On the Maximum Queue Length of the Hyper Scalable Load Balancing Push Strategy

Benny Van Houdt[0000−0002−5955−8493]

Department of Computer Science, University of Antwerp - imec, Belgium

Abstract. In this paper we derive explicit and structural results for the steady state probabilities of a structured finite state Markov chain. The study of these steady state probabilities is motivated by the analysis of the hyper scalable load balancing push strategy when using the queue-at-the-cavity approach. More specifically, these probabilities can be used to determine the largest possible arrival rate that can be supported by this strategy without exceeding some predefined maximum queue length. Contrary to prior work, we study the push strategy when the queue length information updates occur according to a phase-type renewal process with non-exponential inter-renewal times.

Keywords: load balancing · hyper scalable push strategy · bounded queue length.

1 Introduction

Hyper scalable load balancing strategies for large-scale systems have received considerable attention recently [1,2,3]. These strategies further reduce the communication overhead of traditional load balancers such as the power-of-d-choices or join-idle-queue strategies [4,5,6,7], the overhead of which equals at least one message per job. One of the most fundamental hyper scalable load balancing strategies is the push strategy studied in [1,8]. It operates as follows. There is a single dispatcher that maintains an estimate on the queue length of each server. Jobs arrive at the dispatcher at rate λN, where N is the number of servers. Incoming jobs are assigned in a greedy manner, that is, the dispatcher assigns the job to a server with the smallest estimated queue length among all servers and increases its estimate by one. However, the dispatcher is not informed about job completions. Instead the queue length estimates are updated to their actual values at random points in time (meaning the estimates are upper bounds). Whether these queue length updates are triggered by the dispatcher or the servers does not matter in such case. The mean number of updates that occur per incoming job is a control parameter that can be set well below one.

An effective approach to study the performance of large-scale load balancing strategies is the so-called *queue-at-the-cavity* approach [9], which reflects the system behavior as the number of servers tends to infinity assuming asymptotic independence. For the hyper scalable push strategy described above, the corresponding queue-at-the-cavity has a bounded queue length m, the value of which

can be determined by studying a structured finite state Markov chain in case of phase-type distributed job sizes [8]. The maximum queue length m grows as a function of the arrival rate λ and explicit results for the largest possible arrival rate $\lambda(m) \in (0,1)$ that can be supported with a given maximum queue length m were presented in [8] in case of phase-type distributed job sizes with mean one and random server queue length updates. Random updates imply that the time between two updates of the queue length information of a tagged server follows an exponential distribution with some mean $1/\delta$. In this paper we derive explicit results for $\lambda(m)$ when the updates of the queue length of a tagged server follow a renewal process, but the inter-renewal time is not necessarily exponential.

The paper is structured as follows. Section 2 contains the problem statement. Results for the case with exponential job sizes are presented in Section 3, while in Section 4 we focus on non-exponential job sizes. Conclusions are found in Section 5.

2 Problem statement

We assume that the job size distribution Z follows an order k_S phase type distribution characterized by (α, S) with mean $\alpha(-S)^{-1}e = 1$, that is, $P[Z > t] = \alpha \exp(St)e$, where e is a vector of ones of the appropriate size. The time between two updates X of a tagged server follows an order k_T phase type distribution characterized by (β, T) with mean $\beta(-T)^{-1}e = 1/\delta > 1$, meaning $P[X > t] = \beta \exp(Tt)e$. In other words, whenever the dispatcher updates its queue length information for the tagged server, a phase-type distributed timer is started and the dispatcher receives a new update each time this timer expires. Let $t^* = (-T)e$ and $s^* = (-S)e$. All timers and job sizes are assumed to be independent.

In [8, Section 4] it was shown that using the queue-at-the-cavity approach in case of *random updates*, the maximum queue length for the hyper scalable push strategy is the smallest m value such that the steady-state probability of being in the first state of the $1 + k_S m$ state Markov chain with the following rate matrix is less than $1 - \lambda$:

$$\begin{bmatrix} -\delta & & & & & \delta\alpha \\ s^* & S-\delta I & & & & \delta I \\ & s^*\alpha & S-\delta I & & & \delta I \\ & & \ddots & \ddots & & \vdots \\ & & & s^*\alpha & S-\delta I & \delta I \\ & & & & s^*\alpha & S \end{bmatrix}. \tag{1}$$

This Markov chain captures the evolution of the queue length of a tagged server that serves phase-type distributed jobs with representation (α, S) and that jumps up to length m each time an update occurs, where updates occur according to a Poisson process with rate δ.

It is not hard to see that when phase-type distributed timers are used instead to trigger updates, the maximum queue length of a server equals the smallest m

value such that the probability of being in a state part of Ω_0 is less than $1-\lambda$ in the Markov chain with state space

$$\Omega = \Omega_0 \cup (\cup_{\ell=1}^{m}\ \Omega_\ell)\,,$$

where $\Omega_0 = \{(0,j)|j=1,\ldots,k_T\}$ and $\Omega_\ell = \{(\ell,i,j)|i=1,\ldots,k_S, j=1,\ldots,k_T\}$ and rate matrix

$$Q(m) = \begin{bmatrix} T & & & & & \alpha\otimes t^*\beta \\ s^*\otimes I & S\oplus T & & & & I\otimes t^*\beta \\ & s^*\alpha\otimes I & S\oplus T & & & I\otimes t^*\beta \\ & & \ddots & \ddots & & \vdots \\ & & & s^*\alpha\otimes I & S\oplus T & I\otimes t^*\beta \\ & & & & s^*\alpha\otimes I & S\oplus(T+t^*\beta) \end{bmatrix}. \qquad (2)$$

If we set $T=-\delta$, meaning $t^*=\delta$, $\beta=1$ and $k_T=1$, $Q(m)$ coincides with (1).

