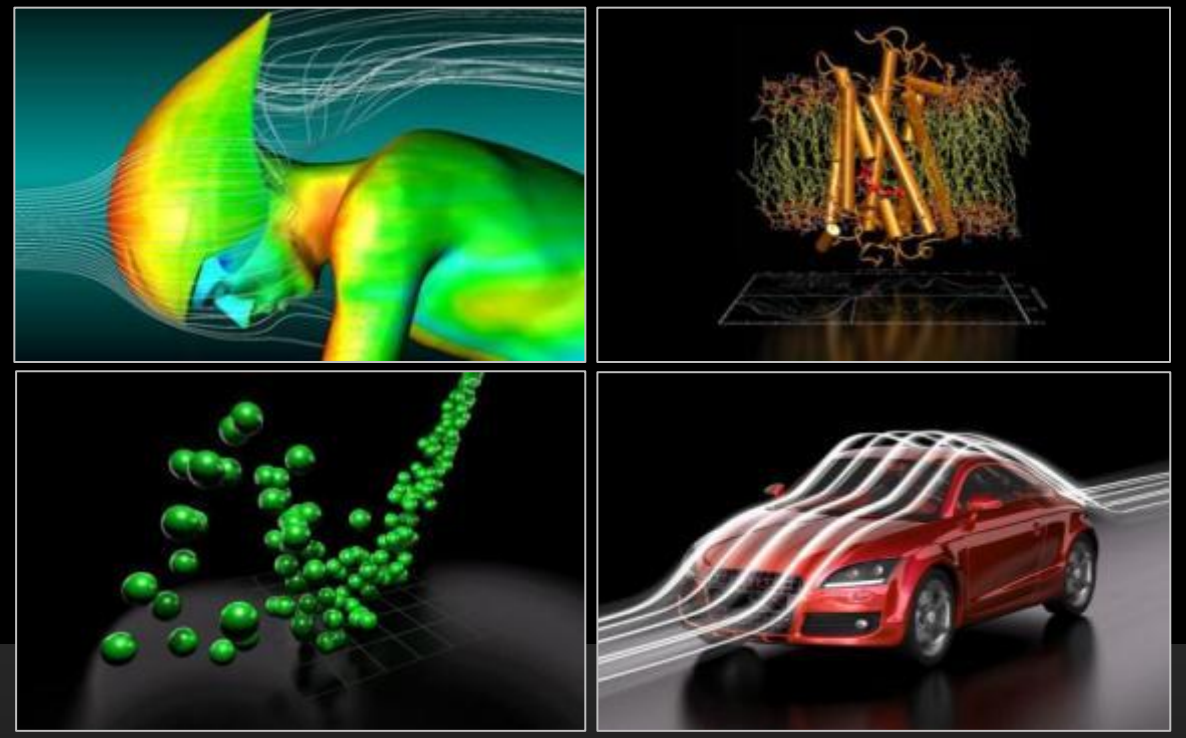


NVIDIA

Atelier Teratec 2011
‘Etat de l’art GPU Computing’


1995
5,000 triangles/second
800,000 transistors GPU

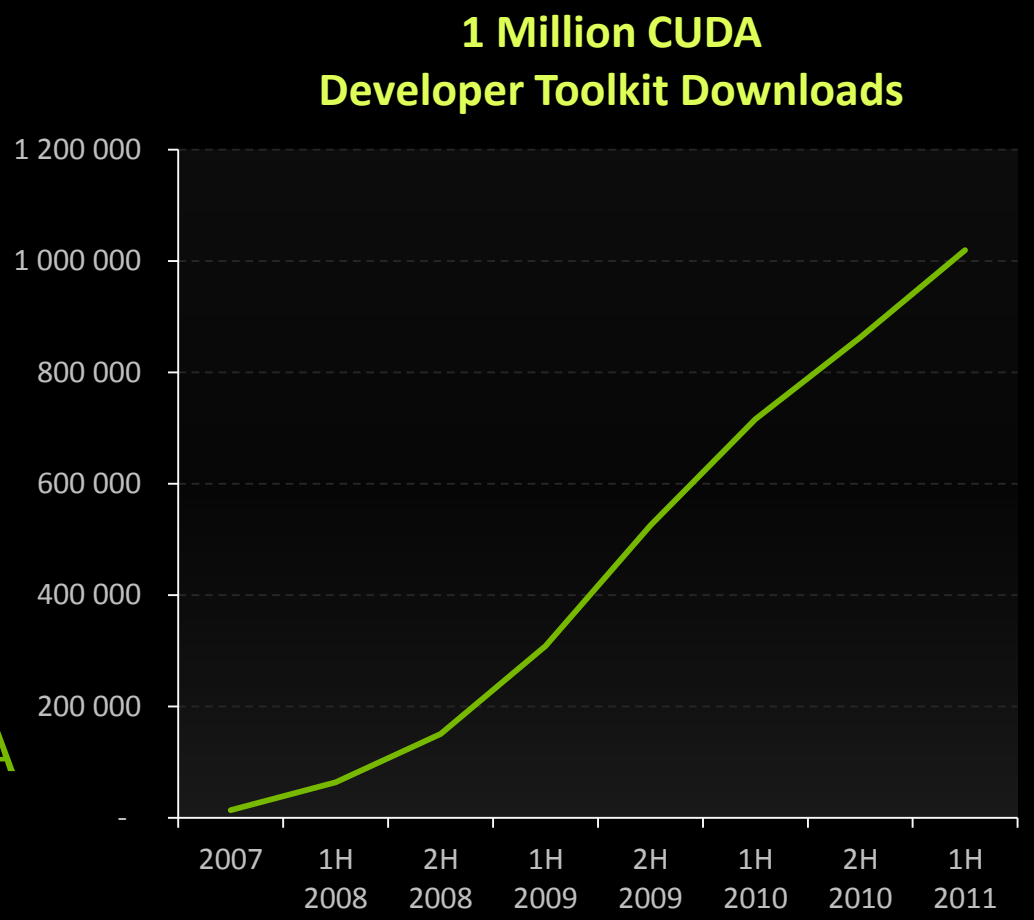
2011

350 Million triangles/second
3 Billion transistors GPU


NVIDIA
NVIDIA
SuperPhones to SuperComputers

GPU Computing Today

300 Million	CUDA Capable GPUs
1 Million	CUDA Toolkit Downloads
150 Thousand	Active CUDA Developers
450	Universities Teaching CUDA



Getting Started with GPUs

TRY

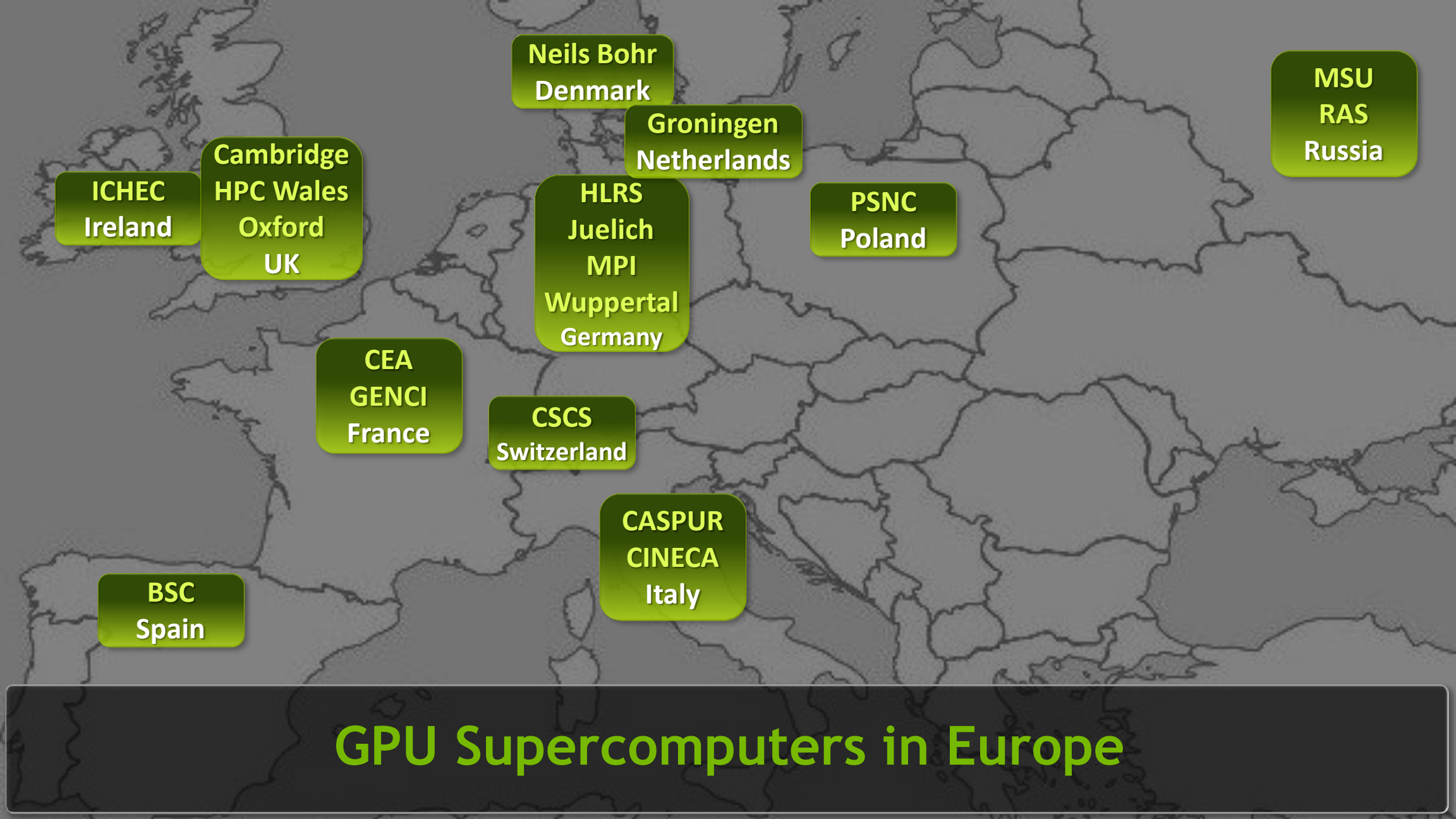
Try CUDA Toolkit on your Notebook or Desktop with CUDA Enabled GPU

