Measurement and Optimization of Robust Stability of Multiclass Queueing Networks: Applications in Dynamic Supply Chains

Michael Schönlein\textsuperscript{a}, Thomas Makuschewitz\textsuperscript{b}, Fabian Wirth\textsuperscript{a}, Bernd Scholz-Reiter\textsuperscript{b}

\textsuperscript{a}Institute for Mathematics, University of Würzburg, Emil-Fischer Straße 40, 97074 Würzburg, Germany (e-mail: \{schoenlein,wirth\}@mathematik.uni-wuerzburg.de).

\textsuperscript{b}BIBA - Bremer Institut für Produktion und Logistik GmbH, University of Bremen, Hochschulring 20, 28359 Bremen, Germany (e-mail: \{bsr,mak\}@biba.uni-bremen.de).

Abstract
Multiclass queueing networks are an essential tool for modeling and analyzing complex supply chains. Roughly speaking, stability of these networks implies that the total number of customers/jobs in the network remains bounded over time. In this context robustness characterizes the ability of a multiclass queueing network to remain stable, if the expected values of the inter-arrival and service times distributions are subject to uncertain shifts. A powerful starting point for the stability analysis of multiclass queueing networks is the associated fluid network. Based on the fluid network analysis we present a measure to quantify the robustness, which is indicated by a single number. This number will be called the stability radius. It represents the magnitude of the smallest shift of the expected value of the interarrival and/or service times distributions so that the associated fluid network looses the property of stability. The stability radius is a worst case measure and is a conceptual adaptation from the dynamical systems literature. Moreover, we provide a characterization of the shifts that destabilize the network. Based on these results, we formulate a mathematical program that minimizes the required network capacity, while ensuring a desired level of robustness towards shifts of the expected values of the interarrival times distributions. This approach provides a new view on long-term robust production capacity allocation in supply chains. The capabilities of our method are demonstrated using a real world supply chain.

Keywords: Robustness and sensitivity analysis, Stochastic processes, Queueing, Uncertainty modeling, Supply chain management

1. Introduction
Multiclass queueing networks are an effective tool for modeling complex dynamic networks, such as supply chains (Meyn (2008)). Typically a multiclass queueing network pictures jobs or customers, arriving over time, that are waiting for service in buffers in front of servers or stations. After service completion, a served job either moves to another buffer at some station or leaves the network. Multiclass queueing networks can capture multiple product lines and model a generic network structure that includes re-entering connections. Furthermore, the servicing sequence of jobs can be controlled by service disciplines, e.g. 'first-in-first-out' or 'processor sharing'.

A multiclass queueing network is called stable, if the underlying Markov process is positive Harris recurrent, cf. Bramson (2008). Dai (1995a) and Stolyar (1995) presented an approach to the investigation of stability of queueing networks using fluid (limit) models. A fluid approximation model is a continuous deterministic analogue of a given discrete stochastic model. The stability of the corresponding fluid limit model implies stability of the original queueing network. However, the behavior of the multiclass queueing model might differ from the behavior of the real world system due to uncertain parameters. A model, which is chosen on the basis of a 'best guess' is referred to as the nominal model. In the mathematical systems theory literature usually a set of parameterized systems is considered, where the parameter vectors lie in a given neighborhood around the parameters of the nominal model, in order to handle uncertainty, see Hinrichsen and
Pritchard (2005). In this context robustness of a nominal model characterizes admissible deviations of the model parameters such that the property of stability is kept. Robustness might also assess the uncertainty in real world parameters for which stability can be guaranteed.

In order to measure and optimize robust stability of a nominal model, derived from a multiclass queueing network, we present a new approach that originates from queueing theory, mathematical systems theory and mathematical optimization. In line with Meyn (2008) we limit our analysis to uncertainties of the interarrival and service times distributions, as uncertainties of the routing parameters concern the flexibility of network. That is, throughout this paper we make the assumption that the topology of the nominal network is fixed.

In the queueing literature, although the term robustness is not used, the related analysis of the stability region has been an active field over the last 15 years. The notion stability region of a fluid network has been introduced by Dai (1995b). Given the topology of the network and the discipline, the set of arrival rates and service rates that lead to a stable fluid network is called the stability region. Furthermore, the investigation of the global stability region has received considerable attraction. The global stability region of a fluid network with given topology is defined as the set of arrival and service rates such that the corresponding fluid network is stable under any work-conserving discipline, cf. Dai (1995b). Chen (1995) showed that the global stability region is monotone with respect to the arrival rates. For two station fluid networks the global stability region is monotone with respect to the arrival rates and service capacities, see Dai and Vande Vate (2000). However, counterexamples given by Dumas (1997) and Bramson (1994) show that, in general, the stability region is neither convex nor monotone with respect to the service rates. In particular, Section 5.4 in Bramson (2008) provides a comprehensive discussion of the global stability region.

Dai et al. (1999) have considered, for fixed arrival rates, the monotone global stability region with respect to the service capacities, defined by the largest monotone subset of the stability region. They have also shown that, in general, the monotone global stability region does not equal the global stability region. In the stability analysis of queueing networks Lyapunov methods are frequently used, see Bramson (2008); Dupuis and Williams (1994); Schönlein and Wirth (2012). In particular, in the context of non-Markovian processes a useful approach is the perturbed Lyapunov function method introduced in Kushner (1984). This method consists of a combination of a standard Lyapunov function as e.g. a measure of queue lengths with perturbation terms introduced to the problem at hand, see e.g. Ying-Chao and Michailidis (2012) for an interesting application.

Robust stability analysis has received considerable attraction in mathematical systems theory over the last 25 years. The fundamental idea of our quantitative approach to robust stability is based on the introduction of a measure for the perturbations that is indicated by a single real number, which is called the stability radius. Given a nominal system and a set of feasible perturbations the stability radius represents the magnitude of the smallest perturbation for which the perturbed system loses the property of stability. The stability radius is a worst case measure in the sense that there might be perturbations larger in magnitude than the stability radius for which the perturbed system is stable. However, the crucial point is that for any perturbation strictly smaller in magnitude of the stability radius, the system is guaranteed to be stable. The notion of the stability radius originates from the dynamical systems literature and was introduced by Hinrichsen and Pritchard (1986a,b). An exposition can be found in Hinrichsen and Pritchard (2005).

We note that specifying the stability region is challenging. Furthermore, in general analytic descriptions of the stability region are hardly available. Therefore, a calculation of the stability radius for such disciplines is not possible. Nevertheless, for some disciplines and networks of special structure there are conditions at hand, which determine the stability region precisely. For instance, the nominal workload condition \( \rho < e \) constitutes the stability region of fluid networks under Head-of-the-Line Proportional Processor Sharing (HLPPS) discipline.

In this paper we consider queueing networks operating under the HLPPS discipline. According to Bramson (1996) the stability region of a HLPPS fluid network with given topology is completely described by the fact that the nominal workload at every station is strictly less than one. We will derive bounds and quantitative information about shifts of the parameters of the fluid network that lead to an unstable network. Loosely speaking, for a stable nominal network we will introduce a measure for the distance to the boundary of the stability region, determining its robustness. Further, we will show that the stability radius of a HLPPS fluid network can be computed by an
optimization problem. Moreover, we characterize perturbations that lead to an unstable network. Besides, the stability radius of HLPPS fluid networks provides an upper bound for the stability radius of the same fluid network operating under any work-conserving discipline. A related approach based on robust optimization for the performance analysis of multiclass queueing networks can be found in Bandi and Bertsimas (2012) and Bertsimas et al. (2011).

