Strain Silicon Optimization for Memory and Logic in Nano-Scale CMOS

Rajani Kuchipudi and Hamid Mahmoodi
School of Engineering, San Francisco State University, San Francisco, CA
{rajanik, mahmoodi}@sfsu.edu

Abstract

Straining of silicon improves mobility of carriers resulting in speed enhancement for transistors in CMOS technology. Traditionally, silicon straining is applied in a similar ad-hoc manner to the whole die including logic and memory. Speed enhancement achieved for both NMOS and PMOS devices is desirable in logic circuits for performance enhancement because both PMOS and NMOS devices lie in critical delay paths. In SRAM cells however PMOS devices are not in the delay path and hence made small to minimize cell area and improve the write stability of the cell. Hence, speed enhancement of PMOS does not result in any reduction in cell access time and in fact it degrades the cell write ability. Hence, optimal method and amount of silicon straining for logic and memory should be different. In this paper, we propose an optimal straining solution for both logic and memory. Based on simulation results in a predictive 45nm process technology, the proposed straining solution enhances circuit performance by 15.6% in SRAM and 39.3% in Logic while satisfying stability requirements. We also propose a co-design optimization methodology that allows optimizing circuit parameters (such as transistor sizing and supply voltage) and process parameters (in this case amount of silicon straining) at the same time for both low power and high performance targets. We found that co-design of supply voltage and silicon straining is very helpful for both low power and high performance targets, whereas co-design of sizing and silicon straining dose not provide any considerable improvements. Our results show that by co-design of supply voltage and silicon straining, power reduction of 38% and 49% is achieved in SRAM and logic, respectively. We also expanded our co-design approach for joint optimization of various circuit and device parameters such as supply voltage, straining, and threshold voltage. The results show that the co-design can reduce leakage by 80% and improve performance by 50%. The developed optimization methodology thus provides a device and circuit co-design framework which is essential as the technology continues to scale to nano-scale regimes.

1. Introduction

The semiconductor industry has recently adopted the concept of silicon straining to enhance the circuit performance. This technology takes advantage of the natural tendency for atoms inside compounds to align with one another. When silicon is deposited on top of a substrate whose atoms are spaced farther apart, the atoms in silicon stretch to align with the atoms beneath, stretching or straining the silicon. In the strained silicon, electrons experience less resistance and flow faster, leading to faster chips without having to shrink the size of transistors [1,2]. Thus, device improvement with strain engineering is nothing but enhancing mobility. So far, silicon straining is applied to both NMOS and PMOS in an ad-hoc manner regardless of the specific circuits. We show that such an approach does not necessarily provide the best of straining technology to all types of circuits. Speed enhancement achieved for both NMOS and PMOS devices is desirable in static CMOS logic circuits for performance enhancement because both PMOS and NMOS devices exist in critical paths. In SRAM however PMOS devices are not in the read delay path and hence designed to be small in order to reduce the cell area and improve the cell write stability. If PMOS strength increases too much by straining, it will degrade SRAM cell write stability without any improvement in the access time. Hence optimal straining of NMOS and PMOS devices is expected to be different for logic circuits and memory cells. Moreover, traditionally, circuit designer have not been considering change of straining since it is a process parameter that seems to be out of control of circuit designers. However, combined circuit and process optimization (device-circuit co-design) can result in a much more optimal design. Thus, a circuit-device co-optimization framework is necessary for future designs.

In this paper, we first propose an optimal straining solution for both logic and memory. We then propose an optimization methodology that allows optimizing circuit parameters (such as supply voltage and transistor sizing) and process parameters (in this case amount of silicon straining) at the same time for both low power and high performance targets. Our results show that co-design of supply voltage and silicon straining is very effective for low power designs, achieving a power reduction of 38% in SRAM and 49% in logic circuits. However, co-design of sizing and silicon straining doe not show any considerable improvements. We also expand our co-design approach for joint optimization of supply voltage, straining, and threshold voltage. The results show that through such a co-design, leakage power can be reduced by 80% and performance can be improved by 50% in SRAM.

