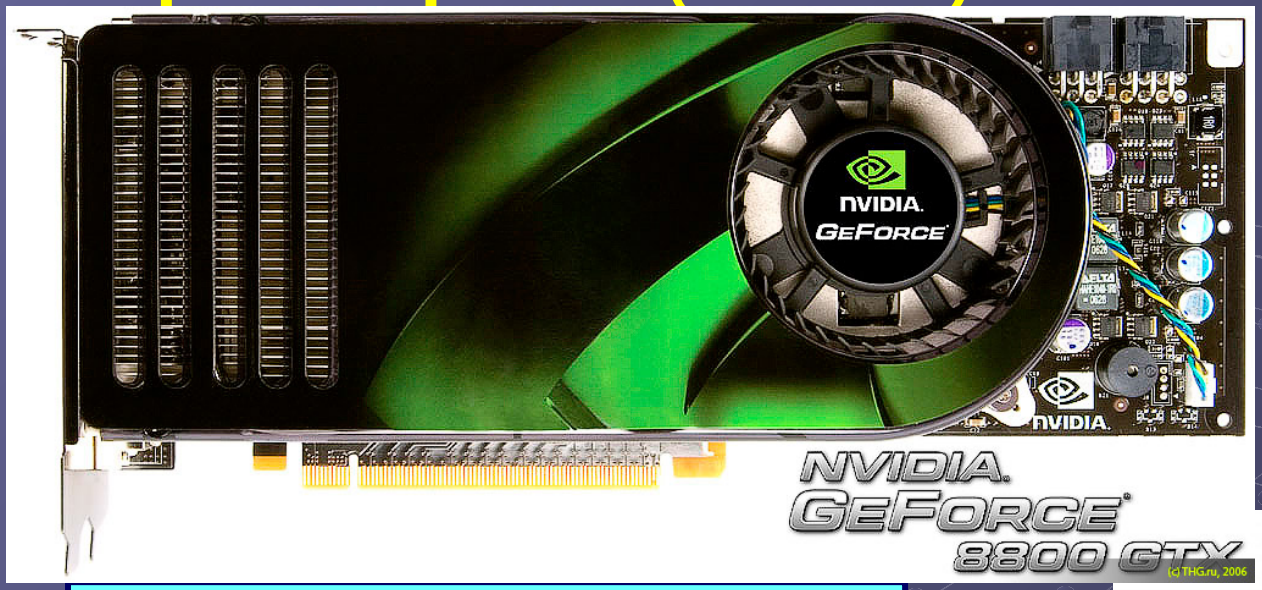
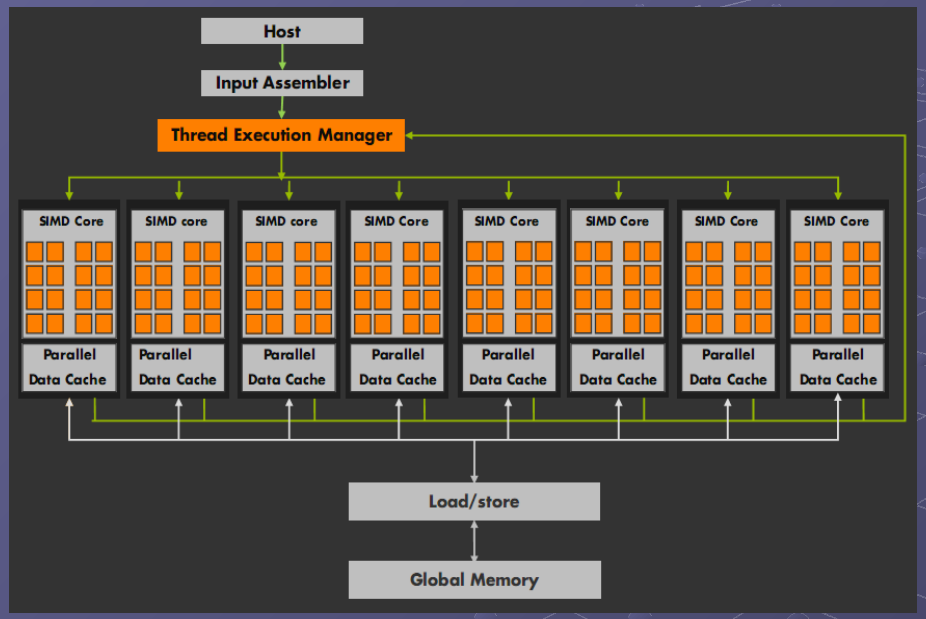
More on GPU

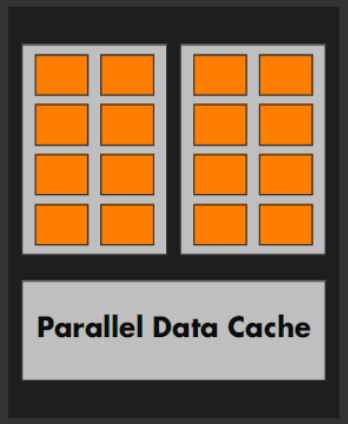
Graphics Processors (GPU) as General Purpose Supercomputers (GPGPU)



2008…
GeForce 9800 GTX, 128 Stream Proc., 512 MB
GeForce 9800 GX2, 256 Stream Proc., 1 GB
GeForce 9800 GT, 64 Stream Proc., 512 MB
[…]
2009: Tesla ~200 Proc., 4GB
2010: Fermi ~400 Proc., 4GB
2013: Kepler K20, ~2500 Procs., 6GB
2016: Kepler K80, ~5000 Procs.
2017/18: Pascal, Volta > 5000 Procs.

Hardware around 2006





Each core

- 8 functional units
- SIMD 16/32 "warp"
- 8-10 stage pipeline
- Thread scheduler
- 128-512 threads/core
- 16 KB shared memory

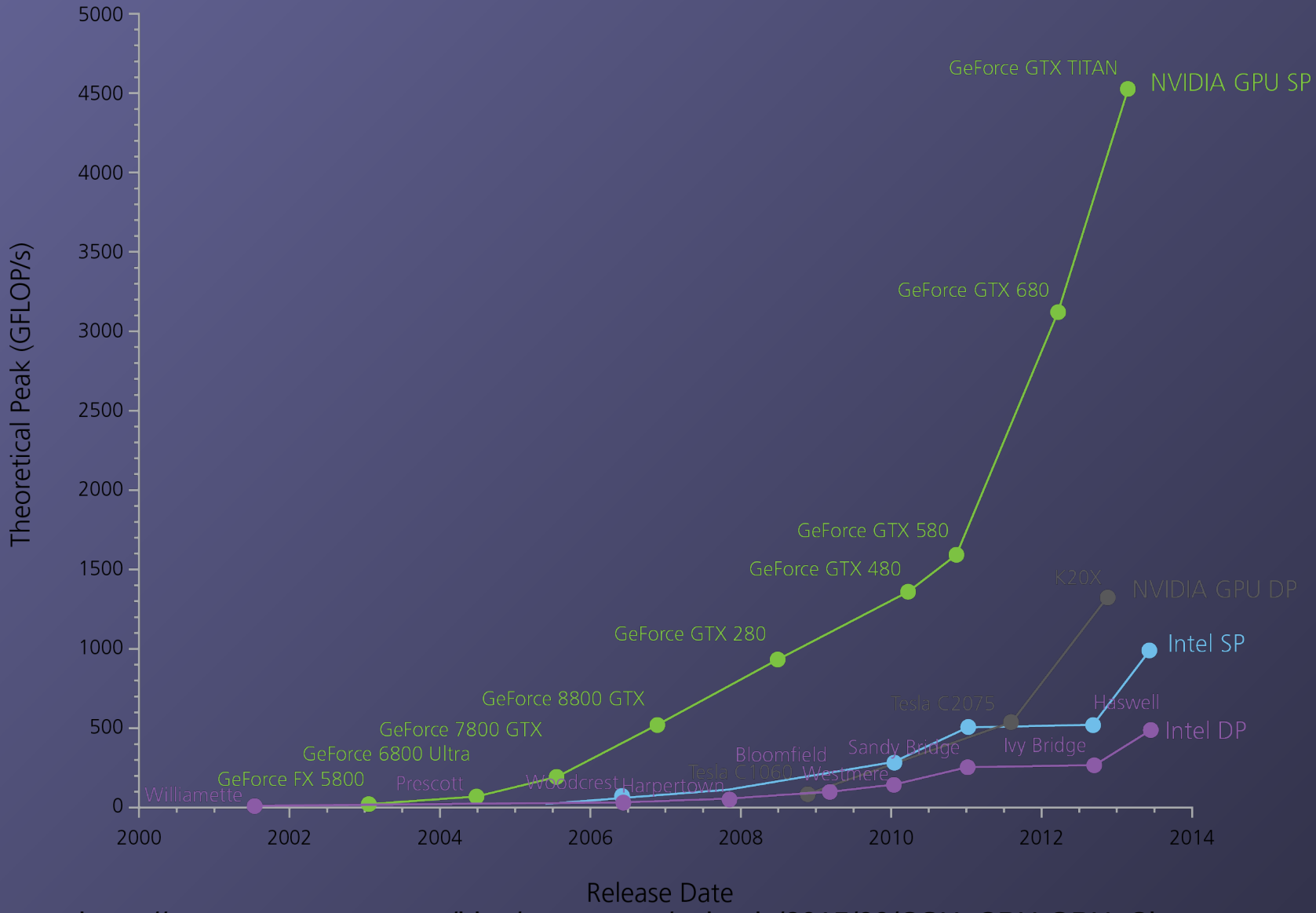
Total #threads/chip

16 * 512 = 8K

GeForce 8800 GTX:

575 MHz * 128 processors * 2 flop/inst * 2 inst/clock = 333 Gflops

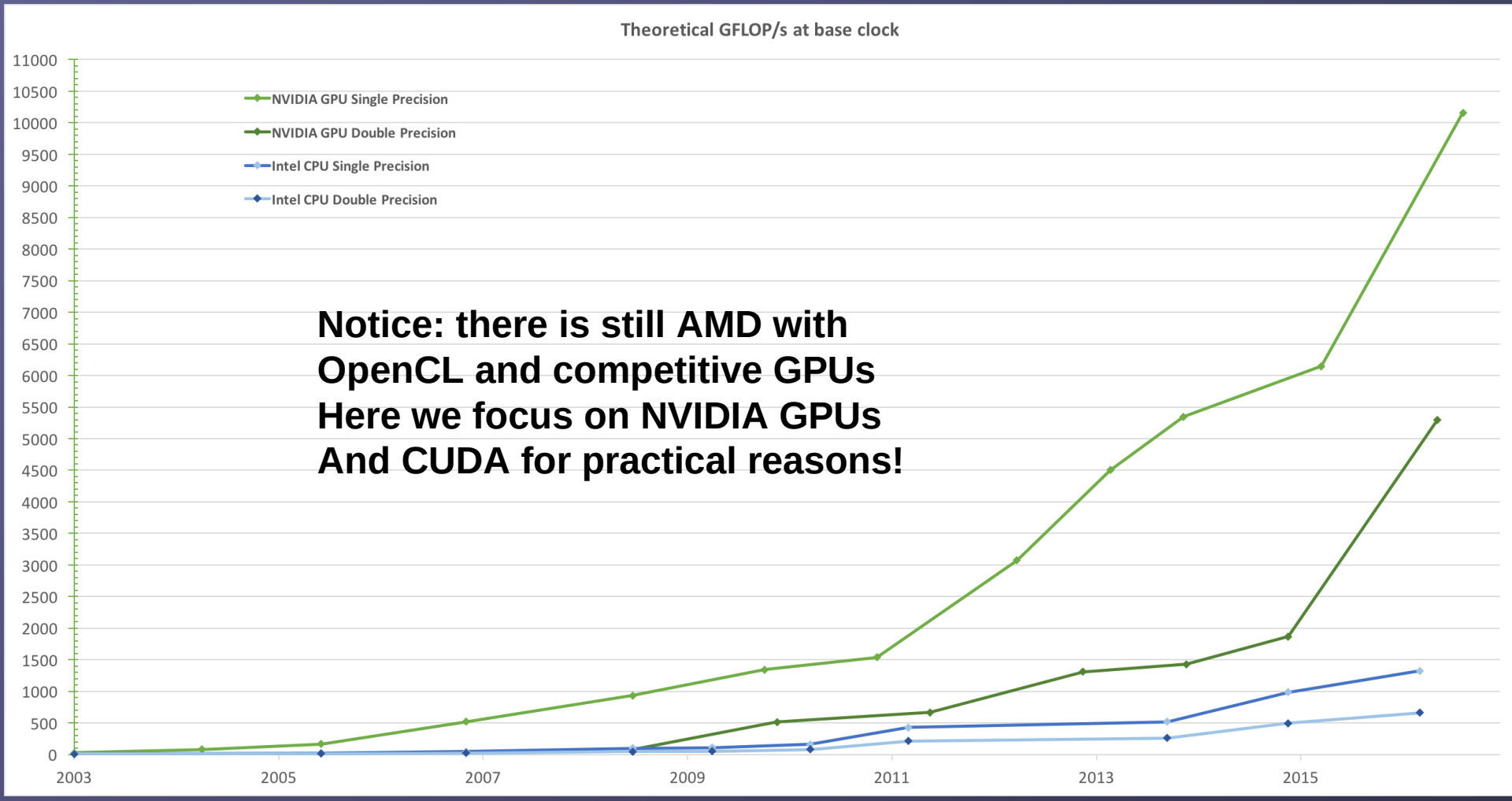
CPU vs. GPU speedup timeline

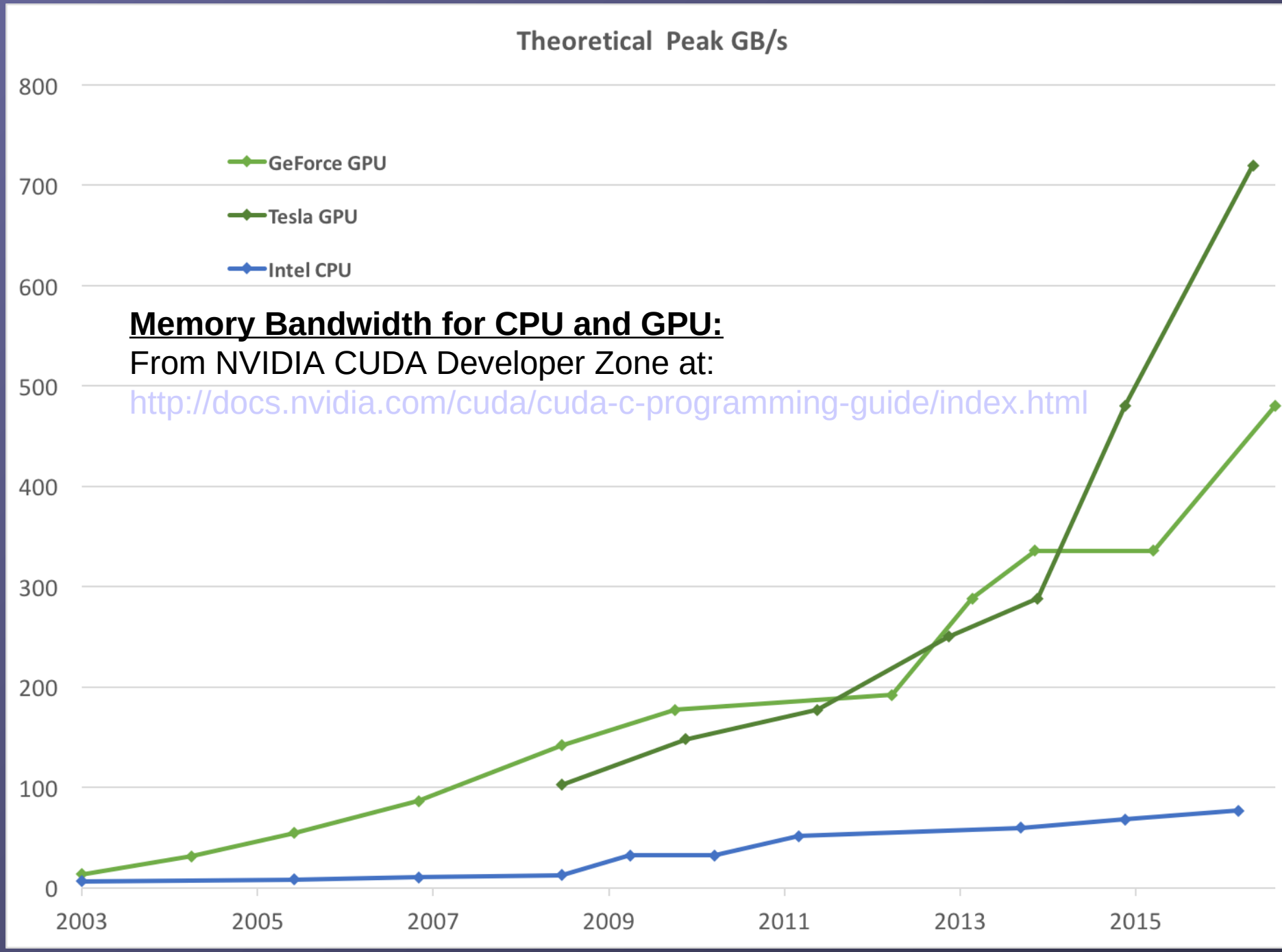


Source: https://www.carestream.com/blog/wp-content/uploads/2015/09/CSH_CPU-GPU_Chart.png

