WHAT'S GOING ON IN HIGH PERFROMANCE COMPUTING

EP WORKSHOP

Dan Stanzione
Executive Director, TACC
Associate Vice President for Research, UT-Austin

Summer School Seminar

June 2022

A QUICK OUTLINE

- A little background about TACC (Which you will see later!).
- What are the big trends in high performance computing?
- What does that mean for how you research and code?
- What am I going to learn from you?

TACC - 2021

LEADERSHIP-CLASS
COMPUTING FACILITY

TACC
TEXAS ADVANCED COMPUTING CENTER

TACC – 2021

- About 170 professionals focused on delivering *RESEARCH* computing and data services.
- Roughly $150M of computing and storage hardware
- 10MW datacenter(s), roughly $100M in physical plant investment.
  - ~80% supported by NSF, other fed agencies.
  - ~10% UT-Austin

TACC FUNDING

- In the next couple of years, we will approach $2B in external funding over our existence.
- NSF Commitment to us as a large scale facility through at least 2036.
- Tens of millions in faculty funding supported each year (maybe up to 25% of all NSF-funding *in Texas*).

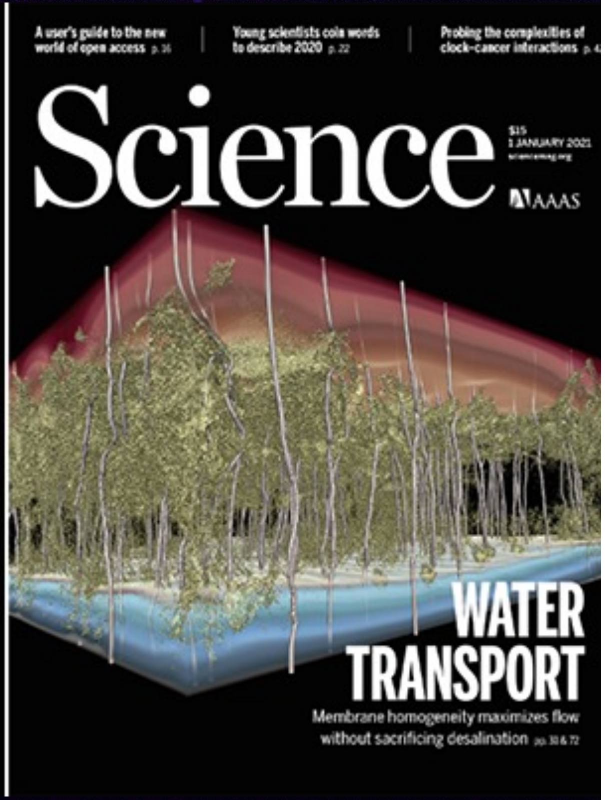
WHAT WE DO

- Provide researchers with:
  - Computing, Data, AI , Software capabilities to support their research
  - The expert help to be able to use it!
  - In the ways they want to consume it
  - Help with grants/strategy
- Computation, AI, Data almost ubiquitous across the sciences.

DESPITE COVID, ANOTHER FANTASTIC YEAR

- Almost *TEN MILLION* jobs delivered to almost 50k people.
- About SEVEN BILLION core hours delivered.
- Lonestar-5 decommissioned, Lonestar-6 delivered.
- Upgrades to Frontera, coming upgrades to Stampede2
- Corral and Stockyard upgraded/replaced
- TACC contributed to the winning Gordon Bell Prize
- Lots of great science. . .

IN 2021



A user's guide to the new world of open access p. 36 | Young scientists coin words to describe 2020 p. 22 | Probing the complexities of clock-cancer interactions p. 4

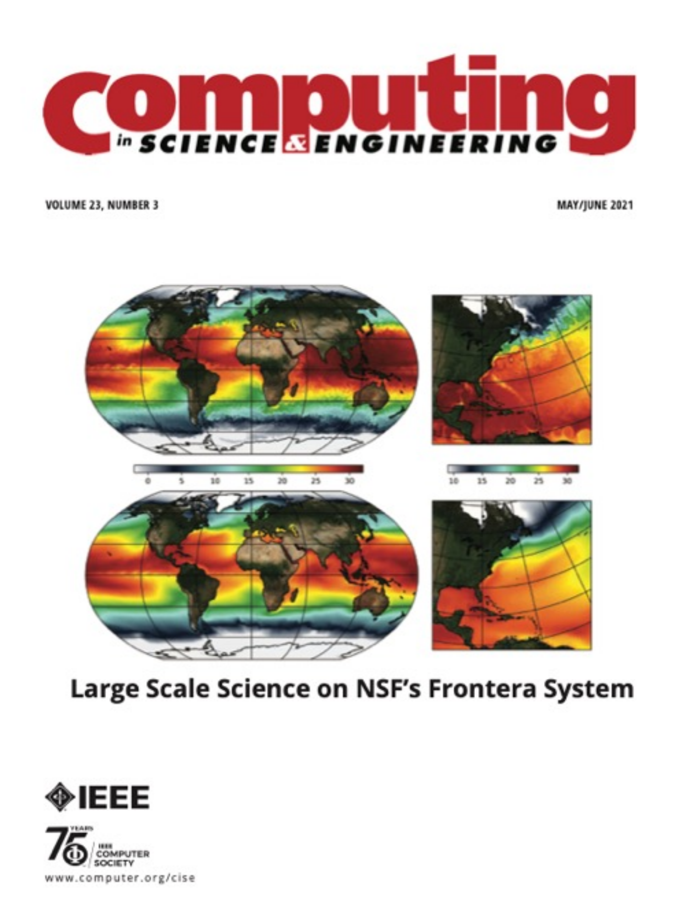
Science

$15
1 JANUARY 2021
sciencemag.org

AAAS

WATER TRANSPORT

Membrane homogeneity maximizes flow without sacrificing desalination pp. 30 & 72



computing in SCIENCE & ENGINEERING

VOLUME 23, NUMBER 3

MAY/JUNE 2021

Large Scale Science on NSF's Frontera System

IEEE

75 YEARS IEEE COMPUTER SOCIETY

www.computer.org/cise

NASA returns to the Moon with a robot armada p. 1188 | Serotonin reduces the risk for cocaine addiction pp. 1197 & 1252 | Strong, transparent, nacre-inspired glass composite p. 1229

Science

$15
10 SEPTEMBER 2021
science.org

AAAS

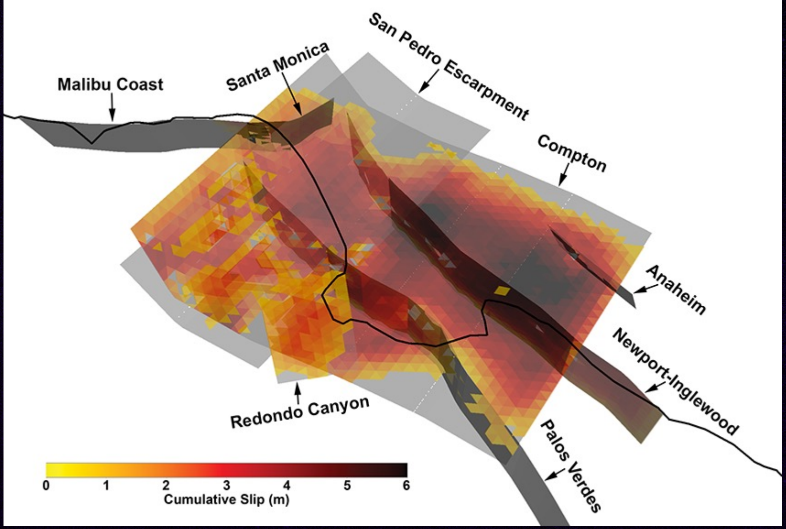
DOUSING THE STRATOSPHERE

Hydraulic jumps above supercell storms pp. 1194 & 1248

SIMULATING 800,000 YEARS OF CALIFORNIA EARTHQUAKE HISTORY TO PINPOINT RISKS

KEVIN MILNER, SCEC – BRUCE SHAW, COLUMBIA

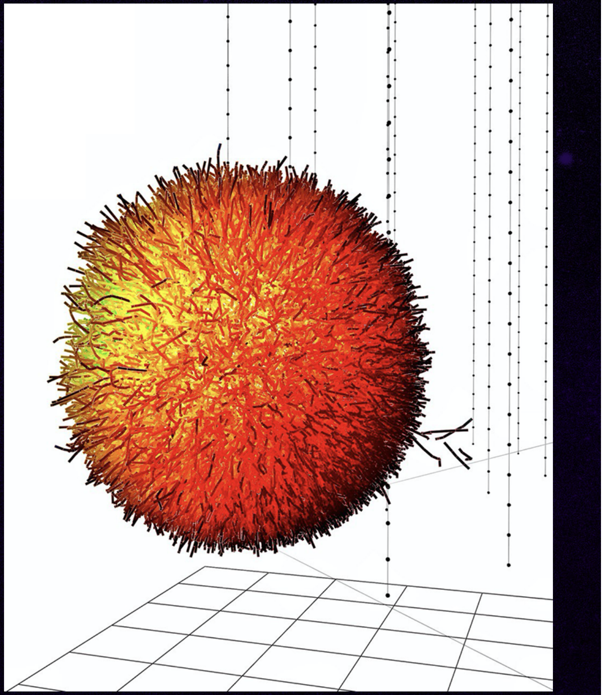
- Rate-State earthquake simulator, coupled to Cybershake
- improves the ability to pinpoint how big an earthquake might occur in a given location,
- *Bulletin of the Seismological Society of America* in January 2021



"We've made a lot of progress on Frontera in determining what kind of earthquakes we can expect, on which fault, and how often," said Christine Goulet, Executive Director for Applied Science at SCEC, also involved in the work. "We don't prescribe or tell the code when the earthquakes are going to happen. We launch a simulation of hundreds of thousands of years, and just let the code transfer the stress from one fault to another."

