

Chronicles of Jockeying in Queueing Systems

ANTHONY KIGGUNDU, German Research Center for Artificial Intelligence, Germany

BIN HAN*, RPTU Kaiserslautern-Landau, Germany

DENNIS KRUMMACKER, German Research Center for Artificial Intelligence, Germany

HANS D. SCHOTTEN, RPTU Kaiserslautern-Landau, German Research Center for Artificial Intelligence, Germany

Emerging trends in communication systems, such as network softwarization, functional disaggregation, and multi-access edge computing (MEC), are reshaping both the infrastructural landscape and the application ecosystem. These transformations introduce new challenges for packet transmission, task offloading, and resource allocation under stringent service-level requirements. A key factor in this context is queue impatience, where waiting entities alter their behavior in response to delay. While balking and reneging have been widely studied, this survey focuses on the less explored but operationally significant phenomenon of jockeying, i.e. the switching of jobs or users between queues. Although a substantial body of literature models jockeying behavior, the diversity of approaches raises questions about their practical applicability in dynamic, distributed environments such as 5G and Beyond. This chronicle reviews and classifies these studies with respect to their methodologies, modeling assumptions, and use cases, with particular emphasis on communication systems and MEC scenarios. We argue that forthcoming architectural transformations in next-generation networks will render many existing jockeying models inapplicable. By highlighting emerging paradigms such as MEC, network slicing, and network function virtualization, we identify open challenges, including state dissemination, migration cost, and stability, that undermine classical assumptions. We further outline design principles and research directions, emphasizing hybrid architectures and decentralized decision making as foundations for re-conceptualizing impatience in next-generation communication systems.

CCS Concepts: • Computing Methodologies → Parallel computing methodologies; Queueing Theory; Jockeying.

Additional Key Words and Phrases: Queueing theory, impatience, jockeying, MEC, Markov decision process.

ACM Reference Format:

Anthony Kiggundu, Bin Han, Dennis Krummacker, and Hans D. Schotten. 2023. Chronicles of Jockeying in Queueing Systems. *ACM Comput. Surv.* XXX, XXX, Article XXX (January 2023), 35 pages. <https://doi.org/10.1145/nnnnnnnnnnnnnnnn>

1 INTRODUCTION

Recent advances in information and communications technology (ICT) have spawned a class of applications that are sensitive to latency and to reliability. Examples include remote control and industrial automation [7, 159], autonomous driving [45, 154, 164], and immersive XR services. These applications place unprecedented demands

*B. Han is the corresponding author

Authors' addresses: Anthony Kiggundu, German Research Center for Artificial Intelligence, Trippstadter Strasse 122, 67663, Kaiserslautern, Germany, anthony.kiggundu@dfki.de; Bin Han, RPTU Kaiserslautern-Landau, Gottlieb-Daimler-Straße 47, 67663, Kaiserslautern, Germany, bin.han@rptu.de; Dennis Krummacker, German Research Center for Artificial Intelligence, Kaiserslautern, Germany, dennis.krummacker@dfki.de; Hans D. Schotten, RPTU Kaiserslautern-Landau, and German Research Center for Artificial Intelligence, Kaiserslautern, Germany, hans_dieter.schotten@dfki.de; schotten@eit.uni-kl.de.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Association for Computing Machinery.

0360-0300/2023/1-ARTXXX \$15.00

<https://doi.org/10.1145/nnnnnnnnnnnnnnnn>

on packet transmission, task offloading, and real time resource allocation [20]. These demands expose limits of classical queueing abstractions and motivate a fresh look at *impatient* queueing behavior, in particular *jockeying*: the act of switching jobs or users between service queues to obtain faster service.

From the behavioral perspective, studies have categorized impatient consumers as those that observe queue status and refrain from joining, a manner termed as **balking** [10, 122]. Then there are those that join a queue and abandon it when the accumulated delay is more than expected (**renege**) [52, 142]. Less studied but operationally important is **jockeying**, where a customer who has already joined a queue relocates to an alternative queue, often driven by perceived differences in waiting time, price or service capability [82]. Most analytical and experimental work on jockeying attaches costs, priorities or class labels to buffers to capture system heterogeneity and tenant preferences [49]. This heterogeneity is the principal rationale for preferring one buffer over another, but it is not the sole trigger for switching.

1.1 Triggers and modeling variants

Generally, the reasons that influence impatience among consumers as covered in most literature [110, 143] can be irrational with no consideration whatsoever for prevailing buffer conditions, or rational [21]. The following are the most widely studied triggers for impatience in queues:

Threshold-based switching: A common rule is a queue length or waiting time threshold; That is, when the difference in queue lengths exceeds a preset margin, an agent may move from the longer to the shorter queue [3, 51, 170] or vice versa in some models [141]. In some findings, jockeying is only permitted when the alternative queue is empty [124, 151]. And in more versatile setups, entities can rationally choose to switch from any position within the queue to the end of an alternative queue [90]. Or more aggressively intermingle randomly in what is referred to as pre-emptive jockeying [59, 65].

Cost and expected delay criteria: In heterogeneous systems like mobile networks or cloud MEC platforms, simple queue length thresholds are often inadequate. Instead, tenants either premise this jockeying behavior on combined constraints on the queue length threshold and the expected waiting time [38, 169]. Other findings compare the expected remaining service time [33, 49] or the distance over which the workload must be migrated [23, 32]. Or setups that consider the expected delay [124]. Essentially, the tenant's decision to move around workload factors in information about changes in the system characteristics [60, 76]. This information in other findings is defined in terms of costs, subscription profiles or network traffic classifications [70] to influence the rationality of the jockeying tenants [141].

1.2 Queue typology and baseline assumptions

Technically, queues are composed of multi-server or single server lines. The canonical $A/S/c$ classification (where A, S, c denoted the arrival rate, service interval and the number of **servers** available for processing respectively) introduced in the queueing literature [80] yields the familiar $M/M/c$, $G/G/c$, and $E/PH/c$ families [88]. Table 1 is a summary of commonly adopted queue notations symbolizing the statistical properties of arrivals, departures and service disciplines. Several standard assumptions recur across jockeying studies in addition to the above nomenclature:

- **Arrival and service processes.** Although studies that differentiate between the properties of the statistical distributions of both arrivals and departures in buffers exist [25, 39], most findings generalize the periodicity of these two activities to obey a Poisson distribution [125, 157]. Others presume this frequency as batch arrivals [24, 165] for example in systems with limited capacity [147], and the performance effect that the periodicity of these two events has on the impatiently queueing consumer has been investigated in

Table 1. Common queueing notation $A/B/X/Y/Z$, originated from [80]

Characteristic	Symbol	Explanation and remarks
Inter-arrival interval (A)	$M^{[S]}$	Exponential with random batch size S , $S \equiv 1$ by default
	D	Deterministic
	E_k	Erlang type k ($k = 1, 2, \dots$)
	H_k	Mixture of k exponentials
	PH	Phase type
Inter-service interval (B)	G	General
Number of servers (X)	1, 2, ..., ∞	$Y = \infty$ by default
Maximal queue capacity (Y)		
Queue discipline (Z)	FCFS	First come, first served (default) [19, 117]
	LCFS	Last come, first served [75, 99]
	RSS	Random selection for service [87]
	PR	Priority [18, 146, 160]
	GD	General discipline

[54, 59, 89]. The dynamics introduced by these events in queues are mostly generalized to form continuous or discrete stochastic chains [148].

- **Service discipline.** First-Come-First-Served is the default service discipline in most work, although according to Table 1 other disciplines have been documented [88]. However, practitioners argue that mission critical operations, such as rescue (where tasks urgently need to be prioritized by jockeying) render the FCFS approach less feasible given the strict latency requirements in such operations [91, 93, 146].
- **Homogeneous setups.** Another wild assumption in most models is that the setups are homogeneous in nature, further limiting the applicability of some modeling techniques in next generation communication systems like 5G and Beyond.
- **Join the Shorter Queue (JSQ).** It has also been argued that joining the shorter queue might not be optimal given the existential differences in queue capacities and workload [158]. Limiting jockeying to occur from the shorter buffer to the longer one raises concerns about such jockeying thresholds as the optimal trigger for switching buffers in heterogeneous setups.
- **Queue length difference as a jockeying threshold.** The use of a preset difference in the size of the buffers (jockeying threshold) as the criterion for the task migration is not practical in heterogeneous systems.
- **Information availability.** A common modeling premise is that arriving agents have access to some form of queue status information [84]. This information guides the decision to join a queue and the common strategy is to join the shorter one [5, 129], or switch to an alternative queue [14, 117]. Technical mechanisms for providing tenants with access to this buffer status information have inspired proposals for broadcasting or subscription to this information in dedicated or shared communication channels [55].

1.3 Why classical assumptions may fail in MEC or 5G

Emerging architectural trends like network softwarization, functional splits in the RAN (Radio, Distributed, Centralized Units and Control or User Planes), network slicing and multi-access edge computing (MEC) are fundamentally altering the latency, control plane and orchestration semantics that underlie modern communication systems [73]. These transformations introduce tight service-level constraints, multi-domain heterogeneity and

new costs (e.g., state transfer latency, orchestration handshakes, and control plane signaling) that violate key assumptions in many classical impatient queueing models.

In particular, common primitives such as Markov modulated queues, threshold-based jockeying rules, Erlang/C waiting-time approximations, implicitly assume fast and reliable dissemination of queue descriptors, negligible migration cost, stationarity and trusted status reports. Under Open Radio Access Network (O-RAN) proposals for disaggregation and multi-vendor slicing, these assumptions break down. This is because information is partial and delayed, migration incurs nontrivial transfer and reconfiguration overhead, and decision policies face incentive and security constraints across administrative domains. Therefore, while the modeling assumptions highlighted in Section 1.2 are useful, the following aspects of modern MEC and 5G systems undermine their validity:

- (1) **Heterogeneity and multi-vendor slices.** Slices assemble diverse RAN, core and MEC components with distinct performance and cost profiles [13, 34, 118]. A simple queue length comparison does not capture differences in processing power, transport latency, or the monetary costs that tenants incur when switching between slices [121, 136, 140, 156].
- (2) **Non-zero migration cost.** Stateful migration of tasks or sessions incurs transfer time proportional to state size and network path characteristics[44, 106]. Therefore, ignoring transfer cost biases models toward excessive switching.
- (3) **Information latency and overhead.** Timely and authenticated dissemination of per-slice descriptors (waiting time, load, price) consumes control plane resources[60]. Yet stale or noisy descriptor information could yield poor decisions and may increase churn [56, 67, 138, 152].
- (4) **Stability and security risks.** Low queue length based jockeying thresholds can induce a "ping-pong" behavior or oscillations that could amplify the load [46, 60]. Furthermore, unauthenticated status reports can be exploited for selfish or adversarial gain, necessitating trust mechanisms [81, 135].

1.4 Implications and objectives of this survey

Proposed Sixth Generation (6G) architectural modifications by the Third Generation Partnership Project (3GPP) fundamentally challenge the assumptions underpinning classical jockeying models. Partial and delayed visibility, state transfer costs, heterogeneous slice performance, and cross-domain trust constraints jointly undermine the tractability of simple threshold or Markovian-based decision rules. In particular, *heterogeneity and multi-vendor slicing* introduce non-uniform queue capacities, service disciplines, and admission rules, rendering traditional homogeneous formulations inapplicable. At the same time, *information latency and dissemination overhead* become critical in MEC and Software Defined Network (SDN)/ Network Functions Virtualization (NFV) deployments, where stale or excessive signaling can negate the benefits of jockeying decisions. Moreover, *stability and security risks*, including oscillatory switching (ping-pong effects) and adversarial or corrupted state reports, threaten SLA compliance and trust in decentralized decision making. Taken together, these transformations render many existing jockeying models infeasible, and they highlight the pressing need for next-generation approaches that explicitly incorporate slice heterogeneity, communication-constrained information flows, and robustness guarantees under adversarial or unstable dynamics.

Motivated by these shifts, this survey consolidates the state of the art in jockeying models for communication systems, identifies the limitations of existing techniques under emerging 5G/6G architectural paradigms, and outlines the critical directions where future modeling and system design efforts must focus. Hence, the contributions of this chronicle can be summarized as follows:

- (1) **Taxonomy and synthesis.** This chronicle catalogs existing jockeying models and control primitives, highlighting their assumptions, analytical tools, and applicability. It provides a retrospective survey of techniques for modeling this impatience behavior, for which we know no equivalent compilation has been published.

- (2) **Quantitative gap analysis:** We evaluate the *practicality* of jockeying under contemporary architectural constraints. That is, we assess the limits of classical models in 5G/6G contexts, focusing on heterogeneity, signaling constraints, and security stability challenges.
- (3) **Integration with emerging architectures:** We analyze how Multi-Access Edge Computing (MEC), SDN/NFV, and network slicing redefines modeling jockeying behavior by introducing heterogeneity, signaling delays, and cross-domain trust constraints. This unravels where classical homogeneous and fully observable impatience modeling no longer applies.
- (4) **Design criteria and Future directions:** From the taxonomy and gap analysis, we outline design principles and research directions for developing robust, communication-aware, and jockeying models suitable for next generation distributed environments. These preeminent challenges include but are not limited to migration latency and bandwidth cost, dissemination delay and staleness, multi-slice heterogeneity, robustness to stale or noisy information, prototype validation on MEC or Non Terrestrial Networks (NTN) testbeds etc.

In Section II of this chronicle, we provide a concise overview of the principal techniques used to model jockeying in queues. Following that, we classify the literature by methodological family and examine individual studies in each class, emphasizing their models, assumptions, and numerical or empirical results. We intentionally avoid lengthy formal proofs; where analytical results are essential to the argument, we summarize the key theorems or lemmas and state their implications for jockeying policies. In the discussion section we then assess how recent architectural proposals (notably those advocated by the O-RAN community) affect the practicality of existing jockeying models and policies. Motivated by that assessment, we identify concrete open problems and promising research directions. Particularly, those that address dissemination overhead, task migration costs, performance stability, security, and the deployment challenges of impatience aware mechanisms in MEC and slice-based deployments. **The following sections systematically review existing jockeying techniques before analyzing their applicability to next-generation networks.**

2 TECHNIQUES FOR MODELING JOCKEYING IN QUEUES

Different approaches have been studied to characterize for equilibrium conditions and optimize queue descriptors for performance enhancement. Table 2 provides a comparative summary about the individual models grouped based on the similarities in techniques while differentiated along performance and descriptor evaluations.

2.1 Stochastic Modeling

Most literature models jockeying as a Markov decision problem: Markov Decision Processes (MDPs) (and their generalizations Partially Observable Markov Decision Processes (POMDPs) or Decentralized Partially Observable Markov Decision Processes (Dec-POMDPs)) formalize sequential decision-making under uncertainty and are natural for planning agent actions in dynamic environments [2, 71, 112, 114, 120, 145]. However, these approaches suffer well-known practical limitations: the exploration-exploitation tradeoff can trap learning in local optima, state-space growth renders exact policy search computationally intractable for realistic multi-agent settings, and partial observability further complicates inference and control [48, 134]. In short, despite strong theoretical grounding, MDP-based methods often struggle to scale and to provide timely, high-quality decisions in large, heterogeneous MEC/NTN deployments.

Game-theoretic formulations (Nash equilibrium and variants) offer an alternative by characterizing selfish participants' strategies and equilibrium outcomes [60, 61, 63, 94, 113]. Yet they too rely on demanding assumptions like players' knowing the strategies of others players and can be slow to converge[22, 28]. These techniques are also sensitive to initialization, or susceptible to saddle-point dynamics in sequential or repeated interactions[66, 74, 167]. Consequently, while both MDP/POMDP and Nash-frameworks yield important insights, their computational and information assumptions limit direct applicability in large, decentralized, latency-sensitive networked systems.

Table 2. Different approaches used for analyzing the jockeying behavior of customers in queueing systems

	Queue model	Jockeying threshold	Other metrics
Stochastic jockeying models			
A Solution for Queues with Instantaneous Jockeying and Other Customer Selection Rules	M/M/C	2	Distance traversed by jockeys
On Jockeying in Queues	M/M/2	1	mean waiting times, impact of jockeying on service rate
Two queues in parallel	M/M/2	1	mean queue sizes
The shortest queue model with jockeying	M/M/C	1	mean queue sizes, mean waiting times
General solutions of the jockeying problem	M/M/3 +	≥ 2	effect of arrival rate on jockeying rate, system utilization
M/M/ ∞			Workload distribution, cost efficiency
Dynamic routing and jockeying controls in a two-station queuing system	M/M/2	≥ 1	
Optimal policy for controlling the two-server queueing systems with jockeying	M/M/2	Hedge point	Cost optimization
Dynamic Load Balancing in Parallel Queueing Systems	M/M/2 +	$\geq N$	frequency of monitoring queues
Strategic Dynamic Jockeying Between Two Parallel Queues	M/M/C		
A new look on the shortest queue system with jockeying	M/M/2		queue lengths limits for jockeying
A Queuing System with Two Parallel Lines, Cost-Conscious Customers, and Jockeying	M/M/C	$\geq N$	Count of jumps before being served
Analysis of job transfer policies in systems with unreliable servers	M/M/2	≥ 1	Costs optimization
Stability analysis of some networks with interacting servers			
Two M/M/1 Queues with Transfers of Customers	2xM/M/1	$\geq N$	Workload expiry threshold
		1	transfer rates, time to migrate workload, magnitude of bulk transfer
Tail asymptotics of two parallel queues with transfers	M/M/2	$\geq N$	
Energy-efficient heuristics for job assignment in server farms	M/M/2		Energy Efficiency
Analytic Models			
Matrix-geometric analysis of the shortest queue problem with threshold jockeying	M/M/C	$\geq N$	mean queue size, mean waiting time
Analysis of two queues in parallel with jockeying and restricted capacities	M/M/2 +	iff $N1=0$ or $N2=0$	traffic intensity
2xM/M/1			
A matrix-geometric solution of the jockeying problem	M/M/C	≥ 2	mean queue sizes, mean waiting time
The shorter queue problem: A numerical study using the matrix-geometric solution	M/M/2	$10 < n > \infty$	mean queue length, mean waiting time
Queuing Analysis of a Jockeying	M/M/C	≥ 1	waiting time, traffic intensity
Transient analysis of two queues in parallel with jockeying	M/M/2	iff $N1=0$ or $N2=0$	traffic intensity
Analysis of the asymmetric shortest queue problem with threshold jockeying	M/M/2	$\geq N$	mean queue length
Behavioral models			
Dynamic load balancing in distributed systems (regeneration-theory)	M/M/2	Delay	mean task completion time
Bi-objective optimization for multi-layer location-allocation with jockeying	M/M/C	Fitness function	Idle time, transfer overhead
Equilibrium strategies and value of information in two-line systems	M/M/2	$3 < n > \infty$	queue status information cost
Impatient queuing for intelligent task offloading in MEC	G/G/1/ ∞	1	Costs, response time and rewards
Resource Allocation in Mobile Networks: A decision model of jockeying in queues	M/M/C	Expected waiting time	jockeying frequency, waiting time
Information Bulletin Strategy in Impatient Queueing	M/M/2	Markov information models	impatience rates, waiting time, service optimality

Analogous to continuous fluid flow, Fluid models are widely used for infinite-capacity queueing approximations and for studying bursty traffic regimes [37, 44, 155]. They have been applied to evaluate loss rates, maximum buffer sizes, busy periods and asymptotic/equilibrium behavior, sometimes using matrix-geometric or diffusion (Brownian) variants [8, 79]. Extensions incorporate feedback of queue-length information so transition rates depend on buffer state [133, 163]. However, fluid descriptions are inherently nonstationary for heterogeneous traffic, produce nonlinear transient equations that are hard to solve, and demand substantial computation for sensitivity analysis, optimization and validation [96].

