ON THE HORIZON:
INTERCONNECTS IN FRONTERA AND THE COMING REPLACEMENT SYSTEM

Dan Stanzione
Executive Director
Associate Vice President for Research
MUG
August 2022

THANKS FOR INVITING ME BACK!

Sorry I can't be there in person, but our first in-person NSF review ended 60 minutes ago.

Looking forward to getting back to Columbus after visits in 2017, 2018, and 2019.

A QUICK TACC REMINDER

- We operate the Frontera, Stampede-2, Jetstream, and Chameleon systems for the National Science Foundation
- Longhorn and Lonestar-6 for our Texas academic and industry users.
- Altogether, ~20k servers, >1M CPU cores, 1k GPUs
- About seven billion core hours over several million jobs per year.

TACC - 2022

LEADERSHIP-CLASS
COMPUTING FACILITY

TACC
TEXAS ADVANCED COMPUTING CENTER

THE THIRD YEAR OF PRODUCTION IS DRAWING TO A CLOSE ON FRONTERA

- And the system has done great!
- In the last 12 months:
  - Uptime of 99.2%
  - Average Utilization of 95.4%
  - ~72M SUs delivered
  - 1.13M jobs delivered
  - Zero security incidents.
- Happy to compare uptime, utilization numbers with any modern supercomputer.
  - On the bright side, we are always full. On the downside, no way to squeeze anything else in.

A LITTLE MORE ON USAGE

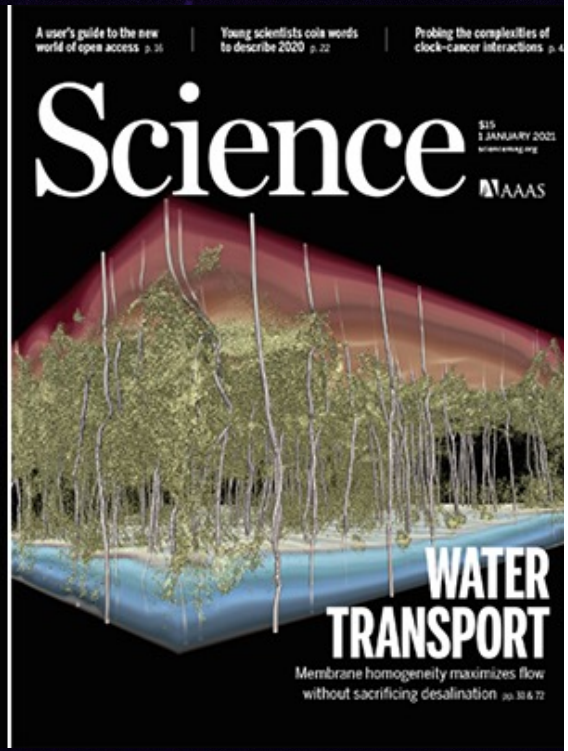
- **>2,000 jobs were >25,000 cores** – about a **quarter** of all cycles on large jobs.
- >100 jobs at half or full system scale (Consider if all jobs were full scale, and averages 24 hours, we'd only run 365 jobs a year, as opposed to 1.1M jobs).
- Flex jobs, used for backfill, represent 20% of the jobs run (>200k), but represent less than 0.5% of SUs delivered (285k out of 70M).
- Small jobs represent ~30% of jobs, but less than 2% of cycles delivered.
  - So **97% of time goes to jobs >2 nodes**.
  - Average jobs size about **6x that of Stampede2** – this machine *is* used differently.
- We tune the scheduling policy multiple times a year… essentially adjusting to demand.

