A QUEUEING-THEORETIC FRAMEWORK FOR PREFILL-DECODE DISAGGREGATED LLM INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid growth of large language models (LLMs) has made efficient large-scale inference a critical systems challenge. However, existing LLM serving research has primarily focused on engineering optimization, while queueing-theoretic guidance for resource provisioning under prefill–decode (P–D) disaggregation remains limited. In this paper, we develop a tandem-queue framework for P–D disaggregated inference. In particular, the interplay of autoregressive generation, evolving KV-cache requirements, continuous batching, and associated constraints in decode gives rise to a distinctive dynamic multi-server queue in LLM inference. We formulate an exact, high-dimensional Markov representation of the proposed queueing model and further develop tractable analytical methods for its stationary population distributions. For the prefill queue, we employ structured Markov analysis and transform-domain methods; for the decode queue, we use saddlepoint approximation, large-deviation theory, and Halfin–Whitt quality-and-efficiency-driven (QED) analysis to classify and characterize regimes dominated by different resource bottlenecks. We find that both the onset of queueing and the transition to resource redundancy exhibit sharp changes, and that when multiple resources simultaneously become bottlenecks, necessary resource slack and the resulting underutilization are critical to system stability. These findings provide analytical tools for system deployment, enabling operators and schedulers to coordinate resource allocation across prefill and decode and to provision concurrency and KV-cache capacity jointly. Such guidance helps avoid underutilizing bottleneck resources or overprovisioning nonbinding resources. Extensive simulations validate the theoretical predictions.

1 INTRODUCTION

Large language models (LLMs) have advanced rapidly, enabling diverse applications (Brown et al., 2020; Chen et al., 2021; Ouyang et al., 2022; OpenAI, 2023) and making efficient large-scale inference a first-order systems concern (Pope et al., 2023). Unlike one-shot prediction workloads, an LLM request first processes its prompt and then generates output tokens autoregressively, corresponding to the prefill and decode phases. Prefill–decode (P–D) disaggregation has emerged as an effective architecture for optimizing LLM inference (Zhong et al., 2024). By deploying the two phases on separate instances, the system enables stage-specific optimization. Prefill is primarily compute intensive and processes the entire prompt in parallel, whereas decode relies on KV caching and batching. KV caching accelerates generation by storing and reusing previous key-value representations, at the cost of substantial memory consumption, while batching enables concurrent processing of multiple requests.

From a systems perspective, P–D disaggregation forms a heterogeneous tandem queue: requests stochastically arrive, wait for prefill service, and then enter the decode queue. Since the two stages have distinct service mechanisms and resource constraints, resource provisioning for P–D inference is fundamentally a queueing problem. Recent studies have started to establish queueing foundations for LLM inference. Existing works mainly characterize stability and throughput, or formulate high-dimensional Markov processes that are difficult to solve directly (Nie et al., 2026; Dai et al., 2025). However, stability and throughput characterizations do not by themselves determine stationary population distributions or congestion tails in a P–D tandem system with coupled decode resource constraints.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Characterizing P–D disaggregated inference through queueing theory introduces three challenges. First, LLM inference is constrained not only by computation but also by concurrency limits and memory capacity. Second, the distinctive KV-cache and batching mechanisms in decode give rise to a complex dynamic multi-server queue that differs from classical queueing models (see Section 3). Third, the KV-cache and concurrency constraints in decode are coupled through request states: prompt length and decoding progress determine the current KV footprint, while advancing a request changes its future memory demand. Therefore, the decode population alone cannot determine instantaneous service capacity, while an exact Markov representation is prohibitively large.

In this paper, we develop a queueing-theoretic framework for stationary and tail analysis of P–D disaggregated LLM inference. We model the system as a tandem queue. For prefill, structured Markov representation and transform-domain analysis yield the stationary population distribution. For decode, we introduce probabilistic resource decoupling based on whether one resource constraint is unlikely to affect service decisions, enabling tractable analysis in different resource-dominant regimes. Our main contributions are as follows.

- **Modeling.** We develop a queueing framework for P–D disaggregated LLM inference and model the two stages as a discrete-time tandem queue. The framework captures the compute constraint of prefill and the hardware concurrency and KV-cache constraints of decode. The model combines the multi-stage structure and multiple resource constraints of P–D inference for stationary population and tail analysis.
- **Analysis.** We characterize the stationary distribution of request populations. Within the prefill queue model, structured Markov and transform-domain analysis yield an exact stationary population distribution. For decode, we develop tractable analytical approximations through probabilistic resource decoupling. The memory-dominated regime is characterized through an associated infinite-server reference system, saddlepoint approximation, and large-deviation analysis, while the concurrency-dominated regime is analyzed through Halfin–Whitt/QED theory.
- **System Insights.** Our distributional analysis translates into concrete guidance for resource provisioning and scheduling. The stationary queue distributions provide queue-pressure signals for allocating resources between prefill and decode, while the redundancy thresholds $B_\theta(M)$ and $M_\theta(c)$ quantify how concurrency capacity and KV-cache capacity should be provisioned relative to each other. We further establish a necessary joint-slack result, showing that satisfying the stability conditions of the two resources independently does not, in general, guarantee stability of the jointly constrained decode system.
- **Validation.** We conduct extensive simulations to validate the theoretical results. Across a wide range of arrival processes, workload distributions, and resource configurations, the theoretical predictions closely match the corresponding simulation results.

The remainder of the paper is organized as follows. Section 3 introduces the system model. Sections 4 and 5 present the queueing analysis of the prefill and decode stages, respectively. Section 6 discusses the resulting implications for system configuration, and Section 7 provides the numerical evaluation.

2 Related Work

Queueing and LLM Inference. Recent work has begun to study LLM inference from a queueing perspective. Yang et al. (2024) use M/G/1 and bulk-service models to analyze how stochastic token lengths, token clipping, and batching policies affect queueing delay. Nie et al. (2026) explicitly model KV-cache memory and derive stability and instability conditions for LLM inference, while Dai et al. (2025) develop a fluid-limit framework to characterize throughput-optimal scheduling under batched service. Queueing models have also been used to optimize reasoning-token allocation (Ozbas & Bastopcu, 2026), and Mitzenmacher & Shahout (2025) survey the broader queueing and scheduling challenges arising in LLM systems. Collectively, these works establish important results on mean delay, throughput, stability, and provisioning, but stationary population distributions, rare-congestion tails, and multi-stage inference under coupled resource constraints remain much less understood.

Theoretical Modeling for LLM Inference. A parallel line of work studies LLM inference through scheduling and optimization. Memory-aware online scheduling is studied by Jaillet et al. (2025); Ao et al. (2025); other work addresses variable prefill/decode lengths (Wang et al., 2025), prediction uncertainty (Chen et al., 2025), prefix reuse (Dexter et al., 2025), and service-level-objective (SLO)-aware serving (Bari et al., 2025). Their emphasis is policy design and performance optimization rather than stationary population characterization.

Multi-Server Queueing Systems. Classical multi-server queueing theory provides several analytical foundations relevant to systems with parallel service capacity. Even for a general GI/G/c queue with a fixed number of servers, exact stationary analysis is difficult, motivating a variety of approximations. In particular, Shore (1988) develops a two-state approximation combining a loaded single-server regime with an infinite-server regime, while Whitt (1993) develops a broader family of approximations for GI/G/m queues. In the many-server heavy-traffic regime, Halfin & Whitt (1981) establish the Halfin–Whitt limit, a classical foundation for QED analysis. Several extensions relax the assumption of a fixed server count. Variable- and queue-dependent-server models activate or release servers according to queue occupancy or idleness (Li & Yang, 2000; Wang & Tai, 2000), while time-varying many-server queues allow staffing capacity to change exogenously over time (Feldman et al., 2008). Related work also studies multiclass queues with heterogeneous per-customer resource requirements (Zychlinski et al., 2022). Stochastic bandwidth-packing models characterize stability under aggregate packing constraints (Gamarnik, 2004).

Unlike classical multi-server models, LLM inference capacity is constrained not only by a concurrency cap but also by KV-cache memory, whose per-request footprint grows as decoding progresses. This dynamic memory usage makes the number of concurrently served requests difficult to characterize. Thus, although classical multi-server theory provides analytical foundations, developing tractable methods for this dynamic, multi-resource queue remains an open and distinct challenge.

3 System Model

To model the LLM inference process, we formulate the LLM inference system consisting of a pair of prefill and decode instances as a tandem queueing system (Appendix A). We consider deployment on modern accelerator clusters equipped with high-bandwidth interconnects, such as InfiniBand or NVLink, as in contemporary disaggregated serving systems (Zhong et al., 2024). We therefore assume that inter-stage communication is not the primary bottleneck and focus on the constraints imposed by computational resources and KV-cache memory capacity.

Arrival and workload process. We consider a discrete-time representation of the LLM inference process. Let A_t denote the number of external requests arriving in time slot t. The arrival process $\{A_t\}_{t\geq 0}$ is assumed i.i.d. and nonnegative integer-valued, with probability mass function $q_a = \Pr\{A_t = a\}$, $a \in \mathbb{Z}_{\geq 0}$.

Each request i is associated with a workload mark (χ_i, ω_i), where χ_i and ω_i denote the prompt length and output length, respectively. The request marks are i.i.d. across requests and independent of the arrival process. We assume that χ_i and ω_i are positive integer-valued random variables with bounded support. We denote their maximum values by $\chi_{\max} \triangleq \max \mathrm{supp}(\chi), \omega_{\max} \triangleq \max \mathrm{supp}(\omega)$. A generic request mark is denoted by (χ, ω).

prefill phase. The prefill phase is typically compute-intensive. We model the prefill instance as a request-level single-server queue and approximate the service time as a linear function of prompt length, consistent with the compute-saturated prefill behavior observed in prior serving studies (Zhong et al., 2024; Yang et al., 2024). The service time S is discretized according to the slot duration δ as $S = \left\lceil \frac{a_{\mathrm{P}}\chi + c_{\mathrm{P}}}{\delta} \right\rceil$, where $a_{\mathrm{P}} > 0$ and $c_{\mathrm{P}} \geq 0$ are model-dependent prefill service coefficients. Let $K_{\mathrm{P}}(t)$ denote the prefill queueing process, which comprises all requests that have not yet departed from the instance.

decode inference process. The decode phase generates tokens autoregressively and must therefore sustain a large number of concurrent requests. We assume a work-conserving policy that prioritizes ongoing requests and admits waiting ones in first-come-first-served (FCFS) order whenever resources permit. This abstraction draws on FCFS scheduling in vLLM and stall-free continuation of active decode requests in Sarathi-Serve (Kwon et al., 2023; Agrawal et al., 2024). Let $K_{\mathrm{D}}(j)$ denote the total decode population immediately before the forward pass of iteration j.

decode continuous batching. Continuous batching operates at the granularity of decode iterations. At each iteration, the scheduler forms an active batch from ongoing requests and waiting requests. After a forward pass, completed requests leave the active batch and release their occupied resources, after which newly admitted requests are added for the next iteration.

We approximate the duration of each decode iteration by a fixed value Δ_{D} under a fixed serving configuration. Because a work-conserving policy keeps the system close to its effective resource limit, continuous batching tends to incur similar computational workloads across iterations. Similar iteration-duration abstractions have been adopted in prior queueing analyses (Nie et al., 2026; Dai et al., 2025). Moreover, P–D disaggregation avoids the execution heterogeneity caused by mixing prefill and decode processing, which further favors the stability of the iteration duration (Appendix H.1).

decode resource constraints. Let $\mathcal{B}_j$ denote the active decode batch at iteration j, and let $g_i(j)$ be the number of output tokens generated by request i before iteration j. We normalize the KV-cache capacity from bytes to tokens and denote the effective admission budget by $M > 0$; the concurrency limit is $B \in \mathbb{Z}_+$. A batch is feasible if

$$R_j \triangleq \sum_{i \in \mathcal{B}_j} \big(\chi_i + g_i(j)\big) \le M, \qquad |\mathcal{B}_j| \le B.$$

Here M is an effective KV-cache admission budget below the physical KV-cache capacity, leaving headroom for cache growth. Practical serving systems expose memory-allocation and concurrency controls, including vLLM's `-gpu-memory-utilization` and `-max-num-seqs`, and SGLang's `-max-running-requests` (vLLM, 2025; SGLang, 2026).

decode as a Dynamic Multi-Server Queueing System. Interpreting the active batch size as the number of simultaneously available servers, continuous batching naturally induces a dynamic multi-server queue. Each request occupies one concurrency position, while its KV-cache footprint depends on both prompt length and decode age. Therefore, systems with the same population size can induce different feasible batch sizes. This resource-coupled, state-dependent service capacity distinguishes decode from existing multi-server queueing models and motivates the two-resource analysis developed in the following sections.

4 Queueing Analysis of the prefill Stage

In this section, we characterize the stationary distribution of the prefill population process K_{P}. The prefill instance is modeled as a discrete-time $G/G/1$ queue. Its structured Markov representation renders the stationary distribution of K_{P} amenable to transform-domain analysis.

Structured Markov process. Let $g_s = \Pr\{S = s\}$ denote the probability mass function (PMF) of the prefill service requirement, and define the probability generating functions (PGFs) $G_A(z) = \sum_{a \ge 0} q_a z^a$ and $G_S(z) = \sum_{s \ge 1} g_s z^s$.

The offered load of the prefill instance is $\rho_{\mathrm{P}} = \mathbb{E}[A_t]\,\mathbb{E}[S]$, and throughout this section we assume the stability condition $\rho_{\mathrm{P}} < 1$.

To obtain a Markov representation, we augment $K_{\mathrm{P}}(t)$ with the residual service time $R_{\mathrm{P}}(t)$ of the request in service, where $R_{\mathrm{P}}(t) = 0$ iff $K_{\mathrm{P}}(t) = 0$. The resulting process $X_{\mathrm{P}}(t) = (K_{\mathrm{P}}(t), R_{\mathrm{P}}(t))$ forms a discrete-time Markov chain on the enlarged state space $\{(0,0)\} \cup \{(k,r) : k, r \ge 1\}$. The transition dynamics are

1. From an empty system, newly arriving requests initiate a service period:
$$\Pr\{(0,0) \to (a,s)\} = q_a g_s, \qquad a \ge 1,\ s \ge 1.$$
2. Before the request in service completes, the residual service decreases deterministically by one slot while new requests accumulate:
$$\Pr\{(k,r) \to (k+a, r-1)\} = q_a, \qquad k \ge 1,\ r \ge 2,\ a \ge 0.$$
3. Upon a service completion, one request departs and, whenever the system remains nonempty, the next request starts service:
$$\Pr\{(k,1) \to (k-1+a, s)\} = q_a g_s, \qquad k-1+a \ge 1,\ s \ge 1.$$

In addition, $\Pr\{(0,0) \to (0,0)\} = q_0$ and $\Pr\{(1,1) \to (0,0)\} = q_0$. The key structural property is the deterministic progression of the residual-service coordinate, which allows this supplementary dimension to be eliminated analytically.

Transform-domain analysis. The composition $G_S(G_A(z))$ is the PGF of arrivals accumulated during a service period. Eliminating the residual-service coordinate yields the following stationary law.

Theorem 1. *Under $\rho_{\mathrm{P}} < 1$, the stationary distribution of the prefill population exists and is given by*

$$\pi_k^{\mathrm{P}} \triangleq \Pr\{K_{\mathrm{P}} = k\} = [z^k]\left[\frac{(1-\rho_{\mathrm{P}})(1-z)G_S(G_A(z))}{G_S(G_A(z)) - z}\right], \qquad k \geq 0, \tag{1}$$

where $[z^k]f(z)$ denotes the coefficient of z^k in the power-series expansion of $f(z)$.

Proof. Please see Appendix B.1 for a detailed proof.

The resulting stationary distribution immediately yields the mean prefill population and tail probabilities $\Pr\{K_{\mathrm{P}} \geq k\}$, which quantify prefill congestion. Prefill departures form the decode input. Let A^{D} denote the number of prefill completions during one decode iteration, with stationary PMF $q_a^{\mathrm{D}} \triangleq \Pr\{A^{\mathrm{D}} = a\}$. This PMF is obtained by marking service-completion transitions of the prefill chain, as detailed in Appendix B.2.