2.2 Analytic Modeling

Matrix-geometric methods provide tractable steady-state solutions for Quasi-Birth-Death and other infinite Markov chains by exploiting the generator's repeating block structure [109, 115]. The idea is to essentially identify the irregular (initial) and regular (repeating) portions of the representative generator matrices that encode the various states a system (re)visits [62, 77]. Under mean-drift conditions one decomposes the generator into initial and repeating blocks and computes the rate (R) matrix whose spectral structure yields equilibrium probabilities [78]. While powerful, there are concerns about using cyclic or logarithmic reduction techniques to solve the rate matrix being computationally complex. Also, the solution for this rate matrix depends on the partitioning scheme between the state sub-levels to compose for initial and boundary states, making the unique solution more intractable.

2.3 Behavioral Modeling

A growing alternative is decentralized, data-driven behavioral modeling that treats impatience and jockeying as local decision problems rather than centrally controlled processes [55, 82]. Artificial neural networks and related connectionist methods can capture high-dimensional, nonlinear relations and predict queue descriptors where classical models fail [11, 85]. Their drawbacks are well known: opacity (“black box”), training costs, over/under-fitting risks, hyperparameter tuning, convergence/stopping issues and additional compute burden at inference time [16, 104, 126, 162]

3 STOCHASTIC MODELS

3.1 Statistical Models

Findings that investigate the behaviour of impatient consumers when queueing up for resources can be traced back to a paper by [51]. There simple setup was composed of two service lines, one “near” and the other “far”. Whether “near” or “far” queue was defined by the queue sizes $X(t), X'(t)$ at the time t of an admission respectively. Such that, the inequality $X(t) \leq X'(t)$ meant new customers preferred to join the shorter buffer (“near”) line. On the other hand, existing consumers were allowed to switch to the “far” queue when deemed beneficial. New customers arrived to the “near” or “far” queue following a Poisson distributed with rate λ ; And were processed at Poisson distributed service rates μ_1 and μ_2 respectively. The switching happened when the queue sizes varied by one and this was always from the longer to the shorter line. The objective was to formulate expressions for steady state conditions in infinite time. Expressions for the rate of change in queue sizes $\frac{\delta p_{xy}(t)}{\delta t}$ were first derived. Then, from the formulations of the queue occupancies $p_{xy}(t)$ as $t \rightarrow \infty$ (where $p_{xy}(t) = \Pr\{X(t) = x, X'(t) = y\}$) evolved Equation (1) as the bi-variate generating function that defined for equilibrium conditions. This was a product over state space changes $s^x s' y$ whose summations characterized for the generating function. Partial derivations of this generating function were then evaluated under variations in queue states $[(s = s' = 0), (s = 1, s' = 0), (s = 0, s' = 1), (s = s' = 1)]$ given arrivals, exits or both events

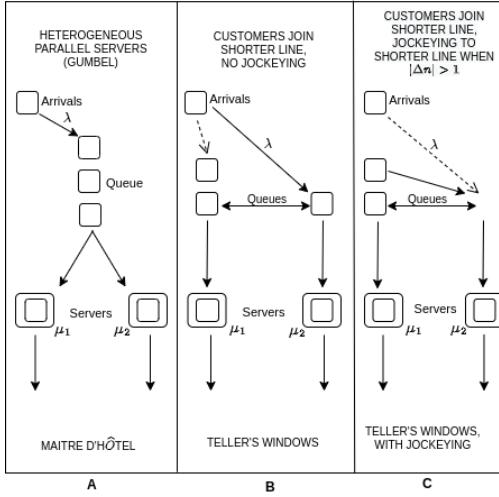


Fig. 1. Jockeying strategies: In the left-most (A) of the illustration is the Maitre d'Hotel queueing strategy where customers waited in a single line and got served when one of the available stations was empty. In the middle (B) is the Tellers' Window strategy where customers joined and waited in the shorter of the two queues and no switching lines was permitted thereafter. In the right-most (C) was the Tellers' Window with Jockeying, a behaviour where despite a new customer having joined the shorter of the two queues, switching to an alternative one was permitted later given a deviation in the sizes by one.

happening simultaneously:

$$\Phi(s, s') = \sum_x \sum_y P_{xy} s^x s'^y, \quad (1)$$

where s and s' denoted the states of the “near” and “far” queues in terms of their queue lengths respectively.

The distributions of the sizes for each queue and the overall system occupancy were formulated for from the representational difference equations. In conclusion, the findings derived for the stability conditions of the queue lengths when customers needed to switch from one line to another. And the solutions for these conditions in each of the states were expressed for in terms of probability that both queues were not occupied.

[90] would later on conduct more concrete studies that compared heterogeneous queue setups. The findings assumed tenants could instantaneously jockey given some threshold on queue length difference (adjacent queue shorter by one) or jockey based on some stochastic computations (based on how the differences in sizes of the queue changed). Figure 1 was a depiction of the heterogeneity in setup (also referred to as strategies). Here, each setup was associated with a set of behavioral rules for the customers. For example, under the “Tellers’ Windows with Jockeying” strategy (Figure 1C), new arrivals were queued to the end of the shorter line. They could however move to the other queue when the difference between the two queues exceeded one i.e. instantaneous strategy. The case of the stochastic-based jockeying strategy on the other was a factor of the rate $k(w_i - w_j)$ (where w_i, w_j as queue sizes) at which the queues varied in size. Specific of the probabilistic jockeying, the structure of the generator function and the evaluations for the steady state conditions exhibited known properties. These properties were descriptive of a phenomenon observed in similar setups where customers simply joined the shortest queue and never left until the end of service (“Teller’s Window”). For the instantaneous jockeying, equilibrium conditions were characterized for under the three presumed occupancy state categories (n, n) , $(n + 1, n)$ and $(n, n + 1)$ (where n was the average number of occupants in a given queue). These categories restricted jockeying within the system to the shorter line always. More complexity was introduced in the “Lane Changing” strategy (setup) where switching was not based on a preset jockeying threshold but the rate at which the two service lines differed in length. It was however observed that under this setup, the number of customers that wanted to jockey grew

exponentially and that equilibrium conditions were a factor of only λ , ($\lambda = \lambda_1 + \lambda_2$) and not λ_i , $i = 1, 2$. In addition to that, it was noted that some of these strategies led to states where one service line was empty while the other had customers queueing up, states in which queue utilization was compromised. Another interesting customer flow "Route changing" was presented where jockeying was probabilistic, depending on how dissatisfied a customer was being at a certain position in the queue. This was given the fact that customer had no access to queue status information. The probability that a customer moved from queue i to j at time t was therein computed as $P_{ij}(t) = \begin{cases} 0, & n_i = 0 \\ 1 - e^{-k_i(n_i-1)t}, & n_i \geq 1 \end{cases}$ where n_i was the number of users in queue i and j was the preferred queue.

From the transition equations, the following system descriptors were characterized for: the equilibrium conditions of the average queue occupancy, the expected number of customers processed and the probability that a queue was not occupied. The work finally presented results from the numerical studies that involved experimenting with varying queue parameters under the aforementioned strategies (setups). Further insight was shared about the queues' performance and quantitative measures on certain queue descriptors.

It would later be revealed that, using the generating function to formulate for the steady-state solution was not the only approach to the problem. According to [32], a closed form solution for some queue descriptors like queue length could be derived from inversion operations on the representational state space matrix. Customers that joined the $M/M/C$ queue system were governed by different rules at admission (like which queue to join based on probabilities or queue size). Jockeying was permitted only when the difference between the sizes of the two queues hit the preset threshold (two). The authors differed in methodology to reason that the underlying Markov process could be modeled using a transition diagram to capture the changes in states. The formulation for the proof followed from the definitions of coefficient matrices that underpinned the transitions in state spaces (Λ_0, Λ_n). The constituent sub-matrices were then partitioned ($\Lambda_{0i}, \Lambda_{ni}, i = 0, 1, \dots, C$) to characterize for this coefficient matrix Λ in equation (2) and the equilibrium conditions $\Lambda P = 0$. Here P denoted a column vector that defined for the probabilities that a queue was in a given state. This matrix Λ was composed of sub-matrices (like $\Lambda_{0i}, \Lambda_{ni}, i = 0, 1, \dots, C$) partitioned along the number of queues and their respective sizes. According to the authors, this matrix was constituent of the regularity property and it was from this property that the solution emanated. It was therefore argued that the solution depended on choosing the proper sub-divisions.

$$\Lambda = \begin{bmatrix} \lambda_{01} & \lambda_{02} & 0 & 0 & 0 & 0 & \dots \\ 0 & \lambda_{11} & \lambda_{12} & 0 & 0 & 0 & \dots \\ 0 & 0 & \lambda_{21} & \lambda_{22} & 0 & 0 & \dots \\ 0 & 0 & 0 & \lambda_{31} & \lambda_{32} & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \quad (2)$$

The matrices of varying dimensions were defined and elements partitioned into vectors to characterize specific queue state probabilities. This iterative technique was followed by an inversion of each matrix (Λ_{n2} and Λ_{02}) to yield derivations for the closed form solution of all state probabilities. It was shown how the preference for a certain buffer when both were equal in size then became a factor of the distance ($n_i - n_k \geq 2$; i, k were queues) the jockey candidate had to traverse to the alternative queue. It was also shown how a closed form solution could exist in scenarios where the coefficient matrix was dependent on the number of customers in a queue.

The idea of [32] about sub-dividing the coefficient matrix motivated further exploratory work with queue descriptors as documented in [36]'s findings. The authors studied how queue parameters like time-until-service and how often newcomers arrived in the system influenced the jockeying behavior. To switch to any shorter line was preconditioned on the size of one of the queues in an $M/M/C$ setup differing by two. Also, a different technique for the partitioning of the state space transitions into sub-matrices was utilized. The first model

considered the case when no jockeying was permitted. The authors formulated for the stability conditions and transition probabilities from the representational differential equations. Analytically, these differential equations were generalized as $AP = 0$. A denoted an $n(C^2 - 1) + 1$ sized matrix whose elements corresponded to the weights (coefficient) of the state probabilities (where n was the queue's capacity). Vector P on the other hand denoted the actual probabilities that a queue was in a certain state and 0 was a vector that constituted non-zero elements. It was this matrix A that was sub-divided into a series of other column vectors and sub-matrices of varying dimensions. Then by applying boundary constraints to the generalized equation AP , expressions that evaluated for the probabilities of the queue being a given state were formulated. The second model then analyzed the jockeying behavior under recalibration in queue descriptors like the arrival λ , the service μ rates and system utilization ρ relative to the state of the a given queue. This was evident in the variation in the representative matrix A and magnitudes of the state probabilities. The performance results revealed that the time a customer had to wait until being processed was greater when jockeying was not allowed compared to when the behavior was permitted. This was affirmative of [32]'s findings about the benefits associated to such impatient behavior. Additionally, it was observed that there was a higher probability of the queues staying idle under the non-jockeying setup. Other performance measures like the predicted system occupancy under varying degrees of the system utilization were documented.

[170] was contentious about the techniques in the preceding studies, arguing that besides the approaches being repetitive, the solution ignored boundary conditions when defining the state space transitions. In their $M/M/C$ model therefore, the findings were aimed at unraveling some hidden dependencies between the transition rate matrix R and the overall queue utilization. New admissions to the shorter of $C \geq 2$ queues tended to a Poisson distribution with rate λ . Each server processed jobs at Identical and Independently Distributed (IID) rates $\mu_1, \mu_2, \dots, \mu_n$. In infinite time, the changes ($S = \{ \vec{i} = (i_1, \dots, i_C | i_j \geq 0) \text{ for } 1 \leq j \leq C \}$ and $|i_k - i_l| \leq 1$) in the server sizes were generalized as a stochastic process $\{(X_1(t), X_2(t), \dots, X_C(t)), t \geq 0\}$. Here k and l denoted the number of customers in given states and stable conditions for this process were defined as $\vec{p}_i = \lim_{n \rightarrow \infty} P_i(t), \vec{i} \in S$. The generator matrix Q and its partitions (sub-matrices A_{ij} with varying dimensions) resulted from arranging the state space, that is, the ordering of state transitions i and j . This was based on a function that defined which state came prior to or after another. The distributions of \vec{p}_i were said to each constitute $2^n - 1$ elements as states that formed a block i with P_0, \vec{p}_1 as the bounds. From this evolved the difference equations essential for the derivation of the equilibrium probabilities as a factor of the traffic intensity ρ . Taking the case of $i \geq 2$ (therefore only interested in the $2^n - 2$ probabilities in \vec{p}_2 since the bounds were known), it was shown that a solution only existed under steady conditions defined by $\vec{p}_{i+1} = \vec{p}_2 \omega^{i-1}, i \geq 1$ only when $\det(A_0 + A_1 \omega + A_2 \omega^2) = 0, 0 < |\omega| < 1$ (where $\omega = \rho^C$). Then building on the theoretical properties of the eigenvalue(s) $r_2^C - 1$ of the rate matrix R in the determinant, it was shown that ρ^C was the only quantity that held true under this conditionality. These theorems were the basis for evaluating for the boundary and equilibrium probabilities of $p(i_1, \dots, i_C)$, \vec{p}_{i+1} (for block $i+1$) and \vec{p}_2 to yield a relation between the rate matrix and the equilibrium probabilities. Then taking the case of C servers, to show that ρ^C was the lone zero in the determinant above, a new chain for the stochastic process ($\bar{S} = \{(i_1, i_2, \dots, i_C) | i_t = 0 \text{ or } 1, 1 \leq t \leq C\} \cup \{m | m > C\}$) consisting of 2^C states was defined. The associated transition state infinitesimal generator matrix \bar{Q} was also constructed. Most ($2^C - 1$) of the columns in both generator matrices (Q and \bar{Q}) were similar except for some few states (for which a specific generator matrix was formulated). The proof showed that for any server k in the server deck n , $(\frac{p^n}{\omega})_i = \text{constant}, i \geq 1$ if and only if $p^C = \omega$ (ω as the only real solution of the determinant). It was from these derivations that quantitative measures in the average system occupancy, the time jobs took until exiting the system, were characterized for. The formulations were validated using numerical experiments under varying configurations of the system descriptors.

Unlike single-customer impatience models, [107] studied bulk workload migration. Arriving jobs were tagged with an expiry time and, if not processed before this time lapsed, were transferred to other servers. The transfer was defined by a policy chosen to minimize holding and migration costs. Arrivals obeyed a Poisson distribution while the service times were exponentially distributed. The paper first analyzed the case of a single-server fallback (where expired jobs reneged) and then generalized to N servers with a transfer probability matrix $Q = (q_{i,j})$. Transfer rates take the form $\beta_i q_{i,j}$ and stationary probabilities $p_{i,j}$ (where i was the server status of up or down and j as the number of jobs) are derived for key cases. An iterative Poisson approximation was used to compute transfer rates and total cost, but the search for optimal Q becomes combinatorial hard as N grew. The authors therefore evaluated heuristic rules (round-robin, fastest-other, etc.) and reported numerical comparisons showing when heuristics approached optimal performance.

More like an extension to his [151] earlier work, [150] investigated the evolution in the transition state space of an asymmetric $M/M/2$ setup of finite capacity L . Therein, two theoretic propositions (randomisation theorem and Runge-Kutta) were essential in the formulation for the dynamics in the state space. Jockeying was only allowed from the longer line only when any of the other service line was empty. The probability of having a certain number of customers at a time t ($p_{i,j} = \Pr(N_1(t) = i, N_2(t) = j)$) in either service lines was defined in a series of difference equations ([150], Equations 1 - 21). The formulation of the model stemmed from the definition of the state probability vectors $P_k(t)$ (and their derivatives $P'_k(t)$, $k = 0, 1, 2, 3, \dots, L$) for all state transitions over L during this time t . The difference equations were then re-defined in terms of these state probability vectors. These equations were re-arranged into a generalized block-matrix formation $P'(t) = QP(t)$ (where $P(t) = (P_0(t), P_1(t), P_2(t), \dots, P_L(t))^T$) constituted by sub-matrices (A, B, C) . The sub-matrices denoted state space partitions to compose the generator matrix Q as:

$$Q = \begin{bmatrix} B_0 & C_1 & 0 & \dots & 0 \\ A_0 & B_1 & C_2 & 0 & 0 \\ & A_1 & B_2 & C_3 & 0 \\ & & & & \dots \\ 0 & \dots & 0 & A_{L-1} & B_L \end{bmatrix} \quad (3)$$

It was argued therein that, the expressions for the rate matrices ($R_L = B_L^{-1}, R_k$) could be ascertained from iterations of computations. This approach was in deviation from the usual technique of calculating for the eigenvalues and eigenvectors of the rate matrix. The modified vector-geometric solution for the equilibrium probabilities of the stochastic process emerged therefore from the evaluations of the block-matrix equations in relation the rate matrix given $P'(t) = 0$. And the solution for state transition probabilities ($P_{i,j}(t)$) in finite state space followed from the randomization theorem [150, Theorem 3.1]. From the subsequent application of the same theorem to the difference equations emanated the recurrence expressions. And based on the properties of these recurring expressions, the distribution of the state transitions was formulated for. It was also shown how the Runge-Kutta method could be manipulated to provide an equivalent evaluation for the transition state distribution. The work then provided numerical and comparative analyses of both methods used in formulating for the densities of selected descriptors like the system occupancy. The studies also assessed how the probabilities and capacities of the queues were influenced by the overall system utilization ρ under variations in queue parameters. Also, the impact of switching queues on the average processing times and sizes of the queues was analyzed versus when no switching queues was possible.

