MULTI-REPACKING IN ASYNCHRONOUS RL POST-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-tail trajectory generation leaves rollout replicas underutilized in asynchronous reinforcement learning post-training. We propose multi-repacking, which jointly redistributes residual trajectories across multiple replicas to release workers earlier. Under exponential service, linear cache growth, and service-preserving migration, we derive finite-system bounds and explicit hierarchy sums for occupied replica time. Under a joint large-system scaling regime and an explicit packing-slack condition, we prove that multi-repacking has a strictly smaller leading-order expected tail occupancy than hierarchical pairwise repacking, with a uniform asymptotic gap. The analysis accounts for dependent anchor ages, asynchronous starts, adaptive packing, and natural releases. With 1,600 asynchronously started replicas, each processing 12,800 trajectories in 3,200 slots, all three merge ratios $m = 3, 4, 5$ outperform hierarchical pairwise in mean tail occupied replica time on both evaluated workloads. Across 160 event-driven runs using paired draws, reductions are 4.74–7.59% under exponential service and 6.34–10.56% under resampled supplied output lengths; all six paired mean improvements have positive pointwise 95% confidence intervals. The length-based runs use the same policy parameters, reserve physical cache headroom, and exhibit no overflow. These finite-system results demonstrate multi-repacking's occupied-time benefit on both exponential and nonexponential workloads.

1 INTRODUCTION

Reinforcement learning (RL) has become a central post-training paradigm for large language models (LLMs), supporting advances in mathematical reasoning, code generation, and agentic decision-making (OpenAI, 2024; Guo et al., 2025; Sheng et al., 2025a). A standard RL post-training pipeline alternates between two stages: *generation*, where rollout workers produce trajectories by responding to prompts or interacting with environments, and *training*, where the collected trajectories are used to update model weights. As rollout workloads scale, efficient trajectory generation becomes increasingly important.

A major bottleneck is the long-tail distribution of trajectory lengths. In reasoning workloads, generation has been reported to account for up to 83.1% of total execution time (Sheng et al., 2025a). Most trajectories finish relatively quickly, while a small fraction continue much longer. During this tail phase, many rollout replicas remain lightly occupied, leaving GPU resources underutilized and limiting the benefit of adding more replicas.

Asynchronous RL systems alleviate this bottleneck by decoupling generation from training and overlapping their execution. Laminar introduces a trajectory-level asynchronous architecture in which individual trajectories can be generated and consumed independently (Sheng et al., 2025a). Its dynamic *repacking* mechanism migrates residual trajectories from underutilized replicas onto fewer workers, allowing released workers to fetch updated weights and begin fresh batches. Repacking therefore improves resource utilization by releasing replicas before all trajectories of a weight version have finished.

Despite their empirical effectiveness, repacking decisions based on preset thresholds and greedy packing rules remain largely heuristic. Quantifying their benefits requires understanding how random trajectory completions and continued KV-cache growth jointly determine consolidation opportunities.

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.

This motivates three basic questions: when repacking becomes feasible, how residual trajectories should be packed, and how much occupied replica time can be saved.

We study trajectory repacking in asynchronous RL post-training through a stochastic model of the draining tail within a single weight version. Under exponential service, linear KV-cache growth, and service-preserving migration, our primary performance metric is expected cumulative tail occupied replica time, which sums occupied time across replicas. Guided by our first-passage analysis, we propose a *multi-repacking* policy that jointly redistributes residual trajectories across multiple replicas, allowing consolidation to become feasible earlier than pairwise merging under suitable packing conditions. The policy first repacks whole blocks, keeping each block's replicas together in every operation. Once the required numbers of rounds have been completed, the remaining replicas can regroup separately with other eligible replicas at the same level. We compare its performance with a hierarchical pairwise policy inspired by Laminar. Our main contributions are:

- A stochastic model and first-passage analysis: we model trajectory completion and KV-cache evolution within a single weight version, derive explicit Lambert W expressions for mean-load threshold-crossing times, and establish concentration and expectation-error bounds for the corresponding random first-passage times, accounting for dependent anchor ages.
- A multi-repacking framework with a strict theoretical advantage: we provide a sufficient packing condition and derive finite-system occupied-replica-time bounds and explicit hierarchy sums for pairwise and multi-repacking, using no repacking as a common baseline. The analysis accounts for adaptive packing, asynchronous starts, and natural releases. Under explicit scaling assumptions and the packing-slack conditions, multi-repacking has a strictly smaller leading term in expected tail occupied replica time than hierarchical pairwise repacking, with a uniform positive gap between the exact hierarchy sums (Proposition 4).
- Consistent finite-system improvements over hierarchical pairwise under independent uniform replica starts: all three tested merge ratios reduce mean tail occupied replica time on both evaluated workloads, by 4.74–7.59% under exponential service and 6.34–10.56% under resampled output lengths, with paired confidence intervals for the mean improvements lying strictly above zero.

2 Related Work

RL training systems. Asynchronous RL systems decouple generation from training to overlap their execution and improve resource utilization. Laminar (Sheng et al., 2025a) combines trajectory-level asynchrony with relay-based weight synchronization and dynamic repacking of residual trajectories. StreamRL, AReaL, AsyncFlow, LlamaRL, and AcceRL also improve resource utilization by allowing training and trajectory generation to run concurrently (Zhong et al., 2025; Fu et al., 2025; Han et al., 2025; Wu et al., 2025; Lu et al., 2026).

Synchronous frameworks such as HybridFlow and OpenRLHF (Sheng et al., 2025b; Hu et al., 2024) and pipeline methods such as GPipe and PipeDream-2BW (Huang et al., 2019; Narayanan et al., 2021a) improve training efficiency, complementing our consolidation of unfinished trajectories during the generation tail.

Llumnix (Sun et al., 2024) implements live request and state migration across serving instances for load balancing, providing migration mechanisms relevant to cross-replica repacking. Distribution-aware speculative decoding (Shao et al., 2025) addresses long-tail generation by accelerating decoding itself, whereas our repacking model preserves trajectory completion times and reduces resource occupancy by consolidating residual trajectories.

Model-based analysis of LLM systems. MELL (Liu et al., 2025) models growing KV caches and request migration, providing competitive guarantees for peak GPU count and bounds on migration frequency. Related scheduling studies analyze inference latency under KV-cache constraints (Jaillet et al., 2025), fluid models for cache-aware admission and scheduling (Ao et al., 2025), throughput-optimal scheduling (Li et al., 2025), and non-clairvoyant completion-time guarantees (Feng et al., 2026). These works analyze provisioning, latency, and throughput. We study expected cumulative occupied replica time during a finite batch's draining tail, using stochastic load crossings and hierarchy sums to compare pairwise and multi-repacking.

Dynamic and stochastic bin packing. Mellou et al. (2024) study migration tradeoffs for total active bin time, an objective closely related to occupied replica time. Their item sizes are fixed and known on arrival, although departure times are unknown. In our setting, current cache loads are observable, but each live trajectory's cache continues to grow until random completion. Packing feasibility therefore depends on both cache growth and the space released by completed trajectories. Hong et al. (2023) allow Markovian, time-varying item sizes and study steady-state active-server minimization under a resource-contention constraint, with jobs remaining on their assigned servers after dispatch. Our analysis instead concerns the transient draining phase of a finite batch, where replicas enter the tail at random asynchronous times and unfinished trajectories can be migrated repeatedly.

3 System Model

We consider one weight version on R replicas, indexed by $\mathcal{R} = [R] := \{1, \dots, R\}$. Replica i starts at prescribed time t_i and processes N trajectories with B parallel slots, immediately replacing completed trajectories while its queue is nonempty. Service durations are independent $\mathrm{Exp}(\lambda)$ variables with rate $\lambda > 0$. The start times are deterministic or sampled independently of service. Theoretical probabilities and expectations are conditional on the start vector; Appendix B.1.2 explains averaging over random starts. Migration preserves trajectory ages and completion times and has zero modeled cost.

KV cache grows at unit rate with trajectory age. Let $C_i(t)$ be the current cache load of physical replica i and $Q_i(t)$ its queued trajectory count. Fix $0 < C_{\mathrm{th}} \le C_{\max}$, where $C_{\max}$ is physical capacity; each accepted output has load at most C_{th}. Since exponential service durations are unbounded and cache grows with age, a fixed capacity reserve cannot guarantee the absence of overflow on every realization. Choosing $C_{\mathrm{th}} < C_{\max}$ leaves room for subsequent growth and reduces overflow risk. The analysis models service without capacity-induced stalling or offloading; capacity limitations are discussed in Section 6. Appendix B.1.1 gives the slot-level dynamics.

Objective. The anchor s_i, defined precisely in Section 4.1, marks replica i's entry into the tail. Only anchored replicas can repack, and a released physical label is never reactivated for the current version. Let $t^{\mathcal{M}}_{i,\mathrm{free}}$ be its release time under mechanism $\mathcal{M}$, through packing or natural completion, and define

$$\tau_i^{\mathcal{M}} := t^{\mathcal{M}}_{i,\mathrm{free}} - s_i, \qquad T^{\mathcal{M}}_{\mathrm{tail}} := \sum_{i \in \mathcal{R}} \tau_i^{\mathcal{M}}.$$

Write $T^{\mathcal{M}}_{\mathrm{total}}$ for total occupied replica time and use superscript no for no repacking. The common pre-anchor execution gives

$$\begin{aligned} T^{\mathcal{M}}_{\mathrm{total}} &= \sum_{i \in \mathcal{R}} (s_i - t_i) + T^{\mathcal{M}}_{\mathrm{tail}}, \\ \Delta^{\mathcal{M}} &:= T^{\mathrm{no}}_{\mathrm{total}} - T^{\mathcal{M}}_{\mathrm{total}} = T^{\mathrm{no}}_{\mathrm{tail}} - T^{\mathcal{M}}_{\mathrm{tail}}. \end{aligned} \tag{1}$$

Thus minimizing $\mathbb{E}[T^{\mathcal{M}}_{\mathrm{tail}}]/R$ maximizes the mean saving per original replica. This is a resource-occupancy metric, not the completion time of the final trajectory.

4 Repacking Framework, Anchor Time, and Occupied-Time Analysis

We first define the anchor-based tail processes and the repacking rules, then relate random consolidation times to deterministic load crossings. The analysis starts with common anchors and extends to asynchronous execution to characterize expected occupied replica time. All proofs are in the appendix.

4.1 Anchor Time and Tail Dynamics

After an original replica's anchor, no new trajectories are dispatched from its queue; only those already running remain to be completed. This gives every policy the same starting state for that

replica's tail. Assume $N \geq B \geq 2$ and put $k := N - B + 1$. Replica i's anchor s_i is its kth completion, immediately after which its queue is empty and exactly $B-1$ trajectories remain, almost surely. Its mean delay is $\mathbb{E}[s_i - t_i] = k/(B\lambda)$. Index these unfinished trajectories by $j \in [B-1]$ independently of future service. Their ages $\tilde{A}_{ij}$ record how long they have already run at s_i, and their residual lifetimes $\tilde{Y}_{ij}$ are their remaining times to completion. The ages can be dependent; the residuals are independent $\mathrm{Exp}(\lambda)$ variables, independent of the collection of originals' own pre-anchor histories. At their respective anchors, the trajectories still running on the original replicas have the same expected age. This expectation and the finite-batch error are

$$\epsilon_N := \left(1 - \frac{1}{B}\right)^{N-B}, \qquad \mu_A := \mathbb{E}[\tilde{A}_{ij}] = \frac{1 - (1 - 1/B)\epsilon_N}{\lambda}, \qquad 0 \leq \lambda^{-1} - \mu_A \leq \frac{\epsilon_N}{\lambda}. \quad (2)$$

The anchor-delay and age calculations are in Appendices B.1.2 and B.1.3. For local time $u = t - s_i \geq 0$, define

$$Z_i(u) := \sum_{j=1}^{B-1} (\tilde{A}_{ij} + u)\mathbf{1}_{\{\tilde{Y}_{ij} > u\}}, \qquad N_i(u) := \sum_{j=1}^{B-1} \mathbf{1}_{\{\tilde{Y}_{ij} > u\}}. \quad (3)$$

After migration, Z_i and N_i still track the trajectories that were on replica i at its anchor, regardless of where they now run. The mean processes are

$$\mathbb{E}[Z_i(u)] = (B-1)e^{-\lambda u}(u + \mu_A), \qquad \mathbb{E}[N_i(u)] = (B-1)e^{-\lambda u}. \quad (4)$$

They determine load crossings and slot feasibility, respectively (Appendix B.1.4). Migration preserves the final completion time $\max_{i,j}(s_i + \tilde{Y}_{ij})$ while changing replica occupancy.

For the synchronous comparison, we shift each replica's tail process so that its anchor occurs at time zero, keeping its anchor ages and residual lifetimes unchanged and preserving their joint law. Asynchronous results retain the actual anchors. We use T_{span} as a deterministic control bound for their spread; its definition and estimates are given in Appendix B.1.2.

4.2 Hierarchical Repacking Mechanisms

We describe two hierarchical repacking mechanisms for reducing the number of active replicas while preserving the unfinished trajectories. The hierarchical pairwise baseline repeatedly merges two same-level replicas into one higher-level replica, whereas multi-repacking jointly redistributes trajectories across several output replicas.

At a decision time t, write $C(r) := C_r(t)$, $Q(r) := Q_r(t)$, $N_{\mathrm{act}}(r)$, and $\ell(r)$ for replica r's cache load, queued count, active count, and hierarchy level. Levels start at zero. Only anchored, unreleased, nonempty replicas can repack. At each completion timestamp, process tied completions and queue dispatches, admit new anchors, and release every started replica with $Q(r) = N_{\mathrm{act}}(r) = 0$.

Hierarchical pairwise baseline. A source and destination must have the same level. Their combined cache load, namely the sum of their two current cache loads, must be at most C_{th}, and their combined active count, namely the total number of running trajectories on the two replicas, must be at most B. Within each level, visit sources in increasing load order and choose the largest-load feasible destination still available. Remove both from the current candidate set, so the resulting pairs are disjoint. Execute the plan by migrating the unfinished trajectories from each source to its paired destination. Release each source after migration and promote each destination by one level. Rebuild the candidates and repeat until no pair is feasible. A level-ℓ replica thus carries the unfinished trajectories from 2^ℓ original replicas. Algorithm 2 gives the complete procedure.

Feasible multi-packing. Multi-repacking uses a multi-replica packing step to redistribute unfinished trajectories across a prescribed number of output replicas. The following sufficient condition guarantees that a collection of trajectories can be redistributed while respecting both the slot and cache constraints.

Let $c_\tau(t)$ be trajectory τ's current cache size and $\mathcal{T}(r)$ the trajectories on replica r. Fix a packing-slack parameter $\alpha \in (0, 1)$. For an integer $p \geq 1$ and a finite trajectory set $\mathcal{T}$, the tests

$$|\mathcal{T}| \leq pB, \qquad c_\tau(t) \leq \alpha C_{\mathrm{th}} \ \ (\tau \in \mathcal{T}), \qquad \sum_{\tau \in \mathcal{T}} c_\tau(t) \leq p(1-\alpha)C_{\mathrm{th}} \quad (5)$$

guarantee that these trajectories can be redistributed among p replicas, each with at most B trajectories and load at most C_{th} (Lemma 9). The construction sorts trajectories by decreasing size and assigns them cyclically to the outputs. The slack absorbs load imbalance.

Multi-repacking. Fix an integer $m \geq 3$. Initially, each original replica forms a singleton block. A block is a logical group of original replicas whose unfinished trajectories are repacked together. The physical replicas currently holding a block's unfinished trajectories are its representatives. Thus, the hierarchy level records the number of original replicas represented by a block, while the representatives describe the current physical replicas used to execute its remaining trajectories. They participate in whole-block repacking together.

After the switch, in independent mode, each nonempty representative can regroup separately with other eligible representatives at the same level. During the whole-block phase, a block is marked terminal and stops participating in repacking if a representative empties naturally or a packing output is empty. Its remaining representatives continue running and may regroup after the switch.

A nonterminal level-ℓ block contains m^ℓ originals represented by $(m-1)^\ell$ physical replicas. A whole-block round combines m eligible same-level blocks and applies (5) with $p = (m-1)^{\ell+1}$. It releases the other $(m-1)^\ell$ representatives and promotes the union block. This phase follows the whole-block reference in Algorithm 3, which never switches to separate regrouping and uses minimum-load slot-feasible groups.

For $R \geq m$, set $K_m := \lfloor \log_m R \rfloor$ and choose a fixed switch depth $d \in \{0, \ldots, K_m\}$, by default $d = \lceil K_m/2 \rceil$. Let J_ℓ count completed whole-block rounds at level ℓ. For $d > 0$, switching requires every level below d to complete its prescribed number of whole-block rounds, as specified in Algorithm 1. At the switch, the remaining nonempty representatives retain their levels; future anchors enter at level zero. Each independent round combines m eligible same-level representatives using (5) with $p = m - 1$, releases the unretained label and empty outputs, and promotes each nonempty output independently, without a level cap.

Algorithm 1 Multi-Repacking with a Certified Switch

Require: Replicas $\mathcal{R}$, $B, C_{\mathrm{th}}, m, \alpha$, and fixed $d \in \{0, \ldots, K_m\}$
1: Initialize level-zero singleton blocks, counters $J_\ell = 0$, and whole-block mode
2: **for** each trajectory-completion timestamp t **do**
3: Process tied completions and dispatches; admit new anchors in the current mode; apply natural releases and, in whole-block mode, terminal rules
4: **if** in whole-block mode **then**
5: **if** $d > 0$ **then**
6: Complete the whole-block packing step of Algorithm 3; update J_ℓ
7: After exhausting feasible whole-block repacking at this timestamp, switch if $J_\ell = \lfloor R/m^{\ell+1} \rfloor$ for all $0 \leq \ell < d$
8: **else**
9: Switch if the first anchor has occurred, before packing
10: If switching, make all anchored, unreleased, nonempty representatives independent at their inherited levels, including nonempty representatives of terminal blocks
11: **if** in independent mode **then**
12: **while** a feasible independent group exists **do**
13: At the smallest feasible level, select a minimum-load slot-feasible group satisfying (5) with $p = m - 1$
14: Pack cyclically into $m - 1$ input representatives; release the unretained label and empty outputs
15: Promote each nonempty output independently

Algorithm 4 expands the event and admission rules.

Scaling assumptions. For the asymptotic conclusions, let $B \to \infty$ with λ fixed. Fix constants $c_-, \eta > 0$ with $c_- \leq 1 - \eta$, and assume

$$0 < c_- \leq \frac{\lambda C_{\mathrm{th}}}{B} \leq 1 - \eta, \tag{6}$$

$$N = \omega\big(B\log(RB)\big), \qquad N = o\left(\frac{B^2}{\log(RB)}\right). \tag{7}$$

Let $T_0 := \max_i t_i - \min_i t_i$ be the startup spread and assume

$$\log R = \Theta(\log B), \qquad \lambda T_0 = o(\log B). \tag{8}$$

For multi-repacking, keep $m \geq 3$ and $\alpha \in (0,1)$ fixed. We call these sufficient conditions the *stochastic asymptotic regime*.

The scale B/λ is the mean aggregate cache load in the stationary reference process with B busy slots. Condition (6) keeps the threshold comparable to, but uniformly below, this reference load. The lower bound on N makes finite-batch approximation errors negligible across all replicas, while the upper bound makes the random contribution to the anchor spread negligible on the $1/\lambda$ time scale. The condition $\log R = \Theta(\log B)$ provides polynomially many replicas for the hierarchy comparison, and $\lambda T_0 = o(\log B)$ allows asynchronous starts whose spread is small relative to the no-repacking tail scale $\log B/\lambda$. For example, $R = B$, $N = \lceil B^{3/2} \rceil$, and $T_0 = O(1/\lambda)$ satisfy these size and startup requirements. Further details are given in Appendix B.1.5. Finite-system bounds and experiments use their own stated hypotheses.

Constants $C > 0$ may change between estimates and depend only on fixed η and, for multi-repacking, m, α; simplified proportional-threshold rates may also depend on c_-. Dependence on λ is explicit.

4.3 Occupied-Time Analysis for Pairwise Repacking

No-repacking baseline. Without repacking, a replica remains occupied until its last anchored trajectory completes. Independent residual lifetimes give

$$\tau_i^{\mathrm{no}} = \max_{j \in [B-1]} \tilde{Y}_{ij}, \qquad \mathbb{E}[\tau_i^{\mathrm{no}}] = \frac{H_{B-1}}{\lambda}, \qquad H_n := \sum_{j=1}^{n} \frac{1}{j}. \tag{9}$$

Random first passage and mean-load crossing. In the pairwise hierarchy, a level-ℓ replica carries unfinished trajectories from 2^ℓ originals. To determine when consolidation becomes possible, we therefore study when the aggregate load of such a group falls below a packing threshold. A useful reference is $C_{\mathrm{th}}/2$: two inputs below this threshold satisfy the combined cache constraint. We approximate the random time at which this happens by the time at which the expected aggregate load reaches the same threshold.

Fix a collection $\mathcal{I}_\ell$ of 2^ℓ original replicas before observing their service paths, and set $n_\ell := 2^\ell(B-1)$ and $Z_\ell(u) := \sum_{i \in \mathcal{I}_\ell} Z_i(u)$ in local tail time. Define the random first-passage time by

$$\sigma_\ell := \inf\{u \geq 0 : Z_\ell(u) \leq C_{\mathrm{th}}/2\}. \tag{10}$$

The corresponding deterministic crossing $\bar{\sigma}_\ell$ is the decreasing-branch solution of

$$n_\ell e^{-\lambda \bar{\sigma}_\ell}(\mu_A + \bar{\sigma}_\ell) = C_{\mathrm{th}}/2. \tag{11}$$

The concentration estimates below quantify the accuracy of this approximation; estimates of the number of unfinished trajectories account for the slot constraint.

Lambert W representation. The Lambert function satisfies $W(z)e^{W(z)} = z$. Its lower real branch $W_{-1}(z) \leq -1$ for $z \in [-1/e, 0)$ selects the decreasing-branch crossing. Define

$$\rho_\ell := \frac{\lambda C_{\mathrm{th}}}{2^{\ell+1}(B-1)}, \qquad \Phi(\rho) := -W_{-1}(-\rho/e) - 1, \quad 0 < \rho \leq 1. \tag{12}$$

Under a fixed margin $\rho_\ell \leq 1 - \eta$ and sufficiently large N/B, Lemma 10 gives

$$\bar{\sigma}_\ell = -\lambda^{-1} W_{-1}(-\rho_\ell e^{-\lambda \mu_A}) - \mu_A, \qquad \left|\bar{\sigma}_\ell - \lambda^{-1}\Phi(\rho_\ell)\right| \leq C\epsilon_N/\lambda. \tag{13}$$

The function Φ therefore expresses the approximate crossing time directly through the model parameters. Define the positive crossing slope

$$\kappa_\ell := \frac{C_{\mathrm{th}}}{2}\left(\lambda - \frac{1}{\mu_A + \bar{\sigma}_\ell}\right). \tag{14}$$

A steeper mean-load crossing produces a smaller timing error for the same load fluctuation.

Lemma 1 (First-passage concentration and expectation error)**.** *For the fixed analytical block above, assume $\rho_\ell \le 1-\eta$ for fixed $\eta > 0$ and sufficiently large N/B. There is a constant $C > 0$, depending only on η, such that, for $x \ge 1$ and $L_\ell(x) := x + \log(2n_\ell)$, whenever $n_\ell e^{-\lambda\bar{\sigma}_\ell} \ge CL_\ell(x)$,*

$$\mathbb{P}\left(|\sigma_\ell - \bar{\sigma}_\ell| > \frac{C(\bar{\sigma}_\ell + \lambda^{-1})}{\kappa_\ell}\left[\sqrt{n_\ell e^{-\lambda\bar{\sigma}_\ell} L_\ell(x)} + L_\ell(x)\right] + \frac{C\epsilon_N}{\lambda}\right) \le 4e^{-x} + n_\ell\epsilon_N. \tag{15}$$

Without this lower bound on the expected number of unfinished trajectories at the mean-load crossing,

$$|\mathbb{E}[\sigma_\ell] - \bar{\sigma}_\ell| \le \frac{C\sqrt{n_\ell}\log(2n_\ell)}{\lambda\kappa_\ell} + \frac{Cn_\ell\log(2n_\ell)}{\lambda}\epsilon_N. \tag{16}$$

The concentration estimate controls the first crossing over the entire interval $u \ge 0$, and the expectation estimate also covers levels with few remaining trajectories. The proof and the simultaneous path estimates used for adaptively formed groups are given in Appendix B.3.3.

From common anchors to asynchronous execution. With common anchors, Lemma 1 relates random first-passage times to mean-load crossings. Its simultaneous path estimates, together with bounds on the number of unfinished trajectories, control the feasible merges formed by the algorithm. Weighting these crossing times by the numbers of replicas released gives a finite hierarchy expression for expected occupied time; blocks left without a partner at their level and natural completions contribute to the remainder (Appendix B.4.1). For actual anchors, T_{span} bounds their spread and the waiting needed to match replicas at the same level. These contributions are lower order under the stochastic asymptotic regime. Converting finite release counts to their limiting weights yields the following asynchronous formula. The finite-system bounds and asymptotic analysis are in Appendices B.4.3, B.4.4, and B.6.2.

Theorem 2 (Tail occupied replica time under asynchronous pairwise repacking)**.** *Under the stochastic asymptotic regime of the Scaling assumptions paragraph, the event-driven hierarchical pairwise rule satisfies*

$$\mathbb{E}[T_{\mathrm{tail}}^{\mathrm{pair}}] = \frac{R}{\lambda}\sum_{\ell=0}^{\infty}\frac{1}{2^{\ell+1}}\Phi\left(\frac{\lambda C_{\mathrm{th}}}{2^{\ell+1}(B-1)}\right) + o(R/\lambda). \tag{17}$$

The factor $2^{-\ell-1}$ is the limiting fraction of physical replicas released at level ℓ.

4.4 Occupied-Time Analysis for Multi-Repacking

Multi-repacking pools trajectories across several outputs, allowing consolidation at higher mean loads. Its occupied-time benefit depends on both the timing of these rounds and the number of replicas they release. We first analyze the whole-block reference and then transfer its leading term to the proposed policy. Write $T_{\mathrm{tail}}^{\mathrm{blk}}(m)$ for the reference occupied time and $T_{\mathrm{tail}}^{\mathrm{multi}}(m,d)$ for the proposed policy with switch depth d. The default-depth quantity is abbreviated $T_{\mathrm{tail}}^{\mathrm{multi}}$.

Block crossings. A level-ℓ whole-block round combines trajectories from $m^{\ell+1}$ originals and packs them into $(m-1)^{\ell+1}$ representatives. The total number of trajectories unfinished at these originals' respective anchors and the aggregate packing threshold are

$$\begin{aligned} n_{\ell,m} &:= m^{\ell+1}(B-1), \qquad \theta_{\ell,m} := (m-1)^{\ell+1}(1-\alpha)C_{\mathrm{th}}, \\ \rho_{\ell,m} &:= \frac{\lambda\theta_{\ell,m}}{n_{\ell,m}} = \left(\frac{m-1}{m}\right)^{\ell+1}\frac{(1-\alpha)\lambda C_{\mathrm{th}}}{B-1}. \end{aligned} \tag{18}$$

For $0 \le \ell < K_m$, fix a collection of $m^{\ell+1}$ originals before observing service and let $Z_{\ell,m}(u)$ be its aggregate local tail load. As in the pairwise analysis, define the random first passage by $\sigma_{\ell,m} := \inf\{u \ge 0 : Z_{\ell,m}(u) \le \theta_{\ell,m}\}$ and the deterministic mean-load crossing by

$$n_{\ell,m}e^{-\lambda\bar{\sigma}_{\ell,m}}(\mu_A + \bar{\sigma}_{\ell,m}) = \theta_{\ell,m}. \tag{19}$$

The corresponding crossing slope is

$$\kappa_{\ell,m} := \theta_{\ell,m}\left(\lambda - \frac{1}{\mu_A + \bar{\sigma}_{\ell,m}}\right) > 0. \tag{20}$$

The Lambert representation (13) and the concentration and expectation estimates of Lemma 1 apply with the corresponding m-indexed quantities. The precise statements are given in Lemmas 11 and 13. The deterministic parameters and crossings are also defined by these formulas for levels beyond the finite hierarchy.

From common anchors to asynchronous regrouping. Applying the first-passage analysis above with the multi-packing tests gives the synchronous whole-block formula and finite-system bound (Appendix B.5.1). Under the stochastic asymptotic regime, the round-span estimates preserve its leading term for actual anchors (Appendix B.5.3). The proposed policy follows this reference until the required whole-block round counts are reached, then allows representatives to regroup separately. At the default switch depth, this change affects expected occupied time by only $o(R/\lambda)$. The switch comparison and conversion to the series below are given in Appendices B.6.1 and B.6.2.

Theorem 3 (Tail occupied replica time of multi-repacking)**.** *Under the stochastic asymptotic regime of the Scaling assumptions paragraph, Algorithm 1 with the default switch depth satisfies*

$$\mathbb{E}[T_{\mathrm{tail}}^{\mathrm{multi}}] = \frac{R}{\lambda}\sum_{\ell=0}^{\infty}\frac{(m-1)^{\ell}}{m^{\ell+1}}\Phi\left(\left(\frac{m-1}{m}\right)^{\ell+1}\frac{(1-\alpha)\lambda C_{\mathrm{th}}}{B-1}\right) + o(R/\lambda). \tag{21}$$

The weight $(m-1)^{\ell}/m^{\ell+1}$ is the limiting fraction of replicas released at level ℓ, and the argument of Φ is the normalized packing threshold. Their combined effect yields the following comparison.

Proposition 4 (Strict improvement over pairwise repacking)**.** *Under the stochastic asymptotic regime of the Scaling assumptions paragraph, fix $m \geq 3$ and suppose*

$$0 < \alpha < 1 - \frac{1}{4}\left(\frac{m}{m-1}\right)^{m}. \tag{22}$$

Then the default multi-repacking policy satisfies

$$\liminf_{B\to\infty}\frac{\lambda}{R}\left(\mathbb{E}[T_{\mathrm{tail}}^{\mathrm{pair}}] - \mathbb{E}[T_{\mathrm{tail}}^{\mathrm{multi}}]\right) \geq \log\left[4(1-\alpha)\left(\frac{m-1}{m}\right)^{m}\right] > 0. \tag{23}$$

Multi-repacking has a strictly smaller occupied-time leading term and a strictly larger saving per original replica. The choice $\alpha = 0.05$ satisfies (22) for $m = 3, 4, 5$. The comparison is proved in Appendix B.7.1.

4.5 Tail Savings and Whole-Batch Interpretation

By (1), the mean saving per replica is $H_{B-1}/\lambda - \mathbb{E}[T_{\mathrm{tail}}^{\mathcal{M}}]/R$; the same finite-system errors apply to the corresponding saving estimates. Under the stochastic asymptotic regime, for pairwise or default multi-repacking,

$$\frac{\mathbb{E}[T_{\mathrm{tail}}^{\mathrm{no}}]}{R} = \frac{H_{B-1}}{\lambda} = \Theta\left(\frac{\log B}{\lambda}\right), \qquad \frac{\mathbb{E}[T_{\mathrm{tail}}^{\mathcal{M}}]}{R} = \Theta\left(\frac{1}{\lambda}\right). \tag{24}$$

The saving is $(\log B + O(1))/\lambda$ per replica, and the saved tail fraction is $1 - \Theta(1/\log B)$. These common orders coexist with the strict leading-term gap in Proposition 4. The whole-batch baseline is $\mathbb{E}[T_{\mathrm{total}}^{\mathrm{no}}]/R = [(N-B+1)/B + H_{B-1}]/\lambda$; the saved whole-batch fraction is $(1+o(1))B\log B/N = o(1)$. Appendix B.7.2 gives the derivation, and Appendix C develops closed-form approximations to the hierarchy sums. The metric quantifies earlier resource release within one version; end-to-end throughput requires a multi-version runtime analysis.

5 Experimental Evaluation

Multi-repacking reduces mean tail occupied replica time on both workloads for every tested merge ratio. Here multi refers to Algorithm 1, except for the unrestricted variant below.

Configuration and workloads. Both settings use $R = 1600$, $B = 3200$, $N = 12800$, $\alpha = 0.05$, and $C_{\text{th}} = 0.95B = 3040$. Time and cache age are expressed in mean-service-time units. In every experiment, replica start times satisfy $t_i \overset{\text{iid}}{\sim} \text{Unif}[0, 1]$, independently of service durations. The startup interval spans one mean service time, corresponding to $[0, 1/\lambda]$ under exponential service. Physical capacity is $C_{\max} = 1.5C_{\text{th}} = 4560$ to accommodate later cache growth; this headroom changes neither the packing threshold nor policy decisions. The simulator preserves trajectory ages and completion times and models zero migration cost. The hierarchical comparison exhausts feasible packing after completions.

Setting I uses independent $\text{Exp}(1)$ service and generates actual finite-batch anchors, dependent anchor ages, and independent residuals; each anchor includes its replica's sampled startup offset. Setting II uses 15,806 positive output lengths from a single software-engineering RL workload: repeated SWE-agent rollouts per task using an internal 198B-parameter Step-3.7-Flash checkpoint on tasks similar to SWE-bench Multilingual. We resample these lengths with replacement after normalization by the dataset mean 460.5780716. Each sampled length is a fixed service duration. Work-conserving FIFO dispatch generates anchors and residuals without an exponential approximation, retaining the supplied marginal length distribution. Both settings use identical system and packing parameters. For Setting II, cache peaks are checked before anchors and throughout the tail, at the left limits before completions and migrations and immediately after packing. No replica exceeds $C_{\max}$ in these runs.

Paired protocol. We compare hierarchical pairwise with multi $m = 3, 4, 5$, using respective depths $d = 3, 3, 2$ from $d = \lceil K_m/2 \rceil$ in both workloads. Each setting has 20 independent draws of startup times and service durations. Within each draw, all four policies share the same start vector and service realizations, giving 160 runs for the hierarchical comparison. The measured cost is tail occupied replica time per original replica, including every natural and packing release. Figures 1a and 1b in Appendix D show sample means with pointwise 95% Student-t intervals across draws. Improvements are calculated relative to the pairwise sample mean; their intervals use within-draw paired differences.

Improvement under exponential service. Pairwise has mean cost 2.6816. Multi $m = 3, 4, 5$ reduce it to 2.5545, 2.4781, and 2.4894, respectively, corresponding to reductions of 4.74%, 7.59%, and 7.17% (Figure 1a). Thus all three ratios improve, with $m = 4$ attaining the lowest mean. The multi policies have 18, 26, and 45 natural releases per run, respectively, only 1.13%, 1.63%, and 2.81% of the original replicas; their full occupied time remains included in the comparison.

Improvement on resampled trajectory lengths. With the same parameters, the pairwise mean is 14.2391, while multi $m = 3, 4, 5$ achieve 13.3358, 12.8014, and 12.7358, respectively. The reductions are 6.34%, 10.10%, and 10.56% (Figure 1b), with $m = 5$ giving the lowest mean. All six paired mean improvements across the two workloads have pointwise 95% confidence intervals strictly above zero.

The size $N/B = 4$ does not verify the batch-size asymptotics, and the exponential theorem is not asserted for the supplied-length law. The measured improvement is in occupied replica time; the fixed trajectories' makespan is unchanged.

Unrestricted implementation. An implementation variant removes same-level matching and whole-block membership while retaining cache and slot feasibility. For both tested workloads with the same independent uniform starts, all three merge ratios of unrestricted multi reduce mean tail occupied replica time per original replica relative to Laminar-style best-fit and an event-driven best-fit control. Appendix D.2 reports the protocol, figures, relative reductions, and paired confidence intervals.