The remainder of this paper is organized as follows. In section 2, impact of silicon straining on transistor characteristics is discussed. In section 3, simulation results obtained from straining SRAM cell and ring oscillator are analyzed. In section 5, optimal straining solutions for memory and logic circuits are discussed. In section 6, different circuit-device optimization techniques proposed for high-performance and low-power applications are discussed. Finally the conclusion of the paper appears in section 7.

2. Impact of Silicon Straining on Transistor Characteristics

In this work, the effect of silicon straining is modeled by multiplying the mobility parameter (U0) in spice model card

Proceedings of the 8th International Symposium on Quality Electronic Design (ISQED'07)
0-7695-2795-7/07 $20.00 © 2007 IEEE

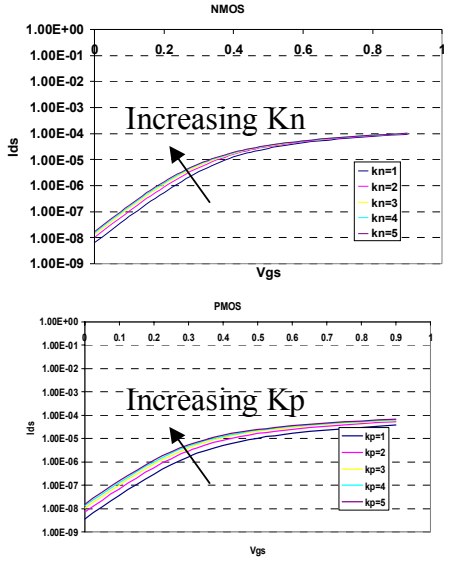


Fig. 1: Impact of silicon straining on Ids/Vgs of (a) NMOS and (b) PMOS

by a new parameter called Kn for NMOS and Kp for PMOS. We use a 45nm predictive technology model [5]. The effect of varying Kn and Kp on Ids/Vgs characteristics of NMOS and PMOS devices is shown in Fig 1. Ioff and Ion both increase by applying straining. By silicon straining, Ioff increases at a faster rate than Ion. Hence, the Ion/Ioff ratio goes down due to silicon straining in both PMOS and NMOS transistors. Fig. 1 also shows that Ioff increase in PMOS is significantly more than Ioff increase in NMOS (as a result of straining). This is a negative aspect from device point of view but it results in a faster device. The increased Ioff can be compensated by straining, Vt, and Vdd co-design approach that will be presented in section 6.3.

3. Silicon Straining for SRAM

Fig 2 provides the schematic of six-transistor SRAM cell. Access to the cell is enabled by the word line (WL) which controls the two access transistors AXR and AXL which, in turn, control whether the cell should be connected to the bit lines: BL and BLC. These bit lines are used to transfer data for both read and write operations.

This section presents the impact of strain silicon parameters of NMOS and PMOS (Kn and Kp) on various design metrics of SRAM cell, which include stability (read and write), leakage power, and access time. Since straining is nothing but increasing mobility of the transistors, it does not really affect the area of the cell. So we are not considering the effect of strain silicon parameters on area.

Access time vs increasing Kn and Kp - Access time is defined as the time delay between read request and the moment the data is available at the output. It represents the delay of the SRAM cell. Fig 3 shows the relationship between access time and increasing Kn and Kp. From Fig. 3, it is clear that straining NMOS transistors (i.e., increase of Kn) in SRAM cell decreases access time. Whereas, straining of PMOS transistors in SRAM cell does not affect access time. The reason can be analyzed as follows. As we increase Kn, the strength of NMOS (NR) and access (AXR) transistors (Fig. 2) increases. So it helps the bit line capacitance to get discharged faster. Thus the access time decreases with increase of Kn. Similarly, PMOS transistor in the driving inverter (Fig. 3) during read access is off and hence it does not affect the discharge of bit line capacitance. Thus, straining of NMOS alone in SRAM reduces delay.

