Rate-Limit Forecast Model

claumon `internal/forecast`

Model version **v2.1** - 2026-06-10

This document specifies the math behind `internal/forecast`. It is normative: implementations must match the formulas here, and deviations are bugs unless this document is updated. The model version above matches `forecast.ModelVersion` in the Go code; retired model versions are preserved under `internal/forecast/archive/v`*N.M*`/`, and `CHANGELOG.md` summarises each bump.

1 What we forecast

For each open rate-limit window (session, weekly, per-model weekly), the forecaster reads the time series of utilization snapshots and produces:

1. A point estimate F of the utilization at reset.
2. An 80% **credible interval** $[F^-, F^+]$ for F. (We use the looser "confidence" wording in user-facing strings only; the posterior is Bayesian, see §5.)
3. For each threshold C_{thr} (e.g. 100%), either a median ETA with an 80% CI, or `nil` if the threshold is unlikely to be reached before reset.

This procedure is an *empirical Bayes* forecast [1]: the prior on the rate r and the path-noise variance σ^2 are both estimated from past data (§5, §7) and then plugged into the posterior update, rather than being themselves marginalized under a hyperprior. The conjugate update on r and the Monte Carlo over rate-and-path uncertainty are genuinely Bayesian; the "empirical" qualifier refers to the prior, not the inference step.

A separate forecast is produced per gauge. The only cross-window sharing is through the calibration constants (§7).

2 The model in one paragraph

Inside the current window, utilization $u(t)$ accumulates at an unknown mean rate r through strictly non-negative increments: tokens are spent, never returned. We estimate r from the recent slope of the snapshots (OLS) and from a prior fit on past sessions, combined by conjugate update, exactly as in v1.x. What changed in v2.0 is the *path law*: instead of Brownian noise (which can drift down), each increment is Gamma-distributed, so simulated trajectories are monotone non-decreasing and bounded below by u_{now}. The point forecast $F = u_{\text{now}} + \hat{r}_{\text{post}}\,\Delta t_{\text{rem}}$ and its moment spread σ_F are unchanged, but the 80% CI and the ETA are now both read off a single Monte Carlo over rate and path uncertainty: the predictive at reset is the right-skewed, u_{now}-floored law the paths trace out, not a symmetric Gaussian z-quantile.

3 Generative model and assumptions

Condition on a non-negative mean rate r. For $s \in [0, \Delta t_{\text{rem}}]$ with $\Delta t_{\text{rem}} = T_{\text{reset}} - t_{\text{now}}$, utilization accumulates as a Gamma process, a non-decreasing Lévy subordinator:

$$u(t_{\text{now}} + s) - u_{\text{now}} \;\sim\; \text{Gamma}\Big(\text{shape} = \tfrac{r^2 s}{\sigma^2_{\text{session}}},\; \text{scale} = \tfrac{\sigma^2_{\text{session}}}{r}\Big),$$

so that, given r,

$$\mathbb{E}\big[u(t_{\text{now}} + s) - u_{\text{now}}\big] = r\, s, \qquad \text{Var}\big[u(t_{\text{now}} + s) - u_{\text{now}}\big] = \sigma^2_{\text{session}}\, s.$$

The mean and variance match the v1.x Brownian term-for-term; only the *shape* changes, from a symmetric Gaussian that can decrease to a right-skewed law with support $[u_{\text{now}}, \infty)$. Increments over disjoint intervals are independent, so the process is monotone non-decreasing by construction. The rate r is itself unknown and non-negative, with mean $\hat{r}_{\text{post}}$ and variance $\max(\tau^2_{\text{post}}, \bar{\tau}^2)$ (§5, §7); the Monte Carlo (§8) draws it from a Gamma matched to those moments. (The §5 posterior on r is Gaussian; we sample a moment-matched Gamma instead so the drawn rate stays positive, since a negative rate is unphysical here and would leave the Gamma increment ill-defined.) Three assumptions:

1. **Constant mean rate.** r is constant within a window. Violated by abrupt mid-window mode shifts; partially absorbed by the recency window in §5.
2. **Gamma path noise, variance linear in time.** As in v1.x the per-hour increment variance is the single calibrated scalar $\sigma^2_{\text{session}}$; the Gamma process is the non-decreasing counterpart whose increment variance also grows linearly in time, so the §7 calibration carries over unchanged.
3. **Exchangeable past sessions.** Historical sessions are iid draws from the same process. Required for the calibration in §7.

