GEFORCE® GTX TITAN

The Ultimate CUDA Development GPU
Introducing GeForce GTX TITAN
The Ultimate CUDA Development GPU

2688 CUDA Cores
4.5 Teraflops Single Precision
1.27 Teraflops Double Precision
288 GB/s Memory Bandwidth
GTX TITAN

Personal Supercomputer on Your Desktop

1 Teraflop < $1000

Develop Anywhere

Ease of Programming with New Kepler Architecture
The Best of Kepler in a PC

Boosts PC with 8x More Performance

More Science, Less Coding
Dynamic Parallelism Makes Parallel Programming Easier

**Quicksort**

Before Kepler

```c
_device_ Workstack stack;
__global__ void quicksort(int *data, int left, int right)
{
    int nleft, nright;
    // Partitions data based on pivot of first element.
    // Returns counts in nleft & nright
    partition(data, left, right, data[left], nleft, nright);
    // If a sub-array needs sorting, push it on the stack
    if(left < nright) stack.push(data, left, nright);
    if(nleft < right) stack.push(data, nleft, right);
}

_host_ void launch_quicksort(int *data, int count)
{
    // Launch initial quicksort to populate the stack
    quicksort<<< ... >>>(data, 0, count-1);
    // Loop more quicksorts until finished
    while(1)
    {
        // Wait for all sorts to complete
        cudaDeviceSynchronize();
        // Copy our stack from the Workstack
        stack_copy = CopyFromDevice(stack);
        // Count of things on stack
        if(stack_copy.size() == 0) break;
        // Pop the stack and launch
        while(stack_copy.size())
        {
            WorkStack elem = stack_copy.top();
            quicksort<<< ... >>>(data, elem.left, elem.right);
        }
    }
}
```

With Kepler

Easier porting for existing codes

No complex CPU & GPU interaction

Easier code in half the lines
2x More Applications, More Customers

- **Irregular Work**
  - Direct N-body
  - Monte Carlo
  - Reverse Time Migration
- **Structured Work**
  - Sparse Matrix
  - Video Transcoding
  - Algebraic Multigrid
- **Data Parallel**
  - Adaptive Mesh (CFD)
  - N-Body Tree Codes (Astrophysics)
  - Sparse Matrix
  - Video Transcoding
  - Algebraic Multigrid
- **Task Parallel**
  - Divide and Conquer (Big Data)
  - Burrows-Wheeler Aligner (Bioinformatics)
## Comparing GTX TITAN and Tesla K20X

<table>
<thead>
<tr>
<th>Features</th>
<th>GeForce GTX TITAN</th>
<th>Tesla K20X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core/Mem clock</td>
<td>837MHz/3GHz (clocks may vary when double precision is on)</td>
<td>732MHz/2.6GHz</td>
</tr>
<tr>
<td>Peak Single Precision</td>
<td>~4.5 Tflops</td>
<td>3.95 TFlops</td>
</tr>
<tr>
<td>Peak Double Precision</td>
<td>~1.27 Tflops (estimate)</td>
<td>1.32 TFlops</td>
</tr>
<tr>
<td>Memory size</td>
<td>6 GB</td>
<td>6 GB</td>
</tr>
<tr>
<td>Memory BW (ECC off)</td>
<td>288 GB/s</td>
<td>250 GB/s</td>
</tr>
<tr>
<td>PCIe</td>
<td>Gen 3 only on Ivy Bridge, Gen 2 on Sandy Bridge</td>
<td>Gen 2</td>
</tr>
<tr>
<td>CUDA Features</td>
<td>Dynamic Parallelism, Hyper-Q For CUDA Streams, GPUDirect Peer to Peer</td>
<td>Dynamic Parallelism, Hyper-Q Proxy for MPI and CUDA Streams, GPUDirect Peer to Peer, and RDMA</td>
</tr>
<tr>
<td>GPU monitoring</td>
<td>None</td>
<td>NVML/NVSMI, OOB, InfoROM, NVHealthmon, TCC</td>
</tr>
<tr>
<td>Cluster monitoring</td>
<td>None</td>
<td>Bright Computing, Ganglia</td>
</tr>
<tr>
<td>ECC Features</td>
<td>No ECC</td>
<td>DRAM, Internal Caches &amp; Reg Files</td>
</tr>
<tr>
<td>Total Board Power</td>
<td>250W</td>
<td>235W</td>
</tr>
</tbody>
</table>
Tesla Advantage: Built for Deployment

Performance

• Fastest DP of 1.31 TFLOPS on Tesla K20X
• Optimized for Infiniband with NVIDIA GPUDirect™
• Hyper-Q for accelerating MPI based workloads
• Tuning and optimization support from NVIDIA experts

Reliability

• ECC protection
• Tested to run 24/7 with real-world workloads
• 3 year warranty and support for bugs/feature requests
• ISVs certify only on Tesla
• NVIDIA technical support
• Longer life cycle for continuity and cluster expansion

Built for HPC

• Integrated solutions & support from OEMs and channel
• Tools for GPU Management (Nvhealthmon, nvidia-smi/nvml)
• Tools for Cluster Management
• Enterprise OS support
• Solution expertise provided by CUDA engineers and technical staff
• Designed, tested and optimized for cluster deployment
Develop with GeForce, Deploy with Tesla
Call to Action

• Promote GTX Titan with OpenACC to new developers
  • Increases # of GPU enabled workloads

• Continue recommending Tesla for workloads in production environment
  • Tesla Advantage: Performance + Reliability + Built for HPC

• Refer to GTX Titan and Tesla FAQ for differentiation
  • GPUDirect, Hyper-Q, GPUBoost, GPUMonitoring
Thank You