Chapter 2
3D Integration Technology

Yuan Xie and Qiaosha Zou

Abstract The emerging three-dimensional (3D) chip architectures, with their intrinsic capability of reducing the wire length, is one of the promising solutions to mitigate the interconnect problem in modern microprocessor designs. To leverage the benefits of fast latency, high bandwidth, and heterogeneous integration capability that are offered by 3D technology, new design methodologies should be developed targeting the unique feature of 3D integration. In this chapter, various approaches to model 3D electrical behavior, handle 3D thermal reliability problems, and design future 3D microprocessors are surveyed.

2.1 Introduction

With continued technology scaling, interconnect has emerged as the dominant source of circuit delay and power consumption. The reduction of interconnect delay and power consumption are of paramount importance for deep-sub-micron designs. Three-dimensional integrated circuits (3D ICs) [11] are attractive options for overcoming the barriers in interconnect scaling, thereby offering an opportunity to continue performance improvements using CMOS technology.

3D integration technologies offer many benefits for future microprocessor designs. Such benefits include: (1) *The reduction in interconnect wire length*, which results in improved performance and reduced power consumption; (2) *Improved*

This chapter includes portions reprinted with permission from the following publications: Qiaosha Zou, Tao Zhang, Eren Kursun, and Yuan Xie. Thermomechanical stress-aware management for 3D IC designs. Proceedings of Design, Automation Test in Europe Conference Exhibition (DATE) (2013). Copyright 2013 IEEE.

Y. Xie (✉)
University of California, Santa Barbara, CA, USA
e-mail: yuanxie@ece.ucsb.edu

Q. Zou
The Pennsylvania State University, State College, PA, USA
e-mail: qszou@cse.psu.edu

R.O. Topaloglu (ed.), *More than Moore Technologies for Next Generation Computer Design*, DOI 10.1007/978-1-4939-2163-8_2

memory bandwidth, by stacking memory on microprocessor cores with TSV connections between the memory layer and the core layer; (3) *The support for realization of heterogeneous integration*, which could result in novel architecture designs. (4) *Smaller form factor*, which results in higher packing density and smaller footprint due to the addition of a third dimension to the conventional two dimensional layout, and potentially results in a lower cost design.

To design the 3D microprocessor that can fully leverage the benefits of fast latency, higher bandwidth, and heterogeneous integration capability that are offered by 3D technology, understanding of 3D electrical behavior is necessary and mature techniques should be developed to handle the unique thermal reliability challenge in 3D technology.

This chapter first presents the background on 3D integration technology, and then reviews the models to capture the 3D electrical behaviors. The challenges of 3D thermal reliability are then presented. Various approaches to design future 3D microprocessors, leveraging the benefits from 3D technology, are surveyed. The challenges for future 3D architecture design are also discussed in the last section.

2.2 3D Integration Technology

The 3D integration technologies [56, 57] can be classified into one of the two following categories. (1) *Monolithic approach*. This approach involves sequential device process. The frontend processing (to build the device layer) is repeated on a single wafer to build multiple active device layers before the backend processing builds interconnects among devices. (2) *Stacking approach*, which could be further categorized as wafer-to-wafer, die-to-wafer, or die-to-die stacking methods. This approach processes each layer separately, using conventional fabrication techniques. These multiple layers are then assembled to build up 3D IC, using bonding technology. Since the stacking approach does not require the change of conventional fabrication process, it is easier to adopt compared to the monolithic approach, and has become the focus of recent 3D integration research.

Several 3D stacking technologies have been explored recently, including wire bonded, microbump, contactless (capacitive or inductive), and *through-silicon vias (TSV)* vertical interconnects [11]. Among all these integration approaches, TSV-based 3D integration has the potential to offer the greatest vertical interconnect density, and therefore is the most promising one among all the vertical interconnect technologies. Figure 2.1 shows a conceptual 2-layer 3D integrated circuit with TSV and microbump.

3D stacking can be carried out using two main techniques [16]: (1) *Face-to-Face (F2F)* bonding: two wafers/dies are stacked so that the very top metal layers are connected. Note that the die-to-die interconnects in face-to-face wafer bonding does not go through a thick buried Silicon layer and can be fabricated as *microbump*. The connections to C4 I/O pads are formed as TSVs; (2) *Face-to-Back (F2B)* bonding: multiple device layers are stacked together with the top metal layer of

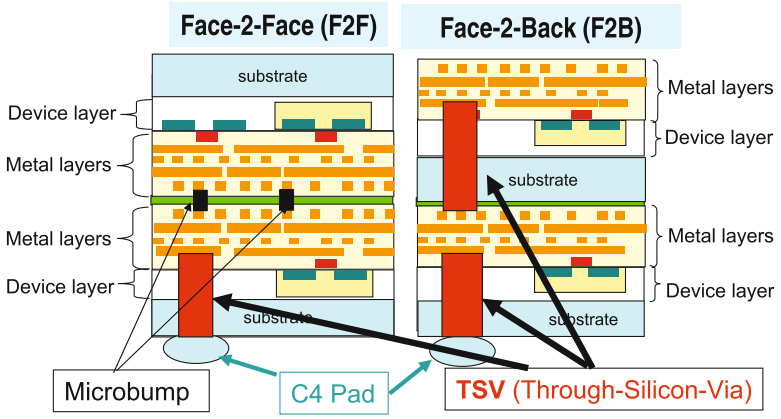

Fig. 2.1 Illustration of F2F and F2B 3D bonding

one die bonding together with the substrate of the other die, and direct vertical interconnects (which are called *through-silicon vias (TSV)*) tunneling through the substrate. In such F2B bonding, TSVs are used for both between-layer-connections and I/O connections. Figure 2.1 shows two conceptual 2-layer 3D ICs with F2F and F2B bonding, with both TSV connections and microbump connections between layers.

All TSV-based 3D stacking approaches share the following three common process steps [16]: (a) *TSV formation*; (b) *Wafer thinning* and (c) *Wafer alignment or die bonding*, which could be wafer-to-wafer(W2W) bonding or die-to-wafer(D2W) bonding. Wafer thinning is used to reduce the area impact of TSVs. The thinner the wafer, the smaller (and shorter) the TSV is (with the same aspect ratio constraint) [16]. The wafer thickness could be in the range of 10–100 μm and the TSV size is in the range of 0.2–10 μm [11].

In TSV-based 3D stacking bonding, the dimension of the TSVs is not expected to scale at the same rate as feature size because alignment tolerance during bonding poses limitation on the scaling of the vias. The vertical connection size, length, and the pitch density, as well as the bonding method (face-to-face or face-to-back bonding, SOI-based 3D or bulk CMOS-based 3D), can have a significant impact on the 3D microprocessor design. For example, relatively large size of TSVs can hinder partitioning a design at fine granularity across multiple device layers, and make the true 3D component design less possible. On the other hand, the monolithic 3D integration provides more flexibility in vertical 3D connection because the vertical 3D via can potentially scale down with feature size due to the use of local wires for connection. Availability of such technologies makes it possible to partition a design at a very fine granularity. Furthermore, face-to-face bonding or SOI-based 3D integration may have a smaller via pitch size and higher via density than face-to-back bonding or bulk-CMOS-based integration. Such influence of the 3D technology parameters on the microprocessor design must be thoroughly studied before an appropriate partition strategy is adopted.

2.3 TSV-Based 3D Electrical Model

For circuit performance (delay, power consumption, and heat dissipation) estimation, RLC model is the most straightforward modeling method by treating the device as composed of resistance, capacitance and inductance. In recently explored 3D stacking approach, TSV is working as the key enabling component, making the electrical modeling of TSV nontrivial. Therefore, this section primarily reviews the previous work that modeling TSV as RLC model under different conditions (frequency, temperature, etc.), followed by the brief introduction of electrical modeling of microbumps and back-end-of-line (BEOL).

2.3.1 TSV RC Model in Low Frequency Region

The TSV geometry description influences the final modeling accuracy. Most papers assume that TSVs are equivalent cylindrical structure [26, 44, 52, 58] and this assumption is examined by paper [44]. The researchers compared the electrical parameters extracted from Ansoft electromagnetic simulation tool with two geometry descriptions. The first one is a cylinder structure containing both top and bottom copper landing pads and another is the proposed simple structure without landing. The results show that only less than a 7 % difference is found in the RLC value, indicating that using simple cylindrical structure is sufficient for TSV modeling. However, this examination is performed with frequency as high as 1 GHz under stationary temperature. The conclusion may not be applicable for higher frequency beyond this point.