6 LIMITATIONS

Later cache growth may exceed capacity, while migration costs are omitted and may reduce the savings. Length resampling ignores request dependence and input-token effects; token lengths are not measured GPU service times. Empirical conclusions concern occupied replica time under the tested configurations; end-to-end RL throughput remains unassessed.

AI USE STATEMENT

Generative AI tools were used to assist with editing and polishing the manuscript, including improving clarity, reducing repetition, and reorganizing the presentation of proofs. The authors take responsibility for the final content of the paper.

REFERENCES

Ruicheng Ao, Gan Luo, David Simchi-Levi, and Xinshang Wang. Optimizing LLM inference: Fluid-guided online scheduling with memory constraints. *arXiv preprint arXiv:2504.11320*, 2025. URL https://arxiv.org/abs/2504.11320v4. Version 4, revised June 2026.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Volodymir Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. Impala: Scalable distributed deep rl with importance weighted actor-learner architectures. *arXiv preprint arXiv:1802.01561*, 2018.

Björn Feldkord, Matthias Feldotto, Anupam Gupta, Guru Guruganesh, Amit Kumar, Sören Riechers, and David Wajc. Fully-dynamic bin packing with little repacking. In *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*, volume 107 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 51:1–51:24. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2018. doi: 10.4230/LIPIcs.ICALP.2018.51. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ICALP.2018.51.

Yiding Feng, Siyu Liu, Zonghan Yang, and Yuhao Zhang. General non-clairvoyant KV-cache scheduling via regime-aware routing. *arXiv preprint arXiv:2607.09248*, 2026. URL https://arxiv.org/abs/2607.09248v1.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, Tongkai Yang, Binhang Yuan, and Yi Wu. Areal: A large-scale asynchronous reinforcement learning system for language reasoning. *arXiv preprint arXiv:2505.24298*, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Zhenyu Han, Ansheng You, Haibo Wang, Kui Luo, Guang Yang, Wenqi Shi, Menglong Chen, Sicheng Zhang, Zeshun Lan, Chunshi Deng, Huazhong Ji, Wenjie Liu, Yu Huang, Yixiang Zhang, Chenyi Pan, Jing Wang, Xin Huang, Chunsheng Li, and Jianping Wu. Asyncflow: An asynchronous streaming rl framework for efficient llm post-training. *arXiv preprint arXiv:2507.01663*, 2025.

Yige Hong, Qiaomin Xie, and Weina Wang. Near-optimal stochastic bin-packing in large service systems with time-varying item sizes. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(3):48:1–48:46, 2023. doi: 10.1145/3626779. URL https://arxiv.org/abs/2209.04123v4.

Jian Hu, Xibin Wu, Zilin Zhu, Xianyu, Weixun Wang, Dehao Zhang, and Yu Cao. Openrlhf: An easy-to-use, scalable and high-performance rlhf framework. *arXiv preprint arXiv:2405.11143*, 2024.

Yanping Huang, Youlong Cheng, Ankur Bapna, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 103–112, 2019.

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. URL https://arxiv.org/abs/2502.07115v5. Version 5, revised January 2026.

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, 2023.

Yueying Li, Jim Dai, and Tianyi Peng. Throughput-optimal scheduling algorithms for llm inference and ai agents. *arXiv preprint arXiv:2504.07347*, 2025.

Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldberg, Joseph Gonzalez, Michael Jordan, and Ion Stoica. Rllib: Abstractions for distributed reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML '18)*, pp. 3053–3062, 2018.

Qianli Liu, Zicong Hong, Peng Li, Fahao Chen, and Song Guo. Mell: Memory-efficient large language model serving via multi-GPU KV cache management. *arXiv preprint arXiv:2501.06709*, 2025. URL https://arxiv.org/abs/2501.06709.

C. Lu, S. Wang, Y. Li, W. Liu, S. Jin, F. Qian, P. Li, B. Sun, and Y. Liu. Accerl: A distributed asynchronous reinforcement learning and world model framework for vision-language-action models. *arXiv preprint arXiv:2603.18464v2*, 2026.

Zhiyu Mei, Wei Fu, Guang Wang, Huanchen Zhang, and Yi Wu. Real: Efficient rlhf training of large language models with parameter reallocation. *arXiv preprint arXiv:2406.14088*, 2024.

Konstantina Mellou, Marco Molinaro, and Rudy Zhou. The power of migrations in dynamic bin packing. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 8(3): 45:1–45:28, 2024. doi: 10.1145/3700435. URL https://arxiv.org/abs/2408.13178.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. *arXiv preprint arXiv:1602.01783*, 2016.

Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I. Jordan, et al. Ray: A distributed framework for emerging ai applications. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI '18)*, pp. 561–577, 2018.

Deepak Narayanan, Aaron Harlap, Amar Phanishayee, et al. Pipedream: Generalized pipeline parallelism for dnn training. In *Proceedings of the 27th ACM Symposium on Operating Systems Principles (SOSP)*, pp. 1–15, 2019.

Deepak Narayanan, Amar Phanishayee, Kaiyu Shi, et al. Memory-efficient pipeline-parallel dnn training. In *International Conference on Machine Learning (ICML)*, pp. 7937–7947, 2021a.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, et al. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC21)*, 2021b.

OpenAI. Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/, 2024.

Sheldon M. Ross. *Introduction to Probability Models*. Academic Press, 11th edition, 2014. See Section 5.3 on order statistics of exponential random variables.

René L. Schilling, Renming Song, and Zoran Vondraček. *Bernstein Functions: Theory and Applications*, volume 37 of *De Gruyter Studies in Mathematics*. De Gruyter, Berlin, 2nd edition, 2012.

Zelei Shao, Vikranth Srivatsa, Sanjana Srivastava, Qingyang Wu, Alpay Ariyak, Xiaoxia Wu, Ameen Patel, Jue Wang, Percy Liang, Tri Dao, Ce Zhang, Yiying Zhang, Ben Athiwaratkun, Chenfeng Xu, and Junxiong Wang. Beat the long tail: Distribution-aware speculative decoding for rl training. *arXiv preprint arXiv:2511.13841*, 2025.

Guangming Sheng, Yuxuan Tong, Borui Wan, Wang Zhang, Chaobo Jia, Xibin Wu, Yuqi Wu, Xiang Li, Chi Zhang, Yanghua Peng, Haibin Lin, Xin Liu, and Chuan Wu. Laminar: A scalable asynchronous rl post-training framework. *arXiv preprint arXiv:2510.12633*, 2025a.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems (EuroSys '25)*, pp. 1279–1297, 2025b.

Biao Sun, Ziming Huang, Hanyu Zhao, Wencong Xiao, Xinyi Zhang, Yong Li, and Wei Lin. Llumnix: Dynamic scheduling for large language model serving. In *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pp. 173–191. USENIX Association, 2024. URL https://www.usenix.org/conference/osdi24/presentation/sun-biao.

Bo Wu, Sid Wang, Yunhao Tang, Jia Ding, Eryk Helenowski, Liang Tan, Tengyu Xu, Tushar Gowda, Zhengxing Chen, Chen Zhu, Xiaocheng Tang, Yundi Qian, Beibei Zhu, and Rui Hou. Llamarl: A distributed asynchronous reinforcement learning framework for efficient large-scale llm training. *arXiv preprint arXiv:2505.24034*, 2025.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Jeff Huang, Chuyue Sun, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Efficiently programming large language models using sglang. *arXiv preprint arXiv:2312.07104*, 2023.

Yinnin Zhong, Zili Zhang, Bingyang Wu, Shengyu Liu, Yukun Chen, Changyi Wan, Hanpeng Hu, Lei Xia, Ranchen Ming, Yibo Zhu, et al. Rlhfuse: Efficient rlhf training for large language models with inter- and intra-stage fusion. *arXiv preprint arXiv:2409.13221*, 2024.

Yinnin Zhong, Zili Zhang, Xiaoniu Song, Hanpeng Hu, Chao Jin, Bingyang Wu, Nuo Chen, Yukun Chen, Yu Zhou, Changyi Wan, Hongyu Zhou, Yimin Jiang, Yibo Zhu, and Daxin Jiang. Streamrl: Scalable, heterogeneous, and elastic rl for llms with disaggregated stream generation. *arXiv preprint arXiv:2504.15930*, 2025.

A Additional Related Work

RL Training Systems and Supporting Infrastructure

Asynchronous RL for LLM post-training. Asynchronous RL systems decouple generation and training to reduce synchronization overhead and improve resource utilization. Laminar (Sheng et al., 2025a) introduces trajectory-level asynchrony through a decoupled architecture with relay workers, fine-grained weight synchronization, and repacking of long-tail trajectories. StreamRL (Zhong et al., 2025) uses disaggregated stream generation with k-step staleness. AReaL (Fu et al., 2025) supports partial rollout, allowing ongoing trajectories to be interrupted and resumed with updated weights. AsyncFlow (Han et al., 2025) and LlamaRL (Wu et al., 2025) develop asynchronous execution frameworks for LLM post-training; AcceRL (Lu et al., 2026) extends asynchronous training, inference, and rollout execution to vision-language-action models. These systems motivate the generation-tail problem studied here. Our theoretical comparison concerns the specified hierarchical pairwise and multi-repacking policies, rather than a comparison of complete training runtimes. Distribution-aware speculative decoding (Shao et al., 2025) attacks long-tail generation by accelerating decoding itself; our repacking model preserves trajectory completion times and measures the replica occupancy saved by changing placement.

Synchronous RL and training pipelines. HybridFlow (Sheng et al., 2025b) combines parallel execution strategies for RLHF, OpenRLHF (Hu et al., 2024) provides scalable RLHF execution, ReaL (Mei et al., 2024) reallocates parameters, and RLHFuse (Zhong et al., 2024) overlaps stages through inter- and intra-stage fusion. In model-parallel training, GPipe (Huang et al., 2019), PipeDream (Narayanan et al., 2019), PipeDream-2BW (Narayanan et al., 2021a), and interleaved pipeline scheduling (Narayanan et al., 2021b) reduce idle time associated with forward–backward dependencies. These execution and pipeline mechanisms are complementary to consolidating residual generation work after local queues drain.

Inference engines, migration, and distributed RL. vLLM (Kwon et al., 2023) improves KV-cache allocation with PagedAttention, SGLang (Zheng et al., 2023) supports efficient LLM execution, and Ray (Moritz et al., 2018) provides distributed execution abstractions. Llumnix (Sun et al., 2024) implements live request and state migration across serving instances for load balancing, resource fragmentation, and differentiated service objectives. Such systems supply mechanisms on which repacking can be implemented; migration costs are outside our occupied-time model. Earlier distributed RL systems include IMPALA (Espeholt et al., 2018), with its actor–learner architecture and importance correction; A3C (Mnih et al., 2016), with asynchronous updates; and RLlib (Liang et al., 2018), with distributed RL abstractions.

Model-Based Analysis of LLM Systems

Growing-cache placement and MELL. MELL (Liu et al., 2025) combines a serving-system implementation with an online allocation analysis. Its model admits request arrivals and departures, observes current cache sizes, and allows migration as caches grow. It minimizes the maximum number of GPUs used over time subject to memory feasibility, and bounds packing quality and the number of migrations per operation. Its implementation also adapts between token transfer and KV-cache transfer to account for communication and computation costs. Thus growing caches, unknown response lengths, and migration are shared features, not novelty claims of this paper.

Our model specializes to a finite batch for one weight version, with prescribed replica starts, fixed parallel-slot limits, exponential service, and linear cache growth. Each replica enters the tail at a random anchor; as completions open space, repacking consolidates survivors and releases replicas for subsequent work. We minimize the expected sum of tail occupancies, rather than peak GPU provisioning or request latency. Our contribution is the explicit stochastic comparison: hierarchy sums, controlled finite-system and asynchronous remainders, and a strict leading-term advantage of multi-repacking under the stated conditions. These are different guarantees from a competitive packing ratio. The model also checks cache thresholds at packing times and preserves service during migration; continuous capacity enforcement and migration overhead remain separate system issues, as discussed in Section 6.

Scheduling under KV-cache and throughput constraints. Jaillet et al. (2025) analyze inference latency under KV-cache constraints using a hindsight benchmark and online scheduling guarantees. Ao et al. (2025) use a fluid model of growing cache requirements to guide admission and scheduling with known or unknown output lengths. Li et al. (2025) characterize throughput-optimal scheduling in a queueing model of LLM inference and agent workloads. Feng et al. (2026) give competitive guarantees for non-clairvoyant completion-time scheduling within a replica under a hard cache budget, with evictions discarding accumulated progress. Unknown output lengths and quantitative LLM-system analysis therefore already have substantial theoretical treatments. Our question concerns how service-preserving migration across replicas changes cumulative occupancy during a batch's tail, and how multi-repacking compares with hierarchical pairwise at the level of leading constants and error terms.

Dynamic and Stochastic Bin Packing

Migration and active bin time. Fully dynamic bin packing studies the tradeoff between the number of bins used and the amount of repacking as fixed-size items arrive and depart (Feldkord et al., 2018). Mellou et al. (2024) study total active bin time with unknown item durations, characterize migration-budget tradeoffs, and also analyze migrations that extend item lifetimes. Their objective is directly related to occupied replica time. The key difference in item information is that our currently observed cache size does not specify the memory required for the remainder of service: a live trajectory grows until its unknown completion time. The cache and active-count tests in our algorithms therefore depend on the evolving survivor population. We analyze stochastic consolidation times and policy-specific occupancy, rather than a worst-case competitive tradeoff with a migration budget.

Stochastic, time-varying item sizes. Hong et al. (2023) explicitly allow resource demand to vary with a job's Markovian service phase. Jobs arrive according to Poisson processes and remain on their selected servers after dispatch. The objective is the steady-state expected number of active servers subject to a resource-contention cost constraint, and their Join-Requesting-Server policy obtains asymptotic optimality in a large service system. This work is relevant both to changing item sizes and to active-server minimization. Our transient finite-batch analysis instead tracks age-dependent cache growth, random tail-entry times, and repeated migrations. It yields explicit occupied-time formulas and a comparison between two repacking mechanisms, rather than a steady-state dispatching optimality result.

B Detailed Definitions, Auxiliary Results, and Proofs

The proof is organized by dependency. We establish the model and packing rules, collect the crossing and concentration tools shared by both hierarchies, and then analyze pairwise and whole-block occupied time. The switch argument transfers the latter result to the implemented policy; the final steps compare hierarchy limits and derive tail bounds.

B.1 Model and anchor facts

B.1.1 Slot dynamics and anchor conventions

For an occupied slot (i, j), let $T_{ij}(t)$ and $L_{ij}(t)$ be its trajectory's start time and lifetime. Its age is $A_{ij}(t) = t - T_{ij}(t)$ until completion at $A_{ij}(t) = L_{ij}(t)$, and is zero when idle. Cache grows at unit rate, so $C_i(t) = \sum_{j=1}^{B} A_{ij}(t)$. Lifetimes need not be observable before completion. They are i.i.d. $\mathrm{Exp}(\lambda)$, and service is work-conserving: each completion immediately admits the next local queued trajectory, until all N trajectories have been processed. If $Q_i(t)$ is the remaining queue length and $\mathbf{1}_{ij}(t)$ the occupancy indicator, the local Markov state is

$$\mathbf{S}_i(t) = \big((A_{ij}(t), \mathbf{1}_{ij}(t))_{j=1}^{B}, Q_i(t)\big).$$

The joint state additionally records levels, block membership, and switch counters. Migration preserves ages and residual lifetimes; transfer overhead is omitted. Packing guarantees output load at most $C_{\mathrm{th}} \le C_{\max}$, whereas subsequent cache growth is unconstrained. Stalling or offloading at the physical limit is outside this service model.

Write $s_{\min} = \min_i s_i$, $s_{\max} = \max_i s_i$, and $k = N - B + 1$. The finite-batch anchor is the k-th aggregate completion and has $B - 1$ survivors. Relabel these survivors by $j \in [B - 1]$ without using their future lifetimes; this index labels trajectories, not physical slots. Their anchor ages $\tilde{A}_{ij}$ can be dependent. By memorylessness, the residuals $\tilde{Y}_{ij}$ are independent $\mathrm{Exp}(\lambda)$ variables. Independence across original replicas makes the full residual vector independent of the collection of their own pre-anchor service histories. At local time $u = t - s_i \ge 0$, define

$$\tilde{Z}_{ij}(u) = (\tilde{A}_{ij} + u)\mathbf{1}_{\{\tilde{Y}_{ij} > u\}}, \qquad Z_i(u) = \sum_{j=1}^{B-1} \tilde{Z}_{ij}(u).$$

The original index i is retained after migration, so Z_i always tracks the same ancestry. A single index j may likewise enumerate a fixed collection of originals, without implying independence of its ages. Lemmas 6 and 7 give the anchor moments and mean-load curves.

The random anchors and ages share their pre-anchor history. We therefore use bounds simultaneous in every original replica and every local time, evaluate them pathwise at $t - s_i$, and intersect with the span event from Lemma 5. We never infer independence by conditioning on that event. Migration leaves the final completion time $\max_{i,j}(s_i + \tilde{Y}_{ij})$ unchanged; the objective instead sums the time that physical replicas remain occupied after their anchors.

Synchronized-anchor model. Generate each original replica's finite-batch anchor data as above, translate its local tail to the common origin $u = 0$, and apply anchor admission and event closure there. This preserves each replica's joint law and independence across replicas. The synchronous comparison uses this constructed model, without conditioning on coincident random anchors. The asynchronous results retain the actual epochs s_i.

B.1.2 Finite-batch anchor-span control

We record the auxiliary estimate used to control the random anchor spread in the asynchronous occupancy analysis. Throughout the theoretical analysis, $\mathbb{P}$, $\mathbb{E}$, and L^p norms are conditional on the prescribed start vector $(t_i)_{i\in[R]}$, which is independent of service. Thus T_0 and T_{span} are deterministic in these bounds. We do not condition on the random anchors. Write $s_{\min} := \min_i s_i$, $s_{\max} := \max_i s_i$, and $k := N - B + 1$. For the startup spread T_0, define

$$T_{\mathrm{span}} := T_0 + \frac{2}{B\lambda}\left[\sqrt{8k\log(8RB)} + 4\log(8RB)\right]. \tag{25}$$

Lemma 5 (Finite-batch anchor-span control)**.** *Under the finite-batch model of Section 4.1, with T_{span} defined in* (25)*,*

$$\mathbb{P}(s_{\max} - s_{\min} > T_{\mathrm{span}}) \le (8RB)^{-3}. \tag{26}$$

Moreover,

$$\|s_{\max} - s_{\min}\|_2 \le C\left[T_0 + \frac{\sqrt{k\log(2R)} + \log(2R)}{B\lambda}\right] \le CT_{\mathrm{span}}. \tag{27}$$

Proof. Until the anchor epoch, a replica always has B active trajectories. Hence, for each $i \in [R]$,

$$s_i - t_i = \sum_{r=1}^{k} E_{ir}, \qquad E_{ir} \overset{\text{i.i.d.}}{\sim} \mathrm{Exp}(B\lambda), \qquad k = N - B + 1. \tag{28}$$

The collections are independent across replicas. Write $G_i := s_i - t_i$ and $\mu_k := k/(B\lambda)$. The standard Chernoff bounds for a sum of k unit-mean exponential variables give, for every $z > 0$,

$$\mathbb{P}\left(G_i - \mu_k \ge \frac{\sqrt{2kz} + z}{B\lambda}\right) \le e^{-z}, \tag{29}$$

$$\mathbb{P}\left(\mu_k - G_i \ge \frac{\sqrt{2kz}}{B\lambda}\right) \le e^{-z}. \tag{30}$$

For completeness, these inequalities follow by applying Chernoff's method to

$$\mathbb{E}e^{\theta(B\lambda G_i - k)} = e^{-k\theta}(1-\theta)^{-k}, \qquad 0 < \theta < 1,$$

and to the analogous Laplace transform with $\theta < 0$, followed by the inequalities $-\log(1-\theta) - \theta \le \theta^2/[2(1-\theta)]$ and $\theta - \log(1+\theta) \le \theta^2/2$.

Let

$$M_R := \max_{i\in[R]} |G_i - \mu_k|.$$

A union bound in (29)– (30) yields

$$\mathbb{P}\left(M_R > \frac{\sqrt{2kz} + z}{B\lambda}\right) \le 2Re^{-z}. \tag{31}$$

Since

$$s_{\max} - s_{\min} \le T_0 + \max_i G_i - \min_i G_i \le T_0 + 2M_R,$$

we may take $z = 4\log(8RB)$ in (31). Because $2R \le 8RB$, the failure probability is at most $(8RB)^{-3}$, and the resulting upper endpoint is precisely (25). This proves (26).

To obtain the second-moment estimate, put $z = \log(2R) + x$ in (31) and integrate the resulting tail in $x \ge 0$. Equivalently, use the standard sub-gamma maximal inequality associated with (29)– (30). It gives

$$\|M_R\|_2 \le \frac{C}{B\lambda}\left[\sqrt{k\log(2R)} + \log(2R)\right].$$

The deterministic inequality $s_{\max} - s_{\min} \le T_0 + 2M_R$ now proves the first inequality in (27); the second follows directly from (25) and $\log(2R) \le \log(8RB)$. □

Independently sampled starts. For unconditional finite-system bounds, use a deterministic upper bound on the startup spread in place of T_0, including in T_{span} and the cutoff conditions. The conditional error bounds then hold uniformly over the start distribution and can be averaged. For unconditional leading-order formulas, require this deterministic upper bound to be $o(\log B/\lambda)$. The error estimates in Appendices B.4.4 and B.5.3 are uniform under this condition, so averaging preserves the $o(R/\lambda)$ remainders. In particular, independent uniform starts on $[0, 1/\lambda]$ have startup spread at most $1/\lambda$.

B.1.3 ANCHOR-AGE MOMENTS

For the anchor data of Section 4.1, the following estimate quantifies the finite-batch correction without assuming independence among different surviving ages.

Lemma 6 (KV-cache Usage at the Anchor Time)**.** *At the anchor time, for each replica i and trajectory j, the KV-cache usage $\tilde{A}_{ij}$ satisfies*

$$\mathbb{E}[\tilde{A}_{ij}] = \frac{1}{\lambda}[1-(1-\frac{1}{B})^{N-B+1}], \qquad 0 \le \frac{1}{\lambda} - \mathbb{E}[\tilde{A}_{ij}] \le \frac{1}{\lambda}\epsilon_N,$$

and

$$\left|\mathrm{Var}(\tilde{A}_{ij}) - \frac{1}{\lambda^2}\right| \le \frac{C}{\lambda^2}\left(1+\frac{N}{B}\right)\epsilon_N.$$

Moreover, $\tilde{A}_{ij}$ and $\tilde{Y}_{ij}$ are independent. No independence among the anchor ages $\{\tilde{A}_{ij}\}_{j=1}^{B-1}$ is asserted.

Proof. The calculation must be made directly at the anchor completion epoch; one cannot substitute this random epoch into a deterministic-time renewal formula. Write

$$k := N - B + 1, \qquad q := 1 - \frac{1}{B}.$$

Until the anchor epoch there are always B active trajectories. Hence the first k aggregate inter-completion times $E_1, \dots, E_k$ are independent $\mathrm{Exp}(B\lambda)$ variables. Their completing-slot marks $M_1, \dots, M_k$ are independent and uniform on $[B]$, independently of the inter-completion times. The k-th completion is the anchor event: its slot becomes idle and the other $B-1$ slots contain the survivors.

Fix one survivor slot j, so that $M_k \neq j$, and let $K_j \in \{1, \dots, k\}$ be the number of aggregate inter-completion intervals since the most recent completion in slot j, truncated at the beginning of the batch. Conditional on $M_k \neq j$,

$$\mathbb{P}(K_j \ge r) = q^{r-1}, \qquad 1 \le r \le k.$$

The marks are independent of the inter-completion times, and, conditional on $K_j = r$, the anchor age is the sum of r independent $\mathrm{Exp}(B\lambda)$ variables. Therefore

$$\mathbb{E}[\tilde{A}_{ij}] = \frac{\mathbb{E}[K_j]}{B\lambda} = \frac{1}{B\lambda}\sum_{r=1}^{k} q^{r-1} = \frac{1}{\lambda}(1-q^k). \tag{32}$$

Since $k = N - B + 1$, this immediately gives the stated mean bound.

For the second moment, the conditional gamma law gives

$$\mathbb{E}[\tilde{A}_{ij}^2 \mid K_j = r] = \frac{r(r+1)}{B^2\lambda^2}.$$

Using $r(r+1) = 2\sum_{a=1}^{r} a$ and the tail-sum formula,

$$\mathbb{E}[\tilde{A}_{ij}^2] = \frac{2}{B^2\lambda^2}\sum_{r=1}^{k} r q^{r-1}.$$

The corresponding infinite sum equals B^2, while

$$\sum_{r=k+1}^{\infty} r q^{r-1} = q^k\big[(k+1)B + qB^2\big].$$

Consequently,

$$0 \le \frac{2}{\lambda^2} - \mathbb{E}[\tilde{A}_{ij}^2] \le \frac{C}{\lambda^2}\left(1+\frac{k}{B}\right)q^k.$$

Combining this estimate with (32) proves the variance bound in the lemma (with a larger universal constant).

Finally, the strong memoryless property at the k-th aggregate completion implies that the residual lifetimes in the $B-1$ surviving slots are independent $\mathrm{Exp}(\lambda)$ variables and are independent of the entire pre-anchor history. In particular, $\tilde{Y}_{ij}$ is independent of $\tilde{A}_{ij}$. The slot marks share the same aggregate event sequence, so this argument does not assert independence among the different anchor ages. □

B.1.4 MEAN TAIL LOAD

Using the anchor mean μ_A from Section 4.1, we record the mean-load identities and their finite-batch expansion.

Lemma 7 (Mean KV-cache load after the anchor)**.** *For each replica i and trajectory j, the expected per-trajectory KV-cache usage at time $t = s_i + u$ is*

$$\mathbb{E}[\tilde{Z}_{ij}(u)] = e^{-\lambda u}(u + \mu_A) = e^{-\lambda u}\left(u + \frac{1}{\lambda} + O\left(\frac{e^{-N/B}}{\lambda}\right)\right).$$

Therefore,

$$\mathbb{E}[Z_i(u)] = (B-1)e^{-\lambda u}(u + \mu_A) = (B-1)e^{-\lambda u}\left(u + \frac{1}{\lambda} + O\left(\frac{e^{-N/B}}{\lambda}\right)\right).$$

Proof. Age–residual independence from Lemma 6 gives $\mathbb{E}[\tilde{Z}_{ij}(u)] = \mathbb{P}(\tilde{Y}_{ij} > u)(u + \mathbb{E}\tilde{A}_{ij}) = e^{-\lambda u}(u + \mu_A)$. Substitute the finite-batch expansion of μ_A and sum over the $B-1$ survivors. □

B.1.5 ROLES OF THE SCALE ASSUMPTIONS

The assumptions are stated in the Scaling assumptions paragraph. The stationary mean cache load of B busy slots is B/λ. The upper margin $\lambda C_{\mathrm{th}}/B \le 1 - \eta$ places crossings on a uniformly regular decreasing branch and provides slot-count slack. Finite-system bounds use this upper margin; the lower bound $c_- > 0$ additionally gives $\lambda C_{\mathrm{th}} = \Theta(B)$ when simplifying their errors. Lower thresholds delay crossings for a fixed group and leave more unused capacity after packing. Physical capacity and subsequent growth are governed by the separate convention in Appendix B.1.1.

The lower batch-size condition $N = \omega(B\log(RB))$ makes $\epsilon_N = (RB)^{-\omega(1)}$, controlling coupling failure across all originals. The upper condition $N = o(B^2/\log(RB))$ gives

$$\lambda(T_{\mathrm{span}} - T_0) = O\left(\frac{\sqrt{N\log(RB)} + \log(RB)}{B}\right) = o(1).$$

Together they imply $\log(RB) = o(\sqrt{B})$. The condition $\log R = \Theta(\log B)$ supplies enough replicas for the uncontrolled frontier to be negligible, while $\lambda T_0 = o(\log B)$ permits growing startup spread and implies $\lambda T_{\mathrm{span}} = o(\log B)$ and $Be^{-\lambda T_{\mathrm{span}}} = B^{1-o(1)}$. The latter leaves enough survivors through the slowly growing cuts verified in the asymptotic proofs; no bounded-span assumption is needed.

These three conditions define the *stochastic asymptotic regime*. Here λ, c_-, η are fixed, as are $m \ge 3$ and $\alpha \in (0,1)$ for multi-repacking. Constants may depend on these fixed dimensionless parameters, with λ displayed explicitly. The regime yields Theorems 2 and 3; Proposition 25 also admits the other switch depths in (166). Finite-system bounds and empirical comparisons use their own stated hypotheses. The small-threshold and large-m limits in Appendix C concern the separate idealized hierarchy model.

B.1.6 NO-REPACKING BASELINE

For the residual lifetimes of Section 4.3, recall $\tau_i^{\mathrm{no}} = \max_{1\le j\le B-1}\tilde{Y}_{ij}$.

Lemma 8 (Expected tail occupied replica time without repacking)**.** *Let $H_{B-1} := \sum_{k=1}^{B-1} 1/k$ denote the $(B-1)$-th harmonic number. Then $\mathbb{E}[\tau_i^{\mathrm{no}}] = H_{B-1}/\lambda$.*

Proof. The ordered residuals satisfy $\tilde{Y}_{i,(0)} = 0$ and $\tilde{Y}_{i,(j)} - \tilde{Y}_{i,(j-1)} \sim \mathrm{Exp}((B-j)\lambda)$ for $1 \le j \le B-1$ (Ross, 2014). Since $\tau_i^{\mathrm{no}} = \tilde{Y}_{i,(B-1)}$ is the sum of these spacings, its expectation is $\sum_{j=1}^{B-1}[(B-j)\lambda]^{-1} = H_{B-1}/\lambda$. □

B.2 PACKING AND EVENT RULES

B.2.1 COMPLETE HIERARCHICAL PAIRWISE ALGORITHM

The following is the complete procedure summarized in Section 4.2. It acts only on anchored labels; the post-completion condition $Q(r) = 0$ and $N_{\mathrm{act}}(r) < B$ identifies anchor admission under the stated dispatch model. Tied completions include their immediate queue dispatches; each plan scans only candidates not already assigned to a pair.

Algorithm 2 Hierarchical Pairwise Repacking (Single Version)

Require: Set of rollout replicas $\mathcal{R}$, number of parallel slots B, migration threshold C_{th}

Initialize $\ell(r) \leftarrow 0$ for all $r \in \mathcal{R}$
for each completion epoch t, after processing tied completions, queue dispatches, and anchor admissions **do**
  Release every started, unreleased replica r with $Q(r) = 0$ and $N_{\mathrm{act}}(r) = 0$, and remove it from the current hierarchy
  while true do
    $\mathcal{S} \leftarrow \emptyset$
    for each started, unreleased $r \in \mathcal{R}$ **do**
      if $Q(r) = 0$ and $1 \le N_{\mathrm{act}}(r) < B$ **then**
        $\mathcal{S} \leftarrow \mathcal{S} \cup \{r\}$
    Sort $\mathcal{S}$ by $(\ell(r),\, C(r))$
    Repacking plan $\mathcal{P} \leftarrow \emptyset$
    for each source r_{src} in the sorted order that remains in $\mathcal{S}$ **do**
      if there exists $r_{\mathrm{dst}} \in \mathcal{S} \setminus \{r_{\mathrm{src}}\}$ such that $\ell(r_{\mathrm{dst}}) = \ell(r_{\mathrm{src}})$, $C(r_{\mathrm{dst}}) + C(r_{\mathrm{src}}) \le C_{\mathrm{th}}$, and $N_{\mathrm{act}}(r_{\mathrm{dst}}) + N_{\mathrm{act}}(r_{\mathrm{src}}) \le B$ **then**
        Select a feasible r^*_{dst} with the largest $C(r^*_{\mathrm{dst}})$
        $\mathcal{P} \leftarrow \mathcal{P} \cup \{(r_{\mathrm{src}}, r^*_{\mathrm{dst}})\}$
        $\mathcal{S} \leftarrow \mathcal{S} \setminus \{r_{\mathrm{src}}, r^*_{\mathrm{dst}}\}$
    if $\mathcal{P} = \emptyset$ **then**
      break
    for each $(r_{\mathrm{src}}, r_{\mathrm{dst}}) \in \mathcal{P}$ **do**
      Migrate the remaining trajectories from r_{src} to r_{dst}
      $\ell(r_{\mathrm{dst}}) \leftarrow \ell(r_{\mathrm{dst}}) + 1$
      Release replica r_{src}

B.2.2 FEASIBLE MULTI-REPLICA PACKING

Lemma 9 (Feasible multi-replica packing)**.** *Let $p \ge 1$ be an integer and let $\mathcal{T}$ be a finite collection of active trajectories such that $|\mathcal{T}| \le pB$ and $c_\tau(t) \le \alpha C_{\mathrm{th}}$ for every $\tau \in \mathcal{T}$. If*

$$\sum_{\tau \in \mathcal{T}} c_\tau(t) \le p(1-\alpha) C_{\mathrm{th}},$$

then all trajectories in $\mathcal{T}$ can be packed into p replicas, each with at most B trajectories and KV-cache load at most C_{th}.

The lemma converts aggregate load, slot count, and individual-size bounds into feasible placement. Its proof sorts trajectories by size and distributes them cyclically; the reserved capacity absorbs the largest possible load imbalance. The whole-block phase uses $p = (m-1)^{\ell+1}$ at level ℓ, and the independent phase uses $p = m - 1$.

Proof. Put $q := |\mathcal{T}|$. If $q = 0$, the assertion is immediate. Otherwise order the item sizes as $x_1 \ge x_2 \ge \cdots \ge x_q$ and assign item j to bin $1 + ((j-1) \bmod p)$. Since $q \le pB$, every bin receives at most B items. It remains to verify the load constraint.

Pad the list by zeros to length pB, and let L_a be the load of bin a. For $a < b$,

$$L_a - L_b = \sum_{r=0}^{B-1} (x_{rp+a} - x_{rp+b}).$$

The intervals of indices in these differences are disjoint, so their sum is at most the total variation $x_1 - 0$. Interchanging a and b therefore gives

$$|L_a - L_b| \le x_1 \le \alpha C_{\mathrm{th}}.$$

In particular, with $\overline{L} = p^{-1} \sum_a L_a$,

$$\max_a L_a \le \overline{L} + x_1 \le (1-\alpha)C_{\mathrm{th}} + \alpha C_{\mathrm{th}} = C_{\mathrm{th}}.$$

Thus both the slot and KV-cache constraints hold. Taking $p = (m-1)^{\ell+1}$ gives the packing step used by a level-ℓ blockwise round. □

B.2.3 WHOLE-BLOCK REFERENCE PROCESS

Algorithm 3 is the auxiliary execution that never switches. Its complete event step is the initial-phase subroutine of Algorithm 1. Given the same service paths and tie rules, the two executions agree until the proposed policy's certificate holds. The block concentration and release-count arguments below concern this reference. Appendix B.6.1 completes the proof for the proposed policy after independent regrouping begins. For a block G, let $\mathcal{R}(G)$ be its current physical representatives and let $\mathcal{T}(G) := \bigcup_{r \in \mathcal{R}(G)} \mathcal{T}(r)$.

