DOING MORE WITH LESS: DEMOCRATIZING COMPUTING ACROSS SCIENTIFIC DOMAINS

Fernanda Foertter, GPU Developer Advocate Healthcare HPC + AI.

PLAY PC GAMES ON YOUR MAC

GeForce NOW™ brings your favorite PC games to your Mac at max settings, with smooth framerates—all powered by GeForce® GTX GPUs in the cloud. Get Game Ready drivers, cloud-sync saved games, and use express install to load games once, in seconds. Over 100 games are supported. Join the free beta today.

LEARN MORE

GEFORCE RTX™

GeForce RTX delivers the ultimate PC gaming experience. Powered by the new NVIDIA Turing™ GPU architecture.

LEARN MORE

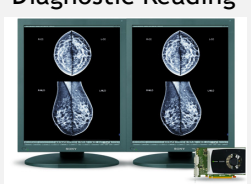
THIN, FOR THE WIN.

Max-Q is an innovative approach to crafting the world's thinnest, fastest, quietest gaming laptops.

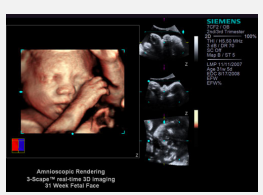
LEARN MORE

TEN YEARS OF NVIDIA IN HEALTHCARE

2007 — NVIDIA Standard Diagnostic Reading

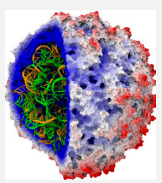


2009 — 1st 3D Stereo Ultrasound



2010 — GPU Computing Innovating CT & MRI

2012 — NGS Breaks $1000 Cost Using GPU Computing

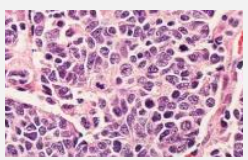


2012 — AlexNet beats expert code by huge margin using GPUs

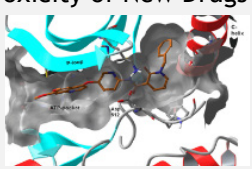
IMAGENET

2012 — NGS Breaks $1000 Cost Using GPU Computing

2014 — IDSIA: DL Detecting Mitosis in Breast Cancer

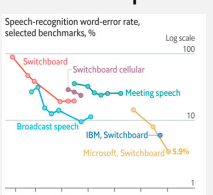


2014 — JKU: DL Predicting Toxicity of New Drugs

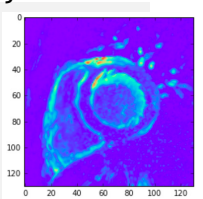


2014 — Google Outperforms Humans in ImageNet

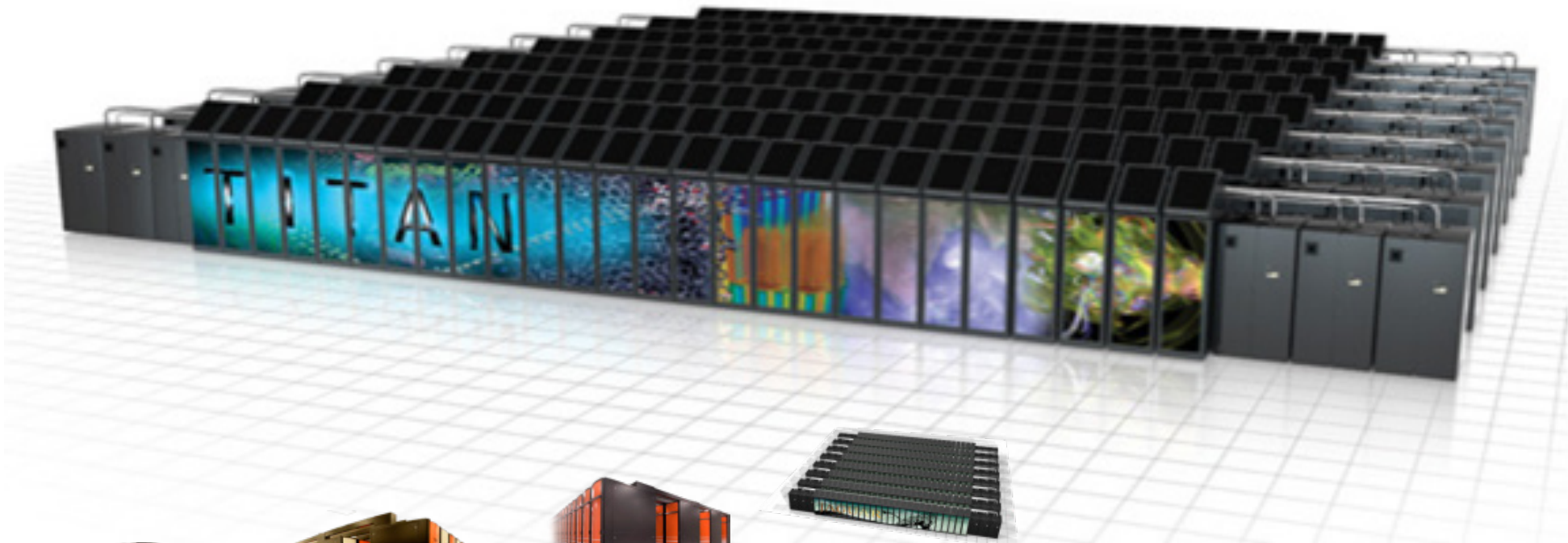
2016 — Microsoft Outperforms Humans in Speech

