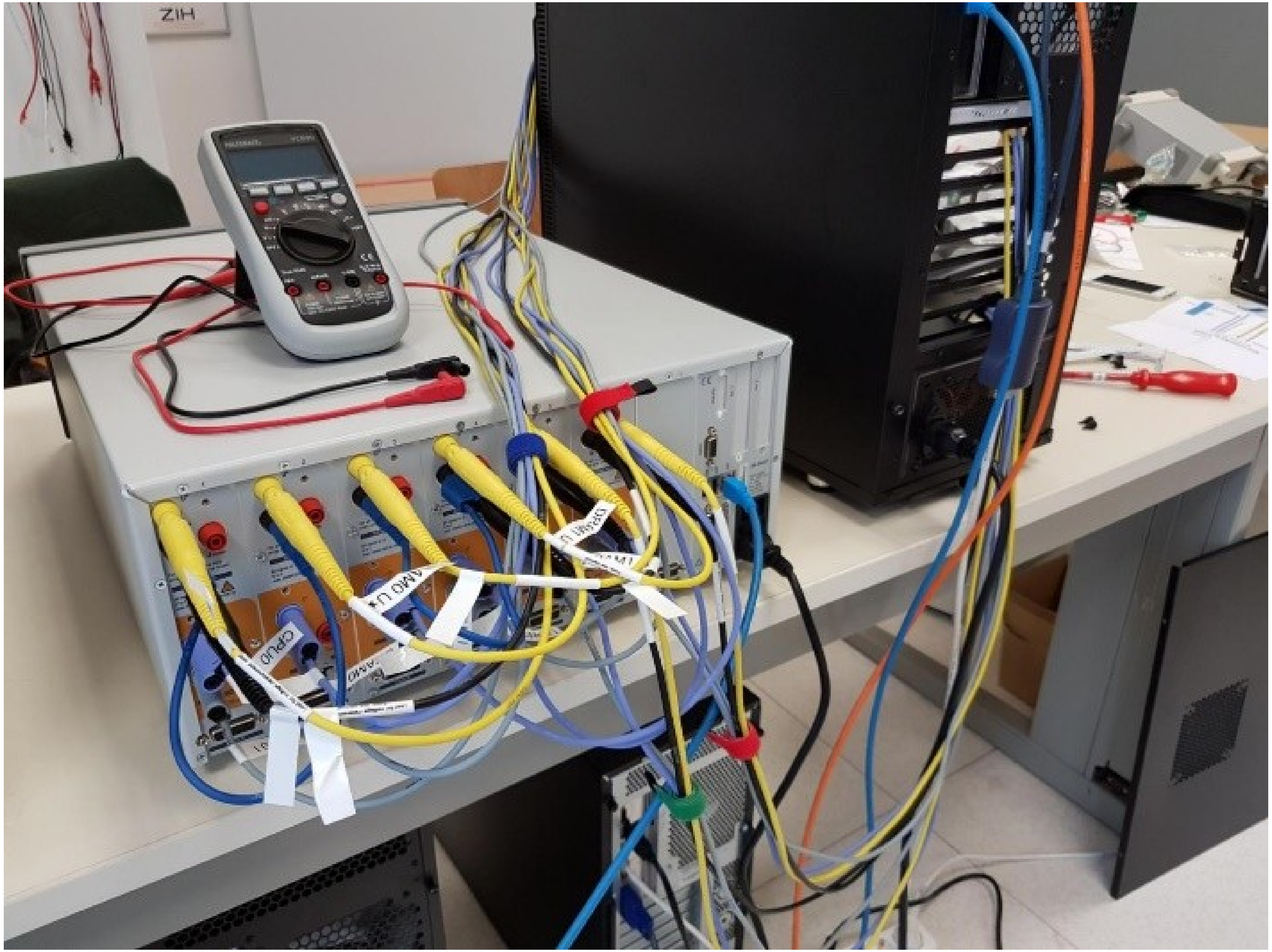
Hannes Tröpgen, Mario Bielert, Thomas Ilsche

Evaluating the Energy Measurements of the IBM POWER9 On-Chip Controller

ICPE '23

Coimbra, 17 April 2023

ZIH
Center for Information Services
and High Performance Computing


ZIH
CPU0

PowerNV Platform @ ZIH

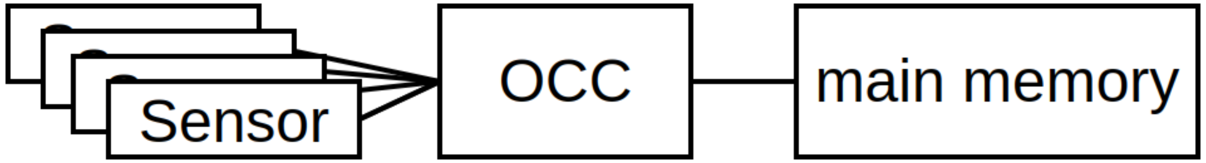
- PowerNV: POWER{8,9,10}
- 32 nodes at our HPC cluster *taurus*
  - AC922 (*Newell*, form. *Witherspoon*) by IBM
- Processor: two POWER9 CPUs per node (Codename *Monza*, 02CY209)
  - 22 cores/88 threads each
- 6 NVIDIA VOLTA V100 GPUs per node
  - 150 GB/s Host GPUs bandwidth via NVLink

On-Chip Controller (OCC)

- Embedded PowerPC 405 processor on PowerNV-CPUs
- Open-source firmware: https://github.com/open-power/occ/
- Objectives:
  - "Keep the system thermally safe"
  - "Keep the system power safe"
  - "Provide [...] sensor data"

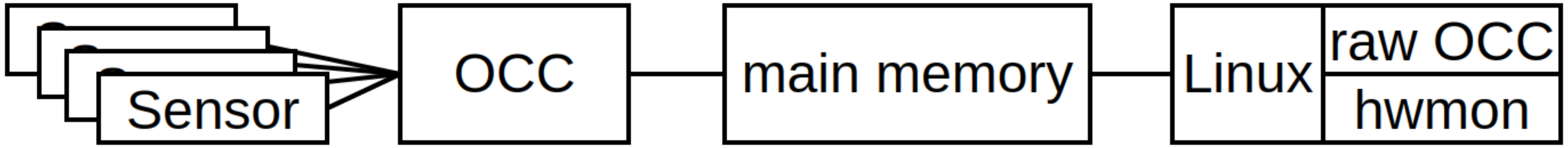
Here: Describe capabilities and limitations of OCC-provided power measurements for energy efficiency analysis.

