Low-power LSI
- Through cooperation among levels -

Takayasu Sakurai
Center for Collaborative Research, and Institute of Industrial Science, University of Tokyo
E-mail: tsakurai@iis.u-tokyo.ac.jp

Abstract

In this paper, methods to achieve low-power and high-speed VLSI's are described with the emphasis on cooperation between levels. To suppress the leakage current in a standby mode, Boosted Gate MOS (BGMOS) is effective, which is based on cooperation between technology level and circuit level. To reduce the power in an active mode, V_{DD}-hopping and V_{TH}-hopping are promising, which are cooperative approaches between circuit and software. Power consumed in interconnect system is another issue in low-voltage deep-submicron designs. A cooperative approach between VLSI and assembly to the interconnect power problem is also discussed.

1. Introduction

Power consumption of VLSI's is ever increasing as is shown in Fig.1 and could be a stumbling block in realizing giga-scale integration. The power consumption of CMOS VLSI's consists of a dynamic charging and discharging current component and a leakage current component shown in Fig.2. Various effective techniques to mitigate the power problem have been proposed at a level of system, algorithm, software, CAD, circuit, technology and assembly. There still remains effective ways to reduce the power consumption through cooperation between levels, which have not been pursued so rigorously because usually it is not easy to establish a cooperative mechanism among different levels of engineers. The treasured sword of CMOS scaling, however, may come to an end due to the power problem and every endeavor to decrease power consumption will be required in the future.

The supply voltage, V_{DD}, is ever decreasing to ensure sufficient reliability of VLSI's with thin gate oxide used in deep submicron transistors. Low-power design in the low- V_{DD} environments is a battle against the ever-increasing leakage current due to the use of low threshold voltage, V_{TH}, to realize high-speed VLSI's. Since low-V_{DD} and high-V_{TH} are advantageous for low power consumption but disadvantageous for speed, there should be some controlling scheme among V_{DD}, V_{TH} and speed. Most of the controlling schemes can be categorized in a tabulated form as in Fig.3. In the table, multiple V_{DD} and multiple V_{TH} signify spatial assignment of V_{DD} and V_{TH}, while variable V_{DD} and variable V_{TH} correspond to temporal assignment of V_{DD} and V_{TH}.

Historically well-known techniques to control V_{DD} and V_{TH} are MTCMOS (Multi-Threshold CMOS) and VTCMOS (Variable Threshold CMOS) [1-3] but these schemes will not be able to stand in the future as shown in Fig.4. In this paper, Boosted Gate MOS (BGMOS) is described to cope with the standby current increase due to the use of low V_{TH}, which is a cooperative approach between a technology level and a circuit level. As for the power reduction in an active mode, V_{DD}-hopping and V_{TH}-hopping are introduced, which are cooperative approaches between circuit and software. Lastly, the paper touches on an interconnect power problem by buffer insertion.

2. Cooperation between technology and circuit: BGMOS

In order to mitigate the leakage problem in a standby mode, it is effective to insert a non-leaking power switch in series to a leaky yet high-speed logic gate block made of low-V_{TH} MOSFET's. The basic idea is the same as MTCMOS but MTCMOS becomes slow when V_{DD} gets less than 1V and stops operating when V_{DD} gets less than 0.5V. This is because the non-leaking power switch can be realized by a high-V_{TH} (0.6V for example) MOSFET that is turned on in an active mode and turned off in a standby mode.

The problem of MTCMOS is solved if higher voltage around 1.5V-2.0V is applied to the gate of the power switch MOSFET. The higher voltage achieves higher drivability and hence higher speed. This scheme is called boosted gate MOS scheme (BGMOS) as shown in Fig.6 [4]. The power switch should have higher oxide thickness to accommodate the higher voltage on the gate. A technology side provides a thicker oxide transistor, while designers think about using the different type of transistors and thus the scheme can be said cooperation between a technology level and a circuit level. MOSFET's tuned for the higher voltage is also helpful in SRAM, I/O and analog designs as shown in Fig.8. The thicker oxide is also beneficial to decrease leakage current caused by direct tunneling though oxide of the power switch (see Fig.5).