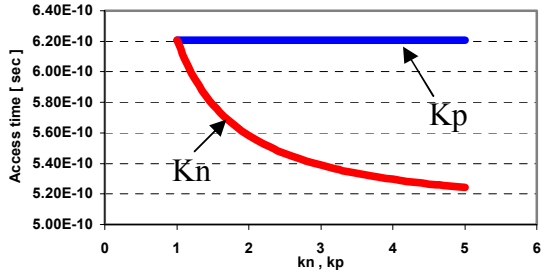


Fig. 3: Access time vs increasing kn and kp

Leakage power vs increasing Kn and Kp - We consider the impact of strain parameters on subthreshold leakage, which dominates SRAM power. Fig 4 shows the relationship between leakage power and increasing Kn and Kp. The increase in leakage power is significantly more when straining is applied to PMOS than NMOS transistor. That is because Ioff increase in PMOS is significantly more than Ioff increase in NMOS as a result of straining (Fig. 1). Thus, leakage power increases more with increase of Kp than Kn.

Read SNM vs increasing Kn and Kp - Read stability is commonly quantified by cell SNM (Static Noise Margin) during read state [6]. It is determined as the side of the maximum square that could fit inside the butterfly curves obtained from the VTC plots of the two cross-coupled inverters [6]. Fig. 5 shows the relationship between Read SNM with increasing Kn and Kp. From Fig 5, it is clear that read SNM decreases (increase) with increase of Kn (Kp). As Kn increases trip point of the inverter falls, and thus read stability decreases as it takes less voltage rise on the node storing zero (Fig. 2) to flip the cell during the read operation. As kp increases, trip point of the inverter rises and thus read SNM increases as it takes larger voltage rise on the node

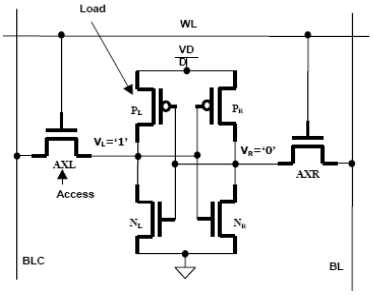


Fig. 2: A Six-Transistor CMOS SRAM cell

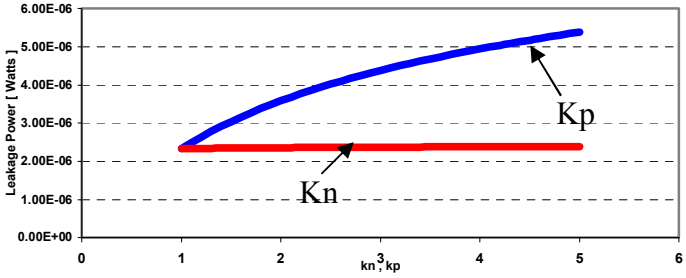


Fig. 4: Leakage power vs increasing Kn and Kp

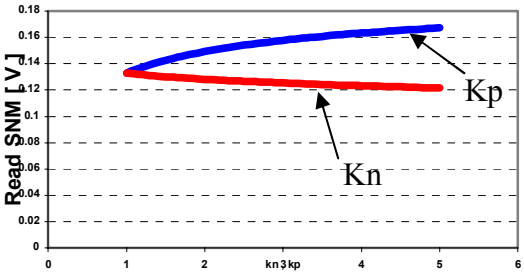


Fig. 5: Read SNM vs increasing Kn and Kp

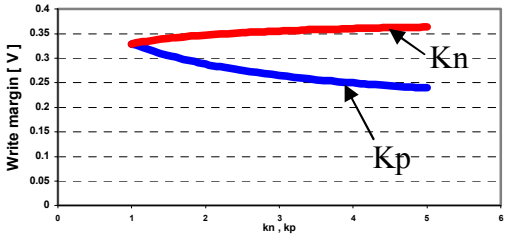


Fig. 6: Write margin vs increasing Kn and Kp

storing zero (Fig 2) to flip the cell during the read operation.