Algorithm 3 Whole-Block Reference Process (Auxiliary)

Require: Replicas $\mathcal{R}$, parallel-slot limit B, threshold C_{th}, merge ratio $m \ge 3$, and $\alpha \in (0,1)$
Ensure: Repacking plan $\mathcal{P}$

Initialize one singleton analytical block $G = \{r\}$ of level zero for each $r \in \mathcal{R}$
for each completion epoch t, after processing tied completions, queue dispatches, and anchor admissions **do**
  $\mathcal{P} \leftarrow \emptyset$
  Among unreleased representatives that have started, release each r with $Q(r) = 0$ and $N_{\mathrm{act}}(r) = 0$; remove it from its block's $\mathcal{R}(G)$ and mark that block terminal
  while true do
    changed $\leftarrow$ false
    for each represented level ℓ in increasing order **do**
      Let $\mathcal{S}_\ell$ contain the nonterminal level-ℓ blocks all of whose constituent original replicas have reached their anchors, whose representatives have empty queues, and whose live trajectories obey $c_\tau(t) \le \alpha C_{\mathrm{th}}$ for every $\tau \in \mathcal{T}(G)$
      while $|\mathcal{S}_\ell| \ge m$ **do**
        Set $p = (m-1)^{\ell+1}$; among all m-block subsets with at most pB active trajectories, choose $G_1, \dots, G_m$ with minimum total load
        if such a subset exists and, for $\mathcal{T} = \bigcup_{a=1}^{m} \mathcal{T}(G_a)$, $\sum_{\tau \in \mathcal{T}} c_\tau(t) \le p(1-\alpha)C_{\mathrm{th}}$ **then**
          Use the balanced construction in Lemma 9 to repack $\mathcal{T}$ into p representatives chosen from $\bigcup_{a=1}^{m} \mathcal{R}(G_a)$
          Release the other $(m-1)^\ell$ physical replicas and replace $G_1, \dots, G_m$ by their level-$\ell+1$ union block
          Immediately release any empty output representative, remove it from the union block's representative set, and mark the union block terminal if such a release occurs
          Add this blockwise repacking record to $\mathcal{P}$, remove $G_1, \dots, G_m$ from $\mathcal{S}_\ell$, and set changed $\leftarrow$ true
        else
          break
    if changed $=$ false **then**
      break

B.2.4 COMPLETE MULTI-REPACKING EVENT RULES

The following expands Algorithm 1, including its boundary depth, natural-release ordering, and future-anchor admission. It uses the whole-block reference defined immediately above.

Algorithm 4 Multi-Repacking (Proposed Policy)

Require: Replicas $\mathcal{R}$, slot limit B, threshold C_{th}, $m \geq 3$, $\alpha \in (0,1)$, and fixed $d \in \{0, \ldots, K_m\}$

Initialize singleton level-zero blocks, completed whole-block round counters, and free $\leftarrow$ false
for each trajectory-completion epoch t **do**
  Process all tied completions and queue dispatches; then admit newly anchored originals at level zero, independently if free $=$ true
  if free $=$ false **then**
    if $d = 0$ **then**
      Apply the reference's natural-release and terminal rules
      if the first anchor has occurred **then**
        free $\leftarrow$ true
    else
      Execute the complete event step of Algorithm 3, including natural releases and exhaustive whole-block packing; update round counters
      if for every $0 \leq \ell < d$, the completed level-ℓ round count equals $\lfloor R/m^{\ell+1} \rfloor$ **then**
        free $\leftarrow$ true
    if free $=$ true **then**
      Make each anchored, unreleased, nonempty representative independent at its inherited level, including nonempty representatives of terminal blocks
  if free $=$ true **then**
    Release each started, unreleased r with $Q(r) = 0$ and $N_{\mathrm{act}}(r) = 0$; never reactivate released labels
    while some level admits a feasible group of m independent representatives **do**
      At the smallest such level, select a minimum-load m-subset of eligible representatives with at most $(m-1)B$ trajectories, each of size at most αC_{th}, and total load at most $(m-1)(1-\alpha)C_{\mathrm{th}}$
      Apply Lemma 9 with $p = m - 1$; release the unretained representative and any empty outputs immediately
      Advance every nonempty output independently by one level

B.3 Common crossing and concentration tools

B.3.1 Fixed analytical blocks and deterministic crossings

We use the fixed analytical block, first-passage time σ_ℓ, and decreasing-branch crossing $\bar{\sigma}_\ell$ defined in Section 4.3. Its mean load is $\mu_\ell(u) := n_\ell e^{-\lambda u}(\mu_A + u)$, with $n_\ell = 2^\ell(B-1)$. The threshold $C_{\mathrm{th}}/2$ ensures cache feasibility for two such inputs; the occupied-time proofs check their combined slot count separately.

In the fixed-block lemmas and the unsimplified pairwise bounds, generic constants C depend only on η and may change from line to line. Under (6), simplified rates may also depend on c_-. Constants are independent of R, B, N, λ, the level, and the start times; all dependence on λ is explicit. Fixed window multipliers are chosen before the survivor constants, as detailed in Appendix B.3.6.

Lemma 10 (Mean-load crossing time)**.** *For each level $\ell \geq 0$, define the normalized threshold ratio*

$$\rho_\ell := \frac{\lambda(C_{\mathrm{th}}/2)}{n_\ell} = \frac{C_{\mathrm{th}}\lambda}{2^{\ell+1}(B-1)}.$$

For $0 < \rho \leq 1$, define

$$\Phi(\rho) := -W_{-1}\left(-\frac{\rho}{e}\right) - 1.$$

Assume that $\rho_\ell \leq 1 - \eta$ for some fixed $\eta > 0$ and that N/B is sufficiently large. Then the decreasing-branch crossing exists uniquely, and the mean-load crossing time $\bar{\sigma}_\ell$ satisfies

$$\bar{\sigma}_\ell = -\frac{1}{\lambda} W_{-1}\left(-\rho_\ell e^{-\lambda\mu_A}\right) - \mu_A.$$

Moreover, the exact expression has the uniform large-batch expansion

$$\left|\bar{\sigma}_\ell - \frac{\Phi(\rho_\ell)}{\lambda}\right| \leq \frac{C}{\lambda}\epsilon_N.$$

Proof. The natural load margin implies $\rho_\ell \le \rho_0 = C_{\mathrm{th}}\lambda/[2(B-1)] \le 1-\eta$ for $B \ge 2$, so it suffices to check level zero. More generally, let $0 < \rho \le 1-\eta$, put $a_N = \lambda\mu_A = 1-(1-1/B)^{N-B+1}$, and consider a decreasing crossing s of $e^{-\lambda s}(\mu_A + s) = \rho/\lambda$. Then

$$w = -\lambda(\mu_A + s), \qquad we^w = -\rho e^{-a_N}, \qquad s = \lambda^{-1}[-W_{-1}(-\rho e^{-a_N}) - a_N].$$

For N/B large, $\rho e^{-a_N} < 1/e$. The branch $w < -1$ is unique and gives $s > 0$, since $a_N < 1$. For uniform sensitivity, write $y_\rho(a) = -W_{-1}(-\rho e^{-a})$ and $G_\rho(a) = y_\rho(a) - a$. Implicit differentiation of $y_\rho e^{-y_\rho} = \rho e^{-a}$ gives

$$G'_\rho(a) = \frac{1}{y_\rho(a) - 1}, \qquad y'_\rho(a) = \frac{y_\rho(a)}{y_\rho(a) - 1}.$$

For $a \in [a_N, 1]$, the fixed margin and $a_N \to 1$ give $y_\rho(a) \ge 1 + c_\eta$ uniformly, including as $\rho \downarrow 0$. Thus both derivatives are bounded, and $|G_\rho(a_N) - G_\rho(1)| \le C(1-a_N) \le C\epsilon_N$. Using $G_\rho(1) = \Phi(\rho)$ and $\rho = \rho_\ell$ proves both claims. □

B.3.2 Deterministic crossings for multi-repacking

Recall the whole-block reference and the threshold, crossing, and slope parameters from Section 4.4. For $K_m = \lfloor \log_m R \rfloor$, define its nominal release counts and remaining-label counts by

$$\nu_{\ell,m} := (m-1)^\ell \left\lfloor \frac{R}{m^{\ell+1}} \right\rfloor, \tag{33}$$

$$Q_h := R - \sum_{\ell=0}^{h-1} \nu_{\ell,m}, \tag{34}$$

for $0 \le \ell < K_m$ and $0 \le h \le K_m$. These counts include unmatched blocks for arbitrary R and concern the whole-block reference. We use the deterministic parameters (18)–(20) also at analytical cutoff levels. The next lemma gives the exact Lambert representation and its uniform finite-batch correction.

Lemma 11 (Mean-load crossing time for multi-repacking)**.** *Assume $\rho_{0,m} \le 1-\eta$ for some fixed $\eta > 0$. For N/B sufficiently large, the decreasing-branch crossing exists uniquely and*

$$\bar\sigma_{\ell,m} = -\frac{1}{\lambda} W_{-1}\left(-\rho_{\ell,m} e^{-\lambda\mu_A}\right) - \mu_A.$$

This is the exact finite-batch formula. Uniformly over all hierarchy levels, its large-batch expansion is

$$\left| \bar\sigma_{\ell,m} - \frac{\Phi(\rho_{\ell,m})}{\lambda} \right| \le \frac{C}{\lambda}\epsilon_N. \tag{35}$$

Proof. The crossing equation divided by $n_{\ell,m}$ is $e^{-\lambda\bar\sigma_{\ell,m}}(\mu_A + \bar\sigma_{\ell,m}) = \rho_{\ell,m}/\lambda$. Since $\rho_{\ell,m} \le \rho_{0,m} \le 1-\eta$, the scalar inversion and uniform sensitivity calculation in Lemma 10 apply with $\rho = \rho_{\ell,m}$, proving the exact formula and (35). For fixed m, α, the natural load margin supplies this condition with a fixed margin once B is large:

$$\rho_{0,m} \le \frac{m-1}{m}(1-\alpha)(1-\eta)\frac{B}{B-1} \le 1 - c_{\eta,m}.$$

Thus only level zero needs to be checked in either hierarchy. □

B.3.3 Proof of Lemma 1

We prove Lemma 1 for all survivors from a fixed collection of original replicas chosen independently of service. Write $n \ge 1$ for their number and take a threshold $c > 0$ with $\lambda c/n \le 1-\eta$. The proof uses no hierarchy-specific property: the pairwise specialization is $n = n_\ell$, $c = C_{\mathrm{th}}/2$; the multi specialization is $n = n_{\ell,m}$, $c = \theta_{\ell,m}$. Actual anchor ages within an original replica share completion histories and need not be independent. We first couple them to independent stationary ages, prove global path and first-passage bounds there, and then transfer probability and expectation estimates separately.

Stationary comparison vector. Fix a deterministic analytical block containing n surviving trajectories at its anchor epochs. For each original replica, place its anchor completion at time zero and consider the Palm version of a two-sided stationary marked Poisson process of aggregate completions. Poisson splitting represents this process as B independent rate-λ slot processes. Conditional on the mark of the completion at zero, the backward recurrence times of the other $B-1$ slot processes are independent $\mathrm{Exp}(\lambda)$ variables. Let A_j° be the corresponding backward recurrence time for the slot containing the j-th surviving trajectory. Different original replicas use independent Palm processes. Consequently, $A_1^\circ, \dots, A_n^\circ$ are independent $\mathrm{Exp}(\lambda)$ variables and are independent of the post-anchor residual lifetimes.

Lemma 12 (Stationary anchor coupling)**.** *Let $\tilde{A}_1, \dots, \tilde{A}_n$ be the actual anchor ages in the fixed analytical block above. The actual and stationary age vectors can be constructed on a common probability space together with an event $\mathcal{C}_n$, measurable with respect to the pre-anchor completion histories, such that*

$$\mathbb{P}(\mathcal{C}_n^c) \le n\epsilon_N. \tag{36}$$

On $\mathcal{C}_n$,

$$\tilde{A}_j = A_j^\circ, \qquad 1 \le j \le n.$$

If the actual and stationary processes use the same residual lifetimes $\tilde{Y}_1, \dots, \tilde{Y}_n$, then, on $\mathcal{C}_n$,

$$X_n(u) = X_n^\circ(u), \qquad \forall u \ge 0,$$

where

$$X_n(u) := \sum_{j=1}^{n} (\tilde{A}_j + u)\mathbf{1}_{\{\tilde{Y}_j > u\}}, \qquad X_n^\circ(u) := \sum_{j=1}^{n} (A_j^\circ + u)\mathbf{1}_{\{\tilde{Y}_j > u\}}.$$

Consequently, the two processes have the same first-passage time at every threshold on $\mathcal{C}_n$.

Proof. Write

$$k := N - B + 1, \qquad q_B := 1 - \frac{1}{B}.$$

Until the anchor epoch there are always B active trajectories. Hence the first k aggregate inter-completion times are independent $\mathrm{Exp}(B\lambda)$ variables; denote them by $E_1, \dots, E_k$. Their slot marks $M_1, \dots, M_k$ are independent and uniform on $[B]$, independently of the inter-completion times. The k-th completion is the anchor event.

We construct the stationary comparison process backward from the anchor, rather than forward from the batch start. Put its Palm event at time zero and give it mark M_k. For $1 \le r \le k-1$, identify the r-th aggregate event before zero with mark M_{k-r} and backward inter-event time E_{k-r+1}. Complete the process farther into the past using independent $\mathrm{Exp}(B\lambda)$ gaps and independent uniform marks. Because the gaps and marks are i.i.d., reversing these first $k-1$ pairs preserves their joint law, and the completed two-sided process has exactly the required Palm distribution.

Let b be a slot that survives the anchor event, so $M_k \ne b$. If b appears among $M_1, \dots, M_{k-1}$ and r_b is its last such occurrence, then both its actual age and its Palm age at the anchor are

$$E_{r_b+1} + \cdots + E_k.$$

Thus the two ages can differ only when b is absent from all of the preceding $k-1 = N-B$ marks. Conditional on $M_k \ne b$, this event has probability

$$\mathbb{P}(M_r \ne b,\ 1 \le r \le k-1) = \epsilon_N.$$

Let $\mathcal{C}_n$ be the event that every surviving slot represented in the fixed analytical block has been selected at least once before its anchor event. A union bound over its n surviving trajectories proves (36), and the backward Palm construction gives $\tilde{A}_j = A_j^\circ$ on $\mathcal{C}_n$.

Finally, the strong memoryless property at the anchor epochs implies that the post-anchor residual lifetimes are independent $\mathrm{Exp}(\lambda)$ variables and are independent of the pre-anchor completion histories. We may therefore use the same residual lifetimes in both processes. Equality of the anchor-age vectors then gives equality of the two paths for every $u \ge 0$, and hence equality of their first-passage times. □

Remark 1. *Lemma 12 does not assert that the actual anchor ages are independent conditional on $\mathcal{C}_n$. Conditioning on $\mathcal{C}_n$ may itself introduce dependence. Instead, all concentration estimates are applied to the unconditional stationary process X_n°. For every measurable path event $\mathcal{E}$, the pathwise equality on $\mathcal{C}_n$ gives*

$$\mathbb{P}\big(\mathcal{E}(X_n)\big) \le \mathbb{P}\big(\mathcal{E}(X_n^\circ)\big) + \mathbb{P}(\mathcal{C}_n^c).$$

Moreover, $\mathcal{C}_n^c$ is determined by the pre-anchor completion histories and is independent of the post-anchor residual lifetimes. This independence will be used below to estimate its contribution to the expectation separately.

Proof of Lemma 1. Use the unconditional stationary process from Lemma 12, with mean $m_n^\circ(u) = ne^{-\lambda u}(u + \lambda^{-1})$. Keep n, c general until the final specialization.

Uniform control of survivor counts. Set $N_n(u) = \sum_{j=1}^n \mathbf{1}_{\{\tilde{Y}_j > u\}}$, $v(u) = ne^{-\lambda u}$, and $L_x = x + \log(2n)$ for $x \ge 0$. The variables $U_j = e^{-\lambda \tilde{Y}_j}$ are independent uniform variables, so $N_n(u) = K_n(e^{-\lambda u})$, where $K_n(q) = \sum_j \mathbf{1}_{\{U_j < q\}}$. At the grid points $q_k = k/n$, Bernstein's inequality gives

$$\mathbb{P}(|K_n(q_k) - k| > b_k) \le 2e^{-L_x}, \qquad b_k = \sqrt{2kL_x} + \tfrac{2}{3}L_x.$$

The endpoints are deterministic. A union bound over $1 \le k < n$ therefore fails with probability at most $2(n-1)e^{-L_x} \le e^{-x}$. On the resulting event, for $k = \lfloor nq \rfloor$ and $q < 1$, monotonicity gives

$$K_n(q) - nq \le b_{k+1} + 1, \qquad nq - K_n(q) \le b_k + 1.$$

Since $k \le nq < k + 1$ and $L_x \ge \log 2$, these inequalities imply

$$|N_n(u) - v(u)| \le C[\sqrt{v(u)L_x} + L_x], \qquad \forall u \ge 0, \tag{37}$$

with probability at least $1 - e^{-x}$. This includes the whole tail $v(u) < 1$: it corresponds to $k = 0$, where $K_n(q) \le K_n(1/n)$. No time truncation is required.

Uniform control of the anchor-age contribution. Order the indices so that $\tilde{Y}_{\pi(1)} \ge \cdots \ge \tilde{Y}_{\pi(n)}$. This permutation depends only on the residual lifetimes, so the variables $A^\circ_{\pi(1)}, \dots, A^\circ_{\pi(n)}$ remain independent exponentials. Indeed, conditionally on the permutation, for $|\theta| \le \lambda/2$,

$$\mathbb{E}\exp\left[\theta\left(A^\circ_{\pi(j)} - \frac{1}{\lambda}\right)\right] = \frac{e^{-\theta/\lambda}}{1 - \theta/\lambda} \le \exp\left(\frac{C\theta^2}{\lambda^2}\right).$$

The exponential Bernstein inequality therefore gives, after choosing the absolute constant C sufficiently large,

$$\mathbb{P}\left(\left|\sum_{j=1}^r \left(A^\circ_{\pi(j)} - \frac{1}{\lambda}\right)\right| > \frac{C}{\lambda}(\sqrt{rL_x} + L_x) \,\middle|\, \pi\right) \le 2e^{-2L_x}.$$

A union bound over $1 \le r \le n$, followed by averaging over π, implies, outside another event of probability at most e^{-x},

$$\left|\sum_{j=1}^r \left(A^\circ_{\pi(j)} - \frac{1}{\lambda}\right)\right| \le \frac{C}{\lambda}\left(\sqrt{rL_x} + L_x\right), \qquad 1 \le r \le n. \tag{38}$$

Since the active indices at time u are precisely the first $N_n(u)$ indices in this order, we have the exact decomposition

$$X_n^\circ(u) - m_n^\circ(u) = \sum_{j=1}^{N_n(u)} \left(A^\circ_{\pi(j)} - \frac{1}{\lambda}\right) + \left(u + \frac{1}{\lambda}\right)[N_n(u) - v(u)]. \tag{39}$$

On the count event, $N_n(u) \le v(u) + C[\sqrt{v(u)L_x} + L_x]$. Hence

$$\sqrt{N_n(u)L_x} \le \sqrt{v(u)L_x} + C\sqrt{L_x[\sqrt{v(u)L_x} + L_x]} \le C[\sqrt{v(u)L_x} + L_x],$$

where the last inequality follows from $L_x\sqrt{v(u)L_x} \le [\sqrt{v(u)L_x} + L_x]^2$. Applying (37) and (38) to (39), and using $u + \lambda^{-1} \ge \lambda^{-1}$, yields

$$\mathbb{P}\left(\exists u \ge 0 : |X_n^\circ(u) - m_n^\circ(u)| > C\left(u + \frac{1}{\lambda}\right)\left[\sqrt{ne^{-\lambda u}L_x} + L_x\right]\right) \le 4e^{-x}. \tag{40}$$

Thus no truncation is used and, in particular, no probability term independent of the deviation level is integrated over an infinite interval.

From path concentration to first passage. We now turn (40) into a global first-passage bound. For a threshold c, let $s^\circ(c)$ be the decreasing-branch solution of $m_n^\circ(s^\circ(c)) = c$, and set

$$v_c := v(s^\circ(c)) = ne^{-\lambda s^\circ(c)}, \qquad \kappa^\circ(c) := -\big(m_n^\circ\big)'(s^\circ(c)).$$

Assume $\lambda c/n \le 1 - \eta$. We now verify the first-passage bound on the entire half-line, rather than only near the deterministic crossing. To keep the choices of constants non-circular, we first use separate constants for the width of the crossing window and for the survivor condition, and absorb them into a single C only at the end.

For brevity, write

$$s := s^\circ(c), \qquad z := \lambda s, \qquad \kappa := \kappa^\circ(c) = \lambda s v_c.$$

The crossing equation is

$$\frac{\lambda c}{n} = (1+z)e^{-z}.$$

Since $z \mapsto (1+z)e^{-z}$ is strictly decreasing on $(0,\infty)$, the regular-crossing condition implies $z \ge z_\eta$ for some $z_\eta > 0$ depending only on η. Fix $a_\eta \in (0, \min\{1, z_\eta/2\})$ and set

$$\theta_\eta := e^{-a_\eta}\frac{1+z_\eta}{1+z_\eta-a_\eta} < 1, \qquad \delta_\eta := 1 - \theta_\eta > 0.$$

The strict inequality follows, for example, from $\log((1+z)/(1+z-a_\eta)) < a_\eta$ when $z > a_\eta$.

Let C_0 be the absolute constant in (40), and denote its pointwise radius by

$$r_x(u) := C_0\left(u + \frac{1}{\lambda}\right)\left[\sqrt{v(u)L_x} + L_x\right].$$

Also put

$$r_x^* := \left(s + \frac{1}{\lambda}\right)\left[\sqrt{v_c L_x} + L_x\right].$$

On the event in (40),

$$|X_n^\circ(u) - m_n^\circ(u)| \le r_x(u), \qquad \forall u \ge 0. \tag{41}$$

We first record uniform local comparisons. If $|u - s| \le a_\eta/\lambda$, then

$$e^{-a_\eta}v_c \le v(u) \le e^{a_\eta}v_c, \qquad u + \lambda^{-1} \le (1+a_\eta)(s + \lambda^{-1}).$$

Moreover, since $a_\eta \le z_\eta/2 \le z/2$ and

$$-\big(m_n^\circ\big)'(u) = \lambda u v(u),$$

there are constants $b_\eta, A_\eta > 0$, depending only on η, such that throughout this local interval

$$-\big(m_n^\circ\big)'(u) \ge b_\eta\kappa, \qquad r_x(u) \le A_\eta r_x^*. \tag{42}$$

Indeed, on the left of s one may use $u \ge s/2$ and $v(u) \ge v_c$, whereas on the right one may use $u \ge s$ and $v(u) \ge e^{-a_\eta}v_c$.

Choose K_η so large that $b_\eta K_\eta \ge 2A_\eta$, and define the working half-width

$$d_x := K_\eta\frac{r_x^*}{\kappa}.$$

Next choose $M_\eta \ge 1$ sufficiently large that

$$\frac{2K_\eta(1+z_\eta)}{z_\eta\sqrt{M_\eta}} \le a_\eta, \qquad C_0\left(M_\eta^{-1/2} + M_\eta^{-1}\right) \le \frac{\delta_\eta}{2}. \tag{43}$$

Suppose that $v_c \geq M_\eta L_x$. Since $\sqrt{v_c L_x} + L_x \leq 2\sqrt{v_c L_x}$, we obtain

$$d_x \leq \frac{2K_\eta}{\lambda}\frac{1+z}{z}\sqrt{\frac{L_x}{v_c}} \leq \frac{a_\eta}{\lambda} \leq \frac{s}{2}. \tag{44}$$

Thus both $s - d_x$ and $s + d_x$ lie in the local interval considered above.

We now exclude every earlier crossing. First consider

$$s - \frac{a_\eta}{\lambda} \leq u \leq s - d_x.$$

By (42),

$$\begin{aligned} m_n^\circ(u) - c &= \int_u^s -\big(m_n^\circ\big)'(t)\,\mathrm{d}t \\ &\geq b_\eta \kappa (s-u) \geq b_\eta \kappa d_x \geq 2A_\eta r_x^* \geq 2r_x(u). \end{aligned}$$

Consequently, (41) gives $X_n^\circ(u) > c$ throughout this local left-hand interval.

It remains to control the part far to the left of the crossing. If $0 \leq u \leq s - a_\eta/\lambda$, monotonicity of m_n° and the explicit formula for the mean give

$$\frac{c}{m_n^\circ(u)} \leq \frac{m_n^\circ(s)}{m_n^\circ(s - a_\eta/\lambda)} = e^{-a_\eta}\frac{1+z}{1+z-a_\eta} \leq \theta_\eta.$$

Here the last inequality uses that $(1+z)/(1+z-a_\eta)$ is decreasing in z and $z \geq z_\eta$. Therefore

$$m_n^\circ(u) - c \geq \delta_\eta m_n^\circ(u).$$

Since $v(u) \geq v_c$ for $u \leq s$, the relative path radius satisfies

$$\begin{aligned} \frac{r_x(u)}{m_n^\circ(u)} &= C_0\left[\sqrt{\frac{L_x}{v(u)}} + \frac{L_x}{v(u)}\right] \\ &\leq C_0\left[\sqrt{\frac{L_x}{v_c}} + \frac{L_x}{v_c}\right] \leq \frac{\delta_\eta}{2}, \end{aligned}$$

where the final inequality follows from $v_c \geq M_\eta L_x$ and (43). Hence

$$X_n^\circ(u) \geq m_n^\circ(u) - r_x(u) > c, \qquad 0 \leq u \leq s - \frac{a_\eta}{\lambda}.$$

Together with the local estimate, this proves

$$X_n^\circ(u) > c, \qquad 0 \leq u \leq s - d_x. \tag{45}$$

Finally, the right endpoint is controlled by the same local slope estimate:

$$\begin{aligned} c - m_n^\circ(s+d_x) &= \int_s^{s+d_x} -\big(m_n^\circ\big)'(t)\,\mathrm{d}t \\ &\geq b_\eta \kappa d_x \geq 2A_\eta r_x^* \geq 2r_x(s+d_x). \end{aligned}$$

Thus (41) implies

$$X_n^\circ(s+d_x) \leq c. \tag{46}$$

Writing

$$\tau_c^\circ := \inf\{u \geq 0 : X_n^\circ(u) \leq c\},$$

equations (45) and (46) yield

$$s - d_x \leq \tau_c^\circ \leq s + d_x.$$

Taking $C := \max\{K_\eta, M_\eta\}$, the assumption $v_c \geq CL_x$ implies the condition used above, while $d_x \leq Cr_x^*/\kappa$. We have therefore proved

$$|\inf\{u \geq 0 : X_n^\circ(u) \leq c\} - s^\circ(c)| \leq C\frac{(s^\circ(c) + \lambda^{-1})[\sqrt{v_c L_x} + L_x]}{\kappa^\circ(c)}. \tag{47}$$

No monotonicity of the random path is used: the lower bound follows from the simultaneous control of every $u \leq s - d_x$, and the upper bound requires only that the path be below c at the single time $s + d_x$.

Let $\bar{s}$ be the decreasing-branch solution of $ne^{-\lambda \bar{s}}(\mu_A + \bar{s}) = c$. Put $q_B = 1 - 1/B$, $a_N = \lambda\mu_A = 1 - q_B\epsilon_N$, and

$$\bar{v} = ne^{-\lambda\bar{s}}, \qquad \bar{\kappa} = \bar{v}[\lambda(\mu_A + \bar{s}) - 1], \qquad \rho = \lambda c/n.$$

Here $\bar{\kappa}$ is the negative derivative of the finite-batch mean at its crossing.

Transfer to the finite-batch process. The functions y_ρ, G_ρ in the proof of Lemma 10 satisfy $\lambda\bar{s} = G_\rho(a_N)$ and $\lambda s = G_\rho(1)$. Their uniformly bounded derivatives give

$$|\bar{s} - s| \le \frac{C}{\lambda} q_B \epsilon_N, \qquad |y_\rho(a_N) - y_\rho(1)| \le C q_B \epsilon_N. \tag{48}$$

The same margin gives $y_\rho(1) - 1 = \lambda s \ge z_\eta > 0$. The exact and stationary slopes are

$$\bar{\kappa} = \bar{v}\,[y_\rho(a_N) - 1], \qquad \kappa = v_c[y_\rho(1) - 1].$$

After increasing the lower bound on N/B if necessary, (48) consequently gives

$$\begin{aligned} C^{-1}\bar{v} \le v_c \le C\bar{v}, \\ C^{-1}\bar{\kappa} \le \kappa \le C\bar{\kappa}, \\ C^{-1}(\bar{s} + \lambda^{-1}) \le s + \lambda^{-1} \le C(\bar{s} + \lambda^{-1}). \end{aligned} \tag{49}$$

In particular, after enlarging the constant in the survivor condition, $\bar{v} \ge CL_x$ implies the stationary condition required in (47), and

$$\frac{(s + \lambda^{-1})[\sqrt{v_c L_x} + L_x]}{\kappa} \le C \frac{(\bar{s} + \lambda^{-1})[\sqrt{\bar{v} L_x} + L_x]}{\bar{\kappa}}. \tag{50}$$

Let τ_c and τ_c° denote the first-passage times of the actual and stationary processes, respectively. The pairwise parameters will be substituted at the end. On the coupling event $\mathcal{C}_n$, $\tau_c = \tau_c^\circ$. Hence, on the intersection of $\mathcal{C}_n$ with the global path event,

$$\begin{aligned} |\tau_c - \bar{s}| &\le |\tau_c^\circ - s| + |s - \bar{s}| \\ &\le C \frac{(\bar{s} + \lambda^{-1})[\sqrt{n e^{-\lambda \bar{s}} L_x} + L_x]}{\bar{\kappa}} + \frac{C}{\lambda}\epsilon_N. \end{aligned}$$

Using (40), (36), and a union bound now gives

$$\begin{aligned} \mathbb{P}\left(|\tau_c - \bar{s}| > C \frac{(\bar{s} + \lambda^{-1})[\sqrt{n e^{-\lambda \bar{s}} L_x} + L_x]}{\bar{\kappa}} + \frac{C}{\lambda}\epsilon_N \right) \\ \le 4e^{-x} + n\epsilon_N. \end{aligned} \tag{51}$$

This is the probability estimate in the lemma.

Expectation of the stationary first-passage error. Set $Z = |\tau_c^\circ - s|$, $L_0 = \log(2n)$, and $Y_{(n)} = \max_{j \le n} \tilde{Y}_j$. The global first-passage bound gives

$$\mathbb{P}(Z > D(x)) \le 4e^{-x}, \qquad D(x) = C \frac{s + \lambda^{-1}}{\kappa}[\sqrt{v_c(x + L_0)} + x + L_0], \tag{52}$$

for $x \ge 1$ with $v_c \ge C(x + L_0)$. If $v_c \ge 4CL_0$, take $x_* = v_c/(2C)$. Integration only over the admissible range gives

$$\mathbb{E}Z \le D(1) + \int_1^{x_*} 4e^{-x} D'(x)\,dx + \mathbb{E}[(Z - D(x_*))_+]. \tag{53}$$

Here $D'(x) = C(s + \lambda^{-1})\kappa^{-1}[\frac{1}{2}\sqrt{v_c/(x + L_0)} + 1]$, so the first two terms are at most $C(s + \lambda^{-1})[\sqrt{v_c L_0} + L_0]/\kappa$. For the last term, $0 \le \tau_c^\circ \le Y_{(n)}$, $\|Y_{(n)}\|_2 \le CL_0/\lambda$, and $s = \lambda^{-1}\log(n/v_c) \le L_0/\lambda$. Cauchy–Schwarz gives

$$\mathbb{E}[(Z - D(x_*))_+] \le \frac{CL_0}{\lambda} e^{-x_*/2} \le \frac{CL_0}{\lambda n}, \tag{54}$$

because $x_*/2 \ge L_0$. The identities $\kappa = z v_c = nze^{-z} \le n/e$ and $(s + \lambda^{-1})\sqrt{v_c} = \lambda^{-1}\sqrt{n}(1 + z)e^{-z/2} \le C\sqrt{n}/\lambda$, together with $v_c \ge 4CL_0$, now yield $\mathbb{E}Z \le C\sqrt{n}L_0/(\lambda\kappa)$.

If $v_c < 4CL_0$, use $\mathbb{E}Z \le (z + H_n)/\lambda$ instead. For $z \le 2L_0$, one has $v_c z(z + H_n) \le CL_0^3 \le C\sqrt{n}L_0$; the last inequality uses $\sup_{n \ge 1} L_0^2/\sqrt{n} < \infty$. For $z > 2L_0$, $v_c = ne^{-z} \le \frac{1}{2}e^{-z/2}$ and $H_n \le CL_0 < Cz$, so the same product is bounded by a constant. Dividing by $\lambda\kappa = \lambda v_c z$ proves the same estimate. Thus, without any survivor assumption,

$$\mathbb{E}|\tau_c^\circ - s| \le C \frac{\sqrt{n}\log(2n)}{\lambda\kappa}. \tag{55}$$

Transfer of the expectation. On $\mathcal{C}_n$, the two first-passage times coincide. On $\mathcal{C}_n^c$, both belong to $[0, Y_{(n)}]$, and $\mathcal{C}_n^c$ is independent of the residual lifetimes. Therefore

$$\begin{aligned}\mathbb{E}|\tau_c - \tau_c^\circ| &\le \mathbb{E}[Y_{(n)}\mathbf{1}_{\mathcal{C}_n^c}] \\ &= \mathbb{P}(\mathcal{C}_n^c)\mathbb{E}Y_{(n)} \\ &\le \frac{nH_n}{\lambda}\epsilon_N \le \frac{Cn\log(2n)}{\lambda}\epsilon_N.\end{aligned} \tag{56}$$

Using (48), (49), (55), and the triangle inequality, we conclude that

$$\begin{aligned}|\mathbb{E}\tau_c - \bar{s}| &\le \mathbb{E}|\tau_c - \tau_c^\circ| + \mathbb{E}|\tau_c^\circ - s| + |s - \bar{s}| \\ &\le C\frac{\sqrt{n}\log(2n)}{\lambda\bar{\kappa}} + \frac{Cn\log(2n)}{\lambda}\epsilon_N.\end{aligned}$$

In the last line we used $q_B\epsilon_N \le \epsilon_N$ and $1 \le Cn\log(2n)$ to absorb the deterministic shift $|s - \bar{s}|$ into the second term. Finally, take $n = n_\ell$ and $c = C_{\mathrm{th}}/2$, so that $\tau_c = \sigma_\ell$, $\bar{s} = \bar{\sigma}_\ell$, and $\bar{\kappa} = \kappa_\ell$. The probability and expectation estimates are exactly those in Lemma 1. □

B.3.4 Concentration for fixed multi-repacking blocks

For the crossing estimate, fix in advance, independently of all completion processes, a collection $\mathcal{B}$ of $m^{\ell+1}$ original replicas. In the synchronized-anchor model, the aggregate process and first-passage time of this deterministic analytical block are

$$Z_{\mathcal{B}}(u) := \sum_{j=1}^{n_{\ell,m}} \tilde{Z}_j(u), \qquad \sigma_{\mathcal{B}} := \inf\{u \ge 0 : Z_{\mathcal{B}}(u) \le \theta_{\ell,m}\}.$$

Lemma 13 (Concentration for a fixed multi-repacking block)**.** *For the deterministic block $\mathcal{B}$ fixed above, under the condition of Lemma 11, there is a constant C such that, for every $x \ge 1$, whenever*

$$n_{\ell,m}e^{-\lambda\bar{\sigma}_{\ell,m}} \ge C[x + \log(2n_{\ell,m})],$$

write $L_{\ell,m}(x) := x + \log(2n_{\ell,m})$. Then

$$\begin{aligned}\mathbb{P}\Bigg(|\sigma_{\mathcal{B}} - \bar{\sigma}_{\ell,m}| > \frac{C(\bar{\sigma}_{\ell,m} + \lambda^{-1})}{\kappa_{\ell,m}}\left[\sqrt{n_{\ell,m}e^{-\lambda\bar{\sigma}_{\ell,m}}L_{\ell,m}(x)} + L_{\ell,m}(x)\right] \\ + \frac{C}{\lambda}\epsilon_N\Bigg) \\ \le 4e^{-x} + n_{\ell,m}\epsilon_N.\end{aligned}$$

The first passage is controlled on $[0, \infty)$, not only near the mean crossing. Without the survivor lower bound,

$$|\mathbb{E}[\sigma_{\mathcal{B}}] - \bar{\sigma}_{\ell,m}| \le C\frac{\sqrt{n_{\ell,m}}\log(2n_{\ell,m})}{\lambda\kappa_{\ell,m}} + \frac{Cn_{\ell,m}\log(2n_{\ell,m})}{\lambda}\epsilon_N. \tag{57}$$

The block is fixed independently of service. Algorithm-selected groups will instead be handled by simultaneous original-replica path events.