5 Queueing Analysis of the Decode Stage

In this section, we characterize the stationary population of the two-resource decode queue. For tractability, we approximate the inter-stage input by i.i.d. per-iteration arrival counts with the prefill-derived PMF q_a^{D} and independent request marks (Appendix H.2). We formulate the resulting Markov chain and apply probabilistic resource decoupling to obtain analytical approximations in the memory-dominated and concurrency-dominated regimes.

5.1 Markov Representation and Probabilistic Decoupling

Markov representation. Define

$$X_{\mathrm{D}}(j) = \big(Q_{\mathrm{D}}(j), \eta_j, Z_1(j), \ldots, Z_B(j)\big),$$

where $Q_{\mathrm{D}}(j)$ is the waiting population and $Z_m(j)$ is either idle (∂) or an active request's prompt length and decode age (p_m, g_m). The coordinate η_j summarizes previous failed admission tests of the current head through $\chi > \eta_j$; it is zero for an unexamined head or an empty queue.

For $x = (q, \eta, z_1, \ldots, z_B)$, write $\mathcal{I}(x) = \{m : z_m \neq \partial\}$. First, each active request independently completes with probability $h_p(g) = \Pr\{\omega = g+1 \mid \omega > g, \chi = p\}$, or advances to $(p, g+1)$. Second, a arrivals join the waiting queue. Finally, FCFS admission proceeds using the survivors' updated KV footprints. With remaining memory budget ℓ and $\bar{F}_\chi(u) = \Pr\{\chi > u\}$, the head is admitted with prompt length p with probability $\frac{\Pr\{\chi=p\}}{\bar{F}_\chi(\eta)}$ for $\eta < p \leq \ell$. Rejection updates η to $\max\{\eta, \ell\}$ and stops admission; success adds $(p, 0)$ and resets η for the next head. Admission also stops when the queue empties or all B positions are occupied.

Let $\mathcal{A}_{B,M}(x' \mid x, a, \mathcal{L})$ denote the resulting admission kernel given a arrivals and completion set $\mathcal{L} \subseteq \mathcal{I}(x)$. Then

$$w_x(\mathcal{L}) = \prod_{m \in \mathcal{L}} h_{p_m}(g_m) \prod_{m \in \mathcal{I}(x)\setminus\mathcal{L}} \big(1 - h_{p_m}(g_m)\big),$$

$$\mathsf{P}_{\mathrm{D}}(x, x') = \sum_{a \geq 0} q_a^{\mathrm{D}} \sum_{\mathcal{L} \subseteq \mathcal{I}(x)} w_x(\mathcal{L})\, \mathcal{A}_{B,M}(x' \mid x, a, \mathcal{L}).$$

Its active-state space grows combinatorially with B and the request phases, making direct stationary computation impractical.

Probabilistic decoupling. The key idea is to approximate the jointly constrained queue by a single-resource system when the omitted constraint rarely changes batch selection. Removing the concurrency constraint yields the *memory-only* reference system, while removing the KV-cache constraint

yields the *concurrency-only* reference system. For $\theta \in (0,1)$, the omitted constraint is θ-*redundant* if reinstating it changes the selected batch with probability at most θ under the corresponding stationary reference law. Accordingly, $B \geq B_\theta(M)$ defines the *memory-dominated* regime, whereas $M \geq M_\theta(c)$ defines the *concurrency-dominated* regime for reference concurrency c. The coupled regime, certified by neither condition, is discussed in Section 6.

5.2 Memory-Dominated decode

Concurrency-redundancy threshold. Let $\lambda_{\mathrm{D}} \triangleq \mathbb{E}\left[A^{\mathrm{D}}\right]$. For the memory-only reference system, let $N^{(M)}$ and $Q_{\mathrm{D}}^{(M)}$ denote the stationary active and waiting populations; the total population is $K_{\mathrm{D}}^{(M)} = N^{(M)} + Q_{\mathrm{D}}^{(M)}$. Assume $\chi \perp \omega$ and define the lifetime memory work

$$W \triangleq \sum_{g=0}^{\omega-1}(\chi + g) = \chi\omega + \frac{\omega(\omega-1)}{2}.$$

Following the two-regime perspective of Shore (1988), we approximate the active population separately according to whether $Q_{\mathrm{D}}^{(M)} = 0$. Both branches are constructed from an infinite-server reference with active count N_∞ and KV occupancy R_∞. Writing $G_{\mathrm{D}}(z) = \mathbb{E}\left[z^{A^{\mathrm{D}}}\right]$, $G_\chi(y) = \mathbb{E}\left[y^\chi\right]$, and $s_g = \Pr\{\omega > g\}$, its joint PGF is

$$\mathcal{G}_\infty(z,y) = \prod_{g\geq 0} G_{\mathrm{D}}(1 - s_g + s_g z y^g G_\chi(y)), \qquad p_{b,r}^\infty \triangleq [z^b y^r]\mathcal{G}_\infty(z,y).$$

Here $[z^b y^r]$ extracts the coefficient of $z^b y^r$ in the bivariate PGF. For the no-backlog branch, retain memory-feasible states through $a_b(M) \triangleq \sum_{0\leq r\leq M} p_{b,r}^\infty$. For the backlog branch, additionally weight each state by the probability that its remaining memory blocks a reference request: $u_b(M) \triangleq \sum_{0\leq r\leq M} p_{b,r}^\infty \bar{F}_\chi(M-r)$, using the prompt tail $\bar{F}_\chi$ defined above. With $Z_0(M) = \sum_b a_b(M)$ and $Z_+(M) = \sum_b u_b(M)$, the normalized weights approximate the two conditional active-count laws. Let $\widehat{R}_0$ and $\widehat{C}_+$ be their respective mean KV occupancies. Matching their mixture to the conserved mean occupancy $\lambda_{\mathrm{D}}\mathbb{E}[W]$ gives the backlog probability estimate $\widehat{p}_+ = \frac{\lambda_{\mathrm{D}}\mathbb{E}[W] - \widehat{R}_0}{\widehat{C}_+ - \widehat{R}_0}$.

Theorem 2. *For a stable memory-only reference system and any $\theta \in (0,1)$, the concurrency-redundancy threshold $B_\theta(M)$ exists. Its two-regime estimate is*

$$\widehat{B}_\theta(M) = \min\left\{ b \in \mathbb{Z}_{\geq 0} : \frac{(1-\widehat{p}_+)\sum_{j>b} a_j(M)}{Z_0(M)} + \frac{\widehat{p}_+ \sum_{j>b} u_j(M)}{Z_+(M)} \leq \theta \right\}. \tag{2}$$

Proof. Please see Appendix C for a detailed proof.

Stationary population distribution. We approximate the waiting population by a discrete-time single-server workload model with service requirement $S^+ = \frac{W}{\widehat{C}_+}$ and stationary PMF π_q^R, $q \geq 0$. Its positive-backlog law is $\frac{\pi_q^R}{1-\pi_0^R}$ for $q \geq 1$. Conditional active/waiting factorization yields a mixture of the no-backlog law and the backlog convolution.

Theorem 3. *For a stable memory-only reference system, the two-regime construction gives the stationary population approximation*

$$\widehat{\pi}_k^{(M)} = \frac{(1-\widehat{p}_+)a_k(M)}{Z_0(M)} + \sum_{b=0}^{k-1} \frac{\widehat{p}_+ u_b(M)\pi_{k-b}^R}{Z_+(M)(1-\pi_0^R)}, \qquad k \geq 0. \tag{3}$$

Proof. Please see Appendix D for a detailed proof.

Beyond the preceding decomposition approximation, we further employ a large-deviation principle to yield the following far-tail characterization.

Corollary 1. *Consider a stable memory-only reference system, with the relevant exponential moments finite. Under the saturated memory-work approximation,*

$$\lim_{k\to\infty} \frac{1}{k} \log \Pr\left\{ K_{\mathrm{D}}^{(M)} \geq k \right\} = -\gamma_M,$$

where $\gamma_M > 0$ *solves* $\gamma_M = \log \mathbb{E}[e^{\frac{\Lambda_{\mathrm{D}}(\gamma_M)W}{M}}]$ *and* $\Lambda_{\mathrm{D}}(s) = \log \mathbb{E}[e^{sA^{\mathrm{D}}}]$.

Proof. Please see Appendix E for a detailed proof.

5.3 Concurrency-Dominated decode

Memory-redundancy threshold. For a stable concurrency-only reference system with c service positions, let $R^{(c)}$ denote its stationary KV occupancy. The threshold $M_\theta(c) \triangleq \min\{m \in \mathbb{Z}_{\geq 0} : \Pr\{R^{(c)} > m\} \leq \theta\}$ is finite since $R^{(c)} \leq c(\chi_{\max} + \omega_{\max} - 1)$, and can be estimated via Monte Carlo simulation.

QED stationary population analysis. When memory is redundant, we use a residual-service QED approximation. Conditional on the preceding scaled state, the next population is approximately Gaussian; averaging its integer-cell probabilities yields the stationary approximation.

Theorem 4. *Consider a sequence of stable concurrency-only systems with* $c \to \infty$ *and arrival rates* $\lambda_{\mathrm{D}} = \lambda_{\mathrm{D}}(c)$ *satisfying* $\beta_c \triangleq \frac{c - \lambda_{\mathrm{D}}\mathbb{E}[\omega]}{\sqrt{c}} \to \beta \in (0, \infty)$. *The finite-*$c$ *approximation for* $\tau_c > 0$ *is*

$$\widehat{\pi}_k^{(c)} = \int \left[\Phi\left(\frac{k + \frac{1}{2} - m_c(x)}{\tau_c} \right) - \Phi\left(\frac{k - \frac{1}{2} - m_c(x)}{\tau_c} \right) \right] \pi_\beta(dx), \qquad k \in \mathbb{Z}_{\geq 0}, \tag{4}$$

where π_β *is the invariant law of the residual-service QED chain with state* $x = (q, \ell_1, \ldots, \ell_{\omega_{\max}})$. *Here* q *is the* $\sqrt{c}$*-scaled waiting population, while* ℓ_j *is the residual-service count with* j *remaining iterations, centered at* $\lambda_{\mathrm{D}} \Pr\{\omega \geq j\}$ *and scaled by* $\sqrt{c}$. *The conditional parameters are*

$$m_c(x) \triangleq \lambda_{\mathrm{D}} \mathbb{E}[\omega] + \sqrt{c} \left(q + \sum_{j=2}^{\omega_{\max}} \ell_j \right), \qquad \tau_c^2 \triangleq c\sigma_{\mathrm{D}}^2, \tag{5}$$

with $\sigma_{\mathrm{D}}^2 \triangleq \frac{\mathrm{Var}(A^{\mathrm{D}})}{\lambda_{\mathrm{D}}\mathbb{E}[\omega]}$, *and* Φ *the standard normal CDF. The case* $\tau_c = 0$ *is given in* (F.49).

Proof. Please see Appendix F for a detailed proof.

6 Implications for System Configuration

prefill–decode resource allocation. Our stationary population laws and large-deviation tails quantify the probability of different congestion levels, which cannot be captured by stability or mean queue length alone. In a P–D disaggregated system, these distributional characteristics reveal where queueing pressure accumulates, thereby providing a theoretical direction for reallocating capacity between prefill and decode.

Load-sensitive capacity provisioning. We observe a sharp transition in waiting congestion as the offered load increases under a fixed resource configuration: a system with little or no waiting can rapidly develop a substantial backlog after a relatively small increase in load (Appendix G.1). This behavior is particularly relevant to latency-sensitive services, where sufficient capacity should be retained to prevent operation near such a congestion transition rather than merely satisfying long-run stability.

KV-cache–concurrency coordination. Within decode, both concurrency and KV-cache capacity exhibit sharp transitions from binding constraints to effectively redundant resources (Appendix G.1). The thresholds $B_\theta(M)$ and $M_\theta(c)$ characterize these transition frontiers and therefore guide joint resource dimensioning. When one resource is scarce, provisioning the other near its redundancy frontier allows the scarce resource to determine system capacity while avoiding substantial overprovisioning of the nonbinding resource.

Necessary joint slack. Let μ_M, μ_B, and $\mu_{M,B}$ denote the maximum stabilizable arrival rates, measured in requests per decode iteration, of the memory-only, concurrency-only, and jointly constrained decode systems, respectively, under the same workload distribution and scheduling policy.

Theorem 5. *There exist bounded request workloads and finite resource configurations such that*

$$\mu_{M,B} < \min\{\mu_M, \mu_B\}. \tag{6}$$

Consequently, for any arrival rate λ satisfying $\mu_{M,B} < \lambda < \min\{\mu_M, \mu_B\}$, both single-resource reference systems are stable, whereas the jointly constrained decode system is unstable.

Proof. The capacity gap arises because stochastic concurrency demand and KV-cache occupancy cannot in general be perfectly aligned over time. Hence, provisioning both resources exactly at their isolated capacity limits is insufficient in general; stable joint operation requires additional resource slack to absorb this stochastic mismatch. A numerical illustration is provided in Appendix G.1.

7 Numerical Evaluation

In this section, numerical results are presented to validate the theoretical analysis. Section 7.1 describes the simulation configurations. Section 7.2 validates Theorem 1, Theorem 3, Corollary 1, and Theorem 4. Section 7.3 further validates Theorem 2.

7.1 Experimental Setup

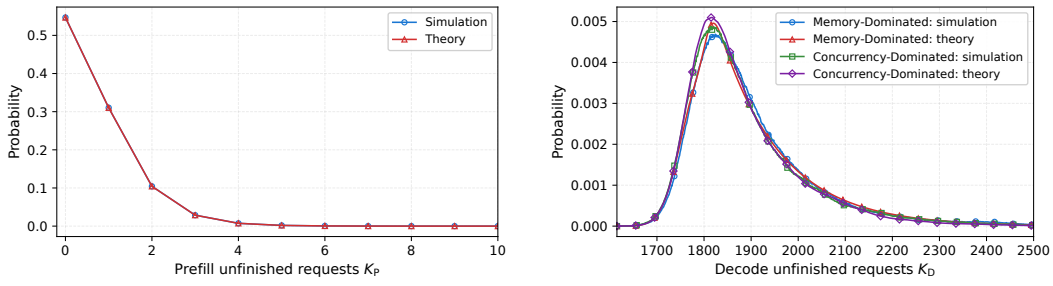
We simulate the system in Section 3 with slot duration $\delta = 0.5\,\mathrm{ms}$, prefill parameters $a_{\mathrm{P}} = 0.006\,\mathrm{ms/token}$ and $c_{\mathrm{P}} = 0.3\,\mathrm{ms}$, and decode iteration duration $\Delta_{\mathrm{D}} = 20\,\mathrm{ms}$. We use $\chi \sim \mathrm{Unif}\{64, \ldots, 192\}$ and $\omega \sim \mathrm{Unif}\{128, \ldots, 384\}$, independently, giving $\mathbb{E}[\chi] : \mathbb{E}[\omega] = 1 : 2$. The memory- and concurrency-dominated configurations use $M = 479{,}250$ and $B = 1{,}800$, respectively, with the secondary constraint made nonbinding. Each end-to-end run contains 4×10^8 slots, with the first 4×10^7 discarded as warm-up.

These configurations are selected to operate the end-to-end system, especially the decode stage, in a high-utilization regime where queueing behavior emerges. In practice, the parameters can be replaced by measurements from specific hardware. We additionally evaluate the framework under other workload and resource configurations; further results are provided in Appendix G.2.

7.2 Queue Distribution Validation

We construct an end-to-end inference system consisting of one prefill instance and one decode instance. In this subsection, external arrivals are Poisson with a rate of 350 requests per second, i.e., $\mathbb{E}[A_t] = 0.175$ requests per base slot and $\lambda_{\mathrm{D}} = 7$ requests per decode iteration. Additional standalone prefill and decode results are provided in Appendix G.2.

Stationary population PMFs. Figure 1 compares the analytical stationary laws with simulation. The prefill solution in Theorem 1 achieves a total-variation (TV) distance of 3.59×10^{-5}. For memory- and concurrency-dominated decode, Theorems 3 and 4 yield TV distances of 2.98×10^{-2} and 2.27×10^{-2}, respectively, with relative mean errors of 0.23% and 0.50%.

