VLSIs in the year 2010 and beyond
- From a designer’s point of view -

Takayasu Sakurai
Center for Collaborative Research and Institute of Industrial Science, University of Tokyo
7-22-1, Roppongi, Minato-ku, Tokyo 106-8558, Japan

Abstract
VLSI designers will face three crises in the coming years: the power crisis, interconnection crisis, and complexity crisis. This paper discusses these crises and possible solutions to them and presents a possible view of future VLSIs.

Introduction
In the last few years, the economic environment surrounding the semiconductor industry has been harsh, but from this year, strong growth is expected to return, as shown in Fig. 1. The market has been changing from being memory oriented to being processor and logic oriented. The market leader is expected to be the System-on-a-Chip (SoC), which integrates processors, logic, and memory. Technically speaking, the history of VLSIs is a history of miniaturization. The gate length of MOSFETs...
has now shrunk to 0.13 µm and the gate oxide is only a few SiO₂ molecule layers thick, as shown in Fig. 2. This miniaturization is backed by the scaling theory, which states that a MOSFET operates at higher speed without any degradation of reliability when the device size is scaled by a factor of $k$ and, at the same time, the operating voltage is scaled by a factor of $k$ as shown in Fig. 3. The circuit cost per function is also decreased by miniaturization. Thus, the cost-performance is rapidly improved as devices are scaled down. That is why miniaturization has been pursued so persistently for the last thirty years and will continue to be pursued in the future. Recently, however, undesirable side-effects of scaling have become noticeable. The power density increases; interconnect-related quantities, such as interconnect delay, current density, and noise, increase; and, since the number of devices on a chip increases, designing and testing VLSIs become more difficult.

In short, there will be three crises in making VLSIs in the coming years: a power crisis, interconnection crisis, and complexity crisis. This paper discusses these crises and possible solutions to them and presents a possible view of future VLSIs.

**Power crisis**

Is the power crisis real? The answer is shown in Fig. 4. A single chip processor in production consumes more than 100 W of power and this value will rise to about 150 W during the next decade, as shown in Fig. 5, which is comparable to an electric light bulb. The power consumption does not sound like an important figure-of-merit, when compared with speed and cost. When we think back to the 1980s, however, we recall that CMOS technology took over the position of dominant technology from NMOS and bipolar technology. CMOS is more expensive than NMOS and slower than bipolar at the device level, but its low-power characteristics led to higher integration that yielded higher performance at the system level. Thus, power consumption is a very important figure-of-merit in predicting the trend of technology.

In addition to the heat generated by the consumed power, the huge current needed to operate VLSIs is also an issue. The increase in current will come from the decrease in supply voltage ($V_{DD}$) to about 0.3 V in 15 years. This huge current will give rise to the $IR$ (current-resistance) voltage drop problem, which is described in the interconnect section. Low-power design of VLSIs is important not only because power consumption is rising dramatically as a
direct result of the scaling law, but also because mobile electronic systems, which will form the infrastructure of the information technology age, need longer-lasting batteries.

The power and delay of a CMOS gate are expressed by

\[
\text{Power} = p_t \cdot f \cdot C_L \cdot V_{DD}^2 + I_0 \cdot 10^{-6} \cdot V_{DD},
\]

\[
\text{Delay} = \frac{C_L \cdot V_{DD}}{f \cdot (V_{DD} \cdot V_{TH})^s}.
\]

where \( p_t \) is switching probability, \( C_L \) is load capacitance, \( V_{TH} \) is threshold voltage, \( v_s \) is a velocity saturation index whose value is about 1.3 for recent MOSFETs \(^{[2]} \), \( f \) is the clock frequency, and \( s \) is an \( s \)-factor whose value is about 0.1 \( V/\text{decade} \) for bulk CMOS technology. The power and delay of a typical CMOS gate are plotted in Fig. 6 and some of the important approaches to achieving low power are summarized in Fig. 7. As can be understood from Fig. 6, if \( V_{DD} \) is lowered, the power is reduced effectively because it depends quadratically on \( V_{DD} \); however, the delay increases. The delay can be shortened by decreasing \( V_{TH} \), but a low \( V_{TH} \) induces a large subthreshold leakage current, which prevents the total power being decreased. Consequently, the trade-off between speed and power should be considered in order to decrease the power. Promising schemes handle this trade-off by controlling \( V_{DD} \) and \( V_{TH} \) in some ways to reduce power wastage and to adaptively use power only when and where it is needed, as shown in Table 1.

