Power Consumption Estimation in CMOS VLSI Chips

Dake Liu and Christer Svensson

Abstract—Power consumption from logic circuits, interconnections, clock distribution, on chip memories, and off chip driving in CMOS VLSI is estimated. Estimation methods are demonstrated and verified. An estimate tool is created. Power consumption distribution between interconnections, clock distribution, logic gates, memories, and off chip driving are analyzed by examples. Comparisons are done between cell library, gate array, and full custom design. Also comparisons between static and dynamic logic are given. Results show that the power consumption of all interconnections and off chip driving can be up to 20% and 65% of the total power consumption respectively. Compared to cell library design, gate array designed chips consume about 10% more power, and power reduction in full custom designed chips could be 15%.

I. INTRODUCTION

Because of the increased circuit speed and density, power consumption in CMOS VLSI chips becomes increasingly important. Therefore, it is important to have a detailed understanding of the power consumption behavior of a chip, such as: What is controlling the power consumption in a digital CMOS VLSI chip? Which is the dominant part of the power consumption? How to estimate total power of a chip or power consumed in different parts of the chip? Finally, how to reduce power consumption during the early design phase?

Many power simulators have been developed at the gate level to analyze the power consumption with statistical methods or with certain stimulus vectors. Based on a netlist, the gate level simulation tools are not suitable for a feasibility study or other early estimations. There are other tools which could be used for the feasibility study or the estimation before circuit design. Powell and Chau, for example, have developed a power consumption model for a class of digital signal processors [1]. Other tools such as those given in [2], [3], are based on cell and gate counts. Yet the power consumption was not analyzed separately for different parts of the chip such as clocking power or power consumed by memory on chip. Therefore, the motivation of this paper is to get a general estimation method for feasibility studies and to investigate the distribution of power between different parts of a chip, to be used as guidelines for early design.

We will discuss these problems in this paper. We will discuss the power consumption divided into five parts: logic circuit, clock distribution, memory, interconnection, and off chip driving. We will compare power consumption between cell library design, gate array design, and full custom design. In the logic circuit part, we will discuss static and domino logic latched by different latches. We will discuss how the power consumption is influenced by logic depth and clock distribution. In the memory part, we will discuss power consumption from different parts of the memory block. In the off chip driving part, we will discuss maximum off chip driving power under certain off chip technologies.

II. ESTIMATION MODELING

A. Model of Logic

We will discuss three kinds of logic circuit styles. The first is the buffered static logic latched by static master slave D flip-flops. The second is domino logic latched by dynamic master slave D flip-flops. The third is domino logic latched by simple dynamic latches. We define the average logic gate function as a three input AND (buffered NAND) gate connected to three-identical AND gates at the output node. Thus, the average fan in and fan out is 3. We simulate random logic circuits with logic depth of $f_{id}$ as in Fig. 1(a). Logic gates are shown in Figs. 1(b) and (c) and latches are shown in Fig. 2. All N transistors in logic gates are of minimum gate width and all PMOST's in logic gates are twice the width of N transistors. $C_t$ is a minimum size NMOST gate capacitance and we assume that the $C_t$ is also a minimum size drain diffusion capacitance. Therefore, one gate capacitance and one diffusion capacitance of a PMOST are both 2$C_t$. When estimating the power consumption of logic gates, we should analyze gates with an environment in the cell of Fig. 1(a). The cell includes combinational logic followed by a latch. The power consumption from the latch is divided into $f_{id}$ pieces and one piece is added to one gate in the cell.

Because the average node duty factor, $f_d$ (node activity ratio) is different between clocked nodes and unclocked nodes, different duty factors are used to sum up equivalent capacitances in logic gates for power consumption estimation. For every logic gate in Fig. 1(a), the total equivalent capacitance for power estimation is divided into three parts. The second and the third part will be discussed later and the first part, in (1), is the logic gate capacitance excluding clock driving nodes.

$$C_{logic} = f_d f_{id} k_3 C_{tr} + f_d k_2 C_{tr} + 3 f_d C_{tr} + k_3 f_d C_{tr}.$$  (1)

where, $f_d$ is average fan in and fan out, $f_{id}$ is the duty factor, $k_1$, the input circuit structure factor, is the number of minimum

Manuscript received October 13, 1993; revised February 28, 1994.

