Scaling Across GPU architectures
$r = 16$, $n=40960$

Scaling Factor vs Method

- H100 (Hopper)
- A100 (Ampere)
- V100 (Volta)