HETEROGENEOUS HPC, ARCHITECTURAL OPTIMIZATION, AND NVLINK

STEVE OBERLIN
CTO, TESLA ACCELERATED COMPUTING
NVIDIA
STATE OF THE ART 2012

18,688 Tesla K20X GPUs
27 PetaFLOPS
FLAGSHIP SCIENTIFIC APPLICATIONS ON TITAN

- Material Science (WL-LSMS)
- Climate Change (CAM-SE)
- Biofuels (LAMMPS)
- Astrophysics (NRDF)
- Combustion (S3D)
- Nuclear Energy (Denovo)
STATE OF THE ART 2017

CORAL Summit System
5-10x Faster
1/5th the Nodes,
Same Energy Use as Titan

Earth Simulator
40.96 TF
Top500 #1 for 3 years

Just 1 Node in Summit
Is the same performance as the
Earth Simulator in 2002
AGENDA: 3 STORIES

- Heterogeneous HPC
  - HPC System Evolution
  - Architecture and Technology Basis for Heterogeneous Processing

- Architectural Optimization and NVLink
  - Tesla Accelerated Computing Platform
  - NVLink

- CORAL Punctuation Mark
  - Optimized Fat node performance projections
ARE YOU CALIBRATED?

.6 MPH  6 MPH  60 MPH  600 MPH

3 ORDERS OF MAGNITUDE
HPC SYSTEM EVOLUTION

MEMORY

CPU
CRAY-1
LATENCY-HIDING SINGLE VECTOR CPU, 160 MFLOPS PEAK
### FIRST LINPACK LIST, JANUARY 1979

<table>
<thead>
<tr>
<th>Facility</th>
<th>Time (micro-sec.)</th>
<th>Computer</th>
<th>Type</th>
<th>Compiler</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCAR</td>
<td>0.049</td>
<td>CRAY-1</td>
<td>S</td>
<td>CFT, Assembly BLAS</td>
</tr>
<tr>
<td>LASL</td>
<td>0.448</td>
<td>CDC 7600</td>
<td>S</td>
<td>FTF, Assembly BLAS</td>
</tr>
<tr>
<td>NCAR</td>
<td>0.519</td>
<td>CRAY-1</td>
<td>S</td>
<td>CFT</td>
</tr>
<tr>
<td>LASL</td>
<td>0.618</td>
<td>CDC 7600</td>
<td>S</td>
<td>FTF</td>
</tr>
<tr>
<td>Argonne</td>
<td>0.216</td>
<td>IBM 370/195</td>
<td>D</td>
<td>H</td>
</tr>
<tr>
<td>Argonne</td>
<td>0.359</td>
<td>CDC 7600</td>
<td>S</td>
<td>Local</td>
</tr>
<tr>
<td>NASA Langley</td>
<td>1.486</td>
<td>CDC Cyber 175</td>
<td>S</td>
<td>FTN</td>
</tr>
<tr>
<td>U. Ill. Urbana</td>
<td>0.500</td>
<td>CDC Cyber 175</td>
<td>S</td>
<td>Ext. 4.6</td>
</tr>
<tr>
<td>LLL</td>
<td>0.534</td>
<td>CDC 7600</td>
<td>S</td>
<td>CHAT, No optimize</td>
</tr>
<tr>
<td>SLAC</td>
<td>0.378</td>
<td>IBM 370/168</td>
<td>D</td>
<td>H Ext., Fast mult.</td>
</tr>
<tr>
<td>Michigan</td>
<td>1.847</td>
<td>AMDahl 470/64</td>
<td>D</td>
<td>H</td>
</tr>
<tr>
<td>Toronto</td>
<td>3.557</td>
<td>IBM 370/185</td>
<td>D</td>
<td>H Ext., Fast mult.</td>
</tr>
<tr>
<td>Northwestern</td>
<td>4.202</td>
<td>CDC 6600</td>
<td>S</td>
<td>FTN</td>
</tr>
<tr>
<td>Texas</td>
<td>5.618</td>
<td>CDC 6600</td>
<td>S</td>
<td>RUN</td>
</tr>
<tr>
<td>China Lake</td>
<td>5.865</td>
<td>Univac 1110</td>
<td>S</td>
<td>V</td>
</tr>
<tr>
<td>Yale</td>
<td>7.001</td>
<td>DEC KL-70</td>
<td>S</td>
<td>F20</td>
</tr>
<tr>
<td>Bell Labs</td>
<td>10.1</td>
<td>Honeywell 6800</td>
<td>S</td>
<td>Y</td>
</tr>
<tr>
<td>Wisconsin</td>
<td>10.1</td>
<td>Univac 1110</td>
<td>S</td>
<td>V</td>
</tr>
<tr>
<td>Iowa State</td>
<td>10.2</td>
<td>Itel A5/5 mod3</td>
<td>D</td>
<td>H</td>
</tr>
<tr>
<td>U. Ill. Chicago</td>
<td>11.9</td>
<td>IBM 370/158</td>
<td>D</td>
<td>GI</td>
</tr>
<tr>
<td>Purdue</td>
<td>16.2</td>
<td>CDC 6500</td>
<td>S</td>
<td>FTF</td>
</tr>
<tr>
<td>U. C. San Diego</td>
<td>36.6</td>
<td>Burroughs 6700</td>
<td>S</td>
<td>H</td>
</tr>
<tr>
<td>Yale</td>
<td>6.23</td>
<td>DEC KA-10</td>
<td>S</td>
<td>F40</td>
</tr>
</tbody>
</table>

