Dynamic Power and Thermal Integrity in 3D Integration

Hao Yu and Lei He

Abstract— This paper presents a state-of-art advance in 3D integrations. It illustrates the need for a high-performance 3D design driven by dynamic power and thermal integrity. The through-silicon-via (TSV) is used to simultaneously deliver power supply and remove heat. More importantly, to cope with the large-scale design complexity, the modern macromodeling technique is applied to handle not only large numbers of dynamic inputs/working-loads but also large sizes of $RLC/RC$ networks distributing the power/heat. Experiments applying the 3D design presented by this paper showed promising results to reduce both runtime and resource.

I. INTRODUCTION

The high-performance VLSI integration by the technology scaling has been confronted with the dramatically increased design cost resulting from noise, power and process variation. Multi-core system integration has emerged as an alternative design paradigm to improve performance by increasing the throughput rate. However, in today’s two dimensional (2D) Systems-on-chip (SoC) integration, the memory is surrounded by logic circuits and its performance, in terms of memory bandwidth, is limited by the length of long interconnects. Thanks to the recent advance in three dimensional (3D) integration [1]–[9], a 3D integration can reduce the physical distance between the memory and the logic circuits and hence, has shown a promising potential to integrate hundreds of cores with scaled performance superior to that of the 2D integration.

Since there are large numbers of devices densely packed in a number of device layers, it brings a significant burden to the heat removal and power (supply voltage) delivery in 3D ICs. This paper discusses an allocation of through-silicon-via (TSV) to simultaneously consider both the dynamic power and thermal integrity in 3D ICs. In Section II, we first illustrate the need of dynamic power and thermal integrity in 3D design, and present a TSV allocation problem and the challenge to solve this problem. In Section III, we discuss how to apply the modern macromodeling technique to reduce design complexity, and present an efficient solution for our TSV allocation. We show the result in Section IV and conclude the paper in Section V.

II. HIGH-PERFORMANCE 3D DESIGN

A. Dynamic Power and Thermal Integrity

Fig. 1 illustrates a typical 3D stacking of multiple device layers within one package. The supply voltage is delivered from the bottom power/ground planes in the package, passed through the vias and C4 bumps, and connected to the on-chip power/ground grid on active device layers. We call the through vias that deliver the supply voltage the power/ground vias. The 3D integration, by definition, has integrated more than one layer of the active device. They draw much larger current from package power/ground planes than 2D ICs. This can obliviously result in the IR drop for the horizontal on-chip power/ground grid. The surge from the injecting current further leads to a large simultaneous switching noise (SSN) for those I/O drivers at the chip package interface. Fig. 2 shows a detailed view of how to place signal and power/ground vias through package planes. They form a number of different sized loop-inductances that have significant couplings with each other. We call the voltage bounce at I/Os power integrity in this paper.

On the other hand, due to increased power density and slow heat-convection at inter-layer dielectrics, heat dissipation is another concern in 3D ICs [1]. The excessively high temperature can significantly degrade the reliability and performance of interconnects and devices [1], [3], [4], [6]–[8], [10]–[12]. We call the temperature gradient at active device layers thermal integrity. As shown in Fig. 1, a heat-sink is placed at the top of device layers and it is the primary heat-removal path to the ambient air. One observation is that there are through vias delivering supply voltages or signals from the bottom package through each active device layer. Since the metal vias are good thermal conductors, the through vias can provide additional heat-removal paths passing the interlayer dielectrics to the top heat-sink. This leads to the concept of adding dummy thermal vias or thermal vias directly inside chips [11] to reduce effective thermal resistances. Its physical arrangement is further studied in [3], [4], [6]–[8].
In modern VLSI designs, dynamic power management such as clock-gating and uncertainty from the workload can lead to time-varying power inputs. This results in a spatially and temporally variant thermal model. The inputs are time-varying thermal power (See Fig. 3) [13], [14] defined by the running-average of the cycle-accurate (often in the range of $ns$) power over several thermal time constants (often in the range of $ms$). They are injected at input ports of each layer. As such, a temporally and spatially variant temperature at output ports can be considered by defining an integrity integral with respect to time and space [6]. As a result, the temperature gradient can have either a sharp-transition with a large peak value, or a time-accumulated impact on the device reliability. In addition, different regions can reach their worst-case temperature at different times.

