SuperCoverage: AI-Guided Full Coverage of Thermal and Power Analysis for SoC Design

Xia (Ivy) Zhu, Jianfang (Olena) Zhu,
Mark J. Gallina, Julien Sebot

62 THE CHIPS TO SYSTEMS CONFERENCE

intel®

SPONSORED BY CEDA® IEEE Council on Electronic Design Automation sigda acm

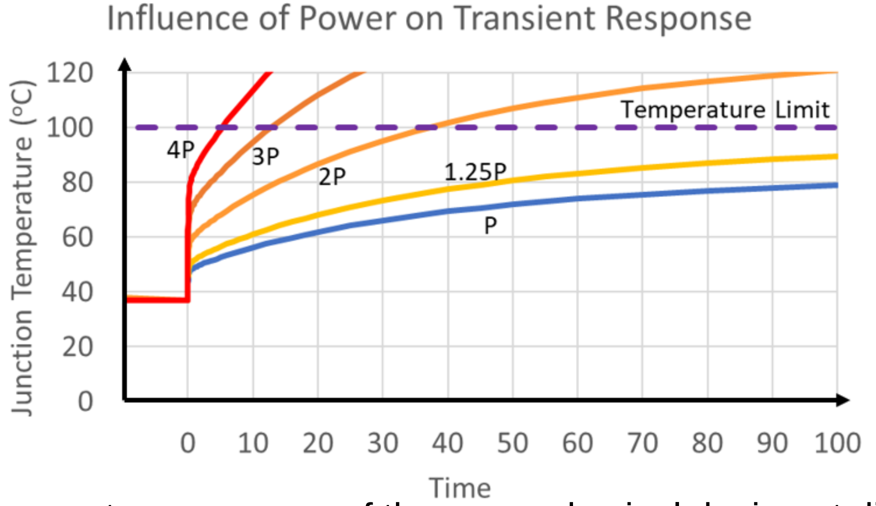
Ever-increasing Power and Thermal Challenges

- Relentless Power Density Increase in core IPs
  - Faster Tj Transients
  - Large In-Core Thermal Gradients
- Further exacerbated by packaging and silicon technology trends
  - 3D Heterogenous Integration
- Workload dependent
  - Different workloads have different thermal response
  - The number of workloads is huge and keeps increasing
- Traditional thermal analysis tools based on steady-state scenarios

Transient temperature response approximately follows an exponential behavior:

