CHAPTER 2

LITERATURE SURVEY

2.1 Introduction

For decades, the silicon industry has continuously scaled CMOS technology to pack billions of transistors within a few millimeters of chip area with enhanced performance. As a result, the packing density of today's SoCs has improved by more than six orders of magnitude, providing systems with lower power consumption, higher speed, and higher computing capability [1]. The tremendous improvement in the processing ability, availability of large data set size, and optimization of algorithms have led to the emergence of modern computing systems with a wide range of functionalities from simple calculative tasks to complex AI tasks such as image and speech recognition [2]-[5]. This improvement in overall performance further enables the support for large number of applications with diverse computational and energy requirements such as laptops, smartphones, personal computers, wearable healthcare tracker systems, and wireless sensor nodes. Moreover, the increasing demand for interconnectivity is giving rise to a new technology paradigm called as Internet of Things, which allow objects to collect and share the data. The common feature in these applications is their 'Always-ON' sensing circuit that continuously monitors and collects the data from the surrounding. Through the collected data, important events are detected, which are further utilized to activate one or more complex sub-systems. For example, in a face detection system, the presence of the human face will activate the recognition system, which typically remains in sleep mode to reduce standby power consumption.

Most of these smart sensor nodes are predominantly powered by either battery or energy harvesting, resulting in smaller power budgets [6]. Therefore, an energy-efficient operation is one of the key challenges in achieving sustainable deployment of these devices. Memory is one of the largest on-chip components in all modern computing systems, making it a major critical component for energy and performance bottleneck. In addition, the increasing demand for more on-chip capacity and bandwidth makes it more challenging to design a low power memory that can meet the requirement of tighter power budgets of IoT systems. Furthermore, IoT defines the new era of computing characterized by the enormous amount of data (zettabytes). Therefore, new form of energy-efficient computing is required to address the fundamental problem of power management in next-generation applications.

In the following sub-section, we first describe the current and future trend of the memory technology, the demands, and challenges faced by the memory system, and then examine some promising research and design directions to overcome challenges posed by memory scaling. Further, we explore alternative approaches to conventional logic operation to improve the energy-efficiency of the computation.

2.2 Embedded Memories in Modern Computing Systems

Memory is one of the most critical components in a modern computing system that stores the data to be processed. Figure 2.1 presents a typical memory hierarchy with increasing speed and decreasing energy consumption. On the other hand, the packing density decreases and manufacturing cost increases as we move up to the top of the hierarchy. Therefore, the frequently accessed data is stored near to the processor in faster cache memories, whereas data that is not accessed frequently is stored further away from memory in slower memories like DRAM and FLASH.

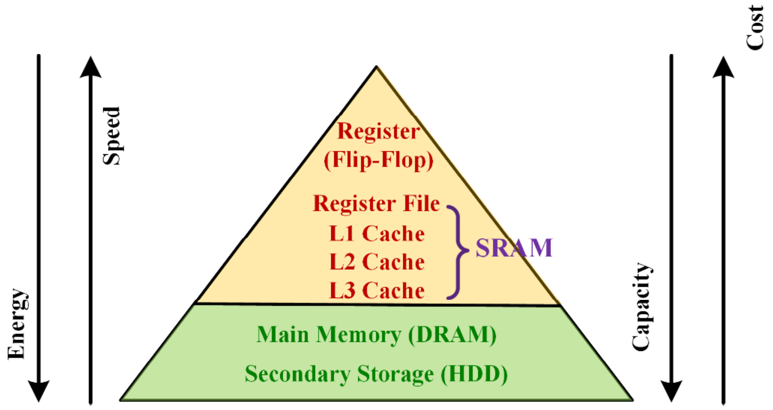

Figure 2.1: Hierarchical memory organization

2.2.1 Memory Technology Trend

With emerging new applications and constantly improving computing systems, the demand for the increased memory capacity, higher performance, lower energy consumption, and reduced cost is increasing at an alarming rate. Such enhanced performance becomes difficult to achieve with the recent trend in systems and application which greatly exacerbate the memory system bottleneck. In particular, at the system/architecture level, the energy consumption of the memory block becomes one of the key design parameters as the memory continues to be responsible for a significant fraction of overall system energy [8]. Moreover, with multicore processing and shared memory system, there is an increased demand for memory bandwidth and capacity along with a relatively new demand for predictable performance and quality of service (QoS) from the memory system [9]-[19].

At the application level, the recent rise in big data applications generates enormous data that needs to be processed in real-time or non-real-time (offline mode) mode to extract important information [20]. The availability of a large amount of data can contribute significantly to numerous research fields such as genomics and postgenomic, drug development, cyber security and intelligence, weather forecasting, etc. For example, next-generation genome sequencing produces a massive amount of raw data related to complex biochemical and regulatory processes in the living being, which can then be used to detect unexpected DNA variation and develop low-cost precision and personalized medicine for improving the health of the patient by individualizing the diagnosis and treatment. The overall performance of a data-intensive application is critically dependent on how efficiently data can be stored and manipulated in a system. However, processing and analyzing such huge amounts of data in the conventional von-neumann based computing system often becomes a major performance bottleneck due to the limited storage and bandwidth of today's memory system [21]-[25]. In addition, there is an increasing trend of integrating heterogeneous systems on the chip, which leads to sharing of memory systems by multiple applications with diverse computational and performance requirements [17]-[26].

At the technology front, one of the major challenges faced by conventional memory technologies, such as SRAM, DRAM, and FLASH memory is, increasing difficulty in scaling the technology to smaller nodes [27][28]. As a result, further improvement in the capacity and performance of the memory technology is not achieved at a lower manufacturing cost [29]-[31]. Some emerging technologies such as phase-change memory [32], [33] and resistive random-access memory [34] have shown promising results because of their unique and attractive features such as scalability, non-volatility, higher speed, lower leakage power consumption and large bandwidth. These memories offer speed like SRAM, density like DRAM and non-volatility like FLASH, which makes the ideal candidate for future computing memories.

2.2.2 Memory System Requirement for Emerging Applications

The ideal requirement of a system designer/architect is high-performing memory with zero latency, unlimited bandwidth, infinite capacity at nearly zero cost. The aforementioned recent trends in the technology, system and application levels not only aggravate the exiting requirement but add some new requirements. We categorize these requirements into two categories: aggravated traditional requirements and new requirements.

The requirement of increased memory performance and capacity is aggravated by the rising demand of the shared memory system due to the integrating heterogeneous applications on a single platform and multi-core processing. Additionally, the increasing difficulty in manufacturing economically smaller technology nodes is further worsening the problem. Over the years, the system and application requirements of performance, capacity, and cost have exacerbated in the following ways: First, these systems require not only high performance and bandwidth but also requires efficient techniques to manage memory interface between multiple cores and various application sharing the memory system to provide an overall performance improvement [35]. Second, high memory pin bandwidth is also required to perform memory-bandwidth-intensive workloads efficiently [36]. Third, in addition to improved performance, the demand for high memory capacity is rapidly increasing due to the placement of multiple data-intensive applications on a single platform and exponentially increasing data sets [37]-[40]. However, the increase in the core count is much faster than DRAM density, and with this trend, it is expected that the memory capacity per core will reduce by 30% every two years [41]. This significantly impacts the performance of the application as most the software innovations and optimizing features can be realized with only high memory capacity. Finally, the difficulty in scaling technology to smaller nodes poses great challenges for the semiconductor industry to develop low-cost and high-density memory systems. Additionally, effective solutions are needed to provide the high reliability and data throughput needs of today’s data-intensive applications.

Furthermore, the relatively new requirement of the system and applications are due to the following reasons. First, the need for more scalable devices has pushed the industry to find alternative technologies which are more compact, consume low power and have low manufacturing cost. Second, the recent rise in multi-core processing and sharing of the memory system, the need for predictive performance, and improved Quality of Service (QoS) are much higher than single-core processors [42]. Finally, to enable more scalable systems where main memory is shared among many application domains, a highly efficient main memory in terms of energy/power/bandwidth is required. The system, application, and technology trends motivate the researchers and designers to develop novel solutions 1) to address the scaling challenges, 2) to enable emerging memory technologies, 3) to improve the predictive performance and quality of the service. The following sub-section focuses on enabling emerging memory technology as a prospective solution to support the requirement of emerging applications.