Available Data



- Per OCC (i.e., per CPU):
  - GPUs, memory, processor itself
  - Processor sub-powers: Vdd (cores), Vdn (nest)
- Once per system:
  - Bulk power
  - 16 *Analog Power Subsystem Sweep* (APSS) channels

Available Interfaces



- hwmon: subsystem for hardware monitoring in Linux kernel
- OCC raw: exposed raw blob (150 KiB per OCC)

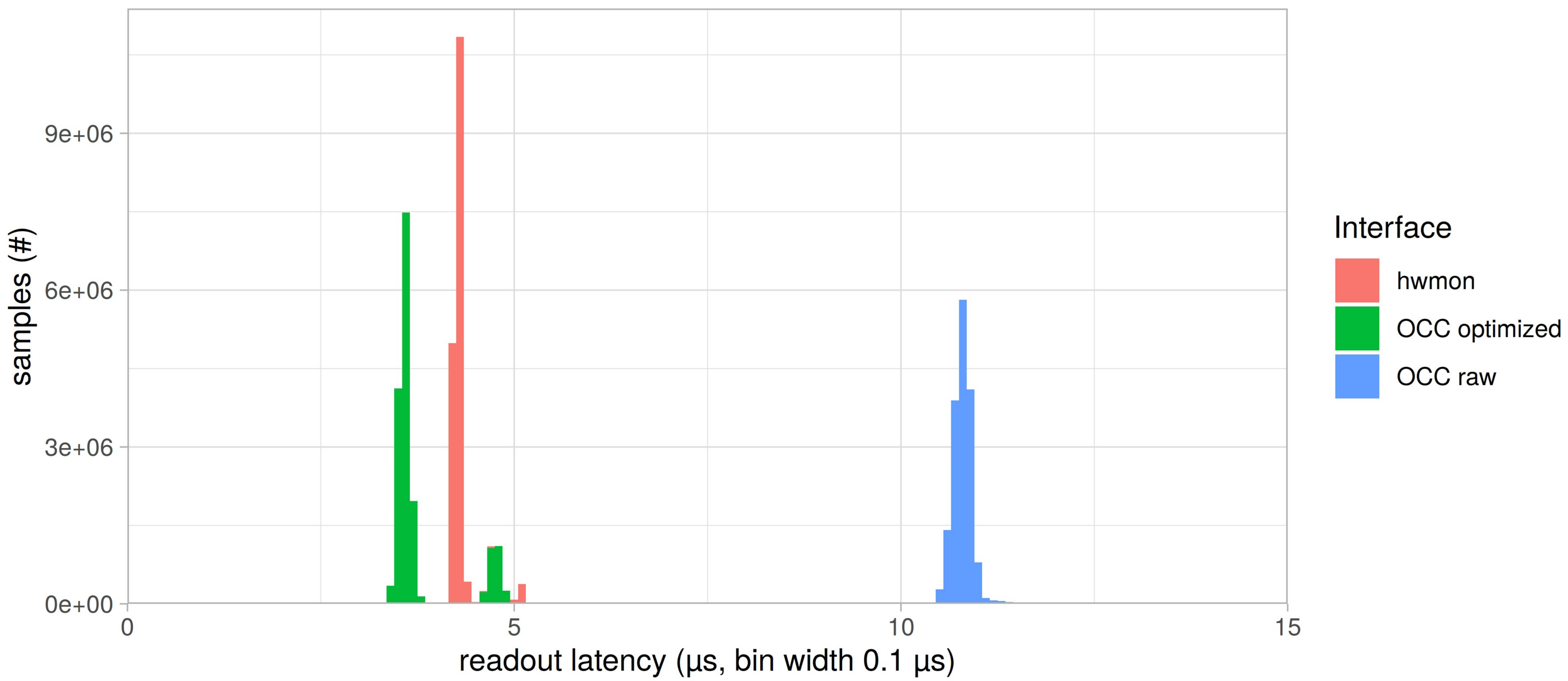
	hwmon	OCC raw
resolution of values	1 W	1 W
current sample	x	x
accumulator		x
timestamp		x

Interface Experiments

- How long does one readout take?
- How often are new values provided?

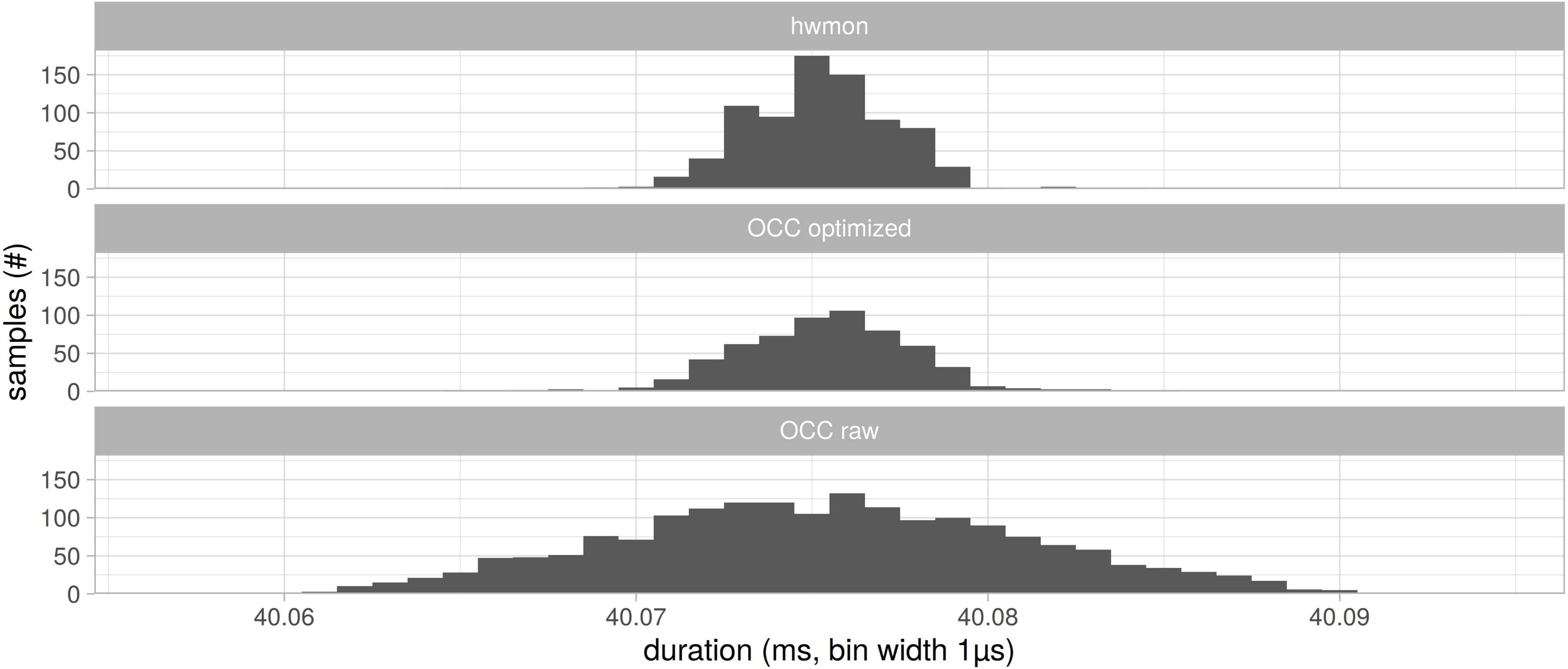
- Read sample & current time, collect 2^{24} samples total
  - hwmon: read sysfs file, parse string
  - OCC raw: read OCC blob, parse header, read data
  - OCC optimized: read OCC blob, parse header *only once*, read data

