On the stochastic and asymptotic improvement of First-Come First-Served and nudge scheduling

On the Stochastic and Asymptotic Improvement of First-Come First-Served and Nudge Scheduling

BENNY VAN HOUDT, University of Antwerp, Belgium

Recently it was shown that, contrary to expectations, the First-Come-First-Served (FCFS) scheduling algorithm can be stochastically improved upon by a scheduling algorithm called *Nudge* for light-tailed job size distributions. Nudge partitions jobs into 4 types based on their size, say small, medium, large and huge jobs. Nudge operates identical to FCFS, except that whenever a *small* job arrives that finds a *large* job waiting at the back of the queue, Nudge swaps the small job with the large one unless the large job was already involved in an earlier swap.

In this paper, we show that FCFS can be stochastically improved upon under far weaker conditions. We consider a system with 2 job types and limited swapping between type-1 and type-2 jobs, but where a type-1 job is not necessarily smaller than a type-2 job. More specifically, we introduce and study the Nudge-K scheduling algorithm which allows type-1 jobs to be swapped with up to K type-2 jobs waiting at the back of the queue, while type-2 jobs can be involved in at most one swap. We present an explicit expression for the response time distribution under Nudge-K when both job types follow a phase-type distribution. Regarding the asymptotic tail improvement ratio (ATIR) , we derive a simple expression for the ATIR, as well as for the K that maximizes the ATIR. We show that the ATIR is positive and the optimal K tends to infinity in heavy traffic as long as the type-2 jobs are on average longer than the type-1 jobs.

CCS Concepts: • **Mathematics of computing** → **Probability and statistics**; • **Networks** → **Network performance modeling**.

Additional Key Words and Phrases: scheduling; First-Come First-Served; Nudge; asymptotic improvement; stochastic improvement; response time distribution

1 INTRODUCTION

Although there exists an abundance of scheduling algorithms, many systems still rely on the First-Come-First-Served (FCFS) scheduling algorithm as FCFS is considered to be a fair scheduling algorithm that is easy to implement and does not require any job size information. There is also theoretical support for selecting FCFS apart from the well-known fact that it minimizes the maximum response time of any finite sequence of jobs. If we denote R as the response time of an arbitrary job under FCFS and make the following technical assumptions:

- Jobs arrive according to a Poisson process.
- The job size distribution X is light-tailed (which means there exists an $\epsilon > 0$ such that $E[e^{-\epsilon X}]$ is finite).
- If $\tilde{S}(s)$ denotes the Laplace transform of the job size distribution, then $\tilde{S}(s)$ has either no singularities or if $s^* < 0$ is its right-most singularity, then $\tilde{S}(s^*) = \infty$.

Then there exist constants $\theta_Z > 0$ and $c_{FCFS} > 0$ such that

$$P[R > t] \sim c_{FCFS} e^{-\theta_Z t},$$

Author's address: Benny Van Houdt, University of Antwerp, Middelheimlaan 1, Antwerp, , B-2020, Belgium, benny.vanhoudt@uantwerpen.be.

where $\sim$ indicates that the ratio of the two quantities converges to 1 as t tends to infinity. Note that the latter two assumptions correspond to a class-I distribution in [1, Section 5] and these distributions include all well-behaved light-tailed distributions such as any phase-type distribution or distribution with finite support (such as truncated heavy-tailed distributions). The constant θ_Z is called the decay rate. Let π be any scheduling algorithm and R_π be its associated response time distribution (in an M/G/1 queue with a class-I job size distribution), then (see [3, Section 3.1]) there exists a constant $M(\pi) \geq 0$ such that

$$\limsup_{t \to \infty} \frac{P[R > t]}{P[R_\pi > t]} \leq M(\pi).$$

This is equivalent to stating that FCFS has the largest possible decay rate in systems subject to Poisson arrivals and class-I job sizes (in fact, θ_Z is equal to the decay rate of the workload distribution Z in the system). Any scheduling algorithm with the largest possible decay rate is called **weakly tail optimal**. In [16] FCFS was conjectured to be **strongly tail optimal** for class-I job sizes, which would imply that $M(\pi) \leq 1$ for any π and FCFS results in the best possible tail behavior for class-I job sizes.

In a recent paper [5] FCFS was shown *not* to be strongly tail optimal by introducing a scheduling algorithm called Nudge such that $M(Nudge) > 1$. Further, contrary to expectations, it was shown that the Nudge scheduling algorithm can stochastically improve upon FCFS. A scheduling algorithm π_1 is said to **stochastically improve** upon an algorithm π_2 if and only if $P[R_{\pi_1} > t] < P[R_{\pi_2} > t]$, for any $t > x_{min}$, where x_{min} is the infimum of the support of the job size distribution X (for $t \leq x_{min}$, we have $P[R_{\pi_1} > t] = P[R_{\pi_2} > t]$). This can be restated by saying that the tail improvement ratio (TIR) in t defined as

$$\text{TIR}(t) = 1 - \frac{P[R_{\pi_1} > t]}{P[R_{\pi_2} > t]},$$

is positive for all $t > x_{min}$. This means that Nudge improves every moment and percentile of the response time of FCFS! Note that it is easy to devise scheduling algorithms that reduce the mean response time of FCFS, but this typically comes at the expense of a worse decay rate [11].

To achieve this stochastic improvement Nudge partitions the jobs into 4 types, say small, medium, large and huge jobs, based on their size using three thresholds x_1, x_2 and x_3. Nudge then operates in the same manner as FCFS, except that whenever a *small* job arrives that finds a *large* job waiting at the back of the queue, Nudge swaps the small job with the large one unless the large job was already involved in an earlier swap. The authors of [5] then showed that it is possible for any continuous class-I job size distribution X, to find appropriate thresholds x_1, x_2 and x_3 (that depend on X) such that Nudge stochastically improves upon FCFS. Simulation experiments further showed that this is often still the case if $x_1 = x_2$ and $x_3 = \infty$, which means in the absence of medium and huge jobs (which were needed for the proofs).

In this paper we consider a system with two types of jobs (see Section 2 for details), where a random type-1 job is not necessarily smaller than a random type-2 job, and a set of scheduling algorithms called Nudge-K, where $K \geq 0$ is an input parameter (that can be set equal to ∞). Under Nudge-K any arriving type-1 job can be swapped with at most K type-2 jobs waiting at the back of the queue, but type-2 jobs can be involved in at most one swap. This means that a type-1 job passes up to K jobs waiting at the back of the queue until it either encounters another type-1 job, a type-2 job that was already swapped or becomes the job waiting at the head of the queue. Note that Nudge-1 coincides with Nudge if we set $x_1 = x_2$, $x_3 = \infty$ and call the small jobs type-1 and the large jobs type-2.

The main contributions of the paper can be summarized as follows:

(1) For the system described in Section 2 we derive an *explicit* expression for the complementary cumulative distribution function (ccdf) of the waiting time (see Theorems 4 and 6) and response time distribution (see Theorems 5 and 7) of type-1 and type-2 jobs. To derive these results we first obtain an explicit expression for the workload in the system (see Theorem 1).

(2) We derive a simple expression for the asymptotic tail improvement ratio of Nudge-K over FCFS defined as

$$\text{ATIR}(K) = 1 - \lim_{t \to \infty} \frac{P[R_{Nudge-K} > t]}{P[R > t]},$$

as well as for the value of K that maximizes the ATIR(K), denoted as K_{opt} (see Theorem 8). We identify simple conditions on when the ATIR(K) > 0, for a given K, for all K and for $K = 1$ (see Theorem 9). We further prove that under heavy traffic, the ATIR(K) > 0 provided that the mean type-2 job size exceeds the mean type-1 job size and that K_{opt} tends to infinity (see Theorems 10 and 11).

(3) We present various novel insights on the stochastic and asymptotic improvement upon FCFS and Nudge in Section 8 using numerical experiments. These show that stochastic improvements of FCFS exist under far weaker conditions that the one considered in [5], that Nudge can be stochastically improved upon, that setting K too large may imply that Nudge-K no longer stochastically improves upon FCFS, that an asymptotic improvement does not necessarily imply a stochastic improvement even if type-2 jobs stochastically dominate type-1 jobs, etc.

The fact that FCFS can be stochastically improved upon under far weaker conditions is important as it is much easier in a real system to identify different types of jobs such that one job type is typically larger and/or has a heavier tail than another type. In such case implementing an algorithm like Nudge-K using these types may improve all percentiles of the response time, without the need of having any indication on the size of individual jobs (being larger or smaller than some threshold) as in [5].

The paper is structured as follows. The exact model considered in the paper is presented in Section 2. A matrix exponential expression for the workload distribution is derived in Section 3, while the mean response time of Nudge-K is analyzed in Section 4. Explicit expressions for the type-2 and type-1 waiting and response time distributions are part of Sections 5 and 6, respectively. Section 7 contains the results for the ATIR, these results are the most elegant results in the paper. Numerical examples and insights are discussed in Section 8. Conclusions are drawn and possible future work is listed in Section 9.

2 THE SYSTEM

We consider a queueing system with Poisson arrivals with rate λ. Arriving jobs are type-1 jobs with probability p, or type-2 jobs with probability $1 - p$. Job types of consecutive jobs are independent. The processing time X_i of a type-i job follows an order n_i phase-type distribution characterized by (α_i, S_i), that is, $P[X_i > t] = \alpha_i e^{S_i t} \mathbf{1}$, where $\mathbf{1}$ is a column vector of ones of the appropriate dimension. Let $E[X_i] = \alpha_i(-S_i)^{-1}\mathbf{1}$ be the mean service time of a type-i job. We assume without loss of generality that $E[X] = pE[X_1]+(1-p)E[X_2] = 1$, with $X = pX_1+(1-p)X_2$ the job size distribution, such that the load of the system is λ. For further use, define $s_i^* = (-S_i)\mathbf{1}$, $\alpha = (p\alpha_1, (1-p)\alpha_2)$ and

$$S = \begin{bmatrix} S_1 & 0 \\ 0 & S_2 \end{bmatrix},$$

such that X has a phase-type distribution characterized by (α, S). Note that $\alpha(-S)^{-1}\mathbf{1} = 1$ as $E[X] = 1$. Denote $\tilde{S}_i(s) = \alpha_i(sI - S_i)^{-1}(-S_i)\mathbf{1}$, for $i = 1, 2$, as the Laplace transform of the size of a type-i job. Let $\tilde{S}(s) = p\tilde{S}_1(s) + (1 - p)\tilde{S}_2(s)$ be the Laplace transform of a random job. As X is a

phase-type distribution, it is a class-I distribution. It is well known that any general positive-valued distribution can be approximated arbitrary close with a PH distribution [9]. Further, various fitting algorithms and tools are available online for phase-type distributions (e.g., [4, 8, 13]).

