Discrete-Time Analysis of Multi-Component Queuing Networks under Renewal Approximation

Stefan Geißler*, Stanislav Lange§, Gerhard Haßlinger¶, Phuoc Tran-Gia*, Tobias Hoßfeld*
*Chair of Communication Networks, University of Würzburg, Germany
Email: {stefan.geissler|trangia|hossfeld}@informatik.uni-wuerzburg.de
§Norwegian University of Science and Technology
Email: stanislav.lange@ntnu.no
¶Deutsche Telekom AG
Email: gerhard.hasslinger@telekom.de

***Abstract*—The analytical and numerical performance evaluation of network components and distributed systems has been a staple in the networking community for many years. However, the ever growing complexity of modern systems and the need to gain detailed insights into systems consisting of many, interconnected components emphasizes the need for an extension to the classical single-component approach, and although approaches like Jackson Networks exist, their limited application scope lags behind the complexity of modern environments. To this end, we revisit existing models of the common Gi/Gi/1-∞ queue, extend them to allow the concatenation of multiple queueing components, and evaluate the approximation error introduced through renewal approximation. We revisit previously performed parameter studies and evaluate the approximation error for a wide range of parameter combinations that we can solve through the power of modern computing equipment and efficient numerical implementations of our models. We show the main impact factors for the linear concatenation of queueing components as well as the split and superposition of processes. Our evaluations show that the renewal approximation can be applied to a wide range of parameters while still obtaining results within acceptable error margins.**

***Index Terms*—discrete-time analysis, queueing theory, numerical computation, queueing networks**

I. Introduction

The complexity of distributed systems is continuously growing. Developments that started years ago in the area of cloud computing, namely virtualization and softwarization, have led to significant changes in the landscape of communication networks. First, the introduction of Software-Defined Networking (SDN) aggregated previously distributed control plane components into a single, centralized software controller. Subsequently, the rise of Network Functions Virtualization (NFV) aims at replacing the remaining hardware appliances in the data plane with more flexible software solutions. This migration from ossified hardware middleboxes to software solutions comes with several advantages, but also introduces new challenges.

On one hand, software solutions are naturally more flexible, can be hosted on virtually any Commercial-off-the-shelf (COTS) hardware component and allow dynamic scaling to accommodate variable system loads. On the other hand, their performance is less predictable and they generally are less performant than their hardware counterparts. To circumvent this performance degradation, previously monolithic middleboxes are decomposed into multiple, lightweight functions that can be scaled independently [1]. This decomposition comes with significant challenges when it comes to the analytical modeling of such systems using queuing theoretical approaches.

Historically, communication systems have often been modeled by abstracting a complex system of many moving parts as a single entity, defined by its processing behavior as well as an external arrival process. This abstraction, however, does not allow the investigation of bottlenecks or the impact that scaling individual components has on the overall system performance. To this end, methodologies that allow the evaluation of queuing networks have been proposed in the past.

However, when considering the analytical or numerical modeling of such systems, the list of constraints quickly limits the practical application of many approaches. Methodologies like Jackson Networks [2, 3] in which service times need to be negative exponentially distributed and events need to be processed on a first-come-first-serve basis do provide product-form solutions to open queuing networks. However, these constraints often do not hold in practice. Even with the extensions provided by Gordon and Newell [4] the application to modern systems remains limited. Further extensions by Baskett *et al.* [5] and Gelenbe [6, 7] do provide solutions under less, or more flexible constraints but still require substantial abstractions when dealing with real world systems.

In the context of extending the list of methodologies available for the performance evaluation of queuing networks, this work proposes models to evaluate waiting time, queue size, and interdeparture time distributions of interconnected queuing components. To this end, we build upon existing models by Tran-Gia and Hasslinger [8, 9] and apply the renewal approximation to allow the interconnection of independent queuing components. The general feasibility of this approach was already evaluated in [10]. In this work, we extend our previous investigation to additional topologies and perform a more detailed parameter study. Specifically, we investigate the impact of the linear concatenation of components, the superposition, and split of processes. For each topology type, we perform an extensive parameter study and compare baseline simulation results against our model prediction. Results from this study

can be used to characterize real world systems, which can be analyzed with the proposed models while incurring a low error.

The remainder of this work is structured as follows. Section II provides an overview of related work and highlights relevant contributions from previous studies. The model used for the computations in this work is detailed in Section III. Section IV describes the methodology applied to obtain the results subsequently presented in Section V. Following is a discussion in Section VI, before Section VII concludes this work and outlines future directions.

II. Related Work

Related to the contributions made in this paper are works regarding the analysis of queueing networks, as well as research investigating the renewal approximation in the context of stochastic processes.

Queueing networks have been subject to research for many years. As early as 1980 Bharath-Kumar [11] investigated the relation between Jackson's result [2] and networks of geometric servers. The initial manuscript was later discussed by Bruneel [12]. At the same time, Whitt introduced the queueing network analyzer (QNA) [13, 14] that was designed to approximate congestion metrics of queueing networks based on a two moment approximation of the initial arrival process. This work has since then been extended [15, 16]. The most recent installment of the QNA is able to generate approximations of the mean steady-state performance at each queue of a queueing network. Shortle *et al.* [17] cover several approximations in the context of queueing networks. The book discusses parametric decomposition, the computation of superimposed and split processes, as well as the computation of departure processes based on arrivals.

The heavy-traffic phenomenon [18] for queueing networks has been investigated by Kim [19]. The authors show that, in general, the renewal approximation of arrival processes does not provide enough information about the dependence among interarrival times. Their numerical experiments show that highly variable external arrival processes cause a bottleneck at the last queue with heavy traffic in nine-station tandem queues, while having little impact in queues with moderate traffic.

