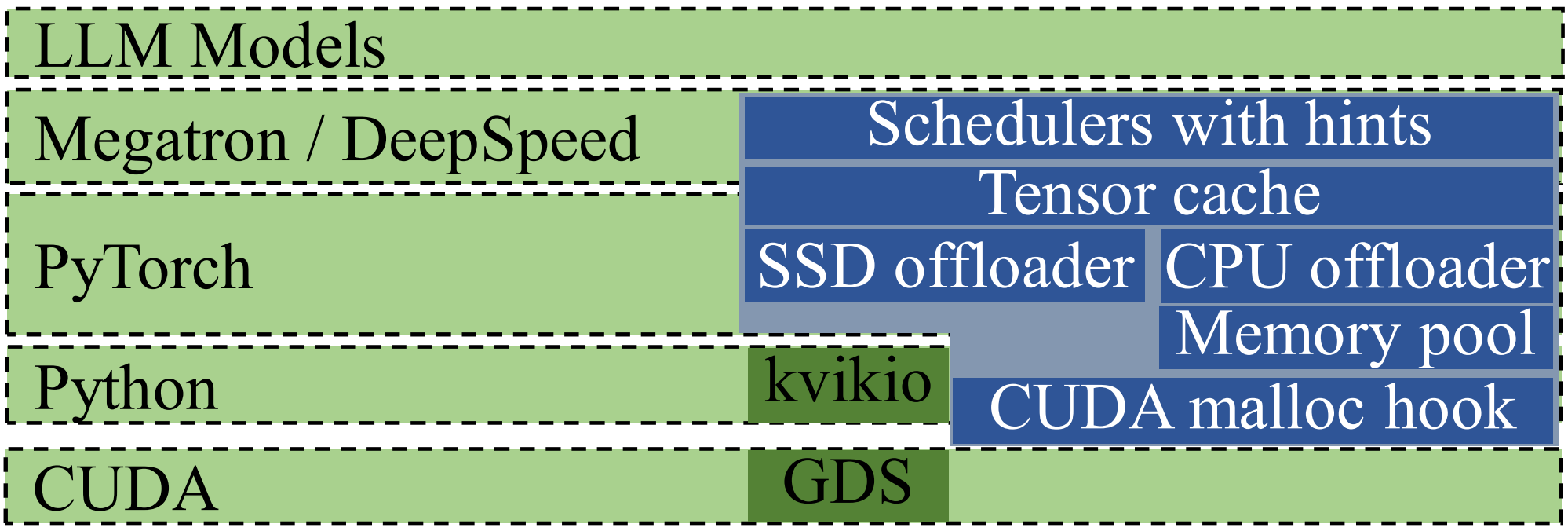
LLM Models
Megatron / DeepSpeed
PyTorch
Python
CUDA
Schedulers with hints
Tensor cache
SSD offloader
CPU offloader
Memory pool
kvikio
CUDA malloc hook
GDS