Chapter 5
Physical Analysis of NoC Topologies for 3-D Integrated Systems

Vasilis F. Pavlidis and Eby G. Friedman

5.1 Introduction

Several new topologies for on-chip interconnect networks are supported by vertical integration. These three-dimensional topologies improve the performance of an on-chip network primarily in two ways. The length of the physical links connecting the switches of the network is shorter. Additionally, the data can be routed across the on-chip network through a smaller number of switches. The three-dimensional (3-D) NoC topologies include two types of physical links implemented with horizontal and vertical interconnects. These links exhibit substantially different physical and electrical characteristics. The different 3-D topologies and timing and power models that describe the performance of the resulting 3-D networks are discussed in this chapter.

These models emphasize the physical characteristics rather than the architectural details of the network. These models provide useful bounds for improving the performance of on-chip networks by exploiting the third dimension. With these models, the topology that minimizes the latency or power consumption of a network can be determined. As described in this chapter, a network topology can typically enhance one of these two primary design objectives.

The thermal behavior of 3-D integrated systems is another important issue due to the increased power densities that can develop. To characterize the thermal effects on the performance of 3-D NoC topologies, the timing and power models are enhanced by including the dependence on temperature of specific parameters, such as the electrical resistance of an interconnect. Consequently, the topology that produces the minimum rise in temperature at the plane located farthest from the heat sink of the system can be selected while satisfying specific performance characteristics.

V. F. Pavlidis (✉)
Integrated Systems Laboratory, EPFL, 1015 Lausanne, Switzerland
e-mail: vasileios.pavlidis@epfl.ch

E. G. Friedman (✉)
University of Rochester, Rochester, NY 14627, USA
e-mail: friedman@ece.rochester.edu

A. Sheibanyrad et al. (eds.), *3D Integration for NoC-based SoC Architectures*,
Integrated Circuits and Systems, DOI 10.1007/978-1-4419-7618-5_5,
© Springer Science+Business Media, LLC 2011

A first-order thermal model is utilized to determine the rise in temperature in each plane within a 3-D system. Elevated temperatures can affect the performance of the processing elements (PEs) in addition to the performance of the network. Consequently, an enhanced analysis methodology including thermal effects, which provides the interconnect architecture employed in the PEs, is described in this chapter. In this methodology, the number of metal layers, pitch of the interconnect in each metal layer, and number of physical planes are considered. In addition, the rise in temperature due to the heat generated by transistor switching and joule heating of the wires is evaluated. In other words, the methodology described in this chapter provides a means to estimate early in the design cycle the behavior of a 3-D topology for an integrated system interconnected with an on-chip network.

This holistic approach in the design of 3-D systems based on networks-on-chip is necessary, as demonstrated in this chapter. Neglecting the effects of the power consumption of the PEs and the related temperature rise can produce a misleading result when selecting the 3-D NoC topology that exhibits the highest performance. The analysis approach presented in this chapter is applied to homogeneous 3-D NoCs (i.e., all of the PEs are assumed identical) exploring diverse objectives, such as speed, power, and temperature. As demonstrated in this analysis, a primary criterion for the design of these 3-D systems is whether the third dimension is used for the on-chip network or the PEs.

In the next section, the 3-D NoC topologies investigated in this chapter are described and some notation is introduced. Timing and power models for the on-chip network are presented in Section 5.3. A technique for determining the wiring resources of the PEs within a 3-D system interconnected with an on-chip network and the resulting rise in temperature in the different 3-D NoC topologies is presented in Section 5.4. Several tradeoffs among different characteristics of the topologies including the network size, number of physical planes, and operating frequency of the PEs are discussed in Section 5.5. The primary objectives of the chapter are summarized in the last section of the chapter.

5.2 Three-Dimensional On-Chip Network Topologies

Primary topologies for 3-D networks are presented and related terminology is introduced in this section. Mesh structures have been a popular network topology for conventional 2-D NoC [1–3]. A fundamental element of a mesh network is illustrated in Fig. 5.1a, where each processing element (PE) is connected to the network through a switch. A PE can be integrated either on a single physical plane (2-D IC) or on several physical planes (3-D IC). Each switch in a 2-D NoC is connected to a neighboring switch in one of four directions. Consequently, each switch has five ports. Alternatively, in a 3-D NoC, the switch typically connects to two additional neighboring switches located on the adjacent physical planes. The switch architecture is considered here to be a canonical switch with input and output buffering [4].

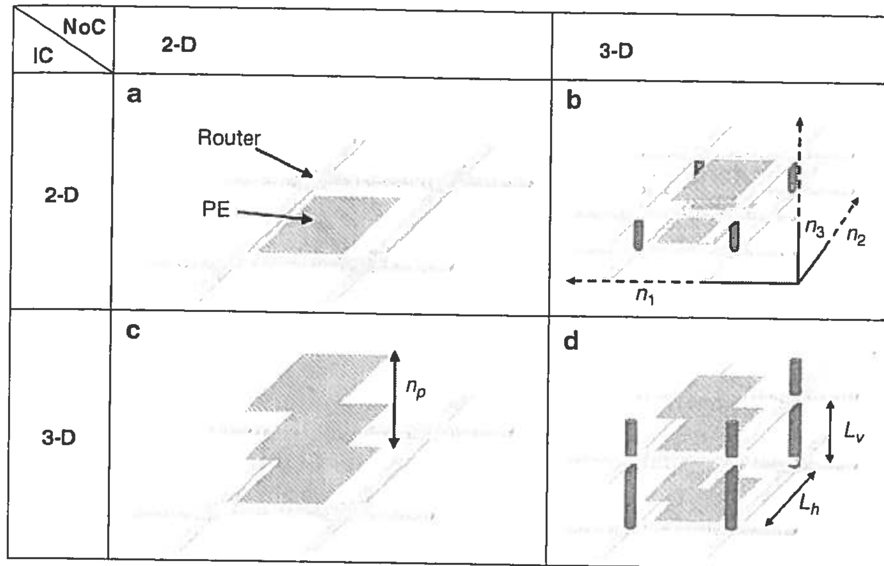

Fig. 5.1 Different NoC topologies (not to scale), (**a**) 2-D IC – 2-D NoC, (**b**) 2-D IC – 3-D NoC, (**c**) 3-D IC – 2-D NoC, and (**d**) 3-D IC – 3-D NoC [8]

Although a network switch can also be designed in a 3-D manner [5], herein, the switches are considered as two-dimensional (i.e., occupy a single physical plane). Note that if 3-D switches are utilized, the performance of each of the targeted topologies will be improved equally. Consequently, a comparison of the 3-D network topologies is independent of whether a 2-D or 3-D switch is used. The combination of a PE and switch is a network node. For a 2-D mesh network, the total number of nodes N is $n_1 \times n_2$, where n_i is the number of nodes included in the ith physical dimension.

Integration in the third dimension introduces a variety of topological choices for NoCs. For a 3-D NoC, as shown in Fig. 5.1b, the total number of nodes is $N = n_1 \times n_2 \times n_3$, where n_3 is the number of nodes in the third dimension. In this topology, each PE is on a single yet possibly different physical plane (2-D IC – 3-D NoC). Alternatively, a PE can be implemented on only one of the n_3 physical planes. The 3-D system, therefore, contains $n_1 \times n_2$ PEs on each of the n_3 physical planes, where the total number of nodes is N. This topology is discussed in [6, 7]. A 3-D NoC topology is proposed in Fig. 5.1c, where the interconnect network is contained within one physical plane (i.e., $n_3 = 1$), while each PE is integrated on multiple planes, notated as n_p (3-D IC – 2-D NoC). Finally, a hybrid 3-D NoC based on the two previous topologies is proposed in Fig. 5.1d. In such an NoC, both the interconnect network and the PEs can span more than one physical plane of the stack (3-D IC – 3-D NoC). In the following section, latency and power expressions for each of the NoC topologies are presented, assuming a zero-load model.

5.3 Timing and Power Model for 3-D NoCs

In this section both timing and power dissipation models for on-chip networks are described. Modeling specific parameters as a function of temperature is also discussed to investigate thermal effects on different 3-D NoC topologies. The latency and power dissipation models are presented in Sections 5.3.1 and 5.3.2, respectively.

5.3.1 Latency Model for 3-D NoC

In this section, analytic models of the zero-load latency of each of the 3-D NoC topologies are described. Zero-load network latency is widely used as a performance metric in traditional interconnection networks [10]. The zero-load latency of a network is the latency where only one packet traverses the network at any one time. Although this model does not consider contention among packets, the zero-load latency can be used to describe the effect of a topology on the performance of a network. The zero-load latency of an NoC with wormhole switching is [10]

$$T_{network} = hops \cdot t_r + t_c + \frac{L_p}{b}, \tag{5.1}$$

where the first term represents the routing delay, t_c is the propagation delay along the wires of the physical link, which is also called a buss here for simplicity, and the third term is the serialization delay of the packet. *hops* is the average number of switches that a packet traverses to reach the destination node, t_r is the switch delay, L_p is the length of the packet in bits, and b is the bandwidth of the buss defined as $b \equiv w_c f_c$, where w_c is the width of the link in bits and f_c is the inverse of the delay of a bit propagating along the longest physical link.

Since the number of planes that can be stacked in a 3-D NoC is constrained by the target technology, n_3 is also constrained. Furthermore, n_1, n_2, and n_3 are not necessarily equal. The average number of hops in a 3-D NoC is

$$hops = \frac{n_1 n_2 n_3 (n_1 + n_2 + n_3) - n_3(n_1 + n_2) - n_1 n_2}{3(n_1 n_2 n_3 - 1)}, \tag{5.2}$$

assuming dimension-order routing to ensure that minimum distance paths are used for routing packets between any source-destination node pair.

