SEMI-MARKOVIAN CAPACITIES
IN PRODUCTION NETWORK MODELS

SIMONE GÖTTLICH* AND STEPHAN KNAPP
University of Mannheim
Department of Mathematics
68131 Mannheim, Germany

Abstract. In this paper, we focus on production network models based on ordinary and partial differential equations that are coupled to semi-Markovian failure rates for the processor capacities. This modeling approach allows for intermediate capacity states in the range of total breakdown to full capacity, where operating and down times might be arbitrarily distributed. The mathematical challenge is to combine the theory of semi-Markovian processes within the framework of conservation laws. We show the existence and uniqueness of such stochastic network solutions, present a suitable simulation method and explain the link to the common queueing theory. A variety of numerical examples emphasizes the characteristics of the proposed approach.

1. Introduction. The mathematical consideration of production processes as well as supply chains implies a variety of modeling approaches. Generally they can be divided into two classes called microscopic and macroscopic models. The key idea of microscopic models is to track each item individually during production. Typical approaches are either based on discrete event simulations [1, 2] or Newton-type dynamics [16]. In contrast, macroscopic models describe the average behavior of the production system in terms of part density or flow. They naturally arise from microscopic limits and are usually based on ordinary differential equations (ODEs) [15], hyperbolic partial differential equations (PDEs) [1, 5, 6, 8, 11] or a mixture of both [4, 7, 14]. We refer to [3] and the references therein for a comprehensive overview of macroscopic production models.

Up to now, most of the macroscopic systems are purely deterministic and stochastic effects are less investigated. However, production networks are influenced by random perturbations as for instance random capacities. In [10, 17], exponentially distributed operating and down times are assumed meaning that a machine is either on and capable to work with full capacity or off leading to a total breakdown of production. Mathematically, the operating and down times can be seen as holding times in the operating and non-operating state of the machine such that the capacity process can be interpreted as a continuous time Markov Chain (CTMC).

2010 Mathematics Subject Classification. Primary: 90B15, 65Cxx; Secondary: 65Mxx.
Key words and phrases. Production networks, conservation laws, semi-Markovian processes.
This work was financially supported by the DAAD project “DAAD-PPP VR China” (Projekt-ID: 57215936).
* Corresponding author: Simone Göttlich.
The theory on CTMCs is well established and also widely used in queueing theory [20, 23, 24], where the distribution of arrival times and service times is not restricted to the exponential distribution only. A natural generalization of CTMCs are semi-Markov processes. These can be used to model and analyze reliability functions with semi-Markov failure rates processes, see e.g. [18, 19]. The theory of semi-Markov processes with discrete state space are intensively discussed [19, 24, 27] and for more arbitrarily state spaces in [21].

In this work, we focus on a macroscopic model [14, 17] and use the concept of semi-Markov processes to reformulate the capacity drop of machines by a capacity process which also allows for intermediate capacity states. Additionally, exponentially distributed operating and down times are replaced by arbitrarily distributed holding times for the corresponding capacity states. This leads to the extension of the original two-state capacity process to a multi-state process.

From a mathematical point of view, we have to fuse the theories of coupled PDE-ODE systems and semi-Markov processes in a way that the well-posedness of the resulting stochastic model is ensured. The numerical treatment and the interpretation of the simulation results compared to the common queueing theory approach are also discussed. A theoretical result as in [10] for the expected density is unfortunately not possible since a Lagrangian particle formulation is not available for network problems.

The paper is organized as follows: In section 2 the reformulation of the capacity drop will be achieved by defining the capacity processes \( \mu^e(t), t \geq 0 \) for all processors \( e \) in the network as a family of independent semi-Markov processes and include the latter in the original coupled PDE-ODE model [14]. The well-posedness of the model follows from the fact that the \( M \)-dimensional process \( ((\mu^1(t), \ldots, \mu^M(t)), t \geq 0) \) has only finitely many jumps in finite time intervals \( P \)-almost surely \( (P \text{-a.s.}) \). In section 3, we will use the independence of the stochastic processes \( \mu^1, \ldots, \mu^M \) to simulate the model. This is done by splitting the generation of sample paths for the capacities and the numerical treatment of the PDE-ODE system, cf. [17]. Therein, a stochastic simulation algorithm [12, 13] and the idea of piecewise-deterministic processes [9] are presented to solve coupled PDE-ODE problems. However, the simulation of sample paths relies on the Markov property of the embedded Markov process and stochastic numerical methods taken from [22]. To close the link between stochastic production network models and queueing theory we analyze expected queue-loads and work in progress for a single queue-processor unit in section 4. Further computational results and comparisons will be provided in section 5.

### 2. Stochastic production networks

The model under consideration is mainly based on [17] while the focus is on a different stochastic process for the machine breakdowns. More precisely, the capacity process originally defined by operating and non-operating states with exponentially distributed durations is replaced by a semi-Markov process. In doing so, there are two benefits: On the one hand, intermediate capacity states are allowed and, on the other hand, generalized distributions for the operating and down times can be applied.

We start with the framework of semi-Markov processes and explain in a second step how these processes can be used in the context of PDE-ODE coupled production networks.
2.1. **Semi-Markov processes.** The general idea of a semi-Markov process with discrete state space basically relies on a continuous-time Markov chain. In the following, we denote by \((H_n, n \in \mathbb{N})\) the holding times, i.e. the duration of the process being in the state \(X_{n-1}\), and by \((J_n, n \in \mathbb{N}_0)\) the jump times, i.e. the times at which the value of the process changes from \(X_{n-1}\) to \(X_n\). Then, using the convention \(J_0 = 0\),

\[
J_n = \sum_{i=1}^{n} H_i
\]

holds for every \(n \in \mathbb{N}_0\). The capacity process \((\mu(t), t \geq 0)\) is thus piecewise-constant and right-continuous satisfying

\[
\mu(t) = X_n \text{ if } t \in [J_n, J_{n+1})
\]

as figure 1 illustrates.

