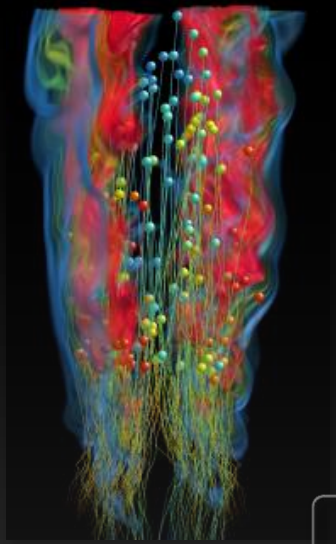
NVIDIA

SC13
GPU Technology Theater

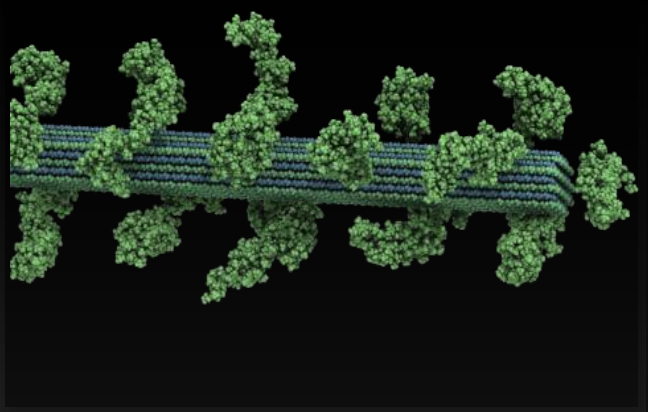
Accelerated Computing: What's Coming Next

