Making the Most of your Hardware

Micro-Benchmarks on NVIDIA GPUs

Ian Lee, UConn
Jose Renau, UCSC
Javi Mahai, UCSC

CUDA

- Computer Unified Device Architecture
- Extension of C
- Programming on Graphical Processing Units (GPUs)
- Thousands of computations in parallel
- More than graphical computations

Passing the Torch

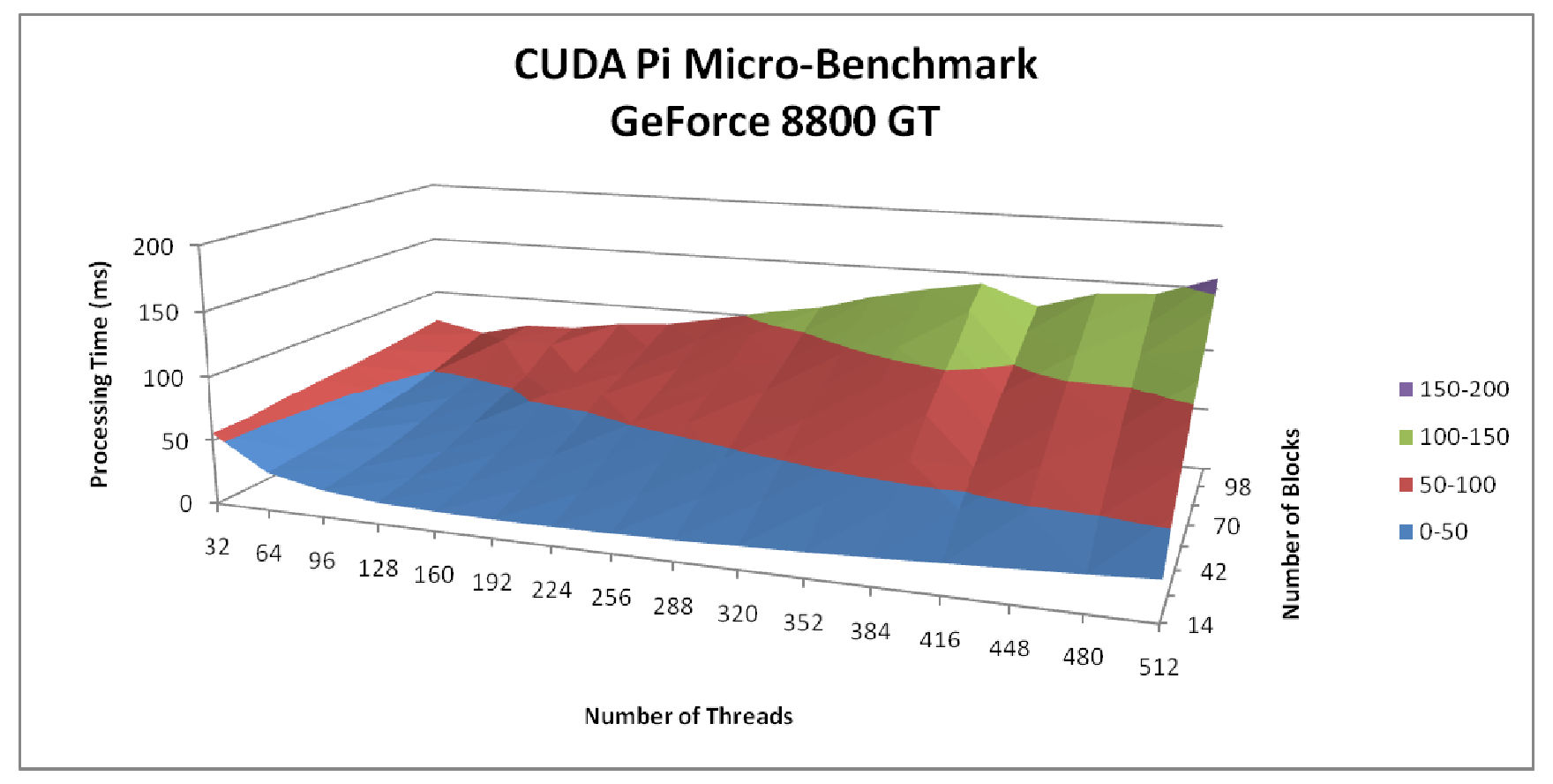
- Today, CPUs are the primary computational units
- Single Instruction Multiple Data (SIMD) ideal for parallelization
  - Graphical Rendering
  - Biological Computation
  - Genetics
  - Database Operations
- Push towards parallelized systems and algorithms

