Polling systems arise as a useful model in many applications, e.g., in computer-communication, production, transportation and maintenance systems. It is well known that among all cyclic service policies, the exhaustive service policy minimizes the total amount of work in the system. However, it does have the disadvantage that it does not put any restrictions on the cycle times. For this reason, the exhaustive service discipline is unsuitable for deadline-critical applications, where some high priority customers need to receive service as quickly as possible, for example, in communication networks in which some messages contain user commands.

A more suitable service policy for deadline-critical polling systems is a $k$-limited service policy. Such a policy makes sure that the number of customers that are served at each queue does not exceed $k$, thus to some extent bounding the cycle time. However, a major drawback of using a $k$-limited service policy is the fact that if the server reaches a very long queue, it will still serve at most $k$ customers, even though it possibly did not have to serve any customers at the last few queues.

In order to overcome this drawback, we propose a flexible $k$-limited service policy, which exploits the fact that the server sometimes visits a queue which has less than $k$ customers and essentially has time to spare. We analyze the performance of this flexible $k$-limited service policy for large-scale symmetric polling systems, comparing it with the exhaustive and traditional $k$-limited service policies. Quantities of interest are the asymptotic queue-length, cycle-time and waiting-time distributions as the number of queues goes to infinity.