BEST PRACTICES FOR DESIGNING, DEPLOYING, AND MANAGING GPU CLUSTERS

Dale Southard

Cluster Design

Data Center Infrastructure

System Solutions

CRAY
DELL
IBM
hp
amazon webservices EC2

Communication

Mellanox TECHNOLOGIES
MVAPICH
OPEN MPI

Infrastructure Management

Adaptive COMPUTING
Bright Computing
IBM PLATFORM COMPUTING

Development

Programming Languages

C/C++
Fortran
OpenACC
python

Development Tools

allinea DDT
PGI
VAMPIR

Software Solutions

Kitware
MATLAB
ROGUE WAVE SOFTWARE

GPU Accelerators

GPU Boost
…

Interconnect

GPU Direct
NVLink
…

System Management

NVML
…

Compiler Solutions

LLVM
…

Profile and Debug

CUDA Debugging API
…

Libraries

cuBLAS
…

Tesla Accelerated Computing Platform

TESLA: PLATFORM WITH OPEN ECOSYSTEM

Pick the CPU that's Correct for You

Libraries

Compiler Directives

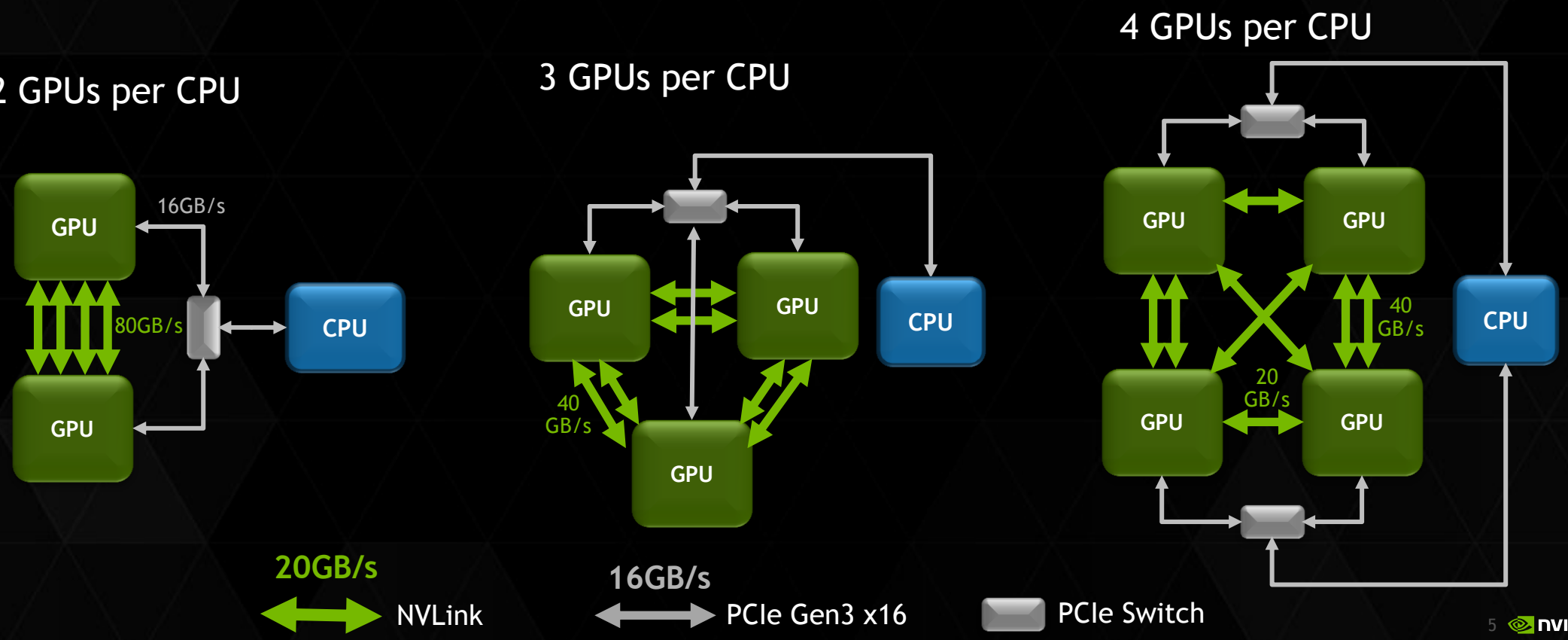
OpenACC

Programming Languages

C/C++ Fortran python Java

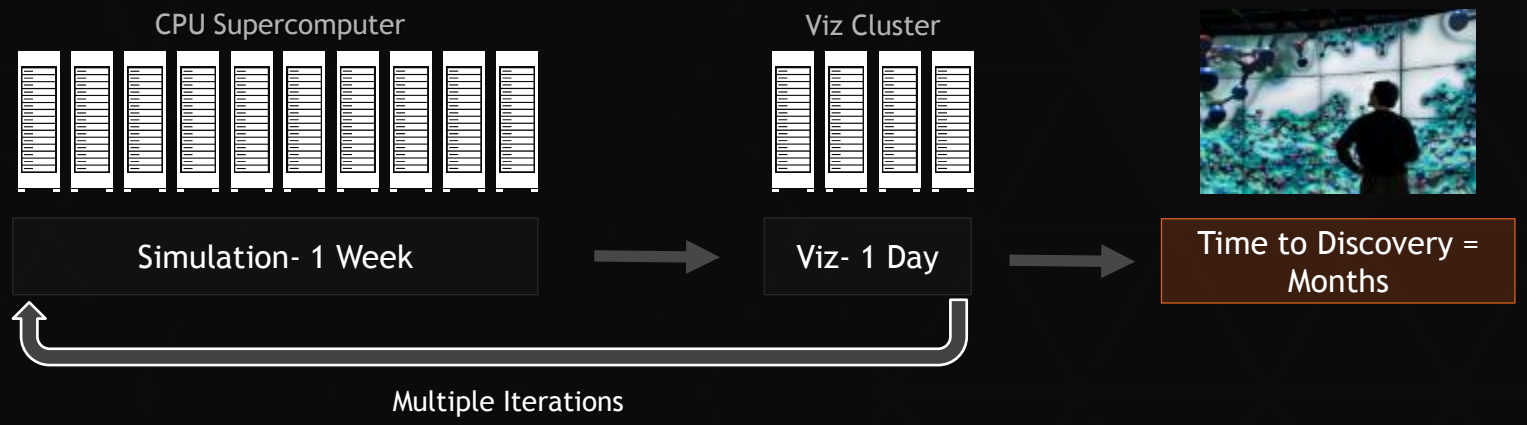
DESIGNING FOR WORKLOAD

Multiple Ways to Balance Parallel and Serial Performance



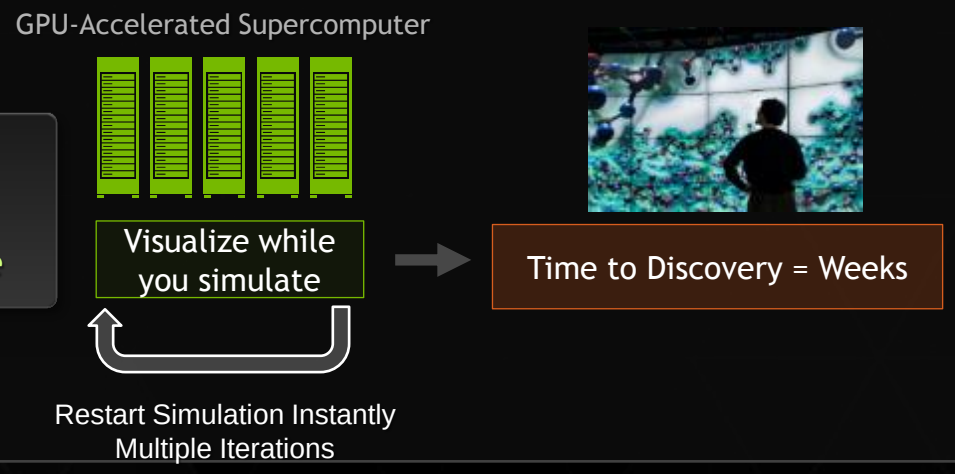
IN SITU VIS – FASTER SCIENCE, LOWER COST

