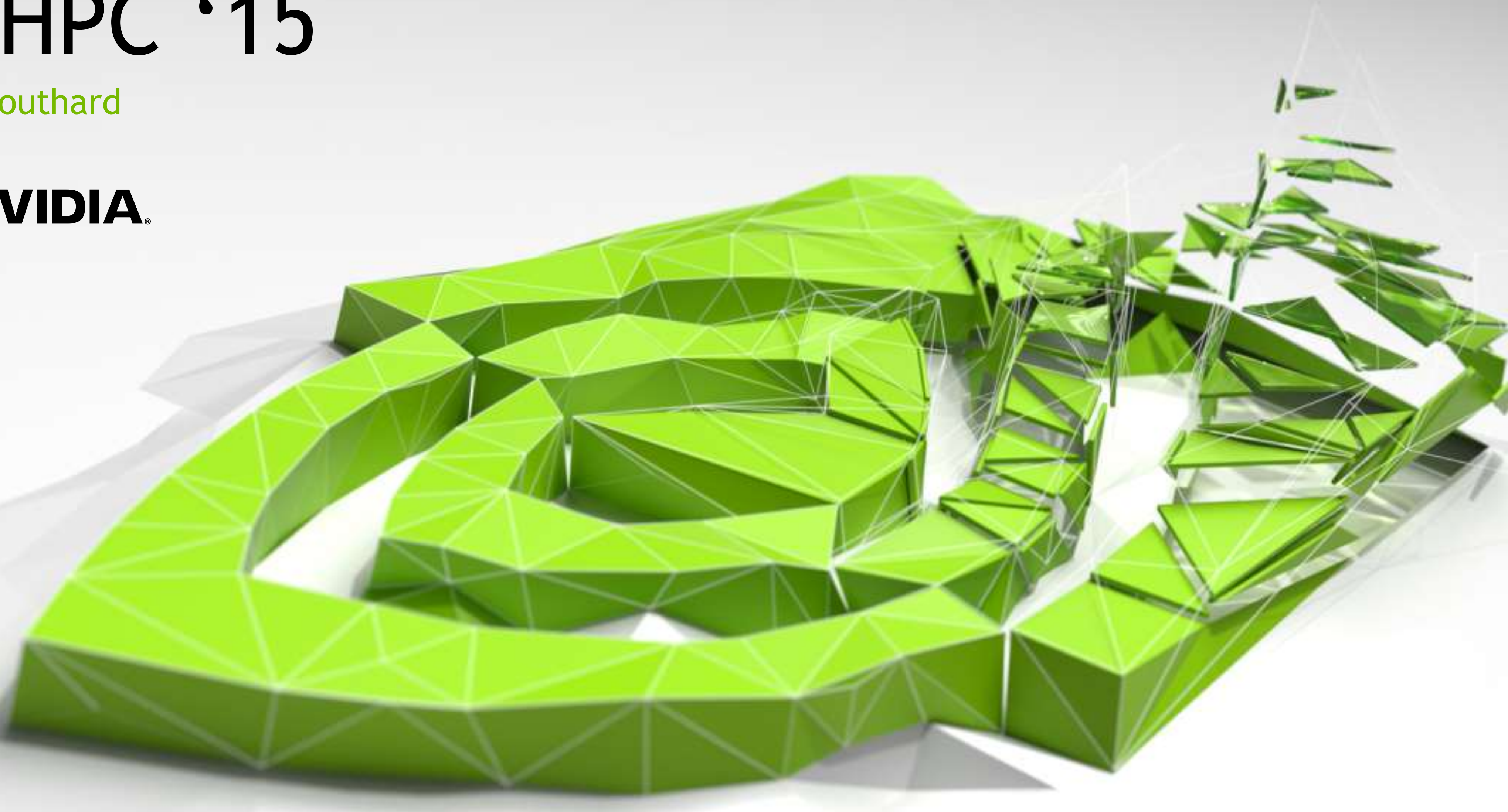
DEHPC '15

Dale Southard

NVIDIA



GAMING

DESIGN

DESKTOP
VIRTUALIZATION

HPC & CLOUD
SERVICE PROVIDERS

AUTONOMOUS
MACHINES

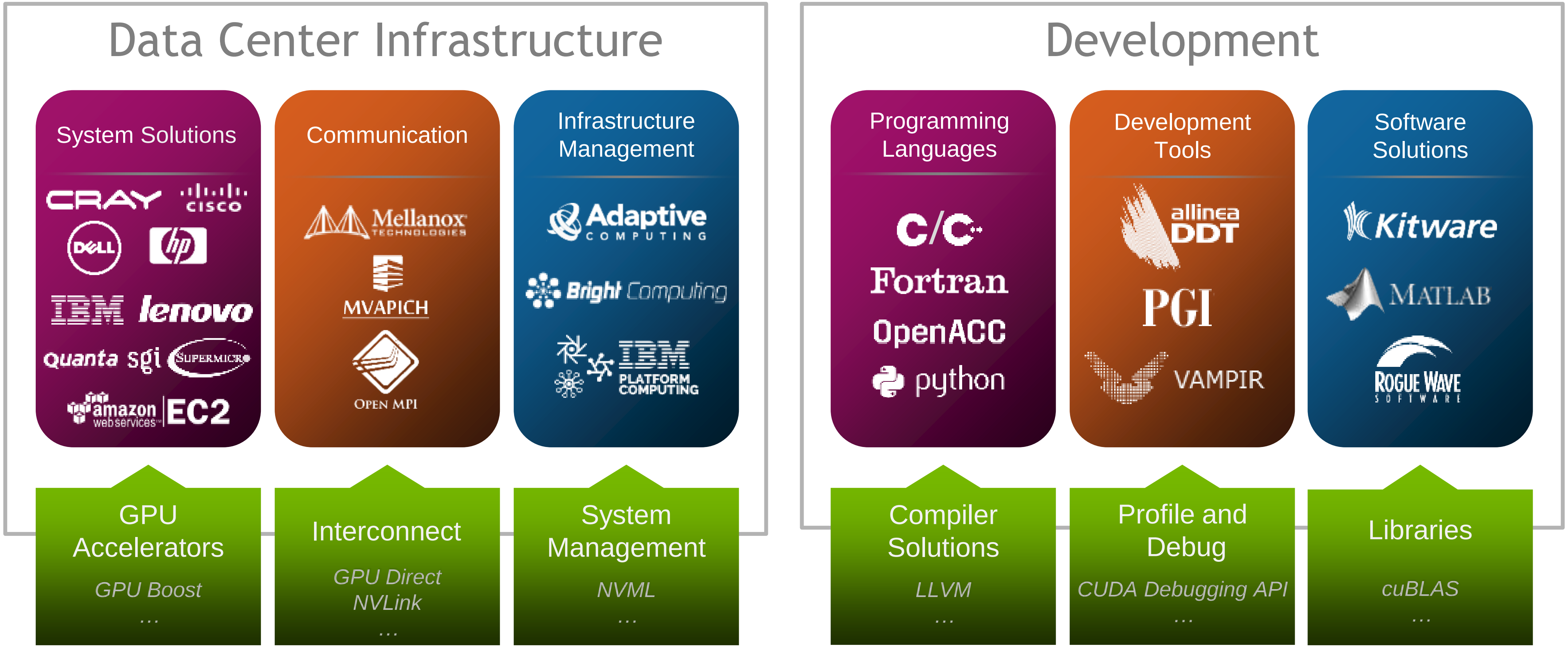
PC

DATA CENTER

MOBILE

The World Leader in Visual Computing

Tesla Accelerated Computing Platform



Data Center Infrastructure

System Solutions

CRAY, cisco, DELL, hp, IBM, lenovo, Quanta, sgi, SUPERMICRO, amazon webservices EC2

Communication

Mellanox Technologies, MVAPICH, Open MPI

Infrastructure Management

Adaptive Computing, Bright Computing, IBM Platform Computing

GPU Accelerators

GPU Boost

…

Interconnect

GPU Direct

NVLink

…

System Management

NVML

…

Development

Programming Languages

C/C++, Fortran, OpenACC, python

Development Tools

allinea DDT, PGI, VAMPIR

Software Solutions

Kitware, MATLAB, Rogue Wave Software

Compiler Solutions

LLVM

…

Profile and Debug

CUDA Debugging API

…

Libraries

cuBLAS

…

“ Accelerators Will Be Installed in More than Half of New Systems ”

“In 2014, NVIDIA enjoyed a dominant market share with 85% of the accelerator market.”