Why a Gamma process (v2.0). Real utilization is monotone non-decreasing in $[0, 1]$: tokens accumulate, the gauge only resets at the window boundary. v1.x used Brownian motion for its closed-form Gaussian CI, at the cost of paths that could drift below u_{now}, above 1, or below 0 (none physical), patched by clipping the displayed CI to $[u_{\text{now}}, 1]$. v2.0 removes the patch by adopting the Gamma process that v1.x's own notes flagged as the principled fix. Three consequences, all improvements:

- **Lower bound for free.** Every increment is ≥ 0, so the 80% CI's lower edge sits at or above u_{now} with no clipping, and the predictive at reset is the right-skewed, u_{now}-floored law the paths trace out.
- **Honest first-passage.** A monotone path crosses a threshold once and stays above it, so the ETA is never the artifact of a Brownian path that dipped below and re-crossed later (the "biased late" effect v1.x flagged in §8).
- **Faithful visualisation.** The forecast-trajectory modal draws the raw §8 paths; under v2.0 they no longer dip.

The cost is that the forecast law is no longer Gaussian, so the CI and the ETA are both read off the Monte Carlo of §8 rather than a closed-form z-quantile. The point forecast F and its moment variance σ^2_F (§6) are unchanged.

4 Notation

Symbol	Meaning
$u(t)$, u_{now}	utilization at time t and at now, $\in [0,1]$
t_{now}, T_{reset}, Δt_{rem}	now, reset, remaining horizon (h)
r	true mean growth rate within window (h^{-1})
$\hat{r}_{\text{OLS}}$, SE_{OLS}	OLS estimate and its standard error
μ_0, τ_0^2	Gaussian prior on r from history
$\hat{r}_{\text{post}}$, τ^2_{post}	posterior on r after the OLS update
$\sigma^2_{\text{session}}$	path-noise variance coefficient (h^{-1})
F, σ_F^2	point forecast and its variance
$z_{0.9}$	$\Phi^{-1}(0.9) \approx 1.2816$
K, Δs	MC trajectories, step size (defaults: 500, 5 min)

5 Estimating the rate

Combine two sources of information about r.

From the current window. Let $\mathcal{R}$ be the snapshots in $[t_{\text{now}} - \tau_{\text{recent}}, t_{\text{now}}]$ and $n = |\mathcal{R}|$. For $n \geq 3$, fit $u_i = \alpha + r\, t_i + \epsilon_i$ by OLS:

$$\hat{r}_{\text{OLS}} = \frac{S_{tu}}{S_{tt}}, \qquad \text{SE}^2_{\text{OLS}} = \frac{\hat{\sigma}^2_\epsilon}{S_{tt}},$$

with $S_{tt} = \sum(t_i - \bar{t})^2$, $S_{tu} = \sum(t_i - \bar{t})(u_i - \bar{u})$, and $\hat{\sigma}^2_\epsilon = \frac{1}{n-2}\sum(u_i - \hat{\alpha} - \hat{r}_{\text{OLS}}\, t_i)^2$. Default $\tau_{\text{recent}} = 30\,\text{min}$.

Strictly, the path residuals are heteroskedastic and serially correlated, so the iid-OLS variance formula above is an approximation. For the short recency window the discrepancy is small and the point estimate is unbiased either way; we accept the approximation and treat SE^2_{OLS} as the likelihood precision in the conjugate update below.