Low power has been pursued by many researchers and engineers at the process, device, circuit, CAD, software, and system levels.

The new trend, however, is to search for a solution utilizing cooperation among different levels, such as device-circuit cooperation and circuit-software cooperation. One such effort is shown in Fig. 8, where the device side provides transistors with multiple oxide thicknesses and multiple threshold voltages, while the circuit side makes use of the various transistors together with multiple voltages to lower the power. For example, the CMOS logic part of a VLSI is built with low-\( V_{TH} \) MOSFETs, which are leaky, but a high-\( V_{TH} \), thick-\( t_{ox} \) device is inserted in series with the logic part to cut off the leakage when the logic is not being used. The leakage is caused not only by the subthreshold current, but also by gate oxide tunneling and the reverse-biased junction current. Boosting the gate voltage of the inserted transistor maximizes the new trend, however, is to search for a solution utilizing cooperation among different levels, such as device-circuit cooperation and circuit-software cooperation. One such effort is shown in Fig. 8, where the device side provides transistors with multiple oxide thicknesses and multiple threshold voltages, while the circuit side makes use of the various transistors together with multiple voltages to lower the power. For example, the CMOS logic part of a VLSI is built with low-\( V_{TH} \) MOSFETs, which are leaky, but a high-\( V_{TH} \), thick-\( t_{ox} \) device is inserted in series with the logic part to cut off the leakage when the logic is not being used. The leakage is caused not only by the subthreshold current, but also by gate oxide tunneling and the reverse-biased junction current. Boosting the gate voltage of the inserted transistor maximizes...
mizes the drain conductance of the inserted transistor, which in turn yields high-speed operation\[^{[3]}\]. The effective use of multiple values of $V_{DD}$, $t_{OX}$, and $V_{TH}$ is the key to prevent power consumption from increasing explosively.

Another example of a promising cooperative approach is found in the circuit and software level, as shown in Fig. 9, namely the $V_{DD}$ hopping scheme. The circuit side provides a processor whose operating frequency and $V_{DD}$ can be varied by software. The software side controls the frequency and $V_{DD}$ adaptively so that the frequency is halved when high-speed operation is not needed. The scheme has been applied to a digital video codec system and the processor power is only one-fourth that of the conventional fixed $V_{DD}$ processor\[^{[4, 5]}\]. The video codec system guarantees real-time operation for any data input but the highest performance is only needed for 6% of the time. How much of the time do you work flat out? Not much, huh? If you worked at maximum performance all the time, you’d soon become burnt out. The same is true for the variable $V_{DD}$ processor. Highest performance, though, defines your capability.

### Interconnect crisis

The interconnect crisis is shown in Fig. 10. The cost, delay, power, reliability, and turnaround time of the future VLSIs will be determined not by transistors but by interconnects. There are many design issues for deep-submicron interconnects. The higher current leads to static and dynamic $IR$ voltage drop problems and reliability degradation due to electromagnetic interference. The smaller geometry and denser pattern lead to an increase in $RC$ delay and signal integrity problems such as high crosstalk noise and large delay fluctuation due to capacitive coupling among adjacent lines. The higher speed causes inductance-related issues and electromagnetic interference problems.

A huge operating current will be required in the future for high-performance VLSIs as shown in Fig. 5. Such a high current creates a voltage drop due to the resistance of the power supply lines. Even a medium-power-consuming chip will need a very thick metal layer, like 10 $\mu$m, to keep the $IR$ drop within an acceptable level as shown in Fig. 11. This thick metal will be implemented in a package. In the future, area pads and co-design of a VLSI and its package will become necessary.

Signal integrity is becoming one of the major design issues due to the increased coupling capacitance between interconnects. The higher aspect ratio of deep sub-micron interconnects increases the coupling capacitance among lines, relative to the grounding capacitance as shown in Fig. 12. The delay of an interconnect may fluctuate by a factor of about 4 between in-phase and anti-phase driving of adjacent lines (see Fig. 13)\[^{[6]}\]. This forces designers to think about the voltage behaviors of adjacent lines in addition to the delay of the target interconnect, which is a nightmare. The delay fluctuation due to the coupling capacitance can be mitigated by using a buffer insertion technique as shown in Fig. 14. The fluctuation will be further reduced by staggering the locations of buffers even in the worst case as shown in Fig. 15\[^{[7]}\].