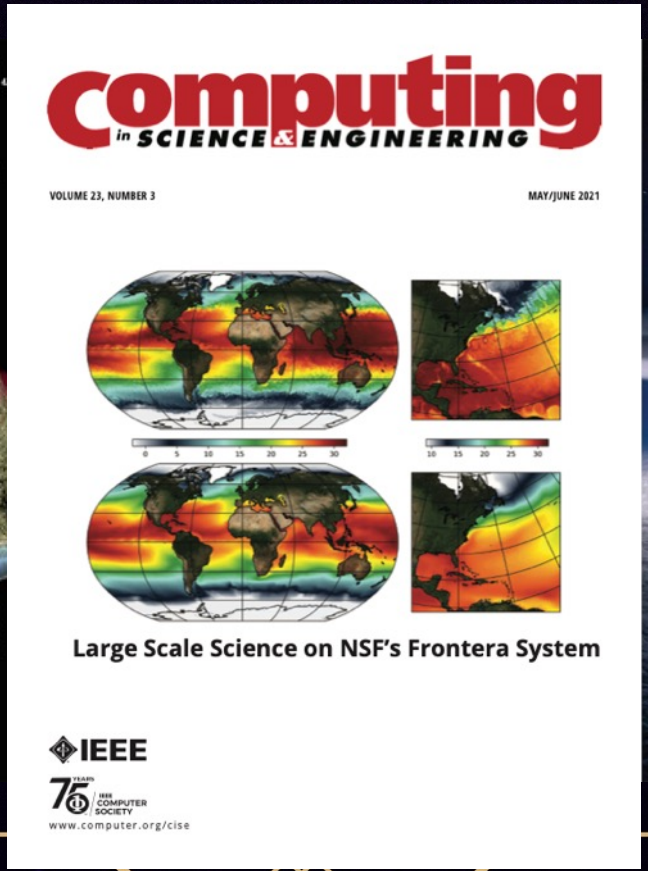
BY QUEUE

Queue	Job Count (2020)	SUs Charged (2020)	Job Count (2021)	SUs Charged (2021)	Job Count (2022)	Sus Charged (2022)
Normal	556,048	38,577,043	906,114	44,157,946	308,476	50,390,674
Development	47,119	183,901	124,526	621,317	153,635	745,604
Flex	457,392	413,471	609,180	271,791	209,706	285,247
Large	2,106	7,989,616	1,769	14,133,257	1,142	13,018,134
RTX	25,872	591,186	82,392	1,623,327	80,477	1,060,014
RTX_DEV	1,676	3,998	10,944	25,578	13,221	20,921
NVDIMM	905	7,954	9,876	90,779	6,920	115,784
Small	--	--	111,380	407,043	316,953	1,253,289
Debug*	--	--	3,793	3,236,969	2,133	2,496,290
Others	--	--	27,696	78,189	40,513	158,224
TOTAL	1,091,118	48,827,566	1,887,670	64,646,197	1,133,176	69,544,181

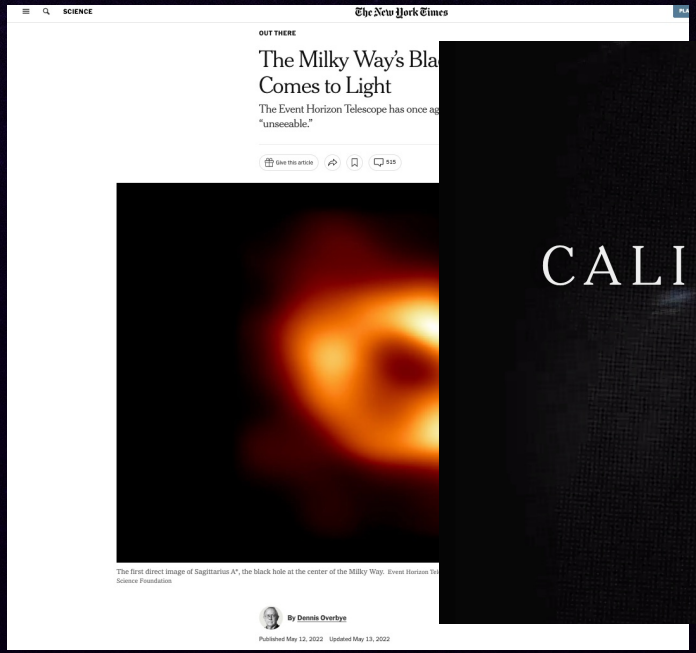
Longhorn had another 3.7M SU charged.
*Texascale jobs are largely in Debug Queue

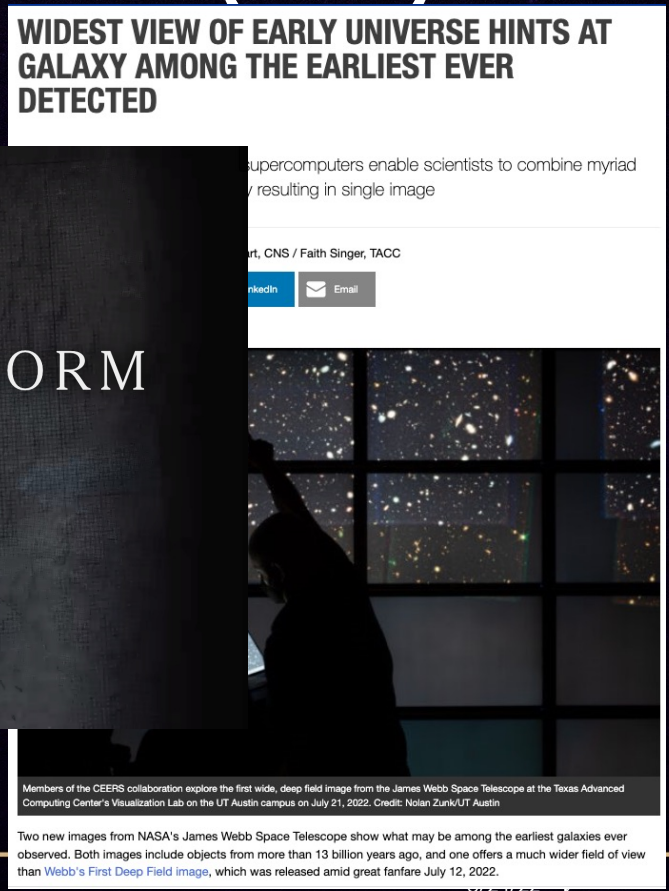
LOTS OF GREAT SCIENCE





ALL FOR UNCLASSIFIED, OPEN SCIENCE (2022)





FIRST IMAGE OF THE BEASTLY BLACK HOLE AT THE HEART OF OUR GALAXY

- Event Horizon Telescope
- Time provided for simulation and data analysis
  - Relatively small user, but relatively large impact!
- Major NSF press drive on this in the Spring (which mentioned Frontera)
- Special issue of the Astrophysical Journal Letters

