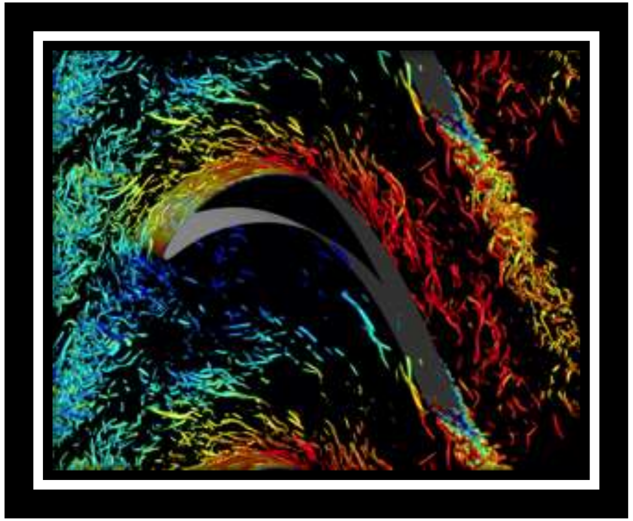
The Accelerated Road to Exascale

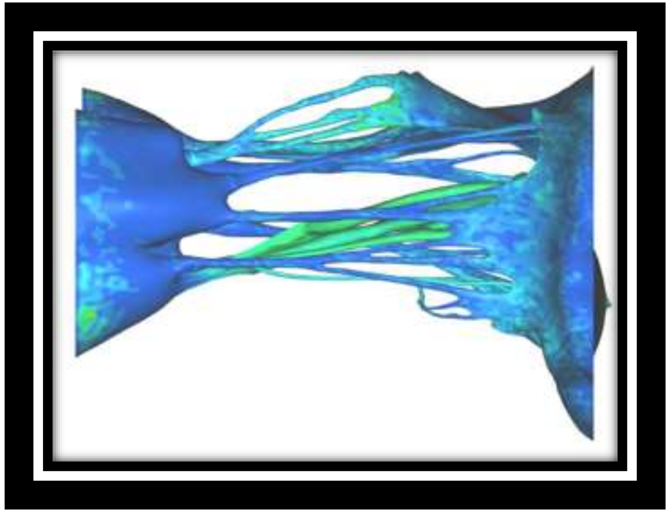
Dale Southard, MVAPICH User Group Meeting 2015

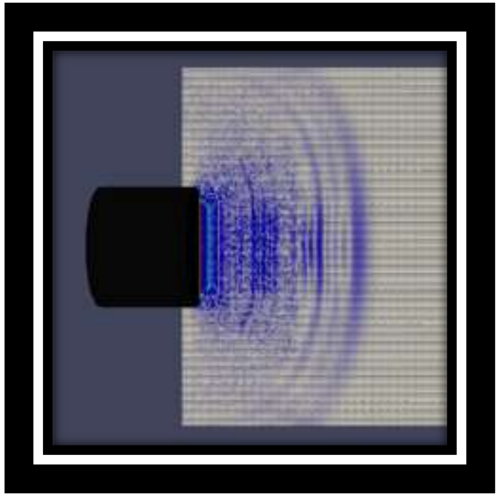
NVIDIA

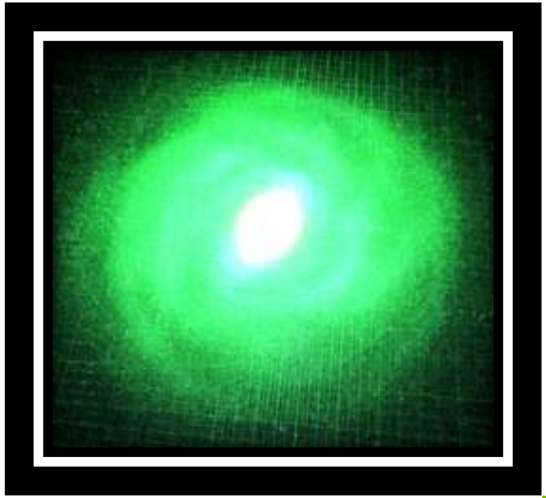
The Future is Big

How Does HPC Touch Your Life?









How Does Your Life Touch HPC?

“<u>Data intensive processing</u>:
High throughput event processing and data capture from sensors, data feeds and instruments”

Pete Ungaro

“<u>Cloud Computing:</u>
App access to converged infrastructure via IP stack.”

Bill Blake

We are the sensors, data feds, and instruments.

The Age of Big Data

2.5 Exabytes of Web Data Created Daily

2.5 Petabytes of Customer Data Hourly

350 Million Images Uploaded a Day

100 Hours Video Uploaded Every Minute

How can we organize, analyze, understand,
and benefit from such a trove of data?

Google “Brain Project”

Building High-level Features Using Large Scale Unsupervised Learning

Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, A. Ng

Stanford / Google

1 billion connections

10 million 200x200 pixel images

1,000 servers(16,000 cores)

3 days to train

Accelerating Machine Learning

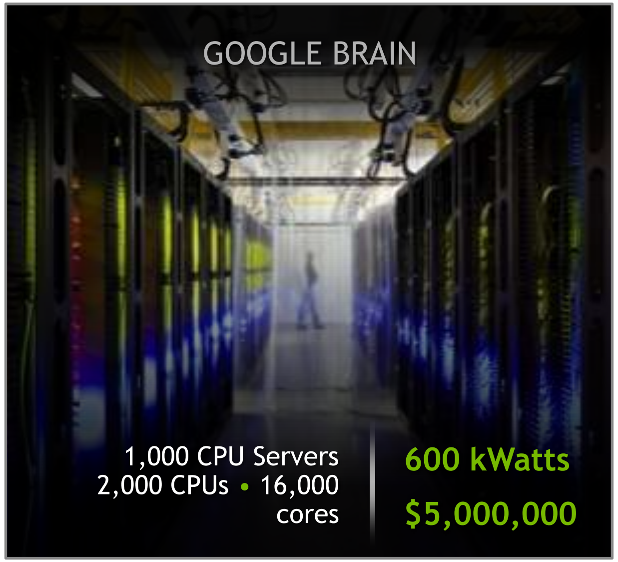
Deep Learning with COTS HPC Systems

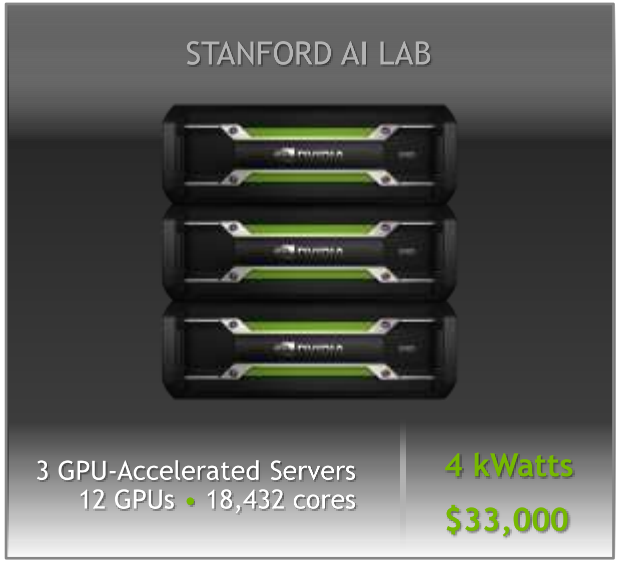
A. Coates, B. Huval, T. Wang, D. Wu,
A. Ng, B. Catanzaro