DEVELOP

Optimize HPC Apps on Compute Workstation with Tesla GPUs

DEPLOY

Run apps in production with GPU Compute Cluster


Neils Bohr
Denmark
Groningen
Netherlands
MSU
RAS
Russia
Cambridge
HPC Wales
Oxford
UK
ICHEC
Ireland
HLRS
Juelich
MPI
Wuppertal
Germany
PSNC
Poland
CEA
GENCI
France
CSCS
Switzerland
CASPUR
CINECA
Italy
BSC
Spain
GPU Supercomputers in Europe

GPU Computing Developer Ecosystem

Numerical Packages

MATLAB/Jacket
Mathematica
NI LabView
pyCUDA

Debuggers & Profilers

cuda-gdb
NV Visual Profiler
Nsight for MS VS
Allinea
TotalView

GPU Compilers

C
C++
Fortran
OpenCL
DirectCompute
Java
Python

Parallelizing Compilers

PGI Accelerator
CAPS HMPP
Par4All
mCUDA

Cluster Management

nv-smi
Bright Cluster
Rocks Roll
Scyid
PBS Pro
Cluster Manager

Libraries

BLAS
FFT
LAPACK
NPP
Video
Imaging

Consultants & Training

acceleware, GASS, STONE RIDGE TECHNOLOGY, Elegant Mathematics, Petapath, TECH-X, CAPS, HPC, EM Photonics, SCALABLE GRAPHICS, WIPRO