TRACKING COSMIC GHOSTS
BENEDIKT REIDEL, ICE CUBE

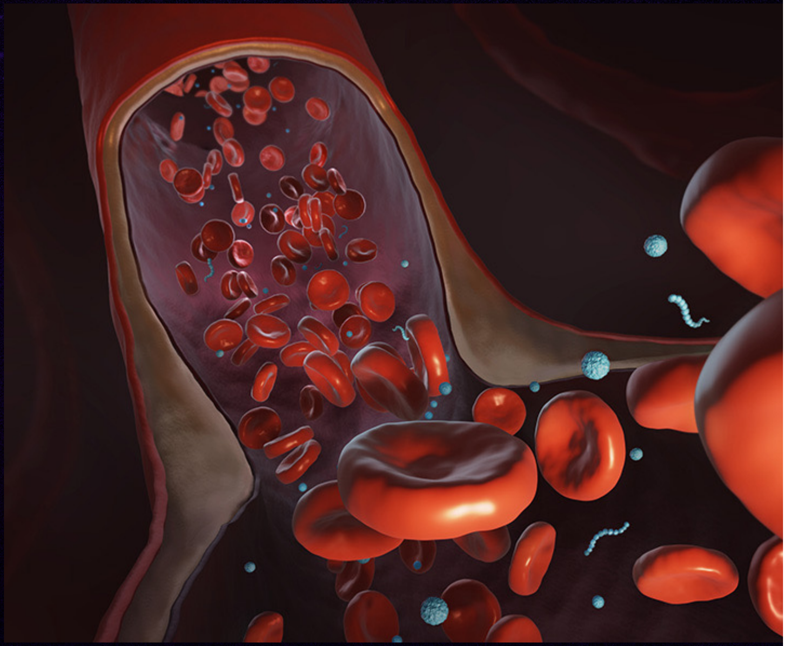
- Ice Cube – a one cubic kilometer block of ice converted into a neutrino observatory.
- “IceCube reveals a slice of Universe we haven't yet observed."
- On March 10, 2021, IceCube announced the detection of a Glashow resonance event, a phenomenon predicted by Nobel laureate physicist Sheldon Glashow in 1960.
  - suggests the presence of electron antineutrinos in the astrophysical flux, while also providing further validation of the standard model of particle physics
- Awarded through LCSP track – mostly ensemble Monte-Carlo simulation run through OSG.
- *Nature* [https://dx.doi.org/10.1038/s41586-021-03256-1]



TARGETING TUMORS WITH NANOWORMS

YING LI, UCONN

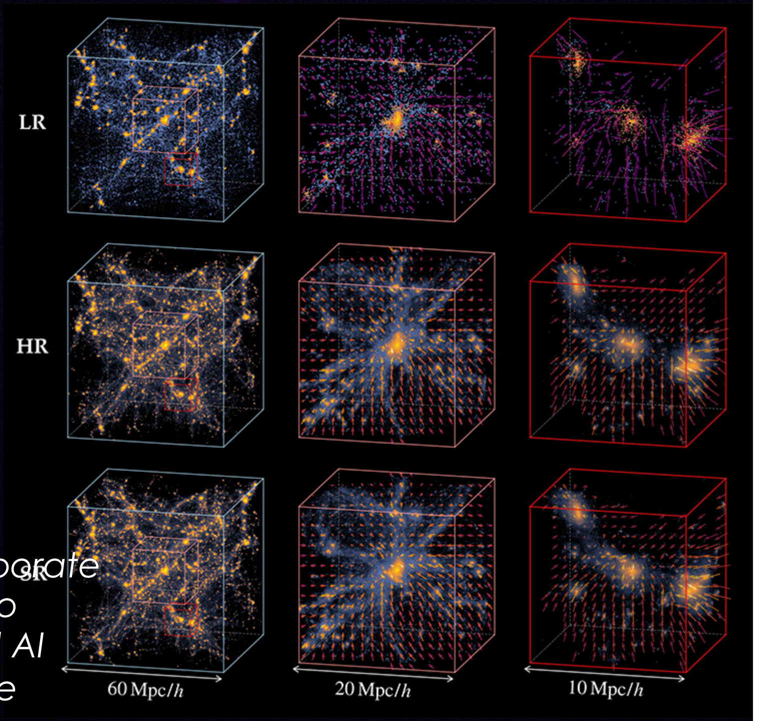
- "My research is centered on how to build high-fidelity, high-performance computing platforms to understand the complicated behaviors of these materials and the biological systems down to the nanoscale,"
- Nanoworms are long, thin, engineered encapsulations of drug contents.
- Modeled how these structures move in blood vessels of different geometries mimicking the constricted microvasculature.
  - Nanoworms can travel more efficiently through the bloodstream, passing through blockages where spherical or flat shapes get stuck.
  - Can use magnetic fields to influence flow.
- Can increase percentage of (highly toxic) drugs delivered directly to tumor.
- Published in *Soft Matter,* 2021.



ARTIFICIAL INTELLIGENCE JUST REMOVED ONE OF THE BIGGEST ROADBLOCKS IN ASTROPHYSICS

TIZIANA DI MATTEO, CMU AND YIN LI, FLATIRON

- Use a Neural net trained on low-res and high-res images of sections of the galaxy.
- Feed the net low-res images to upscale.
- 500x the computational efficiency.
- Published in PNAS, May 2021.

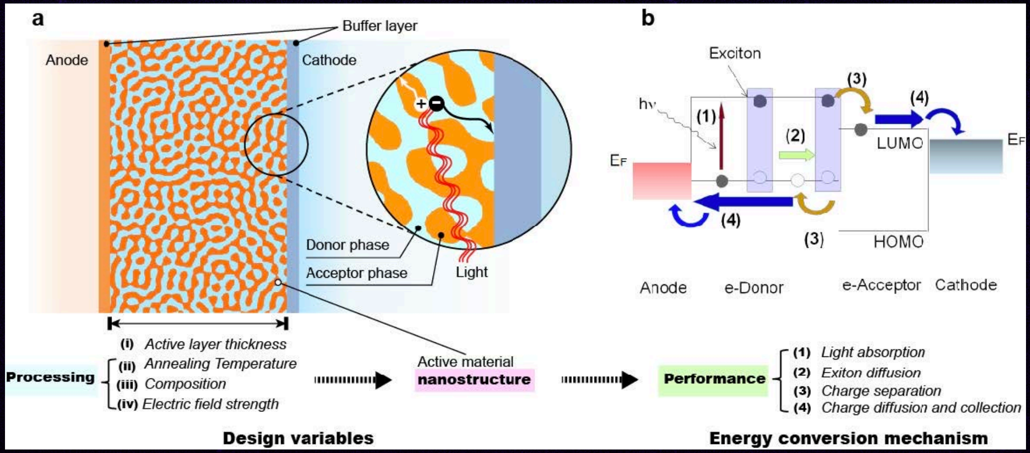


"Our goal is to create models of the entire observable Universe that incorporate information from higher resolution models of individual galaxies," Di Matteo continued. "Frontera is ideal for this: allowing us to couple the physics and AI running on GPUs and CPUs, and enable us to reach detail which would be otherwise impossible."

PHYSICS-INFORMED MACHINE LEARNING FOR SOLAR CELL PRODUCTION

PI: GANESH BALASUBRAMANIAN, LEHIGH UNIVERSITY

- Increase efficiency of organic solar cells (currently ~15%)
- Combines coarse-grained simulation — using approximate molecular models that represent the organic materials — and machine learning.
- Used Support Vector Machines, reduced required computing by 40%.
- Virtual experiments included varying temperature, annealing time, and ratio of donor/receptor molecules in heterojunctions.
- Published in CISE, May 2021, and Computational Materials Science, Feb 2021.



OUR MISSION

- Mission: To enable discoveries that advance science and society through the application of advanced computing technologies.

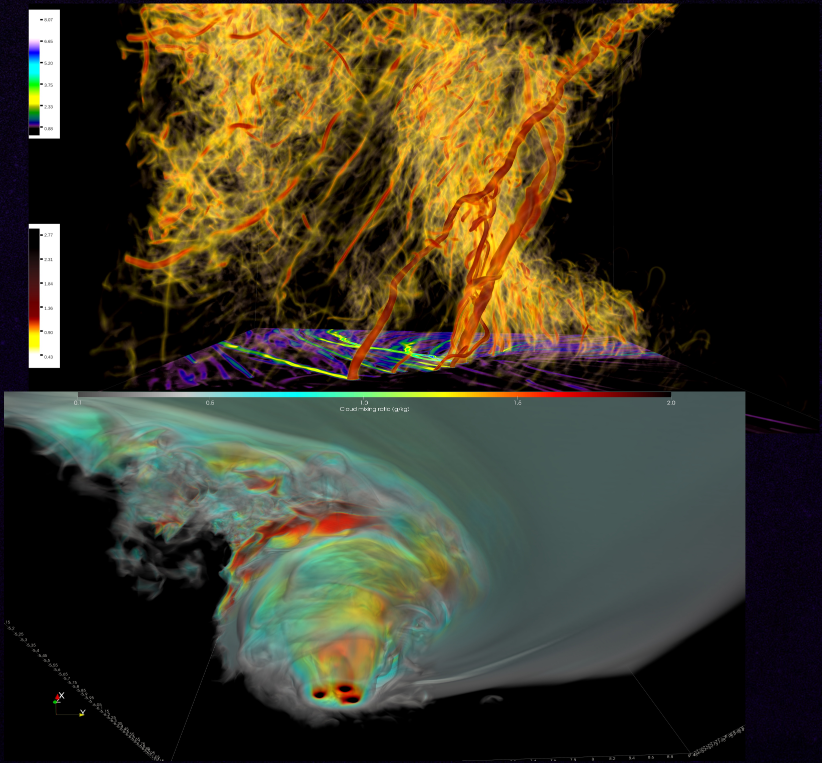


Figure: "World's Most Detailed Tornado Simulation", Leigh Orf, Wisconsin – used more than 200,000 cores on Frontera