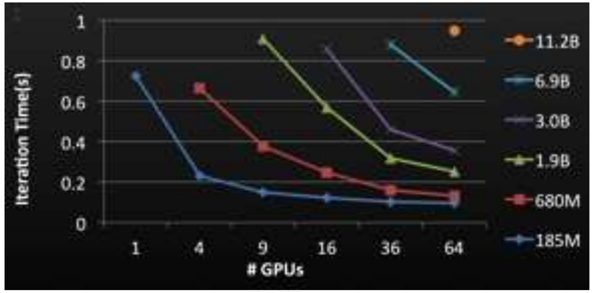
Stanford / NVIDIA • ICML 2013

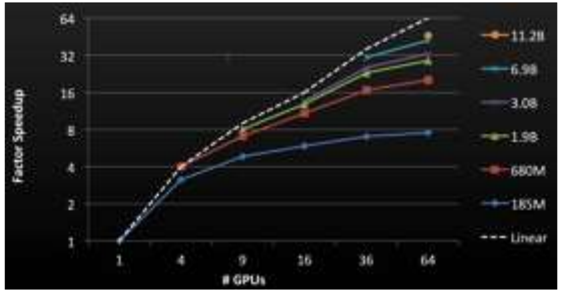
"Now You Can Build Google's $1M Artificial Brain on the Cheap

-Wired









"10 Billion Parameter Neural Networks In Your Basement", Adam Coates
http://on-demand.gputechconf.com/gtc/2014/video/S4694-10-billion-parameter-neural-networks.mp4

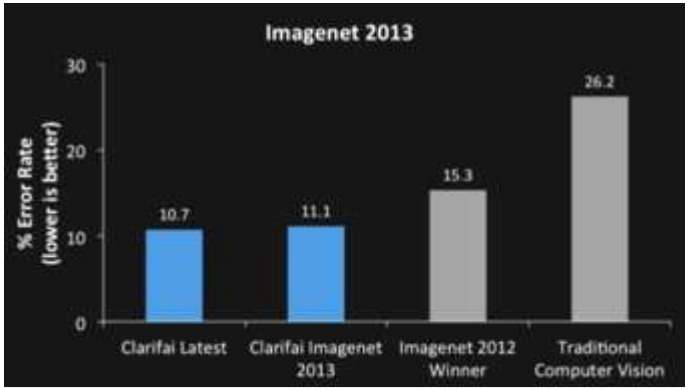
Accelerating Machine Learning

Image Recognition CHALLENGE

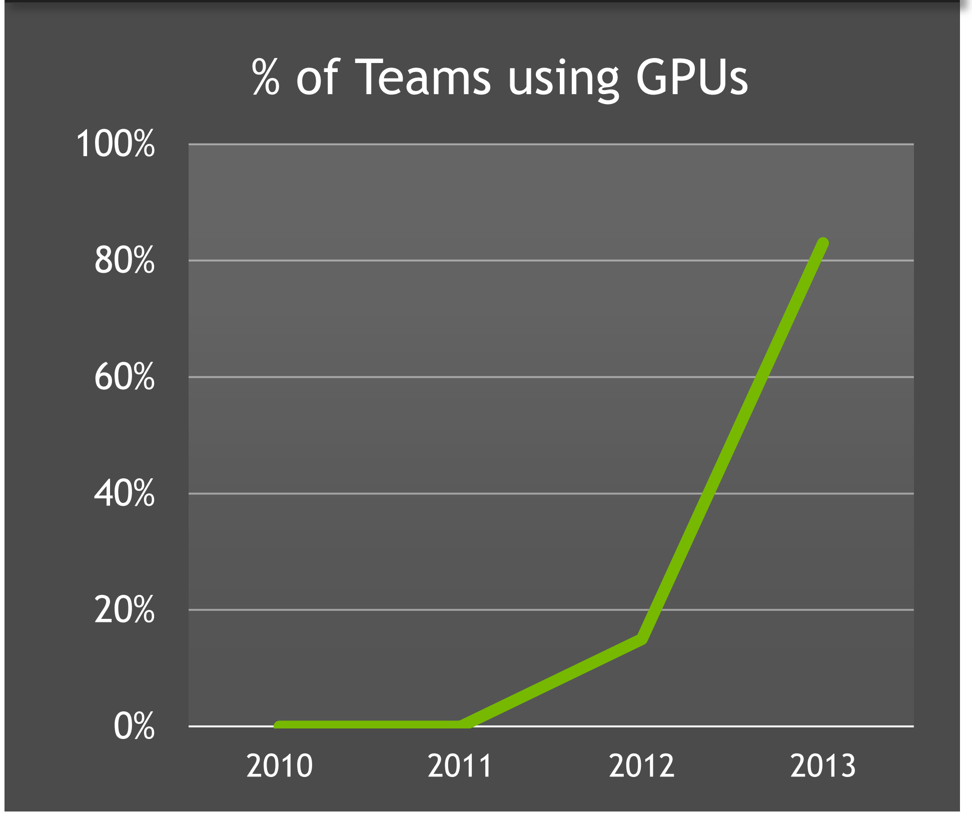
1.2M training images • 1000 object categories

Hosted by

IMAGENET



GPU teams sweep all 3 categories at ILSVRC2013



Clarifai: Enabling Next Generation Intelligent Applications – http://on-demand-gtc.gputechconf.com/gtcnew/on-demand-gtc.php#2957

Machine Learning Comes of Age

Image Detection

Face Recognition

Gesture Recognition

Video Search & Analytics

Speech Recognition & Translation

Recommendation Engines

Indexing & Search

Talks at GTC

facebook

NYU

STANFORD UNIVERSITY

DARPA

DENSO

Carnegie Mellon University

MIT Massachusetts Institute of Technology

Berkeley UNIVERSITY OF CALIFORNIA

Web & Enterprise Companies use GPUs to Accelerate Machine Learning, Data Analytics, And Technical COMPUTING