From history. For each past completed session s with final value u^*_s and duration D_s, define $\rho_s = u^*_s / D_s$. The prior mean is the sample mean of ρ_s. For the prior variance, note that ρ_s is the session's mean rate r_s plus its centered path-noise increment averaged over the window, so under the generative model

$$\text{Var}[\rho_s] = \text{Var}[r_s] + \frac{\sigma^2_{\text{session}}}{D_s}$$

(the path-noise increment over D_s has variance $\sigma^2_{\text{session}} D_s$, so dividing by D_s contributes $\sigma^2_{\text{session}}/D_s$). The raw sample variance of ρ_s therefore overstates $\text{Var}[r_s]$ by the average path-noise contribution. Subtract it off:

$$\mu_0 = \text{mean}_s\, \rho_s, \qquad \tau_0^2 = \max\Big(\text{var}_s\, \rho_s - \sigma^2_{\text{session}} \cdot \text{mean}_s(1/D_s),\ \varepsilon\Big),$$

with $\varepsilon = 10^{-6}$ guarding against negative estimates from small samples. This uses the most recent $\sigma^2_{\text{session}}$ from §7; the two fits depend on each other but only loosely, and the daily refit cycle converges quickly. On the very first fit (before §7 has run), $\sigma^2_{\text{session}} = 0$ and the correction is a no-op.

Fit at startup, refresh daily.

Combine. Normal-normal conjugacy gives the posterior $r \mid \hat{r}_{\text{OLS}} \sim \mathcal{N}(\hat{r}_{\text{post}}, \tau^2_{\text{post}})$:

$$\frac{1}{\tau^2_{\text{post}}} = \frac{1}{\tau_0^2} + \frac{1}{\text{SE}^2_{\text{OLS}}}, \qquad \hat{r}_{\text{post}} = \tau^2_{\text{post}} \left(\frac{\mu_0}{\tau_0^2} + \frac{\hat{r}_{\text{OLS}}}{\text{SE}^2_{\text{OLS}}} \right).$$

Limits: prior dominates when SE_{OLS} is large (early window); data dominates when SE_{OLS} is small (later in the window). For $n < 3$ the OLS step is undefined; use $\hat{r}_{\text{post}} = \mu_0$, $\tau^2_{\text{post}} = \tau_0^2$.

6 Forecast distribution at reset

The point forecast is

$$F = u_{\text{now}} + \hat{r}_{\text{post}}\, \Delta t_{\text{rem}}.$$

By the law of total variance, conditioning on r:

$$\sigma_F^2 = \underbrace{\Delta t_{\text{rem}}^2 \cdot \max(\tau^2_{\text{post}},\, \bar{\tau}^2)}_{\text{rate uncertainty}} + \underbrace{\Delta t_{\text{rem}}\, \sigma^2_{\text{session}}}_{\text{path noise}},$$

where $\bar{\tau}^2$ is the historical-average rate variance recovered in §7 (the quadratic coefficient $\hat{b}$ of that regression). The floor $\max(\tau^2_{\text{post}}, \bar{\tau}^2)$ is new in v1.1 and corrects a systematic under-spread when assumption A1 (constant within-window rate) is violated: the conjugate τ^2_{post} then accurately estimates uncertainty about the *recent* slope but understates uncertainty about the *end-of-session* slope, because the user pivots. The historical $\bar{\tau}^2$, fit from past end-of-session errors, is the right scale for the latter. We take the max rather than always using $\bar{\tau}^2$ because for short windows where τ^2_{post} is large (few snapshots) the conjugate value should win.

Status as a moment summary. The floor is a robustness correction rather than a Bayesian update: $\max(\tau^2_{\text{post}}, \bar{\tau}^2)$ is not the variance of any coherent posterior over r, so σ_F^2 should be read as the predictive variance with τ^2_{post} replaced by an empirical lower bound. In v2.0 the marginal forecast is the right-skewed Monte Carlo law (§8: draw r from a Gamma of mean $\hat{r}_{\text{post}}$ and variance $\max(\tau^2_{\text{post}}, \bar{\tau}^2)$, then accumulate Gamma increments of per-step variance $\sigma^2_{\text{session}}\, dt$), but σ_F^2 remains its variance, with two contributions at different scales: rate uncertainty is quadratic in horizon (dominates at long horizons), path noise is linear (dominates at short).

