Pulse Width Modulation for Reduced Peak Power Full-Swing On-Chip Interconnect

Mackenzie Scott
mrscott@ucdavis.edu

Rajeevan Amirtharajah
ramirtha@ece.ucdavis.edu

Department of Electrical and Computer Engineering
University of California, Davis
Davis, CA 95616

ABSTRACT

A full-swing on-chip interconnect using pulse width modulation (PWM) was designed and fabricated in 0.25μm CMOS for application in a tiled signal processing array architecture targeting wireless sensor nodes. Measurements show a decrease in the worst case power of 7% over traditional binary signaling for 11.8mm long wires at 10Mbps throughput and 0.7V V_{DD}. Power savings are projected to increase to 20% for 30mm long wires, while becoming beneficial for shorter wires at future process nodes without multiple wires or power rails. Power savings occur for wire lengths greater than 1.34mm and reach 20% at 3.03mm at the 32nm node. Savings are also projected to reach 18.2% when used on wires spanning 16 tiles in the target architecture.

Categories and Subject Descriptors

B.7.1 [**Integrated Circuits**]: Types and Design Styles

General Terms

Design

Keywords

Pulse width modulation, low power interconnect, peak power

1. INTRODUCTION

On-chip interconnect consumes an increasingly large fraction of an IC's power budget as CMOS scales because global wire scaling does not keep pace with transistor scaling. Reducing the peak power of wires, defined as the average power consumed during the worst case bit pattern, is important in energy harvesting wireless sensors, which have typical power budgets in the μW range, because exceeding the harvested power for extended periods may quickly discharge any stored energy reserves and shorten operating lifetime. Energy harvesting sensors can be operated at ultra low voltage due to low clock frequency requirements. For voltage-mode signaling, full signal swings are important to maintain adequate noise margins during ultra low voltage operation. In addition, generating multiple supply voltages (e.g., high swing for logic, low swing for interconnect) is often undesirable in these applications due to limited power electronics efficiency.

Previous work has used pulse-width modulation (PWM) to enable energy-efficient high throughput off-chip signaling[6]. This work adapts PWM signaling for on-chip global wires in low power, low throughput systems. Dynamic power is expressed by

$$P_{dyn} = \alpha C V_{DD}^2 f$$

where V_{DD} is supply voltage and f is clock frequency. PWM lowers wire activity factor (α), in exchange for increasing capacitance (C) by the overhead of the PWM circuitry. PWM encodes multiple bits into each return-to-zero (RZ) pulse (corresponding to a single charge-discharge cycle) driven onto a wire. The bits are received by detecting changes in delay and pulse width. Power savings require the PWM overhead capacitance to be a fraction of a conventional binary interconnect's load capacitance assuming that dynamic power dominates and voltage scaling is limited by the required throughput.

2. CIRCUIT DESIGN AND OPERATION

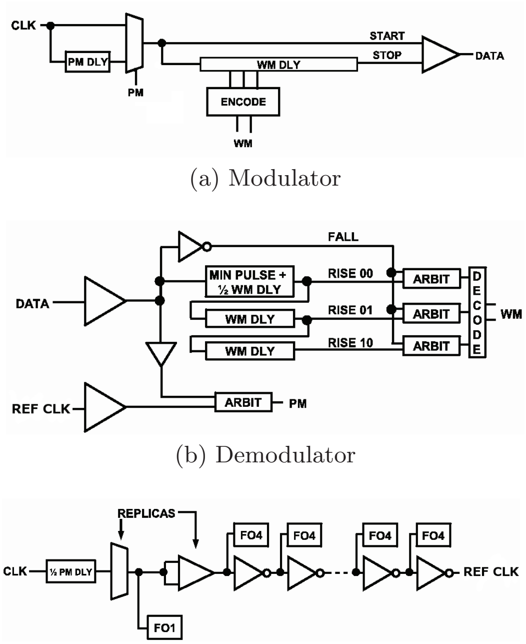
Our design encodes three bits on a single wire, allowing a 3X frequency reduction for fixed throughput. One bit is encoded in the relative delay between a reference clock and a transmitted data pulse. The rising edge of the system clock is used to generate both the reference and construct the data pulse. A further two bits are encoded by altering the transmitted pulse width.

