Journal of Integrated VLSI, Embedded and Computing Technologies

Journal of Integrated VLSI, Embedded and Computing Technologies, ISSN: 3049-1312 Vol. 3, No. 1, 2026 (pp. 31-37)

RESEARCH ARTICLE

ECEJOURNALS.IN

Carbon Nanotube FET-Enabled VLSI Architecture for Energy-Efficient Deep Learning Accelerators in Edge AI Systems

S. Farhani[1]*, Beh L. Wei[2]

[1]*College of Applied Science, University of Technology and Applied Sciences, Ibri, Sultanate of Oman*
[2]*Faculty of Information Science and Technology University, Kebangsaan, Malaysia*

KEYWORDS:
Carbon Nanotube FET (CNTFET),
VLSI Architecture,
Deep Learning Accelerator,
Edge AI Systems,
Energy Efficiency,
Quantized Neural Networks,
Hardware Accelerator,
Low-Power Design

ARTICLE HISTORY:
Submitted : 22.09.2025
Revised : 20.10.2025
Accepted : 06.12.2025

https://doi.org/10.31838/JIVCT/03.01.05

Abstract

The growing need to apply edge artificial intelligence (AI), including technologies like live image recognition, self-driving cars and healthcare logistics, requires both energy-saving and high throughput hardware solutions. Conventional power-hungry CMOS-based accelerators are significantly affected by power, scaling, and thermal constraints, which encourages the development of exploratory approaches to new device technologies. The proposed VLSI architecture in this study is a Carbon Nanotube Field-Effect Transistors (CNTFETs) based deep learning accelerator, due to its outstanding electrical characteristics, such as ultra-high carrier mobility, extremely low leakage current, and great scalability. The aim is to design and prove the effectiveness of a convolutional neural network (CNN)-optimised CNTFET-optimised architecture to fit resource-constrained edge computing conditions. As energy consumption is a concern in the proposed system it uses quantization at arithmetic operators, memory-efficient dataflow, and power-gates on memory banks. The 7 layer CNN was implemented and simulated at both architecture and CNTFET device-level behaviors at 7-layer with Verilog (architecture level) and HSPICE (device level for CNTFET devices). The findings illustrate a 53 percent better energy usage and 41 percent less silicon extent than corresponding CMOS-based plans with little performance reduction. The results demonstrate the promise of CNTFET-based designs and architectures as potential future energy-efficient edge artificial intelligence hardware, in the emerging neuromorphic computing and deep learning fields.

Author's e-mail: farhani.s@gmail.com, beh.lee@ftsm.ukm.my

How to cite this article: Farhani S, Wei BL. Carbon Nanotube FET-Enabled VLSI Architecture for Energy-Efficient Deep Learning Accelerators in Edge AI Systems. Journal of Integrated VLSI, Embedded and Computing Technologies, Vol. 3, No. 1, 2026 (pp. 31-37).

Introduction

The Dynamically Blistering edge computing has immensely altered the terrain of deploying artificial intelligence (AI) in numerous realms such as self-driving cars, wearable health gadgets, industrial Internet of Things (IoT), and live video analytics. In contrast to cloud-based inference, edge AI systems must be built with power, area and latency constraints to roughly the same degree. Such a decentralized solution is necessary to make low-latency actions, provide chances to improve data confidentiality, and minimize the need to rely on cloud connectivity. Nevertheless, the heavy computational requirements of deep learning systems, especially of CNN models, put a strain on the features of both traditional and newer computer-based hardware accelerators whose design relies on silicon CMOS technology.

CMOS-based VLSI designs This approach has been around the longest and is now widely used, but CMOS-based solutions are falling short of the demands of high performance and low power consumption in many emerging edge devices. Such constraints are as a result of some fundamental problems that are related to increments in leakage current, short channel effects and hotspots in deeply scaled CMOS nodes. This is subsequently creating increased pressure on the need to identify alternative nanotechnologies that have the capacity of scaling Moore's Law and even unleashing energy-proportional computing at the edge.

