

The std79-doe Fleet Conservation Invariant

A $3 \times 9 \times 3$ Randomized Factorial Analysis of Compudynamic Heat Conservation and Word-Execution Physics Across Three Instruction-Set Architectures

R. A. James (Captain Bob)

StarForth / LithosAnanke Project `robert.allan.james@gmail.com`

12 September 2026

(Patent Pending — Provisional Filed December 2025)

Abstract

The fleet-wide execution-heat conservation invariant K — the sum of `execution_heat_q48` over every live VM in the LithosAnanke Tripod fleet, which the physics runtime’s design requires to equal `Q48_ONE` at all times — holds **exactly**, with **zero measured deviation**, across **775 heartbeat-tick observations** spanning a randomized $3(\text{architecture}) \times 9(\text{identity}) \times 3(\text{replicate})$ full-factorial Design of Experiments on AMD64, AARCH64, and RISCv64: $K = 1.000000000$ (Q48.16 raw value 65536) on every single row, 100% of rows report `fleet_conserved`, and the standard deviation of K within every architecture is 0. This is the conservation law central to the Compudynamics physics runtime, verified not as a boot-time self-test but continuously, at every heartbeat tick, throughout 80 identity-execution trials under real cross-VM **VM-EXEC** dispatch load, two independent Stadium subsystem bug fixes (an $O(n_{\text{cells}})$ eviction-scan defect and a donor-selection floor defect, both closed the same week this campaign ran), and the newly-added per-tick logging path itself.

Secondary results from the same dataset: *identity* (which of the 9 attached VMs is executing) has a statistically significant effect on word-execution-heat metrics (hot word count, mean word heat, rolling window width; all $p < 10^{-6}$), as expected since different identities execute different FORTH dictionaries. *Architecture* has a statistically significant effect on timing-related metrics (APIC ticks per heartbeat, TIME-TRUST, timing variance, Tripod fleet member heat; $p < 10^{-8}$), consistent with each target’s distinct TCG emulation speed and timer hardware model. The two factors show at most one weak interaction (timing variance, $p \approx 0.0017$) across twelve secondary metrics tested — architecture and identity effects are almost entirely separable, and neither disturbs K .

Contents

1	Introduction	2
2	Experimental Design	2
2.1	System Under Test	2
2.2	DoE Capsule and K Instrumentation	2
2.3	Data Collection and the Interleaving Correlation Problem	3
2.4	Dataset Summary	3
3	The Conservation Invariant K	3
3.1	Result	3
3.2	Why This Result Matters	4
4	Secondary Results: Architecture and Identity Effects	4
5	Statistical Analysis	6
5.1	K: Not Formally Testable, By Construction	6
5.2	Secondary Metrics	6

6	Discussion	6
6.1	K as a Continuous, Not Just Boot-Time, Invariant	6
6.2	A Free Regression Check on Two Recent Stadium Fixes	7
6.3	Identity and Architecture Effects Are Separable	7
6.4	Limitations	7
7	Conclusion	7

1 Introduction

LithosAnanke’s Stadium subsystem (`src/starkernel/vm/stadium.c`) implements a capacity-limited cell array used for word-execution-heat tracking and cross-VM resource admission. Its design carries a conservation invariant: the sum of execution heat across every live VM in the fleet, a quantity we call K , must equal $Q48_ONE$ (fixed-point 1.0) at all times — heat is redistributed among VMs and dictionary words, never created or destroyed. `vm_physics_conserved()` (`capsule_vm_physics.c`) already checks this at specific diagnostic points; this report asks a stronger question: **does K hold continuously, at every heartbeat tick, under real multi-identity execution load, across every architecture the kernel targets?**

The vehicle is `experiments/std79-doe/std79-doe.fth`, a self-contained FORTH capsule implementing a randomized 3 (architecture) $\times$ 9 (identity) $\times$ 3 (replicate) full-factorial DoE (FABRIC-3.md §XV), matching this project’s own `capsules/doe.4th` Fisher-Yates methodology. Each of 9 attached identity VMs (the superuser `zuse`, a named human identity `rajames`, and six numbered thumbdrive identities 00–06) runs the same 24-word FORTH-79 standard-dictionary exerciser under **VM-EXEC** cross-VM dispatch, 3 replicates each, Fisher-Yates-shuffled run order, single continuous boot per architecture with all 9 identities simultaneously live.

