

Zen Inference Accelerator — Datasheet

title: Glow™ — Datasheet

product: Glow™ Inference Accelerator

audience: integrating SoC / platform engineer

created: 2026-09-30

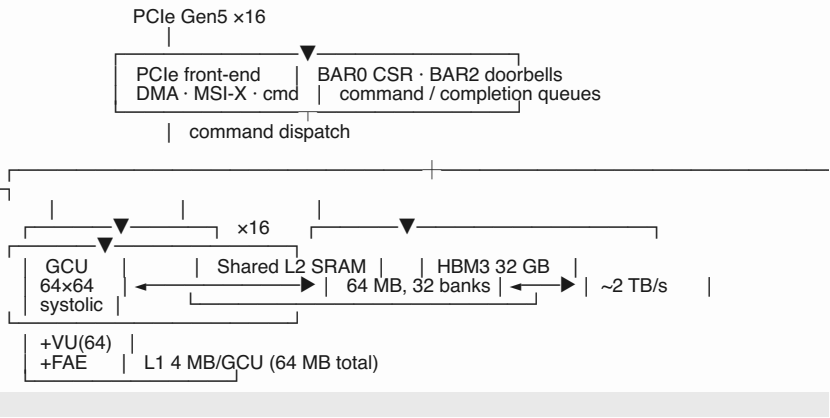
doc-type: datasheet

status: draft — pre-silicon

Glow™ — Datasheet

Product-level specification of the Glow inference accelerator: architecture, interfaces, configuration, performance, and deliverables. **Pre-silicon** — figures are design targets / performance-model outputs; uncharacterized values are **TBC**.

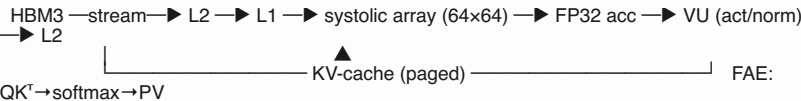
1. Block diagram



- **GCU (Glow Compute Unit) x16** — 64x64 weight-stationary systolic matmul array + 64-lane vector unit + 4 MB L1 scratchpad.
- **Flash Attention Engine (FAE) x16** — tiled attention ($Q \cdot K^T \rightarrow \text{online softmax} \rightarrow P \cdot V$) that never materializes the full score matrix.
- **Shared L2 (64 MB)** — weight cache + KV-cache residency + activation staging.
- **HBM3 (32 GB)** — full model weights and KV overflow.
- **PCIe front-end** — host command/completion queues, DMA, MSI-X, CSR.

2. Functional overview

- Glow serves the two phases of LLM inference in one device: **prefill** (compute-bound, parallel) and **decode** (memory-bound, latency-critical).
- The host submits work as a stream of fixed 128-bit commands via a ring buffer; a dispatcher routes matmul / attention / activation / norm / DMA operations to the GCUs and FAEs. Completions are reported per-fence via a completion queue and MSI-X.
- **Datapath (black-box):** weights stream from HBM3 $\rightarrow$ L2 $\rightarrow$ L1; matmul tiles execute weight-stationary in the systolic array with FP32 accumulation; attention runs on the FAE from L2-resident KV; activations/normalization run on the vector unit. The compiler decides tiling and L1/L2/HBM placement.



3. Interfaces & signals

Group	Summary
Host (PCIe Gen5 x16)	BAR0 = control/status registers; BAR2 = per-queue doorbells; DMA engine for host↔device transfer.
Command / completion	Host-resident ring buffers; 128-byte command descriptors; 32-byte completion entries with fence ID, status, error code.
Control / status registers	DEVICE_ID, REVISION, CAPABILITIES, NUM_GCUS, DEVICE_STATUS, ERROR_CODE, TEMPERATURE, POWER, UTILIZATION, MEM_* telemetry.
Interrupts	MSI-X, 256 vectors, edge-triggered; sources include completion, error, DMA-done, thermal-throttle, watchdog.
Memory	Internal L2 request/response and HBM3 request/response (512-bit data, byte-enables); not exposed to the licensee's SoC except at the device boundary.
Clocks / reset	Core clock 1.5 GHz; PCIe 100 MHz reference; active-low asynchronous reset with synchronized domain crossings.

Full register offsets, field layouts, and the command/completion descriptor formats are the frozen v1 ABI — see the [User Guide](#).

4. Configuration & profiles

Option	Binding	Values	Notes
Number of GCUs	Compile-time	16 (config of record)	Parameterized; 16 is the characterized configuration.
SRAM sizes	Compile-time	L1 4 MB/GCU, L2 64 MB	Fixed at elaboration.
Matmul tile	Compile-time	64 × 64 × K	Matches the systolic array + compiler tiling.
Precision	Per-command	FP8 E4M3 / INT8 / BF16	Selected in the command flags.
Attention mode	Per-command	non-causal / causal / paged-KV	Opcode variant.
Model size	Software	7B–70B tested; broader range targeted	Bounded by weight streaming + KV paging (TBC upper bound).
Batch size	Runtime	1–64 (continuous batching)	Scheduler load-balances across 16 GCUs.
SR-IOV VFs	Profile-optional	up to 8 (roadmap)	«CONFIRM with vendor».

5. Performance

Pre-silicon. Peak is an architectural ceiling; real serving throughput is lower and workload-shape dependent. Targets below are model outputs, not silicon measurements.

Metric	Value	Basis
Peak matmul throughput	196.6 TFLOPS @ 1.5 GHz	as-built single-issue systolic array, 16 GCUs (FP8/INT8/BF16)
Attention engine throughput	~49 TFLOPS/GCU-class	performance-model, not measured (TBC)
First-token latency (7B-class)	<10 ms (target)	design target — TBC
Sustained decode throughput	>1000 tok/s (target)	design target, HBM3 weight streaming — TBC
Perf/watt vs general-purpose GPU	meaningfully higher (target)	from on-chip attention + weight streaming + right-sized compute — TBC
Workload power (modeled)	~146 W prefill / ~79 W decode	pre-silicon power model — TBC

Wire-rate vs goodput: peak (196.6 TFLOPS) assumes fully-packed matmul; latency-critical serving (batch=1, decode) is memory-bound and runs well below peak by design — the trade is latency for utilization. Size deployments on the target latency/throughput figures, not peak.

6. Reliability, availability, serviceability (RAS)

- **Telemetry:** on-die temperature, power, and utilization readable via CSR.
- **Thermal protection:** thermal-throttle interrupt + power capping (thresholds TBC).
- **Watchdog:** watchdog-timeout interrupt for hung-command detection.
- **Error reporting:** per-completion status + global ERROR_CODE register.
- **Memory protection / ECC, DVFS profiles, detailed RAS policy:** TBC (to be characterized).

7. Standards & compliance

- **PCIe Gen5** ×16; **MSI-X** (256 vectors).
- **IOMMU / PASID, ATS;** page sizes 4 KB / 2 MB / 1 GB via resizable BAR.
- **SR-IOV** (optional, roadmap).
- **ML frameworks:** PyTorch, JAX, ONNX, StableHLO.

8. Deliverables

- Synthesizable RTL (compute / memory / attention).
- MLIR compiler toolchain + glow-compile driver.
- Linux kernel driver + libglow (C/C++) + Python runtime.
- Frozen ISA (128-bit fixed format) and register-map ABI specs.
- Per-block architecture specifications.
- Verification collateral (see doc #5) and a cycle-approximate performance model.

Exact packaging, deliverable tiers, and support are per license agreement («CONFIRM with vendor»).

Glow™ Inference Accelerator · Datasheet · © «CONFIRM: vendor legal entity» · pre-silicon draft.