Traditional Workflow



Tesla Platform

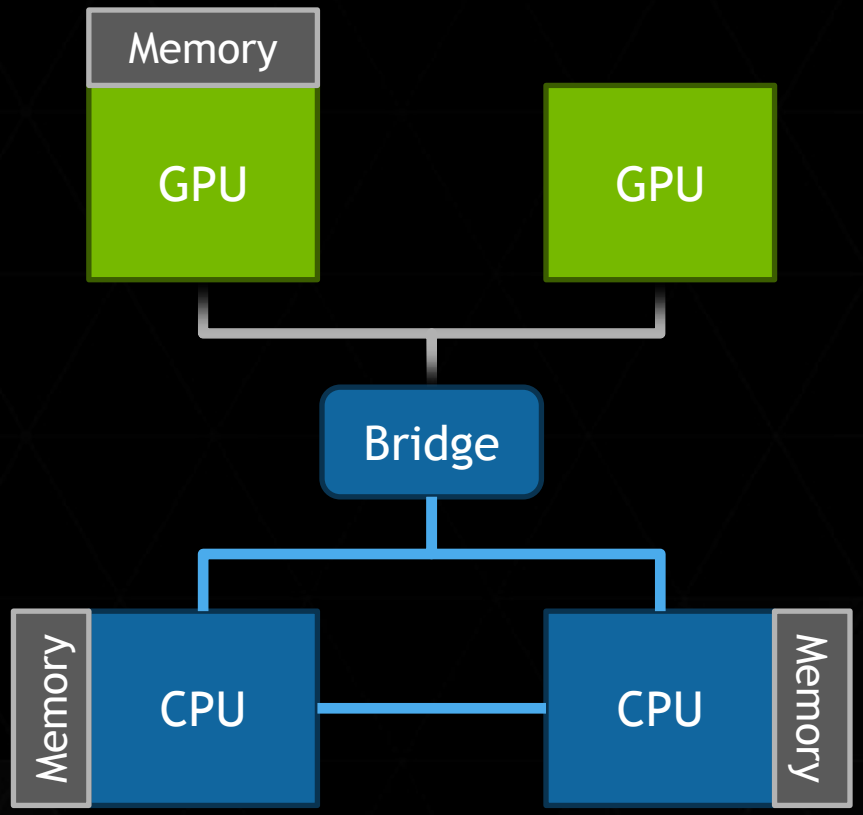
Faster Time to Discovery
Reduced Pressure on Filestore