Efforts are being made to lower the resistance and capacitance of interconnects as...
shown in Fig. 16. Still, the interconnect delay is a big headache in designing a scaled-down interconnect system. If we use the interconnect with the minimum cross-section, the signal cannot propagate a distance of 1 mm in one clock cycle as shown in Fig. 17. The RC delay problem can be mitigated by using the buffer insertion technique already described in Fig. 14. The delay can be reduced by this technique as shown in Fig. 18, but the power increases by about 70% due to the inserted buffers, which will be described in more detail below. Another way to decrease the interconnect delay without increasing the power is to use a thicker and wider metal layer as in Fig. 16 using the superconnect technology described below. If a thick metal layer is available, which could be a layer in a package, by using a 6 μm x 6 μm cross-section interconnect, the RC delay can be reduced to the point where the signal can propagate anywhere within the chip within one clock cycle as shown in Fig. 19. This approach does not increase capacitance and hence power in contrast to the buffer insertion approach. The drawback is the density.

Here, I would like to add an important piece of information concerning the RC delay of an interconnect and its behavior due to the scaling. It is known that inserting buffers (or sometimes they are called repeaters) can lower the delay of a long interconnect. Let us think about the delay of a buffered interconnect system. The delay of an unbuffered interconnect can be approximately expressed as:

$$\text{Delay}_{\text{unb}} = \frac{C_{\text{INT}}}{R_{\text{INT}}} + \frac{C_{\text{T}}}{R_{\text{T}}}$$

where $C_{\text{INT}}$ is the capacitance of the interconnect and $R_{\text{INT}}$ is its resistance, $C_{\text{T}}$ is the gate capacitance of the load, and $R_{\text{T}}$ is the drain effective resistance of the driving transistor (see Fig. 14). If the interconnect is divided into $k$ sections and $(k-1)$ buffers are inserted, the total delay of the buffered interconnect system is expressed as:

$$\text{Delay}_{\text{buf}} = k\left[\frac{C_{\text{INT}}}{R_{\text{INT}}} + \frac{h}{n} h \frac{C_{\text{T}}}{R_{\text{T}}} + \frac{R_{0}}{n} h C_{0}\right]$$

where $h$ denotes the size of the inserted buffer, $C_{0}$ is the gate capacitance of the minimum width transistor, and $R_{0}$ is the gate effective resistance of the minimum width transistor. Here, $k$ and $h$ should be optimized to minimize the delay expressed in the above formula. By differentiating with respect to $h$ and $k$, and setting the derivatives equal to zero, we can easily obtain the optimum $h$, $k_{\text{OPT}}$, and optimum $k$, $k_{\text{OPT}}$.

Then the optimized delay is expressed as:

$$\text{Delay}_{\text{OPT}} = 2\left[p_{1} R_{\text{INT}} C_{\text{INT}} + p_{2} R_{0} C_{0}\right]$$

where $p_{1}$ and $p_{2}$ are fitting parameters.

Here, $(\tau_{\text{INT}} = R_{\text{INT}} C_{\text{INT}})$ is a time constant of the interconnect and $(\tau_{\text{MOS}} = R_{0} C_{0})$ is a time constant of the inserted buffer, which is proportional to the logic gate delay of a certain technology node; $p_{1}$ is 0.377 and $p_{2}$ is 0.693 for the case of the delay from zero to a half $V_{DD}$, but even if these values are different, optimization is possible for the delay from zero to 0.9 $V_{DD}$ or to other intermediate values. In this sense, the formula is quite general. The above...
expression is interesting in that the delay of the buffered interconnect system is the geometric mean of the interconnect delay itself and the delay of the logic gate. Since the scaling factor of the interconnect delay is almost constant and the delay of the logic gate is expected to improve very rapidly as technology advances, scaling of the delay of the buffered interconnect system should improve slowly along with the speed improvement of the logic gates.

In the optimally buffered interconnect, the capacitance of the system increases due to the inserted buffers. The total gate capacitance of buffers is expressed as

$$C_{\text{total}} = \sum C_i$$

This means that the total capacitance is 73% higher than that of a system without buffers. The increase in capacitance in turn increases power consumption.