[169] studied the jockeying behavior and it's applicability to packet routing in multi-beam satellite systems. The earth-stations were ordered as disjoint zones to form up-link and down-link connections and the sequence of the incoming packets was defined by an independent distribution function. New packets were appended to the end of any of the shortest buffers at the satellites. They were then processed at varying Markovian service rates following the First-Come-First-Served (FCFS) rule. The vector-geometric solution was based on the assumption

that the underlying process that characterized for such behavior was non-Markovian. Therefore, the process was segmented to first define an "imbedded Markov chain" ($\{\vec{X}_l = (X_1(t_l), X_2(t_l), \dots, X_c(t_l)) | l = 1, 2, \dots\}$). Then expressions for the probabilities of the buffers' capacities for this chain were ascertained under ergodic conditions. The solution therefore for stable conditions of the buffers arose from dividing the state spaces into groups $B_{<r}$, B_m (maximum size of the buffers and maximum deviation between smallest and largest buffer respectively). The division was premised on the the probability (ω) that there existed only a one-to-one relationship between any two states (with the exception of the boundary states) and that new packets did not necessarily visit all states. Following the sub-division of the state transition matrix was the formulation of the equilibrium equations (both for non-boundary and the boundaries states) that characterized for these sub-divisions. The solution for the queue equations borrowed from earlier findings ([116], lemmas 1.2.4). That is, there existed an eigenvalue $\omega = \omega_0$; $0 < \omega_0 < 1$ of the transition rate matrix R and its determinant $\det(\omega I - \sum_{k=0}^{\infty} \omega^k A_k)$ was 0. Here I denoted an Identity matrix and A_k was a square sub-matrix block of states when there were k packets in the system. This lemma was fundamental to ascertaining whether the vector-geometric solution was valid for the boundary equations . This was by simply taking the probabilities of any of the boundary states ($p_{r,r,\dots,r}$) in \vec{p}_r and evaluating each for redundancy in the equilibrium equations. For the stationary probabilities of this "imbedded Markov chain" X_1 therefore, there existed a geometric parameter $\omega = \sigma$ that defined the uniqueness of the solution. Relatively, the stationary probabilities vector ($\pi_{\vec{i}} = \lim_{t \rightarrow \infty} P\{\vec{X}(t) = \vec{i}\}$, $\vec{i} \in S$, S denoted the state space) of the buffer capacities was also a modified vector-geometric solution governed by similar uniqueness constraints. The studies furthermore defined another "imbedded semi-Markov chain" ($X_i^*(t)$, $i = 1, 2, \dots, c$) that represented the buffer capacities prior to any last packet at any time. It was discovered that the stationary probabilities for this kind of process were similar to those of the "imbedded markov chain". Then, from the sub-division of the stationary probability vectors $\vec{\pi} = (\vec{\pi}_{<r}, \vec{\pi}_r, \vec{\pi}_{r+1}, \dots)$, it was shown that the process assumed a modified vector-geometric solution too. Numerical evaluations of the performance when jockeying was permitted versus when this behavior was prohibited were documented. The performance results showed how the time the packets spent in the buffers varied relative to the processing times, justifying the positive effects of this impatience behavior.

The application of impatience in resource allocation was amplified by [161]. The work derived for optimal rule-sets that controlled packets in an $M/M/2$ setup of multi-beam satellite stations as buffer with infinite capacity. Every packet that joined the server incurred holding costs and moving (instantaneous) a packet from one lane to another generated a jockeying cost but no preemption was allowed. Controlling the admission or jockeying behavior was also conditioned on the state of current service station. This state was defined by the magnitude of the load and the monotonic properties of the function F_{ij} . This control function F_{ij} (optimal rule-set) encapsulated the accumulated costs under discounted or non-discounted service costs and long-run average costs. Therefore, a packet would only be routed to or migrated to another station if that station was in a valid state (x_1, x_2) at time t . In the model, actions were state dependent; that is, given an increase or decrease in the size of either queues. Basically, the work argued for the existence of a function $F(x)$ that defined when it was okay for a new packet to be routed to a given queue, and when or not to move packets. These functions were then adopted in the characterization of the optimal rule-sets for the different behavior. Explicitly, a new packet was routed to lane 2 only if $x_2 \leq F(x_1)$ was true. That is, if F_{12} and F_{21} were true then it was okay to move a packet from lane 1 to 2 ($x_2 \leq F_{12}(x_1)$) or from lane 2 to lane 1 ($x_2 \geq F_{21}(x_1)$). Migrations of packets was not optimal when $F_{12}(x_1) < x_1 < F_{21}(x_1)$ and when $F_{12}(x_1) < x_2 < F_{21}(x_1)$. One restrictive characteristic of the control functions was that the decision to move from a station with lower costs than its alternative was only plausible if the alternative station was idle. This characteristic was evaluated by taking a state of the system when the alternative station was not empty and validating it under the optimal policy when jockeying was permitted. The studies additionally formulated for the asymptotic profiles of the control functions that defined the optimal rule-set F

under discounted cost. These asymptotic characteristics were observed to be generic to the non-discounted mean costing over time. Expressions for the expected reduction in costs ($V_t(x_1, x_2)$ over time t and a predicted mean costs over longer time-span under steady conditions were also documented. It was ascertained though that, the control functions under both costing environments only converged under specific conditions of the both the jockeying and service costs.

This notion of cost-oriented modeling to guide the impatience in queues was further studied by [141]. In their findings, it did not matter whether the shorter or longer queue was joined or jockeyed to as long as the choice for either yielded costs below a preset limit c . These costs were a factor of both the size of the queue (N_i) and processing fee $\beta_i > 0$, ($i \in \{1, 2\}$, i denoted a queue). A control function was also embedded to manage the admission of new customers whose arrival periodicity obeyed a Poisson distribution with rate λ . The arrivals were routed to a queue therefore based on the measure $\alpha_i N_i + \beta_i \leq c$ (where $\alpha_i = 1$ was a weighting measure on queue) of either queue. Alternatively, consumers had the option to stay away from the services when the overall service costs would turn out to be excessively high. Each buffer was finite in capacity to a maximum defined by $K_i = [c - \beta_i + 1]$. The representational Markov chain $X(t) = (X_1(t), X_2(t))$ of the changing queue sizes was irreducible and a factor of the underlying costs. Formulations for the steady-state distribution deriving from theoretical comparisons between selected queue descriptors (K_i, β_i, c) were defined under specific quantities of the system utilization ($\rho = \frac{\lambda}{2\mu}, \rho \neq 1, \rho \neq \frac{1}{2}$). The comparisons were performed for three cases that were distinguished based on the magnitude of $K_1 - K_2$. This measure determined whether a new client had to join or jockeying to either *server1* or *server2*. For each of the cases $\beta_2 - \beta_1 - 1 < K_1 - K_2 < \beta_2 - \beta_1, \beta_2 - \beta_1 < K_1 - K_2 < \beta_2 - \beta_1 + 1$ and $K_1 - K_2 = \beta_2 - \beta_1$ all possible states reachable were defined. And for each of these inequalities, balance equations definitive of the influence of leaving or entering the queue in a given state were shown. Then taking into account what circumstances a specific state was reachable was essential for the proof for equilibrium probabilities ($\pi_{i,j}$) of the system. The solutions for these balance equations evolved from them being re-written as difference equations in relation to the state sequence $s_i, i \in \mathbb{Z}_+$. It was shown from induction principles how $\pi(i, j)$ could be derived for from relations between $\pi(1, 0)$ and $\pi(0, 0)$.

Similar admission-control schemes route arrivals between two service lines (M/M/C, $C = 2$) using Bernoulli-assigned probabilities and charge customers for waiting and for any jockeying between queues [100], building on earlier work [29]. Customers continuously receive queue-state updates and decide whether to stay or switch so as to minimize expected cumulative cost; the system state is $x = (q_1, q_2, \ell)$ and actions $a \in \{0, 1\}$ incur position-dependent costs V_n, V'_n that evolve with time. The authors derive threshold-type (limit) policies: there exist queue-length bounds $q^*(q_1, \ell), \ell^*(q_1, q_2)$ and an admission threshold $\varepsilon(q_1)$ such that a customer switches only when the alternative queue lies below (or above) these limits, and a new arrival prefers one queue when the other exceeds $\varepsilon(q_1)$. Monotonicity properties of the optimal rule set are proved to show that the limits are nondecreasing or nonincreasing in the relevant arguments. So both jockeying and admission rules reduce to simple threshold decisions in the long run – results that echo and refine the limit-policy analysis in [33].

More theoretical models to optimize impatience among jockeying consumers was the subject of [129]’s work. The studies analytically expressed for the frequency of jockeys made from one service station to another before getting served. New arrivals (obeyed a Poisson distribution) were pushed to the shorter station or one of the two stations with equal probability. Each jockeying activity accumulated a cost for moving from the tail-end of a more occupied station to the end of the less occupied station. This was given that the difference between the length of any of $k \geq 2$ servers hit the preset threshold d . (Each station served waiting customers at IID processing times. The transition state space partitioning here was different. It was based on queue length statistics like customers ahead (f) or behind (b) a given customer ($(f, b) \implies b \geq f \geq 1$) in the service line. An additional constraint on the behavior was whether the next move by that customer was a jockey or a forward in the same queue. Inspired by concepts from generating function theory, the objective function encapsulated mappings for

a customer's position in the queue to the possible number of jockeys ($Y_{f,b}$) that the customer would make before being processed. The evaluation of the generating function was a procedural and iterative formulation of the relationships ($\Phi_{f,b}(s) \ 0 \leq s \leq 0$) between these states and actions. This required initial statistical computation for the chance ($P_f(j|b)$, $1 \leq f \leq \min b, j$) that the customer's next state ($f - 1, f$) followed a jockey to the alternative queue or a forward move in the same queue; Such that the customer ended up in state ($f - 1, f$) $f \leq \min j, b$. The expression for the chance that a customer would move to the alternative station (in state ($f - 1, j$), $f \leq j \leq b - 1$ or $j \leq b$) were premised on the prevailing conditions in the current queue. That is, provided that $b > f - 1$ and that the variation between the number of clients behind the current customer that left that station versus the number that wanted to join any of the queues was $b - j$. Equation (4) was this generator function that characterized for the conditions, state probabilities and relations. This function mapped the expected number of jockeys ($Y_{f,b}$) from one queue to another before the customer was serviced. This frequency was dependent on f (the number of people ahead of the current customer being served) and the arrival rate.

$$\Phi_Y(S) = \frac{2 - \rho}{2 + \rho} \left[\frac{2 + \rho}{2} + 2B_0(S) + \frac{4}{\rho} B_1(S) \right] \quad (4)$$

where B_0 and B_1 were linear expressions that denoted aggregations of state relations and service line utilization over n customers in a queue. And $0 \leq s \leq 1$.

Additionally, a generating function ($\Psi_K(\theta)$) that yielded the random distribution of how many customers K ahead of the current customer left the queue was defined. This followed the analysis of the process representative of either actions (joining or leaving) over time, therein referred to as a "difference random walk" (DRW).

The authors in [100] combined static and dynamic admission, service and jockeying controls to derive an equilibrium "hedge-point" policy. Here, incoming jobs did not pre-select a queue but were routed by an admission controller [29]. Each queue operated at exponential service times, and the admission rule $u \in \mathcal{U}$ mapped system state x to actions that traded off holding, service and jockeying costs via a discounted value $J^u(x)$. For optimal control, the value function $V(x) = \max_{u \in \mathcal{U}} J_u(x)$ had to be maximized. This was by iterating a Bellman operator T and retaining only value functions with the required structural properties (sub-modularity or concavity) so that the operator converged. The resulting optimal policy was threshold-style and described by switching functions S_1, S_2, L_i, G_i that governed admissions, routing and migrations. The numerical value-iteration experiments for symmetric servers (equal costs and rates) illustrated the approach and the convergent approximation of $V(x)$.

Building on earlier findings [58, 111], [72]'s work focused on the asymptotic properties of the boundary conditions of the queue length distributions. In an $M/M/2$ setup, the aim was to formulate for the tail probabilities of the two heterogeneous queues. The growth in size of the main buffer was continuously monitored at exponential time intervals. And when the size L of this main queue exceeded a preset threshold K , workload was transferred to the auxiliary queue. For the representative continuous Markov chain $\{(L_1(t), L_2(t)), t \geq 0\}$, steady-state conditions of the queue sizes were first derived for as $\lambda q < \mu_2$ and $\lambda < \mu_1 + \mu_2$. Here, $\lambda < \mu_1, \lambda < \mu_2$ denoted the arrival rates and q the probability that new arrivals joined the auxiliary buffer). The proof developed from adoption of the Foster-Lyapunov[40][108] condition for stability to show that the some process states were revisited in finite time (positive recurrence). For uniformity, the transition probability matrix $P = I + Q$ (where Q represented that rates at which the queue varied in size and I an Identity matrix) was subsequently divided into sub-matrices based on levels. It was observed that one of the sub-matrices $D(\sigma) = A + B\sigma + C\sigma^2$ (where σ as defined by Eq. (5) was the decay rate, and A, B and C were sub-matrices depicting state space splits) was irreducible and inherent of special spectral properties. These properties were fundamental to the derivations, such that a further split of the sub-matrix D along the boundaries was followed with evaluations for its convergence norms

$\overrightarrow{\sigma}, 0 < \sigma < 1$; And verification for the existence of the $\frac{1}{\sigma} - invariant$ measures [72, Lemmas 4.1-4.2].

$$\sigma = \frac{(\lambda p + \mu_1 + \eta) - \sqrt{(\lambda p + \mu_1 + \eta)^2 - 4\lambda p \mu_1}}{2\mu_1} \quad (5)$$

where p was the probability of joining the main queue, η was a distribution of the server polling intervals.

Specifically, when $\lambda q < \mu_2$ and $\lambda < \mu_1 + \mu_2$, it was shown that the asymptotic properties for joint stationary distribution $\pi_{i,j}$ (i, j denoted the level and phase in the state space composition) as approximated from the decay rate decreased geometrically. In essence, it was conclusive that, the decay rate varied with the individual queue sizes such that the number of customers in the main line decreased at some constant ratio in infinite time. This was contrary to the number of customers in the auxiliary line, which stayed constant. Numerical analyses of the these formulations were performed, experimenting with different queue parameter settings for the arrival or processing rates in each queue. At different monitoring intervals, results were compared under changing values of the arrival rate and it was observable that the increase in the decaying rate was linearly proportionally to the rate at which customers joined the queue. It was interesting to also observe how the decaying rate responded to increasing quantities of the processing rates. Of performance significance was the observation that the decaying rate decreased linearly with respect to an increase in the service rates of the queues.

[132]’s work proposed an alternative energy-efficient heuristic for guiding efficient use of energy when allocating jobs in multi-server processor sharing setups. The setup was composed of heterogeneous infrastructure with finite buffer sizes. The job streams followed a Poisson distribution at rate λ to land on any of $j \leq n$ servers for processing at exponentially distributed rates. Each server consumed energy (ε) at a rate (defined by (6)) that was monotonically decreasing with the service rate. The object of the studies was to develop energy efficiency policies when apportioning workload while maximizing throughput.

$$\varepsilon(\mu) = \mu^3 \quad (6)$$

where μ denoted the node’s processing speed.

In the studies, two policies were benchmarked. The baseline heuristic was the insensitive jockeying policy, therein referred to as slowest-server-first (SSF). Jockeyed jobs displaced existing jobs backward or forward and position allocations for jockeys were defined with equal probability to departures. The proposed energy-efficient (EE) rule-set on the other hand allocated tasks to the first $\hat{n} \leq n$ set of busy servers such that the tasks were routed to the least occupied or empty buffers. Then the next $\frac{s}{b}$ servers were selected for task processing if all instances in the \hat{n} set were occupied. Here, b defined for the finite size of a particular queue and the state space $s \in S$ denoted the number of jobs in the queue. The energy efficiency of the servers was then collectively calculated as the ratio of the summed long-run mean throughput T and the expected consumed power E as $\frac{T}{E}$. It was revealed that this server cascading (in states $\hat{n} < s \leq \hat{n}b, s \in S$) yielded relatively lower task processing times to minimize the overall load and balance server utilization. Formulations for comparative analysis of the two heuristics followed from the theoretical propositions that showed conditions for \hat{n}, b, η and μ_j ($\hat{n} \geq 1$) under which the energy-efficient policy was more optimal than the SSF policy. Qualitative measures were then defined by the relative error between the two rule-sets computed using (7). Numerical evaluations of this error were performed in selected quantities of η and μ in a series of experiments. This was purposely to determine the optimal \hat{n} that maximized the servers’ energy use ($\frac{T}{E}$) in the entire set.

$$\Delta \hat{E}_{EE_{\hat{n}}}^{SSF} = \Delta \left(\frac{T_{EE_{\hat{n}}}}{E_{EE_{\hat{n}}}}, \frac{T_{SSF}}{E_{SSF}} \right) \quad (7)$$

It was shown that, a certain measure of \hat{n}_* provided a ceiling such that the mean aggregated processing rates should never exceed the mean arrival rates. Under preset configurations of b, a_i (where a_i was a measure of how two cascaded server groups differed in processing capacity), the EE policy was more optimal energy-wise than

the SSF. However, this efficiency decreased with server utilization or traffic intensity. This meant that EE was only more optimal under limited traffic conditions. Under similar settings though, no significant improvements were recorded in the throughput of the EE ruleset.

