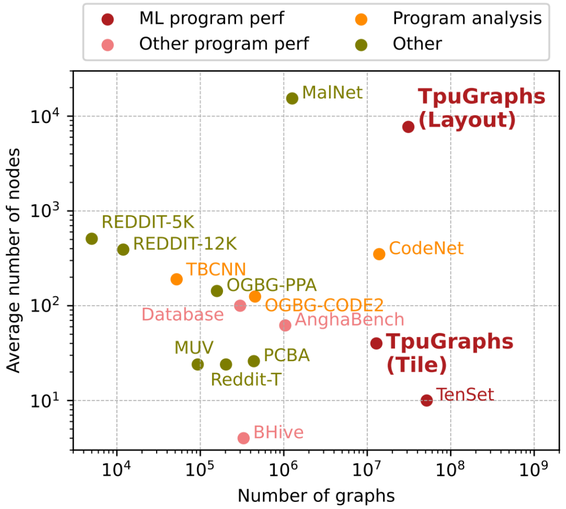
ARKADII

ONE SLIDE TALK

About GNN in TPU and so on...

THIS IS VERY IMPORTANT FOR ALL MANKIND



- TPU is ASIC (high volume of lpc), higher ops/joule, higher bandwidth
- AI is a graph
- Graphs are large (no way to use GNN directly)
- Not too large
- Hardware optimal execution is high profitable (Google used autotuner for XLA, a machine learning compiler, discovered 10–20% speedup on SOTA models)
- Autotuner is expensive
- Compilers need scoring