3D Guard-Layer: An Integrated Agentic AI Safety System for Edge Artificial Intelligence

Eren Kurshan
Princeton University
ekurshan@princeton.edu

Yuan Xie
Hong Kong University
of Science and Technology
yuanxie@ust.hk

Paul Franzon
North Carolina State University
paulf@ncsu.edu

***Abstract*—AI systems have found a wide range of real-world applications in recent years. The adoption of edge artificial intelligence, embedding AI directly into edge devices, is rapidly growing. Despite the implementation of guardrails and safety mechanisms, security vulnerabilities and challenges have become increasingly prevalent in this domain, posing a significant barrier to the practical deployment and safety of AI systems. This paper proposes an agentic AI safety architecture that leverages 3D to integrate a dedicated safety layer. It introduces an adaptive AI safety infrastructure capable of dynamically learning and mitigating attacks against the AI system. The system leverages the inherent advantages of co-location with the edge computing hardware to continuously monitor, detect and proactively mitigate threats to the AI system. The integration of local processing and learning capabilities enhances resilience against emerging network-based attacks while simultaneously improving system reliability, modularity, and performance, all with minimal cost and 3D integration overhead.**

I. INTRODUCTION

In recent years, AI systems have been deployed across a wide range of industries and use cases. The use of edge AI grew rapidly across a spectrum of applications from surveillance systems, to autonomous vehicles, medical devices, wireless communication networks, IoT systems and drones [61][62][63]. This surge in adoption has been accompanied with a corresponding rise in the frequency, variety and severity of attacks targeting edge AI systems [56].

Unlike cloud-based centralized AI, edge AI systems face considerable challenges due to the limitations in network bandwidth, computational and learning resources. The growing list of attacks includes adversarial attacks on sensor inputs, such as manipulating images and LiDAR signals to deceive the model, denial-of-service attacks designed to overwhelm the system, sensor jamming that disrupts the system's sensing capabilities through noise and signal interference, data poisoning and others. Due to limitations in system resources and communication bandwidth, such attacks have a more debilitating impact in edge AI systems, potentially leading to catastrophic consequences.

In addition to the growing security challenges, edge AI systems also face an increasingly complex regulatory landscape [65-69] [64]. Scaling the regulatory and audit capabilities to the range and diversity of edge AI systems and tackling the corresponding complexity locally and in real-time is a key obstacle [64].

In this study, we propose an AI safety architecture, 3D Guard-Layer, that employs 3D integration to tackle emerging edge AI safety challenges. This infrastructure provides shadow AI models and local processing capabilities to address resource and communication limitations respectively. It enables continuous safety monitoring along with failover capabilities through shadow AI models.

By processing input and output streams of the primary edge AI system, the 3D Guard-layer allows continuous behavioral and fine-grain hardware monitoring and anomaly detection to identify potential risks or threats. The agentic architecture enables seamless failover capabilities to mitigate ongoing attacks, along with safety and regulatory agents to locally process input streams, preventing downtime and compliance issues during network failures.

In addition to expanding the computing resources in space-constrained embedded applications, 3D architecture offers significant advantages in fine-grained monitoring by fully aligning floorplans and enabling high-degrees of interconnectivity. It requires minimal modifications to the original edge AI chips, enables heterogeneous layer integration, increases transistor density, as well as improving total system power and costs [19][27][29].

This paper is organized as follows: Section II provides background on emerging AI safety and security challenges impacting edge AI systems, Section III describes the attack detection techniques, Section IV discusses the proposed 3D safety architecture and capabilities, Section V describes the

agentic AI capabilities, Section VI provides 2.5D and 3D configurations, Section VII describes the cost and overhead analysis for edge AI systems. Section VIII highlights the advantages of the 3D guard-layer architecture, and finally Section VIII provides concluding remarks.

II. Emerging Edge AI Vulnerability & Attacks

In recent years, the AI attack landscape has evolved rapidly, with an expanding array of increasingly sophisticated threats. This section aims to highlight a selection of recent attack types, illustrating the breadth and diversity of emerging threats with implications on edge AI systems.

Edge AI system architectures typically include multiple layers though the complexity within the layers vary significantly. All layers are subject to a growing number of attacks.*(i) Application Layer:* AI service layer for specific applications such as health, smart vehicles, smart city applications. In more complex edge AI systems, application programming interfaces, middleware, service and device manager layers are included. The safety challenges in this layer may involve API calls to the applications through HTTP protocol. *(ii) Network Layer:* This layer captures network connectivity and vulnerabilities over RFID, WiFi, 4G/5G, Bluetooth through protocols like HTTPS, common IoT protocols like MQTT. DDoS and other network attacks primarily influence this layer through encryption, authentication and other security vulnerabilities. *(iii) Device Layer:* Device layer involves the user and environmental interactions such as sensors, actuators and other data streams. Device or edge layer involves a number of safety vulnerabilities that range from Sybil attacks to jamming and node failure attacks.

A.Adversarial Attacks on Object Recognition

Attacks on object recognition systems have been among the earliest threats to edge AI systems. These attacks frequently involve introducing subtle perturbations to the data, leading to misclassifications and incorrect decisions. As widely demonstrated in autonomous vehicles, altering the traffic signs through pixel-level perturbations causes the AI system to make incorrect decisions, thereby compromising safety[48].

B. Spoofing Attacks

Spoofing attacks target edge AI system components such as global navigation satellite system (GNSS), global positioning system (GPS), and Lidars [57][50]. These attacks frequently hijack autonomous vehicles as well as drone systems [49].

C. Large Language Model Manipulations

The increasing integration of LLMs into edge AI systems have caused a rise in manipulation attacks such as LLM jailbreaking [81], prompt injection [80] in recent years. These attacks direct the edge AI systems to diverge from safe behaviors and by-pass existing guardrails [61].

D. Multi-Modal Attacks

With the introduction of multi-modal AI systems, developing consistent security measures and guardrails across modalities has emerged as a serious problem. The introduction of a widened attack surface resulted in new and highly effective attack types such as best-of-N (BoN) jailbreaks [47] [52] [72].

