Statistical Thermal Evaluation and Yield Improvement Considering Process Variation for 3D Chip-Multiprocessors

Da-Cheng Juan[†], Siddharth Garg[‡] and Diana Marculescu[†]
[†]Department of Electrical and Computer Engineering, Carnegie Mellon University, PA
[‡]Department of Electrical and Computer Engineering, University of Waterloo, ON, Canada

INTRODUCTION

❑ Thermal concerns for three-dimensional (3D) integrated circuits (ICs) are exacerbated due to higher power density and lower thermal conductivity of inter-tier dielectrics.

❑ Increased leakage power dissipation due to technology scaling further deteriorates these thermal problems.

❖ Interdependency between temperature and leakage power forms a feedback loop, which may lead to thermal runaway.
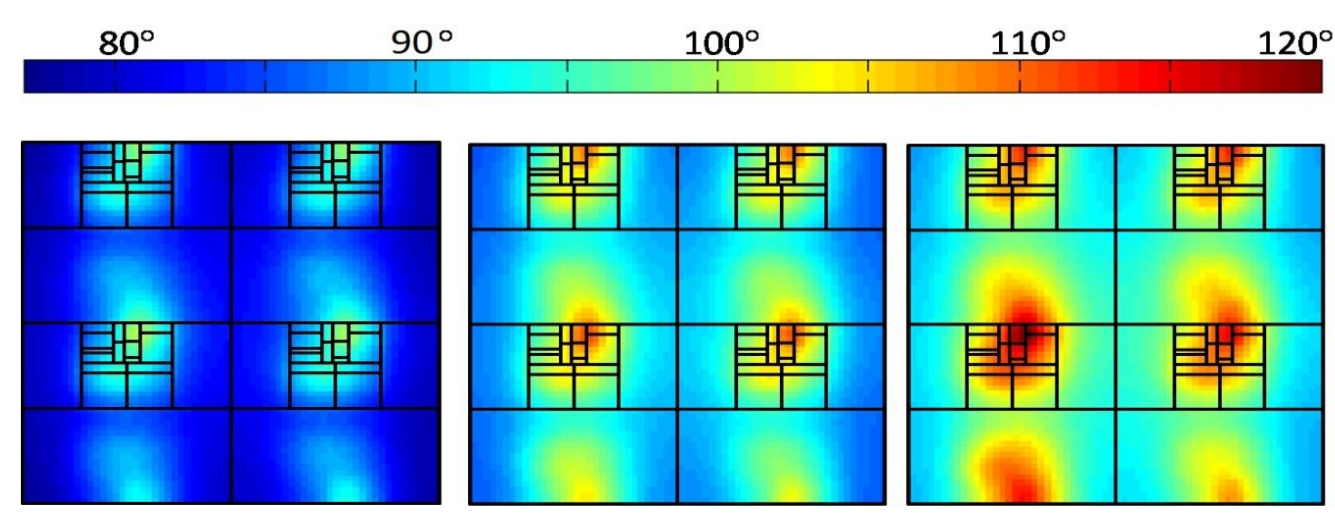
(a)nominal (b) mild variation (c) severe variation
Temperature maps of top tier in a 3D CMP.

MANUFACTURING PROCESS VARIATIONS

❑ Process variations affect several important metrics of an IC, such as leakage power and maximum clock frequency.

❑ Variation of effective transistor channel length in 3D systems can be described as:

❖ $L_{eff} = L_{nom} \pm \Delta_{total}$

❖ $\Delta_{total} = w_1\Delta_{w2w} + w_2\Delta_{spat} + w_3\Delta_{rand}$

❑ In general, the variation assumptions are used:

❖ $\Delta_{total} = 5\% \times L_{nom}$

❖ $\Delta_{w2w}, \Delta_{rand} \sim N(\mu, \sigma)$

❖ Δ_{spat} is a deterministic model by [Cheng et al. DAC'09]

❖ $w_1 : w_2 : w_3 = 0.7 : 1 : 1$ by [J. Sartori et al. ISQED'10]

LEAKAGE POWER VARIATIONS

❑ Leakage power is highly sensitive to process variations and operating temperatures.