It should be noted that there is an optimum thickness for the oxide of the power switch and thus there is optimal gate voltage which is 1.5V-2.0V as shown in Fig.7. This is due to the increase of channel length to suppress a short channel effect when the oxide gets thicker.

3. Cooperation between circuit and software: V_{DD} hopping and V_{TH} hopping

In an active mode, changing V_{DD} and V_{TH} in time in accordance with required performance at every moment is effective for power reduction. If V_{DD} is lowered or V_{TH} is increased, the power decreases but speed is degraded. The difficulty is to find the timing to lower the speed. Hardware cannot know the timing when decreasing the performance of a processor does not affect the system performance. Only software knows when it is possible to decrease the processor performance without sacrificing the system performance. Software should tell hardware when higher performance is needed.

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 lowered to a half when high-speed operation is not needed (see Fig.9). The scheme has been applied to a MPEG4 codec system and the processor power has been reduced to one fourth of the conventional fixed V_{DD} processor [5, 6] (see Fig.11-15). The video codec system guarantees real-time operation for any data input but the highest performance is needed only for 6% of time (see Fig.14).

The algorithm to adaptively change V_{DD} depending on the workload is of importance. Since the workload depends strongly on data, the control should be dynamic in run-time, and should not be static in a compile-time. It is too late to notice that the past task is an easy task which can be done by using the lower voltage because once the task is completed by using the high voltage, energy has been already consumed. On the other hand, it is impossible to predict the workload of the task to be done in the future without error. To solve this problem, the algorithm introduces an application slicing and a software feedback loop. By chopping an application into slices, executing the first slices at the maximum frequency, and checking the current time and the time margin to execute the next slice, the optimum clock frequency and V_{DD} adaptively selected by a software feedback loop. (see Fig.10)

It is to be noted that V_{DD} hopping algorithm works fine for every multimedia application we tried although the switching time between voltage levels requires 0.2ms which is considered to be extraordinary long in terms of processor clock period. In a multimedia application, however, the real-time feature is for humans and human is slow. This is the reason why the V_{DD} hopping works fine in spite of the long transition time between voltage levels. The other point of interest is that the number of voltage levels can be as low as two as is shown in Fig.11.

The V_{DD} hopping scheme can also be applied to multi-tasking real-time operating system [7] (see Fig.16). Since OS knows higher-level information on available time slot assignable to an application, higher efficiency can be realized than application-only case as shown in Fig.16. One example we tried is modified power-conscious μ-ITRON OS running FFT and MPEG4 at the same time and the observed power reduction was 75% while the power saving for FFT alone was only 50%.

When subthreshold leakage becomes dominant in the future, the same software control mechanism can be used in V_{TH} hopping scheme where V_{TH} is changed in time in accordance with the required performance [8] (see Fig.17-19). About 80% power reduction is possible for a multimedia real-time application. It is found that time-domain assignment of V_{TH} is more effective than spatial assignment of V_{TH} for multimedia video application as shown in Fig.20.

4. Cooperation between VLSI and assembly: Superconnect

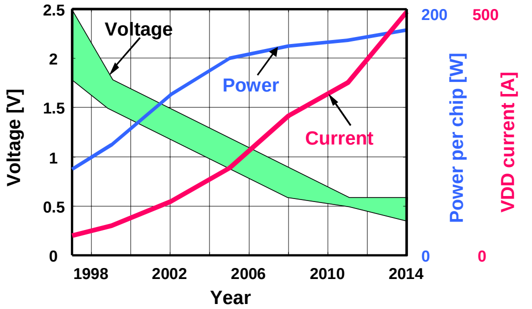
In deep submicron designs, RC delay of interconnect hinders a high-speed operation of VLSI's. A widely used remedy for the interconnect RC delay is a buffer insertion technique. The basic optimization theory is summarized in Fig.21. The drawback of the buffer insertion, however, is that the delay optimization by buffer insertion increases power consumption by 73% as shown in Fig.22. With the use of thicker layer of metals provided by superconnect [9,10], it is possible to decrease interconnect delay without increasing power consumption because buffer insertion is not needed with low resistance interconnects. Co-design between an LSI itself and an assembly structure such as an interposer and a package will be needed. High current expected in low-V_{DD} regime shown in Fig.1 can also be mitigated by the use of the thicker metal layer in an assembly body and small area pads on an LSI.

