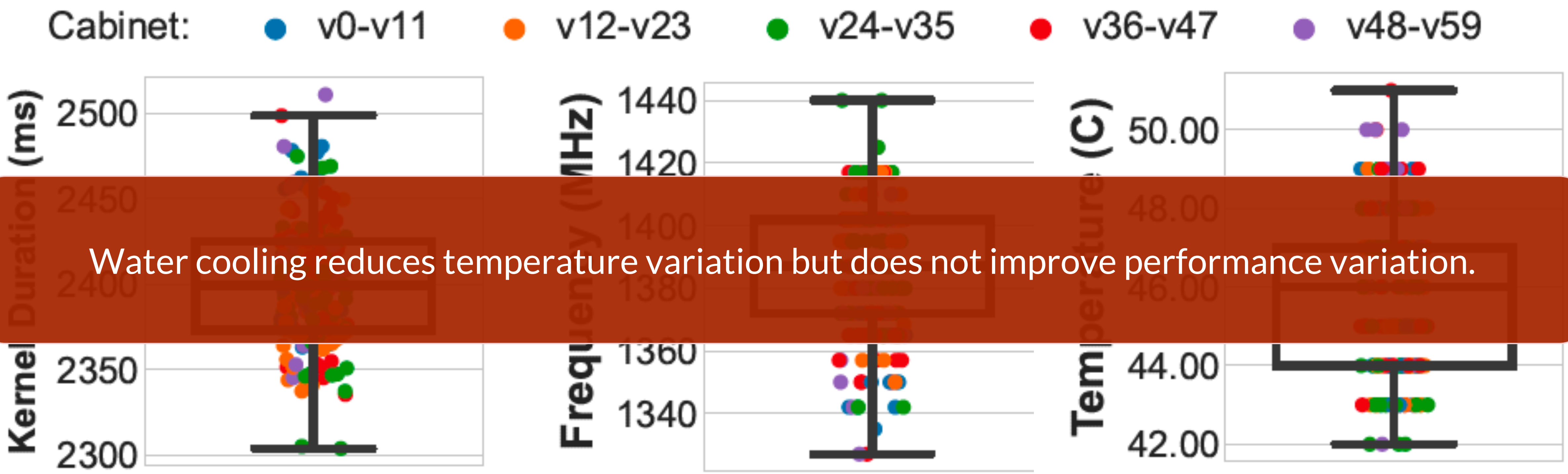


Performance variability
9%

Frequency variability
8.3%

Temperature variability
10°C b/n Q1 and Q3

Water-cooling: SGEMM on SNL Vortex



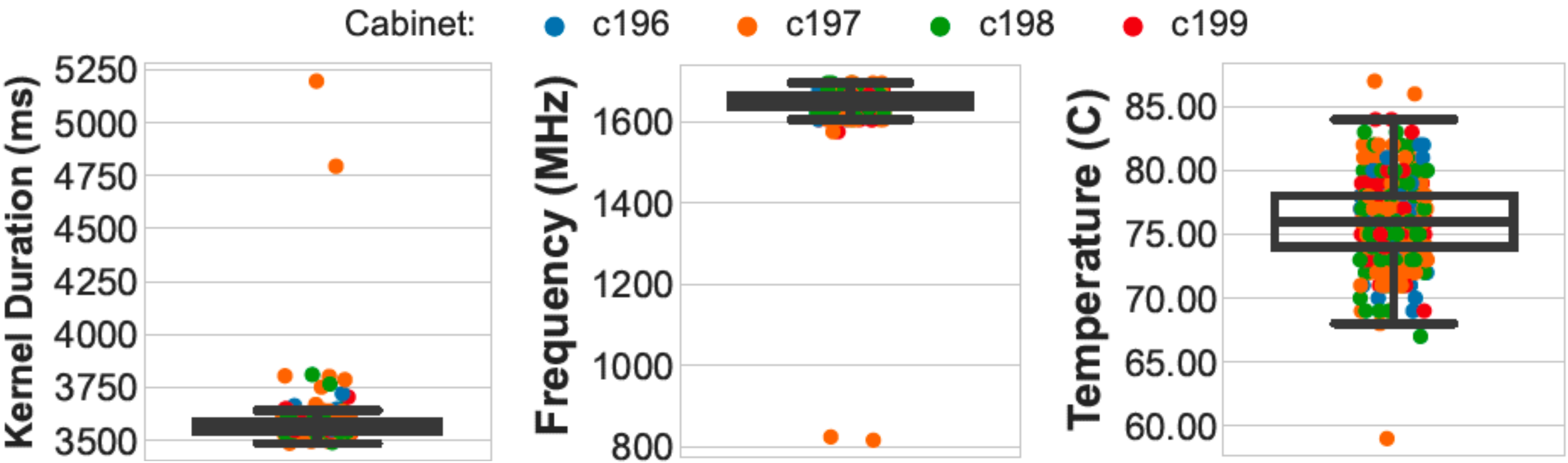
Water cooling reduces temperature variation but does not improve performance variation.