Source: Top 6 predictions for HPC in 2015

Vision: Mainstream Parallel Programming

Enable more programmers to write portable parallel software in their language of choice

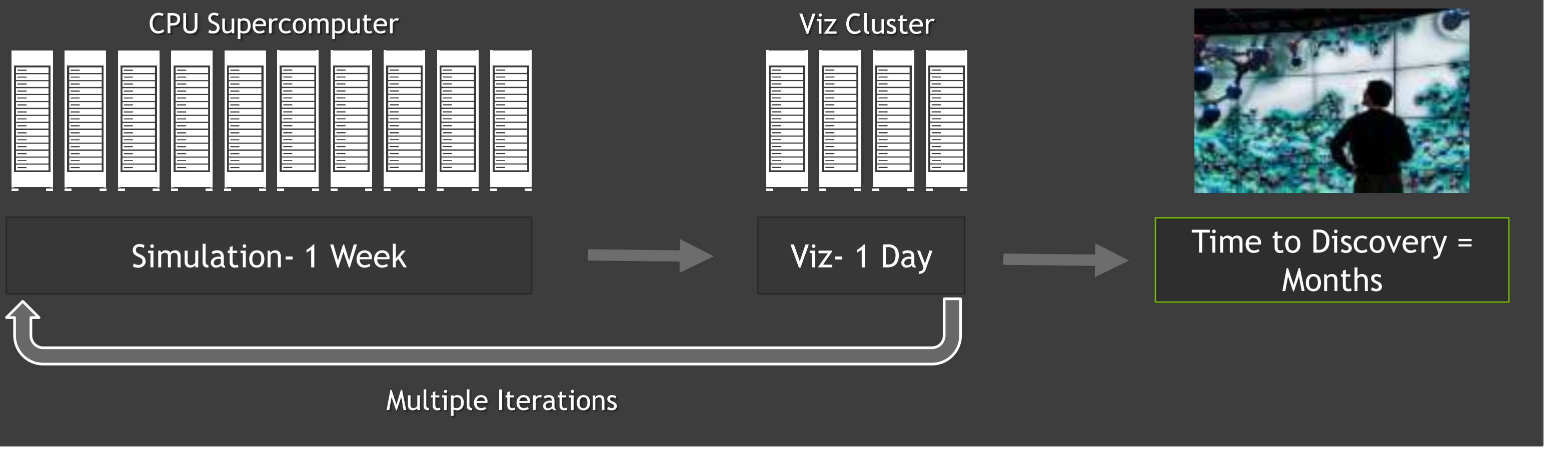
Embrace and evolve standards in key languages

CUDA continues to evolve as the target low-level platform for GPU acceleration

C
THE C++ PROGRAMMING LANGUAGE
OpenACC
Directives for Accelerators
Fortran
R
python
OpenMP
MATLAB
Java
.js
MPI

