

Accelerating Mission-Critical AI: Architectural Strategies for Low-Power, Certifiable CNN Accelerators in Aerospace ASICs

Eren ALTUN^a

^aASIC & FPGA Expert in Aerospace, Toulouse, FRANCE

KEY WORDS

NPU Architecture,
Certifiable AI,
Quantization-Aware Design,
ANSI/VITA, OpenVPX 65.0,
Power-Efficient AI,
DO-254 for Complex Functions,

ABSTRACT

As the aerospace industry moves toward fully autonomous operations, the integration of high-performance Deep Neural Networks (DNNs) into flight-critical systems has become a primary objective. However, traditional GPU-based solutions often fail to meet the stringent power constraints of Unmanned Aerial Mobility (UAM) platforms and the rigorous determinism required for **DO-254** certification. This paper proposes an architectural framework for a **Certifiable AI Accelerator (NPU)** designed for mission-critical environments. We explore the transition from 'Black-Box' commercial architectures to transparent, hardware-centric designs that allow for formal verification of neural kernels. Key strategies discussed include **sparsity-aware data paths** and fixed-point quantization (INT8/INT16) to optimize the Power, Performance, and Area (**PPA**) triad. Furthermore, we detail the implementation of hardware-level 'safety wrappers' that monitor inference integrity in real-time. Finally, we demonstrate how such an ASIC can be seamlessly integrated into modular avionics systems through the ANSI/VITA **OpenVPX 65.0** standard, providing a scalable and certifiable roadmap for the next generation of edge intelligence in the cockpit.

1 Introduction

The aerospace industry is at a critical juncture. The promise of fully autonomous flight, advanced situational awareness, and real-time sensor fusion depends entirely on the deployment of Deep Learning algorithms at the edge. However, a significant technological barrier remains: traditional computing architectures are ill-equipped for the "Zero-Failure" environment of aviation. General-purpose GPUs and high-end CPUs often suffer from non-deterministic execution times and excessive power consumption a major hurdle for the limited thermal budgets of Unmanned Aerial Mobility (UAM) and tactical drones.

To overcome these challenges, the industry is shifting toward dedicated **ASICs** featuring custom **Neural Processing Units (NPU)**. Unlike commercial "*Black-Box*" AI chips, a certifiable *NPU* must be designed from the ground up with transparency and determinism as core requirements. This necessitates a hardware architecture that not only optimizes the **PPA** (Power, Performance, Area) triad through advanced quantization and data-flow management but also aligns strictly with the **DO-254** design assurance flow.

This white paper explores the architectural strategies required to build a certifiable AI accelerator. We will examine how to bridge the gap between complex neural network kernels and the hardware-level safety monitors necessary for **DAL-A** or **DAL-B** compliance. Furthermore, we discuss the strategic importance of integrating these AI-centric ASICs into modular systems via the **ANSI/VITA OpenVPX 65.0** standard, ensuring that the next generation of airborne intelligence is as robust as it is innovative.

2 Sparsity-Aware Architecture and PPA Optimization

In avionics and UAM platforms, energy efficiency is not just about battery life, it is a critical thermal management requirement. Standard AI computing architectures waste significant energy moving non-essential data. Optimizing the **PPA** triad relies on two pillars: Sparsity management and data-movement reduction.

2.1 Leveraging Sparsity-Aware Design

Deep Neural Networks are inherently "*sparse*" a large portion of weights and activations are zeros. A standard processor treats these zeros as normal data, wasting computation cycles and power.

Sparsity-Driven Computation: My NPU architecture detects zero-values at the MAC (Multiply-Accumulate) unit inputs. If a value is zero, the logic is bypassed or "gated," eliminating unnecessary switching power.

Hardware-Level Compression: By using compressed storage formats managed directly by the hardware (e.g., CSR - *Compressed Sparse Row*), we reduce the required memory bandwidth. Fewer data movements between RAM and compute cores result in a drastic drop in thermal output.

3 Data-Flow Optimization

The energy cost of moving data from external memory (DDR) is often 100 to 1000 times higher than the cost of the arithmetic computation itself, therefore the architecture prioritizes near-data processing through a massive local memory hierarchy based on SRAM integrated directly next to the neural cores to significantly reduce off-chip memory accesses while leveraging INT8 and INT16 quantization instead of FP32 in order to reduce silicon area by up to 4× and power consumption by up to 10×, while still maintaining inference accuracy sufficient for navigation and detection missions.