Limitations of Statistical Models in Next Generation Networks: *Statistical modeling of jockeying in queues faces several practical limits in next-generation networks. For example, in such environments telemetry is often partial, delayed or censored, producing biased and high-variance estimates. Also, the system is very dynamic (mobility, auto-scaling, flash crowds) causing rapid concept drifts. New architectural abstractions like slice or tenant heterogeneity invalidates pooled models. And rare but useful tail events like SLA breaches, are underrepresented by standard loss functions. closed-loop effects (predictions changing behaviour) and multi-agent interactions break offline validity; edge compute and latency constraints limit model complexity and freshness; uncertainty quantification is frequently absent or miscalibrated; telemetry can be adversarially manipulated; and ground-truth labels for “beneficial” jockeying are ambiguous—altogether making naïve statistical predictors fragile unless paired with censoring-aware estimators, adaptive retraining, causal or closed-loop evaluation, and robust telemetry authentication.*

3.2 Nash Equilibrium based Models

Nash equilibrium rules have also been applied in studies that model for impatient in buffering systems. Take the case of [60] who was interested in how valuable prior queue status information was to the prospective consumer before joining the system. The experiment in an $M/M/C = 2$ setup were motivated by the notion that such information underpinned optimal usage of the service lines to consequently minimize waiting times. This meant that, for new arrivals, preference for which queue to join was influenced by the precedent customer having bought similar information (externalities). Basically, on arrival a customer purchased a probability value that abstracted the potential benefits given the current state of the queue. The authors sought to ascertain whether this information from prior clients impacted the subsequent client positively or negatively. Analogous to a cost benefit model, the authors deployed Nash-equilibrium strategies that put a value on the purchased status information by evaluating how much benefit a client got from it. Two strategies arose here $p, 0 \leq p \leq 1$ which denoted the probability whether the information was purchased or not respectively. That is, a pure and a mixed strategy. But the strategies ($g(p) = C, g(p) \leq C, g(p) \geq C$) were a factor of the relation between the benefit of the acquired information $g(p)$ and the costs C . The expected benefit $g(p)$ meant knowing how much less time the consumer would be waiting to get serviced given the charges for that information. Following a partitioning of the state space, the authors used the matrix-geometric method [116] to obtain the stationary probabilities $\pi_{i,j}$ (i or j being sizes of either queues) of each queue size given the jockeying threshold $N = 3$ ($3 \leq N \leq \infty$). The stationary probabilities were then characterized for from the eigenvalues/eigenvectors and spectral properties of the rate matrix R . For each assumed position in the therefore, a function $g(p)$ for computing the benefit (expected waiting time) at that position under a given Nash-Equilibrium strategy. The numerical results compared the benefit from purchases under varying measures in the jockeying threshold N and the service line utilization $\rho = \frac{\lambda}{2\mu}$ (where λ, μ were the arrival and service rate respectively). The study conclusively expressed for the magnitude of influence (be it negative or positive) of the actions of prior customers on new arriving customers when they purchased knowledge. The effect was considered positive if the acquired knowledge helped the consumers optimally use the service line and negative if the customer ended up waiting longer than expected. Performance evaluations for the average sojourn time under varying magnitudes in system occupancy revealed that as N grew larger than 4, the benefits of purchasing knowledge were negligible.

Limitations of Nash Equilibrium Models in Next Generation Networks: *Nash models assume well-specified utilities and substantial knowledge (or common priors) about other players, assumptions that fail under partial, delayed, or privacy-constrained telemetry in slice-based MEC or 6G settings. They also require fully rational agents*

and often slow iterative dynamics; in practice agents are bounded-rational, run learned heuristics, and environments change too quickly for equilibria to form. Scalability is problematic—computing equilibria for many tenants, slices and cross-domain constraints is intractable—and enforcement across administrative domains (migration budgets, billing, penalties) is nontrivial. Finally, standard NE ignores adversarial misreporting. Mitigations include Bayesian/mean-field formulations, bounded-rationality or learning dynamics, mechanism design for enforceability, and trace-driven validation, but Nash Equilibrium techniques are mainly useful as a baseline or for small, slow subsystems.

3.3 Fluid Theory based Models

[33]’s fluid model analyses extend jockeying studies by treating bulk arrivals and transfers as continuous flows. In an $M/M/2$ setting with periodic Poisson batches, controllers decided bulk transfers between a low-cost and a high-cost service line to minimize long-run costs (holding plus fixed/variable transfer costs) while avoiding idling. Equation (8) characterized for the average predicted cost of a client starting in a state x . Then adopting a given rule-set π as $g^\pi(x)$, the optimal rule-set π^* was conditioned on $g^{\pi^*}(x) \leq g^\pi(x)$ for all states.

$$g^\pi(x) = \lim_{n \rightarrow \infty} \sup \frac{\mathbb{E}_x^\pi \left\{ \sum_{i=0}^n [k(X_i, Y_i) + \int_{\rho_i}^{\rho_{i+1}} c(X_i, Y_i) \delta t] \right\}}{\mathbb{E}_x^\pi \{\rho_n\}} \quad (8)$$

where X_i was the state at the i^{th} decision step and \mathbb{E}_x^π the expectation (given policy π) that taking action Y_i would accumulate $k(\cdot)$ as the overall cost at a cost rate $c(\cdot)$. Policies π specified the transfer magnitudes and timings and each policies were evaluated by their asymptotic average cost. The optimal rules were then those whose sample-path cost $g_\pi(x)$ was minimal for every state. The fluid model was said to be in equilibrium when all incoming clients had been processed ($\bar{M}(t) = 0$, $t \geq t_0$) by the system (where $\bar{M}(t)$ defined for changing system occupancy given instantaneous arrivals and service completions). And otherwise ($\bar{M}(\delta) \neq 0$; $\delta > 0$) if at a certain time-frame $\delta > 0$ the queues were still occupied. The theory yielded threshold/non-idling conditions under which bulk migrations were justified (move to the costly line only to prevent idling), and enforced symmetry constraints for single or bulk transfers. Numerical experiments for asymmetric multi-queue setups (varying costs and arrival rates) showed how polling interval T , transfer granularity and policy choice affected total cost and system load. And confirmed the benefits of carefully parameterized bulk-transfer rules. These results generalize earlier discrete-control findings by [92] while highlighting computational/design tradeoffs.

In [31]’s studies, fluid theory was the basis for formulating steady-state expressions for a cluster of cascading servers. The endpoints were stacked so that they collaborated to share task executions. Specific classes of customers with IID arrival times were assigned positions in specific queues. But these class specific queues served customers from other classes only when they were idle at IID processing times. Two variations in modeling were presented therein. One of the class of *X-models* as a setup with two servers, such that each represented a class of customers. The other model on the contrary, therein referred to as a "tree-cascade system", was defined with three servers (classes). It was argued that to compute the steady-state conditions for such a networked system, one needed to prove the stability of the underlying fluid limit model. necessary for these conditions was the emergent Markov process X having a non-repeating consistent statistical value [27]. Theoretically therefore, as earlier suggested by [30], given a service line with a lower processing rate ($r_1, r_2 \geq 1$ or $r_1, r_2 < 1$) than the other (given $r_1 = \frac{\mu_1}{\mu_{2,1}}$ and $r_2 = \frac{\mu_2}{\mu_{1,2}}$), then stability for fluid limit based models could only exist under specific conditions of comparative variations in the arrival (λ) and service rates (μ). These conditions were specified by Eq. (9), such that the aggregations over the quantities $\bar{Q}(0), \bar{U}(0), \bar{V}(0)$ of the queue network equations at a time $t \geq 0$ equaled unit, $\bar{Q}(t) = 0$, $t \geq t_1$.

$$\begin{cases} (A_1) & \lambda_1 - \mu_1 + \frac{\lambda_2 - \mu_2}{r_2} < 0, \\ (A_2) & \frac{\lambda_1 - \mu_1}{r_1} + \lambda_2 - \mu_2 < 0. \end{cases} \quad (9)$$

where $U(t)$ = the time before a new arrival seeks to join the server, $V(t)$ = time left to service end for customers and $Q(t)$ = size of the buffer at any time.

The proof for the stability of the X – model then derived from the adoption of the Lyapunov function $f(t)$ [108]. This function encapsulated the relationships between the sizes of the queues($\bar{Q}(1), \bar{Q}(2)$) over time $t \geq 0$. Essentially, it was then necessary to show that under varying conditions of λ_i and μ_1, μ_2 , ($1, 2$ being queues), this function was bounded by the inequality $f(t) \leq -C$ ($C > 0$). The studies were extended to the tree-cascade setup under work-conserving rules. In this setup, jockeying was allowed to any of the servers empty. However, the third station was dedicated to supporting the other two stations. The corresponding Markov process X was similar to the one for the two server setup with the exception that the new state space had a higher dimensionality. The implication of this was slight differences in the system equations that expressed for the interactions within the service stations. The stability of the fluid model in this setup was also premised on ratios of processing times ($r_{1,3} \leq r_{1,2}, r_{2,3}$) and the rate at which customers sought to join a given buffer. Similarly, the respective Lyapunov function $f(t)$ was defined [108]. And this function was inherent of comparable differentials at any time t with the server sizes. The proof for the equilibrium conditions then gathered from evaluation of the inequality $f(t) \leq -C < 0$; $C := \min\{C_1, C_2, C_3\}$ to hold under varying sizes of the queues ($\bar{Q}_i(t), i = 1, 2, 3$). It was shown how the fluid limits on the servers' occupancy hold stable in infinite time. Hence, the conclusion that under such server conditions, the stochastic process was characterized by revisits to some specific states in measurable time, that is, positive Harris Recurrent.

Limitations of Fluid Models in Next Generation Networks: *Fluid models average discrete events into continuous flows, which can obscure tail events and discrete triggers (spikes, deadlines, migration cascades) that drive SLA violations in MEC or 6G networks. These models also assume smooth, slowly varying inputs, yet abrupt network function scaling, fast mobility or rapid slice reconfigurations produce transitions not easy to encapsulate in fluid models. Hard constraints (migration budgets, per-task costs) and event-driven control logic are also awkward to represent, and fluid controllers can become unstable when feedback is delayed or partial. Fluid modeling however remains valuable for aggregate insight and baseline controller design but must be robustified for operational deployments.*

4 ANALYTIC MODELS

4.1 Matrix-geometric models

Modeling of the shortest queue problem using matrix geometric approaches was pioneered by [116]. This was the building block for [46]'s experiments to expressed for the stationary probability vectors of queueing systems. The study simulated an $M/M/C = 2$ airport setup with planes that accessed runways (as queues) at a frequency defined by a Poisson distribution with rate λ . The time at which planes took off was exponentially distributed with rate μ for each runway. The state space transitions evolving from all the dynamics within the systems were encoded by a continuous time markov chain. That encoding morphed into a form of an infinitesimal generator matrix Q defined by Eq.(10). The matrix was fundamentally constituted of sub-matrices (A_0, A_1, A_2 and B_0) representative of the state transitions.

$$Q = \begin{bmatrix} B_0 & A_0 & 0 & 0 & 0 & \dots \\ A_2 & A_1 & A_0 & 0 & \dots & \dots \\ 0 & A_2 & A_1 & A_0 & 0 & \dots \\ 0 & 0 & A_2 & A_1 & A_0 & 0 \dots \\ \vdots & & & & & \end{bmatrix} \quad (10)$$

Ideally, it was suggested that completing a service in either queue and a jockey to a shorter queue summed to a total transition rate of 2μ . This had the direct implication that, the dimensions of the sub-matrices that composed the generator matrix Q . It was therefore argued that, necessary steady state conditions of the stochastic process

existed only if $\lambda < 2\mu$ and a redefinition for the a non-negative rate matrix R that represented the rates at which states changed. This was premised on the proposition ([46], proposition 1) that, given the underlying dynamics in such processes, there was a probability that certain queue states (i, j) in the process were revisited in finite time (positive recurrence [116]). Subsequently expressions for the stationary probability vectors $\pi = (\pi_1, \dots, \pi_n)$ were formulated. And the stationary probability vectors were verified to exist under the steady conditions $\lambda < 2\mu$. The work provided a numerical analysis comparing experiments with variations in parameter settings for the queue intensity ρ . Under varying dimensions of the jockeying threshold n ($n = 5, n = 10, n = 15 \dots$), it was conclusively suggested that as the threshold n increased ($10 < n > \infty$), differences in stationary probability vectors became negligible. The experiments also provided computations for some queue statistics like the means of individual queue lengths and the mean waiting time. The findings were a further justification of earlier revelations [169] that systems where jockeying was permitted performed better than those where the behavior was prohibited.

[77] extended the application of these matrix-geometric techniques to an Markovian/ Markovian/ number of queues (M/M/C) queue setup. This followed the author's counter arguments about the methods used when partitioning the state space in [36]. That, by manipulating the structure of the generator matrix, a more reliable approach for ascertaining the stationary probability vectors for the stochastic chain of events. The experiments assumed multiple servers with an infinite number of tenants that arrived following a Poisson distribution with rate λ . Each server processed tenants at an exponentially distributed rate μ . New arrivals joined the shortest queue but if the queues were equal in length, they joined either queues with the same probability. Jockeying on the other hand was allowed only when the difference in queue sizes was two (jockeying threshold $n = 2$). The representation of such a Quasi-Birth-Death (QBD) process was an infinitesimal generator matrix Q composed of sub-matrices that denote the state transitions of the queue lengths. However, slight modifications were introduced to the structure of the stationary probability vectors. The justification of their argument followed from the proposition that the boundary conditions required a special extension of the probability vectors, hence the need for re-expressing for the rate matrix R . And theoretical findings (theorem 3 [47]) were fundamental for the formulation for this rate matrix R . The studies also revisited [128]'s work, arguing that because the structure of the infinitesimal generator matrix exhibited the likelihood of boundaries state being revisited, there was need for new definitions of the average size L of the queue. Therefore, the representative infinitesimal generator matrix was first re-partitioned. Then, it was proposed that, as part of the state space of the underlying QBD was an absorbing state θ . That is, a state when a new arrival got processed immediately. From the formulations of the Laplace Transform, expressions for the stationary waiting time probabilities and a closed-form solution for the average waiting time were derived. The authors additionally showed how application of randomization methods could yield for the characterization of the stationary waiting time probabilities $W(t)$ for QBD processes. These definitions followed from the assumption that, given a statistical initial state denoted as a vector z , there existed an n^{th} transition step of the stochastic process such that the queue was idle (state θ). The numerical analysis initialized a few system parameters to compute for some properties of rate matrix, the average queue occupancy L plus other system descriptors.

In contrast to earlier suggestions [46] [169] that a solution for the steady state could not be achieved when the number of queues C exceeded 2, [3] disagreed to assert that the solution lay within the state transition space sub-division method used. In the M/M/C setup, new entrants joined the shorter queue and jockeying was permitted given hitting a preset threshold defined around the difference between the queue sizes. It was argued therein that, although existing studies presumed that ergodic conditions could be derived from the sub-division of the state spaces, the solutions did not take the rate matrix into account. And that this necessitated revisiting the state space splitting approach. As a result therefore, the partitioning was based on sub-levels such that sets of sub-levels were mapped to sets of states. The splitting was a factor of whether a sub-level's behavior was regular ($l = T, T + 1, \dots$) or otherwise $(1, \dots, T - 1)$. Here, l denoted a collection of states and T the jockeying

threshold. From these sub-divisions, it was showed that, the condition for ergodicity was only possible when the system utilization $\rho < 1$. After partitioning based on sub-levels, the generator matrix Q took a different form that was irreducible but resolvable given [116](1.7.1)'s theoretical findings. Following the sub-divisions of the stationary probability vectors, it was possible to generate mappings between the sub-levels and the probability vectors using Eqs. (11) and (12).

$$p_l = P_T R^{l-T}, (l < T) \quad (11)$$

$$D_0 + RD_1 + R^2D_2 = 0 \quad (12)$$

where D_0, D_1 and D_2 denoted square sub-matrices constituting states at all sub-levels $\geq T$ and p_l or p_T as stable probability vectors corresponding to the level l (bound on T) that resulted from the split of the stable probability vector p

One of the sub-matrices in the generator Q took different dimensions due to the constituent states in the sub-level set. Then based on these structural deviations, a solution for the rate matrix R could be expressed. Borrowing from [128]'s theoretical formulations for the maximum eigenvalue of the rate matrix, the solution for this matrix was defined by Eq. (13).

$$R = \begin{pmatrix} 0 \\ w \end{pmatrix} \quad (13)$$

where $w = -v(D_1 + \rho^c D_2)^{-1}$, ($w = w_0, \dots, w_{m-1}$), m being number of states at a given level that defined the dimension of the square sub-matrices. c was the number of queues available

Contrary to the mostly homogeneous setups studied, [5] analyzed the behavior of two heterogeneous queues ($M/M/C = 2$), each operating at a different service rate. The studies were purposely to apply matrix-geometric techniques to derive expressions for sizes of the queues under stable conditions. New customers joined the shorter of either queues, or were alternatively routed based on a probability of the sizes of the two queues were equal. It was allowed to switch from a longer to shorter one given some jockeying threshold T . The work built on [4, 6]'s studies, where it was shown that, in infinite time when no jockeying allowed between queues, the equilibrium probabilities ($p_{m,n}$, m, n as queue sizes) of the queue sizes conformed to product-form solutions. These observations were subject to verification on whether the same assertion was true when jockeying was permitted. It was found that the evaluation held true only for a defined portion Q ($\max(m, n) > T$ and (T, T)) of states. Then observing the rates of change in state of the stochastic process for only this portion of states, it was conclusive that the process was irreducible. This meant further analysis of this portion of states as a separate process (with distribution $q_{m,n}$) with a relation to the main process (whose distribution was $p_{m,n} = q_{m,n}P(Q)$). Here $P(Q)$ was the probability that the portion Q included the main process. The product-form solution then derived from the notion of defining a set of metrics. These metrics were defined as a factor of the arrival rate, the service rate of the shorter queue and the queue admission probability. Then, the general purpose principle was the basis for the derivation of equations that resolved for these metrics. This system of equations characterized for the steady-state and ergodicity conditions for this portion of states Q . The work sought to also draw comparisons between the product-form solution and the one ascertained using the matrix-geometric technique. The authors proved that when the size of the larger line exceeded the jockeying threshold T , because of the unique formation at the boundaries, the formulation of the solution was the product of the stationary probabilities of each service line. Hence that given the appropriate sub-division of the state space, the matrix-geometric method yielded the same solution. For the geometric solution, the generator matrix emanated from sub-division of the steady-state probability vector into sub-levels. These sub-levels grouped the states depending on the size l of the longer queue ($l < T$ and $l \geq T$). And each sub-level was associated to a steady-state probability vector \vec{p}_l . It was noticeable that some of sub-matrices (A_0, A_1, A_2) in the generator matrix G were irreducible. Adopting theoretical foundations in [116] (Theorem 1.7.1, 1.7.11) was sufficient for the proof for ergodicity. That is, given knowledge about the

maximum eigenvalue of the R matrix, taking the unique structural properties of A_0 , \vec{p}_l when $l > T$ could be computed for. The the matrix-geometric formulation was found to bear similarities to the product-form solution.

