Performance Heterogeneity in High Performance GPUs

Akhil Guliani, Prasoon Sinha,
Matthew Sinclair, Shivaram Venkatraman

COMPUTER SCIENCES
School of Computer, Data & Information Sciences

Collaborators:

Zhao Zhang, Oscar Hernandez, Clayton Hughes

TACC

Oak Ridge National Laboratory

Sandia National Laboratories

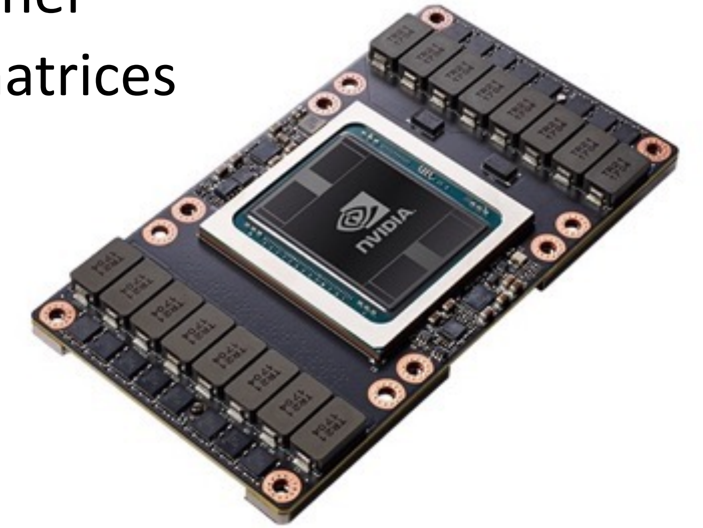
Previous Meetings: Setup

Power bound benchmark [3]

- cuBLAS SGEMM kernel
- Input: 25k by 25k matrices of 32bit floats

Measurements

- NVIDIA's nvprof profiler: Performance, Power, Temperature, Frequency
- Median values of 100 repetitions/run



NVIDIA V100-SXM2

- Set to max Frequency (1530 MHz)
- Set to TDP Power limit (300 W)

[3] Coplin, J., & Burtscher, M. Energy, power, and performance characterization of GPGPU benchmark programs. *In Proceedings - 2016 IEEE 30th International Parallel and Distributed Processing Symposium, IPDPS* 2016 (pp. 1190–1199). (2016). https://doi.org/10.1109/IPDPSW.2016.164

Study Locations

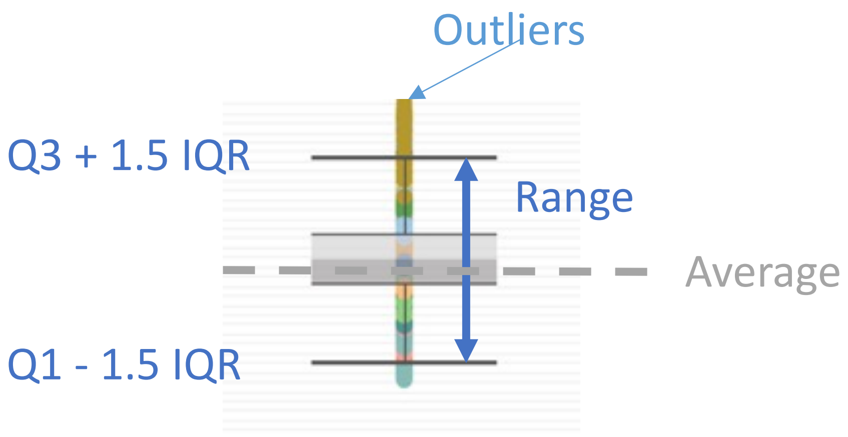
- CloudLab
  - 4 V100 SXM2 GPUs / node
  - Air cooled
  - 10’s of GPUs
- TACC’s Longhorn cluster
  - 4 V100 SXM2 GPUs / node
  - Air cooled + mineral cooled
  - 100’s of GPUs
- SNL’s Vortex cluster
  - 4 V100-SXM2 GPUs / node (Power9)
  - Water cooled
  - 100’s of GPUs
- ORNL’s Summit cluster
  - 6 V100-SXM2 GPUs / node
  - Water cooled
  - 10000’s of GPUs

Let’s define variation

$$\text{Relative range} = \frac{\text{Range}}{\text{Average}} \times 100\ \%$$

Using Inter Quartile Range (IQR) based method

Captures 99.3 % of the Gaussian distribution



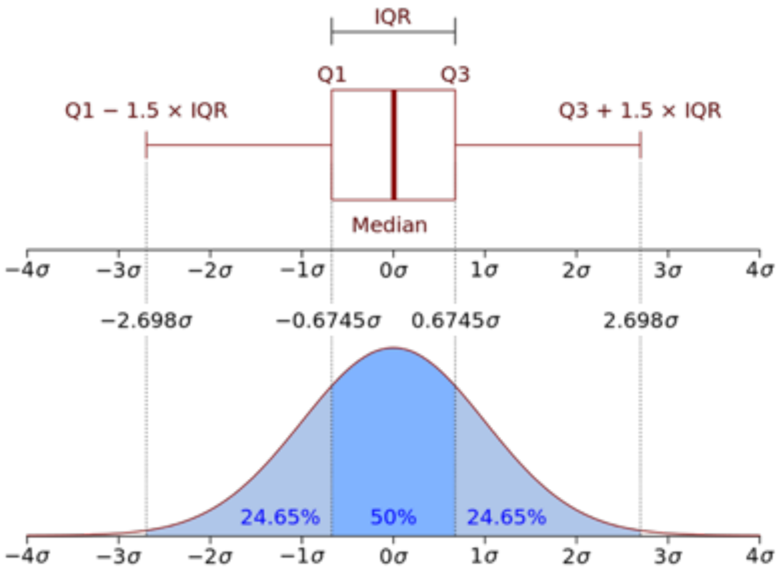
Outliers

Q3 + 1.5 IQR

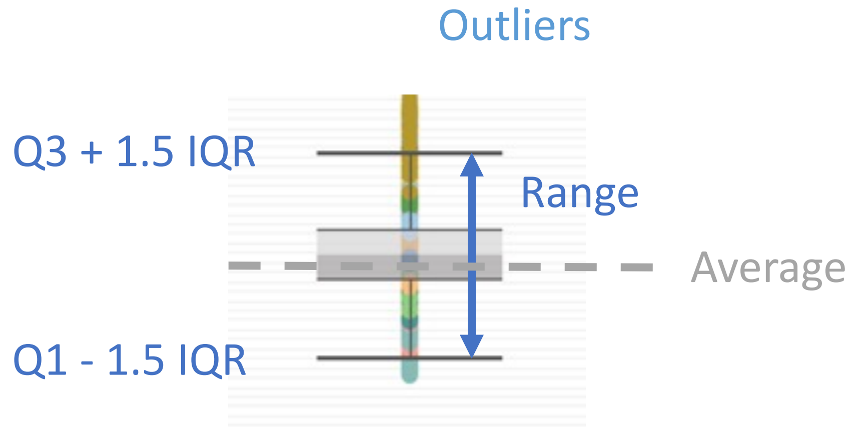
Range

Average

Q1 - 1.5 IQR



Last presentation (TACC)



