BENCHMARKING SUPERCOMPUTERS IN THE POST-MOORE ERA

Dan Stanzione
Executive Director, Texas Advanced Computing Center
Associate Vice President for Research,
The University of Texas at Austin

Bench’19 Conference, Denver
November 2019
Let me first talk about the system we just accepted. . .
  - which means we did performance projections on a bunch of benchmarks
  - then ran all those benchmarks on the real machine to measure against our projections
  - then saw if the benchmarks effectively measured how useful the system would be in production.

And talk about what we did and didn’t learn from them, and what I’d like to see happen in the *next* system benchmarks.
FRONTERA SYSTEM --- PROJECT

- A new, NSF supported project to do 3 things:
  - Deploy a system in 2019 for the largest problems scientists and engineers currently face.
  - Support and operate this system for 5 years.
  - Plan a potential phase 2 system, with 10x the capabilities, for the future challenges scientists will face.

- Frontera is the #5 ranked system in the world – and the fastest at any university in the world.

- Highest ranked Dell system ever, Fastest primarily Intel-based system

- Frontera and Stampede2 are #1 and #2 among US Universities (and Lonestar5 is still in the Top 10).

- On the current Top 500 list, TACC provides 77% of *all* performance available to US Universities.
FRONTERA IS A GREAT MACHINE – AND MORE THAN A MACHINE

ACCESS LAYER
- Login Nodes
- Data Mover Nodes
- Gateway & API Nodes

HIGHSPEED INTERCONNECT INFINIBAND

COMPUTE SYSTEMS
- Primary Compute
  38.7PF Double Precision
  8,008 Xeon Nodes
- Liquid Submerged Subsystem
  4PF Single Precision
  360 NVIDIA Quadro RTX 5000 GPUs submerged in liquid coolant by GRC
- Longhorn Subsystem
  3.5PF Double Precision
  7PF Single Precision
  IBM POWER®-hosted system with 448 NVIDIA V100 GPUs

ROUTER NODES

COMMERCIAL CLOUD STORAGE
- Cloud VDI & Data Access
- RANCH ARCHIVAL SYSTEM
- Other TACC Resources

TACC
HOW DO WE BENCHMARK FRONTERA?

- To gain acceptance, we used a basket of 20 tests, including a suite of full applications, plus some microbenchmarks and reliability measures.
- We passed them all, but the results give some interesting insights into how we do and don’t measure systems, and what is going on architecturally.
Of our 20 numerical measures of acceptance, as outlined in the proposal and project execution plan (PEP), we are “past the post” on all 20.

This represents a mix of full applications, low level hardware performance, and system reliability.
From the solicitation:

- Use the SPP Benchmark
- Target 2-3x Blue Waters (at 1/3 budget) --- 6-9x performance improvement per $ vs. 7 years ago.

The SPP was defined in 2006... 13 years ago.

- Most of the codes still relevant (WRF, MILC, NWChem)
- Some are obsolete
- The *problem sizes* are no longer sufficient for measuring the full capabilities of the machine (though some still pushed us to ~5,000 nodes/250,000 cores).
## APPLICATION ACCEPTANCE TESTS

<table>
<thead>
<tr>
<th>Application</th>
<th>Acceptance Threshold[s]</th>
<th>Frontera Time[s]</th>
<th>% over Threshold</th>
<th>Improvement over Blue Waters</th>
<th>Threshold Node[#]</th>
<th>Frontera Node[#]</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWP-ODC</td>
<td>335</td>
<td>326</td>
<td>1.03</td>
<td>3.2</td>
<td>1366</td>
<td>1366</td>
</tr>
<tr>
<td>CACTUS</td>
<td>1753</td>
<td>1433</td>
<td>1.22</td>
<td>3.3</td>
<td>2400</td>
<td>2400</td>
</tr>
<tr>
<td>MILC</td>
<td>1364</td>
<td>831</td>
<td>1.64</td>
<td>9.5</td>
<td>1296</td>
<td>1296</td>
</tr>
<tr>
<td>NAMD</td>
<td>62</td>
<td>60</td>
<td>1.03</td>
<td>4.0</td>
<td>2500</td>
<td>2500</td>
</tr>
<tr>
<td>NWChem</td>
<td>8053</td>
<td>6408</td>
<td>1.26</td>
<td>3.8</td>
<td>5000</td>
<td>1536</td>
</tr>
<tr>
<td>PPM</td>
<td>2540</td>
<td>2167</td>
<td>1.17</td>
<td>3.6</td>
<td>5000</td>
<td>4828</td>
</tr>
<tr>
<td>PSDNS</td>
<td>769</td>
<td>544</td>
<td>1.41</td>
<td>2.8</td>
<td>3235</td>
<td>2048</td>
</tr>
<tr>
<td>QMCPACK</td>
<td>916</td>
<td>332</td>
<td>2.76</td>
<td>5.5</td>
<td>2500</td>
<td>2500</td>
</tr>
<tr>
<td>RMG</td>
<td>2410</td>
<td>2307</td>
<td>1.04</td>
<td>3.2</td>
<td>700</td>
<td>686</td>
</tr>
<tr>
<td>VPIC</td>
<td>1170</td>
<td>981</td>
<td>1.19</td>
<td>4.3</td>
<td>4608</td>
<td>4096</td>
</tr>
<tr>
<td>WRF</td>
<td>749</td>
<td>635</td>
<td>1.18</td>
<td>5.2</td>
<td>4560</td>
<td>4200</td>
</tr>
<tr>
<td>Caffe</td>
<td>1203</td>
<td>1044</td>
<td>1.15</td>
<td>3.2</td>
<td>1024</td>
<td>1024</td>
</tr>
</tbody>
</table>

