Machine Learning Based Thermal Evaluation for Vertically-Composed Fine-Grained 3D CMOS

Mingyu Li [1], Sourabh Kulkarni [2], Sachin Bhat [2], and Csaba Andras Moritz [2]

[1]University of Massachusetts Amherst
[2]Affiliation not available

October 30, 2023

Abstract

Thermal management in 3D integrated circuits is a critical challenge due to their high computational density. Heat dissipation paths from top circuit layers through bottom layers to substrate are heavily constraining heat extraction. Various thermal management frameworks have been proposed to address thermal issues in different granularities. All these frameworks require a thermal evaluation stage that characterizes the thermal profile of large designs with fast runtime. In this work, we present a machine learning based thermal evaluation method that predicts all standard cell temperatures based on features extracted from circuit CAD files. We have built thermal resistance networks for 10 benchmark circuits. We performed simulations to achieve the thermal data, and trained the thermal model with the data. The model is highly accurate and can identify all over-heated cells that need to be thermally-optimized. Runtime overhead is minimal. For a 435k-cell SPARC T2 core, the runtime for predicting all cell temperatures is as small as 3.12s, which is negligible compared to the runtime of other physical design stages.

Machine Learning Based Thermal Evaluation for Vertically-Composed Fine-Grained 3D CMOS

MINGYU LI, Sourabh Kulkarni, Sachin Bhat, Csaba Andras Moritz

University of Massachusetts, USA

Thermal management in 3D integrated circuits is a critical challenge due to their high computational density. Heat dissipation paths from top circuit layers through bottom layers to substrate are heavily constraining heat extraction. Various thermal management frameworks have been proposed to address thermal issues in different granularities. All these frameworks require a thermal evaluation stage that characterizes the thermal profile of large designs with fast runtime. In this work, we present a machine learning based thermal evaluation method that predicts all standard cell temperatures based on features extracted from circuit CAD files. We have built thermal resistance networks for 10 benchmark circuits. We performed simulations to achieve the thermal data, and trained the thermal model with the data. The model is highly accurate and can identify all over-heated cells that need to be thermally-optimized. Runtime overhead is minimal. For a 435k-cell SPARC T2 core, the runtime for predicting all cell temperatures is as small as 3.12s, which is negligible compared to the runtime of other physical design stages.

1 INTRODUCTION

3D integration is an emerging technology direction to enable surpassing many of the current limitations in traditional CMOS scaling, including interconnection bottlenecks [1] [2] [3] [4] [5] [6] [7] [8]. Most research to date realizes 3D integrated circuits (ICs) with layer-by-layer stacked implementations, utilizing either parallel or monolithic 3D integration. These directions achieve only around a half technology node's PPA (power, performance, area) benefit vs. 2D CMOS, lead to very limited density benefits, and often suffer from reduced routability vs. 2D CMOS due to severe pin / routing congestion [8]. In addition, thermal management in 3D ICs has become a critical challenge [9] [10]. In the 3D technologies to date, thermal management is usually performed at a coarse granularity, optimizing regions containing hundreds or thousands of standard cells.

Confronted with the limitations in routability and PPA benefits, we proposed Skybridge-3D-CMOS (S3DC), a vertically-composed fine-grained 3D CMOS technology, which features much improved pin access, routing flexibility, and is based on fine-grained vertical circuits yielding dramatic efficiencies [11] [12] [13] [14] [15]. Results have shown that S3DC leads to significant benefits in power, performance and density, with 9.7 to 71X PPA benefits vs. the state-of-art transistor-level monolithic 3D approach, while maintaining excellent routability.

In this paper, we present a fast and accurate machine learning based thermal evaluation methodology. It supports thermal management by characterizing all the standard cell temperatures across the design and identifying cells that need to be thermally optimized. In this paper, we focus on evaluating this for the S3DC technology. However, the machine learning approach and lessons learned are applicable to other fine-grained 3D IC directions as well.

The fundamental reasons for the thermal issues in 3D ICs are the higher densities and challenging heat dissipation paths vs. 2D circuits. On one hand, the densely packed transistors in 3D ICs lead to high power density and more generated heat per unit area. On the other hand, the generated heat in the top layers of circuits has to dissipate through the bottom layers before reaching substrate, making the heat dissipation more difficult compared with 2D circuits [9] [10] [16]. These issues are common in various 3D IC directions in general.

Most research to date incorporates thermal considerations as an afterthought and performs optimizations at a coarse granularity [17] [18] [19] [20] [21]. These works can be classified into two directions. One is to focus on reducing the power density in overheated regions, either by inserting empty spaces to lower the local power density of these regions [19], or by re-distributing standard cells or partitions that have high heat generation to neighboring regions [20] [21]. The other way is through *Through Silicon Vias* (TSVs) to improve the heat dissipation in overheated regions, either by placing more signal TSVs at regions with higher temperatures [18] or by inserting additional dummy TSVs that are only for thermal purposes (Thermal TSVs) – this is to increase the number of thermal paths between tiers [17]. In S3DC we have proposed 3-D thermal management fabric components that directly support thermal management at a *fine granularity* [22]. These are specially architected fabric components; they are also integrated as intrinsic parts of the circuits during an electrical-thermal circuit co-design CAD flow.

The first step in all thermal management frameworks is to perform thermal evaluation and achieve the thermal profile of circuits, such that thermal optimization can be performed accordingly. The way most research to date perform thermal evaluation is: 1) divide circuit into smaller blocks / meshes, 2) create a thermal model consisting of lumped thermal resistance for each block / node of the mesh, and 3) solve the network to achieve the temperatures across the design. The modeling granularity is usually coarse-grained, as otherwise it becomes computationally prohibitively expensive for large-scale circuits. However, coarse-grained thermal

evaluation would not be sufficient for technologies with poor lateral heat dissipation, including monolithic 3D [10] [16]. In these technologies, temperatures can change quite abruptly even across a small distance due to the lack of lateral heat dissipation, thus requiring thermal evaluation at finer resolution.

In this paper we describe a machine-learning based thermal evaluation methodology to address the need for characterizing thermal profile across large-scale circuits with small runtime overhead and at a fine granularity. First, we have built physical thermal model in a bottom-up manner for our fabric components and standard cell designs; we present a flow which automatically extracts the thermal resistance network for large-scale 3D circuits, and solves the thermal resistance network through SPICE simulation to generate all cell temperatures across the design. Then, we introduce a different machine learning based thermal evaluation method. This can predict the results of thermal network simulation based on the statically extracted circuit information from CAD files, and achieve all standard cell temperatures without having to solve the thermal network. We have explored 6 different machine learning models; the XGBoost model has achieved the best accuracy and is thus selected. We found the machine learning based approach to be much faster and more scalable compared with the previous simulation-based methods; we also found it to be sufficiently accurate to identify all over-heated cells and support the subsequent thermal-aware circuit design stages.

We have studied 10 circuit benchmarks in S3DC to validate the overall flow. The machine-learning based thermal model has shown good prediction accuracy with less than 1.1% error. The mean absolute error of prediction results is 3.55K compared to simulation-based results. It identifies all the over-heated cells without introducing significant False Positives – only 0.85% cells that are below the maximum allowed temperature are incorrectly predicted as over-heated. The approach is applicable to other 3D technologies as well, simply by creating their own physical thermal models, and continuing the rest of the steps including simulation on thermal resistance network to generate the training data and training the machine learning thermal model as described here.

The rest of the paper is organized as follows. In Section 2 we provide a brief overview of the S3DC technology. In Section 3 we introduce the fabric-level thermal management support. In Section 4 we provide a brief overview of thermal-aware automated circuit design flow and how it is supported by our proposed thermal evaluation methodology. In Section 5 we introduce the baseline thermal evaluation method based on fine-grained thermal network simulation. In Section 6, we present a machine learning based thermal estimation method that is accurate and has much better scalability compared with the baseline method. Section 7 concludes the paper.

2 SKYBRIDGE-3D-CMOS FABRIC OVERVIEW

Skybridge-3D-CMOS (S3DC) is a vertically composed fine-grained 3D CMOS IC technology [11] [12] [13] [14]. It is enabled by a systematic way of designing static CMOS circuits in a skeleton-style nanowire structure. All the circuits are built on the uniform vertical nanowire template, which is pre-doped with p- and n-type horizontal stripes as shown in Fig. 1(A). We place and connect active devices on these nanowires either in series or in parallel to form the pull-up (with p-type transistors in p-doped region) and pull-down (with n-type transistors in n-doped region) networks in static CMOS gates. Series networks are built with devices implemented on one nanowire. Parallel networks are built with devices on different nanowires; these different nanowires are shorted together on both drain and source sides. A specially designed fabric component called Skybridge-Interlayer-Connection (SB-ILC) enables connecting the p-type pull-up and n-type pull-down networks together to generate

the output signal. The SB-ILC structure and materials (shown in Fig. 1(B)) are designed to provide connection between different doping regions with small parasitic resistance and capacitance.

