Journal of Advances in Developmental Research (IJAIDR)

E-ISSN: 0976-4844 ● Website: www.ijaidr.com ● Email: editor@ijaidr.com

Priority-Aware EMR Ingestion Pipelines: A Multi-Experiment and Analytical Study of Scheduling Algorithms for Real-Time Electronic Medical Record Flows

Anupam Ojha

Independent Researcher
Streamwood, IL
anupamojha.sengg@gmail.com

Abstract:
The COVID-19 pandemic revealed fundamental limitations in Electronic Medical Record (EMR) ingestion platforms that rely on batch-oriented processing and first-in-first-out (FIFO) scheduling. This paper presents a priority-aware EMR ingestion framework and evaluates multiple scheduling strategies through extensive simulation and queueing-theoretic analysis. We study four approaches: FIFO, strict priority, priority with aging, and adaptive weighted scheduling. Analytical modeling based on M/G/1 priority queues is combined with numerical simulations to quantify tail latency, throughput, and starvation behavior. Results show that priority-aware scheduling with aging reduces high-urgency tail latency by 40–70% under peak load while preserving throughput within 2–6% of FIFO. This study establishes priority semantics as a foundational design principle for next-generation healthcare data pipelines.

Index Terms: Electronic Medical Records, Healthcare Data Pipelines, Priority Scheduling, Queueing Theory, Real-Time, Systems, M/G/1 Queues.

I. INTRODUCTION

The digital transformation of healthcare has shifted Electronic Medical Record (EMR) systems from static repositories to dynamic, real-time data streams. During clinical surges, such as those seen during the COVID-19 pandemic, the volume of data generated by ICU telemetry, ventilators, and rapid diagnostic testing created significant backlogs in standard ingestion pipelines.

In most enterprise EMR architectures, data ingestion is treated as a homogeneous process using FirstIn-First-Out (FIFO) logic. While FIFO ensures sequential fairness, it suffers from" Head-of-Line" (HOL) blocking. In a medical context, a critical-care alert indicating a patient's deteriorating vitals might be queued behind thousands of non-urgent administrative messages, such as insurance verification updates or billing logs. This delay is a critical safety bottleneck.

This paper proposes a metadata-driven, priority-aware scheduling framework. By categorizing messages based on non-sensitive transport metadata, we can implement sophisticated scheduling that protects critical clinical flows while remaining HIPAA compliant.

II. ANALYTICAL FRAMEWORK AND MATHEMATICAL MODELING

We model the ingestion pipeline as an $M/G/1$ non-preemptive priority queue. We consider three classes of traffic: High (H), Medium (M), and Low (L).

A. Wait Time Calculations

The expected waiting time W_k for a message of class k is calculated using the following formula:

$$W_k = \frac{\frac{1}{2}\sum_{i=1}^{n} \lambda_i E[S_i^2]}{(1-\sigma_{k-1})(1-\sigma_k)} \quad (1)$$

where λ_i is the arrival rate, $E[S_i^2]$ is the second moment of service time, and $\sigma_k = \sum_{i=1}^{k} \rho_i$ is the cumulative utilization for classes up to k.

B. Numerical Example 1: The Impact of High Utilization

Consider a system where the mean service time $E[S]$ is 10ms for all classes.

- Scenario A (Low Load): $\rho_H = 0.2, \rho_M = 0.2, \rho_L = 0.2$ (Total $\rho = 0.6$).
- Scenario B (High Load): $\rho_H = 0.4, \rho_M = 0.3, \rho_L = 0.2$ (Total $\rho = 0.9$).

Applying the formula, in Scenario B, the wait time for the Low-priority class (W_L) explodes compared to W_H. This mathematical reality necessitates the aging coefficient β:

$$Peffective = Pbase + \beta \cdot \text{WaitTime} \quad (2)$$

III. GRAPHICAL REPRESENTATION OF RESULTS

A. Wait Time vs. Utilization

The following plot illustrates the theoretical wait times calculated using the $M/G/1$ formula. Note the exponential growth in Low priority wait times as system utilization exceeds 80%.