Vision: In Situ Vis – Faster Science, Lower Cost

Traditional Workflow



CPU Supercomputer

Simulation- 1 Week

Viz Cluster

Viz- 1 Day

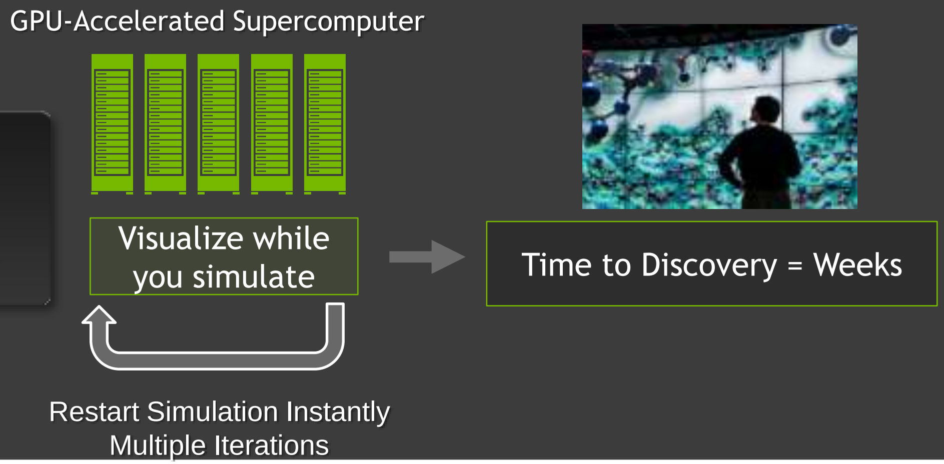
Time to Discovery = Months

Multiple Iterations

Tesla Platform

Faster Time to Discovery

Reduced Pressure on Filestore



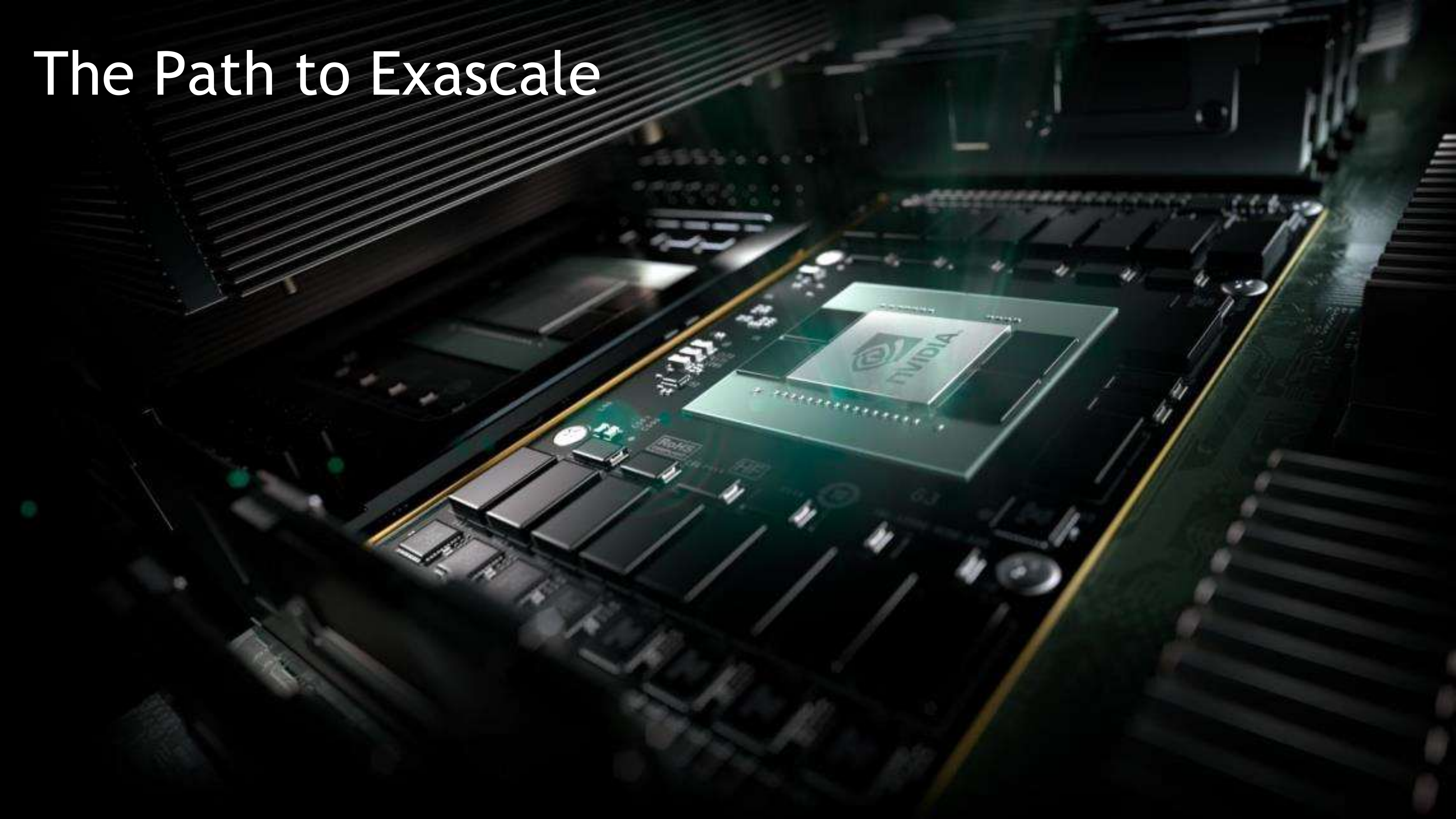
