The Y-Architecture for On-Chip Interconnect: Analysis and Methodology*

Hongyu Chen, Chung-Kuan Cheng, Andrew B. Kahng, Ion Măndoiu,† Qinke Wang and Bo Yao

CSE Department, University of California at San Diego, La Jolla, CA 92093-0114, USA
†CSE Department, University of Connecticut, Storrs, CT 06269, USA
E-mail: {hchen,kuan,abk,qiwang,byao}@cs.ucsd.edu, ion@engr.uconn.edu

Abstract

The *Y-architecture* for on-chip interconnect is based on pervasive use of 0-, 120-, and 240-degree oriented semi-global and global wiring. Its use of three uniform directions exploits on-chip routing resources more efficiently than traditional Manhattan wiring architecture. This paper gives in-depth analysis of deployment issues associated with the Y-architecture. Our contributions are as follows: (1) We analyze communication capability (throughput of meshes) for different interconnect architectures using a multi-commodity flow approach and a Rentian communication model. Throughput of the Y-architecture is largely improved compared to the Manhattan architecture, and is close to the throughput of the X-architecture. (2) We propose a symmetrical Y clock tree structure with better total wire length compared to both H and X clock tree structures, and better path length compared to the H tree. (3) We discuss power distribution under the Y-architecture, and give analytical and SPICE simulation results showing that the power network in Y-architecture can achieve 8.5% less IR drop than an equally-resourced power network in Manhattan architecture. (4) We propose the use of *via tunnels* and *banks of via tunnels* as a technique for improving routability for Manhattan and Y-architectures.

1 Introduction

The *Y-architecture* refers to the use of 0-, 120-, and 240-degree oriented wires for on-chip interconnect, along with supporting methodologies including hexagonal die shapes, hexagonal power and clock distribution, etc. This name is first used in [8] in the same spirit as the "X architecture" for pervasive use of 45- and 135-degree angles [30].

Compared to the traditional Manhattan (M-) architecture, the Y-architecture offers many potential advantages, such as substantially reduced wirelength and power consumption, and increased communication bandwidth for a wide range of demand topologies. Combined with the M-architecture, the Y-architecture can be applied to the upper two layers to improve global interconnects, such as clock and power distribution networks. Moreover, unlike the X-architecture, the Y-architecture supports a regular routing grid and novel means of avoiding via blockage effects.

Two previous series of works examine the potential use of Y-architecture for integrated circuits: a series of LSI Logic patents by Rostoker et al. [22, 23, 24], and a series of works by Cheng and coauthors [7, 8]. Together, these works set out a number of ideas for device architecture, floorplanning, and place-and-route. However, a number of technical gaps still exist, ranging from clock and power distribution methodology to wireability and throughput analysis. In this work, we provide a more complete, technically in-depth analysis of key deployment and methodology issues associated with the Y-architecture. Our main contributions are as follows:

- We give a more realistic throughput analysis using a communication model based on Rent's rule. Our results show that the Y-architecture provides a throughput improvement of about 20% over the M-architecture for a square chip, very close to the throughput of the X-architecture.
- We discuss clock and power distribution under the Y-architecture. For clock distribution we propose a symmetrical Y clock tree structure with better total wire length compared to both H and X clock tree structures, and better path length compared to the H tree. For power distribution we give analytical and SPICE simulation results showing that a mesh power network in Y-architecture can achieve 8.6% less IR drop than an equally-resourced mesh power network in M-architecture.
- To fully utilize the uniform routing grid available in M- and Y-architectures, and to deal with future increases in via demand due to repeaters [25], we propose the use of *via tunnels* and *banks of via tunnels* to improve routability in these architectures. Such techniques are not obvious with the X-architecture.
- We discuss lithography and manufacturing infrastructure needs, particularly in mask write, related to possible adoption of the Y-architecture.

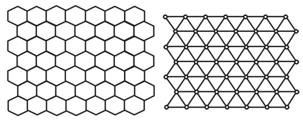
The remainder of the paper is organized as follows. Section 2 presents throughput analysis for square-shaped chips, and also discusses wirelength reduction with hexagonal routing. Sections 3 and 4 examine clock and power distribution, and Section 5 discusses routability issues. The paper concludes in Section 6. Manufacturing issues are discussed in the Appendix.

2 Throughput Analysis and Wirelength Reduction

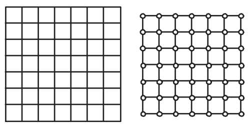
2.1 Communication Throughput in Meshes

A multi-commodity flow (MCF) approach was developed by Chen and coauthors [8] to evaluate communication efficiency of different interconnect architectures. Communication resources are decomposed into a 2D array of slots. A uniform communication requirement is assumed, i.e., every pair of nodes communicates with equal demand and all communications occur at the same time. The throughput, defined as the maximum amount of communication flow simultaneously achievable between every pair of nodes, is computed by a provably good multicommodity flow (MCF) algorithm [12] and is used to measure communication capabilities of different interconnect architectures.

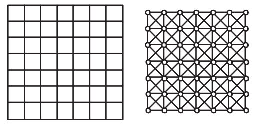
*Work partially supported by Cadence Design Systems, Inc., the California MICRO program, the MARCO Gigascale Silicon Research Center, NSF MIP-9987678 and the Semiconductor Research Corporation. The work of I.M. was performed while he was with the ECE Department at Univeristy of California, San Diego.



(a) A 7 x 7 mesh using Y-architecture.



(b) A 7 x 7 mesh using M-architecture.



(c) A 7 x 7 mesh using X-architecture.

Figure 1: 7×7 meshes with different interconnect architectures.