Most of previous work are on developing simple analytical model for TSV delay estimation. In [26], a physical dimension dependent analytical model for the propagation delay of TSVs is proposed. A lumped element model using dimensional analysis method is proposed with the observation that TSVs have a MOS-like capacitor structure [43]. Similarly, a lumped TSV model and the corresponding TSV propagation delay analysis are demonstrated in [23]. Besides the TSV's structure, the process method also influences the TSV electrical characteristic. An electrical and reliability study based on a fabricated via last TSV is presented in [35]. As it is illustrated in the work, the structural and material parameters both have impact on 3D TSV electrical characteristic. A 3D full wave and SPICE circuit simulation is performed and eye-diagrams at different frequencies are used to study the impact [38].

The analytical models from above mentioned work can provide electrical behavior analysis, however, closed-form equations are needed for real value calculation. The RLC model for single TSV and coupled TSVs with closed-form expressions are given [8, 23, 44, 52]. In [8], in addition to the analysis, a guard ring structure is proposed to suppress the noise coupling in TSVs. Closed-form expressions derived in [44] consider various effects, such as skin effect, therefore, the expression is

relatively complicated with several parameters that need to be determined based on the given operation frequency. The expression is consistent with simulation results up to 2 GHz frequency in the paper. The expressions given in this paper are accurate, however, they are not suitable for fast circuit simulation. Empirical parameters are used in [52], which is more practical for full chip circuit simulation. Due to the relatively large size of TSVs, coupling effect is usually prominent which should be taken into consideration for full chip analysis. In the following section, the low frequency RLC equations for isolated TSV are introduced followed by the expressions for the coupling capacitance and mutual inductance in TSV grid.

2.3.1.1 RLC Model for an Isolated TSV

For an isolated TSV, the RLC model is shown in Fig. 2.2a. Capacitances exist between TSV and the adjacent substrate while resistance and inductance are in series along the TSV.

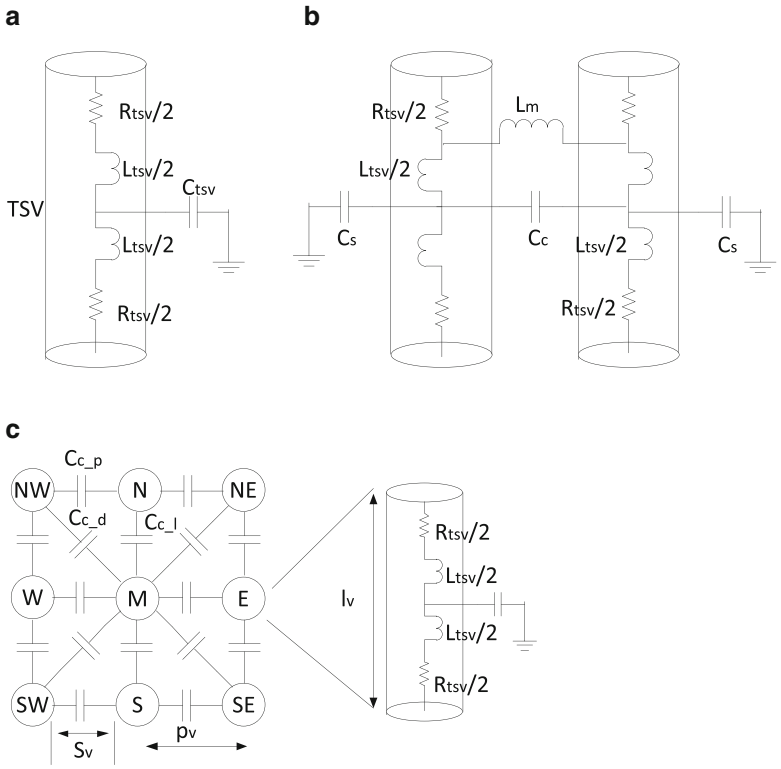

Fig. 2.2 The resistance, inductance, and capacitance components for (**a**) an isolated TSV; (**b**) two coupled TSVs; (**c**) a TSV bundle [52]

The resistance calculation can be described as the function of TSV conductivity (σ), TSV length (l), and radius (r):

$$R_{tsv} = \frac{l_v}{\sigma \pi r_v^2} \tag{2.1}$$

TSV has a MOS-like capacitor structure, therefore, the effective capacitance in TSV is the depletion capacitance and the oxide capacitance acting in series. During TSV formation, a dielectric layer is deposited between TSV metal and surrounding silicon, which makes TSV has the similar MOS structure. Based on the MOS effect modeling, a depletion region appears with introduced depletion capacitance. The depletion region width is determined by the voltage on TSV, threshold voltage (derived from flatband energy), interface charge density, and material properties. The depletion width changes with correspondence to the bias voltage when other conditions are fixed [59]. The final effective capacitance is a function of its geometry and the effective permittivity (ϵ_0) of surrounding dielectric liner. The following expression is based on empirical formula which assumes the thickness of dielectric layer is smaller than 1 μm:

$$C_{tsv} = \frac{63.36\varepsilon_0 l_v}{ln\left(1 + 5.26\frac{l_v}{r_v}\right)} \tag{2.2}$$

When the TSV is treated as a lossy transmission line in the model, the inductance has great impact on signal propagation delay. The propagation delay study in [26] shows that without the presence of inductance in TSVs, the average error is 55.2 % higher than the value of the distributed RLC model. The inductance of an isolated TSV is depended on the geometry parameters. It can be expressed as follows:

$$L_{tsv} = \frac{\mu l_v}{2\pi} ln\left(1 + \frac{2.84}{\pi}\frac{l_v}{r_v}\right) \tag{2.3}$$

All the above empirical closed-form equations are verified by a 3D/2D quasi-static electromagnetic-field solver tool and results show that the maximum error is within 6 %. By using these closed-form expressions, the resistance, capacitance, and inductance values in a single TSV can be easily calculated for fast circuit simulation.

2.3.1.2 RLC Model for Coupled TSVs

The RLC models for two coupled TSVs and a TSV bundle are shown in Fig. 2.2b, c, respectively. For coupled TSVs, the resistance expression is the same as that in an isolated TSV since coupling effect has negligible impact on the resistance.

But for inductance and capacitance, the inter-via coupling effects are prominent. In the following analysis, capacitance and inductance are divided into two parts: self parameter and mutual parameter.

The capacitance of the whole coupled bundle TSVs can be expressed as follows:

$$C_{bundle} = \begin{bmatrix} C_{1,1} & -C_{1,2} & \dots & -C_{1,n} \\ -C_{2,1} & C_{2,2} & \dots & -C_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ -C_{n,1} & -C_{n,2} & \dots & C_{n,n} \end{bmatrix} \tag{2.4}$$

The diagonal element means the sum of self and inter-via coupling capacitances. This capacitance matrix is sparse because only the diagonal elements and elements that represent nearest neighbors contain meaningful values. From the researchers' experiments [52], for a 7×7 TSV bundle, coupling terms for nearest neighbors are more significant than those that are non-adjacent.

The self capacitance formula is different from the isolated TSV, which is given as:

$$C_s = C_{tsv} - k_1 C_{tsv} e^{(k_2 \frac{p_v}{r_v} + k_3 \frac{p_v}{l_v})} \left[k_4 \left(\frac{L_v}{r_v} \right)^{k_5} + k_6 \left(\frac{p_v}{r_v} \right)^{k_7} + k_8 \right] \tag{2.5}$$

where C_{tsv} is the capacitance of an isolated TSV, the parameters from k_1 to k_8 are empirical constants. However, these constants are based on the simulation results and varied with different TSV configurations, making it hard to be directly used in circuit simulation. When k_2 and k_3 are negative and p_v approaches infinity, C_s equals to C_{tsv}.