Although the average number of hops provides a useful expression for a latency model of an NoC, (5.2) does not characterize all possible traffic scenarios within a network. For example, while uniform traffic among the PEs can be modeled by the average number of hops, applications that favor localized traffic will result in a different number of hops from (5.2). This situation does not lessen the applicability of the models described in this section as long as an expression for the number of hops can be determined. In addition, synthesis tools for on-chip networks typi-

cally utilize a zero-load model to determine the number of hops within the network, thereby determining the latency of each synthesized topology [9]. In the case, where the network traffic cannot accurately be described in closed form, the inherent characteristics of the topologies depicted in Fig. 5.1 can guide the selection process for a specific topology. For example, consider a highly localized traffic scenario. The vertical channels can be used primarily for high bandwidth data transfer since the vertical links exhibit a considerably lower delay as compared to the horizontal links. This difference in latency suggests that a 2-D IC – 3-D NoC is a better candidate than a 3-D IC – 2-D NoC topology, where the network includes only horizontal links. The criterion for choosing this topology, in this case, would be the short vertical links, not the reduction in the number of hops (due to the larger number of the PEs connected to each switch).

To describe this difference in latency between the two types of links, the average number of hops in (5.2) can be divided into two components, the average number of hops within the two dimensions n_1 and n_2, and the average number of hops within the third dimension n_3,

$$hops_{2-D} = \frac{n_3(n_1 + n_2)(n_1 n_2 - 1)}{3(n_1 n_2 n_3 - 1)}, \tag{5.3}$$

$$hops_{3-D} = \frac{(n_3^2 - 1) n_1 n_2}{3(n_1 n_2 n_3 - 1)}. \tag{5.4}$$

The delay of the switch t_r is the sum of the delay of the arbitration logic t_a and the delay of the switch t_s, which is assumed to be implemented with a classic crossbar switch [10],

$$t_r = t_a + t_s. \tag{5.5}$$

The delay of the arbiter as described in [11] is

$$t_a = \left(21(1/4)\log_2 p + 14(1/12) + 9\right), \tag{5.6}$$

where p is the number of ports of the switch and τ is the delay of a minimum sized inverter for the target technology. Note that (5.6) exhibits a logarithmic dependence on the number of switch ports. The length of the crossbar switch also depends upon the number of switch ports and the width of the buss,

$$l_s = 2(w_t + s_t) w_c p, \tag{5.7}$$

where w_t and s_t are the width and spacing, respectively, or, alternatively, the pitch of the interconnect and w_c is the width of the physical link in bits. Consequently, the worst case delay of the crossbar switch is determined by the longest path within the switch, which is equal to (5.7). The delay of the physical link t_c is

$$t_c = t_v hops_{3-D} + t_h hops_{2-D}, \tag{5.8}$$

where t_v and t_h are the delay of the vertical and horizontal buss, respectively (see Fig. 5.1b). Note that if $n_3 = 1$, (5.8) describes the propagation delay of a 2-D NoC. Substituting (5.5) and (5.8) into (5.1), the overall zero-load network latency for a 3-D NoC is

$$T_{network} = hops(t_a + t_s) + hops_{2-D} t_h + hops_{3-D} t_v + \frac{L_p}{w_c} t_h. \tag{5.9}$$

To characterize t_s, t_h, and t_v, the models described in [12] are adopted, where repeaters implemented as simple inverters are inserted along the interconnect. According to these models, the propagation delay and rise time of a single interconnect stage for a step input, respectively, are

$$t_{di} = 0.377 \frac{r_i c_i l_i^2}{k_i^2} + 0.693 \left(R_{d0} C_0 + \frac{R_{d0} c_i l_i}{h_i k_i} + \frac{r_i l_i C_{g0} h_i}{k_i} \right), \tag{5.10}$$

$$t_{ri} = 1.1 \frac{r_i c_i l_i^2}{k_i^2} + 2.75 \left(R_{r0} C_0 + \frac{R_{r0} c_i l_i}{h_i k_i} + \frac{r_i l_i C_{g0} h_i}{k_i} \right), \tag{5.11}$$

where $r_i(c_i)$ is the per unit length resistance (capacitance) of the interconnect and l_i is the total length of the interconnect.

The index i is used to notate the interconnect delays included in the network (i.e., $i \in \{s, v, h\}$). h_i and k_i denote the size and number of repeaters, respectively, and C_{g0} and C_0 represent the gate and total input capacitance of a minimum sized device, respectively. C_0 is the summation of the gate and drain capacitance of the device. R_{r0} and R_{d0} describe the equivalent output resistance of a minimum sized device used to determine the propagation delay and transition time of a minimum sized inverter, respectively, where the output resistance is approximated as

$$R_{r(d)0} = K_{r(d)} \frac{V_{dd}}{I_{dn0}}. \tag{5.12}$$

K denotes a fitting coefficient and I_{dn0} is the drain current of an NMOS device where both V_{ds} and V_{gs} are equal to V_{dd}. The saturation current for a MOSFET assuming the alpha-power law model [13, 14] is

$$I_{dsat} = I_{d0} \left(\frac{V_{gs} - V_{th}}{V_{dd} - V_{th}} \right)^{\alpha}, \tag{5.13}$$

where

$$I_{d0} = \frac{\mu_0}{[1 + \theta (V_{dd} - V_{th})][1 + V_{dd}/(E_C L)]} C_{ox} V_{D0} [V_{dd} - V_{th} - (\eta/2) V_{D0}], \tag{5.14}$$

and V_{D0} is the drain source voltage at saturation with V_{gs} equal to V_{dd}. The parameters θ and μ_0 are described in [13], while the technology related constants are from the ITRS report [15] and the MOSFET models [16, 17] are for a 45 nm technology node. The $a_{n(p)}$ parameter of the model is

$$a_{n(p)} = \frac{1}{\ln 2} \ln \left(\frac{2V_{D0}[V_{dd} - V_{th_n(p)} - (\eta/2)V_{D0}]}{V_{Da}[V_{dd} - V_{th_n(p)} - \eta V_{Da}]} \right), \tag{5.15}$$

where V_{Da} is the drain source voltage at saturation with V_{gs} equal to $(V_{dd} + V_{th_n(p)})/2$ [13].

To include the effect of the input slew rate on the total delay of an interconnect stage, (5.10) and (5.11) are further refined by including an additional coefficient γ as in [18],

$$\gamma_r = \frac{1}{2} - \frac{1 - {V_{tn}}/{V_{dd}}}{1 + a_n}. \tag{5.16}$$

By substituting the subscript n with p, the corresponding value for a falling input transition is obtained. The average value of γ_r and γ_f is γ, which is used to determine the effect of the input transition time on the interconnect delay. The overall interconnect delay can therefore be described as

$$t_i = k_i(t_{di} + \gamma t_{ri}) = b_1 \frac{r_i c_i l_i^2}{k_i} + b_2 \left(R_0 C_0 k_i + \frac{R_0 c_i l_i}{h_i} + r_i l_i C_{g0} h_i \right), \tag{5.17}$$

where R_0, b_1, and b_2 are described in [19] and the index i denotes the interconnect structures, such as the crossbar switch ($i \equiv s$), horizontal buss ($i \equiv h$), and vertical buss ($i \equiv v$).

The interconnect delay also depends upon the temperature during circuit operation. To capture these dependencies, the resistance of the interconnects is

$$r_i = \rho_0 \left(1 + \beta_{Cu} \left(T - T_{ref}\right)\right) / A_{wi}, \tag{5.18}$$

where T_{ref} and T are the reference and operating temperature of the circuit, respectively. The resistivity of copper at T_{ref} is ρ_0, where a different resistivity is used for each tier according to the ITRS. β_{Cu} is the temperature coefficient of resistance for copper, $\beta_{Cu} = 3.9 \times 10^{-3}$ 1/°C. A_{wi} is the area of the cross-section of the wire. The MOSFET current described by (5.13) and (5.14) also varies with temperature. Analytic expressions for V_{D0} and V_{th} as a function of temperature have been adapted from the BSIM User's manual [20]. The dependence of $a_{n(p)}$ on temperature is implicitly captured through those analytic expressions describing V_{Da} and V_{th} as a function of temperature [13, 20].

For minimum delay, the size h_i and number k_i of repeaters are determined by setting the partial derivative of t_i with respect to h_i and k_i, respectively, equal to zero and solving for h_i and k_i [21],

$$k_i^* = \sqrt{\frac{a_1 r_i c_i l_i^2}{a_2 R_0 C_0}}, \tag{5.19}$$

$$h_i^* = \sqrt{\frac{R_0 c_i}{r_i C_{g0}}}. \tag{5.20}$$

The expression in (5.17) only considers *RC* interconnects. An *RC* model is sufficiently accurate to characterize the delay of a crossbar switch since the length of the longest wire within the crossbar switch and the signal frequencies are such that inductive behavior is not prominent. For buss lines, however, inductive behavior can appear. For this case, suitable expressions for the delay and repeater insertion characteristics can be adopted from [22, 23]. Additionally, for the vertical buss, $k_v = 1$ and $h_v = 1$, meaning that no repeaters are inserted and minimum sized drivers are utilized. Repeaters are not necessary due to the short length of the vertical buss. Note that the latency expressions include the effect of the input slew rate and temperature. Additionally, since a repeater insertion methodology for minimum latency is applied, any further reduction in latency is due to the network topology.

