

Table 1: Fall 2025 Ring Attention Implementation: Prefill and decode performance for Ring vs Regular attention.

| Strategy | Prompt | TTFT (ms) | TTFT/token | Decode (ms) | Total (ms) | Comm (ms) | Comp/Comm |
|----------|--------|-----------|------------|-------------|------------|-----------|-----------|
| Ring | 256 | 1576.04 | 6.16 | 558.09 | 18318.77 | 12823.92 | 0.43 |
| Regular | 256 | 66.58 | 0.26 | 57.10 | 1779.55 | 0.00 | N/A |
| Ring | 512 | 1513.53 | 2.96 | 565.80 | 18487.42 | 12773.48 | 0.45 |
| Regular | 512 | 71.20 | 0.14 | 55.99 | 1751.03 | 0.00 | N/A |
| Ring | 1024 | 1667.12 | 1.63 | 566.92 | 18674.79 | 12690.79 | 0.47 |
| Regular | 1024 | 108.38 | 0.11 | 56.90 | 1815.44 | 0.00 | N/A |
| Ring | 4096 | 3220.51 | 0.79 | 573.09 | 20413.19 | 13001.37 | 0.57 |
| Regular | 4096 | 453.37 | 0.11 | 47.90 | 1890.37 | 0.00 | N/A |

Table 2: Our Implementation: Prefill performance for Ring vs Regular attention with async communication overlap.

| Strategy | Prompt | TTFT (ms) | TTFT/token | Slowdown |
|----------|--------|-----------|------------|--------------|
| Ring | 256 | 430.61 | 1.68 | $7.06\times$ |
| Regular | 256 | 60.98 | 0.24 | — |
| Ring | 512 | 508.69 | 0.99 | $4.00\times$ |
| Regular | 512 | 127.15 | 0.25 | — |
| Ring | 1024 | 621.78 | 0.61 | $2.56\times$ |
| Regular | 1024 | 242.92 | 0.24 | — |
| Ring | 4096 | 1386.12 | 0.34 | $1.34\times$ |
| Regular | 4096 | 1035.80 | 0.25 | — |
| Ring | 8192 | 2472.77 | 0.30 | $1.16\times$ |
| Regular | 8192 | 2132.00 | 0.26 | — |
| Ring | 16384 | 5414.71 | 0.33 | $1.06\times$ |
| Regular | 16384 | 5085.98 | 0.31 | — |