Floating Point Operations per Second for CPU and GPU:

From NVIDIA CUDA Developer Zone at:
http://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html



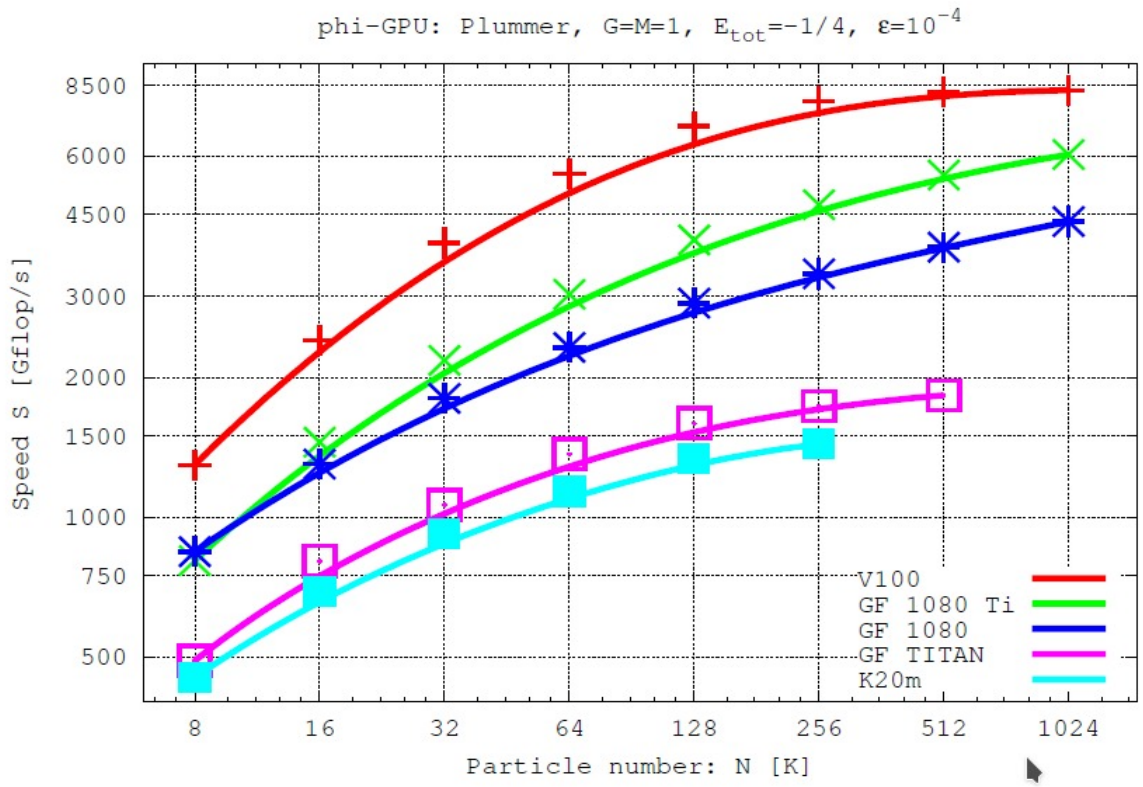


Memory Bandwidth for CPU and GPU:

From NVIDIA CUDA Developer Zone at:

http://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html

Kepler, Pascal, Volta, Scaling, it works...



Volta V100

Pascal GF1080

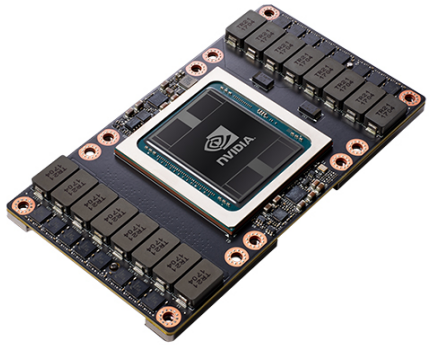
Kepler K20m

Spurzem, Berczik, et al., 2013, LNCS Supercomputing, 2013, pp. 13-25, Springer. (updated unpublished)

Fig. 4. Here we report a preliminary result from a benchmark test of our code on one Kepler K20 card; we compare with the performance on Fermi C2050 (used in the Mole-8.5 cluster), and the oldest Tesla C1060 GPU (used in the laohu cluster of 2009) - the latter is used as a normalization reference. We plot the speed ratio of our usual benchmarking simulation used in the previous figures, as a function of particle number. From this we see the sustained performance of a Kepler K20 would be about 1.4 - 1.5 Tflop/s.

X = first GPU of laohu 2010

NVIDIA Volta V100 GPU, 21 billion transistors, 5120 cores

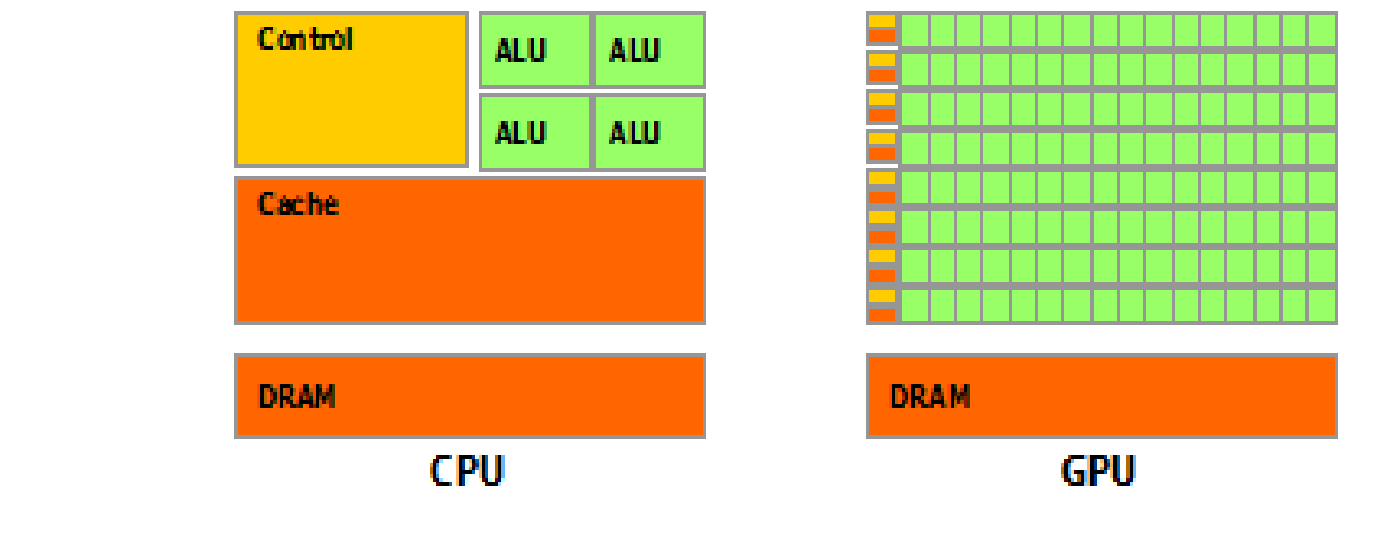


With NVLINK

Without NVLINK

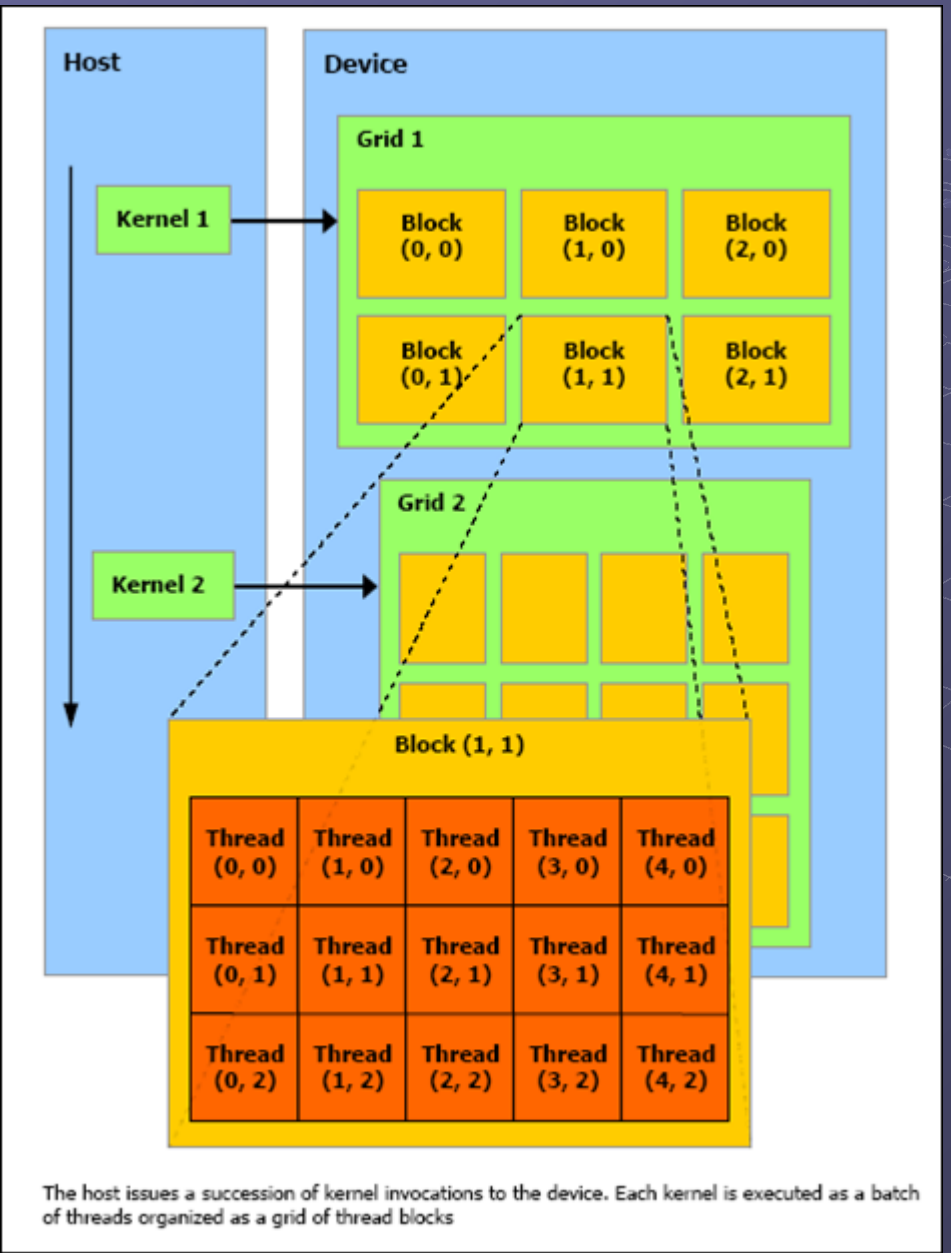
PERFORMANCE with NVIDIA GPU Boost™	DOUBLE-PRECISION 7.8 teraFLOPS	DOUBLE-PRECISION 7 teraFLOPS
	SINGLE-PRECISION 15.7 teraFLOPS	SINGLE-PRECISION 14 teraFLOPS
	DEEP LEARNING 125 teraFLOPS	DEEP LEARNING 112 teraFLOPS
INTERCONNECT BANDWIDTH Bi-Directional	NVLINK 300 GB/s	PCIE 32 GB/s
MEMORY CoWoS Stacked HBM2	CAPACITY 32/16 GB HBM2	
	BANDWIDTH 900 GB/s	
POWER Max Consumption	300 WATTS	250 WATTS

CPU and GPU; from CUDA NVIDIA Developer Zone at
http://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html

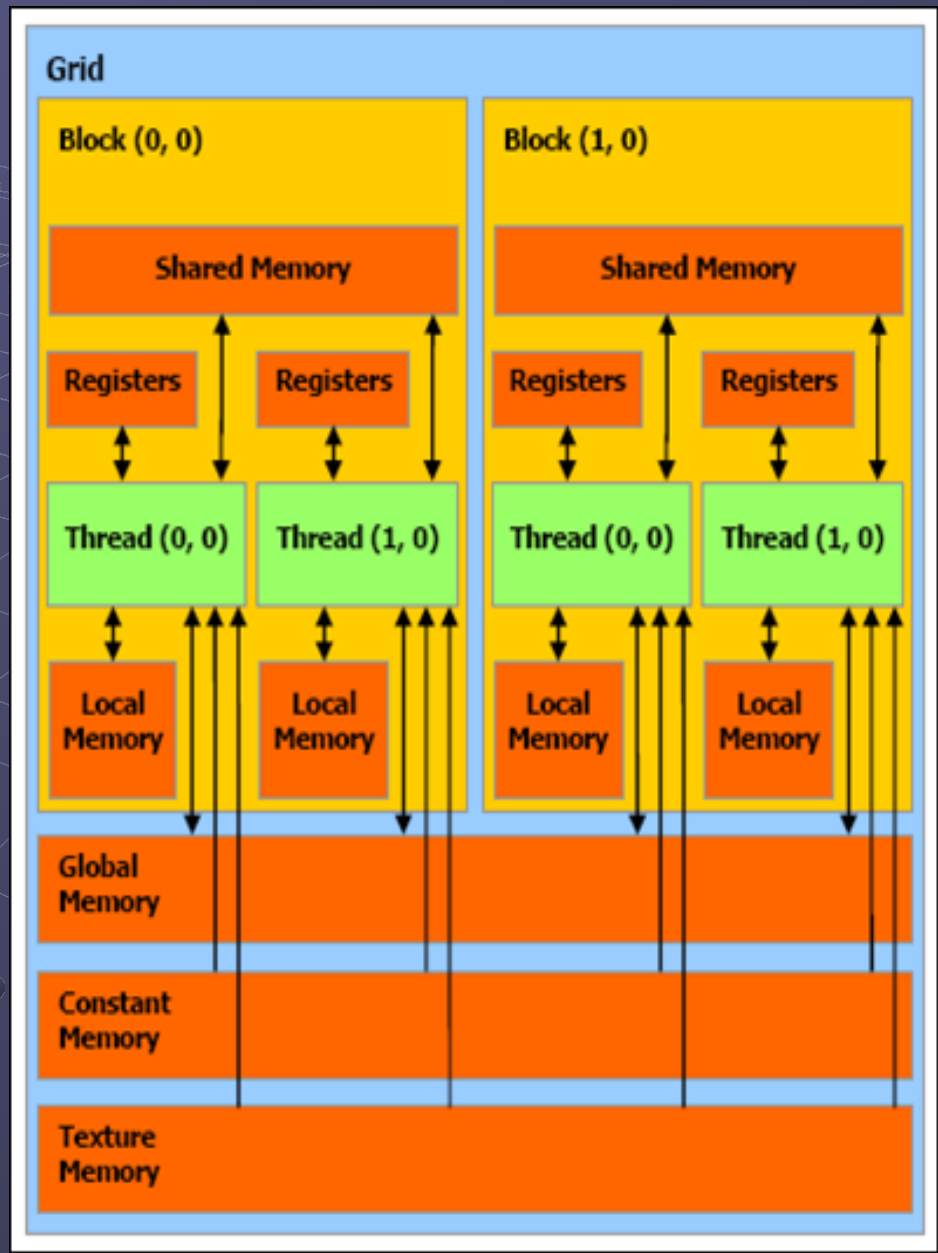


“The GPU devotes more transistors to computing”
“favours data parallel operations”

