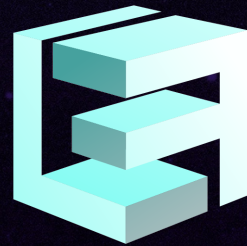




FRONTERA USER'S MEETING LCCF PLANS OVERVIEW

Dan Stanzione
PI, Frontera project
January 2021

VISION



LEADERSHIP-CLASS COMPUTING FACILITY

- ▶ Obviously, a more capable follow-on to Frontera (current code name: Horizon)
- ▶ A more holistic, long term, and collaborative view of how we support “leadership applications”.
- ▶ An NCAR-like leader and anchor for the NSF computational science and engineering community, existing in the context of other NSF and University investments in research computing.
- ▶ A broader view of HPC, with associated systems and services:
 - ▶ Simulation, Analytics, AI, of course.
 - ▶ Instruments/Edge/IoT
 - ▶ Interactive, Urgent, Automated, and Batch
 - ▶ Data Lifecycle and Reproducibility
- ▶ Workforce Development for a diverse technology and science community of researchers
- ▶ Robust Public Outreach

A FEW CRITICAL ACRONYMS

- ▶ LCCF – Leadership Class Computing Facility
- ▶ MREFC – Major Research Equipment and Facilities Construction fund.
- ▶ LFO – Large Facilities Office
- ▶ MFG – Major Facilities Guide
- ▶ CD(R) – Conceptual Design (Review)
- ▶ PD – Preliminary Design
- ▶ FD – Final Design
- ▶ Construction – the phase when MREFC funds are used to *construct* the facility, after FD and before operations (restrictions on activities).
- ▶ Operations – Transition back to CISE directorate funds, when Construction is deemed complete

CONCEPTUAL DESIGN

- ▶ Per the Frontera Solicitation "...10x"
- ▶ MREFC means 3 stage gates, CDR, PDR, FDR prior to construction, then transition to operations
 - ▶ Targeting 2023 construction, 2025 operations
- ▶ We have completed conceptual Conceptual Design.
 - ▶ Given that hardware procurement is still 3+ years out, picking specific technologies would be a bad idea

TIMELINES

- ▶ Pandemics and budgets and politics could re-arrange things (and we can deal with that), but written plans need a written target.
 - ▶ CD – Began August 2019, concluded September 2020.
 - ▶ PD – Begins October 1, 2020, ends December 2021
 - ▶ FD - Begins January 2022
 - ▶ Construction – Tentative late 2023
 - ▶ Operations -Tentative mid 2025

CONCEPTUAL DESIGN FOR LCCF

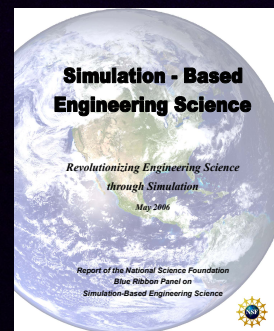
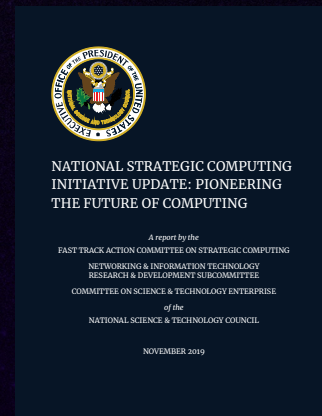
- ▶ **Civil Infrastructure:** +20MW datacenter, 50 offices, public outreach space, associated cooling and utilities
- ▶ **Computing Environment:** 10x primary compute system, interactive/services computing system, fast local scratch pools
- ▶ **Data Environment:** Backing store for fast scratch (think /work), data publication system/services, archive system
- ▶ **Software and Support:** Code improvement/adaptation efforts, all the usual support services
- ▶ **Education and Public Outreach:** Professional training, Fellows and curriculum, K-12, public/congressional outreach.

COLLECTING SCIENCE REQUIREMENTS

- ▶ Fortunately, we aren't the first to think about building a computer to do scientific work.
- ▶ So our process included:
 - ▶ Mining the vast number of reports out there in the community
 - ▶ Directly gathering requirements from stakeholders
 - ▶ Meetings we hosted, many meetings we attended
 - ▶ Gathering data on what actual workloads look like.

REQUIREMENTS – WHAT'S OUT THERE (1)

- ▶ There are a number of national and NSF reports that lay out the case for expanding HPC Investments, from still valid classics of a decade ago to recent updates in the last 12 months.



REQUIREMENTS – WHAT'S OUT THERE

- ▶ There are many, many more papers and workshops in individual scientific disciplines that outline grand challenges – most all of which have computational and data aspects
- ▶ (You can't find one that *doesn't* have computational challenges).