3.1 Impact on DO-254 Certification

Unlike GPU architectures where cache management is unpredictable, this optimized data-flow is deterministic. Every data movement is orchestrated statically or predictably by the hardware controller. This transparency is the key argument for certification auditors: we can prove the Worst-Case Execution Time (WCET) for every AI inference.

4 Hardware-Based Formal Verification for AI Kernels

In a **DAL-A** or **DAL-B** environment, "probabilistic" reliability is not enough. The auditor must see a deterministic proof that the hardware executing the AI will not enter an invalid state.

4.1 Formal Verification of Neural Engines

Standard simulation cannot cover the near-infinite state space of a neural network, so we rely instead on formal verification through property checking. By defining SystemVerilog Assertions (SVA), we establish mathematical proofs that for any possible input, the hardware-level MAC units and activation functions such as ReLU or Sigmoid strictly comply with their underlying mathematical models. In parallel, we formally demonstrate the independence of neural kernels by proving that a fault occurring in one neural layer cannot propagate to the ASIC control registers, thereby ensuring strict logical isolation and preserving system safety.

4.2 Real-Time Safety Monitors (Safety Wrappers)

Since the internal weights and activations of a CNN can evolve into highly complex and non-intuitive states, the AI core is surrounded by dedicated hardware safety monitors. These monitors act as a trusted boundary between the neural computation domain and the safety-critical flight control functions, continuously supervising the behavior of the accelerator at run time.

Range checkers enforce strict safety envelopes by verifying that neuron outputs and intermediate results remain within pre-defined, certified bounds if an anomaly is detected whether due to a transient fault, a radiation-induced upset, or an unexpected input pattern the offending signal is intercepted before it can propagate to the flight control system. In parallel, control flow integrity mechanisms supervise the execution sequence of the NPU, ensuring that each inference cycle follows the expected state progression, with no illegal shortcut, deadlock, or hang, thereby guaranteeing deterministic and fail-safe operation.

5 Integration into OpenVPX 65.0 Systems

To be useful, a certifiable AI ASIC must fit into existing aerospace ecosystems. The **OpenVPX 65.0** standard provides the necessary framework for high-speed interconnects and thermal management.

High-Bandwidth Interconnects: Leveraging PCIe Gen4/5 or 100G Ethernet profiles within the OpenVPX 65.0 backplane to feed sensor data (Lidar, Radar, 4K Cameras) to the ASIC with minimal latency.

Modular Scalability: Using the VITA 65.0 slot profiles allows for a "Plug-and-Play" certification approach, where the AI accelerator card can be upgraded without redesigning the entire mission computer.

5.1 Real-Time Safety Monitors (Safety Wrappers)

Since the internal weights of a CNN can be highly complex, the AI core is encapsulated by dedicated Hardware Safety Monitors that act as a final line of defense between the neural computation and the flight-critical domains. These monitors continuously supervise the numerical behavior of the network at run time, independently of the AI model itself.

Range checkers are implemented to validate that each neuron's output remains within predefined safety envelopes derived from system-level analyses. If an output exceeds these thresholds indicating a potential transient fault, numerical instability, or adversarial input the hardware trigger immediately intercepts or masks the signal before it can propagate to the flight control system. In parallel, Control Flow Integrity mechanisms supervise the execution sequence of the NPU, ensuring that all inference stages are executed in the correct order and timing, and that no unintended shortcut or hang can occur during the inference cycle.

References:

1. RTCA DO-254, *Design Assurance Guidance for Airborne Electronic Hardware*, 2000.
2. SAE ARP4754A, *Guidelines for Development of Civil Aircraft and Systems*, 2010.
3. Sze, V., et al., *Efficient Processing of Deep Neural Networks*, Morgan & Claypool, 2020, pp. 75-110.
4. Sze, V., et al., *Efficient Processing of Deep Neural Networks*, Morgan & Claypool, 2020, pp. 75-110.

5. VITA 65.0-2023, *OpenVPX System Standard*, VITA Standards Organization, 2023.
6. Koren, I. & Krishna, C.M., *Fault-Tolerant Systems*, Morgan Kaufmann, 2007, Ch. 3