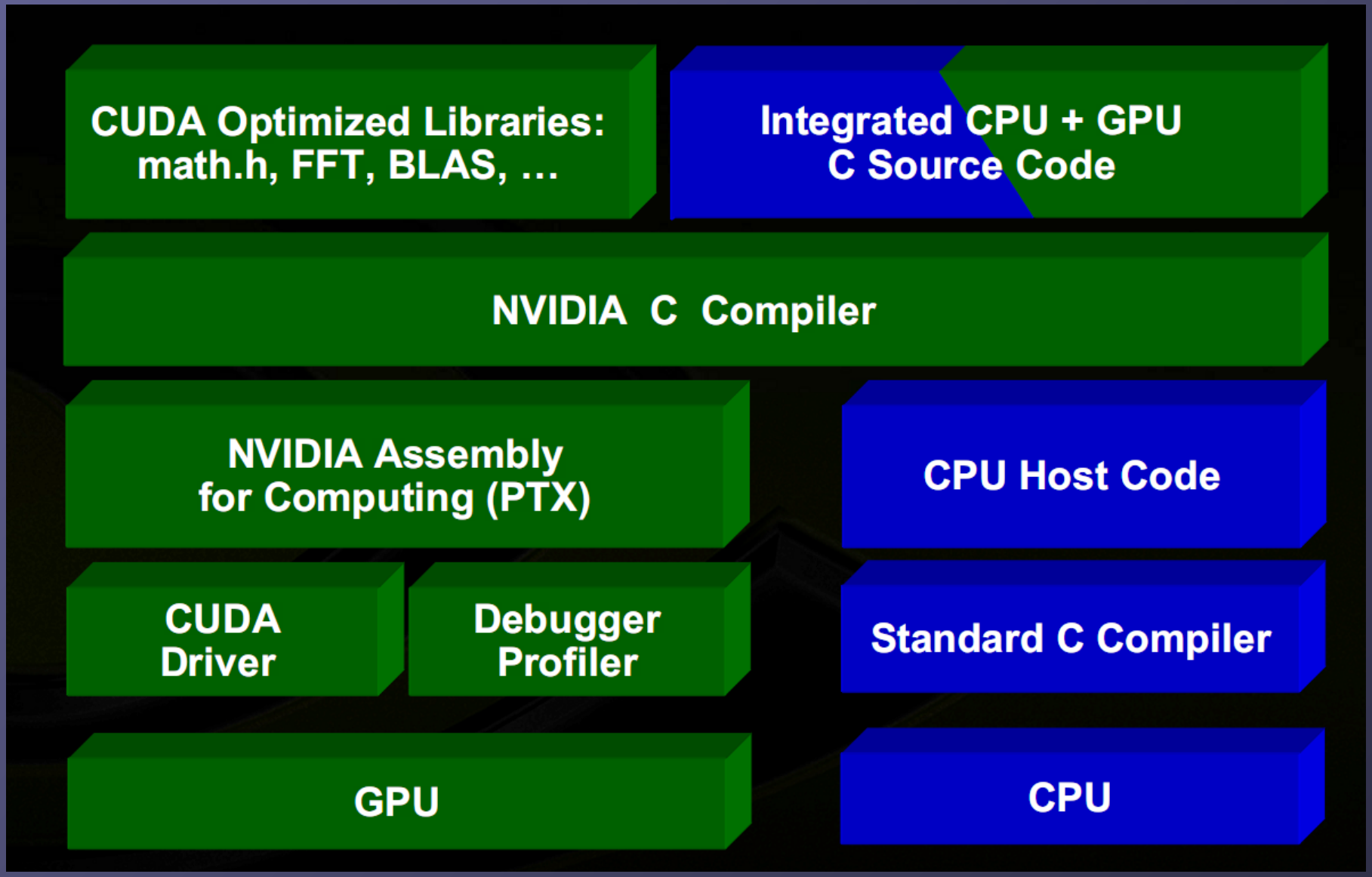
GPU Structure From: http://geco.mines.edu/tesla/cuda_tutorial_mio/



The host issues a succession of kernel invocations to the device. Each kernel is executed as a batch of threads organized as a grid of thread blocks



CUDA


CUDA Optimized Libraries:
math.h, FFT, BLAS, …
Integrated CPU + GPU
C Source Code
NVIDIA C Compiler
NVIDIA Assembly
for Computing (PTX)
CPU Host Code
CUDA
Driver
Debugger
Profiler
Standard C Compiler
GPU
CPU

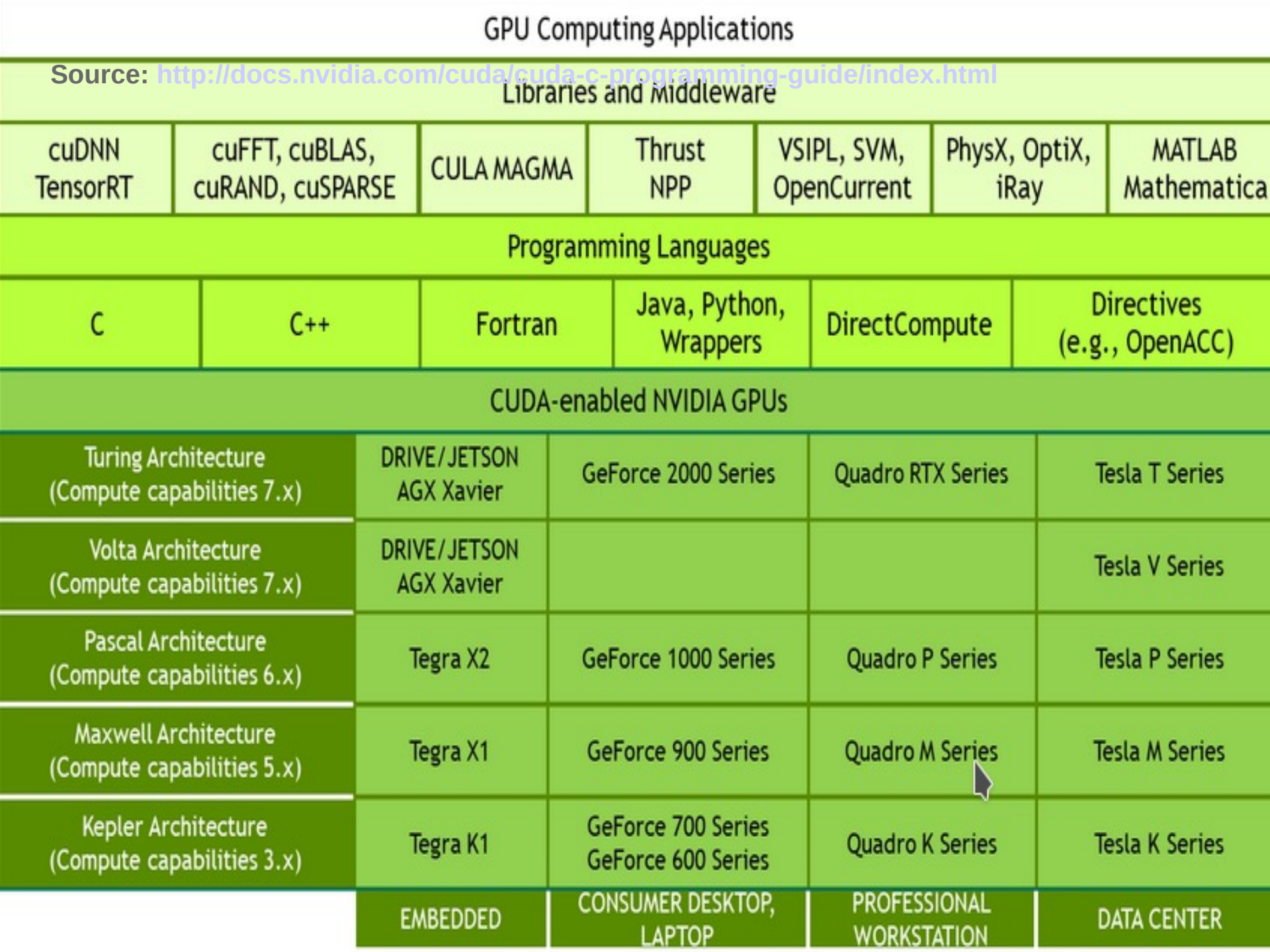
Simple CUDA example

CPU C program

```
void addMatrix(float *a, float *b,
               float *c, int N)
{
  int i, j, index;
  for (i = 0; i < N; i++) {
    for (j = 0; j < N; j++) {
      index = i + j * N;
      c[index]=a[index] + b[index];
    }
  }
}

void main()
{
  .....
  addMatrix(a, b, c, N);
}
```

CUDA C program

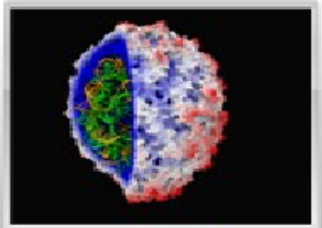
```
__global__ void addMatrix(float *a,float *b,
                          float *c, int N)
{
  int i=blockIdx.x*blockDim.x+threadIdx.x;
  int j=blockIdx.y*blockDim.y+threadIdx.y;
  int index = i + j * N;
  if ( i < N && j < N)
    c[index]= a[index] + b[index];
}


void main()
{
  ..... // allocate & transfer data to GPU
  dim3 dimBlk (blocksize, blocksize);
  dim3 dimGrd (N/dimBlk.x, N/dimBlk.y);
  addMatrix<<<dimGrd,dimBlk>>>(a, b, c,N);
}
```


GPU Computing Applications
Source: http://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html
Libraries and Middleware
cuDNN TensorRT
cuFFT, cuBLAS, cuRAND, cuSPARSE
CULA MAGMA
Thrust NPP
VSIPL, SVM, OpenCurrent
PhysX, OptiX, iRay
MATLAB Mathematica
Programming Languages
C
C++
Fortran
Java, Python, Wrappers
DirectCompute
Directives (e.g., OpenACC)
CUDA-enabled NVIDIA GPUs
Turing Architecture (Compute capabilities 7.x)
DRIVE/JETSON AGX Xavier
GeForce 2000 Series
Quadro RTX Series
Tesla T Series
Volta Architecture (Compute capabilities 7.x)
DRIVE/JETSON AGX Xavier
Tesla V Series
Pascal Architecture (Compute capabilities 6.x)
Tegra X2
GeForce 1000 Series
Quadro P Series
Tesla P Series
Maxwell Architecture (Compute capabilities 5.x)
Tegra X1
GeForce 900 Series
Quadro M Series
Tesla M Series
Kepler Architecture (Compute capabilities 3.x)
Tegra K1
GeForce 700 Series
GeForce 600 Series
Quadro K Series
Tesla K Series
EMBEDDED
CONSUMER DESKTOP, LAPTOP
PROFESSIONAL WORKSTATION
DATA CENTER

Speedups using GPU vs. CPU

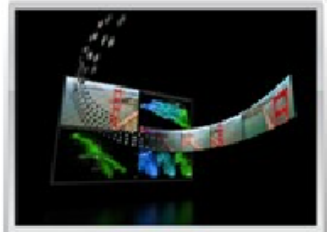
146X

Interactive visualization of volumetric white matter connectivity[1]



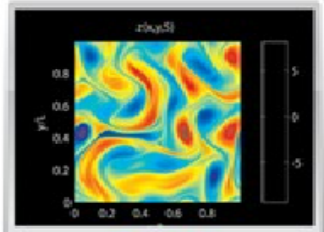
36X

Ionic placement for molecular dynamics simulation on GPU[2]



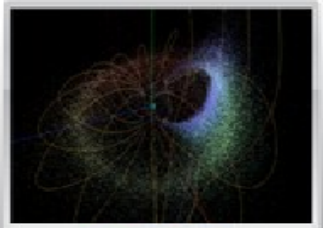
18X

Transcoding HD video stream to H.264 for portable video[3]



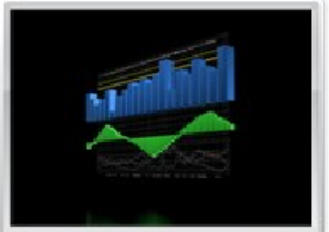
17X

Simulation in Matlab using .mex file CUDA function[4]



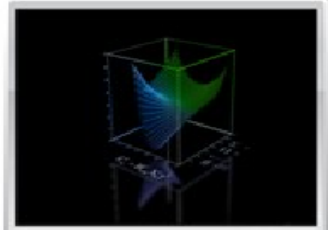
100X

Astrophysics N-body simulation[5]



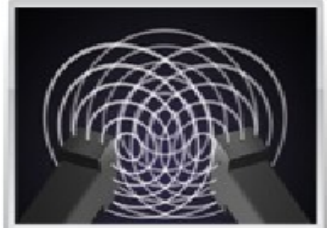
149X

Financial simulation of LIBOR model with swaptions[6]



47X

GLAME@lab: M-script API for linear Algebra operations on GPU[7]

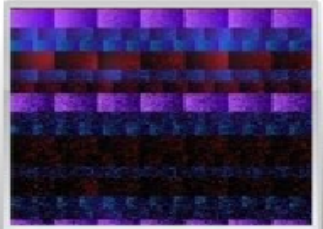


20X

Ultrasound medical imaging for cancer diagnostics[8]

24X

Highly optimized object oriented molecular dynamics[9]



30X

Cmatch exact string matching - find similar proteins & gene sequences[10]

Towards Peta-Scale Green Computation

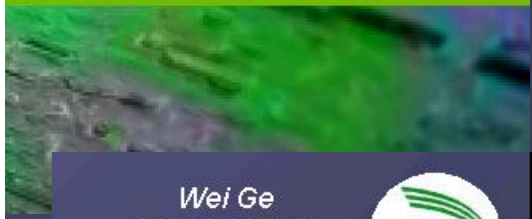
— applications of the GPU supercomputers in CAS

http://www.nvidia.com/gtc2010-content

GPU TECHNOLOGY CONFERENCE

GTC 2010 | Sept 20-23, 2010
San Jose Convention Center, San Jose, California
Watch the Keynote Recordings

- Algorithms & Numerical Techniques
- Astronomy & Astrophysics
- Audio Processing
- Cloud Computing
- Computational Fluid Dynamics
- Computer Graphics
- Computer Vision
- Databases & Data Mining
- Digital Content Creation
- Embedded & Automotive
- Energy Exploration
- Film
- Finance
- General Interest
- GPU Accelerated Internet
- High Performance Computing
- Imaging
- Life Sciences
- Machine Learning & Artificial Intelligence
- Medical Imaging & Visualization
- Mobile & Tablet & Phone
- Molecular Dynamics
- Neuroscience
- Physics Simulation
- Programming Languages & Techniques
- Quantum Chemistry
- Ray Tracing
- Signal Processing
- Stereoscopic 3D
- Tools & Libraries
- Video Processing



Wei Ge
Xiaowei Wang
Inst. of Proc. Eng.

Yunquan Zhang
Inst. of Software

Rainer Spurzem
Nat. Astro. Obs. Chn.

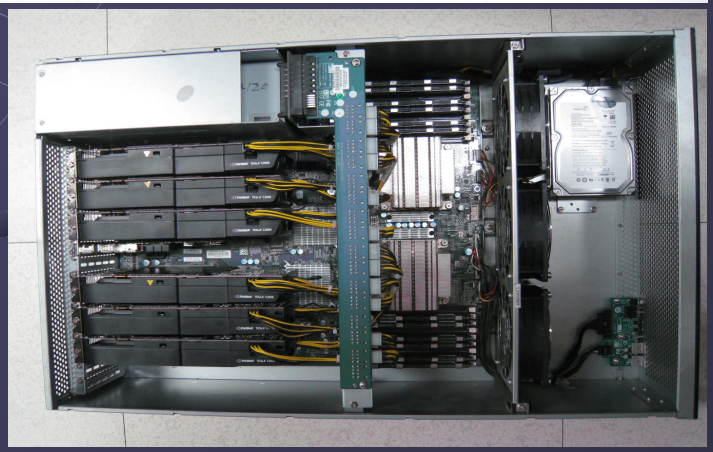
Long Wang
SC Center

Computer Physics - Astrophysics

Molecular Dynamics

Fermi-based GPU supercomputer IPE (2010.04.24)