Solution Providers

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

CUDA GPUs: Support All Languages

CUDA C

CUDA C++

CUDA Fortran

OpenCL

Direct-Compute

C++ AMP

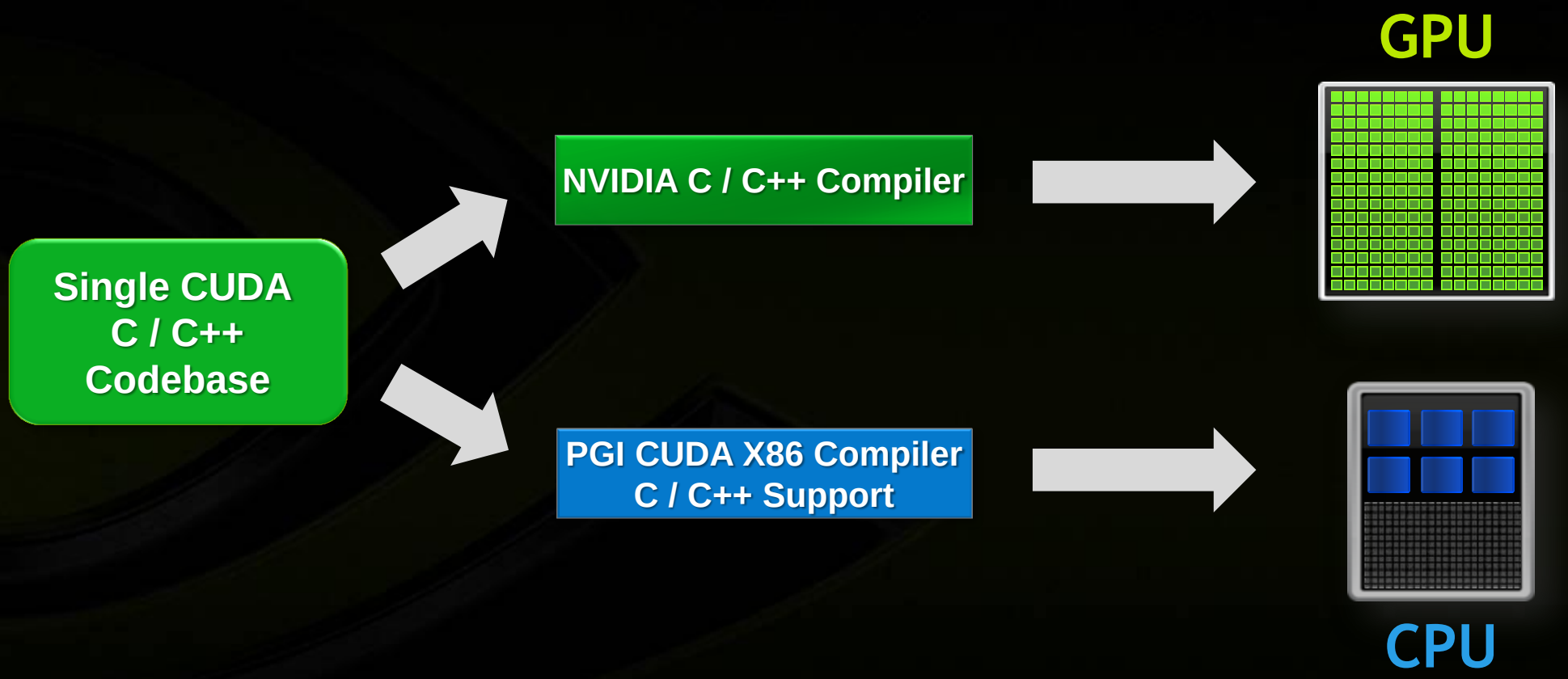
NVIDIA CUDA GPUs

CPU

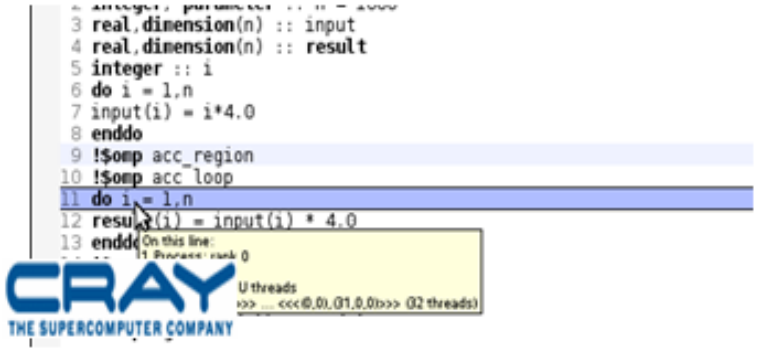
Other Devices

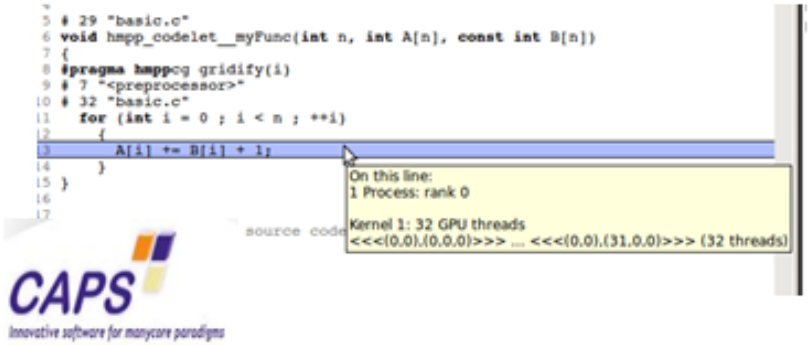
PGI CUDA x86

CUDA Now Available for CPUs and GPUs

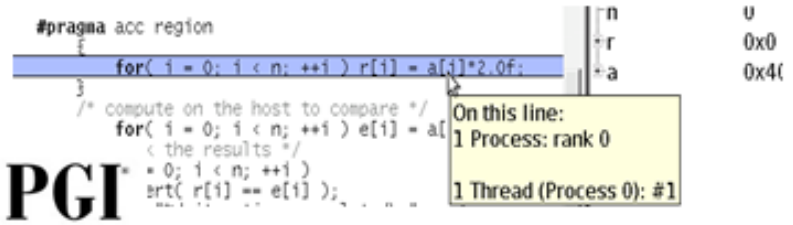


DDT For Directives





- Supporting the environments that you will use for hybrid development



NVIDIA OpenCL Execution

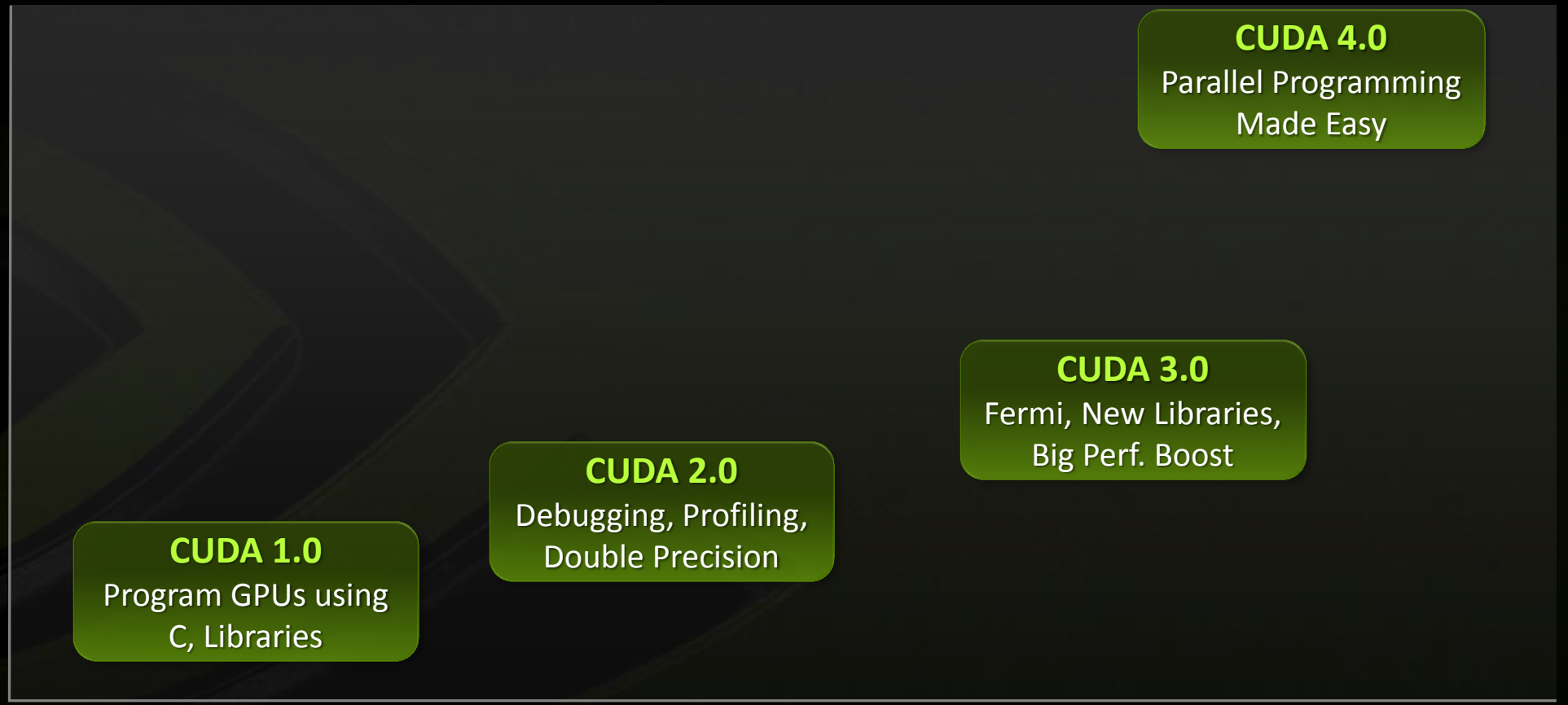
2009

2010

April	June	Aug	Sept	Nov	March
OpenCL Prerelease Driver	OpenCL Conformant Driver	OpenCL Visual Profiler	OpenCL 1.0 R190 UDA	OpenCL 1.0 R195 UDA	OpenCL 1.0 R195 UDA #2
			Conformant release • 2D Imaging • Global atomics • Compiler flags • Compute Query • Byte Addr. Stores	• *Double Precision* • *OpenGL Interop* ++ Performance Enhancements	• ICD • Direct3D9 sharing • Direct3D10 sharing • Direct3D11 sharing • Pragma unroll • Local atomics
OpenCL SDK	OpenCL SDK		OpenCL SDK & CUDA Toolkit 2.3	OpenCL SDK & CUDA Toolkit 2.3	OpenCL SDK & CUDA Toolkit 2.3