E. Network-based Attacks

Network-based attacks on edge AI include classic attack types like man-in-the-middle attacks [74], denial of service attacks (DDoS) [73], communication interception [31], malware and ransomware attacks [86], compromised software updates [75], unauthorized access through password cracking and others [53][54]. Similarly, Sybil attacks, in which an attacker undermines the network's reputation system through a large number of pseudonymous identities use this to mislead the edge system and direct it to behave abnormally have significant implications for edge AI systems [58].

F. Data Poisoning Attacks

Though it has characteristics similar to the network-based attacks, data poisoning is emerging as a growing category of its own [77]. Through data poisoning attacks, the edge system is fed poisoned data such as false sensor data to lead to unsafe decisions and manipulation by the adversaries [55].

G. Hardware Attacks on Edge AI

Hardware attacks, such as electromagnetic fault injection attacks, voltage glitching, hardware Trojans, and side-channel attacks, are significant threats to edge AI systems. While EMFI and voltage glitching can induce faults by manipulating electromagnetic fields and power supply to produce incorrect results or manipulate the inference process in deep neural networks respectively [80], hardware trojans, inserted during manufacturing, can alter system behavior when triggered [81]. Similarly, side-channel attacks exploit physical emissions, such as power consumption or electromagnetic radiation, to extract sensitive information from AI models [82].

III. Attack Detection Techniques

Though the attack types vary significantly attack detection techniques frequently utilize common capabilities such as cross-verification, ensembling and shadow processing, anomaly detection

Object recognition attacks are detected through a number of techniques. Ensemble methods, utilizing multiple classifiers to cross-verify results, facilitate the detection of attacks by identifying inconsistencies across models. Input validation techniques, including preprocessing filters and robustness testing, enable the identification of perturbations in images. Anomaly detection, through statistical modeling of image patterns and tracking of input provenance, further supports the recognition of tampered inputs.

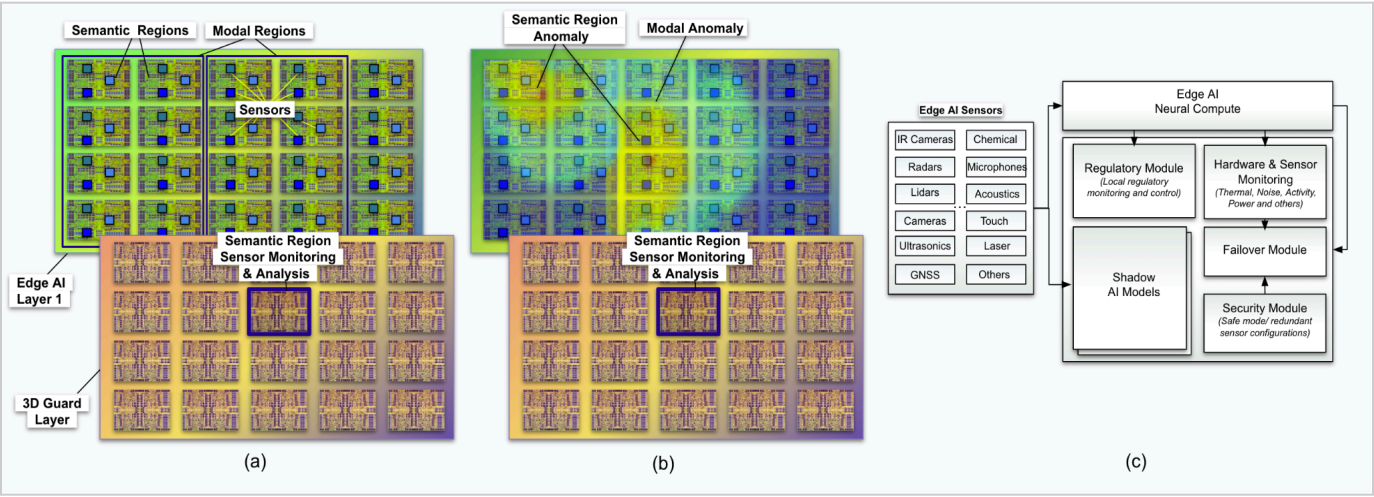


Fig.1 (a) Sensor infrastructure on edge AI layer (top) with semantic, feature and modal analysis, sensor analysis and anomaly detection region in 3D Guard-layer (bottom) (b) Semantic, feature and modal anomaly in edge AI layer (c) High-level architecture of the 3D Guard-Layer with shadow AI models, dedicated regulatory, security and failover modules as well as hardware and sensor monitoring. The guard layer directly monitors all layers

Similarly, spoofing attacks can be detected through anomaly detection models to identify deviations from typical input patterns, flagging suspicious data. Sensor cross-validation across multiple sensors, such as cameras and accelerometers, help identify inconsistencies that may indicate spoofing. Additionally, behavioral analysis of interaction patterns can detect unusual user or device behavior, further enhancing the ability to identify malicious activity. These methods work together to improve the robustness of edge AI systems against spoofing attacks.

LLM attack detection involves several techniques, including input anomaly detection, where unusual text patterns or statistical deviations from normal inputs are flagged. Behavioral monitoring identifies inconsistencies or harmful outputs, such as contradictions or biased responses, indicative of adversarial manipulation.

Network attacks on edge AI systems are frequently detected using a combination of anomaly detection, intrusion detection systems, and behavioral anomaly detection. Anomaly detection compared to baselines of normal network traffic, allows the detection of deviations indicative of potential attacks. In addition, some of the network based attacks also yield hardware behavior anomalies, such as overactivity in communication modules, abnormal processing patterns in malware, ransomware and DDoS attacks etc.

IV. 3D GUARD-LAYER FOR EDGE CO-PROCESSING

This section describes the 3D Guard-layer architecture aiming to improve the overall safety of edge intelligence systems. The 3D Guard-layer is a vertically integrated layer within the primary edge AI system such that it provides: (i) real-time local monitoring, (ii) shadow processing, (iii) failover and (iv) regulatory processing capabilities.

