A Taxonomy of GPGPU Performance Scaling

Abhinandan Majumdar, Gene Wu, Kapil Dev

Joseph L. Greathouse, Indrani Paul, Wei Huang, Arjun Karthik Venugopal, Leonardo Piga, Chip Freitag, Sooraj Puthoor

Goals

- Observe how GPGPU performance scales at different hardware configurations
- Quantitatively determine principal performance scaling trends across 267 GPGPU kernels from 97 programs profiled on 891 GPU configurations
- Performance studied across 5× change in core frequency, 8.3× change in memory bandwidth, and 11× difference in compute units

Platform

AMD FirePro™ W9100 GPU

- 2,816 processing elements (44 CUs) at 930 MHz
- 16 KB of L1 data cache per CU
- 1 MB of L2 cache shared across all CUs
- 16GB GDDR5 GPU memory at 1.25 GHz
- 320 GB/s memory bandwidth

Experimental Setup

- Core frequency variation: 200 MHz to 1 GHz
- Memory frequency: 150 MHz to 1.25 GHz
- Variation in number of CUs: 4 to 44
- June 20, 2014 beta of AMD FirePro™ drivers
- AMD APP SDK version 2.9
- AMD CodeXL version 1.4

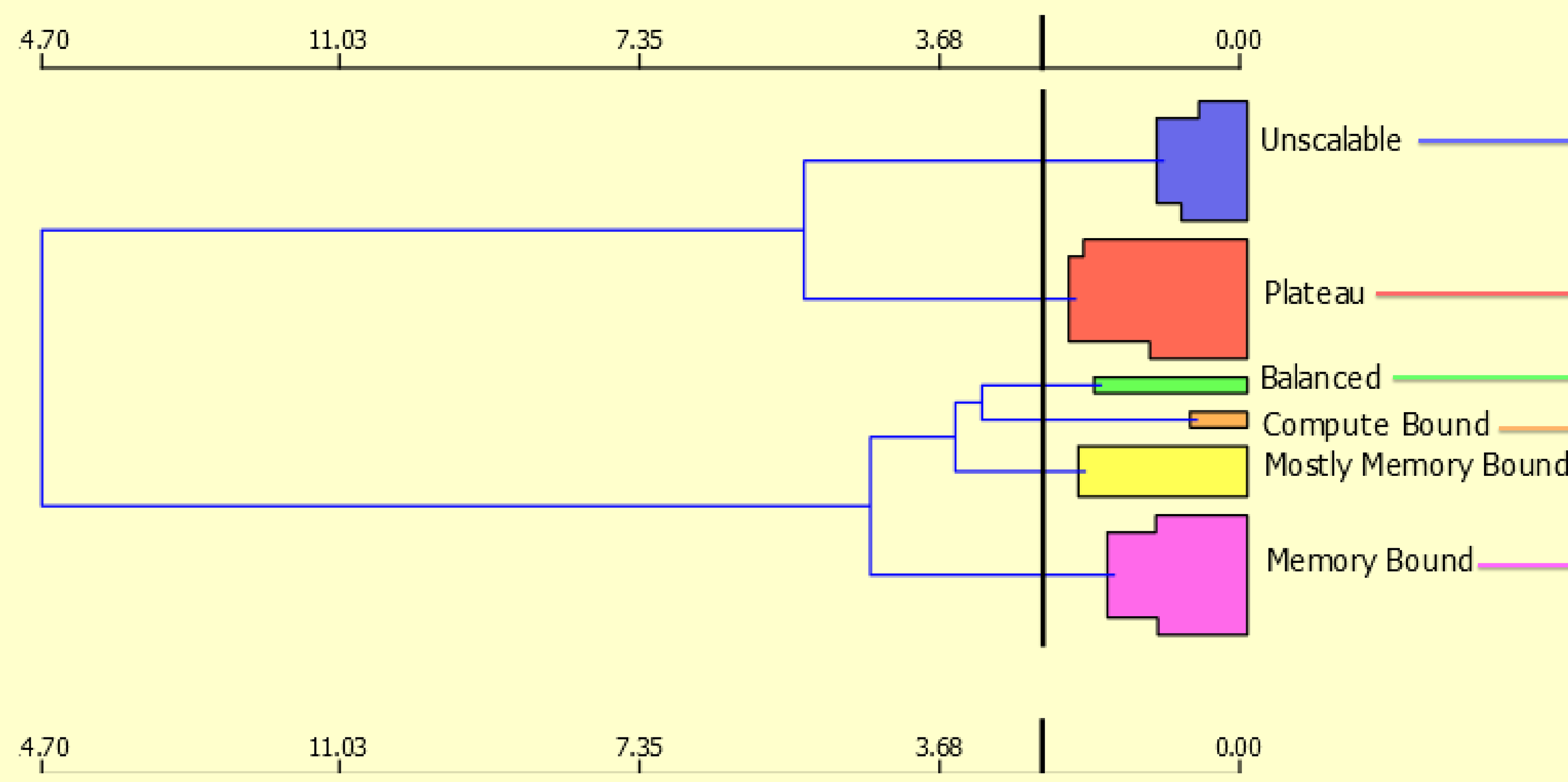
Clustering Methodology

- Quantitatively categorize performance scaling behavior of each kernel sample
- Identify kernels with similar performance scaling
- Principal Component Analysis (PCA) to reduce dimensionality of the data