In addition to the previous works, which assume general processes, there is a large corpus of research based around Jackson's result. Based on methodologies like Jackson Networks [2, 3], extensions by Gordon and Newell [4] partly resolve the strict requirements for the applicability of Jackson's initial approach. Additional work by Baskett *et al.* [5] and Gelenbe [6, 7] further dilute the strict requirements, but are often still not applicable when dealing with real-world queueing networks.

In this work, we aim to extend the already existing, broad research on queueing networks by performing a significant parameter study to quantify the approximation error introduced by applying the renewal approximation. Especially since many works in the area have been conducted in the 1980s, modern computing equipment allows us to conduct significantly larger case studies using numerical approaches.

TABLE I: Notation of random variables and their distributions.

Variable	Description
Input Parameters	
A_i, $a_i(k)$	Interarrival time at component i
B_i, $b_i(k)$	Processing time of component i
Model Output	
W_i, $w_i(k)$	Waiting time of component i
D_i, $d_i(k)$	Interdeparture time of component i
X_i, $x_i(k)$	System size of component i at random time
S, $s(k)$	Sojourn time of total system
V_{A_1,A_2}, $v_{A_1,A_2}(k)$	Superposition of processes defined by interarrival times A_1 and A_2
$G_{p,A}$, $g_{p,A}(k)$	Split processes defined by inclusion probability p and base process defined by interarrival time A

III. Discrete-Time Model

Based on previous works from Tran-Gia [8] and Hasslinger [9], we develop computational models to approximate key characteristics of open queueing networks, including superposition and splitting of processes.

A. Interdeparture Time and Linear Concatenation

We start by computing the interdeparture time of events in the common *Gi/Gi/1*-∞ queue and use the result to realize a simple linear concatenation of queueing components.

To disambiguate between random variables (RVs) and distributions, we use the following convention: uppercase letters such as A denote RVs, their distribution is represented by $a(k)$. The corresponding cumulative distribution functions are denoted as $A(k)$. Accordingly, the model input is composed of the interarrival time distribution $a(k)$, as well as the service time distribution $b(k)$. Based on that, waiting time $w(k)$ and interdeparture time $d(k)$ can be calculated as follows.

To compute the interdeparture time distribution, we first need to compute the waiting time distribution $w(k)$. This can be done using Lindley's equation, in which $*$ denotes the convolution. Note that this computation is performed for each component of a queueing network. We hence omit the index i that denotes the specific component. Instead, the index n describes the n-th service event and hence $w_n(k)$ denotes the waiting time distribution of the n-th processed arrival.

$$w_{n+1}(k) = \pi_0(w_n(k) * c(k)) \quad \text{with} \quad c(k) = a(-k) * b(k) \tag{1}$$

$$\pi_0(x(k)) = \begin{cases} x(k) & k > 0 \\ \sum_{i=-\infty}^{0} x(i) & k = 0 \\ 0 & k < 0 \end{cases}$$

The waiting time in steady state is subsequently defined as

$$w(k) = \lim_{n \to \infty} w_n(k) \, . \tag{2}$$

The waiting time can then be used to compute the idle time distribution $i(k)$ via the virtual unfinished work $u^v(k)$. The idle time distribution describes the time a system is idle after a departure event.

$$u^v(k) = w(k) * c(k) = w(k) * a(-k) * b(k) \tag{3}$$

$$i(k) = K \cdot u^v(-k) \text{ with } K^{-1} = \sum_{j=1}^{\infty} u^v(-j) \tag{4}$$

Finally, the interdeparture time distribution $d(k)$ can be computed using the service time distribution $b(k)$, the idle probability P_E, and the idle time distribution $i(k)$.

$$\begin{aligned} d(k) &= P_E \cdot (i(k) * b(k)) + (1 - P_E) \cdot b(k) \\ P_E &= \frac{E[A] - E[B]}{E[I]} \end{aligned} \tag{5}$$

Additionally, based on the distributional version of Little's Law [20] and the waiting time distribution $w(k)$ we establish the distribution of the number of elements $x(k)$ present in the system at random times. This includes elements currently being processed as well as elements waiting to be processed. We first compute the sojourn time S of elements based on the waiting time and the processing time.

$$s(k) = w(k) * b(k) \tag{6}$$

We then define the number of elements X in the system at a random time as

$$x(k) = x_{S,A}(k) \ . \tag{7}$$

where $x_{S,A}(k)$ describes the probability of observing k arrivals whose interarrival time is distributed according to $a(k)$ during an observation interval whose length is distributed according to the sojourn time S with $s(k)$. The computation of $x_{S,A}(k)$ has been discussed in detail in [21].

Using Equations (1) to (7), all parameters to establish a model for the linear concatenation of *Gi/Gi/1*-∞ queueing systems can be evaluated. Figure 1 shows an exemplary chain of length n, in which the departure process D_i of component i is reused as the arrival process A_{i+1} of component $i+1$. The parameters shown in red indicate model inputs, the ones noted in black represent model outputs.

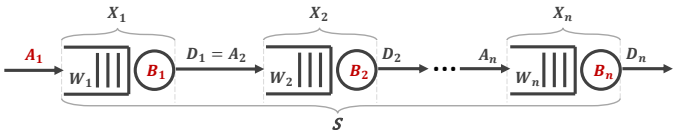


Fig. 1: System and parameter overview in linear concatenation environment.

