A Quasi-Birth-and-Death Process Approach for Integrated Capacity and Reliability Modeling of Railway Systems*

Norman Weik[a,*], Nils Nießen[a]

[a]*Institute of Transport Science, RWTH Aachen University, 52062 Aachen*

Abstract

A railway system's capacity is an important performance indicator allowing to assess different infrastructure variants and to devise market-compliant schedules. Existing approaches in capacity analysis assume the unrestricted availability and peak performance of all system components. Disruptions leading to infrastructure unavailability and reduced system performance are not considered in long- and medium term tactical planning of capacity.

We present a quasi-birth-and-death process approach for the integrated modelling of capacity and reliability. By allowing for phase-type distributed arrival, service and repair processes the model permits to describe a wide range of schedule and operational characteristics. At the same time, the solvability of Markovian processes and the information on the queue length distribution are preserved. The model is solved using a Krylov-subspace method, which allows to effectively deal with large state spaces and transition matrices.

The approach is compatible to existing queueing-based models in the capacity analysis of railway lines and junctions. The functionality of the method is demonstrated in a case study of a mixed service railway line with infrastructure unavailability.

Keywords: Railways, Infrastructure planning, Capacity, Reliability, Performance modelling, Queueing, Quasi-Birth-and-Death-Processes

1. Introduction and Literature Review

A railway system's capacity is generally viewed as the maximum number of trains which can be operated concurrently with market-compliant quality in a predefined time frame (UIC 406 (2013)). While specific requirements for the level of service vary, capacity remains an important performance indicator allowing to assess different infrastructure variants and to devise market-compliant schedules. The most widespread approaches to measure capacity are either based on infrastructure utilization, e.g. UIC schedule compression method (UIC 406 (2013)), or on quality-related criteria such as train punctuality (Graffagnino and Labermeier (2016)) or delays (e.g. Huisman and Boucherie (2001); Schwanhäußer (1974); Hertel (1992); DB Netz AG (2008a)). Depending on the planning stage and input data availability various approaches to determine the feasible number of trains have been described in the literature.

* This is the Authors' Accepted Manuscript of the following article: N. Weik, N. Nießen, A Quasi-Birth-and-Death Process Approach for Integrated Capacity and Reliability Modeling of Railway Systems, Journal of Rail Transport Planning & Management 7(3), pp. 114-126, 2017, which has been published in final form at https://doi.org/10.1016/j.jrtpm.2017.06.001.

* Corresponding author
Email address: weik@via.rwth-aachen.de (Norman Weik)
URL: http://via.rwth-aachen.de/ ()

Our focus is on long and medium term tactical planning where infrastructure topology and basic operational characteristics are fixed, but the exact schedule is not known or may still change. This is where analytic stochastic and queueing based approaches have found widespread application in modelling the expected load and waiting time.

Potthoff introduced a technique to determine the number of tracks required in stations based on the loss probability in $M/D/n/0$-queueing systems (Potthoff (1962)). Hertel (1984) later proposed a method based on $GI/GI/n/\infty$-models to estimate waiting probabilities in stations for the same task.

For capacity analysis of railway lines Schwanhäußer (1974) developed an approach to approximate the average knock-on delays based on the probability distributions of primary delays and buffer times. The method is based on pairwise correlations between trains and the identification of railway lines with $M/D/1/\infty$-queueing systems. The method has recently been revisited in Weik et al. (2016), where model assumptions and limitations are discussed and a mathematically more rigorous derivation is given. An extension of Schwanhäußer's method to station thresholds has been discussed in Nießen (2013). Effective single-channel queueing systems are constructed where service times are reduced according to the share of mutually non-exclusive train runs on the infrastructure (Nießen (2013)). It is also applicable to bottleneck analysis based on a deconstruction of station thresholds into sectional route notes, where all train runs are mutually exclusive (Schwanhäußer (1994)).

Huisman and Boucherie (2001) also establish a relation between train delays and the utilization of railway lines. Here, railway lines are modelled as infinite server resequencing queues. Delays are due to different running times of different train types, which can be taken to be either deterministic or random, hence allowing to consider primary delays within track segments. The model additionally allows for correlations between interarrival times or between interarrival and service times.

Wendler (2007) proposes a queueing model exhibiting more general, Semi-Markovian service processes, which can be used to determine the scheduled waiting times in capacity allocation. The model is closely related to models used to assess runway utilization in view of different aircraft types in aviation (Bäuerle et al. (2007)). For even more general $GI/GI/1/\infty$-queueing networks Wakob (1985) developed an approximation approach to determine the average scheduled waiting times in station threads. The method is built on a statistical regression model for the waiting times based on the first two moments of the arrival and service process. It has been investigated in detail and validated against empirical data by de Kort et al. (1999).

In Huisman et al. (2002), a Jackson queueing network model for the joint analysis of multiple railway lines has been proposed. The railway network is decomposed into station entries, station exits, and line segments, each being modelled as $M/M/1/\infty$-queueing systems, which ensures the factorization of the stationary probability distribution. Unfortunately, the independence property of different queues in Jackson networks is quickly destroyed once correlations between different queues enter.

Analytical models are particularly suited to cope with uncertain or fragmentary input data as they operate on (empirical) probability distributions or moments of probability distributions. Alternative approaches such as traffic simulations or MIP-based approaches require solving a large number of system realizations in order to obtain statistically reliable results. However, existing analytical approaches generally lack precision for two main reasons: First, in order to obtain solvable models, the exactness of either input or output data is reduced. In the first case, arrival and service processes are approximated by more easily tractable

processes, e.g. Markovian ones (Schwanhäußer (1974); Potthoff (1962); Huisman et al. (2002)). In the latter case, more general arrival and service times can be considered, but the output is limited to mean value data and information on the distribution of waiting times is lost (Hertel (1984); Wendler (2007)). Second, the unrestricted availability of all system components is assumed. Disruptions such as train malfunctions or infrastructure breakdowns can only be considered implicitly as far as they affect the distribution of arrival or service times. Still, they are very important for practical capacity investigations as complex systems such as railway networks are constantly subject to failure and maintenance processes limiting the effectively achievable capacity. A promising line of research to include failure and repair processes in Markovian railway capacity analysis models has been discussed in Bär et al. (1988), but has not been further pursued.

The topic of integrated reliability and performance analysis of railway systems has recently received new attention in reliability analysis and asset-management. In Fecarotti et al. (2013), the resilience of operations on a generic railway line with varying track switching possibilities is studied and optimized. To this end, a systematic failure mode analysis using FMECA is performed and a discrete event simulation describing train operations is employed (Fecarotti et al. (2013)). More recently, the train operation simulation model has been exchanged for a Petri-net based approach in Fecarotti et al. (2015). Still, reliability and performance modelling are separated. A list of timed failure events is generated by a reliability subroutine which is then fed to the simulation of train operations. This can be seen as a performance simulation in a random system environment, yet it prohibits the modeling of load-dependent failures such as train malfunctions.

Our present work picks up on Bär et al. (1988) and complements Fecarotti et al. (2013, 2015) in aiming to provide a more realistic representation of railway systems by an integrated modelling of system availability and performance. Unlike in the models described in Bär et al. (1988), which only allow to consider exponential holding times, a Quasi-Birth-and-Death (QBD) process approach is adopted. By allowing for phase-type distributed arrival and service times, which can be fitted to the moments of given empirical data, this concept provides the flexibility to adequately describe a wide range of schedule and operational characteristics. At the same time, the tractability of Markovian models is preserved and the stationary distribution can be obtained. This not only allows to consider the average capacity of the given railway infrastructure, it also allows to determine the probability that the system performs at a given capacity. This information is highly relevant for infrastructure operators which have to ensure contractually defined standards are held.

In the context of railway operations science the utility of phase-type distributions has been demonstrated before. In Meester and Muns (2007) it has been shown that they are well-suited to model delay propagation in railway networks as they provide a good fit of delay distributions and are closed under the mathematical operations governing delay propagation. Büker and Seybold (2012) build on a similar modelling of delay distributions. In addition, an activity-based framework to formalize delay propagation processes is introduced, which allows to analyze large networks like, for instance, Switzerland.

