Bandwidth and latency for different levels of CUDA communication

Plot shows performance for *ideal* scenario with *measurement* errors, but not possible *divergence* in practise

