Efficient Method of Power Management on System on Chip Communication Using Steiner Graph

K. Nirmaladevi and J. Sundararajan

***Abstract*—Power consumption become the major factors limiting the speed of very-large-scale integration (VLSI) circuits, while interconnect is becoming a primary power consumer. These factors bring new demands on the communication architecture of system-on-chips (SoCs). Current bus architectures such as AMBA, Core connect, and Avalon are convenient for designers but not efficient on power. This paper proposes a physical synthesis scheme for on-chip buses and bus matrices to minimize the power consumption, without changing the interface or arbitration protocols. By using a bus gating technique, data transactions can take shortest paths on chip, reducing the power consumption of bus wires to minimal. Experiments indicate that the gated bus from our synthesis flow can save more than 91% dynamic power on average data transactions in current AMBA bus systems, which is about 5–12% of total SoC power consumption, based on comparable amount of chip area and routing resources.**

***Index Term*—Bus gating, System on Chip, AMBA protocol.**

I. Introduction

A system on a chip or system on chip (SoC or SOC) is an integrated circuit (IC) that integrates all components of a computer or other electronic system into a single chip. It may contain digital, analog, mixed-signal, and often radio-frequency functions all on a single chip substrate. The systems are characterized by a high level of parallelism, due to the presence of multiple processors, and large bandwidth requirements, due to the massive scale of component integration. The choice of communication architecture in such systems is of vital importance because it supports the entire inter-component data traffic and has a significant impact on the overall system performance. The Advanced Microcontroller Bus Architecture (AMBA) is used as the on-chip bus in SoC designs. Since its inception, the scope of AMBA has gone far beyond microcontroller devices, and is now widely used on a SoC parts including applications processors used in modern portable mobile devices like Smartphone's. The AMBA protocol is an open standard, on-chip interconnect specification for the connection and management of functional blocks in a SoC. It facilitates right-first-time development of multi-processor designs with large numbers of controllers and peripherals. The first AMBA buses were Advanced System Bus (ASB) and Advanced Peripheral Bus (APB). In its 2nd version, AMBA 2, ARM added AMBA High-performance Bus (AHB) that is a single clock-edge protocol. In 2003, ARM introduced the 3rd generation, AMBA 3, including AXI to reach even higher performance interconnect and the Advanced Trace Bus (ATB) as part of the Core Sight on-chip debug and trace solution.. The objective of the AMBA specification is to facilitate *right-first-time* development of embedded microcontroller products with one or more CPUs, GPUs or signal processors,be technology independent, to allow reuse of IP cores, peripheral and system macro cells across diverse IC processes, encourage modular system design to improve processor independence, and the development of reusable peripheral and system IP libraries minimize silicon infrastructure while supporting high performance and low power on-chip communication.

Manuscript received July 03, 2012; revised September 6, 2012.

K. Nirmaladevi is with the Department of ECE, Paavai Engineering College, Namakkal, TamilNadu, India (e-mail: nirmalnkl03@gmail.com).

J. Sundararajan is with the Pavai College of Technology, TamilNadu, India (e-mail: dharsini_71@yahoo.co.in).

II. Existing System

A. *"Low Power Gated Bus Synthesis using Shortest-Path Steiner Graph for System-on-Chip Communications"[2]*

A low power design technique of gated bus which can greatly reduce power consumption on state-of-the-art bus architectures. By adding demultiplexer and adopting a novel shortest-path Steiner graph, a flexible tradeoff between large power reductions versus small wire length increment.System-on-chips (SoC) are nowadays being developed with increasing complexity and on-chip communication demand. However global on-chip wires which are to meet this demand do not scale well towards 35nm feature size. As a result, global interconnect is becoming a bottleneck of improving system performance and power consumption. Bus architectures are therefore regarded as an important aspect in low power SoC design. Current state-of-the-art bus architectures including AMBA, Core Connect, Avalon, AMBA AHB bus matrix, etc, provide solutions for SoC communications. These bus circuits may consume as much power as other major components such as processor, memory controller and cache. Therefore, reducing power on buses makes significant contribution to the whole system's power consumption. Techniques of clock gating and power gating have been widely and effectively used to reduce power consumption of electronic systems, among which clock gating reduces dynamic switching power, and power gating reduces static leakage power. Both of the techniques save power by masking off signal/power when/where it is not needed. Since a clock distribution network consumes more than 40% of the total power budget of a CMOS circuit, clock gating has become a necessity in most digital circuit designs. Bus connections in current communication architectures are facing a similar (if not the same) situation as clock networks.

