Past, Present, and Future

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

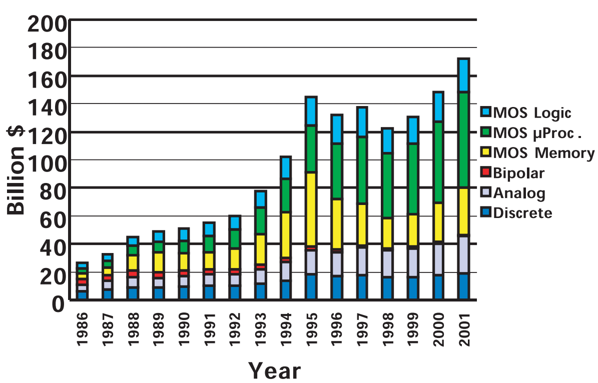


Fig. 1 World semiconductor market. Data is taken from world semicon market statistics.

Transistors		Scaling coefficients	
V_{DD}	[V]	1/k	
Tr. dimensions	[x]	1/k	
Drain current	[I~1/x x/x V^1.3]	$1/k^{0.3}$	
Gate capacitance	[C~1/x xx]	1/k	
Tr. delay	[d~CV/I]	$1/k^{1.7}$	
Tr. power	[P~VI~CVV/d]	$1/k^{1.3}$	
Power density	**[p~P/x/x]**	$k^{0.7}$	
Tr. density	**[n~1/x/x]**	k^2	
Interconnects			
Type Scaling scenario		**Local Scaled**	**Global Anti-scaled**
Line thickness	[T]	1/k	k
Width	[W]	1/k	k
Separation	[S]	1/k	k
Oxide thickness	[H]	1/k	1
Length	[L]	1/k	1
Resistance	[R_{INT}~L/W/T]	k	$1/k^2$
Capacitance	[C_{INT}~LW/H]	1/k	k
RC delay/Tr. delay	**[D~$R_{INT}C_{INT}$/d]**	$k^{1.7}$	-
Current density	**[J~pWL/V /W/T]**	-	$k^{0.7}$
DC noise /V_{DD}	**[N~JWTR/V]**	-	$k^{1.7}$

$$I_{DS} = \frac{\mu\varepsilon}{t_{OX}}\left(\frac{W}{L_{eff}}\right)\frac{(V_{GS}-V_{TH})^{\alpha}}{2}$$

(α ~ 1.3 for short-channel MOSFET)

Fig. 3 Scaling law. k denotes a scaling variable, which is 2 when the size is scaled down to a half. For example, power density increases by a factor of $2^{0.7}$ (~1.6) when the device size is shrunk to a half.

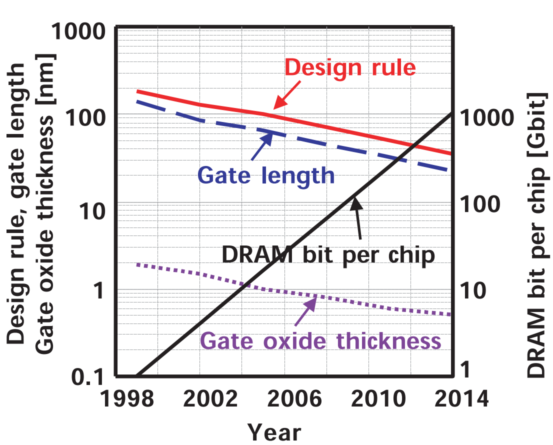


Fig. 2 Trend in design rule of VLSI, gate length of MOS transistors, gate oxide thickness and DRAM bit capacity per chip. Data is taken from ITRS'99 [1].