SYSTEMS UPDATES

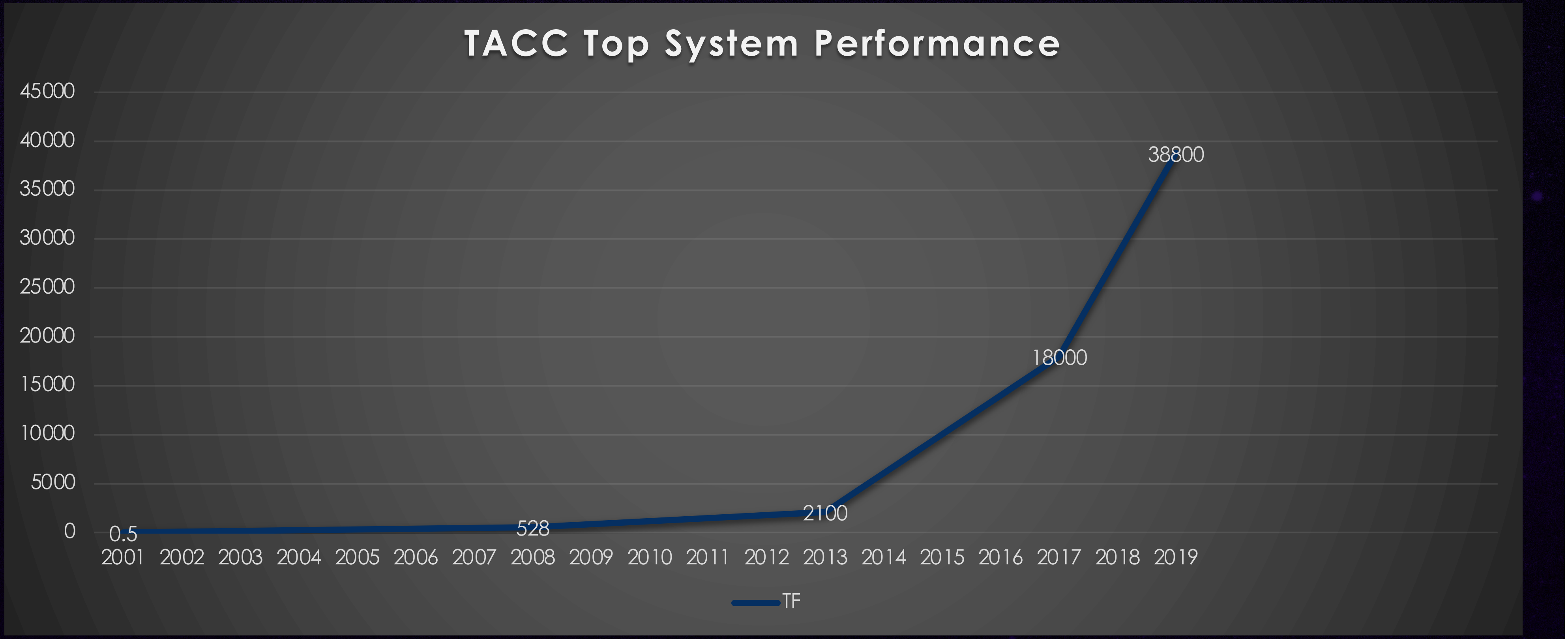
A QUICK REMINDER ON OUR CURRENT MAJOR SYSTEMS

- Frontera, NSF Capability System, 2019-2025 (Currently #12)
- Stampede2, NSF Capacity System, 2017-2023 (Currently #27)
- **Lonestar-6, Texas/Local System 2022-2027**
- **Longhorn – AI/DL GPU System, 2019-2025**
- Jetstream2 - NSF “Cloud” System 2022-2027
- Chameleon – NSF CS Testbed 2015-2024 (multiple HW upgrades)
- Corral, Ranch, Stockyard – Storage Platforms
- *Aggregate: ~75PF, ~16,000 compute nodes, ~350PB*

SYSTEMS UPDATE – LONESTAR-6

- Financial contribution from UTRC
  - Also Oden, CSR, Texas A&M, Texas Tech, UNT, UT-System
- Began early operations in October, full production Jan. 3rd.
- Oil immersion cooling for most of the nodes.
- ~600 AMD “Milan” nodes – 128 cores, 256GB of RAM.
- 32 A100 GPUs (may grow a bit)
- Giga-IO node disaggregation switches (experimental - GPU and NVDIMM)
- Scratch filesystem switched from Lustre to BeeGFS
- OS Switched from Redhat → Rocky Linux
- Experimenting with virtual “small” queue

TACC PERFORMANCE OVER TIME



Our Oversubscription rate is still ~5x, Despite the Growth

TH

(SELECTED) TRENDS IN HPC

- Trends in Technology
- Trends in how technology is used

TRENDS IN HPC - TECHNOLOGY

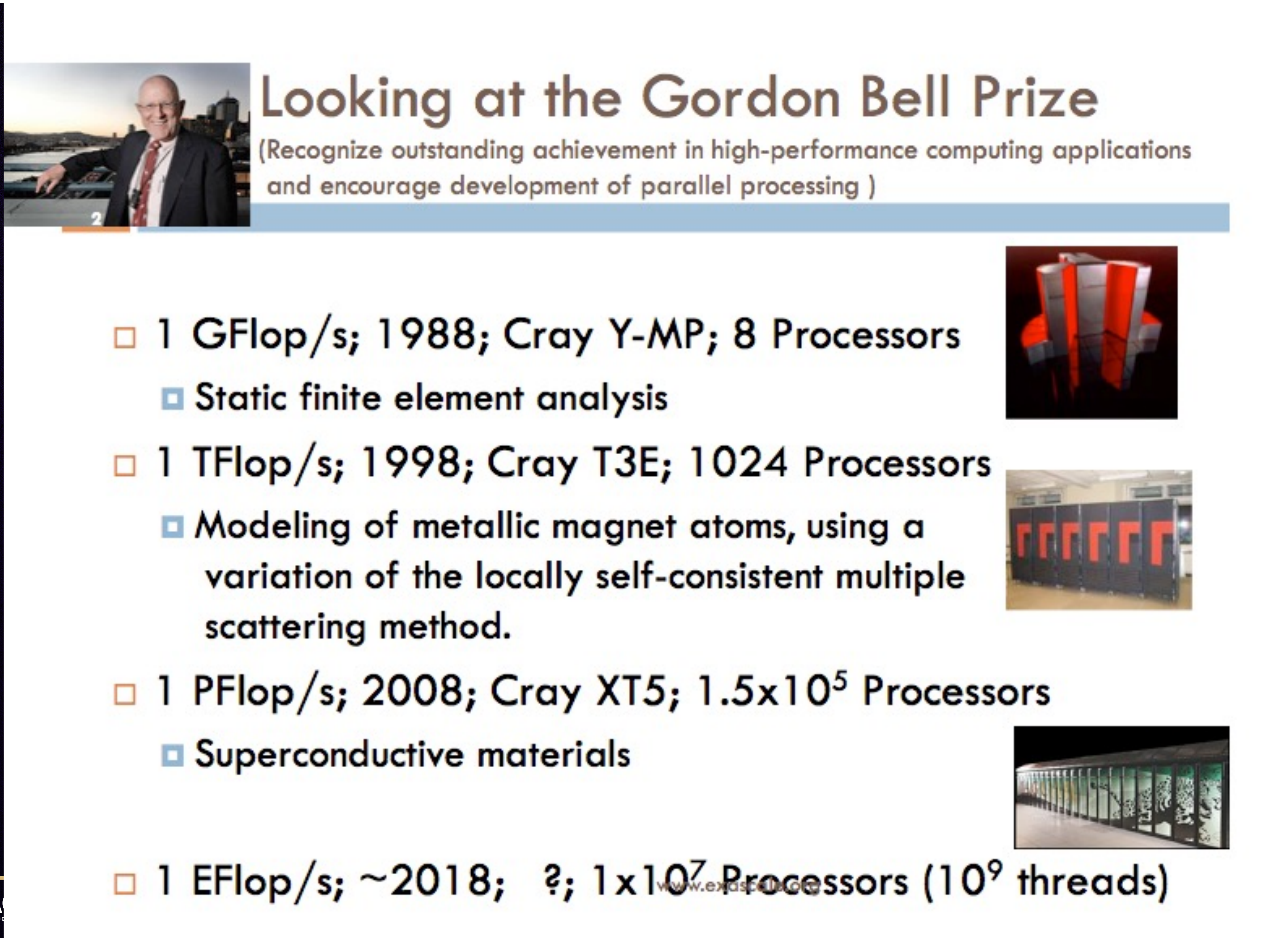
- Power
- Parallelism
- GPU
- End of X86 Hegemony
- POSIX is Dead/Long Live POSIX!

POWER

- In 2011, the goal for exascale was 15MW
- In 2014, the goal for exascale was 20MW
- In 2022, exascale was delivered at >25MW.
- Livermore, Argonne, Oak Ridge expanding to 85MW+
- Chips:
  - In 2015, CPUs would never exceed 130W (per Intel, AMD), GPUs would max out at 300W
  - In current roadmaps, that is 500W and 800W, respectively.
    - One reason GPUs are faster is they have twice as many transistors and twice the power… not much "architectural" about that (it's like claiming bigger buildings have more space).
- Sustainability is a big concern, not to mention operating costs.
  - E.g., Europe wants warm water cooling – that creates a max power per square cm of package!!!
  - Phase changing a substance (liquid->gas) may be the only way to hit densities in ~4 years.