Overhead



samples (#)
9e+06
6e+06
3e+06
0e+00
0
5
10
15
readout latency (µs, bin width 0.1 µs)
Interface
hwmon
OCC optimized
OCC raw

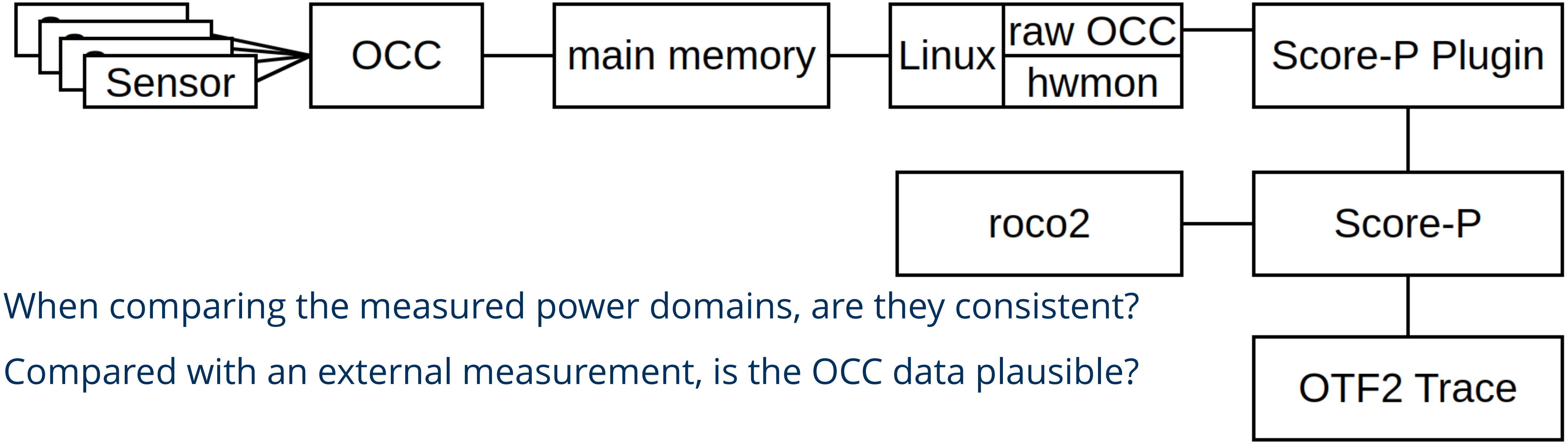
Time Between Updates



Interface Experiments – Results

- How long does one readout take?
  - 4.3 µs (hwmon), 10.8 µs (OCC raw), 3.8 µs (OCC optimized)
- How often are new values provided?
  - Every 40.08 ms, approx. 24.95 Sa/s