❖ Interdependency between temperature and leakage power forms a feedback loop, which may lead to thermal runaway.

❖ $I_{leak} = K' \frac{W}{L_{eff}}(T)^2 e^{-\frac{c \cdot V_{th}}{T}}$
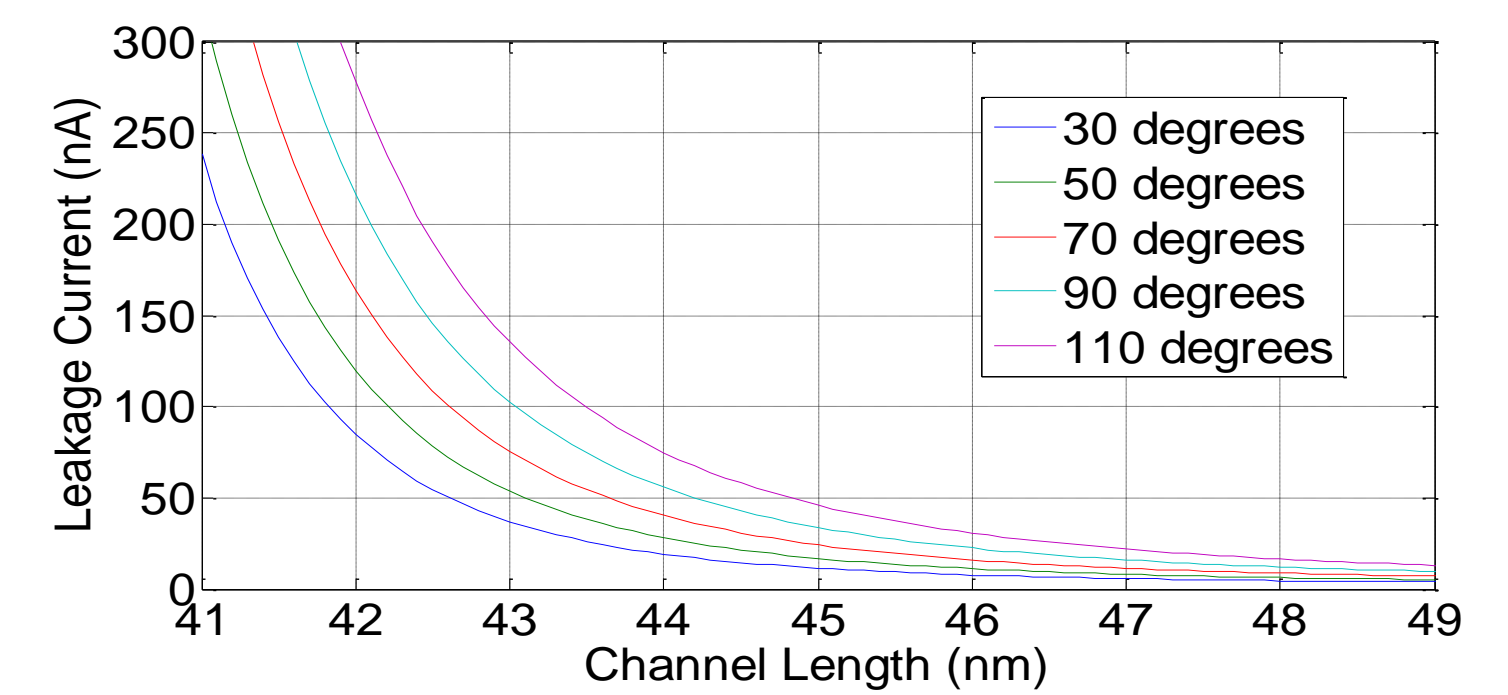
TARGET ARCHITECTURE