We now formally define $\lambda(m)\in(0,1)$ as the largest possible arrival rate λ such that the maximum queue length of the queue-at-the-cavity of the push strategy is bounded by m given a mean job size equal to one.

Definition 1 *Let $\lambda(m)=1-\pi_0(m)$, where $\pi_0(m)$ is the steady probability to be in a state part of the set Ω_0 for the CTMC characterized by $Q(m)$.*

We use the following notations for some special cases:

1. If (β,T) has an Erlang-k_T distribution, we denote this rate as $\lambda_{k_T}(m)$.
2. If the job sizes are exponential, we denote this rate as $\lambda^{(exp)}(m)$.
3. If (β,T) has an Erlang-k_T distribution and job sizes are exponential, hyper exponential, Coxian or Erlang, we denote this rate as $\lambda_{k_T}^{(exp)}(m)$, $\lambda_{k_T}^{(HE)}(m)$, $\lambda_{k_T}^{(Cox)}(m)$ or $\lambda_{k_T}^{(Erl)}(m)$, respectively.

It is possible to develop an algorithm that runs in $O(k_T^3(k_S^3+\log m))$ time to numerically compute $\pi_0(m)$, and thus $\lambda(m)$, by exploiting the structure of $Q(m)$. The main objective of this paper is however to derive closed from results for $\lambda(m)$ and to establish structural results. A first set of closed form results was presented in [3, Corollary 4.5] for exponential timers, that is,

$$\lambda_1(m) = \frac{\delta(1-y(1)^m)}{\delta(1-y(1)^m)+y(1)^{m-1}(1-y(1))}, \qquad (3)$$

where $y(1)=P[Z<X]$ is the probability that the job size Z with $E[Z]=1$ is less than an exponential timer X with mean $1/\delta>1$. Note that $\lambda_1(1)$ simplifies to $\delta/(1+\delta)$ which is independent of the job size distribution Z. The case with $m=1$ is of particular interest as it corresponds to the case where the queue length is bounded by one in the large-scale limit (assuming asymptotic independence), which corresponds to so-called *vanishing waiting times*.

3 Exponential job sizes

In this section we study the arrival rate $\lambda^{(exp)}(m)$ that can be supported such that the maximum queue length is bounded by m in case of exponential job sizes and phase-type distributed timers. We make the following contributions:

1. We prove that $\lambda^{(exp)}(m)$ is maximized over all order k_T phase-type distributions by the Erlang-k_T distribution (due to Theorem 2).
2. We present an explicit formula for $\lambda^{(exp)}_{k_T}(m)$ and prove that $\lambda^{(exp)}_{k_T}(m)$ increases as a function of k_T (see Theorem 3).
3. We derive the limiting expressions as k_T tends to infinity (see Theorem 3).

When the job sizes are exponential the state space reduces to

$$\Omega^{(exp)} = \cup_{\ell=0}^{m} \Omega_\ell^{(exp)} = \cup_{\ell=0}^{m} \{(\ell, j) | j = 1, \ldots, k_T\}.$$

and the rate matrix $Q(m)$ simplifies to

$$Q^{(exp)}(m) = \begin{bmatrix} T & & & & & t^*\beta \\ I & T-I & & & & t^*\beta \\ & I & T-I & & & t^*\beta \\ & & \ddots & \ddots & & \vdots \\ & & & I & T-I & t^*\beta \\ & & & & I & (T+t^*\beta)-I \end{bmatrix}. \tag{4}$$

Theorem 1. *Let $N_1(X)$ denote the number of arrivals of a Poisson process with rate 1 during an (β, T) phase type distributed time X. Denote $\pi_0^{(exp)}(m)$ as the steady state probability that the state of the CTMC with rate matrix $Q^{(exp)}(m)$ is in the set $\Omega_0^{(exp)}$, then*

$$\pi_0^{(exp)}(m) = 1 - \delta E[\min(N_1(X), m)].$$

This implies that $\pi_0^{(exp)}(m)$ is decreasing in m.

Proof. For the CTMC with rate matrix $Q^{(exp)}(m)$ we clearly have a renewal whenever the (β, T) phase-type distributed timer expires. The mean length of a renewal cycle therefore equals $1/\delta$ and its length has an order $(m+1)k_T$ phase-type distribution with initial vector $(\beta, 0, \ldots, 0)$ and subgenerator matrix

$$Q^{(cycle)}(m) = \begin{bmatrix} T-I & I & & & & \\ & T-I & I & & & \\ & & T-I & I & & \\ & & & \ddots & \ddots & \\ & & & & T-I & I \\ & & & & & T \end{bmatrix}, \tag{5}$$

where we reordered the states. In order to express the mean time that the CTMC spends in the set $\Omega_\ell^{(exp)}$ during a cycle, we can focus on the first block row of the matrix $(-Q^{(cycle)}(m))^{-1}$. As $(-Q^{(cycle)}(m))^{-1}$ equals

$$\begin{bmatrix} (I-T)^{-1} & \dots & (I-T)^{-m} & (I-T)^{-m}(-T)^{-1} \\ & \ddots & \vdots & \vdots \\ & & (I-T)^{-1} & (I-T)^{-1}(-T)^{-1} \\ & & & (-T)^{-1} \end{bmatrix},$$

the mean time spend in the set $\Omega_\ell^{(exp)}$, with $\ell > 0$, per cycle can be expressed as $\beta(I-T)^{-(m-\ell+1)}e$ and therefore

$$\pi_0^{(exp)}(m) = \frac{1/\delta - \sum_{\ell=1}^{m} \beta(I-T)^{-(m-\ell+1)}e}{1/\delta} = 1 - \delta \sum_{\ell=1}^{m} \beta(I-T)^{-\ell}e.$$

Furthermore $\beta(I-T)^{-\ell}e$ is also the probability that $N_1(X)$ equals ℓ or more. Hence,

$$\pi_0^{(exp)}(m) = 1 - \delta \sum_{\ell=1}^{m} P[N_1(X) > \ell - 1] = 1 - \delta E[\min(N_1(X), m)].$$