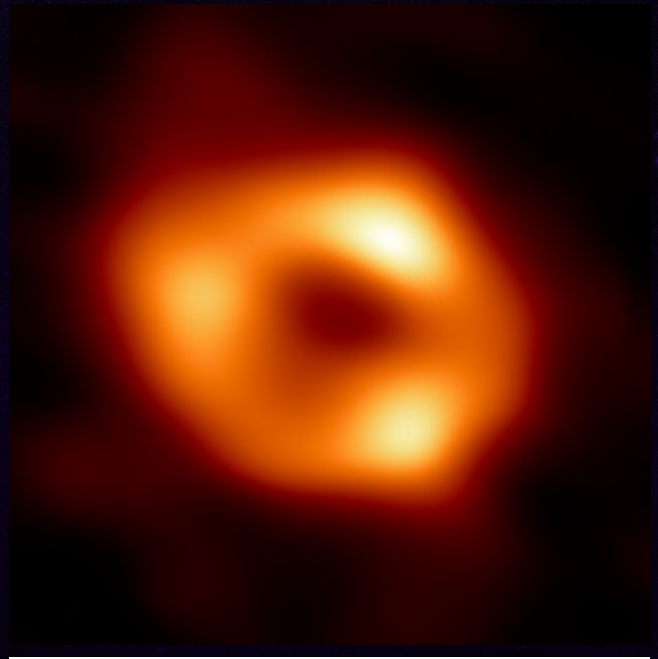


Figure 1 - The first direct image of Sagittarius A*, the black hole at the center of the Milky Way. Credit: Event Horizon Collaboration/National Science Foundation

FRONTERA REVEALS WEAKNESS IN HIV-1 ARMOR

- The viral capsid has to stay stable long enough to take its genetic cargo into the nucleus of the cell. But in the end, it has to break apart to release its genetic material.
- Frontera simulations furthered scientists' understanding of how the HIV-1 virus infects and helped generate the first realistic simulations of its capsid, complete with its proteins, water, genetic material, and a key cofactor called IP6 recently discovered to stabilize and help form the capsid.
- Started with analysis of Cryo-EM data of HIV.
- *"Supercomputers combined with the methods we developed have helped reveal essential elements of the HIV-1 virus that are experimentally extremely difficult to probe at present. I don't think we could have easily done those simulations anywhere else but on Frontera. "*
- Greg Voth, U. of Chicago
- PNAS, March 2022

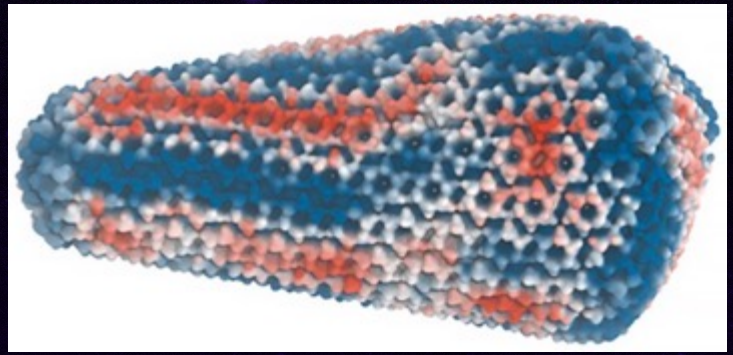


Figure 2 - The HIV-1 capsid encloses its genetic material, traveling all the way to the nucleus of infected white blood cells before it breaks apart to unleash its deadly genetic cargo. Simulations based on experimental evidence were developed on TACC's Frontera supercomputer and revealed stress-strain patterns of the capsid just prior to the critical break-up stage, indicating potential vulnerabilities to exploit for drug design. Image shows strain of the HIV-1 capsid, with red and blue colors corresponding to compressive and expansive strain, respectively. Credit: Yu, et al. DOI:10.1073/pnas.2117781119

HOW THE BRAIN PREPARES TO THINK

- Basic mechanisms of thought are believed to be the result of very fast vesicle fusion between neurons.
- Built a multi-million atom model of the proteins, the membranes, and their environment ---
- "Supercomputers weren't powerful enough to resolve this problem of how transmission was occurring in the brain. So for a long time, I used other methods… *However, with Frontera, I can model 6 million atoms and really get a picture of what's going on with this system.*"
- *"We have a supercomputer system here at the University of Texas Southwestern Medical Center. I can use up to 16 nodes," he said. "What I did on Frontera, instead of a few months, would have taken 10 years"*
- Jose Rizo-Reyes, UT Southwestern Medical Center (Dallas)
- NIH R35, eLife June, 2022

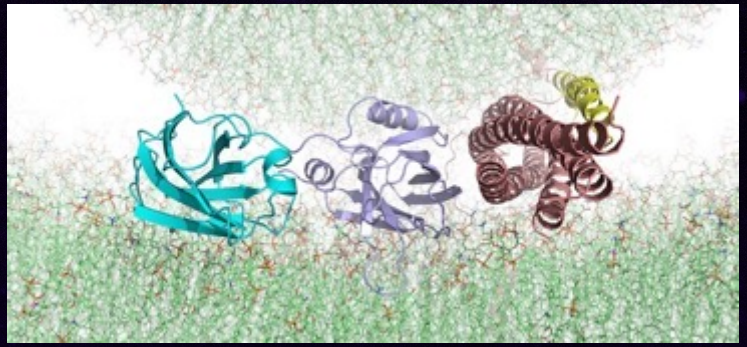


Figure 3 - Configuration of the primed synaptotagmin-SNARE-complexin complex suggested by molecular dynamics simulations. Credit: Rizo-Rey, UT Southwestern Medical Center

CARBON SEQUESTRATION

- Deep injection of supercritical CO2 into rock formations for long-term storage
- Developing ML model to use for prospective sites
- Identified 2 key parameters: Injection Rate and "wettability" of specific formations.
- *International Journal of Greenhouse Gas Control, 12/21*
- *Environmental Science and Technology, 10/21*
- Sahar Bakhsian, UT-Austin BEG

