Although derived from the NUMA CPU factors proposed in DimmWitted , the GPU optimizations are quite different because of the layered parallelism consisting of blocks and threads, the SIMD execution within a warp, and the distinct memory hierarchy optimized for throughput rather than latency. While our initial goal has been to build an analytical model that identifies the optimal execution plan for any data/model configuration automatically, we have found experimentally that this choice is highly-dependent on data and model characteristics. Thus, we limit ourselves to providing practical rules of thumb to guide the optimal choice.

Dimension	**Strategies**
Data access path	row-major round-robin (row-rr) row-major chunking (row-ch) column-major round-robin (col-rr) column-major chunking (col-ch)
Model replication	kernel block thread example
Data replication	no replication (no-rep) k-wise replication (rep-2, rep-5, rep-10)