80% CI. The interval is the 10th and 90th percentiles of the Monte Carlo terminal distribution $\{u_k(T_{\text{reset}})\}$ (§8), not a Gaussian z-quantile. Because every increment is non-negative the lower percentile is $\geq u_{\text{now}}$ automatically (no clip) and the band is right-skewed. Both edges are reported as-is, with no cap at 1 (v2.1): the forecast measures projected *demand*, which can exceed the window limit even though the physical gauge saturates there, and overshoot magnitude is informative. v2.0 capped only the upper edge, which inverted the reported interval whenever the lower percentile itself exceeded 1 (e.g. "80% CI 134%–100%"). σ_F above is retained as a symmetric moment summary (and to parameterise the MC draws); F remains the reported point estimate.

7 Calibration of $\sigma^2_{\text{session}}$

What we learn. One scalar: how much utilization can drift from the trend per hour of waiting. It is the only quantity history contributes to the forecast *spread* (the prior μ_0, τ_0^2 contributes to the *mean*).

Strategy. Run the forecaster on past sessions, where we already know the answer, and measure how wrong it was as a function of horizon. Read $\sigma^2_{\text{session}}$ off the empirical error structure.

Replay. For each past session s with reset value u_s^* at T_s, sample forecast points $t_f \in [t_s + \tau_{\text{recent}},\, T_s - \Delta_{\min}]$ (default 6 per session, $\Delta_{\min} = 30\,\text{min}$). At each t_f, run §5 on snapshots in $[t_f - \tau_{\text{recent}},\, t_f]$ to obtain $\hat{r}_{\text{post}}(t_f)$, and the projection $\hat{F}(t_f) = u(t_f) + \hat{r}_{\text{post}}(t_f)\,(T_s - t_f)$. Record

$$e_f = u_s^* - \hat{F}(t_f), \qquad \delta_f = T_s - t_f.$$

Two sources of error in each residual. The residual e_f mixes (A) the rate estimate being off at t_f, with the error compounded over δ_f, and (B) path noise accumulating over δ_f. By the same decomposition as §5,

$$\mathbb{E}[e_f^2 \mid \delta_f] = \underbrace{\delta_f\, \sigma^2_{\text{session}}}_{\text{path noise (B)}} + \underbrace{\delta_f^2\, \bar{\tau}^2}_{\text{rate uncertainty (A)}},$$

with $\bar{\tau}^2$ the average posterior rate variance at the forecast points. The two contributions scale differently in δ_f, which is what lets us separate them. Naive averaging like $\sigma^2_{\text{session}} \approx \text{mean}_f(e_f^2/\delta_f)$ would not separate them and would bias the estimate upward by a $\delta_f\,\bar{\tau}^2$ contamination.

Joint regression (v1.2: weighted). Fit both unknowns by *weighted* least squares of e_f^2 on $[\delta_f,\, \delta_f^2]$ (no intercept), with weights $w_f = 1/\delta_f^2$:

$$(\hat{a}, \hat{b}) = \arg\min_{a,b} \sum_f w_f \big(e_f^2 - a\,\delta_f - b\,\delta_f^2\big)^2, \qquad \sigma^2_{\text{session}} := \max(\hat{a},\, 10^{-6}).$$

The coefficient $\hat{a}$ on the linear term is the per-hour path-noise variance: that is the quantity we want. The weighting corrects heteroskedasticity: e_f^2 is a squared error whose own variance scales as $(\mathbb{E}[e_f^2])^2$ and so grows steeply with δ_f, so the unweighted OLS of v1.0–v1.1 was dominated by the few long-horizon points and inflated $\hat{a}$ - over-predicting short-horizon spread by roughly an order of magnitude, and (through the noise correction in §5) over-subtracting until τ_0^2 floored to ε, which silently pinned the rate prior. Weighting by $1/\delta_f^2$ is a fixed proxy for $1/\text{Var}[e_f^2]$ that puts short- and long-horizon residuals on a comparable scale; the quadratic term $\hat{b}$ stays identified because only the long-horizon points carry information about δ_f^2.