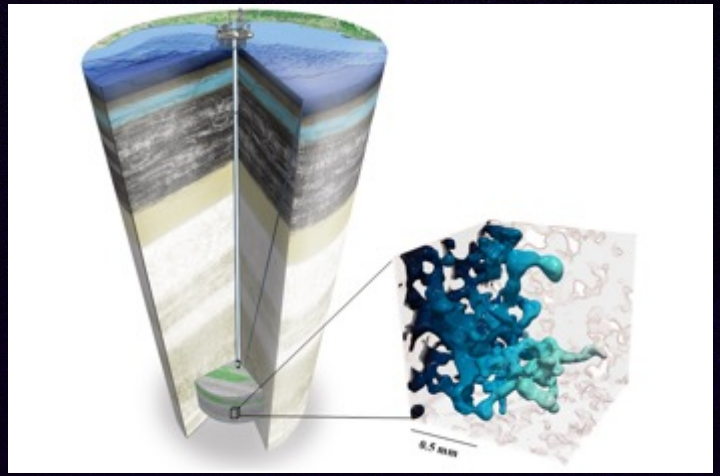
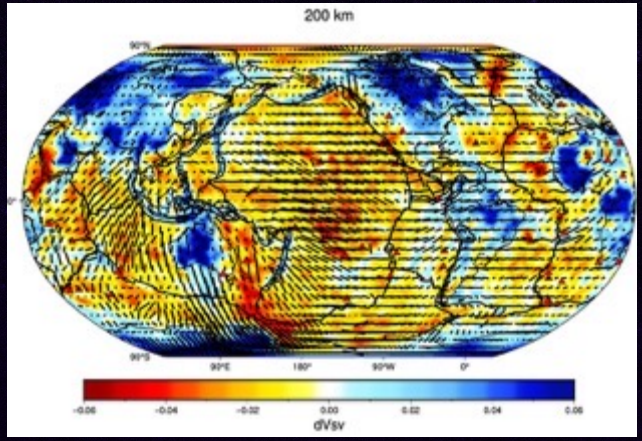


Figure 4 - Left: Subsurface CO2 storage. Right: CO2 migration pattern in a digitized rock sample obtained from pore-scale two-phase flow simulation. The simulation was carried out on the Frontera supercomputer.

AN MRI OF THE EARTH

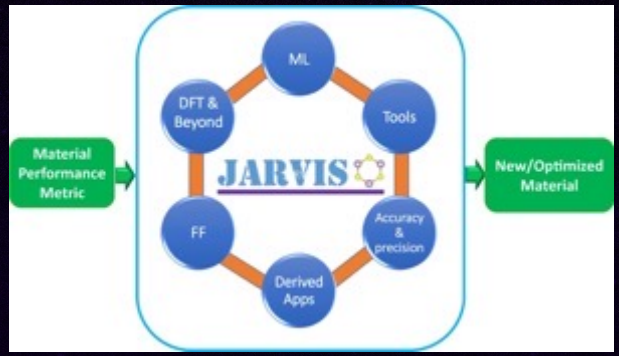
- Full-Wave inversion study of the whole earth (using seismic waves from earthquakes.
- Newest physics modeling subduction zones, hotspots, magma/mantle plumes.
- Data assimilation from marine earthquake detection platforms.
- Ebru Bozdag, Colorado School of Mines
- *Computers and Geosciences*, April 2022
- NSF CAREER Award



Azimuthal anisotropy (black dashed lines showing the fast direction of wave speeds) in the mantle at 200 km depth plotted on top of vertically polarized shear wave speed perturbations (dVsv) after 20 iterations based on global azimuthally anisotropic adjoint tomography. The maximum peak-to-peak anisotropy is 2.3%. Red and blue colors denote the slow and fast shear wave speeds with respect to the mean model which are generally associated with hot and cold materials, respectively.

AN AI ASSISTANT FOR MATERIALS DISCOVERY

- JARVIS – Joint Automated Repository for Various Integrated Simulations.
- AI Model trained by 70k DFT materials simulations.
- Recently used to predict the CO_2 adsorption properties of Metal Organic Frameworks, a class of porous materials that can remove CO_2 from the atmosphere, and to computationally rank leading candidates for experimental synthesis
- NIST Materials Genome Initiative
- *Nature Computational Materials*, December 2021
- David Vanderbilt, Rutgers, NAS
- *"The machine learning field has been around since the 1980s, but the main problem was well-curated datasets," Choudhary said. "We're now approaching 100,000 materials in our database and that was only possible because of Frontera*



For a given materials performance metric, several JARVIS components can work together to design optimized or completely new materials. [Ref: https://www.nature.com/articles/s41524-020-00440-1]

INTERCONNECT

- Mellanox HDR , Fat Tree topology
- 8008 nodes = 88*91 = 91 Compute Racks
- Mellanox ASICS == 40 HDR ports. Chassis switches have 800 ports.
- Each rack is divided in half, with it's own TOR switch:
  - 44 compute nodes at HDR-100 == 22 HDR ports
  - 18 uplink 200Gb HDR ports, 3 lines (600Gb) to each of 6 core switches.
- No oversubscription in higher layers of tree (11-9 in rack).
- No oversubscription to storage, DTN, service nodes (all connected to all 6 switches).
- 8500+ cards, 182 TOR switches, 6 core switches, 50 miles of cable.
- Good news: 8,008 compute nodes use only 3,276 fibers to connect to core.

YOU CAN'T USE AN INTERCONNECT WITHOUT A SOFTWARE STACK

- As always, Frontera is a place where we push and tune MVAPICH at new scales (more nodes, more cores, etc.)
- The MVAPICH team did a lot of work in tuning MVAPICH for HDR, and for Frontera specifically.
  - Some codes always improve dramatically from "out of the box" with MPI tuning.
- We on the expertise of the team here for both tools and research into:
  - runtime introspection,
  - online monitoring,
  - recommendation generation,
  - auto-tuning of MPI parameters

MVAPICH IS ALWAYS HELPFUL!