PARALLELISM

- TACC Frontera is approximately:
  - 8400 nodes/16,800 chips
  - 450,000 cores
  - Each core can issue two fused multiply-add instructions each cycle, which operate on 512 bit vectors.
  - So, every 2.3 billionth of a second, the machine needs (450k*2*2*8) 64 bit FP operations, or twice that many single precision (32 bit), or 4 times as many 16-bit for AI.
    - (Your code needs to expose ~15 *million* simultaneous operations each clock cycle for full performance).
  - Software hasn’t been like this for long – let’s look back a little.



Looking at the Gordon Bell Prize

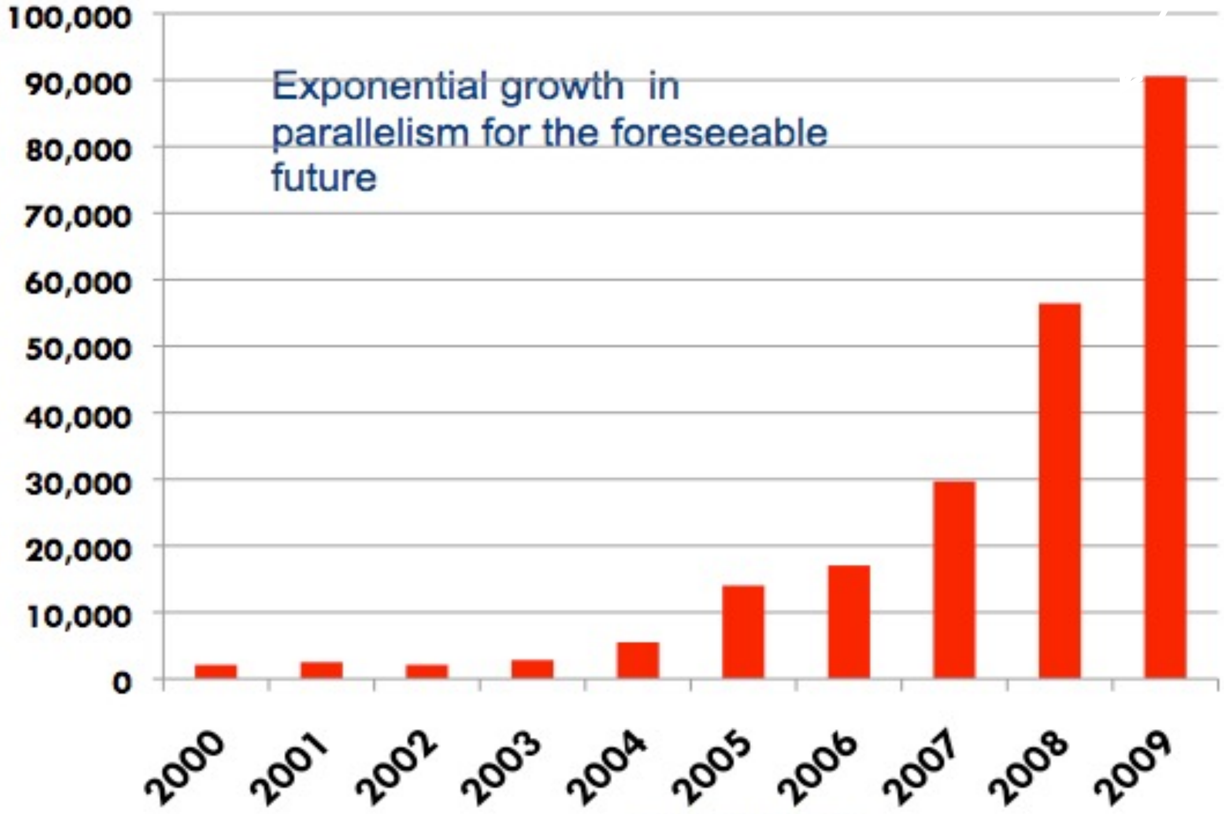
(Recognize outstanding achievement in high-performance computing applications and encourage development of parallel processing)

- □ 1 GFlop/s; 1988; Cray Y-MP; 8 Processors
  - Static finite element analysis
- □ 1 TFlop/s; 1998; Cray T3E; 1024 Processors
  - Modeling of metallic magnet atoms, using a variation of the locally self-consistent multiple scattering method.
- □ 1 PFlop/s; 2008; Cray XT5; 1.5x10^5 Processors
  - Superconductive materials
- □ 1 EFlop/s; ~2018; ?; 1x10^7 Processors (10^9 threads)

www.exascale.org

Average Number of Cores Per Supercomputer

Top20 of the Top500

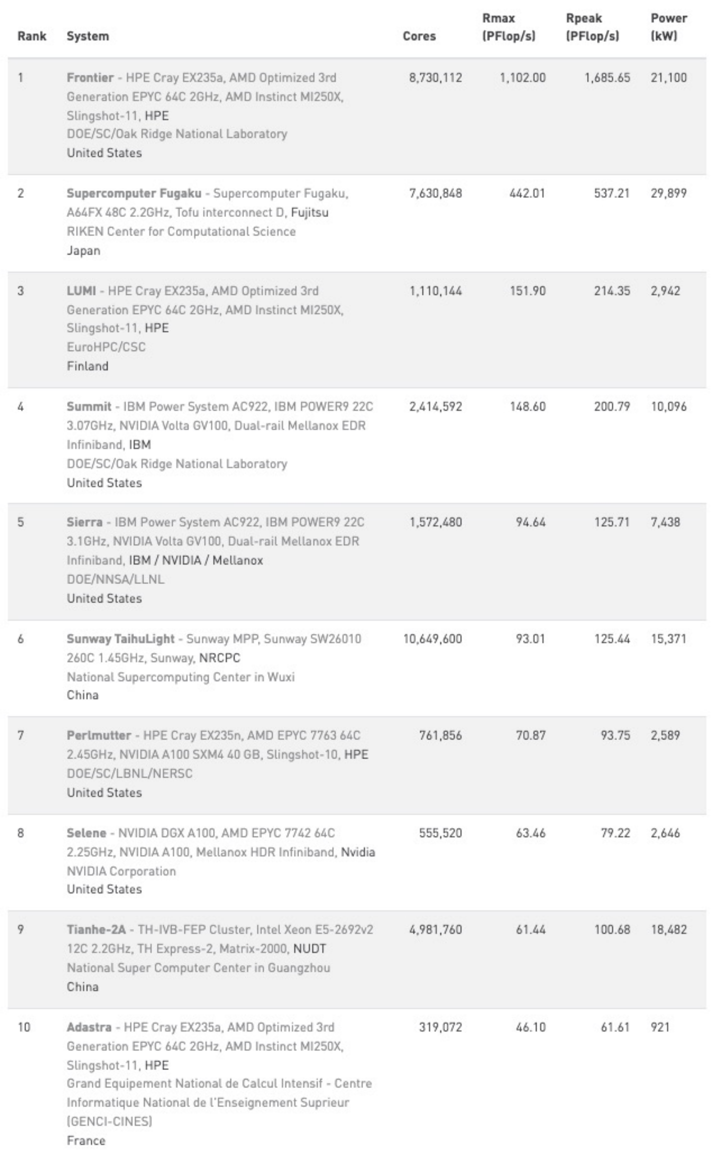


Exponential growth in parallelism for the foreseeable future

www.exascale.org

GPU

- Let’s look at the 10 biggest systems in the world. . . See a trend?



Rank	System	Cores	Rmax (PFlop/s)	Rpeak (PFlop/s)	Power (kW)
1	**Frontier** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE DOE/SC/Oak Ridge National Laboratory United States	8,730,112	1,102.00	1,685.65	21,100
2	**Supercomputer Fugaku** - Supercomputer Fugaku, A64FX 48C 2.2GHz, Tofu interconnect D, Fujitsu RIKEN Center for Computational Science Japan	7,630,848	442.01	537.21	29,899
3	**LUMI** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE EuroHPC/CSC Finland	1,110,144	151.90	214.35	2,942
4	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM DOE/SC/Oak Ridge National Laboratory United States	2,414,592	148.60	200.79	10,096
5	**Sierra** - IBM Power System AC922, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband, IBM / NVIDIA / Mellanox DOE/NNSA/LLNL United States	1,572,480	94.64	125.71	7,438
6	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway, NRCPC National Supercomputing Center in Wuxi China	10,649,600	93.01	125.44	15,371
7	**Perlmutter** - HPE Cray EX235n, AMD EPYC 7763 64C 2.45GHz, NVIDIA A100 SXM4 40 GB, Slingshot-10, HPE DOE/SC/LBNL/NERSC United States	761,856	70.87	93.75	2,589
8	**Selene** - NVIDIA DGX A100, AMD EPYC 7742 64C 2.25GHz, NVIDIA A100, Mellanox HDR Infiniband, Nvidia NVIDIA Corporation United States	555,520	63.46	79.22	2,646
9	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000, NUDT National Super Computer Center in Guangzhou China	4,981,760	61.44	100.68	18,482
10	**Adastra** - HPE Cray EX235a, AMD Optimized 3rd Generation EPYC 64C 2GHz, AMD Instinct MI250X, Slingshot-11, HPE Grand Equipement National de Calcul Intensif - Centre Informatique National de l’Enseignement Suprieur (GENCI-CINES) France	319,072	46.10	61.61	921

TRENDS IN HPC - MARKETS

- Vendor space collapsing.
- Highlander Battle
- Clouds

THE VENDOR SPACE

- I was reading the release notes on a common, well-known open source code recently.   It noted that it supported most large scale supercomputers, including:
  - "IBM Blue Gene, SGI, SUN, and Cray"
  - What all those things have in common is that they **no longer exist**.
- For a long time, you would buy processors from Intel or AMD, Infiniband from Mellanox or Qlogic, and accelerators (if you had them) from NVIDIA.
- (BTW, there are dozens of new AI chips – they still have zero penetration in HPC).