Review | [Open Access](#) | [Published: 07 February 2020](#)
Eleven grand challenges in single-cell data science
[David Lähnemann, Johannes Köster, \[...\] Alexander Schönhuth](#) 
Genome Biology, 21, Article number: 31 (2020) | [Cite this article](#)

Mapping the human brain: comparing the US and EU Grand Challenges†
[Dolores Modic](#) , [Maryann P. Feldman](#) [Author Notes](#)
Science and Public Policy, Volume 44, Issue 3, June 2017, Pages 440–449,
<https://doi.org/10.1093/scipol/scw085>

AMS Grand Challenges in Big Data and Earth Sciences



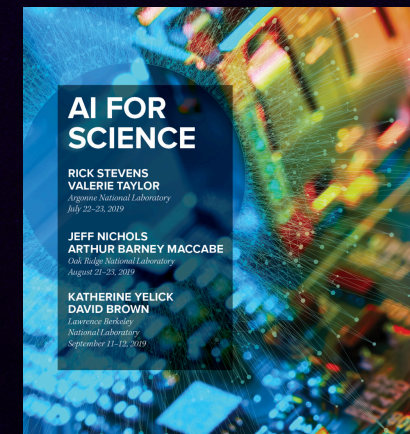
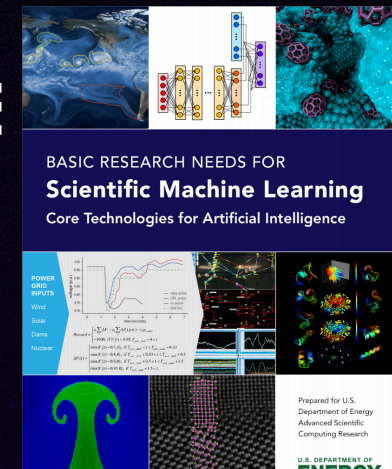
EDITOR'S CHOICE
Grand Challenges in Comparative Physiology: Integration Across Disciplines and Across Levels of Biological Organization 
[Donald L. Mykles](#) , [Cameron K. Ghalambor](#), [Jonathon H. Stillman](#), [Lars Tomanek](#)
Integrative and Comparative Biology, Volume 50, Issue 1, July 2010, Pages 6–16,
<https://doi.org/10.1093/icb/icc015>
 Published: 21 April 2010

 **Procedia Computer Science**
 Volume 108, 2017, Pages 1811–1812
Multiscale Modelling and Simulation, 14th International Workshop
[Derek Green](#) , [Bartosz Bosak](#) , [Valeria Kzhibzhonovskaya](#) , , [Alfons Hoekstra](#) , [Petros Koumoutsakos](#) 

Physics > Atmospheric and Oceanic Physics
 December 21, 2018
Confronting Grand Challenges in Environmental Fluid Dynamics
[T. Dauvois](#), [T. Pascoot](#), [P. Bauer](#), [C.P. Caulfield](#), [C. Cenedese](#), [C. Corlé](#), [G. Haller](#), [G.N. Hey](#), [P.F. Linden](#), [E. Melburg](#), [N. Pinardi](#), [A.A. Sepp](#), [Neves](#), [N.M. Vriend](#), [A. Woods](#)
 Environmental fluid dynamics underlies a wealth of natural, industrial and, by extension, societal challenges. In the coming decades, as we strive towards a more sustainable planet, there are a wide range of grand challenge problems that need to be tackled, ranging from fundamental advances in understanding and modeling of stratified turbulence and consequent mixing, to applied studies of pollution transport in the ocean, atmosphere and urban environment. A workshop was organized in the Les Houches School of Physics in France in January 2018 with the objective of gathering leading figures in the field to produce a road map for the scientific community. Five subject areas were addressed: multiphase flow, stratified flow, ocean transport, atmospheric and urban transport, and weather and climate prediction. This article summarizes the discussions and outcomes of the meeting, with the intent of providing a resource for the community going forward.

REQUIREMENTS – WHAT'S OUT THERE

- ▶ The AI literature is less “settled”, but the role of AI in Science is getting better defined:
 - ▶ DOE AI Town Halls
 - ▶ Scientific ML Report
 - ▶ NSF Workshop on Smart CI (Not yet published)
 - ▶ ASCAC Subcommittee on AI (Not yet published)
- ▶ A quick takeaway:
 - ▶ The demand for training could swamp our current workloads
 - ▶ Reduced precision will play a bigger role
 - ▶ Inferencing in science could take off – but training new datasets for better inferencing a lot fits the scientific mindset way too well (so “inference-only” hardware might be less useful at the datacenter than at the edge).



REQUIREMENTS – TALKING TO THE COMMUNITY

- ▶ While a lot is out there, we spent a fair amount of the year talking to people about thoughts on the LCCF.
- ▶ The two big “events” were the January workshop, consisting of large users in the NSF community, and the BOF at SC – reports from these are in your documents
- ▶ We also had sessions with our project partners, with our on-campus advisory group, with some larger user groups (i.e. Coastal Modeling, FIRE Galaxy simulation), with leaders at other centers, and presentations to CASC, CARCC, XSEDE SPF, etc.
- ▶ Plus our users from Stampede2, Frontera, Wrangler, Chameleon, etc.
- ▶ We also attended/hosted a lot of related meetings.
 - ▶ AI in Natural Hazards, Smart CI, ASCAC hearings, etc, etc, etc.

