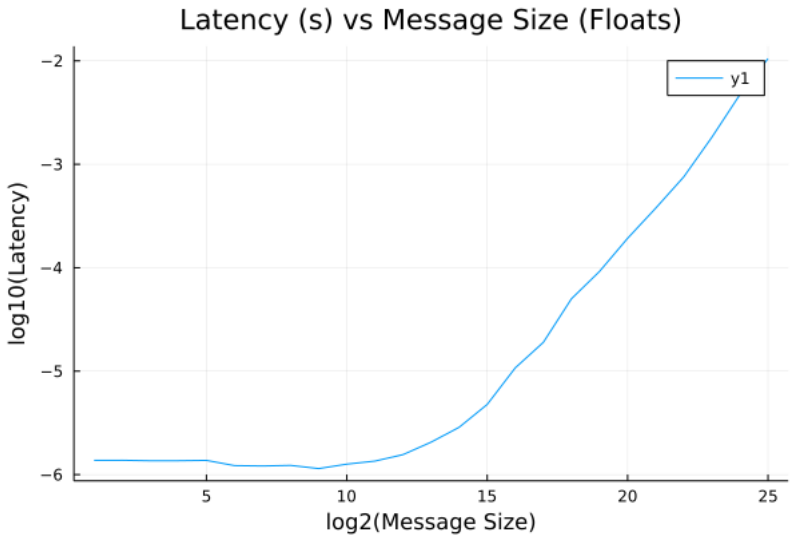
# MIT 18.337

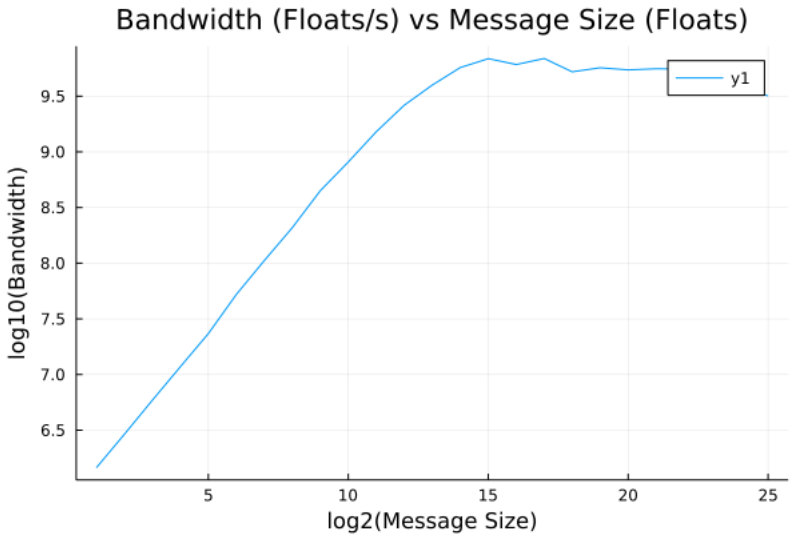
wsunadawong

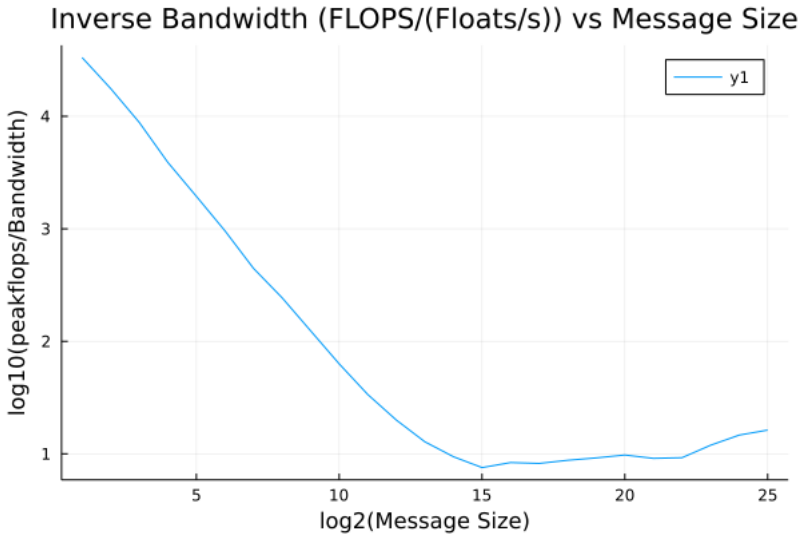
February 2022

Problem 2.

Part 1. When we recreate the Latency and bandwidth plots, we get the following:






These look very similar to the plots from the textbook, but with slightly better performance.

Part 2. Looking at the first plot for the smallest values, the minimal latency is approximately $10^{-6}$ seconds. Based on the second plot, the total bandwidth tops out at approximately $10^{9.7} \approx 5. \times 10^9$ floats/s. If we spend $k$ operations per number, and there are $N$ numbers, then the time spent calculating is

$$t_{calc} = kN/(peakflops)$$

Meanwhile, the time spent passing the message is

$$t_{pass} = N/(bandwidth)$$

Setting $t_{calc} = t_{pass}$ gives $k = peakflops/bandwidth$.

Therefore, the inverse bandwidth plot gives us the threshold for how many operations are necessary for parallelism to be worthwhile. For small messages, we need approximately $10^4$ operations on each number, and for large messages about 10 operations. Otherwise, most of the time will be spent sending messages.