Power consumption and performance of multi-core processors in view of Dark Silicon

Mohanan P R
Research scholar, Department of ship technology,
Cochin University of Science and Technology, Ernakulam, Kerala, India

Mariamma Chacko
Professor, Department of ship technology,
Cochin University of Science and Technology, Ernakulam, Kerala, India

Abstract - Power consumption is now the major technical problem facing semiconductor industry, especially in portable embedded system devices. The two principle sources of power dissipation in present day microprocessors are Dynamic power and Static power. There exists number of techniques to optimize the dynamic power consumption but limited researches are concentrated on static power. Until very recently only dynamic power has been considered as a significant source of power consumption. But when processor technology advances below hundred nanometers, that is when it comes down to smaller geometries, it will exacerbate leakage. So static power begins to dominate the power consumption equation in microprocessor design, this results in the so-called dark silicon problem. Performance and efficiency of Multi-core systems face a new set of challenges at the verge of dark silicon. This paper presents a review of different methods to reduce power consumption of multicore processor through investigations.

Keywords – Power management, Multicore processor, DVFS, Static power, Dark silicon.

I.INTRODUCTION

Energy issues being the most critical one need prime attention and prevention in the field of multiprocessor systems. Today more than 90% of computational applications support multiprocessor systems to meet the performance complexities [1][2]. Extended support for multiprogramming and parallel programming also makes multi-processor systems most sought. The more challenging problem in a multiprocessor system is optimally allocating the jobs and scheduling them. More over the green computing paradigm request low power computation in the operating system level.

The power consumption in embedded multiprocessor systems can be broadly divided into two, Static power and Dynamic power.

$$P_{total} = P_{dynamic} + P_{static} \quad (1)$$

In order to avoid too high power dissipation, a certain part of the chip needs to remain inactive, the inactive part is termed Dark Silicon. The main contributor for static power is Dark silicon. Hence, we have to operate working cores in a multi-core system at less than their full capacity limiting the performance, resource utilization and efficiency of the system. When not switching, CMOS transistors consume negligible power. However, the power they consume has increased dramatically with increases in device speed and chip density.

ITRS (International Technology Roadmap for Semiconductors) projections have predicted that by 2020, designers would face up to 90% of dark silicon, meaning that only 10% of the chip's hardware resources are useful at any given time when high operating frequencies are applied [3]. Until very recently, only dynamic power has been significant

source of power consumption, and Moore's law has helped to control it. Dynamic power is proportional to the square of supply voltage, so reducing the voltage significantly reduces power consumption. Unfortunately , smaller geometries exacerbate leakage, so static power begins to dominate the power consumption equation in microprocessor design[4]. To improve the utility of the embedded system, the power consumption of the entire system has to be minimized. Since the major computations are done by the embedded processor/controller, energy minimization of the processor is also vital for the total power reduction, even though it is relatively small.

II. POWER DISSIPATION IN SEMICONDUCTOR DEVICES

Multicore processors are ubiquitous now both in general purpose and application-specific computing systems starting from smartphones to commercial servers. Power dissipation is an important constraint because it increases the temperature and cooling costs, reduces reliability, and degrades performance.[5]. Power consumed by a processing element can be resolved into three parts [6] as shown

$$P_{tot} = P_{static} + P_{dynamic} + P_{shortcircuit} \quad (2)$$

where P_{tot} is total power dissipation, P_{static} is static or leakage power due to leakage of a transistor's bias currents. $P_{dynamic}$ is dynamic power due to switching of transistors. P_{short} circuit is power dissipation related to concurrent conduction of p type and n type transistors.

$$P_{static} = V \times I_{Leak} \quad (3)$$

V = source voltage,
I_{Leak} = transistor leakage current.
Usually, leakage power contributes 20–40% of the total power dissipation [7].

$$P_{dynamic} = \beta \times C_{Load} V^2 \times f \quad (4)$$

β= activity factor
C_{Load} = effective load capacitance and
f = switching speed.

$$P_{short\ circuit} = V \times I_{SC} \quad (5)$$

I_{sc} = short circuit current flowing from supply to ground.

III. POWER CONSUMPTION IN EMBEDDED PROCESSORS

To improve the utility and performance of portable embedded system, the power consumption of the entire system has to be minimized. Since the major computations are done by the embedded processor/controller, energy minimization of the processor is also vital for the total power reduction, even though it is relatively small[8]. For large and complex applications like smart grid, sensor network etc., where thousands and lakhs of processors are present and are performing computations for almost all the time, the cumulative energy consumption of the entire nodes in the network cannot be neglected.