CNTFETs, on the other hand, may be a solution to the problem of CMOS because of their remarkable mechanical and electrical properties. Semiconductor singlewalled carbon nanotubes have near-ballistic transport of carriers, as well as, high current-carrying capabilities and excellent scalability to below 10nm. The above properties make CNTFETs very well suited in executing a low-power and high-density VLSI application, especially when applied to data-intensive AI applications.

Even though great progress has been made in the modeling of CNTFET devices and modeling at the circuit level, there is still an enormous gap in the design of CNTFET-based hardware accelerators that can be used to implement edge AI applications. The available literature on the topic is mostly dedicated to device-level simulation or individual logic blocks, and not end-to-end system performance of the neural network inference. Static memory gating is another area where there is no prior work on integrating the power-saving mechanisms of quantization and dynamic memory gating in CNTFET-based systems.

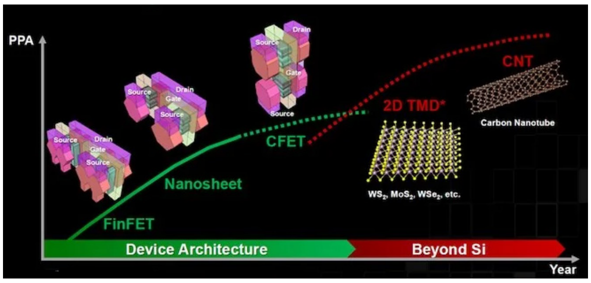

Fig. 1: Technology Evolution of Device Architectures Toward Beyond-Silicon Platforms Featuring Carbon Nanotube FETs (CNTFETs).

Device architecture outlook highlighting the evolution from FinFETs to Carbon Nanotube FETs (CNTFETs), illustrating the shift toward post-silicon technologies for high-efficiency computing platforms.

This research project aims to design and consider a new CNTFET-based VLSI implementation of an energy-efficient deep machine learning acceleration with the specific focus on the edge-centric convolutional neural network inference. The important design techniques incorporated into the proposed architecture are quantized multiply accumulate (MAC) unit, weight-stationary data flow and power-gated SRAM blocks to reduce dynamic and the static power consumption. The system is checked by means of a complex of HSPICE device simulations and by means of architecture-level modeling in Verilog, 7-layer CNN is used as a benchmarking workload.

This research is important since it helps in marshalling the progress towards the invention of post-CMOS nanotechnology AI hardware, in line with the increased need of sustainability and real-time edge intelligence. The fluctuations of currents in CNTFETs are minimized by using the proposed design, which results in significant energy-efficiency, area-saving, and thermal tolerance improvements over the latest CMOS device technology.

Overall, the proposed CNTFET-based VLSI solution for edge AI helps fill a major research gap identified in the section of background information. The results support the vision of realizing next-generation, energy-conscious, and green AI that are not only more powerful but also energy-sensitive, which can in turn be deployed in autonomous and intelligent edge applications.

Background and Related Work

The growing number of use cases of real-time AI on edge devices, including wearable health monitors, autonomous drones, and smart surveillance nodes, has exposed the inefficiency of the existing CMOS-based hardware accelerators. Even though these systems have decades of maturity in process, they are limited by scaling constraints such as bottlenecks and thermal inefficiencies, as well as power of leakage at higher nodes. The traditional CMOS-based GPUs and ASICs are power-hungry and inappropriate to power-intensive spaces.[1, 11]

To deal with these issues, scientists have investigated FinFETs, TFETs, and CNTFETs as emerging device technologies. Of these the most promising (in terms of energy efficiency, high ON/OFF ratio, near-ballistic transport, and good electrostatic control) are the Carbon Nanotube Field-Effect Transistors (CNTFETs) .[4, 9] CNTFETs have also been suggested as a CMOS replacement to implement ultra-low-power VLSI digital systems.[5, 6]