$$T(t,P) \sim (T_i - P * R)e^{-t/\tau} + P * R + T_{ambient}$$

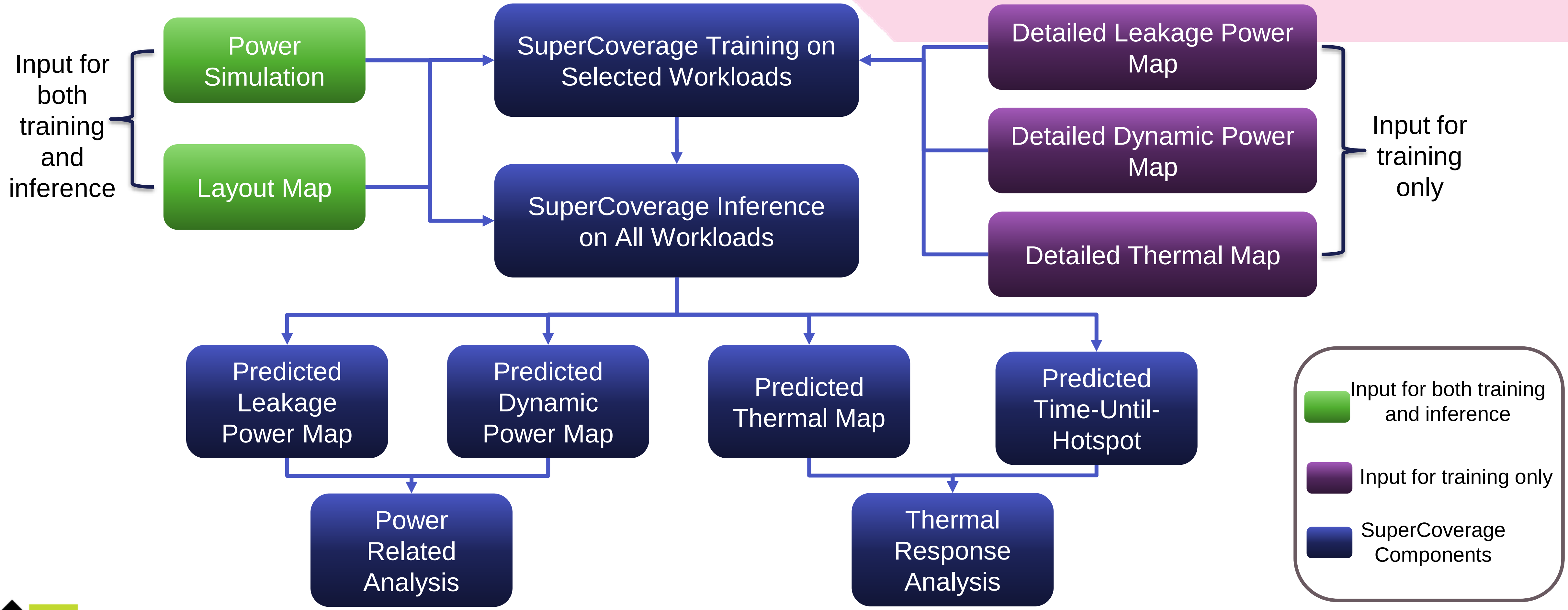
- T: temperature at time t
- t: time



Temperature response of the same physical design at different power levels (P to 4P)

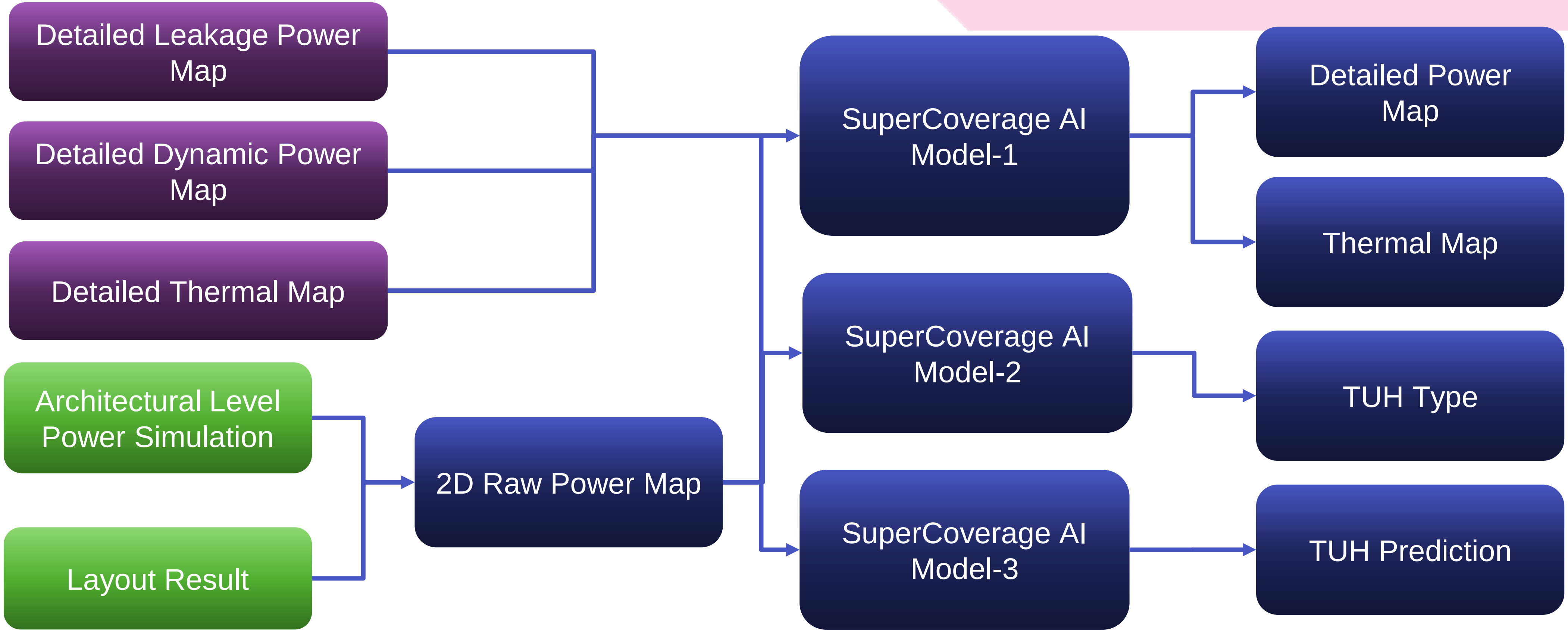
Existing toolset is insufficient to address the complexity of the problem