The two major types of power dissipations occurring in a CMOS circuit are:

- Static dissipation : Due to leakage currents

- Dynamic dissipation: Due to the charging and discharging of capacitance or due to the switching activities of the circuit. As switching activities increase with increased clock frequencies, processors with high operating clock will have more power dissipation [9][7].

The dynamic power consumption of the processor core varies significantly depending on the workload while the leakage power consumption is almost constant. Therefore, leakage power is even a larger percentage of the total power consumption on average. Reducing leakage power can be especially important to battery when a system is idle for a long time, such as for mobile phones.

Low power applications have been driving the supply voltage to become lower and lower, which requires the device threshold to be reduced so as to satisfy performance requirements. This leads to dramatic increase of leakage current due to the exponential relationship between leakage current and threshold voltage V_{th}.

$$I_{off} = K_1 W \, e^{-V_{th} / n V_\theta} \, (1 - e^{-V/V_\theta}) \qquad (6)$$

where K_1 and n are experimentally derived constants, W is transistor width, V_θ is 25 mV. Consequently, leakage power (static power) is no longer negligible in low voltage circuits.

The recent trend indicates that leakage power will contribute as much to total power consumption than dynamic power. Therefore, optimization techniques for leakage power are also necessary. The device and circuits communities have been concerned with increasing leakage power for several generations[10].

3.1 Dynamic Power Consumption

As far as we know, prior works on power reduction at the system level has been focused almost entirely on dynamic power. In order to limit dynamic power consumption, techniques such as clocking gating [11] and cache sub banking [12] have been employed. The goal of these techniques is to reduce the number or frequency of switching devices. Optimization of the supply voltage to reduce the power/performance ratio is also performed [13], this has the added benefit of addressing dynamic power consumption, which is proportional to the square of the supply voltage. However, all of these techniques are hardly used to reduce leakage power.

3.1.1 Dynamic Voltage and Frequency Scaling(DVFS)

DVFS is a technique for altering the voltage and/or frequency of a computing system based on performance and power requirements. For CMOS circuits, dynamic power is related with voltage and frequency as $P \propto FV^2$ and hence, by reducing the frequency, the voltage at which the circuit needs to be operated for stable operation can also be lowered, which leads to energy saving. Several commercial microprocessors support DVFS technology for saving power, e.g. AMD PowerNow and Intel's SpeedStep. The limitation of DVFS is that it harms the performance and hence, it may increase execution time or lead to missed deadlines. Also DVFS requires programmable clock generator and DC-DC converter which incur energy overhead. Further, voltage transitions may require time on the order of tens of microseconds [14]. Due to the trend of using multicore processor, the leakage energy will be increased and returns from DVFS are diminishing.

A genetic algorithm has been employed to integrate task scheduling and voltage scaling under a single iterative optimization loop. The technique searches the solution space to find an assignment and ordering of tasks on each processing element and generates a schedule such that deadline constraints are met and the power consumption is minimized. Further, the technique distributes the slack proportionately to different tasks and uses DVFS to save energy[15]. The method may be used for saving energy in both homogeneous and heterogeneous multiprocessors.

To achieve high task completion ratio(i.e. Low deadline miss ratio), several multimedia applications(e.g. World Wide Web, Interactive TV, Virtual reality etc.) are utilizing the fact that missing some task deadlines can be acceptable since it remains unnoticed to human visual and auditory system [16]. Along with the information on statistical task execution time can save energy in embedded systems by dynamic voltage scaling. It contains two algorithms, first

ensures, achieving highest completion ratio with lowest possible energy consumption. The second algorithm deliberately drops some tasks to create slack for saving additional energy, such that application specific quality-of-service constraint is fulfilled.

The DVFS technique enables achieving a precise energy-performance tradeoff, by using runtime information about the external memory access statistics and chooses the optimal CPU clock frequency and the corresponding minimum voltage level based on the ratio of the on chip computation time to the off-chip access time[17]. This technique lowers the CPU frequency in the memory bound region of a program to keep the performance degradation to a low value.