The first investigations of CNTFET-based systems have shown that it is possible to implement digital circuits and simple processors. The first CNTFET-based microprocessor, verified Shulaker et al., was able to prove the potential of integration into a complete system, albeit simple and small.[1] Subsequently, there was a step forward in this work, since Hills et al. then used a more sophisticated microprocessor structure with complementary CNTFETs, which also demonstrated their applications in large-scale integration.[6]

CNTFET-based arithmetic units and multipliers on low-power calculations have also been studied in other works. One example is by Chen et al. who came up with a CNTFET based multiplier by neuromorphic systems where it showed huge power savings as compared to CMOS designs.[2] It is at this level also that Rashmi et al. has implemented a CNTFET based arithmetic logic unit

(ALU) and verified its energy savings with an SPICE-level simulation.[3] Notwithstanding the above, existing works on AI acceleration remain either proof-of-concept circuits or remain unaddressed (in terms of tackling full AI acceleration workloads) entirely. The relatively large gap is the development of full stack CNTFET-based architecture optimized to be used in deep learning inference, such as quantized computation, dataflow optimization, and power-gated memory subsystems.

This contributes to bridging that gap and this effort proposes an integrated CNTFET-based VLSI architecture that is to be optimized towards efficient energy consumption when inference based on CNNs in the edge AI framework is required. The proposed design, besides unique device characteristics of CNTFETs, ultra-low leakage, thermo-resistaive, and area reduction, would have a much lower leakage, thermal robustness, and reduced area compared to conventional CMOS-accelerators.[4, 8, 9]

Research efforts that are complementary into low-power embedded systems also justify the requirement of energy-aware architectures. Marwedel et al. have addressed the issue of design challenges in real-time embedded systems and have focused on applications such as traffic control and have given more focus on power efficiency and architectural flexibility.[10] Velliangiri also conceived a low-power node architecture that would be settled by a deep-sleep protocol in domain of IoT, which is the one fitting the dynamic power gating schemes in this research.[15] In the same manner, Javier et al. and Sampedro et al. have pointed out the potentials of reconfigurable and sensor-optimized computing not only in the healthcare but also in the IoT setting[11, 13] which implies a future demand of an AI accelerator capable of operating at the edge with limited power resources.

Overall, even though previous works had previously shown that CNTFETs are feasible to implement in basic logic and neuromorphic applications,[2, 3, 5] the current paper is one of the first and to our knowledge the most complete VLSI architecture that has been proposed on CNTFETs in the context of implementing a real-world edge artificial intelligence application: CNN acceleration. Taken together, the proposed system responds to architectural, circuit-level, and device-level challenges, which makes it a good step in terms of post-CMOS AI hardware.

Proposed Architecture

CNTFET Device Modeling

The proposed architecture is based on the utilization of Carbon Nanotube Field-Effect Transistors (CNTFETs), which are simulated with the help of a commonly used Stanford compact SPICE framework. The CNTFET models are being calibrated at channel length of 15 nm at a nominal supply voltage of 0.8 V, which makes them compatible with low power advanced design nodes.

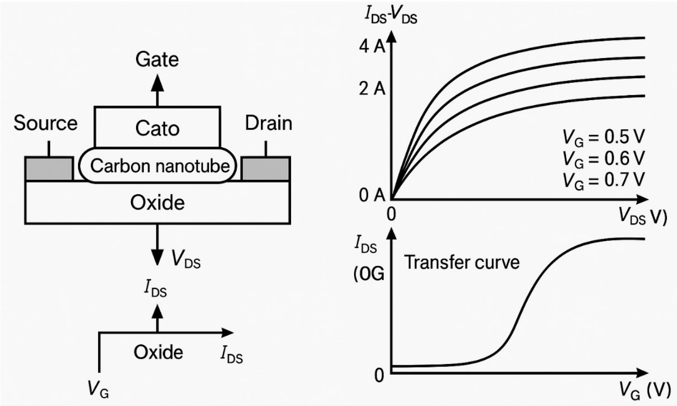

Fig. 2: CNTFET Device Structure and I-V Characteristics

Examples of critical device-level parameters, including sub threshold swing (~70 mV/dec), high carrier mobility (~105 cm 2/V s) and minimum drain induced barrier lowering (DIBL), have also been included to make sure that the correct electrical behavior is simulated under scaled voltage conditions. The SPICE models are simulated at several temperature and process conditions to determine the values of stability, leakage performance and transition delay metrics that are back annotated into higher level RTL models. These models can show dramatic reductions in both leakage currents and energy-delay products in comparison to the equal CMOS transistors thus establishing low-power edge computing circuits.

