XLC/C++ and GPU Programming on Power Systems

Kelvin Li, Kit Barton, John Keenleyside – IBM
{kli, kbarton, keenley}@ca.ibm.com
John Ashley – NVIDIA
jashley@nvidia.com
Agenda

- Overview of POWER8
- Overview of GPU+POWER8
- Overview of CUDA Programming Model
- CUDA and XL Integration
- Technology Exploration
OpenPOWER, a catalyst for Open Innovation

Market Shifts

• Moore’s law no longer satisfies performance gain
• Growing workload demands
• Numerous IT consumption models
• Mature Open software ecosystem

Open Development
open software, open hardware

Collaboration of thought leaders
simultaneous innovation, multiple disciplines

Performance of POWER architecture
amplified capability

New Open Innovation

• Rich software ecosystem
• Spectrum of power servers
• Multiple hardware options
• Derivative POWER chips

The OpenPOWER Foundation is an open ecosystem, using the POWER Architecture to serve the evolving needs of customers.

• 100+ members investing on Power architecture
• Open from the chip through software
• Geographic, sector, and expertise diversity
• 8 workgroups chartered, 12 systems underway
• Hundreds of ISVs serving Linux on POWER
• Partner announced plans, offerings, deployments

Join the conversation at #OpenPOWERSummit

© 2014 OpenPOWER Foundation
IBM and NVIDIA partner to build next generation supercomputers

GPU-Accelerated POWER-Based Systems Available in 2014
POWER S824L (8247-42L)

Open Technology Platform for High Performance Analytics, Big Data, and Java Applications workloads

Incorporating the innovation of the OpenPOWER Community

- Partnership with NVIDIA

Accelerates by GPU

- Exploit the uncompromising performance of proven POWER8 and NVIDIA GPU.

High Performance Analytics, Big Data, Java Application Workhorse

- Aim to deliver a new class of technology that maximizes performance and efficiency for all types of technical compute and high performance analytics workloads as well as Java and Big Data Applications.

http://www.ibm.com/systems/power/
IBM and NVIDIA deliver new acceleration capabilities for analytics, big data, and Java

- Runs pattern extraction analytic workloads 8x faster
- Provides new acceleration capability for analytics, big data, Java, and other technical computing workloads
- Delivers faster results and lower energy costs by accelerating processor intensive applications
- Relieves utilization pressure on CPUs

Power System S824L
- Ubuntu Linux running bare metal
- CUDA Toolkit and libraries (5.5)
- IBM XL and GCC as host compilers
POWER S824L (8247-42L)

Processor
- 2x 10-core 3.42GHz or
- 2x 12-core 3.02GHz

Memory
- Total 16 DDR3 CDIMM slots
- 16,32,64 GB CDIMM @ 1600 Mbps
- 1TB capacity, 384GB/s bandwidth max

Storage
- JBOD, RAID 0,10,5,6
  - 12 SFF Disk Drive, 1 DVD

LAN adapters
- 2x 10GBASE-T adapter or
- 2x 10Gb SFP+ Fiber SR plus 2x 1GE adapter

GPU adapter (1 min or 2 max)
- El Capitan NVIDIA K40 GPU adapter

Power supply
- 2+2 1400W PS

O/S Support
- Ubuntu 14.10

Hypervisor
- OPAL, No virtualization

PCIe Gen3 Slots
- 4 PCIe x16 G3 FHFL slots
- 6 PCIe x8 G3 FHHL slots
- CAPI capable on PCIe x16 slots

Native I/O
- USB 3.0 (2 front, 2 rear)
- System Management 1GE (2 rear)
- System port (rear), USB 2.0 (2 rear)
Power Systems NVIDIA GPU support

<table>
<thead>
<tr>
<th>GPU</th>
<th>K40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double Precision FP Performance</td>
<td>1.43 TFLOPS (Peak)</td>
</tr>
<tr>
<td>Double Precision SP Performance</td>
<td>4.29 TFLOPS (Peak)</td>
</tr>
<tr>
<td>Memory Bandwidth</td>
<td>288 GB/sec</td>
</tr>
<tr>
<td>Memory Size</td>
<td>12 GB</td>
</tr>
<tr>
<td>CUDA Cores</td>
<td>2880</td>
</tr>
</tbody>
</table>

Server support: POWER8 S824L
Amdahl’s Law

Maximum Speedup with n parallel processors = \[ \frac{1}{\%\text{Serial} + \left(\%\text{Parallel} / n\right)} \]

All Serial Execution
- Initialize
- Serial Step
- Serial Step
- ForEach...
  - Loop0
  - Loop1
  - ...
  - Loop8
  - Loop9
- Serial Step
- Finalize

Mixed Execution
- Initialize
- Serial Step
- Serial Step
- ForEach...
  - Loop0
  - Loop1
  - ...
  - Loop8
  - Loop9
- Serial Step
- Finalize

6 Serial Steps, including loop setup
10 Parallel Steps

Time = 16 \rightarrow 7
Speedup 2.28

Join the conversation at #OpenPOWERSummit
Anatomy of a CUDA Application

- **Serial** code executes in a **Host** (CPU) thread
- **Parallel** code executes in many **Device** (GPU) threads across multiple processing elements
A Platform for *Diverse* Parallel Computing

Developers want to build front-ends for Java, Python, R, DSLs ...