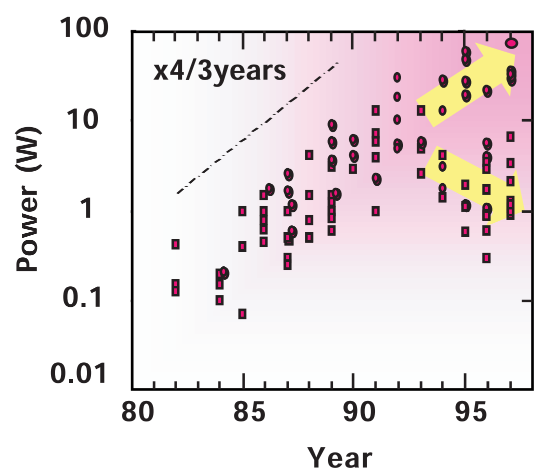


Fig.4 Ever Increasing VLSI Power. Data points are taken from processors reported at the International Solid-State Circuits Conference.

has now shrunk to 0.13μm and the gate oxide is only a few SiO_2 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

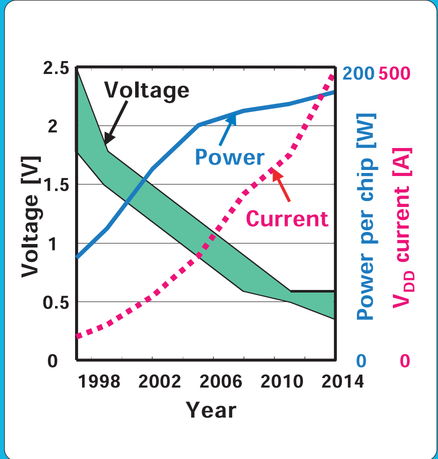


Fig.5 Supply voltage, power and current trend.

Low-voltage
- Multi-V_{TH} , variable V_{TH}
- Multi-V_{DD} , variable V_{DD}
- Ultra low voltage circuit (PLL, less analog, SOI)

Low-swing
- Bus, clock

Low-C_L
- Less # of Tr's,fused digital-analog
- Gate-sizing, low-power cell library
- Low-k (air isolation)
- System on a chip, memory embedding

Low-$p_t f$
- Multi-f, gated clock, low transition coding
- Less glitches (20% power by glitches)

Fig.7 Important technologies for low-power

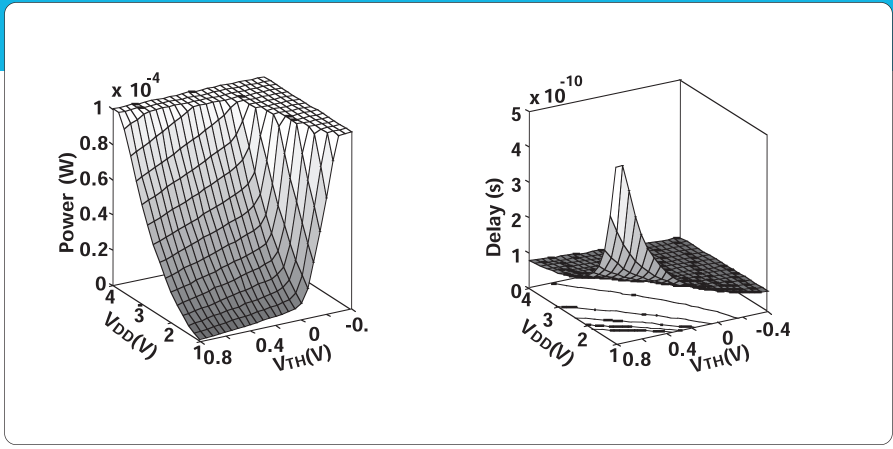


Fig.6 Power and delay dependence on V_{DD} and V_{TH}

	Active	Stand-by
Multiple V_{TH}	Dual-V_{TH}	(MTCMOS)
Variable V_{TH}	VTCMOS Speed adaptive V_{TH} Software V_{TH} cont.	(VTCMOS)
Multiple V_{DD}	Dual-V_{DD}	Boosted gate MOS
Variable V_{DD}	DVS, V_{DD} hopping	Low V_{DD} at sleep