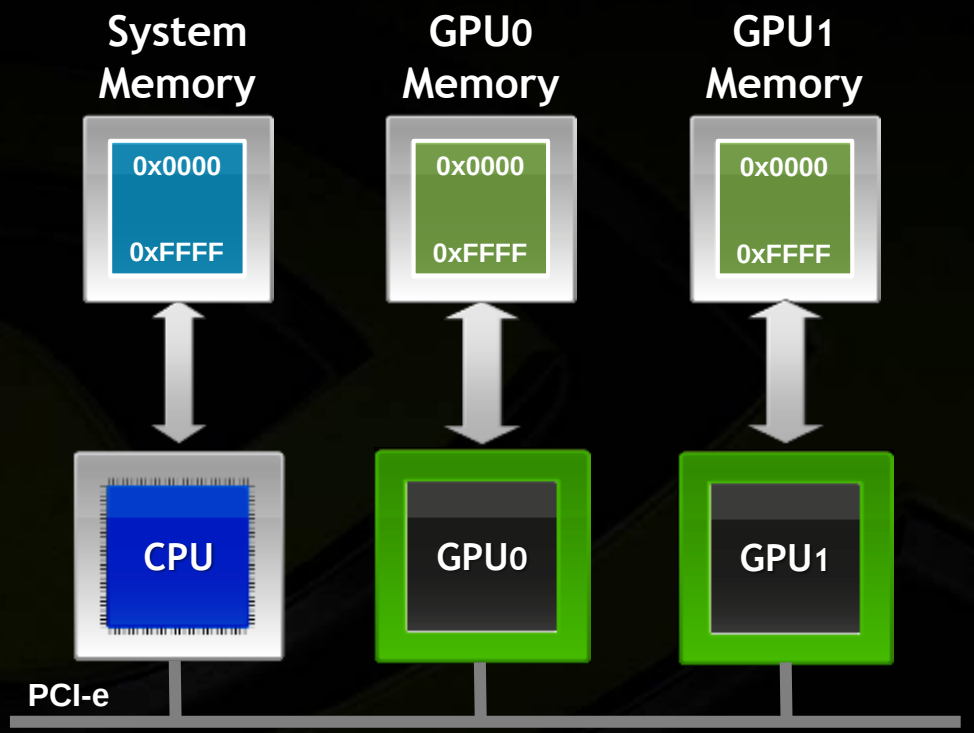
NVIDIA OpenCL Roadmap

CUDA 4.0: Big Leap In Usability

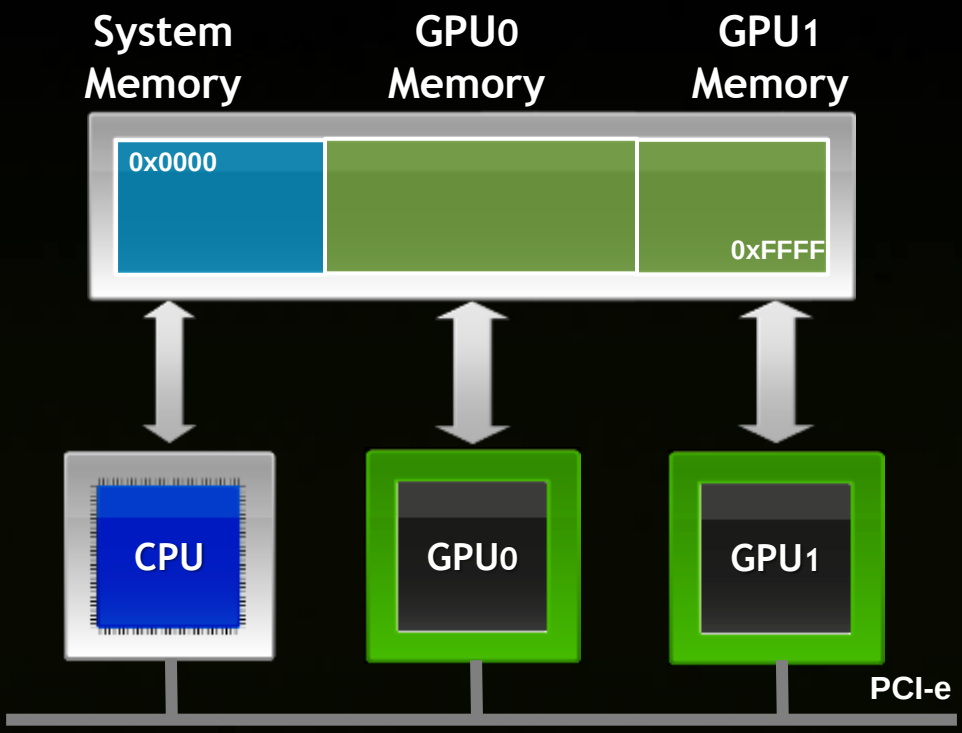


Unified Virtual Addressing

Easier to Program with Single Address Space

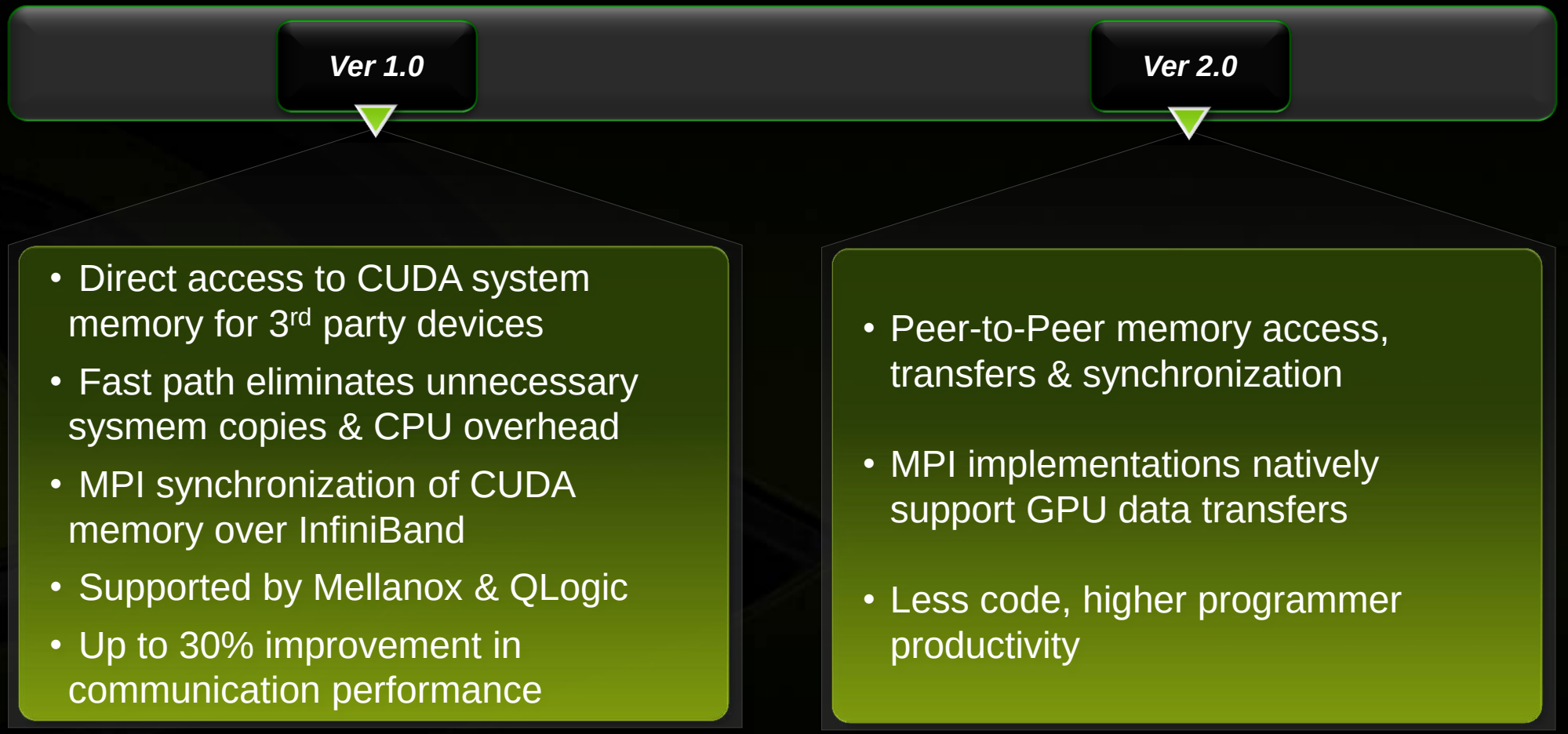


Without UVA



