WHAT’S NEW IN CUDA 8

Siddharth Sharma, Oct 2016
WHAT’S NEW IN CUDA 8
Why Should You Care

>2X

Run Computations Faster*
Solve Larger Problems**
Critical Path Analysis

* HOOMD Blue v1.3.3 Lennard-Jones liquid benchmark
  - K80 and P100 (PCIe); Base clocks; 4 GPUs per PCIe root complex
  - 2x K80 indicates 2-GPU configuration (or 1x K80 board)
  - CUDA 8 GA with r361.79 (K80) and r361.93.02 (P100)
  - Host System: Intel Xeon Broadwell dual socket 22-core E5-2699 v4@2.2GHz 3.6GHz Turbo with CentOS 7.2 x86-64 and 256GB memory

** HPGMG-FV benchmark
  - K80 and P100 (PCIe); Base clocks; 4 GPUs per PCIe root complex
  - 2x K80 indicates 2-GPU configuration (or 1x K80 board)
  - CUDA 8 GA with r361.79 (K80) and r361.93.02 (P100)
  - Host System: Intel Xeon Broadwell dual socket 22-core E5-2699 v4@2.2GHz 3.6GHz Turbo with CentOS 7.2 x86-64 and 256GB memory
WHAT’S NEW IN CUDA 8

PASCAL ARCHITECTURE
- NVLINK
- HBM2 Stacked Memory
- Page Migration Engine

LIBRARIES
- New nvGRAPH library
- Support for FP16, INT8

UNIFIED MEMORY
- Demand Paging
- New Tuning APIs
- Data Coherence & Atomics

DEVELOPER TOOLS
- Critical Path Analysis
- NVCC Compile Time
- OpenACC Profiling
PASCAL ARCHITECTURE
PASCAL ARCHITECTURE

NVLink
GPU Interconnect for Maximum Scalability

HBM2 Stacked Memory
Unifying Compute & Memory in Single Package

Page Migration Engine
Simple Parallel Programming with 512 TB of Virtual Memory

Webinar: “Inside Pascal”
Mark Harris (NVIDIA), Lars Nyland (NVIDIA)
GPU Technical Conference 2016 - ID S6176
CUDA 8 ON P100: >3X FASTER THAN CPUs

Performance may vary based on OS and software versions, and motherboard configuration.

- K80 and P100 (PCIe); Base clocks; 4 GPUs per PCIe root complex
- CUDA 8 GA with r361.79 (K80) and r361.93.02 (P100)
- Host System: Intel Xeon Broadwell dual socket 22-core E5-2699 v4@2.2GHz 3.6GHz Turbo with CentOS 7.2 x86-64 and 256GB memory
UNIFIED MEMORY
UNIFIED MEMORY
Implicit Memory Management

Past Developer View

Starting with Kepler and CUDA 6

System Memory

CPU

Pascal GPU

GPU Memory

Unified Memory
APPLICATIONS: LARGE VARIATIONS IN DATASET SIZES

Graph Analysis
Larger datasets

Combustion
More species & improved accuracy

Quantum Chemistry
Larger systems

Ray-tracing
Larger scenes to render
CUDA 8: PASCAL UNIFIED MEMORY
Easier Memory Management, APIs for High Performance

CUDA 8

EASIER MEMORY MANAGEMENT, APIs FOR HIGH PERFORMANCE

- Oversubscribe GPU memory
- Allocate up to system memory size

SIMPLER DATA ACCESS
- CPU/GPU Data coherence
- Unified memory atomic operations

TUNE UNIFIED MEMORY PERFORMANCE
- APIs for Pre-fetching & Read duplication
- Usage hints via cudaMemAdvise API
CUDA 8 UNIFIED MEMORY — EXAMPLE

Allocating 4x more than P100 physical memory

```c
void foo() {
    // Allocate 64 GB
    char *data;
    size_t size = 64*1024*1024*1024;
    cudaMallocManaged(&data, size);
}
```

64 GB unified memory allocation on P100 with 16 GB physical memory

Transparent - No API changes

Works on Pascal & future architectures
CUDA 8 UNIFIED MEMORY — EXAMPLE

Accessing data simultaneously by CPU and GPU codes

```c
__global__ void mykernel(char *data) {
  data[1] = 'g';
}

void foo() {
  char *data;
  cudaMallocManaged(&data, 2);
  mykernel<<<...>>>(data);
  // no synchronize here
  data[0] = 'c';
  cudaFree(data);
}
```

Both CPU code and CUDA kernel accessing ‘data’ simultaneously

Possible with CUDA 8 unified memory on Pascal

Webinar: “CUDA 8 Unified Memory”
>3X SPEEDUP WITH UNIFIED MEMORY

Performance may vary based on OS and software versions, and motherboard configuration.