DOI: 10.7763/IJIEE.2012.V2.250

B. Efficient Algorithms for the Minimum Shortest Path

SPSA (minimum shortest path Steiner arborescence) problem has various applications in the areas of physical design of very large-scale integrated circuits (VLSI), multicast network communication, and supercomputer message routing.Several heuristics and exact algorithms for the MSPSA problem with applications to VLSI physical design. Experiments indicate that the heuristics generate near optimal results and achieve speedups of orders of magnitude over existing algorithms. The MSPSA problem is a special case of the minimum Steiner arborescence (MSA) problem. The rectilinear version of the MSPSA problem is called the minimum rectilinear Steiner arborescence (MRSA) problem. The MSPSA and MRSA problems have applications to performance-driven VLSI physical design.

Exact methods for the MRSA problem can be classified into:

1) dynamic programming (DP)；
2) integer programming；
3) branch-and-bound (BNB)/enumeration techniques；
4) mincost max-flow (MCMF) technique.

The MSPSA problem cannot be approximated within a factor of times optimal unless deterministic polylog space coincides with nondeterministic polyglot space.

III. Proposed System

We optimize on-chip bus communications on the tradeoffs between minimal power, maximal bandwidth, and minimal total wire length. We use the protocols of AMBA AHB and AXI, since they are most popular in industrial designs. Based on AMBA protocols, we modify the bus structure using a "bus gating" technique, and apply optimizations which is biased toward minimal power, but also favors bandwidth and routing resource. Heuristics are devised to construct a minimal shortest-path Steiner graph, and to reduce its scale with a minimal increment on path lengths. The overall optimization flow can be viewed as three major steps:Step 1: generating the shortest-path Steiner graph H (for minimal power);Step 2: deciding edge weights on H (for adequate bandwidth);Step 3: applying incremental modifications on H(for minimal wire length).

A. Advantages Of Proposed System

The proposed system provides Less Routing resource, Efficiency on bus lines is maximized without the need to redesign system components and IP modules, Routing resource is also reduced without compromising low power.Standard on-chip buses like AMBA were designed to enable fast and convenient integration of system components into the SoC, where simplicity is one of the major objectives. When the bus power consumption comes to a significant level that we cannot afford to ignore, power optimization will be desirable. We introduce a "bus gating" technique to minimize the power on bus lines with a small compromise on design simplicity.

Traditionally used hierarchical shared bus communication architectures such as those proposed by AMBA, CoreConnect and STbus can cost effectively connect few tens of cores but are not scalable to cope with the demands of very high performance systems. Point-to-point communication connection between cores is practical for even fewer components. Network-on-Chip (NoC) based communication architectures have recently emerged as a promising alternative to handle communication needs for the next generation of high performance designs.

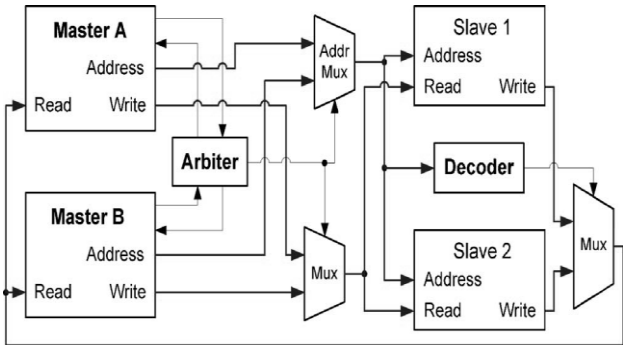


Fig. 1. 1. AMBA AHB bus.