2.1.1 Rentian Communication Demand

The uniform pairwise communication used in [8] is simple and general. However, it is not very realistic, since in a well-designed layout the probability of communication decreases with increasing distance between nodes. Stroobandt and Campenhout [26] derive from Rent's rule an expression for *occupation probability*, i.e., the probability that a given pair of points will be connected by a wire in an optimal physical placement of the circuit. For a hierarchical placement of a circuit with Rent exponent p in a two-dimensional Manhattan grid, the occupation probability of a pair of points with Manhattan distance D between them can be approximated by CD^{2p-4} where C a normalization constant.[1] When only 2-pin nets are considered, the occupation probability indicates the probability of communication between pairs of nodes. In the following, to ensure a fair comparison of the communication throughput capabilities of different interconnect architectures, we assume a Rentian communication demand, i.e., we set the communication demand between any two unit-area slots to be proportional to D^{2p-4}, where D is the Euclidean distance D between them.

2.1.2 Communication Throughput

A widely quoted survey of Bakoglu [3] indicates that the Rent exponent at the chip and module level of high-speed computers is approximately 0.63. We compute the *throughput* – defined to be the maximum fraction of communication demand simultaneously satisfied between every pair of nodes in $n \times n$ square meshes – using the MCF algorithm. The throughput is tightly correlated to routability, and describes communication capabilities of different interconnect architectures. Figure 1 illustrates three 7×7 meshes using different interconnect architectures. For Y-architecture, the shape of each slot is hexagonal, and the enclosing box of the slots is close to square. Although Y-architecture meshes are different from M- and X-architecture meshes, this does not significantly affect the communication demand. For the 17×17 Y-mesh, total communication demand is only 1.8% different from that for other architectures.

In the experiments, total routing area is set to be the same for all meshes. We normalize the computed throughput so that it is independent of the dimension of meshes and total communication demand.[2] Table 1 lists the results for $n \times n$ meshes with n ranging between 9 and 17. Compared to the M-architecture, the Y-architecture provides an average throughput improvement of 19.8% for these meshes, which is comparable to the 21.9% improvement achieved by the X-architecture. For a 17×17 mesh, Y-architecture provides a throughput improvement of 20.6% while X-architecture achieves an improvement of 22.1%.

Table 1: Normalized throughput (and improvement vs. M-architecture) in square chips with Rentian demand.

Nodes	M-architecture	Y-architecture		X-architecture	
	Thrpt	Thrpt	Impr. (%)	Thrpt	Impr. (%)
81	1.989	2.354	18.30	2.412	21.25
100	1.989	2.366	18.92	2.419	21.59
121	1.987	2.374	19.47	2.420	21.78
144	1.986	2.382	19.94	2.423	22.00
169	1.991	2.386	19.84	2.425	21.76
196	1.990	2.392	20.19	2.429	22.02
225	1.988	2.395	20.47	2.429	22.14
256	1.992	2.400	20.44	2.430	21.98
289	1.992	2.402	20.58	2.433	22.11

A rectangular chip has communication bottlenecks on two (horizontal and vertical) middle cut lines. The physical dimension of the middle part of the chip restricts the communication flow and thus prevents us from achieving larger throughput. For M- and Y-architectures, convex-shaped chips (diamond chip for M-architecture and hexagonal chip for Y-architecture) produce better throughput by allowing more wires to cross the original middle cut lines [8].[3] Note that the use of octagonal chips for the X-architecture is undesirable, since the wafer cannot be tiled by octagons without waste.

2.2 Wirelength Reduction

For completeness, here we briefly mention another aspect of the Y-architecture, namely, its potential for wirelength reduction. Because of its restrictions on routing directions, the M-architecture entails significant added wirelength beyond the Euclidean optimum. In the Y-architecture, routing is allowed along three uniform orientations, and total wirelength is expected to be reduced. An accurate cost-benefit analysis of the Y-architecture is impossible without good estimation of the expected wirelength reduction when switching from rectilinear to hexagonal routing. There are some estimates in the literature [20, 14, 15, 22, 27, 8]; unfortunately, most estimates do not adequately address the effect of routing-geometry-aware placement on the overall wirelength improvement.

As shown in [16], Manhattan placers tend to align circuit elements either vertically or horizontally, leaving few opportunities to exploit additional routing directions. A Y-aware placer factors in hexagonal wiring during placement, and results in better placements of nets when such wiring is used to route the nets. Recently, Chen et al. [9] estimated the wirelength improvement achieved by Y-aware or X-aware placement and routing versus Manhattan placement and routing. The estimate is based on a simplified placer which uses simulated annealing driven by hexagonal or octilinear wirelength estimation. According to [9], the Y-architecture achieves a wirelength improvement of up to 8.3% compared to the M-architecture. The X-architecture further reduces total wirelength by up to 11.4% over M-architecture – giving a reduction of about 3.3% over the Y-architecture at the cost of one more routing direction.

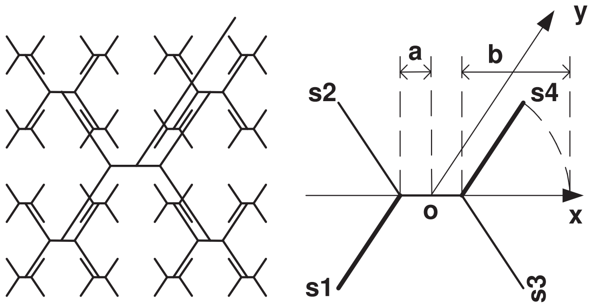
3 Y Clock Tree

Clock distribution networks synchronize the flow of data signals among synchronous data paths. The design of these networks can dramatically affect system-wide performance and reliability. The

[1] C depends on the routing architectures and the underlying distance metric.

[2] For example, the computed throughput on a $n \times n$ mesh using Y-architecture is normalized by $\frac{TD_M}{TD_Y} \cdot D_c/n$, where TD_M and TD_Y are total demand for M- and Y-architectures, respectively, and D_c is the communication demand crossing the horizontal middle cut line on the Manhattan mesh.

[3] Note that it is not necessarily to use a regular hexagon for the Y-architecture: either horizontal or vertical symmetry suffices.



