Accurately Benchmarking Power & Energy

Jae-Won Chung

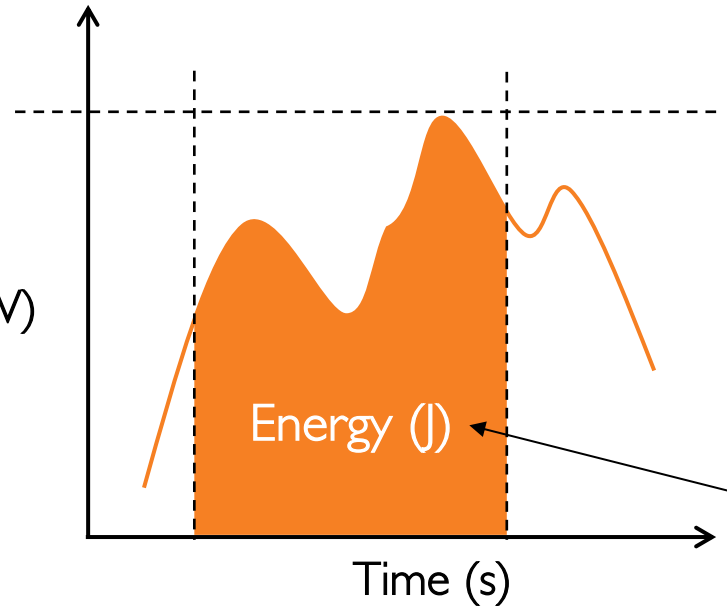
December 2nd, 2025

Outline

- Definitions and metrics
- Tools for measurement
- Measurement common pitfalls

- Hands-on session: Benchmarking image generation on Colab!

Power and Energy



Power capacity/Max power
Any hardware or datacenter has its max power draw ceiling

Total consumption;
what you are billed for.
Unit: Joules (J), Watt-hours (Wh)

Generated by power plants,
delivered by the grid every moment.
Unit: Watts

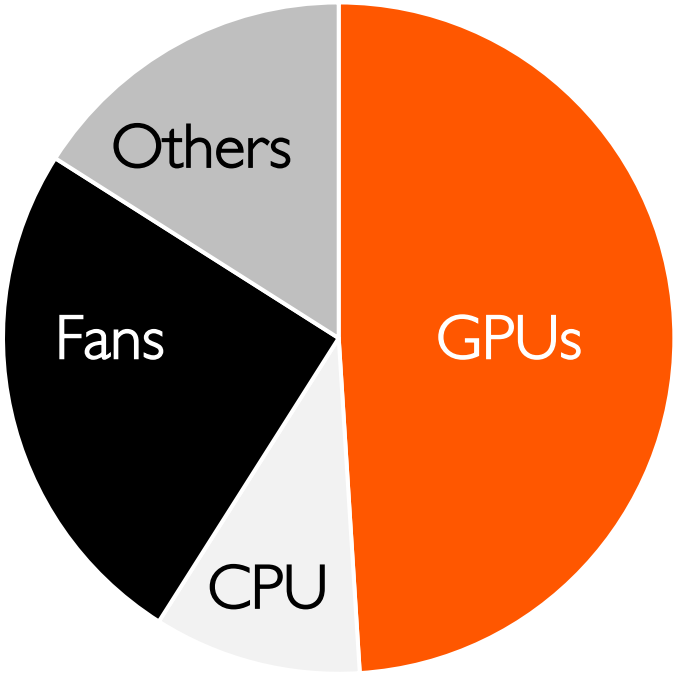
Power and Energy

- Power (W): Rate of energy consumption (Joule/s)
- Throughput (work/s): Rate of work getting done

- Throughput/power: Throughput you get per Watt of power
  - Datacenter capacity is described in terms of power (e.g., 1 GW)
  - Throughput/power = (Work/s) / (Joule/s) = Work/energy

Tools for Measurement

- A computer has a few components
  - CPU
  - Memory
  - GPU
  - Disk
- In a datacenter, there's also
  - Cooling
  - Networking



Provisioned Power
(8xA100-80GB AI Server)

