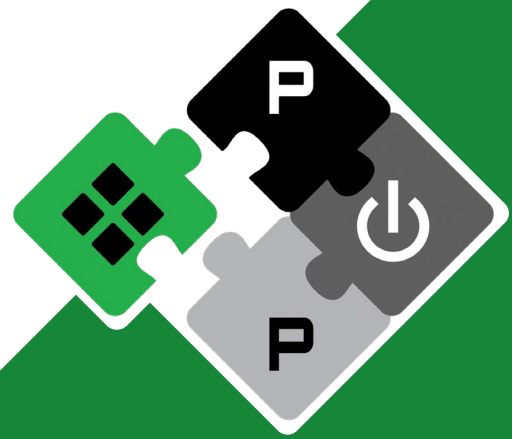
ETH zürich

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

A Data-Driven Approach to Lightweight DVFS-Aware Counter-Based Power Modeling for Heterogeneous Platforms

Sergio Mazzola smazzola@iis.ee.ethz.ch
Thomas Benz
Björn Forsberg
Luca Benini



@pulp_platform

pulp-platform.org

youtube.com/pulp_platform

PULP Platform
Open Source Hardware, the way it should be!

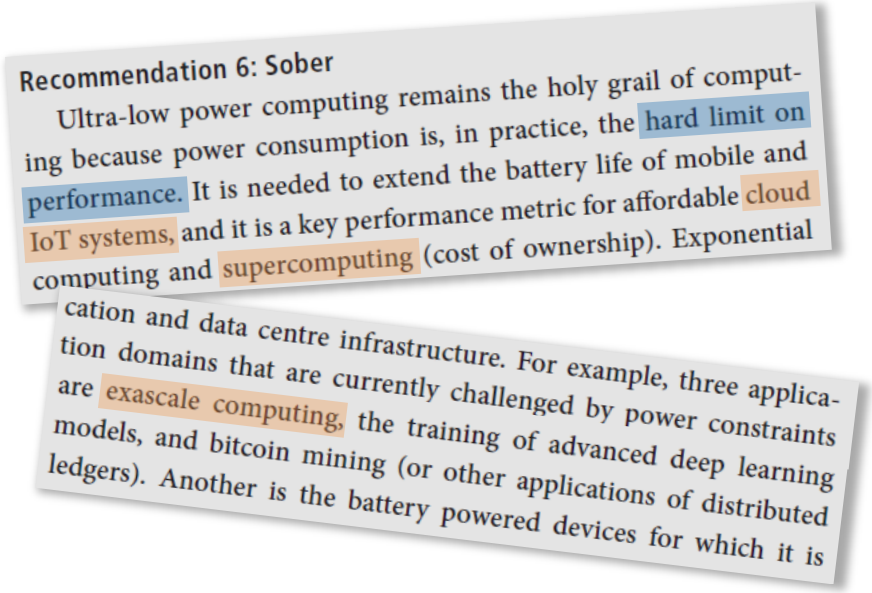
Power-limited computing

- Dennard scaling is over (no breaking news…)
- Buying **energy efficiency** with **area**
  - Multi-cores, GPUs
  - Application-specific hardware accelerators
- Novel **power management** paradigms
  - Robust online DPM control
  - System complexity
  - Parallelism, heterogeneity

Power modeling framework for
parallel and **heterogeneous** systems

HiPEAC

Duranton, Marc, et al. "**HiPEAC Vision 2021**: high performance embedded architecture and compilation." (2021).

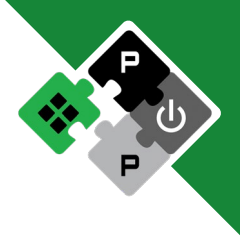


Recommendation 6: Sober

Ultra-low power computing remains the holy grail of computing because power consumption is, in practice, the hard limit on performance. It is needed to extend the battery life of mobile and IoT systems, and it is a key performance metric for affordable cloud computing and supercomputing (cost of ownership). Exponential

cation and data centre infrastructure. For example, three application domains that are currently challenged by power constraints are exascale computing, the training of advanced deep learning models, and bitcoin mining (or other applications of distributed ledgers). Another is the battery powered devices for which it is

Why is it challenging?



- Robust online DPM control
- System complexity
- Parallelism, heterogeneity

Solving ‘em all

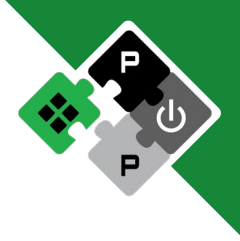
- Robust online DPM control
  - **Fast** + **accurate** power measures
- No analog power sensors
- Hardware **performance counters** (PMCs)

- System complexity
  - Huge **number of model parameters**
  - System **size**, systems **fragmentation**
  - Dynamic V/f scaling (**DVFS**)
- **Data-driven** selection for **optimal** counters
- **Automatic** procedure, **generic** target platform
- One **linear power model** per DVFS state

- Parallelism, heterogeneity
  - Many different sub-systems
- One set of **linear power models** per sub-system
- Model sub-system individually

A Data-Driven Approach to Lightweight DVFS-Aware Counter-Based Power Modeling for Heterogeneous Platforms

Our contributions in a nutshell

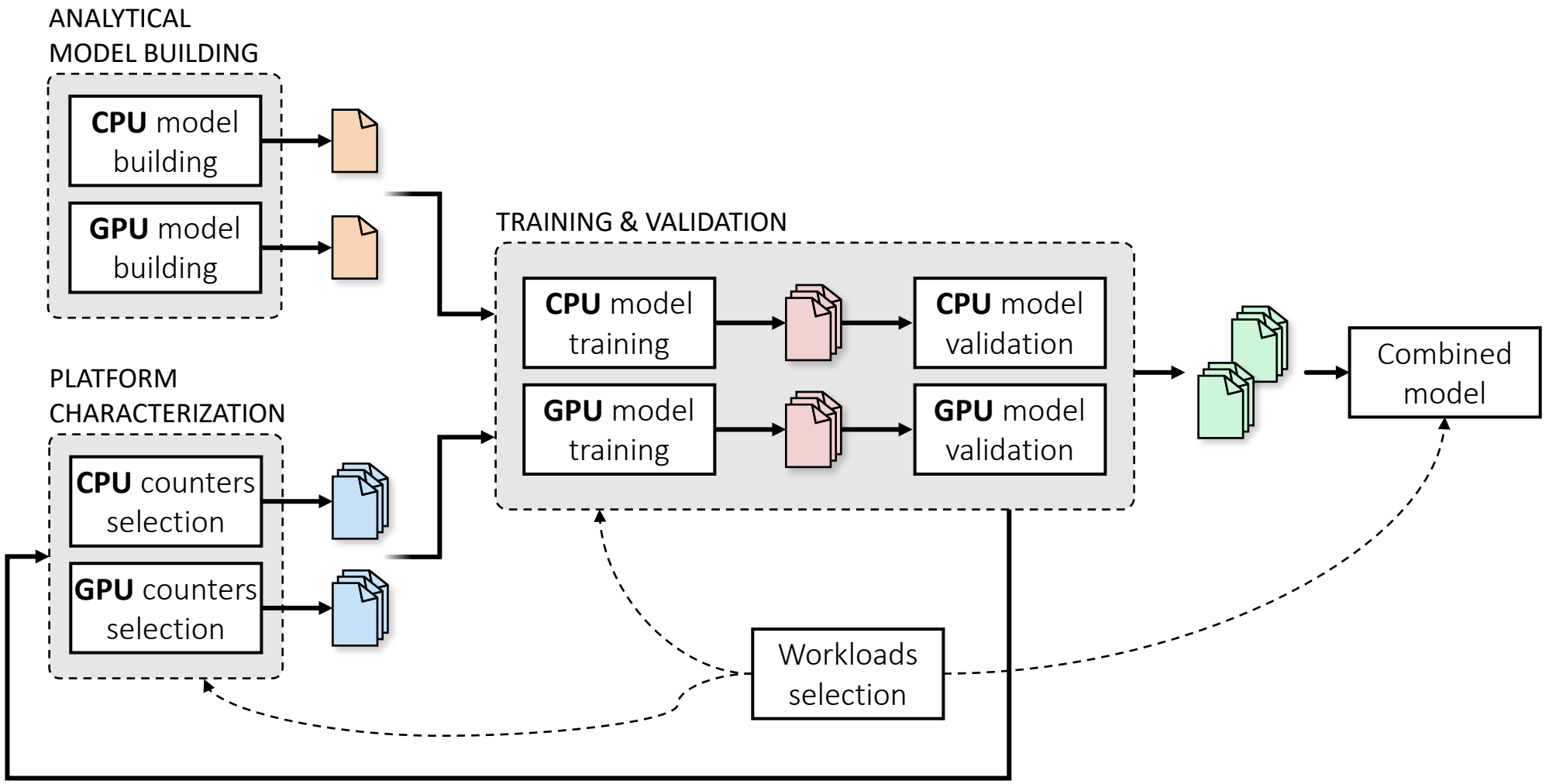


- Statistical **counters selection** for power models (CPU+GPU)
- **Look-up-table** approach to power modeling
  - Addressing **DVFS** + **heterogeneity**
- Validation on NVIDIA Jetson AGX Xavier

