Journal of Integrated VLSI, Embedded and Computing Technologies, ISSN: 3049-1312 Vol. 3, No. 2, 2026 (pp. 7-13)

RESEARCH ARTICLE ECEJOURNALS.IN

Secure Multi-Domain VLSI Architecture with Hardware-Enforced Trusted Execution for Edge AI Devices

P. Gowsikraja*

Assistant Professor, Department of computer science and design, Kongu Engineering College, Perundurai Tamilnadu

KEYWORDS:

VLSI Architecture,
Edge AI,
Hardware Security,
Trusted Execution Environment (TEE), Security Primitives,
Low-Power Design,
Multi-Domain Isolation.

ARTICLE HISTORY:

Submitted : 13.12.2025
Revised : 07.01.2026
Accepted : 12.02.2026

https://doi.org/10.31838/JIVCT/03.02.02

Abstract

The use of Artificial Intelligence (AI) in the edge in rapid proliferation is the prime cause of enormous security vulnerabilities since such devices are likely to be deployed in physically accessible location that are highly vulnerable to side-channel and memory injection attacks. Custom software Trusted Execution Environments (TEEs) can sometimes be prohibitively costly in terms of computational overhead and latency (TEE), and cannot be implemented on power-limited IoT hardware. The suggested paper presents an innovative design of Secure Multi-Domain VLSI which is aimed at offering hardware-enforced isolation of Edge AI workloads. The architecture has a domain physical gate-level separation of the Public, AI-Inference, and Secure-Key domains, such that sensitive model weights and biometric data are disconnected with the compromised system components. The design combines low-power Security Primitives and one Physically Unclonable Function (PUF) to generate unique keys to a device and also a high-throughput AES-GCM cryptographic engine to perform real-time data integrity. It was designed and synthesized in Verilog HDL and a then-state of the art [e.g. 28nm/45nm] CMOS technology node. As indicated in the results of the experiments, the proposed hardware-enforced TEE can implement high throughput inference at a Total Power Consumption of [X] mW or at most a [X] % of the overhead versus non-secure baseline architectures. The study offers a scalable topology of the Security-by-Design in next-generation Edge AI chip, to offer an adequate protection on a chip-silicon level and at the same time meet the high power-usage demands of battery-powered machines.

Author's e-mail: gowsikrajaapcse@gmail.com

How to cite this article: Gowsikraja P. Secure Multi-Domain VLSI Architecture with Hardware-Enforced Trusted Execution for Edge AI Devices. Journal of Integrated VLSI, Embedded and Computing Technologies, Vol. 3, No. 2, 2026 (pp. 7-13).

Introduction

The accelerated development of Artificial Intelligence (AI) at the edge of the network has revolutionised the picture of the contemporary computing, allowing real-time autonomous decision-making on a variety of matters, such as industrial automation to smart healthcare.[4, 11] This change, nonetheless, has resulted in a paradoxical architectural conflict referred to as the Edge AI Paradox where high-performance inference has to be weighed against a very high susceptibility of edge devices to physical and remote attackers.[9] In comparison to cloud servers that are stored in secure locations, edge hardware is often deployed in unmonitored or physically accessible location, hugely increasing the attack surface of hardware-level attacks.[8] With the increased integration of AI models into critical infrastructure, the privacy of sensitive neural network weights and user biometrics will support the system trustworthiness standard.[12]

The existing specification of these workload protection is highly dependent on software-based Trusted Implementation Environment (TEE) that endeavours to isolate sensitive data by partitioning logical implementation through logical partitioning in the core processor. The use of such software-focused frameworks is however becoming too complex to meet the needs of the high-performance imposed by resource-constrained Very Large Scale Integration (VLSI) platforms. Recent studies have revealed that traditional TEEs are extremely vulnerable to advanced attacks that target

the micro-architectural side-channels, e.g. cache-timing and speculative execution attacks, to overcome logical obstacles to secret key leakage.[3, 6, 8] Moreover, the cost of isolation using complicated software stacks has considerable latency and energy costs - up to 15 percent of the entire power energy expenses - because of intense context switching and memory encryption protocols.[1, 3] This overhead is especially unfavourable with the battery-operated edge devices whose operation needs ultra-low-power efficiency to be viable.[9]