Performance variability
9%

Frequency variability
8.3%

Temperature variability
10°C b/n Q1 and Q3

Mineral oil cooling: SGEMM on TACC Frontera

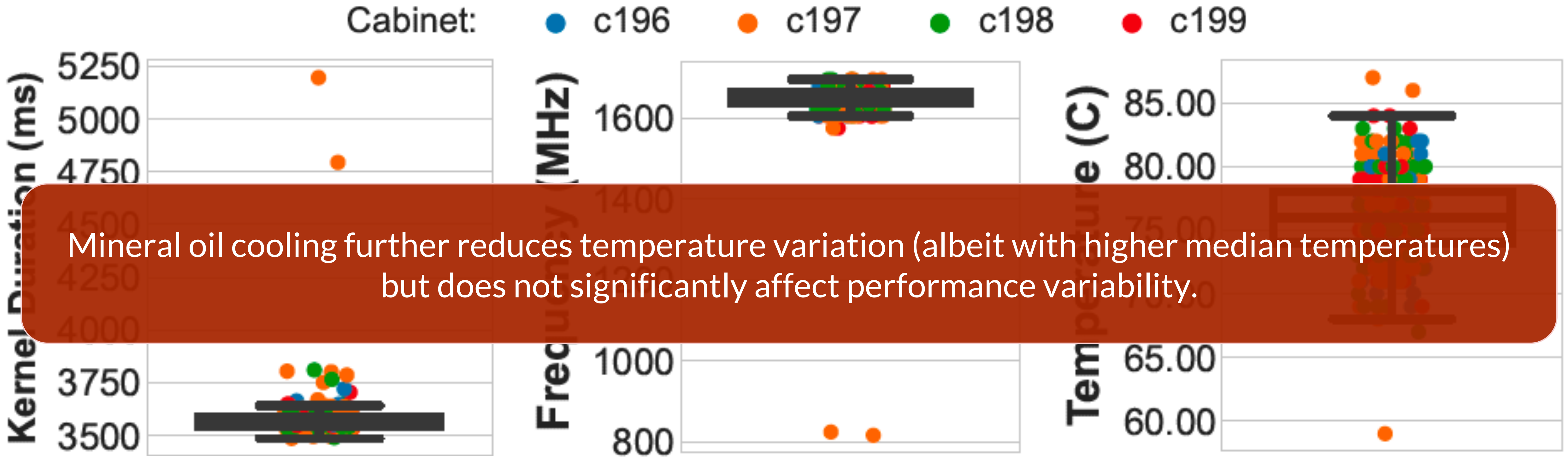


Performance variability
5%

Frequency variability
7%

Temperature variability
4°C b/n Q1 and Q3

Mineral oil cooling: SGEMM on TACC Frontera



Mineral oil cooling further reduces temperature variation (albeit with higher median temperatures) but does not significantly affect performance variability.

Performance variability
5%

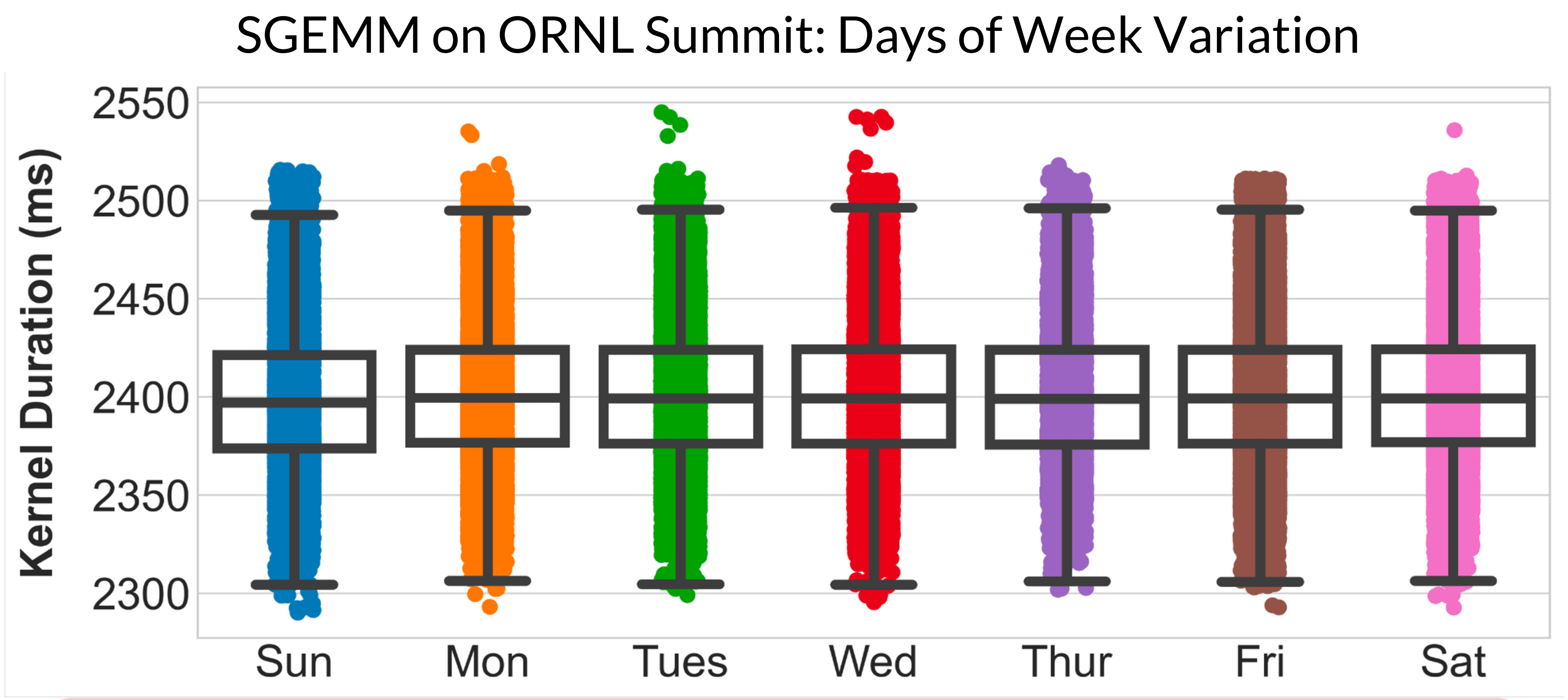
Frequency variability
7%

Temperature variability
4°C b/n Q1 and Q3

Key Questions

1. How much performance variation is there across GPUs?	**9% for SGEMM** **with outliers 1.5x slower than median**
2. Do GPU physical metrics (frequency, power and temperature) vary too?	**Yes**
3. How is variability affected by cluster parameters (size, cooling, GPU vendor)	• **Consistent perf variability across clusters** • **Liquid cooling reduces temperature variation, but not performance variability**
4. Are these variability observations consistent over time?	
5. Is variability application-dependent?	