Adobe

Auto Tagging in Creative Cloud

Baidu

Speech/Image Recognition

flickr

Image Auto Tagging

IBM

Hadoop-based Clustering

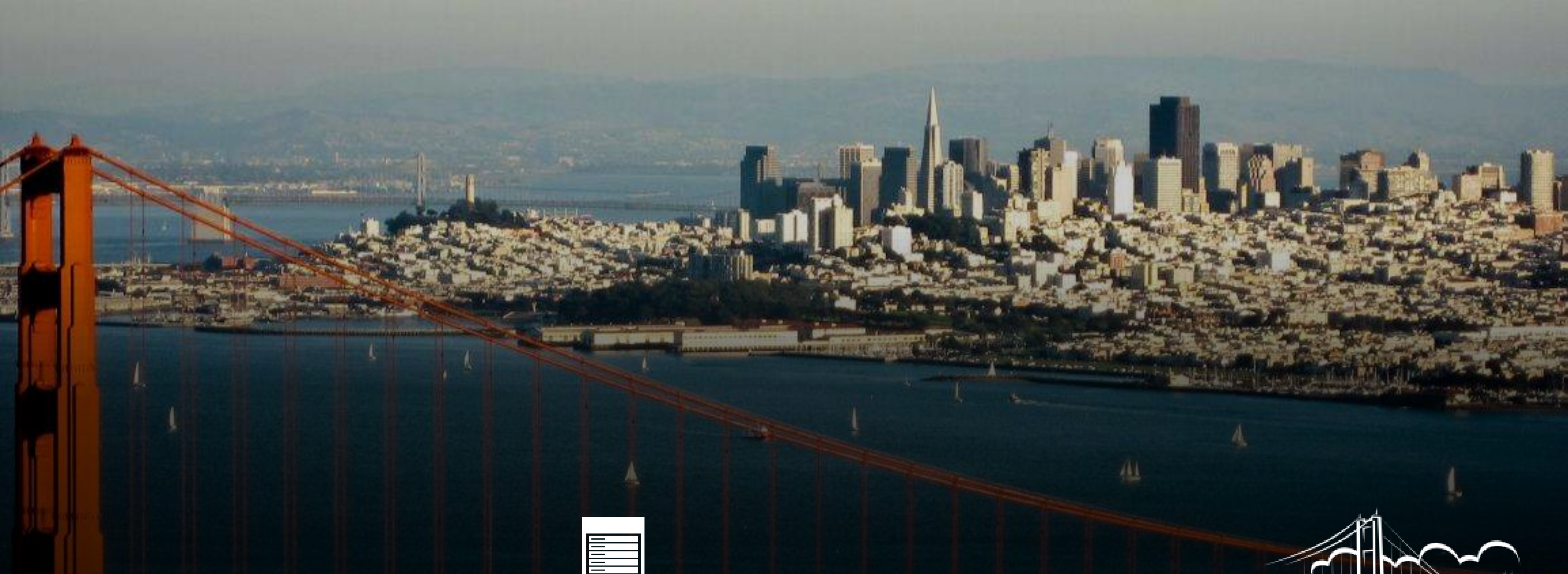
NETFLIX

Recommendation Engine

orange™

Database Queries

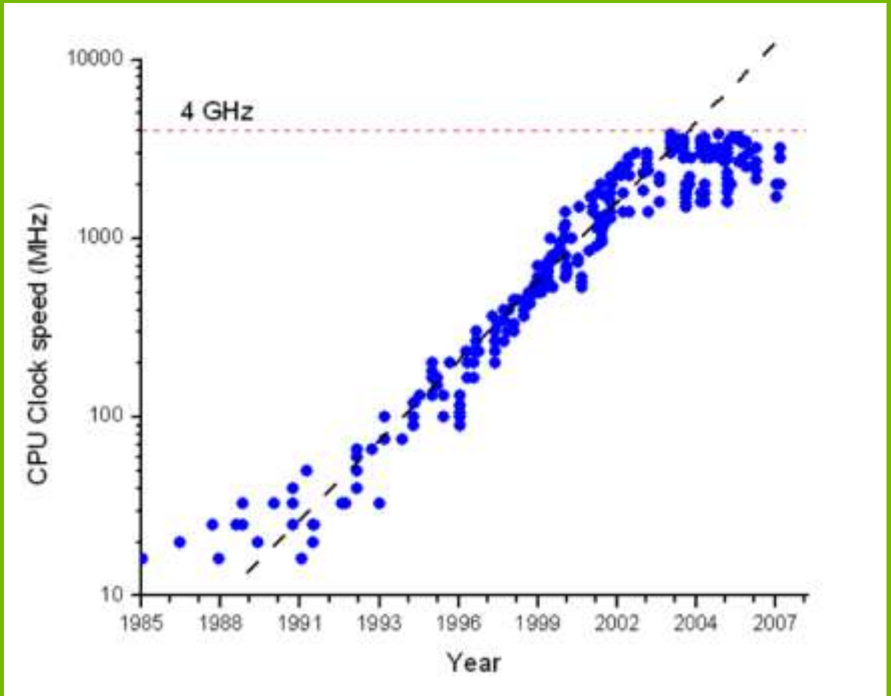
How Do We Meet This Demand?