(a) Y Clock Tree.

(b) A One-Level Y Clock Tree.

Figure 2: Y Clock Tree.

Table 2: Path length and total wirelength of H-tree, X-tree and Y-tree.

	Path Length	Total Wirelength
H-tree	(2^n-1)	$\frac{3}{2}\cdot 2^n(2^n-1)$
X-tree	$\frac{\sqrt{2}}{2}\cdot(2^n-1)$	$\sqrt{2}\cdot 2^n(2^n-1)$
Y-tree	$\frac{1}{2}(1+\frac{\sqrt{3}}{3})\cdot(2^n-1)$	$\frac{1+\sqrt{3}}{2}\cdot 2^n(2^n-1)$

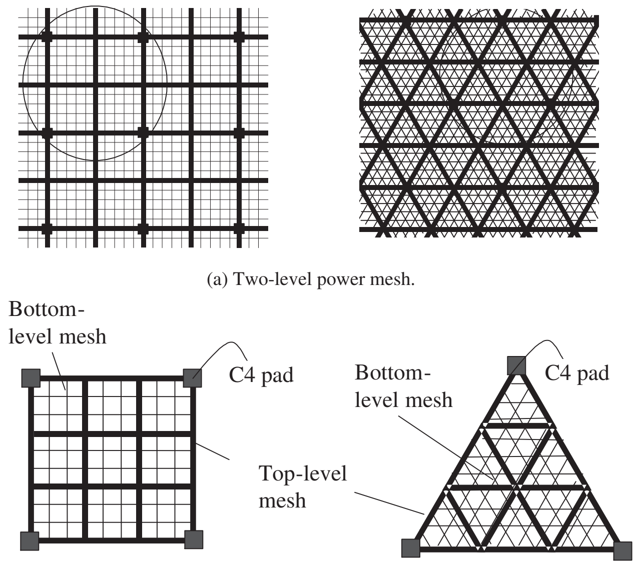
"H" clock tree [4] is widely used in the IC industry. In the H-tree, clock terminals are arranged in a symmetric fashion, and are connected by a planar hierarchy of symmetric "H" structures. When octilinear routing is allowed, the "H" structure can be replaced with an "X" structure, so that source-sink path (i.e., insertion) delay and total wirelength are decreased. However, significant undesirable overlapping (superposition) will occur between parallel interconnect wires in the X-tree.

With three uniform routing directions, a *Y clock tree* can be built as depicted in Figure 2(a), essentially giving a "distorted X-tree" with reduced wirelength and no superposed parallel wires. Let the distance between two adjacent clock terminals be 1. Path length from the clock source to clock terminal, as well as total wirelength, are compared with H-tree and X-tree in Table 2. The Y clock tree has a path length of $.7887\cdot(2^n-1)$, 21.1% less than the H-tree. Its total wirelength is $1.366\cdot 2^n(2^n-1)$, 8.9% less than H-tree, and 3.4% less than X-tree. Actually, the one-level Y-tree shown in Figure 2(b) is the optimal Euclidean Steiner Minimum Tree to connect four adjacent clock terminals $s1, s2, s3, s4$ and the clock source o. Thus the Y clock tree provides minimal total wirelength among all clock trees with similar symmetric structure. The further advantage of Y clock tree is that there is no overlapping of parallel interconnect wires. It can be shown:

Theorem 1 *Let the distance between two adjacent clock terminals be D. The minimum distance between two parallel interconnect wires is* $\frac{\sqrt{3}-1}{4}D$.

Proof. Suppose there is a coordinate system with a 0° x-axis, a 60° y-axis and the origin $(0,0)$ at the center of the main Y-tree structure (see Figure 2(b)). Then in a one-level Y-tree, the two bold interconnect wires that are parallel to the y-axis in the figure have x-coordinates of $\pm a$. In a two-level Y-tree, the lowest-level y-axis-parallel interconnect wires have x-coordinates of $\pm a \pm 2a$ and $\pm a \pm 2(a+b)$. Generally, in an n-level Y-tree, x-coordinates of the lowest-level y-axis-parallel interconnect wires are $\pm a \pm (2a \text{ or } 2(a+b)) \pm \ldots \pm (2^{n-1}a \text{ or } 2^{n-1}(a+b))$.

Since $a = \frac{D}{2}(1-\frac{\sqrt{3}}{3})$, and $(a+b) = \frac{D}{2}(1+\frac{\sqrt{3}}{3})$, the y-coordinates can be written as $(\pm 2^0 \pm 2^1 \pm \ldots \pm 2^{n-1})\cdot\frac{1}{2}D + (\pm 2^0 \pm 2^1 \pm \ldots \pm 2^{n-1})\cdot\frac{\sqrt{3}}{6}D$. These values cannot be zero because the values of $\pm 2^0 \pm 2^1 \pm \ldots \pm 2^{n-1}$ must not be zero, and the minimum absolute value among them is $a = \frac{D}{2}(1-\frac{\sqrt{3}}{3})$. Thus the minimal distance between two parallel interconnect wires in the Y clock tree is $\frac{\sqrt{3}}{2}a = \frac{\sqrt{3}-1}{4}D$.



(a) Two-level power mesh.

(b) Representative areas.

Figure 3: Power distribution networks and representative areas for M- and Y-architectures.

4 Y Power Distribution

Excessive voltage drop in the power grid can slow device switching speed and reduce noise margin. Robust power distribution within available area resource is critical to chip performance and reliability. Hierarchical mesh structures are widely used for power distribution in high performance chips because of their robustness [5]. In this section, we show that power distribution in the Y-architecture is not only natural, but achieves less IR drop than equally-resourced mesh distribution in the M-architecture.

Our comparison is based on the following model of the power distribution network.