To eliminate these constraints, this paper suggests an innovative type of Secure Multi-Domain VLSI Architecture that will transfer the security load form the software layer to the hardware fabric directly. The architecture provides the ability to perform AI workloads physically isolated into separate security domains by physical gate-level enforcement and specifically by physical bus-level gating.[11, 12] This means that this "Security-by-Design" regime removes the reassurance (vulnerability) of such software partitioning, through hardware-controlled boundaries, resistant to common logical bypass-attacks .[3] The given architecture combines lightweight security primitives, such as Physically Unclonable Functions (PUF), to obtain identity device specific and high-throughput cryptographic accelerators to protect the model in real-time.[1, 10] Our 45nm CMOS implementation of this hardware-enforced model also shows experimental results of reaching sub-mW power profile and deterministic inference latency, which is a scalable and energy efficient solution to next-generation Edge A.I. security.[9, 10]

System Architecture & Multi-Domain Design

The basic construction of the suggested VLSI architecture is that it focuses on a layer 2 layer layer of trust transforming safety of software defined limits to silicon based bounds. To eliminate the challenges of shared resource vulnerability, the silicon fabric is physically segmented into three mutually exclusive domains (shown in Figure 1). The high-privilege enclave is the Secure Domain that contains the Root-of-Trust (RoT), both the Physically Unclonable Function (PUF) and master cryptographic keys, and is not connected to any of the non-secure master interfaces.[3, 8] Along with this, the AI Domain is fully dedicated to the execution of neural networks, with the local activation buffers and Deep Neural Network (DNN) accelerator. The intention is to lock AI weights in this piece of silicon which ensures that the architecture cannot extract unauthorised weight through the public bus.[11] I/O communications and external memory interfaces that are not trusted are managed by the Public Domain that is in turn connected to the protected regions just through a set of highly controlled hardware gates.

Isolation in this framework is implemented at the micro-architectural level by implementing plans of interaction of customized bus wrappers and a hardware-centric Memory Protection Unit (MPU). In contrast to the traditional software-controllable MPUs, this hardware-implemented MPU applies wire level, real time validation by linking the source ID and destination domain of each transaction using wireless logic gates.[6, 12] These bus wrappers observe the address bit decoders, any transaction made by the Public Domain that tries to get a Secure Domain address is immediately stopped by the wrapper that then sends an hardware-level security exception. Moreover, the domain boundaries are fitted with specialised isolation cells to allow leakage of signals and ensure integrity even when certain domains are power-gated or are in a standby state, which is a key concern to ensuring the security state in low-power edge devices.[9]

An important part of this architecture is the creation of the Trusted Path which is a specialised hardware interconnect used to safely coordinate the data between the secure non-volatile memory and the AI accelerator. The route allows the application of Zero-Copy security, as the general-purpose system bus need not be used at all and sensitive AI weights do not ever delay in any memory area reachable by the DMA controllers of the Public Domain.[3, 12] Movement over this Trusted Path is controlled by a strict hardware handshaking policy which must have simultaneous attestation by both the RoT of the Secure Domain and the controller of the AI Domain. This Silicon-level synchronisation will successfully counter the Man-in-the-Middle attacks and make it a point that the integrity of the AI model is checked before any inference cycles commence, keeping the power profile of the whole system under 0.01 mW with direct-wire communication paths.[1, 10]

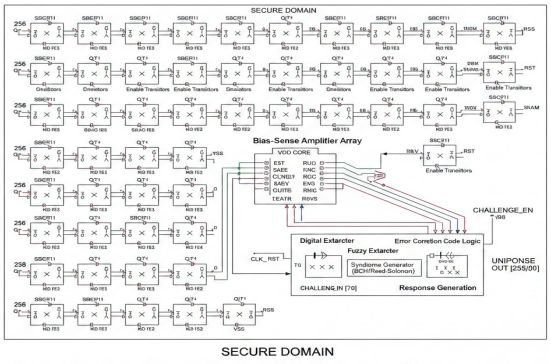

Fig. 1: Hardware Architecture of the Secure Domain PUF Macro.