Processor Parameters	Values
Number of cores	16
Frequency	3.0 GHz
Technology	45nm node with V_{dd}=1.0V
On-chip network	4×4 mesh
L1- I/D caches	64KB, 64B blocks, 2-way SA, LRU
L2 caches	1MB, 64B blocks, 16-way SA, LRU
Pipeline	7 stage deeps, 4 instructions wide
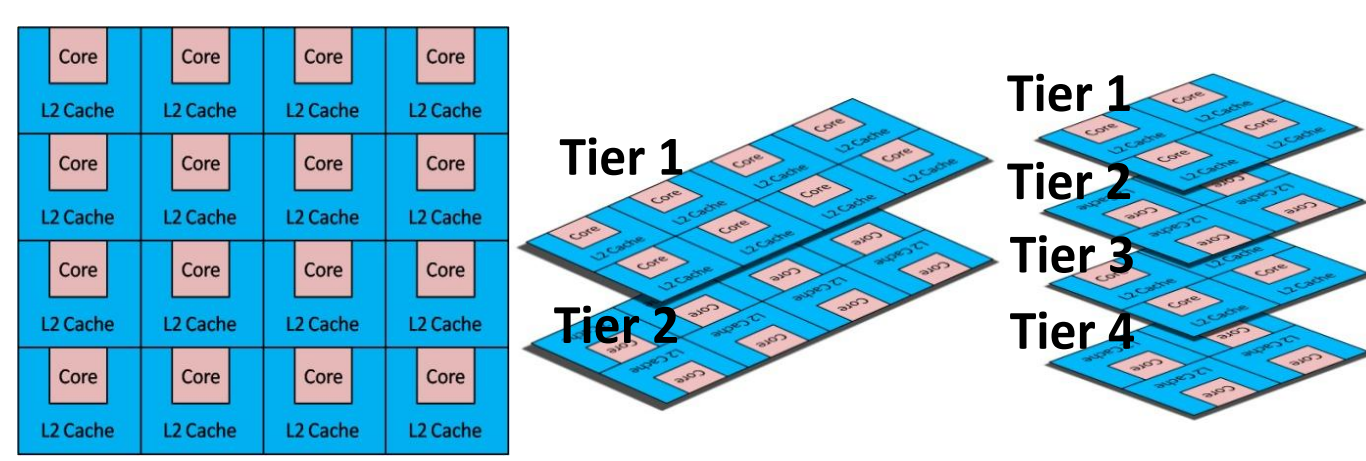
(a) 1-tier CMP (b) 2-tier CMP (c) 4-tier CMP
2D and 3D CMP implementation

MAXIMUM TEMPERATURE PREDICTION

❑ The maximum temperature for each 3D system is not known before fabrication in the presence of leakage variations.

❖ Using Hotspot [Skadron et al, TACO'04] or other simulation-based methods can be too time-consuming.

❑ We propose a learning-based regression model to predict the maximum temperature for the 3D system under steady-state conditions.

❖ In the *learning* phase:

$$(\hat{a}_i, \hat{c}) = argmin\left\{ \sum_{k=1}^{m} \left(T_k^{max} - \sum_{i=1}^{n} a_i P_{leak}^{ik} - c \right)^2 \right\}$$

❖ In the *testing* phase:

$$\hat{T}^{max} = \sum_{i=1}^{n} \hat{a}_i P_{leak}^i + \hat{c}$$

TEMPERATURE PREDICTION RESULTS

❑ Prediction accuracy:

❖ Correlation coefficient = 0.97,

❖ Cross-validation root-mean-square-error < 2%.
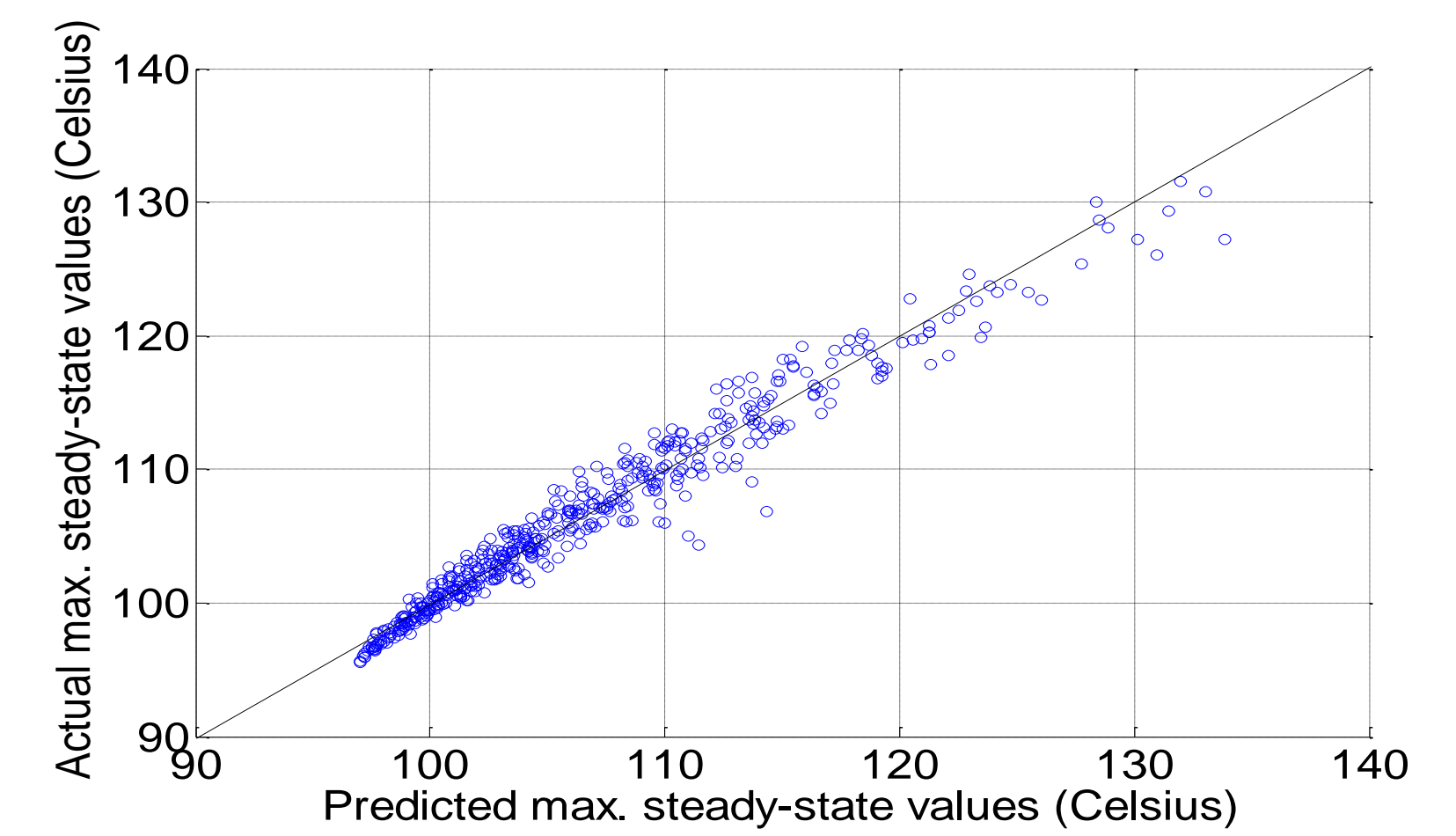
TIER STACKING

❑ $\hat{a}_i$ indicates that the leakage value of each tier has a different impact on the maximum temperature.

❖ Re-stack the tiers based on the leakage values × $\hat{a}_i$ to achieve a potential thermal reduction.

❖ Note that $\hat{a}_i$ is not monotonically decreasing, so this stacking technique is exclusively enable by our learning model.

❖ This stacking technique would only be applicable for symmetric 3D systems.

❑ Searching for the best stacking order for 1,000 4-tier CMPs:

❖ Using Hotspot simulation takes more than 5 days.

❖ Our learning model only needs 4 hours.

❖ A 30X speed-up is achieved.

