



THE FOUNDATIONAL
INTELLIGENCE LAYER

ZEROOTH
— AI —

S2

CUSTOMER WHITEPAPER

Recovering useful AI inference capacity



ProgGen Zeroth AI Limited

Version 1.1

August 2026

01 | Executive summary

S2 is a governed software control layer for AI inference. It sits between application demand and model serving, reducing avoidable physical model calls where eligible workload patterns exist while forwarding work that still requires fresh execution.

Updated evidence posture

S2 has moved from a local mechanism proof into a bounded laptop capacity result plus a first A100/vLLM fixed-demand transfer result. The strongest capacity claim remains specific to Llama 3.1 8B on RTX 5050 + Ollama. The A100 result confirms physical-work reduction at fixed demand but does not yet establish an A100 capacity ceiling.

Completed result	What it supports	Boundary
RTX 5050 + Ollama fixed-demand portability	Fewer physical calls and lower GPU-board energy across four qualified Ollama models.	Self-run laptop, GPU-board energy only.
RTX 5050 + Ollama Llama 3.1 ceiling	At least 50.09% more quality-accepted logical workload than native under the same frozen conditions.	Specific to Llama 3.1 8B Q4_K_M via Ollama on the tested laptop.
A100-SXM4-80GB + vLLM fixed-demand transfer	Completed-response reuse and coalescing reduced physical calls by 26.19% at matched demand.	Fixed-demand result only; no A100 ceiling or whole-system result yet.

02 | The capacity problem

AI inference demand is logical; GPU consumption is physical. Treating the two as identical can waste scarce capacity when retries, repeated demand, simultaneous requests or obsolete work create avoidable physical model calls.

- A logical request is the service obligation presented by an application or user.
- A physical model call is fresh upstream model execution.
- S2 seeks to improve the relationship between accepted logical service and required physical execution.
- Unique, necessary and ineligible work continues to the model-serving path.

The commercial question

Can a suitable workload deliver more quality-accepted logical service from the installed estate by preventing unnecessary work from reaching scarce accelerator capacity?

03 | What S2 does

Control function	Customer meaning	Claim boundary
Accepted-response reuse	A previous accepted outcome may satisfy eligible repeated demand.	Reused service is not newly generated physical computation.
Concurrent coalescing	Eligible simultaneous obligations may share one in-progress generation.	Shared delivery remains quality-gated.
Supersession	Explicitly obsolete work may be prevented from consuming further capacity.	Requires an application-defined replacement relationship.
Priority and admission	Demand is controlled against quality, latency, queue and thermal limits.	Does not overclock or modify GPU hardware.
Fail-open forwarding	Required work continues to the model if S2 cannot optimise safely.	Optimisation failure does not silently drop required inference.

04 | Current completed evidence

RTX 5050 + Ollama: fixed-demand portability

Five valid matched crossover pairs across Llama 3.1 8B, Qwen 2.5 7B, Gemma 3 4B and Llama 3.2 3B produced consistent physical-call reduction and GPU-board energy improvement at matched accepted demand.

Metric	Native	Full S2	Result
Scheduled logical requests	2,476	2,476	Matched
Accepted requests	2,472	2,472	Matched
Physical model calls	2,476	1,778	28.19% fewer
GPU-board energy	134.57 kJ	112.91 kJ	16.10% lower
GPU energy/accepted request	54.44 J	45.67 J	16.10% lower
GPU energy/accepted token	approx. 2.4905 J	approx. 2.0875 J	16.18% lower

RTX 5050 + Ollama: Llama 3.1 capacity ceiling

Measure	Native	Full S2
Sustainable ceiling interval	[0.587240, 0.601563] req/s	[0.902882, 0.924903] req/s
Highest confirmed passing rate	0.587240 req/s	0.902882 req/s
First confirmed failing rate	0.601563 req/s	0.924903 req/s
Binding constraint	80 C thermal rule	80 C thermal rule
Conservative multiplier	-	1.5009x
Conservative capacity reclaim	-	at least 50.09%

The conservative result compares the lowest confirmed S2 passing boundary with the higher native boundary that had already failed. It is a logical service-capacity result, not a claim that the GPU became physically faster.