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

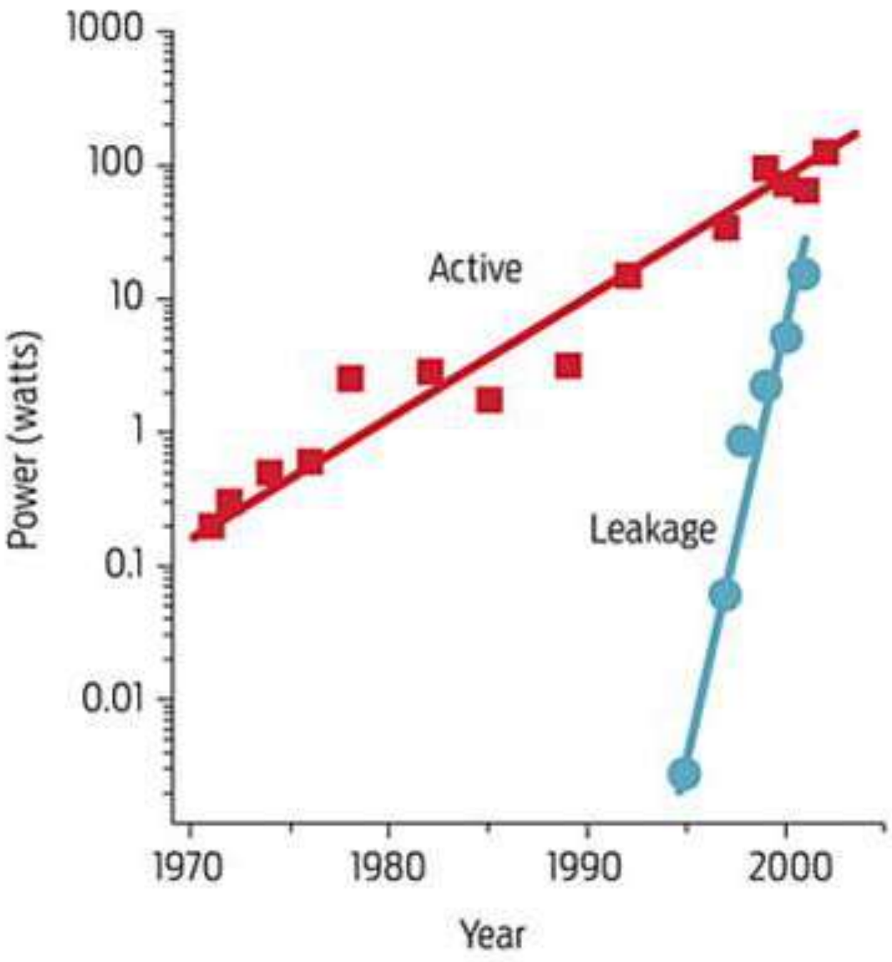
The New Reality

Leakage has limited threshold voltage, largely ending voltage scaling

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

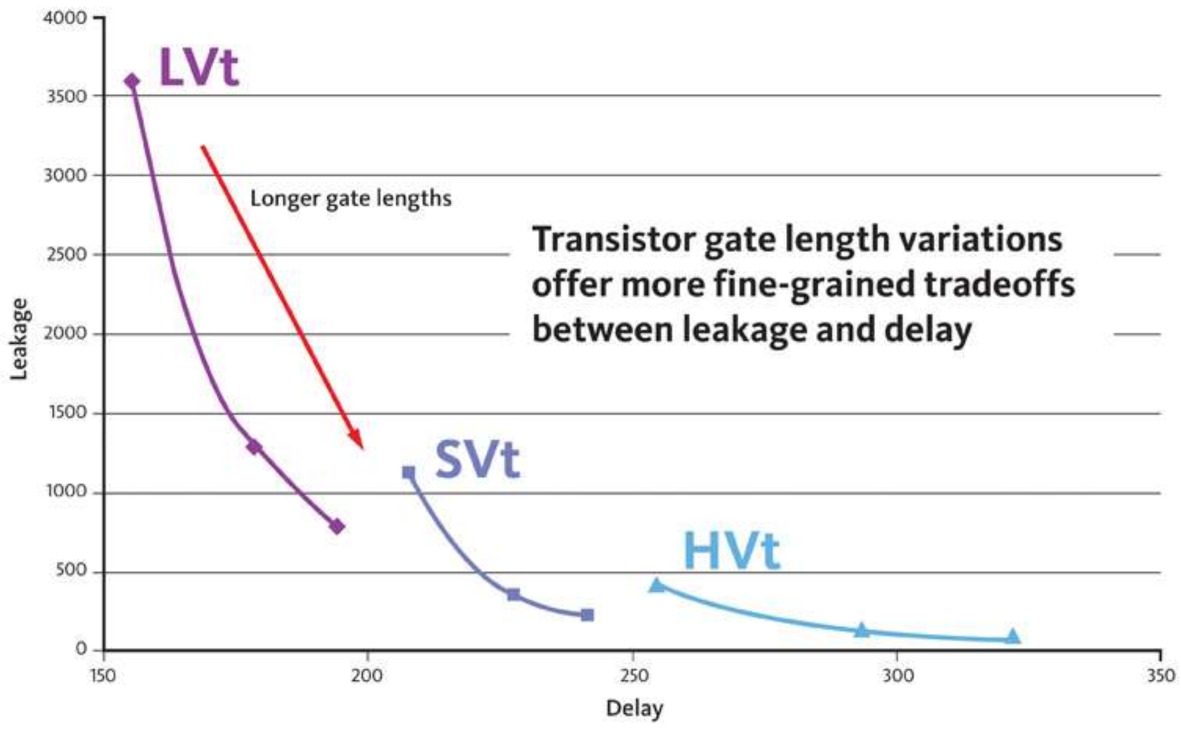
Halve L and get 4x the transistors and 8x the capability for 4x the power,
or 2x the capability for the same power in ¼ the area.

The Rise of Leakage



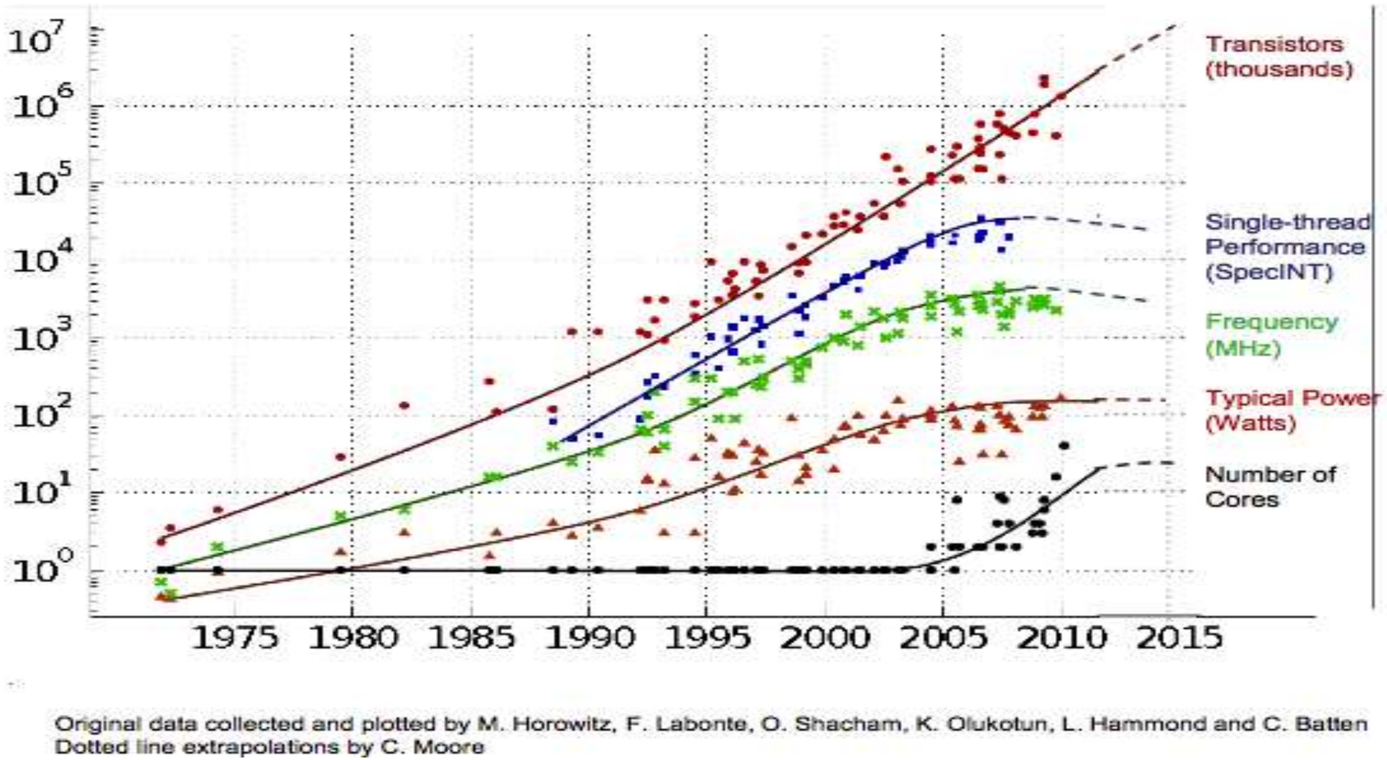
Source: Gordon Moore, Intel; IEEE

Frequency vs. Leakage



Source: Gordon Moore, Intel; IEEE

Parallelism to the Rescue



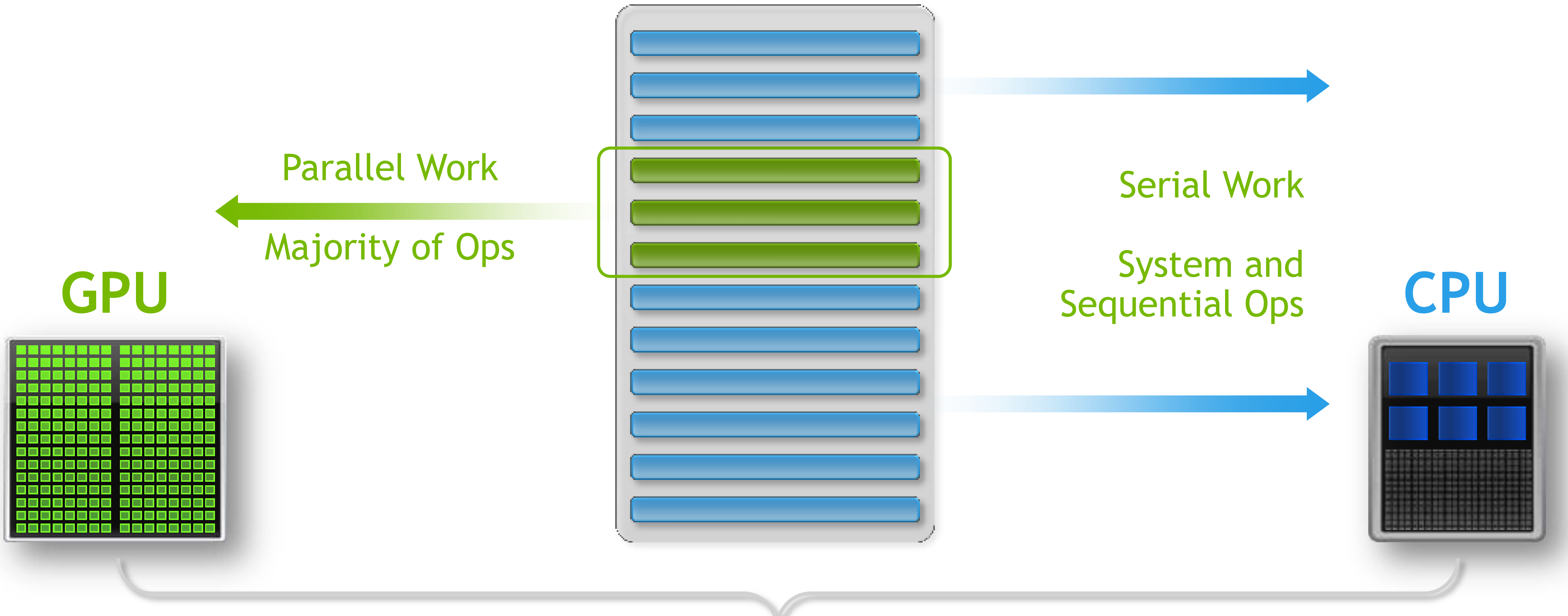
C Moore, *Data Processing in ExaScale-ClassComputer Systems,* Salishan, April 2011

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

- Seymour Cray, 1989

Hint: We want <u>both</u>.