The scheduling algorithm studied in this paper is called the **Nudge-K algorithm**. Under this algorithm a type-1 job can be swapped with at most K type-2 jobs waiting at the back of the queue and any type-2 job can be swapped at most once. In other words, when a type-1 job arrives it can pass up to K waiting type-2 jobs at the back of the queue until it either encounters a type-1 job, a type-2 job that was already passed by another type-1 job or becomes the job waiting at the head of the queue. The job that is being served is never swapped.

3   WORKLOAD AND FCFS RESPONSE TIME DISTRIBUTION

We first provide an **explicit matrix exponential** expression for the workload distribution, which corresponds to the waiting time distribution in case of FCFS. The workload distribution does not depend on the scheduling algorithm, as long as it is work-conserving. Note that if Y is a class-I distribution, the probability $P[Y > t]$ decays exponentially fast and the decay rate θ_Y can be expressed as $\theta_Y = -\lim_{t\to\infty} \frac{1}{t} \log P[Y > t]$.

THEOREM 1. *Let Z be the workload in the system, then*

$$P[Z > t] = \lambda \alpha e^{Tt}(-S)^{-1}\mathbf{1} = \lambda \beta e^{Tt}(-T)^{-1}\mathbf{1}, \tag{1}$$

with $\beta = (1 - \lambda)\alpha$ and

$$T = S + \lambda \mathbf{1}\alpha. \tag{2}$$

Let $\theta_i = -\lim_{t\to\infty} \frac{1}{t} \log P[X_i > t]$ and $\theta_Z = -\lim_{t\to\infty} \frac{1}{t} \log P[Z > t]$, then $0 < \theta_Z < \min(\theta_1, \theta_2)$.

PROOF. It is well known (see (5.41) on p247 in [6]) that the Pollaczek-Khinchin formula for the Laplace transform of the workload $\tilde{Z}(s)$ in an M/G/1 queue can be rewritten as

$$\tilde{Z}(s) = (1 - \lambda) \sum_{n=0}^{\infty} \lambda^n [\tilde{S}_{Res}(s)]^n,$$

where $\tilde{S}_{Res}(s)$ is the Laplace transform of the residual service time. This implies that

$$P[Z > t] = \lambda \sum_{n=1}^{\infty} \lambda^{n-1}(1 - \lambda)P[S_{Res}^{(n*)} > t],$$

where $S_{Res}^{(n*)}$ is the n-fold convolution of the residual service time. The residual service time of a phase-type distribution with representation (α, S) and mean 1, is also phase-type with representation $(\alpha(-S)^{-1}, S)$. The probability $P[Z > t]$ can therefore be expressed as λ times a geometric sum of the phase-type distribution $(\alpha(-S)^{-1}, S)$, which is again phase-type with representation $(\alpha(-S)^{-1}, S + \lambda s^* \alpha(-S)^{-1})$. Hence,

$$P[Z > t] = \lambda \alpha(-S)^{-1} e^{(S+\lambda s^* \alpha(-S)^{-1})t}\mathbf{1}.$$

We now note that

$$(-S)^{-1}e^{(S+\lambda s^*\alpha(-S)^{-1})t} = (-S)^{-1}\sum_{k=0}^{\infty}(S+\lambda s^*\alpha(-S)^{-1})^k\frac{t^k}{k!}$$

$$= \sum_{k=0}^{\infty}\left[(-S)^{-1}(S+\lambda s^*\alpha(-S)^{-1})(-S)\right]^k\frac{t^k}{k!}(-S)^{-1}$$

$$= \sum_{k=0}^{\infty}(S+\lambda\mathbf{1}\alpha)^k\frac{t^k}{k!}(-S)^{-1} = e^{(S+\lambda\mathbf{1}\alpha)t}(-S)^{-1},$$

as $s^* = -S\mathbf{1}$. This shows that

$$P[Z > t] = \lambda\alpha e^{(S+\lambda\mathbf{1}\alpha)t}(-S)^{-1}\mathbf{1}.$$

The second equality in (1) now follows by noting that we can use the Sherman-Morrison formula to find that

$$(-T)^{-1}\mathbf{1} = (-S)^{-1}\mathbf{1} + \frac{\lambda(-S)^{-1}\mathbf{1}\alpha(-S)^{-1}\mathbf{1}}{1-\lambda\alpha(-S)^{-1}\mathbf{1}} = \frac{(-S)^{-1}\mathbf{1}}{1-\lambda}, \tag{3}$$

as $\alpha(-S)^{-1}\mathbf{1} = E[X] = 1$. The fact that $\theta_Z < \min(\theta_1, \theta_2)$ follows by noting that for $\zeta > \max_i |S_{ii}|$ the matrix $T + \zeta I$ is a primitive non-negative matrix with Perron-Frobenius eigenvalue $\zeta - \theta_Z$ [14]. Therefore for any eigenvalue β of a matrix $0 \le B \le T + \zeta I$ with inequality in at least one entry, we have $|\beta| < \zeta - \theta_Z$ by [14, Theorem 1.1(e)]. Setting $B = S + \zeta I$ therefore implies that the real eigenvalue $\zeta - \min(\theta_1, \theta_2)$ of B is strictly smaller than $\zeta - \theta_Z$. □

Remark: $-\theta_Z < 0$ is a real eigenvalue of T and for any other eigenvalue ξ of T we have $Re(\xi) < -\theta_Z$. $-\theta_Z$ may not be equal to the spectral radius of T (that is, $|\xi| \le -\theta_Z$ may not hold), but $-\theta_Z t$ is the spectral radius of e^{Tt}. Further, as T is irreducible, it has a unique right and left eigenvector (up to multiplication by a constant) associated with $-\theta_Z$ and these eigenvectors can be chosen strictly positive [14]. If we denote these unique eigenvectors as u_T and v_T^* and normalize such that $v_T^* u_T = 1$, then

$$\lim_{t\to\infty}e^{\theta_Z t}e^{Tt} = u_T v_T^*,$$

when T is irreducible, meaning

$$c_Z = \lim_{t\to\infty}e^{\theta_Z t}P[Z > t] = \lambda(\beta u_T)(v_T^*(-T)^{-1}\mathbf{1}),$$

To prove the next theorem we rely on the following Lemma:

LEMMA 1 (THEOREM 1 IN [15]). *Let*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix},$$

then

$$e^{At} = \begin{bmatrix} e^{A_{11}t} & \int_0^t e^{A_{11}s}A_{12}e^{A_{22}(t-s)}\,ds \\ 0 & e^{A_{22}t} \end{bmatrix}.$$

THEOREM 2. *Let R be the response time distribution of a job in case of FCFS, then*

$$P[R > t] = (1-\lambda)\alpha e^{St}\mathbf{1} + \lambda(\beta, 0)e^{Ut}\begin{pmatrix}(-T)^{-1}\mathbf{1} \\ \mathbf{1}\end{pmatrix}, \tag{4}$$

with

$$U = \begin{bmatrix} T & \mathbf{1}\alpha \\ 0 & S \end{bmatrix}.$$

Further, $\theta_Z = -\lim_{t\to\infty} \frac{1}{t} \log P[R > t]$ and $c_{FCFS} = \lim_{t\to\infty} e^{\theta_Z t} P[R > t]$ can be expressed as

$$c_{FCFS} = c_Z \tilde{S}(-\theta_Z).$$

PROOF. The response time R in case of FCFS is given by the workload Z plus the job size X. The density of the workload Z is given by $\lambda \beta e^{Ts} \mathbf{1}$ due to (1), hence

$$P[R > t] = P[Z > t] + \lambda \int_0^t \beta e^{Ts} \mathbf{1} \alpha e^{S(t-s)} \mathbf{1} ds + (1-\lambda)\alpha e^{St} \mathbf{1},$$

as a job finds a workload equal to zero with probability $1 - \lambda$. The result for $P[R > t]$ now follows using Lemma 1, (1) and by combining both matrix exponentials.

As the response time R equals the workload Z plus the service time X which is independent of the workload, the Laplace transform of the response time $\tilde{R}(s) = \tilde{Z}(s)\tilde{S}(s)$. Applying the final value theorem to $e^{\theta_Z t} P[R > t]$ we therefore have

$$c_{FCFS} = \lim_{t\to\infty} e^{\theta_Z t} P[R > t] = \frac{1}{\theta_Z} \lim_{s\to 0} s\tilde{R}(s - \theta_Z) = \frac{1}{\theta_Z} \lim_{s\to 0} s\tilde{Z}(s - \theta_Z)\tilde{S}(s - \theta_Z)$$

$$= \lim_{t\to\infty} e^{\theta_Z t} P[Z > t] \tilde{S}(-\theta_Z) = c_Z \tilde{S}(-\theta_Z),$$

where we used the time-domain integration and frequency shifting properties of the Laplace transform (in the 2nd and 4th equality). □

We can also argue that

$$c_{FCFS} = \lim_{t\to\infty} e^{\theta_Z t} P[R > t] = \lambda(\beta, 0) u_U v_U^* \begin{pmatrix} (-T)^{-1}\mathbf{1} \\ 1 \end{pmatrix},$$

where u_U and v_U^* are the unique right and left eigenvectors of U associated with the eigenvalue $-\theta_Z$ such that $v_U^* u_U = 1$.

4 MEAN RESPONSE TIME OF NUDGE-K

Let e_i and e_i^* represent the i-th column and i-th row of the size $K + 1$ identity matrix, respectively.

LEMMA 2. Given that a type-2 job sees a workload of $s > 0$ upon arrival, it is swapped with probability

$$p_{swap}(s) = e_1^* e^{Ms} e_{K+1}, \tag{5}$$

where M is a size $K + 1$ matrix with entry m_{ij} given by

$$m_{ij} = \begin{cases} -\lambda & 1 \le i = j \le K, \\ \lambda(1-p) & 1 \le i < K, j = i+1, \\ \lambda p & 1 \le i \le K, j = K+1, \\ 0 & otherwise \end{cases}.$$

For $K = \infty$, we have $p_{swap}(s) = 1 - e^{-\lambda ps}$.

