NVIDIA HPC Update

Stan Posey; sposey@nvidia.com; NVIDIA, Santa Clara, CA, USA
NVIDIA Core Technologies and Markets

Company Revenue of ~$5B USD; ~9,000 Employees; HPC Growing > 30% CAGR

PC

DATA CENTER

MOBILE

GAMING

DESIGN

ENTERPRISE VIRTUALIZATION

HPC & CLOUD SERVICE PROVIDERS

AUTONOMOUS MACHINES

Tesla®
GPU Technology a Mainstream HPC Platform

Data Center Infrastructure

- System Solutions
- Communication
- Infrastructure Management

Development

- Programming Languages
- Development Tools
- Software Solutions

- GPU Accelerators
  - GPU Boost
- Interconnect
  - GPU Direct
  - NVLink
- System Management
  - NVML
- Compiler Solutions
  - LLVM
- Profile and Debug
  - CUPTI
- Accelerated Libraries
  - cuBLAS

Enterprise Services Support & Maintenance
GPUs Applied Across Mainstream HPC Domains

- Oil & Gas
  - Schlumberger
  - Petrobras
  - Eni
  - Chevron
  - Statoil

- Higher Ed
  - Harvard School of Engineering and Applied Sciences
  - Stanford University
  - Georgia Tech
  - ETH
  - University of Cambridge

- Government
  - Air Force Research Laboratory
  - Raytheon
  - NASA
  - Naval Research Laboratory

- Supercomputing
  - CSCS
  - NCSA
  - Tokyo Institute of Technology
  - Lawrence Livermore National Laboratory

- Finance
  - J.P. Morgan
  - Barclays
  - Standard Life
  - BNP Paribas
  - Murex

- Consumer Web
  - Baidu
  - Salesforce
  - Shazam
  - Amazon
  - Yandex
We love GPU cards. We just use a lot of them.

— Jeff Dean, Google

In five years, we think 50% of queries will be speech or images.

— Andrew Ng, Baidu

50+ Deep Learning Sessions

www.gputechconf.com

<table>
<thead>
<tr>
<th>Developer Labs</th>
<th>Caffe</th>
<th>Torch</th>
<th>Theano</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adobe</td>
<td>Google</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alibaba</td>
<td>iFlyte, Ltd</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baidu</td>
<td>NUANCE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Carnegie Mellon</td>
<td>Stanford Univ</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facebook</td>
<td>UC Berkeley</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flickr / Yahoo</td>
<td>Univ of Toronto</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Broad Use of GPUs in Deep Learning

HPC Drivers for Deep Learning → More Data → Better Models → GPUs

Early Adopters
- Adobe: Image Analytics for Creative Cloud
- Baidu: Speech/Image Recognition
- flickr: Image Classification
- IBM: Hadoop
- Netflix: Recommendation
- Yandex: Search Rankings

Applications
- Image Detection
- Face Recognition
- Gesture Recognition
- Video Search & Analytics
- Speech Recognition & Translation
- Recommendation Engines
- Indexing & Search

Talks @ GTC
- facebook
- NYU
- Stanford University
- DARPA
- DENSO
- Carnegie Mellon University
- MIT
- Massachusetts Institute of Technology
- Berkeley University
## Tesla GPU Progression During Recent Years