Note that this regression treats $\bar{\tau}^2$ as constant across forecast points, while in reality $\tau^2_{\text{post}}(t_f)$ varies (it depends on how many recent snapshots existed at t_f). If $\tau^2_{\text{post}}(t_f)$ correlates with δ_f (and it usually does, since δ_f is anti-correlated with elapsed time in the session), the missing covariance leaks into $\hat{a}$. A tighter alternative is to subtract the known $\delta_f^2\,\tau^2_{\text{post}}(t_f)$ piece as an offset and regress the residual on δ_f alone; we keep the joint regression for simplicity and absorb the bias into the daily refit.

Role of $\hat{b}$ (v1.1: kept as a floor). Conceptually $\hat{b}$ estimates the same quantity as τ^2_{post} in §5 (the rate variance relevant to predicting $u(T_{\text{reset}})$), just as a historical average instead of a fresh per-forecast value. The original v1.0 design discarded $\hat{b}$ on the grounds that the per-forecast τ^2_{post} uses today's actual snapshots and is therefore strictly more informative.

In practice, on real sessions, τ^2_{post} shrinks to one or two orders of magnitude below $\hat{b}$ within a few snapshots of OLS. Empirically, the $\hat{b}$-sized variance is the one that matches the realized end-of-session errors. The reason is assumption A1: τ^2_{post} measures uncertainty in the *recent* slope, which OLS can pin down precisely; but if the rate shifts mid-window, the slope governing the rest of the session has variance closer to the cross-session $\bar{\tau}^2$ than to today's τ^2_{post}.

v1.1 therefore keeps $\hat{b}$ and stores it as $\bar{\tau}^2$ on the calibration. §6 uses $\max(\tau^2_{\text{post}}, \bar{\tau}^2)$ in σ_F^2. The MC in §8 likewise samples r_k with standard deviation $\sqrt{\max(\tau^2_{\text{post}}, \bar{\tau}^2)}$.

The $b\,\delta_f^2$ term has to be present in the regression in any case: without it, the $\delta_f^2\,\bar{\tau}^2$ piece of e_f^2 has nowhere to go but $\hat{a}$, biasing it upward.

Floor and refit. The floor on $\hat{a}$ guards against negative estimates from small samples (OLS does not enforce sign). Refit daily.

8 Threshold ETA

For threshold C_{thr} with $u_{\text{now}} < C_{\text{thr}} \leq 1$, the first-passage time is $T^* = \inf\{t > t_{\text{now}} : u(t) \geq C_{\text{thr}}\}$, with $T^* = \infty$ if no crossing in $[t_{\text{now}}, T_{\text{reset}}]$.

Deterministic ETA (point estimate). On the mean trajectory ($r = \hat{r}_{\text{post}}$, each increment at its expectation):

$$\tilde{T}^* = t_{\text{now}} + \frac{C_{\text{thr}} - u_{\text{now}}}{\hat{r}_{\text{post}}},$$

defined when $\hat{r}_{\text{post}} > 0$ and $\tilde{T}^* \leq T_{\text{reset}}$; else `nil`. Note that $\tilde{T}^*$ is **not** the median of T^*: for the monotone model two effects both push the median *later*, writing $a = C_{\text{thr}} - u_{\text{now}}$.

- *Path skew.* Even at a fixed rate $r = \mu > 0$, $u(t_{\text{now}} + t) - u_{\text{now}}$ is Gamma-distributed and right-skewed, so its median lies below its mean μt. By the monotone-process duality $P(T^* \leq t_{\text{now}} + t) = P(u(t_{\text{now}} + t) - u_{\text{now}} \geq a)$, the median crossing is the t at which median reaches a, which (median < mean) needs $\mu t > a$; hence $\text{median}(T^* \mid r{=}\mu) > t_{\text{now}} + a/\mu = \tilde{T}^*$.
- *Rate uncertainty.* Mixing over the Gamma rate r (mean $\hat{r}_{\text{post}}$) inflates first-passage times: the map $r \mapsto a/r$ is convex on $r > 0$ and steepens sharply as $r \to 0^+$, so the right tail of T^* stretches far more than the left compresses, and near-zero rate draws push crossings past T_{reset} ($T^* = \infty$ within the window).

