Self-timed Regenerators for High-speed and Low-power Interconnect

Jae-sun Seo, Prashant Singh, Dennis Sylvester and David Blaauw
Department of EECS, University of Michigan, Ann Arbor, MI 48109
{jseo,prsingh,dmcs,blaauw}@umich.edu

Abstract

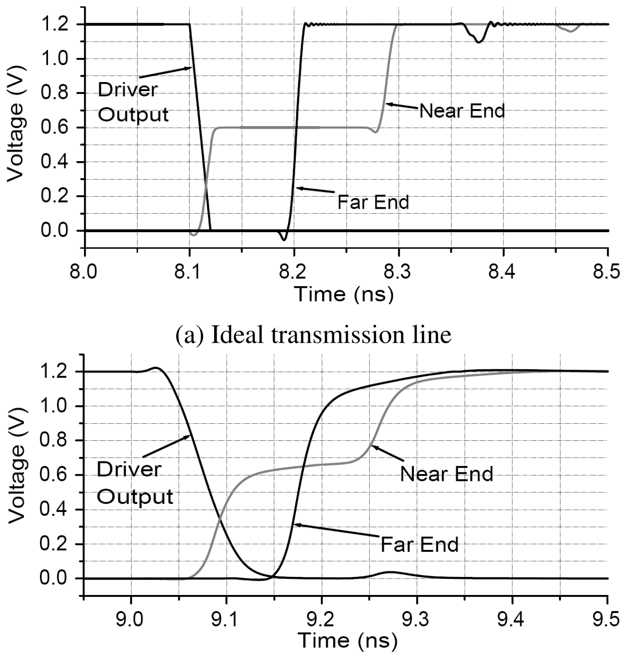
In this paper, we propose a new circuit technique called self-timed regenerator (STR) to improve both speed and power for on-chip global interconnects. The proposed circuits are placed along global wires to compensate the loss in resistive wires and to amplify the effect of inductance in the wires to enable transmission line like behavior. For different wire widths, the number of STR and sizing of the transistors are optimized to accelerate the signal propagation while consuming minimum power. In 90nm CMOS technology, STR design achieved a delay improvement of 14% *over the conventional repeater design. Furthermore,* 20% *power reduction is achieved for iso-delay, and* 8% *delay improvement for iso-power compared with the repeater design.*

1. Introduction

As technology continues to scale, the delay of local wires decreases while the delay of global wires remains the same or even increases making global interconnect a performance bottleneck. Furthermore, the requirement of high clock frequency leads to careful consideration of inductance of the lines, dispersion, and other transmission line effects. On-chip global interconnects are becoming a major bottleneck for circuit design with respect to overall chip performance and power constraints. Until now, repeaters have been the commonly used method to reduce the quadratic dependence of interconnect delay on wire length. However, as CMOS technology scaling continues, the number of repeaters increases dramatically. Deep submicron projections in [1] show that global interconnect distribution (repeaters + wires) will consume $\approx 40\%$ of the total power in $50nm$ technology. In [2] the distance between repeaters is expected to reduce rapidly as scaling continues, and that repeater becomes critical at the synthesizable block level as well. Recently, a microprocessor design [3] reported using as many as 12,900 repeaters showing that power and area overhead due to repeaters is becoming a serious concern.

A number of methods to address interconnect issues without using repeaters has been proposed [4, 5, 6]. The first design uses so-called boosters [4] where extra current is supplied when a transition is detected. However, the fact that it has a stack of two transistors in the charge path limits the speed improvement. In [5] a method is proposed where the receiver biases the voltage at which a transition is detected based on the expected transition direction. Recently, [6] proposed a capacitive coupling accelerator, similar to a booster, to reduce RC delay, but the improvement over repeater design was not as significant. Reducing the voltage swing has been used [7] to improve power, but is problematic in that another power rail has to be present and the delay tradeoff is not highly favorable. Also, driver preemphasis techniques [8] have been used to de-emphasize low frequency part to reduce inter-symbol interference and save power, but reduced signal swing at the receiver input is susceptible to noise and process variations. Finally, alternative approaches include modulated signaling [9] and pulsed current-mode signaling [10]. These methods achieved near speed-of-light latency, but they require wide wire topologies with low loss characteristics and the complexity of these designs makes them difficult to adopt in the industry.