- The power distribution network is constructed by a hierarchy of mesh structures connected by vias at crossing points of wires. Each mesh has equal wire spacing and wire width. Ignoring the resistance of vias,[4] we assume perfect contact at each crossing point.
- On top of metal layers, there are arrays of C4 power pads evenly distributed on the surface of the power mesh.
- Under the bottom-level mesh, there are devices connected to the wires of the bottom-level mesh. The devices are modeled as uniform current sinks and placed at crossing points of the bottom-level mesh.

[4]In practice, high current density on vias often causes reliability problems. In the Y-architecture, assuming same wire width, the area of intersection (overlap) between two adjacent-layer wires is larger than in the M-architecture. Hence, we can place a bigger via between adjacent layers or place more vias in the via array between adjacent layers to reduce resistance and current density for vias. Let A_Y, A_X, and A_M represent this area for Y-, X- and M-architectures, respectively. We have $A_Y = 1.1547A_M$ and $A_X = 1.414A_M$.

In state-of-art designs, there is a fairly large number (> 100) of power pads evenly distributed on the surface of the top-level power mesh [32]. It is reasonable to assume that the whole power mesh is an infinite resistive grid constructed by replicating the area surrounded by adjacent power pads. Figure 3 illustrates two-level power meshes and the *representative areas* in the M- and Y-architectures. Our analysis and circuit simulations consider only the worst-case IR-drop on the representative area. This method is also used in [11].

4.1 IR-Drop on Single-Level Power Mesh

Static IR-drop on a hierarchical power mesh depends largely on the top-level mesh since usually the top-level mesh is wider and coarser and most current flows along the top-level mesh. Here we analyze and compare the worst-case static IR-drop on a single-level power mesh in the M- and Y-architectures.

4.1.1 IR-Drop on Single-Level Power Mesh in the Y-Architecture

A single-level power mesh in the Y-architecture is abstracted as an infinite triangular resistive lattice with edge resistance R_Y.[5] We examine IR-drop in the triangular area with N_Y rows surrounded by three adjacent power pads[6]. In this case, the worst-case IR-drop appears at the center of this representative area. Each power pad supplies a current $I_Y = N_Y^2 i$ to the power mesh, where i is the current drain at each intersection on the mesh.

Assume there is a coordinate system with the origin at the center of the power mesh, and 0-degree and 120-degree lines used as m-axis and n-axis, respectively. We analyze the voltage drop between the node $(0,0)$ and the power pad at $(\frac{N_Y}{3}, -\frac{N_Y}{3})$ by considering currents from power pads and evenly distributed current sinks separately.

IR-drop caused by currents from power pads. Suppose that a current I_Y enters the lattice at the node (m_s, n_s) and leaves at infinity. The voltage drop for any node on the lattice is analyzed in [2]. The voltage drop between (m_s, n_s) and (m,n), denoted as $V_{(m_s,n_s)}(m,n)$, is given by the integral

$$\frac{I_Y R_Y}{2\pi} \int_0^{\pi/2} \left(1 - e^{-|(m-m_s)-(n-n_s)|x}\right) \cos(((m-m_s)+(n-n_s))y)/(\sinh x \cos y)\, \mathrm{d}y, \tag{1}$$

where $2\cosh x \cos y + \cos 2y = 3$. When $|(m-m_s)-(n-n_s)|$ is large, the voltage drop $V_{(m_s,n_s)}(m,n)$ can be approximated as

$$\frac{I_Y R_Y}{4\sqrt{3}\pi}[\ln((m-m_s)^2 + (n-n_s)^2 - (m-m_s)(n-n_s)) + c_1], \tag{2}$$

where $c_1 = 3.6393$ is a constant.

Let $V_{(m_s,n_s)}$ denote the voltage drop between $(0,0)$ and the power pad at $(\frac{N_Y}{3}, -\frac{N_Y}{3})$ caused by the current source at (m_s, n_s). According to the above approximation, we have

- when $(m_s,n_s) = (\frac{N_Y}{3}, -\frac{N_Y}{3})$, $V_{(m_s,n_s)} \approx (I_Y R_Y / 4\sqrt{3}\pi)(2\ln N_Y - \ln 3 + c_1)$;
- when $(m_s,n_s) \neq (\frac{N_Y}{3}, -\frac{N_Y}{3})$, $V_{(m_s,n_s)} = V_{(m_s,n_s)}(0,0) - V_{(m_s,n_s)}(\frac{N_Y}{3}, -\frac{N_Y}{3}) \approx (I_Y R_Y / 2\sqrt{3}\pi)\ln\frac{D_0}{D_s}$, where D_s is the Euclidean distance between (m_s,n_s) and $(\frac{N_Y}{3}, -\frac{N_Y}{3})$, and D_0 is the Euclidean distance between (m_s,n_s) and $(0,0)$. The constant $c_2 = \sum_{(m_s,n_s)\neq(\frac{N_Y}{3},-\frac{N_Y}{3})} \ln\frac{D_0}{D_s}$ can be computed by a simple algorithm, which calculates the summation for all the current sources within a circle around the origin. As the radius of the circle increases, the summation converges to a value of $c_2 = -1.173679$.

Therefore, if only currents from power pads are considered, the voltage drop between $(0,0)$ and the power pad at $(\frac{N_Y}{3}, -\frac{N_Y}{3})$ is

$$V_{source} = \sum_{(m_s,n_s)} V_{(m_s,n_s)} = \frac{I_Y R_Y}{2\sqrt{3}\pi}(\ln N_Y + C_Y), \tag{3}$$

where $C_Y = c_1/2 - \ln 3/2 + c_2 = 0.09666$.