PROOF. Consider the continuous time Markov chain with upper-triangular rate matrix

$$Q = \begin{bmatrix} M & \lambda(1-p)e_{K+1} \\ 0 & 0 \end{bmatrix} = \left[\begin{array}{cccc|cc} -\lambda & \lambda(1-p) & & & \lambda p & \\ & \ddots & \ddots & & \vdots & \\ & & -\lambda & \lambda(1-p) & \lambda p & \\ & & & -\lambda & \lambda p & \lambda(1-p) \\ \hline & & & & 0 & \\ & & & & & 0 \end{array}\right]. \tag{6}$$

This $K + 2$-state Markov chain has two absorbing states: state $K + 1$ and $K + 2$. Entry (i, j) of e^{Qs}, with $i \le j \le K$, represents the probability there are exactly $j - i$ type-2 arrivals and zero type-1 arrivals in an interval of length s. Entry $(i, K + 2)$ holds the probability that more than $K - i$ arrivals occur in an interval of length s and the first $K - i + 1$ arrivals are type-2, while entry $(i, K + 1)$ contains the remaining probability mass and corresponds to the probability that there is a type-1 arrival in an interval of length s that is preceded by at most $K - i$ type-2 arrivals.

A type-2 job is swapped if there is a type-1 arrival that is preceded by at most $K - 1$ type-2 arrivals while it is waiting. Hence, $p_{swap}(s)$ can be expressed as entry $(1, K + 1)$ of e^{Qs}, which is identical to entry $(1, K + 1)$ of e^{Ms} and can be written in matrix form as $e_1^* e^{Ms} e_{K+1}$.

The expression for $K = \infty$ is immediate by noting that a type-2 job is swapped as soon as a single type-1 arrival occurs during its waiting time, irrespective of whether and where type-2 arrivals occur. □

THEOREM 3. *The mean response time $E[R_{Nudge-K}]$ of the Nudge-K algorithm can be expressed as*

$$E[R_{Nudge-K}] = E[R] + (1 - p)p_{swap}(E[X_1] - E[X_2]),$$

with

$$p_{swap} = -\lambda(\beta \otimes e_1^*)(T \oplus M)^{-1}(\mathbf{1} \otimes e_{K+1}),$$

and $E[R] = 1 + \lambda\beta T^{-2}\mathbf{1} = 1 + \frac{\lambda E[S^2]}{2(1-\lambda)}$. When $K = \infty$ we have $p_{swap} = \lambda(1 - \beta(\lambda pI - T)^{-1}\mathbf{1})$.

PROOF. A swap between a type-1 and a type-2 job changes the mean response time by $E[X_1] - E[X_2]$ on average as the response time of the type-2 job increases by $E[X_1]$ on average and that of the type-1 job decreases by $E[X_2]$ on average. The rate at which swaps occur equals λ times $(1 - p)$ times the probability p_{swap} that a random type-2 job is swapped. Making use of Theorem 1 and Lemma 2, we have

$$p_{swap} = \lambda \int_0^\infty \beta e^{Ts} \mathbf{1} p_{swap}(s)ds = \lambda \int_0^\infty (\beta \otimes e_1^*)(e^{Ts} \otimes e^{Ms})(\mathbf{1} \otimes e_{K+1})ds$$

$$= \lambda \int_0^\infty (\beta \otimes e_1^*)e^{(T \oplus M)s}(\mathbf{1} \otimes e_{K+1})ds = -\lambda(\beta \otimes e_1^*)(T \oplus M)^{-1}(\mathbf{1} \otimes e_{K+1}).$$

Finally, the mean response time is found as the mean response time $E[R]$ in case of FCFS plus the mean number of swaps that occur per arrival times the average change $(E[X_1] - E[X_2])$ caused by a swap. The mean number of swaps that occur per arrival is clearly given by the ratio between the swap rate $\lambda(1 - p)p_{swap}$ and the arrival rate λ.

The expression for $E[R]$ in terms of $E[S^2]$ is well known. The other expression follows from Theorem 1 and can also be obtained directly using (3) and the fact that $E[S^2] = 2\alpha(-S)^{-2}\mathbf{1}$. □

Remark: The mean response time of Nudge-K is smaller than the mean response time $E[R]$ under FCFS if and only if $E[X_2] > E[X_1]$. This implies that a stochastic improvement is only possible if type-2 jobs are on average larger than type-1 jobs.

The probability $p_{swap}(s)$ is clearly increasing in K and therefore so is p_{swap}. This implies that setting $K = \infty$ minimizes the mean response time of Nudge-K provided that $E[X_2] > E[X_1]$. We will see however that this choice for K is *never* optimal if we focus on the asymptotic tail improvement ratio.

5 RESPONSE TIME DISTRIBUTION OF TYPE-2 JOBS

We now proceed with the waiting time distribution of type-2 jobs.

THEOREM 4. *Let $W^{(2)}$ be the waiting time distribution of a type-2 job, then*

$$P[W^{(2)} > t] = P[Z > t] + \lambda(\beta \otimes e_1^*, 0)e^{T^{(2)}t}\begin{pmatrix} 0_m \\ 1 \end{pmatrix}, \tag{7}$$

with 0_m a column vector of zeros with the same size as $(T \oplus M)\mathbf{1}$,

$$T^{(2)} = \begin{bmatrix} T \oplus M & \mathbf{1} \otimes e_{K+1}\alpha_1 \\ 0 & S_1 \end{bmatrix},$$

where M is defined in Lemma 2.

Further, $-\lim_{t\to\infty} \frac{1}{t}\log P[W^{(2)} > t] = \theta_Z$ and $c_{W^{(2)}} = \lim_{t\to\infty} e^{\theta_Z t}P[W^{(2))} > t]$ is given by

$$c_{W^{(2)}} = (1-p)^K c_Z + (1 - (1-p)^K)c_Z \tilde{S}_1(-\theta_Z). \tag{8}$$

PROOF. The waiting time of a type-2 job equals the workload Z in the queue when it arrives plus the workload of a type-1 job if the type-2 job is swapped. As $\lambda\beta e^{Ts}\mathbf{1}$ is the density of the workload and a type-2 job that sees a workload of s is swapped with probability $p_{swap}(s) = e_1^*e^{Ms}e_{K+1}$ due to Lemma 2, we get

$$P[W^{(2)} > t] = P[Z > t] + \lambda \int_0^t \beta e^{Ts}\mathbf{1}(e_1^*e^{Ms}e_{K+1})\alpha_1 e^{S_1(t-s)}\mathbf{1}ds,$$

$$= P[Z > t] + \lambda \int_0^t (\beta \otimes e_1^*)(e^{Ts} \otimes e^{Ms})(\mathbf{1} \otimes e_{K+1})\alpha_1 e^{S_1(t-s)}\mathbf{1}ds$$

$$= P[Z > t] + \lambda \int_0^t (\beta \otimes e_1^*)e^{(T\oplus M)s}(\mathbf{1} \otimes e_{K+1})\alpha_1 e^{S_1(t-s)}\mathbf{1}ds.$$

Applying Lemma 1 then yields (7). The first limit for t tending to infinity follows from noting that the eigenvalue of $e^{T^{(2)}t}$ with the largest real part is given by the eigenvalue with the largest real part of the matrices $e^{(T\oplus M)t}$ and $e^{S_1 t}$. As $e^{(T\oplus M)t} = e^{Tt} \otimes e^{Mt}$, the eigenvalues of $e^{(T\oplus M)t}$ are the products of the eigenvalues of e^{Tt} and e^{Mt}. The eigenvalues of e^{Mt} are 1 and $e^{-\lambda t}$ (with multiplicity K), while the eigenvalue with the largest real part of e^{Tt} equals $-\theta_Z t$ by definition. Hence, $-\theta_Z t$ is the eigenvalue with the largest real part of $e^{(T\oplus M)t}$ and therefore also of $e^{T^{(2)}t}$ as $\theta_Z < \theta_1$.

The expression for $c_{W^{(2)}}$ follows from first noting that $p_{swap}(s)$ is increasing in s and therefore

$$e_1^*e^{Ms}e_{K+1} \leq \lim_{s\to\infty} e_1^*e^{Ms}e_{K+1} = 1 - (1-p)^K.$$

Thus for any $\epsilon > 0$ there exist a t_ϵ such that

$$1 - (1-p)^K - \epsilon < e_1^*e^{Ms}e_{K+1} < 1 - (1-p)^K,$$

for $s > t_\epsilon$. Further,

$$\lim_{t\to\infty} e^{\theta_Z t} \int_0^{t_\epsilon} \beta e^{Ts}\mathbf{1}(e_1^*e^{Ms}e_{K+1})\alpha_1 e^{S_1(t-s)}\mathbf{1}ds = 0,$$

as $\alpha e^{S_1 t}\mathbf{1}$ decays faster than $e^{-\theta_Z t}$. This implies that

$$\lim_{t\to\infty} e^{\theta_Z t}P[W^{(2)} > t] =$$

$$\lim_{t\to\infty} e^{\theta_Z t}\left(P[Z > t] + (1 - (1-p)^K)\int_0^t (\lambda\beta e^{Ts}\mathbf{1})\alpha_1 e^{S_1(t-s)}\mathbf{1}ds\right) =$$

$$\lim_{t\to\infty} e^{\theta_Z t}\left((1-p)^K P[Z > t] + (1 - (1-p)^K)P[Z + X_1 > t]\right),$$

from which the expression for $c_{W^{(2)}}$ follows by using the final value theorem in the same manner as in the proof of Theorem 2. □

Remark that when $K = \infty$ we can still use the above approach by simply replacing the matrix M by the 2×2 matrix

$$M = \begin{bmatrix} -\lambda p & \lambda p \\ 0 & 0. \end{bmatrix},$$

and e_{K+1} by e_2. The same remark applies to Theorem 5.