REQUIREMENTS – GATHERING TECHNICAL INFORMATION

- ▶ We have analyzed both published and in-house data about actual utilizations on large scale HPC systems:
 - ▶ Blue Waters
 - ▶ XSEDE
 - ▶ TACC
 - ▶ NERSC/INCITE
- ▶ While demand is always high and keeps increasing, it's nature is remarkably static on decade+ scales.

Blue Waters*
4/2013 – 9/2016

Stampede**
CY2016

Stampede 2**
CY2019

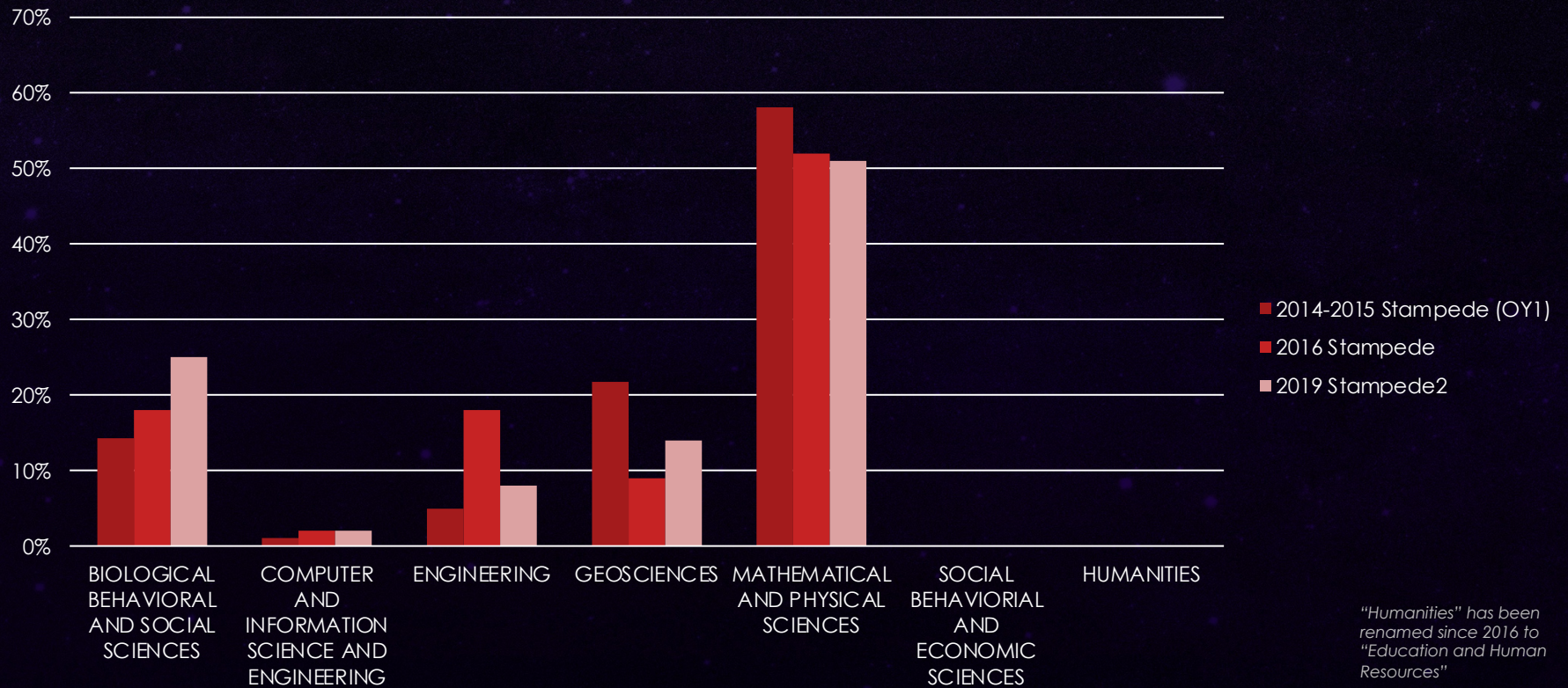
Frontera**
9/2019 – 5/2020

Area	Usage	App	Usage	Area	Usage	App	Usage	Area	Usage	App	Usage	Area	Usage	App	Usage
Quantum Physics				Quantum Physics	17.90%	VASP nemo QE	16.60% 0.70% 0.70%	Quantum Physics	21.60%	VASP BerkeleyGW QE	14.60% 6.20% 0.80%	Quantum Physics	5.40%	epw.x yambo	4.30% 1.20%
Molecular Dynamics	27.9%	NAMD Amber Gromacs	18.0% 7.6% 3.6%	Molecular Dynamics	12.70%	NAMD Gromacs LAMMPS	6.00% 4.00% 2.70%	Molecular Dynamics	15.80%	Gromacs LAMMPS NAMD	7.40% 7.20% 1.30%	Molecular Dynamics	16.00%	LAMMPS NAMD Gromacs	8.40% 6.00% 1.50%
Astrophysics	9.5%	Cactus ChaNGa	5.5% 4.1%	Astrophysics	5.30%	athena orion2 Cactus	3.00% 1.60% 0.60%	Astrophysics	4.60%	athena	4.60%	Astrophysics	24.40%	SpEC harm3d Cactus GIZMO	12.20% 8.80% 2.40% 1.10%
CFD	4.1%	psdns	4.1%	CFD	1.60%	fun3d	1.60%	CFD	3.40%	nek5000 ppm_vortex DNS2d	1.50% 1.30% 0.70%	CFD	17.00%	cdns.x	2.10%
Env. Sciences	2.9%	CESM	2.9%	Env. Sciences	5.70%	WRF ADCIRC ARPS	3.10% 1.40% 1.20%	Env. Sciences	7.90%	WRF SpecFEM3D CESM	5.80% 1.20% 0.90%	Env. Sciences	12.90%	CESM SWWF BASTRUS DNS2d	7.60% 4.40% 1.00% 4.30%