Table 1 Controlling V_{DD} and V_{TH} for low power

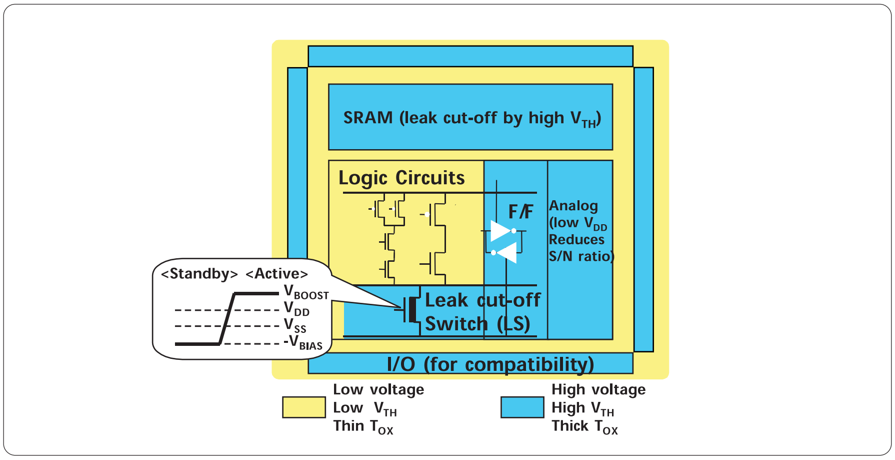


Fig.8 Device-circuit cooperative approach for low power. Multiple use of V_{DD}, V_{TH} and t_{ox} is the key.

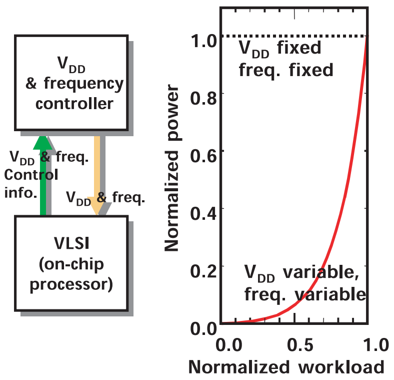


Fig.9 V_{DD} hopping for low-power

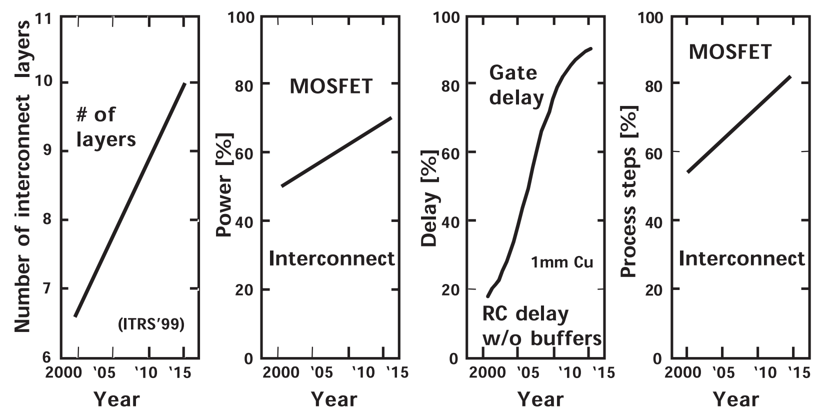


Fig.10 Interconnect determines cost and performance.

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

$$\text{Power} = p_t \cdot f \cdot C_L \cdot V_{DD}^2 + I_0 \cdot 10^{-\frac{V_{TH}}{s}} V_{DD},$$

$$\text{Delay} \propto \frac{Q}{I} = \frac{C_L V_{DD}}{(V_{DD} - V_{TH})^{\alpha}}.$$

where p_t is switching probability, C_L is load capacitance, V_{TH} is threshold voltage, α 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/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.