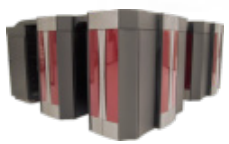


2016 — BAH & NIH Kaggle Breaks accuracy records on Ejection Fraction



USHERING A NEW-ISH ERA





Phoenix X1
- Doubled size
- X1e

2004

Jaguar XT3
- Dual core upgrade

2005

Jaguar XT4
- Quad core upgrade

2007

Jaguar XT5
- 6 core upgrade

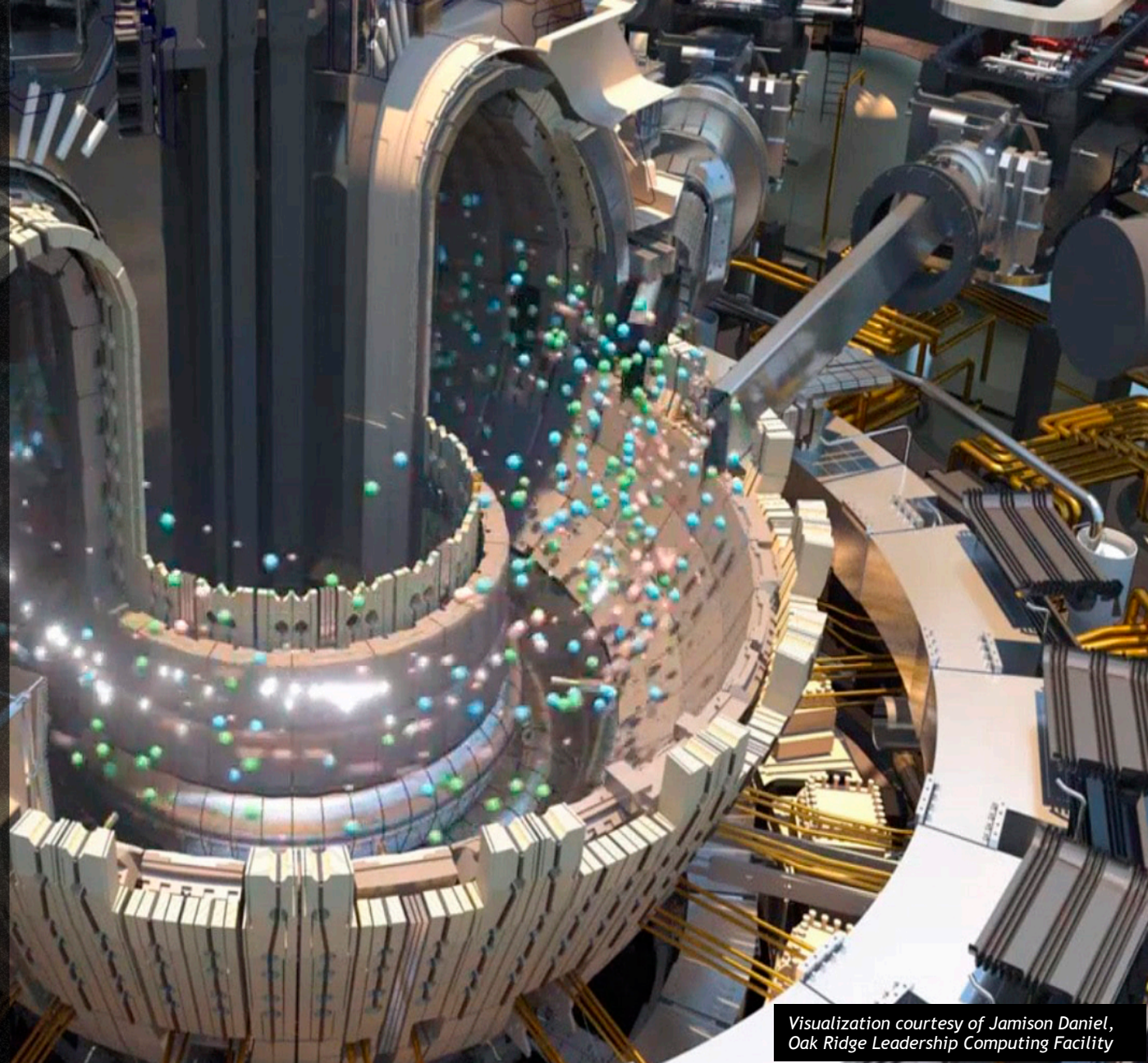
2008

AI IS SPEEDING THE PATH TO FUSION ENERGY

Fusion is the future of energy on Earth. But it's a highly sensitive process where even small environmental disruptions can stall reactions and damage multi-billion machines. Current models can predict the disruptions with 85% accuracy, but ITER will need something more precise.

