High Performance On-Chip Differential Signaling Using Passive Compensation for Global Communication

Ling Zhang¹, Yulei Zhang², Akira Tsuchiya³, Masanori Hashimoto⁴, Ernest S. Kuh⁵ and Chung-Kuan Cheng²
¹.Univ. of California, San Diego, CA ².Kyoto Univ., Kyoto, Japan,
³.Osaka Univ., Osaka, Japan, ⁴.Univ. of California, Berkeley, CA
¹{lizhang@cs.ucsd.edu}, ²{y1zhang,ckcheng}@ucsd.edu,
³tsuchiya@i.kyoto-u.ac.jp, ⁴hasimoto@ist.osaka-u.ac.jp, ⁵kuh@eecs.berkeley.edu

Abstract—To address the performance limitation brought by the scaling issues of on-chip global wires, a new configuration for global wiring using on-chip lossy transmission lines is proposed and optimized. We propose a signaling structure to compensate the distortion and attenuation of on-chip transmission lines, which uses passive compensation and inserts repeated transceivers composing sense amplifiers and inverter chains. An optimization flow for designing this scheme based on eye-diagram prediction and sequential quadratic programming (SQP) is devised. This flow is used to study the latency, power dissipation and throughput performance of the new global wiring scheme as the technology scales from 90 nm to 22 nm. Comparing to repeated RC wire, experimental results demonstrate that at 22 nm technology node, the new scheme can reduce the normalized delay by 80%-96%, the normalized energy consumption by 50%-94%. The normalized latency is 10 ps/m, the energy per bit is 0.2 ps/μm, and the throughput is 15 Gbps/μm. All performance metrics are scalable with technology, which makes this approach a potential candidate to break the “interconnect wall” of digital system performance.

I. INTRODUCTION

As technology scales, interconnect planning has been widely regarded as one of the most critical factors in determining the system performance and total power consumption. According to the prediction of ITRS roadmap [1], the 1 mm global RC wire delay at 45 nm technology is 385 ps, while the 10 level FO4 delay is below 200 ps. Given the fact that global wires with 1 mm length or more are very commonly used for on-chip communication nowadays, a big performance gap exists between the interconnect and the logic gates. Interconnects also consume a significant portion of total power. In [2], Magen et al. found that the interconnect power alone accounts for half the total dynamic power of a 0.13 μm microprocessor that was designed for power efficiency.

The conventional approach to deal with the interconnect delay problem is buffer insertion, which is also referred to as repeated RC wires. By inserting buffers or repeaters along the long wire, the relationship between wire delay and wire length changes from quadratic to linear. Repeater insertion improves the RC wire performance greatly but also introduces overhead in terms of power and wiring complexity. In [3], Zhang et al. compared the repeated RC wires under different design goals across multiple technology nodes. They demonstrated that to minimize delay, the optimum repeated RC wire has equal amount of wire capacitance and gate capacitance, which means half of the dynamic power is dissipated on repeaters.

On-chip transmission line (T-Line) has attracted intensive research focus in recent years. Comparing with repeated RC wires, transmission line delivers signals with speed of light in the medium. It also consumes much less power since the wave propagation eliminates the full swing charge and discharge on wire and gate capacitance. However the intersymbol interference (ISI) can be a barrier for performance, and various approaches have been proposed. [4] and [5] derived the analytical formula for optimal termination resistance. [6], [7] and [8] proposed the surfliner scheme that intentionally inserting shunt resistors along the wire to minimize the distortion. [9], [10], [11] and [12] adopted passive or active equalization schemes to reduce the ISI. To have better understanding of on-chip T-Line performance, [13] predicted the bit-rate of different wire length for future technologies, and [14] and [15] compared the latency and performance of RC wire and T-Line.

In this work, we propose a high performance on-chip global signaling with passive compensation. The proposed scheme is compared with the repeated RC wire in terms of latency, power and bandwidth, and the results are very promising. Our contributions include: 1) an on-chip global signaling scheme with passive compensation, 2) an optimization flow based on SQP method that optimizes the scheme for a given technology and wire dimension, 3) comparison between the proposed on-chip T-Line scheme and repeated RC wire under three different design goals at different technologies.

II. SIGNALING SCHEME FOR GLOBAL WIRING

The signaling scheme we propose is shown in Fig. 1(a), which consists of parallel RC equalizers, differential wires, termination resistance and transceivers. We adopt parallel RC circuit at the driver side to compensate the attenuation in high frequency components. For a given wire, the values of $R_d$, $C_d$ and the termination resistance $R_t$ determine the eye-opening and are optimized in our optimization flow (Section III). Two identical transceivers, which include a double-tail sense amplifier (SA) followed by a differential inverter chain as
indicated in Fig. 1(b), are used at both the driver and receiver sides to recover the signal back to full-swing.