2.2.3 Challenges with the Conventional Memory Technology

Static Random-Access Memory (SRAM) is one of the most popular choices for volatile embedded memory in most of today's computing systems. The memory must be able to operate reliably under the decreased voltage supply of low-power ICs. The low-voltage operation in SRAM is extremely challenging due to increasing process variation and decreasing bitcell stability. At the same time, increasing parallel processing due to placement of multiple cores on a single chip demand more on-chip cache for effectively sharing the data across parallel processing unit. Integrating more memory on-chip provides an efficient way to reduce off-chip access and reduce power consumption for low-power ICs. Recent works on low-voltage SRAM have shown promising results where several techniques have been employed to achieve an almost 50% reduction in leakage power consumption and low energy consumption per access by trading off the performance. The design of low-voltage SRAM has a significant impact on the area and performance of the memory system. Therefore, reducing this area-overhead and further improving the energy consumption will provide new opportunities for low-power electronics in emerging applications such as wearable devices, smart sensor nodes, implantable medical devices, etc.

The basic building block of embedded memory is 6T SRAM cells organized in a grid with horizontal wordline (WL) and vertical bitline (BL and BL_bar). Each cell is connected to one wordline and a pair of bitlines that are globally connected to supply sources. During a read operation, wordline is asserted high, and the pre-charged bitlines, either BL or BL_bar, discharges based on the value stored in the memory cell. The discharging of the bitline generates a differential voltage at the bitline, which is then converted to a valid logic level using a sense amplifier. During the write operation, wordline is asserted, and pre-charged bitline BL and BL_bar take the value depending on the input data. One of the bitline is discharged by the write driver circuit based on the input data value. The value at bitline then overwrites the value stored in the cell. During the hold mode, the worldline is de-asserted low which isolates the cell from the bitlines. As a result, the memory cell holds the value stored in it. One of the most common techniques to evaluate the stability of the 6T SRAM cell is static noise margin. It is defined as maximum noise at the storage nodes q and qc of the memory cell required to flip the content of the cell. It is calculated by plotting the butterfly curve, which contains the voltage transfer characteristic (VTC) of the two cross-coupled inverters. The obtained VTC curves are plotted on the same axis, assuming different conditions for read, write and hold. During the hold, bitlines are driven by the voltage source while the wordline is de-asserted low. Three roots of intersection are obtained from plotting the VTC of two inverters, indicating two stable points and one metastable point. During the read operation, bitlines are connected to the supply source, and the wordline is asserted high. Similar to the hold butterfly plot, three roots of intersection are obtained. Whereas during the write operation, bitline is set to the desired value whereas wordline is asserted high. Only one root of intersection is obtained, indicating successful flipping of the cell-based on bitline polarity. The increased local variation and fluctuation in the threshold voltage, especially at the lower technology nodes (< 65nm), poses a serious challenge while maintaining the stability of 10^6 to 10^9 memory cells within one die [44]. The low power technique of reduced supply voltage severely degrades the read and write stability of the cell. Moreover, single-event upsets due to radiation also result in corrupt data in the SRAM. The alpha particles radiated from the packing material, or the neutron from the space can penetrate the silicon wafers, and their charge can flip the content of the storage node. This error rate increases with a reduction in supply voltage due to reduced charge stored by the cell internal nodes, which can be easily disturbed by the alpha particles or neutrons hitting the memory cell [45]. To address the challenge of soft error caused by radiations, error-correcting code (ECC) is one of the most commonly used techniques in which extra memory cells are added to each word to detect and correct the errors. The complexity of this technique increases significantly if more than one-bit error needs to be detected and corrected. Therefore, to avoid multi-bit soft error, bit interleaving architecture is utilized to interleave multiple words in a single row [46]. By bit interleaving, multiple error show up as a single-bit error in different words. In addition to the soft error, error while sensing the data value from the cell is also a major concern for the memory designer. The variation in the sensing margin is of one the biggest reasons for read failure. For example, in single-ended sensing, the cell storing logic ‘0’ can produce logic ‘1’ at the output of the sense amplifier before the cell storing logic ‘1’ can produce a true logic ‘1’ at the output or for differential sensing, the difference between the BL and BL_bar falls below the offset voltage of the sense amplifier. Moreover, the timing variation in the peripheral circuit worsens the problem. The sense enable signal applied to the sense amplifier is separately generated using the replica bitline technique, which generates a stable delay for the signal to ensure correct value at the BL and hence the valid data at the output. However, this stability is difficult to preserve across various process corners, a wide range of operating temperature, and lower supply voltages [47]. Recent efforts to improve the sense margin of the cell include shortening the bitline at the cost of area overhead. Overall, a low power memory with degraded area efficiency will increase the system cost and potentially reduce the range of applications. The longer bitlines also causes large delay and power penalty. Therefore, beyond a certain length of bitline, it becomes essential to partition the memory into multiple sub-arrays. The area-overhead of the periphery circuit, such as the sense amplifier within the sub-array is compensated by using multiple columns. For example, for 8-bit word memories, the number of columns in a single row is not restricted to 8. This can be highly beneficial for the larger sense-amplifiers, which also mitigates offset variation in scaled CMOS technology. Hence, bit interleaving architecture can be used to provide soft-error immunity as well as area efficiency. Bit interleaving architecture can be realized in the 6T SRAM cell because cells in the unselected column have bitlines floating at VDD, which corresponds to a dummy read condition and will not upset the data stored at the node. However, as the supply voltage decreases, the SRAM eventually fails to operate properly. The memory cell becomes unreadable or unwritable, soft-error increases, sense amplifier doesn’t produce valid logic levels, deviation in the timing of control signal increases. Therefore, circuit and system designers focus on novel solutions to design energy-efficient memories that are reliable for emerging power constraints applications.

2.2.4 Existing Solutions for Low Power Memory Design

With the rising demand for low-power electronics in emerging applications, there have been continuous efforts to develop low-power memory systems. The existing solution improving energy-efficiency of the memory includes assist circuits augmented with the peripherals circuits such as write drivers, sense-amplifier, supply sources to expand the operating range. Another approach includes alternative bitcell architecture that mitigates contention between the read and write stability of the conventional 6T SRAM cell. In addition to the new bitcell structure, new sensing techniques mitigate global and local process variation in the signal path. Furthermore, the data retention mode of SRAM and power on-off techniques drastically reduce the standby power consumption. These existing solutions are discussed in detail in the following sub-section.

2.2.4.1 Alternative Bitcell Design

Various alternative implementations of SRAM bitcell have been proposed in the literature to address the challenges with the conventional SRAM cell. An additional transistor in series with the pull-down NMOS transistor of the conventional 6T SRAM cell breaks the feedback path in the cross-coupled inverter, resulting in improved read and write stability of the cell [48]. Further, to decouple the read path entirely from the write path, additional transistors, and separate read bitline are utilized in the 8T SRAM cell [49]. The 6T core of the proposed bitcell can then be optimized for write operation resulting in improved read and write stability. Additionally, it also results in lower Vmin operation voltage for the SRAM, which can be used to enable near-threshold or sub-threshold operation for low power operation. Another SRAM cell design that works efficiently at sub-Vt voltages is the 8T SRAM cell proposed by Verma and Chandrakasan [58]. The proposed design uses a separate read buffer and write assist circuit to achieve proper read/write operation at low voltages. During the write operation, the wordline voltage is slightly boosted to increase the drivability of the access transistors while the horizontal power supply is reduced to weaken the pmos load transistor. As a result, the assist circuit guarantees the required greater strength of the access transistors than pmos load transistors for successful write operation. For power constraints applications that must retain the state during the extended period of idle mode (e.g., wireless sensor nodes for environmental monitoring), leakage is one of the major concerns. Therefore, operating at min supply voltage will require a solution like 8T SRAM operating at sub-Vt voltages. Furthermore, read buffers in the proposed design are driven high for the unselected rows, thereby reducing the bitline leakages. The major drawback of the design is the large area overhead due to the routing of three additional rails VSS, VWL, and VRWL. On the other hand, the 10T SRAM cell proposed by Calhoun and Chandrakasan [50] utilizes two additional transistors per bitcell to mitigate the BL leakages from the unassessed cells. However, none of the above-mentioned cells (7T, 8T, 10T) can be bit-interleaved because of the horizontally routed control signals and power signals. In order to implement bit interleaving architecture, some researchers propose a read-modify-write scheme for these alternative bit cell topologies. In addition, the single bit-line sensing used in these bitcell topologies suffers from reduced bit-line swing which increases the read failure. The fully differential 10T SRAM cell proposed by Chang et al. [51] improves the sensing margin by employing dynamic differential cascade voltage switch logic level read access path. Moreover, it enables bit interleaving by using column write control signals. The stacking of the access transistor results in performance degradation, thus requires boosted wordline voltages.