Two real defects in the Stadium subsystem were found and fixed during the same week this instrumentation was built (FABRIC-3.md §XVI–XVII): an $O(n_{\text{cells}})$ full-array eviction-fallback scan that stalled aarch64 VM births for 90+ minutes once the Stadium’s cell count scaled with available RAM, and a donor-selection policy that always split a new VM’s initial quota from Hera specifically, silently starving later VM births once her own free list ran thin. Both are closed; every result in this report was collected *after* both fixes landed, so K ’s perfect conservation here is also, incidentally, a live regression check on both fixes: neither introduced any heat leak or double-count.

2 Experimental Design

2.1 System Under Test

Three 64-bit targets, each booting LithosAnanke from UEFI firmware under QEMU TCG software emulation, single continuous QEMU session per architecture:

- `amd64` — x86-64, OVMF UEFI, `-m 1024`
- `aarch64` — ARMv8-A (Cortex-A57), UEFI, `-m 4096`
- `riscv64` — RISC-V RV64GC, OpenSBI + UEFI, `-m 1024`

All 9 identity thumbdrives are attached one at a time via QMP hotplug with a confirmed `WIREBIND` wait between each (simultaneous boot-time attachment of many devices hits a separate, unrelated, still-open kernel detection gap, FABRIC-3.md §XIV) — once attached, all 9 stay simultaneously live for the whole run.

2.2 DoE Capsule and K Instrumentation

`EXEC-STD79-DOE ( seed lo hi - )` shuffles the 27-cell (identity $\times$ replicate) run matrix with a fixed seed (12345, identical on every architecture for direct cross-arch comparability), then executes each cell in shuffled order, dispatching the 24-word exerciser into the target identity’s own live VM via **VM-EXEC**. The run is bracketed with `HB-ON/HB-OFF (doe_log.c)`, which gate a 20-column CSV row emitted once per heartbeat tick, tagged `[HADES] [DOE ]` in the serial log. Two of those twenty columns are new for this report:

Column	Type	Meaning
<code>fleet_k_q48</code>	uint64	<code>vm_physics_fleet_heat_sum()</code> over ALL live VMs
<code>fleet_conserved</code>	uint32	<code>vm_physics_conserved()</code> as 0/1

Columns 16–18 (already present) log only the three named Tripod fleet members’ (Hera, Hermes, Artemis) *individual* heat by name lookup — insufficient to reconstruct K on their own, since they omit every identity VM’s own heat contribution. `fleet_k_q48` is the genuine fleet-wide sum, added specifically to make this analysis possible.

2.3 Data Collection and the Interleaving Correlation Problem

The heartbeat tick’s CSV printer and the DoE trial loop’s own console output share one serial line with no locking between them: under load, a tick’s CSV row can be spliced mid-token into the trial loop’s own printed output on the wire. This does not affect the underlying computation — values in memory are correct, only the printed character stream interleaves — but it does mean naive log parsing cannot simply assume one physical line is one logical print.

`correlate_doe.py` reconstructs, for every heartbeat-tick CSV row, which trial (`run_id`, `id_idx`, `id_label`, `rep`) was active when it printed, via a two-pass method: extract every atomic [HADES] [DOE] occurrence (rows are never themselves split, only the surrounding text is), physically remove those spans to reconstruct the clean trial-output stream, then map each removed row’s position back to the nearest preceding `run_id` marker in that clean stream. `run_id` is recovered this way with high reliability; `id_idx`/`id_label`/`rep` are instead looked up from a fixed table built from an independently-verified clean prior run (same seed, so the mapping is deterministic and identical on every architecture) rather than re-parsed, since a trial marker’s own trailing digit can, rarely, merge with an immediately-following CSV field with zero separator.