The modulator is shown in Fig. 1(a). The **CLK** signal is either delayed or not based on the position modulation (PM) bit and becomes the **START** input to a pulse generator. The signal also passes through a delay line consisting of 8 digitally-controlled current-starved inverters, where the delay is set by two width modulation (WM) bits selecting different combinations of transistor lengths, and becomes the **STOP** input to the pulse generator. A modulator timing diagram can be seen in Fig. 2(a) and the delay element is shown in Fig. 3(a). The pulse generator is a NAND gate with one input inverted.Table 2 provides simulated pulse widths for different WM codings on an extracted layout when running at the designed V_{DD} of 700mV. The difference between each width ranges from 7.15ns to 7.66ns.

The demodulator is shown in Fig. 1(b). An arbiter (Fig. 3(b))

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ISLPED'09, August 19–21,2009,San Francisco, California, USA.
Copyright 2009 ACM 978-1-60558-684-7/09/08 ...$10.00.



(a) Modulator

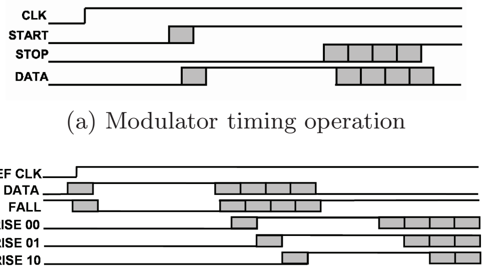
(b) Demodulator

(c) Reference clock generator

Figure 1: Block diagrams of system components.

compares the incoming data pulse to the reference clock to resolve the PM bit. **DATA** is delayed by a minimum pulse width plus a half unit WM delay, then by two unit WM delays. Arbiters compare the inverted falling edge of the received pulse to this delayed signal to create a thermometer code which is then decoded to recreate both WM bits. The primary delay is implemented by applying different control bits to a replica of the modulator delay unit. Other delays are implemented using a mix of current-starved and non-current-starved inverters. Delays may not track equally, leading to a range of allowed operating voltages. For an extended operating range, an alternate demodulator structure consisting of 3 individually controlled delay paths based on replicas of the modulator delay unit may be used, but with an increased power and area overhead. Fig. 2(b) shows demodulator operation.

Correctly resolving the PM bit requires that the reference clock replicate the delay through the modulator and wire path. If the system clock is distributed with low on-chip skew, the replica delay can be at the receiver, making clock forwarding unnecessary. The data pulse delay over the wire and repeaters is replicated using minimum-sized inverters driving fanout of 4 (FO4) loads (Fig. 1(c)). This replica can be made configurable to match reconfigurable interconnect.



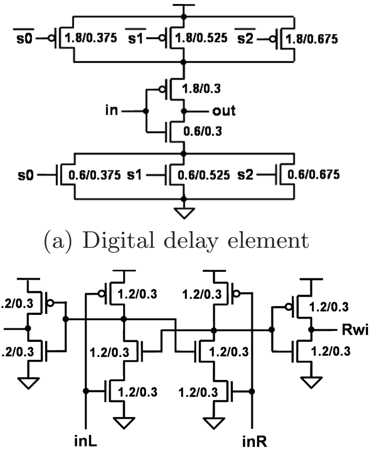
(a) Modulator timing operation

(b) Demodulator timing operation

Figure 2: Timing and operation of system.

Table 1: Pulse Width Timings @ V_{DD} = 700 mV

WM Coding	Width
00	46.72ns
01	31.79ns
10	39.45ns
11	24.64ns



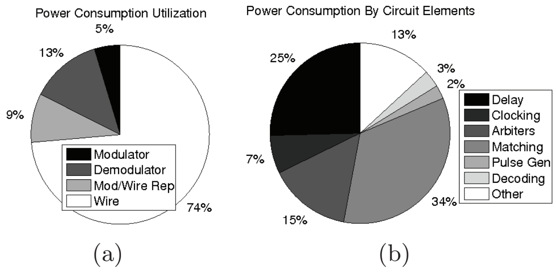
(a) Digital delay element

(b) Arbiter

Figure 3: Key circuit elements. Drawn dimensions shown (μm).

