HORIZON, THE NSF LEADERSHIP COMPUTING FACILITY, AND THE NATIONAL AI RESEARCH RESOURCE

Dan Stanzione
Executive Director, TACC
Associate Vice President for Research, UT-Austin
Dan@tacc.utexas.edu

MVAPICH User Group (MUG): July, 2024

A QUICK OUTLINE

- Where we (@TACC) are now.
- The new Leadership Facility Award…
- …and connecting that to NAIRR
- The challenges we will have that this community can help with.

TACC RESOURCES

- 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
- Typical power ~6MW
  - Max 9.5MW
- Adding 15MW of datacenter capacity for LCCF (25MW total) 2025.

THE NEW TACC RESOURCES

TACC COMPUTE HARDWARE

THE BIG SYSTEMS IN 2024

Resource	CPU type	#Nodes/Sockets/Cores	GPU Type	# GPUs
Frontera	Xeon (Cascade Lake)	8400/16800/470,400	RTX (Volta)	360
Lonestar-6	AMD Epyc	600/1200/76,800	NV A100	255
Stampede-3	Xeon (Sapphire Rapids)	2,024/4,048/150,080	Intel PVC	80
Vista	**ARM/Grace**	**840/1080/77,760**	**NV H100**	**600**
Horizon	Embargo	Like a million	Embargo	Lots

- Rough total peak power, 9.5MW
- Rough total average power, ~6MW
- Plus cooling power

THE INFRASTRUCTURE IS ABOUT TO GET LARGER, MORE LONG LASTING, AND MORE HETEROGENEOUS

SO OUR SOFTWARE/DATA CHALLENGES ARE GOING TO CONTINUE TO GET HARDER

LEADERSHIP-CLASS COMPUTING FACILITY

The National Science Foundation Leadership-Class Computing Facility

Hosted at
The Texas Advanced Computing Center
The University of Texas at Austin

TACC
TEXAS ADVANCED COMPUTING CENTER

NSF

MVAPICH IS STILL A KEY PARTNER

- OSU is a funded partner in LCCF
- We insist on having at least two MPI stacks on every system, regardless of architecture
  - X86: Intel MPI / MVAPICH
  - Arm: OpenMPI /MVAPICH
- A tuned network stack is key to our success.

THE NSF LEADERSHIP CLASS COMPUTING FACILITY

- The original solicitation for this was posted May 10th , 2017
- Proposal was due November 20th, 2017
- Awarded July 10th, 2024
- (Frontera and a few other things in between).

NSF invites proposals for the acquisition and deployment of a High Performance Computing (HPC) system, called the Phase 1 system, with the option of a possible future upgrade to a leadership-class computing facility. The Phase 1 system will serve two important and complementary purposes:

1. *It will serve as a robust, well-balanced, and forward-looking computational asset for a broad range of research topics for which advances in fundamental understanding require the most extreme computational and data analysis capabilities; and*

2. *It will serve as an evaluation platform for testing and demonstrating the feasibility of an upgrade to a leadership-class facility five years following deployment.*

THE NSF LEADERSHIP CLASS COMPUTING FACILITY

- This is a sea change in the way NSF invests in computing
  - Some of that is funding *source*.
  - Some of that is funding *scale*.
- But the big change is:
  - Computing is on a par with the other NSF facilities
  - Computing investments will be on a par with other NSF facilities.
    - Instead of “4 years and gone”.

THE NSF LEADERSHIP CLASS COMPUTING FACILITY
FOUR MAIN COMPONENTS

- A new home for the facility (15MW of new datacenter, new visitor center, etc.)
- Actual Computing and Storage Systems
- Software and Services (including people).
- Education and Outreach

THE NSF LEADERSHIP CLASS COMPUTING FACILITY
A DISTRIBUTED FACILITY

- Frontera/Vista available now.
- Horizon, the first large system, roughly 10x the capability of Frontera, will be in Austin.
- A Quantum system and accelerator testbed will be at NCSA
- A high-throughput data/computing system will be at SDSC
- A storage/data curation system will be at PSC
- An interactive system to support accessibility will be at AUC (physically at Morehouse College).
- People will be distributed across all these sites as well, plus Cornell and Ohio State.
  - And a few other TBA sites for applications work.

THE NSF LEADERSHIP CLASS COMPUTING FACILITY
TIMELINES

- Construction starts now.
- System delivery late in 2025
- User access in 2026
- Horizon will be around until ~2031/2032
  - Expect more systems after that, Congressional funding permitting.