Proof. Apply the general n, c proof in Appendix B.3.3 with $n = n_{\ell,m}$ and $c = \theta_{\ell,m}$. The margin in Lemma 11 gives $\lambda c/n = \rho_{\ell,m} \le 1 - \eta$. Lemma 12 couples the entire block with failure probability at most $n\epsilon_N$. Equations (40) and (47) control the stationary path and exclude every early crossing. The finite-batch comparisons (48)– (50), now with $\bar{s} = \bar{\sigma}_{\ell,m}$ and $\bar{\kappa} = \kappa_{\ell,m}$, give the stated probability radius and survivor condition. For the expectation, (55) already treats both large and small survivor counts. Add the independent coupling-failure cost in (56) and the $C\epsilon_N/\lambda$ deterministic shift to obtain (57). This transfers the full proof, including its terminal-tail bound, without integrating a constant failure probability over an unbounded interval. □

B.3.5 SIMULTANEOUS ORIGINAL-REPLICA BOUNDS

Write $L = \log(8RB)$, $q(u) = (B-1)e^{-\lambda u}$, $m^\circ(u) = q(u)(u+\lambda^{-1})$, and $N_i(u) = \sum_{j=1}^{B-1} \mathbf{1}_{\{\tilde{Y}_{ij} > u\}}$. Apply the coupling and the joint count–path construction above with $n = B-1$ and $x = 4L$ to each original replica. Let $\mathcal{G}$ be the intersection of these stationary path/count events and all anchor-coupling events. A union bound gives

$$\mathbb{P}(\mathcal{G}^c) \le (8RB)^{-3} + R(B-1)\epsilon_N, \tag{58}$$

on which, simultaneously for every $i \in [R]$ and $u \ge 0$,

$$|Z_i(u) - m^\circ(u)| \le C(u + \lambda^{-1})[\sqrt{q(u)L} + L], \tag{59}$$

$$|N_i(u) - q(u)| \le C[\sqrt{q(u)L} + L]. \tag{60}$$

No union bound over algorithm-selected groups is needed: these inequalities can be summed over any ancestry after the path is fixed. For actual asynchronous anchors, let $\mathcal{A} = \{s_{\max} - s_{\min} \le T_{\text{span}}\}$ and $\mathcal{H} = \mathcal{G} \cap \mathcal{A}$. Lemma 5 yields

$$\mathbb{P}(\mathcal{H}^c) \le 2(8RB)^{-3} + R(B-1)\epsilon_N. \tag{61}$$

All later good-event arguments are pathwise intersections, not conditional laws. Thus the bounds remain valid at the random local times $t - s_i$, despite dependence between anchors and anchor ages.

B.3.6 UNIFORM SMALLNESS OF THE CROSSING WINDOWS

The coarse width estimates used to sum occupancy errors do not by themselves show that each window is uniformly narrow. The following consequence of the variance-sensitive radius supplies this separate fact for both hierarchies.

Lemma 14 (Uniform crossing-window bound)**.** *Put $L = \log(8RB)$, $q(u) = (B-1)e^{-\lambda u}$, $m^\circ(u) = q(u)(u+\lambda^{-1})$, and $\bar{m}(u) = q(u)(u+\mu_A)$, and retain the ancestral load Z_i from Section 4.1. Write $N_i(u) := \sum_{j=1}^{B-1} \mathbf{1}_{\{\tilde{Y}_{ij} > u\}}$ for its survivor count. For $0 < a \le b$ with $\lambda b/(B-1) \le 1 - \eta$, denote the stationary and finite-batch decreasing crossings at threshold $c \in [a,b]$ by $\bar{\sigma}^\circ(c)$ and $\bar{\sigma}(c)$. Fix a sufficiently large window multiplier A, independently of the survivor-cutoff constant, and define*

$$d(c) := A\frac{(\bar{\sigma}^\circ(c) + \lambda^{-1})[\sqrt{q(\bar{\sigma}^\circ(c))L} + L]}{-(m^\circ)'(\bar{\sigma}^\circ(c))} + \frac{A}{\lambda}\epsilon_N. \tag{62}$$

For N/B sufficiently large,

$$\lambda \sup_{a \le c \le b} d(c) \le C\left[\sqrt{\frac{L}{q(\bar{\sigma}(a))}} + \frac{L}{q(\bar{\sigma}(a))} + \epsilon_N\right]. \tag{63}$$

The constant here depends on the fixed A and regularity margin, but not on the cutoff constant or the thresholds. If $q(\bar{\sigma}(a)) \ge M_\eta L$ with M_η sufficiently large, then on the simultaneous original-replica path and coupling event, every first passage at $c \in [a,b]$ is within $d(c)$ of $\bar{\sigma}(c)$, and $Z_i(u) > c$ for all $0 \le u < \bar{\sigma}(c) - d(c)$.

Proof. The stationary crossing margin gives $\lambda\bar{\sigma}^\circ(c) \ge z_\eta > 0$. Moreover, $-(m^\circ)'(s) = \lambda s q(s)$, so the dimensionless stationary part of $d(c)$ is exactly

$$A\frac{1 + \lambda\bar{\sigma}^\circ(c)}{\lambda\bar{\sigma}^\circ(c)}\left[\sqrt{\frac{L}{q(\bar{\sigma}^\circ(c))}} + \frac{L}{q(\bar{\sigma}^\circ(c))}\right].$$

The prefactor is uniformly bounded. The finite-batch comparison (48), applied to one original replica at threshold c, gives $\lambda|\bar{\sigma}^\circ(c) - \bar{\sigma}(c)| \le C\epsilon_N$ uniformly in c. Since $\bar{\sigma}(c) \le \bar{\sigma}(a)$ and q decreases, $q(\bar{\sigma}^\circ(c)) \ge c_\eta q(\bar{\sigma}(a))$ once N/B is large enough. Taking the supremum proves (63).

For the path assertion, use $x = 4L$ and $n = B-1$ in (47); then $x + \log(2n) \le 5L$. Choose A to dominate its radius and the finite-batch shift. The global no-early-crossing argument (45) applies on the same path event at every threshold, so no union bound over c is required. The preceding comparison of survivor counts verifies its hypothesis when M_η is large enough. □

In particular, fix a multiplier J and $\delta > 0$. First fix A, J, δ, then choose the survivor constant M_η, and finally take N/B sufficiently large. Equation (63) then gives $J\lambda d(c) \le \delta$ uniformly for $c \in [a, b]$. Indeed, the first two terms are bounded by $M_\eta^{-1/2} + M_\eta^{-1}$, and $\epsilon_N \le e^{-(N-B)/B}$. The same regularity comparison gives

$$c_\eta \lambda \le \frac{-\bar{m}'(\bar{\sigma}(c))}{c} = \lambda - \frac{1}{\mu_A + \bar{\sigma}(c)} \le \lambda. \tag{64}$$

Thus the load error obtained by multiplying a width by the crossing slope is also uniformly small relative to its threshold. All fixed window, endpoint, and no-early-crossing multipliers below are chosen before the survivor constants; a finite number of such requirements can be met by taking their maximum. None of these choices uses a lower bound on $C_{\mathrm{th}}\lambda/B$ or a bounded anchor span.

B.3.7 Persistent readiness and exceptional events

First passage alone does not ensure simultaneous packing feasibility, because cache can grow between completion jumps. The following stronger consequence of the same path bound will be used by both hierarchies.

Lemma 15 (Persistent upper endpoint)**.** *Under Lemma 14, take $c \in [a, b]$, $w = \sup_{a \le v \le b} d(v)$, and $q(\bar{\sigma}(a)) \ge ML$. There are fixed $K, M_0 \ge 1$ depending only on the regularity margin and window multiplier such that, after choosing M sufficiently large and then N/B sufficiently large, the endpoint $r = \bar{\sigma}(c) + Kw$ satisfies $\lambda K w \le \log 2$. On $\mathcal{G}$, for every $D \ge 0$ with $q(r + D) \ge M_0 L$,*

$$Z_i(u) \le c \qquad (i \in [R],\ r \le u \le r + D). \tag{65}$$

Moreover, for every $d \ge 0$,

$$q(r + d) \ge \tfrac{1}{2} q(\bar{\sigma}(c) + d). \tag{66}$$

Any prescribed fixed multiple of λw can be made arbitrarily small by the same order of constant choices.

Proof. Put $s = \bar{\sigma}^\circ(c)$ and $\kappa^\circ = -(m^\circ)'(s)$. The width definition and finite-batch comparison give

$$|s - \bar{\sigma}(c)| \le Cw, \qquad (s + \lambda^{-1})\sqrt{q(s)L} \le C\kappa^\circ w.$$

In a fixed neighborhood $|u - s| \le a_\eta/\lambda$, (42) bounds the mean slope below by $b_\eta \kappa^\circ$ and the path radius above by $C\kappa^\circ w$. Choose K larger than the finite-batch shift constant so that $r \ge s$ and the mean drop from s to r dominates this radius, then use (63) to choose M and N/B so that r lies in that neighborhood and $\lambda K w \le \log 2$. Taking $M \ge 2M_0$ also gives $q(r) \ge q(\bar{\sigma}(c))/2 \ge ML/2 \ge M_0 L$. On $q(u) \ge M_0 L$, absorb the additive L in (59) and set $\mathcal{U}(u) = m^\circ(u) + C(u + \lambda^{-1})\sqrt{q(u)L}$. The preceding comparisons yield

$$\mathcal{U}(r) - c \le -b_\eta \kappa^\circ (r - s) + C\kappa^\circ w \le 0.$$

Throughout $u \ge r$, the regular branch has $\lambda u \ge c_\eta > 0$. Since $-(m^\circ)'(u) = \lambda u q(u)$ and the absolute derivative of $(u + \lambda^{-1})\sqrt{q(u)L}$ is at most $C(1 + \lambda u)\sqrt{q(u)L}$, a sufficiently large M_0 gives

$$\mathcal{U}'(u) < 0 \qquad (u \ge r,\ q(u) \ge M_0 L). \tag{67}$$

This proves (65). Finally, $q(r + d) = q(\bar{\sigma}(c) + d)e^{-\lambda K w}$ proves (66); uniform smallness follows directly from Lemma 14. □

Lemma 16 (Exceptional-event contribution)**.** *Put $Y_* := \max_{i,j} \tilde{Y}_{ij}$ and $V = T_{\mathrm{span}} + \log(2RB)/\lambda$. For any policy considered here, any fixed $C_0 > 0$, and any deterministic $0 \le D_0 \le C_0 RV$, let $\mathcal{C}$ be the simultaneous anchor-coupling event, and let $\mathcal{F}$ be any additional failure event with $\mathbb{P}(\mathcal{F} \cap \mathcal{A}) \le p$. Then*

$$\mathbb{E}[(T_{\mathrm{tail}} + D_0)\mathbf{1}_{\mathcal{A}^c \cup \mathcal{C}^c \cup \mathcal{F}}] \le CRV\left[(8RB)^{-3/2} + R(B-1)\epsilon_N + \sqrt{p}\right]. \tag{68}$$

Here C may also depend on C_0. For synchronized anchors, take $\mathcal{A}$ to be the whole space, use $V = \log(2RB)/\lambda$, and omit the span-failure term.

Proof. Service-preserving migration gives the pathwise bound

$$0 \le T_{\mathrm{tail}} \le R(s_{\max} - s_{\min} + Y_*), \qquad \|Y_*\|_2 \le C\log(2RB)/\lambda. \tag{69}$$

The second bound follows by integrating $\mathbb{P}(Y_* > t) \le \min\{1, R(B-1)e^{-\lambda t}\}$. On $\mathcal{A}^c$, Lemma 5 and Cauchy–Schwarz bound the expectation by $CRV(8RB)^{-3/2}$. On $\mathcal{A} \cap \mathcal{F}$, the same argument with $T_{\mathrm{tail}} \le R(T_{\mathrm{span}} + Y_*)$ gives $CRV\sqrt{p}$. On $\mathcal{A} \cap \mathcal{C}^c$, use that $\mathcal{C}^c$ depends only on pre-anchor histories and is independent of Y_*; its cost is at most $CRV\mathbb{P}(\mathcal{C}^c) \le CRVR(B-1)\epsilon_N$. The deterministic D_0 obeys the same bounds after enlarging C. The synchronized case has zero span. No failure probability is integrated over an unbounded time interval. □

B.4 Pairwise occupied time

B.4.1 Synchronous pairwise bound and complete release counts

An additional superscript sync denotes the synchronized-anchor version of an occupied time. For arbitrary R, define

$$K_2 := \lfloor \log_2 R \rfloor, \qquad \nu_\ell := \left\lfloor \frac{R}{2^{\ell+1}} \right\rfloor, \qquad q_h := R - \sum_{\ell=0}^{h-1} \nu_\ell. \tag{70}$$

for $0 \le \ell < K_2$ and $0 \le h \le K_2$. In the nominal complete hierarchy, ν_ℓ counts level-ℓ repacking releases. The remaining q_h physical labels consist of $\lfloor R/2^h \rfloor$ level-h blocks and at most one unmatched block at each lower level. More explicitly,

$$\begin{aligned} q_h &= \left\lfloor \frac{R}{2^h} \right\rfloor + \sum_{\ell=0}^{h-1} \left(\left\lfloor \frac{R}{2^\ell} \right\rfloor - 2 \left\lfloor \frac{R}{2^{\ell+1}} \right\rfloor \right) \\ &\le \frac{R}{2^h} + h. \end{aligned} \tag{71}$$

These counts agree with the actual controlled hierarchy on the good events proved below. All releases not counted below the cut, including natural completion of unmatched lower-level blocks, are assigned to the q_h frontier labels. For $R = 2^{K_2}$, the formulas reduce to $\nu_\ell = R/2^{\ell+1}$ and $q_h = R/2^h$.

Proposition 17 (Tail occupied replica time under synchronous pairwise repacking)**.** *Suppose $B \ge 2$, $R \ge 2$, the synchronized-anchor model of Section 4.1 is used, and repacking is checked immediately after every trajectory-completion and repacking event. Assume $C_{\mathrm{th}}\lambda/B \le 1 - \eta$ for a fixed $\eta > 0$. Choose an integer $0 \le h \le K_2$. If $h \ge 1$, assume*

$$(B-1)e^{-\lambda\bar{\sigma}_h} \ge C\log(8RB), \tag{72}$$

where the survivor constant is sufficiently large. Then, for N/B sufficiently large,

$$\begin{aligned} &\left| \mathbb{E}[T_{\mathrm{tail}}^{\mathrm{pair,sync}}] - \sum_{\ell=0}^{h-1} \nu_\ell \bar{\sigma}_\ell - \frac{H_{R(B-1)}}{\lambda} \right| \\ &\le C \Bigg\{ \sum_{\ell=0}^{h-1} \frac{\nu_\ell \sqrt{2^\ell B}\log(8RB)}{\lambda\kappa_\ell} + \mathbf{1}_{\{q_h>1\}} \frac{q_h}{\lambda} \left[1 + \log\left(\frac{RB}{q_h} \right) \right] \\ &\qquad + \frac{R\log(2RB)}{\lambda}(8RB)^{-3/2} + \frac{R^2(B-1)\log(2RB)}{\lambda}\epsilon_N \Bigg\}. \end{aligned} \tag{73}$$

Moreover, under the stochastic asymptotic regime of the Scaling assumptions paragraph,

$$\mathbb{E}[T_{\mathrm{tail}}^{\mathrm{pair,sync}}] = \frac{1}{\lambda} \sum_{\ell=0}^{K_2-1} \nu_\ell \Phi(\rho_\ell) + o(R/\lambda). \tag{74}$$

Proof. Fix $0 \le h \le K_2$ satisfying (72) if $h > 0$. Work in the synchronized-anchor model with common local origin zero. Retain the ancestral loads Z_i and survivor counts N_i defined above. For a level-ℓ merge, put

$$c_\ell := \frac{C_{\mathrm{th}}}{2^{\ell+1}}.$$

A level-ℓ physical replica consists of 2^ℓ disjoint ancestral groups. Hence it has load at most $C_{\mathrm{th}}/2$ whenever every one of its ancestral processes is at most c_ℓ.

The crossing function in Lemma 14 satisfies $\bar{\sigma}(c_\ell) = \bar{\sigma}_\ell$. Set $L = \log(8RB)$ and use $x = 4L$ for the original-replica path event. For $\ell < h$, define

$$w_\ell := \sup_{c_{\ell+1} \le c \le c_\ell} d(c), \tag{75}$$

using (62). The cut condition guarantees its survivor requirement, since $c_{\ell+1} = c_\ell/2$. Lemma 14 therefore gives the variance-sensitive estimate

$$\lambda w_\ell \le C\left[\sqrt{\frac{L}{q(\bar{\sigma}_{\ell+1})}} + \frac{L}{q(\bar{\sigma}_{\ell+1})} + \epsilon_N\right], \qquad q(u) = (B-1)e^{-\lambda u}. \tag{76}$$

For summing errors, the crossing equation and $\kappa_\ell = 2^\ell[-((B-1)e^{-\lambda u}(\mu_A + u))']_{u=\bar{\sigma}_\ell}$ give

$$w_\ell \le C\frac{\sqrt{2^\ell B}\log(8RB)}{\lambda\kappa_\ell} + \frac{C}{\lambda}\epsilon_N. \tag{77}$$

It is (76), rather than this coarse sum bound, that makes every fixed multiple of a controlled width small compared with $1/\lambda$. In particular, the regular crossing lower bound allows $w_\ell \le \bar{\sigma}_\ell/3$, and the deterministic level gap allows the separation used below, with the constant choices ordered as in Lemma 14.

Work on the common event $\mathcal{G}$ of (58); denote it also by $\mathcal{G}_{\mathrm{s}}$ in this proof. Its probability is at least

$$1 - (8RB)^{-3} - R(B-1)\epsilon_N. \tag{78}$$

Apply Lemma 15 with $[a, b] = [c_{\ell+1}, c_\ell]$ and define

$$r_\ell := \bar{\sigma}_\ell + K_\eta w_\ell. \tag{79}$$

Dividing consecutive crossing equations gives $\bar{\sigma}_{\ell+1} - \bar{\sigma}_\ell > \lambda^{-1}\log 2$. Choose K_η first, then the cutoff constant and N/B, so that

$$r_\ell < \bar{\sigma}_{\ell+1} - w_{\ell+1} \quad (\ell < h-1), \qquad r_\ell < \bar{\sigma}_{\ell+1} \quad (\ell < h). \tag{80}$$

The same lemma gives the useful shifted-cut comparison

$$q(r_\ell + d) \ge \tfrac{1}{2}q(\bar{\sigma}_\ell + d), \qquad d \ge 0. \tag{81}$$

The synchronous cut ensures $q(r_\ell) \ge CL$. Thus $Z_i(r_\ell) \le c_\ell$, whereas Lemma 14 gives $Z_i(u) > c_\ell$ for $u < \bar{\sigma}_\ell - w_\ell$. Every level-$\ell$ input, with 2^ℓ originals, therefore exceeds $C_{\mathrm{th}}/2$ before the lower endpoint and is at most that threshold at r_ℓ. The count bound (60) and $q(u) \ge q(r_\ell) \ge CL$ also ensure that every original ancestry remains nonempty throughout this controlled interval. No natural release occurs below the cut.

It remains to check the combined slot constraint, rather than merely the individual bounds on the two replicas. Put $q(u) = (B-1)e^{-\lambda u}$ and $L = \log(8RB)$. Summing (37) over the $2^{\ell+1}$ original ancestral groups of any prospective level-ℓ pair gives, on the same simultaneous event,

$$N_{\mathrm{act}}(r_{\mathrm{src}}) + N_{\mathrm{act}}(r_{\mathrm{dst}}) \le 2^{\ell+1}\Big(q(r_\ell) + C[\sqrt{q(r_\ell)L} + L]\Big). \tag{82}$$

Since $r_\ell \ge \bar{\sigma}_\ell$, the crossing equation gives

$$2^{\ell+1}q(r_\ell) \le 2^{\ell+1}q(\bar{\sigma}_\ell) = \frac{C_{\mathrm{th}}}{\mu_A + \bar{\sigma}_\ell} \le B(1 - c_\eta)$$

for N/B sufficiently large, where the last inequality follows from $C_{\mathrm{th}}\lambda/B \le 1 - \eta$ and $\lambda\mu_A \to 1$. The deterministic crossing gap and the small-width property established above give $r_\ell < \bar{\sigma}_{\ell+1}$, so the cutoff ensures $q(r_\ell) \ge CL$; after enlarging its fixed constant, the deviation in (82) is smaller than the remaining margin. Thus the combined active count is at most B. Together with the load bound $C(r_{\mathrm{src}}) + C(r_{\mathrm{dst}}) \le C_{\mathrm{th}}$, this proves that every remaining pair is feasible, and the event-driven rule pairs it by time r_ℓ.

The level separation (80) now ensures that each controlled input is formed before the next level's lower endpoint. If r_ℓ is not an event, feasibility there also holds at the most recent event after formation: between events loads increase and active counts are constant. Exhaustive event closure therefore processes every remaining feasible pair by r_ℓ. It follows inductively that, for every $\ell < h$, exactly $\nu_\ell = \lfloor R/2^{\ell+1} \rfloor$ replicas are released by level-ℓ repacking and all these release times lie between $\bar{\sigma}_\ell - w_\ell$ and r_ℓ. No separation condition is imposed between levels $h-1$ and h: higher-level operations may start as soon as their inputs have been formed, but they cannot affect the remaining lower-level pairs. At each controlled level, an odd number of inputs leaves one unmatched block. It remains occupied until natural completion and does not prevent the other inputs from pairing.

Cut the actual binary merge forest retrospectively at level h. Its frontier $\mathcal{F}_h$ consists of the $\lfloor R/2^h \rfloor$ level-h blocks and all unmatched blocks at lower levels. These disjoint ancestral sets partition the originals and have cardinality q_h by (71); they need not coexist at one physical time. Every release not counted in a level-ℓ round with $\ell < h$ belongs to the forest rooted at this frontier. Denote the sum of these q_h release times by T_{high}.

We assign distinct trajectory witnesses to all of these releases, even when the merge forest is adaptive. Attach to each frontier block the trajectory with the latest completion deadline among its ancestors. At a later merge, both children are nonempty, so both attached deadlines are at least the merge time. Assign the trajectory with the smaller deadline to the current release, and pass the one with the larger deadline to the parent. At a natural completion, assign the remaining trajectory; its deadline equals that release time. This is an analytical assignment and does not change the physical source or destination. Each initial frontier marker is used exactly once, so the q_h witnessing trajectories are distinct and each release time is at most its witness's deadline. Let $\tilde{Y}_{[1]} \geq \cdots \geq \tilde{Y}_{[R(B-1)]}$ be the decreasing order statistics of all residual lifetimes. On the simultaneous good event, the assignment gives

$$\tilde{Y}_{[1]} \leq T_{\mathrm{high}} \leq \sum_{j=1}^{q_h} \tilde{Y}_{[j]}. \tag{83}$$

The lower bound holds because the last natural release is among these releases and occurs at $\tilde{Y}_{[1]}$. In particular, $T_{\mathrm{high}} = \tilde{Y}_{[1]}$ when $q_h = 1$. For P independent $\mathrm{Exp}(\lambda)$ variables,

$$\mathbb{E}\left[\sum_{j=1}^{q} \tilde{Y}_{[j]}\right] = \frac{q}{\lambda}(1 + H_P - H_q) \leq \frac{q}{\lambda}\left[1 + \log\left(\frac{P}{q}\right)\right].$$

Applying the order-statistic identity with $P = R(B-1)$ gives

$$0 \leq \mathbb{E}[(T_{\mathrm{high}} - \tilde{Y}_{[1]})\mathbf{1}_{\mathcal{G}_{\mathrm{s}}}] \leq \mathbf{1}_{\{q_h > 1\}} \frac{q_h}{\lambda}\left[1 + \log\left(\frac{RB}{q_h}\right)\right]. \tag{84}$$

The controlled sum is $\sum_{\ell<h} \nu_\ell \bar{\sigma}_\ell$. Its pathwise error is at most $C \sum_{\ell<h} \nu_\ell w_\ell$, and the frontier accounts for all remaining releases, including natural completion of unmatched blocks.

For the complement, apply Lemma 16 in the synchronized model. The additional path failure has probability at most $(8RB)^{-3}$. The deterministic controlled sum is zero if $h = 0$ and otherwise at most $CR\log(2RB)/\lambda$ by the survivor cut. The lemma controls both this lost sum and the occupied time; the same maximum-lifetime bound controls the contribution of $\tilde{Y}_{[1]}$ on the complement. Hence the two exceptional terms are $CR\log(2RB)(8RB)^{-3/2}/\lambda$ and $CR^2(B-1)\log(2RB)\epsilon_N/\lambda$. Combining these estimates with (77) and (84), and using $\mathbb{E}[\tilde{Y}_{[1]}] = H_{R(B-1)}/\lambda$, gives (73). The finite-batch part of the width sum is absorbed into the coupling term since $\sum_{\ell<h} \nu_\ell \leq R$. In particular, no constant failure probability is integrated over an unbounded time interval. For the asymptotic statement (74), let $L := \log(8RB)$ and choose $h = \lceil 3\log L/\log 2 \rceil$. The stated regime gives $h < K_2$ for sufficiently large B and $2^{-h} = \Theta(L^{-3})$. The proportional threshold and the crossing equation imply $(B-1)e^{-\lambda\bar{\sigma}_h} = \Theta(B2^{-h}/(h+1))$, so (72) holds. After division by R, the concentration sum is $O(L/(\lambda\sqrt{B}))$ and the frontier contribution is $O((2^{-h} + h/R)L/\lambda) = o(1/\lambda)$. The harmonic and exceptional-event terms are also $o(1/\lambda)$. Lemma 10 replaces each controlled crossing by $\Phi(\rho_\ell)/\lambda$ with aggregate error $O(R\epsilon_N/\lambda)$, and the omitted deterministic levels have total weight $O((h+1)2^{-h})$. This proves (74). □

B.4.2 A CONVEX LOWER BOUND FOR HETEROGENEOUS LOCAL AGES

Lemma 18 (A convex lower bound for the residual-load profile)**.** *Let $f(z) = (1+z)e^{-z}$ for $z \geq 0$. The function*

$$\underline{f}(z) = \begin{cases} (7-2z)e^{-2}, & 0 \leq z \leq 2, \\ (1+z)e^{-z}, & z \geq 2 \end{cases} \tag{85}$$

is decreasing and convex, satisfies $\underline{f} \leq f$, and obeys $\underline{f}(3/2) = 4e^{-2} > 1/2$.

The lower bound permits Jensen's inequality to control heterogeneous local ages, including ages where the original load profile is not convex. The proof joins a tangent line to the profile at $z = 2$ and verifies that the result is globally below the profile, decreasing, and convex. This supplies the average-anchor lower bound used in the asynchronous pairwise proof.

Proof. The two branches have the same value $3e^{-2}$ and derivative $-2e^{-2}$ at 2. The first is linear, while $f''(z) = (z-1)e^{-z} \geq 0$ for $z \geq 2$; their derivatives are negative. Thus $\underline{f}$ is convex and decreasing. To verify the lower bound on $[0,2]$, set $g(z) = f(z) - (7-2z)e^{-2}$. Then $g(0) = 1 - 7e^{-2} > 0$, $g(2) = 0$, and $g'(z) = 2e^{-2} - ze^{-z}$. Since ze^{-z} increases on $[0,1]$ and decreases on $[1,2]$, g' changes sign once from positive to negative in $(0,2)$ and vanishes again at 2. Hence the minimum of g on $[0,2]$ occurs at an endpoint and is zero. For $z \geq 2$ the two functions agree. The value at $3/2$ follows by substitution. □

B.4.3 FINITE-SYSTEM PART OF PROPOSITION 19

A positive analytical cutoff h must satisfy

$$(B-1)e^{-\lambda(\bar{\sigma}_h + T_{\mathrm{span}})} \geq C\log(8RB), \tag{86}$$

with a sufficiently large constant; $h = 0$ is always allowed. The explicit per-replica error is

$$\begin{aligned} \mathcal{E}_2(h) := {} & \frac{hT_{\mathrm{span}}}{R} + \frac{1}{R\lambda}\sum_{\ell=0}^{h-1} \frac{\nu_\ell \sqrt{2^\ell B}\log(8RB)}{\kappa_\ell} + \frac{1}{R\lambda}\sum_{\ell=h}^{K_2-1} \nu_\ell \Phi(\rho_\ell) \\ & + \frac{q_h}{R}\left[T_{\mathrm{span}} + \frac{1+\log(RB/q_h)}{\lambda}\right] + \frac{\epsilon_N}{\lambda} \\ & + \left[T_{\mathrm{span}} + \frac{\log(2RB)}{\lambda}\right]\left[(8RB)^{-3/2} + R(B-1)\epsilon_N\right]. \end{aligned} \tag{87}$$

Its terms account for same-level waiting, crossing fluctuations, omitted deterministic levels, remaining physical labels, the finite-batch mean correction, and exceptional events, respectively. It has units of time; the total error is $CR\mathcal{E}_2(h)$. When $R = 2^{K_2}$ and $h = K_2$ is admissible, $q_h = 1$ and the span and frontier contributions per replica are $O(K_2 T_{\mathrm{span}}/R + \log(RB)/(R\lambda))$. For general R, (71) accounts for unmatched blocks. Under the proportional threshold regime, the normalized concentration sum is $O(\log(RB)/(\lambda\sqrt{B}))$.

Proposition 19 (Tail occupied replica time under asynchronous pairwise repacking)**.** *Assume $B \geq 2$, $R \geq 2$, and $C_{\mathrm{th}}\lambda/B \leq 1 - \eta$ for a fixed $\eta > 0$. Use the actual finite-batch anchors, T_{span} from* (25)*, and the event-driven hierarchical pairwise rule. For N/B sufficiently large and every cutoff $0 \leq h \leq K_2$ satisfying* (86) *when $h > 0$,*

$$\left| \frac{\mathbb{E}[T_{\mathrm{tail}}^{\mathrm{pair}}]}{R} - \frac{1}{R\lambda}\sum_{\ell=0}^{K_2-1} \nu_\ell \Phi(\rho_\ell) \right| \leq C\mathcal{E}_2(h). \tag{88}$$

Under the stochastic asymptotic regime of the Scaling assumptions paragraph,

$$\mathbb{E}[T_{\mathrm{tail}}^{\mathrm{pair}}] = \frac{1}{\lambda}\sum_{\ell=0}^{K_2-1} \nu_\ell \Phi(\rho_\ell) + o(R/\lambda). \tag{89}$$

Proof. Fix $0 \leq h \leq K_2$ satisfying (86) if $h > 0$.

Use the common processes Z_i, N_i, q, m° and L from Appendix B.3.5, and work pathwise on $\mathcal{H} = \mathcal{G} \cap \mathcal{A}$. The asynchronous cut implies the synchronous cut and hence defines the same c_ℓ, w_ℓ, r_ℓ. By (80), the cut gives $q(r_\ell + T_{\mathrm{span}}) \geq CL$ after a fixed enlargement of its constant. Lemma 15 therefore yields

$$Z_i(u) \leq c_\ell, \qquad r_\ell \leq u \leq r_\ell + T_{\mathrm{span}}, \quad i \in [R],\ \ell < h. \tag{90}$$

The common no-early-crossing bound gives

$$Z_i(u) > c_\ell, \qquad 0 \leq u < \bar{\sigma}_\ell - w_\ell. \tag{91}$$

These simultaneous bounds apply at random local times without conditioning on the anchor vector.

Consider two level-ℓ replicas that are merged, and let $[x, y]$ be the smallest interval containing the anchor times of all their $2^{\ell+1}$ ancestral replicas. Before $x + \bar{\sigma}_\ell - Cw_\ell$, either one of the two replicas has not yet been formed or every ancestral process is greater than c_ℓ. In the latter case their combined load is greater than $2^{\ell+1}c_\ell = C_{\mathrm{th}}$, so no merge is feasible. At $t^+ = y + r_\ell$, on the other hand,

$$t^+ - s_i \in [r_\ell, r_\ell + y - x] \subseteq [r_\ell, r_\ell + T_{\mathrm{span}}]$$

for every relevant anchor s_i. Hence (90) shows that each of the two physical replicas has load at most $C_{\mathrm{th}}/2$. The combined slot constraint also holds. Indeed, applying (60) to every ancestral group and using that $q(u)$ and its deviation radius decrease over this interval gives

$$N_{\mathrm{act}}(r_{\mathrm{src}}) + N_{\mathrm{act}}(r_{\mathrm{dst}}) \leq 2^{\ell+1} \left\{ q(r_\ell) + C[\sqrt{q(r_\ell)L} + L] \right\} \leq B. \tag{92}$$

The last inequality follows exactly as in (82): $r_\ell \geq \bar{\sigma}_\ell$, the mean count is at most $B(1 - c_\eta)$, and the survivor condition absorbs the deviation. The local formation induction below shows that both replicas already exist by t^+. If t^+ is not itself an event time, look back to the most recent completion or repacking event after both replicas were formed. Loads and active counts can only increase and remain constant, respectively, between that event and t^+, so both feasibility inequalities at t^+ also hold at the preceding event. The repeated event-driven check therefore merges them no later than t^+. Thus every controlled merge satisfies

$$x + \bar{\sigma}_\ell - Cw_\ell \leq \tau_\ell^{\mathrm{merge}} \leq y + \bar{\sigma}_\ell + Cw_\ell. \tag{93}$$

We next justify the formation and candidate assertions used above. If G is an actual level-ℓ block in the event-driven merge tree, write

$$s_+(G) := \max_{i \in G} s_i.$$

We claim that a level-ℓ block is formed no later than $s_+(G) + r_{\ell-1}$. Level-zero blocks exist at their own anchors. For a level-one block, both of its children therefore exist by $s_+(G) + r_0$, and all their ancestral local times then belong to $[r_0, r_0 + T_{\mathrm{span}}]$. They are feasible by (90), and the active-count argument below shows that they are candidates. The most recent event after both children exist and before this endpoint therefore forms the block. Inductively, the two children of a level-ℓ block have already been formed before $s_+(G) + r_{\ell-1}$ because $r_{\ell-2} < r_{\ell-1}$. At that endpoint both children are persistently ready. If they were still unpaired, the repeated feasibility check would pair them (or pair either one elsewhere, in which case G would not be an actual later block). This proves the claim.