GPU-Accelerated Supercomputer

Visualize while you simulate

Time to Discovery = Weeks

Restart Simulation Instantly
Multiple Iterations

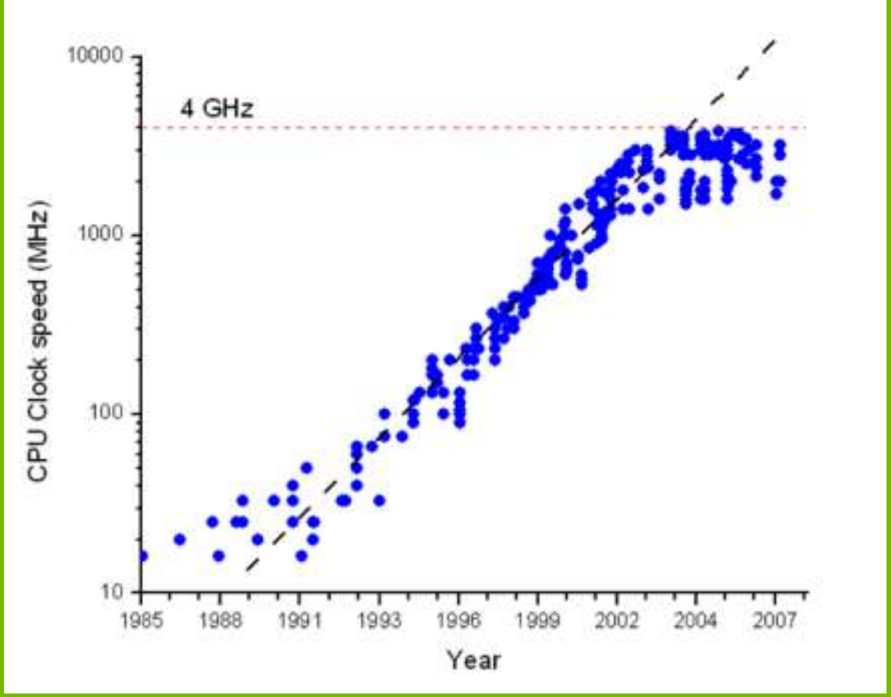
The Path to Exascale



Power for CPU-only
Exaflop Supercomputer = Power for the Bay Area, CA
(San Francisco + San Jose)

HPC’s Biggest Challenge

Hitting a Frequency Wall?