At this point, we introduce the *renewal approximation* as the departure process dictated by D_1 is, in general, not a renewal process, meaning the sequence $D_{1,t_1}, D_{1,t_2}, D_{1,t_3}, ...$ of instances of the RV D_1 are not independent and identically distributed (IID). Instead, for all but the common *M/M/n*-∞ waiting and *M/M/n*-0 loss systems with Markovian arrival and service processes, the departure process D_1 is expected to exhibit some form of autocorrelation. This assumption of the renewal property of D_1 leads to the error introduced in the model, which is investigated and quantified later in Section V.

B. Split

In addition to the linear concatenation of queuing components, we also provide equations to establish the split of departure processes that can be leveraged to model scenarios as the one shown in Figure 2. Here, the departure process D_1 is split into multiple separate processes whereas each element of D_1 may join exactly one of the resulting, partial processes with a specific probability p.

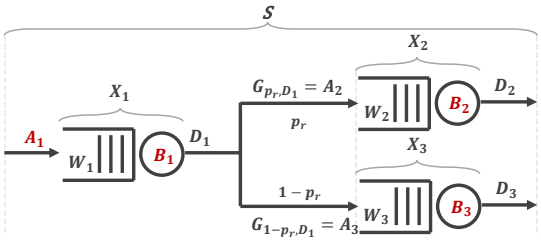


Fig. 2: System and parameter overview in split environment.

The time between occurrences in a resulting process with interarrival time distribution $g_{p,A}(k)$, based on the initial full process A and the probability to join this partial process p, can then be calculated as

$$g_{p,A}(k) = \sum_{i=0}^{\infty} geom_0(i,p) \cdot a^{*i}(k) \ . \tag{8}$$

Thereby, $geom_0(i,p)$ represents the probability of observing i failures before the first success with success probability p. a^{*i} denotes the i-fold convolution of $a(k)$ with itself. Each term of the sum can hence be read as the probability for i events to be removed from the original process multiplied by the probability of the sum of i instances of the RV A to assume k. This dispersion of elements of a non-IID input process A into two or more partial processes results in processes that also violate the IID property. However, a split substream of a renewal process is again renewal.

An alternative, more complex but slightly more efficient computation of the same partial process has been proposed in the past by Hasslinger and Rieger [9].

C. Superposition

Analogously, the superposition, meaning the result of joining two or more processes, can occur in a queuing network, as shown in Figure 3. In this scenario, the elements of two independent processes D_1 and D_2 are merged into a superimposed process with the interarrival time distribution $v_{D_1,D_2}(k)$, that contains all elements of both D_1 and D_2. This process is then used as an arrival process for subsequent queuing components. The time between occurrences in the superimposed process can be computed using the minimum of

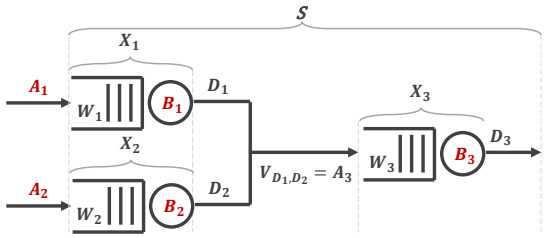


Fig. 3: System and parameter overview in superposition environment.

the recurrence times of the original processes. The recurrence time distribution of a process D can be computed as

$$r(k) = \frac{1 - D(k)}{E[D]} \ . \tag{9}$$

The minima $D = \min(D_1, \ldots, D_n)$ of n RVs $D_1, \ldots, D_n$ has the following cumulative distribution function.

$$D(k) = 1 - \prod_{i=1}^{n} (1 - D_i(k)) \tag{10}$$

Then, V_{D_1,D_2} can be expressed as follows.

$$V_{D_1,D_2} = 1 - \min(R_1, R_2) \tag{11}$$

Thereby, R_1 and R_2 are the RVs denoting the recurrence time of the processes D_1 and D_2, respectively.

IV. Methodology

Based on the computational rules introduced in the previous section, we are now able to compute several relevant parameters of a wide range of queueing networks. In the following, we detail the general methodology of how we obtained the results and cover the exact key performance indicators (KPIs) we take into account to evaluate the quality of our approximation.

A. Experiment Environment

To generate the results presented in Section V, we implemented the computational rules introduced earlier in the R programming language. As most rules imply a numerical solution, we cap the computational accuracy at 0.99999, meaning we terminate the computation once enough probability mass has been accumulated or two compared distributions differ only by 10^{-5} after summing up the probability differences for all values contained in either distribution. Using this implementation, we can compute several KPIs for different, interconnected queueing components. In addition, we perform simulations of equivalent systems using r-simmer [22] to establish baseline values to compare our modeling results against.

B. Evaluated Key Performance Indicators

The first KPI we investigate for all evaluated topologies is the sojourn time distribution $s(k)$ of specific paths of length n in a queuing network. To this end, we compute the sojourn time based on the waiting time distributions $w_i(k)$ and the processing time distributions $b_i(k)$ of all components $1 \leq i \leq n$ on the path.

$$\begin{aligned} s(k) &= w_1(k) * b_1(k) * w_2(k) * b_2(k) * \ldots * w_n(k) * b_n(k) \\ &= s_1(k) * s_2(k) * \ldots * s_n(k) \end{aligned} \tag{12}$$

Note again that the resulting sojourn time neglects potential autocorrelations in $w_i(k)$ due to the assumption of the renewal property regarding all involved processes.

