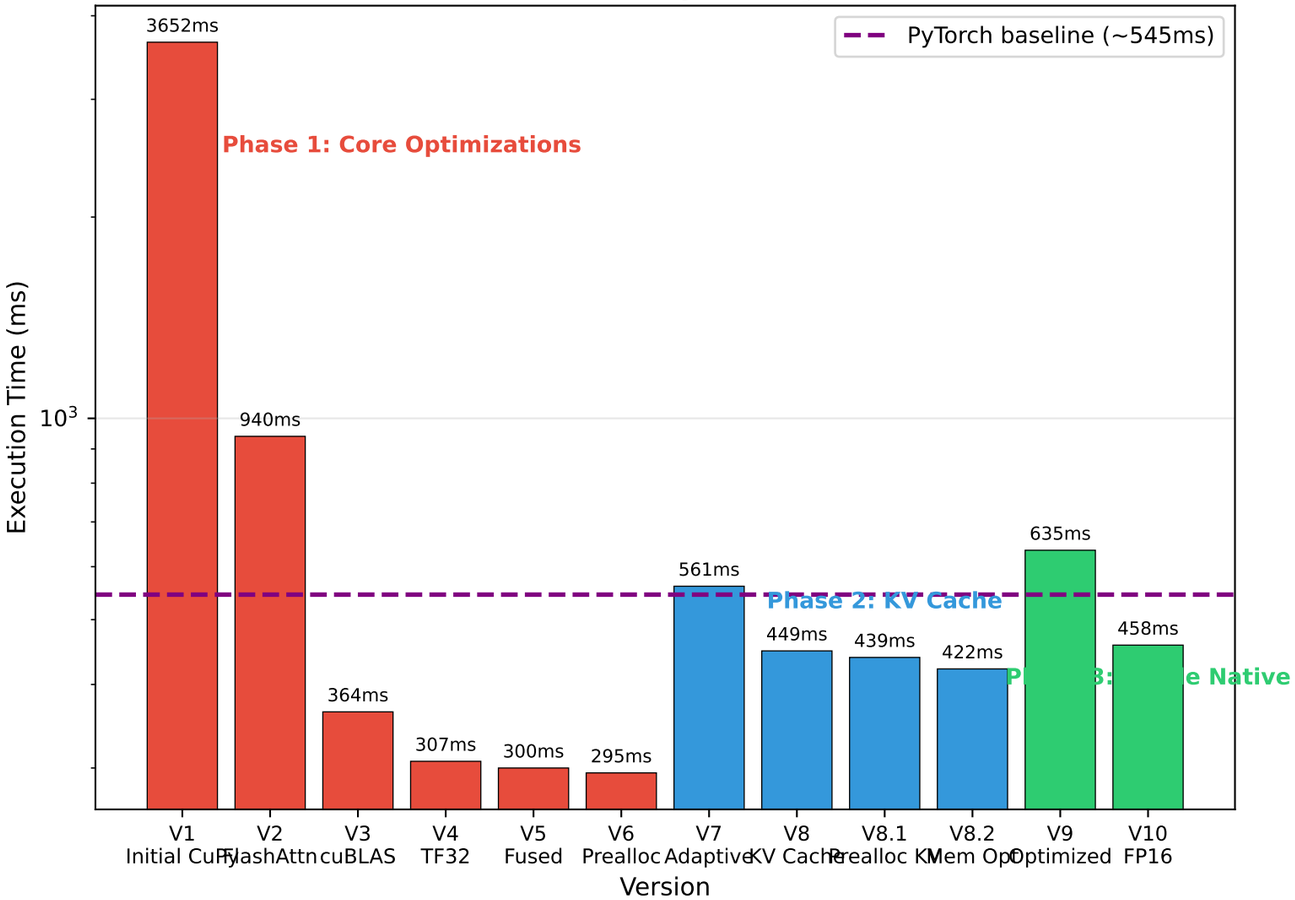
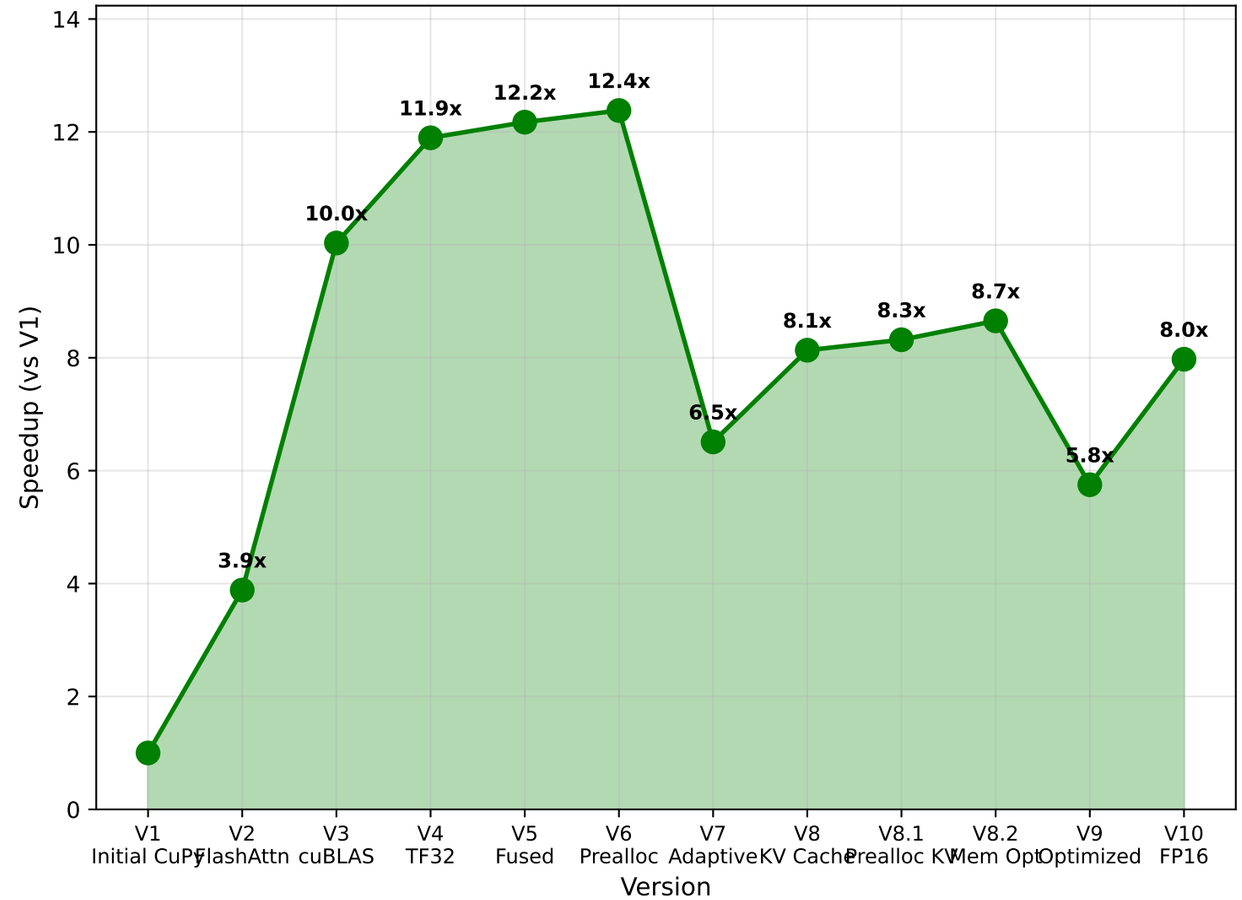
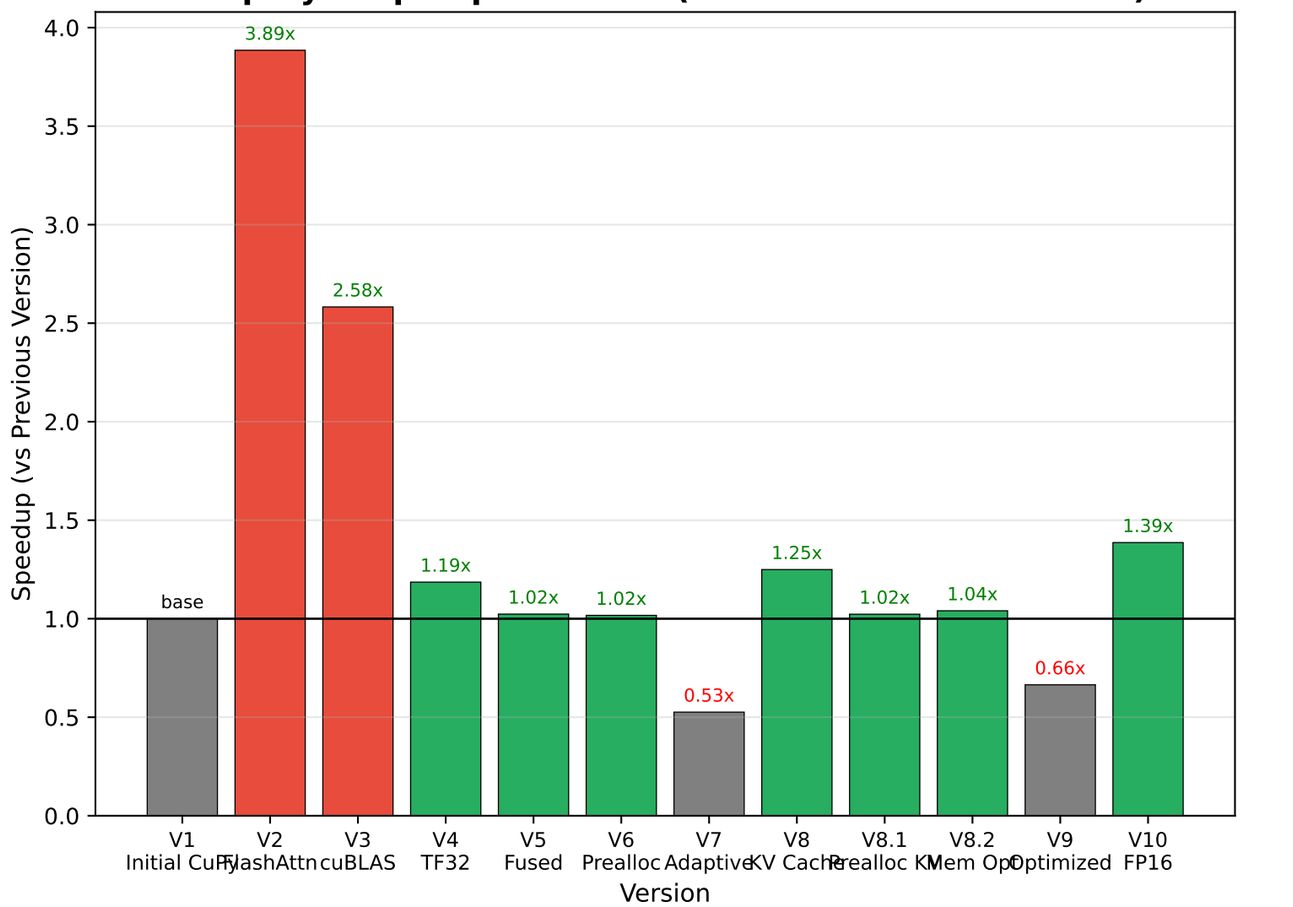
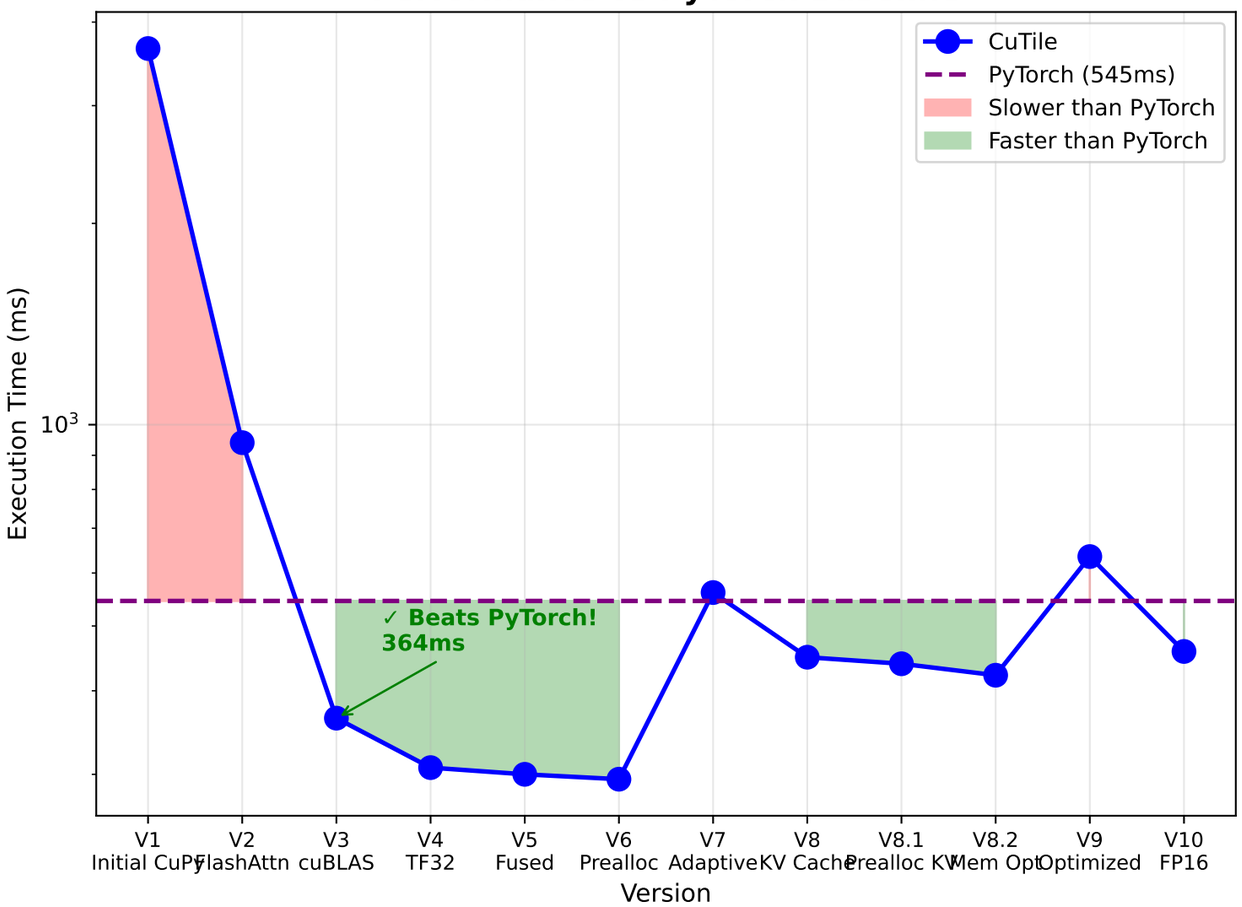

CuTile Optimization Journey: V1 → V10
PyTorch baseline (~545ms)
Phase 1: Core Optimizations
Phase 2: KV Cache
Phase 3: Native
3652ms
940ms
364ms
307ms
300ms
295ms
561ms
449ms
439ms
422ms