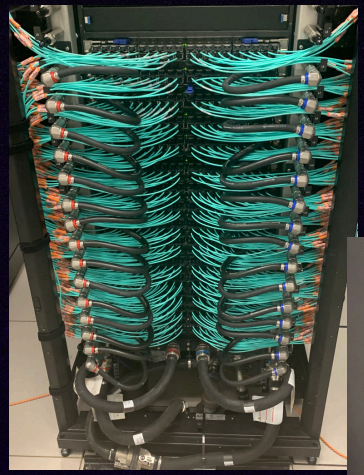
- QMCPACK far outperformed our estimates on Frontera.
  - Why?
    - Dominated by very small messages, in collectives.
    - **MVAPICH TO THE RESCUE!** MVAPICH on IB does substantially better in this scenario than Intel MPI on OPA
      - Validated on older machines.
  - This code is probably 50x faster with a sub-5us interconnect than on a higher latency network, for any large node count.

PHASE 2

LEADERSHIP-CLASS
COMPUTING FACILITY

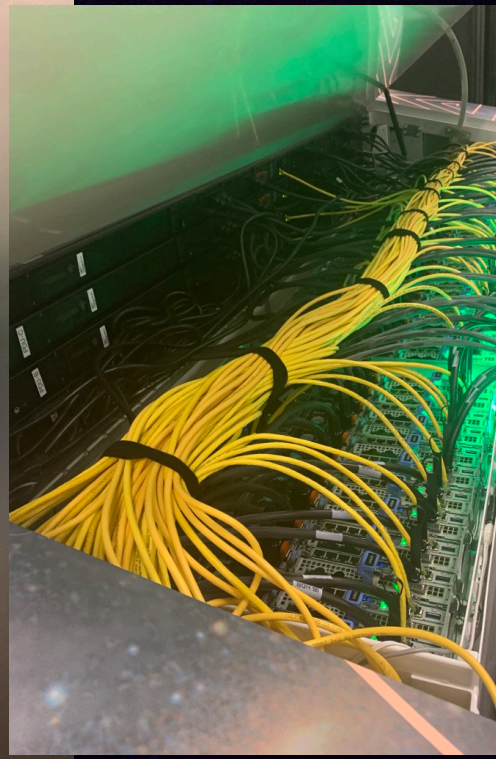
HOW WE SEE SYSTEMS TODAY

- Importantly – we are a user facility. We run *thousands* of applications, and we don't have any real control over any of them (other than occasionally kicking some off). Most of them, like all software, are poorly written crap.
  - We have to be general purpose, and we are a shared, open environment.
  - Stampede2, for instance: 16,000 users have SSH access, another 50k through web services.
- We typically have two interconnects:
  - Ethernet – mostly just for establishing IP-based connections to nodes, ssh to start a session or tunnel etc. Our ethernet is cheap and oversubscribed.
  - Infiniband/Omnipath (and Rockport testbed!) – Fat Tree, little oversubscription. Carries all filesystem traffic, and all node-to-node messaging.
    - 100/200Gbps per node today – many Tbps across the core switches
      - Frontera rack – 36 fibers to core from each rack at 7.2Tbps, *100+ racks.
    - Max latency <1us in rack, less than 2 microseconds across full system



HOW WE SEE SYSTEMS TODAY

- Latency is the dominant performance driver for MPI jobs
  - (which make up 45% of our jobs, but 97% of compute time delivered).
- Bandwidth/IOPS matters more for I/O.
- So naturally both kinds of traffic go over the same network ☺.



LOOKING FORWARD ON INTERCONNECTS. . .

- What are our options for our next system?
- If we "stay the course":
  - Infiniband
  - Resurgent OPA
  - Slingshot
  - Rockport
  - Low-latency ethernet?
- If we evolve somewhat:
  - Could we use disaggregation to replace the traditional interconnect, or would it add a third network?
  - Could we ever justify the cost for this?

CONCERNS IN THE TRADITIONAL PATH

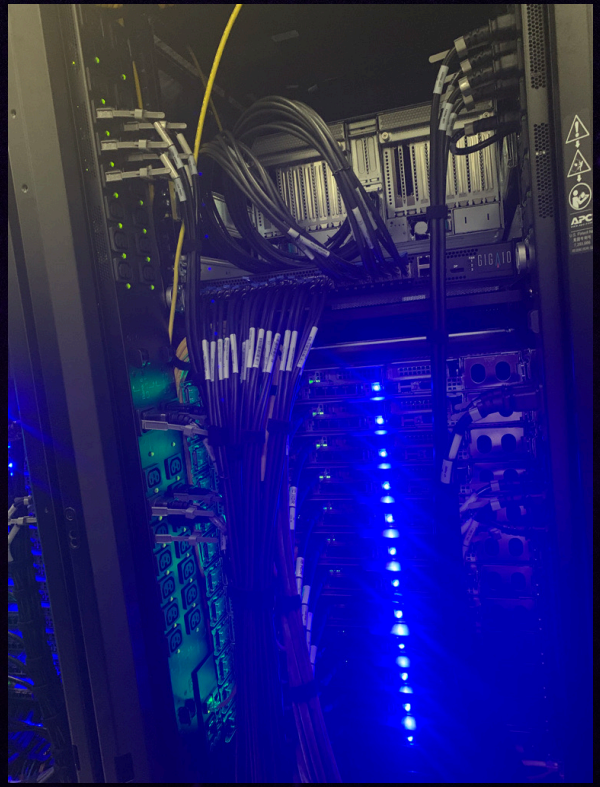
- Vendor consolidation may dictate choice:
  - Will Slinghot play outside of HP-E Systems? Will Mellanox favor NVIDIA? Whither Intel and AMD?
  - These may be more important than any *technical* problems we'd have with any of these otherwise excellent products.
- How many endpoints will future fabrics need?
- What share of the budget will they take?
- Are new options viable?