Consequently, every member of an actual level-ℓ pair whose latest ancestral anchor is y has been formed before $y + r_\ell$, as required in (93). In particular, if $U_\ell := s_{\max} + r_\ell$, all level-$\ell$ blocks not already promoted are formed and persistently ready by U_ℓ, and the repeated feasibility check pairs them by that time. The count bound (60), used at the right end of the interval in (86), also ensures that no ancestral group finishes naturally below the cutoff. Equation (92) verifies the combined slot constraint at each controlled pairing endpoint. Induction therefore shows that there are exactly

$$\nu_\ell = \left\lfloor \frac{R}{2^{\ell+1}} \right\rfloor$$

repacking releases at every $\ell < h$. An odd input count leaves one unmatched block at that level; all such blocks are retained in the frontier accounting below.

We now prove the lower and upper estimates separately. This is the step that replaces the signed carry identity.

Lower bound. Let v be an actual level-ℓ merge below the cutoff, let $\mathcal{B}_v$ be the set of its $2^{\ell+1}$ original ancestral replicas, and write

$$\overline{s}_{\mathcal{B}_v} := \frac{1}{2^{\ell+1}} \sum_{i \in \mathcal{B}_v} s_i.$$

We claim that, on $\mathcal{H}$,

$$t_v \geq \overline{s}_{\mathcal{B}_v} + \bar{\sigma}_\ell - C w_\ell. \tag{94}$$

Unlike (93), this estimate uses the average anchor rather than the earliest anchor.

Here is the deterministic and stochastic justification of the claim. Put

$$f(z) := (1+z)e^{-z},$$

and let $\bar{\sigma}_\ell^\circ$ be defined by

$$2^{\ell+1} m^\circ(\bar{\sigma}_\ell^\circ) = C_{\mathrm{th}}.$$

The finite-batch comparison already proved after (47) gives

$$|\bar{\sigma}_\ell^\circ - \bar{\sigma}_\ell| \leq \frac{C}{\lambda}\epsilon_N, \qquad -2^{\ell+1}(m^\circ)'(\bar{\sigma}_\ell^\circ) \asymp_\eta \kappa_\ell. \tag{95}$$

We now quantify the convexity bound needed for heterogeneous anchors, using Lemma 18. If $h > 0$, the survivor condition implies $B - 1 \geq C \log(8RB)$. After enlarging this fixed constant, $(1-\eta)B/(B-1) \leq 1 - \eta/2$. Hence the load margin gives $\rho_0 \leq (1-\eta/2)/2$ and $\rho_\ell \leq (1-\eta/2)/4$ for $\ell \geq 1$. Since $f(3/2) > 1/2$ and $f(2) > 1/4$, strict monotonicity and continuity of f yield a fixed $c_\eta > 0$ such that

$$f(\lambda\bar{\sigma}_0^\circ) = \rho_0 \leq \frac{1}{2} - c_\eta, \qquad \lambda\bar{\sigma}_0^\circ \geq \frac{3}{2} + c_\eta, \qquad \lambda\bar{\sigma}_\ell^\circ \geq 2 + c_\eta \quad (\ell \geq 1).$$

We claim that, if $M = 2^{\ell+1}$, $u_i \geq 0$, and $\overline{u} = M^{-1}\sum_i u_i \leq \bar{\sigma}_\ell^\circ$, then

$$\sum_{i=1}^{M} m^\circ(u_i) - C_{\mathrm{th}} \geq c_\eta \min\{-M(m^\circ)'(\bar{\sigma}_\ell^\circ)(\bar{\sigma}_\ell^\circ - \overline{u}),\ C_{\mathrm{th}}\}. \tag{96}$$

To prove this for $\ell = 0$, write $\lambda u_1 = z - d$ and $\lambda u_2 = z + d$, where $z = \lambda\overline{u}$ and $0 \leq d \leq z$. Direct calculation gives

$$\frac{f(z-d) + f(z+d)}{2} - f(z) = e^{-z}\big[(1+z)(\cosh d - 1) - d\sinh d\big] \geq 0$$

whenever $z \geq 3/2$. For $d = 0$ this is equality. For $d > 0$, the function $v\coth(v/2)$ increases because, with $a = v/2$, its derivative has the sign of $\sinh(a)\cosh(a) - a > 0$. Thus $d\coth(d/2) \leq z\coth(z/2) \leq 1 + z$; the last inequality is equivalent to $e^z \geq 1 + 2z$, which holds at $z = 3/2$ and persists since $e^z - 2 > 0$ thereafter. Since f is convex on $[1, \infty)$, its tangent at $\lambda\bar{\sigma}_0^\circ$ gives

$$f(z) - f(\lambda\bar{\sigma}_0^\circ) \geq -f'(\lambda\bar{\sigma}_0^\circ)(\lambda\bar{\sigma}_0^\circ - z).$$

This proves the slope term in (96). If instead $z < 3/2$, use the decreasing convex minorant $\underline{f}$ verified in Lemma 18. Jensen's inequality and $f \geq \underline{f}$ give

$$\frac{f(\lambda u_1) + f(\lambda u_2)}{2} \geq \underline{f}(z) \geq \underline{f}(3/2) \geq (1 + c_\eta) f(\lambda\bar{\sigma}_0^\circ),$$

after decreasing the fixed c_η if necessary, since $\underline{f}(3/2) > 1/2$ and $f(\lambda\bar{\sigma}_0^\circ) \leq 1/2 - c_\eta$. This proves the capped term.

For $\ell \geq 1$, Jensen's inequality gives $M^{-1}\sum_i f(\lambda u_i) \geq \underline{f}(\lambda\overline{u})$. If $\lambda\overline{u} \geq 2$, then $\underline{f} = f$ and the same tangent inequality applies on the convex branch. If $\lambda\overline{u} < 2$, monotonicity and $\rho_\ell \leq 1/4$ give

$$\begin{aligned} \underline{f}(\lambda\overline{u}) - f(\lambda\bar{\sigma}_\ell^\circ) &\geq f(2) - \rho_\ell \\ &\geq (12e^{-2} - 1)\rho_\ell. \end{aligned}$$

The coefficient is a fixed positive number. In all cases, use $m^\circ(u) = (B-1)f(\lambda u)/\lambda$ and multiply by $M(B-1)/\lambda$. The tangent bound becomes the slope term, and the relative bound becomes a fixed

positive fraction of C_{th}. Taking the smaller of these fixed constants proves (96) uniformly over all controlled levels.

At a prospective merge of $\mathcal{B}_v$, all its anchors have occurred. Set $u_i = t - s_i$ and $\overline{u} = t - \overline{s}_{\mathcal{B}_v}$. If $\overline{u} \le \bar{\sigma}_\ell^\circ$, then $u_i \le \bar{\sigma}_\ell^\circ + T_{\mathrm{span}}$. Condition (86) and (95), with the same harmless enlargement of C, therefore give

$$(B-1)e^{-\lambda u_i} \ge C\log(8RB) \qquad (i \in \mathcal{B}_v).$$

Thus the logarithmic term in the global path radius is absorbed by its square-root term. If $\sum_i m^\circ(u_i) \le 2C_{\mathrm{th}}$, Cauchy–Schwarz gives

$$\begin{aligned}\sum_{i\in\mathcal{B}_v} |Z_i(u_i) - m^\circ(u_i)| &\le C\sqrt{\log(8RB)\left(\sum_i m^\circ(u_i)\right)\left(\sum_i (u_i + \lambda^{-1})\right)} \\ &\le C\sqrt{C_{\mathrm{th}} 2^{\ell+1}(\bar{\sigma}_\ell^\circ + \lambda^{-1})L}. \end{aligned} \tag{97}$$

Here we used $\sum_i (u_i + \lambda^{-1}) = 2^{\ell+1}(\overline{u} + \lambda^{-1})$, and the survivor condition, which ensures that the logarithmic term has already been absorbed. Crucially, we now use the variance-sensitive definition of w_ℓ, not the coarse upper bound in (77). At the stationary crossing,

$$C_{\mathrm{th}} = 2^{\ell+1} q(\bar{\sigma}_\ell^\circ)(\bar{\sigma}_\ell^\circ + \lambda^{-1}), \qquad \kappa_\ell \asymp_\eta 2^\ell[-(m^\circ)'(\bar{\sigma}_\ell^\circ)].$$

The first-passage radius used to define w_ℓ therefore gives directly

$$\begin{aligned}\sqrt{C_{\mathrm{th}} 2^{\ell+1}(\bar{\sigma}_\ell^\circ + \lambda^{-1})L} &= 2^{\ell+1}(\bar{\sigma}_\ell^\circ + \lambda^{-1})\sqrt{q(\bar{\sigma}_\ell^\circ)L} \\ &\le C\kappa_\ell w_\ell. \end{aligned} \tag{98}$$

Indeed, up to a fixed factor, the expression before the inequality is the stationary first-passage radius at threshold c_ℓ multiplied by the aggregate crossing slope. This radius is one of those over which w_ℓ is maximized, and that slope is comparable to κ_ℓ. Choose first the fixed multiple of w_ℓ required to dominate (98). The earlier uniform estimate (76) applies here because (86); the asynchronous survivor condition is stronger. Moreover, (64) gives $\kappa_\ell w_\ell / C_{\mathrm{th}} \le \lambda w_\ell / 2$. Thus, by choosing the survivor constant after this fixed multiple and then taking N/B sufficiently large, we obtain

$$C\kappa_\ell w_\ell \le \frac{c_\eta}{2} C_{\mathrm{th}}, \tag{99}$$

where c_η is the fixed constant in (96). If instead $\sum_i m^\circ(u_i) > 2C_{\mathrm{th}}$, the relative form of (40), with the survivor constant chosen large enough, gives

$$\sum_i Z_i(u_i) \ge \left(1 - C\sqrt{\frac{L}{\min_i q(u_i)}}\right)\sum_i m^\circ(u_i) > C_{\mathrm{th}}.$$

To make the final comparison explicit, write $\Delta = \bar{\sigma}_\ell^\circ - \overline{u}$. Let $c_1, c_2 > 0$ and $C_1 > 0$ be fixed constants such that $-M(m^\circ)'(\bar{\sigma}_\ell^\circ) \ge c_1\kappa_\ell$, the coefficient in (96) is c_2, and (97) together with (98) bounds the total fluctuation by $C_1\kappa_\ell w_\ell$. First choose a fixed $D > 2C_1/(c_1c_2)$. The cutoff choice in (99) can then ensure $C_1\kappa_\ell w_\ell \le c_2 C_{\mathrm{th}}/2$. If $\Delta \ge D w_\ell$ and $\sum_i m^\circ(u_i) \le 2C_{\mathrm{th}}$, the deterministic gap is at least

$$c_2 \min\{c_1\kappa_\ell\Delta, C_{\mathrm{th}}\} \ge 2C_1\kappa_\ell w_\ell,$$

and therefore strictly exceeds the fluctuation. The case $\sum_i m^\circ(u_i) > 2C_{\mathrm{th}}$ was covered by the relative bound. Finally, the definition of w_ℓ absorbs the finite-batch shift $|\bar{\sigma}_\ell^\circ - \bar{\sigma}_\ell| \le w_\ell$. Consequently,

$$t < \overline{s}_{\mathcal{B}_v} + \bar{\sigma}_\ell - (D+1)w_\ell \quad \Longrightarrow \quad \sum_{i\in\mathcal{B}_v} Z_i(t - s_i) > C_{\mathrm{th}}.$$

Such a merge is infeasible, proving (94). The event $\mathcal{G}$ is simultaneous over every original replica and every local time, while $\mathcal{A}$ supplies the pathwise bound on the actual anchor span. Hence this argument remains valid for the random blocks selected by the algorithm; no independence between anchor times and anchor states, and no union bound over possible merge trees, is required.

For arbitrary R, the blocks merged at one level need not cover every original replica. We therefore cancel anchor terms locally at each merge. Give every current analytical block G a bookkeeping

anchor $\overline{s}_G = |G|^{-1}\sum_{i\in G} s_i$. If equal-size children A, B merge into P, then $\overline{s}_A + \overline{s}_B = 2\overline{s}_P$. Charge $\overline{s}_P$ to the current release and pass another copy of $\overline{s}_P$ to the parent. This preserves the sum of all assigned anchors, irrespective of which physical child is released.

Apply this averaging only to merges below level h. The resulting frontier $\mathcal{F}_h$ contains $\lfloor R/2^h \rfloor$ level-h blocks and every unmatched lower-level block, for a total of q_h labels. Above the cut, propagate these labels without changing their anchor values: at a later merge assign the released child's label to that release and pass the other to the survivor; at a natural completion assign the remaining label. Every frontier label is used exactly once. The associated release time t_G^{rel} is at least the formation time of G, hence at least $\overline{s}_G$. Conservation of the anchor sum gives the exact identity

$$\begin{aligned} T_{\mathrm{tail}}^{\mathrm{pair}} = &\sum_{\ell=0}^{h-1} \sum_{v:\mathrm{level}(v)=\ell} (t_v - \overline{s}_{\mathcal{B}_v}) \\ &+ \sum_{G\in\mathcal{F}_h} (t_G^{\mathrm{rel}} - \overline{s}_G). \end{aligned} \tag{100}$$

The last sum is nonnegative and includes natural completion of unmatched blocks. Applying (94) to the first sum therefore yields

$$T_{\mathrm{tail}}^{\mathrm{pair}} \geq \sum_{\ell=0}^{h-1} \nu_\ell \bar{\sigma}_\ell - C\sum_{\ell=0}^{h-1} \nu_\ell w_\ell. \tag{101}$$

Thus the lower bound pays no multiple of T_{span}, without requiring R to be divisible by 2^h.

Upper bound. We first remove the source–destination ambiguity by a bookkeeping relabeling. If two current representatives with anchors a and b merge at time t and the survivor is later released at time u, then

$$(t-a) + (u-b) = (t-b) + (u-a).$$

Thus swapping the two representative labels at a merge does not change total occupied time. Recursively make the child with the larger value of $s_+(G)$ the bookkeeping survivor and charge the other child at the current merge. The surviving bookkeeping representative of every block G then has anchor $s_+(G)$.

Continue the ancestral load and active count of a block counterfactually after migration, so that readiness does not depend on the physical host. Write

$$C_G(t) := \sum_{i\in G} Z_i(t - s_i), \qquad N_G(t) := \sum_{i\in G} N_i(t - s_i).$$

For a level-ℓ block G, let $a_\ell(G)$ be the first event time after its formation from which, through $U_\ell = s_{\max} + r_\ell$, both

$$C_G(u) \leq \frac{C_{\mathrm{th}}}{2}, \qquad N_G(u) \leq \frac{B}{2} \tag{102}$$

hold. We now verify that this time exists within the required window, rather than using load readiness alone. At $t_G := s_+(G) + r_\ell$, all ancestral local times lie in $[r_\ell, r_\ell + T_{\mathrm{span}}]$, so (90) gives the load inequality in (102). Since the survivor count is nonincreasing in local time, (60) gives

$$N_G(t_G) \leq 2^\ell \left\{ q(r_\ell) + C[\sqrt{q(r_\ell)L} + L] \right\} \leq \frac{B}{2}. \tag{103}$$

The last inequality is exactly one half of the calculation in (92); it uses $C_{\mathrm{th}}\lambda/B \leq 1 - \eta$ for the deterministic margin and the survivor condition to absorb the deviation. The count can only decrease after t_G. If t_G is not an event time, the most recent event after the formation of G has no larger load or count, because the load increases and the count remains constant between events. It follows that both inequalities in (102) hold from some event no later than t_G. Therefore the formation induction and (90) imply

$$a_\ell(G) \leq s_+(G) + \bar{\sigma}_\ell + Cw_\ell. \tag{104}$$

If S_v is the bookkeeping source charged at a level-ℓ merge v, define

$$I_v := \begin{cases} [a_\ell(S_v), t_v], & t_v > a_\ell(S_v), \\ \varnothing, & t_v \leq a_\ell(S_v). \end{cases}$$

Then (104) gives the one-sided inequality

$$t_v - s_+(S_v) \le \bar{\sigma}_\ell + Cw_\ell + |I_v|. \tag{105}$$

At a fixed level, the interiors of the nonempty intervals I_v are disjoint. Indeed, at the later left endpoint of two overlapping intervals, the two corresponding blocks still exist and are unpaired. By (102), each has load at most $C_{\mathrm{th}}/2$ and active count at most $B/2$. They are therefore a feasible same-level pair under both algorithmic constraints. The exhaustive event-driven check must merge at least one of them at that event, contradicting an interior overlap. This argument would not follow from load readiness alone. The no-early bound (91) and the upper merge window place every such interval inside

$$[s_{\min} + \bar{\sigma}_\ell - Cw_\ell,\ s_{\max} + \bar{\sigma}_\ell + Cw_\ell].$$

Therefore

$$\sum_{v:\mathrm{level}(v)=\ell} |I_v| \le T_{\mathrm{span}} + Cw_\ell, \tag{106}$$

and summing (105) gives

$$\sum_{v:\mathrm{level}(v)=\ell} \big(t_v - s_+(S_v)\big) \le \nu_\ell \bar{\sigma}_\ell + C\big(T_{\mathrm{span}} + \nu_\ell w_\ell\big). \tag{107}$$

This is where same-level disjointness replaces the crude $\nu_\ell T_{\mathrm{span}}$ by one copy of T_{span}. No quantity is transferred recursively between levels.

After the bookkeeping relabeling above, cut the merge forest at level h and retain one label for each of the q_h frontier blocks, including the unmatched lower-level blocks. Let $T_{\mathrm{high}}^{\mathrm{asyn}}$ be the sum of their eventual release charges: each release time is measured from its bookkeeping anchor, with labels propagated through later merges. The frontier blocks need not coexist at one physical time. Relabeling preserves total occupied time, so the exact decomposition is

$$T_{\mathrm{tail}}^{\mathrm{pair}} = \sum_{\ell=0}^{h-1} \sum_{v:\mathrm{level}(v)=\ell} \big(t_v - s_+(S_v)\big) + T_{\mathrm{high}}^{\mathrm{asyn}}. \tag{108}$$

For a pathwise bound on the last term, use the distinct-witness assignment from the synchronous proof, now ordering trajectories by their immutable global deadlines $s_i + \tilde{Y}_{ij}$. At each frontier block choose the latest deadline, assign the smaller child deadline to each later merge release, and pass the larger one to the parent. Natural completion uses the remaining deadline. Thus all q_h releases have distinct witnesses, even though the blocks were selected adaptively. Every bookkeeping anchor a is an original anchor and lies in $[s_{\min}, s_{\max}]$. A release with witness (i, j) therefore has charge at most $(s_i - a) + \tilde{Y}_{ij} \le T_{\mathrm{span}} + \tilde{Y}_{ij}$ on $\mathcal{H}$. Let $\tilde{Y}_{[1]} \ge \cdots \ge \tilde{Y}_{[R(B-1)]}$ denote the decreasing residual-lifetime order statistics. We obtain

$$0 \le T_{\mathrm{high}}^{\mathrm{asyn}} \le q_h T_{\mathrm{span}} + \sum_{j=1}^{q_h} \tilde{Y}_{[j]}. \tag{109}$$

The exponential order-statistic identity then gives

$$\mathbb{E}[T_{\mathrm{high}}^{\mathrm{asyn}} \mathbf{1}_{\mathcal{H}}] \le q_h \left[T_{\mathrm{span}} + \frac{1 + \log(RB/q_h)}{\lambda} \right]. \tag{110}$$

This remainder includes every release not counted below the cut, including natural completion of unmatched lower-level blocks and the last natural release. The cut is by hierarchy level, not by physical time.

Combining (101), (108), and (107) yields on $\mathcal{H}$

$$\begin{aligned} \sum_{\ell=0}^{h-1} \nu_\ell \bar{\sigma}_\ell - C \sum_{\ell=0}^{h-1} \nu_\ell w_\ell &\le T_{\mathrm{tail}}^{\mathrm{pair}} \\ &\le \sum_{\ell=0}^{h-1} \nu_\ell \bar{\sigma}_\ell + C \left[h T_{\mathrm{span}} + \sum_{\ell=0}^{h-1} \nu_\ell w_\ell \right] + T_{\mathrm{high}}^{\mathrm{asyn}}. \end{aligned} \tag{111}$$

The width term in this display contains only parameters already introduced:

$$\sum_{\ell=0}^{h-1} \nu_\ell w_\ell \le C \left[\sum_{\ell=0}^{h-1} \frac{\nu_\ell \sqrt{2^\ell B} \log(8RB)}{\lambda \kappa_\ell} + \frac{R}{\lambda} \epsilon_N \right]. \tag{112}$$

Likewise, (95) and Lemma 10 give

$$\sum_{\ell=0}^{h-1} \nu_\ell \left| \bar{\sigma}_\ell - \frac{\Phi(\rho_\ell)}{\lambda} \right| \le \frac{CR}{\lambda} \epsilon_N. \tag{113}$$

To pass from $\mathcal{H}$ to expectations, use Lemma 16 with additional path failure probability $p \le (8RB)^{-3}$. The controlled deterministic sum is zero for $h = 0$; otherwise the survivor cut and $\sum_{\ell<h} \nu_\ell \le R$ give

$$\sum_{\ell<h} \nu_\ell \bar{\sigma}_\ell \le CR \log(2RB)/\lambda. \tag{114}$$

Thus both the occupied time on the complement and the controlled sum lost in the lower bound cost at most

$$CR \left[T_{\mathrm{span}} + \frac{\log(2RB)}{\lambda} \right] \left[(8RB)^{-3/2} + R(B-1)\epsilon_N \right].$$

This keeps the coupling cost linear in its probability and uses no independence between the anchor span and anchor ages.

Finally, take expectations in (111), use (110), (112), and (113), and subtract the deterministic levels $h, \dots, K_2 - 1$ on the lower side. The resulting right-hand side is $CR\mathcal{E}_2(h)$ with (87). Dividing by R proves (88). Under the proportional regime (6), $\kappa_\ell \ge c_\eta B$ for every controlled level, and the geometric summation gives

$$\sum_{\ell=0}^{K_2-1} \frac{\sqrt{2^\ell B} \log(8RB)}{2^{\ell+1} \lambda \kappa_\ell} \le C \frac{\log(RB)}{\lambda \sqrt{B}}.$$

Here $\nu_\ell/R \le 2^{-\ell-1}$ justifies the same normalized concentration bound for arbitrary R. When $R = 2^{K_2}$ and $h = K_2$ is admissible, $q_h = 1$, giving the special case stated after the theorem. Subtracting from the no-repacking value RH_{B-1}/λ proves the saving formula and completes the proof. □

B.4.4 Asymptotic part of Proposition 19

Proof. Assume the stochastic asymptotic regime of the Scaling assumptions paragraph. Write $L = \log(8RB)$ and $S = \lambda T_{\mathrm{span}}$ only within this proof. The batch-size scaling gives $\lambda(T_{\mathrm{span}} - T_0) = o(1)$, so $S = o(\log B)$. Moreover, $\log R = \Theta(\log B)$ implies $L = \Theta(\log B)$ and $R \ge B^c$ for some fixed $c > 0$ and all sufficiently large B. Choose

$$h = \left\lceil \frac{3 \log\log B}{\log 2} \right\rceil.$$

Then $h < K_2$ eventually and $2^{-h} = O((\log B)^{-3})$.

We first verify that this cut is admissible. Under the proportional threshold regime, the crossing equation gives uniformly $\lambda \bar{\sigma}_\ell \le C(\ell+1)$ and $\Phi(\rho_\ell) \le C(\ell+1)$. Consequently,

$$(B-1) e^{-\lambda(\bar{\sigma}_h + T_{\mathrm{span}})} = \frac{C_{\mathrm{th}} e^{-S}}{2^{h+1}(\mu_A + \bar{\sigma}_h)} \ge \frac{cBe^{-S}}{2^{h+1}(h+1)} \gg L.$$

Since the crossings increase with the level, the survivor condition holds at every level below h, so the cut is admissible in (86).

It remains to verify $\lambda \mathcal{E}_2(h) = o(1)$ in (87). The controlled-level span term is $hS/R = o(1)$. The slope bound $\kappa_\ell \ge cB$ gives a normalized concentration term at most $CL/\sqrt{B} = o(1)$. By (71), the frontier contribution satisfies

$$\frac{q_h}{R} \left[S + 1 + \log(RB/q_h) \right] \le \left(2^{-h} + \frac{h}{R} \right) (S + 1 + L) = o(1).$$

The omitted deterministic levels satisfy

$$\frac{1}{R}\sum_{\ell=h}^{K_2-1}\nu_\ell\Phi(\rho_\ell)\le C\sum_{\ell=h}^{\infty}2^{-\ell-1}(\ell+1)\le C2^{-h}(h+1)=o(1).$$

Finally, the lower batch-size condition implies $\epsilon_N=(RB)^{-\omega(1)}$. All remaining normalized terms are bounded by

$$C\left[(S+L)(8RB)^{-3/2}+\{1+RB(S+L)\}\epsilon_N\right]=o(1).$$

Every term in the finite bound therefore vanishes after normalization, proving (89). □

B.5 WHOLE-BLOCK REFERENCE OCCUPIED TIME

B.5.1 SYNCHRONOUS BOUND FOR THE WHOLE-BLOCK REFERENCE

For fixed $m\ge 3$, the nominal final frontier obeys

$$Q_{K_m}=O_m\big((m-1)^{K_m}\big)=O_m\Big(R^{\log_m(m-1)}\Big)=o(R). \tag{115}$$

Choose an integer $0\le h\le K_m$. For $h\ge 1$, require

$$(B-1)e^{-\lambda\bar\sigma_{h,m}}\ge C\log(8RB), \tag{116}$$

where the survivor constant depends only on the fixed parameters and is sufficiently large. No survivor condition is required for $h=0$.

Proposition 20 (Synchronous occupied time of the whole-block reference)**.** *Suppose $B\ge 2$, $R\ge m$, the synchronized-anchor model of Section 4.1 is used, $C_{\rm th}\lambda/B\le 1-\eta$, and blockwise repacking is checked immediately after every completion and repacking event. Then, for N/B sufficiently large and B sufficiently large in terms of the fixed m,α,η, for every such cut h,*

$$\begin{aligned}
&\left|\mathbb{E}[T^{\rm blk,sync}_{\rm tail}(m)]-\sum_{\ell=0}^{K_m-1}\frac{\nu_{\ell,m}\Phi(\rho_{\ell,m})}{\lambda}\right|\\
&\le C\Bigg\{\sum_{\ell=0}^{h-1}\frac{\nu_{\ell,m}\sqrt{m^{\ell+1}\theta_{\ell,m}/\lambda}\log(8RB)}{\kappa_{\ell,m}}+\sum_{\ell=h}^{K_m-1}\frac{\nu_{\ell,m}\Phi(\rho_{\ell,m})}{\lambda}\\
&\qquad+Q_h\frac{1+\log(2RB)}{\lambda}+\frac{R}{\lambda}\epsilon_N\\
&\qquad+\frac{R\log(2RB)}{\lambda}(8RB)^{-3/2}+\frac{R^2(B-1)\log(2RB)}{\lambda}\epsilon_N\\
&\qquad+\frac{R\log(2RB)}{\lambda}\sqrt{R(B-1)(1+\lambda\alpha C_{\rm th})e^{-\lambda\alpha C_{\rm th}}}\Bigg\}.
\end{aligned}\tag{117}$$

Moreover, under the stochastic asymptotic regime of the Scaling assumptions paragraph,

$$\mathbb{E}[T^{\rm blk,sync}_{\rm tail}(m)]=\frac{1}{\lambda}\sum_{\ell=0}^{K_m-1}\nu_{\ell,m}\Phi(\rho_{\ell,m})+o(R/\lambda). \tag{118}$$

For the counting identities used in both multi-repacking proofs, write $b_\ell:=\lfloor R/m^\ell\rfloor$ and $d_\ell:=b_\ell-mb_{\ell+1}\in\{0,\dots,m-1\}$. They count nominal level-ℓ blocks and unmatched inputs, respectively. Since $\nu_{\ell,m}=(m-1)^\ell b_{\ell+1}$, telescoping gives

$$Q_h=\sum_{r=0}^{h-1}d_r(m-1)^r+b_h(m-1)^h. \tag{119}$$

Indeed, the right-hand side equals R at $h=0$ and decreases by $(m-1)^hb_{h+1}=\nu_{h,m}$ when h increases by one. Each unmatched level-r block carries $(m-1)^r$ physical labels, so this identity accounts for every label outside the controlled rounds. As $d_r,b_{K_m}\le m-1$, it also proves (115) by a finite geometric sum.

Proof. Work in the synchronized-anchor model with common local origin zero.

Use the ancestral processes and notation of Appendix B.3.5. Fix $0 \le h \le K_m$ satisfying (116) if $h > 0$. Under the stationary anchor coupling, A°_{ij} and $\tilde{Y}_{ij}$ are independent $\mathrm{Exp}(\lambda)$ variables. Their sum has tail $(1+\lambda x)e^{-\lambda x}$ at $x \ge 0$, so a union bound over the $R(B-1)$ ancestral trajectories gives

$$\mathbb{P}\left(\max_{i,j}(A^\circ_{ij} + \tilde{Y}_{ij}) > \alpha C_{\mathrm{th}}\right) \le R(B-1)(1+\lambda\alpha C_{\mathrm{th}})e^{-\lambda\alpha C_{\mathrm{th}}}. \tag{120}$$

On the coupling event, every live trajectory satisfies $\tilde{A}_{ij} + u = A^\circ_{ij} + u \le A^\circ_{ij} + \tilde{Y}_{ij}$. Thus this bound controls the individual-size requirement at all local times; coupling failure is accounted for separately.

Let $\mathcal{P}_m$ be the intersection of the common event $\mathcal{G}$ with the item-size event in (120). On it, the common load and count bounds hold and every live trajectory has size at most αC_{th}. We first obtain readiness pathwise on this event, then prove output survival under the current stopping-time history before intersecting with it. The final intersection is denoted by $\mathcal{G}_m$.

Put

$$c_{\ell,m} := \frac{\theta_{\ell,m}}{m^{\ell+1}} = \left(\frac{m-1}{m}\right)^{\ell+1}(1-\alpha)C_{\mathrm{th}}. \tag{121}$$

Using the crossing functions of Lemma 14, one has $\bar{\sigma}(c_{\ell,m}) = \bar{\sigma}_{\ell,m}$. Write $\bar{\sigma}^\circ_{\ell,m} := \bar{\sigma}^\circ(c_{\ell,m})$ and, for $\ell < h$, define

$$w_{\ell,m} := \sup_{c_{\ell+1,m} \le c \le c_{\ell,m}} d(c), \tag{122}$$

with d from (62), using the fixed regularity margin established in Lemma 11. The window multiplier may therefore depend on the fixed m, α, η, but is independent of the survivor constant. The cutoff and Lemma 14 give

$$\lambda w_{\ell,m} \le C\left[\sqrt{\frac{L}{q(\bar{\sigma}_{\ell+1,m})}} + \frac{L}{q(\bar{\sigma}_{\ell+1,m})} + \epsilon_N\right]. \tag{123}$$

This controls every fixed multiple of the width, uniformly in the controlled level. For the separate coarse estimate used to sum errors, the identity

$$\kappa_{\ell,m} = m^{\ell+1}\left[-\frac{\mathrm{d}}{\mathrm{d}u}(B-1)e^{-\lambda u}(\mu_A + u)\right]_{u=\bar{\sigma}_{\ell,m}},$$

together with the crossing equation and $\bar{\sigma}_{\ell,m} + \lambda^{-1} \le C\lambda^{-1}\log(2RB)$ gives

$$w_{\ell,m} \le C\frac{\sqrt{m^{\ell+1}\theta_{\ell,m}/\lambda}\,\log(8RB)}{\kappa_{\ell,m}} + \frac{C}{\lambda}\epsilon_N. \tag{124}$$

Indeed, at the crossing $q(\bar{\sigma}_{\ell,m})(\mu_A + \bar{\sigma}_{\ell,m}) = c_{\ell,m}$, so multiplying the one-original path radius by $m^{\ell+1}$ turns its square-root part into $O(\sqrt{m^{\ell+1}\theta_{\ell,m}/\lambda}\,L)$. The logarithmic part is absorbed by the survivor condition. This explains the factor $(m-1)^{(\ell+1)/2}$ that would be missing if $\theta_{\ell,m}$ were incorrectly treated as level independent.

The no-early-crossing assertion of Lemma 14 gives, with the multiplier already included in the definition of $w_{\ell,m}$,

$$Z_i(u) > c_{\ell,m}, \qquad 0 \le u < \bar{\sigma}_{\ell,m} - w_{\ell,m}. \tag{125}$$

Apply Lemma 15 to $[c_{\ell+1,m}, c_{\ell,m}]$ and set

$$r_{\ell,m} := \bar{\sigma}_{\ell,m} + K_{\eta,m}w_{\ell,m}. \tag{126}$$

With $K_{\eta,m}$ fixed first, enlarge the survivor constant and then take N/B large enough. The lemma gives $q(r_{\ell,m}) \ge q(\bar{\sigma}_{\ell,m})/2 \ge CL$ and

$$Z_i(r_{\ell,m}) \le c_{\ell,m} \qquad (i \in [R]). \tag{127}$$

Thus a prospective level-ℓ superblock, which contains $M_\ell = m^{\ell+1}$ originals, has total load at most $M_\ell c_{\ell,m} = \theta_{\ell,m}$ at this upper time.

Dividing consecutive crossing equations gives $m/(m-1) < e^{\lambda(\bar{\sigma}_{\ell+1,m}-\bar{\sigma}_{\ell,m})}$, hence

$$\bar{\sigma}_{\ell+1,m} - \bar{\sigma}_{\ell,m} > \lambda^{-1} \log \frac{m}{m-1}. \tag{128}$$

Uniform window smallness, with the same order of constant choices, therefore ensures

$$r_{\ell,m} < \bar{\sigma}_{\ell+1,m} - w_{\ell+1,m} \quad (\ell < h-1), \qquad r_{\ell,m} < \bar{\sigma}_{\ell+1,m} \quad (\ell < h). \tag{129}$$

The common shifted-cut comparison also gives

$$q(r_{\ell,m} + d) \geq \tfrac{1}{2} q(\bar{\sigma}_{\ell,m} + d), \qquad d \geq 0. \tag{130}$$

We next verify the slot constraint that is absent from a capacity-only packing argument. Set $p_\ell = (m-1)^{\ell+1}$. At the upper time, (60) gives

$$\sum_{i \in \mathcal{B}} N_i(r_{\ell,m}) \leq M_\ell q(r_{\ell,m}) + C M_\ell [\sqrt{q(r_{\ell,m})L} + L].$$

At the deterministic crossing,

$$M_\ell q(\bar{\sigma}_{\ell,m}) = \frac{\theta_{\ell,m}}{\mu_A + \bar{\sigma}_{\ell,m}} \leq p_\ell B (1 - c_{\eta,m})$$

for N/B sufficiently large, using $C_{\mathrm{th}}\lambda/B \leq 1 - \eta$. The survivor condition makes the deviation a smaller fixed fraction of the right-hand side. Hence

$$\sum_{i \in \mathcal{B}} N_i(r_{\ell,m}) \leq p_\ell B. \tag{131}$$

Together with (127) and the item-size part of $\mathcal{P}_m$, Lemma 9 therefore makes every such superblock feasible, provided its input blocks have formed and remain nonterminal.

We next establish output survival as an implication whose premise is measurable at the packing time. Let $\mathcal{F}_t$ contain the initial anchor states, the observed completions up to time t, and all nonanticipating packing decisions made by then. Each actual packing time is a stopping time. Consider a level-ℓ packing at time t, with $0 \leq \ell < h-1$, and let $Q(t)$ be its number of live items. Write $M_\ell = m^{\ell+1}$ and $p_\ell = (m-1)^{\ell+1}$. The next controlled deadline is the deterministic time $U_\ell := r_{\ell+1,m}$. Define the $\mathcal{F}_t$-measurable event

$$\mathcal{E}_t := \left\{ t \leq r_{\ell,m}, \quad Q(t) \geq \tfrac{1}{2} M_\ell q(t) \right\}. \tag{132}$$

We do not yet assume that every packing satisfies $\mathcal{E}_t$; this will follow from the formation induction below.