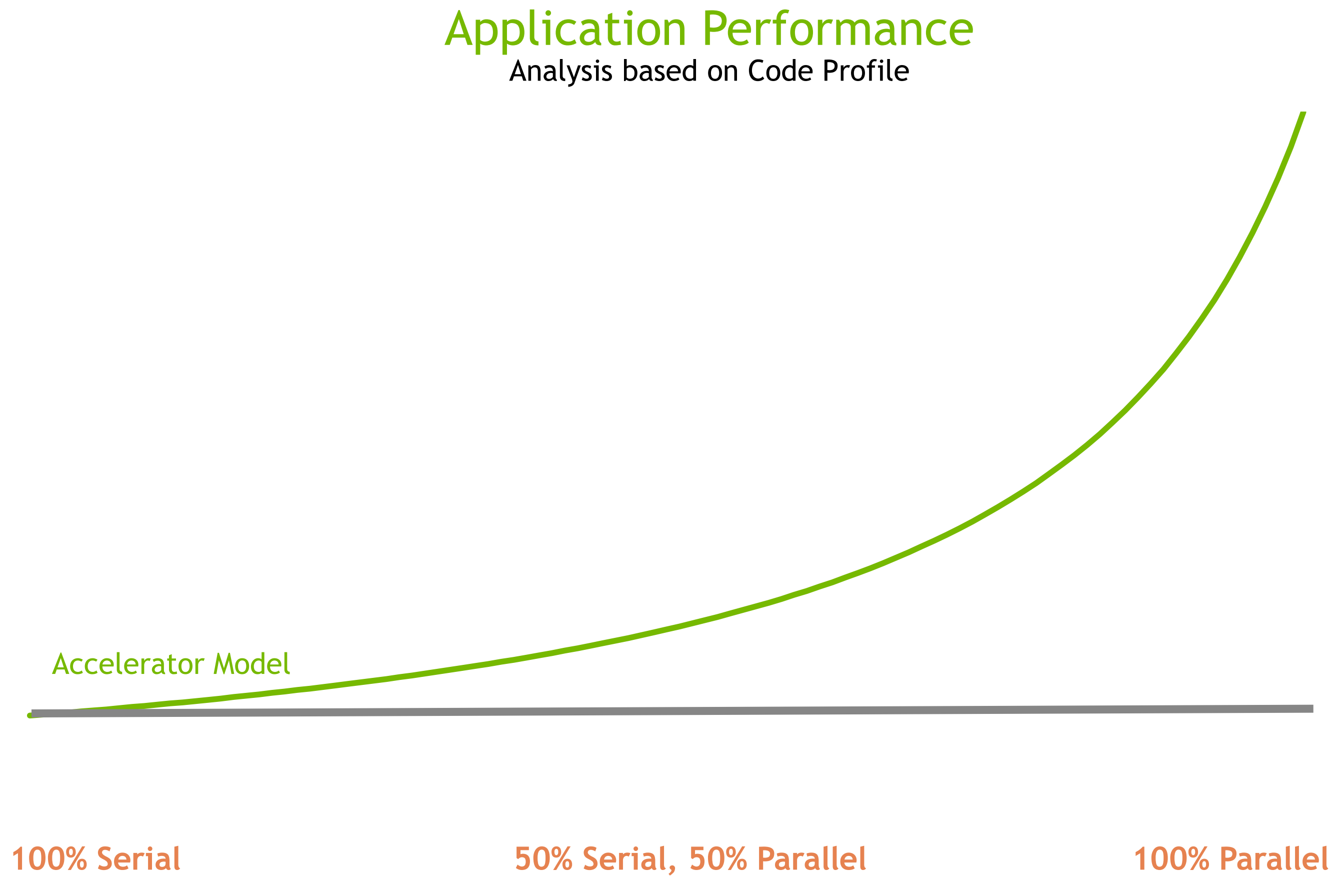
Optimizing Serial/Parallel Execution



Tightly-Coupled Heterogeneous Architecture

Accelerator Model And Amdahl’s Law

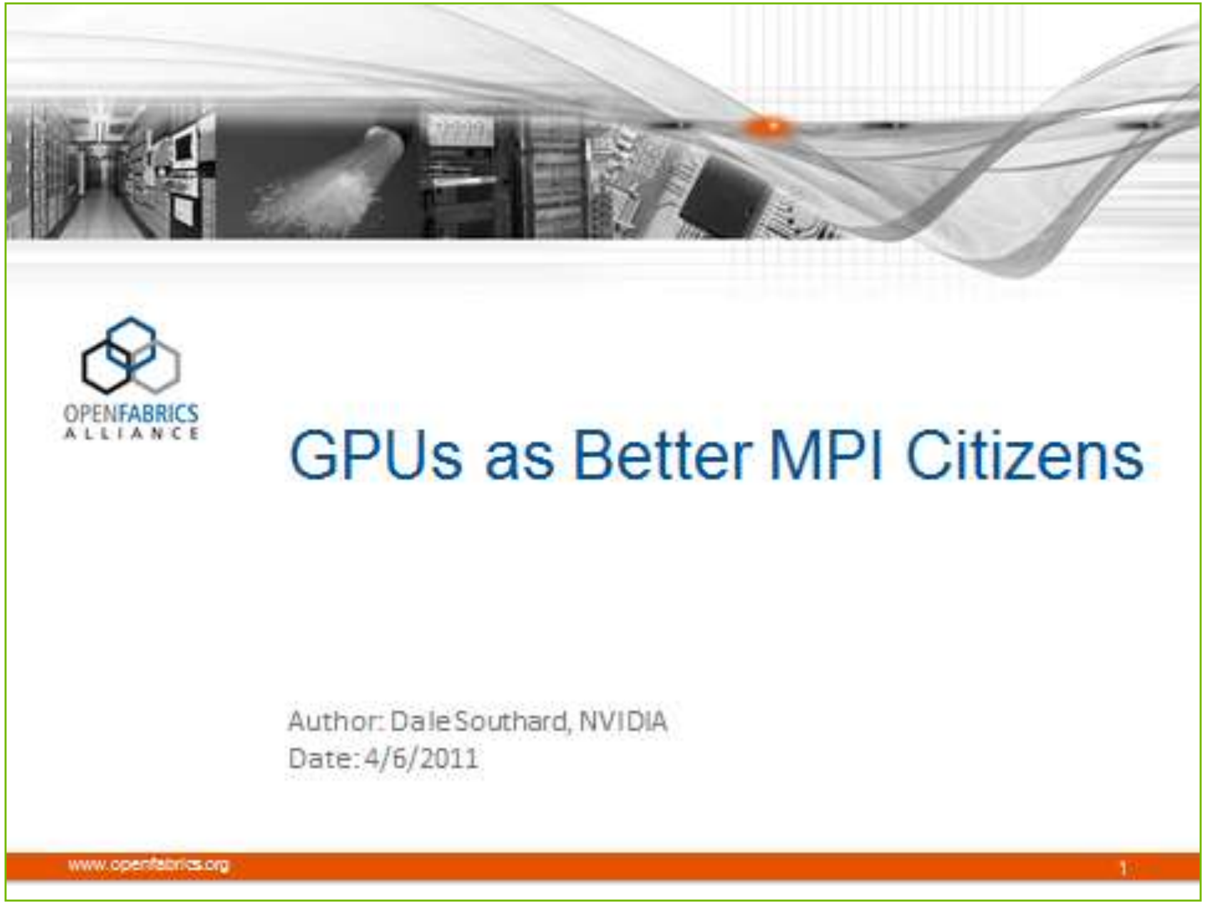
Code Runs on Optimal Processor



- Always faster than CPU
- Serial code runs on CPU
- Parallel code offloads to optimized accelerator

Nice, but What About Messaging….

