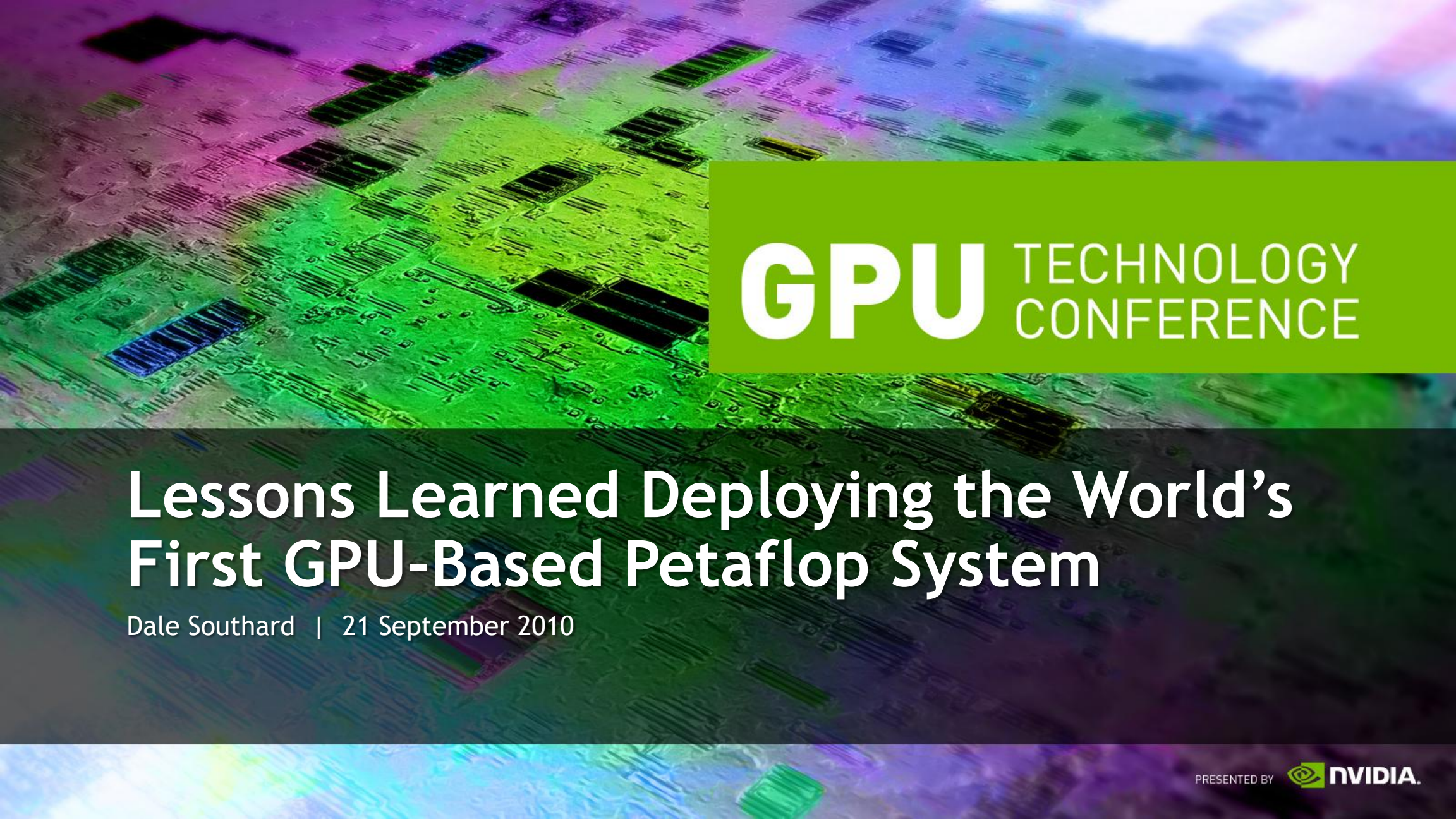

Lessons Learned Deploying the World's First GPU-Based Petaflop System

Dale Southard | 21 September 2010

About the Speaker and You

[Dale] is a senior solution architect with NVIDIA (aka a professional debugger). In the past I was a HW architect in the LLNL systems group designing the vis/post-processing solutions and on-call for capability systems. I spent three weeks in Tianjin China assisting with Nebula benchmarking.