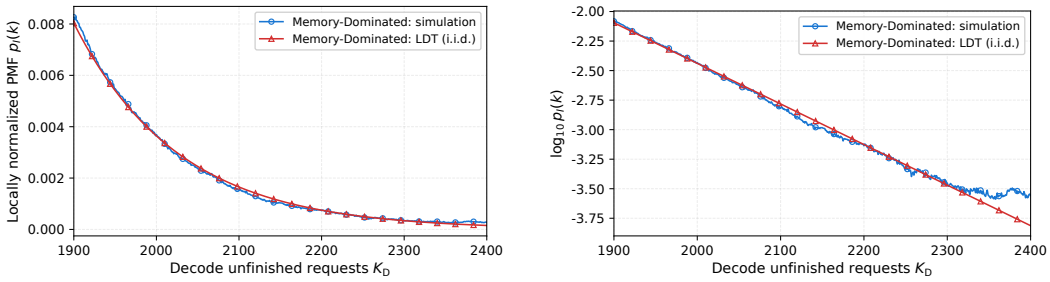


(a) prefill stationary PMF.

(b) decode stationary PMFs.

Figure 1: End-to-end validation of stationary population PMFs.

Large-deviation tail. Figure 2 compares the locally normalized decode PMF over $\{1900, \ldots, 2400\}$ with the exponential approximation from Corollary 1. The theoretical exponent $\gamma_M = 0.00791$ has a relative error of 7.44% with respect to the fitted empirical exponent 0.007362. This discrepancy largely stems from the extremely small probability mass near 2400, which makes the simulation estimates in that regime subject to pronounced statistical fluctuation.



(a) Locally normalized tail PMF.

(b) Locally normalized tail PMF (log scale).

Figure 2: End-to-end validation of the memory-dominated tail exponent.

7.3 Concurrency-Redundancy Threshold Validation

We test Theorem 2 at $M = 479{,}250$ with deterministic ($A^{\mathrm{D}} = 7$), Poisson (mean seven), and discrete uniform ($\{0, \ldots, 14\}$) per-iteration arrivals. The Monte-Carlo (MC) threshold B_θ^{MC} is the smallest integer cap whose empirical stationary active-batch exceedance probability is at most θ. Table 1 compares the predicted and MC thresholds at three tolerances, with absolute relative error $\frac{|\widehat{B}_\theta - B_\theta^{\mathrm{MC}}|}{B_\theta^{\mathrm{MC}}}$. Errors remain below 1.64% and decrease with θ, supporting the two-regime approximation under the tested workloads.

Table 1: Concurrency-redundancy thresholds under three arrival models.

	Deterministic			Poisson			Uniform		
θ	MC	$\widehat{B}_\theta$	Err. (%)	MC	$\widehat{B}_\theta$	Err. (%)	MC	$\widehat{B}_\theta$	Err. (%)
0.01	1837	1807	1.633	1861	1832	1.558	1865	1847	0.965
0.10	1815	1802	0.716	1829	1816	0.711	1833	1825	0.436
0.20	1807	1800	0.387	1817	1809	0.440	1820	1815	0.275

8 Conclusion

This paper presents a discrete-time queueing framework for P–D disaggregated LLM inference. We derive the prefill stationary population law and develop tractable decode population and tail approximations through probabilistic resource decoupling. Resource-redundancy thresholds and a necessary joint-slack result provide guidance for coordinated concurrency and KV-cache provisioning. Numerical simulations support the predictions, establishing a quantitative basis for congestion-aware system configuration.

Our work has limitations that suggest promising future research directions. First, we approximated decode batch-processing time as a constant; a promising extension is to develop accurate yet tractable stability approximations under more sophisticated service-time models. Second, deriving analytical population distributions when both decode resource constraints bind with non-negligible probability remains an important extension. Third, communication is assumed not to be a bottleneck. Under limited communication resources, transfer delays and additional communication queues may become significant, motivating an extension to more complex queueing networks.

AI Use Statement

Large Language Models (ChatGPT) were employed exclusively as a writing aid for language polishing. The model did not participate in research ideation, experimental design, data analysis, or interpretation of results. The authors take full responsibility for the entire content.

REFERENCES

Amey Agrawal, Nitin Kedia, Ashish Panwar, Jayashree Mohan, Nipun Kwatra, Bhargav Gulavani, Alexey Tumanov, and Ramachandran Ramjee. Taming Throughput-Latency tradeoff in LLM inference with Sarathi-Serve. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pp. 117–134, Santa Clara, CA, July 2024. USENIX Association. ISBN 978-1-939133-40-3. URL https://www.usenix.org/conference/osdi24/presentation/agrawal.

Ruicheng Ao, Gan Luo, David Simchi-Levi, and Xinshang Wang. Optimizing LLM inference: Fluid-based online scheduling under memory constraints. In *NeurIPS 2025 Workshop MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making*, 2025. URL https://openreview.net/forum?id=LFlsYLfkM8.

Agrim Bari, Parikshit Hegde, and Gustavo de Veciana. Optimal scheduling algorithms for llm inference: Theory and practice. *Proc. ACM Meas. Anal. Comput. Syst.*, 9(3), December 2025. doi: 10.1145/3771574. URL https://doi.org/10.1145/3771574.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Zixi Chen, Yinyu Ye, and Zijie Zhou. Adaptively robust LLM inference optimization under prediction uncertainty. *arXiv preprint arXiv:2508.14544*, 2025. doi: 10.48550/arXiv.2508.14544. URL https://arxiv.org/abs/2508.14544.

J.-G. Dai, Tian-Ze Deng, Yueying Li, and Tianyi Peng. Throughput-optimal scheduling algorithms for llm inference and ai agents. *ArXiv*, abs/2504.07347, 2025. URL https://api.semanticscholar.org/CorpusID:277667715.

Gregory Dexter, Shao Tang, Ata Fatahibaarzi, Qingquan Song, Tejas Dharamsi, and Aman Gupta. LLM query scheduling with prefix reuse and latency constraints. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=HKfZwLjSwQ.

Zohar Feldman, Avishai Mandelbaum, William A. Massey, and Ward Whitt. Staffing of time-varying queues to achieve time-stable performance. *Management Science*, 54(2):324–338, 2008. doi: 10.1287/mnsc.1070.0821. URL https://doi.org/10.1287/mnsc.1070.0821.

David Gamarnik. Stochastic bandwidth packing process: Stability conditions via lyapunov function technique. *Queueing Systems*, 48(3–4):339–363, 2004. doi: 10.1023/B:QUES.0000046581.34849.cf. URL https://doi.org/10.1023/B:QUES.0000046581.34849.cf.

David Gamarnik and Petar Momčilović. Steady-state analysis of a multiserver queue in the Halfin–Whitt regime. *Advances in Applied Probability*, 40(2):548–577, 2008.

Peter W. Glynn and Ward Whitt. Large deviations behavior of counting processes and their inverses. *Queueing Systems*, 17:107–128, 1994a.

Peter W. Glynn and Ward Whitt. Logarithmic asymptotics for steady-state tail probabilities in a single-server queue. *Journal of Applied Probability*, 31(A):131–156, 1994b.

Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations Research*, 29(3):567–588, 1981. doi: 10.1287/opre.29.3.567. URL https://doi.org/10.1287/opre.29.3.567.

Patrick Jaillet, Jiashuo Jiang, Konstantina Mellou, Marco Molinaro, Chara Podimata, and Zijie Zhou. Online scheduling for LLM inference with KV cache constraints. *arXiv preprint arXiv:2502.07115*, 2025. doi: 10.48550/arXiv.2502.07115. URL https://arxiv.org/abs/2502.07115.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, pp. 611–626, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702297. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Hui Li and Tao Yang. Queues with a variable number of servers. *European Journal of Operational Research*, 124(3):615–628, 2000. ISSN 0377-2217. doi: https://doi.org/10.1016/S0377-2217(99)00175-7.

Michael Mitzenmacher and Rana Shahout. Queueing, predictions, and large language models: Challenges and open problems. *Stochastic Systems*, 15(3):195–219, 2025. URL https://doi.org/10.1287/stsy.2025.0106.

Chengyi Nie, Nian Si, and Zijie Zhou. A queueing-theoretic framework for stability analysis of LLM inference with KV cache memory constraints. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=7X91AAKL5B.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. URL https://arxiv.org/abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Emre Ozbas and Melih Bastopcu. Queueing-aware optimization of reasoning tokens for accuracy-latency trade-offs in llm servers. In *2026 IEEE International Symposium on Information Theory (ISIT)*, pp. 1–6, 2026. doi: 10.1109/ISIT62367.2026.11653787.

Reiner Pope, Sholto Douglas, Aakanksha Chowdhery, Jacob Devlin, James Bradbury, Anselm Levskaya, Jonathan Heek, Ke-Fan Xiao, Shivani Agrawal, and Jeff Dean. Efficiently scaling transformer inference. *Proceedings of machine learning and systems*, 5:606–624, 2023.

SGLang. SGLang server arguments. SGLang documentation, 2026. URL https://docs.sglang.io/docs/advanced_features/server_arguments. Accessed September 21, 2026.

Haim Shore. Simple approximations for the GI/G/c queue—I: The steady-state probabilities. *Journal of the Operational Research Society*, 39(3):279–284, 1988. doi: 10.1057/jors.1988.45. URL https://doi.org/10.1057/jors.1988.45.

Frank Spitzer. A combinatorial lemma and its application to probability theory. *Transactions of the American Mathematical Society*, 82(2):323–339, 1956. URL https://doi.org/10.2307/1993051.

vLLM. vLLM engine arguments. vLLM v0.7.1 documentation, 2025. URL https://docs.vllm.ai/en/v0.7.1/serving/engine_args.html. Accessed September 21, 2026.

Kuo-Hsiung Wang and Keng-Yuan Tai. A queueing system with queue-dependent servers and finite capacity. *Applied Mathematical Modelling*, 24(11):807–814, 2000. ISSN 0307-904X. doi: https://doi.org/10.1016/S0307-904X(00)00013-5.

Meixuan Wang, Yinyu Ye, and Zijie Zhou. Llm serving optimization with variable prefill and decode lengths. *ArXiv*, abs/2508.06133, 2025. URL https://api.semanticscholar.org/CorpusID:280561678.

Ward Whitt. Approximations for the gi/g/m queue. *Production and Operations Management*, 2(2):114–161, 1993. doi: https://doi.org/10.1111/j.1937-5956.1993.tb00094.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1937-5956.1993.tb00094.x.

Yuqing Yang, Lei Jiao, and Yuedong Xu. A queueing theoretic perspective on low-latency llm inference with variable token length. In *2024 22nd International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pp. 273–280, 2024.

Yinmin Zhong, Shengyu Liu, Junda Chen, Jianbo Hu, Yibo Zhu, Xuanzhe Liu, Xin Jin, and Hao Zhang. DistServe: Disaggregating prefill and decoding for goodput-optimized large language model serving. In *18th USENIX Symposium on Operating Systems Design and Implementation*, pp. 193–210, Santa Clara, CA, July 2024. USENIX Association. ISBN 978-1-939133-40-3. URL https://www.usenix.org/conference/osdi24/presentation/zhong-yinmin.

Noa Zychlinski, Carri W. Chan, and Jing Dong. Managing queues with different resource requirements. *Operations Research*, 71(4):1387–1413, 2022. doi: 10.1287/opre.2022.2284. URL https://doi.org/10.1287/opre.2022.2284.

A TANDEM QUEUEING MODEL

Figure A.1 illustrates the tandem system in Section 3. Requests first receive prefill service and then join decode. The decode instance admits concurrent requests subject to both the concurrency cap B and the shared KV-cache budget M; requests awaiting admission remain in the decode queue.

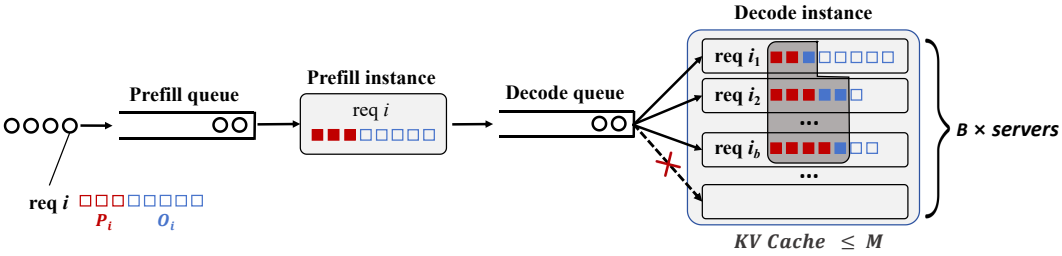


Figure A.1: Queueing representation of P–D disaggregated inference. Each request has prompt length χ_i and output length ω_i. Prefill is represented by a single service position, whereas decode shares its concurrency and KV-cache resources across active requests.

B PREFILL ANALYSIS

B.1 PROOF OF THEOREM 1

We use the slot-boundary state and transition order in Section 4. Arrivals are i.i.d. across slots and independent of the i.i.d. positive integer service requirements. Write

$$a(z) = G_A(z), \tag{B.1}$$
$$b(u) = G_S(u), \tag{B.2}$$
$$H(z) = b(a(z)), \tag{B.3}$$
$$\lambda_{\mathrm{P}} = \mathbb{E}\left[A_t\right], \tag{B.4}$$
$$\rho_{\mathrm{P}} = \lambda_{\mathrm{P}}\mathbb{E}\left[S\right] < 1. \tag{B.5}$$

Stationarity. Let T_t be the total unfinished service work at epoch t, including the residual work in service, and write $x^+ = \max\{x, 0\}$. Revealing the service requirements of waiting requests gives

$$T_{t+1} = (T_t - 1)^+ + U_{t+1}, \tag{B.6}$$
$$U_{t+1} = \sum_{i=1}^{A_{t+1}} S_{t+1,i}, \tag{B.7}$$
$$\mathbb{E}\left[U_{t+1}\right] = \rho_{\mathrm{P}}. \tag{B.8}$$

Here $S_{t,i}$ are i.i.d. copies of S, independent of the arrival counts. For $T_t \geq 1$, the conditional drift is $\rho_{\mathrm{P}} - 1 < 0$. The expected drift at zero is finite, so the accessible workload class is positive recurrent. Empty epochs regenerate the FCFS queue, which therefore has a stationary law on the class accessible from the empty state. Since $S \geq 1$ and $\rho_{\mathrm{P}} < 1$, $q_0 > 0$. The empty-state self-loop makes this class aperiodic. For $\lambda_{\mathrm{P}} = 0$ the population is identically zero; the derivation below considers $\lambda_{\mathrm{P}} > 0$.

Elimination of the residual-service coordinate. Define

$$p_0 = \Pr\left\{K_{\mathrm{P}} = 0\right\}, \tag{B.9}$$
$$p_{k,r} = \Pr\left\{K_{\mathrm{P}} = k, R_{\mathrm{P}} = r\right\}, \tag{B.10}$$
$$F_r(z) = \sum_{k\geq 1} p_{k,r} z^k, \tag{B.11}$$

so that

$$\Pi_{\mathrm{P}}(z) = p_0 + \sum_{r\geq 1} F_r(z). \tag{B.12}$$

An empty state is reached only from $(0, 0)$ or from $(1, 1)$, in either case with no new arrival. Hence

$$p_0 = q_0(p_0 + p_{1,1}). \tag{B.13}$$

A residual phase r is reached by decrementing phase $r + 1$ or by starting a request of service length r. Therefore

$$F_r(z) = a(z)F_{r+1}(z) + g_r C(z), \tag{B.14}$$
$$\begin{aligned} C(z) &= p_0(a(z) - q_0) + \frac{a(z)F_1(z)}{z} - q_0 p_{1,1} \\ &= p_0(a(z) - 1) + \frac{a(z)F_1(z)}{z}, \end{aligned} \tag{B.15}$$

where the last equality follows from (B.13). Backward substitution over the bounded service support gives

$$F_r(z) = C(z)\sum_{s\geq r} g_s a(z)^{s-r}, \tag{B.16}$$
$$F_1(z) = \frac{C(z)H(z)}{a(z)}. \tag{B.17}$$

Substitution into (B.15) yields

$$C(z)(H(z)-z) = p_0 z(1-a(z)). \tag{B.18}$$

Summing the residual phases then gives

$$\sum_{r\geq 1} F_r(z) = C(z)\sum_{s\geq 1} g_s \sum_{r=1}^{s} a(z)^{s-r} = \frac{C(z)(1-H(z))}{1-a(z)} = \frac{p_0 z(1-H(z))}{H(z)-z}. \tag{B.19}$$

Combining this identity with (B.12),

$$\Pi_{\mathrm{P}}(z) = \frac{p_0(1-z)H(z)}{H(z)-z}. \tag{B.20}$$

These identities are derived for $0 < z < 1$ and extended to $z = 0$ by continuity; in particular, $F_1(0) = 0$.