THEOREM 5. *Let $R^{(2)}$ be the response time distribution of a type-2 job, then*

$$P[R^{(2)} > t] = (1 - \lambda)\alpha_2 e^{S_2 t} \mathbf{1} + \lambda(\beta, 0) e^{U_2^{(2)} t} \begin{pmatrix} (-T)^{-1}\mathbf{1} \\ \mathbf{1} \end{pmatrix} + \lambda(\beta \otimes e_1^*, 0) e^{U_1^{(2)} t} \begin{pmatrix} 0_m \\ \mathbf{1} \end{pmatrix}, \tag{9}$$

with

$$U_1^{(2)} = \begin{bmatrix} T \oplus M & \mathbf{1} \otimes e_{K+1}\alpha_1 & -\mathbf{1} \otimes e_{K+1}\alpha_2 \\ 0 & S_1 & s_1^*\alpha_2 \\ 0 & 0 & S_2 \end{bmatrix},$$

and

$$U_2^{(2)} = \begin{bmatrix} T & \mathbf{1}\alpha_2 \\ 0 & S_2 \end{bmatrix}.$$

Further, $-\lim_{t\to\infty} \frac{1}{t} \log P[R^{(2)} > t] = \theta_Z$ *and* $c_{R^{(2)}} = \lim_{t\to\infty} e^{\theta_Z t} P[R^{(2)} > t] = c_{W^{(2)}} \tilde{S}_2(-\theta_Z)$.

PROOF. The result follows from noting that

$$P[R^{(2)} > t] = P[Z > t] + \lambda \int_0^t \beta e^{Ts} \mathbf{1}(e_1^* e^{Ms} e_{K+1}) P[X_1 + X_2 > t - s] ds,$$

$$+ (1 - \lambda)\alpha_2 e^{S_2 t} \mathbf{1} + \lambda \int_0^t \beta e^{Ts} \mathbf{1}(1 - e_1^* e^{Ms} e_{K+1}) \alpha_2 e^{S_2(t-s)} \mathbf{1} ds, \tag{10}$$

with

$$P[X_1 + X_2 > t - s] = (\alpha_1, 0) e^{\begin{bmatrix} S_1 & s_1^*\alpha_2 \\ 0 & S_2 \end{bmatrix}(t-s)} \mathbf{1}.$$

The result now follows using Lemma 1 and combining some of the matrix exponentials. An alternate proof exists in making use of the fact that

$$P[R^{(2)} > t] = (1 - \lambda)\alpha_2 e^{S_2 t} \mathbf{1} + P[W^{(2)} > t] + \int_0^t \left(-\frac{\partial}{\partial s} P[W^{(2)} > s] \right) \alpha_2 e^{S_2(t-s)} \mathbf{1} ds.$$

The equality $c_{R^{(2)}} = c_{W^{(2)}} \tilde{S}_2(-\theta_Z)$ follows from the final value theorem. □

Apart from (8) we can also express $c_{W^{(2)}}$ as

$$c_{W^{(2)}} = c_Z + \lambda(\beta \otimes e_1^*, 0) u_{T^{(2)}} v_{T^{(2)}}^* \begin{pmatrix} 0_m \\ \mathbf{1} \end{pmatrix},$$

where $u_{T^{(2)}}$ and $v_{T^{(2)}}^*$ are the unique right and left eigenvectors of $T^{(2)}$ associated with the eigenvalue $-\theta_Z$ such that $v_{T^{(2)}}^* u_{T^{(2)}} = 1$.

6   RESPONSE TIME DISTRIBUTION OF TYPE-1 JOBS

In this section we present an approach for the waiting time and response time distribution of a type-1 job. The approach exists in computing the **workload** distribution seen by a type-1 job from the **queue length** distribution of the FCFS queue. The following observation make this possible.

(1) Whenever a type-2 job is waiting in the FCFS queue, it is also waiting in the Nudge-K queue.
(2) If the FCFS queue contains i type-2 jobs waiting at the back of the queue, these i jobs are also waiting at the back of the Nudge-K queue (as there have been no type-1 arrivals after these type-2 arrivals).
(3) If the $i + 1$ jobs waiting at the back of the FCFS queue are a type-1 job, say job j, followed by i type-2 jobs, then any new type-1 arrival passes exactly $\min(i, K)$ type-2 jobs in the Nudge-K queue. Note that job j may have passed one or multiple type-2 jobs in the Nudge-K queue, but in such case the $(i + 1)$-th last job in the Nudge-K queue is a type-2 job that was already swapped.

Hence, we conclude that the number of type-2 jobs that a tagged type-1 job passes under Nudge-K is equal to the minimum of K and the number of type-2 jobs that are **waiting at the back of the FCFS queue** when the tagged type-1 job arrives. Note that this argument fails if a type-2 job can be passed by more than one type-1 job, which is not the case under Nudge-K.

THEOREM 6.   *Let $W^{(1)}$ be the waiting time distribution of a type-1 job. Let $R = -\lambda(S - \lambda I + \lambda \mathbf{1}\alpha)^{-1}$ and $\pi_1 = (1 - \lambda)\alpha R$. Further define*

$$\pi_0^{(1)} = \pi_1(I - (1 - p)^{K+1} R^{K+1})(I - (1 - p)R)^{-1}, \tag{11}$$

$$\pi_1^{(1)} = \pi_1(1 - p)^K R^{K+1}, \tag{12}$$

$$\pi_0^{(2)} = \pi_1 R(I - (1 - p)^K R^K)(I - (1 - p)R)^{-1} p, \tag{13}$$

then

$$P[W^{(1)} > t] = v_1 e^{(S^\top \otimes I + (s^*\alpha)^\top \otimes R)t} \xi + (\pi_0^{(1)} - \pi_1^{(1)} R^{-1})e^{St}\mathbf{1} - \pi_0^{(2)} R^{-1} e^{St} \begin{pmatrix} \mathbf{1}/p \\ 0 \end{pmatrix}, \tag{14}$$

where $X^\top$ denotes the transposed of X,

$$v_1 = \mathbf{1}^\top \otimes [(\pi_1^{(1)} + \pi_0^{(2)})(I - R)^{-1} + \pi_1^{(1)} R^{-1}] + (\mathbf{1}^\top, 0)/p \otimes \pi_0^{(2)} R^{-1},$$

and ξ is a size $(n_1 + n_2)^2$ vector obtained from stacking the columns of the size $n_1 + n_2$ unity matrix.
Further, $-\lim_{t \to \infty} \frac{1}{t} \log P[W^{(1)} > t] = \theta_Z$ and $c_{W^{(1)}} = \lim_{t \to \infty} e^{\theta_Z t} P[W^{(1)}) > t]$ is given by

$$c_{W^{(1)}} = c_Z(1 - p)^K \tilde{S}(-\theta_Z)^{-K} + c_Z p \frac{\tilde{S}_1(-\theta_Z)}{\tilde{S}(-\theta_Z)} \frac{1 - (1 - p)^K \tilde{S}(-\theta_Z)^{-K}}{1 - (1 - p)\tilde{S}(-\theta_Z)^{-1}}. \tag{15}$$

PROOF.   Let $P[Q^{(FCFS)} = (q, i)]$ be the steady state probability that the FCFS queue contains q jobs and the server is in service phase $i \in \{1, 2, \dots, n_1 + n_2\}$, for $q > 0$. The probability that the queue is empty is clearly $1 - \lambda$. It is well known [10] that

$$P[Q^{(FCFS)} = (q, i)] = (\pi_1 R^{q-1})_i,$$

with $R = -\lambda(S - \lambda I + \lambda \mathbf{1}\alpha)^{-1}$ and $\pi_1 = (1 - \lambda)\alpha R$.

As noted earlier, the workload seen by a tagged type-1 job corresponds to the work present at a random point in time in the FCFS queue if we remove up to K type-2 jobs that are waiting at the back of the FCFS queue. Note that if fewer than K jobs are removed because of the presence of another type-1 job, then this should be taken into account when computing the workload.

We now define two reduced queue length distributions. The first Q_1 corresponds to the case where either

(1) The FCFS queue contains $1 \le k \le K+1$ jobs upon arrival of the tagged type-1 job and the $k-1$ waiting jobs are type-2 jobs (that are swapped with the tagged type-1 job by the Nudge-K algorithm). This occurs when the FCFS server is in phase i with probability $(\pi_k)_i(1-p)^{k-1}$.

(2) The FCFS queue contains exactly $q+1+K$ jobs, with $q > 0$, upon arrival of the tagged type-1 job and the last K waiting jobs are type-2 jobs (that are swapped with the tagged type-1 job). This happens when the FCFS server is in phase i with probability $(\pi_{K+1+q})_i(1-p)^K$.

In both cases the **workload** seen by the tagged type-1 job is nonzero and corresponds to the sum of $q \ge 0$ jobs plus a remaining service time which if the current phase equals i, has a phase-type distribution given by (e_i, S). Hence,

$$P[Q_1 = (q, i)] = 1[q = 0] \sum_{k=1}^{K+1} (\pi_k)_i(1-p)^{k-1} + 1[q > 0](\pi_{K+1+q})_i(1-p)^K. \tag{16}$$

The second Q_2 corresponds to the case where the tagged job is swapped with $0 \le k < K$ jobs and there is at least one type-1 job waiting in the FCFS queue. In this case the workload consists of the sum of $q \ge 0$ jobs, one type-1 job and a remaining service time with phase-type distribution (e_i, S), if the server is in phase i:

$$P[Q_2 = (q, i)] = \sum_{k=0}^{K-1} (\pi_{q+k+2})_i(1-p)^k p. \tag{17}$$

Note that Q_1 and Q_2 have a matrix geometric form with the same rate matrix R as the FCFS queue. More specifically, let the vectors $\pi_q^{(j)}$ contain the probabilities $P[Q_j = (q, i)]$, for $j = 1, 2$, then $\pi_q^{(1)} = \pi_1^{(1)} R^{q-1}$ for $q > 0$ and $\pi_q^{(2)} = \pi_0^{(2)} R^q$ for $q \ge 0$. The expressions in (11), (12) and (13) follow from (16) and (17).