The 3D Guard-layer architecture aims to provide the basic hardware infrastructure support to detect and remediate a wide range of attacks enabling a wide range of safety monitoring, attack detection and remediation techniques. 3D Guard-layer variants can be further customized for specific application families and integrated in a modular fashion.

A. Hardware Infrastructure for Real-Time Monitoring and Anomaly Detection

1) System-level Behavioral Monitoring: 3D Guard-layer architecture uses a shadow AI layer to continuously monitor the behavior of the primary AI chip(s) for safety threats and behavioral anomalies. Co-location of the guard-layer provides direct access to the activity on primary computing layers for anomalous behavioral pattern detection. As an example, in an autonomous vehicle use case driving patterns may be tracked to detect anomalous events by the shadow layer.

2) Hardware Monitoring: The system leverages the sensor monitoring infrastructure on the main AI chip for hardware monitoring. This may further be enhanced by additional sensor support on the primary AI system such as block-level activity monitors, power, noise and thermal sensors. As shown in Fig.1(a) the semantic tiles incorporate hardware sensors in the primary edge AI layer(s), which are tightly interconnected through TSV and uC4 to the adjacent 3D Guard-layer with sensor monitoring and anomaly detection regions. Common attack types such as DDoS, flooding and malware attacks can be detected through the anomalous thermal sensor readings and activity monitor patterns.

3) Semantic Monitoring: The popularity of architectures with semantic specialization such as YOLO [26], Deeplab [11], faster R-CNN [96] and others provides the opportunity to monitor semantic block, region level and modal monitoring to detect anomalies that involve activations of incorrect

modalities or semantic regions during attacks. As shown in Fig.1(b) anomalies in semantic regions or modalities are picked up by the sensor networks in edge AI layers and the semantic monitoring and analysis regions in 3D Guard-layer.

Recent attack types can be detected through anomalous computational demands or excessive activity in specific regions. For example, multimodal jailbreaking induces sudden spikes in computational load in specific modalities, while poisoning attacks result in abnormal computational demands across the system [72]. Indirect prompt injections embedded in image or audio inputs often lead to unexpected and disproportionate activity within the corresponding modal regions [78]. Poisoning, malware, and takeover attacks induce abnormal behaviors in unusual regions. While anomalous patterns exhibit in a range of hardware configurations, semantic monitoring provides more insights in neuromorphic hardware. Similarly, the infrastructure enables feature analysis by examining activation patterns and conducting feature space comparisons to identify discrepancies in the recognition process.

B. Hardware System Support for Shadow Processing to Mitigate Attacks and Failover

In the event that an on-going attack or a security compromise is detected, the 3D guard-layer provides infrastructure support for remediation techniques.

1) Shadow AI Models: Shadow AI models are integrated in the 3D security layer serving functions such as:

(i) Shadow Co-Processing for Safety: In some configurations shadow models use alternative redundant sensors for co-processing to detect anomalous behaviour. As shown in Fig.2 the 3D Guard-layer architecture includes monitoring, failover, security and regulatory modules as well as shadow AI models that input sensory data from sensor devices associated with the edge AI system. For higher-criticality events, and hardware attacks the intervention is shadow chip taking over the decision making processes over the primary chip. As edge AI systems are equipped with redundant sensors, the system reconfigures the use of active sensors to mitigate the on-going attack.

(ii) Confidence Scoring: Stochastic nature of large language models present significant safety challenges, which the 3D Guard-Layer architecture addresses by utilizing shadow and ensemble models to detect inaccuracies through entropy-based divergence quantification. Specifically, semantic entropy can be employed to estimate the entropy of meaning distributions in free-form responses, enabling the identification of incorrect or low confidence outputs [59].

(iii) Learning Attack Patterns: Shadow AI models provide continuous learning of attack patterns over time as well as coordinating with backend cloud-based AI systems to track novel attack types, security vulnerabilities and remediation techniques to be implemented directly in the learning hardware. Privacy preserving federated learning solutions are also enabled through the distributed learning infrastructure.

2) Failover Capabilities: The 3D Guard-layer architecture includes failover capabilities through a dedicated failover agent. In the event of a model compromise, shadow models are used to detect model compromise through anomalous pattern analysis. The corresponding agent changes the system configuration to replace the primary edge AI model(s) with shadow models to thwart the ongoing attack. This capability supports full offline processing capabilities for DDoS and other network based attacks.

C. Regulatory Processing AI Agents

With the introduction of new AI regulations around the globe, the need to perform real-time regulatory monitoring has become a key capability for AI systems. Due to their mobility, edge AI systems may be used in and move through a wide range of jurisdictions with different regulatory requirements. The 3D Guard-layer provides real-time regulatory monitoring, audit and intervention capabilities for both AI and application-specific regulations through shadow processing and continuous monitoring. While the primary model(s) process the data and make decisions, shadow regulatory agent(s) continuously monitor and remediate potential compliance issues.

D. 3D Configurations

The proposed system can be implemented through (i) silicon interposers and (ii) 3D configurations to capture the varying degrees of complexity of edge AI systems. Current edge AI systems span the full range from single chip systems to highly complex multi-chip modules. Moreover, 3D ICs are commonly used in edge AI scenarios. This provides an opportunity to integrate 3D Guard-layer with the existing 3D stacks. The 3D floor plan alignment and co-location provides advantages in seamless failover mode activation, shadow security processing, and fine-grain sensory analysis. Through-silicon-via or micro-C4 based access to the main processor layer provides monitoring capabilities at a fine grain. The 3D Guard-layer can be integrated through face-to-back (guard-on-top/edge-on-top) and face-to-face configuration (as shown in Fig.2). 2.5D interposers provide support for existing multi-chip modules (Fig.3).