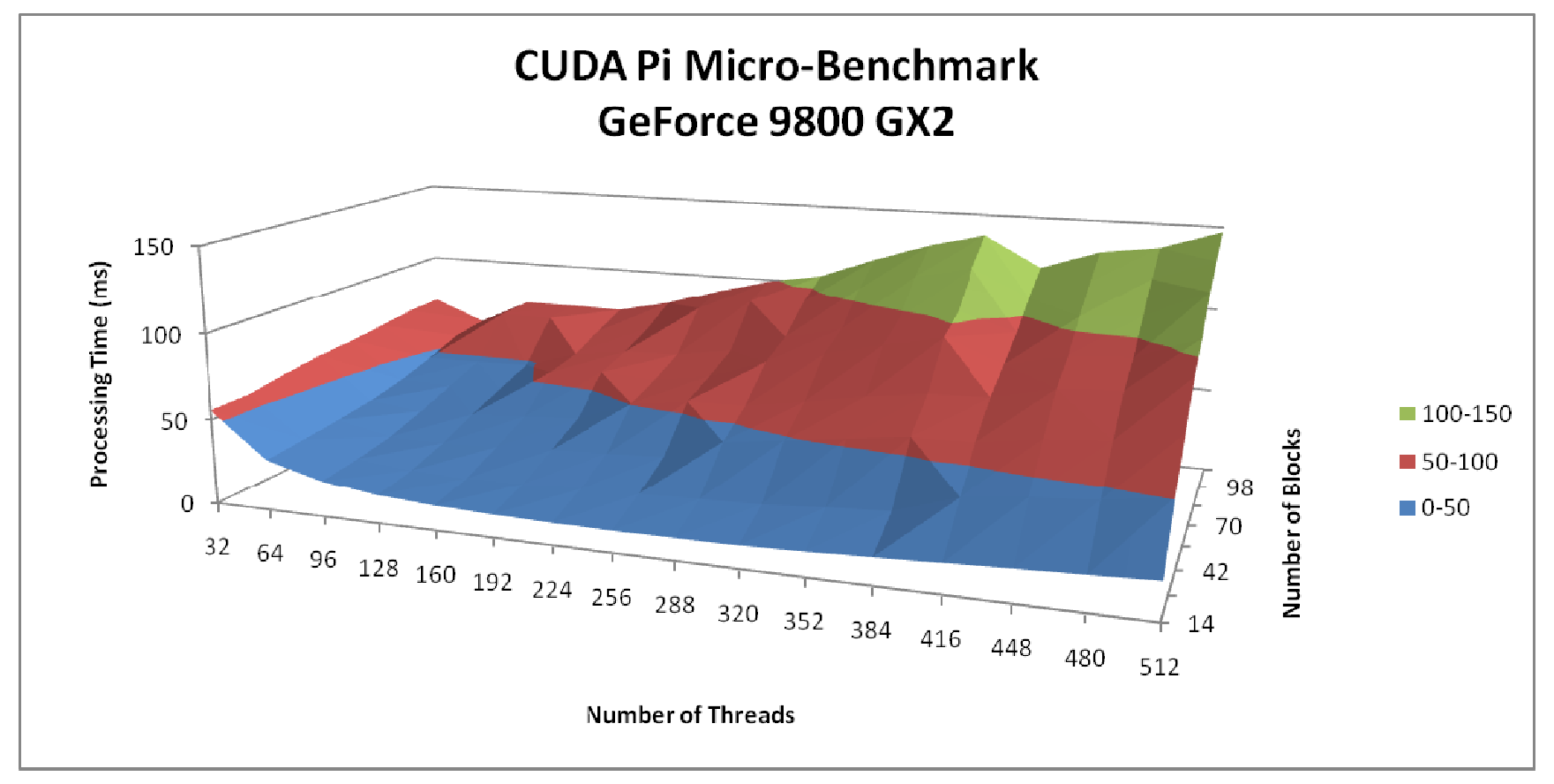
Future Work

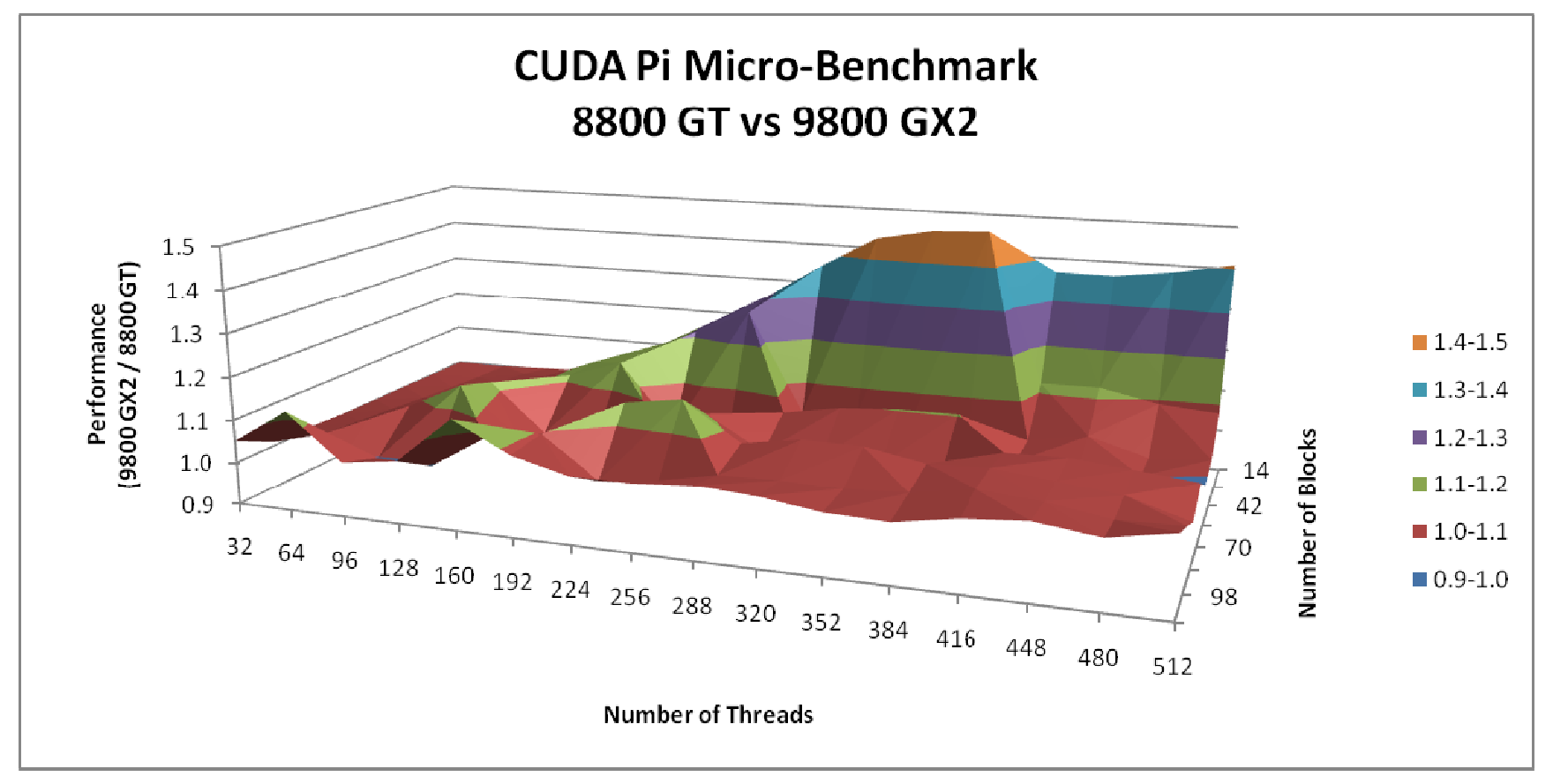
- Creation of a full Benchmark Suite
- Performance Analysis on various GPUs
- Further parallelization of Micro-Benchmark Kernels
- Thermal Imaging GPUs under different loads
- Performance Ranges within Benchmarks