Researchers at Princeton University have developed the Fusion Recurrent Neural Network (FRNN) using deep learning and NVIDIA GPUs with CUDA to predict disruptions and make adjustments to minimize damage and downtime. Even a 1% improvement in the prediction accuracy can be transformative considering the immense scale and cost of fusion science. FRNN has achieved 90% accuracy and is on the path to achieving its goal of 95% accuracy for ITER's tests.

PRINCETON UNIVERSITY



Visualization courtesy of Jamison Daniel, Oak Ridge Leadership Computing Facility

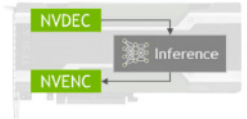
GPU ACCELERATED LIBRARIES

“Drop-in” Acceleration for Your Applications

DEEP LEARNING

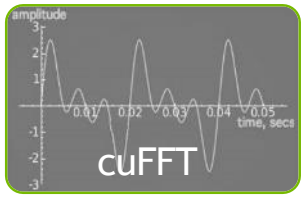
cuDNN


TensorRT



DeepStream SDK

SIGNAL, IMAGE & VIDEO



cuFFT

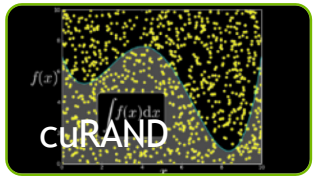

NVIDIA NPP


CODEC SDK

LINEAR ALGEBRA


cuBLAS

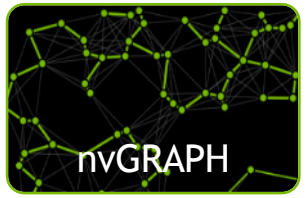

cuSPARSE

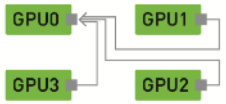

cuRAND


cuSOLVER


CUDA
Math library

PARALLEL ALGORITHMS

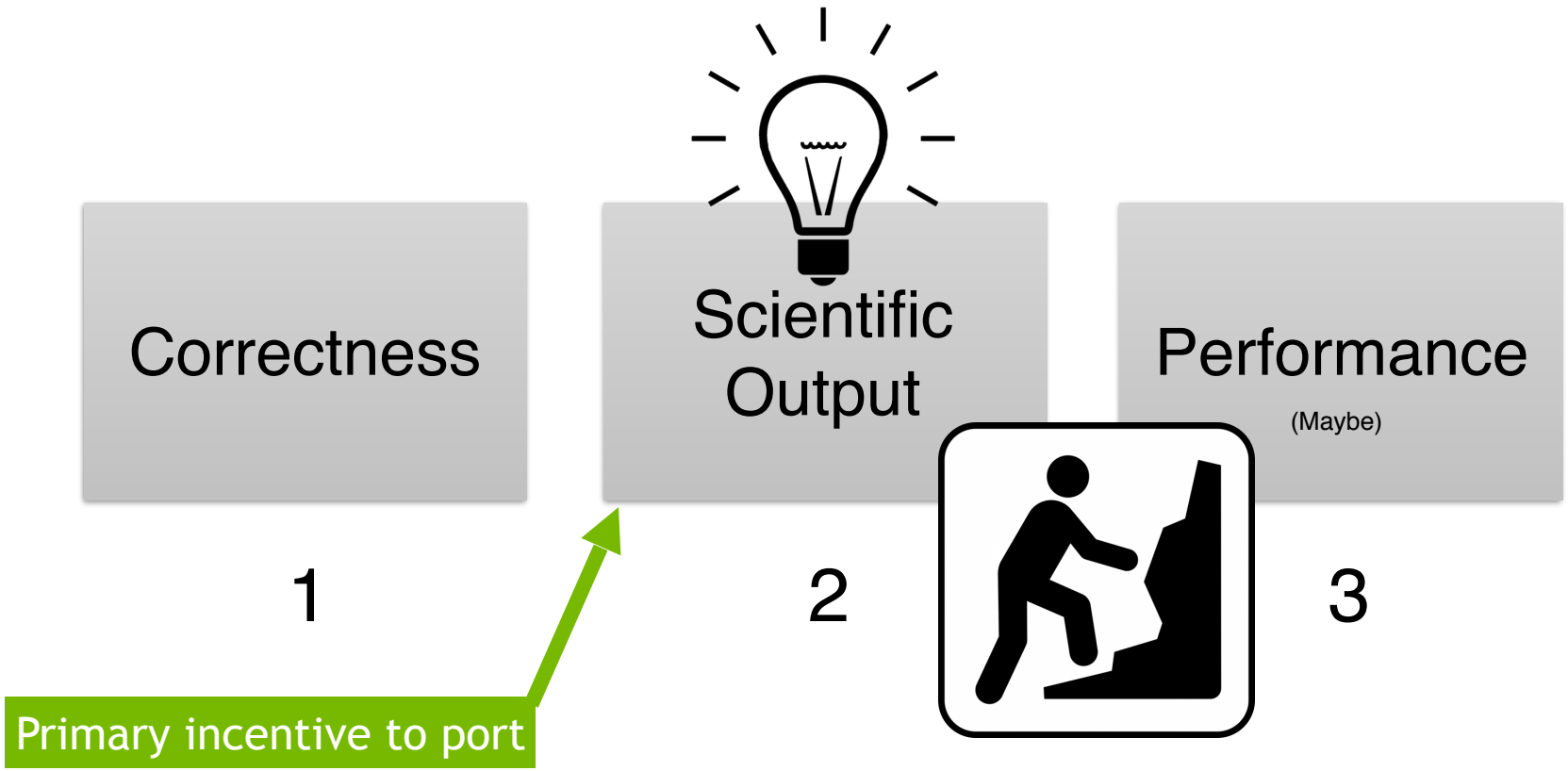

nvGRAPH



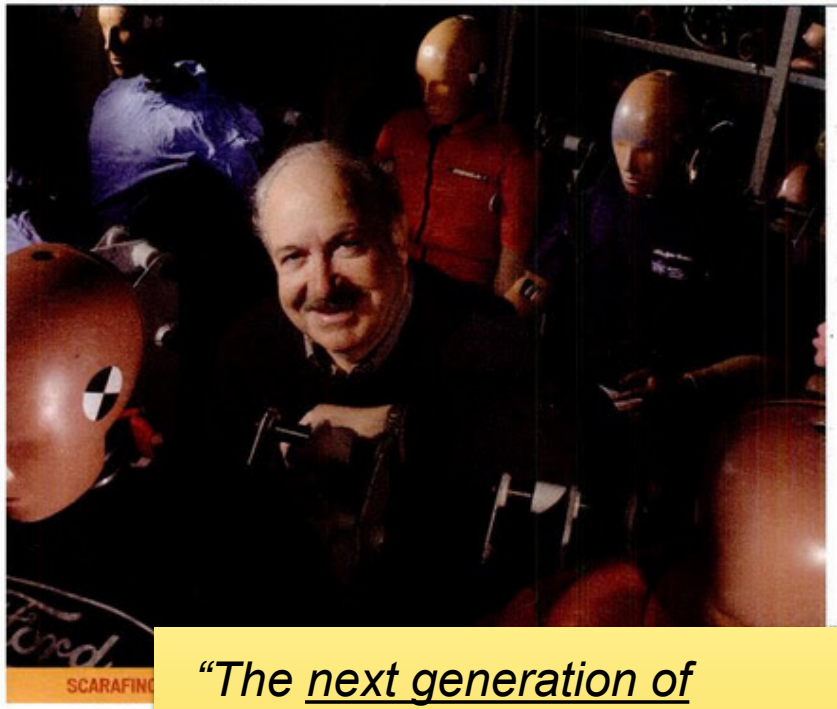
NCCL

LESSON: PERFORMANCE DOESN'T MATTER**

- Hypothesis: Performance is not the primary** concern of a domain scientist



2003



SCARAFINO

"The next generation of supercomputers will most likely be similar to the last generation of supercomputers built in the early to mid-90s…but significantly faster."

— Vincent Scarafino, Ford Motor Co

ment over the last vector supercomputer we made here in the mid-1990s, the Cray T-90. Japanese auto companies are formidable competitors. We don't need to hand them yet another advantage.

What should the federal government do to boost U.S. supercomputing technology? Fund high-end processor design and supporting system components. The goal would be ultrafast processors with memory and I/O systems well matched to the computational speeds.