<table>
<thead>
<tr>
<th></th>
<th>2012 (Fermi) M2075</th>
<th>2014 (Kepler) K20X</th>
<th>2014 (Kepler) K40</th>
<th>2014 (Kepler) K80</th>
<th>Kepler / Fermi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak SP</td>
<td>1.03 TF</td>
<td>3.93 TF</td>
<td>4.29 TF</td>
<td>8.74 TF</td>
<td>4x</td>
</tr>
<tr>
<td>Peak SGEMM</td>
<td>2.95 TF</td>
<td>3.22 TF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peak DP</td>
<td>.515 TF</td>
<td>1.31 TF</td>
<td>1.43 TF</td>
<td>2.90 TF</td>
<td>3x</td>
</tr>
<tr>
<td>Peak DGEMM</td>
<td>1.22 TF</td>
<td>1.33 TF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Memory size</td>
<td>6 GB</td>
<td>6 GB</td>
<td>12 GB</td>
<td>24 GB (12 each)</td>
<td>2x</td>
</tr>
<tr>
<td>Mem BW (ECC off)</td>
<td>150 GB/s</td>
<td>250 GB/s</td>
<td>288 GB/s</td>
<td>480 GB/s (240 each)</td>
<td>2x</td>
</tr>
<tr>
<td>Memory Clock</td>
<td>2.6 GHz</td>
<td>3.0 GHz</td>
<td>3.0 GHz</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCIe Gen</td>
<td>Gen 2</td>
<td>Gen 2</td>
<td>Gen 3</td>
<td>Gen 3</td>
<td>2x</td>
</tr>
<tr>
<td># of Cores</td>
<td>448</td>
<td>2688</td>
<td>2880</td>
<td>4992 (2496 each)</td>
<td>5x</td>
</tr>
<tr>
<td>Board Power</td>
<td>235W</td>
<td>235W</td>
<td>235W</td>
<td>300W</td>
<td>0% – 28%</td>
</tr>
</tbody>
</table>

**Note:** Tesla K80 specifications are shown as aggregate of two GPUs on a single board.
# Features of Pascal GPU Architecture – 2016

<table>
<thead>
<tr>
<th>NVLink</th>
<th>Stacked Memory</th>
<th>Unified Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interconnect at 80 GB/s (Speed of CPU Memory)</td>
<td>4x Higher Bandwidth ~1 TB/s</td>
<td>Lower Development Effort (Available Today in CUDA6)</td>
</tr>
<tr>
<td><strong>TESLA GPU</strong></td>
<td>3x Capacity, 4x More Efficient</td>
<td></td>
</tr>
<tr>
<td><strong>CPU (POWER, ?)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBM 1 TB/s</td>
<td><strong>NVLink</strong> 80 - 200 GB/s</td>
<td></td>
</tr>
<tr>
<td>DDR4 50-75 GB/s</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stacked Memory</td>
<td>UVM</td>
<td>DDR Memory</td>
</tr>
</tbody>
</table>

- HBM: 1 TB/s
- DDR4: 50-75 GB/s

---

**Interconnect at 80 GB/s**

- *Speed of CPU Memory*
GPU and CPU Platforms and Configurations

2014 – Kepler GPU

2016 – Pascal GPU

<table>
<thead>
<tr>
<th>Attribute</th>
<th>CPU</th>
<th>x86</th>
<th>POWER</th>
<th>ARM64</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Performance</td>
<td></td>
<td>green</td>
<td>green</td>
<td>red</td>
</tr>
<tr>
<td>Power Consumption</td>
<td>red</td>
<td>yellow</td>
<td></td>
<td>green</td>
</tr>
<tr>
<td>GPU Interconnect Speed</td>
<td>red</td>
<td>green</td>
<td></td>
<td>?</td>
</tr>
<tr>
<td>Development Environment</td>
<td>green</td>
<td>yellow</td>
<td></td>
<td>red</td>
</tr>
</tbody>
</table>
GPU Programming Developments for CPU Platforms

Availability for x86 Today, POWER and ARM Under Development

Libraries
- AmgX
- cuDNN
- cuBLAS
- OpenCV
- Thrust

Compiler Directives
- OpenACC

Programming Languages
- C/C++
- Fortran
- python
- Java

Platforms
- ARM
- x86
IBM Power + NVIDIA GPU Accelerated HPC

Next-Gen IBM Supercomputers and Enterprise Servers

Long term roadmap integration

POWER CPU + Tesla GPU

OpenPOWER Foundation

Open ecosystem built on Power Architecture

First GPU-Accelerated POWER-Based Systems Available Since 2015

& 30+ more…
DOE 2017 CORAL Systems Based on Power + GPU

US DOE CORAL Systems
- Summit (ORNL) and Sierra (LLNL)
- Installation 2017 at ~150 PF each
- Nodes of POWER 9 + Tesla Volta GPUs
- NVLink Interconnect for CPUs + GPUs

ORNL Summit System
- Approximately 3,400 total nodes
- Each node 40+ TF peak performance
- About 1/5 of total #2 Titan nodes (18K+)
- Same energy used as #2 Titan (27 PF)
ACME: Accelerated Climate Model for Energy
- DOE Labs: Argonne, LANL, LBL, LLNL, ORNL, PNNL, Sandia
- Co-design project using US DOE LCF systems
- Branch of CESM using CAM-SE (atm) and MPAS-O (ocn)