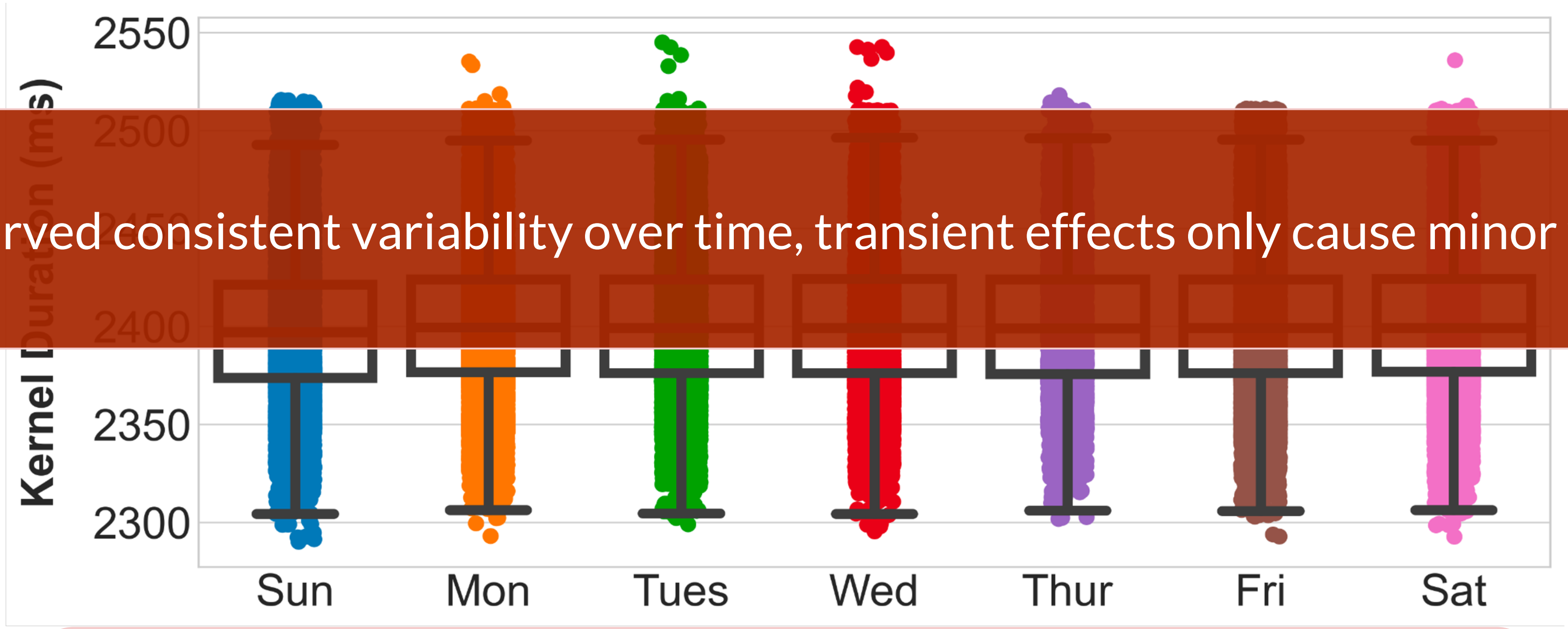
Variation over time?



Every day of the week has **8%** average performance variability

Variation over time?