[151]'s work then offered a performance comparison with slight adjustments in the $M/M/C = 2$ system. The arrivals were simulated to imitate a Poisson distribution with rate λ , choosing the shorter of the two service lines. Each operated at exponentially distributed service rates (μ_1 and μ_2) and the capacity of each queue was restricted to L . If none of the service lines was shorter than the other, preference for one would follow probabilities (α or β , $\alpha + \beta = 1$). Instantaneous jockeying was possible when one of the service line was empty. For the Markovian chain, the author defined a sequence of difference equations that characterized the probability $(P_{i,j})$ of queues (i,j) being in equilibrium state. The compressed form of the difference equations $(A_{k-1}P_{k-1} + B_kP_k + C_{k+1}P_{k+1} = 0, k = 2, 3, 4, \dots, L-1$ and $A_{L-1}P_{L-1} + B_LP_L = 0, k = L)$ were first re-arranged to compose for a matrix block $A_1P_1 + B_2P_2 + C_3P_3 = 0$. Here A, B and C were sub-matrices that encapsulated the dynamics in queue sizes and traffic intensities. Then the evaluations for the sub-matrices were computed from the definition of column vectors that encoded state transition probabilities over all positions in both queues. Theoretically, the proof required that the sub-matrix B_k was invertible. This required computing the inverse of the sub-matrix and its determinants. Such that, there existed an inverse of this sub-matrix only if its determinant did not evaluate to 0 for all values of the traffic intensity ρ . Then taking P_k^+ as the theoretical solution for stability conditions, $A_{L-1}P_{L-1} + B_LP_L = 0, k = L$ was conditional for the existence of this determinant. This solution characterized for probability that the system was occupied to maximum capacity ($P_L = -R_LA_{L-1}P_{L-1}, k = L$) when $R_L = B_L^{-1}$ (R being the rate matrix). Iteration of computations for R evolved into solutions for the differential equations expressed in terms of $P_{0,0}$ (probability that both queues were idle). The proof by induction on k was essential to relate the queue intensities and the total number of customers in the system as $g_k = \rho^{k-2}g_2$ for $k = 2, 3, \dots, 2L$. Based on this relation, equilibrium probabilities of the $M/M/2$ system were formulated for as the capacity doubled ($2L$) or as $L \rightarrow \infty$. A numerical analysis initialized the arrival and service rates to different values. Then comparisons were made with earlier results from the Conolly's model[26]. Performance evaluations for the effect of the system utilization on the equilibrium probability g_n under different queue setups showed that the author's model was quantitatively better.

The more practical and challenging aspects arise when multiple tenants seeking to switch buffers in an instant. [64] setup such a scene as two parallel $M/M/1$ servers. The portion of workload K ($0 < K < L$) would be moved from the longer to the shorter of either queues (q_1 to q_2 or vice versa) when the difference between their sizes hit the preset threshold L . Admissions obeyed a Poisson distribution with rate λ_i while the processing times of either servers were exponentially distributed at rates μ_1 and μ_2 . This activity was generalized as QBD process $((q_1(t), q_2(t)), t \geq 0)$ with state space $(n_1, n_2) : n_1 \geq 0, |n_1 - n_2| < L$ (where $n_i, i = 1, 2$ were the number of customers in a given queue). According to the authors, the structure of the stochastic process made it hard to analyze. However, the process constituted special properties $(q_1 - q_2 \nmid L - 1)$ which, coupled with inherent recurrence properties of such processes, allowed for the sub-division of the state space $\{(q_1(t), (q_2(t)), t \geq 0\}$. This sub-division was essentially to convert the process into a valid QBD to which matrix-geometric techniques could be applied for a solution to the equilibrium probabilities. The studies then built on the theoretical propositions of [64, 86], that for such processes, steady-state conditions could only exist under specific conditions of the traffic intensity ($\rho < 1$). This therefore meant re-organizing the state space until the resultant stochastic process $(X(t), J(t), t \geq 0)$ (individual queue occupancy as a level variable and the difference $J(t) = q_1 - q_2$ in sizes as the phase variable) was irreducible and not dependent on the system occupancy $q(t) = q_1 + q_2$. The matrix geometric solution was similar to [116],[97] and sought to resolve for the eigenvalues (one with the largest modulus - Perron-Frobesius eigenvalue) of the R matrix at each of these different sub-division levels in L . Besides formulating for the stable conditions of common descriptors like system redundancy or number of processed tasks, the rate at which bulk load was moved around in the system was also expressed for. Expressions for the rate

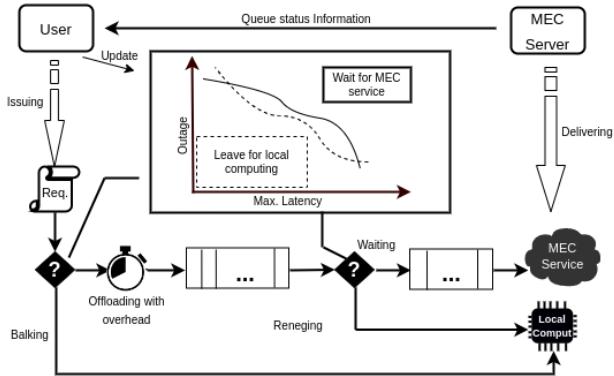


Fig. 2. Illustrating the impatient customer that had the option to either process a task on a MEC server or locally depending on the latency requirements of the underlying application

at which the distribution of system occupancy diminished were formulated. From this, it was deducted that this decay rate was neither affected by the workload transfer threshold L nor the number of workload migrations K under varying inequalities of the traffic intensity ρ . The analytic solution was validated by numerical evaluations that involved experimentation with variations in queue design parameters. The results revealed evidence of the convexity properties characteristic of the relationship between the service rates and the rates at which customers were transferred from one queue to another. It was conclusively suggested that, choosing the right processing rate for each queue was requisite for even load distribution. The parameter that exerted much influence on these decisions was the deviation in queue utilization (traffic intensity).

Limitations of Analytic Modeling in Next Generation Networks: *Matrix-geometric methods yield exact solutions for structured, stationary Markovian systems. However, emerging 5G/6G architectures that are characterized by slice heterogeneity, non-stationary dynamics, partial or delayed observations, migration/signaling costs, federated multi-vendor policies etc, invalidate the the assumptions in these analytic models.*

5 BEHAVIORAL MODELS

5.1 Modeling based on the Value of Information

Recent growth in latency-sensitive applications forces operators to prioritize traffic and manage selfish, impatient tenants. Han et al. [55] modeled this impatience (reneging/balking, akin to jockeying) in an FCFS setting. As depicted in Figure 2, the customers had the option to continuously weigh the risk related to either process jobs locally (on their devices) or forward them to a cloud server. The preference for either local or cloud server was a factor of the latency requirements of the jobs to be processed. These requirements were defined in terms of the expected cloud server response times τ_c . Additionally, the predicted waiting time ($\tau_{w,k}$) of a user landing in position k at joining time also influenced the decision for a processing platform. A reward u that decays with total delay Δt drives decisions: users maximize expected reward while minimizing risk P_0 from varying network conditions. With accurate channel information the model shows a user will not immediately withdraw a submitted task, since predicted remaining latency decreases as earlier jobs are served (see [55], Lemmas 1–2). The paper further studies decision behavior under imperfect information: learned experience and estimator error change perceived risk and can make users either overly cautious or rash. Numerical experiments quantified the regret, learning gain and reward or loss under correct versus partial system knowledge, highlighting how queue-state information quality shapes offload and jockeying choices.

[124]’s work then provided an extended analysis of the value of buffer status information and its applicability in balancing the load within distributed compute systems. Here, the state of a node’s knowledge about prevailing queue setup was fundamental for admitting tasks to the queues and transferring them around to alternative queues. The nodes cooperated on task executions by migrating excess load (total system load minus load on node) to other nodes. Equation (14) (the case of $n \geq 3$ for example) was definitive of the excess load partitioning and distribution to $n - 1$ nodes. The transfer of this excess load depended on the size of the partition such that the least loaded node received the bigger partition.

$$p_{ij} = \begin{cases} \frac{1}{n-2} \left(1 - \frac{\lambda_{d_i}^{-1} Q_i(t - \eta_{ji})}{\sum_{l \neq j} \lambda_{d_l}^{-1} Q_l(t - \eta_{jl})} \right), & \sum_{l \neq j} Q_l(t - \eta_{jl}) > 0 \\ \frac{\lambda_{d_i}}{\sum_{k \neq j} \lambda_{d_k}}, & \text{otherwise,} \end{cases} \quad (14)$$

where η_{jl} was the expected lag when node l and node j communicated, λ_{d_i} was the rate at which a countable number of tasks departed the queue i . Also λ_{d_k} denoted the departure rate at preset values of the k^{th} load balancing instant. And $Q_l(t - \eta_{jl})$ was node j ’s assumption about the number of tasks running on node l which depended on the communication lag not exceeding time t .

The load balancing algorithm resident on each node was executed before accepting any collaboration in the task processing. The deployment of the algorithm followed from a prior broadcast from all nodes about their current buffer sizes as an update to the local state of the knowledge. Two load balancing policies were evaluated, i.e. centralized one-shot and dynamic load balancing. For the centralized one-shots rule-set, the proof evolved from adoption of principles of conditional expectation and regeneration-event decomposition. From these theoretic findings, it was possible to characterize for the average overall completion time (AOCT). The centralized one-shots rule-set was extended for distributed environments as a sender-initiated dynamic load balancing (DLB) policy capable of meeting the dynamic processing speeds of the infrastructure. For the one-shot centralized policy, experiments were setup with two servers with the aim of optimizing the overall time to complete (AOCT) tasks. It was observed that load balancing actions taken when t_b increased beyond one second evolved into the slower node carrying more load. This resulted in larger measures in the AOCT because of delayed updates to the knowledge state. The performance of the DLB policy was evaluated in terms of the time it took to complete a given task within a defined time frame. This metric, therein referred to as the mean task completion time (ACTT) was quantified from the processing, queueing and transfer time. From the analysis of the two qualitative measures (system processing rate (SPR) and ACTT) for both policies under different configurations of the load balancing gain K , the following generalizations were drawn: In either policies, improvements in SPR were recorded under lower measures in K . However, more transfer activity in higher measures of $k \in K$ or excessive load migration delays that led to higher ACTT for the static load balancing policy (one-shot). The DLB policy on the other hand yielded lower queueing transfer delays to reduce the ACTT. Further comparative assessment of the proposed policies against classical DLB policies like Shortest-Expected-Delay (SED) or Never-Queue (NQ) revealed measurable improvements in the ACTT with the DLB over NQ and SED.

In a shift from centralized to decentralized control in multi-tenancy MEC environments, [82] pioneered work in behavioral modeling of impatience in queueing systems. Their decision model sought to understand the benefits switching queues brings to the impatient tenant. The authors propose for decentralized control of the impatient tenant’s behavior. This follows from the argument that current centralized control of this behavior might not be practical in next generation communication systems. That the inherent dynamics in such systems further explode the state space leading to intractable stochastic models. Hence, that the rationale to move workload from one queue to another should be made by the individual tenants after assessing the up-to-date availed information about the expected waiting time. The Monte Carlo experiments assumed a setup of network slices arranged as queues in an $M/M/C$, $C = 2$ configuration. The arrivals that obeyed a Poisson distribution with rate λ and joined

the shorter of the two heterogeneous buffer lines. This was assuming the new tenants were availed with given prior knowledge about queue lengths. The tasks were processed at exponentially distributed service times. Then at each completion of a given task in either queue, tenants evaluated whether to stay in their current position(s) k or to jockey to the alternative buffer line. Here, there was disregard for whether jockeying was to the shorter queue or not. Instead, the rational was premised on the expected waiting time that the jockeyed task would take. However, the position and therefore the effective expected waiting time of the jockey in the alternate queue was a factor of what portion of the new arrivals $\beta \leq \lambda$ would prefer the same queue as the jockey. To manifest this competitiveness therefore, the jockey's final position followed a shuffle operation with the portion β of new arrivals. Assuming $N \geq 1$ departures at an instant and $\beta \leq \lambda$ arrivals seeking to join the preferred queue at that point in time, then switching buffers was only if the expected waiting time in the preferred queue (at position τ) was less than when the tenant stayed put (at position k). (15) was definitive of this behavioral decision following from these sojourn time-position dependencies.

$$F_{T_w|\tau}^{\vec{i}j}(t_w|\tau) = \begin{cases} T_w & | \quad P_{Q_{i,j}}^{\beta}(t+1) \quad \text{if } \beta, N \geq 1 \\ T_w & | \quad P(N \geq 1) \quad \text{if } \beta = 0 \\ T_w & \quad \quad \quad \text{if } \beta, n = 0 \end{cases} \quad (15)$$

where T_w was the expected waiting time and $P_{Q_{i,j}}^{\beta}(t+1)$ denoted the probability that β new arrivals joined the preferred queue at $t+1$ when the jockey decision was to be taken.

Formulations for the number of times that tenants switched from one queue to another then followed from the adoption of principles of conditional probability theory (Baye's theorem). The numerical analysis then compared the empirical measures from the Monte Carlo experiment to these analytical expression given the underlying dependencies. It was observed that tenants that jockeyed more than once ended up waiting less until service in comparison to tenants that did not jockey at all.

Dispatching queue descriptor information as Markov models at varying intervals was investigated in [83] to assess what kind of information or at what interval does this information about queue states lead to optimal system performance. The study involves disseminating two Markov models, one for the service rates of either queue and another Markov model of the how often the queue sizes change. The queued tenant then weighs the decision to jockey or renege based on the comparison of the Markov encapsulations of the service times and queue length dynamics for both queues. A tenant reneges from the queue to prefer the local processing if the estimated remaining waiting time at a given position must be less than the time it takes to process the task locally given that there's no waiting involved. The jockeying tenant compares the statistical distributions of the service times using First-order Stochastic Dominance (FSD). A rule based policy is then devised such that the queue learns from the tenant impatience and adjusts its service rates to minimize the delay and impatience. In their Monte-Carlo setup, the information models are disseminated at intervals of 3,5,7,9 seconds and for each interval is characterized by hundreds of iterations. Numerical results reveal that indeed the interval at which broadcasting takes places has an effect on the eventual impatience of the tenants. The rule-based heuristic yields less volatility with regard to the impatience and sub-optimal service operations.

5.2 Artificial Neural Networks modeling

Although not yet prevalent in jockeying modeling literature, active queue management techniques ideally embed Artificial Intelligence (AI) algorithms to optimize routing and workload distribution. One such study was in Facility location problems (FLP) problems where [23] studied a cooperative multi-layered setup of queues. The authors were interested in dynamically availing the facilities in each layer effectively so as to meet the emerging demand in small and large systems. The demand for service queues followed a Poisson distribution. A job was processed through each layer by those facilities in closest proximity to the job's location. The service rates at

each facility were exponentially distributed and each facility in a given layer participated in the processing of a jockeyed applicant. The optimal solution therefore involved determining the set of facilities that could partake in the servicing of the job request. Starting with a solution for small-scale systems, expressions that characterized the objective functions (reduce jockeying plus mean waiting times and keeping all facilities busy) were formulated. For the small scale scenarios, an augmented ϵ -constraint method was used to validate the global solution to the multi-objective model. And for medium to large scale service systems, an Non-Dominated Sorting Genetic Algorithm (NSGA)-II was deployed. The behavior of each entity in the population and system components like layers or facilities were abstracted as chromosomes. How worth a chromosome was for parenting the next generation was determined by associating that chromosome to penalties defined by $p(x) = U * \text{Max}\{0, \frac{g(x)}{b} - 1\}$ (*such that* U *was a constant, $g(x)$ the constraint and $p(x)$ the punishment on chromosome x*). The Taguchi scheme was found relevant for the adjustment of input parameters used for the initialization of the population. The final mating population evolved from the iterative application of conventional genetic operations (like selection, cross-over, etc) until the fitness function quantitatively yielded no better results for a series of selection runs. The authors concluded the investigations with a numerical analysis of the model by setting up a manufacturing system. Performance comparisons were done when jockeying was permitted versus when no jockeying activity within the layers. Results further gravitated the benefits of jockeying not only at the system level but also at the consumer level. At the system level, the job transfer overhead was minimized plus the time queues were idle. This was because only facilities within the proximity of a potential jockey were those considered to participate in a job's sub-processing. Yet still at the user-level, the jockeying behavior was beneficial in terms of the overall time the user waits before being served.

Limitations of Behavioral Modeling in Next Generation Networks: *These models assume timely, low-cost and accurate queue-state information and stationary service laws—assumptions that break down under mobility, NTN links, strict privacy constraints and heterogeneous, virtualized SDN stacks.*

6 DISCUSSION, CONCLUSION AND FUTURE WORK

The dynamic migration of queued jobs between service pools has shown measurable performance improvements in MEC and heterogeneous 5G/Beyond environments [57, 171]. By leveling demand spikes, it raises average server utilization toward optimality. In the MEC, this is achieved by routing tasks across heterogeneous access points[17, 50, 69], slices and compute tiers to mitigate performance bottlenecks under variable traffic conditions [9, 53]. In end-to-end slice setups, task offloading that employs jockeying behavior reduces processing delays and improves service performance [98, 102]. Numerical studies show that adaptive workload redistribution reduces the mean-sojourn by up to 20 – 30% to accelerate task completion for latency sensitive MEC applications [33, 38, 124]. The benefits extend to real-time load balancing use cases where jockeying heuristics reallocate jobs to minimize buffer growth and packet loss. When combined with network-level state information, they further improve application-specific resource allocation and utilization [101, 145]. Additionally, under SLA regimes, jockeying can prioritize ultra low-latency traffic [105] and improve QoS/QoE when integrated with packet schedulers [43, 168], slice controllers or sub-band selection mechanisms [1, 53, 139, 144]. And reducing customer holding times or aggregate waiting minimizes service costs in energy constrained or SLA-penalized settings such as UAV assisted MEC [42, 53].

