Robust Optimization approach to characterize and design Globally Stable Queueing Networks

Stability and throughput maximization are fundamental properties of queueing networks and have been a subject of continuous research, with fluid-based analysis providing a powerful tool since the work of [7] and [3]. An interesting (and somewhat unexpected) insight of the research in this field is that having the load on each resource $j$, $\rho_j$, be strictly less than one – often referred to as the natural load condition – is not generally sufficient for the stability of a multiclass network. The Rybko-Stolyar and Lu-Kumar (see Figure ??) networks are simple instances in which priority policies destabilize the network eventhough the natural load conditions are met. The literature does propose and identify controls that stabilize a multi-class networks under the natural load conditions. Head of the Line Process Sharing (HLPPS) and the family of maximum pressure policies and versions thereof provide instances of such rules.

Maximum pressure is a centralized control that requires the coordination of various resources: the priorities resources assign to activities must take into account the state of queues that they do not process. Head of the line Proportional Processor Sharing (HLPPS) is a local discipline that stabilizes the network under the natural load conditions but, as Maximum Pressure, leave no freedom to the different servers to determine their local priorities. In certain processing networks such freedom is important – physicians in emergency rooms, for example, make decisions following clinical considerations. From both the intellectual and the practical perspectives it is thus important to study networks and conditions on parameters (e.g. mean service times) that render a network stable under any work conserving policy, i.e., globally stable. This agenda was pursued towards the end of 90s. [5], [6], [4] and [2] identify several networks that are stable under any work conserving policy as long as the natural load conditions are met. In this paper, we seek to study networks and condi-
tions on parameters (e.g. mean service times) that render a network stable under any work conserving policy, i.e., globally stable.

**Model:** We consider a queueing network with a set $\mathcal{J} = \{1, \ldots, J\}$ of single-server stations and, a set $\mathcal{K} = \{1, \ldots, K\}$ of customer classes (with $K \geq J$). We will interchangeably use the terms class and activity to mean the same thing. We will similarly use the terms station and server. The many-to-one mapping from classes to stations is described by a $J \times K$ constituency matrix $C$ where $C_{jk} = 1$ if class $k$ is served at station $j$, and it equals 0 otherwise. We assume naturally that the set $\mathcal{C}(j) := \{k \in \mathcal{K} : C_{jk} = 1\}$ is non-empty. For $k \in \mathcal{K}$, let $s(k)$ be the station at which class $k$ is served, i.e., $s(k)$ is the unique $j \in \mathcal{J}$ such that $C_{jk} = 1$. For a set $\mathcal{B} \subseteq \mathcal{J}$, $\mathcal{C}(\mathcal{B}) := \{k \in \mathcal{K} : C_{jk} = 1\}$ for some $j \in \mathcal{B}$ is the set of customer classes served by station in $\mathcal{B}$. We denote by $M$ the $K \times K$ diagonal matrix with the mean service time of class $k$, $m_k$, as the $k^{th}$ diagonal element. The parameter $\mu_k = 1/m_k$ then stands for the long-run average rate at which class-$k$ customers would be served if the server in station $s(k)$ were never idle and worked exclusively on class $k$.

**Contributions:**

In this paper, we re-visit the question of global stability in open multi-class queueing networks. Characterizing these conditions is a difficult task and results are known for specific networks and fully-understood only for 2-station networks. We propose here a new approach into this problem. We make two relaxations: a) rather than aiming at the full characterization of global stability for given networks, we seek to identify assembly (LEGO) operations, such as pooling of networks, and assembly rules-of-thumb that, when followed, create a globally stable network from globally stable building blocks; b) Use a robust optimization approach to model the queueing primitives.

To this end, we assemble an analysis infrastructure based on queue-ratio policies, linear attraction to state space collapse and the stability of the network Skorohod problem (SP) that captures the dynamics of the “collapsed” network. The infrastructure produces sufficient conditions and lends itself to LEGO in a relatively natural way. Key in the analysis is a semi-definite programming formulation of the key problem that naturally gives rise to modularity of the stability analysis problem. The key step in this approach is to consider
the problem at the Skorohod (“collapsed”) problem level, and to relate global stability to robust optimization where the uncertainty set is the family of values that the queue ratio matrix $\Delta$ can obtain. This is robust optimization over policy parameters. Building on this relationship we prove that it suffices to consider static priority policies – these are mathematically (and not just conceptually) the extreme points of queue-ratio policies. We, similarly establish the sufficiency of state space collapse for priority policies towards state space collapse. To summarize, the problem of checking whether a given open multi-class multi-server queueing system is globally stable is reduced to a convex optimization problem, in particular, a semi-definite optimization problem. This leads to both structural results as well as allows tractable computation.

References


