

CHRONOS Prefill-Decode Fabric: An Evidence-Aware Architecture, Prototype Evaluation, and Integration Roadmap

Aaron Singh

Abstract—This paper presents CHRONOS Prefill-Decode Fabric, a project in a twenty-project AI infrastructure portfolio. The current repository is an active prototype with implemented behavior, automated correctness tests, and explicit boundaries around unverified hardware or production claims. We describe the system model, current implementation, goals, and evaluation method, then propose five integrations that advance the project toward reproducible system-level validation. A local audit executed 3 project tests successfully. No accelerator, deployment, or performance conclusion is inferred unless a corresponding committed benchmark or profiler artifact exists.

Index Terms—CHRONOS, AI infrastructure, reproducibility, prototype validation, systems evaluation, hardware–software co-design

I. INTRODUCTION

CHRONOS Prefill-Decode Fabric (P09) addresses a bounded problem within the portfolio’s track-03-distributed-infrastructure track. Its declared status is *Active Prototype* [1]. The project validates its central mechanism before adding device-specific acceleration, production orchestration, or real-world hardware. This ordering matters because optimization without a trusted reference can make incorrect behavior appear successful.

The project goals are to establish deterministic behavior, encode correctness as tests, create machine-readable evidence, identify limiting resources through measurement, and integrate results with adjacent portfolio systems without losing provenance. The repository contains one paper named exactly after its singular folder: `P09-chronos-prefill-decode-fabric.tex`.

II. TECHNICAL CONTEXT

Distributed inference depends on scheduling, capacity, communication, reliability, and observability. Later multi-node experiments must record topology, synchronization costs, retries, and tail latency. The portfolio benchmark standard requires environment, workload, method, metric, artifact, reproduction, and limitation fields; unknown values remain explicitly unmeasured [3].

III. SYSTEM MODEL AND ARCHITECTURE

The prototype is organized around the following domain model:

$$T_i = W_i^{pre} + S_i^{pre} + W_i^{dec} + S_i^{dec} \quad (1)$$

The equation is a design and test abstraction rather than a claimed empirical law. It supports invariants and expected-value checks while later implementations replace synthetic inputs with representative workloads or devices.

The software architecture contains an input/configuration layer, a deterministic core, validation and evidence output, and a local visualization. The principal inspected source artifacts are `python/scheduler.py`. Unsupported real-world conditions are surfaced as limitations rather than silently simulated.

IV. DETAILED SCRIPT OPERATION AND RATIONALE

The scheduler models separate prefill and decode service stages. It advances queue timing, guarantees that each stage does not overlap itself, ensures decode follows prefill, and emits machine-readable timing data for later disaggregated-serving experiments.

The execution path is:

- 1) Create deterministic requests with prefill and decode work.
- 2) Queue and service prefill without overlapping the prefill worker.
- 3) Release completed prefill work into the decode queue.
- 4) Service decode without overlap and enforce stage ordering.
- 5) Emit per-request timing for future transfer-cost and parallel-worker models.

V. IMPLEMENTED PROTOTYPE

The metadata and source audit found these completed features [2]:

- Python two-stage scheduler
- Prefill/decode queue timing
- Unit tests and JSONL output

`python3 -m unittest discover -s tests -p 'test_*.py'` completed successfully with 3 tests. The inspected test artifacts are `tests/test_scheduler.py`. This is evidence of local correctness for encoded cases, not production scale or hardware performance.

VI. TESTING METHODOLOGY AND OBSERVED RESULTS

Testing uses Python’s standard `unittest` discovery and exercises the public behavior of the reference implementation. The audit reran the suite from the project folder with `python3 -m unittest discover`

TABLE I
IMPLEMENTATION ARTIFACTS AND WHY THEY EXIST

Artifact	Observed responsibility	Engineering rationale
python/scheduler.py	Job, simulate, sample, main	Implements the inspectable, unit-tested project core.
scripts/reproduce.sh	Fixed test and demonstration entry point	Gives another developer one command for local reproduction.
PROJECT.yaml	and	Separates declared intent from evidence-backed implementation.
ANALYSIS.md		Makes outputs inspectable without upgrading simulation into a hardware claim.
streamlit_app.py		

`-s tests -p 'test_*.py'`. All 3 discovered tests passed. The result establishes correctness only for the encoded local cases; it does not establish accelerator correctness, real-device behavior, production reliability, or benchmark completion.