Ian Buck, General Manager, Accelerated Computing, NVIDIA

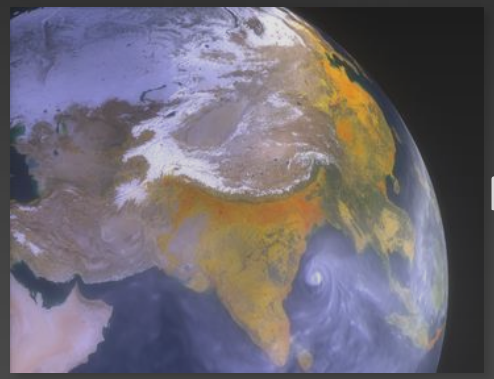


S3D
Model Combusion for higher efficiency fuels & engines

LAMMPS
Model biofuels. Reduce carbon emissions & dependence on foreign oil

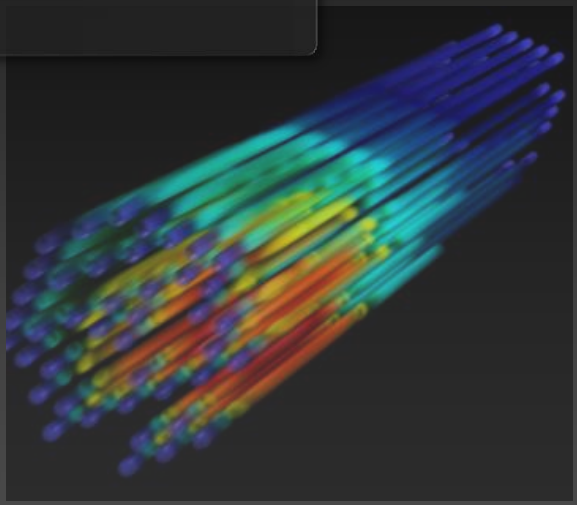


Supercomputing Science

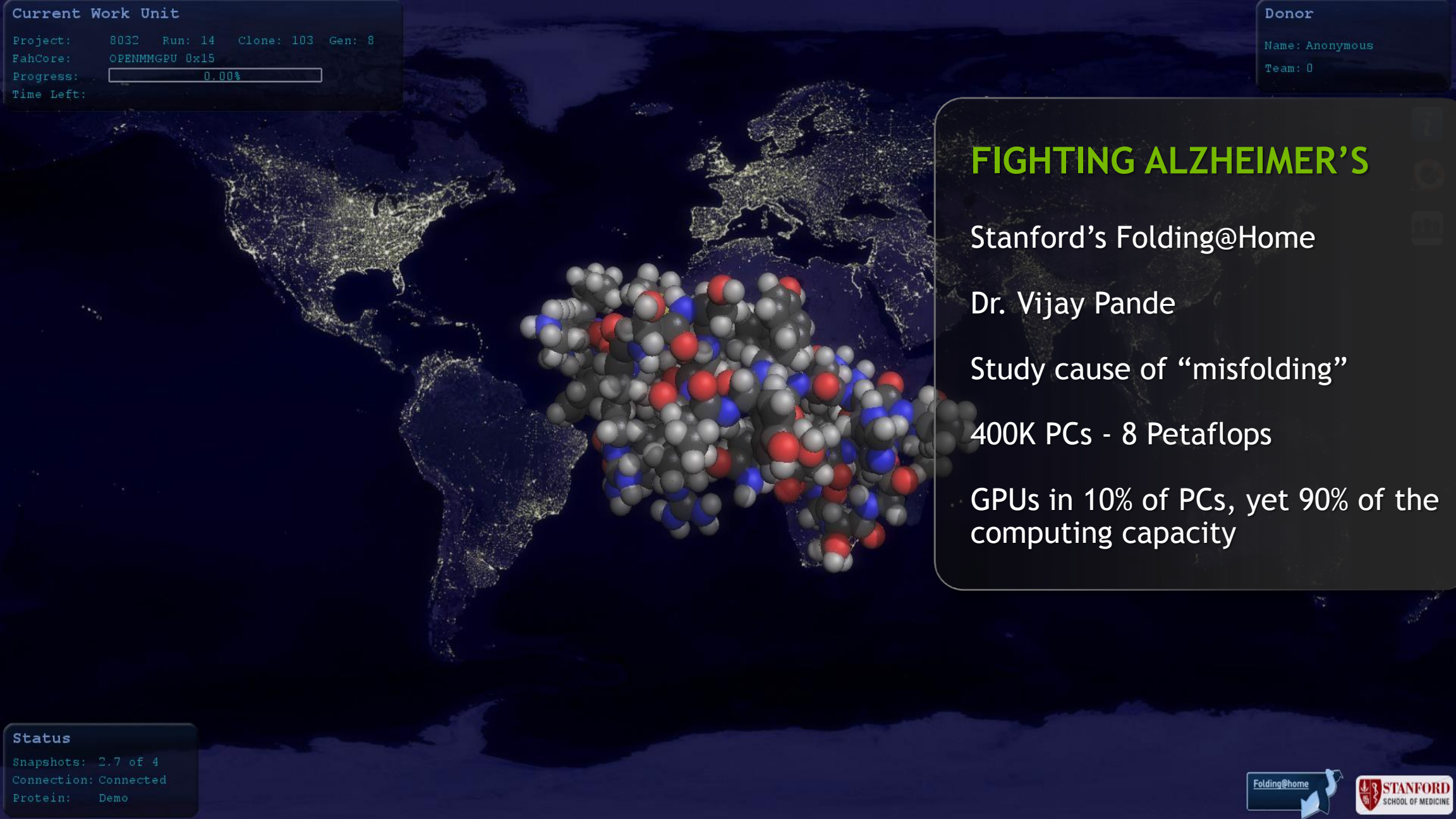


CAM-SE
Model global climate change & explore mitigation strategies

Denovo
Simulate radiation transport for safe, clean, fusion energy



Source: ORNL


Current Work Unit
Project: 8032 Run: 14 Clone: 103 Gen: 8
FahCore: OPENMMGPU 0x15
Progress: 0.00%
Time Left:
Donor
Name: Anonymous
Team: 0
FIGHTING ALZHEIMER'S
Stanford's Folding@Home
Dr. Vijay Pande
Study cause of "misfolding"
400K PCs - 8 Petaflops
GPUs in 10% of PCs, yet 90% of the computing capacity
Status
Snapshots: 2.7 of 4
Connection: Connected
Protein: Demo
Folding@home
STANFORD SCHOOL OF MEDICINE