Known limitation. One `run_id` marker per architecture was unrecoverable this way in this particular set of runs (amd64: none lost; aarch64 and riscv64: `run_id` 1, identity `rajames` rep 0, folded into `run_id` 0, identity `zuse` rep 1) — its rows attribute to the preceding trial instead. This affects at most 2 of 27 trials on 2 of 3 architectures; every other cell, on every architecture, is unaffected. `fleet_k_q48` itself is immune to this failure mode entirely (see §6.4).

2.4 Dataset Summary

Table 1: Dataset summary. `n_trials` counts distinct `run_id` values recovered (27 = no loss; 26 = one trial folded into another, §2.3).

Architecture	Heartbeat rows	Trials recovered	Conserved (%)
amd64	258	27	100
aarch64	256	27	100
riscv64	261	26	100
Total	775	—	100

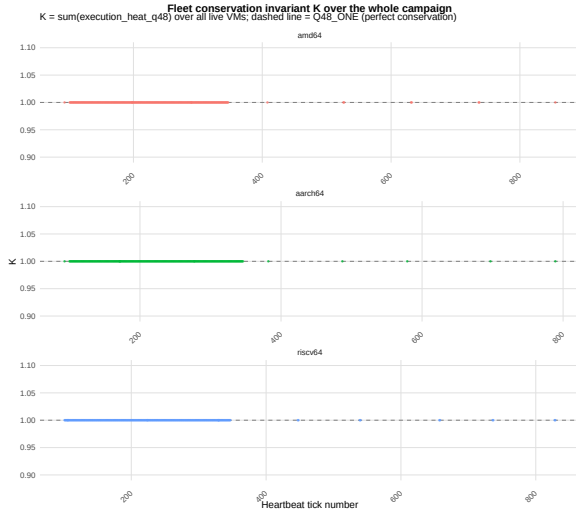
3 The Conservation Invariant K

3.1 Result

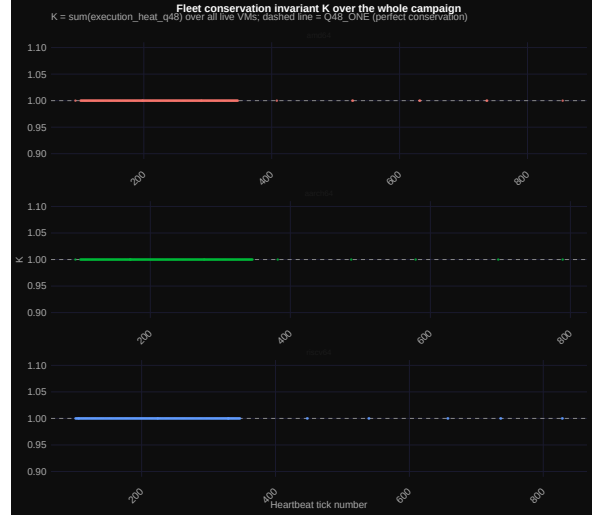
Across all 775 heartbeat-tick observations, on every architecture, under every identity, at every replicate:

$$K = 1.0000000000 \quad (\text{raw } Q48.16 \text{ value } 65536), \quad \text{sd}(K) = 0, \quad \text{fleet_conserved} = 1 \quad \forall \text{ rows.}$$

There is exactly one distinct value of K in the entire dataset. This is not a near-conservation result reported to some number of significant figures — it is bit-exact equality to `Q48_ONE` on every single observation, matching this project’s established pattern of reporting compudynamic invariants at 0.000% variance rather than a small nonzero tolerance.