A multiprocessor scheduling technique is there to save energy. By using the earliest deadline first scheduling, schedules periodic real time tasks. Which ensures meeting the deadline of all tasks while minimizing energy consumption [18]. For saving energy in soft real time systems, the DVFS technique find an optimal frequency for a task assuming availability of continuous range of frequencies. From the actually available frequencies, the closest frequencies which are smaller and larger than the optimal frequency are chosen and fraction of time each of them should be used to meet the deadline is decided [19].

3.2 The Dark Silicon Phenomenon

The main contributor for the leakage power is Dark silicon. Too high power density is the chief contributor to dark silicon. It is to be noted that trends in computer design that are technology node centric have a significant impact on area, voltage, frequency, power, performance, energy and reliability issues. Most of these parameters are cyclically dependent on one another in a way that improving one aspect will deteriorate the other. Computational intensity of future applications such as deep machine learning, virtual reality, big data, etc. demands further technology scaling, leading to further rise in power densities and dark silicon issue. Increase in power density leads to thermal issues[20], hampering the chip's performance and functionality. Subsequently, issues of reliability and ageing have come up, in addition to limitation in performance. Figure shows the amount of usable logic on a chip as per projections of ITRS, [2][20][21]. Increasing dark silicon directly reflects on performance to a point that many core scaling provides zero gain [21]. Thermal awareness and dark silicon sensitive resource allocation are key techniques that can address these challenges.

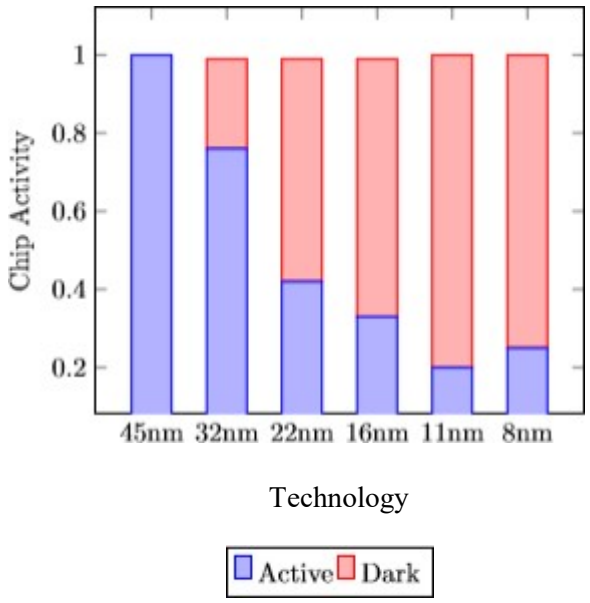

Figure1: Projections of Dark silicon

Moore's Law (the doubling of transistors on chip every 18 months) has been a fundamental driver of computing[22]. For the past three decades, through device, circuit, microarchitecture, architecture, and compiler advances, Moore's law coupled with Dennard scaling, has resulted in commensurate exponential performance increases. The recent shift to multicore designs has aimed to increase the number of cores along with transistor count increases, and continue the proportional scaling of performance. As a result, architecture researchers have started focusing on 100-core and 1000-core chips and related research topics and called for changes to the undergraduate curriculum to solve the parallel programming challenge for multicore designs at these scales.

3.2.1 Dennard's Scaling and Its Breakdown

It observes that transistor dimensions could be scaled by-30% (0.7x) every technology generation, thus reducing their area by 50%. This would reduce circuit delays by 30% (0.7x) and therefore increase operating frequency by about 40% (1.4x). Finally, to keep the electric field constant, voltage is reduced by 30%, reducing energy by 65% and power (at 1.4x frequency) by 50%. Therefore, in every technology generation, if the transistor density doubles, the circuit becomes 40% faster, and power consumption (with twice the number of transistors) stays the same[23].

The dynamic (switching) power consumption of CMOS circuits is proportional to frequency [24]. Historically, the transistor power reduction afforded by Dennard scaling allowed manufacturers to drastically raise clock frequencies from one generation to the next without significantly increasing overall circuit power consumption.

Since around 2005–2007 Dennard scaling appears to have broken down. As of 2016, transistor counts in integrated circuits are still growing, but the resulting improvements in performance are more gradual than the speed-ups resulting from significant frequency increases [25][26]. The primary reason cited for the breakdown is that at small sizes, current leakage poses greater challenges and also causes the chip to heat up, which creates a threat of thermal runaway and therefore further increases energy costs [25][26].