Both effects push $\text{median}(T^*)$ above $\tilde{T}^*$, so for the monotone model $\tilde{T}^*$ is an early deterministic anchor, not an estimator of the median. The MC percentiles below are the authoritative summary.

CI via Monte Carlo. The first-passage time of the Gamma process with random rate has no tractable closed form. Simulate $K = 500$ trajectories (an MC run with the same seed supplies the terminal values for the §6 CI):

```
tau_eff^2 = max(tau_post^2, bar_tau^2)              # rate-variance floor
for k in 1..K:
    r_k    = Gamma(mean=r_hat_post, var=tau_eff^2)  # >= 0
    u_k[0] = u_now
    for j in 1..M, dt = step size of jth bucket:
        u_k[j] = u_k[j-1] + Gamma(mean=r_k*dt, var=sigma_session^2*dt)
    T*_k = linear interpolation of first j with u_k[j] >= C_thr
         = infinity if none
    U_k  = u_k[M]                                   # terminal -> forecast CI
```

The rate and every increment are non-negative, so each path is monotone: once u_k crosses C_{thr} it stays above, and the first crossing is final (no dip and re-cross). A near-zero rate draw simply never reaches C_{thr} within the window, yielding $T_k^* = \infty$ - the correct outcome under the model, and the non-negative analogue of v1.x's downward-drift paths. The Gamma sampler (Marsaglia-Tsang [2]) reuses the same seeded RNG, so trajectories stay reproducible.

Reported summaries. Let p_∞ be the fraction of trajectories with $T_k^* = \infty$. Define $\tilde{T^*}_{\text{MC}} = \text{median}(T_k^*)$ over the **full** sample (treating ∞ as larger than any finite value):

- If $p_\infty \geq 0.5$: the true median is ∞. Report ETA as `nil`.
- If $p_\infty < 0.5$ but $p_\infty \geq 0.1$: the median is finite but the 90th percentile is ∞. Report median, lower bound = 10th percentile of finite T_k^*, upper bound = `nil` (open-ended).
- If $p_\infty < 0.1$: report median and full 80% CI from the 10th and 90th percentiles of finite T_k^*.

RNG seed derived from (t_{now}, T_{reset}, u_{now}, $\hat{r}_{\text{post}}$, τ^2_{post}, $\sigma^2_{\text{session}}$, $\bar{\tau}^2$, C_{thr}) for test reproducibility.

9 Edge cases

Case	Handling
$n < 3$ snapshots in $\mathcal{R}$	Use prior alone: $\hat{r}_{\text{post}} = \mu_0$, $\tau^2_{\text{post}} = \tau_0^2$
Prior empty ($\lvert\mathcal{S}\rvert < 2$)	Suppress forecast; UI shows "collecting data"
Reset detected in $\mathcal{R}$ ($u_{i+1} < u_i$)	Drop pre-reset snapshots; refit
$C_{\text{thr}} \leq u_{\text{now}}$	Threshold already crossed; return $T^* = t_{\text{now}}$
$\hat{r}_{\text{post}} \leq 0$	Gamma rate degenerates to 0 (no growth): MC paths stay flat, so $p_\infty \to 1$ and the ETA is `nil`
$F > 1$	Report as-is: overshoot is projected demand beyond the limit and its magnitude is signal; only the gauge ring saturates at 1

10 Worked example

5-hour session window: $t_{\text{start}} = 11{:}00$, $T_{\text{reset}} = 16{:}00$, $t_{\text{now}} = 13{:}00$, $u_{\text{now}} = 0.30$, so $\Delta t_{\text{rem}} = 3\,\text{h}$.