QBD processes are mathematically well understood and have found widespread application in modelling computer systems, communication networks and supply chains (see e.g. Bolch (2006)). The additional modelling flexibility provided by phase-type distributed holding times comes at the expense of an increase of the size of the state space. Hence, a numerically delicate, memory efficient approach is required to incorporate system availability in the QBD model.

The procedure we propose to handle these difficulties relies on the following four major aspects:

- Different failure modes are aggregated to avoid redundancy if they have comparable effects on service times and exhibit similar failure and repair rates.
- Service times, arrival times and restoration times are fitted by hypoexponential distributions. Hypoexponential distributions are the most memory efficient class of phase type distributions with variation coefficient smaller than 1 (Sommereder (2011)), which is typical for railway applications.
- The sparse transition matrix of the QBD is built from its smaller constituting blocks by exploiting the repetitive tridiagonal block structure. The blocks themselves are initialized as sparse matrices making use of repetitive sub-block structures whenever possible.
- A non-standard approach to determine the stationary distribution is pursued. Rather than employing matrix-analytical techniques such as Neuts' R-matrix approach (Neuts (1981)), which would require the inversion of sparsely populated submatrices of the transition matrix, we opt to solve the Kolmogorow equations directly using a projection-based iterative method (Philippe et al. (1992); Saad and Schultz (1986)). The main effort of this method is due to matrix-vector multiplications which can be performed very efficiently given the sparseness of the transition matrix.

The approach is presented in more detail in the following sections. In Section 2 we describe the fundamentals of the QBD modelling approach, also providing a brief introduction to the concept of phase type distributions and QBD processes in general and hypoexponential distributions in particular. Section 3 discusses the initialization and implementation of the model. In addition, an automated approach to define and aggregate infrastructure availability states based on their effects on service times is presented. In Section 4, the model is applied to a mixed service railway line as a case study.

2. Method

2.1. Phase-Type Distributions

One of the major shortcomings of queueing-based railway capacity analysis models is their limitation to Markovian arrival and service processes. Interarrival times and minimum headways between trains generally exhibit variation coefficients significantly smaller than 1 (Schwanhäußer (1974); de Kort et al. (1999)). Hence, Markovian models tend to overrate the variance of service and arrival processes. The same holds true for restoration processes after failures. This is where phase-type distributions introduce additional flexibility w.r.t. the variation of holding times and provide a versatile tool to approximate general distributions.

2.1.1. Definition

A continuous-time phase-type distribution with m phases is defined by the distribution of the time to absorption of a continuous-time Markov chain (CTMC) with m transient states and one absorbing state. The generator of the CTMC hence takes the form

$$\begin{pmatrix} \mathbf{S} & \mathbf{S}^0 \\ 0 \dots 0 & 0 \end{pmatrix},$$

where $\mathbf{S}$ is an $m \times m$-Matrix describing transitions within the set of transient states and $\mathbf{S}^0$ is an $m \times 1$ vector featuring the transition rates to the absorbing state.

Phase-type distributions find widespread use for mainly two reasons: First, phase-type distributions are dense in the space of all positive-valued distributions (Asmussen (2008)). Hence, for any positive-valued probability distribution, a suitable phase-type distribution correctly representing this distribution can be given. The number of required phases, however, especially in case of small variation coefficients, may be very large (Aldous and Shepp (1987)). At the same time, expanding the state space by interpreting the transient phases as additional states, a CTMC description of processes with phase-type distributed holding times is obtained. The solvability of Markov processes is thus preserved at the expense of a larger state space.

2.1.2. Hypoexponential Distributions

In our model, train service times, interarrival times and infrastructure restoration times are modelled by hypoexponential distributions, a subclass of phase-type distributions which are the most memory efficient for approximating distributions with variation coefficient smaller than 1 (Sommereder (2011)).

For hypoexponential distributions $\mathbf{S}$ and $\mathbf{S}^0$ take the form (Sommereder (2011))

$$\mathbf{S} = \begin{pmatrix} -\mu_1 & \mu_1 & 0 & 0 & \cdots \\ 0 & -\mu_2 & \mu_2 & 0 & \cdots \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & \cdots & 0 & -\mu_{m-1} & \mu_{m-1} \\ 0 & \cdots & 0 & 0 & -\mu_m \end{pmatrix}, \quad \mathbf{S}^0 = \begin{pmatrix} 0 \\ \\ \vdots \\ \\ 0 \\ \mu_m \end{pmatrix}. \tag{1}$$

This means phases are visited consecutively in a given direction without skipping or re-visiting. Erlang(μ, m)-distributions are a subclass of hypoexponential distributions which are obtained if all transition rates between phases are identical ($\mu = \mu_1 = \mu_2 = \ldots = \mu_m$).

2.1.3. Fitting Hypoexponential Distributions to Empirical Data

Methods for fitting phase-type distributions to empirical data can be classified into three categories: Moment-matching, maximum-likelihood and entropy-maximization approaches (Elmaghraby et al. (2010); Reinecke et al. (2012)). In this work, a moment fitting approach based on the first two moments of the hypoexponential distribution is pursued. The approach, which is described in Sommereder (2011), is based on the composition of a hypoexponential distribution from two Erlang distributions and can be implemented very efficiently. The two-moment fitting approach is also consistent to the input required in current queueing-based models in railway capacity analysis, which are based on expectation value and variance of arrival and service times, respectively (Nießen (2014)).

2.2. Quasi-Birth-and-Death Processes

The quasi-birth-and-death (QBD) process model we propose is a joint availability-queueing model allowing to consider failure/restoration-processes and train arrival/service-processes simultaneously. This allows

to represent a wide variety of interactions between the two modelling aspects including availability-dependent service rates or load-dependent failure rates. It also incorporates transitory effects connected to disruptions such as the build-up of queues during failures and the time required to reduce queues after restoration of availability. These effects cannot be considered in an individual capacity assessment of different infrastructure availability states.

Markovian queueing models are widely used in performance analysis of complex systems, including railway systems. Mathematically, these models are birth-and-death processes, a subclass of continuous-time Markov chains with only two types of state transitions: Births (arrivals of customers) and Deaths (completion of service), increasing, respectively decreasing population size by 1. As a consequence, the transition matrix of the process is triangular.

A quasi-birth-and-death (QBD) process is an extension of this model class allowing for a block-tridiagonal structure of the transition matrix. The blocks are referred to as the "levels" of the process whereas states within one block are generally called "phases" (Latouche (2011)). "Levels" correspond to the population size, "phases" are interior system transitions not resulting in a customer entering or leaving the system. The first level deviates from the bulk blocks as it corresponds to the system being empty. This often implies different system behavior manifesting in block size, possible state transitions and transition rates. The transition matrix $\mathbf{Q}$ of a level-independent QBD takes the form

$$\mathbf{Q} = \begin{pmatrix} B_{00} & B_{01} & 0 & \cdots & & & \\ B_{10} & A_1 & A_2 & 0 & \cdots & & \\ 0 & A_0 & A_1 & A_2 & 0 & \cdots & \\ 0 & 0 & A_0 & A_1 & A_2 & 0 & \cdots \\ & & & \ddots & \ddots & \ddots & \end{pmatrix}, \tag{2}$$

where $B_{00}, B_{01}, B_{10}, A_1, A_2$, and A_0 are matrices. If A_1, A_2, and A_0 also depend on the level, the process is called a level-dependent QBD.

If the QBD is viewed as a process operating on a state space consisting of both levels and phases the model becomes a continuous-time Markov chain and can be solved accordingly. Thence, solvability properties of Markovian processes are preserved. The QBD can also be seen as a stochastic process operating on levels only. In this case, the QBD is a Semi-Markovian process, where holding times are governed by an underlying environment process. The process is Markovian only at the jump instances as the direction of a level transition is stochastically independent of the previous state.

2.3. QBD Model for Integrated Capacity and Availability Analysis

The QBD formulation can be used to describe queueing systems where the transitions between different levels, i.e. arrival of new customers and completion of service of customers, follow more general processes. In the context of this paper, the service process is given by a Semi-Markovian process with phase-type distributed holding times depending on infrastructure availability. The arrival process is generalized analogously, which is particularly relevant if periodic schedules are to be represented. An illustrative representation of the QBD queueing system is given in Figure 1, below.

