A Survey Addressing on High Performance On-Chip VLSI Interconnect

C. Mohamed Yousuff, V. Mohamed Yousuf Hasan, and M. R. Khan Galib

Abstract—With the rapid increase in transmission speeds of communication systems, the demand for very high-speed low-power VLSI circuits is on the rise. Although the performance of CMOS technologies improves notably with scaling, conventional CMOS circuits cannot simultaneously satisfy the speed and power requirements of these applications. In this paper we survey the state of the art of on-chip interconnect techniques for improving performance, power and delay optimization and also comparative analysis of various techniques for high speed design have been discussed.

Keywords—current-mode signaling, hybrid current/voltage mode circuits, on-chip interconnects delay and power, voltage mode signaling

I. INTRODUCTION

In VLSI technology, interconnect delay reduction is of important concern. As anticipated by Moore's law, the number of transistors in an integrated circuit (IC) has doubled every two to three years. The device size and the switching delay have shrank continuously. Figure 1 shows the highest frequency achieved by micro-processors for each technology node. At present, 22 nm technologies is in production and microprocessor clock frequencies are above a GHz. The speed of an electrical signal in an IC is governed by two components. The first component is the switching time of an individual transistor, known as transistor gate delay, and the second one is the signal propagation time between transistors, known as wire delay or interconnect delay. Since many generations ago, the device switching speed no longer limits the circuit performance, but the wire delay becomes dominant. There are different techniques have been proposed by different authors for efficient low power-high speed CMOS circuits and other parameters along the interconnects. An extensive research survey has been presented in this paper illustrating pros and cons of the various research articles and a method to overcome that.

The paper is arranged as follows. Section I introduces the topic. Section II deals with voltage mode signaling (VMS) and comparative analysis of various VMS techniques is discussed. Section III describes the current mode (CM) interconnects. In Section IV describes the most efficient way of reducing delay by combining the VM and CM signalings. Finally, conclusions are drawn in Section V.

Fig. 1. VLSI technology scaling (Source: ITRS Roadmap, Intel Corporation).

A. Voltage Mode Interconnects

In voltage mode, the receiver presents the interconnect with a high impedance to changes over a full voltage swing and the sensing circuit at the destination determines the signal state by using this capacitive termination ($R_L \approx \infty$). The speed on the interconnection value. A simple example of Voltage mode signaling is presented in Fig. 2 where inverter drives an inter-connect, charging the wire capacitance which builds up a voltage along the line. Another inverter senses the voltage and provides a high impedance termination.

In voltage mode logic (VML) circuits value of the input voltage(s) to switch the different gate transistors to either the ON or the OFF state, creating a path from the output node to one of the supply rails. This is different from the reduced voltage swing circuits, where some transistors are completely ON while the other transistors are partially ON. Such transistors do not work properly in VML. Current mode logic, where all the transistors are ON (partially or fully), is preferred for such cases.

Fig. 2. Voltage mode signaling on interconnect.
Using Copper and Low-k Dielectric Materials

C. Ryu [1] proposed a method to overcome the bottleneck in performance due to on-chip wires, by introducing new interconnect technologies such as copper and low-k dielectric materials to reduce wire resistance and capacitance. The resistivity of copper is 40% lower than that of aluminum. Since the dielectric constant of SiO₂ is 3.9 and new materials, such as Fluorinated Silica Glass (FSG) and Black Diamond, have dielectric constants approaching 3.0. Other materials with even lower dielectric constants are being studied extensively. These reduced resistivity and low dielectric constant materials reduce the wire delay, capacitive crosstalk noise, and switching power. Figure 3 shows the interconnect lengths at which the wire delay is equal to the gate delay for local, intermediate and global interconnects as predicted in the ITRS roadmap [2].

This figure indicates that, as technology scales, the wire length at which the wire delay starts to dominate becomes less. Due to the growing importance of wire delay, circuit designers have been putting increasing efforts on wire design and analysis. Sometimes on-chip inductive effects may be more significant with the technology scaling and increase of clock frequencies. Some researchers have found that inductive impedance of the on-chip wires become comparable to or larger than the gate delay. For instance, in a 0.25 μm technology the gate delay is about 25 psec, however, a 5 mm long aluminum wire with 0.25 μm × 0.25 μm cross section has a delay of 1 nsec. It is possible to reduce the propagation delay by introducing intermediate buffers, generally known as repeaters shown in Fig. 4 in the interconnect line. Repeater insertion is a classical solution that changes the delay dependence on the wire length from quadratic to linear. A delay analysis method has been presented by Bakoglu [3], wherein the repeater is modeled by discrete resistance and capacitance while the interconnect is modeled as distributed RC elements. Several other repeater insertion methods have been proposed [4]–[8]. In all of the above repeater insertion methods, the repeater is modeled by discrete resistance and capacitance elements.