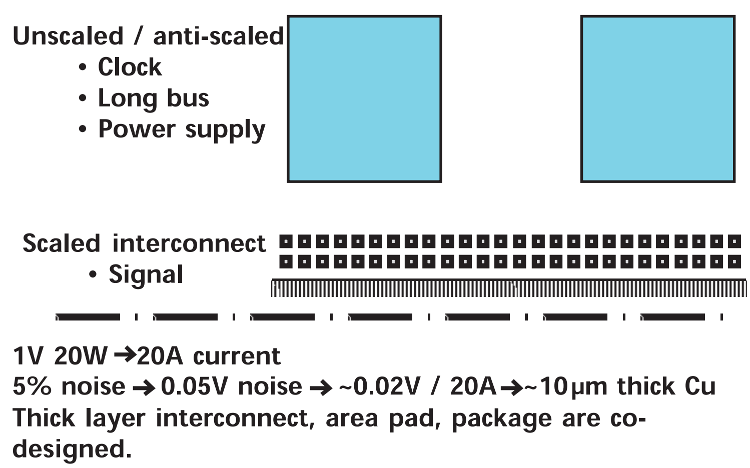


Fig.11 Interconnect cross-section and *IR* drop.

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 maxi-

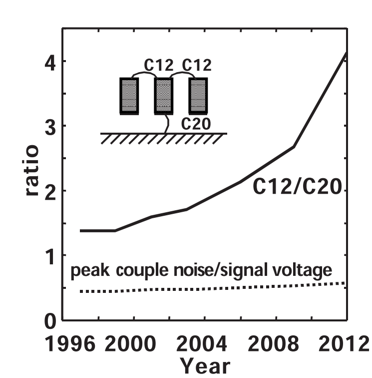
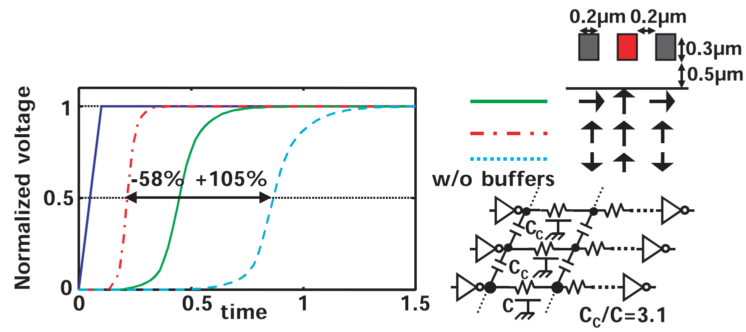


Fig.12 Trend in coupling capacitance.



$$\frac{t_{pd}}{RC}=\frac{1}{2}+2\eta-\frac{1}{\sqrt{6}}\log\frac{e}{2}\sqrt{1+8\eta+6\eta^2}$$
$$\approx 1.63\eta+0.37(\eta\leq 2)\ (\eta=C_c/C)$$

Fig.13 Delay fluctuation due to coupling capacitance [6]

Fig.14 Delay reduction by buffer insertion. a) without buffers and b) with buffers.

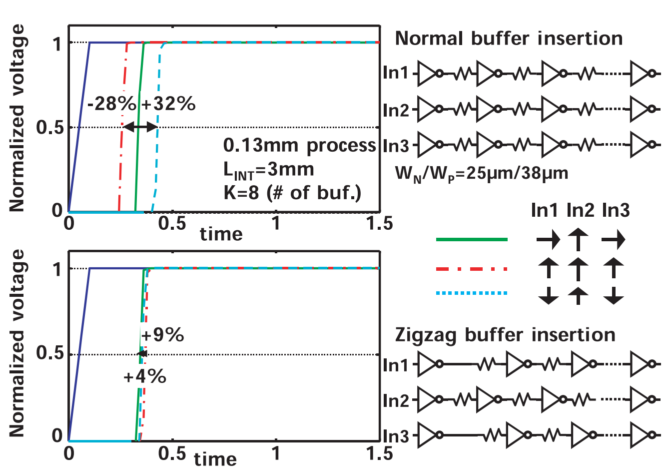


Fig.15 Delay fluctuation due to coupling capacitance with non-staggered and staggered buffers.