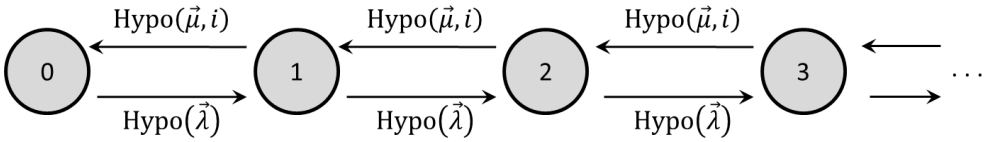

Figure 1: Schematic representation of the QBD queueing model. Transitions between levels are hypoexponentially distributed, where service rates $\vec{\mu}$ depend on availability state i.

We subsequently assume that traffic volume is independent of the system state, i.e. arrival rates do not depend on the availability of the infrastructure. However, the model is not limited to that case, such that, e.g., a restriction of traffic flow to infrastructure segments with failures could be incorporated.

2.3.1. Notation

Let $L = \{0, \dots, L_{max}\}$ be the set of levels, where the maximal possible queue length considered in the model is $L_{\max}$. Furthermore, let I denote the space of different infrastructure availability scenarios (also referred to as failure modes) and $P := \bigcup_{i \in I} \{P(i) = (P_1(i), P_2(i), \dots P_{|p|}(i))\}$ denote the set of phases in the service process, where $|p|$ refers to the number of phases. Accordingly, $T := \{T_1, \dots, T_{|t|}\}$ denotes the set of phases in the arrival process. Finally, the number of elements in a given set X will subsequently be denoted by $|X|$.

2.3.2. Fundamentals

The state space of the QBD model is characterized by tuples $\{(l, i, p, t) : l \in L, i \in I, p \in P(i), t \in T\}$. In order to obtain a two-dimensional representation of the transition matrix the multidimensional array is transformed into an ordered one-dimensional array with hierarchical ordering (level $\rightarrow$ arrival phases $\rightarrow$ service phases $\rightarrow$ infrastructure availability states). This can be seen as a hierarchical state-space with an exterior queueing model for capacity estimation and an integrated interior reliability model. The ordering does not mean the queueing model is given preference over the reliability model as transition rates in the reliability model can be made dependent on the level, for example. However, it is an efficient structure reducing matrix bandwidth given the fact that availability influences service rates, but does not act on the number of customers, i.e. the levels, or the arrival process.

The transition matrix then takes the form in (2). B_{00} is an $|I| \cdot |T| \times |I| \cdot |T|$-matrix describing the state transitions in the first level. As the first level corresponds to the queueing system being empty, no service of customers is performed and B_{00} only contains availability transitions as well as transitions between arrival phases. For any other level, the $|R| \cdot |S| \cdot |T| \times |R| \cdot |S| \cdot |T|$-matrix A_1 describes intra-level transitions. Those transitions comprise changes of the availability state as well as transitions between transient phases in the arrival process and the service process of customers currently receiving service. B_{01} is an $|I| \cdot |T| \times |R| \cdot |S| \cdot |T|$-matrix describing the arrival of a customer to the empty system. Accordingly, B_{10} is an $|R| \cdot |S| \cdot |T| \times |I| \cdot |T|$-matrix and describes the completion of the service of the last customer in the system. A_2 and A_0 describe the arrival of customers and completion of services for all other levels and, like A_1, are $|R| \cdot |S| \cdot |T| \times |R| \cdot |S| \cdot |T|$-matrices.

We subsequently discuss the structure of the submatrices in more detail. The discussion is split in two parts, the first one dealing with the design of the exterior queueing model for capacity modelling, the second one dealing with the substructure of the incorporated reliability model for availability prognosis.

2.3.3. Queueing Submodel

For notational clarity the number of arrival phases is assumed to be 1 in the subsequent discussion of QBD block matrices. In this case, the block-matrices A_1, A_2, and A_0 have the following form:

$$A_1 = \begin{pmatrix} S_1 & S_{12} & 0 & \cdots & \\ 0 & S_2 & S_{23} & 0 & \cdots \\ & & \ddots & \ddots & \\ 0 & \cdots & 0 & S_{|p|-1} & S_{|p|-1,|p|} \\ 0 & \cdots & 0 & 0 & S_{|p|} \end{pmatrix}, \tag{3}$$

$$A_0 = \begin{pmatrix} 0 & 0 \\ S_{|p|,1} & 0 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} \lambda & 0 & \cdots & \\ 0 & \lambda & 0 & \cdots \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & \lambda \end{pmatrix}. \tag{4}$$

It can be seen that the shape of A_1 resembles the shape of the transition matrix of the hypoexponential distribution in (1). However, due to the integration of the reliability model, S_k, $S_{k,k+1}$ ($k \in \{1, \ldots, |p|-1\}$) are $|I| \times |I|$-matrices.

$$S_k = R - \begin{pmatrix} \mu_{k,k+1}^{(1)} + \lambda & 0 & \cdots & \\ 0 & \mu_{k,k+1}^{(2)} + \lambda & 0 & \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & \mu_{k,k+1}^{(|I|)} + \lambda \end{pmatrix} \tag{5}$$

$$S_{k,k+1} = \begin{pmatrix} \mu_{k,k+1}^{(1)} & 0 & \cdots & \\ 0 & \mu_{k,k+1}^{(2)} & 0 & \\ \vdots & & \ddots & \\ 0 & \cdots & 0 & \mu_{k,k+1}^{(|I|)} \end{pmatrix} \tag{6}$$

The super-index denotes the infrastructure availability state upon which service rates depend. $S_{k,k+1}$ denotes transitions between subsequent service phases k and $k+1$. The first summand R describes the availability changes within service phase k and will be discussed in more detail in the following section. The second summand is a diagonal matrix containing the exit rates of phase k in level l. Exits either occur due to completion of a service phase $\mu_{k,k+1}$ or due to the arrival of a new customer λ, leading to a level change. Note that for numerical treatment, where only a final number of levels can be considered, the matrices S_k

in the last level will slightly differ. As no additional arrivals are possible any more in this level S_k will not feature arrival rates on the diagonal.

Upon completion of the final phase of service, the service of the current customer is finished and he leaves the system. Hence, the level is reduced by 1. These level transitions are modelled by A_0. As the service of the next customer starts in phase 1 again, A_0 only has a non-zero block of size $|I| \times |I|$ in the lower left corner. The structure of $S_{|p|,1}$ is identical to the structure of $S_{k,k+1}$ and contains the service rates $\mu_{|p|}^{(i)}$, $(i \in I)$ in the final phase on the diagonal.

To generalize to $|T|$ arrival phases, the matrices A_1, A_2 and A_0 in Equations 3 and 4 have to be diagonally repeated $|T|$-times with corresponding values λ_i, $i \in \{1, \ldots, |T|\}$ in S_k, $S_{k,k+1}$.

2.3.4. Reliability Submodel

The reliability submodel describes the transitions between different availability states $i \in I$. In reliability analysis Markovian models have found widespread application in modelling system availability. The transition matrix R hence is an arbitrary CTMC transition matrix. Its layout depends on the failure and repair processes defining the possible transitions between different availability states. Failure processes are typically well-described by Markovian processes with exponentially distributed times between failures. As repair processes – like service times in the queueing context – generally exhibit variation coefficients smaller than 1, a phase-type distribution is more suitable in that case. As this results in a CTMC formulation with extended state space model characteristics remain unchanged. Note that, in spite of the use of phase type distributions in this case, the reliability submodel is generally no QBD process as there is no ascending order of availability states.

2.4. Solving the Equilibrium Equations

The stationary probability distribution π of the availability-capacity model is obtained by solving the generating equations

$$0 = \pi \cdot Q, \tag{7}$$

subject to normalization $\sum_i \pi_i = 1$.

For QBD processes, matrix-analytic techniques exploiting the block-tridiagonal structure provide an efficient way to determine π, see e.g. (Grassmann and Standford (2000)). The approach goes back to (Neuts (1981)) who showed that the solution can be obtained iteratively by a Matrix-geometric series

$$\pi_{n+1} = R \cdot \pi_n,$$

where the Matrix R only depends on the sub-blocks of the QBD transition matrix Q. As the calculation of R requires inverting $(I - A_1)$ the method is particularly well-suited to analyze QBDs with small and moderate block size. In our case, the blocks encompass the entire availability submodel, which makes inverting block-matrices very costly. In addition, this ports a high risk of running out of memory as the inverse of $(I - A_1)$ is no longer sparse. For an explicit calculation of R-matrices in case of queueing systems with hypoexponential service times see Marin and Rola-Bulo (2014).