In addition to the sojourn time, we also provide results for the distributions of the system size, meaning the total number of customers in the last component of the system, as well as the interdeparture time distribution $d_i(k)$ of specific components in a queuing network as computed in Equation (1).

As for the system size X, we apply the distributional variant of Little's Law [20] to compute the number of customers in the system at random times, as defined in Equation (7).

For all three KPIs, we examine the Kolmogorov-Smirnov distance between distributions obtained via simulations and the model, respectively. Note that additional metrics have been computed and evaluated, but are omitted here due to space constraints and to allow a concise presentation of results.

V. Parameter Study

After detailing the computational rules required to calculate approximations for several key performance indicators, the following section presents results obtained by means of an extensive parameter study. The goal is to improve our understanding of the relationship between the approximation error and the parameters of the queuing network to be investigated with respect to various KPIs. To this end, we present results comparing model approximations with baseline simulation results to quantify the error for a broad range of topologies and queuing system parameters.

A. Linear Concatenation

The first mode of interconnection investigated is the linear connection of multiple queuing components. Figure 1 shows the structure of linearly concatenated systems investigated in this section.

Input parameters that are defined during the parameter study are shown in red and encompass the initial arrival process A_1 as well as the service processes of each of the queuing components B_1 to B_n. Additionally, we vary the total number of queuing elements to be concatenated. Table II shows the range of parameter values evaluated for the linear concatenation scenario. For both processes A_1 and B_i, the negative binomial distribution has been selected for its ease of parameterization using the mean and coefficient of variation. The mean value for the initial arrival process A_1 has been kept at $E[A] = 100$ for all experiments to eliminate unforeseen interactions between $E[A]$ and $E[B]$. By varying only $E[B]$, we can adjust the load experienced by the system while maintaining control over process interactions. Note that all experiments have been repeated while multiplying all means

TABLE II: Parameter values used during the parameter study.

RV, Value	Parameter Values
A_1	Negative binomial distribution $c_A \in \{0.5, 1, 2, 3, 5\}$ $E[A] = 100$
B_i	Negative binomial distribution $c_B \in \{0.5, 1, 2, 3, 5\}$ $E[B] \in \{10, 30, 50, 70, 90\}$
No. of linear components	$n \in \{1, 2, ..., 20\}$
Prob. to remain during split	$p_r \in \{0.1, 0.2, ..., 0.9\}$

by a factor of 10, and it was established that the absolute values have no impact on the observations presented in this work. Note further that, in order to keep the number of parameter combinations manageable, we assign the same $E[B]$ and c_B to all processing units in a chain of n components. In addition to the mean service times $E[B]$, we vary both the coefficient of variation of the initial arrival process c_A and of the service process c_B. For the parameter study, we conduct experiments using all available parameter combinations that can be formed by the ranges provided in Table II, ultimately evaluating 2,500 parameter combinations with 10 simulation repetitions each, resulting in 25,000 data points.

a) Sojourn Time: Figure 4 shows aggregated results regarding the sojourn time of various linearly concatenated systems in green. We show the main effects plot for all input parameters in Figure 4a. The y-axis shows the mean and 95% confidence intervals for the Kolmogorov-Smirnov distance (KSD) between the simulation and model output. The different input parameters are shown along the x-axis, starting with the system load in the top left. The data shows a continuously growing KSD with an increasing slope as the load approaches 1. These observations are due to the growing effects of the autocorrelation within the system due to queuing that become more prominent at higher loads. Next, in the top right, we show the KSD trend for increasing coefficients of variation of the arrival process c_A. Here we see that, except for $c_A = 1$, higher coefficients of variation lead to a larger approximation error. The lower error for $c_A = 1$ occurs due to the fact that for a coefficient of variation of 1 the negative binomial distribution is identical to the geometrical distribution and hence exhibits memorylessness, effectively resulting in a Markov arrival process. The rate at which the error grows is declining as the coefficient of variation approaches higher values. This is explained by the effect high values of c_A have on the system. With increasing c_A, the overall variability of all arrival, and hence departure, processes increases as well. This in turn reduces the impact of the error introduced by the renewal approximation, as the inherent variability dilutes the neglected autocorrelation of departure processes.

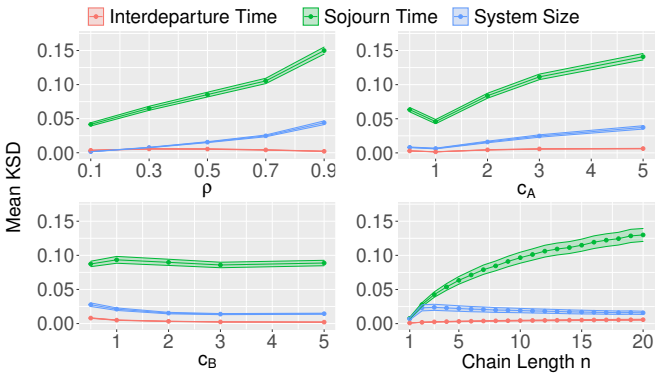
Next, the impact of the length of a concatenated chain is shown in the bottom right. Similar to the coefficient of variation of the arrival process, the chain length induces a continuously growing approximation error that grows quickly for shorter chains and exhibits a decreasing growth rate as the chain length approaches higher values. This is most likely the most intuitive result, as longer chains simply provide more opportunity to introduce approximation errors. For each component of a linear system, the model introduces the same assumption of no relevant autocorrelation. The longer the chain, the larger the error produced through this assumption becomes. Similarly to increasing c_A, when increasing the number of components in the chain, the additional error due to one additional component declines. Again, this is due to the fact that the impact of the neglected autocorrelation decreases for components later in the queue. Hence, longer queues tend to converge against a maximum error instead of diverging.