Normalization. The composition $H(z)$ is the PGF of arrivals during one service:

$$H(z) = \mathbb{E}\left[z^{\sum_{u=1}^{S} A_u}\right], \tag{B.21}$$

$$H(1) = 1, \tag{B.22}$$

$$H'(1) = \rho_{\mathrm{P}}. \tag{B.23}$$

In one-sided limits, $\uparrow$ denotes approach from below. Thus

$$H(z) - z = (\rho_{\mathrm{P}} - 1)(z-1) + o(|z-1|), \tag{B.24}$$

$$\lim_{z\uparrow 1} \frac{(1-z)H(z)}{H(z)-z} = \frac{1}{1-\rho_{\mathrm{P}}}. \tag{B.25}$$

Since $\Pi_{\mathrm{P}}(1) = 1$, $p_0 = 1 - \rho_{\mathrm{P}}$, and coefficient extraction proves Theorem 1:

$$\Pi_{\mathrm{P}}(z) = \frac{(1-\rho_{\mathrm{P}})(1-z)G_S(G_A(z))}{G_S(G_A(z)) - z}, \tag{B.26}$$

$$\pi_k^{\mathrm{P}} = [z^k]\Pi_{\mathrm{P}}(z), \quad k \geq 0. \tag{B.27}$$

B.2 Departure-count interface

Let $\mathsf{P}^{(0)}$ and $\mathsf{P}^{(1)}$ retain the one-slot prefill transition probabilities with zero and one departure, respectively, with all other entries set to zero. Both matrices are indexed by the joint states $(K_{\mathrm{P}}, R_{\mathrm{P}})$. Every transition from residual phase 1, including one into the empty state, is included in $\mathsf{P}^{(1)}$; all other transitions are included in $\mathsf{P}^{(0)}$. For the PGF variable $v \in [0,1]$, define the marked transition matrix

$$\mathsf{P}_v = \mathsf{P}^{(0)} + v\mathsf{P}^{(1)}. \tag{B.28}$$

At $v = 1$, $\mathsf{P}_1 = \mathsf{P}^{(0)} + \mathsf{P}^{(1)}$ is the full prefill transition matrix, whereas $\mathsf{P}^{(1)}$ contains only one-departure transitions. Let $\boldsymbol{\pi}_{\mathrm{P}}$ be its stationary row vector, whose components are p_0 and $p_{k,r} = [z^k]F_r(z)$ from the preceding derivation. Then

$$\boldsymbol{\pi}_{\mathrm{P}}\mathsf{P}_1 = \boldsymbol{\pi}_{\mathrm{P}}. \tag{B.29}$$

Write $\mathbf{1}$ for an all-ones column vector with dimensions determined by context; here its entries are indexed by the joint prefill states.

Suppose $\Delta_{\mathrm{D}} = d\delta$ with $d \in \mathbb{Z}_{\geq 1}$ and aligned iteration boundaries. Let D_d count prefill departures over d consecutive slots in stationarity, so that $A^{\mathrm{D}} \overset{\mathrm{d}}{=} D_d$. Each departure contributes a factor v in a product of marked transition probabilities. Summing over the initial and final states gives

$$G_{\mathrm{D}}(v) = \mathbb{E}\left[v^{D_d}\right] = \boldsymbol{\pi}_{\mathrm{P}}\left(\mathsf{P}_v\right)^d \mathbf{1}. \tag{B.30}$$

Here, $(\mathsf{P}_v)^d$ is the d-fold matrix product, corresponding to d successive slot transitions. Consequently,

$$q_a^{\mathrm{D}} = [v^a]G_{\mathrm{D}}(v), \qquad a \in \mathbb{Z}_{\geq 0}, \tag{B.31}$$

$$\lambda_{\mathrm{D}} = G'_{\mathrm{D}}(1) = d\lambda_{\mathrm{P}}, \tag{B.32}$$

$$\mathrm{Var}(D_d) = G''_{\mathrm{D}}(1) + G'_{\mathrm{D}}(1) - \left[G'_{\mathrm{D}}(1)\right]^2. \tag{B.33}$$

The mean identity follows from stationary request-rate conservation.

For two consecutive d-slot blocks, let $D_d^{(1)}$ and $D_d^{(2)}$ denote the departure counts in the first and second blocks, respectively, and let $v_1, v_2 \in [0,1]$ be their PGF variables. Both counts have the same marginal law as D_d, but need not be independent. Their joint PGF is

$$\mathbb{E}\left[v_1^{D_d^{(1)}} v_2^{D_d^{(2)}}\right] = \boldsymbol{\pi}_{\mathrm{P}} \left(\mathsf{P}_{v_1}\right)^d \left(\mathsf{P}_{v_2}\right)^d \mathbf{1}. \tag{B.34}$$

The intermediate prefill state is carried across the block boundary, so this expression retains dependence between the two counts. The decode approximation instead uses independent per-iteration samples from the marginal law (B.30), with independent request marks.

C Proof of Theorem 2

Reference quantities. Consider the stable, lossless memory-only reference system of Section 5.2, with uninterrupted service for admitted requests. Immediately after admission and before decode iteration j, let $N_j^{(M)}$, $Q_j^{(M)}$, and $R_j^{(M)}$ denote the active population, waiting population, and active KV occupancy, respectively. The total number of unfinished requests is

$$K_{\mathrm{D},j}^{(M)} = N_j^{(M)} + Q_j^{(M)}. \tag{C.1}$$

We omit the time index for stationary random variables and write $Q^{(M)} = Q_{\mathrm{D}}^{(M)}$ and $K_{\mathrm{D}}^{(M)} = N^{(M)} + Q^{(M)}$.

The analytical construction uses i.i.d. per-iteration arrival counts A_j^{D}, independent of the i.i.d. request marks (χ_i, ω_i). Write

$$q_m^{\mathrm{D}} = \Pr\left\{A^{\mathrm{D}} = m\right\}, \tag{C.2}$$

$$G_{\mathrm{D}}(z) = \sum_{m \geq 0} q_m^{\mathrm{D}} z^m, \tag{C.3}$$

$$\lambda_{\mathrm{D}} = G_{\mathrm{D}}'(1) > 0. \tag{C.4}$$

As in Section 5.2, assume $\chi \perp \omega$ and bounded positive integer supports. Define

$$\chi_{\min} = \min \operatorname{supp}(\chi), \tag{C.5}$$

$$G_\chi(y) = \mathbb{E}\left[y^\chi\right], \tag{C.6}$$

$$\bar{F}_\chi(\ell) = \Pr\left\{\chi > \ell\right\}, \tag{C.7}$$

$$s_g = \Pr\left\{\omega > g\right\}. \tag{C.8}$$

We take M to be an integer admission budget; otherwise, use $\lfloor M \rfloor$ for the integer-valued occupancy tests and sums. For end-to-end evaluation, the prefill departure-count PGF in (B.30) supplies G_{D} under the i.i.d. batch-input abstraction described in Appendix B.2.

At active age $g \in \{0, \ldots, \omega - 1\}$, a request occupies $\chi + g$ KV units. Its total resource-time contribution is therefore

$$W = \sum_{g=0}^{\omega-1} (\chi + g) = \chi\omega + \frac{\omega(\omega-1)}{2}, \tag{C.9}$$

$$M_W(v) = \mathbb{E}\left[e^{vW}\right]. \tag{C.10}$$

For request i, write $W_i = \chi_i \omega_i + \frac{\omega_i(\omega_i - 1)}{2}$. In particular,

$$\mathbb{E}[W] = \mathbb{E}[\chi]\,\mathbb{E}[\omega] + \frac{\mathbb{E}\left[\omega^2\right] - \mathbb{E}[\omega]}{2}. \tag{C.11}$$

Infinite-server law. Remove both admission constraints and let each request begin decode immediately on arrival. Denote its stationary active count and KV occupancy by (N_∞, R_∞). Consider the cohort that arrived g iterations before the observation epoch. Each request survives exactly when

$\omega > g$, in which case it contributes one request and $\chi + g$ KV units. Writing $\mathbb{I}_E$ for the indicator of event E, its joint PGF is

$$f_g(z, y) = \mathbb{E}\left[\mathbb{I}_{\{\omega \le g\}} + \mathbb{I}_{\{\omega > g\}} z y^{\chi+g}\right] = 1 - s_g + s_g z y^g G_\chi(y). \tag{C.12}$$

Conditional on a cohort size m, the cohort PGF is $f_g(z, y)^m$. Averaging over that size gives $G_{\mathrm{D}}(f_g(z, y))$. Since the cohorts are independent and no request survives beyond $\omega_{\max}$ iterations,

$$\mathcal{G}_\infty(z, y) = \mathbb{E}\left[z^{N_\infty} y^{R_\infty}\right] = \prod_{g=0}^{\omega_{\max}-1} G_{\mathrm{D}}\left(1 - s_g + s_g z y^g G_\chi(y)\right). \tag{C.13}$$

Consequently, the joint reference probabilities are

$$p^\infty_{b,r} = [z^b y^r]\mathcal{G}_\infty(z, y) = \Pr\left\{N_\infty = b, R_\infty = r\right\}. \tag{C.14}$$

We set $p^\infty_{b,r} = 0$ outside the nonnegative integer support. In particular, $p^\infty_{b,r} = 0$ whenever $r < b\chi_{\min}$.

Differentiating the PGF at $(z, y) = (1, 1)$ yields

$$\mathbb{E}\left[N_\infty\right] = \lambda_{\mathrm{D}} \sum_{g=0}^{\omega_{\max}-1} s_g = \lambda_{\mathrm{D}} \mathbb{E}\left[\omega\right], \tag{C.15}$$

$$\mathbb{E}\left[R_\infty\right] = \lambda_{\mathrm{D}} \sum_{g=0}^{\omega_{\max}-1} \mathbb{E}\left[(\chi + g)\mathbb{I}_{\{\omega > g\}}\right] = \lambda_{\mathrm{D}} \mathbb{E}\left[W\right]. \tag{C.16}$$

Conditional active-population laws. For the no-backlog branch, we approximate the joint distribution of $(N^{(M)}, R^{(M)})$ conditional on $Q^{(M)} = 0$ using the memory-feasible conditional distribution of the infinite-server reference. For $b \in \mathbb{Z}_{\ge 0}$ and $r \in \{0, \dots, M\}$, the reference conditional PMF is

$$\Pr\{N_\infty = b, R_\infty = r \mid R_\infty \le M\} = \frac{p^\infty_{b,r}}{\Pr\{R_\infty \le M\}}. \tag{C.17}$$

Its unnormalized active-count weights and normalizing constant are

$$a_b(M) = \sum_{r=0}^{M} p^\infty_{b,r}, \tag{C.18}$$

$$Z_0(M) = \sum_{b \ge 0} a_b(M) = \Pr\left\{R_\infty \le M\right\}. \tag{C.19}$$

Marginalization and expectation under this conditional reference distribution give

$$\widehat{f}_0(b) = \frac{a_b(M)}{Z_0(M)}, \tag{C.20}$$

$$\widehat{R}_0 = \frac{\sum_{b \ge 0} \sum_{r=0}^{M} r p^\infty_{b,r}}{Z_0(M)}. \tag{C.21}$$

For the backlog branch, introduce an independent reference prompt χ° with the same law as χ. Given a feasible occupancy R_∞, the prompt is blocked when it exceeds the remaining memory. Define

$$\Xi = \{R_\infty \le M,\ \chi^\circ > M - R_\infty\}. \tag{C.22}$$

For $u_b(M) = \Pr\left\{N_\infty = b, \Xi\right\}$, conditioning on $R_\infty = r$ and using the independence of χ° gives

$$u_b(M) = \sum_{r=0}^{M} p^\infty_{b,r} \bar{F}_\chi(M - r) = \sum_{\ell=0}^{\min(M, \chi_{\max}-1)} p^\infty_{b,M-\ell} \bar{F}_\chi(\ell). \tag{C.23}$$

The last expression follows by setting $\ell = M - r$ and using $\bar{F}_\chi(\ell) = 0$ for $\ell \geq \chi_{\max}$. Normalization yields

$$Z_+(M) = \sum_{b \geq 0} u_b(M) = \Pr\{\Xi\}, \tag{C.24}$$

$$\widehat{f}_+(b) = \frac{u_b(M)}{Z_+(M)}, \tag{C.25}$$

$$\widehat{C}_+ = \frac{\sum_{b \geq 0} \sum_{\ell=0}^{\min(M, \chi_{\max}-1)} (M-\ell) p^{\infty}_{b, M-\ell} \bar{F}_\chi(\ell)}{Z_+(M)}. \tag{C.26}$$

For the backlog branch, we approximate the joint distribution of $(N^{(M)}, R^{(M)})$ conditional on $Q^{(M)} > 0$ using the infinite-server reference distribution conditioned on Ξ. Its conditional PMF is

$$\Pr\{N_\infty = b, R_\infty = r \mid \Xi\} = \frac{p^{\infty}_{b,r} \bar{F}_\chi(M-r)}{Z_+(M)}, \qquad 0 \leq r \leq M. \tag{C.27}$$

Thus $\widehat{f}_+$ and $\widehat{C}_+$ are, respectively, the active-count PMF and mean occupancy of this conditional reference distribution.

Conservation and mixing weights. Let T_i^{adm} be the admission iteration of request i and define the age-g reward of the cohort admitted at iteration t by

$$V_{t,g} = \sum_{i : T_i^{\mathrm{adm}} = t} (\chi_i + g) \mathbb{I}_{\{\omega_i > g\}}. \tag{C.28}$$

Since admitted requests receive uninterrupted decode service, the occupancy at iteration j decomposes pathwise as

$$R_j^{(M)} = \sum_{g=0}^{\omega_{\max}-1} V_{j-g,g}. \tag{C.29}$$

Taking expectations and applying stationarity gives

$$\mathbb{E}\left[R_j^{(M)}\right] = \sum_{g=0}^{\omega_{\max}-1} \mathbb{E}\left[V_{0,g}\right] = \mathbb{E}\left[\sum_{i : T_i^{\mathrm{adm}} = 0} \sum_{g=0}^{\omega_i - 1} (\chi_i + g)\right] = \mathbb{E}\left[\sum_{i : T_i^{\mathrm{adm}} = 0} W_i\right] = \lambda_{\mathrm{D}} \mathbb{E}\left[W\right]. \tag{C.30}$$

The last equality follows from marked rate conservation: each arriving request is admitted once with the same lifetime work.