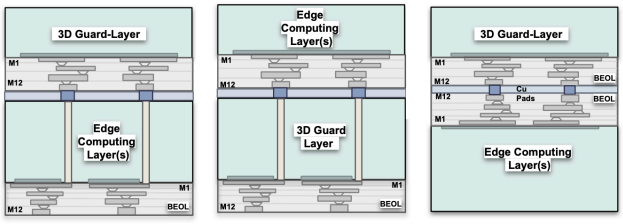
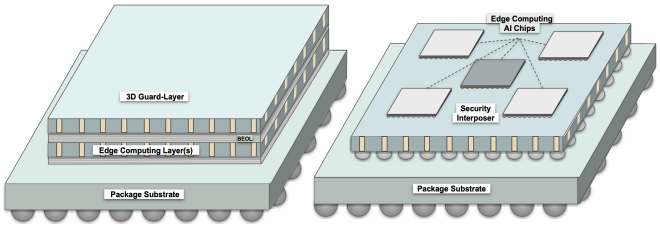


Fig.2 (i) Face-to-Back 3D Guard-layer configuration with edge intelligence layer codesign (ii) Face-to-Back configuration with limited changes to edge intelligence layers (iii) Face-to-Face Configuration

As shown in Fig. 3 the edge AI chips can be integrated with the guard-layer interposer for sensor-based monitoring, passive and active redundant models, and failover hardware.



(i) (ii)

Fig.3 (i) 2.5D Silicon interposer configuration with multiple edge intelligence chips and a dedicated Guardlayer interposer connected to all AI chips (ii) 3D Guard layer configuration with one or more edge intelligence layers and stacked 3D guardlayer.

VI. AGENTIC AI IMPLEMENTATION

This section presents the agentic implementation of a 3D GuardLayer configuration. The framework incorporates 5 agents: (i) Behavioral monitoring agent, (ii) Hardware monitoring agent, (iii) Shadow Processing Agent, (iv) Failover System Agent, (v) Regulatory Compliance Agent. The agentic AI framework details are described below.

```
# Agent 1: Behavioral Monitoring Agent
Agent role ← Behavioral Monitoring Specialist
goal ← Monitor the primary AI system for behavioral anomalies and safety threats
if (agent in [BehavioralMonitoringAgent]) then
  agent = context(instruction) ← monitor_behavior
  tool = [anomaly_detection_algorithm, alert_system]
end if
End Agent

procedure (Behavioral Monitoring)
  role ← Behavioral Monitoring Specialist
  goal ← Detect deviations from expected behavior in the AI system
  tool = [anomaly_detection_algorithm, alert_system]
  task ← monitor_behavior
end procedure

# Agent 2: Hardware Monitoring Agent
Agent role ← Hardware Monitoring Specialist
goal ← Detect hardware-level anomalies in AI system (thermal, power, noise)
if (agent in [HardwareMonitoringAgent]) then
  agent = context(instruction) ← monitor_hardware_activity
  tool = [sensor_data_analysis, hardware_alert_system]
end if
End Agent

procedure (Hardware Monitoring)
  role ← Hardware Monitoring Specialist
goal ← Detect anomalous hardware behavior such as abnormal power consumption, thermal spikes, or noise
tool = [sensor_data_analysis, hardware_alert_system]
task ← monitor_hardware_activity
end procedure

# Agent 3: Shadow AI Model Agent
Agent role ← Shadow AI Model Specialist
goal ← Take over the decision-making process in case of anomalous behavior or attack
if (agent in [ShadowAIModelAgent]) then
agent = context(instruction) ← trigger_shadow_intervention
tool = [shadow_model_activation, redundant_sensors]
end if
End Agent

procedure (Shadow Processing)
role ← Shadow AI Model Specialist
goal ← Take over decision-making if the primary AI system is compromised
tool = [shadow_model_activation, redundant_sensors]
task ← trigger_shadow_intervention
end procedure
```

```
# Agent 4: Failover System Agent
Agent role ← Failover System Specialist
goal ← Enable failover from primary AI system to backup systems in case of security breach
if (agent in [FailoverSystemAgent]) then
agent = context(instruction) ← initiate_failover
tool = [failover_module, offline_mode]
end if
End Agent

procedure (Failover)
role ← Failover System Specialist
goal ← Ensure smooth transition to backup models and offline mode during attacks
tool = [failover_module, offline_mode]
task ← initiate_failover
end procedure

# Agent 5: Regulatory Compliance Agent
Agent role ← Regulatory Compliance Officer
goal ← Ensure continuous compliance with industry-specific regulations
if (agent in [RegulatoryComplianceAgent]) then
agent = context(instruction) ← monitor_compliance
tool = [compliance_checker, audit_logging]
end if
End Agent

procedure (Regulatory Compliance)
role ← Regulatory Compliance Officer
goal ← Continuously monitor the AI system for compliance with industry regulations
tool = [compliance_checker, audit_logging]
task ← monitor_compliance
end procedure

# Agent orchestration: A security threat is detected, and the agents work together
if (anomaly_detected or hardware_anomaly_detected):
trigger(BehavioralMonitoringAgent) ← monitor_behavior
trigger(HardwareMonitoringAgent) ← monitor_hardware_activity
if (anomaly_detected):
trigger(ShadowAIModelAgent) ← trigger_shadow_intervention
trigger(FailoverSystemAgent) ← initiate_failover
end if
trigger(RegulatoryComplianceAgent) ← monitor_compliance
end if
```

Fig.5 Agentic AI implementation of the 3D Guard-layer system.

VII. 3D GUARD-LAYER SYSTEM-LEVEL BENEFITS

The proposed 3D Guard-layer architecture provides a number of key system-level benefits in edge AI systems. This section discusses a selection of 3D Guard-layer advantages.

A. Real-Time Response to Attacks

Current edge AI systems primarily rely on cloud-based backend systems for security and safety matters as they possess limited safety capabilities beyond encryption due to the local resource limitations. The resulting reliance on the cloud systems and communication networks introduces associated delays, thereby limiting the AI system's response time to emerging attacks. It increases the safety risks and catastrophic failures due to these extended response times. The 3D Guard-layer architecture provides real-time local detection and safety capabilities as well as protection from network failure and network related attacks.

