The 3rd International Conference on Materials Chemistry and Environmental Engineering
DOI: 10.54254/2755-2721/3/20230422

Evolution of Silicon Integrated Circuit: from Scaling to Lowering Power Consumption

Zhiyuan Zhang[1,†], Haoyang Zhang[2,†], Zonghai Jing[3,*,†], Xuquan Liu[4,†]

[1]School of advanced technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China
[2]Beijing eighty school, Beijing 100020, China
[3]ZJU-UIUC Institute, Haining 314400, China
[4]Institute of Physics and Electronic Science, Hunan Normal University, Changsha, 410081, China

zonghai.20@intl.zju.edu.cn
[†]All the authors contributed equally to this work and should be considered co-first authors.

Abstract: Semiconductor industry as a measure of a country's light industry level, has its own development history. This article will give a systematic account of the history of the technology used to produce integrated circuits on chips used in silicon electronics, as well as concepts and solutions to power consumption problems, for those who want to enter this industry. In the former case, we will quote data and compare and analyze them to highlight the characteristics of the development over a certain period. For the latter, we will argue and explain more from the principle point of view.

Keywords: Integration**,** Lithography, Power consumption, Dynamic power, Quiescent power.

1. Introduction

To understand the iteration of integrate circuit process technology, we must first understand Moore's Law. Introduced in 1965 by Gordon Moore, one of the founders of Intel, is that the number of transistors that can fit on an integrated circuit doubles every 18-24 months. The support behind it is Dennard Scaling: the operating voltage current of a chip shrink in equal proportion to the shrinking transistor size. Size reduction by 30% (i.e., becomes 0.7 times the original size). Then the length and width become 0.7X, 0.7Y, and the area is reduced by 0.7X * 0.7Y = 0.49XY (about 50% reduction). The size reduction reduces the effective capacitance, so the circuit delay is reduced by 30% (t → 0.7t) and the operating frequency becomes 1.4x (1/t → 1/0.7t = 1.4). Power consumption W=Pt, a 50% decrease. Some people may consider that the Moore's law is not proper nowadays because Dennard Scaling law failed from 2005-2010 due to leakage current. In fact, developments in manufacturing technology, particularly photolithography, do not appear to be well suited to keeping manufacturing costs low as device sizes are scaled down. Innovation in semiconductor manufacturing will certainly continue, but traditional feature size scaling methods are reaching their limits [1].

According to Wind data, the global market size of integrated circuits in 2021 is $460.8 billion,

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/3/20230422

accounting for 83% of the size of the semiconductor. Since the number of students entering the industry is gradually increasing, the number of people who need to understand the history of the industry to grasp the future trend is also increasing. Thus, we take a holistic approach to help beginners acquire a general knowledge.

In the work, we focus on the history of CMOS scale development, the successive generations of lithography applications, and the concept of power consumption problems and available solutions. We also discuss the industry outlook.

2. Scaling

2.1 CMOS development history:

The 50-micron lithography process is a semiconductor process technology used by early semiconductor companies in the mid-1960s. Where 50 micron refers to the effective channel length between the source and drain being approximately 50 microns. For example, Fairchild's typical wafer size using this process was 0.875 inches (22 mm).

The 16 μm lithography process was the semiconductor process technology used by semiconductor companies during the late 1960s. This process had an effective channel length of roughly 16 μm between the source and drain (Poly-SI channel implant).

The 10-micron lithography process was the semiconductor process technology used by major semiconductor companies between 1967 and 1973. At this point in time, the PMOS or NMOS technology used by most companies. And the wafer size now is 2-inch (take Intel as an example)

The 6-micron lithography process was a semiconductor process technology used by several semiconductor companies in the early to mid-1970s. From now on, 3-inch wafer is the standard size.

For a more visual representation, the following will be presented in a table.

Table 1. Evolution of CMOS.