A. On-Chip T-Line

1) Basic theory of on-chip T-Line: On-chip T-Line is very lossy due to the miniaturization of the wire cross section, and it can operate in either RC or LC region given different frequencies [17]. In RC region, the frequency is low, which makes \( \omega L \ll R \). Generally \( G \approx 0 \) for on-chip wires, and the propagation constant can be written as \( \gamma = \frac{R}{2\sqrt{L/C}} + j\omega \sqrt{LC} \). In RC region, both the attenuation and the phase velocity depend on frequency. The condition of \( \omega L \ll R \) is usually satisfied up to 10 GHz.

If the frequency increases such that \( \omega L \gg R \), the wire operates at LC region and the propagation constant becomes \( \gamma = \frac{R}{2\sqrt{L/C}} + j\omega \sqrt{LC} \). Therefore the attenuation constant is

\[
\alpha = \frac{R}{2\sqrt{L/C}} = \frac{R}{2Z_0}
\]

where \( Z_0 \) is the characteristic impedance of T-Line, and the phase velocity \( v = \frac{\omega}{\beta} = \frac{1}{\sqrt{LC}} \). In LC region, both the attenuation and the phase velocity are independent of frequency.

Two parameters need to be considered in modeling the wire. The critical wire length distinguishes lumped-element region and distributive-element region, which can be computed as follows [17]:

\[
L_{\text{critical}} = \frac{0.25}{\sqrt{(R + j\omega L)(j\omega C)}}
\]

The other is the corner frequency \( f_{\text{LC}} \) between RC region and LC region, which is defined as:

\[
f_{\text{LC}} = \frac{1}{2\pi} \frac{R_{\text{DC}}}{L}
\]

where \( R_{\text{DC}} \) is the DC resistance of the wire.

2) Transmission line geometries: We use the differential stripline as shown in Fig. 2. We restrict the wire thickness versus width \( W \) (the aspect ratio) to be 2 according to the data given by ITRS roadmap [1], and we define the vertical clearance equals wire thickness for simplicity. In this work, we choose 5 mm length on-chip global communication. Considering noise issue, we assume the SA has an input threshold voltage \( V_{\text{min}} \) (the half of the differential input voltage), and the input voltage level is 1 V, then the attenuated voltage for a single wire can be no less than \( V_{\text{min}} \):

\[
e^{-\alpha l} \geq V_{\text{min}}
\]

Noticing that the resistance of wire is determined by \( H \) if \( W = 0.5H \), we can have the lower bound of \( H \) from Eq. (1) that satisfies the eye-opening constraint:

\[
H \geq \sqrt{\frac{2\rho_{\text{Cu}}}{-Z_0 \ln V_{\text{min}}}}
\]

We list the lower bound of \( H \) with \( V_{\text{min}} = 25 \text{ mV} \) for each technology in Table I, row 8. In the optimization flow (Section III) we always use this minimum \( H \) since it gives us higher wire density. For each technology and a given \( H \), larger spacing increases \( Z_0 \), reduces the attenuation and generates better eye. Therefore we vary the \( S \) so that \( S = 0.5H,1.5H,2H \) to observe the performance change (last row in Table I). The corresponding critical length and corner frequency for \( S = 0.5H \) case are also listed in Table I. It can be seen that \( f_{\text{LC}} \) has small variations because wire inductance does not change much and wire resistance is tuned to be very similar by selecting the lower bound \( H \). At the same time, \( L_{\text{critical}} \) is much smaller than the wire length (5 mm), and hence we can safely model the wires as T-Line in LC region. When \( S \) increases, \( f_{\text{LC}} \) decreases as \( L \) increases, which pushes the wire into LC region further.

3) Delay and power models: The wire delay has two parts: time of flight \( T_{\text{flight}} \) and \( T_c \). Once the input signal arrives at far end, it requires some time to rise up to \( V_{\text{out}} \geq V_{\text{min}} \) to trigger the SA. For the “1010” input pattern, the rise time can be no longer than cycle time \( T_c \). Therefore the wire delay can be written as:

\[
D_{\text{wire}} = T_{\text{flight}} + T_c
\]
For a given wire dimension and an operating frequency, the average total power of the wire with RC equalizer and termination resistance $R_l$ is a function of variables $R_s$, $R_d$, $C_d$ and $R_l$. We model the relationship using the following polynomial function:

$$P_{wire} = \sum_{k=1}^{P} a_k R_s^i R_d^j C_d^m R_l^n, \quad i + j + m + n \leq N$$

(7)

The number of terms $P$ is determined by the order $N$. We run circuit simulations to collect the power values for different variables combinations, and use min-square-error curve fitting method to find the coefficients $a_k$. We found that when $N = 4$, the error is less than ±6%.