IN THE END, THERE CAN BE ONLY ONE.

- NVIDIA bought Mellanox, and started making CPUs (and, you might notice, systems).
- Intel bought up Qlogic, made IB into Omnipath, killed it, then spun it off as Cornelis again, but still pretty tightly coupled. Intel also has started making HPC GPUs, and pushing DAOS for filesystems.
- AMD is now making CPUs and GPUs. As no interconnect is aligned, at large scale they are exclusively using HP-E Slingshot. Curiously, HP-E seems to be *only* delivering AMD systems at scale this year.
- See the trend? Good luck buying Infiniband for your AMD GPU cluster. (You can get it, but since delivered price<<1/5th list price, there are. . .complications).
- This is probably. . . Not good for innovation.

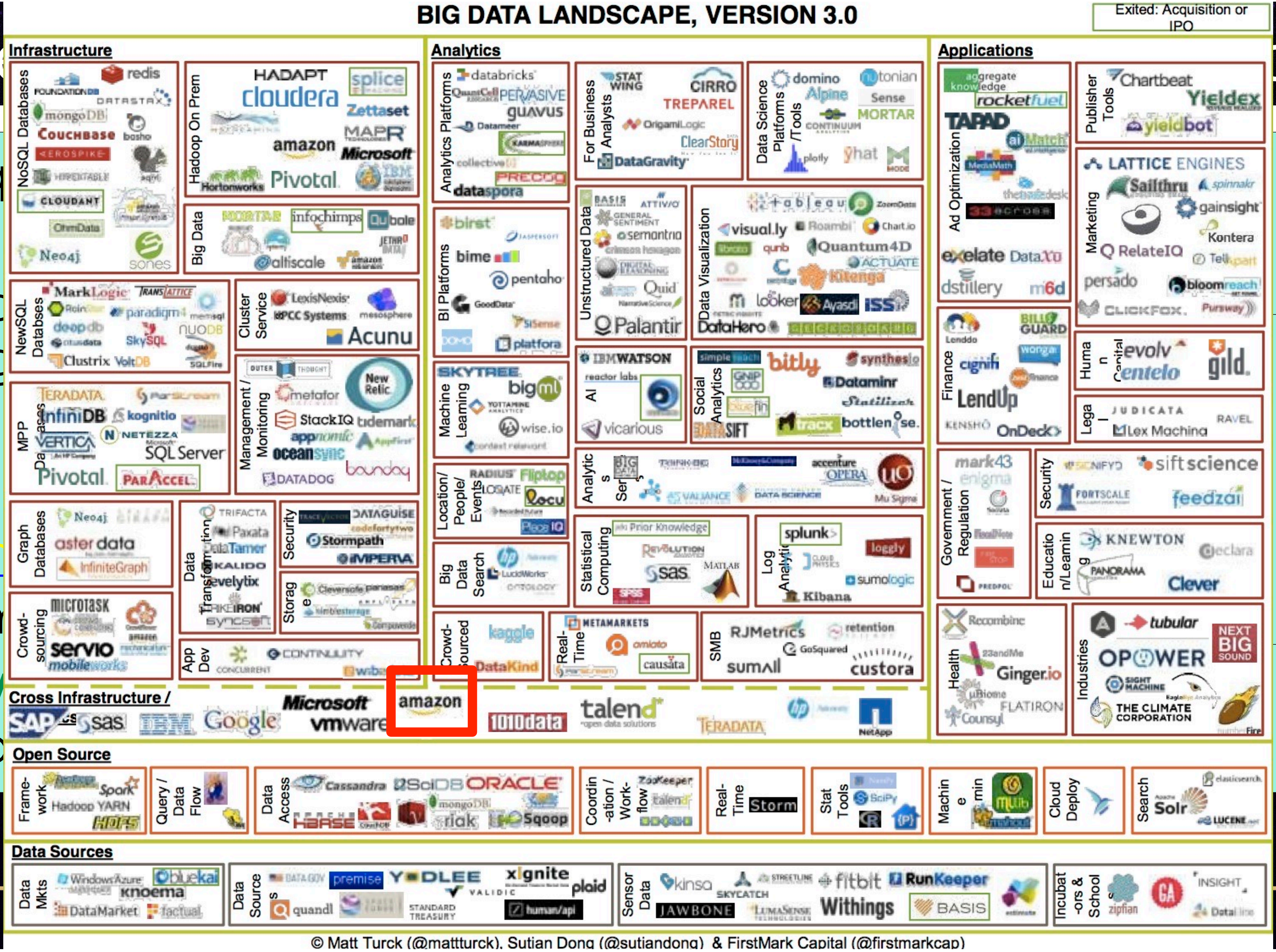
THEN THERE IS THE CLOUD, WHICH MEANS AMAZON, GOOGLE, OR MICROSOFT JUST MAGICALLY DELIVER YOU SERVICES…

What r cture)

Midterm q

- □ You
- □ Proc meg
- □ The
- □ Use
- □ h
- □ Con
- □ How
- □ (Sho

100



© Matt Turck (@mattturck), Sutian Dong (@sutiandong) & FirstMark Capital (@firstmarkcap)

TACC TEXAS ADVANCED COMPUTING CENTER
TEXAS The University of Texas at Austin

HAPPY BIRTHDAY, IKEA!
HERE'S YOUR CAKE!

TRENDS IN HPC - USERS

- Modes of Compute
- On-Demand
- Higher Levels of Abstraction

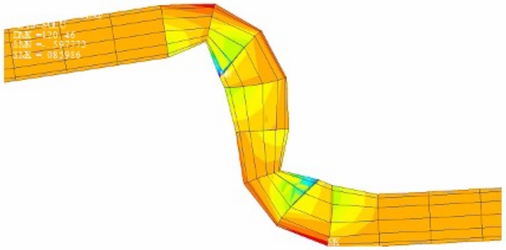
NEW MODES OF COMPUTE

- Our “all batch” model doesn’t quite work anymore (though it is not going away, either).
  - Interactive
    - Not just impatience
    - Data exploration, hyperparameter tuning.
  - Real-Time
    - Time to solution matters!
      - Hurricane forecasts, planning tumor-specific dosages of proton therapy, control of a fusion reactor. . .
  - Persistence
    - Connection to web services, responding to automated requests.

Workflow for seismic risk assessment of gas pipelines

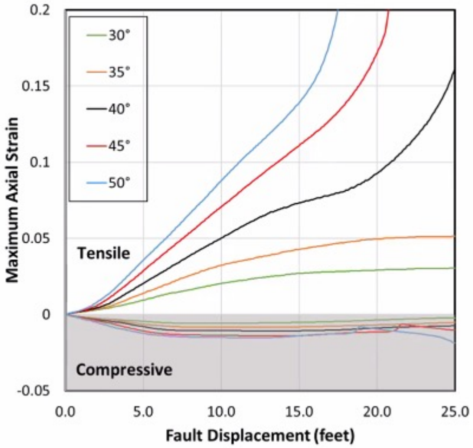
Main point: comprehensively understand the **pipeline** response subjected to **natural hazards** (fault displacement and others) to improve future **risk assessment** and **decision making**