GPUs Power World's Largest Artificial Brain

2012
Google Datacenter

1000 CPU Servers

1.7 billion parameter
neural network

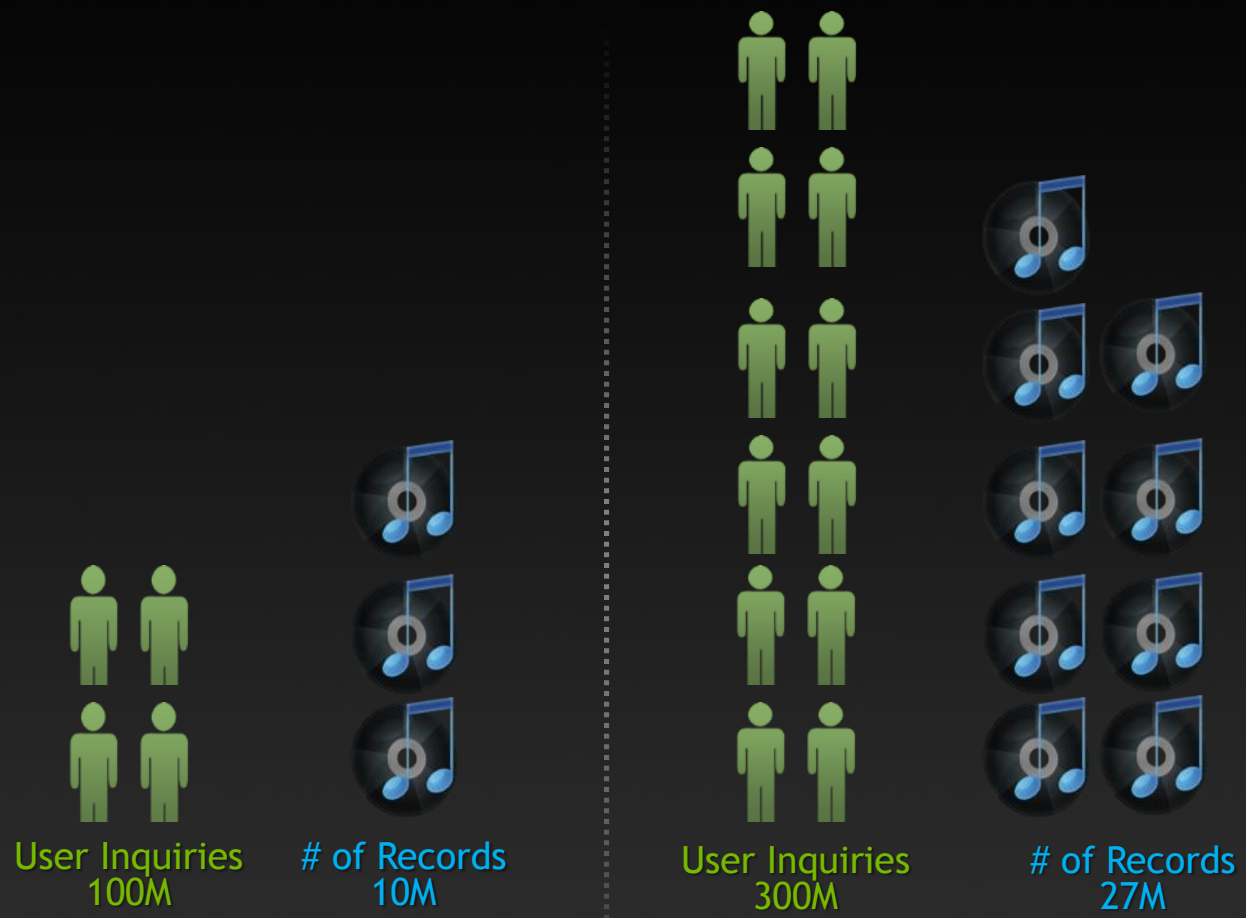
Today
Stanford AI Lab

16 GPU-Accelerated Servers

11.2 billion parameter
neural network

6.5x Bigger Artificial Brain

Shazam: 300M GPU Accelerated Searches

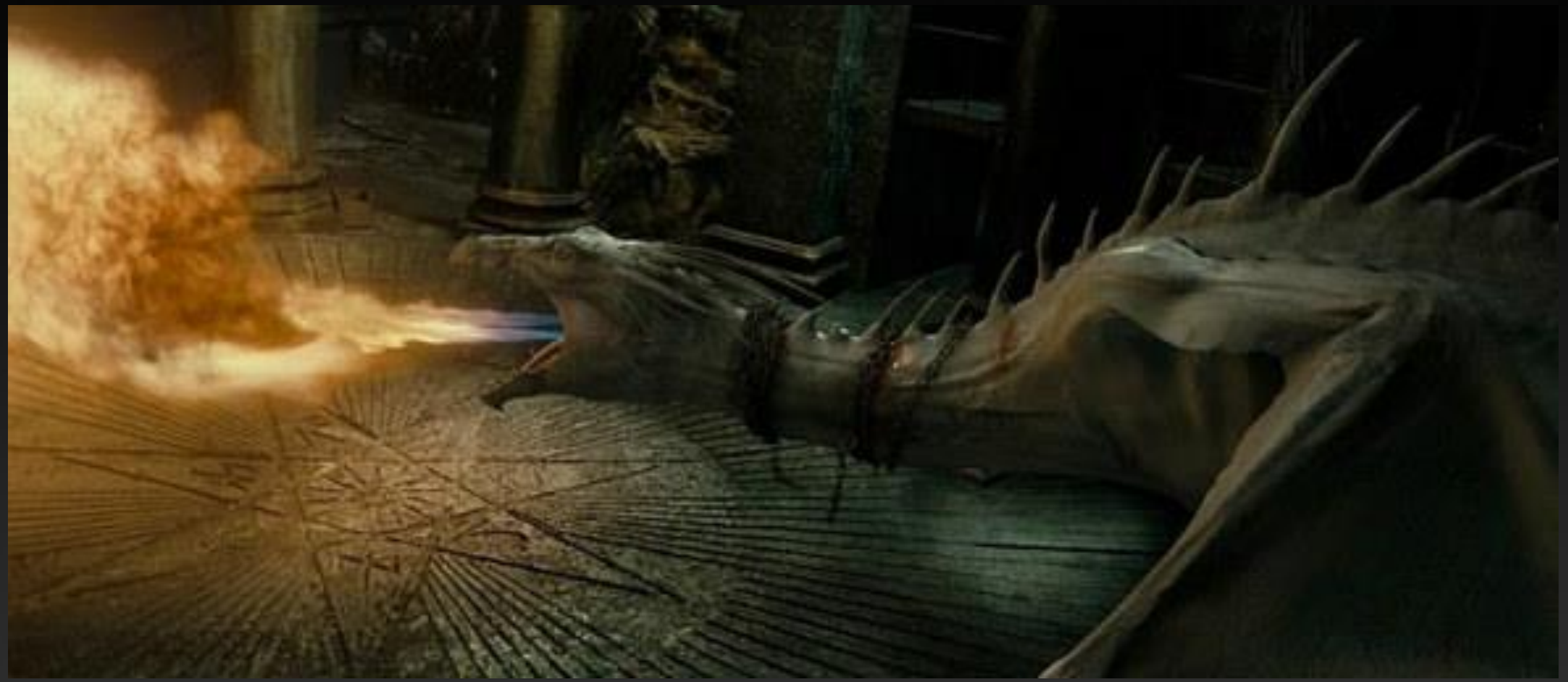


Hundreds
of GPUs in Datacenter

GPUs Enable Scalable Growth

User Inquiries averaged per Month

Double Negative



CUDA: World's Most Pervasive Parallel Programming Model

14,000 Institutions with CUDA Developers

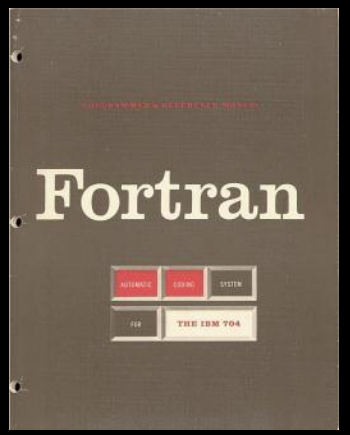
2,000,000 CUDA Downloads

487,000,000 CUDA GPUs Shipped

700+ University Courses
In 62 Countries

Parallelism in Mainstream Languages

- Enable more programmers to write parallel software
  - CUDA as a platform target for standard languages / APIs / libraries
- Drive and embrace standards in key languages



