Timing Optimization in Logic with Interconnect

Arkadiy Morgenshtein
arkadiy@tx.technion.ac.il

Eby G. Friedman
friedman@ece.rochester.edu

Ran Ginosar
ran@ee.technion.ac.il

Avinoam Kolodny
kolodny@ee.technion.ac.il

VLSI Systems Research Center, Electrical Engineering Department
Technion – Israel Institute of Technology, Haifa, Israel

ABSTRACT

Timing optimization in logic paths with wires has become an important issue in the VLSI circuit design process. Existing techniques for minimizing delay treat only the relatively rare cases of logic without wires (logical effort) or logic with a long resistive wire (repeater insertion). The techniques described in this paper address the fundamental questions of optimal sizing, the number and location of the gates. The Unified Logical Effort (ULE) method supports fast and precise optimal sizing of gates in the presence of interconnect based on intuitive closed-form expressions. The optimal number of repeaters is determined by the Gate-terminated Sized Repeater Insertion (GSRI) technique, resulting in lower delay as compared to standard repeater insertion methodologies. The Logic Gates as Repeaters (LGR) method is used for optimal wire segmenting and gate location, suggesting a distribution of logic gates over interconnect rather than using logically-redundant repeaters. The combination of these techniques provides solution for a wide variety of design issues.

Categories and Subject Descriptors: B.8.2 [Performance and Reliability]: Performance Analysis and Design Aids

General Terms: Performance, Design

1. INTRODUCTION

The general timing optimization problem can be defined as reducing the delay of a logic path propagating over a distance from point *A* to point *B* while performing a logical function *F* (see Figure 1a). Existing timing optimization techniques address the following cases: *(i)* Circuits where the output wire is absent or relatively short (Figure 1b) use the Logical Effort method [1][2] that incorporates gate sizing and buffer addition; *(ii)* Circuits where the output drives a high impedance wire (Figure 1c) use the repeater insertion method [7][8] that is based on interconnect segmentation by optimally scaled inverters. Extensive research has focused on improving the precision and power efficiency of Logical Effort [4]–[6] and Repeater Insertion [9]–[13] methods.

The particular cases treated by the existing techniques are relatively rare in modern circuits. The general timing optimization problem is based on a practical model, which includes the wires between the gates of function *F* (Figure 1d).

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, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SLIP'08, April 5–6, 2008, Newcastle, United Kingdom.
Copyright 2008 ACM 978-1-59593-918-0/08/04...$5.00.

Three interrelated fundamental questions are:

1. What is the optimal size of the gates?
2. What is the optimal number of gates/repeaters?
3. Where should the gates be located along the wire?

A unified timing optimization approach that solves these general design problems, and converges to the aforementioned existing techniques is described in this paper. The techniques described in this paper address the fundamental questions of timing optimization for any practical circuit structure. The proposed techniques can be combined to provide the best solution for a wide variety of design objectives.

The paper is composed of the following sections. The Unified Logical Effort (ULE) method is presented in Section 2 for optimal sizing of gates in the presence of interconnect. The question of optimal number of repeaters is analyzed in Section 3, where the Gate-terminated Sized Repeater Insertion (GSRI) technique is described. In Section 4, an approach for optimal wire segmenting and gate location is described based on the Logic Gates as Repeaters (LGR) method. The proposed techniques are accompanied by examples and a discussion of power-efficient applications. Finally, a summary of the paper is provided in Section 5.

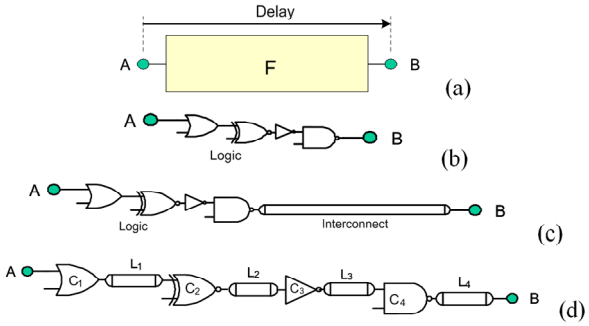

Figure 1. Classification of circuit configurations in timing optimization: (a) general timing problem, (b) logic with short wires, treated by Logical Effort, (c) logic with a long wire at the output, treated by Repeater Insertion, (d) general case including significant wire delays between the gates.

2. UNIFIED LOGICAL EFFORT (ULE)

The first fundamental question of timing optimization regards gate sizing. In current technologies the delays caused by wires and gates along a logic path are tightly coupled and cannot be treated separately. Wire delays are not correlated with the delay of the driver gates; thereby the standard Logical Effort (LE) model cannot be used. Furthermore, optimal gate sizing in the presence of interconnect does not correspond to equal effort of all of the

stages along a path (as in standard LE) [1][2]. The Unified Logical Effort (ULE) method addresses delay minimization in logic paths with general gates and *RC* wires.

2.1. Delay Model of Logic Gates with Wires

The logical effort model is modified here to include the interconnect delay. This change is achieved by extending the logical effort delay to include the wire delay, establishing a *Unified Logical Effort* (ULE) model.

A circuit composed of logic gates with wires is shown in Figure 2. The interconnect is represented by a π -model. The Elmore delay model [14] is used to describe the wire delay. The total combined delay expression is

$$D_i = R_i \cdot \left(C_{p_i} + C_{w_i} + C_{i+1} \right) + R_{w_i} \cdot \left(0.5 \cdot C_{w_i} + C_{i+1} \right), \quad (1)$$

where R_i is the effective output resistance of the gate i, C_{p_i} is the parasitic output capacitance of gate i, C_{w_i} and R_{w_i} are, respectively, the wire capacitance and resistance of segment i, and C_{i+1} is the input capacitance of gate $i+1$.