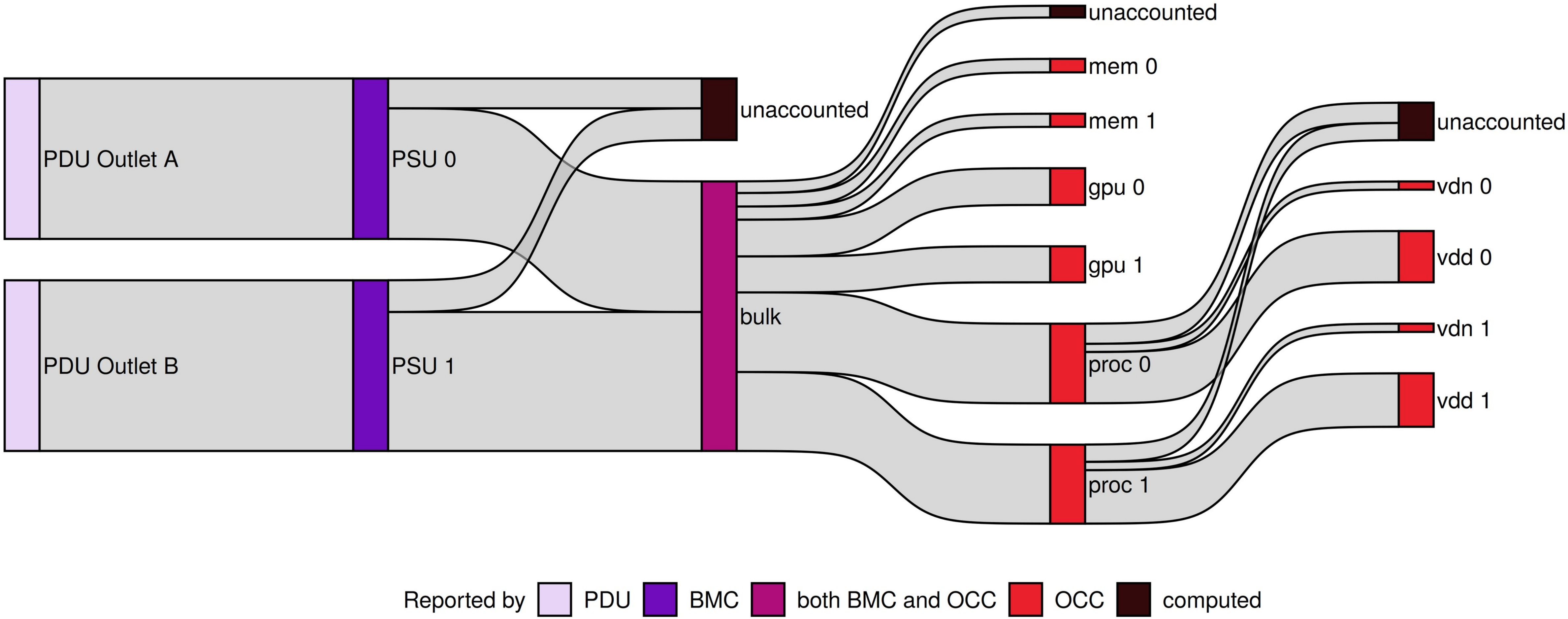
Measurement Across Power Domains



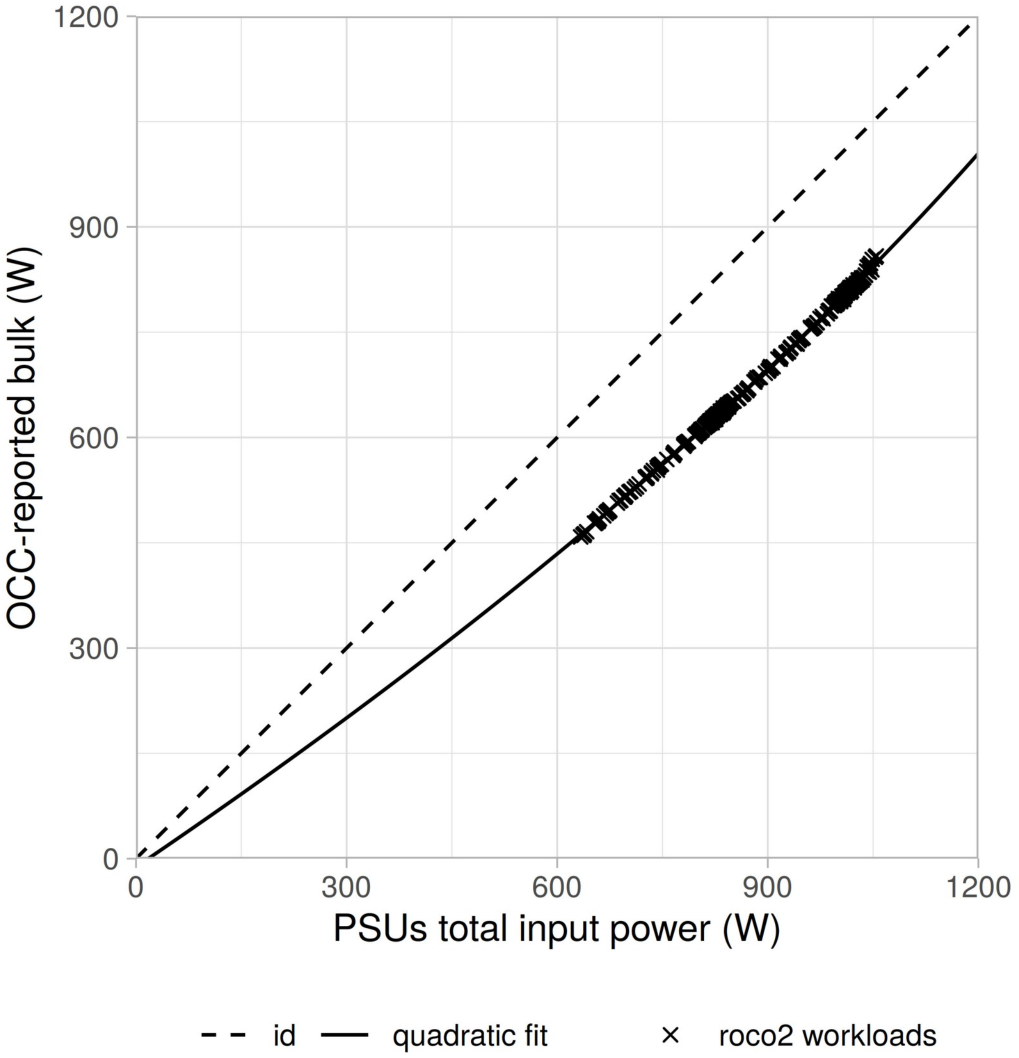
- When comparing the measured power domains, are they consistent?
- Compared with an external measurement, is the OCC data plausible?