B. Reduced Network Attack & Failure Risk

3D Guard-layer reduces communication delays and demands by locally processing safety related tasks. Given that a large percentage of edge attacks exploit communication channels, this improves safety and provides resilience against network attacks Moreover, the system exhibits enhanced resilience to network failures and outages, improving the operations and safety during such events.

C. Improved Privacy & Data Protection

Network-based attacks also pose significant threats to data protection and privacy. The local data processing reduces the need to transmit sensitive data to centralized servers. This results in improved data privacy and protection which is critically important in high-security use cases such as medical applications.

E. Modularity & Customization of Local Safety Capabilities

Edge AI systems span a broad spectrum of applications making them vulnerable to a variety of attack types. However, customizing edge AI systems for specific attack types for an application is costly, which results in reliance on centralized cloud infrastructures. The proposed system offers a cost-effective solution for integrating custom AI safety features into edge devices, tailored to diverse operational and environmental conditions. By utilizing baseline 3D infrastructure or silicon interposer support, a range of safety capabilities can be seamlessly incorporated into individual edge AI systems.

F. Deep Inspection of Attacks

3D stacking enables fine-grain monitoring of the underlying computation in the hardware at a block or macro-level through a high-bandwidth interconnect to power, thermal, and activity sensors. This provides deeper insights on the source and semantic details of the attacks, modality of the attacks, risky objects for recognition attacks and other characteristics.

G. System Reliability and Performance Improvement

The 3D Guard-layer serves as an embedded redundancy for learning hardware attacks and failures by enabling the redirection of the computation from the AI layer to shadow guard layer or layers. This improves system-level reliability, robustness - key to applications like cybersecurity, medical systems, military systems and other mission critical system infrastructure. Furthermore, 3D Guard-layer resources can be

used to boost performance in adversarial environments, highly unpredictable scenarios with changing system demands.

H. Resilience Against Hardware Attacks

Hardware-based attacks, such as electromagnetic fault injection, hardware Trojans, voltage glitching, and side-channel analysis, have proven effective in compromising edge AI systems. Guard-layer hardware provides the ability to shield the critical device layers from electromagnetic radiation commonly used in trojan triggering, EMFI, side channel and other attacks. The 3D vertical power grid enables protection from power glitching attacks.

I. Post-Quantum Hardware Demand and Side-Channel

Increased computational demands of post-quantum cryptography techniques such as recently standardized Dilithium and Kyber, increases the demand for hardware support on the device. Furthermore, these algorithms require additional protections due to their underlying vulnerabilities to side-channel attacks [10] [38] [54]. The side channel vulnerabilities of PQC standards like Kyber and Dilithium due to their lattice characteristics include power analysis, timing analysis and memory access patterns, the 3D guard layer becomes an essential component for hardware support for advanced cryptography techniques.

VII. COST & OVERHEAD ANALYSIS

This section presents a cost and overhead analysis for the 3D Guard-layer architecture and its possible configurations. As discussed earlier, the 3D Guard-layer can be implemented in a range of baseline edge AI systems with single chip, multi-chip, 2.5D and 3D packages. The system itself provides flexible implementation through 2.5D silicon interposer, face-to-face and face-to-back 3D stacking. Table 1 shows a list of commonly used commercial edge AI systems with a range of CPU, GPU, TPU, AI accelerator and neuromorphic architectures.

In terms of 3D vertical interconnect overhead, 75% of the configurations, namely (i) 2.5D, (ii) 3D face-to-face and (iii) 3D face-to-back edge-on-top configurations do not incur any TSV overhead. While these configurations are preferred, TSV overhead for the face-to-back guard-on-top configuration can be estimated as shown in Table 1. Using state-of-the-art commercial 3D technology assumptions such as 3µm size and 10µm pitch, Table 1 provides an interconnectivity sweep from 10K to 100K.

System	Package	Architecture	Area (mm^2)	10K TSV %	20K TSV %	50K TSV %	100K TSV%	Tech Node
System 1 [93]	MCM	GPU + CPU	5133	0.00%	0.00%	0.01%	0.01%	8nm
System 2 [94]	MCM	GPU + CPU	3107	0.00%	0.00%	0.01%	0.02%	8nm
System 4 [96]	SCM	Neuromorphic	400	0.02%	0.04%	0.09%	0.18%	7nm/ 16nm
System 5 [97]	MCM (3D*)	FPGA + CPU + AI Engines	2025	0.00%	0.01%	0.02%	0.04%	7nm
System 6 [98]	SCM	Hybrid (CPU, GPU, DLAs)	289	0.02%	0.05%	0.12%	0.24%	16nm
System 7 [99]	SCM	CPU + GPU + NPU	196	0.04%	0.07%	0.18%	0.36%	14nm
System 8 [100]	SCM	Neuromorphic	800	0.01%	0.02%	0.04%	0.09%	12nm
System 9 [101]	MCM/ SCM	ASIC (Tensor Processing Unit)	150	0.05%	0.09%	0.24%	0.47%	12nm

Table.1 Through-silicon-via overhead estimations for the only configuration with overhead (F2B 3D configuration) for commonly used edge AI hardware systems in the market

3D manufacturing cost overheads depend on the baseline 3D edge AI system hardware manufacturing costs. The original system costs for the most common edge AI hardware systems exhibit a 20x variation. Similarly, these baseline system packages span multi-chip modules (MCM) and single-chip modules (SCM). As discussed earlier, using the 2.5D guard-layers for MCMs and 3D for SCMs, 3D guard-layer cost overheads range from 3.75% for high-end to mid-range edge computing systems to close to 60% in the lowest cost options due to the low price point of the baseline system. With the rapidly growing complexity of edge AI systems, this manufacturing cost overhead is likely to decline.

VIII. CONCLUSIONS

AI is rapidly adopted in edge computing systems, which caused a similar growth in the number and diversity of attacks. Edge AI hardware and communication resources already overutilized by rapidly growing AI demands face compounding challenges from safety risks and emerging attacks.

This paper proposes a novel agentic AI architecture, 3D Guard-layer, that utilizes 3D integration and agentic AI capabilities to tackle the critical issues in edge AI safety. It introduces key infrastructure capabilities, fine-grain real-time monitoring shadow AI models, offline/failover mode detect and mitigate a wide range of attacks. The infrastructure also provides benefits such as improved system reliability, modularity, speed and performance improvements with minimal cost and 3D overheads.

