Latency Budget Spec

The Latency Budget Spec translates user experience and operational requirements into per-task-class time ceilings that constrain model tier selection, cache strategy, and batch eligibility in e6-routing-policy. Without declared latency budgets, the routing policy has no signal for when a cheaper-but-slower model tier is acceptable and when a faster path is required regardless of cost. This spec also identifies the principal bottleneck in each task class so optimization effort is applied to the right layer. Latency budget enforcement is a joint responsibility: the routing policy enforces per-call ceilings; this spec declares the end-to-end ceilings from the user's perspective.

Task Class Taxonomy

Each workflow is broken into task classes by the latency sensitivity of their triggering interaction. Three primary classes cover most agent use cases:

Interactive — A human is actively waiting for a response. The user perceives latency directly as responsiveness. Ceiling is typically 3–8 seconds for first substantive token; 15–30 seconds for a full structured response. Premium tier is acceptable only if it fits within the ceiling; if not, the routing policy must use a faster tier or a streaming approach.

Asynchronous — A human or system expects a result in a defined window (minutes to hours) but is not blocking on it. The user does not perceive sub-minute latency variations. Ceilings in the range of 2–30 minutes. Premium tier is almost always viable within an async window.

Batch / scheduled — Runs on a fixed cadence (nightly, weekly) or as a background sweep. The output is consumed later; latency ceiling is hours to a day. These runs should always use the cheapest viable tier and be aggregated for cache pre-warming.

Fillable Fields

```
agent_id: <matches s1-operating-context-canvas>
latency_budget_version: <semver>
effective_date: <ISO-8601>

task_classes:
  - task_class_id: <kebab-case>
    description: <one sentence>
    interaction_type: <interactive | async | batch>
    user_expectation_source: <"user research" | "SLA agreement" | "regulatory" | "internal
policy">
    ceiling_ms: <integer — end-to-end from request receipt to response delivery>
    p95_target_ms: <integer — acceptable 95th-percentile observed latency>
    p99_ceiling_ms: <integer — hard ceiling at p99; breaches open an incident>
    model_tier_ceiling: <premium | standard | cheap | deterministic — fastest tier that
fits the budget AND ceiling>
    bottleneck_layer: <model-inference | retrieval | tool-call | orchestration | network |
human-review>
    bottleneck_mitigation: <one sentence>
    cache_eligible: <true | false>
    batch_eligible: <true | false>
    streaming_required: <true | false>
```

```
latency_budget_alerts:
  p95_breach_action: <log-only | soft-alert | hard-gate>
  p99_breach_action: <log-only | soft-alert | hard-gate>
  sustained_breach_threshold_pct: <integer — % of runs above ceiling before escalation>
  escalation_target_ref: <s1-threshold-escalation-spec>
```

Worked Example: TownOracle Community-Q&A Agent (illustrative)

