Table **??** contains the speedup in time per iteration generated by parallel CPU over sequential CPU and GPU over parallel CPU, respectively. For synchronous SGD, this also represents the speedup in time to convergence. Given that parallel CPU uses 56 threads, we expect a speedup in the range of 56 over sequential CPU. This is the case for `news`. The speedup for `rcv1` is slightly below 56—due to the large size of the dataset which does not allow for efficient caching even in L3. We obtain super-linear speedup on `covtype`, `w8a`, and `real-sim`—on `w8a` the speedup is more than 400X for SVM. The reason for this is the improved cache behavior when all the cores are in use. `w8a` can be entirely cached in L1 due to its small size, while `real-sim` and `covtype` are cached in L2 and L3, respectively. None of these datasets can be cached on a single core when executing the sequential code. GPU improves further over parallel CPU by a factor of 1 to 5X.

task	**dataset**	**cpu-seq/cpu-par**	**cpu-par/gpu**
LR	covtype	112.54	1.23
	w8a	323.80	1.24
	real-sim	200.46	2.47
	rcv1	46.34	1.52
	news	65.27	5.66
SVM	covtype	99.63	1.32
	w8a	428.84	1.03
	real-sim	164.36	2.00
	rcv1	42.76	2.31
	news	58.23	5.63