![Figure 1. General idea of a semi-Markov process](image)

To embed these modeling assumptions into the theory of semi-Markov processes, we present basic definitions and relevant results taken from [25] and [27]. We always assume a finite set of capacities and to ease the notation we use integer values as possible states of the process to capture different capacities, i.e., state \(k\) corresponds to capacity \(c_k\). Let \(m \in \mathbb{N}\) be the number of states and \(Q : \mathbb{R} \to \mathbb{R}^{m \times m}\) be a matrix-valued function. Each component function is assumed to be non-decreasing and right-continuous. Furthermore, the following properties should hold:

\[
Q_{ij}(t) = 0 \text{ for every } t \leq 0,
\]

\[
L_i(t) := \sum_{j=1}^{m} Q_{ij}(t) \text{ for every } t \in \mathbb{R} \text{ and}
\]

\[
\lim_{t \to \infty} L_i(t) = 1.
\]

Let us now define the so-called \((X, H)\)-process, see [27].

**Definition 2.1.** (\((X, H)\)-process) The \((X, H)\)-process is a two-dimensional time-discrete stochastic process \((X_n, H_n), n \in \mathbb{N}_0\) with values in \(\{1, \ldots, m\} \times \mathbb{R}\) on some probability space \((\Omega, \mathcal{A}, P)\) satisfying

\[
P(H_0 = 0) = 1,
\]

\[
P(X_0 = k) = \pi_k
\]
for a given initial distribution vector $\pi$ and
\[ P(X_{n+1} = k, H_{n+1} \leq t | X_0, H_0, \ldots, X_n, H_n) = Q_{X,k}(t) \] (7)
P-a.s. for every $t \in \mathbb{R}$, $n \in \mathbb{N}_0$ and every $k = 1, \ldots, m$.

From equations (2)-(4) we can deduce that all $L_i$ are distribution functions and consequently describe the distribution of the holding times given the state of the process. The matrix $Q$ implies the common transition probabilities of the holding time process $(H_n, n \in \mathbb{N}_0)$ and the state process $(X_n, n \in \mathbb{N}_0)$. Consequently, we can choose some matrix-valued function with entries given by weighted cumulative distribution functions (cdf’s) such that the assumptions (2)-(4) are satisfied. Note that there is no specific choice of the cdf’s and the exponential distribution forms a special case.

A relevant result for the numerical simulation of such processes is given by the following theorem, see [27]:

**Theorem 2.2.** The stochastic process $(X_n, n \in \mathbb{N}_0)$ is a Markov chain and the $(X, J)$-process $((X_n, J_n), n \in \mathbb{N}_0)$ with $J_n$ defined by (1) is a Markov process.

Using the considerations from figure 1, the capacity process is induced by the $(X, H)$-process. This is the so-called $Z$-process or semi-Markov process which we define as follows:

**Definition 2.3** (Semi-Markov process). Let a $(X, H)$-process on $(\Omega, \mathcal{A}, P)$ be given. We define the $Z$-process as a time-continuous stochastic process $(Z(t), t \geq 0)$ on $(\Omega, \mathcal{A}, P)$ with values in $\{1, \ldots, m\}$ by
\[ Z(t) = X_n \text{ if } t \in [J_n, J_{n+1}) \] (8)
for every $t \geq 0$. The process is called semi-Markov process.

The semi-Markov process $(Z(t), t \geq 0)$ is in general no Markov process but a characterization as an embedded Markov process in the sense of a $(X, J)$-process is straightforward.

### 2.2. Network definition and modeling equations.

We briefly recall the PDE-ODE production network model from [14] and consider for simplicity the special case of a serial network without branches. The network consists of $M$ queue-processor units $e \in E = \{1, \ldots, M\}$, where each processor possesses a single queue (or buffer) in front. There is a given external inflow into the network at the first queue and the total outflow is considered as the output of processor $M$.

The queue of each processor is modeled by an ordinary differential equation (ODE) that measures the difference between incoming and outgoing flux. In detail, the queue-load at time $t$ is denoted by $q^e(t)$ and computed as
\[ \partial_t q^e(t) = g^e_{in}(t) - g^e_{out}(t), \] (9)
\[ q^e(0) = q^e_0 \] (10)
for an initial value $q^e_0 \in \mathbb{R}_{\geq 0}$.

If the queue is non-empty, the outflux $g^e_{out}$ is determined by the capacity $\mu^e(t)$ at time $t$ of the associated processor $e$. Otherwise, if the queue is empty, the outflow is the minimum of the incoming flow $g^e_{in}(t)$ and the capacity $\mu^e(t)$:
\[ g^e_{out}(t) = \begin{cases} \mu^e(t) & \text{if } q^e(t) > 0, \\ \min \{ \mu^e(t), g^e_{in}(t) \} & \text{if } q^e(t) = 0. \end{cases} \] (11)
Every processor $e \in E$ is characterized by its individual capacity $\mu^e(t)$ and its processing velocity $v^e$. The part density $\rho: [a^e, b^e] \times \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ is governed by the conservation law
\[
\partial_t \rho^e(x, t) + \partial_x f^e(t, \rho^e(x, t)) = 0, \tag{12}
\]
where the flux function is
\[
f^e(t, \rho) = \min\{v^e \rho, \mu^e(t)\}, \tag{13}
\]
and initial and boundary values are
\[
\rho^e(x, 0) = \rho^e_0(x), \tag{14}
\]
\[
\rho^e(0, t) = \frac{g^e_{\text{out}}(t)}{v^e}. \tag{15}
\]
The flux into a queue is consequently
\[
g^e_{\text{in}}(t) = \begin{cases} G^e_{\text{in}}(t) & \text{if } e = 1, \\ f^{e-1}(t, \rho^{e-1}(b^{e-1}, t)) & \text{else} \end{cases} \tag{16}
\]
with $G^e_{\text{in}}$ being the time-dependent inflow into the serial production network. We state the following

**Definition 2.4** (Stochastic production network model). Let $\{(\mu^e(t), t \geq 0), e \in E\}$ be a family of independent semi-Markov processes on some probability space $(\Omega, \mathcal{A}, P)$ having finite state spaces $C^e \subset \mathbb{R}_{\geq 0}$. The stochastic production network model is then given by the equations (9)-(16) with capacities $\mu^e(t), e \in E$.

Analogously to the deterministic case, i.e., all processors are working with their maximal capacities $\mu^e$, we have to prove that the stochastic network model is well-defined. Therefore, the following definition is needed.

