Technology Leverage for Ultra-Low Power Information Systems

J.M.C. Stork

ULSI Lab

Hewlett Packard Laboratories[1]

Abstract

Many applications for future generations of logic and memory chips will be requiring highly sophisticated computing functions at low cost. Small form factors, portability, and low cost will require low power operation. While continued scaling of silicon technology to dimensions below quarter micron devices and interconnections appears technically feasible, higher levels of integration and operation at higher speed have been driving the power consumption of logic chips up instead of down. This paper discusses how scaled submicron silicon technology can provide leverage to reduce power, while gaining in throughput for logic chips, and in capacity for memory functions. Strong reductions in voltage supply have to accompany shrinking dimensions. Materials limits such as tunneling currents through ultra-thin silicon-dioxide gate dielectrics and electromigration in minimum pitch interconnections emerge to be key challenges to realize low power 0.1 μm level CMOS circuits. A more than 10x gain in productivity as measured by the energy*delay product can be realized by shrinking from 0.5 to 0.125 μm CMOS device technology.

---

[1] Much of this work was accomplished at the IBM T.J. Watson Research Center

Introduction

The largest leverage and perhaps one of the most important justifications for submicron silicon technology appear to be hand-held products [1]. These products require all of the features of deep submicron CMOS, namely high integration and performance at low power. The cost of computing has decreased by several orders of magnitude during the past two decades as indicated in Fig.1 [2]. With the advancement of technology, the mainstream computing platform has changed from traditional mainframes and minicomputers, to desktops and laptops. With the obvious reduction in size, power and cost have also diminished. Whereas it took $10M and 10 kW to afford and power up a 10 MIPS machine twenty years ago, next generation game machines will have microprocessors that make the same computing power available at only $10 and with less than one Watt of peak power consumption. These revolutionary changes in functionality have come about in part through new architectures, such as server-client system connections and RISC (Reduced Instruction Set Computing), and design methodologies, such as custom, dynamic CMOS circuits, but would certainly not have been possible without the reduction in feature size of the semiconductor chips.

Note the trend shown in Fig. 2 of the area used by microprocessor chips. With the reduction in feature size, the area required to implement a "typical" 64 bit general purpose processor has decreased to the point that it is now possible to build a very complex processor on a single chip. If the increase in chip size continuous as indicated (see also[3]), it will be possible by the end of this decade to integrate the circuitry required to build a mainframe general purpose processor on a single chip! It is extremely unlikely that these applications will drive sufficient demand to build billion dollar fabrication facilities (so-called megafabs), with millions of chips per month capacity. Indeed, it is very likely instead that volume production in the semiconductor fabs will follow the application trend: as most of the technology was developed and built for mainframes in the early days, and as desktops are clearly driving the microprocessor and memory (especially Dynamic Random Access Memory) markets today, so will long term future growth for semiconductor parts be the "computing appliance" market. This forecast signifies a dramatic change to the figures-of-merit for developing new generations of silicon technology: the technology for consumer oriented products will no longer be the second generation, the one behind the state-of-the-art, cheap general purpose process. This technology will be the leading edge itself. One may even speculate about whether business applications will continue to be the entry point for new technologies or design and architecture. Likely, high-end game products could be attractive entries as well. In any case, in addition to speed, power per function will be of primary importance to many future applications. In addition, low cost requirements will drive towards integration of logic, memory and analog functions, while maintaining or better reducing the complexity of integration. But, since a generic process that can satisfy all but the most demanding tasks may be too complex and costly, differentiating process options should be developed in close synergy with design to enable what is exercised in the automobile industry as "mass customization". It is the ability to use economies of scale to reduce product cost, but maintain the flexibility to differentiate for unique, almost individual, taste and usage.

Since hand-held products have limited space for displays, keyboards and cables, these conventional connections will be at first be amended, and eventually perhaps replaced, by speech and pen for the human interface, and with RF/infrared signals for tele- and data-communication.

The focus of this paper is on the chip technology, and therefore no attention is devoted to future display, storage and power (battery) technologies. To level our thinking about establishing the technology demands of future products, let us briefly review how scaled technology could impact a speech recognition system. This will provide a guide and a connection for the subsequent sections, that study in more detail the scaling path of the logic processor, the trends in memory, and at last, some highlights of the key accomplishments and challenges of 0.1 μm CMOS and Giga-bit memory to date. The requirements for analog functions, and the potential difficulty of integrating these with the digital parts on a single chip will be discussed briefly before concluding.

Communication with small, usually portable, devices is most easily accomplished by speech, were it not that a present implementation requires a significant amount of space and power[2]. For example, in order to implement a 20,000-word dictation vocabulary (or equivalently a 1,000 words continuous speech) system, a full board is needed, and the required power with today's components is on the order of 20 Watts at peak operation. A generic speech recognition system requires use of the computational power of a typical host microprocessor, specialized DSP (Digital Signal Processor) functions, DRAM (Dynamic Random Access Memory) and (optionally) EEPROM (Electrically Erasable and Programmable Read Only Memory) chips, and finally also analog-to-digital (A/D) and digital-to-analog (D/A) converters (see Fig. 3). A large fraction of the power requirement results today from the assumed use of 4 Megabit DRAM chips. As will be shown later, these may consume as much as 1 Watt per chip. Using instead a byte wide organized 256 Mega-bit chip, the total chip power needs reduce to only 100 mW, and

[2] Development of algorithms has not been perfected either, but recent efforts have focussed on the practical implementation of "adequate" speech vocabularies.

provides increased memory. Within the same two to three generations of scaling as the memory, the same processor and DSP functions could be implemented on only a few square millimeters of Si area, and consume less than 100 mW in power. This will be discussed first in the next section, using a first-order microprocessor power/performance model. It forms the core section of this paper. As the required area to implement the original speech system shrinks, off-chip I/O power may become the dominant power consuming function [4], unless a significant amount of memory (usually Static RAM) is provided on chip, as well as the analog-to-digital conversion functions. However, although the latter function may only require about five percent of the transistor count, the additional technology and circuit design complexity that arises from combining analog and digital signals on the same silicon die presents a difficult trade-off for the technology developers. Perhaps novel substrate technology, such as SOI (Silicon On Insulator) with its much better cross-talk characteristics [5] will enable easier integration of analog and digital circuits. These promises and challenges of SOI and conventionally scaled CMOS technology will be discussed in more detail in the last section.

One key point needs to be re-emphasized here, namely that of final cost. The technological capability of providing a speech system on one or two chips (or as an added function to the ubiquitous microprocessors on microcontrollers) will only be economically attractive if the cost will be reduced as well as the size and power. Only then is there an opportunity to increase the market volume for the semiconductor parts, generating the required revenue stream for continued technology scaling. In summary, using scaled technology for the logic and memory chips, both form-factor and power consumption can be reduced drastically. With the right cost structure, and undoubtedly with many advances in design and system aspects, such a technology will be a key enabler for the pervasive application of speech, image and other sophisticated technologies not only in conventional computers, but in any appliance or business tool, including very small portable, battery powered, information systems.

Microprocessor scaling

The power need of microprocessors has grown instead of decreased as would perhaps be expected (see Fig. 4). Even the recent voltage reduction from 5 to 3.3 Volt seems to have had little impact. The higher clock frequency, larger chip size and higher density of devices keeps the power of the chip high. It seems that there is a plateau at about 30 Watts, which could be a direct consequence of the escalating packaging and cooling costs for power densities on the order of 50 Watts/cm^2 or higher. Such high power severely limits the application market, especially the growing low power segment. In order to understand the trend in Fig 4. and possible alternative future directions, the next two sections discuss first how today's microprocessor would operate when implemented in next generation technologies, followed by how the scaling to higher integration might develop.

Constant Functionality Scaling

A simple, hypothetical, geometric scaling scenario for a general purpose processor (e.g PowerPC or Pentium[3] chips) is shown in the table in Fig. 5. Because no additional functions are added, the speed performance of the chip can easily be estimated. As all chip dimensions shrink uniformly, the wire delays stay constant. The device limited circuit delays reduce approximately proportional to the scaling factor [6]. Assuming for simplicity of argument that the original

[3] PowerPC is a trademark of the IBM, Motorola, Apple Corporations, and Pentium is a trademark of the Intel Corporation.

design was balanced in wire and device delay, then the global clock speed could increase approximately two times in four generations. At the other extreme, if wire delay was negligible in the original design, a factor of four should be realizable. Calculations taking the PowerPc 604 as example, indicate a factor of 2.3 (3) is possible in three (four) generations.

For a constant system performance, processor speed can be kept steady, allowing a further reduction in size and power as indicated in the lower half of the table in Fig. 5. The methodology to arrive at these results can be described as follows. Assuming that the starting design was optimized for speed, the design point for the next generation must be changed to maintain the same (low) clock speed. This allows a significant reduction in power dissipation, because the full device performance potential of the next generation is not utilized. This in turn allows a reduction in channel width and, if density is not completely limited by wireability, a smaller cell pitch. Local capacitance is then reduced at the same time, calling for another iteration to optimize on this new low power design point. Note that this process takes advantage of the improved power*delay product that scaled technology offers not only to reduce power, but also to reduce chip size and therefore cost as well. Unfortunately, I/O power and packaging present an increased challenge for such small chip dimensions. Especially for the increased performance die, I/O performance will degrade if the pad and off-chip capacitance is not improved. Even in the constant speed performance scenario, the I/O power will become a larger, dominating the total power need [4]. These simple estimates shown in Fig. 5 indicate the substantial power and cost reduction possible by technology only. No improvements in circuit, system, architecture and other higher levels of abstraction have been assumed.

Logic Power and Delay Sizing Model

The model used to generate the data in the table in Fig. 5 in the previous section is an extension of previously published analytical power and delay models [4,6]. Its key features and assumptions will be described in this section for the expert and/or interested reader. Without much loss of continuity one can skip to the next section, which discusses the results of several analyses.

The analytical details of the model are developed after [4,6] to which the reader is referred for complete details; here a brief summary follows of the major features included - specifically those added to ref [6]. The methodology of the model is to make an approximate calculation of the speed performance and power requirements for a "typical" logic system by calculating the total RC delay of a series of locally connected gates and one global interconnection. The delay calculations therefore take into account the technology specifications of the transistors, as well as of the local and global interconnections, through calculating their effective resistances and capacitances. To determine the local interconnect RC, a typical length is derived from establishing the total chip area requirements, taking into account the number of logic circuits, the number and pitch of the interconnection layers, as well as the amount of on-chip SRAM memory. The global interconnection is assumed to be made on the top level (largest cross-sections) layers across the diagonal of the chip. Power calculations include the dynamic power of the logic circuits, the clock distribution, the off-chip drivers, and the off-current leakage currents. The off-power is calculated from summing the off-current over all circuits. The logic, clock and off-chip power calculations require the calculation of the total logic circuit capacitance (device and interconnections), the total clock capacitance, and the sum of the I/O drivers and pads capacitance. These capacitances are then multiplied with the obtained clock frequency (inverse of the sum of the RC delays described above) and the supply voltage squared to obtain the maximum power. The actual power is then reduced by a prefactor, an effective duty cycle, which was adjusted once to match the data of the PowerPC chip.

The two independent variables are supply voltage and interconnection pitch. Supply voltage ranges from 5.0 V down to 1.25 V, and minimum dimensions correspond to generations ranging from 0.7 um to 0.18 um CMOS. Threshold voltage is related to the supply voltage, while interconnect pitch and effective channel length are derived from the gate length dimensions to complete the specification of the first order technology parameters.

Here a brief summary follows of the major features included, specifically those added to ref [6]. The dimensions of the CMOS gate length are related to the interconnect and lithography generation in the following way. The drawn gate length (L_{gate}) is set at 70% of half the minimum interconnection pitch, and the effective channel length (L_{eff}) is taken at 70% of L_{gate}. For example, 0.5 um (L_{gate}) CMOS implies L_{eff}=0.35 µm and minimum interconnect pitch of 1.4 µm. With regard to the representation of the device performance, velocity saturation is assumed to determine the voltage dependence of the drain current. A constant (per unit gate width) source/drain resistance is added in the calculation of the effective device resistance. Unless explicitly mentioned, a constant W/L ratio of 10 is used. The threshold voltage is reduced proportional to the square root of the supply voltage, from 0.8 V for a 5.0 Volt supply to 0.4 V for a 1.25 Volt supply. The impact of the increased off current with lower threshold voltage is included in the estimation of the total standby power. The total number of circuits is increased proportional to the square of the dimensional reduction, resulting in a slightly increasing chip size as minimum dimensions shrink with future generations. The number and pitch of the interconnection levels is calculated as follows: starting with 3 levels of interconnect - the bottom two levels with a 2 μm pitch and a top layer with 3 μm - for the 0.7 μm CMOS generation, each successive generation[4] adds another layer at the Si (bottom) level, with a 0.7x reduced pitch. At the 0.18 μm generation there are therefore 7 levels with the smallest wires being 0.25x0.5 μm^2 in cross-section for a 0.5 μm pitch. The effective interconnection length used to determine the delay of local resistance and capacitances, is dominated by the smallest wire cross-section. The current density through the smallest wires is limited to $10^6 A/cm^2$ to take into account a possible impact of electromigration on the design rules. Note that the total sum of all the local interconnection wires makes up most of the capacitance and is thus mostly responsible for the dynamic power consumption of the logic circuits. This is in contrast to the dominant role of the global wires on any critical path delay. Therefore, these global connections are assumed to be implemented on the upper levels, limiting the impact of long distance connection delays relative to the (increased) clock frequency.

Memory speed and power are not analyzed in detail; only the area requirements are accounted for (having an indirect effect on delay and power by growing the chip size). Starting with 16 KByte (8 bits per byte) cache at the 0.7 μm generation, total cache size doubles every generation. Although this is a conservative estimate, it is also assumed that the required 6 transistor cell area continuous to be constant in term of the number of minimum feature squares,

---

4 One generation is defined as a square root of 2 reduction in dimension and supply voltage

consistent with the assumption above of no novel circuit design concepts. The net result is that in the following examples, the distribution of memory versus logic area remains roughly constant with every generation. Although recent data suggests that on-chip cache has been growing slightly more rapidly, since the amount of 1st or even 2nd level cache is strongly coupled with the chosen architecture, optimizing the amount of memory is beyond the scope of this model, which focuses on the impact of technology alone.

Scaling Scenarios

In this section the impact of technology scaling (voltage and size reduction) on the delay and power of a generic logic systems will be discussed, highlighting that even with a quadratic increase in circuitry per generation, computational productivity as measured by the improvement in speed and energy*delay product can expect gains of an order of magnitude over the four generations of technology considered. In other words, for a sixteen-fold increase in circuits, a 2x speed increase can be obtained at five times lower energy. The simplicity of the analysis neglects the fact that improvements in circuit design and architecture could easily affect these estimates by a factor two or more as well. The sole objective of the analysis is to show the effects of technology improvements separately. A serious attempt is made to include all important technology parameters, such as scaled device characteristics and limits at lower voltage (more about that in the last section), embedded SRAM area requirements, scaled interconnection characteristics and limits, clock distribution and off-chip (I/O) requirements. The basic parameters have been calibrated against published data of the PowerPc 604 chip to provide a credible starting point.