Raw Data
200,000 * 10GB



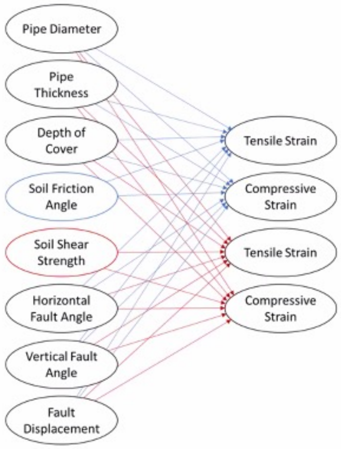
2000 TB

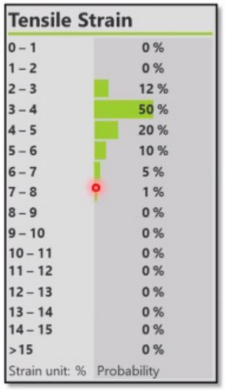
Strain curves
200,000 * 1MB



200 GB

4 Strain Models
4 * 3MB

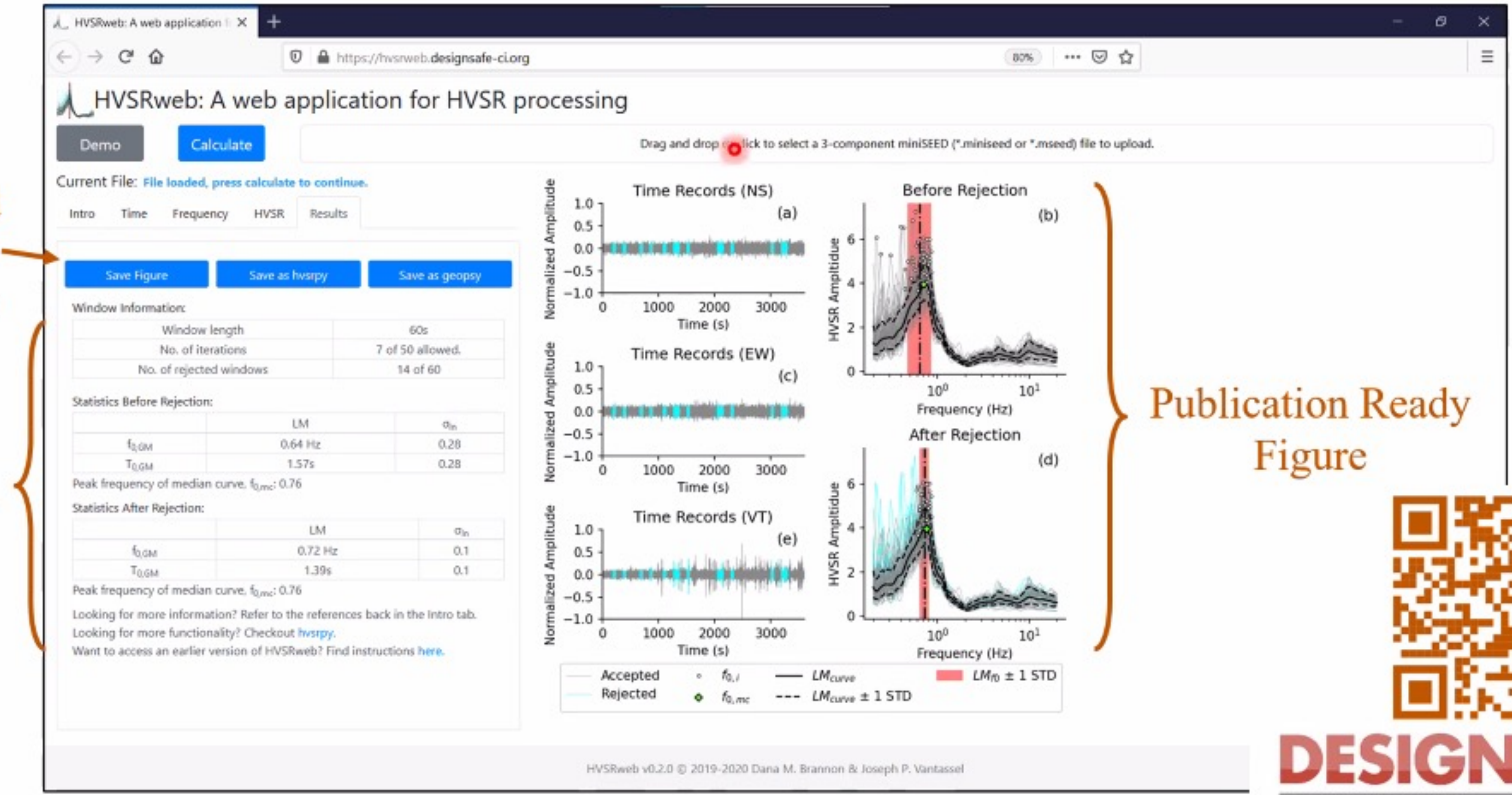




Tensile Strain	
0 – 1	0 %
1 – 2	0 %
2 – 3	12 %
3 – 4	50 %
4 – 5	20 %
5 – 6	10 %
6 – 7	5 %
7 – 8	1 %
8 – 9	0 %
9 – 10	0 %
10 – 11	0 %
11 – 12	0 %
12 – 13	0 %
13 – 14	0 %
14 – 15	0 %
>15	0 %
Strain unit: %	Probability

12 MB

Measuring a Site's Resonant Frequency



HVSRweb: A web application for HVSR processing

Drag and drop or click to select a 3-component miniSEED (*.miniseed or *.mseed) file to upload.

Current File: File loaded, press calculate to continue.

Window Information:

Window length	60s
No. of iterations	7 of 50 allowed.
No. of rejected windows	14 of 60

Statistics Before Rejection:

	LM	σ_{ln}
$f_{0,GM}$	0.64 Hz	0.28
$T_{0,GM}$	1.57s	0.28

Peak frequency of median curve, $f_{0,mc}$: 0.76

Statistics After Rejection:

	LM	σ_{ln}
$f_{0,GM}$	0.72 Hz	0.1
$T_{0,GM}$	1.39s	0.1

Peak frequency of median curve, $f_{0,mc}$: 0.76

Looking for more information? Refer to the references back in the Intro tab.
Looking for more functionality? Checkout hvsrpy.
Want to access an earlier version of HVSRweb? Find instructions here.

HVSRweb v0.2.0 © 2019-2020 Dana M. Brannon & Joseph P. Vantassel

Download Figure & Summary Files

Detailed Summary

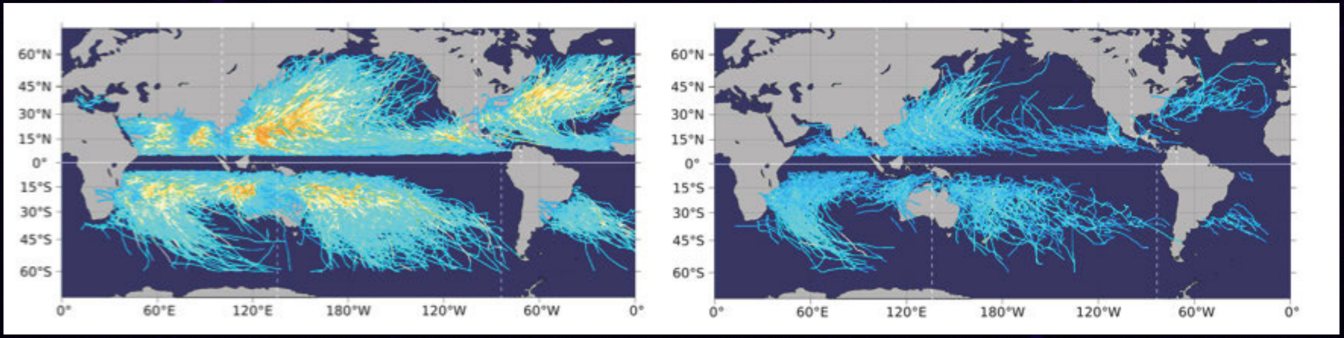
Publication Ready Figure

DESIGNSAFE-CI
A NATURAL HAZARDS ENGINEERING COMMUNITY

FRONTERA BRIDGES TO THE COMMERCIAL CLOUD:

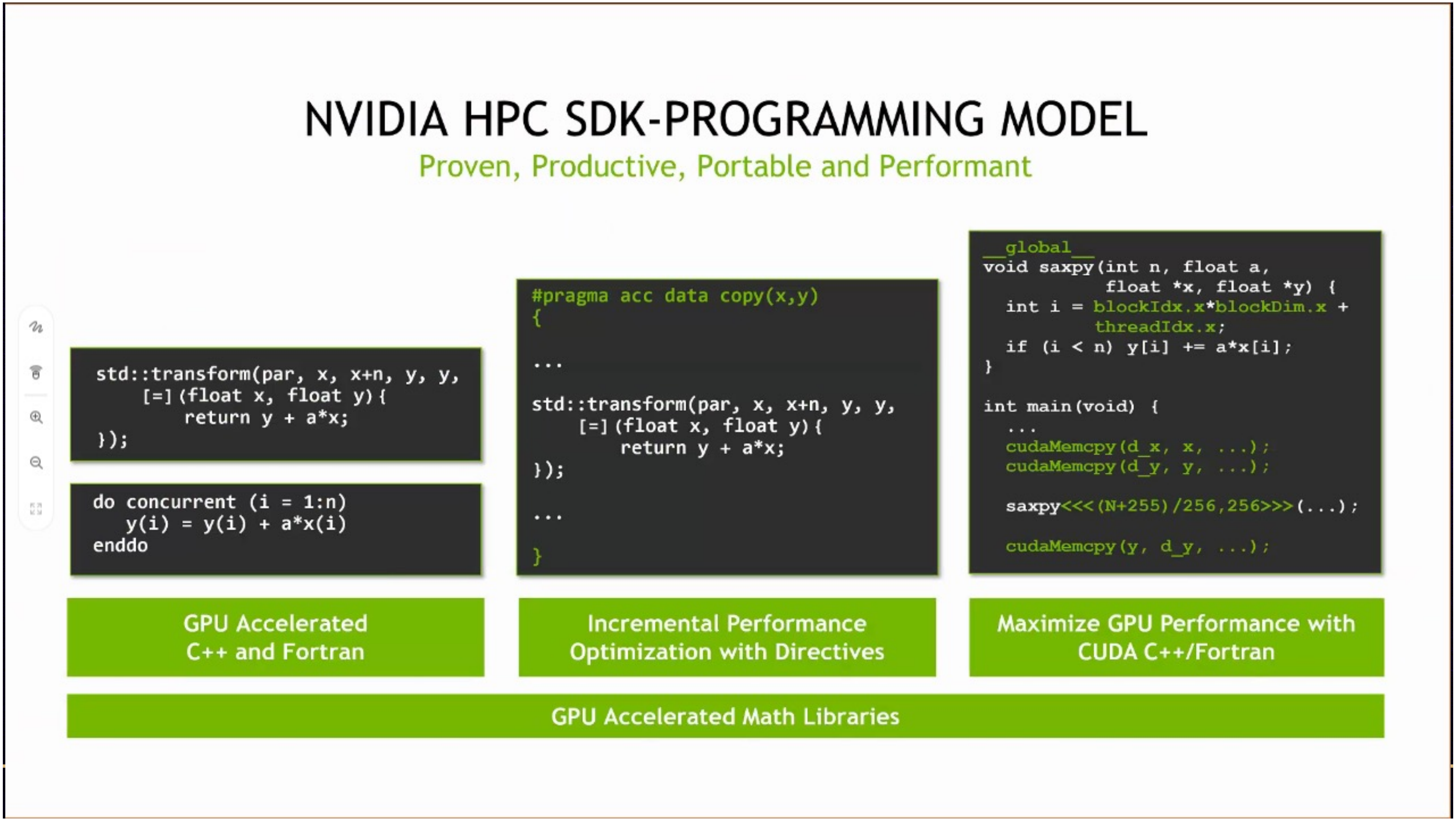
NEW HIGH-RESOLUTION GLOBAL CLIMATE MODEL PREDICTS CHANGES TO EXTREME WEATHER, OCEAN CURRENTS

- Ping Chang, Texas A&M (collaborators at NCAR and QNLM).
- CESM, 25KM resolution
- One of the UN-supported HighResMIP simulations – the official climate change forecasts.
- In addition to large simulations, also published the data results to Azure, where they do online analytics on ~550GB of data with the "PanGeo" Cloud platform.



