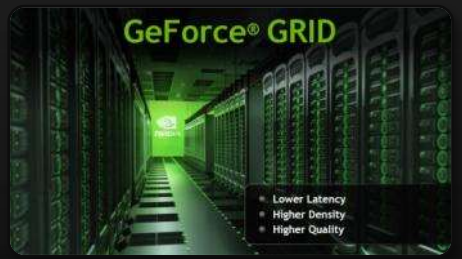
GPU Computing

Axel Koehler
Sr. Solution Architect HPC

NVIDIA

NVIDIA: Parallel Computing Company

GPUs: GeForce, Quadro, Tesla

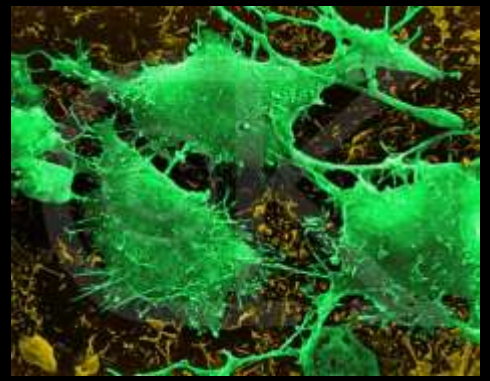


ARM SoCs: Tegra

Continued Demand for Ever Faster Supercomputers

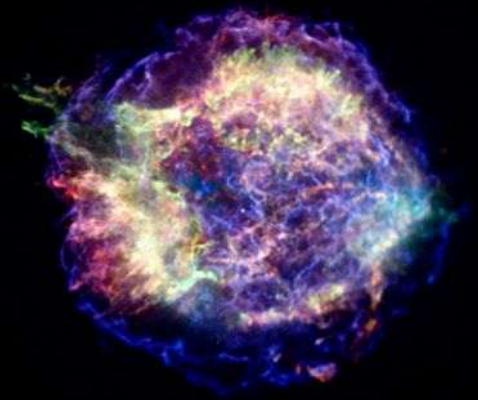
Comprehensive Earth System Model at 1KM scale, enabling modeling of cloud convection and ocean eddies.

First-principles simulation of combustion for new high-efficiency, low-emision engines.

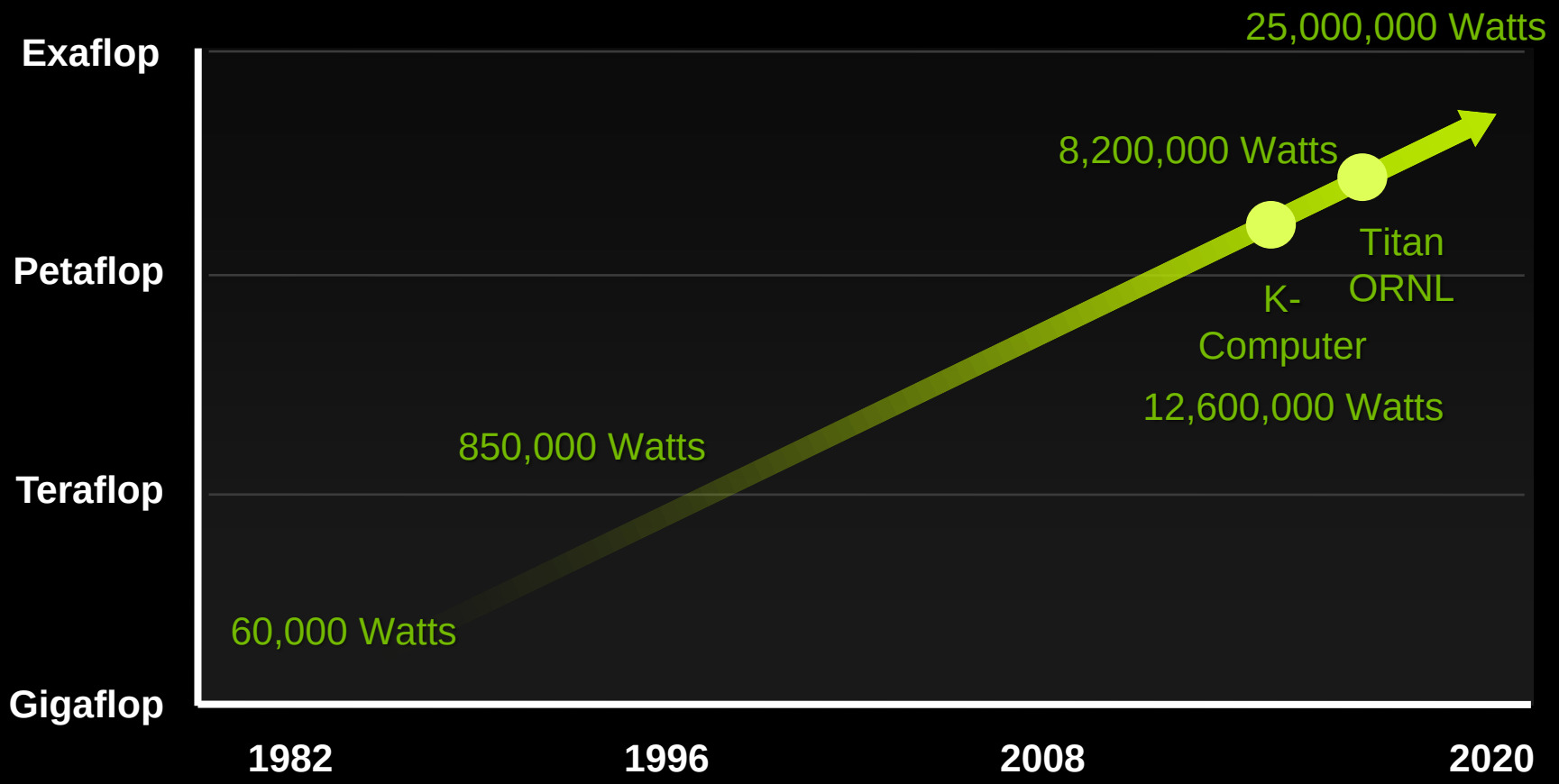


Coupled simulation of entire cells at molecular, genetic, chemical and biological levels.

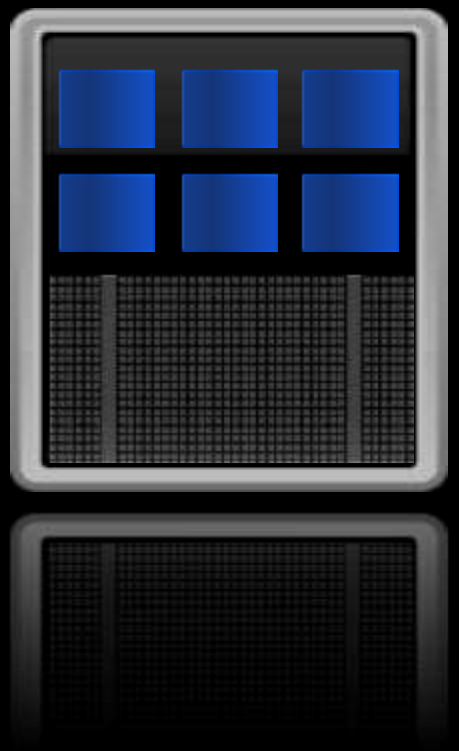
Predictive calculations for thermonuclear and core-collapse supernovae, allowing confirmation of theoretical models.



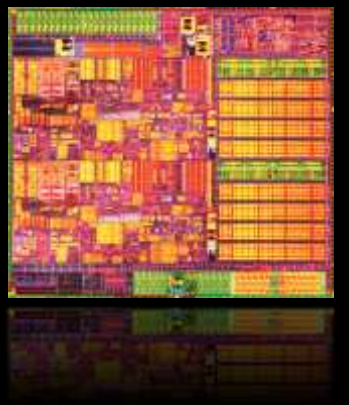
Power Crisis in Supercomputing



Multi-core CPUs



- Industry has gone multi-core as a first response to power issues
  - Performance through parallelism, not frequency
- But CPUs are fundamentally designed for single thread performance rather than energy efficiency
  - Fast clock rates with deep pipelines
  - Data and instruction caches optimized for latency
  - Superscalar issue with out-of-order execution
  - Lots of predictions and speculative execution
  - Lots of instruction overhead per operation

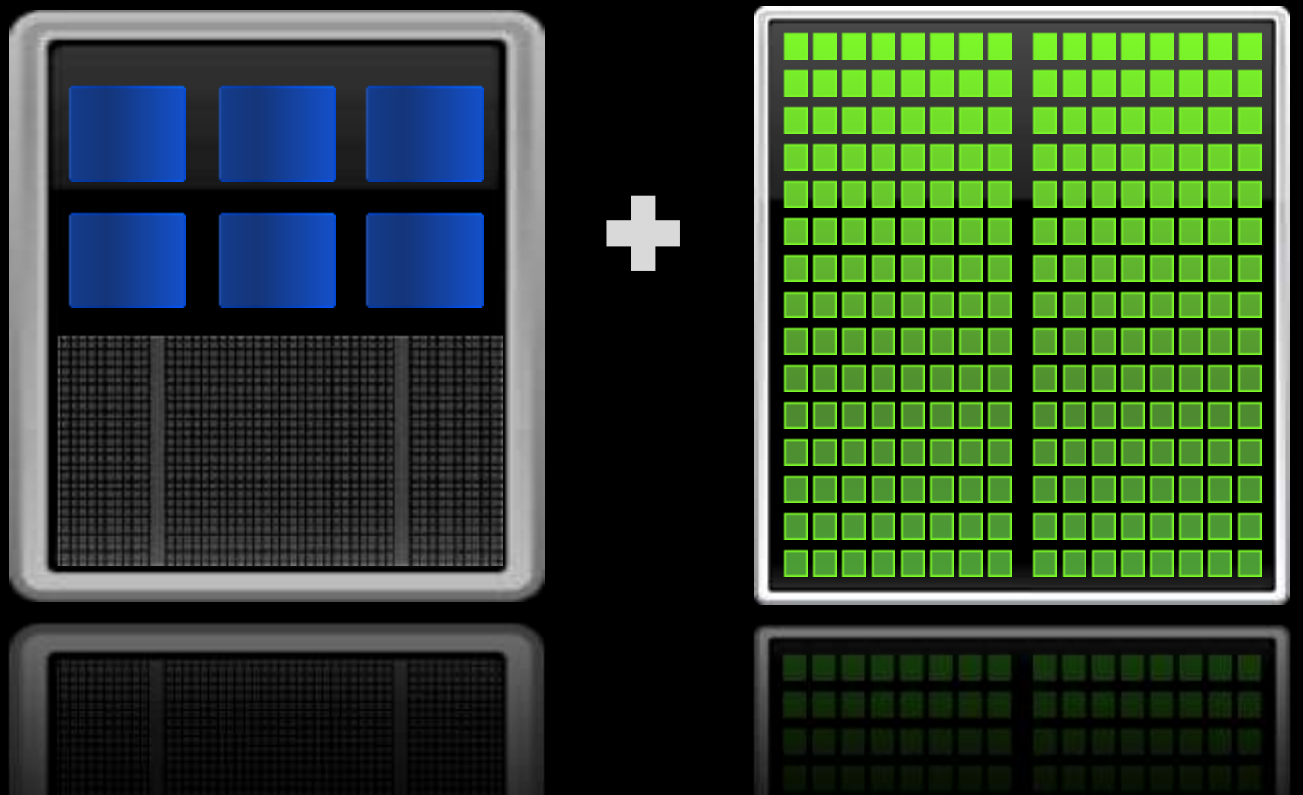


Less than 2% of chip power today goes to flops.

