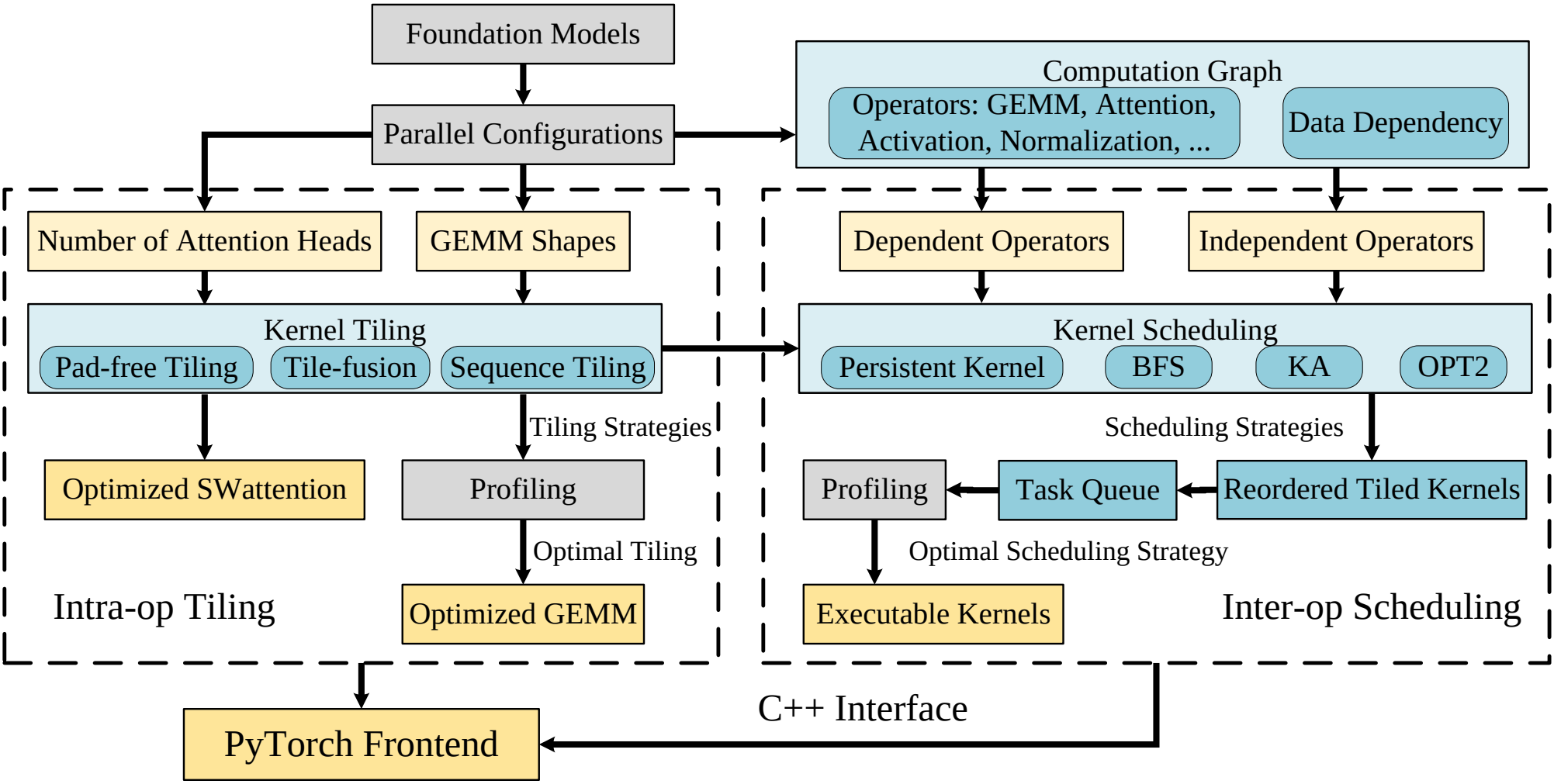
Foundation Models
Parallel Configurations
Computation Graph
Operators: GEMM, Attention, Activation, Normalization, ...
Data Dependency
Number of Attention Heads
GEMM Shapes
Dependent Operators
Independent Operators
Kernel Tiling
Pad-free Tiling
Tile-fusion
Sequence Tiling
Kernel Scheduling
Persistent Kernel
BFS
KA
OPT2
Tiling Strategies
Scheduling Strategies
Optimized SWattention
Profiling
Profiling
Task Queue
Reordered Tiled Kernels
Optimal Tiling
Optimal Scheduling Strategy
Intra-op Tiling
Optimized GEMM
Executable Kernels
Inter-op Scheduling
C++ Interface
PyTorch Frontend