You are here because you are interested in what it takes to deploy large HPC systems that use GPUs for compute acceleration.

Introduction and a Disclaimer

Nebulae Summary

- First GPU Computing system to exceed one petaflop
- First deployment to use Fermi-based Tesla™
- COTS –based
- Built in three months
- GPUs added in the last few weeks

Petaflop -- so easy that everyone can do it!

COTS Clustering by Size

- 10 nodes is easy
- 100 nodes is shrink-wrapped
- 1000 nodes requires your own tools
- Anything larger is a learning experience

Mental Preparation

- Petascale computing is difficult
  - Things that happen rarely on smaller systems happen constantly
  - Things that should never occur do occur

Expected Linpack perf for a new petascale system is zero.

Choose the Correct Components

Teslas™ Teslas™ Everywhere

GPUs are a good fit for HPC, but choose the right form-factor

- Cxxxx are workstation products
- Mxxxx are server products
- Sxxxx are external systems

The Thing You Plug a GPU Into….

Sever-class node hardware

- Continuous operation
- Efficient racking
- Management features

Everything Else

- Cluster fabric
- Management network
- Diskfull/diskless
- Physical plant (footprint, power, cooling)
- Infrastructure (fileservice, authentication, etc)
- Maintenance/repair

Get Your Tools in Order

Parallel Shell

pdsh is the predominant “parallel distributed shell”

- supports backend remote shell systems like ssh and rsh
- Includes `dshbak` to consolidate output

Console Output

- Netconsole is insufficient
- Persistently log all console output from all nodes
- Provide interactive use as well as logging

System Monitoring

- You need it
- Should provide a persistent record
- Lightweight and out-of-band whenever possible

ECC

- You want EDAC (aka Bluesmoke)
- EDAC state is in the file below, you want “S4ECD4ED”, also known as chipkill

/sys/devices/system/edac/mc/mc*/csrow*/edac_mode

- EDAC also detects PCI errors:

modprobe edac_mc

echo 1 > /sys/devices/system/edac/pci/check_pci_parity

Resource Manager

- SLURM, RMS, others
- Lower level than a batch system
- Provide a way to quickly remove bad nodes from service

Things That Should Not Happen

Nodes Fail in Strange Ways

- CPU speed (/proc/cpuinfo)
- CPU corecount (/proc/cpuinfo)
- BIOS version (dmidecode)
- Correct amount of memory (free)
- Is ECC still working (told you already, pay attention)
- Hard disk errors (something besides S.M.A.R.T.)

You need to proactively check for failures.

Infiniband

- Firmware version (ibstat)
- Correct link speed (ibstat, state AND rate)

Keeping IB fabric “clean” can be a major challenge…

Job Setup and Cleanup

- Cgroups (containers) may be helpful
- Timeshare scheduling is asking for trouble
- CUDA exclusive mode is provided for a reason
- Resource Managers

Performance Tuning

Homework Assignment

"The Case of the Missing Supercomputer Performance"
(Petrini et al)

Node Quiescing

- Stop all unnecessary daemons
- Beware of your monitoring software
- If you have to monitor in-band, coalesce
- Synchronicity

Processor Affinity

- Multi-socket nodes are NUMA, not shared memory
- Ditto for multi-bridge (multi-IOH) designs
- Trend appears to be towards less uniformity
- Some RM's now support affinity, more tools on the way

The Horrors

Part Tracking

- Labels
- Process
- History

Out of Memory

- OOM is no longer a recoverable condition
- OOM-killer is insufficient

PCI Errors

- PCI transport is not infallible
- Without EDAC, you may not know

MCElog

Designed to move some ECC/health functions into userspace.

- Latency issues
- General notification problems
- Hair-pulling issues

Light at the End of the Tunnel?