The model described in more detail in the previous section, allows one to study the impact of device and interconnect optimization on the speed and power of a typical delay path in a microprocessor. Similar work has been reported in [4,6]. Supply voltage and gate length are chosen as two independent variables, ranging from 5.0 Volt to 1.25 V and from 0.7 μm to 0.18 μm respectively. Threshold voltage is assumed the scale with the square root of supply voltage reduction, from a high 0.8 V to a low of 0.4 Volt. The effective channel length of the devices is taken to be 70 % of the gate length, while the gate length itself is 70 % of half the minimum interconnection pitch. The clock frequency is calculated from summing the RC delays of a series of local gates (the number of the series being defined as the logic depth, taken to be 15 in the following discussions) and a single global interconnection running across the diagonal of the chip using the top level, largest cross-sectional wires. The local RC delays include contributions from the device capacitance and resistance, as well as the local interconnection dimensions and typical lengths. The latter are a function of the total number and size of gates, the number and density of the interconnection layers, as well as the area required by the on-chip memory array. Maximum active power is calculated using $f_c C_{tot} V^2$, where f_c is the calculated clock frequency, and where the total capacitance C_{tot} is determined for the logic circuits, the clock distribution, and the off-chip drivers and pads. Actual power is reduced by the duty cycle, assumed to be 15 percent, which provided a decent match with available data. The off-power is calculated as well by summing the off-current over all gates. In Figs. 6a and 6b the clock frequency and power contours are plotted versus gate length and supply voltage. To maximize the leverage of low voltage operation, the CMOS device design is adjusted for the supply voltage by reducing the gate oxide thickness linearly from 14 nm at 5.0 Volt to 3.5 nm at 1.25 Volt, independently from the channel length. Drain current density is likewise increased. However, limiting our focus on low power, the shortest channel lengths are not redesigned to be consistent with the highest voltages, at least not in the simple model used here. Therefore, the data points farthest away from the diagonal at the high voltage supply have little credibility. Because the number of gates was chosen to increase quadratically with reducing the minimum feature size, the die size is approximately constant. As can be seen from Fig. 6a, at larger dimensions, the clock frequency is only very weakly decreasing with voltage, mostly a result of the scaled oxide thickness with voltage. In fact at the smallest geometries considered, a reverse trend is observed: the highest speed is achieved at the lowest voltage. Because of velocity saturation, the potential increase in drain current with supply voltage is significantly reduced. Combined with the fact that the local interconnect capacitance does not scale and has therefore increased relative to the gate capacitance at the smallest features, the local gate delay actually increases slightly with voltage. Considering that the power is reduced approximately quadratically with voltage as shown in Fig. 6b, low voltage operation presents also a much more energy (energy is simply power*delay) efficient design point. Finally, from a point of device reliability low voltage design is again the most attractive design point. The more difficult question is however what combination of voltages (supply and threshold) and feature size (channel length and interconnection pitch) provides the most power efficient design point. This cannot be answered in isolation from the application constraints, but it will be useful to clarify what metrics have the most relevance for high efficiency or high throughput designs.

Consider first Fig. 7 where clock frequency is now plotted against power. As expected, the highest frequency is reached at about constant power density for every generation, but note that the optimum shifts from the highest power at large groundrules, to the lowest power at the smallest groundrules. The lowest energy design points, in terms of mW/MHz, always lie at the low power end of all curves, with significant reduction in speed at today's feature sizes. Energy is clearly not an adequate figure of merit in general, since it provides no optimization or boundary condition for either power or delay. In order to better evaluate the trade-off between energy and delay, it has been suggested before to use the energy*delay product as a figure-of-merit [7]. This choice is physically interesting, because it implements in a simplistic way the calculation of the action (has dimension of energy*delay) of the system, which according to the physical laws should be at an extreme value [8]. When the relation between power and delay is not simply linear, which is mostly true at the high speed part of the design curves (see Fig. 7), the simplest interpretation of a design point with "least action" can be described as that beyond which one has to spend more power to gain the same improvement in delay or below which the reduction in power causes a larger than proportional degradation in speed. Figs. 8a and 8b show an example by plotting clock frequency and action (energy*delay) versus power. In these figures, power is changing as a result of varying the width to length ratio of the CMOS devices. Voltage and feature size are scaling proportionately in these examples, as opposed to the cases in Figs. 6 and 7, where the W/L ratio was kept constant. From Fig. 8b it is apparent that the least action is obtained for W/L ratios close to ten - the fixed ratio in all the other examples. Very little frequency improvement is seen for the largest W/L ratios, while the action increases rapidly, indicating sharply increasing power caused by high voltage as well as less density (higher capacitance).

In summary, in order to optimize the technology for the widest set of applications, it is important to consider the energy*delay product as a figure-of-merit. This is particularly useful as scaling progresses from "today's" 0.7 µm features, where the highest speed is always achieved at the highest possible supply voltage, to future geometries of sub 0.25 µm, where maximum frequency of operation occurs at the lower possible voltages.

Low Threshold Voltage for Low Power

As a final case study, threshold voltage scaling was decoupled from the supply voltage to understand the trade-off between increased off-current and higher overdrive for the devices. This study was executed in more detail using FIELDAY and SPICE simulations, calibrated with experimental hardware [9]. The circuit considered here is a 3-way, loaded, static CMOS NAND gate. Conceptually, the design space is bounded by moving the lines of constant active (Pact) and standby power (Pstd) in one direction, and that of constant clock frequency (Fclk) in another, as shown in Fig. 9a. The higher the performance, the smaller the design freedom for choosing threshold and supply voltage levels. The next question then is again one of minimizing action, namely, to find how much power can be saved by limiting the speed to more modest value. The results are shown in Fig. 9b for two scaling scenarios, namely with a 2x and 1x increase in speed between the 0.5 and 0.175 µm generations of technology. The power savings are substantial. This case can be improved further by evaluating the effect of making two different NMOS threshold voltages available: one low enough to get about the same current drive as in the earlier generation at higher voltage, the other high enough to enable one to turn-off blocks or sections of a chip timely and effectively. Although this early study has improved our understanding of the possible new design trade-off's with multiple threshold voltages at deep submicron devices, much further work is needed to understand what circuit design techniques or changes in algorithms and architecture would reap the most benefit from this approach.

In closing this section, it is useful to address the key limits that power dissipation poses upon the scalability of CMOS technology. The most important in the author's opinion is that the key applications to support continued development of future generations of technology thrive from the higher level of integration in Si, by providing more function at higher speed with less energy. These product markets appear to be more consumer and mass market oriented, and therefore do not allow, from a cost and space point of view, a dissipation of more than about one or at most a few Watts of power. Secondly, the implications of higher power on reliability are very dramatic, because of the high temperature dependence of electromigration. Current densities will be driven as high as possible even in very low voltage designs, so lower operating temperature is imperative. Finally, the issue of supplying high current at low voltage is not trivial. Small voltage drops cause significant inefficiencies. A simple example will illustrate this: suppose the chosen design point was to provide 250 MHz with a 0.7 μm pitch technology at 2.5 Volt, the current input to be supplied to the chip is 5 Amperes! Power distribution with less than 10% voltage drop (250 mV) requires an interconnect resistance of less than 50 milli-Ohm. This is quite a technical challenge for thin film wiring, and certainly difficult to implement at low cost.

Low Power Memory

During operation of a speech recognition system, our earlier application example, most of the instantaneous peak power is dissipated by the DRAM memory chips. Future generations reduce the power requirement for a fixed amount of memory considerably. Fortunately, steady state power dissipation is often not the limiting factor for DRAM intensive applications, and the refresh cycles of new DRAM generations continue to get longer. Because key reduction in DRAM power has been obtained from the development of next generations, it is instructive to review the historical trends and a few of the key limits that will be faced in the near future. With the many new organizations introduced recently [10], it goes beyond the scope of this paper to discuss all the possibilities of lowering the power in memory chips.

DRAM Power Consumption Trends

In general, the power consumption of DRAMs has decreased rapidly during the past several generations. While operating current has stayed nearly constant, the internal voltage has been reduced steadily, while the number of bits quadrupled with every generation. As shown in Fig.10 , the power per chip has dropped by one order of magnitude, and the number of bits has grown from 1Mbit to 256Mbit over the past decade. The improvement per bit is an astounding factor of 10,000 or approximately 10x per generation. Where 8MByte of DRAM implemented with 4Mbit chips required as much as 8 Watts, an implementation with 256Mbit chips provides four times the memory for less than 100 mW. The continuation of this trend will probably extend through the 1Gbit generation, but the challenge to develop even higher densities remains unsolved at this point. One of the key reasons is that new architectures will be needed to reduce the area per bit for the next lithography step. As shown in Fig. 11, an extrapolation of the bit area required, and available (assuming a yieldable, cost effective die size [11]) bit area are essentially overlapping at the 1Gbit generation (180 nm lithography). In other words, no design margin is left to implement a conventional capacitor-and-transistor cell in the conventional folded bitline architecture. The number of minimum lithography squares can be reduced from 8 to about 6 by introduction of a hierarchical bitline organization [10,12], but this introduces a new challenge to minimize the capacitance per addressed line to keep the power low and speed high. Ultimately one might expect that memory cells would need only slightly more area than 4 minimum squares, making it likely that conventional DRAMs could be extended to 1 and 4 Gbit density.

The most effective way to reduce power of any given memory size is to lower the voltage and increase the effective capacitance to maintain sufficient charge in the cell. Lowering the voltage on the cell plate also reduces the voltage on the sense amplifier, particularly when the bitline capacitance is reduced at the same time. Because of the capacitive voltage division between storage and bitline capacitor, the ratio of Cbit/Ccell has to decrease to guarantee a large enough signal to sense and amplify. This effect is schematically presented in Fig. 12 (adapted from [10]). The specific capacitance of the storage cell can be increased by reducing the dielectric thickness and increasing the dielectric constant of the material. Recent materials studies have shown the promise of new materials like $BaSrTiO_2$ for use in decoupling and memory capacitors. The success of these and other new materials depends on two critical factors: leakage current and conformality. By equating the required charge capacity and the leakage rate in a given time, a key relation between dielectric constant and material resistivity is derived as follows: The charge per unit area that can be stored equals simply $\varepsilon_r V/d$, and the charge that leaks off through the dielectric in time τ_{refr} equals $\tau_{refr} V/\rho d$ where d is the thickness of the dielectric. Equating both results in the requirement that the product of $\varepsilon_r \rho$ must exceed the refresh cycle time constant. A plot of the recently available data is given in Fig. 13. It is worth noting that this requirement is independent of voltage and capacitor area. To accommodate the minimum required charge, the area per cell can be smaller for a higher dielectric material than for say SiO_2. On the other hand, with continued demand for higher density, even the high dielectric materials will need to be deposited on some topography to increase the available area in the third dimension. In addition, therefore, to good contacting metallurgy and etch characteristics, conformal deposition is another processing requirement. The high dielectric materials have the potential to extend the future of DRAM generations, but have otherwise no direct impact on the power consumption.

SRAM Cells Comparison

SRAM cache memories have become a necessary function in today's microprocessors. The cell size is of significant importance as the cache size adds to the chip area and thus increases cost and delay. The scalability of SRAM cell performance follows essentially that of the logic technology. Especially for the six transistor cell , which is used in all high performance applications, the cell performance and size directly follows that of the logic devices and of the narrowest interconnect layers respectively. Cell performance is impacted by the current drive of the NMOS and PMOS devices, and cell density by the isolation and interconnect specifications. For the above reasons, the SRAM cell has become a key driving factor for the density specifications of the technology. A denser technology generally implies a lower power technology, thus in that regard a more densely packed cell should be considered as an alternative for very low power applications. On the other hand, as supply voltage scales to values close to 1 Volt, the stability and noise margin (under dynamic as well as static conditions) requires a high internal cell gain. Cells which use high resistance polysilicon resistors, easier to fabricate and denser than the 6 transistor cell, provide therfore insufficient gain at low voltage to allow stable cell operation. Thin Film Transistors as load devices continue to improve and might be a good compromise for those designs where very large cache sizes are needed, but unless the TFT characteristics improve by another order of magnitude[5], the six transistor cell will probably continue to be the preferred choice for integration with other logic technology. And with more integration, more function at lower power is feasible, as argued earlier.

[5] These improvements must be made both in higher on-current and lower off-current

Future Technology Challenges

The performance of CMOS technology is very well established and predicted by the well-known scaling theory. Technologically there are challenges to build FETs with channel lengths below 0.1 μm but the so-called scaling "limits" have been revisited in the last two years with the result that currently accepted values are a gate length of 50 nm for conventional structures and 30 nm for double gate devices on SOI [13,14].

The design challenges for sub 100 nm devices are concentrated on maintaining high transconductance with low supply voltage. Fig. 14 shows an empirical trend of supply and threshold voltage versus effective channel length [15], very close to the calculated values used in the scaling scenario section earlier. It is clear that as the supply voltage is reduced and benefits lower active power, the gap with the threshold voltage is shrinking and off-power is increasing as the off-current increases exponentially at the same time. Since the off-current is an exponential function of structural device properties such as geometry and doping variations, control of threshold voltage and its short channel effects is increasingly difficult and important. Reduced margins between supply and threshold voltage can lead to very high sensitivity of delay. Recent studies [16,17] conclude that 150 mVolt variations in V_t lead to 30% delay degradation at 1.5 Volts as shown in Fig. 15. Although a high V_t is desirable to keep off-current as low as possible, when the ratio of V_t to V_{dd} increases above about 75% the sensitivity of delay (and indirectly thus power) to V_t variations is intolerable. In addition to providing better performance, a threshold voltage that is scaled with supply voltage also has a wider design window. Also shown in Fig. 14 is the gate oxide thickness as a function of supply voltage. It is very important to reduce the effective gate dielectric thickness to maintain the current drive as high as possible. Unfortunately, in the case of SiO_2, when the thickness reaches 3.5 nm or less, direct tunneling current through the oxide becomes very significant [15]. Our understanding of the detailed mechanisms controlling the transport through the oxide is still limited, but it is clear that interface properties are as important as the bulk effects at those dimensions. In any case, with improved growth and/or deposition techniques, the gate dielectric will probably be scaleable to an effective SiO2 thickness of 2.5 - 3.5 nm for 1-1.5 Volt operation.

