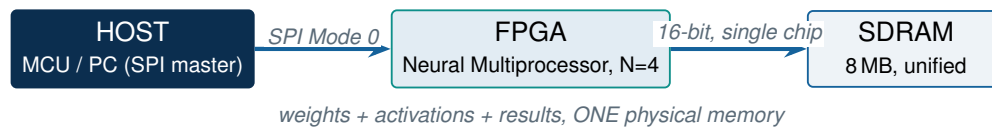


FPGA – Neural V2

INT8 Neural Multiprocessor, unified single-SDRAM architecture

Parametric N-processor hardware accelerator – Datasheet and reference manual

Rev. B0 • September 2026



Reference target device: Lattice ECP5 LFE5U-45F-8BG381 (speed grade –8, CABGA381, 72×MULT18X18D, ≈44k LUT).

Baseline configuration: INT8/INT32, N_PROCESSORS=4, P_IN=8, single external SDRAM Alliance Memory AS4C4M16SA-6TIN (weights + activations + results, unified). No PSRAM anywhere in the V2 physical path.

Clock: 16 MHz board oscillator → real ECP5 EHXPLL → 64 MHz system clock, timing-closed across 8 real place-and-route seeds (worst 68.51 MHz, best 74.17 MHz).

Host: real SPI protocol engine (`spi_host_bridge.v`) – WRITE_JOB / WRITE_MEM / READ_MEM / STATUS / RESET.

Status: RTL verified in simulation (Verilator, bit-exact), real synthesis (Yosys) and real place-and-route (nextpnr-ecp5). No fabricated hardware exists – see §8 for the precise, itemized readiness state. Document describing the project as of September 2026.

Project author: Michele Bigi • MIKILAB / manvalan.

This datasheet documents the RTL code, documentation and benchmarks present in the repository github.com/manvalan/FPGA-Neural, ported from the V1 datasheet's own LaTeX structure (same preamble and macros, content rewritten for V2).

FPGA-Neural V2 — General description and features

FPGA-Neural V2 is a **parametric hardware neural multiprocessor** built around `N_PROCESSORS` independent INT8 MAC engines dispatched by a real dependency-graph scheduler (Dependency Manager → Neural Director → per-slot Memory Manager), streaming weight/activation tiles from, and writing results to, a **single external SDRAM chip** – no PSRAM, no second memory device anywhere in the V2 physical path. Job registration is via a real SPI host protocol engine; the host never participates in the compute datapath.

Features

- **INT8 × INT8 → INT32** datapath per processor, `P_IN=8`-wide parallel MAC, balanced adder tree.
- **N_PROCESSORS** independent Neural Processors (frozen reference: `N=4`; `N=2` also fully validated; `N=8` is a future evolution).
- **Unified single-SDRAM memory**: weights, activations, AND results all share ONE physical Alliance Memory `AS4C4M16SA-6TIN` (8 MB, 16-bit), through one `sdram_unified_backend` arbitrating a read-only, cached weight (W) port and a read/write, byte-maskable activation+result (AR) port.
- Real dependency-graph scheduling: the Dependency Manager tracks per-node WAITING/READY/DISPATCHED state and producer/consumer wake-up; the Neural Director allocates the first free processor slot.
- **Real SPI host protocol engine** (`spi_host_bridge.v`): `WRITE_JOB` (job registration), `WRITE_MEM/READ_MEM` (raw SDRAM access), `STATUS`, `RESET`.
- **Real ECP5 clock generation**: 16 MHz oscillator → `EHXPLLL` (real Project Trellis `ecppll` parameters) → 64 MHz system clock.
- Real reset architecture (`reset_sync.v`): asynchronous assertion, synchronous deassertion, gated by external POR and PLL lock.
- Verified in **simulation** (Verilator, bit-exact) and real synthesis + place-and-route (Yosys + nextpnr-ecp5).

Applications

- Deterministic low-latency inference as an SPI peripheral of a microcontroller or SoC host.
- Reusable dependency-graph compute fabric, not a single fixed network topology.
- Edge AI on compact, INT8-quantized dense/DAG-structured networks.
- Off-loading neural workload from a constrained host CPU to dedicated hardware with predictable, measured throughput.

