

From Scheduled to Real Overlap

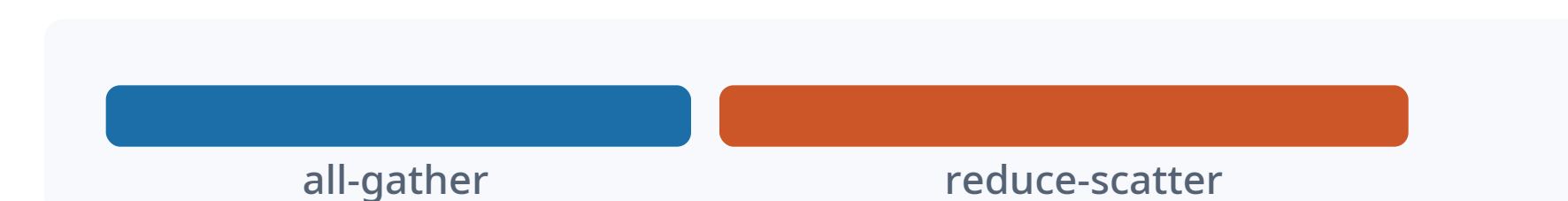
All-Gather / Reduce-Scatter in PyTorch Distributed

Jeffrey Mahou · Sheng Fu

1. Problem and mechanism

FSDP gathers parameters and reduce-scatters gradients. A shared data-parallel group can serialize them despite overlapping calls.

SERIAL PATH



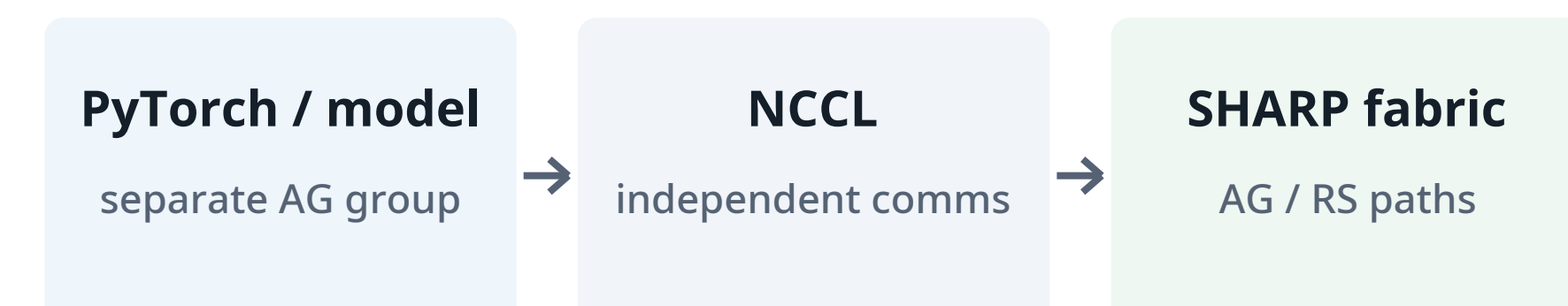
CONCURRENT PATH



Separate AG and RS communicators let both collectives progress in the same window.

Schematic: the pair finishes when the slower RS call finishes.

WHAT CHANGES



SHARPV4 GA target: January 2027

WHY THESE PATHS CAN OVERLAP

- **All-gather: multicast fan-out**
Receive-heavy delivery of parameter shards.
- **Reduce-scatter: in-network reduction**
Send-heavy contribution of gradient shards.

