We consider single- and multi-server FCFS systems, where jobs have a maximum waiting time (deadline) defined, e.g., by a service level agreement. The task is to minimize the long-run cumulative deadline violations. Job sizes (service durations) are observed upon arrival, and current queue backlogs are known. For a single FCFS server, the optimization task is to find the optimal admission policy that may reject a job upon arrival if admitting it would cause in future one or more deadlines to be violated (in expectation). For parallel FCFS servers, the policy must (i) either accept or reject a job upon arrival, and if accepted, (ii) assign it to one of the servers. For a single server, we obtain the optimal admission policy. For dispatching to parallel servers, we develop efficient heuristic admission and dispatching policies, whose performances are evaluated by means of numerical examples. Additionally, we give some exact closed-form results for heavy-traffic limits.