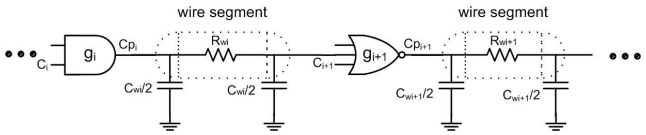

Figure 2. Cascaded logic gates with RC interconnect.

This expression is rewritten by introducing the delay of a minimum size inverter as a technology constant $\tau = R_0 \cdot C_0$, where R_0 and C_0 are the output resistance and input capacitance of a minimum sized inverter, respectively,

$$D_i = \tau \cdot d_i$$
$$= \tau \cdot \left[\frac{R_i}{R_0} \cdot \frac{\left(C_{w_i} + C_{i+1} + C_{p_i} \right)}{C_0} + \frac{R_{w_i}}{R_0 \cdot C_0} \cdot \left(0.5 \cdot C_{w_i} + C_{i+1} \right) \right]. \quad (2)$$

The stage delay, normalized with respect to a minimum inverter delay τ, is expressed using logical effort (LE) terms,

$$d_i = g_i \cdot \left(h_i + \frac{C_{w_i}}{C_i} \right) + \frac{R_{w_i} \cdot \left(0.5 \cdot C_{w_i} + C_{i+1} \right)}{\tau} + p_i, \quad (3)$$

where $g_i = \left(R_i \cdot C_i \right) / \left(R_0 \cdot C_0 \right)$ is the logical effort related to the gate topology, $h_i = C_{i+1} / C_i$ is the electrical effort describing the driving capability, and $p_i = \left(R_i \cdot C_{p_i} \right) / \left(R_0 \cdot C_0 \right)$ is the delay factor of the parasitic impedance. The capacitance and resistance of the gate are related to the scaling factor x_i as $C_i = C_0 \cdot g_i \cdot x_i$ and $R_i = R_0 / x_i$, respectively.

The *capacitive interconnect effort* h_w and the *resistive interconnect effort* p_w are, respectively,

$$h_{w_i} = \frac{C_{w_i}}{C_i}, \quad (4)$$

$$p_{w_i} = \frac{R_{w_i} \cdot \left(0.5 \cdot C_{w_i} + C_{i+1} \right)}{\tau}. \quad (5)$$

As shown in (4), h_w expresses the influence of the wire capacitance on the electrical effort of the gate. The component p_w in (5) is the delay of the loaded wire in terms of the gate delay τ.

The final expression of the ULE delay for a single stage is

$$d = g \cdot \left(h + h_w \right) + \left(p + p_w \right). \quad (6)$$

The ULE delay expression for an *N* stage logic path with wires is

$$d = \sum_{i=1}^{N} g_i \cdot \left(h_i + h_{w_i} \right) + \left(p_i + p_{w_i} \right). \quad (7)$$

Note that in the case of short wires, the resistance R_w of the wire may be neglected, eliminating p_w and only leaving the capacitive interconnect effort h_w in the expression. The extended delay expression reduces to the standard LE delay equation when no significant interconnect impedance exist along the logic path.

2.2. Delay Minimization Using Unified Logical Effort

As the first step in path delay optimization, a two-stage portion of a logic path with wires (as shown in Figure 2) is considered. In this case, the ULE expression of the total delay is

$$d = g_i \left(h_i + h_{w_i} \right) + \left(p_i + p_{w_i} \right) + g_{i+1} \left(h_{i+1} + h_{w_{i+1}} \right) + \left(p_{i+1} + p_{w_{i+1}} \right), \quad (8)$$

where the electrical effort of each stage is $h_i = C_{i+1} / C_i$ and $h_{i+1} = C_{i+2} / C_{i+1}$. Substituting $C_{i+1} = h_i \cdot C_i$ into (8) in the presence of resistive interconnect, the delay can be expressed in terms of h_i as

$$d = g_i \cdot \left(h_i + \frac{C_{w_i}}{C_i} \right) + p_i + \frac{R_{w_i} \cdot \left(0.5 \cdot C_{w_i} + h_i \cdot C_i \right)}{R_0 \cdot C_0} +$$
$$+ g_{i+1} \cdot \left(\frac{C_{i+2} + C_{w_{i+1}}}{h_i \cdot C_i} \right) + p_{i+1} + p_{w_{i+1}}. \quad (9)$$

The condition for optimal gate sizing is determined by equating the derivative of the delay with respect to the gate size to zero (see [3] for derivation details),

$$\left(g_i + \frac{R_{w_i} \cdot C_i}{R_0 \cdot C_0} \right) \cdot h_i = g_{i+1} \cdot \left(h_{i+1} + h_{w_{i+1}} \right). \quad (10)$$

To provide an intuitive interpretation of the expression, it can be rewritten by multiplying by $R_0 \cdot C_0$ and using the relationships $h_i = C_{i+1} / C_i$, $C_i = C_0 \cdot g_i \cdot x_i$ and $R_i = R_0 / x_i$. The resulting optimum condition is

$$\left(R_i + R_{w_i} \right) \cdot C_{i+1} = R_{i+1} \cdot \left(C_{i+2} + C_{w_{i+1}} \right). \quad (11)$$

The meaning of (11) is that the optimum size of gate $i+1$ is achieved when the delay component $\left(R_i + R_{w_i} \right) \cdot C_{i+1}$ due to the gate capacitance is equal to the delay component $R_{i+1} \cdot \left(C_{i+2} + C_{w_{i+1}} \right)$ due to the effective resistance of the gate.

A schematic model describing the related delay components is shown in Figure 5. Note that other delay components ($R_i \cdot C_{w_i}$, $0.5 \cdot R_{w_i} \cdot C_{w_i}$, and $R_{w_{i+1}} \cdot \left(0.5 \cdot C_{w_{i+1}} + C_{i+2}\right)$) are independent of the size of gate $i+1$ and do not influence the optimum size. Also note that in the presence of wires, the condition for minimum path delay does not correspond to equal delay or to equal effort at every stage along the path.