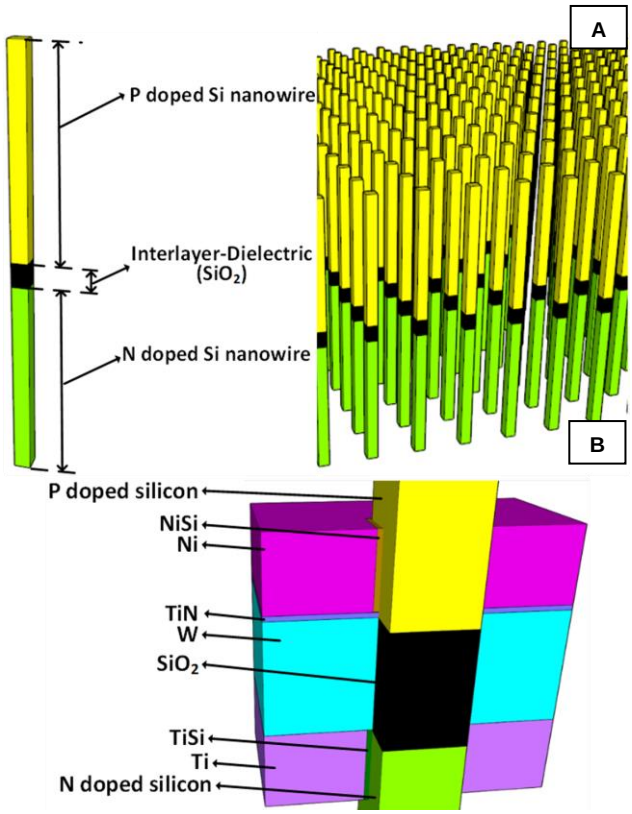

Figure 1: a) One Single Nanowire with Striped Doping and a Uniform Vertical Nanowire Template; (b). SB-ILC Allows Routing between Various Doping Layers without MIVs.

Other S3DC fabric components are shown in Fig. 2. (i) An n-type Vertical Gate-All-Around (V-GAA) Junctionless transistor structure is shown in Fig. 2(A). The source, channel, and drain regions are based on heavily doped vertical nanowires. The channel is surrounded by gate electrodes and dielectric layers. (ii) Fig. 2(B) shows the routing structures. Routing Bridges are horizontal metal wires connecting adjacent vertical nanowires. Routing Nanowires are vertical nanowires that can also act as routing elements since they are heavily doped and silicided, having high conductivity. Coaxial Routing structures are metal layers formed along the vertical nanowires to add connectivity in vertical directions.

Fig. 3 shows an example of logic-implementing circuit utilizing the above concepts. It consists of 3 NAND2X1 gates and implements $\overline{\overline{A \cdot B} \cdot \overline{C \cdot D}}$. The p-type transistors on the top are connected at the source side by VDD, and on the drain side by the SB-ILCs. Thus, the pull-up network is parallel. The n-type transistors at the bottom are connected in series by the vertical nanowire. They form the pull-down network. SB-ILCs connect the pull-up and pull-down networks to generate the output signal, which is conducted out by the Bridges. The outputs

of 1st-stage-NAND2-gates are accessed through Ohmic contact inserted in the middle of NWs and fed to the gate electrodes of 2nd-stage NAND2 gates through Bridges.

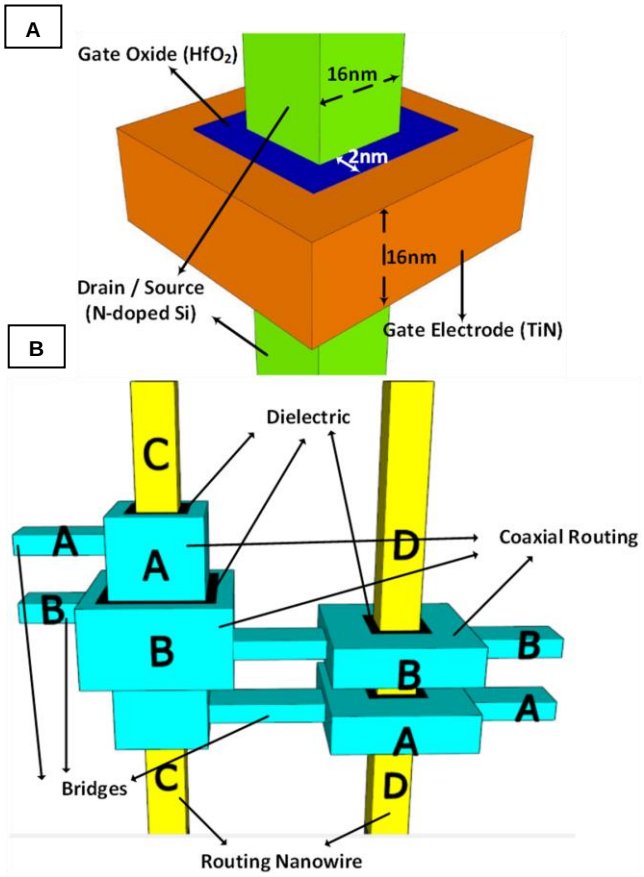

Figure 2: (a). An n-type V-GAA Junctionless Transistor in 16nm S3DC Technology; (b). 3D Connections within One Doping Layer Realized by Bridges, Coaxial Routings, and Routing Nanowires; Four Signals A, B, C, D are Carried in This Example.

We have developed a device-to-system level design flow incorporating commercial CAD tools [12] [13]. The design flow allows us to perform large-scale benchmarking in S3DC to quantify its benefits vs. baseline technologies, including 2D and the state-of-art 3D directions such as transistor-level monolithic 3D (M3D). We have evaluated routability, as well as performance, power, and area (PPA). As compared to the usually severely congested monolithic 3D implementations, S3DC eliminates the routing congestions in all benchmarks studied. Further results, for the implemented benchmarks, show 56%-77% reductions in power consumption, 4X-30X increase in density, and 20% loss to 9% benefit in best operating frequencies compared with the transistor-level monolithic 3D technology. The loss is likely caused by less optimized S3DC devices vs industrial and is primarily in smaller circuits where better routability benefits are not that accentuated. Today, Skybridge technologies have attracted attention from several leading vendors (under NDA) who initiated a partnership relationship. A full scalable manufacturing flow based on industrial requirements and processes is under development in

collaboration, through the utilizing of state-of-art process development tool Synopsys Sentaurus™ Process Explorer.

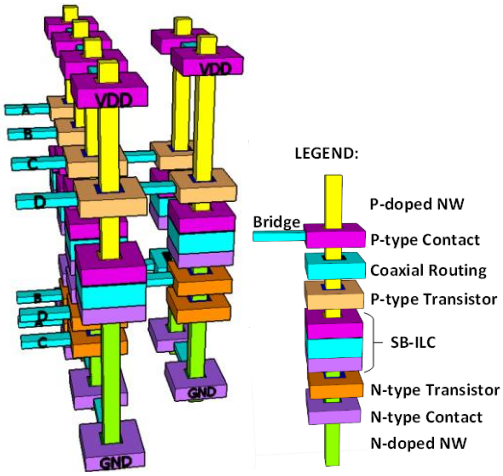

Figure 3: S3DC $\overline{\overline{AB} \cdot \overline{CD}}$ Circuit Layout (Dielectric for Isolation between Components and For Structural Support Not Shown)

3 S3DC FABRIC-LEVEL THERMAL MANAGEMENT SUPPORT

In this section, we provide a briefly overview the thermal management challenges in 3D IC directions including S3DC. Then we propose thermal management fabric components that provide fabric-level support for fine-grained thermal management of S3DC circuits.

3.1 Thermal Management Challenges in 3D IC

Heat is generated on chip mainly due to the electron-phonon scattering in transistors [23]. When current flows through a transistor, charged carriers are accelerated by the electric field across source and drain. They interact with silicon lattice vibration (phonons), exchange energy with the lattice and increase the lattice temperature. The main heat generation location is near the drain side of a transistor, since that is where electric field peaks. The generated heat raises the local temperature and forms a temperature gradient that drives heat to flow towards neighboring cooler regions, and eventually dissipate through the substrate and heat sink.

Thermal management is a critical challenge for all 3D IC directions for two reasons. On one hand, the power density of 3D ICs is usually much higher than for planar CMOS circuits, leading to higher generated heat per unit area. For example, power density in the 4-tier monolithic 3D ICs can be up to 3.4X higher compared with 2D ICs [24]. In S3DC, although S3DC designs are much more energy efficient vs. 2D CMOS, the power density of S3DC circuits can still be 3.2-3.5X higher due to the ultra-high-density of S3DC designs [13]. On the other hand, heat dissipation in 3D ICs is more constrained due to the longer heat dissipation paths from the top-tier circuits through bottom tiers to substrate. As Fig. 4 shows, in S3DC, the heat generated in the top layers of transistors flow through bottom layers to the substrate through the vertical silicon nanowires. This heat path is inefficient since a silicon nanowire has reduced thermal conductance due to its confined geometry [25] [26].

3.2 Thermal Management Components in 3D IC

We have developed intrinsic fabric-level support to enable the fine-grained thermal management in S3DC. We have designed several thermal management fabric components, including the Thermal Contact, Thermal Bridge, and Thermal Pillar. They are inserted as a part of S3DC circuit physical designs through unified electrical and thermal designs. These components are shown in Fig. 5.