□

Theorem 2. *Let $\pi_0^{(exp)}(m)$ be the steady state probability of the CTMC with rate matrix $Q^{(exp)}(m)$ to be in the set $\Omega_0^{(exp)}$, then $\pi_0^{(exp)}(m)$ is minimized over all order k_T phase type distributions (β, T) by the Erlang-k_T distribution. Moreover $\pi_0^{(exp)}(m)$ is decreasing in k_T when (β, T) is an Erlang-k_T distribution.*

Proof. Let X follow any order k_T phase type distribution characterized by (β, T) and let Y have an Erlang-k_T distribution. By Theorem 3 in [10] we have $Y \leq_{cx} X$ where $\leq_{cx}$ is the usual convex ordering between random variables with the same mean (see [11]). By Theorem 3.A.40 in [11] we therefore have $N_1(Y) \leq_{cx} N_1(X)$. Using 3.A.5 in [11], we have

$$E[\max(N_1(Y), m)] \leq E[\max(N_1(X), m)],$$

for any m. Clearly,

$$E[N_1(X)] = E[\max(N_1(X), m)] + E[\min(N_1(X), m)] - m,$$

and the same holds for Y, while both $E[N_1(X)]$ and $E[N_1(Y)]$ equal $1/\delta$. This allows us to conclude that

$$E[\min(N_1(Y), m)] \geq E[\min(N_1(X), m)],$$

and by the previous theorem $\pi_0^{(exp)}(m)$ is minimized by the Erlang-k_T distribution over all order k_T phase-type distributions.

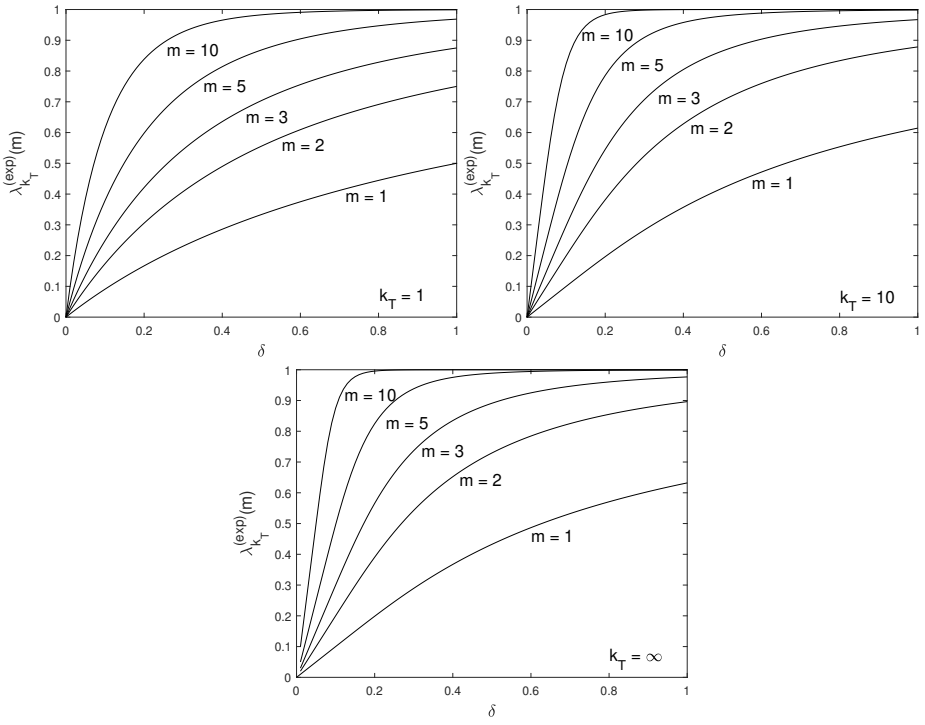

Fig. 1: Illustration of Theorem 3: exponential job sizes and Erlang-k_T timers with $k_T = 1, 10$ and ∞.

The fact that $\pi_0^{(exp)}(m)$ is decreasing in k_T when (β, T) has an Erlang-k_T distribution follows by noting that the Erlang k_T distribution also has an order $k_T + 1$ phase-type representation and therefore $Z_{k_T+1} \leq_{cx} Z_{k_T}$ with Z_k an Erlang-k random variable. Following the same argument as above implies

$$E[\min(N_1(Z_{k_T+1}), m)] \geq E[\min(N_1(Z_{k_T}), m)],$$

which concludes the proof. □

Theorem 3. *For an Erlang-k_T timer and exponential job sizes we have*

$$\lambda_{k_T}^{(exp)}(m) = \delta m - \delta \left(\frac{\delta k_T}{\delta k_T + 1}\right)^{k_T} \sum_{n=0}^{m-1} \frac{m-n}{(\delta k_T + 1)^n} \binom{k_T + n - 1}{n} \tag{6}$$

$$= 1 - \frac{1}{(\delta k_T + 1)^m} \sum_{j=0}^{k_T - 1} \left(1 - \frac{j}{k_T}\right) \binom{m + j - 1}{j} \left(\frac{\delta k_T}{\delta k_T + 1}\right)^j \tag{7}$$

Further,

$$\lim_{k_T\to\infty} \lambda_{k_T}^{(exp)}(m) = \delta m - \delta e^{-1/\delta} \sum_{n=0}^{m-1} \frac{m-n}{\delta^n n!} \tag{8}$$

and $\lim_{m\to\infty} \lambda_{k_T}^{(exp)}(m) = 1$.