THINKING ABOUT ENDPOINTS

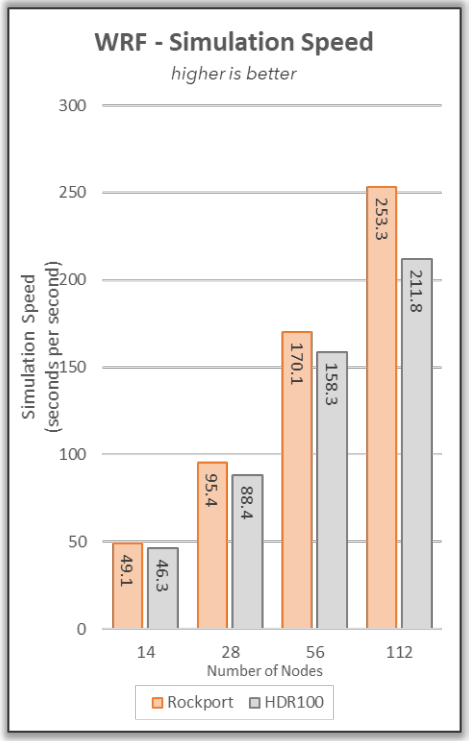
- Lately, heterogeneous systems have seen node counts actually decline. . .
- But rails per node going *up*.
  - Are we better off with a quad-CPU, quad-GPU node with 4 network rails, or one of each?
  - The "one of each" might be cheaper and simpler… but you have to adopt distributed memory (more on that later).
- Regardless, that might mean a 4k (node) system would have 16k network endpoints.
- And if you did a 16k "cheap" node system, but disaggregated the accelerators, storage and remote memory. . .
  - Would 32k or more network endpoints be unrealistic?

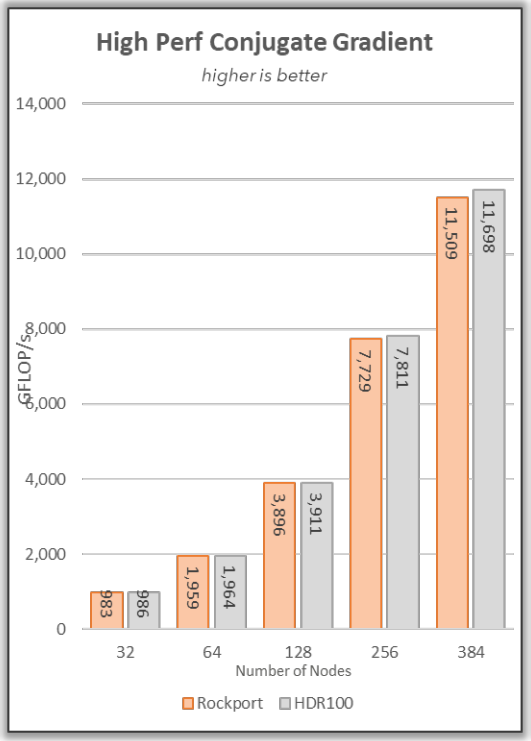
WHERE ARE WE ON DISAGGREGATION?

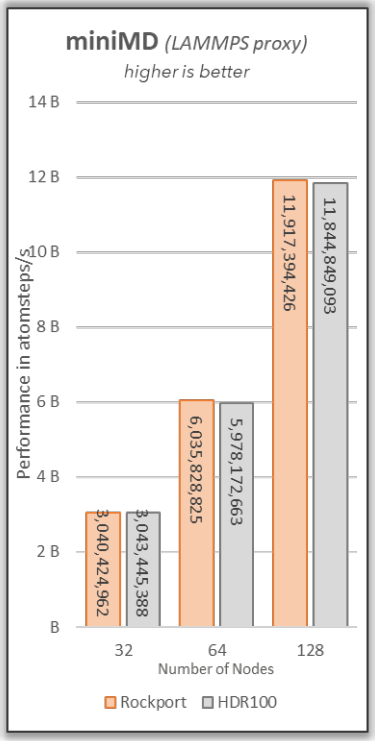
- Testbeds:
  - Giga-IO: Lonestar-6
  - Liqid - Chameleon (UT/Argonne)
    - Faster (TX A&M)
  - Fungible? (coming soon)
- We also have a pretty big Rockport testbed
- And a DPU testbed
  - Those 2 may play a role in future versions of composability



Rockport Testing @TACC







Rockport provides equivalent uncongested performance to HDR100

- Results from Rockport's Center of Excellence at TACC on Frontera
- Performance is consistent, predictable
- Performance is equivalent to IB HDR100, typically performs better under load
- These results do not yet take advantage of Rockport's advanced capabilities

April 2022 Texascale event
CONUS dataset from 2005-06-04_06:00:00 to 2005-06-04_09:00:00

July 2022 Texascale event

August testing activities
536,870,912 Atoms

rockport.

OUR HISTORY IN PCI-E DISAGGREAGATION

- Wrangler, proposed in 2013 and in production in 2015, used DSSD NAND Flash arrays connected over an external PCI fabric to the compute machines.
  - 96 servers had about 0.5PB of shared flash (then 20PB of shared disk back end).
- Note, this machine was deployed *prior* to the release of the NVMe standard, and *years* prior to the NVMe-OF standard (2017).
- The Wrangler configuration gave us unprecedented IOPS capability, and a lot of interfaces (Hadoop briefly mattered in 2013-2015. . .).