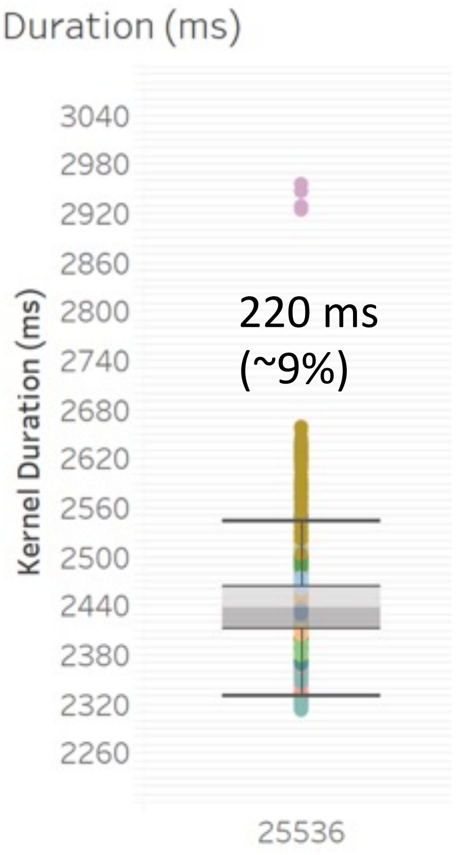
Outliers

Q3 + 1.5 IQR

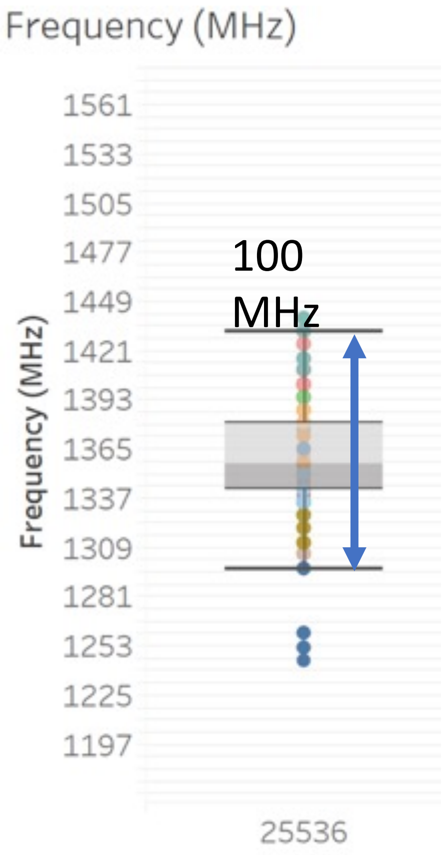
Range

Average

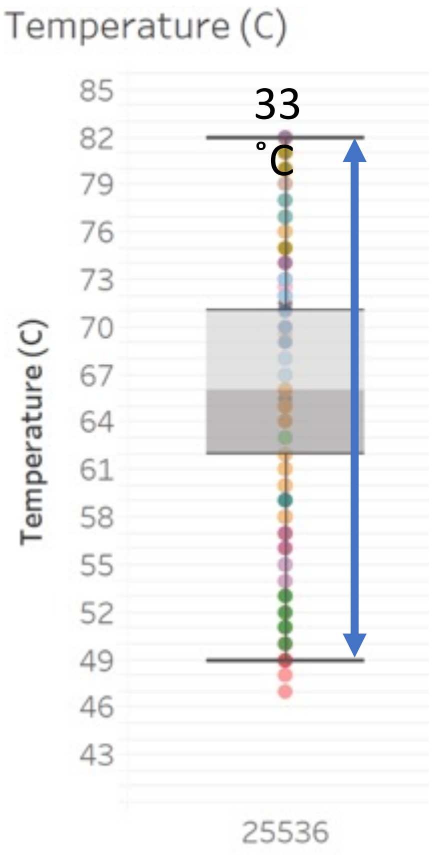
Q1 - 1.5 IQR



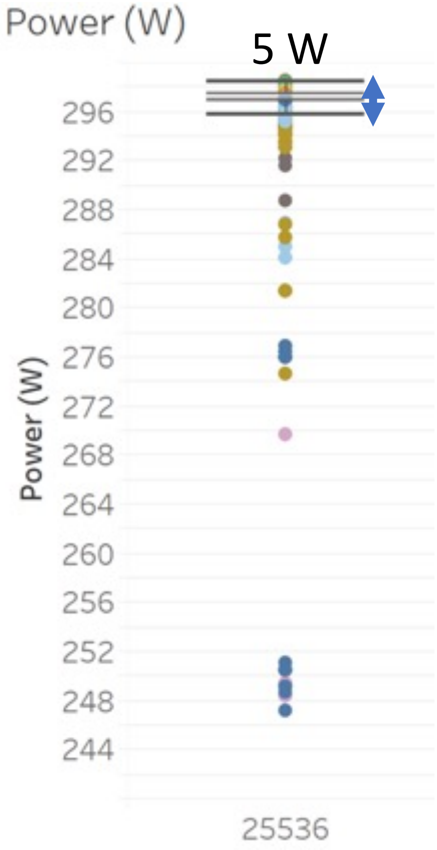
220 ms (~9%)