3. EXPERIMENTAL RESULTS AND COMPARISON

The proposed design was fabricated in 0.25μm CMOS. For testing, five different wire length paths are implemented with capacitance equivalent to 6.57, 7.89, 9.2, 10.5, and 11.8mm minimum-spaced wires on Metal 3, with repeaters inserted every 1.3mm. Correct operation was verified over a range of V_{DD}, throughput, and wire lengths, and power consumption was measured. Fig. 5 shows that PWM can reduce peak power on wires longer than 8.78mm. A breakdown of the power consumption distribution of an 11.8mm wire is shown in Fig. 4. Fig. 6 shows how modulated and binary



(a)

(b)

Figure 4: Power consumption distribution over an 11.8mm wire at 10Mbps.

signaling power vary with wire length and V_{DD} at 10Mbps throughput. Fig. 7 shows a surface plot of the crossover point (length at which PWM reduces peak power) variation with voltage and throughput. The relatively flat surface shows that this technique is applicable to reasonably-sized die and is not sensitive to V_{DD} and throughput once the threshold length is reached.

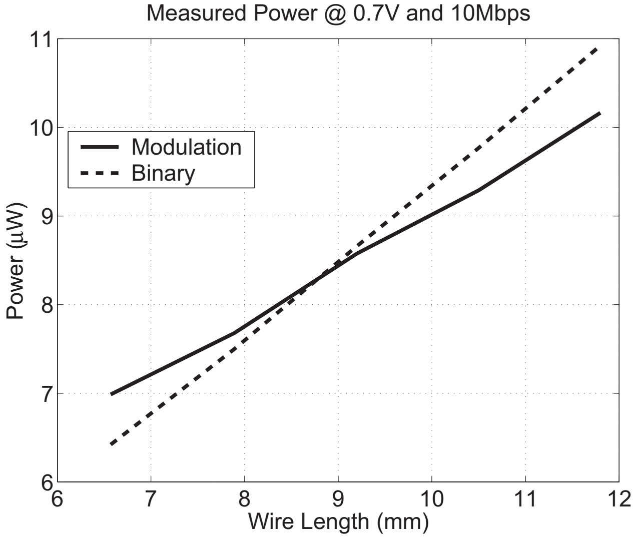


Figure 5: Measured power vs. length.

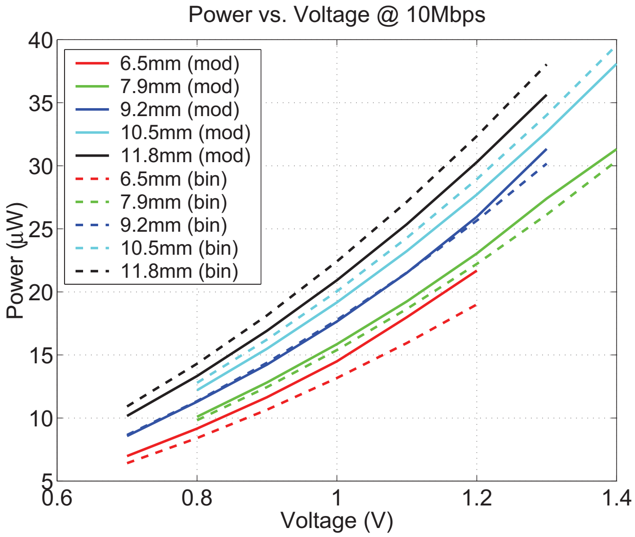


Figure 6: Measured power vs. voltage.