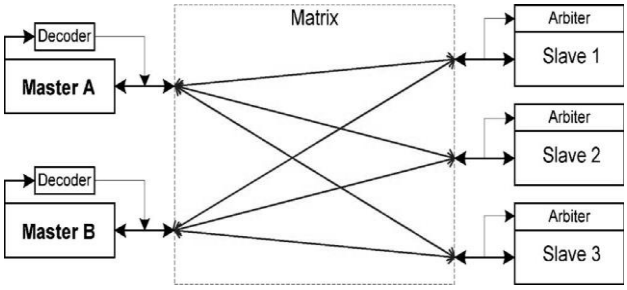


Fig. 1. 2. AMBA AXI full bus matrix (sketch).

Gated bus is a solution to save the wasted dynamic power. The simplest way is to add a de-multiplexer after each multiplexer , and add a de-multiplexer after each master, so that the signals only propagate to where they are needed. This method works in a similar way as clock gating and can be even more effective because the signal receivers here have much less complex behaviors than in a clock tree.

IV. Single Bus Architecture

A. Single Bus

Tree structured buses, distributing the multiplexer and de-multiplexer into the wire net helps to save both power and wires. While the single multiplexer needs independent lines from every sender, Lines can be shared with distributed multiplexers and form Steiner arborescence and arborescence is a Directed tree such that every root-to-leaf path is shortest. On the receivers' side with distributed de-multiplexers, the bus lines change from a rectilinear Steiner minimum tree to a minimum rectilinear Steiner arborescence (MRSA).total bus wire length can be reduced by the distributing the multiplexer/de-multiplexers ,while the dynamic power can also be reduced at the same time. Based On the same tree topology, effective bus gating can be applied by distributing the control over the entire tree.

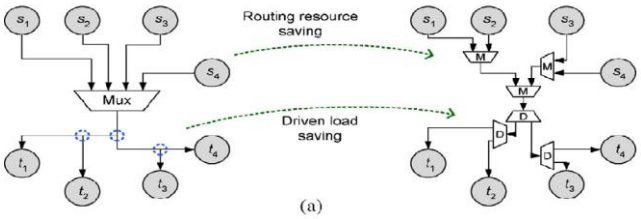


Fig. 2. Single bus: bus gating using distributed mux and de-mux.

B. Bus Matrix Structure

On bus matrices, however, simply adding de-multiplexers may increase the total wire length, because when the number of master-to-slave paths becomes large, each path will need its own bus wires. To reduce wire length in the bus matrix, also to further reduce power on the basic bus, we adopt the structures of Steiner graphs. A Steiner graph is a generalization of Steiner trees, without the limitation of tree structure that there is only one root placed at a certain point, which cannot be on the shortest path of every connection. By removing the constraint of tree topologies, we gain higher freedom to choose shortest paths for reduced power on data transactions, and to let the paths share wires for reduced routing congestion.

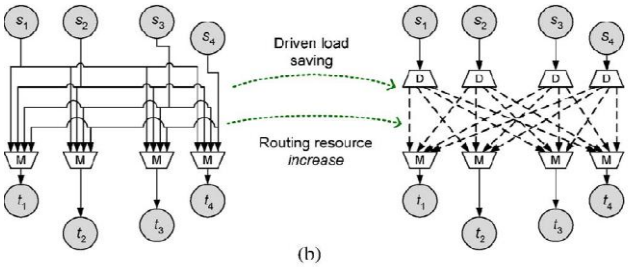


Fig. 3. On bus matrix: Bus gating using distributed mux and de-mux.

C. Steiner Graph Reduction:

Iterative graph reduction process by parallel segment merging operations. Effect of merging parallel segments in narrow rectangles. The total edge length is greatly reduced, while the increment on average path length is relatively small. Although fewer edges will generally result in larger edge weight, the total weighted edge length can still be reduced by this merging operation due to improved wire sharing among paths. If the added total wire length is really reduced, we keep the merging operation and continue to the next iteration, otherwise discard the operation. Eventually, there will be no positive wire length reduction in the graph, and we have a series of bus matrix graphs with decreasing wire length and increasing path lengths, where a comprise can be chosen. Reduction on total edge length is by combining the vertical segments of length h and changing the lengths of related horizontal connections.