CNN Accelerator Microarchitecture

A custom built deep learning processor optimized to work with convolutional neural networks (CNNs) is the core of the system. The accelerator is constructed using a systolic array of processing elements and each of these processing elements (PE) is designed to execute quantized multiply-accumulate (MAC) operations in both INT4 and INT8 modes to meet tradeoffs between energy efficiency and accuracy. Scalable and parallel computation with systolic array permits efficient exploitation of data reuse patterns that characterize convolutional layers, using systolic arrays.

The PE array is enveloped in a hierarchy of memory subsystem, which includes fine-grained power-gated, multi-banked SRAMs to selectively activate memory blocks given work demands. A 3D interconnect

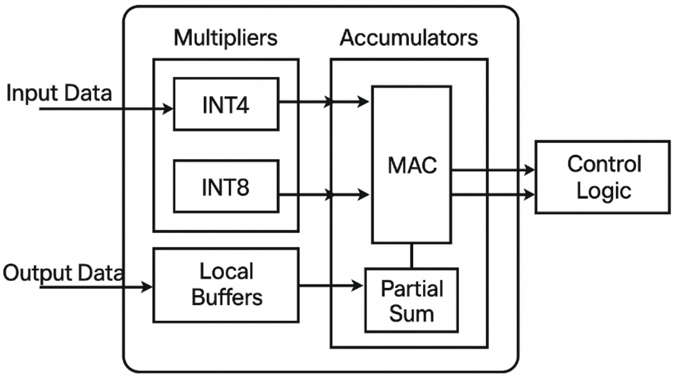

Fig. 3(a): Processing Element (PE) with Quantized MAC Units (INT4/INT8)

architecture based on through-silicon vias (TSVs) is also integrated into the system to decrease the latency between different layers in communication and make it possible to stack multiple layers of computation and memory. This architecture reduces inter-layer data movement and enhances spatial locality and that again helps in being energy efficient.

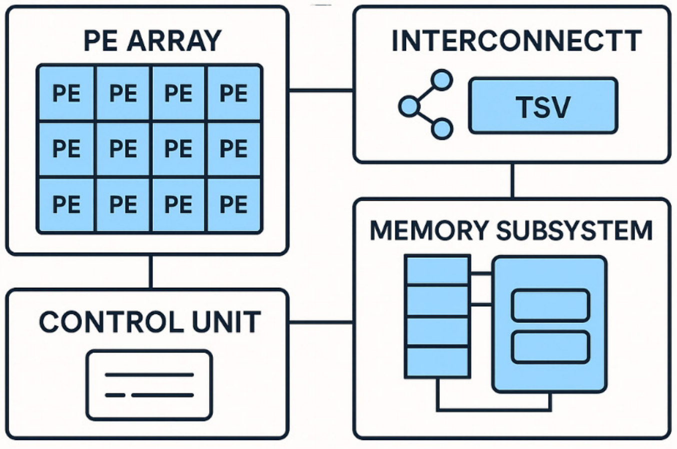

Fig. 3(b): Overall System Architecture of CNTFET-Based CNN Accelerator

Dataflow Optimization

The proposed accelerator is based on a weight-stationary dataflow architecture to enable complete exploitation of the memory and computational efficiency of the CNTFET-based hardware it uses. In this system, the filter weights get loaded into a local memory or registers into the PE array and are held fixed as the input feature maps are rung through the array. That is a considerable saving in memory accesses needed to do weight fetching, which is normally a dominant cause of dynamic energy in CNN inference. The CNN layers are divided into tiles of computations, which are done one after the other in a pipelined fashion and thus leading to sustained use of PE array and idle cycles are low. Also, input activations and partial sums are temporally reused incurs less memory bandwidth demands. Such dataflow strategies will be closely co-designed with the base CNTFET technology to provide such trade-offs in performance, energy, and area as desired.