Average runtime improvement vs. Blue Waters: **4.3**
APPLICATION IMPROVEMENT – PER NODE

- For these applications (with their associated caveats) per node performance is 8.5x Blue Waters
  - Which is better than we projected – yet still somewhat disappointing for the industry in a broad context.

<table>
<thead>
<tr>
<th>Application</th>
<th>Blue Waters Nodes</th>
<th>Frontera Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>AWP-ODC</td>
<td>2048</td>
<td>1366</td>
</tr>
<tr>
<td>CACTUS</td>
<td>4096</td>
<td>2400</td>
</tr>
<tr>
<td>MILC</td>
<td>1296</td>
<td>1296</td>
</tr>
<tr>
<td>NAMD</td>
<td>4500</td>
<td>2500</td>
</tr>
<tr>
<td>NWChem</td>
<td>5000</td>
<td>1536</td>
</tr>
<tr>
<td>PPM</td>
<td>8448</td>
<td>4828</td>
</tr>
<tr>
<td>PSDNS</td>
<td>8192</td>
<td>2048</td>
</tr>
<tr>
<td>QMCPACK</td>
<td>5000</td>
<td>2500</td>
</tr>
<tr>
<td>RMG</td>
<td>3456</td>
<td>686</td>
</tr>
<tr>
<td>VPIC</td>
<td>4608</td>
<td>4096</td>
</tr>
<tr>
<td>WRF</td>
<td>4560</td>
<td>4200</td>
</tr>
<tr>
<td>Caffe (BW GPU/Ftr CPU)</td>
<td>1024</td>
<td>1024</td>
</tr>
</tbody>
</table>
A FEW LOOKS AT THIS PERFORMANCE LEVEL

- If you consider the SPP applications representative:
  - Frontera has 3x the "SPP Throughput" of Blue Waters, despite 1/3rd the nodes.
  - 9x the “SPP Throughput per dollar” of Blue Waters.
  - 4.7x the “SPP Throughput per watt” of Blue Waters
THAT’S ALL GOOD, BUT. . .

- 50% of peak performance improvement is not captured across our application suite.
- How do the microbenchmarks stack up?
HPL COMPARISON

I don’t have per node benchmarking for Blue Waters on HPL, but let’s look at Stampede 1 from roughly the same area:

- Intel Sandy Bridge, 8 core, 2.7GHZ, dual-socket nodes
- (Frontera: Intel Cascade Lake, 28 core, 2.7Ghz, dual-socket).

On Stampede 1 (just the CPU part) we got about 90% of peak on a large run.

- Per node Peak: 345.6GF  System Peak: 2.2PF  HPL Peak: 2.1PF
TANGENT: A FEW WORDS ON HPL

- The “Golden Age” of Linpack was probably the Intel Sandy Bridge processor, when we could get 90% of theoretical peak on a large system.
- Since then, many systems have fallen to 60-65% of peak.
- Unfortunately, not only has % of peak fallen, the *definition* of peak has changed. . .
- The old way:  (Clock rate)*Sockets*(Core Count)*(Vector Length)*FMA* (# of Simultaneous issues)
  - Frontera: 2.7*2*28*8*2*2 = 4,834GF per node.
  - 8,008 nodes = **38.8 PF**
- That is the current official, peak performance – also a lie.
The headline clock rate is not the peak clock rate – which is much higher.

If you do the computations you used to do the math in peak (FMA, 512bit Vectors, two issues per cycle), there is no theoretical way to run at the nominal clock rate.