2.2.4.2 Assist Circuit

The SRAM assist circuits are utilized to achieve maximum energy efficiency for a given performance requirement of an application. A dynamic peripheral assist circuit proposed by Zhang et al. [52] changes the relative strength of the bitcell device depending on the mode of operation, resulting in improved margins and lower Vmin. The supply voltage is reduced below WL and BL for the write operation, whereas it is raised above WL and BL for the read operation. The alternative bitcell topologies with horizontal control signals face difficulty while implementing column bit interleaving. To realize bit interleaving architecture in these bitcell, a write-back scheme is proposed, in which the addressed row is read and buffered before performing the write operation. Hence, selected columns are overwritten with new values, whereas unselected columns are written using the buffered values. This scheme incurs an extra read operation for every write operation resulting in degraded performance and increased area overhead and power consumption. Some researchers use static biasing to enhance the performance of the SRAM. For example, wordline WL voltage can be reduced to achieve higher read stability. However, under-driven wordline degrades the write stability. Hence, Nho et al. [55] propose an adaptive wordline underdrive scheme in which wordline underdrive is optimized for a die based on read limited or write limited application. A sensor is employed for fine-grained wordline underdrive strength for each die to improve energy efficiency across a wide range of variations. Self-repairing SRAM implements body biasing technique to improve the read and write margin, but it does not address the Vth ratio of PMOS and NMOS device, necessary for correct SRAM operation. Addressing this issue, Yamaoka et al. [57] propose an SRAM with NMOS and PMOS device separately biased to achieve the required margin by maintaining the appropriate Vth ratio between NMOS and PMOS device. Alternatively, the supply voltage of the cell can be enhanced to improve the stability, however, writing into the cell with lower bitline voltage becomes very challenging. Further, Pilo et al. [54] propose a read redundancy scheme in which the sense amplifier is also connected to the bitline to write back the data to the cell with the original value. As a result, flipping of the cell during the read operation is avoided, which reduces the failure rate. Another approach of buffered bitline is utilized by Cosemans, Dehaene, and Catthoor [58], which utilizes short local bitlines to improve the read current and read performance. A major advantage of this architecture is efficient resource utilization as it doesn’t require a sense amplifier per column rather one set of sense amplifiers for the entire memory.

2.2.4.3 Sensing Innovations

The single-ended sensing required in most alternative bitcell topologies requires a change in traditional differential sense amplifier design. In addition, the single-ended sensing imposes a constraint on the weakest bitcell that it must overpower the leakage current by the strongest off-transistor. Moreover, the midpoint differentiating the two logic states can be determined by using a strobe signal, pseudo-differential sense amplifier, and dynamic conversion of BL to a static voltage level. However, at low voltages, the difference between the on and off bitline voltage becomes almost negligible thereby, determining a midpoint is very challenging. Several solutions have been proposed in the literature to address the challenges with single-ended sensing. A redundant sensing approach is proposed by Verma and Chandrakasan [59], where bitline in each column is connected to N different sense amplifier and using a selection logic circuit, only one sense amplifier which can correctly read the data is enabled. Another redundant sensing approach was proposed by Sinangil, Verma and Chandrakasan [60], where two individual sense amplifiers optimized for different voltage ranges are employed for ultra-dynamic voltage scaling systems. For high voltage read operation, the voltage at bitline is closer to Vdd and can be efficiently sensed with NMOS input. On the other hand, for low voltage read operation, the voltage at bitline drops to Vss, therefore, it can be efficiently sensed with a PMOS transistor. Alternatively, Cosemans, Deahene, and Catthoor [58] employ two sense amplifiers, out of which only one with the lowest offset is selected at the run time. Another approach for obtaining large margins at lower voltage supply is proposed by subthreshold SRAM design proposed by Kim et al. [61] which involves a replica circuit to track the optimum trip point for the read buffers. The read stack of the 10T SRAM cell is modified to obtain data-independent bitline leakage current by forcing leakage current to flow from a cell into the bitline irrespective of the data value. The logic low at the read bitline (RBL) is determined by the ratio of the leakage current by the unaccessed cell pulling up the RBL and leakage current by the accessed cell pulling down the RBL, whereas logic high is close to vdd since both the accessed cell and the unaccessed cell contribute to pulling up the bitline voltage to vdd. As a result of data independent leakage current on RBL, it is possible to automatically track the voltage generated through a replica column technique and determine an optimum trip voltage for the read buffers. Another class of sense amplifier involves an offset compensation method to improve the stability of the read operation. With advanced technology nodes and low voltage operation, the gap between the logic states on the BL is diminishing, hence there is an increased emphasis on variation-tolerant sense amplifier. One such variation tolerant design is ACSA (AC coupled sense amplifier) which controls the threshold voltage of the amplifying PMOS transistor connected to the BL, hence suppressing the variation at the output.

2.2.4.4 Data Retention Mode

The emerging low-power portable and wearable electronics devices in IoT network aim to reduce overall power consumption to meet the energy requirements of energy autonomous systems. For long idle periods, lowering the supply voltage helps in reducing the sub-threshold and gate leakage current. As a result, a significant portion of the standby power consumption is reduced. The supply voltage can be reduced to a limit of hold stability, also known as data retention voltage (DRV). However, DRV is highly impacted by process and temperature variation. The exiting solution includes using PMOS or NMOS diodes in series with the supply voltage to reduce the leakage current. Recent work by Pilo et al. [63] emphasized regulating and applying the fine-grained retention bias to the memory array to minimize the standby power consumption. Moreover, dynamic biasing of individual subarrays is done on a cycle-by-cycle basis. Another approach involves the prior prediction of DRV, which helps in aggressively reducing the standby power consumption. Takeyama et al. [64] utilizes the replica memory devices to determine the threshold voltage of the cell and apply twice the threshold voltage during standby mode for successful data retention.

2.3 Embedded Non-Volatile Memory Technology

The emerging IoT systems employ intelligent power on-off schemes to reduce standby power consumption. The increasing standby power consumption is an important issue when dealing with battery-powered or batteryless IoT such as low power wearable or implantable devices equipped with nanometer chips susceptible to large power consumption because of exponentially increasing leakage current at advanced technology node. Recent work employs on-chip non-volatile memory (eNVM) to store critical data during the idle mode while the system is powered off to reduce the standby power consumption [65]. For most IoT applications, a large capacity of eNVM is required to store the long code and large volume of data. After the device is powered ON, the program stored in eNVM is transferred to an on-chip instruction buffer while data is transferred to on-chip volatile memories (e.g., SRAM or eDRAM) [66]-[67]. During the normal operating mode, the on-chip volatile memories are used to provide access to the data required for computation, whereas eNVM is used for periodic or on-demand backup. Before the power down period, all the critical data is stored in the eNVM. To reduce the chip area, most low-cost IoT devices use eNVM to store data during the power down period and provide access to the program code during the power-on period. As a result, eNVM frequently perform read operation to access code during the normal operation and infrequently perform write operation to store the critical data before power off.

Thus, to reduce overall power consumption, the read operation must be optimized to consume minimum energy. The standby power consumption in a normally-off system constitutes two components: 1) Energy consumption during the backup operation performed before the power-off; 2) Energy consumption during the restore operation performed after the power on. The power off-on technique is beneficial compared to memory in data retention mode when the duration of the idle period exceeds the break-even time (BET). The break-even time is defined as the time duration for which the energy overhead due to backup and restore required for power off-on technique compensates the standby power consumption by the memory operated in data retention voltage. BET is smaller for the smaller technology nodes due to increasing SRAM leakages. The power on-off technique is more beneficial than the data retention mode of SRAM for applications with extended standby periods such as IoT.