LESSONS LEARNED FROM WRANGLER

- The dev cycle/supply chain for PCI switch chips is not like server-class processors that make $Billions.
  - The last of our original spec'ed components for 2014 delivery I believe finally shipped in 2020.
    - Obviously, we re-engineered with much smaller switches, that limited the connection of a single server to arrays of 4 DSSD devices.
- All the promised interfaces worked as advertised. We presented to users:
  - A "normal" parallel POSIX Filesystem
  - HDFS
  - DB services (mostly PostgreSQL)
  - The Object store API.

LESSONS LEARNED FROM WRANGLER

- As you might have guessed:
  - 90%+ of users *only* used the Filesystem interface.
    - In the best case, an end user workflow got 12x faster, but that wasn't the norm.
  - The Database interface got the best acceleration, which mattered for all 3 users that relied on DB performance in scientific workflows.
  - HDFS performance was better than anything else, but the rest of Hadoop sucked so much it hardly mattered.
  - Zero end-user applications built on the API.
- So, users saw some benefits, but:
  - For most, only when it was ***completely transparent*** to all aspects of their workflow
  - For many, fixing I/O bottlenecks only resulted in limited improvement without fixing all the *other* bottlenecks exposed.
- Cabling was a nightmare (400 thick PCI cables in addition to the usual ethernet and IB, in a 96 node system).
- Commodity technology (namely NVMe) caught up and made the advantage less relevant over time.
  - Though we have all-flash filesystems in every system now.
- The retail price of a "boutique" solution ultimately outweighed the performance advantage (not that we paid that).

COULD DISAGGREGATION MAKE THINGS BETTER?

- YES
- Picking system configs is among our hardest and most important tasks
  - We have all kinds of workloads
  - We have limited ability to push software changes to our users.
- Right now, we tend to put a massive amount of hardware in one homogeneous partition (Frontera -- 8,400 CPU compute nodes) and much smaller amounts in specialized subsystems (also Frontera – 16 large mem nodes, 90 quad-GPU nodes).
- Often, load conditions are such that some subsystem has idle capacity while others have wait times – this is obviously not the *best* possible thing.
  - *Caveat* -- Deep Learning Workloads are still essentially immature and all over the map. Some are limited by the shared GPU address space. It is possible (likely) this is an artifact of the tools and not the algorithm… actually, let's take a 2 slide detour on that, because it's a relevant lesson. . .

THE "HOW MANY GPUS PER NODE DEBATE"

- Also known as "Welcome to problems we solved in the 1990s, but used different words, so the AI guys must figure it out again on their own".
- Once upon a time, there was a huge debate about whether we should build giant shared-memory machines, or lots of smaller machines in distributed memory clusters.
- In the Shared Memory camp, there were many, many players:
  - Silicon Graphics Data General 8-way Itanium servers
  - Unisys Convex
  - Honeywell Sequent
- In the Distributed Memory camp, there were companies **who aren't currently dead.**
- So we know how that came out. Why?

THE "HOW MANY GPUS PER NODE DEBATE"

- Shared memory systems are inherently easier to write software for, and easier to optimize/performance tune.
  - So all application writers would like them.
- At the same time, making the buses and especially cache coherency protocols gets exponentially more expensive as you scale up the number of processing elements.
  - And if you sacrifice that, and go heavy-NUMA "distributed shared" memory, those same programmers can fall in all kinds of performance traps without ever knowing why.
- So, the 10x cheaper hardware wins out over the 2x more efficiency and ease.
- And, it turns out, much to our surprise, a huge amount of the algorithmic space can be formulated to use distributed memory approaches.
- Horovod, or some other approach to model parallelism will come along, and we will wonder why we ever built $100k+ super-GPU nodes.

DOES IT ACTUALLY WORK?

- YES
- With both of our PCI fabric evals, and with our look at Rockport, we can verify that things function like they are supposed to – we can compose nodes, even without rebooting (!!!!), and make things appear as single large nodes.

MY TAKE ON USE CASES

- OK, there are lots of things we can disaggregate (and each will have its own value proposition):
  - Accelerators – typically GPUs, but really any PCI compute device (FPGA, IPU, Vector Engine, AI Accelerator, etc.).
  - Storage – pool remote storage devices into locally appearing block devices or filesystems (aka Wrangler).
  - Memory – Use either PCI-attached memory devices (really, CXL) or memory from a remote node.
- Keep in mind current PCI implementations are a waystation on the way to CXL (etc.) kinds of future fabrics
- Let's dive into these separately. . .

MY TAKE ON USE CASES

- Storage – Dynamically composing storage is great; but do we need PCI/CXL level latency?
  - If not, we could do this over our conventional fabrics (NVMeOF).
  - Better software layers are needed, but roughly the entire storage industry is working on this.
  - Lessons of the past – BW/IOPS are the driver, not latency, so we probably don't need a PCI fabric for this right now.
- Remote memory
  - Here the opposite is true – we can see huge differences going from L2 to L3 cache in application performance. Latency is what matters most when finding memory in a NUMA system!!
  - CXL may bring this down some.
  - I am somewhat skeptical we will ever see great performance here.
  - *But*, in a small fraction of our systems, we have ridiculously inefficient largemem nodes, because sometimes, you just need the answer.
  - So this is probably more of a niche use case, but I'd want to have it on, say, 5% of my nodes.