**Definition 2.5** (Solution). We call $X = (\mu^e, \rho^e, q^e, e \in E)$ a solution of the stochastic production network model given by definition 2.4 if there exists a set $N \in \mathcal{A}$ with $P(N) = 0$ such that for every $\omega \in N^c$ the equations (9)-(16) hold in the sense of a deterministic network solution, see [3].

We obtain the following existence and uniqueness result.

**Theorem 2.6** (Existence and uniqueness of a solution). Let a stochastic production network model be given. Furthermore, we assume that
\[
TV_{[a^e, b^e]}(\rho^e_0) := \sup \left\{ \sum_{i=0}^{N-1} |\rho^e_0(x_{i+1}) - \rho^e_0(x_i)| : \{x_0, \ldots, x_N\} \text{ partition of } [a^e, b^e] \right\}
\]
and
\[
TV_{[t_0, T]}(G^e_{\text{in}}) := \sup \left\{ \sum_{j=0}^{T-1} |G^e_{\text{in}}(t_{j+1}) - G^e_{\text{in}}(t_j)| : \{t_0, \ldots, t_T\} \text{ partition of } [t_0, T] \right\}
\]
are bounded for every $e \in E$. Then, there exists a unique solution of the stochastic network model on the time interval $[t_0, T]$.

**Proof.** The proof comprises two parts. First, we show that the $M$-dimensional stochastic process $(\mu^1, \ldots, \mu^M)$ consists of finite jumps in finite time intervals $P$-a.s. Let $N(t)$ be the number of jumps for the multidimensional process up to time...
and let \( N^e(t) \) be the number of jumps of the \( e \)--th semi-Markov process for some \( t \in [t_0, T] \) and for all \( e \in E \). Due to the inequality

\[
N(t) \leq \sum_{e \in E} N^e(t) \leq M \max\{N^e(t), e \in E\}
\]

we can conclude

\[
P(N(t) \geq n) \leq P(M \max\{N^e(t), e \in E\} \geq n)
= 1 - P\left(\max\{N^e(t), e \in E\} < \frac{n}{M}\right)
= 1 - \prod_{e \in E} P\left(N^e(t) < \frac{n}{M}\right)
\]

for every \( n \in \mathbb{N}_0 \) following from the independence of the processes \( N^e \). Since a semi-Markov process with finite state space only consists of finitely many jumps in finite time \( P\)-a.s., see e.g. [27], it follows that

\[
\lim_{n \to \infty} P\left(N^e(t) < \frac{n}{M}\right) = 1
\]

for every \( e \in E \). Hence, we get

\[
\lim_{n \to \infty} P(N(t) \geq n) = 0
\]

which implies that the number of jumps of the multidimensional process are also \( P\)-a.s. finite in every finite time interval.

Second, we can prove the existence of a solution by setting \( N \in \mathcal{A} \) as the set of all events such that \((\mu^1, \ldots, \mu^M)\) has infinitely many jumps in \([t_0, T]\). From the first part of the proof we know that \( P(N) = 0 \). For every \( \omega \in N^e \) we obtain an integer \( K(\omega) \in \mathbb{N} \) and finitely many jump times \( \{J_1(\omega), \ldots, J_{K(\omega)}(\omega)\} \) such that we can construct the solution piecewise in-between. The main assumptions for the existence and uniqueness of solutions in the deterministic case [3] are \( \text{TV}_{[a^e, b^e]}(\rho_0^e) < \infty \), \( \text{TV}_{[t_0, T]}(G_{tn}) < \infty \) and are also fulfilled here. Thus, we get the existence of a unique solution for the realization \( \omega \in N^e \) on \([t_0, J_1(\omega)]\) with \( \text{TV}_{[a^e, b^e]}(\rho^e(\cdot, J_1(\omega), \omega)) < \infty \).

Then, the assumptions on the total variation are again satisfied and \( \mathcal{X}(J_1(\omega), \omega) \) can be used as a new initial value for a network with capacities \( \mu^e(J_1(\omega), \omega) \). A unique solution for this realization can be obtained by applying this procedure iteratively up to \([J_K(\omega)(\omega), T]\). Since the process has \( P\)-a.s. a finite amount of jumps, we can conclude the existence of a unique solution in the sense of definition 2.5.

**Remark 1.** The bound from [3] on the total variation of a deterministic network solution reads in our context as follows

\[
\sum_{e=1}^M \text{TV}_{[a^e, b^e]}(\rho^e(\cdot, t)) + \sum_{e=1}^M |\partial_t q^e(t)|
\leq \text{TV}_{[0, t]}(G_{tn}) + \sum_{e=1}^M \text{TV}_{[a^e, b^e]}(\rho^e(\cdot, 0)) + \sum_{e=1}^M |\partial_t q^e(0)|.
\]

Regarding the proof of theorem [27], the total variation of the network solution can increase through the jumps in the capacity processes and a dependence of the total variation on this number of jumps is indispensable. For a sample path \( \omega \in N^e \) we
postulate the following bound on the total variation

\[
\sum_{e=1}^{M} TV_{[a_e, b_e]}(\rho_e(\cdot, t, \omega)) + \sum_{e=1}^{M} |\partial_t q_e(t, \omega)| \leq TV_{[0, t]}(G_m) + \sum_{e=1}^{M} TV_{[a_e, b_e]}(\rho_e(\cdot, 0)) + \sum_{e=1}^{M} C_{e, \text{max}} N_e(t, \omega),
\]

where we set \(C_{e, \text{max}} = \max C_e\) and \(N_e(t, \omega)\) as the number of jumps of \((\mu_e(t, \omega), t \in [0, T])\) until time \(t\).

3. Numerical treatment. As proposed in [17], we use the fact that a PDE-ODE system can be solved deterministically between the random switching times in the sense of a piecewise-deterministic process [9]. Therefore, we have to face two issues to derive a suitable numerical approximation of the new stochastic network model, in particular the generation of semi-Markov capacities and the approximation of the deterministic PDE-ODE system between the random times. The discretization of the conservation law [12] is done using a standard left-sided Upwind method while the ordinary differential equation for the queue is discretized by an explicit Euler method. However, the reformulation of the simulation algorithm [17] for the capacity process to a semi-Markov process is not straightforward and will be discussed in detail.