Accelerated Computing

Add GPUs: Accelerate Applications

CPUs: designed to run a few tasks quickly.



GPUs: designed to run many tasks *efficiently*.

Energy efficient GPU

Performance = Throughput

- **Fixed function hardware**
  - Transistors are primarily devoted to data processing
  - Less leaky cache
- **SIMT thread execution**
  - Groups of threads formed into warps which always executing same instruction
- **Cooperative sharing of units with SIMT**
  - eg. fetch instruction on behalf of several threads or read memory location and broadcast to several registers
- **Lack of speculation reduces overhead**
- **Minimal Overhead**
  - Hardware managed parallel thread execution and handling of divergence

Supercomputing

Weather / Climate Modeling
Molecular Dynamics
Computational Physics

Life Sciences

Biochemistry
Bioinformatics
Material Science

Manufacturing

Structural Mechanics
Comp Fluid Dynamics (CFD)
Electromagnetics

Defense / Govt

Signal Processing
Image Processing
Video Analytics

Oil and Gas

Reverse Time Migration
Kirchoff Time Migration

Tesla K20/K20X

Kepler GK110

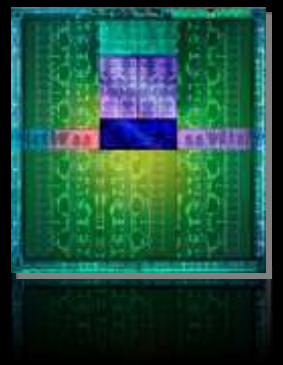
Tesla K10

Kepler GK104

Product Name	K10	K20	K20X
GPU Architecture	Kepler: GK104	GK110	GK110
# of GPUs	2	1	1
Peak Single Flops Peak SGEMM	4.58 TF (2.3 TF per GPU) 2.98 TF	3.52 TF 2.61 TF	3.95 TF 2.90 TF
Peak Double Flops Peak DGEMM	0.19 TF (0.095 TF per GPU) 0.12 TF	1.17 TF 1.10 TF	1.32 TF 1.22 TF
Memory size	8 GB (4GB per GPU)	5 GB	6 GB
Memory BW (ECC off)	320 GB/s (160GB/s per GPU)	208 GB/s	250 GB/s
New CUDA Features	GPUDirect w/ RDMA	GPUDirect (RDMA), Hyper-Q, Dynamic Parallelism	
ECC Features	External DRAMs only	DRAM, Caches & Reg Files	
# CUDA Cores	3072 (1536 per GPU)	2496	2688
Total Board Power	225W	225W	235W
Board Type	PCI-e Passive	PCI-e Passive, Active, SXM	PCI-e Passive SXM

Kepler GK110 Block Diagram

- 7.1B Transistors
- 15 SMX units
- 1.3 TFLOP FP64
- 1.5 MB L2 Cache
- 384-bit GDDR5
- PCI Express Gen3 compliant



Kepler GK110 SMX vs Fermi SM

3x sustained perf/W

Ground up redesign for perf/W

6x the SP FP units

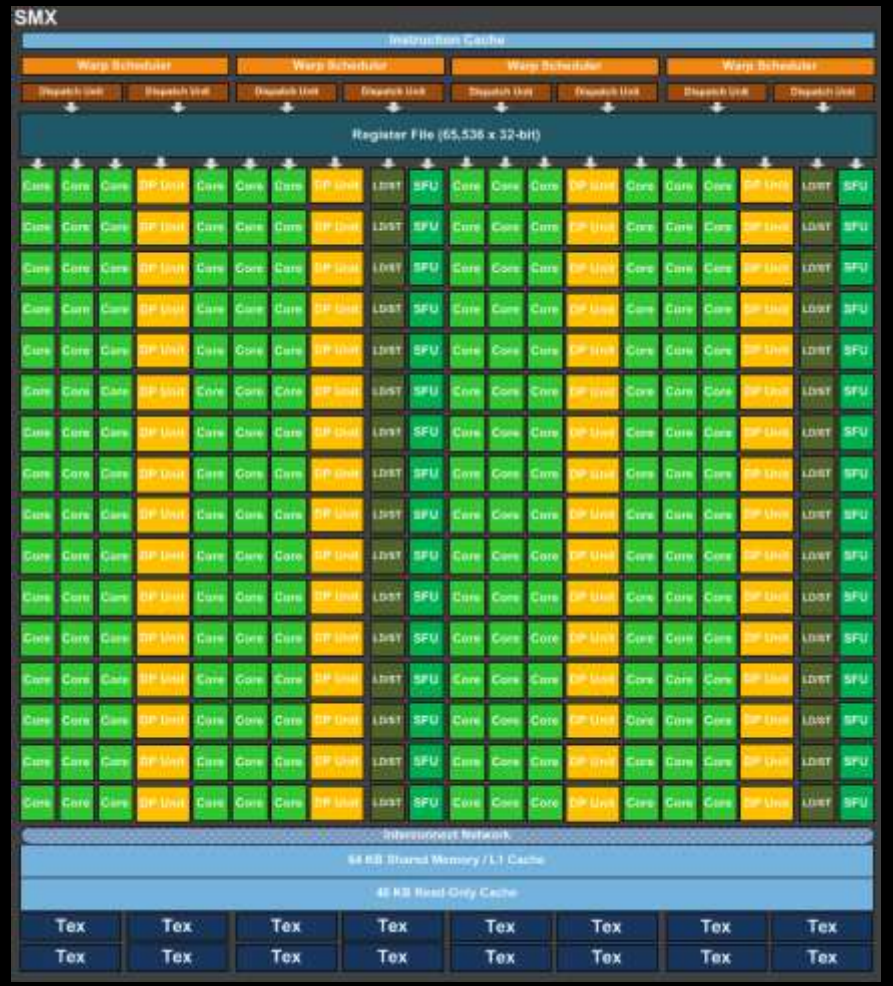
4x the DP FP units

Slower FU clocks

~4x the overall instruction throughput

2x register file size (64K regs)

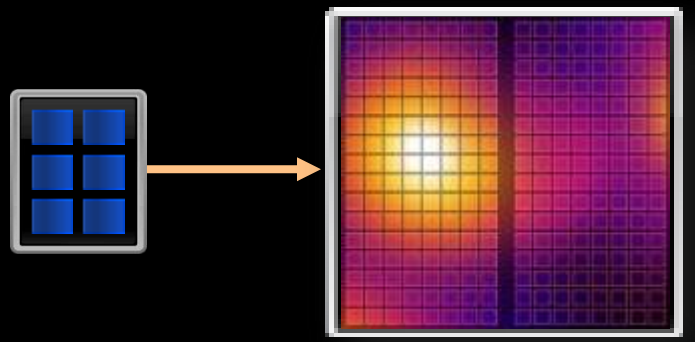
2x threadblocks (16) & 1.33x threads (2K)