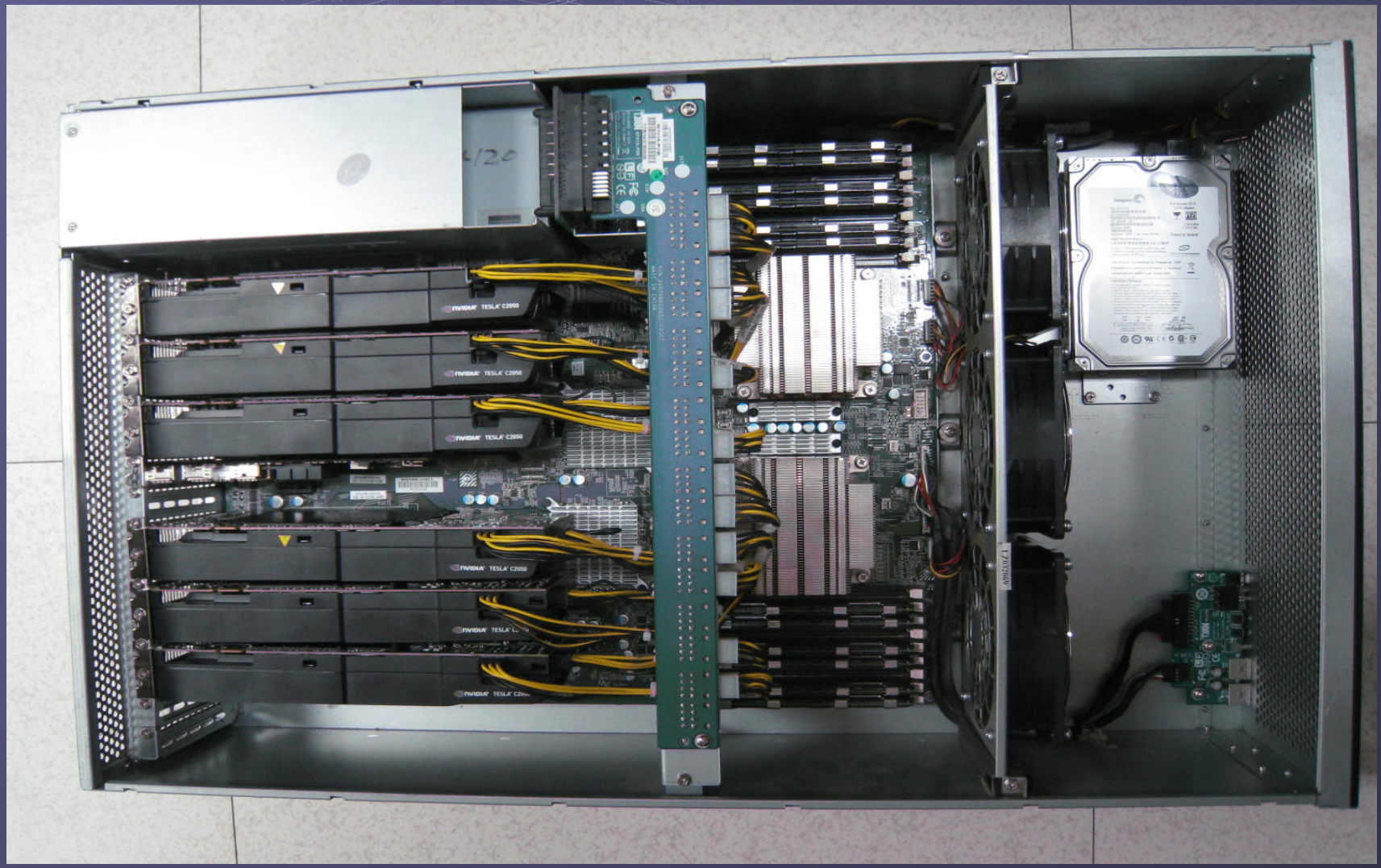
Rpeak SP：	2Pflops
Rpeak DP：	1Pflops
Linpack:	207.3T (Top500 19th)
Mflops/Watt:	431 (Green500 8th)
Total RAM：	17.2TB
Total VRAM：	6.6TB
Total HD：	360TB
Inst. Comm.：	H3C GE
Data Comm.：	Mellanox QDR IB
Occupied area：	150 sq.m.
Weight：	12.6 tons
Max Power：	600kW(computing) 200kW(cooling)
System：	CentOS 5.4, PBS
Monitor：	Ganglia, GPU monitor
Languages：	C, C++, CUDA 3.1 , OpenCL



中国科学院过程工程研究所
Institute Of Process Engineering, Chinese Academy Of Sciences

IPE CAS 372 node 6xC2050 cluster

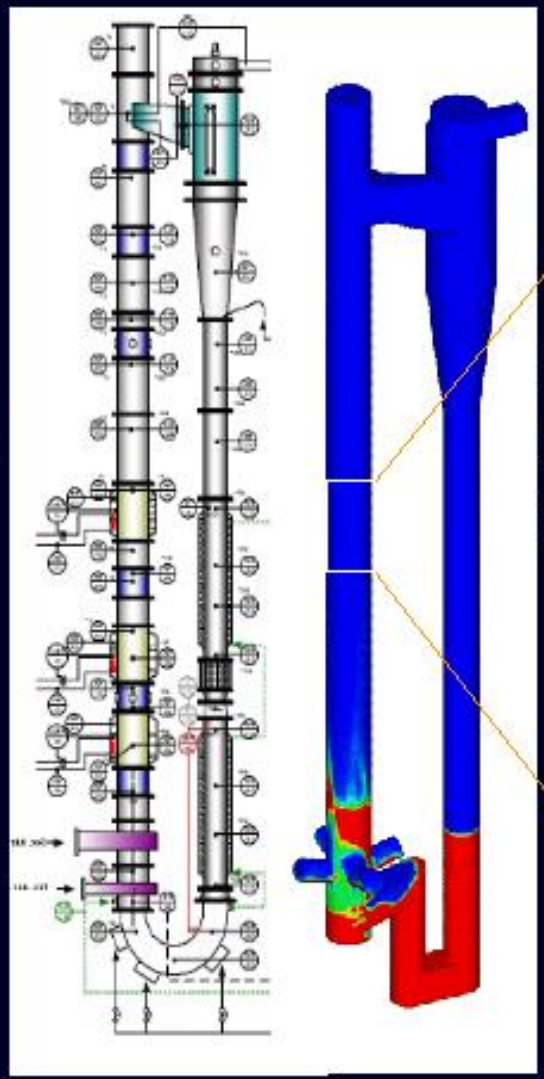
2232 GPU = 2.2 Pflops SP / 1.1 Pflops DP



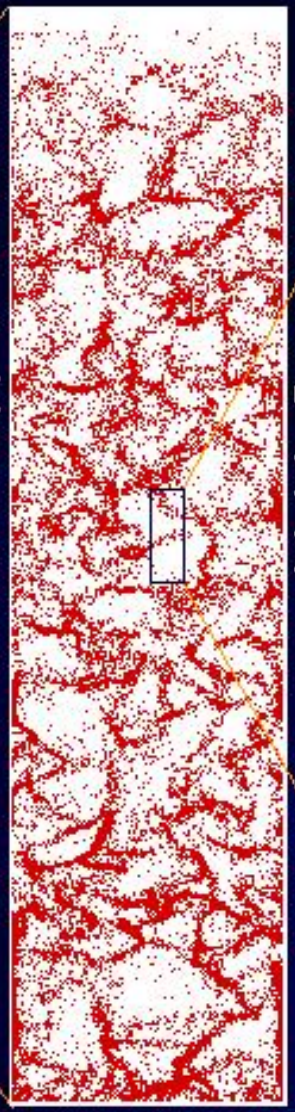
DNS of gas-solid flow : >20x speedup (1C1060/1E5430 core)

120K Particles + 400M pseudo-particles

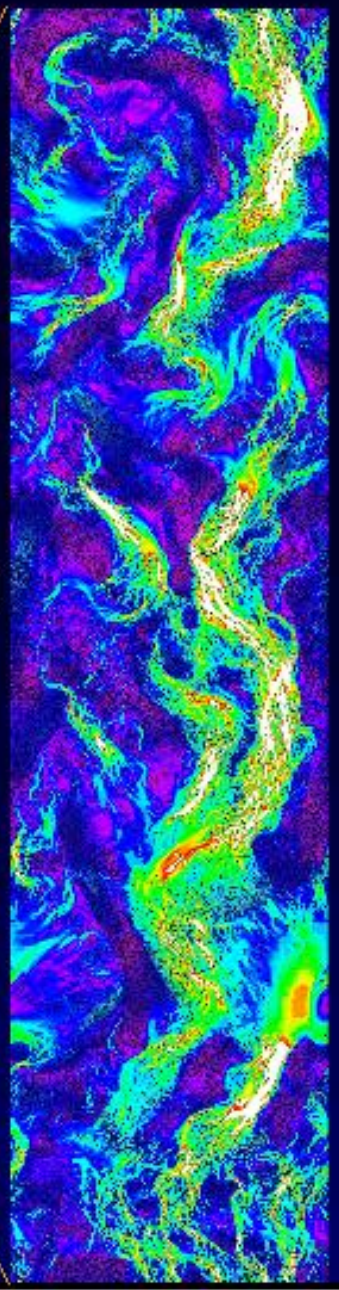
Reactor:
0.4*20m
3D



Section:
0.4*1m
2D



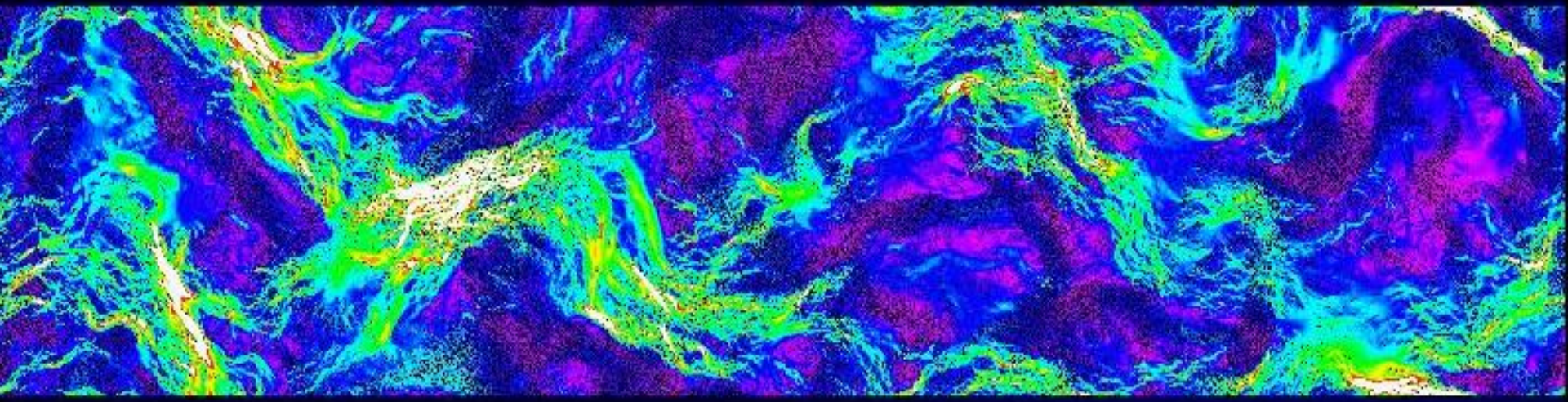
Cell:
2*10cm
2D



Animation Challenge:

9600x2400 → 1200x300 pixels

1000 → 17 frames



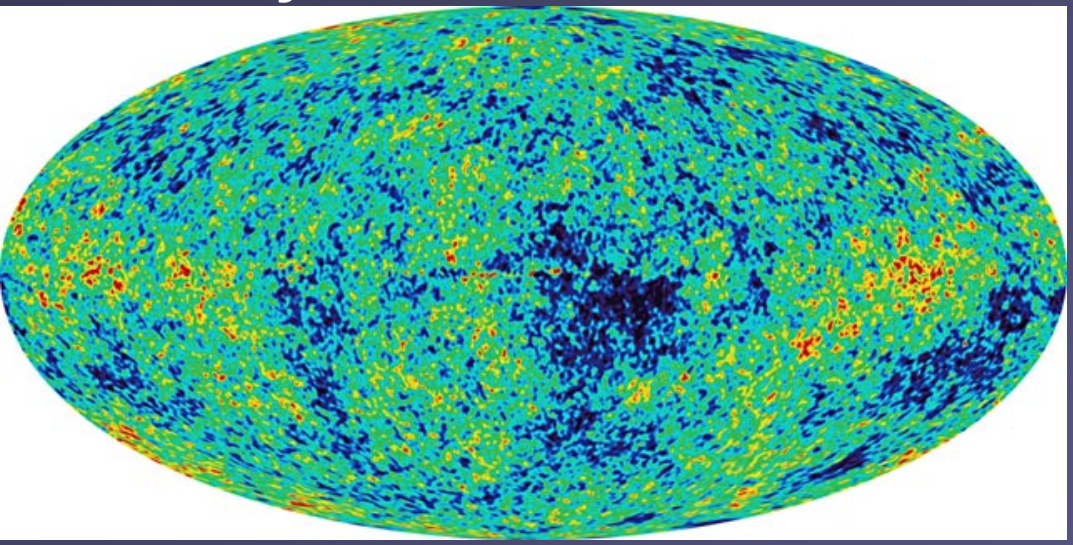
Computer Physics - Astrophysics

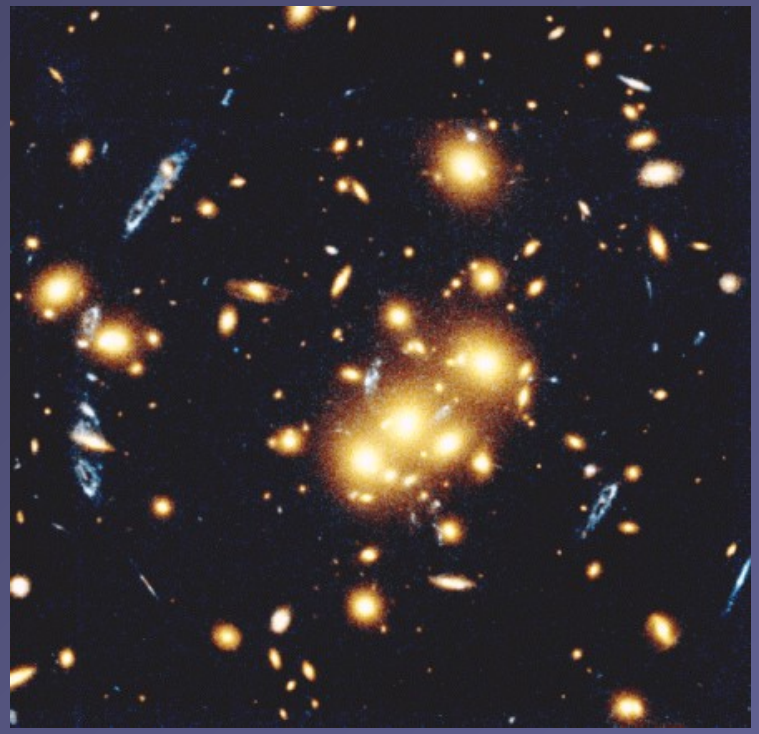
Cosmology

Computer Physics – Astrophysics

- Structure Formation in the Universe

In the year 100.000....

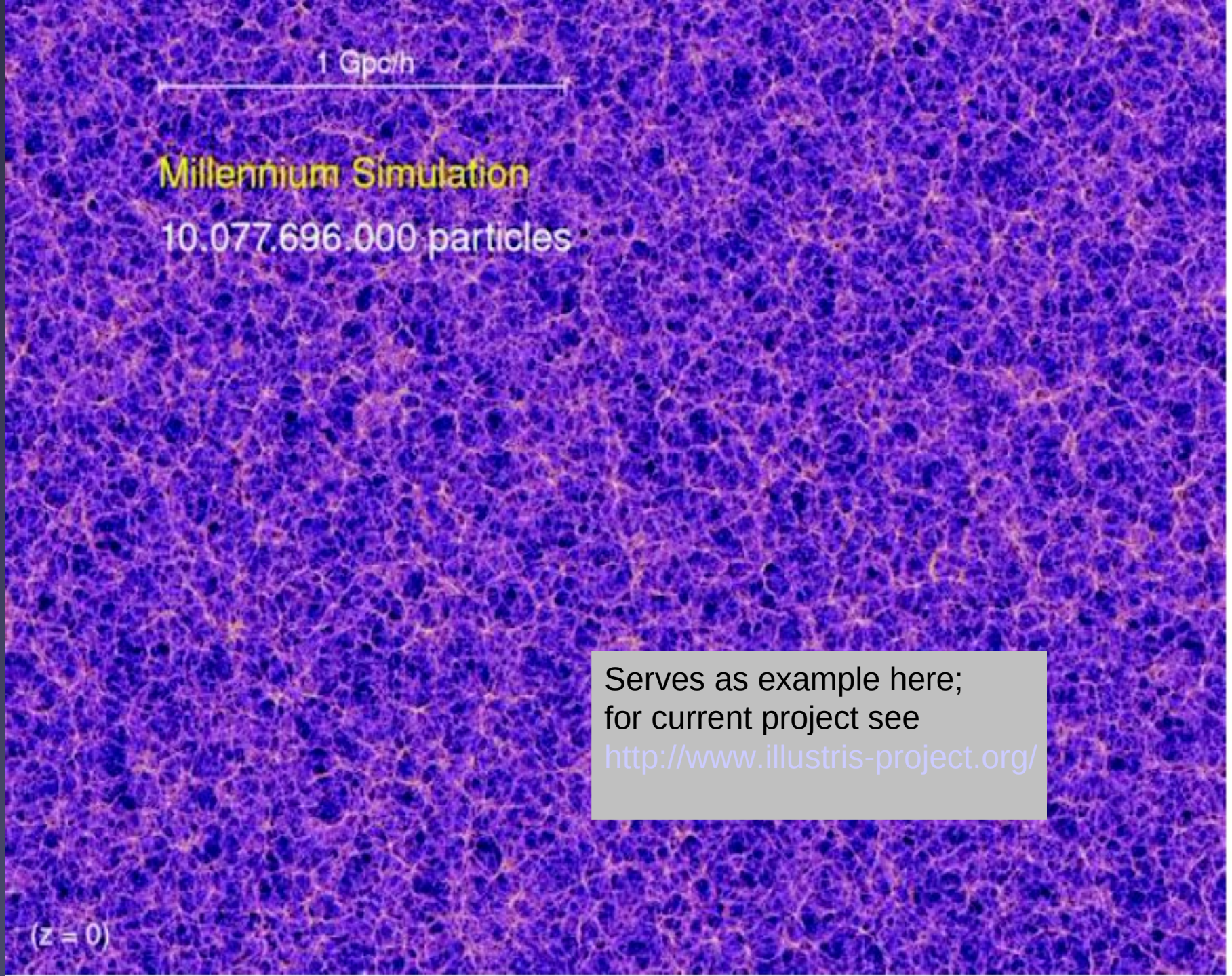




- Wilkinson Microwave Anisotropy Probe (WMAP)

