QUEUEING SYSTEM TOPOLOGIES WITH LIMITED FLEXIBILITY

John N. Tsitsiklis, jnt@mit.edu
Kuang Xu, kuangxu@mit.edu
Laboratory for Information and Decision Systems (LIDS), MIT, Cambridge, MA, USA

We study a multi-server model with n *flexible* servers and rn queues, connected through a fixed bipartite graph, where the level of flexibility is captured by the graph's average degree, $d(n)$. Applications in content replication in data centers, skill-based routing in call centers, and flexible supply chains are among our main motivations.

We focus on the scaling regime where the system size n tends to infinity, while the overall traffic intensity stays fixed. We show that a large capacity region (robustness) and diminishing queueing delay (performance) are jointly achievable even under very limited flexibility ($d(n) \ll n$). In particular, when $d(n) \gg \ln n$, a family of random-graph-based interconnection topologies is (with high probability) capable of stabilizing all admissible arrival rate vectors (under a bounded support assumption), while simultaneously ensuring a diminishing queueing delay, of order $\ln n/d(n)$, as $n \to \infty$. Our analysis is centered around a new class of virtual-queue-based scheduling policies that rely on dynamically constructed partial matchings on the connectivity graph.