GPU TECHNOLOGY CONFERENCE

Clusters with GPUs under Linux and Windows HPC

Massimiliano Fatica (NVIDIA), Calvin Clark (Microsoft)

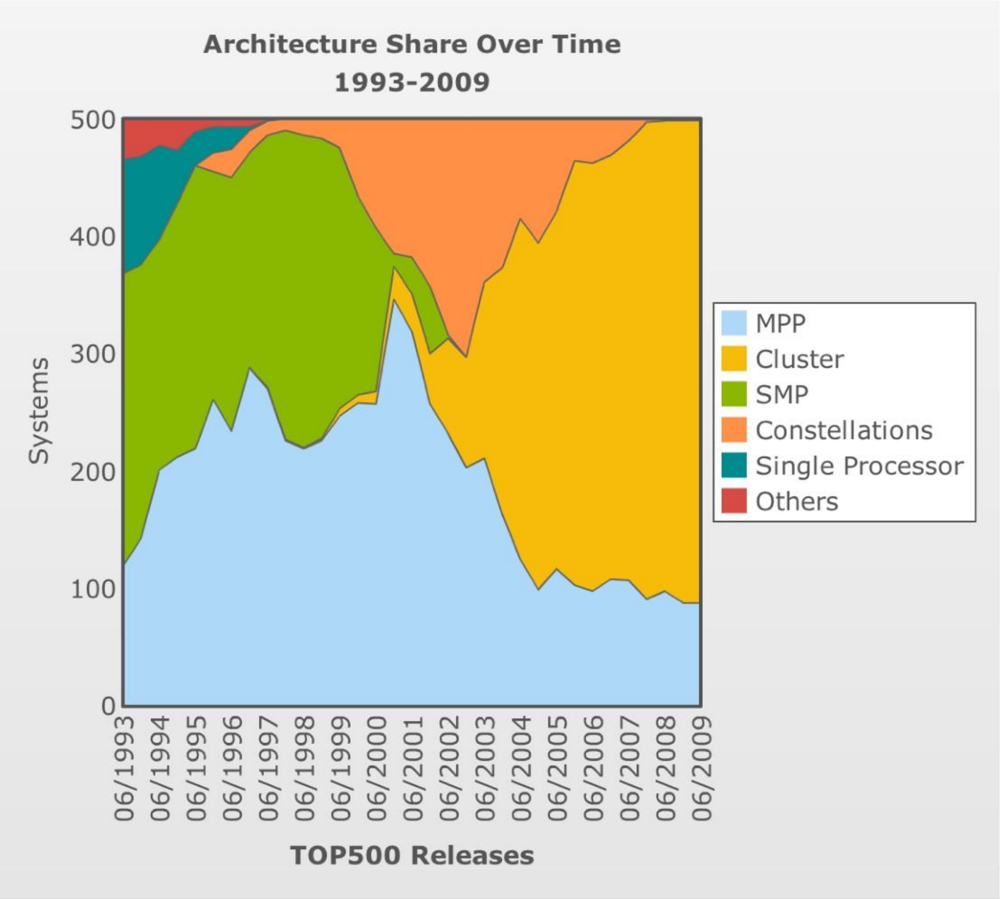
Hillsborough Room | Oct 2 2009

NVIDIA.

Agenda

- Overview
- Requirements for GPU Computing
- Linux clusters
- Windows HPC clusters
- Q & A

HPC Clusters



- Clusters are very popular in HPC due to their flexible configurations and excellent price/performances

- Clusters with GPUs are the latest trend (Titech, NCSA, CEA, CSIRO, Petrobras, HESS, Bloomberg, …)

HPC Clusters with GPUs

- The right configuration is going to be dependent on the workload

- NVIDIA Tesla GPUs for cluster deployments:
  - Tesla GPU designed for production environments
  - Memory tested for GPU computing
  - Tesla S1070 for rack-mounted systems
  - Tesla M1060 for integrated solutions

- Minimum requirements for clusters with Tesla S1070 GPUs:
  - One CPU core per GPU
  - One PCI-e x8 slot (x16 Gen2 highly recommended)