The authors are with the Department of Physics and Measurement Technology, Linköping University, S-581 83 Linköping, Sweden.

IEEE Log Number 9402130.

0018-9200/94$04.00 © 1994 IEEE
Fig. 1. Logic circuit styles used in this paper. (a) A cell built by a latch and random logic gates. (b) Domino logic. (c) Buffered static logic.

Fig. 2. Latches used in this paper. (a) Static MS D flip-flop. (b) Dynamic MS D flip-flop. (c) Dynamic simple latch.

Fig. 3. Memory structure and a six transistor memory cell.

size gate capacitances in one input node, and \( k_2 \) is the inverter buffer node factor. It is the number of the minimum size gate or diffusion capacitances on the buffer input node. The third term of (1) is the equivalent capacitance at the output node of a logic gate. Because loads of a gate are input nodes in other gates, their power consumption has been counted in the following gate. Therefore, loads of the output nodes are not included. \( k_3 \) is the factor used to calculate the equivalent capacitance on no-output nodes 1 to 5.

The second part of the equivalent logic gate capacitance is the capacitance on nodes driven by the clock. To decide the global clock buffer size, this is classified as loads of the clock distribution. The third part of the equivalent logic gate capacitance is \( 1/f_{sd} \) of the unclocked capacitances in a latch following the logic gates:

\[ C_{\text{latch}} = f_d k_4 C_{tr}, \]

where, \( k_4 = 6/f_{sd} \) for a dynamic master slave \( D \) latch, \( k_4 = 3/f_{sd} \) for a simple dynamic latch, and \( k_4 = 12/f_{sd} \) for a master slave static \( D \) latch. The total equivalent logic gate capacitance for a logic gate is:

\[ C_{\text{total}} = C_{\text{logic}} + C_{\text{latch}}. \]

Finally the upper limit of the global clock frequency \( f_{c,max} \) is given by (4) [3]. \( f_{c,max} \) is also the system clock frequency when measuring the maximum system power consumption. \( T_g \) is the gate delay, \( D_c \) is the chip dimension without memories, and \( v_c \) is the propagation speed of the electromagnetic wave.

\[ f_{c,max} = \left( \frac{f_s T_g + r_{int} C_{int}}{2} + \frac{D_c}{v_c} \right)^{-1} \]

\[ T_g = f_g R_{gout} C_{gout} + f_g R_{gout} C_{gout} \]

\[ + r_{int} C_{int} \]

\[ \text{excl.} P_{\text{total}} = 0.5 f_{c,max} C_{\text{total}} V_{dd}^2 \]

B. Model of On-Chip Memory

The purpose of this paper is to give an overview of the power consumption in an ASIC system. Therefore, we do not need a deep description of a specific memory structure or a certain cache in a special system. Instead, we just use a well-accepted memory structure and its power consumption as a typical case.

A typical memory [5] is divided into four parts: the memory cell, the row decoder, the column selection, and read/write circuits. The memory arrangement is in Fig. 3. To compare the power consumption with different row widths, we define the storage array as a \( 2^n \times 2^{n-k} \) matrix with \( 2^n \) memory cells, \( 2^{n-k} \) rows, and \( 2^k \) columns. A typical control circuit of a bit line and a sense amplifier is shown in Fig. 4 [5]. Address-to-row-select decoder and Address-to-column selector are shown in Fig. 5. Just before an operation, \( \Phi_1 \) is high, both bit and bit lines are precharged high. During an operation cycle, \( \Phi_2 \) is on and precharge is off. All memory cells on one row select line are connected to their bit and bit lines because this row line is selected high. But only one sense amplifier is on and sends out its detected low signal through the column select circuit.
The total capacitance in the buffer chain is about 0.3 of its total load. Thus, the last inverter drives the load with total capacitance of 1/4 of its load. This is calculated from the assumption that the size ratio in the inverter chain is 4. Thus, the last inverter drives the load with total capacitance of 1/4 of its load. The total capacitance ratio of the inverter chain is 1/3. A \( A_i \), and \( n-kA_i \) address lines. The wire capacitance is \( c_{int} \).