D. Bus Matrix Control Design

Slave device has an arbiter which handles the requests from masters and decides the connection. The result is sent to the central switch control unit, where all the connection paths are stored. Depending on the set of active paths, the central switch control sends control signals to all the switches on each path, which together instantly create the master-to-slave connection requested by the master device. A bus switch is basically a crossbar plus an auxiliary local control which remembers each path going through. The local control handles two types of requests from the central switch control, create connection and dispose connection.

Design Flow with Gated Bus Synthesis: The bus gating technique may bring some additional complexity in the design flow. Traditionally, physical level design starts from gate level net list, and goes through placement, routing, timing analysis, verification, and so on. With bus gating, since the buses are usually included in the system with wires and control units, floor plan and placement depend on the bus connections, while the topology of the gated bus may depend on floor plan and placement. To resolve this loop of dependency, we need to change the design flow by inserting the bus gating stage into placement and routing. Since the updates are limited to the bus or bus matrix part of the system, the process can be controlled in small scale and mostly automated. So provided with appropriate algorithms, the impact of bus gating on design flows can be minimal.

Maximum Bandwidth Bus Matrix Formulation: To meet the demand of the communication graph GC, we define the bus matrix graph based on a Steiner graph of GC. Every connection path should take the shortest rectilinear path for minimal communication power, i.e., the path from a to b has the length of Manhattan distance _P(a) − P(b)_1. Path definition is natural. Definition 3: For communication graph GC = (Vs, Vt,A) and placement function P : Vs _Vt _→ R2, a bus matrix graph is a weighted graph _ = (V, E, ω) with placement P _ : V _→ R2.The objective is to find the bus matrix graph with minimal total wire length L(_) = _(u,v)∈E ω((u, v))_P(u) − P(v)_1. The bus matrix graph is defined above to have the capability of efficient communications. Constraints i) and ii) ensure that the graph covers all the devices. Constraint iii) dictates that for any set of disjoint arcs in A, there is a set of connection paths _, where each path is shortest (by iii-b) and the weighted edges in _ can hold all the paths in _ (by iii-c). The total weighted edge length, i.e., total wire length is to be minimized. So the bus matrix we are looking for should support all possible communication patterns, consume minimal power, and use minimal routing resources. Fig. 4 shows an example of a bus matrix graph connecting four masters' s0, s1, s2, s3 and three slaves t1, t2, t3. Five communication arcs are present: s1 may access t2 and t3, and t1 may be accessed by s0, s2, and s3. The single weight edges in Fig. 4 (by solid segments) are adequate for this requirement. Notice that (s0, t1) is the only arc having more than one shortest paths. And when its connection is on, s2 and s3 cannot access t1 at the same time, i.e., bus lines "s2 ↔ t1" and "s3 ↔ t1" are both open.

Depending on s1's connection, since s1 can take at most one of "s1 ↔ s2" and "s1 ↔ s3," the connection from s0 can always choose the one other than s1's and find an open path to t1.This formulation defines an ideal high bandwidth low power on-chip communication solution, but with limited practicality. Because first, minimization on the wire length of _ is computationally expensive due to the exponentially increasing combinations of arc subset A _. And even if we pre-compute the optimal solution, it is still impractical to store the path sets for all the subsets, or to compute the path

set _ in real time. Another problem is that, if the communication pattern changes dynamically, when some connections are still on but need to change paths, it may induce extra delay or timing issues. To make a practical bus matrix achievable with fast response, small silicon footprint and good power efficiency, we slightly weaken the formulation as follows.