SGEMM on ORNL Summit: Days of Week Variation

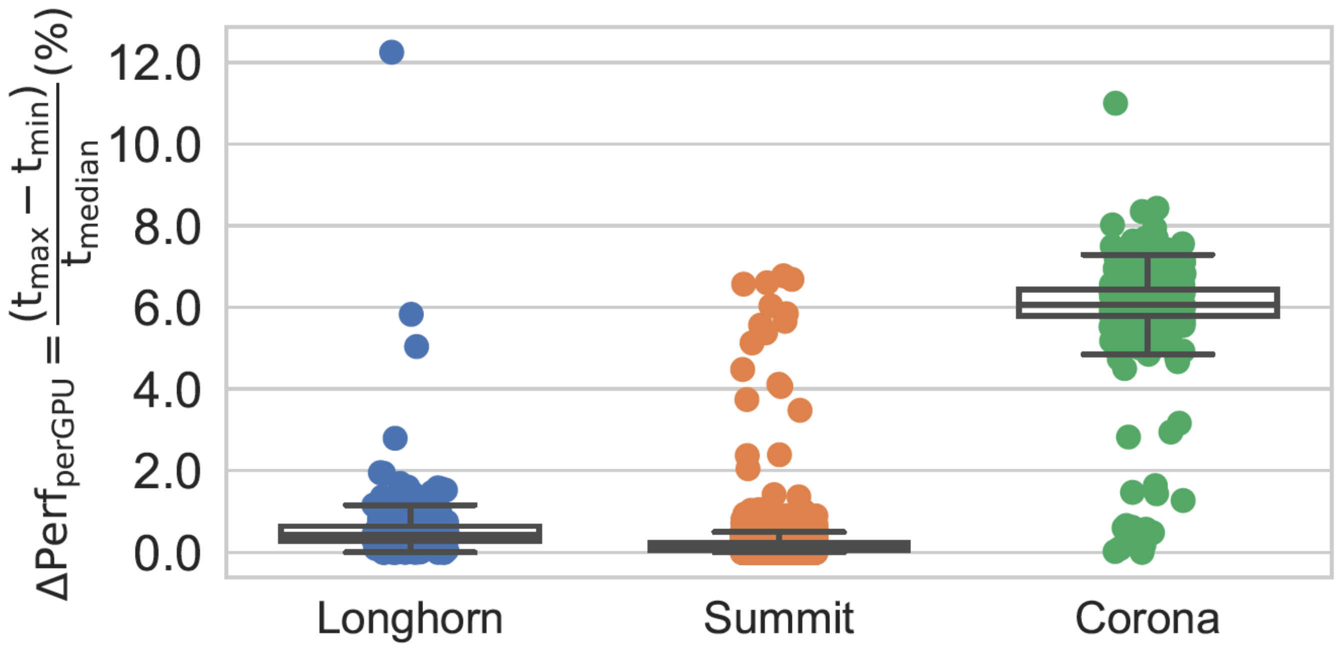


We observed consistent variability over time, transient effects only cause minor variations

Every day of the week has **8%** average performance variability

Variation over time?

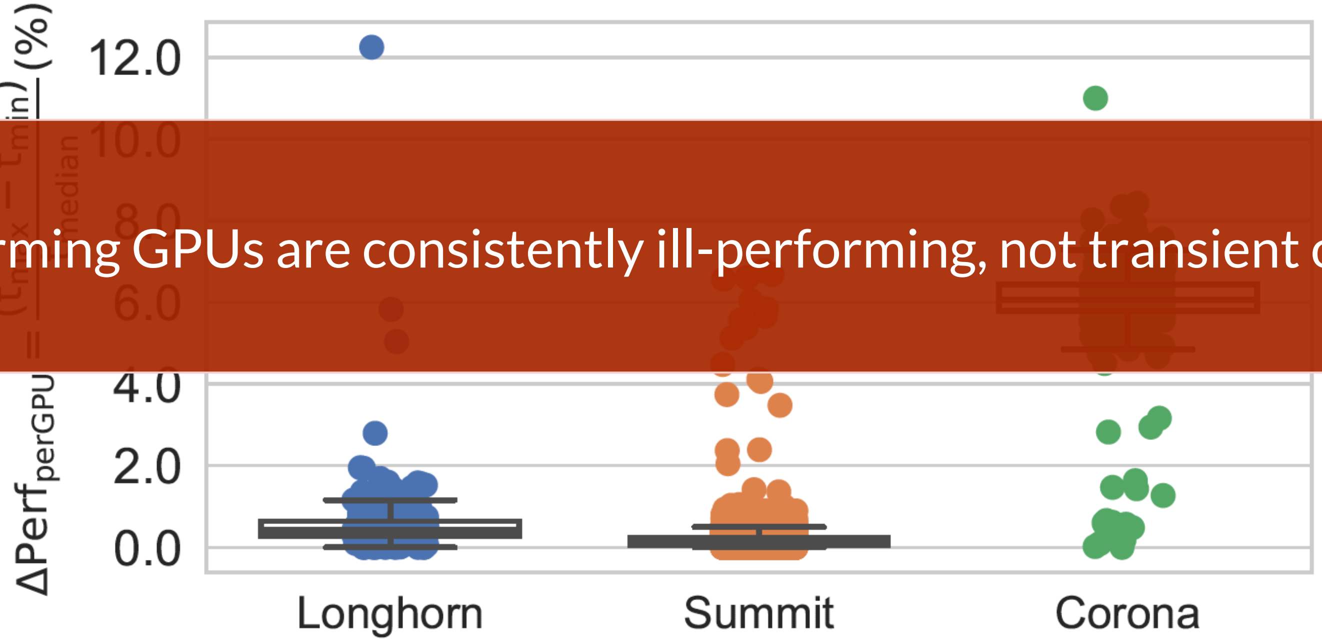
Per-GPU variation in performance over time
Each point shows normalized Δ(Performance) across 5 runs of SGEMM on same GPU



0.44% **0.12%** **6.06%**

Variation over time?

Per-GPU variation in performance over time
Each point shows normalized Δ(Performance) across 5 runs of SGEMM on same GPU



Ill-performing GPUs are consistently ill-performing, not transient outliers.

0.44% **0.12%** **6.06%**

Key Questions

1. How much performance variation is there across GPUs?	**9% for SGEMM with outliers 1.5x slower than median**
2. Do GPU physical metrics (frequency, power and temperature) vary too?	**Yes**
3. How is variability affected by cluster parameters (size, cooling, GPU vendor)	**consistent perf variability across clusters (other details in paper)**
4. Are these variability observations consistent over time?	**Yes, variability is not just a transient effect - consistent over time**
5. Is variability application-dependent?	

Comparing Applications (on NVIDIA GPUs)

Benchmark	**SGEMM**
Input Size	**25536 x 25536**

Comparing Applications (on NVIDIA GPUs)

Benchmark	**SGEMM**	**ResNet-50**
Input Size	**25536 x 25536**	Training set: **1.2M** images Batch Size: **64**

Comparing Applications (on NVIDIA GPUs)

Benchmark	**SGEMM**	**ResNet-50**	**BERT**
Input Size	**25536 x 25536**	Training set: **1.2M** images Batch Size: **64**	Training set: **30K words** Batch Size: **64**

Comparing Applications (on NVIDIA GPUs)

Benchmark	**SGEMM**	**ResNet-50**	**BERT**	**LAMMPS**
Input Size	**25536 x 25536**	Training set: **1.2M** images Batch Size: **64**	Training set: **30K words** Batch Size: **64**	(x,y,z) = (**8,16,16**)

Comparing Applications (on NVIDIA GPUs)