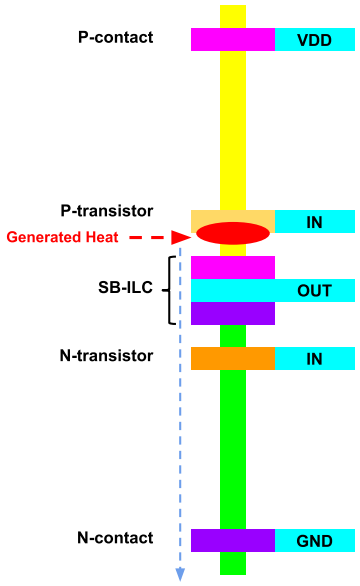

Figure 4: S3DC Inverter Layout Sideview and Heat Dissipation (when Output is 1 and only P-type Transistor is ON)

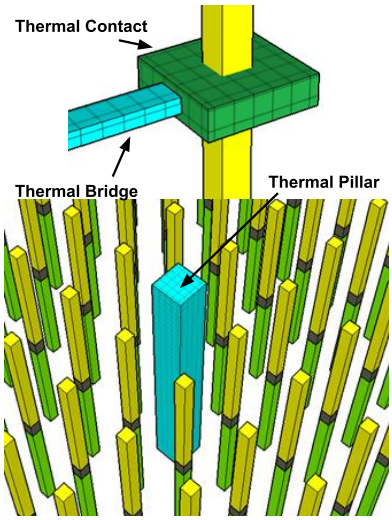

Figure 5: Thermal Components in S3DC (Thermal Structures Shown in Grided-Shapes)

(i). Thermal Contacts (Fig. 5(A)) are specialized junctions that are designed for extracting heat from hotspots on logic nanowires. This achieves efficient thermal extraction without interfering with the electrical operation of circuits.

(ii). Thermal Bridge (Fig. 5(A)) connects a Thermal Contact on one end and a Thermal Pillar on the other. It is routed in the metal layer similarly as a signal Bridge through routing tracks. It conveys heat flow from Thermal Contacts to Thermal Pillars in the lateral direction.

(iii). Thermal Pillars (Fig. 5(B)) are vertical metal pillars that are larger in cross-section area than vertical silicon nanowires, and thus have lower thermal resistance and provide good heat dissipating paths down to the substrate in vertical direction. They are inserted in the gaps between placed S3DC gates and are sparsely located on chip. Each pillar can be connected to several Thermal Bridges extracting heat from different S3DC gates. In addition to thermal management, these pillars also serve as part of the Power Delivery Network (PDN) and are connected to the VDD power rails [27].

Fig. 6 shows the added heat dissipation paths formed by inserted thermal components in a thermal-aware S3DC inverter design connected to a Thermal Pillar. The primary heat dissipation path is through Thermal Contacts, Thermal Bridge, Thermal Pillar and finally the substrate. In the thermal domain, to dissipate heat efficiently, all the components in this path need to be thermally conductive. On the other hand, in the electrical domain, the added thermal components should not interfere with the electrical operation of a circuit, which requires good isolation between a logic nanowire and corresponding Thermal Bridge. Consequently, the electrical conductance and the capacitance across a Thermal Contact should be kept minimal.

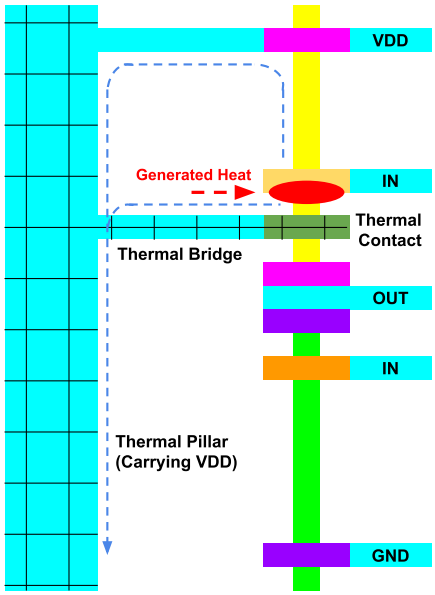

Figure 6: S3DC Inverter Layout Sideview (Thermal Components Shown in Gridded Shapes) and Its Heat Dissipation Paths through Inserted Thermal Components (when P-type Transistor is ON)

We design the thermal components such that they meet the afore-mentioned thermal and electrical requirements. We choose Tungsten as the Thermal Pillar and Thermal Bridge material. On one hand, Tungsten

has superior thermal conductivity (167 $Wm^{-1}K^{-1}$). On the other hand, by sharing the same material type with other fabric components (signal Bridge, Coaxial Routing structure), manufacturing complexity is reduced. The material type of Thermal Contacts is chosen as Al_2O_3 since it is both electrically insulating and thermally conductive.

To validate that the designed Thermal Contact meets the electrical requirements, we have performed TCAD simulation to characterize the resistance / conductance and capacitance through Thermal Contacts. TCAD process simulation is performed to simulate the process steps for building the Thermal Bridge – Thermal Contact – logic nanowire structure. TCAD device simulation achieves the IV and CV characteristics of the structure. The maximum current flowing through Thermal Contact is 2.1E-19A, which is at least 6 orders of magnitude lower than the state-of-art transistor leakage current. The added parasitic capacitance on the logic nanowire due to the Thermal Contact insertion is 4.5E-18F, which is more than one order of magnitude smaller than the minimum gate capacitance of state-of-art transistors. These results prove that the inserted thermal components pose negligible influence on the electrical properties of S3DC circuits.

4 S3DC THERMAL-AWARE PHYSICAL DESIGN

In this section, we briefly present how our proposed thermal evaluation methodology supports the thermal-aware design flow. The flow targets to optimize the over-heated cells leveraging the concepts of thermal fabric components and thermal-aware standard cell library. The list of over-heated cells is generated during thermal evaluation stage, which is introduced in more details in Section 5 and 6.

The thermal components are inserted together with the Place & Route of the electrical circuits in a unified flow as shown in Fig. 7. The flow takes the post-placement design, performs thermal evaluation to achieve the cell temperatures across the design, and generates the list of cells that needs to be thermally optimized. For these over-heated cells, we insert Thermal Contacts at the hotspots of the circuit and connect them to inserted Thermal Pillars through Thermal Bridges to improve their heat dissipation. The Thermal Pillar locations need to be optimized in order to lower the impact on routing. Then the physical design files are updated to include the implemented thermal features. The Place & Route tool restores the updated design and continues with the remaining physical design steps.

5 STEADY-STATE THERMAL EVALUATION

In this section, we present our material-to-system evaluation methodology which yields the hotspot temperature of all cells in a S3DC physical design. We build a thermal resistance network that models the steady-state heat generation and dissipation in a bottom-up manner for large-scale S3DC circuit designs. We then solve the thermal network through simulation to obtain the hotspot temperature of each standard cell in the design.

5.1 Steady-State Thermal Analysis

To calculate the steady-state temperatures in a circuit layout, we need to estimate the heat generation and dissipation in steady-state. For heat generation, we need to estimate the generated heat at all transistor drains. For heat dissipation, as the heat flows through structures, we can apply Fourier's law of heat conduction to calculate the temperature drop across the structure with:

$$\Delta Temp = R_{thermal} * Q_{flow} \quad (1)$$

where $R_{thermal}$ is the thermal resistance of the structure that heat flows through, $\Delta Temp$ is the temperature difference across the structure, and Q_{flow} is the amount of heat flow through the structure. Thermal resistance of the heat conductor depends on the material property and geometry, and can be calculated with:

$$R_{thermal} = \frac{L}{K*A} \quad (2)$$

where L is the length of the heat conductor, K represents the material thermal conductivity, A is the cross-section area. These thermal resistors are interconnected to form a thermal resistance network according to how the heat-conducting structures are attached to each other in the circuit layout. Example thermal networks of S3DC fabric components will be shown later.

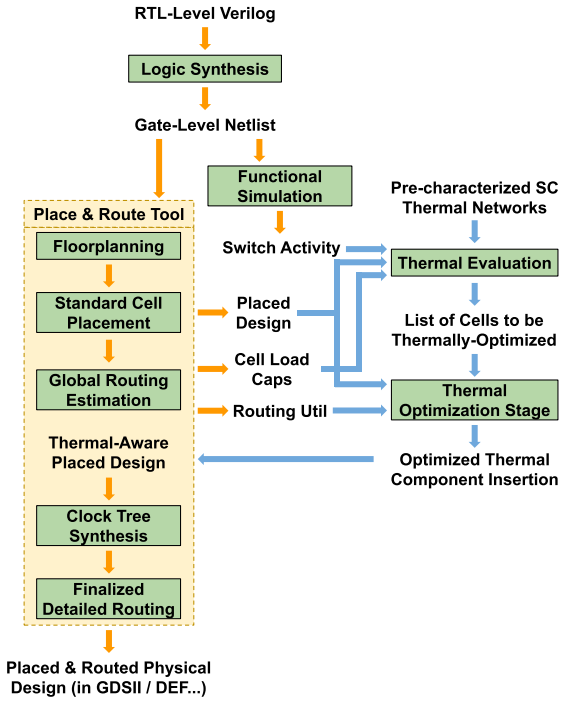

Figure 7: Thermal-aware Automated Circuit Design (SC: Standard Cell)

The temperatures in the thermal resistance network can be solved similarly to solving the voltages in an electrical resistance network, since the concepts in thermal domain are analogous to those in the electrical domain. Heat flow, temperature, and thermal resistance are analogous to current flow, voltage, and electrical resistance, respectively. In an electrical resistance network, if all the electrical resistance values of resistors and current values of current sources are known, and we assume a GND at a reference node in the circuit, we are able to solve the voltage at all nodes by applying Ohm's law and Kirchhoff's Voltage and Current Laws. Similarly, in a thermal resistance network, if all the thermal resistance values and the generated heat in transistors are calculated, we are able to calculate the temperatures of all the nodes after assuming an ambient temperature at the substrate or heat sink. Table 1 summaries these analogous concepts and principles.