The delay of a line is

$$\tau_L = \frac{3.56 K_{ox} \varepsilon_0 \rho L^2}{\lambda^2} \quad (1)$$

By breaking a long interconnect line into n smaller lines the propagation delay of each line is reduced quadratically

$$\tau_{L/n} = \frac{3.56 K_{ox} \varepsilon_0 \rho (L/n)^2}{\lambda^2} \quad (2)$$

The total wire delay is thus

$$(\tau_{L/n} + \tau_G) = \frac{3.56 K_{ox} \varepsilon_0 \rho (L/n)^2}{\lambda^2} + n\tau_G \quad (3)$$

where

- $\tau_G$ – gate delay,
- $\lambda$ – feature size,
- $L$ – Line length,
- $\tau_L$ – Line delay,
- $K_{ox}$ – Dielectric constant,
- $\varepsilon_0$ – permittivity of free space,
- $\rho$ – interconnect resistivity,
- $n$ – no. of smaller lines.

As long as the gate delay is small the total wire delay is reduced substantially. This gain results at the cost of increased chip area occupied and extra power consumed by the repeaters.

Using Repeaters for Delay Reduction

A first option for reducing RC delays is to use better interconnect materials when they are available and appropriate. However, for very long wires the delay can be substantially larger than the gate delay. For instance, in a 0.25 μm technology the gate delay is about 25 psec, however, a 5 mm long aluminum wire with 0.25 μm × 0.25 μm cross section has a delay of 1 nsec. It is possible to reduce the propagation delay by introducing intermediate buffers, generally known as repeaters shown in Fig. 4 in the interconnect line. Repeater insertion is a classical solution that changes the delay dependence on the wire length from quadratic to linear. A delay analysis method has been presented by Bakoglu [3], wherein the repeater is modeled by discrete resistance and capacitance while the interconnect is modeled as distributed RC elements. Several other repeater insertion methods have been proposed [4]–[8]. In all of the above repeater insertion methods, the repeater is modeled by discrete resistance and capacitance elements.

The delay of a line is

$$\tau_L = \frac{3.56 K_{ox} \varepsilon_0 \rho L^2}{\lambda^2} \quad (1)$$

By breaking a long interconnect line into n smaller lines the propagation delay of each line is reduced quadratically

$$\tau_{L/n} = \frac{3.56 K_{ox} \varepsilon_0 \rho (L/n)^2}{\lambda^2} \quad (2)$$

The total wire delay is thus

$$(\tau_{L/n} + \tau_G) = \frac{3.56 K_{ox} \varepsilon_0 \rho (L/n)^2}{\lambda^2} + n\tau_G \quad (3)$$

where

- $\tau_G$ – gate delay,
- $\lambda$ – feature size,
- $L$ – Line length,
- $\tau_L$ – Line delay,
- $K_{ox}$ – Dielectric constant,
- $\varepsilon_0$ – permittivity of free space,
- $\rho$ – interconnect resistivity,
- $n$ – no. of smaller lines.

As long as the gate delay is small the total wire delay is reduced substantially. This gain results at the cost of increased chip area occupied and extra power consumed by the repeaters.

Using Repeaters for Delay Reduction

A first option for reducing RC delays is to use better interconnect materials when they are available and appropriate. However, for very long wires the delay can be substantially larger than the gate delay. For instance, in a 0.25 μm technology the gate delay is about 25 psec, however, a 5 mm long aluminum wire with 0.25 μm × 0.25 μm cross section has a delay of 1 nsec. It is possible to reduce the propagation delay by introducing intermediate buffers, generally known as repeaters shown in Fig. 4 in the interconnect line. Repeater insertion is a classical solution that changes the delay dependence on the wire length from quadratic to linear. A delay analysis method has been presented by Bakoglu [3], wherein the repeater is modeled by discrete resistance and capacitance while the interconnect is modeled as distributed RC elements. Several other repeater insertion methods have been proposed [4]–[8]. In all of the above repeater insertion methods, the repeater is modeled by discrete resistance and capacitance elements.