Benchmark	**SGEMM**	**ResNet-50**	**BERT**	**LAMMPS**	**PageRank**
Input Size	**25536 x 25536**	Training set: **1.2M** images Batch Size: **64**	Training set: **30K words** Batch Size: **64**	(x,y,z) = (**8,16,16**)	**642661** nodes **2785421** edges

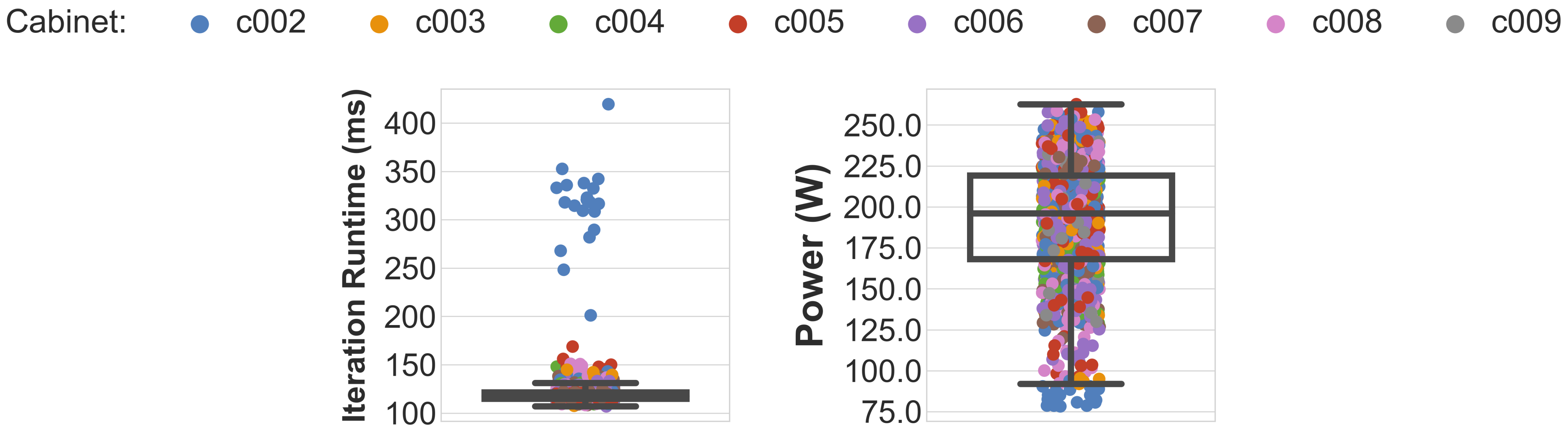
Comparing Applications (on NVIDIA GPUs)

	compute-intensive			memory-intensive	
Benchmark	**SGEMM**	**ResNet-50**	**BERT**	**LAMMPS**	**PageRank**
Input Size	**25536 x 25536**	Training set: **1.2M** images Batch Size: **64**	Training set: **30K words** Batch Size: **64**	(x,y,z) = (**8,16,16**)	**642661** nodes **2785421** edges

Comparing Applications (on NVIDIA GPUs)

	compute-intensive			memory-intensive	
Benchmark	**SGEMM**	**ResNet-50**	**BERT**	**LAMMPS**	**PageRank**
Input Size	**25536 x 25536**	Training set: **1.2M** images Batch Size: **64**	Training set: **30K words** Batch Size: **64**	(x,y,z) = (**8,16,16**)	**642661** nodes **2785421** edges
Single/Multi-GPU	single	multi-GPU (4-GPU)	multi-GPU (4-GPU)	single	single