Practical Bus Matrix Synthesis Formulation: Given a communication graph GC and its placement, we define another bus matrix graph with fixed paths for the arcs in GC, i.e., each pair of master-slave connection always takes the same path regardless of other connections. Definition : For communication graph GC = (Vs, Vt,A) and placement function P : Vs _Vt _→ R2, a bus matrix graph is a weighted graph H = (V, E, ω) with placement P _ : V _→ R2 and a set of fixed path ρ : A .In this formulation, the constraints are similar to those in definition 3, except that the connection paths no longer depend on the communication pattern. The same case in Fig. 4 has a different solution here because of the reduced flexibility on paths. Since the path from s0 to t1 is fixed, it conflicts with either path (s1, t2) or path (s1, t3). For the objective of minimal wire length, we should assign weight 2 on edge (s1, t3), and arc (s0, t1) always takes path "s0 ↔ s1 ↔ s3 ↔ t1." We can see that the total wire length of the optimal solution L(Hm) may be larger than that of the ideal formulation by definition 3. However, the fixed paths can be easily stored in bus control units, so that the entire communication protocol can be implemented with very small consumption on area and power. Therefore, the formulation based on definition 4 is more preferable for practical use. Algorithms based on this formulation are elaborated in the following two sections.

V. Bus Matrix Graph Construction

The flow we use is to first construct a shortest-path Steiner graph based on the given placement of Vs ∪ Vt and communication graph GC, and then decide the weight ω(e) on each edge. The single-source case is the MRSA problem, which is well studied in previous work such as. Our shortest-path Steiner graph is constructed by multiple iterations of a revised MRSA construction.

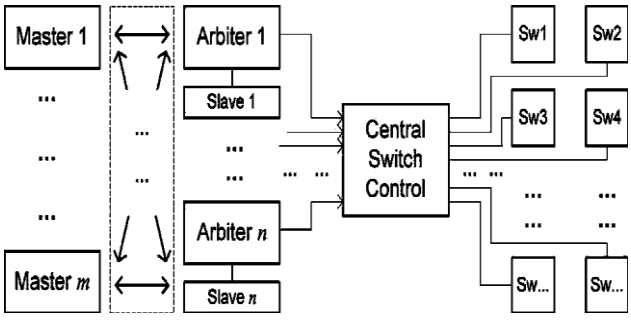


Fig. 4. Control on switches in a bus matrix.

Steiner Graph Reduction: Since high bandwidth bus matrices will need significantly more wires to support parallel communications across the chip, routing resource may become another limitation as more components are integrated into SoCs and interactions increase. Especially when the components are placed in irregular placement instead of cell arrays, the shortest-path Steiner graph generated by the algorithm may contain a lot of loops, which bring additional wire length. Thus look for changes in the graph structure which can significantly reduce the wires, while preserving the short paths at the same time.There are some long and narrow rectangles formed by the graph edges, and the long double edges are necessary if we want all the connections to be shortest. But when the double edges are geometrically very close to each other, combining them into one edge only slightly increases the length of some connections, while possibly saving much more wire length. The total edge length is greatly reduced, while the increment on average path length is relatively small. Although fewer edges will generally result in larger edge weight, the total weighted edge length (wire length) can still be reduced by this merging operation due to improved wire sharing among paths.

VI. Conclusion

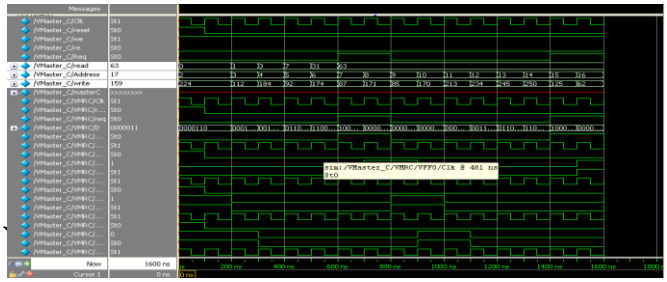
The system-on-chip communications referring to the AMBA AHB bus (matrix) architecture, weaknesses of original bus matrices, such as low power efficiency and low wire efficiency, are resolved by using a Steiner graph structure. Therefore, we believe bus matrix architectures will be widely applied for efficient communications in various future systems. Thus the approach for the synthesis of bus matrix communication architecture for SoC designs with high bandwidth requirements. The synthesis approach satisfies all throughput performance constraints in the design, while generating an optimal bus matrix topology having a minimal number of busses, as well as values for parameters such as bus speeds Results from the synthesis of an AMBA AXI based bus matrix. The results show promising potentials of bus matrices for low power and high performance on-chip communications.

