The Scalable HeterOgeneous Computing (SHOC) Benchmark Suite

A. Danalis C. McCurdy J. Meredith P. Roth K. Spafford V. Tipparaju J. Vetter L. Wesolowski
Future Technologies Group, Oak Ridge National Lab | http://ft.ornl.gov/doku/shoc/start

What is SHOC?

SHOC is a benchmark suite for heterogeneous systems focused on scientific computing workloads including common kernels like GEMM, FFT, and Stencil computations.

It's implemented in CUDA and OpenCL for a 1:1 comparison.

It's distributed with MPI, so you can test multiple GPUs and GPU-accelerated clusters.

It's open source, and version 1.0 will release on October 1. A beta version is available for download now.

- Level 0 – "Feeds and Speeds"
  - PCIe Bandwidth
  - Device Memory Bandwidth - Global, Shared, and Texture Memory
  - Sustained FLOPS
  - OpenCL Kernel Compilation
  - OpenCL Queuing Delay
- Level 1 – Basic Algorithms and Parallel Primitives
  - **F**ast **F**ourier **T**ransform
  - MD – Computation of Lennard-Jones Potential Using a Neighbor-List Algorithm
  - Reduction
  - GEMM
  - Scan – a.k.a. Parallel Prefix Sum
  - Sort
  - SpMV – Sparse Matrix-Vector Multiply
  - Stencil2D
  - Triad
- Level 2 – Real Application Kernels
  - S3D – Chemical rates computation used in the simulation of combustion

Stability Test

SHOC has an FFT-based stress test based on Prime95's famous "Torture Test."

This test is designed to stress GPU hardware and identify any errors due to insufficient cooling, bad memory, or other hardware problems.

The test alternates between forward and inverse transforms, and performs the correctness check on the GPU. This in-place computation keeps the GPU hot and is very sensitive to any innaccuracy introduced by hardware errors.

SpMV

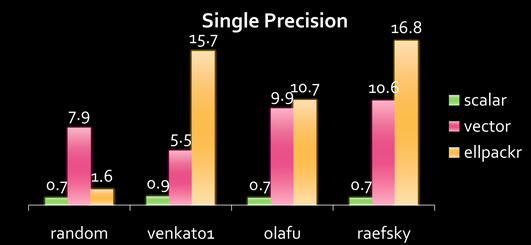
There are a couple of good algorithms out there for sparse matrix-vector multiply, but there's no clear winner—the best algorithm depends on the structure of the data.

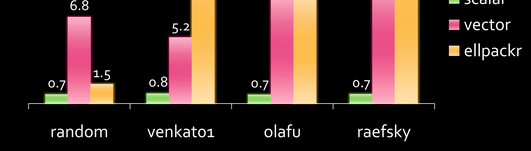
SHOC contains implementations of SpMV based on Baskaran ('09), Bell ('09), and Vasquez ('09).

Run SpMV on your matrix to decide which algorithmic strategy is best for you code.

Approaches:
Scalar – One thread per row
Vector – One warp per row
ELLPack-R – Use an alternative data structure

Results in GFLOPS/s from Tesla C2050, CUDA 3.1, ECC On

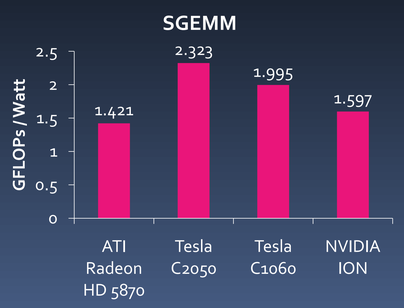


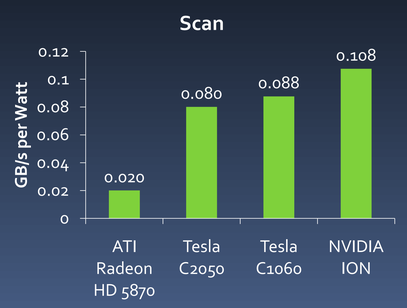


Power Efficiency

We also use SHOC to evaluate the energy efficiency of GPUs. The following graphs show energy efficiency across several devices using the OpenCL version of the benchmarks.

These results were measured using NVIDIA's GPU Computing SDK 3.0 and AMD's Stream Computing SDK v2.0. Power measurements are the manufacturer's thermal design point (TDP).





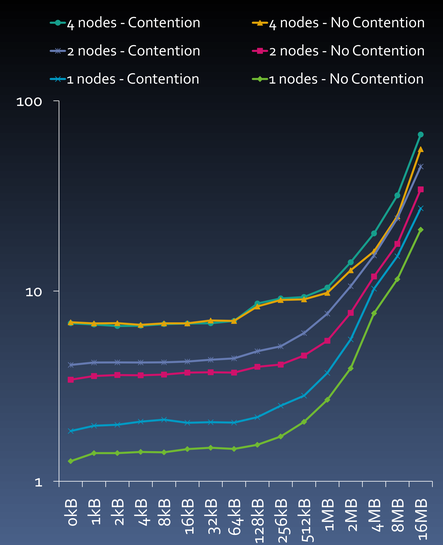
Resource Contention

GPUs and HCAs typically connect to the host processor via the PCIe bus. Sometimes, if a program uses multiple GPUs and the HCA at the same time, there will be contention on the PCIe bus.

SHOC includes a test that measures the impact on MPI message latency when GPUs and HCAs are used at the same time.

In this test, several MPI tasks are started on the host. To get a baseline, they first measure message latency without using the GPUs. Then, they start transferring data over the PCIe bus and sending MPI messages simultaneously.

The chart at the right shows the difference from sequential (no contention) and simultaneous (contention) PCIe access on the Lens cluster at ORNL.



GPU and API Comparison

SHOC benchmarks span a variety of computational and memory access patterns, and are useful for the common task of comparing the performance of varying GPU devices. Scan and FFT results are shown on the right, in units of GB/sec and GFLOPS, respectively.

As all benchmarks have CUDA and OpenCL versions, you can also compare the performance of these two APIs. The chart below shows this comparison on the Tesla C1060 using NVIDIA GPU Computing SDK 3.0. S3D and SGEMM results are measured in GFLOPS, and Scan and Reduction in GB/sec.