* TIME(100) = (100/75)**3 SGEFA(75) + (100/75)**2 SGESL(75)
TOP500 SUPERCOMPUTER LIST 1993-2012

9 ORDERS OF MAGNITUDE
HPC SYSTEM EVOLUTION
CRAY-2
4 LATENCY-HIDING VECTOR CPUS, 2 GFLOPS PEAK, 1985
HPC SYSTEM EVOLUTION
The Attack of the Killer Micros

E. D. Brooks III
Massively Parallel Computing Initiative
Lawrence Livermore National Laboratories
brooks@{maddog.llnl.gov, maddog.uucp}

Presented at: Supercomputing ’89
Reno, Nov 13-17 1989
CRAY T3D
DEC ALPHA EV4 MICROPROCESSORS, 1 TFLOPS PEAK, 1993
HETEROGENEOUS RESULTS

![Graph showing heterogeneous results for NASPB BT](image1)

![Graph showing heterogeneous results for NASPB CG](image2)
CRAY T3E
DEC ALPHA EV5 MICROPROCESSORS, 1 TFLOPS SUSTAINED, 1995
“There’s hasn’t really been anything truly new for 20 years... except GPUs.”

-- Thomas Schulthess, Director, CSCS
COMING SOON TO A NODE NEAR YOU

- **Cray T3E**
  - 2.4 TFLOPS peak performance (2K processors)
  - ~128 GB/s bisection bandwidth

- **CRAY T90**
  - 56 GFLOPs peak performance (32 processors)
  - 1 TB/s bisection bandwidth

- *Pascal GPU (1H16) has >T3E peak with T90 memory BW*
OPTIMIZING SERIAL/PARALLEL EXECUTION