Various existing eNVM technologies include embedded flash and one-time programmable (OTP) devices. At present, embedded flash (e-FLASH) is one of the most mature and reliable technology for embedded non-volatile memory (eNVM). The embedded flash has two type of structure: split gate and stack gate. These structures are based on a floating gate that stores the charge (either electrons or holes) to alter the threshold voltage of the device [68]. A compact 1T stack gate cell is most commonly utilized in commercially available NOR, and 2D NAND flash memory. For writing into the flash memory, two operations, program, and erase are performed. During the program operation, charges are stored in the floating gate resulting in increased threshold voltage and reduced cell current. The storage of the charge in the floating gate is achieved through channel hot electron (CHE) generated through source-drain current [68]. During the erase operation, charges are removed from the floating gate, resulting in lower threshold voltage and substantial current. A high voltage at the substrate initiates the Fowler-Nordheim tunneling mechanism between the floating gate and channel. The stored charges at the floating gate are then removed through the bottom tunneling oxide. The flash memory cell employs either a floating poly-gate or charge trapping technology [69]. However, 1T stack gate cell suffers from high program current and over-erase issues. A high program current requirement to store the charges in the floating gate results in large power consumption. In addition, charge pumps are required to create potential difference across source-drain to generate hot electrons, increasing the area overhead of the memory. The over-erase in the flash memory results in ultra-low threshold voltage, which in turn results in considerable bitline leakage and degraded sensing margin. To address this issue, program-verify schemes are employed. However, due to area constraints of eNVM, these techniques are precluded, which aggravate the issue of over-erase in eNVM. Addressing this issue, a split-gate memory cell was proposed [70]. In a conventional split-gate memory cell, the floating gate is split with one floating gate placed beneath the control gate to control the main floating gate in addition to the control gate. As a result, the split-gate flash memory cell acts as two transistors in series equivalent to 1.5T per cell. With a 1.5T split-gate, it is possible to turn off the unselected over-erased cell, thus eliminating the need for boosted wordline voltage to perform read operation. In recent split-gate memory cells, a large floating gate is placed beneath the control gate. A side-wall select gate or erase gate can also be employed within the flash memory cell to facilitate the inclusion of high voltage and low voltage areas within the spit gate [70]. In addition, several techniques have been introduced to reduce power consumption in split-gate flash cells, such as source-side injection (SSI) program operation, rapid erase using poly-to-poly FN tunneling, or hot hole injection using band-to-band tunneling [71]. The advantages offered by the split-gate flash cell, such as small program current, rapid erase, and immunity to over-erase, make it a very promising candidate for the on-chip/embedded applications, especially for low-power applications IoT. However, challenges with the flash memory are: 1) improving the performance of the flash memory by exploiting the low latency and high parallelism of the flash device [72], 2) lack of transactional support for better flexibility [73], 3) degraded reliability at the scaled technology node (beyond 20nm) [72], [73]. There has been tremendous research to provide solutions to overcome the reliability and endurance challenges of flash memory.

The flash memory suffers from four fundamental errors: retention errors, program interface errors, read errors and erase errors [74]. Furthermore, the impact of these errors significantly increases for the scaled technology nodes (< 20nm). From the experimental data, the relationship between the various errors is derived, and it is demonstrated that the retention errors are the dominant error affecting the reliability of the memory. Therefore, several techniques have been explored in the literature to address the issue of retention errors. One such technique is Flash Correct-and-Refresh (FCR), which refers to reading each page of the flash memory periodically and correcting the errors using the error-correcting codes (ECC) [75]. The correct values of each page are then either relocated to a different location or reprogrammed at the same location. More efficient error correction methods such as low-density parity-check (LDPC) codes can improve reliability in future flash memories [76]. Another challenge with the downscaling of the flash memory cell is increased cell-to-cell interference due to large parasitic capacitance between the floating cell. Recent work by Y.

Cia et al. [29] proposes a read retry mechanism to characterize and model the threshold voltage distribution of the NAND flash chip from 24-30nm technology nodes. The results are then used to obtain a cell-to-cell program interface under various programming conditions. Further, to improve the reliability and reduce the error in the flash memory, Neighbor-cell Assisted Correction (NAC) mechanism has been proposed [25]. It is based on the empirical observation that identifying the value of immediate neighbor helps determines the correct value of cell currently being read. Therefore, the value of the neighboring cell is used in NAC to correct the errors in the current page. The key idea is to re-read the flash memory page with reference to threshold voltage distribution assuming a cell neighbor value and then correct the cell with that neighbor value. More accurate and detailed characterization of errors mechanism is required to effectively understand and develop error-tolerant solutions for sub-20 nm flash memories. A promising direction is developing a predictive model that can predict the error and take preventive measures to avoid it. Flash-correct-refresh [77], read reference voltage prediction [78], and retention optimized reading mechanisms [29] are some of the predictive techniques found in the literature that are used to increase error tolerance in sub-20nm flash memory. Another promising direction involves exploiting the application and memory access characteristic to optimize the performance, life, and cost of the flash memory. Furthermore, techniques such as data-characteristic-aware management mechanisms for flash will likely aid in scaling flash memory technology into the future.

2.4 Logic Circuit in Modern Computing System

With technology dimensions reaching the atomic-scale limit, achieving energy-efficient computing has become one of the biggest concerns for the research community. Koomey's law state that the computation per kilowatt-hour must double every 1.57 years [79]. However, today it takes almost three years for peak efficiency to double [80]. This slowdown is due to the exponentially increasing amount of data to be processed. The emerging IoT applications generates enormous amount of data which creates new challenges for the systems to achieve energy-efficient computing. The big data is identified by the '3-V model" which is described as follows: 1) Volume: It is the amount of bulk data generated with high information potential 2) Velocity: It is the frequency by which information changes in an application. 3) Variety: It is an aggregation of data from heterogenous sources [81]. The huge amount of data collected is utilized to provide support for new service platforms to serve humanity. It is predicted that by the year 2020, the IoT network will generate 40 Zettabytes of data per year [80]. The zettabyte era requires infrastructure to analyze the data in real-time and create great opportunities for new applications. One of the biggest challenges faced by the big data era of IoT is energy-efficient computing. Moreover, there exists a theoretical bound on the minimum energy consumption per computation. According to Landauer's principle, a logical irreversible manipulation of data requires atleast E_{min} energy dissipation for each bit of data lost [82]. The energy dissipated while completing a logic operation is defined by k.T.ln2, where k is the Boltzmann's constant and T is the temperature. However, logic circuits employed in current computing systems dissipate energy consumption which is three orders of magnitude higher than the fundamental limit. Therefore, new computing paradigms are needed to address this challenge of large computational power. In the following sub-section, we discuss exiting low-power computing approaches suitable for IoT applications.

2.4.1 Reversible Logic

According to Landauer's theory, losing information during data manipulation results in energy consumption [82]-[95]. Therefore, to reduce energy consumption, data should not be lost or erased. In recent time, reversible logic operations have attracted a lot of attention as it does not require any erasing operation, resulting in lower power consumption [85]. In addition, adiabatic circuits are used to improve hardware security [97]-[106]. The adiabatic circuits based on reversible logic utilizes two approaches to reduce the power consumption: 1) constant current source for charging output capacitor 2) power supply capable of recovering the charge stored in the capacitor [95]-[96]. The constant current source is realized through a ramp voltage source which minimizes the potential difference between the source and drain, thereby reducing the power consumption during switching. The recent work on the adiabatic circuit can be broadly divided into two categories: *Fully-Adiabatic logic family* with nearly zero power consumption and *Quasi-Adiabatic logic family* with non-zero power consumption. In fully-adiabatic logic circuit, theoretically, the energy consumption must be zero. However, due to the non-zero threshold voltage of the transistor, energy consumption is not zero. S. G Younis et al. [92] proposes a fully adiabatic buffer with CMOS logic and transmission gate pipeline. The buffer design is easy to implement and has a small area overhead. Another buffer design proposed by Jeanniot et al. [105] is based on dual rail circuit that takes complementary inputs and produces complementary output. The dual rail circuit mitigate the problem of energy recovery while cascading the adiabatic circuits by using a separate recovery path. The area overhead of the proposed design is large due to the increased number of transistors. Another dual rail buffer proposed by Kramer et al. [108] uses a cross-coupled transmission gate instead of conventional CMOS logic. The advantage of the proposed design is that it can be fully pipelined. However, using a cross-coupled transmission gate increases the area overhead. On the other hand, in quasi-adiabatic logic, the total energy consumption is not zero. Moon et al. [109] propose dual rail circuits with PMOS in the pull-up network to pull the output to vdd, and NMOS is pull-down network to build the logic. The energy consumption occurs because the energy recovery path is cut off from the source. Whereas, Blotti et al. [107] propose a dual-rail circuit that utilizes PMOS transistors to build logic, effectively reducing the output capacitance, resulting in lower energy consumption.

