

System Technology

"Under the Hood of Hardware Systems for Accelerator-Aware Workloads: From Lab to Fleet"

The rapid growth of AI-driven, memory-intensive, and data-centric cloud workloads is increasing the demand for sustained memory bandwidth, efficient data movement, and predictable performance at scale. However, conventional host-based control-plane architectures can introduce resource contention and scheduling interference, resulting in higher latency, performance jitter, and reduced workload determinism. **Solution Overview:** This paper presents a policy-driven accelerator offload framework that decouples data-plane execution from control-plane activities, enabling a scalable path from laboratory validation to fleet-scale production deployment. **Details:** The framework is evaluated using representative cloud workloads that emphasize memory-intensive processing and data movement. Accelerators perform bulk data transfers, streaming I/O, and in-path data transformations closer to the data source, reducing host overhead and improving throughput for analytics and AI workloads. To support predictable multi-tenant operation, the design incorporates resource partitioning, bandwidth governance, workload prioritization, queue-level policy enforcement, and congestion-aware throttling across compute, memory, accelerator, and network domains. The framework also includes a comprehensive reliability and observability architecture spanning user-space enablement, operating system integration, telemetry, platform diagnostics, structured error propagation, interrupt management, and standardized error reporting. **Key Benefits:** The proposed approach improves throughput, reduces latency and jitter, and enables predictable performance through policy-based resource isolation. Its key contributions include a scalable accelerator integration model that unifies compute, memory, and network offload; a comprehensive telemetry and reliability framework; and a robust error-resilience strategy that strengthens end-to-end system reliability, transforming tightly integrated accelerators from simple hardware offload engines into scalable, energy-efficient, and performance-isolated cloud services.

- Vidya Samyuktha / *Principal Hardware Systems Engineer* / **Microsoft**