SOI Technology

Silicon On Insulator (SOI) substrate technology has recently received a great deal of attention because of its high potential for lower power operation [18]. Not only as a result of the reduced device capacitance, but possibly also because more transconductance is available at a given voltage, and because isolation is easier and denser, circuit density and therefore power and speed can be improved. Furthermore, the sub-threshold slope is typically steeper than in bulk devices, so smaller threshold voltage or off-current can be obtained. These features have all been demonstrated (see papers in [18] for an overview of the state-of-the-art of SOI) to a greater or lesser degree, but few if any complex applications have demonstrated that all these advantages can be achieved with a high degree of control and repeatability. The promises of SOI are plenty, and so are its challenges. First the material quality and cost must continue to be improved. Both SIMOX (Separation by Implantation of Oxygen) and BESOI (Bond and Etchback SOI), the two leading substrate manufacturing approaches, have grown very rapidly in the past 5 years, but unless the wafer preparation can be simplified even further, cost alone will limit the use of SOI to specialized applications, choking off any potential economies of scale. Defect densities, oxide pinholes, surface roughness, thickness control from run to run, wafer warpage during processing, wafer sizes to 300 mm, etc., all these qualities have to come close to the manufacturing specifications of bulk-on-epi silicon wafers. These are necessary but certainly not sufficient conditions for widespread use of SOI technology. Device optimization will be different for fully and non-fully depleted devices, each different from bulk CMOS technology. Kink-effects in the drain current must be eliminated or at least precisely controlled. Substrate contacts may need to be provided, so their impact on layout rules must be understood. Series resistance in source and drain junctions/contacts [19] must be limited in devices built on very thin silicon layers (typically fully depleted). New junction and silicidation technology must then be developed. Threshold voltage setting and control appears significantly more difficult than in bulk devices [20]. Circuit models are less complete than in bulk devices because of the additional back-gate device effects possible under a variety of static and dynamic conditions. Isolation is probably achieved in a much simpler way than with bulk technology, but then also causes the layout rules to differ. The bottom line of this (incomplete) summation of issues is that SOI optimization will differ from bulk technology and therefore can not be a simple variant of the bulk process and design. The changes will need to made from material to device to process integration qualification, to circuit model development, design rules, etc.

These issues can all be overcome since they mostly concern engineering and not fundamental "show stoppers". Indeed, the more fundamental advantage of SOI is that it is rooted in and takes full advantage of the silicon technology investment. Moreover, its applications extend to high power, high temperature, and radiation hard systems, and most importantly to mixed signal applications. The noise coupling between digital and analog circuits on the same die is an order of magnitude less than with conventional bulk. And finally, the scalability to smaller dimensions appears from Computer Aided Design studies to go beyond that of bulk CMOS devices [21].

To estimate the expected improvement in speed and power with an SOI technology, the curves of Fig. 7 were recalculated with the following parameter changes. The device current was increased by 20 percent, the parasitic device capacitance and the capacitance of the bottom two layers of the interconnect were reduced by 30 percent, and the density of the gates was improved by 25 percent. Finally, the threshold voltage has been lowered to obtain the same off-current, while the subtreshold slope has been improved from 80 to 60 mV per decade (for the NMOS). Fig. 16 shows and compares the SOI and the Si results. It is important to notice that in today's technology (0.7 um CMOS) most of the leverage is in a potential power reduction for fixed speed. It appears that the same power benefit is not so easily obtained in future generations. A redesign at (even) lower supply voltage is difficult because of the already low threshold voltage. Some improvement is possible with smaller W/L ratio and resulting higher circuit density, but wireability constraints limit any significant benefit, unless extra levels of interconnections are added. But this brings us back to the cost question again. Finally, it should be mentioned that the SOI device structure has been simulated to have better scaling properties than its bulk counterpart.

The most scaleable, and by some considered to be an ultimate form of SOI device, is the double-gate or gate-all-around transistor structure [13]. Because the channel is pinched between two (in principal) identical gates, it demonstrates the best turn-off characteristics possible, namely 60 mV/dec, and eliminates essentially any short channel effects of the threshold voltage, since there is no parasitic path outside the channel through which the electric field can penetrate from drain to source. This structure thus allows the shortest possible channel length, probably less than 50 nm even at 1.5 V and room temperature. Moreover, since the gate is effectively doubled in width (one half over, one half underneath), the transconductance is superior, allowing at least twice as small device widths as single gate devices[6]. Finally, the channel in this structure is undoped, the V_t is set by the work-functions of the gate, and doping fluctuations as reported in [22] are avoided. Obviously, the double gate structure is not easy to construct when one wishes to preserve the flexibility of various gate lengths and widths, as well as planarity. Even for conventional CMOS the technological challenges on the device structure such as gate stack lithography and etching, while maintaining low sheet resistance, ultra-thin gate dielectric growth, shallow junction formation with controlled lateral profiles, and channel profile optimization present enormous challenges [15]. However, if the highest possible performance is not needed, parasitic resistance deserves less emphasis, and density and simplicity are more effective to extend the usability of conventional CMOS devices and circuits. Again SOI technology provides a potentially simpler fabrication process.

Interconnection Technology

Interconnection technology has mainly received negative attention because of the limited scalability in the case of ever-larger growing chips. Integrating more functions onto a larger chip cannot be accomplished with the same interconnection wires that were scaled down to

---

[6] The transconductance improves more than 2x because the whole channel volume inverts, little vertical electric field acts on the carriers, leading to superior carrier density and mobility.

accommodate the higher device density. A hierarchy of levels should be used to allow for effective wireability at the densest levels, and acceptable performance at the uppermost levels [23]. Wireability at the lowest levels is important to take full advantage of the reduced cell pitch for lower power designs, but capacitance reduction is even more key. Note that this can be achieved not only with lower dielectric materials, but also by smaller cross-sectional wires if the resistivity of the conductor is improved and the electromigration resistance is higher. Since the electromigration resistance of Cu is an order of magnitude higher than for Al, this might become the most important driving factor to move from Al to Cu based alloys [23]. Lower power as well as higher performance will be facilitated by progressing from oxide to lower dielectric constant, such as organic or cellulose-structure, interdielectric materials. The most compromising feature of a Cu-based metallurgy is the need for a barrier metal between the Cu and the insulator. In addition to the process complexity/cost issue, the reduction in cross-sectional area is significant. It can be easily shown that based upon the bulk material resistivities of Cu and Al, the liner thickness in the Cu line cannot be more than 10 percent of the line width (and height assuming a square line) to completely annihilate the resistance advantage of Cu over Al. In other words, for the 0.18 μm generation, with 0.25 μm wide interconnection lines and spaces, an effective, conformal barrier layer technology of 25 nm thickness has to be developed to make such small Cu interconnects have in fact a lower sheet resistance than non-barrier-cladded Al lines of the same dimension.

Discussion

The leverage of technology in the evolution of semiconductor based products has been and is expected to continue to be significant. As a result of the changing markets toward more mobile, small form-factor, appliance type systems, power and cost-performance will be more important than ever in the development of the next generations (0.25 and 0.18 μm) of Si CMOS technology. Designing devices and interconnections for low voltage and low capacitance should be balanced with the desire for higher speed. Higher integration, more functionality is what drives these applications. Conventional bulk CMOS may be supplemented with SOI technology to provide solutions for critical low power or mixed signal applications. Both approaches appear to be scaleable to at least 100-50 nm effective gate length, at least one generation beyond the 193 nm optical lithography capability.

The demand for more throughput will continue to drive increases in circuit density, and faster execution per function. Increasingly important will be efficiency and functionality. Digital and analog, logic and memory functions tightly integrated to reduce power and cost. These factors present enormous challenges for the technology designers. Although the density capability exists already at the 0.25 μm generation to implement all the logic and memory requirements for a typical speech recognition system, combining both the analog and digital functions presents another challenge. Analog and digital technologies are diverging with respect to voltage optimization. Analog functions are not easily compatible with low supply voltage, low current, and short CMOS gate lengths [24].

The success and utilization of Ultra Large Scale Integration will depend on our ability to match logic, memory and analog functionality in a cost effective way. It will not only enhance existing applications, but holds tremendous promise for novel applications.

Acknowledgements

Without the technical contributions of J. DeBrosse, R. Dennard, D. Frank, J-Y. Mii, H. Shin, J-Y. Sun, Y. Taur, T. Theis and H-S. Wong of the IBM T.J. Watson Research Center, this work would not have been possible. The author also acknowledges the discussions with J. Barnes, K. Cham, Y. Nishi and P. Raje of the HP Laboratories.

References

[1] Private communication: T.N. Theis, J.M.C. Stork, P. Vernes, R.H. Dennard, T.H. Ning, G.L. Chiu, and E.P. Harris, "Low-power technology development and insertion into dual-use applications," prepared for Advanced Research and Projects Agency, July 1993.

[2] P.K. Chatterjee and G.B. Larrabee, "Gigabit age microelectronics and their manufacture," IEEE Trans. on Very Large Scale Integration (VLSI) Systems, Vol. 1, No. 1, pp 7-21, March 1993.

[3] Semiconductor Industry Association Workshop Conclusions, issued by the SIA, San Jose, CA, (408) 246-2711.

[4] D.K. Liu and C. Svensson, "Power Consumption Estimation in CMOS VLSI Chips," IEEE Journal of Solid-State Circuits, Vol. 29, No. 6, pp 663-670, June 1994.

[5] R.B. Merril, W.M. Young and K. Brehmer, "Effect of Substrate Material on Crosstalk in Mixed Analog/Digital Circuits," in IEDM digest of Tech Papers, pp 433-436, 1994 IEDM.

[6] H.B. Bakoglu, Circuits Interconnection and Packaging for VLSI, Addison Wesley, 1990.

[7] J.B. Burr, "Ultra Low Power CMOS Technology," Stanford University Report, 1991.

[8] The Feynman Lectures on Physics, Vol. 2, Chapter 19: "The Principle of Least Action," Addison-Wesley Publishing Co., 1964

[9] Y. Mii, S. Wind, Y. Taur, Y. Lii, D. Klaus, and J. Bucchignano, "An ultra-low power 0.1μm CMOS," 1994 Symposium on VLSI Technology, Hawaii, pp 9-10, June 1994.

[10] M. Takada, "Low Power Memory Design," 1993 IEDM Short Course Program.

[11] C.A. Warwick and A. Ourmazd, "Trends and limits in monolithic integration by increasing the die area," IEEE Transactions on Semiconductor Manufacturing, Vol. 6, No. 3, pp 284-289, Aug 93.

[12] K. Shibahara, H. Mori, S. Ohnishi, R. Oikawa, Y. Kojima, H. Yamashita, K. Itoh, S. Kamiyama, H. Watanabe, T. Hamada and K. Koyama, "1 G DRAM Cell with Diagonal Bit-line (DBL) Configuration and Edge Operation (EOS) FET," in 1994 IEDM Digest of Tech Papers, pp 639-642, 1994 IEDM.

[13] D.J. Frank, S.E. Laux, and M.V. Fischetti, "Monte-carlo simulation of a 30 nm dual-gate MOSFET: how short can Si go?," in IEDM Digest of Tech Papers, pp 553-558, 1992.

[14] C.A. Mead, "Scaling of MOS technology to submicrometer feature sizes," to appear in The Journal of VLSI Signal Processing.

[15] Y. Taur, Y. Mii, D. Frank, H-S. Wong, D. Buchanan, G. Sai-Halasz, S. Wind, S. Rishton, and E. Nowak, "0.1μm CMOS and beyond," to be published in IBM Journal of Research and Development.

[16] T. Kobayashi and T. Sakurai, "Self-adjusting threshold-voltage scheme (SATS) for low-voltage high-speed operation," IEEE 1994 Custom Integrated Circuits Conference, pp.271.

[17] S.W. Sun and P.G.Y. Tsui, "Limitation of CMOS supply-voltage scaling by MOSFET threshold-voltage variation," IEEE 1994 Custom Integrated Circuits Conference, pp.267.

[18] 1994 IEEE International SOI Conference Proceedings, IEEE Cat No 94CH 35722, Oct. 1994.

[19] L.T. Su, M.J. Sherony, H. Hu, J.E. Chung and D.A. Antoniadis, "Optimization of Series Resistance in sub-0.2 μm SOI MOSFET's," IEEE Electron Device Letters, Vol.15, No.5, pp. 145-147, May 1994.

[20] L.T. Su, J.B. Jacobs, J.E. Chung, and D.A. Antoniadis, "Deep-Submicrometer Channel Design in Silicon-On-Insulator (SOI) MOSFET's," IEEE Electorn Device Letters, Vol.15, No.5, pp.183-185, May 1994.

[21] C. Fiegna, H. Iwai, T. Wada, M. Saito, E. Sangiorgi, and B. Ricco, "Scaling the MOS Transistor Below 0.1 μm: Methodology, Device Structures, and Technology Requirements," IEEE Transactions on Electron Devices, Vol. 41, No. 6, pp. 941-951, June 1994

[22] H.S. Wong and Y. Taur, "Three-dimensional 'atomistic' simulation of discrete random dopant distribution effects in sub-0.1μm MOSFET's," IEDM Digest of Tech Papers, pp 705-708, 1993 IEDM.

[23] Harper et al., "Materials Issues in Copper Interconnections," MRS Bulletin, p.23, Aug 1994.

[24] A. Matsuzawa, "Low voltage mixed analog/digital circuit design for portable equipment," 1993 Symp. on VLSI Circuits, Digest of Tech Papers, pp. 49-54.

Figure Captions

Fig. 1

Both power/MIPS and price/MIPS have decreased by orders of magnitude over the past two decades (adapted from [2]).

Fig. 2

Microprocessor chip area versus year as reported at ISSCC conferences. The trend is compared with the area required to implement a 64 bit processor, and with an extrapolation of the mainframe board area.

Fig. 3

A general speech recognition system is implemented with a general purpose micro-processor, a specialized DSP chip, SRAM as well as DRAM and EEPROM for memory, and also A/D and D/A functions. Scaled technology has the capability to integrate the whole system on one or two chips, if the various technology elements can be merged economically.

Fig. 4

Recent data reported on power versus frequency for microprocessors. Only a slight shift is observed due to the transition from 5 to 3.3 Volt supply voltage.

Fig. 5

Technology scaling of a constant function (microprocessor) provides an order of magnitude improvement in power*delay product, whih can be utilized for increased speed, as well as very low power.

Fig. 6a,b

Calculations of clock frequency and power contours versus gate length and supply voltage:

a) Frequency optimum lies at higherst practical voltages for the smallest dimensions

b) Power increases rapidly with voltage and reaches impractical values at the high voltage, small feature size designs

Fig. 7

Clock frequency versus power for various supply voltages and gate length dimensions. Maximum performance stays at about the same power per chip.

Fig. 8a,b

Clock frequency a) and action b) versus power for various W/L ratio's of the devices. The gate length and supply voltage are scaled proportionally. A least action design point is observed for a value of W/L of approximately 10.

Fig. 9a,b

Design curves of fixed active and standby power, and speed versus threshold and supply voltage:

a) The arrows indicated the desired scaling direction. The design space enclosed by the curves is diminishing.

b) The curves show the threshold and supply voltage combination for which constant performance as well as a factor two increase in speed can be obtained with scaling from 0.35 to 0.175 μm CMOS generations.

Fig. 10

Active power trends for DRAM and SRAM chips. Factors of 10 in power-per-bit have been achieved per generation for the DRAM. The power need for SRAM has stayed below DRAM until this time but is driven up by high speed requirements.