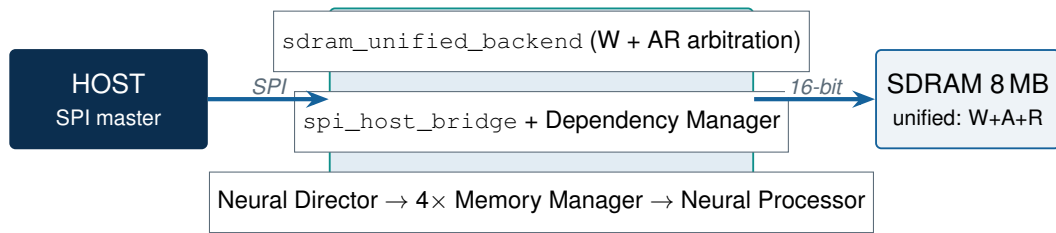
Target & toolchain

- FPGA: Lattice ECP5 `LFE5U-45F-8BG381` (–8, CABGA381).
- Synthesis: Yosys; place&route: nextpnr-ecp5.
- Simulation: Verilator (trusted per this project's own DEC-0004 protocol) and Icarus Verilog (cross-checked).
- SDRAM: Alliance Memory `AS4C4M16SA-6TIN` (4M×16, 8 MB).

Key parameters (frozen V2 reference configuration)

Quantity	Value	Notes
Data precision	INT8 (signed)	<code>DATA_WIDTH=8</code>
Accumulator	INT32 (signed)	<code>ACC_WIDTH=32</code>
Processors	4 (<code>N=2</code> also validated)	<code>N_PROCESSORS</code>
MAC width per processor	8	<code>P_IN=8</code>
Activation	ReLU + INT8 saturate	fixed, matches golden model
External memory	1 × SDRAM, 8 MB	<code>AS4C4M16SA-6TIN</code> , unified
System clock	64 MHz (real P&R, 8/8 seeds PASS)	16 MHz osc. → <code>EHXPLLL</code>
Bit-exact regression	256/256, <code>N=2</code> and <code>N=4</code>	Verilator, D-Stress workload
Host interface	real SPI (4 pins)	<code>spi_host_bridge.v</code>

System block diagram



Weights, activations and results all share the single physical SDRAM; the host only registers jobs and reads back results over SPI.

Contents

1	Overview	1
1.1	Document scope and status labels	1
1.2	Frozen reference configuration	1
2	Architecture	2
2.1	Compute: Neural Multiprocessor	2
2.2	Unified single-SDRAM memory	2
2.3	Read-timing discipline (ERR-0025)	3
3	Clock and reset architecture	4
3.1	Clock generation	4
3.2	Reset architecture	4
4	Host interface (SPI)	5
4.1	Opcodes	5
4.2	Verification	5
5	Pinout, place-and-route, and timing	6
5.1	Final pinout	6
5.2	Synthesis (real, this revision)	6
5.3	Place and route: 8-seed timing table	6
6	Power, SDRAM electrical, and configuration	8
6.1	Power architecture – status: OPEN	8
6.2	SDRAM electrical / timing sign-off – status: PARTIAL	8
6.3	Configuration – status: OPEN	8
7	Benchmarks	9
7.1	Internal benchmark: D-Stress workload	9
7.2	Host-interface pacing benchmark	9
7.3	Software reference comparison	9
8	Chip readiness and roadmap	11
8.1	Precise readiness checklist	11
8.2	Final answer	11
8.3	Roadmap	11

CHAPTER 1

Overview

FPGA-Neural V2 succeeds V1 with three architectural changes, all frozen and verified in this revision:

1. **Neural multiprocessor:** N_PROCESSORS (4, frozen reference) independent Neural Processors instead of V1's single engine, scheduled by a real dependency-graph Dependency Manager and Neural Director.
2. **Unified single-SDRAM memory:** one physical Alliance Memory AS4C4M16SA-6TIN SDRAM serves weights, activations, *and* results. V1's PSRAM (IS66WVE4M16EBLL-70BLI) and its controller are not part of V2's physical path at all.
3. **Real physical host interface:** a from-scratch SPI protocol engine (`spi_host_bridge.v`) replaces the internal 110-pin `reg_*` testbench bus as the board-level top's own physical interface.