5. Summary

Power consumption of LSI's tend to increase due to the scaling law and due to the leakage increase including sub-threshold, gate tunneling, and junction leakage.

New trend for low-power LSI's is to pursue cooperative approaches among levels: BGMOS to cut-off standby leakage, V_{DD} / V_{TH} hopping to reduce operating power, and super-connect to reduce I/O power are some of the examples.

One of the biggest barriers to the scaling is the leakage power increase and solutions are yet to be discovered, though several good trials have been made.



International Technology Roadmap for Semiconductors 1999 update sponsored by the Semiconductor Industry Association in cooperation with European Electronic Component Association (EECA) , Electronic Industries Association of Japan (EIAJ), Korea Semiconductor Industry Association (KSIA), and Taiwan Semiconductor Industry Association (TSIA)

Fig.1 V_{DD}, Power and Current Trend

In deep submicron designs, RC delay of interconnect

References

[1] S. Mutoh, et al., "1V High-Speed Digital Circuit Technology with 0.5um Multi-Threshold CMOS," in Proc. IEEE 1993 ASIC Conf., 1993, pp. 186-189.

[2] T.Kuroda, et al., "A 0.9V 150MHz 10mW 4mm^2 2-D Discrete Cosine Transform Core Processor with Variable-Threshold-Voltage Scheme," in ISSCC, pp. 166-167, Feb. 1996.

[3] H.Mizuno, K.Ishibashi, T.Shimura, T.Hattori, S.Narita, K.Shiozawa, S.Ikeda and K.Uchiyama, "A 18uA-Standby-Current 1.8V 200MHz Microprocessor with Self Substrate-Biased Data-Retention Mode," 1998 ISSCC Digest of Tech. Papers, pp.280-281, Feb.1999.

[4] 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.

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

[6] S.Lee, and T.Sakurai, "Run-Time Voltage Hopping for Low-Power Real-Time Systems",Proceedings of Design Automation Conference, pp. 806-809,June 2000.

[7] Y.S.Shin, H.Kawaguchi, T.Sakurai, "Cooperative Voltage Scaling (CVS) between OS and Applications for Low-Power Real-Time Systems," CICC'01, pp.553-556, May 2001.

[8] K.Nose, M.Hirabayashi, H.Kawaguchi, S.Lee, and T.Sakurai, "V_{TH}-Hopping Scheme for 82% Power Saving in Low-Voltage Processors," CICC, May 2001.

[9] T.Sakurai, "Superconnect Technology (Invited)," Trans. C of IEICE, to be published, 2002.

[10] T.Sakurai, "Interconnection from Design Perspective," Advanced Metallization Conference, Oct. 2000.

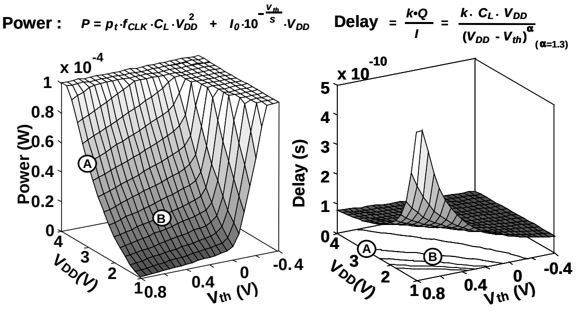


Fig.2 Power & Delay Dependence on V_{DD} & V_{TH}

Low power → Low V_{DD} → Low speed → Low V_{TH}
→ High leakage → V_{DD}-V_{TH} control

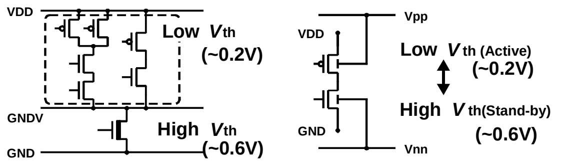
	Active	Stand-by
Multiple V_{TH}	Dual-V_{TH}	MTCMOS
Variable V_{TH}	V_{TH} hopping	VTCMOS
Multiple V_{DD}	Dual-V_{DD}	Boosted gate MOS
Variable V_{DD}	V_{DD} hopping	

Software-hardware cooperation

Technology-circuit cooperation

*) MTCMOS: Multi-Threshold CMOS
*) VTCMOS: Variable Threshold CMOS
- Multiple : spatial assignment
- Variable : temporal assignment

Fig.3 Controlling V_{DD} and V_{TH} for low power



MTCMOS [1]

- Tunneling leakage cannot be cut-off.
- Area penalty increases when VDD < 1V.

