Low Power Multilevel Interconnect Networks Using Wave-Pipelined Multiplexed (WPM) Routing

Ajay Joshi, Vinita Deodhar and Jeffrey Davis
School of ECE, Georgia Institute of Technology, Atlanta, GA 30332, USA
{joshi, vinita, jeff}@ece.gatech.edu

Abstract

A low power multilevel interconnect architecture that uses wave-pipelined multiplexed (WPM) interconnect routing is proposed in this paper. WPM takes advantage of existing interconnect idleness and implements low-overhead wire sharing to reduce the number of routing tracks required for intra-macrocell communication. It is shown that the extra available routing area could then be used to redesign the interconnect network to substantially reduce coupling capacitance and driver sizes. System-level simulation reveals that the systematic application of WPM reduces total power of a 40M transistor macrocell by almost 28% without any loss in performance.

1. Introduction

For future deep sub-micron (DSM) technology generations, the International Technology Roadmap for Semiconductors (ITRS) [1] has projected an exponential increase in the number of transistors and operating frequency. This commensurate increase could significantly increase dynamic and static power dissipation of future gigascale integration (GSI) systems. Furthermore, a continuous increase in the transistor count and system complexity could increase the number of wires required for communication, which in turn can further increase power dissipation of a multilevel wire network. However, it is a supposition of this work that deliberate coordination between novel interconnect circuit implementation and multilevel interconnect network design can help reduce the power dissipation of future GSI products.

A large number of low power design techniques have been proposed to reduce the dynamic and static power of future GSI systems. To reduce static power, circuit-level techniques such as reducing supply voltage, using multiple threshold voltages, gating of supply voltage or clock, etc. can be used. Similarly, dynamic power reduction by reducing switching capacitance, scaling of supply voltage, reducing the activity factor, etc. has also been proposed.

The insertion of repeaters on interconnects to reduce power has been proposed by [2] and [3]. The use of repeaters enables scaling of wire dimensions and device sizes, which helps to reduce power. The authors in [4] propose to decrease the wire-limited macrocell area by insertion of repeaters in an n-tier multilevel interconnect architecture. Reducing the macrocell area decreases the average wiring capacitance of a system, which reduces the average driver size. This results in a decrease in both dynamic and static power dissipated by a macrocell.

This paper proposes the use of a novel wire sharing technique called wave-pipelined multiplexed (WPM) routing, in combination with repeater insertion, to reduce the total power of large logic macrocells. Like repeater insertion, not only can this technique be applied to both semi-global and global interconnects, but also to shorter interconnects. In addition, it can be seamlessly incorporated into the existing multilevel interconnect network design methodologies. In this paper, circuit design techniques are integrated with system design methodologies to create a realistic and accurate framework to study this low power design technique. Total power (dynamic + static) reduction in a 2-wire circuit and a 40M transistor system using WPM routing is presented. Section 2 explains the application of WPM routing to reduce power in a 2-wire circuit. A case study exhibiting the power savings by application of WPM routing to a 40M transistor system is presented in Section 3.

2. Wave-pipelined multiplexed (WPM) circuit-level optimization

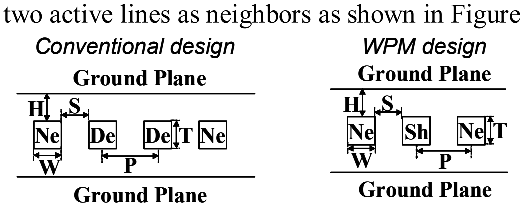
Wave-pipelined multiplexed (WPM) interconnect routing is proposed in [5]. This technique takes advantage of the inherent interconnect idleness and sends multiple signals in a pipelined fashion on a single shared routing channel, in a single clock cycle. This wire sharing technique can be easily applied to both global and semi-global interconnects, and it has been designed such that there is no reduction in the overall throughput performance of the system. The WPM circuit uses 2:1 multiplexer, 1:2 demultiplexer, buffers and some delay circuitry for correct sampling and routing of data. With the use of WPM routing, two dedicated interconnects can be replaced by a single shared interconnect. The primary advantage of this wire sharing technique is the reduction in the number of wires that need to be routed. This reduction in the number of wires reduces the required wire area, resulting in an underutilization of available routing channels. A direct result of this is the reduction in the number of metal levels that are required for routing the wires. It is shown in [5] that the application of the WPM routing technique to a 40M transistor logic core reduces the number of metal levels by almost 20% without any loss of throughput performance. However, the use of WPM to reduce the number of metal levels can result in an

This material is based upon work supported by the National Science Foundation under Grant No. 0092450.

increase in power dissipation and transistor area due to the overhead circuitry required to implement WPM routing.

In this paper, it is shown that this reduction in required routing track count can be harnessed to reduce overall power dissipation. Instead of reducing metal levels, the spacing between the interconnects can be increased so as to fill up all available routing area. This increase in wire spacing will decrease the coupling capacitance between the neighboring interconnects. As a result, smaller sized drivers and receivers can be used, resulting in a decrease in the total device capacitance. Hence, there is an opportunity to reduce both dynamic and static power dissipation with this technique.

To illustrate this concept, consider two dedicated wires, each of length 1.0 cm and designed to operate at 1.3 Ghz. It is assumed that these two wires have active lines as their neighbors as shown in Figure 1. The dimensions of these two wires are designed such that they require 70% of the clock period for data transmission. It is assumed that a buffer of 20% of the clock period is necessary to account for clock skew and signal guardbands. Even though each wire remains idle for only 10% of the clock period, this time is enough to schedule a second signal in a pipelined fashion, using the WPM technique [5], on either wire without any loss of throughput performance. Hence, we use WPM routing and replace these two interconnects by a single shared interconnect, which uses the overhead circuitry described in [5]. The WPM design will have a single shared interconnect with two active lines as neighbors as shown in Figure 1.



W = Wire width | De = Dedicated wire
S = Wire spacing | Sh = Shared wire using WPM
T = Wire thickness | Ne = Active neighboring wire
H = Dielectric thickness | Spwire = S/W
P = Wire pitch | Hoxide = H/W

Figure 1. Cross-sectional view of the wiring network.

2.1 Minimum wire area optimization

This optimization represents the more classic application of WPM to reduce wire area only and will represent a baseline to compare to other optimizations in this section. Figure 2 shows the variation in the wire area with the number of repeaters for the conventional design with two dedicated interconnects and WPM design with a single shared interconnect. One can observe from Figure 2 that a simple application of WPM (no change in wire spacing or dielectric thickness) decreases the total wire area by 50%. This reduction in wire count decreases repeater count, and therefore decreases active transistor area. Figure 3 shows the variation in the transistor area for different number of repeaters. Even with WPM overhead, one can get more than 15% reduction in transistor area at the optimal design point. An increase in static and dynamic power of the

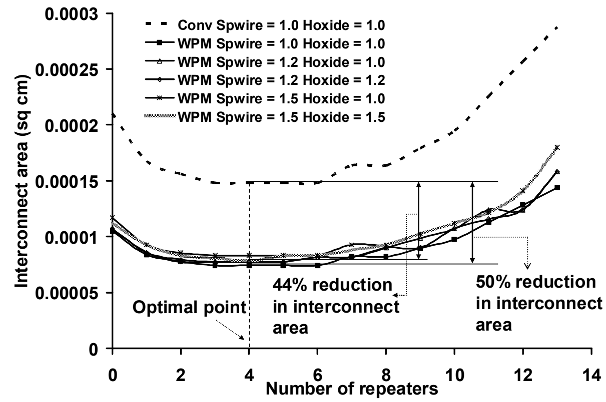


Figure 2. Interconnect area vs Number of repeaters.

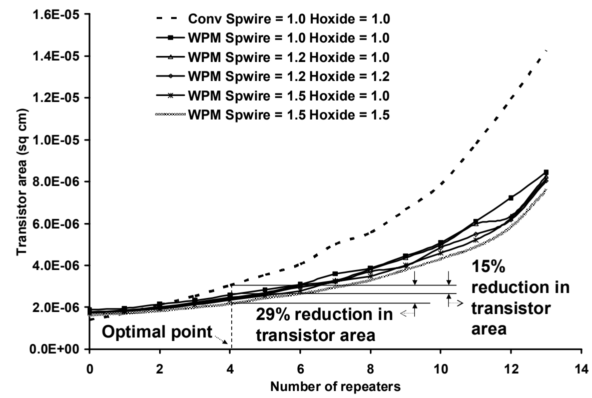


Figure 3. Transistor area vs Number of repeaters.

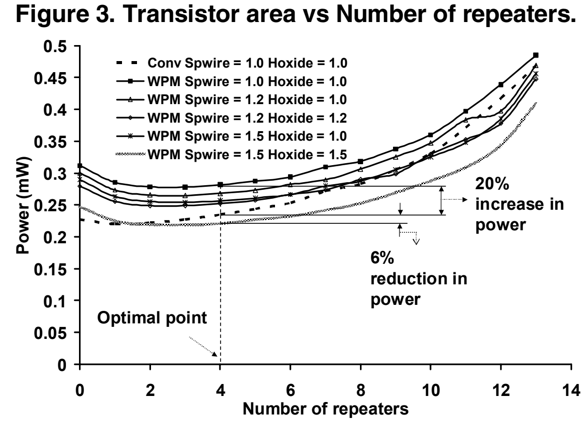


Figure 4. Total power vs Number of repeaters.

system is observed after application of WPM routing. The dynamic power increases due to the increase in the activity factor of the shared resources. In addition, the switching capacitance of the WPM overhead circuitry also contributes to the power equation. As a result, there is an increase in the total power dissipated by the WPM design. At the minimum wire area design point, close to 20% increase in the total power is observed as shown in Figure 4.

2.2 Low power optimization

The elimination of interconnects resulting from WPM increases available routing area and provides an opportunity to increase wire spacing, which can result in lower wire capacitance and smaller driver sizes. Furthermore, because of crosstalk constraints, increasing wire spacing can enable an increase in dielectric thickness, which reduces the effective ground capacitance. The wire spacing and dielectric thickness are increased in same proportions so that the crosstalk constraints and processing constraints are not violated. Figures 2, 3 and 4 also show the variation in interconnect area, transistor area and power with the number of repeaters for various wire spacings and dielectric thicknesses. The increase in wire spacing and dielectric thickness decreases wire capacitance, which decreases wire delay to less than 70% of the clock period. Hence, the pitch and width of the interconnect can be proportionately decreased so that the delay will be equal to 70% of the clock period (as in the conventional design). This enables the use of smaller sized drivers/receivers. The resulting decrease in interconnect capacitance and device capacitance decreases the total power dissipated by the system creating an opportunity for a low power design. For Spwire = 1.5 and Hoxide = 1.5, a 6% reduction in power can be observed at the optimal point in Figure 4. The decrease in wire count and driver/receiver sizes decreases the wire area and transistor area, respectively, of the system. A 44% decrease in wire area and 29% decrease in transistor area can be observed at the optimal point.

2.3 Minimum power design

It is possible to further increase wire spacing so that the wire area of the WPM circuit is the *same* as that of the conventional circuit. This would represent a minimum power design for the 2-wire circuit. In this case study, for Spwire = 4.9 and Hoxide = 2.0 the total wire area of WPM circuit is equal to the wire area of the conventional circuit. Here, it is assumed that Hoxide can have a maximum value of 2.0 due to manufacturing constraints. The WPM circuit dissipates 26% less power than the conventional circuit at the optimal point. In addition, more than 36% reduction in the total transistor area is observed.

3. Wave-pipelined multiplexed (WPM) system-level optimization

Consider a digital logic core consisting of 40M CMOS transistors and designed using 0.1μ technology parameters with a 3-input (six transistors) NAND gate chosen to represent the average standard gate. Copper and low-k (ε_r = 2.0) dielectric material are used to design the multilevel interconnect stack. A suboptimal number of repeaters are inserted as [4] shows that inserting 50% of Bakoglu's optimal number of repeaters [6] results in only 10% performance penalty. Repeater sizing is assumed to be suboptimal because Bakoglu's expression [6] for optimal sizing of repeaters overestimates the required transistor size [7]. Repeaters are inserted beginning from the topmost tier and are successively inserted on the lower tiers based on the availability of silicon area. The macrocell area for the system is optimized so that an even number of metal levels are used for routing interconnects and the transistor area is less than 60% of the macrocell area.

The logic core is simulated using an enhanced version of a multilevel interconnect network design simulator (MINDS) [4] that uses stochastic wiring distribution [8] to estimate wire lengths. This new simulator uses HSPICE and RAPHAEL to model wire transients instead of compact models and is referred to as HR-MINDS. For the case study, the core is assumed to operate at 1.3 Ghz, the number of metal levels is fixed at 8 and a wiring efficiency of 40% is assumed. A macrocell area of 1.125 cm^2 is required to completely utilize available routing tracks in 8 metal levels for a conventional design. Spwire = 1.0 and Hoxide = 1.0 are used across all metal levels for the conventional design.

The application of the WPM routing technique decreases the number of wires that need to be routed. Hence, available routing area in the various tiers increases and provides an opportunity to increase the wire spacing. The dielectric thickness is increased in proportion to the wire spacing to maintain a constant crosstalk ratio. As explained in Section 2, an increase in wire spacing decreases the wire coupling capacitance, which in turn can be used to reduce wire thickness and driver sizes. As discussed previously this results in a decrease in both dynamic and static power dissipation.

In this case study, after application of WPM routing, the entire multilevel interconnect architecture of the logic core is redesigned with increased wire spacing; however, it is constrained by the same macrocell area, operating frequency and metal level count as the conventional design. The overhead resulting from the circuitry required to implement WPM routing is accounted in the system-level simulation while assigning wires to the different tiers. Figure 5 shows the variation of power reduction for the 40M transistor logic core with cutoff lengths. Here, cutoff length is the lower limit of the range of interconnect lengths to which the WPM routing technique is applied [5]. It is assumed that 60% of all wires greater than the cutoff length have been implemented using WPM routing. Table 1 shows the various interconnect parameter values required for designing the conventional system with 40M transistors, and the corresponding WPM system with cutoff length = 0.08215 cm. For this WPM system, it is possible to increase the Spwire to 1.821 in all metal levels by using WPM routing. Hoxide is also increased to 1.821 to maintain conventional crosstalk constraints. In addition to increasing the wire spacing, the wire width is decreased to satisfy the performance requirements. This facilitates the use of smaller drivers and receivers. All this contributes to a reduction in the total power dissipated by the system. It can be seen from Table 1 that there is almost 28% reduction in the total power of the system. Figure 6 shows the power components of a conventional system and the corresponding WPM system with cutoff length of 0.08215 cm. As expected, the dynamic power is greater than the leakage power, which is greater than the short-circuit power. It can be seen in Figure 6 that the application of WPM routing reduces *all three components of power*.

4. Conclusion

It is shown in this paper that a pervasive application of WPM circuits can significantly reduce power if wire dimensions are redesigned to reduce wire capacitance. The application of WPM reduces the number of interconnects that need to be routed. This provides an opportunity to increase wire spacing so as to utilize all the available wiring tracks. This helps decrease wire dimensions and device sizes for a given operating frequency, which in turn reduces the total power dissipated by the system.

For individual 2-wire circuits, a striking example of WPM and wire co-design illustrates that total power can be reduced by up to 26% without any loss in communication performance or area. Moreover, using rigorous system-level simulation the impact of WPM on large logic macrocells is explored. In case of a 40M transistor logic core, the application of WPM routing decreases the total power by almost 28% without any loss of performance or any change in the macrocell area and metal level count. The application of WPM reduces all three components of power: dynamic power by 26.75%, leakage power by 26.96% and short-circuit power by 31.48%.

5. References

[1] International Technology Roadmap for Semiconductors (ITRS) http://public.itrs.net/.

[2] V. Adler et al., "Repeater design to reduce delay and power in resistive interconnect," *IEEE Trans. Circuits Syst. I*, vol. 45, pp. 607–616, May 1998.

[3] A. Nalamalpu et al., "A practical approach to DSM repeater insertion: satisfying delay constraints while minimizing area and power," *Proc. IEEE ASIC/SOC*, pp. 152–156, 2001.

[4] R. Venkatesan, et al., "Optimal *n-tier* multilevel interconnect architectures for gigascale integration (GSI)," *IEEE Trans. VLSI Systems,* Vol. 9, No. 6, pp.899-912, Dec. 2001.

[5] A. Joshi, et al., "Wave-pipelined multiplexed (WPM) routing for gigascale integration (GSI)," *IEEE Tran. VLSI System,* vol. 13, pp. 899-910, Aug. 2005.

[6] H. Bakoglu, et al., "Optimal interconnection circuits for VLSI," *IEEE Trans. Electron Dev*, Vol. ED-32, pp.903-909, May 1985.

[7] Y. Cao, et al., "Effects of global interconnect optimizations on performance estimations of deep submicron design," *Proc. ICCAD 2000*, pp. 56-61, 2000.

[8] J. Davis, et al., "A stochastic wire-length distribution for gigascale integration (GSI)—Parts I and II," *IEEE Trans. Electron Dev.,* vol.45, pp. 580-597, Mar. 1998.

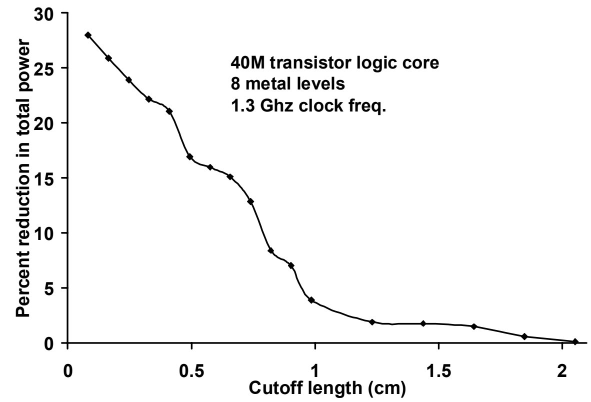


Figure 5. Percent reduction in power by increasing wire spacing and dielectric thickness.

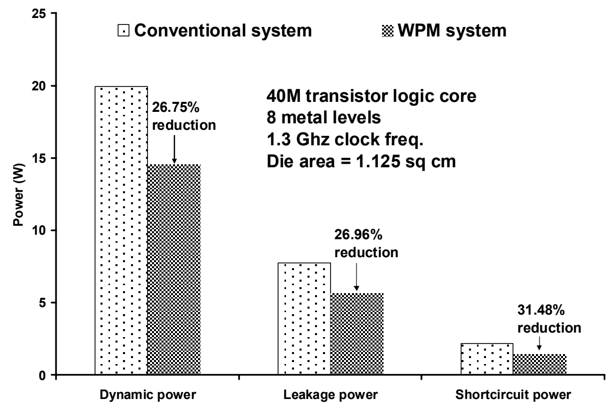


Figure 6. Power in conventional and WPM system.

Table 1. Multilevel interconnect network design parameters for a conventional system and a WPM system.

Tier #	No. of levels/tier	Length of the longest wire on the tier (cm)	Interconnect Width (µm)	Interconnect Spacing (µm)	Number of repeaters	NFET aspect ratio Logic (avg.)	NFET aspect ratio Repeaters	Total power (W)
Conventional system: 40M transistors; Suboptimal number of repeaters, Macrocell area = 1.125 sq cm; Operating frequency = 1.3 Ghz; Number of metal levels = 8.								
4	2	2.1213	0.718	0.718	34844	12.59	120.60	29.84
3	2	0.7125	0.235	0.235	297848		41.73	
2	2	0.2840	0.138	0.138	0		0	
1	2	0.0587	0.100	0.100	0		0	
WPM system: 40M transistors; Suboptimal number of repeaters, Macrocell area = 1.125 sq cm; Operating frequency = 1.3 Ghz; Number of metal levels = 8; Cutoff length = 0.08215 cm.								
4	2	2.1213	0.564	1.03	30749	8.72	83.67	21.49
3	2	0.6487	0.177	0.322	311069		26.30	
2	2	0.1975	0.100	0.182	0		0	
1	2	0.0253	0.100	0.182	0		0	