Performance Analysis using the Roofline Model

Samuel Williams (SWWilliams@lbl.gov), Charlene Yang, Khaled Ibrahim, Thorsten Kurth, Nan Ding, Jack Deslippe, Leonid Oliker
CRD/NERSC, Lawrence Berkeley National Laboratory

Introduction
- Roofline is a throughput-oriented performance model
- Tracks rates not times
- Independent of ISA and architecture
- Applies to CPUs, GPUs, Google TPUs, FPGAs, etc...
- Defines Good Performance

Roofline on GPUs
- Developed a Roofline methodology POC for analyzing applications running on NVIDIA GPUs
- Use NVProf to collect Roofline-related metrics (FLOPs, cache/DRAM data movement, etc...)
- BerkeleyGW (Materials)
  - nw increases data reuse in inner loop
  - More fops for fixed data movement
  - Understand cache effects
  - Quantity effects of FMA-MUL ratio (disable FMA in compiler)
- Observations...
  - High correlation with HBM BW
  - FMA doesn't hit FMA ceiling
  - High RF and L2 Locality
  - Minimal increases in L1 locality

Roofline for TensorFlow
- Multiple variants of GSRB smoother...
  - GSRB_FP does 2k the work but is trivial to implement
  - STRIDE2 requires more complex logic, memory access and prediction
- Observations...
  - High correlation with HBM BW for large problem sizes (level=5)
  - Moderate L1 cache locality
  - Low reuse in the L2 cache for GSRB_FP variant
  - STRIDE2 performance changes due to decline in intensity

Scaling Trajectories
- Performance as a function of thread concurrency provides little insight
- Need better approach to understand turn overs in performance
- Use Roofline to analyze thread scalability
  - "Roofline Scaling Trajectories"
    - 2D scatter plot of performance as a function of intensity and concurrence
    - Identify loss in performance due to increased cache pressure (data movement)
- NAS Parallel Benchmarks
- Intensity (data movement) varies with concurrency and problem size
- Large problems (green and red) move more data per thread, and exhaust cache capacity
- Falling Intensity → hit the bandwidth ceiling quickly and degrade

Integration in Intel Advisor
- Roofline has been integrated into Intel's Advisor Performance Tool.
  - Automatically instruments applications (one dot per loop nest/function)
  - Computes FLOPS and AI for each function / loop nest (Integrated Cache Simulator (hierarchical roofline)
  - Automatically benchmarks target system (calculates ceilings)
  - AVX-512 support including vector masks
  - Full integration with existing Advisor capabilities
- Fully supported on NERSC's Edison and Cori (Haswell and Knights Landing Systems)

Community Engagement
- Strong collaboration with NERSC, Intel, and NVIDIA
- We've run Roofline tutorials at SC'17, SC'18, SC'19, ECP'18, ECP'19, ISC'18, ISC'19, NERSC, etc...

Publications
- https://crd.lbl.gov/roofline/publications