Digits of Pi

- Computation of Pi has been explored for thousands of years
- Many algorithms and methods of computing
- Gregory-Leibniz Series computes whole value of Pi, rather than digits
- CPU takes ~1 second perform same calculation

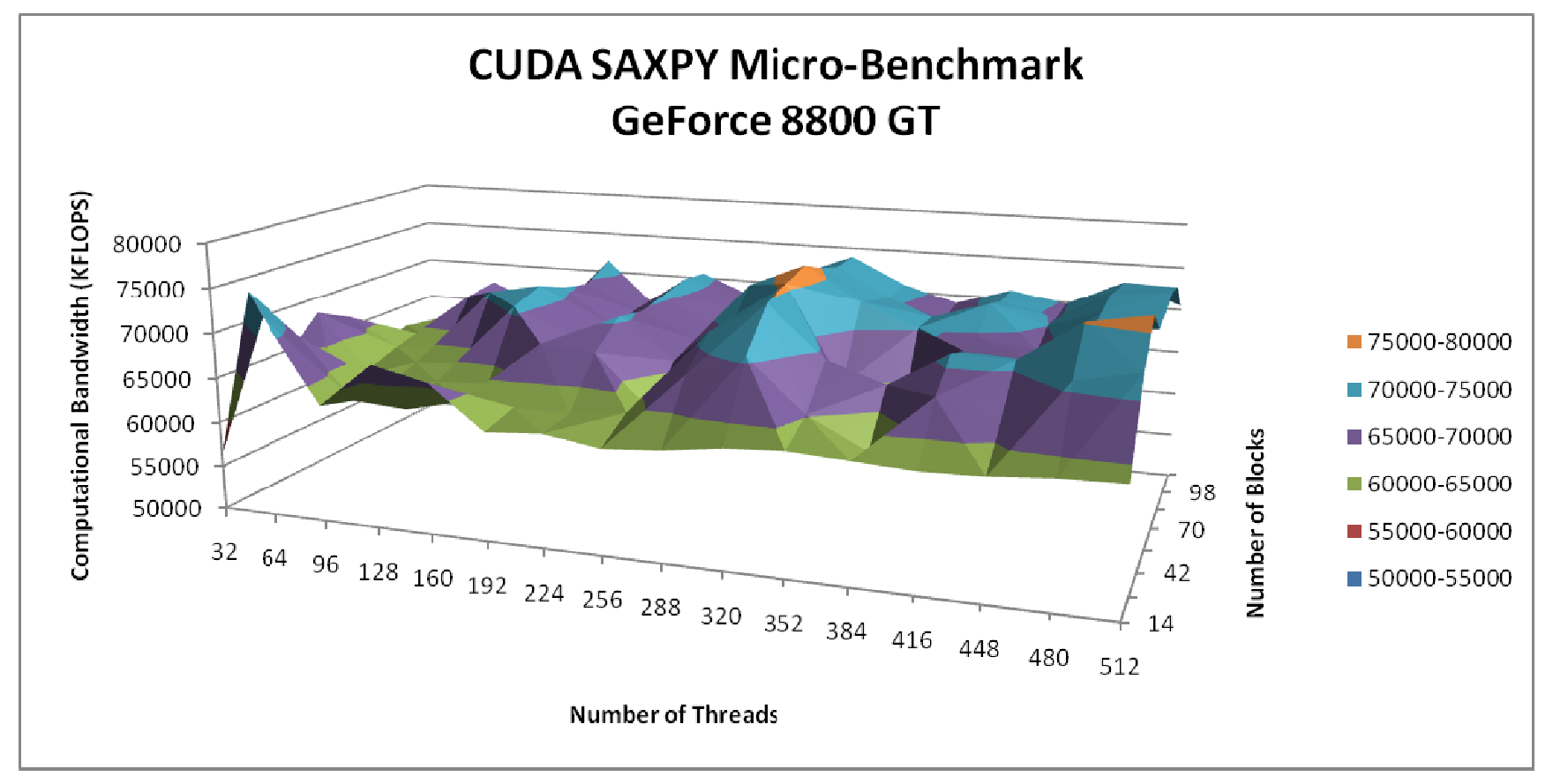


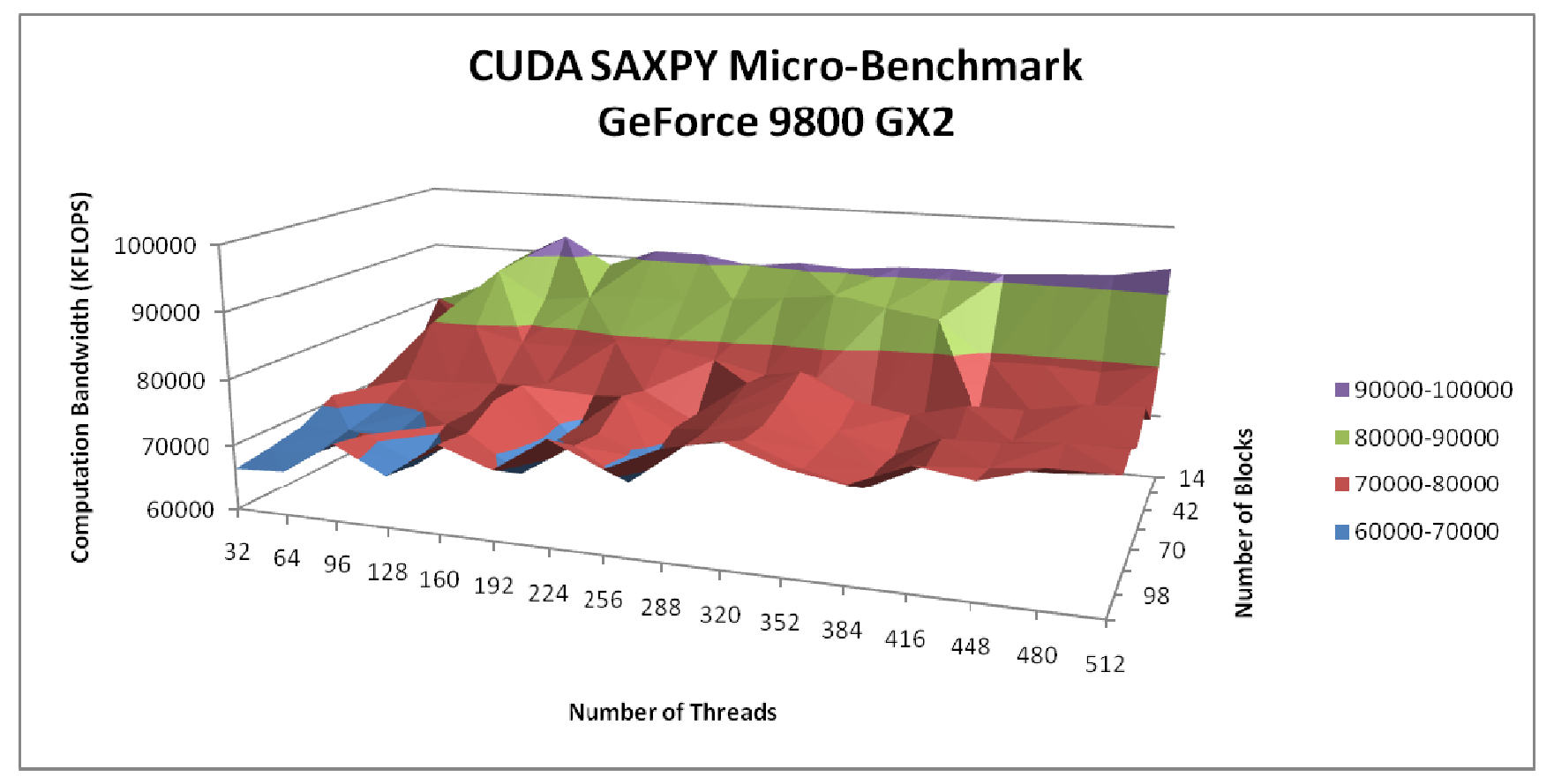