MPI Interop is Not a New Conversation



The DMA/RDMA Problem Circa 2009

Two Processors, Two Memory Spaces

- CUDA driver allocates its own pinned memory region for DMA transfers to/from GPU
- IB driver allocates its own pinned memory region for RDMA transfers to/from IB card
- GPU can only access system memory
- IB can only access system memory
- MPI stack has no knowledge of GPU

Enter GPUDirect

GPUDirect is a long-term effort by NVIDiA

- The goal is tighter integration of GPU and Network Fabrics
- GPUDirect is not tied to specific network hardware
- GPUDirect is not tied to specific network middleware or programming APIs

GPUDirect “version 1”

- Intial release of GPUDirect focused on memory pinning
  - Allowed OFED stack and GPU driver to agree on pinned regions
  - Maximized RDMA throughput for both GPU and IB
- Implemented as userspace calls to register/deregister memory
- Drove further improvements in driver and CUDA
  - Unified Virtual Addressing (single VA space spanning GPU and CPU memories)
  - Unified Memory (more later)

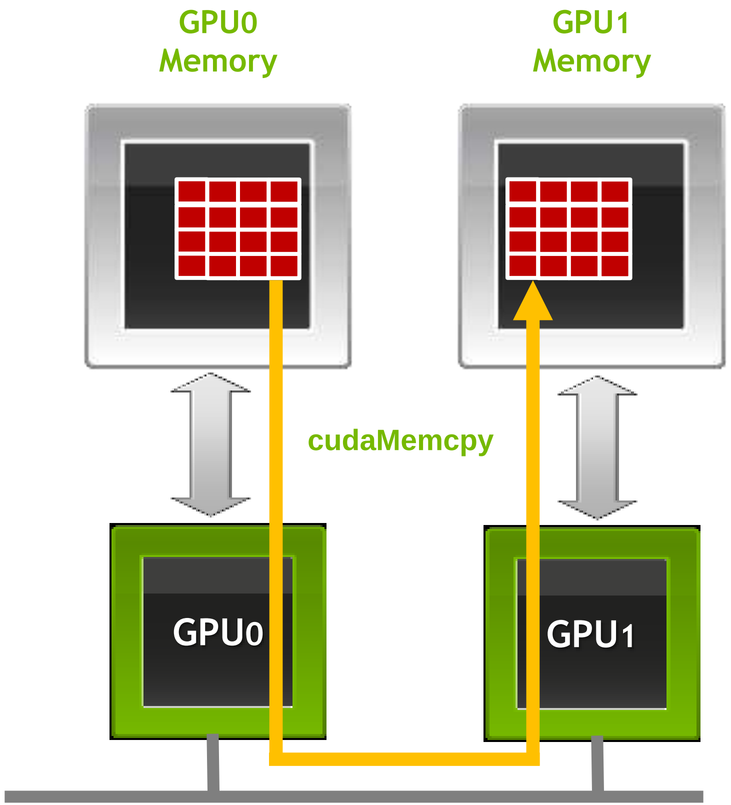
GPUDirect v2.0: Peer-to-Peer

Starting to look like Messaging

GPUDirect v2 introduced P2P features

- Direct Access (dereferencing pointers to remote GPU memory over PCIe)
- Direct Transfer (memCopy between GPUs on the Same PCIe complex)

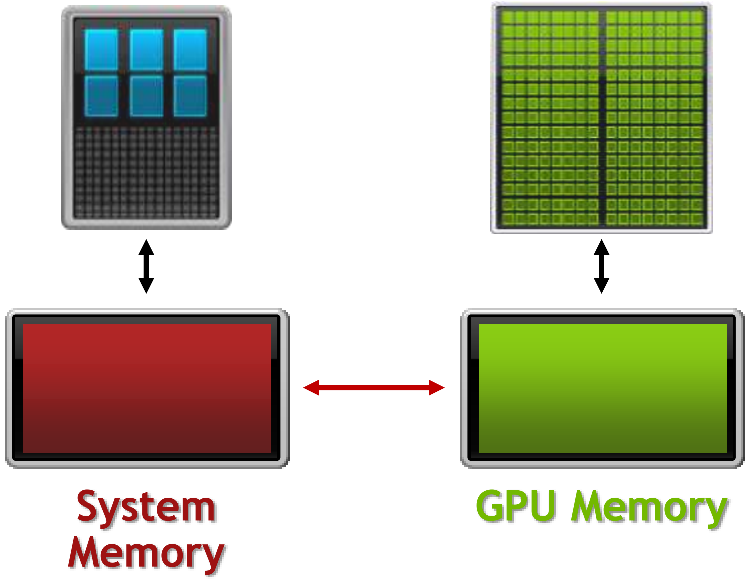
MPI implementations leveraged the latter as a message delivery path (eliminating need to bounce-buffer in host)