Write margin vs increasing Kn and Kp - The write margin can be measured as the maximum BLC (zero bit-line) voltage, (Fig 2) which is able to flip the cell state, while BL is kept high (vdd) [5]. Fig 6 shows the impact of straining of NMOS and PMOS transistors in SRAM cell on write margin. From Fig. 7, it is clear that by straining NMOS transistors in SRAM cell, the write stability increases. Whereas, straining of PMOS transistors in SRAM cell, decreases write stability. The reason may be analyzed as follows. By increasing Kn, the strength of the access transistor (AXR in Fig 2) increases. Thus, the node storing '1' (VL) can be more easily discharged. Hence, it takes more BLC (Fig 2) voltage before the write to the cell fails. As we increase Kp, PMOS transistor strength increases, and discharging of the node storing '1' (VL) through the access transistors (AXR) thus becomes more difficult. So, less BLC voltage is required to be able to flip or write the cell state. Therefore, write margin decreases with increasing Kp.

From the above observations, it is clear that straining of NMOS transistors in SRAM cell improves the performance.

4. Straining for Logic

To study the impact of silicon straining on general logic circuits, we use a ring oscillator (Fig. 7) as a test bench. Ring oscillator is a good representative of general logic circuits.

4.1 Simulations results obtained from straining Logic

This section discusses the impact of strain silicon parameters (mobility parameters) of NMOS and PMOS on various design metrics of a ring oscillator. Logic design metrics include delay, power, noise immunity, and area. Straining does not affect the area of the circuit. Therefore, we are not considering the effect of strain silicon parameters on area. For the ring oscillator circuit, silicon straining is applied to both NMOS and PMOS transistors at the same time, since both transistors lie in the critical path.

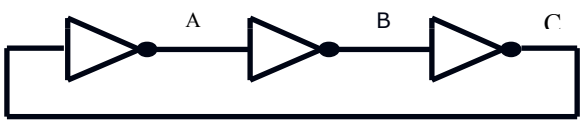


Fig. 7: Ring oscillator as a test bench for logic circuits

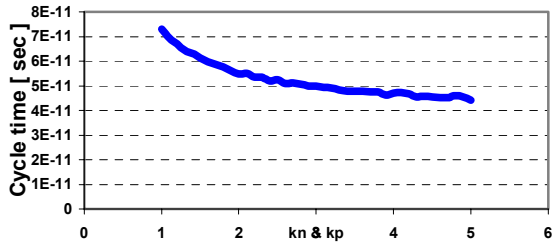


Fig. 8: Cycle time vs Kn and Kp

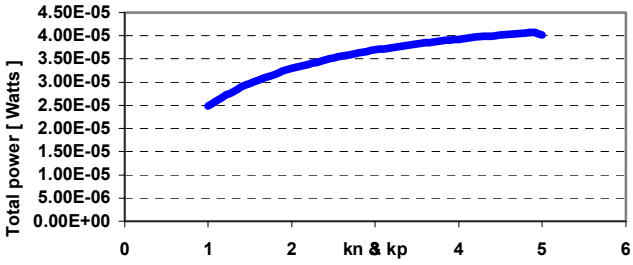


Fig. 9: Total power vs Kn and Kp

Cycle time vs Kn and Kp – Fig. 8 shows the impact of simultaneous straining of NMOS and PMOS transistors of ring oscillator on its oscillation cycle time which represents the delay of the ring oscillator. From Fig. 8, it is clear that straining reduces delay for logic circuits. The reason for reduction in delay can be analyzed as follows. Increase of straining (Kn and Kp) for both NMOS and PMOS transistors results in improving the performance of both the transistors and since both the transistors in a ring oscillator are in critical paths, the total delay is reduced.

Total power vs Kn and Kp – Fig. 9 shows the impact of straining on total power dissipation of the ring oscillator. It clearly shows that straining both NMOS and PMOS transistors increases the total power. The results can be analyzed as follows. The total power of the ring oscillator is composed of leakage power and switching power. Silicon straining increases leakage power (Fig. 1). Frequency increases due to decrease in delay and since switching power ($f.C.Vdd^2$) is proportional to frequency, the switching power also increases. Thus the total power increases due to increase of both the leakage power and switching power.