**Complexity crisis**

It is quite impossible to design a VLSI with billions of transistors from scratch. The complexity crisis can only be solved by sharing design data and re-using it. By doing so, we can design an electronic system at a higher level of abstraction. The so-called IP (Intellectual Property)-based System-on-a-Chip (SoC) design style will be preferable. Here, the IP is transferable design data related to VLSIs. The virtual components are put together on a silicon chip to build billion-transistor VLSIs, which can be compared to the present system implementation with separately packaged VLSI components and printed circuit boards.

However, SoC issues have been getting clear as the VLSI industry has pursued extensively the SoC. Some issues are undistributed IPs (i.e., CPU, DSP of a certain company), huge initial investment for masks and development, IP testability, upfront IP test cost, process-dependent memory IPs, difficulty in embedding high-precision analog IPs due to noise, and process incompatibility with non-Si materials and/or MEMS. The mask count increases greatly if different types of technologies are included on a single chip as shown in Fig. 20. Moreover, the embedding technologies should be developed for each generation and if the types of technologies are diverse, the required engineer-
predictions are taken from the ITRS [1]. The overall future prospects for VLSIs in 2014 are shown in Fig. 24. Sensors and actuators can be built on a chip. The design rule will be around 0.035 \( \mu m \), integrating more than 3 billion silicon transistors on a chip. Some mechanism to control \( V_{TH} \) and \( V_{DD} \) will be necessary to cope with the high power consumption problem. The clock frequency is predicted to be around 17 GHz locally and globally the chip will be operated asynchronously. VLSIs, packages, and higher assembly structures will be co-designed to raise the performance of electronic systems.

**VLSIs in coming years**

Some of the important parameters related to VLSIs in 2014 are summarized in Table 2. These

![Fig.24 Possible electronic system in the future.](image)

**Table 2** Predicted important parameters related to VLSIs in 2014. Factor is the ratio between the values in 2014 and those in 1999.

<table>
<thead>
<tr>
<th>Year</th>
<th>Unit</th>
<th>1999</th>
<th>2014</th>
<th>Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design rule</td>
<td>( \mu m )</td>
<td>0.18</td>
<td>0.015</td>
<td>0.2</td>
</tr>
<tr>
<td>Tr. Density</td>
<td>( /cm^2 )</td>
<td>6.2M</td>
<td>190M</td>
<td>10</td>
</tr>
<tr>
<td>Chip size</td>
<td>( mm^2 )</td>
<td>140</td>
<td>900</td>
<td>6.6</td>
</tr>
<tr>
<td>Tr. Count per chip</td>
<td>(pF)</td>
<td>21M</td>
<td>3.6G</td>
<td>170</td>
</tr>
<tr>
<td>DRAM. capacity</td>
<td></td>
<td>1G</td>
<td>1T</td>
<td>1000</td>
</tr>
<tr>
<td>Local clock on a chip</td>
<td>Hz</td>
<td>1.2G</td>
<td>17G</td>
<td>14</td>
</tr>
<tr>
<td>Global clock on a chip</td>
<td>Hz</td>
<td>1.2G</td>
<td>3.7G</td>
<td>3.1</td>
</tr>
<tr>
<td>Power</td>
<td>W</td>
<td>90</td>
<td>183</td>
<td>2.0</td>
</tr>
<tr>
<td>Supply voltage</td>
<td>V</td>
<td>1.5</td>
<td>0.17</td>
<td>0.2</td>
</tr>
<tr>
<td>Current</td>
<td>A</td>
<td>60</td>
<td>494.6</td>
<td>8</td>
</tr>
<tr>
<td>Interconnection levels</td>
<td></td>
<td>6</td>
<td>10</td>
<td>1.7</td>
</tr>
<tr>
<td>Mask count</td>
<td></td>
<td>22</td>
<td>28</td>
<td>1.3</td>
</tr>
<tr>
<td>Cost /tr. (packaged)</td>
<td>( \mu )cents</td>
<td>175S</td>
<td>22</td>
<td>0.01</td>
</tr>
<tr>
<td>Chip to board clock</td>
<td>Hz</td>
<td>500M</td>
<td>1.5G</td>
<td>3.0</td>
</tr>
<tr>
<td># of package pins</td>
<td></td>
<td>810</td>
<td>2700</td>
<td>3.3</td>
</tr>
<tr>
<td>Package cost</td>
<td>( cents /pin )</td>
<td>1.61</td>
<td>0.75</td>
<td>0.5</td>
</tr>
</tbody>
</table>

---

**References**