100 MHz



33 ˚C



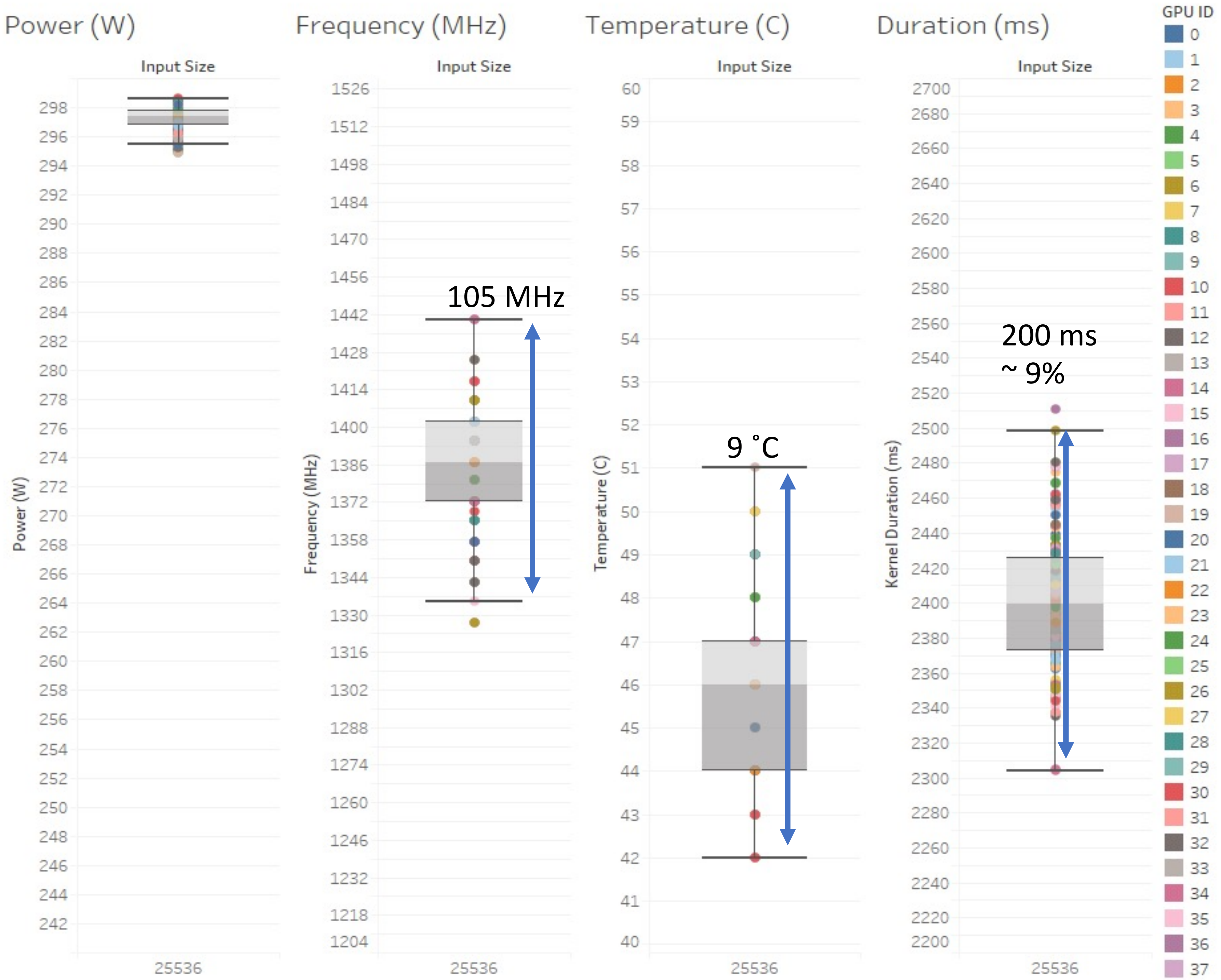
5 W

Last Presentation Vortex (SNL)

There is a 9% variation in kernel duration for the same workload

This variation is correlated to variation in operating frequency

Power management could be probable cause of this frequency variation

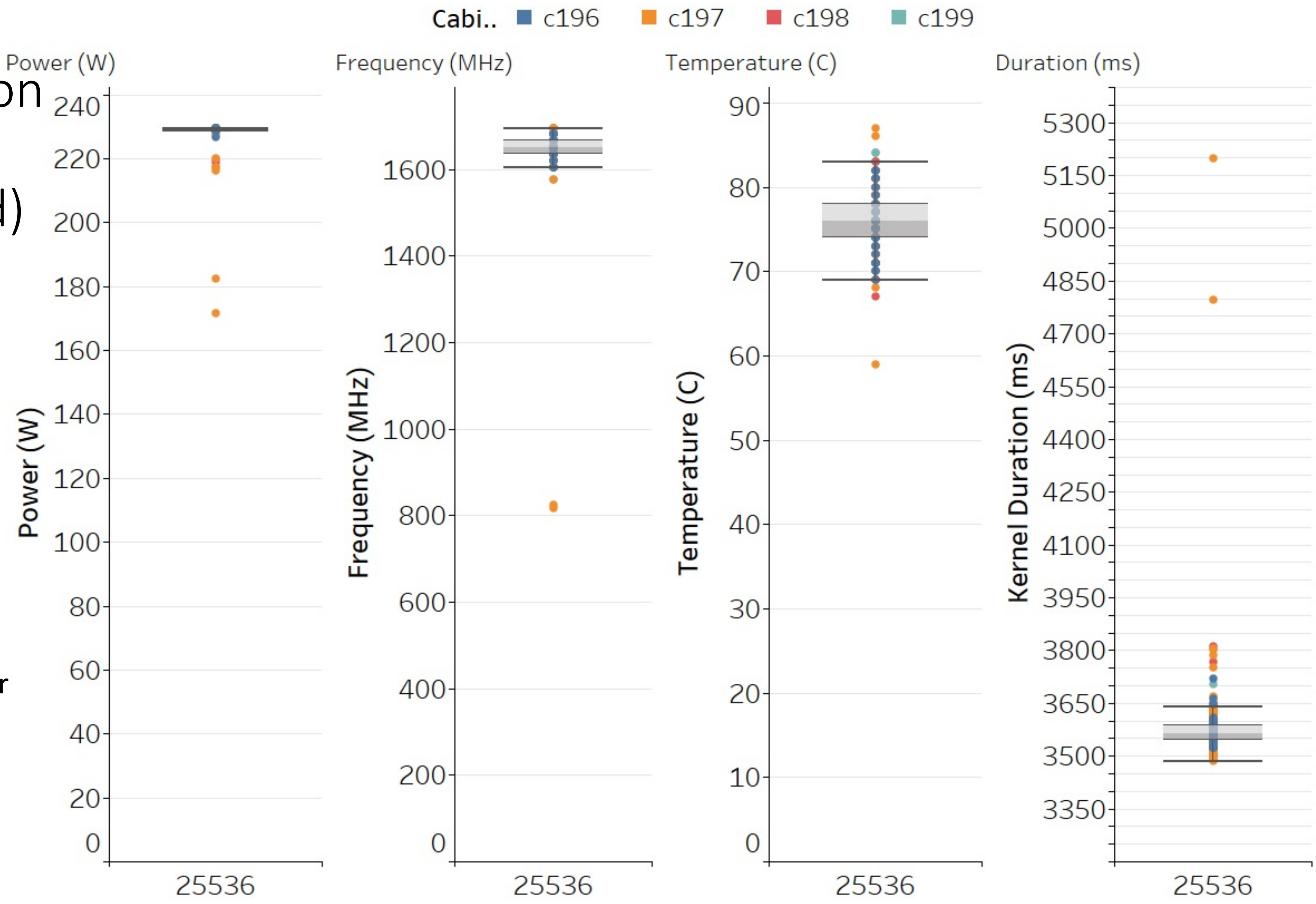


105 MHz

9 ˚C

200 ms
~ 9%

Last presentation Frontera (mineral cooled)