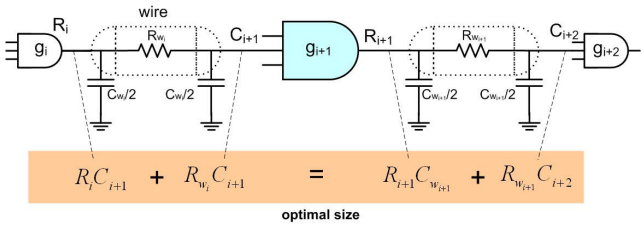

Figure 3. Delay components in ULE characterization

The intuitive optimum condition (11) can be further developed for any gate i based on the characteristic that the total delay D_i is comprised of the sum of the upstream delay D_{R_i} and the downstream delay D_{C_i} ,

$$\begin{aligned} D_{C_i} &= \left(R_{i-1} + R_{w_{i-1}}\right) \cdot C_i = \left(R_{i-1} + R_{w_{i-1}}\right) \cdot C_0 \cdot g_i \cdot x_i \ , \\ D_{R_i} &= R_i \cdot \left(C_{i+1} + C_{w_i}\right) = \frac{R_0}{x_i} \cdot \left(C_{i+1} + C_{w_i}\right) \quad , \\ D_i &= D_{C_i} + D_{R_i} + const \quad . \end{aligned} \tag{12}$$

When the total delay is minimum, the sum of the differential of the delay components with respect to the sizing factor x_i is equal to 0, leading to the expression for the optimal sizing factor $x_{i_{opt}}$,

$$x_{i_{opt}} = \sqrt{\frac{R_0}{\left(R_{i-1} + R_{w_{i-1}}\right)} \cdot \frac{\left(C_{i+1} + C_{w_i}\right)}{C_0 \cdot g_i}} \ . \tag{13}$$

When $x_{i_{opt}}$ is substituted into (11), a general optimum condition can be determined,

$$\begin{aligned} &\left(R_{i-1} + R_{w_{i-1}}\right) \cdot C_i = R_i \cdot \left(C_{i+1} + C_{w_i}\right) = \\ &= \sqrt{\left[\left(R_{i-1} + R_{w_{i-1}}\right) \cdot C_0 \cdot g_i\right] \cdot \left[R_0 \cdot \left(C_{i+1} + C_{w_i}\right)\right]} \ . \end{aligned} \tag{14}$$

An intuitive interpretation of (14) is that the minimum delay is achieved when the downstream delay component (due to C_i) and the upstream delay component (due to R_i) of an optimally sized gate are both equal to the geometric mean of the upstream and downstream delays obtained if the gate (with logical effort g_i) is minimally sized,

$$D_{R_{i_{opt}}} = D_{C_{i_{opt}}} = D^* \equiv GM\left[D_{R_{i_{\min}}}, D_{C_{i_{\min}}}\right] . \tag{15}$$

For a logic path without wires ($h_{w_i} = 0$, $R_{w_i} = 0$), the optimum condition of ULE (10) converges to the optimum of LE [1]: $g_i \cdot h_i = g_{i+1} \cdot h_{i+1}$.

The gate sizes based on ULE can be iteratively determined along the path while applying the optimum condition (13) to each capacitance along the path. An example of ULE optimization in the logic path is shown in Figure 4 where the ULE technique has been applied to a logic path consisting of nine identical stages. Parameters [15] for a 65 nm CMOS technology are used. The input capacitance of the first and last gates are $10 \cdot C_0$ and $100 \cdot C_0$, respectively. The size of the logic gates along the path is shown in Figure 4 for several values of wire length L between each stage. All of the solutions range between two limits (the bold lines in the plot): (a) for zero wire lengths, the solution converges to LE optimization [1], and (b) for long wires, the gate size in the middle stages of the path converges to a fixed value, $x_{opt} \cong 50$ (the dashed line), similar to repeater insertion methods [7],[13].

2.3. ULE Gate Sizing for Long Wires

As shown in Figure 4, in the case of long wire segments, the gate sizing optimization process converges to the scale factor x_{opt} . When long wires are assumed, the impedances C_{w_i} and $R_{w_{i-1}}$ of (13) become dominant as compared to the gate impedances. A schematic model of this case is shown in Figure 5.

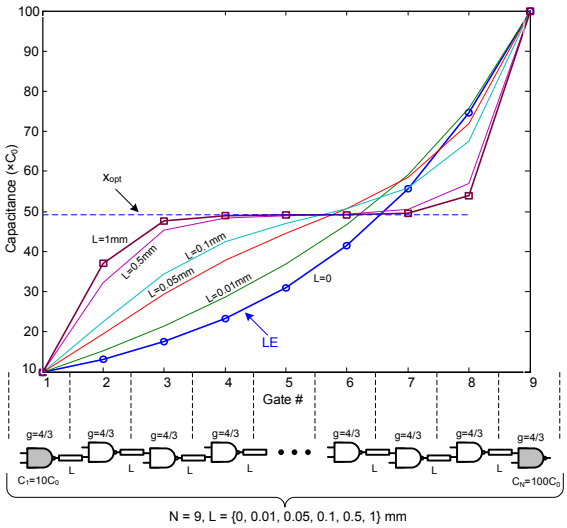

Figure 4. Optimization of ULE sizing (normalized with respect to C0) for a chain of nine NAND gates with equal wire segments for a variety of lengths.

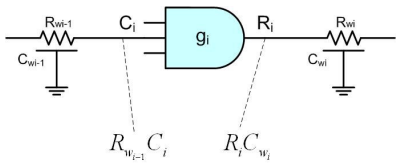

Figure 5. Delay components of optimum ULE for long wires

The scale factor of a general gate is determined from (13) for the case of long wires,

$$x_{opt_i} \cong \sqrt{\frac{R_0 \cdot C_{w_i}}{R_{w_{i-1}} \cdot C_0 \cdot g_i}} = \underbrace{\sqrt{\frac{c_w \cdot R_0}{r_w \cdot C_0 \cdot g_i}}}_{\text{constant}} \cdot \sqrt{\frac{L_i}{L_{i-1}}} \quad , \tag{16}$$

using the relationships, $C_{w_i} = c_w \cdot L_i$, and $R_{w_i} = r_w \cdot L_i$, where r_w and c_w are the resistance and capacitance of the wire per unit length, respectively, and L_i and L_{i-1} are the length of the wires before and after the logic gate g_i, respectively. Note that the scale factor of the gate in the case of long wires only depends upon the ratio of the adjacent wire lengths.