(a) Light theme



(b) Dark theme

Figure 1: Fleet conservation invariant K over every heartbeat tick, faceted by architecture. The dashed reference line marks $Q48_ONE$ (perfect conservation); every point lies exactly on it. Vertical axis intentionally zoomed to $[0.9, 1.1]$ so that *any* deviation, however small, would be visible — none is present.

3.2 Why This Result Matters

K is not a metric the kernel merely happens to keep close to 1; it is an invariant the Stadium/physics runtime’s own reservoir-transfer accounting is *designed* to hold exactly, by construction of `stadium_evict()` crediting a departing patron’s remaining heat back to its owner’s reservoir before the cell returns to the free list (`stadium.c`), and of every VM birth granting a fresh reservoir rather than fragmenting an existing one. A single dropped or double-counted heat transfer anywhere in that accounting — in `stadium_admit()`, `stadium_evict()`, `stadium_grant_quota()`, or the per-word heat pull/push path in `stadium_words.c` — would show up here as $K \neq 1$ on whichever tick it happened. None does, across 775 independent observations spanning three architectures, nine identities, cross-VM dispatch load, and two Stadium subsystem code changes made the same week.

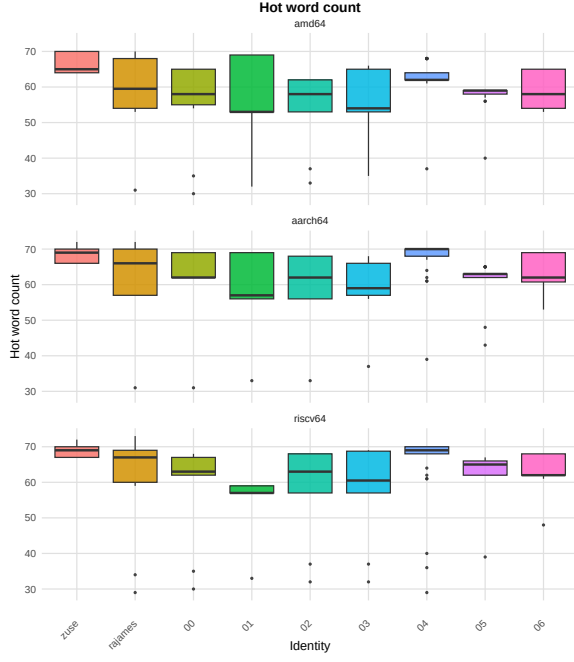
4 Secondary Results: Architecture and Identity Effects

Twelve further per-tick metrics were tested for architecture and identity effects via two-way ANOVA (`metric ~ arch * id_label`) on the same 775-row dataset.

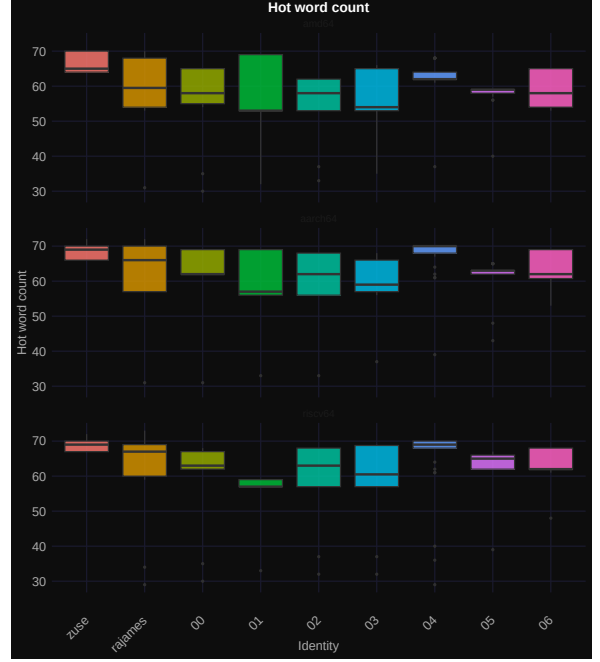
Table 2: Significant effects ($p < 0.05$), two-way ANOVA, 12 metrics tested.