The probability $P[W^{(1)} > t]$ that the workload seen by a tagged type-1 job exceeds t can now be computed as

$$P[W^{(1)} > t] = \sum_{q=0}^{\infty} P[Q^{(1)} = (q, i)]P[X^{(q*)} + R_i > t]$$

$$+ \sum_{q=0}^{\infty} P[Q^{(2)} = (q, i)]P[X^{(q*)} + X_1 + R_i > t], \tag{18}$$

where R_i is a random variable with phase-type distribution (e_i, S), $X^{(q*)}$ is the sum of q independent copies of the job size with phase-type representation (α, S) and X_1 is the type-1 job size with phase-type representation (α_1, S_1).

The probabilities $P[X^{(q*)} + R_i > t]$ and $P[X^{(q*)} + X_1 + R_i > t]$ can be expressed using the following two observations. $P[X^{(q*)} + R_i > t]$ is the probability that there are less than $q + 1$ renewals in $[0, t]$ for the phase-type renewal process with inter-renewal time (α, S) that starts in phase i at time zero. The probability $P[X^{(q*)} + X_1 + R_i > t]$ can be expressed as $P[X^{(q*)} + R_i > t] + P[X^{(q*)} + R_i < t, X^{(q*)} + R_i + X_1 > t]$. Let $P(k, t)$ be the matrix such that entry (i, j) contains the probability that k renewals occur in $[0, t]$ given that the initial phase equals i and the phase at

time t is j of the phase-type renewal process with inter-renewal time (α, S). Then,

$$\sum_{q=1}^{\infty} P[Q^{(1)} = (q,i)]P[X^{(q*)} + R_i > t] = \sum_{q=1}^{\infty} \pi_1^{(1)} R^{q-1} \sum_{k=0}^{q} P(k,t)\mathbf{1}$$

$$= \pi_1^{(1)} \sum_{q=1}^{\infty} R^{q-1} \sum_{k=0}^{q-1} P(k,t)\mathbf{1} + \pi_1^{(1)} R^{-1} \sum_{q=1}^{\infty} R^q P(q,t)\mathbf{1}$$

$$= \pi_1^{(1)} (I-R)^{-1} \sum_{k=0}^{q-1} R^k P(k,t)\mathbf{1} + \pi_1^{(1)} R^{-1} \left(\sum_{k=0}^{\infty} R^k P(k,t)\mathbf{1} - P(0,t)\mathbf{1} \right)$$

$$= \pi_1^{(1)} ((I-R)^{-1} + R^{-1}) \sum_{k=0}^{\infty} R^k P(k,t)\mathbf{1} - \pi_1^{(1)} R^{-1} e^{St} \mathbf{1}.$$

Using the same reasoning as in the proof of [12, Theorem 2], we can show that

$$a \sum_{k=0}^{\infty} R^k P(k,t) b = (b^\top \otimes a) e^{(S^\top \otimes I + (s^*\alpha)^\top \otimes R)t} \xi, \tag{19}$$

for any row vector a and column vector b as

$$\frac{\partial}{\partial t} P(0,t) = SP(0,t),$$

$$\frac{\partial}{\partial t} P(k,t) = SP(k,t) + s^*\alpha P(k-1,t),$$

with $P(0,0) = I$ and $P(k,0) = 0$ for $k > 0$. Setting $b = \mathbf{1}$ and $a = \pi_1^{(1)} ((I-R)^{-1} + R^{-1})$ implies that

$$\sum_{q=0}^{\infty} P[Q^{(1)} = (q,i)]P[X^{(q*)} + R_i > t] = (\pi_0^{(1)} - \pi_1^{(1)} R^{-1}) e^{St} \mathbf{1}$$

$$+ (\mathbf{1}^\top \otimes \pi_1^{(1)} ((I-R)^{-1} + R^{-1})) e^{(S^\top \otimes I + (s^*\alpha)^\top \otimes R)t} \xi. \tag{20}$$

For the second sum in (18) it is worth noting that the events $X^{(q*)} + R_i \leq t$ and $X^{(q*)} + X_1 + R_i > t$ occur simultaneously if there are exactly $q+1$ renewals for the phase-type renewal process in $[0,t]$ starting from phase i given that the $(q+2)$-th inter-renewal time corresponds to a type-1 job. In other words $P[X^{(q*)} + R_i \leq t, X^{(q*)} + X_1 + R_i > t] = \sum_{j=1}^{n_1} P(q+1,t)_{ij}/p$ as the first n_1 phases correspond to a type-1 job and the fraction of type-1 jobs equals p. With this observation we can express the second sum in (18) as

$$\sum_{q=0}^{\infty} P[Q^{(2)} = (q,i)]P[X^{(q*)} + X_1 + R_i > t]$$

$$= \sum_{q=0}^{\infty} \sum_i (\pi_0^{(2)} R^q)_i \left(P[X^{(q*)} + R_i > t] + P[X^{(q*)} + R_i \leq t, X^{(q*)} + X_1 + R_i > t] \right)$$

$$= \sum_{q=0}^{\infty} \pi_0^{(2)} R^q \left(\sum_{k=0}^{q} P(k,t)\mathbf{1} + P(q+1,t) \binom{1/p}{0} \right)$$

$$= \pi_0^{(2)} (I-R)^{-1} \sum_{k=0}^{\infty} R^k P(k,t)\mathbf{1} + \pi_0^{(2)} R^{-1} \left(\sum_{k=0}^{\infty} R^k P(k,t) - P(0,t) \right) \binom{1/p}{0}.$$

Due to (19) we have

$$\sum_{q=0}^{\infty} P[Q^{(2)} = (q, i)] P[X^{(q*)} + X_1 + R_i > t]$$

$$= [(\mathbf{1}^{\top} \otimes \pi_0^{(2)} (I - R)^{-1}) + ((\mathbf{1}^{\top}/p, 0) \otimes \pi_0^{(2)} R^{-1})] e^{(S^{\top} \otimes I + (s^*\alpha)^{\top} \otimes R)t} \xi - \pi_0^{(2)} R^{-1} e^{St} \begin{pmatrix} 1/p \\ 0 \end{pmatrix}.$$

$$(21)$$

Combining (18), (20) and (21) yields (14).

As the workload decays at rate θ_Z and $\pi_1 \sum_{k=0}^{\infty} R^k P(k, t) \mathbf{1}$ is the probability that the workload exceeds t, (19) implies that $e^{(S^{\top} \otimes I + (s^*\alpha)^{\top} \otimes R)t}$ decays at rate θ_Z, while the other terms decay faster. The expression for $c_{W^{(1)}}$ can be derived by noting that for any vector a we have

$$a \sum_{k=0}^{q} P(k, t) \mathbf{1} = (a, 0) e^{\Omega t} \mathbf{1},$$

where the block matrix

$$\Omega = \begin{bmatrix} S & s^*\alpha & & \\ & S & \ddots & \\ & & \ddots & s^*\alpha \\ & & & S \end{bmatrix},$$

has $q + 1$ diagonal blocks. We therefore have that $a \sum_{k=0}^{q} P(k, t) \mathbf{1}$ for any fixed q has a decay rate equal to $\min(\theta_1, \theta_2)$. The probability $P[R_i > t]$ also decays at this rate. We may therefore when computing $\lim_{t \to \infty} e^{\theta_Z t} P[W^{(1)} > t]$ start the summations in (18) at K. This means that as far as this limit is concerned, a tagged type-1 job passes N type-2 jobs, where N is a truncated geometric distribution, that is, $P[N = k] = p(1 - p)^k$ for $k < K$ and $P[N = K] = (1 - p)^K$. Applying the final value theorem in the same manner as before therefore yields

$$c_{W^{(1)}} = c_Z (1 - p)^K \tilde{S}_2(-\theta_Z)^{-K} \frac{\tilde{S}_2(-\theta_Z)^K}{\tilde{S}(-\theta_Z)^K} + c_Z \sum_{k=0}^{K-1} p(1 - p)^k \tilde{S}_2(-\theta_Z)^{-k} \frac{\tilde{S}_2(-\theta_Z)^k \tilde{S}_1(-\theta_Z)}{\tilde{S}(-\theta_Z)^{k+1}}$$

$$= c_Z \left((1 - p)^K \tilde{S}(-\theta_Z)^{-K} + p \frac{\tilde{S}_1(-\theta_Z)}{\tilde{S}(-\theta_Z)} \frac{1 - (1 - p)^K \tilde{S}(-\theta_Z)^{-K}}{1 - (1 - p)\tilde{S}(-\theta_Z)^{-1}}, \right),$$

where the fraction $c_Z \tilde{S}_2(-\theta_Z)^K / \tilde{S}(-\theta_Z)^K$ is used to get the workload conditioned on having K type-2 jobs in the back, while the fractions $c_Z \tilde{S}_2(-\theta_Z)^k \tilde{S}_1(-\theta_Z) / \tilde{S}(-\theta_Z)^{k+1}$ are needed to get the workload conditioned on the fact that the last $k + 1$ jobs are a type-1 job followed by k type-2 jobs. □

THEOREM 7. *Let $R^{(1)}$ be the response time distribution of a type-1 job, then*

$$P[R^{(1)} > t] = P[W^{(1)} > t] + (1 - \lambda)\alpha_1 e^{S_1 t} \mathbf{1} - (v_1, 0) e^{A^{(1)} t} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$- (\pi_0^{(1)} - \pi_1^{(1)} R^{-1}, 0) e^{A^{(3)} t} \begin{pmatrix} 0 \\ 1 \end{pmatrix} + (\pi_0^{(2)} R^{-1}, 0) e^{A^{(2)} t} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad (22)$$

where

$$A^{(1)} = \begin{bmatrix} S^\top \otimes I + (s^*\alpha)^\top \otimes R & (S^\top \otimes I + (s^*\alpha)^\top \otimes R)\xi\alpha_1 \\ 0 & S_1 \end{bmatrix},$$

$$A^{(2)} = \begin{bmatrix} S & (Se)\alpha_1 \\ 0 & S_1 \end{bmatrix},$$

$$A^{(3)} = \begin{bmatrix} S & S\begin{pmatrix} e/p \\ 0 \end{pmatrix}\alpha_1 \\ 0 & S_1 \end{bmatrix}.$$

Further, $-\lim_{t\to\infty} \frac{1}{t} \log P[R^{(1)} > t] = \theta_Z$ *and* $c_{R^{(1)}} = \lim_{t\to\infty} e^{\theta_Z t} P[R^{(1)} > t] = c_{W^{(1)}} \tilde{S}_1(-\theta_Z).$

PROOF. The result follows from Theorem 6 and the fact that $P[R^{(1)} > t]$ can be expressed as

$$P[R^{(1)} > t] = (1 - \lambda)\alpha_1 e^{S_1 t}\mathbf{1} + P[W^{(1)} > t] + \int_0^t \left(-\frac{\partial}{\partial t} P[W^{(1)} > s]\right)\alpha_1 e^{S_1(t-s)}\mathbf{1}ds,$$

combined with Lemma 1.   □

7 ASYMPTOTIC TAIL IMPROVEMENT RATIO

In this section we present results for the asymptotic tail improvement ratio. Most of the results are expressed in terms of $\tilde{S}(-\theta_Z)$, $\tilde{S}_1(-\theta_Z)$, and $\tilde{S}_2(-\theta_Z)$, where θ_Z is the decay rate of the workload Z. It is worth noting at this stage that

$$\tilde{S}(-\theta_Z) = \frac{\lambda + \theta_Z}{\lambda}, \tag{23}$$

which follows from [2, Equation (4)] by noting that $\lambda/(\lambda + \theta_Z)$ is the Laplace transform of the inter-arrival time evaluated in $s = \theta_Z$.