The breakdown of Dennard scaling and resulting inability to increase clock frequencies significantly has caused most CPU manufacturers to focus on multicore processors as an alternative way to improve performance. An increased core count benefits many (though by no means all) workloads, but the increase in active switching elements from having multiple cores still results in increased overall power consumption and thus worsens CPU power dissipation issues [2][27]. The end result is that only some fraction of an integrated circuit can actually be active at any given point in time without violating power constraints. The remaining (inactive) area is referred to as dark silicon.

3.2.2 Leakage Power

Leakage current is split into two components - sub-threshold leakage and gate oxide leakage, as in Equation

$$P_{static} = VI_{leak} = V \times (I_{sub} + I_{ox}) \quad (7)$$

Leakage current is the static power drawn even at an off state (not switching), meaning that a chip can still consume appreciable power without being active [31][24]. Off-state leakage current is experimentally derived in [28], and is represented as in Equation (8).

$$I_{off} = K_1 W e^{-V_{th}/nV_\theta} (1 - e^{-V/V_\theta}) \quad (8)$$

Where K_1 and n are experimentally derived constants, W is transistor width, V_θ is 25mV, with V_θ being much smaller than V.

This indicates that leakage current is low when the exponential component is small. In other words, leakage current increases exponentially with a decrease in threshold voltage (8). Reduction in transistor physical parameters reduces voltage and thus also reduces threshold voltage. This results in higher leakage current and thus higher static power being drawn. It can be said that with every technology node generation, leakage current increases exponentially with a decrease in threshold voltage [33][29]. Reduction in transistor physical dimensions also reduces oxide thickness, to an extent that electrons start to leak also through the oxide resulting in leakage current [30]. Oxide thickness is reaching around 1.5nm, results in significant leakage current (I_{ox}) [31]. Gate oxide leakage current is shown in Equation (9).

$$I_{ox} = K_2 W (V/T_{ox})^2 e^{-\alpha T_{ox}/V} \quad (9)$$

where K_2 and α are experimental constants.

The squared denominator and negative exponential component of Equation shows that gate oxide leakage increases exponentially with decrease in oxide thickness. Leakage increases exponentially at higher temperatures, implying that a chip would have far more leakage if it operates to its full potential[32].

3.2.3. Consequences of Dark Silicon

Increase in the number of processing elements (According to Moore's Law[23]) on a chip led to emergence of hardware redundancies in processor architectures mainly to allow performance increase. Theoretically, it can be proven that scaling down the transistor size and simultaneously decreasing the operating voltage keeps the power density of the chip constant. This is known as Dennard scaling [4]. High power density causes more heat generation that needs to be dissipated via different cooling methods. Otherwise high temperature may affect the functionality of the chip causing less reliable computation and even burning of the chip. Therefore, to keep the power density at a tolerable level, some parts of the chip need to be kept inactive. Such inactive parts are called Dark Silicon [33]. Due to the phenomena of dark silicon the following problems may arise.

3.2.3.1. Performance Degradation

With a section of the chip being inactive, expected performance from many-core systems can never be realized in practice. Transistor scaling has paved the way to integrate more cores at a lower power consumption, effectively increasing compute capacity per area [40][23]. Inherent parallelism, if present, in workloads could take advantage of multi and many-core computers to show improved performance to a certain extent [34]. Dark silicon changes this consensus, since we cannot power up all the available resources at any given time [35].

3.2.3.2. Energy Efficiency

Deterioration in performance due to dark silicon affects energy efficiency of the system [36]. With static power dominating the total power consumption, achievable performance in a fixed energy budget i.e., performance per watt, is declining. Around half of the energy budget is wasted towards static power, literally implying spending energy for doing nothing.

3.2.3.3. Thermal Management

High power density and accumulation of heat result in an increase of average and peak temperatures of the chip. This opens the window for manifesting soft errors and unreliability in computation [37], in addition to decrease in life time of the chip[38]. This ages the chip faster and puts an additional penalty on manufacturing cost per reliable chip. A chip working at full throttle can dissipate more heat and thus high temperature, which in turn increases leakage too [32]. Again, the omnipresent cyclic dependency of temperature, leakage, performance and energy efficiency surface here. Further, dissipating heat requires a heat sink and a fan that blows air through it, adding to manufacturing costs. For a data centric level compute platform, more sophisticated cooling solutions have to be used, further increasing maintenance costs.

3.2.4 Some Solutions to Dark Silicon Problem

Architectural and run-time management techniques can mitigate dark silicon to a large extent and improve energy efficiency.