After the observed sizes determine the balanced assignment, let n_b be the number assigned to output bin b. The cyclic construction gives $n_b \geq Q(t)/p_\ell - 1$. On $\mathcal{E}_t$, one has $t \leq r_{\ell,m} < U_\ell$ and

$$\begin{aligned} n_b e^{-\lambda(U_\ell - t)} &\geq \frac{M_\ell}{2p_\ell} q(U_\ell) - 1 \\ &\geq \frac{M_\ell}{4p_\ell} q(\bar{\sigma}_{\ell+1,m}) - 1 \geq 5L. \end{aligned} \tag{133}$$

Here the first step uses $q(t)e^{-\lambda(U_\ell - t)} = q(U_\ell)$; the second uses (130) at the controlled level $\ell + 1$. The cutoff and $M_\ell/p_\ell \geq 1$ give the last step after one fixed enlargement of the survivor constant.

Independence at the anchors and exponential memorylessness at the stopping time imply that, conditional on $\mathcal{F}_t$, the remaining lifetimes of these items are independent $\mathrm{Exp}(\lambda)$ variables. Let $F_{t,b}$ be the event that all items assigned to bin b at time t complete by U_ℓ, irrespective of any later migration. On $\mathcal{E}_t$,

$$\begin{aligned} \mathbb{P}(F_{t,b} \mid \mathcal{F}_t) &= \left(1 - e^{-\lambda(U_\ell - t)}\right)^{n_b} \\ &\leq \exp[-n_b e^{-\lambda(U_\ell - t)}] \leq (8RB)^{-5}. \end{aligned} \tag{134}$$

Consequently the tower property yields $\mathbb{P}(\mathcal{E}_t \cap F_{t,b}) \leq (8RB)^{-5}$. No conditioning on $\mathcal{P}_m$ or a future formation event has been used. Outside this failure, whenever $\mathcal{E}_t$ holds, the physical bin stays nonempty until U_ℓ or its next repacking, whichever comes first.

Define $\mathcal{G}_m$ by removing from $\mathcal{P}_m$ all events $\mathcal{E}_t \cap F_{t,b}$ for the outputs of the actual rounds with $\ell < h-1$. A level-ℓ round with $k = (m-1)^\ell$ repacking releases creates $(m-1)k$ output bins, while the total number of repacking releases is at most R. Thus at most $(m-1)R$ bins are tested. Enumerate them in their order of creation, apply the conditional estimate at each actual stopping time, and use the tower property before summing. This bounds their total failure probability by $(m-1)R(8RB)^{-5} \le C_m(8RB)^{-3}$ without any independence between rounds. Together with the original-replica path, coupling, and item-size bounds, this gives

$$\begin{aligned}\mathbb{P}(\mathcal{G}_m^c) \le C_m(8RB)^{-3} + R(B-1)\epsilon_N \\ + R(B-1)(1+\lambda\alpha C_{\mathrm{th}})e^{-\lambda\alpha C_{\mathrm{th}}}.\end{aligned} \tag{135}$$

We now close the formation and survival induction on $\mathcal{G}_m$. If $h = 0$, there are no controlled rounds and the induction is empty. For $h > 0$, all $b_0 = R$ singleton blocks exist at time zero. The count bound and the cutoff give $N_i(r_{0,m}) \ge q(r_{0,m})/2 > 0$, so each singleton remains nonterminal until $r_{0,m}$ or its first repacking.

Suppose first that $\ell \ge 1$ and all controlled levels below ℓ have been processed as claimed. Their b_ℓ level-ℓ output blocks have formed by $r_{\ell-1,m}$, and every physical bin needed in them survives until $r_{\ell,m}$ or its next repacking. The separation (129) and the no-early bound (125) exclude a level-ℓ round before $r_{\ell-1,m}$. Thus all inputs are available before level ℓ begins; the singleton argument supplies the same premise for $\ell = 0$.

At time $r_{\ell,m}$ every remaining group of m level-ℓ blocks passes the load, count, and item-size tests. The exhaustive event-driven check must therefore leave fewer than m such blocks. To justify this even if $r_{\ell,m}$ is not an event time, take the most recent completion or repacking event at or before $r_{\ell,m}$ after all level-ℓ inputs formed, including the common anchor event at zero when $\ell = 0$. Between this event and $r_{\ell,m}$, loads and live item sizes increase, and counts stay constant. Feasibility at $r_{\ell,m}$ would then imply feasibility at that event, contradicting exhaustion if m inputs remained. There are consequently exactly $b_{\ell+1} = \lfloor b_\ell/m \rfloor = \lfloor R/m^{\ell+1} \rfloor$ rounds by $r_{\ell,m}$ and $d_\ell = b_\ell - mb_{\ell+1}$ unmatched inputs.

For each actual round so obtained, its time satisfies $t \le r_{\ell,m}$. On $\mathcal{P}_m$, monotonicity of q and the cutoff give $N_i(t) \ge q(t)/2$ for every ancestor, so $Q(t) \ge M_\ell q(t)/2$. If $\ell < h-1$, this verifies $\mathcal{E}_t$; the survival part of $\mathcal{G}_m$ now keeps every output bin nonempty until $r_{\ell+1,m}$ or its next repacking. This proves the survival premise for the next level using only the formation conclusion already established at the current level. No input needed at a controlled level triggers the terminal fallback.

At the last controlled level $\ell = h-1$, no survival to a later deadline is required. Nevertheless, the same observed count bound gives $n_b \ge M_\ell q(r_{\ell,m})/(2p_\ell) - 1 > 0$ at creation. Thus this round also releases exactly $(m-1)^\ell$ replicas and creates the prescribed number of nonempty frontier bins. Their subsequent natural releases are included in the frontier remainder. The induction proves the nominal multiplicity $\nu_{\ell,m}$ at every $\ell < h$. Finally, (125) and the established upper time place all these repacking releases in $[\bar{\sigma}_{\ell,m} - w_{\ell,m}, r_{\ell,m}]$.

On $\mathcal{G}_m$, cut the block forest at level h. Its frontier consists of the d_r unmatched level-r blocks for $0 \le r < h$ and the b_h level-h blocks. A level-r frontier block carries $(m-1)^r$ physical labels, so the frontier contains exactly Q_h labels by (34). Propagate the level-h labels through later repacking operations, assigning the labels removed at an operation to its released physical replicas; labels in unmatched lower-level blocks end at natural completion. Every frontier label is assigned exactly once, either to a later repacking release or to a natural completion. Therefore all releases not counted below the cut, including the natural releases of unmatched lower-level blocks, have total time at most

$$T_{\mathrm{high}}^{\mathrm{sync}} \le Q_h \max_{i,j} \tilde{Y}_{ij}.$$

Taking $T_{\mathrm{high}}^{\mathrm{sync}} = 0$ off $\mathcal{G}_m$ if needed, it follows that

$$\mathbb{E}[T_{\mathrm{high}}^{\mathrm{sync}} \mathbf{1}_{\mathcal{G}_m}] \le \frac{CQ_h[1+\log(2RB)]}{\lambda}. \tag{136}$$

The error from the controlled levels is at most $C\sum_{\ell<h} \nu_{\ell,m} w_{\ell,m}$ on $\mathcal{G}_m$. Since $\sum_{\ell<h} \nu_{\ell,m} \le R$, (124) yields

$$\sum_{\ell<h} \nu_{\ell,m} w_{\ell,m} \le C\left[\sum_{\ell<h} \frac{\nu_{\ell,m}\sqrt{m^{\ell+1}\theta_{\ell,m}/\lambda}\log(8RB)}{\kappa_{\ell,m}} + \frac{R}{\lambda}\epsilon_N\right]. \tag{137}$$

For the complement of $\mathcal{G}_m$, apply Lemma 16 in the synchronized model. The path and destination-survival failures contribute $C_m(8RB)^{-3}$ to p, and the item-size failure contributes $R(B-1)(1+\lambda\alpha C_{\mathrm{th}})e^{-\lambda\alpha C_{\mathrm{th}}}$. The deterministic controlled sum is at most $CR\log(2RB)/\lambda$ by the cut (and is zero for $h=0$), so the lemma controls both sides. It yields precisely the three exceptional terms in (117), keeping the coupling cost linear. Use (35) for the mean correction, (136) for the frontier, and subtract the omitted deterministic levels $h,\dots,K_m-1$. Together with (137), this proves the finite bound. For (118), set $\beta := (m-1)/m$ and $L := \log(8RB)$, and choose $h = \lceil 3\log L/\log(1/\beta)\rceil$. Under the stated regime, $h < K_m$ for sufficiently large B, $\beta^h = \Theta(L^{-3})$, and the crossing equation gives $(B-1)e^{-\lambda\bar{\sigma}_{h,m}} = \Theta(B\beta^h/(h+1))$. Thus (116) holds. The frontier decomposition (119) implies $Q_h/R \le C_m\beta^h$, while $\nu_{\ell,m}/R \le \beta^\ell/m$ and $\Phi(\rho_{\ell,m}) = O(\ell+1)$. Consequently, after division by R, the omitted-level and frontier terms are $O(((h+1)\beta^h + L\beta^h)/\lambda) = o(1/\lambda)$. The normalized concentration term is $O(L/(\lambda\sqrt{B}))$; the finite-batch and exceptional-event terms also vanish relative to $1/\lambda$. Equation (117) therefore yields (118). □

B.5.2 Finite-system part of Proposition 21

A positive analytical cutoff h must satisfy

$$(B-1)e^{-\lambda(\bar{\sigma}_{h,m}+T_{\mathrm{span}})} \ge C\log(8RB); \tag{138}$$

with a sufficiently large constant; $h=0$ is always allowed. The proof cutoff h is distinct from the algorithm's switch depth d. For the reference and multi-repacking bounds, the explicit per-replica error is

$$\begin{aligned}\mathcal{E}_m(h) := {} & \frac{T_{\mathrm{span}}}{R}\sum_{\ell=0}^{h-1}\nu_{\ell,m} + \frac{1}{R}\sum_{\ell=0}^{h-1}\frac{\nu_{\ell,m}\sqrt{m^{\ell+1}\theta_{\ell,m}/\lambda}\log(8RB)}{\kappa_{\ell,m}} \\ & + \frac{1}{R\lambda}\sum_{\ell=h}^{K_m-1}\nu_{\ell,m}\Phi(\rho_{\ell,m}) + \frac{Q_h}{R}\left[T_{\mathrm{span}} + \frac{1+\log(2RB)}{\lambda}\right] + \frac{\epsilon_N}{\lambda} \\ & + \left[T_{\mathrm{span}} + \frac{\log(2RB)}{\lambda}\right]\Big[(8RB)^{-3/2} + R(B-1)\epsilon_N \\ & \qquad + \sqrt{R(B-1)(1+\lambda\alpha C_{\mathrm{th}})e^{-\lambda\alpha C_{\mathrm{th}}}}\Big].\end{aligned} \tag{139}$$

The terms have the same roles as in (87), with roundwise waiting and an additional individual-size failure term. The latter accounts for the observed size test; it does not impose a deterministic bound on every trajectory. This route retains the aggregate error bound without imposing a new bounded-startup assumption. It does not assert a uniformly smaller remainder or a pathwise ordering between the two executions. Under (6), the concentration term remains $O(\log(RB)/(\lambda\sqrt{B}))$; the displayed item-size failure term is exponentially small when $\log(RB) = o(B)$ and its time prefactor is subexponential.

Proposition 21 (Asynchronous occupied time of the whole-block reference)**.** *Suppose $B \ge 2$, $R \ge m$, and $C_{\mathrm{th}}\lambda/B \le 1-\eta$. Let the anchors be the actual finite-batch epochs, let T_{span} be given by* (25)*, and use the blockwise event-driven rule of Algorithm 3. For N/B sufficiently large and B sufficiently large in terms of the fixed m,α,η, every cut $0 \le h \le K_m$ satisfying* (138) *when $h>0$ obeys*

$$\left|\frac{\mathbb{E}[T^{\mathrm{blk}}_{\mathrm{tail}}(m)]}{R} - \frac{1}{R\lambda}\sum_{\ell=0}^{K_m-1}\nu_{\ell,m}\Phi(\rho_{\ell,m})\right| \le C\mathcal{E}_m(h). \tag{140}$$

The same bound holds for the mean saving per original replica when its reference value is H_{B-1}/λ minus the deterministic term in (140)*. Moreover, under the stochastic asymptotic regime of the Scaling assumptions paragraph,*

$$\mathbb{E}[T^{\mathrm{blk}}_{\mathrm{tail}}(m)] = \frac{1}{\lambda}\sum_{\ell=0}^{K_m-1}\nu_{\ell,m}\Phi(\rho_{\ell,m}) + o(R/\lambda). \tag{141}$$

Proof. Fix $0 \le h \le K_m$ satisfying (138) if $h>0$.

The asynchronous cut implies (116), so use the same widths, endpoints, and item-size event as in the synchronous proof, together with the common $q, m^\circ, L, \mathcal{G}, \mathcal{A}$. Let $\mathcal{G}^{\mathrm{a}}_m$ intersect $\mathcal{G}$, the item-size

event, and the asynchronous output-survival events constructed below; put $\mathcal{H}_m = \mathcal{G}_m^{\mathrm{a}} \cap \mathcal{A}$. More precisely, the path, coupling, item-size, and conditional empty-bin bounds give

$$\begin{aligned}\mathbb{P}\big((\mathcal{G}_m^{\mathrm{a}})^c \cap \mathcal{A}\big) \le\ & C_m(8RB)^{-3} + R(B-1)\epsilon_N \\ & + R(B-1)(1+\lambda\alpha C_{\mathrm{th}})e^{-\lambda\alpha C_{\mathrm{th}}}.\end{aligned}$$

Combining this estimate with Lemma 5 gives

$$\begin{aligned}\mathbb{P}(\mathcal{H}_m^c) \le\ & (C_m+1)(8RB)^{-3} + R(B-1)\epsilon_N \\ & + R(B-1)(1+\lambda\alpha C_{\mathrm{th}})e^{-\lambda\alpha C_{\mathrm{th}}}.\end{aligned} \tag{142}$$

All deterministic comparisons below are made on $\mathcal{H}_m$. The conditional survival estimates are instead proved at the actual stopping times under the current history, before intersecting with $\mathcal{A}$ or the global path events. Because the path bounds are simultaneous in i and $u \ge 0$, they may be evaluated at $u = t - s_i$ without conditioning on the anchor vector or assuming that the anchor epochs are independent of the anchor ages.

Use $r_{\ell,m}$ and the level separation (129). The asynchronous cut ensures $q(r_{\ell,m} + T_{\mathrm{span}}) \ge CL$, so the common persistent endpoint and no-early-crossing lemmas give

$$Z_i(u) \le c_{\ell,m}, \qquad r_{\ell,m} \le u \le r_{\ell,m} + T_{\mathrm{span}}, \quad i \in [R],\ \ell < h, \tag{143}$$

and $Z_i(u) > c_{\ell,m}$ for $u < \bar{\sigma}_{\ell,m} - w_{\ell,m}$, as in (125). The count version of the same argument gives

$$\sum_{i\in\mathcal{B}} N_i(t - s_i) \le (m-1)^{\ell+1}B \tag{144}$$

whenever $\mathcal{B}$ is a prospective level-ℓ superblock and $t - s_i \in [r_{\ell,m}, r_{\ell,m} + T_{\mathrm{span}}]$ for all $i \in \mathcal{B}$. The proof is the calculation leading to (131), with the largest count taken at the left endpoint. The margin $C_{\mathrm{th}}\lambda/B \le 1 - \eta$ absorbs the uniform count deviation.

We first establish a pathwise round window under the formation and survival properties verified by induction below. Consider a prospective level-ℓ round v whose m children have formed by the candidate upper endpoint below and remain nonterminal until then, unless already promoted in another round. Let $\mathcal{B}_v$ be its $M_\ell = m^{\ell+1}$ original ancestors and write

$$x_v := \min_{i\in\mathcal{B}_v} s_i, \qquad y_v := \max_{i\in\mathcal{B}_v} s_i, \qquad \overline{s}_v := \frac{1}{M_\ell}\sum_{i\in\mathcal{B}_v} s_i.$$

Before $x_v + \bar{\sigma}_{\ell,m} - Cw_{\ell,m}$, either some input block has not been formed or every relevant ancestral process that has started is above $c_{\ell,m}$. In the latter case the total load is greater than $M_\ell c_{\ell,m} = \theta_{\ell,m}$, so the algorithm's sufficient packing test cannot pass.

At $t^+ = y_v + r_{\ell,m}$, every local time belongs to $[r_{\ell,m}, r_{\ell,m} + T_{\mathrm{span}}]$. Equations (143) and (144) give total load at most $\theta_{\ell,m}$ and at most $(m-1)^{\ell+1}B$ active trajectories. The item-size condition and Lemma 9 make the round feasible. If the exhaustive check were to consume one of them in another feasible group first, $\mathcal{B}_v$ would not be the ancestral set of a later actual round; otherwise the m blocks themselves are processed no later than t^+. If t^+ is between events, feasibility there implies feasibility at the most recent event after formation because loads only increase between completion events. Under these formation and survival properties, the two-sided window is

$$x_v + \bar{\sigma}_{\ell,m} - Cw_{\ell,m} \le t_v \le y_v + \bar{\sigma}_{\ell,m} + Cw_{\ell,m}. \tag{145}$$

We next prove the survival estimates under the current history. Let $\mathcal{F}_t$ be the global filtration generated by observed service events and the nonanticipating packing decisions. Each packing time is a stopping time. At every anchor or packing time t, the earliest anchor $s_{\min}$ has already occurred and is $\mathcal{F}_t$-measurable: every anchor not yet observed must exceed t.

Consider an actual level-ℓ packing with $\ell < h - 1$, performed at time t, and write $\mathcal{B}$ for its original ancestors and $p = (m-1)^{\ell+1}$ for its output count. Define the deadline measurable with respect to the current history

$$D := s_{\min} + T_{\mathrm{span}} + r_{\ell+1,m}. \tag{146}$$

On $\mathcal{A}$, this deadline is at least $s_{\max} + r_{\ell+1,m}$. Moreover, for every $i \in \mathcal{B}$, $D - s_i \le T_{\text{span}} + r_{\ell+1,m}$. Let $Q(t)$ be the number of items being packed, and work on the $\mathcal{F}_t$-measurable event

$$\mathcal{E}_t := \left\{ t \le D, \quad Q(t) \ge \frac{1}{2} \sum_{i \in \mathcal{B}} q(t - s_i) \right\}.$$

We impose no conditioning on $\mathcal{A}$ or on a future path event here. The balanced construction assigns at least $Q(t)/p - 1$ items to each output bin b; denote its actual item count by n_b. On $\mathcal{E}_t$,

$$\begin{aligned} n_b e^{-\lambda(D-t)} &\ge \frac{1}{2p} \sum_{i \in \mathcal{B}} q(D - s_i) - 1 \\ &\ge \frac{m^{\ell+1}}{2p} q(T_{\text{span}} + r_{\ell+1,m}) - 1 \ge 5L. \end{aligned}$$

The last inequality follows from the cutoff and (130), after one fixed enlargement of the cutoff constant. In particular, the same span T_{span} suffices.

Conditional on $\mathcal{F}_t$, the future residual lifetimes of the assigned items are independent $\text{Exp}(\lambda)$ variables. Let $F_{t,b}$ be the event that all items assigned to bin b at time t complete by D, regardless of any later change of physical location. On $\mathcal{E}_t$,

$$\begin{aligned} \mathbb{P}(F_{t,b} \mid \mathcal{F}_t) &= \left(1 - e^{-\lambda(D-t)}\right)^{n_b} \\ &\le \exp\left[-n_b e^{-\lambda(D-t)}\right] \le (8RB)^{-5}. \end{aligned} \tag{147}$$

The tower property gives $\mathbb{P}(\mathcal{E}_t \cap F_{t,b}) \le (8RB)^{-5}$. Outside this failure event, whenever $\mathcal{E}_t$ holds, the bin cannot disappear by natural completion before D or its next repacking, whichever comes first.

For the initial singletons, when $h > 0$, apply the same argument at $t = s_i$ with $B - 1$ items and deadline $s_{\min} + T_{\text{span}} + r_{0,m}$. On the event that this deadline is at least s_i, the conditional expected survivor count is at least $q(T_{\text{span}} + r_{0,m}) \ge 5L$, using (129) and the cutoff. This event always holds on $\mathcal{A}$. Thus its intersection with the event that all these items complete by the deadline also has probability at most $(8RB)^{-5}$.

Define the survival part of $\mathcal{G}_m^{\text{a}}$ by excluding these failures for the initial singletons and the outputs of the actual rounds with $\ell < h - 1$. A round releasing k replicas creates $(m-1)k$ output bins, and the total number of repacking releases is at most R. Hence there are at most mR bins to check, including the initial singletons. Applying the tower property successively at the actual stopping times and summing over their current output bins bounds the total survival failure probability by $mR(8RB)^{-5} \le C_m(8RB)^{-3}$. This uses no independence between rounds and justifies the survival contribution in (142).

We now close the formation and survival induction on $\mathcal{H}_m$. When $h = 0$, there are no controlled rounds and no such induction is needed. Otherwise, the singleton estimate keeps every initial replica nonterminal until $s_{\max} + r_{0,m}$ or its first repacking. For a prospective parent G, write $s_+(G) := \max_{i \in G} s_i$. Its singleton children exist by $s_+(G)$. At level $\ell \ge 1$, the induction hypothesis forms its children no later than $s_+(G) + r_{\ell-1,m}$ and keeps them nonterminal until $s_{\max} + r_{\ell,m}$ or their next repacking. Since $r_{\ell-1,m} < r_{\ell,m}$, the formation and survival properties used in (145) hold. The exhaustive check forms G by $s_+(G) + r_{\ell,m}$ or consumes a child in another round, in which case G is not an actual later block.

Every actual level-ℓ round therefore occurs by $s_{\max} + r_{\ell,m}$. If $\ell < h - 1$, its packing time satisfies $t \le D$ on $\mathcal{A}$, because the endpoints $r_{\ell,m}$ increase with ℓ. Each ancestral local time lies between zero and $T_{\text{span}} + r_{\ell,m}$, where the uniform count bound implies $Q(t) \ge \frac{1}{2} \sum_{i \in \mathcal{B}} q(t - s_i)$ after enlarging the cutoff constant. Thus $\mathcal{E}_t$ holds, and the survival estimate keeps every output needed at the next controlled level nonterminal until $s_{\max} + r_{\ell+1,m}$ or its next repacking. This closes both induction claims. At the last controlled level $\ell = h - 1$, the same count bound and the balanced assignment give, at creation,

$$n_b \ge \frac{m^h}{2(m-1)^h} q(T_{\text{span}} + r_{h-1,m}) - 1 > 0.$$

The strict inequality follows from the cut condition and (130). Thus these rounds also create the prescribed number of nonempty outputs. They are frontier representatives and need no further survival estimate.

By $s_{\max} + r_{\ell,m}$, every remaining level-ℓ block is formed, nonterminal, and persistently ready. The exhaustive check leaves fewer than m such blocks. Induction therefore gives exactly $b_{\ell+1} = \lfloor R/m^{\ell+1} \rfloor$ rounds at each controlled level, with release multiplicity $(m-1)^\ell$, and (145) applies to every actual round.

We next convert the round windows into total occupied time without assuming any false disjointness. At a fixed level the sets $\{\mathcal{B}_v\}_v$ are disjoint and their union, denoted by $\mathcal{I}_\ell$, contains exactly $m^{\ell+1} b_{\ell+1}$ original replicas. Hence

$$\sum_{v:\mathrm{level}(v)=\ell} \overline{s}_v = \frac{1}{m^{\ell+1}} \sum_{i \in \mathcal{I}_\ell} s_i. \tag{148}$$

Since $x_v \geq \overline{s}_v - T_{\mathrm{span}}$ and $y_v \leq \overline{s}_v + T_{\mathrm{span}}$, (145) implies

$$|t_v - \overline{s}_v - \bar{\sigma}_{\ell,m}| \leq T_{\mathrm{span}} + C w_{\ell,m}. \tag{149}$$

Each such round releases $k_\ell = (m-1)^\ell$ physical replicas, so summing (149) at level ℓ produces exactly

$$\nu_{\ell,m} \left[T_{\mathrm{span}} + C w_{\ell,m} \right].$$

This separate-round estimate does not use the minimum-total-load choice. Appendix B.5.3 improves its span contribution by summing the individual round spans before substituting the global bound.

For completeness, we now give the exact anchor accounting, including the releases above the cutoff and the unmatched blocks below it. Initially, the single label of a level-zero block $\{i\}$ is assigned anchor s_i. Suppose inductively that a level-r block G carries $(m-1)^r$ labels, each assigned the block average

$$\overline{s}_G := \frac{1}{|G|} \sum_{i \in G} s_i.$$

If m such blocks are repacked into a parent P, their assigned anchors have total

$$(m-1)^r \sum_{G \text{ child of } P} \overline{s}_G = m(m-1)^r \overline{s}_P,$$

because the children have equal cardinality. We may therefore reassign the common value $\overline{s}_P$ to all $m(m-1)^r$ incoming labels without changing the total anchor charge. Charge the $(m-1)^r$ labels released in this round at time t_v and pass the remaining $(m-1)^{r+1}$ labels to the parent. Thus every controlled level-r round contributes exactly $(m-1)^r (t_v - \overline{s}_v)$.

After all rounds below level h, the frontier consists of the d_r unmatched level-r blocks for $0 \leq r < h$ and the b_h level-h blocks. It contains exactly Q_h physical labels by (34). Propagate the level-h labels through all later blockwise repackings; labels removed at an operation are assigned to its released replicas, and terminal natural releases receive all remaining labels, including those in unmatched lower-level blocks. This is a bijection between the Q_h frontier labels and all release events not counted below the cut, with each event repeated according to its release multiplicity. Define

$$T_{\mathrm{high}}^{\mathrm{asyn}} := \sum_{\text{frontier labels } a} \left(t_{\mathrm{release}}(a) - \overline{s}_{\mathrm{block}(a)} \right).$$

Each assigned release occurs after its frontier block was formed, so $T_{\mathrm{high}}^{\mathrm{asyn}} \geq 0$. Moreover, on $\mathcal{H}_m$ every release occurs no later than $s_{\max} + \max_{i,j} \tilde{Y}_{ij}$, and every frontier average is at least $s_{\min}$. Therefore, pathwise,

$$0 \leq T_{\mathrm{high}}^{\mathrm{asyn}} \leq Q_h \left(T_{\mathrm{span}} + \max_{i,j} \tilde{Y}_{ij} \right). \tag{150}$$

The label construction preserves the total assigned anchor at every round; it starts from $\sum_i s_i$. Consequently, the definition of T_{tail} gives the exact decomposition

$$T_{\mathrm{tail}}^{\mathrm{blk}}(m) = \sum_{\ell=0}^{h-1} \sum_{v:\mathrm{level}(v)=\ell} (m-1)^\ell (t_v - \overline{s}_v) + T_{\mathrm{high}}^{\mathrm{asyn}}. \tag{151}$$

Combining (149), (150), and (151) yields on $\mathcal{H}_m$

$$T^{\mathrm{blk}}_{\mathrm{tail}}(m) \geq \sum_{\ell=0}^{h-1} \nu_{\ell,m}\bar{\sigma}_{\ell,m} - \sum_{\ell=0}^{h-1} \nu_{\ell,m}[T_{\mathrm{span}} + Cw_{\ell,m}],$$
$$T^{\mathrm{blk}}_{\mathrm{tail}}(m) \leq \sum_{\ell=0}^{h-1} \nu_{\ell,m}\bar{\sigma}_{\ell,m} + \sum_{\ell=0}^{h-1} \nu_{\ell,m}[T_{\mathrm{span}} + Cw_{\ell,m}] + T^{\mathrm{asyn}}_{\mathrm{high}}. \tag{152}$$

This accounts for all natural completions, not merely the last one.

It remains to pass to expectations. From (150),

$$\mathbb{E}[T^{\mathrm{asyn}}_{\mathrm{high}}\mathbf{1}_{\mathcal{H}_m}] \leq C_m Q_h \left[T_{\mathrm{span}} + \frac{1+\log(2RB)}{\lambda}\right]. \tag{153}$$

Use (137) at the chosen cut h for the low-level widths. Also, (35) and $\sum_{\ell<h}\nu_{\ell,m} \leq R$ give

$$\sum_{\ell<h} \nu_{\ell,m}\left|\bar{\sigma}_{\ell,m} - \frac{\Phi(\rho_{\ell,m})}{\lambda}\right| \leq \frac{CR}{\lambda}\epsilon_N.$$

Apply Lemma 16 with

$$p \leq C_m(8RB)^{-3} + R(B-1)(1+\lambda\alpha C_{\mathrm{th}})e^{-\lambda\alpha C_{\mathrm{th}}}.$$

This accounts for path, destination-survival, and item-size failures on $\mathcal{A}$; span failure and the independent coupling failure are handled separately in that lemma. The controlled deterministic sum is at most $CR\log(2RB)/\lambda$ by the cut, so its loss on the complement obeys the same estimate. Expanding $\sqrt{p}$ gives exactly the exceptional-event terms of (139).

Take expectations in (152), apply (153) and (137), replace the controlled exact crossings using (35), and subtract the deterministic levels $\ell = h, \ldots, K_m - 1$. The resulting error is $CR\mathcal{E}_m(h)$ with (139); dividing by R proves (140). To verify the simplified concentration order quoted after the theorem, observe that under (6), uniform regularity gives $\kappa_{\ell,m} \asymp_{\eta,m} \lambda\theta_{\ell,m}$. Consequently,

$$\frac{1}{R}\sum_{\ell<h} \frac{\nu_{\ell,m}\sqrt{m^{\ell+1}\theta_{\ell,m}/\lambda}\log(8RB)}{\kappa_{\ell,m}}$$
$$\leq \frac{C\log(8RB)}{\lambda\sqrt{B}}\sum_{\ell\geq 0}\left(\frac{m-1}{m}\right)^{(\ell+1)/2} \leq C\frac{\log(RB)}{\lambda\sqrt{B}}.$$

Subtracting from the no-repacking value RH_{B-1}/λ gives the saving formula. □

B.5.3 Refined asynchronous comparison for the whole-block reference

The auxiliary comparison below uses the proportional threshold and batch-size conditions together with an admissible cut. The additional replica-count and startup conditions in (8) are invoked only for the limiting argument at the end of this subsection.

Proposition 22 (Refined asynchronous comparison)**.** *Retain the model and algorithm in the finite-system part of Proposition 21, with the fixed parameters and proportional threshold regime in* (6)*. Assume the batch-size scaling* (7)*. There is a constant $C > 0$, depending only on the fixed model parameters, such that, for every integer $1 \leq h \leq K_m$ satisfying* (138)*,*

$$\left|\frac{\lambda}{R}\mathbb{E}[T^{\mathrm{blk}}_{\mathrm{tail}}(m)] - \frac{1}{R}\sum_{\ell=0}^{K_m-1}\nu_{\ell,m}\Phi(\rho_{\ell,m})\right|$$
$$\leq C\left[(1+\lambda T_{\mathrm{span}})e^{C(h+1)^2}\left(\frac{\log(8RB)}{Be^{-\lambda T_{\mathrm{span}}}}\right)^{1/4} + \frac{Q_h}{R}\log(8RB)\right] + o(1). \tag{154}$$

Neither $\lambda T_0 = O(1)$ nor $\lambda T_{\mathrm{span}} = O(1)$ is assumed. The constant C and the $o(1)$ term are uniform over h and all start-time vectors admitting such a cut. In particular, the approximation error is $o(R/\lambda)$ whenever both displayed terms tend to zero for some choice of h. If no positive cut satisfies (138)*, the finite-system bound in Proposition 21 still applies with $h = 0$.*

This refinement replaces the charge of a global span to every controlled release by a bound on the sum of actual round spans. The proof orders blocks by formation time and uses an exchange argument: a sufficiently earlier unselected block would reduce both load and active count if it replaced a later selected block, contradicting the minimum-load rule. The resulting control of anchor-order inversions absorbs the global-span contribution into the frontier remainder, with the random error displayed in (154). Its effective survivor scale $Be^{-\lambda T_{\mathrm{span}}}$ keeps the dependence on an unbounded span explicit.

Lemma 23 (Availability and feasible exchange)**.** *Under the hypotheses of Proposition 22, fix* $1 \le h \le K_m$ *satisfying* (138) *and a path in the event* $\mathcal{H}_m$ *of* (142)*. Put* $L := \log(8RB)$*,* $S := \lambda T_{\mathrm{span}}$*,* $U := S + \lambda r_{h-1,m}$*, and* $\epsilon := C\sqrt{Le^U/(B-1)}$*, where* C *depends only on the fixed model parameters and is sufficiently large for the uniform path bounds. At a tied timestamp, "before" refers to the processing order of Algorithm 3. For every actual level-*ℓ *round with* $0 \le \ell < h$*:*

(i) Every level-ℓ *block formed before this decision and not previously consumed belongs to the current candidate set* $\mathcal{S}_\ell$*. All its originals are anchored, its* $(m-1)^\ell$ *representatives have empty queues and are nonempty, the block is nonterminal, and every live trajectory has size at most* αC_{th}*.*

(ii) Let H *be a selected block and* G *an unselected block from (i). Write* $M = m^\ell$*. If, for some* $\delta \ge 0$*,*

$$\lambda\left(\max_{i\in G} s_i - \min_{j\in H} s_j\right) \le \delta, \qquad D := 64M\delta + 32\sqrt{M\epsilon} \le \tfrac{1}{2},$$

then

$$\lambda\left(\max_{j\in H} s_j - \min_{i\in G} s_i\right) \le D.$$

Otherwise, replacing H *by* G *preserves every packing test in* (5) *and strictly reduces the selected total load, contradicting the minimum-load rule.*

Part (i) supplies candidate membership on the good event; part (ii) checks load and active-count feasibility for the entire replacement group. Formation alone does not assert that a group is feasible.

Proof. Candidate membership. By (145), the round time t is at most $s_{\max} + r_{\ell,m}$. The formation and survival induction in the proof of Proposition 21 applies on $\mathcal{H}_m$: initial singletons survive until $s_{\max} + r_{0,m}$ or their first repacking, and the outputs forming a level-ℓ block, $1 \le \ell < h$, all survive until $s_{\max} + r_{\ell,m}$ or that block's next repacking. This uses (147); its history-measurable deadline in (146) dominates this endpoint on the anchor-span event. Consequently, a formed block not yet consumed retains all its nonempty representatives at the current decision. Neither a natural release nor an empty output can have made it terminal. Singleton formation is anchor admission, and higher blocks are unions of anchored inputs, so all their queues are empty. The item-size component of $\mathcal{H}_m$, constructed using (120), controls every live trajectory at every time, including after migration. These are exactly the membership conditions for $\mathcal{S}_\ell$. Within a level's packing loop only selected blocks are removed, and a level-ℓ block can be formed only during lower-level processing. Thus a block formed before this decision and not consumed is present when the current group is selected, also for tied formations. This proves (i).