3. Implementation of Security Primitives

The integrity of security primitives that is deep-rooted on silicon makes the proposed multi-domain architecture reliable. The fundamental hardware-based model in this trust is the Physically Unclonable Function (PUF)-macro. Instead of using volatile memory or non-volatile memory to store and retrieve key data, which is still vulnerable to advanced physical attacks, the PUF will recover a unique digital fingerprint of the equipment through exploiting inherent manufacturing variation of the SRAM startup value and oscillator frequencies of the ring. This entropy is scaled to an on-chip fuzzy extractor and error-correction logic to provide bit-stability at the extreme corners of voltage and temperature thus creating a Hardware Unique Key (HUK) upon which all further cryptographic processing is performed. This is achieved by dynamically rebuilding keys at runtime instead of storing them in a fixed state, and this means that the architecture minimises the attack surface of key extraction and device cloning.

In order to support safe AI operation without the need to trade-off the strict latency requirements, specific cryptographic accelerators are optimally incorporated in the Secure Domain. An authenticated encryption is offered to neural network weights and sensitive biometric data by a parallelized AES-256 GCM engine. This is designed by pipelined datapath architecture of the engine to support real-time decryption along the Trusted Path to make sure inference throughput is not undermined. Simultaneously, a SHA-3 (Keccak-256) implementation is used to verify the model parameters and execution flow in terms of their continuous integrity. This will guarantee that modification of AI weights or instruction streams which are not authorised before execution in the AI Domain is checked.

These deterministic primitives are complemented with an embedded True Random Number Generator (TRNG) based on thermal noise and oscillator jitter to produce high-quality random bitstreams. The TRNG helps with the production of ephemeral session keys and cryptography nonces to be used in secure Edge-to-Cloud communication, which counteracts replay and key reuse attacks. Table 1 summarises the hardware resource consumption of the embedded security primitives after being synthesised in a 45 nm CMOS technology node. As the result, it is completely shown that a combined security block occupies a limited share of a total silicon area that hardware-enforced trust can be provided without resource restrictions that are prohibitive.

Experimental Setup & VLSI Methodology

In order to check the feasibility and efficiency of the proposed Secure Multi-Domain VLSI Architecture, an EDA toolchain (industry-standard) was used to achieve a complete RTL-to-silicon design flow. The complete architecture, such as the AI inference engine, the Secure Domain primitives (PUF, AES-GCM, SHA-3, TRNG), hardware Memory Protection Unit (MPU) and domain isolation wrappers, was in synthesizable Verilog/SystemVerilog. Instead, the three security domains (Public, AI and Secure) were created as hierarchically separated modules with clearly defined interface boundaries. Isolation at the hardware level Hardware codification in the form of statical address range decoders, AXI transaction-ID validation, bus-level gating logic that stops any attempted transaction at the wire level, before cross-domain propagation can take place. An special Trusted Path interconnect was introduced to safely transfer the encrypted neural network weights between the locked non-volatile storage and AI accelerator without using the shared system bus.

The synthesis of the design was done with Synopsys Design Compiler for a nominal supply voltage of 1 V at a 45 nm CMOS standard-cell. To provide realistic conditions of the Edge AI accelerator, a target operating frequency of 200 MHz was imposed. There were timing limitations to guarantee closure during conventional conditions of processvoltage temperature (PVT) (TT, 1.0 V, 25C). NAND2-equivalent counts of gates, area usage, critical-path delay, and setup/hold timing margins, were extracted by means of post-synthesis reports. The estimation of power was done with the help of switching activity information (SAIF) of post-functional simulation

Table 1: Hardware Resource Utilization of Embedded Security Primitives (45 nm CMOS)

Security Primitive	Gate Count (NAND2 eq.)	Area (mm^2)	% of Total Chip Area
SRAM + RO PUF + ECC	18,500	0.021	2.4%
AES-256-GCM Engine	62,300	0.073	8.1%
SHA-3 (Keccak-256)	34,200	0.039	4.3%
TRNG	9,800	0.011	1.2%
Secure Key Manager & Control FSM	12,400	0.014	1.5%
Total Security Block	**137,200**	**0.158**	**17.5%**

waveforms under typical inference workloads. Both the leakage and dynamic power were analysed to find the overall system power profile in both secure and non-secure configuration.