In this paper, we present a new circuit technique to achieve high performance, repeater-less propagation for global interconnects. We propose a transmission line configuration where the driver is perfectly matched to the line impedance, to supply reflections, and where the receiver is lightly loaded to amplify the propagated wave to full swing transitions. The main challenge is the loss inherent in the narrow lossy wires found on-chip. Rather than attempting to use very wide wires to achieve low loss throughout the interconnect lines, we propose the use of active circuits that are positioned periodically along the line to compensate for the attenuation and achieve fast signal propagation. One advantage of such a repeater-less propagation is its suitability to point-to-point bidirectional signaling. A unique feature of the proposed regenerator is its self-timed design. Hence, once the regenerator has sensed and amplified a transition, it automatically resets itself to a transition monitoring mode. In this mode, it does not actively drive the wire and is ready to detect the next transition. This is in contrast to the booster design [4] where the signal line is actively held by the booster at the existing state until a transition to the opposite state is detected. This introduces a so-called fight between the booster and the transition in the early part of the transition and makes the booster behavior less efficient. The proposed regenerator design avoids this “fight” while detecting and accelerating the transition, thereby allowing a much stronger amplification path and obtaining improved



(a) Ideal transmission line

(b) Wide interconnect line (width : $40\mu m$)

Figure 1. Interconnect with transmission line behavior.

delay and power properties. The proposed design was implemented and tested for a number of interconnect structures. For a $0.3\mu m$ wide, $10mm$ long interconnect, using STRs instead of optimal repeaters, total power consumption is reduced by 19.8% for the same delay. Alternatively, for the same power, the delay is improved by 7.7%, and maximum delay improvement over repeater design is 14.0%.

The remainder of the paper is organized as follows. Section 2 describes the operation of the self-timed regenerator circuit. Experimental results are found in Section 3. The paper is concluded in Section 4.

2. Self-timed Regenerator Design

2.1. Transmission Line Configuration

We first consider a lossless transmission line where the driver is perfectly matched with the line impedance and the receiver is sufficiently small to present a negligible load. When the driver input transitions, a wave of V_{DD}/2 will be propagated along the transmission line toward the far end, due to the matching of the driver impedance. When the propagated wave reaches the receiver, this voltage will be doubled to the full rail due to the light loading of the receiver, as shown in Figure 1(a). Note that while a reflected wave is sent back through the transmission line, this reflected wave will be absorbed completely by the driver since it is perfectly matched. This configuration is particularly advantageous for point-to-point signaling in VLSI designs, for instance between the processor and the cache. Taking advantage of reflection at the receiver termination to obtain a full swing signal allows the new design to be easily incorporated in existing design methodologies.

For a wire that is sufficiently wide, the resistance is insignificant and we obtain behavior similar to that of an ideal transmission line, as shown in Figure 1(b) for a $40\mu m$ wide wire. However, when the wire becomes thinner, the resistance becomes significant and the signal attenuates as it propagates through the wire. In this case, the signal swing at the receiver may not be sufficient to detect the transition reliably and signal propagation speed is also degraded. To compensate for this signal degradation, our design enhances the transition by properly supplying additional current, while still utilizing the impedance matching and the receiver reflection.

2.2. Circuit Operation

The self-timed regenerator (STR) is designed to quickly detect and accelerate the transition for a certain amount of time with a certain amount of current from the rail. Figure 2 shows the self-timed regenerator (STR) circuit on both sides of the interconnect. The upper part is the pull-up circuitry for accelerating the rising transition and the lower part is the pull-down circuitry for the falling transition, each one being complementary to the other. When there is a low-to-high transition, the pull-up circuit is triggered and the pull-down circuit remains insensitive to the signal line. Similarly, the pull-down circuit is triggered and the pull-up circuit remains insensitive in case of a high-to-low transition.