Fig. 11

DRAM cell area versus lithography generation. The extrapolated trends and requirement lines cross over at the 1 Gbit generation, indicating vanishing design margins. Reducing cell area requirements to less than 8 lithographic squares provides for continued scaling of the conventional transistor+capacitor cell.

Fig. 12

Bitline voltage appearing to the sense amplifier decreases with reduced supply/plate voltage. The bitline capacitance needs to be scaled faster than the storage capacitor (adapted from [10]).

Fig. 13

Required dielectric resistivity and permittivity to determine the feasibility of new insulators for DRAM cell applications. Few examples meet these (minimum) criteria.

Fig. 14

The ratio between power supply and threshold voltage is decreasing for sub 0.25 µm devices. Gate oxide thickness is approaching fundamental tunneling current limits below 3 nm [14].

Fig. 15

Delay variations due to threshold voltage shift, assuming a constant worst-case V_t of 0.4 Volt. The delay sensitivity to becomes intolerable when Vt increases to more than 75% of Vdd (adapted from [10]).

Fig. 16

Clock frequency versus power comparing Si and SOI technology for 3 generations. Higher speed at constant power or lower power at constant speed may be realized.

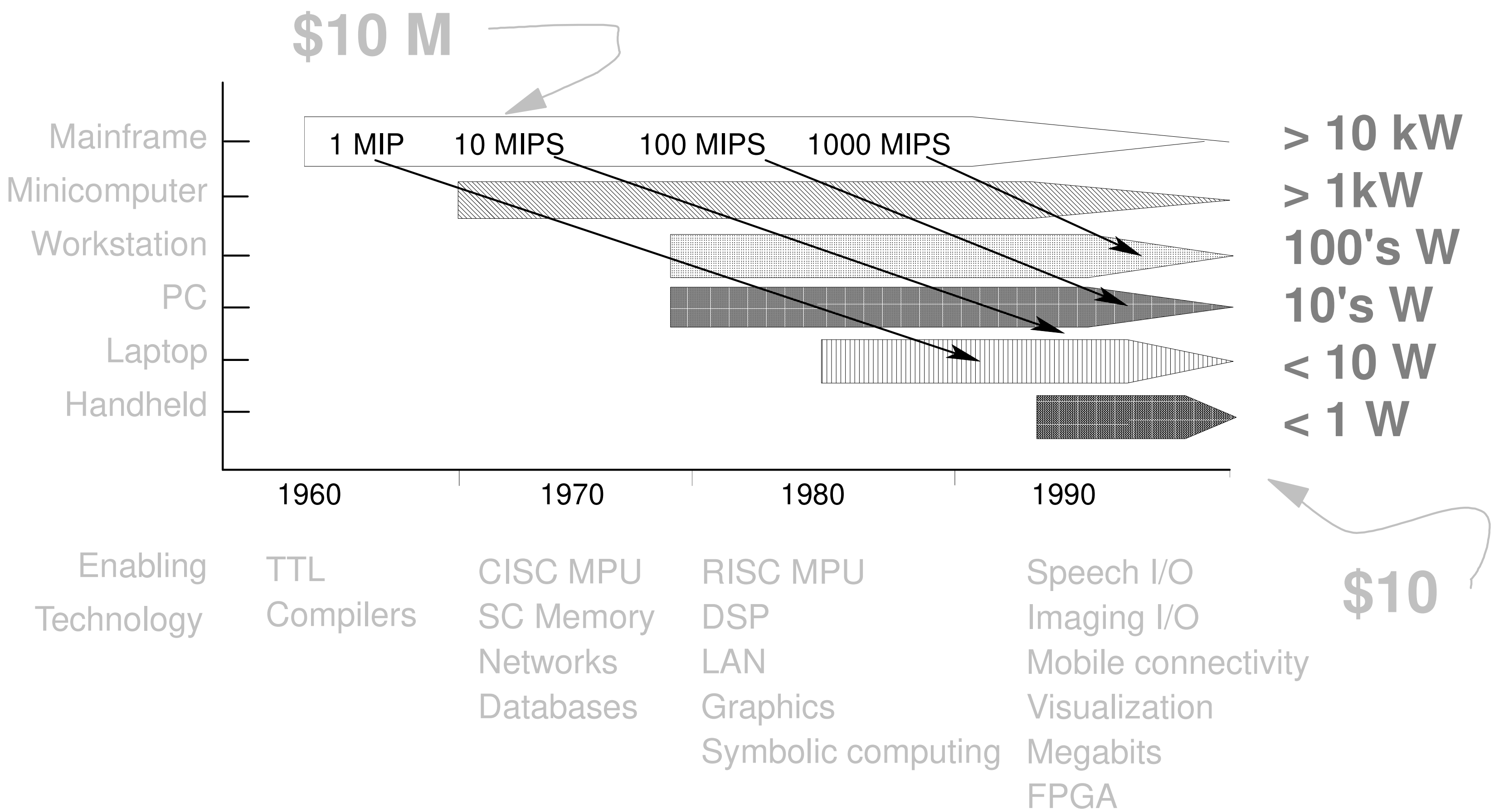
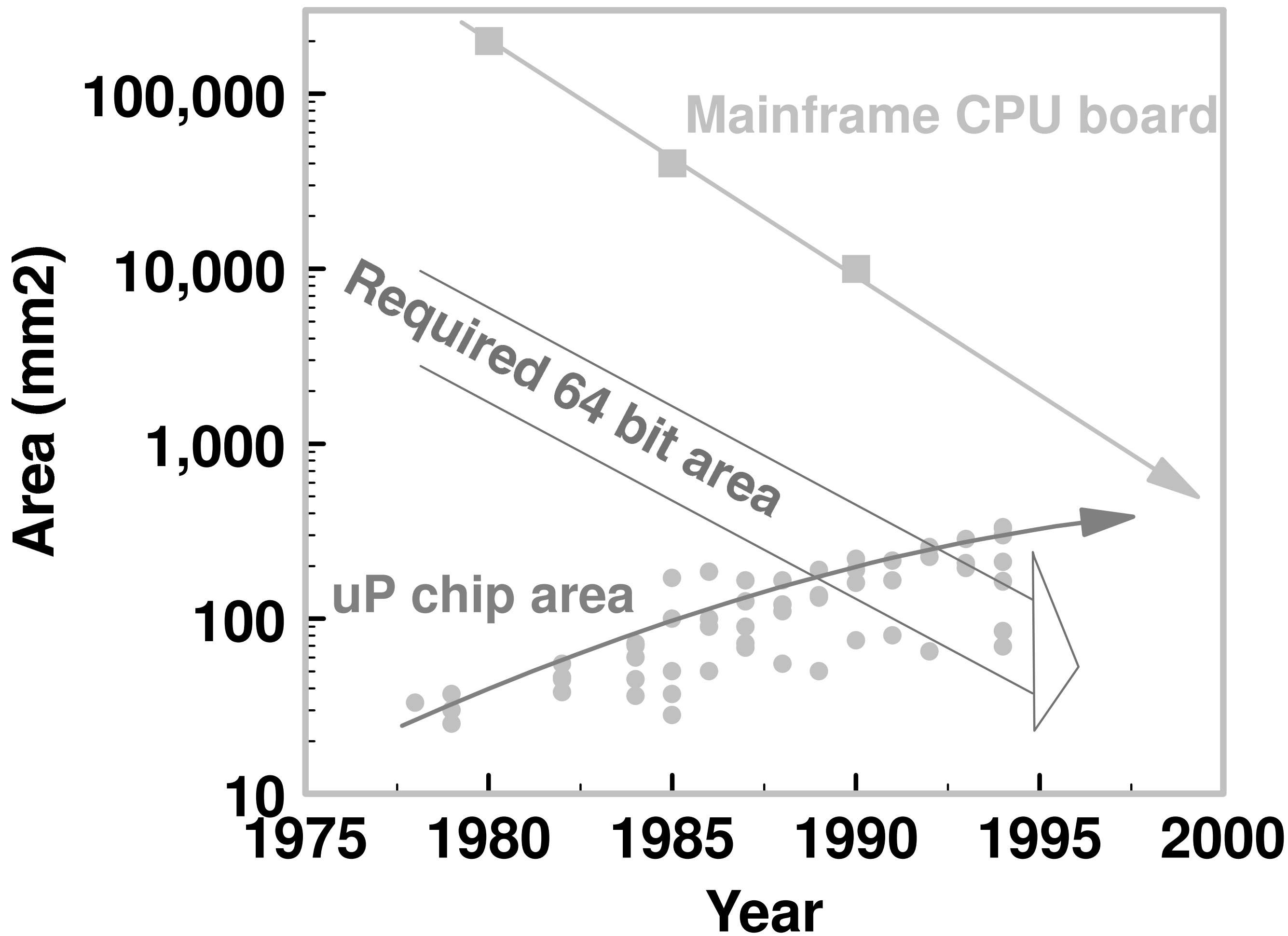

$10 M
Mainframe
Minicomputer
Workstation
PC
Laptop
Handheld
1 MIP
10 MIPS
100 MIPS
1000 MIPS
> 10 kW
> 1kW
100's W
10's W
< 10 W
< 1 W
1960
1970
1980
1990
$10
Enabling
Technology
TTL
Compilers
CISC MPU
SC Memory
Networks
Databases
RISC MPU
DSP
LAN
Graphics
Symbolic computing
Speech I/O
Imaging I/O
Mobile connectivity
Visualization
Megabits
FPGA

Fig 1



Mainframe CPU board
Required 64 bit area
uP chip area
100,000
10,000
1,000
100
10
1975
1980
1985
1990
1995
2000
Area (mm2)
Year

Fig 2

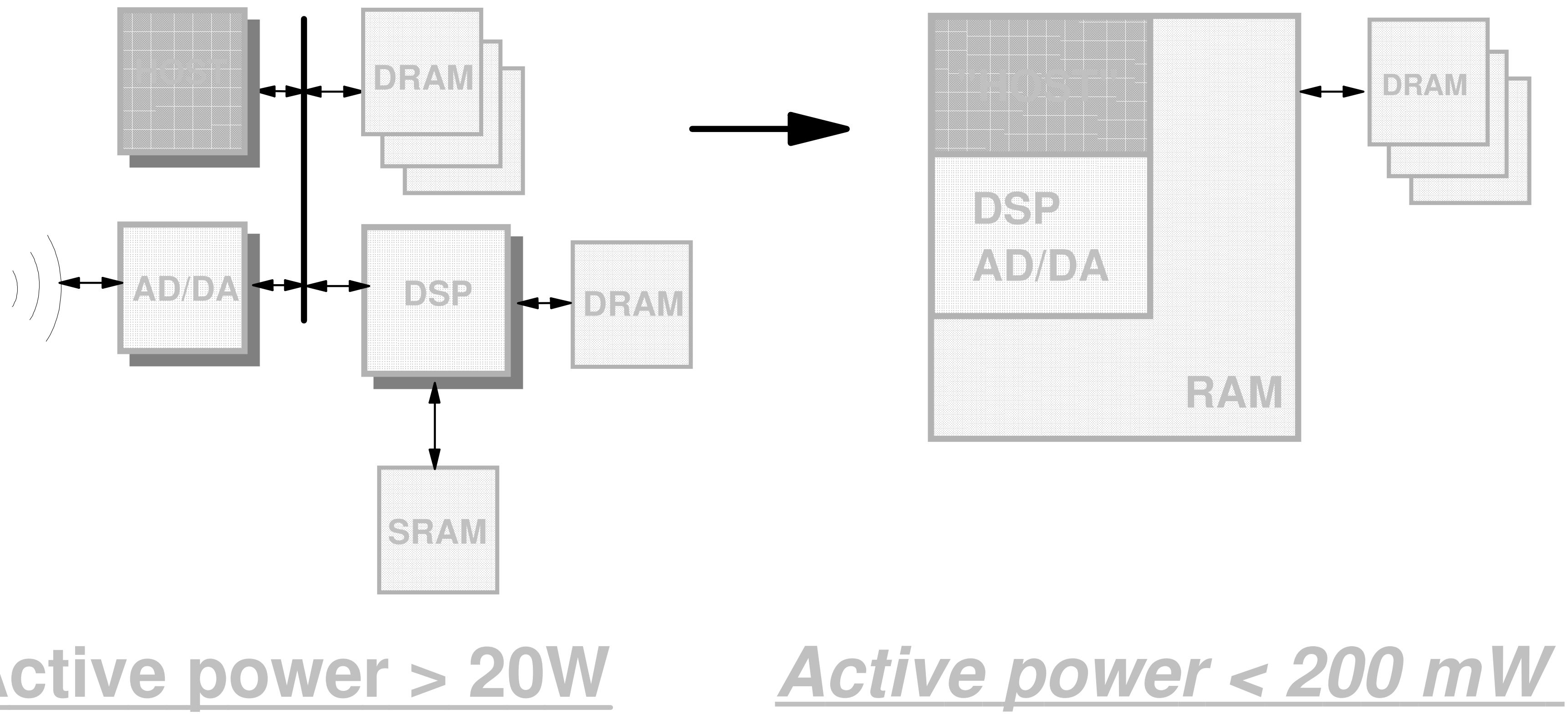

HOST
DRAM
AD/DA
DSP
DRAM
SRAM
HOST
DSP
AD/DA
RAM
DRAM
Active power > 20W
Active power < 200 mW