Architecture and Implementation Perspective

Operating core with same instruction set architecture (ISA) but different microarchitectures, possibly at different voltage and frequency levels result in asymmetric cores with different power-performance characteristics. Such an asymmetric multi-core system offers the choice of executing specific application on a core with suitable power-performance characteristics that can improve energy efficiency. Near threshold computing (NTC) [39], where the chip's supply voltage is scaled down to a slow as threshold voltage offers ultra-low power consumption, at the loss of performance. Near threshold computing leads to better per-chip-throughput cores, also termed Dim-Silicon. Energy efficiency can be improved by selectively down scaling the voltage, inducing Dim Silicon over darksilicon.

Run-time Resource Management-Computational Perspective

Run-time management operates in an observe-decide-act loop to enhance the energy efficiency of underlying hardware. Observe phase involves monitoring critical parameters of chip's performance including instantaneous power consumption, active and dark cores, network congestion, temperature accumulation, etc. [40]. The Decide phase uses a pro-active strategy, relying on application mapping, resource allocation and scheduling to reduce potential dark silicon. Act phase employs a reactive strategy that applies different actuators such as clock and power gating, voltage and frequency scaling, hardware re-configuration, task migration, to mitigate dark silicon. Run-time estimation of power budget subject to the set of applications currently executing on the chip and their spatial alignment can provide a thermally safe upper bound on power budget, variable at run-time. This variable upper bound of Thermal Safe Power (TSP) improves utilization of available power budget and minimizes darksilicon.

Design and Management-Communication Perspective

Power management actuators, like power gating and voltage and frequency scaling, can also be applied to interconnect components at both design time and run-time, giving fine grained control. An optimized network with suitable energy efficient and reliable interconnect components can increase overall energy efficiency of the system[41].

Operating a section of the chip at highest possible voltage and frequency for a brief period of execution time is known as computational sprinting. Sprinting can accelerate performance with an increase in power consumption, possibly violating the safe upper bound. However, the sprinting time is kept low enough so that increase in power does not reflect directly in temperature accumulation. Since thermal safety is maintained, relatively high performance and utilization can be achieved despite dark silicon.

IV. CONCLUSION

Review of power consumption and performance of multi-core systems in view of dark silicon phenomenon is carried out and presented. The importance of keeping power consumption within the limits to achieve green computing paradigm are analyzed. It is found from the survey, reducing leakage power is important when a system is idle for a long time, such as for mobile communication devices. Since the main contributor for leakage power is dark silicon novel ideas can be introduced to reduce the Dark silicon phenomena in multi-core systems.

REFERENCES

[1] A. Kanduri, A. M. Rahmani, P. Liljeberg, A. Hemani, A. Jantsch, and H. Tenhunen, "A perspective on dark silicon," in *The Dark Side of Silicon: Energy Efficient Computing in the Dark Silicon Era*, Springer International Publishing, 2017, pp. 3–20.

[2] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," *ACM SIGARCH Comput. Archit. News*, vol. 39, no. 3, pp. 365–376, 2011.

[3] G. Grau, "ROADMAPPING VS. S-CURVES: HOW TO SWITCH TO THE NEXT S-CURVE Analyzed using the example of the semiconductor industry," *Interdiscip. Manag. Res. Ix*, vol. 9, no. Moore 1965, pp. 173–182, 2013.

[4] R. Dennard, F. Gaensslen, W.-N. Yu, L. Rideout, E. Bassous, and A. Le Blanc, "Design of Ion-Implanted Small MOSFET ' S Dimensions with Very," *IEEE J. Solid State Circuits*, vol. 9, no. 5, pp. 257–268, 1974.

[5] K. Nagalakshmi and N. Gomathi, "Analysis of power management techniques in multicore processors," *Adv. Intell. Syst. Comput.*, vol. 517, no. July, pp. 397–418, 2017, doi: 10.1007/978-981-10-3174-8_35.

[6] S. Zhuravlev, J. C. Saez, S. Blagodurov, A. Fedorova, and M. Prieto, "Survey of energy-cognizant scheduling techniques," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 7, pp. 1447–1464, 2013, doi: 10.1109/TPDS.2012.20.

[7] D. Chinnery and K. Keutzer, "Overview of the factors affecting the power consumption," *Closing Power Gap Between ASIC Cust. Tools Tech. Low Power Des.*, pp. 11–53, 2007, doi: 10.1007/978-0-387-68953-1_2.