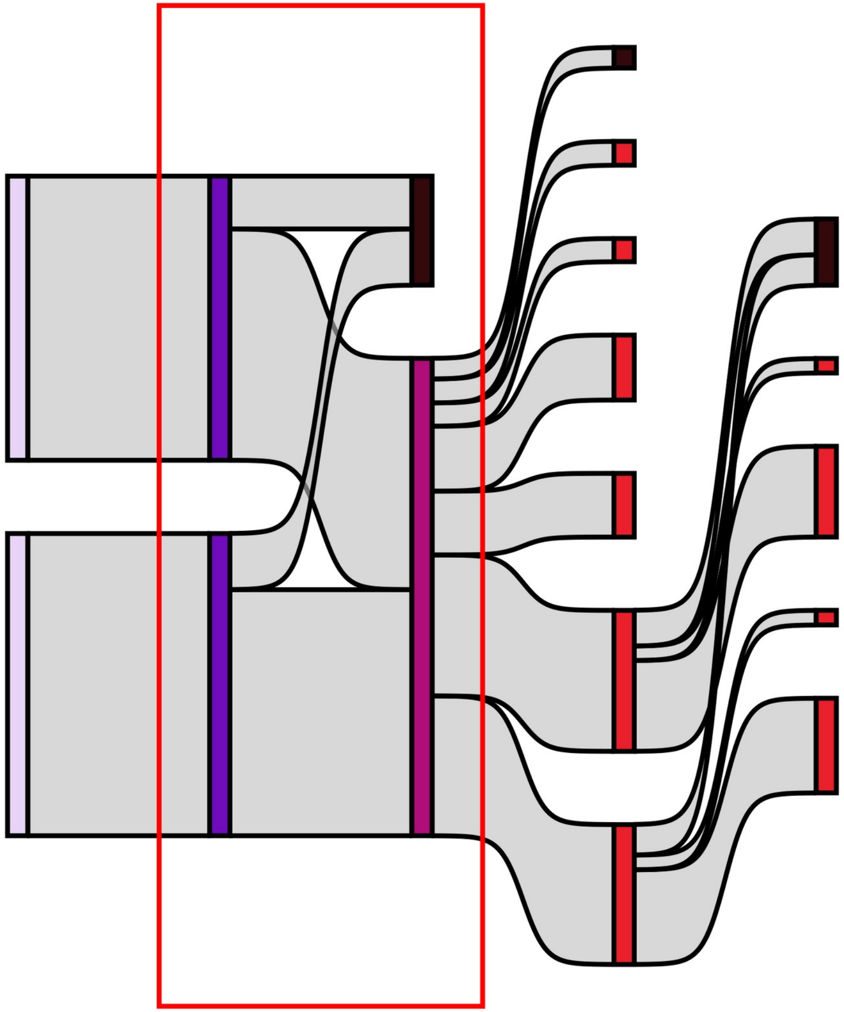
- Generate different workloads & power levels with roco2
- Use ZIH co-developed toolchain: Score-P, OTF2 trace format
- We developed the IBM PowerNV Score-P Plugin:
- https://github.com/score-p/scorep_plugin_ibmpowernv