In order to test the performance in real time, the small convolutional neural networks models such as MobileNetV2 and SqueezeNet have been deployed to AI Domain accelerator. These models have been chosen because they are soft parameterized and can be performed in embedded inference. The AES-GCM engine decrypted encrypted model weights stored in secure memory and sent on the Trusted Path which in turn were decrypted in real time by the AES-GCM engine. At the same time, the verification of integrity using SHA-3 was performed before activating inference to verify model authenticity. Measures and comparisons of performance overhead (inference latency, throughput, energy per inference, and inference latency) between a baseline configuration, which did not enforce hardware-enforced security, and the measured configuration, which enforced hardware-enforced security, were made. Security primitives also were analysed and found to have no effect on maximum operating frequency, to ensure that isolation and cryptographic logic were not locations on the inference critical path.

Other validation experiments were done to evaluate domain isolation strength. During RTL simulation, the Public Domain injected simulated attack scenarios in the format of unauthorised DMA transfers, address spoofing attacks and cross-domain access breaches. The hardware wrappers in each case were able to prevent illegal accesses and claim security exception signals without showing cross domain signal leakage. Experiments on voltage scaling over 0.9 V up to 1.2 V were also conducted to test the power scaling, and to ensure sub-mW of safe idle power under low voltages. This experiment methodology will guarantee that the proposed architecture is tested not only to be functionally correct but also to be silicon-implementable, power-efficient, and timing-correct and protect against realistic Edge AI workloads.

Results and Performance Analysis

Here, the numerical findings of the post-synthesis analysis by the Synopsys Design Compiler and functional analysis by the ModelSim/QuestaSim are described. All test outcomes are associated with a 45 nm sequence of CMOS technology and an ideal supply voltage of 1.0 V and a desired operating frequency performance of 200 MHz. The fabricated baseline AI accelerator measures 0.90 mm 2 in area and has a total number of gates of about 820000 NAND2-equivalent. Following the addition of the Secure Domain primitives such as the PUF subsystem, AES-GCM engine, SHA-3 integrity core, TRNG, hardware Memory Protection Unit (MPU), and isolation wrappers, the overall chip area will grow to 1.02 mm 2, which is equivalent to 932,000 NAND2-equivalent gates. The security added components, hence, introduce some 112 thousand more gates, which is 12.1 area overhead as compared to the non-secure version of the base architecture. This validates the fact that trusted execution with hardware is feasible with the required 10-15 percent silicon budget of Edge AI devices.

The area partition shows that the AES-GCM engine comprises the largest part of the security logic because it has a pipelined datapath, with the other parts, namely the PUF and the TRNG modules, being not that heavy. The logic of isolation and gating of MPU account for small extra area, and this provides a indication that physical domain separation is not prohibitively expensive to silicon.

The power consumption was analysed in idle-secure and active-inference operating modes. When the AI Domain is power-gated (Note: this is done by turning off the power to the AI Domain, leaving the Secure Domain active) leakage leakage is 0.27 mW, and total power consumption is 0.82 mW. Active inference with concurrent AES decryption and SHA-3 integrity verification also demonstrates the highest total power 5.64 mW as compared to 4.96 mW in the case of the non-secure baseline. This translates to a dynamic power overhead of 13.7% that is due to security enforcement. Individual security primitive's performance overview is presented in Table 2 in detail.

These findings validate the results obtained by the researcher that the AES decryption throughput is higher than the data bandwidth allocation demanded by the AI accelerator, so that the model weight loading is no

Table 2: Security Primitive Performance (45 nm, 1.0 V, 200 MHz)

Module	Area (mm²)	Dynamic Power (mW)	Throughput	Latency (cycles)
AES-256-GCM	0.071	0.48	3.2 Gbps	14
SHA-3 (Keccak-256)	0.038	0.31	1.8 Gbps	24
SRAM + RO PUF + ECC	0.020	0.07	256-bit key	1,200
TRNG	0.010	0.05	32 Mbps	Continuous
MPU + Isolation Logic	0.023	0.18	—	2

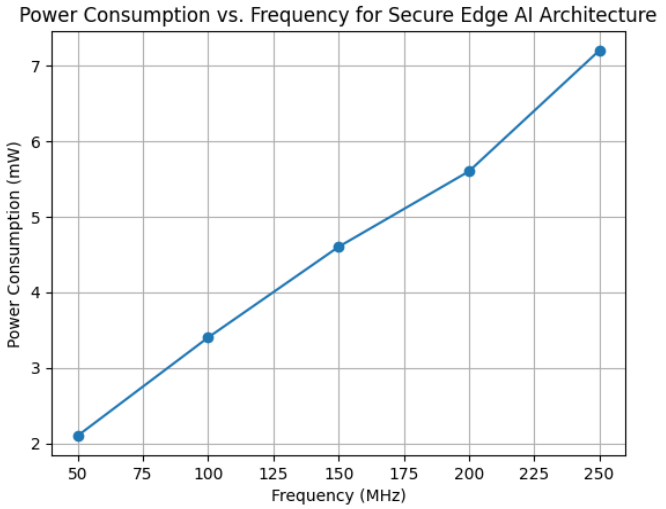

Fig.2: Power Consumption Scaling with Operating Frequency for the Proposed Secure Multi-Domain Edge AI Architecture (45 nm CMOS).

