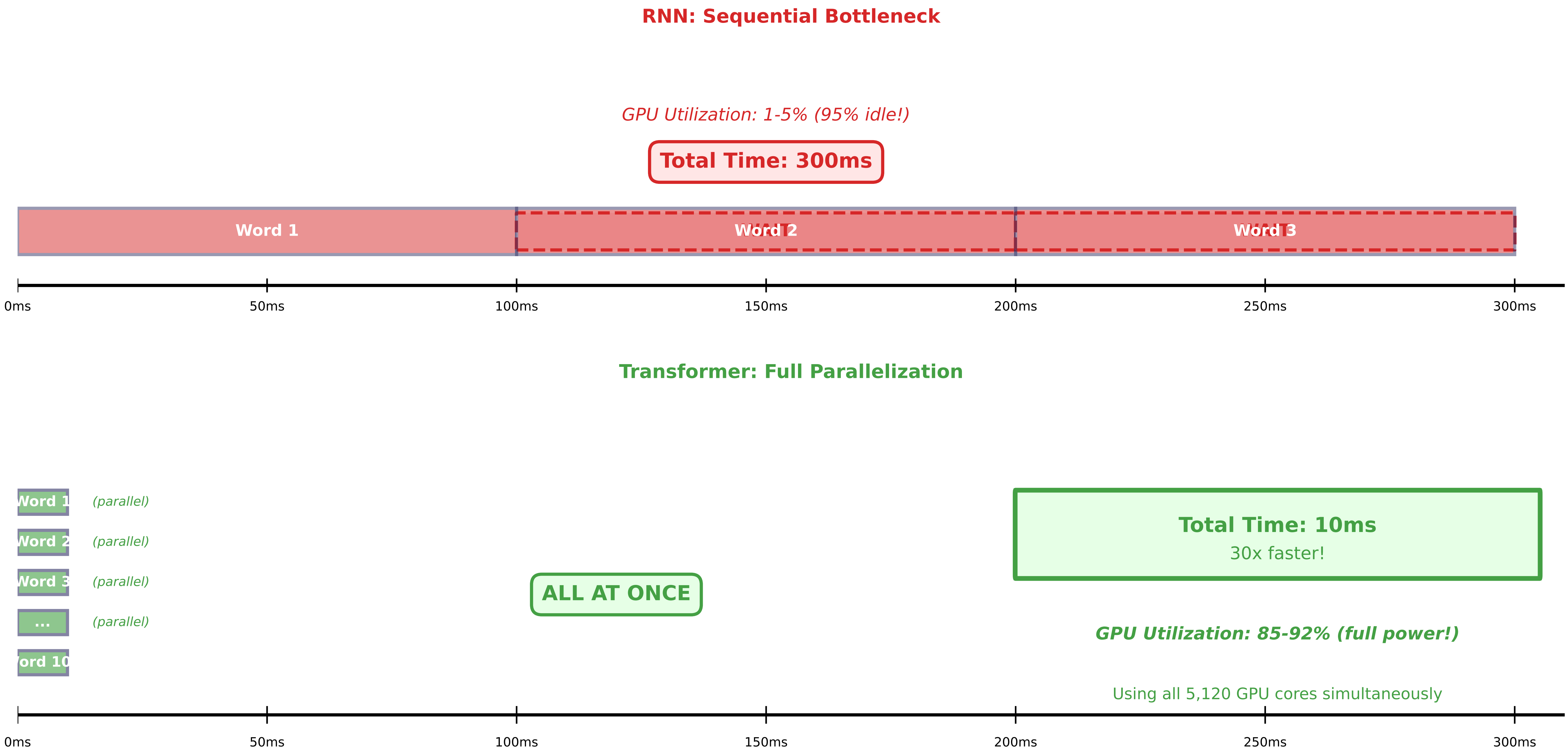
Processing Timeline: RNN (Sequential) vs Transformer (Parallel)


RNN: Sequential Bottleneck
GPU Utilization: 1-5% (95% idle!)
Total Time: 300ms
Word 1
Word 2
WAIT
Word 3
WAIT
0ms
50ms
100ms
150ms
200ms
250ms
300ms
Transformer: Full Parallelization
Word 1
(parallel)
Word 2
(parallel)
Word 3
(parallel)
(parallel)
ord 10
ALL AT ONCE
Total Time: 10ms
30x faster!
GPU Utilization: 85-92% (full power!)
Using all 5,120 GPU cores simultaneously
0ms
50ms
100ms
150ms
200ms
250ms
300ms