The formula for the coupling capacitance where $i \neq j$ in the matrix is given as follows:

$$C_{coupled} = \frac{k_1 \epsilon_0 l_v}{ln(k_2 \frac{pv}{rv})} \left[1 + k_5 \left(\frac{L_v}{r_v} \right)^{k_6} + k_3 \left(\frac{p_v}{r_v} \right)^{k_4} + k_7 \left(\frac{p_v}{l_v} \right)^{k_8} \right] \tag{2.6}$$

The coupling inductance terms is defined similarly to the coupling capacitance. Different from capacitance, inductive coupling effect has long range, therefore, the matrix is not sparse. The mutual inductance between any two TSVs can be captured with the following formula:

$$L_m = 0.199 \mu l_v ln \left(1 + 0.438 \frac{d_v}{l_v} \right) \tag{2.7}$$

where d_v is the center-to-center distance between two TSVs.

Based on the simulation results, the maximum error for coupling capacitance and inductance are within 6 and 8 %, respectively.

2.3.2 TSV RLCG Model in High Frequency Region

Previous content introduces previous work that use simple RLC model in low frequency region. In the following section, the relatively complex RLCG model is introduced for high frequency operation which is up to 100 GHz [41, 59]. In this model, the substrate is not assumed as ideal conductor, therefore, the impact of substrate resistance is taken into consideration. The RLCG model contains two components: admittance per unit TSV height which consists of conductance and susceptance; impedance per unit TSV height which is composed of resistance and reactance.

Skin effect and eddy currents in silicon should be also taken into consideration for TSV modeling at high frequencies [59]. Skin effect means the current density drops by a certain factor below the surface of a conductor. It has great impact on the high frequency resistance.

The RLCG model developed in [59] is introduced in detail. The equivalent distributed circuit model is shown in Fig. 2.3a, the simplified model is given in Fig. 2.3b. The impedance which is represented by Z is inside TSV, similar to the resistance and inductance in series in RLC model. Capacitance $C1$ resides between TSV and substrate representing the final effective capacitance (oxide capacitance and depletion capacitance in series). Admittance, represented by Y, exists in the silicon substrate between two adjacent TSVs. In this figure, Y_{open} is the input admittance between ports 1 and 2 if ports 3 and 4 are open while Z_{short} represents the impedance between ports 1 and 2 when ports 3 and 4 are short circuited.

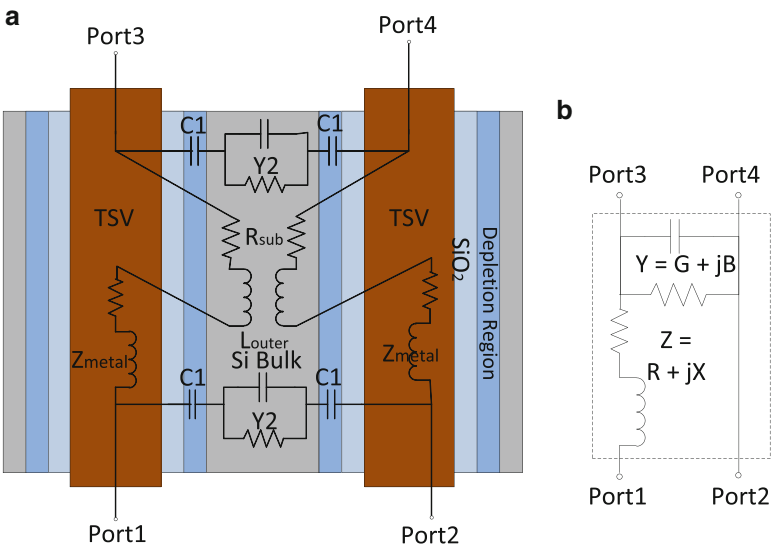

Fig. 2.3 RLCG model for TSV, (**a**) the equivalent distributed RLCG model of two coupled TSVs; (**b**) a simplified distributed transmission line model [59]

2.3.2.1 Admittance of TSV in RLCG Model

The admittance (CG) per unit TSV height can be treated as two components working in series, one is the effective capacitance (C_1) and the other is the coupling admittance (Y_2) due to the bulk silicon. The admittance expression is shown in the following equation:

$$Y = [2(j\omega C_1)^{-1} + Y_2^{-1}]^{-1} \tag{2.8}$$

where ω is the radial frequency. Since there are two TSVs contributing to C_1 in series with Y_2, the equation contains a factor of 2 before C_1. The detailed equations to calculate C_1 and Y_2 can be found in [59].

The CG model is verified with a 2-D quasi-electrostatic simulation tool. The results suggest that at low frequencies, if the depletion region is not considered, the error is not negligible, however this difference is not so significant at high frequencies.

2.3.2.2 Impedance of TSV in RLCG Model

The serial impedance (RL) per unit height is not so straightforward. The final expression results are shown here without detailed deduction steps. For simplicity, the serial impedance can be treated as the sum of three components: the inner impedance of TSV (Z_{metal}), the outer inductance (L_{outer}), and the resistance due to eddy currents in silicon substrate (R_{sub}), where the equations for these three components can be found in [59]:

$$Z = 2Z_{metal} + j\omega L_{outer} + R_{sub} \tag{2.9}$$

The model is compared with the simulation tool and the results indicate skin effect in TSV is of great importance for high frequency analysis when the higher frequency resistance is dominant over DC resistance.

2.3.2.3 TSV Electrical Performance with RLCG Model

As technology scales, the diameter and pitch of TSVs shrink, however, the substrate thickness almost remains the same as predicted by the ITRS. When the radius of TSVs reduce, C, G and L do not change much due to the proportional scaling of geometrical parameters. Nevertheless, resistance increases significantly when the frequency reaches the region of tens of GHz due to the decreasing TSV cross sectional area.

In terms of circuit performance sensitivity, capacitance has the most important impact on circuit behavior while resistance is of the least importance. The interconnect exhibits the short-transmission line behavior on signal propagation, which

indicates that simple RLC model is enough for delay and signal rise/fall calculation. However, the L and G are crucial factors for the estimation of voltage variations in V_{DD} and *GND*. More accurate whole circuit performance evaluation can only be done with all RLCG models.

2.3.3 TSV RC Model with Temperature Consideration

Although 3D stacking provides a number of benefits over traditional 2D circuits, 3D exacerbates the thermal dissipation problems due to higher power density in smaller footprint. The temperature gradients on chip result in TSV electrical characteristic variation. Several work have explored the temperature dependent TSV modeling [22, 24, 55].

Due to the complexity of temperature-dependent TSV modeling, only semi-analytical capacitance model and empirical RC model are brought out in previous work. First, the semi-analytical capacitance model is briefly introduced. Then the empirical RC model formulations are given for straightforward RC value computation.

2.3.3.1 Semi-Analytical Temperature-Dependent Capacitance Model

This model is called semi-analytical because the close-form expression is not given, instead, a four-step algorithm is given to calculate the capacitance until the convergence conditions are satisfied. Normally, the behavior of TSV is similar to a MOS capacitor, and the analytical expression for TSV capacitance is derived by solving a 1D Poisson equation in the radial direction in a cylindrical coordinate system. Considering the depletion region, a semi-analytical algorithm for depletion capacitance calculation is proposed [24].

The algorithm first identifies the initial maximum depletion radius. The assumption neglects the hole and electron charges. The initial maximum depletion radius can be obtained from the following equation:

$$\frac{qN_aR_{OX}^2}{4\varepsilon_{Si}} - \frac{qN_aR_{max}^2}{2\varepsilon_{Si}}Ln(R_{OX}) + \frac{qN_aR_{max}^2}{4\varepsilon_{Si}}(2ln(R_{max}) - 1 = \psi_s \quad (2.10)$$

with the assumption that the surface potential ψ_s equals to $2(K_BT/q)ln(N_a/n_i)$. In the equation, q is the electron charge, N_a is the density of ionized acceptors or the doping concentration, ε is the silicon permittivity, and $\psi(r)$ represents the electrostatic potential with respect to the radius.

The second step is trying to identify electron-hole densities in the substrate from the potential with the initial depletion radius calculated from previous step. The potential at every point is calculated as follows:

$$\psi(r) = \frac{qN_a r^2}{4\varepsilon_{Si}} - \frac{qN_a R_{max}^2}{2\varepsilon_{Si}} ln(r) + \frac{qN_a R_{max}}{4\varepsilon_{Si}} (2ln(R_{max} - 1) \tag{2.11}$$

The value of the potential at distance r is used to compute the hole and electron charge densities in the substrate using $p(r) = p_{Po} exp(-\beta\psi(r))$ and $n(r) = n_{Po} exp(\beta\psi(r))$.