3.1. Generation of a semi-Markov process. The general procedure to simulate a semi-Markov process is based on the Markov property of the underlying \((X, J)\)-process. This means, information on the states \(X_n\) and \(J_n\) is sufficient to compute the states \(X_{n+1}\) and \(J_{n+1}\). Once we are able to determine the transition of the \((X, J)\)-process, we can find a corresponding path of the semi-Markov process as the step function given by the \((X, H)\)-path. By defining

\[
P = \lim_{t \to \infty} Q(t)
\]

we get a stochastic matrix which generates the time-discrete Markov chain \((X_n, n \in \mathbb{N}_0)\). We observe the following relation as in [27] by using the definition of conditional probabilities

\[
Q_{lk}(t) = P(H_{n+1} \leq t, X_{n+1} = k|X_n = l) = P(H_{n+1} \leq t|X_{n+1} = k, X_n = l)P(X_{n+1} = k|X_n = l) = F_{lk}(t)P_{lk}
\]

with cumulative distribution function \(F_{lk}(t) := P(H_{n+1} \leq t|X_{n+1} = k, X_n = l)\). Using this notation, the distribution function of the holding time reads

\[
P(H_{n+1} \leq t|X_n = l) = \sum_{k=1}^{m} F_{lk}(t)P_{lk}.
\]

The property

\[
\sum_{k=1}^{m} P_{lk} = 1
\]

ensures to use the composition method to simulate the holding times, see e.g. [22].
Altogether, we end up with Algorithm 1 for the transition of a $(X, J)$-process given by $P, F$ and the values $X_n, J_n$. Therein, the routine $\text{RandGenP}(X_n)$ is a random number generator to the probability vector $(P_{l1}, \ldots, P_{lm})$ and $\text{RandGen}_F(l, X_{n+1})$ generates a random number with cumulative distribution function $F_{li}$.

**Algorithm 1: Transition step for an $(X, J)$-process**

Data: $X_n, J_n$

begin

\[
 X_{n+1} = \text{RandGenP}(X_n);
 H_{n+1} = \text{RandGen}_F(l, X_{n+1});
 J_{n+1} = J_n + H_{n+1};
\]

end

3.2. **Performance measures.** To evaluate the performance of the stochastic network and its numerical behavior, performance measures need to be defined. For our purposes, these are the mean queue-load, the mean outflow, the mean production time and eventually relevant higher moments of the random variables. Accordingly to [17] the outflow of the network is given by

\[
 G_{\text{net}out}(t) = \int_{t_0}^{t} f_M(s, \rho^M(b_M, s))ds
\]

and additionally we define

\[
 G_{\text{net}in}(t) = \int_{t_0}^{t} G_{in}(s)ds,
 q^{\text{net}}(t) = \sum_{e \in E} \int_{t_0}^{t} q^e(s)ds,
 \tau_{\text{prod}}(T) = \inf\{t \geq t_0 : G_{\text{net}out}(t) \geq G_{\text{net}in}(T)\}
\]

as the network inflow, the sum of all queue-loads and the production time of all parts which have entered the production up to time $T$.

Since the measures above are random variables with unknown distributions, some estimators (e.g. moments) are required. We denote as usual by

\[
 \overline{X} := \frac{1}{n} \sum_{i=1}^{n} X_i
\]

the mean value estimator and by

\[
 \sigma^2(X) := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2
\]

the variance estimator for the family of independently, identically distributed (iid) random variables $X_1, \ldots, X_n$. The strong convergence of these estimators follows directly from the law of large numbers, i.e. $G_{\text{net}out}(t), q^{\text{net}}(t)$ are finite for finite time horizons and $\tau_{\text{prod}}(T)$ is $P$-a.s. finite for reasonable parameters or distributions,
respectively. For given, independent generated realizations \(x_1, \ldots, x_n\) of the random variable \(X\), the estimated mean value is given by

\[
\frac{1}{n} \sum_{i=1}^{n} x_i
\]

and the estimated variance of the sample by

\[
\frac{1}{n-1} \sum_{i=1}^{n} \left( x_i - \frac{1}{n} \sum_{i=1}^{n} x_i \right)^2.
\]

To ease the notation, we denote them also by \(\bar{X}\) and \(\sigma^2(X)\), respectively.

4. Relation to queueing theory. In this section, we analyze under which conditions the macroscopic model described in section 2 is related to a classical queueing system and how the latter can be used to compute the expected queue-load and the expected work in progress (WIP) in the steady state of the macroscopic model. We restrict our analysis to a production network consisting of a single queue-processor unit and the comparison with an \((M,G,1)\) queue, i.e., arrivals are Markovian \(M\), service times have a General distribution \(G\) and a single server 1. The processor is assumed to have a two-state capacity process \((\mu(t), t \geq 0)\) with values in \(\{0,c\}\) for some fixed capacity \(c > 0\) such that it is on or off, respectively. Let \(\lambda > 0\) be the constant arrival rate for the Markovian arrivals in the queueing system. This means the inter-arrival times \((H^1_n, n \in \mathbb{N})\) are exponentially distributed with mean \(\lambda^{-1}\) such that the arrival process can be seen as a Poisson process with parameter \(\lambda\). To transfer the arrival process into the macroscopic framework, we set

\[
G_{in}(t) = \sum_{n=0}^{\infty} \frac{1}{H^1_{n+1}} \mathbb{I}_{[J^1_n, J^1_{n+1})}(t),
\]

with

\[
J^1_n = \sum_{k=1}^{n} H^1_k.
\]