QCD	21.3%	Chroma MILC Hisq	9.8% 8.0% 3.5%	QCD	1.40%	qlua	1.40%	QCD	3.60%	Chroma	3.60%	QCD	4.00%	Chroma ks_spectrum*	2.10% 1.90%
Geophysics				Geophysics	0.80%	SpecFEM3D	0.80%	Geophysics	0.80%	SpecFEM3D	0.80%	Geophysics	1.50%	RSQSim	1.50%

*Limited to sample of the top 10 BW applications

**Top 20 applications truncated to exclude 0.6% use

Utilization by NSF Directorate



SCIENCE REQUIREMENTS

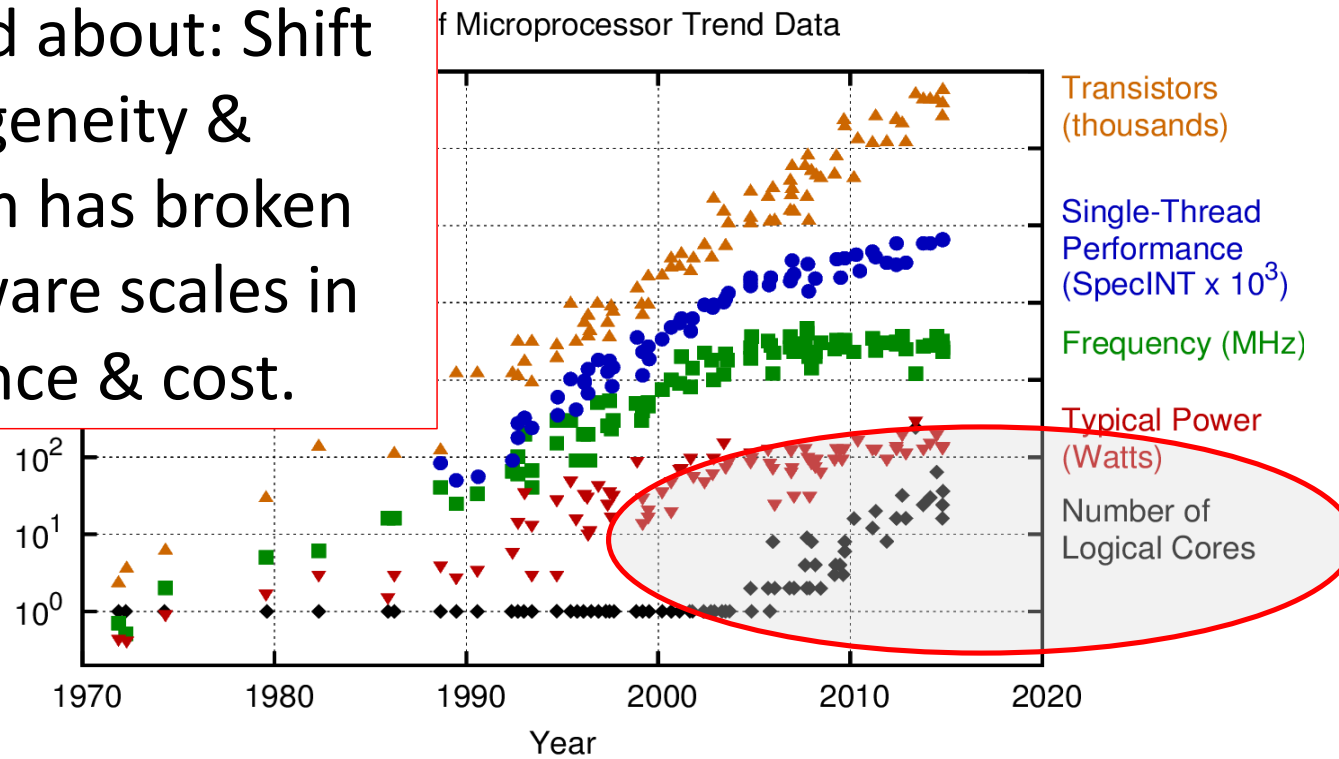
- ▶ Numerous individual science tasks have demand for vastly greater computational scale and time; as such, a 10x baseline improvement in application performance should be considered a minimum for the Facility
- ▶ Much of the work of extracting science and engineering results from the workload happens *outside* the main simulation or analysis run, and is done in analysis of the produced results later, over a much longer time. Support for this “expanded” workflow at the appropriate scale is therefore critical for the facility, including throughout the data lifecycle.
- ▶ The workload is evolving, with increased AI/ML focused workloads, and increased emphasis on throughput at scale – this work is in addition to, rather than replacing simulation at Scale.
- ▶ Gradual evolution of code is possible, with proper incentives and sustained investments. Radical or rapid change to software in order to support new hardware is hard; the transition may outlive the hardware.