A dynamic thermal-integrity constraint is thereby needed to accurately guide the physical level resource allocation. Since the active device layer at the bottom (See. Fig. 1) has the longest path to the heat-sink at the top, in this paper, a dynamic thermal integrity is defined as the integrated temperature fluctuation at $p_o$ output ports on the bottom device layer. As shown in [6], [7], [10]--[12], a dynamic thermal integrity can accurately capture not only the sharp-transition of temperature change due to dynamic power management, but also time-accumulated temperature impact that can affect device reliability. Similar to static thermal-integrity analysis, dynamic thermal integrity assumes the worst-case input from a limited number of thermal-power inputs. However, since the dynamic integrity has a more accurate transient temperature profile, it leads to a smaller allocation when compared to the static thermal-integrity based design [3], [4]. Note that the dynamic power-integrity has already been employed in many on-chip or off-chip power integrity verifications and designs [15]--[17]. A similar dynamic power integrity in this paper is defined as the time-integrated voltage bounce at power/ground I/Os, which are located on the interface between the bottom device layer and the package.

B. TSV Allocation Problem

We notice that previous thermal via allocations [3], [4], [6], [8] assume adding dummy vias to conduct heat. They ignore the fact that power/ground vias can help remove heat as well. Therefore, the reusing of the power/ground via as the thermal via can save the routing resource for signal nets. More importantly, the allocation of power/ground vias can minimize not only dynamic power integrity, i.e., voltage bounce for those I/Os at package and chip interface, but also thermal integrity, i.e., the temperature gradient at those active device layers.

Identical to [4], this paper assumes that the via allocation is after the placement and global routing but before the detailed routing of the signal nets. The power ground vias are placed at centers of tiles between two layers, and follow an aligned path from the bottom package I/Os to the top heat-sink. We call those aligned paths vertical tracks or tracks. As vias are aligned, the $p_o$ tracks pass both $p_o$ output ports of the electrical-RLC model and $p_o$ output ports of the thermal-RC model. The density of power/ground vias at each track is the primary design parameter considered in this paper. The density adjusts to satisfy two requirements at output ports. The first is the integrity constraint of the temperature gradient and voltage bounce. The second is the resource constraint with provided signal net congestion.

Accordingly, we have the following problem formulation: Formulation 1: Given the targeted voltage bounce $V_t$ for $p_o$ output ports at power/ground I/Os, and the targeted temperature gradient $T_t$ for $p_o$ output ports at bottom device layer, the via-allocation problem is to minimize the total via number, such that the temperature gradient $f^T$ is smaller than $T_t$ and the voltage bounce $f^V$ is smaller than $V_t$.

Such a via-allocation problem simultaneously driven by power and thermal integrity can be represented by

$$
\begin{align*}
\min & \sum_{j=1}^{p_o} n_j \\
\text{s.t.} & f^V \leq V_t, \ f^T \leq T_t \\
& n_{min} \leq n_j \leq n_{max}
\end{align*}
$$

Note that $n_j$ is the via density at the $j$th track and $V_t$ and $T_t$ are the targeted voltage bounce and temperature gradient. $f^V$ and $f^T$ are the metrics of power integrity and thermal integrity, defined by the spatially-averaged time-integral of the transient $V(t)$ [17] and $T(t)$ [6], respectively.

As discussed later in Section III, $n_j$ is decided according to the power and thermal sensitivities obtained from the

---

**Fig. 2.** The power delivery by vertical power/ground vias and its impact to inductive current loops.

**Fig. 3.** The definitions of the cycle-accurate power, transient thermal-power, and maximum thermal-power at the different scale of time-constant.
macromodel. As our power/ground vias are allocated after the placement and global routing of signal nets at each active device layer, the densities of those inter-layer signal nets are available to calculate a maximum density $n_{max}$ for the power/ground vias. In addition, for the sake of the reliability concern for the large current, the via density $n_t$ on the other hand can not be smaller than the minimum density $n_{min}$. These parameters ($n_{max}, n_{min}, V_t, T_t$) can be estimated and provided by users.

III. INTEGRITY OPTIMIZATION WITH MACROMODEL

We represent the 3D ICs by two distributed models: a thermal-RC model for the heat-removal and an electrical-RLC model for the power-delivery. They (without power/ground vias) can be described in the state-space by

$$G x(t) + C \frac{dx(t)}{dt} = B I(t), \quad y(t) = L^T x(t)$$

or in frequency (s) domain

$$(G + sC) x(s) = B I(s), \quad y(s) = L^T x(s).$$

Note that $B$ is the topology matrix to describe $p_i$ input ports with injected input sources, and $L$ is the one to describe $p_o$ output ports for probing integrity and adjusting via density.