The third step is to model the power from row decoding. It is the power consumed in all loads on one horizontal row line. Because only one row line is active in one operation, the total row driving power is:

\[
P_{row\text{-driving}} = \frac{1}{2} \{ 2^k(2C_{tr}) + 2(n-k)C_{tr} + c_{int}[8(n-k)W_{int} + l_{row}] \} V_{dd}^2, \quad (8)
\]

where, \( 2^k(2C_{tr}) \) is loads from memory cells on one row line. \( 2(n-k)C_{tr} \) is loads of drain capacitances in the row decoding matrix. \( 8(n-k)W_{int} + l_{row} \) is the total row select line length.

The fourth step is to model the power consumption from all column select parts. It includes all power used for column selections, and power from one sense amplifier. Power for one column selection includes: 1) power from the address buffer, 2) power from a half of all NMOST gates in the column selection matrix, and 3) power from interconnections in this part. It is modeled as:

\[
P_{column\text{-select}} = \frac{1.3}{2} \left( \sum_{i=1}^{k-1} 2^{k-i}C_{tr} + k c_{int} l_{column} \right) V_{dd}^2, \quad (9)
\]

where, 1.3 is the buffer chain power ratio, and the sum term is the power from a half of all gates in the column selection matrix. Power consumed from the sense amplifier and the readout inverter includes static power and dynamic power. Static power is from the differential amplifier, and dynamic part is from the read out inverter. The power of one memory operation in this part is:

\[
P_{sensamp.+load} = \frac{V_{dd} I_{sens}}{f_{mem\text{-clock}}} + \frac{1.25}{2} c_{int} \times \left( 8kW_{int} + \frac{l_{row}}{2} \right) V_{dd}^2, \quad (10)
\]

where, the first part is static power and the second part is dynamic power. \( I_{sens} \) is twice maximum drain current of a minimum size NMOST. \( f_{mem\text{-clock}} \) is the memory clock frequency. The second part power includes power from the inverter and power from all interconnections after the read out inverter.

The memory power in one operation is the total power from the structure in Fig. 3. It is the sum from (6) to (10).

C. Model of the Local and Intermediate Interconnections

We define two categories of interconnections in CMOS VLSI, local and intermediate interconnections, and global buses. The local interconnection could be defined as interconnections within a logic gate. The intermediate interconnections are used for connections between gates or sub systems. The global bus includes data, control, and address buses. The interconnection width is assumed to be minimum wire width.

Authorized licensed use limited to: Rice University. Downloaded on January 3, 2009 at 17:19 from IEEE Xplore. Restrictions apply.
excluding the clock distribution, as the wire RC constant does not change with wire width.

The lengths of local and intermediate interconnections could be modeled based on Rent's rule. Rent's rule is given by [3]:

\[ N_p = K_p N_g^\beta \] (11),

where, \( N_p \) is the number of external signal connections to a logic block, \( N_g \) is the number of logic gates in the block, \( K_p \) is a constant, and \( \beta \) is the Rent's rule constant. The local and intermediate interconnection length of a logic gate is:

\[ l_{av} = \bar{R} d_g, \] (12)

where, \( \bar{R} \) is the average gate pitch ratio, and \( d_g \) is the average gate dimension. Following [3], \( \bar{R} \) is derived from Rent's rule and the assumption of hierarchical layout placement is in Fig. 6(a) [8].

\[ \bar{R} = 2 \left( \frac{N_g^{\beta-0.5}}{4^{\beta-0.5}} - 1 \right) \left( \frac{N_g^{1.5}}{4^{1.5}} - 1 \right)^{-1} - 1 
\] (13)

For an interconnection limited chip, the gate dimension is:

\[ d_g = (T_e + W_e) \frac{f_g R_{pw}}{c_w n_w}, \] (14)

where, \( T_e \) is the average transistor dimension weight factor in a gate. It is determined by area and numbers of minimum size \( N \) and PMOST's as well as metal connection area on drain/source diffusion in a gate. \( W_e \) is the interconnection factor of a logic gate. The sum of \( W_e \) and \( T_e \) is 1, \( f_g R_{pw} / c_w n_w \) is the wiring ability determined gate dimension. \( p_w \) is the wire pitch, \( c_w \) is the utilization efficiency of the chip interconnections, and \( n_w \) is the number of wire layers in the chip. This model is suitable for cell library and gate array design. In full custom design, the gate dimension in (15) is limited by transistor area. It is a transistor packing density limited gate.