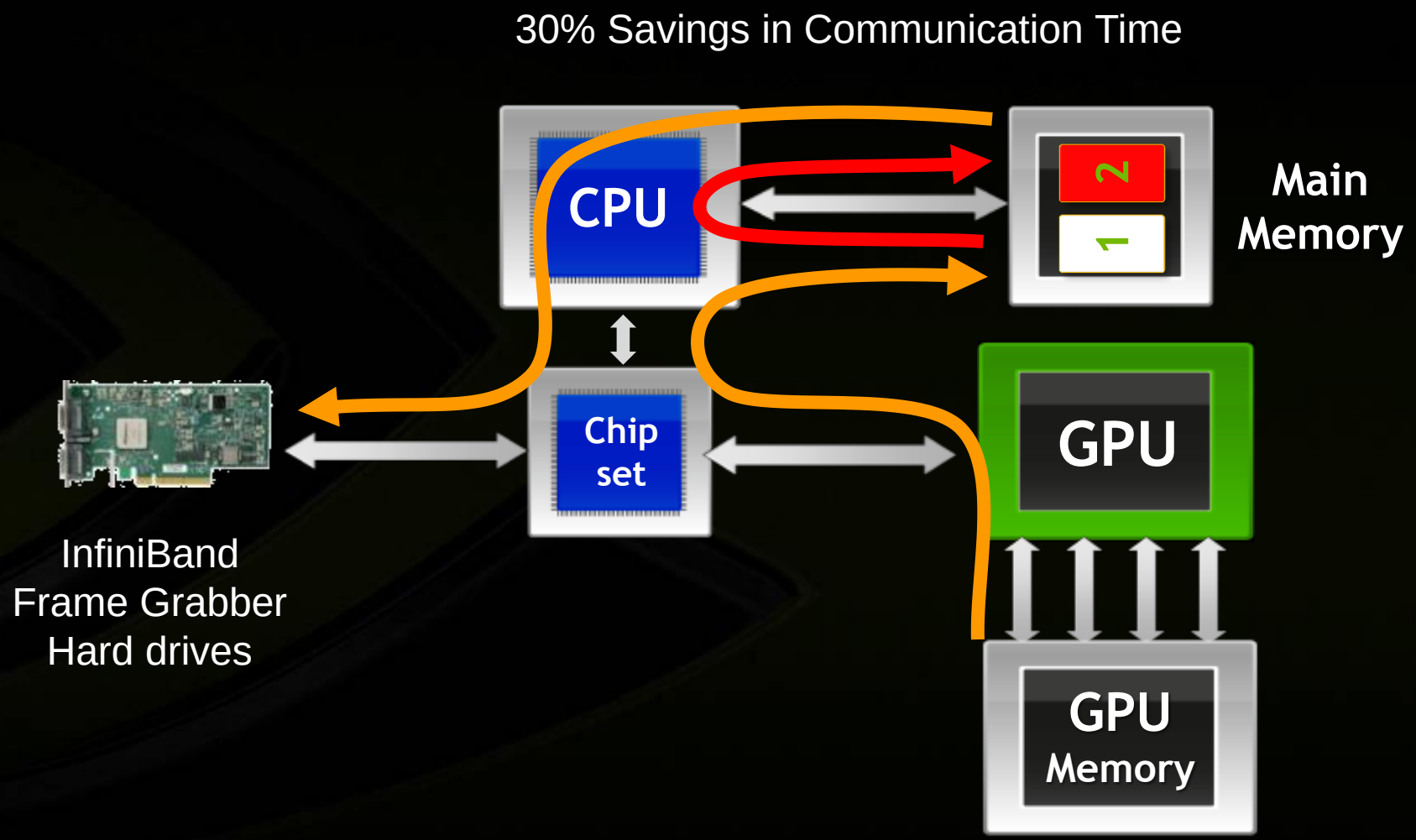
With UVA

NVIDIA GPUDirect™

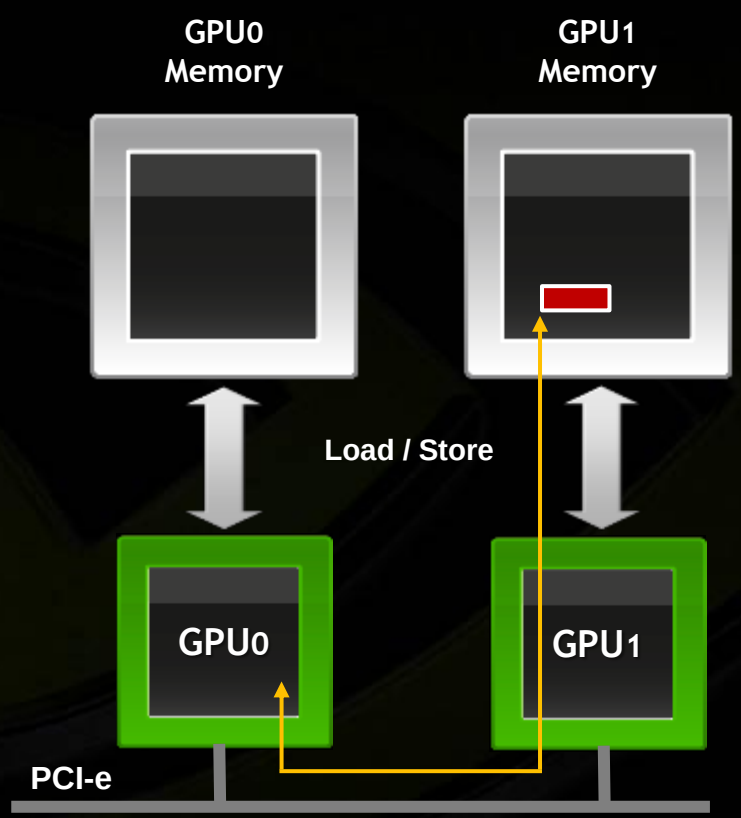


Details @ http://developer.nvidia.com/object/gpudirect.html

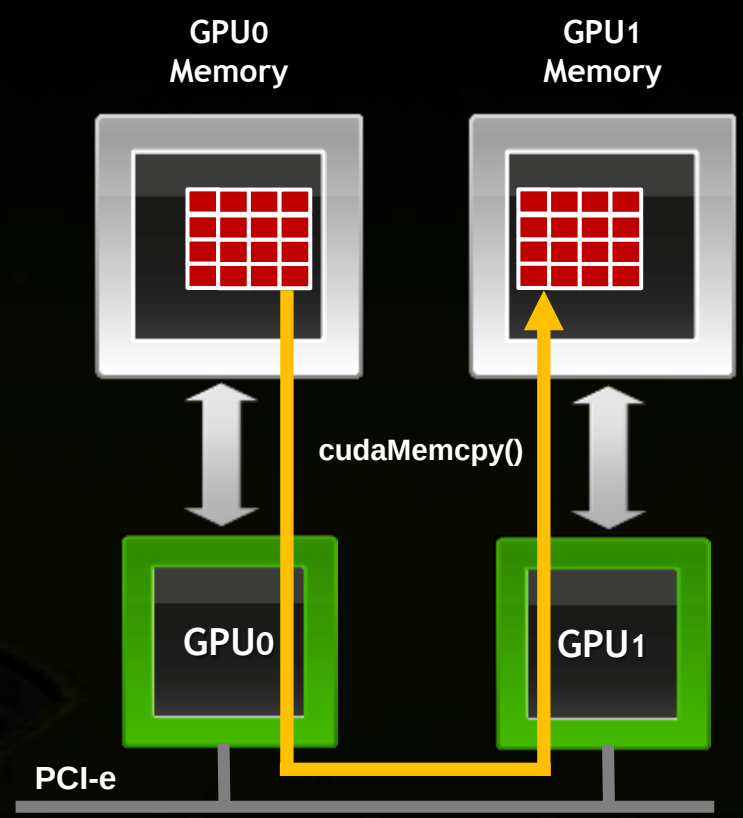
NVIDIA GPUDirect 1.0: Faster Communication with GPUs