V1 itself (`hardware/v1/**`) is untouched and remains the golden, independently certified reference design; V2 shares no RTL file with it.

1.1 Document scope and status labels

This datasheet distinguishes explicitly between:

- **RTL verified** – verified through RTL simulation only.
- **Simulation verified** – verified in simulation for the specific configuration named.
- **Synthesis verified** – synthesizes cleanly for the target FPGA (Yosys, real run, zero unintended latches or multi-driver issues beyond a small, previously-reviewed benign set).
- **P&R verified** – successfully placed, routed, and timing-analyzed by nextpnr-ecp5 for a real, ball-assigned LPF.
- **Hardware designed** – schematic/PCB electrically designed but not fabricated.
- **Hardware validated** – physically tested on fabricated hardware.
- **Silicon validated** – the actual physical FPGA has been programmed and tested.

As of September 2026, V2 is **RTL verified**, **simulation verified** (bit-exact), **synthesis verified**, and **P&R verified**. No item in this datasheet is presented as hardware- or silicon-validated – no fabricated V2 board exists. See Chapter 8 for the complete, itemized status.

1.2 Frozen reference configuration

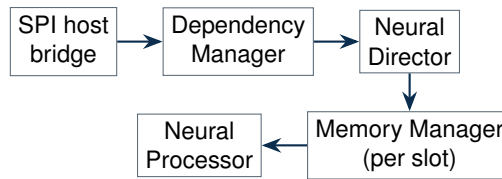
Parameter	Value
FPGA	Lattice LFE5U-45F-8BG381, speed grade -8, CABGA381
N_PROCESSORS	4 (frozen reference; N=2 also fully validated)
P_IN	8 (MAC width per processor)
ACC_WIDTH	32
External memory	1 × SDRAM, Alliance Memory AS4C4M16SA-6TIN, 8 MB
Oscillator	16 MHz (board-level, external)
System clock	64 MHz, generated by a real ECP5 EHXPPLL
Host interface	real SPI (4 pins), <code>spi_host_bridge.v</code>

CHAPTER 2

Architecture

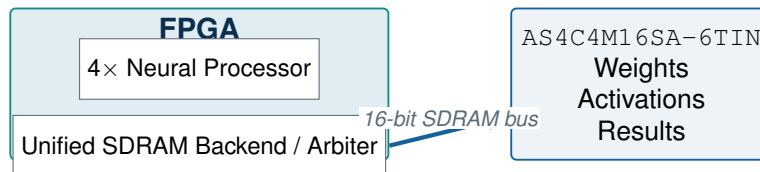
2.1 Compute: Neural Multiprocessor

Each of the `N_PROCESSORS` Neural Processors is an independent INT8 datapath: `P_IN=8` parallel multipliers feeding a balanced binary adder tree into a 32-bit accumulator, followed by ReLU and INT8 saturation – byte-for-byte identical to `neural_processor.v`, unmodified since before the V2 single-SDRAM freeze. Job dispatch is a real dependency-graph pipeline:



The Dependency Manager holds a per-node table (state, required/ resolved producer count, job descriptor fields) and hands READY nodes to the Director one at a time (valid/ready, backpressure-safe). The Director allocates the first free processor slot (first-free, not load-balanced) and frees it the instant that slot's job completes.

2.2 Unified single-SDRAM memory



`sdram_unified_backend.v` owns exactly one `sdram_controller.v` instance and presents two logical ports:

- **W** (weight fetch): 64-bit, read-only, a 4-entry fully-associative “other half” cache (round-robin eviction – safe under any sizing, since an evicted-too-early entry only costs an extra real fetch, never wrong data).
- **AR** (activation fill + result write-back): 16-bit, read/write, byte-maskable via real SDR SDRAM DQM semantics (`lb_n/ub_n`).