We therefore calculate π using GMRES (**G**eneralized **M**inimal **Res**idual Method) (Saad and Schultz (1986)). GMRES is an iterative projection-based solution technique to linear systems of equations. The

problem is projected to a Krylov-subspace $\mathcal{K}_s(Q, r) = \langle r, Qr, Q^2r, \ldots, Q^{s-1}r\rangle$ and solved on this significantly smaller subspace. The cutoff criterion for the Krylov-space dimension s is based on the norm of the residual r, hence the name of the method. As the numerical effort is predominantly due to matrix-vector multiplications Krylov-space projection methods are particularly well-suited for large sparse systems. To increase convergence speed a preconditioning of the equations based on an incomplete LU-factorization (ILU) is performed (cf. Saad and Schultz (1986)).

GMRES can even be used for singular systems in Markov process application. Applications to Markov chain modelling have been discussed in Saad (1995) and Philippe et al. (1992), for example. In this work, MATLAB's GMRES-routine (MATLAB (2016)), which is based on Walker (1988), has been used. It has been slightly modified in order to start the iteration instead of returning the trivial solution in case of singular systems. More recently, modifications ensuring the GMRES algorithm does not get trapped in a subspace of the solution space have been discussed (Reichel and Ye (2005)). However, we found MATLAB's GMRES routine performs well and did not encounter any solvability issues if ILU preconditioning is used.

3. Implementation

This section deals with the implementation of the QBD model. Input data requirements as well as the initialization of the transition matrix are discussed. In addition, a fully automated approach to set up the reliability submodel in the absence of expert knowledge or detailed data on the significance of different failure modes is discussed. It only requires failure and repair statistics of individual system components and provides a model reduction to cope with the problem of an exploding availability state space if large systems are considered.

3.1. Initialization of the Model

In order to initialize the QBD model the transition matrix Q has to be populated. For a given infrastructure, the input data required for our approach consists of

- the definition of failure modes and possible transitions between failures,
- mean times to failure and the first two moments of repair times,
- first two moments of the distribution of interarrival times,
- the first two moments of the distribution of train service requirements for each failure mode, and
- a basic operations concept specifying train types, required stops and the frequencies of the train types in operations.

Train service requirements are equated with minimum headway times on the infrastructure, which is a widely used modelling approach in railway capacity analysis (Pachl (2008); Nießen (2014)).

The failure rates in the reliability submatrix R within the QBD transition matrix are given as the inverse of the mean time to failure (MTTF) and can be inserted directly. Repair, arrival and service processes are approximated by phase-type distributions. Hence, the corresponding rates are obtained as the result of

fitting a hypoexponential distribution to match the first two moments according to the discussion in Section 2.1.3. Depending on the variation, the number of required phases is specified. It is given by $k := \left\lceil \frac{1}{v_x} \right\rceil$, where v_x is the variation coefficient of the corresponding data x.

Currently, our method works with a fix number of service (and arrival) phases for each infrastructure state i, such that all subblocks $S_k, S_{k,k+1}$ in A_1 (cf. Equation (3)) have the same length and the repetitive structure of the submatrices in Q can be exploited. If the number of required service or repair phases varies for different infrastructure states (failure modes) i, k is set to the maximum number of required phases. For other availability states, remaining phases are filled with transitions rates several orders of magnitude larger, such that those phases are traversed by the process quasi-instantaneously.

To control memory usage, a system-dependent cutoff on the maximally admissible number of phases is implemented. Apart from computing resources it notably depends on the number of infrastructure states $i \in I$ as well as the sparsity of the transition matrix R in the interior CTMC process modelling availability changes.

Up to this point it has implicitly been assumed that empirical data for failure and maintenance processes is available. In particular, it has been claimed minimum headway times are known for each failure mode. While infrastructure managers keep extensive maintenance databases registering type and frequency of component failures it generally cannot be assumed that minimum headway statistics are available for each failure mode. As minimum headway times are only required in operational planning and capacity analysis tasks it is highly unlikely minimum headways have been calculated for each combination of individual component failures. This would imply a strong entanglement between asset management and operational planning, which is only starting to emerge in today's railway industry. We therefore subsequently present a fully automated approach to set up the model even if no information about service capabilities in case of failures (Section 3.2) or the importance of individual failures or combinations of failures (Section 3.3) are available.

3.2. Service Time Calculation

Based on an XML-based infrastructure data exchange format such as railML (Nash et al. (2004)) or the XML-ISS format used in Germany (Brünger and Gröger (2003)), the infrastructure is read into MATLAB (2016) and an infrastructure graph is constructed. In the present work XML-ISS has been used. It comprises information about all components of the interlocking system including possible train paths and track clearance detection equipment. Failures of system components such as signals, switches or track vacancy detection equipment resulting in inaccessibility of track segments or velocity reductions can be considered by removing elements resp. adding velocity constraints in the infrastructure graph.

For a given failure mode, i.e. a given availability state of the system, trains are routed automatically on the infrastructure graph using Dijkstra's algorithm (Dijkstra (1959)). Routing is performed based on path length as a distance measure. For the railway line considered in the case study in Section 4 pathlength and train running time are equivalent. In general, train running time might be preferable as a distance measure. For the present work, however, the first approach is pursued as the information about track velocities is stored pointwise in the input data (in terms of velocity changes). Hence, an additional data pre-processing

step would be required to attribute graph edges with velocities, such that running time could be used as a distance measure.

Apart from a rudimentary train operating concept providing train categories (traction characteristics, train start and end points, required stops and holding times in stations) and their relative frequencies in operations no additional information is required. After trains have been routed the subgraphs of infrastructure elements jointly used by two trains are extracted and minimum headway times between trains can be calculated. By accounting for all combinations of succeeding trains the statistics of minimum headway times is obtained.

In case of infrastructure failures trains frequently have to be rerouted to track segments also used by trains travelling in the opposite direction (cf. Figure 2). This entails waiting times due to ongoing occupations of the bidirectional track segment by crossing traffic.

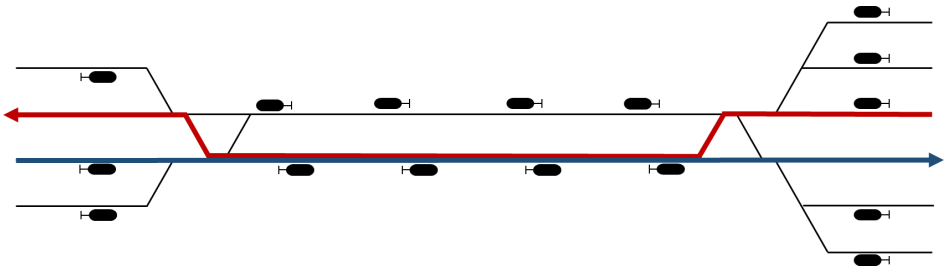

Figure 2: Bidirectional segment on a double track railway line.

The corresponding waiting times add to the running time of trains and hence to minimum headway times of unidirectional trains. It is assumed trains travelling in different direction are uncorrelated, such that – on average – the waiting time corresponds to $\frac{t_{occ}}{2}$, where t_{occ} is the occupation time of the commonly used track segment by crossing traffic. Note that, in practice, bunches of same-directional trains will be channeled through single-track bottlenecks to optimize capacity usage. Hence, waiting times will generally depend on the operating strategy, as well. This could be considered by using conditional probability distributions to determine waiting times, for example.

3.3. Setup of the Reliability Submodel

As the number of fallible components grows the number of different failure combinations grows exponentially, such that the QBD model cannot possibly cope with all combinations of infrastructure failures. In addition, different failures often have identical effects on system performance such that a full-scale analysis of all possible failure combinations would come at the expense of a high degree of redundancy. We therefore include a model reduction step in the availability modelling which consists of the following procedure:

A hierarchical ordering of failure modes is introduced, where Level 0 corresponds to the system operating at peak performance, Level 1 to single component failures, Level 2 to simultaneous failures of two components, etc. For the present work, a cutoff is introduced at Level 3. That means it is assumed that no trains can be run on a railway line if more than two components have failed. As railway infrastructure components are required to be highly reliable simultaneous failures of 3 and more components are highly unlikely. Neglecting this type of events hence corresponds to cutting the tails of the distribution of failures.