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 electro-migration. 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 μm, 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

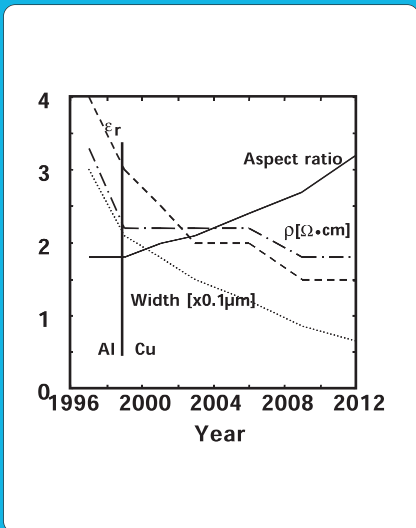


Fig.16 Trend in interconnect related parameters. Data is taken from ITRS'97.

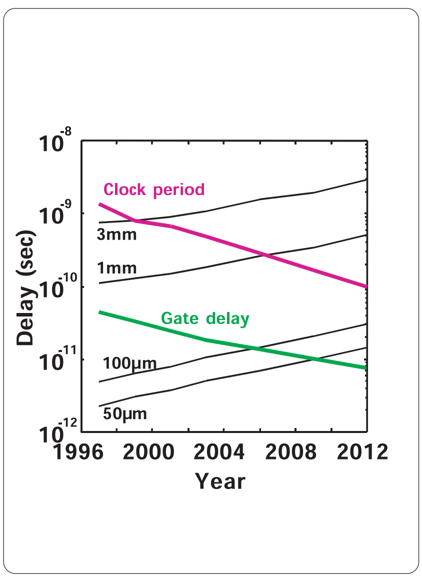


Fig.17 *RC* delay and gate delay.

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 a 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

$t_{05} \approx 0.377 R_{INT} C_{INT} + 0.693 (R_T C_T + R_T C_{INT} + R_{INT} C_T)$,

where C_{INT} is the capacitance of the interconnect and R_{INT} is its resistance, C_T is the gate capacitance of the load, and R_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

$$\text{Delay} \approx k\left[p_1 \frac{R_{INT}}{k}\frac{C_{INT}}{k} + p_2\left(\frac{R_0}{h} hC_0 + \frac{R_0}{h}\frac{C_{INT}}{k} + \frac{R_{INT}}{k} hC_0\right)\right],$$

where h denotes the gate 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, h_{OPT}, and optimum k, k_{OPT}.

$$\frac{\partial \text{Delay}}{\partial h} = 0 \rightarrow h_{OPT} = \sqrt{\frac{C_{INT} R_0}{R_{INT} C_0}}$$

$$\frac{\partial \text{Delay}}{\partial k} = 0 \rightarrow k_{OPT} = \sqrt{\frac{p_1}{p_2}} \sqrt{\frac{R_{INT} C_{INT}}{R_0 C_0}}$$

Then the optimized delay is expressed as

$$\text{Delay}_{OPT} = 2(\sqrt{p_1 p_2} + p_2)\sqrt{R_{INT} C_{INT} R_0 C_0}$$
$$\approx 2.4\sqrt{\tau_{INT}\tau_{MOS}}$$

Here, (τ_{INT} (=$R_{INT}C_{INT}$) is a time constant of the interconnect and (τ_{MOS} (=R_0C_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