Power Delivery Scheme

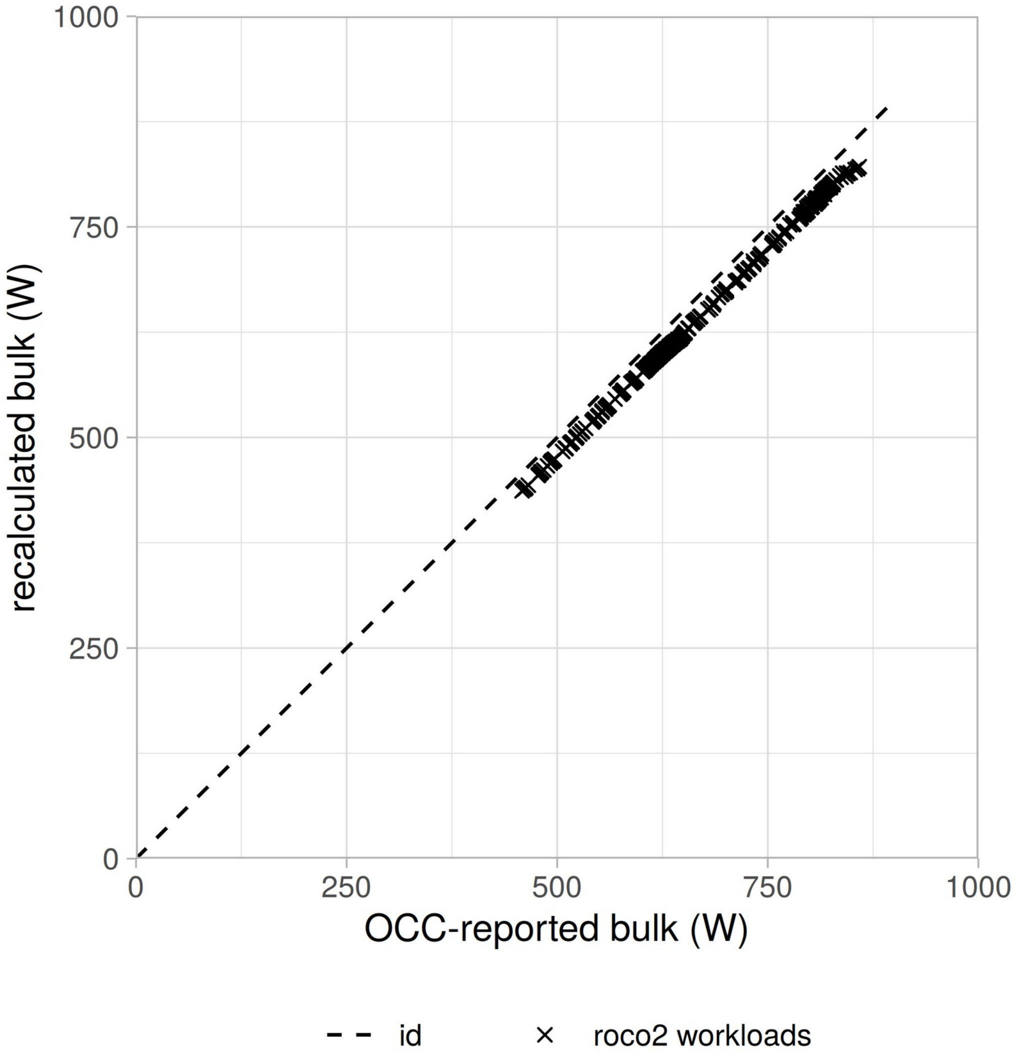


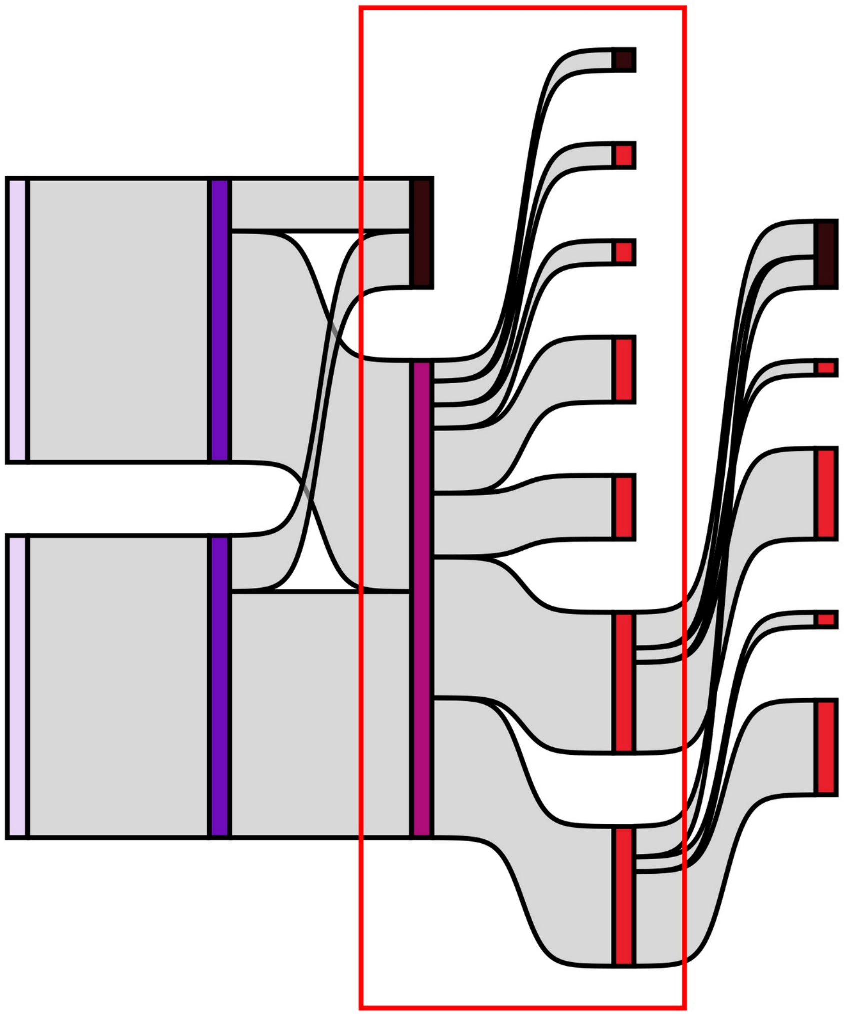
PSU Inputs vs OCC bulk





OCC Bulk Recalculation

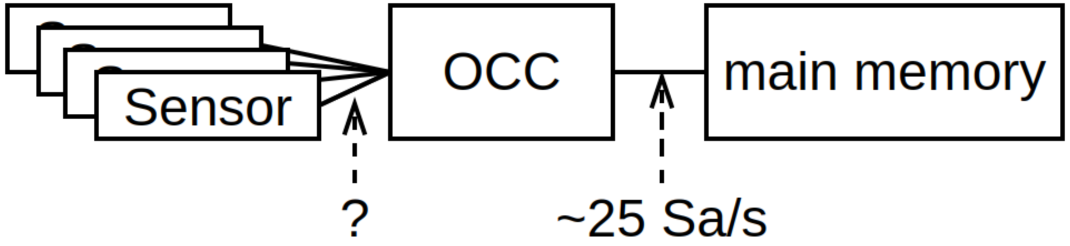




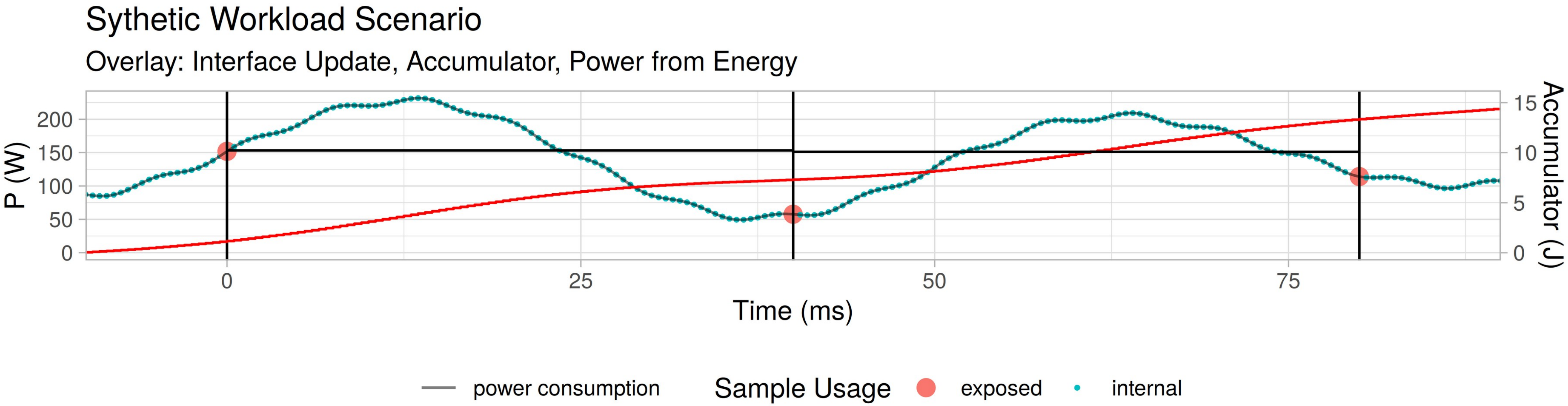
Power Domains Experiments – Results

- When comparing the measured power domains, are they consistent?
  - Yes, but discrepancy visible (3.8 % MAPE, 25.5 W MAE)
- Compared with an external measurement, is the OCC data plausible?
  - Yes, no workload bias visible
  - Plausible mean efficiency of 77%

Internal Update Rate



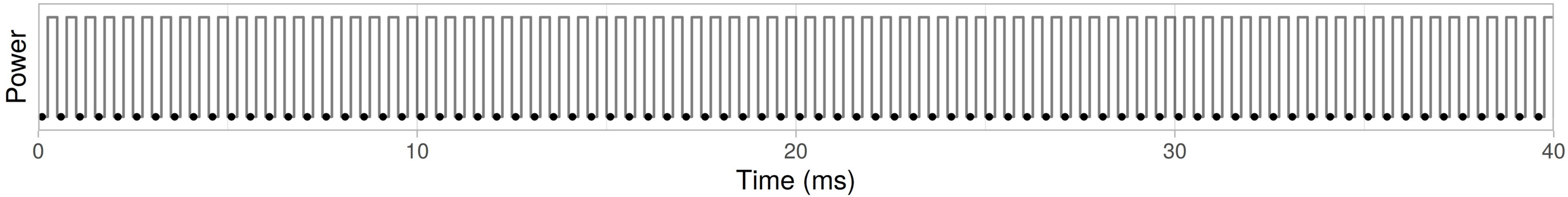
- “sample time” in documentation is e.g. 500 µs → 2 kSa/s