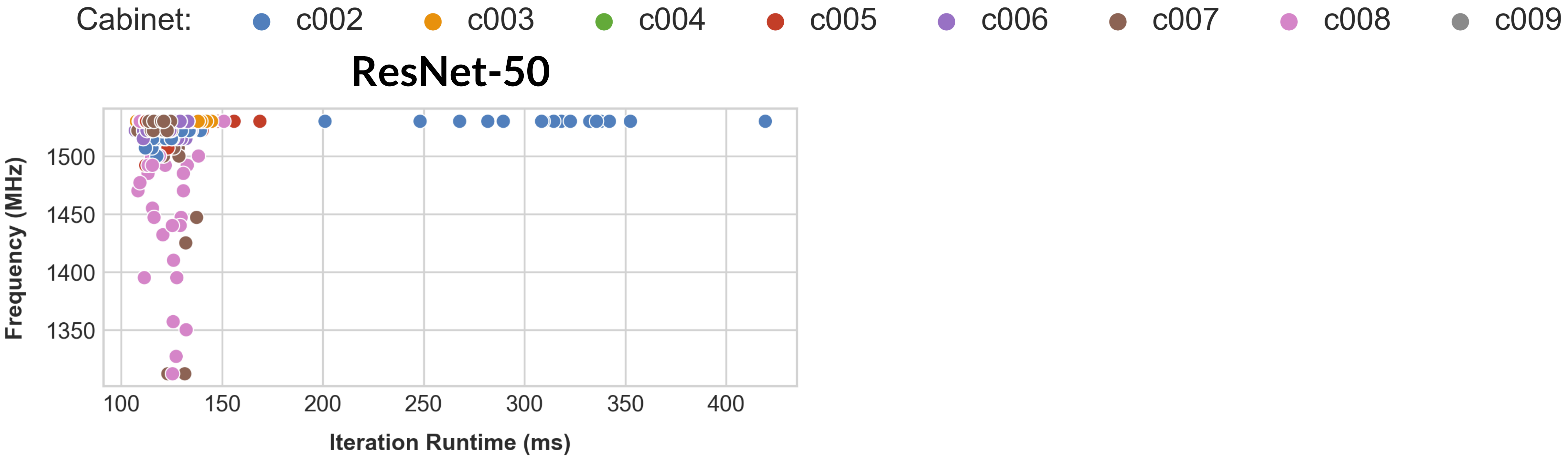
Multi-GPU ResNet-50 on Longhorn



performance
variability
22%

power
variability
104%

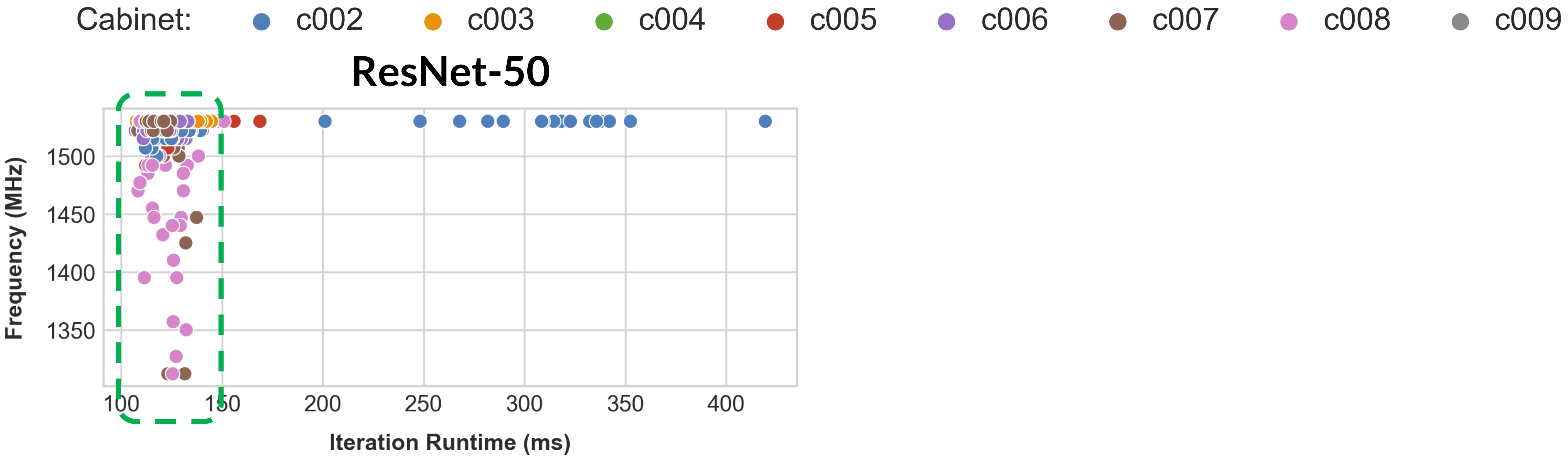
Frequency-Performance correlation



Weak inverse correlation

$\rho = -0.46$

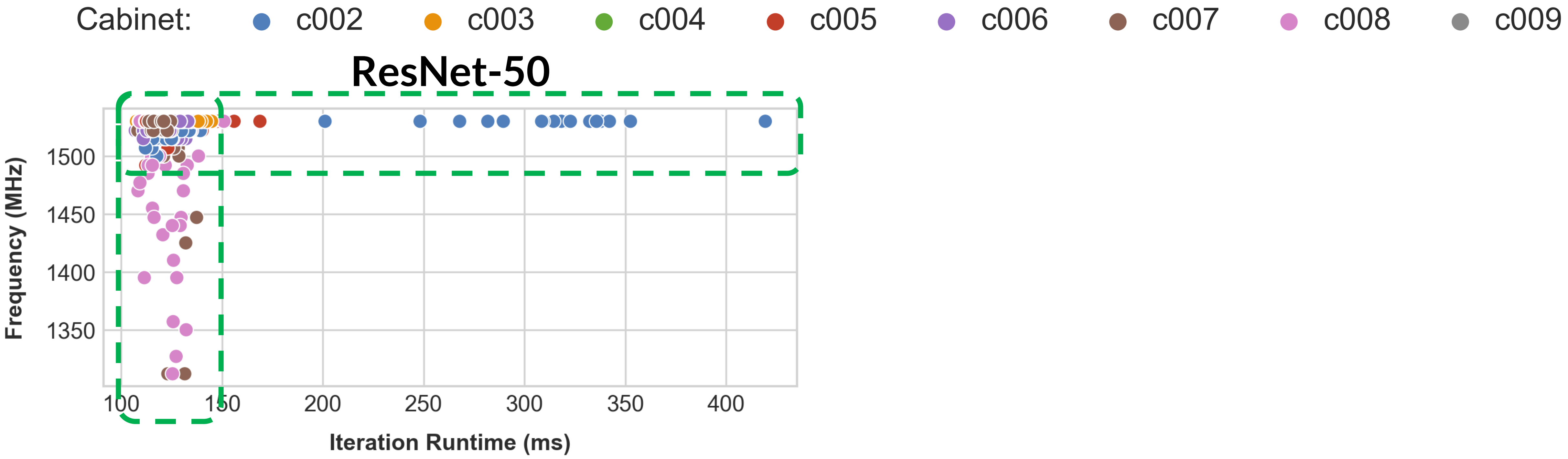
Frequency-Performance correlation



Weak inverse correlation

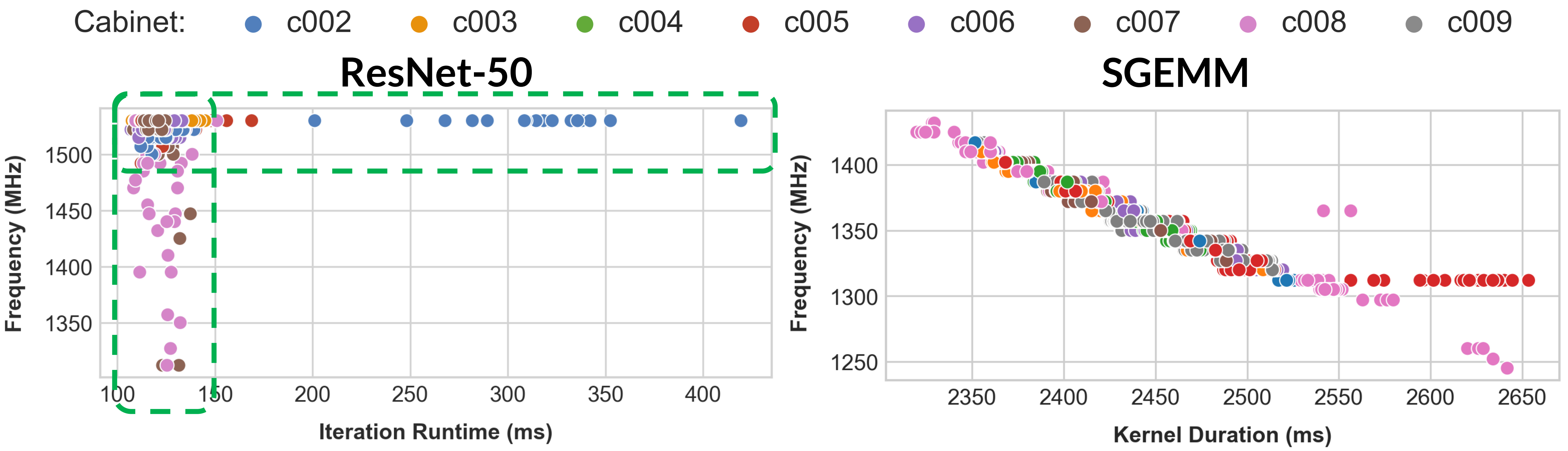
$\rho = -0.46$

Frequency-Performance correlation



Weak inverse correlation
$\rho = -0.46$