Proof. We first establish (6). Using Theorem 1 we know that $\lambda(m)$ can be written as $\delta E[\min(N_1(X), m)]$, which is equivalent to stating that

$$\begin{aligned}
\lambda(m)/\delta = E[\min(N_1(X), m)] &= \sum_{j=0}^{m-1} P[N_1(X) > j] \\
&= m - \sum_{n=0}^{m-1} (m-n) P[N_1(X) = n] \\
&= m - \left(\frac{\delta k_T}{\delta k_T + 1}\right)^{k_T} \sum_{n=0}^{m-1} \frac{m-n}{(\delta k_T + 1)^n} \binom{k_T + n - 1}{n},
\end{aligned}$$

as $P[N_1(X) = n]$ if there are n arrivals (these occur at rate 1) and $k_T - 1$ phase changes of the timer (these occur at rate δk_T) in the next $n + k_T - 1$ events, followed by the expiration of the timer.

To obtain (7) we look at the mean time that the order $(m+1)k_T$ phase-type distribution characterized by $(\beta, 0, \ldots, 0)$ and $Q^{(cycle)}$ given by (5) spends in the set of states $\Omega_0^{(exp)}$. The probability that this set is reached during a cycle via state $(0, j+1) \in \Omega_0^{(exp)}$ is given by

$$\binom{m+j-1}{j} \left(\frac{1}{\delta k_T + 1}\right)^m \left(\frac{\delta k_T}{\delta k_T + 1}\right)^j,$$

and the mean time that we spend in the set $\Omega_0^{(exp)}$ given that we are in state $(0, j)$ equals $(k_T - j)/(\delta k_T)$. As $\pi_0^{(exp)}(m)$ equals δ times this mean time, we have

$$1 - \lambda(m) = \frac{1}{(\delta k_T + 1)^m} \sum_{j=0}^{k_T - 1} \left(1 - \frac{j}{k_T}\right) \binom{m+j-1}{j} \left(\frac{\delta k_T}{\delta k_T + 1}\right)^j.$$

The expression for the limit in (8) is immediate from (6) as $(\delta k_T/(\delta k_T+1))^{k_T}$ converges to $e^{-1/\delta}$ and $\binom{k_T+n-1}{n}/(\delta k_T + 1)^n$ converges to $1/(\delta^n n!)$ as k_T tends to infinity. The second limit is immediate from (7) as $\lim_{m\to\infty} \binom{m-j+1}{j}/(\delta k_T + 1)^m = \lim_{m\to\infty} m^{j-1}/j!(\delta k_T + 1)^m = 0$. □

Remarks:

1. By (6) we have vanishing waits with Erlang-k_T timers and exponential job sizes if and only if

$$\lambda < \lambda_{k_T}^{(exp)}(1) = \delta - \delta \left(\frac{\delta k_T}{\delta k_T + 1} \right)^{k_T}, \tag{9}$$

   which increases to $\delta(1 - e^{-\delta})$ as k_T tends to infinity.
2. Using (7) we can compute $\lambda(m)$ in $O(k_T + \log m)$ time. Indeed, $(\delta k_T + 1)^m$ can be computed in $\log m$ time and the sum in $O(k_T)$ time as $\binom{m+j-1}{j} = \binom{m+j-2}{j-1}(m + j - 1)/j$. Similarly (6) allows us to compute $\lambda(m)$ in $O(m + \log k_T)$ time. Theorem 3 is illustrated in Figure 1.

4 Vanishing waiting times: $m = 1$

When both the job sizes and timers follow a general phase-type distribution, it appears hard to find elegant closed form results. As such we limit ourselves to the case where $m = 1$. Recall that this case is of particular interest as it corresponds to vanishing waiting times in the large-scale limit for the hyper scalable push strategy (assuming asymptotic independence). For exponential timers we know that $\lambda_1(1) = \delta/(\delta + 1)$, which does not depend on the job size distribution. This insensitivity is however lost when timers are not exponential. Given the results in the previous section, we focus on Erlang k_T distributed timers, that is, we focus on $\lambda_{k_T}(1)$ in this section for various job size distributions.

This section contains the following contributions for hyper exponential (HE), Coxian (Cox) and Erlang (Erl) job sizes:

1. We present an explicit formula for $\lambda_{k_T}^{(HE)}(1)$ (see Theorem 4).
2. Tight lower and upper bounds for $\lambda_{k_T}^{(HE)}(1)$ are presented that hold for any HE job size distribution (see Theorem 5).
3. An explicit formula for $\lambda_{k_T}^{(Cox)}(1)$ is derived (see Theorem 6).
4. An explicit formula for $\lambda_{k_T}^{(Erl)}(1)$ is presented (see Theorem 7).

We first present a lemma that is based on the following observation. A renewal occurs for the Markov chain characterized by $Q(m)$ each time the timer expires and the state is in Ω_0. During such a cycle the timer may expire a number of times C_N before the set Ω_0 is reached. Once this set is reached, the cycle ends when the timer expires one more time. Let Y_n denote the service phase of the job when the timer expires for the n-th time during a cycle in the event that $C_N \geq n$.

Lemma 1 *The arrival rate* $\lambda(1)$ *can be expressed as*

$$\lambda(1) = \delta \Big/ 1 + \sum_{s=1}^{k_S} \sum_{n \geq 1} P[Y_n = s, C_N \geq n] \, .$$

Proof. A renewal occurs each time that the set Ω_0 is left. For $m = 1$ the time spend in the set Ω_1 clearly equals 1 as the state remains in Ω_1 until the job completes. The value of $\lambda(1)$ can therefore be expressed as 1 divided by the mean length of a cycle, which we denote as $E[C]$.

A cycle ends when the timer expires and the state is in the set Ω_0. As $1/\delta$ is the mean time for the timer to expire, the mean cycle length can be expressed as

$$E[C] = \frac{1}{\delta}(E[C_N] + 1),$$

where C_N reflects the number of times that the timer expires before the job completes. This number can be expressed as

$$E[C_N] = \sum_{n\geq 1} P[C_N \geq n] = \sum_{s=1}^{k_S} \sum_{n\geq 1} P[Y_n = s, C_N \geq n],$$

where Y_n is the phase of the job when the timer expires for the n-th time in a cycle. Note that Y_n is well defined when $C_N \geq n$. □

4.1 Hyper Exponental job sizes

The next theorem allows us to compute $\lambda_{k_T}^{(HE)}(1)$ in $O(k_S \log k_T)$ time.