Step three calculates the new maximum depletion radius with consideration of the hole and electron charge densities derived from the previous step. The new maximum depletion radius is calculated from the following equation:

$$\frac{q(N_a + p - n)R_{OX}^2}{4\varepsilon_{Si}} - \frac{q(N_a + p - n)R_{max}^2}{2\varepsilon} ln(R_{OX}) + \frac{q(N_a + p - nR_{max}^2)}{4\varepsilon_{Si}} (2ln(R_{max}) - 1) = \psi_s \tag{2.12}$$

The last step calculates the depletion capacitance by using equation $C_{dep} = 2\pi\varepsilon_{Si} L_{TSV} / ln(2R_{max}/\phi_{TSV})$. The final depletion capacitance is obtained by continuing these four steps until the new depletion radius approaches the initial maximum depletion radius. The total TSV capacitance can be viewed as the oxide capacitance and depletion capacitance in series.

Comparison between the semi-analytical results and the measurement results shows that the error is within 3 %. When the temperature rises, the TSV capacitance increases due to the reduction of maximum depletion radius.

2.3.3.2 Empirical Temperature-Dependent RC Model

Besides the TSV capacitance, resistance also changes with temperature variation. Lumped RC model should be enhanced by considering TSV capacitance and resistance change due to temperature variation [22]. However, due to lack of close-form expression for temperature-dependent resistance, [22] builds a 2D/3D ring oscillator to measure the model parameters at different temperatures. Thus, an empirical RC model of TSV is discussed. This RC model is similar to the simple signal-transmission line model at DC and low frequencies without inductances.

The expressions of resistance and capacitance from empirical data are given in the following:

$$R_{TSV}(T) = R_0(1 + \alpha(T - T_0)) \tag{2.13}$$

$$C_{TSV}(T) = 0.0007T^2 - 0.0333T + 44.4 \tag{2.14}$$

The measurement results suggest that with temperature rise, substantial increment in TSV capacitance and resistance can be seen.

The temperature-dependent RLC model is still far beyond maturity in current research. Furthermore, from these work, we can see that the resistance and capacitance have great dependency on temperature. Moreover, these dependencies can be translated into further influence on the on-chip temperature by producing Joule heating [55]. Accurate modeling and electrical-thermal co-analysis framework are required for precise circuit performance and on-chip temperature estimation.

2.3.4 RC Model for Microbumps, RDL, and BEOL

In addition to TSV which is the key enabling component in 3D ICs, other components (microbumps, redistribution layers, C4, etc.) are necessary for electrical modeling to perform full chip and package analysis. Several work modeled the redistribution layer (RDL), back-end-of-line (BEOL) and microbumps [1, 40, 54].

The BEOL and RDL can be treated as traditional metal layers for signal transmission with resistance and capacitance. The theoretical values of resistance for RDL can be calculated from [40]:

$$R_{Th} = R_{RDL} + \frac{R_{Ground}}{2} = \alpha \frac{1}{w_{RDL} * t_{RLD}} \frac{3}{2} \tag{2.15}$$

where w_{RDL} and t_{RDL} are the width and the oxide thickness of the metal layers. The resistance value of BEOL can be obtained in a similar way. The capacitance calculation is absent in previous work.

Microbump can not be simply treated as metal layers due to the structure difference. In [54], the microbump has the RLC model similar to the TSV and shares the same expressions. The microbump model contains resistance and inductance in series from input to output port. Two capacitances reside between microbump and connected tiers. One of the capacitance represents the capacitance between microbump and substrate of the first tier while the other captures the capacitance between microbump and the substrate of the second tier.

2.4 Thermal Stress-Aware Design for 3D ICs

Stacked chips on 3D architecture increase the packaging density and thermal resistances, which results in higher on-chip temperatures. Plenty of studies have focused on the 3D thermal modeling, analysis [6, 7, 19, 51], and thermal-aware design methodology [9, 10, 17] to manage the on-chip thermal issues of 3D ICs. These work, however, failed to consider the TSV lateral thermal blockage effect and thermomechanical stress. Moreover, they used TSVs as thermal vias to build

vertical heat dissipation path, which in turn results in increased thermal load on TSVs as well as thermomechanical stresses, and thus weakens the reliability. On the other hand, prior work on analyzing the mechanical stresses in 3D ICs [2, 21] only consider the static stress management by adjusting TSV keep-out zone size, TSV placement, or TSV structure. In [61], the work not only accounts for the static (design-time) management of TSV thermal stress and thermal load but also takes into account the run-time TSV stress analysis and management.

For better thermal management, profound understandings of 3D heat transfer and cycling effects are necessary. Accurate 3D ICs thermal modelings have been conducted. An analytical and numerical model for temperature distribution in a 3D stack considering multiple heat sources is developed to help 3D thermal analysis [19]. An analytical thermal model for the top layer in 3D architecture with TSVs vertical thermal conductivity model is proposed to determine TSVs density during design time [51]. TSV is one of the most important component in 3D ICs, precise thermal modeling of TSVs can significantly improve the thermal analysis of 3D architectures. The equivalent thermal conductivity model [7] and lateral thermal blockage model [6] of TSVs are demonstrated. The thermal modeling for both silicon devices and TSVs in vertical and horizontal directions should be used for precise temperature analysis in 3D architectures.

Based on the thermal modeling and analysis of 3D architectures, several work have performed thermal-aware design. Thermal-aware 3D design placement techniques with TSVs for thermal vias are introduced to alleviate the on-chip temperature [9, 10, 17]. These work, however, fail to take the TSV lateral thermal blockage effects into consideration. As thermal vias, TSVs are likely to place near hotspot for vertical thermal dissipation, but the lateral blockage effects may worsen the thermal problem in horizontal direction.

The above mentioned work on thermal-aware design only makes effort on reducing the on-chip temperature without considering the thermomechanical stresses related reliability issues in 3D architectures. Analysis of reliability problems induced by thermomechanical stresses and strains is performed but it includes one single TSV [3]. Full-chip thermomechanical stresses and reliability analysis tool is generated to alleviate the reliability problems in 3D ICs [21]. The 3D FEA (finite element analysis) simulations are performed to examine the effect of TSV structure and liner material/thickness on TSV radial stress. Superposition method which is proved to be effective is applied for full chip analysis with TSV bundles. Besides the thermal stress analysis, stress-aware reliability schemes are also developed. Both design-time and run-time thermal stress management strategies are developed with the consideration of TSV horizontal thermal blockage effects in [61]. During design time, the management scheme tries to reduce the thermal load on TSV to reduce the TSV thermal stress, preventing the early time TSV interfacial delamination and wafer cracking. Moreover, thermal cycling effect is considered during run-time and thermal control mechanism is used to provide mechanical equilibrium for whole chip reliability.

The detailed TSV thermal stress model and 3D thermal cycling effect are introduced in the following subsection.

2.4.1 Analysis of TSV Thermal Stress

In 3D IC fabrication, copper (Cu) is usually used as TSV filling material. Copper has more than five times larger coefficient of thermal expansion (CTE) than silicon. The CTE mismatch between TSV and silicon substrate in turn introduces mechanical stresses that can lead to high probability of die cracking and interfacial delamination [33, 34, 45]. The coefficients of thermal expansion of TSV materials and silicon are listed in Table 2.1. The CTE of four possible TSV materials are all larger than the CTE of silicon substrate. As an example, Fig. 2.4 illustrates the potential cracking and delamination damage. Once heating is applied to the die, TSVs tend to expand much faster than silicon; this finally results in TSVs stretching out of silicon substrate. As a consequence, damage is generated in back-end-of-line (BEOL) and wire layers [33]. On the other hand, the contracted TSVs pull the surface of surrounding silicon during the cooling process, causing surface delamination and tensile stress in the surrounding region. Since silicon substrate is thinned drastically to expose TSVs, it is more vulnerable to mechanical stresses than 2D circuits.