ACME Project Roadmap
(page 2 in report)

2017
- CORAL: Summit, Sierra, Aurora
- Trinity: NERSC8 (Cori)

2014
- TITAN – OLCF [AMD + GPU]
- Mira – ALCF [Blue GeneQ]
# GPU Developments for Atmosphere Models

## Global

<table>
<thead>
<tr>
<th>Organization</th>
<th>Location</th>
<th>Model</th>
<th>GPU Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORNL, SNL, NCAR</td>
<td>US</td>
<td>CAM-SE</td>
<td>OpenACC (migration from CUDA-F)</td>
</tr>
<tr>
<td>NOAA ESRL</td>
<td>US</td>
<td>FIM/NIM</td>
<td>OpenACC, F2C-ACC</td>
</tr>
<tr>
<td>NOAA GFDL</td>
<td>US</td>
<td>FV3</td>
<td>OpenACC</td>
</tr>
<tr>
<td>NASA GSFC</td>
<td>US</td>
<td>GEOS-5</td>
<td>OpenACC (migration from CUDA-F)</td>
</tr>
<tr>
<td>NCAR</td>
<td>US</td>
<td>MPAS-A</td>
<td>OpenACC</td>
</tr>
<tr>
<td>ECMWF</td>
<td>UK</td>
<td>IFS (Arpege)</td>
<td>Libs + OpenACC</td>
</tr>
<tr>
<td>MetOffice, STFC</td>
<td>UK</td>
<td>UM/GungHo</td>
<td>OpenACC</td>
</tr>
<tr>
<td>DWD, MPI-M, CSCS</td>
<td>DE, CH</td>
<td>ICON</td>
<td>DSL – dycore, OpenACC – physics</td>
</tr>
<tr>
<td>JMA, Hitachi</td>
<td>JP</td>
<td>JMA-GSM</td>
<td>OpenACC</td>
</tr>
<tr>
<td>JAMSTEC, UT, RIKEN</td>
<td>JP</td>
<td>NICAM</td>
<td>Libs + OpenACC</td>
</tr>
<tr>
<td>NCAR; SSEC</td>
<td>US</td>
<td>WRF-ARW</td>
<td>(i) OpenACC, (ii) CUDA</td>
</tr>
<tr>
<td>CSCS, MCH</td>
<td>CH</td>
<td>COSMO</td>
<td>DSL – dycore, OpenACC – physics</td>
</tr>
<tr>
<td>Bull, MFR</td>
<td>FR</td>
<td>HARMONIE</td>
<td>OpenACC</td>
</tr>
<tr>
<td>JMA, Hitachi; TiTech</td>
<td>JP</td>
<td>ASUCA</td>
<td>(i) OpenACC; (ii) Hybrid-Fortran</td>
</tr>
</tbody>
</table>

## Regional

<table>
<thead>
<tr>
<th>Organization</th>
<th>Location</th>
<th>Model</th>
<th>GPU Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bull, MFR</td>
<td>FR</td>
<td></td>
<td>OpenACC</td>
</tr>
<tr>
<td>JMA, Hitachi; TiTech</td>
<td>JP</td>
<td>ASUCA</td>
<td>(i) OpenACC; (ii) Hybrid-Fortran</td>
</tr>
</tbody>
</table>
ECMWF Project for Exascale Weather Prediction

Expectations towards Exascale: Weather and Climate Prediction – P. Bauer, ECMWF, EESI-2015

ESCAPE*, Energy efficient SCalable Algorithms for weather Prediction at Exascale:
- Next generation IFS numerical building blocks and compute intensive algorithms
- Compute/energy efficiency diagnostics
- New approaches and implementation on novel architectures
- Testing in operational configurations

From Berkeley Dwarfs for Numerical Computing ... 