Hyper-Q

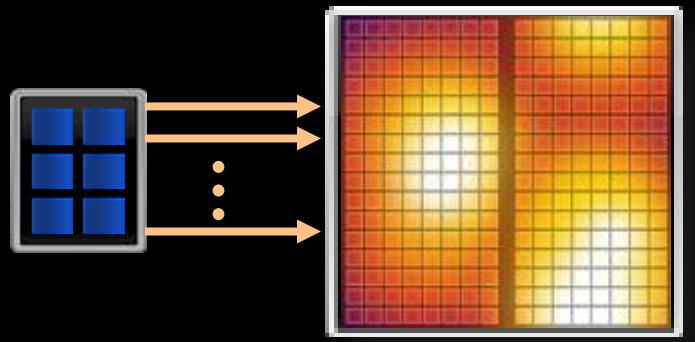
FERMI

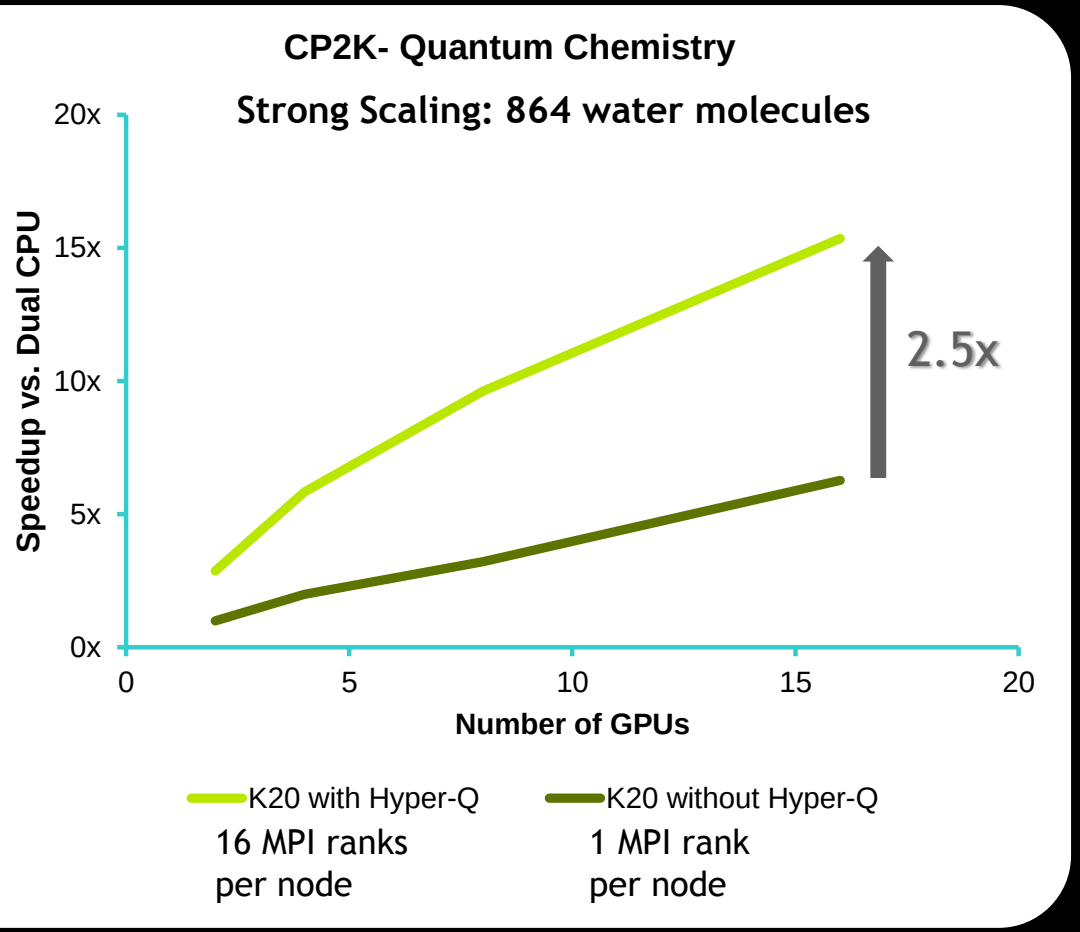
1 Work Queue



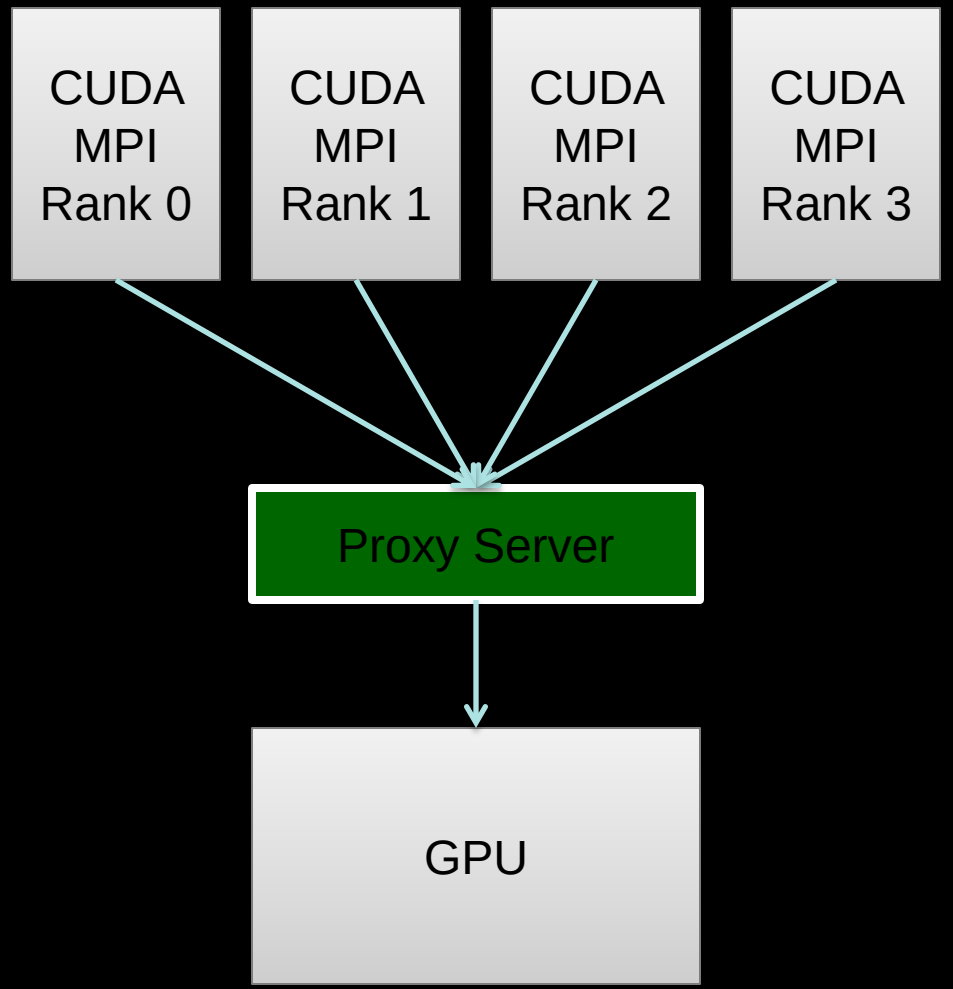
KEPLER

32 Concurrent Work Queues





Proxy - A Multi-Process Runtime for MPI

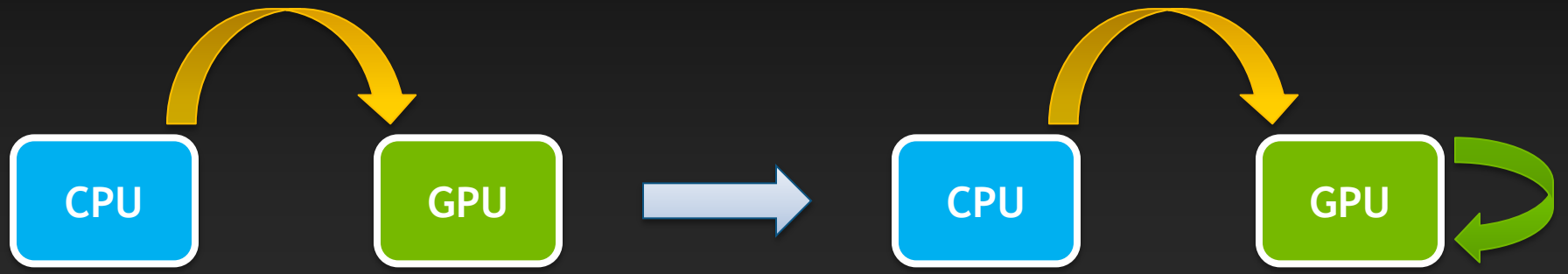


- **Why**
  - **Speedups for MPI programs with low-GPU utilization**
- **How**
  - **multiple CPU processes on a single GPU simultaneously**
  - **client-server architecture**
  - **client processes share the same CUDA context**
- **When**
  - **Currently on Cray; Production on Linux with CUDA 5.5**

What is Dynamic Parallelism?

The ability to launch new kernels from the GPU

- Dynamically - based on run-time data
- Simultaneously - from multiple threads at once
- Independently - each thread can launch a different grid