Table 1: Analogous concepts in thermal vs. electrical domain analysis

Thermal	Electrical
generated heat at transistor drain	current source
temperature gradient ($\Delta Temp$)	potential difference
thermal resistance ($R_{thermal}$)	electrical resistance
heat flow through heat conductor (Q_{flow})	current flow through electrical resistor
Fourer's law of heat conduction	Ohm's law

5.2 Thermal Resistance Network of S3DC Fabric Components

Fig. 8(A) shows the structure of an S3DC transistor and its thermal resistance network. To build the network, we look at the structure of the transistor, divide it into elements, calculate the thermal resistance for the elements, and connect them according to how they are attached to each other in the fabric component structure. Note that nanoscale thermal effects are captured with calibrated thermal conductivity parameter K values. For example, K is 147 $Wm^{-1}K^{-1}$ in bulk silicon and only 13 $Wm^{-1}K^{-1}$ in thin silicon layer like nanowires [25] [26]. We add a source injecting heat at the drain side of the transistor to represent the generated heat.

We have built the thermal resistance network for all the other fabric components as well, including vertical nanowire, Ohmic contact, Coaxial Routing structure and Skybridge-Interlayer-Connection. Fig. 8(B)-(D) shows three types of interconnection structures and their thermal resistance network as examples.

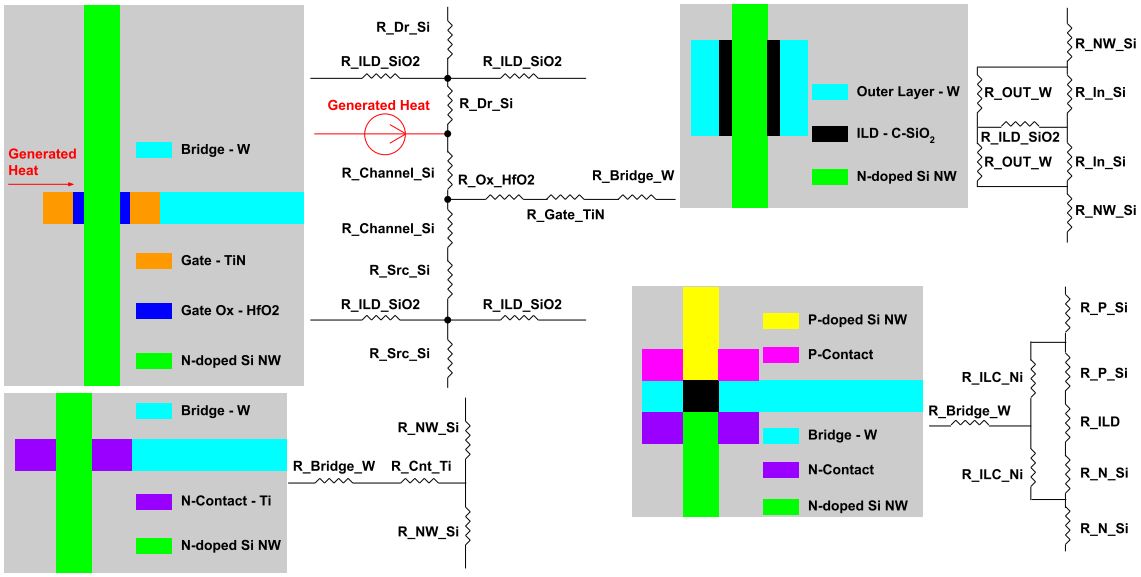

Figure 8: Thermal Networks of S3DC Fabric Components: Upper Left: N-type Transistor; Upper Right: Coaxial Routing Structure; Bottom Left: Ohmic Contact; Bottom Right: Interlayer Connection

5.3 Thermal Evaluation of S3DC Gates

With all the thermal networks built for S3DC fabric components, we can assemble the thermal networks for S3DC standard cells. We need to estimate the generated heat for each standard cell based on its actual power

consumption in the circuits. We look at the load capacitance (C_L) of the standard cells, and estimate the energy converted to heat per switching as $0.5*C_LVDD^2$ [28]. Also, to estimate the number of switches per unit time, we multiply circuit operating frequency and switching activity (SA) - the probability of an output being switched in each cycle. Generated heat can then be estimated by multiplying energy converted to heat per switch and number of switches per unit time, i.e. $0.5*C_LVDD^2*SA*freq$. C_L and SA needs to be obtained for each standard cell in the circuit.

We have manually built the thermal resistance networks for all the standard cells in S3DC library, including both baseline layout designs and thermally-optimized designs, *i.e.*, layouts with Thermal Contacts inserted at hotspots. SPICE simulations were performed on these networks; we measure temperatures on all nodes, and generate the hotspot temperature for each gate. Table 2 shows the hotspot temperatures in several example S3DC standard cells with baseline vs. thermally-optimized designs. As we can see, in the baseline design without inserted thermal components, the hotspot temperature can be as high as 526K, which is far above the industrial maximum allowed temperature of 398K. In the thermally-optimized design the hotspot temperatures are lowered to 334-344K.

Table 2: Hotspot temperatures in S3DC standard cells

	Inverter	2-in NAND	3-in NAND	2-in NOR	AOI21	AOI22
Not Thermally-Optimized	526K	499K	478K	505K	512K	501K
Thermally-Optimized	334K	339K	342K	341K	341K	344K

5.4 Thermal Evaluation of S3DC Large-Scale Circuits

The next step is to perform thermal evaluation for standard-cell-based large-scale S3DC circuits. Fig. 9 shows the thermal evaluation flow for a post-placement design to generate a thermal profile that guides the later thermal-aware design steps. The thermal evaluation block (in light blue color) extracts circuit information from CAD files, creates the thermal resistance network, and solves the network through SPICE simulation to obtain the temperatures of all standard cells. The intra-cell networks are from the manually-built pre-characterized thermal networks for each cell. The generated heat of each standard cell is estimated according to its load capacitance C_L (parasitic RC estimated based on Global Routing performed by the Place & Route tool), switching activity (from switching activity propagation or functional simulation results), and the estimated best frequency of the circuit. The inter-cell heat paths are estimated based on the information on how cells are interconnected and placed. This information is achieved from the gate-level netlist and the physical design descriptions (DEF file in this work) generated by the Place & Route tool. After the thermal resistance network is built, we perform SPICE simulation to solve the network, measure the temperatures of all nodes in each standard cell, and get the hotspot temperature of all cells.

We have performed thermal evaluation for 10 benchmarks including S13207, S38584, B14, B22, SPI, TV80N, WB_DMA, SHA1, SYSTEMCAES, and GNG [29] [30] [31]. Fig. 10 shows the temperature distribution across all designs. As we can see, despite most cells operating at a low temperature, there are still significant number of cells with high temperatures, which necessitates a thermal-aware design flow. The highest temperature is found to be 586K, exceeding the industrial maximum allowed temperature by 188K.

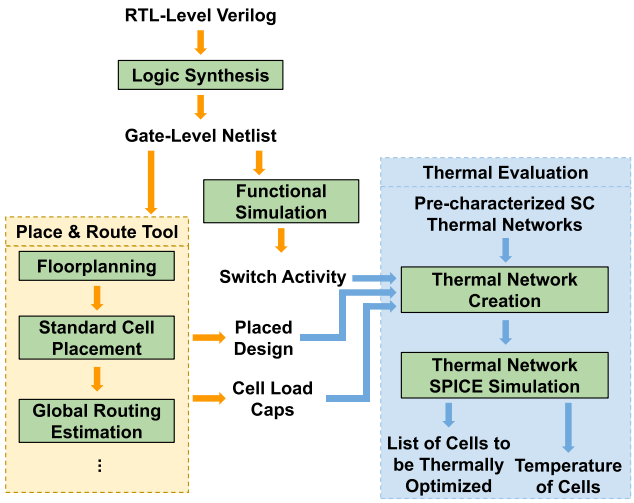

Figure 9: Thermal Evaluation Flow for Post-Placement Design based on HSPICE Simulation of Thermal Resistance Network (SC: Standard Cell)

