Cluster Design
Data Center Infrastructure

- System Solutions
- Communication
- Infrastructure Management

- GPU Accelerators
- Interconnect
- System Management

- GPU Boost
- GPU Direct
- NVLink
- NVML

Development

- Programming Languages
- Development Tools
- Software Solutions

- Compiler Solutions
- Profile and Debug
- Libraries

- C/C++
- Fortran
- OpenACC
- Python

- PGI
- VAMPiR

- LLVM
- CUDA Debugging API
- cuBLAS
TESLA: PLATFORM WITH OPEN ECOSYSTEM

Pick the CPU that’s Correct for You

Libraries
- AmgX
- cuDNN
- cuBLAS
- OpenCV
- Thrust
- cuBLAS

Compiler Directives
- OpenACC

Programming Languages
- C/C+
- Fortran
- Python
- Java
- x86
DESIGNING FOR WORKLOAD

Multiple Ways to Balance Parallel and Serial Performance

2 GPUs per CPU

3 GPUs per CPU

4 GPUs per CPU

GPU

CPU

GPU

GPU

GPU

GPU

GPU

80GB/s

16GB/s

20GB/s

80GB/s

40 GB/s

40 GB/s

40 GB/s

NVLink

PCIe Gen3 x16

PCIe Switch
IN SITU VIS - FASTER SCIENCE, LOWER COST

Traditional Workflow

CPU Supercomputer
Simulation - 1 Week

Viz Cluster
Viz - 1 Day

Multiple Iterations

Time to Discovery = Months

Tesla Platform

Faster Time to Discovery
Reduced Pressure on Filestore

GPU-Accelerated Supercomputer

Visualize while you simulate

Restart Simulation Instantly
Multiple Iterations

Time to Discovery = Weeks
Cluster Deployment
NODE TOPOLOGY CIRCA 2008

Life Was Good

- Memory
- GPU
- Bridge
- Memory
- CPU
- CPU
NODE TOPOLOGY IN 2014 AND BEYOND

Better, but More Choices
TOPOLOGY-AWARE RESOURCE MANAGERS

$\texttt{CUDA\_VISIBLE\_DEVCIES}$ today, cgroups coming soon
Cluster Management
TESLA GPU DEPLOYMENT KIT

Command-line and Library

- nvidia-smi provides a command-line interface
- NVML provides API access

COMPUTE MODE

Insurance Against Scheduling Issues

The Compute Mode setting controls simultaneous use

- **DEFAULT** allow multiple simultaneous processes
- **EXCLUSIVE_THREAD** allows only one context
- **EXCLUSIVE_PROCESS** one process, but multiple threads
- **PROHIBITED**

Can be set by command-line (nvidia-smi) & API (NVML)
RESOURCE LIMITS

Use cgroups, Not ulimit

- UVA depends on allocating virtual address space

- Virtual address space ≠ physical ram consumption

Several batch systems and resource managers support cgroups, either directly or via plugins.
MONITORING

Open Source, Commercial, DIY

- Environmental and utilization metrics are available
- Tesla cards may provide OoB access to telemetry via BMC
- NVML support has been integrated into many monitoring systems
GPU ACCOUNTING

- Per-process accounting of GPU usage by PID

- Accessible by library (NVML) or command-line (nvidia-smi)

- Enable accounting mode:
  ```
  sudo nvidia-smi -am 1
  ```

- Human-readable accounting
  ```
  nvidia-smi -q -d ACCOUNTING
  ```

- CSV accounting data
  ```
  nvidia-smi -query-accounted-apps=gpu_name,gpu_util -format=csv
  ```
Summary
Pick the HW mix that’s right for your site

Topology Matters in design and in resource management

Rich ecosystem of management hooks and tools
Thanks