The characterizations of destabilizing perturbations provide a valuable starting point for robust production capacity allocation in dynamic supply chains. Suppose that the nominal model and the real world system are in line, then the robustness can be interpreted as the shift of relevant parameters of the real world system before it becomes unstable. This may model the evolution of relevant external and internal parameters, which is not known a priori and makes long-term capacity planning a challenging task. Here, production capacity levels have to be sufficient for managing operational disruptions as well as requirements arising from long-term trends of key parameters. Such trends have to be anticipated by the supply chain management while allocating the required production capacity to each production location in the network (Van Landeghem and Vanmaele, 2002).

Nowadays strategic management decisions related to network design and the allocation of production capacities are usually based on future scenarios. These scenarios project the evolution of key parameters, e.g. customer demand, and allow finding a reasonable network configuration with stochastic programming and robust optimization, cf. Bai et al. (1997), Mulvey et al. (1995) and Roy (2010). Since the evolution of key parameters is not known a priori, the creation of reasonable future scenarios is a challenging task, which has a significant impact on planning results.

In the context of dynamic supply chains the stability radius is interpreted as the smallest shift/trend of average customer demand and/or available production capacity that destabilizes the supply chain. This characteristic is very powerful, since it allows guaranteeing robustness for all future parameter scenarios that do not exceed the stability radius. Optimized parameter settings of the supply chain that comply with desired robustness characteristics can be obtained with mathematical programs. First results of this approach and a four step framework for robust production capacity allocation were presented in Scholz-Reiter et al. (2011):

1. Set-up of a multiclass queueing network modelling the real world supply chain.
2. Approximate the queueing network by its fluid model.
3. Minimize the required production capacities with a mathematical program that is based on the stability radius, so that some robustness property is kept.
4. Transfer of the results to the real world supply chain.

The remainder of the paper is organized as follows. In Section 2 we briefly describe multiclass queueing networks and introduce the associated fluid network. In particular, we recall the basic assumptions that allow for a stability analysis based on the associated fluid network. In Section 3 we recall the definition of stability for fluid networks and summarize known stability results. Furthermore, we provide a formal definition of the stability radius for fluid networks and present its characterization for three different kinds of parameter uncertainty. In Section 4 we introduce different objectives for the optimal parameter adjustment of robust stability in networks and formulate a mathematical program that minimizes the required service rates for keeping a desired level of robustness. Subsequently, our results are applied to a real world supply chain of a European pump set manufacturer in Section 5. Conclusions and an outlook on future research are given in Section 6.

2. Queueing networks and associated fluid networks

We recall briefly the description of multiclass queueing networks. For further details, the interested reader is referred to Dai (1995a) and Bramson (2008). A multiclass queueing network consists of $J$ service stations and $K$ classes of jobs. The interarrival times for jobs of class $k \in \{1,...,K\}$ are given by positive random variables $a_k(n)$, with $n = 1,2,3,...$, and the service times of class $k$ jobs are given by positive random variables $s_k(n)$, with $n = 1,2,3,...$. Each job class
is exclusively served at a certain station. The many-to-one mapping \( c : \{1, \ldots, K\} \to \{1, \ldots, J\} \) determines which job class is served at which station. The corresponding \( J \times K \) matrix \( C = (C_{jk}) \), called the constituency matrix, is defined by \( C_{jk} = 1 \) if \( c(k) = j \) and \( C_{jk} = 0 \) else. For station \( j \in \{1, \ldots, J\} \), the set \( C(j) = \{ k \in \{1, \ldots, K\} : c(k) = j \} \) is the collection of all job classes that are served at the station \( j \). After a class \( k \) job received service at the station \( c(k) \) its routing is given by a \( K \) dimensional Bernoulli random variable \( \phi^k \). To be precise, each component of \( \phi^k(n) \) is either 0 or 1, but the entry 1 appears at most once. Let \( a \) primitive increments of the queueing network. Based on the only finitely many stations before leaving the network. Then, the \( n \)th served class \( k \) job at station \( c(k) \) becomes a class \( l \) job after service completion if \( \phi^k(n) = e_l \) and the job leaves the network if \( \phi^k(n) = 0 \). Further, the buffer at each station is assumed to have infinite capacity. Before summarizing the standing assumptions, we recall that the spectral radius of a matrix \( M \in \mathbb{R}^{K \times K} \) is defined by \( \rho(M) := \sup\{|\lambda| : \exists x \neq 0 \in \mathbb{R}^K : Ax = \lambda x\} \).

The following general assumptions are made:

(A1) The interarrival and service times and the Bernoulli routing are assumed to be i.i.d. and mutually independent.

(A2) The first moments satisfy for all \( k \in \{1, \ldots, K\} \) that \( \mathbb{E}[a_k(1)] = \alpha_k^{-1} < \infty \) \( \mathbb{E}[s_k(1)] = \mu_k^{-1} < \infty \) and \( \mathbb{E}[\phi^k(1)] = P_k \) and the spectral radius of the matrix \( P = [P_1 \ldots P_K] \) is strictly less than one.

(A3) The interarrival times are unbounded and spread out, cf. Dai (1995a).

The parameter \( P_{kl} \) reflects the probability that a class \( k \) job becomes a class \( l \) job and, hence, \( 1 - \sum_{l=1}^{K} P_{kl} \) represents the probability that a class \( k \) job is leaving the network after service completion. Since the spectral radius of \( P \) is strictly less than one, almost surely every job visits only finitely many stations before leaving the network.

Next, we introduce the performance processes and the dynamic equations describing the evolution of the queueing network. Based on the primitive increments \((a, s, \phi)\) we define the primitive cumulatives \((E, S, R)\), where \( E_k(t) \) counts the arrivals of class \( k \) jobs from outside up to time \( t \), i.e.

\[
E_k(t) := \max\{n \in \mathbb{Z}_+ : a_k(1) + \ldots + a_k(n) \leq t\}.
\]

The process \( S(t) \) counts the service completions in the time period \([0, t]\), where \( S_k(t) \) is defined by

\[
S_k(t) := \max\{n \in \mathbb{Z}_+ : s_k(1) + \ldots s_k(n) \leq t\}.
\]

The process \( R(n) \) is called the routing process. For class \( k \) the routing process is defined by

\[
R_k(n) := \sum_{i=1}^{n} \phi^k(i).
\]

Furthermore, the evolution of the queueing network is also determined by the allocation process, denoted by \( T := \{T(t), t \geq 0\} \). Precisely, \( T_k(t) \) denotes the cumulative amount of time that station \( c(k) \) has spent on processing class \( k \) products in the time period \([0, t]\). The allocation process \( T \) is determined by the service discipline. Given the allocation process \( T \), the number of service completions of class \( k \) customers up to time \( t \) is given by \( S_k(T_k(t)) \). Further, the number of class \( l \) customers that have routed from class \( l \) to class \( k \) in the time period \([0, t]\) is given by \( R^k_l(S_l(T_l(t))) \). The queue length of class \( k \) jobs at time \( t \), denoted by \( Q_k(t) \), can be described by the following balance equation

\[
Q_k(t) = Q_k(0) + E_k(t) + \sum_{l=1}^{K} R^k_l(S_l(T_l(t))) - S_k(T_k(t)).
\]

where \( Q_k(0) \) denotes the initial number of jobs in the network. A complete description of the evolution of the queueing network requires additional equations, see Bramson (2008) and Dai (1995a). There, a precise definition for stability of multiclass queueing networks can be found. Roughly speaking, a queueing network is said to be stable if the total number of jobs in the network remains bounded over time. This can also be interpreted in the way that the long-run
input rate of the network equals the long-run output rate. The stability of multiclass queueing networks may be analyzed by rescaling the primitive cumulatives and taking the limits, i.e. for $q \to \infty$ consider

$$
\frac{1}{q} E_k(qt) \to \alpha_k t, \quad \frac{1}{q} S_k(qt) \to \mu_k t, \quad \frac{1}{q} R_k^i(qt) \to p_{lk} t.
$$

Then, the associated fluid network is obtained by replacing the primitive cumulatives $(E, S, R)$ with their first moments $(\alpha, \mu, P)$. The associated flow-balance equation in the continuous deterministic associated fluid network is of the form

$$
Q_k(t) = Q_k(0) + \alpha_k t + \sum_{i=1}^{K} p_{lk} \mu_i T_i(t) - \mu_k T_k(t).
$$

Also, let $Q(t) = (Q_1(t) \ldots Q_K(t))^T$. Again, there are additional conditions on $Q$ and $T$ that are specific to the individual service discipline. In summary, Dai (1995a) has shown that, under the assumptions (A1)-(A3), the stability of the associated fluid network is sufficient for the stability of the multiclass queueing network. Consequently, there is a purely deterministic criterion for the stability of a queueing network.