Fully realizing these qualitative benefits in next-generation systems demands careful attention to MEC and slice deployment constraints. That is, impatience manifested as jockeying introduces coordination overhead. The behavior can also destabilize scheduling given that it is sensitive to physical layer effects and mobility dynamics. Understanding and mitigating these factors is therefore essential before jockeying can move from a promising technique to a dependable production strategy in heterogeneous, multi-vendor networks. Besides understanding

why workload migration improves general system utilization, the previous sections in addition expose which techniques and how each model abstracts this practice.

However, two higher-level questions motivate reopening this line of inquiry for modern networks: (1) what architectural and technological changes are emerging in communication systems, and (2) are the classical modeling approaches still adequate under these changes, or do new issues arise that require different techniques? In response to these queries, Section 6.1 examines how these architectural trends in next-generation networks challenge traditional jockeying assumptions. We then identify concrete modeling and engineering challenges that arise from these architectural transformations.

6.1 Architectural transformations and the expected Operational challenges

6.1.1 Network Slicing Implications: Third Generation Partnership Project (3GPP) recommends a major evolution in how RAN and Core networks are structured. This change is called the functional split. Proposals hint decoupling the traditional Baseband Unit (BBU) and Remote Radio Unit (RRU) into two flexible components, that is, the Distributed Units (DU) and the Centralized Units (CU). These units have separate paths for the Control Plane and User Plane [95, 123]. The objective is to decompose the network functionality into modular software components and service chains that can be instantiated, scaled and composed by multiple vendors [13, 119]. These modular components, when combined into an end-to-end network slice expose a slice specific set of compute or network resources, control policies and SLA metrics (latency, reliability, cost) [35, 41]. From a queueing theory perspective, each slice (or slice sub-component, e.g., an edge compute pool) can be modeled as a logical service queue (or small queue network) with its own service rate characteristics and pricing. Tenants must then choose the best slice to minimize both cost and latency. Because slices differ in both service parameters and cost schedules, this choice incentivizes impatience behaviors (reneging or jockeying) when slices become transiently adverse to the tenant’s objective[121]. Allowing tenants to pick between vendor slices creates conditions where jockeying is common. This behavior has been found to help improve resource sharing between multiple vendors [140, 156].

6.1.2 Multi-vendor Challenges: Unlike abstract queue networks, real MEC deployments must consider physical topologies and user mobility since MEC nodes span fronthaul and/or backhaul links. Hence, network delays and user movement complicates the optimal selection of target queues. Embedding jockeying policies within network slicing frameworks necessitates joint optimization of compute placement and communication cost. Moreover, designing jockeying-aware schedulers that guarantee latency and reliability SLAs across network slices is critical for future edge applications. In such heterogeneous MEC environments, jockeying becomes a tool for dynamically shifting workloads between slices or queues to minimize perceived delay and operational expenses. However, to realize the full benefits in these environments, we find the following operational challenges that still have to be overcome.

- **Real-time status dissemination:** Broadcasting up-to-date queue lengths or time-of-flight estimates can congest control channels. While timely information aids routing [68, 138, 153], it adds overhead and may require lightweight, event-driven or piggybacked updates. Partial, inaccurate or maliciously altered status data could induce inefficient switching or mimic denial-of-service patterns.
- **Topology and mobility:** Unlike abstract queue networks, MEC deployments must account for fronthaul and/or backhaul latency, user mobility, and multi-hop paths between collaborating nodes [44, 106]. These factors affect task transfer times and can undermine the timeliness of jockeying decisions.
- **Batch versus individual switching:** Bulk migrations of queued jobs may improve throughput but risks transient overloads. Conversely, overly frequent switching at low jockeying thresholds can cause “ping-pong” behavior that wastes resources.

Addressing these issues demands cost-sensitive migration rules, authenticated state dissemination and context-aware policies that jointly optimize queue selection, compute placement, and communication cost under realistic

network constraints. This necessitates quantifying the trade-offs between the performance gains from jockeying and the control or transfer overheads it incurs. The operational challenges identified above, in addition to migration and signaling costs, and cross-domain policy constraints, motivate a pragmatic reappraisal of design choices. In Section 6.2, we synthesize these challenges to guide suggestions for more practical hybrid architectures and propose mitigation techniques as part of the future work.

6.2 Conclusion and Future Work

Third Generation Partnership Project (3GPP) positions Quality of Experience (QoE) as a key metric for 5G and Beyond. In such networks where users expect predictable QoS or QoE, modeling selfish but impatient tenants is essential for slice admission and resource allocation policy design. Existing implementations ignore the need to separate state dissemination from decision control and therefore two design aspects must be distinguished:

- **Centralized dissemination**, involves aggregation and broadcasting of queue metrics (e.g., queue lengths, service rates, waiting times) from a central point to ensure system-wide visibility of queue states [12, 82, 138]. While centralized dissemination offers consistency, it can incur significant control-channel overhead. Conversely, centralized control without fresh visibility of system states risks delayed or suboptimal responses in dynamic MEC environments.
- **Centralized control**, in which a single authority enforces or overrides user decisions (forbids jockeying, enforces admission). Centralized control can be optimal when latency and freshness are guaranteed, but it suffers single point delays, scaling limits and fragility in large, mobile, or NTN deployments. Decentralized decision-making, where each user or local agent autonomously evaluates and acts upon available information, can still function effectively provided that dissemination of state information—centralized or otherwise—is sufficiently timely and accurate. Decentralized decision making scales better but depends critically on the timeliness and fidelity of disseminated state [137].

A practical hybrid architecture separates the roles of publishing and decision making. Queue descriptor metrics (length, mean/or variance of wait, tenant tier, signature etc) are collected by a telemetry aggregator and published through a topic-based publish/subscribe broker. Then the edge agents, such as MEC nodes, base-station proxies, or UE proxies, subscribe based on the QoS requirements. Centrally, the slice orchestrator retains slice-level enforcement (migration budgets, cool-downs, admission rules), while individual jockeying or reneging decisions are decentralized or remain local to meet latency constraints. To reduce the overhead, state could be piggybacked on existing control messages (e.g., Radio Resource Control /N2 signaling between the g-NodeB and control-plane function or lightweight periodic keepalives). Then the updates can be event-triggered such that they are transmitted only when metrics cross meaningful thresholds. This hybrid design minimizes decision latency and avoids single point performance bottlenecks. Centralized dissemination ensures global visibility, decentralized agents act quickly on local conditions, and the slice orchestrator can enforce global constraints. Evaluating such designs requires measuring decision latency, control plane throughput, and the trade-offs or performance gap compared to centralized optimal policies across diverse topologies and mobility patterns.

We advocate for modeling impatience behavior through the Value-of-Information (VoI) approach. In this approach, high-fidelity updates are triggered only when the expected utility gain outweighs the communication cost [60, 166]. A lightweight VoI surrogate can be used to map state, uncertainty, and action costs to the expected utility gain. Bayesian methods, such as ensembles or predictive intervals, provide estimates of information value and determine when refreshes or on-demand requests are necessary. Predictors are expected to deliver both point estimates and calibrated uncertainty. For simple settings, lightweight Markov models are suitable, while in richer contexts neural network ensembles or Bayesian neural networks offer greater accuracy [103, 127, 149]. The learning capacity of these neural models has motivated studies that move beyond statistical quantification of queueing descriptors. Instead, these data-driven methods increasingly make deductive inferences that support

automation processes [11, 127]. Their performance is qualitatively assessed using the Root Mean Square Error (RMSE), calibration error, decision accuracy, and marginal QoE gains per signaling byte.

Incorporating jockeying in next-generation networks requires integration with orchestration frameworks and strong safety controls. Migration decisions must account for service-level improvements relative to migration costs, with slice budgets, cooldown intervals, and fairness schemes to ensure stability [15, 130, 131]. Telemetry and control channels need authenticated communication and safeguards against malicious reporting, while adversarial testing helps assess the resilience of these defenses. While these mechanisms establish the foundation for secure and policy compliant jockeying, the dynamics of unconstrained switching introduce new risks. Excessive jockeying can lead to unnecessary oscillations, wasted resources, and unstable scheduling. Stability can be improved with hysteresis thresholds, cool-down timers, explicit accounting of migration costs such as transfer delay and energy expenditure etc. Evaluating these safeguards requires measuring the number of switches per session, signaling reduction, migration overhead, and worst-case oscillations.

There is need for empirical validation of the above proposals through practical investigations in multi-access edge computing (MEC) testbeds. These testbeds should embed telemetry brokers, MEC decision agents with predictors and value-of-information logic, orchestrator hooks, containerized migration, and radio emulation environments such as srsRAN or OpenAirInterface. And the testbeds should allow for controlled experiments that compare dissemination strategies (periodic, event-driven, or piggyback). The testbeds also need to incorporate mechanisms for assessing the role of value-of-information, uncertainty, and study performance under conditions of high mobility, flash crowds, or adversarial reporting. In addition, complementary large scale simulations, driven by real traffic traces, could help quantify migration latency, control plane overhead, energy costs, latency percentiles, task completion rates, and fairness indices. Then analytical studies, including bounds on minimum update frequency to maintain decision error within certain bounds or large-deviation asymptotics for rare events, provide further theoretical grounding. Together, these mechanisms constitute a key performance indicator suite (covering service quality, signaling overhead, computational footprint, and migration costs) that can guide robust deployment of jockeying in real-world networks.

In conclusion, for systems where Quality of Service (QoS) and Quality of Experience (QoE) determine resource entitlement and pricing, selfish and impatient tenants pose a harder control problem. Their incentives to minimize completion time or maximize utility interact with slice-level policies and pricing, producing complex, feedback-driven behavior that simple rules cannot reliably manage. This chronicle shows that classical jockeying models remain informative but are increasingly inadequate for 5G/6G deployments. Future models must explicitly represent slice heterogeneity and multi-vendor stacks, where non-uniform queue capacities, service disciplines, and admission rules invalidate homogeneous assumptions. They must also treat information flow as a constrained resource: latency, sampling rate, and signaling cost all affect whether a jockeying decision is actionable. Finally, models should include stability and security guarantees so controllers can limit oscillatory switching (ping-pong), bound migration overhead, and resist adversarial or corrupted state reports. Addressing these factors is necessary to produce robust, efficient, and audit-able impatience models suitable for research and system design in next-generation networks.

ACKNOWLEDGMENTS

This work is supported in part by the German Federal Ministry of Education and Research (BMBF) within the project **Open6GHub** under grant numbers 16KISK003K and 16KISK004.

REFERENCES

[1] Md. Abu Baker Siddiki Abir, Mostafa Zaman Chowdhury, and Yeong Min Jang. 2023. A Software-Defined UAV Network Using Queueing Model. *IEEE Access* 11 (2023), 91423–91440. <https://doi.org/10.1109/ACCESS.2023.3281421>

- [2] Mohammad Abu Alsheikh, Dinh Thai Hoang, Dusit Niyato, Hwee-Pink Tan, and Shaowei Lin. 2015. Markov Decision Processes With Applications in Wireless Sensor Networks: A Survey. *IEEE Communications Surveys and Tutorials* 17, 3 (2015), 1239–1267. <https://doi.org/10.1109/COMST.2015.2420686>
- [3] Ivo Adan, J. Wessels, and W. Henk M. Zijm. 1993. Matrix-geometric analysis of the shortest queue problem with threshold jockeying. *Operations Research Letters* 13 (03 1993), 107–112. [https://doi.org/10.1016/0167-6377\(93\)90037-H](https://doi.org/10.1016/0167-6377(93)90037-H)
- [4] Ivo J. B. F. Adan, Wessels Jaap., and W. Henk M. Zijm. 1991. Analysis of the Asymmetric Shortest Queue Problem. *Queueing Syst. Theory Appl.* 8, 1 (2 1991), 1–58. <https://doi.org/10.1007/BF02412240>
- [5] Ivo J. B. F. Adan, Jaap Wessels, and W. Henk M. Zijm. 1991. Analysis of the asymmetric shortest queue problem with threshold jockeying. *Communications in Statistics. Stochastic Models* 7, 4 (1991), 615–627. <https://doi.org/10.1080/15326349108807209>
- [6] J. B. F. Ivo Adan, Wessels Jaap, and W. Henk M. Zijm. 1990. Analysis of the symmetric shortest queue problem. *Communications in Statistics. Part C, Stochastic Models* 6 (1990), 691–713. <https://doi.org/10.1080/15326349908807169>
- [7] Anders Ahlen, Johan Akerberg, Markus Eriksson, Alf J. Isaksson, Takuya Iwaki, Karl Henrik Johansson, Steffi Knorn, Thomas Lindh, and Henrik Sandberg. 2019. Toward Wireless Control in Industrial Process Automation: A Case Study at a Paper Mill. *IEEE Control Systems Magazine* 39, 5 (2019), 36–57. <https://doi.org/10.1109/MCS.2019.2925226>
- [8] Soohan Ahn and V. Ramaswami. 2006. Matrix-geometric algorithms for stochastic fluid flows. In *Proceeding from the 2006 Workshop on Tools for Solving Structured Markov Chains (SMCtools '06)*. Association for Computing Machinery, 11–es. <https://doi.org/10.1145/1190366.1190376>
- [9] Moayad Aloqaily, Venkatraman Balasubramanian, Faisal Zaman, Ismaeel Al Ridhawi, and Yaser Jararweh. 2018. Congestion Mitigation in Densely Crowded Environments for Augmenting QoS in Vehicular Clouds. In *Proceedings of the 8th ACM Symposium on Design and Analysis of Intelligent Vehicular Networks and Applications (DIVANet'18)*. Association for Computing Machinery, 49–56. <https://doi.org/10.1145/3272036.3272038>
- [10] Clinton J. Ancker and A. V. Gafarian. 1963. Some Queueing Problems with Balking and Reneging. II. *Operations Research* 11, 6 (1963), 928–937. <https://doi.org/10.1287/opre.11.6.928>
- [11] Alex Aussem, Antoine Mahul, and Raymond Marie. 2000. Queueing network modelling with distributed neural networks for service quality estimation in B-ISDN networks. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, Vol. 5. IEEE, 392–397.
- [12] A. Aussem, A. Mahul, and R. Marie. 2000. Queueing network modelling with distributed neural networks for service quality estimation in B-ISDN networks. 5 (2000), 392–397 vol.5. <https://doi.org/10.1109/IJCNN.2000.861501>
- [13] Alcardo Alex Barakabitzé, Arslan Ahmad, Rashid Mijumbi, and Andrew Hines. 2020. 5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges. *Computer Networks* 167 (2020), 106984.
- [14] Ward Whitt Benjamin Melamed. 1990. On Arrivals That See Time Averages. *Operations Research* 38 (1990), 156–172. Issue 1. <https://doi.org/10.1287/opre.38.1.156>
- [15] B. Bensaou, D.H.K. Tsang, and King Tung Chan. 2001. Credit-based fair queueing (CBFQ): a simple service-scheduling algorithm for packet-switched networks. *IEEE/ACM Transactions on Networking* 9, 5 (2001), 591–604. <https://doi.org/10.1109/90.958328>
- [16] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. 2011. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems* 24 (2011).
- [17] Jyoti Bisht and Venkata Subrahmanyam Vampugani. 2022. Load and cost-aware min-min workflow scheduling algorithm for heterogeneous resources in fog, cloud, and edge scenarios. *International Journal of Cloud Applications and Computing (IJCAC)* 12, 1 (2022), 1–20.
- [18] Andreas Brandt and Manfred Brandt. 1999. On a Two-Queue Priority System with Impatience and Its Application to a Call Center. *Method. Comput. Appl. Prob.* 1, 2 (9 1999), 191–210. <https://doi.org/10.1023/A:101009304213>
- [19] Alexandre Brandwajn and Thomas Begin. 2019. First-come-first-served queues with multiple servers and customer classes. *Perform. Eval.* 130, C (apr 2019), 51–63. <https://doi.org/10.1016/j.peva.2018.11.001>
- [20] Mark Campbell, Magnus Egerstedt, Jonathan P How, and Richard M Murray. 2010. Autonomous driving in urban environments: approaches, lessons and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 368, 1928 (2010), 4649–4672.
- [21] Delphine Caruelle, Line Lervik-Olsen, and Anders Gustafsson. 2023. The clock is ticking—Or is it? Customer satisfaction response to waiting shorter vs. longer than expected during a service encounter. *Journal of Retailing* 99, 2 (2023), 247–264. <https://doi.org/10.1016/j.jretai.2023.03.003>
- [22] Sofia Ceppi, Nicola Gatti, Giorgio Patrini, and Marco Rocco. 2010. Local Search Methods for Finding a Nash Equilibrium in Two-Player Games. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, Vol. 2. 335–342. <https://doi.org/10.1109/WI-IAT.2010.57>
- [23] Amir Eshaghi Chaleshtori, Hamed Jahani, and Abdollah Aghaie. 2020. Bi-objective optimization approach to a multi-layer location-allocation problem with jockeying. *Computers and Industrial Engineering* 149 (2020), 106740. <https://doi.org/10.1016/j.cie.2020.106740>