The previous theorems yield the following result for the asymptotic tail improvement ratio:

THEOREM 8. *The asymptotic tail improvement ratio (ATIR) is equal to*

$$\begin{aligned} ATIR(K) &= 1 - \lim_{t\to\infty} \frac{P[R_{Nudge\text{-}K} > t]}{P[R > t]} \\ &= w_1(\tilde{S}_2(-\theta_Z) - 1)w\frac{1 - w^K}{1 - w} - (1 - w_1)(\tilde{S}_1(-\theta_Z) - 1)(1 - (1 - p)^K) \end{aligned} \tag{24}$$

with $w_1 = p\tilde{S}_1(-\theta_Z)/\tilde{S}(-\theta_Z) \in (0, 1)$, $w = (1 - p)/\tilde{S}(-\theta_Z) \in (0, 1)$ *and* $w_1 + w \in (0, 1)$.

Further, the integer K that maximizes ATIR(K) is given by

$$K_{opt} = \left\lfloor \log\left(\frac{\tilde{S}_1(-\theta_Z)(\tilde{S}_2(-\theta_Z) - 1)}{\tilde{S}_2(-\theta_Z)(\tilde{S}_1(-\theta_Z) - 1)}\right) \middle/ \log \tilde{S}(-\theta_Z) \right\rfloor, \tag{25}$$

if $K_{opt} \geq 0$, otherwise setting $K = 0$ is optimal.

PROOF. By definition

$$\text{ATIR}(K) = 1 - \frac{pc_{R^{(1)}} + (1 - p)c_{R^{(2)}}}{c_{FCFS}} = 1 - \frac{pc_{W^{(1)}}\tilde{S}_1(-\theta_Z) + (1 - p)c_{W^{(2)}}\tilde{S}_2(-\theta_Z)}{c_Z \tilde{S}(-\theta_Z)}.$$

Combined with (8) and (15), this yields

$$\text{ATIR}(K) = 1 - p\left(w^K + w_1\frac{1-w^K}{1-w}\right)\frac{\tilde{S}_1(-\theta_Z)}{\tilde{S}(-\theta_Z)}$$

$$- (1-p)\left((1-p)^K + (1-(1-p)^K)\tilde{S}_1(-\theta_Z)\right)\frac{\tilde{S}_2(-\theta_Z)}{\tilde{S}(-\theta_Z)},$$

$$= (1-w_1) + w_1(1-w^K) - w_1^2\frac{1-w^K}{1-w}$$

$$- (1-w_1)\left((1-p)^K + (1-(1-p)^K)\tilde{S}_1(-\theta_Z)\right),$$

as $1-w_1 = (1-p)\tilde{S}_2(-\theta_Z)/\tilde{S}(-\theta_Z)$. (24) now follows by verifying that $1-w-w_1 = w(\tilde{S}_2(-\theta_Z)-1)$. To see that $w \in (0,1)$, it suffices to note that $\tilde{S}(-\theta_Z) > 1$ as $\theta_Z > 0$. Similarly $\tilde{S}_i(-\theta_Z) > 1$, which implies that $w_1 \in (0,1)$ as $\tilde{S}(-\theta_Z) = p\tilde{S}_1(-\theta_Z) + (1-p)\tilde{S}_2(-\theta_Z)$. Further, $w + w_1 = ((1-p) + p\tilde{S}_1(-\theta_Z))/\tilde{S}(-\theta_Z)) < 1$.

Setting the derivative of the ATIR(K) equal to zero implies

$$\frac{(1-p)^K}{w^K} = \frac{w_1}{1-w_1}\frac{w}{1-w}\frac{\log w}{\log(1-p)}\frac{\tilde{S}_2(-\theta_Z)-1}{\tilde{S}_1(-\theta_Z)-1},$$

which shows that the ATIR(K) has a unique stationary point.

Define $\Delta_{\text{ATIR}}(K) = \text{ATIR}(K+1) - \text{ATIR}(K)$. Recalling that $1-w_1 = (1-p)\tilde{S}_2(-\theta_Z)/\tilde{S}(-\theta_Z)$, we have

$$\Delta_{\text{ATIR}}(K) = w_1(\tilde{S}_2(-\theta_Z)-1)w^{K+1} - p(1-w_1)(\tilde{S}_1(-\theta_Z)-1)(1-p)^K,$$

$$= \left(\frac{\tilde{S}_1(-\theta_Z)(\tilde{S}_2(-\theta_Z)-1)}{\tilde{S}(-\theta_Z)^{K+2}} - \frac{\tilde{S}_2(-\theta_Z)(\tilde{S}_1(-\theta_Z)-1)}{\tilde{S}(-\theta_Z)}\right)p(1-p)^{K+1}. \tag{26}$$

As $\lim_{K\to-\infty}\Delta_{\text{ATIR}}(K) = +\infty$ and $\Delta_{\text{ATIR}}(K) < 0$ for K sufficiently large, the unique stationary point of ATIR(K) is a maximum and the optimal integer K is located in the ceil of the unique root of $\Delta_{\text{ATIR}}(K)$, which yields (25). □

Remarks: The expression in (26) shows that the increase in the ATIR(K) decreases with K as long as it remains positive. Thus the gain obtained by increasing K by one decreases with K until the optimal K is reached.

It is worth noting that if the type-i jobs have an exponential job size distribution with parameter μ_i, then (25) simplifies to

$$K_{opt} = \lfloor\log(\mu_1/\mu_2)/\log(\tilde{S}(-\theta_Z))\rfloor, \tag{27}$$

which implies that K_{opt} is non-decreasing in λ when the job sizes are exponential and $\mu_1/\mu_2 = E[X_2]/E[X_1] > 1$ (as $\tilde{S}(-\theta_Z)$ decreases in λ). As we will demonstrate in Section 8, this property does not necessarily hold when the type-2 jobs are no longer exponential even if X_2 stochastically dominates X_1.

THEOREM 9. *The ATIR(K) > 0 if and only if*

$$\left(1 - \frac{1}{\tilde{S}_2(-\theta_Z)}\right)\bigg/\left(1 - \frac{1}{\tilde{S}_1(-\theta_Z)}\right) > \left(1 + \frac{\theta_Z}{\lambda p}\right)\frac{1-(1-p)^K}{1 - \left(\frac{\lambda(1-p)}{\lambda+\theta_Z}\right)^K}, \tag{28}$$

meaning the ATIR(1) > 0 if and only if

$$\left(1 - \frac{1}{\tilde{S}_2(-\theta_Z)}\right) \bigg/ \left(1 - \frac{1}{\tilde{S}_1(-\theta_Z)}\right) > 1 + \frac{\theta_Z}{\lambda}, \tag{29}$$

and the ATIR(K) > 0 for any K if and only if

$$\left(1 - \frac{1}{\tilde{S}_2(-\theta_Z)}\right) \bigg/ \left(1 - \frac{1}{\tilde{S}_1(-\theta_Z)}\right) > 1 + \frac{\theta_Z}{\lambda p}. \tag{30}$$

PROOF. From (24) and the definition of w_1, we have ATIR(K) > 0 if and only if

$$p\tilde{S}_1(-\theta_Z)(\tilde{S}_2(-\theta_Z) - 1)w\frac{1 - w^K}{1 - w} > (1 - p)\tilde{S}_2(-\theta_Z)(\tilde{S}_1(-\theta_Z) - 1)(1 - (1 - p)^K).$$

This condition can be restated as

$$\left(1 - \frac{1}{\tilde{S}_2(-\theta_Z)}\right) \bigg/ \left(1 - \frac{1}{\tilde{S}_1(-\theta_Z)}\right) > \frac{(1 - w)\tilde{S}(-\theta_Z)}{p}\frac{1 - (1 - p)^K}{1 - w^K}.$$

Using $w = (1 - p)/\tilde{S}(-\theta_Z)$ and (23) we obtain (28). Setting $K = 1$ and taking the limit for K to infinity yield (29) and (30). The result for $K = 1$ is also immediate from (26) as ATIR(1) = $\Delta_{\text{ATIR}}(0)$. $\square$

Remarks: The condition in (29) is very similar to Theorem 4.3 in [5]. Moreover when $K = 1$, we have

$$\begin{aligned}
\text{ATIR}(1) &= p(1 - p)\left(\frac{\tilde{S}_1(-\theta_Z)}{\tilde{S}(-\theta_Z)}\frac{\tilde{S}_2(-\theta_Z) - 1}{\tilde{S}(-\theta_Z)} - \frac{\tilde{S}_2(-\theta_Z)}{\tilde{S}(-\theta_Z)}(\tilde{S}_1(-\theta_Z) - 1)\right) \\
&= \frac{p(1 - p)}{\tilde{S}(-\theta_Z)}\left(\tilde{S}_2(-\theta_Z) - \frac{\tilde{S}_1(-\theta_Z)}{\tilde{S}(-\theta_Z)} - \tilde{S}_1(-\theta_Z)\tilde{S}_2(-\theta_Z)(1 - \tilde{S}(-\theta_Z)^{-1})\right) \\
&= \frac{\lambda p(1 - p)}{\lambda + \theta_Z}\left(\tilde{S}_2(-\theta_Z) - \frac{\lambda}{\lambda + \theta_Z}\tilde{S}_1(-\theta_Z) - \frac{\theta_Z}{\lambda + \theta_Z}\tilde{S}_1(-\theta_Z)\tilde{S}_2(-\theta_Z)\right),
\end{aligned}$$

which is again similar in form to Theorem 4.3 in [5].