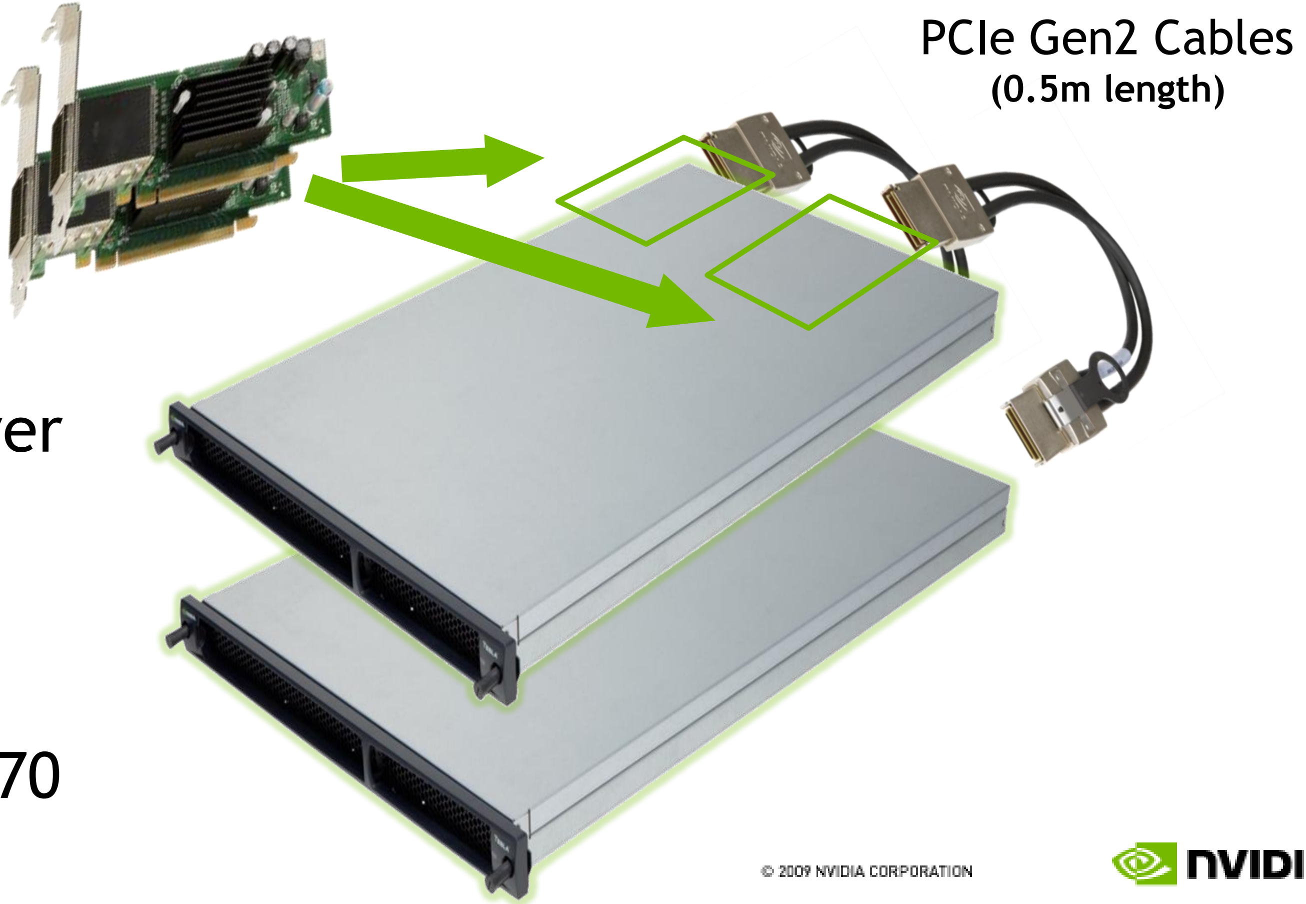
CUDA software requirements

- **Driver**: required component to run CUDA applications.
- **Toolkit**: compiler, runtime, libraries (BLAS and FFT).
- **SDK**: collection of examples and documentation