For Level 1, all individual failures are analyzed and clustered according to their effects on the distribution of minimum headway times, which are calculated according to the procedure discussed in Section 3.2. The clustering is performed based on the first two moments of minimum headway times, which determine the hypoexponential service times in the queueing model. The distance criterion for the clustering of failure modes with similar minimum headway times is tunable and can be adjusted to system characteristics and computational resources. It can also be set to zero such that only failure combinations with identical moments of service times are aggregated.

As a third model reduction step, a (statistically significant) sample of s cluster elements is drawn from each Level 1 – cluster. For the hence generated subset of infrastructure components, simultaneous failures of pairs of elements are analyzed and aggregated to clusters in the same way as on Level 1. The sampling mainly concerns large clusters, which usually correspond to failures in siding tracks in stations leaving minimum headway times practically unchanged.

If no cutoff is introduced at Level 3 the same procedure can be transferred to higher levels to account for triples of simultaneous component failures, etc. The hierarchical ordering and clustering of failure modes is illustrated in Figure 3. Routing between clusters on different levels is determined based on the relative occurrence frequency of cluster elements in clusters on the next higher level. The corresponding failure and restoration rates are aggregated from the failure statistics of the clusters' individual elements. This sets up a reduced availability state space consisting of failure modes clustered according to the effects on operations and the corresponding transition rates.

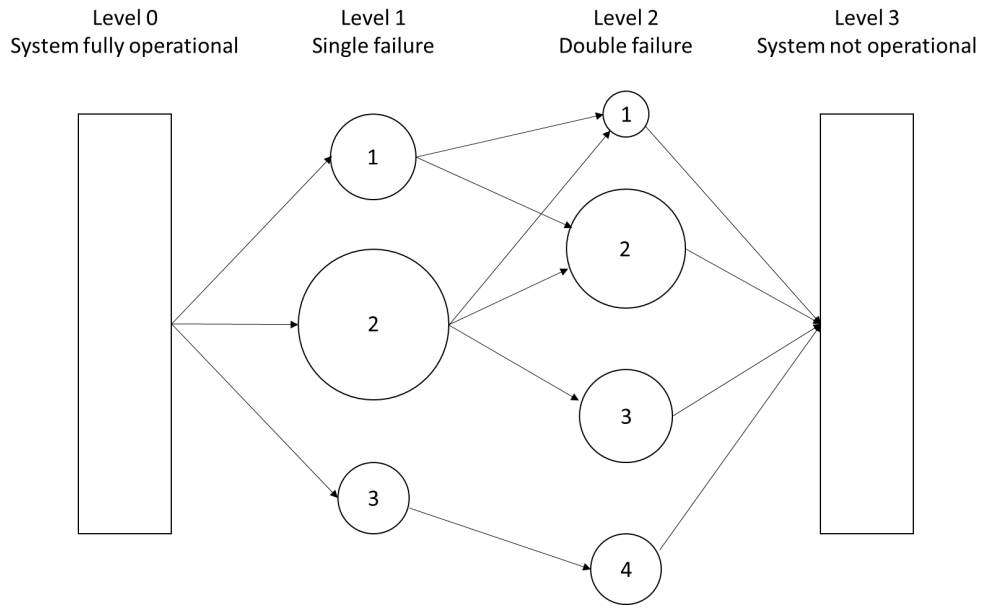

Figure 3: Clustering of failure modes and transitions in the reliability submodel.

4. Case Study – Capacity of a Railway Line in view of Infrastructure Unavailability

4.1. Infrastructure Layout and Train Program

As a test case we consider a generic railway line which has been constructed in LUKS® (Janecek and Weymann (2010)) with reference to German line standards for mixed service lines defined in DB Netz AG (2008b). It is a double track line, which is approximately 120 km long and consists of 4 medium size stations with at least 6 tracks and 9 siding stations with 4 tracks (see Figure 4) and has a maximal admissible velocity of 230 km/h. The distance between sidings is 8 to 15 km and the block length varies between 1 and 2 km.

The train program consists of four different train types including long distance high speed trains, regional trains, local commuter trains and freight trains. The driving characteristics (acceleration) of the different train types are taken according to Steimel (2006).

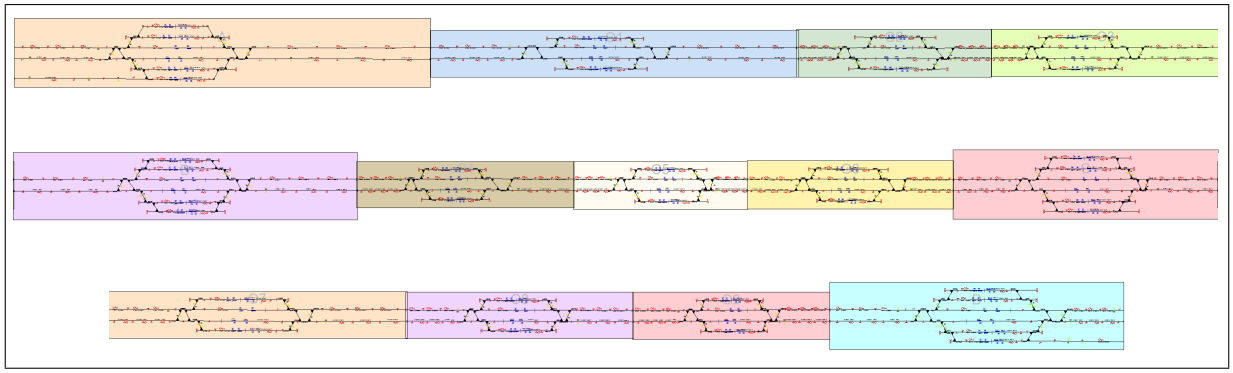

Figure 4: Railway line. The line has been cut in three segments for the illustration. The right hand side in the first two lines is connected to the left hand side in the following line.

4.1.1. Failures

For the test case failures corresponding to inaccessibilities of track segments between switches are considered. These segments represent the smallest elements of network topology required for train routing. Their unavailability can best be represented by failures of corresponding links in the infrastructure graph.

While rail track failures certainly cause this type of failure they may also be provoked by switch failures entailing the inaccessibility of adjacent track segments. As switch lifetime is generally shorter and failures are more frequent than rail track failures (cf. turnout failure statistics in British railways in Hassankiadeh (2011) we consider switch failures to have a major influence on the availability of track segments. Still, considering track segments rather than point elements has the advantage that partial failures of switches, e.g. blocking in locked state, where some adjacent track segments remain accessible, can be accounted for.

The frequency of switch failures is known to be highly dependent on the conditions of their use and exterior influences such as weather (Hassankiadeh (2011)). We therefore subsequently consider three different failure scenarios for track segments and compare the corresponding results to the optimal performance of the system, where the infrastructure is fully available. Repair processes in all three scenarios are assumed to be identical for better comparison. The basic parameters of the three scenarios are given in Table 1.

Scenario 1 corresponds to high reliability. Track segments are assumed to fail once every 100000 hours, on average. In Scenario 2 the mean time between failures is 10000 hours, which corresponds to approximately 1 failure per year and is motivated by typical inspection and maintenance intervals, which are of the order of 3 – 12 months (DB Netz AG (1998)). Finally, Scenario 3 exhibits an average time between failures of just 1700 hours. Such (temporary) low reliability could become relevant in case of adverse weather, for example. The repair time statistics is assumed to be identical in all three cases with an average time to restoration of 3 hours and a coefficient of variation of 0.73.

Table 1: Input parameters for the reliability model. $\mathbf{v}_{\text{restoration}}$ denotes the variation coefficient of restoration times.

	mean time to failure	**mean time to restoration**	$\mathbf{v}_{\text{restoration}}$
Scenario 1	100 000 h	3 h	0.73
Scenario 2	10 000 h	3 h	0.73
Scenario 3	1 700 h	3 h	0.73

The main focus of this work is the integrated modeling of capacity and reliability. A detailed analysis of failure modes in railway systems exceeds the scope of the paper. However, the the model layout as well as the clustering routine presented in Section 3.3 can easily be adapted to a more refined reliability model incorporating various types of failures and a more rigorous failure mode characterization. For the failure classification, a top-down approach based on basic infrastructure elements such as track segments, signals or velocity constraints, which act on train routing and running times, seems most appropriate. Supplementing FMEA used by Fecarotti et al. (2013, 2015) this would also allow to consider multiple simultaneous failures in reliability modelling, which we expect to be particularly important for the quality of train operations.