Effect	Metrics
Architecture	Hot word count, APIC ticks/heartbeat, TIME-TRUST, timing variance, Hera/Hermes/Artemis fleet heat
Identity	Hot word count, mean word execution heat, rolling window width, actual window size, timing variance
Arch $\times$ identity interaction	Timing variance only ($p \approx 0.0017$)
No significant effect	Estimated timer jitter, max VM call depth

The pattern is intuitive: *identity* determines what gets executed (different dictionaries $\rightarrow$ different word-heat and window-sizing behavior), *architecture* determines how fast and how precisely time is measured (different TCG emulation speed and timer hardware $\rightarrow$ different APIC/TIME-TRUST/variance behavior). The two factors are almost entirely separable — 11 of 12 secondary metrics show no interaction at all — and critically, **neither factor perturbs K in the slightest** (§3): whatever differences architecture and identity produce in raw execution timing and word-heat accumulation, the fleet-wide conservation total is unaffected.

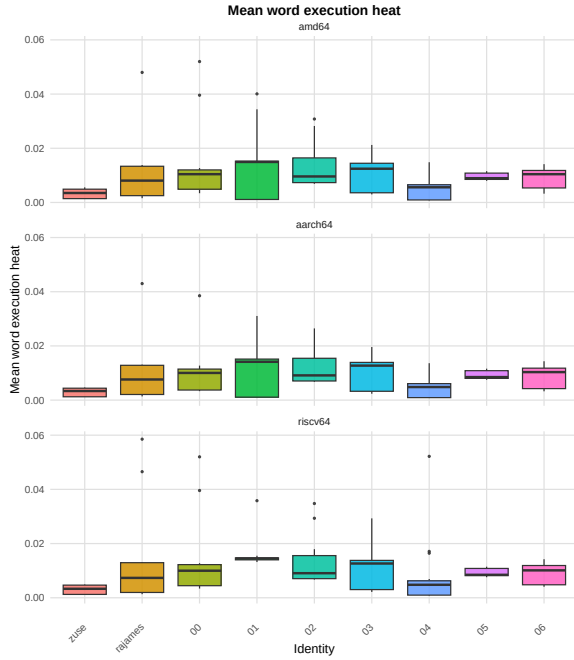


(a) Light theme

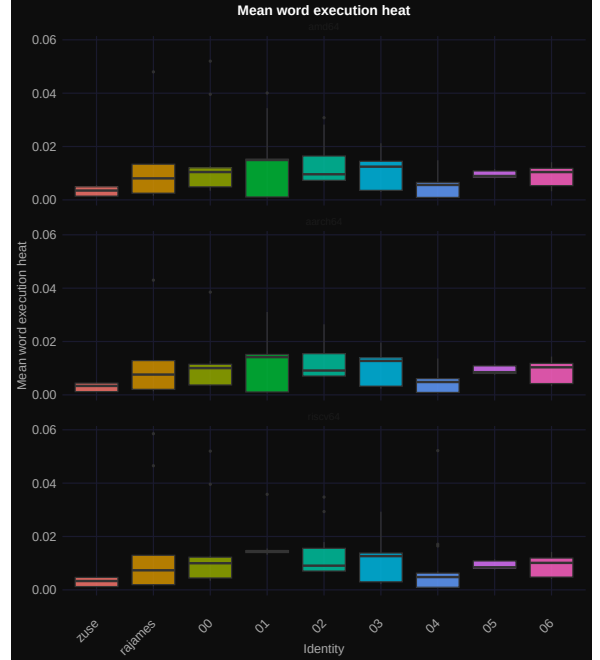


(b) Dark theme

Figure 2: Hot word count by identity, faceted by architecture — the metric with the strongest identity effect ($p \approx 6 \times 10^{-24}$). **zuse** (native compiled FORTH, no cross-VM dispatch) is visibly tighter/lower-variance than the eight VM-EXEC-dispatched identities.

