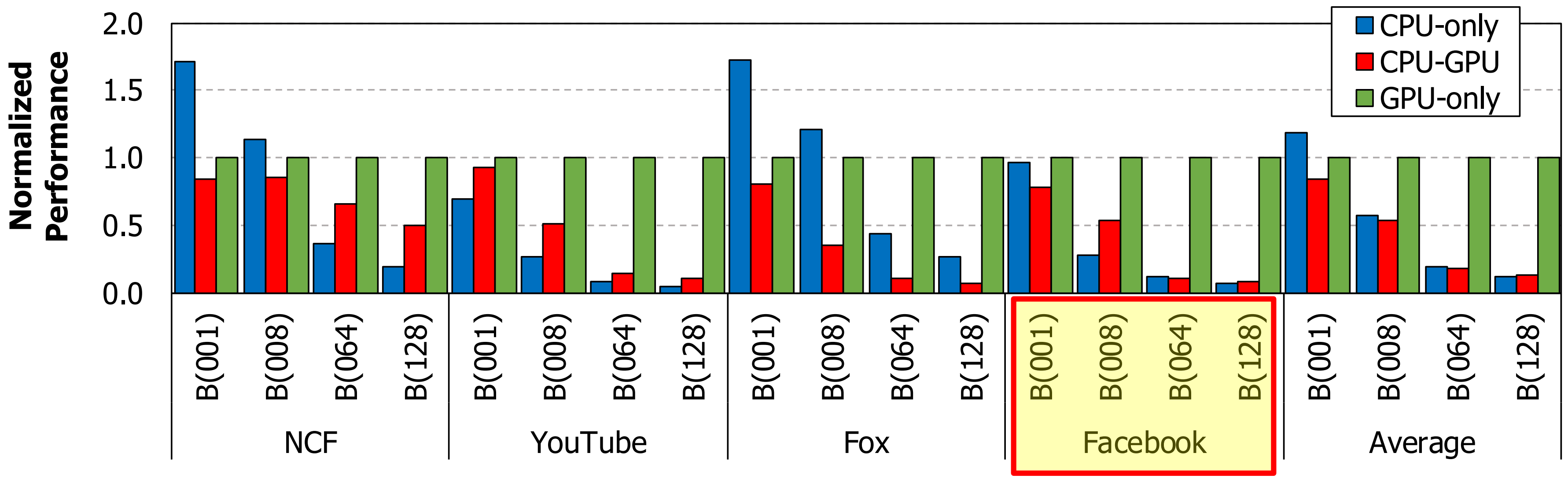
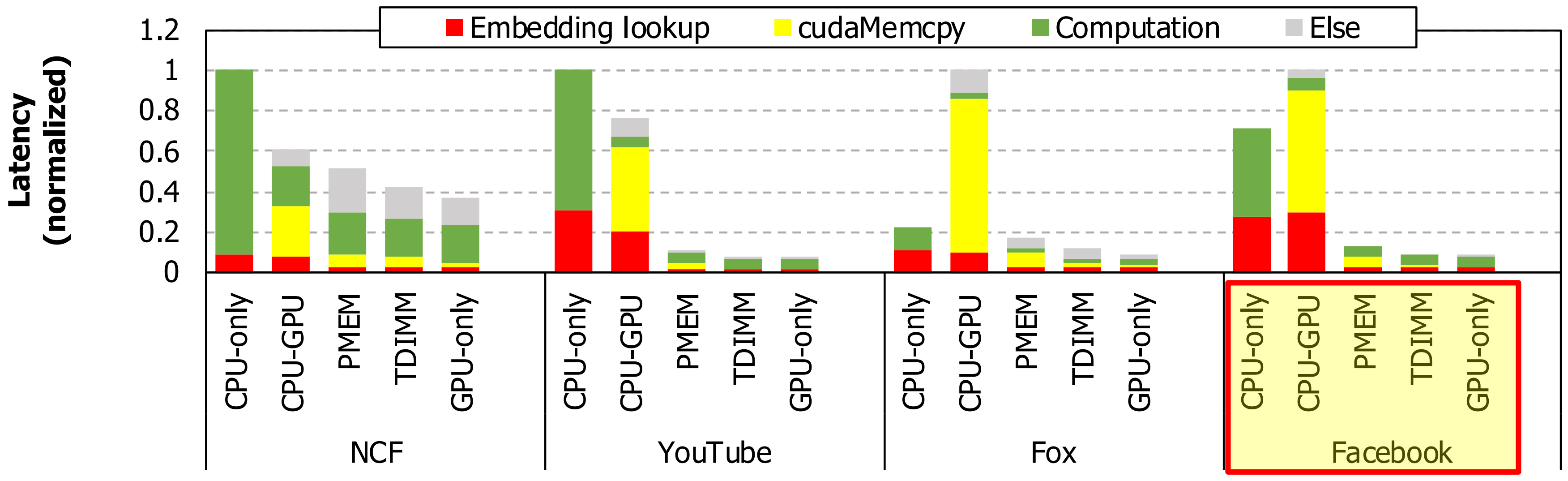
[1] Evaluated benchmarks and default configurations

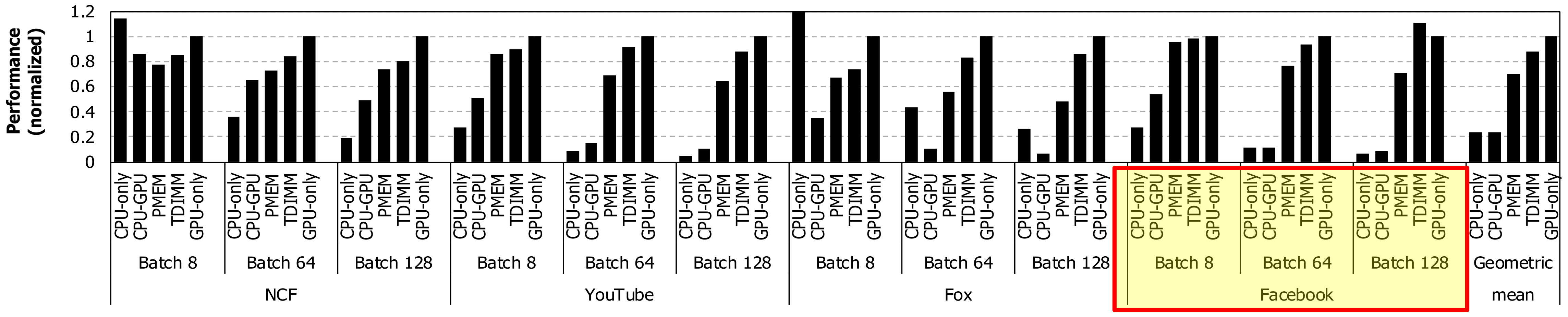
Network	Lookup tables	Max reduction	FC/MLP layers
NCF	4	2	4
YouTube	2	50	4
Fox	2	50	1
Facebook	8	25	6



[2] Performance of baseline `CPU-only` and hybrid `CPU-GPU` versions of recommender system, normalized to an oracular `GPU-only` version. Figure follows the same format and evaluation setting as assumed in Figure 4 of the submitted manuscript.



[3] Breakdown of latencies during an inference with batch 64, normalized to the slowest design point (i.e., `CPU-only` or `CPU-GPU`). Figure follows the same format and evaluation setting as assumed in Figure 13 of the submitted manuscript.



[4] Performance of the five design points of recommender systems, normalized to the oracular GPU (`GPU-only`). Figure follows the same format and evaluation setting as assumed in Figure 14 of the submitted manuscript.