Fig 3

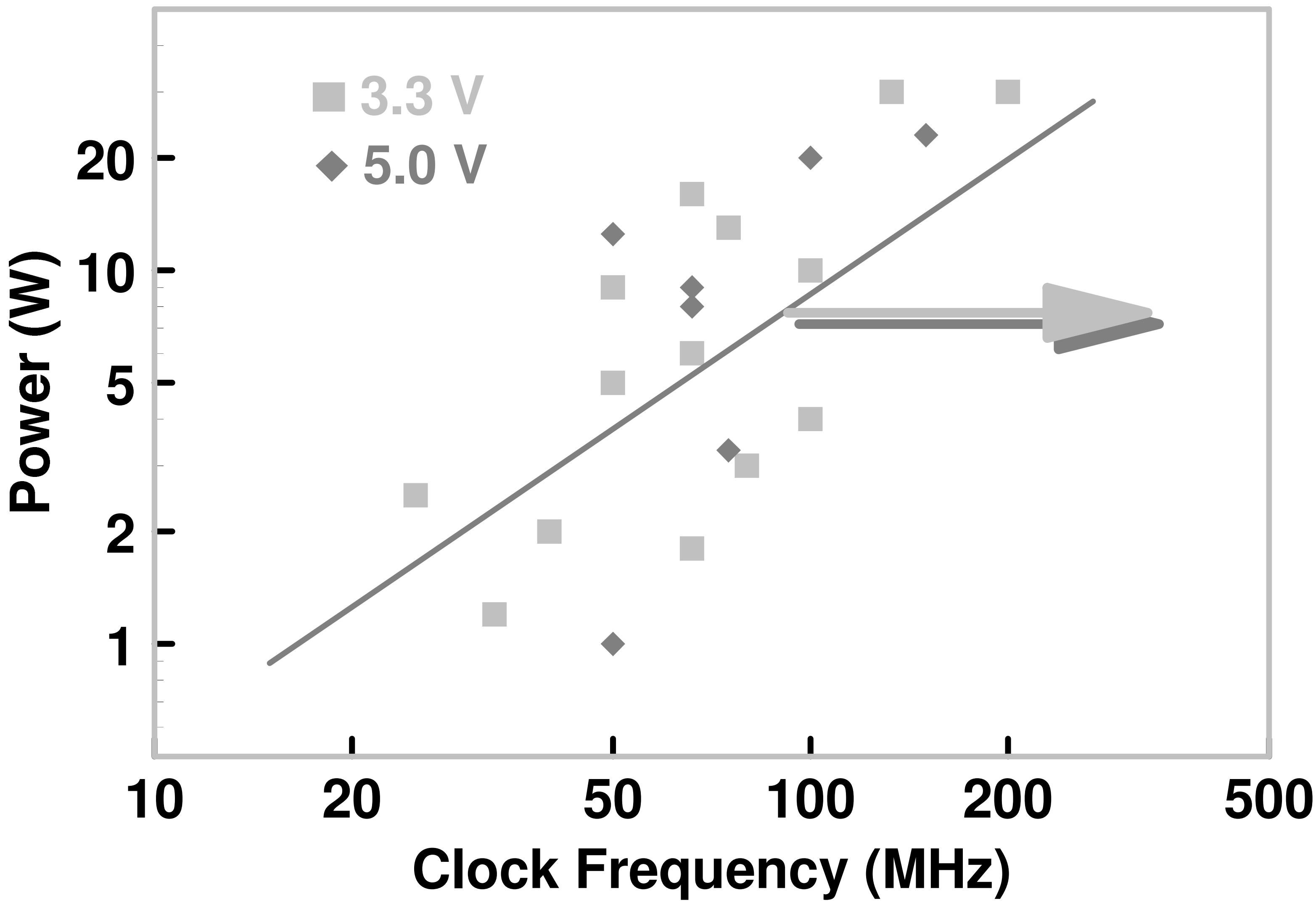

3.3 V
5.0 V
20
10
5
2
1
Power (W)
10
20
50
100
200
500
Clock Frequency (MHz)

Fig 4

Pitch	1.4	1.0	0.70	0.50	um
Lgate	**0.50**	**0.35**	**0.25**	**0.175**	**um**
Leff	**0.35**	**0.25**	**0.175**	**0.125**	**um**
Vdd	**3.3**	**2.5**	**1.75**	**1.25**	**Volt**
Clock	**100**	**145**	**190**	**230**	**MHz**
Area	**8x10**	**5.6x7**	**4x5**	**2.5x3**	**mm2**
Power	**5.0**	**3.2**	**1.9**	**1.0**	**Watt**
@constant 100 MHz:					
Area	**8x10**	**4.5x6**	**3.2x4.2**	**2x2.5**	**mm2**
Power	**5.0**	**2.2**	**1.0**	**0.45**	**Watt**
Energy	**50**	**22**	**10**	**4.5**	**mW/MHz**

Fig 5

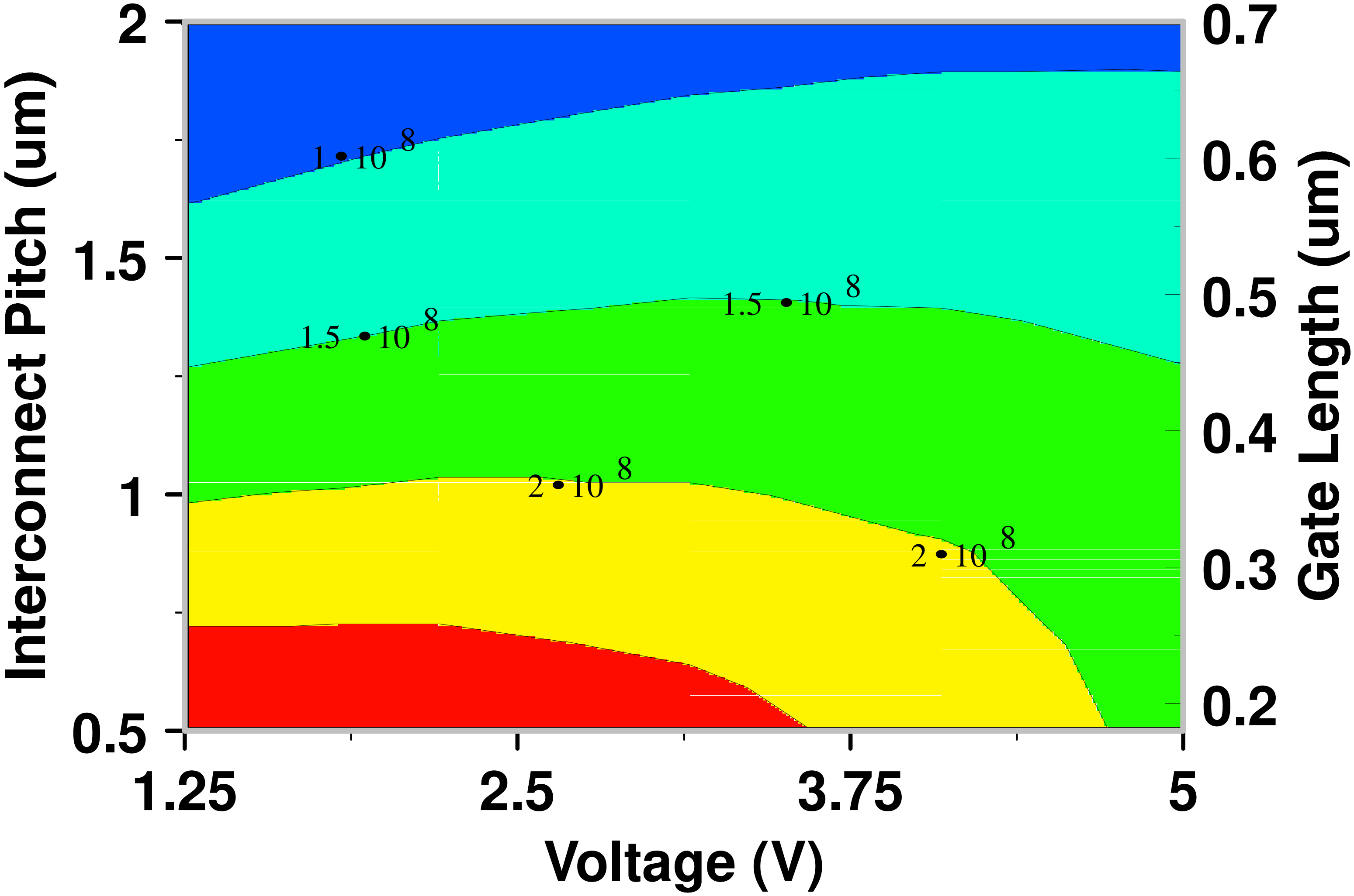
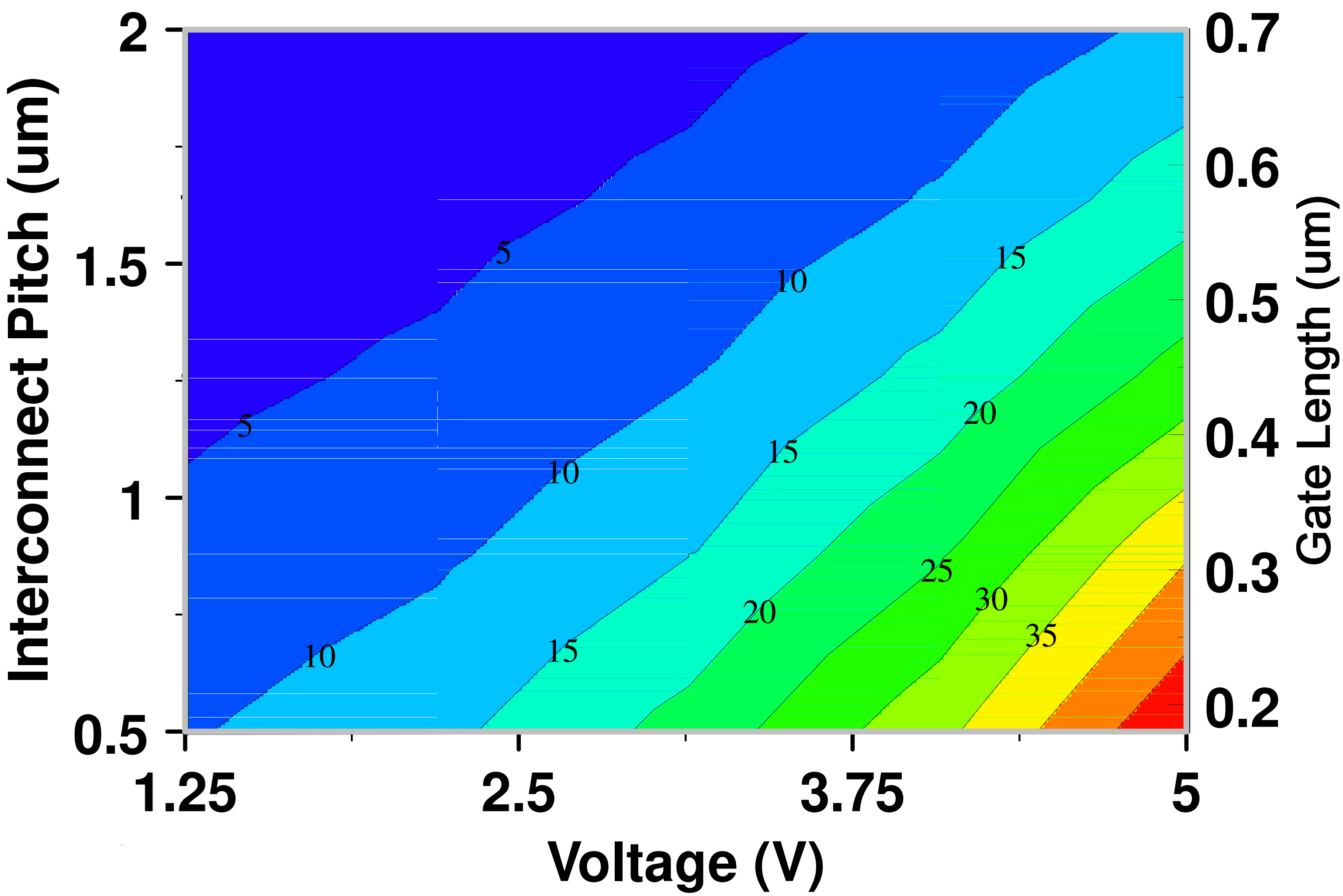
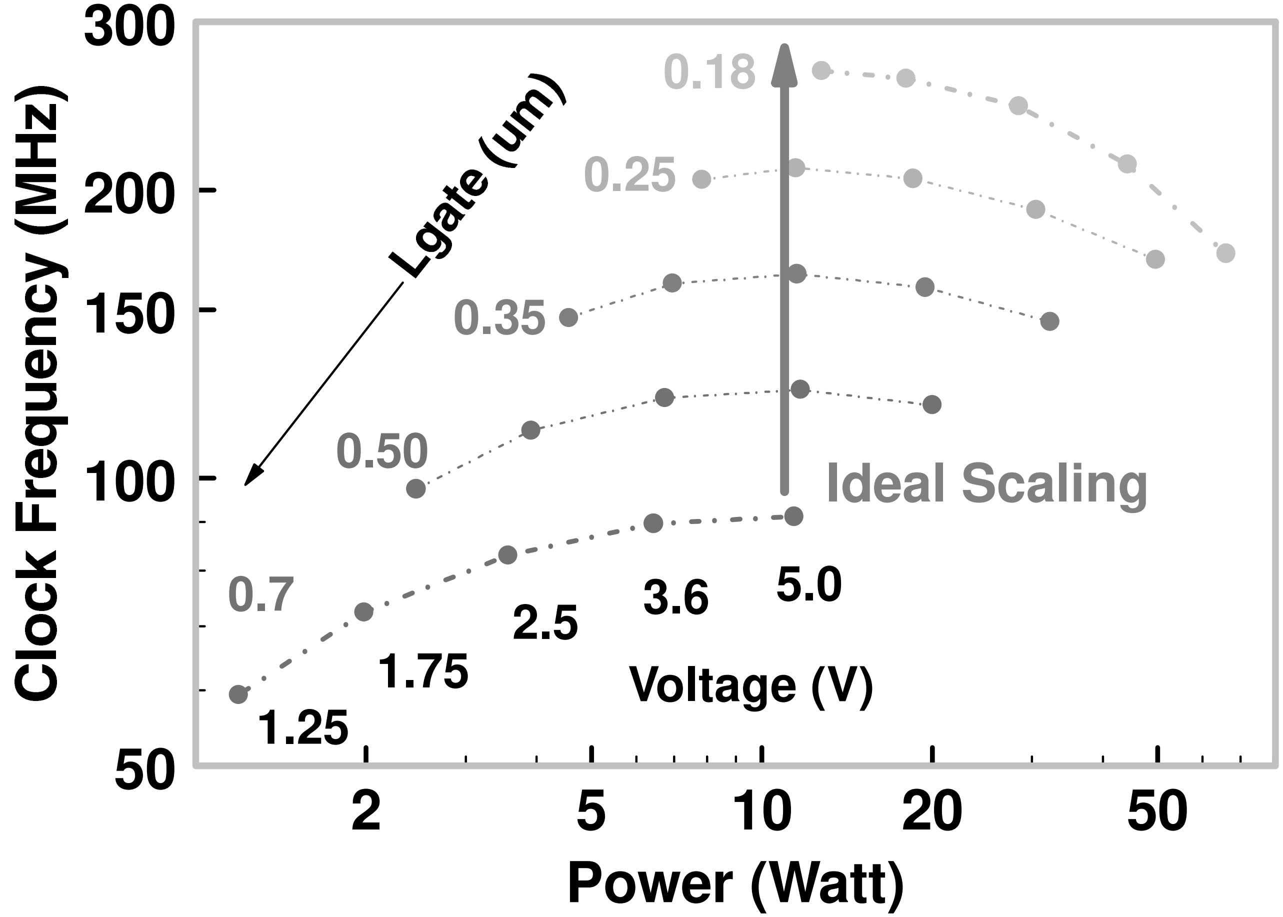

2
1.5
1
0.5
Interconnect Pitch (um)
0.7
0.6
0.5
0.4
0.3
0.2
Gate Length (um)
1.25
2.5
3.75
5
Voltage (V)
1•10 8
1.5•10 8
1.5•10 8
2•10 8
2•10 8

Fig 6a



Interconnect Pitch (um)
2
1.5
1
0.5
1.25
2.5
3.75
5
Voltage (V)
0.7
0.6
0.5
0.4
0.3
0.2
Gate Length (um)
5
10
15
20
25
30
35