635ms
458ms
Execution Time (ms)
10^3
V1 Initial CuPy
V2 FlashAttn
V3 cuBLAS
V4 TF32
V5 Fused
V6 Prealloc
V7 Adaptive
V8 KV Cache
V8.1 Prealloc KV
V8.2 Mem Opt
V9 Optimized
V10 FP16
Version


Cumulative Speedup Relative to V1
Speedup (vs V1)
3.9x
10.0x
11.9x
12.2x
12.4x
6.5x
8.1x
8.3x
8.7x
5.8x
8.0x
V1 Initial CuPy
V2 FlashAttn
V3 cuBLAS
V4 TF32
V5 Fused
V6 Prealloc
V7 Adaptive
V8 KV Cache
V8.1 Prealloc KV
V8.2 Mem Opt
V9 Optimized
V10 FP16
Version


Step-by-Step Improvement (Each Version vs Previous)
Speedup (vs Previous Version)
base
3.89x
2.58x
1.19x
1.02x
1.02x
0.53x
1.25x
1.02x
1.04x
0.66x
1.39x
V1 Initial CuPy
V2 FlashAttn
V3 cuBLAS
V4 TF32
V5 Fused
V6 Prealloc
V7 Adaptive
V8 KV Cache
V8.1 Prealloc KV
V8.2 Mem Opt
V9 Optimized
V10 FP16
Version


Performance vs PyTorch Baseline
CuTile
PyTorch (545ms)
Slower than PyTorch
Faster than PyTorch
✓ Beats PyTorch!
364ms
Execution Time (ms)
10^3
V1 Initial CuPy
V2 FlashAttn
V3 cuBLAS
V4 TF32
V5 Fused
V6 Prealloc
V7 Adaptive
V8 KV Cache
V8.1 Prealloc KV
V8.2 Mem Opt
V9 Optimized
V10 FP16
Version