The main idea is to generate a pulse at node B and C which would turn on P3 and N6 for a time equal to the width of the pulse. When transistors P3 and N6 are turned on, additional current is supplied from the power rail to the propagating signal to expedite the transition. Transistors N1 and P4 are low threshold transistors which turn on quickly according to the polarity of the signal. P1 and N4 are weak transistors which are present only to establish and maintain initial conditions at nodes B and C. The delay set by the odd-number inverter chain determines the width of the pulse. The number and size of inverters in the chain can be optimized for different wires and constraints. This enables self-timing of the pulse width.

The initial state of the internal nodes of the circuit should be known. When the signal line is at low-voltage steady state, transistor P1 is driving node B to V_{DD}. Node D is also set at V_{DD}, making the pull-up circuit just ready to detect low-to-high transition while lower circuit remains insensitive to any rising transition. If any noise pulls node D down to GND, P4 and P5 charges node C and after going through a chain of inverters, P6 actively drives node D back to V_{DD}, which is the desired initial condition. Similarly, when the signal line is at high-voltage steady state, node C and node D is set at GND.

Upon a transition, the circuit works as follows and the timing diagram for each transition is shown at Figure 3. Note that node A is the interconnect line itself. When the wire is initially at GND, the next transition will be a rising transition. Since transistor P5 is off, the pull-down circuit would be insensitive to this transition. When a rising transition is detected by N1, node B is pulled down to GND immediately as can be seen in Figure 3(a). P3 turns on and

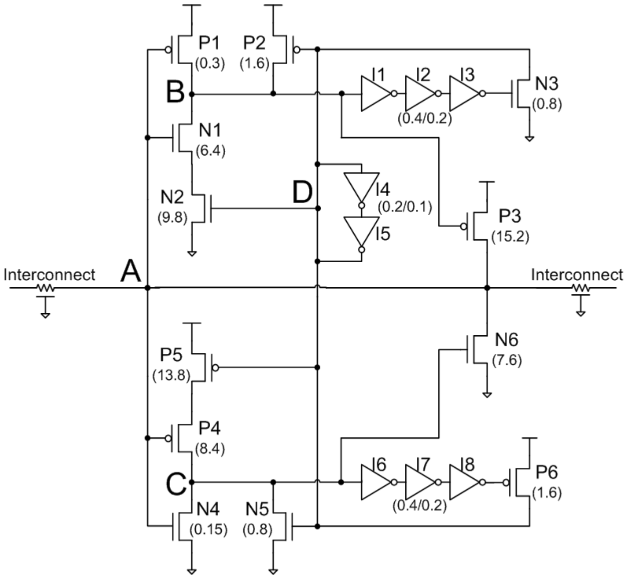


Figure 2. Self-timed regenerator circuit. Optimal sizing(unit: μm) for power reduction when 5 STRs are placed for a 0.45μm wire is shown.

