Chapter4

Power-Speed Optimization in Multiplier-Accumulator Circuit Design

Previously, the high-speed micro-architecture of MAC has been developed. In this chapter, we will base on it to proceed MAC circuit level or transistor level implementation, targeting for power-speed tradeoffs.

In Sec. 4.1, we calibrate the primitive gates which be used in MAC circuit design. A minimum achievable delay of Booth recoder in circuit level design will be presented in Sec.4.2. Furthermore, Sec. 4.3 shows a proposed method for the power-speed circuit level optimization of the column compression stage as well as final addition stage. Finally, some leakage control techniques will be presented in Sec. 4.4.

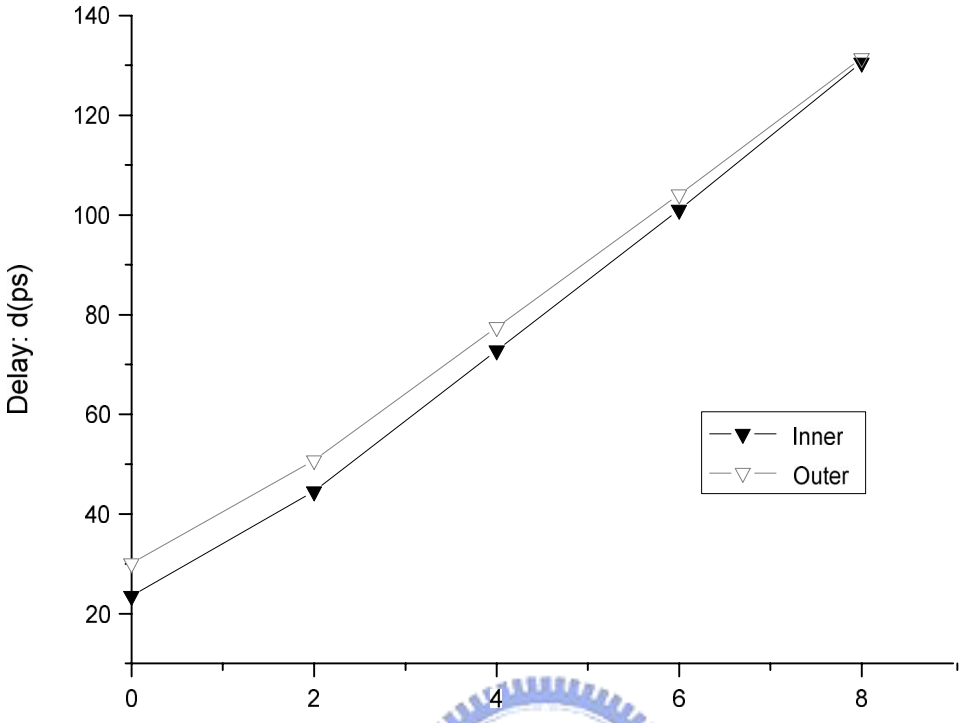
4.1 Primitive Gates Calibration

The calibration technique which we have discussed in Sec. 2.2.3 will be used to calibrate other gates such as NAND2, NAND3, NOR3, XOR2. Figure 4.1 shows the calibration data of a two input NAND gate.

Designing test circuits with logic gate besides inverters requires deciding which input signals be used to propagate signals along the circuit. The unused inputs must be wired to proper logic level so that the gate's output will be controlled by the input signal. For example, unused NAND gate inputs are wired to high, and unused NOR gate inputs are wired to low.

4.2 Low Power Booth Recoder Design

There are two parts of Booth recoder, one is the encoder the other is the selector. The Booth encoder encodes three bits from multiplier and generates five control signal to drive the Booth selector, then produce correct partial product. In our 16X16 MAC, there are 8 partial product have to be generated, each one contains 120 Booth selector. Therefore, to design Booth selector properly will affect area and power significantly.



h		*d*	*g*	*p*
0	inner	23.51		1.95
	outer	30.09		2.51
2		44.56	0.88	
		50.78	0.86	
4		72.80	1.03	
		77.50	0.99	
6		101.03	1.08	
		104.13	1.03	
8		130.58	1.11	
		131.47	1.05	

avg ginner =1.025 g*outer* = *0.98*

Figure 4.1 Simulated delay of NAND2 driving various loads. Results from tsmc 0.13μm 1.2v process.

4.2.1 Booth Selector Design

The Booth selector is responsible for generating one bit of the partial product according to the five control signals from Booth encoder. One bit Booth selector is shown in Figure 4.2. We use the transmission gate as a multiplexer because it provides balanced rise time and fall time and full swing output voltage. The size of PMOS and NMOS transistors in the transmission gate can be equal because both transistors operate in parallel when driving the output. We will determine the size for each transistor as follow:

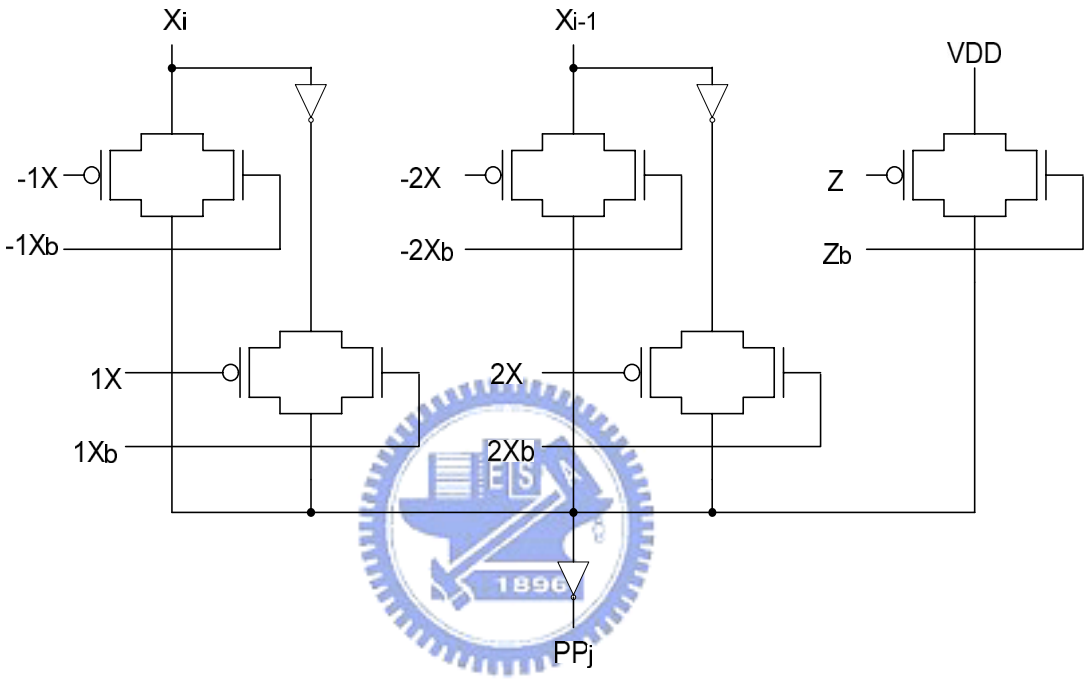


Figure 4.2 One bit Booth selector

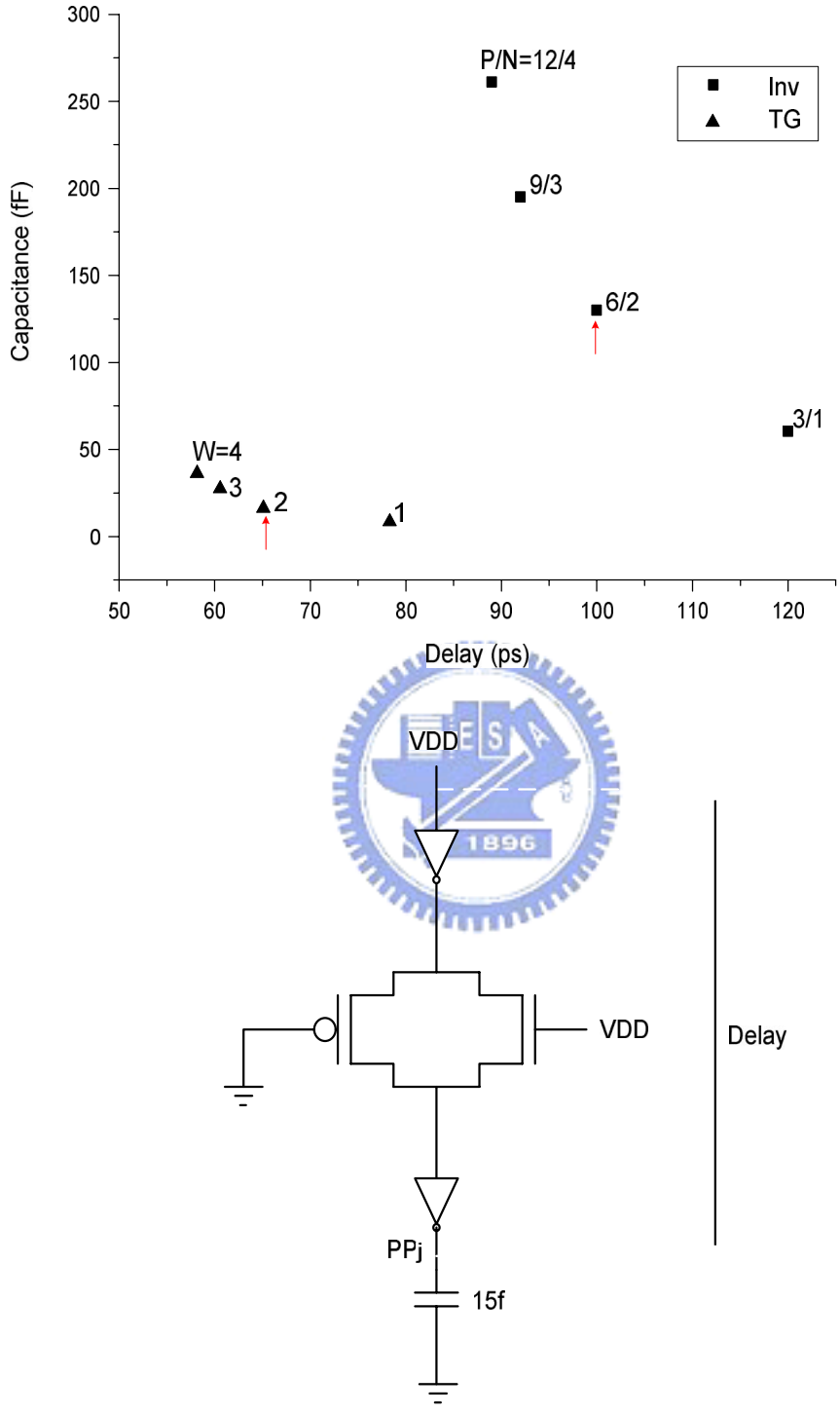


Figure 4.3 Test circuit and results of deciding the size of transmission gate and inverter

4.2.2 Booth Encoder Design

The encoder circuits, as Figure 3.9 shows, must be able to drive large loads in the order of 20fF-40fF as they drive 15 NMOS and PMOS gates. We use path effort to design and estimate the encoder for achieving minimum delay. Figure 4.4 shows the hand calculation result and its simulation result. No more than 5% error should be generated between hand calculation and the simulation result.