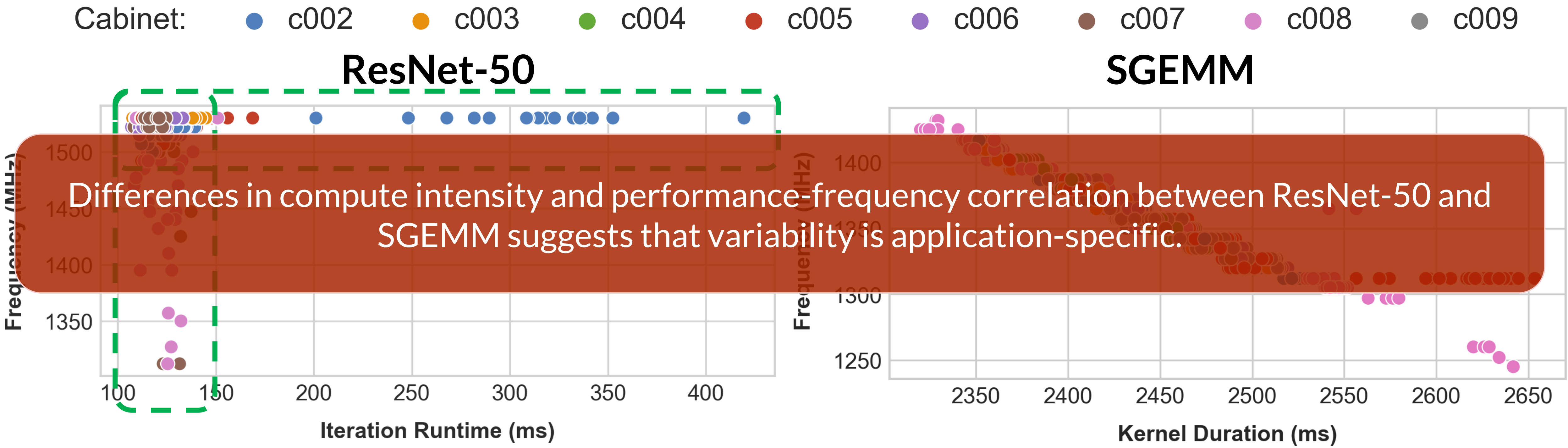
Frequency-Performance correlation



Weak inverse correlation
$\rho = -0.46$

Strong inverse correlation
$\rho = -0.94$

Frequency-Performance correlation



Differences in compute intensity and performance-frequency correlation between ResNet-50 and SGEMM suggests that variability is application-specific.

Weak inverse correlation
$\rho = -0.46$

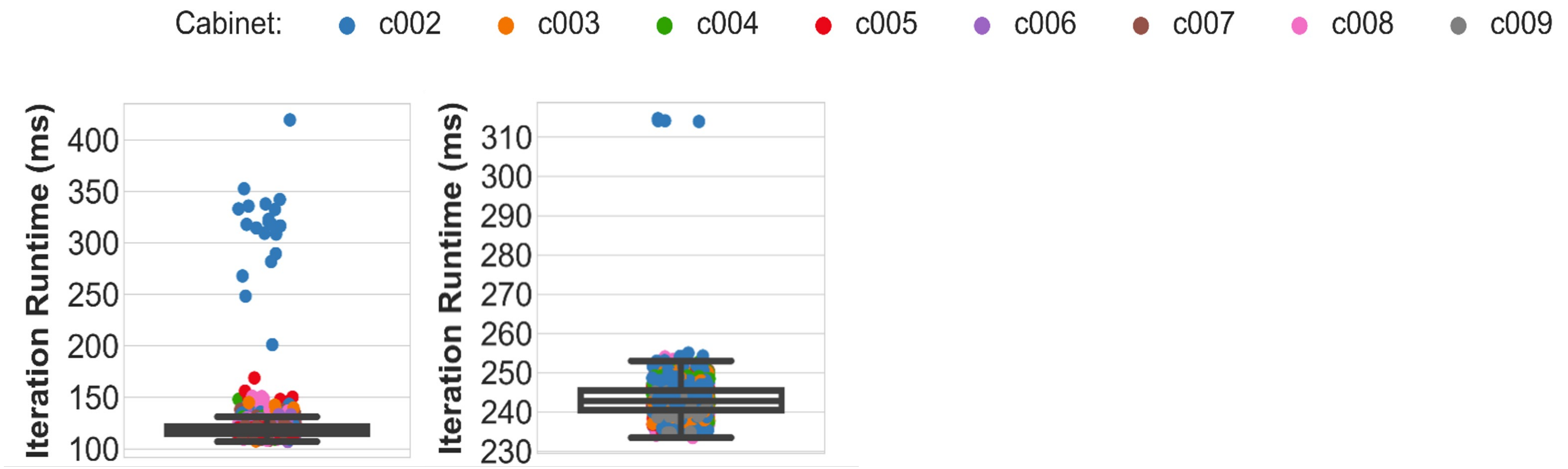
Strong inverse correlation
$\rho = -0.94$