- HPGMG AMR on 1xK40, 1xP100 (PCIe) with CUDA 8 (r361)
- CPU measurements with Intel Xeon Haswell dual socket 10-core E5-2650 v3@2.3 GHz 3.0 GHz Turbo, HT on
- Host System: Intel Xeon Haswell dual socket 16-cores E5-2630 v3@2.4GHz 3.2GHz Turbo
LIBRARIES
... and much more: Parallel Computing, Recommender Systems, Fraud Detection, Voice Recognition, Text Understanding, Search
nvGRAPH
GPU Accelerated Graph Analytics

Parallel Library for Interactive and High Throughput Graph Analytics

Solve graphs with up to 2.5 Billion edges on a single GPU (Tesla M40)

Includes — PageRank, Single Source Shortest Path and Single Source Widest Path algorithms

Semi-ring SPMV operations provides building blocks for graph traversal algorithms

<table>
<thead>
<tr>
<th>PageRank</th>
<th>Single Source Shortest Path</th>
<th>Single Source Widest Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>Search</td>
<td>Robotic Path Planning</td>
<td>IP Routing</td>
</tr>
<tr>
<td>Recommendation Engines</td>
<td>Power Network Planning</td>
<td>Chip Design / EDA</td>
</tr>
<tr>
<td>Social Ad Placement</td>
<td>Logistics &amp; Supply Chain Planning</td>
<td>Traffic sensitive routing</td>
</tr>
</tbody>
</table>
> 200X SPEEDUP ON PAGERANK VS GALOIS

Performance may vary based on OS and software versions, and motherboard configuration.

- nvGRAPH on M40 (ECC ON, r352), P100 (r361), Base clocks, input and output data on device
- GraphMat, Galois (v2.3) on Intel Xeon Broadwell dual-socket 22-core/socket E5-2699 v4 @ 2.22GHz, 3.6GHz Turbo
- Comparing Average Time per Iteration (ms) for PageRank
- Host System: Intel Xeon Haswell single-socket 16-core E5-2698 v3 @ 2.3GHz, 3.6GHz Turbo
- CentOS 7.2 x86-64 with 128GB System Memory
HIGHER THROUGHPUT THROUGH LOWER PRECISION COMPUTATION

Deep Learning

cuBLAS: FP16 and INT8 GEMMS

Radio Astronomy

cuFFT: native FP16 operations

Fluid Dynamics

cuSPARSE: FP16 CSRMM
DEVELOPER TOOLS
DEPENDENCY ANALYSIS

Easily find the critical kernel to optimize

The longest running kernel is not always the most critical optimization target
IDENTIFY BOTTLENECKS ON CRITICAL PATH

Visual Profiler and NVPROF

Unguided Analysis

Dependency Analysis

Functions on critical path
IDENTIFY BOTTLENECKS ON CRITICAL PATH

Visual profiler and NVPROF

Launch copy_kernel

Memcpy HtoD [sync]

MemcpyDtoH [sync]

Inbound dependencies

Outbound dependencies
OpenACC PROFILING

OpenACC->Driver API->Compute correlation

OpenACC timeline

OpenACC->Source Code correlation

OpenACC Properties
PROFILE CPU CODE + GPU CODE IN VISUAL PROFILER

Profile execution times on host function calls

View CPU code function hierarchy
PROFILE UNIFIED MEMORY

MONITOR NVLINK BANDWIDTH

Webinar: “CUDA 8 Tools Webinar”
COMPILE NVCC 2X FASTER
Improved Developer Productivity

Average total compile times (per translation unit)
- Host system: Intel Core i7-3930K 6-cores @ 3.2GHz
- CentOS x86_64 Linux release 7.1.1503 (Core) with GCC 4.8.3 20140911
- GPU target architecture sm_52

Webinar: “CUDA 8 Performance Report”
# NEW PLATFORMS SUPPORTED

<table>
<thead>
<tr>
<th>Platform</th>
<th>Operating Systems</th>
<th>Compilers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linux</td>
<td>Fedora 23, Ubuntu 16.04, SLES 1</td>
<td>PGI C++ 16.1/16.4, Clang 3.7, ICC 16.0</td>
</tr>
<tr>
<td>MAC</td>
<td>OS X 10.12</td>
<td>GCC 5.x</td>
</tr>
</tbody>
</table>
WHAT’S NEW IN CUDA 8

PASCAL ARCHITECTURE

- NVLINK
- HBM2 Stacked Memory
- Page Migration Engine

LIBRARIES

- New nvGRAPH library
- Support for FP16, INT8

UNIFIED MEMORY

- Demand Paging
- New Tuning APIs
- Data Coherence & Atomics

DEVELOPER TOOLS

- Critical Path Analysis
- NVCC Compile Time
- OpenACC Profiling
CUDA 8 - DOWNLOAD TODAY!
Everything You Need to Accelerate Applications

- CUDA Applications in your Industry: [www.nvidia.com/object/gpu-applications-domain.htm](www.nvidia.com/object/gpu-applications-domain.htm)

- Additional Webinars:
  - Inside PASCAL
  - CUDA 8 Performance Report
  - CUDA 8 Tools
  - CUDA 8 Unified Memory


developer.nvidia.com/cuda-toolkit
THANK YOU