Both ports are, in turn, each arbitrated across the `N_PROCESSORS` slots by a generic, reused `slot_mem_arbiter` (a proven, pending-latch-based, single-owner-until-ready design used identically throughout this project).

Official V2 memory map

Region	Base address	Notes
Weights	0x010000	1 MB-aligned
Activations	0x200000	1 MB-aligned
Results	0x300000	1 MB-aligned

All three regions are non-overlapping within the single 8 MB (0x000000–0x7FFFFFF) SDRAM address space. Base addresses are host-programmable per job (via `WRITE_JOB`'s own `x_base/w_base/result_addr` fields), not hard-coded in the datapath.

2.3 Read-timing discipline (ERR-0025)

The shared weight and activation SRAMs (`nms_weight_packed.v`, `nms_activation_replicated.v`) use **combinational** read ports, matching the exact 1-cycle latency assumption of the per-slot Memory Manager's own pipelined read-ahead consumer. An earlier, registered-read implementation added one uncounted cycle of latency that a busy, multi-tile job's own prefetch lead time always absorbed invisibly, but that an uncontested single-tile job exposed on its first (only) tile – found and fixed via this revision's own board-level SPI integration testing, with zero regression to the existing bit-exact regression suite (identical cycle counts before and after the fix).

CHAPTER 3

Clock and reset architecture

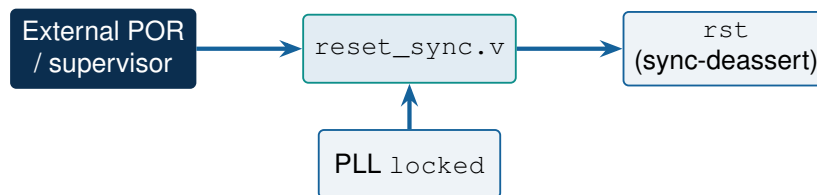
3.1 Clock generation



`ecp5_pll_sys_clk.v` instantiates a real EHXPPLL, with parameters generated by Project Trellis's own `ecppll` utility (not hand-derived): `CLKI_DIV=1`, `CLKFB_DIV=4`, `CLKOP_DIV=9`, `FEEDBK_PATH=CLKOP`, giving a real VCO frequency of 576 MHz (within the ECP5's documented 400–800 MHz range) and an exact, zero-error 64 MHz output ($16 \times 4/1$, divided by 9 at the VCO). 64 MHz was chosen over 80 MHz specifically because it is the highest frequency at which *all* measured P&R seeds close timing with real margin – see Chapter 5 for the full 8-seed table.

Simulation note: EHXPPLL has no open, licensable behavioral model (Lattice ships it only inside encrypted simulation libraries). `ecp5_pll_sys_clk.v` therefore provides a declared, simulation-only bypass under a ``SIM` define (`clk_sys` tied directly to `clk_16mhz`, `locked` tied high) – this is not, and does not claim to be, a simulation of real PLL lock timing.

3.2 Reset architecture



`reset_sync.v` is a standard async-assert/sync-deassert bridge: `rst` is asserted *immediately* (combinationally) whenever either the external POR (`ext_rst_n`, active-low) is asserted *or* the PLL has not yet reported lock, and is released only after two flip-flops of the system clock following both conditions clearing – so no downstream synchronous logic (SDRAM controller, compute datapath, SPI bridge) ever sees an asynchronous release edge. This is a real, RTL-implemented mechanism, not a placeholder: it is included, unmodified, in every synthesis and P&R run reported in this datasheet.

CHAPTER 4

Host interface (SPI)

`spi_host_bridge.v` is the real, physical host interface for V2's board-level top (`fpga_neural_v2_top.v`). It replaces the internal 110-pin `reg_*` bus (a simulation/testbench convenience that must never be represented as a physical board interface) with 4 real pins: `spi_sclk`, `spi_mosi`, `spi_miso`, `spi_cs_n`. Mode 0 (CPOL=0/CPHA=0), MSB-first, one opcode per CS-low period.