The graph in Fig. 8 shows the length required for a given wire's activity factor to yield a net power savings using the proposed modulation scheme. The required length asymptotically approaches infinity at an activity factor $\alpha = 0.66$ represented by the dashed line, while savings at realistic wire lengths occur with activity factors larger than ~0.8 ($\alpha = 1$ corresponds to a stream of alternating 1's and 0's). This can be seen in the following derivation , which compares the dynamic power for binary (P_{bin}) and modulated signaling (P_{mod}):

$$\begin{aligned}
P_{bin} &\geq P_{mod} \\
\alpha_{bin} C_{bin} V_{DD}^2 f_{bin} &\geq \alpha_{mod} C_{mod} V_{DD}^2 f_{mod} \\
\alpha_{bin} C_{wire} V_{DD}^2 f_{bin} &\geq \alpha_{mod} C_{wire} V_{DD}^2 f_{mod} \\
\alpha_{bin} f_{bin} &\geq \alpha_{mod} f_{mod} \\
\frac{\alpha_{bin}}{2} f_{clk} &\geq \frac{\alpha_{mod}}{n} f_{clk} \\
\alpha_{bin} &\geq \frac{2}{n}
\end{aligned}$$

The derivation makes the assumption that wire capacitance C_{wire} dominates circuit overhead for driving global wires for both binary and modulated signaling. Frequencies for both types of data signaling are normalized to the global clock frequency, where n is the number of bits encoded for PWM signals. $f_{bin} = f_{clk}/2$, corresponding to an alternating stream of 1's and 0's. The result shows that each extra bit encoded into a pulse decreases the necessary activity factor, leading to increased power savings. Although the necessary activity factor for power savings at realistic wire lengths remains high in this design, encoding of a fourth bit allows the graph in Fig. 8 to asymptotically approach 0.5, extending the range of applicable activity factors. This is achieved relatively easily by encoding an additional PM bit. However, each additional bit leads to diminishing returns, while the proposed design demonstrates the minimum amount of hardware necessary to achieve power savings of any kind.

Fig. 9 shows the projected power savings possible given extremely long wires, eventually approaching 29% at infinity. These projections are extrapolated from the measured results shown in Fig. 5. Fig. 10 shows the projected power consumption of reconfigurable interconnect for various wire lengths in a signal processing array for sensors described by Guo [3]. The FPGA-style interconnect consists of configurable switch boxes and connection boxes [10]. Switch boxes contain multiple instances of a unit consisting of a pass-gate mux followed by a tri-state buffer and a transistor to ground to keep unused nodes from floating. With PWM, the projected worst case power of a connection spanning 16 tiles would be reduced by 18.2% or 2.8μW. Fig. 13 shows an oscilloscope trace of the WM1 bit toggled on and off during operation. A die photo and summary of the circuit is shown

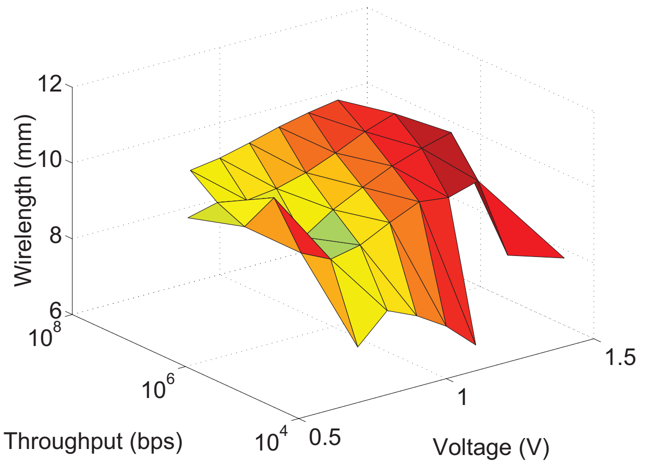


Figure 7: Crossover point variation.