Power is around 235W (Quadro RTX)

Temperature band is narrower but also higher

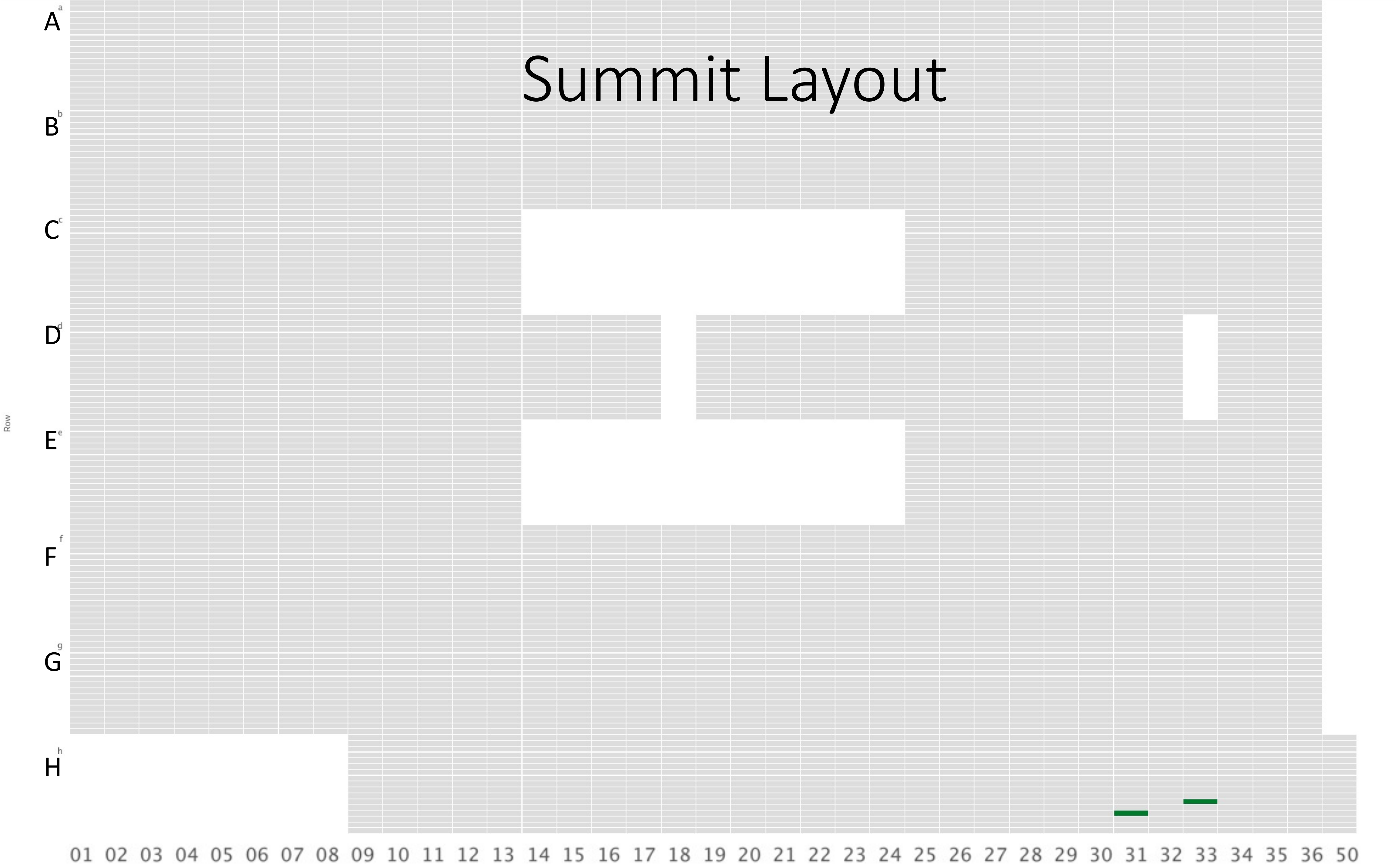
Kernel duration is higher (Quadro RTX)

Frequency is higher (Quadro RTX?)