GPUDirect v2.0: Peer-to-Peer Communication

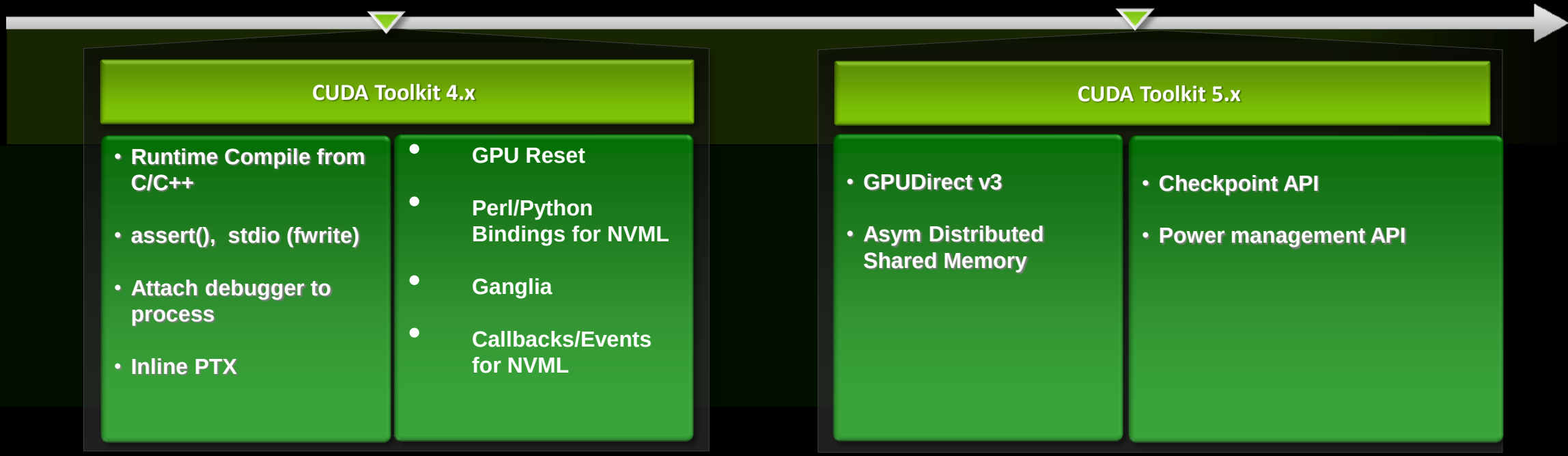


Direct Access



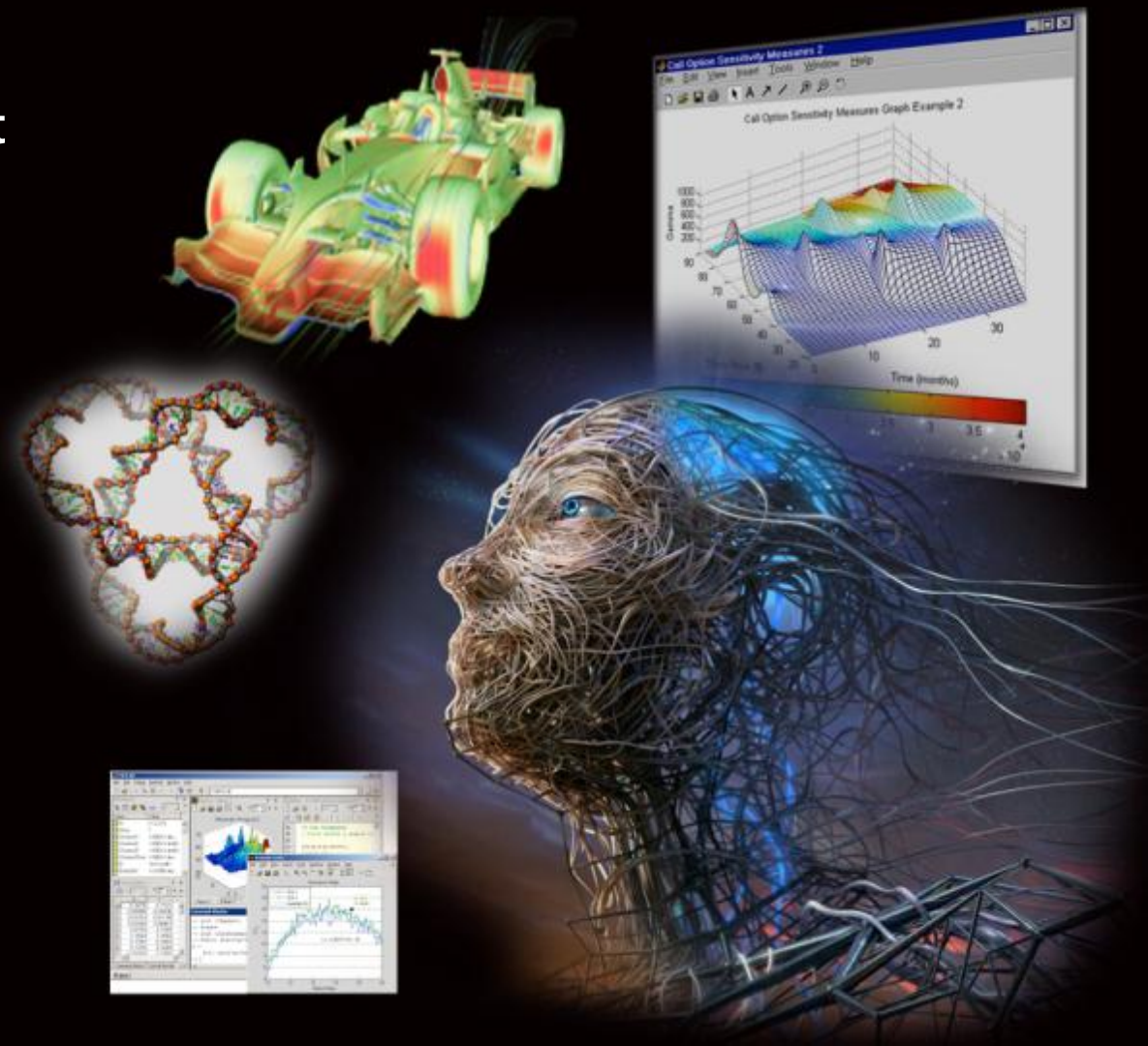
Direct Transfers

CUDA Roadmap



Strategic Focus on Applications

- Senior-level relationship and market managers
- Dedicated technical resources
- More than 150 people devoted to libraries, tools, application porting and market development
- Worldwide focus



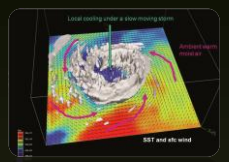
Wide Adoption of Tesla GPUs

Oil and gas

Reverse Time Migration
Kirchoff Time Migration
Reservoir Sim

Chevron
TOTAL
PETROBRAS
Paradigm
Schlumberger

Edu/Research



Astrophysics
Molecular Dynamics
Weather / Climate Modeling

Georgia Institute of Technology
HARVARD School of Engineering and Applied Sciences
UNIVERSITY OF CAMBRIDGE

Government

Signal Processing
Satellite Imaging
Video Analytics
Synthetic Aperture Radar

NAVAIR
BAE SYSTEMS
NASA
ARL
DIGITALGLOBE

Life Sciences

Bio-chemistry
Bio-informatics
Material Science
Sequence Analysis
Genomics

Boston Scientific
Delivering what's next.
MGH
MASS GENERAL HOSPITAL
WISCONSIN
BECKMAN COULTER
MAX-PLANCK-GESELLSCHAFT

Finance

Risk Analytics
Monte Carlo
Options Pricing
Insurance modeling

Bloomberg
BNP PARIBAS
NumeriX
SciComp Inc.

Manufacturing

Structural Mechanics
Computational Fluid Dynamics
Machine Vision
Electromagnetics

ANSYS
SIMULIA
ACUSIM SOFTWARE
CST
Agilent
MIT EC
Autodesk

Strategic Partners

CAD/ CAM/ CAID	CAE/ EDA	Computational chemistry	Computational Finance	Defence & Intelligence	Digital Content creation	Physical Sciences	Seismic processing and visualization
Autodesk	Ansys	Amber	MATLAB	Ikena	Adobe	Quda (L-QCD)	Schlumberger
Dassault Systemes: CATIA Solidworks	Dassault Systemes: Simulia	NAMD	Mathematica	Intergraph	Autodesk M&E	WRF	Landmark
PTC	Nastran	Gromacs	NAG	ESRI	Avid	ACUSA	Paradigm
Siemens	LSTC	Lammps	Murex	Manifold	MainConcept	HOMME	
	Synopsys	GAMESS			Sony	HYCOM	

CUDA GPU Roadmap

NVIDIA Tesla M2070-Q

With full Quadro OpenGL

- Quadro Enhanced performance profile
- Quadro Optimized profiles for professional applications
- Certified applications
- Remote OGL Visualization Support
- 3rd Party Remote Support: MS RemoteFX, Citrix HDX
- AXE – Application Acceleration Engines
- Multi GPU: SLI Multi- OS, SLI Mosaic

Same dimensions, same thermal as Tesla M2070
Dual slot width, no video out connector