Table 2.1 Coefficient of thermal expansion of TSV materials and silicon [42]

Material	CTE (ppm/K)
Silicon	2.3
Copper	17
Aluminium	20
Tungsten	4.4
Nickel	13

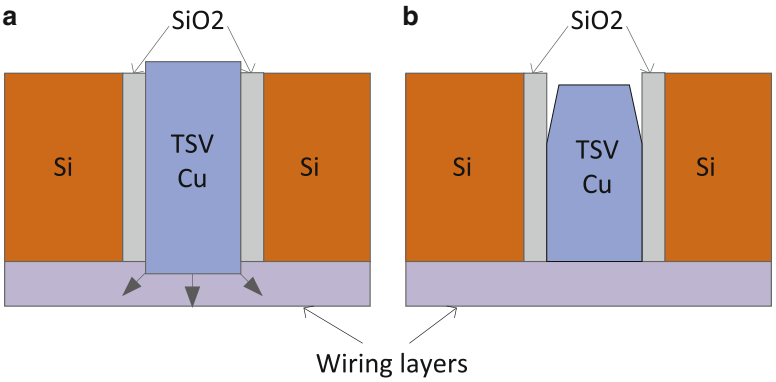

Fig. 2.4 TSV thermal expansion and delamination due to CTE mismatch. (**a**) TSV expansion during heating; (**b**) the delamination between TSV and silicon during cooling [33]

To minimize thermomechanical stresses, TSV farms should be placed smartly during design time. Therefore, the corresponding analysis on the thermal stresses around TSVs is critical to the solution. Several work [32, 42] have targeted thermal stress analysis showing that the stress field in TSVs is uniform and can be represented by radial, circumferential, and axial stresses. The stresses can be expressed as following:

$$\sigma_r = \sigma_\theta = \frac{-E(\alpha_{tsv} - \alpha_{si})T_{tsv}}{2 - 2\upsilon}, \sigma_z = 2\sigma_\theta \tag{2.16}$$

where σ_r, σ_θ, and σ_z are radial, circumferential, and axial stresses, respectively. α_{tsv} is the CTE of TSVs and α_{si} represents the CTE of silicon. T_{tsv} is the thermal load on TSV, E is the Young's modulus and υ is the Poisson's ratio.[1]

TSVs thermal load estimation during design-time is usually based on accurate thermal modeling of TSVs and TSV temperature is used to represent the corresponding thermal load assuming the stress-free temperature is at room temperature. Both vertical high thermal conductivity and lateral thermal blockage effect [6] should be considered in the TSV model for more accurate temperature modeling. The lateral thermal blockage effect is due to the relatively low thermal conductivity of dielectric layer surrounding TSV. For example, the normal dielectric layer material is SiO_2 with thermal conductivity of $1.4\,\mathrm{W\,m^{-1}\,K}$ compared to the silicon thermal conductivity of $149\,\mathrm{W\,m^{-1}\,K}$. Therefore, the thermal resistances exist between lateral TSV walls and all neighboring blocks, resulting in high thermal resistances on the lateral thermal dissipation path.

In general, the thermal resistance of TSV farms can be captured by:

$$R_{TSV} = \frac{h}{k \cdot A} \tag{2.17}$$

where h is the material thickness, k is the thermal conductivity of the material per volume, and A is the cross sectional area where heat flow passes through. Note that this equation can be used to calculate both vertical and lateral thermal resistance of TSVs. For vertical thermal resistance calculation, the TSV metal thermal conductivity is used, otherwise, lateral TSV farm thermal conductivity (including low thermal conductivity insulator) is adopted. To this end, the lateral heat blockage effect has been taken into account during design-time floorplan and the TSV thermal stress is proportional to the thermal load on TSVs.

[1] In this formula, the difference of elastic between materials is omitted for simplicity.

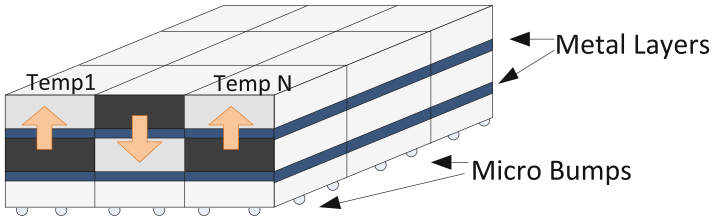

Fig. 2.5 Stack level thermal cycling effect in 3D structure. Thermal stresses are pointing from hot blocks (*dark color*) to cool blocks (*light color*). Alternating direction of stresses (the *arrows*) easily cause cracking on thinned substrate

2.4.2 3D Thermal Cycling Effects

Thermal cycling effect is another factor that can cause reliability issues for 3D ICs [36, 61]. Particularly, the thermal cycling effects in 3D ICs become prominent because the dynamic thermal gradients and stresses in x, y, z directions during run-time can no longer be ignored. Moreover, the thermal cycling effects are more complicated since each functional block now has two more proximity blocks in the vertical direction. As shown in Fig. 2.5 [61], the generated thermal expansion forces are highlighted by arrows, which are from the hotter blocks (dark color) to the cooler blocks (light color). When the force direction varies in the stacked chips, it makes the thinned silicon substrate more vulnerable to damage. A run-time thermal cycling management scheme should be proposed to eliminate the damaging thermal cycling pattern.

Most of the traditional 2D and 3D techniques can not mitigate the problem because they only strive to minimize the peak temperature on chip but disregard the thermomechanical stresses. Sometimes the traditional thermal management techniques can even worsen the problem by wrongly forcing the thermal patterns to exert maximum stress on the device layers in checkerboard configurations, where cold and hot structures are overlaid. In [61], the analysis of vertical thermal cycling pattern and temperature gradients between neighbors in x, y, z directions is employed as part of the thermomechanical stresses management scheme. As a result, the management scheme can alleviate the temperature gradients on cell granularity and achieve mechanical equilibrium.

2.5 Designing 3D Processor Architecture

The following subsections discuss various architecture design approaches that leverage different benefits that 3D integration technology can offer, namely, wirelength reduction, high memory bandwidth, heterogeneous integration, and cost reduction. They also briefly review 3D network-on-chip architecture designs.

2.5.1 Wirelength Reduction

Designers have resorted to technology scaling to improve microprocessor performance. Although the size and switching speed of transistors benefit as technology feature sizes continue to shrink, global interconnect wire delay does not scale accordingly with technologies. The increasing wire delays have become one major impediment to performance improvement.

Three-dimensional integrated circuits (3D ICs) are attractive options for overcoming the barriers to interconnect scaling, thereby offering an opportunity to continue performance improvements using CMOS technology. Compared to a traditional two dimensional chip design, one of the important benefits of a 3D chip over a traditional two-dimensional (2D) design is the reduction on global interconnects. It has been shown that three-dimensional architectures reduce wiring length by a factor of the square root of the number of layers used [20]. The reduction of wire length due to 3D integration can result in two obvious benefits: *latency improvement* and *power reduction*.

2.5.1.1 Latency Improvement

Latency improvement can be achieved due to the reduction of average interconnect length and the critical path length.

Early work on fine-granularity 3D partitioning of processor components shows that the latency of 3D components could be reduced. For example, since interconnects dominate the delay of cache accesses which determines the critical path of a microprocessor, and the regular structure and long wires in a cache make it one of the best candidates for 3D designs, 3D cache design is one of the early design example for fine-granularity 3D partition [56]. Wordline partitioning and bitline partitioning approaches divide a cache bank into multiple layers and reduce the global interconnects, resulting in fast cache access time. Depending on the design constraints, the 3DCacti tool [47] automatically explores the design space for a cache design, and finds out the optimal partitioning strategy, and the latency reduction can be as much as 25 % for a two-layer 3D cache. 3D arithmetic-component designs also show latency benefits. For example, various designs [15, 37, 39, 48] have shown that the 3D arithmetic unit design can achieve around 6–30 % delay reduction due to the wire length reduction. Such fine-granularity 3D partitioning was also demonstrated by Intel [4], showing that by targeting the heavily pipelined wires, the pipeline modifications resulted in approximately 15 % improved performance, when the Intel Pentium-4 processor was folded onto 2-layer 3D implementation.

Note that such fine-granularity design of 3D processor components increases the design complexity, and the latency improvement varies depending on the partitioning strategies and the underlying 3D process technologies. For example, for the same Kogge–Stone adder design, a partitioning based on logic level [48] demonstrates that the delay improvement diminishes as the number of 3D layers increases;

while a bit-slicing partitioning [37] strategy would have better scalability as the bit-width or the number of layers increases. Furthermore, the delay improvement for such bit-slicing 3D arithmetic units is about 6 % when using a bulk-CMOS-based 180 nm 3D process [15], while the improvement could be as much as 20 % when using a SOI-based 180 nm 3D process technology [37], because the SOI-based process has much smaller and shorter TSVs (and therefore much smaller TSV delay) compared to the bulk-CMOS-based process.