(Cosmic Microwave Background)

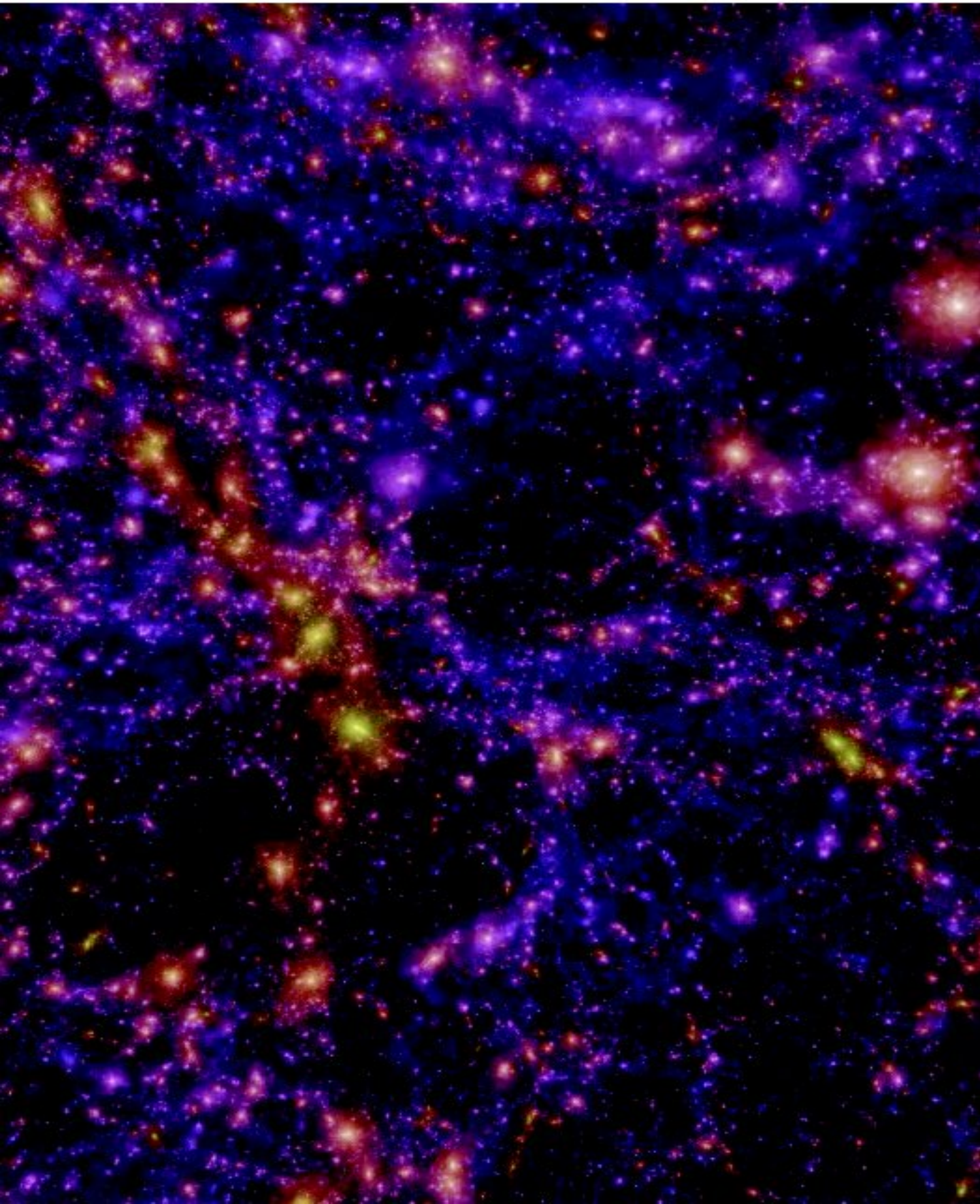
...and ``today´´

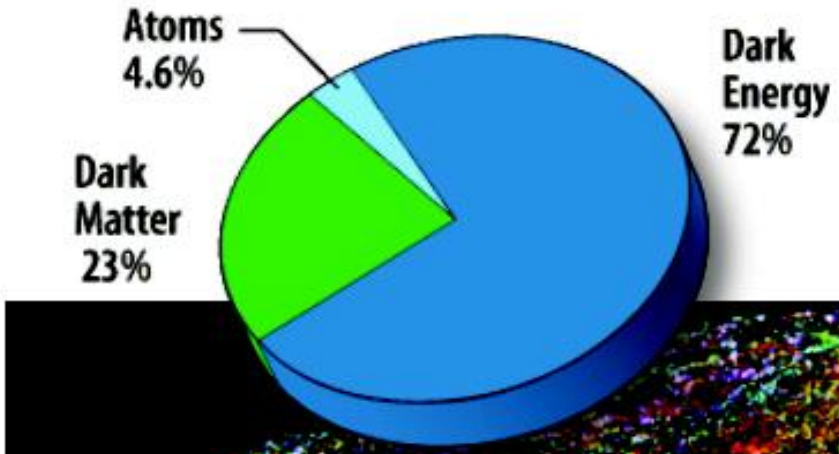


Serves as example here;
for current project see
http://www.illustris-project.org/

Millenium Simulaiton (Springel et al.)

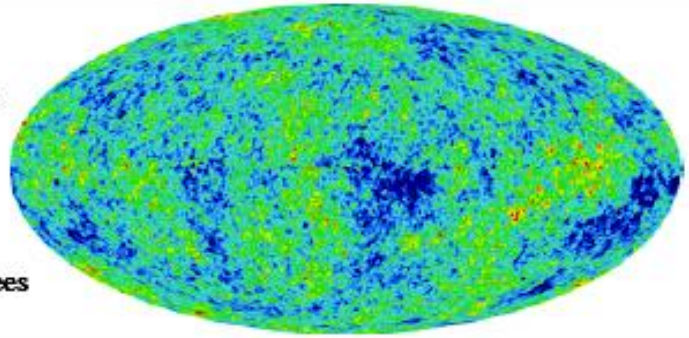
Millenium Simulaiton (Springel et al.)



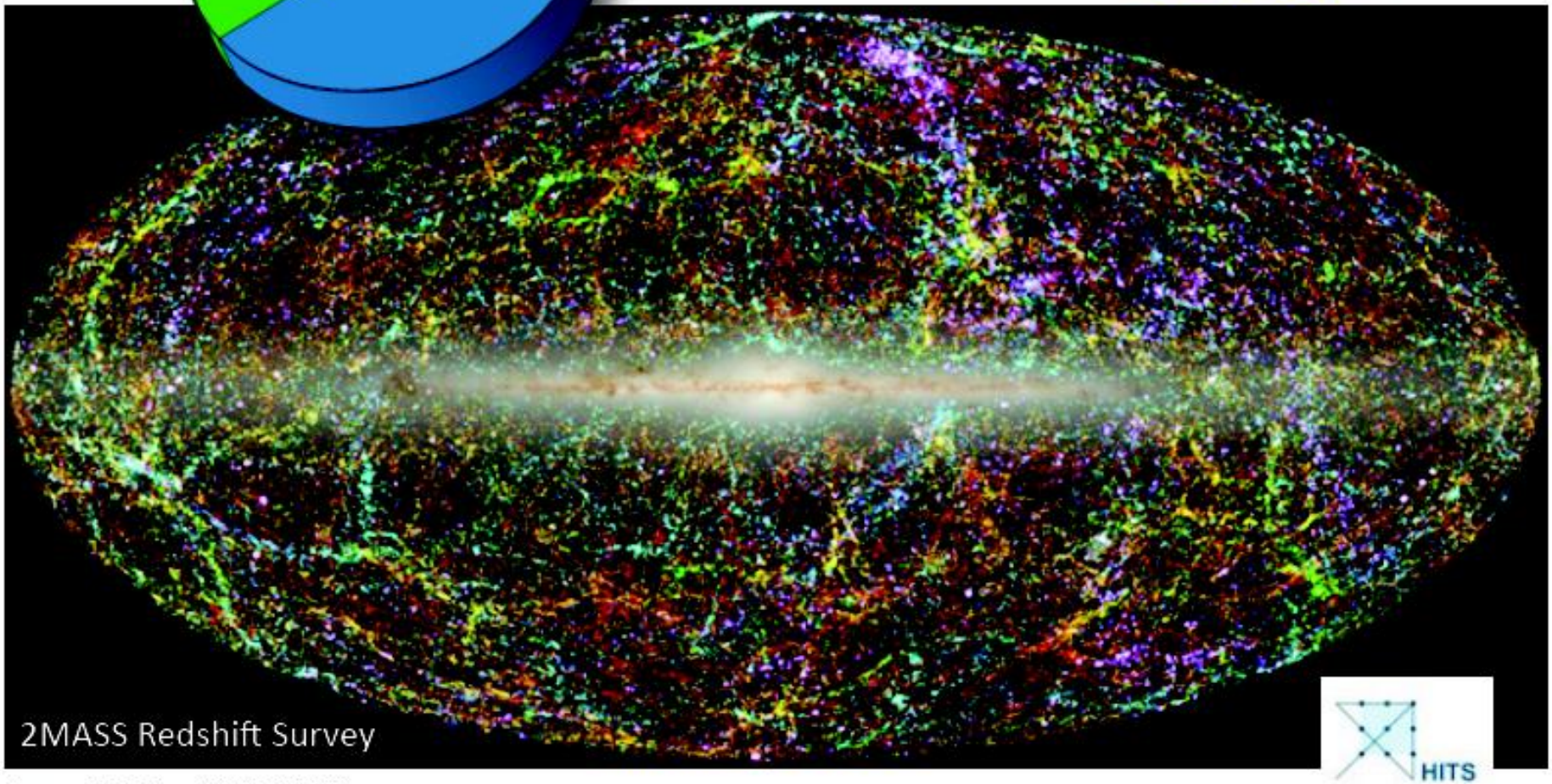


WMAP

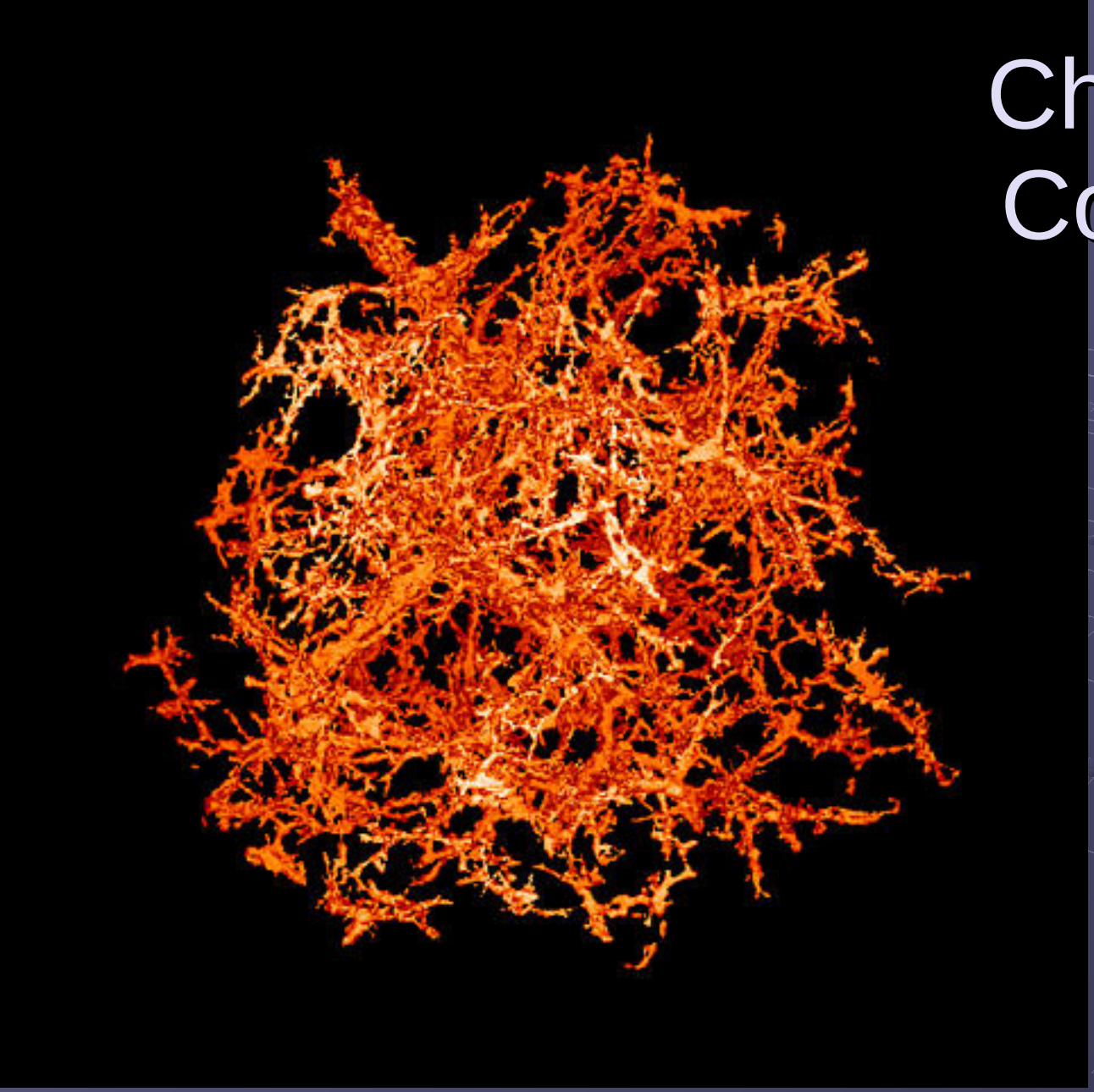
2.725 Kelvin



0.0002 degrees



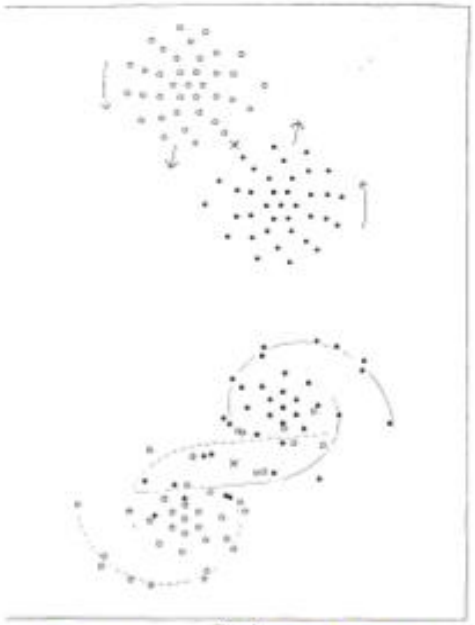
(Image: T.H. Jarrett (IPAC/SSC))



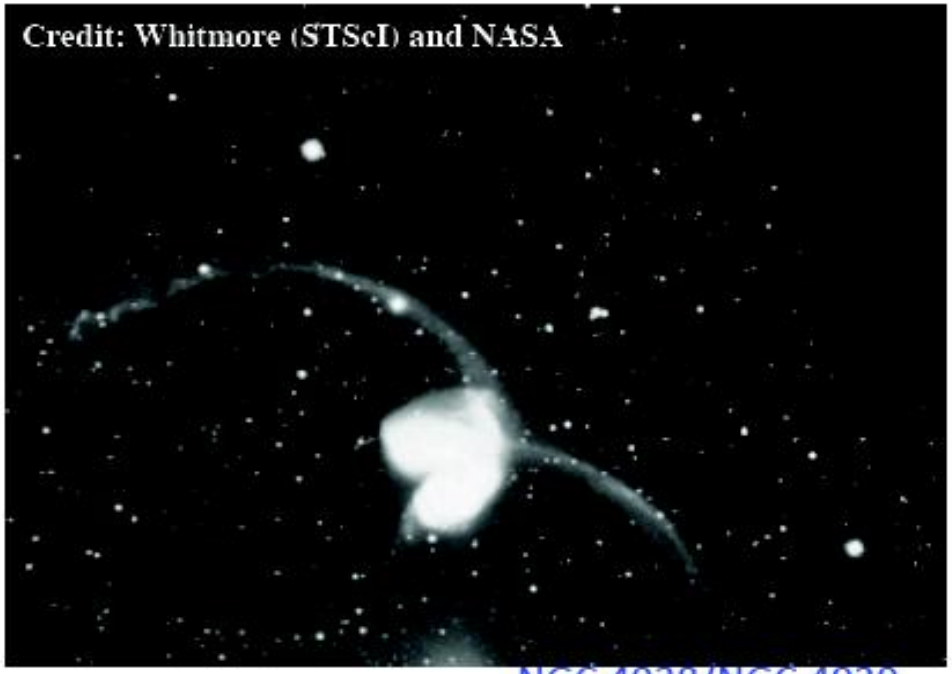
Challenges: Cosmology

Structure of Voids and Filaments in 3D

(From Virgo-Webpages)



Holmberg, 1937/1941



NGC 4038/NGC 4039

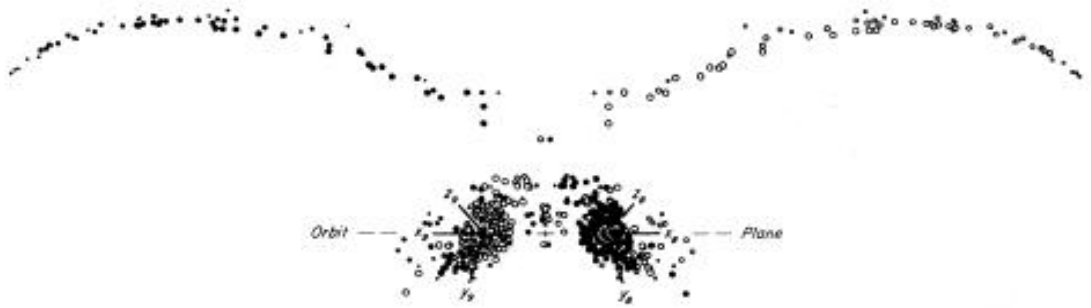


FIG. 23.—Symmetric model of NGC 4038/9. Here two identical disks of radius $0.75R_{min}$ suffered an $e \approx 0.5$ encounter with orbit angles $i_8 = i_9 = 60°$ and $\omega_8 = \omega_9 = -30°$ that appeared the same to both. The above all-inclusive views of the debris and remnants of these disks have been drawn exactly normal and edge-on to the orbit plane; the latter viewing direction is itself 30° from the line connecting the two pericenters. The viewing time is $t = 15$, or slightly past apocenter. The filled and open symbols again disclose the original loyalties of the various test particles.