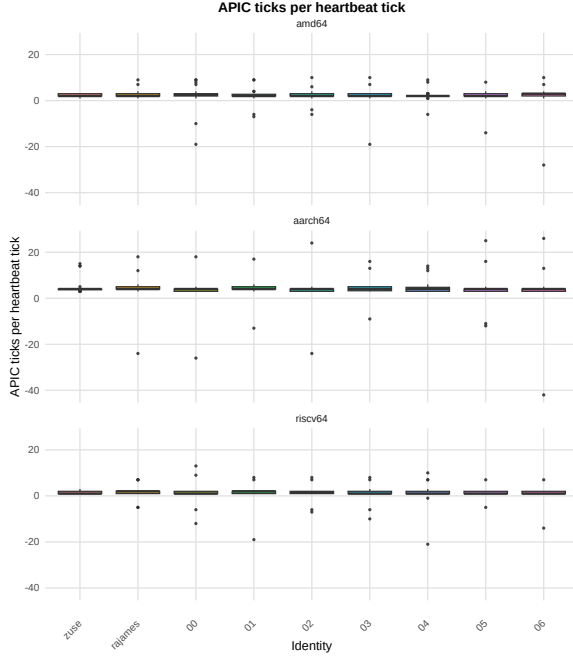


(a) Light theme

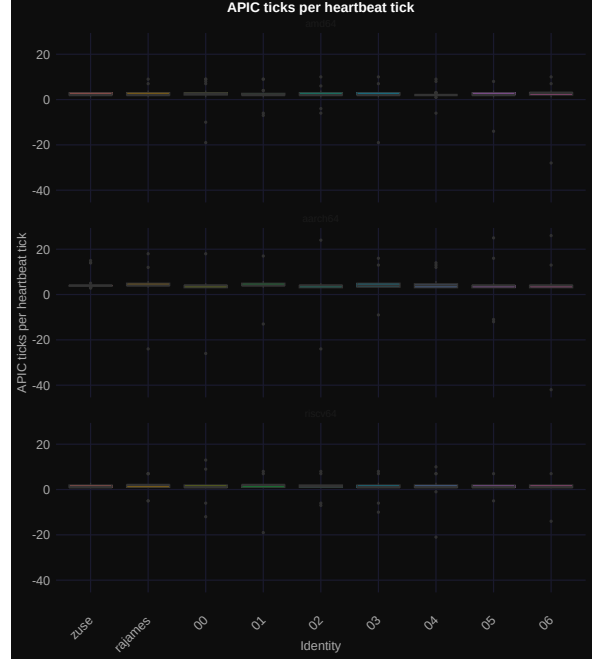


(b) Dark theme

Figure 3: Mean word execution heat by identity, faceted by architecture — the metric with the strongest effect of any kind in this dataset ($p \approx 5 \times 10^{-28}$ for identity).



(a) Light theme



(b) Dark theme

Figure 4: APIC hardware ticks consumed per heartbeat tick, by identity, faceted by architecture — the clearest architecture-only effect (no meaningful identity structure within a facet), reflecting each target’s distinct TCG emulation speed and timer hardware.

5 Statistical Analysis

5.1 K : Not Formally Testable, By Construction

No ANOVA is reported for K itself: with zero variance in the response (every one of 775 observations exactly 1.0), the F-statistic is undefined (division by a zero within-group sum of squares) rather than merely large. This absence of a test statistic *is* the result — the same convention this project uses elsewhere when a metric is reported as architecture-independent to machine precision rather than statistically indistinguishable at some p -value.

5.2 Secondary Metrics

Full per-cell (architecture $\times$ identity) means, standard deviations, and coefficients of variation, and the full ANOVA table for all twelve secondary metrics, are in `tables/cell_summary.csv` and `tables/anova.csv` alongside this report’s own source. Selected results, representative of the pattern in Table 2:

Table 3: Hot word count: per-architecture mean (averaged across all 9 identities), the metric showing both a strong architecture and a strong identity effect.