2.5.1.2 Power Reduction

Interconnect power consumption becomes a large portion of the total power consumption as technology scales. The reduction of the wire length translates into power savings in 3D IC design. For example, 7–46 % of power reduction for 3D arithmetic units were demonstrated in [37]. In the 3D Intel Pentium-4 implementation [4], because of the reduction in long global interconnects, the number of repeaters and repeating latches in the implementation is reduced by 50 %, and the 3D clock network has 50 % less metal RC than the 2D design, resulting in a better skew, jitter and lower power. Such 3D stacked redesign of Intel Pentium 4 processor improves performance by 15 % and reduces power by 15 % with a temperature increase of 14°. After using voltage scaling to lower the peak temperature to be the same as the baseline 2D design, their 3D Pentium 4 processor still showed a performance improvement of 8 %.

2.5.2 Memory Bandwidth Improvement

It has been shown that circuit limitations and limited instruction level parallelism will diminish the benefits of modern superscalar microprocessors by increased architectural complexity, which leads to the advent of Chip Multiprocessors (CMP) as a viable alternative to the complex superscalar architecture. The integration of multi-core or many-core microarchitecture on a single die is expected to accentuate the already daunting memory-bandwidth problem. Supplying enough data to a chip with a massive number of on-die cores will become a major challenge for performance scalability. Traditional off-chip memory will not suffice due to the I/O pin limitations. Three-dimensional integration has been envisioned as a solution for future micro-architecture design (especially for multi-core and many-core architectures), to mitigate the interconnect crisis and the "memory wall" problem [18,30,31]. It is anticipated that memory stacking on top of logic would be one of the early commercial uses of 3D technology for future chip-multiprocessor design, by providing improved memory bandwidth for such multi-core/many-core microprocessors. In addition, such approaches of memory stacking on top of

core layers do not have the design complexity problem as demonstrated by the fine-granularity design approaches, which require re-designing all processor components for wire length reduction.

Intel [4] explored the memory bandwidth benefits using a base-line Intel Core2 Duo processor, which contains two cores. By having memory stacking, the on-die cache capacity is increased, and the performance is improved by capturing larger working sets, reducing off-chip memory bandwidth requirements. For example, one option is to stack an additional 8 MB L2 cache on top of the base-line 2D processor (which contains 4 MB L2 cache), and the other option is to replace the SRAM L2 cache with a denser DRAM L2 cache stacking. Their study demonstrated that a 32 MB 3D stacked DRAM cache can reduce the cycles per memory access by 13 % on average and as much as 55 % with negligible temperature increases.

PicoServer project [25] follows a similar approach to stack DRAM on top of multi-core processors. Instead of using stacked memory as a larger L2 cache (as shown by Intel's work [4]), the fast on-chip 3D stacked DRAM main memory enables wide low-latency buses to the processor cores and eliminates the need for an L2 cache, whose silicon area is allocated to accommodate more cores. Increasing the number of cores by removing the L2 cache can help improve the computation throughput, while each core can run at a much lower frequency, and therefore result in an energy-efficient many core design. For example, it can achieve a 14 % performance improvement and 55 % power reduction over a baseline multi-core architecture.

As the number of the cores on a single die increases, such memory stacking becomes more important to provide enough memory bandwidth for processor cores. Recently, Intel [49] demonstrated an 80-tile terascale chip with network-on-chip. Each core has a local 256 KB SRAM memory (for data and instruction storage) stacked on top of it. TSVs provide a bandwidth of 12 GB/s for each core, with a total about 1 TB/s bandwidth for Tera Flop computation. In this chip, the thin memory die is put on top of the CPU die, and the power and I/O signals go through memory to CPU.

Since DRAM is stacked on top of the processor cores, the memory organization should also be optimized to fully take advantages of the benefits that TSVs offer [29, 31]. For example, the numbers of ranks and memory controllers are increased, in order to leverage the memory bandwidth benefits. A multiple-entry row buffer cache is implemented to further improve the performance of the 3D main memory. Comprehensive evaluation shows that a 1.75× speedup over commodity DRAM organization is achieved [31]. In addition, the design of MSHR was explored to provided a scalable L2 miss handling before accessing the 3D stacked main memory. A data structure called the Vector Bloom Filter with dynamic MSHR capacity tuning is proposed. Such structure provides an additional 17.8 % performance improvement. If stacked DRAM is used as the last-level caches (LLC) in chip multiple processors (CMPs), the DRAM cache sets are organized into multiple queues [29]. A replacement policy is proposed for the queue-based cache

to provide performance isolation between cores and reduce the lifetimes of dead cache lines. Approaches are also proposed to dynamically adapt the queue size and the policy of advancing data between queues.

The latency improvement due to 3D technology can also be demonstrated by such memory stacking design. For example, Li et al. [28] proposed a 3D chip multiprocessor design using network-in-memory topology. In this design, instead of partitioning each processor core or memory bank into multiple layers (as shown in [47, 56]), each core or cache bank remains to be a 2D design. Communication among cores or cache banks are via the network-on-chip (NoC) topology. The core layer and the L2 cache layer are connected with TSV-based bus. Because the short distance between layers, TSVs provide a fast access from one layer to another layer, and effectively reduce the cache access time because of the faster access to cache banks through TSVs.

2.5.3 Heterogenous Integration

3D integration also provides new opportunities for future architecture design, with a new dimension of design space exploration. In particular, the heterogenous integration capability enabled by 3D integration gives designers new perspective when designing future CMPs.

3D integration technologies provide feasible and cost-effective approaches for integrating architectures composed of heterogeneous technologies to realize future microprocessors targeted at the "More than Moore" technology projected by ITRS. 3D integration supports heterogeneous stacking because different types of components can be fabricated separately, and layers can be implemented with different technologies. It is also possible to stack optical device layers or non-volatile memories [such as magnetic RAM (MRAM) or phase-change memory (PCRAM)] on top of microprocessors to enable cost-effective heterogeneous integration. The addition of new stacking layers composed of new device technology will provide greater flexibility in meeting the often conflicting design constraints (such as performance, cost, power, and reliability), and enable innovative designs in future microprocessors.

2.5.3.1 Non-volatile Memory Stacking

Stacking layers of non-volatile memory technologies such as Magnetic Random Access Memory (MRAM) [13] and Phase Change Random Access Memory (PRAM) [53] on top of processors can enable a new generation of processor architectures with unique features. There are several characteristics of MRAM and PRAM architectures that make them promising candidates for on-chip memory. In addition to their non-volatility, they have zero standby power, low access power and are immune to radiation-induced soft errors. However, integrating these non-volatile

memories along with a logic core involves additional fabrication challenges that need to be overcome (for example, MRAM process requires growing a magnetic stack between metal layers). Consequently, it may incur extra cost and additional fabrication complexity to integrate MRAM with conventional CMOS logic into a single 2D chip. The ability to integrate two different wafers developed with different technologies using 3D stacking offers an ideal solution to overcome this fabrication challenge and exploit the benefits of PRAM and MRAM technologies. For example, Sun et al. [46] demonstrated that the optimized MRAM L2 cache on top of multi-core processor can improve performance by 4.91 % and reduce power by 73.5 % compared to the conventional SRAM L2 cache with similar area.

2.5.3.2 Optical Device Layer Stacking

Even though 3D memory stacking can help mitigate the memory bandwidth problem, when it comes to off-chip communication, the pin limitations, the energy cost of electrical signaling, and the non-scalability of chip-length global wires are still significant bandwidth impediments. Recent developments in silicon nanophotonic technology have the potential to meet the off-chip communication bandwidth requirements at acceptable power levels. With the heterogeneous integration capability that 3D technology offers, one can integrate optical die together with CMOS processor dies. For example, HP Labs proposed a Corona architecture [50], which is a 3D many-core architecture that uses nanophotonic communication for both inter-core communication and off-stack communication to memory or I/O devices. A photonic crossbar fully interconnects its 256 low-power multithreaded cores at 20 TB/s bandwidth, with much lower power consumption.