VTCMOS [2]

- Junction leakage increases due to band-to-band tunneling.
- Tunneling leakage is not suppressed.

[1] S.Mutoh *et al*. IEEE, JSSC, 1995. [2] T.Kuroda *et al*. IEEE, JSSC, 1996.

Fig.4 Previous circuit schemes

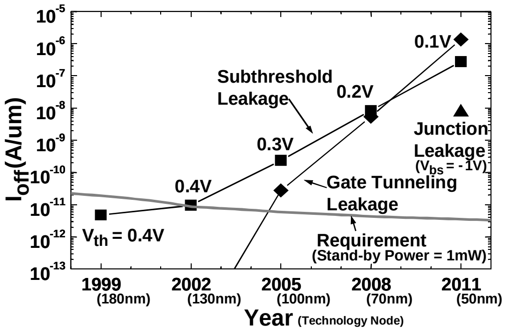


Fig.5 Transistors go leaky

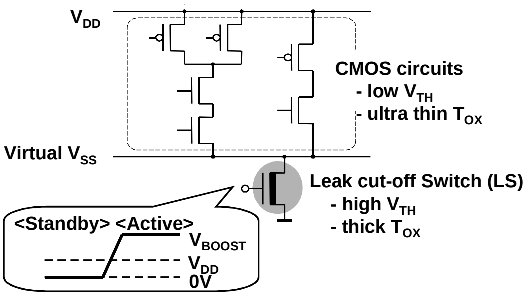


Fig.6 Boosted gate MOS scheme

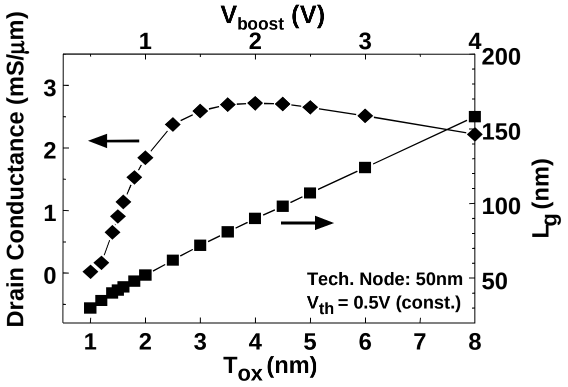


Fig.7 Power switch optimization

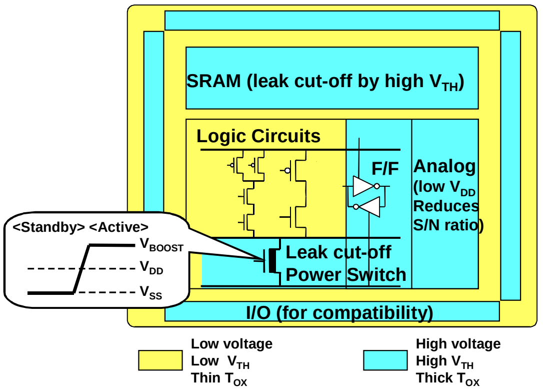


Fig.8 GSI's in deep-submicron era

Energy consumption is proportional to the square of V_{DD}.

V_{DD} should be lowered to the minimum level which ensures the real-time operation.