Application Code

GPU

Parallel Work
Majority of Ops

Serial Work
System and Sequential Ops

CPU

+
TWO COMPUTING MODELS FOR ACCELERATORS

Many-Weak-Cores (MWC) Model
Single CPU Core for Both Serial & Parallel Work

Xeon Phi (And Others)
Many Weak Serial Cores

Heterogeneous Computing Model
Complementary Processors Work Together

CPU
Optimized for Serial Tasks

GPU Accelerator
Optimized for Parallel Tasks
AMDAHL’S LAW ANALYSIS

98% Parallel Work

Minutes Run Time

- 1 GPU + 1 CPU
- 2 x MWC (.25x CPU)
- Work 1x CPU

Serial
Parallel

GPU TECHNOLOGY CONFERENCE
AMDAHL’S LAW ANALYSIS

90% Parallel Work

- **1 GPU + 1 CPU**
- **2 x MWC (.25x CPU)**
- **Work 1x CPU**

Minutes Run Time

Serial
Parallel

GPU TECHNOLOGY CONFERENCE
AMDAHL’S LAW ANALYSIS

80% Parallel Work

- 1 GPU + 1 CPU
- 2 x MWC (.25x CPU)
- Work 1x CPU

Minutes Run Time
AMDAHL’S LAW ANALYSIS

70% Parallel Work

- 1 GPU+1 CPU: 3 minutes run time (2.5 minutes parallel)
- 2 x MWC (.25x CPU): 10.5 minutes run time (2.5 minutes parallel)
- Work 1x CPU: 6 minutes run time (4.5 minutes parallel)
AMDAHL’S LAW ANALYSIS

60% Parallel Work

- 1 GPU+1 CPU
- 2 x MWC (.25x CPU)
- Work 1x CPU

Minutes Run Time

Serial
Parallel
AMDAHL’S LAW ANALYSIS

50% Parallel Work

- 1 GPU+1 CPU
- 2 x MWC (.25x CPU)
- Work 1x CPU

Minutes Run Time

GPU TECHNOLOGY CONFERENCE
AMDAHL’S LAW ANALYSIS

40% Parallel Work

- 1 GPU + 1 CPU
- 2 x MWC (.25x CPU)
- Work 1x CPU

Minutes Run Time
AMDAHL’S LAW ANALYSIS

30% Parallel Work

1 GPU+1 CPU

2 x MWC (.25x CPU)

Work 1x CPU

Minutes Run Time

Serial
Parallel
AMDAHL’S LAW ANALYSIS

20% Parallel Work

- 1 GPU+1 CPU
- 2 x MWC (.25x CPU)
- Work 1x CPU

Minutes Run Time

Serial
Parallel
AMDAHL’S LAW ANALYSIS

10% Parallel Work

- 1 GPU+1 CPU
- 2 x MWC (.25x CPU)
- Work 1x CPU

Minutes Run Time

Serial
Parallel
TESLA K80

WORLD’S FASTEST ACCELERATOR FOR DATA ANALYTICS AND SCIENTIFIC COMPUTING

**2x Faster**
2.9 TF | 4992 Cores | 480 GB/s

**Double the Memory**
Designed for Big Data Apps

**Maximum Performance**
Dynamically Maximize Performance for Every Application

Deep Learning: Caffe

![Bar Chart Showing Caffe Performance](chart.png)

Double the Memory

![Memory Diagram](memory_diagram.png)

Maximum Performance

![Performance Boost](performance_boost.png)

Caffe Benchmark: AlexNet training throughput based on 20 iterations, CPU: E5-2697v2 @ 2.70GHz. 64GB System Memory, CentOS 6.2
FOCUS ON THROUGHPUT PERFORMANCE

Peak Double Precision FLOPS

Peak Memory Bandwidth

GFLOPS

GB/s

NVIDIA GPU

x86 CPU

M1060

M2090

K20

K40

K80

Westmere

Sandy Bridge

Ivy Bridge

Haswell

2008

2009

2010

2011

2012

2013

2014

2008

2009

2010

2011

2012

2013

2014
10X FASTER THAN CPU ON APPLICATIONS

CPU: 12 cores, E5-2697v2 @ 2.70GHz. 64GB System Memory, CentOS 6.2
GPU: Single Tesla K80, Boost enabled
TESLA PLATFORM ENABLES OPTIMIZATION