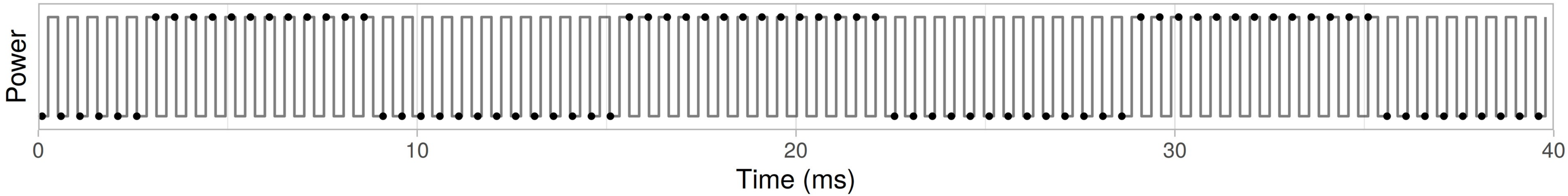
Internal Update Rate

- What is the internal update rate?
  - Idea: Use Aliasing
  - Alisaing produces artifacts in data → Workload frequency matches sampling rate

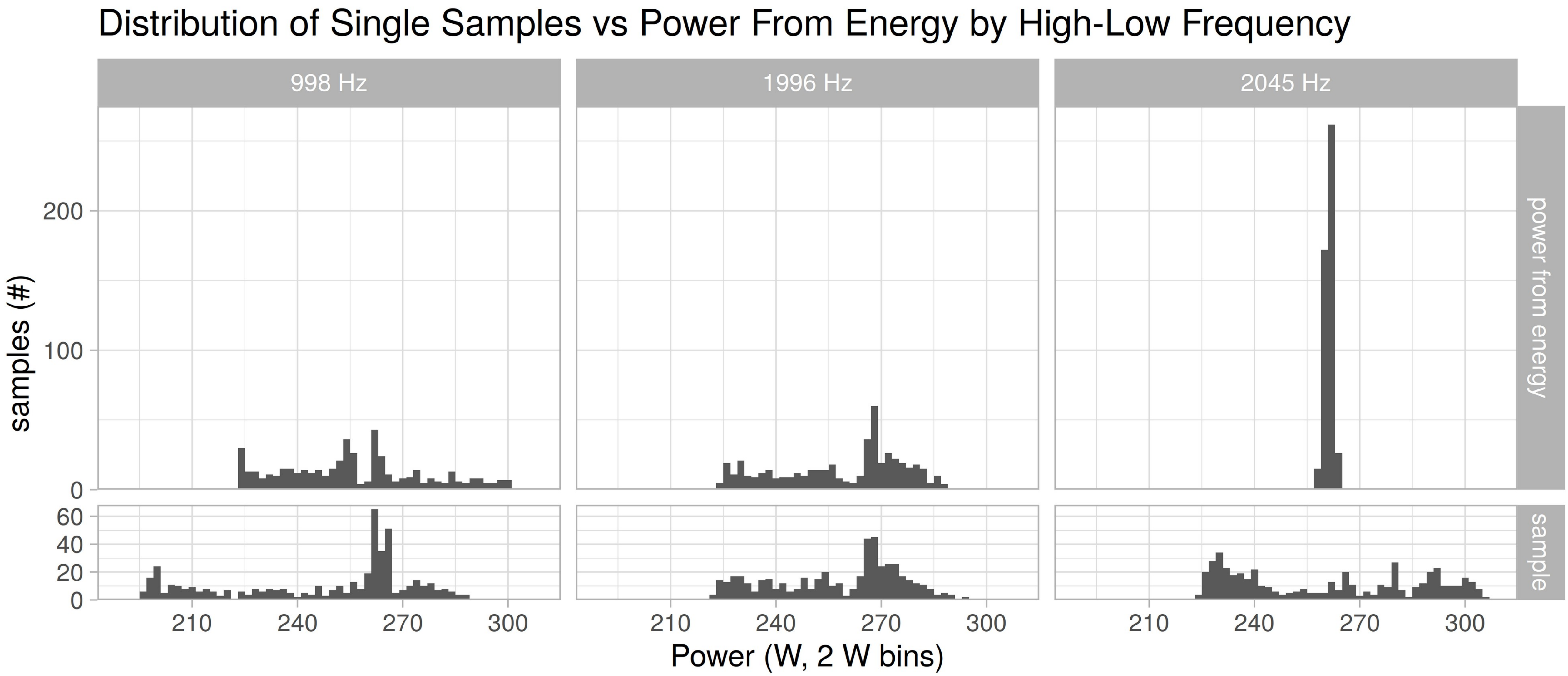
Workload Frequency and Internal Sampling Rate Match Exactly



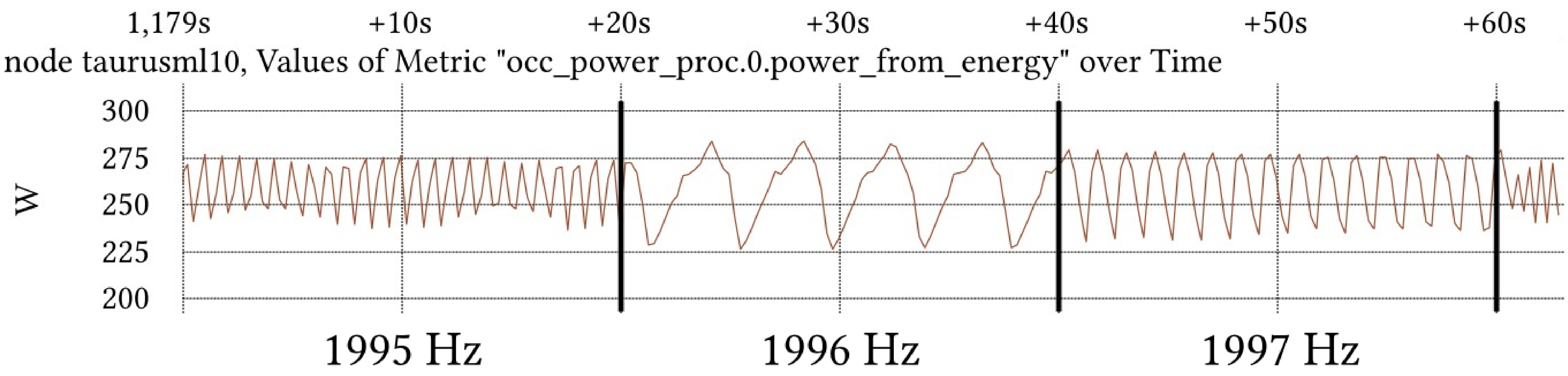
Workload Frequency and Internal Sampling Rate do NOT Match Exactly