SIZE	TIME	THINGS SHOULD BE KNOWN
3-micron	Mid 1970-Mid1980	Wafer size standard: 4-inch; x86 processor
1.5-micron	Mid 1980	Wafer size standard become 6-inch
800-nano	1990s	BICMOS, CHMOS V are being applied
600-nano	1990s	A full node semiconductor manufacturing process
500-nano	1992	A full node semiconductor manufacturing process
350-nano	1995	From now on using DUV photolithography
250-nano	1997	Polysilicon electrode (Intel)
180-nano	1998	First time to use Cu as interconnect
150-nano	2000	A half-node semiconductor manufacturing process
130-nano	2001	A full node semiconductor manufacturing process
90-nano	2003	12-inch wafer till now; ArF193nm photolithography
45-nano	2007	First time to use High-k and metal gate
22-nano	2008	22nm process for memory (2008) and MPUs (2012)
10-nano	2014-2018	Last generation before advance process
7-nano	2018	From DUV to EUV
5-nano	2020	A technology node following 7nm
3-nano	2022	From FinFET to GAA (Samsung)

2.2 High integration:

Integration is a very significant index which is usually used to judge the level of IC manufacturing technology. Integration refers to the number of components contained on a single chip. Usually, integration is negatively related to line width. In a MOS circuit, line width is usually defined by the

DOI: 10.54254/2755-2721/3/20230422

length of gate. Therefore, for a MOS circuit, the higher integration is, the smaller gate length is. With the development of IC fabrication technology, the integration of integrated circuits is becoming higher and higher. These improvements bring humans many benefits, however, at the same time, there are also some disadvantages of higher integration.

There are many advantages of higher integration. First of all, higher integration means a chip could contain more elements with same area. Therefore, for chips with similar area, higher integration usually means stronger computing power, leading to higher price. In addition, chips with higher integration could contain same amounts of elements with those lower integrated circuits with less area. Thus, for chips with similar complexity, higher integrated ones will take up less area, decreasing the cost of production. Moreover, in most cases, higher integration means better process, which results in lower power consumption. This is because the drive voltage is reduced, and the dynamic power consumption is proportional to the square of the voltage.

However, higher integration can also lead to some problems. First of all, since higher integration usually requires higher resolution ratio, the cost of research and development will increase. What's more, as more and more silicon circuits are integrated in the same small space, more and more heat is generated, decreasing the use period of the chip.

2.3 Different photo etching levels:

In this day and age, electronic devices have been put in use in various fields, including medical care, transportation and so on. With the rapid development of science and technology, the expectations and requirements of size, computing power and power consumption are becoming higher and higher. Therefore, one of the basic requisites is patterning, which is done by lithography [2]. Lithography plays a very important role in the production of integrated circuits. Humans have done much work on different lithography techniques. Figure 1 shows the percentage of work done on different lithography techniques in the last 5 years [3].

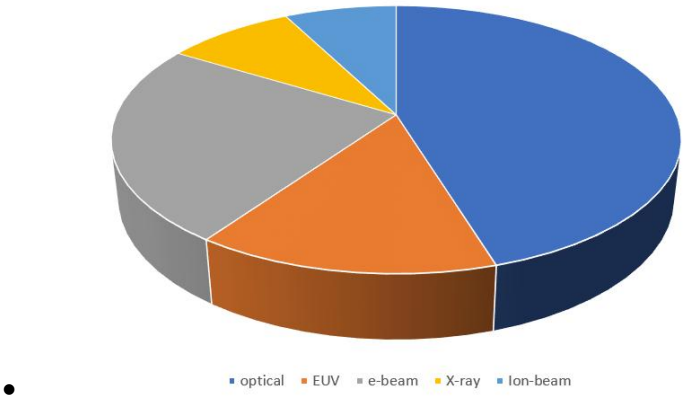


Figure 1. Percentage of work done on different lithography techniques in the last 5 years. Image is adapted from [3].

The data shows though some new techniques emerge, optical lithography is still the most popular one that humans study.

2.3.1 Optical lithography. In IC industry, optical lithography is the most widely used technique. The exposure system can be calculated by Rayleigh equation:

$$R = \frac{k_1 \lambda}{NA} \#(1)$$

$$DOF = \frac{k_2 \lambda}{NA^2} \#(2)$$

DOI: 10.54254/2755-2721/3/20230422

In the expressions, λ is the wavelength of the illuminated radiation, R is resolution, DOF is the depth of focus, k_1 and k_2 are Rayleigh constants, and NA is the numerical aperture [3]. Therefore, it is obvious that higher resolution can be achieved by lowering wavelength of light source or increasing numerical aperture. In the past few decades, there were numerous attempts made to reduce the wavelength. In the earliest time, since there were no compatible resist materials for DUV, the reduction of NA was developed faster. With the help of a step and repeat system designed for reducing projection exposure, a higher NA could be obtained. Not long after the appearance of stepper, a g-line source of mercury lamp under high pressures was adopted. The wavelength of this illuminated radiation was 436 nm. In 1970s, the numerical aperture was increased to 0.28 and therefore a 1μm resolution was obtained [3]. In middle 1980s, NA was increased to 0.5. In late 1980s, the light source i-line of 365nm was introduced. After that, in 1990, the first scanner was invented due to the demand for good resolution and low distortion [4]. Then in 1995, Nikon has developed 248nm excimer laser illuminated scanner with resolution reaching 25 μm [4]. Later the wavelength has been extended to 193nm in 1998 and 193nm water immersion version in 2004 by ASML [4]. At the same time with the development of immersion version, there was an extra method extending 193nm laser to 157nm with help of F2 laser. However, it failed in the competition with 193nm water immersion version.

2.3.2 EUV lithography. In most cases, optical lithography is used for integrated circuits over 100 nm. When it comes to devices with smaller size, a more advanced technique called “extreme ultraviolet lithography” is usually used. EUVL usually uses 10-14nm extreme ultraviolet light source [3]. EUV lithography technology has been developed since 1980s. In 1991 the practicability of EUVL was proved due to the successful experiment on the laser-plasma source [3]. There were some significant improvements in exposure tools in this century. Full-field exposure tool was prepared by the EUV LLC in 2001 [5]. This tool has 4 mirrors and a field area of $24 \times 32.5mm^2$. This tool enables resolution to achieve 100 nm. Then in 2004 and 2005, Exitech and Nikon developed new exposure tools which could decrease the resolution to 30 nm and 32 nm respectively. Later in 2010, NXE3100 was introduced by the famous corporation ASML in Holland. This tool has a NA of 0.25 which gives a resolution of 28 nm [3]. Then ASML developed a more advanced tool NXE3300B, which is able to help obtain a resolution of 13 nm. In addition, a resolution of 9 nm can be achieved with the help of double patterning technology [3]. New exposure tools are developed by ASML and ZEISS with high NA such as 0.55 targeting a resolution up to 8 nm [6].

3. Power consumption

With the continuous development of technology, the scale and integration of chips are constantly improving, and in this process, new obstacles and challenges are constantly emerging. Only by making greater breakthroughs in these fields can chip technology make great progress again. One of the challenges is power consumption.

According to Chenming [7], “By making the transistors and the interconnects smaller, more-circuits can be fabricated on each silicon wafer and therefore each circuit becomes cheaper. Miniaturization has also been instrumental to the improvements in speed and power consumption of IC s”.

In fact, the power consumption problem is actually a problem of energy conversion. According to the law of conservation of energy, energy is immortal, the sum of energy is unchanged, but energy can constantly change from one form to another form. The problem of power consumption is that the energy in the chip is not consumed in the form of useful work, but converted into heat. If the power consumption of a chip is too large, it is easy to cause the temperature to be too high, resulting in function failure, and even transistor failure. Therefore, it can be seen that reducing the power consumption of the chip is very important for maintaining the stable operation of the chip and reducing the power loss. There are two main power consumption sources, static power consumption and dynamic power consumption. Let's first make a general introduction to power consumption from these two parts.

DOI: 10.54254/2755-2721/3/20230422

3.1 Dynamic power consumption

The first is dynamic power consumption, the dynamic power consumption is derived from two parts, switch power consumption (Psw) and short circuit power consumption (Psc).

3.1.1 Switch power consumption. When the gate is flipped, the load capacitor is charged and discharged, which is called the switch power consumption.

3.1.2 Short circuit power consumption. When the gate is flipped, the load capacitor is charged and discharged, which is called the switch power consumption, as can be seen in figure 2.

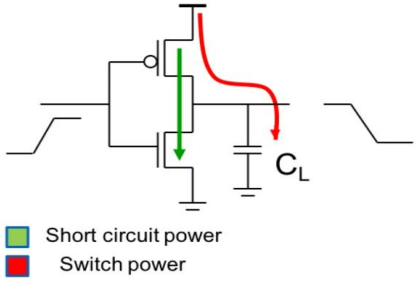


Figure 2. Switch power consumption and short circuit power consumption.