VII. Future Enhancement

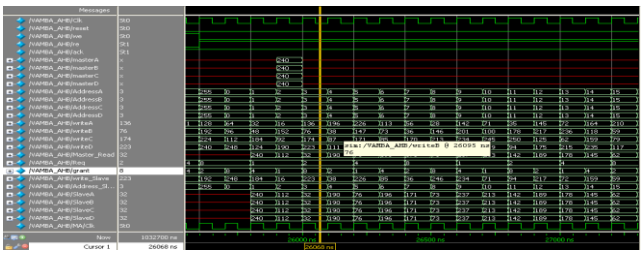
In the future work, the power can be reduced by using the advanced techniques of bus gating and the efficiency of wire can be reduced by using the advanced algorithm of Steiner Graph. More improvements can be explored in future works on formulations, algorithms, and the overall optimization flow with advanced AMBA matrix based architecture.

VIII. Screen Shots

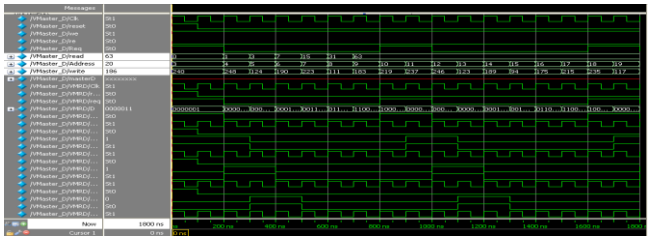
A. Master A-power is Reduced:



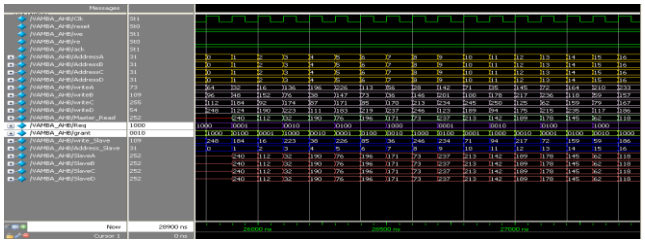
B. Master B when wire is Applied:



C. Simulation Of Arbitre:



D. Simulation:



Total wire length vs. total edge length along parallel segment merging operations.

First decreasing (less edges)

Then increasing (longer paths)