A general optimum condition is determined, similar to (14),

$$R_{w_{i-1}} \cdot C_i = R_i \cdot C_{w_i} = \sqrt{\left[R_{w_{i-1}} \cdot C_0 \cdot g_i\right] \cdot \left[R_0 \cdot C_{w_i}\right]} \,. \tag{17}$$

In the special case of equal wire segments, the capacitance and resistance of all the segments are equal to C_w and R_w, respectively. In this case, the scaling factor x_{opt} is independent of the wire length since the component C_w / R_w is independent of the wire length. The optimum condition can be rewritten as a function of the capacitance and resistance per unit length c_w and r_w,

$$x_{opt_i} = \sqrt{\frac{R_0 \cdot c_w}{r_w \cdot C_0 \cdot g_i}} \,. \tag{18}$$

For the special case of inverter-based repeater insertion (with an electrical effort $g = 1$), the condition of (18) reduces to optimal repeater scaling, as described by Bakoglu in [7]. The best sizing of a repeater is achieved when the delay component $R_w \cdot C_{rep}$ due to the repeater capacitance is equal to the delay component $R_{rep} \cdot C_w$ due to the effective resistance of the repeater.

The application of ULE to repeater insertion provides a solution to some specific design problems. Two examples are presented here:

- *Layout constraint:* given a wire of total length L comprising two segments of lengths L_1 and L_2, the optimal size of the repeater located between the segments is

$$x_{rep_{opt}} = \sqrt{\frac{c_w \cdot R_0}{r_w \cdot C_0 \cdot g_i}} \cdot \sqrt{\frac{L_2}{L_1}} \,. \tag{19}$$

- *Cell size constraint:* given a repeater of size x_{rep} dividing a wire of total length L into two segments, a ratio of the optimal segment lengths $L_{1_{opt}}$ and $L_{2_{opt}} = L - L_{1_{opt}}$ is

$$\frac{L_{2_{opt}}}{L_{1_{opt}}} = x_{rep}^2 \Big/ \left(\frac{c_w \cdot R_0}{r_w \cdot C_0 \cdot g_i}\right). \tag{20}$$

ULE optimization has been verified by comparison to the results of a commercial numerical optimizer which uses a circuit simulator to estimate the delay [3]. The Cadence Virtuoso® Analog Optimizer [16] is used as the reference tool. The delay after ULE optimization is within 9% of the Analog Optimizer tool. The low complexity and fast run time of ULE makes the algorithm a competitive alternative for integration into EDA toolsets that optimize complex logic structures with interconnect. The run time of ULE is orders of magnitude shorter than the run time of Analog Optimizer.

2.4. ULE Gate Sizing for Power-Delay Product Minimization

Sizing gates for minimum delay can result in a large size dissipating significant power. A power-delay product as the minimization goal results in a smaller gate size while trading off delay and power.

The delay of a two stage chain (see Figure 2) is described in (9) and is a function of h_i. The dynamic power is represented by the capacitance of the gate $i+1$ and the wire capacitance,

$$P \propto \left(C_{i+1} + C_{w_{i+1}}\right) = C_i \cdot h_i + C_{w_{i+1}} \,. \tag{21}$$

The optimal condition is determined by setting the derivative of the power-delay product to zero, resulting in the following expression (see [3] for the complete derivation) for the optimal input capacitance C_i,

$$\begin{aligned}
&C_i^3 \cdot a_1 + C_i^2 \cdot a_2 + C_i \cdot a_3 + a_4 = 0 \,, \\
&a_1 = 2\left(g_{i-1} + \frac{R_{w_{i-1}} \cdot C_{i-1}}{\tau}\right) \\
&a_2 = \left(\begin{array}{l} g_{i-1} \cdot \left(C_{w_{i-1}} + C_{w_i}\right) + \\ + \dfrac{R_{w_{i-1}} \cdot C_{i-1} \cdot \left(0.5 \cdot C_{w_{i-1}} + C_{w_i}\right)}{\tau} + p_{w_i} \cdot C_{i-1} \end{array}\right) \\
&a_3 = 0 \\
&a_4 = -\left(g_i \cdot C_{w_i} \cdot C_{i-1} \cdot \left(C_{i+1} + C_{w_i}\right)\right) \,.
\end{aligned} \tag{22}$$

The polynomial has a single positive real root. The optimization can be performed iteratively, similarly to the ULE delay minimization technique.

3. GATE-TERMINATED SIZED REPEATER INSERTION (GSRI)

The second fundamental question of timing optimization addresses the optimal number of gates. This problem is particularly important in the case of long wires, where sizing gates along a logic path does not sufficiently reduce the delay. In such cases, repeater insertion is used to minimize the delay.

Standard repeater insertion methodologies (herein named RI) include several assumptions that lead to elegant expressions for the optimal number and size of the repeaters [7]. The following assumptions are made:

1. The gates at the wire edges are similar to repeaters.
2. The size of the repeaters is constant and depends only on process technology parameters.
3. The size of the repeaters is equal.

These assumptions may be unjustified. The wires are usually located between logic gates that are different in type and size from repeaters. Moreover, different repeater sizes may be chosen to maintain specific design rules, or to target power efficiency. To address these issues, a Gate-Terminated Sized Repeater Insertion (GSRI) technique is developed here for timing optimization under realistic circuit constraints.

3.1. Delay Model of Logic Path with Repeaters

The general case of repeater insertion in a wire between two logic gates is illustrated in Figure 6. In this case, uniformly distributed equally sized repeaters are assumed.

