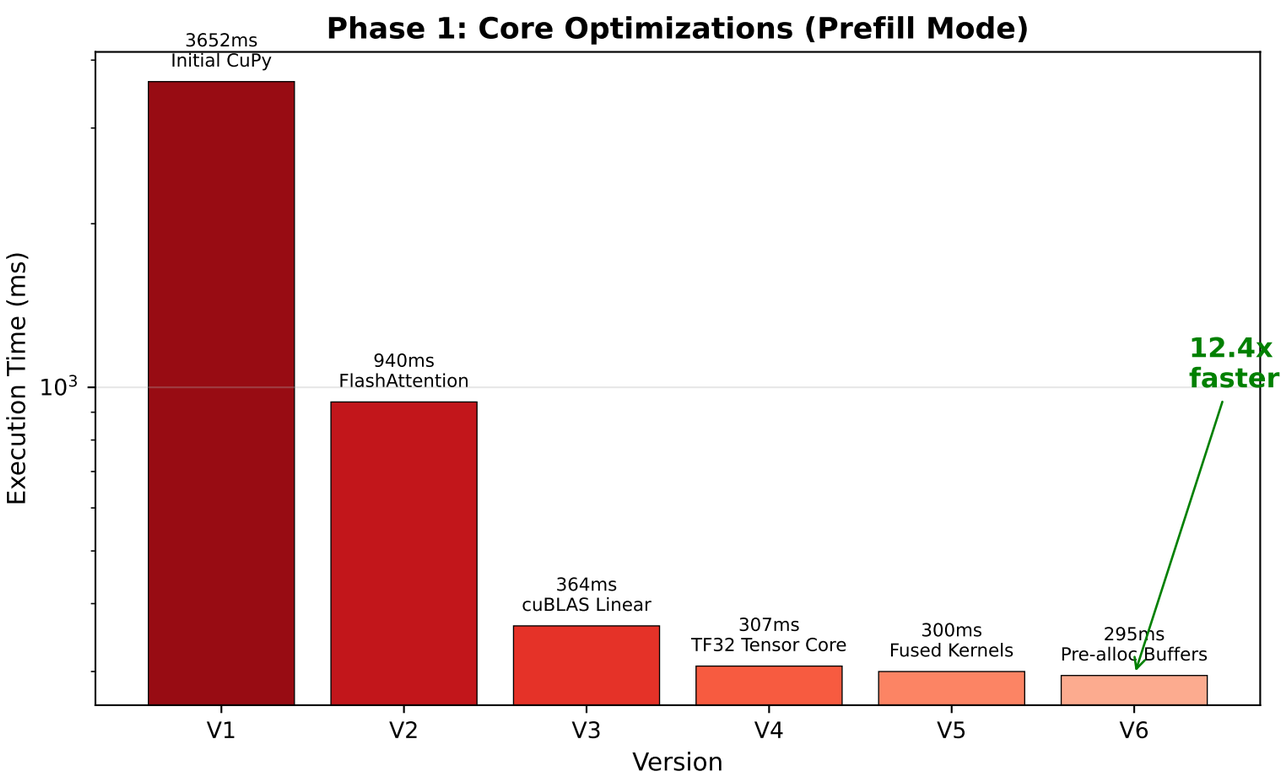
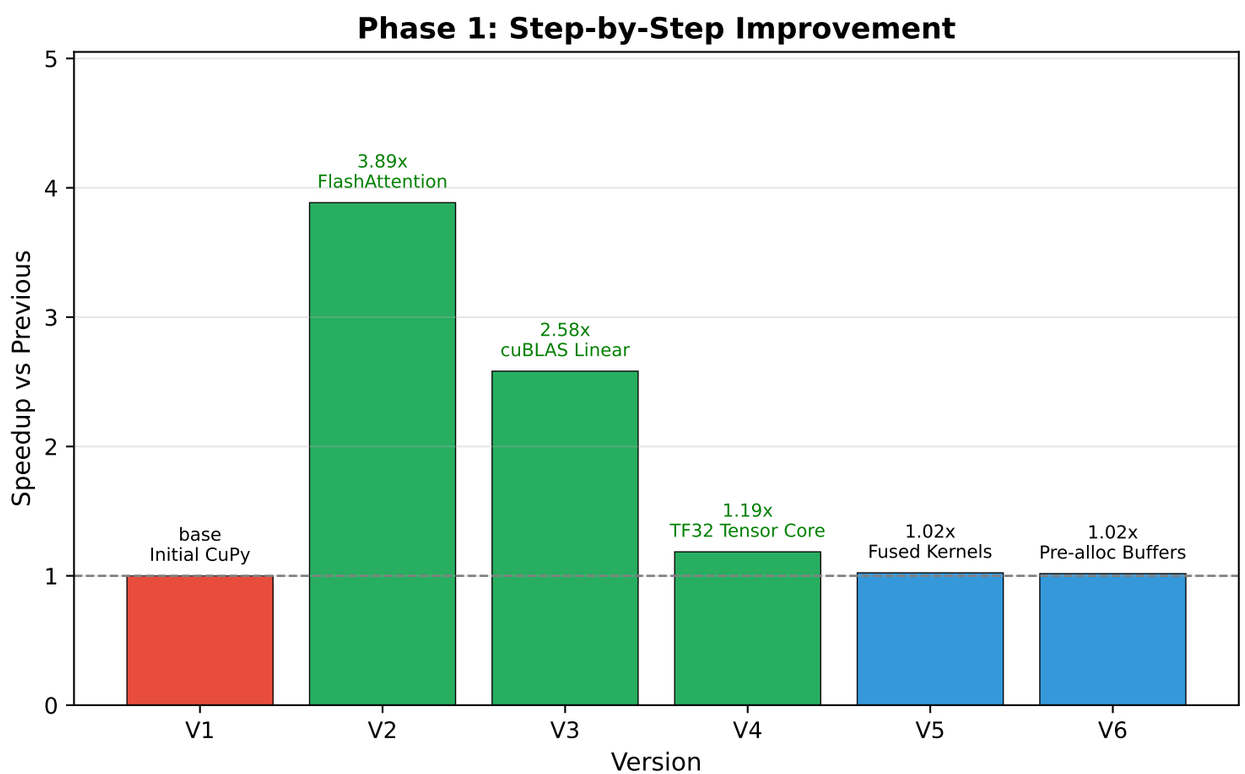
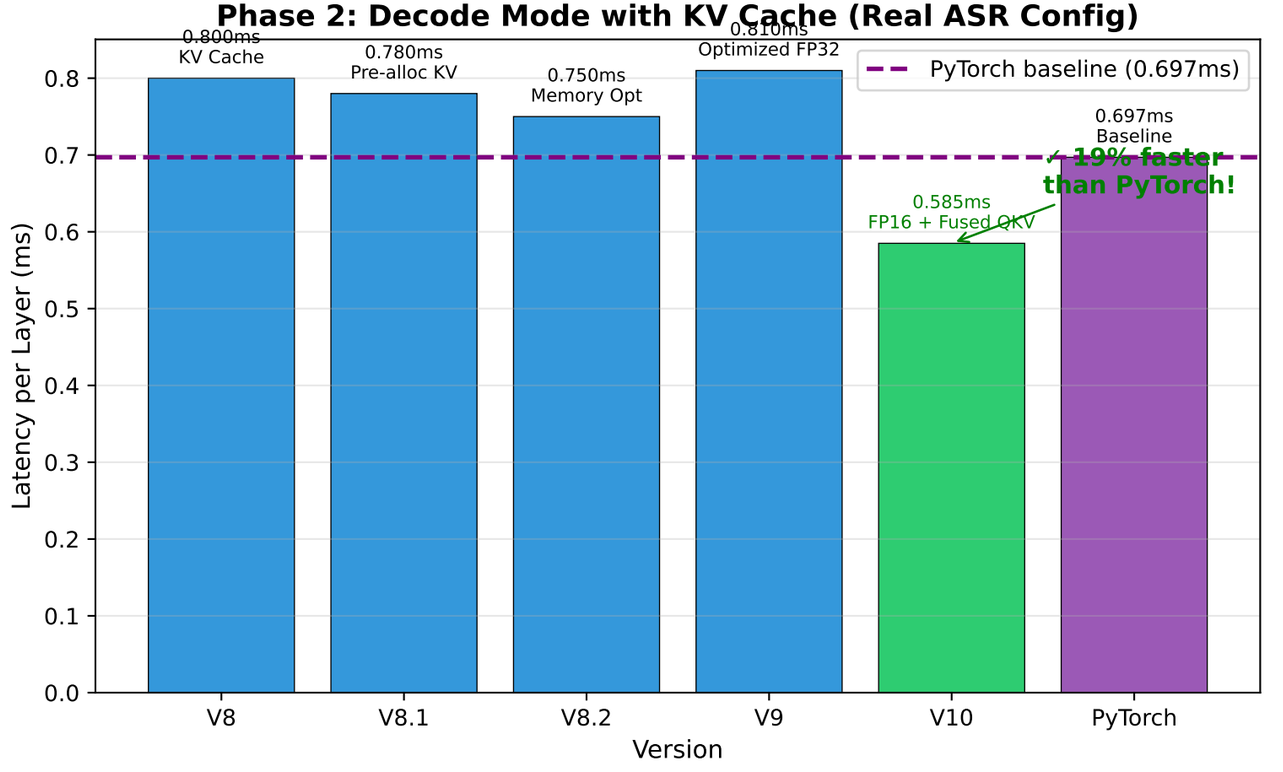
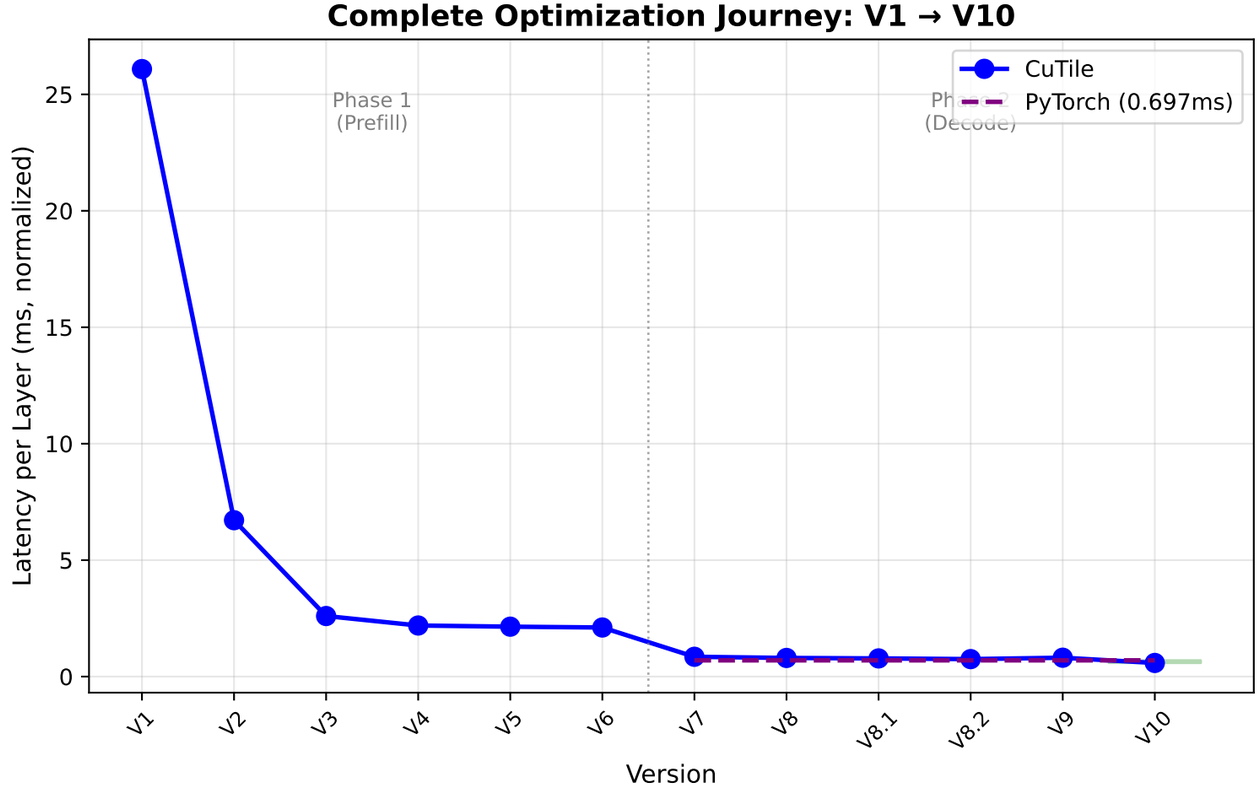

Phase 1: Core Optimizations (Prefill Mode)
Execution Time (ms)
10^3
3652ms
Initial CuPy
940ms
FlashAttention
364ms
cuBLAS Linear
307ms
TF32 Tensor Core
300ms
Fused Kernels
295ms
Pre-alloc Buffers
12.4x
faster
V1
V2
V3
V4
V5
V6
Version


Phase 1: Step-by-Step Improvement
Speedup vs Previous
0
1
2
3
4
5
base
Initial CuPy
3.89x
FlashAttention
2.58x
cuBLAS Linear
1.19x
TF32 Tensor Core
1.02x
Fused Kernels
1.02x
Pre-alloc Buffers
V1
V2
V3
V4
V5
V6
Version


Phase 2: Decode Mode with KV Cache (Real ASR Config)
Latency per Layer (ms)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.800ms
KV Cache
0.780ms
Pre-alloc KV
0.750ms
Memory Opt
0.810ms
Optimized FP32
0.585ms
FP16 + Fused QKV
0.697ms
Baseline
PyTorch baseline (0.697ms)
✓ 19% faster
than PyTorch!
V8
V8.1
V8.2
V9
V10
PyTorch
Version


Complete Optimization Journey: V1 → V10
Latency per Layer (ms, normalized)
0
5
10
15
20
25
Phase 1
(Prefill)
Phase 2
(Decode)
CuTile
PyTorch (0.697ms)
V1
V2
V3
V4
V5
V6
V7
V8
V8.1
V8.2
V9
V10
Version