No numerical performance result is promoted by this test run. Where scripts emit JSON or JSONL, those outputs remain raw or simulation-specific until a reviewed summary includes hardware, software, workload, warm-up, repetition, correctness threshold, Git revision, and limitations.

VII. CLAIM BOUNDARIES AND RISKS

The project records these unverified or excluded claims:

- Simulation does not prove serving performance

The main risk is confusing synthetic or modeled behavior with deployed-system behavior. Other risks include incomplete workloads, platform-dependent timing, missing failure injection, and interfaces not yet exercised across device boundaries. Performance claims require a reviewed record meeting the portfolio standard.

VIII. EVALUATION PLAN

Evaluation proceeds through correctness tests, deterministic reproduction with Git and environment metadata, repeated benchmarks reporting latency/throughput/memory/error metrics, and a named profiler capture tied to exact hardware and source revision. Success requires reference equivalence within a documented tolerance, preservation of safety and resource invariants, clear failures, and evidence reproducible from a clean environment.

IX. GOALS, MILESTONES, AND SUCCESS CRITERIA

The project goals are staged so that correctness precedes performance and integration. A goal is complete only when its proof artifact is committed or otherwise reviewable; prose or a simulated number alone is insufficient.

X. FUTURE WORK AND INTEGRATIONS

The five project-specific next steps are:

- 1) Parallel workers
- 2) Network transfer cost
- 3) vLLM adapter
- 4) Add KV-cache transfer, backpressure, and recovery state machines.

- 5) Deploy the fabric on P10's two-node testbed.

The early items complete declared evidence; the later items connect downstream portfolio consumers. Each integration should add tests and a reviewable artifact such as JSONL evidence, a report, profiler capture, deployment manifest, trace, or labeled data set.

XI. CONCLUSION

CHRONOS Prefill-Decode Fabric is an evidence-aware active prototype: its implemented behavior and tests are real, while unbuilt hardware, deployment, and performance goals remain labeled. Completing the five integrations in dependency order will advance it from a learning artifact toward a credible portfolio component.

REFERENCES

- [1] Aaron Singh, "CHRONOS Prefill-Decode Fabric PROJECT.yaml," local portfolio repository.
- [2] Aaron Singh, "CHRONOS Prefill-Decode Fabric: README, ANALYSIS, source, and test artifacts," local portfolio repository.
- [3] Aaron Singh, "AI Infrastructure Portfolio Benchmark Standard," local portfolio repository.

TABLE II
AUDITED TEST MATRIX

Test artifact and case	Behavior being checked	Observed result
tests/test_scheduler.py:test_stages_do_not_overlap_themselves	Stages do not overlap themselves.	Pass (local)
tests/test_scheduler.py:test_decode_follows_prefill	Decode follows prefill.	Pass (local)
tests/test_scheduler.py:test_bad_rate	Bad rate.	Pass (local)

TABLE III
DECLARED STATUS VERSUS AUDITED EVIDENCE

Audit field	Finding	Evidence source
Declared status	Active Prototype	PROJECT.yaml
Evidence-backed status	Active local prototype; 3 tests passed	Source plus local unittest run
Accepted measured results	None recorded in measured_results	PROJECT.yaml
Mismatch / claim boundary	Simulation does not prove serving performance	PROJECT.yaml and ANALYSIS.md
Next proof required	Parallel workers; Network transfer cost	Planned features

TABLE IV
PROJECT GOALS AND REQUIRED PROOF

ID	Goal	Completion evidence
G1	Parallel workers	Passing tests and a reviewed source artifact
G2	Network transfer cost	Machine-readable result with reproduction metadata
G3	vLLM adapter	Reference-equivalence or domain-correctness report
G4	Model transfer and parallel workers	Named profiler, deployment, or integration artifact
G5	Validate a disaggregated serving adapter	Dashboard/report link preserving provenance and limitations