GPU TECHNOLOGY:
PAST, PRESENT, FUTURE

Marc Hamilton, Vice President,
Solution Architecture and Engineering
A Decade Of GPU Computing

- From Scientific Computing To Machine Learning
- Mobile Is More Than Just Phones
- GPU Architecture & CUDA Roadmap
- Grid & The Last Mile of Virtualization
From Scientific Computing To Machine Learning
Giving Drones the Vision to Help Fight Fires

A Breakthrough in HIV Research

Early, Accurate Detection of Breast Cancer
<table>
<thead>
<tr>
<th>Rank</th>
<th>NodeID</th>
<th>Site</th>
<th>Computer</th>
<th>Total Power (kW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>358.02</td>
<td>Keys Center, Tokyo Institute of Technology</td>
<td>TSUBAME KFC - LX-10 HET/HT7/HT8/HT10 Cluster, Intel Xeon E5-2620 v3 2.40GHz in warm, NVIDIA K20</td>
<td>34.48</td>
</tr>
<tr>
<td>2</td>
<td>167.78</td>
<td>Cambridge University</td>
<td>XSEDE - Peer Cluster, Intel Xeon E5-2695 v2 2.40GHz in warm, NVIDIA K20</td>
<td>53.02</td>
</tr>
<tr>
<td>3</td>
<td>578.28</td>
<td>Center for Computational Sciences, University of Tsukuba</td>
<td>KASPAR TCA - Cepheus 26256 - 80K Cluster, Intel Xeon E5-2690 v3 2.60GHz in warm, NVIDIA K20</td>
<td>79.77</td>
</tr>
<tr>
<td>4</td>
<td>459.41</td>
<td>SFB/Fairs</td>
<td>Carini Accelerator - Bnl BK5 cluster, Intel Xeon E5-2678 v3 2.50GHz in warm, NVIDIA K20</td>
<td>44.42</td>
</tr>
<tr>
<td>5</td>
<td>316.91</td>
<td>Serbia National Supercomputing Centre (CSCS)</td>
<td>SC4Agni - Corp yCIC, Intel Xeon E5-2670 v3 2.40GHz in warm, NVIDIA K20</td>
<td>1.735.68</td>
</tr>
<tr>
<td>6</td>
<td>317.86</td>
<td>ROMEO HPC Center - Champagne-Ardenne</td>
<td>xenon - Bull R4F-63 Cluster, Intel Xeon E5-2690 v3 2.60GHz in warm, NVIDIA K20</td>
<td>81.41</td>
</tr>
<tr>
<td>7</td>
<td>319.72</td>
<td>CIERC</td>
<td>CIERC GPU16x8 - 64K in warm, NVIDIA K20, Intel Xeon E5-2690 v3 2.60GHz in warm, NVIDIA K20</td>
<td>82.02</td>
</tr>
<tr>
<td>8</td>
<td>807.15</td>
<td>GIG Center, Tokyo Institute of Technology</td>
<td>TSUBAME 2.5 - Cluster Pabellon 02, 1024K in warm, Intel Xeon E5-2670 v2 2.50GHz in warm, NVIDIA K20</td>
<td>327.08</td>
</tr>
<tr>
<td>9</td>
<td>813.14</td>
<td>Exploration &amp; Production - OPE S.A.</td>
<td>KPC2 - Graphite DEMI4, Intel Xeon E5-2680 v3 2.90GHz in warm, NVIDIA K20</td>
<td>1.074.49</td>
</tr>
<tr>
<td>10</td>
<td>2675.41</td>
<td>France/Institution</td>
<td>Grafia 5 6M, Intel Xeon E5-2690 v3 2.80GHz in warm, NVIDIA K20</td>
<td>84.65</td>
</tr>
<tr>
<td>11</td>
<td>2628.42</td>
<td>France/Institution</td>
<td>Grafia 5 6M, Intel Xeon E5-2690 v3 2.80GHz in warm, NVIDIA K20</td>
<td>85.25</td>
</tr>
<tr>
<td>12</td>
<td>2628.42</td>
<td>France/Institution</td>
<td>Grafia 5 6M, Intel Xeon E5-2690 v3 2.80GHz in warm, NVIDIA K20</td>
<td>85.25</td>
</tr>
<tr>
<td>13</td>
<td>2628.42</td>
<td>France/Institution</td>
<td>Grafia 5 6M, Intel Xeon E5-2690 v3 2.80GHz in warm, NVIDIA K20</td>
<td>85.25</td>
</tr>
<tr>
<td>14</td>
<td>2628.42</td>
<td>France/Institution</td>
<td>Grafia 5 6M, Intel Xeon E5-2690 v3 2.80GHz in warm, NVIDIA K20</td>
<td>85.25</td>
</tr>
<tr>
<td>15</td>
<td>2628.42</td>
<td>Max-Planck-Gesellschaft MPG</td>
<td>Grafia 5 6M, Intel Xeon E5-2690 v3 2.80GHz in warm, NVIDIA K20</td>
<td>269.04</td>
</tr>
</tbody>
</table>

TSUBAME KFC
#1 OF “TOP 15” GREEN SUPERCOMPUTERS
POWERED BY CUDA GPUS
MACHINE LEARNING

Branch of Artificial Intelligence
Computers that learn from data
THREE TRENDS CONVERGING

- Torrent of Data
- Deep Neural Networks
- GPU Computing

Exabytes of unstructured data

SOURCE: IDC
Deep Learning with COTS HPC Systems
A. Coates, B. Huval, T. Wang, D. Wu, A. Ng, B. Catanzaro
Stanford / NVIDIA • ICML 2013

“Now You Can Build Google’s $1M Artificial Brain on the Cheap”