Toomre & Toomre,1972, ApJ, 178, 623

HITS

Computer Physics - Astrophysics

Black Holes in Star Clusters

VIRGO – Pisa 3km
LIGO – Livingston, LA
Hanford, WA
1km
GEO600 – Hannover
600m
AIGO – Australien
(planned, 5 km)

http://www.ligo-la.caltech.edu/
http://www.ego-gw.it
http://www.geo600.uni-hannover.de

Outreach to 50 Millionen light years (Neutron Stars)

Example: VIRGO Detector in Cascina near Pisa, Italy

GW Detection Frequency Time Diagram
Top: Our simulation (Wang et al. 2016, Sobolenko et al. In prep.)
Down: Abott et al. 2016 LIGO measurement

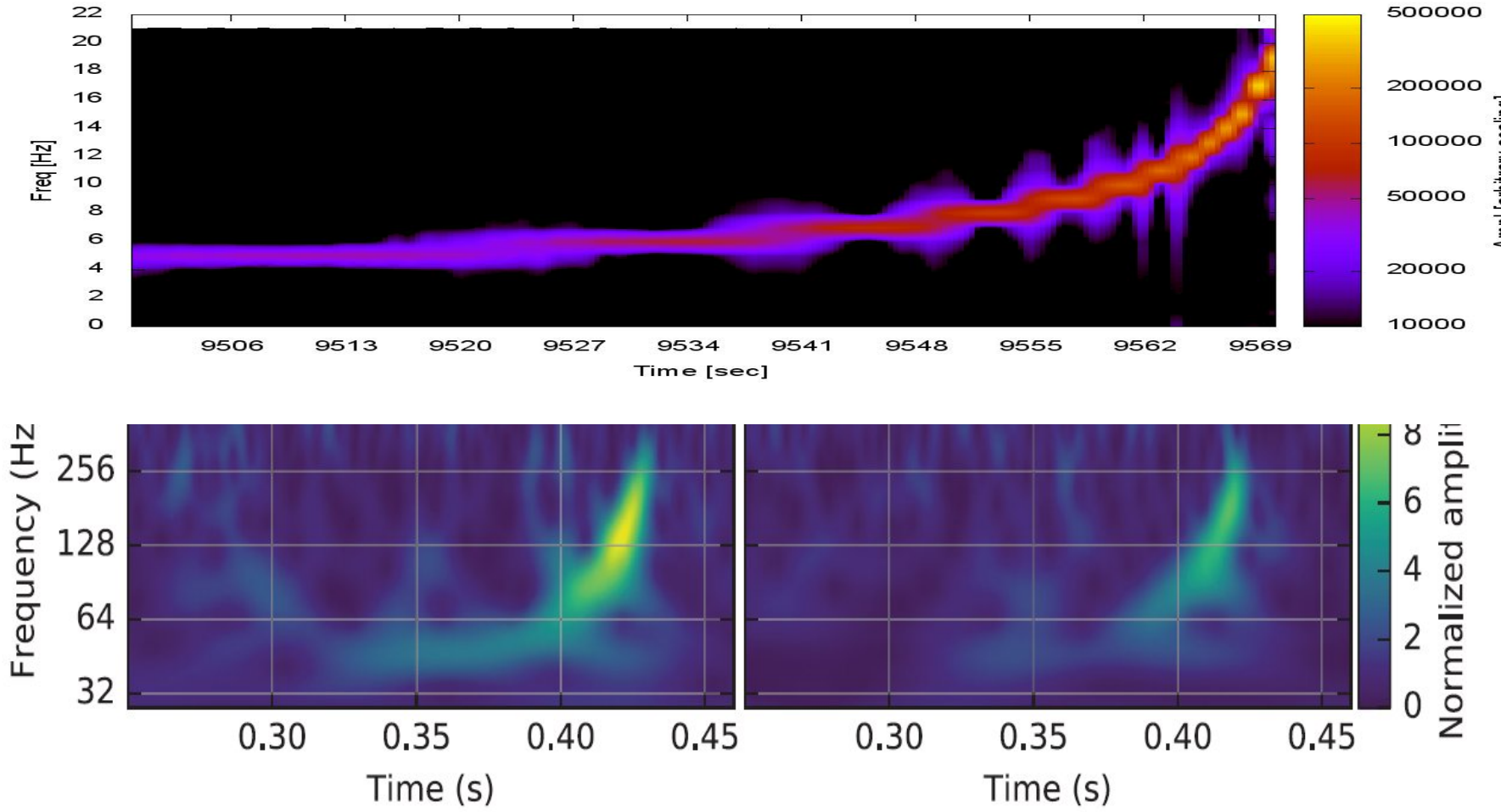


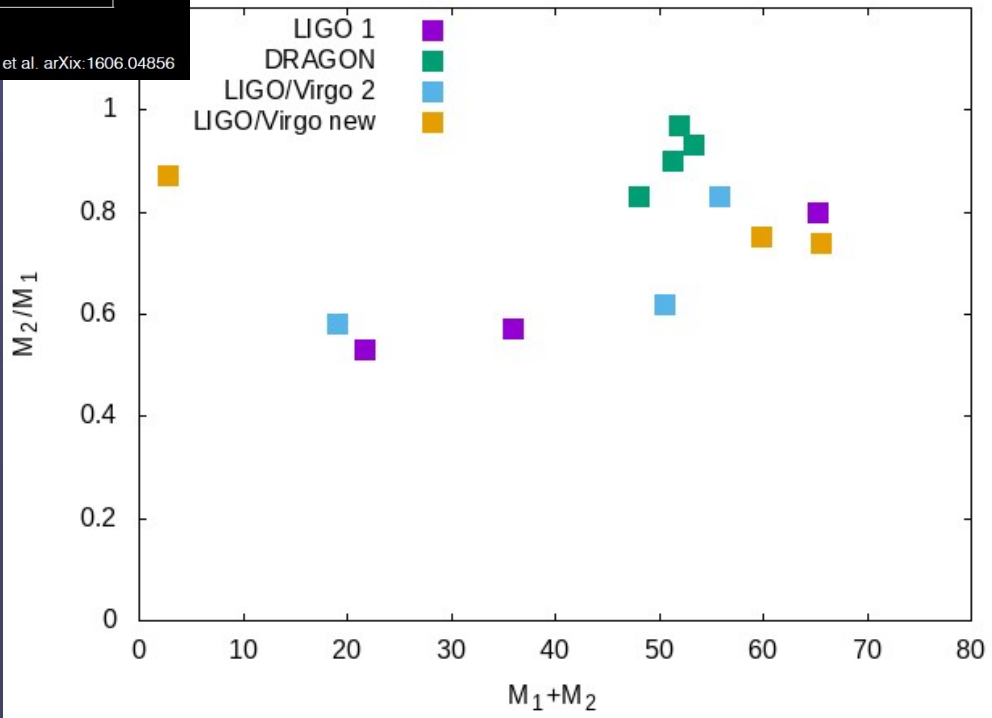
FIG. 1. The gravitational-wave event GW150914 observed by the LIGO Hanford (H1, left column panels) and Livingston (L1, right column panels) detectors. Times are shown relative to September 14, 2015 at 09:50:45 UTC. For visualization, all time series are filtered

The Observed LIGO Events – Table from Brown's Talk at KITP (2)

	GW150914	GW151226	LVT151012
Source Mass 1	$36.2^{+5.2}_{-3.8}\ M_\odot$	$14.2^{+8.3}_{-3.7}\ M_\odot$	$23^{+18}_{-6}\ M_\odot$
Source Mass 2	$29.1^{+3.7}_{-4.4}\ M_\odot$	$7.5^{+2.3}_{-2.3}\ M_\odot$	$13^{+4}_{-5}\ M_\odot$
Luminosity Distance	420^{+150}_{-180} Mpc	440^{+180}_{-190} Mpc	1000^{+500}_{-500} Mpc

Abbott, ..., DAB, et al. arXix:1606.04856

The Observed LIGO/Virgo Events…
The DRAGON events in the supercomputer...



MPA Garching
Highlight March 2016

http://www.mpa-garching.mpg.de/328833/hl201603

HIGHLIGHT: MARCH 2016

The DRAGON globular cluster simulations: a million stars, black holes and gravitational waves

March 01, 2016

An international team of experts from Europe and China has performed the first simulations of globular clusters with a million stars on the high-performance GPU cluster of the Max Planck Computing and Data Facility. These – up to now - largest and most realistic simulations can not only reproduce observed properties of stars in globular clusters at unprecedented detail but also shed light into the dark world of black holes. The computer models produce high quality synthetic data comparable to Hubble Space Telescope observations. They also predict nuclear clusters of single and binary black holes. The recently detected gravitational wave signal might have originated from a binary black hole merger in the center of a globular cluster.

RGB image of a simulated globular cluster
© MPA

Globular clusters are truly enigmatic objects. They consist of hundreds of thousands luminous stars and their remnants, which are confined to a few tens of parsecs (up to 100 lightyears) – they are the densest and oldest gravitationally bound stellar systems in the Universe. Their central star densities can reach a million times the stellar density near our Sun. About 150 globular clusters orbit the Milky Way but more massive galaxies can have over 10,000 gravitationally bound globular clusters. As their stars have mostly formed at the same time but with different masses, globular clusters are ideal laboratories for studies of stellar dynamics and stellar evolution.

The dynamical evolution of globular clusters, however, is very complex. Unlike in galaxies, the stellar densities are so high that stars can interact in close gravitational encounters or might even physically collide with each other. Because of these interactions there are more tightly bound binary stars than for normal galactic field stars. Moreover, in a process called mass-segregation more massive stars sink to the center of the system.

The Kavli Institute for Astronomy and Astrophysics at Peking University
北京大学科维理天文与天体物理研究所

简体中文 English

HOME | ABOUT KIAA | PEOPLE | ACTIVITIES | NEWS | SCIENCE | JOB OPPORTUNITIES | OUTREACH | VISITOR INFO | INTERNAL

http://kiaa.pku.edu.cn **News...**

Home » The DRAGON globular cluster simulations: a million stars, black holes and gravitational waves

Search @ KIAA-PKU

Search

Upcoming Events

Pulsars and FRBs: Recent Developments
Speaker: Richard N. Manchester (CSIRO Astronomy and Space Science, Australia)
3 Nov 2016 - 4:00pm
KIAA-PKU Auditorium

The role of vortices in multi-dimensional protoplanetary disks
Speaker: Hui Li
7 Nov 2016 - 12:00pm
DoA, Rm 2907

TBD
Speaker: Jessy Jose
8 Nov 2016 - 12:00pm
KIAA-PKU

more

Navigation

Biblio

The DRAGON globular cluster simulations: a million stars, black holes and gravitational waves

By shuyan on Mon, 2016-06-27 08:55

An international team of experts from China and Europe has performed the first simulations of globular clusters with a million stars on the high-performance GPU

1. ***Wang, Long; Spurzem, Rainer; Aarseth, Sverre; Nitadori, Keigo; Berczik, Peter; Kouwenhoven, M. B. N.; Naab, Thorsten***
NBODY6++GPU: ready for the gravitational million-body problem
2015, MNRAS, 450, 4070
Source

2. ***Wang, Long; Spurzem, Rainer; Aarseth, Sverre; Giersz, Mirek; Askar, Abbas; Berczik, Peter; Naab, Thorsten; M. B. N. Kouwenhoven, Riko Schadow***
The DRAGON simulations: globular cluster evolution with a million stars

