

Edge Inference Gateway v2

Document owner	Platform Architecture
Review window	March 2026
Status	Draft for architecture review

This design review proposes a new edge gateway tier for Basethree's LLM inference platform. The goal is to reduce cold-start latency for high-volume chat workloads, isolate noisy tenants before they reach regional control planes, and make rollout decisions easier to audit after Google Docs import.

Design Goals

- Reduce p95 first-token latency for interactive traffic without changing the regional serving contract.
- Keep the blast radius bounded when `edge-router` or `session-cache` signals degrade.
- Make canary decisions legible for reviewers who rely on shared markdown and generated DOCX artifacts.
- Preserve a clear fallback path to regional-only routing during the first launch wave.

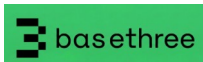
Non-Goals

- We are not replacing the regional control plane in this review.
- We are not promising per-tenant dedicated edge clusters in the first release window.
- The earlier `global-write-through-cache` concept remains out of scope for this RFC.

Options Compared

Option	p95 latency	Rollout complexity	Failure isolation	Recommendation
Regional-only routing	1.00x	Low	Low	Stable baseline, but does not improve user-perceived responsiveness enough
Edge cache plus regional fallback	0.72x	Medium	High	Recommended default for interactive chat traffic
Dedicated per-tenant edge clusters	0.64x	High	Medium	Too expensive and operationally noisy for the first release window

Canary Rollout Plan



The rollout plan is intentionally written as a nested operational checklist so reviewers can validate both structure and traceability in the generated document.

1. Control plane readiness

- Validate tenant routing tags, rollout labels, and reviewer metadata.
- Freeze routing-config churn for the final 24 hours before the canary begins.

If reviewer annotations are missing from the weekly change log, stop here and do not expand the canary.

Gate	Threshold	Primary owner
Gateway 5xx rate	< 0.5%	Inference
p95 TTFT	< 750 ms	Platform
Audit tag coverage	100%	Developer Productivity

2. Enterprise canary

2.1 Cohort selection

- Pick one low-variance tenant cohort with predictable request shapes.
- Keep one opt-out route pointed at the regional baseline.

2.2 Regional burn-in

- Observe latency, cache churn, and rollback frequency for three business days.
- Record every routing override in the [incident playbook](#).

3. General rollout

- Expand to broader traffic only after the review committee signs off on the canary notes.
- Keep one emergency switch for immediate regional-only routing.

Reviewer note: prefer predictable rollback behavior over maximum cache hit rate during the first two launch waves.

Example Edge Policy

```
router:
  mode: edge-first
  fallback: regional-only
  cache:
    strategy: session-aware
    ttl_seconds: 90
  rollback:
    trigger: p95_ttft_ms > 750 or gateway_5xx_rate > 0.5%
```

Rollback Drill

```
basethree rollout status edge-gateway-v2
basethree rollout shift edge-gateway-v2 --mode regional-only
basethree alerts ack inference/edge-gateway
```

API Surface Proposal

The edge gateway should be legible not only as infrastructure, but also as a public-facing API surface that product teams can integrate with directly.

Endpoint	Method	Purpose	Auth
/v2/chat/completions	POST	Session-aware chat inference with optional edge caching	API key
/v2/responses/{id}	GET	Fetch a previously created response by identifier	API key
/v2/models	GET	Enumerate deployable model IDs and rollout metadata	API key

Request Fields

Field	Required	Type	Notes
model	yes	string	Must match a model that is enabled in the tenant's rollout policy
messages	yes	array	Ordered chat messages passed to the gateway
metadata.trace_id	no	string	Helps support teams correlate edge decisions with logs
cache_control.mode	no	string	Accepts prefer-cache, bypass, or inherit

Example Request

```
{
  "model": "bt-reasoner-2",
  "messages": [
    {"role": "system", "content": "You are a routing-aware assistant."},
    {"role": "user", "content": "Summarize the latest latency regression."}
  ],
  "metadata": {
    "trace_id": "trace_01hzy0example"
  },
  "cache_control": {
    "mode": "prefer-cache"
  }
}
```

Example Response

```
{
  "id": "resp_01hzy0example",
  "model": "bt-reasoner-2",
  "gateway": {
    "route": "edge-sjc1",
    "cache": "hit",
    "fallback_used": false
  },
  "output": [
    {"type": "message", "role": "assistant", "content": "Latency stabilized after the regional rebalance."}
  ]
}
```

Error Catalog

Code	Meaning	Suggested action
tenant_not_enabled	The tenant is not yet in the edge rollout cohort	Retry against the regional baseline or wait for rollout approval
cache_mode_invalid	The caller passed an unsupported cache mode	Validate the request against the published API schema
gateway_overloaded	The edge tier refused the request to protect latency SLOs	Retry with backoff or force regional-only routing

Risk Register

Risk	Trigger	Mitigation	Owner
Cache stampede	Sudden tenant burst after a release	Cap concurrent fills and bias to regional fallback	Inference
Uneven capacity	Regional pool already near saturation	Hold edge rollout behind demand-aware admission checks	Capacity
Audit ambiguity	Reviewers cannot explain why a request was routed differently	Log decision tags and persist rollout annotations in weekly review docs	Platform
Rollback confusion	Operators disagree on when the gateway should stand down	Keep one published threshold table and one named owner per gate	Architecture



Brand Mark In Review Deck

The following standalone image verifies anchored body-image sizing separately from the page-header treatment.



Basethree reference mark used in the design review deck

Launch Decision

Proceed with a canary in one region, one enterprise tenant cohort, and one internal dogfood route. Success should be measured against p95 first-token latency, gateway saturation, and the time required to explain routing outcomes during review.

APPENDIX A - ROLLOUT METRICS

Metric	Green	Yellow	Red	Owner
p95 TTFT	< 650	650-750	> 750	Inference
Gateway 5xx rate	< 0.2%	0.2-0.5%	> 0.5%	Platform
Cache reuse rate	> 30%	20-30%	< 20%	Inference
Reviewer annotation coverage	100%	95-99%	< 95%	Dev Prod

- Primary scorecard:
 - p95 TTFT by route
 - cache reuse rate by workload type
 - rollback frequency per tenant cohort
- Qualitative review prompts:
 - Can a support engineer explain the route that was selected?
 - Can a reviewer map the routing behavior back to the markdown source and sidecar overrides?