4.2. Results

For the three failure scenarios the model is initialized according to Section 3 and solved for the stationary queue length distributions using the GMRES-solver with ILU-precon-ditioning.

In our implementation 16 levels, 128 service phases, 2 arrival phases and exact clustering of failure modes, i.e. only failure states with identical moments of minimum headway times are aggregated, have been considered. The exact clustering introduces approximately 750 different failure modes in the reliability submodel, which results in about 1500 infrastructure availability states for two interior phases in the repair process. The resulting transition matrices of the entire QBD process are of size $5 \cdot 10^6 \times 5 \cdot 10^6$. Still, the solution of the preconditioned Kolmogorow equations up to a relative residual of $\mathcal{O}(10^{-9})$ on a computer with Intel i5-4590 dual core (3.3 GHz) chip and 8 GB RAM was obtained within 9 to 10 iterations in less than $10s$. The ILU-factorization itself took about $30s$ for matrices of this size. This shows the model is extremely competitive in spite of the large state space. Given an occupation ratio of 0.6% of the reliability submatrix R memory requirements would allow for an additional increase of the state space by a factor of 2 on the computer system mentioned before, such that even more infrastructure states or phases could be considered.

For the model parameters discussed above, arrival rate $\lambda = 0.08$ and $v_{arr} = 0.8$, and the failure scenarios given in Table 1, the results are presented in the following. Figures 5, 6 and 7 show the queue length distribution in linear (left) and semi-logarithmic (right) scaling for the three scenarios, respectively. For comparison, the queue length distribution in the reference scenario, where the system is operating at peak efficiency, is added. The (red) vertical line in the linear plots corresponds to the level of service for the admissible queue length according to German railway operation guidelines (DB Netz AG (2008a)).

Note that "queue length" in the QBD context refers to the number of trains in the system including those currently receiving service. Hence, queue length 0 corresponds to the system being empty, queue length 1 to one train being in service and 0 trains waiting, etc. It is only for queue lengths larger than 1 that trains are waiting for their service to start. This fine difference of definitions of queue lengths has to be observed when comparing to the level of service defined in DB Netz AG (2008a).

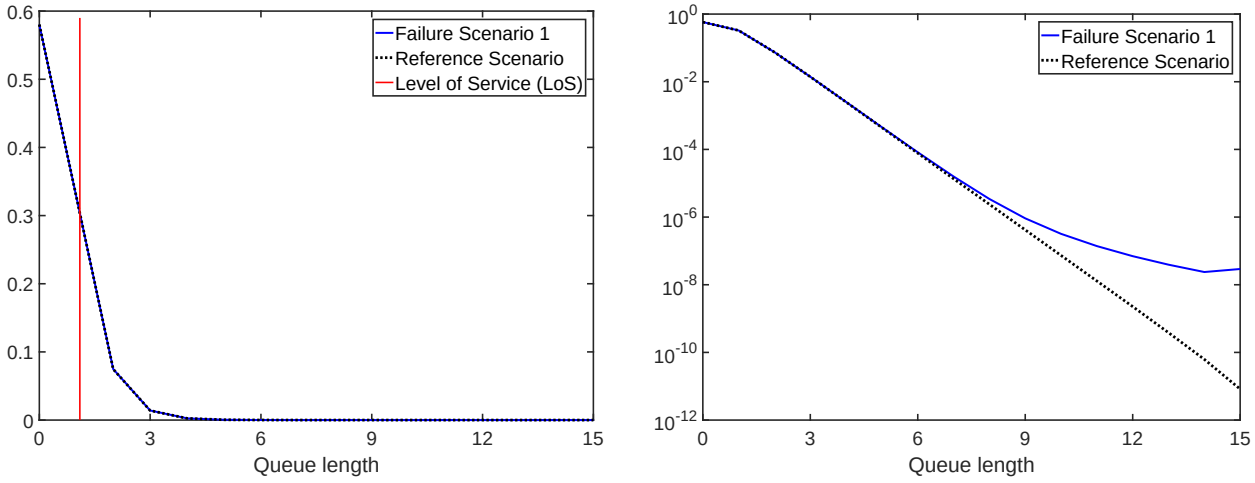

Figure 5: Queue length distribution for Failure Scenario 1 in linear (left) and semi-logarithmic (right) scale

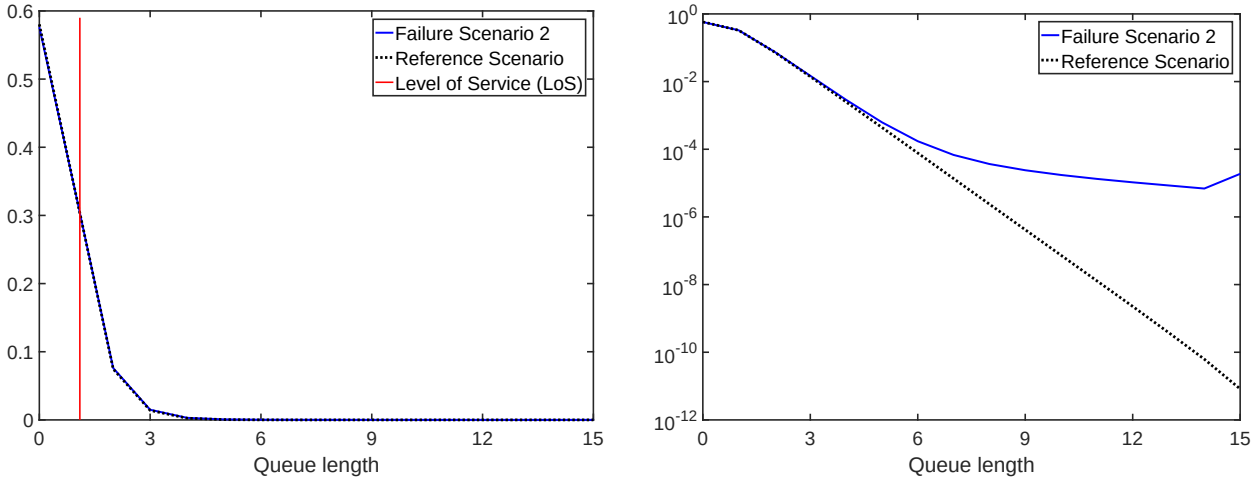

Figure 6: Queue length distribution for Failure Scenario 2 in linear (left) and semi-logarithmic (right) scale

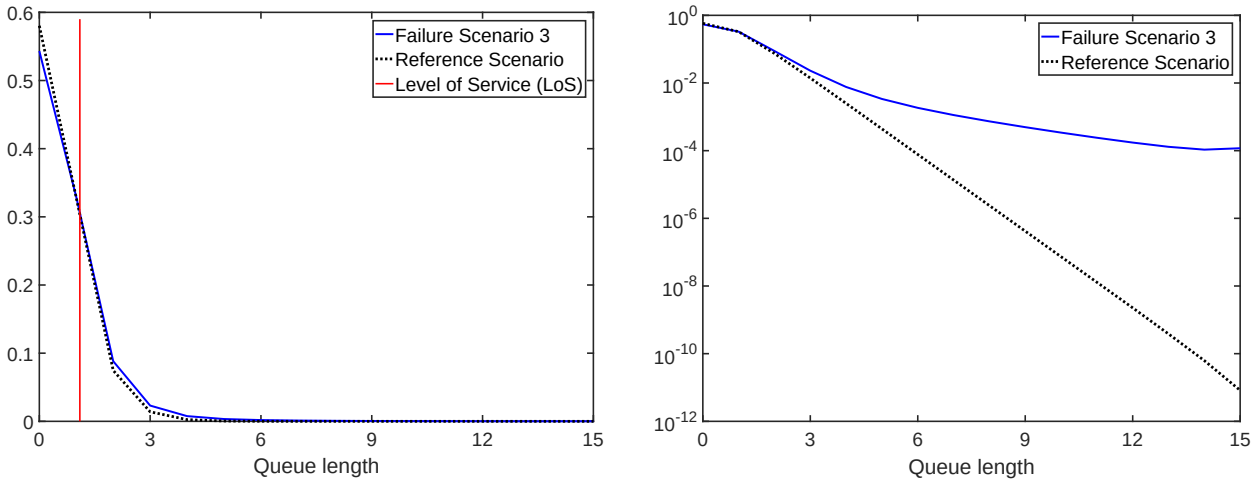

Figure 7: Queue length distribution for Failure Scenario 3 in linear (left) and semi-logarithmic (right) scale