The government used to do just that, sponsoring development of high-end supercomputer architectures like the Cray vector machines. But now it seems to favor huge clusters of commodity microprocessors. Yes, in the mid-1990s they said that microprocessors were getting faster and faster, and we just need to put a whole bunch of them together and we've got a supercomputer. Well, it doesn't work quite that way. Microprocessors are fast at computing, but in order to run real difficult problems, they have to have real fast access to memory and be able to do I/O quickly. And memory subsystems are extremely expensive.

If you look at the very large machines made up of off-the-shelf components, they get about 5% of their theoretical peak performance. But if you look at the Earth Simulator, you see numbers from the high 30s to mid-50s.

Are there some applications for which the commodity-based clusters of microprocessors are a good approach? They provide ex-

to actually compute what kind of damage is done to human organs — the brain or liver, for example. Today's analyses with test dummies are very crude. They find at a gross level whether that kind of crash is survivable. But [occupant injury analysis] takes much more computing power than is available now.

What else would you like to be able to do? Try to understand how exotic materials would work, well enough to understand if they'd work in vehicles. These composite materials are very strong, but understanding how they would react in a failure mode is a difficult problem to solve with today's computers.

What will the next generation of supercomputers look like? The next generation of supercomputers will most likely be similar to the last generation of supercomputers built in the early to mid-1990s. But they will be significantly faster and able to execute difficult algorithms at speeds much closer to theoretical peak rates than commodity-based machines are able to do.

Will there be any breakthroughs in software over the next five years? There has been significant progress in the area of parallel processing during the last eight years. I would expect continued evolution. I am not aware of any specific areas that seem ripe for breakthroughs, but these things are difficult to predict. Software cannot substitute for raw processing speed.