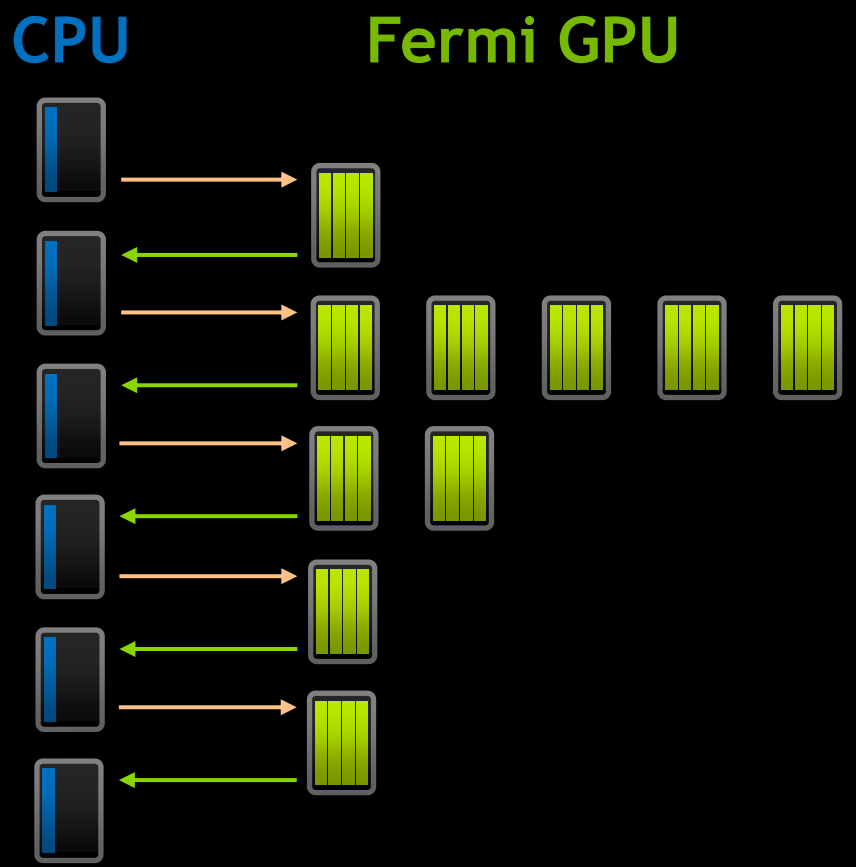


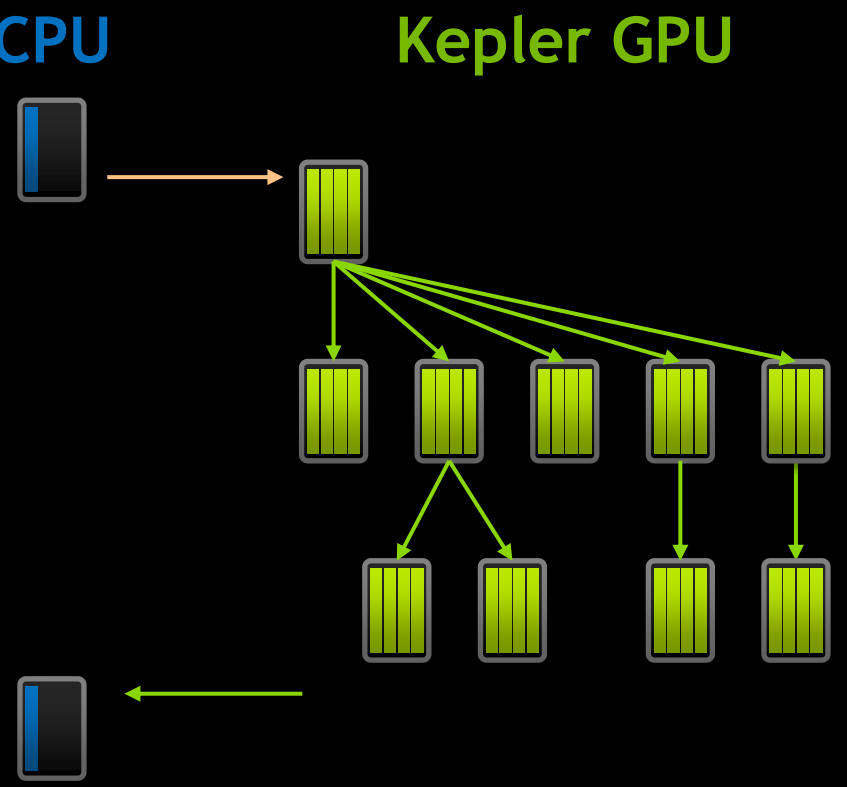
Fermi: Only CPU can generate GPU work

Kepler: GPU can generate work for itself

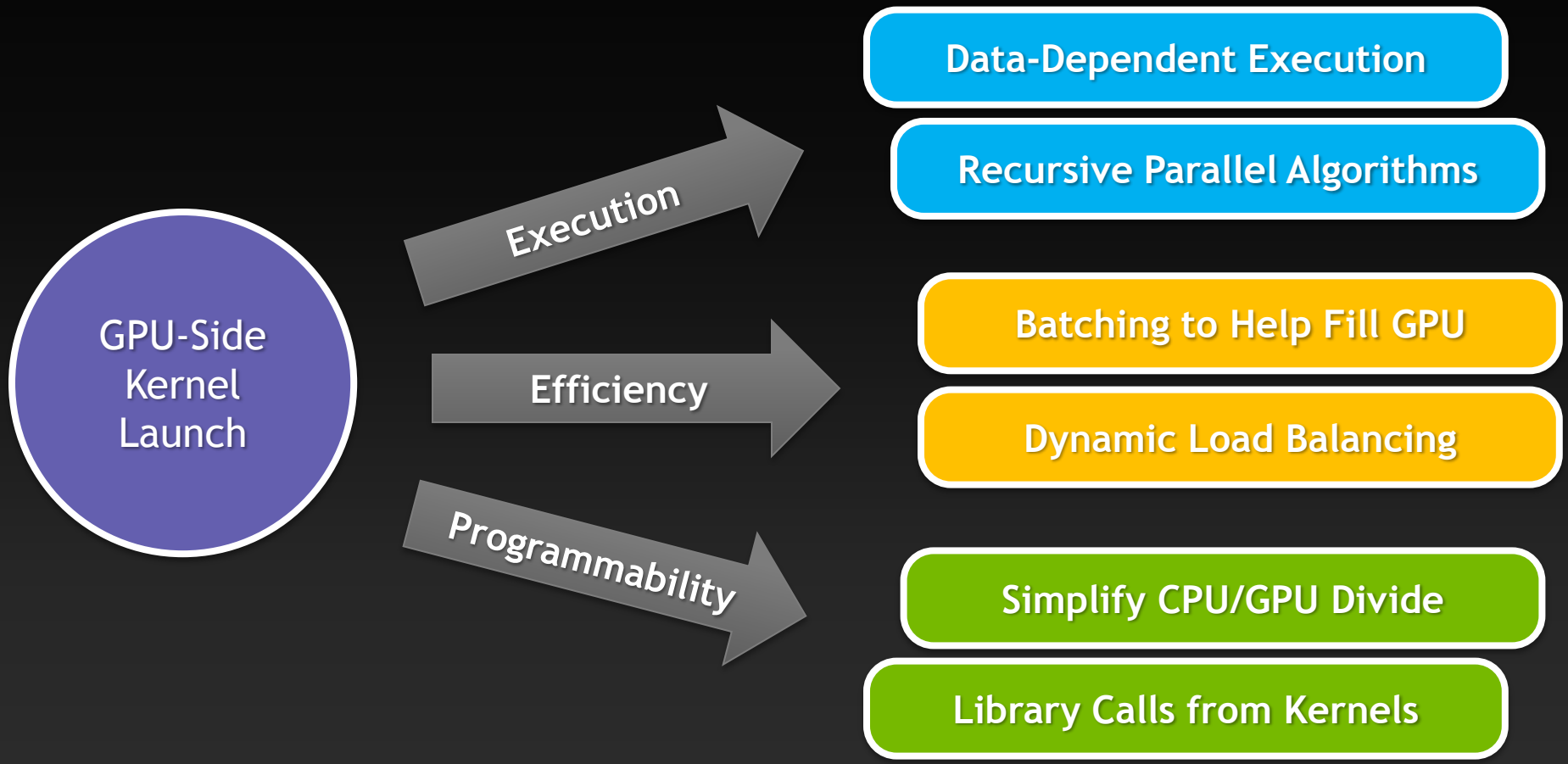
Dynamic Parallelism

Simpler Code, More General, Higher Performance





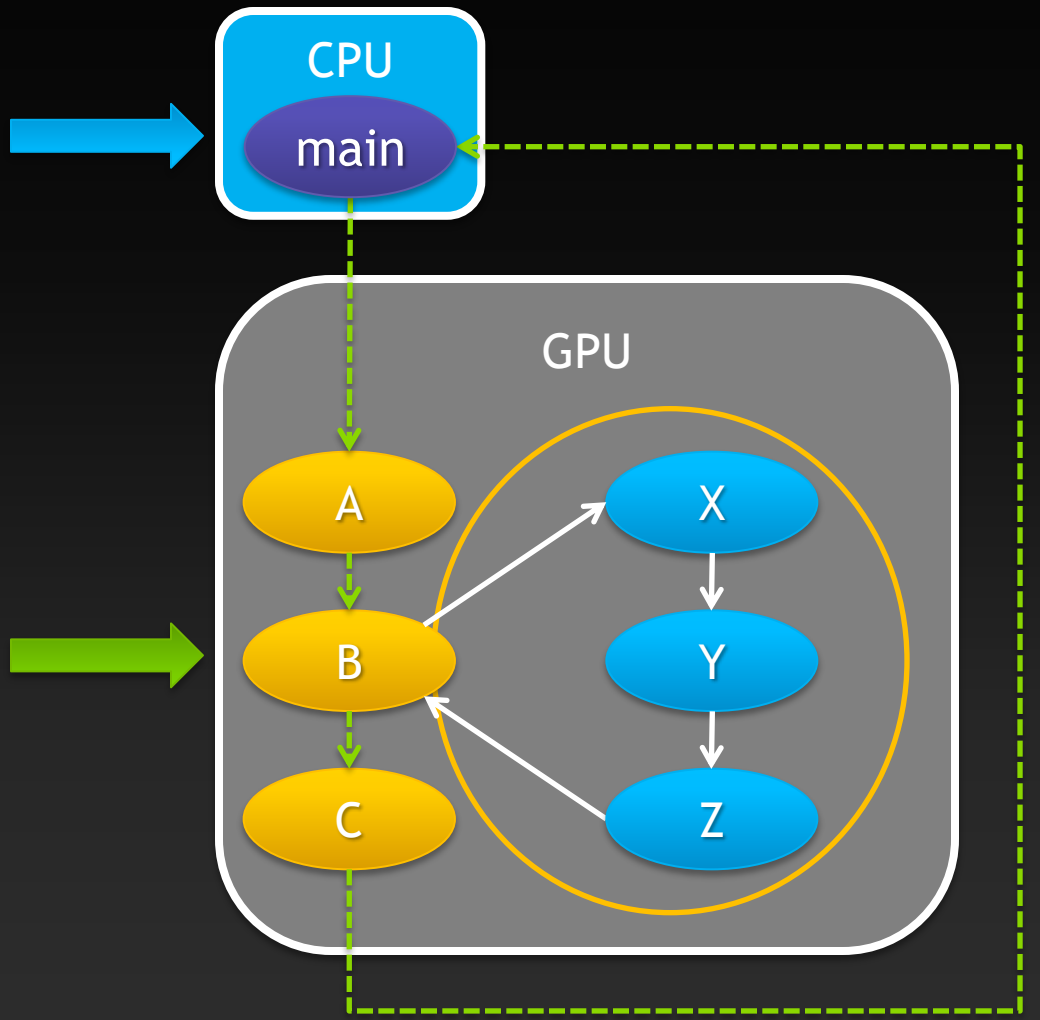
Dynamic Parallelism



Familiar Syntax and Programming Model

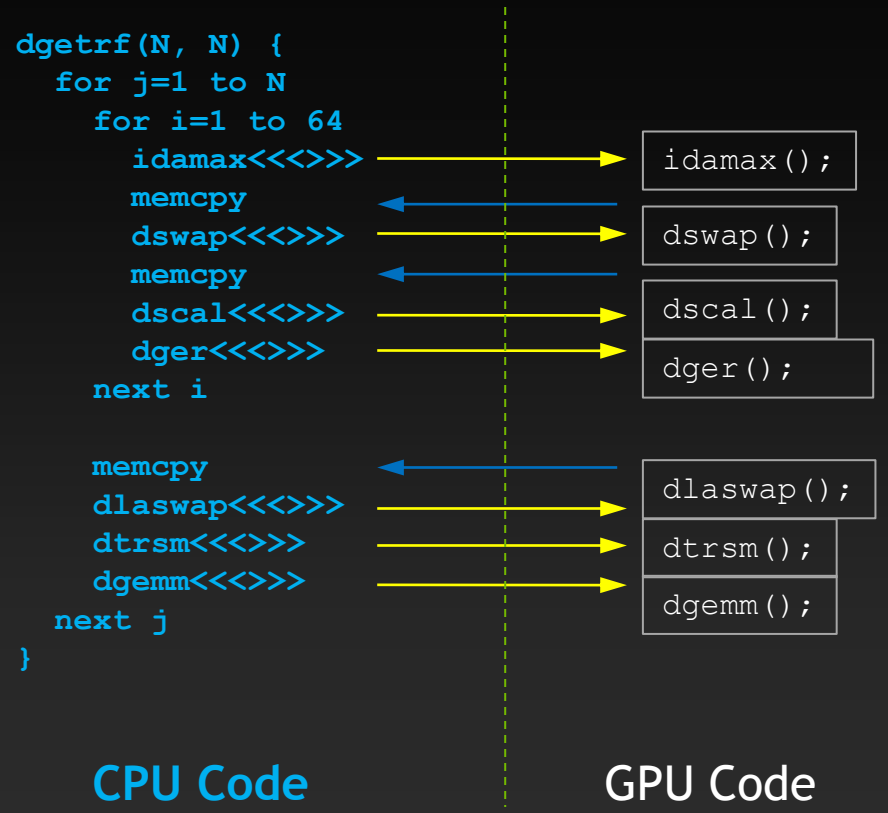
```
int main() {
    float *data;
    setup(data);

    A <<< ... >>> (data);
    B <<< ... >>> (data);
    C <<< ... >>> (data);

    cudaDeviceSynchronize();
    return 0;
}
```

```
__global__ void B(float *data)
{
    do_stuff(data);

    X <<< ... >>> (data);
    Y <<< ... >>> (data);
    Z <<< ... >>> (data);
    cudaDeviceSynchronize();

    do_more_stuff(data);
}
```