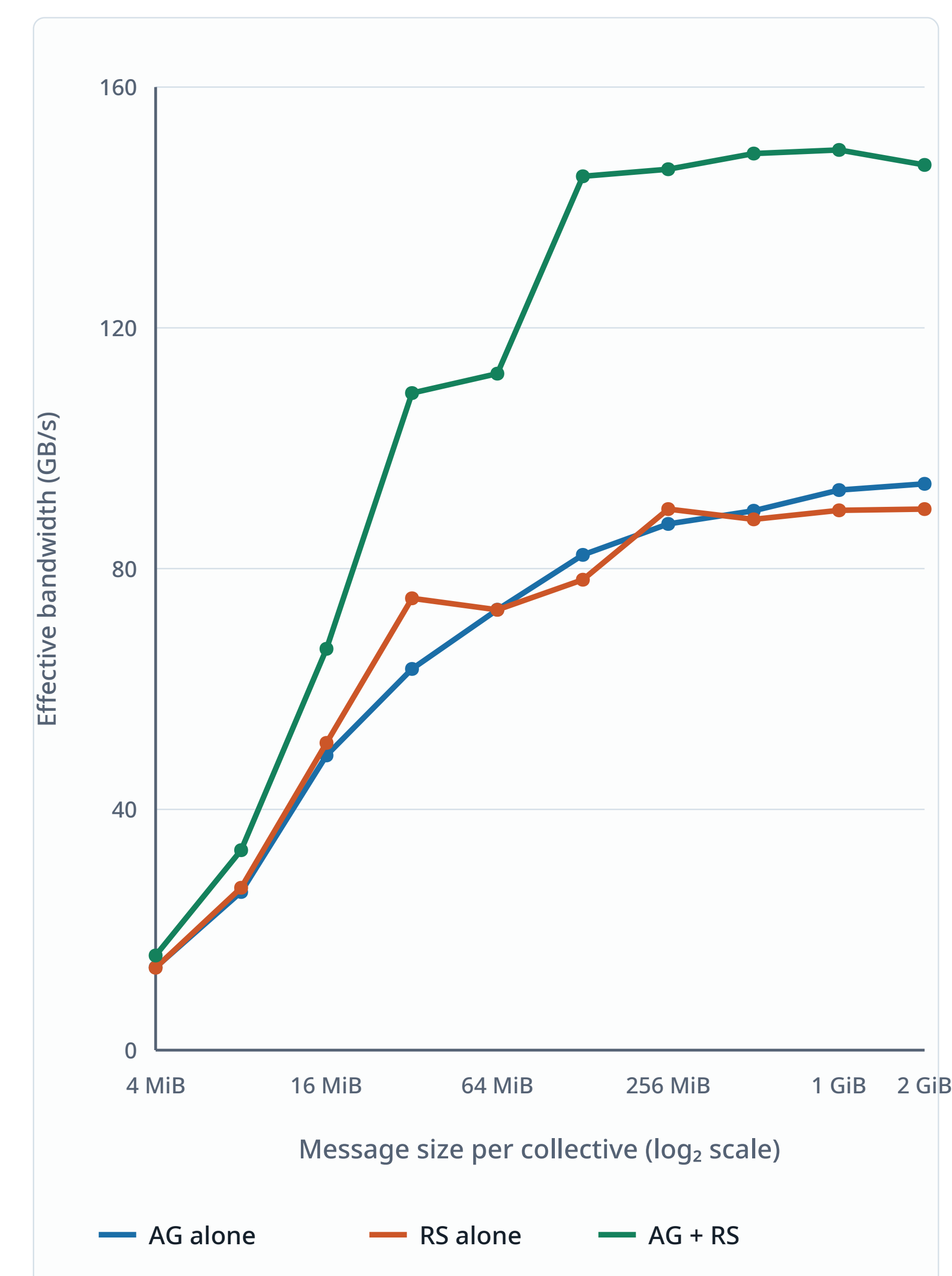
These directions can make progress together.

Megatron-LM PR #2663 · github.com/NVIDIA/Megatron-LM/pull/2663

2. Maestro: measured concurrent progress

EAST-WEST · 64 RANKS · 16 GB200 NODES · FP16

10 sizes · registered · SHARP DEFAULTS.