In this paper, we consider the head-of-the-line proportional processor sharing discipline (HLPPS). HL means that a station processes only one job of each type at any time. Under the HLPPS discipline all nonempty job classes present at a station are served simultaneously proportional to their queue length. Thus, if $\sum_{j \in C(u)} Q_j(t)$ is positive, then the allocation rate $T_k(t)$ for class $k$ jobs is $T_k(t) = \frac{Q_k(t)}{\sum_{j \in C(u)} Q_j(t)}$. Note that, even if $\sum_{j \in C(u)} Q_j(t) = 0$ the allocation rate of class $k$ jobs may still be positive. To state the dynamic equations completely, we have to introduce further processes. The idle process is denoted by $I = \{I_j(t) : t \geq 0\}$, where $I_j(t)$ is the cumulative time that station $j$ idles in the interval $[0, t]$. The immediate workload process is given by $W(t) = C M^{-1} Q(t)$, where $M = \text{diag}(\mu)$. We assume that the network has the work-conserving property, i.e. the idle time of class $k$ jobs can only increases if the immediate workload of station $j$ is zero. Moreover, we note that the processes introduced above are Lipschitz, cf. Chen (1995) and Dai (1995a). Thus the processes are differentiable almost everywhere (a.e.). Using $e := \sum_{k=1}^{K} e_k = (1 \ldots 1)^T$ the dynamic equations of the fluid network under HLPPS discipline can be summarized as follows

$$
Q(t) = Q(0) + \alpha t - (I - P^T)MT(t) \geq 0, \quad (1)
$$

$$
I(t) = et - CT(t), \quad (2)
$$

$$
0 = \dot{I}_j(t) W_j(t) \quad \text{a.e., for all } j \in \{1, \ldots, J\}, \quad (3)
$$

$$
\dot{T}_k(t) = \frac{Q_k(t)}{\sum_{j \in C(u)} Q_j(t)} \quad \text{if } \sum_{j \in C(u)} Q_j(t) > 0. \quad (4)
$$

In the following we denote a fluid network by $(\alpha, \mu, P, C, \pi)$, where $\pi$ indicates the discipline. Any pair $(Q(\cdot), T(\cdot))$ satisfying the associated fluid network equations is called a pair of $(\alpha, \mu, P, C, \pi)$ and the set of all fluid level processes is denoted by $Q(\alpha, \mu, P, C, \pi)$, in short we write $\mathcal{Q}$.


In this section we present a framework for the analysis of robust stability of fluid networks. To this end, we recall the definition of their stability and define a measure for robustness with respect to uncertain arrival rates and service capacities. In particular, we present a scheme to calculate the stability radius for HLPPS fluid networks.

**Definition 1** A fluid network $(\alpha, \mu, P, C, \pi)$ is called stable, if there exists a finite time $\tau \geq 0$ such that $Q(\tau + \cdot) \equiv 0$ for any fluid level process $Q(\cdot) \in \mathcal{Q}$ with $\|Q(0)\| = 1$.

We call a fluid network $(\alpha, \mu, P, C, \pi)$ unstable if it is not stable. The nominal workload is a variable that does not depend on the discipline and follows directly from the parameters of the
network. It is defined by \( \rho := CM^{-1}(I - P^T)^{-1} \alpha \). The latter expression contains the effective arrival rate \( \lambda \) defined by \( \lambda = (I - P^T)^{-1} \alpha \). The nominal workload can be used to characterize the stability of HLPPS fluid networks, see Bramson (2008).

**Remark 1** (a) A HLPPS fluid network is stable if and only if \( \rho < e \), see Bramson (1996).

(b) The nominal workload condition \( \rho < e \) is necessary for the stability of any work-conserving fluid network, see Chen (1995).

(c) Sufficient stability conditions depend on the individual service discipline of the network. In particular, a fluid network may be stable under one discipline but not under another, see Bramson (2008).

To analyze the robustness of a stable fluid network \((\alpha, \mu, P, C)\) under a discipline \(\pi\), we consider a perturbed fluid network \((\alpha, \mu, P, C, \pi, \Delta)\), where \(\Delta \in \mathbb{R}^K_+\) denotes a feasible perturbation. At this point we do not yet specify the perturbation. In order to define a measure for the perturbation we define for \(\gamma \in \mathbb{R}^K_+\) the \(\gamma\)-weighted norm by \(\|x\|_\gamma := \sum_{k=1}^K |\gamma_k x_k|\). Here, \(\mathbb{R}^K_+\) denotes the positive orthant, i.e., \(\mathbb{R}^K_+ = \{x \in \mathbb{R}^K_+ : x_i > 0 \forall i = 1, ..., K\}\). Also, we denote \(B_\gamma(x, r) := \{y \in \mathbb{R}^K : \|x - y\|_\gamma < r\}\) with \(x \in \mathbb{R}^K\) and \(r > 0\). For the special case \(\gamma = e\) we simply write \(\| \cdot \|_e = \| \cdot \|\).

**Definition 2** Let \(\gamma \in \mathbb{R}^K_+\) be fixed. The \(\gamma\)-weighted stability radius of the fluid network \((\alpha, \mu, P, C, \pi)\) is defined by

\[
r_\gamma(\alpha, \mu, P, C, \pi) = \inf \left\{ \|\Delta\|_\gamma : \Delta \text{ is feasible and } (\alpha, \mu, P, C, \pi, \Delta) \text{ is unstable} \right\}.
\]

We use the \(\gamma\)-weighted stability radius to measure the robustness of the fluid network. Since \(\rho < e\) provides a necessary condition for stability of any work-conserving discipline, the nominal workload condition allows for a derivation of an upper bound of the stability radius that is valid for any discipline. We denote the upper bound by \(r^0_\gamma(\alpha, \mu, P, C)\). As the upper bound is not affected by the particular discipline, we omit the letter \(\pi\). In terms of the nominal workload condition \(\rho < e\), the upper bound is given by

\[
r^0_\gamma(\alpha, \mu, P, C) := \inf \left\{ \|\Delta\|_\gamma : \Delta \text{ is feasible, } \rho(\Delta) \neq e \right\},
\]

where \(\rho(\Delta)\) denotes the nominal workload of the fluid network \((\alpha, \mu, P, C, \Delta)\) that is subject to a perturbation \(\Delta\). The stability of fluid networks under HLPPS and FIFO of Kelly type disciplines can be equivalently characterized by the nominal workload condition \(\rho < e\), see Bramson (2008). In the remainder of this paper we consider fluid networks under HLPPS disciplines. Consequently, in this case \(r^0_\gamma(\alpha, \mu, P, C)\) is equal to the stability radius of the network. Moreover, for \(x, y \in \mathbb{R}^+_+\) the notation \(x \not\sim y\) means that \(x_i \geq y_i\) for at least one component \(i \in \{1, ..., K\}\). This reflects the fact that for some station \(j \in \{1, ..., J\}\) the nominal workload is at least one and, thus, the network is unstable.