Patel et al., ASPLOS, 2024

Tools for Measurement

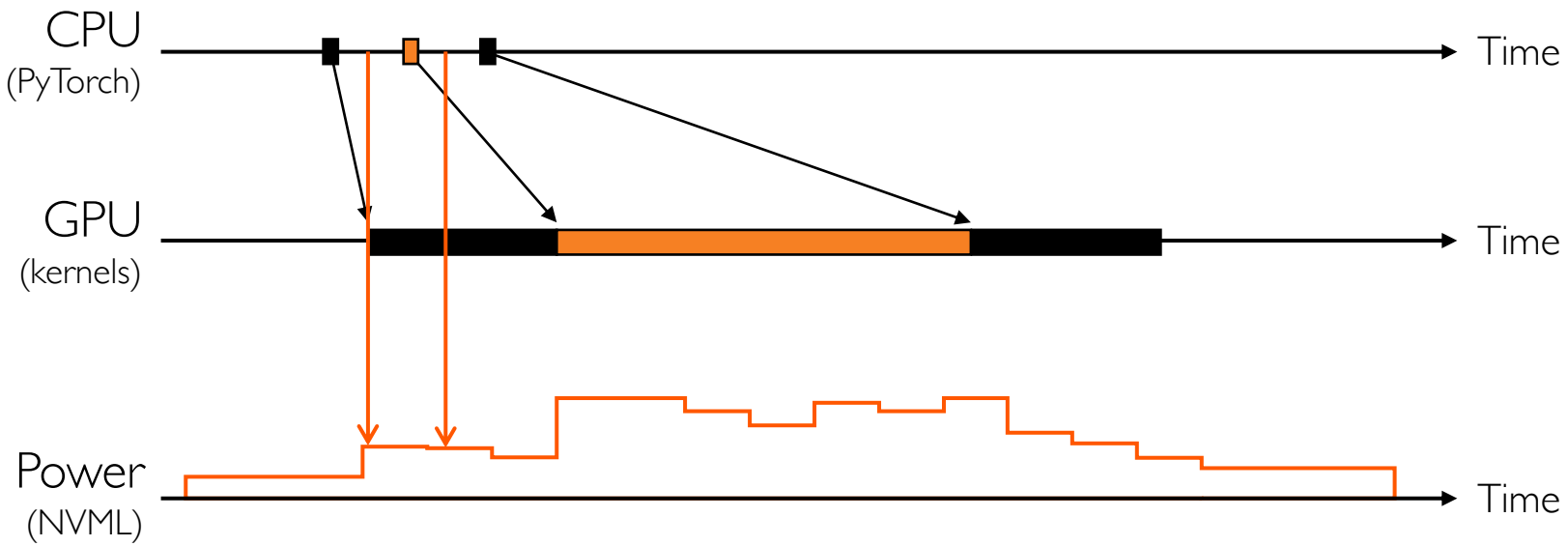
- CPU and DRAM: Intel RAPL
  - Counters tracked and exposed by the CPU & sysfs
  - CPU package
    - Supported by most CPUs (e.g., Intel, AMD)
  - DRAM
    - Supported by some CPUs (e.g., Intel)

Tools for Measurement

- GPU: Vendor-specific management libraries
  - NVML (NVIDIA) and AMDSMI (AMD)
    - C++ libraries with Python bindings
  - NVML counters
    - Power draw
      - Windowed average (1 second) or instant power draw
      - For the whole module (e.g., Grace Hopper chips), GPU, or HBM
    - Total energy consumption
      - Cumulative energy since driver load
      - Subtracting two points gives energy consumption between a time window