Clock speed is dynamic, based on power and thermal, and adjust independently on each socket, with a 1ms interval

On 16,016 sockets, on a 10 hour HPL run, there are 577,152,000,000 opportunities for the clock frequency to change on a processor

When you exceed a certain % of AVX instructions the chip hypothetically runs at the AVX frequency (for Frontera 1.8Ghz).
PEAK PERFORMANCE FALLACIES

- In reality, if you have thermal and power space, AVX instructions can run above AVX frequency – we observe 2Ghz most of the time.
- Then there is the other gaming you can do (that we don’t do) – i.e. lower the memory controller speed to free up more watts for AVX.
- If you computed % of peak on AVX frequency, it would be **25.8PF**.
- For obvious reasons, they will never market it this way, so “% of peak” has become another deceptive metric for how systems are tuned in the last ~4 years.

- We hit 22+ PF in the Top 500 – prior to applying a number of fixes to the system.
BACK TO OUR COMPARISON

- Per node Peak Flop Comparison (Frontera/Stampede1) : 4834/345 = 14
- Per node HPL 2.9TF/310GF = ~9
- So, HPL implies we’ve captured only around 64% of performance improvement
- Our Application Suite implies we’ve captured around 53% of performance improvement.

- FOR ALL OUR CRITICISM OF HPL, IT’S ACTUALLY A FAIR PREDICTOR OF APPLICATION IMPROVEMENT
  - And infinitely easier than developing representative test cases in 10 apps and tuning and running them all.

- I did not expect this result – we may give HPL way too hard of a time. . .
- Or possibly, our choice of applications sucks almost as much in almost the same ways!
WHAT IF WE USED THE SYSTEM PEAK INSTEAD OF PER NODE?

- Again, I don’t have the data I need for BW, but we can roughly guess what 22000 nodes of AMD Bulldozer would have peaked at.
  - We know that we had 3x the “SPP throughput” of Blue Waters.
  - Let’s assume BW (CPU only) would have had an HPL in that era close to the peak – let’s call it 8PF.
  - If we use the “theoretical peak” of 39PF for Frontera, this is 5x higher. But we get 3x, so again the implication is we only captured 60% of the peak performance improvement, broadly consistent with other measures.
  - However if we use the *AVX Frequency* peak of 26PF, the ratio is about 3x, which is what we got.
- So that means...
ANOTHER SURPRISING RESULT

- If we report the ratio of peak performance based on the *actual* frequencies of the chips, it turns out the peak ratio is *almost exactly predictive* of the application speedup.

- This tells me two things:
  - Damn, maybe we don’t need benchmarks at all (I’m still skeptical).
  - Maybe we haven’t actually lost anything in the architecture – that any perceived loss in code efficiency is a result of how we *market* performance.
AND THE COROLLARY

- If this is true – that we aren’t actually suffering a loss of performance due to architectural changes, but a loss versus how performance is marketed.
- We don’t really need big software changes to use future chips, but, we don’t really have 16x socket improvement over the last 4 years, we have more like 8x.
- And our progress in chips has slowed even further than we have feared. . .
SO WHAT ABOUT FUTURE BENCHMARKS?

Well, what will we run?
The nuclear pore complex serves as a gatekeeper, regulating the transport of biomolecules in and out of the nucleus of a biological cell.

To uncover the mechanism of such selective transport, the Aksimentiev lab at UIUC constructed a computational model of the complex.

The team simulated the model using memory-optimized NAMD 2.13, 8tasks/node, MPI+SMP.

Ran on up to 7,780 nodes on Frontera.

One of the largest biomolecular simulations ever performed.

Scaled close to linear on up to half of the machine.

Plan to build a new system twice as large to take advantage of large runs.
Mature HIV-1 capsid proteins self-assemble into large fullerene-cone structures.

These capsids enclose the infective genetic material of the virus and transport viral DNA from virion particles into the nucleus of newly infected cells.

On Frontera, Voth’s team simulated a viral capsids containing RNA and stabilizing cellular factors in full atomic detail for over 500 ns.

First molecular simulations of HIV capsids that contain biological components of the virus within the capsid.

The team ran on 4,000 nodes on Frontera.

Measured the response of the capsid to molecular components such as including genetic cargo and cellular factors that affect the stability of the capsid.

“State-of-the-art supercomputing resources like Frontera are an invaluable resource for researchers. Molecular processes that determine the chemistry of life are often interconnected and difficult to probe in isolation. Frontera enables large-scale simulations that examine these processes, and this type of science simply cannot be performed on smaller supercomputing resources.”