G Bell, *History of Supercomputers*, LLNL, April 2013

The End of Voltage Scaling

The Good Old Days

Leakage was not important, and voltage scaled with feature size

$L' = L/2$
$V' = V/2$
$E' = CV^2 = E/8$
$f' = 2f$
$D' = 1/L^2 = 4D$
$P' = P$

Halve L and get 4x the transistors and 8x the capability for the same power

The New Reality

Leakage has limited threshold voltage, largely ending voltage scaling

$L' = L/2$
$V' = \sim V$
$E' = CV^2 = E/2$
$f' = 2f$
$D' = 1/L2 = 4D$
$P' = 4P$

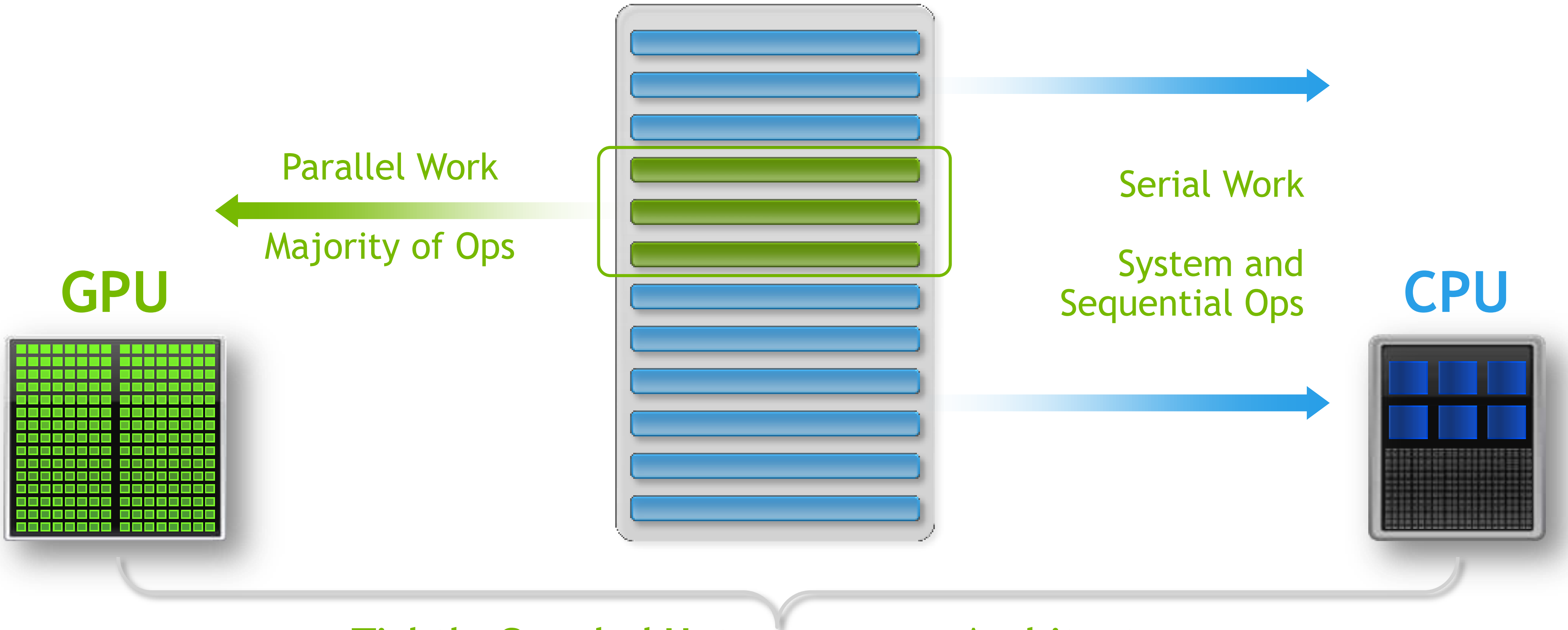
Halve L and get 4x the transistors and 8x the capability for 4x the power, or 2x the capability for the same power in ¼ the area.

“If you want to plow a field, which would you rather use, 4 strong oxen or 1024 chickens?”

- Seymour Cray, 1989

Hint: We want both.