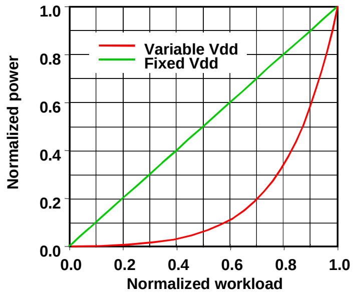


Fig.9 If you don't need to hussle,
V_{DD} should be as low as possible

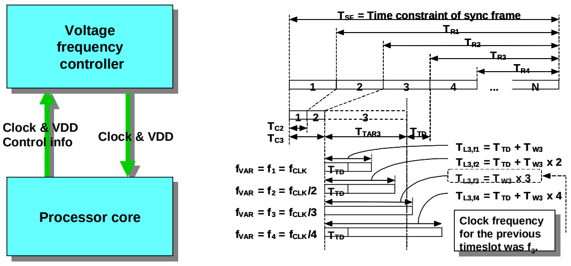

Voltage frequency controller
Clock & VDD Control info
Clock & VDD
Processor core
T_{SF} = Time constraint of sync frame
T_{R1}
T_{R2}
T_{R3}
T_{R4}
1
2
3
4
...
N
T_{C2}
T_{C3}
T_{TAR3}
T_{TD}
$T_{L3,f1} = T_{TD} + T_{W3}$
$T_{L3,f2} = T_{TD} + T_{W3}$ x 2
$T_{L3,f3} = T_{W3}$ x 3
$T_{L3,f4} = T_{TD} + T_{W3}$ x 4
$f_{VAR} = f_1 = f_{CLK}$
$f_{VAR} = f_2 = f_{CLK}/2$
$f_{VAR} = f_3 = f_{CLK}/3$
$f_{VAR} = f_4 = f_{CLK}/4$
Clock frequency for the previous timeslot was f_3.

Fig.10 Application slicing and software feedback loop in V_{DD} hopping

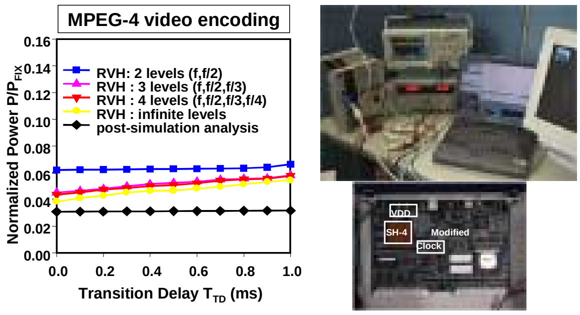


Fig.11 Run-time V_{DD} hopping reduces power to less than 1/10

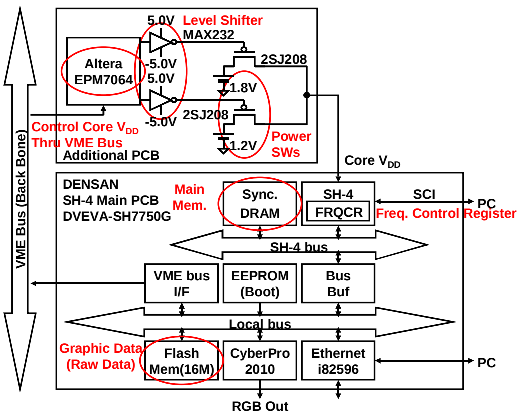


Fig.12 Block diagram

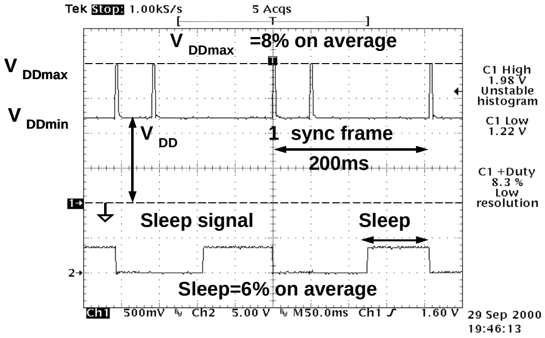
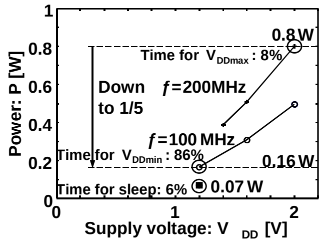