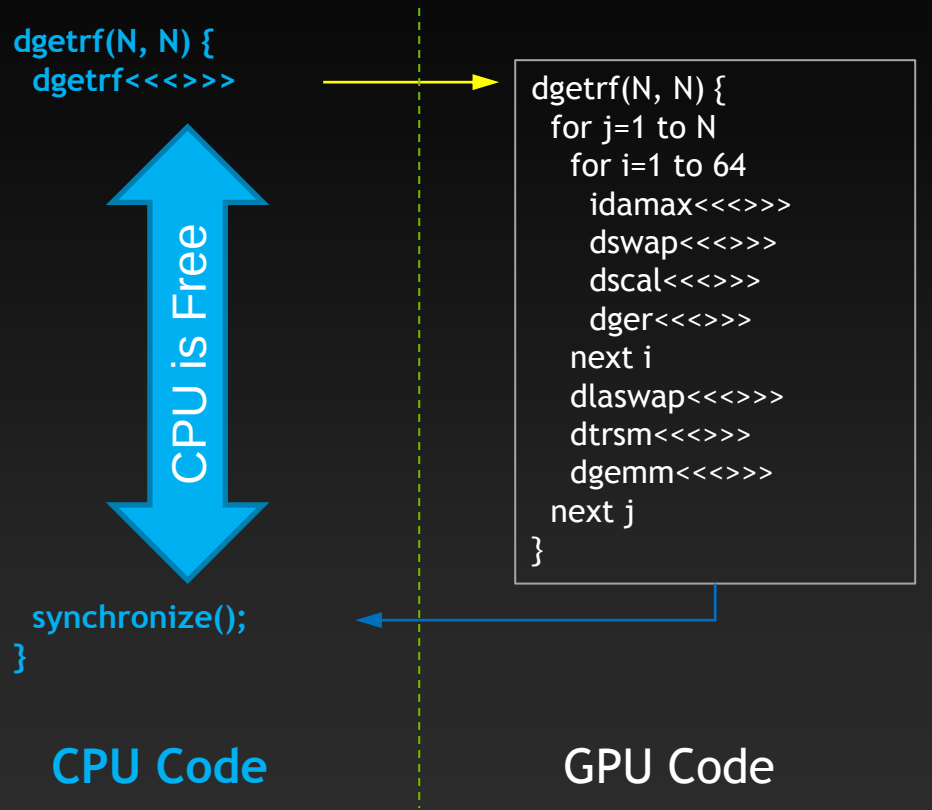


Simpler Code: LU Example

LU decomposition (Fermi)

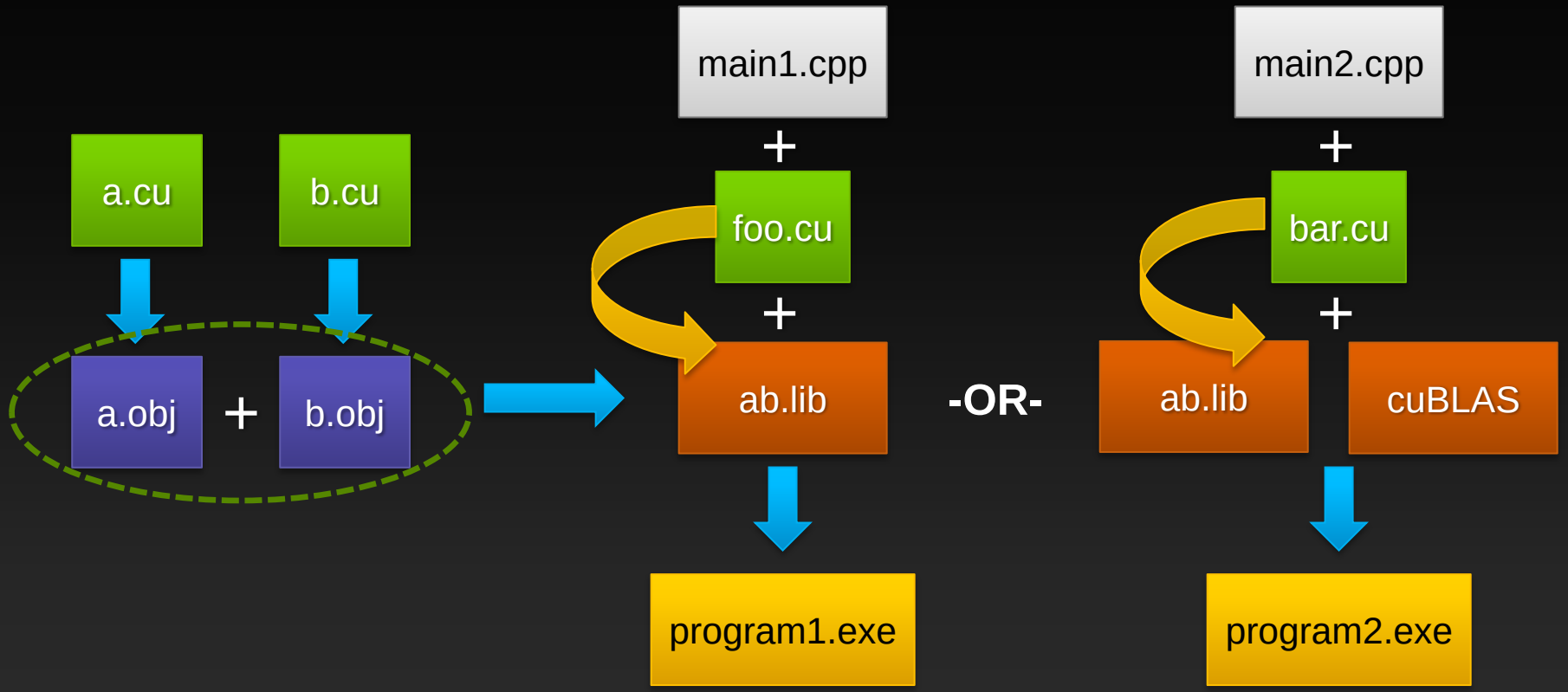


LU decomposition (Kepler)



GPU Callable Libraries

Direct Access to BLAS and other Libraries from GPU Code

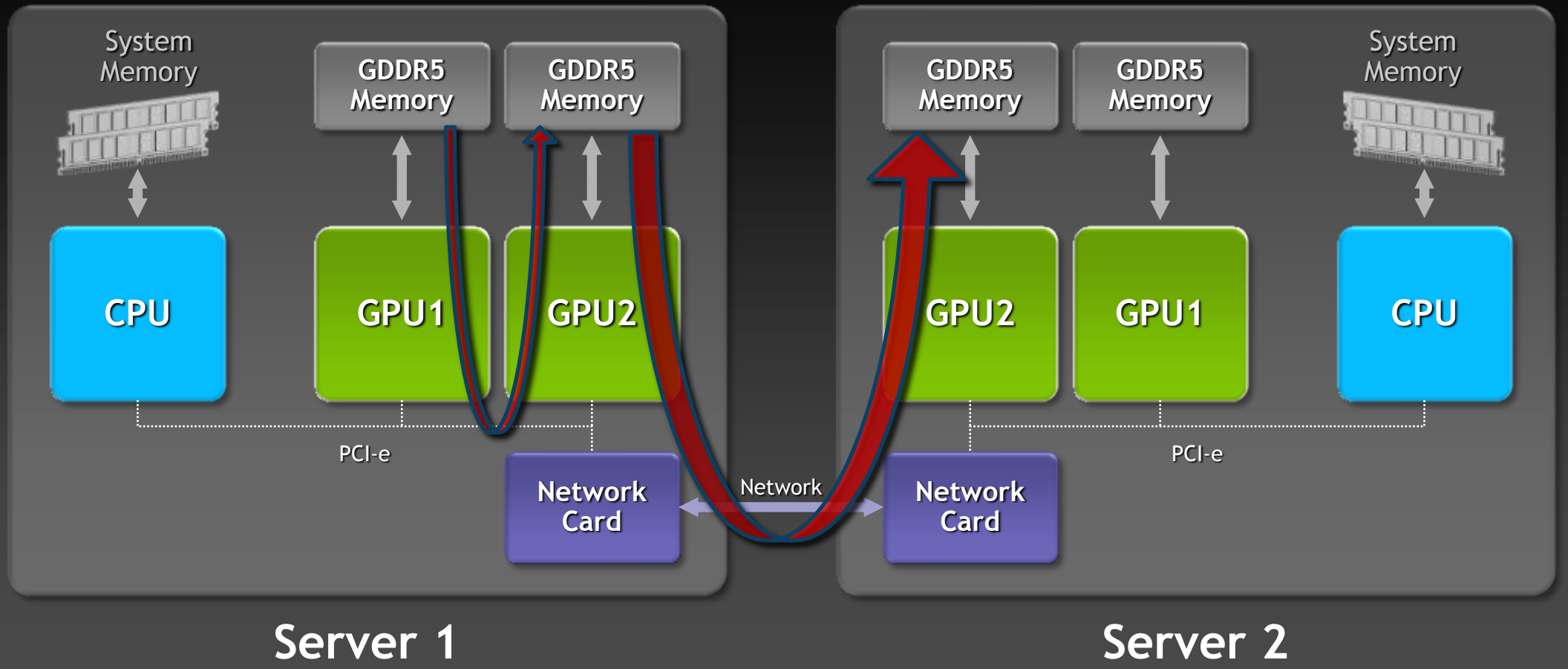


Source files compiled separately to create independent object files

Linker creates GPU Callable Libraries and links with CUDA code

NVIDIA® GPUDirect™ Support for RDMA

Direct Communication Between GPUs and PCIe devices



GPU-aware MPI

- Support GPU to GPU communication through standard MPI interfaces without exposing low level details to the programmer (make MPI implementations aware of GPU pointers)
  - e.g. enable MPI_Send, MPI_Recv from/to GPU memory
  - Made possible by Unified Virtual Addressing (UVA) in CUDA 4.0
  - MVAPICH2, OpenMPI, Platform MPI

Code without MPI integration

At Sender:

```
cudaMemcpy(s_buf, s_device, size, cudaMemcpyDeviceToHost);
MPI_Send(s_buf, size, MPI_CHAR, 1, 1, MPI_COMM_WORLD);
```

At Receiver:

```
MPI_Recv(r_buf, size, MPI_CHAR, 0, 1, MPI_COMM_WORLD, &req);
cudaMemcpy(r_device, r_buf, size, cudaMemcpyHostToDevice);
```

