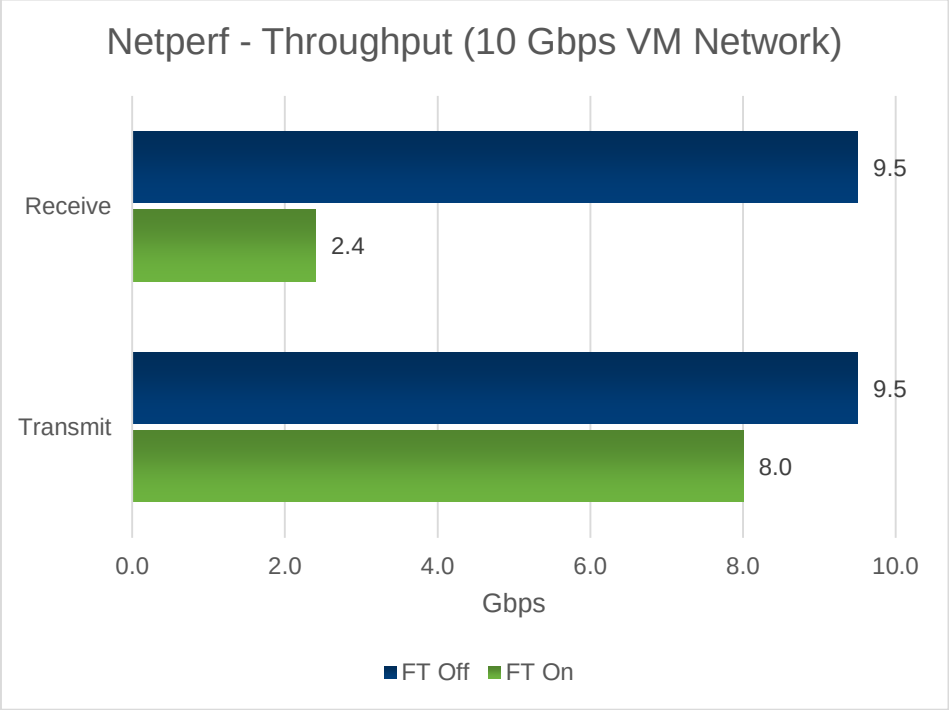

Figure 4. Netperf throughput on a 10 Gbps VM network for receive and transmit (higher is better)

Latency

This netperf experiment measures TCP/IP latency under FT. Fault tolerance introduces some delay to network output (measureable in milliseconds, as seen in Figure 5). The latency occurs because an FT-protected primary withholds network transmissions until the secondary acknowledges to the primary that it has reached an identical runtime state.

In this experiment, netperf was run with the TCP_RR configuration (single stream, no parallel streams) and the round-trip latency is reported here (it is the inverse of the round-trip transaction rate). TCP_RR is a pure latency benchmark: the sender transmits a 1-byte message and blocks waiting for a 1-byte response, the benchmark counts the number of serial transactions completed in a unit time, and there is no parallelization.

In a pure latency benchmark, all latency increases drop throughput (for example, latency increases 57 times and throughput drops identically). In real-world applications, however, workload performance is generally robust against network latency increases.

This is because normal server applications are not pure latency benchmarks. They handle multiple connections at a time, and each connection will transmit several packets worth of data before pausing to hear a response. The result is that real-world applications tolerate network latencies without dropping throughput. The netperf throughput experiment is an example of this, and the client/server workloads examined in this paper demonstrate the same thing. In all these experiments, while low-level network latency may increase, it typically results in slight visible effects on perceived UI interactivity, latency for database queries, and overall application throughput.

One aspect not measured by netperf is jitter and latency fluctuation. FT-protected virtual machines can vary widely in latency depending on the workload, and over time within a given workload. This can cause significant

jitter. Highly latency-sensitive applications, such as high frequency trading (HFT), or some voice-over-IP (VoIP) applications may experience high overhead with FT. However, some voice applications, where the bulk data is carried by separate processes where only call management traffic is FT-protected, would perform well.

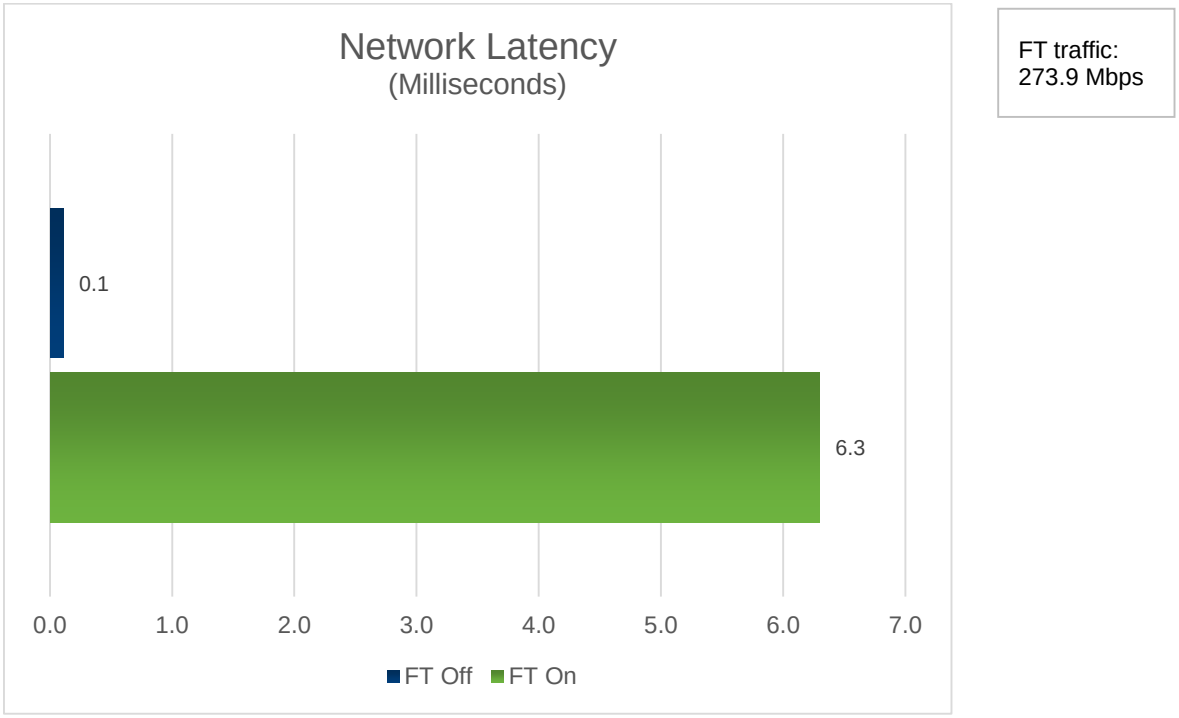

Figure 5. Network latency for VMs with FT off and on (lower is better)

IOMeter

IOMeter is an I/O subsystem measurement and characterization tool for Microsoft Windows. It is designed to produce a mix of operations to stress the disk. This benchmark ran random I/Os of various types. Figure 6 shows the FT-protected virtual machine achieves nearly as much throughput as the non-protected virtual machine.