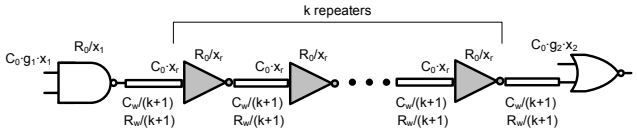

Figure 6. Logic path with wire segmented by repeaters.

Note that in this case the number of repeaters is k, while the number of wire segments after repeater insertion is $k+1$. This case is unlike [7], where both values are k, since the first gate is also a repeater. The size of the gates is represented by x_1, x_2 for the logic gates at the edges, and x_r for the repeaters.

The total delay of the scheme is:

$$D = \underbrace{0.7\cdot\frac{R_0}{x_1}\cdot\left(\frac{C_w}{k+1}+C_0\cdot x_r\right)}_{\text{first gate driving wire segment and repeater}} + \underbrace{(k-1)\cdot\frac{R_0}{x_r}\cdot\left(0.4\cdot\frac{C_w}{k+1}+0.7\cdot C_0\cdot x_r\right)}_{\text{k-1 repeaters driving wire segment and following repeater}} + \underbrace{\frac{R_0}{x_r}\cdot\left(0.4\cdot\frac{C_w}{k+1}+0.7\cdot C_0\cdot g_2\cdot x_2\right)}_{\text{last repeater driving wire segment and last gate}} + \underbrace{k\cdot\frac{R_w}{k+1}\cdot\left(0.4\cdot\frac{C_w}{k+1}+0.7\cdot C_0\cdot x_r\right)}_{\text{resistance of k wires driving wire capacitance and repeater}} + \underbrace{\frac{R_w}{k+1}\cdot 0.7\cdot C_0\cdot g_2\cdot x_2}_{\text{resistance of last wire-gate pair}} \quad (23)$$

The delay expression contains factors of 0.7 and 0.4 for lumped and distributed devices, respectively (similarly to [7]).

3.2. Delay Minimization Using GSRI

The optimal number of repeaters is determined by setting the differential of (23) as a function of k to zero and performing a substitution $\mathrm{K} = k+1$, which leads to

$$\mathrm{K}^3\cdot a_1 + \mathrm{K}^2\cdot a_2 + \mathrm{K}\cdot a_3 + a_4 = 0\,, \quad (24)$$

where the coefficients are

$$\begin{aligned} a_1 &= 0.7\cdot C_0\cdot R_0 \\ a_2 &= 0 \\ a_3 &= -R_0C_w\left(\frac{0.7}{x_1}-\frac{0.4}{x_r}\right)+0.7R_wC_0\cdot\left(x_r-g_2\cdot x_2\right)-0.4R_wC_w \\ a_4 &= 2\cdot 0.4\cdot R_w\cdot C_w \end{aligned} \quad (25)$$

The optimal solution can be obtained by choosing the minimal real root greater than one. If no such roots exist (e.g. when real roots are negative or smaller than one), no repeater insertion is performed. After the optimal solution of (25) is determined, the number of repeaters is found from $k = \mathrm{K}-1$. Since the number of repeaters is an integer, the value of k is usually rounded.

The optimal number of repeaters determined from (25) is dependant on the size of the first and last gates, as well as the size of the repeaters. This behavior is different from [7] and reduces the delay. Note that the optimum number of repeaters from (25) converges to the expression in [7] in those cases where the basic repeater insertion assumptions are maintained (long wires, or wires with gates similar to repeaters).

3.3. GSRI Examples

Repeaters insertion is performed on a critical path of an ALU circuit containing the following gates, INV(×10), NAND3, INV, NOR2, NAND4, and INV(×10). Parameters [15] for a 65 nm CMOS technology are used. Equal wire segments are located between each pair of gates. The path is optimized using ULE prior to inserting repeater for various wire lengths.

In both techniques, the repeater size is ×258, according to the optimal sizing factor of RI [7]. RI and GSRI methods produce a different number of repeaters. The number of repeaters in GSRI is not equal for each wire (although the wires are the same). This behavior is due to the difference that exists in the gates between the wires. For gates with higher electrical effort (smaller gate driving a larger gate), the number of repeaters will be higher. RI optimization is effective only for wires longer than 2 mm, while GSRI allows optimization of shorter wires.

A comparison of the resulting delay is presented in Figure 7. The circuit is initially optimized using ULE sizing of the gates without repeater insertion. RI and GSRI techniques are then applied. GSRI produces up to a 25% delay reduction as compared to RI. Note that the increase in the delay in 0.5 mm wires by GSRI is a result of the quantization of the number of the repeaters and the large uniform repeaters driven by a small first gate. As shown later, the delay can be further reduced by ULE size optimization of the repeaters.

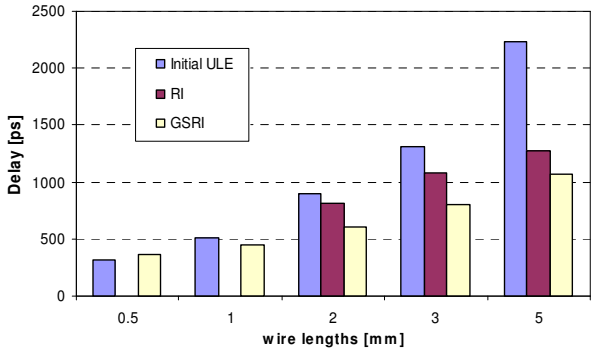

Figure 7. Comparison of resulting delay after using ULE sizing of gates, RI and GSRI, as a function of wire lengths.

The GSRI technique can also be successfully applied in those cases where the circuit requires uniform repeaters with different sizes than RI (usually smaller). The number of repeaters inserted for each size, as well as the resulting delay and power as compared to RI using standard sizes in the case of 3mm wires is listed in Table 1.