Measurement Pitfalls

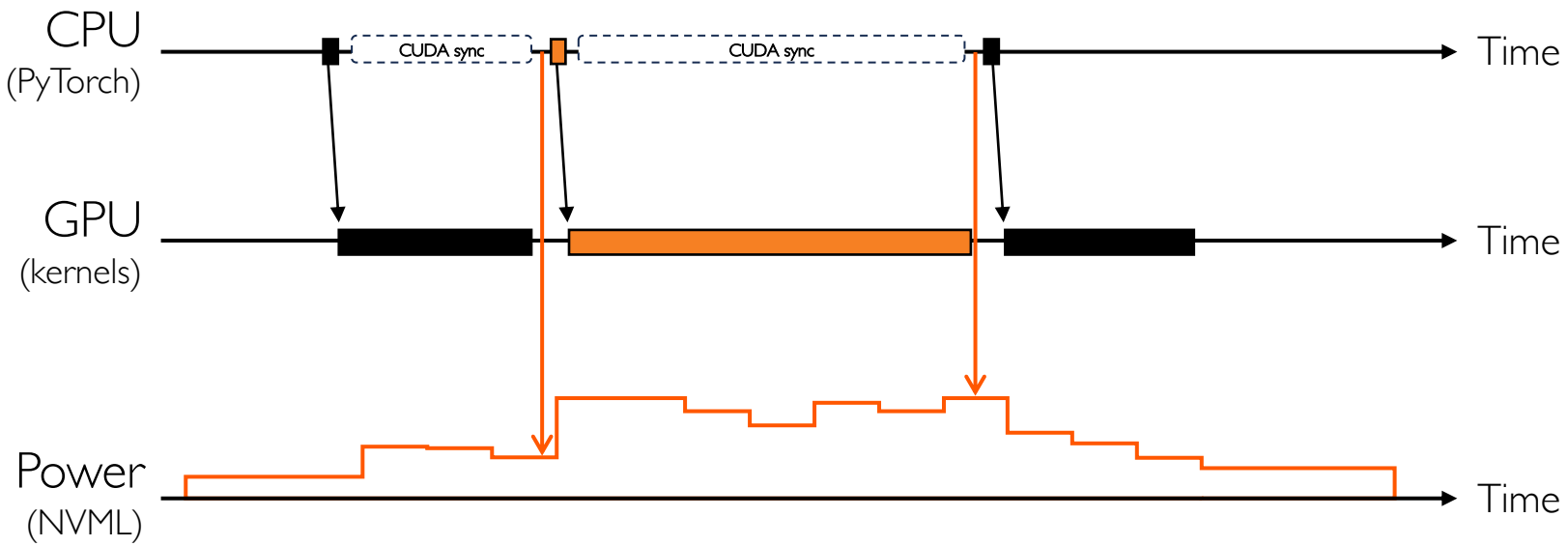
- CUDA synchronization
  - NVML reflects what's going on in the GPU, not the CPU



Read power from the wrong range of time!

Measurement Pitfalls

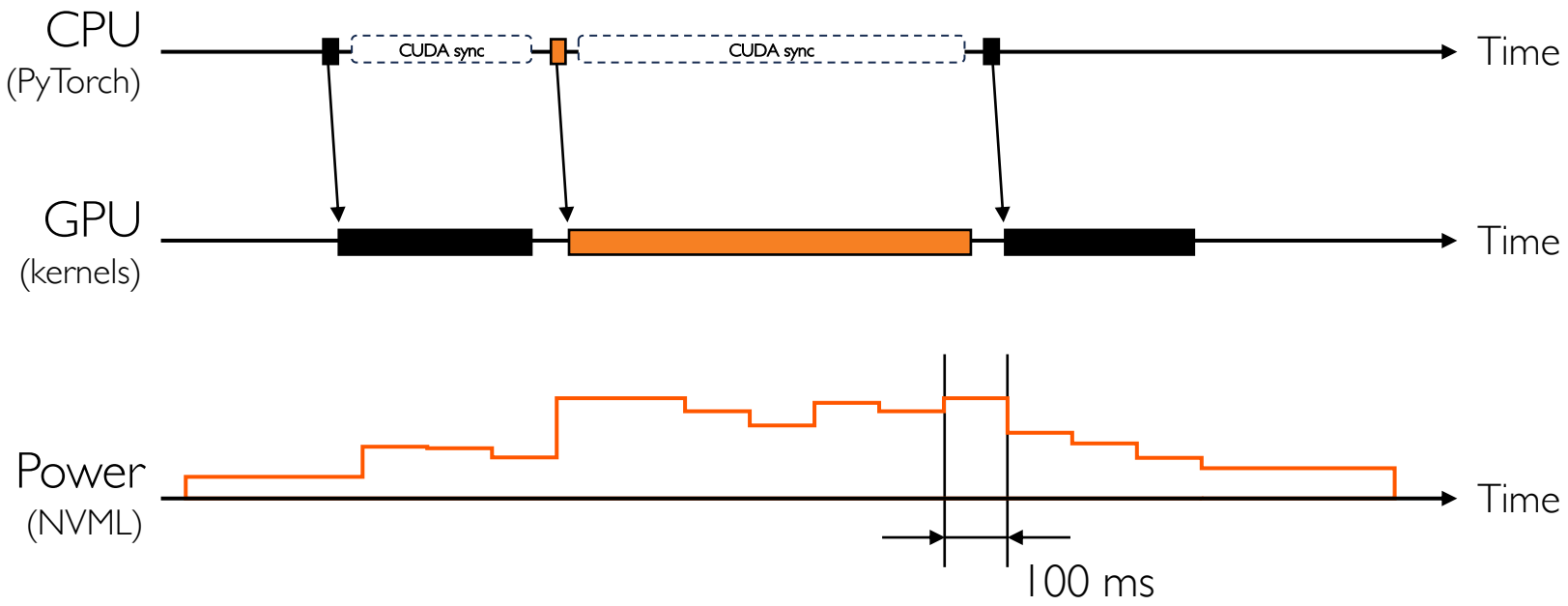
- CUDA synchronization
  - NVML reflects what's going on in the GPU, not the CPU