- Hierarchical agglomerative clustering (HAC) to cluster kernels with similar scaling behavior
- Performance scaling similarity represented by a dendrogram

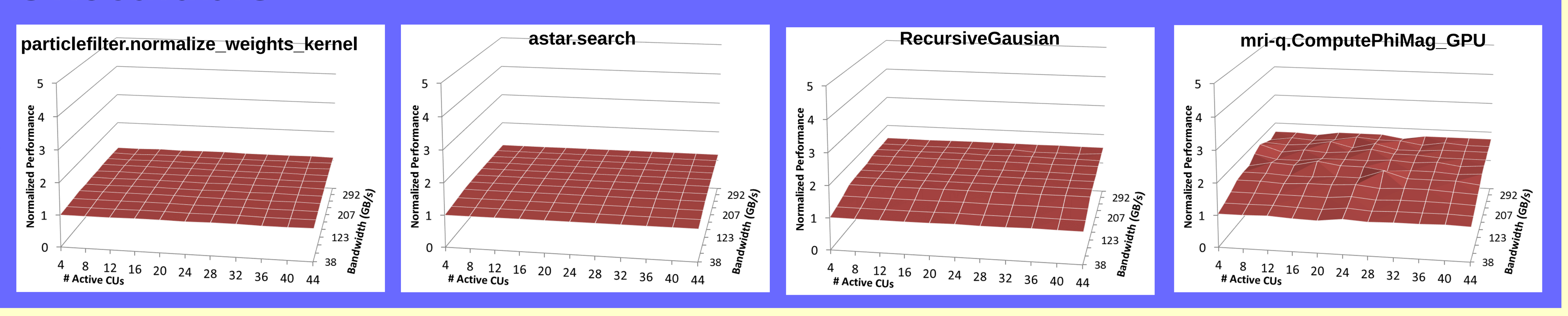
Benchmarks

Benchmark Suite	Benchmarks
Rodinia	backprop, bfs-rodinia, b+tree, cfd, dwt2d, gaussian, heartwall, hotspot, hybridsort, kmeans, lavaMD, leukocyte, lud, myocyte, nn, nw, particlefilter, pathfinder, srad, streamcluster
SHOC	DeviceMemory, MaxFlops, BFS-SHOC, FFT, GEMM, MD, Reduction, Sort, Spmv, Stencil2D, Triad, S3D
Phoronix	juliaGPU, mandelbulbGPU, smallptGPU, MandelGPU
OpenDwarfs	astar, bwa_hmm, crc, csr, nqueens, swat, tdm, gemnoui
Pannotia	bc, color, fw, mis, prk, sssp
AMD APP SDK	NBody, BlackScholes, BinomialOption, BitonicSort, BoxFilter, DCT, DwtHaar1D, EigenValue, FastWalshTransform, FluidSimulation2D, GaussianNoise, HDRToneMapping, Histogram, ImageOverlap, KmeansAutoClustering, Mandelbrot, MatrixMultiplication, MatrixTranspose, MersenneTwister, MonteCarloAsian, PrefixSum, QuasiRandomSequence, RadixSort, RecursiveGaussian, Reduction, ScanLargeArrays, SimpleConvolution, SobelFilter, UnsharpMask, URNG
Parboil	pb-bfs, stencil, mri-gridding, lbm, sad, histo, mri-q, cutcp, pb-sgemm, pb-spmv, tpacf
Exascale Proxy Applications	CoMD, CoMD-LJ, lulesh, miniFE, XSBench
Other applications	BPT, graph500