SuperCoverage Design



SuperCoverage is designed to improve speed and coverage of power/thermal analysis

SuperCoverage Training



TUH: Time Until Hotspot

A set of models are trained for power/thermal analysis

SuperCoverage and Test Results

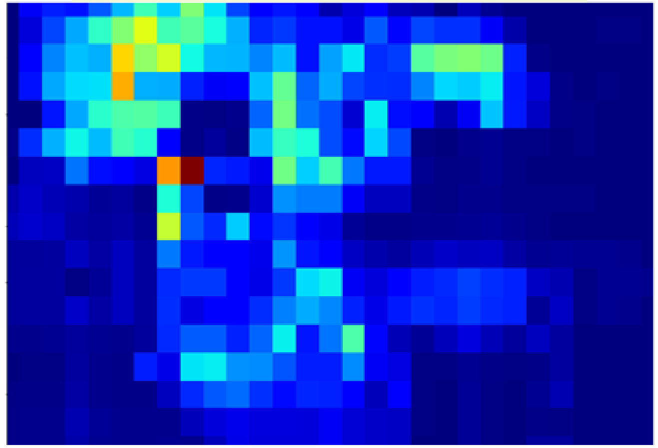
- Customized neural network design
- Configurable network complexity for various dataset
- Special designed loss function with multiple objectives on spatial and frequency domain
  - Temperature delta
  - Structural similarity
  - Least absolute deviations (L1) and least square errors (L2)
- Train on selected workloads
  - 70% train
  - 15% validation
  - 15% test
- Inference on all workloads in trace repo

Metric	AI Prediction Result
Thermal $\boldsymbol{R}^2$	0.9963
Thermal gradient $\boldsymbol{R}^2$	0.9973
Thermal WAPE	0.3%
Power $\boldsymbol{R}^2$	0.9972
Power gradient $\boldsymbol{R}^2$	0.9978
Power WAPE	2.3%

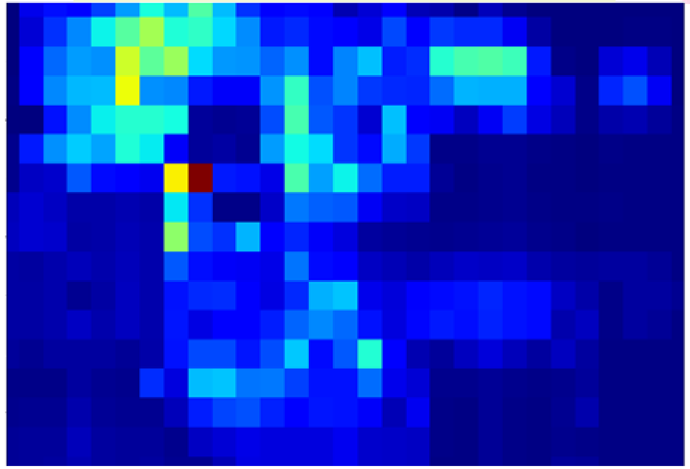
WAPE: Weighted Absolute Percentage Error