4.1 Opcodes

Opcode	Name	Function
0x10	WRITE_JOB	15-byte payload: node id, required dependency count, producer ids, <code>x_base</code> , <code>w_base</code> , <code>n_tiles</code> , <code>result_addr</code> . Registers one Dependency Manager job; <code>reg_valid</code> is held until the same-cycle <code>reg_valid && reg_ready</code> acceptance fires.
0x01	WRITE_MEM	word-addressed raw SDRAM write, via a second, arbitrated AR port.
0x02	READ_MEM	word-addressed raw SDRAM read.
0x20	STATUS	1-byte status: job-busy, mem-busy, sticky last-job-accepted.
0x0F	RESET	pulses a soft-reset into the compute+memory core (not the SPI bridge's own state, avoiding a self-reset hazard).

4.2 Verification

The protocol engine is verified in two independent ways:

- **Isolated:** `tb_spi_host_bridge.v`, 18/18 PASS – every opcode, including a real DQM-style byte-masked `WRITE_MEM`.
- **End-to-end:** `tb_fpga_neural_v2_top_smoke.v`, 11/11 PASS – a single job, two jobs back-to-back, two jobs across a realistic $\sim 85\mu\text{s}$ SPI-paced gap, and a parametric sweep of inter-job gaps (100 ns / 5 000 ns / 50 000 ns), all checked bit-exact against a software golden model via a real SDRAM backdoor readback.

Both regressions pass with **zero regression** to the existing N=2/N=4 D-Stress bit-exact baseline (identical cycle counts before and after every fix applied during this interface's own development).

CHAPTER 5

Pinout, place-and-route, and timing

5.1 Final pinout

All 44 top-level signals of `fpga_neural_v2_top.v` carry a real ball assignment (`v2_board_top.lpf`), sourced from the official Lattice pinout CSV (rev. 3.0) and confirmed by a real, successful `nextpnr-ecp5` P&R run – **no placeholders**.

Signal	Ball	Bank	Notes
<code>osc_clk</code>	H5	–	reused from V1's own validated LPF
<code>ext_rst_n</code>	B4	–	reused from V1's own validated LPF
<code>spi_sclk</code>	L3	6/7	real, plain GPIO
<code>spi_mosi</code>	M3	6/7	real, plain GPIO
<code>spi_miso</code>	L2	6/7	real, plain GPIO
<code>spi_cs_n</code>	N2	6/7	real, plain GPIO
<code>pll_locked</code>	L1	6/7	real, plain GPIO (bring-up/debug)
<code>sdram_*</code> (37 signals)	see LPF	6/7	control/address/data/mask bus

All signals are assumed LVCMOS33, matching banks 6/7's own real VCCIO range and the SDRAM device's own 3.3 V requirement (not yet independently cross-verified at the schematic/PCB level – a disclosed WARNING, not a blocker).

5.2 Synthesis (real, this revision)

Yosys `synth_ecp5`, target `fpga_neural_v2_top`, real run, zero CHECK-pass problems:

Resource	Count
TRELLIS_FF	6 322
TRELLIS_COMB (LUT4 equiv.)	7 084
MULT18X18D	32 (= 4 processors × 8-wide MAC)
EHXPLL	1 (real PLL, confirmed present)
DP16KD	0 (all small SRAMs synthesize to distributed RAM)

38 unique warnings (43 total), all matching this project's own previously-reviewed, benign base-line set (a known `genvar` multi-driver artifact in `neural_processor.v`, small-array-to-register unrolling, and the real SDRAM `DQ` tristate bus) – no new warnings introduced by the SPI bridge, PLL, or reset synchronizer.

5.3 Place and route: 8-seed timing table

Real `nextpnr-ecp5` P&R, same final top and LPF, 8 distinct seeds, target 64 MHz:

Seed	Fmax (MHz)	Result	Slack @ 64 MHz
1	73.17	PASS	+1.958 ns
2	68.90	PASS	+1.111 ns
3	72.10	PASS	+1.755 ns
4	68.51	PASS	+1.029 ns (worst)
5	69.29	PASS	+1.193 ns
6	73.03	PASS	+1.931 ns
7	74.17	PASS	+2.143 ns (best)
8	70.10	PASS	+1.360 ns

8/8 seeds PASS at 64 MHz – worst 68.51 MHz, best 74.17 MHz, mean 71.16 MHz. Every seed also fits with real routing margin: `TRELLIS_IO=44/245` (17%), zero unrouted nets, zero placement/routing errors, across all 8 runs.