Relative load and count comparison. Every original in such a block has local age $z = \lambda(t - s_i)$ in $[0, U]$. The proportional threshold regime, crossing equations, and controlled-width bound give $\lambda\bar{\sigma}_{\ell,m} \le C(\ell+1)$ and $\lambda(r_{\ell,m} - \bar{\sigma}_{\ell,m}) \le C$, hence $U \le S + C(h+1)$. By (129) and the cutoff, $(B-1)e^{-U} \ge CL$. These facts use the preceding finite-system proof, without any anchor-order conclusion. Let

$$f(z) = (1+z)e^{-z}, \qquad g(z) = e^{-z}.$$

Normalize (59) and (60) by their stationary mean curves. The additive L term is absorbed by the survivor bound just noted, so, simultaneously for every original replica and $0 \le z \le U$,

$$\begin{aligned} \left|\frac{\lambda Z_i(z/\lambda)}{B-1} - f(z)\right| &\le \epsilon f(z), \\ \left|\frac{N_i(z/\lambda)}{B-1} - g(z)\right| &\le \epsilon g(z). \end{aligned} \tag{155}$$

Using relative errors is essential: the exchange comparison below has numerical constants independent of U.

Here is that deterministic comparison. Suppose two vectors $(u_i)_{i=1}^M, (v_i)_{i=1}^M$ lie in $[0, U]^M$ and satisfy $\min_i u_i \geq \max_i v_i - \delta$, where $\delta \geq 0$. Put $d = \max_i u_i$, $c = \min_i v_i$, and $\Delta = d - c$. If

$$D := 64M\delta + 32\sqrt{M\epsilon} \leq \tfrac{1}{2}, \qquad \Delta > D, \tag{156}$$

then

$$\begin{aligned} (1+\epsilon)\sum_i f(u_i) &< (1-\epsilon)\sum_i f(v_i), \\ (1+\epsilon)\sum_i g(u_i) &< (1-\epsilon)\sum_i g(v_i). \end{aligned} \tag{157}$$

To prove this, set $\beta = \min\{\Delta, 1\}$ and $w = \max_i v_i$. The hypotheses imply $w \leq d+\delta$, $\delta < d/(64M)$, $\delta < \beta/(64M)$, and $\delta \leq 1/(128M)$. Pair d with c and pair the remaining entries arbitrarily. For each remaining pair the magnitude of a possible negative contribution to $f(v_i) - f(u_i)$ is at most

$$f((w-\delta)_+) - f(w) \leq \delta e^{\delta} \frac{w}{1+w} f(w) \leq 2\delta \frac{d}{1+d} f(c).$$

For the extreme pair,

$$\log \frac{f(c)}{f(d)} = \int_c^d \frac{z}{1+z}\,\mathrm{d}z \geq \frac{\Delta d}{2(1+d)}, \qquad f(c) - f(d) \geq \frac{\beta}{4}\frac{d}{1+d} f(c).$$

The latter bound follows from $1 - e^{-x} \geq x/2$ for $0 \leq x \leq 1$, applied with $x = \beta d/(2(1+d))$. Consequently, using $d \geq \beta$,

$$\sum_i [f(v_i) - f(u_i)] \geq \frac{\beta}{8}\frac{d}{1+d} f(c) \geq \frac{\beta^2}{16} f(c) > 3M\epsilon f(c).$$

For the counts, the extreme pair contributes at least $\beta g(c)/2$, whereas each possible negative contribution is at most $\delta e^{\delta} g(w) \leq 2\delta g(c)$. Hence

$$\sum_i [g(v_i) - g(u_i)] \geq (\beta/2 - 2M\delta) g(c) \geq \frac{\beta^2}{4} g(c) > 3M\epsilon g(c).$$

Both strict inequalities use $\beta > D \geq 32\sqrt{M\epsilon}$. Finally, $\sum_i [f(u_i) + f(v_i)] \leq M(1+e^{\delta}) f(c) < 3Mf(c)$, and the same bound holds for g. Thus each gap exceeds ϵ times the sum of both sides, proving (157).

Suppose the asserted bound in (ii) fails. Apply the deterministic comparison to the local ages of G and H at the current decision, with $u_i = \lambda(t - s_i)$ for ancestors of G and $v_j = \lambda(t - s_j)$ for ancestors of H. The anchor hypotheses are exactly $\min_i u_i \geq \max_j v_j - \delta$ and $\max_i u_i - \min_j v_j > D$. Since whole-block packing preserves all live ancestral trajectories, the current total load and active count of a nonterminal block G are

$$C_G(t) := \sum_{i \in G} Z_i(t - s_i), \qquad N_G(t) := \sum_{i \in G} N_i(t - s_i).$$

Equations (155) and (157) therefore give $C_G(t) < C_H(t)$ and $N_G(t) < N_H(t)$ simultaneously. Let C_{sel} and N_{sel} be the total load and active count of the selected group, and set $p = (m-1)^{\ell+1}$. Replacing H by G gives

$$\begin{aligned} N_{\mathrm{sel}} - N_H(t) + N_G(t) &< N_{\mathrm{sel}} \leq pB, \\ C_{\mathrm{sel}} - C_H(t) + C_G(t) &< C_{\mathrm{sel}} \leq p(1-\alpha) C_{\mathrm{th}}. \end{aligned}$$

All replacement blocks are distinct current candidates by (i), and their items satisfy the size test. The replacement is therefore among the slot-feasible m-subsets over which the algorithm minimizes load, and it also passes the total-load test. Its strict load decrease contradicts the selected group's minimality, proving (ii). The argument is pathwise on $\mathcal{H}_m$ and adds no conditioning on future anchors or events. □

Proof of Proposition 22. Fix $1 \le h \le K_m$ satisfying (138) and use the same good event $\mathcal{H}_m$ as in Proposition 21. All selected-block arguments below are pathwise on this event. Use L, S, U, ϵ from Lemma 23. Constants depend only on the fixed model parameters. In particular, they do not depend on S, which need not stay bounded. The existence of a positive controlled cut already implies, by (138),

$$(B-1)e^{-S-\lambda\bar{\sigma}_{1,m}} \ge CL, \qquad S \le \log B \le L. \tag{158}$$

This observation will absorb span factors in the frontier and bad-event estimates; it is not an additional startup assumption.

Lemma 23 provides the uniform relative bounds (155) and the feasible exchange comparison, with $U \le S + C(h+1)$.

Anchor order through the hierarchy. Order the level-zero blocks by their anchor times, which are their availability times. Order every higher-level collection by the actual formation order, breaking simultaneous formations according to the algorithm's processing order. Define, only for this proof,

$$\delta_0 = 0, \qquad \delta_{\ell+1} = (1 + 64m^\ell)\delta_\ell + 32\sqrt{m^\ell \epsilon}. \tag{159}$$

First suppose $\delta_h \le 1/2$. We claim that, whenever a level-ℓ block G precedes a distinct block H in this order and $\ell \le h$,

$$\lambda\left(\max_{i\in G} s_i - \min_{j\in H} s_j\right) \le \delta_\ell. \tag{160}$$

The claim is immediate for singletons. For the induction step, consider two parent blocks P, P' formed in this order and let H be the last formed child selected for P. The induction hypothesis gives $\lambda \max_{i\in P} s_i \le \lambda \max_{j\in H} s_j + \delta_\ell$. Take any child G of P'. If G follows H in the child order, applying the induction hypothesis once more bounds $\lambda(\max_{i\in P} s_i - \min_{j\in G} s_j)$ by $2\delta_\ell$, which is at most $\delta_{\ell+1}$. If G precedes H, it was formed before the decision creating P, including in the processing order at a tied timestamp. It has not been consumed: whole blocks never split or reappear, and G is an actual child of the later parent P'. Part (i) of Lemma 23 therefore places G in the candidate set at this decision, with all representatives alive and all size tests satisfied. The induction hypothesis supplies the anchor condition of part (ii). Apply that part with $M = m^\ell$, $\delta = \delta_\ell$, and $D = \delta_{\ell+1} - \delta_\ell \le 1/2$, obtaining $\lambda(\max_{j\in H} s_j - \min_{i\in G} s_i) \le D$. Adding δ_ℓ proves the same bound $\delta_{\ell+1}$. Taking the minimum over the children of P' completes the induction. Children formed after the earlier parent fall in the first case, which uses no availability claim about them at the earlier time.

Expanding the recursion and using $1 + 64m^\ell \le 65m^\ell$ gives

$$\delta_h \le Ch65^h m^{h(h-1)/2}\sqrt{\epsilon} \le Ce^{C(h+1)^2}\sqrt{\epsilon}.$$

The auxiliary restriction $\delta_h \le 1/2$ can now be removed. If it holds, every inversion through level h is at most δ_h. If it fails, every inversion is at most $S \le 2S\delta_h$ simply because all anchors lie in one interval of length S/λ on $\mathcal{H}_m$. Thus in either case a common radius is $2(1+S)\delta_h$. Thus (160) holds at every $\ell \le h$ with δ_ℓ replaced by $2(1+S)\delta_h$. The definition of ϵ and $U \le S + C(h+1)$ give

$$2(1+S)\delta_h \le C(1+S)e^{C(h+1)^2}\left(\frac{Le^S}{B}\right)^{1/4}. \tag{161}$$

No factor depending on S has been incorporated into the constants.

Summing the actual round spans. For a level-ℓ round v, retain $x_v, y_v, \bar{s}_v$ from the preceding proof. The local window (145) gives the sharper form

$$|t_v - \bar{s}_v - \bar{\sigma}_{\ell,m}| \le y_v - x_v + Cw_{\ell,m}. \tag{162}$$

List these rounds as $v_1, \dots, v_b$ in formation order, where $b = b_{\ell+1}$. Applying the common anchor-order bound to their output blocks yields

$$\lambda\sum_{j=1}^{b}(y_{v_j} - x_{v_j}) = \lambda(y_{v_b} - x_{v_1}) + \lambda\sum_{j=1}^{b-1}(y_{v_j} - x_{v_{j+1}}) \le S + 2(b-1)(1+S)\delta_h.$$

Thus the global span is paid once at each level, with the release multiplicity $(m-1)^\ell$, rather than once for every round. Summing gives

$$\begin{aligned}
&\sum_{\ell<h}\sum_{v:\mathrm{level}(v)=\ell}(m-1)^\ell(y_v-x_v)\\
&\quad\le T_{\mathrm{span}}\sum_{\ell<h}(m-1)^\ell+\frac{2(1+S)\delta_h}{\lambda}\sum_{\ell<h}\nu_{\ell,m} \qquad (163)\\
&\quad\le C_mQ_hT_{\mathrm{span}}+\frac{2R(1+S)\delta_h}{\lambda}.
\end{aligned}$$

Here $h\le K_m$ implies $b_h\ge 1$, so $Q_h\ge(m-1)^h$, and $\sum_{\ell<h}(m-1)^\ell\le C_m(m-1)^h$ for fixed $m\ge 3$.

The exact anchor accounting (151) applies at every admissible cut h. Combining it with (162)–(163) controls both sides of the comparison with the controlled hierarchy sum. The remaining natural and repacking releases still obey (153). The extra Q_hT_{span} in (163) is absorbed there. By (158), their normalized contribution is at most CQ_hL/R. The omitted deterministic levels have the same bound: $\Phi(\rho_{\ell,m})\le C(\ell+1)\le CL$ and $\sum_{\ell\ge h}\nu_{\ell,m}\le Q_h$.

The width sum, exact-crossing correction, and bad-event estimates are unchanged. Using $S\le L$, after multiplication by λ/R they are bounded by a constant times the sum of

$$\frac{L}{\sqrt{B}},\qquad RBL\epsilon_N,\qquad L(8RB)^{-3/2},\qquad L\sqrt{RB(1+\lambda\alpha C_{\mathrm{th}})e^{-\lambda\alpha C_{\mathrm{th}}}}.$$

All four tend to zero. Indeed, the batch-size scaling gives $L^2/B=(BL/N)(NL/B^2)=o(1)$ and $N/(BL)\to\infty$, while $\lambda\alpha C_{\mathrm{th}}\ge\alpha c_-B$. These bounds are uniform in h and in the start-time vectors admitting a positive controlled cut. Taking expectations, using (161), and including the omitted deterministic levels proves (154). No new probability event or conditioning on the random anchor vector has been introduced. □

Asymptotic part of Proposition 21. Assume the stochastic asymptotic regime of the Scaling assumptions paragraph. Write $L=\log(8RB)$ and $S=\lambda T_{\mathrm{span}}$ within this proof. The batch-size scaling and (25) give $S=\lambda T_0+o(1)=o(\log B)$. Under $\log R=\Theta(\log B)$, one has $L=\Theta(\log B)$ and $K_m=\Theta(\log B)$. Choose

$$h=\left\lceil\frac{3\log\log B}{\log(m/(m-1))}\right\rceil.$$

Then $h<K_m$ for large B, and $((m-1)/m)^h$ is within fixed constant factors of $(\log B)^{-3}$. The exact crossing equation and proportional threshold scale give

$$(B-1)e^{-\lambda(\bar\sigma_{h,m}+T_{\mathrm{span}})}\ge\frac{cBe^{-S}}{h+1}\left(\frac{m-1}{m}\right)^{h+1}\ge\frac{cBe^{-S}}{(\log B)^3\log\log B}.$$

Since $e^{-S}=B^{-o(1)}$, this is larger than any fixed multiple of L for all sufficiently large B. The crossing times increase with the level, so the survivor condition holds at every level below h, and h satisfies (138). Equation (34), with $h\le K_m$, implies

$$\frac{Q_h}{R}\le C_m\left(\frac{m-1}{m}\right)^h,\qquad\frac{Q_h}{R}L=O((\log B)^{-2}).$$

Finally,

$$(1+S)e^{C(h+1)^2}\left(\frac{Le^S}{B}\right)^{1/4}=B^{-1/4+o(1)}\longrightarrow 0.$$

Proposition 22 now gives (141). □

B.6 Transfer to the implemented policy and hierarchy limits

B.6.1 Multi-repacking comparison and general switch depths

The following result combines Proposition 21 with a pathwise comparison for Algorithm 1.

Lemma 24 (Switch-transfer bound). *Fix $0 \le d \le K_m$ and couple Algorithms 1 and 3 using the same original service paths, physical labels, priorities, and tie-breaking up to the switch. Let τ be multi-repacking's first certified switching epoch, with $\tau = \infty$ if there is no switch. On every such realization with finite service durations, the two occupied times agree when $\tau = \infty$, and in all cases*

$$\left|T_{\mathrm{tail}}^{\mathrm{multi}}(m,d) - T_{\mathrm{tail}}^{\mathrm{blk}}(m)\right| \le Q_d \left(s_{\max} - s_{\min} + \max_{i,j} \tilde{Y}_{ij}\right). \tag{164}$$

This comparison requires neither a good event nor a bound on the anchor spread. Future anchors are included in the count Q_d.

The lemma bounds the difference of the two occupancy integrals after their common prefix; it makes no claim about which policy is faster on an individual path.

Proof. Every event closure is finite because each packing round releases at least one physical label and released labels never return. For $d > 0$, the common prefix ends after the full reference closure at the first event satisfying the round-count certificate; for $d = 0$, it ends after the first anchor and tied natural releases, before either policy packs at that timestamp. These are precisely the switching conventions in Algorithm 4. The test uses only observed events and completed counters, so τ is determined by the current history. If it never succeeds, both policies execute the reference throughout and their occupied times coincide.

Suppose $\tau < \infty$, and let $\mathcal{U}$ be the set of original physical labels not yet released at the end of this common prefix, before any divergent operation at time τ. This includes labels that have not started or have not reached their anchors. The two executions have exactly the same set $\mathcal{U}$ at this point. To count it, the whole-block invariant holds without a probabilistic event: an eligible level-ℓ block has $(m-1)^\ell$ nonempty representatives. Indeed, it starts from anchored singletons, and a round on m such blocks retains $(m-1)^{\ell+1}$ representatives and releases the other $(m-1)^\ell$. Any empty output is released in addition and makes the parent terminal, so it cannot invalidate this invariant at a later accepted round. Natural releases likewise make a block terminal. For $d > 0$, the certified rounds therefore release at least

$$\sum_{\ell=0}^{d-1} (m-1)^\ell \left\lfloor \frac{R}{m^{\ell+1}} \right\rfloor = R - Q_d$$

distinct physical labels. Distinctness across rounds follows from the no-reactivation rule. Natural releases, empty outputs, and higher-level rounds only increase this count. Thus $M := |\mathcal{U}| \le Q_d$; the same conclusion holds for $d = 0$ because $Q_0 = R$.

After the switch, each policy's unreleased set is a subset of $\mathcal{U}$: packing uses only its current unreleased representatives, and natural releases remove labels permanently. A future anchor admits its original label, already counted in $\mathcal{U}$, rather than creating a new one. For either policy $\mathcal{M}$, define its tail occupancy by

$$A_{\mathcal{M}}(t) := \sum_{i \in [R]} \mathbf{1}_{\{s_i \le t < t_{i,\mathrm{free}}^{\mathcal{M}}\}}.$$

The anchor epochs are common to both executions because only anchored replicas can repack. The definition of the objective gives $T_{\mathrm{tail}}^{\mathcal{M}} = \int_{\mathbb{R}} A_{\mathcal{M}}(t)\,dt$. The two integrands agree for $t < \tau$, and each lies in $[0, M]$ for $t \ge \tau$. Migration preserves service completion times, so both vanish by $F := \max_{i,j}(s_i + \tilde{Y}_{ij})$; by then every anchor has occurred and every remaining replica is empty and released. Since $s_{\min} \le \tau \le F$, we obtain

$$\begin{aligned}
\left|T_{\mathrm{tail}}^{\mathrm{multi}}(m,d) - T_{\mathrm{tail}}^{\mathrm{blk}}(m)\right| &= \left|\int_\tau^F \left(A_{\mathrm{multi}}(t) - A_{\mathrm{blk}}(t)\right) dt\right| \\
&\le M(F - \tau) \\
&\le Q_d \left(s_{\max} - s_{\min} + \max_{i,j} \tilde{Y}_{ij}\right).
\end{aligned}$$

The bound also holds on no-switch paths, where the difference is zero. All statements are pathwise; no future anchor maximum or residual lifetime is conditioned on the switching history. □

Proposition 25 (Tail occupied replica time of multi-repacking)**.** *Suppose $B \geq 2$, $R \geq m$, and $C_{\mathrm{th}}\lambda/B \leq 1-\eta$ for a fixed $\eta > 0$. Use the actual finite-batch anchors, T_{span} from* (25)*, and Algorithm 1 with deterministic depth $0 \leq d \leq K_m$. For N/B sufficiently large and B sufficiently large in terms of fixed m, α, η, every cutoff $0 \leq h \leq d$ satisfying* (138) *when $h > 0$ obeys*

$$\left| \frac{\mathbb{E}[T_{\mathrm{tail}}^{\mathrm{multi}}(m,d)]}{R} - \frac{1}{R\lambda} \sum_{\ell=0}^{K_m-1} \nu_{\ell,m}\Phi(\rho_{\ell,m}) \right| \leq C\mathcal{E}_m(h). \tag{165}$$

Under the stochastic asymptotic regime of the Scaling assumptions paragraph, if

$$\frac{Q_d}{R}\log(RB) \longrightarrow 0, \tag{166}$$

then

$$\mathbb{E}[T_{\mathrm{tail}}^{\mathrm{multi}}(m,d)] = \frac{1}{\lambda} \sum_{\ell=0}^{K_m-1} \nu_{\ell,m}\Phi(\rho_{\ell,m}) + o(R/\lambda). \tag{167}$$

The default $d = \lceil K_m/2 \rceil$ and $d = K_m$ satisfy (166) *for fixed m. The proposition neither requires d to satisfy the cutoff condition nor assumes a switch on every path.*

Proof. Apply Lemma 24 with the prescribed depth d. It covers both switching and no-switch paths and gives (164) without any cutoff assumption.

In the exponential model the $R(B-1)$ anchor residual lifetimes are i.i.d. $\mathrm{Exp}(\lambda)$, so their maximum has mean $H_{R(B-1)}/\lambda$. Taking expectations in the pathwise bound gives

$$\frac{\left|\mathbb{E}T_{\mathrm{tail}}^{\mathrm{multi}}(m,d) - \mathbb{E}T_{\mathrm{tail}}^{\mathrm{blk}}(m)\right|}{R} \leq \frac{Q_d}{R}\left[\mathbb{E}(s_{\max} - s_{\min}) + \frac{H_{R(B-1)}}{\lambda}\right]. \tag{168}$$

There is no independence requirement between τ, the anchor spread, and the maximum residual lifetime. Lemma 5, Cauchy–Schwarz, and $H_n \leq 1 + \log n$ bound the right-hand side by $C(Q_d/R)[T_{\mathrm{span}} + (1+\log(RB))/\lambda]$. Combining this estimate with Proposition 21 yields, for every admissible proof cutoff h,

$$\begin{aligned} &\left| \frac{\mathbb{E}T_{\mathrm{tail}}^{\mathrm{multi}}(m,d)}{R} - \frac{1}{R\lambda} \sum_{\ell=0}^{K_m-1} \nu_{\ell,m}\Phi(\rho_{\ell,m}) \right| \\ &\qquad \leq C\mathcal{E}_m(h) + C\frac{Q_d}{R}\left[T_{\mathrm{span}} + \frac{1+\log(RB)}{\lambda}\right]. \end{aligned} \tag{169}$$

The counts Q_r decrease with r. Therefore, if $h \leq d$, then $Q_d \leq Q_h$ and the additional term in (169) is bounded by a constant times the existing frontier term of $\mathcal{E}_m(h)$. This proves the stated multi-repacking finite-system bound (165) with the same aggregate error and a larger constant. Subtracting each expected occupied time from RH_{B-1}/λ gives the asserted saving estimate. For the leading-order claim use the already proved refined asymptotic part of Proposition 21, without requiring $\mathcal{E}_m(h) = o(1/\lambda)$. The standing regime gives $\lambda T_{\mathrm{span}} = o(\log B)$ and $\log R = \Theta(\log B)$. Hence (166) makes the additional comparison error $o(1/\lambda)$. Combining it with (141) proves (167) for the proposed policy.

For completeness, the asserted switch-depth choices follow from the exact finite-system counts. Write the base-m digits below depth d as $a_0, \ldots, a_{d-1} \in \{0, \ldots, m-1\}$ and $b_d = \lfloor R/m^d \rfloor$. Equation 119 gives

$$Q_d = \sum_{\ell<d} a_\ell (m-1)^\ell + b_d(m-1)^d \leq \frac{m-1}{m-2}(m-1)^d + R\left(\frac{m-1}{m}\right)^d.$$

Since $m^d \leq R$ for $d \leq K_m$, this implies $Q_d/R \leq C_m((m-1)/m)^d$. If $\beta_m := \log_m(m/(m-1)) > 0$, then

$$d = \lceil K_m/2 \rceil \quad \Longrightarrow \quad \frac{Q_d}{R} \leq C_m R^{-\beta_m/2}, \qquad d = K_m \quad \Longrightarrow \quad \frac{Q_d}{R} \leq C_m R^{-\beta_m}.$$

Both satisfy (166) under $\log R = \Theta(\log B)$. Conversely, for any fixed d and fixed m, $Q_d/R \to ((m-1)/m)^d > 0$ as $R \to \infty$, so a fixed early switch does not meet that sufficient condition. Finally, no requirement that d be a valid concentration cutoff has been used: Proposition 21 is applied to the reference process, and the proposed policy changes only after the actual observable certificate. This completes the transfer proof. □

B.6.2 From finite hierarchies to the main formulas

The finite-system propositions already account for unmatched blocks and natural releases. It remains to replace their deterministic, integer-count sums by the series displayed in the main text.

Lemma 26 (Rounding and hierarchy-tail errors)**.** *Under* (6)*, for sufficiently large B, the pairwise counts satisfy, for every integer $R \geq 2$,*

$$\begin{aligned} 0 &\leq \sum_{\ell=0}^{\infty} 2^{-\ell-1}\Phi(\rho_\ell) - \frac{1}{R}\sum_{\ell=0}^{K_2-1} \nu_\ell \Phi(\rho_\ell) \\ &\leq C\frac{(1+\log R)^2}{R}. \end{aligned} \tag{170}$$

For fixed $m \geq 3$ and $\alpha \in (0,1)$, the multi counts satisfy, for every integer $R \geq m$,

$$\begin{aligned} 0 &\leq \sum_{\ell=0}^{\infty} \frac{(m-1)^\ell}{m^{\ell+1}}\Phi(\rho_{\ell,m}) - \frac{1}{R}\sum_{\ell=0}^{K_m-1} \nu_{\ell,m} \Phi(\rho_{\ell,m}) \\ &\leq C\frac{1+\log R}{R^{1-\log_m(m-1)}}. \end{aligned} \tag{171}$$

The constants are independent of R, B, N, λ, depending only on c_-, η and, in the multi bound, fixed m, α.

Proof. The threshold upper margin ensures $0 < \rho_\ell \leq 1-\eta$. For multi, $\rho_{0,m} \leq ((m-1)/m)(1-\alpha)(1-\eta)B/(B-1) < 1$ for sufficiently large B, and the ratios decrease with ℓ. Thus all evaluations of Φ lie on its real branch. If $z = \Phi(\rho) \geq 0$, then

$$(1+z)e^{-z} = \rho, \qquad \log(1/\rho) = z - \log(1+z).$$

Since $\log(1+z) \leq z/2 + \log 2$ for $z \geq 0$,

$$0 \leq \Phi(\rho) \leq 2\log(1/\rho) + 2\log 2.$$

The threshold lower bound gives

$$\rho_\ell \geq c_- 2^{-\ell-1}, \qquad \rho_{\ell,m} \geq (1-\alpha)c_- \left(\frac{m-1}{m}\right)^{\ell+1}.$$

Consequently both $\Phi(\rho_\ell)$ and $\Phi(\rho_{\ell,m})$ are bounded by $C(\ell+1)$, uniformly over all levels and system sizes under consideration. In particular, both series converge.

For pairwise, subtracting the finite sum splits the difference exactly into a rounding term and a hierarchy tail:

$$\frac{1}{R}\sum_{\ell=0}^{K_2-1}\left(\frac{R}{2^{\ell+1}} - \left\lfloor \frac{R}{2^{\ell+1}} \right\rfloor\right)\Phi(\rho_\ell) + \sum_{\ell=K_2}^{\infty} 2^{-\ell-1}\Phi(\rho_\ell).$$

Both terms are nonnegative. The rounding term is at most $CR^{-1}\sum_{\ell<K_2}(\ell+1) \leq C(K_2+1)^2/R$. The tail is at most $C(K_2+2)2^{-K_2}$. Since $2^{K_2} \leq R < 2^{K_2+1}$, these bounds give (170).

For multi, the corresponding exact decomposition is

$$\begin{aligned} &\frac{1}{R}\sum_{\ell=0}^{K_m-1}(m-1)^\ell\left(\frac{R}{m^{\ell+1}} - \left\lfloor \frac{R}{m^{\ell+1}} \right\rfloor\right)\Phi(\rho_{\ell,m}) \\ &\qquad + \sum_{\ell=K_m}^{\infty}\frac{(m-1)^\ell}{m^{\ell+1}}\Phi(\rho_{\ell,m}). \end{aligned}$$

The rounding term is bounded by

$$\frac{C}{R}\sum_{\ell<K_m}(\ell+1)(m-1)^\ell \leq \frac{C(K_m+1)(m-1)^{K_m}}{R}.$$

Writing $r = (m-1)/m \in (0,1)$ within this proof, the tail is at most $C(K_m+1)r^{K_m}$. Moreover,

$$\frac{(m-1)^{K_m}}{R} \le r^{K_m} \le r^{-1} R^{\log_m(m-1)-1},$$

using $m^{K_m} \le R < m^{K_m+1}$. This proves (171). □

Proof of Theorems 2 and 3. Under the stochastic asymptotic regime, $R \to \infty$. Proposition 19 gives

$$\mathbb{E}[T_{\text{tail}}^{\text{pair}}] = \frac{1}{\lambda} \sum_{\ell=0}^{K_2-1} \nu_\ell \Phi(\rho_\ell) + o(R/\lambda).$$

Multiplying (170) by R/λ changes this finite sum to its infinite-series expression at a cost $O((1+\log R)^2/\lambda) = o(R/\lambda)$. Substitution of $\rho_\ell = \lambda C_{\text{th}}/[2^{\ell+1}(B-1)]$ proves (17).

For default multi-repacking, Appendix B.6.1 proves that $d = \lceil K_m/2 \rceil$ satisfies (166). Proposition 25 therefore gives (167). Multiplying (171) by R/λ adds at most

$$\frac{CR}{\lambda} \frac{1+\log R}{R^{1-\log_m(m-1)}} = o(R/\lambda),$$

because m is fixed and $1 - \log_m(m-1) > 0$. Substituting the expression for $\rho_{\ell,m}$ from (18) proves (21). The same series conversion applies to every depth satisfying (166); no additional scaling assumption is used. □

B.7 Strict comparison and tail implications

B.7.1 Strict comparison of multi and pairwise repacking

The comparison uses the exact infinite hierarchy sums, with a positive gap uniform in the normalized threshold.

Lemma 27 (Uniform gap between the hierarchy sums)**.** *Fix $m \ge 3$ and α satisfying* (22)*. Put $q = (m-1)/m$ and $\delta := \log[4(1-\alpha)q^m] > 0$. For $0 < \rho \le 1$, define*

$$\begin{aligned} S_2(\rho) &:= \sum_{k=1}^{\infty} 2^{-k} \Phi(\rho\, 2^{-k}), \\ S_m(\rho,\alpha) &:= \sum_{k=1}^{\infty} (1-q) q^{k-1} \Phi((1-\alpha)\rho q^k). \end{aligned} \tag{172}$$

Then $S_2(\rho) - S_m(\rho,\alpha) \ge \delta$.

Proof. Set $g(u) := \Phi(e^{-u})$ for $u \ge 0$, with $g(0) = 0$. Its defining identity is $g(u) - \log(1+g(u)) = u$. For $u > 0$, $g'(u) = 1 + 1/g(u)$. Writing $v(u) = 1/g(u)$ gives $v' = -v^2 - v^3$. Inductively, $(-1)^n v^{(n)}(u) = P_n(v(u))$, where $P_0(z) = z$ and $P_{n+1}(z) = (z^2+z^3)P_n'(z)$. Every P_n has nonnegative coefficients, so v is completely monotone. Since $g(u) - u = \log(1+g(u)) \ge 0$ and its derivative is v, the Bernstein representation theorem (Schilling et al., 2012, Theorem 3.2) gives a positive measure μ on $(0,\infty)$ such that

$$g(u) = u + \int_0^\infty (1 - e^{-tu})\, \mu(dt), \qquad \int_0^\infty \min\{1,t\}\, \mu(dt) < \infty. \tag{173}$$

The constant and linear parts of $g(u) - u$ vanish because $g(0) = 0$ and $v(u) \to 0$ as $u \to \infty$. Also, $\log(1+z) \le z/2 + \log 2$ for $z \ge 0$ implies $g(u) \le 2u + 2\log 2$.

Let $x = -\log\rho$, $a = -\log q$, $b = \log 2$, and $s = -\log(1-\alpha)$, so $0 < a < b$. Take auxiliary geometric variables V_2, V_m with $\mathbb{P}(V_2 = k) = 2^{-k}$ and $\mathbb{P}(V_m = k) = (1-q)q^{k-1}$ for $k \ge 1$. The two sums become

$$S_2(\rho) = \mathbb{E}[g(x + bV_2)], \qquad S_m(\rho,\alpha) = \mathbb{E}[g(x + s + aV_m)]. \tag{174}$$

Both expectations are finite by the linear bound on g. The corresponding Laplace transforms are

$$\begin{aligned} L_2(t) &:= \mathbb{E}[e^{-tbV_2}] = \frac{e^b - 1}{e^{(t+1)b} - 1}, \\ L_m(t) &:= \mathbb{E}[e^{-t(s+aV_m)}] = e^{-st}\frac{e^a - 1}{e^{(t+1)a} - 1}, \qquad t \geq 0. \end{aligned} \tag{175}$$

For $F(t) := \log(L_m(t)/L_2(t))$, direct differentiation gives $F(0) = 0$ and

$$F'(0) = 2b - ma - s = \delta, \qquad F''(t) = \frac{a^2}{4\sinh^2((t+1)a/2)} - \frac{b^2}{4\sinh^2((t+1)b/2)} > 0.$$

The last inequality follows because $z/\sinh z$ is strictly decreasing for $z > 0$ and $a < b$. Hence $F(t) \geq \delta t \geq 0$, so $L_m(t) \geq L_2(t)$ for every $t \geq 0$.

Apply (173) to (174). Tonelli's theorem and finiteness of the expectations justify integrating first and then subtracting. Since $\mathbb{E}[bV_2] - \mathbb{E}[s + aV_m] = 2b - ma - s = \delta$,

$$S_2(\rho) - S_m(\rho, \alpha) = \delta + \int_0^\infty e^{-tx}\big(L_m(t) - L_2(t)\big)\,\mu(dt) \geq \delta.$$

This proves the uniform gap. □

Proof of Proposition 4. Under (6), $\rho := \lambda C_{\mathrm{th}}/(B-1)$ belongs to $(0,1)$ for all sufficiently large B. Reindexing (17) and (21) by $k = \ell + 1$ gives

$$\frac{\lambda}{R}\left(\mathbb{E}[T_{\mathrm{tail}}^{\mathrm{pair}}] - \mathbb{E}[T_{\mathrm{tail}}^{\mathrm{multi}}]\right) = S_2(\rho) - S_m(\rho, \alpha) + o(1).$$

Lemma 27 bounds the right-hand side below by $\delta + o(1)$, proving (23). In particular, for all sufficiently large systems multi-repacking saves at least $\delta R/(2\lambda)$ additional occupied replica time relative to pairwise. The saving comparison follows from the common no-repacking term in (1). The switch error is already included in (21); no additional switching assumption is needed. □

B.7.2 Derivation of the Tail and Whole-Batch Implications

For completeness, the exact implications of the main-text comparison are recorded here. Under the stochastic asymptotic regime,

$$\frac{\mathbb{E}[\Delta^{\mathcal{M}}]}{R} = \frac{\log B + O(1)}{\lambda}, \qquad \frac{\mathbb{E}[\Delta^{\mathcal{M}}]}{\mathbb{E}[T_{\mathrm{tail}}^{\mathrm{no}}]} = 1 - \Theta\left(\frac{1}{\log B}\right). \tag{176}$$

For every $N \geq B \geq 2$, common pre-anchor execution gives the baseline and upper bound

$$\begin{aligned} \frac{\mathbb{E}[T_{\mathrm{total}}^{\mathrm{no}}]}{R} &= \frac{1}{\lambda}\left(\frac{N - B + 1}{B} + H_{B-1}\right), \\ \frac{\mathbb{E}[\Delta^{\mathcal{M}}]}{\mathbb{E}[T_{\mathrm{total}}^{\mathrm{no}}]} &\leq \frac{H_{B-1}}{(N - B + 1)/B + H_{B-1}}. \end{aligned} \tag{177}$$

In the stochastic asymptotic regime this yields

$$\frac{\mathbb{E}[\Delta^{\mathcal{M}}]}{\mathbb{E}[T_{\mathrm{total}}^{\mathrm{no}}]} = (1 + o(1))\frac{B \log B}{N} = o(1). \tag{178}$$

This subsection justifies the consequences stated in Section 4.5; it adds no scaling assumption.