Theorem 4. *For Erlang-k_T timers and HE job sizes we have*

$$\lambda_{k_T}^{(HE)}(1) = \delta \Bigg/ \sum_{i=1}^{k_S} \frac{p_i}{1 - \left(\frac{\delta k_T}{\delta k_T + \mu_i}\right)^{k_T}}, \tag{10}$$

where p_i is the probability that a job has an exponential size with mean $1/\mu_i$.

Proof. We make use of Lemma 1. With probability p_i the service phase equals i and remains the same as long as the job is in service. Hence,

$$P[Y_n = i, C_N \geq n] = p_i \left(\frac{\delta k_T}{\delta k_T + \mu_i}\right)^{nk_T}.$$

Summing over n and i and writing $1 = \sum_{i=1}^{k_S} p_i$ yields the result. □

Lemma 2 *The function $\xi(x)$ given by*

$$\xi(x) = \frac{1}{1 - (\delta k_T)^{k_T}/(\delta k_T + 1/x)^{k_T}}, \tag{11}$$

is convex on $(0, \infty)$.

Proof. We have

$$\xi''(x) = \frac{k_T(\delta k_T + 1/x)^{k_T}(\delta k_T)^{k_T}}{x^2}\frac{\eta(x)}{\delta k_T + x}\left(\frac{1}{(\delta k_T + 1/x)^{k_T} - (\delta k_T)^{k_T}}\right)^3,$$

with

$$\begin{aligned}\eta(x) = (k_T+1)(\delta k_T)^{k_T} + (k_T-1)(\delta k_T + 1/x)^{k_T} \\ - 2(\delta k_T)x\left((\delta k_T + 1/x)^{k_T} - (\delta k_T)^{k_T}\right).\end{aligned}$$

It therefore suffices to show that $\eta(x) \geq 0$ on $(0,\infty)$. Expanding the k_T-th powers implies that $\eta(x)$ equals

$$\begin{aligned}&(k_T+1)(\delta k_T)^{k_T} + (k_T-1)\sum_{j=0}^{k_T}\binom{k_T}{j}\frac{(\delta k_T)^{k_T-j}}{x^j} - 2\sum_{j=1}^{k_T}\binom{k_T}{j}\frac{(\delta k_T)^{k_T-j+1}}{x^{j-1}} \\ &= \frac{k_T-1}{x^{k_T}} + \sum_{j=1}^{k_T-1}\left((k_T-1)\binom{k_T}{j} - 2\binom{k_T}{j+1}\right)\frac{(\delta k_T)^{k_T-j}}{x^j},\end{aligned}$$

which is non-negative on $(0,\infty)$ as

$$(k_T-1)\binom{k_T}{j} - 2\binom{k_T}{j+1} \geq 0,$$

if and only if $(j-1)(k_T+1) \geq 0$.

□

Theorem 5. *For Erlang-k_T timers and HE job sizes we have for $\lambda_{k_T}^{(HE)}(1)$ that*

$$\frac{\delta}{1+\delta} = \lambda_1^{(exp)}(1) = \lambda_1^{(HE)}(1) \leq \lambda_{k_T}^{(HE)}(1) \leq \lambda_{k_T}^{(exp)}(1) = \delta\left(1 - \frac{(\delta k_T)^{k_T}}{(\delta k_T+1)^{k_T}}\right)$$

Proof. The result follows by noting that

$$\lambda_{k_T}^{(HE)}(1) = \delta \Big/ \sum_{i=1}^{k_S} p_i \xi(1/\mu_1)\ .$$

Therefore by the convexity of $\xi(x)$ on $(0,\infty)$, we have

$$\lambda_{k_T}^{(HE)}(1) = \delta \Big/ \sum_{i=1}^{k_S} p_i \xi(1/\mu_i) \leq \delta/\xi(\sum_{i=1}^{k_S} p_i/\mu_i) = \delta/\xi(1) = \lambda_{k_T}^{(exp)}(1).$$

□

Remarks:

1. The upper bound is clearly tight, while the lower bound for a fixed k_T is also tight by using the 2-phase HE distribution with $p_1 = 1-\epsilon$, $p_2 = \epsilon$, $\mu_1 = (1-\epsilon)/\epsilon$ and $\mu_2 = \epsilon/(1-\epsilon)$ as in such case $\lim_{\epsilon\to 0}\lambda_{k_T}^{(HE)}(1) = \delta/(\delta+1)$.

4.2 Coxian job sizes

In this section we consider Coxian job sizes, meaning $\alpha = (1, 0, \ldots, 0)$ and

$$S = \begin{bmatrix} -\mu_1 & \mu_1 p_1 & & & \\ & -\mu_2 & \mu_2 p_2 & & \\ & & \ddots & \ddots & \\ & & & \mu_{k_S-1} & \mu_{k_S-1} p_{k_S-1} \\ & & & & \mu_{k_S} \end{bmatrix},$$

with μ_i for $i = 1, \ldots, k_S$ and $0 < p_i < 1$ for $i = 1, \ldots, k_S - 1$. We note that any acyclic phase-type distribution, that is, any phase-type distribution where S is upper triangular, can be represented as a Coxian distribution [12].