Title
nun
con
Mot

Obs
cen
per
the
Simulator for climate

best machines? Advanced su-

Change is hard

1991

Thinking Machines thinks big

Speed enters new dimension with massively parallel supercomputer

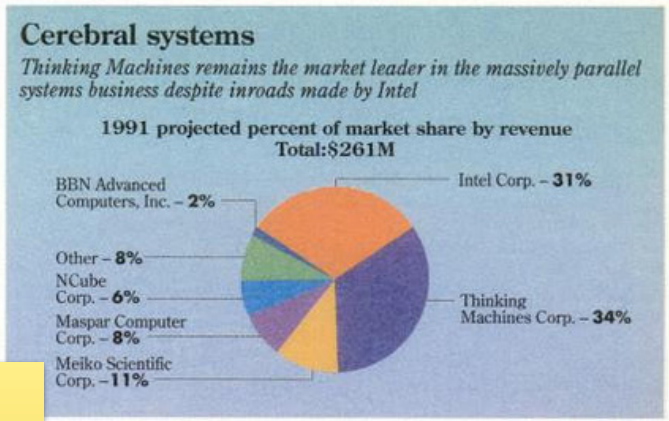
BY MICHAEL ALEXANDER
CW STAFF

CAMBRIDGE, Mass. — Thinking Machines Corp. introduced a radically new massively parallel processing supercomputer last week with a peak performance

The building block of Thinking Machines' CM-5 Connection Machine supercomputer is Sun's reduced instruction set computing (RISC) chip-based Scalable Processor Architecture (Sparc) microprocessor. "We used a Sparc processor because it cur-

hindered the acceptance of massively parallel processing machines in the business world, Hillis claimed.

The parallel processing industry has been divided into two camps on the issue of whether all processors should run the same



"The firm also announced that it has inked a pact with IBM and Sun Microsystems to develop a programming standard that will allow the same application written in Fortran to run unmodified on workstations, mainframes and supercomputers. The joint effort to develop common software standard helped clinch the sale…The firm had seriously considered Intel Corp until learning of the pact"— Steve Cone, Senior VP Amex

ve Cone, senior vice presi-
t of direct marketing at
erican Express Travel Relat-
Services Co. The company
"seriously considered" an
l Corp. parallel processing
hine until learning of the
t, Cone said.

rking together

're pleased IBM and Think-
Machines are working to-
er because that will allow us
move applications from our
mainframes to the Connec-
Machines," Cone said.

The CM-5 computers will be used to enhance customer service by speeding the collection of billing data for card members and merchants, Cone added. He declined to elaborate further.

The largest machine on order is a 1,024-node CM-5 that is being built at a cost of $25 million for the Los Alamos National Laboratory in Los Alamos, N.M. Officials at Schlumberger Ltd. as well as eight federal goverment and university research centers also announced plans to acquire the new machines.