REFERENCES

[1] Kursun E., Three Dimensional Integration Technology: A Micro-architectural Outlook, IBM Research Technical Report, Report, RC24147, 2007

[3] Knickerbocker J.U. et.al.,Three Dimensional Silicon Integration Using Fine Pitch Interconnection, Silicon Processing and Silicon Carrier Packaging Technology, IEEE Custom Integrated Circuits Conference, 2005

[4] Black B., Annavaram M., Brekelbaum E., DeVale J., Jiang L., Loh G.H., McCauley D., Morrow P., Nelson D.W. , Pantuso D. , Reed P., Rupley J., Shankar S., Shen J. P., Webb C., Die Stacking (3D) Microarchitecture, International Symposium on Microarchitecture, 2006

[6] Zhang R., et.al Power Trends and Performance Characterization of 3-Dimensional Integration, Proceedings of IEEE International Symposium on Circuits and Systems, Vol.4, pp.414-417, 2001
[7] Reif R., Fan A., Chen K-N , Das S., Fabrication Technologies for Three-Dimensional Integrated Circuits, International Symposium on Quality Electronic Design, 2002
[9] Emma P., Kursun E., Is 3D Silicon the Next Growth Engine after Moore's Law,or is it Different?, IBM Journal of Research and Development
[10] Kursun E. et.al. Processor Emulator Design, IBM East Fishkill Technical Conference
[11] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L., Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE transactions on pattern analysis and machine intelligence, 40(4), pp.834-848, 2017.
[12] Emma, P., Buyuktosunoglu, A., Healy, M., Kailas, K., Puente, V., Yu, R., Hartstein, A., Kursun, E., Bose, P. and Moreno, J., 3D stacking of High-Performance Processors. In 2014 IEEE 20th International Symposium on High Performance Computer Architecture (HPCA) (pp. 500-511). IEEE.
[13] Zhou, H., Li, X., Cher, C.Y., Kursun, E., Qian, H. and Yao, S.C.,An Information-Theoretic Framework for Optimal Temperature Sensor Allocation and Full-Chip Thermal Monitoring. Design Automation Conference (pp. 642-647).
[14] Chen, Y., Kursun, E., Motschman, D., Johnson, C. and Xie, Y., 2013. Through Silicon via aware design planning for thermally efficient 3-D integrated circuits. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 32(9), pp.1335-1346.
[15] Emma, P.G., Kursun, E., Qureshi, M.K. and Srinivasan, V., International Business Machines Corp, 2013. Processor core stacking for efficient collaboration. U.S. Patent 8,417,917.
[16] Emma, P. and Kursun, E., 2009. Opportunities and challenges for 3D systems and their design. IEEE Design & Test of Computers, 26(5), pp.6-14.
[17] Kurshan, E., Li, H., Seok, M. and Xie, Y., 2020. A case for 3D integrated system design for neuromorphic computing and AI applications. International Journal of Semantic Computing, 14(04), pp.457-475.
[18] Gu, P., Stow, D., Barnes, R., Kursun, E. and Xie, Y., 2016, October. Thermal-aware 3D design for side-channel information leakage. In 2016 IEEE 34th International Conference on Computer Design (ICCD) (pp. 520-527). IEEE.
[19] Sun, G., Kursun, E., Rivers, J.A. and Xie, Y., 2013. Exploring the vulnerability of CMPs to soft errors with 3D stacked nonvolatile memory. ACM Journal on Emerging Technologies in Computing Systems (JETC), 9(3), pp.1-22.
[20] Kursun, E., Wakil, J. and Iyengar, M., 2010, June. Analysis of spatial and temporal behavior of three dimensional multi-core architectures towards run-time thermal management. In 2010 12th IEEE Itherm (pp. 1-8). IEEE.
[21] Kursun, E., Wakil, J., Farooq, M. and Hannon, R., 2012. Spatial and temporal thermal characterization of stacked multicore architectures. ACM Journal on Emerging Technologies in Computing Systems (JETC), 8(3), pp.1-17.
[22] Qian, H., Sapatnekar, S.S. and Kursun, E., 2012. Fast Poisson solvers for thermal analysis. ACM Transactions on Design Automation of Electronic Systems (TODAES), 17(3), pp.1-23.
[23] Bose, P., Buyuktosunoglu, A. and Kursun, E., International Business Machines Corp, 2009. Methods for thermal management of three-dimensional integrated circuits. U.S. Patent 7,487,012.
[24] Zou, Q., Kursun, E. and Xie, Y., 2017. Thermomechanical stress-aware management for 3-D IC designs. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 25(9), pp.2678-2682.
[25] Chen, Y., Kursun, E., Motschman, D., Johnson, C. and Xie, Y., Analysis and mitigation of lateral thermal blockage effect of through-silicon-via in 3D IC designs. In IEEE/ACM International Symposium on Low Power Electronics and Design (pp. 397-402).
[26] Jiang, P., Ergu, D., Liu, F., Cai, Y. and Ma, B., 2022. A Review of Yolo Algorithm Developments. Procedia Computer Science, 199, pp.1066-1073.