Internal Update Rate – Observations



Internal Update Rate – Results



$f_{workload}$ (Hz)	$f_{pattern}$ (Hz)	$f_{sampling}$ (Hz)
1995	1.24	1996.24
1996	0.24	1996.24
1997	0.77	1996.23

- Min: 225 W
- Max: 285 W
- → Assume 255 W mean
- → Measurement deviates up to 12% from that

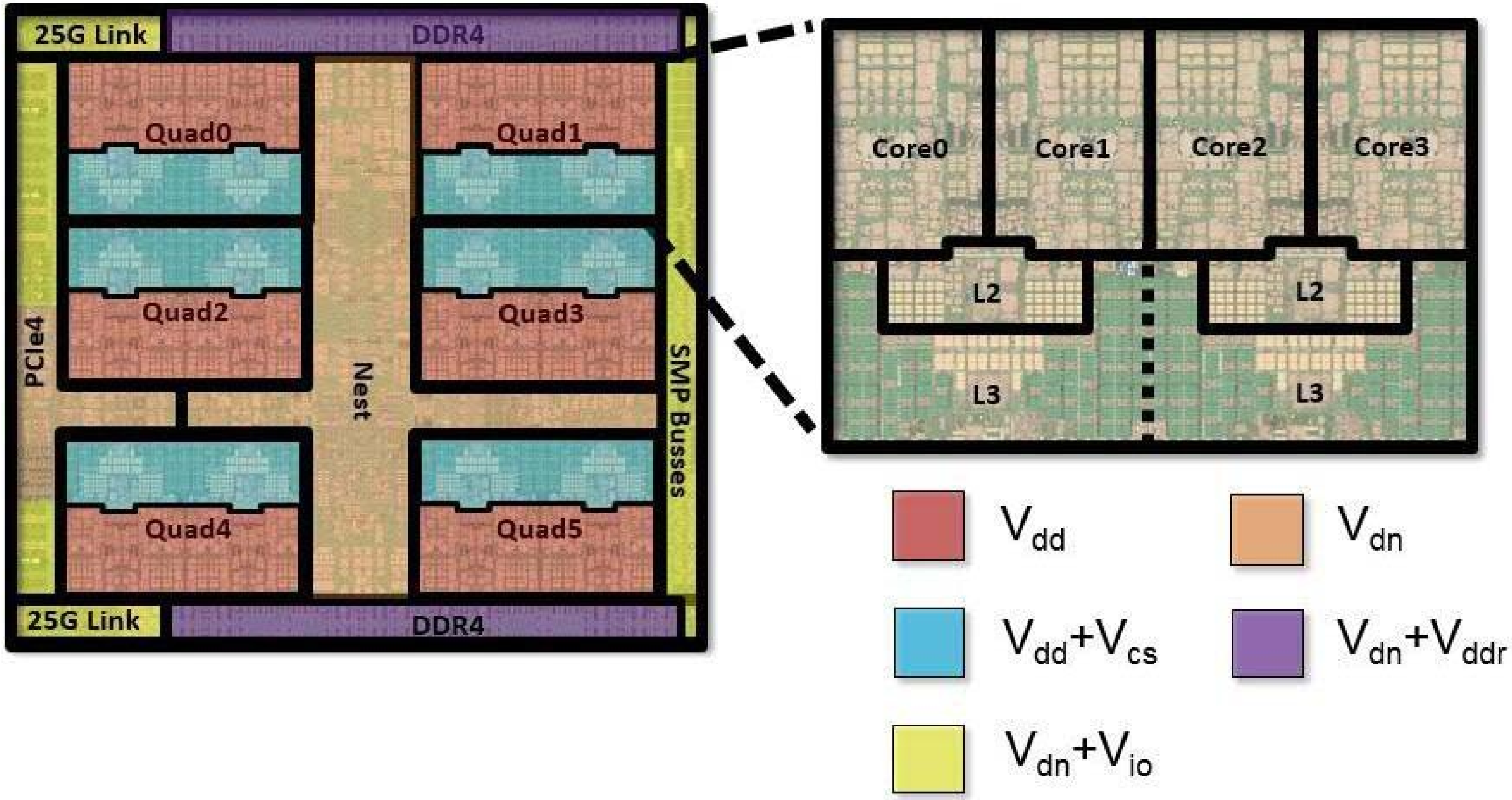
Summary & Future Work

- OCC provides power measurements for CPU (incl. sub-powers), GPUs, memories, system bulk
- Provided measurements are consistent with PSUs and with themselves
  - Calibrated reference measurement lacking
  - Only tested for CPU workloads
- Two interfaces: hwmon, raw OCC
  - Overhead: 3.8 µs to 10.8 µs; (external) update rate: 24.95 Sa/s
- Power from energy derived from accumulator uses more samples
  - 1996 Sa/s experimentally verified
  - Experimental worst-case deviation: 12% for one processor

Thank you for your attention!

Backup-Slides

Processor Power Sub-Domains



- Not shown: $V_{REF}/V_{SB}/V_{DPLL}\ /V_{AVDD}/V_{I2C}$

Fig. 1 from C. Gonzalez et al., "The 24-Core POWER9 Processor With Adaptive Clocking, 25-Gb/s Accelerator Links, and 16-Gb/s PCIe Gen4," in IEEE Journal of Solid-State Circuits, vol. 53, no. 1, pp. 91-101, Jan. 2018, doi: 10.1109/JSSC.2017.2748623.