02 cdn warm

git a4aed0a | 2026-09-20 | Jermaine Harris, Hutts Enterprises / Ferrus Labs

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

Jermaine Harris

Hutts Enterprises / Ferrus Labs

date: "2026-09-20"

license: "CC BY 4.0"

keywords: "cs.PF, cs.NI, CDN, Cloudflare, static site generation"

Abstract

A multilingual static-site campaign for Animal Start warms Cloudflare from extra client hosts so locale HTML is HIT at the edge without overloading origin. We report a 2026-08-14 throughput matrix (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 (`400×2` clients) because hotter cells and `800×4` configurations have fail-closed with Cloudflare `NONE`/`429`/`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.

Keywords

cs.PF; CDN; Cloudflare; static generation; cache warming

1. Introduction

RQ. 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. Table of 16 historical matrix cells (Table 1) from `logs/warm_matrix_bench.log` (mtime 2026-08-14).
- C2. Documented abort rules now encoded in `bench_cf_warm_matrix.py`.
- C3. A 2026-09-20 `CF-Cache-Status` probe (Table 2) that is *not* mixed into Table 1.

2. System under test

`fleet-topic-pipeline` at `854caf16a7958bf0beed85fd00e9dd737b5c78ab`. Domain under test: `animalstart.com`. English origin role **W1**; locale disk origin **W2**; warm clients should be **L1** (and similar), never W1/W2/utility-one (403 BYPASS / origin / Cloudflare 1106). Warmer user-agent is `fleet-ssg-cf-warm/2.0`; a bare bot UA is Bot Fight 403.

Instrument: `infra/fleet-ssg/bench_cf_warm_matrix.py`. Abort if `/es/` ≠ 200, load ≥ 7.5, or `/es/` slower than 2.0 s.

3. Method

Historical cells (2026-08-14). Parsed from the bench log summary JSON. Each cell: host count n, per-lang rate, observed aggregate rps, max load, `/es/` status, serving_ok. Burst length used `--limit 2000` in that campaign. That run included **W2 as a warm client**, which later policy forbids. We keep the numbers and label the threat.

This campaign (2026-09-20). (1) Extract CSV only. (2) `curl` `https://animalstart.com/es/` and `/` from the operator box and record `cf-cache-status`. (3) No `--apply`, no `FLEET_SSG_WARM_ALLOW_UNCAPPED=1`, no 800×4.

4. Results

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

n hosts	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

Docs for the same week: sustained campaign stays ~800 rps (`400×2`); `800×4` (~3.2k) fail-closes with mass `NONE`/`429`/`BYPASS` even when a short bench burst still shows `/es/` 200.

Table 2. Probe 2026-09-20 from operator box (DFW). Not a warm matrix.

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` on `/es/`.

5. Discussion and threats

Short bursts in Table 1 never tripped the load abort (all max_load ≤ 4.28 ≪ 7.5). That is why policy is not “whatever the matrix survived for 2000 URLs.” Cloudflare Bot Fight and cache fill behave differently over a full language. Mixing Table 1 agg rps with Table 2 MISS would be wrong: the probe is one-off from a residential/office network.

Threats. Historical host set includes W2. Slug export in 2026-09 was empty (pruned locale articles), so we could not ethically re-bench. CF status histogram was not in the 2026-08-14 logger (only serving_ok / es / load / rps).

6. Related work

Cloudflare cache response headers [cloudflare-cache]. This is operational measurement of one fleet, not a comparison to other CDNs.

7. 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).

8. Conclusion

For this fleet, the safe *sustained* envelope remains ~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 `800×4` as a campaign. New matrices need live slug sources and blocklist-compliant hosts.

9. Artifact appendix

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

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

References

- Cloudflare cache responses. https://developers.cloudflare.com/cache/concepts/cache-responses/