COLLECTING TECHNICAL REQUIREMENTS/ CAPABILITIES

- ▶ Endless meetings with vendors, large and small
- ▶ Meetings with colleagues at the DOE, internationally, etc.
- ▶ More conferences, committees and reports (e.g. AI for Science Town Halls)
- ▶ Extrapolation from large scale systems current and future
- ▶ Talking to the CI community (not only for technical, but workforce requirements)
- ▶ Start of prototyping. . .

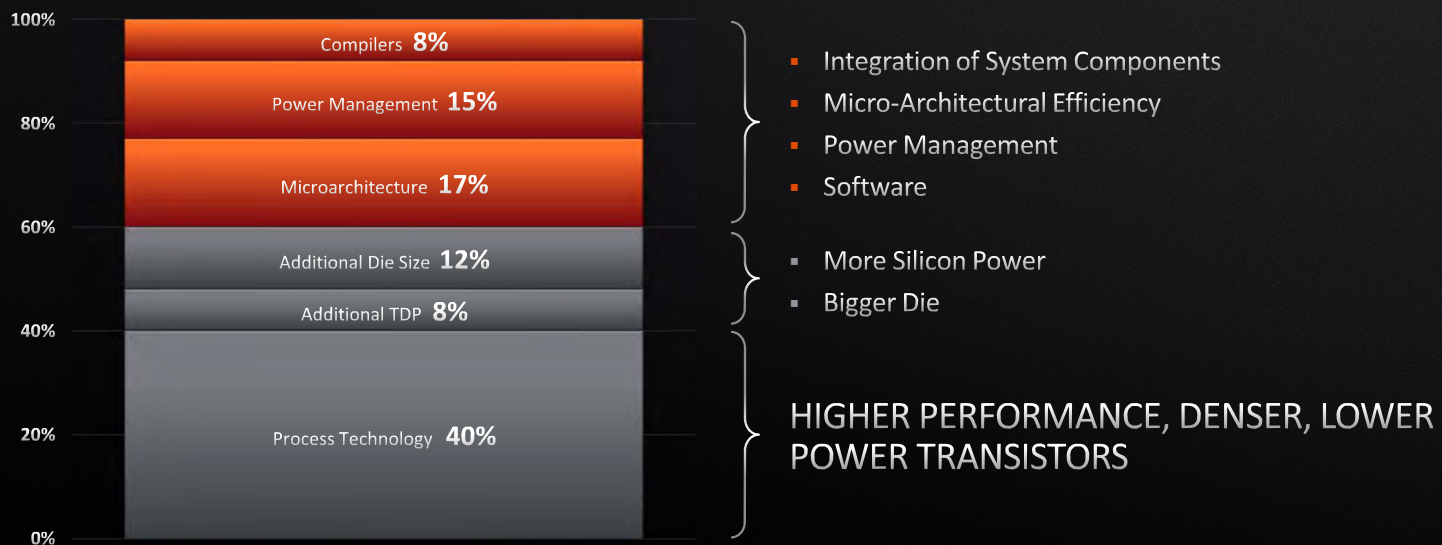
Not News: End of Decades of Moore's Law scaling

Less talked about: Shift to heterogeneity & parallelism has broken how software scales in performance & cost.



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

PERFORMANCE GAINS OVER THE PAST DECADE



ELEMENTS OF 2x IN 2.5 YEAR PERFORMANCE GAIN OVER THE PAST DECADE

A FEW THINGS WE CAN RULE OUT

▶ Quantum

- ▶ We are investing in quantum *people*, and in running the various simulators on the system.
 - ▶ Replacing those we've lost to Quantum startups. . .
 - ▶ We are investing in the Atos quantum simulation box with Stanford.
 - ▶ The time is here to invest in algorithms, programming models, etc., but,
 - ▶ We see no chance quantum is a mainline scientific computing technology in 2024, though it may have a few niches where it is effective commercially by then.
- ▶ The more exotic neuromorphic devices are also probably greater than 3.5 years out.
- ▶ The "AI chips" are probably viable, but reduced precision makes it tough to make them the *primary* capability of any general-purpose computing system. More data still needed.

