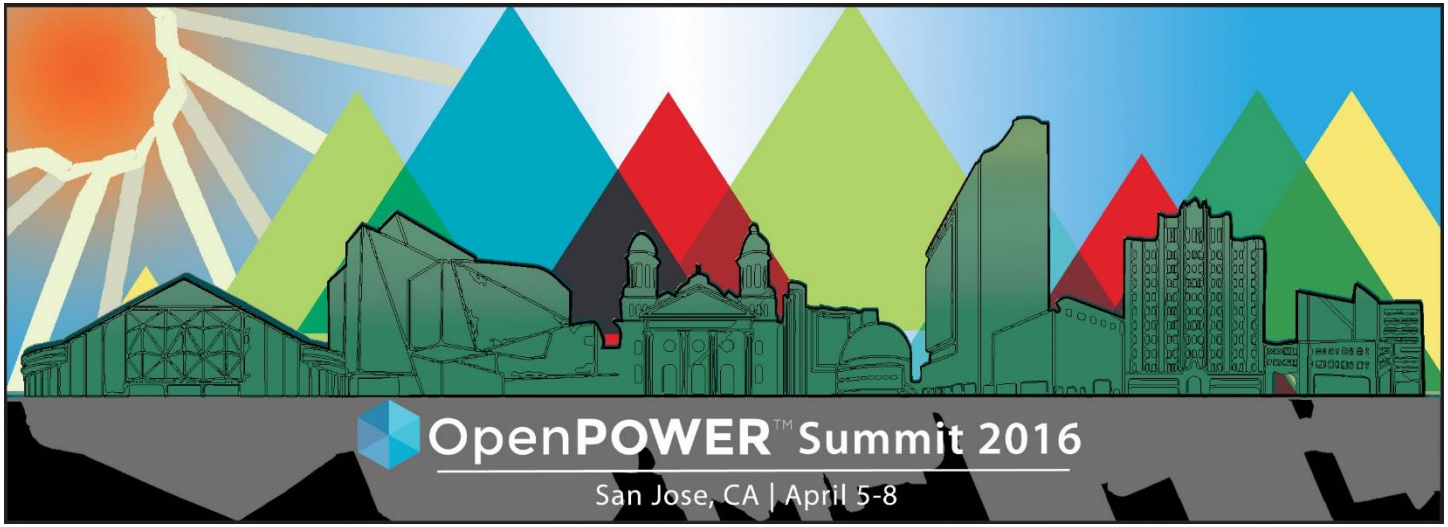
OpenPOWER™

Programming Forward for NVIDIA Pascal Architecture on OpenPOWER Platforms

John Ashley, Senior IBM Developer Relationships Manager, NVIDIA

Revolutionizing the Datacenter



Join the Conversation #OpenPOWERSummit

Key Takeaways

- From Power8™, Tesla® K80 to Power 8+™, Tesla P100
  - GPU accelerated code ***at least* 2x** faster.
  - CPU-GPU data transfers ***at least* 2x** faster.
  - Deep Learning ***more than 2x***!
- Easier coding: Tesla P100, CUDA® 8 Unified Memory
- Faster execution: Power 8+, NVLink™, Tesla P100

Programming Forward

- NVIDIA Tesla P100 -- More FLOPS, more Memory BW, More RAM
- NVLink -- Breaking the PCIe Speed Limit
- Taking advantage of Unified Memory
- Getting Started

NVIDIA Tesla P100

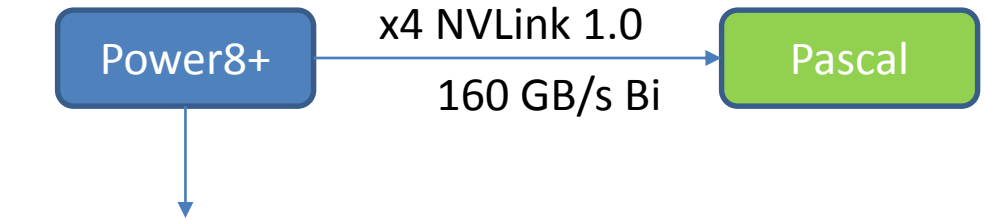
- Newest & Fastest GPU generation but…
  - …similar to Maxwell from a programmer's perspective