- [24] Gautam Choudhury and Kailash C Madan. 2004. A two phase batch arrival queueing system with a vacation time under Bernoulli schedule. *Appl. Math. Comput.* 149, 2 (2004), 337–349.
- [25] T. Collings and C. Stoneman. 1976. The $M/M/\infty$ Queue with Varying Arrival and Departure Rates. *Operations Research* 24, 4 (1976), 760–773.
- [26] Brian W. Conolly. 1984. “The Autostrada Queueing Problem.” *Journal of Applied Probability*. *JSTOR* 21, no. 2 (1984), 394–403. <https://doi.org/10.2307/3213648>
- [27] Jim G. Dai. 1995. On Positive Harris Recurrence of Multiclass Queueing Networks: A Unified Approach Via Fluid Limit Models. *The Annals of Applied Probability* 5, 1 (1995), 49–77.
- [28] Constantinos Daskalakis, Paul Goldberg, and Christos Papadimitriou. 2006. The complexity of computing a Nash equilibrium. *SIAM J. Comput.* 39, 71–78. <https://doi.org/10.1145/1461928.1461951>
- [29] Amin Dehghanian, Jeffrey P. Kharoufeh, and Mohammad Modarres. 2016. Strategic dynamic jockeying between two parallel queues. *Probability in the Engineering and Informational Sciences* 30, 1 (2016), 41–60. <https://doi.org/10.1017/S0269964815000273>
- [30] Rosario Delgado and Evsey Morozov. 2014. Stability analysis of cascade networks via fluid models. *Performance Evaluation* 82 (2014), 39–54. <https://doi.org/10.1016/j.peva.2014.10.001>
- [31] Rosario Delgado and Evsey V. Morozov. 2014. Stability Analysis of Some Networks with Interacting Servers. *International Conference on Analytical and Stochastic Modeling Techniques and Applications* ASMTA 2014 (2014), 1–15. https://doi.org/10.1007/978-3-319-08219-6_1
- [32] William E. DISNEY, Ralph L; MITCHELL. 1970. A Solution for Queues with Instantaneous Jockeying and Other Customer Selection Rules. *Naval Research Logistics* 17 (9 1970), 315–325. <https://doi.org/10.1002/nav.3800170308>
- [33] Douglas G. Down and Mark E. Lewis. 2006. Dynamic load balancing in parallel queueing systems: Stability and optimal control. *European Journal of Operational Research* 168, 2 (2006), 509–519. <https://doi.org/10.1016/j.ejor.2004.04.041> Feature Cluster on Mathematical Finance and Risk Management.
- [34] Salah Eddine Elayoubi, Sana Ben Jemaa, Zwi Altman, and Ana Galindo-Serrano. 2019. 5G RAN Slicing for Verticals: Enablers and Challenges. *IEEE Communications Magazine* 57, 1 (2019), 28–34. <https://doi.org/10.1109/MCOM.2018.1701319>
- [35] Salah Eddine Elayoubi, Sana Ben Jemaa, Zwi Altman, and Ana Galindo-Serrano. 2019. 5G RAN slicing for verticals: Enablers and challenges. *IEEE Communications Magazine* 57, 1 (2019), 28–34.
- [36] Elsayed A. Elsayed and Ali S. Bastani. 1985. General solutions of the jockeying problem. *European Journal of Operational Research* 22, 3 (1985), 387–396. [https://doi.org/10.1016/0377-2217\(85\)90258-9](https://doi.org/10.1016/0377-2217(85)90258-9)
- [37] Anwar I. Elwalid and Debasis Mitra. 1991. Analysis and design of rate-based congestion control of high speed networks, I: stochastic fluid models, access regulation. *Queueing Systems* 9, 1 (03 1991), 29–63. <https://doi.org/10.1007/BF01158791>
- [38] D.J Evans and W.U.N Butt. 1993. Dynamic load balancing using task-transfer probabilities. *Parallel Comput.* 19, 8 (1993), 897–916. [https://doi.org/10.1016/0167-8191\(93\)90073-T](https://doi.org/10.1016/0167-8191(93)90073-T)
- [39] Huei-Wen Ferng. 2001. Departure processes of BMAP/G/1 queues. *Queueing Systems* 39 (10 2001), 109–135. <https://doi.org/10.1023/A:1012786932415>
- [40] Robert D. Foley and David R. McDonald. 2001. Join the Shortest Queue: Stability and Exact Asymptotics. *The Annals of Applied Probability* 11, 3 (2001), 569–607.
- [41] X. Foukas, A. Elmokashfi, G Patoureas, and MK Marina. 2017. Network Slicing in 5G: Survey and Challenges. *IEEE Communications Magazine* 55, 5 (2017), 94–100. <https://doi.org/10.1109/MCOM.2017.1600951>
- [42] Jing Fu, Jun Guo, Eric W. M. Wong, and Moshe Zukerman. 2015. Energy-efficient heuristics for job assignment in processor-sharing server farms. In *2015 IEEE Conference on Computer Communications (INFOCOM)*. 882–890. <https://doi.org/10.1109/INFOCOM.2015.7218459>
- [43] Tom Z.J. Fu, Jianbing Ding, Richard T.B. Ma, Marianne Winslett, Yin Yang, and Zhenjie Zhang. 2015. DRS: Dynamic Resource Scheduling for Real-Time Analytics over Fast Streams. (2015), 411–420. <https://doi.org/10.1109/ICDCS.2015.49>
- [44] R. Gaeta, M. Gribaudo, D. Manini, and M. Sereno. 2006. Analysis of resource transfers in peer-to-peer file sharing applications using fluid models. *Performance Evaluation* 63, 3 (2006), 149–174. <https://doi.org/10.1016/j.peva.2005.01.001> P2P Computing Systems.
- [45] David Garcia-Roger, Edgar E González, David Martín-Sacristán, and Jose F Monserrat. 2020. V2X support in 3GPP specifications: From 4G to 5G and beyond. *IEEE access* 8 (2020), 190946–190963.
- [46] Ilya B. Gertsbakh. 1984. The shorter queue problem: A numerical study using the matrix-geometric solution. *European Journal of Operational Research* 15 (1984), 374–381.
- [47] Fablette Gillent and Guy Latouche. 1983. Semi-explicit solutions for M/PH/1-like queueing systems. *European Journal of Operational Research* 13, 2 (1983), 151–160. [https://doi.org/10.1016/0377-2217\(83\)90077-2](https://doi.org/10.1016/0377-2217(83)90077-2)
- [48] Judy Goldsmith and Martin Mundhenk. 1998. Complexity Issues in Markov Decision Processes.. In CCC. 272–280.
- [49] Harold Gumbel. 1960. Waiting Lines with Heterogeneous Servers. *Operations Research* 8, 4 (1960), 504–511.
- [50] Mian Guo, Quansheng Guan, Weiqi Chen, Fei Ji, and Zhiping Peng. 2022. Delay-Optimal Scheduling of VMs in a Queueing Cloud Computing System with Heterogeneous Workloads. *IEEE Transactions on Services Computing* 15, 1 (2022), 110–123. <https://doi.org/10.1109/TSC.2019.2920954>
- [51] Frank A. Haight. 1958. Two Queues in Parallel. *Biometrika* 45, 3/4 (1958), 401–410.

- [52] Frank A. Haight. 1959. Queueing with reneging. *Metrika* 2, 1 (12 1959), 186–197. <https://doi.org/10.1007/BF02613734>
- [53] Bin Han, Vincenzo Sciancalepore, Xavier Costa-Pérez, Di Feng, and Hans D. Schotten. 2020. Multiservice-Based Network Slicing Orchestration With Impatient Tenants. *IEEE Transactions on Wireless Communications* 19, 7 (2020), 5010–5024. <https://doi.org/10.1109/TWC.2020.2988644>
- [54] Bin Han, Vincenzo Sciancalepore, Di Feng, Xavier Costa-Perez, and Hans D. Schotten. 2019. A Utility-Driven Multi-Queue Admission Control Solution for Network Slicing. (2019), 55–63. <https://doi.org/10.1109/INFOCOM.2019.8737517>
- [55] Bin Han, Vincenzo Sciancalepore, Yihua Xu, Di Feng, and Hans D. Schotten. 2023. Impatient Queueing for Intelligent Task Offloading in Multiaccess Edge Computing. *IEEE Transactions on Wireless Communications* 22, 1 (2023), 59–72. <https://doi.org/10.1109/TWC.2022.3191287>
- [56] Bin Han, Vincenzo Sciancalepore, Yihua Xu, Di Feng, and Hans D. Schotten. 2023. Impatient Queueing for Intelligent Task Offloading in Multiaccess Edge Computing. *IEEE Transactions on Wireless Communications* 22, 1 (2023), 59–72. <https://doi.org/10.1109/TWC.2022.3191287>
- [57] Yinghua Han, Qinjin Xu, Qiang Zhao, and Fangyuan Si. 2023. Queue-aware computation offloading for UAV-assisted edge computing in wind farm routine inspection. *Journal of Renewable and Sustainable Energy* 15, 6 (2023).
- [58] Lani Haque, Yiqiang Q. Zhao, and Liming Liu. 2005. Sufficient conditions for a geometric tail in a QBD process with many countable levels and phases. *Stochastic Models* 21, 1 (2005), 77–99. <https://doi.org/10.1081/STM-200046489>
- [59] Mor Harchol-Balter. 2010. *Performance modeling and design of computer systems*. Imperial college press.
- [60] Refael Hassin and Moshe Haviv. 1994. Equilibrium strategies and the value of information in a two line queueing system with threshold jockeying. *Stochastic Models - STOCH MODELS* 10 (01 1994), 415–435. <https://doi.org/10.1080/15326349408807302>
- [61] Refael Hassin and Moshe Haviv. 1997. Equilibrium Threshold Strategies: The Case of Queues with Priorities. *Operations Research* 45, 6 (1997), 966–973.
- [62] B.R. Haverkort. 1995. Matrix-geometric solution of infinite stochastic Petri nets. In *Proceedings of 1995 IEEE International Computer Performance and Dependability Symposium*. 72–81. <https://doi.org/10.1109/IPDS.1995.395815>
- [63] Moshe Haviv and Liron Ravner. 2015. Strategic bidding in an accumulating priority queue: equilibrium analysis. *Annals of Operations Research* 244 (2015), 505–523.
- [64] Qiming He and Marcel F. Neuts. 2002. Two M/M/1 Queues with Transfers of Customers. *Queueing Systems* 42 (2002), 377–400.
- [65] Qi-Ming He. 2013. Analysis of queueing systems with customer interjections. *Queueing Systems* 73 (01 2013), 79–104.
- [66] Samid Hoda, Andrew Gilpin, Javier Peña, and Tuomas Sandholm. 2010. Smoothing Techniques for Computing Nash Equilibria of Sequential Games. *Mathematics of Operations Research* 35, 2 (2010), 494–512.
- [67] Longbo Huang. 2017. The Value-of-Information in Matching With Queues. *IEEE/ACM Transactions on Networking* 25, 1 (2017), 29–42. <https://doi.org/10.1109/TNET.2016.2564700>
- [68] Longbo Huang. 2017. The Value-of-Information in Matching With Queues. *IEEE/ACM Transactions on Networking* 25, 1 (2017), 29–42. <https://doi.org/10.1109/TNET.2016.2564700>
- [69] R. Jain. 1986. A Timeout-Based Congestion Control Scheme for Window Flow-Controlled Networks. *IEEE Journal on Selected Areas in Communications* 4, 7 (1986), 1162–1167. <https://doi.org/10.1109/JSAC.1986.1146431>
- [70] Xingguo Ji and Qingmin Meng. 2020. Traffic Classification Based on Graph Convolutional Network. In *2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications (AEECA)*. 596–601. <https://doi.org/10.1109/AEECA49918.2020.9213630>
- [71] Jiekai Jia, Anam Tahir, and Heinz Koepll. 2022. Decentralized Coordination in Partially Observable Queueing Networks. In *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*. 1491–1496. <https://doi.org/10.1109/GLOBECOM48099.2022.10001584>
- [72] Tao Jiang and Li-Wei Liu. 2016. Tail Asymptotics of Two Parallel Queues with Transfers of Customers. *Journal of the Operations Research Society of China* 4 (04 2016). <https://doi.org/10.1007/s40305-016-0127-1>
- [73] Wei Jiang, Bin Han, Mohammad Asif Habibi, and Hans Dieter Schotten. 2021. The Road Towards 6G: A Comprehensive Survey. *IEEE Open Journal of the Communications Society* 2 (2021), 334–366. <https://doi.org/10.1109/OJCOMS.2021.3057679>
- [74] J.S. Jordan. 1993. Three Problems in Learning Mixed-Strategy Nash Equilibria. *Games and Economic Behavior* 5, 3 (1993), 368–386. <https://doi.org/10.1006/game.1993.1022>
- [75] Michella H. Kallmes, D. Towsley, and Christos G. Cassandras. 1989. Optimality of the last-in-first-out (LIFO) service discipline in queueing systems with real-time constraints. *Proceedings of the 28th IEEE Conference on Decision and Control*, 2, 0 (1989), 1073–1074. <https://doi.org/10.1109/CDC.1989.70296>
- [76] Clement Kam, Joseph P. Molnar, and Sastry Kompella. 2018. Age of Information for Queues in Tandem. In *MILCOM 2018 - 2018 IEEE Military Communications Conference (MILCOM)*. 1–6. <https://doi.org/10.1109/MILCOM.2018.8599728>
- [77] Edward P.C. Kao and Chiunsin Lin. 1990. A matrix-geometric solution of the jockeying problem. *European Journal of Operational Research* 44, 1 (1990), 67–74. [https://doi.org/10.1016/0377-2217\(90\)90315-3](https://doi.org/10.1016/0377-2217(90)90315-3)
- [78] Stella Kapodistria and Zbigniew Palmowski. 2017. Matrix geometric approach for random walks: Stability condition and equilibrium distribution. *Stochastic Models* 33, 4 (2017), 572–597. <https://doi.org/10.1080/15326349.2017.1359096>

[79] Rajeeva L. Karandikar and Vidyadhar G. Kulkarni. 1995. Second-Order Fluid Flow Models: Reflected Brownian Motion in a Random Environment. *Operations Research* 43, 1 (1995), 77–88.

[80] David G. Kendall. 1953. Stochastic Processes Occurring in the Theory of Queues and their Analysis by the Method of the Imbedded Markov Chain. *The Annals of Mathematical Statistics* 24, 3 (1953), 338–354.

[81] Hisham A. Kholidy, Andrew Karam, James Sidoran, Mohammad A. Rahman, Mohammed Mahmoud, Mahmoud Badr, Maqsood Mahmud, and Ahmed F. Sayed. 2022. Toward Zero Trust Security IN 5G Open Architecture Network Slices. In *MILCOM 2022 - 2022 IEEE Military Communications Conference (MILCOM)*. 577–582. <https://doi.org/10.1109/MILCOM55135.2022.10017474>

[82] Anthony Kiggundu, Bin Han, Dennis Krummacker, and Hans D. Schotten. 2024. Resource Allocation in Mobile Networks: A Decision Model Of Jockeying in Queues. In *2024 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*. 311–316. <https://doi.org/10.1109/MeditCom61057.2024.10621432>

[83] Anthony Kiggundu, Bin Han, and Hans D. Schotten. 2025. Information Bulletin Strategy in Impatient Queuing. In *2025 IEEE Conference on Standards for Communication Networks (CSCN)*.

[84] Bara Kim, Jeongsim Kim, and Ole Bueker. 2023. Equilibrium analysis of a partially observable priority queue. *Computers and Industrial Engineering* 182 (2023), 109434. <https://doi.org/10.1016/j.cie.2023.109434>

[85] Minsu Kim, Muhammad Jaseemuddin, and Alagan Anpalagan. 2021. Deep Reinforcement Learning Based Active Queue Management for IoT Networks. *J. Netw. Syst. Manage.* 29, 3 (jul 2021), 28 pages. <https://doi.org/10.1007/s10922-021-09603-x>

[86] John Frank C. Kingman. 1961. Two Similar Queues in Parallel. *The Annals of Mathematical Statistics* 32, 4 (1961), 1314–1323.

[87] John Frank C. Kingman. 1962. On queues in which customers are served in random order. *Mathematical Proceedings of the Cambridge Philosophical Society* 58 (1962), 79 – 91.

[88] Leonard Kleinrock. 1975. *Queueing Systems. Volume I: Theory*. John Wiley and Sons, New York.

[89] Leonard Kleinrock. 1976. *Queueing Systems. Volume II: Computer Applications*. John Wiley & Sons, New York.

[90] Ernest Koenigsberg. 1966. On Jockeying in Queues. *Management Science* 12 (1966), 412–436.

[91] Yaakov Kogan, Yonatan Levy, and Rodolfo A. Milito. 1997. Call routing to distributed queues: Is FIFO really better than MED? *Telecommunication Systems* 7 (06 1997), 299–312.

[92] Ger Koole. 1996. On the Pathwise Optimal Bernoulli Routing Policy for Homogeneous Parallel Servers. *Mathematics of Operations Research* 21, 2 (1996), 469–476.

[93] V. Kumar and N.S. Upadhye. 2022. On first-come, first-served queues with three classes of impatient customers. *International Journal of Advances in Engineering Sciences and Applied Mathematics* 13 (03 2022), 368–382.

[94] Corine M. Laan, Judith Timmer, and Richard J. Boucherie. 2021. Non-cooperative queueing games on a network of single server queues. *Queueing Systems* 97, 3 (4 2021), 279–301. <https://doi.org/10.1007/s11134-020-09681-9>

[95] Line M. P. Larsen, Aleksandra Checko, and Henrik L. Christiansen. 2019. A Survey of the Functional Splits Proposed for 5G Mobile Crosshaul Networks. *IEEE Communications Surveys and Tutorials* 21, 1 (2019), 146–172. <https://doi.org/10.1109/COMST.2018.2868805>

[96] Toni Lassila, Andrea Manzoni, Alfio Quarteroni, and Gianluigi Rozza. 2014. *Model Order Reduction in Fluid Dynamics: Challenges and Perspectives*. Springer International Publishing, 235–273. https://doi.org/10.1007/978-3-319-02090-7_9

[97] Guy Latouche and V. Ramaswami. 1987. *Introduction to Matrix Analytic Methods in Stochastic Modelling*. Society for Industrial and Applied Mathematics, Philadelphia.

[98] Lei Li, Quansheng Guan, Lianwen Jin, and Mian Guo. 2019. Resource Allocation and Task Offloading for Heterogeneous Real-Time Tasks With Uncertain Duration Time in a Fog Queueing System. *IEEE Access* 7 (2019), 9912–9925. <https://doi.org/10.1109/ACCESS.2019.2891130>

[99] Vlada Limic. 2001. A LIFO Queue in Heavy Traffic. *The Annals of Applied Probability* 11, 2 (2001), 301–331.