If \(t \in [J_k, J_{k+1})\), we get the amount of arrivals until time \(t\) as

\[
\int_0^t G_{in}(s) ds = k + \frac{t - J^1_k}{H^1_{k+1}}
\]

which equals the value of the Poisson process at the times \(t = J^1_k\). The major difference between queueing theory and the macroscopic model arises from the different interpretation of the service time \(S\) which corresponds to the production time in this context. In an \((M,G,1)\) queue the production time can be arbitrarily distributed assuming \(S \geq 0\) and is defined as the time needed to complete exactly one good. If we define the production time in the macroscopic model as the time needed to produce one amount of goods, a processor with large length \(l\) would have a large production time. In this case, the processor is able to absorb parts from the queue although there is no outflow up to now. This leads to a totally different size of the queue in the macroscopic model compared to the queueing system. For this reason, we define the production time \(S\) in the macroscopic model as the time needed to reduce the queue by one amount of goods assuming no inflow during this time. Let
a filled queue with \( q(0) = 1 \) without any inflow be given, then the production time \( S \) is the first time such that

\[
q(S) = 1 - c \int_0^S r(s) \, ds = 0
\]

holds, where \( r(t) = \mu(t)/c \in \{0, 1\} \) is the normalized capacity. By defining

\[
Y(t) = \int_0^t r(s) \, ds,
\]

we can reformulate the problem to

\[
Y(S) = \frac{1}{c}
\]

which has no solution in the usual sense since the function \( Y \) is increasing but not strictly. Hence, to define a solution to equation (17), we define the pseudo inverse

\[
Y^{-1}(x) := \inf\{t \geq 0 : Y(t) \geq x\}
\]

like for cumulative distribution functions and obtain the production time

\[
S = Y^{-1}\left(\frac{1}{c}\right).
\]

Figure 2 shows one realization of \( Y \) and includes the pseudo inverse \( Y^{-1} \) of this sample path.

If we subtract the identity from the inverse, we obtain a piecewise-constant jump process, where the holding times are given by the on-times \((\tau_{on}^k, k \in \mathbb{N})\) and off-times \((\tau_{off}^k, k \in \mathbb{N})\) characterize the jump sizes. Therefore, the inverse can be written as

\[
Y^{-1}(x) = x + \sum_{k=1}^{N_x} \tau_{off}^k
\]
with \((N_x, x \geq 0)\) satisfying
\[
N_x = k \Leftrightarrow \sum_{l=1}^{k} \tau_i^{on} \leq x < \sum_{l=1}^{k+1} \tau_i^{on}
\]
for every \(k \in \mathbb{N}_0\). Now, we can use well-known results from queueing theory, see e.g. [20], to get the average number of parts in the queue \(L_q\) in steady state. From the literature, we know that
\[
L_q = \frac{\varrho + \lambda^2 \text{Var}(S)}{2(1-\varrho)},
\]
\[
\varrho = \lambda E[S]
\]
under the condition \(\varrho < 1\). The first two moments of the random variable \(S\) are
\[
E[S] = \frac{1}{c} + \mathbb{E}[\tau_1^{off}] \mathbb{E}[N_1],
\]
\[
\text{Var}(S) = \text{Var}(\tau_1^{off}) \mathbb{E}[N_1] + \mathbb{E}[\tau_1^{off}]^2
\]
for \(E[(\tau_i^{on})^2] < \infty\) and \(E[(\tau_i^{off})^2] < \infty\). If we now assume that \((\tau_n^{on}, n \in \mathbb{N})\) are iid exponentially distributed random variables, the stochastic process \((N_x, x \geq 0)\) is a Poisson process and \(N_x\) is Poisson distributed with parameter \(\frac{x}{E[\tau_1^{on}]}\). This leads to the more condensed formulas
\[
E[S] = \frac{1}{c} \left( 1 + \frac{\mathbb{E}[\tau_1^{off}]}{E[\tau_1^{on}]} \right),
\]
\[
\text{Var}(S) = \frac{1}{c} \frac{E[(\tau_1^{off})^2]}{E[\tau_1^{on}]}
\]
and together with equations (18)–(19) we are able to calculate the expected number of parts in the queue in steady state.

Next, we are interested in the computation of the expected number of parts being in the system that is given by the sum of the expected number of parts in the queue \(L_q\) and currently in progress \(\varrho\), i.e.,
\[
\hat{L} = L_q + \varrho.
\]
Equation (20) is truly valid for a queueing system but might fail for the determination of an expected work in progress (WIP) in the macroscopic model. We define the WIP as the number of parts in the processor, i.e.,
\[
\text{WIP}(t) = \int_a^b \rho(x,t)dx.
\]
Similar to the calculations above, we need a reformulation of the WIP in the macroscopic model in terms of the random variable \(S\). Therefore, we derive conditions on the length and velocity of a processor such that equation (20) is satisfied and furthermore deduce an equation which is valid for arbitrary parameters. We use again the interpretation of the production time which is defined in the queueing system as the time needed to produce one part. In other words, we need to calibrate a processor such that the production time \(S\) of the macroscopic model equals the time to produce one amount of goods. Let the processor be characterized by its length \(\tilde{l}\), its processing velocity \(\tilde{v}\) and the capacity process \(\mu(t)\) from above. Due to the definition of the queue in the macroscopic model, an initial datum \(\tilde{v} \rho_0 \leq c\) implies the reduction of the conservation law to an advection equation with linear
flux $f(t, \rho) = \hat{v}r(t)\rho$. Let $x$ be the characteristic curve of the advection equation starting in the origin, then we get

$$x(t) = \hat{v} \int_0^t r(s) ds = \hat{v}Y(t)$$

for every $t \geq 0$. Then, the time $\tilde{S}$ to empty the processor fulfills

$$\tilde{x}(\tilde{S}) = \tilde{l} \iff Y(\tilde{S}) = \frac{\tilde{l}}{\hat{v}}. \tag{21}$$

The time $\tilde{S}$ should match with the production time $S$ such that a comparison of equations (17) and (21) implies the relation

$$\frac{\tilde{l}}{\hat{v}} = \frac{1}{c}.$$

If we fix the processing velocity $\hat{v}$ to $v$, the length is

$$\tilde{l} = \frac{v}{c}.$$

Then, equation (20) holds true and the expected WIP for this processor is given by

$$\varrho = \lambda \mathbb{E}[S].$$

In steady state, the expected density is constant and not affected by the length of the processor. Thus, for $\bar{\varrho}$ being the expected steady state density of a processor described by the capacity process $\mu(t)$ and the processing velocity $v$, we have

$$\varrho = \tilde{l}\bar{\varrho} = \frac{v}{c}\bar{\varrho}$$

and for a processor with general length $l$ we have

$$\hat{\varrho} = l\bar{\varrho} = \frac{lc}{v}\varrho.$$

Summarizing, applying known results from queueing theory, the expected amount of goods in the macroscopic queue-processor unit is

$$L := L_q + \hat{\varrho} = \frac{\lambda \mathbb{E}[S] + \lambda^2 \mathbb{V}(S)}{2(1 - \lambda \mathbb{E}[S])} + \frac{lc\lambda \mathbb{E}[S]}{v}. \tag{22}$$

Remark 2. 1. For a single processor in the macroscopic case with stochastic inflow and off-capacities, we are able to calculate the expected queue size and WIP in steady state with equations (18) and (22). The assumption of exponentially distributed on-times is not necessary but allows to state an explicit expression of the moments for the stochastic process $N_x$.