Four-hour S2 endurance

Endurance measure	Result
Offered / accepted rate	0.857738 / 0.857708 req/s
Scheduled / submitted / accepted	12,351 / 12,351 / 12,351
Rejected / timed out / drain completions	0 / 0 / 0
Physical calls	8,646
Physical-call reduction	29.998%
Quality and structured acceptance	100% / 100%
SLO attainment	99.903%
P95 / P99	1.764 s / 2.696 s
GPU energy/request	33.064 J
Mean / peak GPU temperature	60.23 C / 77 C

A100-SXM4-80GB + vLLM: fixed-demand transfer

Three matched Qwen2.5-7B-Instruct BF16 pairs were completed on a full NVIDIA A100-SXM4-80GB allocation with MIG disabled, using vLLM 0.23.0. The result establishes transfer of the physical-work reduction mechanism to data-centre-class hardware and a production-oriented runtime, but not an A100 capacity ceiling.

Metric	Native	Full S2	Change
Physical calls	44,976	33,198	26.19% fewer
GPU energy / accepted request	8.166 J	8.023 J	1.76% lower
GPU energy / accepted token	0.677 J	0.665 J	1.75% lower
P50 latency	0.312 s	0.282 s	9.72% faster
P95 latency	0.617 s	0.611 s	0.91% faster
Quality / SLO	approx. 100%	100%	Equal or better

A100 interpretation

A 26.19% physical-call reduction translated into a modest 1.76% GPU-board energy reduction at the tested fixed rate. The A100 capacity-ceiling campaign is required to determine whether avoided physical work converts into additional sustainable service near the native vLLM limit.

05 | Workload fit

S2 is most relevant when demand contains governed recoverable patterns and the service boundary can be measured.

Positive indicator	Why it matters
Retries or repeated requests are visible	Eligible accepted outcomes may serve repeated obligations.
Concurrent requests ask for the same outcome	Single-flight coalescing may reduce fresh generations.
Agent or workflow work can become obsolete	Explicit supersession may prevent wasted execution.
Queue or latency pressure appears during demand spikes	Avoided physical calls may release headroom.
Hardware expansion is being considered	A measured assessment can identify avoidable demand before procurement.

06 | Measurement and validation standard

- Verify the hardware and serving environment.
- Establish the strongest stable native baseline.
- Freeze S2 policies before comparison.
- Run matched native and Full S2 conditions.
- Report accepted logical work, physical calls, latency, queue state, GPU-board energy and hardware state separately.
- Preserve passing, failing, invalid, interrupted and excluded runs.
- Publish only evidence-supported conclusions.

07 | Commercial value

The customer value case is evidence-led: identify avoidable work, prove what can be recovered inside the agreed boundary and decide what the installed estate can support. The result is customer-specific and depends on workload composition, model behaviour, quality requirements, service constraints and deployment conditions.

No automatic multiplier

The 50.09% laptop capacity result is not a customer promise and is not an A100 result. Any customer claim must stay inside that customer assessment’s model, workload, serving, energy and service boundary.

08 | Claim boundaries

- S2 does not claim to make a GPU physically faster.
- Reused or shared service is not described as newly generated computation.
- GPU-board energy is not whole-system, server-input, rack, cooling or facility energy.
- The current work is not production-cluster validation.
- The current work is not independent validation.
- One traffic profile is not assumed to represent every workload.
- The A100 fixed-demand result does not establish an A100 capacity multiplier.
- A customer-specific capacity claim requires matched validation under that customer's conditions.

09 | Assessment path

Step	Purpose
1. Non-confidential workload-fit discussion	Identify visible recoverable demand patterns and the operating constraint.
2. Confidentiality agreement	Protect detailed traces, architecture and operational data.
3. Technical assessment	Define model, serving, workload, tenant, quality and energy boundaries.
4. Matched validation	Compare strongest native operation with Full S2 under frozen rules.
5. Evidence-led recommendation	Proceed, refine, pause or reject based on supported results.

10 | Next step

Primary next step

Request an S2 Capacity Assessment. The current completed evidence supports a structured assessment, but not a universal savings promise.