Unprecedented combination of:

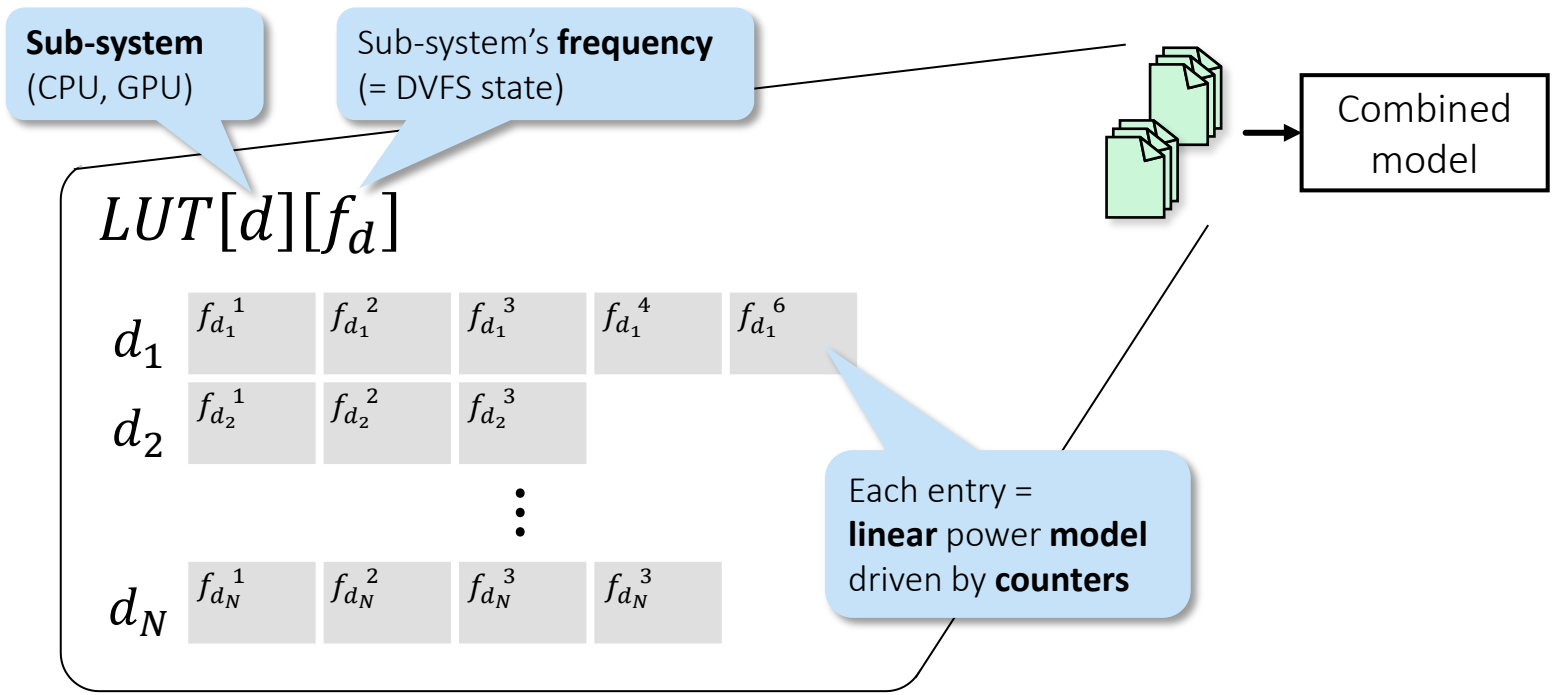
- general applicability, automation
- **low-overhead** model
- state-of-the-art **accuracy** for modern systems