Code with MPI integration

At Sender:

```
MPI_Send(s_device, size, …);
```

At Receiver:

```
MPI_Recv(r_device, size, …);
```

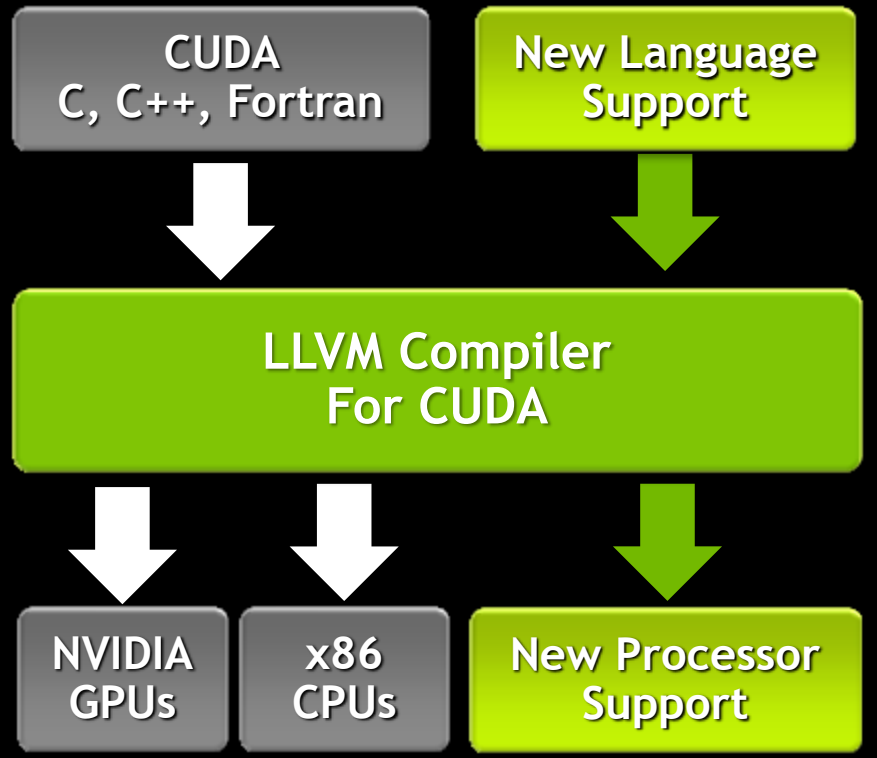
CUDA Compiler Contributed to Open Source LLVM

Developers want to build front-ends for

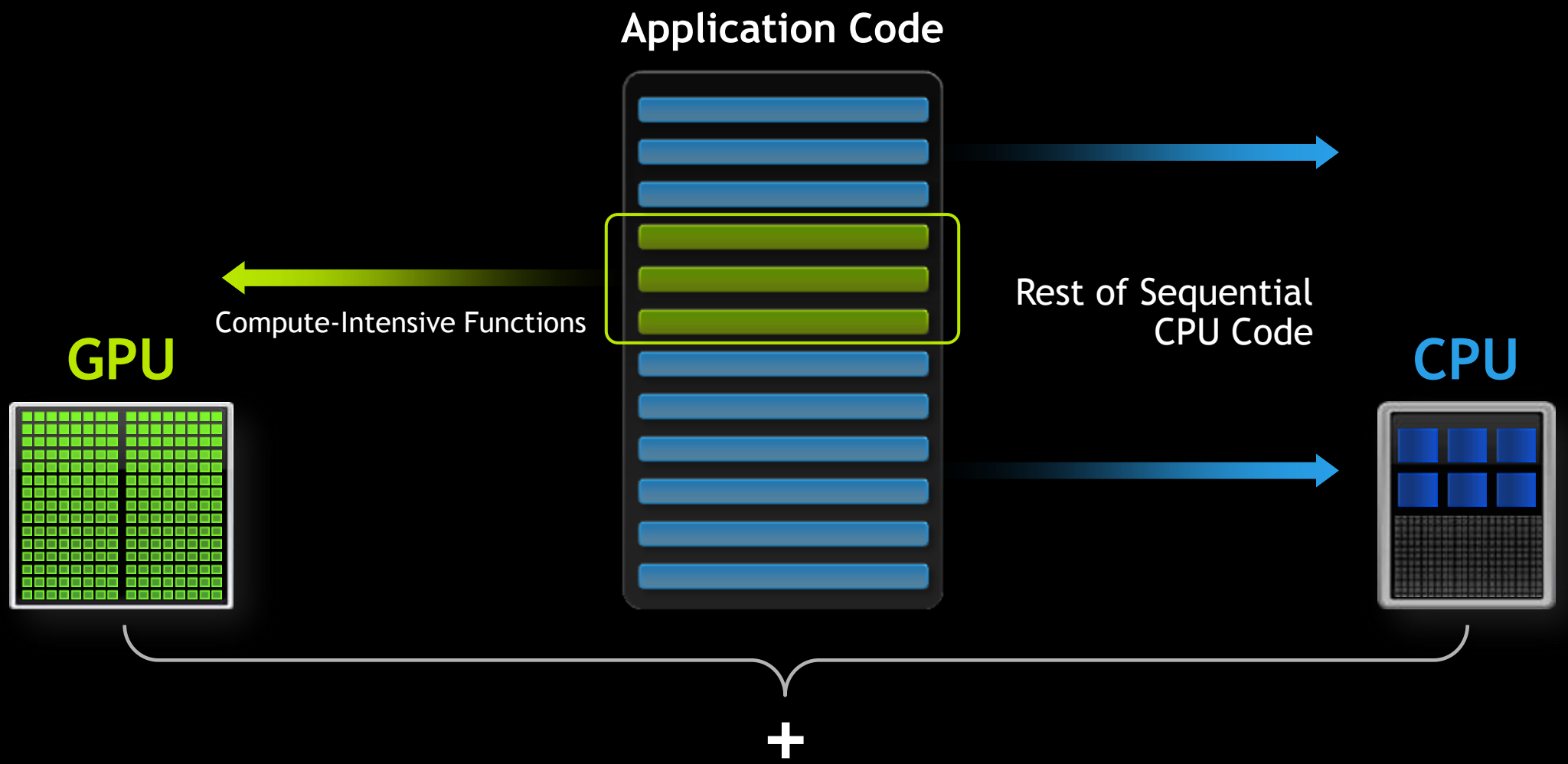
Java, Python, R, DSLs

Target other processors like

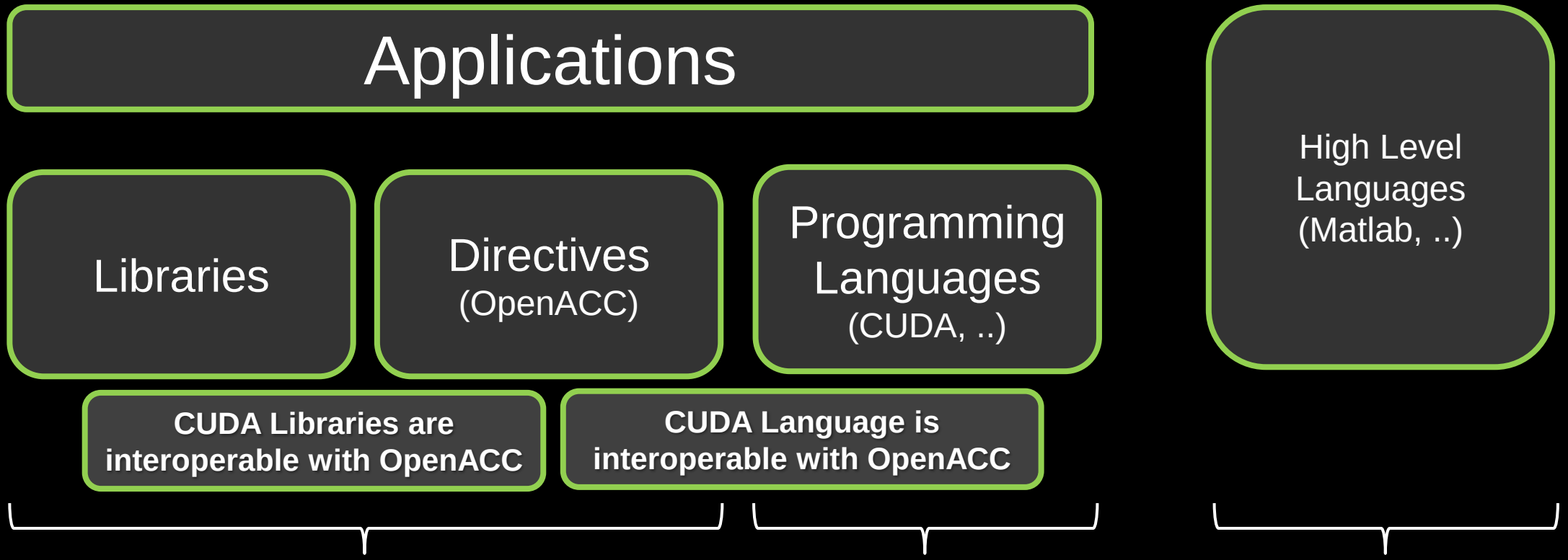
ARM, FPGA, GPUs, x86



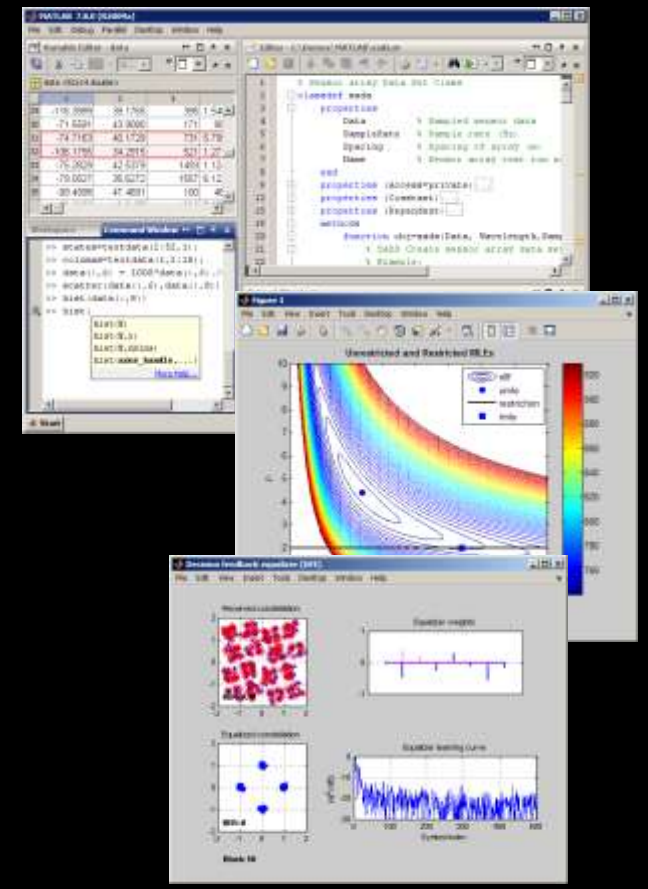
Minimum Change, Big Speed-up



Ways to Accelerate Applications



MATLAB Parallel Computing
On-ramp to GPU Computing

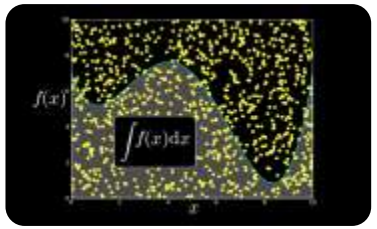


- **Most popular math functions on GPUs**
  - Random number generation
  - FFT
  - Matrix multiplications
  - Solvers
  - Convolutions
  - Min/max
  - SVD
  - Cholesky and LU factorization
- **MATLAB Compiler support (GPU acceleration without MATLAB installed)**
- **GPU features in Communications Systems Toolbox**

