4.

a. 100 * (10ms + 1ms) + 1ms + 1s = 2.101 seconds

b. Thread A => switch to B => 8 ms run => switch to A => 1 ms run:

   (0 + 1 + 8 + 1 + 1)ms * 100 = 1100ms

   1100ms + 200 ms = 1.3 seconds

c. The multi-thread program in part (b) is more efficient because it achieves the goal of responsiveness that is often exemplified with a server hosting multiple clients. Over time, the program described in part (a) could take longer than that in part (b) if more threads, with similar time deficits, are running.