Study Locations

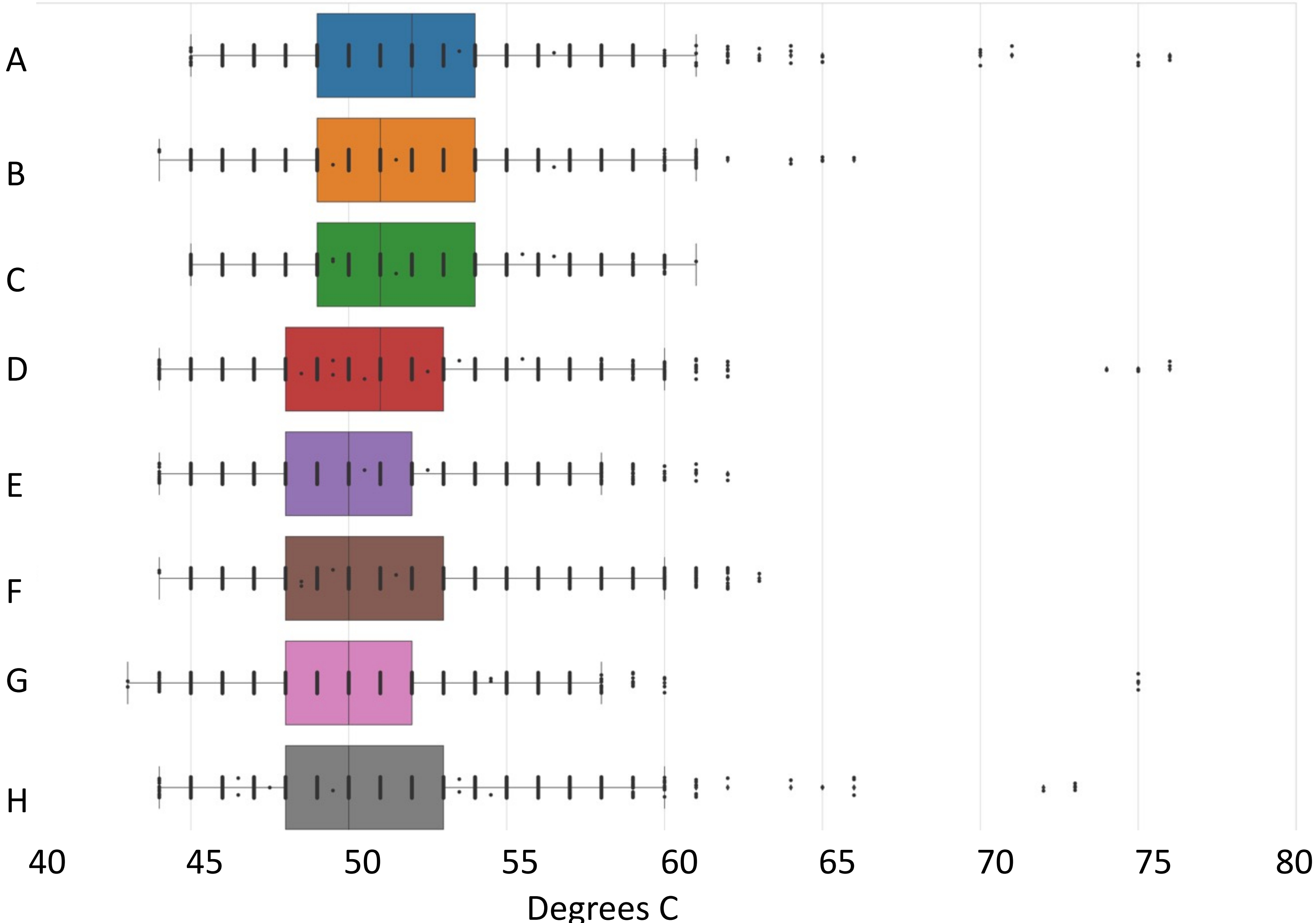
- CloudLab
  - 4 V100 SXM2 GPUs / node
  - Air cooled
  - 10’s of GPUs
- TACC’s Longhorn cluster
  - 4 V100 SXM2 GPUs / node
  - Air cooled + mineral cooled
  - 100’s of GPUs
- SNL’s Vortex cluster
  - 4 V100-SXM2 GPUs / node (Power9)
  - Water cooled
  - 100’s of GPUs
- ORNL’s Summit cluster
  - 6 V100-SXM2 GPUs / node
  - Water cooled
  - 10000’s of GPUs