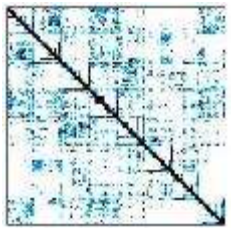
GPU Accelerated Libraries

"Drop-in" Acceleration for Your Applications

NVIDIA cuBLAS



NVIDIA cuRAND



NVIDIA cuSPARSE

NVIDIA NPP

GPU VSIPL

Vector Signal Image Processing

CULA tools

GPU Accelerated Linear Algebra

MAGMA

Matrix Algebra on GPU and Multicore

NVIDIA cuFFT

ROGUE WAVE SOFTWARE

IMSL Library

CUSP

Sparse Linear Algebra

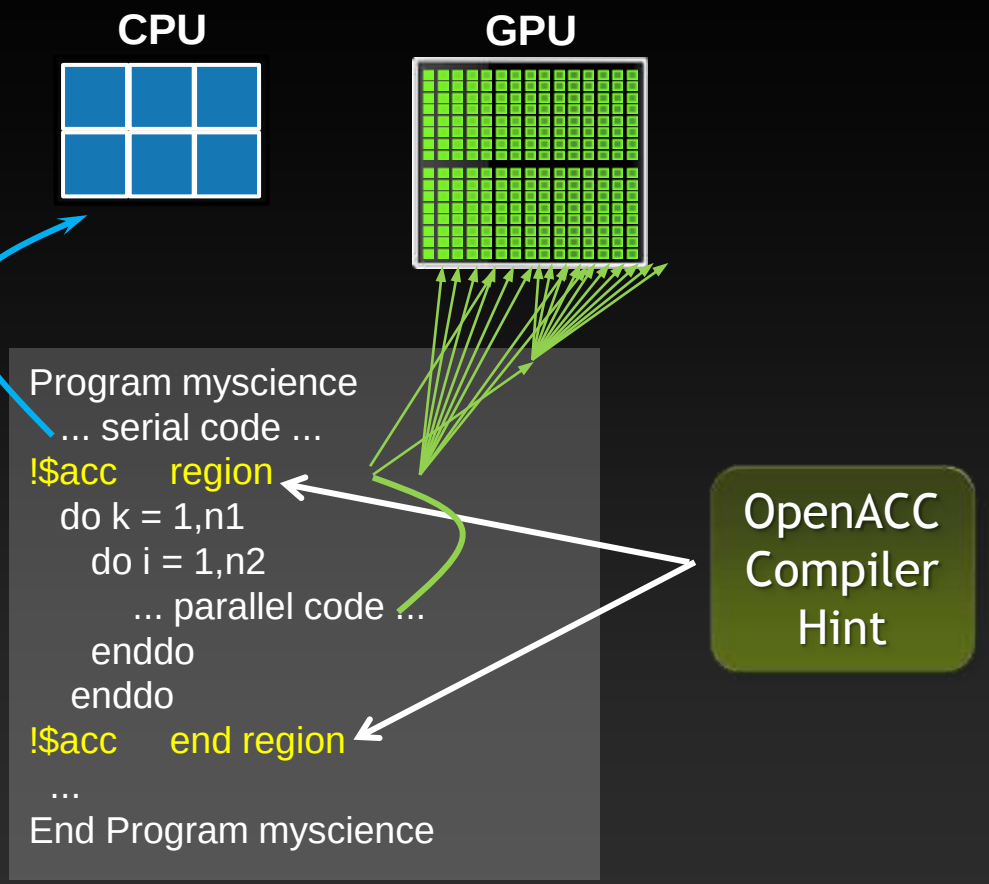
libjacket

Building-block Algorithms for CUDA

Thrust

C++ STL Features for CUDA

OpenACC Directives



Easy, Open, Powerful

- Simple Compiler hints
- Works on multicore CPUs & many core GPUs
- Compiler Parallelizes code
- Future Integration into OpenMP standard planned

http://www.openacc.org

NVIDIA CRAY PGI CAPS

OpenACC Directives Example

```
!$acc data copy(A,Anew)
 iter=0
 do while ( err > tol .and. iter < iter_max )

  iter = iter +1
  err=0._fp_kind

!$acc kernels
   do j=1,m
    do i=1,n
     Anew(i,j) = .25_fp_kind *( A(i+1,j  ) + A(i-1,j  ) &
                               +A(i  ,j-1) + A(i  ,j+1))
     err = max( err, Anew(i,j)-A(i,j))
    end do
   end do
!$acc end kernels
   IF(mod(iter,100)==0 .or. iter == 1)    print *, iter, err
   A= Anew

 end do
!$acc end data
```

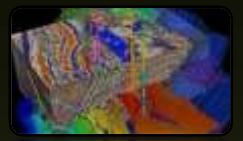
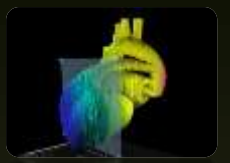
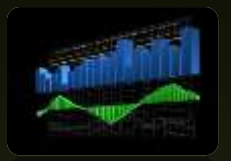
Copy arrays into GPU memory within data region

Parallelize code inside region

Close off parallel region

Close off data region, copy data back

Wide Adoption of Tesla GPUs

Oil and gas	Edu/Research	Government	Life Sciences	Finance	Manufacturing
					
Reverse Time Migration Kirchoff Time Migration Reservoir Sim	Astrophysics Lattice QCD Molecular Dynamics Weather / Climate Modeling	Signal Processing Satellite Imaging Video Analytics Synthetic Aperture Radar	Bio-chemistry Bio-informatics Material Science Sequence Analysis Genomics	Risk Analytics Monte Carlo Options Pricing Insurance modeling	Structural Mechanics Computational Fluid Dynamics Machine Vision Electromagnetics

Over 200 GPU-Accelerated Applications

http://www.nvidia.com/teslaapps/

POPULAR GPU-ACCELERATED APPLICATIONS

Application	Description	Supported Features	Expected	Multi-GPU	Release Status
Molecular Dynamics					
Abalone	Models molecular dynamics of biopol simulations of proteins, DNA and liga				
ACEMD	Simulation of mechanics force fields, & explicit solvent on CUDA				
AMBER	Suite of programs to simulate molecu dynamics on biomolecules				
DL-POLY	Simulate macromolecules, polymers, systems, etc on a distributed memory parallel computer				
GROMACS	Simulation of biochemical molecules complicated bond interactions				
HOOMD-Blue	Particle dynamics package written g up for GPUs				
LAMMPS	Classical molecular dynamics packag				
NAMD	Designed for high-performance simu of large molecular systems				
Quantum Chemistry					
GAMESS-US	Computational chemistry suite used t simulate atomic and molecular electr structure				
MWChem	Computational chemistry package de for HPC clusters				
Q-CHEM	Computational chemistry package de for HPC clusters				
TeraChem	Quantum chemistry software designe to run on NVIDIA GPU				
Materials Science					
LSMS	Materials code for investigating the ef of temperature on magnetism				
QMCPACK	Solves the many-body Schrodinger e for electronic structures using a quan Monte Carlo method				
Quantum-Espresso/PWscf	An integrated suite of computer code electronic structure calculations and modeling at the nanoscale				
VASP	First principles materials code that c electronic structures and quantum-mechanical molecular dynamics				
Visualization & Docking Software					
Amira 5	A multifaceted software platform for visualizing, manipulating, and unders life sciences and bio-medical data				
Core Hopping	Rapid screening of novel cores to imp drug properties				
FastROCS	3D molecular shape comparison				
VMD	Visualizing and analyzing large bio-m systems in 3-D graphics				

POPULAR GPU-ACCELERATED APPLICATIONS, Continued

Application	Description	Supported Features	Expected	Multi-GPU	Release Status
Weather & Climate Forecasting					
ASUCA	Weather forecasting model fully optimized for GPUs				
CAM / SE	Community Atmospheric Model is a g atmosphere model for weather and c research				
GEOS-5	Weather modeling and forecasting ap used by NASA				
HIRLAM	Weather forecasting model fully optim for GPUs				
HOMME	Weather modeling tool for atmospher scientists				
HYCOM	Weather forecasting model using ico horizontal grid				
MITgcm	Numerical model designed for study atmosphere, ocean, and climate				
NIM	Weather forecasting model using ico horizontal grid				
WRF	Weather and Ocean modeling applica				
Editing and Effects					
Adobe Premiere Pro	Video editing				
Avid Media Composer	Video editing				
GenArts Sapphire	Effects plug-in for video editing				
Sony Vegas Pro	Video editing				
Animation					
Autodesk 3Ds Max	3D modeling, animation, and rendering				
Autodesk Maya	3D modeling, animation, and rendering				
Defense & Intelligence					
DigitalGlobe Advanced Ortho Series	Geospatial Visualization				
Elsevex iRaze Terra	Geospatial Visualization				
Exelis (ITT) ENVI	Geospatial Visualization				
GeoEye Analytics Signature Analyst	Geospatial Visualization				
GeoWebSol Desktop	Geospatial Visualization				
Incogna GIS	Geospatial Visualization				
Intergraph Motion Video Analyst	Video filters and mosaicing — Geo-l FMV analytics with intelligence data				
Intrinsium Panoptes 3.0	Video Analytics				