Let

$$p_+ = \Pr\left\{Q^{(M)} > 0\right\}. \tag{C.31}$$

Conditioning the conservation identity on the backlog event gives

$$\lambda_{\mathrm{D}} \mathbb{E}\left[W\right] = (1 - p_+) \mathbb{E}\left[R^{(M)} \mid Q^{(M)} = 0\right] + p_+ \mathbb{E}\left[R^{(M)} \mid Q^{(M)} > 0\right]. \tag{C.32}$$

Substituting the two reference means from (C.21) and (C.26) reduces this identity to

$$\lambda_{\mathrm{D}} \mathbb{E}\left[W\right] \approx \widehat{R}_0 + p_+ (\widehat{C}_+ - \widehat{R}_0). \tag{C.33}$$

Solving for the mixing weights yields

$$\widehat{p}_+ = \frac{\lambda_{\mathrm{D}} \mathbb{E}\left[W\right] - \widehat{R}_0}{\widehat{C}_+ - \widehat{R}_0}, \tag{C.34}$$

$$1 - \widehat{p}_+ = \frac{\widehat{C}_+ - \lambda_{\mathrm{D}} \mathbb{E}\left[W\right]}{\widehat{C}_+ - \widehat{R}_0}. \tag{C.35}$$

The two-branch mixture is used with

$$Z_0(M) > 0, \tag{C.36}$$
$$Z_+(M) > 0, \tag{C.37}$$
$$0 \le \widehat{R}_0 < \widehat{C}_+, \tag{C.38}$$
$$\widehat{R}_0 \le \lambda_{\mathrm{D}} \mathbb{E}[W] \le \widehat{C}_+. \tag{C.39}$$

These conditions ensure that both conditional laws and the mixture are normalized and nonnegative.

Concurrency threshold. From a given reference state, a concurrency cap b changes the memory-only batch selection precisely when that selected batch contains more than b requests. Hence

$$\Pr\{\text{cap } b \text{ changes the selected batch}\} = \Pr\left\{N^{(M)} > b\right\}, \tag{C.40}$$

and

$$B_\theta(M) = \min\left\{b \in \mathbb{Z}_{\ge 0} : \Pr\left\{N^{(M)} > b\right\} \le \theta\right\}, \qquad 0 < \theta < 1. \tag{C.41}$$

The strict exceedance reflects that a batch of exactly b requests is feasible under a cap of b.

Set

$$n_M = \left\lfloor \frac{M}{\chi_{\min}} \right\rfloor. \tag{C.42}$$

At each successful admission, the resulting active set $\mathcal{B}$ satisfies

$$|\mathcal{B}|\chi_{\min} \le \sum_{i \in \mathcal{B}} (\chi_i + g_i) \le M. \tag{C.43}$$

Between admissions its cardinality cannot increase. Starting from an empty configuration, induction over admissions and completions therefore gives $N_j^{(M)} \le n_M$, and the same bound holds under the stationary law. In particular, $\Pr\left\{N^{(M)} > n_M\right\} = 0$. The set in (C.41) is nonempty and has an integer minimum, with

$$0 \le B_\theta(M) \le n_M. \tag{C.44}$$

By total probability,

$$\Pr\left\{N^{(M)} = b\right\} = (1 - p_+) \Pr\left\{N^{(M)} = b \mid Q^{(M)} = 0\right\} + p_+ \Pr\left\{N^{(M)} = b \mid Q^{(M)} > 0\right\}. \tag{C.45}$$

Using the branch laws and the weight already derived gives

$$\widehat{\pi}_N^{(M)}(b) = \frac{(1 - \widehat{p}_+) a_b(M)}{Z_0(M)} + \frac{\widehat{p}_+ u_b(M)}{Z_+(M)}. \tag{C.46}$$

The PMF is nonnegative and sums to one. Moreover, both components vanish for $b > n_M$, since $p_{b,r}^{\infty} = 0$ for $r < b\chi_{\min}$. Its strict upper tail, $\widehat{T}_M(b) \triangleq \sum_{j>b} \widehat{\pi}_N^{(M)}(j)$, is

$$\widehat{T}_M(b) = \frac{(1 - \widehat{p}_+) \sum_{j>b} a_j(M)}{Z_0(M)} + \frac{\widehat{p}_+ \sum_{j>b} u_j(M)}{Z_+(M)}. \tag{C.47}$$

Consequently,

$$\widehat{T}_M(b) - \widehat{T}_M(b+1) = \widehat{\pi}_N^{(M)}(b+1) \ge 0, \tag{C.48}$$
$$\widehat{T}_M(n_M) = 0. \tag{C.49}$$

Substituting this tail into the threshold definition yields

$$\widehat{B}_\theta(M) = \min\left\{b \in \mathbb{Z}_{\ge 0} : \frac{(1 - \widehat{p}_+) \sum_{j>b} a_j(M)}{Z_0(M)} + \frac{\widehat{p}_+ \sum_{j>b} u_j(M)}{Z_+(M)} \le \theta\right\}. \tag{C.50}$$

This is Theorem 2. The selected threshold is the smallest integer satisfying the strict-tail constraint.

Evaluation of the reference law. The reference probabilities are defined by coefficient extraction in (C.14). For positive contour radii r_z, r_y inside the domain of convergence, the corresponding coefficient formula is

$$p^{\infty}_{b,r} = \frac{r_z^{-b} r_y^{-r}}{(2\pi)^2} \int_{-\pi}^{\pi} \int_{-\pi}^{\pi} \mathcal{G}_{\infty}(r_z e^{iu}, r_y e^{iv}) e^{-i(bu+rv)}\, du\, dv. \tag{C.51}$$

For saddlepoint evaluation, introduce the joint cumulant generating function

$$\Gamma(\alpha, \xi) = \log \mathcal{G}_{\infty}(e^{\alpha}, e^{\xi}) = \sum_{g=0}^{\omega_{\max}-1} \log G_{\mathrm{D}}\left(1 - s_g + s_g e^{\alpha+g\xi} G_{\chi}(e^{\xi})\right). \tag{C.52}$$

The tilting parameters for a target (b, r) satisfy

$$\partial_{\alpha}\Gamma(\alpha, \xi) = b, \tag{C.53}$$

$$\partial_{\xi}\Gamma(\alpha, \xi) = r. \tag{C.54}$$

Substituting the reference probabilities into (C.18)–(C.34) determines the common quantities used here and in the population PMF.

D Proof of Theorem 3

We use the active-count laws, occupancy means, and mixture weight from Appendix C. It remains to obtain the waiting-population law π_q^R and combine it with those quantities.

Workload reduction. Use the lifetime work W from (C.9) and the congested occupancy $\widehat{C}_+$ from (C.26). Their ratio defines the normalized service requirement:

$$S_i^+ = \frac{W_i}{\widehat{C}_+}, \tag{D.1}$$

$$Y_t = \sum_{i=1}^{A_t^{\mathrm{D}}} S_{t,i}^+, \tag{D.2}$$

$$X_t = Y_t - 1. \tag{D.3}$$

Let V_t be the unfinished surrogate work immediately before the batch at integer epoch t. Work is processed at rate one between successive batch epochs, so

$$V_{t+1} = (V_t + Y_t - 1)^+. \tag{D.4}$$

Here t counts decode iterations, $S_{t,i}^+$ are i.i.d. copies of $S^+ = \frac{W}{\widehat{C}_+}$, and the workload is generally noninteger.

Using the transform M_W defined in (C.10) and independence of arrival batches and marks gives

$$\varphi_X(v) = \mathbb{E}\left[e^{vX_t}\right] = e^{-v} G_{\mathrm{D}}\left(M_W\left(\frac{v}{\widehat{C}_+}\right)\right), \tag{D.5}$$

$$\mathbb{E}\left[X_t\right] = \rho_+ - 1, \tag{D.6}$$

$$\rho_+ = \frac{\lambda_{\mathrm{D}} \mathbb{E}\left[W\right]}{\widehat{C}_+} < 1. \tag{D.7}$$

We use $\rho_+ < 1$ together with the mixture conditions (C.36)–(C.39).

Let $T_n^X = X_1 + \cdots + X_n$ and $T_0^X = 0$. Negative drift implies $T_n^X \to -\infty$ almost surely, and the stationary solution of (D.4) has the random-walk maximum representation

$$V \overset{d}{=} \sup_{n \geq 0} T_n^X. \tag{D.8}$$

Iterating the recursion backward and reversing its i.i.d. increments gives this representation.

Let $\mathcal{M}_n = \max_{0\le j\le n} T_j^X$. The generating-function form of Spitzer's identity (Spitzer, 1956) is

$$\sum_{n\ge 0} r^n \mathbb{E}\left[e^{-s\mathcal{M}_n}\right] = \exp\left(\sum_{n\ge 1} \frac{r^n \mathbb{E}\left[e^{-s(T_n^X)^+}\right]}{n}\right), \qquad 0\le r<1,\ s\ge 0. \tag{D.9}$$

Multiplying by $1-r = \exp\left(-\sum_{n\ge 1}\frac{r^n}{n}\right)$ and taking $r\uparrow 1$ gives

$$L_V(s) = \mathbb{E}\left[e^{-sV}\right] = \exp\left(\sum_{n\ge 1} \frac{\mathbb{E}\left[e^{-s(T_n^X)^+}\right]-1}{n}\right). \tag{D.10}$$

The left side converges by the Abelian limit and bounded convergence; on the right, monotone convergence applies to $\sum_{n\ge 1}\frac{r^n\left(1-\mathbb{E}\left[e^{-s(T_n^X)^+}\right]\right)}{n}$.

Waiting-population law. Within a batch, let r requests precede a tagged request. Its time to service start in the pooled surrogate is

$$Z_r = V + \sum_{i=1}^{r} S_i^+, \tag{D.11}$$

$$L_{Z_r}(s) = L_V(s) L_S(s)^r, \tag{D.12}$$

$$L_S(s) = M_W\left(-\frac{s}{\widehat{C}_+}\right). \tag{D.13}$$

Here V is independent of the new batch and its marks, and an empty sum is zero. Since the queue excludes a request once it starts service, the relevant event after elapsed time x is $\{Z_r > x\}$.

The tail transform follows directly by nonnegative integration:

$$\int_0^\infty e^{-sx}\Pr\{Z_r > x\}\,dx = \mathbb{E}\left[\int_0^{Z_r} e^{-sx}\,dx\right] = \frac{1-L_V(s)L_S(s)^r}{s}, \qquad s>0. \tag{D.14}$$

To account for service progress within a normalized workload interval, let U be independent and uniform on $[0,1)$ and observe the surrogate at phase U. A request from the batch at epoch $-n$ has then been in the system for $n+U$ time units. Define

$$H_{r,n} = \Pr\{Z_r > n+U\} = \int_n^{n+1} \Pr\{Z_r > x\}\,dx, \qquad n\ge 0, \tag{D.15}$$

$$I_r(x) = \int_0^x \Pr\{Z_r > u\}\,du, \tag{D.16}$$

$$H_{r,n} = I_r(n+1) - I_r(n), \tag{D.17}$$

$$\int_0^\infty e^{-sx} I_r(x)\,dx = \frac{1-L_V(s)L_S(s)^r}{s^2}. \tag{D.18}$$

Use the arrival probabilities q_m^{D} defined in (C.2). For the total count in n subsequent arrival batches, define

$$\alpha_n(v) = [z^v] G_{\mathrm{D}}(z)^n, \tag{D.19}$$

$$\alpha_0(v) = \mathbb{I}_{\{v=0\}}, \tag{D.20}$$

$$\alpha_n(v) = 0 \quad (v<0). \tag{D.21}$$

Using the kernel in (D.15), define

$$\psi_{n,j} = \sum_{m\ge j} q_m^{\mathrm{D}} H_{m-j,n}, \qquad j\ge 1, \tag{D.22}$$

$$\mathcal{C}_q = \sum_{n\ge 0}\sum_{j=1}^{q} \alpha_n(q-j)\psi_{n,j}, \qquad q\ge 1. \tag{D.23}$$

The following FCFS counting argument establishes that $\mathcal{C}_q$ is the tail probability of the surrogate waiting population Q^R.

Fix $q \geq 1$ and enumerate requests backward from the most recent arrival. Almost surely, there is a unique batch containing the qth most recent request. Suppose this batch arrived at epoch $-n$, has size m, and contains exactly j requests from the tagged request through the end of that batch. The n subsequent batches must contain $q-j$ requests, an event of probability $\alpha_n(q-j)$, and the tagged request has $m-j$ predecessors in its own batch. By FCFS, the waiting population is at least q if and only if this request has not started service. At the selected observation phase, this event is $Z_{m-j} > n+U$. The workload preceding the old batch and its marks are independent of all later batches and of U. Averaging over U, this case has probability $\alpha_n(q-j)\, q_m^{\mathrm{D}}\, H_{m-j,n}$. The cases are disjoint because the batch and position of the tagged request are unique. Summing over m, n, j proves $\mathcal{C}_q = \Pr\left\{Q^R \geq q\right\}$. All sums are justified by nonnegativity. Consequently,

$$\pi_0^R = 1 - \mathcal{C}_1, \tag{D.24}$$

$$\pi_q^R = \mathcal{C}_q - \mathcal{C}_{q+1}, \quad q \geq 1. \tag{D.25}$$

These probabilities telescope to one because the stationary surrogate has a finite waiting population almost surely.

The resulting phase-averaged workload count is the waiting law used in the two-regime approximation.

Total-population mixture. For $q \geq 1$, define

$$\widehat{g}_+(q) = \frac{\pi_q^R}{1-\pi_0^R}, \tag{D.26}$$

$$\widehat{g}_+(0) = 0, \tag{D.27}$$

provided $1-\pi_0^R > 0$. The last closure is

$$\Pr\left\{N^{(M)} = b, Q^{(M)} = q \mid Q^{(M)} > 0\right\} \;\approx\; \widehat{f}_+(b)\widehat{g}_+(q). \tag{D.28}$$

This is the conditional active/waiting factorization used in the two-regime approximation.

The exact law of total probability, followed by the specified closures, now gives

$$\begin{aligned}\Pr\left\{K_{\mathrm{D}}^{(M)} = k\right\} &= (1-p_+)\Pr\left\{N^{(M)} = k \mid Q^{(M)} = 0\right\} \\ &\quad + p_+ \sum_{b=0}^{k-1} \Pr\left\{N^{(M)} = b, Q^{(M)} = k-b \mid Q^{(M)} > 0\right\} \\ &\approx (1-\widehat{p}_+)\widehat{f}_0(k) + \widehat{p}_+ \sum_{b=0}^{k-1} \widehat{f}_+(b)\widehat{g}_+(k-b).\end{aligned} \tag{D.29}$$

Substituting (C.20), (C.25), (C.34), and (D.26) gives

$$\widehat{\pi}_k^{(M)} = \frac{(1-\widehat{p}_+)a_k(M)}{Z_0(M)} + \sum_{b=0}^{k-1} \frac{\widehat{p}_+ u_b(M)\pi_{k-b}^R}{Z_+(M)(1-\pi_0^R)}, \qquad k \geq 0. \tag{D.30}$$

For $k=0$ the sum is empty. Normalization follows from

$$\sum_{k\geq 0} \widehat{\pi}_k^{(M)} = (1-\widehat{p}_+)\sum_{k\geq 0}\widehat{f}_0(k) + \widehat{p}_+ \sum_{b\geq 0}\widehat{f}_+(b)\sum_{q\geq 1}\widehat{g}_+(q) = 1. \tag{D.31}$$

Nonnegativity follows from (C.36)–(C.39). If $\widehat{p}_+ = 0$, the expression is defined directly as $\widehat{\pi}_k^{(M)} = \widehat{f}_0(k)$ without evaluating an unnecessary positive-queue normalization. For $\widehat{p}_+ > 0$, the positive-queue component is used under the condition $1-\pi_0^R > 0$.

This proves the approximation in Theorem 3.

E PROOF OF COROLLARY 1

Fix $M > 0$ and use the saturated memory-work representation, in which a backlogged system processes M units of aggregate work per decode iteration. The request work W and its moment generating function M_W are defined in (C.9) and (C.10). Let A_n^{D} be i.i.d. iteration arrival counts, independent of the i.i.d. request works W_i. Throughout this proof, t counts integer decode iterations and logarithms are natural. Recall

$$G_{\mathrm{D}}(z) = \mathbb{E}\left[z^{A^{\mathrm{D}}}\right], \tag{E.1}$$

$$\Lambda_{\mathrm{D}}(s) = \log G_{\mathrm{D}}(e^s), \tag{E.2}$$

$$\Lambda_W(a) = \log M_W(a). \tag{E.3}$$

The bounded positive request lengths imply

$$0 < w_{\min} \triangleq \min \operatorname{supp}(W) \le W \le w_{\max} \triangleq \max \operatorname{supp}(W) < \infty. \tag{E.4}$$

Arrival and potential-service transforms. For the cumulative arrival count $\mathcal{A}(t) = \sum_{n=1}^{t} A_n^{\mathrm{D}}$, independence yields

$$\mathbb{E}\left[e^{s\mathcal{A}(t)}\right] = \prod_{n=1}^{t} \mathbb{E}\left[e^{sA_n^{\mathrm{D}}}\right] = e^{t\Lambda_{\mathrm{D}}(s)}. \tag{E.5}$$

Let $S_0^W = 0$ and $S_n^W = \sum_{i=1}^{n} W_i$. For a work budget $u \ge 0$, define the renewal count

$$\mathcal{N}(u) = \max\left\{n \ge 0 : S_n^W \le u\right\}. \tag{E.6}$$

A work budget Mt completes the first n requests exactly when $S_n^W \le Mt$, so the potential service count is

$$D_M(t) = \mathcal{N}(Mt). \tag{E.7}$$

Thus both $\mathcal{A}(t)$ and $D_M(t)$ are measured in requests.