IR-drop caused by evenly distributed current sinks. Next, we consider the voltage drop caused by current sinks at the intersections of the power mesh. If the voltage between $(0,0)$ and (m,n) is denoted by $V_{sink}(m,n)$, by a combination of Ohm's and Kirchhoff's Laws we have

$$\begin{aligned} &V_{sink}(m-1,n) + V_{sink}(m+1,n) + V_{sink}(m,n+1) \\ &+V_{sink}(m,n-1) + V_{sink}(m-1,n-1) \\ &+V_{sink}(m+1,n+1) - 6V_{sink}(m,n) = iR_Y. \end{aligned} \tag{4}$$

If the resistive lattice is regarded as a discrete approximation to a continuous resistive medium, we will obtain a potential function proportional to D^2, where D is the Euclidean distance from the origin. Therefore, we assume the following representation for the voltage between $(0,0)$ and (m,n):

$$V_{sink}(m,n) = k\,(m^2 + n^2 - mn), \tag{5}$$

where k is a constant. Equation (4) then yields

$$V_{sink}(m,n) = \frac{iR_Y}{6}\,(m^2 + n^2 - mn). \tag{6}$$

When only current sinks are considered, the voltage drop between $(0,0)$ and the power pad at $(\frac{N_Y}{3}, -\frac{N_Y}{3})$ is

$$V_{sink} = V_{sink}(\frac{N_Y}{3}, -\frac{N_Y}{3}) = \frac{I_Y R_Y}{18}. \tag{7}$$

Verification of Worst-Case IR-Drop. From the above analysis, we obtain the voltage drop at the center:

$$\mathbf{V_Y = V_{source} + V_{sink} \approx \frac{I_Y R_Y}{18} + \frac{I_Y R_Y}{2\sqrt{3}\pi}(\ln N_Y + C_Y)}, \tag{8}$$

where $C_Y = 0.09666$.

To verify the above formula for worst-case IR-drop on the single-level power mesh, we use HSpice to simulate various power meshes with different values of N_Y's. Since the problem is linear in nature, in our experiments the resistance of each wire segment R_Y is simply set to be $1K\Omega$, and the total current drain in the area I_Y is set to be $1mA$. We list simulation results for N_Y from 3 to 21 in Table 3, and compare them with the estimated values from the formula. The results show that the formula is accurate, with error less than 1%.

4.1.2 Comparing IR-Drop on Single-Level Power Mesh

For a single-level power mesh in the M-architecture, worst-case IR-drop is analyzed and verified in [34]. Suppose the power mesh has edge resistance R_M, number of rows within the representative area

[5] Note that for a uniform mesh with fixed total routing area, the edge resistance is independent of the number of metal lines on the mesh. When the number lines increases, wire pitch and wire width decreases with the same ratio, and the edge resistance remains the same.

[6] E.g., for the top-level Y-architecture mesh shown in Figure 3(b), N_Y is equal to 3.

Table 3: Simulation results for worst-case IR-drop on the single-level power mesh in the Y-architecture, compared to estimated values (mV).

N_Y	IR-Drop	Estimated IR-Drop	Error
3	166.67	165.39	1.28
6	229.17	229.08	0.09
9	266.36	266.34	0.02
12	292.78	292.77	0.01
15	313.28	313.27	0.01
18	330.03	330.03	0.00
21	344.20	344.19	0.00

Table 4: IR-drop improvements in single-level Y-mesh vs. M-mesh.

N_M	Estimated IR-Drop (mV) in M-mesh	IR-Drop Impr. (%) with Y-mesh
2	214.25	10.78
3	278.78	8.28
4	324.56	7.11
5	360.08	6.41
6	389.09	5.93
7	413.63	5.58
8	434.88	5.31
9	453.63	5.09

N_M and current supplied by each power pad I_M, the worst-case IR-drop on the single-level Manhattan (M-) mesh is:

$$\mathbf{V_M} \approx \frac{\mathbf{I_M R_M}}{\mathbf{8}} + \frac{\mathbf{I_M R_M}}{\mathbf{2\pi}}(\ln \mathbf{N_M} + \mathbf{C_M}), \qquad (9)$$

where $C_M = -0.1324$.

To fairly compare the Y-mesh and M-mesh, we constrain the two meshes to have the same wire material and thickness, cover the same area (same total current drain) with the same wiring resource, and have the same number of crossing points and power pads. Therefore, we have $R_Y = \sqrt{3}R_M$, $I_Y = I_M$, and $N_Y = N_M$. According to Equations (8) and (9), worst-case IR-drop on the single-level Y-mesh is less than that on the M-mesh by

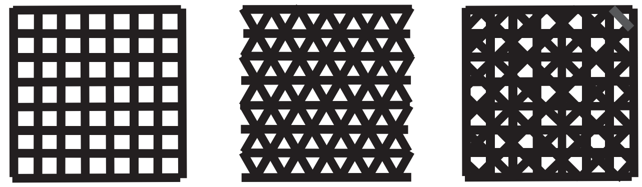
$$\Delta \mathbf{V} = \mathbf{V_M} - \mathbf{V_Y} = \mathbf{c I_M R_M}, \qquad (10)$$

where $c = 0.02309$. We list IR-drop improvements with Y-mesh for different values of N_M. The number of wire lines between two adjacent power pads on the top-level power mesh is usually small [34]. When $N_M = 4$, static IR-drop improvement of the Y-mesh over M-mesh is 7.1%.

4.2 IR-Drop on Hierarchical Power Mesh

In practice, power is distributed through a hierarchy of six or more metal layers. In this section, we simulate hierarchical power networks for the Y- and M-architectures using HSpice, explore different configurations of power networks, and compare the best solutions. We assume an equal sum of routing resources (i.e., total routing area) for Y- and M-architecture power distribution across layers M6, M5 and M4. In our experiment below, we set the total wiring area of M6, M5 and M4 to be 52% of the total representative area. The representative area for the Manhattan mesh is set to be a 1.2mm by 1.2mm square. To achieve the same power pad density, the representative area for the Y power grid is an equilateral triangle with edge length 1.289mm. Further details of our comparison are as follows.

- Layer thickness and resistivity parameters of a 6-layer process are taken from TSMC 0.13μm copper process information [28]. Layer thicknesses are 0.33μm for M1, 0.36μm for M2-5, and 1.02μm for M6.