[8] A. Pillai and T. Isha, "Factors causing power consumption in a embedded processor-a study," *Int. J. Appl. or Innov.*, vol. 2, no. 7, pp. 300–306, 2013, [Online]. Available: http://www.ijaiem.org/volume2issue7/IJAIEM-2013-07-23-077.pdf.

[9] "Gary K. Yeap Practical Low Power Digital VLSI Design – World of Digitals." https://www.world-of-digitals.com/gary-k-yeap-practical-low-power-digital-vlsi-design-ebook-pdf (accessed May 19, 2021).

[10] H. Fahad, M. Hasan, G. Li, and M. Hussain, "Thermoelectricity from wasted heat of integrated circuits," *Appl. Nanosci.*, vol. 3, no. 3, pp. 175–178, 2013, doi: 10.1007/s13204-012-0128-2.

[11] Y. P. You, C. Lee, and J. K. Lee, "Compiler analysis and supports for leakage power reduction on microprocessors," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2481 LNCS, no. 89, pp. 45–60, 2005, doi: 10.1007/11596110_4.

[12] L. Benini and G. De Micheli, "Energy-efficient system-level design."

[13] I. Hong, D. Kirovski, G. Qu, M. Potkonjak, and M. B. Srivastava, "Power optimization of variable voltage core-based systems," *Proc. - Des. Autom. Conf.*, pp. 176–181, 1998, doi: 10.1145/277044.277088.

[14] W. Kim, M. S. Gupta, G. Wei, and D. Brooks, "High Performance Computer Architecture, 2008, HPCA 2008, IEEE 14th International Symposium on : date, 16-20 Feb. 2008.," pp. 123–134, 2008.

[15] V. Kianzad, S. S. Bhattacharyya, and G. Qu, "CASPER: An integrated energy-driven approach for task graph scheduling on distributed embedded systems," *Proc. Int. Conf. Appl. Syst. Archit. Process.*, pp. 191–197, 2005, doi: 10.1109/asap.2005.23.

[16] S. Hua and G. Qu, "Approaching the Maximum Energy Saving on Embedded Systems with Multiple Voltages," *IEEE/ACM Int. Conf. Comput. Des. Dig. Tech. Pap.*, pp. 26–29, 2003, doi: 10.1109/iccad.2003.159666.

[17] K. Choi, R. Soma, and M. Pedram, "Fine-grained dynamic voltage and frequency scaling for precise energy and performance trade-off based on the ratio of off-chip access to on-chip computation times," *Proc. - Des. Autom. Test Eur. Conf. Exhib.*, vol. 1, no. 1, pp. 4–9, 2004, doi: 10.1109/date.2004.1268819.

[18] C. Xian, Y. H. Lu, and Z. Li, "Energy-aware scheduling for real-time multiprocessor systems with uncertain task execution time," *Proc. - Des. Autom. Conf.*, pp. 664–669, 2007, doi: 10.1109/DAC.2007.375248.

[19] E. Y. Y. Kan, W. K. Chan, and T. H. Tse, "Leveraging performance and power savings for embedded systems using multiple target deadlines," *Proc. - Int. Conf. Qual. Softw.*, pp. 473–480, 2010, doi: 10.1109/QSIC.2010.59.

[20] J. Kong, S. W. Chung, and K. Skadron, "Recent thermal management techniques for microprocessors," *ACM Comput. Surv.*, vol. 44, no. 3, Jun. 2012, doi: 10.1145/2187671.2187675.

[21] E. For, "I Nternational T Echnology R Oadmap for R Adio F Requency and a Nalog / M Ixed - Signal T Echnologies for," 2011.

[22] V. Tiwari, D. Singh, S. Rajgopal, G. Mehta, R. Patel, and F. Baez, "Reducing power in high-performance microprocessors," *Proc. - Des. Autom. Conf.*, pp. 732–737, 1998, doi: 10.1145/277044.277227.

[23] G. E. Moore, "Cramming More Components onto Integrated Circuits."

[24] S. Borkar and A. A. Chien, "The future of microprocessors," *Commun. ACM*, vol. 54, no. 5, pp. 67–77, 2011, doi: 10.1145/1941487.1941507.