3.1.3 *Methods to reduce dynamic power consumption.* For these two kinds of dynamic power consumption, there are relatively good methods to reduce their influence.

For switch power consumption, we can reduce the power consumption by dividing the voltage domain and reducing the load capacitance. This is because the dynamic power consumption has a square relationship with the voltage, and reducing the supply voltage can significantly reduce the power consumption. The chip is divided into multiple voltage domains, each of which can be optimized according to the needs of a particular circuit. For example, a high supply voltage is used for memory to ensure the stability of the memory unit, a medium voltage is used for the processor, and a low voltage is used for the IO peripheral circuit which runs at a low speed. Capacitance comes from the wires in the circuit and the transistors. Shorten the cable length, good plane planning and layout can reduce the cable capacitance. Choosing smaller logical series and smaller transistors can reduce the turn-over capacitance of the device. According to Chenming [7], "Thanks to the reduction in C and V_{dd} , power consumption per chip has increased only modestly per node in spite of the rise in switching frequency, f and the doubling of transistor count per chip at each technology node".

For short circuit power consumption, we can also reduce its influence by increasing the input signal turnover rate. If the input signal turnover rate is relatively slow, the pull-up and pull-down networks will be on for a long period of time at the same time, and the short-circuit power consumption will be relatively large. Increasing the load capacitance can reduce the short-circuit power consumption, because when the load is large, the output will only flip a small amount during the input jump. When the input edge changes rapidly, the short circuit power consumption only accounts for 2%-10% of the switch power consumption.

It is worth noting that in the past, dynamic power consumption has always been the main source of power consumption problem, so people mainly focus on reducing dynamic power consumption. However, with the increase of chip integration and scale, as well as the continuous maturity of dynamic power processing, people now pay more and more attention to the impact of another power source, that is, static power consumption.

3.1.4 *Analysis of dynamic power consumption questions.* As we are looking deeper into what can

DOI: 10.54254/2755-2721/3/20230422

dynamic power consumption and quiescent power consumption can bring to us, we are basically talking about how can these properties relate to real-world applications. To discuss, we focus on three main topics: The reason of looking for lowering down power consumptions, the difficulties we face in the process, and the methods to solve the difficulties.

To begin with, main problems often occur when we reduce the scale of chips. When chips are growing smaller, we will unfortunately but undoubtedly hit on the wall of high-power consumption of chips on unit areas of chips, since less space are available for those integrated circuits to settle in. The power-consumption in unit-area increases most dramatically when the chip shrinks from ultra-deep submicron level to nano level. In this case, we would require way lower power consumptions to make smaller chips possible to be produced. Simply saying, it is the key problem for both modern academy field and industry field. To speak from another prospect, reducing power-consumption is still significant in real-world applications. Take wireless fidelity as an example. We install routers in our homes to make sure we are able to connect to the internet. However, to maintain the functionality of WIFI routers, we will be requiring chips in those routers to keep their function operating consistently for years, or even decades. If we don't put the power inside chips at their lowest state, we can prolong the lifecycle of WIFI routers.

In general, reducing power-consumption is extremely useful to applications whether in scientific researches or in financial analyzes, which is also why techniques about power-consumption in chips have been a hotspot in the field for decades.

As what have been introduced in the preceding passages, power-consumption has two types: dynamic and quiescent. Quiescent power consumption mainly comes from dissipated electric current and counter current when the switch of semiconductor transistors turns from 1 to 0 or 0 to 1. Then, these escaped electric energy transformed into heat energy, increased the temperature of the entire system. When the temperature exceeded a certain limit, the integrated circuits burned out. However, sadly, the existence of such things as dissipated current can only be erased in an ideal model but not in a real-world situation. In order to not let the chip gets on fire, we will need devices to bring down the temperature, which is the computer water cooling system.

This system is in fact not complex at all. It mainly serves to cool down CPUs and GPUs by utilizing water's greater specific heat capacity. The heat is conducted from computer components to water, and then water carries heat to heat sink, so that water can be cool again to conduct heat. The cycle repeats to achieve the goal of cooling temperature. The main advantage of water-cooling system is that it makes barely no noise, and it has a pretty long maintenance period; in contrast, air fan-cooling system makes great noises and often need fixing. However, water-cooling system is comparatively speaking more expensive.