Before we will present a scheme for the calculation of the stability radius, we have to collect some notations and preliminary results concerning convex analysis. A set \(A \subseteq \mathbb{R}^n\) is called _convex_ if \((1 - c)x + cy \in A\) whenever \(x, y \in A\) and \(c \in (0, 1)\). A point \(z\) of a convex set \(A\) is called an _extreme point_ if there is no way to express \(z\) as a convex combination \((1 - c)x + cy\) such that \(x, y \in A\) are distinct and \(c \in (0, 1)\). The set of all extreme points of \(A\) is denoted by \(\text{ext}(A)\). The _convex hull_ of a set \(A\), denoted by \(\text{conv}(A)\), is the smallest convex set that contains \(A\). The _boundary_ and the _interior_ of a set \(A\) is denoted by \(\partial A\) and \(\text{int}(A)\), respectively. Moreover, we will use the following auxiliary result. A proof can be found in Schönhlein et al. (2011).

**Proposition 1** Consider a closed and convex set \(B \subseteq \mathbb{R}^n\) and let \(A \subseteq B\) be convex and compact such that \(\partial A \cap \partial B \neq \emptyset\). Then, \(\text{ext}(A) \cap \partial B \neq \emptyset\).
3.1. Perturbations of Arrival Rates

In this section we investigate the situation where the service capacities $\mu$, the routing matrix $P$ and the constituency matrix $C$ are fixed, while the arrival rates $\alpha$ are subject to perturbations. Since the stability region is monotone with respect to the arrival rates we consider perturbations of the form $\Delta \in \mathbb{R}_+^k$. Precisely, we investigate the stability of the fluid network $(\alpha + \Delta, \mu, P, C)$ in terms of the perturbation $\Delta$. We will derive estimates on the size of $\Delta$ that characterize the stability region of the fluid network $(\alpha + \Delta, \mu, P, C)$. The nominal workload of the fluid network $(\alpha + \Delta, \mu, P, C)$ is given by

$$\rho(\Delta) = CM^{-1} (I - PT)^{-1} (\alpha + \Delta).$$

Then, the stability radius can be expressed by

$$r_0(\alpha, \mu, P, C) = \min\{ \|\Delta\|_\gamma : \Delta \in \mathbb{R}_+^k \text{ and } \rho(\Delta) \neq e \}. \quad (7)$$

To describe the geometric perspective of the stability radius consider a stable single station fluid network $(\alpha, \mu, P, C)$ that serves two classes of fluid. According to Remark 1 (a) the nominal workload is strictly less than one. In Figure 1 the light gray domain represents the stability region $D_0$. Let $B_\gamma(x, r) = \{ z \in \mathbb{R}^k : \|x - z\|_\gamma \leq r \}$. The stability radius can be illustrated as follows: It is the largest neighborhood $B_\gamma(\alpha, r)$ around $\alpha$, which is completely contained in the stability region $D_0$, where the size of the neighborhood is measured by the $\gamma$-weighted norm $\|\cdot\|_\gamma$. For arrival rates $\alpha'$ in the interior of $B_\gamma(\alpha, r)$, depicted by the dark gray domain in Figure 1, the fluid network $(\alpha', \mu, P, C)$ is stable, while for arrival rates on the boundary $\partial B_\gamma(\alpha, r)$ the fluid network $(\alpha', \mu, P, C)$ might be unstable.

![Figure 1: Illustration of the stability radius for a single station fluid network serving two job classes.](image)

In the following, we will derive a scheme for the calculation of the stability radius. Based on the geometric interpretation, our approach is to calculate (7) by means of an optimization problem. For this reason, the constraint on the perturbation $\Delta \in \mathbb{R}_+^k$ is that for at least one station $j \in \{1, \ldots, J\}$ of the fluid network the nominal workload $\rho_j(\Delta)$ is at least one. In terms of the fluid network $(\alpha + \Delta, \mu, P, C)$ this can be expressed as $\max_{j=1, \ldots, J} \rho_j(\Delta) \geq 1$. Consequently, the stability radius $r_0(\alpha, \mu, P, C)$ is the optimal value of the optimization problem

$$\begin{align*}
\text{minimize} \quad & \gamma^T \Delta \\
\text{subject to} \quad & \max_{j=1, \ldots, J} \rho_j(\Delta) = 1, \\
& 0 \leq \Delta.
\end{align*} \quad (8)$$

The following statement characterizes the solutions of the optimization problem.
**Theorem 1** Consider the perturbed fluid network \((α + Δ, µ, P, C)\) and the corresponding stability radius \(r_γ^0(α, µ, P, C)\). If \(r_γ := r_γ^0(α, µ, P, C) = \|Δ⁺\|_γ > 0\) is the optimal value of (8), then a minimizing \(Δ⁺\) can be chosen such that at most one of its components is strictly positive. That is, \(Δ⁺ = r_γ γ⁻¹ k_e\) for some \(k \in \{1, ..., K\}\).

**Proof.** Let \(N = CM^\dagger(I - PT)^\dagger\) and \(N_j\) be the \(j\)th row of \(N\). Then, the workload condition (6) reads as \(NΔ \not< c - Nα =: c\). The optimization problem can equivalently be stated as

\[
\text{minimize } \gamma^TΔ \\
\text{subject to } 0 \leq Δ, \\
N_jΔ \leq c_j, \quad j = 1, ..., J, \\
∃ j \in \{1, ..., J\} : N_jΔ = c_j.
\]

Let \(r_γ\) be the optimal value of (9). We note that, if \(Δ⁺ = r_γ γ⁻¹ k_e\) is feasible it is optimal.

To show the claim, suppose to the contrary that all \(\{r_γ γ⁻¹ k_e : k = 1, ..., K\}\) are not feasible. Then, at least one of the constraints in (9) is not satisfied. That is, for all \(k \in \{1, ..., K\}\), there is a \(j \in \{1, ..., J\}\) such that \(r_γ γ⁻¹ N_j k_e > c_j\), or

(1) there is a \(j \in \{1, ..., J\}\) such that \(r_γ γ⁻¹ N_j k_e < c_j\).

To consider the first case, let \(\overline{K} \in \{1, ..., K\}\) be such that for some \(j \in \{1, ..., J\}\) it holds that \(c_j < r_γ γ⁻¹ N_{\overline{K}} γ\). In particular, \(N_{\overline{K}} γ > 0\). Choose \(\overline{Δ} := c_j N_{\overline{K}} γ k_e γ\). Then, we have \(N_{\overline{K}} γ k_e γ\) and \(γ^T k_e γ N_{\overline{K}} γ < r_γ\). Hence, \(\overline{Δ}\) satisfies the equality constraints in (9) for \(j\). Further, if \(\overline{Δ}\) is not feasible then \(β \overline{Δ}\) is feasible for some \(β \in (0, 1)\). Then, it holds that \(β γ^T k_e γ N_{\overline{K}} γ < r_γ\), which yields a contradiction.

Thus, we have shown that the first case is not possible. It remains to consider the case that for all \(k \in \{1, ..., K\}\) and for all \(j \in \{1, ..., J\}\) it holds that

\(N_j k_e < c_j γ⁻¹\gamma k_e\). (10)

In the optimum \(Δ = (Δ_1 ... Δ_K)^T\) there is a \(j \in \{1, ..., J\}\) so that \(N_j Δ = c_j\). Then, by (10) it holds that

\[c_j = \sum_{k=1}^K N_j k_e Δ_k < \sum_{k=1}^K c_j γ k_e γ⁻¹ Δ_k = c_j γ⁻¹ \sum_{k=1}^K γ k_e Δ_k,
\]

or equivalently \(r_γ < γ^T Δ\), which yields a contradiction and the assertion follows. ■

Using \(N(μ) = CM(μ)^{-1}(I - PT)^{-1}\) and the fact that a destabilizing perturbation is of the form \(Δ⁺ = r_γ γ⁻¹ k_e\) for some \(k \in \{1, ..., K\}\), for some \(j \in \{1, ...., J\}\) it holds that

\(N(μ)_{jk} r_γ γ⁻¹ k_e = 1 - (N(μ)α)_{jk}\).