Fig 6b



Clock Frequency (MHz)
300
200
150
100
50
Power (Watt)
2
5
10
20
50
Lgate (um)
0.18
0.25
0.35
0.50
0.7
Ideal Scaling
1.25
1.75
2.5
3.6
5.0
Voltage (V)

Fig 7

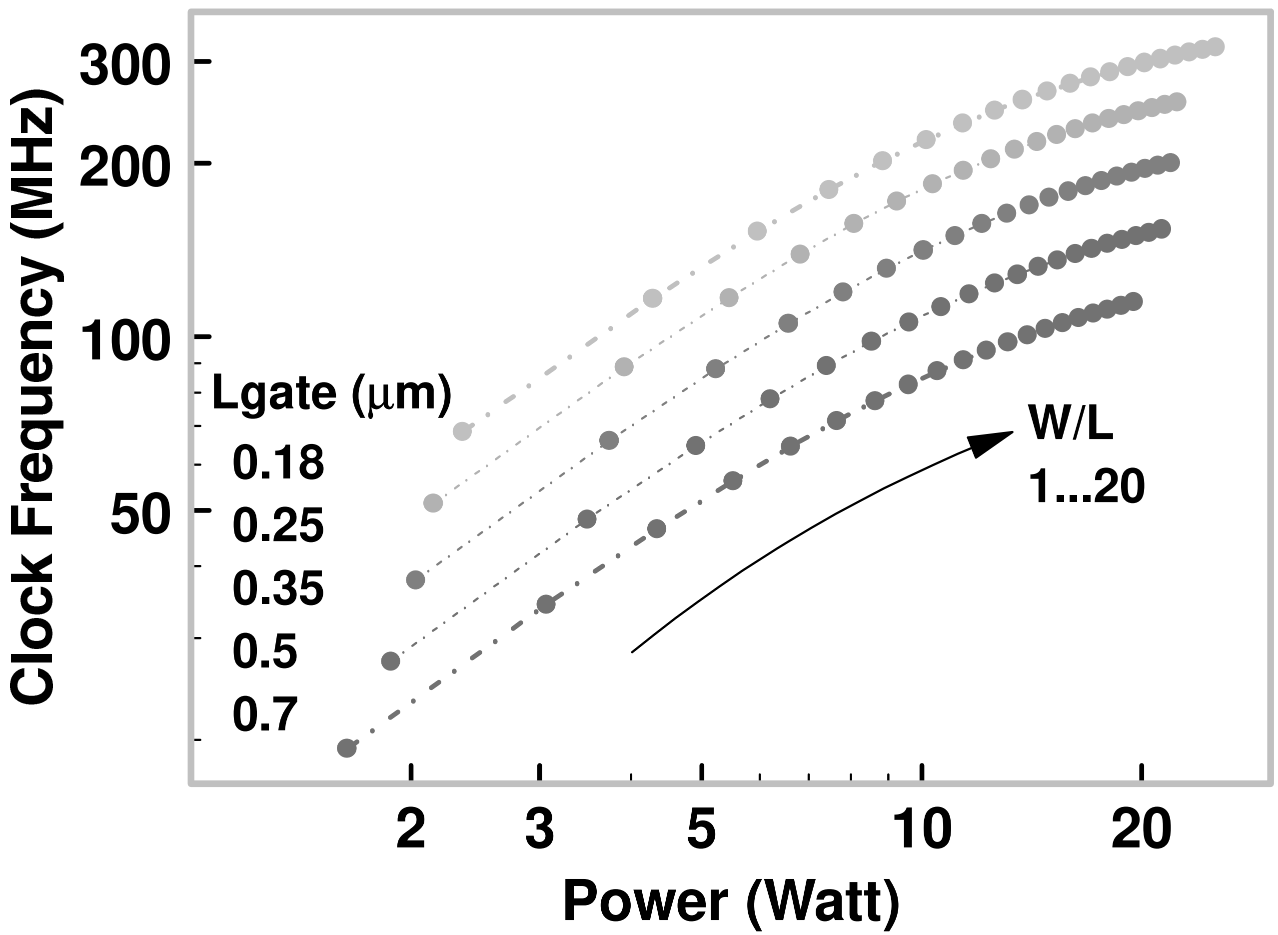
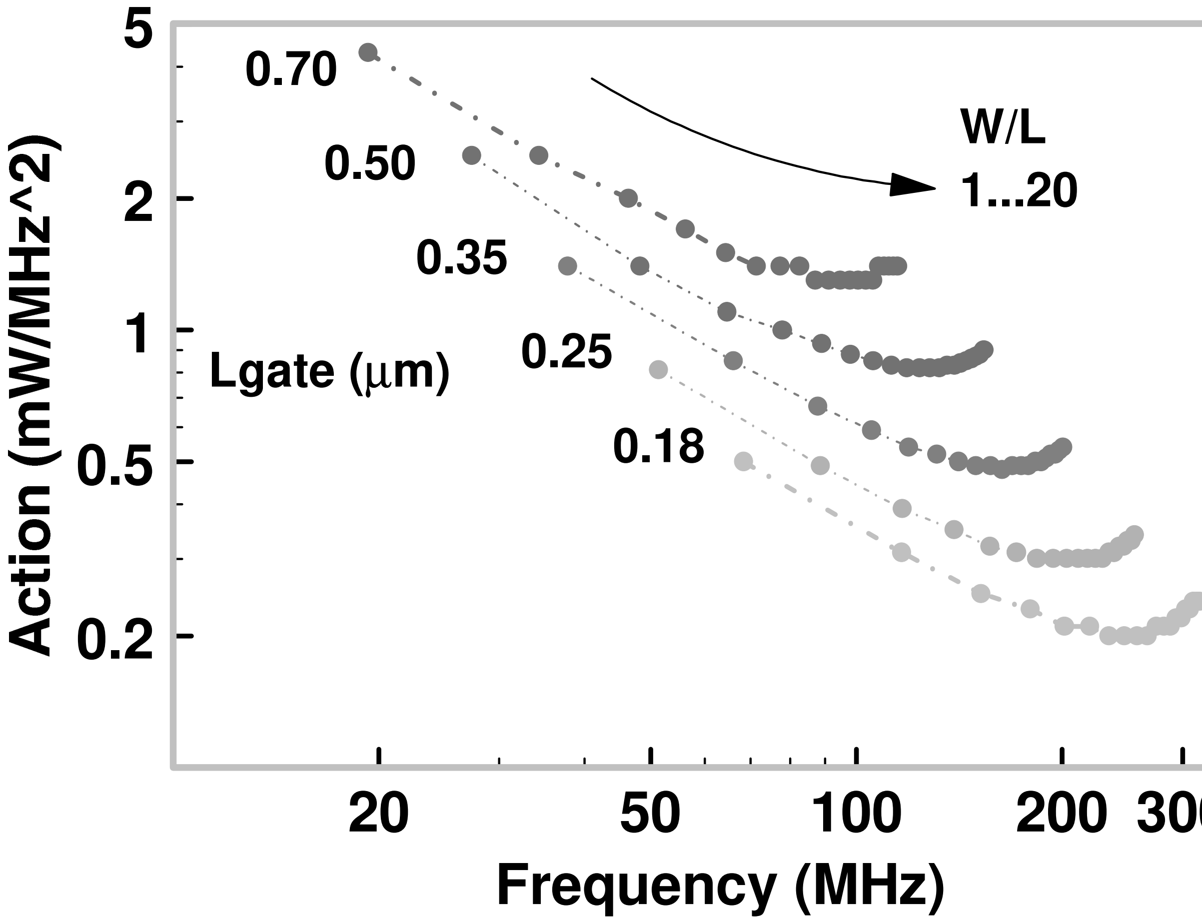

Clock Frequency (MHz)
300
200
100
50
Lgate (μm)
0.18
0.25
0.35
0.5
0.7
W/L
1...20
2
3
5
10
20
Power (Watt)

Fig 8a



5
2
1
0.5
0.2
Action (mW/MHz^2)
0.70
0.50
0.35
0.25
0.18
Lgate (μm)
W/L
1...20
20
50
100
200
300
Frequency (MHz)

Fig 8b

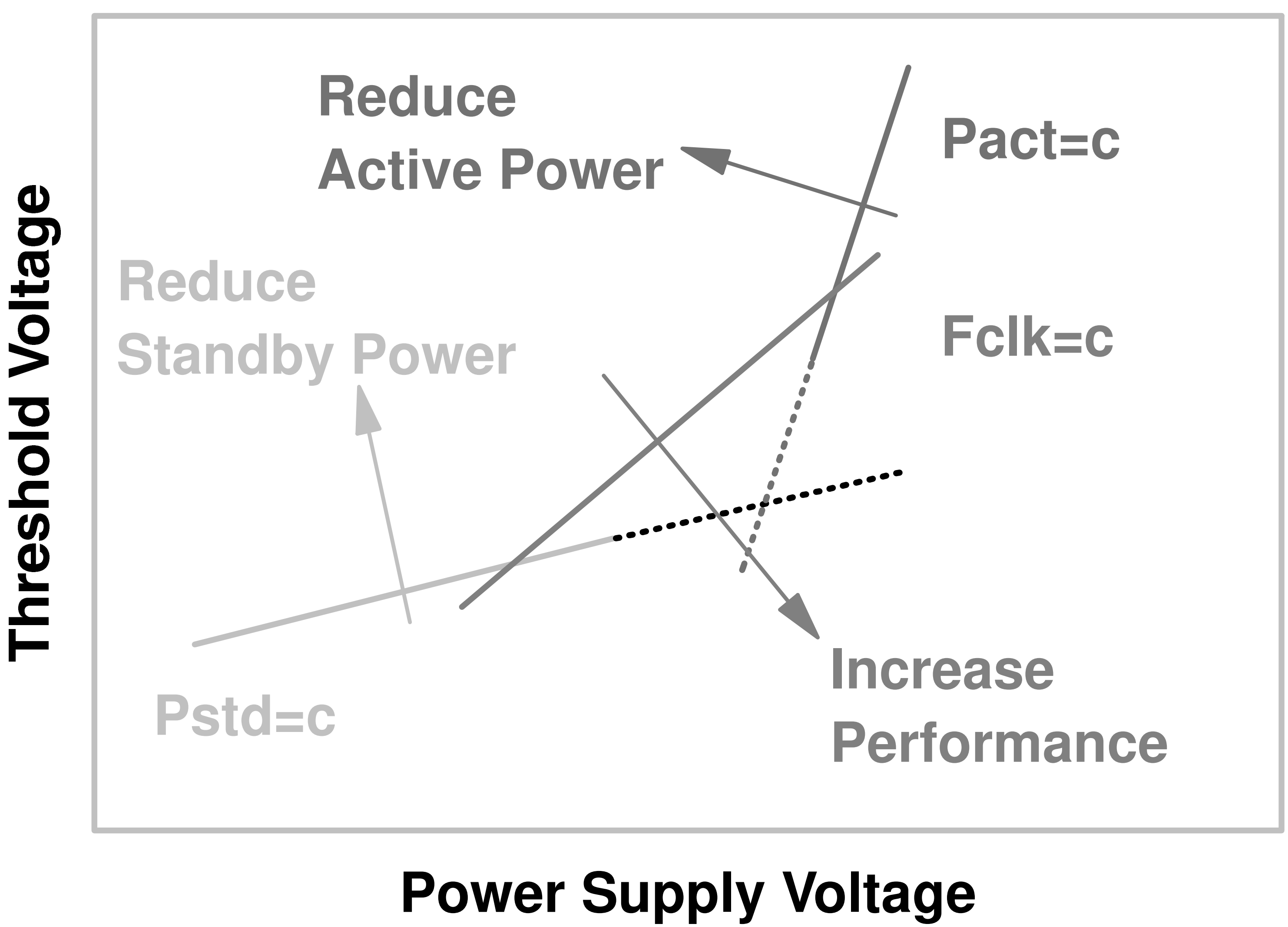
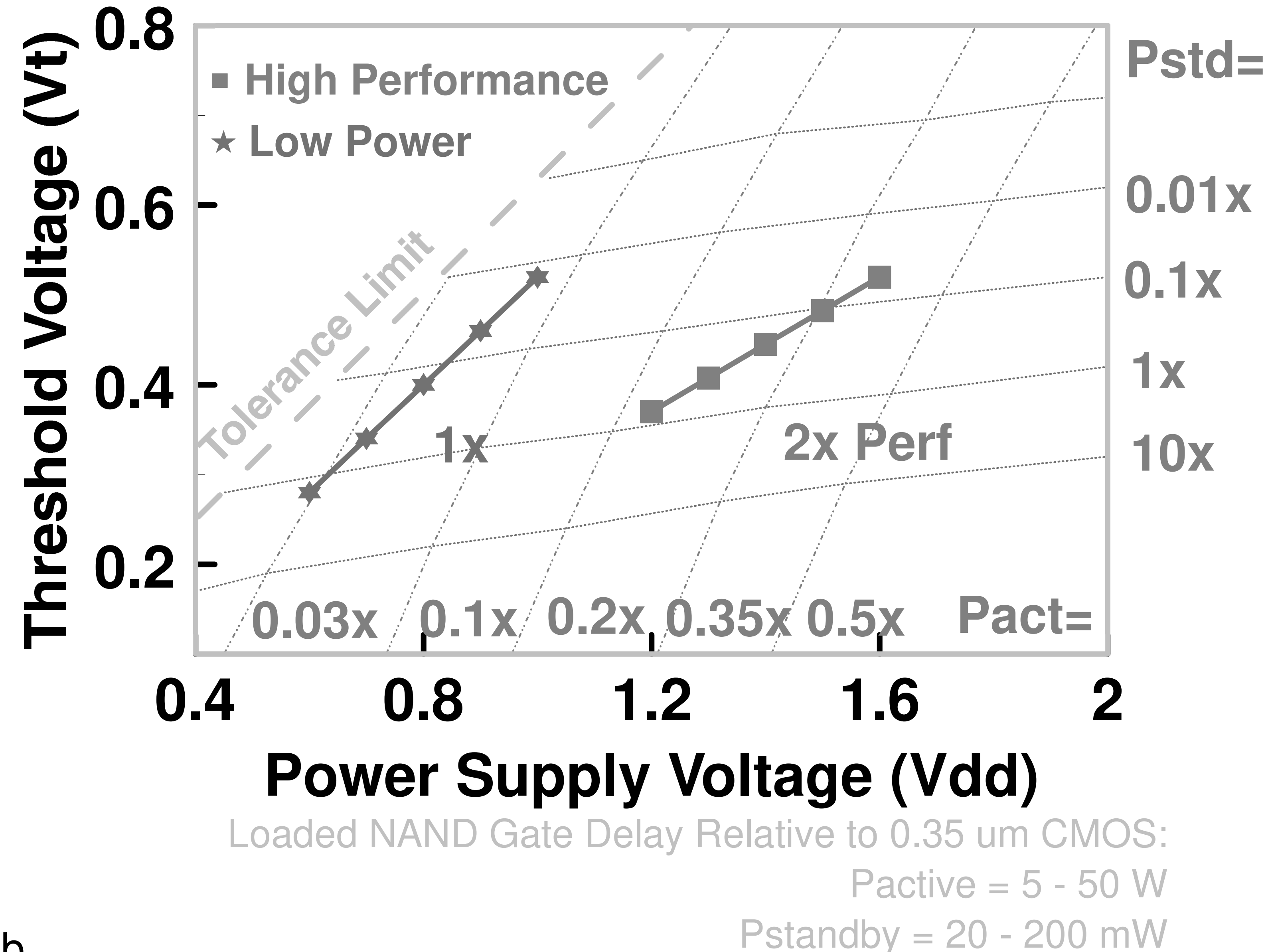