Figure 2.6 illustrates such a 3D heterogenous processor architecture, which integrates non-volatile memories and optical die together through 3D integration technology.

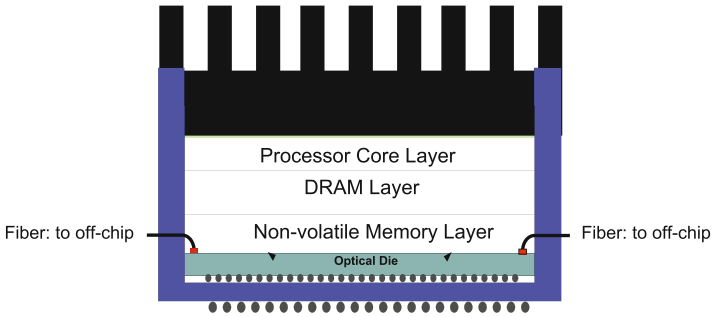

Fig. 2.6 An illustration of 3D heterogeneous architecture with non-volatile memory stacking and optical die stacking

2.5.4 Cost-Effective Architecture

Increasing integration density has resulted in large die size for microprocessors. With a constant defect density, a larger die typically has a lower yield. Consequently, partitioning a large 2D microprocessor to be multiple smaller dies and stacking them together may result in a much higher yield for the chip, even though 3D stacking incurs extra manufacturing cost due to extra steps for 3D integration and may cause a yield loss during stacking. Depending on the original 2D microprocessor die size, it may be cost-effective to implement the chip using 3D stacking [12], especially for large microprocessors. The heterogenous integration capability that 3D provides can also help reduce the cost.

In addition, as technology feature size scales to reach the physical limits, it has been predicted that moving to the next technology node is not only difficult but also prohibitively expensive. 3D stacking can potentially provide a cost-effective integration solution, compared to traditional technology scaling.

2.5.5 3D NoC Architecture

Network-on-chip (NoC) is a general purpose on-chip interconnection network architecture that is proposed to replace the traditional design-specific global on-chip wiring, by using switching fabrics or routers to connect processor cores or processing elements (PEs). Typically, the PEs communicate with each other using a packet-switched protocol. Even though both 3D integrated circuits and NoCs are proposed as alternatives for the interconnect scaling demands, the challenges of combining both approaches to design three-dimensional NOCs have not been addressed until recently [14, 27, 28]. Researchers have studied various NoC router design with 3D integration technology. For example, various design options for the NoC router for 3D NoC has been investigated: (1) symmetric NoC router design with a simple extension to the 2D NoC router; (2) NoC-bus hybrid router design which leverages the inherent asymmetry in the delays in a 3D architecture between the fast vertical interconnects and the horizontal interconnects that connect neighboring cores; (3) True 3D router design with major modification as dimensionally-decomposed router [27]; (4) Multi-layer 3D NoC router design which partitions a single router to multiple layers to boost the performance and reduce the power consumption [14]. 3D NoC topology design was also investigated [60]. More details can be found in [5].

2.6 Challenges for 3D Architecture Design

Even though 3D integrated circuits show great benefits, there are several challenges for the adoption of 3D technology for future architecture design: (1) *Thermal management.* The move from 2D to 3D design could accentuate the thermal concerns due to the increased power density. To mitigate the thermal impact, thermal-aware design techniques must be adopted for 3D architecture design [56]; (2) *Design Tools and methodologies.* 3D integration technology will not be commercially viable without the support of EDA tools and methodologies that allow architects and circuit designers to develop new architectures or circuits using this technology. To efficiently exploit the benefits of 3D technologies, design tools and methodologies to support 3D designs are imperative [57]; (3) *Testing.* One of the barriers to 3D technology adoption is insufficient understanding of 3D testing issues and the lack of design-for-testability (DFT) techniques for 3D ICs, which have remained largely unexplored in the research community.

References

1. Alam S, Jones R, Rauf S, Chatterjee R. Inter-strata connection characteristics and signal transmission in three-dimensional (3D) integration technology. In: International symposium on quality electronic design, 2007.
2. Athikulwongse K, Chakraborty A, Yang JS, Pan D, Lim SK. Stress-driven 3D-IC placement with TSV keep-out zone and regularity study. In: International conference on computer-aided design, 2010.
3. Barnat S, Fremont H, Gracia A, Cadalen E, Bunel C, Neuilly F, Tenailleau J. Design for reliability: Thermo-mechanical analyses of stress in through silicon via. In: International conference on thermal, mechanical multi-physics simulation, and experiments in microelectronics and microsystems, 2010.
4. Black B, et al. Die stacking 3D microarchitecture. In: MICRO, 2006. pp. 469–79.
5. Carloni L, Pande P, Xie Y. Networks-on-chip in emerging intercoonect paradigms: advantages and challenges. In: Intl. symp. on networks-on-chips, 2009.
6. Chen Y, Kursun E, Motschman D, Johnson C, Xie Y. Analysis and mitigation of lateral thermal blockage effect of through-silicon-via in 3D IC designs. In: International symposium on low power electronics and design, 2011.
7. Chien HC, Lau JH, Chao YL, Tain RM, Dai MJ, Lo WC, Kao MJ. Estimation for equivalent thermal conductivity of silicon-through vias TSV used for 3D IC integration. In: International microsystems, packaging, assembly and circuit technology conference, 2011.
8. Cho J, Song E, Yoon K, Pak JS, Kim J, Lee W, Song T, Kim K, Lee J, Lee H, Park K, Yang S, Suh M, Byun K, Kim J. Modeling and analysis of through-silicon via (TSV) noise coupling and suppression using a guard ring. IEEE Trans Compon Packag Manuf Technol. 2011;1:220–33.
9. Cong J, Luo G, Wei J, Zhang Y. Thermal-aware 3D IC placement via transformation. In: Asia and South Pacific design automation conference, 2007.
10. Cong J, Luo G, Shi Y. Thermal-aware cell and through-silicon-via co-placement for 3D ICs. In: Design automation conference, 2011.
11. Davis WR, Wilson J, Mick S, Xu J, Hua H, Mineo C, Sule AM, Steer M, Franzon PD. Demystifying 3D ICs: the pros and cons of going vertical. IEEE Des Test Comput. 2005;22(6):498–510.