Target other processors like ARM, FPGAs, GPUs, x86, POWER ...

Cuda
C, C++, Fortran

LLVM Compiler
For CUDA

New Language Support

New Processor Support

NVIDIA GPUs
x86 CPUs

Join the conversation at #OpenPOWERSummit
CUDA Compiler Architecture (4.1, 4.2, 5.0 ...)

NVVM = LLVM with NVIDIA enhancements

PTX = NVIDIA Virtual Machine Assembly
Overview of the XL compilers

- **Target Linux, AIX on Power**
  - Common technology for Blue Gene/Q, and zOS (XL C/C++ only for zOS)
- **Advanced optimization capabilities**
  - Full platform exploitation
  - Loop transformations, SIMDization and vectorization
  - Parallelization
  - Whole program optimization (IPA)
  - Profile-directed optimization (PDF)
- **Language standard compliance**
  - C99, selected C11 features
  - C++98 and subsequent TRs, selected C++11 features
  - Fortran 2003, selected Fortran 2008 features
- **Fully backward compatible with objects complied with older compilers**
- **GCC affinity**
  - Partial source and full binary compatible with GCC
Objectives:
- Accelerate SQL queries
- Compare DB2 BLU with/without GPU
- Show CPU offload

Highlights:
- ~9x higher throughput
- 96 concurrent queries
- 3 minutes to run DB2 BLU non-GPU
- less than 20 Seconds on DB2 BLU with GPU
- CPU is freed up to 30% when we use GPU

OpenPOWER: Open architecture
- POWER8 Server(S824L, 2s/24c, 512GB RAM)

NVIDIA GPUs: Massively parallel
- Thousands of cores, 2880 in NVIDIA K40
- Up to 288 GB/sec memory bandwidth
Regular Expression Parsing

Potential applications:
- Database – predicates
- Stream processing – pattern matching
- Distributed processing
- Security – packet inspection and analysis

Highlights: ~2-3x higher throughput
- Regex pattern compiled to executable representation
- Execute multiple regular expression operations concurrently on multiple fragments
- Based on open source re2 library
- Current GPU implementation not suitable for latency sensitive applications
GPU Accelerated Text Search

Technology Exploration

- Enterprise applications:
  - Content management systems
  - Databases
  - Security
- Prototype:
  - Search index cached on K40 GPU
  - Supports multi-GB indices
  - Subset of Lucene query syntax
    - Boolean, term, span queries
- Highlights: ~10x reduction in query latency
  - 2000 queries, 4 terms per search
  - Measured query latency and throughput

Join the conversation at #OpenPOWERSummit
GPU Accelerated Image Processing / Analytics

- **Enterprise applications:**
  - Smarter Planet – surveillance, law enforcement, counter fraud
  - Content Management systems
  - Security
- **Image processing opportunities:**
  - Highly data parallel operations
  - Compute intensity grows with image size/resolution
- **Face Detection Prototype:** ~4.75x performance improvement
  - Haar Cascade processes 10-20 QVGA streams at 12-15 FPS used on CPU
  - Binary Brightness Feature implemented for GPU can process 75 QVGA streams at 2-15 FPS

---

**Technology Exploration**

**FPS vs algorithm/platform**

- **Concurrent face detection on multiple QVGA image streams with and without GPU**

  - SVS (cpu)
  - BBF (GPU)

Join the conversation at #OpenPOWERSummit
BACKUP
How GPU Acceleration Works

Application Code

Most Compute-Intensive Functions

Small Highly Parallel Part of Code

Original Serial CPU Code

GPU

CPU

Minimal code changes for maximum performance.

Join the conversation at #OpenPOWERSummit
3 Ways to Accelerate Applications

Applications

Libraries

“Drop-in” Acceleration

OpenACC Directives

Easily Accelerate Applications

Programming Languages

Maximum Flexibility

Join the conversation at #OpenPOWERSummit
NVIDIA Supplied Libraries

- cuFFT – Fast Fourier Transforms Library
- cuBLAS – Complete BLAS Library
- cuSPARSE – Sparse Matrix Library
- cuRAND – Random Number Generation (RNG) Library
- NPP – Performance Primitives for Image & Video Processing
- Thrust – Templated Parallel Algorithms & Data Structures
- math.h – C99 floating-point library
- cuDNN – Deep Neural Net building blocks

Included in the CUDA Toolkit (free download): developer.nvidia.com/cuda-toolkit
For more information on CUDA libraries: developer.nvidia.com/gpu-accelerated-libraries