Finally, the bottom left shows the impact of the coefficient of variation of the service units c_B. Note that all components in a chain of length n exhibit the same service distribution to keep the number of parameter combinations in check. The data shows that c_B has negligible impact, as the observed KSD remains largely stable for all values. This is due to the fact that the effects of coefficient of variation of the service process c_B even out as customers travel through several service units.

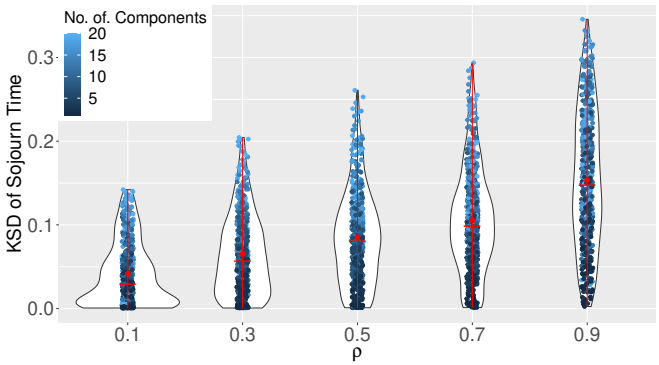
In addition to the main effects plot, Figure 4b shows a violin plot to outline the distribution of KSD values for each of the investigated load levels. In addition, the separate data points are shown, and their color indicates the number of concatenated components. The violins show the density of values along the y-axis, the data points indicate trends regarding which chain lengths contribute mass to the distribution at which KSD values. The different load levels are shown along the x-axis. In addition, the red markers indicate the mean (dot) and median (dash). This means the red dots in this plot correspond to the values shown in the top left subplot of Figure 4a. The data shows that the distribution of KSD values gets wider as load increases. However, even for $\rho = 0.9$, parameter combinations resulting in small KSD values exist. Naturally, the trivial combinations, consisting of only one component, so no concatenation, generate low error values with a KSD of 0.002 for $c_A = c_B = 0.5$. The first non-trivial parameter combination has been observed with $n = 2$ components and $c_A = c_B = 0.5$. The resulting KSD was 0.01. In addition, the color coded data points show that longer chains tend to generate higher approximation errors, confirming the observation made earlier.

To summarize, ρ, c_A, and n are key contributors to the error when one is interested in estimating the sojourn time in linearly concatenated systems.

b) System Size: Next, we analyze the system size, meaning the total number of customers in a service unit including waiting customers as well as customers currently being processed. Figure 4a shows the main effects plot of the KSD of the system size regarding the last component in a chain of length n. The blue curve indicates the mean observed error as well as the 95% confidence interval. The top two subplots, showing the load ρ and the coefficient of variation of the arrival process c_A, exhibit largely the same behavior already observed for the sojourn time in Figure 4a, but overall values are around three times higher for the sojourn time. However, the coefficient of



(a) Main effects plot.



(b) Violin plot for increasing load.

Fig. 4: Kolmogorov-Smirnov distance between the sojourn time, interdeparture time and system size distribution for the linear concatenation scenario obtained by means of simulation and numerical approximation. Main effects plot and violin plot for increasing load. The red markers indicate the mean (dot) and median (dash) in the violin plot.

variation of the service time c_B as well as the chain length n show slightly different behaviors. While the KSD remains largely stable for increasing c_B when examining the sojourn time, when it comes to the system size, the KSD assumes the highest observed values for $c_B = 0.5$ and declines from there. This is an intuitive result, as systems with deterministic service times will result in systems with maximum autocorrelation regarding their departure process. In return, systems with high processing time variations, and hence high c_B, will generally exhibit lower natural autocorrelation and hence decreased approximation errors. Regarding the chain length n, similar behavior can be observed. After the minimal observed KSD for $n = 1$, the data shows a trend of declining values while increasing n from two to 20. Since, for the system size X_n, we are only interested in the last component of a chain of length n, the $n-1$ components ahead of the last one are diluting the error introduced through concatenation. Hence, we observe a high error for short chains as the error of the last component is more prominent, as it had fewer preceding stages of dilution, while the error for longer chains declines due to the increased variability of the service process resulting from service in more components.

Finally, the curve depicts the error observed regarding the interdeparture time of events departing from the final component in a chain of length n. The data shows that the error remains small for all evaluated parameter combinations. This is again an intuitive result, as the computation of the departure distribution introduced in this work produces exact results. The error is only introduced by neglecting the autocorrelation of the process when taking into account the time component. However, as the distribution only takes into account the probability of interdeparture times occurring, this is not mirrored in the results.

In order to illustrate the magnitude of the introduced error, Figure 5 shows the ECDFs of the difference of means. The color indicates the load ρ, each ECDF encompasses all other parameter combinations for that load level. The figure shows the error regarding the prediction via the model. Hence, a

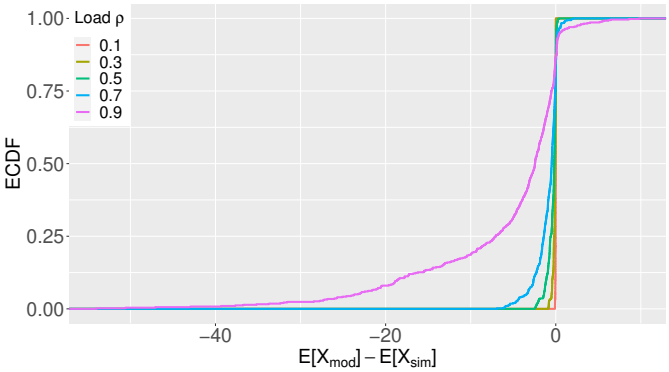


Fig. 5: Difference of means between model and simulation for the system size at the last component in a chain.

value of $E[X_{mod}] - E[X_{sim}] = -40$ means the model underestimates the mean by 40. Note that the model does predict the full distribution instead of only the mean.