EASCAPE HPC Goals:
- Standardized, highly optimized kernels on specialized hardware
- Overlap of communication and computation
- Compilers/standards supporting portability

Project Approach:
- Co-design between domain scientists, computer scientists, and HPC vendors

*To be funded by EC H2020 framework, Future and Emerging Technologies – High-Performance Computing; Partners: ECMWF, Météo-France, RMI, DMI, Meteo Swiss, DWD, Loughborough U, PSNC, Bull, NVIDIA, Optalysys

15
### NOAA NGGPS: NH Model Dycore Candidates (5)

<table>
<thead>
<tr>
<th>Model</th>
<th>Organization</th>
<th>Numeric Method</th>
<th>Grid</th>
</tr>
</thead>
<tbody>
<tr>
<td>NIM</td>
<td>NOAA/ESRL</td>
<td>Finite Volume</td>
<td>Icosahedral</td>
</tr>
<tr>
<td>MPAS</td>
<td>NCAR/LANL</td>
<td>Finite Volume</td>
<td>Icosahedral/Unstructured</td>
</tr>
<tr>
<td>NEPTUNE</td>
<td>Navy/NRL</td>
<td>Spectral Element</td>
<td>Cubed-Sphere with AMR</td>
</tr>
<tr>
<td>HIRAM/FV3</td>
<td>NOAA/GFDL</td>
<td>Finite Volume</td>
<td>Cubed-Sphere, nested</td>
</tr>
<tr>
<td>NMM-UJ</td>
<td>NOAA/EMC</td>
<td>Finite difference</td>
<td>Cubed-Sphere</td>
</tr>
<tr>
<td>GFS-NH</td>
<td>NOAA/EMC</td>
<td>Semi-Lagrangian/Spectral</td>
<td>Reduced Gaussian</td>
</tr>
<tr>
<td>IFS (RAPS13)</td>
<td>ECMWF</td>
<td>Semi-Lagrangian/Spectral</td>
<td>Reduced Gaussian</td>
</tr>
</tbody>
</table>
**MeteoSwiss GPU-Driven Weather Prediction**

---

**MeteoSwiss COSMO NWP Configurations Since 2008**

Before GPUs

- IFS from ECMWF
  - 2 per day, 10 day forecast

- COSMO 7 (6.6 KM)
  - 3 per day, 3 day forecast

- COSMO 2 (2.2 KM)
  - 8 per day, 24 hr forecast

---

**MeteoSwiss COSMO NWP Configurations During 2016**

With GPUs

- IFS from ECMWF
  - 2 per day, 10 day forecast

- COSMO E (2.2 KM)
  - 2 per day, 5 day forecast

- COSMO 1 (1.1 KM)
  - 8 per day, 24 hr forecast

---

“New configurations of higher resolution and ensemble predictions possible owing to the performance-per-energy gains from GPUs” – X. Lapillonne, MeteoSwiss; EGU Assembly, Apr 2015
# WRF GPU Speedups of Physics Modules