For $s \ge 0$, define $\alpha(s)$ by

$$\Lambda_W(\alpha(s)) = s. \tag{E.8}$$

Since W is bounded and strictly positive, Λ_W is finite and differentiable on $\mathbb{R}$, with $\Lambda_W(0) = 0$, $\Lambda_W'(a) > 0$, and $\Lambda_W(a) \to \infty$ as $a \to \infty$. Hence $\alpha(s)$ exists uniquely, with $\alpha(0) = 0$ and $\alpha(s) > 0$ for $s > 0$. Moreover,

$$\frac{1}{n} \log \mathbb{E}\left[e^{aS_n^W}\right] = \Lambda_W(a). \tag{E.9}$$

The inverse-process relation of Glynn & Whitt (1994a, Theorem 1), applied to these bounded positive renewal intervals, gives

$$\lim_{u\to\infty} \frac{\log \mathbb{E}\left[e^{-s\mathcal{N}(u)}\right]}{u} = -\alpha(s). \tag{E.10}$$

Substituting $u = Mt$ therefore yields the potential-service scaled cumulant generating function

$$\kappa_M(-s) \triangleq \lim_{t\to\infty} \frac{\log \mathbb{E}\left[e^{-sD_M(t)}\right]}{t} = -M\alpha(s). \tag{E.11}$$

Net-input exponent and root equation. Define the cumulative net input in request units by

$$X_M(t) = \mathcal{A}(t) - D_M(t). \tag{E.12}$$

Arrival counts and request works are independent, so

$$\mathbb{E}\left[e^{sX_M(t)}\right] = \mathbb{E}\left[e^{s\mathcal{A}(t)}\right] \mathbb{E}\left[e^{-sD_M(t)}\right]. \tag{E.13}$$

Combining (E.5) and (E.11) gives

$$\Psi_M(s) \triangleq \lim_{t\to\infty} \frac{\log \mathbb{E}\left[e^{sX_M(t)}\right]}{t} = \Lambda_{\mathrm{D}}(s) - M\alpha(s). \tag{E.14}$$

The regular positive-root case in the corollary assumes a finite $\gamma_M > 0$ such that

$$\Psi_M(\gamma_M) = 0, \tag{E.15}$$

$$\Psi_M(s) < 0 \quad (0 < s < \gamma_M), \tag{E.16}$$

$$\Psi'_M(\gamma_M) > 0, \tag{E.17}$$

with Λ_{D} finite on $(\gamma_M - \varepsilon, \gamma_M + \varepsilon)$ for some $\varepsilon \in (0, \gamma_M)$. Differentiating (E.8) gives

$$\alpha'(s) = \frac{1}{\Lambda'_W(\alpha(s))}, \tag{E.18}$$

$$\alpha''(s) = -\frac{\Lambda''_W(\alpha(s))}{[\Lambda'_W(\alpha(s))]^3}, \tag{E.19}$$

$$\Psi''_M(s) = \Lambda''_{\mathrm{D}}(s) + \frac{M\Lambda''_W(\alpha(s))}{[\Lambda'_W(\alpha(s))]^3} \geq 0. \tag{E.20}$$

If either the arrival-count or request-work law is nondegenerate, the curvature is strictly positive, so the positive root is unique.

Substituting (E.14) into the root condition gives

$$\alpha(\gamma_M) = \frac{\Lambda_{\mathrm{D}}(\gamma_M)}{M}. \tag{E.21}$$

Using (E.8), we obtain

$$\begin{aligned} \gamma_M &= \Lambda_W\left(\frac{\Lambda_{\mathrm{D}}(\gamma_M)}{M}\right) \\ &= \log \mathbb{E}\left[\exp\left(\frac{\Lambda_{\mathrm{D}}(\gamma_M)W}{M}\right)\right]. \end{aligned} \tag{E.22}$$

Equivalently, the request-work expression in (C.9) gives

$$\gamma_M = \log \mathbb{E}\left[\exp\left(\frac{\Lambda_{\mathrm{D}}(\gamma_M)\left(\chi\omega + \frac{\omega(\omega-1)}{2}\right)}{M}\right)\right]. \tag{E.23}$$

Stationary tail and total population. The count $\mathcal{N}$ starts at a renewal epoch. Let $\mathcal{N}_{\mathrm{eq}}(u)$ count renewals in $(0, u]$ for an equilibrium version of the renewal process, independent of the arrivals. Let R_* be the work distance from the observation origin to its first subsequent renewal. Then $0 < R_* \leq w_{\max}$. Coupling the subsequent i.i.d. full work intervals with those of $\mathcal{N}$ gives

$$\mathcal{N}_{\mathrm{eq}}(u) = \begin{cases} 0, & 0 \leq u < R_*, \\ 1 + \mathcal{N}(u - R_*), & u \geq R_*. \end{cases} \tag{E.24}$$

A work interval of length at most $w_{\max}$ contains only a bounded number of renewals because each full interval is at least $w_{\min}$. Consequently, uniformly in $u \geq 0$,

$$|\mathcal{N}_{\mathrm{eq}}(u) - \mathcal{N}(u)| \leq c_W, \tag{E.25}$$

$$c_W = 2 + \left\lceil \frac{w_{\max}}{w_{\min}} \right\rceil. \tag{E.26}$$

Define

$$D_M^{\mathrm{eq}}(t) = \mathcal{N}_{\mathrm{eq}}(Mt), \tag{E.27}$$

$$X_M^{\mathrm{eq}}(t) = \mathcal{A}(t) - D_M^{\mathrm{eq}}(t). \tag{E.28}$$

The coupling implies $|X_M^{\mathrm{eq}}(t) - X_M(t)| \leq c_W$. Thus, for $s \geq 0$ with finite transforms,

$$e^{-sc_W}\mathbb{E}\left[e^{sX_M(t)}\right] \leq \mathbb{E}\left[e^{sX_M^{\mathrm{eq}}(t)}\right] \leq e^{sc_W}\mathbb{E}\left[e^{sX_M(t)}\right]. \tag{E.29}$$

Taking logarithms and dividing by t shows that X_M^{eq} has the same scaled cumulant Ψ_M. Its increments are stationary, since the renewal process is in equilibrium and the arrival sequence is i.i.d. and independent of that process.

Let $Q^{\mathrm{LD}}_{M,t}$ be the reflected count process, started from $Q^{\mathrm{LD}}_{M,0} = 0$,

$$Q^{\mathrm{LD}}_{M,t+1} = \left[Q^{\mathrm{LD}}_{M,t} + X^{\mathrm{eq}}_M(t+1) - X^{\mathrm{eq}}_M(t)\right]^+, \tag{E.30}$$

and let Q^{LD}_M denote its stationary waiting population in the memory-work representation. For every finite integer $t \geq 1$,

$$\mathbb{E}\left[e^{\gamma_M X^{\mathrm{eq}}_M(t)}\right] \leq \mathbb{E}\left[e^{\gamma_M \mathcal{A}(t)}\right] = e^{t\Lambda_{\mathrm{D}}(\gamma_M)} < \infty. \tag{E.31}$$

Together with the stationary increments, the cumulant limit (E.14), and the regular positive-root conditions (E.15)–(E.17), this verifies the hypotheses of Glynn & Whitt (1994b, Theorem 1). Therefore,

$$\lim_{k\to\infty} \frac{\log \Pr\{Q^{\mathrm{LD}}_M \geq k\}}{k} = -\gamma_M. \tag{E.32}$$

Finally, the total population in this representation is

$$K^{\mathrm{LD}}_M = Q^{\mathrm{LD}}_M + N^{(M)}. \tag{E.33}$$

For fixed M, the active-population bound (C.42) gives $0 \leq N^{(M)} \leq n_M$, where $n_M = \left\lfloor \frac{M}{\chi_{\min}} \right\rfloor$. The active population may be random and dependent on the waiting population; the pathwise bound nevertheless implies, for $k > n_M$,

$$\Pr\{Q^{\mathrm{LD}}_M \geq k\} \leq \Pr\{K^{\mathrm{LD}}_M \geq k\} \leq \Pr\{Q^{\mathrm{LD}}_M \geq k - n_M\}. \tag{E.34}$$

Since $\frac{k-n_M}{k} \to 1$, both bounding probabilities have the logarithmic decay rate in (E.32). Hence

$$\lim_{k\to\infty} \frac{\log \Pr\{K^{\mathrm{LD}}_M \geq k\}}{k} = -\gamma_M. \tag{E.35}$$

Combining this limit with (E.22) yields Corollary 1 under the saturated memory-work approximation.

F Proof of Theorem 4

We derive the residual-service QED approximation from the discrete-time concurrency-only queue. The residual-service construction adapts the many-server framework of Gamarnik & Momčilović (2008) to i.i.d. iteration-level batch arrivals.

Finite-system recursion. Let c be the number of concurrent service positions, with no memory admission restriction, and let

$$p_j = \Pr\{\omega = j\}, \tag{F.1}$$

$$\widetilde{p}_j = \Pr\{\omega \geq j\}, \tag{F.2}$$

$$d_\omega = \omega_{\max}, \tag{F.3}$$

$$\mu = \frac{1}{\mathbb{E}[\omega]}. \tag{F.4}$$

The service distribution is fixed as c varies, has finite support, and is independent across requests and independent of arrival counts.

Here t indexes decode iterations, and $A^{(c)}_t$ denotes the i.i.d. arrival count in the system with concurrency c. The system is observed immediately after admission and before the forward pass. Let $Q^{(c)}_t$ be the waiting population and let $L^{(c)}_{t,j}$ count active requests requiring exactly j further decode iterations. Then

$$K^{(c)}_t = Q^{(c)}_t + \sum_{j=1}^{d_\omega} L^{(c)}_{t,j}, \tag{F.5}$$

$$\sum_j L^{(c)}_{t,j} \leq c, \tag{F.6}$$

$$Q^{(c)}_t \left(c - \sum_j L^{(c)}_{t,j}\right) = 0. \tag{F.7}$$

Thus $K_t^{(c)}$ is the concurrency-only counterpart of $K_{\mathrm{D}}(t)$. During the next transition, the $L_{t,1}^{(c)}$ requests with one remaining iteration complete. The remaining active population is $\sum_{j=2}^{d_\omega} L_{t,j}^{(c)}$. With $A_{t+1}^{(c)}$ new arrivals, the number of requests admitted to service is exactly

$$J_{t+1}^{(c)} = \min\left\{ Q_t^{(c)} + A_{t+1}^{(c)}, c - \sum_{j=2}^{d_\omega} L_{t,j}^{(c)} \right\}, \tag{F.8}$$

$$Q_{t+1}^{(c)} = \left[Q_t^{(c)} + A_{t+1}^{(c)} + \sum_{j=2}^{d_\omega} L_{t,j}^{(c)} - c \right]^+. \tag{F.9}$$

Service requirements can be revealed when a request starts service, because the FCFS admission rule in this memory-free system does not inspect its unrevealed service requirement. Let $\mathcal{H}_{t+1}^{(c)}$ be the σ-field generated by $Q_s^{(c)}$ and $L_{s,j}^{(c)}$ for $0 \le s \le t$, $1 \le j \le d_\omega$, and by $A_s^{(c)}$ for $1 \le s \le t+1$. It represents the information available before revealing the service requirements of the newly admitted requests, and $J_{t+1}^{(c)}$ is $\mathcal{H}_{t+1}^{(c)}$-measurable. If $J_{t+1,j}^{(c)}$ counts new starts of length j, then conditional on $\mathcal{H}_{t+1}^{(c)}$,

$$\left(J_{t+1,1}^{(c)}, \ldots, J_{t+1,d_\omega}^{(c)}\right) \sim \mathrm{Multinomial}(J_{t+1}^{(c)}; p_1, \ldots, p_{d_\omega}). \tag{F.10}$$

Consequently,

$$L_{t+1,j}^{(c)} = L_{t,j+1}^{(c)} + J_{t+1,j}^{(c)}, \quad 1 \le j < d_\omega, \tag{F.11}$$

$$L_{t+1,d_\omega}^{(c)} = J_{t+1,d_\omega}^{(c)}. \tag{F.12}$$

In particular, summing the finite-system equations gives the pathwise identity

$$K_{t+1}^{(c)} = Q_t^{(c)} + A_{t+1}^{(c)} + \sum_{j=2}^{d_\omega} L_{t,j}^{(c)}. \tag{F.13}$$

Scaling and Gaussian innovations. Here $\to$ denotes limits of deterministic quantities, $\Rightarrow$ denotes convergence in distribution (weak convergence for probability measures), and $\xrightarrow{\mathrm{p}}$ denotes convergence in probability. Unless stated otherwise, limits are taken as $c \to \infty$. Let $\lambda_c = \mathbb{E}\left[A_t^{(c)}\right] = \lambda_{\mathrm{D}}(c)$ and define the offered active load and safety-capacity parameter by

$$\mathcal{R}_c = \lambda_c \mathbb{E}\left[\omega\right], \tag{F.14}$$

$$\rho_c = \frac{\mathcal{R}_c}{c}, \tag{F.15}$$

$$\beta_c = \frac{c - \mathcal{R}_c}{\sqrt{c}} \to \beta \in (0, \infty). \tag{F.16}$$

Thus $\frac{\lambda_c}{c} \to \mu$. In a stationary stable system, rate conservation gives $\mathbb{E}\left[J_t^{(c)}\right] = \lambda_c$; taking expectations in (F.12) then gives

$$\mathbb{E}\left[L_{t,j}^{(c)}\right] = \lambda_c \widetilde{p}_j. \tag{F.17}$$

This motivates the scaled state

$$\widehat{Q}_t^{(c)} = \frac{Q_t^{(c)}}{\sqrt{c}}, \tag{F.18}$$

$$\widehat{L}_{t,j}^{(c)} = \frac{L_{t,j}^{(c)} - \lambda_c \widetilde{p}_j}{\sqrt{c}}, \tag{F.19}$$

$$\widehat{A}_t^{(c)} = \frac{A_t^{(c)} - \lambda_c}{\sqrt{c}}. \tag{F.20}$$

Write $\widehat{\boldsymbol{L}}_t^{(c)} = (\widehat{L}_{t,1}^{(c)}, \ldots, \widehat{L}_{t,d_\omega}^{(c)})^\top$ and $\widehat{X}_t^{(c)} = (\widehat{Q}_t^{(c)}, \widehat{\boldsymbol{L}}_t^{(c)})$ for the complete scaled state. We assume i.i.d. arrival counts across iterations and the triangular-array limit

$$\widehat{A}_t^{(c)} \;\Rightarrow\; Z_t^A \sim N(0, \sigma_{\mathrm{D}}^2). \tag{F.21}$$

The limiting arrival innovations are i.i.d. across iterations.