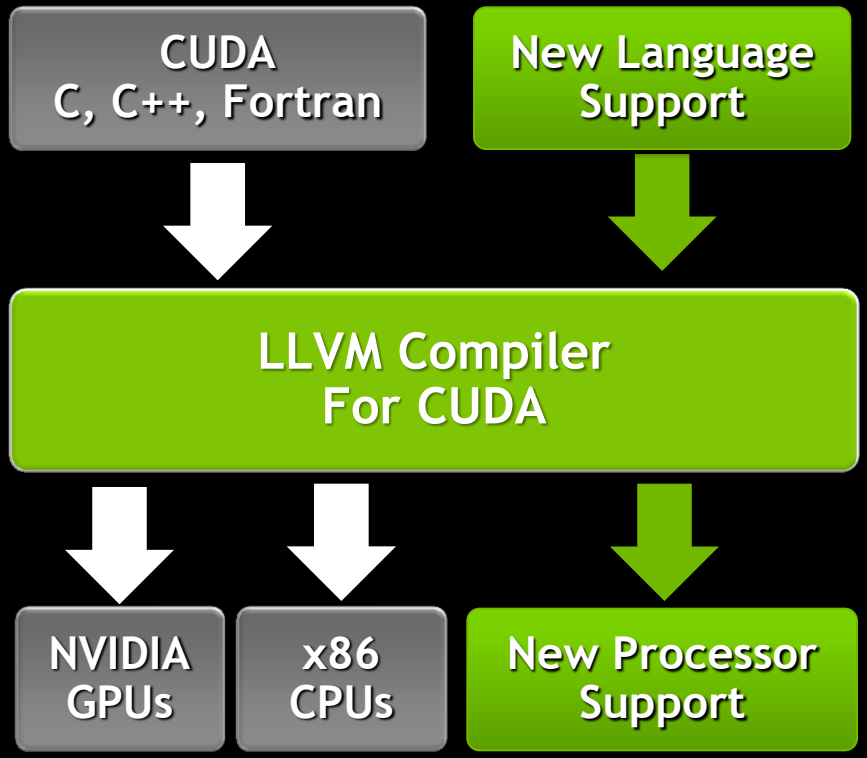
Enabling More Programming Languages

Developers want to build front-ends for

Python, Java, R, DSLs …

Target other processors like

ARM, FPGAs, GPUs, x86 …



Linux GCC Compiler to Support GPU Accelerators

Open Source

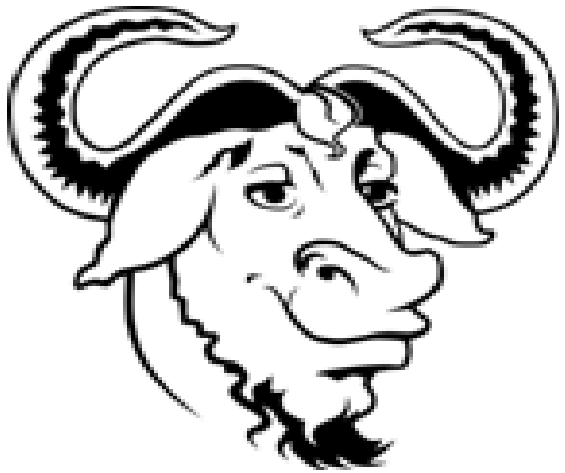
OpenACC in GCC by Mentor Graphics & Samsung

Pervasive Impact

Free to all Linux users

Mainstream

Most Widely Used HPC Compiler



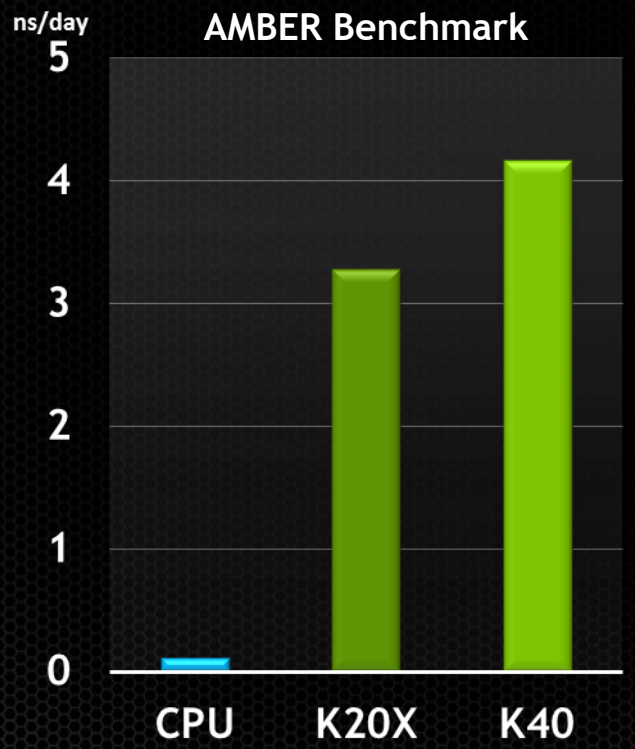
> *"Incorporating OpenACC into GCC is an excellent example of open source and open standards working together to make accelerated computing broadly accessible to all Linux developers."*
>
> Oscar Hernandez
> Oak Ridge National Laboratory