IMPLEMENTATION FLOW
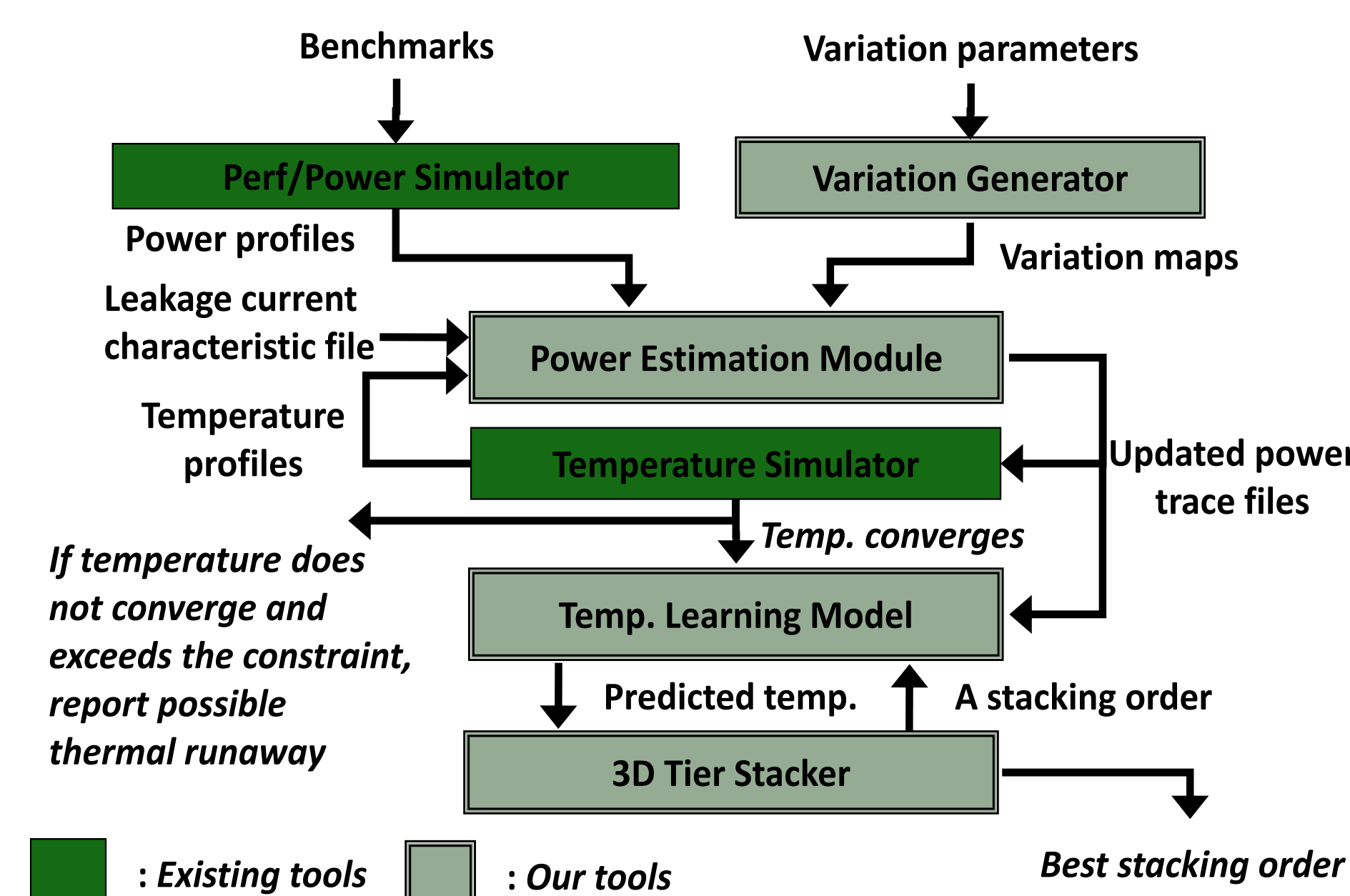
TRANSIENT THERMAL BEHAVIOR

❑ Leakage variations may alter the time point when the maximum temperature occurs.