TownOracle.AI serves South Walton community members with answers to local government questions. The agent must feel responsive to a constituent on a mobile device while also running nightly batch sweeps to pre-warm its knowledge cache. All latency figures are (illustrative).

```
agent_id: townoracle-community-qa
latency_budget_version: 0.1.0
effective_date: 2026-05-29

task_classes:
  - task_class_id: live-constituent-answer
    description: Real-time answer to a constituent question submitted via the TownOracle
chat interface.
    interaction_type: interactive
    user_expectation_source: user research
    ceiling_ms: 8000        # (illustrative) 8 seconds total
    p95_target_ms: 5500     # (illustrative)
    p99_ceiling_ms: 11000   # (illustrative)
    model_tier_ceiling: standard
    bottleneck_layer: retrieval
    bottleneck_mitigation: Pre-embed the top-200 community FAQs on each nightly sweep;
retrieval for cached queries completes in ~80ms.
    cache_eligible: true
    batch_eligible: false
    streaming_required: true

  - task_class_id: permit-status-brief
    description: Asynchronous briefing for a clerk on the status of a specific permit
application; delivered in the ops dashboard within 10 minutes.
    interaction_type: async
    user_expectation_source: internal policy
    ceiling_ms: 600000      # (illustrative) 10 minutes
    p95_target_ms: 240000   # (illustrative) 4 minutes
    p99_ceiling_ms: 480000  # (illustrative) 8 minutes
    model_tier_ceiling: premium
    bottleneck_layer: tool-call
    bottleneck_mitigation: GIS and permit-API calls are serialized; parallelize read-only
lookups where permit schema allows.
    cache_eligible: false
    batch_eligible: false
    streaming_required: false

  - task_class_id: nightly-faq-refresh
    description: Nightly sweep that re-embeds updated FAQs, refreshes the vector store, and
pre-warms the cache for the next day's interactive queries.
    interaction_type: batch
    user_expectation_source: internal policy
    ceiling_ms: 10800000    # (illustrative) 3 hours
    p95_target_ms: 7200000  # (illustrative) 2 hours
    p99_ceiling_ms: 10800000
    model_tier_ceiling: cheap
    bottleneck_layer: retrieval
    bottleneck_mitigation: Parallelize embedding generation across FAQ sections; rate-limit
embedding API calls to avoid 429 errors.
    cache_eligible: false
    batch_eligible: true
```

```
    streaming_required: false

  - task_class_id: enforcement-notice-draft
    description: Draft an enforcement notice citing applicable South Walton County
ordinance sections; always followed by clerk HITL review.
    interaction_type: async
    user_expectation_source: regulatory
    ceiling_ms: 120000      # (illustrative) 2 minutes to draft stage; HITL adds separate
time
    p95_target_ms: 90000    # (illustrative)
    p99_ceiling_ms: 180000  # (illustrative)
    model_tier_ceiling: premium
    bottleneck_layer: model-inference
    bottleneck_mitigation: Premium models have higher variance latency; reserve a dedicated
routing slot during business hours. Enforce streaming for progressive disclosure to
reviewing clerk.
    cache_eligible: false
    batch_eligible: false
    streaming_required: true

latency_budget_alerts:
  p95_breach_action: soft-alert
  p99_breach_action: hard-gate
  sustained_breach_threshold_pct: 5
  escalation_target_ref: s1-threshold-escalation-spec
```

Latency Budget vs. Cost Trade-offs

The table below captures the design tension the routing policy must resolve:

Trade-off axis	Interactive tasks	Async tasks	Batch tasks
Model tier	Standard/cheap preferred; premium only if ceiling permits	Premium acceptable	Cheap always
Streaming	Required for ceiling compliance	Optional	Never
Cache benefit	High — hit rate directly cuts p95	Medium	Low — batch re-populates cache
Bottleneck priority	Retrieval and first-token latency	Tool-call round-trips	Throughput (tokens/sec at scale)
Cost sensitivity	Medium — latency is primary	Low — time budget is wide	High — batch runs are bulk spend

Bottleneck Identification Protocol

When a task class consistently misses its p95 target, the routing policy should trigger a bottleneck investigation before adjusting tier or ceiling. Investigation order:

1. Retrieval — Check embedding query time and vector-search time in C7 est_latency_ms records. If retrieval exceeds 30% of ceiling, pre-embed or reduce chunk count.

2. Tool calls — Check tool-call latency_ms in s2-audit-trail-schema. Serialize vs. parallelize analysis. Connector timeouts often dominate; see failure mode tool-timeout in s1-failure-mode-register.

3. Model inference — Actual inference latency varies by tier and time-of-day. Measure observed vs. estimated in C7 records; persistent gaps indicate the estimate in e6-cost-model needs updating.

4. Orchestration — Token-passing overhead between planner and executor can accumulate in multi-step workflows. Check step_id durations in the C5 provenance records from e4-workflow-artifacts.

5. Network — Egress to external connectors. Rarely the dominant factor but important for cross-region deployments.

Open an optimization task in e6-waste-reduction-playbook for any bottleneck that is not addressed within one release cycle.

Latency Budget Maintenance

Review this spec on every router configuration change, every new model tier onboarding, and every quarter. Observed p95 values from the C7 audit records are the authoritative input to the next review. If the observed p95 for any task class exceeds its p99_ceiling_ms, escalate via s1-threshold-escalation-spec immediately rather than waiting for the next review cycle.