Our holistic methodology

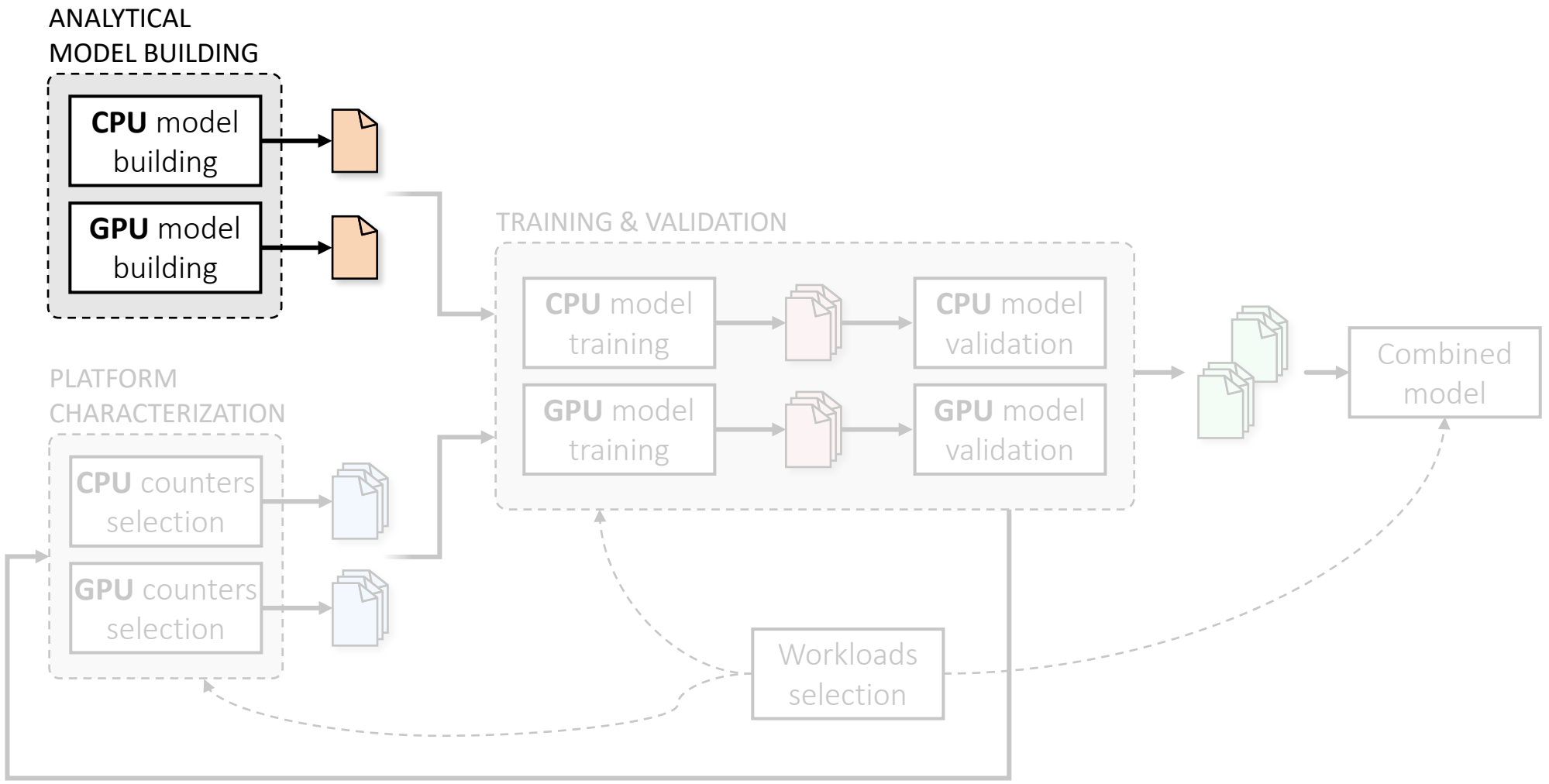


Our holistic methodology

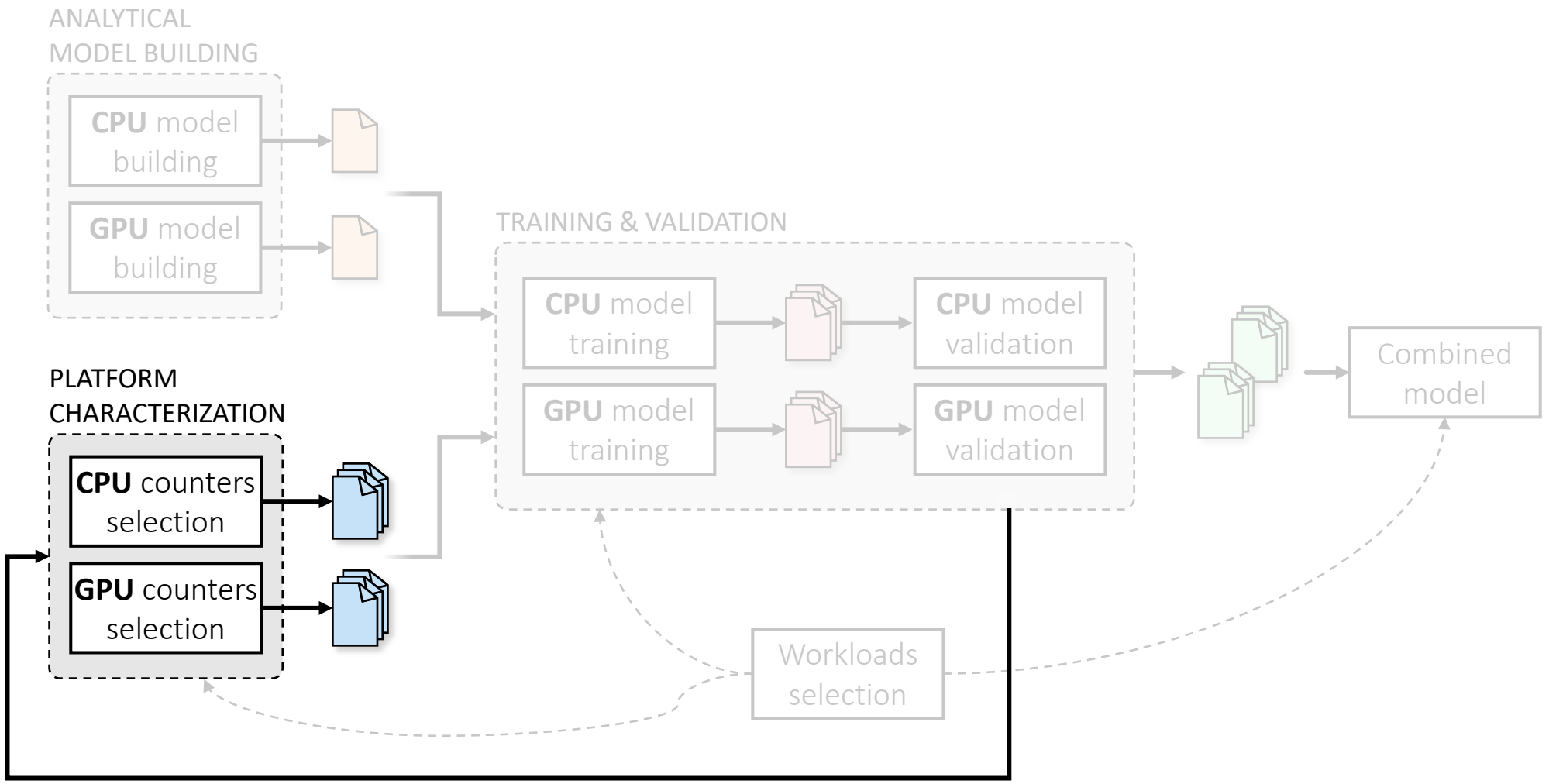
- **LUT-based** system-level power model
- Each sub-system's model built **individually**