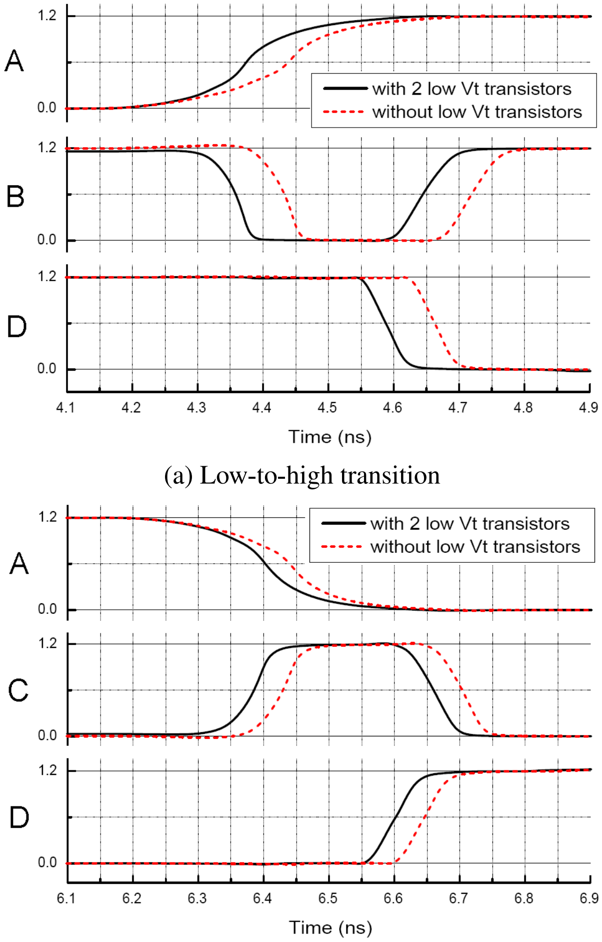
it enhances the transition of the signal. After some delay, N3 turns on and node D is grounded, also shown in Figure 3(a). P2 charges node B to V_{DD}, so P3 turns off. After some time, N3 turns off but node D is maintained at GND by the cross-coupled inverters. Now, the pull-up circuit becomes insensitive to the high-to-low transition as N2 is turned off.

In case of a falling transition, the timing waveforms of the internal nodes are shown in Figure 3(b). Since node D is held at GND, P5 is turned on and P4 is waiting for the falling transition, while the pull-up circuit now remains insensitive to the transition. As soon as the falling transition occurs, P4 and N6 is turned on to quicken the transition. After the inverter chain delay, node C is back at GND, turning off N6, and node D is charged to V_{DD} again as shown in Figure 3(b). Now, the pull-down circuit becomes insensitive and the pull-up circuit is ready to detect a low-to-high transition. Figure 3 also compares the waveforms between the case when we use low Vt transistors for N1 and P4 and when we do not. Considerable performance improvement is observed by using 2 low Vt transistors in the STR.

One of the key features of our design is that transistors (P2, N1, N2) and (P5, P4, N5) are never turned on simultaneously. This eliminates the fight during a transition which degrades the performance improvement of the regenerator and results in additional short circuit current, resulting in power reduction. In addition, the signal line is accelerated by a single transistor in series between the signal line and the supply rail. This allows for a high drive current which results in improved signal acceleration.

2.3. Sizing of the Circuit

Sizing of the transistors in the STR circuit should be done carefully to facilitate the desired operation. The optimal sizing of STR for maximum power reduction while maintaining same delay with repeaters when 5 STRs are placed on a 0.45μm wide wire is shown in Figure 2. Transistors N1, N2, P3 and P4, P5, N6 are relatively larger transistors than the others. As the number of STRs placed on a wire increases, their sizes get reduced. The size of transistors P3 and N6 determines the amount of current supplied to the signal line. Sizes of N1, N2 and P4, P5 determine the response time of the circuit to the propagating wave. The faster these transistors are, more quickly transistors P3, N6 get triggered and the better output waveform is at the far end of the wire. The rest of the transistors are not critical in terms of speed and are sized relatively smaller than these six transistors to minimize power consumption. N3 and P6 should be strong enough to switch the state of cross-coupled inverters. Sizes of P2 and N5 determine the slope of the trailing edge of the pulse. The sizing of STR is optimized for every combination of wire width and number of STRs placed on the line, resulting in different sizing for each different situation. More details of STR sizing in the overall interconnect system comparing to the repeater scheme is further discussed in section 3.1.



(a) Low-to-high transition

(b) High-to-low transition

Figure 3. Timing diagrams of STR at rising and falling transition. Speedup due to 2 low Vt transistors is shown.