Thus, a representation of the stability radius that expresses the dependency on \(μ\) is given by

\[r_γ(μ) = \min_{j=1, ..., J} \left\{ \frac{1 - (N(μ)α)_{jk}}{\max_{k=1, ..., K} γ k_e γ⁻¹ N(μ)_{jk}} \right\}, \quad (11)
\]

The significance of (11) is that the stability radius of a network subject to perturbations of the arrival rates can be calculated based on \(α, μ, P\) and \(C\).
3.2. Perturbations of Service Capacities

In the following we assume that the arrival rates $\alpha$, the routing matrix $P$ and the constituency matrix $C$ are fixed, while the service capacities $\mu$ are subject to perturbations of the form $0 \leq \delta < \mu$. The inequalities have to be understood componentwise. For the perturbed fluid network $(\alpha, \mu - \delta, P, C)$ the matrix of service capacities and the nominal workload are denoted by $M(\delta) = \text{diag}(\mu - \delta)$ and $\rho(\delta) = CM(\delta)^{-1}(I - PT)^{-1}\alpha$, respectively. Note that perturbations $\delta < 0$ would lead to increased service capacities, which is not of interest for real world systems. Also, disturbances $\delta \geq \mu$ yield negative service capacities that are neglected. Then, the stability radius can be expressed as

$$ r^0_{\gamma}(\alpha, \mu, P, C) = \min \{ \|\delta\|_\gamma : 0 \leq \delta < \mu, \max_{j=1,\ldots,J} \rho_j(\delta) = 1 \}. $$

(12)

As a geometric illustration for this case, we consider once again a single station fluid network serving two classes of fluids. Given the arrival rates $\alpha$ and routing matrix $P$, the effective arrival rates $\lambda$ are determined. The light gray domain in Figure 2 represents the stability region $D_0$ for fixed arrival rates. By Remark 1 (a) service capacities in the interior of the stability region provide a stable network. For service capacities on the boundary $\partial D_0$ of the stability region the corresponding fluid network might be unstable. Using $B_r(\mu, r) = \{ \bar{\mu} \in \mathbb{R}^2_+ : \|\bar{\mu} - \mu\|_\gamma \leq r \}$ the stability radius can be described as the radius of the largest neighborhood $B_r(\mu, r)$ around $\mu$ that is completely contained in the stability region $D_0$ such that at least one edge of $\partial B_r(\mu, r)$ intersects the boundary $\partial D_0$ of the stability region. In Figure 2 the neighborhood $B_r(\mu, r)$ is illustrated by the dark gray domain. The stability radius $r^0_{\gamma}(\alpha, \mu, P, C)$ can be calculated by the following optimization problem

$$ \begin{align*}
\text{minimize} & \quad \gamma^T \delta \\
\text{subject to} & \quad \max_{j=1,\ldots,J} \rho_j(\delta) = 1 \\
& \quad 0 \leq \delta < \mu.
\end{align*} $$

(13)

To calculate the stability radius we split the problem into $J$ subproblems. We consider the optimization problem (13) for each station $j \in \{1,\ldots,J\}$ individually. The corresponding solution is simply denoted by $r_j$. The smallest solution in magnitude represents the solution to (13). In the sequel, we describe how to solve the optimization problem for a single station $j \in \{1,\ldots,J\}$. That is, for each $j$ the task is of the form

$$ \begin{align*}
\text{minimize} & \quad \gamma^T \delta \\
\text{subject to} & \quad \rho_j(\delta) = 1 \\
& \quad 0 \leq \delta_k < \mu_k, \quad k \in C(j).
\end{align*} $$

(14)
The minimal value is denoted by \( r_j \). Based on the subproblems (14) the stability radius of the fluid network is given by the minimal solution of all subproblems, i.e. \( r^0_j(\alpha, \mu, P, C) = \min_{j=1, \ldots, J} r_j \).

For some \( c > 0 \) the general form of the optimization problem (14) for a single station is

\[
\begin{align*}
\text{minimize} & \quad \gamma_1 \delta_1 + \ldots + \gamma_n \delta_n \\
\text{subject to} & \quad \frac{\lambda_1}{x_1 - \delta_1} + \ldots + \frac{\lambda_n}{x_n - \delta_n} = c \\
& \quad 0 \leq \delta_k < x_k, \quad k = 1, 2, \ldots, n.
\end{align*}
\] (15)

The minimal value is denoted by \( r_\gamma \). The subsequent statement characterizes the structure of the solutions of the optimization problem (15).

**Theorem 2** Consider the perturbed fluid network \((\alpha, \mu - \delta, P, C)\) and the corresponding stability radius \( r_\gamma := r^0_\gamma(\alpha, \mu, P, C) \). If \( r_\gamma = \|\delta\|_\gamma > 0 \) is a the value of (15), then a minimizing \( \delta \) can be chosen such that at most one of its components is strictly positive. That is, \( \delta_* = r_\gamma \gamma_k^{-1} e_k \) for some \( k \in \{1, 2, \ldots, n\} \).

**Proof.** Assume that \( \delta = (\delta_1, \delta_2, \ldots, \delta_n)^T \) is a solution to (15), i.e. \( r_\gamma = \|\delta\|_\gamma \) and \( \frac{\lambda_1}{x_1 - \delta_1} + \frac{\lambda_2}{x_2 - \delta_2} + \ldots + \frac{\lambda_n}{x_n - \delta_n} = c \). The statement is shown by induction. For \( n = 2 \) consider the optimization problem

\[
\begin{align*}
\text{minimize} & \quad \gamma_1 \delta_1 + \gamma_2 \delta_2 \\
\text{subject to} & \quad \frac{\lambda_1}{x_1 - \delta_1} + \frac{\lambda_2}{x_2 - \delta_2} = c \\
& \quad 0 \leq \delta_k < x_k, \quad k = 1, 2.
\end{align*}
\] (16)

For given \( \lambda_1, \lambda_2 \) and \( c > 0 \) the set \( \mathcal{M} = \{y = (y_1, y_2) \in \mathbb{R}_{>0}^2 : \frac{\lambda_1}{y_1} + \frac{\lambda_2}{y_2} \leq c\} \) is closed, convex and the boundary \( \partial \mathcal{M} \) equals the set of extreme points \( \text{ext}(\mathcal{M}) \). The minimal distance \( r_{\gamma} \) from \( x = (x_1, x_2) \in \mathcal{M} \) to the boundary \( \partial \mathcal{M} \) can be described by \( r_\gamma = \max \{r : B_r(x, r) \subset \mathcal{M} \} \). Further, for every \( \varepsilon > 0 \) it holds that \( B_{r_{\gamma}}(x, r_{\gamma} + \varepsilon) \notin \mathcal{M} \) and this implies that \( \partial \mathcal{M} \cap \partial B_{r_{\gamma}}(x, r_{\gamma} + \varepsilon) \neq \emptyset \). By Proposition 1 we have that \( \partial \mathcal{M} \cap \text{ext}(B_{r_{\gamma}}(x, r_{\gamma} + \varepsilon)) \neq \emptyset \). Hence, \( \delta \) can be chosen as \( \delta = r_{\gamma} \gamma_k^{-1} e_k \) for some \( k \in \{1, 2\} \).