北京大学 PEKING UNIVERSITY

Computational and Computer Science

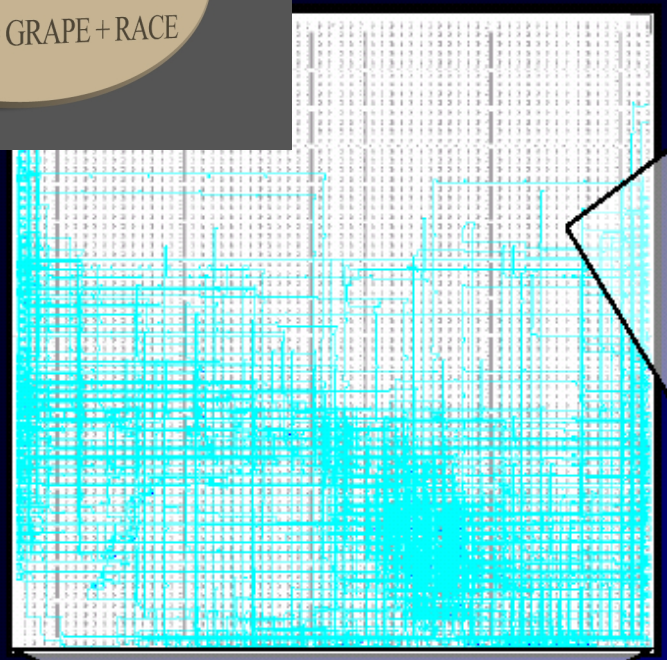
More About the Future

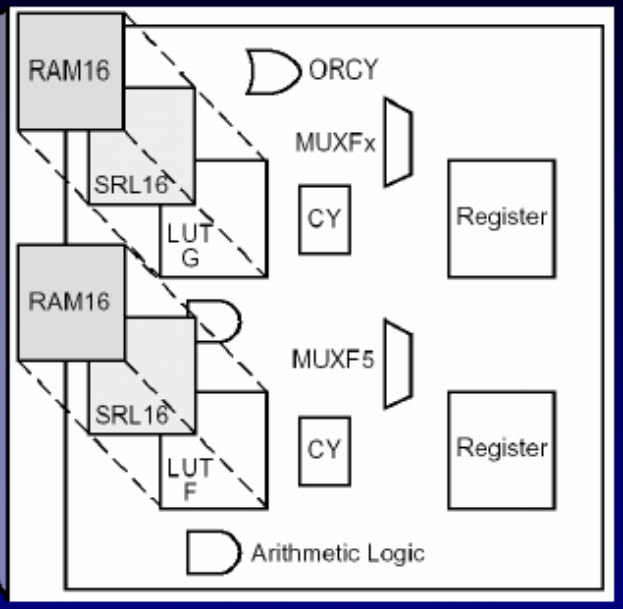
GRACE

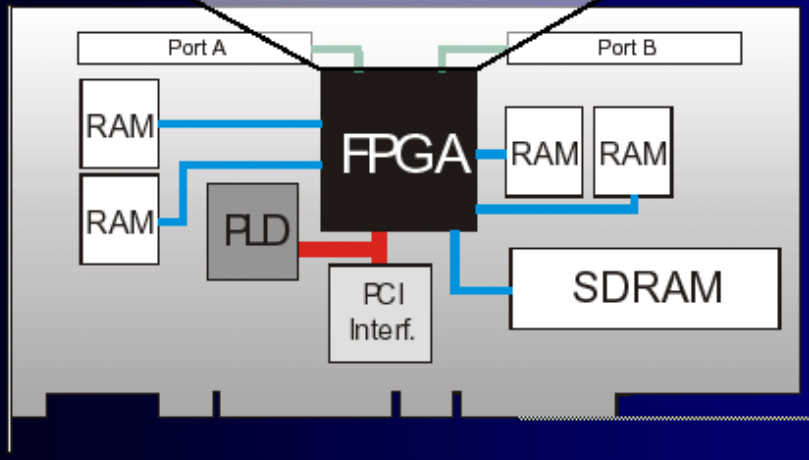
GRACE = GRAPE + RACE

FPGA-Plattform MPRACE

2005 - 2009







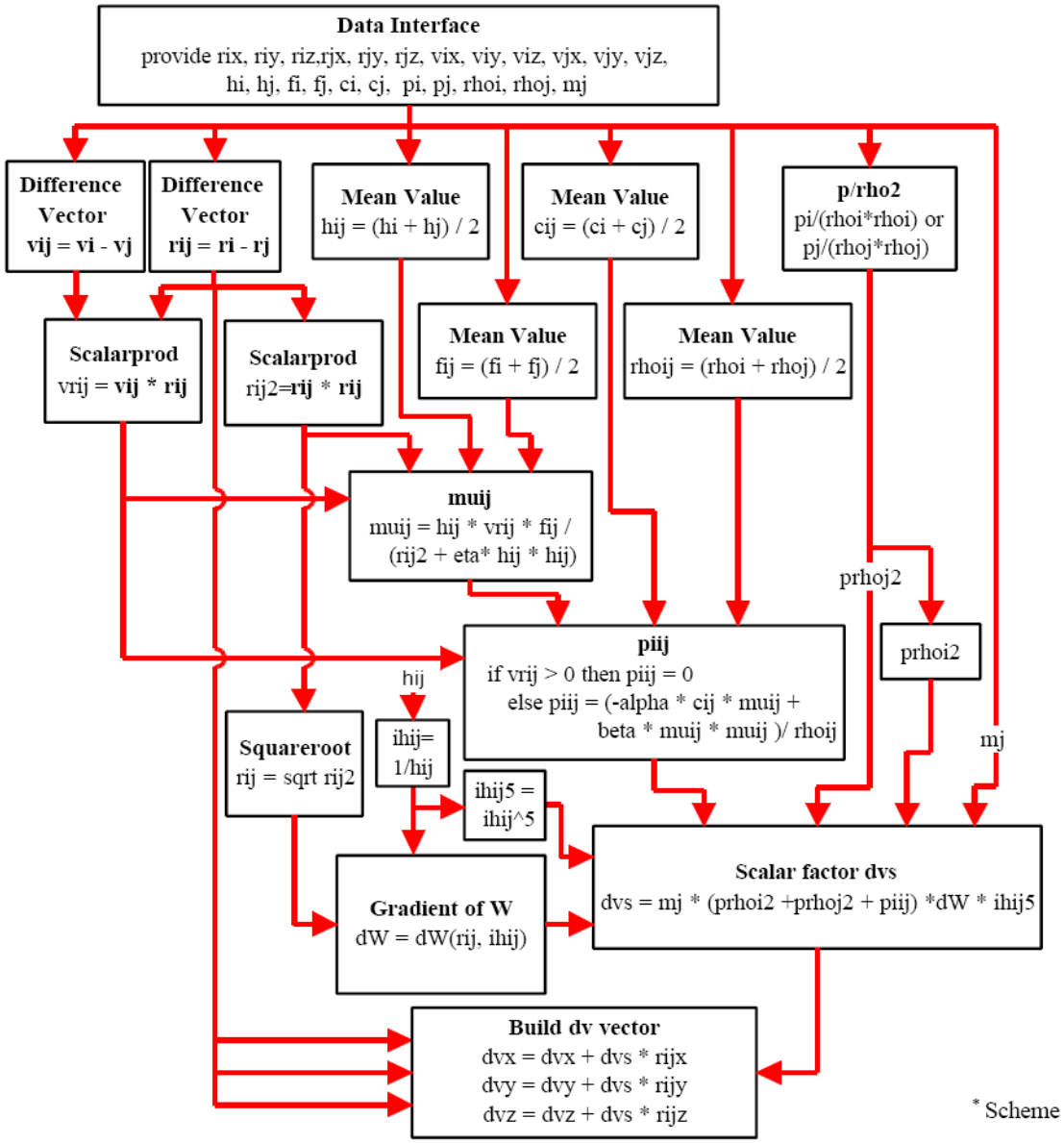
lienhart@ti.uni-mannheim.de

GRACE

GRACE = GRAPE + RACE

FPGA...

Pressure force pipeline:



* Scheme doesn't show energy term

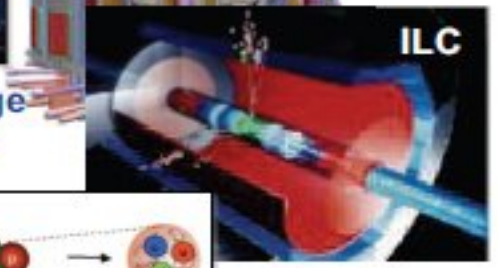
Exascale simulation will enable fundamental advances in basic science

BERKELEY LAB

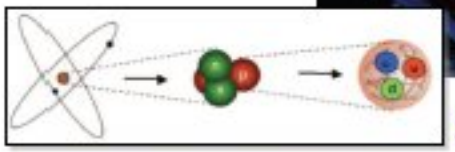
- High Energy & Nuclear Physics
  - Dark-energy and dark matter
  - Fundamentals of fission fusion reactions
- Facility and experimental design
  - Effective design of accelerators
  - Probes of dark energy and dark matter
  - ITER shot planning and device control
- Materials / Chemistry
  - Predictive multi-scale materials modeling: observation to control
  - Effective, commercial technologies in renewable energy, catalysts, batteries and combustion
- Life Sciences
  - Better biofuels
  - Sequence to structure to function

These breakthrough scientific discoveries and facilities require exascale applications and resources

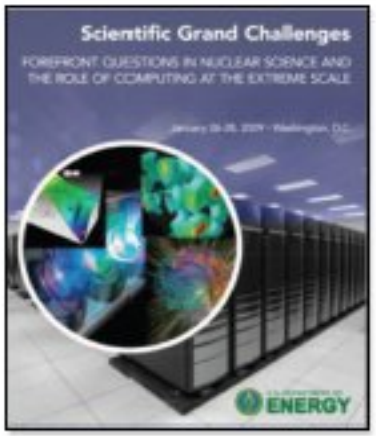
ITER

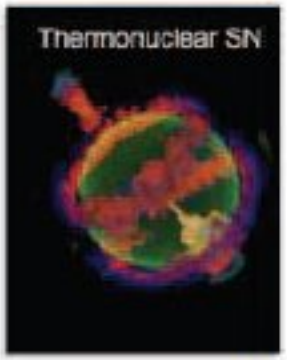


Hubble image of lensing



Structure of nucleons





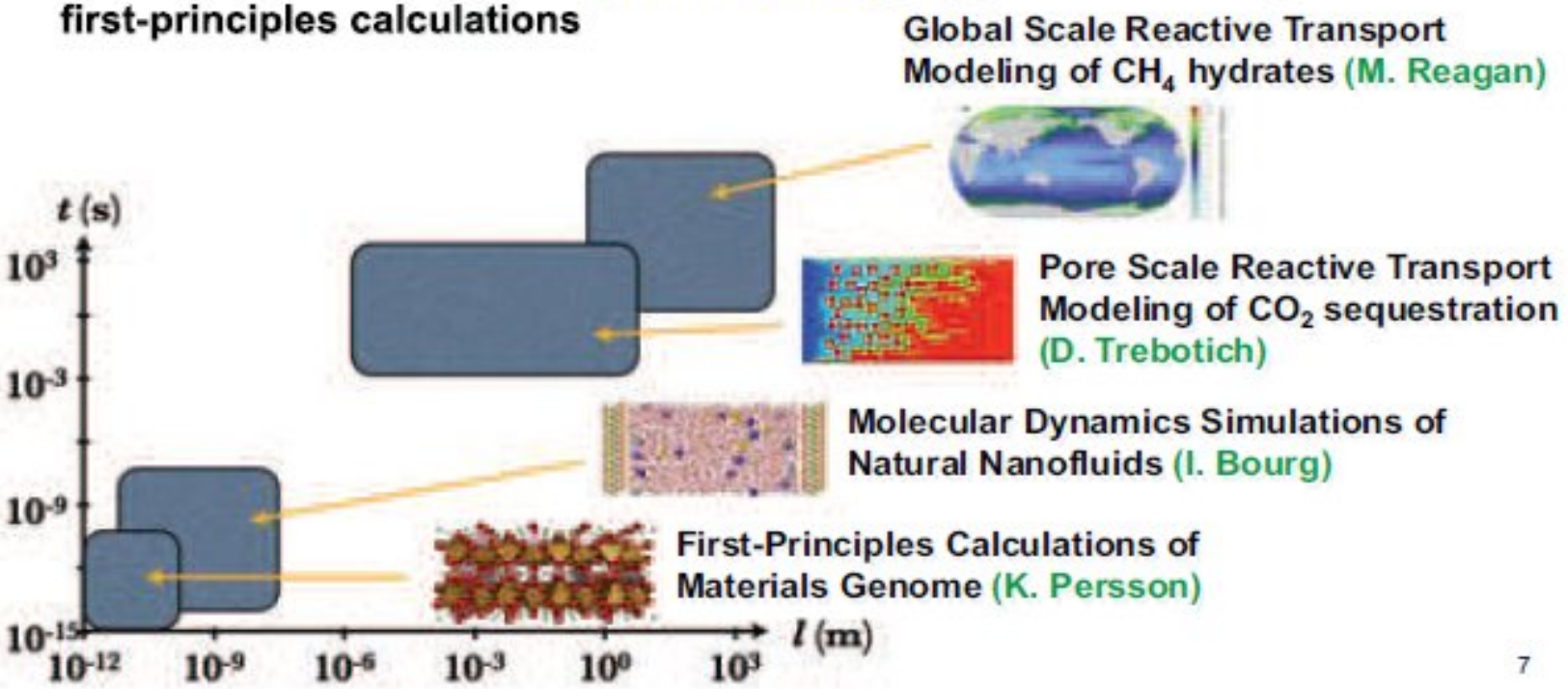
Slide: Horst Simon, Exascale Computing

Advanced Computation in Energy Science at LBNL

Probe natural systems under constraints that are difficult or impossible to impose in the field or laboratory

Reveal the manner in which large-scale phenomena arise from smaller-scale properties

Discover new materials for green technology applications through first-principles calculations

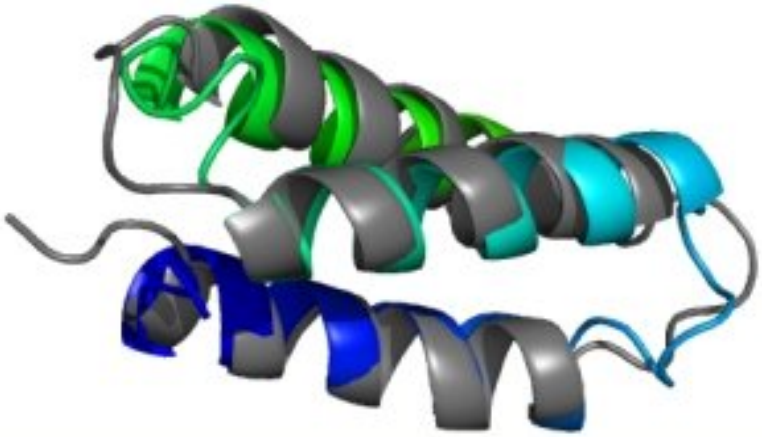


Slide: Horst Simon, Exascale Computing

Research

http://www.fz-juelich.de/ias/jsc/EN/Research/research_node.html

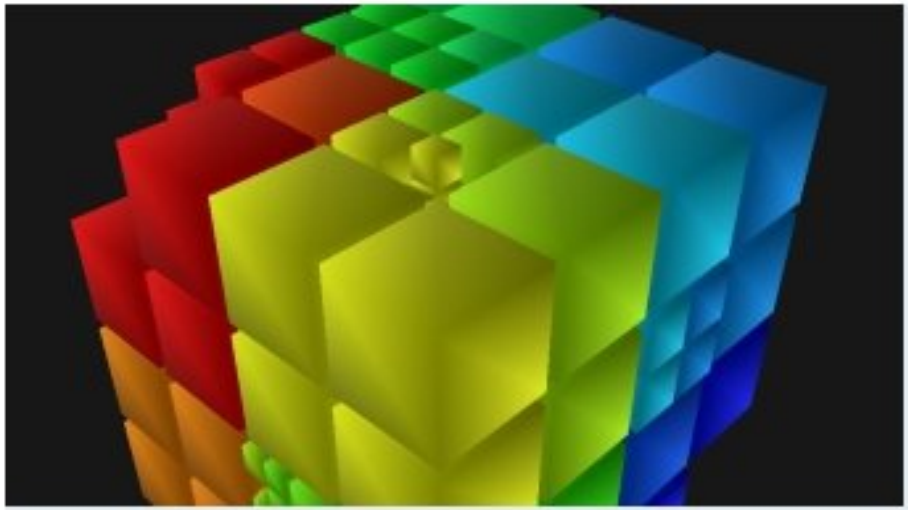
JSC's research and development concentrates on mathematical modelling and numerical, especially parallel algorithms for quantum chemistry, molecular dynamics and Monte-Carlo simulations. The focus in the computer sciences is on cluster computing, performance analysis of parallel programs, visualization, computational steering and grid computing.



Modelling and Simulation

The simulation of complex systems in natural science or engineering depends on the development of adequate mathematical models. Thus the development of realistic and yet efficient models is a core activity at JSC. Examples of simulations are:

- Computational Plasma Physics
- Protein Folding
- Quantum Information Processing
- Civil Security and Traffic



Algorithms and Methods

Efficient simulations need powerful algorithms and methods. JSC focusses on the development of the following methods:

- Fast Coulomb Solvers
- Parallel-In-Time Integration
- Fast Multipole Method
- Parallel I/O

HPCI: High Performance Computing Infrastructure

RIST

- Established as Japanese integrated high performance computing infrastructure in 2011
- Variety of computer systems are connected via high speed academic backbone network and provided as ***HPCI*** resources to users in ***Japan and overseas,*** Also it will be a platform for international collaborations.

FY2017 Allocated computing resources

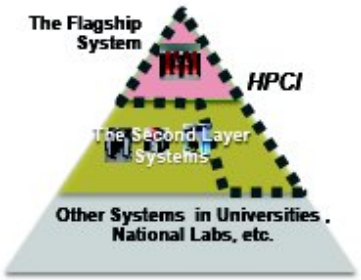
~9.6 PFlops x Yr.

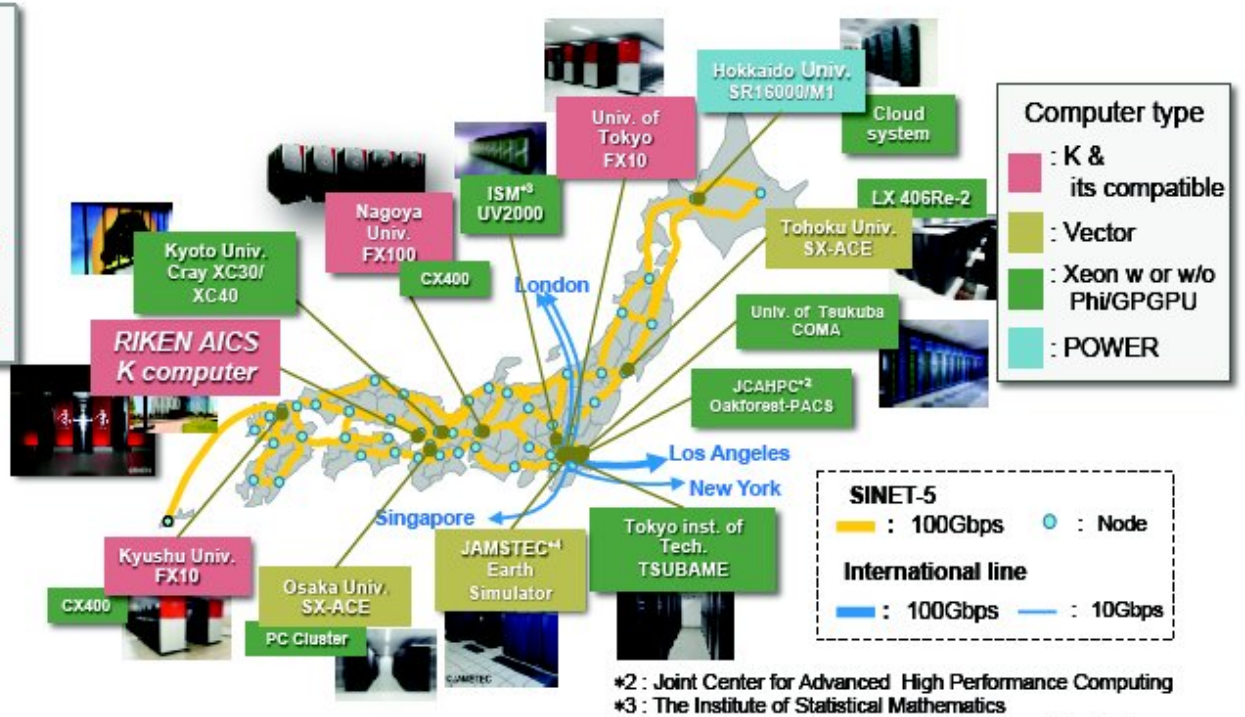
K computer

~4 PFlops x Yr.