OAK RIDGE National Laboratory

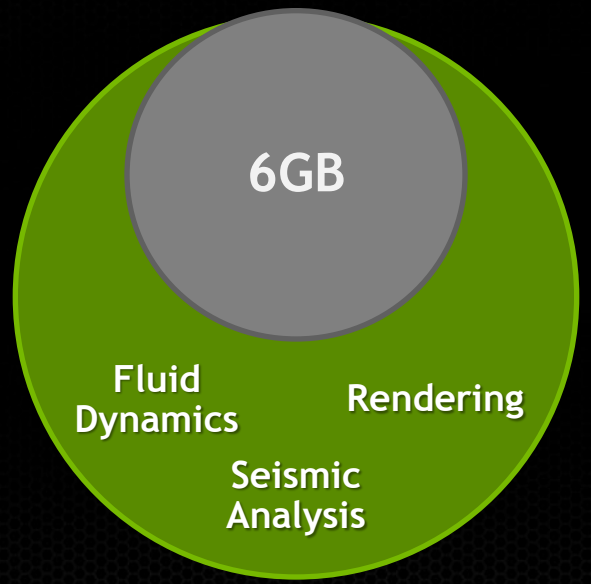
OpenACC Directives for Accelerators

Tesla K40
World's Fastest Accelerator

FASTER
1.4 TF| 2880 Cores | 288 GB/s



LARGER
2x Memory Enables More Apps



12GB

SMARTER
Unlock Extra Performance
Using Power Headroom

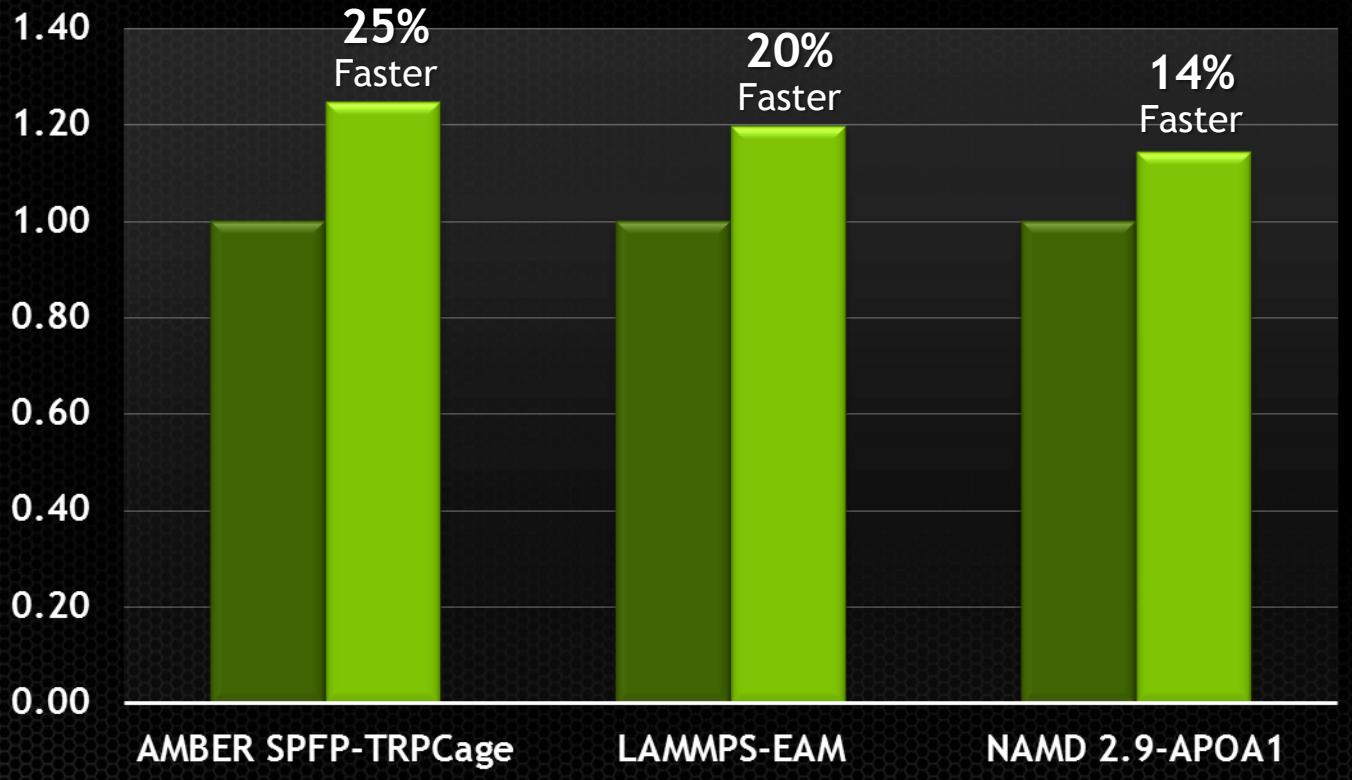
GPU Boost

AMBER Benchmark: SPFP-Nucleosome
CPU: Dual E5-2687W @ 3.10GHz, 64GB System Memory, CentOS 6.2, GPU systems: Single Tesla K20X or Single Tesla K40

