

SYMATICS

Technical Overview

State-resident compute architecture for long-context inference and multi-agent systems

Public document · 14 October 2026 · symaticslab.com

Status: pre-silicon. Gen1 architecture defined; architecture simulation and validation POC completing end of October 2026; multi-node FPGA proof-of-concept build targeted end of November 2026. Concept launch at the AI for Good Summit, Shangri-La Singapore, 5–7 November 2026.

1. The thesis: the next bottleneck is state

In autoregressive inference, every generated token re-reads the active weights and state. As models move to long context and multi-agent workloads, the dominant cost is no longer arithmetic but data movement: weights stream from memory on every decode step, and context state grows from megabytes to tens of gigabytes per sequence.

Symatics is building the MPU, a state-resident compute architecture: silicon where information stays local, structured and continuously available. The MPU keeps weights resident in local memory, moves only bounded state between chips over fixed-latency links, and fixes the schedule at compile time rather than at runtime.

2. Gen1 at a glance

Gen1 design targets. These are architecture targets, not silicon-measured values, and may change as validation proceeds.

Parameter	Gen1 design target
Compute array	32,768 compute units per MPU (256 × 128 baseline), FP16 datapath
Process	TSMC N5P-class
On-chip memory	~384 MB SRAM per MPU (baseline)
External memory	Commodity DRAM (GDDR7-class). No HBM
Packaging	Standard packaging; no advanced packaging required
Interconnect	SIF state fabric on standard PCIe-class SerDes PHYs; fixed latency, compile-time configured; <200 ns single hop within a domain
Scale	64 MPUs per state domain (one rack); the compiler sees one array
Scheduling	Compile-time. No runtime routing, no cache coherence, no transaction layer in the fabric

3. Validation status

Gen1 exists to prove the core principles of state-resident structural computing on real hardware before any tape-out commitment. Status as of 14 October 2026:

Done	In progress	Next
<ul style="list-style-type: none">• Gen1 architecture defined: compute array, state domain, fabric topology• Design partners engaged for front-end (Taiwan) and implementation (Canada)• FPGA platform selected: AMD Alveo V80 compute nodes, AMD Versal VPK120 fabric node	<ul style="list-style-type: none">• Architecture specification and cycle-level C-model• Architecture simulation & validation POC — completing end of October 2026• Proof-of-concept compiler for compile-time state-path scheduling• Multi-node FPGA proof-of-concept build — targeted end of November 2026	<ul style="list-style-type: none">• First FPGA results: fixed-latency state transfer between nodes, published as measured• Latency-determinism and state-residency measurements• Scaling toward a 64-chip domain equivalent

No measured results are reported in this document. We will publish FPGA measurements, with their configuration, when they exist.

Independent review — in progress

An independent validation statement covering the Gen1 architecture simulation & validation POC is in preparation. Earl Mai, CTO of ePIC, reviews the hardware platform and measurement methodology; a postdoctoral researcher in a University of Toronto AI laboratory reviews the AI-workload methodology. The signed statement will be published at symaticslab.com when complete.

4. Where the MPU fits, stated honestly

Single-stream decode throughput is bounded by memory bandwidth: $\text{tokens/s} \approx \text{memory bandwidth} \div \text{bytes per token}$. The MPU case is conditional and specific:

Workload	Assessment	Basis
Dense models served at large batch	Not the wedge	Same bandwidth wall as GPU; HBM-based GPUs win on bandwidth per dollar at batch
Mixture-of-experts / sparse	Core opportunity	Weights stay resident; only bounded state crosses chips. Full expert-gating support is a Gen2 target
Long context	Candidate opportunity	Context state held in local DRAM while it fits; no descent to slow storage
Single-stream, latency-bound	Candidate opportunity	Deterministic schedule; no queueing in the fabric

Performance figures shared with partners are analysis-derived projections under stated configurations, not benchmark results. They are available under NDA.

5. Roadmap

Gen1: architecture validation	Prove state-resident structural computing on FPGA; close the architecture	In progress: sim & validation POC (Oct) → FPGA proof-of-concept build (Nov)
Gen2: production silicon	Production tape-out for MoE, long-context and agentic workloads	Planning; target of record set after Gen1 closes
Gen1 closes the architecture. Gen2 moves to production silicon.		

6. Singapore concept launch

Symatics shows its architecture publicly for the first time at the AI for Good Summit, hosted by Alliance for Good at Shangri-La Singapore, 5–7 November 2026. The booth exhibit presents one state domain at rack scale: state leaves a compute node, crosses the fabric and a fixed-latency optical link, and arrives at the next node — on the same path and for the same duration, every cycle. Visitors can lengthen the fiber and see only the link segment stretch, or inject a fault and see it flagged and dropped with no retry. FPGA emulation of the full path is planned to follow.

7. Company and contact

Symatics Technology Pte. Ltd. is a fabless semiconductor company headquartered in Singapore, with an engineering office in Taipei. It is led by founder and CEO Mo Zining and co-founder and CTO Wayne Wu, a 25-year semiconductor veteran whose career spans ATI/AMD, Lightelligence (HK:01879), eTopus, Atmel, Western Digital and HP.

Detailed specifications, analysis and architecture documents are available under NDA.

General: hello@symaticslab.com · **Partnerships:** partnerships@symaticslab.com · **Investors:** investors@symaticslab.com · symaticslab.com

© 2026 Symatics Technology Pte. Ltd. Provided for general information. Specifications are design targets, not measurements from silicon.