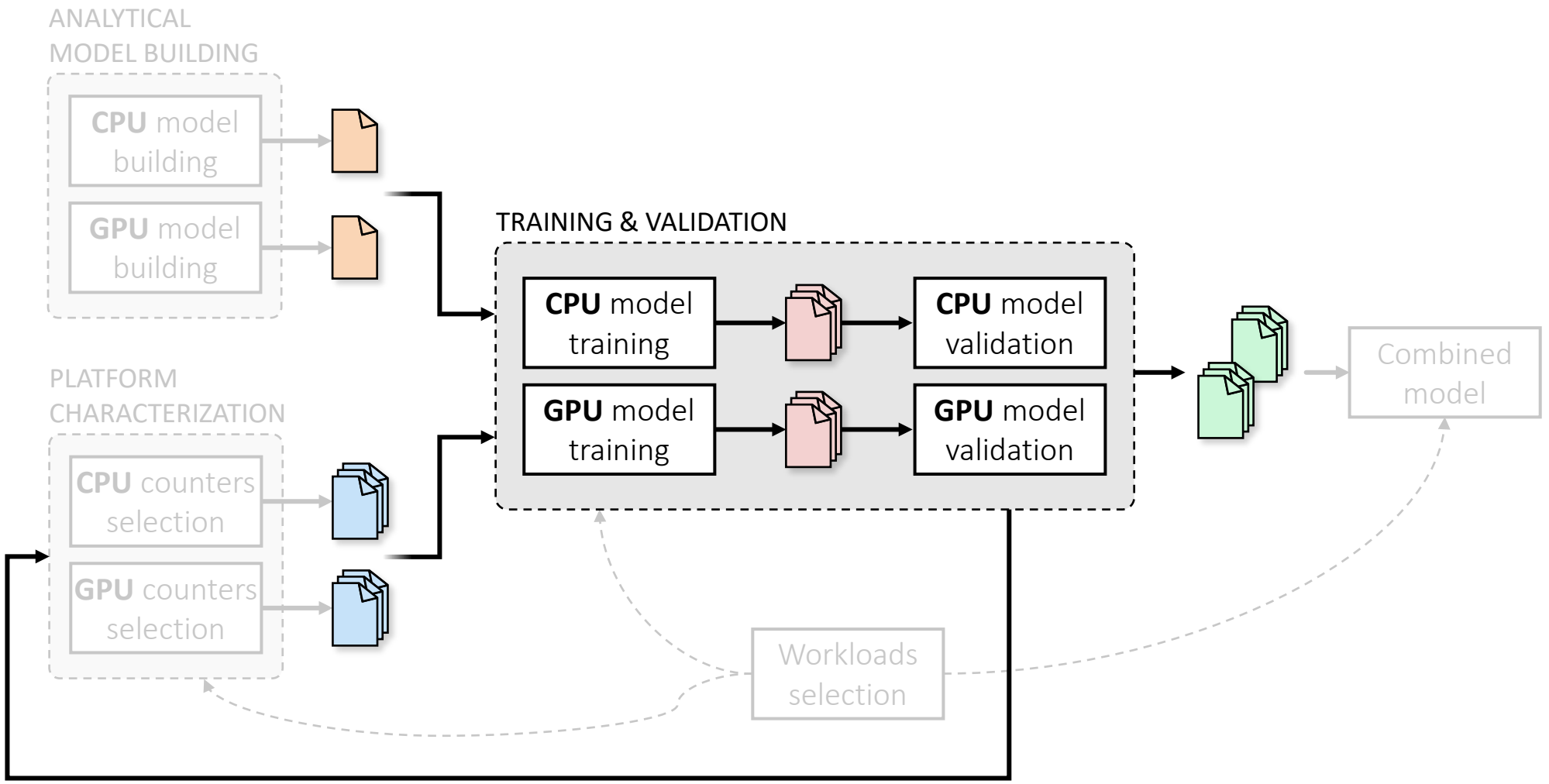
Analytical power model building



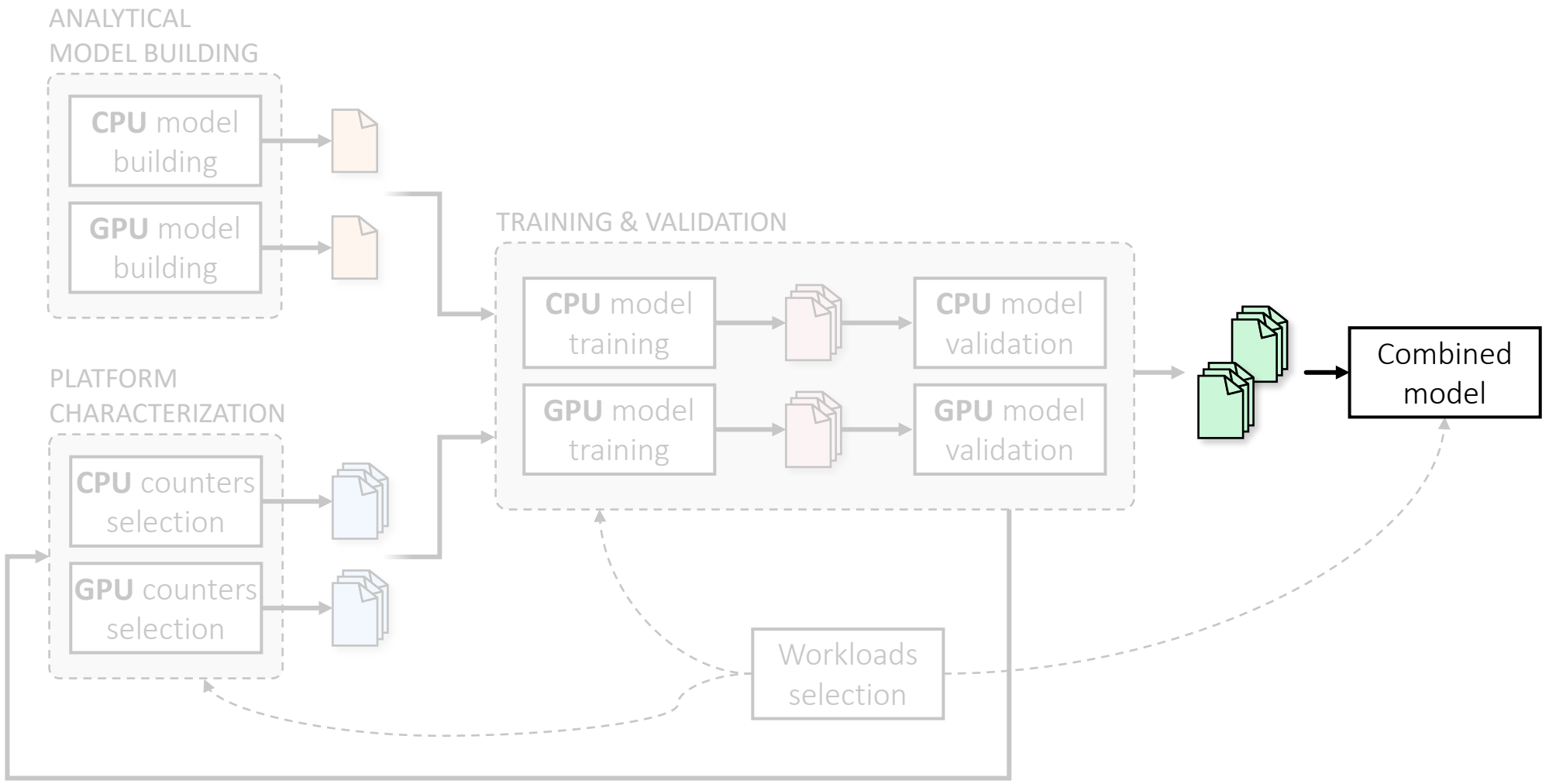
Platform characterization



Training and validation

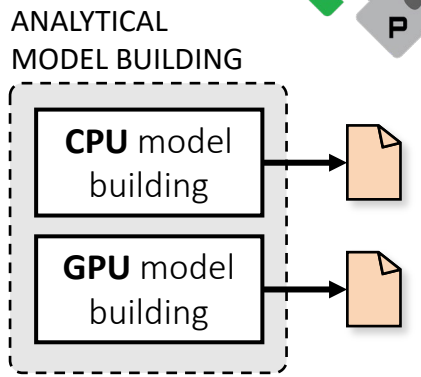


Training and validation



Analytical power model building

- Based on CMOS static/dynamic power consumption
- **Linear** models, thanks to LUT
  - Capture **non-linear DVFS** behaviors
  - Low overhead, high accuracy
- **CPU model**
  - Per-core model, per-cluster DVFS
  - Supports power gating (requires *cycle counter*)
  - Dynamic power (*activity*) modeled by counters
- **GPU model**
  - Static + dynamic component



static | per-core leakage | per-core activity

$$P_{CPU} = L + \sum_{i=1}^{\#cores} \left(g_i \cdot G_i + \sum_{j=i}^{\#PMCs} x_{ij} \cdot A_{ij} \right)$$

$$P_{GPU} = K + \sum_{j=i}^{\#PMCs} x_j \cdot B_j$$

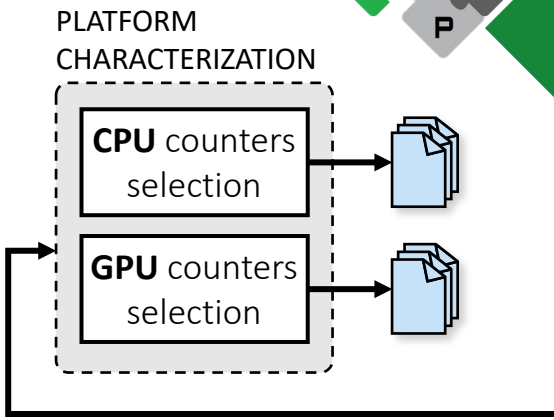
Platform characterization

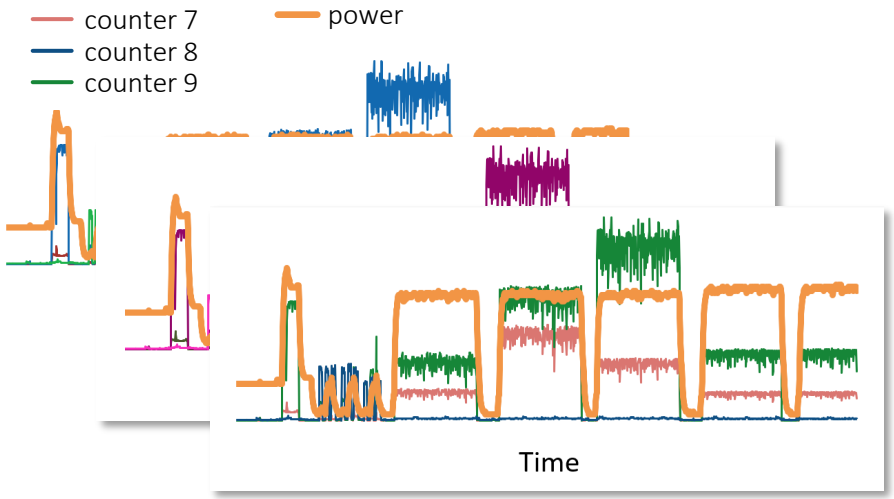
- Statistical data-driven method
  - **One time** per platform sub-system
  - Reproducible on any given platform
- Procedure
  1. profile all counters + power
  2. compute PCC of all counters
  3. select best counters