[27] Dennard, Robert H et.al (October 1974). "Design of ion-implanted MOSFET's with very small physical dimensions". IEEE Journal of Solid-State Circuits. SC-9 (5): 256–268
[28]Xu, Z., Pemberton, O., Roy, S.S., Oswald, D., Yao, W. and Zheng, Z., 2021. Magnifying side-channel leakage of lattice-based cryptosystems with chosen ciphertexts: The case study of kyber. IEEE Transactions on Computers, 71(9), pp.2163-2176.
[29] Knickerbocker, J.U.,et.al.. Three-dimensional silicon integration. IBM Journal of Research and Development, 52(6), pp.553-569.
[30] An, H., Ehsan, et.al. 2019. Monolithic 3D neuromorphic computing system with hybrid CMOS and memristor-based synapses and neurons. Integration, 65, pp.273-281.
[31] Zou, Y. and Wang, G., 2015. Intercept Behavior Analysis of Industrial Wireless Sensor Networks in the Presence of Eavesdropping Attack. IEEE Transactions on Industrial Informatics, 12(2), pp.780-787.
[32]Bose, P., Kursun, E., Rivers, J.A. and Zyuban, V., International Business Machines Corp, 2009. Method and arrangement for enhancing process variability and lifetime reliability through 3d integration. U.S. Patent Application 11/947,207.
[33]Cher, C.Y., Kursun, E., Maier, G.W. and Robertazzi, R.P., International Business Machines Corp, 2013. Three-dimensional (3D) stacked integrated circuit testing. U.S. Patent 8,542,030.
[34] Burns, J., Carpenter, G., Kursun, E., Puri, R., Warnock, J. and Scheuermann, M., 2011, June. Design, CAD and technology challenges for future processors: 3D perspectives. In Proceedings of the 48th Design Automation Conference (pp. 212-212).
[35] Jiménez, V., Cazorla, F.J., Gioiosa, R., Valero, M., Boneti, C., Kursun, E., Cher, C.Y., Isci, C., Buyuktosunoglu, A. and Bose, P., 2010, September. Power and thermal characterization of POWER6 system. In Proceedings of the 19th international conference on Parallel architectures and compilation techniques (pp. 7-18).
[36] Emma, P.G., Kursun, E., Qureshi, M.K. and Srinivasan, V., International Business Machines Corp, 2013. Processor core stacking for efficient collaboration. U.S. Patent 8,417,917.
[37]Bose, P., Emma, P.G., Kursun, E. and Rivers, J.A., International Business Machines Corp, 2013. Low overhead dynamic thermal management in many-core cluster architecture. U.S. Patent 8,595,731.
[38] Qian, H., Sapatnekar, S.S. and Kursun, E., 2012. Fast Poisson solvers for thermal analysis. ACM Transactions on Design Automation of Electronic Systems (TODAES), 17(3), pp.1-23.
[39] Dofe, J., Gu, P., Stow, D., Yu, Q., Kursun, E. and Xie, Y., Security threats and countermeasures in three-dimensional integrated circuits. Great Lakes Symposium on VLSI 2017 (pp. 321-326).
[40] Kursun, E., Reinman, G., Sair, S., Shayesteh, A. and Sherwood, T., 2005. Low-overhead core swapping for thermal management. In Power-Aware Computer Systems: 4th International Workshop, PACS 2004, (pp. 46-60).
[41] Emma, P.G., Kursun, E., Qureshi, M.K. and Srinivasan, V., International Business Machines Corp, 2013. Processor core stacking for efficient collaboration. U.S. Patent 8,417,917.
[42] Emma, P. and Kursun, E., 2010, June. 3D system design: A case for building customized modular systems in 3D. In 2010 IEEE International Interconnect Technology Conference (pp. 1-3). IEEE.
[43]Dennard, R.H. and Kursun, E., International Business Machines Corp, 2013. Power delivery in a heterogeneous 3-D stacked apparatus. U.S. Patent 8,473,762.
[44]Ravi, P., Chattopadhyay, A., D'Anvers, J.P. and Baksi, A., 2024. Side-channel and fault-injection attacks over lattice-based post-quantum schemes (Kyber, Dilithium): Survey and new results. ACM Transactions on Embedded Computing Systems, 23(2), pp.1-54.
[45]Vyas, A., et.al., 2020. Fast transformers with clustered attention. Advances in Neural Information Processing Systems, 33, pp.21665-21674.
[46] Al Maashri et.al.3D GPU architecture using cache stacking: Performance, cost, power and thermal analysis. In 2009 IEEE International Conference on Computer Design (pp. 254-259). IEEE.
[47] Hughes, J., Price, S., Lynch, A., Schaeffer, R., Barez, F., Koyejo, S., Sleight, H., Jones, E., Perez, E. and Sharma, M., 2024. Best-of-N jailbreaking.
[48] Chowdhury, A., Karmakar, G., Kamruzzaman, J., Jolfaei, A. and Das, R., 2020. Attacks on self-driving cars and their countermeasures: A survey. IEEE Access, 8, pp.207308-207342.
[49] Arteaga, S.P., Hernández, L.A.M., Pérez, G.S., Orozco, A.L.S. and Villalba, L.J.G., 2019. Analysis of the GPS spoofing vulnerability in the drone 3DR solo. IEEE Access, 7, pp.51782-51789.
[50] Eiza, M.H. and Ni, Q., 2017. Driving with sharks: Rethinking connected vehicles with vehicle cybersecurity. IEEE Vehicular Technology Magazine, 12(2), pp.45-51.