40% Smaller Tesla Module

Tesla X

Tesla M-series

Fastest Supercomputer in Russia

Moscow State University

1.3 Petaflops Peak

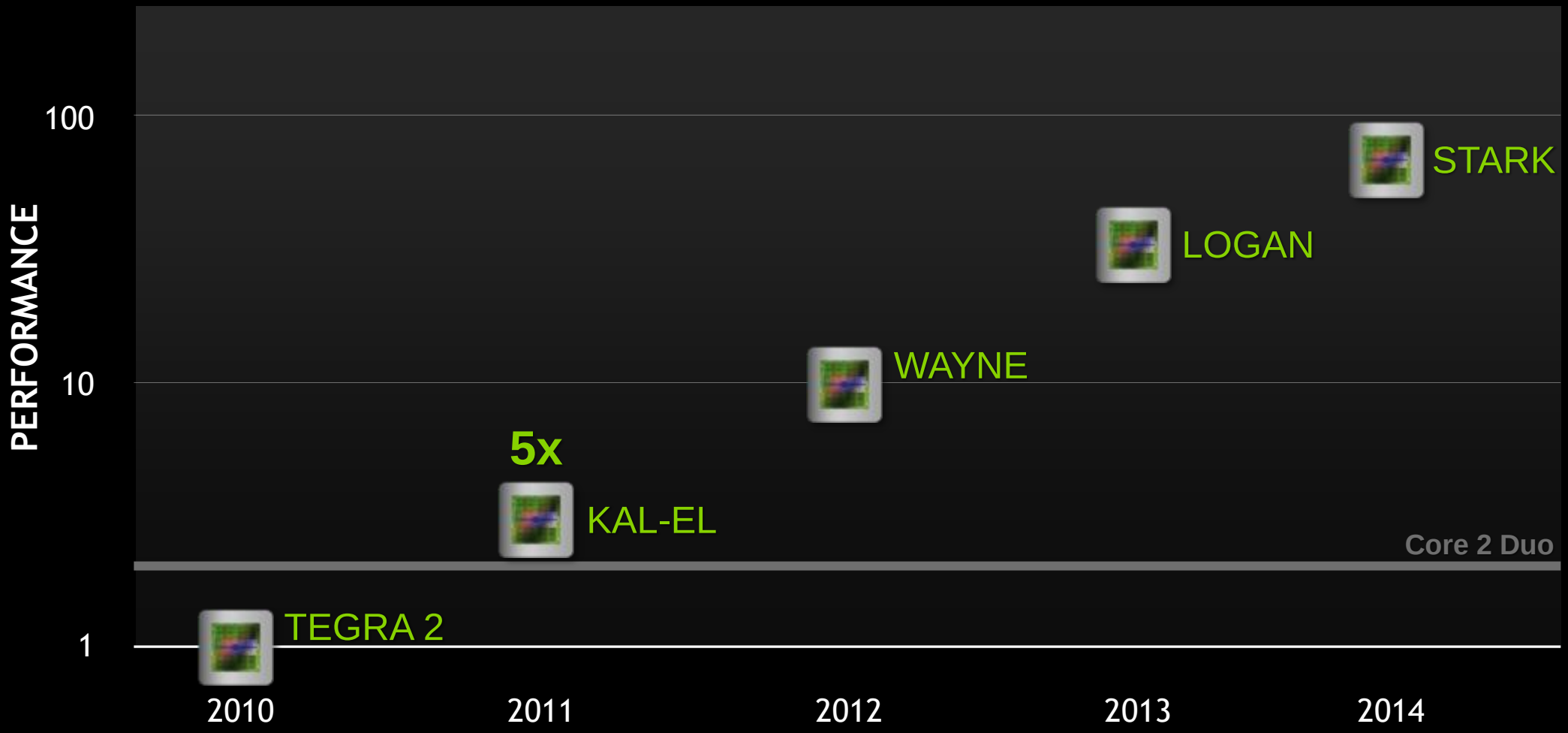
1554 Tesla X2070 GPUs

+1554 Intel CPUs

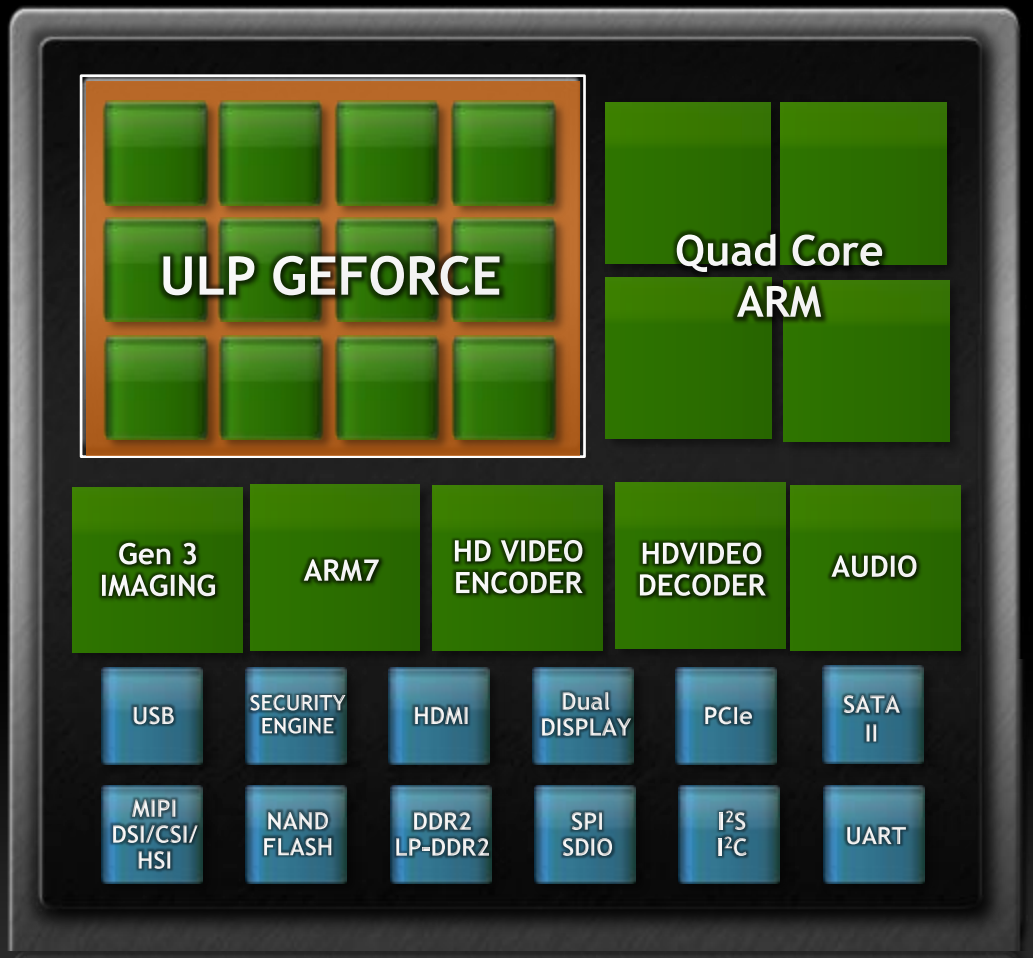
Research Topics

- **Ocean Modeling**
- **Climate Change**
- **Genomic Medicine**
- **Galaxy Formation**

Tegra Roadmap

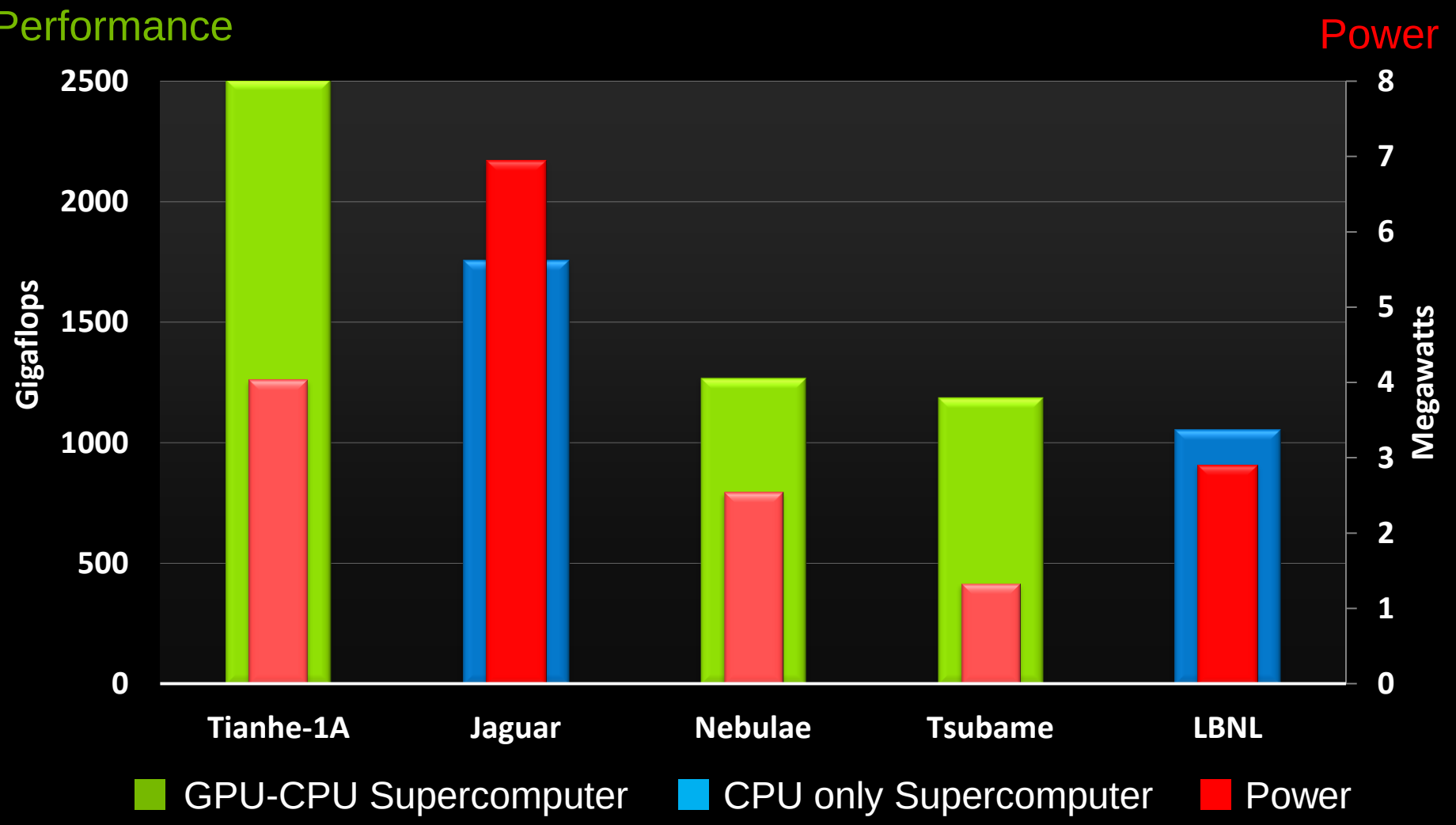


Kal-El — Independent, Dedicated Processing



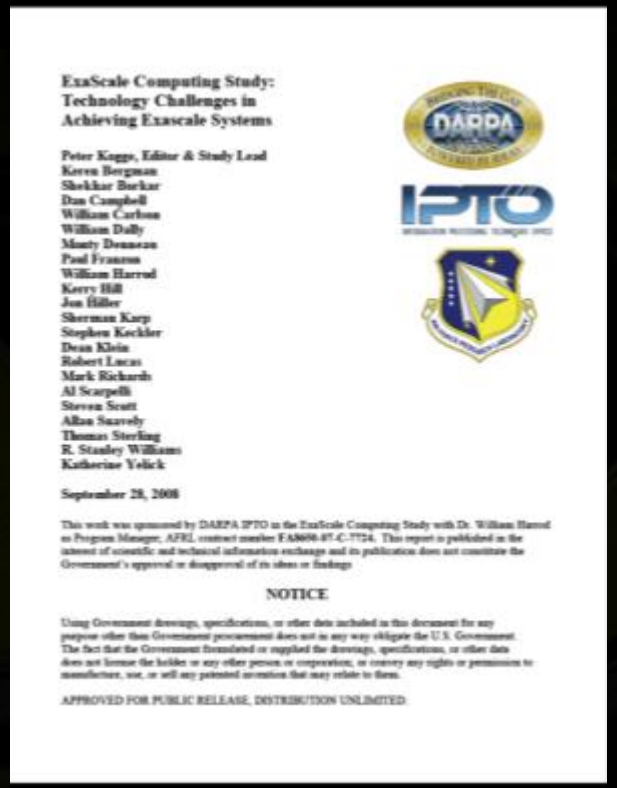
CPU	3X Performance *Quad Core up to 1.5GHz, NEON*
POWER	20x Lower Power *Due to ULP mode*
VIDEO	4X Complexity *1080i/p High Profile*
GRAPHICS	3X Performance *12 Core, Dual Pixel Pipe*
MEMORY	3X bandwidth *DDR3L up to 1600 data rate*
IMAGING	Better noise reduction & color rendition *Two simultaneous streams*
AUDIO	HD Audio *7.1 channel surround*
STORAGE	2 - 6X faster *e.MMC 4.4 and SATA-II*