Reduce
Active Power
Pact=c
Reduce
Standby Power
Fclk=c
Pstd=c
Increase
Performance
Threshold Voltage
Power Supply Voltage

Fig 9a



Threshold Voltage (Vt)
0.8
0.6
0.4
0.2
High Performance
Low Power
Tolerance Limit
Pstd=
0.01x
0.1x
1x
10x
1x
2x Perf
0.03x
0.1x
0.2x
0.35x
0.5x
Pact=
0.4
0.8
1.2
1.6
2
Power Supply Voltage (Vdd)
Loaded NAND Gate Delay Relative to 0.35 um CMOS:
Pactive = 5 - 50 W
Pstandby = 20 - 200 mW

Fig 9b

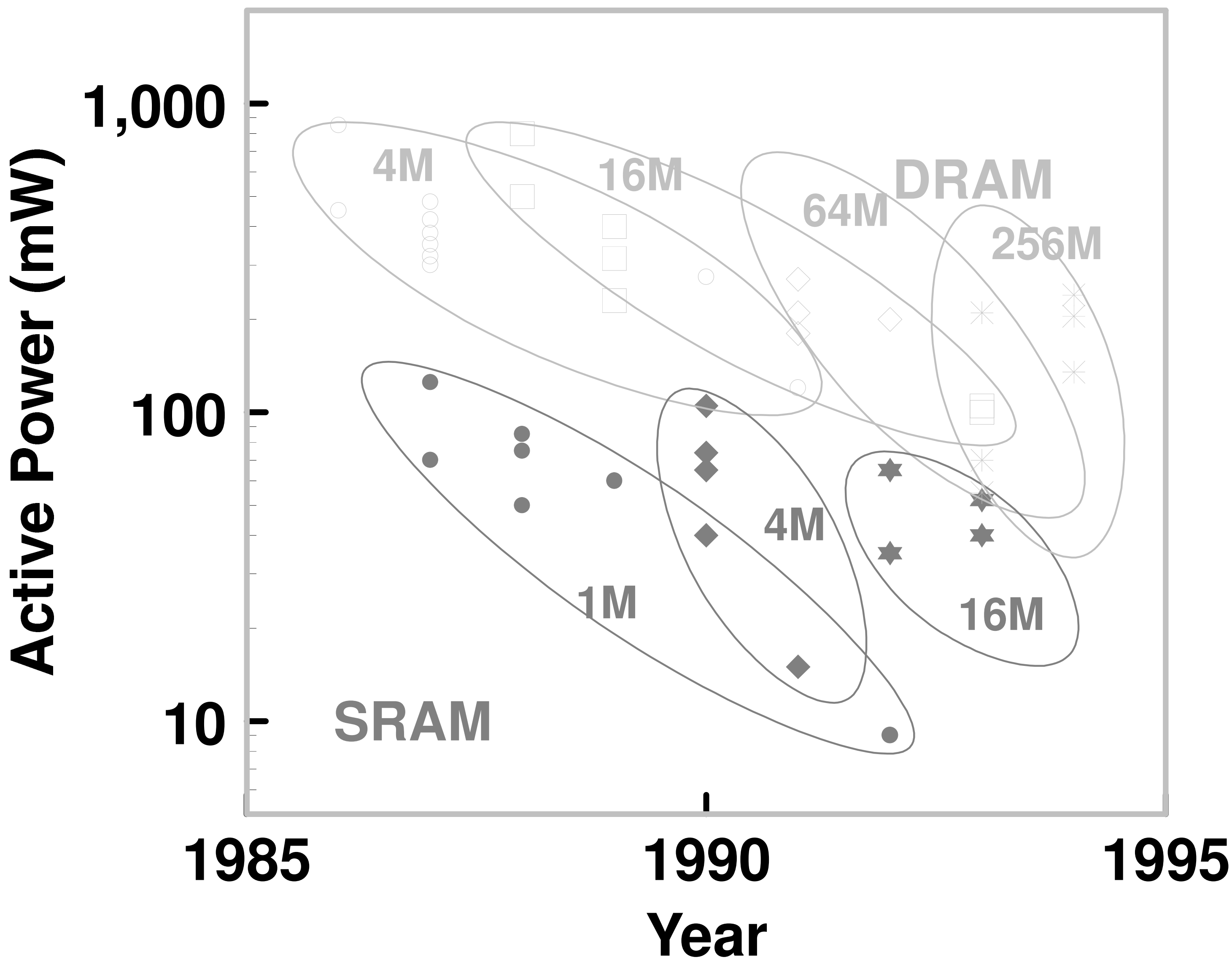


Fig 10

Adapted from [2]

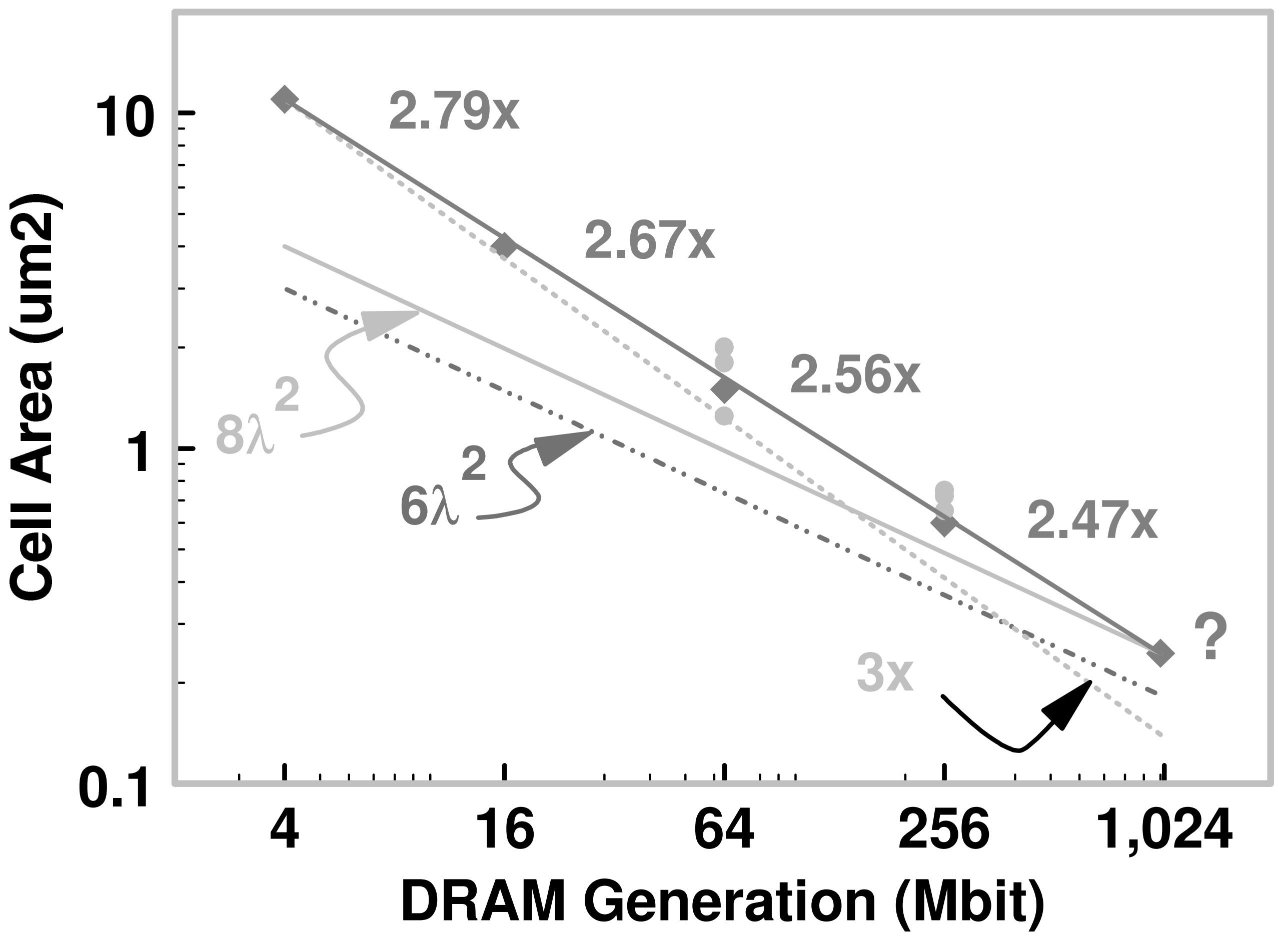
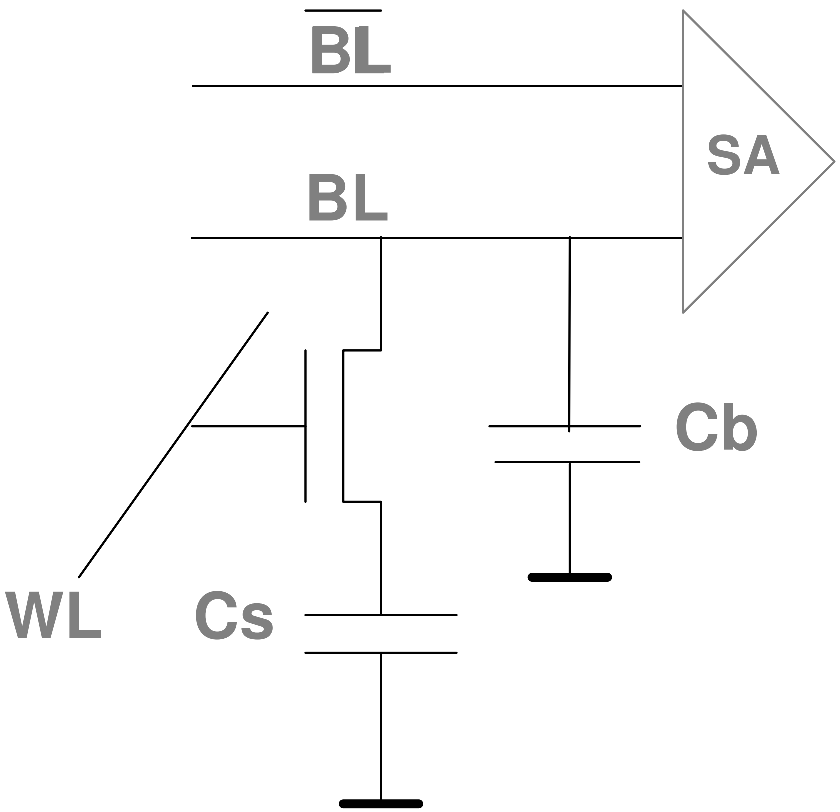
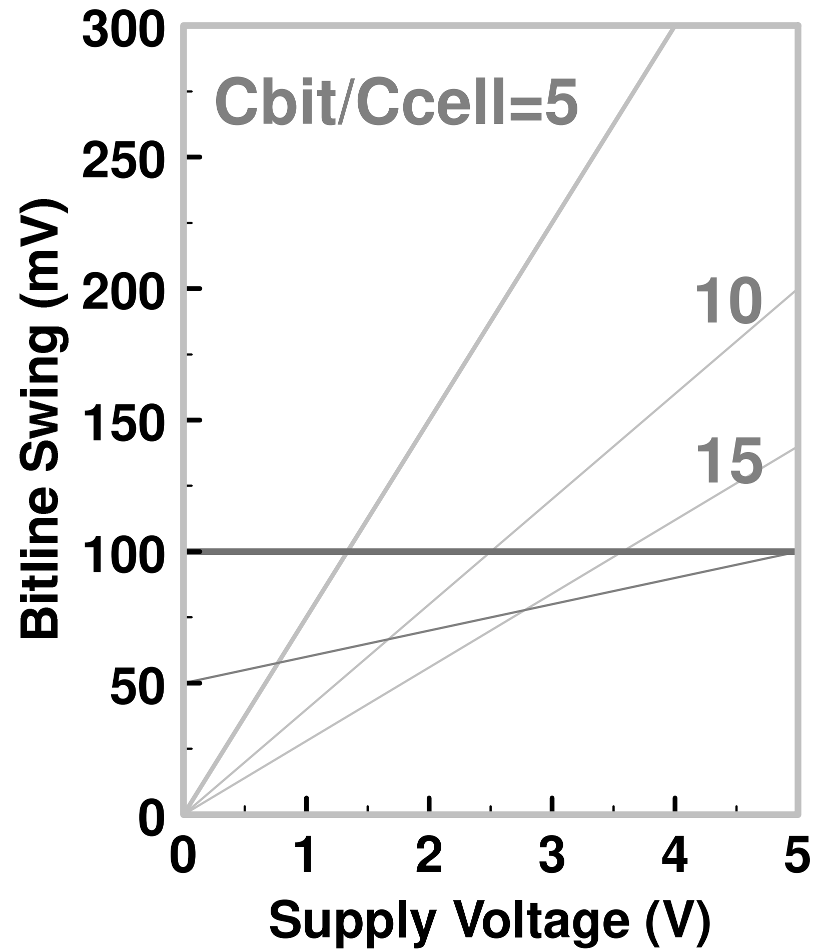

Cell Area (um2)
10
1
0.1
4
16
64
256
1,024
DRAM Generation (Mbit)
2.79x
2.67x
2.56x
2.47x
?
3x
$8\lambda^2$
$6\lambda^2$

Fig 11





Adapted from [2]

Fig 12

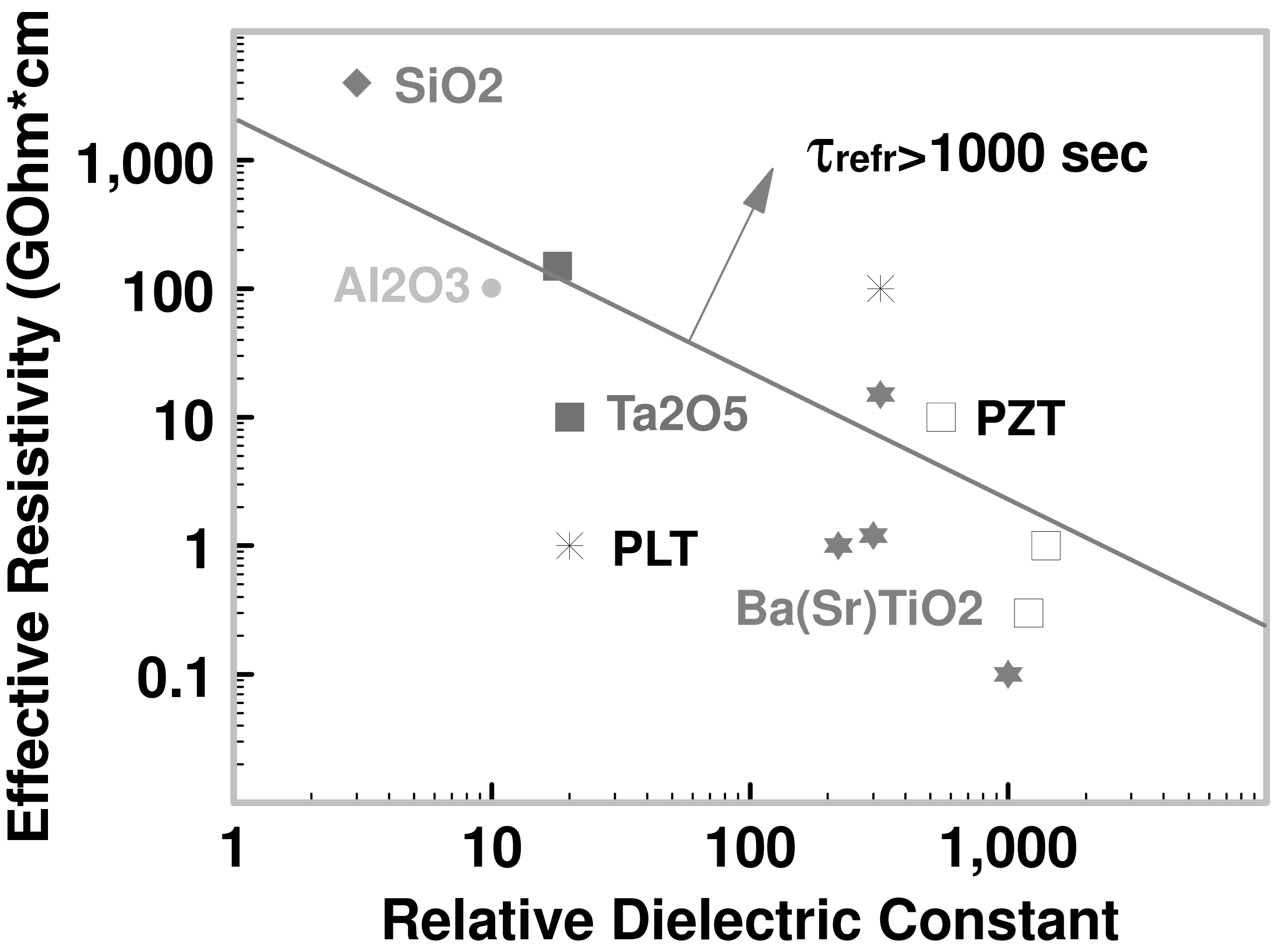

Effective Resistivity (GOhm*cm
1,000
100
10
1
0.1
SiO2
Al2O3
Ta2O5
PLT
PZT
Ba(Sr)TiO2
τ_{refr}>1000 sec
1
10
100
1,000
Relative Dielectric Constant

Fig 13

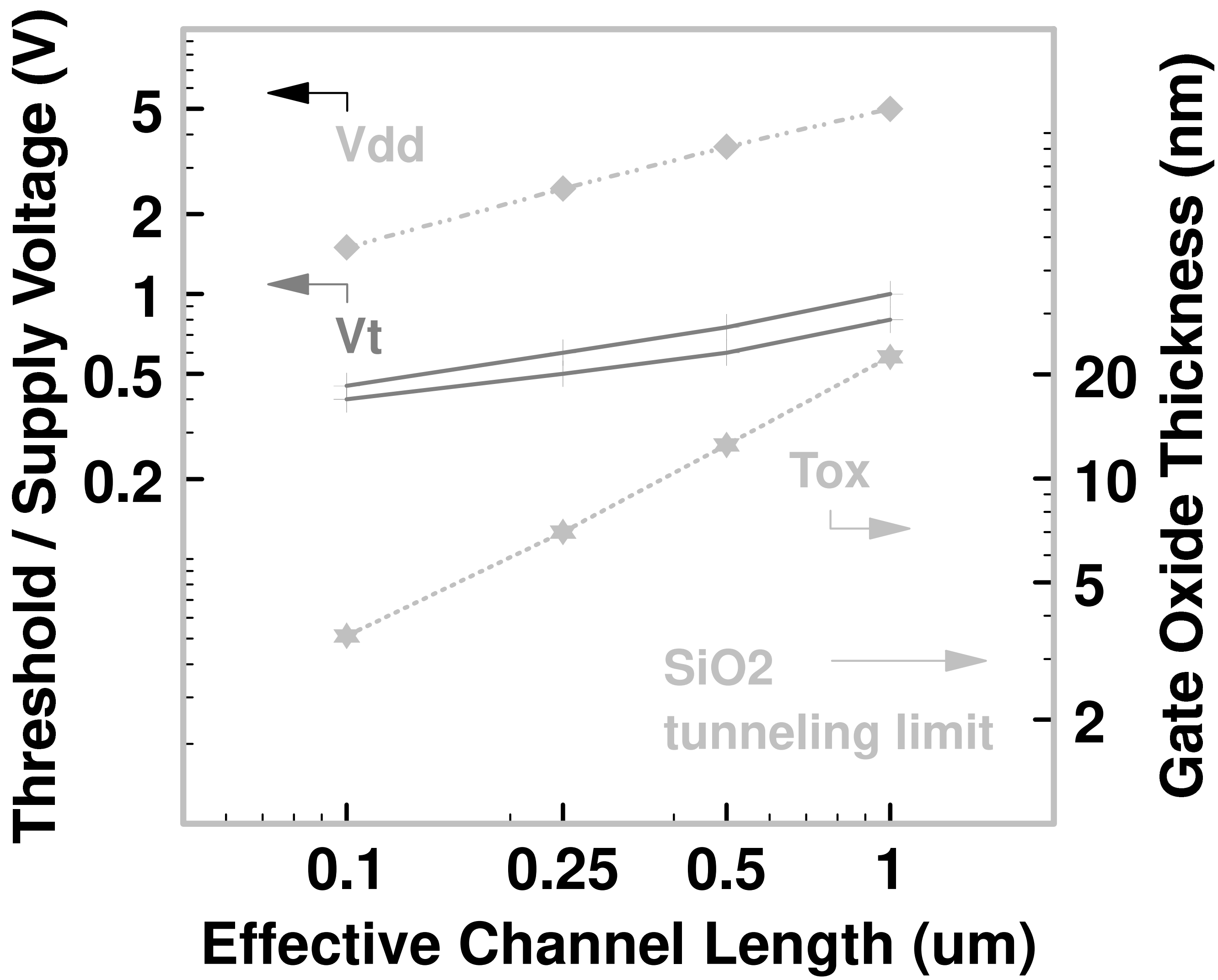

Threshold / Supply Voltage (V)
5
2
1
0.5
0.2
Vdd
Vt
Tox
SiO2
tunneling limit
20
10
5
2
Gate Oxide Thickness (nm)
0.1
0.25
0.5
1
Effective Channel Length (um)

Fig 14

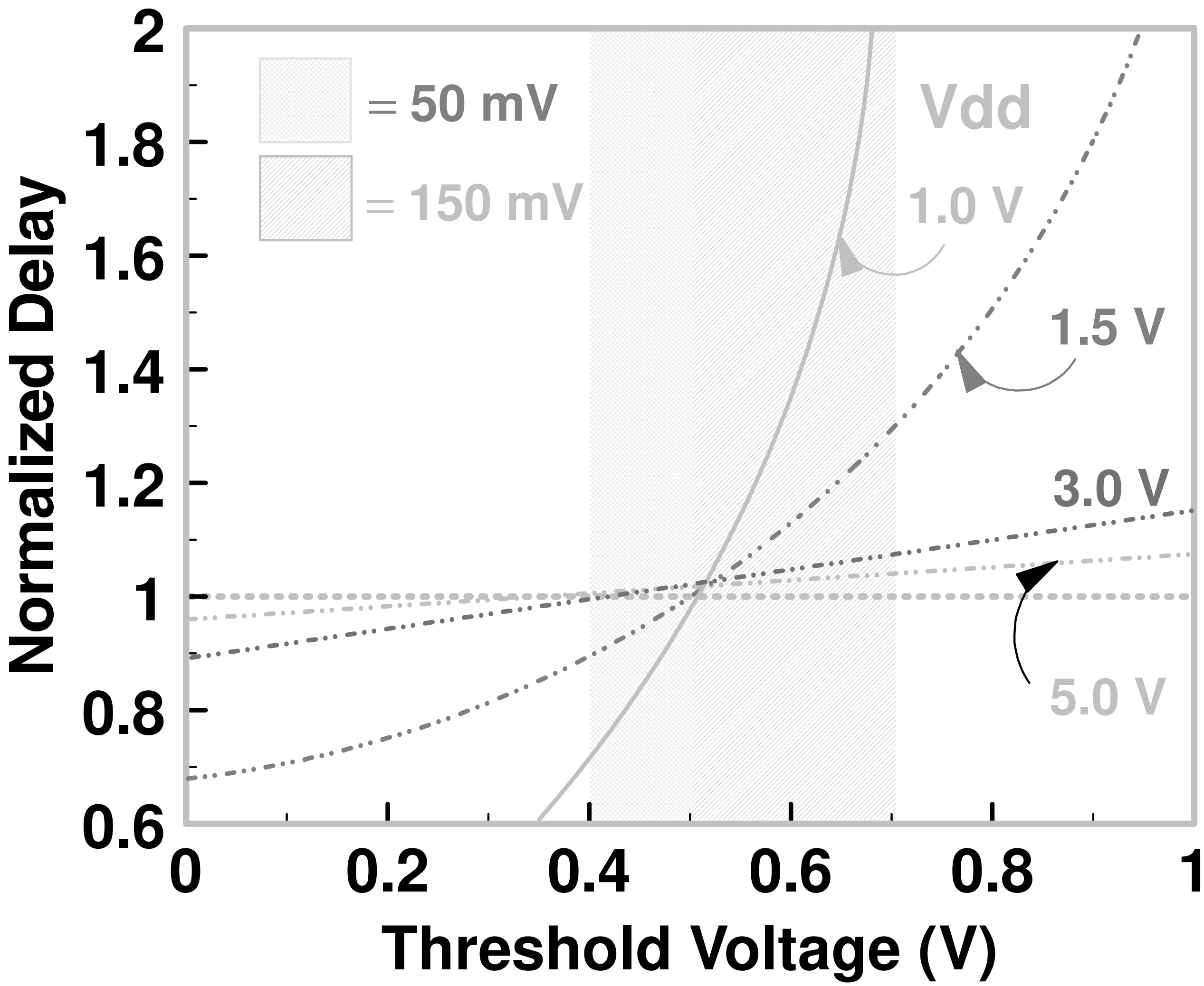
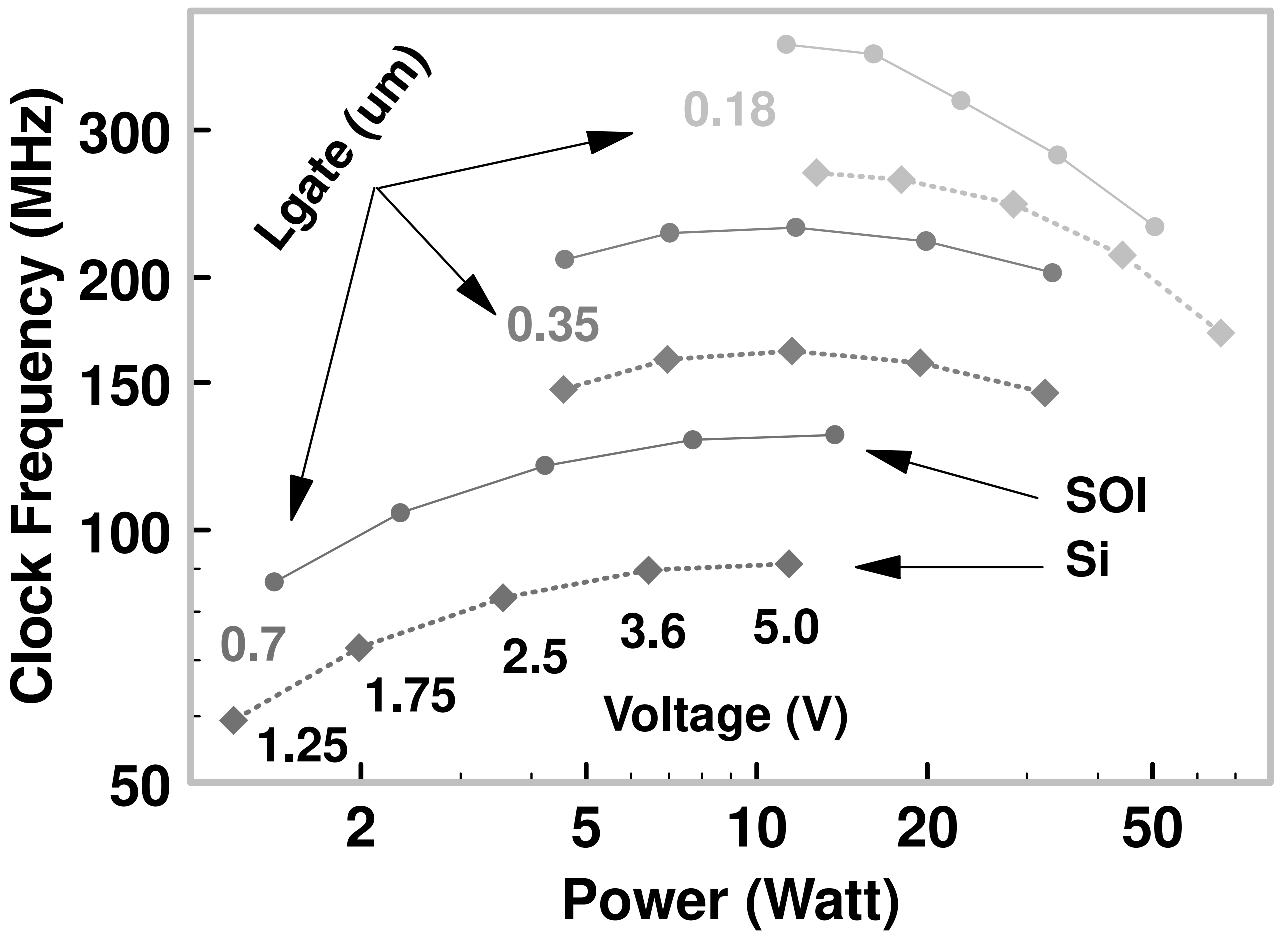


Fig 15

Adapted from [3]



Lgate (um)
0.18
0.35
0.7
SOI
Si
1.25
1.75
2.5
3.6
5.0
Voltage (V)
Clock Frequency (MHz)
300
200
150
100
50
2
5
10
20
50
Power (Watt)

Fig 16