In the inductive step we assume that the claim is valid for \( n \) and consider the case \( n + 1 \). So, consider the condition

\[
\frac{\lambda_1}{x_1 - \delta_1} + \frac{\lambda_2}{x_2 - \delta_2} + \ldots + \frac{\lambda_n}{x_n - \delta_n} = c - \frac{\lambda_{n+1}}{x_{n+1} - \delta_{n+1}}.
\]

By hypothesis it holds that \( r_{\gamma} = \gamma_i \delta_i \) for some \( i = 1, 2, \ldots, n \). Without loss of generality, let \( \gamma_1 \delta_1 = r_{\gamma} \). Hence,

\[
\frac{\lambda_1}{x_1 - \delta_1} + \frac{\lambda_{n+1}}{x_{n+1} - \delta_{n+1}} = c - \frac{\lambda_2}{x_2} - \ldots - \frac{\lambda_n}{x_n}.
\]

Thus, by induction hypothesis, it follows that \( \gamma_1 \delta_1 = r_{\gamma} \) or \( \gamma_{n+1} \delta_{n+1} = r_{\gamma} \), which shows the assertion. \( \blacksquare \)

**3.3. Perturbations of Arrival Rates and Service Capacities**

In the following we consider fluid networks that are subject to perturbations of the arrival rates as well as perturbations of the service capacities. That is, given a nominal fluid network \((\alpha, \mu, P, C)\), we consider fluid networks \((\alpha + \Delta, \mu - \delta, P, C)\). The disturbances \( \Delta \) and \( \delta \) are of the form \( \Delta \in \mathbb{R}_+^K \) and \( 0 \leq \delta < \mu \), respectively. The nominal workload of the perturbed fluid network is given by \( \rho(\Delta, \delta) = CM(\delta)^{-1}(I - PT)^{-1}(\alpha + \Delta) \). Following the approach of the previous sections, we consider the stability radius that corresponds to the stability region. We regard

\[
r^0_j(\alpha, \mu, P, C) = \min_{j=1, \ldots, J} \min\{\|\Delta\|_\gamma + \|\delta\|_\gamma : \rho_j(\Delta, \delta) = 1\}.
\]
Based on this, the stability radius can be calculated by an optimization problem of the following form. For brevity, let $A := (I - P^T)^{-1}$ and consider

$$r^0_\gamma (\alpha, \mu, P, C) = \min_{j \in \{1, \ldots, J\}} r_j,$$

where $r_j$ is the minimal value to

$$\begin{align*}
\text{minimize} & \quad \gamma^T (\Delta + \delta) \\
\text{subject to} & \quad \rho_j (\Delta, \delta) = 1, \\
& \quad 0 \leq \delta < \mu, \\
& \quad 0 \leq \Delta.
\end{align*}$$

(17)

For each station $j \in \{1, \ldots, J\}$ the general form of the latter setting is of the following form

$$\begin{align*}
\text{minimize} & \quad \sum_{k=1}^K \gamma_k (\Delta_k + \delta_k) \\
\text{subject to} & \quad \sum_{k \in C(j)} \frac{\lambda_k + \sum_{l=1}^K a_{kl} \Delta_l}{\mu_k - \delta_k} = c \\
& \quad 0 \leq \delta_k < \mu_k, \quad k \in C(j), \\
& \quad 0 \leq \Delta_k, \quad k \in \{1, \ldots, K\}.
\end{align*}$$

(18)

In the sequel, we will characterize the perturbations $\Delta_*$, $\delta_*$ that provide solutions to (18). This is the content of the following theorem.

**Theorem 3** Consider the perturbed fluid network $(\alpha + \Delta, \mu - \delta, P, C)$ and the corresponding stability radius $r_* := r^0_\gamma (\alpha, \mu, P, C)$. If $r_\gamma = \|\Delta_*\|_{\gamma} + \|\delta_*\|_{\gamma} > 0$ is the value of (18), then minimizing perturbations $\Delta_*$ and $\delta_*$ can be chosen such that either for $\Delta_*$ and $\delta_*$ at most one of its components is strictly positive. That is, either $\Delta_* = r_\gamma \gamma_k e_k$ and $\delta_* = 0$ or $\Delta_* = 0$ and $\delta_* = r_\gamma \gamma_k^{-1} e_k$ for some $k \in \{1, \ldots, K\}$.

**Proof.** Suppose that the perturbations $(\Delta_*, \delta_*)$ are a solution to (18). Then, by Theorem 1 and 2, it follows that $\Delta_* = \Delta_m e_m$ for some $m \in \{1, \ldots, K\}$ and $\delta_* = \delta_n e_n$ for some $n \in \{1, \ldots, K\}$. If $n \neq m$, the optimization problem is given by

$$\begin{align*}
\text{minimize} & \quad \gamma_m \Delta_m + \gamma_n \delta_n \\
\text{subject to} & \quad \sum_{k \in C(j), \ k \neq n} \frac{\lambda_k}{\mu_k - \delta_k} + \frac{\lambda_n}{\mu_n - \delta_n} + \sum_{k \in C(j), \ k \neq n} \frac{a_{km} \Delta_m}{\mu_k} + \frac{a_{nm} \Delta_m}{\mu_n - \delta_n} = c, \\
& \quad 0 \leq \delta_n < \mu_n, \\
& \quad 0 \leq \Delta_m.
\end{align*}$$

(19)

In the following we consider the first constraint in (19), which can be written as

$$\frac{\lambda_n}{\mu_n - \delta_n} + \Delta_m \sum_{k \in C(j), \ k \neq n} \frac{a_{km}}{\mu_k} + \frac{a_{nm} \Delta_m}{\mu_n - \delta_n} = c - \sum_{k \in C(j), \ k \neq n} \frac{\lambda_k}{\mu_k}.$$

(20)

Since $\mu_n - \delta_n > 0$ and using for brevity $q := \sum_{k \in C(j), \ k \neq n} \frac{a_{nm} \Delta_m}{\mu_k}$ and $\bar{c} := c - \sum_{k \in C(j), \ k \neq n} \frac{\lambda_k}{\mu_k}$, equation (20) reads as

$$\bar{c} \delta_n + (a_{nm} + q\mu_n) \Delta_m - q \Delta_m \delta_n = c\mu_n - \lambda_n.$$

(21)

Moreover, using $\bar{d} := a_{nm} + q\mu_n$ and $p := \bar{c}\mu_n - \lambda_n$ condition (21) can be written in compact form as

$$\bar{c} \delta_n + \bar{d} \Delta_m - q \delta_n \Delta_m - p = 0.$$

(22)
By the definition of $\bar{d}, q$ it follows that $\frac{\bar{d}}{q} > \mu_n > \delta_n$. So, (22) can be written as

$$\Delta_m = h(\delta_n) := \frac{p - \bar{c} \delta_n}{d - q \delta_n} = \frac{\bar{c}}{q} \cdot \frac{p}{\bar{d}} - \delta_n, \quad \delta_n \in [0, \mu_n).$$

The following instances may occur.

Suppose that $pq > \bar{c} \bar{d}$. Then, it holds that $\frac{\partial}{\partial \delta_n} \Delta_m(\delta) = \frac{pq - \bar{c} \bar{d}}{(d - q \delta_n)^2} > 0$. Rewriting the problem (19) as

$$\begin{align*}
\text{minimize} & \quad m(\delta_n) := \gamma_m h(\delta_n) + \gamma_n \delta_n \\
\text{subject to} & \quad 0 \leq \delta_n < \mu_n
\end{align*}$$

and using that $\frac{\partial m(\delta)}{\partial \delta_n} = \gamma_m \frac{d h(\delta)}{d \delta_n} + \gamma_n > 0$, the minimum of $m$ is attained for $\delta_n = 0$. Thus, the minimum of (19) is attained for $(\Delta_m, \delta_n) = (\frac{\bar{c}}{q}, 0)$.

