INTRODUCING TESLA P100
New GPU Architecture to Enable the World’s Fastest Compute Node

Pascal Architecture
Highest Compute Performance

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

Unified Memory
CPU
Tesla P100
GIANT LEAPS IN EVERYTHING

3x Compute

5x GPU-GPU BW

3x GPU Mem BW
TESLA P100 PERFORMANCE DELIVERED
NVLink for Max Scalability, More than 45x Faster with 8x P100

![Graph showing speed-up vs Dual Socket Haswell for various applications with different GPU configurations.]

- 2x K80 (M40 for Alexnet)
- 2x P100
- 4x P100
- 8x P100

Applications tested:
- Caffe/Alexnet
- VASP
- HOOMD-Blue
- COSMO
- MILC
- Amber
- HACC

2x Haswell CPU
PASCAL ARCHITECTURE
TESLA P100 GPU: GP100

- 56 SMs
- 3584 CUDA Cores
- 5.3 TF Double Precision
- 10.6 TF Single Precision
- 21.2 TF Half Precision
- 16 GB HBM2
- 720 GB/s Bandwidth
## GPU PERFORMANCE COMPARISON

<table>
<thead>
<tr>
<th></th>
<th>P100</th>
<th>M40</th>
<th>K40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Double Precision TFlop/s</td>
<td>5.3</td>
<td>0.2</td>
<td>1.4</td>
</tr>
<tr>
<td>Single Precision TFlop/s</td>
<td>10.6</td>
<td>7.0</td>
<td>4.3</td>
</tr>
<tr>
<td>Half Precision Tflop/s</td>
<td>21.2</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Memory Bandwidth (GB/s)</td>
<td>720</td>
<td>288</td>
<td>288</td>
</tr>
<tr>
<td>Memory Size</td>
<td>16GB</td>
<td>12GB, 24GB</td>
<td>12GB</td>
</tr>
</tbody>
</table>
**GP100 SM**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUDA Cores</td>
<td>64</td>
</tr>
<tr>
<td>Register File</td>
<td>256 KB</td>
</tr>
<tr>
<td>Shared Memory</td>
<td>64 KB</td>
</tr>
<tr>
<td>Active Threads</td>
<td>2048</td>
</tr>
<tr>
<td>Active Blocks</td>
<td>32</td>
</tr>
</tbody>
</table>
P100 SM

Maxwell SM

More resources per core

- 2x Registers
- 1.33x Shared Memory Capacity
- 2x Shared Memory Bandwidth
- 2x Warps

Higher Instruction Throughput
IEEE 754 FLOATING POINT ON GP100
3 sizes, 3 speeds, all fast

<table>
<thead>
<tr>
<th>Feature</th>
<th>Half precision</th>
<th>Single precision</th>
<th>Double precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layout</td>
<td>s5.10</td>
<td>s8.23</td>
<td>s11.52</td>
</tr>
<tr>
<td>Issue rate</td>
<td>pair every clock</td>
<td>1 every clock</td>
<td>1 every 2 clocks</td>
</tr>
<tr>
<td>Subnormal support</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Atomic Addition</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
HALF-PRECISION FLOATING POINT (FP16)

- 16 bits
  - 1 sign bit, 5 exponent bits, 10 fraction bits

- $2^{40}$ Dynamic range
  - Normalized values: 1024 values for each power of 2, from $2^{-14}$ to $2^{15}$
  - Subnormals at full speed: 1024 values from $2^{-24}$ to $2^{-15}$

- Special values
  - +/- Infinity, Not-a-number

USE CASES

- Deep Learning Training
- Radio Astronomy
- Sensor Data
- Image Processing
NVLink
P100 supports 4 NVLinks

Up to 94% bandwidth efficiency

Supports read/writes/atomics to peer GPU

Supports read/write access to NVLink-enabled CPU

Links can be ganged for higher bandwidth

NVLink on Tesla P100
NVLINK - GPU CLUSTER

Two fully connected quads, connected at corners

160GB/s per GPU bidirectional to Peers

Load/store access to Peer Memory

Full atomics to Peer GPUs

High speed copy engines for bulk data copy

PCIe to/from CPU
Fully connected quad

120 GB/s per GPU bidirectional for peer traffic

40 GB/s per GPU bidirectional to CPU

Direct Load/store access to CPU Memory

High Speed Copy Engines for bulk data movement
TESLA P100 PHYSICAL CONNECTOR
With NVLink
HBM2 STACKED MEMORY
HBM2: 720GB/SEC BANDWIDTH

And ECC is free
UNIFIED MEMORY
PAGE MIGRATION ENGINE
Support Virtual Memory Demand Paging

49-bit Virtual Addresses
  Sufficient to cover 48-bit CPU address + all GPU memory

GPU page faulting capability
  Can handle thousands of simultaneous page faults

Up to 2 MB page size
  Better TLB coverage of GPU memory
KEPLER/MAXWELL UNIFIED MEMORY

CUDA 6+

Kepler GPU  CPU

Unified Memory

Allocate Up To GPU Memory Size

Simpler Programming & Memory Model

- Single allocation, single pointer, accessible anywhere
- Eliminate need for *explicit copy*
- Greatly simplifies code porting

Performance Through Data Locality

- Migrate data to accessing processor
- Guarantee global coherency
- Still allows explicit hand tuning
PASCAL UNIFIED MEMORY
Large datasets, simple programming, High Performance

CUDA 8

Pascal GPU

CPU

Unified Memory

Allocate Beyond GPU Memory Size

Enable Large Data Models
Oversubscribe GPU memory
Allocate up to system memory size

Tune Unified Memory Performance
Usage hints via cudaMemAdvise API
Explicit prefetching API

Simpler Data Access
CPU/GPU Data coherence
Unified memory atomic operations
INTRODUCING TESLA P100

New GPU Architecture to Enable the World’s Fastest Compute Node

Pascal Architecture

- Highest Compute Performance

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

More P100 Features: compute preemption, new instructions, larger L2 cache, more...

Find out more at http://devblogs.nvidia.com/parallelforall/inside-pascal