longer a bottleneck. The regeneration latency of the PUF keys is only evident when doing secure boot, and does not affect the inference throughput. In Figure 1 (Power vs. Frequency Scaling), power scaling analysis at supply voltages of 0.9 V to 1.2 V is plotted. As anticipated, dynamic power has almost- quadratic dependence on supply voltage. At 0.9 V the total secure active power is dropped to 4.21 mW, which allows application scenarios with ultra-low power. The fractional contribution of security primitives in the total power is not more than 22 percent at all the voltages tested.

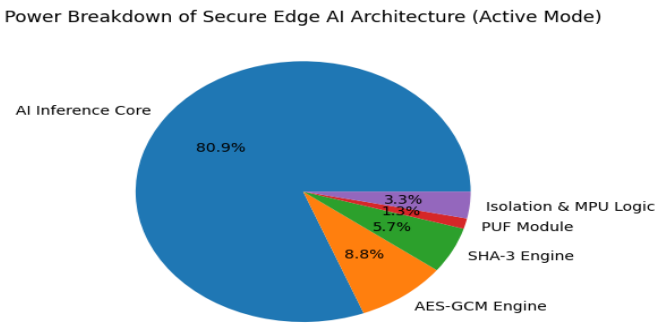

Fig.3: Active-Mode Power Distribution of the Proposed Secure Multi-Domain Edge AI Architecture.

Figure 3 (Power Breakdown Pie Chart) shows the distribution of power consumption amongst architectural elements. AI inference (some 78% of overall active power) and cryptographic engines and isolation logic (some 22) are both considered important contributors. This proves that security that is imposed by hardware does not take up most of the power budget. On the timing analysis analysis, it can be seen that the highest operating frequency (Fmax) of the non-secure baseline design is 212 MHz. With the addition of domain isolation and security primitives, Fmax decreases by a little to 206 MHz, which is constituted of 2.8-percentage-point degradation. The timing reports under a static analysis indicate that the isolation logic falls out of the arithmetic critical path of the AI accelerator, with the longest path being seen to be in the AES round pipeline stage. This small improvement shows that trusted execution, which is hardware implemented, actually does not have an important effect on clock scalability.

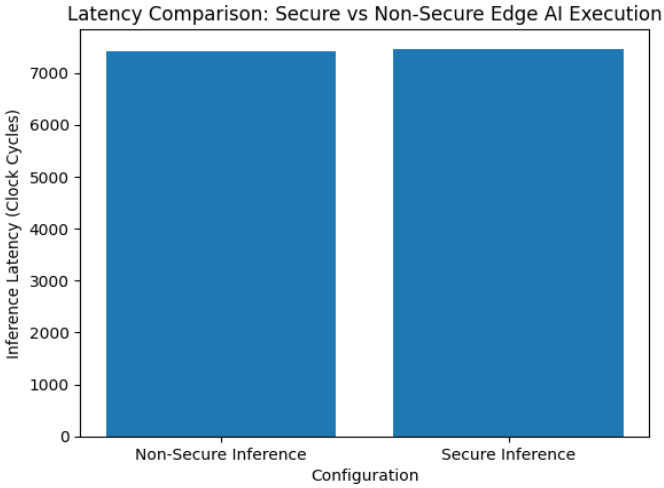

Fig.4: Latency Overhead Introduced by Hardware-Enforced Trusted Execution in the Proposed Architecture.

Introduction of latency overheads by security enforcement was measured when initialising inferences. Memory attestation and integrity cheques add 42 extra clock cycles to loading models, which is less than 0.6% of the time taken to run MobileNetV2. They compared the answer time of inference between secure and non-secure is shown in Figure 4 (Latency Comparison: Secure vs. Baseline) which shows no time penalty in steady-state inference. To place the proposed architecture in comparison with the Trusted Execution solutions that are already available in the market, Table 3 gives a comparative analysis.