If $\frac{\bar{c}}{q} = \frac{p}{\bar{d}}$, it holds that

$$\Delta_m = h(\delta_n) = \frac{\bar{c}}{q} \cdot \frac{p}{\bar{d}} - \delta_n \equiv \frac{\bar{c}}{q} = \frac{p}{\bar{d}} \cdot \delta_n, \quad \delta_n \in [0, \mu_n).$$

So, the solution to

$$\begin{align*}
\text{minimize} & \quad m(\delta_n) := \gamma_m \frac{\bar{c}}{q} + \gamma_n \delta_n \\
\text{subject to} & \quad 0 \leq \delta_n < \mu_n
\end{align*}$$

is $(\Delta_m, \delta_n) = (\frac{\bar{c}}{q}, 0)$.

Finally, suppose that $pq < \bar{c} \bar{d}$ or equivalently $\frac{\bar{c}}{q} < \frac{\bar{d}}{\bar{q}}$. Then, the constraint set is given by

$$\{(\Delta_m, \delta_n) : \Delta_m = h(\delta_n), \delta_n \in [0, \frac{\bar{d}}{\bar{q}}]\}.$$

Further, $\gamma_m \Delta_m + \gamma_n \delta_n = r_\gamma$ defines a straight line, where $\Delta_m, \delta_n$ also satisfy $\Delta_m = h(\delta_n)$. So, $\delta_n \in [0, \frac{\bar{d}}{\bar{q}}]$ and $\Delta_m \in [0, \frac{\bar{c}}{q}]$. Hence, we have that either $\Delta_m = \frac{\bar{c}}{q}$ and $\delta_n = 0$ or $\Delta_m = 0$ and $\delta_n = \frac{\bar{d}}{\bar{q}}$.

The case $n = m$ follows the same line of reasoning. This shows the assertion. \hfill \blacksquare

In the following remark we discuss the implications of the results in this section.

**Remark 2** Consider a HLPPS fluid network with $(\alpha, \mu, P, C)$.

(a) Based on Theorems 1, 2 and 3 the stability radius can be calculated as follows:

1. Consider each station $j \in \{1, \ldots, J\}$ separately.
2. Let $\chi$ be a feasible perturbation of either the arrival rate or the service capacity of one fluid class. Then, for each $k \in \{1, \ldots, K\}$ solve the optimization problem:

$$\begin{align*}
\text{minimize} & \quad \gamma_k \chi_k \\
\text{subject to} & \quad \rho_j(\chi_k) = 1.
\end{align*}$$

Let $r_{j, k}^\gamma$ denote the corresponding optimal value.

3. The stability radius is given by the minimal value of all optimal values, i.e.

$$r_\gamma(\alpha, \mu, P, C) = \min \left\{ r_{j, k}^\gamma : j \in \{1, \ldots, J\}, k \in \{1, \ldots, K\}, \chi \in \{\Delta, \delta\}, \right\}.$$

(b) A HLPPS fluid network $(\alpha + \Delta, \mu - \delta, P, C)$ keeps the property of stability for all perturbations $(\Delta, \delta)$ satisfying

$$\|\Delta\|_\gamma + \|\delta\|_\gamma < r_\gamma^0(\alpha, \mu, P, C).$$

(c) Let $(\alpha, \mu, P, C, \pi)$ be a fluid network under the work-conserving discipline $\pi$. Then, its stability radius $r_\gamma(\alpha, \mu, P, C, \pi)$ is bounded from above by $r_\gamma^0(\alpha, \mu, P, C)$.

In summary, the findings of Section 3 allow for characterizing the stability radius of a given fluid network that is subject to perturbations of the arrival and/or service capacities.
4. Optimal parameter adjustment for robust stability

In this section we assume that the nominal model and the real world system are congruent/in line and relax the assumption that all network parameters are given in advance. The application of the results for the stability radius of a fluid network can be used in order to adjust the parameters of the real world system towards disturbed arrival rates as well as service capacities, while ensuring certain robustness properties of the network. Since there exist various combinations of objectives and constraints, we briefly name three typical problems.

(i) Let the arrival rates $\alpha$ and the network topology $P, C$ be given and suppose that for each station $j$ there is a $z_j$ such that $\sum_{k \in C(j)} \mu_k \leq z_j$. We call $z_j$ the maximal service capacity of station $j$. Given $z = (z_1 \ldots z_J)^T$, we can determine the optimal allocation of the services capacities $\mu_k$ to the different job classes, while maximizing the robustness $r = r^0(\alpha, \mu_*, P, C)$ of the whole network.

(ii) Given the arrival rates $\alpha$, the network topology $P, C$, and a desired stability radius $r$ of the network, we can specify the minimal required service capacities $\mu$ that comply with $r$.

(iii) Let the service capacities $\mu$, the network topology $P, C$, and a desired stability radius $r$ of the network be given, then we can specify the maximal arrival rates $\alpha$ that comply with $r$.

In the sequel, we focus on the case (ii) and formulate a mathematical program that allocates the service capacities. The program aims for a sustainable capacity allocation that enables the network to handle increased arrival rates with a minimum of additional service capacities. In line with Remark 2 the program contains $S \subseteq \{1, \ldots, K\}$ instances of the network, where in each instance exactly one arrival rate $\alpha_k$ is disturbed. The resulting mathematical formulation is a mixed integer nonlinear optimization program.

4.1. Nomenclature

For convenience we recall the required notation for the mathematical program.

Sets

$K$ Job classes ($k, l \in K$)
$J$ Stations ($j \in J$)
$S$ Instances, where the arrival rate is disturbed ($s \in S \subseteq K$)
$C(j, k)$ Assignment of job class $k$ to station $j$

Parameters

$\alpha_k$ Arrival rate of job class $k$
$C_i,k$ Constituency matrix; job class $k$ is processed at station $j$
$I_{l,k}$ Identity matrix of job classes
$L$ Large scalar (big $M$)

Variables

$\mu_k$ Service capacity for job class $k$ at its service station
$\rho_{j,s}$ Nominal workload of $j$ in $s$, where $\alpha_k$ of $k = s$ is disturbed
$CM_{j,k}$ Auxiliary matrix $CM^{-1}$
$B_{j,k}$ Aux. matrix $CM^{-1}(I - P^T)^{-1}$

Binary variable

$X_{j,k}$ Binary variable denoting that station $j$ has a nominal workload $\rho_{j,k} = 1$ if the service capacity $\mu_k$ of job class $k$ is disturbed by $r$

4.2. Mathematical model

In (25) the required processing time per job of class $k$ is calculated. These times depend on the allocated service capacity $\mu_k$, via

$$CM_{j,k} = C_{j,k}\mu_k^{-1} \quad (j \in J; k \in K). \quad (25)$$
Constraint (26) relates these times with information about the routing, i.e.,

$$B_{j,k} = \sum_{l} CM_{j,l} A_{l,k}$$  \hspace{1cm} (j \in J; k \in K). \hspace{1cm} (26)$$

For each station the nominal workload $\rho$ is given by the sum of the products of effective arrival rate and service time over all job classes present at $j$, i.e.,

$$\rho_{j,s} = \sum_{k} B_{j,k} (\alpha_k + r \cdot 1_{s,k})$$  \hspace{1cm} (j \in J; s \in S). \hspace{1cm} (27)$$

Here, $|S|$ instances are created, where each time exactly one arrival rate $\alpha_k$ is increased by $r$. To this end, we use the notation of the Kronecker symbol, i.e., $1_{s,k} = 1$ if $s = k$ and $1_{s,k} = 0$ else. Each instance comprises all stations. In line with Remark 2 the service capacity allocation $\mu_k$ in (25) should be chosen such that the nominal workload takes for a given $r$ at least in one instance for one station a value of one. In this case, the network is able to handle any increase $\Delta$ of the arrival rates such that $\|\Delta\| < r$. This is enforced by

$$\rho_{j,s} \leq 1$$  \hspace{1cm} (j \in J; s \in S), \hspace{1cm} (28)$$

$$\rho_{j,s} \geq 1 - (1 - X_{j,s})L$$  \hspace{1cm} (j \in J; s \in S) \hspace{1cm} (29)$$

and

$$\sum_{j} \sum_{s} X_{j,s} = 1$$  \hspace{1cm} (30)$$

The objective function (31) minimizes the required service capacities for the different job classes at the stations, by

$$\text{Min} \sum_{k} \mu_k.$$  \hspace{1cm} (31)$$

5. Application in dynamic supply chains

Supply chains are frequently subject to a dynamic and uncertain parameter evolution. In this section we apply our approach to robust stability to a real world supply chain and demonstrate its properties.