Optimizing Serial/Parallel Execution



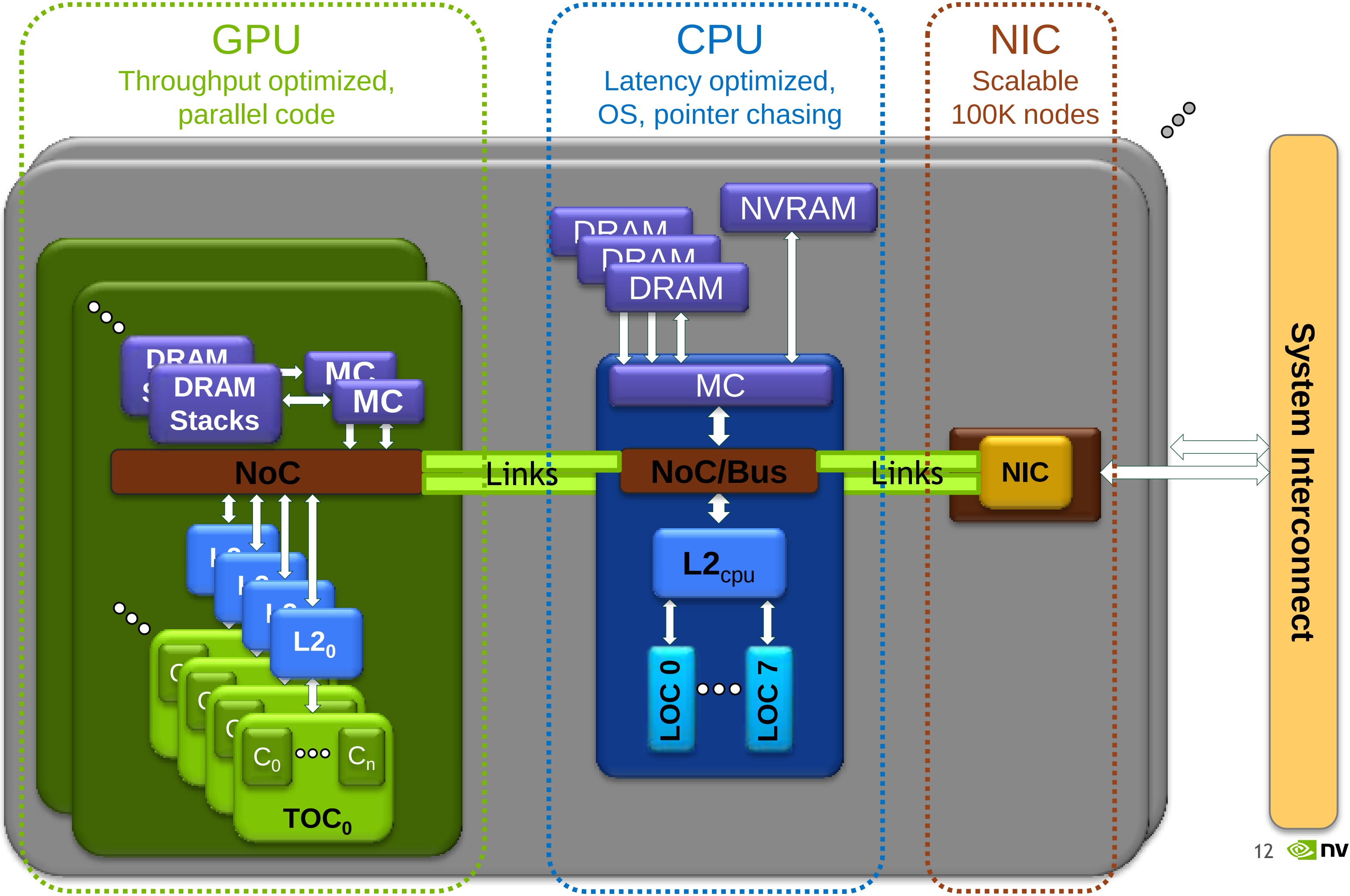
Tightly-Coupled Heterogeneous Architecture

Generic Future Node Model

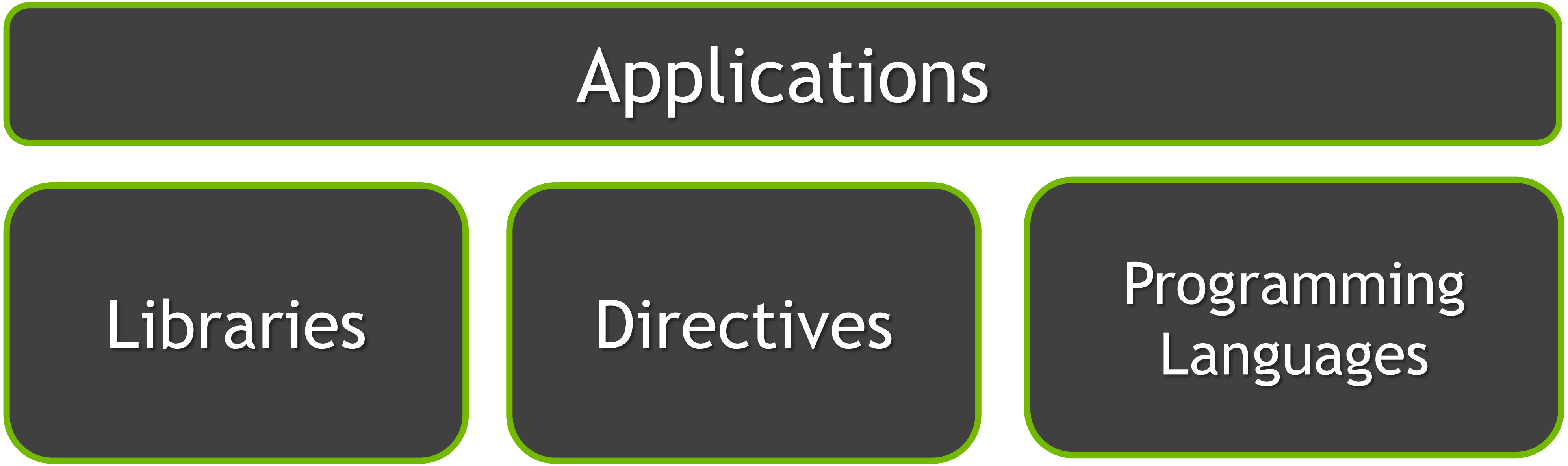
Three Building Blocks (GPU, CPU, Network)