The critical path is routing-dominated (typ. 15–20% logic / 80–85% routing) and alternates, seed to seed, between two comparably-tight structures already documented in this project's own prior timing investigations: `dependency_manager.v`'s own wide priority-encoder scan, and `sdram_unified_backend.v`'s own weight-cache hit-index logic – neither is a new defect introduced by this revision.

Hold timing was not separately reported by this toolchain's standard summary output and was not independently analyzed this round (an honest, disclosed OPEN item, not a fabricated PASS).

CHAPTER 6

Power, SDRAM electrical, and configuration

6.1 Power architecture – status: OPEN

Rail *voltages* are real, DATASHEET VALUEs from the actual ECP5 and AS4C4M16SA-6TIN datasheets; regulator selection, current budget, and decoupling are **not** finalized in this revision.

Rail	Nominal	Status
VCC (core)	1.1 V	DATASHEET VALUE; regulator TBD
VCCAUX	2.5 V	DATASHEET VALUE; regulator TBD
VCCIO 6/7 (SDRAM)	3.3 V (assumed)	matches SDRAM's own LVCMOS33 need; not independently re-verified
SDRAM VDD/VDDQ	3.3 V	DATASHEET VALUE (AS4C4M16SA-6TIN)

No real power-estimation tool was run against the actual synthesized netlist this revision; a real, computed current budget therefore remains an OPEN item, not an invented number.

6.2 SDRAM electrical / timing sign-off – status: PARTIAL

AS4C4M16SA-6TIN is a standard JEDEC SDR SDRAM; the controller (`sdram_controller.v`) implements a real power-up wait, mode register set, and periodic AUTO REFRESH, verified functionally in simulation (init/refresh/read/write/burst/masked-write, 40 real refresh events per D-Stress run, zero corruption). A byte-for-byte cross-check of the controller's own timing constants against the device's real datasheet parameters (CAS latency, t_{RCD} , t_{RP} , refresh interval, setup/hold) was **not** independently re-performed this revision – simulation-level correctness is established; a dedicated datasheet-parameter audit remains an OPEN item before board layout.

6.3 Configuration – status: OPEN

Standard ECP5 JTAG (`TDI/TDO/TCK/TMS`) and configuration (`PROGRAMN/INITN/DONE/CCLK`) pins are identified and standard (real balls listed in `docs/pinouts.md`); JTAG remains available regardless of boot mode. No configuration-flash part number has been selected, and no SPI-flash-boot-vs-JTAG-only decision has been made this revision.

CHAPTER 7

Benchmarks

7.1 Internal benchmark: D-Stress workload

256 independent neurons, 128 inputs each (16 tiles of `P_IN=8`), one shared activation vector – the project’s own existing, reused benchmark, run against the final, frozen RTL via Verilator.

Config	Cycles	Correctness	Latency @ 64 MHz
N=2	49 788	256/256 PASS	777.9 μ s
N=4	49 771	256/256 PASS	777.7 μ s

Cycle counts are **simulation-verified** (Verilator, bit-exact) and unchanged before/after the ERR-0025 read-timing fix. Latency at 64 MHz is a **derived** figure (cycles / clock), not an independently re-simulated wall-clock measurement, and is not a measured-hardware number (no fabricated board exists).

N=4 barely improves over N=2 on this specific workload (49 771 vs. 49 788 cycles) because the workload is dominated by shared-SDRAM and shared-activation-fill traffic, not by raw MAC throughput – consistent with this project’s own prior STEP17/18 memory-bound-vs-compute-bound findings, not a new anomaly.