Summit Layout



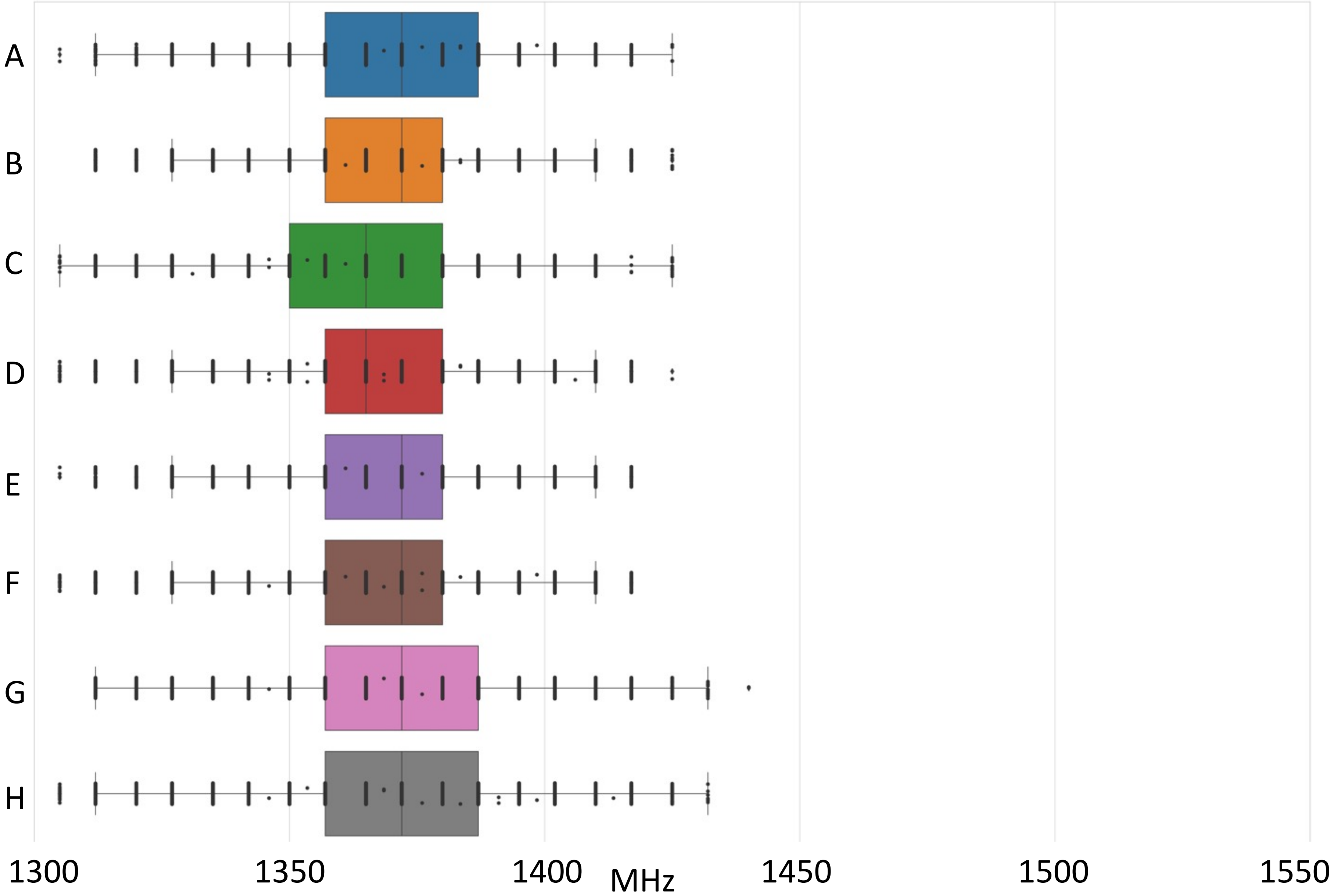
Summit Analysis (by row): Temperature

Takeaway:
Rows a, h see more outliers



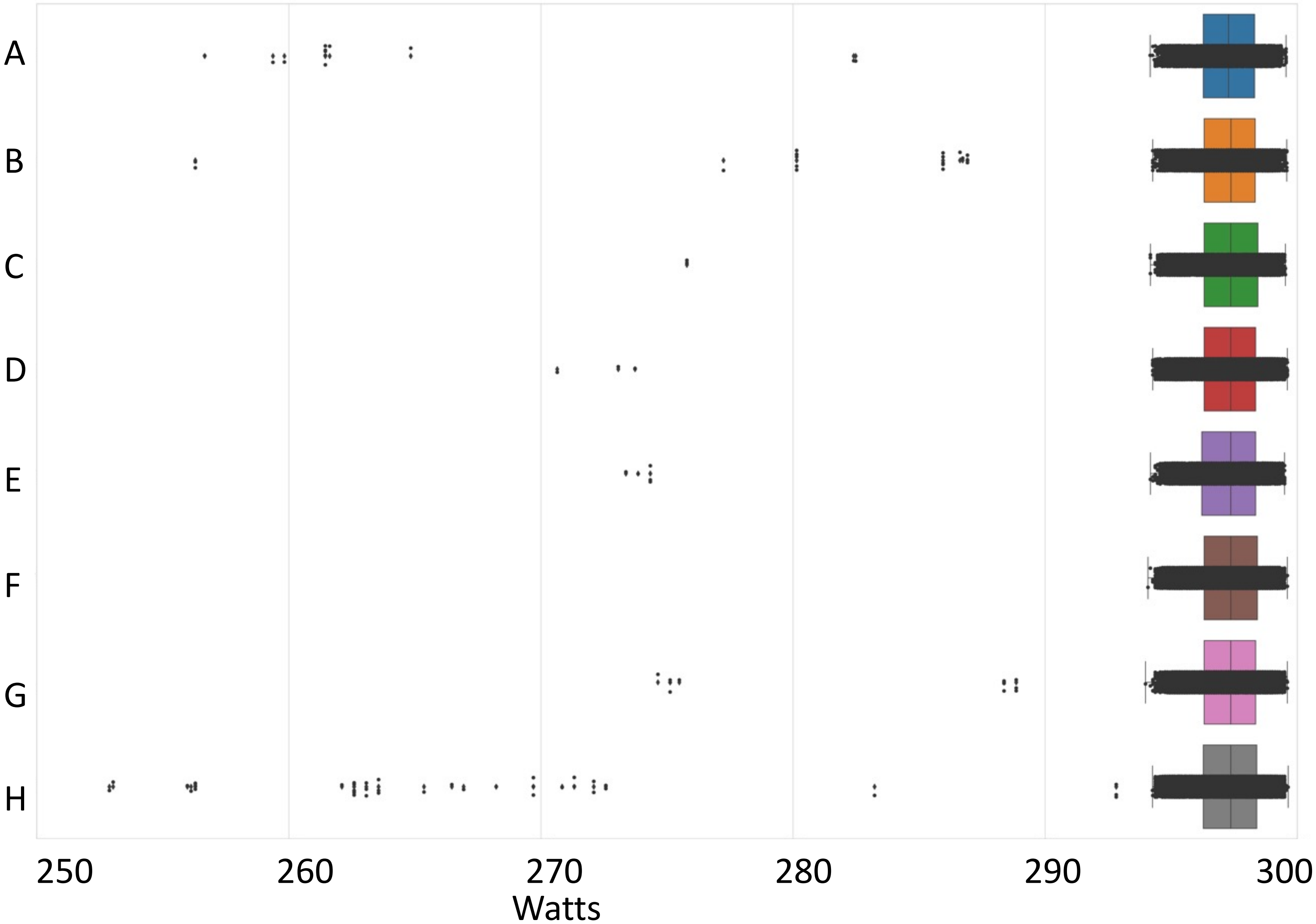
Summit Analysis (by row): Frequency

Takeaway: Outliers across many rows, but outlier range is not the same in each row