SCIENCE REQUIREMENTS AND TECH LANDSCAPE DRIVE ENGINEERING REQUIREMENTS:

- ▶ Power/Cooling requirements for LCCF-1 ******will fall between 18-25MW******. LCCF-2 should exist within the same power envelope.
- ▶ The LCCF system should target a higher multiple on memory bandwidth versus the baseline system than FLOPS multiple
- ▶ Capability should exist for interactive and visualization workflows, and support for the data lifecycle at scale beyond core simulation.
- ▶ Aggregate throughput of the system is as important to science productivity as the single job peak.
- ▶ Continuity in programming model required for science productivity.
- ▶ **Substantial support for application and library teams both prior to start of production and during system life is required for success.**

WHAT CAN WE SAY ABOUT COMPUTING SYSTEM FOOTPRINT AND PERFORMANCE

- ▶ Pick the ends of the architecture spectrum, make NDA-informed extrapolations.
- ▶ Take what we know about performance, and scale to 10x.
 - ▶ Establish a baseline density, price, and footprint.

COMPUTING ENVIRONMENT GETTING TO 10X

- ▶ The basic assumptions we are making:
 - ▶ Over 5+ years from Frontera, we will get roughly 3x improvement per node in “effective” performance (Moore’s law would imply 8x).
 - ▶ We actually think it would be worse, except for expected improvements in memory BW.
 - ▶ We will simply put in twice the \$\$\$ of Frontera to get another 2x in improvement, which gets us a 6x faster system (headline peak will probably be 10x).
- ▶ The remaining ~1.5x will have to happen from improvements in software, algorithms, or methods (i.e., use of surrogate models).

COMPUTING ENVIRONMENT

- ▶ Carrying two options forward at CDR (though the likely result will be neither, but more of a hybrid).
 - ▶ A “pure CPU” option, linearly extrapolated from Frontera
 - ▶ An “all accelerated option”, extrapolated from Longhorn.
- ▶ Though one is 4x the node count of the other, conveniently they are the same power and number of cabinets, and roughly the same cost.
 - ▶ (Roughly speaking, a 4,000 watt \$35,000 node versus a 1,000 watt \$8,500 node).
 - ▶ 20 or 80 nodes per rack, 80KW a cabinet.
- ▶ Baseline sizing is 200 racks for the “main” compute system.

THE LCCF ECOSYSTEM - COMPUTE

- ▶ Compute Systems:
 - ▶ The “10x” Computing System
 - ▶ An additional system, at about 1/10th scale, for non-batch services (interactive, persistent)
 - ▶ Prototypes
 - ▶ Early access system
 - ▶ Continued prototypes after construction
 - ▶ Distributed Systems (extended capabilities at 3 other sites).
- ▶ (Initial power footprint ~18MW).

THE LCCF ECOSYSTEM - DATA

- ▶ Data Systems
 - ▶ We want to make changes we've been evolving towards in storage.
 - ▶ Both our "one giant scratch" model and the cloud "no persistent filesystem" model are broken.
- ▶ Proposed Hierarchy (4 tiers):
 - ▶ Per-project-ish solid state "scratch" volumes. (Minimal sharing – sized at 3x system RAM)
 - ▶ A "backing store", roughly equivalent to our current Stockyard (/work) filesystem.
 - ▶ Persistent, POSIX, a sort of longer-term scratch, mounted everywhere in the ecosystem.
 - ▶ From there, move data to one of two "permanent" tiers:
 - ▶ Publication - high integrity, public access, DOI and metadata support (evolve from Corral)
 - ▶ Archive – evolve from Ranch, likely still a tape tier behind significant spinning, focused on lowest-cost.
- ▶ Support for protected/data encryption across all tiers.

PROTOTYPING ACTIVITIES

- ▶ We have tested or acquired a number of prototypes in 2020 we will make more widely available in 2021
 - ▶ We added the NVDIMM subsystem to Frontera in 2020.
 - ▶ DAOS testing
- ▶ New in-house prototypes:
 - ▶ NVIDIA DGX Ampere A100 (2 nodes, x8 GPUS per node)
 - ▶ Fujitsu ARM
 - ▶ NEC Vector accelerator
 - ▶ AMD CPU/GPU
 - ▶ Quantum simulator (Stanford)
 - ▶ NextSilicon (Q2)
- ▶ External access, but still available to users:
 - ▶ NextSilicon
 - ▶ IBM Q network
 - ▶ Through Argonne (Cerebus, Groq).
- ▶ Expect more on this through the next two years. . .

IS SOFTWARE IMPROVEMENT REALISTIC TO EXPECT?

- ▶ What are potential sources?
 - ▶ Change Algorithms
 - ▶ Optimize code implementation
 - ▶ Improve system software
 - ▶ Runtime tuning on the system

1. CHANGE ALGORITHMS

► Manuela Campanelli, yesterday:

A new Multi-Patch Scheme for Accreting BBH + Jets

How do we efficiently simulate 10^7 - 10^8 cells for 10^6 - 10^7 steps?

- PatchworkMHD – Avara+ 2020 in prep
New software infrastructure for problems of discrepant physical, temporal, scales and multiple geometries.

