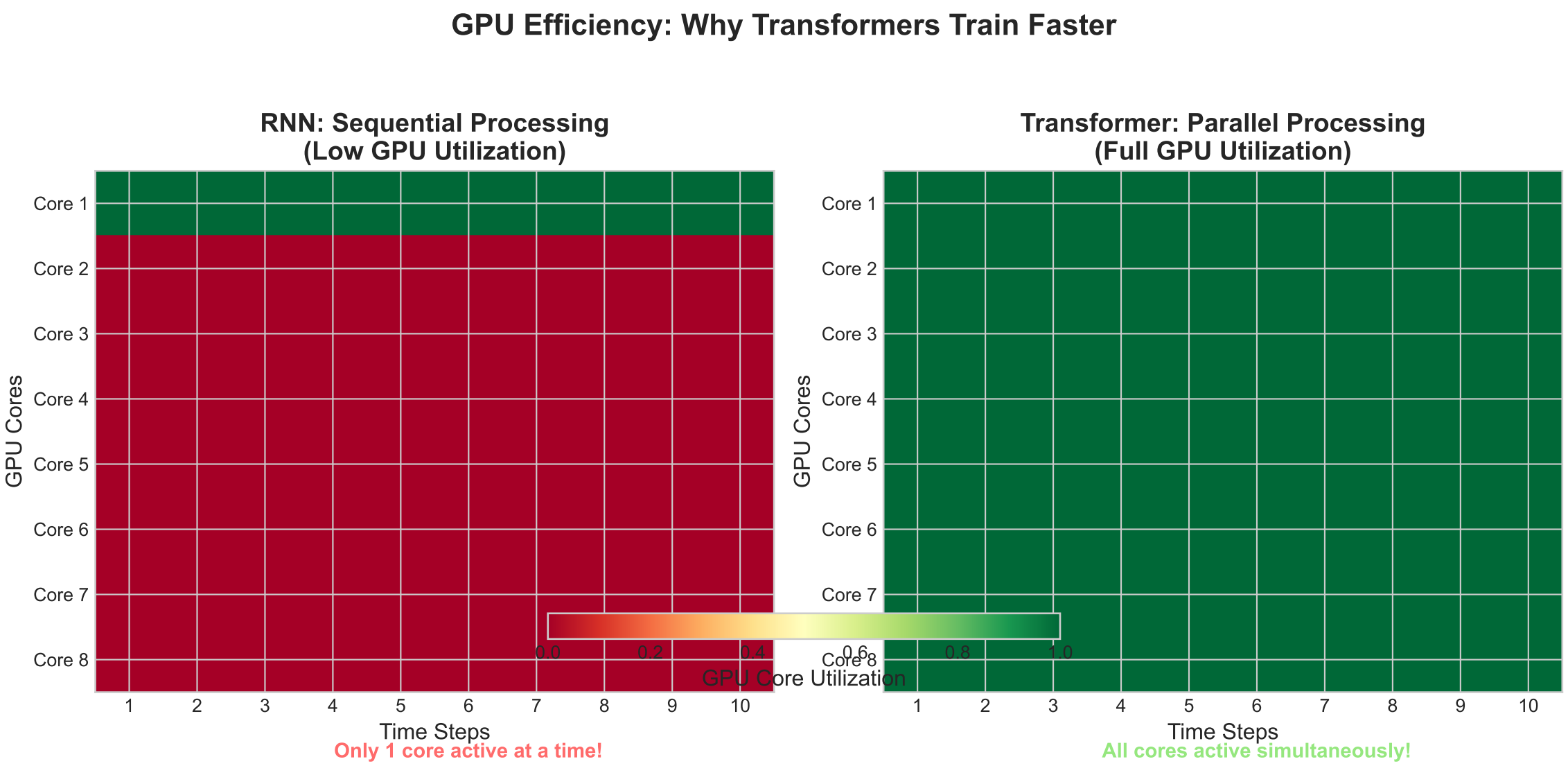
GPU Efficiency: Why Transformers Train Faster
RNN: Sequential Processing
(Low GPU Utilization)
Transformer: Parallel Processing
(Full GPU Utilization)
GPU Cores
Core 1
Core 2
Core 3
Core 4
Core 5
Core 6
Core 7
Core 8
1
2
3
4
5
6
7
8
9
10
Time Steps
0.0
0.2
0.4
0.6
0.8
1.0
GPU Core Utilization
Only 1 core active at a time!
All cores active simultaneously!