❑ An unexpected high thermal peak occurs, which is completely different from the thermal behavior of Nominal.
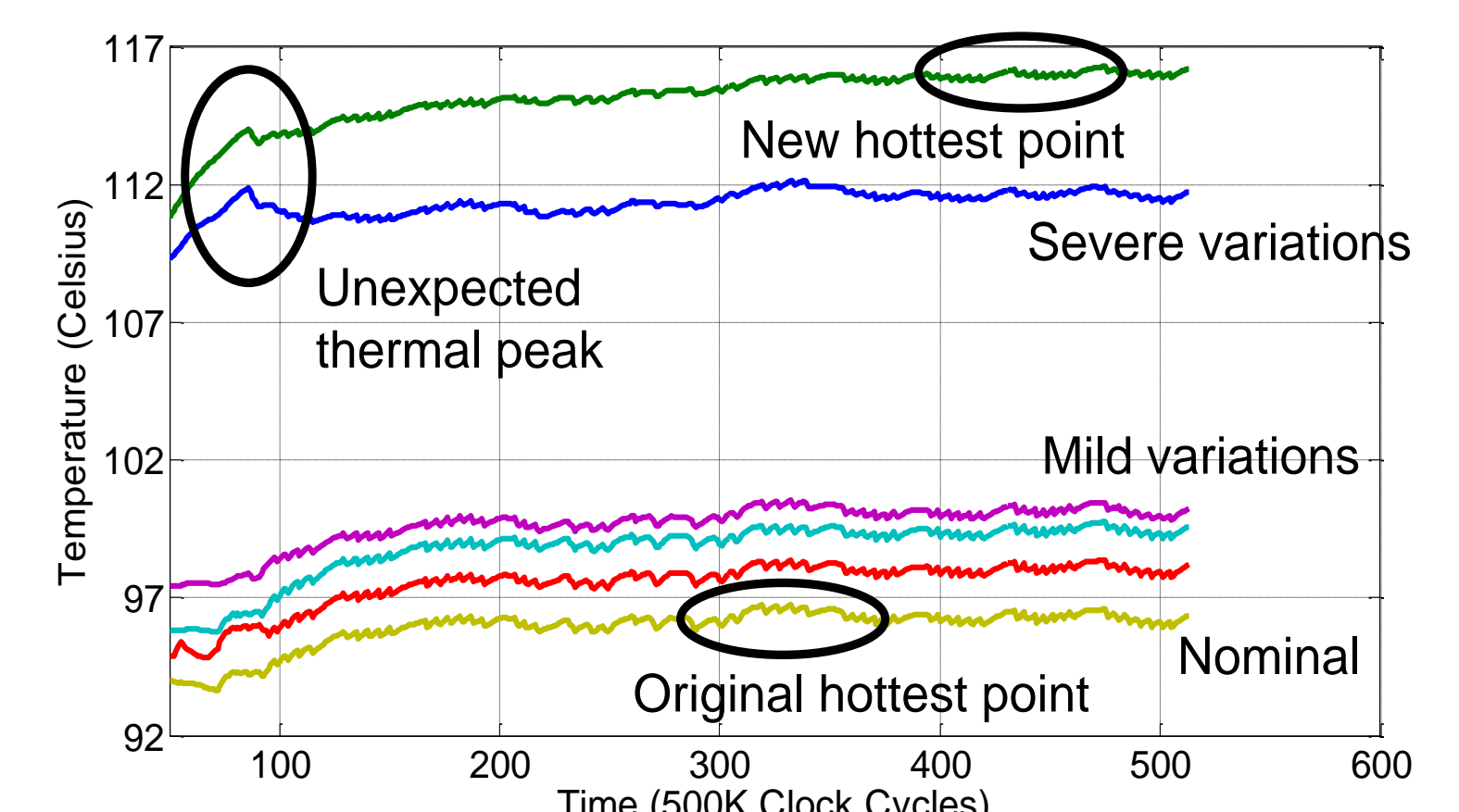

MAXIMUM TEMPERATURE DISTRIBUTION

❑ The distribution for the 2D implementation is very narrow with a standard deviation of only 0.11˚C.