To consider dynamic integrity during the design optimization of our TSV allocation problem, main difficulties to apply the above state-space equation come from three-fold. Firstly, there are so many inputs to try and so many outputs to probe. Secondly, the dimension of the distributed thermal-RC and electrical-RLC models are too large to analyze. Lastly, for the sake of design optimization, we are more interested in the sensitivity than the nominal response. In the following, we will show how to compress the I/Os and further generate an effective structure and parameterized macromodel for the design automation.

A. Compression of I/Os

Generally, there can be thousands of thermal-power sources injected at each active layer or hundreds of switching-current sources injected at I/Os. The size of the macromodel increases with the number of ports, and hence the computational cost to solve the macromodel is still high. Since the electrical signals may share the same clock and operate within a similar logic function, their waveforms in the time-domain at certain input ports can show a correlation. Similarly, the thermal power may differ significantly between those regions with and without the clock gating, but can be quite similar inside the region with the same mode since inputs have similar duty-cycles over time. Based on the correlation, we can reduce the redundancy in I/Os by identifying those principal ports.

We call this phenomenon input similarity. As the input vector

$$I(t) = [I_1 \quad I_2 \quad \cdots \quad I_{p_i}] \in R^{p_i \times 1},$$

is usually known during the physical design, they can be represented by taking a set of 'snapshots' sampled at $N$ time-points

$$[I_1(t_0) \quad \cdots \quad I_1(t_N)] \quad [I_2(t_0) \quad \cdots \quad I_2(t_N)] \quad \cdots \quad [I_{p_i}(t_0) \quad \cdots \quad I_{p_i}(t_N)]$$

in a sufficiently long period $[0, T]$. The sampling cycle is in a different time-scale for the thermal-power (ms) and switching-current (ns). According to the POD analysis [18], the similarity can be mathematically described by a correlation matrix (or Grammian), estimated by a co-variance matrix:

$$R = \frac{1}{N} \sum_{\alpha=1}^{N} (I(t_{\alpha}) - \bar{I})(I(t_{\alpha}) - \bar{I})^T \in R^{p_i \times p_i}.$$  

$\bar{I}$ is a vector of mean values defined by:

$$\bar{I} = \frac{1}{N} \sum_{\alpha=1}^{N} I(t_{\alpha})$$

Usually, the input vector $I(t)$ is periodic and the waveform in each period can be approximated by the piecewise-linear model.

An output similarity is defined for responses at output ports and measured by a output correlation matrix. To extract the output correlation matrix that is independent on the inputs, we assume that $p_i$ inputs in the input vector $I(s)$ are all the unit-impulse source $h(s)$ and define an input-port vector $J(s)$ by

$$J = BI(s), \quad \in R^{1\times N}_s,$$

which has $p_i$ non-zero entries with the unit-value ‘1’. Accordingly, the $p_o$ output responses $y(s)$ are calculated by

$$y(s) = L^T (G + sC)^{-1} J = [y_1(s) \quad y_2(s) \quad \cdots \quad y_{p_o}(s)] \in R^{p_o \times 1}_s.$$  

The according output correlation matrix is extracted in the frequency-domain. Similarly, the output signals can be represented by taking a set of ‘snapshots’ sampled at $N$ frequency points

$$[y_1(s_0) \quad \cdots \quad y_1(s_N)] \quad [y_2(s_0) \quad \cdots \quad y_2(s_N)] \quad \cdots \quad [y_{p_o}(s_0) \quad \cdots \quad y_{p_o}(s_N)]$$

in a sufficiently wide band $[0, s_{max}]$. The $s_{max}$ locates in a low-frequency range for the temperature and in a high-frequency range for the voltage. A co-variance matrix is defined in the frequency-domain as follows

$$R = \sum_{\alpha=1}^{N} (y(s_{\alpha}) - \bar{y})(y(s_{\alpha}) - \bar{y})^T \in R^{p_o \times p_o}$$

to estimate the correlation matrix among $p_o$ outputs. $\bar{y}$ is a vector of mean values defined by:

$$\bar{y} = \frac{1}{N} \sum_{\alpha=1}^{N} y(s_{\alpha})$$
Let $V = [v_1, v_2, ..., v_K] (\in R^{N \times K})$ as the first K singular-value vectors of the input correlation matrix $R$, and $W = [w_1, w_2, ..., w_K] (\in R^{N \times K})$ as the first K singular-value vectors of the output correlation matrix $R$. All singular-value vectors are obtained from the singular-value decomposition (SVD) of $(V, W)$. A rank-K matrix $P_i$ can be constructed by $P_i = Y_iV^T$, and a rank-K matrix $P_o$ can be constructed by $P_o = W_iW^T$. As shown in [18], the correlation matrix $(R, R)$ is essentially the solution that minimizes the least-square between the original states $(I(t), y(s))$ and their rank-K approximations $(P_i \cdot I(t), P_o \cdot y(s))$. As a result, both the input signals $I(t)$ and the output signals $y(s)$ can be approximated by an invariant (or dominant) subspace spanned by the orthonormalized columns of $V$ and $W$, respectively:

$$I = YI_K, \quad y = Wy_K.$$  

(13)

Based on (13), it leads to the following equivalent system equation

$$(G + sC)x_K(s) = B_KI_K(s), \quad y_K(s) = L_K^Tx_K(s)$$

where

$$L_K^T = W^T L^T, \quad B_K = BV.$$  

(15)

Therefore, both the dimensions of $L (\in R^{N \times p_o})$ and $B (\in R^{N \times p_i})$ are greatly reduced when $K << p_i$ and $p_o$. We call $I_K$ and $y_K$ principal inputs and outputs identified by principal input-port and output-port matrices $B_K$ and $L_K$, respectively.

B. Dynamic Sensitivity by Structured and Parameterized Macromodel

Recall that the design parameter in our problem formulation is the via density at one track. Blindly allocating the via by searching all kinds of combinations would be computationally expensive if not impossible. Therefore, we decide the via density based on the changes at outputs, i.e., sensitivities, caused by the change of via density.

Let’s first parameterize the nominal system 3. The added via is described by two parameters: $n_j$ the via density and $X_j$ the topological matrix that connects the via into the nominal system. As such, a parameterized state-space description can be obtained by

$$(G + sC + \sum_{j=1}^{p_o} n_j g_j + s \sum_{j=1}^{p_o} n_j c_j)x(n, s) = B_KI_K(s), \quad y_K(n, s) = L_K^Tx(n, s).$$  

(16)

Similar to [6], [7], [17], we expand $x(n, s)$ in the Taylor series with respect to $n_j$, and introduce a new state variable $x_{ap}$

$$x_{ap} = [x^{(0)} \cdot x^{(1)} \cdot \ldots \cdot x^{(1)}]^T,$$  

(17)

It contains both the nominal response $x^{(0)}$ and its first-order sensitivities $x^{(1)} \cdot \ldots \cdot x^{(1)}$ with respect to $p_o$ parameters $[n_1, \ldots, n_{p_o}]$. The overall responses is obtained by

$$x = x^{(0)} + \sum_{j=1}^{p_o} x_j^{(1)}.$$  

Substituting (17) into (16), (16) can be reformulated into a parameterized system with an augmented dimension by

$$(G_{ap} + sC_{ap})x_{ap} = B_{ap}I_K(s), \quad y_{ap} = L_{ap}^Tx_{ap},$$  

(18)

where $G_{ap}$ and $C_{ap}$ show a lower-triangular-block structure and hence $x_{ap}$ can be solved from block-backward-substitution.

To further compress the dimension of the state-matrices $G_{ap}$ and $C_{ap}$, we first construct a lower-dimensioned subspace $Q_{ap}$ from the moment expansion of (18), and then transform $Q$ into the block-diagonal form $Q_{ap}$. After the block-orthonormalization of $Q_{ap}$, we apply a two-side projection to (18) by $Q_{ap}$ and obtain a dimensioned-reduced system with preserved lower-triangular-block structure [6], [7], [16], [17]. The accuracy of the macromodel is preserved to match the dominant moments of the original model. More importantly, due to the structure-preservation, both of the nominal response and the sensitivity with regard to the via-density change, can be calculated simultaneously. As such, we can easily embed such a structured and parameterized macromodel into the optimization flow of our TSV allocation problem.