The length of the vertical communication channel for the 3-D NoC shown in Fig. 5.1 is

$$l_v = \begin{cases} L_v, & \text{for 2D IC} - \text{3D NoC} & (5.21a) \\ n_p L_v, & \text{for 3D IC} - \text{3D NoC} & (5.21b) \\ 0, & \text{for 2D IC} - \text{2D NoC and 3DIC} - \text{2D NoC}, & (5.21c) \end{cases}$$

where L_v is the length of a through-silicon (interplane) via connecting two switches on adjacent physical planes. n_p is the number of physical planes used to integrate each PE. The length of the horizontal buss is

$$l_h = \begin{cases} \sqrt{A_{PE}} & \text{for 2D IC} - \text{2D NoC and 2D IC} - \text{3D NoC} & (5.22a) \\ coef\sqrt{A_{PE}/n_p}, & \text{for 3D IC} - \text{2D NoC and 3D IC} - \text{3D NoC}(n_p > 1), & (5.22b) \end{cases}$$

where A_{PE} is the area of the processing element. The area of all of the PEs and, consequently, the length of each horizontal link are assumed to be equal. For those cases where the PE is implemented in multiple physical planes ($n_p > 1$), a coefficient *coef* is used to consider the effect of the interplane vias on the reduction in wirelength due to utilization of the third dimension. This coefficient is based on the layout of a crossbar switch designed [24] with the FDSOI 3-D technology from MIT Lincoln Laboratory (MITLL) [25]. In the following section, expressions for the power consumption of a network with delay constraints are presented.

$$k_i^* = \sqrt{\frac{a_1 r_i c_i l_i^2}{a_2 R_0 C_0}}, \tag{5.19}$$

$$h_i^* = \sqrt{\frac{R_0 c_i}{r_i C_{g0}}}. \tag{5.20}$$

The expression in (5.17) only considers *RC* interconnects. An *RC* model is sufficiently accurate to characterize the delay of a crossbar switch since the length of the longest wire within the crossbar switch and the signal frequencies are such that inductive behavior is not prominent. For buss lines, however, inductive behavior can appear. For this case, suitable expressions for the delay and repeater insertion characteristics can be adopted from [22, 23]. Additionally, for the vertical buss, $k_v = 1$ and $h_v = 1$, meaning that no repeaters are inserted and minimum sized drivers are utilized. Repeaters are not necessary due to the short length of the vertical buss. Note that the latency expressions include the effect of the input slew rate and temperature. Additionally, since a repeater insertion methodology for minimum latency is applied, any further reduction in latency is due to the network topology.

The length of the vertical communication channel for the 3-D NoC shown in Fig. 5.1 is

$$l_v = \begin{cases} L_v, & \text{for 2D IC} - \text{3D NoC} \quad (5.21a) \\ n_p L_v, & \text{for 3D IC} - \text{3D NoC} \quad (5.21b) \\ 0, & \text{for 2D IC} - \text{2D NoC and 3DIC} - \text{2D NoC}, \quad (5.21c) \end{cases}$$

where L_v is the length of a through-silicon (interplane) via connecting two switches on adjacent physical planes. n_p is the number of physical planes used to integrate each PE. The length of the horizontal buss is

$$l_h = \begin{cases} \sqrt{A_{PE}} & \text{for 2D IC} - \text{2D NoC and 2D IC} - \text{3D NoC} \quad (5.22a) \\ coef\sqrt{A_{PE}/n_p}, & \text{for 3D IC} - \text{2D NoC and 3D IC} - \text{3D NoC}(n_p > 1), \quad (5.22b) \end{cases}$$

where A_{PE} is the area of the processing element. The area of all of the PEs and, consequently, the length of each horizontal link are assumed to be equal. For those cases where the PE is implemented in multiple physical planes ($n_p > 1$), a coefficient *coef* is used to consider the effect of the interplane vias on the reduction in wirelength due to utilization of the third dimension. This coefficient is based on the layout of a crossbar switch designed [24] with the FDSOI 3-D technology from MIT Lincoln Laboratory (MITLL) [25]. In the following section, expressions for the power consumption of a network with delay constraints are presented.

5.3.2 Power Consumption Model for 3-D NoC

Power dissipation is a critical issue in three-dimensional circuits. Although the total power consumption of a 3-D system is expected to be lower than that of an equivalent 2-D circuit (since the global interconnects are shorter [26]), the increased power density is a challenging issue for this novel design paradigm. Therefore, those 3-D NoC topologies that offer low power characteristics are of significant interest.

The different power consumption components for interconnects with repeaters are briefly discussed in this section. Due to specified performance characteristics, a low power design methodology with delay constraints for the interconnect within an NoC is adopted from [19]. An expression for the power consumption per bit of a packet transferred between a source destination node pair is used as the basis for characterizing the power consumption of an NoC for the proposed topologies.

The power consumption components of an interconnect line with repeaters are:

(a) *Dynamic power consumption* is the dissipated power due to the charge and discharge of the interconnect and input gate capacitance during a signal transition, and can be described by

$$P_{di} = a_{s_noc} f (c_i l_i + h_i k_i C_0) V_{dd}^2, \tag{5.23}$$

where f is the clock frequency and a_{s_noc} is the switching factor [27].

(b) *Short-circuit power* is due to the DC current path that exists in a CMOS circuit during a signal transition when the input signal voltage changes between V_{tn} and $V_{dd} + V_{tp}$. The power consumption due to this current is described as short-circuit power and is modeled in [28] by

$$P_{si} = \frac{4 a_{s_noc} f I_{d0}^2 t_{ri}^2 V_{dd} k_i h_i^2}{V_{dsat} G C_{effi} + 2 H I_{d0} t_{ri} h_i}, \tag{5.24}$$

where I_{d0} is the average drain current of the NMOS and PMOS devices operating in the saturation region and the value of the coefficients G and H are described in [29]. Due to resistive shielding of the interconnect capacitance, an effective capacitance is used in (5.23) rather than the total interconnect capacitance. Note that resistive shielding results in a smaller capacitive load as seen from the interconnect driver (i.e., $C_{eff} \leq C_{total}$). This effective capacitance is determined from the methodology described in [30, 31].

(c) *Leakage power* is comprised of two power components, the subthreshold and gate leakage currents. Subthreshold power consumption is due to current flowing where the transistor operates in the cut-off region (below threshold), causing I_{sub} current to flow. The gate leakage component is due to current flowing through the gate oxide, denoted as I_g. The total leakage power can be described as

$$P_{li} = h_i k_i V_{dd} (I_{sub0} + I_{g0}), \tag{5.25}$$

where the average subthreshold I_{sub0} and gate I_{g0} leakage current of the NMOS and PMOS transistors is considered in (5.25).

The total power consumption with delay constraint T_0 for a single line of a crossbar switch P_{stotal}, horizontal buss P_{htotal}, and vertical buss P_{vtotal} is, respectively,

$$P_{stotal}(T_0 - t_a) = P_{di} + P_{si} + P_{li}, \tag{5.26}$$

$$P_{htotal}(T_0) = P_{di} + P_{si} + P_{li}, \tag{5.27}$$

$$P_{vtotal}(T_0) = P_{di} + P_{si} + P_{li}. \tag{5.28}$$

The power consumed by the arbitration logic is not considered in (5.26)–(5.28) since most of the power is consumed by the crossbar switch and the buss interconnect, as discussed in [32]. Note that for a crossbar switch, the additional delay of the arbitration logic poses a stricter delay constraint on the switch. The minimum power consumption with delay constraints is determined by the methodology described in [19], which is used to determine the optimum size h^*_{powi} and number k^*_{powi} of repeaters for a single interconnect line. Consequently, the minimum power consumption per bit between a source destination node pair in an NoC with a delay constraint is

$$P_{bit} = hopsP_{stotal} + hops_{2-D}P_{htotal} + hops_{3-D}P_{vtotal}. \tag{5.29}$$

Note that the proposed power expression includes all of the power consumption components in the network, not only the dynamic power. The effect of resistive shielding is also considered in determining the effective interconnect capacitance. Additionally, the effect of temperature on each of the power dissipation components is considered. Furthermore, since the repeater insertion methodology described in [19] minimizes the power consumed by the repeater system, any additional decrease in power consumption is only due to the network topology. In the following section, a technique for analyzing the power dissipation of a PE and the related rise in temperature within an NoC system is described.

5.4 Thermal-Aware Analysis Methodology

The 3-D NoC topologies and related timing and power models presented in the previous section emphasize performance improvements achieved by including the third dimension within the on-chip network. Vertical integration, however, can significantly enhance the performance of the PEs [33, 34], in addition to the performance of the network. Redesigning a planar PE into several physical planes can greatly decrease the power consumed by the PEs [35]. This reduction, in turn, lowers the temperature rise within the stack leading to tangible benefits in the overall behavior of the system. For example, since the temperature rise is limited, the corresponding increase in the interconnect resistance is lower, decreasing the interconnect delay within the NoC. Excessive increases in leakage power will also be avoided. In this

section, a methodology for analyzing the overall behavior of a 3-D system interconnected with an on-chip network is presented.