12. Dong X, Xie Y. Cost analysis and system-level design exploration for 3D ICs. In: Asia and South Pacific design automation conference, 2009.
13. Dong X, Wu X, Sun G, Xie Y, Li H, Chen Y. Circuit and microarchitecture evaluation of 3D stacking Magnetic RAM (MRAM) as a universal memory replacement. In: Design automation conference, 2009. pp. 554–9.
14. Dongkook P, Eachempati S, Das R, Mishra AK, Xie Y, Vijaykrishnan N, Das CR. MIRA: a multi-layered on-chip interconnect router architecture. In: International symposium on computer architecture, 2008. pp. 251–61
15. Egawa R, Tada J, Kobayashi H, Goto G. Evaluation of fine grain 3D integrated arithmetic units. In: IEEE international 3D system integration conference, 2009.
16. Garrou P. Handbook of 3D integration: technology and applications using 3D integrated circuits. Wiley-CVH, chap Introduction to 3D integration, 2008.
17. Goplen B, Sapatnekar S. Thermal via placement in 3D ICs. In: International symposium on physical design, 2005.
18. Jacob P, et al. Mitigating memory wall effects in high clock rate and multi-core CMOS 3D ICs: processor memory stacks. Proc IEEE 2008;96(10):5.
19. Jain A, Jones R, Chatterjee R, Pozder S. Analytical and numerical modeling of the thermal performance of three-dimensional integrated circuits. IEEE Trans Compon Packag Technol. 2010;33(1):56–63.
20. Joyner J, Zarkesh-Ha P, Meindl J. A stochastic global net-length distribution for a three-dimensional system-on-a-chip (3D-SoC). In: International ASIC/SOC conference, 2001.
21. Jung M, Mitra J, Pan D, Lim SK. TSV stress-aware full-chip mechanical reliability analysis and optimization for 3D IC. In: Design automation conference, 2011
22. Katti G, Mercha A, Stucchi M, Tokei Z, Velenis D, Van Olmen J, Huyghebaert C, Jourdain A, Rakowski M, Debusschere I, Soussan P, Oprins H, Dehaene W, De Meyer K, Travaly Y, Beyne E, Biesemans S, Swinnen B. Temperature dependent electrical characteristics of through-si-via (TSV) interconnections. In: International interconnect technology conference, 2010.
23. Katti G, Stucchi M, De Meyer K, Dehaene W. Electrical modeling and characterization of through silicon via for three-dimensional ICs. IEEE Trans Electron Devices. 2010;57:256–62.
24. Katti G, Stucchi M, Velenis D, Soree B, De Meyer K, Dehaene W. Temperature-dependent modeling and characterization of through-silicon via capacitance. IEEE Electron Device Lett. 2011;32:563–5.
25. Kgil T, D'Souza S, Saidi A, Binkert N, Dreslinski R, Mudge T, Reinhardt S, Flautner K. PicoServer: using 3D stacking technology to enable a compact energy efficient chip multiprocessor. In: ASPLOS, 2006. pp. 117–28.
26. Khalil D, Ismail Y, Khellah M, Karnik T, De V. Analytical model for the propagation delay of through silicon vias. In: International symposium on quality electronic design, 2008.
27. Kim J, Nicopoulos C, Park D, Das R, Xie Y, Vijaykrishnan N, Das C. A novel dimensionally-decomposed router for on-chip communication in 3D architectures. In: International symposium on computer architecture, 2007.
28. Li F, Nicopoulos C, Richardson T, Xie Y, Vijaykrishnan N, Kandemir M. Design and management of 3D chip multiprocessors using network-in-memory. In: International symposium on computer architecture, 2006.
29. Loh G. Extending the effectiveness of 3D-stacked DRAM caches with an adaptive multi-queue policy. In: International symposium on microarchitecture, 2009.
30. Loh G, Xie Y, Black B. Processor design in three-dimensional die-stacking technologies. IEEE Micro. 2007;27(3):31–48.
31. Loh GH. 3D-stacked memory architectures for multi-core processors. In: International symposium on computer architecture, 2008.
32. Lu KH, Zhang X, Ryu SK, Im J, Huang R, Ho P. Thermo-mechanical reliability of 3-D ICs containing through silicon vias. In: Electronic components and technology conference, 2009.
33. Lu KH, Ryu SK, Zhao Q, Zhang X, Im J, Huang R, Ho PS. Thermal stress induced delamination of through silicon vias in 3D interconnects. In: Electronic components and technology conference, 2010.

34. Lu KH, Ryu SK, Im J, Huang R, Ho P. Thermomechanical reliability of through-silicon vias in 3D interconnects. In: International reliability physics symposium, 2011.
35. Majeed B, Sabuncuoglu Tezcan D, Vandevelde B, Duval F, Soussan P, Beyne E. Electrical characterization, modeling and reliability analysis of a via last TSV. In: Electronics packaging technology conference, 2010.
36. Noritake C, Limaye P, Gonzalez M, Vandevelde B. Thermal cycle reliability of 3D chip stacked package using PB-free solder bumps: Parameter study by FEM analysis. In: International conference on thermal, mechanical and multi-physics simulation and experiments in micro-electronics and microsystems, 2006.
37. Ouyang J, Sun G, Chen Y, Duan L, Zhang T, Xie Y, Irwin M. Arithmetic unit design using 180 nm TSV-based 3D stacking technology. In: international 3D system integration conference, 2009.
38. Pak JS, Ryu C, Kim J. Electrical characterization of through silicon via (TSV) depending on structural and material parameters based on 3D full wave simulation. In: International conference on electronic materials and packaging, 2007.
39. Puttaswamy K, Loh GH. Scalability of 3D-integrated arithmetic units in high-performance microprocessors. In: Design automation conference, 2007.
40. Roullard J, Capraro S, Farcy A, Lacrevaz T, Bermond C, Leduc P, Charbonnier J, Ferrandon C, Fuchs C, Flechet B. Electrical characterization and impact on signal integrity of new basic interconnection elements inside 3D integrated circuits. In: Electronic components and technology conference, 2011.
41. Ryu C, Chung D, Lee J, Lee K, Oh T, Kim J. High frequency electrical circuit model of chip-to-chip vertical via interconnection for 3-D chip stacking package. In: Topical meeting on electrical performance of electronic packaging, 2005.
42. Ryu SK, Lu KH, Zhang X, Im JH, Ho P, Huang R. Impact of near-surface thermal stresses on interfacial reliability of through-silicon vias for 3-D interconnects. IEEE Trans Device Mater Reliab. 2011;11:35–43.
43. Salah K, El Rouby A, Ragai H, Amin K, Ismail Y. Compact lumped element model for TSV in 3D-ICs. In: International symposium on circuits and systems, 2011.
44. Savidis I, Friedman E. Closed-form expressions of 3-D via resistance, inductance, and capacitance. IEEE Trans Electron Devices. 2009;56:1873–81.
45. Selvanayagam C, Lau J, Zhang X, Seah S, Vaidyanathan K, Chai T. Nonlinear thermal stress/strain analyses of copper filled TSV (through silicon via) and their flip-chip microbumps. IEEE Trans Adv Packag. 2009;32(4):720–8.
46. Sun G, Dong X, Xie Y, Li J, Chen Y. A novel 3D stacked MRAM cache architecture for CMPs. In: International symposium on high performance computer architecture, 2009.
47. Tsai YF, Wang F, Xie Y, Vijaykrishnan N, Irwin MJ. Design space exploration for three-dimensional cache. IEEE Trans Very Large Scale Integr VLSI Syst. 2008;16(4):444–55.
48. Vaidyanathan B, Hung WL, Wang F, Xie Y, Narayanan V, Irwin MJ. Architecting microprocessor components in 3D design space. In: Intl. conf. on VLSI design, 2007.
49. Vangal S, et al. An 80-tile Sub-100-W TeraFLOPS processor in 65-nm CMOS. IEEE J Solid State Circuits. 2008;43(1):29–41.
50. Vantrease D, Schreiber R, Monchiero M, McLaren M, Jouppi NP, Fiorentino M, Davis A, Binkert N, Beausoleil RG, Ahn JH. Corona: system implications of emerging nanophotonic technology. In: international symposium on computer architecture, 2008.
51. Wang F, Zhu Z, Yang Y, Wang N. A thermal model for the top layer of 3D integrated circuits considering through silicon vias. In: International conference on ASIC, 2011.
52. Weerasekera R, Grange M, Pamunuwa D, Tenhunen H, Zheng LR. Compact modelling of through-silicon vias (TSVs) in three-dimensional (3-D) integrated circuits. In: International conference on 3D system integration, 2009.
53. Wu X, Li J, Zhang L, Speight E, Xie Y. Hybrid cache architecture. In: International symposium on computer architecture, 2009.

54. Wu X, Zhao W, Nakamoto M, Nimmagadda C, Lisk D, Gu S, Radojcic R, Nowak M, Xie Y. Electrical characterization for intertier connections and timing analysis for 3-D ICs. IEEE Trans Very Large Scale Integr VLSI Syst. 2012;20:186–91.
55. Xie J, Chung D, Swaminathan M, Mcallister M, Deutsch A, Jiang L, Rubin B. Electrical-thermal co-analysis for power delivery networks in 3D system integration. In: International conference on 3D system integration, 2009.
56. Xie Y, Loh G, Black B, Bernstein K. Design space exploration for 3D architectures. ACM J Emerg Technol Comput Syst. 2006;2:65–103.
57. Xie Y, Cong J, Sapatnekar S. Three-dimensional integrated circuit design: EDA, design and microarchitectures. Springer: New York, 2009.
58. Xu C, Li H, Suaya R, Banerjee K. Compact ac modeling and analysis of Cu, W, and CNT based through-silicon vias (TSVs) in 3-D ICs. In: International electron devices meeting, 2009.
59. Xu C, Li H, Suaya R, Banerjee K. Compact ac modeling and performance analysis of through-silicon vias in 3-D ICs. IEEE Trans Electron Devices. 2010;57:3405–17.
60. Xu Y, et al. A low-radix and low-diameter 3D interconnection network design. In: Intl. symp. on high performance computer architecture, 2009.
61. Zou Q, Zhang T, Kursun E, Xie Y. Thermomechanical stress-aware management for 3D IC designs. In: Design, automation test in Europe conference exhibition, 2013.