-Wired
**CUDA FOR MACHINE LEARNING**

<table>
<thead>
<tr>
<th>Early Adopters</th>
<th>Use Cases</th>
<th>Prominent Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adobe</td>
<td>Image Detection</td>
<td>facebook</td>
</tr>
<tr>
<td></td>
<td>Speech/Image Recognition</td>
<td>STANFORD UNIVERSITY</td>
</tr>
<tr>
<td>Baidu</td>
<td>Face Recognition</td>
<td>DARPA</td>
</tr>
<tr>
<td></td>
<td>Gesture Recognition</td>
<td>NYU</td>
</tr>
<tr>
<td>Flickr</td>
<td>Video Search &amp; Analytics</td>
<td>DENSO</td>
</tr>
<tr>
<td></td>
<td>Speech Recognition &amp; Translation</td>
<td>Carnegie Mellon University</td>
</tr>
<tr>
<td>Netflix</td>
<td>Recommendation Engines</td>
<td>MIT Massachusetts Institute of Technology</td>
</tr>
<tr>
<td>Yandex</td>
<td>Indexing &amp; Search</td>
<td>Berkeley University</td>
</tr>
</tbody>
</table>

**Use Cases**
- Image Detection
- Face Recognition
- Gesture Recognition
- Video Search & Analytics
- Speech Recognition & Translation
- Recommendation Engines
- Indexing & Search
Mobile - More Than Just Phones
TEGRA TK1

UNIFIED ARCHITECTURE

TEGRA K1 - MOBILE SUPER CHIP

BREAKTHROUGH EXPERIENCES
JETSON TK1 DEV KIT
1ST MOBILE SUPERCOMPUTER FOR EMBEDDED SYSTEMS

192 CUDA cores
326 GFLOPS
VisionWorks SDK
EVOLUTION OF COMPUTING IN THE CAR

Virtual Cockpit

Infotainment

Autonomous Driving

Tegra 3

Tegra 4

Tegra K1
TEGRA TK1 SUPERCOMPUTER FOR DRIVER ASSISTANCE

Pedestrian Detection  
Blind Spot Monitoring  
Lane Departure Warning

Collision Avoidance  
Traffic Sign Recognition  
Adaptive Cruise Control

Optical Flow  
Histogram  
Feature Detection
COMPUTER VISION ON CUDA

Feature Detection / Tracking
~30 GFLOPS @ 30 Hz

Object Recognition / Tracking
~180 GFLOPS @ 30 Hz

3D Scene Interpretation
~280 GFLOPS @ 30 Hz
GPU Architecture & CUDA Roadmap
BANDWIDTH BOTTLENECKS

<table>
<thead>
<tr>
<th>Component</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCI Express</td>
<td>16GB/sec</td>
</tr>
<tr>
<td>CPU Memory</td>
<td>60GB/sec</td>
</tr>
<tr>
<td>GPU Memory</td>
<td>288GB/sec</td>
</tr>
</tbody>
</table>
INTRODUCING NVLINK

Differential with embedded clock
PCle programming model (w/ DMA+)
Unified Memory
Cache coherency in Gen 2.0
5 to 12X PCle
5X More Bandwidth for Multi-GPU Scaling
3D MEMORY

3D Chip-on-Wafer integration
Many X bandwidth
2.5X capacity
4X energy efficiency
PASCAL

- **NVLink**: 5 to 12X PCIe 3.0
- **3D Memory**: 2 to 4X memory BW & size
- **Module**: 1/3 size of PCIe card

![Diagram of PASCAL module with components labeled: GPU Chip, Power Regulation, HBM Stacks.](image)
CUDA-ENABLED GPUS: 522M
CUDA DOWNLOADS: 2.5M
ACADEMIC PAPERS: 58K
UNIVERSITY COURSES: 770

CUDA EVERWHERE
GOALS FOR THE CUDA PLATFORM

Simplicity
- Learn, adopt, & use parallelism with ease

Productivity
- Quickly achieve feature & performance goals

Portability
- Write code that can execute on all targets

Performance
- High absolute performance and scalability
UNIFIED MEMORY
DRAMATICALLY LOWER DEVELOPER EFFORT

Developer View Today

Developer View With Unified Memory

System Memory

GPU Memory

Unified Memory
REMOTE DEVELOPMENT TOOLS

- Local IDE, remote application
  - Edit locally, build & run remotely
  - Automatic sync via ssh
  - Cross-compilation to ARM

- Full debugging & profiling via remote connection
EXTENDED (XT) LIBRARY INTERFACES

Automatic Scaling to multiple GPUs per node

cuFFT 2D/3D & cuBLAS level 3

Operate directly on large datasets that reside in CPU memory

developer.nvidia.com/cublasxt

16K x 16K SGEMM on Tesla K10
GRID Graphics Accelerated VDI
The Original Graphics GPU Returns To The Data Center
NVIDIA Grid
GPUs Power
Enterprise
Virtualization
2.0
IMPORTANCE OF A GPU

COMMERCIAL MARKETS

- DESIGNERS: 25M
- POWER USERS: 200M
- KNOWLEDGE WORKERS: 400M
- TASK WORKERS: 100M

MUST HAVE

- 3D Engineering & Design Apps
  - Autodesk
  - SolidWorks
  - PTC
- PLM & Volume Design
  - Teamcenter
  - Autodesk
- Media Rich Web
  - HTML5

INCREASINGLY NICE TO HAVE

- Office Productivity
  - Cerner
  - Epic
- Medical Records
  - McKesson
  - Siemens
NIGHT AND DAY DIFFERENCE

Without GPU

With GPU
GRID ACCELERATED GRAPHICS
Thank You