\[ d_g = F \sqrt{A_g}, \] (15)

where, \( F \) is feature size and \( A_g \) is the unit gate area extracted from experimental layouts. The average interconnection length of a gate with a fan out of \( f_g \) is represented as:

\[ l_{av} = f_g l_{av}. \] (16)

The total average interconnection capacitances are:

\[ C_{av} = N_g f_g C_{int}. \] (17)

where, \( C_{int} \) is the unit wire capacitance per unit length with a minimum wire width of \( w_{int} \). \( C_{int} \) is 2 pf/cm when \( w_{int} \) is less than 3 \( \mu \)m [3]. The power consumed in local and average wires is:

\[ P_{local,average} = \frac{1}{2} f_e C_{av} V_{dd}^2. \] (18)

D. Model of the Global Clock Distribution

Different systems may have different clock distributions. One example of clock arrangement with low clock skew is a H-tree [3] [9], in Fig. 6(b). The H-tree and the clock driver are matched at the source end. Therefore, the width of a clock wire is half its incoming width before the branching points. If the far end clock wire has minimum width, the global clock wire capacitance is:

\[ C_{clockwire} = \left( \frac{16}{2} D_t + \frac{1}{2} D_t + \frac{2 \times 4}{2} D_t + \frac{4 \times 2}{2} D_t + \frac{8 \times 1}{2} D_t \right) C_{int} 
\] (19)

where, the chip dimension is \( D_t = \sqrt{D_G^2 + D_M^2} \). The first part, \( D_t = d_g \sqrt{N_g} \) is the chip dimension of logic part and \( D_M = 2^{n/2} d_m + n(4W_{int}) \) is the memory block dimension. The global clocking load includes all clocked transistors, all clocking wire loads, and all clock drivers:

\[ C_{total, clock} = (1 + k_{driver}) \times \left( L_{clk} N_g C_{tr} + 2^{k+1+1+2} C_{tr} + C_{clockwire} \right), \] (20)

where, \( k_{driver} \) is the clock driver ratio. It means that the total clock driver capacitance is \( k_{driver} \) times of the clock driven loads. It could be 0.3 for a conventional system. If the system needs very high speed, \( k_{driver} \) could be larger. \( L_{clk} \) is the number of clock driven transistor gates in a logic gate (in Fig. 1(b), (c)) and in a latch (in Fig. 2). \( L_{clk} = 3 + 6 / f_{dd} \) for domino logic with dynamic MS D flip-flops, \( L_{clk} = 3 + 3 / f_{dd} \) for domino logic with simple latch, and \( L_{clk} = 12 / f_{dd} \) for static logic. The second term is from memory precharge and control. In the term of \( 2^{k+1+1+2} \), the first 1 is the PMOST's sizing factor, the second 1 means that there are two PMOST's for bit and bit, and 2 means that there are four clocked transistors in the control circuit. The total power consumption from clock distribution is:

\[ P_{clock} = f_e C_{total, clock} V_{dd}^2. \] (21)

E. Model of the Global Bus

The bus consumes power from three parts, the power from the bus wire capacitance, the power from the bus loads, and the
power consumed in bus drivers. The total bus wire capacitance is:

\[ C_{\text{buswire}} = (k_{\text{bus}} + W_{\text{bus}})(D_c + D_M)C_{\text{int}} \]  

(22)

where, \( W_{\text{bus}} \) is the bus width, the number of parallel bits. It may include width of data and address. \( k_{\text{bus}} \) is the equivalent global control bus number. An example is \( k = 3 \), one for read/write control, one for data/address control, and one for other controls. The bus loads include ALU’s, memory access ports, register blocks, inputs of bus drivers, and others.

\[ C_{\text{busload}} = 3C_{\text{tr}}W_{\text{bus}}N_{\text{total,b,load}} \]