2.4.2 Near-Threshold Logic

Near-Threshold Computing (NTC) is a computing paradigm in which logic circuits are operated close to the transistor threshold voltage [110]-[111]. Since the dynamic power consumption has a quadratic relationship with the power supply, significant power saving (upto 10x) is achieved by reducing the power supply to near-threshold voltages. The lower power consumption offered by NTC makes it attractive for power constraint IoT applications. However, reducing the power supply severely impacts the reliability and performance of the system due to increased process variations. Moreover, designing a robust circuit that is immune to the process, voltage, and temperature (PVT) variations at such low voltages is extremely challenging. Several techniques have been proposed to address the degraded reliability issue of the circuits operating in near-threshold voltages. One such technique involves using PVT compensation circuits augmented with the main circuit to increase its reliability. These circuits have a feedback path to adjust the power supply to prevent any meta-stabilities. Some examples of the compensation circuits are canary circuits and razor flip-flops. The canary circuits are the replication of critical paths designed to adjust the power supply [112]-[113]. However, canary circuits can only track the variation in the global process. The information about the local variations can not be obtained using canary circuits since these replica circuits are placed at a different location than actual path. Another compensation circuit is proposed to address this issue and track both the local and global variation, also known as razor flip-flop circuit [114]. The razor flip-flop lowers the power supply to a critical point. The disadvantage of operating at such low power supply is increased error rate due to timing violations [115]. To address this issue, the razor circuit restores the correct value by using a shadow flip-flop. The shadow flip-flops are operated using a delayed clock signal; therefore, they preserve the correct values and restore them to the real datapath. An alternative compensative technique that is very effective for circuit operating in near-threshold voltages involves body biasing. However, modifying the body biasing of the transistor only impacts the leakage power consumption, whereas the approaches mentioned above impact both the leakage and dynamic power consumption.

2.5 Summary

IoT devices are implemented on various technology nodes to accommodate a wide range of applications and cost structures. The stringent requirement of energy consumption in most IoT applications, present great challenges, and opportunities for the advances in memory and computing block to enable their usage in IoT network. Energy-efficient logic combined with low-power memory significantly improve the performance of batteryless IoT applications. As the power budgets of IoT devices are extremely small, embedded memories are required to operate in low power mode during the access (read and write) and standby periods. However, the traditional approaches such as ultra-dynamic voltage scaling to address the rising power consumption have slowed down with the increasing scaling challenges. Hence advanced circuit techniques are needed for the energy-efficient and reliable operation of devices in IoT domains. At the same time, the new and emerging workload of data-intensive IoT applications demands increasing memory capacity and bandwidth. Both the research and academia community focus on providing novel solutions to design a low-power memory design. Several circuit level techniques have been proposed which modifies either the peripheral circuit or the bitcell structure to address the large power consumption while minimizing area overhead or performance penalty. Additionally, with the development of new bitcell structures, assist circuits, and sensing techniques, CAD methodologies must also be improved to accurately characterize random fluctuation present from die to die and within the die. The static random-access memory continues to play a critical role in a system design because of its dominant on-chip area and its always-on power supply needed to retain the data. The state-of-the-art embedded flash (eFlash) is a non-volatile technology that can be utilized to retain the data during the idle period while volatile counterparts such as register and cache memory are powered off to reduce the critical standby power consumption. However, the requirement of a large programming current for storing the charge in the floating gate results in significant power consumption. Addressing these challenges, emerging non-volatile devices emerge as the most promising candidates for low-power IoT applications. They have lower write energy compared to eFlash memories due to lower write voltage and faster write time. At the same time, it is becoming extremely challenging to integrate conventional floating point-based e-flash memory cells with the advanced logic processes. This has led to the need for on-chip NVM solutions that are logic process compatible. Hence, the emerging resistive non-volatile technology is one of the most promising solutions for eNVM in the nanometer process. Furthermore, logic circuits must be optimized for minimum energy points for obtaining a low power system. New computing paradigms such as near-threshold computing significantly lowers the computational power. However, to realize NTC and achieve correct functionality at low voltages, additional transistors are incorporated with the main circuit. An alternative approach for low power computations is adiabatic circuits based on reversible logic, which significantly reduces the switching energy. In the end, it can be concluded that reducing the overall power consumption of the system requires a balance between energy-efficient logic and memory.

REFERENCES

[1] G. Moore, “Cramming More Components onto Integrated Circuits," *Proceedings of the IEEE*, vol. 86, no. 1, pp. 82-85, Jan 1998.

[2] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the Gap to Human-Level Performance in Face Verification," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1701-1708.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems* vol. 25, 2012, pp. 1097-1105.

[4] G. Hinton et al., “Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82-97, Nov 2012.

[5] R. Sarikaya, G. E. Hinton, and A. Deoras, “Application of Deep Belief Networks for Natural Language Understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 778-784, April 2014.

[6] Y. K. Ramadass and A. P. Chandrakasan, “A battery-less thermoelectric energy harvesting interface circuit with 35 mv startup voltage," *IEEE Journal of Solid- State Circuits*, vol. 46, no. 1, pp. 333-341, Jan 2011.

[7] H. Yamauchi, “Embedded SRAM Design in Nanometer-Scale Technologies," in *Embedded Memories for Nano-Scale VLSIs*, K. Zhang, Ed. New York: Springer, 2009.

[8] C. Lefurgy, K. Rajamani, F. Rawson, W. Felter, M. Kistler and T. W. Keller, "Energy management for commercial servers," in *Compute*r, vol. 36, no. 12, pp. 39-48, Dec. 2003, doi: 10.1109/MC.2003.1250880.

[9] E. S. Chung, P. A. Milder, J. C. Hoe and K. Mai, "Single-Chip Heterogeneous Computing: Does the Future Include Custom Logic, FPGAs, and GPGPUs?," *2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture,* 2010, pp. 225-236, doi: 10.1109/MICRO.2010.36.

[10] K. Van Craeynest, A. Jaleel, L. Eeckhout, P. Narvaez and J. Emer, "Scheduling heterogeneous multi-cores through performance impact estimation (PIE)," *2012 39th*

Annual International Symposium on Computer Architecture (ISCA), 2012, pp. 213-224, doi: 10.1109/ISCA.2012.6237019.

[11] R. Ausavarungnirun, K. K. Chang, L. Subramanian, G. H. Loh and O. Mutlu, "Staged memory scheduling: Achieving high performance and scalability in heterogeneous systems," *2012 39th Annual International Symposium on Computer Architecture (ISCA),* 2012, pp. 416-427, doi: 10.1109/ISCA.2012.6237036.

[12] S. Eyerman and L. Eeckhout, "Modeling critical sections in amdahl's law and its implications for multicore design," *ISCA '10: The 37th Annual International Symposium on Computer Architecture,* 2010, pp. 362–370.

[13] J. A. Joao et al., "Bottleneck identification and scheduling in multithreaded applications," in *ASPLOS, 2012,* doi: https://doi.org/10.1145/2248487.2151001.

[14] J. A. Joao et al., "Utility-based acceleration of multithreaded applications on asymmetric CMPs," in *ISCA*, 2013, doi: https://doi.org/10.1145/2508148.2485936.

[15] M. A. Suleman et al., "Accelerating critical section execution with asymmetric multi-core architectures," in *ASPLOS, 2009,* doi: https://doi.org/10.1145/1508244.1508274.

[16] M. A. Suleman et al., "Accelerating critical section execution with asymmetric multi-core architectures," *IEEE Micro* vol. 30, no. 1, 2010.

[17] L. Subramanian et al., "MISE: Providing performance predictability and improving fairness in shared main memory systems," in *HPCA*, 2013, pp. 639-650, doi: 10.1109/HPCA.2013.6522356.

[18] T. Moscibroda and O. Mutlu, "Memory performance attacks: Denial of memory service in multi-core systems," in *USENIX Security*, 2007, pp. 1-18.

[19] O. Mutlu and T. Moscibroda, "Stall-time fair memory access scheduling for chip multiprocessors," in *MICRO*, 2007, pp. 146-160, doi: 10.1109/MICRO.2007.21.

[20] R. Bryant, "Data-intensive supercomputing: The case for DISC," *CMU CS Tech. Report* 07-128, 2007, doi: https://doi.org/10.1184/R1/6604628.v1.

[21] C. Alkan et al., "Personalized copy-number and segmental duplication maps using nextgeneration sequencing," in *Nature Genetics*, 2009.

[22] D. Lee et al., "Fast and accurate mapping of complete genomics reads," in *Methods*, 2014.

[23] T. Treangen and S. Salzberg, "Repetitive DNA and next-generation sequencing: computational challenges and solutions," in *Nature Reviews Genetics*, 2012.

[24] H. Xin et al., "Accelerating read mapping with Fast HASH," in *BMC Genomics*, 2013.

[25] Hongyi Xin, et al., "Shifted Hamming distance: a fast and accurate SIMD-friendly filter to accelerate alignment verification in read mapping" *Bioinformatics*, Volume 31, Issue 10, 2015, pp. 1553–1560.

[26] L. Tang et al., "The impact of memory subsystem resource sharing on datacenter applications," in *ISCA*, 2011, doi: https://doi.org/10.1145/2000064.2000099.

[27] B. C. Lee et al., "Architecting phase change memory as a scalable DRAM alternative," in *ISCA*, 2009, doi: https://doi.org/10.1145/1555754.1555758.

[28] "International technology roadmap for semiconductors (ITRS)," 2011.

[29] Y. Cai, E. F. Haratsch, O. Mutlu and K. Mai, "Error patterns in MLC NAND flash memory: Measurement, characterization, and analysis," 2012 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2012, pp. 521-526.

[30] Y. Cai et al., "Flash correct-and-refresh: Retention-aware error management for increased flash memory lifetime," 2012 IEEE 30th International Conference on Computer Design (ICCD), 2012, pp. 94-101.

[31] U. Kang et al., "Co-architecting controllers and DRAM to enhance DRAM process scaling," in *The Memory Forum*, 2014. [Online] Available: https://www.cs.utah.edu/ thememoryforum/kang.pdf

[32] B. C. Lee et al., "Architecting phase change memory as a scalable DRAM alternative," in *Proceedings of the 36th annual international symposium on Computer architecture, ISCA*, 2009.

[33] B. C. Lee et al., "Phase change memory architecture and the quest for scalability," *Communications of the ACM*, vol. 53, no. 7, 2010.

[34] H. -. P. Wong et al., "Metal–Oxide RRAM," in *Proceedings of the IEEE*, vol. 100, no. 6, pp. 1951-1970, June 2012, doi: 10.1109/JPROC.2012.2190369.

[35] T. Moscibroda and O. Mutlu, "Memory performance attacks: Denial of memory service in multi-core systems," in *Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium*, 2007.

[36] S. Hong, "Memory technology trend and future challenges," *2010 International Electron Devices Meeting*, 2010, pp. 12.4.1-12.4.4, doi: 10.1109/IEDM.2010.5703348.

[37] A. Jog et al., "OWL: Cooperative thread array aware scheduling techniques for improving GPGPU performance," in *ASPLOS*, 2013, doi: https://doi.org/10.1145/2451116.2451158.

[38] T. L. Johnson, M. C. Merten and W. W. Hwu, "Run-time spatial locality detection and optimization," *Proceedings of 30th Annual International Symposium on Microarchitecture*, 1997, pp. 57-64, doi: 10.1109/MICRO.1997.645797.

[39] R. Iyer, "CQoS: a framework for enabling QoS in shared caches of CMP platforms," in *ICS, 2004*, doi: https://doi.org/10.1145/1006209.1006246.

[40] R. Iyer et al., "QoS policies and architecture for cache/memory in CMP platforms," in *SIGMETRICS*, 2007, doi: https://doi.org/10.1145/1269899.1254886

[41] K. Lim et al., "Disaggregated memory for expansion and sharing in blade servers," in *ISCA*, 2009, doi: https://doi.org/10.1145/1555754.1555789.

[42] T. Moscibroda and O. Mutlu, "Memory performance attacks: Denial of memory service in multi-core systems," *in USENIX Security*, 2007.

[43] N.A. Kurd et al., ''Westmere: A Family of 32nm IA Processors,'' *Proc. IEEE Int'l Solid-State Circuits Conf. (ISSCC 10)*, 2010, pp. 96-97

[44] K.J. Kuhn, ''Reducing Variation in Advanced Logic Technologies: Approaches to Process and Design for Manufacturability of Nanoscale CMOS,'' *Proc. IEEE Int'l Electron Devices Meeting (IEDM 07)*, IEEE Press, 2007, pp. 471-47.

[45] T. Karnik et al., ''Scaling Trends of Cosmic Ray Induced Soft Errors in Static Latches beyond 0.18m,'' *Proc. Symp. VLSI Circuits*, 2001, pp. 61-62.

[46] T. Suzuki et al., ''A Sub-0.5-V Operating Embedded SRAM Featuring a Multi-bit-Error-Immune Hidden-ECC Scheme,'' *IEEE J. Solid-State Circuits,* vol. 41, no. 1, pp. 152-160.12, 2006.

[47] K. Osada et al., ''Universal-Vdd 0.65-2.0-V 32-kB Cache Using a Voltage-Adapted Timing-Generation Scheme and a Lithographically Symmetrical Cell,'' *IEEE J. Solid-State Circuits,* vol. 36, no. 11, pp. 1738-1744, 2011.

[48] K. Takeda et al., ''A Read-Static-Noise-Margin-Free SRAM Cell for Low-VDD and High-Speed Applications,'' *IEEE J. Solid-State Circuits,* vol. 41, no. 1, pp. 113-121, 2006.

[49] L. Chang et al., ''Stable SRAM Cell Design for the 32 nm Node and Beyond,'' P*roc. Symp. VLSI Tech.,* IEEE Press, pp. 128-129, 2005.

[50] B.H. Calhoun and A.P. Chandrakasan, ''A 256-kb 65-nmSub-threshold SRAM Design for Ultra-Low-Voltage Operation,'' *IEEE J. Solid-State Circuits,* vol. 42, no. 3, pp. 680-688, 2007.

[51] I.J. Chang et al., ''A 32 kb 10T Sub-threshold SRAM Array with Bit-Interleaving and Differential Read Scheme in 90 nm CMOS,'' *IEEE J. Solid-State Circuits,* vol. 44, no. 2, pp. 650-658, 2009.

[52] K. Zhang et al., ''A 3-GHz 70-mb SRAM in 65-nm CMOS Technology with Integrated Column-Based Dynamic Power Supply,'' *IEEE J. Solid-State Circuits,* vol. 41, no. 1, pp. 146-151, 2006.

[53] S. Ohbayashi et al., ''A 65-nm SoC Embedded 6T-SRAM Designed for Manufacturability with Read and Write Operation Stabilizing Circuits,'' *IEEE J. Solid-State Circuits,* vol. 42, no. 4, pp. 820-829, 2007.

[54] H. Pilo et al., ''An SRAM Design in 65-nm Technology Node Featuring Read and Write-Assist Circuits to Expand Operating Voltage,'' *IEEE J. Solid-State Circuits,* vol. 42, no. 4, pp. 813-819, 2007.

[55] H. Nho et al., ''A 32nm High-k Metal Gate SRAM with Adaptive Dynamic Stability Enhancement for Low-Voltage Operation,'' *Proc. IEEE Int'l Solid-State Circuits Conf. (ISSCC 10),* IEEE Press, pp. 346-347, 2010.

[56] J. Pille et al., ''Implementation of the Cell Broadband Engine in 65 nm SOI Technology Featuring Dual Power Supply SRAM Arrays Supporting 6 GHz at 1.3 V,'' *IEEEJ. Solid-State Circuits,* vol. 43, no. 1, pp. 163-171, 2008.

[57] M. Yamaoka et al., ''65nm Low-Power High-Density SRAM Operable at 1.0V under 3sSystematic Variation Using Separate Vth Monitoring and Body Bias for NMOS and PMOS,'' *Proc. IEEE Int'l Solid-State Circuits Conf. (ISSCC 08),* IEEE Press, pp. 384-385, 622, 2008.

[58] S. Cosemans, W. Dehaene, and F. Catthoor, ''A 3.6 pJ/Access 480 MHz, 128 kb On-Chip SRAM with 850 MHz Boost Mode in 90 nm CMOS with Tunable Sense Amplifiers,'' *IEEE J. Solid-State Circuits,* vol. 44, no. 7, pp. 2065-2077, 2009.

[59] N. Verma and A.P. Chandrakasan, ''A 256 kb 65 nm 8T Subthreshold SRAM Employing Sense-Amplifier Redundancy,'' *IEEE J. Solid-State Circuits,* vol. 43, no. 1, pp. 141-149, 2008.

[60] M.E. Sinangil, N. Verma, and A.P. Chandrakasan, ‘‘A Reconfigurable 8T Ultra-Dynamic Voltage Scalable (U-DVS) SRAM in 65 nm CMOS,’’ *IEEEJ. Solid-State Circuits*, vol. 44, no. 11, pp. 3163-3173, 2009.

[61] T.-H. Kim et al., ‘‘A 0.2 V, 480 kb Subthreshold SRAM with 1k Cells per Bitline for Ultra-Low-Voltage Computing,’’ *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 518-529, 2008.

[62] M. Qazi et al., ‘‘A 512kb 8T SRAM Macro Operating down to 0.57V with an AC-Coupled Sense Amplifier and Embedded Data-Retention-Voltage Sensor in 45nm SOI-CMOS,’’ *Proc. IEEE Int'l Solid-State Circuits Conf. (ISSCC 10*), 2010, pp. 350-351.

[63] H. Pilo et al., ‘‘A 450ps Access-Time SRAM Macro in 45nm SOI Featuring a Two-Stage Sensing-Scheme and Dynamic Power Management,’’ *Proc. IEEE Int'l Solid-State Circuits Conf. (ISSCC 08),* 2008, pp. 378-379.

[64] Y. Takeyama et al., ‘‘A Low Leakage SRAM Macro with Replica Cell Biasing Scheme,’’ *Proc. IEEE Symp. VLSI Circuits*, IEEE Press, 2005, pp. 166-167

[65] J. Li, R.K. Montoye, M. Ishii, L. Chang, “1 Mb 0.41 µm2 2T-2R cell nonvolatile TCAM with two-bit encoding and clocked self-referenced sensing” *IEEE J. Solid State Circuits* Volume 49, pp. 896–907, 2014.

[66] F. Frustaci, M. Khayatzadeh, D. Blaauw, D. Sylvester, M. Alioto, “A 32 kb SRAM for error-free and error tolerant applications with dynamic energy-quality management in 28 nm CMOS” *IEEE International Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers*, 2014, pp. 244–245.

[67] Y.-H. Chen, W.-M. Chan, W.-C. Wu, et al., “A 16 nm 128 Mb SRAM in high-κ metal-gate FinFET technology with write-assist circuitry for low-VMIN applications” in *IEEE International Solid-State Circuits Conference (ISSCC) Dig. Tech. Papers,* 2014, pp. 238–239.

[68] K. Zhang, “Embedded memories for nano-scale VLSIs” Springer, New York, 2009.

[69] T. Kono, T. Ito, T. Tsuruda, T. Nishiyama, T. Nagasawa, T. Ogawa, Y. Kawashima, H. Hidaka, T. Yamauchi, “40-nm embedded split-gate MONOS (SG-MONOS) flash macros for automotive with 160-MHz random access for code and endurance over 10 M cycles for data at the junction temperature of 170” *IEEE J. Solid State Circuits,* Volume 49, pp. 154–166, 2013.

[70] B. Eitan, P. Pavan, I. Bloom, E. Aloni, A. Frommer,D. Finzi, “NROM: a novel localized trapping, 2-bit nonvolatile memory cell” *IEEE Electron Device Lett.*, Volume 21, Issue 11, pp. 543–545, 2000.

[71] V. Vasudevan et al., “Using vector interfaces to deliver millions of IOPS from a networked key-value storage server,” *SoCC, 2012*, doi: https://doi.org/10.1145/2391229.2391237.

[72] Y. Lu et al., “LightTx: A lightweight transactional design in flash-based SSDs to support flexible transactions,” *ICCD, 2013*, pp. 115-122, doi: 10.1109/ICCD.2013.6657033.

[73] Y. Cai et al., “Error patterns in MLC NAND flash memory: Measurement, characterization, and analysis,” in *DATE 2012*, pp. 521-526, doi: 10.1109/DATE.2012.6176524.

[74] Y. Cai et al., “Flash Correct-and-Refresh: Retention-aware error management for increased flash memory lifetime,” in *ICCD 2012,* pp. 94-101, doi: 10.1109/ICCD.2012.6378623.

[75] Y. Cai et al., “Program interference in MLC NAND flash memory: Characterization, modeling, and mitigation,” in *ICCD*, 2013, pp. 123-130, doi: 10.1109/ICCD. 2013.6657034.

[76] A. Bhavnagarwala et al., ‘‘Fluctuation Limits & Scaling Opportunities for CMOS SRAM Cells,’’ *Proc. IEEE Int’l Electron Devices Meeting (IEDM 05),* IEEE Press, 2005, pp. 659-662.

[77] J. Pille et al., ‘‘Implementation of the Cell Broadband Engine in 65 nm SOI Technology Featuring Dual Power Supply SRAM Arrays Supporting 6 GHz at 1.3 V,’’ *IEEEJ. Solid-State Circuits,* vol. 43, no. 1, 2008, pp. 163-171.

[78] T.-H. Kim et al., ‘‘A 0.2 V, 480 kb Subthreshold SRAM with 1k Cells per Bitline for Ultra-Low-Voltage Computing,’’ *IEEE J. Solid-State Circuits,* vol. 43, no. 2, 2008, pp. 518-529.

[79] J. Koomey, S. Berard, M. Sanchez and H. Wong, "Implications of Historical Trends in the Electrical Efficiency of Computing," in *IEEE Annals of the History of Computing*, vol. 33, no. 3, pp. 46-54, March 2011, doi: 10.1109/MAHC.2010.28.

[80] A. M. Ionescu, "Energy efficient computing and sensing in the Zettabyte era: From silicon to the cloud,*" 2017 IEEE International Electron Devices Meeting (IEDM),* 2017, pp. 1.2.1-1.2.8, doi: 10.1109/IEDM.2017.8268307.

[81] R. Bergelt, M. Vodel and W. Hardt, "Energy efficient handling of big data in embedded, wireless sensor networks," *2014 IEEE Sensors Applications Symposium (SAS),* 2014, pp. 53-58, doi: 10.1109/SAS.2014.6798916.

[82] J. Liu, M. B. Clavel and M. K. Hudait, "TBAL: Tunnel FET-Based Adiabatic Logic for Energy-Efficient, Ultra-Low Voltage IoT Applications," in *IEEE Journal of the Electron Devices Society,* vol. 7, pp. 210-218, 2019, doi: 10.1109/JEDS.2019.2891204

[83] N. S. Kim et al., "Leakage current: Moore's law meets static power," in *Computer,* vol. 36, no. 12, pp. 68-75, Dec. 2003, doi: 10.1109/MC.2003.1250885.

[84] E. P. DeBenedictis, J. K. Mee and M. P. Frank, "The Opportunities and Controversies of Reversible Computing," in Computer, vol. 50, no. 6, pp. 76-80, 2017.

[85] M.P. Frank, Foundations of generalized reversible computing, in: International Conference on Reversible Computation, Springer, 2017, pp. 19–34.

[86] A. P. Chandrakasan, S. Sheng and R. W. Brodersen, "Low-power CMOS digital design," in IEEE Journal of Solid-State Circuits, vol. 27, no. 4, pp. 473-484, April 1992, doi: 10.1109/4.126534.

[87] V. Anantharam, M. He, K. Natarajan, H. Xie, M.P. Frank, “Driving fully-adiabatic logic circuits using custom high-q mems resonators” *ESA/VLSI,* 2004, pp. 5–11.

[88] W.C. Athas, L.J. Svensson, J.G. Koller, N. Tzartzanis, E.Y.-C. Chou, “Low-power digital systems based on adiabatic-switching principles”, *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.,* Volume 2, no. 4, 398–407, 1994.

[89] J. Fischer, E. Amirante, A. Bargagli-Stoffi, D. Schmitt-Landsiedel, “Improving the positive feedback adiabatic logic family”, *Adv. Radio. Sci.*, Vol. 2, pp. 221–225, 2005.

[90] D. Maksimovic, V.G. Oklobdzija, B. Nikolic, K.W., “Current, Clocked cmos adiabatic logic with integrated single-phase power-clock supply”, *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* vol. 8, no. 4, pp. 460–463, 2000.

[91] H H. Mahmoodi-Meimand and A. Afzali-Kusha, "Low-power, low-noise adder design with pass-transistor adiabatic logic," *ICM 2000. Proceedings of the 12th International Conference on Microelectronics*, 2000, pp. 61-64.

[92] S.G. Younis, “Asymptotically Zero Energy Computing Using Split-Level Charge Recovery logic”, Tech. Rep., *Massachusetts Inst of Tech* Cambridge Artificial Intelligence Lab, 1994.

[93] R. Landauer, "Irreversibility and Heat Generation in the Computing Process," in *IBM Journal of Research and Development,* vol. 5, no. 3, pp. 183-191, July 1961, doi: 10.1147/rd.53.0183.

[94] T. Indermaur and M. Horowitz, "Evaluation of charge recovery circuits and adiabatic switching for low power CMOS design," *Proceedings of 1994 IEEE Symposium on Low Power Electronics,* 1994, pp. 102-103.

[95] I. Hnninen, H. Lu, E.P. Blair, C.S. Lent, G.L. Snider, “Reversible and adiabatic computing: energy-efficiency maximized”, *Field-Coupled Nanocomputing*, Springer, 2014, pp. 341–356.

[96] A. Galisultanov, Y. Perrin, H. Samaali, H. Fanet, P. Basset, G. Pillonnet, Contactless four-terminal mems variable capacitor for capacitive adiabatic logic” *Smart Mater. Struct.*, vol. 27, no. 8, 2018.

[97] Y. Perrin, A. Galisultanov, H. Samaali, P. Basset, H. Fanet, G. Pillonnet, “Contactless capacitive adiabatic logic”, *Nanoengineering: Fabrication, Properties, Optics, and Devices XIV, vol. 10354, International Society for Optics and Photonics*, 2017.

[98] Ayrat Galisultanov, Yann Perrin, Hervé Fanet, Gaël Pillonnet, “Capacitive-Based Adiabatic Logic Reversible Computation”, *9th International Conference, RC 2017*, Kolkata, India, July 6-7, 2017,

[99] S. Houri, G. Billiot, M. Belleville, A. Valentian, H. Fanet, “Limits of cmos technology and interest of nems relays for adiabatic logic applications”, *IEEE Trans. Circuit Syst. I*, vol. 62, no. 6, pp. 1546–1554, 2015.

[100] S S. D. Kumar, H. Thapliyal and A. Mohammad, "EE-SPFAL: A Novel Energy-Efficient Secure Positive Feedback Adiabatic Logic for DPA Resistant RFID and Smart Card," in *IEEE Transactions on Emerging Topics in Computing,* vol. 7, no. 2, pp. 281-293, 2019.

[101] S. D. Kumar, H. Thapliyal and A. Mohammad, "FinSAL: A novel FinFET based Secure Adiabatic Logic for energy-efficient and DPA resistant IoT devices," *2016 IEEE International Conference on Rebooting Computing (ICRC), 2016*, pp. 1-8.

[102] M.P. Frank, “Introduction to reversible computing: motivation, progress, and challenges”, in: *Proceedings of the 2nd Conference on Computing Frontiers, ACM*, 2005, pp. 385–390.

[103] J. Lim, D.-G. Kim, S.-I. Chae, “Reversible energy recovery logic circuits and its 8-phase clocked power generator for ultra-low-power applications”, *IEICE Trans. Electron.*, vol. 82, no. 4, 646–653, 1999.

[104] N. Jeanniot, A. Todri-Sanial, P. Nouet, G. Pillonnet, H. Fanet, “Investigation of the power-clock network impact on adiabatic logic”, *2016 IEEE 20th Workshop on Signal and Power Integrity (SPI), IEEE*, 2016, pp. 1–4.

[105] N. Jeanniot, G. Pillonnet, P. Nouet, N. Azemard, A. Todri-Sanial, “Synchronised 4-phase resonant power clock supply for energy efficient adiabatic logic”, *2017 IEEE International Conference on Rebooting Computing (ICRC), IEEE*, 2017, pp.1–6.

[106] M. P. Frank, "The physical limits of computing," in Computing in Science & Engineering, vol. 4, no. 3, pp. 16-26, May-June 2002, doi: 10.1109/5992.998637.

[107] A. Blotti, S. Di Pascoli, R. Saletti, “Simple model for positive-feedback adiabatic logic power consumption estimation”, *Electron. Lett.*, Volume 36, Issue 2, p. 116 – 118, 2000.

[108] A. Kramer, J.S. Denker, B. Flower, J. Moroney, “2nd order adiabatic computation with 2n-2p and 2n-2n2p logic circuits”, *Proceedings of the 1995 International Symposium on Low Power Design, ACM*, 1995, pp. 191–196.

[109] Y. Moon, D.-K. Jeong, “An efficient charge recovery logic circuit”, *IEEE J. Solid State Circ.*, vol. 31, no. 4, pp. 514–522, 1996.

[110] M. Capra, R. Peloso, G. Masera, M. Ruo Roch, M. Martina, “Edge Computing: A Survey On the Hardware Requirements in the Internet of Things World” *Future Internet*, Volume 11, Issue 4., 2019.

[111] B. Chen, P. Chou, Y. Fang, L. Yong, T. Lin and J. Wang, "Design of ultra-low-leakage near-threshold dynamic circuits in nano CMOS for IoT applications," *2016 IEEE 16th International Conference on Nanotechnology (IEEE-NANO)*, 2016, pp. 537-540, doi: 10.1109/NANO.2016.7751533.

[112] B. H. Calhoun and A. P. Chandrakasan, "Standby power reduction using dynamic voltage scaling and canary flip-flop structures," in *IEEE Journal of Solid-State Circuits*, vol. 39, no. 9, pp. 1504-1511, Sept. 2004, doi: 10.1109/JSSC.2004.831432

[113] J. Wang and B. H. Calhoun, "Techniques to Extend Canary-Based Standby V_DD Scaling for SRAMs to 45 nm and Beyond," in *IEEE Journal of Solid-State Circuits*, vol. 43, no. 11, pp. 2514-2523, Nov. 2008, doi: 10.1109/JSSC.2008.2005814.

[114] D. Ernst et al., "Razor: circuit-level correction of timing errors for low-power operation," in *IEEE Micro*, vol. 24, no. 6, pp. 10-20, Nov.-Dec. 2004, doi: 10.1109/MM.2004.85

[115] D. Ernst et al., "Razor: a low-power pipeline based on circuit-level timing speculation," Proceedings. *36th Annual IEEE/ACM International Symposium on Microarchitecture, 2003. MICRO-36.*, 2003, pp. 7-18, doi: 10.1109/MICRO.2003.1253179.