Long term simulation covering the full domain with PWMHD, now 30 times our prior efficiency
Avara+2021, in prep

2. OPTIMIZE CODES

► Elias Most, Yesterday:

Efficient use of the HPC architecture

Skylake-X workstation

Elapsed Time: 105.481s **~30% run time speed up!**
 SPGFLOPS: 177.005 **On average: 20.5% peak performance**

Effective Physical Core Utilization: 97.5% (7,803 out of 8)
 Effective Logical Core Utilization: 94.4% (15,129 out of 16)
 Effective CPU Utilization Histogram

Memory Bound: 42.8% of Pipeline Slots
 Cache Bound: 15.2% of Cycles
 DRAM Bound: 36.7% of Cycles
 DRAM Bandwidth Bound: 31.2% of Elapsed Time
 NUMA % of Remote Accesses: 0.0%

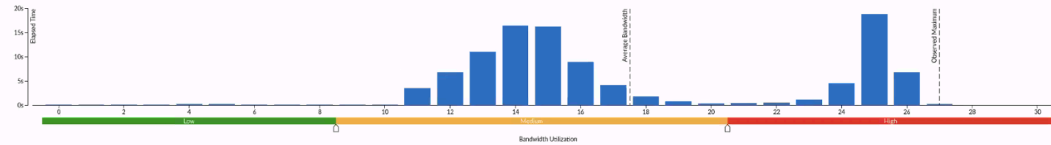
Memory limited: 42.8%!!

Bandwidth Utilization Histogram
 Capture bandwidth utilization over time using the histogram and identify memory objects or functions with maximum contribution to the high bandwidth utilization.

Bandwidth Domain: DRAM Capacity

Bandwidth Utilization Histogram

This histogram displays the wall time the bandwidth was utilized by certain values. Use sliders at the bottom of the histogram to derive thresholds for Low, Medium and High utilization levels. You can use these bandwidth utilization types in the Bottom-up view to group data and see all functions executed during a particular utilization type. To learn bandwidth capabilities, refer to your system specifications or run appropriate benchmarks to measure them; for example, Intel Memory Latency Checker can provide maximum achievable DRAM and Interconnect bandwidths.



Top Functions with High Bandwidth Utilization

This section shows top functions, sorted by LLC Misses that were executing when bandwidth utilization was high for the domain selected in the histogram area.

Function	LLC Miss Count
Loop at line 1043 in amrex:BaseFab::doReco[...]	8.7%
Loop at line 271 in amrex:MultiFab::Saxpy_omp_k[...]	6.0%
Loop at line 582 in amrex:FabIter::amrex::FabIter::ParallelCopy_omp_k[...]	5.3%
Loop at line 342 in amrex:MultiFab::SetConst_omp_k[...]	4.4%
Loop at line 1704 in amrex:FabArray::amrex::FabArray::mult::mult::enable_0[...]	3.3%
Global	1.5%

Memory access dominated by AMReX routines!

FPU Utilization: 5.1%

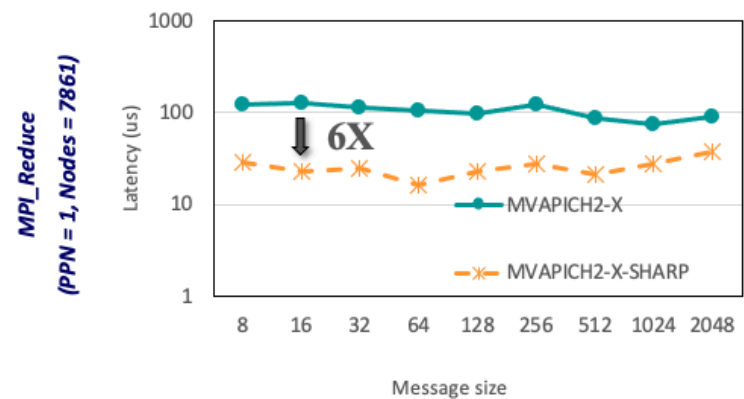
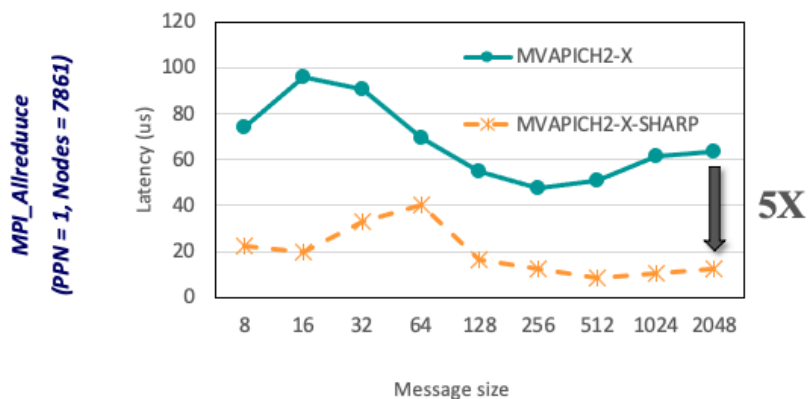
SP FLOPs per Cycle: 3.022 Out of 64
 Vector Capacity Usage: 99.2%
 FP Instruction Mix:
 % of Packed FP Instr.: 99.3%
 % of 128-bit: 0.0%
 % of 256-bit: 1.5%
 % of 512-bit: 88.4%
 % of Scalar FP Instr.: 0.1%

AVX512: 99.1% vectorization!