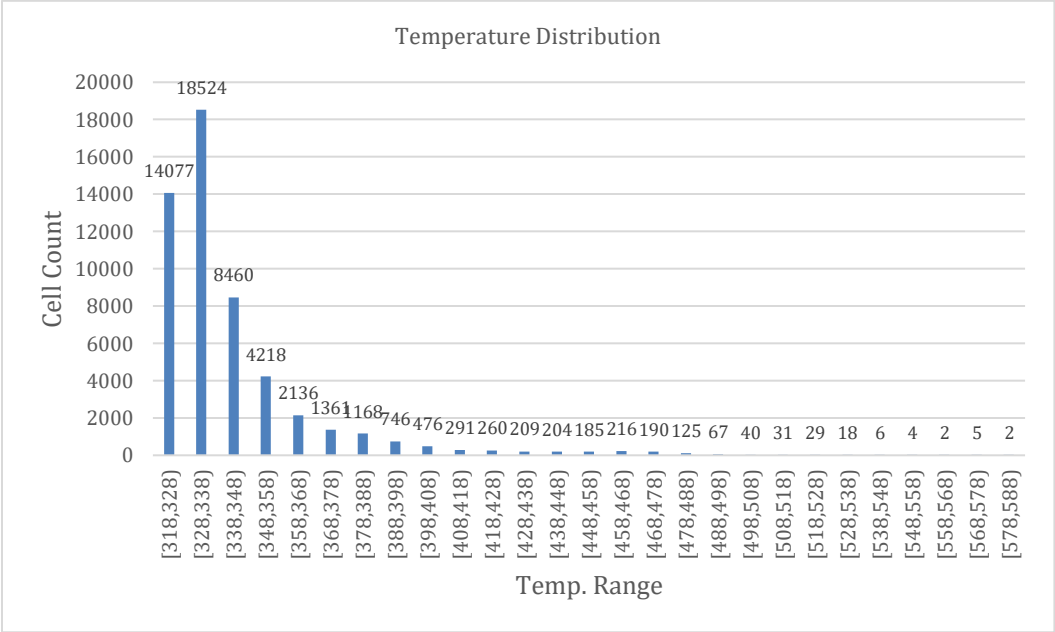

Figure 10: Temperature Distribution of Standard Cells in 10 Tested Benchmarks

Fig. 11 shows the thermal map of WB_DMA design. As we can see, one distinction of S3DC's thermal profile is that cell temperatures change abruptly across neighboring cells. This is due to the lack of thermal coupling between adjacent cells, as they are located on different nanowires that are separated by thermally-insulating dielectric materials. It shows that current thermal optimization techniques targeting at optimizing large areas of the design are not fine-grained enough to optimize over-heated standard cells in S3DC.

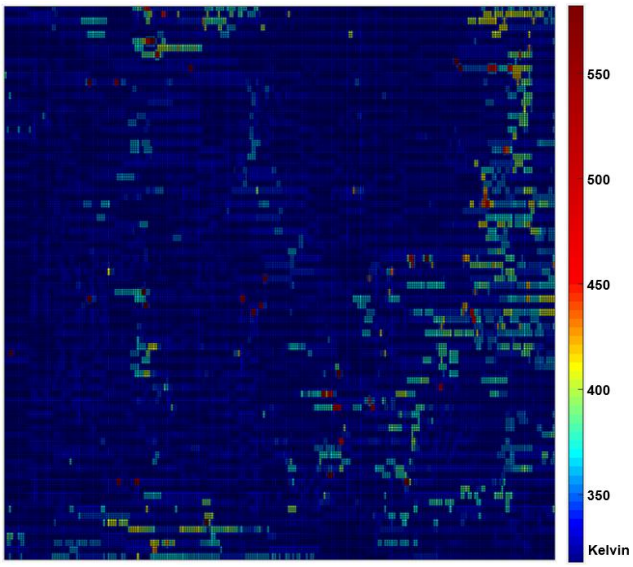

Figure 11: Thermal Map of WB_DMA Benchmark

One challenge of the presented thermal evaluation method is its large runtime overhead. The SPICE simulation of the thermal-resistance networks for large designs takes very long runtime, which is due to the large and complex inter-connected thermal resistance network. As an example, the SPICE simulation of B22 (contains 12,887 cells) thermal resistance network takes 127 seconds, which is 2.3X longer than its placement runtime.

6 MACHINE LEARNING BASED THERMAL MODEL

The thermal evaluation method based on solving thermal resistance network using SPICE simulation becomes intractable as the design size increases. In this section, we present another method to achieve the cell hotspot temperatures with better scalability. We have developed a thermal model that predicts hotspot temperatures of each cell using machine-learning regression method. The developed thermal model takes the circuit information extracted statically for each cell in the S3DC circuits as input features, and predicts the hotspot temperature of each cell efficiently without running SPICE simulation.

6.1 Training Procedure of Thermal Model

To train the thermal model, we need to obtain the training data set with input features and targets of the thermal model: input features are statically extracted circuit information of each cell, and targets are hotspot temperatures of each cell. We obtain the training data set from 10 benchmarking results using our simulation-based thermal evaluation flow that was introduced in Section 5. These benchmarks contain 53050 cells in total. Each cell is one sample in the data set. Extracted circuit information that correlates to the hotspot temperature in one cell includes estimated heat generation of this cell, cell type (Inverter, NAND2, NAND3, NOR2, AOI21, AOI22, Buffer, D Flip Flop), estimated heat generation of its neighboring cells, nearby empty nanowires, and so on. These information are the candidate input features to the thermal model. We have performed analysis to

achieve good accuracy while minimizing the dimension of input features. Table 3 shows a list of the selected input features and provides intuitions on how these features affect the cell temperature. Fig. 12 shows the flow of training and thermal model.

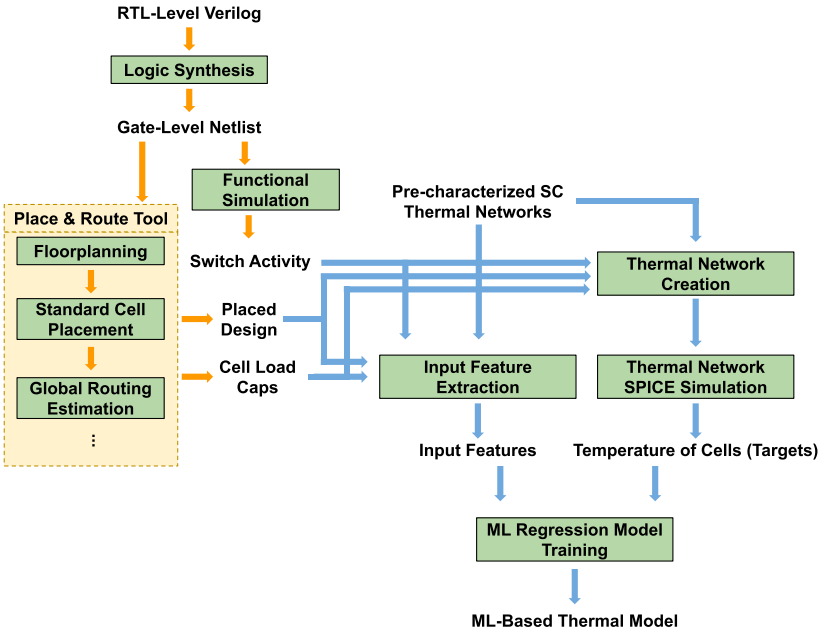

Figure 12: ML-Based Thermal Model Training Flow (SC: Standard Cell)

Table 3: Selected Input Features of Thermal Model to Determine a Cell Hotspot Temperature

Input Feature	**Impact on Cell Temperature**
Cell Type	determines intra-cell thermal network
Estimated Heat Gen of This Cell	determines the heat generated inside this cell that needs to be dissipated
Neighboring Routing NW Count	adjacent routing NW helps w/ lower cell temperature because: 1). no transistor on routing NW -> no generated heat from routing NW 2). added heat path from this cell through dielectric to adjacent routing NWs to substrate
Distance to Nearest Thermal Pillar	determines length (and thus thermal resistivity) of heat path from cell through Thermal Bridge to Thermal Pillar
Avg Estimated Heat Gen in Cells that Are Connected to This Cell (with Various Levels of Distance)	affects temperature of this cell due to thermal coupling through signal Bridges connecting in between; shorter signal Bridge length means stronger thermal coupling
Avg Estimated Heat Gen in Neighboring Cells (with Various Levels of Distance)	affects temperature of this cell due to thermal coupling through dielectric between NWs; nearer cells have stronger thermal coupling
Ratio of Total Thermal Pillar Count to Total Design Area	determines the density of heat paths from PDN through Thermal Pillar to substrate; higher density leads to lower overall temperature in the design

We split the data into 80% training and 20% test sets. During training we apply a 5-fold cross validation with the training set. The testing set is reserved only for testing purpose to evaluate if over-fitting has occurred. We have multiple choices of regression models, including support vector regressor (based on radial basis function and polynomial kernel function), gaussian process regressor, random forest regressor, nearest neighbor regressor, voting, multi-layer perceptron, as well as gradient boosting models including XGBoost and AdaBoostRegressor. We have tested these models, and have achieved best accuracy with the XGBoost model as Fig. 13 shows. Thus, XGBoost is selected for developing our thermal model.