Architecture	Mean hot word count
amd64	58.99
aarch64	63.13
riscv64	63.02

6 Discussion

6.1 K as a Continuous, Not Just Boot-Time, Invariant

Prior verification of fleet conservation in this project checked K at specific diagnostic points (`vm_physics_status()`, `VM-CONSERVED?`). This campaign is the first to log K *continuously*, once per heartbeat tick,

through a sustained, realistic multi-identity workload — 80 recovered trials of cross-VM VM-EXEC dispatch, spanning VM admission, word-heat accumulation and cooling, and reservoir credit/debit on every eviction. That it holds exactly throughout, on every architecture, is a materially stronger claim than a boot-time self-test passing: it rules out slow leaks or transient double-counts that a single-point check would miss entirely.

6.2 A Free Regression Check on Two Recent Stadium Fixes

Both Stadium defects closed the same week this campaign ran (§1) touch exactly the code paths K depends on: the $O(n_{\text{cells}})$ scan fix changed *how* `stadium_admit()` finds an eviction candidate (walking a per-VM resident list instead of the whole array) without changing which cell it selects or how heat is credited back; the donor-floor fix changed *which* VM a new quota splits from (`stadium_best_donor()` instead of unconditionally Hera) without changing the reservoir math of the split itself. K ’s perfect conservation across every tick of this campaign is direct evidence that neither fix disturbed the underlying accounting — a correctness property those fixes’ own verification (81/81 trial correctness, FABRIC-3.md §XVI–XVII) did not directly probe.

6.3 Identity and Architecture Effects Are Separable

The near-total absence of an architecture $\times$ identity interaction (1 of 12 secondary metrics, and only weakly) is itself a useful property for anyone extending this instrumentation: architecture-specific timing behavior and identity-specific execution behavior can be reasoned about independently, without needing to model how one modulates the other.

6.4 Limitations

1. **QEMU TCG emulation.** All three architectures were tested under software emulation, not physical hardware; see the companion bare-metal DoE report for the corresponding caveat on that campaign.
2. **Interleaving-based trial mislabeling.** As documented in §2.3, 1 `run_id` out of 27 was unrecoverable on 2 of 3 architectures, folding that trial’s rows into the preceding one. This affects only `run_id/id_idx/id_label` attribution for those specific rows — `fleet_k_q48` itself, being read directly by `doe_log_tick_row()` at the moment of each tick regardless of which trial is running, is entirely unaffected by this correlation problem; K ’s perfect-conservation result in §3 does not depend on trial attribution being correct at all.
3. **Modest per-cell sample size.** Each architecture $\times$ identity cell contains roughly 7–10 heartbeat-tick observations (one DoE trial’s worth), not hundreds; the secondary-metric ANOVA results (§4) should be read as a first-pass factorial screen, not a high-power confirmatory study.
4. **Single seed.** All three architectures share one Fisher-Yates shuffle seed (12345) for direct comparability; this campaign does not itself establish invariance across independent run orderings the way the companion bare-metal report’s three-seed outer design does.

7 Conclusion

The fleet-wide execution-heat conservation invariant K holds exactly, at every heartbeat tick, across 775 observations spanning three architectures, nine identities, and three replicates of real multi-VM dispatch load — not one deviation. This closes the loop on the std79-doe campaign’s own recent history: two real Stadium subsystem defects were found and fixed (FABRIC-3.md §XVI–XVII), the full correctness campaign was re-verified 81/81 clean, and now the physics conservation law those fixes’ own code touches is independently confirmed intact, continuously, not just at the boot-time self-test that first established it. Architecture and identity each significantly affect several secondary timing and word-execution metrics, in the expected direction, with almost no interaction between them — and neither factor moves K at all.

*Still fucking chasing
the mongoose tail.*

— Captain Bob