In normal scaling, the queue length distributions show only minor deviations from the queue length distribution in the reference scenario. Only in the third scenario, corresponding to failure rates of 1/1700 h, the weight of the queue length distribution in the admissible region below the level of service starts to visibly decrease and is shifted to the tail of the distribution. In the semi logarithmic depiction in the right hand side of Figure 5 to 7, however, it can be seen that even in Scenarios 1 and 2 the shape of the queue length distributions deviates profoundly from the reference scenario. Whereas the decrease of the probability distribution in the reference case is exponential, the consideration of failures results in a much slower decrease in all three scenarios. This corresponds to a significant increase of the probability that long queues form. By comparing the three failure scenarios it can also be seen that failure rates have huge effects on the probability distribution in this region. Between Scenario 1 and 2 an increase of the stationary probability for $L_Q = 15$ of several orders of magnitude is observed.

Additionally, a small peak becomes visible in the semi logarithmic plot at $L_Q = 15$ in all three scenarios. Analyzing the probability distribution of availability states in the last level an increased weight in states with total blockage of the railway line was observed. In this case, service can only recommence once the blockage is lifted by repairs. As repair rates are much smaller than train arrival rates long queues form and the system saturates in the final state due to the finite system size.

Quantitatively, the stationary probability for the system being in a state below the maximal admissible queue length is 90.83% in the reference scenario. It reduces moderately to 90.81% and 90.42% in Scenario 1 and 2 and to 87.23% in Scenario 3. While the decrease of weight in the admissible region seems insignificant in Scenario 1 and 2 the increase of weight in the tail region does lead to remarkable changes of average queue length and waiting time of trains. Identifying the queue length with the number of waiting trains according to German capacity analysis rules defined in DB Netz AG (2008a) $E[L_Q]$ in the three failure scenarios increases by 0.8%, 12.4% and more than 85%, respectively. Note that traffic density in the reference scenario has been adjusted to a point slightly beyond the optimal level of service, but still in the optimal region ($L_{Q,opt} \pm 20\%$) of capacity usage on railway lines according to DB Netz AG (2008a).

A detailled comparison of the statistical properties of the queue length distributions in the three failures scenarios is given in Table 2. Quantiles are given by $Q_c := \min\{x : P(X \leq x) \geq c\}$ and mean waiting times of trains are calculated from mean queue lengths using Little's law.

Table 2: Queue length statistics. $E[L_Q]$ = avg. number of waiting trains, $E[T_W]$ = avg. waiting time [min], Q_c =c-Quantiles, $L_{Q,opt}$ = optimal queue length.

	$\mathbf{E[L_Q]}$	$\mathbf{E[T_W]}$	$\mathbf{Q_{98}}$	$\mathbf{Q_{99}}$	$\mathbf{L_{Q,opt}}$	**unsatisf. LoS**
Ref. Scenario	0.112	1.343	1	2	0.097	9.17%
Scenario 1	0.113	1.354	1	2	0.097	9.19%
Scenario 2	0.126	1.510	1	2	0.097	9.58%
Scenario 3	0.208	2.496	2	3	0.097	12.77%

It can be seen that failures indeed significantly affect railway capacity. In the present work only a single type of failures – failures of track segments between switches – has been considered. We expect the effect of reduced infrastructure availability on capacity to be larger, and to become signficantly visible even in the high reliability scenario, once additional failure types such as e.g. signal failures are incorporated.

5. Conclusion

Following up on previous work by Bär et al. (1988) a new QBD-model for the joint modelling of capacity and system reliability has been discussed. By matching phase-type distributions to the first two moments of minimum headway, interarrival and repair times it provides an accurate modelling of train operation and failure characteristics. The integrated modelling of capacity and infrastructure availability allows to consider correlations such as state-dependent service rates or load-dependent failure rates.

In a test case the model was applied in capacity analysis of a double track mixed service railway line subject to the blockage of track segments. It was shown that preconditioned GMRES allows to efficiently solve the Kolmogorow equations for the stationary distribution in spite of the large state space resulting from the incorporation of the reliability subroutine and the use of phase-type distributions. The model can easily be adapted to capacity analysis of stations and station threads, where it can supplement existing queueing based approaches. As our approach preserves the information about the queue length distribution it can be used to estimate the waiting probability of trains, for instance.

The queue length distribution can also be used to estimate the share of time the system is operating at a prescribed level of service. The interplay of queue length and availability states is expected to provide insights into critical infrastructure components w.r.t. the quality of operations. Furthermore, the model allows to study the effects of different maintenance strategies on system capacity. This could benefit maintenance and infrastructure planning by facilitating cost-revenue analysis.

A desirable next step would be to complement the QBD modelling approach with fault tree analysis, hence allowing for a more systematic classification and importance sampling of failure modes. This is especially relevant if a large number of different component failures and multiple failure types are to be analyzed and clustered according to their effects.

Acknowledgments

This work was supported by German Research Foundation (DFG) grant NI 1597/2-1 "Integral capacity and reliability analysis of guided transport systems based on analytical models".

References

Aldous, D. and Shepp, L., 1987. "The least variable phase type distribution is Erlang", *Commun. in Statistics. Stochastic Models*, vol. 3 (3), pp. 467–473, https://doi.org/10.1080/15326348708807067.

Asmussen, S., 2008. "Applied Probability and Queues". In *Stochastic Modelling and Applied Probability*, vol. 51, Springer, New York.

Bär, M., Fischer, K., Hertel, G., 1988. "Leistungsfähigkeit, Qualität, Zuverlässigkeit", Transpress VEB Verlag für Verkehrswesen, Berlin.

Bäuerle, N., Engelhardt-Funke, O., Kolonko, M., 2007. "On the waiting time of arriving aircrafts and the capacity of airports with one or two runways", *European Journal of Operational Research*, vol. 177 (2), pp. 1180–1196, https://doi.org/10.1016/j.ejor.2006.01.002.

Bolch, G., Greiner, S., de Meer, H., and Trivedi, K.S., 2006. "Queueing networks and Markov chains: modeling and performance evaluation with computer science applications", 2nd edition, John Wiley & Sons, New York.

Brünger, O. and Gröger, Th., 2003. "Fahrplantrassen managen und Fahrplanerstellung simulieren". In: *Proceedings of 19. Verkehrswissenschaftliche Tage (VWT)*, Dresden, Germany.

Büker, Th., and Seybold, B. "Stochastic modelling of delay propagation in large networks", *Journal of Rail Transport Planning & Management*, vol. 2 (1–2), pp. 34–50. https://doi.org/10.1016/j.jrtpm.2012.10.001.

DB Netz AG, 2008a. Richtlinie 405 – Fahrwegkapazität. Berlin.

DB Netz AG, 2008b. Richtlinie 413 – Infrastruktur gestalten. Frankfurt.

DB Netz AG, 1998. Richtlinie 821 – Oberbau inspizieren. Frankfurt.

De Kort, A.F., Heidergott, B., van Egmond, R.J., and Hooghiemstra, G., 1999. "Train Movement Analysis at Railway Stations: Procedures & Evaluation of Wakob's approach", Delft University Press, Delft (NL).

Dijkstra, E.W., 1959. "A note on two problems in connexion with graphs", *Numerische Mathematik*, 1 (1), pp. 269–271.

Elmaghraby, S. E., Benmansour, R., Artiba, A., and Allaoui, H., 2010. "On the approximation of arbitrary distributions by phase-type distributions", In: *Proceedings 3rd International conference on information systems, logistics and supply chains ILS 2010*, Casablanca, Morocco.

Fecarotti, C., Andrews, J., Remenyte-Prescott, R., 2013. "Modelling Railway Service Reliability", In: *Proceedings 20th Advances in Risk and Reliability Technology Symposium*, pp. 259–273, Loughborough, UK.