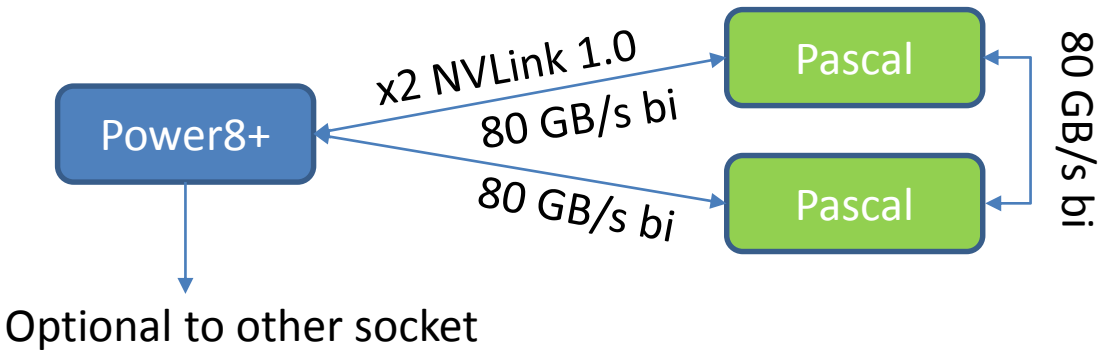
Feature	Value	Feature	Value
FLOPS@boost		Mem BW	> 700 GB/s
Double	> 5TF	PCIe BW, Peak	32 GB/s Bidirect
Single	> 10TF	NVLink BW, Peak	160 GB/s Bidirect
Half	> 20TF	Core Counts	3840
Resources / Core generally >= Maxwell or Kepler			

NVLink

- New interconnect for GPU-GPU and CPU-GPU communications
  - GPU requires Pascal generation GPU
  - CPU requires Power8+
- Transparent to the developer – just 2-3x faster than PCIe!
- Technical Details
  - Point to Point Interconnect
  - 4x sub links each 20 GB/s Peak per direction, about 16 GB/s observed
  - Topology determined by motherboard

Some Possible NVLink Systems





Many other options including mixed PCIe and NVLink interconnects

Unified Memory

- Speeds and eases development ; can result in very fast code as well
- First Available in CUDA 6 with Kepler
  - Allows opt-in allocation of variables for automatic memory management
  - No-overallocation of GPU memory – duplicate allocation on both CPU & GPU
  - Synchronous copy to GPU
  - Demand paging back from CPU
- Significant upgrade with CUDA 8 & Pascal
  - Still opt-in
  - Can now over-allocate GPU memory
  - Demand paging from either side
  - You can provide hints for pre-fetch, preferred location, etc

OpenPower w/ NVLink is 2-3x faster than over PCIe

OpenACC

- Directives based approach
- Power & GPU production version ships late 2016
- LOTS more details in Doug's talk at 4:25 !

Getting Started

- Code is GPU accelerated on Power today
  - Will just run faster on P100 and Power8+
  - Profile and talk to NVIDIA for advanced tuning help
- Code is GPU accelerated on x86 today
  - Port to Power8 & get Kepler baseline
  - Many IBM and NVIDIA resources
- Code is not GPU accelerated today
  - Examine current profile for performance opportunities
  - Look at similar codes that may already be GPU enabled
  - Look at Libraries, OpenACC, Unified Memory
  - Contact IBM & NVIDIA

NEW IBM NVIDIA Accelerator Lab!

Additional Resources

- IBM NVIDIA Accelerator Lab
  - accellab@us.ibm.com
- NVIDIA Tesla P100
  - …Coming soon…
- CUDA 8
  - **S6224 - Featured Presentation: CUDA 8 and Beyond**
  - **S6531 - CUDA® Debugging Tools in CUDA 8**
  - **S6810 - Optimizing Application Performance with CUDA® Profiling Tools**
- OpenACC
  - OPS: PGI Accelerator Fortran/C/C++ Compilers for OpenPOWER+Tesla
  - **S6410 - Comparing OpenACC 2.5 and OpenMP 4.5**
- Unified Memory
  - **S6216 - The Future of Unified Memory**
  - **S6134 - High Performance and Productivity with Unified Memory and OpenACC: A LBM Case Study**

OR always
jashley@nvidia.com