-Alvin Yu, Postdoctoral Scholar in Voth Group
Ab initio numerical simulations of quantum chromodynamics (QCD) help obtain precise predictions for the strong-interaction environment of the decays of mesons that contain a heavy bottom quark.

Compare predictions with results of experimental measurements to look for discrepancies that point the way to new fundamental particles and interactions.

Carried out the very initial steps in the shuffle for the Exascale-size lattice during Frontera large-scale capability demonstration.

16x larger problem than anything they had previously calculated.

Ran on 3,400+ nodes.

The capability demonstration showed that given sufficient resources the team can run an Exascale-level calculation on Frontera.

“In addition to demonstrating feasibility, we obtained a useful result. We are now in good position for a future exascale run. We have working code and a working starting gauge configuration file.”

- Carlton DeTar, University of Utah
• Simulated fluid-structure interactions relevant to hypersonic vehicles.
• Simulations replicated a companion experiment performed at NASA Langley in their 20-inch Mach 6 tunnel.
• Frontera runs used 2 MPI ranks per node (one per socket) and 26 OpenMP threads per MPI rank.
• Saw superlinear speedup on up to 2,000+ nodes by fitting into cache rather than fetching from main memory.
• Linear speedup up to 4,000 nodes.
The project's goal is to study the process of Convective Boundary Mixing (CBM) and shell mergers in massive stars.

The computational plan includes a sequence of brief three-dimensional simulations alternating with longer one-dimensional simulations.

Ran on 7,300+ nodes for more than 80 hours during Frontera large-scale capability demonstration.

Saw 588 GFlop/s/node — or 4 Petaflops of sustained performance — for more than 3 days!
ISSUES FOR NEXT GENERATION BENCHMARKS (MY OPINIONS ONLY)

- Let’s assume that peak performance is not actually sufficient for future systems
  - Who knows what future distortions may creep in.
  - Doesn’t really help us when we need to compare across architectures.
- We have benchmarks that measure things like memory balance (STREAM, HPCG), but they don’t tell us anything about sensitivity to system balance.
  - What can our benchmarks tell us about the right I/O ratio, or sensitivity to memory bandwidth, or if we can tune our architectures correctly independent of code?
  - Can we do this for ML/AI and Analytics as well as simulation? Can we even put simulation codes into classes?
- Will reduced precision break us all?
We are already tasked with designing the Facility that will replace Frontera, with 10x the capability (for some definition) in 2024.

What are the challenges that we will face?

What are the mix of computing, data, and human resources that will be required to tackle them?
THE LCCF BENCHMARK SUITE

- We need to build a new benchmark suite for phase 2
  - Replace the SPP
  - Be relevant to the science challenges of 2025-2030
  - 10x performance of Frontera is the fixed requirement
- This leaves a lot of room to maneuver
  - Though the Post-Moore’s law world makes this hard to achieve.
- What apps and workflows would you include???
Some thoughts/questions on how we might specify/measure a 10x improvement in <5 years.

- What if we fix input/output, but not code?
  - Require the same answer, but not the same implementation
  - Better algorithms/software improvements could count towards the target.
  - Replacing part of the computation with a surrogate model could count, if the accuracy was the same.
  - We could capture portability of code between architectures by measuring changes from the baseline vs. the target.

- What if we don't simply scale the problem *size*, but improve the uncertainty range?
  - Do Uncertainty Quantification through inverse methods – massively increases amount of computation, but makes scaling simpler.

- Should we include data movement and locality in a full workflow in the production target?
  - Should we include complex workflows that might involve Analytics + AI/ML + Simulation + Vis or other post-processing?

- I would really, really like to avoid trying to measure *human* productivity in defining the 10x
PLEASE GIVE ME YOUR THOUGHTS AND REQUIREMENTS

- Here, at SC, later, send me an email, give me a call...
- Come to the BOF next Thursday!
- dan@tacc.utexas.edu
Humphry Davy, Inventor of Electrochemistry, 1812

(Pretty sure he was talking about our machine).

Nothing tends so much to the advancement of knowledge as the application of a new instrument. The native intellectual powers of men in different times are not so much the causes of the different success of their labours, as the peculiar nature of the means and artificial resources in their possession.

Humphry Davy

PICTUREQUOTES.com
THANKS!!

The National Science Foundation
The University of Texas
Peter and Edith O’Donnell
Dell, Intel, and our many vendor partners
Cal Tech, Chicago, Cornell, Georgia Tech, Ohio State, Princeton, Texas A&M, Stanford, UC–Davis, Utah

Our Users – the thousands of scientists who use TACC to make the world better.

All the people of TACC