3. SYSTEM SOFTWARE IMPROVEMENTS

► Hari Subramoni, yesterday:

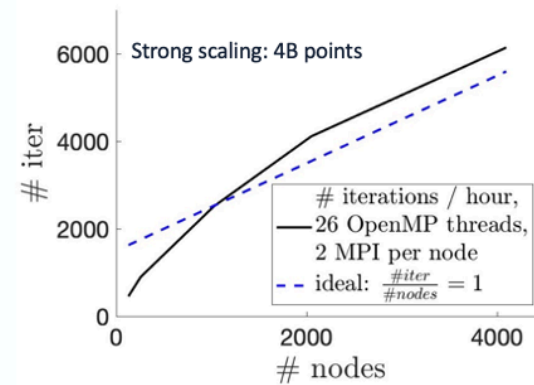
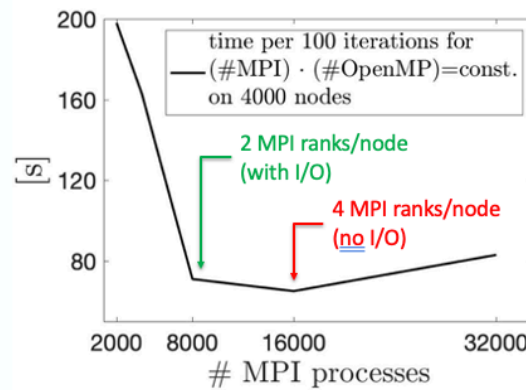
Performance of Collectives with SHARP on TACC Frontera



4. TUNING FOR THE SYSTEM

► Dan Bodony, yesterday:

Performance on Frontera



TRANSLATING SCIENCE REQUIREMENTS TO ACCEPTANCE

- ▶ The exact workload of the machine, while it can be categorized by field of science, etc, is somewhat unknowable.
- ▶ We know enough about *some* of the problems that we can provide useful baselines to determine if the system is “what is promised” – scientifically, performance wise, etc.
- ▶ Given the nature of the evolution of science right now, a complete set of “traditional” benchmarks makes less sense at 4 years distant, for HW that will be used over the next 9-10 years.
- ▶ We will propose a set of *problems* the machine will address (i.e., simulate the galaxy at 10x the current resolution).

CHARACTERISTIC SCIENCE APPLICATIONS

- ▶ The new solicitation is a chance to not only help us solve our benchmarking issue, but also to begin to meet our requirements to *co-evolve* with large user application teams.
- ▶ We want a *problem* to solve (not just a code to scale, though that may be the gist of how we do it).
 - ▶ The problem can (should?) be one that we can't do on current machines.
 - ▶ It can be an end-to-end workflow, not just a single core simulation run (ensemble, data processing, etc.).
 - ▶ We do need to know what the current state of the code(s) are, can we describe the inputs and outputs, etc.
- ▶ Evaluating on Significance, Suitability, and Representation.

THE CSA PROCESS

- ▶ Application to be considered, open NOW thru Feb. 26th.
 - ▶ <https://lccf.tacc.utexas.edu/application-partners/>
 - ▶ This is a very short (~900 word) application to sign up for a more in-depth *evaluation* with our team.
- ▶ Post-evaluation, we will construct a full application to be considered at a review in ~May.
- ▶ An NSF review will examine the full slate of selected applications in ~June
- ▶ Then funding will begin (~1 TACC FTE, ~1 CSA team FTE) summer 2021 (10-15 teams)
 - ▶ Continued for a second year with sufficient progress.
- ▶ Downselect to most successful ~8 teams to continue funding in construction phase late 2023-2025
 - ▶ Up to 4+ years total funding
- ▶ The problems that move forward in the construction phase will become part of the acceptance suite for the new machine.

THE CSA PROCESS

- ▶ Remember, the CSA problems serve multiple goals:
 - ▶ To explain to NSF/OMB/Congress the unique value of this facility as we go through the appropriation process (Why is this the highest priority? What science will/won't happen if we build this?)
 - ▶ To serve as acceptance benchmarks, i.e. to prove that the facility does what we promised it would do.

THE LCCF WILL BE A *UNIQUE* FACILITY

- ▶ The NSF mission is much broader, in terms of S&E applications, than the DOE missions.
 - ▶ LCCF will have much wider use cases and capabilities than the DOE Exascale Facilities.
- ▶ The commercial cloud is much *less* focused on science capabilities and science relationships.
 - ▶ The LCCF will be much more of a “science cloud” than the cloud – in hardware, software, and personnel capabilities.

AND A UNIQUE OPPORTUNITY

- ▶ We believe this is a rare chance to significantly impact how NSF is serving the computational science and engineering community.
 - ▶ More than any single system grant, this could change the model and scale of that support.
 - ▶ And create a sustainable base for NSF to continue this.
- ▶ We are humbled and excited by the opportunity to plan this project.