Measurement taken from reasonable time range

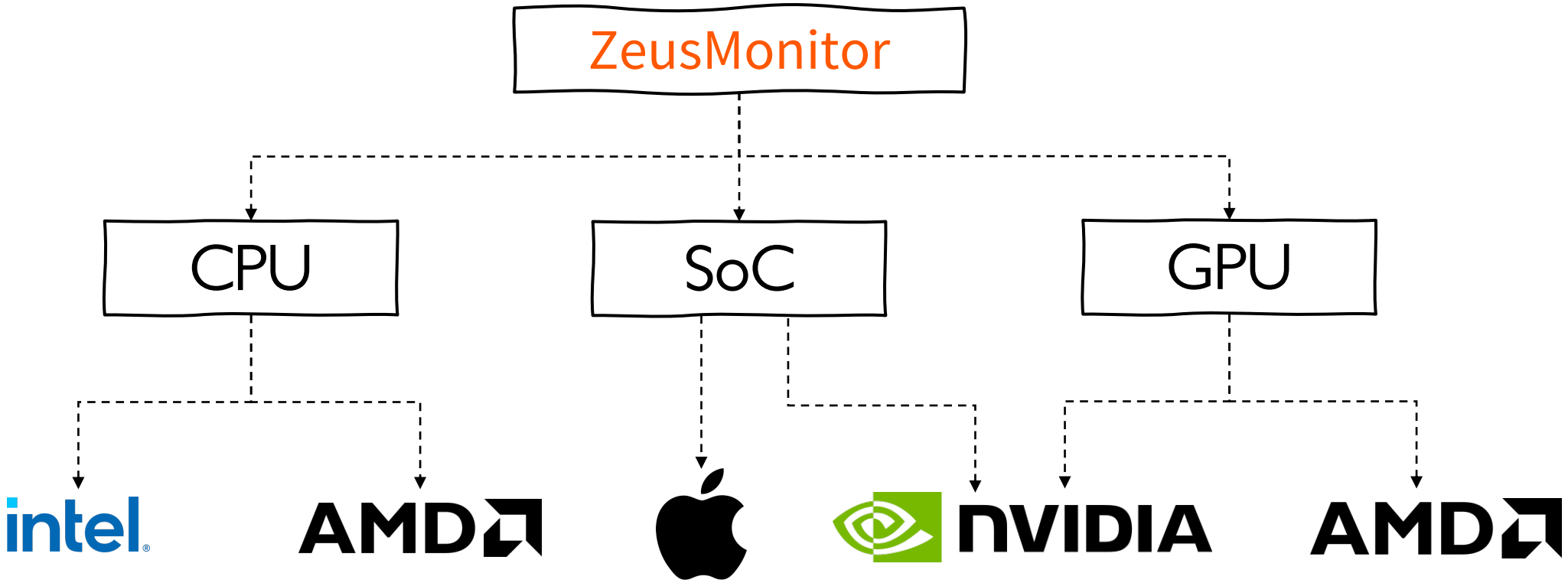
Measurement Pitfalls

- NVML counter fidelity
  - On modern GPUs, NVML counters update once every 100 ms
  - Similar or shorter duration computation should be repeated



Tools for Measurement

- Abstraction layer over devices & sync



Hands-On Session

https://tinyurl.com/measure-energy