Energy/cycle vs Kn and Kp – Fig. 10 shows the impact of straining both PMOS and NMOS transistors of ring the oscillator on energy dissipated per cycle. Energy dissipated per cycle is the metric that determines battery lifetime in portable electronic devices. From Fig. 8, 9, and 10 we can

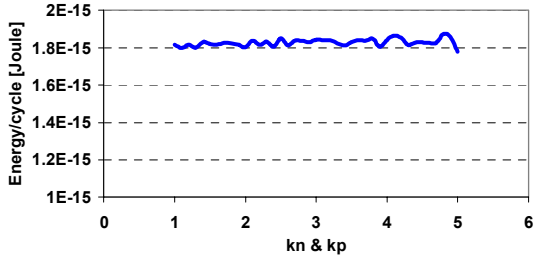


Fig. 10: Energy/cycle vs Kn and Kp

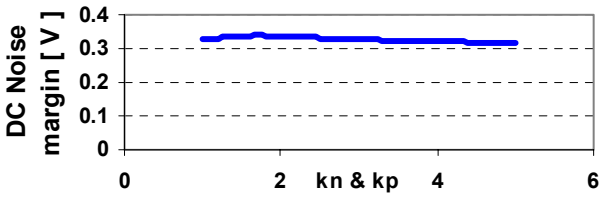


Fig. 11: DC noise margin vs Kn and Kp

conclude that applying silicon straining to logic circuits improves performance while keeping energy/cycle same.

DC noise margin vs Kn and Kp - The DC noise margin of an inverter versus Kn and Kp is shown in Fig. 11. It shows that the noise margin remains almost constant with straining. The results may be analyzed as follows. Since Kn and Kp are increased at the same rate there would be no change in strengths of NMOS and PMOS transistors. Hence, trip point of the inverter remains same. Thus the DC noise margin remains constant. Thus, we can conclude that straining of both NMOS and PMOS transistors improves design quality in logic circuits (improving performance with no penalty in energy/cycle and DC noise margin).

5. Optimal Straining for Logic and Memory on Same Die

From the obtained simulation results so far, we can conclude the following

- Straining of NMOS alone improves performance of SRAM cell.
- Applying straining to both NMOS and PMOS improves performance and overall quality of logic circuits.

Hence, applying straining to all NMOS transistors regardless of logic or memory and applying straining to all PMOS transistors only in case of logic but not memory boosts the performance of the chip. In many applications such as processor chips, logic and memory are integrated on the same die. In that case, we propose to strain all NMOS transistors regardless of logic or memory, whereas straining of PMOS transistors should only be applied in logic part but not memory part of the die. This requires an extra mask for PMOS transistors in memory which increases the processing cost. Given the improvement obtained in design quality, the increased cost may be justified. In the next section, we further maximize the benefit of straining by proposing a circuit-device co-design approach.

6. Circuit-Device Co-Design and Optimization

Traditionally, circuit designers do not consider change of straining since it is a process parameter that seems to be out of control of circuit designers. However, merged circuit and process optimization (device-circuit co-design) will result in a much more optimal design. Currently there is no circuit-device optimization framework in place. We propose an optimization methodology that allows optimizing circuit parameters (such as supply voltage or transistor sizings) and process parameters (in this case amount of silicon straining and threshold voltage) at the same time for both low power and high performance targets. We first look at co-design of voltage and straining and then expand it to include threshold voltage. Our results show that co-design of sizing and straining does not result in any improvement. Hence, we do not consider transistor sizing.

