The Impact of Wave Pipelining on Future Interconnect Technologies

Jeff Davis, Vinita Deodhar, and Ajay Joshi
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250

ABSTRACT

In the era of gigascale integration, both interconnect technologist and interconnect circuit designers must work together closely to ensure that the integrated circuit (IC) industry will overcome current and future interconnect limits on system performance, power dissipation, noise, and cost. This paper will review wave-pipelined interconnect circuits that are used to enhance wire performance and density. The impact of wave-pipelined interconnect circuits on interconnect material integration decisions over the next 10-12 years are explored.

INTRODUCTION

A survey of natural and man-made highly connected complex systems quickly reveals that current integrated circuit (IC) technology is one of the more significant engineering feats of the last half-century. The physical implementation of microscopic IC wiring networks is unique in comparison to current macroscopic communication networks (e.g. telephony, internet, PWB, etc.) in two very striking ways. First, IC wire networks have an enormously large number of wires per unit area, which is growing exponentially and commensurate with Moore's Law. Furthermore, unlike macroscopic wire networks, on-chip IC wire networks do not have traditional ground planes or co-axial structures that are used to enhance signal integrity and performance. These key differences have created both a myriad of opportunities and problems in VLSI designs. As we continue to increase computing complexity, solutions to the problems caused by IC wiring must come from a variety of research directions. Both interconnect technologist and circuit designers must work together closely to ensure that VLSI interconnect limits will not derail the current evolution of digital computation.

Traditional motivations for interconnect technology innovations for on-chip interconnects are driven by a desire to reduce wire latency (i.e. delay) and to increase wire reliability; however, developing interconnect technologies that enable significant and robust wire sharing must also influence technology decisions. It is argued in this paper that extensive use of wave pipelining for global across-chip wires can be used to create robust interconnect circuits that can increase performance even without further development of new inter-level dielectric (ILD) materials. However, shorter semi-global interconnects, such as a bypass bus that is needed to quickly resolve data hazards in a superscalar architecture, must continue to maintain the lowest possible latency. Even with the interconnect technology changes prescribed by the International Technology Roadmap for Semiconductors (ITRS) [1], it is difficult to continue aggressive latency scaling with conventional interconnect circuits that have both their cross-sectional wire dimensions and axial length reduced every technology generation. However, it is argued in this paper that an alternate wave-pipelined multiplexed (WPM) scaling technique [2] could be used to achieve aggressive latency constraints on semi-global interconnects while maintaining continued increases in effective wire density.

INTERCONNECT WAVE PIPELINING

Wave pipelining was originally proposed for logic design [3,4] as an alternative to pipeline register insertion. Instead of pipeline registers, wave-pipeline designs balance path delays to ensure that data signals travel as coherent "waves" propagating simultaneously through a combinational circuit. A key challenge with this technique in logic design is the large variability and data-dependent wave front velocities. Using wave pipelining in VLSI wires is potentially more viable because signal paths are unique and data-dependent wave velocities can be controlled through coplanar shielding, careful routing, or staggered repeater insertion [5]. Unlike lossless or low-loss transmission lines, however, most global and semi-global wires do not naturally support wave propagation because they are essentially distributed RC lines. To produce a type of wave propagation on VLSI wires, one must periodically insert inverters along the length of the line. These inverters enable effective wave propagation in the same way that waves of data propagate through combinational circuits in logic wave pipelining. Hence, this high-speed serialization technique for VLSI wires is referred to as interconnect wave pipelining [5,6].

Global Wire Pipelining

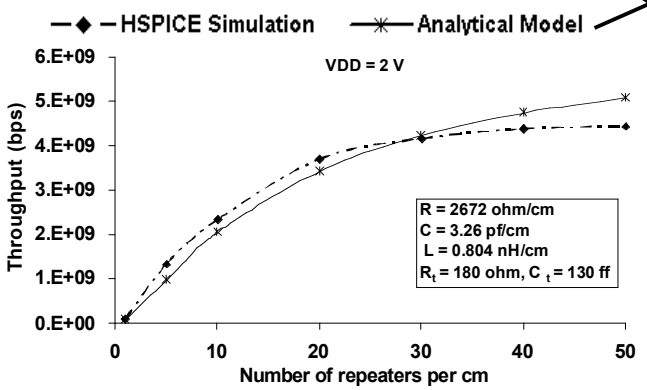
To increase global wire bit rate, which is also referred to as wire throughput, a conventional design strategy would involve the insertion of latches and repeaters periodically along a global wire. This strategy results in an improvement in the global wire throughput by breaking a longer interconnect into smaller segments. The interconnect throughput is limited by the delay of the individual wire segments and the setup and hold times of the inserted latches. However, even though the bit rate can improve significantly with this type of strategy, the overall latency can significantly increase to several clock periods. In addition, the extra latch overhead can consume significant silicon area and power dissipation. Table I shows the attributes of a 1 cm length interconnect that is designed using traditional latch insertion to have a data rate of 3Gbps, which illustrates the limits of 180 nm technology used in this example. In contrast, we can also use wave pipelining on these interconnects to produce high-speed serialization of the data. Table I also shows the design attributes of the same 1 cm length interconnect using wave pipelining. Both the wave-pipelined interconnect and the latch inserted interconnect are simulated using level-49 HSPICE transistor models [7]. It is assumed that the wire dimensions for this example are consistent with metal 5 dimensions in [8] (i.e. W = S = 800 nm, H_ρ /W = 2).

A comparison of traditional pipeline approaches to interconnect wave-pipelined approaches is shown in Table I. This comparison reveals the powerful allure of a wave-pipelined design style because it outperforms latch insertion in several design metrics. Perhaps one of the most significant advantages is that wave pipelining can achieve 3Gbps with the latency of each signal being approximately a third of the latency of the traditional pipeline structure. This fact illustrates that wave pipelining can provide a significant amount of data movement while still maintaining relatively low transmission latency, which is important for architectural performance metrics. In addition, the total power, which includes dynamic, short circuit, and leakage power, is almost cut in half. For leakage concerns it is worth noting that the silicon area for the wave-pipelined interconnect circuit is a fifth of the area of the traditional pipeline structure.

To gain physical insight and to calculate the trends in interconnect wave pipelining, a closed-form analytical expression for the maximum communication throughput of a global wire

Table I. Design of a 1cm interconnect circuit implemented in 180 nm technology

	Latch Insertion	Wave pipelining
Target Throughput	3 Gbps	
Number of latches/repeaters	20	18
Latency	3.33 ns	0.97 ns
Total Power	6.23 mW	3.77 mW
Wire Pitch	1.6 micron	1.6 micron
Silicon Area	5.07e-5 cm^2	1.05e-5 cm^2



Maximum throughput [6]	$T_{max} = \frac{1}{\sigma_{RCseg} \ln\left(\frac{K_1}{1-v_1}\right) + 0.693 R_t C_t}$
Sakurai time constant [12]	$\sigma_{RCseg} = \sigma_1 = R_t C_t + R_t C_{seg} + C_t R_{seg} + 0.4 R_{seg} C_{seg}$
Sakurai coefficient [12]	$K_1 = 1.01\left[\frac{R_t C_{seg} + R_{seg} C_t + R_{seg} C_{seg}}{R_t C_{seg} + R_{seg} C_t + \frac{\pi}{4} R_{seg} C_{seg}}\right]$
Voltage swing v_1 calculated recursively [6]	$v_n = 0.9$ $v_{n-1} = 1/(2 - v_n)$

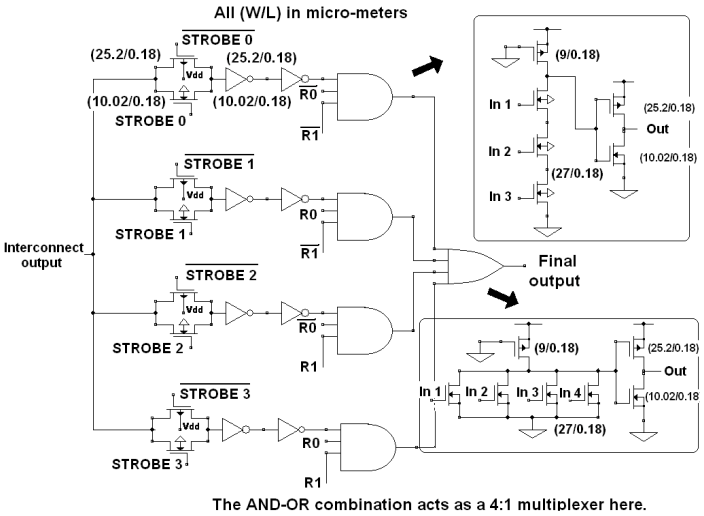
R_t =Transistor equivalent output resistance
C_t = Transistor equivalent input capacitance
R_{seg} = Wire segment resistance
C_{seg} = Wire segment capacitance
n = number wire segments

Figure 1. Comparison between results of closed-form analytical expression and HSPICE simulations.

with repeaters can be used and appears in Figure 1 [6]. The values of throughput using this analytical model also are compared with simulation using from HSPICE level-49 models [7]. This 250nm x 250nm interconnect is modeled in HSPICE using a distributed resistance-inductance-capacitance (RLC) network, and the RLC values are extracted using RAPHEAL.

A global VLSI interconnect is defined as an interconnect that sends information across a microprocessor or ASIC die. For the discussions in this paper, it is assumed that Moore's Law will be maintained through the development of monolithic multiprocessor cores with large on-chip caches. Under this paradigm, global interconnects would then send information from core to core and from core to memory. The goal of these global wire networks are to send as much data as possible every clock period without significant latency penalties. However, global interconnect communication will be plagued by synchronization issues cause by manufacturing variations, power supply variations, temperature variations, global clock skew and jitter [1]. To synchronize wave-pipelined data, a dedicated clock line must accompany a group of signal lines to be used to extract data using a FIFO re-timer. A FIFO re-timer circuit, which appears in Figure 2(a), can be used to synchronize two cores that have the same nominal clock frequency with completely random phase difference[11]. This FIFO re-timer does increase the communication latency, but it allows for extremely robust global communication. Figure 2(b) shows the order for which each signal is written to and read from the FIFO re-timer.

(a)



(b)

S_1	S_0	STROBE0	STROBE1	STROBE2	STROBE3	Data written to	R_1	R_0	Data read from
0	0	1	0	0	0	M0	1	1	M3 (x)
0	1	0	1	0	0	M1	0	0	M0
1	0	0	0	1	0	M2	0	1	M1
1	1	0	0	0	1	M3	1	0	M2

Writing of data with CLK_S	Reading of data with CLK_R		
	ΔPhase = 0°	ΔPhase =180°	ΔPhase = 360°
M0	M3	X	X
M1	M0	M3	X
M2	M1	M0	M3
M3	M2	M1	M0

Figure 2. FIFO re-timer circuit in (a) is used to synchronize data between two cores that have the same frequency but random phase difference, and (b) shows the order of the signal written to and read by FIFO re-timer.

Wave-Pipelined Multiplexed (WPM) Routing

Another circuit type that utilizes the wave pipelining ideology along with simple transmitter and receiver circuits is referred to as wave-pipeline multiplex (WPM) routing [2]. This technique has the advantage of being able to significantly reduce wire area while still maintaining stringent latency constraints. The timing for WPM is unique in that two bits are sent in rapid succession every clock period in a wave-pipelined fashion; thus allowing two neighboring lines to be combined into one wire channel. The receiver and transmitter circuits contain a multiplexer, demultiplexer, and dynamic latches and are illustrated in Figure 3. HSPICE simulations have been used to verify this circuit and to confirm that the delay of multiplexer and demultiplexer is small in comparison to the wire delay for semi-global and global wires.

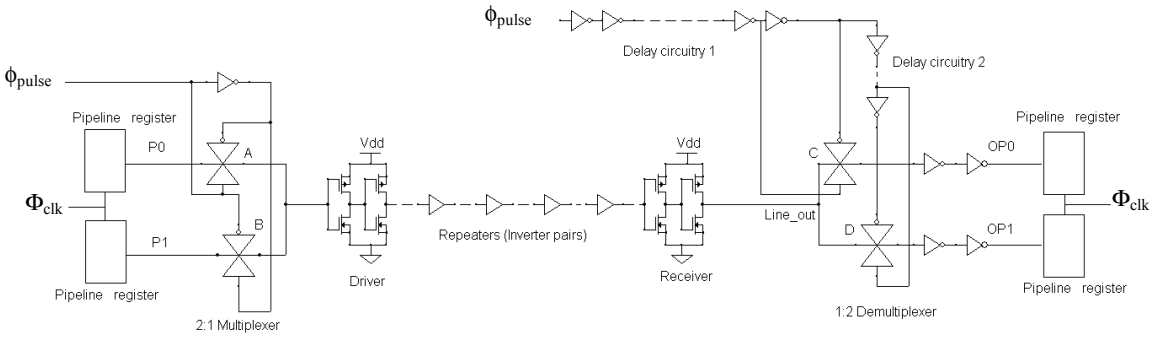


Figure 3. Circuit schematic diagram for WPM routing

Having access to a highly synchronized local clock is key to generate the necessary pulse clock, ϕ_{pulse}, seen in Figure 3 to multiplex and de-multiplex each line successfully. It is assumed that such synchronization is possible for semi-global interconnects where intra-core clock synchronization is more precise. Unlike wave pipelining for global interconnects, a FIFO re-

timer cannot be used for WPM routing because of the overhead and latency constraints. The design of WPM routing circuits can be challenging to produce a robust interconnect circuit; however, as illustrated in the next section, the possible benefits could justify this increase design effort.

WAVE-PIPELINED TECHNOLOGY PROJECTIONS

Historically, the clock frequency for microprocessors has approximately doubled every technology generation [1]. However, due to the power dissipation limitations and other design limitations described in [9], it is assumed in this paper that future clock frequency advances will come at a slower pace than projected by the ITRS. In this section it is assumed that the clock frequency will increase 1.4x every technology generation and is due primarily to transistor performance enhancements from constant field scaling. Both global and semi-global interconnect scaling scenarios that attempt to keep pace with these projected clock frequencies are presented in this section.

Global Wave-Pipeline Scaling

As mentioned, global interconnect design is most concerned with moving large amounts of data rapidly across the die. Significant increases in wire bit rate can be achieved by using wave pipelining even without any further advances in ILD materials. Figure 4(a) shows that if the dielectric constant remain fixed at k=3.1 beyond 90nm, then the number of repeaters in a 1cm length interconnect, for example, can be increased and wave pipelined signaling can be used to produce bit rates that keep pace with projected clock frequencies. In addition, Figure 4(b) illustrates the impact these design will have on overall signal latency. For this example, the latency would increase to roughly 2.5x the clock periods at the 22nm technology node. FIFO re-timer circuits could be used to synchronize this data across chip; however, this would add at least two more clock delays to the values in Figure 4(b). Interconnect dimensions are assumed to be large global dimensions on the order of approximately 1000nm across all technology generations.

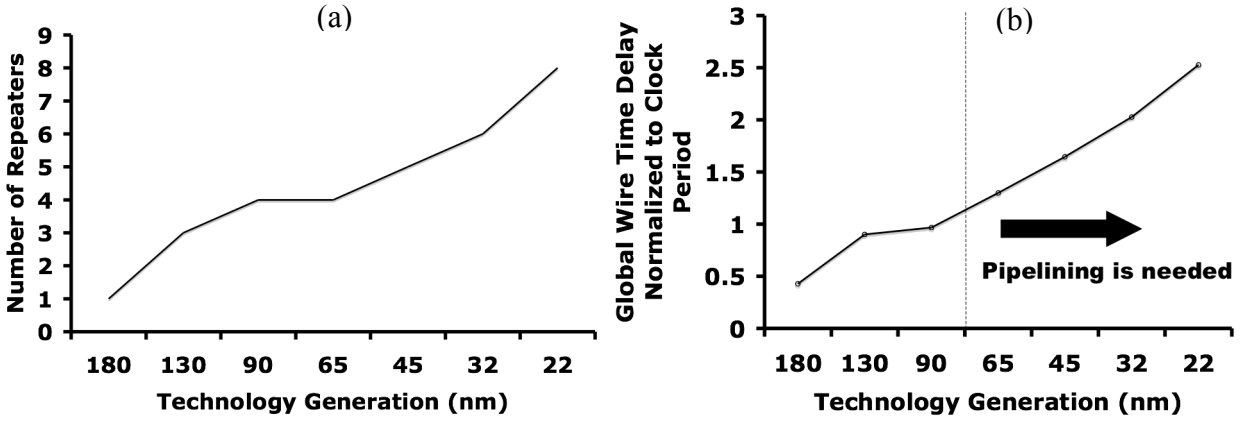


Figure 4. Global interconnect design to meet target bit rates with a constant ILD dielectric constant (k=3.1) appear in (a), and global interconnect latency for each wave-pipelined circuit appears in (b).

Semi-Global WPM Wire Scaling

In high performance microprocessor design there will always be a set of interconnects that will need to communicate information with extremely short latency to enhance architectural performance. A semi-global bypass bus is a prime example of this type of wire and is needed to forward register values to resolve data hazards in a computational pipeline [10]. Therefore, it is assumed that future scaling of semi-global interconnects must consider both latency and wire density. Ideal scaling of semi-global wires needs to decrease wire area by a factor of one-half every technology generation so that the number of semi-global wire levels does not increase. This can be achieved by scaling both the interconnect length and pitch by 0.7x every technology generation. In addition, to maintain the projected clock frequencies these critical semi-global interconnects must have a 0.7x reduction in wire delay every technology generation as well. Meeting future latency targets are especially difficult because of the wire sizing effects on resistivity of intermediate length wires [1].

WPM scaling in this paper refers to the strategy of increasing the wire cross-sectional dimensions by a factor of 2x over a non-WPM wire channel. This increase in wire area is completely offset by then sending two bits per channel using WPM. Increasing the interconnect cross-sectional dimensions helps reduce the intrinsic latency so that a second bit can be sent along the same line such that the effective 2-bit latency is in some cases is less than the conventional 1-bit latency scheme. Because there is no increase in effective wire area, WPM scaling could be an effective method to achieving both low latency and low wire area even without significant changes in the ILD materials.

Figure 5(a) shows how a bypass bus length might change with ideal scaling over the next 10-12 years. This bus length at 180 nm technology is consistent with a critical bypass bus in the Itanium-2 processor and has approximately a 2 mm length [10]. Figure 5(b) illustrates that even with ITRS material changes and aggressive use of repeaters, the scaled latency constraints for this bypass bus beyond 65nm cannot be met. However, Figure 6(a) illustrates that WPM wire scaling can significantly reduce wire latency along the roadmap even without changing the dielectric constant (k=3.1). Figure 6(b) illustrates that WPM scaling with ITRS material changes can meet all latency constraints until the 22nm technology node.

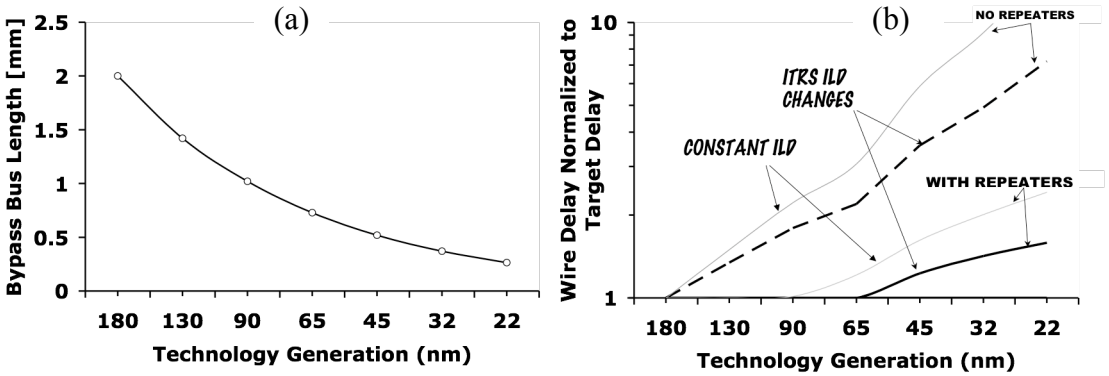


Figure 5. Bypass bus length scaling scenario appears in (a), and the corresponding wire delays assuming constant ILD and ITRS prescribed changes to ILD appears in (b).

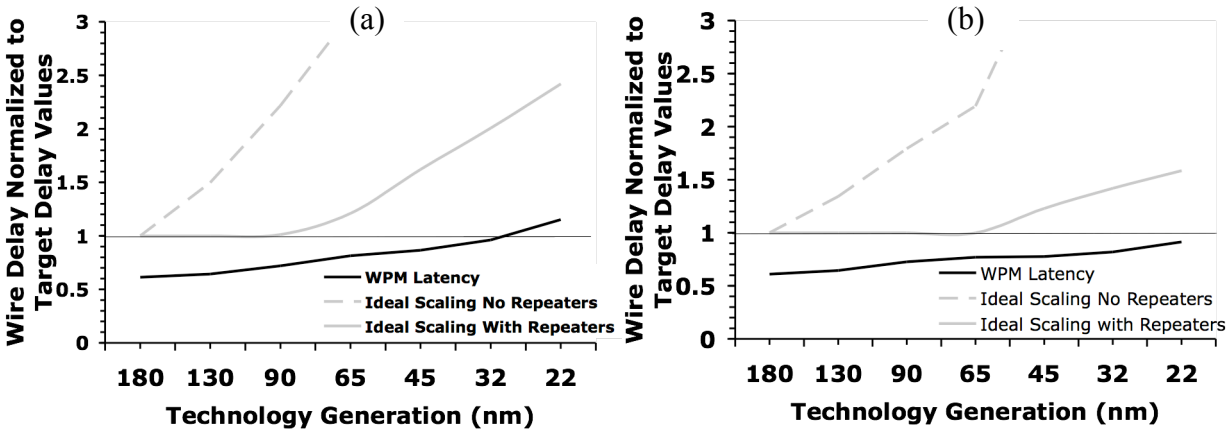


Figure 6. Bypass bus wire delays using WPM wire scaling with no changes in dielectric constant beyond 90nm (i.e. k = 3.1) in (a) and with ITRS dielectric projections in (b).

CONCLUSIONS

Advances in dielectric materials always reduce the latency of global wires, but it is shown in this paper that wave-pipelined circuits can be used to meet the bit-rate targets for future global wires without changes in dielectric materials. In addition, for semi-global wires, wave-pipelined multiplexed (WPM) routing could be used to create interconnect circuits that have both low latency and low wire area. It is illustrated that WPM wire scaling could be used to meet aggressive delay requirements on a bypass bus for the next three technology generations without further advances in ILD materials. Even though WPM scaling could provide significant benefits, this design technique comes at a cost of careful custom circuit design to ensure robust operation.

ACKNOWLEGEMENTS

The authors would like to thank the National Science Foundation for the support of this research (NSF # 0092450).

REFERENCES

[1] *2004 International Technology Roadmap for Semiconductors (ITRS)* (http://public.itrs.net)
[2] A. Joshi and J. Davis, *IEEE Trans. VLSI Systems*, 13 (2005) (in press).
[3] L. Cotton, Proc. *Proc. of AFIPS Spring Joint Computer Conf.,* 881 (1969).
[4] C. Grey, et al, *Wave Pipelining: Theory and CMOS Implementation*, (Kluwer Academic Publishers, 1994).
[5] J. Xu and W. Wolf, *Proceedings of 11th Symposium on HOTI*, 10-14 (2003).
[6] V. Deodhar and J.A. Davis, *IEEE Trans. VLSI Systems*, 13, 308-318, (2005).
[7] MOSIS Service (http://www.mosis.org) for HSPICE models
[8] S. Yang, et. al, *Proc. of IEDM*, 197-200 (1998).
[9] V. De and S. Borkar, *Proc. of ISLPED*, 163-168 (1999).
[10] E. Fetzer, et al, *IEEE J. of Solid-State Circuits*, 37, 1433-1440 (2002).
[11] V. Deodhar and J.A. Davis, *Proc. of ISQED*, 592-597 (2005).
[12] T. Sakurai, IEEE Trans. on Electron Devices, 40(1), 118-124 (1993).