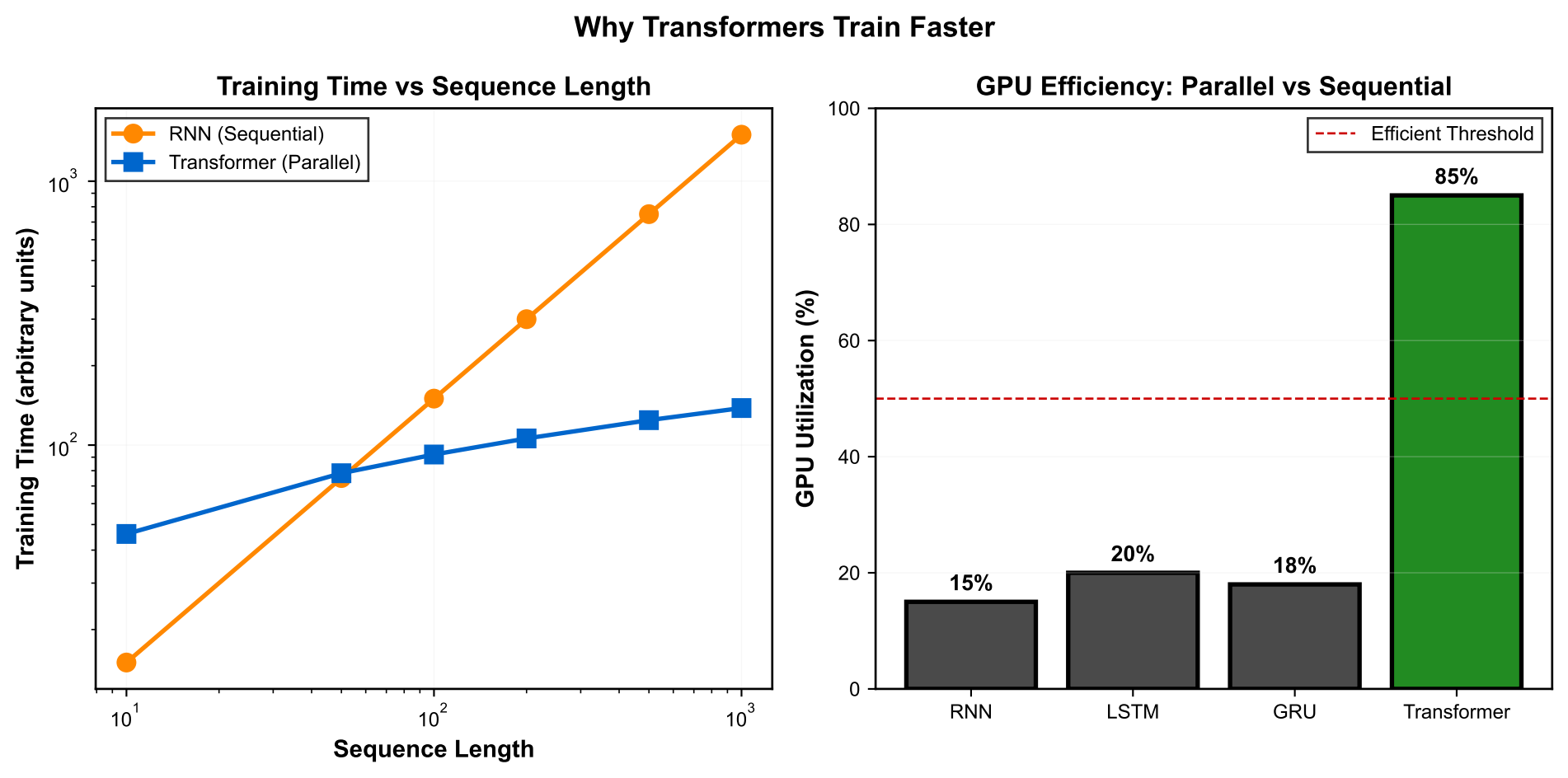
Why Transformers Train Faster
Training Time vs Sequence Length
RNN (Sequential)
Transformer (Parallel)
Training Time (arbitrary units)
Sequence Length
GPU Efficiency: Parallel vs Sequential
Efficient Threshold
GPU Utilization (%)
15%
20%
18%
85%
RNN
LSTM
GRU
Transformer