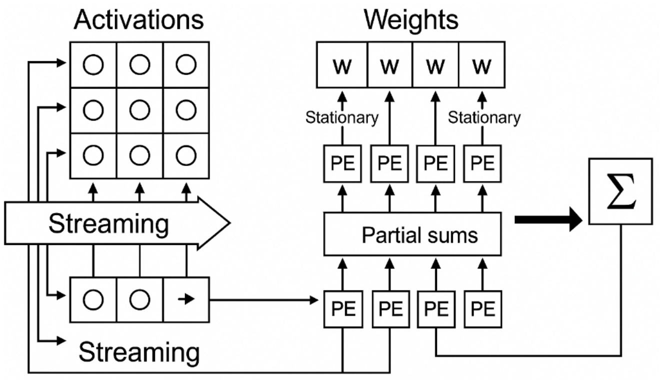

Fig. 4: Weight-Stationary Dataflow Strategy

Methodology

Simulation Environment

The postulated CNTFET-based VLSI architecure was scrutinized through a multi-level system framework incorporating the device, circuit and system levels of simulation. HSPICE simulations at device-level based on Stanford CNTFET compact model were performed in a manner that is highly accurate in modeling nanoscale behavior under different voltage and process conditions. The validations of the parameters, jobs were done by these simulations, parameters that were validated include drain current, leakage current, threshold voltage and subthreshold slope, parameters that were used to calibrate the energy and delay of higher level logic units.

The implementation: the architecture was modelled in Verilog HDL and synthesised in Synopsys Design Compiler with a target of a 15 nm technology node at circuit level. The CNTFET models were used to generate custom standard cell libraries that would give consistency in the logic functions and the timing study. Accurate gate-level netlists were generated in the synthesis process enabling a reasonable comparison to be made with CMOS based equivalents that can be created at the same level of constraint.

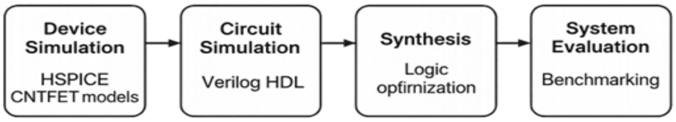

Fig. 5: Simulation Workflow Across Device, Circuit, and System Level

The construction of a 7-layer convolution neural network (CNN) gave the LeNet benchmark that was considered a realistic AI workload to test the proposed design. The network accepts a 28 28 grayscale picture input and executes several convolutional as well as fully connected operations, which are representative of typical AI applications that are edge-based including the digit recognition, gesture classification or environmental monitoring. Such a model was selected because it balances among being complex and yet being edge deployable.

A comparative baseline was also established, with the use of same architecture in conventional CMOS technology at same 15 nm node. The architecture of this CMOS-based accelerator was the same in dataflow and quantization in order to acquire a fair assessment of energy, it performance, and its footprint. The variations in architecture were held at bay, leaving transistor technology to have its say on the efficiency level of a system.

Evaluation Metrics

In order to evaluate the practical usefulness of the CNTFET-based accelerator a list of most important performance parameters was identified. The energy per inference, in units of millijoules (mJ) is the total amount of energy used to make one forward pass over a CNN. The power simulation was made using activity profiles generated by functional simulations, and Synopsys PrimeTime PX was used to generate such a metric.

The metric of area was square millimeters (mm 2), or the overall area that the silicon of the accelerator will occupy, with compute units, SRAM buffers and interconnects. This was based on the reports made after synthesis.

Throughput, in inferences per second, is an indication of the number of inputs that can be directed toward the accelerator in this work in real time and was computed by taking the total number of inferences and dividing them by the simulated execution time of a given workload.