<table>
<thead>
<tr>
<th>WRF Module</th>
<th>GPU Speedup (w-wo I/O)</th>
<th>Technical Paper Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kessler MP</td>
<td>70x / 816x</td>
<td>J. Comp. &amp; GeoSci., 52, 292-299, 2012</td>
</tr>
<tr>
<td>Purdue-Lin MP</td>
<td>156x / 692x</td>
<td>SPIE: doi:10.1117/12.901825</td>
</tr>
<tr>
<td>WSM 3-class MP</td>
<td>150x / 331x</td>
<td>JSTARS, 5, 1256-1265, 2012</td>
</tr>
<tr>
<td>WSM 5-class MP</td>
<td>202x / 350x</td>
<td>SPIE: doi:10.1117/12.976908</td>
</tr>
<tr>
<td>Eta MP</td>
<td>37x / 272x</td>
<td>JSTARS, 5, 625-633, 2012</td>
</tr>
<tr>
<td>WSM 6-class MP</td>
<td>165x / 216x</td>
<td>Submitted to J. Comp. &amp; GeoSci.</td>
</tr>
<tr>
<td>Goddard GCE MP</td>
<td>348x / 361x</td>
<td>Accepted for publication in JSTARS</td>
</tr>
<tr>
<td>Thompson MP*</td>
<td>76x / 153x</td>
<td>JSTARS, 5, 1256-1265, 2012</td>
</tr>
<tr>
<td>SBU 5-class MP</td>
<td>213x / 896x</td>
<td>JSTARS, 5, 625-633, 2012</td>
</tr>
<tr>
<td>WDM 5-class MP</td>
<td>147x / 206x</td>
<td>J. Atmos. Ocean. Tech., 30, 2896, 2013</td>
</tr>
<tr>
<td>WDM 6-class MP</td>
<td>150x / 206x</td>
<td>JSTARS, 5, 555-562, 2012</td>
</tr>
<tr>
<td>RRTMG LW*</td>
<td>123x / 127x</td>
<td>JSTARS, 7, 3660-3667, 2014</td>
</tr>
<tr>
<td>RRTMG SW*</td>
<td>202x / 207x</td>
<td>Submitted to J. Atmos. Ocean. Tech.</td>
</tr>
<tr>
<td>Goddard SW</td>
<td>92x / 134x</td>
<td>JSTARS, 5, 555-562, 2012</td>
</tr>
<tr>
<td>Dudhia SW*</td>
<td>19x / 409x</td>
<td>Submitted to JSTARS</td>
</tr>
<tr>
<td>MYNN SL</td>
<td>6x / 113x</td>
<td>Submitted to GMD</td>
</tr>
<tr>
<td>TEMF SL</td>
<td>5x / 214x</td>
<td>Submitted to JSATRS</td>
</tr>
<tr>
<td>Thermal Diffusion LS</td>
<td>10x / 311x</td>
<td>Submitted to JSATRS</td>
</tr>
<tr>
<td>YSU PBL*</td>
<td>34x / 193x</td>
<td>Submitted to GMD</td>
</tr>
</tbody>
</table>

* Modules for initial NVIDIA-funded integration project

## Hybrid WRF Customer Benchmark Capability

**Staring in 2H 2015**

**Hardware and Benchmark Case**

- **CPU**: Xeon Core-i7 3930K, 1 core use;
- **Benchmark**: CONUS 12 km for 24 Oct 24; 433 x 308, 35 levels
WRF Modules Available in OpenACC

- Project to integrate CUDA modules into latest full model WRF 3.6.1
- Several dynamics routines including all of advection
- Several physics schemes:
  - Planetary boundary layer – YSU, GWDO
  - Cumulus – KFETA
  - Microphysics – Kessler, Morrison, Thompson
  - Radiation – RRTM
  - Surface physics – Noah

```plaintext
dyn_em/module_advect_em.OpenACC.F
dyn_em/module_bc_em.OpenACC.F
dyn_em/module_big_step_utilities_em.OpenACC.F
dyn_em/module_diffusion_em.OpenACC.F
dyn_em/module_em.OpenACC.F
dyn_em/module_first rk step part1.OpenACC.F
dyn_em/module_first rk step part2.OpenACC.F
dyn_em/module_small step em.OpenACC.F
dyn_em/module_stoch.OpenACC.F
dyn_em/solve_em.OpenACC.F
dyn_em/start_em.OpenACC.F
frame/module_dm.OpenACC.F
frames/module_domain_extra.OpenACC.F
frames/module_domain.OpenACC.F
frames/module_domain_types.OpenACC.F
phys/module_bl_gwdo.OpenACC.F
phys/module_bl_ysu.OpenACC.F
phys/module_cu_kfeta.OpenACC.F
phys/module_microphysics_driver.OpenACC.F
phys/module_microphysics_zero_out.OpenACC.F
phys/module_mp_kessler.OpenACC.F
phys/module_mp_morr_two_moment.OpenACC.F
phys/module_mp_thompson.OpenACC.F
phys/module_pbl driver.OpenACC.F
phys/module_physics_addtendc.OpenACC.F
phys/module_physics_init.OpenACC.F
phys/module_ra_rrtm.OpenACC.F
phys/module_ra_sw.OpenACC.F
phys/module_sf_noahslm.OpenACC.F
phys/module_sf_sfclayrev.OpenACC.F
share/module_bc.OpenACC.F
share/wrf_bdyin.OpenACC.F
```
GPUs a mainstream platform for the HPC data center

More CPU choices: x86, ARM64, OpenPOWER (NVLink)

NVIDIA focus on applications: Deep Learning; Earth Science
Thank You

Stan Posey; sposey@nvidia.com; NVIDIA, Santa Clara, CA, USA