HACKATHONS

Five Days + **Two** Mentors + **HPC** App + **Your** Team

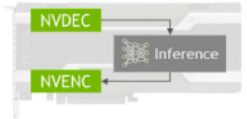
GPU ACCELERATED LIBRARIES

“Drop-in” Acceleration for Your Applications

DEEP LEARNING

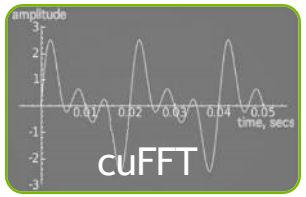
cuDNN


TensorRT



DeepStream SDK

SIGNAL, IMAGE & VIDEO



cuFFT

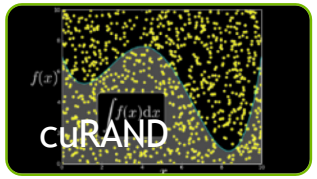

NVIDIA NPP


CODEC SDK

LINEAR ALGEBRA


cuBLAS

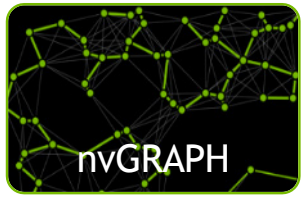

cuSPARSE

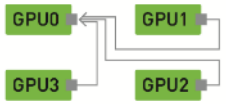

cuRAND


cuSOLVER


CUDA
Math library

PARALLEL ALGORITHMS


nvGRAPH



NCCL

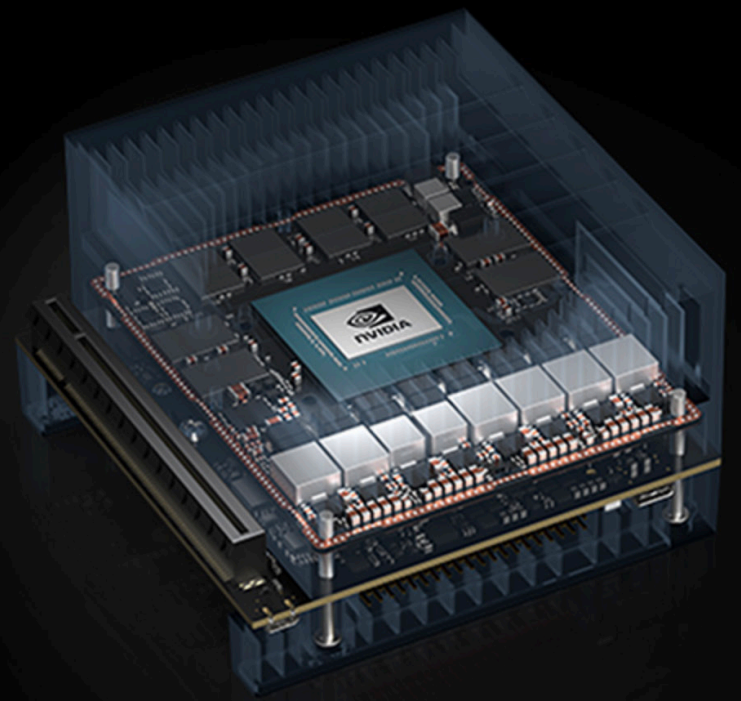
NVIDIA

AUTONOMOUS MACHINES INDUSTRIES PRODUCTS FOR DEVELOPERS

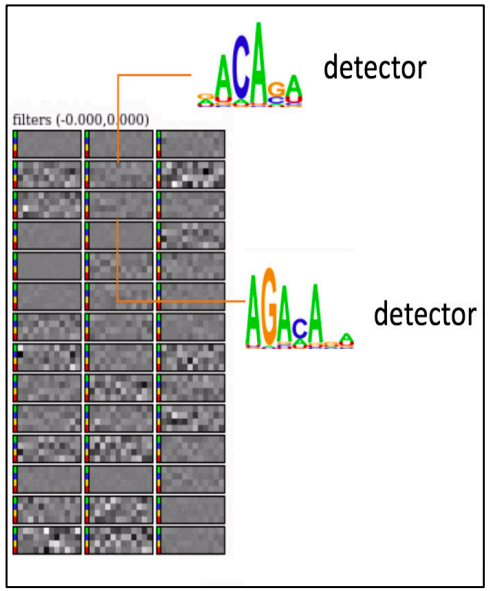
NVIDIA® JETSON AGX XAVIER™ DEVELOPER KIT

Powering AI in Autonomous Machines

BUY NOW

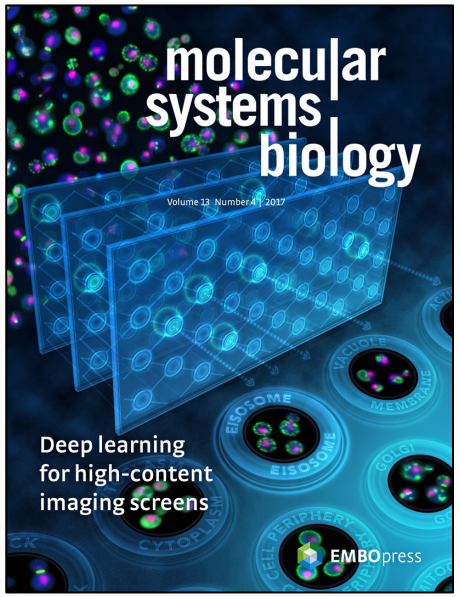


DEEP LEARNING IN DRUG DEVELOPMENT



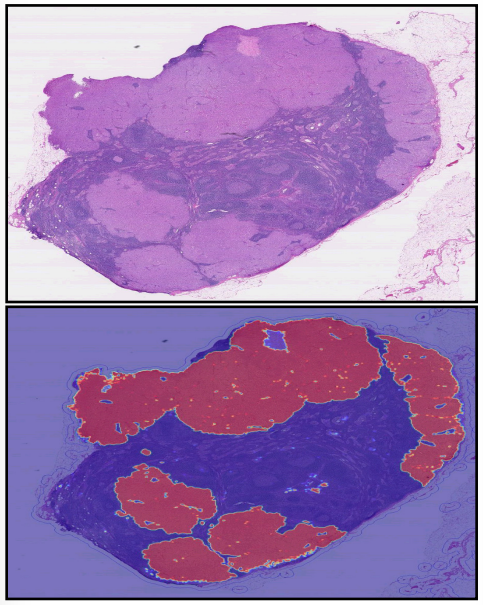
deep genomics

DISCOVERY
VARIANT CALLING



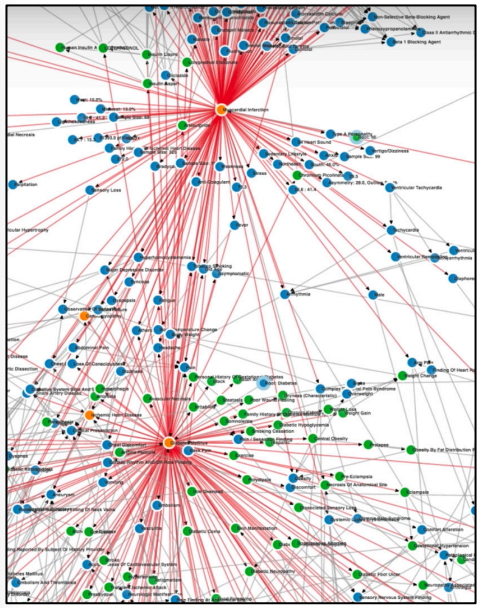
PHENOMIC AI
BIOIMAGE INFORMATICS

DISCOVERY