The delay of a line is

$$\tau_L = \frac{3.56 K_{ox} \varepsilon_0 \rho L^2}{\lambda^2} \quad (1)$$

By breaking a long interconnect line into n smaller lines the propagation delay of each line is reduced quadratically

$$\tau_{L/n} = \frac{3.56 K_{ox} \varepsilon_0 \rho (L/n)^2}{\lambda^2} \quad (2)$$

The total wire delay is thus

$$(\tau_{L/n} + \tau_G) = \frac{3.56 K_{ox} \varepsilon_0 \rho (L/n)^2}{\lambda^2} + n\tau_G \quad (3)$$

where

- $\tau_G$ – gate delay,
- $\lambda$ – feature size,
- $L$ – Line length,
- $\tau_L$ – Line delay,
- $K_{ox}$ – Dielectric constant,
- $\varepsilon_0$ – permittivity of free space,
- $\rho$ – interconnect resistivity,
- $n$ – no. of smaller lines.

As long as the gate delay is small the total wire delay is reduced substantially. This gain results at the cost of increased chip area occupied and extra power consumed by the repeaters.

Using Repeaters for Delay Reduction

A first option for reducing RC delays is to use better interconnect materials when they are available and appropriate. However, for very long wires the delay can be substantially larger than the gate delay. For instance, in a 0.25 μm technology the gate delay is about 25 psec, however, a 5 mm long aluminum wire with 0.25 μm × 0.25 μm cross section has a delay of 1 nsec. It is possible to reduce the propagation delay by introducing intermediate buffers, generally known as repeaters shown in Fig. 4 in the interconnect line. Repeater insertion is a classical solution that changes the delay dependence on the wire length from quadratic to linear. A delay analysis method has been presented by Bakoglu [3], wherein the repeater is modeled by discrete resistance and capacitance while the interconnect is modeled as distributed RC elements. Several other repeater insertion methods have been proposed [4]–[8]. In all of the above repeater insertion methods, the repeater is modeled by discrete resistance and capacitance elements.

The delay of a line is

$$\tau_L = \frac{3.56 K_{ox} \varepsilon_0 \rho L^2}{\lambda^2} \quad (1)$$

By breaking a long interconnect line into n smaller lines the propagation delay of each line is reduced quadratically

$$\tau_{L/n} = \frac{3.56 K_{ox} \varepsilon_0 \rho (L/n)^2}{\lambda^2} \quad (2)$$

The total wire delay is thus

$$(\tau_{L/n} + \tau_G) = \frac{3.56 K_{ox} \varepsilon_0 \rho (L/n)^2}{\lambda^2} + n\tau_G \quad (3)$$

where

- $\tau_G$ – gate delay,
- $\lambda$ – feature size,
- $L$ – Line length,
- $\tau_L$ – Line delay,
- $K_{ox}$ – Dielectric constant,
- $\varepsilon_0$ – permittivity of free space,
- $\rho$ – interconnect resistivity,
- $n$ – no. of smaller lines.

As long as the gate delay is small the total wire delay is reduced substantially. This gain results at the cost of increased chip area occupied and extra power consumed by the repeaters.
Even though repeater insertion is a very simple and popular method but it suffers from the following disadvantages:

- They are effective for capacitively loaded lines and they add their own switching delay.
- With technology scaling, the number of repeaters on a chip is increasing exponentially and lead to large source of leakage powers in future technologies.
- Repeaters are needed to be placed at regular interval with uniform sized otherwise repeater chain performance degrades.
- Inserting repeaters along the wire makes the wire unidirectional.

Bidirectional buffers consume additional serious signal integrity problem because of electrostatic coupling between long wires. Inter-signal interference can lead to unpredictable delay variations. Grounded shielding wires must often be inserted to avoid interference. This leads to extra capacitance and CV $^{2}$f power loss. In order to include inductive effects for reducing signal propagation delay, Moursy and Friedman [9], [10] introduced exponential wire shaping on RLC lines which includes the inductance effects. But still since the operation is with voltage mode signaling, considerable amount of power dissipated due to full swing voltage. Hence there should be an alternative to compromise the drawbacks that results in voltage mode scheme.

II. CURRENT MODE SIGNALING (CMS) FOR INTERCONNECT DELAY REDUCTION

The limitation with VMS scheme is that, voltage has to swing from rail to rail over the entire length of the wire. This leads to large transient currents consuming more power, larger delay, and it also generates power-supply noise [11]. The routing area (extra devices and control signal) and power resources.

A. Using Shield Insertion for Delay Reduction

However using, repeaters along the line does not show significance performance for crosstalk delay reduction. Author in [12], [13] introduces a method of reducing crosstalk in inductively coupled interconnects using distributed shield insertion, he presented various conditions like prior to shield insertion, after shield insertion and after additional ground tap insertion at shield terminal for reducing crosstalk noise and signal delay uncertainty. Shielding in high speed digital circuits is one of the effective and common ways to reduce crosstalk noise and signal delay uncertainty. Shield is a wire directly connected to Vdd or Gnd. One of the effective methods of shielding is placing ground or power lines at the sides of a victim signal line to reduce noise and delay uncertainty. This can be easily explained with the help of Fig. 5. Inserting a shield line is necessary to retain proper signal integrity. However, shield insertion consumes more power, increases routing area and add to interconnect routing complexity [14].

As interconnect wire separation is reduced, there is an optimal repeater insertion technique [15] used in voltage-mode signaling was developed to reduce the wire delay and improve the performance of global interconnections. However, with the increase in number and density of interconnects with technology scaling; the number of repeaters necessary would increase considerably, presenting significant overhead in terms of power and area. The other type of transmission system is current mode signaling has shown significant improvement over the VMS scheme. The key to current-mode signal transporting is the low impedance termination at the receiver which results in reduced signal swings without the need of separate voltage references and increased bandwidth performance. Also this low-impedance termination shifts the dominant pole of the system and leads to a smaller time constant and thus, to a smaller delay. It can operate at a much lower noise margin than the voltage-mode network, and at a much lower swing as well due to its immunity to power supply noise. All these translate into increased bandwidth performance [16], decreased delay and dynamic power dissipation and higher noise immunity. For these reasons, CMS technique becomes a better alternative than VMS scheme for on temporary and future high-speed noise-prone single chip systems.

B. Use of MOS Current Mode Logic (MCML)

MOS current mode logic (MCML) has emerged as a logic style that can achieve the much needed high speeds while consuming less power than conventional CMOS circuits at these high frequencies [17] and also provides considerable potentials for improving signal integrity in digital logic circuits. Current-sensing or current-mode signaling determines the logic value transmitted on a wire based on the current through the wire wherein 1’s and 0’s be signaled by the presence or absence of a current and not by a high or a low voltage. This is in direct contrast to voltage-mode which defines logic levels as voltages on the nodes. Current-sensing techniques in digital CMOS technology were first proposed for sense amplifiers in memories. Seevinck [18] proposed a sense amplifier for SRAMs and presented an analysis comparing the performance of current-mode sensing with the conventional sensing. Current mode signaling circuits will have low impedance nodes [19], where the resultant output voltage swing is also small. This low impedance transforms them into low time constant circuits and increases the bandwidth. But MOS current mode logic (MCML) is not widely used in digital design because of its static power dissipation and design complexity.
C. Dynamic Current Mode Logic

Allam and Elmasry [20], introduces new logic-style DyCML family combines the advantages of both MCML logic circuits and dynamic logic styles. The current mode signaling scheme is recommended at high operating frequencies. However, consumes more power at low frequencies. Not suitable for power-down modes because of the dc current source. A simple current mode circuits for interconnects have been shown in Fig. 6. It uses constant current source to generate current on the line and also current mode signaling with constant current source suffers static power dissipation. This method can remove this drawback. A major advantage of the DyCML is the dynamic current source, which achieves smaller delays compared to the basic MCML circuits. Other advantages inherited from MCML are high performance, noise immunity, and robustness to supply voltage scaling. DyCML gates reduce power dissipation by reducing the output voltage swing and thus DyCML circuits achieve high speed with low-power dissipation. Also current rise time is limited by inductance rather than capacitance. Inductive effects are much smaller than capacitive effects. Therefore, coupling due to inductance effect is less compared with the coupling due to capacitive effect. Also this small output swing of CML circuits reduces the cross talk between the adjacent signals; this in turn reduces the dynamic power dissipation [20].

D. Static Power Reduction by Bridge Resistor Termination

Static power dissipation can be reduced by different techniques which reduce leakage currents. For the current-sensing CM circuit architecture, a static current path always exists between the driver and receiver stages even if there is no data activity on the interconnect and hence static power dissipation. Differential bus with bridge resistor termination structure proposed in [21] can be used to reduce static power. Also, static power dissipation is mitigated by using transition encoded current sensing and current pulse signaling proposed in [22]. This leads to an extremely low power and high performance circuit solution for on-chip interconnects. Transition encoded current-sensing is 52% faster and 56% lower power than the repeater insertion method. Current-pulse signaling further reduces power consumption and saves 48% of power over transition encoded current-sensing. Current-pulse signaling is however more noise sensitive than transition encoded current-sensing.