(23)

where \( N_{\text{total,b,load}} \) is the total load number in one bit bus. \( 3C_{\text{tr}} \) is an inverter input capacitance. Though there might be many groups of bus drivers on one common bus, only one group is active in one piece of time. Thus, the total capacitances of bus drivers are:

\[ C_{\text{busdriver}} = 0.3(C_{\text{busload}} + C_{\text{buswire}}) \]

(24)

where, 0.3 was shown after (7). Finally, the total bus capacitances are the sum of (22), (23) and (24). The total power consumed by bus is given by (25), and the total chip interconnection power \( P_{\text{wire}} \) is the sum of (18) and (25):

\[ P_{\text{bus}} = \frac{1}{2}I_c(C_{\text{buswire}} + C_{\text{busload}} + C_{\text{busdriver}})V_{dd}^2 \]

(25)

\[ F. \ Model \ of \ Off \ Chip \ Driving \]

Off chip driving power is consumed in two parts. One is the power used to drive off chip capacitance, bonding wires, and the pad capacitance. The other is the power consumed by the driver itself, an inverter driving chain. The first part is not given by the silicon chip technology. It is determined by the package technology and printed circuit or multichip technology. We define three kinds of off chip technologies [4]. One is traditional package with traditional printed circuit board (PCB), and the total off chip capacitance, \( C_{\text{offchip}} \), is 50 pF. The second is advanced package and advanced PCB, and the total off chip capacitance is 30 pF. The third is multichip module technology with a total off chip capacitance of 10 pF. The width decrease ratio in the inverter chain is 4. Therefore, a total inverter chain capacitance is \( C_{\text{offchipdriver}} = 0.3C_{\text{offchip}} \). The power consumption of off chip driving is:

\[ P_{\text{offchip}} = \frac{1}{2}I_c(C_{\text{offchip}} + C_{\text{offchipdriver}})V_{dd}^2 \]

(26)

\[ \text{G. Model of Three Kinds of Layout Design Strategies} \]

There are three kinds of layout design strategies: full custom design, cell library design, and gate array design. Because of the difference of the design tools and designer’s skill, the same circuit function could be designed much differently by different designers. Under this restriction, we can only define some average situations. For example, we define the full custom design as being designed by an experienced designer, so that the local and intermediate interconnections are minimized, the silicon area is best compacted, and logic gates are reasonably organized. Thus, the gate dimension is limited by transistor dimension and the gate pitch ratio is minimized to \( \sqrt{2} \). That means the intermediate interconnection wiring space between two gates is about 0.4 of their sizes. The cell library design and gate array design are finished with automatic design tools. Gate positions are arranged by a computer, the gate dimension is interconnection limited, the gate pitch is defined by Rent’s rule constant. For cell library design, \( \beta = 0.43 \) and for gate array design, \( \beta = 0.5 \) [3].

\[ \text{III. POWER ESTIMATIONS AND DISCUSSIONS} \]

This part is divided into four sections. In the first section, we will verify all estimation models in this paper by comparisons of two real designs. In the second section we will define a conventional ASIC system as an example chip and describe its parameters. In section three, we will give a total power discussion of the example chip to show how the power consumption is distributed in different parts of the chip and to compare the power consumed in full custom design, cell library design, and gate array design. In section four, we will discuss in detail power consumptions of interconnections, clock distribution, on-chip memory, off-chip driving and power versus logic depth based on different system parameters.

\[ \text{A. Verification of Estimation Models} \]

Verification of the estimation models developed above has been done in this section by comparing two sets of published data. The first is the Alpha 21064 microprocessor [14], [15], and the second is Intel 80386 microprocessor [3]. Comparisons are given in Table I between parameters from both reference papers and our estimations. In our estimation of Alpha 21064, process technology parameters are feature size of 0.75 \( \mu m \), gate oxide thickness of 10.5 nm, minimum interconnection width of 0.75 \( \mu m \), average wire pitch of 2.625 \( \mu m \), and memory cell area of \( 10 \times 10 \mu m \). The total clock load is 3.2nF driven by a driver chain with ratio \( k_{\text{driver}} \) of 0.37. The system parameters are a supply voltage of 3.3 V, threshold voltages of 0.5 V, average logic depth of 7, and average node activity ratio of 0.3. In the estimation of Intel 80386, process technology parameters are feature size of 1.5 \( \mu m \), gate oxide thickness of 30 nm, minimum wire width of 3 \( \mu m \), and wire pitch of 6 \( \mu m \). The system parameters are a supply voltage of 5 V, threshold voltages of 0.7 V, average logic depth of 25, and average node activity ratio of 0.3. Estimations show that the estimated system parameters (for example, the maximum system clock frequency, the total power consumption, and the chip dimension) are close to parameters given in references. This shows that models proposed in this paper is reasonably accurate.