5.1. Real world test case

The considered test case scenario is based on the structure and data of a European pump set manufacturer supply chain. This supply chain produces highly specialized industrial pump sets for different industries. Suppliers are responsible for the production of pump set components like motor, coupling, coupling guard, separator and base plate. The pump set manufacturer focuses on the production of three different pump types (type 1, type 2 and type 3), which have different capabilities and characteristics, and the assembly of pump sets. Different production facilities process these orders in line with customer specifications. In particular, each production facility is specialized in the production of one pump type. However, the pump set assembly can be carried out by every distribution or service-center.

For convenience we limit the section of the considered real world supply chain to three main production and assembly facilities. Each of these facilities produces one pump type and assembles all three types of pump sets for local customers. A sketch of the studied network is given in Figure 3. For instance, Location 1 (Germany) produces pumps of type 1 and assembles these as type 1 pump sets for the local market. Furthermore, Location 1 performs the final assembly of pump sets that are based on pumps of type 2 and 3. The local label for these pump sets is 5 and 9. Customers in France and Spain also demand for pump sets of type 1. Hence these are shipped to
Figure 3: Real world test case.

Location 2 and 3 for their final adaptation to local requirements. Since the facilities have limited resources, they need to allocate their production capacity to the various operations related to the different pumps and pump sets.

The real world material flow within the considered network and customer demand was analyzed for a given observation period. In detail, the assignment of manufactured pump set type to a production facility is given by the constituency matrix $C$. The material flow provided information about the internal routing of orders and the need for reworking, which is captured by the routing matrix $P$. Furthermore, customer orders were analyzed for each market and pump set type. As a result distributions characterizing the observed customer demand were derived. The parameters used in the following are given by

$$P = \begin{bmatrix}
0.05 & 0 & 0 & 0.45 & 0 & 0 & 0.06 & 0 & 0 \\
0 & 0.05 & 0 & 0 & 0.51 & 0 & 0 & 0.10 & 0 \\
0 & 0 & 0.1 & 0 & 0 & 0.16 & 0 & 0 & 0.47 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix},$$

$$\alpha = \begin{bmatrix}
2.88 & 5.27 & 3.61 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}^T,$$

$$C = \begin{bmatrix}
1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0
\end{bmatrix}.$$
5.2. Computational results

The dynamics of a supply chain can be modeled by a multiclass queuing network. In this case we consider the arrival rate as customer demand and the service rate as production capacity that is required for processing customer orders. Our approach to robust stability enables the adjustment of the production capacity towards different objectives (see Section 4), subject to shifts of the expected value of customer demand, available production capacity or combined shifts. In the following we apply the mathematical model presented in Section 4 and show the properties of our approach.

To this end the dynamics of the supply chain are modeled in a discrete stochastic model as well as by a deterministic fluid model, both implemented with Vensim 5.8 simulation software. The mathematical optimization model was implemented in GAMS 23.6 and solved by BARON 9.0.6. In addition, constraint (27) was modified in a way that the external arrival rate \( \alpha_k \) is only disturbed, if there is a positive real world customer demand for class \( k \). Hence, only three instances are created for an increased order arrival of pump type 1, 2 and 3. We assume that the pump set manufacturer wants to prepare the network for an increased total demand of 10%, which corresponds to a desired stability radius of \( r = 1.18 \). The application of the mathematical program yields the following production capacities \( \mu \):

\[
\mu = \begin{bmatrix} 10.59 & 11.91 & 8.03 & 5.36 & 8.61 & 3.66 & 1.48 & 2.59 & 7.03 \end{bmatrix}^T.
\]

In the following we investigate the evolution of the stock level of the 9 pre-products of the supply chain for three different demand scenarios. To this end we assume an initial queue length as follows:

\[
Q(0) = \begin{bmatrix} 2 & 3 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T.
\]

First, we study a situation in which the original customer demand is applied. The evolution of the multiclass queueing and fluid model of the supply chain is shown in Figure 4. Here, the first line presents the discrete evolution of the queueing model, where the stock level at the different locations accumulates sometimes up to 5 pump sets waiting for production. In general the process appears to be stable, since the queue length drops back to zero from time to time. The evolution of the continuous fluid model looks completely different, since all stock levels become zero after some time. This evolution implies stability of the overall model.

![Figure 4: Stock level of the undisturbed queueing and fluid network (Location 1 to Location 3).](image)

Secondly, we apply an increased customer demand by \( r = 1.18 \) for \( \alpha_1 \). Such a demand puts the network at the edge of instability. The evolution of the fluid model is shown in Figure 5. As a characteristic of this case one can see that the stock level remains at the same level after a short period of initialization.

![Figure 5: Stock level of the disturbed queueing and fluid network (Location 1 to Location 3).](image)
Thirdly, we apply an increased demand of $1.1r = 1.3$ for $\alpha_1$. In this case the network is unstable in the sense that it is no longer able to process all arriving orders. Hence orders accumulate in the buffers over time. This is shown in Figure 6 for the fluid model. Note that in this case the stock level of the queueing network grows over any bound (the stock becomes infinitely large).

In summary, these three scenarios show that it is possible to adjust the parameters of a stochastic supply chain towards a desired level of robustness, which corresponds to shifts of the expected value of customer demand and available production capacity. This opens a new perspective for the strategic supply chain management. Since the underlying planning problem is not NP-hard, large problem instances with several facilities and products can be solved.

6. Conclusions and outlook

In this paper we considered the question how the stability of multiclass queueing networks is affected by changes of the expectations of the distributions of the interarrival and service times. As the stability of the associated fluid network provides a sufficient stability criterion for stability of the multiclass queueing network, we used the associated fluid network to define the weighted stability radius in order to measure the robustness of stability of the fluid network with respect to uncertainties of the arrival rates and/or the service capacities. The weighted stability radius is a real number and represents the magnitude of the smallest shift of an expected value so that the fluid network is no longer stable. Furthermore, using the fact that the stability of a fluid network under head-of-the-line proportional processor sharing (HLPPS) discipline is equivalent to a nominal workload strictly less than one, we discussed a scheme to calculate the weighted stability radius for HLPPS fluid networks. This scheme is applicable for three different kinds of uncertainties. Based on this characterization we discussed possible applications to optimal parameter adjustments. More precisely, we presented briefly reasonable scenarios and studied one of them in detail. To this end we formulated a mathematical program so that, for predefined arrival rates, structure, routing and a desired level of robustness in terms of the stability radius, the service capacities are allocated so that the total service capacity is minimized. Moreover, this program was applied to a real world supply chain of a European pump set manufacturer. Here it was shown that shifts of the customer demand below the stability radius do not destroy stability of the supply chain in contrast to shifts that are larger than the stability radius. Furthermore, precise knowledge about the robustness of a supply chain can be used as an indicator for identifying the need of future production capacity adjustments.
In future research other service disciplines have to be investigated, since the calculation scheme for the stability radius depends on the discipline under consideration. However, we note that the calculation scheme provided here yields an upper bound for the stability radius for any work-conserving discipline. From the application perspective, this framework might be embedded in larger design problems.

Acknowledgment

This research was funded by the Volkswagen Foundation.

References