2. The condition $\varrho < 1$ in (19) can be used to analyze whether the macroscopic production model achieves a steady state or increases infinitely in the sense of expectations.

Example. We illustrate the results obtained from (18) and (22) numerically for one example. To do so, we choose a queue-processor unit with length $l = 0.5$, processing velocity $v = 1.5$, and maximal capacity $c = 2$. The capacity process is described by exponentially distributed on-and off-times such that $\mathbb{E}[\tau^m] = 0.9$ and $\mathbb{E}[\tau^f] = 0.1$. This yields a mean availability of 0.9 and we can randomly choose the initial state for the capacity process according to a Bernoulli distribution with $p = 0.9$ to accelerate the convergence to the expected steady states. We run our simulations using $10^4$ Monte-Carlo iterations, time step size $\Delta t = 1/400$ and spatial discretization such
that the CFL condition $\Delta t \leq \Delta x/v$ is satisfied with equality. Figures 3-6 show the sample mean queue-load $\bar{q}$ and total amount of goods $\bar{q} + \text{WIP}$ of the macroscopic system with different arrival rates $\lambda$. The dashed horizontal line represents the value $L_q$ and $L$ calculated by equations (18) and (22). We observe a good approximation of the steady state values in all cases and see a relatively fast convergence towards them independent of the different arrival rates.

![Figure 3. Comparison with $\lambda = 0.1$.](image1)

![Figure 4. Comparison with $\lambda = 0.5$.](image2)

5. **Computational results.** Starting from the original stochastic production network model [17] we have made two essential extensions, namely arbitrary holding times and more general state spaces. The resulting semi-Markov processes are different compared to the CTMCs with simpler state spaces leading to problem-adapted performance measures. Our goal is the numerical comparison of the extended model to the original one and comments on differences and similarities.

5.1. **Impact of state space.** As in section 4 we consider a network only consisting of one queue-processor unit with length $l = 1$ and velocity $v = 1$. To analyze the impact of the state space on the performance measures we distinguish two cases: In the first case, the processor can randomly breakdown and lead to two possible
capacities. In the second case we assume that the capacity can be randomly reduced but not to zero resulting in three possible capacities. Generally, these two capacity processes \( \mu_{\text{Case 1}}(t), t \geq 0 \) and \( \mu_{\text{Case 2}}(t), t \geq 0 \) are diverse and will produce different results. However, to compare both, we choose the transition rates such that the expected stationary capacities are equal. To do so, we use well-known results from the theory of CTMCs, see e.g. [26], and exponentially distributed holding times. The transition rates of the two CTMCs \( \mu_{\text{Case 1}}(t), t \geq 0 \), \( \mu_{\text{Case 2}}(t), t \geq 0 \) are illustrated in figure 7 with a resulting expected stationary capacity of 0.9 in both cases. In detail, the stationary probabilities for the first case are

\[
P(\mu_{\text{Case 1}}(\infty) = 1) = \frac{9}{10} \quad \text{and} \quad P(\mu_{\text{Case 1}}(\infty) = 0) = \frac{1}{10}
\]

and in the second case

\[
P(\mu_{\text{Case 2}}(\infty) = 1) = \frac{9}{11}, \quad P(\mu_{\text{Case 2}}(\infty) = 0.5) = \frac{9}{55} \quad \text{and} \quad P(\mu_{\text{Case 2}}(\infty) = 0) = \frac{1}{55}.
\]

By a suitable choice of the parameters such that the expected stationary capacity equals, we expect the same stationary density and queue-load for both cases. In fact, for a constant inflow of \( G_{\text{in}}(t) = 0.8 \) which is below the expected stationary capacity, we observe this behavior in figures 8, 9 and 10 as a result of a simulation with \( 10^4 \) Monte-Carlo iterations. In figure 8 we show the difference of the first and
second case in terms of sample mean and variance. A fast convergence behavior for increasing time $t$ can be observed. This behavior can be also seen for the sample mean and variance of the queue-loads in figure 9. We suppose that the steady state of the first two moments of the density and the queue-load only depend on the mean capacity in these scenarios.

Due to the fact that

$$G_{\text{net}}(t) = \min\{v\rho(b,t), \mu(t)\}$$

is directly affected by the capacity processes with different variances, there occur different sample variances for the outflow as figure 10 illustrates.

---

**Figure 7.** Graph representation of the CTMCs

**Figure 8.** Difference of the sampled mean an variance densities

**Figure 9.** Sampled mean and variance of the queue-loads
Figure 10. Sample mean and variance of the network outflow

If we focus on the local behavior in time, we conclude from figure 8 a significant difference between the both cases at the boundaries of the two densities. This motivates the analysis of a non-stationary situation, i.e., a time-dependent inflow. We choose

\[ G_{in}(t) = 0.9|\cos(t)| \]

and observe an impact on the sample means and variances as figures 11, 12 and 13 indicate. From the periodic differences in figure 11 we conclude different reaction times of the model for case one and two that lead to different production times.

This motivates the analysis of the production times in both cases, where we set the inflow to

\[ G_{in}(t) = \mathbb{1}_{[0,1]}(t). \]

We might expect a lower production time in the second case arising from the stationary producing probability of $54/55 \approx 0.98$ that is higher than 0.9 from the first case. Table 1 shows the sample mean and variance of the production times. As we can see, they are very close for both cases which is an unexpected result.

Additionally, by comparing the histograms of the production times in figure 14 as an approximation of their distribution, we detect no severe differences between the two cases. One reason for these similar distributed production times may be effects which cancel out, i.e., the stationary probability to have a capacity lower
than one is in case one (1/10) and hence smaller than case two (10/50) but the probability to produce nothing is in the first case higher (1/10) than in the second one (1/55).