Last four snapshots ($\tau_{\text{recent}} = 30\,\text{min}$):

i	t_i (h from 12:30)	u_i
1	0.000	0.270
2	0.167	0.280
3	0.333	0.295
4	0.500	0.300

OLS: $\hat{r}_{\text{OLS}} \approx 0.0630\,\text{h}^{-1}$, $\text{SE}^2_{\text{OLS}} \approx 6.30 \times 10^{-5}$.

Prior (suppose, from history, after the §5 noise correction): $\mu_0 = 0.080\,\text{h}^{-1}$, $\tau_0^2 = 3.6 \times 10^{-3}$.

Posterior: $\tau^2_{\text{post}} \approx 6.19 \times 10^{-5}$, $\hat{r}_{\text{post}} \approx 0.0633\,\text{h}^{-1}$ (data dominates).

Calibration: $\sigma^2_{\text{session}} = 2.5 \times 10^{-3}\,\text{h}^{-1}$, $\bar{\tau}^2 = 3.6 \times 10^{-3}\,\text{h}^{-2}$ (approximately matches τ_0^2 in a stable regime, as expected).

Forecast:

$$F = 0.30 + 0.0633 \cdot 3 \approx 0.490,$$

$$\max(\tau^2_{\text{post}}, \bar{\tau}^2) = \max(6.19 \times 10^{-5}, 3.6 \times 10^{-3}) = 3.6 \times 10^{-3} \quad \text{(floor active)},$$

$$\sigma_F^2 \approx 9 \cdot 3.6 \times 10^{-3} + 3 \cdot 2.5 \times 10^{-3} \approx 3.24 \times 10^{-2} + 7.5 \times 10^{-3} \approx 4.0 \times 10^{-2},$$

$$\sigma_F \approx 0.200.$$

80% CI: the Monte Carlo terminal quantiles for this scenario are [0.30, 0.78] - asymmetric, with the lower edge resting at $u_{\text{now}} = 0.30$ (with σ_F this large the Gamma rate has shape $\hat{r}^2_{\text{post}} / \max(\tau^2_{\text{post}}, \bar{\tau}^2) \approx 1.1$, so a sizeable share of draws are near-zero and those paths barely move) and a longer reach upward. Display: "Projected 49% by reset (80% CI: 30%–78%)". For comparison the retired v1.x symmetric z-quantile was $0.490 \pm 1.282 \cdot 0.200 = [0.23, 0.75]$, clipped up to $u_{\text{now}} = 0.30$ on the left; v2.0 produces the floor and the right-skew directly from the monotone process.

Rate uncertainty dominates path noise by 4.3×. The recent slope is locally well-estimated, but historically the rate has varied a lot across sessions, and the v1.1 floor honors that.

ETA to 100%: gap is 0.70, deterministic crossing at $13{:}00 + 0.70/0.0633 \approx 24{:}04$, far beyond reset. Returned as `nil`. The MC confirms by producing infinite crossings for almost all trajectories.

11 Limitations and upgrade paths

- **Heavy-tailed / bursty usage.** v2.0's Gamma process fixes the monotonicity and the lower-tail floor and adds a right skew, but its tail is still relatively light and its index of dispersion is tied to one scalar. If the MC CI still undercovers in the right tail (check the `UnderspreadX` and per-horizon coverage diagnostics on real replays), the upgrades are a heavier subordinator (inverse Gaussian) or a compound-Poisson increment that models idle-then-burst usage with an explicit mass at "no further growth".
- **Constant-rate assumption.** A mode shift within the window (deep work → meetings) breaks (A1). The recency window τ_{recent} absorbs slow drift but not abrupt shifts. The principled upgrade is a Kalman filter where r is a slowly-varying latent process.
- **Global $\sigma^2_{\text{session}}$.** Pivot variance may depend on day-of-week or project type. Stratified calibration is the upgrade, but requires more sessions per stratum.

References

[1] G. Casella. An introduction to empirical Bayes data analysis. *The American Statistician*, 39(2):83–87, 1985.

[2] G. Marsaglia and W. W. Tsang. A simple method for generating gamma variables. *ACM Transactions on Mathematical Software*, 26(3):363–372, 2000.