Table 1. STR power and performance comparison

Wire width	Opt. Repeaters	Power reduction (Iso-delay)	Opt. Regen	Delay improvement (Iso-power)	Opt. Regen	Delay improvement (Best case)	Opt. Regen
0.3μm	10	19.8%	6	7.7%	9	14.0%	13
0.45μm	8	17.6%	5	6.8%	6	13.8%	12
1μm	5	16.9%	3	7.4%	5	13.3%	10
4μm	2	10.9%	1	8.1%	1	11.8%	9

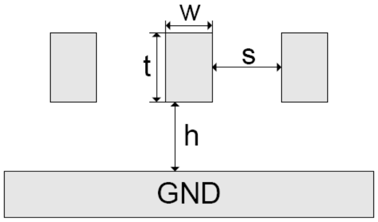


Figure 4. Structure of global interconnect.

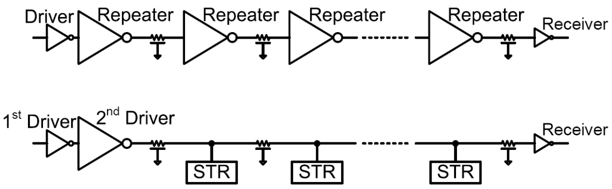


Figure 5. Repeater/STR implementation scheme.

3. Experimental Results

The 90nm technology results are obtained from SPICE simulations using industrial device models. The simulation is done for a relatively broad range of widths including 0.3μm, 0.45μm, 1μm, and 4μm. We modeled the interconnect line as a top global layer metal with shielding wires on either side, as shown in Figure 4. For simulation of different wire widths, width w in Figure 4 is changed and all the other parameters such as spacing, thickness, and distance from the ground plane are fixed. The resistance, inductance, and capacitance values for a distributed interconnect have been extracted using FastHenry [11] and Predictive Technology Model [12] for a 10mm line, and the line length is fixed throughout this paper.

We also compared the proposed STR and repeater technique against a traditional booster design [4]. However, the booster design was not able to improve on the performance of the repeater design even after extensive optimization of the transistor sizing of the booster topology. Hence, no comparison of our proposed approach against the booster design is given since the gains will be more than that compared to repeater designs. Previous reported measurements for the booster design were performed by measuring delay from the output of the inverter driving the interconnect to the input of the receiver gate. This method of delay measurement ignores the delay due to the loading of the initial driver and hence might not be accurate.

3.1. Repeater and STR Design Scheme

The overall scheme to compare repeater and STR is shown in Figure 5. To make the comparison fair, we have identical initial drivers both in the repeater and STR scheme. In the repeater scheme, the first repeater is placed after the initial driver. In the STR scheme, the initial driver is followed by a stronger driver, which is sized properly to match the impedance of the line. As a result, the optimum size of this 2nd driver increases as the wire becomes wider. Throughout the interconnect line, repeaters and STRs are inserted regularly. Clearly, we cannot just make the driver arbitrarily strong because this will add more delay to the initial driver. The delay is measured from the initial driver input to the final receiver output.

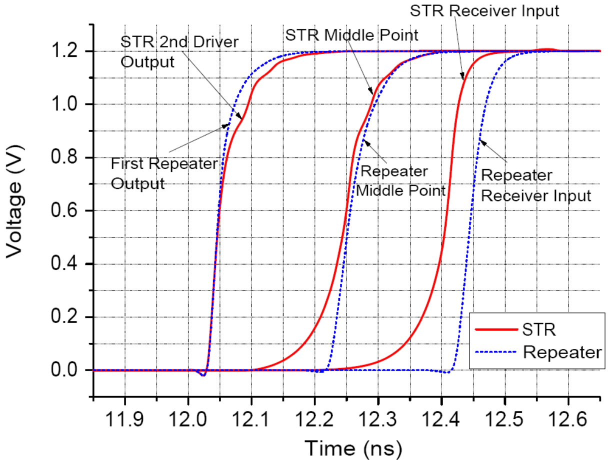


Figure 6. STR and repeater simulation waveforms of 0.3μm wide interconnect.