SO, WHAT DOES THIS MEAN FOR YOU, THE RESEARCHER/USER?

- Your code should be really, really parallel (and do I/O well!!!).
- Parallelism means instruction level, thread level, task level. And thinking about where the memory is.
- Follow some best practices (NVIDIA version next slide).



NVIDIA HPC SDK-PROGRAMMING MODEL

Proven, Productive, Portable and Performant

```
std::transform(par, x, x+n, y, y,
    [=](float x, float y){
        return y + a*x;
});
```

```
do concurrent (i = 1:n)
   y(i) = y(i) + a*x(i)
enddo
```

GPU Accelerated
C++ and Fortran

```
#pragma acc data copy(x,y)
{

...

std::transform(par, x, x+n, y, y,
    [=](float x, float y){
        return y + a*x;
});

...

}
```

Incremental Performance
Optimization with Directives

```
__global__
void saxpy(int n, float a,
           float *x, float *y) {
  int i = blockIdx.x*blockDim.x +
          threadIdx.x;
  if (i < n) y[i] += a*x[i];
}

int main(void) {
  ...
  cudaMemcpy(d_x, x, ...);
  cudaMemcpy(d_y, y, ...);

  saxpy<<<(N+255)/256,256>>>(...);

  cudaMemcpy(y, d_y, ...);
```

Maximize GPU Performance with
CUDA C++/Fortran

GPU Accelerated Math Libraries

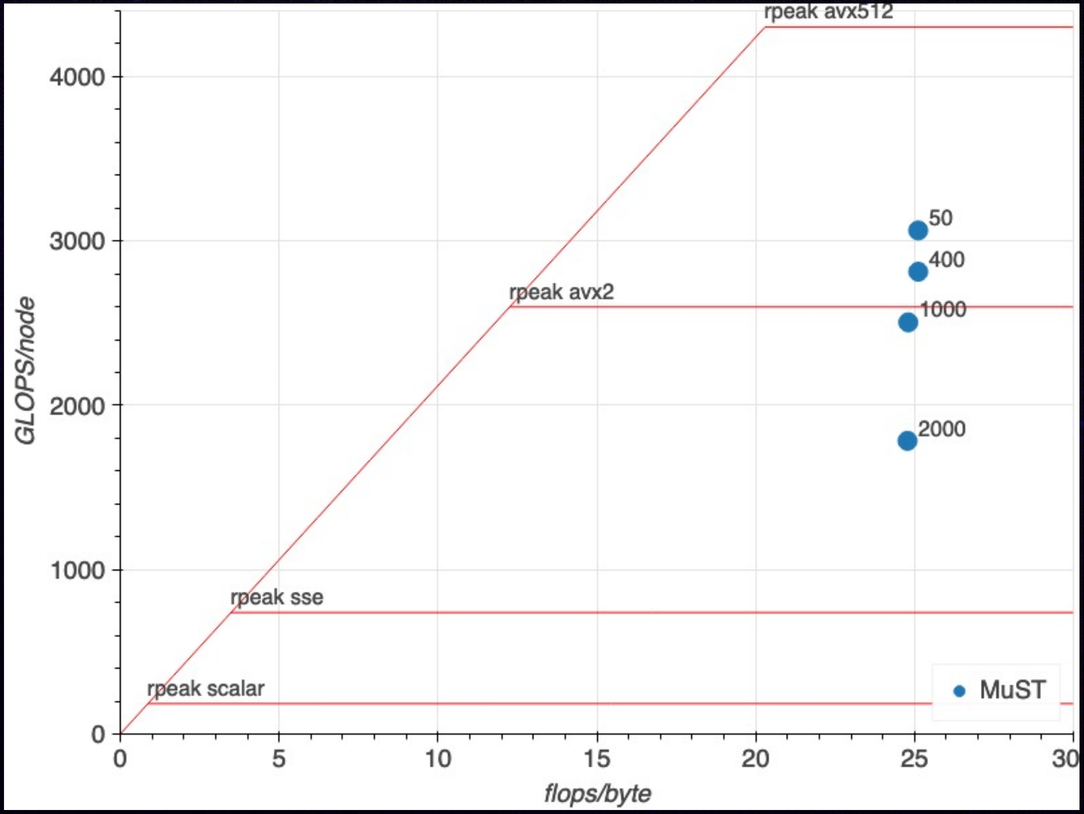
SO, WHAT DOES THIS MEAN FOR YOU, THE RESEARCHER/USER?

- ”Ready for Future” – use new parallel constructs in C++/Fortran/Python
  - (Expect this not to be “most performant” in the near term).
- Otherwise
  - Use Directives for thread-level parallelism (OpenMP/CUDA/OpenACC)
  - Use MPI for task-level parallelism, plus one of the above thread systems.
- Align MPI tasks with layout of chips (one AMD chiplet per task, one per GPU. . .), try to pin to sockets so memory doesn’t move.
- Think about scale – the checkpoint interval on your laptop is not the same as on 100k cores.
- Codes can work at this scale… but many still don’t.

SOME SUCCESSFUL EXAMPLES

MuST

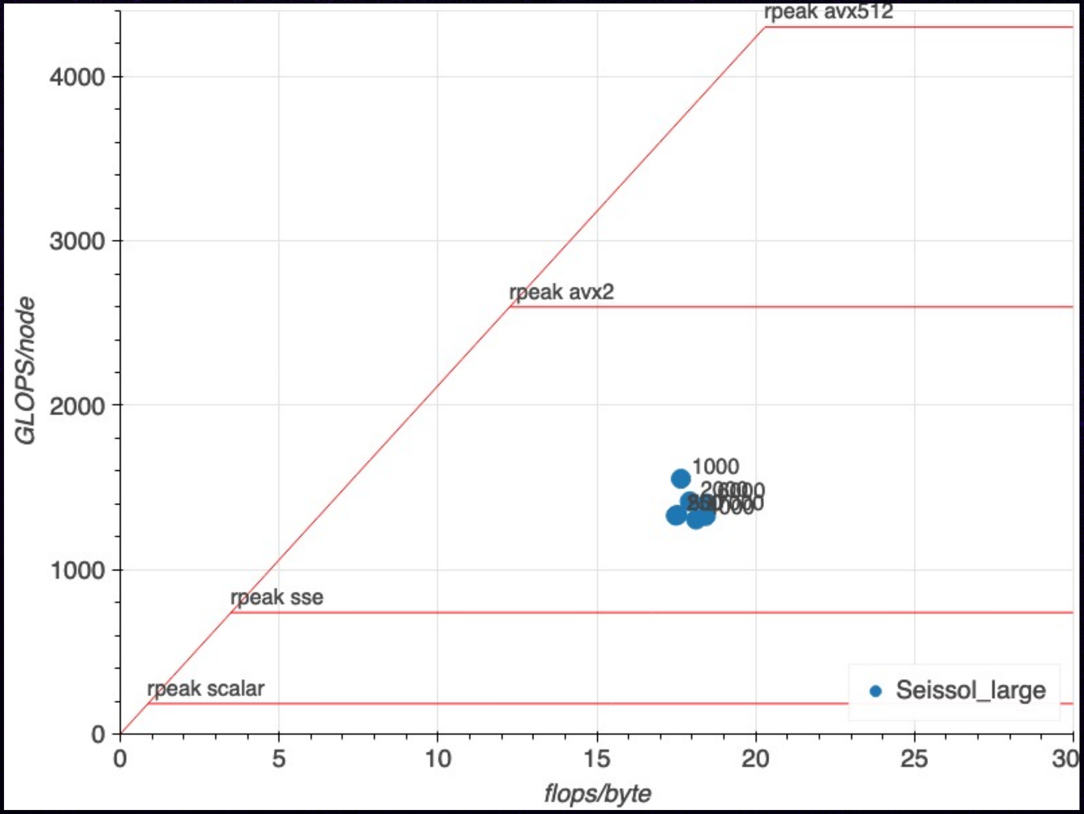
- Ab initio electronic structure code
- MPI/OpenMP
- GPU version also
- ~70% peak performance (50 nodes)
- Potential for scaling improvement due to complicated communication pattern
- Performance per node should improve with hardware