Table 2: On-Chip Interconnect Implementation Comparison

Implementation	Technology	pJ/bit/mm	Throughput	Wire Length (mm)	Voltage (V)	Voltage Swing (V)
Zhang '00 [11] $^{\triangleleft}$	0.5μm	0.192	416 Mbps	10	2	0.5
Tzartzanis '05 [8]	0.11μm	0.171	1.73 Gbps	10	1.2	-
Venkatraman '06 [9]	65nm	0.216	2.5 Gbps	5	1.1	0.084 - 0.107
Ho '08 [4] $^{\triangleleft}$	0.18μm	0.28	200 Mbps	14	1.8	0.2
Akl '08 [1] $^{\triangleleft}$	90nm	0.120 †	2 Gbps	9	1.1	-
Singh '08 [7] $^{\triangleleft}$	90nm	0.45 †	2.33 Gbps	10	1.2	-
This work	0.25μm	0.143	100 Mbps	11.8	0.9	-
This work	0.25μm	0.0856	10 Mbps	11.8	0.7	-
This work	0.25μm	0.0862	1 Mbps	11.8	0.7	-
Binary Signaling	0.25μm	0.153 †	100 Mbps	11.8	0.9	-
Binary Signaling	0.25μm	0.0924 †	10 Mbps	11.8	0.7	-
Binary Signaling	0.25μm	0.0926 †	1 Mbps	11.8	0.7	-

$^{\triangleleft}$ Simulated results. † Peak energy quoted.

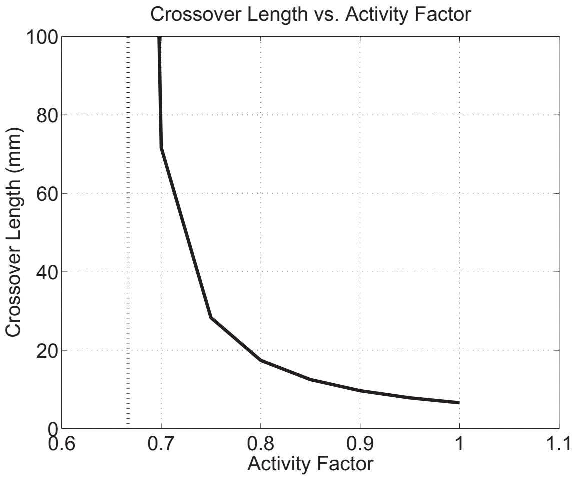


Figure 8: Activity factor crossover.

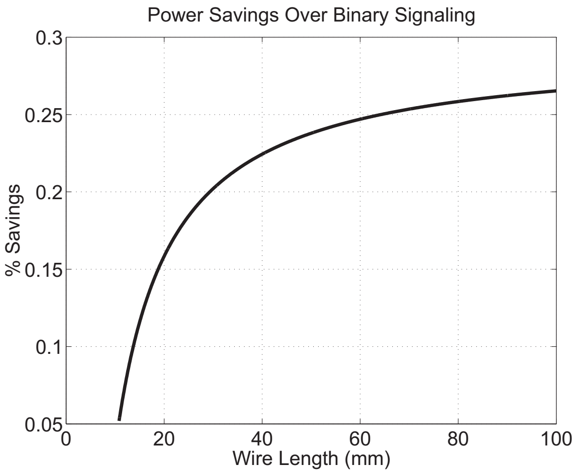


Figure 9: Projected potential power savings.

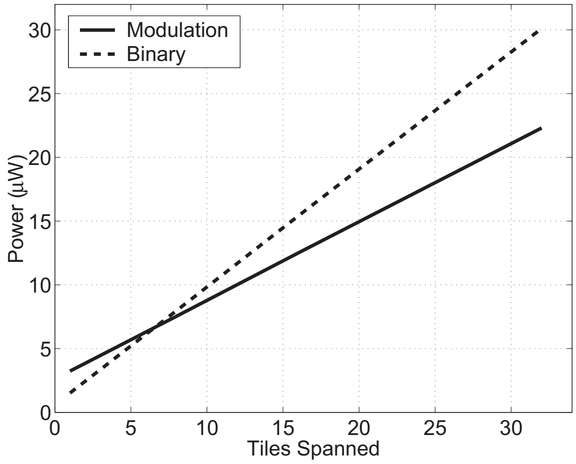
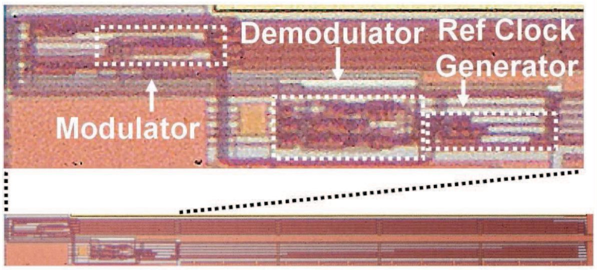


Figure 10: Reconfigurable interconnect power.

in Fig. 12. Leakage was measured to be ~ 2.3nA at 1V V_{DD}.