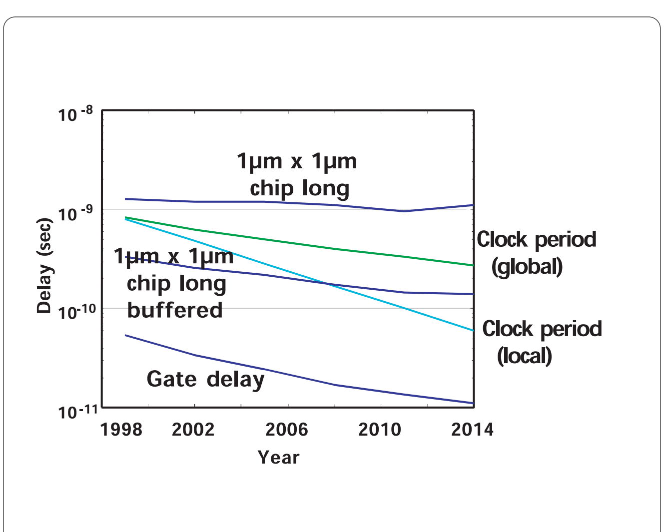


Fig.18 Trend in interconnect delay with buffer insertion.

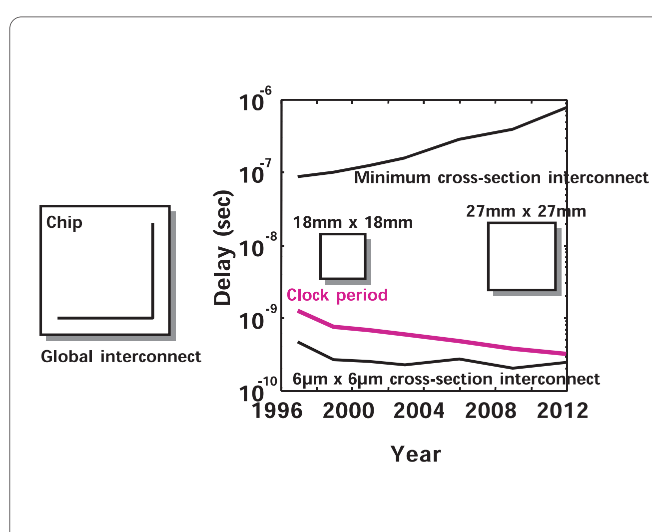


Fig.19 *RC* delay of global interconnects.

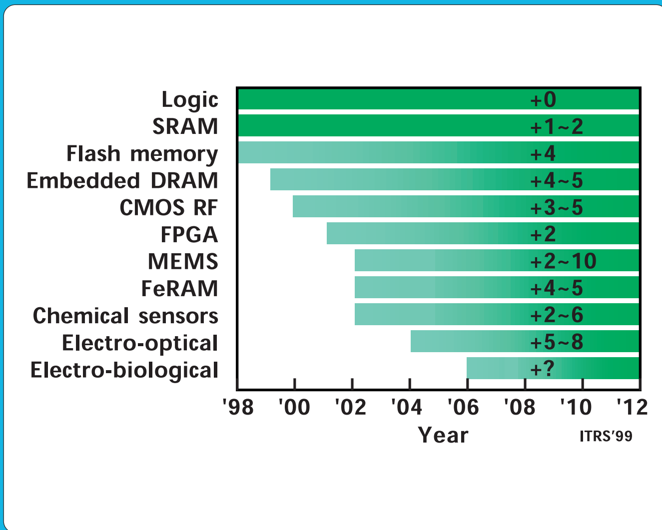


Fig.20 Technologies integrated on a chip. The numbers in the bars show the increase of mask steps extra to a logic process.