The data in Figure 5 shows that there exist both scenarios in which the model underestimates and overestimates. The absolute error largely scales with system load, which is an intuitive observation, as higher load levels generally lead to higher total mean values regarding the system size X in both simulation and model. When examining the data, we observed that the maximum errors were indeed observed for the worst case parameter combinations inferred from the main effects plot. Thus, the maximum observed mean deviation of $E[X_{mod}] = 5.17$ and $E[X_{sim}] = 11.05$ (-53%) was obtained for $c_A = 5$, $c_B = 0.5$, $n = 3$ and $\rho = 0.9$.

To summarize, short linear chains with high c_A, low c_B and high overall system load generally lead to the highest approximation error regarding the system size X.

B. Split

In addition to the linear concatenation of queuing elements, scenarios in which a process is divided into two or more separate streams are relevant in many scenarios like load balancing or internal request routing. To evaluate the accuracy

of the renewal approximation in these scenarios, we investigate the impact of various parameters on the approximation accuracy. Figure 2 shows the component layout for the process splitting scenario. We evaluate a system consisting of three queuing components that are arranged such that the initial arrival process is being processed by the first component. The resulting departure process is then divided into two separate processes based on the probability to remain p_r that dictates the likelihood of customers being processed by component two, whereas customers are processed by component three with the remaining probability of $1 - p_r$.

Figure 6 shows the main effects plot for the observed KSD. The approximation errors are shown in red for the interdeparture time of the last component, green for the sojourn time and blue for the system size of the last component. Instead of the chain length n, the bottom right plot shows the staying probability p_r.

First, looking at the sojourn time, the data shows a near-linear impact for c_A, c_B and p_r, while the load ρ exhibits roughly constant behavior up until $\rho = 0.7$. For higher loads, a significant increase of the approximation error has been observed. Note that the overall values are significantly smaller compared to the values shown in Figure 4a. As the number of components included in this scenario is small ($n = 2$) and constant over all parameter combinations, the overall observed errors are expected to be smaller. The data for ρ, c_A and c_B follows the intuitive expectation, as both high load and high coefficients of variation lead to an increased waiting probability, which in return leads to increased autocorrelation. This autocorrelation is the source of the approximation error observed in the data. Similarly, probabilistically removing events from a process generally also removes autocorrelations, shifting the system closer to our IID assumption. Hence, the observed KSD increases with a growing probability for elements to remain a part of the process p_r, as shown in the bottom right facet.

Next, the interdeparture time of customers as they depart the last component of the system D_2, as shown in Figure 2 is shown in red. It can be seen that for both high and low load levels, the approximation error is small, while medium load levels tend to generate higher errors. This is explained by the fact that, for low load levels, the departure process converges towards the arrival process while for high load levels the departure process converges towards the service process. For medium load levels, the system behavior is "unpredictable", and hence leads to larger errors.

c_A and c_B have opposing effects on the interdeparture time distributions. While increasing c_A reduces the overall approximation error, increasing c_B tend to increase the observed error.

The staying probability p_r, once again intuitively affects the resulting error, as high values lead to lower errors. Since for $p_r = 1$ the split scenario converged towards a linear concatenation scenario, the system then follows the same trends as observed before. Hence, for higher values of p_r, the error introduced through the splitting process becomes more and more negligible.

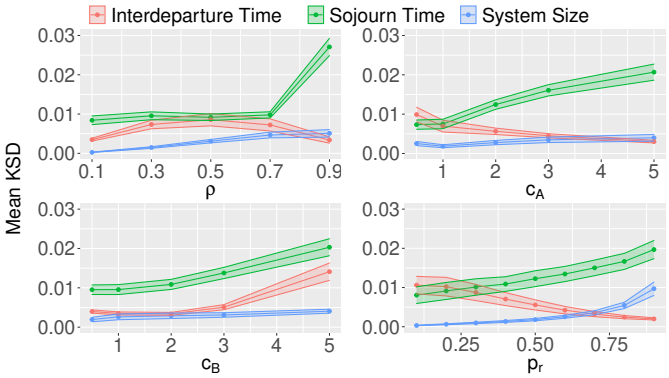


Fig. 6: Main effects plot of KSD values after process splitting.

Finally, the blue curve shows the approximation error when it comes to the system size of the last component in the chain, meaning component two in Figure 2. Generally, the system size behaves similarly to the sojourn time, whereas the absolute error values are significantly lower compared to the sojourn time. Further, the system size is far less sensitive to high load levels as well as coefficients of variation.

C. Superposition

Finally, the last synthetic topology investigated in this work results in the superposition of processes, as shown in Figure 3. Here, two independent arrival processes, defined by A_1 and A_2 experience service by components one and two, respectively. The resulting departure processes are then superimposed before the resulting combined process is serviced by component three. We again investigate the impact of various system parameters on the observed approximation error.