Table 2 presents a comparison of this work to a number of recently reported low-power on-chip coding and circuit techniques. Fig. 11 plots a trend line of recent works and reveals an upward trend in pJ/bit/mm as technology scales, demonstrating the increasing need for energy-efficient interconnect. This work is shown to be competitive to others in terms of energy-efficiency in addition to limiting peak power consumption. It is important to note that the proposed technique can easily be used combined with other low-power interconnect techniques. For example, simple low-swing signaling circuits or other transmitter/receiver pairs are easily incorporated at the output of the modulator and input of the demodulator. Unlike other techniques, generating additional power rails is not required. Full-swing levels decrease the need for differential signaling (to increase robustness to noise) and shielding (to improve crosstalk isolation).

The ITRS roadmap projects continued transistor scaling will reduce the switched capacitance of circuits with each technology node [5]. Global wire lengths and capacitance will increase relative to logic and thus the relative overhead of PWM will decrease. PWM interconnect power projections based on measured results, Predictive Technology Models [12, 2], and ITRS numbers show that as CMOS scales, the crossover wire length decreases down to 1.34mm



Technology	0.25 μm CMOS
Area	11,454 μm^{2*}
Transistors	503*
Vdd Range	0.7-1.2 V
Power	10.1 μW**
* Modulator, demodulator, and ref clock generator	
** 0.7V, 10Mbps, 11.8mm	

Figure 12: Die photo and summary of test circuit.

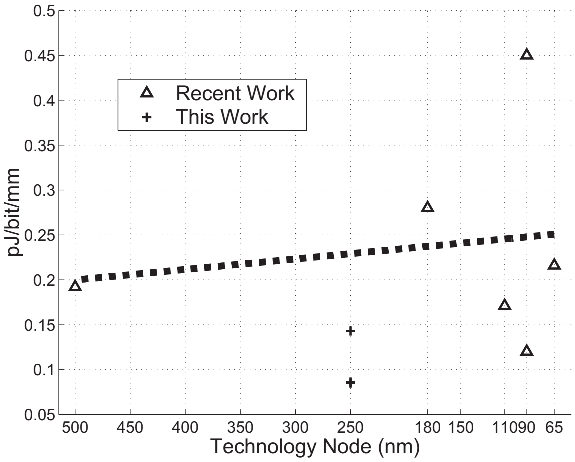


Figure 11: Trend line of pJ/bit/mm vs. technology node of recent low power on-chip interconnects.

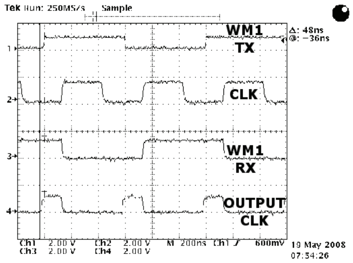


Figure 13: WM1 bit oscilloscope trace.

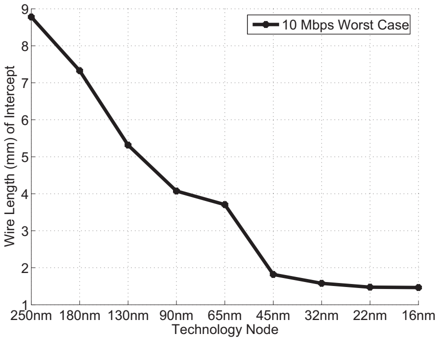


Figure 14: Crossover length scaling.

at the 32nm node (Fig. 14) with 20% power savings at 3.03mm, making the proposed technique increasingly applicable.

