Information Bulletin Strategy in Impatient Queueing

Anthony Kiggundu*, Bin Han† and Hans D. Schotten*†
*German Research Center for Artificial Intelligence (DFKI), Germany
†RPTU University of Kaiserslautern-Landau, Germany

Abstract

We study an information bulletin strategy for decentralized decision-making in 6G multi-tenant systems. Queues periodically broadcast descriptor information as two Markov models (queue length inter-change dynamics and service-time distributions) that tenants use to decide whether to renege or jockey. The queues observe tenant responses and adapt their processing rates via a learning loop, with the goal of minimizing aggregate delay and impatience while respecting service constraints.

***Index Terms*—6G, Queuing theory, Jockeying and Reneging, Behavioral modeling, Performance evaluation**

I. System Setup

Our study assumes two parallel queues with incoming jobs that obey a Poisson distribution at rate λ_i and get processed following a first come first serve service discipline. The service rates μ_i, μ_j are heterogeneous as characterized by $\mu_i = \frac{\lambda+\delta\lambda}{2}, \mu_j = \frac{\lambda-\delta\lambda}{2}$. Two information models are broadcast to all queued requests, i.e. the Markov model of how frequently the queue length changes and the Markov model of the service times. Similar attempts to use this queue descriptor information have been explored [1], [2]. and the other an arrival or departure event can occur to end the queue in state $\lambda_i + \mu_i$. in steady state given these events is define $R_i = \sum_{n=0}^{\infty} \pi_{i,n} \left(\lambda_i + \mu_i \cdot 1_{n\geq 1}\right)$ (*where* $1_{n\geq 1}$ *is an indicator function for measurability around* $n \geq 1$ *since a departure event only occurs when* $n \geq 1$). Given $\rho_i \mu_i = \lambda_i$ implies $R_i = \lambda_i + \lambda_i = 2\lambda_i$ as the number of changes introduced by these events and the expected time between successive changes in the sizes of the queue is $T_i^{\mathrm{ICD}} = \frac{1}{2\lambda_i}$. Additionally, in steady state, the size of any of these $M/M/1$ queues i or j follows a birth-death continuous Markov chain with K states and rates $\{\mu_i\}_{i=1}^K$, $\{\mu_j\}_{j=1}^K$ respectively. We let X and Y denote these two stationary service rate distributions over $\{\mu_1 < \cdots < \mu_K\}$ with steady state probabilities $\pi_i^X = Pr\{\mu = \mu_i\}$ and $\pi_j^Y = Pr\{\mu = \mu_j\}$. Then the effective service rates as averaged over the entire Markov chain are defined by $\bar{\mu}_X = \sum_{i=1}^K \pi_i^X \mu_i$ and $\bar{\mu}_Y = \sum_{j=1}^K \pi_j^Y \mu_j$. To abandon the queue, the local processing T_{local} (deterministic and no waiting time involved) must be less than the estimated remaining waiting time at position ℓ. This time is defined for as $\mathbb{E}[W_i|\ell] = \sum_{i=0}^{\ell} \frac{1}{\mu_i} = \frac{\ell}{\mu_i} = \sum_{i=0}^{\infty} \pi_\ell \frac{\ell}{\mu_i}$ We summarize the impatience conditionalities in algorithm 1.

Algorithm 1: Impatience Algorithm

```
Input: arrival rate(s) {λ_i}, service rates {μ_i, μ_j}, dispatch interval r etc
Output: mean wait, reneging rate, jockeying rate, learned policy if any
Initialize: two queues Q_1, Q_2, ICD-model, FSD-model ;
while t < SimulationTime do
    if t is a dispatch epoch (every r seconds) then
        for each queue Q_i do
            Broadcast ICD-model_i and FSD-model_i to all queued
              requests;
        end
    end
    for each queued request u in each queue Q_i do
        if FSD-model_j (i.e., F_j(x) ≤ F_i(x) ∀x ) then
            if True then
                Move request u from Q_i to tail of Q_j ;
                continue to next request;
            end
        end
        Compute expected remaining waiting time at position ℓ: if
          Ŵ_rem(ℓ) > T_local then
            Request u reneges; ;
        end
    end
end
return collected performance metrics and (if predictive) learned policy
  {μ_i(t)};
```

II. Preliminary Results

III. Conclusion and Future Work

To achieve global optimality can be generalized into an optimization problems where we can tune μ_i and μ_j to minimize weighted sum of average waiting time and the overall impatience.

References

[1] A. Kiggundu, B. Han, D. Krummacker, and H. D. Schotten, "Resource allocation in mobile networks: A decision model of jockeying in queues," in *2024 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*, 2024, pp. 311–316.

[2] S. R. Mahabhashyam and N. Gautam, "On queues with markov modulated service rates," *Queueing Systems*, vol. 51, pp. 89–113, 2005.

[4]This work is supported in fully by the German Federal Ministry of Research, Technology and Space (BMFTR) within the Open6GHub project under grant numbers *16KISK003K and 16KISK004.*

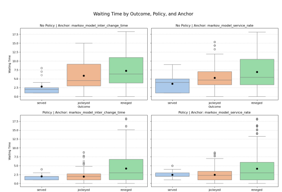

Fig. 1: The box plot showing the median waiting time when a policy is embedded versus otherwise for both Markov information models. The waiting time and spread here are high when no policy is embedded for both impatience behavior.