[51] Robey, A., Ravichandran, Z., Kumar, V., Hassani, H. and Pappas, G.J., 2024. Jailbreaking llm-controlled robots. arXiv preprint arXiv:2410.13691.
[52] Wu, C.H., Koh, J.Y., Salakhutdinov, R., Fried, D. and Raghunathan, A., 2024. Adversarial attacks on multimodal agents. arXiv e-prints, pp.arXiv-2406.
[53] Ahmad, F., Adnane, A., Franqueira, V.N., Kurugollu, F. and Liu, L., 2018. Man-in-the-middle attacks in vehicular ad-hoc networks: Evaluating the impact of attackers' strategies. Sensors, 18(11), p.4040.
[54] Dibaei, M., Zheng, X., Jiang, K., Maric, S., Abbas, R., Liu, S., Zhang, Y., Deng, Y., Wen, S., Zhang, J. and Xiang, Y., 2019. An overview of attacks and defences on intelligent connected vehicles. arXiv preprint arXiv:1907.07455.
[55]Chen, Y., Zhu, X., Gong, X., Yi, X. and Li, S., 2022. Data poisoning attacks in internet-of-vehicle networks: Taxonomy, state-of-the-art, and future directions. IEEE Transactions on Industrial Informatics, 19(1), pp.20-28.
[56]Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D., 2016. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
[57]K. Lim, K. M. Tuladhar, and H. Kim, ''Detecting location spoofing using ADAS sensors in VANETs,'' in Proc. 16th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC), Jan. 2019, pp. 1–4.
[58]A. Hakkala and O. I. Heimo, ''Automobile automation and lifecycle: How digitalisation and security issues affect the car as a product and service,'' in Proc. SAI Intell. Syst. Conf., Springer, 2019, pp. 121–137.
[59] Farquhar, S., Kossen, J., Kuhn, L. and Gal, Y., 2024. Detecting hallucinations in large language models using semantic entropy. Nature, 630(8017), pp.625-630.
[60] Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C. and Dai, D., 2024. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437.
[61] Deng, S., Zhao, H., Fang, W., Yin, J., Dustdar, S. and Zomaya, A.Y., 2020. Edge intelligence: The confluence of edge computing and artificial intelligence. IEEE Internet of Things Journal, 7(8), pp.7457-7469.
[62]Hua, H., Li, Y., Wang, T., Dong, N., Li, W. and Cao, J., 2023. Edge computing with artificial intelligence: A machine learning perspective. ACM Computing Surveys, 55(9), pp.1-35.
[63] Singh, R. and Gill, S.S., 2023. Edge AI: a survey. Internet of Things and Cyber-Physical Systems, 3, pp.71-92.
[64] FALCO, Gregory, et al. Governing AI safety through independent audits. Nature Machine Intelligence, 2021, 3.7: 566-571.
[65] Regulation (EU) 2024/1689 of the European Parliament and of the Council, 2024/1689, Official Journal of the European Union, 2024
[66] Cyberspace Administration of China (CAC). (2023). The Interim Measures for the Management of Generative Artificial Intelligence Services. Cyberspace Administration of China (CAC).
[67] Japanese Ministry of Economy, Trade, and Industry (METI). (2024). Japan's AI Implementation Guidelines. METI.
[68] Innovation, Science and Economic Development Canada. (2024). The Artificial Intelligence and Data Act (AIDA). Government of Canada.
[69] Personal Data Protection Commission of Singapore. (2020). Singapore AI Governance Framework. Latest Release 2020.
[70] Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., Liu, Y., Wang, H., Zheng, Y. and Liu, Y., 2023. Prompt Injection attack against LLM-integrated Applications. arXiv preprint arXiv:2306.05499.
[71]Xu, Z., Liu, Y., Deng, G., Li, Y. and Picek, S., 2024. LLM Jailbreak Attack versus Defense Techniques--A Comprehensive Study. arXiv e-prints, pp.arXiv-2402.
[72] Shayegani, E., Dong, Y. and Abu-Ghazaleh, N., 2023. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. arXiv preprint arXiv:2307.14539.
[73] Mirkovic, J. and Reiher, P., 2004. A taxonomy of DDoS attack and DDoS defense mechanisms. ACM SIGCOMM Computer Communication Review, 34(2), pp.39-53.
[74]Conti, M., Dragoni, N. and Lesyk, V., 2016. A survey of man in the middle attacks. IEEE communications surveys & tutorials, 18(3), pp.2027-2051.
[75] Karthik, T., Brown, A., Awwad, S., McCoy, D., Bielawski, R., Mott, C., Lauzon, S., Weimerskirch, A. and Cappos, J., 2016, November. Uptane: Securing Software Updates For Automobiles. In International Conference on Embedded Security in Car (Vol. 11).
[76] Kharraz A et.al.. Cutting the Gordian Knot: A Look Under the Hood of Ransomware Attacks. In Detection of Intrusions and Malware, and Vulnerability Assessment: International Conference, DIMVA 2015
[77] Steinhardt, J., Koh, P.W.W. and Liang, P.S., 2017. Certified Defenses For Data Poisoning Attacks. Advances In Neural Information Processing Systems, 30.
[78] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023, November). Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (pp. 79-90).
[79] B. Mahfouz, H. Le, P. Jamieson, and P. Choudhury, "Fault Injection in FPGA-Based Devices via Electromagnetic Radiation," *Proceedings of the 2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, Florence, Italy, 2018, pp. 1-4, doi: 10.1109/ISCAS.2018.8352048.
[80]K. Zhang, Y. Zhang, and R. Karri, "Voltage Glitching Attack on Deep Neural Networks: Vulnerabilities and Mitigation," *Proceedings of the 2020 IEEE European Symposium on Security and Privacy (EuroS&P)*, London, UK, 2020, pp. 195-210, doi: 10.1109/EuroSP48016.2020.00033.
[81] Warnecke, A., Speith, J., Möller, J.-N., Rieck, K., & Paar, C. (2023). Evil from Within: Machine Learning Backdoors through Hardware Trojans. *arXiv*.
[82]Xiang, Y., Chen, Z., Chen, Z., Fang, Z., Hao, H., Chen, J., ... & Yang, X. (2020). Open DNN box by power side-channel attack. IEEE Transactions on Circuits and Systems II: Express Briefs, 67(11), 2717-2721.
[83]Jetson AGX Orin Series: Data Sheet and Technical Overview, NVIDIA Corporation, 2022.
[84]Jetson Xavier NX Module Data Sheet, NVIDIA Corporation, 2020.
[85]Hailo-8 AI Processor: High-Performance, Power-Efficient Deep Learning at the Edge, Hailo Technologies Ltd., 2020.
[86]Versal AI Edge Series: Product Table and Technical Brief, Xilinx Inc., 2021.
[87]TDA4VM Processor: Jacinto™ 7 Family of Processors for ADAS and Vision Analytics, Texas Instruments Inc., 2020.
[88]i.MX 8M Plus Applications Processor: Multicore Arm®
[89]Cortex®-A53 with Neural Processing Unit, NXP Semiconductors N.V., 2021.
[90]NorthPole: A 12nm Energy-Efficient Digital Architecture for Brain-Inspired AI, IBM Research, 2023.
[91]Google Edge TPU: High-Efficiency Machine Learning at the Edge, Google LLC, 2019.