❑ In 3D CMPs, the standard deviation of the maximum temperature distribution is significantly larger. For a 4-tier CMP, the standard deviation dramatically increases to approximately 40 times higher than that of a planar CMP.
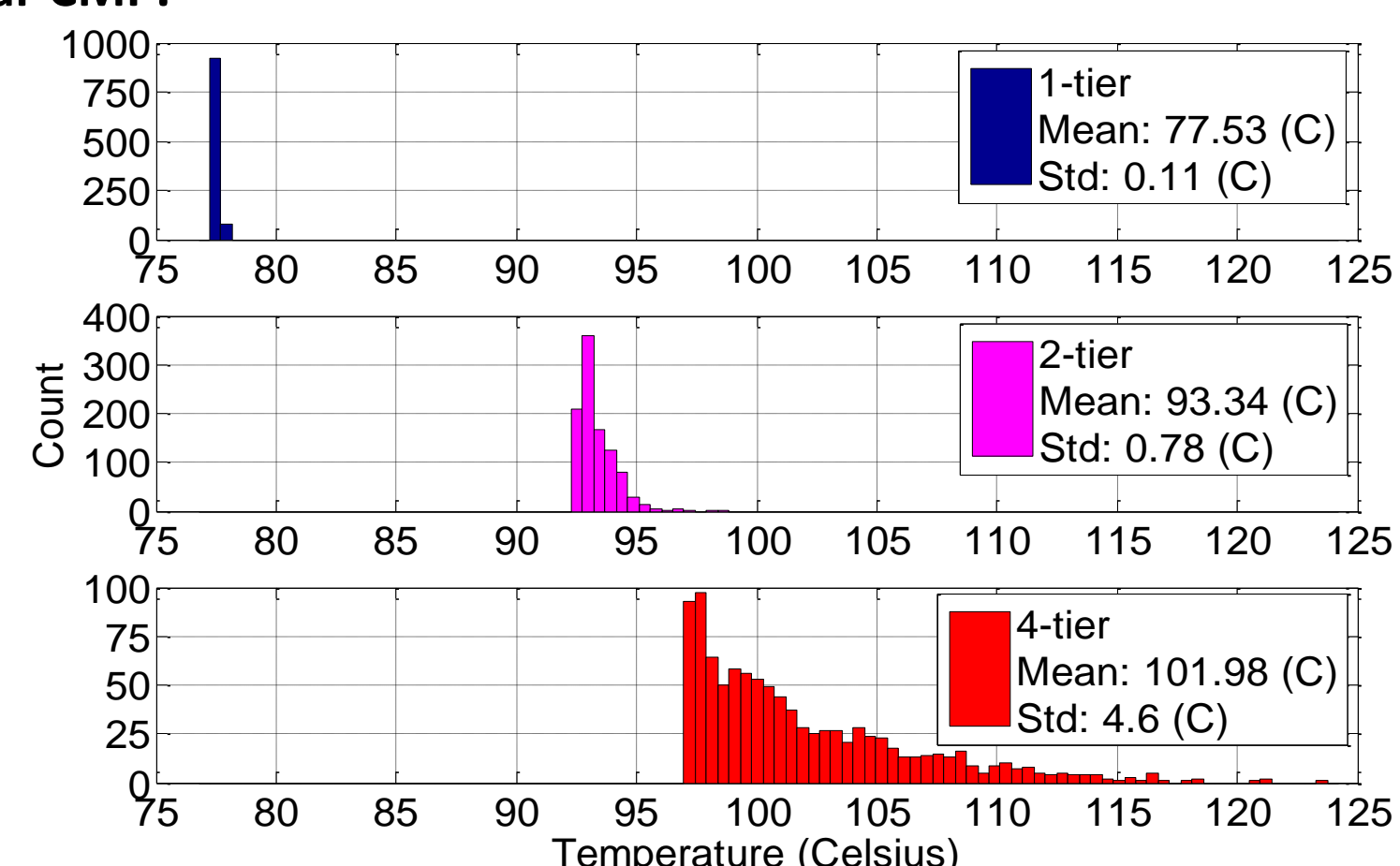
TIER STACKING IMPROVEMENT

❑ The standard deviation of the maximum temperature distribution is reduced by 54%, from 4.6˚C to 2.11˚C.