Note that the number of repeaters increases as the sizes decrease. The change in the number of the repeaters for ×200 and ×150 sizes is insignificant due to quantization. As can be seen, the delay penalty for using smaller repeaters is relatively low for sizes down to ×100. Using repeaters with a size of ×100 still provides smaller delay than RI.

Smaller sizes dictate a higher number of repeaters, keeping the total area almost unchanged. The power consumption of the path with repeaters, however, may decrease while using a higher

number of smaller repeaters. Note that for ×100 repeaters, there is a delay reduction of 17% and power reduction of 15% as compared to RI. This effect can be explained by the reduction in the short-circuit power of the repeaters, as the size of the repeaters are reduced and the number increased [11][21][22]. This power reduction is achieved due to the reduced transition times of the signals.

Table 1. Results for repeaters with sizes different than RI

	RI	*GSRI*					
repeater sizes	*258*	*258*	*200*	*150*	*100*	*50*	*20*
# of repeaters	*10*	*17*	*20*	*20*	*24*	*30*	*44*
delay [ps]	*1084*	*804*	*804*	*813*	*896*	*1322*	*2673*
energy [pJ]	*18.0*	*21.4*	*19.5*	*16.9*	*15.3*	*13.8*	*14.1*

3.4. Non-uniform Repeater Sizing by ULE

The delay of the path can be further decreased by the ULE sizing of the repeaters after GSRI. There are two alternatives for ULE sizing:

– *Sizing of the repeaters* without sizing the gates. This alternative is most suitable for circuits with fixed logic gates, as well as for power-efficient circuits.

– *Sizing of the entire path*, including the gates and the repeaters. This alternative provides the lowest possible delay.

The two alternatives are compared in Table 2. The delay and power of GSRI with uniformly sized repeaters vs. the ULE sizing in the case of 1mm wires is shown.

Table 2. Delay minimization in GSRI using ULE sizing

	GSRI, uniform repeaters	*ULE sizing of repeaters*	*ULE sizing of all gates*
delay [ps]	*447*	*368*	*345*
energy [pJ]	*9.9*	*6.8*	*8.4*

Note that the ULE sizing of repeaters may provide lower delay and power consumption than inserting equally sized repeaters. Note that the ULE sizing of the path including the logic gates results in an additional reduction in delay at higher power. The resulting power consumption can be lower than equally sized repeaters.

4. LOGIC GATES AS REPEATERS (LGR)

The usage of repeaters implies a significant cost in power and area, without contributing to the logical computation performed by the circuit. A study in [18] claims that in the near future, up to 40% of chip area will be used by inverters operating as repeaters and buffers. The use of numerous logically-redundant repeaters (Figure 8b) seems to be a waste, because the logic gates themselves may function as repeaters due to their amplifying nature. The *Logic Gates as Repeaters (LGR)* concept was proposed in [13] suggesting a distribution of logic gates over interconnect, which allows driving the partitioned interconnect without adding inverters to serve as repeaters (Figure 8c).

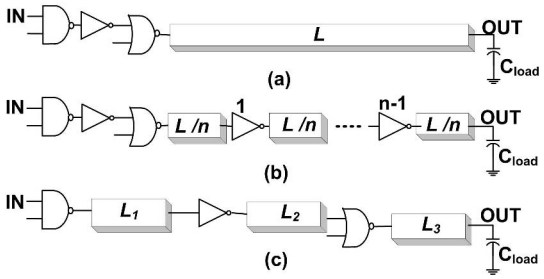

Figure 8. (a) A logic path driving a long interconnect wire. (b) Repeater insertion on the long interconnect (c) LGR optimization: the logic gates are distributed over the interconnect and serve as repeaters.

After the distribution of logic gates over interconnect is performed, each logic gate has a related interconnect segment, as presented in Figure 8c. After segmentation, the delay of each pair of logic-interconnect segment can be calculated separately.

The overall delay is the sum of delays of all the combined logic-interconnect segments,

$$D_{tot} = \sum_{i=1}^{N} \left[\begin{array}{l} \tau \left(g_i \left(\frac{C_{i+1} + L_i C_w}{C_i} \right) + p_i \right) + \\ + \left(0.5 L_i^2 R_w C_w + L_i R_w C_{i+1} \right) \end{array} \right] \quad (26)$$

where N is the number of gates and C_{N+1} is the load capacitance at the output of the circuit.

4.1. Optimization Methods

4.1.1. Optimal Segmenting

The total length of the interconnect along the logic path is denoted by L. The goal is to divide L into segments such that the delay expression (26) is minimized. The optimal length of each segment is derived by partial differentiation of the delay expression, performed for each of the segment lengths L_i.

There are two constraints on the goal function. The first constraint is

$$L_1 + L_2 + ... + L_n = L \quad . \quad (27)$$

Since the length of each segment must be non-negative due to its physical nature, the second constraint applied is

$$\forall i \quad L_i \geq 0 \, . \quad (28)$$

Applying differentiation on (26) with constraint (27), and equating to zero, the resulting optimal length of the i-th segment is

$$L_{i_{opt}} = \frac{L}{N} + \frac{L\left(R_{av} - R_i\right)}{R_w} + \frac{L\left(C_{av} - C_{i+1}\right)}{C_w} \, , \quad (29)$$

where the R_{av} and C_{av} are the average output resistance and input capacitance of the gates, respectively.

The first term represents equal partitioning of the total length, and the other terms represent corrections required because of different driving abilities and different input gate capacitances. If the driving gate is large (R_i is small), the segment to be driven will be increased. Similarly, when the driven gate is large (C_{i+1} is large), the segment should be decreased to reduce loading on the driving gate and wire segment. Note that in the case where all gates are of the same type and size, equal segmentation is obtained from (29).