### B. Transceiver design and modeling

The transceiver stage consists of a sense amplifier (SA) and differential inverter chain, as shown in Fig. 1(b). For the SA, we adopt a state-of-the-art configuration called double-tail latch-type scheme based on [18]. This scheme could provide more flexibility for designer to balance the trade-offs of performance metrics, which is suitable for on-chip interconnect application. In this design, we tune the size of transistors to minimize the SA delay for given technologies.

For the inverter chain, size of the last inverter is computed according to the requirement of inverter chain output resistance $R_s$. During the design, we fix the number of inverter stage to 6, and sweep the size of first inverter to achieve the optimal total delay of transceiver stage. Simulation results show that, while the SA input voltage $\Delta V_{in} = 50 \text{ mV}$ and $R_s = 50 \Omega$, at 90 nm node, the optimal delay of transceiver stage is 87.08 ps and the power consumption is 518 $\mu$W, whereas the delay and power consumption will decrease to 5.61 ps and 44.8 $\mu$W at 22 nm node as the technology scales. Our simulation results also indicate, the power consumed on the wire between transceivers (including $R_d$ and $R_l$) is dominant in the total power consumption, so we optimize the transceiver stage in terms of minimized delay, and use this design in the overall optimization of whole scheme.

The transceiver stage at driver side could be modeled as a voltage source $V_i$ with an output resistance $R_s$, where $V_i$ is a full swing pulse signal with rise time equal to 10% of the cycle time. The delay and power consumption of transceiver stage at receiver side is modeled using non-linear fitting method. We extract the delay and power data from SPICE simulation results to build a look-up table with index of $\Delta V_{in}$ and $R_s$, and fit the data using non-linear functions as follows:

$$\text{delay}(\Delta V_{in}, R_s) = a_1 \Delta V_{in} a_2 + a_3 R_s a_4 + a_5$$

$$\text{power}(\Delta V_{in}, R_s) = \frac{b_1 + b_2 \Delta V_{in} + b_3 R_s}{1 + b_4 \Delta V_{in} + b_5 R_s} + b_6$$

(8)

(9)

where $a_i (i = 1 \sim 5), b_i (i = 1 \sim 6)$ are the fitting coefficients. The relative error of this fitting model is within ±2% and ±5% for delay and power, respectively.

### III. Problem Formulation and Optimization Flow

We formulate the optimization problem as a constrained non-linear programming problem, and adopt Sequential Quadratic Programming (SQP) method ([16]) to solve it. The design goals include minimized latency, minimized latency-power product and minimized latency$^2$-power product, which are referred as min-$d$, min-$dp$ and min-$d^2p$ respectively. The optimization variables are $R_s, R_d, C_d$ and $R_l$.

For a given technology node and a given wire dimension, this formulation can be written as:

$$\min f = f_0 + a(V_0 - V_{eye})$$

$$s.t. \quad R_s^{min} \leq R_s \leq R_s^{max}$$

$$R_d^{min} \leq R_d \leq R_d^{max}$$

$$C_d^{min} \leq C_d \leq C_d^{max}$$

$$R_l^{min} \leq R_l \leq R_l^{max}$$

(10)

(11)

where $f_0$ is the design objects that we want to minimize, including latency, latency-power product and latency$^2$-power product. $a$ and $V_0$ are constants. We add the exponential term to handle the constraint on the eye opening of far end of T-Line. When the eye opening $V_{eye}$ is smaller than $V_0$, the exponential term dominates and force the flow to find a larger $V_{eye}$, otherwise the $f_0$ term dominates and the design goal will be minimized.

The overall optimization flow is shown in Fig. 3. It follows the idea presented in [12], which models the wire and transceiver separately. However the signaling scheme, optimization method and modeling approaches are different. The flow inputs are technology node and wire dimensions. Based on these design parameters, we then build the wire model and transceiver stage model, respectively. For the wire model, we employ the 2D field solver to generate RLGC tabular model, which could be simulated in SPICE. For the transceiver model, we do the optimization and fit the simulation data into non-linear functions as mentioned in...
Section II-B.