Cluster Deployment

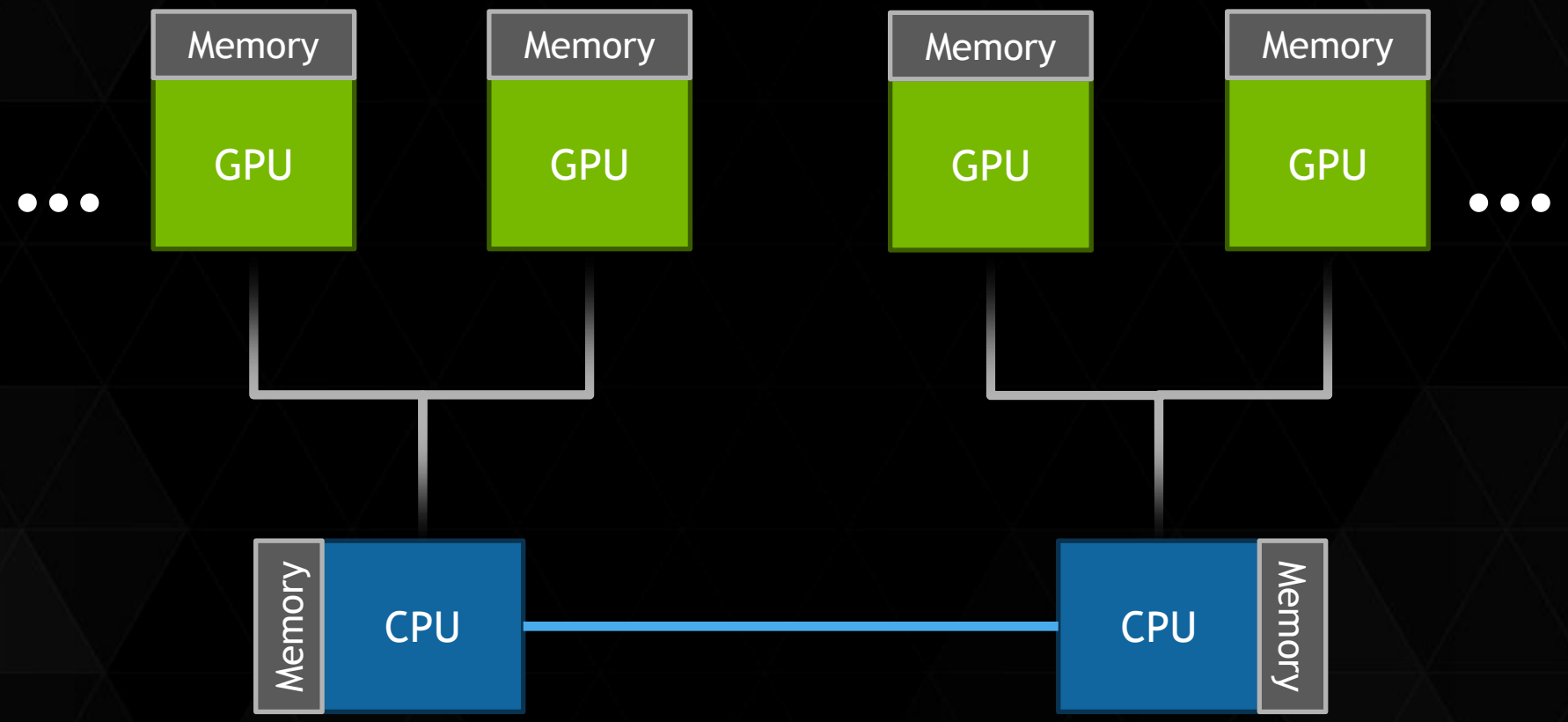
NODE TOPOLOGY CIRCA 2008

Life Was Good



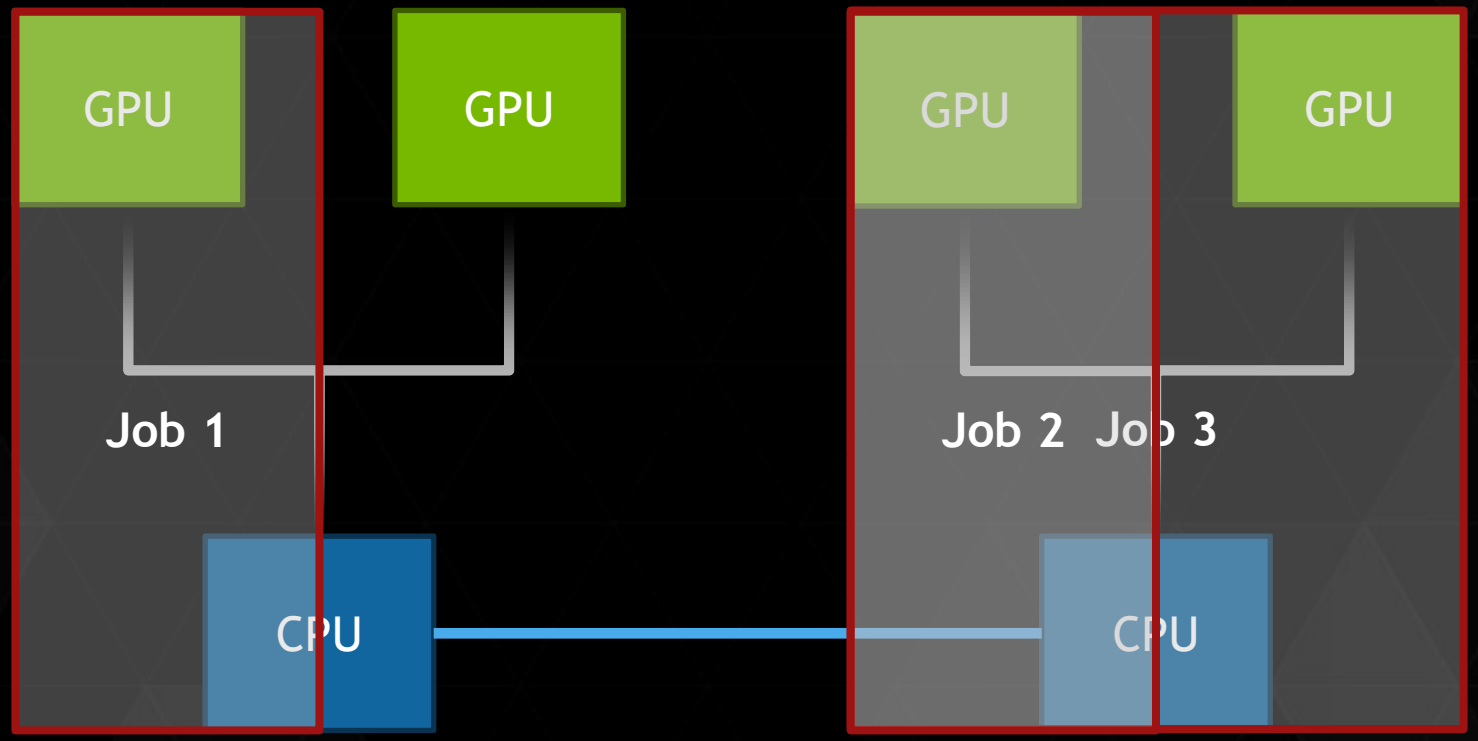
NODE TOPOLOGY IN 2014 AND BEYOND

Better, but More Choices



TOPOLOGY-AWARE RESOURCE MANAGERS

$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

https://developer.nvidia.com/gpu-deployment-kit

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 mangers 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

TESLA ACCELERATED COMPUTING PLATFORM

Libraries, Languages, Tools

- 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