Direct Evolution

- Programming Model Continuity
- Specialized Cores
  - GPU for parallel work
  - CPU for serial work
- Coherent memory system with Stacked, Bulk, & NVRAM
- Amortize non-parallel costs
  - Increase GPU:CPU
  - Smaller CPU
- *Can be one chip or MCM or multiple sockets*



3 Ways to Program GPUs



Applications

Libraries

Directives

Programming Languages

“Drop-in” Acceleration

Easily Accelerate Applications

Maximum Flexibility

GPU DEVELOPER ECO-SYSTEM

Numerical Packages

MATLAB
Mathematica
NI LabView
pyCUDA

Debuggers & Profilers

cuda-gdb
NV Visual Profiler
Parallel Nsight
GPU Wizard
Allinea
TotalView

GPU Compilers

C
C++
Fortran
Java
Python

Auto-parallelizing & Cluster Tools

OpenACC
GPUDirect
OpenMP
Ocelot

Libraries

BLAS
FFT
LAPACK
NPP
Video
Imaging
GPULib

Consultants & Training

acceleware
GASS
STONE RIDGE TECHNOLOGY
ANEO
GPU Tech
TECH
ArrayFire
HPC PROJECT
EM Photonics
SCALABLE GRAPHICS
WIPRO

OEM Solution Providers

DELL
hp
IBM
CRAY
ASUS
SUPERMICRO
sgi
FUJITSU
Bull
APPRO
lenovo 联想
NEC

DEVELOP ON GEFORCE, DEPLOY ON TESLA

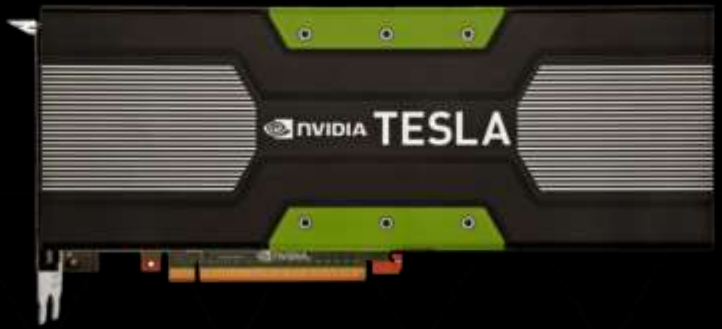
GeForce GPUs

Designed for Gamers & Developers

Available Everywhere

https://developer.nvidia.com/cuda-gpus

Tesla K40/K80



Designed for Cluster Deployment

ECC
24x7 Runtime
GPU Monitoring
Cluster Management
GPUDirect-RDMA
Hyper-Q for MPI
3 Year Warranty
Integrated OEM Systems, Professional Support

CUDA: WORLD'S MOST PERVASIVE PARALLEL PROGRAMMING MODEL

14,000	Institutions with CUDA Developers
2,000,000	CUDA Downloads
487,000,000	CUDA GPUs Shipped

700+ University Courses
In 62 Countries

NVIDIA