Figure 7 shows the data obtained for the superposition scenario. We again show the main effects plot of the interdeparture time, sojourn time and system size for the relevant scenario parameters. In this case, this includes the system load ρ and the coefficients of variation of the arrival processes c_A as well as the service processes c_B. Note that the processes dictated by A_1 and A_2, as shown in Figure 3, are identical in all evaluated scenarios. Similarly, the coefficient of variation for the distributions of B_1, B_2 and B_3 are identical. Finally, in all parameter combinations we chose $E[B_3] = 0.5 \cdot E[B_1] = 0.5 \cdot E[B_2]$, as this achieves the same system load ρ for all three components.

In addition to the main effects plot, the top right facet shows the violin plot for the observed KSD values between the sojourn times observed in simulation and model, respectively. To show a more detailed picture of the observed distribution instead of only mean and confidence intervals, we show the specific distributions for each of the investigated load levels. The green markers show the mean (dot) and median (dash) values, the black markers indicate single observed data points.

The observations in this scenario are, for the most part, in line with what was observed earlier. System load ρ being the most significant factor for the approximation accuracy of the sojourn time as well as system size, with both c_A and c_B only having minor, linear effects. The interdeparture

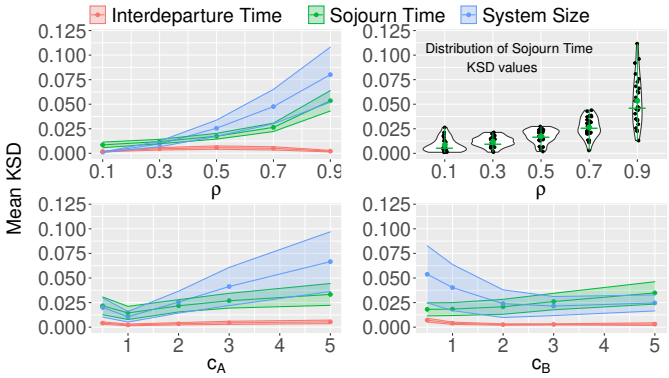


Fig. 7: Main effects plot of KSD values after process superposition. Violin plot for observed sojourn time KSD values.

time distribution is largely unaffected by all three parameters. What stands out is the behavior of the system size in general. Where the KSD for the system size X was continuously smaller than the KSD for the sojourn time, it exceeds the sojourn time for the first time in this scenario. In addition, the observed confidence interval is significantly larger for all three parameters. The latter is explained by the increased impact of the investigated parameters. As all three parameters have significant impact on the system size approximation accuracy during the superposition of processes, the observed confidence intervals naturally grow. As for the overall observed values, it would be expected for the superposition of processes to generate relatively small errors, as the superposition of non-synchronized processes generally moves the result closer to memorylessness than either of the input processes. This has for example been shown in [23] and [24] for the superposition of deterministic processes. However, the inaccuracy introduced by the approximation is higher than the "advantage" gained through superposition, hence the errors observed in our parameter study.

VI. Discussion and Limitations

In this work, we examine the approximation error for linear concatenation, split, and superposition. However, when dealing with general, open queueing networks, we also need the ability to incorporate direct and indirect feedback. Hence, elements may be processed by one of the network's components and be either fed back into the same or other components. In the context of this work, we also examined the impact of various parameters on the approximation error for this feedback scenario. However, the results show that no parameter combination yields acceptable approximation results. For these reasons, we decide not to include the model as well as the results on feedback in this work. In order to provide usable approximations, further research on modeling the feedback behavior required by modern systems needs to be performed. In short, the autocorrelation introduced by feedback is vastly more complex compared to the other concatenation types. For example, in a single component system with feedback, a single event may cycle multiple times before any "fresh" events arrive at the system. This significantly impacts the departure process as in an interval like this, several interdeparture times would follow the same distribution as the service time. Other periods of the departure process may follow other distributions, depending on system load, coefficients of variation and other parameters such as feedback distance (i.e., how many components are passed before events are fed back into the system). Hence, this evaluation is omitted here. We invite the community to tackle this interesting and challenging problem.

Furthermore, our conducted parameter study comes with one major limitation. In all evaluated scenarios, we assume all queueing components of a network to have the same service time distribution. Obviously, this would not hold true in reality. However, it allowed us to keep the number of parameter combinations in check. Even with this limitation, we calculated over 250,000 data points. When allowing every queueing component (e.g. of a chain of length 20) to exhibit varying service time distributions, we would end up with an exponentially higher number of parameter combinations. The impact of varying processing time distributions within a queueing network remains for future work. Similarly, the impact of more complex processing, as well as arrival processes, remains content for future research. These include arrival specific processing, in which processing times differ between arrivals, but are constant for a single arrival and all components in the network, as well as arrival processes exhibiting explicit autocorrelation.

VII. Conclusion

In this work, we introduce models that allow the linear concatenation of queueing components as well as the splitting and superposition of random processes. These models are based on the assumption that the departure processes of a general Gi/Gi/1-∞ queue exhibit the renewal property. This assumption, however, does not generally hold in practice for arbitrary processes. According to Burke [25], the only systems that formally fulfill this property are M/M/1, M/M/n and M/M/∞ waiting systems as well as their geometric counterparts in the discrete-time domain. In this work, however, we investigate the impact of this assumption when applying the same principle to the common Gi/Gi/1-∞ waiting system.

We revisit previously proposed models by Tran-Gia [8] and Hasslinger [9] and extend them to allow the concatenation of elements. We perform an extensive parameter study to investigate a broad set of system parameters on the resulting approximation error. To this end, we design three distinct scenarios to evaluate the impact of various system parameters in systems that feature linear concatenation of queues as well as splitting and superposition of processes. For all scenarios, we have performed simulations as well as computed our model output to generate statistically significant comparisons.