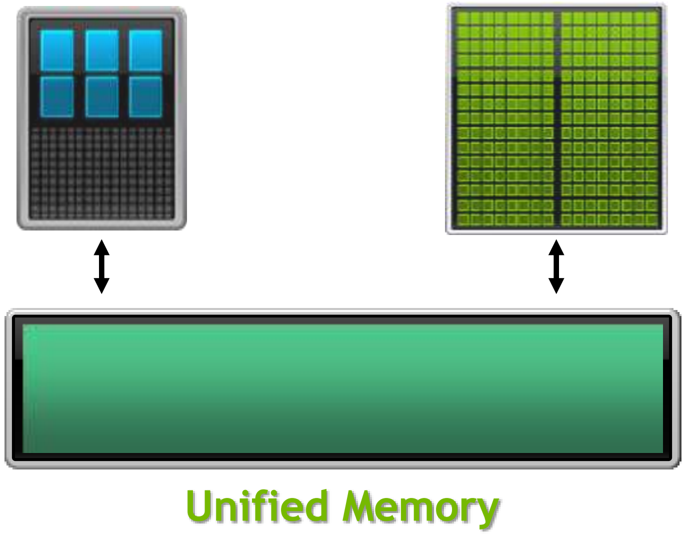
P2P Direct Transfers

Unified Memory

Dramatically lower developer effort



Developer View before UM



Developer View With Unified Memory

The Future is Parallel

To Exascale and Beyond

With CEO Math

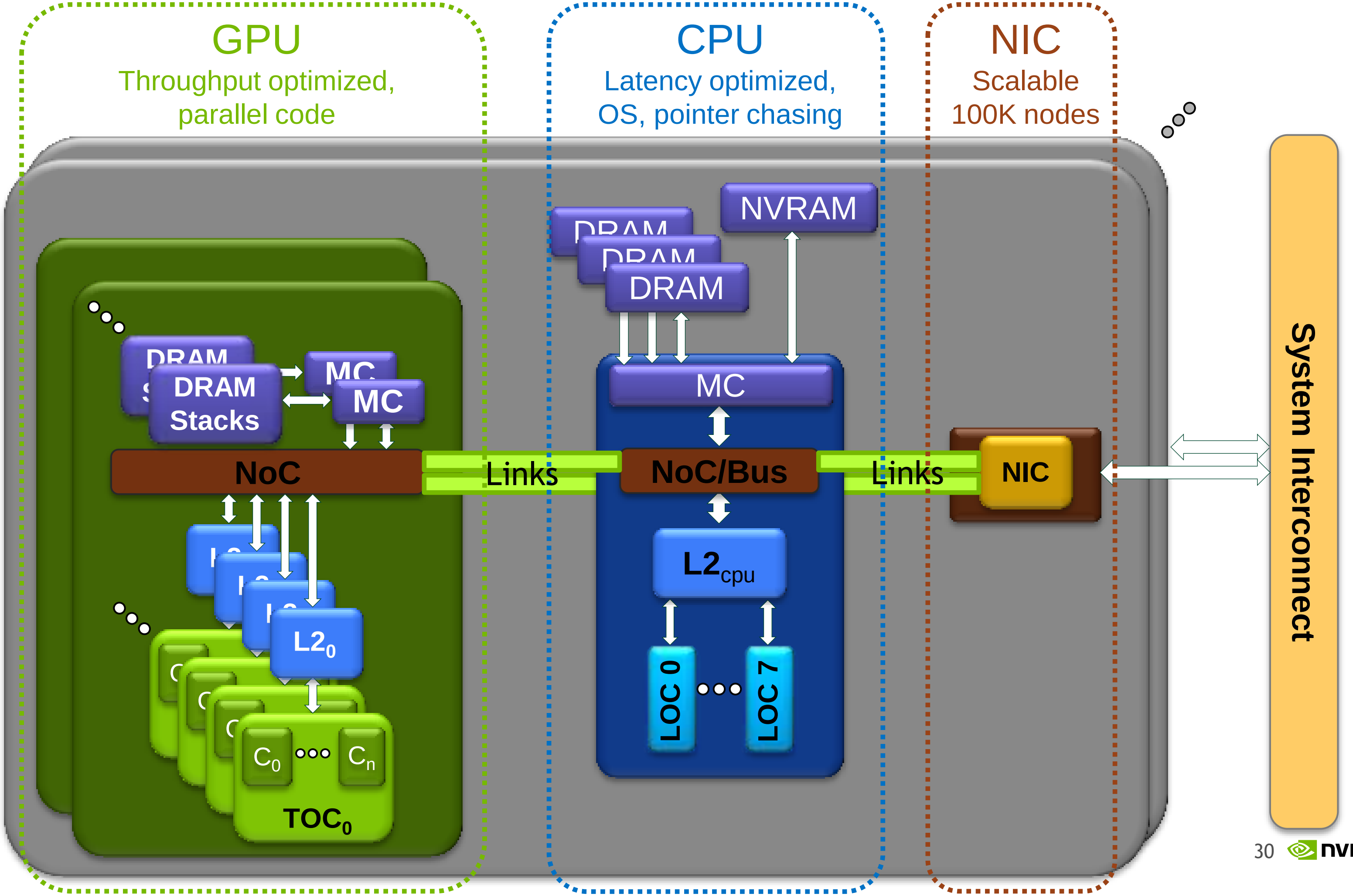
- Achieving exascale performance will require immense parallism
  - Combined thread-level and node-level ~10^10 way parallism
  - Multiple Approaches (MPI+X, PGAS)
- This is not (just) a HW problem
- This is not (just) a SW problem
- Other Technical Computing areas looking for different design points

Generic Future Node Model

Three Building Blocks (GPU, CPU, Network)