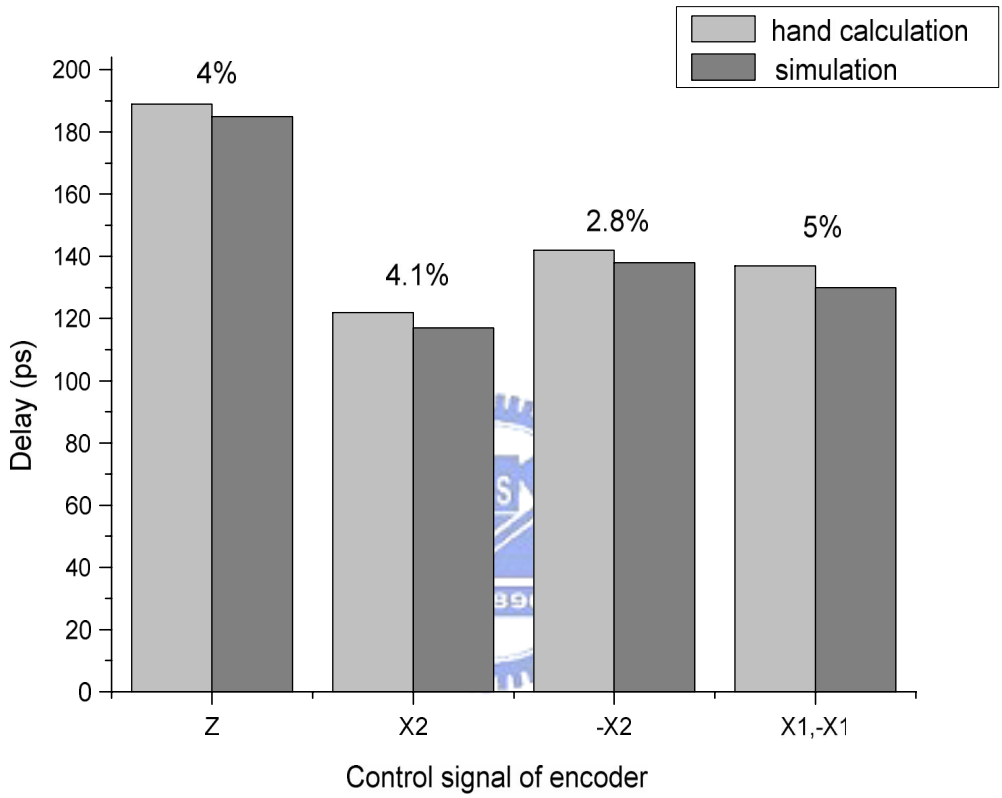
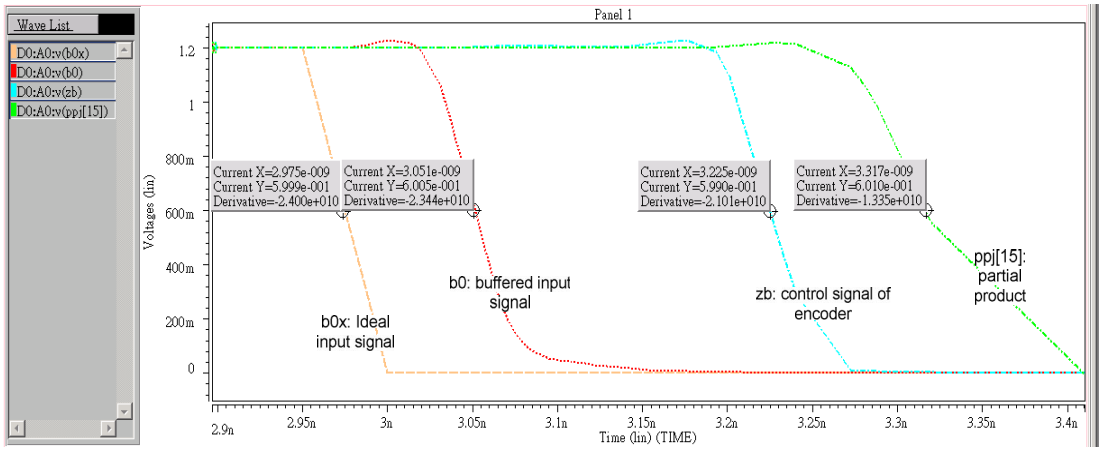


Figure 4.4 Encoder delay estimation and simulation

Figure 4.5 shows the simulation waveform of Booth recoder and its critical path, the control signal Z. The effectiveness of transistor sizing for low power in Booth recoder can be estimated from the profile of power dissipation in the circuit. Figure 4.6 shows the power distribution and profile of full macro Booth recoder. As Figure 4.6 (a) shows, the major power source is the Booth selector which occupies over 40%. We trade speed for lower power consumption by downsizing the transistor width of non-critical path in Booth recoder. Nearly 50% power saving can be reached accompany with 10% delay overhead, Figure 4.6 (b) shows the numerical result. Figure 4.6 (c), (d) depicts the overall power profile of Booth recoder under minimum delay and under 1.1X minimum delay.



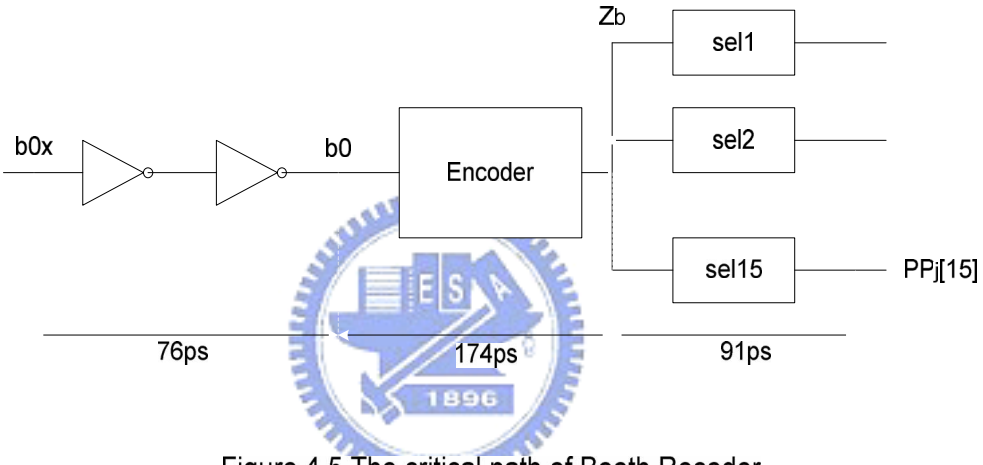


Figure 4.5 The critical path of Booth Recoder

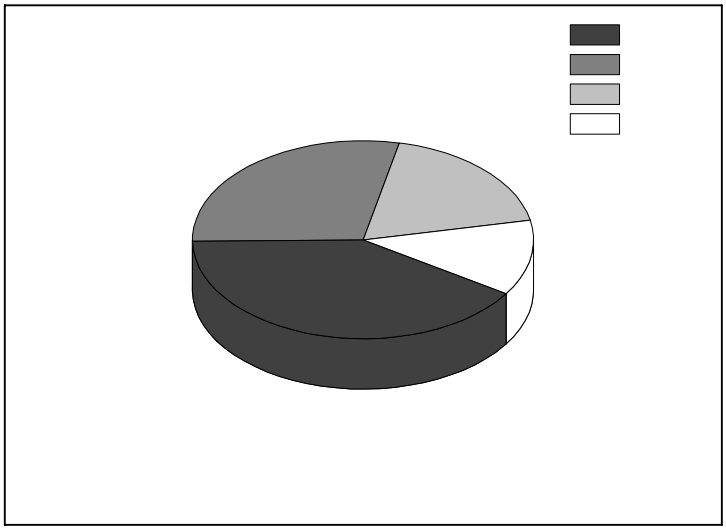


Figure 4.6 (a) Power distribuiton of Booth Recoder under MIN delay

	Power consumption (uW)	
	MIN Delay	1.1 X MIN Delay
Buffer	2428	1174
Selector	3440	1695
Encoder	1560	672
Sparse	1045	949
total	8473	4490

Figure 4.6 (b) Trade speed for power reduction

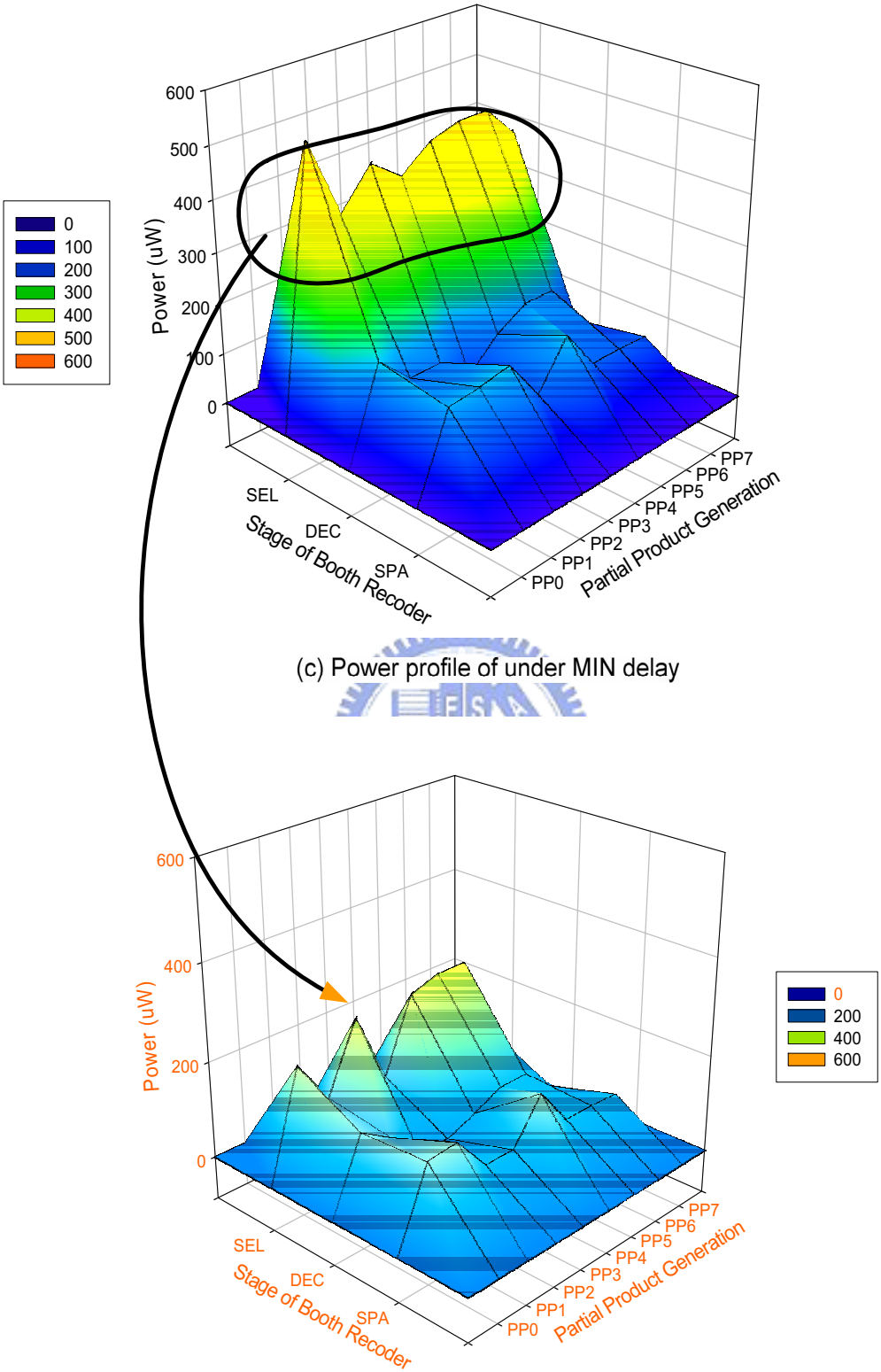


Figure 4.6 (c) (d) Power distribuiton and profile of Booth Recoder

4.3 Power-Speed Optimization of MAC

4.3.1 Evaluation of XOR gate

The computation kernel of column compression stage is the XOR gate, because all of the 5-2 compressor family, the 5-2, 4-2 compressor, and 3-2 adder, are XOR-based circuit structure. Therefore, an efficient XOR gate design is a must.

There are numerous efficient XOR gates have been proposed in the past. Figure 4.7 summarized six single-rail XOR designs. We evaluated these single-rail XOR gates as a template for designing efficient compressors later. Design Figure 4.7 (a) is composed of two transmission gates and two inverters which is the TG-CMOS logic style. Design Figure 4.7 (b) uses six transistors and is based on the transmission function theory. Figure 4.7 (c) is an inverter-based design, which has non-full swing problem. Design Figure 4.7 (d) is presented in [4.1]. It uses only four transistors. Design Figure 4.7 (e) is presented in [4.2], which uses seven transistors is suitable for low voltage design. Design Figure 4.7 (f) is the one proposed by [4.3], two Vt loss problem also makes it a unfavorable design.

We have evaluated four comparable XOR gates, Figure 4.7 (a), (b), (d), and (e) will be performed by using the calibration method of logical effort, the power will also be measured, as shown in Figure 4.8. The simulation result from Figure 4.8 shows that, the design Figure 4.7 (d) presents lowest logical effort, electrical effort, and power consumption. However, it suffers by the weak ground path results in “pool 0”. We can solve this problem by introducing swing restore circuit to recover its voltage level. The design Figure 4.7 (e) provides good output voltage swing and suitable for low voltage design, but it consumes largest power when drives large loading. The design Figure 4.7 (a) and (b) presents slower speed but have good output voltage swing and low power consumption, which are suitable for driving large loading.

In next section, we will use design Figure 4.7 (c) and (d) to implement dual rail XOR gates, and add swing restore circuit to form a computation kernel of compressors. Design Figure 4.7 (b) will be served as the final stage of compressors to drive large loading. And then, compare delay and power using different computation kernel in compressors design.

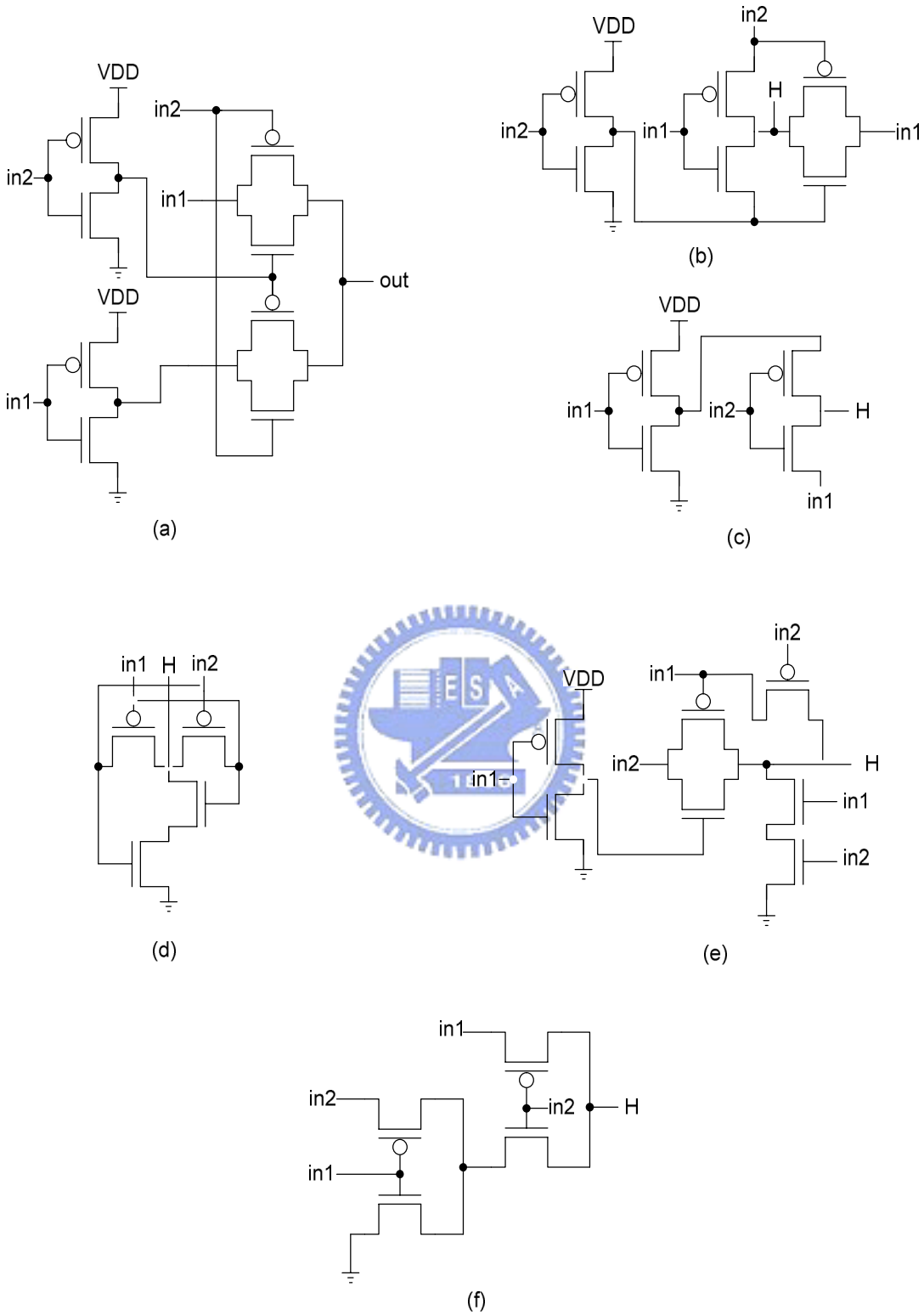
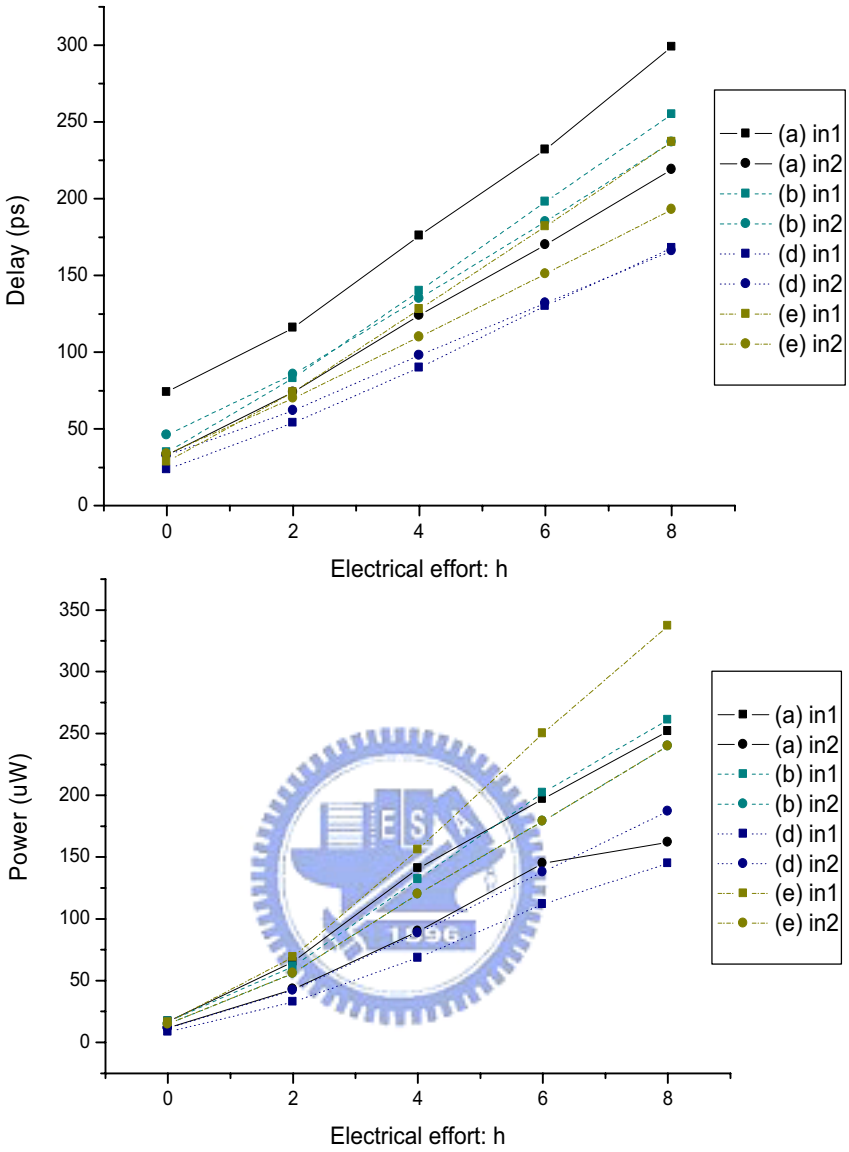


Figure 4.7 Six different designs of XOR gates



Gate type	Input	*Logical effort*	*Parasitic Delay*
Fig. (a)	1	2.1	6.17
	2	1.86	2.75
Fig. (b)	1	2.19	2.9
	2	1.86	3.8
Fig. (d)	1	1.41	1.96
	2	1.33	2.73
Fig. (e)	1	2.05	2.39
	2	1.6	2.8

Figure 4.8 Evaluations of XOR gates using logical effort

4.3.2 The 5-2 Compressor's Family

Micro-Architectural Comparison

The 5-2 compressor's family is composed of the 3-2 adder, the 4-2 compressor, and 5-2 compressor, each component must contain two types of gates, a XOR gate and a MUX. Table 4.1 compared three possible designs of 5-2 compressor. FA-based design is hardly comparable to that of others because its critical path was in excess of two XOR delays. The inner fanout node and the total amount of component in the compressor Comp1 are much higher than in the compressor Comp2, resulting in higher area and higher power consumption. Thus, our favorable choice is the Comp2 compressor. Figure 4.9 (a), (b), and (c) show the Comp2 circuit structure for the 5-2, 4-2 compressor and the 3-2 adder respectively.

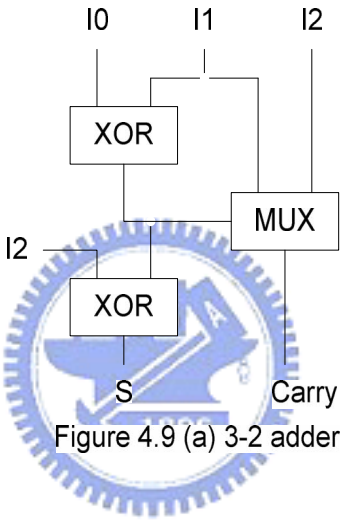


Figure 4.9 (a) 3-2 adder

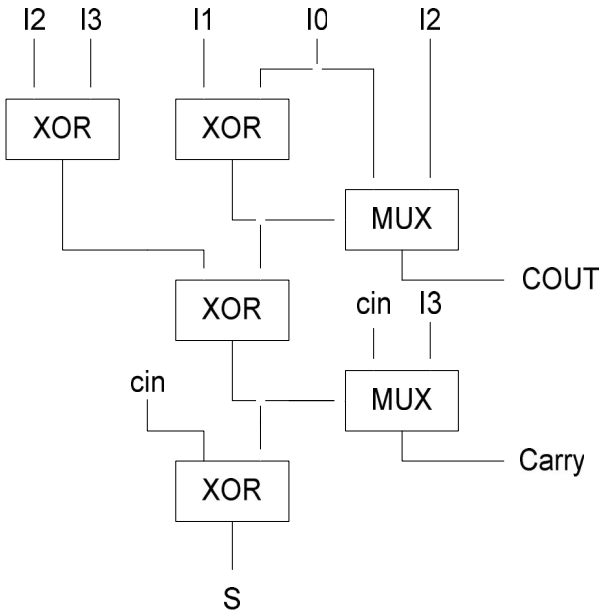


Figure 4.9 (b) 4-2 compressor

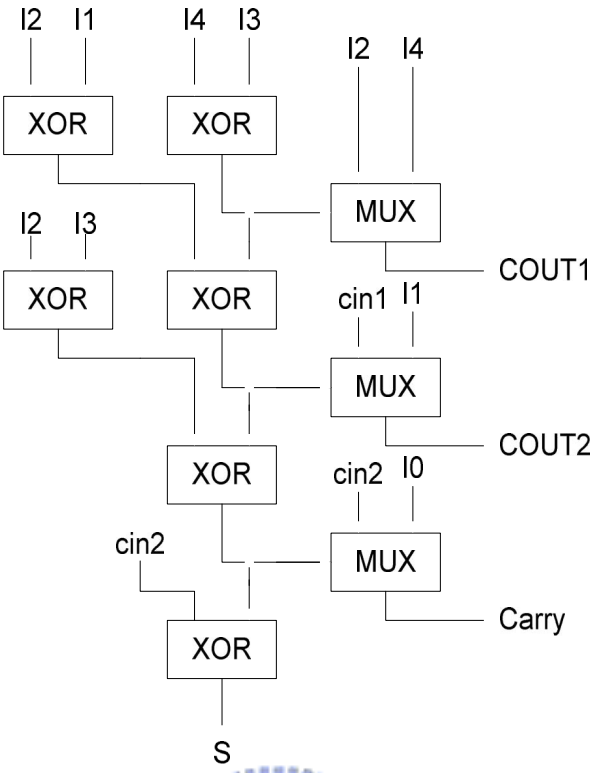


Figure 4.9 (c) 5-2 compressor

Efficient Circuit Level Implementation

There are many efficient 4-2, and 5-2 compressor designs have been presented [4.4] [4.5] [4.6] [4.7] [4.8] [4.9] [4.11]. In [4.8] [4.9] [4.10] [4.11], these novel circuit level designs of 4-2, and 5-2 compressors are well discussed. The dual rail XOR gates in pass-transistor logic style have been used extensively. Figure 4.10-1 divides a 5-2 compressor into three modules. Module1 which is the computation kernel determines the critical path delay. Design Figure 4.10-1 (a) is the design Figure 4.7 (e). In the Design Figure 4.10-1 (b), compact XOR gate is proposed by [4.12]. Design Figure 4.10-1 (c) [4.11] is a combination of design Figure 4.7 (c) and a swing restore back-to-back latch [2.12]. Design Figure 4.10-1 (d) is composed of design Figure 4.7 (d) and a swing restore circuit. Design Figure 4.10-1 (e) is the DCVSPG logic style XOR gate [4.13]. Module2 and Module3 must have large driving capability and good output voltage swing, are shown in Figure 4.10-2. In the same way, the 4-2 compressor has the same circuit structure, and hence, can reuse the same module of the 5-2 compressor.

Figure 4.11 shows the simulation results of a 5-2 compressor using different computation kernels. Obviously, The design Figure 4.10 (a) XOR gate has lowest power-delay product. However, if the speed is the big concern, the DCVSPG logic style is a best choice. Table 4.2 shows these numerical results of the critical path delay of the 5-2 compressor using different XOR gates.

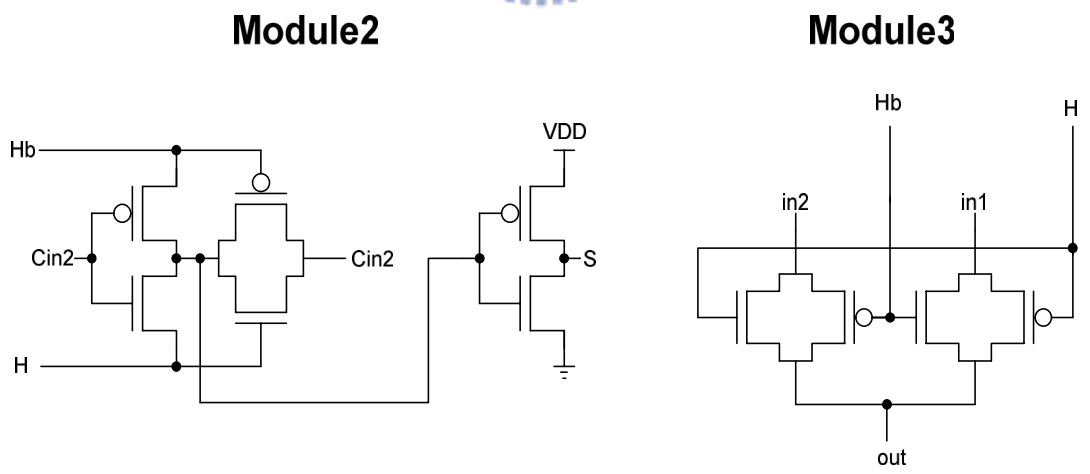


Figure 4.10-2 Decomposition of the 5-2 Compressor: module2 and module3

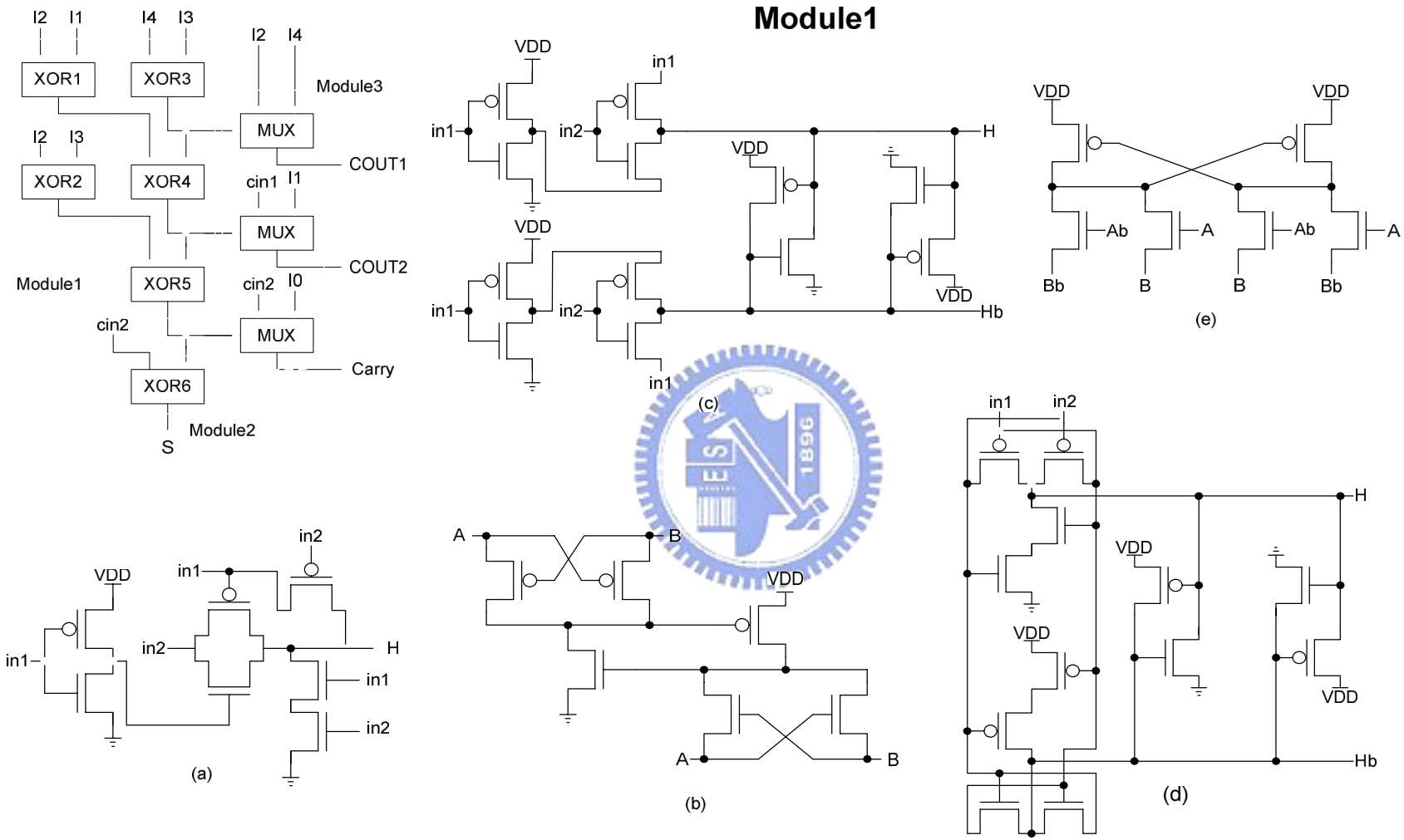


Figure 4.10-1 Decomposition of the 5-2 Compressor: module1

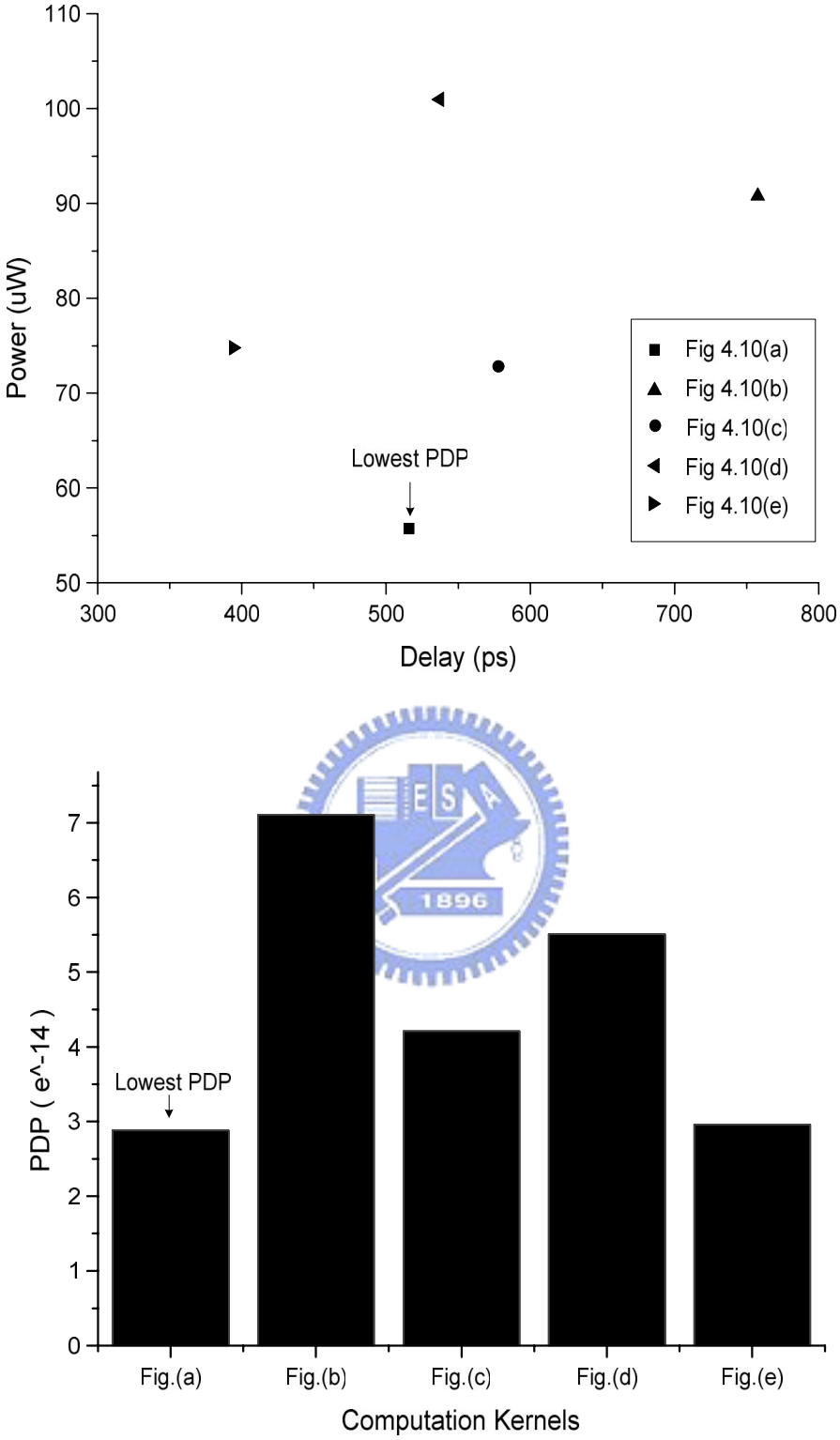


Figure 4.11 High speed and low power computation kernel of compressors

Table 4.2 Numerical results of different implementations of 5-2 compressors

5-2 comp	Fig 4.10 (a)	Fig 4.10 (b)	Fig 4.10 (c)	Fig 4.10 (d)	Fig 4.10 (e)
Worse case delay (ps)	517	767	578	541	398
Power (μW)	55.8	92.7	73	102	74.3
PDP	2.8822E-14	7.1120E-14	4.2170E-14	5.5074E-14	2.9564E-14

4.3.3 Proposed Practical Power-Speed Optimization Procedure

After designing efficient micro-architecture and circuit design of compressors, we proposed a practical optimization procedure to optimize power and speed. There is an assumption should be make, the initial power-delay point has been designed with minimum delay. Figure 4.12 shows proposed optimization procedure. Initially, it starts at minimum delay design point. In the second step, three tuning variables, the transistor sizing, the threshold voltage, and the supply voltage, which were scaled separately in order to find optimization curves on the power and delay design space. The third step overlaps these simulated curves on the same power and delay design space. Finally, the design point is moved to approach optimum design point as follow:

1. Downsizing the non-critical path to decrease power without sacrificing speed.
2. Select proper supply voltage to further lower the power, while not results in adding delay overhead too much.
3. Select suitable threshold voltage, move the design point to achieve optimum.

4.3.4 Power-Speed Optimization of Column Compression Stage

In the column compression step of a MAC operation, the partial product reduction topology is an inverted triangle, as shown in Figure 3.19. This topology leads to different combinations of 4-2, 5-2 compressors and 3-2 adders, and to form unbalanced path delays. In order to achieve power-delay tradeoffs effectively, we have to identify the critical path in this stage, at first, and downsizing the non-critical path to save active power without sacrificing speed, Figure 4.13 shows the critical path in the column compression step. As figure shows, path 1, path 2 and path 3 are the critical paths, others are non-critical. We downsize these non-critical paths without sacrificing speed. But, however, the delay of critical path should increase somewhat, because the non-critical path affects the delay of carry of the critical path. The power saving in this region has great potential. As we downsize the critical path, the delay penalty increases rapidly, the power saving potential becomes moderate. Therefore, the largest circuit sensitivity of transistor sizing appears when we downsize the non-critical path. Figure 4.14 shows the simulation results by using transistor sizing as the tuning variable.

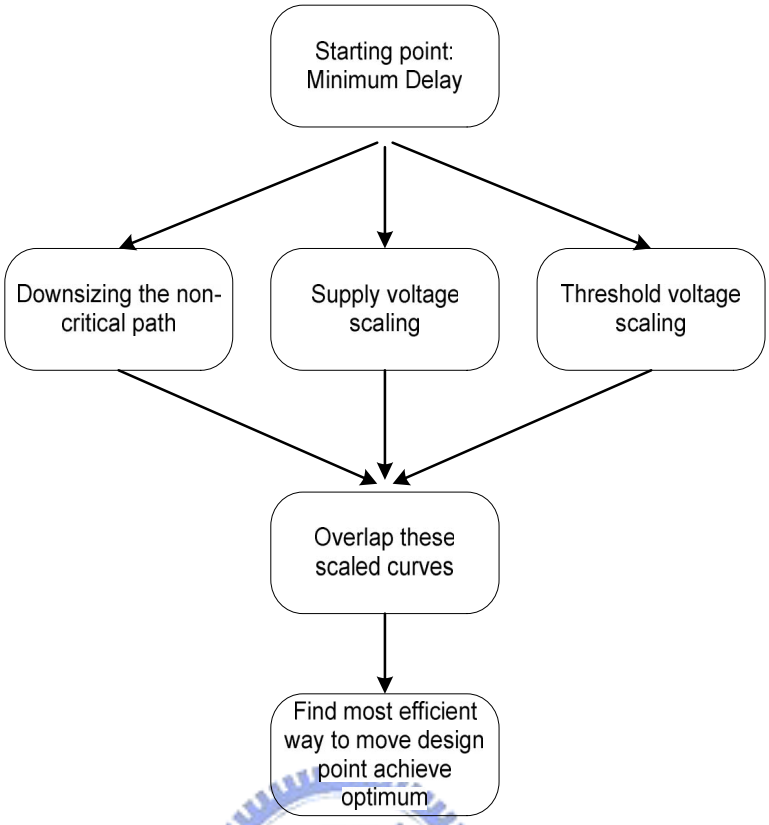


Figure 4.12 Proposed power-speed optimization procedure

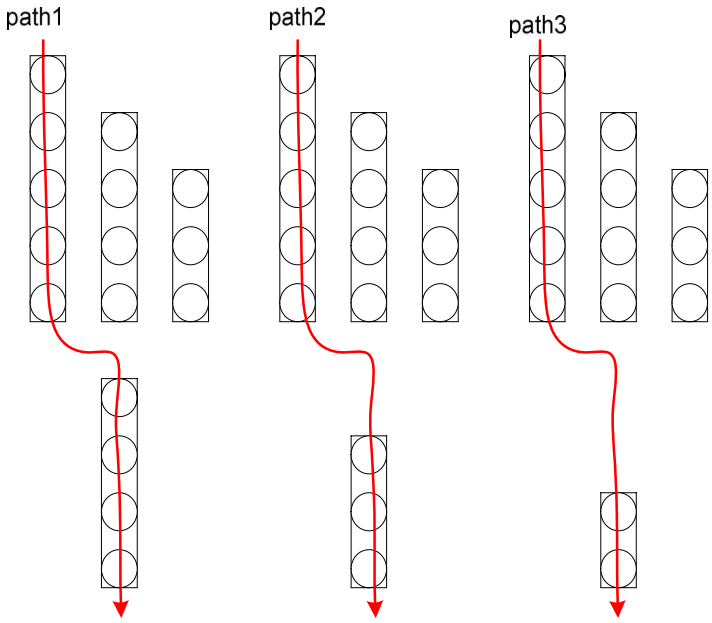


Figure 4.13 Critical path in the column compression stage

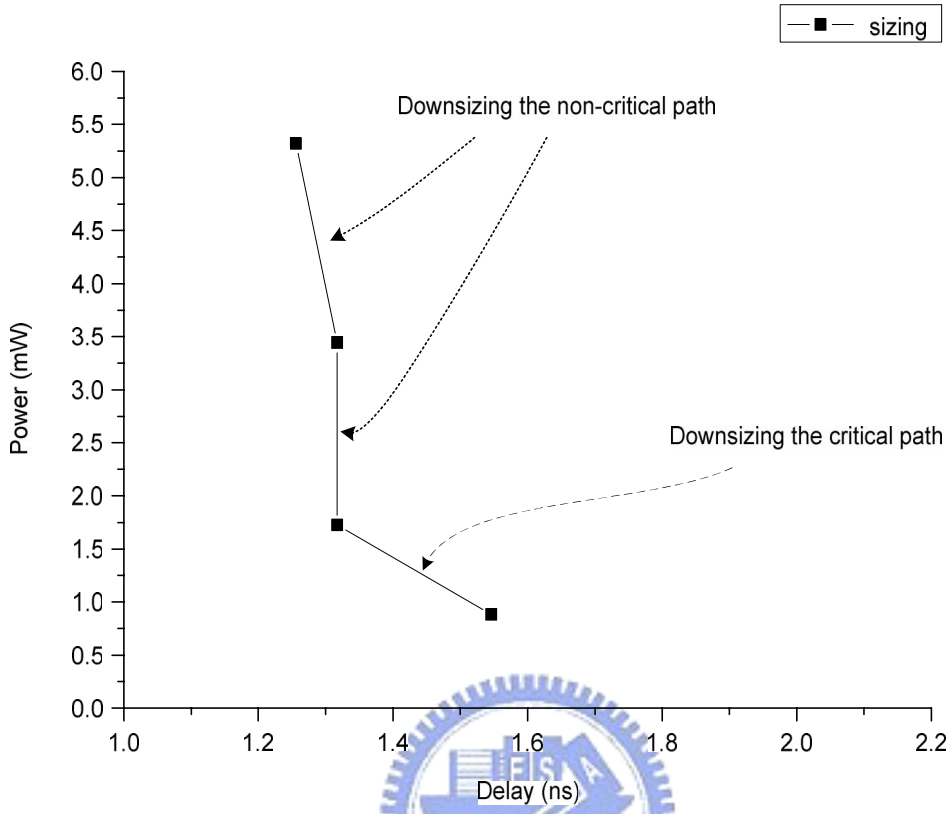
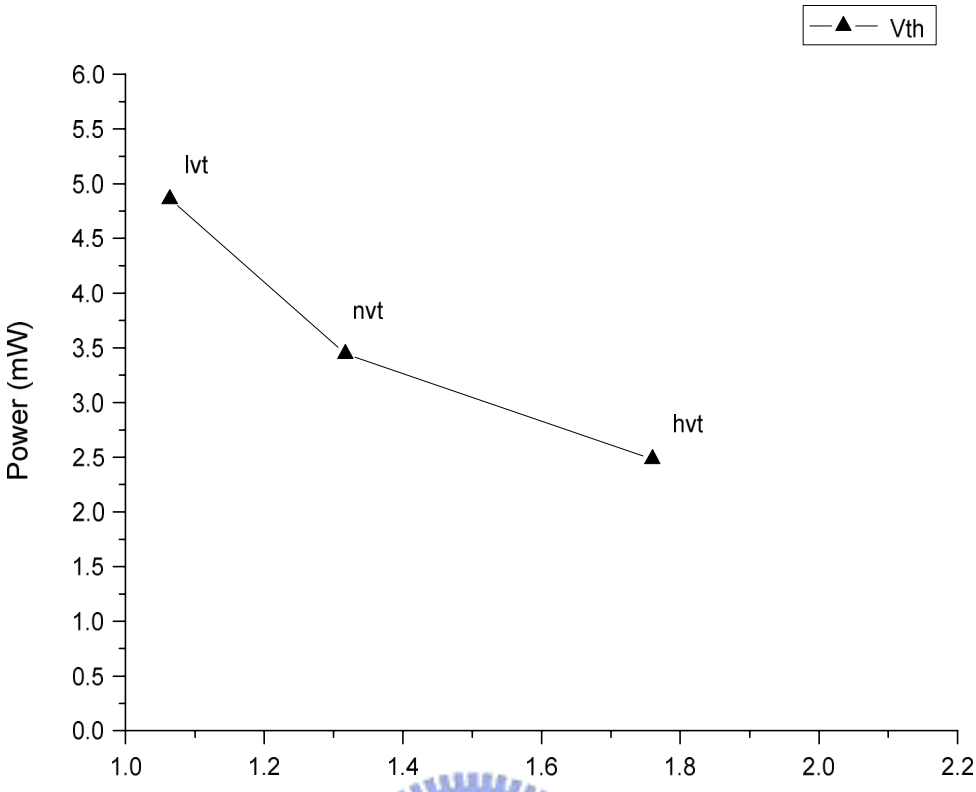


Figure 4.14 Power-speed optimization via transistor sizing

We introduce the second design control variable, the threshold voltage, to achieve the goal of power-speed optimization. Three options of threshold voltages, the low threshold voltage, the normal threshold voltage, and the high threshold voltage, which is provided by tsmc 0.13μm technology, are used to carry out the simulation of column compression macro. Figure 4.15 lists the voltage value of three types of threshold voltages and shows the simulation results. The region between lvt and nvt has higher circuit sensitivity than the region between nvt and hvt. Thus, lvt and nvt are the superior design point in the demand of power-speed optimization.



Threshold voltage of tsmc 0.13um technology (mV)

	Low Vth	Normal Vth	High Vth
N-ch	284.6	372.6	432.1
P-ch	-261.6	-370.9	-439.1

Figure 4.15 Power-speed optimization via threshold voltage scaling

The third tuning variable is the supply voltage. We use the normal threshold voltage as a basis for simulating the column compression macro. Our supply voltage is ranging from 0.8V ~ 1.5V in tsmc 0.13μm technology. Under the voltage scaling simulation, we can find the optimum value of power and speed. Figure 4.16 shows power, delay, and the power delay product for each supply voltage. The lowest PDP point is located at 0.9v, but its delay penalty is too heavy. If our concern is to optimize the speed, the 1.1v supply voltage is a suitable point which has moderate speed and power consumption.

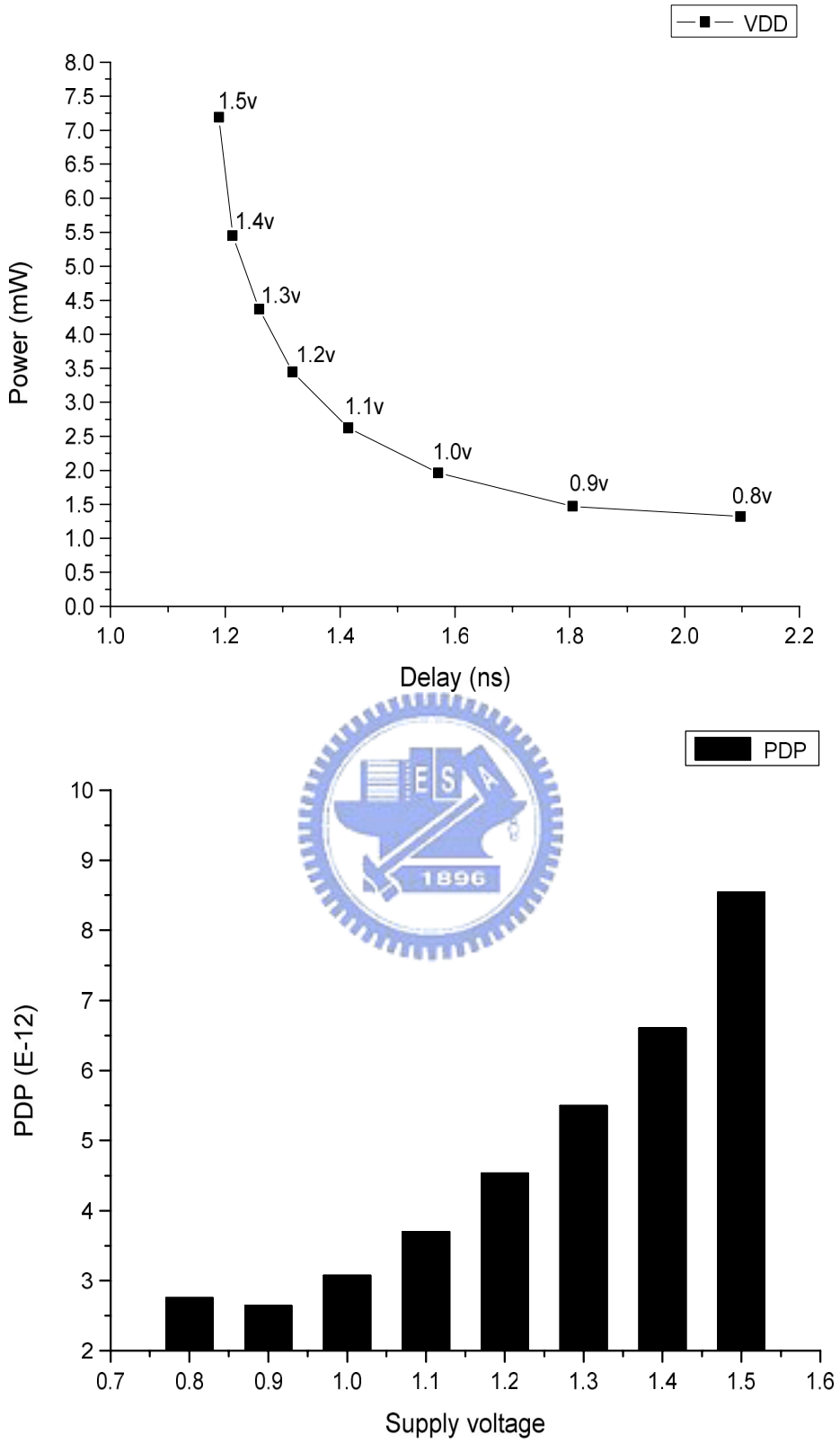


Figure 4.16 Power-speed optimization via supply voltage scaling

Finally, we have combined three tuning variables, the transistor sizing, the threshold voltage, and the supply voltage, to optimize the column compression macro jointly. The first step is to downsize the non-critical path. The objective of this step is to reduce the power as possible while without increasing the delay. The second step is to lower the supply voltage, thus the power consumption increases further, while the delay penalty not likely to increase too much. The third step lower the threshold voltage, therefore the delay will be decreased substantially, while the power consumption increases gently. Figure 4.17 (a) shows the simulation results of individual optimization of the tuning variables, while Figure 4.17 (b) shows the steps of power-delay optimization.

Simulation results show that the optimum design point outperforms the normal design point. After optimizing three tuning variables, in terms of the power consumption it achieves 45% of improvement, and in case of delay it achieves 18.2% speed improvement.

4.3.5 Power-Speed Optimization of Final Adder

For the last stage of the MAC circuit design, we implemented a 32-bit K-S adder at circuit level, its circuit structure is shown in Figure 3.22. In order to achieve low power goal and increases its speed at the time, the strategy that we used in final adder is the same as previous section. By using three tuning variables optimization, the power consumption achieves about 50% reduction, and speed achieves 17% of improvement. Simulation results are shown in Figure 4.18 (a) and Figure 4.18 (b).

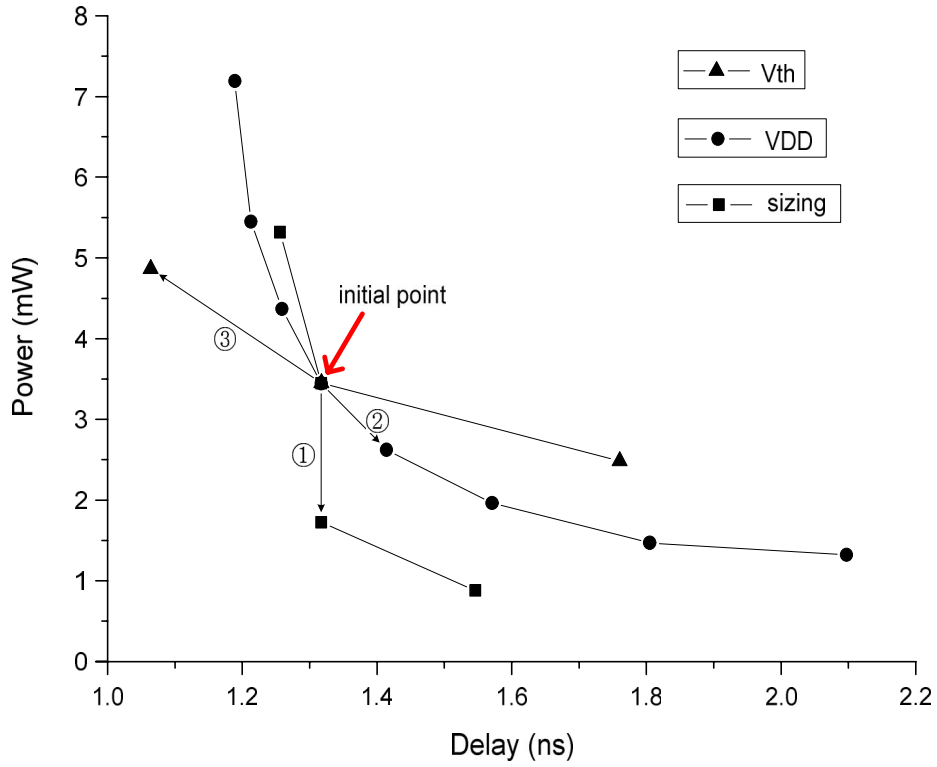


Figure 4.17 (a) Individual optimization of column compression stage: sizing, VDD, and Vth

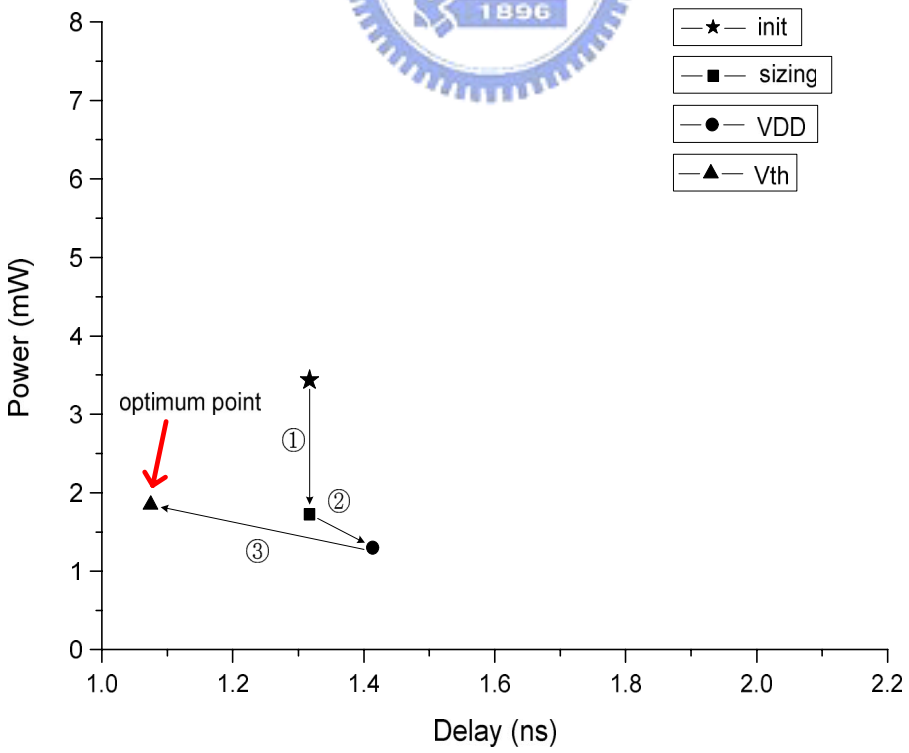


Figure 4.17 (b) Joint optimization of column compression stage: sizing, VDD, and Vth

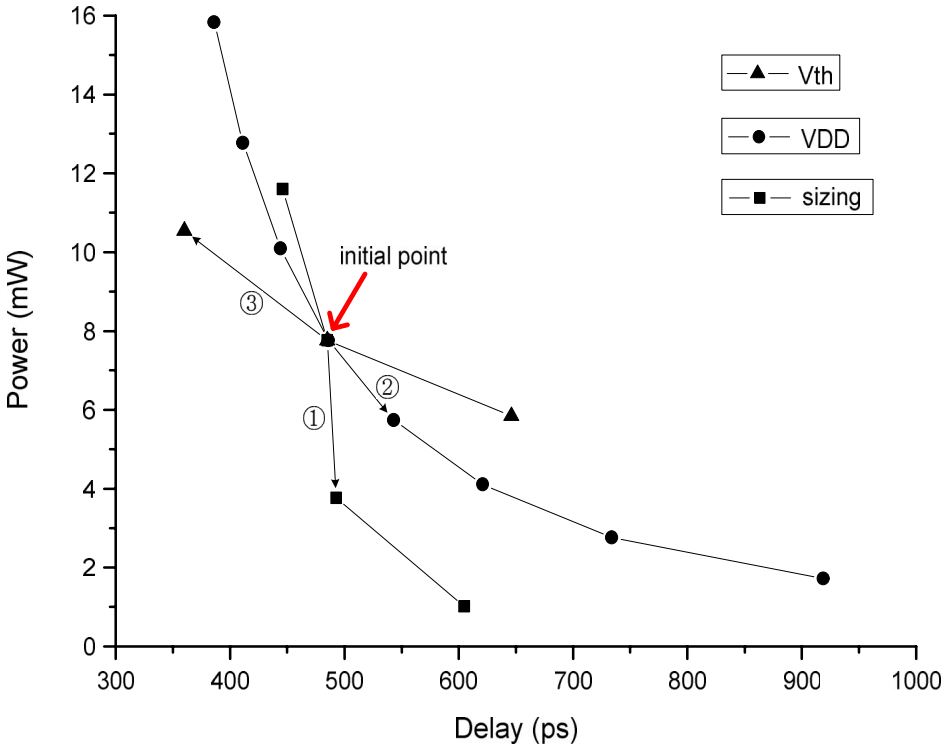


Figure 4.18 (a) Individual optimization of K-S adder: sizing, VDD, and Vth

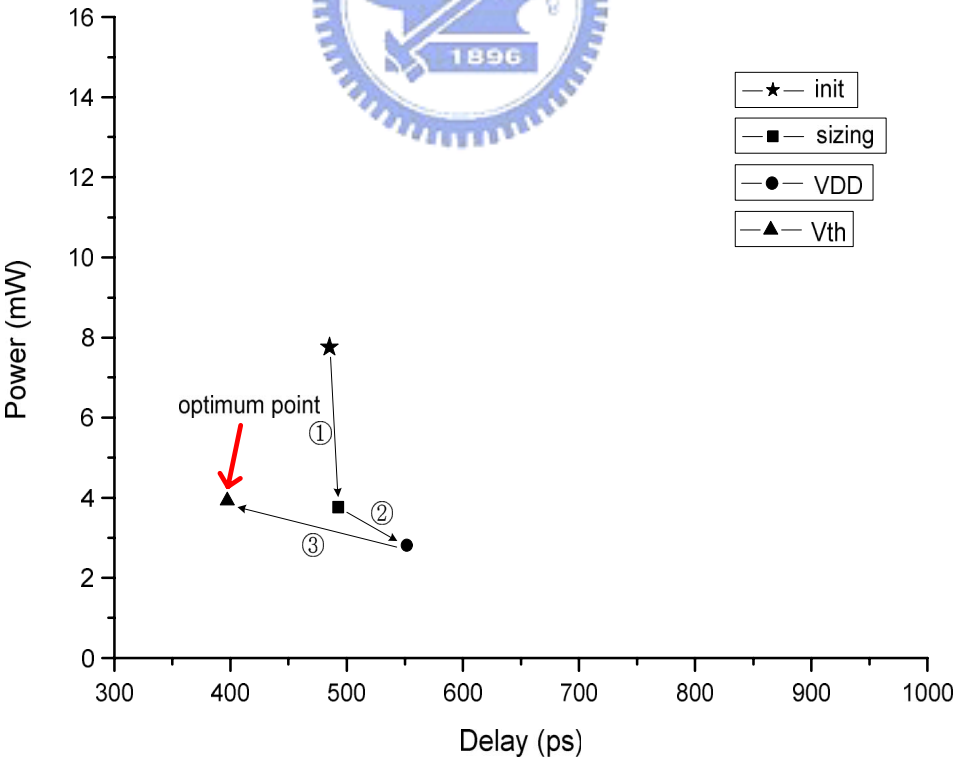


Figure 4.18 (b) Joint optimization of K-S adder: sizing, VDD, and Vth

4.4 Managing Standby and Active Mode Leakage Power

In this section, we describe circuit design techniques for managing leakage power, both during standby mode and active mode. Firstly, the leakage components in modern MOSFETs are described. Secondly, we concern about design techniques to limit standby leakage. Lastly, the approaches of limiting active leakage are examined.

CMOS circuit power has different dominant components, as discussed in chapter 2.3.1, depending on the operation mode, these are:

- ✓ The active power component dominants during active mode.
- ✓ There are two major leakage components, the active component and that during standby mode.
- ✓ Constant term: "Always on" circuit contributes in both active and standby mode operation.

4.4.1 Leakage Components

Figure 4.19 illustrates leakage components of MOS devices. Leakage currents can be broken down into various components such as PN junction reverse bias current, gate-induced drain leakage current (GIDL), oxide tunneling current, hot carrier injection, and sub-threshold leakage current.

- ✓ The largest contributor is the sub-threshold leakage (I_{off}). Scaled transistors reduce the threshold voltage to maintain performance by keeping gate overdrive voltage as supply voltage is lowered. I_{off} is also increased by drain induced barrier lowering (DIBL).
- ✓ For gate oxide thicknesses below 3nm [4.14] quantum mechanical tunneling current becomes significant. I_{gate} can be between the gate and source, drain, and channel. It can be suppressed by operating in weak inversion to diminish tunneling to the channel region.
- ✓ Another important leakage effect is gate induced drain leakage (GIDL) at the NMOS gate-drain edge. For a gate having a 0V bias with the drain at V_{dd}, significant band bending occurs in the drain region, allowing electron-hole pair creation mechanisms. The advent of this mechanism can be alleviated by limiting the drain to gate voltage.
- ✓ Diode area components from both the source-drain diodes and the well diodes are generally negligible compare to I_{off} and GIDL components despite having an approximately 1000x/100C temperature coefficient [4.15].

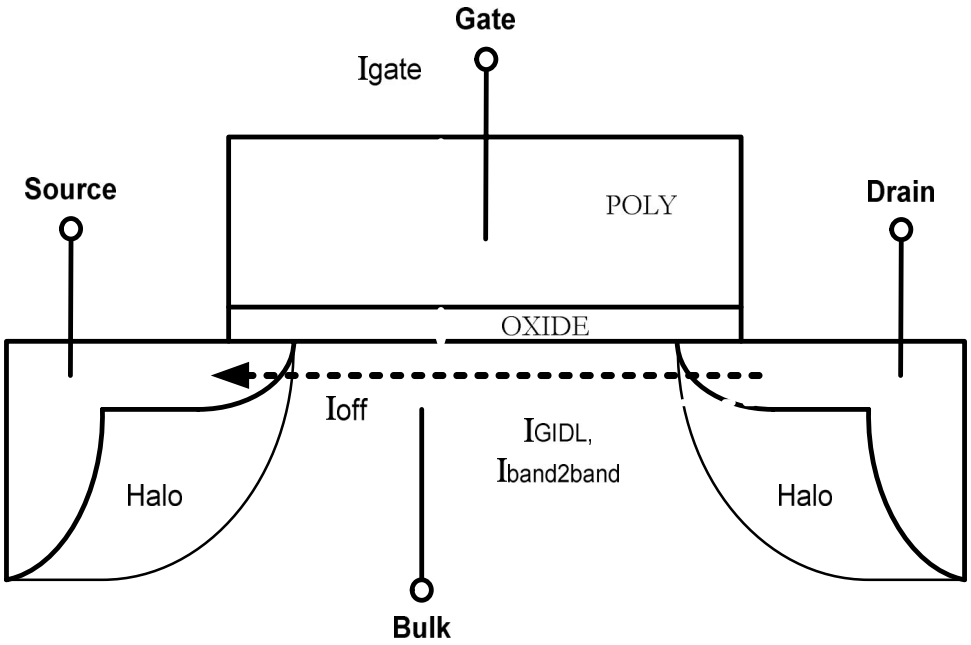


Figure 4.19 MOSFET leakage components

Transistor Stacks (Self-Reverse Bias)

Sub-threshold leakage current flowing through a stack of series-connected transistors reduces when more than one transistor in the stack is turned off. Figure 4.20 (a) illustrates this stacking effect, when both M1 and M2 are turned off, the voltage at the intermediate node VM is positive due to small train current. Positive potential at intermediate node has great effect for reducing sub-threshold leakage current [4.16]. Therefore, the leakage of a two-transistor stack is an order of magnitude less than the leakage in a single transistor. And hence, the sub-threshold leakage through a logic gate depends on the applied input vector.

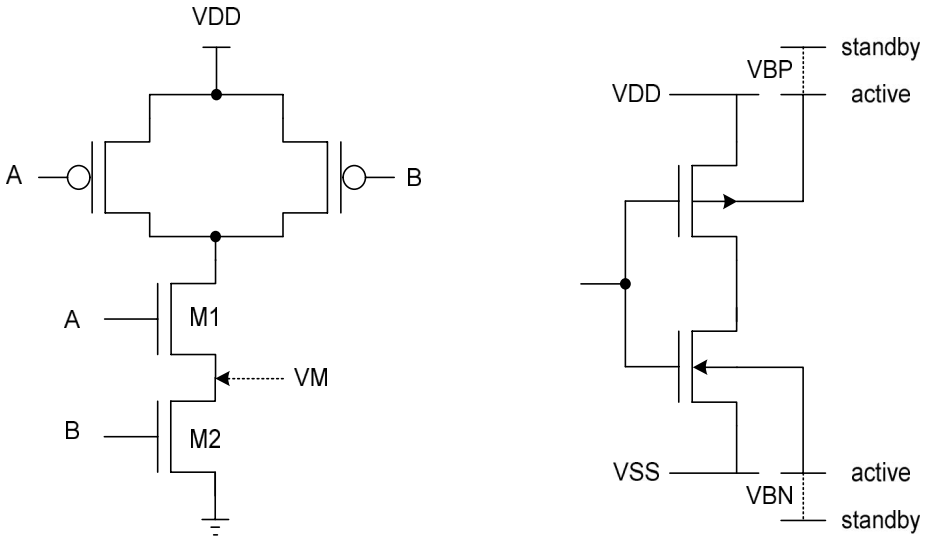


Figure 4.20 Staking effect and VTCMOS

Multithreshold Voltage CMOS (MTCMOS)

Multithreshold voltage CMOS (MTCMOS) reduces the leakage by inserting high-threshold devices in series to low-Vth circuitry [4.17]. This technique has been discussed in section 2.3.5.

Variable Threshold Voltage CMOS (VTCMOS)

Variable threshold CMOS (VTCMOS) is a reverse body-biasing design technique. Figure 4.20 (b) shows the VTCMOS scheme. To achieve different threshold voltages, a self-substrate bias circuit is used to control the body bias. In the active mode, a zero body bias is applied. While in the standby mode, a proper reverse body bias is applied to increase the threshold voltage and suppress the leakage current.

4.4.3 Active Mode Leakage Control

In this sub-section, three major circuit techniques for active leakage suppression – namely, dual threshold CMOS, dynamic threshold scaling, supply voltage scaling – are described.

Dual threshold CMOS

For a logic circuit, a higher threshold voltage can be assigned to some transistor in non-critical paths to reduce the leakage current, while the performance is maintained due to the use of low threshold transistors in the critical path. Dual threshold technique is good for leakage power reduction during both standby and active modes without delay and area overhead [4.18].

Dynamic threshold Scaling

Dynamic threshold voltage scaling is a technique to adjust the active leakage power based on the desired frequency of operation. This scheme utilizes dynamic adjustment of frequency through back gate bias control depending on the work load of a system. When the workload decreases, the threshold voltage is increased and less power is consumed. Figure 4.21 shows a block diagram of the scheme and its feedback loop [4.19]. A clock speed scheduler, which is embedded in the operating system, determines the (reference) clock frequency at run-time. The DVTS controller adjusts the PMOS and NMOS body bias so that the oscillator frequency of the VCO tracks the given reference clock frequency. The error signal, which is the difference between the reference clock frequency and the oscillator frequency, is fed into the feedback controller. The continuous feedback loop also compensates for variation in temperature and supply voltage.

Supply Voltage Scaling

Supply voltage scaling was originally the method for dynamic power reduction. However, it is also a effective method that helps reduce leakage power, since the sub-threshold leakage due to DIBL decreases as the supply voltage is scaled down. For lowering active leakage power, there are two ways of lowering supply voltage, static supply scaling and dynamic supply scaling. In static supply scaling, multiple supply voltages which are the practical ways of implementation has been discussed in section 2.2.3, while the dynamic supply scaling has been discussed in section 2.2.4.

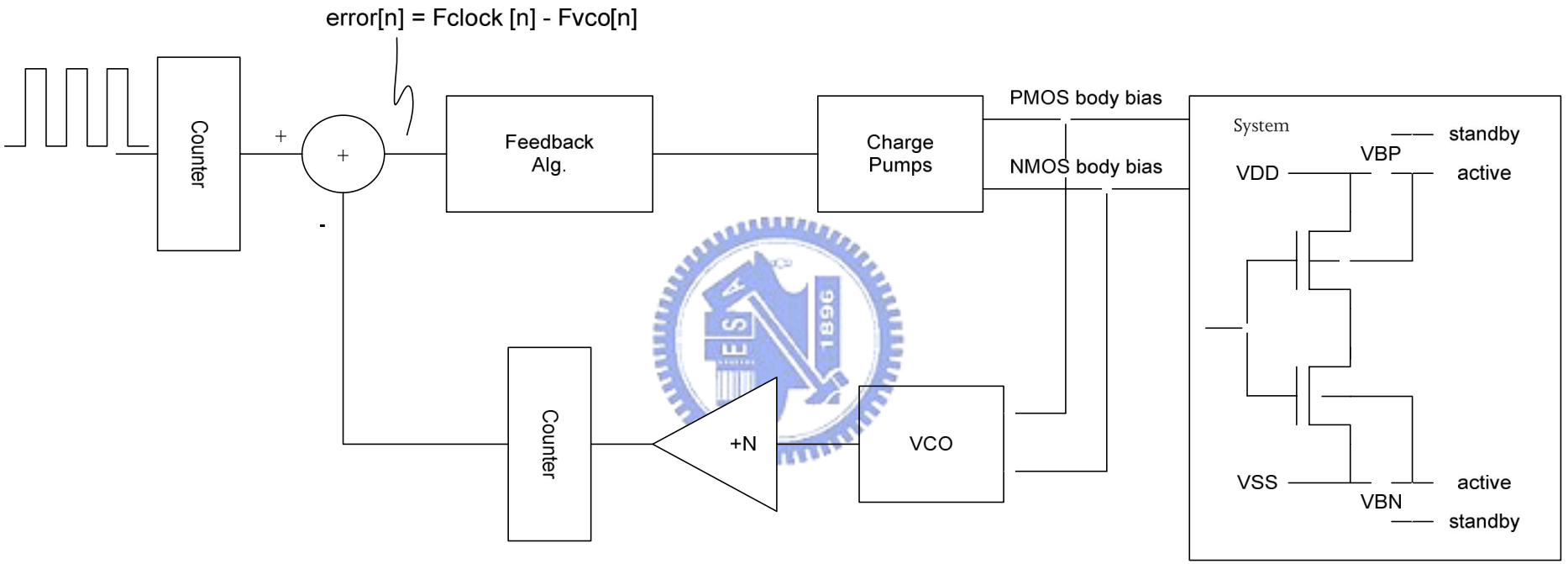


Figure 4.21 Schematic of DVTS hardware [4.19]

4.5 Conclusions

In this chapter, first we designed a low power Booth recoder by using transistor sizing. We used logical effort to characterize the efficient XOR gate as the subsequent circuit design of 5-2 compressors. Simulation results of 5-2 compressors show that dual rail DCVSPG XOR gates have attractive high speed but however, consume much more power than single rail design. After designing the efficient circuit topology and circuit style of compressors, we proposed a practical optimization procedure to carry out power and speed optimization of the column compression stage as well as the final addition stage. By adopting proposed optimization procedure, power saving of about 50% while speed improvement of about 18% can be achieved, respectively.