POPULAR GPU-ACCELERATED APPLICATIONS, Continued

Application	Description	Supported Features	Expected	Multi-GPU	Release
Electronic Design Automation and CEM					
Agilent Technologies ADS	Simulation tool for design of RF, micr and high speed digital circuits				
Agilent Technologies EMPro	Modeling and simulation environmen analyzing 3D EM effects of high speed RF/Microwave components				
ANSYS Nexxim	Circuit simulation engine for RF/anal mixed-signal IC design; IBIS-AMI ana speedup with GPU computing				
CST Microwave Studio (MWS)	High frequency electromagnetic field simulation				
Gauda OPC, OPV	Collection of several software tools fo computational lithography running on Gauda hardware platform				
Remcom XF-dtd	3D EM modeling and simulation				
Rocketick RocketSim	Verilog simulation				
SPEAG SEMCAD-X	3D EM modeling and simulation				
CAD					
CATIA V6 – Live Rendering	Photorealistic rendering				
Bunkspeed Pro Suite	Easy to use photorealistic rendering s				
RTT DeltaGen 10.x	Photorealistic rendering used for des				
RTT DeltaPix	Photorealistic rendering with integrat TeamCenter and RTT formats				
Numerical Analysis					
Jacket AccelerEyes	GPU acceleration for MATLAB				
Mathematica Wolfram	Symbolic math analysis				
MATLAB Mathworks	Technical computing language and integrated development environment (MATLAB PCT, MDCS)				

POPULAR GPU-ACCELERATED APPLICATIONS, Continued

Application	Description	Supported Features	Expected	Multi-GPU	Release
Oil&Gas					
Acceleware RTM	Seismic Processing				
CGGVeritas RTM	Seismic Processing				
ffa SWI Pro	Seismic Interpretation				
Headwave Suite	Seismic Imaging				
Geotrix	Seismic Processing/Interpretation				
Paradigm EarthStudy360	Reservoir Modeling				
Paradigm Echos RTM	Seismic Processing				
Paradigm SKUA	Reservoir Modeling				
Paradigm VoxelGeo	Seismic Interpretation				
Schlumberger WesterGeco Omega2 RTM	Seismic Processing				
Seismic City Prestack Interpretation	Seismic Processing				
SpectraSeis	Seismic Processing / Imaging				
Stoneridge Reservoir Simulation	Reservoir Simulation				
Tsunami RTM	Seismic Processing				
Computational Finance					
Hanweck Associates	Real-time options analytical engine f				
MATLAB Mathworks	Data parallel mathematics (MATLAB PCT, MDCS)				
Murex	Risk analytics (MACS)				
Numerical Algorithms Group	Random Number Generators				
SciComp, Inc	Derivative pricing (SciFinance)				
Wolfram Mathematica	Mathematical Development Environ				

*GPU performance compared against multi-core x86 CPU socket. kernel to kernel performance comparison. Performance results

POPULAR GPU-ACCELERATED APPLICATIONS, Continued

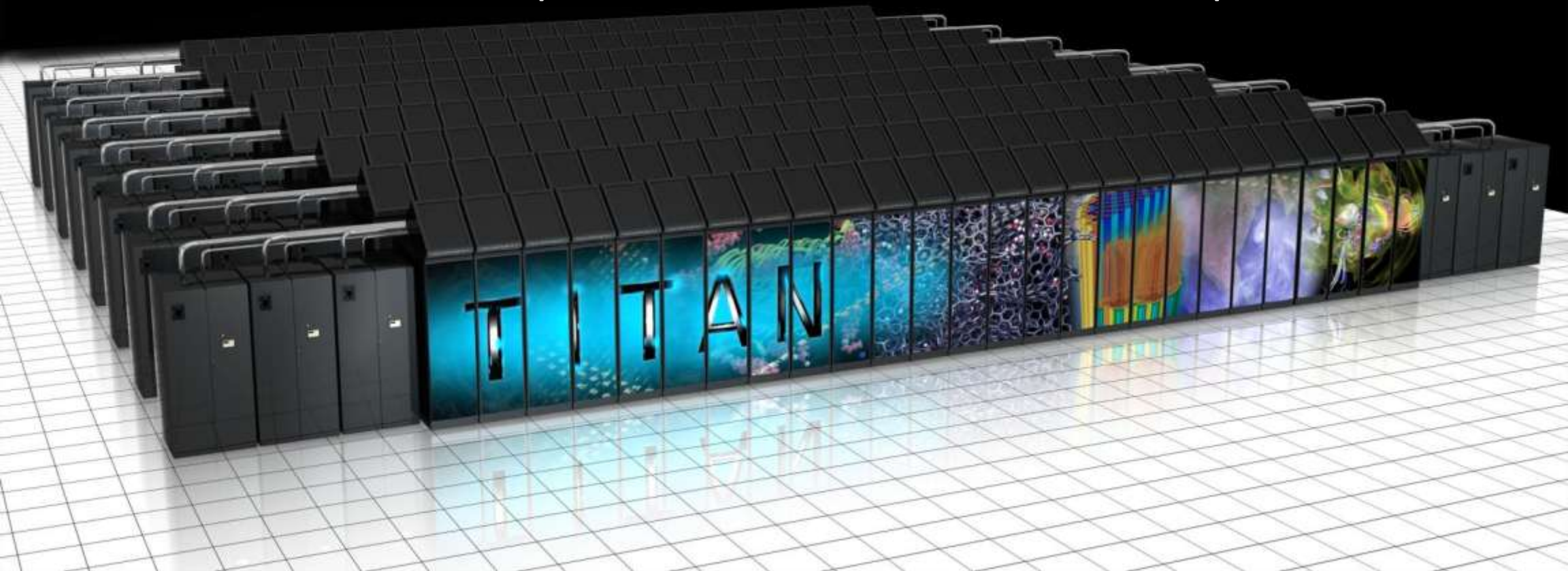
Application	Description	Supported Features	Expected Speed Up*	Multi-GPU Support	Release Status
Physics					
Chroma	General purpose LQCD application	Wilson-clover fermions, Krylov solvers, Domain-decomposition	5-8x	Yes	Available now
MILC	General purpose LQCD application	Staggered fermions, Krylov solvers, Gauge-link fattening	5-8x	Yes	Available now
Computational Fluid Dynamics					
Altair AcuSolve	General purpose CFD flow solver	Linear equation solver	2x	Yes	Available now
Autodesk Moldflow	Optimize design of plastic parts and injection molds	Linear equation solver	2x	Single Only	Available now
FEFLO (GMU-Lutoser)	Navier-Stokes flow solver based on unstructured grids for modeling both compressible and incompressible flows	Explicit solver	10x	Yes	In Development
FluiDyna LBultra	Computing physical flows in and around solid bodies	LBM, particle CFD	20x	Yes	Available now
FluiDyna Culises-OpenFOAM	Computing physical flows with Culises — a software library with special algorithms for solving systems of equations	Linear equation solvers	3x Solver	Single Only	Available now
Prometech Particleworks	Fluid simulation for free surface flow like Tsunami, material processing and liquids	MPS, Particle CFD	4x-9x	Yes	Available now
SGI (Sandia NL-SIO)	Massively parallel direct numerical solver (DNS) for the full compressible Navier-Stokes	Chemistry kernel	3x SP, 5x DP kernel	Yes	In Development
Turbostream	Ultrafast CFD solver for turbomachines	Explicit solver	19x	Yes	Available now
Vratis SpeedIT-OpenFOAM Solver	Set of accelerated solvers for sparse linear systems of equations	Linear equation solvers	3x Solver	Yes	Available now
Computational Structural Mechanics					
Abaqus/Standard	Simulation and analysis tool for structural mechanics	Linear equation solver	1.5-2.5x	Single Only	Available now
ANSYS Mechanical	Simulation and analysis tool for structural mechanics	Linear equation solver	2x	Single Only	Available now
Impetus Afea	Predicts large deformations of structures and components exposed to extreme loading conditions	Linear equation solver, SPH	10x SPH, 2x Total	Yes	Available now
LS-DYNA Implicit	Multiphysics simulation package used	Linear equation solver	3x	Yes	In Development
MSC Nastran	Simulation and analysis tool for structural mechanics	Linear equation solver	1.4-2x	Yes	Available now
Marc	Simulation and analysis tool for structural mechanics	Linear equation solver	1.5x	Yes	In Development
RADIOSS Implicit	Used to maximize durability, NVH, crash, safety, manufacturability and fluid-structure interaction performance	Linear equation solver	2x	Single Only	In Development

Titan: World's Fastest Supercomputer

18,688 Tesla K20X Accelerators

27 Petaflops Peak: 90% of Performance from GPUs

17.59 Petaflops Sustained Performance on Linpack



Tesla CUDA Architecture Roadmap

How Will GPUs Evolve Over This Decade?

- Integration (memory, processor types, network)
- Further concentration on locality (both HW and SW)
- Reducing overheads (intra- and inter-GPU)
- Continued convergence with consumer technology

Echelon

NVIDIA's Extreme-Scale Computing Project
DARPA UHPC Program
Targeting 2018

Fast Forward Program

Which Takes More Energy?

Performing a 64-bit floating-point FMA:

893,500.288914668
× 43.90230564772498
= 39,226,722.78026233027699
\+ 2.02789331400154
= 39,226,724.80815564

Or moving the three 64-bit operands 20 mm across the die:

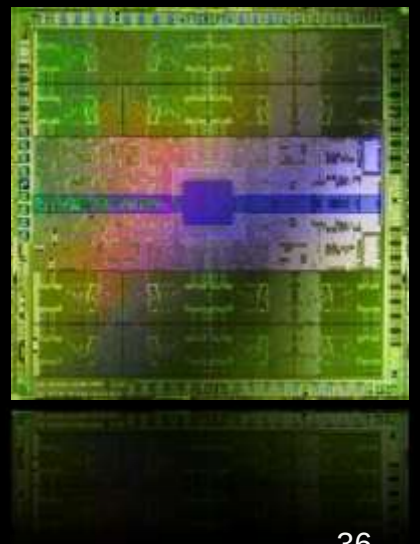
This one takes over 4.7x the energy today (40nm)!

It's getting worse: in10nm, relative cost will be 17x!

Loading the data from off chip takes >> 100x the energy.

Summary

- **Power is the main HPC constraint**
  - Vast majority of work *must* be done by cores designed for efficiency
- **NVIDIA GPU's are already designed for energy efficiency**
- **Data movement dominates the power**
  - **Locality at all levels and reduction of overhead is necessary**
- **GPU computing has a sustainable model**
  - Aligned with technology trends, supported by consumer markets
- **GPUs are the path to the tightly-coupled hybrid processor future**




Thank you.
Questions?
Axel Koehler
Sr. Solution Architect HPC
akoehler@nvidia.com
NVIDIA