In each iteration of optimization, we utilize the model derived in Section II-A to get the delay/power of the wire. Meanwhile, we simulate the step response of T-Line for given design variables, and then use [19] to estimate the eye opening, which corresponds to the $\Delta V_{in}$ of the following transceiver stage. With the $\Delta V_{in}$ and $R_s$, delay and power of transceiver stage are given using non-linear formulas defined in (8) and (9). According to the different design goals, the cost function is evaluated by combining the delay and power of both wires and transceivers, which is utilized by the SQP routine to do optimization. The flow finally outputs optimal values of design variables in terms of optimal design goal, and also provides the performance metrics, including delay/power/throughput at this optimal situation.

IV. EXPERIMENTAL RESULTS

We optimize the proposed signaling scheme under three different design goals with four different spacings. The performance scaling of on-chip T-Line is compared with the repeated RC wires, which are also optimized as in [3] for the three design goals.

A. Experiment settings

For RC wire, data in Table I is used to calculate the necessary parameters. The method proposed in [3] is implemented in MATLAB to optimize the repeated RC wire for technologies from 90 nm to 22 nm. The optimal repeated wires are verified with HSPICE simulation using the predictive transistor model [20].

For T-Line, we use the 2D EM solver CZ2D in EIP tool suite from IBM [21] to extract the frequency dependent RLGC tabular model. We use PowerSPICE [22] to simulate the transmission lines and HSPICE with predictive transistor model [20] to simulate the transceiver delay and power. Regression method in MATLAB is adopted to model the T-Line power and the transceiver delay and power. The optimization flow is implemented in MATLAB.

B. Metrics definitions

We compare the latency, power consumption and throughput of T-Line and repeated RC wires. We use the wire length normalized delay to define the latency because we are investigating the scalability of these two wires across different technologies and we want the latency be independent of wire length. The normalized delay is written as:

$$delay_n = \frac{\text{propagation delay}}{\text{wire length}}$$ (12)

The propagation delay includes the wire delay and gate delay. The gate delay refers to repeater delay for RC wire or transceiver delay for T-Line.

To demonstrate the scaling trend of power consumption as technology shrinks, we use the normalized energy per bit as follows:

$$power_n = \frac{\text{energy per bit}}{\text{wire length}} = \frac{\text{power}}{\text{frequency} \times \text{wire length}}$$ (13)

C. Comparison of repeated RC wires and transmission lines

The frequency of RC wire is determined by the propagation delay since one bit is transmitted only after the previous bit reaches destination. For T-Line, the operating frequency is determined by the bandwidth of transceiver as shown in Table I.

The throughput of T-Line is defined as

$$throughput_n = \frac{\text{frequency}}{\text{wire pitch}}$$ (14)

which reflects the amount of data can be transmitted for a given cross area and a given period. We assume the RC wire is without pipelining, and the throughput is defined as

$$throughput_n = \frac{1}{\text{delay}_n \times 5\text{mm} \times \text{wire pitch}}$$ (15)

Fig. 4. Normalized delay comparison between repeated RC wire and proposed on-chip T-line

Fig. 5. Normalized energy consumption between repeated RC wire and proposed on-chip T-line
we also show the performance of min-d at minimum pitch, which is labeled as min-pitch.

Fig. 4 shows that design goal has very little impact on T-Line latencies. The latency of T-Line has three components: time of flight $T_{flight}$ (only depends on technology since the wire length is fixed to 5 $\mu$m), cycle time $T_c$ and transceiver latency. As technology advances, $T_{flight}$ reduces as dielectric constant gets smaller; $T_c$ decreases as a result of faster SA switching speed. The transceiver latency varies with design goal, because different designs choose optimal $R_s$, which determines the last inverter size of the inverter chain, and therefore affects the sizing of the whole inverter chain. Generally speaking, the variations of transceiver latency (Table. II) are insignificant comparing to total latency. For example, at 90 $nm$ technology, the time of flight for the 5 $\mu$m wire is around 30 ps, and the cycle time is 150 ps. For different design goals, transceiver latency ranges from 70 to 80 ps, and the total latency varies from 250 to 260 ps. The corresponding delay $\alpha$ changes from 50 to 52 ps/$\mu$m. Table. II also shows that as technology scales, transceiver latency improves because of faster device switching. Consequently, the latency of T-Line shows decreasing trend with technology.