Fig.13 Measured voltage waveforms

Total power = 0.8 x 0.08 + 0.16 x 0.86 + 0.07 x 0.06 = 0.2W



VDD hopping can cut down power consumption to 1/4

Fig.14 Measured power characteristics

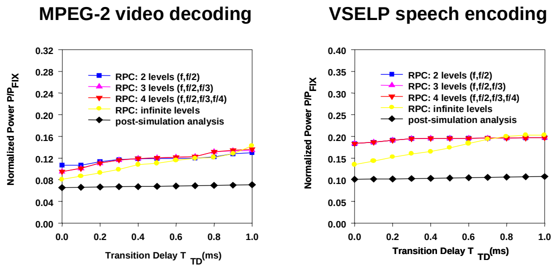


Fig.15 Simulation results

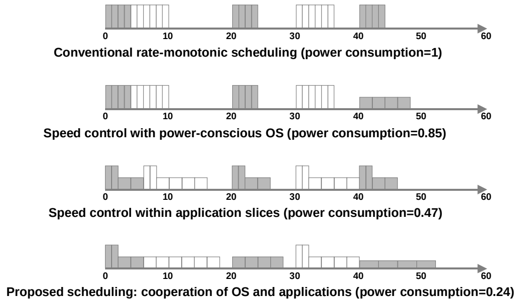


Fig.16 Power Conscious OS & Application Slicing

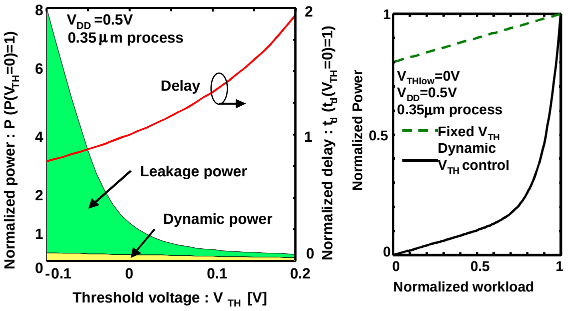


Fig.17 V_{TH}-hopping

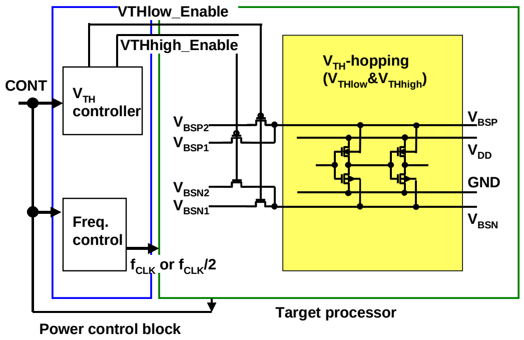
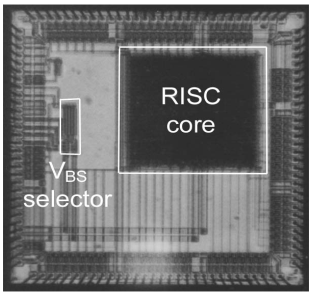


Fig.18 Schematic of V_{TH}-hopping



0.6µm process

Overhead of V_{TH}-hopping
= 14%

RISC core
= 2.1mm x 2.0mm
V_{BS} selector
= 0.2mm x 0.6mm

Fig.19 Microphotograph of RISC processor

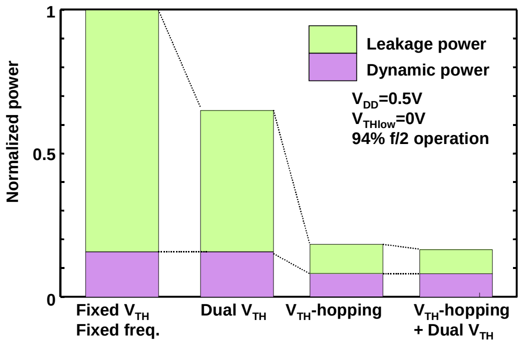
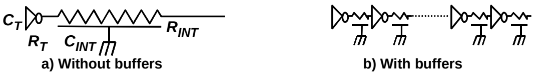


Fig.20 Power comparison



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

C_0 :*Gate capacitance of minimum MOSFET*

R_0 :*Gate effective resistance of minimum MOSFET*

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

$$\frac{\partial Delay}{\partial h} = 0 \rightarrow h_{OPT} = \sqrt{\frac{C_{INT} R_0}{R_{INT} C_0}} : \textit{Optimized size of buffer inverter}$$

$$\frac{\partial Delay}{\partial k} = 0 \rightarrow k_{OPT} = \sqrt{\frac{p_1}{p_2}}\sqrt{\frac{R_{INT} C_{INT}}{R_0 C_0}} : \textit{Optimized number of stages}$$

$$Delay_{OPT} = 2\left(\sqrt{p_1 p_2} + p_2\right)\sqrt{R_{INT} C_{INT} R_0 C_0} \approx 2.4\sqrt{\tau_{INT}\tau_{MOS}}$$

$$Cap.\ of\ gates = k_{OPT} h_{OPT} C_0 = \sqrt{p_1 / p_2}\, C_{INT} = 0.73 C_{INT}$$