\[ \text{B. Parameters} \]

In the following sections, we will estimate power consumption based on an ASIC system example. This example contains 10000 logic gates and 32 kbits (4Kbytes) of memory. The logic circuit styles are buffered static logic, domino logic latched by dynamic MS flip-flops, or domino logic latched by simple dynamic latches. The logic depth is 10. Two layers of wires are used in layout. We define the bus width as 32 bits, and the
TABLE I
ESTIMATION MODEL VERIFICATION

<table>
<thead>
<tr>
<th>Alpha 21064</th>
<th>Intel 80386</th>
</tr>
</thead>
<tbody>
<tr>
<td>f_{cmax} (MHz)</td>
<td>214</td>
</tr>
<tr>
<td>P_{tot} (W)</td>
<td>32</td>
</tr>
<tr>
<td>P_{gates} (W)</td>
<td>4.6</td>
</tr>
<tr>
<td>P_{pe} (W)</td>
<td>10.2</td>
</tr>
<tr>
<td>P_{pchip} (W)</td>
<td>11.9</td>
</tr>
<tr>
<td>P_{memory} (W)</td>
<td>0.73</td>
</tr>
<tr>
<td>Gate number</td>
<td>263666</td>
</tr>
<tr>
<td>Tr. number</td>
<td>1.68 M</td>
</tr>
<tr>
<td>Memory</td>
<td>16 k</td>
</tr>
<tr>
<td>Logic depth</td>
<td>7</td>
</tr>
<tr>
<td>Chip size</td>
<td>1.57 x 1.57</td>
</tr>
</tbody>
</table>

E 3.5-

TABLE II
CMOS TECHNOLOGY PARAMETERS

<table>
<thead>
<tr>
<th>Feature size (μm)</th>
<th>2 μm</th>
<th>1 μm</th>
<th>0.5 μm</th>
<th>0.25 μm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gate width W (μm)</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>1.2</td>
</tr>
<tr>
<td>Gate overlap SL (μm)</td>
<td>0.30</td>
<td>0.15</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Gate oxide thickness (nm)</td>
<td>40</td>
<td>25</td>
<td>15</td>
<td>8.5</td>
</tr>
<tr>
<td>Width of metal (Wint) (μm)</td>
<td>3</td>
<td>2</td>
<td>1.5</td>
<td>1</td>
</tr>
<tr>
<td>Metal pitch pW (μm)</td>
<td>6</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

number of off chip drivers as 64. f_{cmax} is the system clock frequency and off chip data rate. The memory data access rate is 1/4 of the system clock frequency. The average activity ratio for all unclocked nodes is 0.2. The off chip capacitance is 50 pF. Layout design style is cell library, gate array, or full custom design.

3 V supply voltage is used to get the same comparing condition, though 5 V could be used in 2 and 1μm technology. Feature sizes are 2, 1, 0.5 and 0.25 μm. The CMOS technology parameters are given in Table II. Metal wire width and pitch are minimum sizes. There are other experimental parameters, such as gate area factor A_g, transistor site factor T_s, local interconnection factor W_c, and memory size d_m, not being shown in this table, but being extracted from experimentally layouts made in GDT (Mentor tool).

C. A System Example

Estimation results based on parameters in Section 2 are shown in Fig. 7(a). As f_{cmax} increases with downscaling the total power will increase with scaling. In Fig. 7(b) we give the system power in mW/MHz. The thick lines give data for a system designed with static logic in cell library design. It is used as references in the following.