(a) (b) (c)

Figure 4: Routing grids in M-, Y- and X-architectures.

- M1-M3 power distribution is native to library cells and blocks, requiring a common interface (0-degree) at M4. Power routing in M1-3 has the same pitch in both the Y and Manhattan solutions: M1 has pitch of 8μm and wire width of 2μm, M2 has pitch of 60μm and wire width of 4μm, and M3 has pitch of 60μm and wire width of 4μm. M4 pitch is fixed at 75μm to enable matchup with M1-3 macros and an apples-to-apples comparison.
- Allowed values of wiring separations (= pitches) on M5 and M6, denoted by S5 and S6, are $\{600\mu m, 300\mu m, 150\mu m, 75\mu m\}$. Allowed percentages of total wiring area used on M4 and M5, denoted as P4 and P5, are $\{10\%, 20\%, 30\%, 40\%, \ldots, 80\%\}$.
- $1V$ voltage sources are placed at the corners of representative areas. Each current sink on M1 (between two adjacent vias) is 5.21×10^{-7}A.

All combinations of wire pitch and wire width of M4, M5, and M6 are exhaustively searched. In the best M-architecture configuration, M6 has wire pitch of 300μm and uses 70% of the power routing resource; M5 has wire pitch of 75μm and uses 20% of the resource. The IR-drop produced by this configuration is 38.5mV. In the best Y-architecture configuration, M6 has pitch 600μm and uses 70% of the power routing resource, while M5 has pitch 150μm and again uses 20% of wiring area. The IR-drop is 35.2mV, which is 8.6% smaller than that of the best M-architecture solution. Ongoing research seeks a more general and formal comparison.

5 Routability in the Y-Architecture

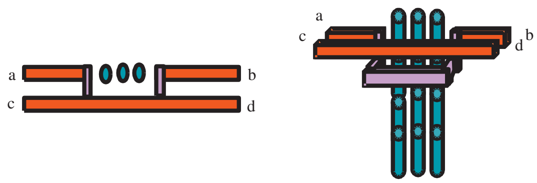
5.1 Uniform Routing Grid

A nice property of the Y-architecture is that there is a natural, uniform routing grid. Figure 4(a)(b) illustrates the routing grid in the M- and Y-architectures, wherein each routing layer has exactly the same wiring pitch. Figure 4(c) shows the X-architecture grid, where identical layer pitches imply that wire intersection points are not coincident. It is therefore difficult to find a natural, resource-efficient, uniform wiring grid in the X-architecture.

A uniform routing grid is expected to benefit large VLSI designs for three main reasons. (1) It enables continued use of today's dominating gridded routing algorithms. (2) Most advanced manufacturing processes require uniform width and spacing for M2 through M5, e.g., to simplify determination of legal via locations. Uniform pitch and dimension is also increasingly required for printability in subwavelength lithography. (3) The uniform routing grid can permit integral coordinates (even if absolute positions have irrational coordinates!), significantly simplifying detailed routing and design rule checking algorithms.

5.2 Via Tunnels and Via Tunnel Banks

Another advantage of the uniform global routing grid is that we can utilize *via tunnels* and *via tunnel banks* to avoid the fragmen-



(a) Top view.

(b) Side view.

Figure 5: Via tunnel in M-architecture.

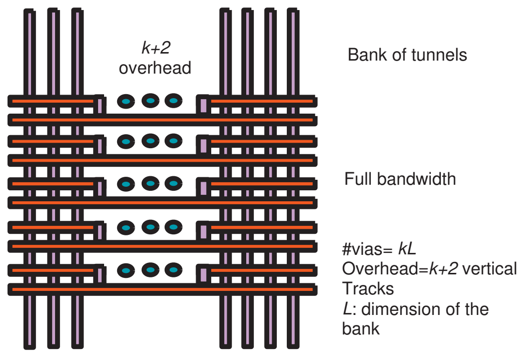


Figure 6: Bank of via tunnels in M-architecture.

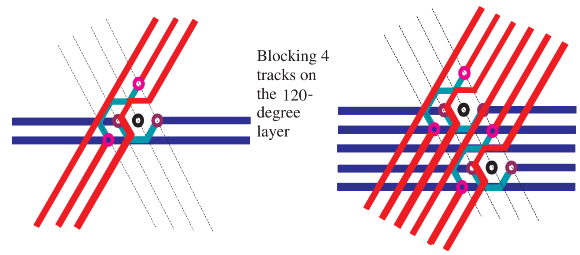
tation of routing resources caused by vias; this improves overall chip routability. In multi-layer routing, wire tracks are blocked on the layers that a via passes through. Traditional routing schemes scatter vias all over the chip, and this fragmentation of routing resources may cause serious wireability problems; this is called "via blockage effect". As we approach the 65nm technology node, this effect becomes more serious, since buffering of global wires introduces many via chains that go through all the way from the top-level metal down to the gate layer. We believe that the proposed use of via tunnels and via tunnel banks will reduce the via blockage effect and thus improve routability and wiring density.

Figure 5 shows an example of a *via tunnel* in Manhattan architecture. Figure 5 (a) is the birds-eye view and Figure 5(b) is a 3-D side view. There are two routing layers shown in the figure: the upper layer is for horizontal routing and the lower layer is for vertical routing. Terminals a and b are connected by detouring the horizontal wires around the via using the space on the vertical layer. Because the detour happens on the lower layer, it will not affect the wire between terminals c and d on the upper layer.

By aligning a number of *via tunnels* in vertical direction, we obtain a *bank of via tunnels*, which is shown in Figure 6. Suppose each *via tunnel* have k vias arranged in a horizontal line (in Figure 6, $k = 3$) and we align L *via tunnels* into a bank. In the resulting bank, all the horizontal tracks are free to route, and only $k+2$ vertical tracks are blocked. Note that there are a total of kL vias in the bank; without *bank of via tunnels* up to kL tracks could be blocked on each layer that the vias pass through. The use of *via tunnel banks* can significantly reduce the "via blocking" effect.