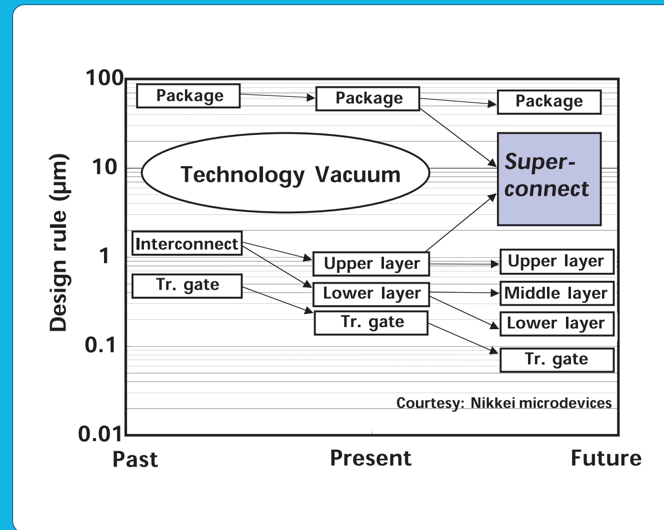


Fig.22 Super-connect technology.

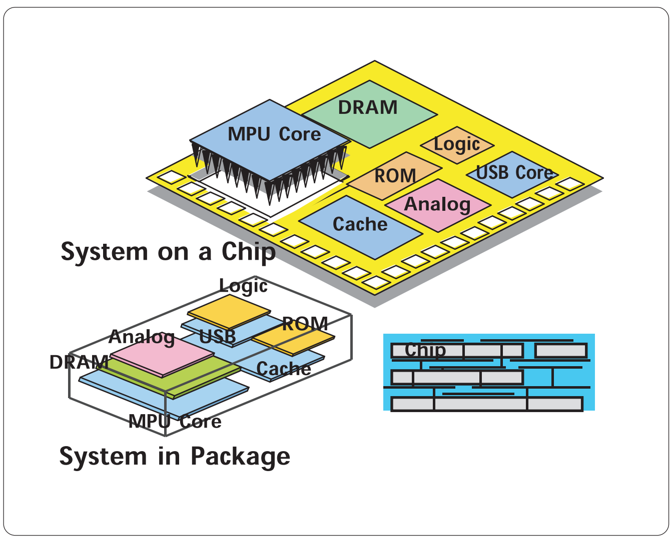


Fig.21 System-on-a-Chip vs. System-in-a-Package.

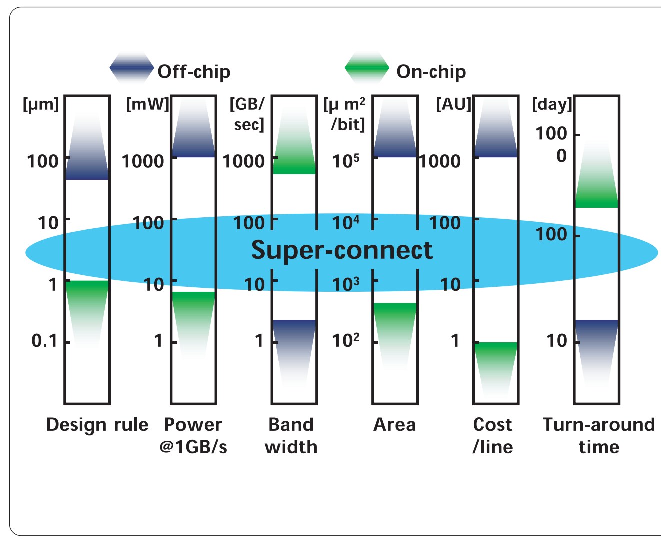


Fig.23 Performance gap between off-chip and on-chip interconnects.

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

Total capacitance of gates =

$$k_{OPT}h_{OPT}C_0=\sqrt{p_1/p_2C_{INT}}=0.73C_{INT}.$$

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-

Year	Unit	1999	2014	Factor
Design rule	µm	0.18	0.035	0.2
Tr. Density	/cm2	6.2M	390M	30
Chip size	mm2	340	900	2.6
Tr. Count per chip (µP)		21M	3.6G	170
DRAM capacity		1G	1T	1000
Local clock on a chip	Hz	1.2G	17G	14
Global clock on a chip	Hz	1.2G	3.7G	3.1
Power	W	90	183	2.0
Supply voltage	V	1.5	0.37	0.2
Current	A	60	494.6	8
Interconnection levels		6	10	1.7
Mask count		22	28	1.3
Cost / tr. (packaged)	µcents	1735	22	0.01
Chip to board clock	Hz	500M	1.5G	3.0
# of package pins		810	2700	3.3
Package cost	cents/pin	1.61	0.75	0.5

Table 2 Predicted important parameters related to VLSI's in 2014. Factor is the ratio between the values in 2014 and those in 1999.