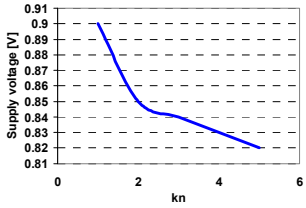
6.1 Supply voltage and straining co-design for memory and logic for low-power targets

This section discusses the implementation of circuit-device optimization for low-power targets in SRAM cell and ring oscillator. From the results obtained so far, we observed that straining of NMOS improves performance of SRAM and straining of both PMOS and NMOS improves performance of logic. Thus for low-power targets where speed requirement is not high the delay improvement obtained by straining is not directly useful. However, the obtained delay reduction by straining can be used to reduce supply voltage for reducing power under a given delay constraint. Hence, silicon straining can be used for scaling of supply voltage, which in turn reduces both leakage and switching power. The proposed optimization approach thus provides a device and circuit co-design framework optimizing both the circuit parameters (supply voltage in this case) and process parameters (silicon straining in this case). The simulation results obtained in a 45nm predictive technology [5] are described as follows.

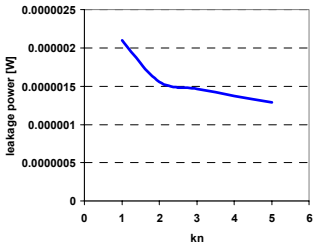
6.1.1 Simulation results for supply and straining co-design in SRAM memory

Supply voltage vs Kn - As mentioned in the previous section, we can keep the delay constant, by using the speed enhancement achieved through straining for reducing the supply voltage. Under a constant delay, the trend of supply voltage scaling and the impact on leakage power and energy/cycle with increasing Kn is shown in Fig. 12.

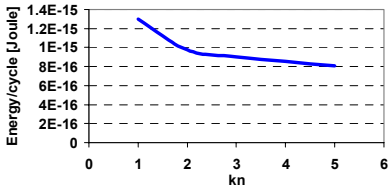
From Fig 12, it is clear that for low power targets, maximum straining should be applied so that minimum supply voltage can be used for the given delay constraint. As observed from



(a)



(b)



(c)

Fig. 12: (a) Supply voltage scaling vs Kn in SRAM under constant delay and (b) impact on leakage power and (c) energy/cycle

Table 1: Supply voltage and straining co-design impact on read and write margins.				
Kn	Cycle time [pS]	Supply voltage [V]	Read SNM [V]	Write margin [V]
1	621	0.9	0.133	0.216
2	621	0.85	0.123	0.217
3	621	0.84	0.119	0.222
4	621	0.83	0.116	0.223
5	621	0.82	0.113	0.221

Fig. 12, reduction in supply voltage from 0.9 to 0.82 reduces power dissipation (leakage power) and energy/cycle by 38%.

We also studied the effect of co-design on read and write margins and the results are as shown in Table 1. From table 1, it is clear that lowering supply voltage by co-design results in read margin (SNM) reduction but no write margin penalty. The reduction in read margin is insignificant compared to the leakage reduction and energy/cycle reduction. With the proposed circuit-device co-optimization of SRAM for low power targets, leakage power and energy/cycle are reduced by 38% with a read margin penalty of 15% and no write margin penalty.

6.1.2 Simulation results for supply voltage and straining co-design in logic

Supply voltage vs Kn – In this optimization, we keep the delay constant by using the speed enhancement achieved through straining for reducing the supply voltage. From Fig. 13, it is clear that the lowest power is achieved for lowest supply voltage obtained by applying maximum straining. In this way, supply voltage scales from 0.9 to 0.66, resulting in total power reduction by 48.38% and energy/cycle reduction by 48.71%. The effect of this co-design on DC noise margin is shown in Table 2. It is observed that this co-design results in DC Noise Margin reduction of 19%. However, the reduction in noise margin is insignificant compared to the reduction is power and energy/cycle. Thus, supply voltage and straining co-design is proved to be efficient for low-power targets in logic. The proposed circuit-device co-optimization of logic for low-power targets reduced leakage power and energy/cycle by 49% with DC noise margin penalty of 19%.

6.2 Supply voltage and straining co-design in SRAM and logic for high-performance applications

From the results obtained in sections 3 and 4 it is clear that Straining improves the speed of SRAM cell by 15.6% with some read penalty. Similarly in logic, straining both NMOS and PMOS improves performance by 39.3%. Supply voltage also needs to be maximized to achieve the maximum speed in both logic and memory. Thus, the optimal supply voltage and straining co-design solution for high performance targets would be to use maximum supply voltage and keep Kn maximum in SRAM. In logic, however, the optimal straining solution would be to keep Kn, Kp, and supply voltage at their maximum value.

Table 2: Supply voltage and straining co-design impact on DC Noise Margin.			
Kn and Kp	Cycle time [pS]	Supply voltage [V]	DC noise margin [V]
1	735	0.9	0.328
2	735	0.76	0.293
3	735	0.70	0.276
4	735	0.67	0.265
5	735	0.66	0.263

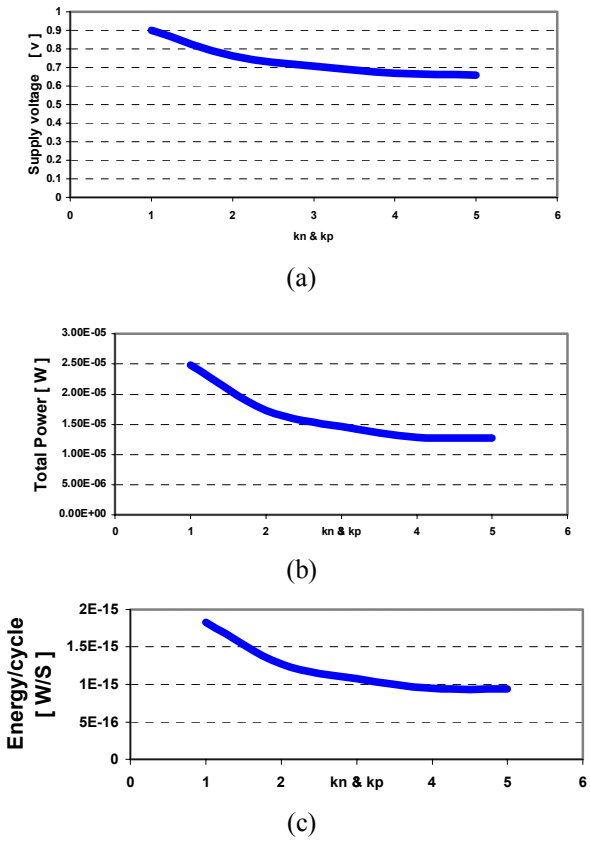


Fig. 13: (a) Supply voltage scaling vs Kn and Kp in logic under constant delay and (b) impact on total power and (c) energy/cycle

6.3 Supply voltage, threshold voltage, and straining co-design in SRAM

In this section, we extended the co-design approach by including another device parameter, threshold voltage. In section 2, we observed that leakage power increases by straining. In SRAM, leakage power is exponentially proportional to Vdd and threshold voltage. Thus leakage power can be reduced by doing joint optimization of supply voltage, threshold voltage, and straining. We started with an SRAM cell that is optimal w.r.t transistor sizing and then performed joint optimization of supply voltage, threshold voltage, and straining. The optimization is performed for low power (minimum leakage) or high performance (minimum access time) targets under some constraints on cell stability metrics (read margin and write margin) and leakage power or access time depending on low power or high performance optimization targets.

6.3.1 Supply voltage, threshold voltage, and straining co-design for low-power targets

As mentioned above, we started with an optimally sized design and then applied our proposed co-design approach. The optimization is done to minimize leakage under constraints of

Table 3: Low power optimization of SRAM cell: comparison of leakage power using conventional and co-design approach.

	Kn	Kp	Supply voltage [V]	Dvtn [V]	Dvtp [V]	Leakage power [µW]
Conventional Approach (Optimally sized)	1	1	0.9	0	0	80.683
Co-design Approach (optimally co-designed)	2	1	0.8	0.06	0.06	13.238

Table 4: High performance optimization of SRAM cell: comparison of access time using conventional and co-design approach.

	Kn	Kp	Supply Voltage [V]	Dvtn [V]	Dvtp [V]	Access Time [pS]
Conventional Approach (optimally sized)	1	1	0.9	0	0	432.93
Co-design Approach (optimally co-designed)	5	2	1.2	-0.06	0.02	214.74

cell stability (read and write margin) and access time. The threshold voltage is modified by changing the threshold voltage parameter (Vth0) is Spice model file. We added the default value of Vth0 to a new parameter, Dvtn for NMOS and Dvtp for PMOS. We developed software programs in java and perl for joint optimization using an exhaustive search in the design space. The results are shown in table 3 where Kn and Kp represents the amount of straining applied to NMOS and PMOS, respectively. Dvtn, Dvtp represents the change in threshold voltage of the PMOS and NMOS transistors. Table 3 shows that using the proposed co-design approach reduces leakage power significantly by 80% compared to conventional approach. It is observed the increase in Kn is used to raise Vt and reduce Vdd while maintaining the access time. The substantial reduction in leakage is due to increase in Vt and reduction in Vdd. Thus co-design is very essential for low power targets.

6.3.2 Supply voltage, threshold voltage, and straining co-design for high-performance targets

As mentioned above, we started with an optimally sized SRAM cell and then applied our proposed co-design approach. Since this optimization is done for high performance targets, the optimization goal is to minimize access time under constraints of stability (read and write margin) and leakage power. The optimal solution was found by developing software programs in java and perl for exhaustive search in the design space. The results are shown in Table 4. Co-design approach has reduced delay significantly by 50% compared to conventional approach. It is observed that Kn is increased, Dvtn is reduces, and supply voltage is increased to minimize access time. Leakage power constraint is satisfied by increasing Dvtp.

The proposed circuit-device co-design approach can be further extended to include more device and circuit parameters in order to get more optimal solutions.

6.3 Transistor sizing and straining co-design in SRAM and logic

We also considered a co-design approach that optimizes transistor sizing and silicon straining at the same time. The obtained results however did not show any considerable improvement for low-power or high-performance targets. Therefore sizing and straining co-design is not effective for both SRAM and logic regardless of high-Performance or low-power targets.

7. Conclusions

In this paper, we proposed a separate silicon straining approach for SRAM and logic. Straining of NMOS alone improves the performance of SRAM and straining of both NMOS and PMOS in logic improves performance keeping energy/cycle same. Moreover, we proposed a device and circuit co-design framework for low power and high-performance targets. The proposed circuit-device co-optimization can optimize circuit parameters (such as transistor sizing and supply voltage) along with device parameters (such as silicon straining and threshold voltage). Co-optimization of supply voltage and silicon straining reduced leakage power and energy/cycle by 38% with no penalty on write margin and with a read margin penalty of 15% in SRAM for low-power targets. Supply voltage and silicon straining co-optimization in logic for low-power targets reduced leakage power and energy/cycle by 49% with DC noise margin penalty of 19%. We also found that sizing and straining co-design is beneficial regardless of logic or memory. We also extended this co-design approach to threshold voltage as a device parameter and found that the joint optimization of supply voltage, straining, and threshold voltage reduced leakage power by 80% and increased performance by 50%.

References

1. IBM Corporation "IBM's Strained Silicon Breakthrough Image page" June 2001. [Online]. Available: http://www.research.ibm.com/resources/press/strainedsilicon
2. Intel Corporation "Strained Silicon Yields Transistor Performance Gains" August 2002. [online]. Available: http://www.intel.com/technology/silicon/si12031.htm
3. Victor chan, Ken Rim, et.al.,"Strain for CMOS performance improvement" IEEE Custom Integrated circuits conference, sepetember 2005.
4. Zheng Guo, Sriram Balasubramanian, et.al., " Finfet Based SRAM Design " University of California, Berkeley.
5. Berkeley Predictive Technology Model (BPTM) "45nm BSIM4 model card for bulk CMOS V1.0".
6. Evert Seevnick, Frans.J.List, et.al.,"Static-Noise Margin Analysis of MOS SRAM cells " IEEE Journal of Solid-State Circuits, vol.22, October 1987.