5.2. Impact of holding times. Before we analyze the influence of the holding time distribution on the performance measures, we comment on the impact of the holding time distribution. We know that the parameter of the exponential distribution is determined by the mean, i.e., it holds

$$E[X] = \lambda^{-1}$$ for $X \sim \text{Exp}(\lambda)$, $\lambda > 0$

and for the variance

$$\text{Var}(X) = \lambda^{-2} = E[X]^2.$$

If we think about the modeling of an event occurring every 30 days in mean and having a standard deviation of two days, the exponential ansatz would always imply
a standard deviation of 30 days and hence fails. There is a need for alternative distributions with positive support such as e.g. the gamma or Weibull distribution. Here, we focus on the gamma distribution and a comparison with the exponential distribution. The latter is in fact a special case of the gamma distribution. This allows us to vary the parameters near the exponential distribution to investigate the influence on the overall performance of the production network. We follow the convention of MATLAB\textsuperscript{1} for the gamma distribution, this means that
\[
E[X] = d_2 \cdot d_1 \quad \text{and} \quad \text{Var}(X) = d_2^2 \cdot d_1
\]
with \(X \sim \gamma(d_1, d_2), d_1, d_2 > 0\). Given the expected value and the variance, the parameters are determined by
\[
d_1 = \frac{E[X]^2}{\text{Var}(X)} \quad \text{and} \quad d_2 = \frac{\text{Var}(X)}{E[X]}.
\]
Our plan for the analysis of the influence caused by the holding time distribution is keeping the expected value constant and varying the variance of the gamma distribution. An illustration of probability density functions (pdf’s) for the gamma distribution is given in figure 15. An increase of the variance implies an increasing behavior in the expected queue-loads and higher production times. In contrast to

\textsuperscript{1}http://mathworks.com/help/stats/gamma-distribution.html
this, a small variance close to zero implies a concentration of the distribution at the expected value and a more deterministic behavior with deterministic time spans of processing or failing.

\[
\begin{bmatrix}
0 & 0.2 & 0.4 & 0.6 & 0.8 & 1
\end{bmatrix}
\]

**Figure 16.** Serial network with five processors

*Example: A Chain of Five Processors.* We consider a chain of five processors as in [17], where each processor has a length of 0.2 while the random dynamics in the capacities are different. The state space of each capacity process is \( C = \{0, 2\} \) and table 2 shows the Mean Time Between Failures (MTBF) and the Mean Repair Time (MRT) of each processor. Apparently, processor 3 is the bottleneck in the network. The inflow at the first processor is constant \( G_{in}(t) = 1.8 \) and the processing velocity of every processor is \( v = 1 \).

| Processor | 1    | 2    | 3    | 4    | 5    |
|-----------|------|------|------|------|------|
| MTBF      | 0.95 | \( \infty \) | 0.85 | 1.9  | 0.95 |
| MRT       | 0.05 | 0.15 | 0.10 | 0.05 |

**Table 2.** Parameters of the network model with five processors

This information is sufficient for the exponential case whereas further assumptions are needed for the gamma distributed one. Therefore we set for every \( \alpha > 0 \) the Variance of the Time Between Failures (VTBF) and the Variance of the Repair Time (VRT) as

\[
\text{VTBF}(\alpha) = \alpha \cdot \text{MTBF}^2, \\
\text{VRT}(\alpha) = \alpha \cdot \text{MRT}^2.
\]

The case \( \alpha = 1 \) corresponds to the exponential distribution such that we can examine the behavior of the performance measures by varying the variance, i.e., varying \( \alpha \).

The parameters for the gamma distribution are

\[
\alpha_1^{\text{TBF}}(\alpha) = \frac{1}{\alpha}, \quad \alpha_2^{\text{TBF}}(\alpha) = \alpha \cdot \text{MTBF}
\]

for the times between failures and

\[
\alpha_1^{\text{RT}}(\alpha) = \frac{1}{\alpha} \quad \text{and} \quad \alpha_2^{\text{RT}}(\alpha) = \alpha \cdot \text{MRT}
\]

in the case of repair times.

The sampled mean of the density in the exponential case is shown in figure 17, where we use \( 10^4 \) iterations for the Monte-Carlo simulation. If we compare this result to the choice of \( \alpha = 4 \) and \( \alpha = \frac{1}{4} \) in figure 18, we observe a smoother shape for higher variance and vice versa.

As one may expect, the queue-loads and its variance are positive correlated to the variance of the holding time distribution, see table 3. Analogously, the sample
mean and the sample variance of the outflow differ by varying the variances. This can be seen figure 19.

\[
\begin{array}{|c|c|c|c|}
\hline
\alpha & q_{net}(4) & \sigma^2(q_{net}(4)) \\
\hline
\frac{1}{4} & 1.3361 & 0.3278 \\
1 & 1.7391 & 1.2588 \\
4 & 2.8382 & 5.1694 \\
\hline
\end{array}
\]

**Table 3.** Sample mean and variance of the network queue-load

Next, we investigate a range of configurations for \( \alpha \) to capture one fourth up to four times the exponential variance \( \alpha \in \{2^x : x = -2, -1.5, \ldots, 1.5, 2\} \) and compare the performance measures. To see the sensitivity of the model with respect to the parameter \( \alpha \), we plot each result of the simulation in terms of \( \log_2(\alpha) \). If \( \log_2(\alpha) = 0 \) holds, we are in the exponential case again. The number of Monte-Carlo iteration is \( 10^4 \) and we assume an empty system at the beginning.

Figure 20 shows the estimated mean and the estimated variance of the first capacity process \( (\mu^1(t), t \in [0, 4]) \) for different \( \alpha \). The computed convergence of both quantities to a value which is independent of \( \alpha \) is very conspicuous. We suppose that the one dimensional distribution, i.e. the distribution of the random
variable $\mu(t)$ is not severely influenced by the variance of the holding times for large times $t$.

The sampled mean network outflow in figure 21 shows the negative influence of a higher variance although the mean of the distribution is fixed. Together with the positive dependence of the expected queue-load and the variance this implies a relevant connection between the performance of the production network and the chosen probability distribution.