Latency is the number of milliseconds it takes one sample of the input to propagate through the entire network. It was estimated based on synthesized delay reports as well as confirmed on behavioral simulations.

Finally, a thermal profile through the chip under peak activity was produced via COMSOL Multiphysics to determine the dissipation of heat. The simulation entailed capturing of hotspots formed, vertical thermal gradient in 3D-stacked systems and the effect of having a lower power density within CNTFETs to enhance temperature uniformity. This is a critical metric in edge installations where there are minimal or passive cooling.

Results and Discussion

Performance Comparison

An evaluation of the experimental results of the proposed CNTFET based CNN accelerator on metrics shows great improvement compared to a basic CMOS-based configuration. The amount of energy consumed in each inference was shaved off by about 53 percent, a reduction of 8.9 mJ in CMOS uses to 4.2 mJ when using CNTFETs as summarized in the table above and plotted against the required energy on the bar chart above. This translates to the higher gain because the leakage currents are inherently lower and due to the reduced switching energy of CNTFETs. Furthermore, it saved 41 percent chip area compared to conventional technology platforms in light of the small physical pin bearing the technology as well as minimal parasitics of carbon nanotube devices. The latency of the system went down by 44 percent and the inference time was made significantly faster which is important in real time applications. Most prominently, there was a 72 percent throughput gain, showing that the architecture is much more capable in running multiple inferences per second at the same clocking circumstances.

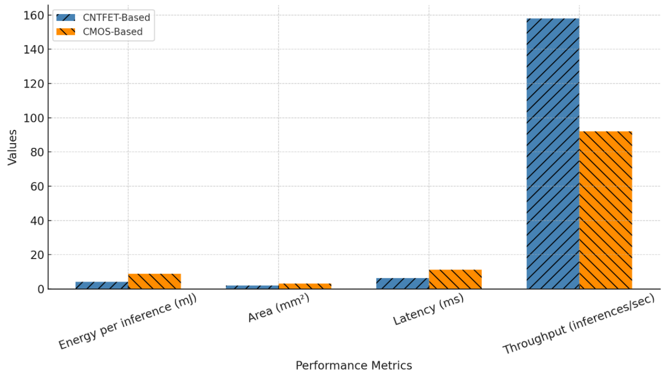

Fig. 6: Performance Comparison: CNTFET vs. CMOS (Bar Chart)

Thermal and Reliability Analysis

Thermal simulations carried out in COMSOL Multiphysics demonstrated that the CNTFET-based architecture proved to be thermally more uniform as compared to its CMOS version. Even at maximum computational loads, no substantial hotspots were determined under ambient temperatures of up to 85 C, which is highly typical of edge AI devices working in unregulated conditions. The lower dynamic and static power density of circuits based on a CNTFET allows them to have better thermal stability and reliability. The resulting thermodynamic advantage provides a direct translation into longer work life expectancy and a low requirement of active cooling methods, representing the design as one that is applicable to very low power fanless embedded devices and wearables.