SOME SUCCESSFUL EXAMPLES

SeisSol

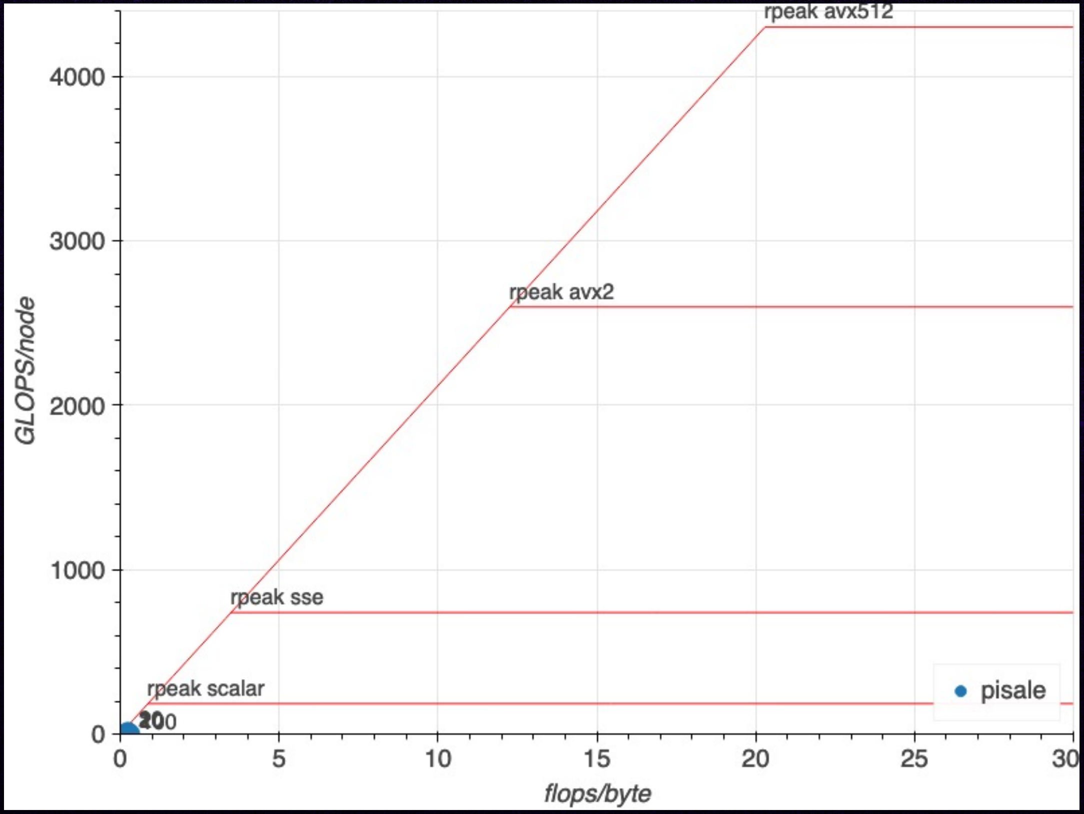
- Seismic wave propagation code
- MPI/OpenMP
- GPU version also
- ~40% peak performance on CLX
- Consistent from 250 – 7000 nodes
- GPU version doesn't scale as well
- Potential for performance increase
  - More nodes/cores
  - Improved MBW



SOME LESS SUCCESSFUL EXAMPLES

PISALE

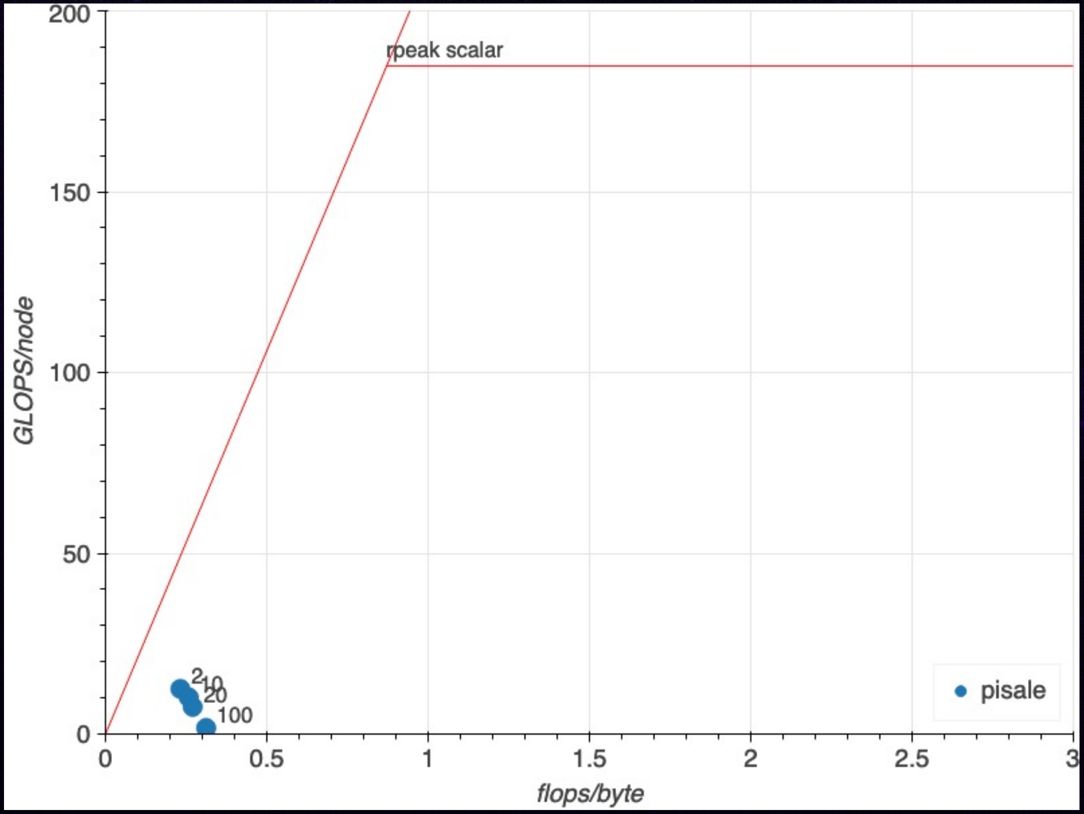
- Arbitrary Lagrangian Eulerian code
- MPI/RAJA(SAMRAI)
- GPU version dependent on SAMRAI
- Very low performance on CLX
- Scales only to 10 nodes
- Code not ready for LCCF scale



SOME LESS SUCCESSFUL EXAMPLES

PISALE

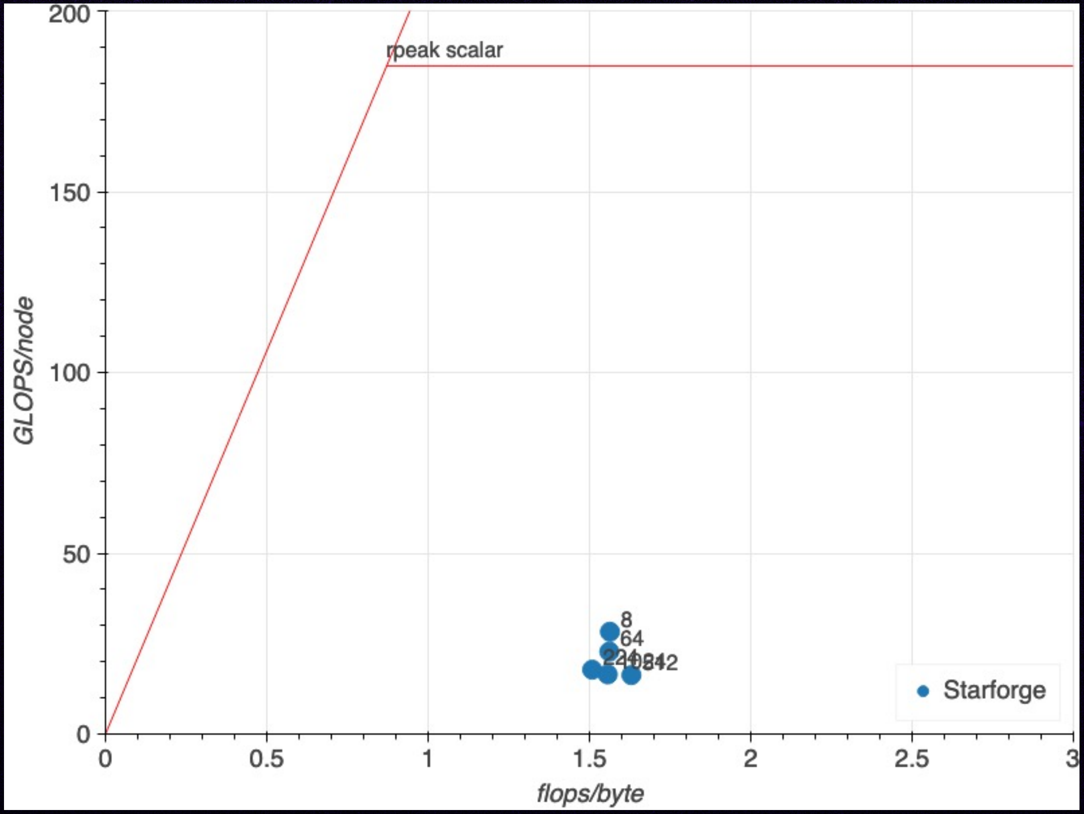
- Arbitrary Lagrangian Eulerian code
- MPI/RAJA(SAMRAI)
- GPU version dependent on SAMRAI
- Very low performance on CLX
- Scales only to 10 nodes



SOME LESS SUCCESSFUL EXAMPLES

StarForge

- Meshless MHD code
- MPI/OpenMP
- Communications bound
- Very low performance on CLX
- Poor MBW



FRONTERA
TACC
NSF
TEXAS

MY TURN. . .

- (Programming) Language of choice?
- Incorporating AI?
- Ever think about I/O strategies?
- Do you use Jupyter?

- How much of your time goes to code/data (and related issues) vs. science/engineering?

A FEW QUICK COMMENTS ON THE CLOUD.

- The cloud is a wonderful thing, that has its place(s) in the computing ecosystem, but. . .
  - Frontera costs last year: $24M ($12M Operations, $12M HW Depreciation).
  - Frontera delivered 64.6M Intel Cascade Lake Node hours, and 3.6M Quad-V100 GPU hours.
  - Commercial rates (using AWS, September 2021) for equivalent node hours:
    - CPU: c5n.18xlarge , which is not as good but close enough.
    - GPU: p3.8xlarge

A FEW QUICK COMMENTS ON THE CLOUD.

- Let's assume:
  - Cloud storage if free *(it's not)*.
  - Cloud Parallel Filesystem is free *(it's very much not)*
  - Cloud networking/transfer is free *(it's not)*
  - Tightly coupled interconnects don't matter *(they do)*.
  - You would get 1-1 performance for large jobs on clouds vs. Frontera *(you won't)*.
  - There is zero effort to get diverse technical workloads to run on the cloud *(there is a lot)*.

A FEW QUICK COMMENTS ON THE CLOUD.

- Even with those ridiculously optimistic assumptions:
  - The equivalent cloud cost for delivered cycles on Frontera for FY2021 is ***$291,000,000***.
- Turns out, $291M > $24M.   Any questions?