To embed a given finite-system batch law A^{D} into a QED sequence, take independent sources

$$A_t^{[m]} = \sum_{r=1}^{m} A_{t,r}^{\mathrm{D}}, \tag{F.22}$$

$$c_m = m\lambda_{\mathrm{D}}\mathbb{E}\left[\omega\right] + \beta\sqrt{m\lambda_{\mathrm{D}}\mathbb{E}\left[\omega\right]} + O(1), \tag{F.23}$$

where $A_{t,r}^{\mathrm{D}}$ are independent copies of A^{D} and c_m is rounded to an integer. For $0 < \lambda_{\mathrm{D}} < \infty$ and $\mathrm{Var}(A^{\mathrm{D}}) < \infty$, the ordinary i.i.d. CLT gives, as $m \to \infty$,

$$\frac{A_t^{[m]} - m\lambda_{\mathrm{D}}}{\sqrt{c_m}} = \frac{\sqrt{\frac{m}{c_m}}\sum_{r=1}^{m}(A_{t,r}^{\mathrm{D}} - \lambda_{\mathrm{D}})}{\sqrt{m}} \Rightarrow N\left(0, \frac{\mathrm{Var}(A^{\mathrm{D}})}{\lambda_{\mathrm{D}}\mathbb{E}\left[\omega\right]}\right). \tag{F.24}$$

Therefore the manuscript's variance parameter is

$$\sigma_{\mathrm{D}}^2 = \frac{\mu\,\mathrm{Var}(A^{\mathrm{D}})}{\mathbb{E}\left[A^{\mathrm{D}}\right]}. \tag{F.25}$$

For deterministic batch sizes, this variance is zero.

Let $\boldsymbol{p} = (p_1, \ldots, p_{d_\omega})^\top$ and $\boldsymbol{J}_{t+1}^{(c)} = (J_{t+1,1}^{(c)}, \ldots, J_{t+1,d_\omega}^{(c)})^\top$. Define

$$\Sigma = \mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top, \tag{F.26}$$

$$\boldsymbol{Z}_{t+1}^{J,(c)} = \frac{\boldsymbol{J}_{t+1}^{(c)} - J_{t+1}^{(c)}\boldsymbol{p}}{\sqrt{c}}. \tag{F.27}$$

For states with tight scaled coordinates and a tight arrival innovation, (F.8) implies $J_{t+1}^{(c)} = \lambda_c + O_{\mathrm{Pr}}(\sqrt{c})$, and hence $\frac{J_{t+1}^{(c)}}{c} \xrightarrow{\mathrm{p}} \mu$. Here $O_{\mathrm{Pr}}(\sqrt{c})$ denotes fluctuations bounded in probability after division by $\sqrt{c}$.

The limiting composition noise follows directly from the conditional characteristic function. If a categorical vector E equals the jth coordinate vector with probability p_j, then for fixed $u \in \mathbb{R}^{d_\omega}$,

$$\mathbb{E}\left[\exp\left(\frac{iu^\top(E - \boldsymbol{p})}{\sqrt{c}}\right)\right] = 1 - \frac{u^\top\Sigma u}{2c} + O(c^{-\frac{3}{2}}). \tag{F.28}$$

With $J = J_{t+1}^{(c)} \le c$, independence of the J categorical draws conditional on $\mathcal{H}_{t+1}^{(c)}$ yields

$$\mathbb{E}\left[e^{iu^\top \boldsymbol{Z}_{t+1}^{J,(c)}} \,\middle|\, \mathcal{H}_{t+1}^{(c)}\right] = \exp\left(-\frac{Ju^\top\Sigma u}{2c} + O(c^{-\frac{1}{2}})\right) \xrightarrow{\mathrm{p}} \exp\left(-\frac{\mu u^\top\Sigma u}{2}\right). \tag{F.29}$$

The deterministic conditional limit and fresh service marks give innovations independent of the preceding state, arrival innovation, and previous composition innovations:

$$\boldsymbol{Z}_t^J \overset{\mathrm{iid}}{\sim} N(0, \mu\Sigma), \tag{F.30}$$

$$\{\boldsymbol{Z}_t^J\} \perp \{Z_t^A\}, \tag{F.31}$$

$$\mathbf{1}^\top \boldsymbol{Z}_t^J = 0 \quad \text{almost surely.} \tag{F.32}$$

The zero sum reflects preservation of the total number of new starts.

Limiting recursion. Define the shift operator $\mathcal{T}(\ell_1, \ldots, \ell_{d_\omega}) = (\ell_2, \ldots, \ell_{d_\omega}, 0)$ and $\widehat{J}_{t+1}^{(c)} = \frac{J_{t+1}^{(c)} - \lambda_c}{\sqrt{c}}$. Using $\sum_{j=2}^{d_\omega}\widetilde{p}_j = \mathbb{E}\left[\omega\right] - 1$, the number of available positions in (F.8) becomes

$$c - \sum_{j=2}^{d_\omega} L_{t,j}^{(c)} = \lambda_c + \sqrt{c}\left(\beta_c - \sum_{j=2}^{d_\omega}\widehat{L}_{t,j}^{(c)}\right). \tag{F.33}$$

Consequently, the exact scaled equations are

$$\widehat{J}_{t+1}^{(c)} = \min\left\{\widehat{Q}_t^{(c)} + \widehat{A}_{t+1}^{(c)}, \beta_c - \sum_{j=2}^{d_\omega} \widehat{L}_{t,j}^{(c)}\right\}, \tag{F.34}$$

$$\widehat{Q}_{t+1}^{(c)} = \left[\widehat{Q}_t^{(c)} + \widehat{A}_{t+1}^{(c)} + \sum_{j=2}^{d_\omega} \widehat{L}_{t,j}^{(c)} - \beta_c\right]^+, \tag{F.35}$$

$$\widehat{\boldsymbol{L}}_{t+1}^{(c)} = \mathcal{T}\widehat{\boldsymbol{L}}_t^{(c)} + \widehat{J}_{t+1}^{(c)}\boldsymbol{p} + \boldsymbol{Z}_{t+1}^{J,(c)}. \tag{F.36}$$

The last equation uses $\widetilde{p}_j = \widetilde{p}_{j+1} + p_j$, with $\widetilde{p}_{d_\omega+1} = 0$, to cancel the deterministic centering terms.

Combining (F.21) and (F.29) with the continuity of the minimum, positive-part, and shift maps gives the limiting recursion

$$\widehat{J}_{t+1} = \min\left\{\widehat{Q}_t + Z_{t+1}^A, \beta - \sum_{j=2}^{d_\omega} \widehat{L}_{t,j}\right\}, \tag{F.37}$$

$$\widehat{Q}_{t+1} = \left[\widehat{Q}_t + Z_{t+1}^A + \sum_{j=2}^{d_\omega} \widehat{L}_{t,j} - \beta\right]^+, \tag{F.38}$$

$$\widehat{\boldsymbol{L}}_{t+1} = \mathcal{T}\widehat{\boldsymbol{L}}_t + \widehat{J}_{t+1}\boldsymbol{p} + \boldsymbol{Z}_{t+1}^J. \tag{F.39}$$

For convergent initial scaled states, induction using the conditional CLT and the continuous mapping theorem establishes convergence over any fixed number of transitions to this recursion.

The limiting state space is

$$\mathcal{E}_\beta = \left\{(q, \boldsymbol{\ell}) : q \geq 0,\ \mathbf{1}^\top \boldsymbol{\ell} \leq \beta,\ q(\beta - \mathbf{1}^\top \boldsymbol{\ell}) = 0\right\}. \tag{F.40}$$

Indeed, the sum of the new residual coordinates is $\min\{\sum_{j\geq 2} \ell_j + q + Z^A, \beta\}$, while the next queue is the positive part of the same unconstrained quantity minus β. This verifies preservation of the capacity and non-idling constraints.

Stationary limit. Let ν_c be the stationary law of $\widehat{X}_t^{(c)}$. For the stationary QED passage, assume that $\{\nu_c\}$ is tight and the limiting recursion has a unique invariant probability measure π_β. To identify any subsequential limit $\nu_c \Rightarrow \nu$, initialize each finite queue in stationarity, independently of its future primitive variables. If F_β denotes the limiting update and Z its independent Gaussian innovation, the one-step limit gives

$$(\widehat{X}_0^{(c)}, \widehat{X}_1^{(c)}) \Rightarrow (X, F_\beta(X, Z)), \tag{F.41}$$

$$X \sim \nu. \tag{F.42}$$

Both finite-system marginals have law ν_c, hence both limiting marginals have law ν. For every bounded continuous function f,

$$\int f(x)\nu(dx) = \int \mathbb{E}\left[f(F_\beta(x, Z))\right]\nu(dx). \tag{F.43}$$

Thus ν is invariant. The scaled capacity constraints place its support in $\mathcal{E}_\beta$, and uniqueness implies $\nu = \pi_\beta$. Every subsequence has the same limit, so $\nu_c \Rightarrow \pi_\beta$.

Population projection. Define the centered total population in the limiting chain by

$$\widehat{Y}_t = \widehat{Q}_t + \sum_{j=1}^{d_\omega} \widehat{L}_{t,j}. \tag{F.44}$$

Since the service-composition noise sums to zero and $\widehat{Q}_{t+1} + \widehat{J}_{t+1} = \widehat{Q}_t + Z_{t+1}^A$, summing the limit equations gives

$$\widehat{Y}_{t+1} = \widehat{Q}_t + \sum_{j=2}^{d_\omega} \widehat{L}_{t,j} + Z_{t+1}^A. \tag{F.45}$$

Conditional on an invariant state $x = (q, \ell_1, \ldots, \ell_{d_\omega}) \sim \pi_\beta$, the next scaled total population is Gaussian with mean $q + \sum_{j=2}^{d_\omega} \ell_j$ and variance σ_{D}^2.

For a target finite system, substitute $\mathcal{R}_c = \lambda_{\mathrm{D}} \mathbb{E}[\omega]$ and define

$$m_c(x) = \mathcal{R}_c + \sqrt{c} \left(q + \sum_{j=2}^{d_\omega} \ell_j \right), \tag{F.46}$$

$$\tau_c = \sqrt{c\sigma_{\mathrm{D}}^2}. \tag{F.47}$$

Assigning the real-valued QED population to the integer cell $\left(k - \frac{1}{2}, k + \frac{1}{2}\right]$ gives the raw cell weight

$$w_c(k) = \int \left[\Phi\left(\frac{k + \frac{1}{2} - m_c(x)}{\tau_c} \right) - \Phi\left(\frac{k - \frac{1}{2} - m_c(x)}{\tau_c} \right) \right] \pi_\beta(dx), \qquad k \in \mathbb{Z}, \tag{F.48}$$

when $\sigma_{\mathrm{D}}^2 > 0$. Since π_β is invariant, these are also the stationary Gaussian-cell weights.

When $\sigma_{\mathrm{D}}^2 = 0$, the conditional distribution is a point mass and division by τ_c must be avoided:

$$w_c(k) = \int \mathbb{I}_{\{k - \frac{1}{2} < m_c(x) \le k + \frac{1}{2}\}} \pi_\beta(dx). \tag{F.49}$$

The half-open interval fixes the rounding convention at any atoms.

For $k \ge 0$ and $\tau_c > 0$, the resulting finite-c approximation is

$$\widehat{\pi}_k^{(c)} = \int \left[\Phi\left(\frac{k + \frac{1}{2} - m_c(x)}{\tau_c} \right) - \Phi\left(\frac{k - \frac{1}{2} - m_c(x)}{\tau_c} \right) \right] \pi_\beta(dx). \tag{F.50}$$

Together with (F.46)–(F.47) and (F.25), this is the formula in Theorem 4. When $\tau_c = 0$, use the indicator in (F.49) instead.

For numerical evaluation, the weights can be normalized on the nonnegative integers. To distinguish this operation from the theorem's cell formula, write

$$\mathcal{Z}_c = \sum_{k \ge 0} w_c(k), \tag{F.51}$$

$$\widetilde{\pi}_k^{(c)} = \frac{w_c(k)}{\mathcal{Z}_c}, \quad k \ge 0. \tag{F.52}$$

For $\tau_c > 0$, telescoping the Gaussian CDF differences yields

$$\mathcal{Z}_c = \int \Phi\left(\frac{m_c(x) + \frac{1}{2}}{\tau_c} \right) \pi_\beta(dx). \tag{F.53}$$

For $\tau_c = 0$, the integrand is $\mathbb{I}_{\{m_c(x) > -\frac{1}{2}\}}$. Since $\frac{\mathcal{R}_c}{\sqrt{c}} \to \infty$ and the limiting population fluctuation is finite almost surely, $\mathcal{Z}_c \to 1$. For a finite target configuration, evaluate the limiting recursion at its specified β_c, using π_{β_c} in the population formulas above.

G Supplementary Numerical Results

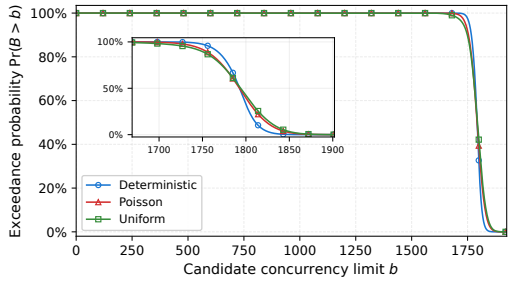
G.1 Resource Provisioning and Coupled Constraints

The experiments in this subsection simulate decode directly, with one step corresponding to a decode iteration of duration $\Delta_{\mathrm{D}} = 20\,\mathrm{ms}$. Unless specified otherwise, $\chi \sim \mathrm{Unif}\{64, \ldots, 192\}$ and $\omega \sim \mathrm{Unif}\{128, \ldots, 384\}$, independently. Memory budgets below are effective KV-cache admission capacities.

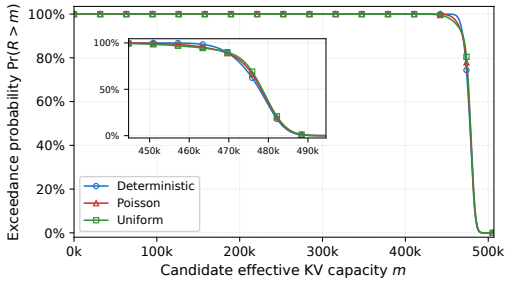
Resource redundancy. Figures G.1a and G.1b show the strict exceedance probabilities of the active batch size in the memory-only reference system with $M = 479{,}250$, and of KV occupancy in the concurrency-only reference system with $c = 1{,}800$, respectively. Each panel compares deterministic arrivals of seven requests per iteration, $\mathrm{Poisson}(7)$ arrivals, and discrete uniform arrivals on $\{0, \ldots, 14\}$. The rapid decline of these tails identifies the resource ranges in which an additional constraint becomes unlikely to bind. Their crossings of a tolerance θ determine $B_\theta(M)$ and $M_\theta(c)$.

Load-sensitive provisioning. Figure G.1c fixes $(M, B) = (479{,}250, 1{,}800)$ and varies the Poisson mean per iteration over $\{0, 1, 2, 2.25, 2.5, \ldots, 7\}$. The three curves use prompt supports $\{64, \ldots, 192\}$, $\{128, \ldots, 384\}$, and $\{256, \ldots, 768\}$, with the same output distribution. The plotted quantity is the fraction of measured iterations in which new requests remain waiting after admission. The transition to persistent waiting moves to lower arrival rates as prompts become longer. Each point uses 2×10^5 warm-up iterations followed by 10^7 measurement iterations.

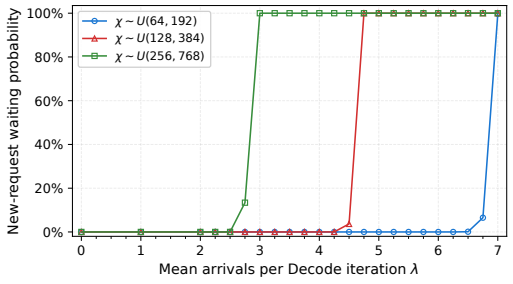
Joint resource constraints. Figure G.1d compares cumulative completion rates under saturated demand. The memory-only, concurrency-only, and joint configurations use $(M, B) = (479{,}250, 10{,}000)$, $(1{,}080{,}000, 1{,}800)$, and $(479{,}250, 1{,}800)$, respectively. Each rate is the number of post-warm-up completions divided by the number of elapsed decode iterations. The two single-resource estimates are above seven requests per iteration, whereas the joint estimate is slightly below seven. This separation illustrates the non-composability of the two isolated resource margins discussed in Section 6.