SuperCoverage Prediction Results

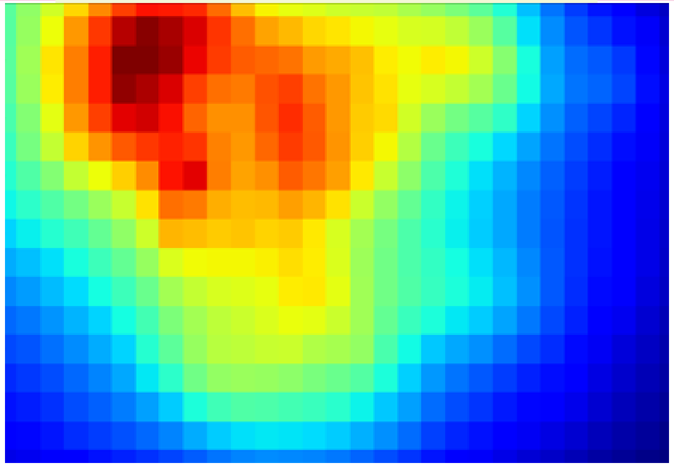
Input Raw Power Map



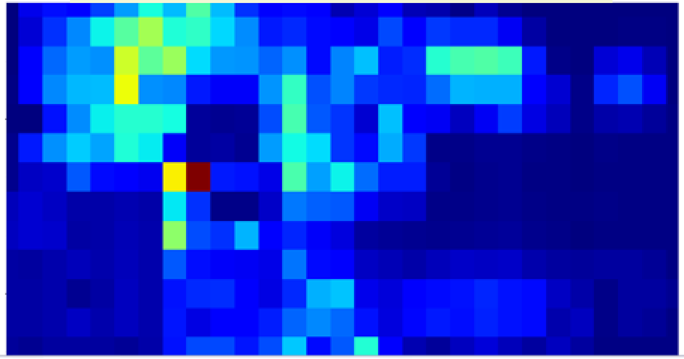
Simulated Detailed Power Map



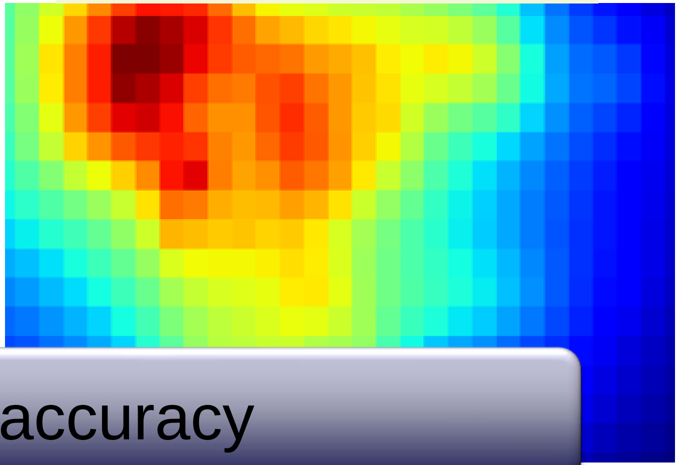
Simulated Thermal Map



Predicted Detailed Power Map



Predicted Thermal Map



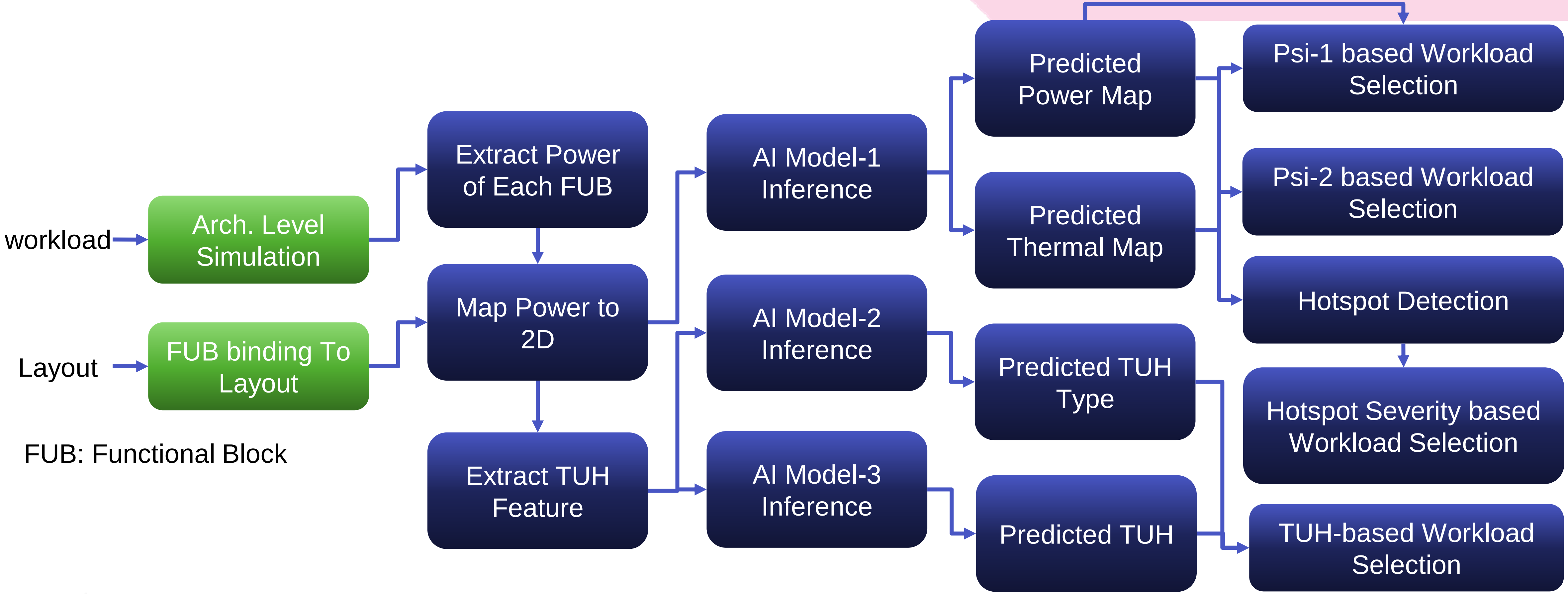
SuperCoverage achieves high accuracy

Thermal Analysis based on SuperCoverage Prediction

- Spatial Temperature Gradient based
  - Global Temperature Gradient based
    - Thermal resistance: $\psi_1 = (T_{max} - T_{ambient})/P_{total}$
    - Thermal resistance: $\psi_2 = (T_{max} - T_{min})/P_{total}$
  - Local Temperature Gradient based
    - Hotspot detection
    - Hotspot severity calculation [1]
      - Maximum Localized Temperature Difference (MLTD)
      - $sev(x, y) = \sigma_{df}(T_{x,y}) + \sigma_M(MLTD_{x,y}) \times \sigma_T(T_{x,y})$
      - $\sigma(x_0, y_0, s, a) = \frac{a}{1+e^{-s(x-x_0)}} + y_0$
      - σ_{df} to model device failure at $T_{failure}$
      - σ_M and σ_T to model marginal contributions of MLTD and temperature for timing issues.
- Temporal Temperature Gradient based
  - Time-Until-Hotspot (TUH) classification
  - Time-Until-Hotspot (TUH) prediction

[1]. Alexander Hankin, David Werner, Maziar Amiraski, Julien Sebot, Kaushik Vaidyanathan, and Mark Hempstead. "Hotgauge: A methodology for characterizing advanced hotspots in modern and next generation processors." In *2021 IEEE International Symposium on Workload Characterization (IISWC)*, pp. 163-175. IEEE, 2021.