At each wire width, both the number of repeaters and the repeater size are varied to achieve best performance. For a certain number of repeaters placed on the wire, the size of the NMOS transistor in the repeater is swept until it reaches maximum performance while the P/N ratio of the repeater is kept at 2. Among all the combinations, optimal number of repeaters and optimal size which results in the best case delay is chosen, and this serves as the baseline of comparison for the STR designs. Similarly, for a given wire geometry, both the sizing of STR and number of STRs along the interconnect is varied and optimized to achieve better energy and delay compared to the repeater scheme. For simplicity, when multiple repeaters or STRs are placed on a wire, all repeaters and STRs are sized identically.

Figure 6 shows the waveform of intermediate nodes of 0.3μm wide wire simulation for STRs and repeaters. For

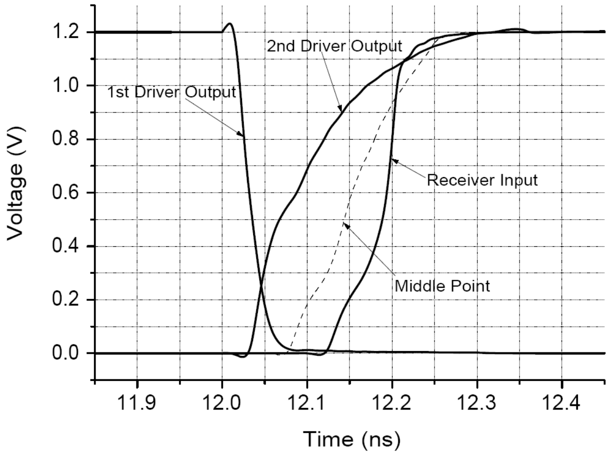


Figure 7. STR simulation waveform of 4μm wide interconnect.

this lossy wire, we see significant delay improvement using STRs compared to the repeater design. The waveform of a 4μm wide interconnect is shown in Figure 7. Since the resistance of the wire is now reduced significantly, we start to see the transmission line effects with reflections at the far end. This results in faster transition time at the output of the interconnect line than an intermediate point of the line.

The overall power and performance comparison is summarized in Table 1. "Opt. Repeaters" is the optimal number of repeaters for each wire width, and "Opt. Regen" is the optimal number of STRs for each corresponding scenario. In Table 2, delay, energy, and total device width comparison is shown for the iso-delay case of each wire width. Total device width is the total width of the transistors in the drivers, receivers, STRs and repeaters.

3.2. Power

To measure how much power we can save with STRs, the power conparison of the two designs is performed with the same delay. For iso-delay, power reduction up to 19.8% is achieved in the STR design. This is first due to the fact that smaller numbers of STRs are needed than that of repeaters at the same delay constraint. Also, the STR circuit need not be oversized to produce equivalent delay with repeaters. Therefore, the total area is reduced significantly as shown in Table 2. Furthermore, the short-circuit current is minimized in the STR design because there is no strongly conducting direct path from V_{DD} to GND at any given time.

In Table 1, we observe that power savings of STRs comparing to the repeaters decreases as the wire width increases. This is because the capacitance dominates the interconnect parasitics in wide wires, and therefore sizing down the STR cannot reduce the power dissipated by the capacitance by a large amount.

3.3. Performance

Similarly, to fairly compare the performance of STRs with repeaters, a power constraint is imposed. The power consumption of both STR and repeater design is set to be the same, and the delay is measured in each case. Across different wire widths, the maximum delay improvement is 8.1% for iso-power.

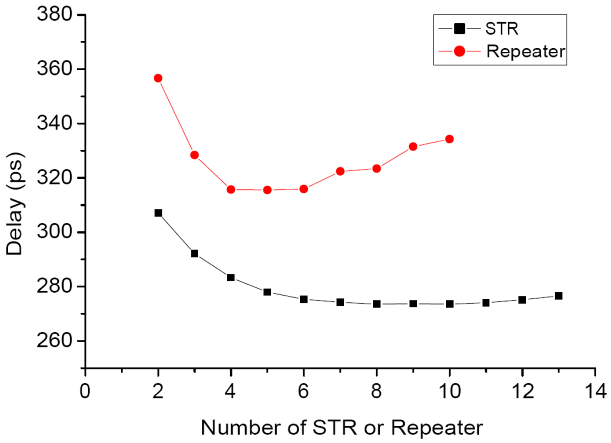


Figure 8. Delay comparison with different numbers of STRs and repeaters (width : 1μm). Sizing is optimized for each different number of STR and repeater.