Lemma 3 *Let $x_1 \neq \ldots \neq x_s \in \mathbb{R}$, then for $n \geq 1$*

$$\sum_{\substack{j_1,\ldots,j_s \geq 0 \\ j_1+\ldots+j_s = n-1}} \prod_{i=1}^{s} x_i^{j_i} = \sum_{i=1}^{s} x_i^{n+s-2} \prod_{\substack{\ell=1 \\ \ell \neq i}}^{s} \frac{1}{x_i - x_\ell} \tag{12}$$

Proof. The equality in (12) is a known identity for complete homogeneous symmetric functions [13, Ex 7.4]. □

Remarks:

1. The result also holds for $n = 0$, that is,

$$\sum_{i=1}^{s} x_i^{s-2} \prod_{\substack{\ell=1 \\ \ell \neq i}}^{s} \frac{1}{x_i - x_\ell} = 0. \tag{13}$$

   This is easily checked for $s = 2$. For $s > 2$ this follows from Lagrange's interpolation formula as the polynomial $p(x) = x^{s-2}$ interpolates the points (x_i, x_i^{s-2}) for $i = 1, \ldots, s-1$ and therefore

$$p(x_s) = x_s^{s-2} = \sum_{i=1}^{s-1} x_i^{s-2} \prod_{\substack{\ell=1 \\ \ell \neq i}}^{s-1} \frac{x_s - x_\ell}{x_i - x_\ell} = -\sum_{i=1}^{s-1} x_i^{s-2} \frac{\prod_{\ell=1}^{s-1}(x_s - x_\ell)}{\prod_{\substack{\ell=1 \\ \ell \neq i}}^{s}(x_i - x_\ell)}.$$

Theorem 6. *For Erlang-k_T timers and Coxian job sizes with probabilities $p_1, \ldots, p_{k_S-1}$ and rates $\mu_1 \neq \ldots \neq \mu_{k_S}$ we have*

$$\lambda_{k_T}^{(Cox)}(1) = \delta \Bigg/ \sum_{i=1}^{k_S} \frac{\hat{p}_i}{1 - \left(\frac{\delta k_T}{\delta k_T + \mu_i}\right)^{k_T}}, \tag{14}$$

where $\hat{p}_i$ for $i = 1, \ldots, k_S$ is given by

$$\hat{p}_i = \sum_{s=i}^{k_S} \left(\prod_{j=1}^{s-1} \mu_j p_j \right) \prod_{\substack{\ell=1 \\ \ell \neq i}}^{s} \frac{1}{\mu_\ell - \mu_i}. \tag{15}$$

Proof. We rely on Lemma 1. Clearly, as the initial service phase is one when a cycle starts, we have $P[Y_n = 1, C_N \geq n] = \left(\frac{\delta k_T}{\delta k_T + \mu_1}\right)^{nk_T}$ which yields

$$1 + \sum_{n \geq 1} P[Y_n = 1, C_N \geq n] = 1 \Bigg/ 1 - \left(\frac{\delta k_T}{\delta k_T + \mu_1}\right)^{k_T}.$$

For $s > 1$, we get the more involved expression

$$\begin{aligned}
&P[Y_N = s, C_N \geq n] \\
&\quad = \sum_{\substack{j_1,\ldots,j_s \geq 0 \\ j_1+\ldots+j_s = nk_T - 1}} \left(\prod_{i=1}^{s-1} \left(\frac{\delta k_T}{\delta k_T + \mu_i}\right)^{j_i} \frac{\mu_i p_i}{\delta k_T + \mu_i}\right) \left(\frac{\delta k_T}{\delta k_T + \mu_s}\right)^{j_s+1} \\
&\quad = \frac{\delta k_T}{\delta k_T + \mu_s} \left(\prod_{j=1}^{s-1} \frac{\mu_j p_j}{\delta k_T + \mu_j}\right) \sum_{\substack{j_1,\ldots,j_s \geq 0 \\ j_1+\ldots+j_s = nk_T - 1}} \prod_{i=1}^{s} x_i^{j_i}, \\
&\quad = \frac{\delta k_T}{\delta k_T + \mu_s} \left(\prod_{j=1}^{s-1} \frac{\mu_j p_j}{\delta k_T + \mu_j}\right) \sum_{i=1}^{s} x_i^{nk_T+s-2} \prod_{\substack{\ell=1 \\ \ell \neq i}}^{s} \frac{1}{x_i - x_\ell} \qquad (16)
\end{aligned}$$

with $x_i = \delta k_T/(\delta k_T + \mu_i)$ due to Lemma 3. For $s > 1$ we therefore have

$$\begin{aligned}
&\sum_{n \geq 1} P[Y_N = s, C_N \geq n] \\
&\quad = \frac{\delta k_T}{\delta k_T + \mu_s} \left(\prod_{j=1}^{s-1} \frac{\mu_j p_j}{\delta k_T + \mu_j}\right) \sum_{n \geq 0} \sum_{i=1}^{s} x_i^{nk_T+s-2} \prod_{\substack{\ell=1 \\ \ell \neq i}}^{s} \frac{1}{x_i - x_\ell}, \qquad (17)
\end{aligned}$$

where the sum may start in $n = 0$ due to (13).

By definition of x_i, we have for $i \leq s$

$$x_i^{s-1} \prod_{\substack{\ell=1 \\ \ell \neq i}}^{s} \frac{1}{x_i - x_\ell} = \prod_{\substack{\ell=1 \\ \ell \neq i}}^{s} \frac{\delta k_T + \mu_\ell}{\mu_\ell - \mu_i},$$

which combined with (17) implies for $s > 1$

$$\begin{aligned}
&\sum_{n \geq 1} P[Y_N = s, C_N \geq n] \\
&\qquad = \frac{\delta k_T}{\delta k_T + \mu_s} \left(\prod_{j=1}^{s-1} \frac{\mu_j p_j}{\delta k_T + \mu_j}\right) \sum_{n \geq 0} \sum_{i=1}^{s} x_i^{nk_T-1} \prod_{\substack{\ell=1 \\ \ell \neq i}}^{s} \frac{\delta k_T + \mu_\ell}{\mu_\ell - \mu_i} \\
&\qquad = \left(\prod_{j=1}^{s-1} \mu_j p_j\right) \sum_{i=1}^{s} \left(\sum_{n \geq 0} x_i^{nk_T}\right) \prod_{\substack{\ell=1 \\ \ell \neq i}}^{s} \frac{1}{\mu_\ell - \mu_i}.
\end{aligned}$$

Therefore,

$$\sum_{n\geq 1} P[Y_N = s, C_N \geq n] = \sum_{i=1}^{s} \left(\prod_{j=1}^{s-1} \mu_j p_j \right) \prod_{\substack{\ell=1 \\ \ell \neq i}}^{s} \frac{1}{\mu_\ell - \mu_i} \Bigg/ \left(1 - \left(\frac{\delta k_T}{\delta k_T + \mu_i} \right)^{k_T} \right),$$

and

$$\begin{aligned} 1 + E[C_N] &= 1 + \sum_{s=1}^{k_S} \sum_{n\geq 1} P[Y_N = s, C_N \geq n] = \frac{1}{1 - \left(\frac{\delta k_T}{\delta k_T + \mu_i} \right)^{k_T}} \\ &\quad + \sum_{s=2}^{k_S} \sum_{i=1}^{s} \left(\prod_{j=1}^{s-1} \mu_j p_j \right) \prod_{\substack{\ell=1 \\ \ell \neq i}}^{s} \frac{1}{\mu_\ell - \mu_i} \Bigg/ \left(1 - \left(\frac{\delta k_T}{\delta k_T + \mu_i} \right)^{k_T} \right) \\ &= \sum_{i=1}^{k_S} \hat{p}_i \Bigg/ \left(1 - \left(\frac{\delta k_T}{\delta k_T + \mu_i} \right)^{k_T} \right) \end{aligned}$$

□

Remarks:

1. The sum of the $\tilde{p}_i$ equals one as

$$\sum_{i=1}^{k_S} \tilde{p}_i = \sum_{s=1}^{k_S} \left(\prod_{j=1}^{s-1} \mu_j p_j \right) \sum_{i=1}^{s} \prod_{\substack{\ell=1 \\ \ell \neq i}}^{s} \frac{1}{\mu_\ell - \mu_i},$$

   as the latter sum equals zero for $s > 1$. However $\tilde{p}_i$ is not necessarily between 0 and 1. For instance, when $\mu_1 = 3/2, p_1 = 1$ and $\mu_2 = 3$, we get $\tilde{p}_1 = 2$ and $\tilde{p}_2 = -1$. This also indicates that Theorem 6 does not have a simple probabilistic interpretation.
2. When $\tilde{p}_i \in [0, 1]$ for $i = 1, \ldots, k_S$ the Coxian job size distribution corresponds to an HE distribution where the job is exponential with parameter μ_i with probability $\tilde{p}_i$. In fact, any HE distribution can be represented as a Coxian distribution [12,14] and as such Theorem 6 can be regarded as a generalization of Theorem 4.
3. Using (14) we can compute $\lambda_{k_T}^{(Cox)}(1)$ in $O(k_S^2 + k_S \log k_T)$ time, where the computation of the $\hat{p}_i$ values require $O(k_S^2)$ time.
4. If some of the μ_i are identical we can still derive an explicit expression by taking limits. For instance, for an order 2 Coxian distribution with service rate μ in both phases this leads to the following formula:

$$\lambda_{k_T}^{(Cox)}(1) = \delta \Bigg/ \frac{1}{1 - \left(\frac{\delta k_T}{\delta k_T + \mu} \right)^{k_T}} + \frac{p_1 \mu k_T}{\delta k_T + \mu} \frac{\left(\frac{\delta k_T}{\delta k_T + \mu} \right)^{k_T}}{\left(1 - \left(\frac{\delta k_T}{\delta k_T + \mu} \right)^{k_T} \right)^2}.$$

4.3 Erlang k_S job sizes

While results for Erlang distributed job sizes can in principle be derived from Theorem 6 by taking limits, this leads to very involved expressions for larger values of k_S. We therefore present an alternate approach in this section.

Lemma 4 *For $s \geq 1$ and $|x| < 1$, we have*

$$\sum_{n\geq 1} \binom{kn+s-2}{s-1} x^{kn} = \frac{1}{k} \sum_{j=1}^{k} \frac{w_k^j x}{(1-w_k^j x)^s}, \tag{18}$$

with $w_k^j = \cos 2\pi j/k + \mathrm{i} \sin 2\pi j/k \in \mathbb{C}$ the k-th roots of unity.

Proof. The orthogonal relation for the k-th roots of unity states that

$$\frac{1}{k} \sum_{j=1}^{k} w_k^{jn} = \begin{cases} 1 \text{ if } n \text{ is a multiple of } k \\ 0 \text{ otherwise.} \end{cases}$$

This implies

$$\begin{aligned} \sum_{n\geq 1} \binom{kn+s-2}{s-1} x^{kn} &= \sum_{n\geq 1} \binom{n+s-2}{s-1} x^n \left(\frac{1}{k} \sum_{j=1}^{k} w_k^{jn} \right) \\ &= \frac{1}{k} \sum_{j=1}^{k} \sum_{n\geq 1} \binom{n+s-2}{s-1} (x w_k^j)^n = \frac{1}{k} \sum_{j=1}^{k} \frac{w_k^j x}{(1-w_k^j x)^s}, \end{aligned}$$

as $\sum_{n\geq 0} \binom{n+s-2}{s-1} x^n = x/(1-x)^s$. □

Theorem 7. *For Erlang-k_S job sizes and Erlang-k_T timers, we have*

$$\lambda_{k_T}^{(Erl)}(1) = \delta \Bigg/ 1 + \frac{1}{k_T} \sum_{j=1}^{k_T} w_{k_T}^j \sum_{s=1}^{k_S} \frac{x(1-x)^{s-1}}{(1-w_{k_T}^j x)^s}, \tag{19}$$

with $x = \delta k_T/(\delta k_T + k_S)$ and $w_k^j = \cos 2\pi j/k + \mathrm{i} \sin 2\pi j/k \in \mathbb{C}$.