We are also interested in how the variance of the queue-loads and the network outflow behave if we double or half the variance of the intermediate times. In the log$_2$-log$_2$ diagram of figure 22 we observe a linear relationship with a slope of approximately one, i.e., doubling the variance implies a doubled variance of the outflow and the net queue-load. Indeed, we cannot observe this behavior if only one capacity process is changed in a more complex network. In particular, this processor would then imply the slope of the linear relationship. The production time does not follow the linear relationship as figure 23 shows. In fact, the dependence is even in the log$_2$-log$_2$ scaling exponential and therefore very sensitive with respect to the distribution. Summarizing, we can conclude that the choice of the probability distribution for the holding-times strongly affects the performance measures of the production system.
Conclusion. We have extended the stochastic time-continuous production network model with two state CTMC capacities to a setting of capacity processes consisting of independent semi-Markov processes having values in a finite set of possible capacities. The relation of the stochastic model to queueing theory models is explained.
from an analytical and computational point of view. The influence of holding time distributions as well as the impact of the state space have been numerically investigated for different examples. For a better presentation of the results, we have restricted the modeling to serial networks. However, an extension to general networks is straightforward.

Future work includes on the one hand the extension of the stochastic model to load dependent capacities and on the other hand the construction of the transition Matrix \( Q \), for example the analysis of the entries in \( Q \) if the capacity process \((\mu(t), t \geq 0)\) can be represented as the sum of independent capacity processes.

REFERENCES

[1] D. Armbruster, P. Degond and C. Ringhofer, A model for the dynamics of large queuing networks and supply chains \( SIAM J. Appl. Math. \), 66 (2006), 896–920.

[2] J. Banks, J. S. Carson, B. L. Nelson and D. M. Nicol, \( Discrete-Event System Simulation \), 5th edition, Pearson, 2010.

[3] C. D’Apice, S. Göttlich, M. Herty and B. Piccoli, \( Modelling, Simulation, and Optimization of Supply Chains \) 1st edition, SIAM, Philadelphia, 2010.

[4] C. D’Apice, P. I. Kogut and R. Manzo, On relaxation of state constrained optimal control problem for a PDE-ODE model of supply chains \( Netw. Heterog. Media \), 9 (2014), 501–518.

[5] C. D’Apice, R. Manzo and B. Piccoli, Modelling supply networks with partial differential equations \( Quart. Appl. Math. \), 67 (2009), 419–440.

[6] C. D’Apice, R. Manzo and B. Piccoli, Existence of solutions to Cauchy problems for a mixed continuum-discrete model for supply chains and networks \( J. Math. Anal. Appl. \), 362 (2010), 374–386.

[7] C. D’Apice, R. Manzo and B. Piccoli, Optimal input flows for a PDE-ODE model of supply chains \( Commun. Math. Sci. \), 10 (2012), 1225–1240.

[8] C. D’Apice, R. Manzo and B. Piccoli, Numerical schemes for the optimal input flow of a supply chain \( SIAM J. Numer. Anal. \), 51 (2013), 2634–2650.

[9] M. H. A. Davis, Piecewise-deterministic Markov processes: A general class of nondiffusion stochastic models, \( J. Roy. Statist. Soc. Ser. B \), 46 (1984), 353–388.

[10] P. Degond and C. Ringhofer, Stochastic dynamics of long supply chains with random breakdowns \( SIAM J. Appl. Math. \), 68 (2007), 55–79.

[11] L. Forestier-Coste, S. Göttlich and M. Herty, Data-fitted second-order macroscopic production models \( SIAM J. Appl. Math. \), 75 (2015), 999–1014.

[12] D. T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions \( J. Computational Phys. \), 22 (1976), 403–434.

[13] D. T. Gillespie, Approximate accelerated stochastic simulation of chemically reacting systems \( J. Chem. Phys. \), 115 (2001), 1716–1733.

[14] S. Göttlich, M. Herty and A. Klar, Network models for supply chains \( Commun. Math. Sci. \), 3 (2005), 545–559.

[15] S. Göttlich, M. Herty and C. Ringhofer, Optimization of order policies in supply networks \( European J. Oper. Res. \), 202 (2010), 456–465.

[16] S. Göttlich, A. Klar and S. Tiwari, Complex material flow problems: A multi-scale model hierarchy and particle methods \( J. Engrg. Math. \), 92 (2015), 15–29.

[17] S. Göttlich, S. Martin and T. Sickenberger, Time-continuous production networks with random breakdowns \( Netw. Heterog. Media \), 6 (2011), 695–714.

[18] F. Grabski, Semi-Markov failure rates processes \( Appl. Math. Comput. \), 217 (2011), 9956–9965.

[19] F. Grabski, \( Semi-Markov Processes: Applications in System Reliability and Maintenance \) 1st edition, Elsevier, Amsterdam, 2015.

[20] D. Gross, J. F. Shortle, J. M. Thompson and C. M. Harris, \( Fundamentals of Queueing Theory \) 4th edition, John Wiley & Sons, Inc., Hoboken, NJ, 2008.

[21] B. Harlamov, \( Continuous Semi-Markov Processes \) 1st edition, ISTE, London; John Wiley & Sons, Inc., Hoboken, NJ, 2008.

[22] M. Kolonko, \( Stochastische Simulation \) (German) \( Stochastic Simulation \), 1st edition, Vieweg+Teubner Verlag, Wiesbaden, 2008.

[23] A. M. Lee, \( Applied Queueing Theory \) Reprint edition, Macmillan [u.a.], London [u.a.], 1966.
L. Lipsky, *Queueing Theory*, 2nd edition, Springer, New York, 2009.

J. Medhi, *Stochastic Processes*, 3rd edition, New Age Science, Tunbridge Wells, 2010.

J. R. Norris, *Markov Chains*, Reprint edition, Cambridge University Press, Cambridge, 1998.

R. Pyke, *Markov renewal processes: Definitions and preliminary properties*, *Ann. Math. Statist.*, 32 (1961), 1231–1242.

Received July 2016; revised December 2016.

E-mail address: goettlich@uni-mannheim.de
E-mail address: stknapp@mail.uni-mannheim.de