Fig.21 Buffer insertion

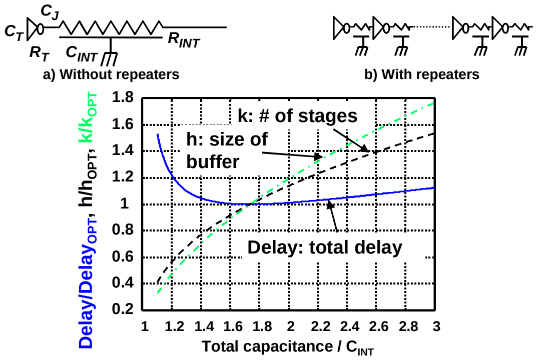


Fig.22 Power delay optimization

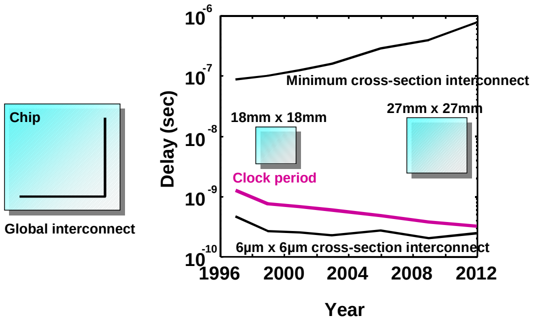
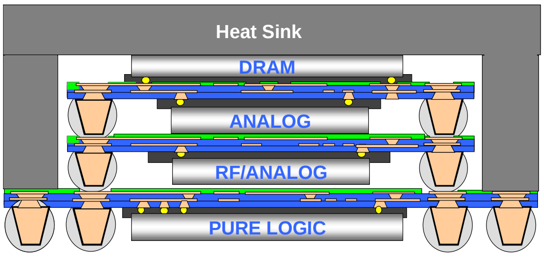


Fig.23 RC delay of global interconnections



K.Ohsawa, H.Odaira, M.Ohsawa, S.Hirade, T.Iijima, S.G.Pierce, "3-D Assembly Interposer Technology for Next-Generation Integrated Systems," ISSCC Digest of Tech. Papers, pp.272-273, Feb.2001.

Fig.24 3D Integration

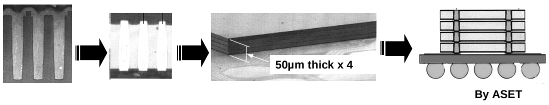


Fig.25 3D Integration

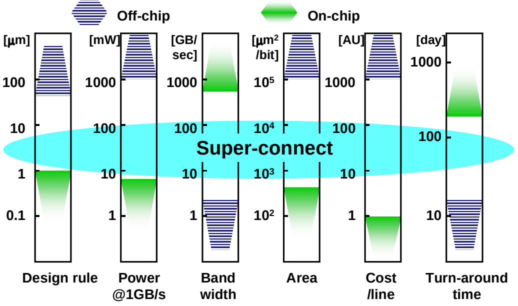
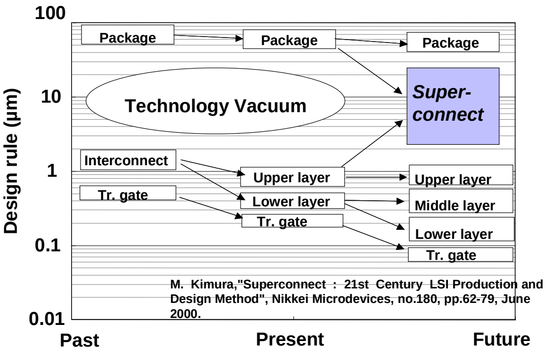


Fig.26 Superconnect



M. Kimura,"Superconnect : 21st Century LSI Production and Design Method", Nikkei Microdevices, no.180, pp.62-79, June 2000.

Fig.27 Superconnect technology