Proof. Assume first the stochastic asymptotic regime. For $x = \Phi(\rho) \geq 0$, the defining equation $(1 + x)e^{-x} = \rho$ gives $x - \log(1 + x) = \log(1/\rho)$. Since $\log(1 + x) \leq x/2$ for $x \geq 3$,

$$0 \leq \Phi(\rho) \leq 3 + 2\log(1/\rho), \qquad 0 < \rho \leq 1.$$

The proportional threshold lower bound and fixed m, α imply $\log(1/\rho_\ell) = O(\ell + 1)$ and $\log(1/\rho_{\ell,m}) = O(\ell + 1)$ uniformly in B sufficiently large. The upper threshold margin ensures that

both level-zero ratios are bounded away from one. Since Φ is positive and decreasing on $(0,1)$, their level-zero crossings are bounded below by a positive constant. Moreover,

$$\frac{\nu_\ell}{R} \le 2^{-\ell-1}, \qquad \frac{\nu_{\ell,m}}{R} \le \frac{1}{m}\left(\frac{m-1}{m}\right)^{\ell}, \qquad \frac{\nu_0}{R} \longrightarrow \frac{1}{2}, \qquad \frac{\nu_{0,m}}{R} \longrightarrow \frac{1}{m}.$$

Thus each normalized finite hierarchy sum is bounded above by a convergent series of a linear factor times geometric weights, and bounded below by its positive level-zero contribution. Applying (89) and (167), whose normalized errors are $o(1/\lambda)$, proves the repacking statement in (24). The exact no-repacking expectation (9) and $H_{B-1} = \log B + O(1)$ prove its baseline statement and (176). The multi leading formula used here is already the conclusion for the proposed policy, proved in Appendix B.6.1. In particular, the saving is positive in expectation for sufficiently large B in this regime; no pathwise nonnegative-saving assumption has been used.

For the whole-batch calculation, the first $k = N - B + 1$ completion gaps of each replica are independent $\mathrm{Exp}(B\lambda)$ variables, as there are B busy slots before the kth completion. Hence, exactly,

$$\mathbb{E}[s_i - t_i] = \frac{N - B + 1}{B\lambda}.$$

This expectation holds for every $N \ge B$ and does not require independence between the anchor time and the subsequent occupied time. Linearity of expectation in (1) and the harmonic baseline give the first identity in (177). Since $T^{\mathcal{M}}_{\mathrm{tail}} \ge 0$, one has $\Delta^{\mathcal{M}} \le T^{\mathrm{no}}_{\mathrm{tail}}$ pathwise, which gives the displayed upper bound after taking expectations and dividing by the positive baseline expectation. This upper bound applies even to mechanisms with a negative expected saving at a finite size.

Finally, in the stochastic asymptotic regime, $\log(RB) = \Theta(\log B)$ and $N/(B\log(RB)) \to \infty$, so

$$\frac{N - B + 1}{B} + H_{B-1} = \frac{N}{B}(1 + o(1)).$$

Substituting (176) in the exact ratio gives

$$\frac{\mathbb{E}[\Delta^{\mathcal{M}}]}{\mathbb{E}[T^{\mathrm{no}}_{\mathrm{total}}]} = \frac{\log B + O(1)}{(N - B + 1)/B + H_{B-1}} = (1 + o(1))\frac{B\log B}{N} \longrightarrow 0,$$

which is (178). □

C Deterministic Comparison

C.1 Leading-Order Tail Occupied Replica Time

We compare the deterministic hierarchy terms for no repacking, pairwise repacking, and multi-repacking. Theorems 2 and 3 identify the infinite series as leading terms under the stochastic asymptotic regime. Propositions 19 and 25 retain the finite-system bounds. This section develops further approximations to the deterministic terms; outside the stated regime, they remain hierarchy-model comparisons.

Recall that

$$\Phi(\rho) := -W_{-1}\left(-\frac{\rho}{e}\right) - 1.$$

Without repacking, every initial replica remains occupied until all of its $B - 1$ residual trajectories are completed. The corresponding expected tail occupied replica time is

$$\widehat{T}_{\mathrm{no}} = R\frac{H_{B-1}}{\lambda}.$$

For the deterministic comparison, we replace the integer release counts by continuous weights and extend both finite hierarchy sums to infinite series, continuing the formulas for ρ_ℓ and $\rho_{\ell,m}$ algebraically to every integer $\ell \ge 0$. For pairwise repacking, this replaces $\nu_\ell = \lfloor R/2^{\ell+1} \rfloor$ by $R/2^{\ell+1}$ at every $\ell \ge 0$. The resulting idealized tail occupied replica time is

$$\widehat{T}_{\mathrm{p}} = \frac{R}{\lambda}\sum_{\ell \ge 0}\frac{1}{2^{\ell+1}}\Phi\left(\rho_\ell\right).$$

The multi leading term inherited from the reference uses the nominal release multiplicity $\nu_{\ell,m}$ from (33). The continuous-hierarchy relaxation replaces the floor by its continuous value,

$$\nu_{\ell,m}^{\mathrm{cont}} = \frac{R(m-1)^{\ell}}{m^{\ell+1}},$$

and uses this expression for all $\ell \geq 0$. Thus, for merge ratio $m \geq 3$, the leading-order tail occupied replica time is

$$\widehat{T}_m = \frac{R}{\lambda} \sum_{\ell \geq 0} \frac{(m-1)^{\ell}}{m^{\ell+1}} \Phi\left(\rho_{\ell,m}\right).$$

For a generic mechanism $\mathcal{M}$ among these cases, $\widehat{T}_{\mathcal{M}}$ denotes the corresponding leading-order tail occupied replica time.

C.2 Closed-Form Approximation of the Hierarchy Sums

The closed forms summarize the geometric hierarchy weights through their mean levels: 1 for pairwise repacking and $m-1$ for multi-repacking. These means determine the expansion centers below.

Define

$$\Lambda_{\mathrm{p}} := 1 - \log\left(\rho_1\right)$$

and, for $m \geq 3$,

$$a_m := \log \frac{m}{m-1}, \qquad \Lambda_m := 1 - \log\left(\rho_{m-1,m}\right).$$

Also define the limiting normalized threshold ratio and its logarithmic transform by

$$\rho_{\infty} := \frac{(1-\alpha)C_{\mathrm{th}}\lambda}{B-1}, \qquad \Lambda_{\infty} := 2 - \log \rho_{\infty}.$$

Theorem 28 (Closed-form approximation of tail occupied replica time)**.** *Assume $0 < \rho_0 < 1$ and, for each finite merge ratio under consideration, $0 < \rho_{0,m} < 1$. For the large-merge-ratio statement, assume additionally that $\rho_{\infty} \in (0,1)$. Under the continuous-hierarchy relaxation, the pairwise leading-order tail occupied replica time admits the following approximation:*

$$\begin{aligned}
\widetilde{T}_{\mathrm{p}} := \frac{R}{\lambda}\Bigg\{ & \Lambda_{\mathrm{p}} - 1 + \log \Lambda_{\mathrm{p}} + \frac{\log \Lambda_{\mathrm{p}}}{\Lambda_{\mathrm{p}}} \\
& + \frac{\log \Lambda_{\mathrm{p}} \left(2 - \log \Lambda_{\mathrm{p}}\right)}{2\Lambda_{\mathrm{p}}^2} \\
& + (\log 2)^2 \Bigg[-\frac{1}{\Lambda_{\mathrm{p}}^2} + \frac{2\log \Lambda_{\mathrm{p}} - 3}{\Lambda_{\mathrm{p}}^3} \\
& \qquad + \frac{-3\left(\log \Lambda_{\mathrm{p}}\right)^2 + 11 \log \Lambda_{\mathrm{p}} - 6}{\Lambda_{\mathrm{p}}^4} \Bigg] \Bigg\}.
\end{aligned} \tag{179}$$

More precisely, as $\Lambda_{\mathrm{p}} \to \infty$,

$$\widehat{T}_{\mathrm{p}} = \widetilde{T}_{\mathrm{p}} + O\left(\frac{R}{\lambda} \frac{\left(\log \Lambda_{\mathrm{p}}\right)^3}{\Lambda_{\mathrm{p}}^3} \right). \tag{180}$$

For multi-repacking with merge ratio $m \geq 3$, the corresponding approximation is

$$\begin{aligned}
\widetilde{T}_m := \frac{R}{\lambda}\Bigg\{ & \Lambda_m - 1 + \log \Lambda_m + \frac{\log \Lambda_m}{\Lambda_m} \\
& + \frac{\log \Lambda_m \left(2 - \log \Lambda_m\right)}{2\Lambda_m^2} \\
& + \frac{m(m-1)}{2} \left(\log \frac{m}{m-1} \right)^2 \Bigg[-\frac{1}{\Lambda_m^2} + \frac{2\log \Lambda_m - 3}{\Lambda_m^3} \\
& \qquad + \frac{-3\left(\log \Lambda_m\right)^2 + 11 \log \Lambda_m - 6}{\Lambda_m^4} \Bigg] \Bigg\}.
\end{aligned} \tag{181}$$

For each fixed m, as $\Lambda_m \to \infty$,

$$\widehat{T}_m = \widetilde{T}_m + O_m\left(\frac{R}{\lambda}\frac{(\log \Lambda_m)^3}{\Lambda_m^3}\right). \tag{182}$$

The large-merge-ratio limit admits the approximation

$$\begin{aligned}\widetilde{T}_\infty := \frac{R}{\lambda}\Bigg\{ & \Lambda_\infty - 1 + \log \Lambda_\infty + \frac{\log \Lambda_\infty}{\Lambda_\infty} \\ & + \frac{\log \Lambda_\infty \,(2 - \log \Lambda_\infty)}{2\Lambda_\infty^2} \\ & + \frac{1}{2}\Bigg[- \frac{1}{\Lambda_\infty^2} + \frac{2\log \Lambda_\infty - 3}{\Lambda_\infty^3} \\ & \qquad + \frac{-3\,(\log \Lambda_\infty)^2 + 11 \log \Lambda_\infty - 6}{\Lambda_\infty^4}\Bigg]\Bigg\}. \end{aligned} \tag{183}$$

Writing $\widehat{T}_\infty := \lim_{m\to\infty} \widehat{T}_m$, the exact continuous-hierarchy limit exists, and, as $\Lambda_\infty \to \infty$,

$$\widehat{T}_\infty = \widetilde{T}_\infty + O\left(\frac{R}{\lambda}\frac{(\log \Lambda_\infty)^3}{\Lambda_\infty^3}\right). \tag{184}$$

These formulas compare the idealized infinite hierarchies; their remainder bounds measure closed-form approximation error within that model. Relating them to a finite algorithm additionally requires the occupancy bounds and frontier corrections above. The proof expands the Lambert representation and then averages a second-order expansion about the mean hierarchy level; third-order remainder bounds control the error and justify the large-m limit uniformly. See Appendix C.4.1.

C.3 Numerical Evaluation of the Deterministic Leading Terms

We now evaluate the idealized hierarchy model under a representative parameter configuration. Throughout this subsection, we set

$$\alpha = 0.05.$$

This choice requires each trajectory in a candidate group to occupy at most 5% of the migration threshold. It is an eligibility test applied in both phases of the proposed policy; the probability of an item-size violation is accounted for explicitly in the stochastic bounds.

Following the threshold convention in (6), we set

$$C_{\mathrm{th}} = 0.95\,\frac{B}{\lambda}.$$

This gives the fixed upper margin $\eta = 0.05$. We evaluate the deterministic expressions under

$$R = 64, \qquad B = 330, \qquad \lambda = \frac{1}{200}, \qquad C_{\mathrm{th}} = 0.95\,\frac{B}{\lambda} = 62700.$$

The no-repacking baseline is

$$\widehat{T}_{\mathrm{no}} = R\frac{H_{B-1}}{\lambda}.$$

The pairwise and finite-m multi-repacking values reported below are obtained by directly evaluating the relaxed infinite hierarchy sums in the preceding subsection, rather than by using the closed-form approximations in Theorem 28.

All reported values and percentage reductions compare the mechanisms within this idealized hierarchy model, which uses continuous release weights and infinitely many levels. Quantitative predictions for a finite system additionally require accounting for integer release counts, the finite hierarchy, and the natural release times of the remaining replicas.

The numerical accuracy of the closed-form approximation is examined in Appendix C.4.1. Among the cases reported in Table 2, the largest observed relative discrepancy is approximately 2.62%. The logarithmic truncation error decreases as $\Lambda = 1 - \log \rho$ increases, while the remainder of

Table 1: Comparison in the idealized hierarchy model, obtained by direct numerical evaluation of the relaxed infinite hierarchy sums.

Method	$\widehat{T}_{\mathcal{M}}$	Reduction from no repacking	Reduction from pairwise
No repacking	81,597.34	0.00%	–
Pairwise repacking	34,485.25	57.74%	–
Multi-repacking, $m = 3$	32,213.70	60.52%	6.59%
Multi-repacking, $m = 4$	30,927.93	62.10%	10.32%
Multi-repacking, $m = 5$	30,229.31	62.95%	12.34%
Multi-repacking, $m = 6$	29,789.36	63.49%	13.62%
Multi-repacking, $m = 8$	29,265.35	64.13%	15.14%
Multi-repacking, $m = 10$	28,963.62	64.50%	16.01%
Multi-repacking, $m = 16$	28,527.40	65.04%	17.28%
Multi-repacking, $m = 32$	28,177.85	65.47%	18.29%
Multi-repacking, $m = 64$	28,007.57	65.68%	18.78%
$m \to \infty$ (relaxed limit)	27,840.15	65.88%	19.27%

the second-order hierarchy expansion also depends on the effective dispersion and higher centered moments of the weighted hierarchy levels.

For a repacking mechanism $\mathcal{M}$, define its relative reduction in tail occupied replica time compared with no repacking by

$$\mathrm{Red}_{\mathrm{no}}(\mathcal{M}) := \frac{\widehat{T}_{\mathrm{no}} - \widehat{T}_{\mathcal{M}}}{\widehat{T}_{\mathrm{no}}} \times 100\%.$$

For multi-repacking with merge ratio m, define its additional reduction relative to pairwise repacking by

$$\mathrm{Red}_{\mathrm{pair}}(m) := \frac{\widehat{T}_{\mathrm{p}} - \widehat{T}_{m}}{\widehat{T}_{\mathrm{p}}} \times 100\%.$$

The calculation of the exact $m \to \infty$ value reported below is provided in the proof of Theorem 28.

Within this idealized model, Table 1 shows that pairwise repacking reduces the deterministic tail occupied replica time by 57.74% relative to no repacking. Multi-repacking provides a further reduction for every reported merge ratio. For example, $m = 3$ reduces the tail occupied replica time by 60.52% relative to no repacking and by 6.59% relative to pairwise repacking. These reductions increase to 63.49% and 13.62%, respectively, when $m = 6$.

Across the reported merge ratios, the model's tail occupied replica time decreases, with gradually diminishing improvements. At the largest finite merge ratio reported in the table, $m = 64$, the continuous-hierarchy relaxation gives reductions of 65.68% relative to no repacking and 18.78% relative to pairwise repacking. The corresponding relaxed limit as $m \to \infty$ yields reductions of 65.88% and 19.27%, respectively.

The $m \to \infty$ value is the limiting benchmark of the same idealized hierarchy model. An implementable finite system requires $m \le R$.

C.4 Proofs of the deterministic approximation

C.4.1 Proof and numerical accuracy of Theorem 28

Proof. Let

$$g(\Lambda) := \Lambda - 1 + \log\Lambda + \frac{\log\Lambda}{\Lambda} + \frac{\log\Lambda(2 - \log\Lambda)}{2\Lambda^2}.$$

Put $h(\Lambda) := \Phi(e^{1-\Lambda})$. To justify the expansion, write $y = h(\Lambda) + 1$ and $y_0 = g(\Lambda) + 1$. The defining equation is $y - \log y = \Lambda$ on $y > 1$. Substituting y_0 and expanding $\log y_0 = \log\Lambda + \log(1 + (y_0 - \Lambda)/\Lambda)$ through its quadratic term gives

$$y_0 - \log y_0 - \Lambda = O\left(\frac{(\log\Lambda)^3}{\Lambda^3}\right).$$

Indeed, $(y_0 - \Lambda)/\Lambda = O(\log \Lambda/\Lambda)$, and the terms of orders Λ^{-1} and Λ^{-2} cancel by the definition of g. Both y and y_0 are at least 2 for large Λ, where the derivative of $y - \log y$ is at least $1/2$. The mean-value theorem gives the same order for $|y - y_0|$. Increasing the constant covers the remaining compact interval starting at 2; taking the supremum over $\Lambda \geq \Lambda_0$ gives, uniformly for $\Lambda_0 \geq 2$,

$$|h(\Lambda) - g(\Lambda)| \leq C \frac{(\log \Lambda_0)^3}{\Lambda_0^3}. \tag{185}$$

A direct calculation gives

$$g''(\Lambda) = -\frac{1}{\Lambda^2} + \frac{2\log \Lambda - 3}{\Lambda^3} + \frac{-3(\log \Lambda)^2 + 11 \log \Lambda - 6}{\Lambda^4}.$$

Differentiating once more also gives

$$\sup_{\Lambda \geq \Lambda_0} |g'''(\Lambda)| \leq C \frac{1 + (\log \Lambda_0)^2}{\Lambda_0^3}, \qquad \Lambda_0 \geq 2. \tag{186}$$

Pairwise repacking. Let J_{p} have the hierarchy distribution

$$\mathbb{P}(J_{\mathrm{p}} = \ell) = \frac{1}{2^{\ell+1}}, \qquad \ell \geq 0.$$

Then

$$\mathbb{E}[J_{\mathrm{p}}] = 1, \qquad \mathrm{Var}(J_{\mathrm{p}}) = 2.$$

Since

$$1 - \log(\rho_\ell) = \Lambda_{\mathrm{p}} + (\ell - 1)\log 2,$$

the hierarchy sum has the exact representation

$$\frac{\lambda}{R}\widehat{T}_{\mathrm{p}} = \mathbb{E}[h(\Lambda_{\mathrm{p}} + (J_{\mathrm{p}} - 1)\log 2)]. \tag{187}$$

Every argument in (187) is at least $\Lambda_{\mathrm{p}} - \log 2 \geq \Lambda_{\mathrm{p}} - 1$. As $\Lambda_{\mathrm{p}} \to \infty$, this lower endpoint is comparable to Λ_{p}. Thus (185) replaces h by g with error $O((\log \Lambda_{\mathrm{p}})^3/\Lambda_{\mathrm{p}}^3)$. Taylor's theorem with integral remainder gives

$$\begin{aligned}&\left|\mathbb{E}g(\Lambda_{\mathrm{p}} + (J_{\mathrm{p}} - 1)\log 2) - g(\Lambda_{\mathrm{p}}) - (\log 2)^2 g''(\Lambda_{\mathrm{p}})\right| \\ &\quad \leq \frac{(\log 2)^3}{6}\mathbb{E}|J_{\mathrm{p}} - 1|^3 \sup_{\Lambda \geq \Lambda_{\mathrm{p}} - 1} |g'''(\Lambda)|.\end{aligned}$$

The linear term vanishes because $\mathbb{E}[J_{\mathrm{p}} - 1] = 0$, and the quadratic coefficient uses $\mathrm{Var}(J_{\mathrm{p}}) = 2$. Since $\mathbb{E}|J_{\mathrm{p}} - 1|^3 < \infty$, equations (185) and (186) prove (180). Substituting g and g'' gives (179).

Multi-repacking. Let J_m have the multi-repacking hierarchy distribution

$$\mathbb{P}(J_m = \ell) = \frac{1}{m}\left(\frac{m-1}{m}\right)^\ell, \qquad \ell \geq 0.$$

Hence,

$$\mathbb{E}[J_m] = m - 1, \qquad \mathrm{Var}(J_m) = m(m-1).$$

Define

$$a_m := \log \frac{m}{m-1}.$$

Since

$$1 - \log(\rho_{\ell,m}) = \Lambda_m + (\ell - (m-1))a_m,$$

we likewise have the exact representation

$$\frac{\lambda}{R}\widehat{T}_m = \mathbb{E}[h(\Lambda_m + (J_m - (m-1))a_m)]. \tag{188}$$

Every argument is at least $\Lambda_m - (m-1)a_m > \Lambda_m - 1$, since $\log(1+x) < x$ for $x > 0$. Thus the lower endpoint is comparable to Λ_m as $\Lambda_m \to \infty$. Taylor's theorem and (186) give

$$\begin{aligned}&\left|\mathbb{E}g(\Lambda_m + (J_m - (m-1))a_m) - g(\Lambda_m) - \frac{m(m-1)a_m^2}{2}g''(\Lambda_m)\right| \\ &\quad \leq \frac{a_m^3}{6}\mathbb{E}|J_m - (m-1)|^3 \sup_{\Lambda \geq \Lambda_m - 1} |g'''(\Lambda)|.\end{aligned}$$

For fixed m, the centered third absolute moment is finite. Combining this inequality with (185) proves (182); substituting g and g'' gives (181).

Large-m limit. Let

$$\rho_\infty := \frac{(1-\alpha)C_{\mathrm{th}}\lambda}{B-1},$$

and assume that $0 < \rho_\infty \le 1$, as satisfied by the parameter configuration considered in Section C.3. Then

$$\Lambda_m = 1 - \log\left[\rho_\infty \left(\frac{m-1}{m}\right)^m\right] \longrightarrow 2 - \log\rho_\infty = \Lambda_\infty.$$

Moreover,

$$\frac{m(m-1)}{2}\left(\log\frac{m}{m-1}\right)^2 \longrightarrow \frac{1}{2}.$$

For the centered level displacement $X_m := a_m(J_m - (m-1))$, $\mathbb{E}[J_m^3] = (m-1)(6m^2-6m+1)$ and $|J_m - (m-1)|^3 \le 4(J_m^3 + (m-1)^3)$. Since $a_m \le 1/(m-1)$, these inequalities imply

$$\sup_{m\ge 3} \mathbb{E}|X_m|^3 < \infty.$$

Furthermore, the exact minimum argument is $1 - \log\rho_{0,m} = \Lambda_\infty - 1 + a_m \ge \Lambda_\infty - 1$. Using this common lower endpoint in both remainder estimates, together with the uniform third-moment bound, makes the Lambert and Taylor remainders uniform in m. Since $\Lambda_\infty - 1$ is comparable to Λ_∞ as $\Lambda_\infty \to \infty$, the same error order holds with Λ_∞ in the denominator. Taking the limit of the center and variance terms gives (183), while the uniform remainders prove (184).

For use in the numerical comparison, we also calculate the exact large-m limit of the relaxed infinite hierarchy sum. Let

$$q_m := \frac{m-1}{m}.$$

Then

$$\frac{\lambda}{R}\widehat{T}_m = \sum_{\ell\ge 0}\frac{1}{m}q_m^\ell \Phi\left(\rho_\infty q_m^{\ell+1}\right).$$

To justify the limit, note that $a_m = -\log q_m$, $(1/m)/a_m \to 1$, and

$$\Phi(\rho_\infty e^{-x}) \le C_{\rho_\infty}[1 + x + \log(1+x)], \qquad x \ge 0.$$

Thus the corresponding step functions are dominated, after multiplication by e^{-x}, by an integrable function. The dominated Riemann-sum theorem therefore gives

$$\lim_{m\to\infty}\frac{\lambda}{R}\widehat{T}_m = \int_0^\infty e^{-x}\Phi\left(\rho_\infty e^{-x}\right)\,\mathrm{d}x = \frac{1}{\rho_\infty}\int_0^{\rho_\infty}\Phi(\rho)\,\mathrm{d}\rho.$$

Let $y = \Phi(\rho)$. From the definition of Φ,

$$\rho = (1+y)e^{-y}, \qquad \mathrm{d}\rho = -ye^{-y}\,\mathrm{d}y.$$

Writing $y_\infty := \Phi(\rho_\infty)$, we obtain

$$\begin{aligned}\frac{1}{\rho_\infty}\int_0^{\rho_\infty}\Phi(\rho)\,\mathrm{d}\rho &= \frac{1}{\rho_\infty}\int_{y_\infty}^\infty y^2 e^{-y}\,\mathrm{d}y \\ &= \frac{e^{-y_\infty}\left(y_\infty^2 + 2y_\infty + 2\right)}{(1+y_\infty)e^{-y_\infty}} \\ &= y_\infty + 1 + \frac{1}{y_\infty+1}.\end{aligned}$$

Therefore,

$$\lim_{m\to\infty}\widehat{T}_m = \frac{R}{\lambda}\left[\Phi(\rho_\infty) + 1 + \frac{1}{\Phi(\rho_\infty)+1}\right].$$

This completes the proof. □

Numerical accuracy. Let $\widetilde{T}_{\mathcal{M}}$ denote the corresponding closed-form approximation in Theorem 28. Define

$$\mathrm{RelErr}(\mathcal{M}) := \frac{\left|\widetilde{T}_{\mathcal{M}} - \widehat{T}_{\mathcal{M}}\right|}{\widehat{T}_{\mathcal{M}}} \times 100\%.$$

Table 2 compares the closed-form approximations with the reference values of the relaxed infinite sums; its relative errors measure approximation accuracy within the idealized hierarchy model. The parameters are

$$\alpha = 0.05, \qquad R = 64, \qquad B = 330, \qquad \lambda = \frac{1}{200}, \qquad C_{\mathrm{th}} = 0.95\,\frac{B}{\lambda}.$$

Table 2: Numerical accuracy of the closed-form approximation to the idealized hierarchy model.

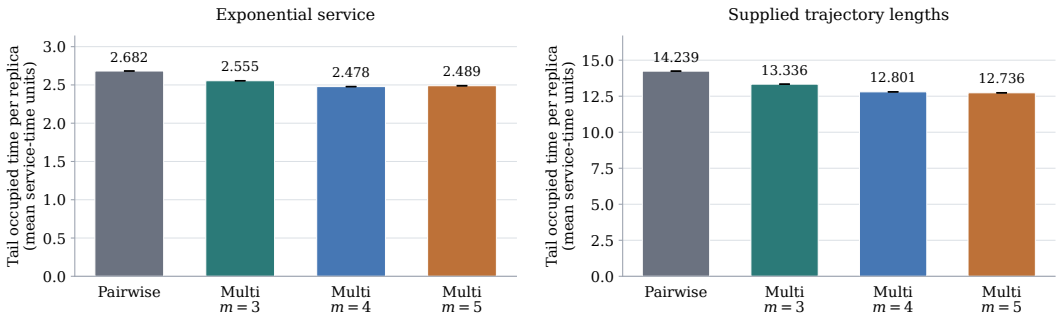
Method	Reference value	Closed-form approximation	Relative error
Pairwise repacking	34,485.25	34,185.48	0.869%
Multi-repacking, $m = 3$	32,213.70	31,796.20	1.296%
Multi-repacking, $m = 4$	30,927.93	30,428.52	1.615%
Multi-repacking, $m = 5$	30,229.31	29,681.84	1.811%
Multi-repacking, $m = 6$	29,789.36	29,210.44	1.943%
Multi-repacking, $m = 8$	29,265.35	28,647.86	2.110%
Multi-repacking, $m = 10$	28,963.62	28,323.42	2.210%
Multi-repacking, $m = 16$	28,527.40	27,853.73	2.361%
Multi-repacking, $m = 32$	28,177.85	27,476.83	2.488%
Multi-repacking, $m = 64$	28,007.57	27,293.03	2.551%
$m \to \infty$ (relaxed limit)	27,840.15	27,112.19	2.615%

All relative errors are computed using the unrounded values. Among the reported cases, the maximum relative discrepancy is approximately 2.62%, attained at the relaxed large-m limit.

D Additional Experimental Results

D.1 Hierarchical policy comparison

The figures below report the 160 runs described in Section 5. All occupied-time costs include natural releases. Replicas start independently and uniformly over $[0, 1]$ in mean-service-time units, and each draw's start vector and service realizations are shared by all policies.



(a) Exponential service: up to 7.59% lower.

(b) Supplied lengths: up to 10.56% lower.

Figure 1: All three merge ratios improve over hierarchical pairwise on both workloads under independent uniform starts. Bars show mean tail occupied replica time per original replica in mean-service-time units, with pointwise 95% intervals over 20 shared draws. Lower is better. All four policies remain within the reserved physical capacity on the supplied-length runs.

D.2 Removing the hierarchy restrictions

Implementation and scope. The unrestricted implementation admits each nonempty anchored replica to one common pool. It removes same-level matching, whole-block membership, the switch certificate, and terminal fallback. A round chooses the minimum-load slot-feasible set of m replicas, with deterministic ties, and applies the same individual-size and aggregate-load tests and the same descending-size cyclic placement into $m-1$ outputs. Every nonempty output immediately returns to the common pool. Slot and cache feasibility remain enforced: each output contains at most B trajectories and has cache load at most C_{th}. This implementation is an empirical extension; the hierarchical proofs do not establish its performance. All policies retain the same trajectory birth times and completion deadlines. Migration has zero simulated time cost.

Laminar-style best-fit baseline. We implement the repacking rule of Laminar (Sheng et al., 2025a) in the same single-version simulator. At a packing check, a replica is eligible if it is anchored and nonempty, has fewer than B live trajectories, and satisfies $C_{\mathrm{used}} < \min(C_{\mathrm{th}}, C_{\mathrm{prev}})$. Sources are visited in ascending observed cache load and assigned to the largest projected feasible destination. A destination may receive multiple sources in one plan; there is no hierarchy or disjoint-pair restriction. We initialize C_{prev} at anchor admission and update it from observations immediately before each plan. If a receiving destination subsequently becomes a source, feasibility uses its entire projected load and count, including earlier incoming trajectories. This explicit bookkeeping convention prevents chained plans from undercounting resources. The shared packing limit is C_{th}; the larger physical capacity is reserved headroom for later cache growth.

Two check schedules. The periodic comparison uses checks spaced by 0.1 mean service times, with a common grid starting at the earliest prescribed start. Both Laminar-style best-fit and unrestricted multi use this grid. Natural releases still occur immediately at completions. This fixed interval is not a calibration of Laminar's wall-clock check interval. To test whether an advantage depends on sampled ramp-down eligibility, a second comparison checks after completion events and removes the $C_{\mathrm{used}} < C_{\mathrm{prev}}$ filter from best-fit. All anchored nonempty representatives are then eligible subject to slot and cache feasibility. Both event-driven policies exhaust feasible operations at each check. Together these comparisons distinguish performance against the sampled Laminar-style rule from performance against a more responsive best-fit control. Neither comparison evaluates the full multi-version Laminar runtime.

Paired data and reporting. We use 20 paired draws for each of the two workloads, with the same $R, B, N, \alpha, C_{\mathrm{th}}$ and physical capacity as in Section 5. All comparisons use independent $\mathrm{Unif}[0,1]$ replica starts, independent of service durations. Within each draw, the start vector and service realizations are shared across policies and check schedules. Four policies on each of two check schedules and two workloads give 320 additional runs. We report every $m = 3, 4, 5$ result; there is no performance-based selection of draws or ratios. Pointwise 95% Student-t intervals for improvements use the 20 within-draw absolute cost differences. Percentage reductions are ratios of sample means. They are not confidence intervals for ratios, and no multiplicity correction is applied. The modified event engine exactly reproduces the original four policies on draw zero of each workload; deterministic fixtures check slot feasibility, regrouping across histories, chained plans, and tied events.

Occupied-time comparison. Under periodic checks, Laminar-style best-fit has mean tail occupied replica time per original replica 2.6819 for exponential service and 14.2453 for the supplied-length workload. Unrestricted $m = 3, 4, 5$ reduce these means by 4.89–10.55% and 7.11–13.76%, respectively. Under completion-event checks, best-fit has respective means 2.6758 and 14.2365, with reductions of 6.60–12.28% and 7.64–14.27%. All twelve paired improvement intervals are strictly above zero under these asynchronous starts. Among the tested ratios, $m = 5$ has the lowest mean in each comparison. Figure 2 reports policy means, and Table 3 gives relative reductions and paired improvement intervals.

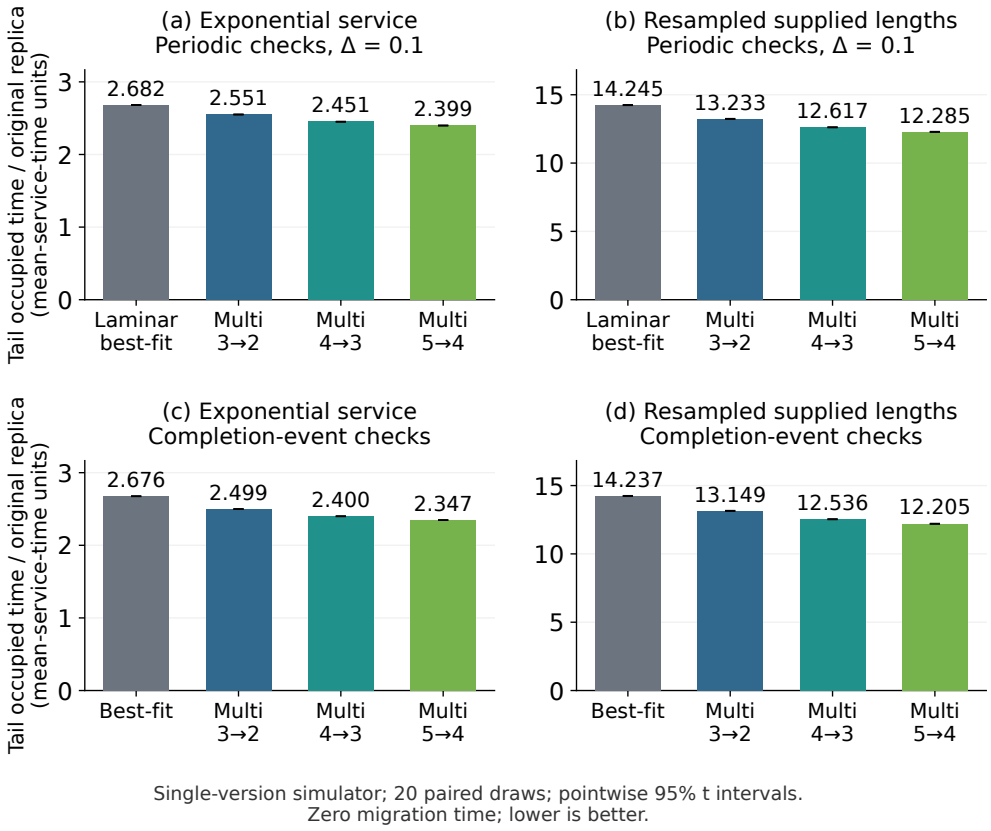


Figure 2: Unrestricted multi-repacking under periodic (top) and completion-event (bottom) checks. The periodic baseline retains Laminar-style sampled ramp-down eligibility; the event-driven control removes this filter. All policies share the sampled startup times and cache and slot limits. Bars show mean tail occupied replica time per original replica and pointwise 95% Student-t intervals over 20 paired draws; migration time is zero.

Workload	Checks	m	Reduction	Paired improvement (95% CI)
Exponential	Periodic	3	4.89%	0.1312 [0.1305, 0.1318]
		4	8.59%	0.2304 [0.2298, 0.2310]
		5	10.55%	0.2830 [0.2824, 0.2836]
Exponential	Event	3	6.60%	0.1766 [0.1763, 0.1769]
		4	10.31%	0.2758 [0.2754, 0.2761]
		5	12.28%	0.3285 [0.3280, 0.3289]
Supplied lengths	Periodic	3	7.11%	1.0126 [1.0074, 1.0177]
		4	11.43%	1.6279 [1.6232, 1.6325]
		5	13.76%	1.9599 [1.9555, 1.9644]
Supplied lengths	Event	3	7.64%	1.0876 [1.0808, 1.0943]
		4	11.94%	1.7005 [1.6935, 1.7075]
		5	14.27%	2.0318 [2.0245, 2.0390]

Table 3: Paired improvements for unrestricted multi (positive favors multi), in mean-service-time units per original replica. Percentage reductions use ratios of sample means; 95% intervals use within-draw absolute differences and are pointwise, without multiplicity correction.