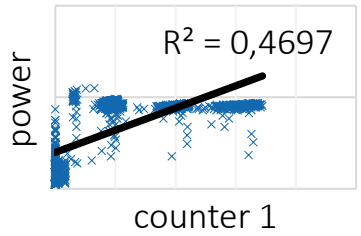
- **No manual** intervention
- As **architecture-agnostic** as possible

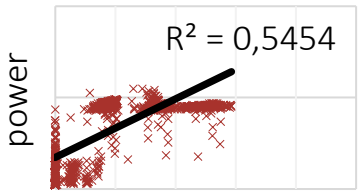
Per **sub-system**, per **frequency**

- Aware of **PMU constraints**
- **Accuracy/overhead** trade-off



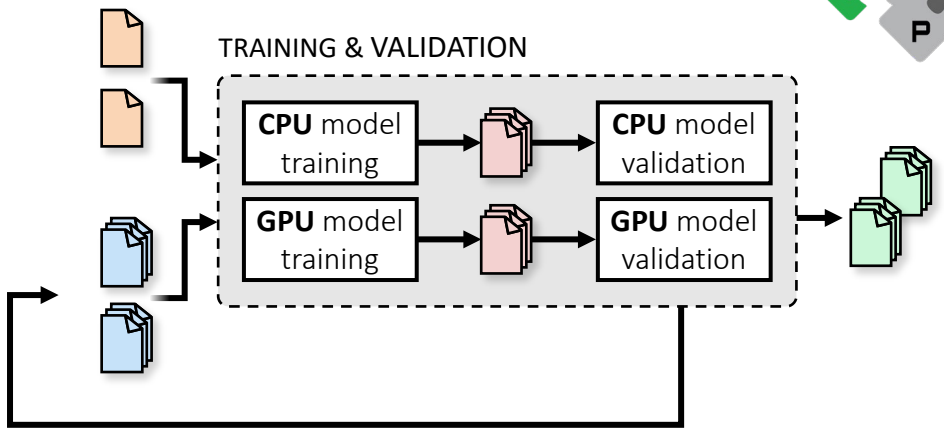






Training and validation

- Training: **NNLS**
  - Non-negative weights physically meaningful
  - Robust to **multi-collinearity** and overfitting
- Output
  - **Complete LUT** (power models for all sub-systems @ all frequencies)

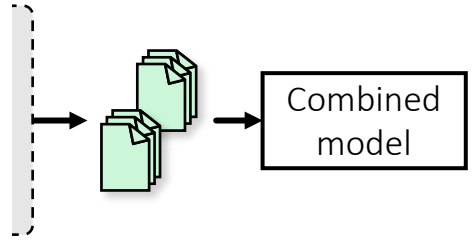


$$LUT[d][f_d]$$

CPU	f_{CPU}^{1}	f_{CPU}^{2}	f_{CPU}^{3}	…	f_{CPU}^{N}
GPU	f_{GPU}^{1}	f_{GPU}^{2}	…	f_{GPU}^{M}	

Combined system-level model

- Sub-systems power estimates combination
  - Fix a frequency for each sub-system
  - Sum power estimates from all sub-systems' models
- Individual sub-systems models up until here
  - Reduced model complexity and overhead
  - More general applicability
  - Limitation discussed later…



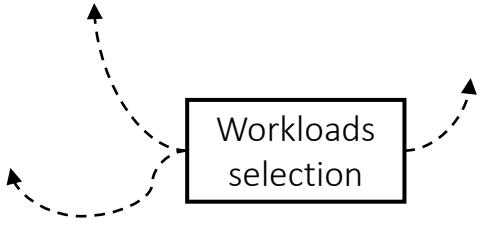
$$LUT[d][f_d]$$

$$CPU \quad f_{CPU}^{x}$$

$$+$$

$$GPU \quad f_{GPU}^{y}$$

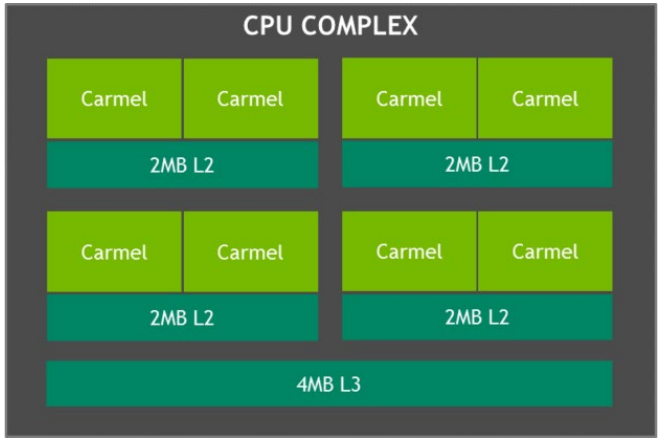
Workloads selection

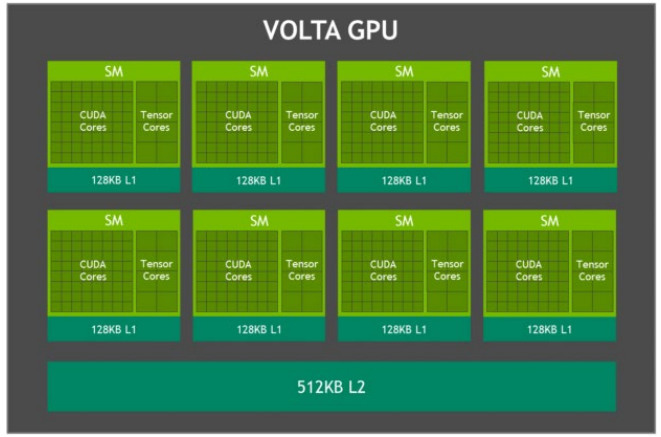


- Employed for
  - Platform characterization
  - Sub-system models training and validation
  - Combined model validation
- How to select
  - Coverage of targeted **sub-systems**: heterogeneity
  - Variety of **behaviors**: workload- and platform- independent
- Employed benchmarks
  - Rodinia
  - Synthetic benchmarks for CPU

Experimental setup – **NVIDIA Jetson** AGX Xavier

- 8-core 64-bit **ARM SoC**
  - Per-cluster DVFS
  - 29 nominal frequencies (115 MHz – 2.3 GHz)
  - Employed: 730 MHz, 1.2 GHz, 2.3 GHz
  - PMU: *3 events + 1 fixed for clock cycles*
- 512-core NVIDIA **Volta GPU**
  - 14 nominal frequencies (115 MHz – 1.4 GHz)
  - PMU compatibility constraints
- 2 on-board **power monitors** (INA3221)
  - Analog current sensors
  - Useful to build better models