Since the systems described in this chapter are presumed to be homogeneous, the power consumed by the entire system can be straightforwardly determined if the power dissipated by an elemental unit, as depicted in Fig. 5.1, is known. The power consumed by a PE and the two physical links that surround the PE is another useful metric for characterizing the different 3-D NoC topologies,

$$P_{total} = P_{PE} + 2L_p P_{bit}, \tag{5.30}$$

where P_{PE} is the power consumed by a single PE. P_{PE} includes all of the different energy components described in Section 5.3.2. Different expressions, however, can be used to describe these components. The dynamic power consumption P_{PE_dyn} is separated into the power constituents, P_g and P_{int}, for driving the capacitance of the logic gates and interconnects, respectively. The dynamic power consumption can therefore be written as

$$P_{PE_dyn} = P_g + P_{int} = f_{PE} \cdot \alpha_s \left(N_{tot} C_{gate} + C_{int_loc} + C_{int_semi} + C_{int_glob} \right) V_{dd}^2, \tag{5.31}$$

where N_{tot} is the number of gates within a PE, a_s is the switching activity within a PE, and f_{PE} is the operating frequency of the PE. Note that this frequency is typically different from the clock frequency f_c of the NoC. C_{int_loc}, C_{int_semi}, and C_{int_glob}, are, respectively, the total capacitance of the local, semi-global, and global interconnects driven by the N_{tot} gates within the PE. The leakage and short-circuit power dissipation of a PE can be determined similarly to (5.24)–(5.25). The leakage power of the PE is

$$P_{PE_leak} = W_{eff}(I_{sub0} + I_{g0}) \cdot V_{dd}, \tag{5.32}$$

where W_{eff} is the total width of the transistors within a PE. Assuming that the PEs consist of four transistor gates and the transistors are sized to equalize the rise and fall transitions, the width of the four devices is a function of the minimum feature size. Multiplying this width by the number of gates within the PE, W_{eff} is obtained. To determine the interconnect capacitance described in (5.31), the distribution of the interconnect within a PE is required. An enhanced wire distribution model for either 2-D or 3-D circuits based on [36] is utilized in this analysis. This model is further integrated into the methodology described by [37] to produce the number of tiers (i.e., local, semi-global, global) and the pitch of each metal layer. A pair of metal layers routed orthogonally is assumed to comprise a tier. A specific constraint for the interconnect delay is set for each tier to determine the maximum interconnect length that can be placed within that tier. This constraint is set to 25 and 90% of the clock frequency f_{PE} for the local and other tiers, respectively. An interconnect is considered to be placed on the next tier whenever the length of the wire does not satisfy the aforementioned constraint.

Although the effect of temperature is considered in these power expressions, the methodology described in [37] does not consider thermal issues. To consider

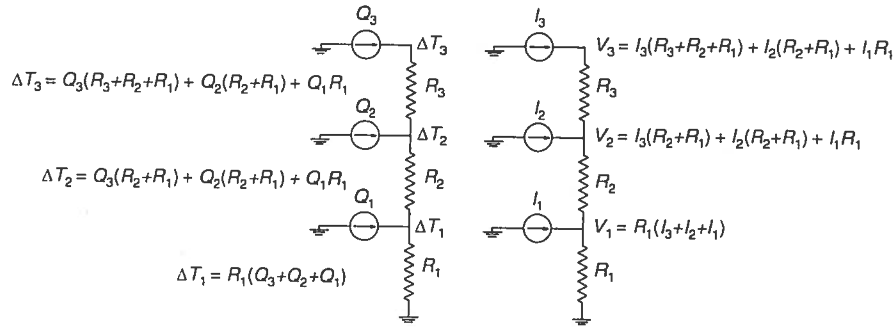

Fig. 5.2 An example of the duality of thermal and electrical systems

these important 3-D circuit effects, a first-order thermal model of a 3-D circuit has been integrated into this methodology. This model assumes a one-dimensional (1-D) flow of heat throughout a 3-D stack [38, 39]. Note that the model includes both the heat generated by the devices and interconnect. The assumption of a 1-D heat flow is justified as the path towards the heat sink exhibits the lowest thermal resistance, facilitating the removal of heat from the circuit to the ambient. This path is the primary path for the heat to flow in a 3-D circuit [39]. By exploiting the electro-thermal duality, as illustrated in Fig. 5.2, the temperature in each plane and each metal layer within a physical plane of a 3-D circuit can be determined.

The number of metal layers, metal pitch within these layers, temperature rise, and power dissipation for a PE can all be determined early in the design cycle by utilizing this thermal-aware interconnect architecture design methodology. The analysis method is depicted in Fig. 5.3. The input parameters are the target technology node, the number of gates N_{tot} within the PE, the number of physical planes n_p used for the PE, and the clock frequency f_{PE}. Most of the interconnect and related parameters can be extracted for the target technology node. The complete methodology proceeds as follows:

1. For the clock frequency f_{PE} and nominal temperature, an initial number of metal layers and the interconnect pitch are determined to satisfy the aforementioned delay constraints.
2. For this interconnect architecture and assuming the same average current density for the wires on each metal layer, the rise in temperature is determined.
3. Based on this increase in temperature, the electrical wire resistance, leakage power, and other temperature dependent parameters are updated. The interconnect delay is again evaluated against the input delay constraints. If these specifications are not satisfied, a new interconnect architecture is produced.
4. The iterative process terminates once the circuit has reached a steady state temperature and the delay constraints for each tier is satisfied. The output is the number of metal layers, interconnect pitch, and temperature of the circuit.

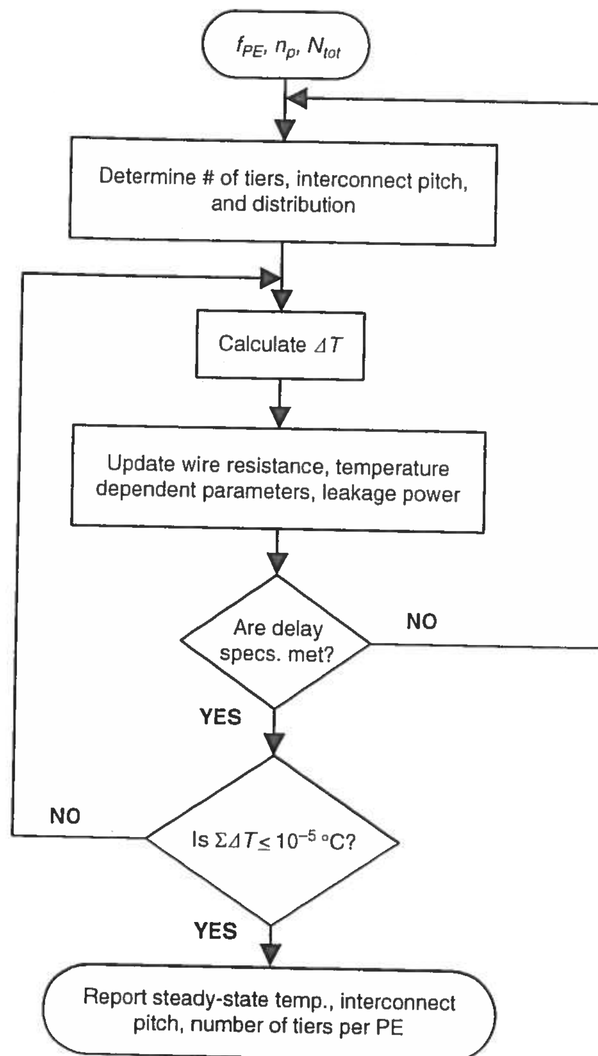

Fig. 5.3 Analysis flow diagram that produces a first-order interconnect architecture and the steady-state temperature of a PE consisting of n_p physical planes within a 3-D stack of n planes ($n_p \leq n$)

Once the temperature and interconnect length at each tier have been determined, the power consumed by the PE is readily determined from (5.30)–(5.32). In this manner, the topology that produces the lowest power dissipation for the entire 3-D system rather than the physical link of the on-chip network is determined. Following this procedure, the different 3-D topologies discussed in Section 5.2 are evaluated in terms of the latency and power dissipation of the NoC, the total power dissipation of the system, and the rise in temperature. Various tradeoffs inherent to these topologies are also discussed for a NoC-based 3-D system.

5.5 Latency, Power, and Temperature Tradeoffs in 3-D NoC Topologies

The improvement in the performance of traditionally planar on-chip networks by introducing the third-dimension is discussed in this section. Those topologies that produce the highest network performance, the lowest power dissipation of the network and system, and the lowest rise in temperature are evaluated. Different topologies are demonstrated that satisfy each of these objectives. Consequently, the analysis methodology presented in the previous section can be a useful tool to evaluate the improvement in a specific objective offered by a 3-D topology. The effect of the 3-D topologies on the speed, power, and temperature rise of a network is discussed in Sections 5.5.1, 5.5.2, and 5.5.3, respectively. The latency, power, and rise in temperature of a 2-D mesh network with the same number of nodes is used as a reference for the comparisons throughout this section.

In all of the on-chip networks, a 45 nm technology node, as described in the ITRS, is assumed [15]. Consequently, specific interconnect and device parameters, such as the minimum pitch of the horizontal interconnects, the maximum power density, and the dielectric constant k are adopted from this report. A TSV technology is assumed to provide the vertical interconnects, with a TSV pitch of 5 μm and an aspect ratio of five [33]. Finally, a synchronous on-chip network is assumed with a clock frequency of $f_c = 1$ GHz.

A small set of parameters are considered as variables throughout the analysis of the 3-D topologies. These variables include the network size, number of physical planes making up the 3-D system, clock frequency of the PEs f_{PE}, and area of the PEs A_{PE}. The range of these variables is listed in Table 5.1. For multi-processor SoC networks, sizes of up to $N = 256$ are expected to be feasible within the near future [7, 40], whereas for NoC with a finer granularity, where each PE corresponds to a hardware block of approximately 100 thousand gates, network sizes over a few thousands nodes are predicted at the 45 nm technology node [41]. Furthermore, to apply the thermal model discussed in [38, 39], the nominal temperature is considered to be 300 K and the side and topmost surfaces of the 3-D stack are assumed to be adiabatic. In other words, the heat is assumed to flow from the uppermost to the lowest plane of the multi-plane system.