GPU Supercomputers: More Power Efficient



To ExaScale and Beyond

DARPA Study Identifies Four Challenges for ExaScale Computing



ExaScale Computing Study:
Technology Challenges in
Achieving Exascale Systems

Peter Kogge, Editor & Study Lead
Keren Bergman
Shekhar Borkar
Dan Campbell
William Carlson
William Dally
Monty Denneau
Paul Franzon
William Harrod
Kerry Hill
Jon Hiller
Sherman Karp
Stephen Keckler
Dean Klein
Robert Lucas
Mark Richards
Al Scarpelli
Steven Scott
Allan Snavely
Thomas Sterling
R. Stanley Williams
Katherine Yelick

September 28, 2008

This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod as Program Manager; AFRL contract number FA8650-07-C-7724. This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings

NOTICE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED

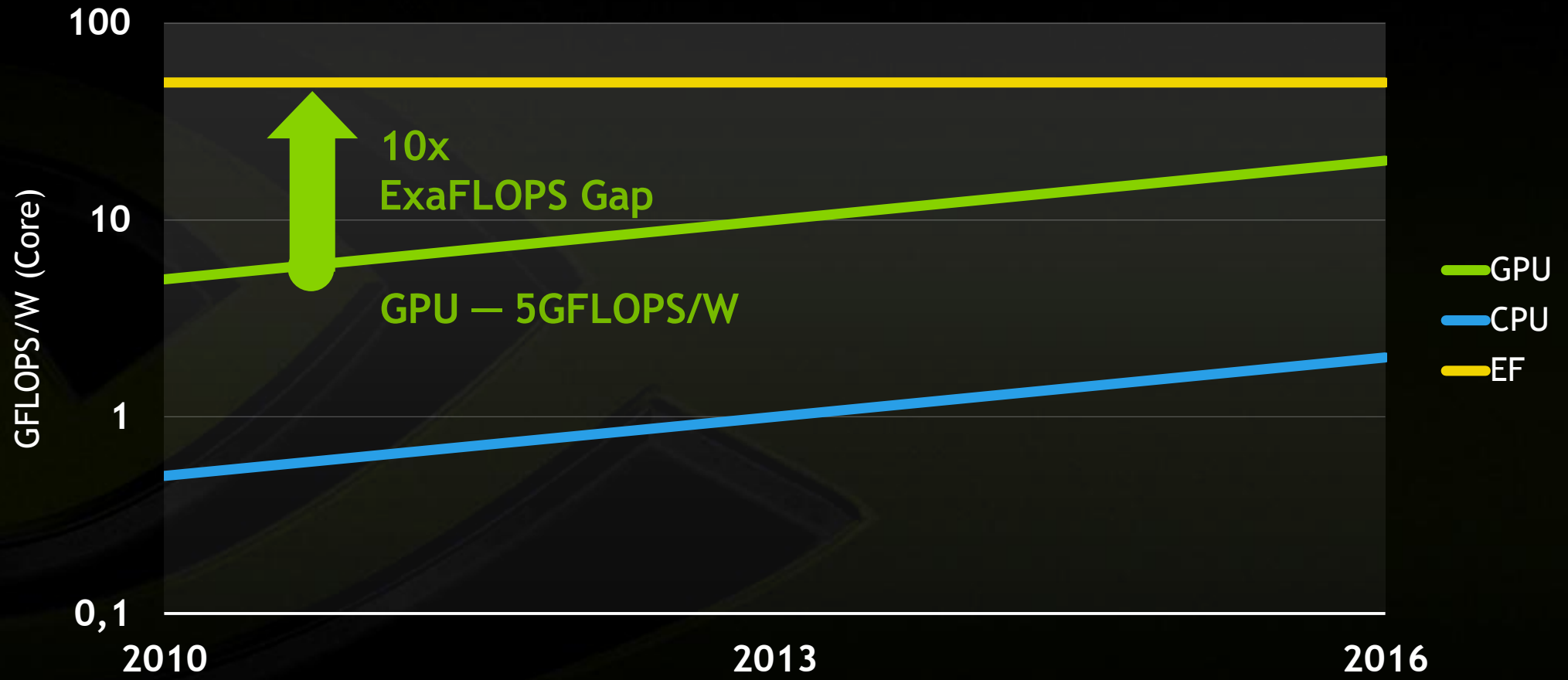
Available at
www.darpa.mil/ipto/personnel/docs/ExaScale_Study_Initial.pdf

Report published September 28, 2008:

- **Four Major Challenges**
  - Energy and Power challenge
  - Memory and Storage challenge
  - Concurrency and Locality challenge
  - Resiliency challenge
- **Number one issue is *power***
  - Extrapolations of current architectures and technology indicate over 100MW for an Exaflop!
  - Power also constrains what we can put on a chip

ExaFLOPS at 20MW = 50GFLOPS/W

50GFLOPS/W
10x Energy Gap for Today's GPU

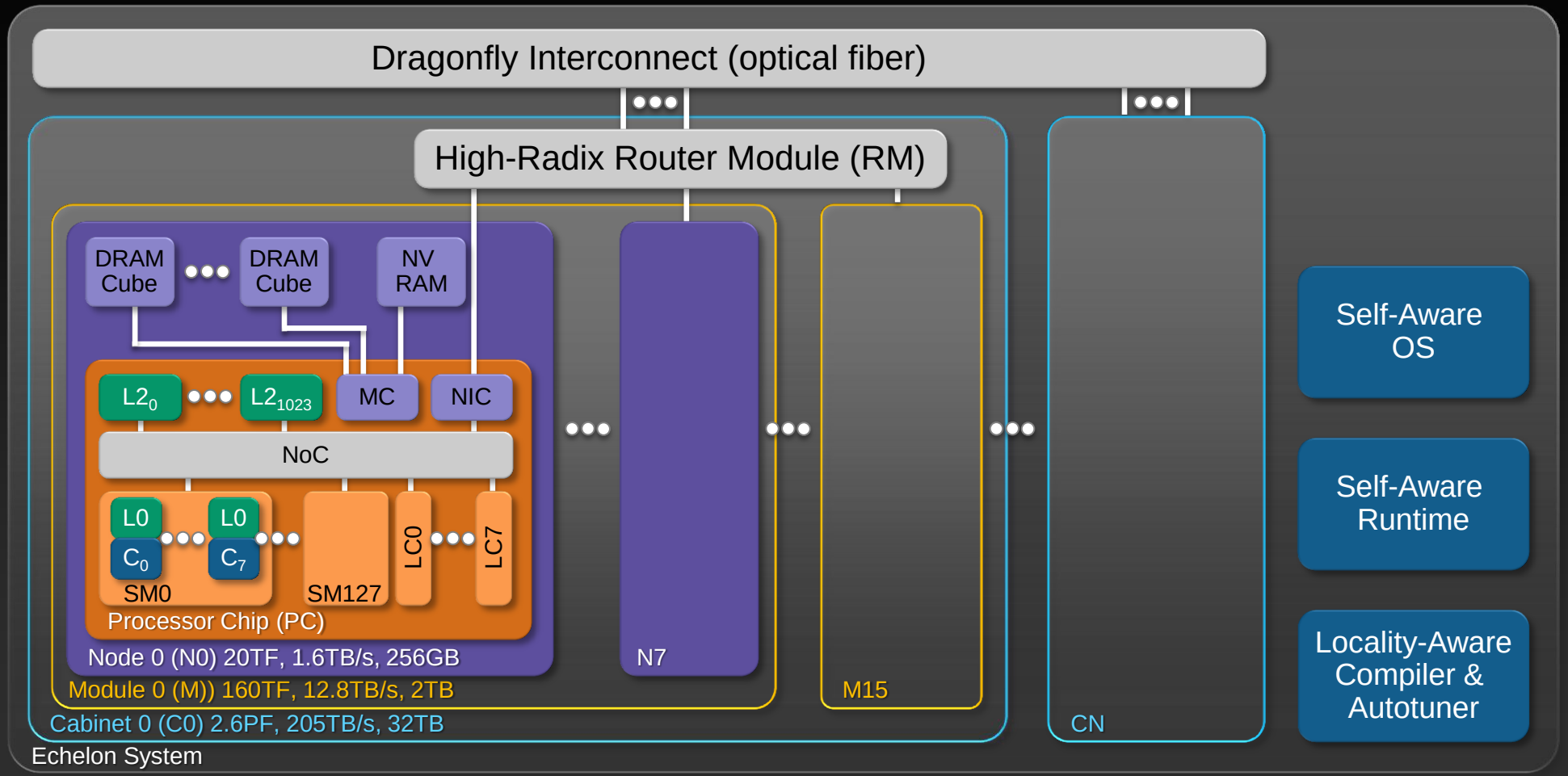


Echelon

NVIDIA’s Extreme-Scale Computing Project

Echelon Team

System Sketch



Jean-Christophe Baratault
jbaratault@nvidia.com
Cell +33 6 8036 8483

NVIDIA