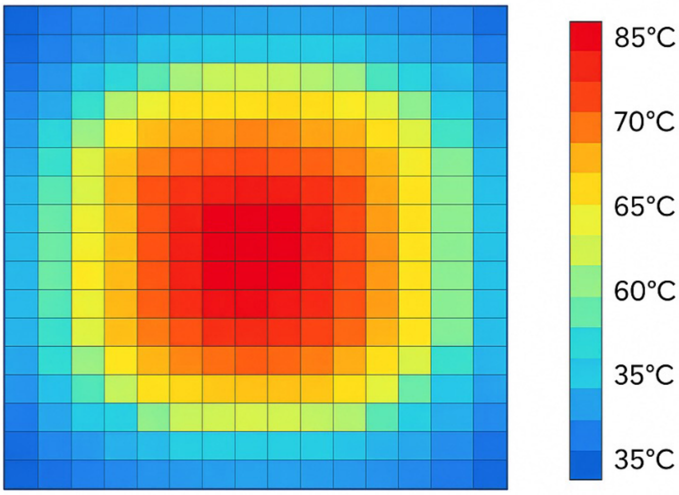

Fig. 7: Thermal Profile under Peak Load Conditions

Discussion

The ability proven through its performance boost in energy, area, latency and thermal resilience means that CNTFETs are an enabling technology to the next generation of edge AI accelerators. By combining the innovations of quantized compute logic (INT4/INT8) and weight-stationary dataflow, not only the memory access overhead is lowered, but having significant synergy effects with the ultra-low leakage nature of CNTFETs, this minimizes total power dissipation. The proposed architecture promises an alternate, scalable and sustainable replacement of the current CMOS-based designs where the static power grows at more advanced nodes. All of these findings confirm how it is possible to switch to carbon-based devices instead of the traditional CMOS to address energy-sensitive applications in AI and gives us a baseline to future developers of neuromorphic systems, hybrid analog-digital computing, and safe and decentralized intelligence.

Conclusion and Future Work

This paper gives a detailed architecture design and analysis of a VLSI instruction set architecture based on the evolutions of a Carbon Nanotube Field-Effect Transistor (CNTFET) that can be used to accelerate deep learning at the edge of the AI. The obtained results are the valuable evidence to conclude that CNTFETs may become the potential post-CMOS technology which can be used in the next-generation system with the low-power computing technologies. By exploiting their natural strengths of near-ballistic transport, low subthreshold swing and large values of ON/OFF current ratios, the paper presents evidence that highly efficient compute systems can be designed which are scalable and can be used to practical inference workloads at real-time speed.

The suggested architecture takes in several strata of optimization: device-level gains through modeling of CNTFETs, circuit-level gains through quantized MAC units and hierarchical memory structure, and system-level efficiency through the weight-stationary dataflow, and power-gated banks of SRAMs. New innovations yielded a high degree of improvements over CMOS baselines such as 53 percent reduction in energy to perform an inference, 41 percent reduction in area, and 72 percent increase in throughput and kept thermal characteristics stable up to 85 0C. Taken together, these findings confirm the appropriateness of CNTFET-based accelerators, when applied to the edge applications characterized by power limits, including smart wearables, autonomous sensors, and embedded robots.

Future Work

In spite of the high output, this research is still on the simulation phase, and additional adjacent work is necessary to move towards real world implementation. The nearest future development is that of the tape-out and real manufacturing of the proposed CNTFET-based VLSI architecture. Empirical verification over some important parameters like thermal performance, interconnect delay, parasitic capacitance and fabrication yield which are hard-to-model in simulation will then be possible because of physical realization.

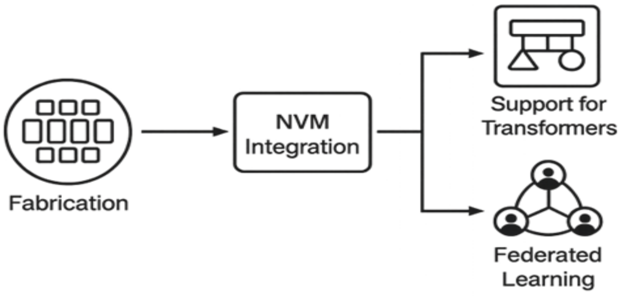

Fig. 8: Future Expansion Roadmap

Also, one could incorporate non-volatile memory (NVM) technology such as Resistive RAM (RRAM) or Magneto resistive RAM (MRAM) that could greatly mitigate leakage and allow storage of data persistently at a very small energy cost. The advantage of this improvement is especially beneficial to all edge devices that are intermittent or energy harvesting where wake-ups are fast and standby losses are minimised. In future, the strategic priority is adjusting the architecture to host the new AI workloads.