7.2 Host-interface pacing benchmark

The board-level SPI smoke test (§4.2) exercises single-job, back-to-back, and gap-swept (100 ns/5 000 ns/50 000 ns/ $\sim$ 85 μ s) dispatch patterns – all 11/11 bit-exact – establishing that realistic host pacing does not, itself, change correctness (only the STEP19-era registered-read bug did, and that is fixed).

7.3 Software reference comparison

No physical ESP32 (or other embedded MCU) hardware is available in this environment. A real ESP32 benchmark was *not* performed, and no ESP32 number is estimated or invented here – this is a disclosed OPEN item, not a claimed result.

As an honest, clearly-labeled illustrative data point only, the identical D-Stress arithmetic (256 neurons $\times$ 128 INT8 MACs, ReLU + INT8 saturate) was compiled (`cc -O2`) and run on the *development machine itself* (Apple M4, arm64 – **not** an embedded target):

Metric	Value
Platform	Apple M4 (arm64), <code>cc -O2</code> , single-threaded scalar C
Latency, full 256-neuron batch	1.28 μ s
Throughput	2.56×10^{10} MAC/s

On this specific, tiny workload, a modern, auto-vectorizing, multi-GHz superscalar CPU outperforms the current FPGA design’s *simulated* cycle count in raw latency. This is reported honestly rather than omitted: FPGA-Neural V2’s real value proposition is a deterministic, low-power,

standalone SDRAM-attached accelerator for hosts that do *not* have this class of CPU (e.g. an ESP32-class microcontroller) – not a claim of outperforming a desktop-class processor on this workload size. A real, meaningful software baseline for THAT comparison requires the actual embedded target hardware, which remains unavailable this revision.

CHAPTER 8

Chip readiness and roadmap

8.1 Precise readiness checklist

Per this project's own standing rule: no item below is checked unless genuinely verified; no OPEN item is masked as a future enhancement.

Item	Status
RTL frozen, zero V1 dependency	OK
Bit-exact regression (N=2, N=4)	OK (256/256, Verilator)
Real SPI host interface, end-to-end	OK (11/11, zero regression)
SDRAM validation (init/refresh/masked-write)	OK
Synthesis (final top, real Yosys run)	OK (0 CHECK-pass problems)
Place&route (8 real seeds)	OK (8/8 fit, routed, zero errors)
Timing closure @ 64 MHz	OK (8/8 seeds PASS, worst 68.51 MHz)
Final pinout (44/44 top-level signals)	OK (real balls, P&R-confirmed)
Hold-time analysis	OPEN (tool limitation, see PRE_PCB_VERIFICATION.md §6)
Power: rail voltages	OK (real datasheet values)
Power: regulator selection / current budget	OPEN
SDRAM datasheet-parameter cross-check	OK (real AS4C4M16SA-6TIN Table 17 audit; found and fixed ERR-0026, tMRD unit mismatch)
Configuration flash selection	OK (Winbond W25Q32JVSSIQ, SOIC-8, frozen)
Real KiCad schematic + ERC	NOT STARTED
PCB layout	NOT STARTED
BOM (sourced, purchasable parts)	NOT STARTED
Physical board bring-up	NOT APPLICABLE (no board fabricated)
Real embedded-target (ESP32) software baseline	NOT AVAILABLE (no hardware)

8.2 Final answer

NOT SILICON READY

The RTL, simulation, synthesis, place-and-route, timing, and pinout gates are all genuinely closed for the final board-level top. The remaining, real, disclosed blockers are entirely in the physical domain – schematic capture, PCB layout, sourced BOM, and (necessarily) physical fabrication and bring-up – none of which were claimed complete in this revision.

8.3 Roadmap

- Real KiCad schematic capture + ERC.

- PCB layout (BGA escape, SDRAM routing, power integrity).
- Sourced, purchasable BOM (regulators, configuration flash, connectors).
- Real power current-budget estimate against the actual synthesized netlist.
- First physical board fabrication and the 15-step bring-up procedure already documented in `FIRST_POWER_ON.md`.
- A real embedded-target (ESP32-class) software baseline, once hardware is available, to replace this revision's honestly-labeled desktop-CPU illustrative comparison.