We have designed similar *via tunnel* and *bank of via tunnels* for the Y-architecture.

- Figure7(a) shows the bird view of a simple *via tunnel* design in the Y-architecture. In this example, we have three layers. From top to bottom, the routing direction of each layer is 60-degree, 120-degree, and 0-degree, respectively. The circle in the center represents a through via. We use the space in the middle layer to detour wires around the via. We can achieve blockage-free routing on the top and bottom layers, and have four tracks blocked on the middle layer.



(a) A via tunnel in Y-architecutre.

(b) Two tunnels aligned together.

Figure 7: Via tunnels and bank of via tunnels in Y-architecture.

- Similar to the construction of *banks of via tunnels* in M-architecture, we align the *via tunnels* together to obtain a *bank of via tunnels* in Y-architecture. Figure 7(b) illustrate how two *via tunnels* shown in Figure 7(a) are aligned along the 120-degree direction.

6 Conclusions

In this paper, we have examined key issues concerning the potential use of Y-architecture for semiconductor ICs, including throughput analysis, estimates of wirelength savings, clock and power distribution methodology, wireability, and manufacturing. We have not discussed such issues as graphics engine changes, computational-geometric data structures, number and coordinate systems, calibration of parasitic extraction (especially capacitance extraction) models, etc. Such "mundane" issues are part of the necessary groundwork for the eventual deployment of the Y-architecture, and the subject of ongoing work in our group, but are beyond the scope of the present paper.

Further research directions include: (1) theoretical analysis and high-impact designs or codes to demonstrate Y-architecture advantages; (2) more accurate estimations of expected wirelength improvement which formalizes interactions between nets; and (3) interfaces to current library cells and new Y-specific library cells. Many parts of a commercially successful Y-architecture methodology remain open. The Y-architecture also has applications beyond the die, e.g., it may be valuable on laminates used for multi-die integration, and on the buildup layers (e.g., BBUL [31]) that will replace traditional packages.

References

[1] F. Abboud, S. Babin, V. Charkarian, et al., "Design Considerations for an Electron-Beam Pattern Generator for the 130-nm Generation of Masks", *SPIE Symp. on Photomask and X-Ray Mask Technology VI*, SPIE Vol. 3748, 1999, pp. 385-399.

[2] D. Atkinson and F. J. van Steenwijk, "Infinite Resistive Lattices", *Am. J. Phys.* 67 (1999), pp. 486-492.

[3] H. B. Bakoglu, *Circuits, Interconnections, and Packaging for VLSI*, Addison-Wesley, 1990.

[4] H. B. Bakoglu, J. T. Walker and J. D. Meindl, "A Symmetric Clock Distribution Tree and Optimized High-Speed Interconnections for Reduced Clock Skew in ULSI and WSI Circuits", *Proc. IEEE Int. Conf. Computer Design*, Oct. 1986, pp. 118-122.

[5] S. Boyd, L. Vandenberghe and A. El Gamal, "Design of Robust Global Power and Ground Networks", *Proc. ACM/SIGDA Int. Symp. Physical Design*, 2001, pp. 283-288.

[6] P. Buck, Dupont Photomasks, *personal communication*, Nov. 2002.

[7] H. Chen, B. Yao, F. Zhou and C. K. Cheng, "Physical Planning of On-Chip Interconnect Architectures", *Proc. IEEE Int. Conf. Computer Design*, Sep. 2002, pp. 30-35.

[8] H. Chen, B. Yao, F. Zhou and C. K. Cheng, "The Y-Architecture: Yet Another On-Chip Interconnect Solution", *Proc. Asia and South Pacific Design Automation Conf.*, 2003, pp. 840-846.

[9] H. Chen, C.-K. Cheng, A. B. Kahng, I. Măndoiu and Q. Wang, "Estimation of Wirelength Reduction for λ-Geometry vs. Manhattan Placement and Routing", *Proc. ACM/IEEE Workshop on System Level Interconnect Prediction*, 2003, pp. 71-76.

[10] H. Chen, C.-K. Cheng, A. B. Kahng, I. Măndoiu, Q. Wang and B. Yao, "The Y-Architecture for On-Chip Interconnect: Analysis and Methodology", *Proc. Int. Conf. Computer Aided Design*, 2003, to appear.

[11] A. Dharchoudhury and R. Panda, "Design and Analysis of Power Distribution Networks in POwerPC Microprocessors", *Proc. Design Automation Conf.*, 1998, pp. 738-743.

[12] N. Garg, and J. Konemann, "Faster and Simpler Algorithms for Multicommodity Flow and other Fractional Packing Problems", *Proc. 39th Annual Symp. Foundations of Computer Science*, 1998, pp. 300-309.

[13] M. Igarashi, T. Mitsuhashi, A. Lee, et al., "A Diagonal-Interconnect Architecture and Its Application to RISC Core Design", *Proc. Int. Solid-State Circuits Conf.*, 2002, pp. 166-167.

[14] A. B. Kahng, I. I. Măndoiu and A. Z. Zelikovsky, "Highly Scalable Algorithms for Rectilinear and Octilinear Steiner Trees", *Proc. Asia and South Pacific Design Automation Conf.*, 2003, pp. 827-833.

[15] C. K. Koh and P. H. Madden, "Non-Manhattan Routing", *IEEE Trans. Computer-Aided Design*, to appear.

[16] P. H. Madden, "Congestion Reduction in Traditional and New Routing Architectures", to appear.

[17] T. Hildebrandt, "An Annotated Placement Bibliography", *ACM SIGDA Newsletter*, Dec. 1985, pp. 12-21.