When type-i jobs have an exponential distribution with mean $1/\mu_i$, for $i = 1, 2$, we have $\tilde{S}_i(s) = \mu_i/(\mu_i + s)$ and (29) simplifies to

$$\frac{\mu_1}{\mu_2} > \tilde{S}(-\theta_Z) = 1 + \theta_Z/\lambda.$$

As $\lim_{\lambda \to 1^-} \tilde{S}(-\theta_Z) = 1$, this means that for $\mu_1/\mu_2 = E[X_2]/E[X_1] > 1$, there exists a $\bar{\lambda} \in (0, 1)$ such that ATIR(1) > 0 for $\lambda > \bar{\lambda}$. The next theorem shows that this result holds in general for phase-type distributions:

THEOREM 10. *When $E[X_2] > E[X_1]$, then ATIR(K) > 0 for λ sufficiently close to one for any K. Further, if $E[X_2] > E[X_1]$, then $\lim_{\lambda \to 1^-} K_{opt} = \infty$.*

PROOF. Using the Taylor series expansion of the exponential function, we readily see that

$$\tilde{S}_i(-\theta_Z) = \sum_{k=0}^{\infty} \frac{\theta_Z^k E[X_i^k]}{k!} = 1 + \theta_Z E[X_i] + o(\theta_Z^2), \tag{31}$$

for $i = 1, 2$ and similarly $\tilde{S}(-\theta_Z) = 1 + \theta_Z E[X] + o(\theta_Z^2)$. This implies that

$$\lim_{\lambda \to 1^-} \left(1 - \frac{1}{\tilde{S}_2(-\theta_Z)}\right) \bigg/ \left(1 - \frac{1}{\tilde{S}_1(-\theta_Z)}\right) = \frac{E[X_2]}{E[X_1]}. \tag{32}$$
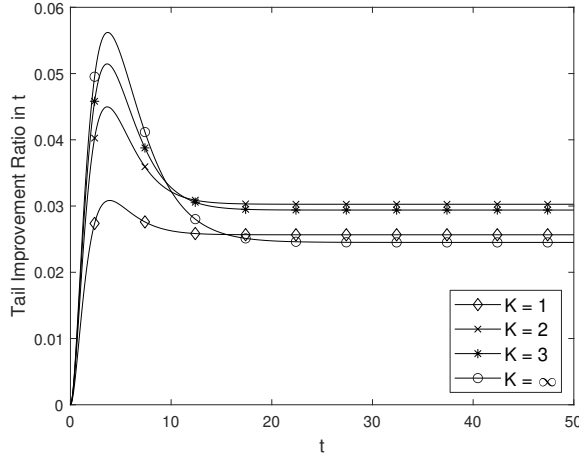
Fig. 1. The tail improvement ratio of Nudge-K over FCFS for expo jobs with $E[X_2]/E[X_1] = 2, \lambda = 3/4$ and $p = 1/2$.

as θ_Z tends to zero when λ tends to one. Hence, if $E[X_2]/E[X] > 1$, then ATIR$(K) > 0$ for any K for λ close enough to 1 due to (30) as $1 + \theta_Z/\lambda$ tends to one.

To determine $\lim_{\lambda \to 1^-} K_{opt}$ we note that (32) implies that

$$\lim_{\lambda \to 1^-} K_{opt} = \lim_{\lambda \to 1^-} \left\lfloor \frac{\log(E[X_2]/E[X_1])}{\log(\tilde{S}(-\theta_Z))} \right\rfloor. \tag{33}$$

This proves that K_{opt} tends to ∞ when $E[X_2]/E[X_1] > 1$, while K_{opt} tends to $-\infty$ if $E[X_2]/E[X_1] < 1$. □

THEOREM 11. *For λ close to one and $E[X_2] > E[X_1]$, we have*

$$K_{opt} \approx \left\lfloor \frac{\log(E[X_2]/E[X_1])E[X^2]}{2(1-\lambda)} \right\rfloor \approx \left\lfloor \log\left(\frac{E[X_2]}{E[X_1]}\right) E[Z] \right\rfloor, \tag{34}$$

where $\log()$ *is the natural logarithm.*

PROOF. Using (33) and the fact that $\tilde{S}(-\theta_Z) = 1 + \theta_Z + o(\theta_Z^2)$ (as $E[X] = 1$), we have for λ close to one

$$K_{opt} \approx \left\lfloor \frac{\log(E[X_2]/E[X_1])}{\log(1+\theta_Z)} \right\rfloor.$$

The result therefore follows from the classic heavy-traffic limit of Kingman [7] for the GI/G/1 queue (when both the inter-arrival time I and service time distribution X has finite variance), which states that the decay rate θ_Z of the waiting time in the heavy traffic limit equals

$$\frac{2(E[I] - E[X])}{Var[I] + Var[X]} = \frac{2(\frac{1}{\lambda} - 1)}{\frac{1}{\lambda^2} + pE[X_1^2] + (1-p)E[X_2^2] - 1}, \tag{35}$$

and the Taylor series expansion of $\log(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} x^n/n$. The expression using $E[Z]$ is due to the fact that $E[Z] = \lambda E[X^2]/(2(1-\lambda))$. □
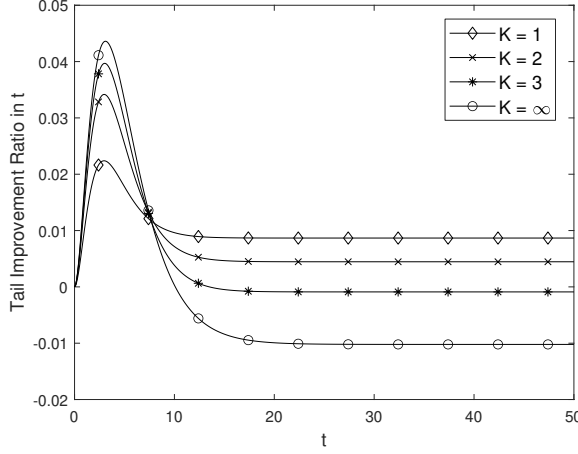
Fig. 2. The tail improvement ratio of Nudge-K over FCFS for expo jobs with $E[X_2]/E[X_1] = 3/2, \lambda = 3/4$ and $p = 1/2$.

8 NUMERICAL RESULTS AND INSIGHTS

In this section we present various new insights on stochastic and asymptotic improvements of FCFS. We start with the stochastic improvements.

8.1 On stochastic improvements

Figure 1 plots the tail improvement ratio of Nudge-K over FCFS for $\lambda = 3/4$ when both type-1 and type-2 jobs have an exponential distribution with $E[X_2]/E[X_1] = 2$ and $p = 1/2$. A number of observations can be made:

(1) Nudge-K stochastically improves FCFS for $K = 1, 2, 3$ and ∞, even though type-1 jobs are not necessarily smaller than type-2 jobs. This illustrates that FCFS can be stochastically improved upon under far weaker conditions than in [5].

(2) Nudge-2 and Nudge-3 both stochastically improve Nudge-1, but neither stochastically improves the other. From (25) we also know that setting $K = 2$ optimizes the ATIR(K). This implies that there does not exist a K that minimizes $P[R_{\text{Nudge}-K} > t]$ for all t.

(3) Setting $K = \infty$ is best for reducing $P[R_{\text{Nudge}-K} > t]$ for small t, but does not stochastically improve Nudge-K for $K \in \{1, 2, 3\}$.

Figure 2 plots the tail improvement ratio of Nudge-K over FCFS for the same setting as Figure 1, except that $E[X_2]/E[X_1] = 3/2$. The main observation in this plot is that while $K = 1$ and 2 results in a stochastic improvement over FCFS, setting $K \geq 3$ does not (as the ATIR(K) decreases in K beyond K_{opt}). In other words, *in some cases the stochastic improvement over FCFS can be lost if we allow that type-1 jobs can pass too many type-2 jobs.*

Figure 3 considers the scenario with $p = 0.7$, $\lambda = 0.7$, $E[X_2]/E[X_1] = 1.2$. Type-1 jobs are exponential, while the type-2 jobs follow an order-2 hyper-exponential distribution with $SCV = 2$ and shape parameter $f = 9/10$. This means that type-2 jobs are a mixture of two classes of exponential jobs: one with mean $1/\mu_{21} \approx 1.034$ and one with mean $1/\mu_{22} \approx 7.674$, where 10% of the workload is offered by the jobs belonging to the class with the larger mean (that is, $\alpha_2 \approx (0.985, 0.015)$). It is important to note that both $1/\mu_{21}$ and $1/\mu_{22}$ are larger than 1, while the mean of
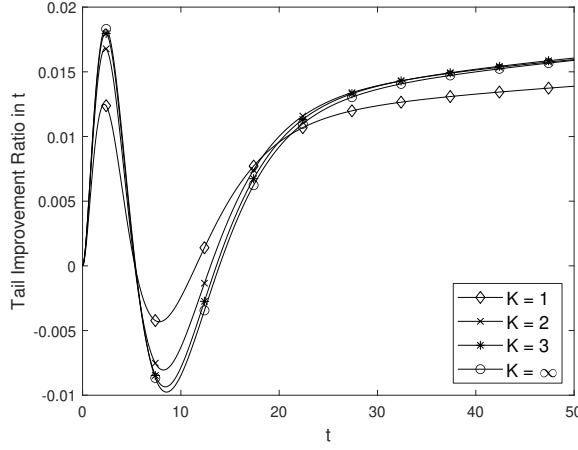
Fig. 3. The tail improvement ratio of Nudge-K over FCFS for expo type-1 and H2 ($SCV = 2, f = 9/10$) type-2 jobs with $E[X_2]/E[X_1] = 1.2, \lambda = 0.7$ and $p = 0.7$.

the type-1 jobs is $100/106 < 1$. As a result X_2 stochastically dominates X_1 (in the first order), that is

$$P[X_2 > t] > P[X_1 > t],$$

for any $t > 0$. The plot shows that while all the considered K values result in an *asymptotic* tail improvement ratio, none of them stochastically improves FCFS. This example shows that *having an asymptotic improvement does not imply a stochastic improvement in general even if X_2 stochastically dominates X_1*. The intuition is that while it is good to swap type-1 jobs with the type-2 jobs with mean $1/\mu_{22}$, the swaps with the jobs with mean $1/\mu_{21}$ are not beneficial as their mean is fairly close to the mean of the type-1 jobs.