IV. RESULTS

Experiments are implemented in C and MATLAB and run on a Sun-Fire-V250 workstation with 2G RAM. We call the separated allocation of thermal vias and power/ground vias the sequential optimization, and call our allocation of power/ground vias for both power and thermal integrity the simultaneous optimization. Moreover, the steady-state analysis is employed to calculate a static integrity [3], [4]. We use the sequential optimization with the static integrity as the baseline, in comparison to the sequential optimization with the dynamic integrity and the simultaneous optimization with the dynamic integrity proposed in this paper. Table I and Table II summarize the used electrical and thermal constants and dimensions. The targeted voltage violation $V_t$ is 0.2V and the targeted temperature $T_t$ is 52°C. One modest 3D stacking is assumed with 2-device-layer/2-dielectric-layer. Moreover, there are 1-heat-sink and 2-P/G-plane used.

<table>
<thead>
<tr>
<th></th>
<th>Silicon</th>
<th>Copper</th>
<th>Dielectric</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma$</td>
<td>NA</td>
<td>$5.96 \times 10^5$ S/m</td>
<td>NA</td>
</tr>
<tr>
<td>$\epsilon_r$</td>
<td>NA</td>
<td>NA</td>
<td>3.3</td>
</tr>
<tr>
<td>$\mu_r$</td>
<td>NA</td>
<td>NA</td>
<td>1.0</td>
</tr>
<tr>
<td>$\kappa_K$</td>
<td>1000 W/m · K</td>
<td>400 W/m · K</td>
<td>50 W/m · K</td>
</tr>
<tr>
<td>$\kappa_C$</td>
<td>$1.75 \times 10^8$ J/m$^3$ · K</td>
<td>$3.55 \times 10^8$ J/m$^3$ · K</td>
<td>0.7 W/m · K</td>
</tr>
</tbody>
</table>

**TABLE I**

Elecrical and thermal constants.

<table>
<thead>
<tr>
<th>layer</th>
<th>size</th>
<th>material</th>
</tr>
</thead>
<tbody>
<tr>
<td>heat-sink</td>
<td>2cm × 2cm × 1mm</td>
<td>copper</td>
</tr>
<tr>
<td>device-layer</td>
<td>1cm × 1cm × 4um</td>
<td>silicon</td>
</tr>
<tr>
<td>inter-layer</td>
<td>1cm × 1cm × 1um</td>
<td>dielectric</td>
</tr>
<tr>
<td>P/G plane</td>
<td>2cm × 2cm × 10um</td>
<td>copper</td>
</tr>
</tbody>
</table>

**TABLE II**

Dimensions of 3D ICs layers.
We compare the runtime and the number of vias in Table III. In Table III, column 2-3 show the runtime and the number of allocated vias for the baseline, and column 4-8 show the results for the optimizations using the dynamic integrity. In detail, column 4 shows the runtime of transient analysis using macromodels without the port-compression, and column 5 shows the number of allocated vias under the sequential optimization. Column 6 shows the runtime of transient analysis using macromodels with the port-compression, and column 7-8 shows the number of allocated vias under the sequential and simultaneous optimizations, respectively.

The use of macromodels reduces the computational cost to solve power and thermal integrity and their sensitivities. Compared to the macromodel without the port-compression, the macromodeling with the port-compression reduces the overall runtime up to 16X with similar allocation results. Compared to steady-state analysis with full-matrix analysis, our macromodel with the port-compression has a 127X smaller runtime. Additionally, the steady-state analysis can not complete the largest example in a reasonable runtime. The maximum transient-waveform difference introduced by the macromodel is about 7% when compared to the exact transient waveform.

We further compare the sequential thermal/power optimization with the simultaneous thermal/power optimization. Here both methods allocate vias with the use of dynamic integrity. Our simultaneous optimization reduces the via-cost up to 34% when compared to the sequential optimization with static integrity, and up to 22% when compared to the sequential optimization with dynamic integrity. This demonstrates that the reusing of power/ground vias can reduce the via cost when compared to allocating the dummy thermal vias separately from the power/ground vias.

V. CONCLUSION

This paper explains the need for dynamic power and thermal integrity for the high-performance 3D integration, using an example of the through-silicon-via (TSV) allocation. To cope with design complexity, an effective macromodel is employed to abstract the physical level detail for the system level design. It includes the I/O compression and the structured and parameterized model order reduction, which efficiently calculate power/thermal integrity and their sensitivity with respect to via density. Compared to the design without the use of the dynamic integrity, experiments show that our approach reduces the the number of TSVs up to 38%, yet achieves the speedup by hundreds of times.

REFERENCES