[18] Carl Sechen, "Placement and Global Routing of Integrated Circuits Using Simulated Annealing", Ph.D. Dissertation, U. California, Berkeley, 1987, Chapter 2.

[19] M. Lemke, J. Gramss, H. J. Doering, et. al., "Advanced Writing Strategies for High-End Mask Making", *Proc. SPIE*, Vol. 3996, 2000, pp. 166-172.

[20] B. K. Nielsen, P. Winter and M. Zachariasen, "An Exact Algorithm for the Uniformly-Oriented Steiner Tree Problem", *Proc. 10th European Symp. on Algorithms*, Springer LNCS Vol. 2461, 2002, pp. 760-772.

[21] C. Progler, Photronics Inc., *personal communication*, Nov. 2002.

[22] M. D. Rostoker et al., "Hexagonal Architecture", *U.S. Patent*, No. US6407434B1, June 2002.

[23] M. D. Rostoker et al., "CAD for Hexagonal Architecture", *U.S. Patent*, No. US5822214, Oct. 1998.

[24] R. Scepanovic et al., "Microelectronic Integrated Circuit Structure and Method Using Three Directional Interconnect Routing Based on Hexagonal Geometry", *U.S. Patent*, No. US5578840, Nov. 1996.

[25] P. Saxena, N. Menezes, P. Cocchini and D. A. Kirkpatrick, "The Scaling Chanllenge: Can Correct-by-Construction Design Help?", *Proc. Intl. Symp. Physical Design*, 2003, pp. 51-58.

[26] D. Stroobandt and J. V. Campenhout, "Accurate Interconnection Length Estimations for Predictions Early in the Design Cycle", *VLSI Design, Special Issue on Physical Design in Deep Submicron* 10(1) (1999), pp. 1-20.

[27] S. Teig and J. L. Ganley, "Method and Apparatus for Considering Diagonal Wiring in Placement", *Int. Patent Application*, No. WO 02/47165 A2, June 2002.

[28] TSMC 0.13μm Design Rules. `http://www.tsmc.com.`

[29] S. Teig, "The X Architecture", *Proc. ACM/IEEE Workshop on System Level Interconnect Prediction*, 2002, pp. 33-37.

[30] `http://www.xinitiative.org.`

[31] Intel Research Webpage on Packaging. `http://www.intel.com/research/silicon/packaging.htm.`

[32] The ITRS Assembly and Packaging roadmap. `http://public.itrs.net.`

[33] WaterJet-Guided Laser In Wafer Cutting – Synova SA. `http://www.gemcity.com/downloads/synova01.pdf.`

[34] H. Chen, C.-K. Cheng, A. B. Kahng, I. I. Măndoiu, Q. Wang and B. Yao, "Optimal Sizing Analyses for Mesh-Based Power Plans", *unpublished manuscript*, 2003.

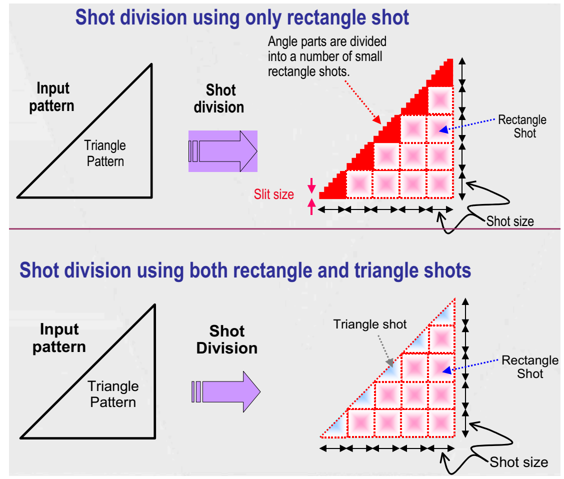


Figure 8: Toshiba machine triangle shots [30].

Appendix. Manufacturing and Other Issues

As is well-known from the example of the X Initiative [30], any new back end of the line (BEOL) architecture requires engagement throughout the mask and process infrastructure. According to our discussions with domain experts [6, 21], the Y-architecture presents a number of generic challenges to manufacturing; there are no show-stoppers, but engineering efforts will be required across several domains. Space limits preclude detailed discussion here, but we sketch several main points.

With respect to mask making, Vector Shaped Beam (VSB) ebeam lithography tools [1] create "shots" of varying shape and size by imaging the overlap of two apertures, typically both square. This allows a range of rectangular shots to be created and exposed on the mask. Existing Toshiba ebeam lithography systems can produce 45-degree pattern at high speed through the combination of one rectangular aperture and one with 45- and 135-degree edges [30]. The new JEOL JBX3030 tool [19] also has apertures to produce 45- and 135-degree edges. These new tools mitigate the write time implications of angled data since they provide an alternative to approximating an angled line with a series of small rectangles; Figure 8 illustrates mask fracturing using both rectangle and triangle shots versus mask fracturing using only rectangle shots. With successful experiences with 45-degree edges in mind, 60- and 120-degree edges can be printed with the availability of 30- and 60-degree angles in apertures.

Current support for angular edges is really focused on small edge segments rather than long lines. To produce long lines efficiently, it is necessary to have a pair of rectangular apertures rotated to still produce rectangular shots, but rotated to the desired angle. On the other hand, if the Y architecture is applied only to the upper, lower resolution metal layers - as we have proposed - the write time issue could be solved if the masks could be made with optical (laser) lithography (e.g., ETEC Alta writers), where throughput is independent of angular edges.

The potential of non-rectangular die also presents challenges to package I/O design and dicing. Current side-to-side die sawing cannot cut hexagonal dies due to the silicon lattice structure. New technologies, such as waterjet-guided laser [33], are emerging to confront the challenges.

There are other challenges related to inspection, exposure, repair, metrology and pattern compensation. Ultimately, the deployment of the Y-architecture will depend on careful engineering, and provable cost reductions vis-a-vis achievable design quality with pervasive 60- and 120-degree wiring.