Fecarotti, C., Andrews, J., Remenyte-Prescott, R., 2015. "Modelling railway service reliability in the event of failures". In: *Proceedings of the 25th European Safety and Reliability Conference (ESREL 2015) annual conference*, Zurich, Switzerland.

Graffagnino, Th., and Labermeier, H., 2016. "A definition of timetable stability for a long-term timetable". In: Brebbia, Mera, Tomii, Tzieropoulos (eds.), *Computers in Railways XV*, pp. 91–100, WITPress, Southampton.

Grassmann, W.K., and Stanford, D.A., 2000. "Matrix Analytic Methods". In: Grassmann, W.K. (ed.), *"Computational Probability"*, pp. 153–203. Springer US, Boston.

Hassankiadeh, S.J., 2011. "Failure Analysis of Railway Switches and Crossings for the purpose of Preventive Maintenance", *Master Thesis*, Royal Institute of Technology, Stockholm.

Hertel, G., 1984. "Exakte Lösung zur Berechnung der Wartegleiszahl vor im Einrichtungsbetrieb befahrenen Streckengleisen bei Nicht-Poisson-Ankünften (G/M/1–Wartesystem)", *Wissenschaftliche Zeitschrift Hochschule für Verkehrswesen*, vol. 31, pp. 195-205.

Hertel, G., 1992. "Die maximale Verkehrsleistung und die minimale Fahrplanempfindlichkeit auf Eisenbahnstrecken", *Eisenbahntechnische Rundschau (ETR)*, vol. 41 (10), pp. 665–672.

Huisman, T., and Boucherie, R.J., 2001. "Running times on railway sections with heterogeneous train traffic", *Transportation Research Part B: Methodological*, vol. 35 (3), pp. 271–292, https://doi.org/10.1016/S0191-2615(99)00051-X.

Huisman, T., Boucherie, R.J., van Dijk, N.M., 2002. “A solvable queueing network model for railway networks and its validation and applications for the Netherlands”, *European Journal of Operational Research*, vol. 142 (1), pp. 30–51, https://doi.org/10.1016/S0377-2217(01)00269-7.

Janecek, D., Weymann, F., 2010. “LUKS – Analysis of lines and junctions”. In: *Proceedings of the 12th World Conference on Transport Research (WCTR)*, Lissabon, Portugal.

Latouche, G., 2011. “Level-Independent Quasi-Birth-and-Death Processes”, Wiley Encyclopedia of Operations Research and Management Science. John Wiley, New York.

Marin, A., and Rota Bulo, S., 2014. “Explicit solutions for queues with Hypo- or Hyper-Exponential service time distribution and application to product-form approximations”, *Performance Evaluation* vol. 81, pp. 1–19, https://doi.org/10.1016/j.peva.2014.07.021.

Matlab R2016a (version 9.0.0). The MathWorks Inc., Natick, Massachusetts.

Meester, L.E., and Muns, S. “Stochastic delay propagation in railway networks and phase-type distributions”, *Transportation Research Part B: Methodological*, vol. 41 (2), pp. 218–230 https://doi.org/10.1016/j.trb.2006.02.007.

Nash, A., Huerlimann, D., Schütte, J., and Krauss, V.P., 2004. “RailML – a standard data interface for railroad applications”, In: Hansen, I.A. (Ed.) *“Timetable Planning and Information Quality”*, pp. 3–10. WIT Press, Southampton

Neuts, M.F., 1981. “Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach”, John Hopkins University Press, Baltimore.

Nießen, N., 2014. “Queueing” In: Hansen, I., and Pachl, J. (eds.) *Railway timetabling and operations*, DVV Media Group Eurailpress, Hamburg.

Nießen, N., 2013. “Waiting and loss probabilities for route nodes”, In: *Proceedings of The 5th International Seminar on Railway Operations Modelling and Analysis (RailCopenhagen2013)*, Lyngby, Denmark.

Pachl, J., 2008. “Timetable Design Principles”, In: Hansen, I.A., and Pachl, J. (eds.), *Railway Timetable and Traffic: Analysis, Modelling, Simulation*, EurailPress, Hamburg.

Philippe, B. and Saad, Y. and Stewart, W. J., 1992. “Numerical methods in Markov chain modeling”. *Operations Research*, vol. 40 (6), pp. 1156–1179, https://doi.org/10.1287/opre.40.6.1156.

Potthoff, G., 1962. “Verkehrsströmungslehre Band 1: Die Zugfolge auf Strecken und in Bahnhöfen”, *Transpress VEB*, 1st edition.

Reichel, L. and Ye, Q., 2005. “Breakdown-free GMRES for singular systems”. *SIAM Journal on Matrix Analysis and Applications*, vol. 26 (4), pp. 1001-1021, https://doi.org/10.1137/S0895479803437803.

Reinecke, P., Bodrog, L. and Danilkina, A., 2012. “Phase-type Distributions”. In *“Resilience Assessment and Evaluation of Computing Systems”*, pp. 85–113, Springer, Heidelberg, Berlin.

Saad, Y., 2003. *“Iterative methods for sparse linear systems”, 2nd edition*, SIAM, Philadelphia.

Saad, Y., 1995. “Preconditioned Krylov Subspace Methods for the Numerical Solution of Markov Chains” In: Stewart, W.J (ed.), *”Computations with Markov Chains: Proceedings of the 2nd International Workshop on the Numerical Solution of Markov Chains”*. Springer US, Boston, https://doi.org/10.1007/978-1-4615-2241-6_4.

Saad,Y., and Schultz, M.H., 1986. “GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems”. *SIAM Journal of Scientific and Statistical Computing*, vol. 7 (3), pp. 856–869, https://doi.org/10.1137/0907058.

Schwanhäußer, W., 1978. “Die Ermittlung der Leistungsfähigkeit von großen Fahrstraßenknoten und von Teilen des Eisenbahnnetzes”, *Archiv für Eisenbahntechnik*, vol. 33, pp. 7–18.

Schwanhäußer, W., 1984. “Die Leistungsfähigkeit moderner Eisenbahnstrecken”, *Internat. Verkehrswesen*, vol. 36, pp. 32–37.

Schwanhäußer, W., 1994. “The status of German railway operations management in research and practice”, *Transportation Research Part A: Policy and Practice*, vol. 28 (6) pp. 495–500, https://doi.org/10.1016/0965-8564(94)90047-7.

Schwanhäußer, W., 1974. Die Bemessung der Pufferzeiten im Fahrplangefüge der Eisenbahn. *Veröffentlichungen des Verkehrswissenschaftlichen Instituts der RWTH Aachen*, 20, Aachen.

Sommereder, M., 2011. “Modelling of Queueing Systems with Markov Chains: An Introduction to Basic and Advanced Modelling Techniques”, BoD, Norderstedt, Germany.

Steimel, A., 2006. *“Elektrische Triebfahrzeuge und ihre Energieversorgung: Grundlagen und Praxis”*, 2nd edition. Oldenbourg Industrieverlag, München, Germany.

UIC, 2013. Code 406 – Capacity, 2nd edition.

Wakob, H., 1985. “Ableitung eines generellen Wartemodells zur Ermittlung der planmäßigen Wartezeiten im Eisenbahnbetrieb unter besonderer Berücksichtigung der Aspekte Leistungsfähigkeit und Anlagenbelastung“, *Veröffentlichungen des Verkehrswissenschaftlichen Instituts der RWTH Aachen*, 36, Aachen.

Walker, H.F., 1988. ”Implementation of the GMRES Method Using Householder Transformations”. *SIAM Journal of Scientific and Statistical Computing*, vol. 9, pp. 152–163, https://doi.org/10.1137/0909010.

Weik, N., Niebel, N., Nießen, N., 2016. “Capacity analysis of railway lines in Germany – A rigorous discussion of the queueing based approach”, *Journal of Rail Transport Planning & Management*, vol. 6 (2), pp. 99–115, https://doi.org/10.1016/j.jrtpm.2016.06.001.

Wendler, E., 2007. “The scheduled waiting time on railway lines”, *Transportation Research Part B: Methodological*, vol. 41 (2), pp. 148–158, https://doi.org/10.1016/j.trb.2006.02.009.