[25] E. Total, P. Dissipation, and B. On, "Estimating Power Dissipation in Application an-154 Introduction Estimating Total Power Dissipation," *Power*, pp. 99–103, 1996.

[26] R. Dennard, F. Gaensslen, H. Yu, V. L. Rideout, and E. Bassous, "TECHNICAL ARTICLES ore than three decades have passed since the," *IEEE SSCS Newsl.*, pp. 11–13, 2007.

[27] J. Hruska, "The death of CPU scaling: From one core to many - and why we're still stuck - ExtremeTech," *ExtremeTech*, p. 3, 2012, [Online]. Available: https://www.extremetech.com/computing/116561-the-death-of-cpu-scaling-from-one-core-to-many-and-why-were-still-stuck.

[28] A. Chandrakasan, W. J. Bowhill, and F. Fox, "Design of high-performance microprocessor circuits," *Des. High-Performance Microprocess. Circuits*, pp. 1–557, 2000, doi: 10.1109/9780470544365.

[29] S. Borkar, "Getting Gigascale Chips," *Queue*, vol. 1, no. 7, pp. 26–33, 2003, doi: 10.1145/957717.957757.

[30] D. Gate and O. Si, "Direct-Tunneling Gate Oxide Si MOSFET's," vol. 43, no. 8, pp. 1233–1242, 1996.

[31] C. H. Choi, K. Y. Nam, Z. Yu, and R. W. Dutton, "Impact of gate direct tunneling current on circuit performance: A simulation study," *IEEE Trans. Electron Devices*, vol. 48, no. 12, pp. 2823–2829, 2001, doi: 10.1109/16.974710.

[32] G. Sery, S. Borkar, and V. De, "Life is CMOS: Why chase the life after?," *Proc. - Des. Autom. Conf.*, pp. 78–83, 2002, doi: 10.1109/dac.2002.1012598.

[33] H. Esmaeilzadeh, E. Blem, R. St. Amant, K. Sankaralingam, and D. Burger, "Dark silicon and the end of multicore scaling," *IEEE Micro*, vol. 32, no. 3, pp. 122–134, 2012, doi: 10.1109/MM.2012.17.

[34] G. M. Amdahl, "Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities, Reprinted from the AFIPS Conference Proceedings, Vol. 30 (Atlantic City, N.J., Apr. 18amp;#x2013;20), AFIPS Press, Reston, Va., 1967, pp. 483amp;#x2013;485, when Dr," *IEEE SolidState Circuits Newsl.*, vol. 12, no. 3, pp. 19–20, 2007, [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4785615.

[35] G. Venkatesh, J. Sampson, N. Goulding-Hotta, S. K. Venkata, M. B. Taylor, and S. Swanson, "QsCores: Trading dark silicon for scalable energy efficiency with quasi-specific cores," *Proc. Annu. Int. Symp. Microarchitecture, MICRO*, pp. 163–174, 2011, doi: 10.1145/2155620.2155640.

[36] M. B. Taylor, "Is Dark Silicon Useful ? Harnessing the Four Horsemen of the Coming Dark Silicon Apocalypse U " liza " on Wall : Dark Silicon ' s Effect on Mul " core Scaling," *Proc. Des. Autom. Conf.* -, pp. 1131–1136, 2012.

[37] J. Srinivasan, S. V. Adve, P. Bose, and J. A. Rivers, "The case for lifetime reliability-aware microprocessors," *Conf. Proc. - Annu. Int. Symp. Comput. Archit. ISCA*, vol. 31, pp. 276–287, 2004, doi: 10.1145/1028176.1006725.

[38] J. Srinivasan, S. V. Adve, P. Bose, and J. A. Rivers, "The impact of technology scaling on lifetime reliability," *Proc. Int. Conf. Dependable Syst. Networks*, pp. 177–186, 2004.

[39] M. B. Taylor, "A landscape of the new dark silicon design regime," *IEEE Micro*, vol. 33, no. 5, pp. 8–19, 2013, doi: 10.1109/MM.2013.90.

[40] A. S. Pillai and T. B. Isha, "Optimal task allocation and scheduling for power saving in multiprocessor systems," *2014 Power Energy Syst. Conf. Towar. Sustain. Energy, PESTSE 2014*, no. Pestse, 2014, doi: 10.1109/PESTSE.2014.6805324.

[41] IEEE, "International Roadmap for Devices and Systems 2017 Edition : More Moore," *IEEE Adv. Technol. Humanit.*, 2017.