Others in total

~5.6 PFlops x Yr.





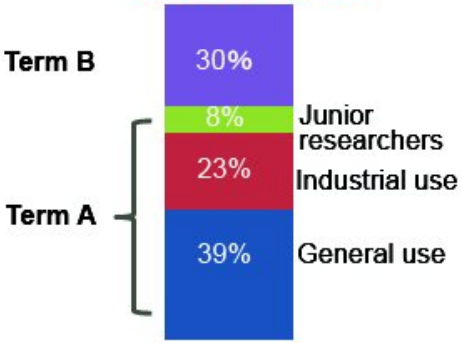
*2 : Joint Center for Advanced High Performance Computing
*3 : The Institute of Statistical Mathematics
*4 : Japan Agency for Marine-Earth Science and Technology

Resources allocation and Awarding results of FY 2017

RIST

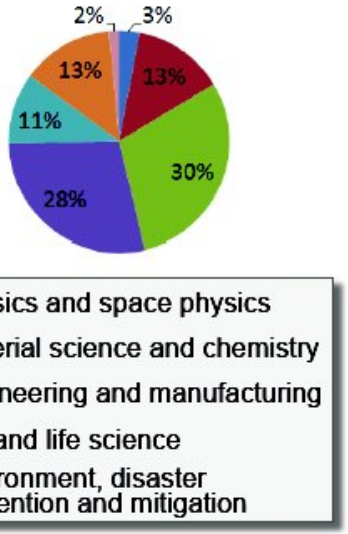
■ *K computer*

~ 4 PFlops・year (corresponding to 45% of total K resource)



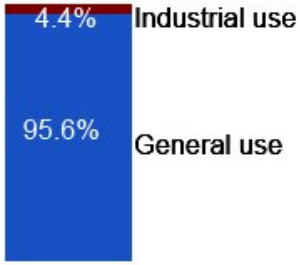
Submitted	96
Awarded	67
Ratio	70%

Major Research areas

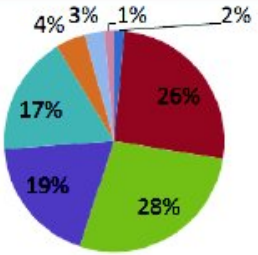


■ Other ***HPCI*** system

~ 5.6 PFlops・year



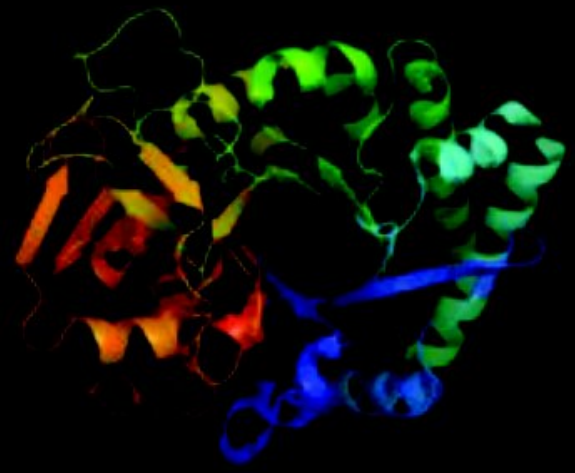
Submitted	155
Awarded	69
Ratio	45%



Deep Learning Is Getting Real Now …

Deep learning algorithm does as well as dermatologists in identifying skin cancer

Artificial intelligence could build new drugs faster than any human team



AMD | RADEON

MILS: Machine Intelligence Led Services

Mills

Information Revolution

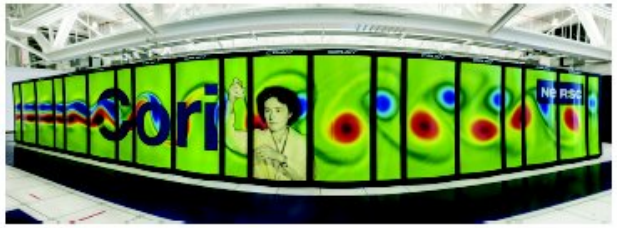
MILS

Intelligence Too Big for a Single Machine

"We're seeing a rebirth of artificial intelligence driven by the cloud, huge amounts of data and the learning algorithms of software," Larry Smarr, founding director of the California Institute for Telecommunications and Information Technology

http://bits.blogs.nytimes.com/2014/06/11/intelligence-too-big-for-a-single-machine/

Deep Learning in Science

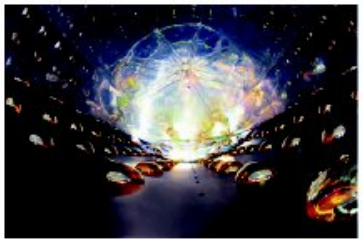


Cray XC40 system at NERSC

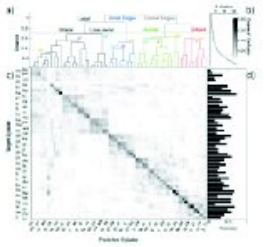
NERSC

BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY

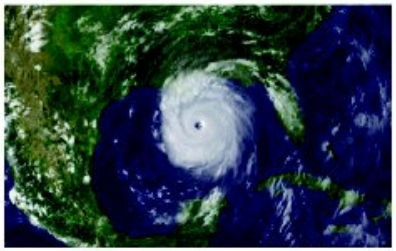
Modeling galaxy shapes



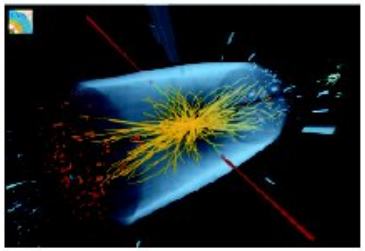
Clustering Daya Bay events



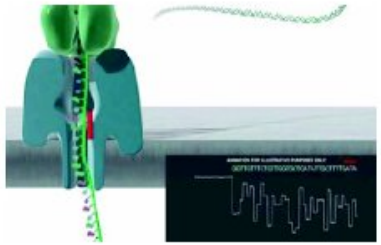
Decoding speech from ECoG



Detecting extreme weather



Classifying LHC events



Oxford Nanopore sequencing

Opportunities to apply DL widely in support of classic HPC simulation and modelling

...the new hype: HPC and AI … here by T. Maruyama, ISC17:

Processor Designed for Deep Learning

FUJITSU

K computer

Utilizing technologies derived from the K computer

Features of DLU

- Architecture designed for Deep Learning
- Low power consumption design
- Optimized precision

→ Goal: 10x Performance / Watt compared to competitors

- Scalable design with Tofu interconnect technology

→ Ability to handle large-scale neural networks

FY2018 ~

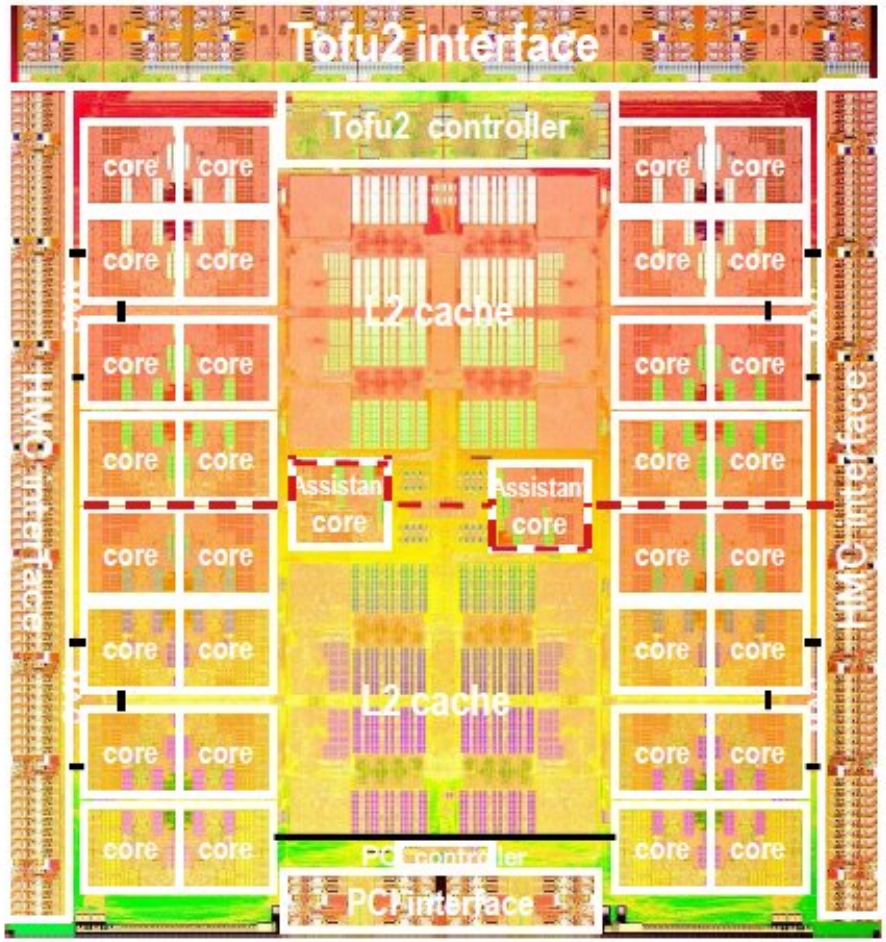
DLU™

(Deep Learning Unit)

FUJITSU DLU™ Deep Learning Unit

SPARC64™ XIfx Chip (HPC)

FUJITSU

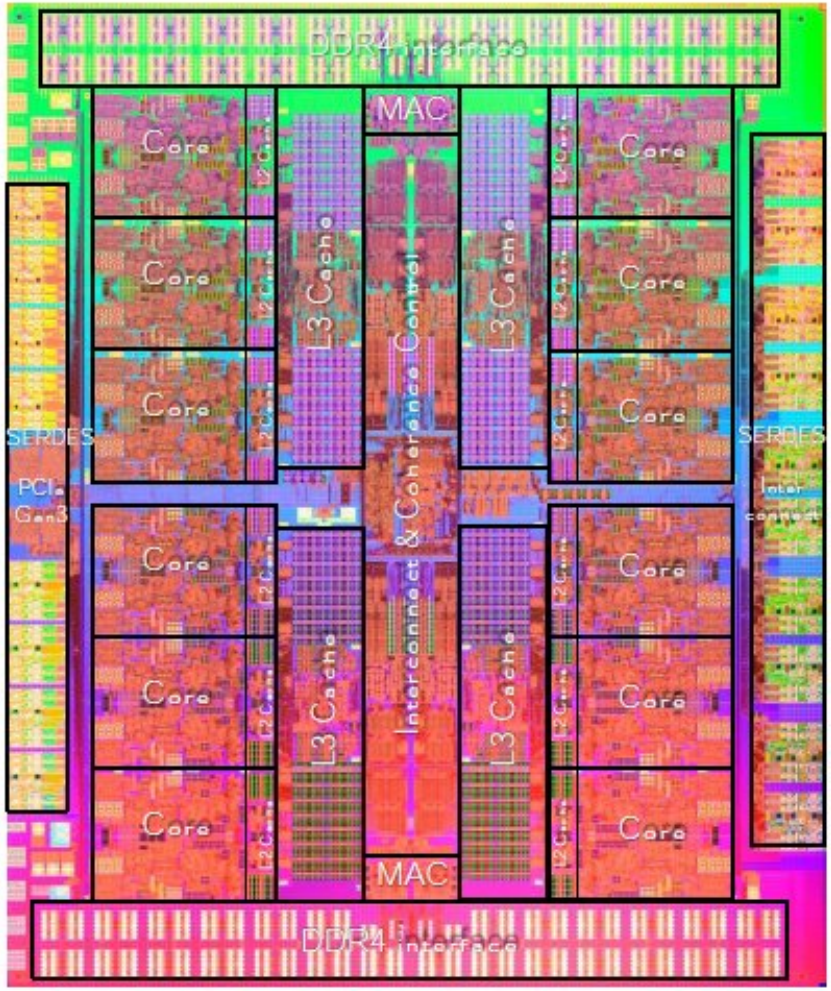


Many (32+2) cores, Medium CPU GHz

- Architecture Features
  - 32 computing cores + 2 assistant cores
  - HPC-ACE2 (256bit SIMD) Fujitsu's ISA enhancements
  - Sector Cache: Cache with SW controllability
  - 24 MB L2 cache
- 20nm CMOS
  - 3,750M transistors
  - 2.2GHz
- Performance (peak)
  - 1.1TFlops
  - HMC 240GB/s x 2 (in/out)
  - Tofu2 125GB/s x 2 (in/out)

SPARC64™ XII Chip (UNIX)

FUJITSU

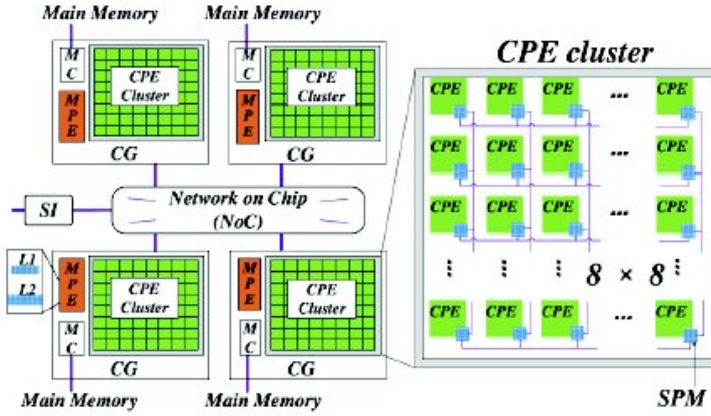


Multiple big cores, High CPU GHz

- Architecture Features
  - 12 cores x 8 threads
  - SWoC ("Software on Chip") Fujitsu's ISA enhancements
  - 32MB L3 cache
  - Embedded MAC and IOC
- 20nm CMOS
  - 25.8mm x 30.8mm
  - 5,450M transistors
  - 4.25GHz (up to 4.35GHz with "High Speed Mode" enabled)
- Performance (peak)
  - 417GIPS / 835GFlops
  - 153GB/s memory throughput

SUNWAY TAIHULIGHT

- SW26010 processor (Chinese design, ISA, & fab)
- 1.45 GHz
- Node = 260 Cores (1 socket)
  - 4 – core groups
  - 32 GB memory
- 40,960 nodes in the system
- 10,649,600 cores total
- 1.31 PB of primary memory (DDR3).
- 125.4 Pflop/s theoretical peak
- 93 Pflop/s HPL, 74% peak
- 15.3 Mwatts water cooled
- 3 of the 6 finalists for Gordon Bell Award@SC16



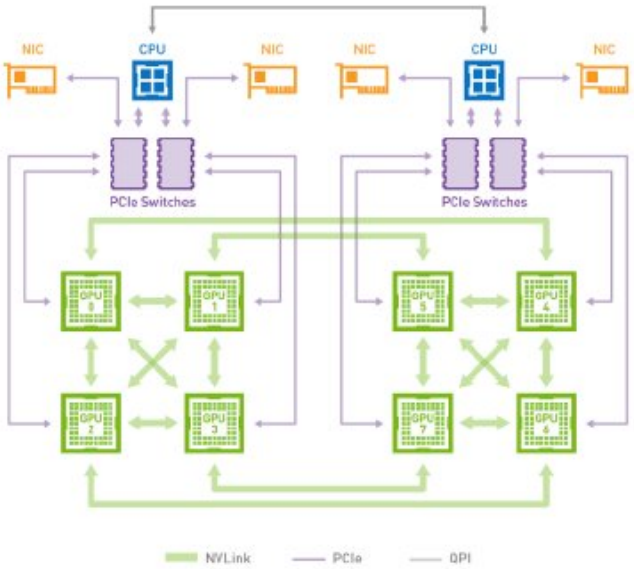
SYSTEMS APPROACHES TO EXASCALE

More GPUs, Fewer CPUs:

Titan: 1GPU/CPU

Summit: 3 GPUs/CPU

Exascale: ?



Faster Serial Processing (~~MANY CORE~~):

Run 8x Fewer Cores @ 2x Speed

Denser Packaging:

Move Networking to Faster Local Networks: NVLINK

EXASCALE: "50X FASTER THAN TITAN"

Per-GPU -hardware- speedups will be less than 50x

	2013 Kepler	2016 Pascal	2017 Volta	2021*	Speedup
FP64 Tflop/s	1.5	4.5	7	7-21	5-15
Memory GB/s	288	720	900	900-4000	3-14
I/O BW GB/s	7	80	150	150-500	20-70
Deep Learning FP16 Tflop/s	3	20	112	112-500	37-166
Deep Learning BW GB/s	576	2880	3600	3600-16000	6-27

*Extremely Fuzzy Public Projections for 2021