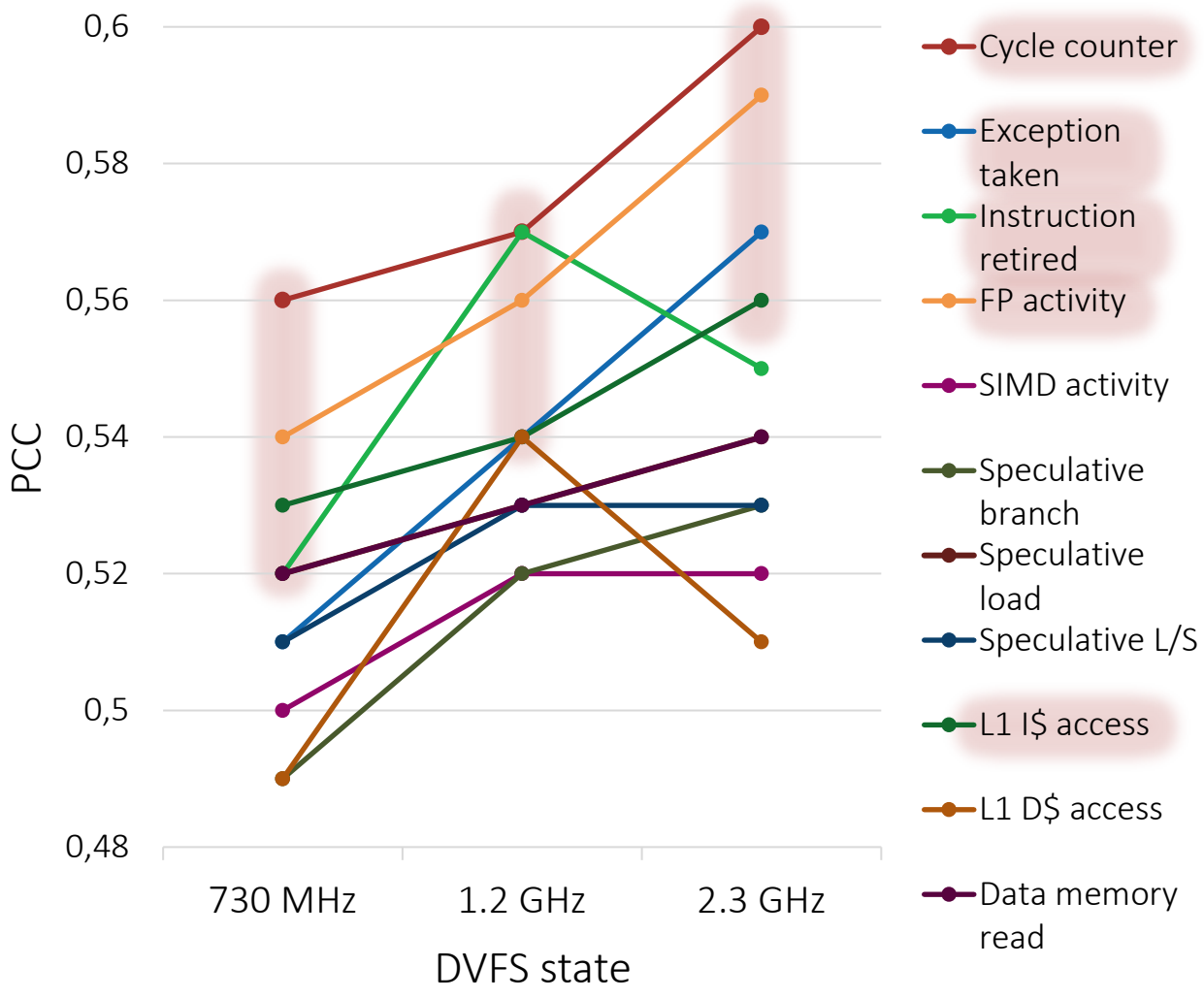




Source: https://developer.nvidia.com/blog

Results of platform characterization – **CPU**

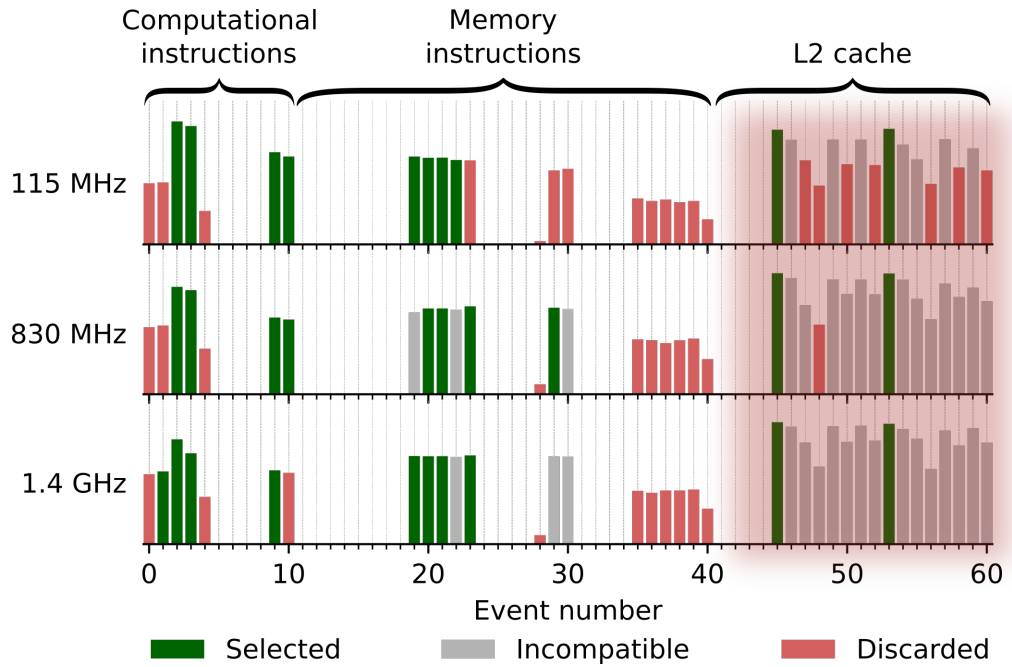
- PMU no compatibility constraints
- Best **3** PMCs @ each frequency
- **+1** fixed clock cycle counter
- **Per-cluster DVFS**: same PMCs for each core

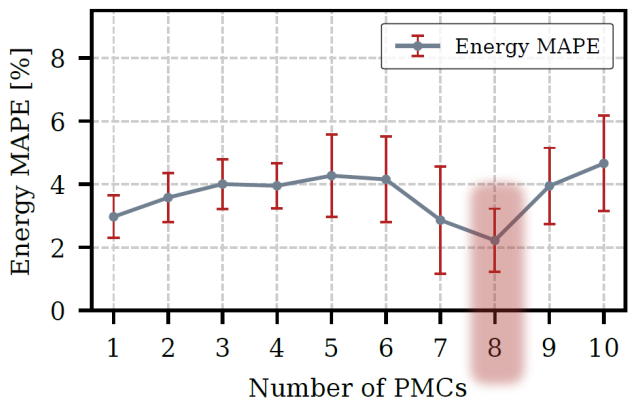


Results of platform characterization – GPU

- PMCs **compatibility** constraints
  - Some events not countable simultaneously
  - No fixed max number
- **PMU-aware** algorithm
  - Select best counters considering PMU constraints
  - **Heterogeneous** choice of counters
- Overhead/accuracy trade-off
  - Number of PMCs

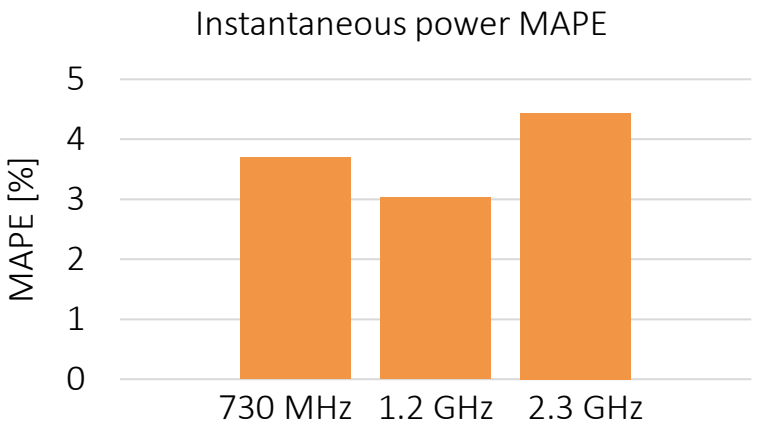
Energy MAPE (mean abs percentage error)