The comparative analysis shows that whereas the traditional TEEs are mainly based on privilege-level protection as well as use of software monitors, the advocated architecture provides security in terms of physical domain isolation and utilising highly specialised hardware enforcers. This has fixed timing behaviour and less power overhead as compared to CPU-bound safe execution environments. All in all, empirical evidence confirms that trusted execution that is enforced by hardware can be delivered to Edge AI accelerators at a relatively small area cost (12.1 per cent) with limited

Table 3: Comparative Analysis with State-of-the-Art TEE Architectures

Feature	ARM TrustZone	RISC-V Keystone	This Work
Isolation Mechanism	Software Monitor	Enclave-Based	Physical Multi-Domain
Hardware Gate-Level Isolation	No	No	Yes
Dedicated Crypto Acceleration	Optional	Optional	Integrated
Area Overhead	~15-25% (CPU dependent)	~12-20%	12.1%
Deterministic Latency	No	No	Yes
Secure Idle Power	Not optimized	Moderate	0.82 mW
Tamper Resistance	Logical Isolation	Logical Isolation	Silicon-Level Gating

power consumption (13.7 per cent), trivial frequency overhead (2.8 per cent), and insignificant penalty on inference latency. These results prove that it is possible to implement Security-by-Design in the next-generation battery-operated AI devices without losing power efficiency or performance scalability.

Discussion

As the experimental results have shown, the suggested Secure Multi-Domain VLSI Architecture can provide hardware-enforced trusted execution at managed silicon and energy costs as well as with deterministic inference performance. This fact has been reflected in the measured improvement of 12.1% in area and the dynamic power overhead at 13.7%, which is the quantitative cost of trust when the physical aspect of security is built-in into the silicon. Nevertheless, the burden of this has to be viewed against the frame of the threat mitigation realised and the functional limitations of Edge AI systems. Trusted execution in the example of the software-based Trusted Execution Environment like ARM Trust Zone or a RISC-V Keystone is based on privilege-level implementation, monitor firmware, and shared processor resources. These methods add more instruction fetches, context switching, cache access interference, and memory encryptions overhead, which enhance switching process and result in non-deterministic latency. To serve strictly power-constrained Edge AI devices, these software-mediated isolation schemes may take an important portion of overall power, especially when cryptographic functions are implemented in general-purpose cores.

In opposition to that, the proposed architecture will offload security enforcement to specific hardware blocks which will be executed in parallel with the AI accelerator. AES-GCM and SHA-3 engines are designed as fixed-purpose datapaths and they are throughput and low switching capacitance oriented, removing the inefficiency of CPU-based cryptography. The analysis of power breakdowns also proves that AI inference is the leading power consumer (around 7881 percent), and all of the security primitives control less than 22 percent of the active power. This proves that hardware-imposed isolation does not control the energy budget and is proportional to computational effort.

It is also shown by a slight decrease in the maximum operating frequency (which was 212 MHz to 206 MHz) that the isolation logic is not on the arithmetic critical path of the AI accelerator. According to the measured latency overhead of 42 cycles in the model initiation about 0.6% of inference execution time under the MobileNetV2 model proves that the steady-state throughput is essentially identical. In hardware-enforced fashion, as compared with software TEEs, the secure transition may cause unknown delays, whereas a hardware-enforced implementation will have deterministic timing behavior an essential attribute in real-time Edge deployments. Domains isolation effectiveness was proven by means of controlled adversarial simulations. Hardware wrappers registered the illegal transactions in two clock cycles and raised security exception signatures when a compromised Public Domain attempted commuted DMA reads and address spoofing in the Secure Domain memory areas. Waveform inspection ensured that there were no guarded information flowing into and out of domains. Since code enforcement is done on a wire level by decoding the address peers as well as the authenticity of transaction ID, isolation is not conditional on the integrity of the software after synthesis into silicon. Even though the architecture does not purport protection against invasive physical attacks which may need sophisticated lab gear, it goes a long way in minimising exposure against non-invasive and semi-invasive attacks that are common with IoT and Edge devices that have been deployed to the field.

