CHAPTER 8

CONCLUSION AND FUTURE SCOPE

8.1 Conclusion

This thesis aims to propose novel circuits based on emerging technologies and various non-volatile designs that can support the new generation of harvesting platforms and autonomous nodes that contribute to the revolution called the Internet of Things. In this thesis, we have chosen to explore the design space of memory cells in terms of data retention during power off/standby mode, power consumption, delay, and power-delay product during instant on/off functions in low-power IoT nodes. Further, we have chosen to optimize logic circuits in terms of energy per computation, delay, and computational accuracy to meet the requirement of emerging data-intensive workloads of energy-constraints IoT applications.

First, in this thesis, several techniques have been explored to mitigate the large leakage power consumption and achieve a low power datapath for future embedded applications such as wireless sensors nodes or IoT nodes. Based on the literature review, normally-off computing has shown promising results in mitigating the large standby power consumption by switching off the component when not in use. However, in the absence of power supply, embedded SRAM and flip-flops memory cell lose their content, without which the system is unable to resume the computation upon restoration of power.

To address this challenge, we first modify the traditional 6T SRAM cell to incorporate non-volatility using emerging magnetic technology, which offers quick wake-up times. To quantitatively analyze and evaluate the effectiveness of the proposed SRAM cell at the circuit level, we simulate the SRAM cells at different technology nodes and compare them in terms of power consumption and access latency. The proposed SRAM cell is compared to the conventional 6T SRAM counterpart indicating similar read and write performance. Also, a comparison with the existing SRAM cells for normally-off computing applications shows a 51-78% reduction in the backup power consumption.

Secondly, we propose a multi-storage SRAM cell which offers increased storage capacity, lower power consumption during multi-context computing and data retention during the power-off period in IoT applications. To verify and effectively analyze the proposed multi-storage cell, rigorous SPICE simulations are performed. It is worth noting that the proposed cell shows a reduction of energy consumption by 72% during the store operation and a reduction of 68% during the context switch operation when compared to the existing multi-storage SRAM cell. We also propose an assist circuit for the proposed SRAM cell, which offers 35% of power-saving when compared to SRAM cell without an assist circuit.

Further, we propose a state retentive D flip-flop that can store the data before power-down mode, such that the operation can be resumed from the pre-standby state. The active and sleep mode power consumptions of the proposed state retentive D-flip-flop are analyzed using extensive SPICE simulations. It is observed that the proposed flip-flop consumes almost negligible power during the sleep mode. The power consumption was also evaluated for FinFET and FDSOI technology. It is observed that the FinFETs and SOI have reduced leakage power consumption, thus provide promising alternatives to deep submicron MOSFETs suffering from large leakages. This is attributed to FinFETs vertical channel, which is wrapped around upto three sides by gate, thereby enhancing the gate control over the channel. Whereas SOI has a buried oxide layer, which prevents the formation of leakage path far away from the gate.

Furthermore, we explore asynchronous circuits to overcome the challenges faced by state-of-the-art synchronous systems in energy-autonomous applications. We explore basic asynchronous blocks such as c-element and half-buffer, which employ the event-based behavior and consume power only if an event needs to be processed. A detailed analysis of different implementations of volatile c-element is performed. The circuit-level simulations validate the proper functionality of the proposed designs. Moreover, we compare the performance of the proposed hybrid designs with their volatile counterparts.

Finally, we design an energy-efficient logic and arithmetic circuit which leverages the relaxation in computational accuracy to reduce the energy per computation. We have performed rigorous SPICE simulations for the proposed circuits, and results have confirmed that all the proposed designs perform the correct functionality during the read, write, power-off and basic memory/logic operations. The simulation results show that the proposed logic and adder circuit can efficiently perform operation in 4n and 6ns while consuming 0.47 pJ, and 0.7 pJ of energy.

Overall, in this thesis, we demonstrate significant improvements in performance and energy for on-chip storage elements (SRAM & flip-flops) and logic circuits.

8.2 Future Scope of Work

We plan to extend our future work on high-performance bitwise in/near memory CNN accelerators. These new architectures could be leveraged in the substantial advancement in Artificial Intelligence and can help provide more efficient emulation of human intelligence with great speed & accuracy. In recent years, the utilization of the machine and deep learning techniques has gained rapid popularity as a viable solution to large-scale image processing, computer vision, cognitive tasks, and information analysis applications.

However, the data-intensive workload due to the large amount of computation poses a great challenge for the conventional von-neumann based computing systems. The inevitable data movement between memory and processor has become a major performance bottleneck for the conventional systems, most commonly known as von-neumann bottleneck. This bottleneck is due to the fact that over the years, the performance of the processor has improved significantly compared to the memory, which results in higher energy consumption during data transportation than computation. In addition to substantial power consumption, the limited bandwidth of conventional memories results in heavy I/O traffic and large access latency which significantly degrades the throughput of a system.

In the future, we aim to mitigate the data transfer overhead by bringing processing closer to the memory, where we will explore different feasible architectures to support near/in-memory computing. We also plan to revisit conventional memory hierarchy to investigate the best possible multibanked architecture (mix of volatile and non-volatile banks), which can cater to a large range of Machine/deep learning applications and provide high bandwidth, massive parallelism and high energy efficiency.

The proposed work can be further augmented by investigating various methods to integrate the novel memory architectures with existing systems by extending instruction set architecture (ISA) with custom instructions using existing FPGA boards and also extend on-chip buses to support such instructions.