Table **??** gives the hardware specifications of the NUMA machine and the NVIDIA Tesla K80 GPU used in this paper. While the number of cores and threads is much larger for the GPU, the numbers for the NUMA machine are quite high compared to previous CPU generations, e.g., 56 independent threads can run concurrently on a single machine. Although the amount of memory available on the CPU is 20X larger than on the GPU, the L2 cache on the GPU is 6X larger. This reflects the throughput emphasis of the GPU memory hierarchy as opposed to the latency optimization for CPU.

	NUMA	**GPU**
CPU/MP	2	13
cores	14 per CPU	192 per MP
blocks	-	16 per MP
threads	28 per CPU	2048 per MP
L1 cache	32+32 KB	48 KB
L2 cache	256 KB	1.5 MB
L3/shared	35 MB	48 KB
RAM/global	256 GB	12 GB