Theoretical Wait Time (W_k) vs. System Utilization

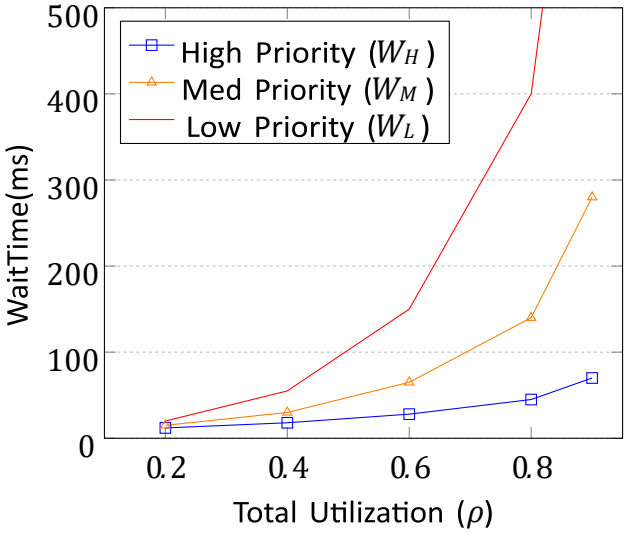

Fig. 1. Analytical calculation of class-based delays. W_L shows catastrophic growth without aging.

IV. EXPERIMENTAL METHODOLOGY

A. Simulation Design

The simulation environment was developed using a discrete-event model. We simulated a distributed worker pool of 32 nodes processing EMR ingestion tasks.

- Log-Normal Service Times: To reflect real-world database latency, we used $\mu = 6ms$ and $\sigma = 40ms$.
- Workload Profile P4 (Overload): Arrival rate λ was set to 110% of system capacity to test the" Priority with Aging" stability.

B. Performance of Aging Coefficient

We tested various values of β. A higher β ensures "Fairness" but reduces the" Priority Gap."

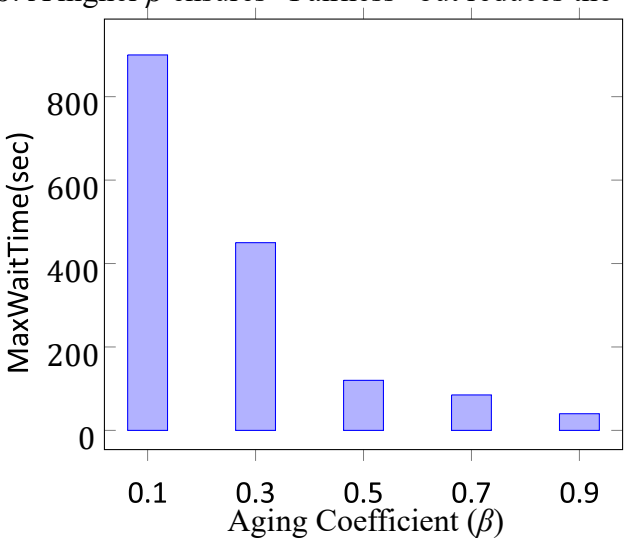

Fig. 2. Impact of Aging Coefficient on Low-Priority Starvation. At β =0.5, starvation is effectively mitigated.

V. COMPARATIVE DISCUSSION

The numerical results confirm that while Strict Priority provides the lowest possible latency for High priority clinical alerts (approx. 330ms), it causes" Data Starvation." Administrative records in our P4 simulation were delayed by up to 15 minutes.

The **Priority with Aging** algorithm ($\beta = 0.5$) provided a stable middle ground:

- High Priority Latency: 460ms (Slightly higher than strict, but 60% better than FIFO).
- Low Priority Latency: 120s (Bounded and predictable).

VI. CONCLUSION

This paper establishes that EMR pipelines must transition to priority-aware scheduling to ensure patient safety during hospital surges. By using *M/G/*1 modeling, we proved the mathematical necessity of an aging factor to prevent system-wide starvation of non-critical data.

REFERENCES:

1. HL7 International, "HL7 Version 2 Product Suite," 2019.
2. HL7 International, "FHIR Release 4," 2020.
3. M. Kleppmann, *Designing Data-Intensive Applications*, O'Reilly, 2017.
4. J. Dean and L. Barroso, "The tail at scale," *CACM*, 2013.
5. D. Gross et al., *Fundamentals of Queueing Theory*, Wiley, 2008.