E. Power Reduction by Multi Level Current Signaling

Current mode signaling may cause power consumption problem due to a constant current flow. Therefore, in order to save power, three current levels are required, two to differentiate between consecutive same symbols transmissions and third one is, zero current level to indicate the wire is idle were proposed in [23]–[25]. Differential current signaling proposed by Atul Maheshwari, Wayne Burleson [15] occupies more routing area Vs repeater. Single ended current sensing receivers utilize less wire usage and hence routing area shortcomings of differential signaling is eliminated. Author in [22] has proposed phase coding as a multi-bit signaling technique which encodes multiple bits on a single wire in terms of phase information. Transmitting multiple bits saves power, double the bandwidth and delay is comparable to buffer insertion. Multi level current signaling specifically attempts to reduce the number of interconnect wires [26]–[29]. The interconnect is terminated by a low impedance receiver, and the signals sensed by four current mirror comparators. Figure 7 shows the overall block diagram of the multi-level signaling system. The driver encodes the two bits of signals into four current levels and transmits. The currents propagate through the interconnect and are compared at the receiver by reference currents. The receiver converts the four current levels into thermometer codes and the decoder recovers the original signal.

F. Using Pulsed Current Mode Signaling

Author in [30] has proposed pulsed current mode signaling, in which generates return-to-zero (RZ) codes and does not consume static power achieves near speed of light latency and low bit energy through low swing current mode operation. And this signal modulates the transmitter energy to higher frequencies, where the effect of wire inductance can be minimized. Schematics of the on-chip pulsed-current mode transmission line interconnects (PTLI) are shown in Fig. 8.

PTLI [31] consists of pre buffers that generate differential signals, stacked-switch Txs, an on-chip differential transmission line (DTL), and an Rx. Rail to rail signals are input into the PTLI, and Txs convert rail-to-rail signals into pulse-shaped differential RZ-signals. RZ signals propagate in the DTL at the speed of electromagnetic waves. Vcom stabilizes common-mode voltages of the Tx output. Rx amplifies the pulse signals and converts the RZ signals into NRZ (non-return-to-zero) signals. This work can be extended for better performance if differential bipolar current mode links are employed.
III. HYBRID CURRENT/VOLTAGE MODE CIRCUITS FOR DELAY AND POWER REDUCTION

An adaptive bandwidth bus architecture based on hybrid current/voltage mode repeaters has been proposed in [32], [33] for long global RC interconnect static busses that achieve high-data rates while minimizing the static power dissipation associated with current-mode (CM) signaling. An adaptive bandwidth bus scheme [22] that uses CM sensing to allocate interconnection bandwidth when input data transitions are sensed and otherwise remains in low-bandwidth voltage-mode to compensate for the increase in static power dissipation associated with CM signaling. And differential CM signaling that uses driver pre-emphasis circuit [32] reduce intersymbol interference (ISI) and increase channel bandwidth by emphasizing the high-frequency signal components or attenuating the low-frequency components to improve the delay and power performance Measurement results indicate a 62% improvement in static power dissipation over CM sensing techniques while achieving 40% increase in maximum data rate over VM signaling.

IV. CONCLUSIONS AND FUTURE WORK

Continued CMOS process scaling and system integration continues to increase the on-chip communication demands beyond what conventional digital signaling can efficiently provide. To this end, a survey of various efficient interconnects implementation techniques was presented comparing both voltage mode signaling (VMS) and current mode signaling (CMS) schemes for improving performance based on delay and power reduction. VMS techniques have the limitations to operate beyond 5GHz whereas CMS scheme will provide faster and reliable performance at these high frequencies and also have the potential to improve both speed and dynamic power consumption. It consumes much less power compared to the improved repeater circuits. The analysis shows that CMS scheme is at least an order times faster than VMS scheme. The only advantage of VMS is that it can interface directly with circuits and no converters are required and hence power efficient. In the recent SoCs, with high packing density and complex circuits, one has to opt for both CMS and VMS signaling technique based on the circuit scheme, how the far signal has to travel etc. Hybrid Current mode and Voltage mode circuits based on good floor planning and placement practice is planned to be proposed which can result in most favorable solution for the performance improvement of SoCs. Hybrid VM/CM signaling with different CM signaling scheme can be analyzed for performance improvement. Furthermore, signal integrity parameter (like impedances, crosstalk, power loss, attenuation, reflections etc.) for high speed interconnects may also be considered for further improvement.

REFERENCES