Table 5.1 Parameters of the investigated NoCs

Parameter	Values
N	16, 32, 64, 128, 256, 512
A_{PE} (mm^2)	0.64, 0.81, 1.00, 2.25
f_{PE} (GHz)	1, 2, 3
Max. number of planes, n_{max}	8

5.5.1 Latency of 3-D NoCs

Utilizing the third dimension to implement an on-chip network (i.e., 2-D IC – 3-D NoC topology) by simply stacking the PEs (i.e., $n_3 > 1$, $n_p = 1$) and using network switches with seven ports decrease the average number of hops for packet switching, thereby improving the network latency. Alternatively, utilizing the third dimension to decrease the length of the physical links between adjacent network switches (i.e., 3-D IC – 2-D NoC topology) also reduces the latency of the network. The reduction in buss length is achieved by implementing the PEs in multiple physical planes (i.e., $n_3 = 1$, $n_p > 1$), thereby reducing PE area. Finally, a hybrid topology (i.e., 3-D IC – 3-D NoC topology), which uses the third dimension both for the on-chip network and the PEs (i.e., $n_3 > 1$, $n_p > 1$), can result in the greatest reduction in latency.

Note that the effect of the various topologies on only the speed of the network, described by f_c, is considered, while the operating frequency of the PEs f_{PE} can be different. Although this approach is convenient from an architectural perspective, certain physical design issues can arise due to the multiple clock domains that can co-exist in these topologies. Each of these topologies faces different synchronization challenges.

A multi-plane on-chip network can be implemented with various synchronization schemes ranging from a fully synchronous approach (as assumed herein) to an asynchronous network. A potent non-synchronous approach that has been used for PE-to-network communication in planar systems is the globally asynchronous locally synchronous (GALS) approach. An immediate extension of the GALS approach into three physical dimensions could include a number of clock domains equal to the number of planes comprising a 3-D circuit. This synchronization scheme is suitable for the 2-D IC – 3-D NoC topology.

Alternatively, for the 3-D IC – 2-D NoC topology where the network is planar, a synchronous clocking scheme is a simpler and efficient solution. The synchronization challenge for this topology is related to the multi-plane PEs. The primary issue is how to efficiently propagate the clock signal across a PE occupying several planes. Recently, several synthesis techniques that produce bounded skew and testable clock trees prior and after bonding have been reported [42]. In addition, preliminary experimental results on multi-plane clock networks demonstrate that operating frequencies in the gigahertz regime are feasible, while the clock skew is manageable [24]. Although functional testing for each plane of a multi-plane PE remains an important problem, early results are encouraging.

For the 3-D mesh networks discussed in this chapter, fully synchronous schemes are assumed due to the simplicity of these approaches. In this way, the effect of the topological characteristics rather than the synchronization mechanism on the performance of the network is evaluated, which is the objective of the physical analysis flow discussed in this chapter. The latency for different network sizes and a PE area of 0.64 mm^2 and 2.25 mm^2 is illustrated in Fig. 5.4a, b, respectively. Note

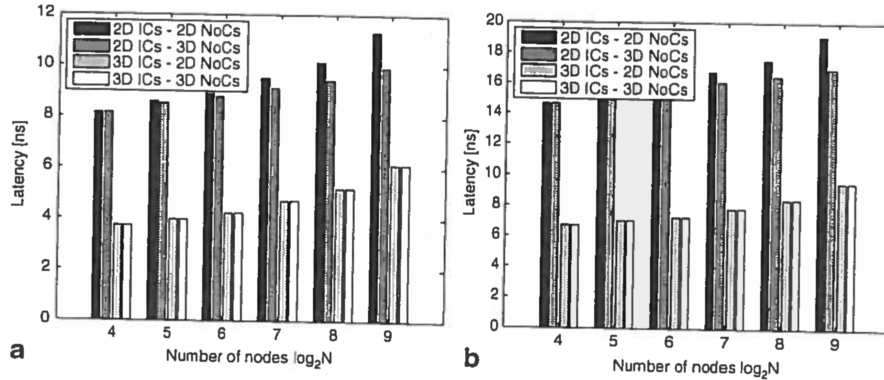

Fig. 5.4 Zero-load latency for various network sizes where **(a)** A_{PE} = 0.64 mm² and **(b)** A_{PE} = 2.25 mm² for f_{PE} = 1 GHz

that the temperature rise due to stacking or folding the PEs is also considered by the methodology presented in Section 5.4.

The latency of the 2-D IC – 3-D NoC topology decreases with increasing network size. For example, for N = 16, the third dimension does not improve the latency, since the number of hops required for packet switching is small. The increase in the delay of the network switch (due to the increase in the number of switch ports) can outweigh the decrease in latency due to the reduction in the number of hops, which is negligible for this network size. As the network size increases, however, the decrease in latency offered by this topology progressively increases. The improvement in latency increases from 1.08% for $N = 32$ to 6.79% for $N = 256$, where A_{PE} = 0.64 mm². In addition, the area of the PE has no significant effect in this improvement, as depicted in Fig. 5.4, since this topology only alters the number of hops for packet switching. Note, however, that the absolute network latency increases for this topology, since the length of the busses increases with the area of the PEs.

The 3-D IC – 2-D NoC exhibits the opposite behavior, since this topology reduces the length of the physical links. Thus, the improvement in latency increases in those networks with a large PE area. The latency decreases by 48.97% for A_{PE} = 0.64 mm² and 60.39% for A_{PE} = 2.25 mm² for a network size of $N = 256$. The reduction in network latency for this topology decreases with increasing network size. As the network size increases, the greatest portion of the latency as described by (5.9) is due to the larger number of hops rather than the buss delay. Consequently, the benefits offered by the reduction in length of the busses decrease with network size for the 3-D IC – 2-D NoC topology. For example, the improvement in latency decreases from 54.12% for $N = 32$ to 50.83% for N = 128, where A_{PE} = 0.64 mm².

The hybrid topology 3-D IC – 3-D NoC demonstrates the greater improvement in latency as compared to a 2-D network, since the third dimension decreases both the length of the busses and the number of hops [43]. Depending upon the network size, the area of the PE, and the interconnect impedance characteristics of the bus-

ses, n_3 and n_p ensure that the 3-D IC – 3-D NoC topology can support either the 2-D IC – 3-D NoC (i.e., $n_3 = n_{max}$) or the 3-D IC – 2-D NoC (i.e., $n_p = n_{max}$).

The results shown in Fig. 5.4 include the increase in delay caused by the rise in temperature within a 3-D stack. Based on the methodology discussed in the previous section, the resulting temperature rise does not significantly affect the improvement in latency provided by the 3-D topologies. The resulting higher temperatures within the 3-D system cause a small 3% increase in the interconnect latency for all of the investigated networks. The thermal effects are similar to those discussed in [44, 45]. Consequently, the overall effect of the 3-D topologies is that the network latency is significantly decreased, although an inevitable increase in temperature will consume a small portion of this improvement. The effect of the third dimension on the power consumed by the network and the PEs is discussed in the following section.

5.5.2 *Power Dissipation of 3-D NoCs*

The decrease in the power of a conventional 2-D on-chip network achieved by the 3-D topologies is presented in this section. Two different power consumption metrics are used to characterize the benefits of these topologies. First, the 3-D topology that minimizes the power consumed by the network is described by (5.29), ignoring the power of the PEs. For the second metric, the overall power dissipation of the system, including both the power of the network and the PEs, is described by (5.30). Those topologies that minimize each of these two metrics are determined. Furthermore, the distribution of the network nodes in terms of the physical dimensions (i.e., n_1, n_2, n_3, and n_p) can be quite different for the same 3-D topology.

The power consumed by these 3-D topologies is illustrated in Fig. 5.5, where the power dissipated by the PEs is ignored. Similar to the discussion related to latency, the 2-D IC – 3-D NoC and 3-D IC – 2-D NoC topologies lower the power dissipated

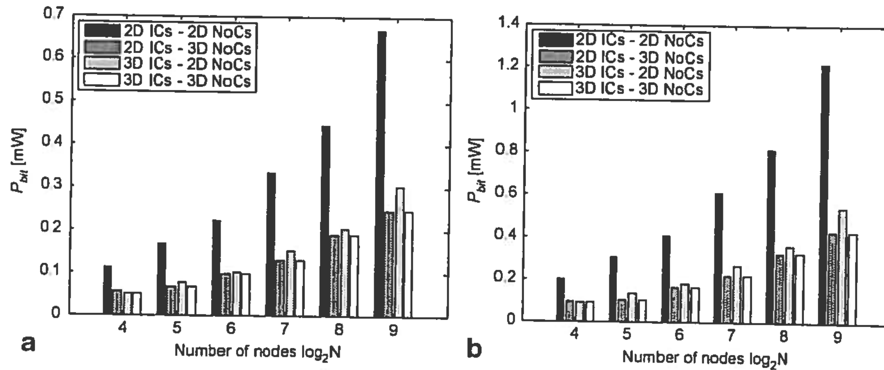

Fig. 5.5 Power consumed by the network with delay constraints (f_c = 1 GHz) for several network sizes where **(a)** A_{PE} = 0.64 mm² and **(b)** A_{PE} = 2.25 mm² for f_{PE} = 1 GHz

by the network through a reduction in the number of hops and the capacitance of the wires, respectively. Note that the *y-axis* in Fig. 5.5 corresponds to the power required to transfer a single bit over an average distance within the network where this distance is determined by the number of hops for packet switching, as described by (5.2). Comparing Fig. 5.5a, b the power consumed by a planar on-chip network increases with the area of the PEs interconnected by this network. For example, the power almost doubles for the same network size as the area of the PE increases from 0.64 mm^2 to 2.25 mm^2.