In Figure 4 we consider the same scenario as in Figure 3 with $K = 1$, but we vary the SCV of the H2 type-2 traffic from 1 to 100. For all of these settings, X_2 stochastically dominates X_1. The experiment shows that while there is no stochastic or asymptotic improvement for low SCV (that is, when the SCV equals 1 or 1.2). Larger SCV values do result in an asymptotic improvement, but not in a stochastic improvement. We further see that the asymptotic improvement tends to zero as the SCV tends to infinity. This can be understood by noting that when the SCV tends to infinity, there are very few swaps as the type-2 jobs that have an exponential distribution with mean $1/\mu_{22}$ are rare (as $1/\mu_{22}$ is large).

8.2 On asymptotic tail improvements

In this subsection we address two issues. First, we noted that if both the type-1 and type-2 jobs are exponential, then K_{opt} is non-decreasing in λ if $E[X_2] > E[X_1]$ (see (27)). We now demonstrate that if we make the type-2 jobs hyper-exponential, this is not necessarily the case even if X_2 stochastically dominates X_1. Second, for exponential job sizes having $E[X_1] > E[X_2]$ implies that $K_{opt} = 0$, meaning we cannot asymptotically improve upon FCFS. We illustrate that when type-2 jobs are no longer exponential and $E[X_1] > E[X_2]$, we can still achieve an asymptotic tail improvement in some cases.

The scenario considered is similar to Figure 3, that is, $p = 0.7$, type-1 jobs are exponential and type-2 jobs are hyper-exponential with $SCV = 2$ and shape parameter $f = 0.9$. We vary λ from 0.01 to 0.99 and $E[X_2]/E[X_1]$ from 0.4 to 2 (in Figure 3 these were set at 0.7 and 1.2, respectively).
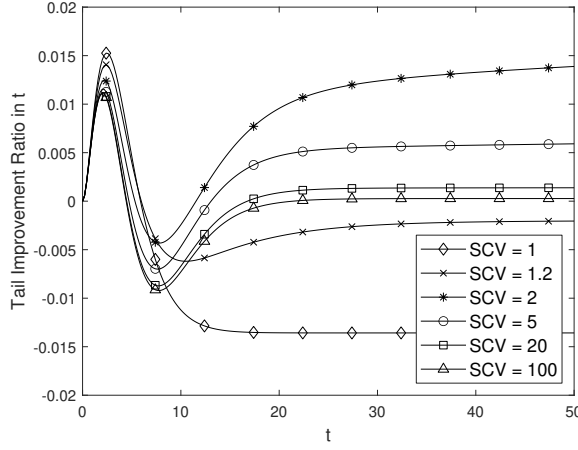
Fig. 4. The tail improvement ratio of Nudge-K with $K = 1$ over FCFS for expo type-1 and H2 ($f = 9/10$) type-2 jobs with $E[X_2]/E[X_1] = 1.2, \lambda = 0.7$ and $p = 0.7$.
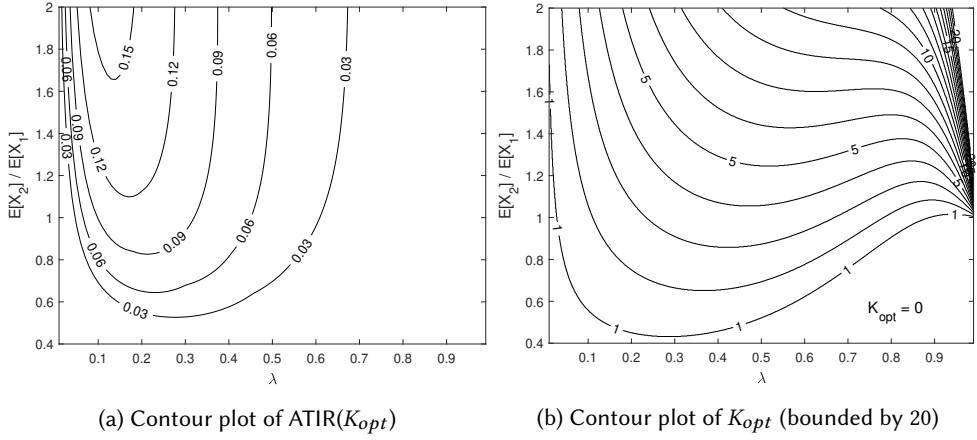

(a) Contour plot of ATIR(K_{opt})

(b) Contour plot of K_{opt} (bounded by 20)

Fig. 5. The asymptotic tail improvement ratio of Nudge-K over FCFS (a) and optimal K (b) for $p = 0.7$ with exponential type-1 and hyper-exponential type-2 jobs ($SCV = 2$ and $f = 9/10$).

Figure 5 presents two contour plots: one for the ATIR(K_{opt}) with contour lines from 0.03 to 0.15 in steps of 0.03 and one for K_{opt} with contour lines in $1, 2, \ldots, 20$.

Looking at the region where $E[X_2]/E[X_1] < 1$ clearly shows that we can have an asymptotic tail improvement even when $E[X_1] > E[X_2]$. If we focus on the line with $E[X_2]/E[X_1] = 1.2$ in the contour plot of K_{opt}, we note that K_{opt} first increases to 4, then drops to 3 and finally starts to increase (without bound due to Theorem 10) as λ tends to one. This shows that K_{opt} can decrease as a function of λ. Further note that when $E[X_1]/E[X_2] = 1.2$, then X_2 stochastically dominates X_1 (as explained when discussing Figure 3).

9 CONCLUSIONS

In this paper we demonstrated that the First-Come-First-Served scheduling algorithm can be stochastically improved upon by the Nudge-K algorithm under far weaker conditions that the ones

considered in [5]. This is practically relevant as it indicates that it may suffice to identify certain job types, where jobs belonging to one type are typically larger than jobs belonging to another, in order to improve all of the response time percentiles of First-Come-First-Served scheduling. We did this by deriving explicit expressions for the response time of Nudge-K for the system defined in Section 2. In addition we presented a number of elegant results on the asymptotic tail improvement ratio, such as the expression for the optimal K. We also presented various new insights on stochastic and asymptotic improvements upon First-Come-First-Served scheduling.

The current results can be extended in a number of ways. It is not too difficult to include a third job type that cannot be swapped with either type-1 or type-2 jobs, though we expect smaller gains in such a scenario and the notations become somewhat heavier. The results on the ATIR(K) presented in Section 7 and their proofs remain mostly valid if we relax the assumptions that type-1 and type-2 job sizes are phase-type and simply demand that $X = pX_1 + (1-p)X_2$ is phase-type. In such case we could also define the type-1 and type-2 jobs using the job size (as in [5]).

It is also worthwhile seeing what happens if we consider a larger class of Nudge-like policies, for instance, suppose that we allow that type-2 jobs can be involved in at most L swaps instead of just one. When $L > 1$, the analysis performed for the type-1 jobs in Section 6 fails as we can no longer make a similar connection with the FCFS queue. We believe that in such case it is possible to rely on the framework of Markov modulated fluid queues to study the response time of a type-1 job. In fact initially we used this approach for the analysis of the response time of a type-1 job of Nudge-K before coming up with the more elegant approach presented in Section 6. We note that both the Markov modulated fluid queue approach and the more elegant approach in Section 6 yielded the same numerical results.

ACKNOWLEDGEMENT

REFERENCES

[1] J. Abate, L.G. Choudhury, and W. Whitt. Waiting-time tail probabilities in queues with long-tail service-time distributions. *Queueing Systems*, 16:311––338, 1994.

[2] J. Abate, L.G. Choudhury, and W. Whitt. Exponential approximations for tail probabilities in queues, i: Waiting times. *Operations Research*, 43:885–901, 1995.

[3] O. Boxma and B. Zwart. Tails in scheduling. *SIGMETRICS Perform. Eval. Rev.*, 34(4):13–20, mar 2007.

[4] A. Feldmann and W. Whitt. Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, 31(3):245 – 279, 1998.

[5] I. Grosof, K. Yang, Z. Scully, and M. Harchol-Balter. Nudge: Stochastically improving upon FCFS. *Proc. ACM Meas. Anal. Comput. Syst.*, 5(2), jun 2021.

[6] D. Gross and C.M. Harris. *Fundamentals of Queueing Theory*. John Wiley and Sons, New York, 1974.

[7] J. F. C. Kingman. On queues in heavy traffic. *Journal of the Royal Statistical Society. Series B (Methodological)*, 24(2):383–392, 1962.

[8] J. Kriege and P. Buchholz. *PH and MAP Fitting with Aggregated Traffic Traces*, pages 1–15. Springer International Publishing, Cham, 2014.

[9] G. Latouche and V. Ramaswami. *Introduction to Matrix Analytic Methods and stochastic modeling*. SIAM, Philadelphia, 1999.

[10] M.F. Neuts. *Matrix-Geometric Solutions in Stochastic Models, An Algorithmic Approach*. John Hopkins University Press, 1981.

[11] M. Nuyens, A. Wierman, and B. Zwart. Preventing large sojourn times using smart scheduling. *Operations Research*, 56(1):88–101, 2008.

[12] Toshihisa Ozawa. Sojourn time distributions in the queue defined by a general QBD process. *Queueing Systems and its Applications*, 53(4):203–211, 2006.

[13] A. Panchenko and A. Thümmler. Efficient phase-type fitting with aggregated traffic traces. *Perform. Eval.*, 64(7-8):629–645, August 2007.

[14] E. Seneta. *Non-negative Matrices and Markov Chains*. Springer-Verlang, New York, 1981.

[15] C. Van Loan. Computing integrals involving the matrix exponential. *IEEE Transactions on Automatic Control*, 23(3):395–404, 1978.

[16] A. Wierman and B. Zwart. Is tail-optimal scheduling possible? *Operations Research*, 60(5):1249–1257, 2012.