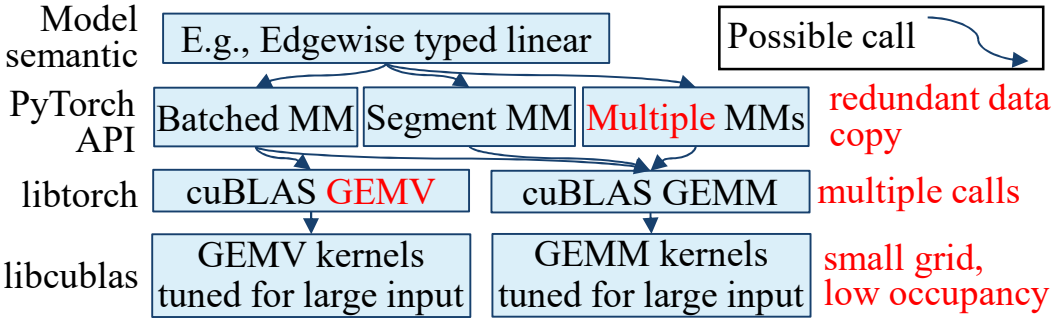
Model semantic
E.g., Edgewise typed linear
Possible call
PyTorch API
Batched MM
Segment MM
Multiple MMs
redundant data copy
libtorch
cuBLAS GEMV
cuBLAS GEMM
multiple calls
libcublas
GEMV kernels tuned for large input
GEMM kernels tuned for large input
small grid, low occupancy