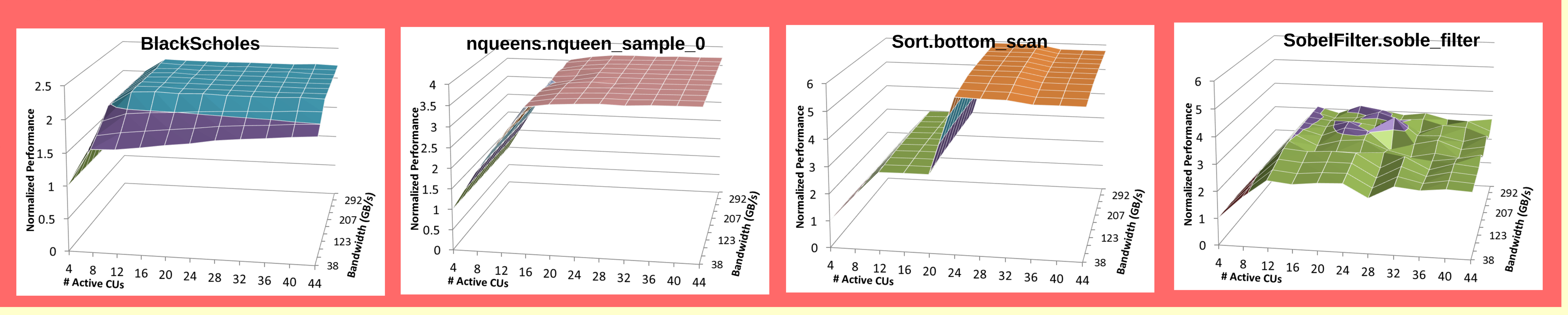
Performance Scaling Trends



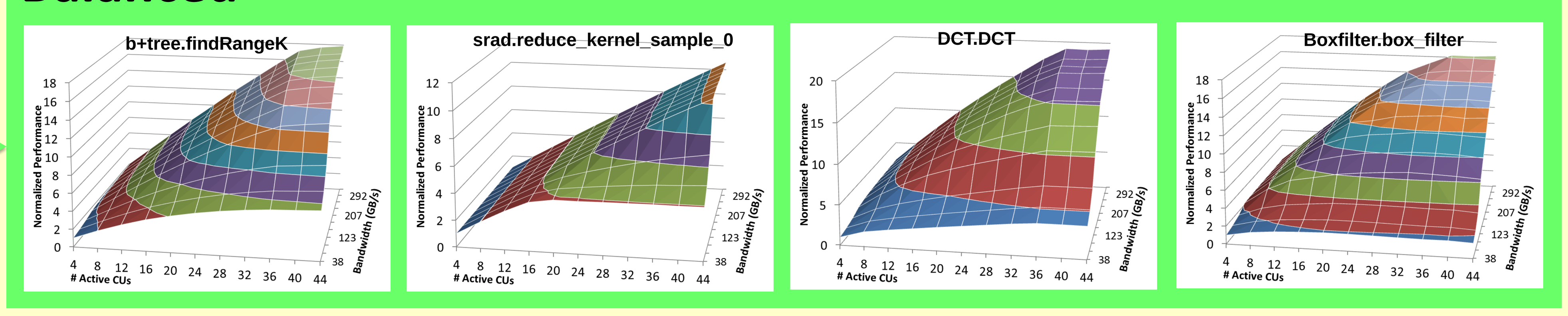
Unscalable



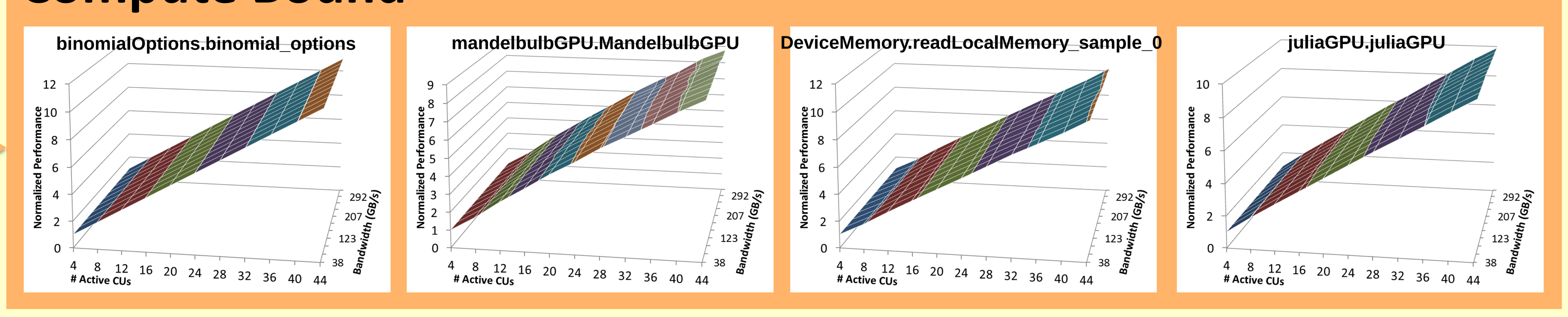
Plateau



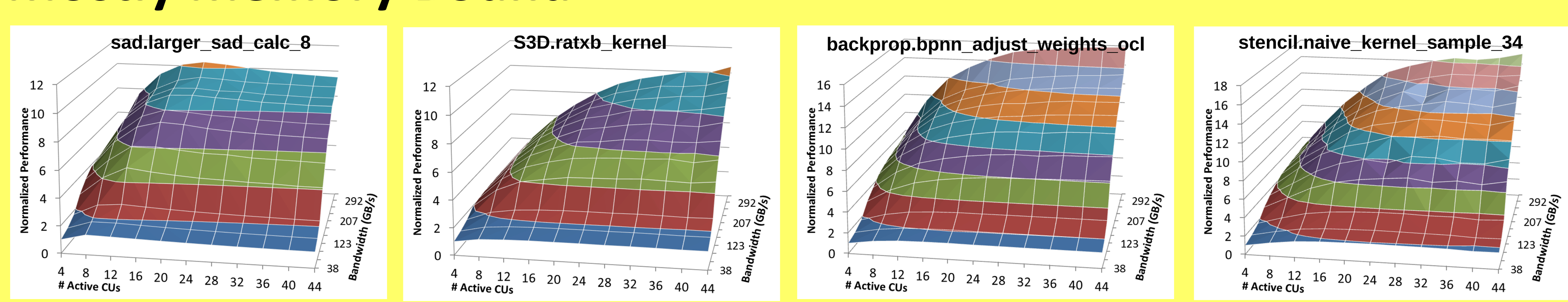
Balanced



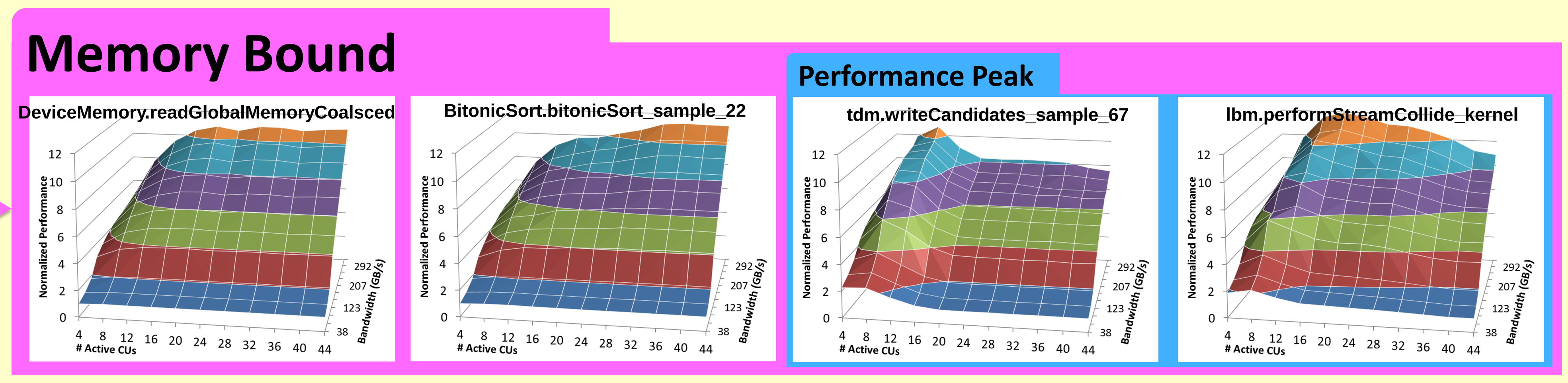
Compute Bound



Mostly Memory Bound



Memory Bound



Conclusion and Future Work

- The performance of many kernels scales as more CUs are added
- Around 40% of the kernel iterations do not scale to modern GPU sizes
- Future studies should examine whether existing benchmarks (and input sets) are representative GPGPU workloads
- Future work could also consider other hardware configurations, such as cache sizes or double precision rate, and could characterize power and energy

Cornell University TEXAS The University of Texas at Austin BROWN

AMD