Downloadable from http://www.nvidia.com/cuda

Connecting Tesla S1070 to hosts

Host Interface Cards (HIC) or Graphic Host Interface Cards (GHIC): PCI-e Gen2 x8 or x16



PCIe Gen2 Cables
(0.5m length)

Host Server

Tesla S1070

Linux for GPU clusters

Deploying CUDA on Linux clusters

Several cluster management systems are now CUDA enabled
(Rocks, Platform Computing, Clustervision, Scyld Clusterware)

If you want to deploy on your preferred software stack:

-Install and load the driver on each node (required):

- Use the silent and no-network flags in the installer (-s -N)
- Use the script in the release notes to load the driver without having to start X

-Install the toolkit on each node (required)

-Install the SDK on each node (optional):

- Use deviceQuery and bandwidthTest to check that the GPUs are properly configured and the bandwidth is consistent among nodes (--noprompt flag)

System management for Tesla S1070

nvidia-smi is a software tool providing:

- Thermal Monitoring:

  GPU temperatures, chassis inlet/outlet temperatures

- System Information:

  Unit serial number, firmware revision, configuration info

- System Status

  System fan states (e.g. failure), GPU faults

  Power system fault , cable fault

Exclusive access mode

nvidia-smi can set up access policies for the GPUs:

```
#nvidia-smi --loop-continuously --interval=60 --filename=/var/log/nvidia.log &

#nvidia-smi -g 0 -c 1 (Set GPU 0 in exclusive access mode)
#nvidia-smi -g 1 -c 1 (Set GPU 1 in exclusive access mode)

#nvidia-smi -g 1 -s
                    Compute-mode rules for GPU=0x1: 0x1
#nvidia-smi -g 0 -s
                    Compute-mode rules for GPU=0x0: 0x1
```

This simplify interaction with job scheduling (GPUs become consumable resources, similar to tapes and licenses)

Windows HPC for GPU clusters

Current limitation:
Requires an NVIDIA GPU for the display (S1070 + GHIC) or an host system graphic chipset with WDDM driver

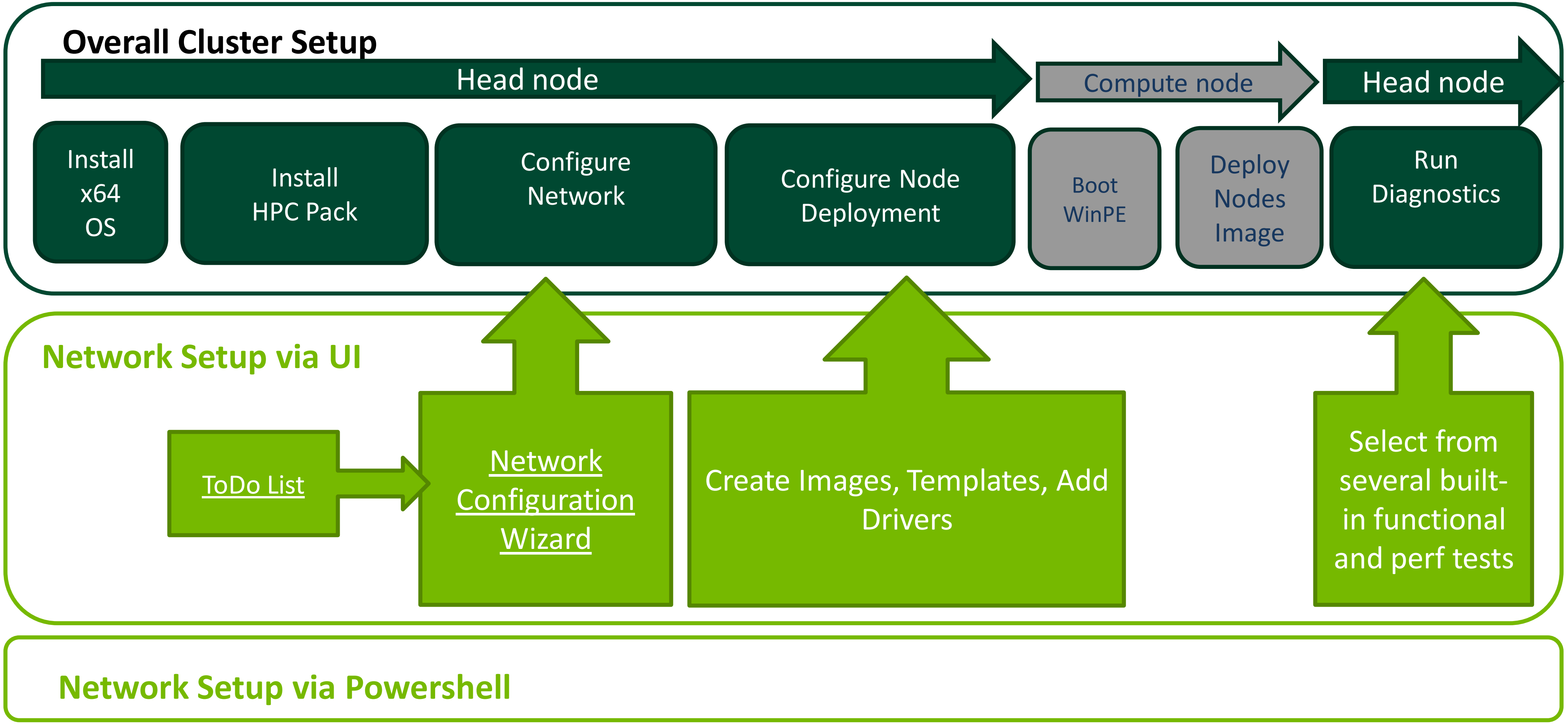
What is Windows HPC Server?