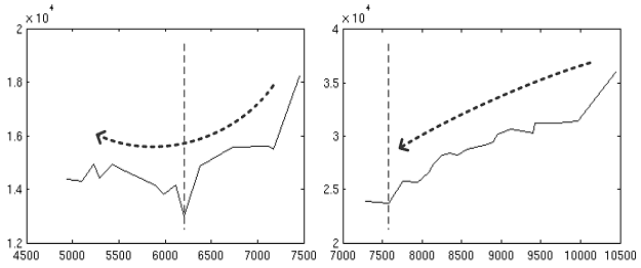


Fig. 5. (a). Tradeoff between wire & bandwidth

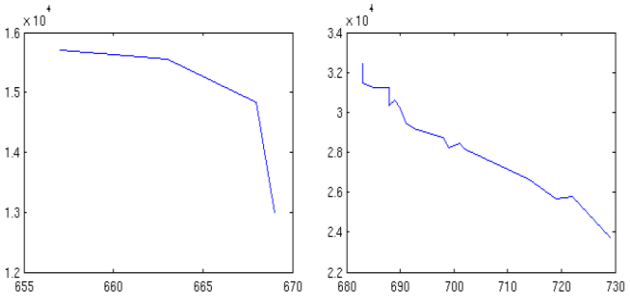


Fig. 5. (b). Tradeoff between wire & Power

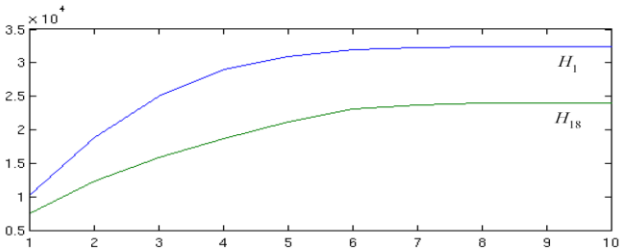


Fig. 5. (c). Total wire length vs. total edge length along parallel segment merging operations

REFERENCE

[1] Renshen Wang, Yulei Zhang, *Student Member, IEEE*, Nan-Chi Chou, Evangeline F. Y. Young, Chung-Kuan Cheng, *Fellow, IEEE*, and Bus Matrix Synthesis Based on Steiner Graphs for Power Efficient System-on-Chip Communications Ronald Graham, 2011.

[2] Bollob´as, D. Coppersmith, and M. Elkin, "Sparse distance preservers and additive spanners," SIAM J. Discrete Math., vol. 19, no. 4, pp. 1029–1055, 2005.

[3] J. Y. Chen, W. B. Jone, J. S. Wang, H. I. Lu, and T. F. Chen, "Segmented bus design for low power systems," *IEEE Trans. Very Large Scale Integr .Syst.*, vol. 7, no. 1, pp. 25–29, Mar. 1999.

[4] J. Cong, A. B. Kahng, and K.S. Leung, "Efficient algorithms for the minimum shortest path Steiner arborescence problem with applications to VLSI physical design," *IEEE Trans. Comput.-Aided Design*, vol. 17, no. 1, pp. 24–39, Jan. 1998.

[5] J. Cong, A. B. Kahng, G. Robins, M. Sarrafzadeh, and C. K. Wong, "Provably good performance-driven global routing," *IEEE Trans. Comput.-Aided Design*, vol. 11, no. 6, pp. 739–Jun. 1992.

[6] J. Griffith, G. Robins, J. Salowe, and T. Zhang, "Closing the gap: Nearoptimal Steiner trees in polynomial time," *IEEE Trans. Comput.-Aided Design*, vol. 13, no. 11, pp. 1351–1365, Nov. 1994.

[7] L. A. Ca, Q. Wu, M. Pedram, and X. Wu, "Clock-gating and its application to low power design of sequential circuits," in *Proc. IEEECustom Integr. Circuits Conf.*, vol. 47. Mar. 2000, pp. 415–420.

[8] M. Donno, A. Ivaldi, L. Benini, and E. Macii, "Clock-tree power optimization based on RTL clock-gating," in *Proc. ACM/IEEE Des. Autom. Conf.*, Jun. 2003, pp. 622–627.

[9] R. Ho, K. W. Mai, and M. A. Horowitz, "The future of wires," in *Proc. IEEE*, vol. 89, no. 4, pp. 490–504, Apr. 2001.

[10] W. Dally, "Keynote: The end of denial architecture and the rise of throughput computing," in *Proc. ACM/IEEE Des. Autom. Conf.*, Jul. 2009, p. xv.

K. Nirmaladevi received her M.Tech Degree in the field of VLSI Technology from SASTRA University, TamilNadu, India. She is currently pursuing her PhD Degree in Anna University, Coimbatore. She is currently working as a Associate Professor in Paavai Engineering College, Namakkal, TamilNadu, India. Her Research Interests are VLSI SystemOnChip.

J. Sundararajan received his PhD Degree from Anna University, Chennai, India in the field of Biomedical Signal Processing and Instrumentation. He is the Principal in Pavai College of Technology, Namakkal, TamilNadu, India. He has published more than 50 international papers in the field of Bio Signal Processing. His Research Interests are Bio Signal Processing, Computer Networks, Neural Networks and Image Processing.