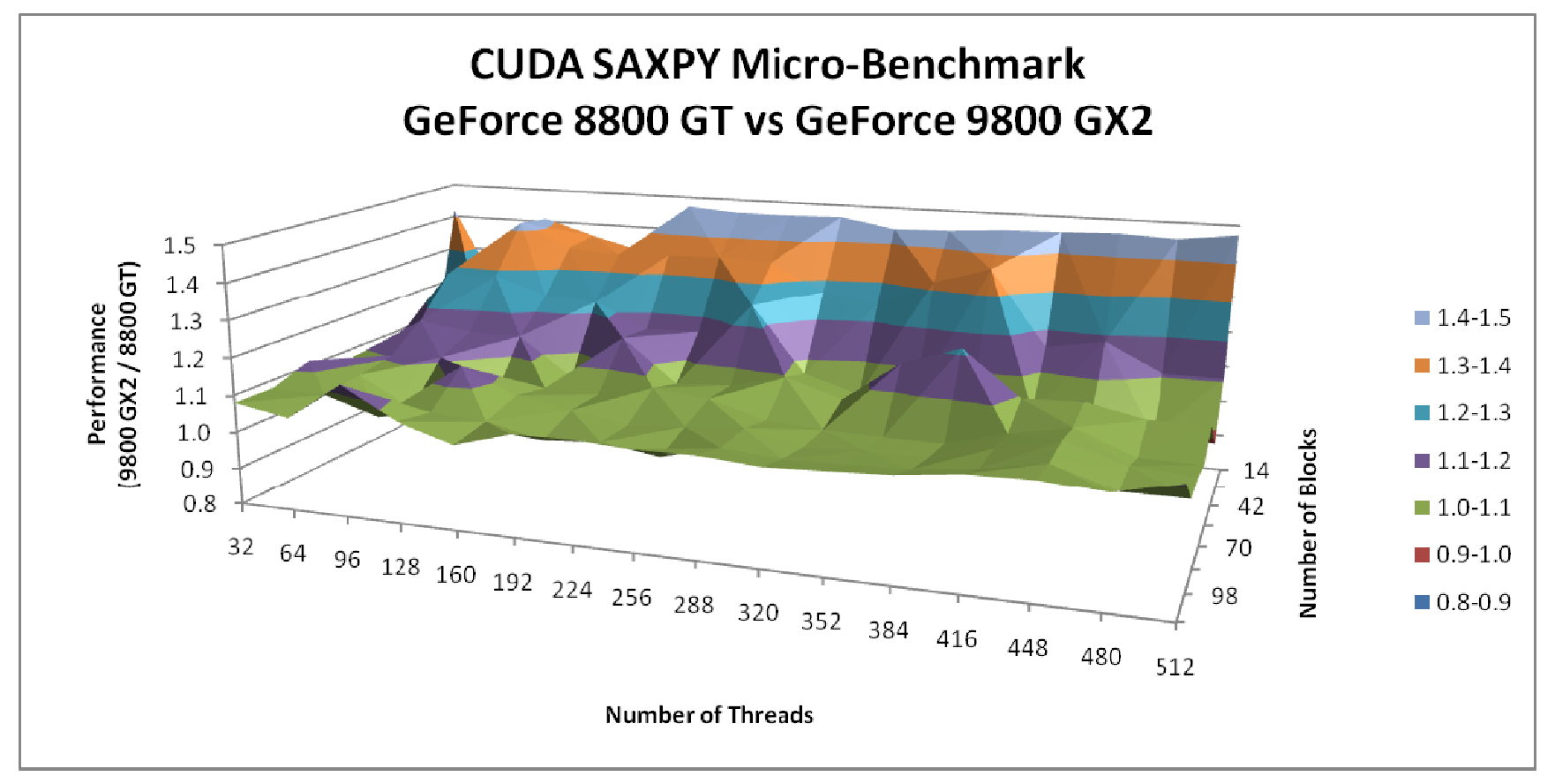


Scalar Multiplication of Vectors

- Current CPU Benchmarks adaptable to GPU
- Scalar Alpha X Plus Y (SAXPY) of Single Precision Floats
  - α X + Y
- Pseudo Randomly generated vectors X and Y, as well as scalar α
- Comparisons between GPUs