- Windows HPC Server consists of:
  - A Windows Server x64 OS installation
    - An inexpensive SKU called “HPC Edition” can be volume-licensed for clusters dedicated to HPC applications
  - The HPC Pack, which provides services, tools and runtime environment support for HPC applications
    - Management
    - Job Scheduling
    - Diagnostics
    - MPI Stack

Deployment Means…

1. Getting the OS Image on the machines. Options?
   - Manual Installation
   - 3rd-Party Windows Deployment Tools
     - Includes some solutions for mixed Linux/Windows clusters
   - PXE Boot from Head Node
2. Configuring the HPC Pack
   - Step-by-step wizard for interactive installation
   - XML-based configuration for automated, reproducible deployments

Deployment Process for Network Boot



Overall Cluster Setup

Head node

Compute node

Head node

Install x64 OS

Install HPC Pack

Configure Network

Configure Node Deployment

Boot WinPE

Deploy Nodes Image

Run Diagnostics

Network Setup via UI

ToDo List

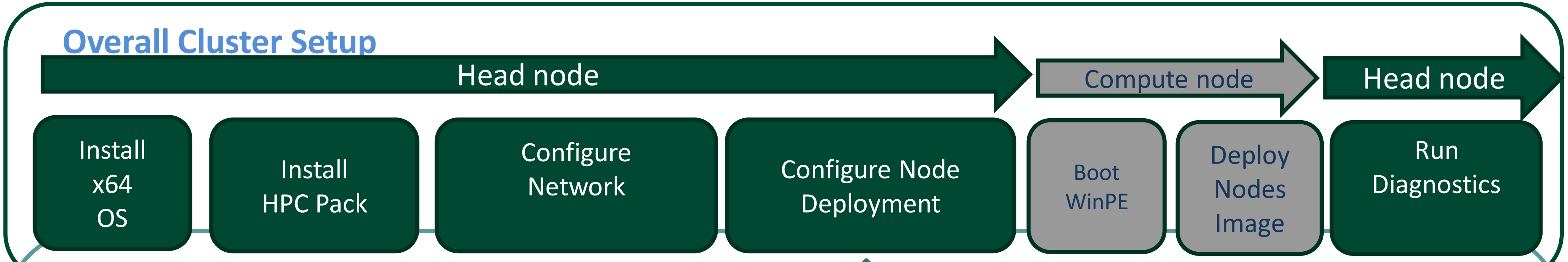
Network Configuration Wizard

Create Images, Templates, Add Drivers

Select from several built-in functional and perf tests

Network Setup via Powershell

Network Drivers Management



Network Driver Setup

HW cards and drivers must be installed manually on head node for each network (public, private, MPI)

Use Deployment Management UI to install and update network drivers on nodes (includes locating and staging).

