ABSTRACT

Is it possible to reduce the expected response time of every request at a web server, simply by changing the order in which we schedule the requests? That is the question we ask in this final project.

This paper proposes a method for improving the performance of web servers servicing static HTTP requests. The idea is to give preference to requests for small files.

The implementation is at the kernel level and involves controlling the order in which socket buffers are drained into the network. Experiments are implemented in a LAN environment.

We use the Linux operating system and the Apache web servers.

Results indicate that the size-based scheduling of connections yields some reductions in delay at web server. These results are in both mean response time and mean slowdown. Significantly, and counter to intuition, the requests for large files are only negligibly penalized.

Keywords: *size based scheduling, response time, slowdown.*