Our results have shown that for all concatenation types — linear, split and superposition — the system load ρ is generally the most influential factor. As the system load approaches $\rho = 1$, the error introduced through approximation becomes gradually more significant. However, even for load levels close to 1, the approximate models are able to generate results within

reasonable margins. The largest deviation observed in our studies was -53% with $E[X_{mod}] = 5.17$ and $E[X_{sim}] = 11.05$ when concatenating three queueing systems in a scenario with high load and an interarrival time distribution with a large coefficient of variation. Most errors, however, are substantially smaller. In general, the mean KSD for the interdeparture time of the last component in a system was smaller than 0.023 across the board. When looking at the total system sojourn time, the maximum KSD was 0.34 for a chain of length $n = 20$, $\rho = 0.9$, and high coefficients of variation for interarrival and service time distributions, $c_A = 5$ and $c_B = 2$. Finally, the maximum approximation error observed regarding the system size, meaning the number of elements present, at the last element within a concatenated system was 0.22 for $\rho = 0.9$ and $c_A = 5$, $c_B = 0.5$.

Generally, the approximation allows the evaluation of complex queueing systems, like microservice architectures, VNF chains or other interconnected, distributed systems. However, when applying this approximation, we always have to keep in mind the error we are willing to accept and the metric we are interested in. While the renewal approximation has only minor impact on the predicted number of elements present in a specific component at one time, the sojourn time may include a significant approximation error, depending on the system configuration and input parameters. For the future, in addition to the open issues outlined in the discussion, we aim at extending our parameter study and want to establish a library of topologies and configurations that allows the estimation of the expected error beforehand.

REFERENCES

[1] S. R. Chowdhury, M. A. Salahuddin, N. Limam, and R. Boutaba, "Re-architecting nfv ecosystem with microservices: State of the art and research challenges," *IEEE Network*, 2019.

[2] J. R. Jackson, "Networks of Waiting Lines," *Operations research*, 1957.

[3] ——, "Jobshop-like Queueing Systems," *Management Science*, 1963.

[4] W. J. Gordon and G. F. Newell, "Closed Queuing Systems with Exponential Servers," *Operations research*, 1967.

[5] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios, "Open, Closed, and Mixed Networks of Queues with Different Classes of Customers," *Journal of the ACM (JACM)*, 1975.

[6] E. Gelenbe, "G-networks by Triggered Customer Movement," *Journal of applied probability*, 1993.

[7] E. Gelenbe and J.-M. Fourneau, "G-networks with Resets," *Performance Evaluation*, 2002.

[8] P. Tran-Gia, "Discrete-time analysis for the interdeparture distribution of gi/g/1 queues," in *Proc. of the international seminar on Teletraffic analysis and computer performance evaluation*, 1986.

[9] G. Hasslinger and E. S. Rieger, "Analysis of open discrete time queueing networks: A refined decomposition approach," *Journal of the Operational Research Society*, 1996.

[10] S. Geißler, S. Lange, P. Tran-Gia, and T. Hoßfeld, "Discrete-time Analysis of Multicomponent GI/GI/1 Queueing Networks," in *International Conference on Networked Systems (NetSys)*, 2021.

[11] K. Bharath-Kumar, "Discrete-time queueing systems and their networks," *IEEE Transactions on Communications*, 1980.

[12] H. Bruneel, "Comments on" discrete-time queueing systems and their networks"," *IEEE Transactions on Communications*, 1983.

[13] W. Whitt, "The queueing network analyzer," *The bell system technical journal*, 1983.

[14] ——, "Performance of the queueing network analyzer," *Bell System Technical Journal*, 1983.

[15] W. Whitt and W. You, "A robust queueing network analyzer based on indices of dispersion," *arXiv preprint arXiv:2003.11174*, 2020.

[16] ——, "A robust queueing network analyzer based on indices of dispersion," *Naval Research Logistics (NRL)*, 2022.

[17] J. F. Shortle, J. M. Thompson, D. Gross, and C. M. Harris, *Fundamentals of queueing theory*. John Wiley & Sons, 2018.

[18] S. Suresh and W. Whitt, "The heavy-traffic bottleneck phenomenon in open queueing networks," *Operations Research Letters*, 1990.

[19] S. Kim, "The heavy-traffic bottleneck phenomenon under splitting and superposition," *European Journal of Operational Research*, 2004.

[20] R. Haji and G. F. Newell, "A relation between stationary queue and waiting time distributions," *Journal of Applied Probability*, 1971.

[21] S. Gebert, T. Zinner, S. Lange, C. Schwartz, and P. Tran-Gia, "Discrete-time analysis: Deriving the distribution of the number of events in an arbitrarily distributed interval," University of Wuerzburg, Tech. Rep., Jun. 2016.

[22] I. Ucar, B. Smeets, and A. Azcorra, "Simmer: Discrete-event simulation for r," *arXiv preprint arXiv:1705.09746*, 2017.

[23] F. Wamser, P. Tran-Gia, S. Geißler, and T. Hoßfeld, "Modeling of traffic flows in internet of things using renewal approximation," in *Advances in Optimization and Decision Science for Society, Services and Enterprises*, Springer, 2019.

[24] P. Tran-Gia and T. Hoßfeld, *Performance Modeling and Analysis of Communication Networks, A Lecture Note*. Würzburg University Press, 2021. DOI: 10.25972/WUP-978 - 3 - 95826 - 153 - 2. [Online]. Available: https : / / modeling.systems.

[25] P. J. Burke, "The output of a queuing system," *Operations research*, 1956.