Driver packaging format	WinPE	OS	Mechanisms beneath UI
.inf	UI	UI	.WIM injection, WDS, PnP
.MSI	N/A	UI	via .MSI Inst (no .MSI inst for WinPE)

A word about WDS

- Windows Deployment Service
  - Standard Deployment Solution for Network Boot of current Windows Client and Server Operating Systems
  - Supports Multicast IP
- HPC Pack automatically configures and drives WDS for common HPC cluster scenarios
  - Substantially reduced learning curve
  - Many admins don’t need to learn anything about WDS to deploy their first Windows HPC cluster

Images, drivers, and all that

- Adding drivers is easy!
  - Browse to location where .INF files are stored, and drivers will automatically be injected into the images to be deployed (*unpack or install the driver on the head node*)
- Some advanced users might like to learn about ImageX, WinPE and the WAIK (Windows Automated Installation Kit)
  - Capturing “golden image” of compute node
  - Pre-configuring Windows Roles and Features, and settings, OEM Branding

After Deployment…

- Custom software can be added with post-deployment steps, configurable in the UI and persisted in XML
  - It helps to have an unattended command-line way to install the software, to avoid prompting user
  - Examples:
    - Windows Debugger
    - CUDA Toolkit, CUDA SDK
- Steps to control how and when OS patches are deployed are built into the management tools

CUDA Toolkit

- To automate the toolkit deployment
  - Generate a setup.iss file:

    cudatoolkit -r

    This will generate the setup.iss in C:\Windows
  - Use this file for unattended installation on all the nodes:

    cudatoolkit -s -f1”fullpathto\file.iss”

- Same steps for the CUDA SDK.

Looking forward

- Windows HPC Server “V3”
  - CTP2 is now available for trial at http://connect.microsoft.com
  - Future builds will include many enhancements, but of particular interest are:
    - Diskless boot via iSCSI
    - Improved support for > 1000 node deployments
    - Extensibility of Diagnostics for software and hardware partners

Leveraging the GPU

- A special environment variable is needed:
    - **HPC_ATTACHTOCONSOLE** ***or***
    - **HPC_ATTACHTOSESSION**
  - – Required because normally, Session 0 processes like Windows Services can’t access the GPU
  - – *Session 0 != Console Session any more*
  - – ***User launching job needs to be logged in to corresponding session***
- *Example:*
  - – `Job submit /env:HPC_ATTACHTOCONSOLE=TRUE mygpgpu.exe`
- See whitepaper at http://resourcekit.windowshpc.net titled “GPU Computing in Windows HPC Server 2008” for tips on automating this

Ongoing Windows HPC GPU Work

- Accelerator
  - High-level data-parallel library written in C# that can be called from .Net applications
  - Leverages DirectX
  - http://research.microsoft.com/en-us/projects/Accelerator/
- Tokyo Institute of Technology
  - Drs. Satoshi Matsuoka and Yutaka Akiyama
  - 32-node Windows HPC Cluster
  - Using CUDA for Advanced Structural Proteomics
- High Performance Discrete Fourier Transforms on Graphics Processors
  - Naga K. Govindaraju et al. (SuperComputing 2008)

Thank You!

- www.nvidia.com/cuda
- www.microsoft.com/hpc
- Special thanks to:
  - Microsoft Team
    - Ryan Waite, Ken Oien, Alex Sutton, Parmita Mehta, Wenming Ye, Joshua Shane, Kerry Hammil, Chris Crall and many others
- Q&A