Similar to the network latency, the power consumption decreases in the 2-D IC – 3-D NoC topology by reducing the number of hops for packet switching. Again, the increase in the number of ports adds to the power consumed by the crossbar switch; however, the effect of this increase in power is not as significant as the corresponding increase in the latency of the network. A three-dimensional network can therefore reduce power even in small networks. The power savings achieved with this topology is greater in larger networks. This situation occurs because the reduction in the average number of hops for a three-dimensional network increases in larger network sizes.

With the 3-D IC – 2-D NoC topology, the number of hops in the network is the same as for a two-dimensional network. The horizontal buss length, however, is shortened by implementing the PEs in more than one physical plane. The greater the number of physical planes that can be integrated in a 3-D system, the larger the power savings; meaning that the optimum value for n_p with this topology is always n_{max} regardless of the network size and operating frequency (if temperature is not the target objective). The savings is practically limited by the number of physical planes that can be integrated in a 3-D technology. For this type of NoC, the topology resulting in the maximum speed is identical to the topology minimizing the power consumption, as the key element of either objective originates solely from the shorter buss length. Finally, the 3-D IC – 3-D NoC can achieve the minimum power consumption for a 3-D on-chip network by properly adjusting n_3 and n_p depending upon the interconnect impedance characteristics, the available number of physical planes, and the clock frequency of the network.

Interestingly, when the power metric described in (5.30) is utilized, the topologies that minimize the total power are different, as illustrated in Fig. 5.6. The distribution of the network nodes within those topologies also changes. The total power of a network-based 3-D system is plotted in Fig. 5.6, where the clock frequency of the network is $f_c = 1$ GHz and the area of the PE is $A_{PE} = 0.64$ mm^2 and $A_{PE} = 2.25$ mm^2. The clock frequency of the PE is equal to f_c in this case.

A common characteristic of Fig. 5.6a, b is that for larger network sizes, the topology that produces the lowest power dissipation changes from the 3-D IC – 2-D NoC to the 2-D IC – 3-D NoC topology (for this specific example, the 3-D IC – 3-D NoC coincides with either of these topologies). For small networks and PE area (see Fig. 5.6a), the reduction in power originates from the shorter buss length of the network and the shorter interconnects within the PEs since the PEs are implemented in multiple planes. As the network size increases, however, the number of hops increases considerably, making the power dissipated by the network the dominant

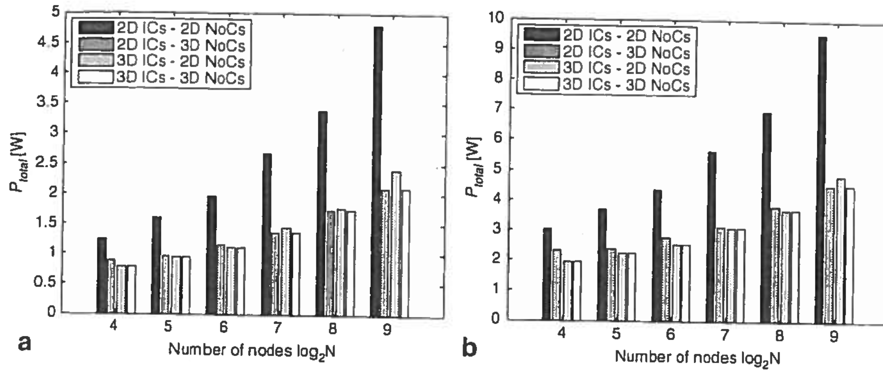

Fig. 5.6 Total power consumed by a PE and the adjacent network busses according to (5.30) with delay constraints (f_c = 1 GHz) for several network sizes where **(a)** A_{PE} = 0.64 mm² and **(b)** A_{PE} = 2.25 mm² for f_{PE} = 1 GHz

power component consumed by the system. Consequently, the 3-D IC – 2-D NoC does not offer the maximum power savings; the maximum savings is now achieved with the 2-D IC – 3-D NoC topology.

If the PE is larger, the network size at which the optimum topology changes increases. This behavior occurs since larger PEs include a greater number of gates leading to additional longer interconnections within the PEs, as described by the interconnect distribution model presented in Section 5.4. The greater number and length of the wires within a PE are the primary power component of the entire system. The 3-D IC – 2-D NoC topology offers a greater improvement for even larger network sizes before the power caused by the increasing number of hops starts to dominate. This behavior occurs since the 3-D IC – 2-D NoC topology reduces the length of the interconnects within the PEs in addition to the length of the network busses.

Another interesting result is that the clock frequency of the PE f_{PE} affects the overall power dissipation, a factor typically ignored when evaluating the performance of a network-based integrated system. In Fig. 5.7, f_{PE} increases from 1 to 3 GHz. This increase has a profound effect on the overall power of the system. To satisfy this aggressive timing specification while limiting the interconnect power consumption, the 3-D IC – 2-D NoC topology exhibits the best results for most of the network sizes depicted in Fig. 5.7a. Note that this behavior is more pronounced for larger PE areas, as depicted in Fig. 5.7b, where the 3-D IC – 2-D NoC topology performs better than the 2-D IC – 3-D NoC topology for any network size. Furthermore, the 3-D IC – 3-D NoC topology can lead to the lowest power consumption with appropriate adjustment of the parameters n_3 and n_p.

To demonstrate the distribution of the networks nodes within the three physical dimensions in addition to the effect on the topology, the node distribution is listed in Table 5.2 for specific network and PE characteristics. From Table 5.2, both the operating frequency and the area of the PEs affect the distribution of nodes within the NoC. Large PE areas (A_{PE} = 2.25 mm²) and high operating frequencies (f_{PE} = 3 GHz) require 3-D NoC topologies where some of the physical planes are

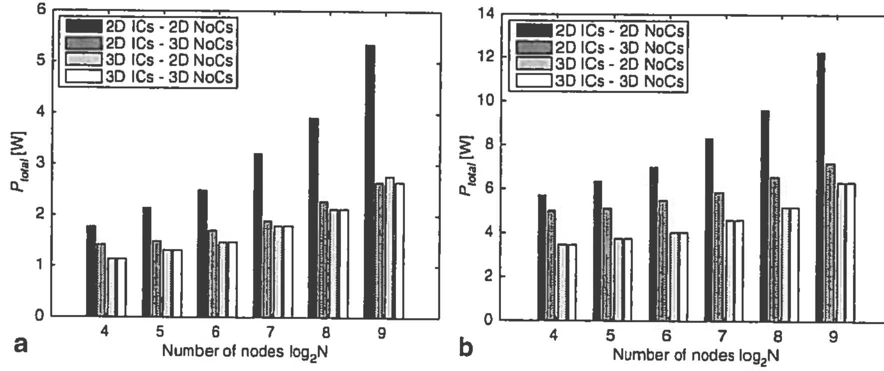

Fig. 5.7 Total power consumed by a PE and the adjacent network busses according to (5.30) with delay constraints (f_c = 1 GHz) for several network sizes where **(a)** A_{PE} = 0.64 mm^2 and **(b)** A_{PE} = 2.25 mm^2 for f_{PE} = 3 GHz

Table 5.2 Node distribution for minimum power consumption of different network sizes

N	f_{PE} = 1 GHz								f_{PE} = 3 GHz							
	A_{PE} = 0.64 mm^2				A_{PE} = 2.25 mm^2				A_{PE} = 0.64 mm^2				A_{PE} = 2.25 mm^2			
	n_1	n_2	n_3	n_p	n_1	n_2	n_3	n_p	n_1	n_2	n_3	n_p	n_1	n_2	n_3	n_p
16	4	4	1	8	4	4	1	8	4	4	1	8	4	4	1	8
32	8	4	1	8	8	4	1	8	8	4	1	8	8	4	1	8
64	8	8	1	8	8	8	1	8	8	8	1	8	8	8	1	8
128	4	4	8	1	16	8	1	8	16	8	1	8	16	8	1	8
256	8	4	8	1	16	16	1	8	16	16	1	8	16	16	1	8
512	8	8	8	1	8	8	8	1	8	8	8	1	32	16	1	8

used for the PEs (i.e., $n_p > 1$). For small PEs (A_{PE} = 0.64 mm^2) and low operating frequencies (f_{PE} = 1 GHz), a simple 3-D network (i.e., $n_3 > 1$ and $n_p = 1$) is typically the best choice. Note that the selection of the optimum topology for either a latency or power objective depends strongly on the interconnect and device characteristics of the specific technology node. Consequently, even for system level exploratory design, the analysis methodology presented in Section 5.4 provides a first estimate of the behavior of a network-based 3-D system. The related temperature rise for these 3-D topologies, which is another design objective for this type of integrated system, is discussed in the following section.

5.5.3 Temperature in 3-D NoCs

Elevated temperatures are expected to become an important challenge in vertical integration, specifically where several high performance circuits form a multi-plane

integrated system. The increased power densities per unit volume can potentially increase the operating temperature of the system to prohibitive levels, greatly affecting the performance characteristics and severely degrading the reliability of this system. Consequently, the temperature rise resulting from these 3-D topologies is of primary interest. Based on the methodology described in Section 5.4, the temperature of the substrate and each metal layer within a physical plane is determined assuming a one-dimensional flow of heat towards the heat sink. The heat sink is assumed to be attached to the lowest plane within the 3-D stack. The change in temperature rise due to the different 3-D topologies, area and operating frequency of the PEs, and number of physical planes are discussed in this section.