We have then tried to study the power distribution between different parts (logic circuit, interconnections, clock driving, and off-chip driving) of the system in Figs. 8 and 9. In Fig. 8 we compare the differences of the power distribution between static logic and dynamic logic based on cell library design. The total power is in Fig. 8(a) and the power without off chip driving is in Fig. 8(b). In a similar way, we compare the differences of the power distribution between three design styles (cell library design, gate array design, and full custom design) based on static logic in Fig. 9. Again data is given for as well a complete system as for a system without off chip drivers.

As the off-chip driving power cannot scale down, the off-chip driving power could be up to 70% of the total chip power.
The power used for precharge and evaluation in gate dimension and more PMOST's of static logic. The total consumption when using different latches in domino logic. The logic depth is large, there is not much difference in power that when the logic depth is larger than this value, the dynamic logic will consume more power than static logic. When logic depth is larger than a certain value, dynamic logic consumes more power than static logic. This is because the relative number of clock driven transistors is decreased in static logic when increasing the logic depth. The critical logic depth, in Fig. 10 for domino logic with simple latches, is about 6. If dynamic master slave D flip-flops are used in dynamic logic, the critical logic depth is only 4. The critical logic depth means that when the logic depth is larger than this value, the dynamic logic will consume more power than static logic. When logic depth is large, there is not much difference in power consumption when using different latches in domino logic. The reason is that the power used for precharge and evaluation in the gates is the dominant part of the clock power consumption.

If we analyze the wire power in detail, we find that the local and intermediate interconnection power is the dominant part (about 90%) of the total wire power. The total wire power is scaled down with the scale of the wire pitch instead of the feature size.

Based on the same logic style (for example, domino logic with simple latch) and the same logic depth, the difference of the clock power consumption is very small comparing cell library design, gate array design and full custom design. The reason is that the dominant clock power is consumed in silicon gate loads.

When memory size is more than 4 kbytes, 80-70% of the memory power is consumed in a row of memory cells driving bit/bit lines. The power consumption is not linearly scaled by scaling the feature size. This is mainly because that the unit interconnection capacitances in the memory block do not decrease when wire width is less than 3 μm. When we use different number of rows under the same memory size, the power consumption is just changed less than 6% of the total memory power. This is because the dominant power, used to drive and precharge bit/bit lines, is not changed.

We estimate the maximum power consumed from total off chip driving in Table III. Because this power does not depend on silicon technology but depends on off-chip technologies, it is not scaleable. In table 3, 50 pF, 30 pF, and 10 pF are one bit total off chip load capacitances, % is the percentage of the off-chip power in a chip, mpm is the total off chip power in mW/MHz, and W is the maximum off chip power in watt under $f_{c_{max}}$. If we use multi-chip module, the off chip power consumption could have a great reduction from 62.5% to 25% of the total chip power, when feature size is 0.25 μm.

IV. CONCLUSION AND RECOMMENDATIONS

We have developed a method and a tool for power modeling of CMOS VLSI chips in this paper. The method makes it possible to estimate the power consumption of a chip based on gate count, memory size, logic, and layout styles. The tool is not as accurate as gate level simulators, but it gives a fast estimation far before circuit and layout design. A few verifications with known chips indicate that the models give reasonable results.

We have further used these models to describe the power consumption of a schematic example, with the goal to find
TABLE III

<table>
<thead>
<tr>
<th>Off Chip Power (mpm=mW/MHz)</th>
<th>2 µm, 31 MHz</th>
<th>1 µm, 50 MHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit</td>
<td>% mpm W</td>
<td>% mpm W</td>
</tr>
<tr>
<td>50 pF</td>
<td>3.18 3.84 0.12 44.3 3.84 0.19 53.2 3.84 0.39 62.5 3.84 0.83</td>
<td></td>
</tr>
<tr>
<td>30 pF</td>
<td>2.19 2.3 0.07 32.3 2.3 0.12 40.6 2.3 0.23 50.0 2.3 0.50</td>
<td></td>
</tr>
<tr>
<td>10 pF</td>
<td>8.54 0.77 0.02 13.7 0.77 0.04 18.5 0.77 0.77 25.0 0.77 0.17</td>
<td></td>
</tr>
</tbody>
</table>

a strategy for low power design. From this example we thus conclude:

- The total power increased with downsizing. This is caused by the increased clock frequency and indicates that high-end designs will have more severe power problems in the future. The only way to handle this problem is to reduce supply voltage [6], [11], [16].