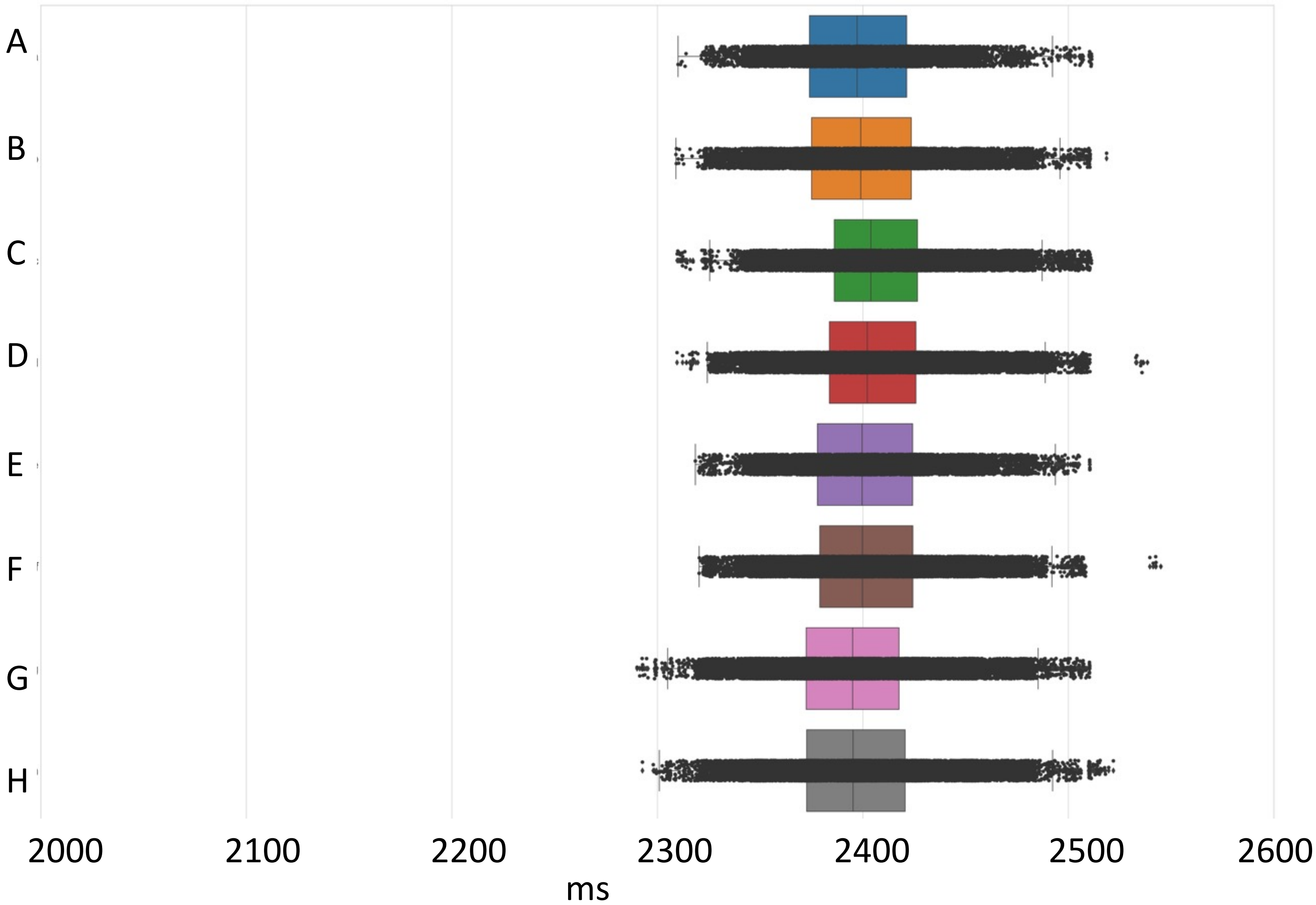
Summit Analysis (by row): Power

Takeaway: Similar to TACC some machines in row a, row h are at around 250W

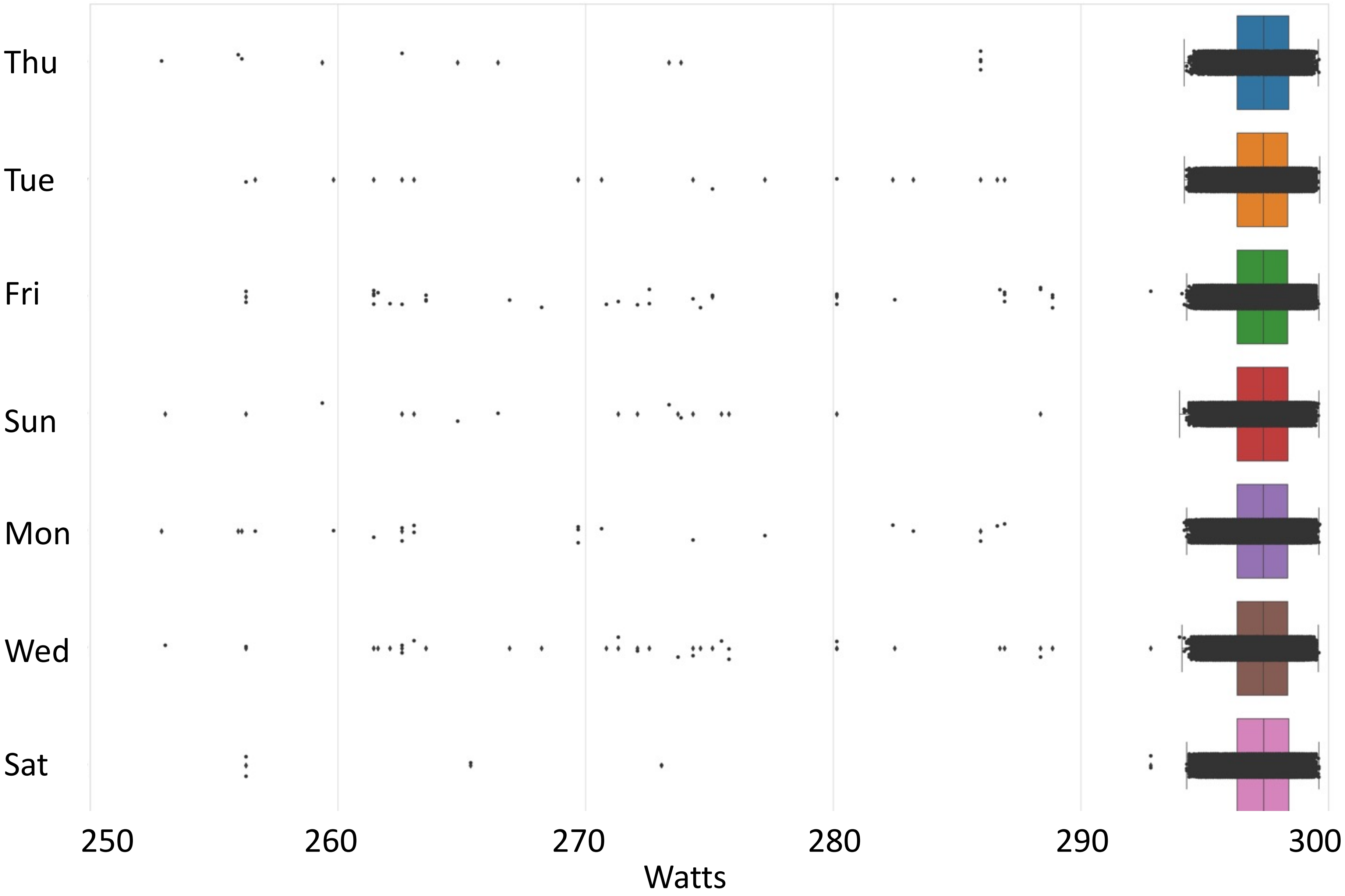


Summit Analysis (by row): Performance

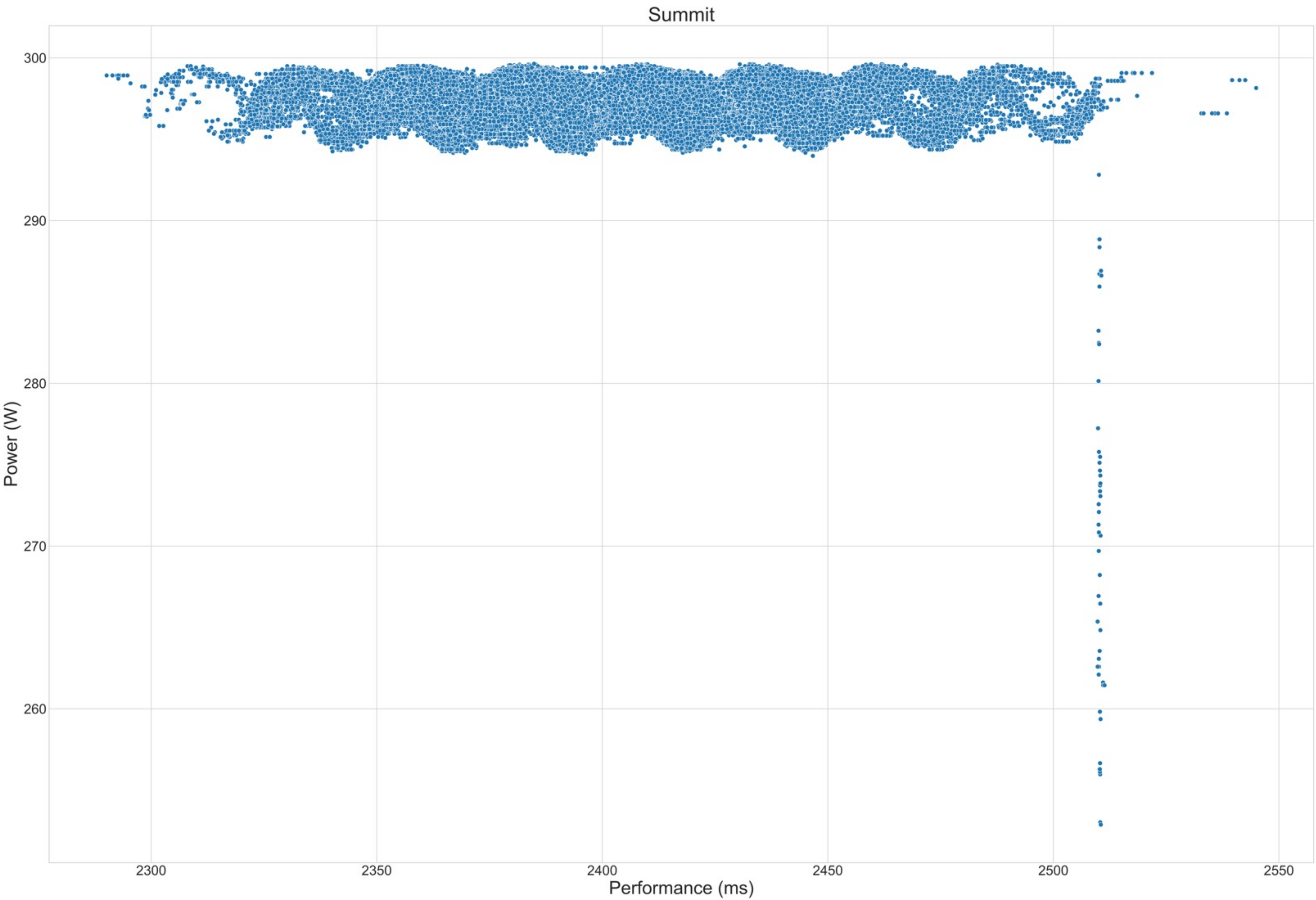
Takeaway: Performance outliers in row d and f are most severe



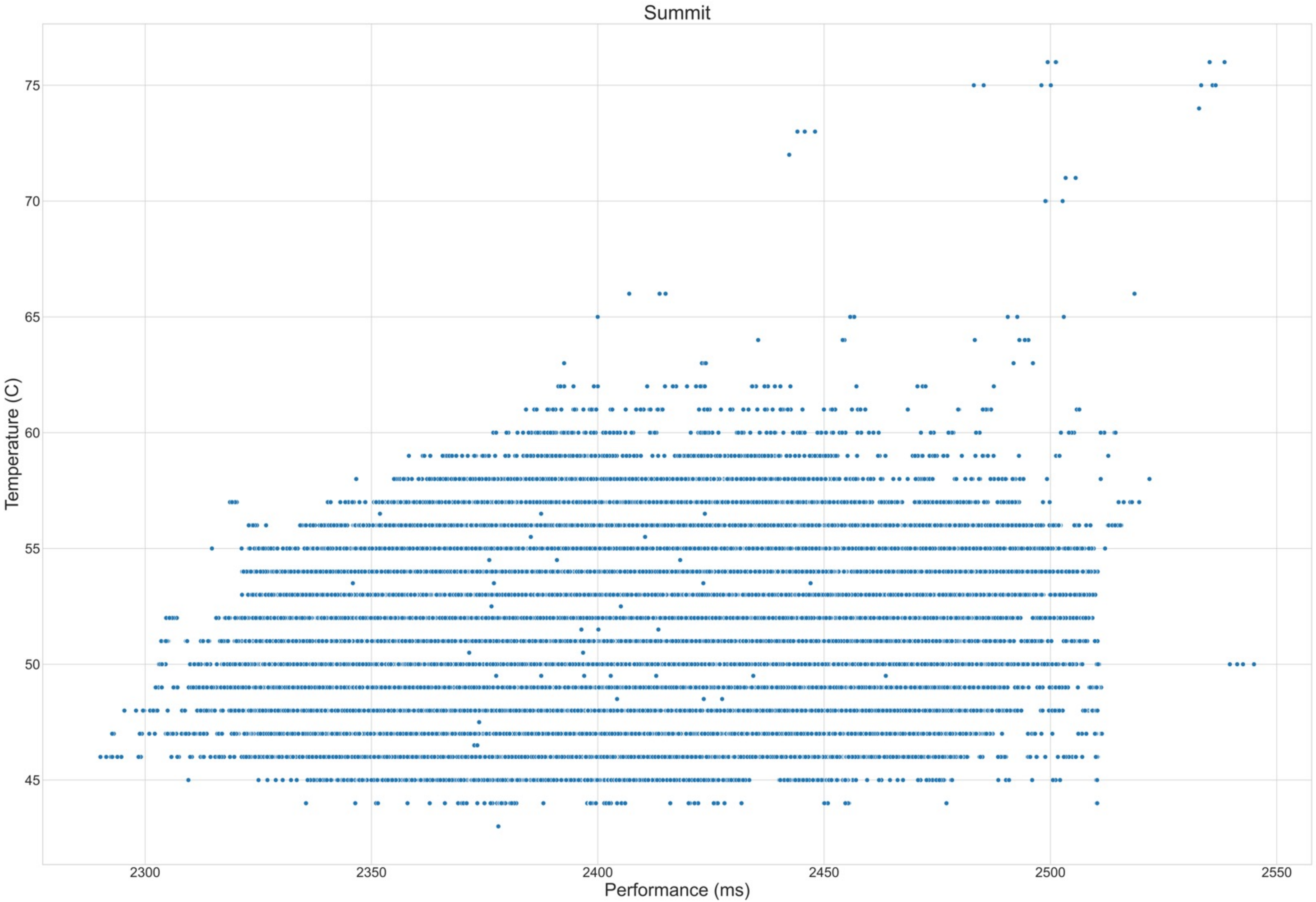
Summit Analysis: Day of the week (Power)



Perf vs Power



Perf vs Temperature



Work In Progress

- More diverse workloads (DOE Proxy Apps, Graph Analytics, ML):
  - Compute-Intensive: HACC, LAMMPS (EAM, ReactFF)
    - Irregular: Quicksilver
  - Memory-Intensive: AMR, LULESH, ML Training, QMCPack, SNAP, STREAM, XSBench
    - Irregular: Nekbone, Graph Analytics
    - Shared Memory Bound: Finite Element, Kripke
    - Latency Bound: Pennant (lots of pointer chasing)
  - Balanced: CoMD, LAMMPS (EAM), HPGMG