The latency of repeated RC wire (Fig. 4) is strongly affected by design goals, and increases as technology shrinks. The design goal of minimizing delay chooses wire pitch and width that are much larger than minimum pitch and width, therefore the wire resistance, coupling capacitance and delay are greatly reduced. As technology advances, wire pitch decreases but aspect ratio grows, and wire becomes more resistive and more heavily coupled, which results in larger latency.

Comparing the latency of RC wire and T-Line under three design goals, T-Line has larger latency only at 90 $nm$ (roughly 50 ps/$\mu$m versus 35 ps/$\mu$m, which is 1.5X). The newer the technology, the more advantageous the T-Line in terms of normalized latency.

As shown in Fig. 5, the energy per bit of T-Line is much lower than RC wire, and it decreases as technology advances, since frequency increases and the energy per bit is inverse proportional to the working frequency (as shown in (13)). The power consumption of transceiver is insignificant compared to the power consumed on the metal wires and resistors ($R_l$ and $R_d$), as illustrated in Table. III. As apposed to T-Line, the energy per bit of RC wire is strongly coupled with design goals and the value is much higher than T-Line. Under the min-dp design goal, the energy per bit of RC wire varies from twice (90 $nm$) to 4.3X (22 $nm$) of those of T-Line, and for 90 $nm$ technology, the energy per bit ranges from more than 400 $pJ/m$ for min-pitch to around 50 $pJ/m$ for min-dp.

Fig. 6 demonstrates the trend of the normalized throughput of T-Line and RC wire. At 90 $nm$ technology, RC wire has higher throughput than T-Line, while at 22 $nm$, T-Line with min-pitch enjoys the highest throughput of around 15 Gbps/$\mu$m. According to the definition in (15), the normalized throughput of T-Line relies on the cycle time and the wire pitch. As discussed above, $T_c$ is determined by technology regardless of the design goal, therefore, for the same technology, the normalized throughput only depends on the wire pitch. As we will show in Section IV-D, the wires with the largest spacing $S = 2H$ give all the optimal design goals, which result the three design goals have the same throughput in Fig. 6 and the min-pitch has higher throughput. The throughput improves by 6X from 90 $nm$ to 22 $nm$ due to the scaling of $T_c$, as shown in Fig. 6.

Since we assume the RC wire is without pipelining, the throughput of RC wire relies on the latency rather than operating frequency. Consequently the min-d and min-pitch RC wires have higher throughput than min-dp and min-ddp design goals. Under the design goals of min-dp and min-ddp, the throughput of RC wires decreases from 90 $nm$ to 45 $nm$ technology due to the usage of smaller repeaters with larger intervals and smaller wire width to reduce the power, which increase the latency. Before 45 $nm$ technology, RC wire with min-dp and min-ddp have substantial advantage comparing to T-Line with min-d, min-dp and min-ddp. After 45 $nm$ technology, the throughput of the two schemes are very close, and at 22 $nm$, the throughput of T-Line is even
Fig. 7. Performance comparison of different wire spacing

higher than RC wire, which means that as technology scales, T-Line becomes more appealing in terms of throughput.

D. Performance comparison of different wire spacing

The effects of wire spacing upon different design goals are shown in Fig. 7 using 45 nm technology as an example. The left y axis shows the performance degradation, defined as the minimum design goals at $S = 2H$ divided by the design goals at different spacing. The right y axis is the normalized throughput as defined in (15). For all three design goals, the optimums are achieved when $S = 2H$. The reason is larger spacing gives higher $Z_0$, which reduces the attenuation along the wire. As a result, to have the same eye opening as wires with larger attenuation, larger $R_s$ can be used in this situation. Larger $R_s$ reduces transceiver delay because it reduces the size of inverter chain, and it also produces lower power consumption on both transceiver and wires. However, large spacing gives smallest throughput and there are tradeoffs between different design goals and the desired throughput.

V. CONCLUSIONS

In this paper, a new signaling scheme using on-chip lossy T-Line for global interconnects is proposed. The modeling and optimization of the proposed scheme are discussed. An SQP based method is adopted to find the optimal design variables under three different design goals across five technologies. The experimental results demonstrate that at 22 nm, the new scheme can improve the delay by 80%-95% and the normalized energy consumption by 50%-94% comparing to the repeated RC wire. At 22 nm, the normalized latency is 10 ps/mm, the energy per bit is 20 pJ/μm, and the throughput is 15 Gbps/μm.

VI. ACKNOWLEDGMENT

The authors would like to acknowledge the support of the NSF CCF-0811794 and California MICRO Program.

REFERENCES