[100] Bing Lin, Yuchen Lin, and Rohit Bhatnagar. 2022. Optimal Policy for Controlling Two-Server Queueing Systems with Jockeying. *Journal of Systems Engineering and Electronics* 33, 1 (2022), 144–155. <https://doi.org/10.23919/JSEE.2022.000015>

[101] Frank Po-Chen Lin and Zsehong Tsai. 2020. Hierarchical Edge-Cloud SDN Controller System With Optimal Adaptive Resource Allocation for Load-Balancing. *IEEE Systems Journal* 14, 1 (2020), 265–276. <https://doi.org/10.1109/JSYST.2019.2894689>

[102] Chen-Feng Liu, Mehdi Bennis, and H. Vincent Poor. 2017. Latency and Reliability-Aware Task Offloading and Resource Allocation for Mobile Edge Computing. (2017), 1–7. <https://doi.org/10.1109/GLOCOMW.2017.8269175>

[103] Yuanhe Liu and Ruiming Quan. 2023. Research and Analysis on the Interaction between Queueing Theory and Artificial Intelligence. (2023), 391–401. https://doi.org/10.2991/978-94-6463-300-9_40

[104] Aleksander Lodwich, Yves Rangoni, and Thomas Breuel. 2009. Evaluation of robustness and performance of Early Stopping Rules with Multi Layer Perceptrons. In *2009 International Joint Conference on Neural Networks*. 1877–1884. <https://doi.org/10.1109/IJCNN.2009.5178626>

[105] Lingling Lv, Zongyu Wu, Lei Zhang, Brij B. Gupta, and Zhihong Tian. 2022. An Edge-AI Based Forecasting Approach for Improving Smart Microgrid Efficiency. *IEEE Transactions on Industrial Informatics* 18, 11 (2022), 7946–7954. <https://doi.org/10.1109/TII.2022.3163137>

[106] Mamoru Maekawa. 1977. Queueing Models for Computer Systems Connected by a Communication Line. *J. ACM* 24, 4 (oct 1977), 566–582.

- [107] Simon P. Martin and Isi Mitrani. 2008. Analysis of job transfer policies in systems with unreliable servers. *Annals of Operations Research* 162 (2008), 127–141.
- [108] Sean P. Meyn and Richard L. Tweedie. 1993. Stability of Markovian Processes III: Foster-Lyapunov Criteria for Continuous-Time Processes. *Advances in Applied Probability* 25, 3 (1993), 518–548.
- [109] Isi Mitrani and Ram Chakka. 1995. Spectral expansion solution for a class of Markov models: application and comparison with the matrix-geometric method. *Performance Evaluation* 23, 3 (1995), 241–260. [https://doi.org/10.1016/0166-5316\(94\)00025-F](https://doi.org/10.1016/0166-5316(94)00025-F)
- [110] Banwari Mittal. 2016. Retrospective: why do customers switch? The dynamics of satisfaction versus loyalty. *Journal of Services Marketing* 30 (09 2016), 569–575. <https://doi.org/10.1108/JSM-07-2016-0277>
- [111] Masakiyo Miyazawa and Yiqiang Q. Zhao. 2004. The Stationary Tail Asymptotics in the GI/G/1-Type Queue with Countably Many Background States. *Advances in Applied Probability* 36, 4 (2004), 1231–1251.
- [112] George E. Monahan. 1982. State of the Art—A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algorithms. *Management Science* 28, 1 (1982), 1–16. <https://doi.org/10.1287/mnsc.28.1.1>
- [113] John F. Nash. 1950. Equilibrium Points in n-Person Games. *Proceedings of the National Academy of Science* 36, 1 (1 1950), 48–49. <https://doi.org/10.1073/pnas.36.1.48>
- [114] M. Natarajan and A. Kolobov. [n. d.]. *Planning with Markov Decision Processes: An AI Perspective*. Springer International Publishing.
- [115] Marcel F. Neuts. 1986. The caudal characteristic curve of queues. *Advances in Applied Probability* 18, 1 (1986), 221–254. <https://doi.org/10.2307/1427244>
- [116] Marcel Fernand Neuts. 1994. *Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach*. Dover Publications, Newyork.
- [117] David M. Nicol. 1988. Parallel discrete-event simulation of FCFS stochastic queueing networks. *SIGPLAN Not.* 23, 9 (jan 1988), 124–137. <https://doi.org/10.1145/62116.62128>
- [118] Jose Ordóñez-Lucena, Pablo Ameigeiras, Diego Lopez, Juan J. Ramos-Munoz, Javier Lorca, and Jesus Folgueira. 2017. Network Slicing for 5G with SDN/NFV: Concepts, Architectures, and Challenges. *IEEE Communications Magazine* 55, 5 (2017), 80–87. <https://doi.org/10.1109/MCOM.2017.1600935>
- [119] Jose Ordóñez-Lucena, Pablo Ameigeiras, Diego Lopez, Juan J Ramos-Munoz, Javier Lorca, and Jesus Folgueira. 2017. Network slicing for 5G with SDN/NFV: Concepts, architectures, and challenges. *IEEE Communications Magazine* 55, 5 (2017), 80–87.
- [120] Joni Pajarinen, Ari Hottinen, and Jaakko Peltonen. 2013. Optimizing spatial and temporal reuse in wireless networks by decentralized partially observable Markov decision processes. *IEEE Transactions on Mobile Computing* 13, 4 (2013), 866–879.
- [121] Apostolos Papageorgiou, Adriana Fernández-Fernández, Leonardo Ochoa-Aday, Miguel Silva Peláez, and Muhammad Shuaib Siddiqui. 2020. SLA Management Procedures in 5G Slicing-based Systems. (2020), 7–11. <https://doi.org/10.1109/EuCNC48522.2020.9200904>
- [122] Amit I. Pazgal and Sonja Radas. 2008. Comparison of customer balking and reneging behavior to queueing theory predictions: An experimental study. *Comput. Oper. Res.* 35, 8 (2008), 2537–2548.
- [123] Jordi Pérez-Romero, Oriol Sallent, Antoni Gelonch, Xavier Gelabert, Bleron Klaiqi, Marcus Kahn, and David Campoy. 2023. A Tutorial on the Characterisation and Modelling of Low Layer Functional Splits for Flexible Radio Access Networks in 5G and Beyond. *IEEE Communications Surveys and Tutorials* 25, 4 (2023), 2791–2833. <https://doi.org/10.1109/COMST.2023.3296821>
- [124] J. E. Pezoa, M. M. Hayat, D. A. Bader, C. Yang, and S. Dhakal. 2007. Dynamic Load Balancing in Distributed Systems in the Presence of Delays: A Regeneration-Theory Approach. *IEEE Transactions on Parallel and Distributed Systems* 18, 04 (04 2007), 485–497. <https://doi.org/10.1109/TPDS.2007.1009>
- [125] S.D. Poisson and S.D. Poisson. 1837. *Recherches sur la probabilité des jugements en matière criminelle et en matière civile: précédées des règles générales du calcul des probabilités*. Bachelier, online.
- [126] Lutz Prechelt. 1998. Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks* 11, 4 (1998), 761–767. [https://doi.org/10.1016/S0893-6080\(98\)00010-0](https://doi.org/10.1016/S0893-6080(98)00010-0)
- [127] Rezaur Rahman and Samiul Hasan. 2021. Real-time signal queue length prediction using long short-term memory neural network. *Neural Comput. Appl.* 33, 8 (apr 2021), 3311–3324.
- [128] V. Ramswami and Guy Latouche. 1986. A general class of Markov processes with explicit matrix-geometric solutions. *Operations-Research-Spektrum* 8 (1986), 209–218.
- [129] Rachel Ravid. 2021. A new look on the shortest queue system with jockeying. *Probability in the Engineering and Informational Sciences* 35, 3 (2021), 557–564. <https://doi.org/10.1017/S0269964819000469>
- [130] David Raz, Benjamin Avi-Itzhak, and Hanoch Levy. 2006. Fairness considerations of scheduling in multi-server and multi-queue systems. In *Proceedings of the 1st International Conference on Performance Evaluation Methodologies and Tools (valuetools '06)*. Association for Computing Machinery, New York, NY, USA, 39–es. <https://doi.org/10.1145/1190095.1190145>
- [131] David Raz, Hanoch Levy, and Benjamin Avi-Itzhak. 2004. A resource-allocation queueing fairness measure. *SIGMETRICS Perform. Eval. Rev.* 32, 1 (jun 2004), 130–141. <https://doi.org/10.1145/1012888.1005704>
- [132] Zvi Rosberg, Yu Peng, Jing Fu, Jun Guo, Eric W. M. Wong, and Moshe Zukerman. 2014. Insensitive Job Assignment With Throughput and Energy Criteria for Processor-Sharing Server Farms. *IEEE/ACM Transactions on Networking* 22, 4 (2014), 1257–1270. <https://doi.org/10.1109/TNET.2013.2276427>

- [133] Werner Scheinhardt, Nicky van Foreest, and Michel Mandjes. 2005. Continuous feedback fluid queues. *Operations Research Letters* 33, 6 (2005), 551–559. <https://doi.org/10.1016/j.orl.2004.11.008>
- [134] William T. Scherer, Stephen Adams, and Peter A. Beling. 2018. On the Practical Art of State Definitions for Markov Decision Process Construction. *IEEE Access* 6 (2018), 21115–21128. <https://doi.org/10.1109/ACCESS.2018.2819940>
- [135] Dimitrios Schinianakis, Ruben Trapero, Diomidis S. Michalopoulos, and Beatriz Gallego-Nicasio Crespo. 2019. Security Considerations in 5G Networks: A Slice-Aware Trust Zone Approach. In *2019 IEEE Wireless Communications and Networking Conference (WCNC)*. 1–8. <https://doi.org/10.1109/WCNC.2019.8885658>
- [136] Vincenzo Sciancalepore, Konstantinos Samdanis, Xavier Costa-Perez, Dario Bega, Marco Gramaglia, and Albert Banchs. 2017. Mobile traffic forecasting for maximizing 5G network slicing resource utilization. (2017), 1–9. <https://doi.org/10.1109/INFOCOM.2017.8057230>
- [137] Flore Sentenac, Etienne Boursier, and Vianney Perchet. 2021. Decentralized Learning in Online Queuing Systems. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 18501–18512.
- [138] Hsien-Po Shiang and Mihaela van der Schaar. 2008. Queuing-Based Dynamic Channel Selection for Heterogeneous Multimedia Applications Over Cognitive Radio Networks. *IEEE Transactions on Multimedia* 10, 5 (2008), 896–909. <https://doi.org/10.1109/TMM.2008.922851>
- [139] Hsien-Po Shiang and Mihaela van der Schaar. 2008. Queuing-Based Dynamic Channel Selection for Heterogeneous Multimedia Applications Over Cognitive Radio Networks. *IEEE Transactions on Multimedia* 10, 5 (2008), 896–909. <https://doi.org/10.1109/TMM.2008.922851>
- [140] Zhaogang Shu and Tarik Taleb. 2020. A novel QoS framework for network slicing in 5G and beyond networks based on SDN and NFV. *IEEE Network* 34, 3 (2020), 256–263.
- [141] Wolfgang Stadje. 2009. A Queueing System with Two Parallel Lines, Cost-Conscious Customers, and Jockeying. *Communications in Statistics - Theory and Methods* 38:16–17 (2009), 3158–3169. <https://doi.org/10.1080/03610920902947618>
- [142] Robert E. Stanford. 1979. Reneging Phenomena in Single Channel Queues. *Mathematics of Operations Research* 4, 2 (05 1979), 162–178.
- [143] Oded Stark, Wiktor Budzinski, and Grzegorz Kosiorowski. 2019. Switching queues, cultural conventions, and social welfare. *European Journal of Operational Research* 278, 3 (2019), 837–844. <https://doi.org/10.1016/j.ejor.2019.02.053>
- [144] René Brandborg Sørensen, Dong Min Kim, Jimmy Jessen Nielsen, and Petar Popovski. 2017. Analysis of Latency and MAC-Layer Performance for Class A LoRaWAN. *IEEE Wireless Communications Letters* 6, 5 (2017), 566–569. <https://doi.org/10.1109/LWC.2017.2716932>
- [145] Anam Tahir, Bastian Alt, Amr Rizk, and Heinz Koeppl. 2023. Load Balancing in Compute Clusters With Delayed Feedback. *IEEE Trans. Comput.* 72, 6 (jun 2023), 1610–1622. <https://doi.org/10.1109/TC.2022.3215907>
- [146] Hisashi Takagi. 1991. *Queueing Analysis: A Foundation of Performance Evaluation, Part 1: Vacation and Priority Systems*. North-Holland, Amsterdam.
- [147] Hisashi Takagi. 1993. *Queueing Analysis: A Foundation of Performance Evaluation, Part 2: Finite Systems*. North-Holland, Amsterdam.
- [148] Hisashi Takagi. 1993. *Queueing Analysis: A Foundation of Performance Evaluation, Part 3: Discrete-Time Systems*. North-Holland, Amsterdam.
- [149] Junichi Takinami, Yutaka Matsumoto, and Norio Okino. 1993. Performance Evaluation of Neural Networks Applied to Queueing Allocation Problem. In *Artificial Neural Nets and Genetic Algorithms*. Springer Vienna, 316–323.
- [150] Ahmed M.K. Tarabia. 2009. Transient analysis of two queues in parallel with jockeying. *Stochastics* 81, 2 (2009), 129–145. <https://doi.org/10.1080/17442500802200658>
- [151] Ahmed M. K. Tarabia. 2008. Analysis of two queues in parallel with jockeying and restricted capacities. *Applied Mathematical Modelling* 32 (2008), 802–810.
- [152] Rafael Tolosana-Calasanz, Javier Diaz-Montes, Omer F. Rana, and Manish Parashar. 2017. Feedback-Control & Queueing Theory-Based Resource Management for Streaming Applications. *IEEE Trans. Parallel Distrib. Syst.* 28, 4 (April 2017), 1061–1075. <https://doi.org/10.1109/TPDS.2016.2603510>
- [153] Rafael Tolosana-Calasanz, Javier Diaz-Montes, Omer F. Rana, and Manish Parashar. 2017. Feedback-Control and Queueing Theory-Based Resource Management for Streaming Applications. *IEEE Transactions on Parallel and Distributed Systems* 28, 4 (2017), 1061–1075. <https://doi.org/10.1109/TPDS.2016.2603510>
- [154] Montanaro Umberto, Dixit Shilp, Fallah Saber, Dianati Mehrdad, Stevens Alan, Oxtoby David, and Alexandros Mouzakitis. 2019. Towards connected autonomous driving: review of use-cases. *Vehicle System Dynamics* 57, 6 (2019), 779–814. <https://doi.org/10.1080/00423114.2018.1492142>
- [155] James S. Vandergraft. 1983. A Fluid Flow Model of Networks of Queues. *Management Science* 29, 10 (1983), 1198–1208.
- [156] Qi Wang, Jose Alcaraz-Calero, Ruben Ricart-Sanchez, Maria Barros Weiss, Anastasios Gavras, Navid Nikaein, Xenofon Vasilakos, Bernini Giacomo, Giardina Pietro, Mark Roddy, et al. 2019. Enable advanced QoS-aware network slicing in 5G networks for slice-based media use cases. *IEEE transactions on broadcasting* 65, 2 (2019), 444–453.
- [157] Ward Whitt. 1984. Departures from a Queue with Many Busy Servers. *Mathematics of Operations Research* 9, 4 (1984), 534–544.

- [158] Ward Whitt. 1986. Deciding Which Queue to Join: Some Counterexamples. *Operations Research* 34, 1 (1986), 55–62.
- [159] Martin Wollschlaeger, Thilo Sauter, and Juergen Jasperneite. 2017. The future of industrial communication: Automation networks in the era of the internet of things and industry 4.0. *IEEE industrial electronics magazine* 11, 1 (2017), 17–27.
- [160] Jingui Xie, Qi-Ming He, and Xiaobo Zhao. 2008. Stability of a priority queueing system with customer transfers. *Operations Research Letters* 36 (11 2008), 705–709. <https://doi.org/10.1016/j.orl.2008.06.007>
- [161] Susan H. Xu and Y. Quennel Zhao. 1996. Dynamic routing and jockeying controls in a two-station queueing system. *Advances in Applied Probability* 28, 4 (1996), 1201–1226. <https://doi.org/10.2307/1428170>
- [162] Li Yang and Abdallah Shami. 2020. On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing* 415 (2020), 295–316.
- [163] Mehmet Akif Yazici and Nail Akar. 2013. Analysis of continuous feedback Markov fluid queues and its applications to modeling Optical Burst Switching. In *Proceedings of the 2013 25th International Teletraffic Congress (ITC)*. 1–8. <https://doi.org/10.1109/ITC.2013.6662952>
- [164] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. 2020. A Survey of Autonomous Driving: Common Practices and Emerging Technologies. *IEEE Access* 8 (2020), 58443–58469. <https://doi.org/10.1109/ACCESS.2020.2983149>
- [165] A. Badamchi Zadeh. 2015. A batch arrival multi phase queueing system with random feedback in service and single vacation policy. *OPSEARCH* 52, 4 (12 2015), 617–630. <https://doi.org/10.1007/s12597-015-0206-9>
- [166] V. M. Zaiats, O. M. Rybytska, and M. M. Zaiats. 2019. An Approach to Assessment of the Value and Quantity of Information in Queueing Systems Based on Pattern Recognition and Fuzzy Sets Theories. *Cybernetics and Systems Analysis* 55 (07 2019), 638–648. Issue 4. <https://doi.org/10.1007/s10559-019-00172-1>
- [167] Yan Ma Zaiming Liu and Zhe George Zhang. 2015. Equilibrium Mixed Strategies in a Discrete-Time Markovian Queue Under Multiple and Single Vacation Policies. *Quality Technology and Quantitative Management* 12, 3 (2015), 369–382. <https://doi.org/10.1080/16843703.2015.11673387>
- [168] Yuan Zhang, Peng Du, Jiang Wang, Teer Ba, Rui Ding, and Ning Xin. 2019. Resource Scheduling for Delay Minimization in Multi-Server Cellular Edge Computing Systems. *IEEE Access* 7 (2019), 86265–86273. <https://doi.org/10.1109/ACCESS.2019.2924032>
- [169] Y. Zhao and Grassmann K. Winfried. 1995. Queueing Analysis of a Jockeying Model. *Operations Research* 43, 3 (1995), 520–529.
- [170] Yiqiang Q. Zhao and Winfried K. Grassmann. 1990. The shortest queue model with jockeying. *Naval Research Logistics* 37 (1990), 773–787.
- [171] Jianshan Zhou, Daxin Tian, Zhengguo Sheng, Xuting Duan, and Xuemin Shen. 2021. Distributed Task Offloading Optimization With Queueing Dynamics in Multiagent Mobile-Edge Computing Networks. *IEEE Internet of Things Journal* 8, 15 (2021), 12311–12328. <https://doi.org/10.1109/JIOT.2021.3063509>

Received 09 January 2023