Increasing problems due to crosstalk with advanced technologies may become a concern for the technique due to the possibility of premature signal pulse termination causing incorrect arbiter edge detection. Additionally, large far-end pulse delay times from crosstalk may cause position detection errors, requiring the need for signal integrity techniques such as increased wire spacing, staggered repeaters, or shielding.

Future technology design will also require a focus on designing for variability. Variation affecting delays can be overcome by designing minimum changes in pulse widths to accommodate decreasing delay margins. Although added delay will affect maximum operating throughput and frequency, they should still remain acceptable within the targeted sensor node applications. The digital nature of delay creation also allows for post fabrication calibration at the expense of additional logic and configuration bits, further accounting for variability.

4. SUMMARY

A PWM technique for low peak power full-swing on-chip interconnect has been demonstrated in 0.25μm CMOS. The design encodes 3 bits into a single RZ pulse on a wire, enabling a reduction in activity factor in exchange for additional capacitive overhead for modulation and demodulation. It was shown to effectively reduce peak power when driving global wires longer than ~9mm. This length was roughly constant over both throughput and voltage, and is projected to decrease at more advanced process nodes (1.34mm at 32nm node). Measured results show a decrease in peak power of 7% over traditional binary signaling for wires of length 11.8mm at 10Mbps and 0.7V V_{DD}. Power reduction is projected to increase to 20% for 30mm long wires, and 18.2% for wires spanning 16 tiles in the targeted tiled DSP array architecture.

5. ACKNOWLEDGMENTS

This work was supported by the Focus Center Research Program Interconnect Focus Center, Subcontract No. B-12-M06-S12.

6. REFERENCES

[1] C. Akl and M. Bayoumi. Transition skew coding for global on-chip interconnect. In *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, volume 16, pages 1091–1096, Aug. 2008.

[2] Y. Cao, T. Sato, D. Sylvester, M. Orshansky, and C. Hu. New paradigm of predictive mosfet and interconnect modeling for early circuit design. In *IEEE Custom Integrated Circuits Conference*, pages 201–204, 2000.

[3] L. Guo, M. Scott, and R. Amirtharajah. An energy scalable computational array for sensor signal processing. In *IEEE Custom Integrated Circuits Conference (CICC)*, pages 317–320, 2006.

[4] R. Ho, T. Ono, R. Hopkins, A. Chow, J. Schauer, F. Liu, and R. Drost. High speed and low energy capacitively driven on-chip wires. In *Solid-State Circuits, IEEE Journal of*, volume 43, pages 52–60, Jan. 2008.

[5] ITRS. International technology roadmap for semiconductors. 2007.

[6] T. Simon, R. Amirtharajah, J. Benham, J. Critchlow, and J. Knight, T.F. A 1.6gb/s/pair electromagnetically coupled multidrop bus using modulated signaling. In *IEEE International Solid-State Circuits Conference*, pages 184–487 vol.1, 2003.

[7] P. Singh, J.-s. Seo, D. Blaauw, and D. Sylvester. Self-timed regenerators for high-speed and low-power on-chip global interconnect. In *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, volume 16, pages 673–677, June 2008.

[8] N. Tzartzanis and W. Walker. Differential current-mode sensing for efficient on-chip global signaling. In *Solid-State Circuits, IEEE Journal of*, volume 40, pages 2141–2147, Nov. 2005.

[9] V. Venkatraman, M. Anders, H. Kaul, W. Burleson, and R. Krishnamurthy. A low-swing signaling circuit technique for 65nm on-chip interconnects. In *SOC Conference, 2006 IEEE International*, pages 289–292, Sept. 2006.

[10] W. Wolf. *FPGA-Based System Design*. Prentice Hall, 2004.

[11] H. Zhang, V. George, and J. Rabaey. Low-swing on-chip signaling techniques: effectiveness and robustness. In *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, volume 8, pages 264–272, Jun 2000.

[12] W. Zhao and Y. Cao. New generation of predictive technology model for sub-45nm early design exploration. In *IEEE Transactions on Electron Devices*, pages 2816–2823, Nov. 2006.