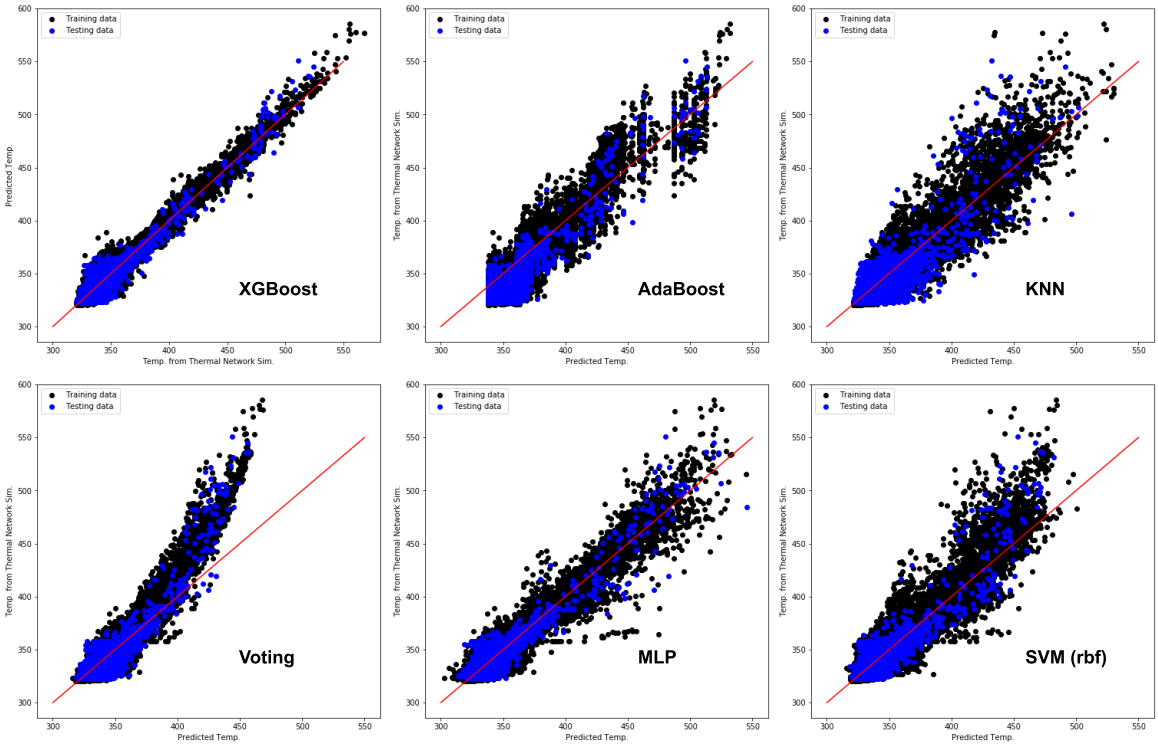

Figure 13: Scattered Plot from Different Machine Learning Models: Y: Predicted Temperatures, X: Temperatures from Thermal-Network-Simulation-Based Thermal Evaluation (Blue: Testing Data, Black: Training Data)

Fig. 14 shows the flow of cell hotspot temperature estimation based on our machine learning based thermal model. The flow requires the same input CAD files with the simulation-based thermal evaluation flow, including post-placement physical design, pre-extracted thermal networks for standard cells, switching activity, and load capacitance of each cell. These inputs provide all the information that is needed to statically extract the input features of the thermal model. The thermal model predicts the hotspot temperature for each cell based on the input features. The predicted cell temperatures can be used to determine which cells are over-heated and thus need to adopt their thermally-optimized version of layout in the thermal-aware design flow.

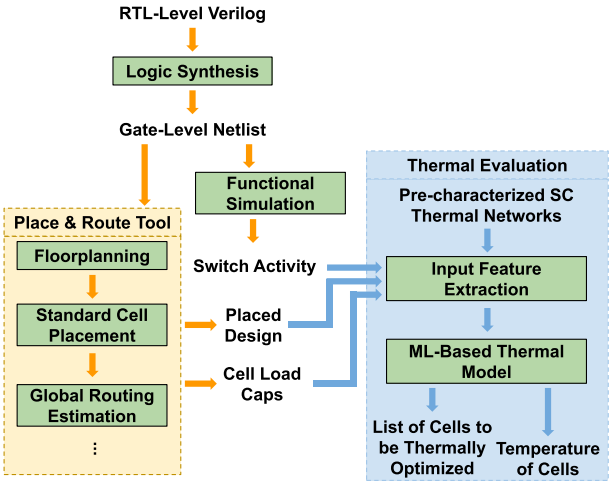

Figure 14: Thermal Evaluation Flow with Machine-Learning-Based Thermal Model (SC: Standard Cell)

6.2 Standard Cell Temperature Prediction Results

Compared with the simulation-based thermal evaluation method, using this machine learning based thermal model to estimate cell temperatures takes much shorter runtime. Among the large benchmarks we have predicted cell temperatures using our developed thermal model, the thermal model runtime is less than 0.04% of the placement runtime of the benchmark. In addition to the 10 afore-mentioned benchmarks, we have implemented more larger-scale circuits and used our developed thermal model to predict the cell temperatures. The largest benchmark we have implemented - the SPARC T2 core containing 435K cells [32], takes 3.12 second for predicting all cell temperatures, which is only 0.17% of the runtime for performing cell placement for the design.

Fig. 15 shows the thermal maps of WB_DMA benchmark circuits generated by simulation-based thermal evaluation method and machine learning based thermal model, showing good agreement. The predictions from trained model have a mean absolute error of 3.55K in training set and 3.61K in testing set, proving very minor overfitting in our trained thermal model. The error distribution is shown in Fig. 16.

Fig. 17 shows the trend of temperature prediction accuracy across benchmarks with different design sizes. The observation is that the accuracy is generally better in larger designs. The reason is that in our case, the accuracy of cells closer to boundaries are not as good as those away from boundaries. In larger designs, more cells are further away from boundaries, leading to better overall accuracy in larger designs. Consequently, we can train the thermal model based on the data from a number of smaller designs with similar technology assumption and be able to predict the cell temperatures for a larger design

6.3 Dealing with Prediction Errors

The standard cell temperatures generated by thermal model are compared with the industrial maximum allowed temperature (398K) to determine if a cell is over-heated. Based on the cell temperatures achieved from thermal network simulation, we classify all the standard cells into two sets, simulated negative or simulated positive. Similarly, the cell temperatures predicted by machine-learning thermal model divide all the standard

cells into two sets, predicted negative or predicted positive. In the problem that targets to flag over-heated cells, the True Positive, True Negative, False Positive, False Negative are visualized by Fig. 18.

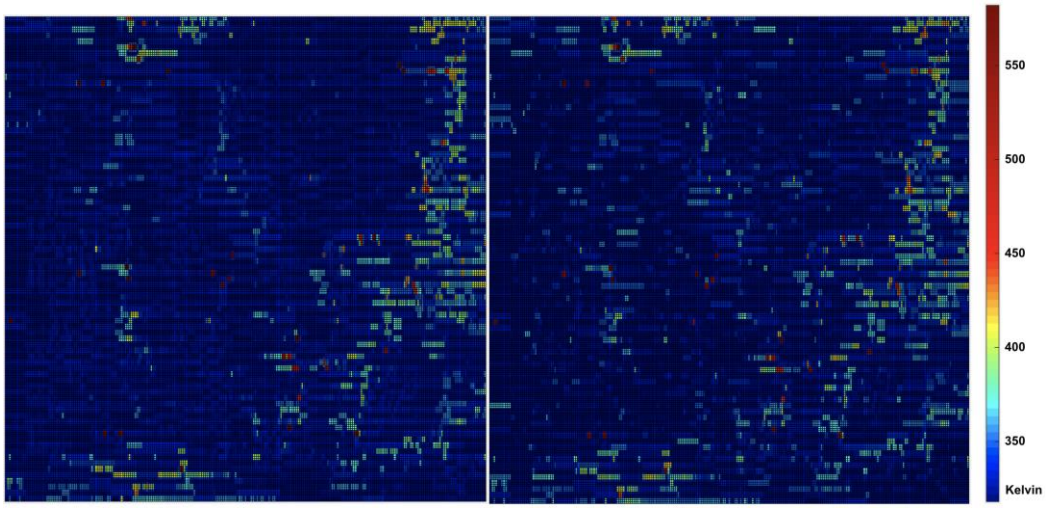

Figure 15: Thermal Map from HSPICE Thermal Network Simulation (Left) and Machine-Learning-Based Thermal Model (Right)

