ParaGraph: An application-simulator interface and toolkit for hardware-software co-design

Mikhail Isaev1, Nic McDonald2, Jeff Young1, Rich Vuduc1

1Georgia Institute of Technology, 2Nvidia

Research problem
System co-design crosses many domain boundaries
Most HW/SW co-design studies fall under one of the categories:
  - Optimizing software on existing hardware;
  - Designing future hardware systems with limited and simplistic application models.
Lack of infrastructure to model both future hardware AND applications with appropriate fidelity

Cost Model
High-Level Language or Framework
Application
ParaGraph
Algorithm Translation
IR
Graph IR
Runtime
Simulator

ParaGraph goal – a real software model for hardware people and future hardware model for software people
ParaGraph for simulators is what LLVM is for real hardware

ParaGraph Workflow

Case study 1: co-design on unified HW/SW search space
ParaGraph allows SW engineers to model system-level SW, such as communication libraries, before system deployment
With ParaGraph we navigate landscape of SW and HW parameters simultaneously

Case study 2: model accuracy analysis
ParaGraph is validated against MLPerf training trace from 64-cores TPU v3 system
ParaGraph and TPU run the same HLO code
Network DES model matches the trace
Compute rooftop allows modeling longer application run

Case study 3: in-network reduction analysis
ParaGraph provides actual applications to future hardware engineers
ParaGraph helps assessing actual performance benefits of future hardware, and discovering potential performance bottlenecks, for example, performance with fat tree tapering

Conclusion
ParaGraph is a versatile co-design tool that effectively decouples application from hardware modeling, mutually beneficial to hardware and software engineers
ParaGraph supports various modeling workflows spanning several frontends (TF, JAX), backends (SuperSim, n10), and approaches (trace-driven, execution-driven, motif)
Find us on https://github.com/paragraph-sim

Acknowledgement
Portions of this work were performed during internships at Google and Nvidia.