Next, let us focus on a scientifically higher-level method of reducing power-consumption: Dynamic Voltage Scaling, or simply DVS. This is a mathematics-based algorithm that intuitively reduces power-cost. To illustrate the principle of DVS, we would need to know a formula first. The expression of electric power in integrated circuits has one kind of expression:

$$P = \alpha CV^2 f \#(3)$$

In this equation, P represents the power consumption, and V is the source voltage. C, $\boldsymbol{\alpha}$ are constants, and f is the frequency of the circuit changing from 0 to 1 or 1 to 0 [8]. This formula brings us a connection between V_{dd}, the voltage of the source: battery, and the dynamic power of integrated circuit. And the square on top of V_{dd} is what stands for the basement of

DOI: 10.54254/2755-2721/3/20230422

DVS system. Because of the square, we can manage the power consumption efficiently from controlling source voltage. Also, the DVS system linearly decreases power consumption by reducing the flipping frequency to reduce p-consumption. As we know, the workload of chips changes in real-time, therefore we don't need the chip to maintain a peak performance for the entire time when it is functioning. Hence, by using DVS, we can dynamically switch the chip's required working voltage and frequency to operate at the moment.

DVS was a hotspot in academic field. In recent years, however, many researches about DVS realized that adjusting source voltage can only be accomplished in a certain range. In other words, it has limits. In the last few years, UDVS (Ultra Dynamic Voltage Scaling) has been invented [9]. As an upgrade of DVS, it dramatically amplified the range of system's working voltage, therefore is able to lower the source voltage to subthreshold region under the situation of making sure the electrical system is working in the right way, and the parameters are correct, and then decrease the p-consumption.

However, even powerful as UDVS has disadvantages. UDVS can only be used by adding special circuit structures into integrated circuits to control the source voltage. Unfortunately, when people build integrated circuits, there are two ways: Full-custom IC and Half-custom IC. Full-custom IC is the kind that people designed first and then construct. It is more precise and stronger than half-custom, but it cannot be changed once the construction completed. In contrast, Half-custom IC is a kind that is assembled by many small electrical components. Compared to Full-custom, Half-custom IC is more flexible and costs lower, but it is also easier to breakdown. For UDVS, it can only be applied to Full-custom integrated circuits, but is rarely applied to large-scale IC, because of technical difficulties.

Scientists have been working on reducing power consumption for many decades, but there are still innumerable problems waited to be solved.

3.2 Static power consumption

The static power consumption mainly comes from 4 sources: sub-threshold leakage current flowing through a cut-off transistor, leakage current flowing through the gate medium, leakage current at P-N section in source-leakage diffusion zone, and the competitive current in a specific circuit.

3.2.1 *Sub-threshold leakage current.* It is the current flowing through the transistor when it should be cutoff. As can be seen in figure 3, before the 90nm node, the leakage power consumption is mainly considered in sleep mode because it is negligible compared to the dynamic power consumption. However, in the process with low threshold voltage and thin gate oxygen, the leakage current accounts for one third of the total power consumption. According to Chenming [7], "At the 22 nm node, new transistor structures may be used to reverse the trend of increasing I_{off}, which is the source of a serious power consumption issue".

DOI: 10.54254/2755-2721/3/20230422

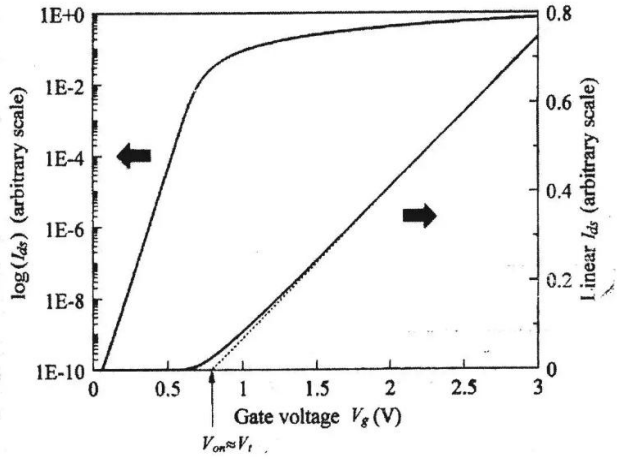


Figure 3. Sub-threshold swing.