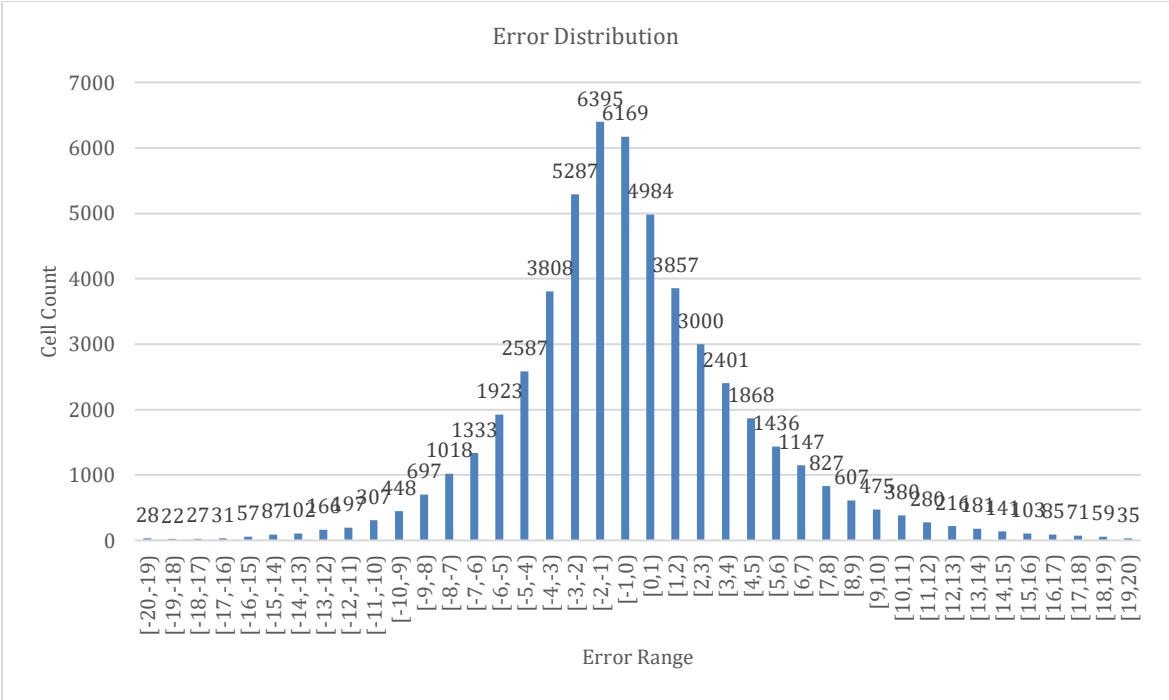

Figure 16: Error Distribution (X-axis: error range, Y-axis: number of cells in the error range)

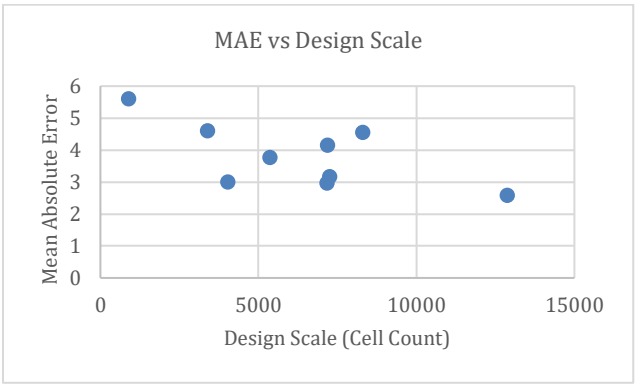

Figure 17: Mean Absolute Error vs Design Scale by Partitions

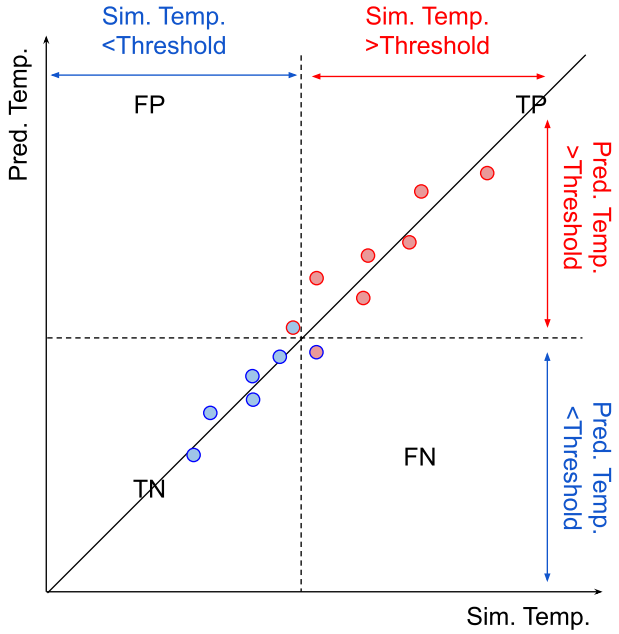

Figure 18: Visualizing True Positive (TP) / True Negative (TN) / False Positive (FP) / False Negative (FN). X axis: Temperature from Simulation Results, Y axis: Predicted Temperatures. Data Point Color Scheme: Blue for Below Threshold Temperature, Red for Above Threshold Temperature; Edge Color for Predicted Temperature, Filling Color for Simulated Temperature

Considering predicted cell temperatures guide the thermal-aware design flow, False Negatives are more catastrophic compared to False Positives as is explained here. False Positive causes over-design by making a standard cell that is below the threshold temperature being unnecessarily adopting its thermally-optimized layout, leading to over-design in thermal domain. False Negative, on the other head, leaves over-heated cells not being thermally optimized, causing thermal issues in the design and thus leading to potential reliability issue, timing degradation and even functional failures. Therefore during the training of thermal model we need to spend more effort on eliminating False Negatives than False Positives.

To make sure these False Negative cases are addressed in presence of the prediction errors in our thermal model, one straightforward method is to introduce additional guard band by applying a static margin below the original threshold temperature. In this way we cover more False Negatives at the cost of introducing more False Positives. The concepts are visualized in Fig. 19. In our tested benchmarks, a 10.7K margin is needed to address all the False Negatives, meaning all cells with predicted hotspot temperatures above 387.3K would be targeted for thermal optimization during the thermal-aware design flow. In this case, the portion of False Positive among all standard cells is 1.48%.

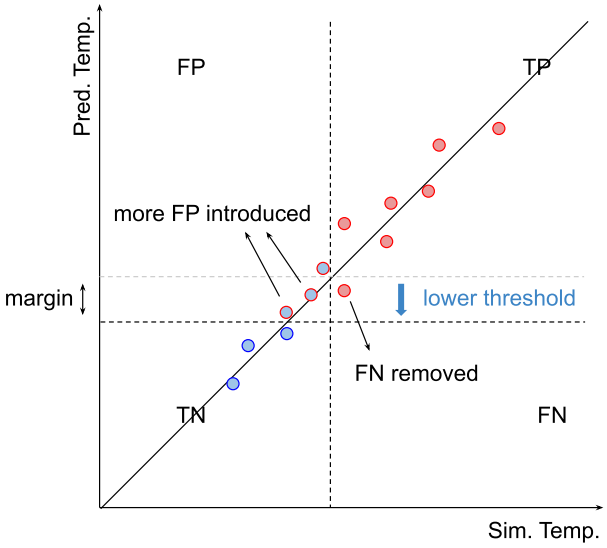

Figure 19: Addressing False Negative with Additional Margin. By Applying the Additional Margin, We Classify if a Cell is Predicted Positive or Predicted Negative based on a New Lower Threshold, Leading to More Eliminated False Negatives at the Cost of Introducing Additional False Positives

The other way to eliminate the False Negative is to customize the loss function – the function that indicates the inaccuracy and is minimized during the training of the machine-learning model. We can customize the loss function to make it asymmetric such that under-estimating cell temperatures is penalized more during training. We have experimented different asymmetric loss functions that scales the loss function by 5X-100X when it is under-estimation. Fig. 20 shows the scattered plots of predicted temperatures vs simulated temperatures when scaling the loss function by 40X when under-estimating. As we can see, the under-estimation of cell temperatures has been significantly suppressed compared to the default loss function, making it much easier to eliminate the False Negative. In the case of scaling up loss function for under-estimation by 40X, a 2.2K static margin can eliminate all the False Negatives, while the portion of False Positive cells is only 0.85%

7 RESULTS AND DISCUSSION

S3DC technology is a promising vertically-composed fine-grained 3D technology that achieves significant benefits in PPA while maintaining routability. It features the fine-grained 3D thermal management framework with fabric-level support. Our developed machine learning based thermal estimation method provides fast and accurate feedback on the thermal profile of current design and guides the unified automated co-design of

thermal and electrical circuits. The thermal evaluation runtime is only up to 0.04% of placement runtime, posing negligible runtime overhead to the physical design flow. It is accurate enough to identify all over-heated cells with a small 2.2K static margins, which leads to minor over-design in thermal domain on 0.85% of the standard cells. The proposed thermal evaluation method is also potentially applicable to other 3D technologies and promising to support their thermal-aware design.

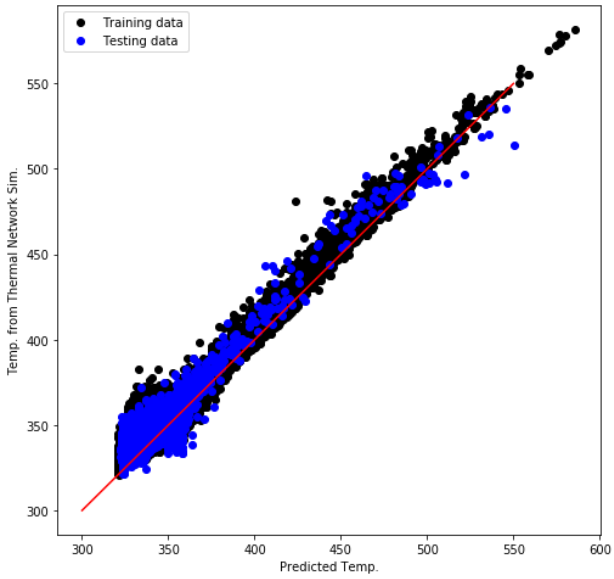

Figure 20: Scattered Plot: Y: Predicted Temperatures, X: Golden Temperatures from Custom Asymmetric Custom Function that Penalize More on Under-Estimation (Blue: Testing Data, Black: Training Data)

REFERENCES

[1] J. A. Burns, B. F. Aull, C. K. Chen, Chang-Lee Chen, C. L. Keast, J. M. Knecht, V. Suntharalingam, K. Warner, P. W. Wyatt, and D.-R. W. Yost. 2006. A wafer-scale 3-D circuit integration technology. In *Proceedings of IEEE Trans. on Electron Devices*, 53, 10 (2006), 2507-2516.