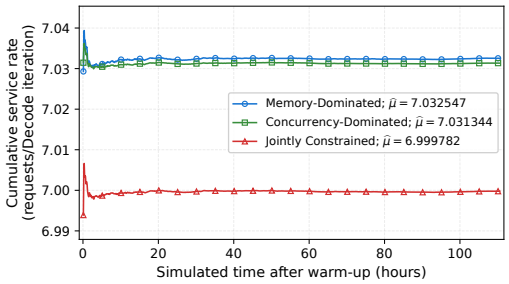
(a) Concurrency redundancy, $M = 479{,}250$.



(b) Memory redundancy, $c = 1{,}800$.

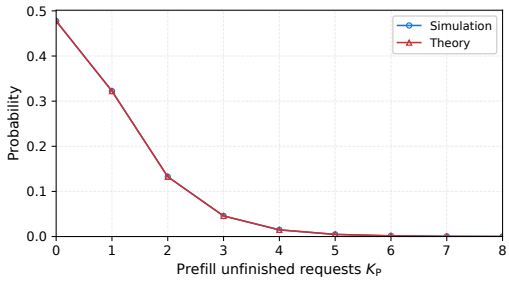


(c) Load sensitivity under joint constraints.

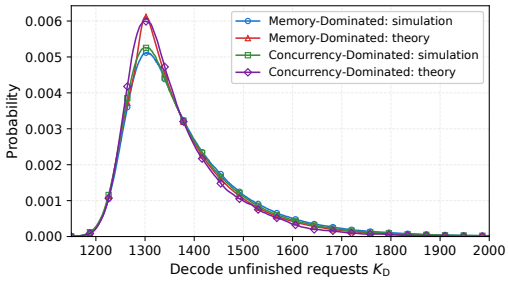


(d) Saturated completion rates.

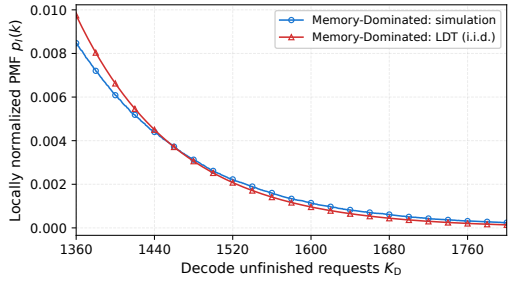
Figure G.1: Numerical illustrations of resource provisioning. The first row shows the two redundancy transitions, with insets enlarging their transition regions. The second row shows load-sensitive waiting and the completion-rate gap under joint constraints. In (a), the plotted B is the reference active batch size; in (d), the slot unit is a decode iteration and $b_{\max}$ denotes the concurrency cap B.



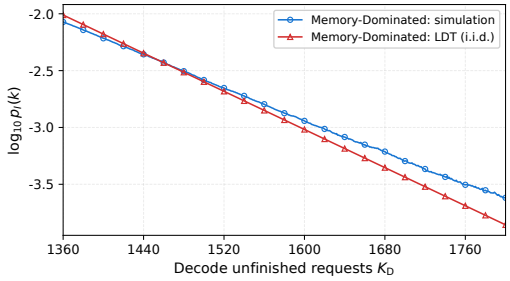
(a) prefill stationary PMF.



(b) decode stationary PMFs.



(c) Locally normalized tail PMF.



(d) Locally normalized tail PMF (log scale).

Figure G.2: Supplementary end-to-end validation with $\chi, \omega \sim \mathrm{Unif}\{128, \ldots, 384\}$. The decode PMFs are displayed over $[1150, 2000]$; the tail is locally normalized over $\mathcal{I} = \{1360, \ldots, 1800\}$, shown on a linear scale in (c) and on a base-10 logarithmic scale in (d). The tail curves use the prefill-derived marginal arrival-count distribution as i.i.d. decode input with mean five, giving $\gamma_M \simeq 0.00965924$.

G.2 Additional Queue Distribution Experiments

All experiments below use discrete time and independent, integer-valued request marks. Uniform distributions include both endpoints. Full-distribution PMFs retain their global normalization. For a displayed tail window $\mathcal{I}$, the LDT comparisons use the local normalization

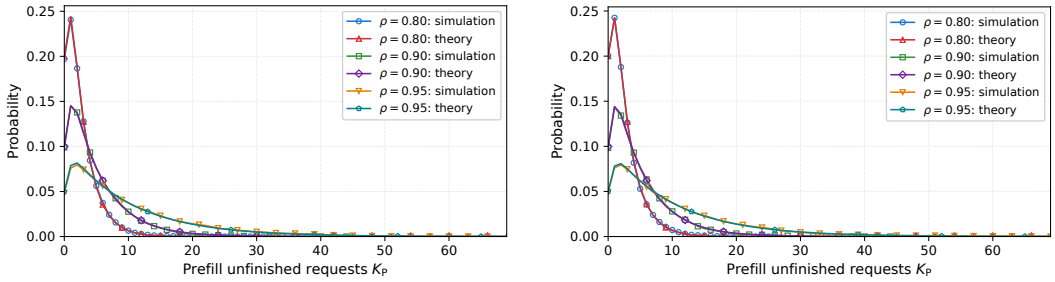
$$p_{\mathcal{I}}(k) = \frac{\Pr\{K_{\mathrm{D}} = k\}}{\sum_{\ell \in \mathcal{I}} \Pr\{K_{\mathrm{D}} = \ell\}}, \tag{G.1}$$

$$p_{\mathcal{I}}^{\mathrm{LD}}(k) = \frac{e^{-\gamma_M k}}{\sum_{\ell \in \mathcal{I}} e^{-\gamma_M \ell}}, \qquad k \in \mathcal{I}. \tag{G.2}$$

The simulated PMF supplies the empirical probabilities in the first expression. The second uses the model-based exponent, so its base-10 logarithmic slope is $-\frac{\gamma_M}{\log 10}$.

End-to-end validation with longer prompts. We set $\chi, \omega \sim \mathrm{Unif}\{128, \ldots, 384\}$ and use external Poisson arrivals at 250 requests per second, or 0.125 requests per base slot. The mean decode input is five requests per iteration. The memory-dominated configuration uses $(M, B) = (507{,}150, 5{,}000)$; the concurrency-dominated configuration uses $(M, B) = (1{,}080{,}000, 1{,}286)$. The remaining timing parameters follow Section 7.1. For each configuration, ten independent runs each contain 4×10^8 base slots, of which the first 4×10^7 are discarded. Figure G.2 compares the shared prefill PMF, the two decode PMFs, and the memory-dominated tail approximation. The PMF approximations capture the main concentration and decay patterns under this changed prompt distribution.

Standalone prefill under different loads. Figure G.3 compares two prompt distributions at $\rho_{\mathrm{P}} \in \{0.80, 0.90, 0.95\}$. These runs use $a_{\mathrm{P}} = \frac{0.16\,\mathrm{ms}}{\mathrm{token}}$, $c_{\mathrm{P}} = 45\,\mathrm{ms}$, and $\delta = 0.5\,\mathrm{ms}$, with $S = \left\lceil \frac{a_{\mathrm{P}}\chi + c_{\mathrm{P}}}{\delta} \right\rceil$. The Poisson mean per slot is selected as $\lambda_{\mathrm{P}} = \frac{\rho_{\mathrm{P}}}{\mathbb{E}[S]}$. Each run discards 2×10^6 warm-up slots and measures 5×10^7 subsequent slots. The theoretical curves track the broadening of the population distribution as utilization increases for both prompt supports.



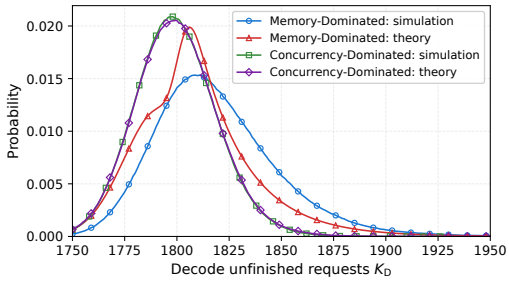
(a) $\chi \sim \mathrm{Unif}\{64, \ldots, 192\}$.

(b) $\chi \sim \mathrm{Unif}\{128, \ldots, 384\}$.

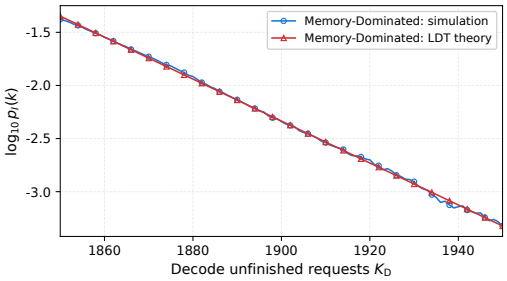
Figure G.3: Standalone prefill stationary PMFs under Poisson arrivals. Each panel compares simulation and theory at three utilizations; the legend parameter ρ denotes ρ_{P}, not the per-slot arrival mean.

Standalone decode under different arrival laws. We use $\chi \sim \mathrm{Unif}\{64, \ldots, 192\}$ and $\omega \sim \mathrm{Unif}\{128, \ldots, 384\}$, with direct decode arrivals given by $A^{\mathrm{D}} = 7$, $A^{\mathrm{D}} \sim \mathrm{Poisson}(7)$, or $A^{\mathrm{D}} \sim \mathrm{Unif}\{0, \ldots, 14\}$. The memory-dominated and concurrency-dominated configurations are $(M, B) = (479{,}250, 10{,}000)$ and $(M, B) = (1{,}080{,}000, 1{,}800)$, respectively. The deterministic case uses 10^7 iterations including 2×10^5 warm-up iterations; each stochastic case uses 10^8 iterations including 10^6 warm-up iterations. The left column of Figure G.4 compares the two PMF approximations with simulation. The largest bulk-distribution discrepancy occurs in the deterministic memory-dominated case; the concurrency approximation and the Poisson and uniform comparisons are closer.

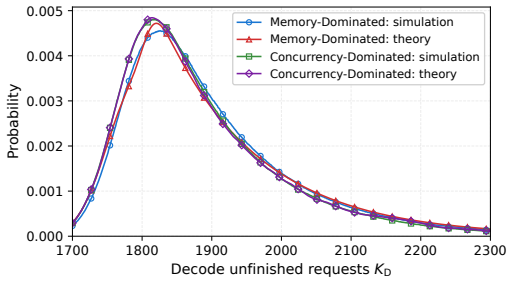
Memory-dominated tail decay. The right column of Figure G.4 uses the same memory-dominated simulation samples to examine the logarithmic tails. For deterministic, Poisson, and uniform arrivals, respectively, the windows are $\{1850, \ldots, 1950\}$, $\{1900, \ldots, 2300\}$, and $\{2100, \ldots, 3000\}$. The corresponding theoretical exponents are approximately 0.0454754, 0.00744109, and 0.00311735. The approximately linear empirical profiles follow the predicted slopes, with slower tail decay under the more variable arrival laws.



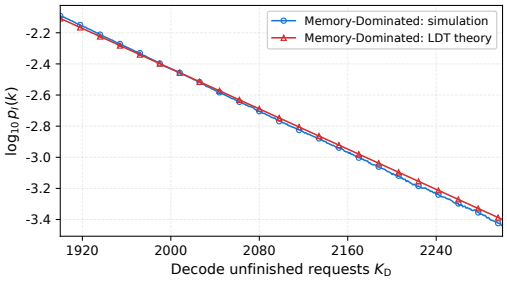
(a) Deterministic arrivals: stationary PMFs.



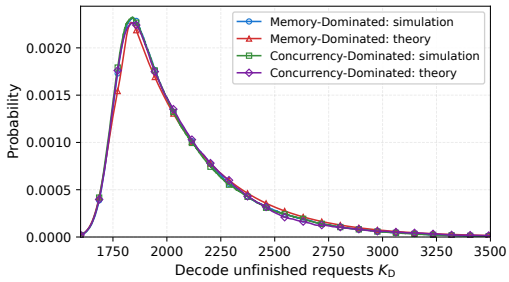
(b) Deterministic arrivals: tail PMF (log scale).



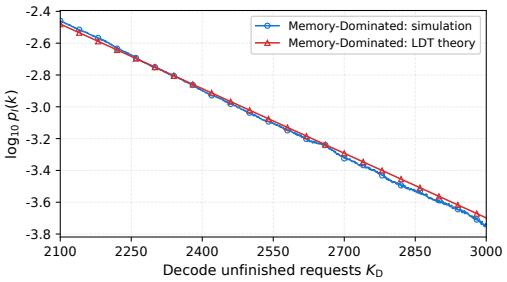
(c) Poisson arrivals: stationary PMFs.



(d) Poisson arrivals: tail PMF (log scale).



(e) Uniform arrivals: stationary PMFs.



(f) Uniform arrivals: tail PMF (log scale).

Figure G.4: Standalone decode validation at mean input seven requests per iteration. Rows correspond to deterministic, Poisson, and discrete uniform arrivals. Left: globally normalized PMFs in the memory- and concurrency-dominated regimes. Right: locally normalized memory-dominated PMFs on a base-10 logarithmic scale, compared with the LDT approximation.

H JUSTIFICATION OF THEORETICAL MODELS

H.1 CONSTANT DECODE BATCH-PROCESSING TIME

In practical LLM serving systems, the duration of an individual decode iteration may vary due to differences in active batch composition, request states, and hardware execution effects. In this work, we approximate the iteration duration by a fixed value Δ_{D} to obtain a tractable queueing model. This approximation is motivated by the execution characteristics of continuous batching and the long-run perspective of queueing analysis.

Under continuous batching, the scheduler continuously admits waiting requests until either the KV-cache or concurrency constraint becomes active. Therefore, in high-utilization regimes, the active batch typically remains close to the effective resource boundary. As a result, the computational and memory-access workload of successive decode iterations varies within a relatively limited range. Moreover, P–D disaggregation isolates decode from prefill execution, avoiding the additional execution-time variability caused by mixing the two heterogeneous phases within the same batch.

From a queueing perspective, the stationary population and congestion behavior are primarily governed by the long-run service capacity rather than individual iteration-level fluctuations. If the iteration duration contains bounded random perturbations around its mean value, these fluctuations average out over many iterations, and the empirical service rate concentrates around its characteristic value. Therefore, the constant-duration abstraction captures the dominant service-rate behavior relevant to stationary queueing analysis. Similar characteristic iteration-time abstractions have also been adopted in prior queueing analyses of LLM inference (Nie et al., 2026; Dai et al., 2025).

H.2 INDEPENDENT PER-ITERATION DECODE ARRIVAL COUNTS

In the exact P–D tandem system, decode arrivals are generated by the departure process of the upstream prefill queue. Let D_t^{P} denote the number of prefill departures during slot t. With $\Delta_{\mathrm{D}} = d\delta$ and aligned iteration boundaries, the actual inter-stage arrival count is

$$\widetilde{A}_j^{\mathrm{D}} = \sum_{\ell=0}^{d-1} D_{jd+\ell}^{\mathrm{P}}.$$

These counts may exhibit temporal dependence across iterations. In principle, this dependence can be retained in a joint prefill–decode Markov chain, whose stationary law is characterized by its balance equations. However, direct computation becomes prohibitively expensive because of the request-level phase information required by decode.

To obtain a tractable stage-wise analysis, we approximate the inter-stage process by i.i.d. per-iteration arrival counts A_j^{D} with the same stationary marginal:

$$\Pr\left\{A_j^{\mathrm{D}} = a\right\} = q_a^{\mathrm{D}} = \Pr\left\{\widetilde{A}_j^{\mathrm{D}} = a\right\}.$$

Here, q_a^{D} is obtained from the prefill departure-count interface in Appendix B.2. The approximation preserves the full one-iteration count distribution while neglecting temporal dependence across iterations.

Several considerations motivate this approximation. First, a decode iteration aggregates departures over multiple prefill slots, which smooths slot-level variability through temporal aggregation. Moreover, in practical deployments with multiple prefill instances, the decode input is the superposition of multiple departure streams; such aggregation provides an additional smoothing effect. These mechanisms motivate the use of an independent arrival-count surrogate for stage-wise stationary analysis.

Most directly, the end-to-end simulations in Section 7 retain the original prefill-generated arrival stream rather than resampling independent arrival counts. The agreement between theoretical and simulated decode population distributions provides empirical support for the approximation under the evaluated configurations.