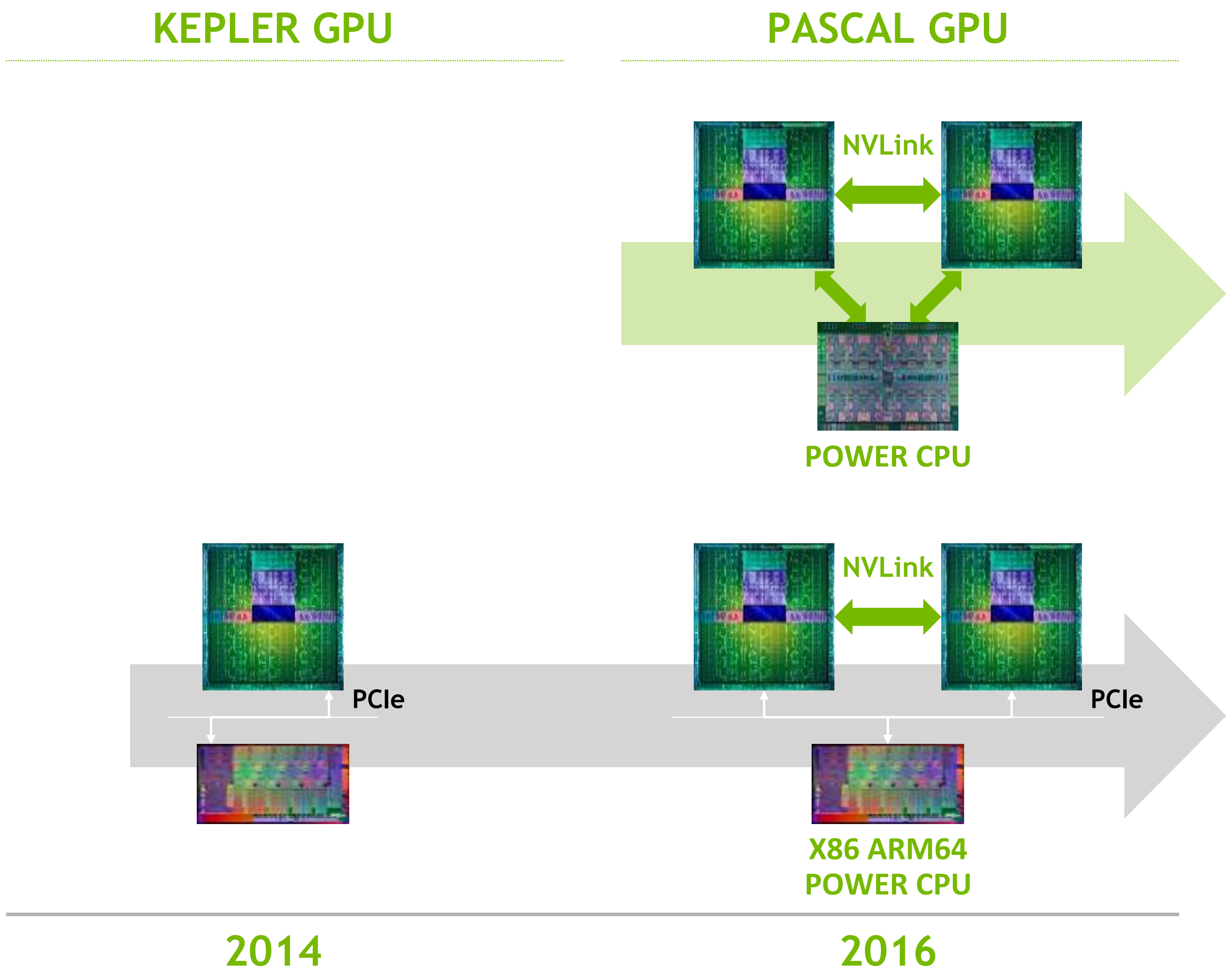
Direct Evolution

- 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*

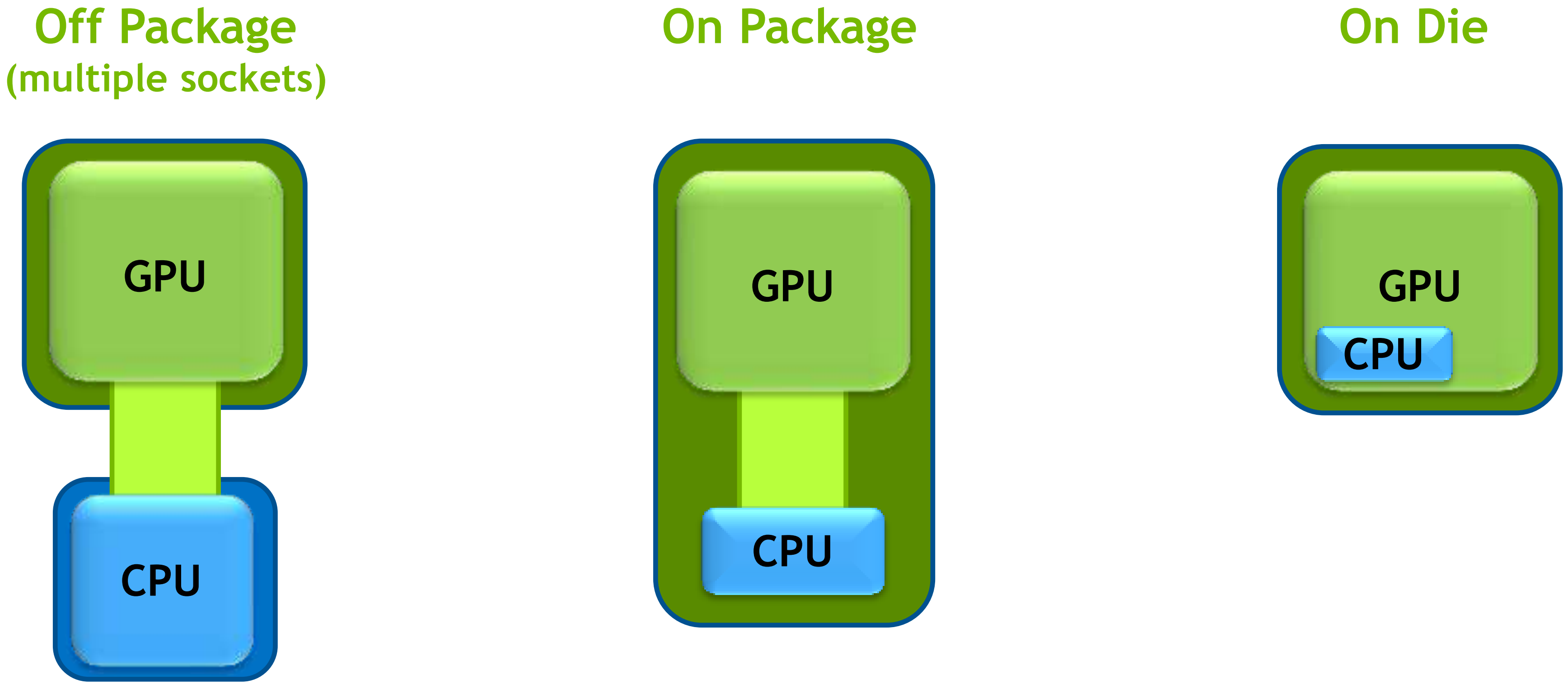


NVLink

High-Speed GPU Interconnect



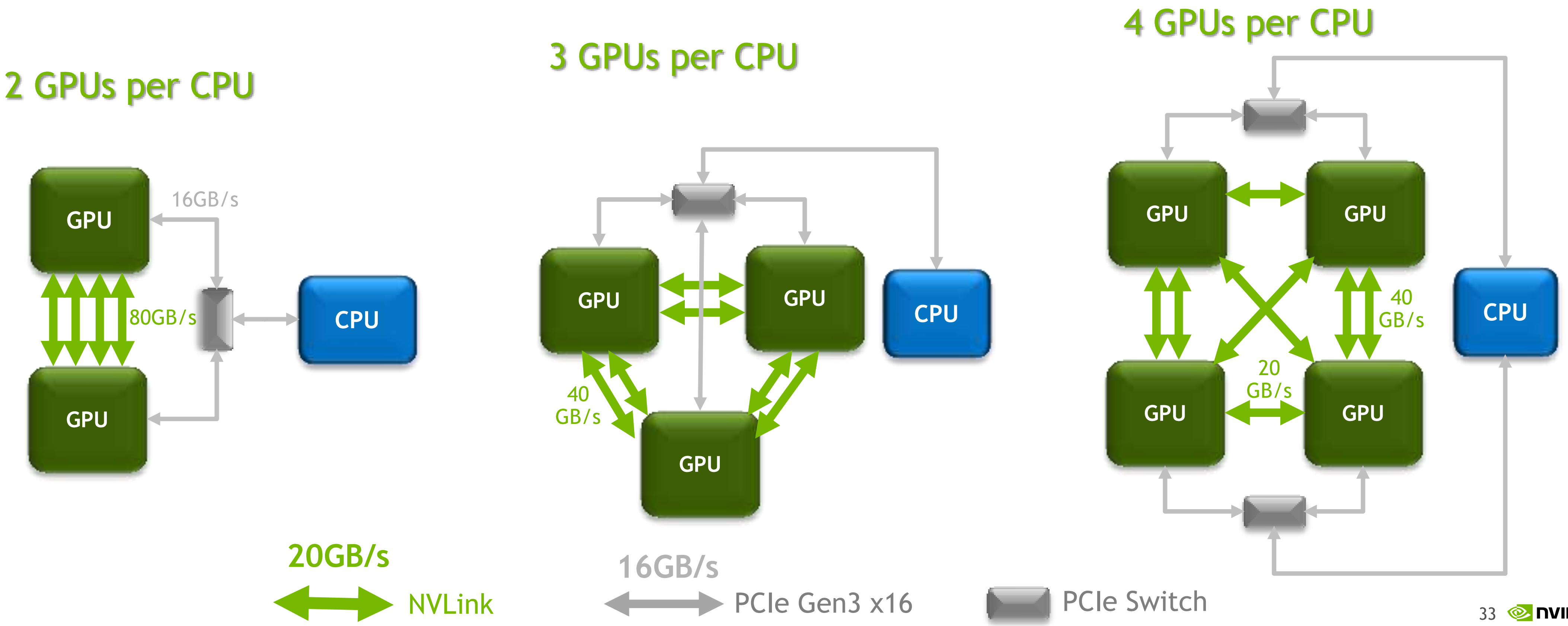
Integration



Packaging does not change programming model

Scaling Node for Workload

Multiple Ways to Balance Parallel and Serial Performance



Scaling the Network

- Future Networks need to scale to over 100K nodes
- Multiple vendors (but some common HW and middleware)
- Evolving towards put/get (or direct load/store)
- Evolving away from “flat” topologies (towards dragonfly)

End result is systems that leverage data locality at the expense of uniformity

The Role of MPI

…in the age of NVLink and PGAS capable networks

MPI (and MVAPICH) will continue to play an important role

- Provide a single messaging API that is fabric agnostic
- Decouple messaging granularity from node design
- Utilize heuristics to select the "best" delivery path

MVAPICH does all of these things today.

We are counting on them for exascale.