4.1.2. Scaling and Segmenting

Additional speed-up may be obtained by enlarging each of the gates in the logic chain by a constant factor *s*. Uniform value of *s* is assumed for all the gates. The delay expression for a logic chain with gates enlarged by factor *s* is:

$$D_{tot} = \sum_{i=1}^{N} \left[\begin{array}{l} \tau\left(g_i \left(\frac{s \cdot C_{i+1} + L_i c_w}{s \cdot C_i} \right) + p_i \right) + \\ + \left(0.5 \cdot L_i^2 \cdot r_w \cdot c_w + L_i \cdot r_w \cdot s \cdot C_{i+1} \right) \end{array} \right]. \quad (30)$$

The optimal scaling factor *s* is obtained by differentiation of (30),

$$s = \sqrt{\left(\sum_{i=1}^{N} R_i \cdot C_{w_i} \right) \Big/ \left(\sum_{i=1}^{N} R_{w_i} \cdot C_{i+1} \right)}. \quad (31)$$

Note that in the special case where all gates are inverters and the interconnect is equally segmented (31) yields the scaling factor

$$s = \sqrt{(C_w R_0)/(C_0 R_w)}, \quad (32)$$

which is similar to the scaling factor presented by Bakoglu [7] in the context of optimally sized repeaters.

The optimal segment lengths and optimal scaling factor can be obtained by iterative calculation of (29) and (31). In experiments, convergence to within 1% of the optimal delay is reached in a few steps, usually less than three.

4.2. LGR Examples

LGR optimization is characterized and compared with Repeater Insertion. A circuit of a 8 to 256 decoder is analyzed. The symmetric structure of the decoder is suitable for LGR, since all the paths are simultaneously improved. The critical path of the decoder was optimized according to the LGR methodology. The results of segmenting optimization are listed in Table 3. The simple distribution of the critical path logic gates over the interconnect produces timing improvement of up to 27%.

The LGR segmenting and scaling results are compared with traditional repeater insertion and presented in Table 4. For intermediate lengths of interconnect the LGR produces 55% improvement over Repeater Insertion. For long interconnect, where a significant number of additional repeater stages are required, Repeater Insertion outperforms LGR by up to 70%. However, RI requires 44 additional functionally useless repeaters. Generally, in the case of a short logic chain, the LGR optimization technique is preferred for intermediate interconnect length. For long interconnect, where many repeaters are required, LGR can be combined with the addition of some repeater stages.

Table 3. 8-to-256 Decoder delay for segmenting

	Unoptimized	*LGR Segmenting*
Low-tier 1.5mm	*2.28 nsec*	*2.15 nsec*
Low-tier 15mm	*34.6 nsec*	*25.2 nsec*
High-tier 1.5mm	*3.62 nsec*	*3.47 nsec*
High-tier15mm	*36.4 nsec*	*34.9 nsec*

Table 4. 8-to-256 Decoder delay for segmenting and scaling

	LGR	*Repeater Insertion*
Low-tier 1.5mm	*0.188 nsec*	*0.268 nsec*
Low-tier 15mm	*5.45 nsec*	*1.65 nsec*
High-tier 1.5mm	*0.086 nsec*	*0.194 nsec*
High-tier15mm	*0.557 nsec*	*0.542 nsec*

4.3. Power Considerations in LGR

As a result of aggressive sizing, the circuit area and the power dissipated by up-scaled gates are considerably increased. Hence, in some cases, repeater insertion may be preferred over LGR for power and area considerations, because an inverter consumes the smallest possible area in comparison with other gates having the same current drive capability. Here, an analytic comparison between the LGR and repeater insertion is presented for dynamic power considerations, assuming that similar path delay is obtained by both techniques.

The dynamic power is related to the total capacitance of the system. Hence, the total capacitance of the LGR method and the traditional repeater insertion technique provides an estimate of the power dissipation. The total capacitance of the circuit optimized by LGR and Repeater Insertion is

$$\begin{aligned} C_{LGR} &= C_w + C_{gates} \cdot s_{LGR} \quad , \\ C_{rep} &= C_w + C_{gates} + C_0 \cdot N_{rep} \cdot s_{rep} \quad , \end{aligned} \quad (33)$$

where s_{LGR} is the optimal scaling factor for gates in LGR technique (31), and s_{rep} is the optimal scaling factor for inverter-based repeaters by (32), N_{rep} is the optimal number of optimally scaled repeaters for a wire of length L, as derived in [7], C_{gates} is the total capacitance of the initial circuit (prior to scaling) and C_w is a wire capacitance assumed to be the same for both optimizations (considering the critical path).

LGR is preferable in terms of power if,

$$N_{rep} > \sum_{i=1}^{N} C_i \cdot \left[\sqrt{\tau \left(\sum_{i=1}^{N} L_i \cdot \frac{g_i}{C_i} \right) \Big/ \left(\sum_{i=1}^{N} L_i \cdot C_{i+1} \right)} - 1 \right] \Big/ \sqrt{\frac{R_0}{C_0}} \quad (34)$$

In particular, for a chain of *N* identical gates with logical effort *g*, LGR is preferable in terms of power if

$$N_{rep} > N \cdot \sqrt{g} \quad (35)$$

In terms of delay, it would be beneficial to combine the two techniques: use smaller wire segments and add some repeaters. For short interconnect with a substantial number of gates N in the logic chain, LGR will be less efficient than repeater insertion in terms of dynamic power. In this case, the scaling of all gates will waste area and power. Still, LGR can be modified to be advantageous over classical Repeater Insertion, if a subset of the gates in the chain is used as the repeaters.

5. SUMMARY

Timing optimization in logic paths with wires has become an important issue in the VLSI circuit design process, as large logic blocks contain significant wire delays within critical timing paths. The existing techniques for minimizing delay treat only the particular cases of logic without wires (Logical Effort) or logic with a long resistive wire (Repeater Insertion). These particular cases are relatively rare in modern circuits. The general timing optimization problem should be based on more realistic models, which includes wires between the gates.