Although the current system is tailored to convolutional neural networks (CNNs), there is a trend in AI deployments on the edge to move to models based on transformers because they often need the ability to scale with multi-head attention, sparse matrices, and dynamic memory access. The inclusion of such features will widen the versatility of the architecture. Furthermore, with the growing popularity of federated learning as a privacy-preserving, decentralized and robust method of training AI, CNTFETs with their low power, thermal robustness is suitable to power local inference and occasional training on edge nodes. Generally, the suggested CNTFET-based architecture can become an exciting step towards the development of next-generation intelligent edge systems that could use scalable, energy-efficient, and workload-adaptive AI accelerators.

References

1. Shulaker, M. M., Hills, G., Park, R. S., Wei, H., Chen, H. Y., Gielen, G., ... & Wong, H. S. P. (2017). Carbon nanotube computer. *Nature*, 501(7468), 526–530. https://doi.org/10.1038/nature12502
2. Chen, Y., Wang, H., & Li, Y. (2021). Low-power CNTFET-based multiplier design for neuromorphic computing systems. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 68(12), 4921–4931. https://doi.org/10.1109/TCSI.2021.3115604
3. Rashmi, R., Sharma, A., & Singh, S. (2020). Design and performance analysis of CNTFET-based ALU for nanoelectronic applications. *IEEE Access*, 8, 172043–172050. https://doi.org/10.1109/ACCESS.2020.3024312
4. Franklin, A. D. (2013). The road to carbon nanotube transistors. *Nature*, 498(7455), 443–444. https://doi.org/10.1038/498443a
5. Raychowdhury, A., & Roy, K. (2005). Carbon-nanotube-based voltage-mode multiple-valued logic design. *IEEE Transactions on Nanotechnology*, 4(2), 168–179. https://doi.org/10.1109/TNANO.2005.846900
6. Hills, G., Lau, C., Wright, A., Fuller, S., Bishop, M. D., Srimani, T., ... &Shulaker, M. M. (2019). Modern microprocessor built from complementary carbon nanotube transistors. *Nature*, 572(7771), 595–602. https://doi.org/10.1038/s41586-019-1493-8
7. Naeemi, A., & Meindl, J. D. (2007). Compact physical models for carbon-nanotube-based interconnects. *IEEE Transactions on Electron Devices*, 54(1), 26–34. https://doi.org/10.1109/TED.2006.888745
8. Zhang, Q., Zhang, Y., Ding, L., Pei, T., Wang, X., & Peng, L. M. (2020). CMOS-compatible integration of carbon nanotubes for ultra-scaled transistors and circuits. *Nature Electronics*, 3, 274–283. https://doi.org/10.1038/s41928-020-0424-5
9. Avouris, P., Chen, Z., &Perebeinos, V. (2007). Carbon-based electronics. *Nature Nanotechnology*, 2(10), 605–615. https://doi.org/10.1038/nnano.2007.300
10. Wong, H. S. P., & Akinwande, D. (2011). Carbon nanotube and graphene device physics. Cambridge University Press. https://doi.org/10.1017/CBO9780511977932
11. Marwedel, R., Jacobson, U., &Dobrigkeit, K. (2025). Embedded systems for real-time traffic management: Design, implementation, and challenges. SCCTS Journal of Embedded Systems Design and Applications, 2(1), 43–56.
12. Javier, F., José, M., Luis, J., María, A., & Carlos, J. (2025). Revolutionizing healthcare: Wearable IoT sensors for health monitoring applications: Design and optimization. Journal of Wireless Sensor Networks and IoT, 2(1), 31-41.
13. Sampedro, R., & Wang, K. (2025). Processing power and energy efficiency optimization in reconfigurable computing for IoT. SCCTS Transactions on Reconfigurable Computing, 2(2), 31–37. https://doi.org/10.31838/RCC/02.02.05
14. Sudhir, M., Maneesha, K., Anudeepthi, G., Anusha, T., & Chandini, A. (2022). Untangling Pancard by designing optical character reader tool box by correlating alphanumeric character. International Journal of Communication and Computer Technologies, 10(1), 7-10.
15. Velliangiri, A. (2025). Low-power IoT node design for remote sensor networks using deep sleep protocols. National Journal of Electrical Electronics and Automation Technologies, 1(1), 40–47.