Cumulative Speedup Progression

Log-scale cumulative speedup across optimization phases

Cumulative Speedup (log scale)

100x
30x
10x
3x
1x

1.0x
(33.0 min)

12.0x
(2.8 min)

30.0x
(1.1 min)

42.0x
(0.8 min)

46.0x
(0.7 min)

Baseline (CPU, sequential)

Batch-parallel forward

Mixed precision (FP16)

Batch size 128

Async GPU transfers

9x more data (2367 games)