The techniques described in this work address the fundamental questions of optimal sizing and number and location of the gates. Unified Logical Effort (ULE) method allows fast determination of optimal sizing of gates in the presence of interconnect, while using intuitive closed-form conditions. The question of optimal number of repeaters is addressed by the Gate-terminated Sized

Repeater Insertion (GSRI) technique resulting in smaller delay and enhanced design flexibility as compared to standard repeater insertion. Logic Gates as Repeaters (LGR) method is used for optimal wire segmenting and gate location, suggesting a distribution of logic gates over interconnect, for driving the partitioned interconnect without adding many logically-redundant repeaters.

The combination of the proposed techniques provides solutions for a wide variety of design considerations. The proposed techniques enrich the toolbox of timing optimization in VLSI circuits by overcoming the limitations of the existing techniques and addressing a broad range of logic gate and wire combinations.

6. REFERENCES

[1] I. Sutherland, B. Sproull, and D. Harris, Logical Effort - Designing Fast CMOS Circuits, Morgan Kaufmann, 1999.

[2] I.E. Sutherland and R.F. Sproull, "Logical Effort: Designing for Speed on the Back of an Envelope," Proc. of the University of California/Santa Cruz Conference on Advanced Research in VLSI (ARVLSI), pp. 1-16, 1991.

[3] A. Morgenshtein, E.G. Friedman, R. Ginosar, and A. Kolodny, "Unified Logical Effort - A Method for Delay Evaluation and Minimization in Logic Paths with RC Interconnect", CCIT Technical Report #612, EE Pub. no. 1569, Technion, 2007. http://www.ee.technion.ac.il/matrics/papers/Unified%20Logical%20Effort-tr.pdf

[4] B. Lasbouygues, S. Engels, R. Wilson, P. Maurine, N. Azemard, and D. Auvergne, "Logical Effort Model Extension to Propagation Delay Representation," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 9, pp. 1677-1684, September 2006.

[5] A. Kabbani, D. Al-Khalili and A.J. Al-Khalili, "Delay Analysis of CMOS Gates Using Modified Logical Effort Model," *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, no. 6, pp. 937-947, June 2005.

[6] J. Ebergen, J. Gainsley, and P. Cunningham, "Transistor Sizing - How to Control the Speed and Energy Consumption of a Circuit," *Proc. of the IEEE International Symposium on Asynchronous Circuits and Systems*, pp. 51-61, April 2004

[7] H.B. Bakoglu, Circuits, Interconnections and Packaging for VLSI, Adison-Wesley, pp. 194-219, 1990.

[8] H.B. Bakoglu and J.D. Meindl, "Optimal Interconnection Circuits for VLSI," *IEEE Trans. on Electron Devices*, vol. ED-32, pp. 903-909, May 1985.

[9] S. Srinivasaraghavan and W. Burleson, "Interconnect Effort - A Unification of Repeater Insertion and Logical Effort," *Proc. of the IEEE Computer Society Annual Symposium on VLSI*, pp. 55-61, February 2003.

[10] A. Nalamalpu and W.Burleson, "Repeaters Insertion in Deep Submicron CMOS: Ramp-based Analytical Model and Placement Sensitivity Analysis," *Proc. of the IEEE Int'l Symposium on Circuits and Systems*, pp. 766-769, May 2000.

[11] V. Adler and E.G. Friedman, "Repeater Design to Reduce Delay and Power in Resistive Interconnect," IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing, vol. 45, no. 5, pp. 607-616, May 1998.

[12] K. Venkat, "Generalized Delay Optimization of Resistive Interconnections through an Extension of Logical Effort," *Proc. of the IEEE Int'l Symposium on Circuits and Systems*, pp. 2106-2109, May 1993.

[13] M. Moreinis, A. Morgenshtein, I. Wagner, and A. Kolodny, "Logic Gates as Repeaters (LGR) for Area-Efficient Timing Optimization," *IEEE Trans. on Very Large Scale Integration Systems,* vol. 14, no. 11, pp. 1276- 1281, November 2006.

[14] W. C. Elmore, "The Transient Response of Damped Linear Networks with Particular Regard to Wide Band Amplifiers," *J. of Applied Physics*, vol. 19, no. 1, pp. 55-63, January 1948.

[15] Predictive Technology Model, http://www.eas.asu.edu/~ptm/

[16] Virtuoso Advanced Analysis Tools User Guide, http://www.ece.uci.edu/eceware/cadence/aatoolsuser/chap3.html

[17] A. Cao, R. Lu, and C.K. Koh, "Post-Layout Logic Duplication for Synthesis of Domino Circuits with Complex Gates," *Proc. of the Conference on Asia South Pacific Design Automation*, pp. 260 - 265, January 2005.

[18] J.A. Davis, R. Venkatesan, K. A. Bowman and J. D. Meindl, "Gigascale Integration (GSI) Interconnect Limits and N-tier Multilevel Interconnect Architectural Solutions," *Proc. of the Int'l Workshop on System Level Interconnect Prediction*, San Diego, pp. 147-148, April 2000.

[19] C. Chu and D. F. Wong, "Closed Form Solution to Simultaneous Buffer Insertion / Sizing and Wire Sizing," *ACM Trans. on Design Automation of Electronic Systems*, vol. 6, no. 3, pp. 343-371, July 2001.

[20] Yu Cao, C. Hu, X. Huang, A.B. Kahng, S. Muddu, D. Stroobandt, D. Sylvester, "Effects of Global Interconnect Optimizations on Performance Estimation of Deep Submicron Designs," Digest of Tech. Paper, *ICCAD*, pp. 56-61, November 2000.

[21] G. Chen and E. G. Friedman, " Low Power Repeaters Driving RC and RLC Interconnects with Delay and Bandwidth Constraints," *IEEE Trans. Very Large Scale Integration (VLSI) Systems*, vol. 14, no. 2, pp. 161-172, February 2006.

[22] K. Nose and T. Sakurai, "Analysis and Future Trend of Short-Circuit Power," *IEEE Trans. Computer-Aided Design*, vol. 19, pp. 1023-1030, November 2000.