One might think that latency and throughput are related - if a server can answer N queries per second, doesn't that mean answering each query must take $1/N$ seconds on average? In fact, there is no such simple or obvious mathematical relationship between throughput and latency - it is possible for a caching server to have excellent latency characteristics but abysmal throughput or vice vers

up, crash, or drop packets as a result of running out of state storage space.

You will also need an appropriate testing tool. The current tool of choice is "resperf", which was specifically designed for measuring caching server throughput and works much better than dnsperf for this purpose. The resperf program is distributed with dnsperf (see Appendix A). Please read the resperf(8) man page for detailed instructions.