GPU Boost

Up to 25% Extra Performance on Applications

Use Power Headroom to Run at Higher Clocks



Paradigm — 17% Faster

OpenEye — 13% Faster

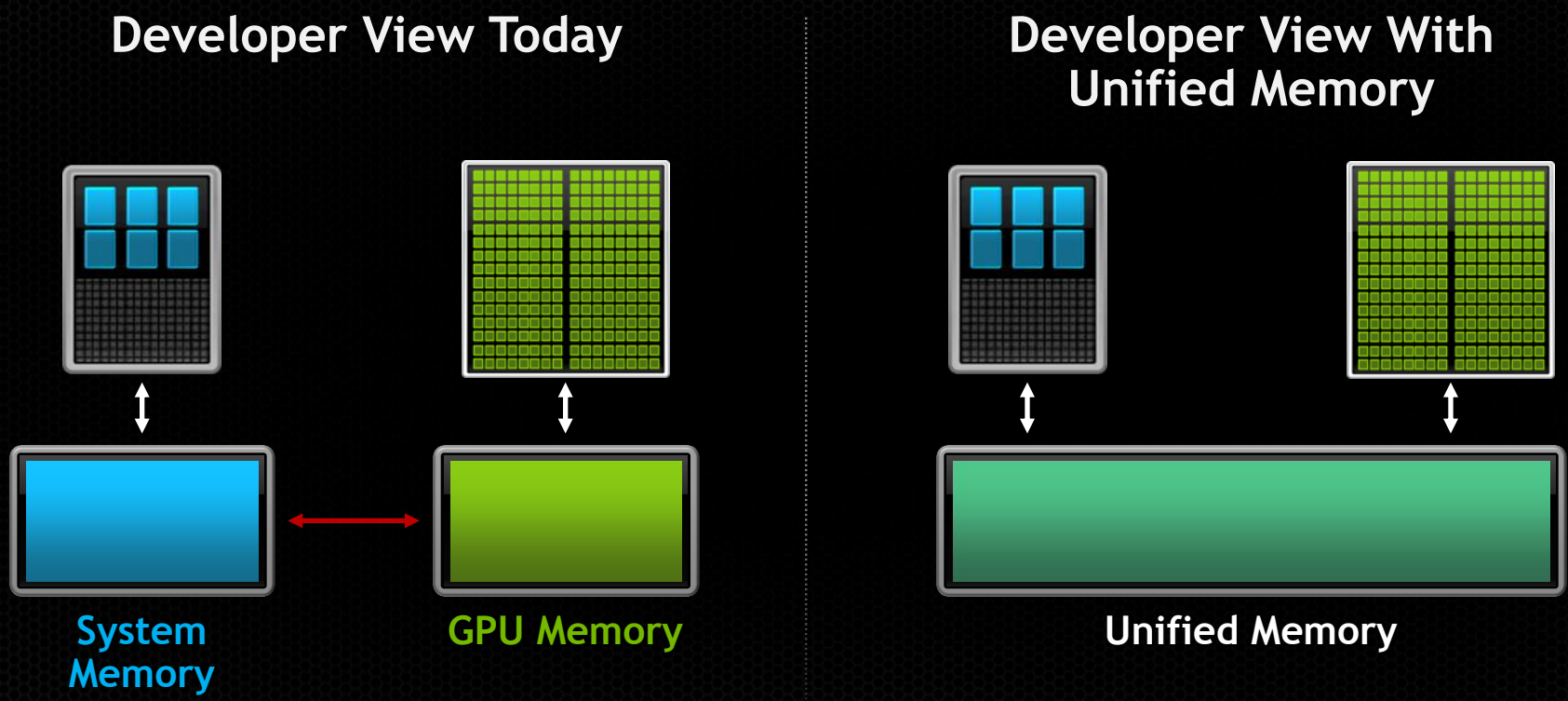
ANSYS — 11% Faster

ANNOUNCING

Unified Memory

CUDA 6

Unified Memory
Dramatically Lower Developer Effort



Super Simplified Memory Management Code

CPU Code

```
void sortfile(FILE *fp, int N) {
  char *data;
  data = (char *)malloc(N);

  fread(data, 1, N, fp);

  qsort(data, N, 1, compare);


  use_data(data);

  free(data);
}
```

CUDA 6 Code with Unified Memory

```
void sortfile(FILE *fp, int N) {
  char *data;
  cudaMallocManaged(&data, N);

  fread(data, 1, N, fp);

  qsort<<<...>>>(data,N,1,compare);
  cudaDeviceSynchronize();

  use_data(data);

  cudaFree(data);
}
```