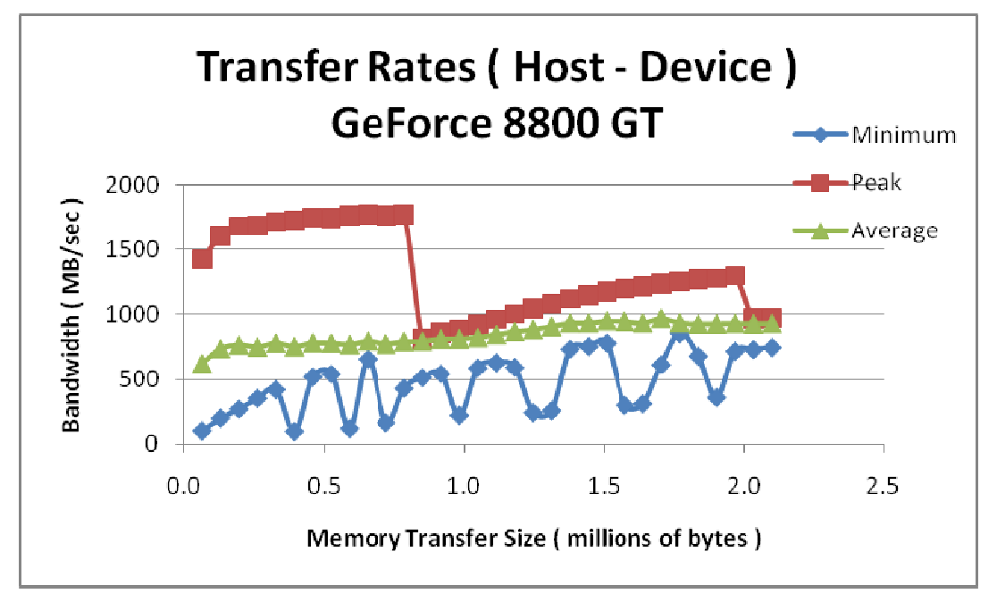


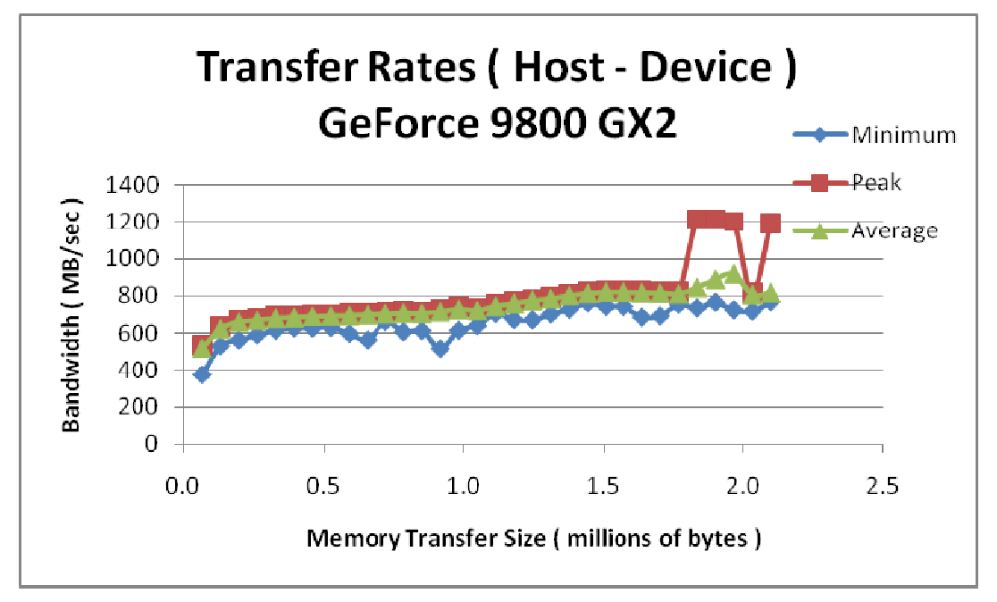


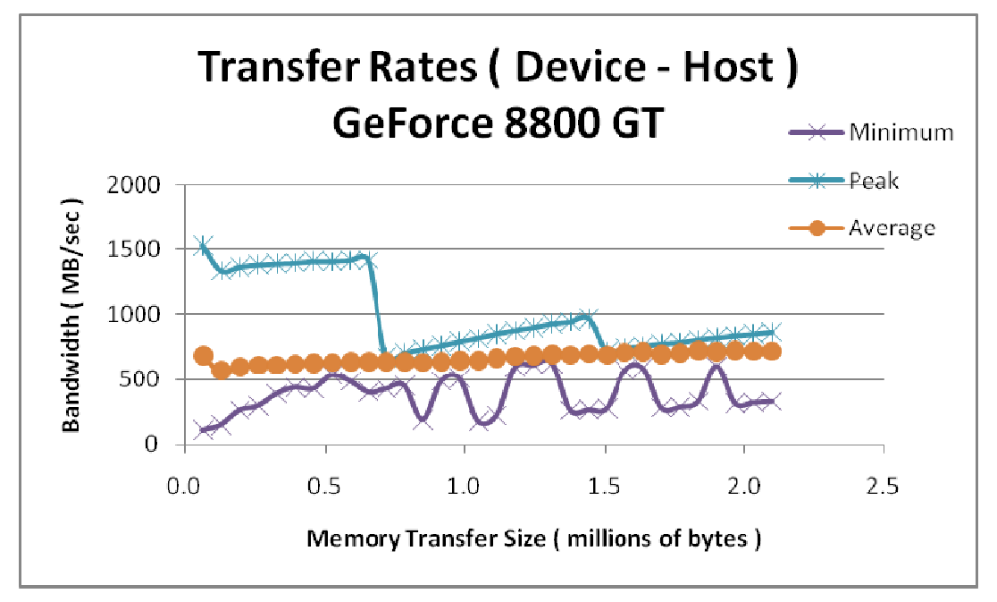


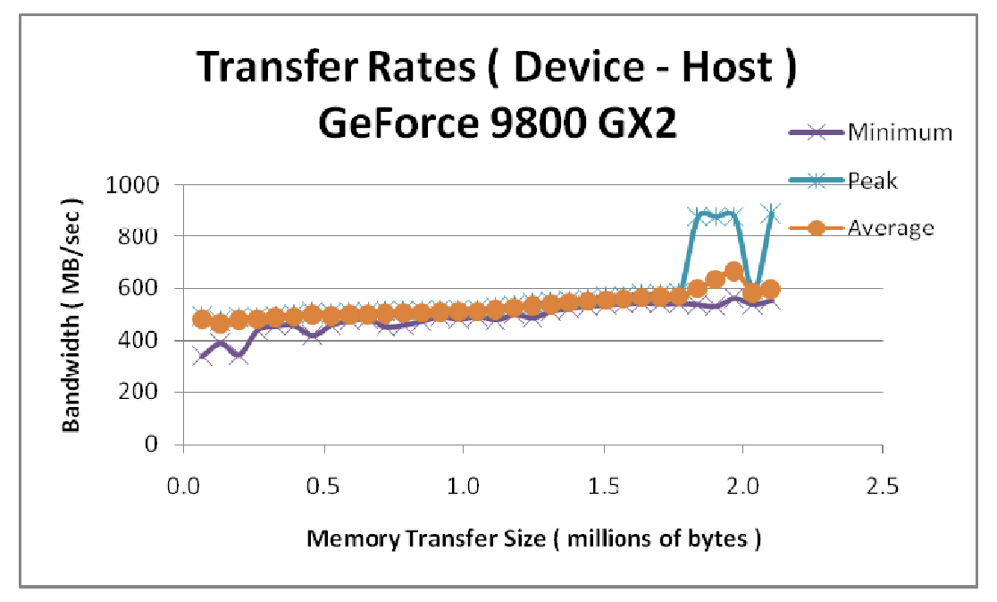
Bus Bandwidth

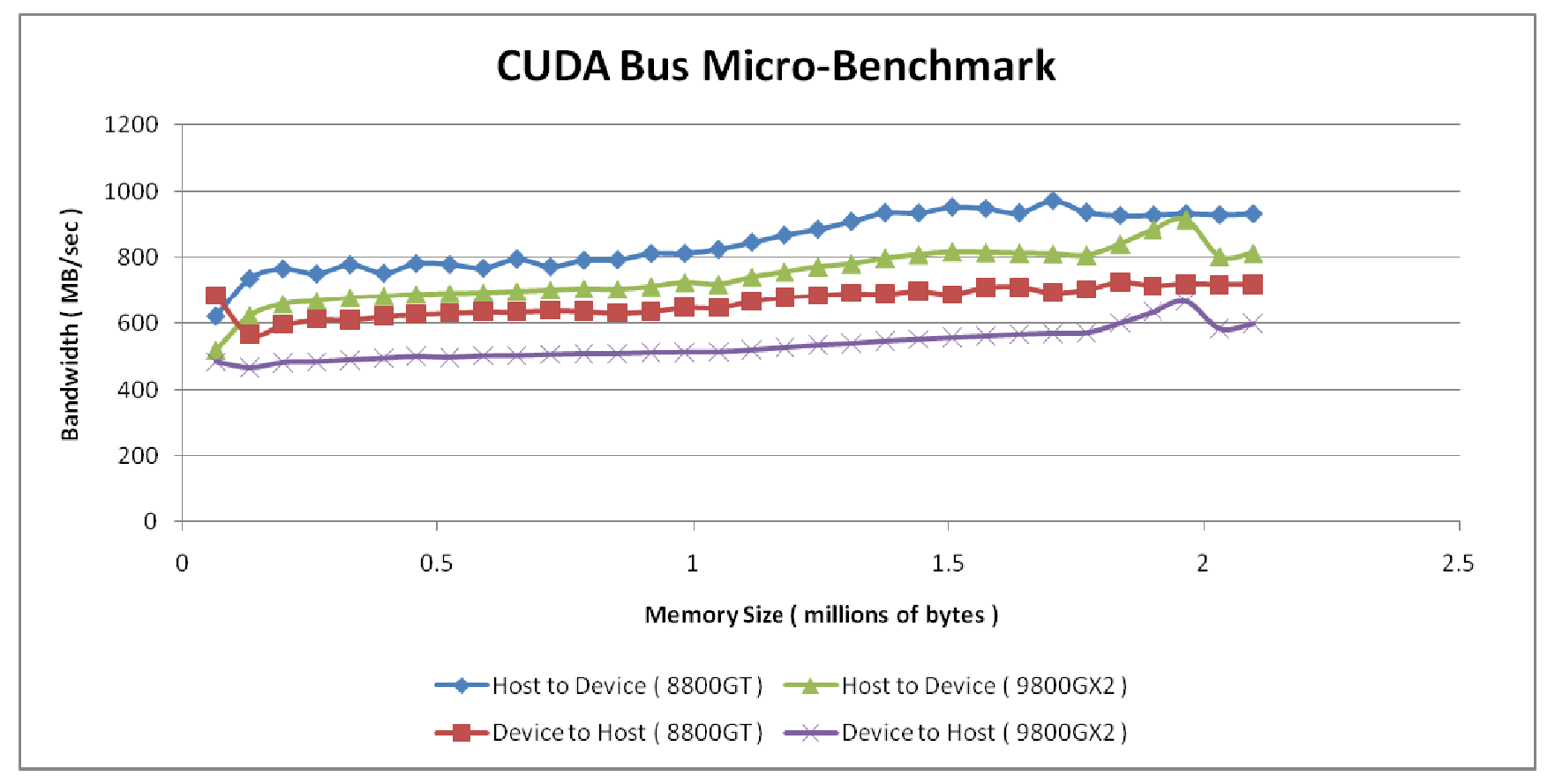
- Computation on the GPU requires data to be transferred to the device
- Limited by hardware transfer rates
- Average Bandwidth varies by transfer direction
- Outlying cases vary by model of GPU