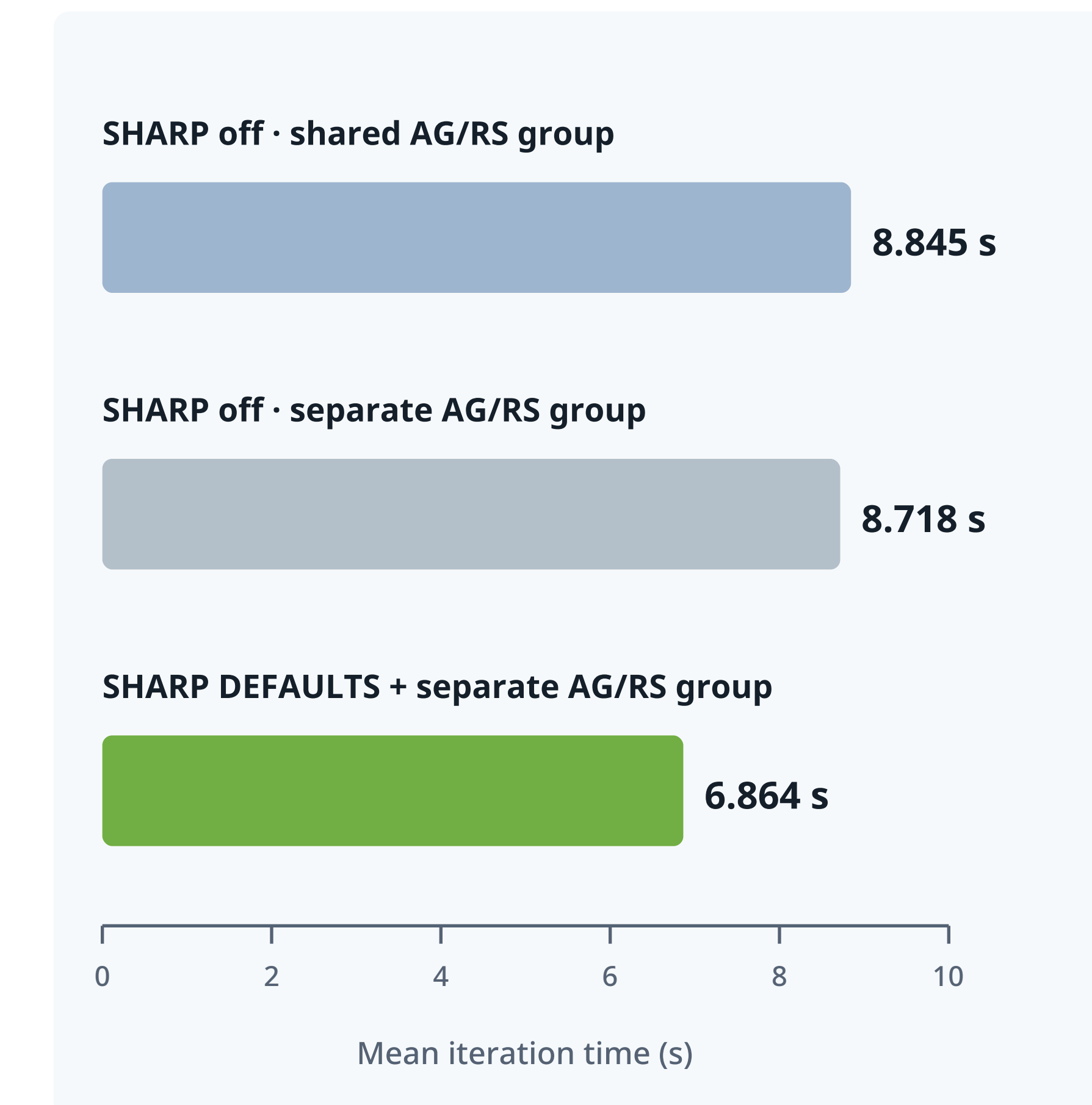


Maestro PR #2702 · github.com/pytorch/benchmark/pull/2702

3. Qwen at training scale

EAST-WEST · QWEN3-235B-A22B · 32 GB200 NODES

Megatron FSDP + DeepEP · 128 steps · synthetic data.
Mean training iteration time; lower is better.



22.4% lower · SHARP defaults + separate group

1.4% lower observed · separate group, SHARP off

Separate AG and RS groups enable overlap; Maestro traces verify it; Qwen measures the combined setup.