Variability across applications

ResNet-50
22%

performance variability

Variability across applications



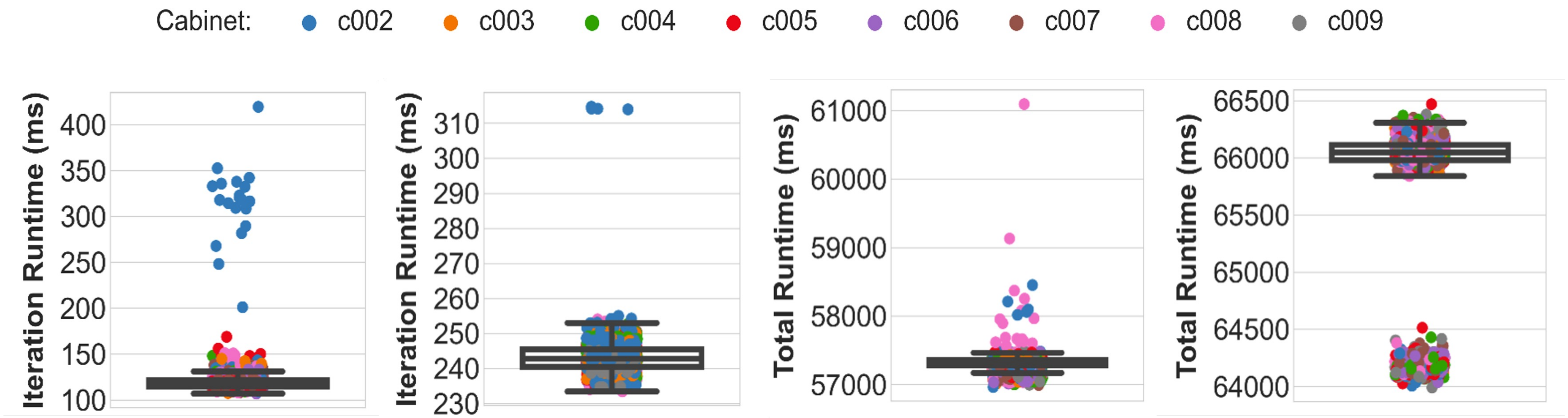
ResNet-50
22%

BERT
8%

performance variability

Variability across applications

- ❑ Variability is application-specific
- ❑ Ill-performing GPUs are consistently ill-performing
- ❑ Memory intensive apps see lower variability



ResNet-50
22%

BERT
8%

LAMMPS
<1%

PageRank
<1%

performance variability

Summary of findings

The following weren't covered in the talk, but are in the paper:	• *Variability with cluster scale* • *Variability with different GPU vendors (AMD/NVIDIA)* • *Effect of varying GPU Power Limit* • *Comparing single-GPU ResNet vs multi-GPU ResNet*
1. How much performance variation is there across GPUs?	9% for SGEMM with outliers 1.5x slower than median
2. Do GPU physical metrics vary too?	Yes
3. How is variability affected by cluster parameters	consistent perf variability across clusters
4. Is variability consistent over time?	Yes
5. Is variability application-dependent?	Yes, compute-intensive applications see more performance variability than memory intensive ones

Summary of findings

The following weren't covered in the talk, but are in the paper:	• *Variability with cluster scale* • *Variability with different GPU vendors (AMD/NVIDIA)* • *Effect of varying GPU Power Limit* *Comparing single-GPU ResNet vs multi-GPU ResNet*
1. How much performa	an
2. Do GPU physical me	
3. How is variability aff	usters
4. Is variability consistent over time?	Yes
5. Is variability application-dependent?	Yes, compute-intensive applications see more performance variability than memory intensive ones

What can we do?

Harnessing variability to our advantage

- **Blacklisting and Maintenance**
  - Periodic variability benchmarking
  - Identifying and performing targeted maintenance
- **Application-Aware Frameworks**
  - Variability-aware allocation and scheduling frameworks
  - Application-specific, dynamic allocation
- **New hardware and system design**
  - Standards for exposing variability info from HW to SW and runtime layers
  - Global power management techniques

Questions?

Summary

- Significant performance variability across clusters – 7-9% on average for SGEMM
- Variability much larger for compute-intensive workloads (ResNet-50 - 22%) than memory-intensive workloads (PageRank - 1%)
- Air-cooled clusters have larger temperature variation; water/mineral-oil cooling reduces temperature variation but does not help performance and power variation

Artifact

You can reproduce our experiments using:
https://github.com/hal-uw/gpu_variability_sc22_artifact

ArXiV

extended version of the paper
https://arxiv.org/abs/2208.11035

Contact/About

rnjain@cs.wisc.edu

https://pages.cs.wisc.edu/~rnjain/