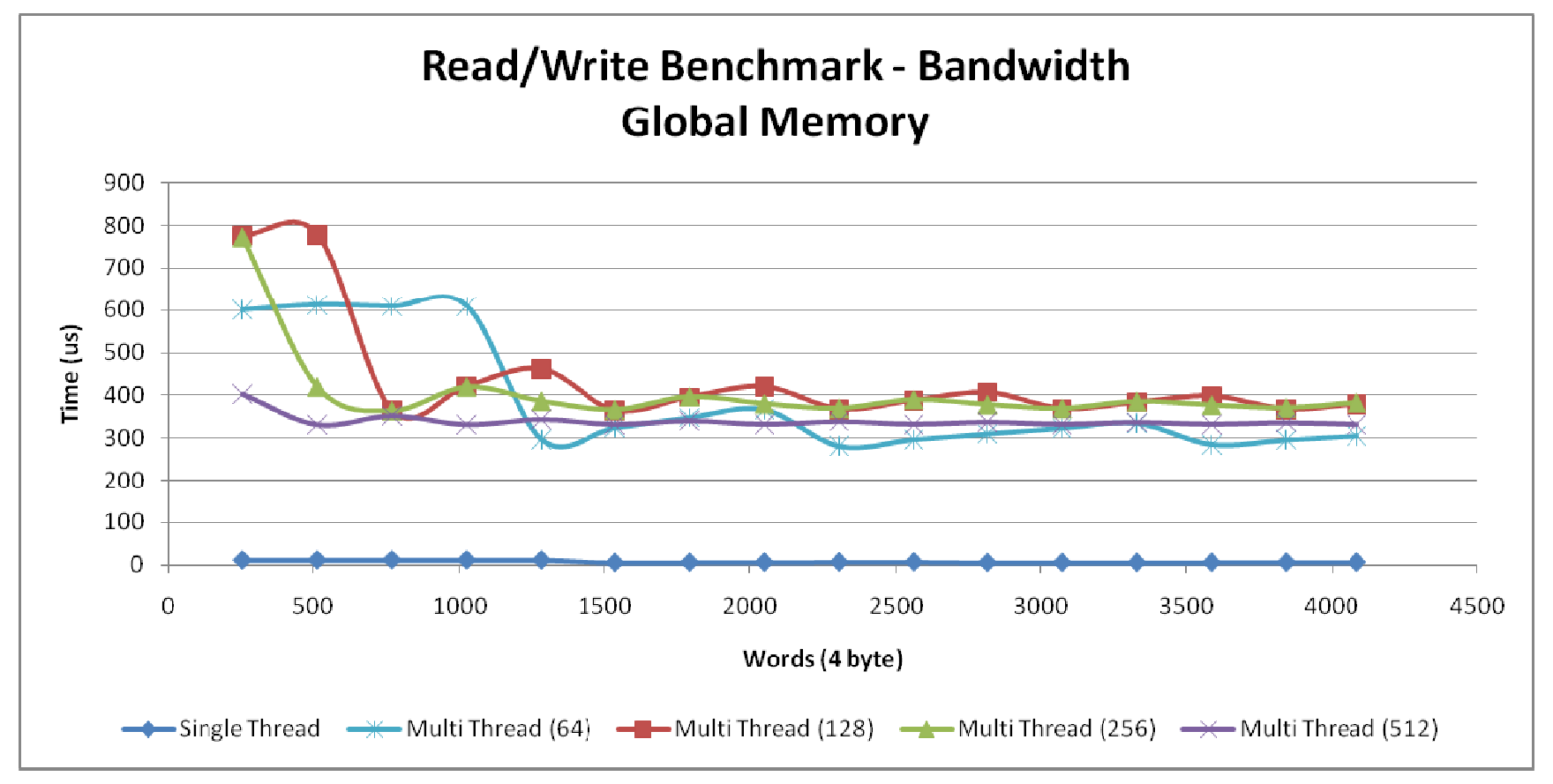


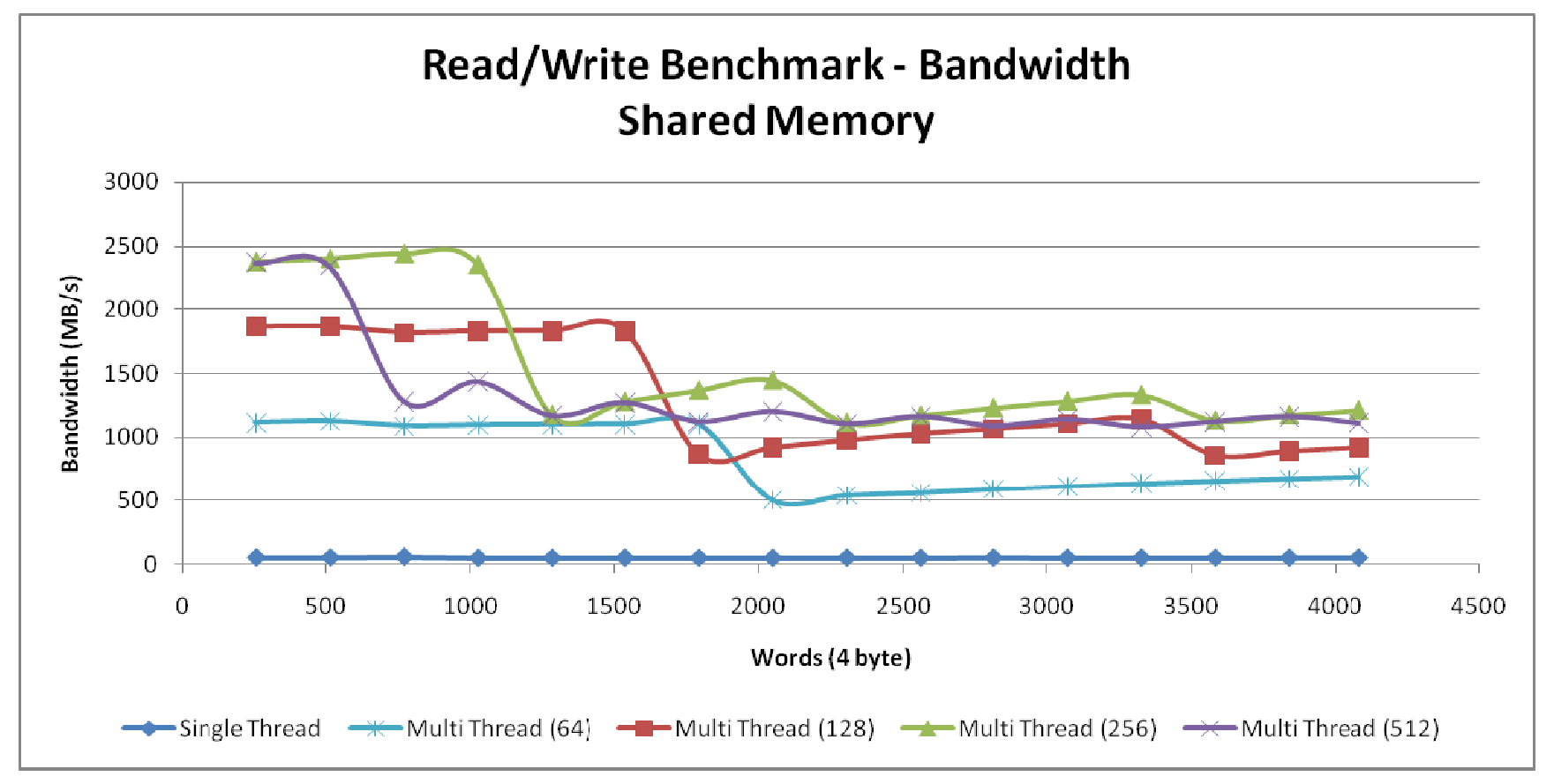




GPU Memory Access

- Read, Write, and Copy are most common
- GPU Architecture modeled after CPU
- Global Memory
  - Like Primary Memory in CPU Architecture
  - 256 – 1024 MB on Modern GPUs
- Shared Memory
  - Similar to Cache of CPUs
  - Typically 16 KB per Streaming Multi-Processor (SM)
  - 112 – 192 SMs per GPU


Read/Write Benchmark - Bandwidth
Global Memory
Time (us)
Words (4 byte)
Single Thread
Multi Thread (64)
Multi Thread (128)
Multi Thread (256)
Multi Thread (512)


Read/Write Benchmark - Bandwidth
Shared Memory
Bandwidth (MB/s)
Words (4 byte)
Single Thread
Multi Thread (64)
Multi Thread (128)
Multi Thread (256)
Multi Thread (512)