Proof. The proof makes use of Lemma 1. For Erlang k_S job sizes we note that the expression for $P[Y_n = s, C_N \geq n]$ for $s \geq 1$ becomes

$$P[Y_n = s, C_N \geq n] = \binom{nk_T+s-2}{s-1} \left(\frac{k_S}{\delta k_T + k_S} \right)^{s-1} \left(\frac{\delta k_T}{\delta k_T + k_S} \right)^{nk_T}.$$

Lemma 4 now suffices to complete the proof. □

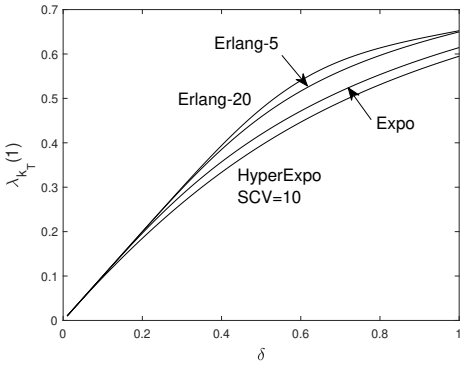

Fig. 2: Illustration of Theorem 4 and 7: Erlang, Exponential and Hyper Exponential job sizes (with balanced means) and Erlang-10 timers.

Remarks:

1. By means of (19) we can compute $\lambda_{k_T}^{(Erl)}(1)$ in $O(k_T k_S)$ time.
2. The result can easily be generalized to mixtures of Erlang distributions (mErl). Suppose that with probability p_i the job size is an order $k_{S,i}$ Erlang with rate μ_i, for $i = 1, \ldots, v$, then

$$\lambda_{k_T}^{(mErl)}(1) = \delta \Bigg/ 1 + \frac{1}{k_T} \sum_{j=1}^{k_T} w_{k_T}^j \sum_{i=1}^{v} p_i \sum_{s=1}^{k_{S,i}} \frac{x_i (1-x_i)^{s-1}}{(1 - w_{k_T}^j x_i)^s} \ ,$$

   with $x_i = \delta k_T / (\delta k_T + \mu_i)$.
3. Theorem 4 and 7 are illustrated in Figure 2. Less variable job sizes imply that higher rates can be supported while still having vanishing waiting times.

5 Conclusions

In this paper we studied the steady state probabilities to be in the set Ω_0 of the structured finite state Markov chain with rate matrix $Q(m)$. The study of this Markov chain was motivated by the largest possible arrival rate $\lambda(m)$ that can be supported by the hyper scalable load balancing push strategy such that the queue length is bounded by some predefined maximum m.

More specifically, the following contributions were made. For exponential job sizes we showed that $\lambda(m)$ is maximized among all order k_T phase type distributions by the Erlang k_T distribution and presented explicit formulas for this maximum $\lambda_{k_T}(m)$. For non-exponential job sizes we focussed on the setting with vanishing waiting times, i.e., $m = 1$ and derived closed form expressions for $\lambda_{k_T}(1)$ for various job size distributions such as hyper exponential, Coxian and Erlang distributions.

References

[1] M. van der Boor, S. Borst, and J. van Leeuwaarden, "Hyper-scalable JSQ with sparse feedback," *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, vol. 3, no. 1, pp. 1–37, 2019.

[2] ——, "Optimal hyper-scalable load balancing with a strict queue limit," *Performance Evaluation*, p. 102217, 2021.

[3] T. Hellemans, G. Kielanski, and B. Van Houdt, "Performance of load balancers with bounded maximum queue length in case of non-exponential job sizes," *To appear in IEEE/ACM Transactions on Networking*.

[4] M. Mitzenmacher, "The power of two choices in randomized load balancing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 12, pp. 1094–1104, October 2001.

[5] N. Vvedenskaya and B. Tsybakov, "Random multiple access of packets to a channel with errors," *Problemy Peredachi Informatsii*, vol. 19, no. 2, pp. 69–84, 1983.

[6] Y. Lu, Q. Xie, G. Kliot, A. Geller, J. R. Larus, and A. Greenberg, "Join-idle-queue: A novel load balancing algorithm for dynamically scalable web services," *Perform. Eval.*, vol. 68, pp. 1056–1071, 2011.

[7] A. Stolyar, "Pull-based load distribution in large-scale heterogeneous service systems," *Queueing Systems*, vol. 80, no. 4, pp. 341–361, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11134-015-9448-8

[8] T. Hellemans, G. Kielanski, and B. Van Houdt, "Performance of load balancers with bounded maximum queue length in case of non-exponential job sizes," *arXiv preprint arXiv.org/abs/2201.03905*, 2022.

[9] M. Bramson, Y. Lu, and B. Prabhakar, "Randomized load balancing with general service time distributions," in *ACM SIGMETRICS 2010*, 2010, pp. 275–286. [Online]. Available: http://doi.acm.org/10.1145/1811039.1811071

[10] C. O'Cinneide, "Phase-type distributions and majorizations," *Annals of Applied Probability*, vol. 1, no. 2, pp. 219–227, 1991.

[11] M. Shaked and J. G. Shanthikumar, *Stochastic Orders and their Applications*. Associated Press, 1994.

[12] A. Cumani, "On the canonical representation of homogeneous markov processes modelling failure - time distributions," *Microelectronics Reliability*, vol. 22, no. 3, pp. 583 – 602, 1982. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0026271482900336

[13] R. P. Stanley and S. Fomin, *Enumerative Combinatorics*, ser. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1999, vol. 2.

[14] B. Van Houdt, "Global attraction of ODE-based mean field models with hyperexponential job sizes," *Proc. ACM Meas. Anal. Comput. Syst.*, vol. 3, no. 2, p. Article 23, June 2019. [Online]. Available: https://doi.org/10.1145/3326137