- The power used for off-chip driving is very dominating and becomes more dominating with scaling (Table III). Figs. 8(a), 9(d). Up to 70% of the power may be due to off-chip driving. To reduce power the most important thing is therefore to reduce the power used for off-chip driving. This can be done by using a more advanced off-chip technology [17], by reducing the off-chip swing [12], [13] or by reducing the external bandwidth of the chip by proper high level partitioning (e.g., by using single chip solutions) [1], [3].

- If off-chip power is excluded, the power related to wires could be up to 46% of the chip power. This share increases with downsizing. Considerable reduction (70%) of wire power can be achieved by using full custom design style.

- Comparing design styles, gate array style uses about 10% more power than cell library style and full custom style uses about 15% less than cell library style (at fixed speed) (Fig. 9). Full custom design utilizes power best as it has a low share of interconnection power (Fig. 9(c)).

- Comparing logic forms we need to consider the sum of power, clock power and wire power (Fig. 10). Static logic uses more power than domino logic for small logical depth. However, for logical depths larger than six static logic uses less power than domino logic.

- The clock power is about twice the logic power for static logic and about three times the logic power for dynamic logic. Therefore, when using domino logic a larger share of the total logic power must be supplied through the clock generator.

The power used for off-chip driving is very dominating, and becomes more dominating with scaling (Table III). Figs. 8(a), 9(d). Up to 70% of the power may be due to off-chip driving. To reduce power the most important thing is therefore to reduce the power used for off-chip driving. This can be done by using a more advanced off-chip technology [17], by reducing the off-chip swing [12], [13] or by reducing the external bandwidth of the chip by proper high level partitioning (e.g., by using single chip solutions) [1], [3].

- If off-chip power is excluded, the power related to wires could be up to 46% of the chip power. This share increases with downsizing. Considerable reduction (70%) of wire power can be achieved by using full custom design style.

- Comparing design styles, gate array style uses about 10% more power than cell library style and full custom style uses about 15% less than cell library style (at fixed speed) (Fig. 9). Full custom design utilizes power best as it has a low share of interconnection power (Fig. 9(c)).

- Comparing logic forms we need to consider the sum of power, clock power and wire power (Fig. 10). Static logic uses more power than domino logic for small logical depth. However, for logical depths larger than six static logic uses less power than domino logic.

- The clock power is about twice the logic power for static logic and about three times the logic power for dynamic logic. Therefore, when using domino logic a larger share of the total logic power must be supplied through the clock generator.

REFERENCES


Duke Liu was born in Heilongjiang, China, in 1957. He received the B.S. and M.Sc. degrees in electronic engineering from Northern Jiatong University, Beijing, China in 1982 and 1987, respectively. He received the Swedish Technology Licentiate degree in 1993 and is now working toward the Ph.D. degree in electronic devices at the Department of Physics and Measurement, Linköping University, Linköping, Sweden. From 1982 to 1990, he did his research work in the field of automatic control and EMC in the Department of Communication and Control Engineering, Northern Jiatong University, Beijing.

Christer Svensson was born in Borås, Sweden in 1941. He received the M.S. and Ph.D. degrees in electrical engineering from Chalmers University of Technology, Gothenburg, Sweden, in 1965 and 1970, respectively. He was with the Research Laboratory of Electronics at Chalmers University from 1965 to 1978, where he performed research on MOS transistors, MNOS memory transistors, Pd-MOS gas-sensitive transistors, and the physics and chemistry of the MOS system. Since 1978, he has been with the Department of Physics and Measurement Technology, Linköping University, Linköping, Sweden, first as Lecturer, and since 1983, Professor of Electronic Devices. At Linköping, he started a new research group on integrated circuit design. He is now performing research on high performance analog and digital CMOS circuit techniques, on timing verification and optimization of CMOS circuits, and on integrated electronic and sensor systems. He has published 80 papers in international journals and conferences and holds 501–50 patents. Dr. Svensson was awarded the Solid-States Circuits Council 1988–89 Best Paper Award. He is a member of the Royal Swedish Academy of Engineering Sciences.