Table 2. Energy and area comparison (iso-delay)

Wire width	Scheme	Delay (ps)	Energy (pJ)	Total device width (μm)
0.3μm	Repeater	482.0	5.64	738
	STR	481.9	4.52	451
0.45μm	Repeater	424.6	5.49	666
	STR	424.1	4.52	421
1μm	Repeater	315.5	5.28	543
	STR	315.7	4.39	317
4μm	Repeater	231.4	5.03	282
	STR	231.2	4.48	130

Finally, we tried to obtain the maximum performance with the STR design when performance has a higher priority than power consumption. In Figure 8, delay comparison of the two designs with 1μm wide wire shows that the performance of the proposed STR design dominates that of the repeater design. The more STRs added along the line, we get better performance up to 14.0%. More delay improvement is achieved for thinner wires, and the performance improvement slightly decreases for wider wires.

Figure 9 shows energy vs. delay for STRs and repeaters. The data points in this plot are the minimum energy points obtainable with the given delay for STRs and repeaters. We can see that the STR energy-delay curve exists in the left-bottom side than that of the repeater. The data points for iso-delay and iso-power are also shown in the plot.

3.4. Low Vt Repeaters and Leakage power

Since we are exploiting the fast behavior of 2 low Vt transistors in STRs, we also optimized repeaters with low Vt devices to achieve a comprehensive comparison. Furthermore, we measured leakage power of high Vt repeater,

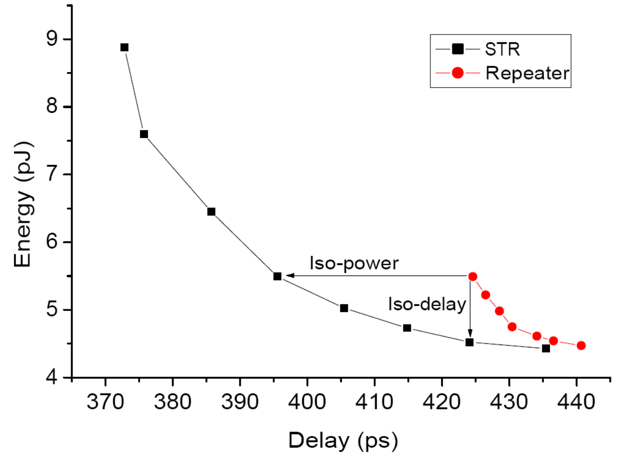


Figure 9. Energy vs. Delay of STR and repeater (width : 0.45μm).

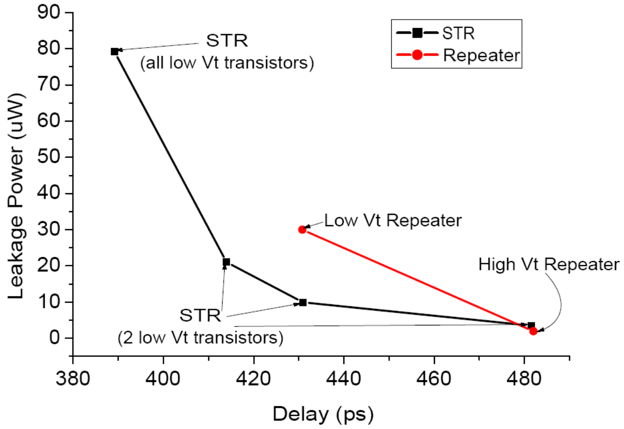


Figure 10. Leakage power comparison with different Vt assignments (width : 0.3μm).

low Vt repeater, and the STR design. When repeaters are designed with low Vt devices, the number of repeaters to obtain optimal delay for each wire width was found to be the same as with high Vt devices while the sizing was different. Using low Vt transistors in the repeater design, 10$\sim$15% speed improvement was achieved, while leakage power increased significantly compared to high Vt repeater design.