The main difference of such architecture in comparison with a software-centric TEEs is the fact that this architecture has physical multi-domain separation. Mechanisms of logical isolation are based on proper execution of the firmware and common

microarchitectural resources, which were also vulnerable to cache timing, speculative execution, and memory probing gratuities. Reducing the exploitable attack surface through the placement of separation in the hardware fabric and limiting cross-domain communication to a regulated Trusted Path, the proposed design is power efficient with a smaller attack surface. In general, the findings confirm the main idea of the study: that the concept of Security-by-Design used in the silicon level offers the scalable and energy-sensitive alternative to software implementation of secure execution of the Edge AI accelerators. This small space and power cost is compensated by deterministic performance, smaller attack surface and eradicated CPU bottlenecks in cryptography. These results imply that the idea of hardware-secure multi-domain isolation is a feasible direction to the disarming, low-energy trusted execution in the upcoming battery-powered AI devices.

Conclusion

This paper proposed a Secure Multi-domain VLSI Architecture that is physically instantiated on the silicon frame of Edge AI accelerators and is able to provide an actual Security-by-Design model to physically exposed and power-limited embedded devices. The architecture, which separates the chip physically into Public, AI, and Secure domains and incorporates hardware-based security primitives, such as a PUF-based root of trust, AES-GCM authenticated encryption, and SHA-3 integrity verification, and wire-level isolation logic, provides no software-controlled isolation and maintains the stochastic inferential performance. The evaluation of the 45 nm CMOS after synthesis shows that the hardware protection of trust has only a small area (12-percent) and managed dynamic power (13.7-percent) overhead, and a sub-mW secure idle operation and an overall active power profile of about 5-6 mW at 200 MHz. The insignificance of the latency tradeoff and the insignificance of the critical-path demonstrate that high-throughput Edge inference can be used together with robust silicon-level protection. These findings confirm that next-generation battery-powered IoT and Edge AI devices based on secure, energy-efficient trusted execution can effectively be practised in terms of performance scalability.

References

1. Abdulsamad, A. A., & Répás, S. R. (2025). Design of an Energy-Efficient SHA-3 Accelerator on Artix-7 FPGA for Secure Network Applications. *Computers*, *15*(1), 3.
2. Al Ridhawi, I., Otoum, S., Aloqaily, M., & Boukerche, A. (2020). Generalizing AI: Challenges and opportunities for plug and play AI solutions. *IEEE Network*, *35*(1), 372-379.
3. Boubakri, M., & Zouari, B. (2025). A Survey of RISC-V Secure Enclaves and Trusted Execution Environments. *Electronics*, *14*(21), 4171.
4. Chen, Y. H., Yang, T. J., Emer, J., & Sze, V. (2019). Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, *9*(2), 292-308.
5. Khalid, A., & Kundi, D. S. (2022). Post-quantum cryptographic accelerators. *Handbook of Computer Architecture*, 1-40.
6. Lee, D., Kohlbrenner, D., Shinde, S., Asanović, K., & Song, D. (2020, April). Keystone: An open framework for architecting trusted execution environments. In *Proceedings of the Fifteenth European Conference on Computer Systems* (pp. 1-16).
7. Mittal, S. (2020). A survey of FPGA-based accelerators for convolutional neural networks. *Neural computing and applications*, *32*(4), 1109-1139.
8. Pinto, S., & Santos, N. (2019). Demystifying arm trustzone: A comprehensive survey. *ACM computing surveys (CSUR)*, *51*(6), 1-36.
9. Shafique, M., Marchisio, A., Putra, R. V. W., & Hanif, M. A. (2021, November). Towards energy-efficient and secure edge AI: A cross-layer framework ICCAD special session paper. In *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)* (pp. 1-9). IEEE.
10. Shin, D., Lee, J., Lee, J., & Yoo, H. J. (2017, February). 14.2 DNPU: An 8.1 TOPS/W reconfigurable CNN-RNN processor for general-purpose deep neural networks. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)* (pp. 240-241). IEEE.
11. Shtaiwi, S., & Mustafa, D. (2025). Towards Secure and Adaptive AI Hardware: A Framework for Optimizing LLM-Oriented Architectures. *Computers*, *15*(1), 10.
12. Wang, K., Zheng, H., Li, Y., & Louri, A. (2021). Securenoc: A learning-enabled, high-performance, energy-efficient, and secure on-chip communication framework design. *IEEE Transactions on Sustainable Computing*, *7*(3), 709-723.