❑ If the temperature constraint is set to 105˚C, the improved yield is 98.0% compared to the original yield of 78.1%.
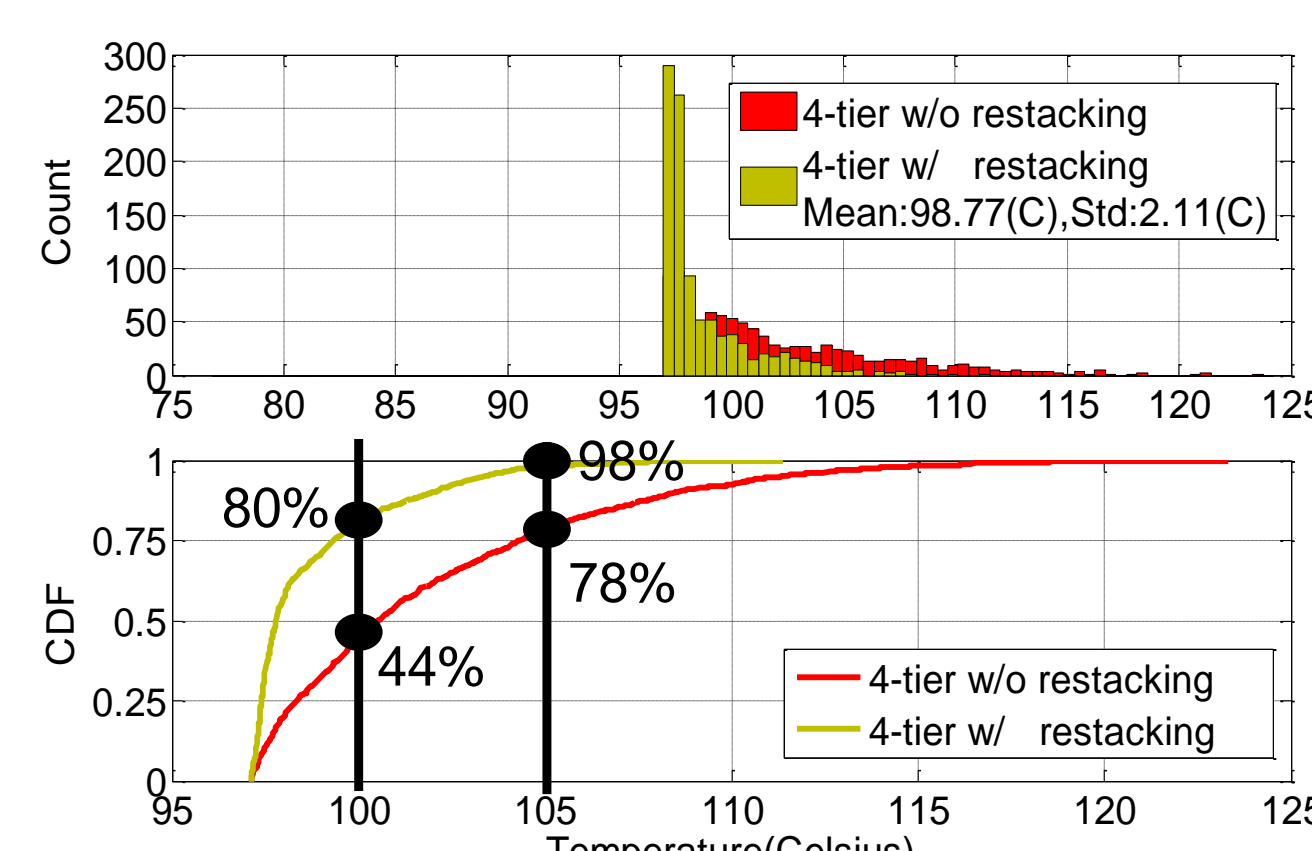
CONCLUSION

❑ We perform statistical thermal evaluation for 3D ICs.

❖ 3D systems are much more susceptible to process variations than their 2D counterparts.

❑ We propose an accurate learning-based regression model to predict the maximum steady-state temperature.

❖ No extra time-consuming simulations required after the coefficients are learnt.

❖ Highly accurate (RMSE < 2%) and can be used in an iterative design exploration environment for improving thermal yield.

❑ We propose an effective algorithm to determine the best tier stacking order that minimizes the maximum temperature and maximizes the thermal yield.