Figure 10 shows the leakage power comparison with STR using two low Vt devices, repeaters with high Vt, and repeaters with low Vt for a 0.3μm wide wire. The data points in Figure 10 for repeaters are the best case delay for each configuration. It is shown that although low Vt devices are used for all repeaters, it cannot reach the speed of STR, which has only 2 low Vt transistors. In a 0.3μm wide wire, the STR still achieves 4% delay improvement compared to low Vt repeaters, while consuming less leakage power. When STR performance reaches the same speed as low Vt repeaters, the leakage power of STRs is 3X lower than that of repeaters. When the STR is performing at the same speed as high Vt repeaters, the leakage power of STR is comparable with that of high Vt repeaters. Again, as we saw in Table 2, the total area in the STR scheme is only 50$\sim$60% of that in the repeater scheme for similar delay. Since the subthreshold leakage current is proportional to the transistor device size, this reduces leakage power although there are more transistors in STR comparing to a repeater. Also, these results show that our design has a very specific critical path so that we gain considerable speed improvement by using a few low Vt devices without sacrificing leakage power. As the wire becomes wider and the capacitance dominates the interconnect parasitics, we observed that leakage power improvement over the repeater design diminishes.

4. Conclusions

In this paper, we presented a new circuit technique to improve delay and save power for global interconnects. To enable fast propagation of the signal without using repeaters, we added active circuits regularly along the interconnect which accelerate transition. For a 10mm line at reasonable wire widths, relatively few active circuits can be used along the line comparing to the number of repeaters to obtain iso-delay or iso-power with the repeaters. We could achieve 20% lower power at the fastest repeater speed achievable, 14% faster speeds than repeaters, and up to 3X leakage power reduction than the repeaters with same delay.

5. Acknowledgments

This work was supported by NSF and SRC.

References

[1] D. Sylvester and K. Keutzer, "Getting to the bottom of deep submicron II: A global wiring paradigm," *Proc. ISPD*, pp. 193-200, 1999.

[2] P. Saxena et. al., "The scaling challenge: can correct-by-construction design help?" *Proc. ISPD*, pp. 51-57, 2003.

[3] T. Takayanagi et. al., "A dual-core 64-bit ultrasparc microprocessor for dense server applications," *IEEE J. Solid-state Circuits*, Jan. 2005.

[4] A. Nalamalpu et. al., "Boosters for driving long onchip interconnects—Design issues, interconnect synthesis, and comparison with repeaters," *IEEE Trans. on CAD*, Jan. 2002.

[5] H. Kaul and D. Sylvester, "Transition aware global signaling (TAGS)," *Proc. ISQED*, pp. 10-14, 2002.

[6] H. Huang and S. Chen, "Interconnect accelerating techniques for sub-100-nm gigascale systems," *IEEE Trans. on Very Large Scale Integration (VLSI) Systems*, Nov. 2004.

[7] R. Ho et. al., "Efficient on-chip global interconnects," *Symp. VLSI Circuits Dig. Tech. Papers.*, pp. 271-274, 2003.

[8] L. Zhang et. al., "Driver pre-emphasis techniques for on-chip global buses," *Proc. ISLPED*, pp. 186-191, 2005.

[9] R. T. Chang et. al., "Near speed-of-light signaling over on-chip electrical interconnects," *IEEE J. Solid-State Circuits*, pp. 834-838, May 2003.

[10] A. P. Jose et. al., "Near speed-of-light on-chip interconnects using pulsed current-mode signaling," *Symp. VLSI Circuits Dig. Tech. Papers.*, pp. 108-111, 2005.

[11] M. Kamon et. al., "Fasthenry: A multipole-accelerated 3-d inductance extraction program," *IEEE Trans. on Microwave Theory and Techniques*, Sep. 1994.

[12] Y. Cao et. al., "New paradigm of predictive mosfet and interconnect modeling for early circuit simulation," *Proc. CICC*, pp. 201-204, 2000.