Considering the 3-D NoC topologies discussed in this chapter, the 2-D IC – 3-D NoC topology will result in higher temperatures as compared to the 3-D IC – 2-D NoC topology since the former topology simply stacks the PEs, while the latter topology utilizes more than one plane to implement a PE. The 2-D IC – 3-D NoC topology leads to higher temperatures for two reasons. Several PEs, determined by n_3, are placed adjacent to the vertical direction. Consequently, the power density generated by both the devices and metal layers increases. In addition, each of these PEs is implemented in one physical plane (i.e., $n_p = 1$) and, hence, no reduction in power density is possible. Alternatively, the 3-D IC – 2-D NoC topology utilizes more than one plane for each PE, reducing the interconnect load capacitance and, consequently, the temperature within the 3-D system.

The temperature rise resulting from the different 3-D topologies is illustrated in Fig. 5.8 for different number of planes. These temperatures correspond to the temperature rise at the topmost metal layer of the uppermost physical plane over the nominal temperature (here assumed to be 27°C). From Fig. 5.8, as the number of planes increases, the temperature naturally increases for the 2-D IC – 3-D NoC topology (for this topology $n_p = 1$). Alternatively, when some or all of the physical planes are used to implement the PEs, as occurs for the 3-D IC – 3-D NoC and 3-D

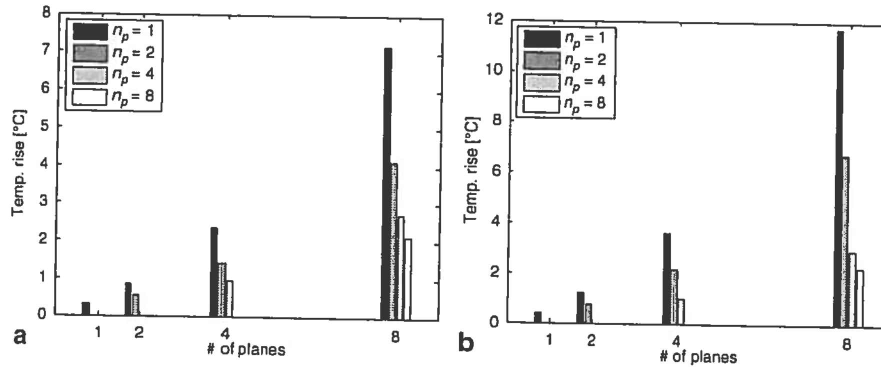

Fig. 5.8 Temperature rise within the 3-D topologies for different combinations of n_3 and n_p. For all of the topologies, $n_3 \times n_p = n$, where n is the number of planes within the 3-D stack. A maximum number of planes $n_{max} = 8$ is assumed, according to Table 5.1. The clock frequency of the PE is $f_{PE} = 1$ GHz and the area is **(a)** $A_{PE} = 0.64$ mm^2 and **(b)** $A_{PE} = 2.25$ mm^2

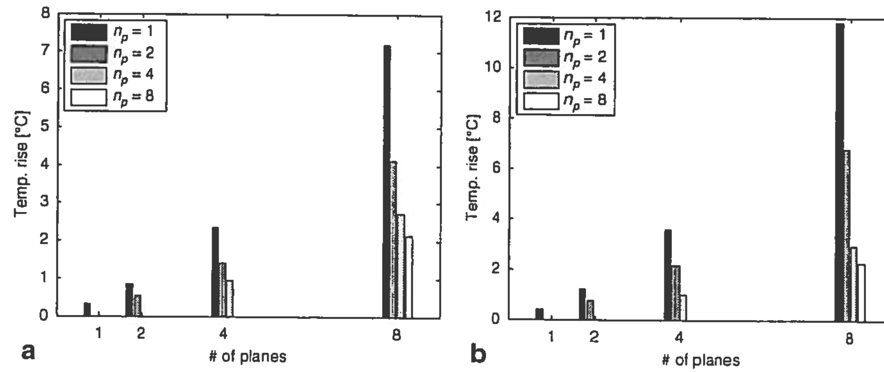

Fig. 5.9 Temperature rise within the 3-D topologies for different combinations of n_3 and n_p. For all of the topologies, $n_3 \times n_p = n$; where n is the number of planes in the 3-D stack. A maximum number of planes $n_{max} = 8$ is assumed according to Table 5.1. The area of the PE is $A_{PE} = 0.64\ \text{mm}^2$ and the clock frequency is **(a)** $f_{PE} = 1$ GHz and **(b)** $f_{PE} = 3$ GHz

IC – 2-D NoC topologies, the temperature rise is considerably smaller. Note, for example, that a 3-D system consisting of eight planes exhibits comparable temperatures to another system comprised of only four planes, as long as the former system uses two physical planes for the PE. This behavior is more pronounced for PEs with larger area, as depicted in Fig. 5.8b. For this case, using more than one plane for the PE significantly reduces the power density per plane. Most importantly, however, the number of metal layers required for a two-plane PE can be smaller [36]. This construction decreases the length and resistance of the vertical thermal path to remove the heat within the 3-D stack.

An increase in temperature also occurs when the operating frequency of the PEs increases, as illustrated in Fig. 5.9. This behavior can be explained by noting that an increase in frequency produces a linear increase in the (dynamic) power consumed by the 3-D system. Note that the temperature rise for higher frequencies within the PEs is comparable to the increase observed for PEs with larger areas. In Fig. 5.8, the area of the PE is almost quadrupled, while in Fig. 5.9 the operating frequency is tripled, resulting in approximately the same rise in temperature. This behavior can be explained as follows. A larger PE includes additional gates that require additional wiring resources. Alternatively, tighter timing constraints can be satisfied, in this example, by increasing the wire pitch. If in either case, an additional tier is required, the thermal resistance of the heat flow path increases. Additionally, for both cases, the power consumption increases, resulting in higher temperatures.

The increase in temperature shown in Figs. 5.8 and 5.9 is for the highest metal layer of the uppermost physical plane within a 3-D system. Although this increase may not be catastrophic, the timing specifications for the PE or the network can possibly not be satisfied if temperature is ignored. To better explain this situation, the different metal pitches for the PEs considering thermal effects are listed in Table 5.3 for the 2-D IC 3-D NoC topology where $n_3 = 8$. In columns 2 to 7, thermal effects are not considered in the analysis flow diagram depicted in Fig. 5.3, while

Table 5.3 Pitch of the interconnect layers for each plane for the 2-D IC – 3-D NoC topology where $n_3 = 8$, $A_{PE} = 1\ mm^2$, and $f_{PE} = 3$ GHz. Two cases are considered, where the system operates at nominal T_0 and at temperature $T_0 + \Delta T$. At the uppermost plane $\Delta T = 20.1°C$

T_0							$T_0 + \Delta T$					
Plane	# of tiers	Metal pitch (nm)					# of tiers	Metal pitch (nm)				
		Tier 1	Tier 2	Tier 3	Tier 4	Tier 5		Tier 1	Tier 2	Tier 3	Tier 4	Tier 5
1	5	90	270	900	1,440	2,000	5	90	270	900	1,280	2,250
2	5	90	270	900	1,440	2,000	5	90	270	900	1,280	2,500
3	5	90	270	900	1,440	2,000	5	90	270	900	1,280	2,500
4	5	90	270	900	1,440	2,000	5	90	270	900	1,280	2,750
5	5	90	270	900	1,440	2,000	5	90	270	900	1,280	2,750
6	5	90	270	900	1,440	2,000	5	90	270	900	1,280	3,000
7	5	90	270	900	1,440	2,000	5	90	270	900	1,280	3,000
8	5	90	270	900	1,440	2,000	5	90	270	900	1,280	3,000

in columns 8–13, thermal effects are considered. Note that a different temperature is determined for each tier in a plane according to the flow diagram shown in Fig. 5.3. For the uppermost plane, the maximum rise in temperature is $\Delta T = 20.1°C$. As reported in Table 5.3, neglecting the rise in temperature, particularly in the upper planes, results in a smaller interconnect pitch which is insufficient to satisfy the timing requirements. Another tier (not shown in Table 5.3) should be used for the network and, therefore, separate timing specifications would apply for this tier. The pitch of this global interconnect tier is not determined by the analysis procedure described in Section 5.4, but a small pitch is selected to constrain the area allocated for the physical links within the network.

The power consumption and related temperature rise also depend upon the switching activity of both the network a_{s_noc} and the PEs a_s. The relative magnitude of these two parameters can greatly affect the behavior of a 3-D topology. In these examples, $a_{s_noc} = 0.25$ and $a_s = 0.15$ have been assumed. These parameters do not affect those traits of the 3-D topologies that improve the performance of a conventional 2-D network but can considerably affect the extent to which each of the 3-D topologies can improve a specific design objective.

5.6 Summary

3-D NoC are a natural evolution of 2-D NoC, exhibiting superior performance. Several 3-D NoC topologies are discussed. Models for the zero-load latency and power consumed by a network are presented for these 3-D topologies. Expressions for the power dissipation of the entire system including the PEs are also provided. A methodology that predicts the distribution of the interconnects within a system based on an on-chip network is extended to accommodate the 3-D nature of the investigated topologies. Thermal effects of the interconnect distribution are also considered in this analysis methodology.

In 3-D NoCs, the minimum latency and power consumption can be achieved by reducing both the number of hops per packet and the length of the communication channels. The topology that best achieves this reduction, however, changes according to the design objective. The network size, speed, and gate count of the PEs, as well as the particular 3-D technology are some important aspects that need to be considered when a 3-D topology is chosen. Selecting a topology that minimizes the power dissipated by an on-chip network does not necessarily guarantee that the power dissipated by the overall system will be minimized. Consequently, the analysis methodology described in this chapter can be a useful tool for exploring early in the design cycle the topological and architectural choices of a 3-D NoC-based system.