SuperCoverage Inference & Analysis

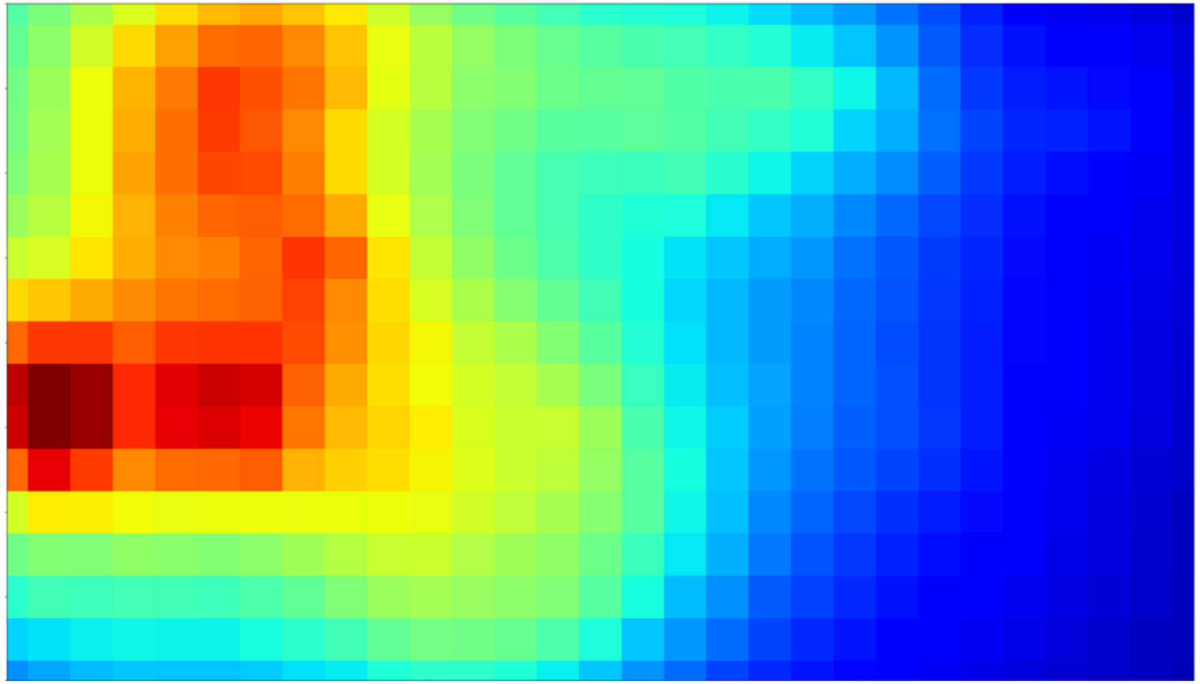


FUB: Functional Block

Enables various analysis, new insight discovery, and critical workload selection

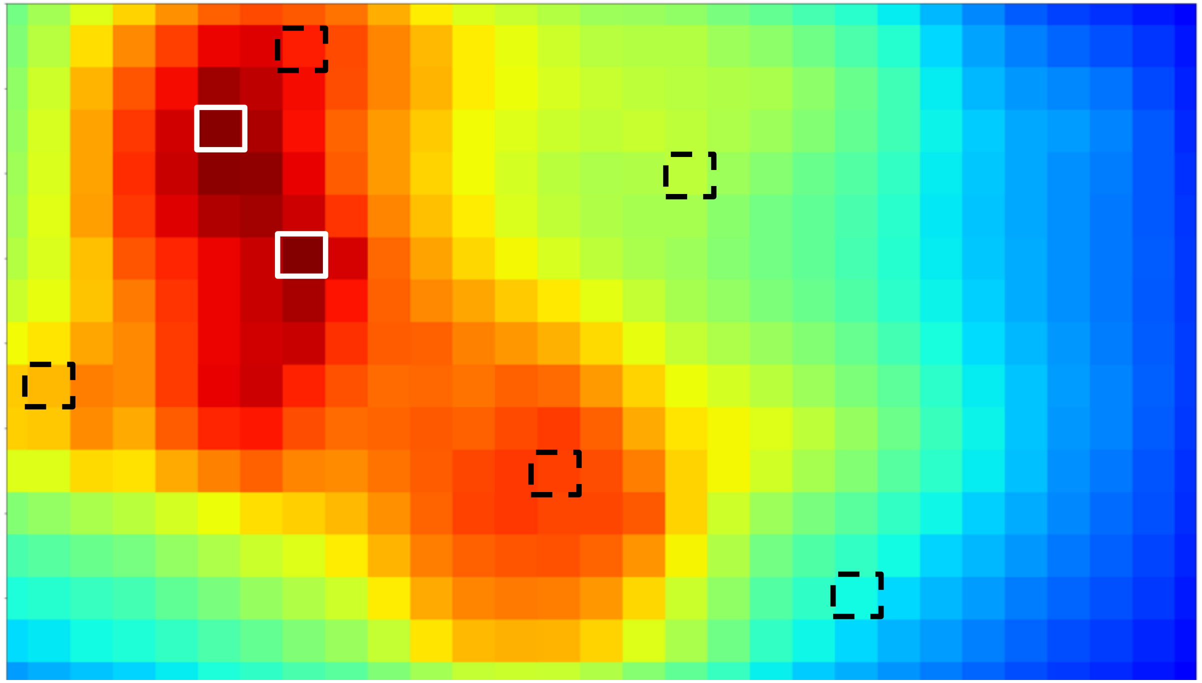
SuperCoverage Results – Identify Hotspots

Thermal ramp map of workload A



SuperCoverage identified workload(s) with up to 4% higher temperature rise than original study list

Thermal ramp map of workload B



SuperCoverage identified new hotspot location. Dashed lines indicate previously identified hot spot locations

Summary

- AI-guided power and thermal analysis can expand workload coverage by ~100X, speedup map generation by ~1000X
  - Largely expand power and performance workload selection
  - Easily identify thermal critical workloads based on spatial & temporal temperature gradient
  - Discover new insights (e.g., higher Tjmax, new hotspots) obtained from full workload analysis, some are non-intuitive
  - Refine workload selection based on different metric thresholds to fit for different product SKUs
  - Speedup thermal & power map generation by ~1000X.

Xia (Ivy) Zhu

Senior Principal AI Engineer
Intel Corporation

Xia (Ivy) Zhu

Intel Senior Principal Engineer

Jianfang (Olena) Zhu

Intel Senior Principal Engineer

Mark J. Gallina

Intel Principal Engineer

Julien Sebot

Intel Fellow

Thank You!

AI

Security

Systems

EDA

Design

THE CHIPS TO SYSTEMS CONFERENCE

SPONSORED BY CEDA sigda