[2] J. Van Olmen, A. Mercha, G. Katti, C. Huyghebaert, J. Van Aelst, E. Seppala, Z. Chao, S. Armini, J. Vaes, R. C. Teixeira, M. Van Cauwenberghe, P. Verdonck, K. Verhemeldonck, A. Jourdain,W. Ruythooren, M. de Potter de ten Broeck, A. Opdebeeck, T. Chiarella, B. Parvais, I. Debusschere, T. Y. Hoffmann, B. De Wachter, W. Dehaene, M. Stucchi, M. Rakowski, P. Soussan, R. Cartuyvels, E. Beyne, S. Biesemans, and B. Swinnen. 2008. 3D Stacked IC Demonstration using a Through Silicon Via First Approach. In *Proceedings of IEEE Int. Electron Devices Meeting*, 1-4.

[3] M. Motoyoshi. 2009. Through-Silicon Via (TSV). In *Proceedings of IEEE*, 97, 1 (2009), 1-4.

[4] P. Batude, M. Vinet, A. Pouydebasque, C. Le Royer, B. Previtali, C. Tabone, J.-M. Hartmann, L. Sanchez, L. Baud, V. Carron, A. Toffoli, F. Allain, V. Mazzocchi, D. Lafond, O. Thomas, O. Cueto, N. Bouzaida, D. Fleury, A. Amara, S. Deleonibus, and O. Faynot. 2009. Advances in 3D CMOS Sequential Integration. In *Proceedings of IEEE Int. Electron Devices Meeting*, 1-4.

[5] Y.-J. Lee, P. Morrow, and S. K. Lim. 2012. Ultra High Density Logic Designs Using Transistor-Level Monolithic 3D Integration. In *Proceedings of IEEE/ACM Int. Conf. on Comput.-Aided Design*, 539-546.

[6] M. S. Ebrahimi, G. Hills, M. M. Sabry, M. M. Shulaker, H. Wei, T. F. Wu, S. Mitra, and H.-S. Philip Wong. 2014. Monolithic 3D integration advances and challenges: From technology to system levels. In *Proceedings of SOI-3D-Subthreshold Microelectronics Technology Unified Conf.*, 1-2.

[7] M. M. Shulaker, T. F. Wu, A. Pal, L. Zhao, Y. Nishi, K. Saraswat, H.-S. Philip Wong, and S. Mitra. 2014. Monolithic 3D integration of logic and memory: Carbon nanotube FETs, resistive RAM, and silicon FETs. In *Proceedings of IEEE Int. Electron Devices Meeting*, 27.4.1-27.4.4.

[8] Y.-J. Lee, D. Limbrick, and S. K. Lim. 2013. Power benefit study for ultra-high density transistor-level monolithic 3D ICs. In *Proceedings of Design Automation Conf.*, 1-10.

[9] A. Todri, S. Kundu, P. Girard, A. Bosio, L. Dilillo, and A. Virazel. 2013. A Study of Tapered 3-D TSVs for Power and Thermal Integrity. In *IEEE Trans. on Very Large Scale Integration Syst*., 21, 2 (2013), 306-319.

[10] S. K. Samal, S. Panth, K. Samadi, M. Saedi, Y. Du, and S. K. Lim. 2014. Fast and Accurate Thermal Modeling and Optimization for Monolithic 3D ICs. In *Proceedings of ACM/EDAC/IEEE Design Autom. Conf*., 1-6.

[11] M. Li, J. Shi, M. Rahman, S. Khasanvis, S. Bhat, and C. A. Moritz. 2016. Skybridge-3D-CMOS: A Vertically-Composed Fine-Grained 3D CMOS Integrated Circuit Technology. In *IEEE Comput. Soc. Annu. Symp. on VLSI*, 403-408.

[12] J. Shi, M. Li, S. Khasanvis, M. Rahman, and C. A. Moritz. 2016. Routability in 3D IC Design: Monolithic 3D vs. Skybridge 3D CMOS. In *IEEE/ACM Int. Symp. on Nanoscale Architectures*, 145-150.

[13] M. Li, J. Shi, M. Rahman, S. Khasanvis, S. Bhat, and C. A. Moritz. 2017. Skybridge-3D-CMOS: A Fine-Grained 3D CMOS Integrated Circuit Technology. In *IEEE Trans. Nanotechnol*., 16, 4 (2017), 639-652.

[14] M. Li, J. Shi, M. Rahman, S. Khasanvis, S. Bhat, and C. A. Moritz. 2017. Vertically-composed fine-grained 3D CMOS. In *IEEE SOI-3D-Subthreshold Microelectronics Technol. Unified Conf*., 1-2.

[15] M. Li, S. Khasanvis, J. Shi, S. Bhat, M. Rahman, and C. A. Moritz. 2016. Towards automatic thermal network extraction in 3D ICs. In *Proceedings of IEEE/ACM Int. Symp. on Nanoscale Architectures*, 25-30.

[16] C. Santos, P. Vivet, S. Thuries, O. Billoint, J.-P. Colonna, P. Coudrain, and L. Wang. 2016. Thermal performance of CoolCube™ monolithic and TSV-based 3D integration processes. In *IEEE Int. 3D Syst. Integr. Conf*., 1-5.

[17] B. Goplen, and S.S. Sapatnekar. 2006. Placement of thermal vias in 3-D ICs using various thermal objectives. In *IEEE Trans. on Comput.-Aided Design Integr. Circuits Syst*., 25, 4 (2006), 692-709.

[18] Y.-J. Lee, and S. K. Lim. 2011. Co-Optimization and Analysis of Signal, Power, and Thermal Interconnects in 3-D ICs. In *IEEE Trans. on Comput.-Aided Design Integr. Circuits Syst*., 30, 11 (2006), 1635-1648.

[19] W. Liu, A. Calimera, A. Macii, E. Macii, A. Nannarelli, and M. Poncino. 2013. Layout-Driven Post-Placement Techniques for Temperature Reduction and Thermal Gradient Minimization. In *IEEE Trans. on Comput.-Aided Design Integr. Circuits Syst*., 32, 3 (2013), 406-418.

[20] G. Luo, Y. Shi, and J. Cong. 2013. An Analytical Placement Framework for 3-D ICs and Its Extension on Thermal Awareness. In *IEEE Trans. on Comput.-Aided Design Integr. Circuits Syst*., 32, 4 (2013), 510-523.

[21] S. K. Samal, S. Panth, K. S, M. Saeidi, Y. Du, and S. K. Lim. 2016. Adaptive Regression-Based Thermal Modeling and Optimization for Monolithic 3-D ICs. In *IEEE Trans. on Comput.-Aided Design Integr. Circuits Syst*., 35, 10 (2016), 1707-1720.

[22] M. Rahman, S. Khasanvis, J. Shi, M. Li, and C. A. Moritz. 2015. Architecting 3-D integrated circuit fabric with intrinsic thermal management features. In *Proceedings of IEEE/ACM Int. Symp. on Nanoscale Architectures*, 157-162.

[23] E. Pop, S. Sinha, and K. E. Goodson. 2006. Heat Generation and Transport in Nanometer-Scale Transistors. In *Proceedings of IEEE*, 94, 8 (2006), 1587-1601.

[24] K. M. Kim, S. Sinha, B. Cline, G. Yeric, and S. K. Lim. 2016. Four-tier Monolithic 3D ICs: Tier Partitioning Methodology and Power Benefit Study. In *Proceedings of ACM/IEEE Int. Symp. on Low Power Electron. Design*, 70-75.

[25] D. Li, Y. Wu, P. Kim, L. Shi, P. Yang, and A. Majumdar. 2003. Thermal conductivity of individual silicon nanowires. In *Appl. Phys. Lett*., 83, 14 (2003), 2934-2936.

[26] E. Pop. 2010. Energy Dissipation and Transport in Nanoscale Devices. In *Nano Research*, 3, 3 (2010), 147-169.

[27] J. Shi, M. Li, and C. A. Moritz. 2017. Power-delivery network in 3D ICs: Monolithic 3D vs. Skybridge 3D CMOS. In *Proceedings of IEEE/ACM Int. Symp. on Nanoscale Architectures*, 73-78.

[28] J. Rabaey. 2009. *Low Power Design Essentials*. (2009) Boston, MA: Springer.

[29] OpenCores. http://opencores.org.

[30] F. Brglez, D. Bryan, and K. Kozminski. 1989. Combinational Profiles of Sequential Benchmark Circuits. In *IEEE Int.Symp. on Circuits and Systems*, 1929-1934.

[31] ITC99 Benchmark. https://www.cerc.utexas.edu/itc99-benchmarks/bench.html

[32] OpenSPARC T2. https://www.oracle.com/servers/technologies/opensparc-t2-page.html