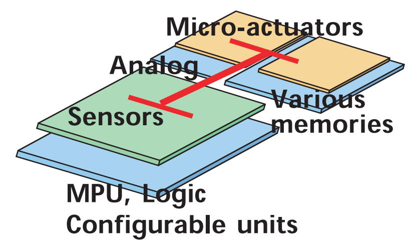


Fig.24 Possible electronic system in the future.

ing efforts are too big to sustain.

Another way to cope with the complexity crisis is to make use of an advanced assembly technology. Here, components are real components instead of virtual components and we can design an electronic system by assembling the real components; that is, we can design in higher abstraction and thus escape from the complexity crisis. A new type of assembly technology called System-in-a-Package has been proposed as shown in Fig. 21. This will use 'super-connect' technology, as shown in Fig. 22, with an interconnect thickness on the order of 10 µm. Super-connect technology will fill the technology vacuum between the design rule orders of 1 µm for on-chip interconnects and 100 µm for off-chip interconnects. Super-connects in a package used in cooperation with on-chip interconnects will solve the *IR* voltage drop problem, the clock distribution problem, and other problems of future VLSIs. The co-design of on-chip interconnects and super-connects in a package will become important, but it will require the development of a new set of EDA tools.

The super-connect technology fills the gap between off- and on-chip interconnects not only in terms of design rules but also in terms of power, bandwidth, area, cost, and turn-around-time, as shown in Fig. 23. The major issue in making the System-in-a-Package, however, is to establish a method for selecting known good dies before assembly. It is very difficult to test a bare chip at its operating speed without a package, since probing needles used for the wafer test cannot handle signals more than a hundred MHz. Recently, however, a new test method using a pseudo-package called an interposer has been proposed. The pseudo-package enables a chip to be tested at full operating speed, which may solve the known good die problem. Assembly and packaging technology is becoming vital to VLSIs as evidenced by the following passage from the International Technology Roadmap for Semiconductors (ITRS): "There is an increased awareness in the industry that assembly and packaging is becoming a differentiator in product development."

VLSIs in coming years

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

JSAP

References

[1] International Technology Roadmap for Semiconductors: 1999 edition. Austin, TX: International SEMATECH,1999.

[2] T. Sakurai and A. Newton, "Alpha-power law MOSFET model and its application to CMOS inverter delay and other formulas," IEEE JSSC, vol. 25, no. 2, pp. 584-594, Apr. 1990.

[3] T. Inukai, M. Takamiya, K. Nose, H. Kawaguchi, T. Hiramoto and T. Sakurai, "Boosted Gate MOS (BGMOS): Device/Circuit Cooperation Scheme to Achieve Leakage-Free Giga-Scale Integration," CICC'00, p. 409, May 2000.

[4] S. Lee and T. Sakurai, "Run-time Power Control Scheme Using Software Feedback Loop for Low-Power Real-time Applications,"ASPDAC'00, A5.2, pp. 381-386, Jan. 2000.

[5] S. Lee and T. Sakurai, "Run-time Voltage Hopping for Low-power Real-time Systems," DAC'00, June 2000.

[6] H. Kawaguchi and T. Sakurai, "Delay and Noise Formulas for Capacitively Coupled Distributed RC Lines," 1998 ASPDAC, Digest of Tech. Papers, pp. 35-43, Feb. 1998.

[7] Koichi Nose, private communications.