Ecosystem Industry Standard CPUs and Interconnects

- ARM64
- POWER
- x86
- Ethernet
- Cray
- InfiniBand
- Others

Industry-Driven Solutions
LOGICAL VS. PHYSICAL INTEGRATION

OPTIMIZING FOR EFFICIENCY + FLEXIBILITY

Latency-Optimized

Throughput-Optimized

System Memory

GPU Memory

NVLink

CPU

GPU
UNIFIED MEMORY

DRAMATICALLY LOWER DEVELOPER EFFORT

Exposed Developer View
(Pre-CUDA 5.5)

Developer View With Unified Memory
NVLink: Node Integration Network

- 5x PCIe bandwidth
- Move data at CPU memory speed
- 3x lower energy/bit

**Tesla GPU**

- HBM
  - 1 Terabyte/s
- Stacked Memory
- Throughput Optimized

**Power or ARM CPU**

- DDR4
  - 50-75 GB/s
- DDR Memory
- Latency Optimized

NVLink 80 GB/s
NVLink
HIGH-SPEED NODE NETWORK
NVLink Multi-GPU Performance

- Over 2x Application Performance Speedup
- When Next-Gen GPUs Connect via NVLink Versus PCIe

### Key Points:

- GPUs Interconnected with NVLink
- NVLink 1.00x, 1.25x, 1.50x, 1.75x, 2.00x, 2.25x
- 5x Faster than PCIe Gen3 x16

### Applications:

- ANSYS Fluent
- Multi-GPU Sort
- LQCD QUDA
- AMBER
- 3D FFT

### Performance Speedup:

- 3D FFT, ANSYS: 2 GPU configuration, All other apps comparing 4 GPU configuration
- AMBER Cellulose (256x128x128), FFT problem size (256^3)
TESLA PLATFORM ENABLES OPTIMIZATION

Scalable Nodes, ISA Choice

NVLink

2 GPUs per Node

3 GPUs per Node

4 GPUs per Node

x86 CPU

NVLINK 20GB/s

PCIe Gen3 x16
NVLINK-ENABLED HETEROGENEOUS NODE
LOGICAL INTEGRATION + FLEXIBILITY + EFFICIENCY

IBM POWER CPU
Most Powerful Serial Processor

NVIDIA NVLink
Node Integration Interconnect

80-200 GB/s

NVIDIA Volta GPU
Most Powerful Parallel Processor
Approximately 3,400 nodes, each with:

- IBM POWER9 CPUs and multiple NVIDIA Tesla® Volta GPUs
- CPUs and GPUs integrated on-node with high speed NVLink
- Large coherent memory: over 512 GB (HBM + DDR4)
  All directly addressable from the CPUs and GPUs
- An additional 800 GB of NVRAM, burst buffer or as extended memory
- Over 40 TF peak performance/node(!)
OPTIMIZED HETEROGENEOUS NODE

CORAL Application Performance Projections

Scalable Science Benchmarks

Throughput Benchmarks
SUMMARY:

- Heterogeneous acceleration is powerful and efficient
  - Latency-optimized cores for serial and system work
  - Throughput-optimized cores for on-node parallel work
  - High-performance integration

- NVLink provides a compelling node integration advantage
  - Flexible configuration of resources in application-driven proportions
  - Scalable performance into the future

- CORAL is awesome, you can buy nodes just like it.
  - OpenPower is surprisingly affordable.
  - It’s just one example of optimized architecture for an application set and budget.
NVIDIA REGISTERED DEVELOPER PROGRAMS

- Everything you need to develop with NVIDIA products
- Membership is your first step in establishing a working relationship with NVIDIA Engineering
  - Exclusive access to pre-releases
  - Submit bugs and features requests
  - Stay informed about latest releases and training opportunities
  - Access to exclusive downloads
  - Exclusive activities and special offers
  - Interact with other developers in the NVIDIA Developer Forums

REGISTER FOR FREE AT: developer.nvidia.com
THANK YOU

JOIN THE CONVERSATION
#GTC15