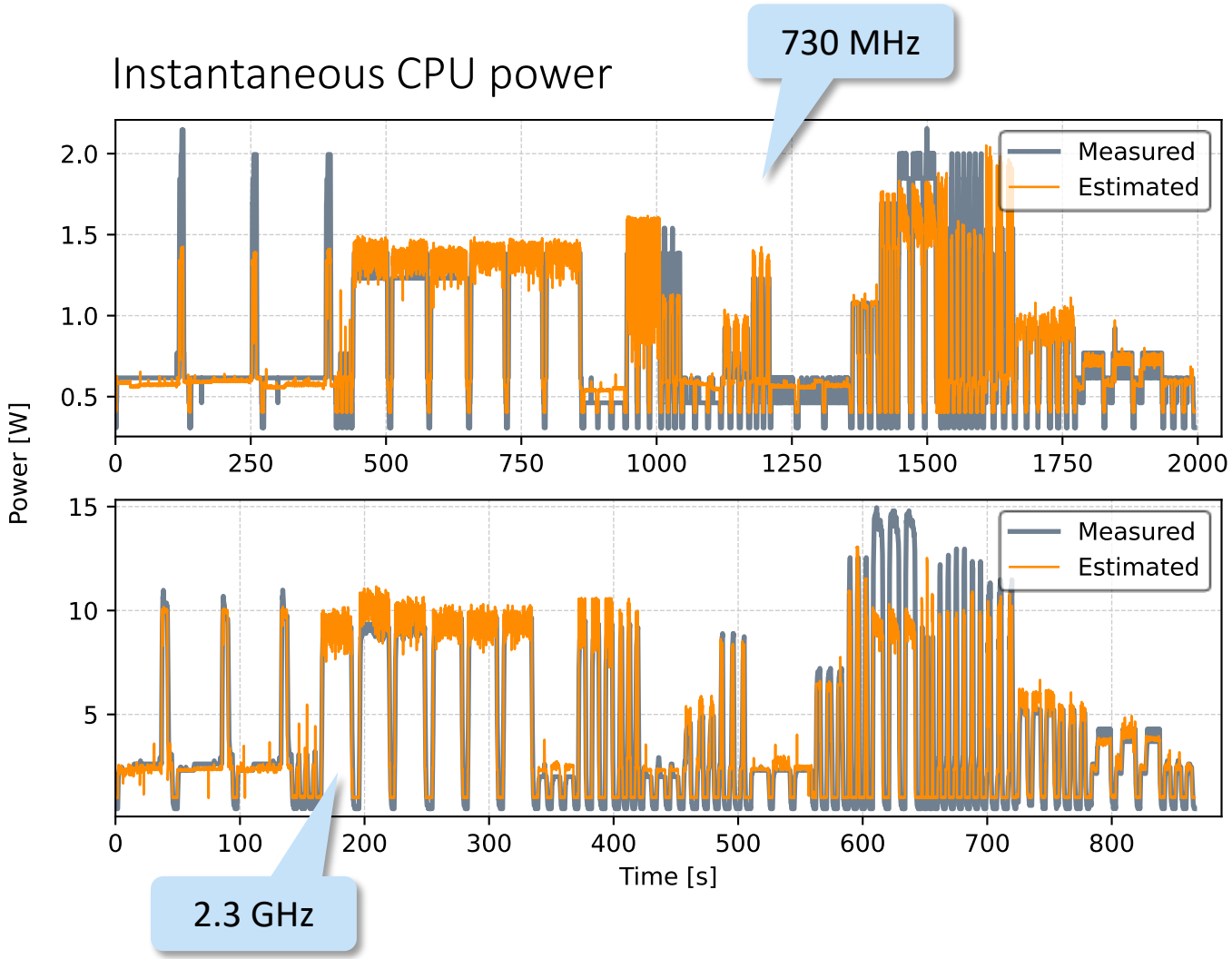




Sub-system models validation – **CPU**

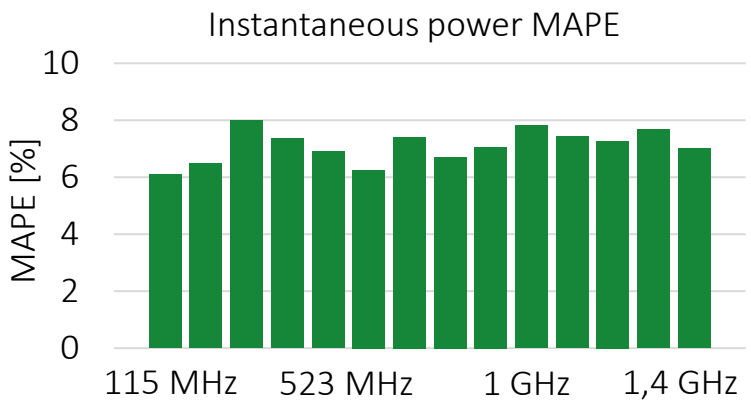
- Power tracked overtime
- **Instantaneous power** MAPE
  - Avg over frequencies = **~4%**
- Total **energy estimation** error
  - Avg over frequencies = **~4%**