HIGH CONTENT SCREENING



PathAI

PRE & CLINICAL
PATHOLOGY

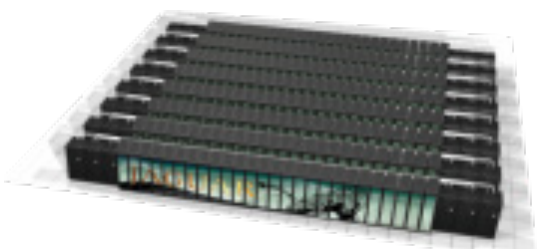


CLOUDMEDX®

REAL WORLD EVIDENCE
DATA ANALYTICS

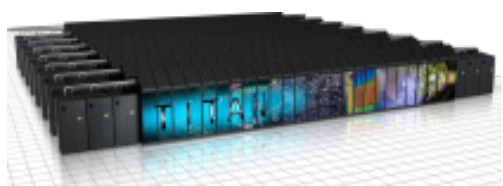
PATH OF FUTURE ARCHITECTURES SHOWS INCREASING PARALLELISM

- Hierarchical parallelism
- Hierarchical data spaces



Jaguar XT5
- 6 core upgrade

2008

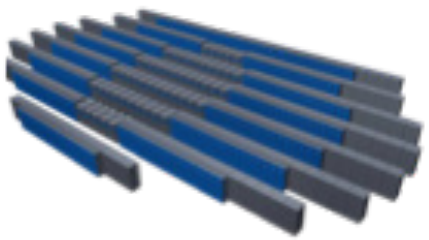


Titan XK7
- 16 cores
- GPU upgrade

2012

Summit
- Hybrid Accelerated

2017



OLCF5
- ? ? ?

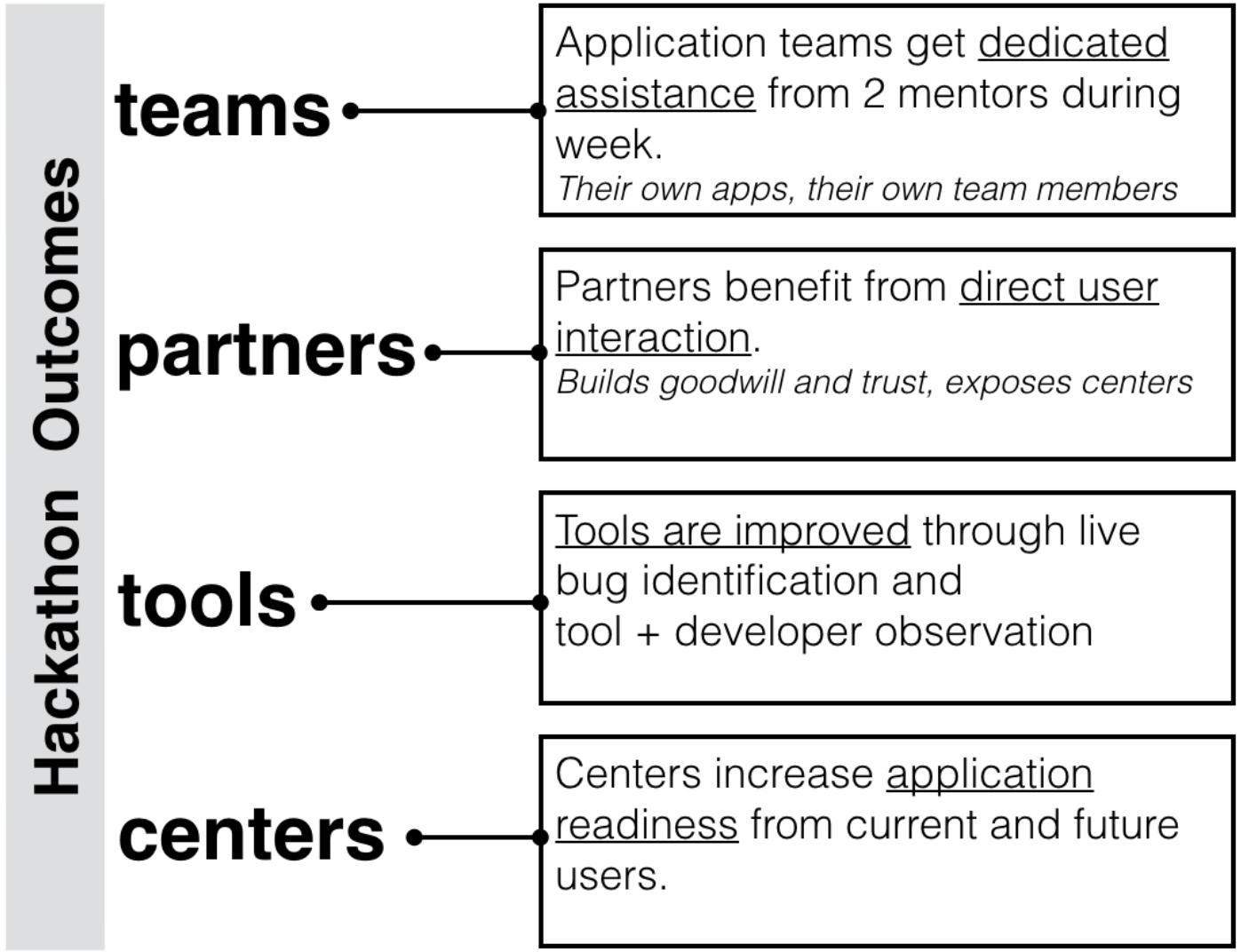
2022

A 21st CENTURY PLANNING TOOL BUILT ON AI

With the Earth's population at 7 billion and growing, understanding population distribution is essential to meeting societal needs for infrastructure, resources and vital services. Using GPUs with CUDA and deep learning, Oak Ridge National Laboratory can quickly process high-resolution satellite imagery to map human settlements and changing urban dynamics. With the ability to process a major city in minutes, ORNL can provide emergency response teams critical information that used to take days to create.

