We study the scheduling policies for order-optimal delay in single-hop queueing networks with context-switching overhead. In 60GHz wireless networks with directional antennas, base stations need to train and reconfigure their beam patterns whenever they switch from one client to another. This context-switching overhead can result in significant capacity loss, and needs to be explicitly addressed in the design of scheduling policies. Considerable context-switching overhead can also be observed in many other queueing networks such as transportation networks.

We first consider single-hop networks with only one server and multiple queues. While the celebrated Max-Weight policy achieves order-optimal average delay for systems without context-switching overhead, it fails to preserve throughput-optimality when context-switching overhead is taken into account. We propose a queue-length-based scheduling policy that explicitly takes context-switching overhead into account. We prove that our policy not only is throughput-optimal, but also achieves order-optimal delay when the traffic load in the system approaches the boundary of stability region. Next, we extend our to single-hop ad hoc networks with arbitrary interference constraints. We show that our policy remains throughput-optimal, and it can be made arbitrary close to the asymptotic lower bound of average delay.