References

1. G. De Micheli and L. Benini, *Networks on Chips: Technology and Tools*, Morgan Kaufmann, San Francisco, CA, 2006.
2. A. Jantsch and H. Tenhunen, *Networks on Chip*, Kluwer Academic, San Francisco, CA, 2003.
3. M. Millberg et al., "The Nostrum Backbone—A Communication Protocol Stack for Networks on Chip," *Proceedings of the IEEE International Conference on VLSI Design*, pp. 693–696, January 2004.
4. J. M. Duato, S. Yalamanchili, and L. Ni, *Interconnection Networks: An Engineering Approach*, Morgan Kaufmann, San Francisco, CA, 2003.
5. D. Park et al., "MIRA: A Multi-Layered On-Chip Interconnect Router Architecture," *Proceedings of the IEEE International Symposium on Computer Architecture*, pp. 251–261, June 2008.
6. C. Addo-Quaye, "Thermal-Aware Mapping and Placement for 3-D NoC Designs," *Proceedings of the IEEE International System-on-Chip Conference*, pp. 25–28, September 2005.
7. F. Li et al., "Design and Management of 3D Chip Multiprocessors Using Network-in-Memory," *Proceedings of the IEEE International Symposium on Computer Architecture*, pp. 130–142, June 2006.
8. V. F. Pavlidis and E. G. Friedman, "Three-Dimensional (3-D) Topologies for Networks-on-Chip," *Proceedings of the IEEE International System-on-Chip Conference*, pp. 285–288, September 2006.
9. C. Seiculescu, S. Murali, L. Benini, and G. De Micheli, "SunFloor 3D: A tool for Networks on Chip Topology Synthesis for 3D Systems on Chips," *ACM/IEEE Design, Automation and Test in Europe Conference and Exhibition*, pp. 9–14, April 2009.
10. W. J. Dally and B. Towles, *Principles and Practices of Interconnection Networks*, Morgan Kaufmann, San Francisco, CA, 2004.
11. L.-S. Peh and W. J. Dally, "A Delay Model for Router Microarchitectures," *IEEE Micro*, Vol. 21, No. 1, pp. 26–34, January/February 2001.
12. T. Sakurai, "Closed-Form Expressions for Interconnection Delay, Coupling, and Crosstalk in VLSI's," *IEEE Transactions on Electron Devices*, Vol. 40, No. 1, pp. 118–124, January 1993.
13. K. A. Bowman et al., "A Physical Alpha-Power Law MOSFET Model," *IEEE Journal of Solid States Circuits*, Vol. 34, No. 10, pp. 1410–1414, October 1999.
14. S. L. Garverick and C. G. Sodini, "A Simple Model for Scaled MOS Transistors that Includes Field-Dependent Mobility," *IEEE Journal of Solid States Circuits*, Vol. SC-22, No. 2, pp. 111–114, February 1987.
15. *The International Technology Roadmap for Semiconductors* Reports, 2009 [Online]. Available: http://www.itrs.net/Links/2008ITRS/Home2008.htm

16. Predictive Technology Model [Online]. Available: http://www.eas.asu.edu/~ptm
17. W. Zhao and Y. Cao, "New Generation of Predictive Technology Model for Sub-45 nm Design Exploration," *Proceedings of the IEEE International Symposium on Quality Electronic Design*, pp. 585–590, March 2006.
18. T. Sakurai and A. R. Newton, "Alpha-Power Law MOSFET Model and Its Applications to CMOS Inverter Delay and Other Formulas," *IEEE Journal of Solid State Circuits*, Vol. 25, No. 2, pp. 584–594, April 1990.
19. G. Chen and E. G. Friedman, "Low-Power Repeaters Driving *RC* and *RLC* Interconnects with Delay and Bandwidth Constraints," *IEEE Transactions on Very Large Integration (VLSI) Systems*, Vol. 12, No. 2, pp. 161–172, February 2006.
20. X. Xi et al., *BSIM4.5.0 MOSFET Model User's Manual*, University of California, Berkeley, CA, 2004.
21. H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, Addison-Wesley, Reading, MA, 1990.
22. Y. I. Ismail, E. G. Friedman, and J. L. Neves, "Equivalent Elmore Delay for *RLC* Trees," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 19, No. 1, pp. 83–97, January 2000.
23. Y. I. Ismail, E. G. Friedman, and J. L. Neves, "Figures of Merit to Characterize the Importance of On-Chip Inductance," *IEEE Transactions on Very Large Integration (VLSI) Systems*, Vol. 7, No. 4, pp. 442–449, December 1999.
24. V. F. Pavlidis, I. Savidis, and E. G. Friedman, "Clock Distribution Networks for 3-D Integrated Circuits," *Proceedings of the IEEE International Conference on Custom Integrated Circuits*, pp. 651–654, September 2008.
25. Massachusetts Institute of Technology Lincoln Laboratory, *FDSOI Design Guide*, Cambridge, 2006.
26. H. Hua et al., "Performance Trends in Three-Dimensional Integrated Circuits," *Proceedings of the International IEEE Interconnect Technology Conference*, pp. 45–47, June 2006.
27. K. Banerjee and A. Mehrotra, "A Power-Optimal Repeater Insertion Methodology for Global Interconnects in Nanometer Design," *IEEE Transactions on Electron Devices*, Vol. 49, No. 11, pp. 2001–2007, November 2002.
28. H. J. M. Veendrick, "Short-Circuit Dissipation of Static CMOS Circuitry and Its Impact on the Design of Buffer Circuits," *IEEE Journal of Solid State Circuits*, Vol. SC-19, No. 4, pp. 468–473, August 1984.
29. K. Nose and T. Sakurai, "Analysis and Future Trend of Short-Circuit Power," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 19, No. 9, pp. 1023–1030, September 2000.
30. G. Chen and E. G. Friedman, "Effective Capacitance of Inductive Interconnects for Short-Circuit Power Analysis," *IEEE Transactions on Circuits and Systems I: Brief Papers*, Vol. 55, No. 1, pp. 26–30, January 2008.
31. P. R. O'Brien and T. L. Savarino, "Modeling the Driving-Point Characteristic of Resistive Interconnect for Accurate Delay Estimation," *Proceedings of the International IEEE/ACM Conference on Computer-Aided Design*, pp. 512–515, April 1989.
32. H. Wang, L.-S. Peh, and S. Malik, "Power-Driven Design of Router Microarchitectures in On-Chip Networks," *Proceedings of the IEEE International Symposium on Microarchitecture*, pp. 105–116, December 2003.
33. V. F. Pavlidis and E. G. Friedman, *Three-Dimensional Integrated Circuit Design*, Morgan Kaufmann, San Francisco, CA, 2009.
34. V. F. Pavlidis and E. G. Friedman, "Interconnect-Based Design Methodologies for Three-Dimensional Integrated Circuits," *Proceedings of the IEEE, Special Issue on 3-D Integration Technology*, Vol. 97, No. 1, pp. 123–140, January 2009.
35. J. W. Joyner and J. D. Meindl, "Opportunities for Reduced Power Distribution Using Three-Dimensional Integration," *Proceedings of the IEEE International Interconnect Technology Conference*, pp. 148–150, June 2002.

36. J. W. Joyner et al., "Impact of Three-Dimensional Architectures on Interconnects in Gigascale Integration," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 9, No. 6, pp. 922–927, December 2000.
37. R. Venkatesan, J. A. Davis, K. A. Bowman, and J. D. Meindl, "Optimal *n-tier* Multilevel Interconnect Architectures for Gigascale Integration (GSI)," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 9, No. 6, pp. 899–912, December 2001.
38. T.-Y. Chiang, S. J. Souri, C. O. Chui, and K. C. Saraswat, "Thermal Analysis of Heterogeneous 3D ICs with Various Integration Scenarios," *Proceedings of the IEEE International Electron Device Meeting*, pp. 681–684, December 2001.
39. T.-Y. Chiang, K. Banerjee, and K. C. Saraswat, "Analytical Thermal Model for Multilevel VLSI Interconnects Incorporating Via Effect," *IEEE Electron Device Letters*, Vol. 23, No. 1, pp. 31–33, January 2002.
40. C. Marcon et al., "Exploring NoC Mapping Strategies: An Energy and Timing Aware Technique," *Proceedings of the ACM/IEEE Design, Automation and Test in Europe Conference and Exhibition*, Vol. 1, pp. 502–507, March 2005.
41. P. P. Pande et al., "Performance Evaluation and Design Trade-Offs for Network-on-Chip Interconnect Architectures," *IEEE Transactions on Computers*, Vol. 54, No. 8, pp. 1025–1039, August 2005.
42. X. Zhao, D. L. Lewis, H.-S. H. Lee, and S. K. Lim, "Pre-Bond Testable Low-Power Clock Tree Design for 3D Stacked ICs," *Proceedings of the IEEE/ACM International Conference on Computer Aided Design*, pp. 184–190, November 2009.
43. V. F. Pavlidis and E. G. Friedman, "3-D Topologies for Networks-on-Chip," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 15, No. 10, pp. 1081–1090, October 2007.
44. A. H. Ajami, K. Banerjee, and M. Pedram, "Modeling and Analysis of Nonuniform Substrate Temperature Effects on Global ULSI Interconnects," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, Vol. 24, No. 6, pp. 849–861, June 2005.
45. J. C. Ku and Y. Ismail, "Thermal-Aware Methodology for Repeater Insertion in Low-Power VLSI Circuit," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, Vol. 15, No. 8, pp. 963–970, August 2007.