New Drop-in NVBLAS Library

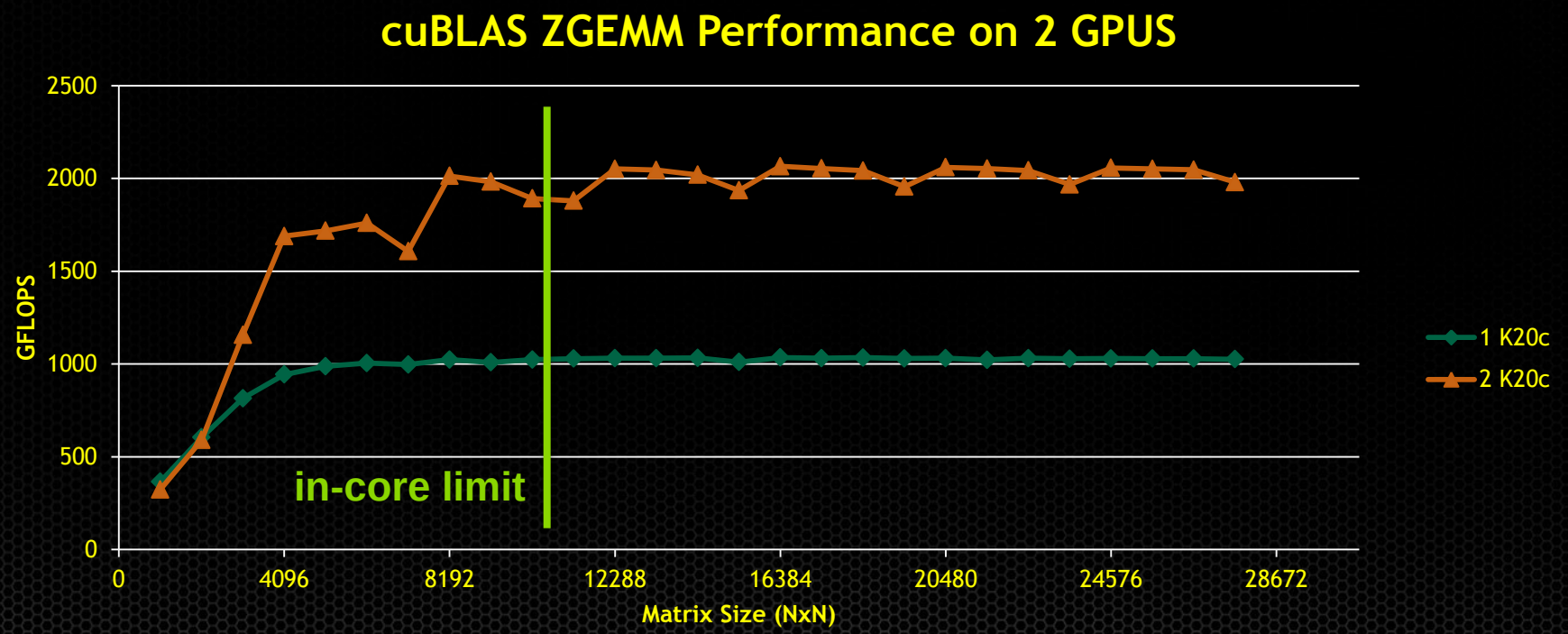
- Drop-in replacement of CPU-only BLAS
  - Automatically routes standard BLAS3 calls to cuBLAS
  - Optionally configure which routines and matrix sizes are accelerated
  - User provides CPU-only BLAS dynamic library location

- Simply re-link or change library load order

```
gcc myapp.c –lnvblas -lmkl_rt -o myapp - or -
env LD_PRELOAD=libnvblas.so myapp
```

Multi-GPU cuBLAS

- Single function call automatically spreads work across two GPUs
- Source and result data in system memory
- Supports matrices > size of memory (out-of-core)
- All BLAS Level-3 routines



Fastest Supercomputer In Europe

6.27 PetaFLOPS (80% Linpack Efficiency)

Piz Daint

Greenest Petascale System

3110 MFLOPS/W

#2: JUQUEEN: 2176 MFLOPS/W

Production-Grade Weather Forecasts: COSMO

7 National Weather Agencies

Germany | Greece | Italy | Poland | Russia | Romania | Switzerland

Greenest Supercomputer in the World

Tokyo Tech KFC System

4000+ MFLOPS per Watt

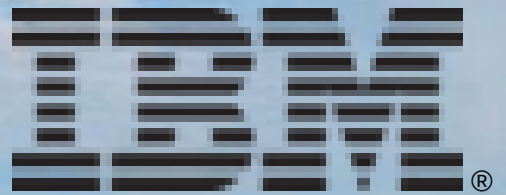
25% Higher than #1 Green500 System

160 Tesla K20X GPUs

Oil Immersion Technology

Current Green500 #1: CINECA Eurora System, Italy, 3208 MF/W

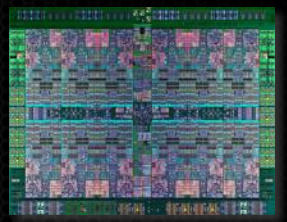
A new era of computation and analytics



IBM & NVIDIA Accelerating Computing

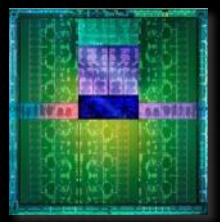
Next-Gen IBM Supercomputers and Enterprise Servers

Long term roadmap integration



POWER CPU

\+



Tesla GPU

OpenPOWER Consortium

Open ecosystem built on Power Architecture

NVIDIA

IBM

Google

Mellanox TECHNOLOGIES

TYAN

Enterprise & Data Analytics Software

Applications, Tools, Algorithms, Libraries, Languages, Compilers

Java

OpenMP

OpenACC Directives for Accelerators

GPU-Accelerated POWER-Based Systems Available in 2014

Dave Turek

Vice President, Advanced Computing, IBM

A new era of computation and analytics