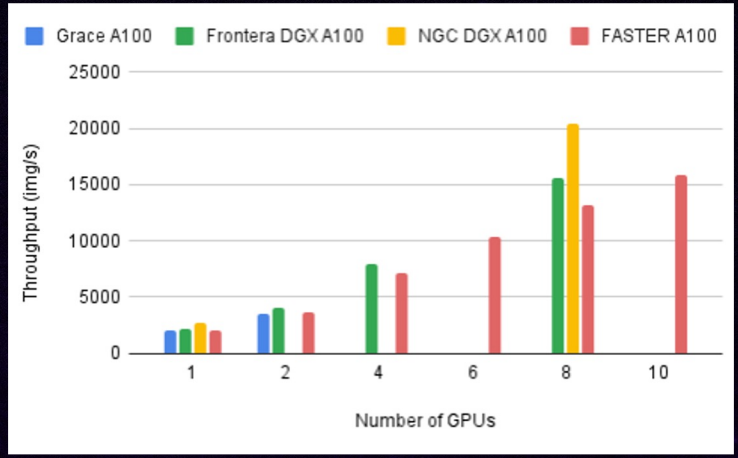
MY TAKE ON USE CASES

- Accelerators
  - If there is to be a “killer app” for composability, it’s probably accelerators.
  - As previously noted, for better or worse, the current state of DL software is “fit in the address space of the GPUs on one node”.
  - >4 GPU nodes carry a premium price.
- I’ll focus the rest of the slides on the accelerator (really, GPU) use case only.

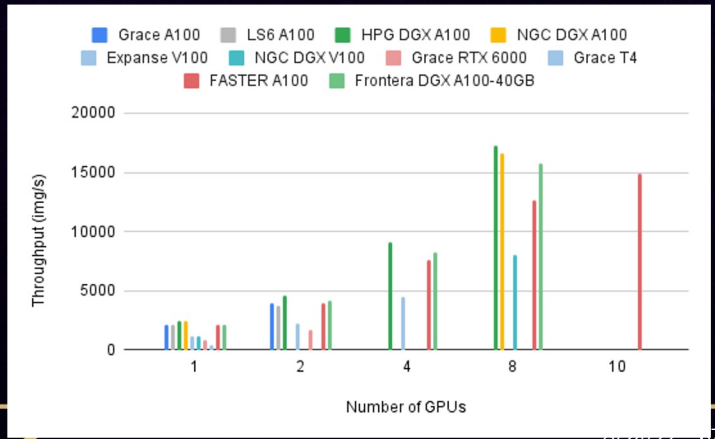
DOES IT PERFORM?

- YES, pretty well.
- Again, at least for accelerators.
- Comparisons with LIQID, all A100 GPUs, on traditional machines, DGX with 8 GPUs.
- ~80% of tuned DGX performance at 8 GPUs, scaling out to 10 GPUs.

RESNET over TensorFlow



RESNET over PyTorch



CAN WE MAKE IT USABLE?

- YES
- While algorithms for scheduling are still fun, we can provide basic Slurm integration, do the orchestration for users (transparency!), and run jobs. So, cool.
- Still messing with Kubernetes a bit, but we also have used OpenStack successfully, no reason to believe K8s won't work too (hey, it's probably the *primary* use case).
- We can also manage the physical/install rack layout stuff, so no usability barriers to introduce this.

DO THE ECONOMICS WORK?

- And here is where it gets interesting – it works, and is worth something – how much???
  - i.e. can it be sold profitably at good value for enough use cases?
  - Getting this wrong has sunk many a promising technology/company.
  - Still some work to do here.
- True Facts:
  - GPUs are expensive
  - CPUs in GPU nodes can be underutilized resources.
  - Different codes need different size nodes, and are not particularly malleable.

DO THE ECONOMICS WORK?

- More true facts:
  - GPUs are expensive. You have to buy them either way. (Though high counts per node cost non-linearly more – see SMPs).
  - CPUs in GPU nodes are underutilized, but are a tiny fraction of overall node performance.
  - Software/workloads can slowly change over time.
  - This *can* be a *third* fabric you are incorporating into a system. Which most users will never notice, but inevitably will have to be debugged at some point.
  - HPC people never pay list price (HPC=Half Price Computing)

SO THAT MEANS. . .

- OK, so it seems fair to say, if you do some more experiments, that unless you can fill every node type all the time, you will see utilization improvements. Sometimes small, sometimes large, but maybe 15% for a largely heterogeneous system.
- There are many confounding factors:
  - What if you can run a bigger job (e.g. 10 GPU) than you could before – what is that worth?
  - What if we could *replace* one of the fabrics with the PCI/CXL fabric – e.g. not have infiniband in every node?
    - Tough in our "little oversubscription" environment, but the IB/OPA network typically is 15% of our system cost (HCA, ports, cables).

SO, WHERE ARE WE NOW?

- On LCCF, I don't think it's *likely* that we will use disaggregation across *all* the system, though it may fit in some niches.
- Our look at workloads means we probably still won't have much oversubscription in our fabric – so replacing one of the traditional interconnects with CXL/NVLINK/PCI Fabric/Whatever probably doesn't make a lot of sense (different workloads will have different answers on this. . . What if you only need 1Tb/s into a rack?).
- Our decision on our "fast" fabric is likely not going to come down to *only* technical factors among IB/OPA/Slingshot/(Rockport or other emerging) due to conditions in the markets.
- We will need a great MPI implementation on that fabric regardless!

THANKS!!

- **The National Science Foundation**
- The University of Texas
- Our many vendor and university partners.
- The MVAPICH Team!!!!
- **Our Users – the thousands of scientists who use TACC to make the world better.**
- All the people of TACC


FRONTERA
TACC
NSF
TEXAS