3.2.2 *Gate leakage current.* It occurs when a voltage is applied to the gate and a carrier is driven through a thin gate medium. This leakage current is strongly related to the thickness of the medium and also depends on the gate voltage.

3.2.3 *Junction leakage current.* It occurs when the source or leakage diffusion zone is at a potential different from that of the substrate. Junction leakage currents are usually small when compared to other leakage currents.

3.2.4 *Competitive current.* Although static CMOS circuits do not have any competing currents, certain other circuits draw currents themselves even when static. Current-mode logic and many analog circuits also draw static currents.

3.2.5 *Methods to reduce static power consumption.* For static power consumption, there are also some common ways to reduce its impact, such as using variable threshold voltage and input vector control.

Variable threshold voltage refers to the threshold voltage modulated by volume effect. Apply a reverse body bias in sleep mode to reduce leakage. Use a forward body bias in operating mode to improve performance. In this way, static power consumption can be reduced.

Input vector control is the application of a set of input patterns to minimize module leakage when the module is placed in sleep mode. Stacking effects and input ordering can cause changes in sub-threshold leakage and gate leakage. So, the leak of a logic module is related to the gate input. And we can reduce static power consumption by reducing leakage.

4. Conclusion

Generally, the whole passage has discussed two main things: Techniques of producing IC on chips that are used in silicon electronic components, and concepts and solutions of power consumption problems. For the former we discussed from both technical prospect and financial limitation and preferences prospect. For the later we discussed through basic concepts of dynamic and quiescent power consumption and how can Dynamic power consumption be reduced in real situation.

References

[1] Mack, C. A. (2011). Fifty Years of Moore's law. IEEE Transactions on Semiconductor Manufacturing, 24(2), 202–207. https://doi.org/10.1109/tsm.2010.2096437.

[2] Pease, R. F., & Chou, S. Y. (2008). Lithography and Other Patterning Techniques for Future

DOI: 10.54254/2755-2721/3/20230422

Electronics. Proceedings of the IEEE, Proc. IEEE, 96(2), 248–270. https://ieeexplore-ieee-org.ez.xjtlu.edu.cn/document/4432794

[3] Sharma, E., Rathi, R., Misharwal, J., Sinhmar, B., Kumari, S., Dalal, J., & Kumar, A. (2022). Evolution in Lithography Techniques: Microlithography to Nanolithography. Nanomaterials, 12(16), 2754. https://doi.org/10.3390/nano12162754

[4] Wu, Q., Li, Y., & Zhao, Y. (2020). The Evolution of Photolithography Technology, Process Standards, and Future Outlook. 2020 IEEE 15th International Conference on Solid-State & Integrated Circuit Technology (ICSICT), Solid-State & Integrated Circuit Technology (ICSICT, 2020 IEEE 15th International Conference On, 1–5. https://doi-org.ez.xjtlu.edu.cn/10.1109/ICSICT49897.2020.9278164

[5] Tichenor, D.A.; Ray-Chaudhuri, A.K.; Replogle, W.C.; Stulena, R.H.; Kubiaka, G.D.; Rockett, P.D.; Klebanoffa, L.E.; Jeffersona, K.L.; Leunga, A.H.; Wronoskyb, J.B.; et al. (2001). System integration and performance of the EUV engineering test stand. Proc. Spie, 4343. https://doi.org/10.1117/12.436665

[6] Van Schoot, J.; van Setten, E.; Troost, K.; Bornebroek, F.; van Ballegoij, R.; Lok, S.; Stoeldraijer, J.; Finders, J.; Graeupner, P.; Zimmermann, J.; et al. (2019). High-NA EUV lithography exposure tool progress. Proc. Spie, 10957, 1095707. https://doi.org/10.1117/12.2515205

[7] Chenming, C. Hu. (2009). Modern Semiconductor Devices for Integrated Circuits. Upper Saddle River: Prentice Hall.

[8] Shen, C. (2011). Research on Key Technologies of Super Dynamic Voltage Regulation for Low Power SoC Chips (in Chinese). (Doctoral dissertation, South-east University).

[9] Huang, J. (2009). Low-power computing for embedded mobile system Software with DVS support. (Doctoral dissertation, Zhejiang University).