OAK RIDGE National Laboratory

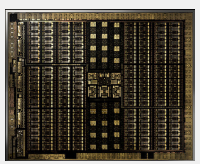
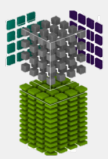
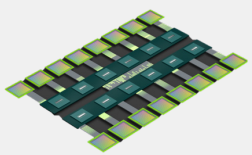
COLLABORATION IS KEY



CUDA TOOLKIT 10.0

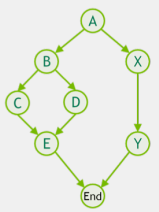
TURING AND NEW SYSTEMS

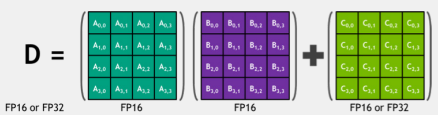
New GPU Architecture, Tensor Cores, NVSwitch Fabric


CUDA PLATFORM

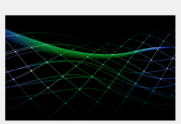
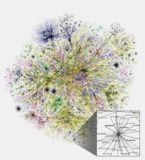
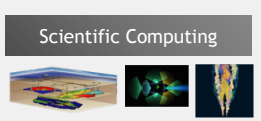
CUDA Graphs, Vulkan & DX12 Interop, Warp Matrix





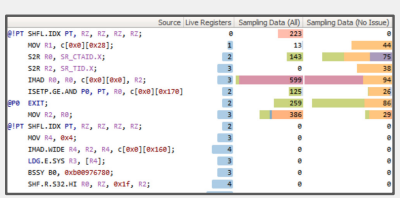
LIBRARIES

GPU-accelerated hybrid JPEG decoding,
Symmetric Eigenvalue Solvers, FFT Scaling

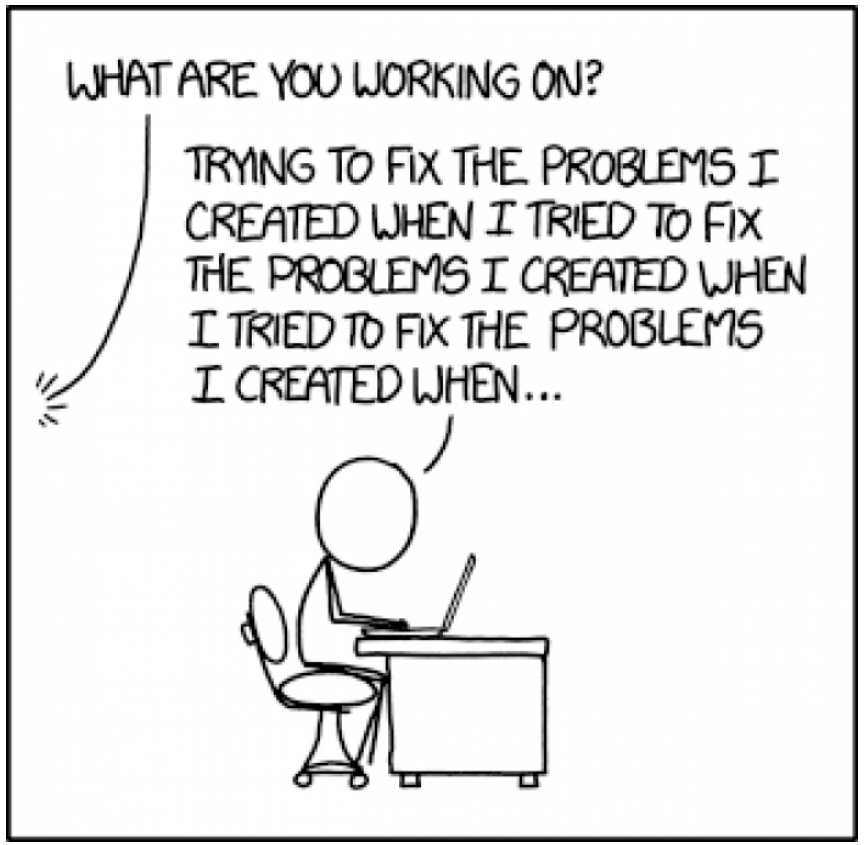
DEVELOPER TOOLS

New Nsight Products – Nsight Systems and Nsight Compute


BACKUP

CREATE PLAYGROUNDS



Try many ideas and APIs; avoid making costly “singular roadmap” decisions

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov