Warming a Multilingual Static Fleet at the Edge: Throughput, Origin Load, and Fail-Closed Abort Rules

Jermaine Harris
Hutts Enterprises / Ferrus Labs
2026-09-20 · CC BY 4.0
Keywords: cs.PF, cs.NI, CDN, Cloudflare, static site generation
Source git 6ef9c2d · technical report

Abstract

Animal Start is a multilingual static fleet. Locale HTML is generated to disk origin, then filled into Cloudflare so later readers see HIT without asking origin for every slug. Warming is done from extra client hosts with a known user-agent and a per-language token bucket. The operational question is not “how fast can we go,” it is which (host count, rate) pairs stay healthy.

We report a 2026-08-14 throughput matrix of 16 cells extracted from the campaign log, plus a 2026-09-20 probe of CF-Cache-Status on two public URLs. Short bursts reached about 3.1k aggregate requests/s with origin load still under 4 and /es/ HTTP 200. Operational policy nevertheless caps sustained warm at about 800 rps (400 x 2 clients) because hotter cells and 800 x 4 configurations have fail-closed with Cloudflare NONE, 429, and BYPASS.

We did not rerun an --apply matrix in 2026-09: locale article slugs on disk origin were pruned, default hosts include blocklisted machines, and an uncapped warm is forbidden. The contribution is the abort envelope and labeled measurements, not a new CDN algorithm.

1. Introduction

A static locale fleet has two clocks. Generation writes HTML. The edge cache forgets. If operators wait for organic traffic, first readers pay origin. If operators warm too hard, Cloudflare Bot Fight and origin load both fail closed: 403 BYPASS, 429, Cloudflare 1106 from the wrong IP, or load that starves WordPress/SSR siblings on the same box.

Research question. What warm-client parallelism and per-language token-bucket rates keep a large locale static fleet healthy (origin load, /es/ 200) without tripping Bot Fight or saturating Cloudflare?

Contributions:

- C1. Sixteen historical matrix cells (Table 1) from logs/warm_matrix_bench.log (mtime 2026-08-14).
- C2. Abort rules now encoded in bench_cf_warm_matrix.py: /es/ must stay 200, load below 7.5, /es/ TTFB below 2.0 s.
- C3. A 2026-09-20 CF-Cache-Status probe (Table 2) that is not mixed into Table 1.

2. Operator roles

Hosts in this fleet are typed. English origin is W1. Locale disk origin is W2. Warm clients

should be L1 and similar extra boxes. W1, W2, and utility-one are blocklisted as warm clients: those IPs look like origin or trip Cloudflare 1106 / 403 BYPASS. The warmer user-agent is fleet-ssg-cf-warm/2.0. A bare bot UA is Bot Fight 403.

That split is easy to violate in a bench. The 2026-08-14 log used W2 as a warm client. Later policy forbids it. We keep the numbers and label the threat instead of deleting the only dated matrix.

3. System under test

fleet-topic-pipeline at 854caf16a7958bf0beed85fd00e9dd737b5c78ab. Domain: animalstart.com. Instrument: infra/fleet-ssg/bench_cf_warm_matrix.py. Cells are (n hosts, rate per language). Aggregate rps is observed, not the product n * rate, because languages and token buckets interact.

Abort if /es/ is not 200, origin load is at least 7.5, or /es/ is slower than 2.0 s. Uncapped warm requires FLEET_SSG_WARM_ALLOW_UNCAPPED=1, which this writing campaign did not set.

4. Method

Historical cells (2026-08-14). Parsed from the bench log summary JSON. Each cell records host count n, per-lang rate, observed aggregate rps, max load, /es/ status, serving_ok. Burst length used --limit 2000 in that campaign. That is a short fill, not a full-language crawl.

This campaign (2026-09-20). (1) Extract CSV only via harness/cdn-warm/extract_historical.py. (2) curl https://animalstart.com/es/ and / from the operator box (DFW) and record cf-cache-status and TTFB. (3) No --apply, no uncapped flag, no 800 x 4.

We did not re-bench because the 2026-09 slug export on disk origin was empty (locale articles had been pruned). Warming an empty catalog would measure Cloudflare against 404s or a stub hub, not the campaign the log describes.

5. Results

Table 1. Historical warm matrix, 2026-08-14, animalstart.com. Source: results/02/warm_matrix_cells.csv.

n	rate	agg rps	max load	/es/	ok
2	400	789	2.85	200	yes
2	500	991	3.38	200	yes
2	650	1275	4.02	200	yes
2	800	1561	4.28	200	yes
3	400	1187	3.79	200	yes
3	500	1484	3.75	200	yes
3	650	1919	3.30	200	yes
3	800	2338	2.96	200	yes
4	400	1586	3.51	200	yes

4	500	1980	3.00	200	yes
4	650	2555	2.81	200	yes
4	800	3109	2.98	200	yes
5	400	1846	3.23	200	yes
5	500	1979	3.14	200	yes
5	650	2562	3.81	200	yes
5	800	3115	3.85	200	yes

Every short burst stayed under the load abort (max 4.28 vs 7.5). Aggregate rps scales roughly with n * rate until about 3.1k, then 5 x 800 is indistinguishable from 4 x 800 (3115 vs 3109). That plateau is a Cloudflare or client-side ceiling, not origin CPU: load did not jump.

Docs for the same week: sustained campaign stays about 800 rps (400 x 2). 800 x 4 (about 3.2k) fail-closes with mass NONE / 429 / BYPASS even when a short bench burst still shows /es/ 200. Table 1 is therefore an existence proof that origin survived 2000 URLs, not a license for a full-language 3k rps campaign.

Table 2. Probe 2026-09-20 from operator box (DFW, cf-ray DFW). Not a warm matrix. Source: results/02/cf_status_probe_2026-09-20.json.

URL	HTTP	CF-Cache-Status	TTFB (s)
https://animalstart.com/es/	200	MISS	0.361
https://animalstart.com/	200	MISS	0.240

Headers also showed x-fleet-locale-disk: w2 and x-fleet-locale-origin: w2-disk on /es/. English / last-modified was 14 Sep 2026; /es/ last-modified was the probe instant. A MISS from DFW on a quiet Sunday is expected when nobody has warmed; it is not evidence that Table 1 cells are stale HIT.

6. How to read the two tables together

Mixing Table 1 aggregate rps with Table 2 MISS would be wrong. Table 1 is a 2026-08-14 burst from warm clients with a dedicated UA. Table 2 is one-off from a residential/office network. Cloudflare cache fill and Bot Fight behave differently over a full language than over --limit 2000.

The policy rule that survives both tables: cap sustained warm near 800 aggregate rps; abort on load and /es/; never use origin or utility-one as a client; do not set the uncapped flag because a matrix cell looked green.

7. Discussion and threats

Short bursts never tripped the load abort. That is why policy is not “whatever the matrix survived for 2000 URLs.” A language with tens of thousands of slugs keeps connections and Bot Fight state that a 2000-URL burst never sees.

Threats. (1) Historical host set includes W2. (2) 2026-09 slug export was empty, so we could not ethically re-bench. (3) CF status histogram was not in the 2026-08-14 logger (only

serving_ok, /es/, load, rps). (4) Single domain (animalstart.com); other locales and other properties may have different Bot Fight sensitivity. (5) Load is origin 1-minute load, not Cloudflare CPU. (6) Probe TTFB includes WAN to DFW PoP.

8. Related work

Cloudflare documents cache response headers (HIT, MISS, BYPASS, DYNAMIC, EXPIRED, NONE). This paper is operational measurement of one fleet, not a comparison to Fastly, Cloudfront, or academic CDN traces. Token-bucket warmers are a standard operations pattern; we do not claim a new algorithm.

9. Ethics and operational safety

We did not warm production in this writing campaign. Abort rules exist to protect origin and avoid 403 BYPASS storms. Utility-one must not be a warm client (Cloudflare 1106). Uncapped warm is a break-glass flag, not a default.

10. Conclusion

For this fleet, the safe sustained envelope remains about 800 aggregate rps with fail-closed abort on load and /es/. Burst matrices can look fine near 3k rps; that is not permission to run 800 x 4 as a campaign. New matrices need live slug sources and blocklist-compliant hosts.

Appendix A. Artifact

```
python3 harness/cdn-warm/extract_historical.py
```

Do not pass --apply unless slugs exist and hosts omit W1, W2, and utility-one.

References

- Cloudflare cache responses. https://developers.cloudflare.com/cache/concepts/cache-responses/
- fleet-topic-pipeline (pinned). git SHA 854caf16a7958bf0beed85fd00e9dd737b5c78ab