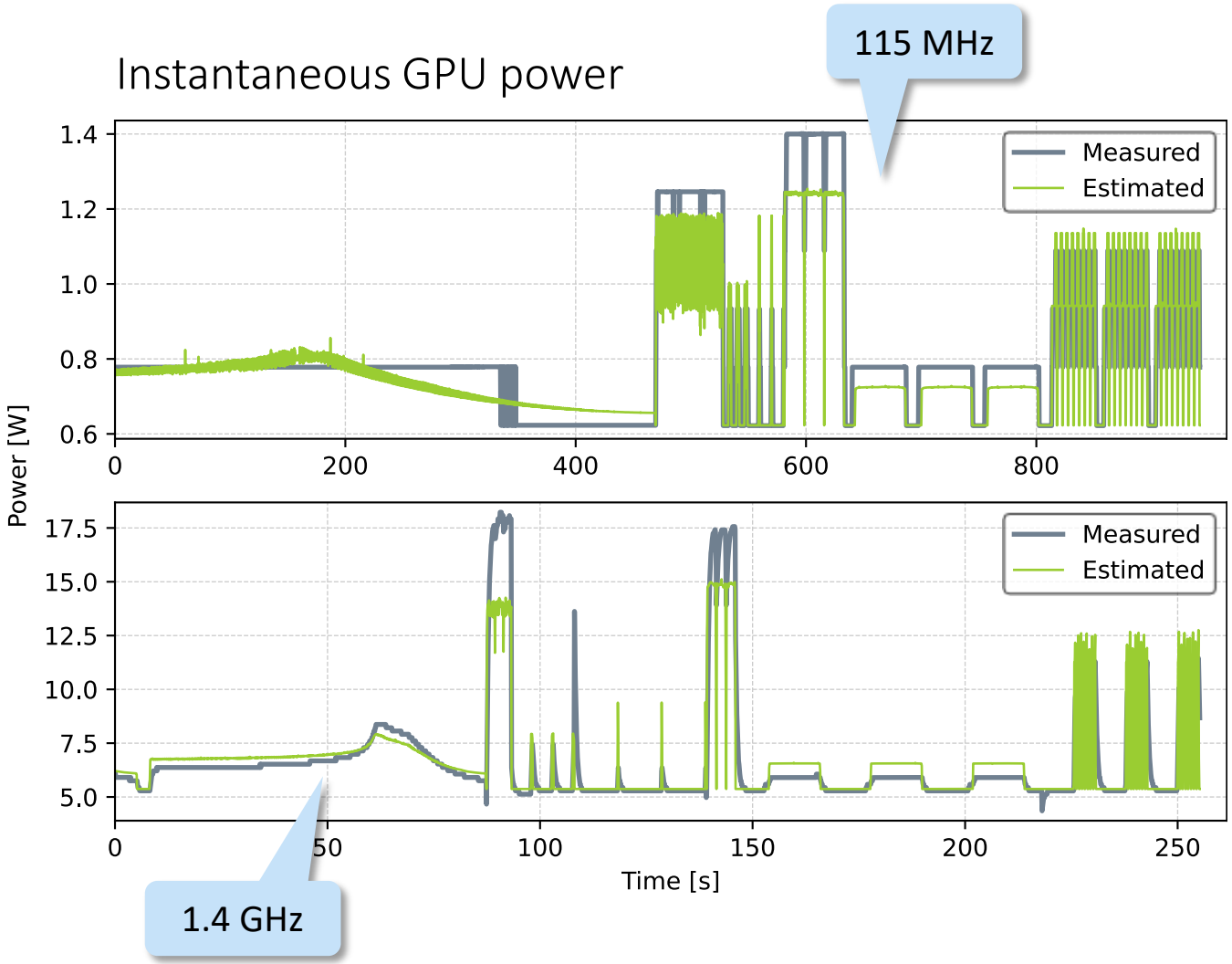




Sub-system models validation – **GPU**

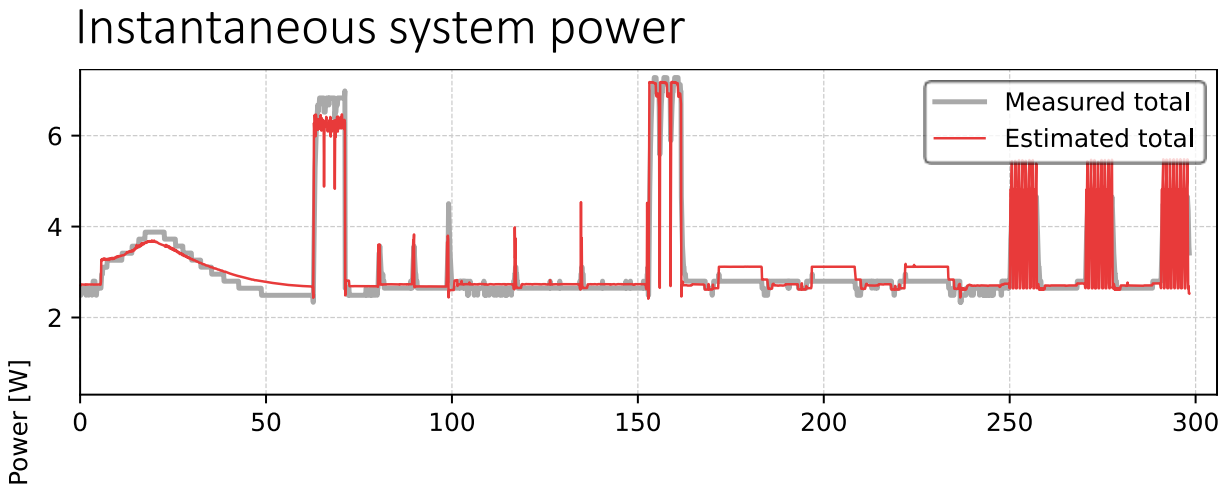
- **Instantaneous power** MAPE
  - Avg over frequencies = **~7%**
- Total **energy estimation** error
  - Avg over frequencies = **~2.2%**
  - Max = ~5.5%



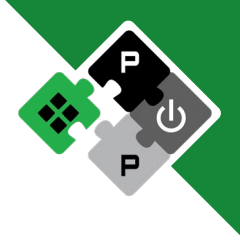


Combined system-level model validation

- LUT-based model not effective with mismatching frequencies
- Limiting to $f_{GPU} > 600\ MHz$
  - Instantaneous **power** MAPE = **~7.5%**
  - Total **energy** estimation error = **~1.3%**
- Target: online DPM policies
  - Floating-point model implemented for online execution
  - Max runtime = **500 ns** @ 730 MHz



Conclusions



- Systematic statistical data-driven counters selection
  - Little manual intervention, general applicability
- LUT-based approach & linear models
  - Addressing heterogeneity + DVFS
- State-of-the-art accuracy + low overhead
  - 1.3% average energy estimation error
- Further work
  - Further validation of our methodology with broader benchmarks and target platforms
  - Online policies based on the proposed models

Acknowledgments

Partially supported by the European projects

- AMPERE
- Fractal

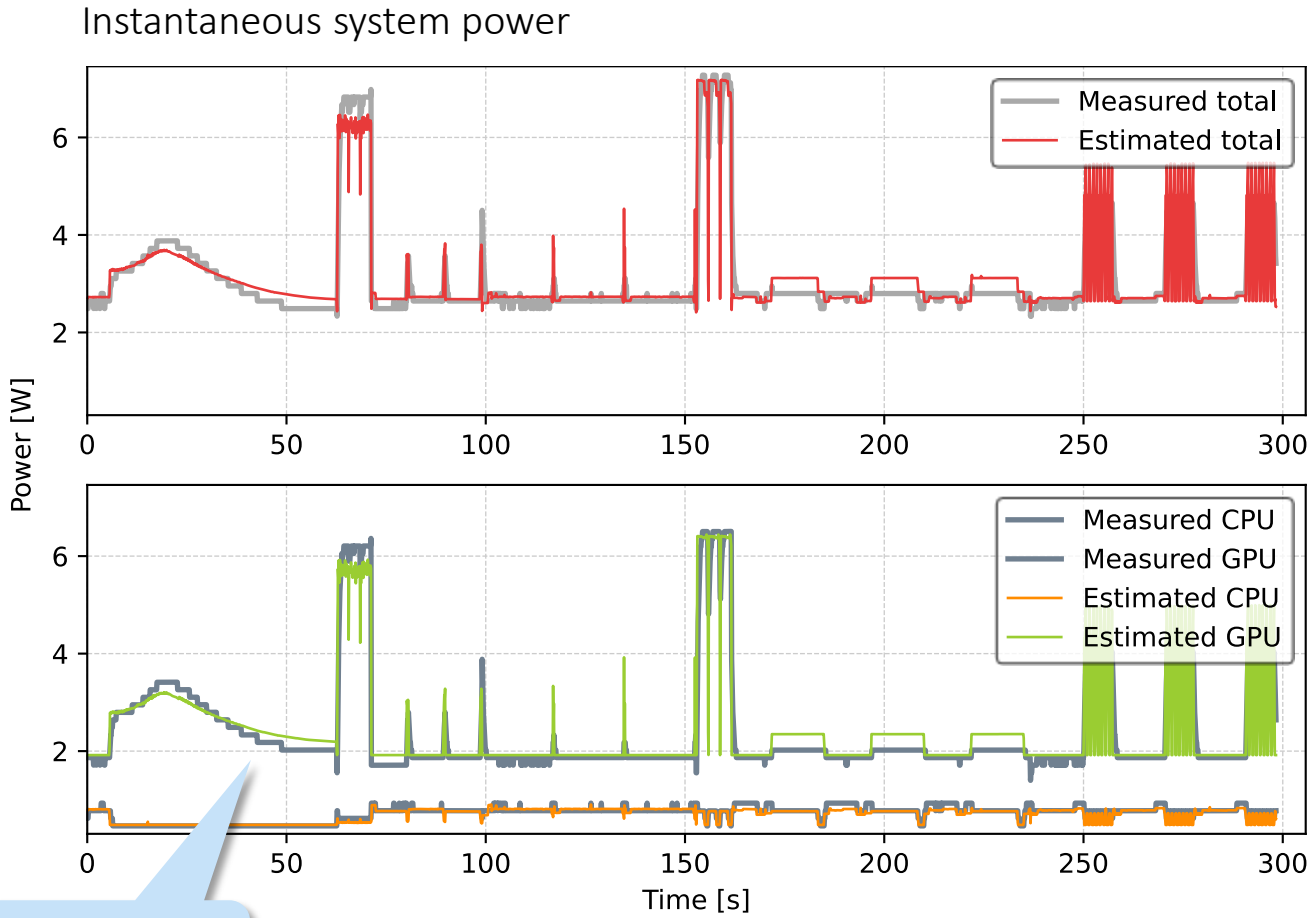
Thank you!

Experimental setup

- **CPU** frequencies
  - 730 MHz, 1.2 GHz, 2.3 GHz
- **GPU** frequencies
  - All 14 frequencies: 115 MHz – 1.4 GHz
- Activity + power **profiler**
  - Entirely runs on the target platform
  - **Continuous sampling** mode, $T_S = 100\ ms$
  - **Time correlation** between activity and power samples

Combined system-level model validation

- Considering all CPU/GPU frequency combinations
- **Instantaneous power** MAPE
  - Avg over frequencies = **~8.6%**
- Total **energy estimation** error
  - Avg over frequencies = **~2.5%**
  - Inconsistent across different frequencies combinations



Combined system-level model validation

- Sub-systems modeled individually
  - Low modeling complexity
  - Broad applicability, flexibility
  - Trade-off with accuracy
- LUT-based model limitations
  - Not effective if frequencies mismatch
  - Complex interactions not grasped
  - But… very edge cases, rarely useful
- Limiting to $f_{GPU} > 600\ MHz$
  - Instantaneous **power** MAPE = **~7.5%**
  - Total **energy** estimation error = **~1.3%**