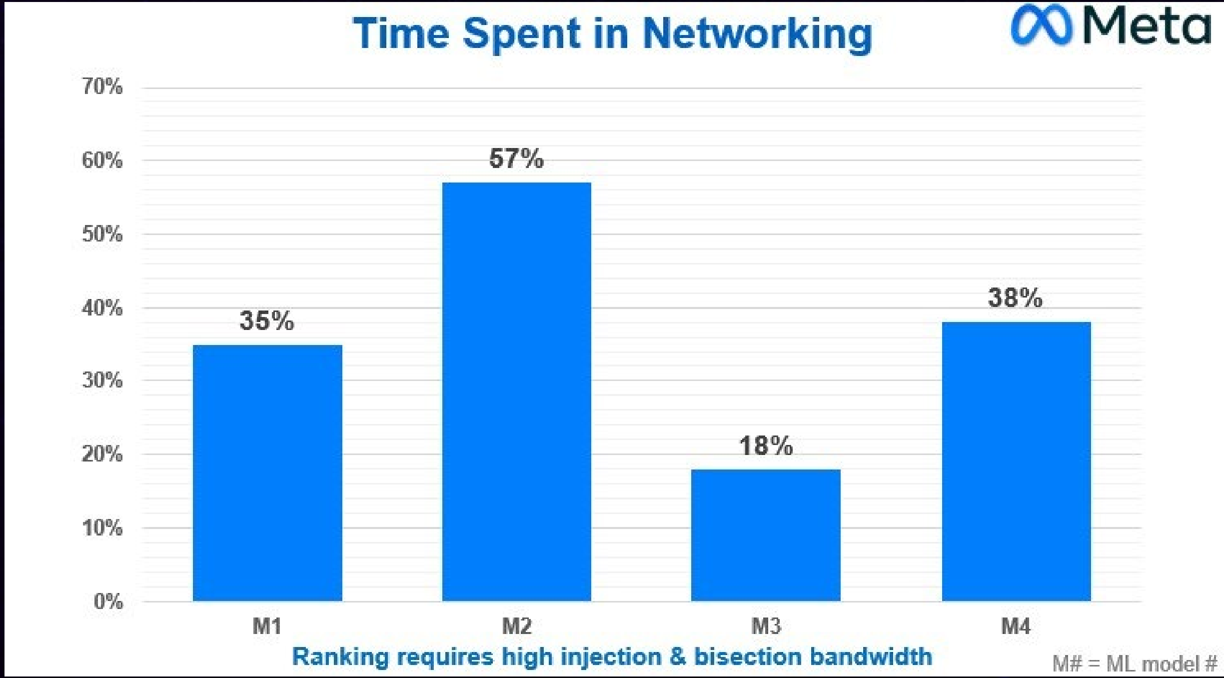
THE NATIONAL AI RESEARCH RESOURCE

- A pilot infrastructure for NAIRR is now underway.
  - But it's within existing funding lines, no new money yet.
- At some point, it is projected to expand greatly.
- As Horizon is funded, and will have a fair amount of GPU capability, expect it to play a large role. . . Especially if lots of new money isn't as forthcoming.
- NAIRR is also envisioned as a stable stream of funding, with resources running on six year cycles.

SO, WE WILL HAVE LARGER RESOURCES COMING

- And, they are going to have longer individual hardware lives
- We know user demand is going to keep driving the data sizes and computation challenges through the roof.
- There are many topics we will need to explore, but let’s focus on a couple that this community can help improve:
  - Interconnects for large distributed AI (and other) applications.
  - Exploiting AI hardware
  - Climate/Sustainability challenges

INTERCONNECTS ARE ONLY GROWING IN IMPORTANCE – AI



- Often, one network rail per GPU
- Both latency *and* bandwidth seems to matter.
- The need for good interconnect is even *more* important than in HPC.
- And AI is the 800lb gorilla to HPC's modest sized chimp.
- This is unleashing new investments in networking.

I STARTED USING THAT INTERCONNECT SLIDE ABOUT A YEAR AGO.

- Since then, I've made it a point to ask the cloud/AI vendors what matters more to boost AI efficiency – bandwidth or latency?
  - Remarkably, no one seems to be sure.
  - This seems like a question worth answering.
- Conventional AI wisdom is we need lots of bandwidth system wide, but even more locally (see: NVLINK/DGX architectures).
  - I'm not sure anybody has validated that assumption at scale.
- Would like to know those answers before we make more nine figure investments in systems.

AI HARDWARE WILL DOMINATE

- AI has led to a new investment in interconnect, and that's great… but it may not be the interconnects HPC users need.
- Similarly, processors and filesystems:
  - The forecast HPC market is $10B/year
  - The forecast AI market is $300B/year.
- We know where hardware vendors will focus.

ADAPTING TO THE MARKET

- This isn't actually a new problem in supercomputing.
  - And academics tend to lead the market on this.
- In 1991, the cold war was ending, which was killing the unlimited government budgets for vector-based custom silicon supercomputers. Cray, SGI, Thinking Machines, Convex, Raytheon Supercomputing, many other companies were falling apart – most didn't survive.
- At NASA Goddard, Thomas Sterling and Don Becker started the "Beowulf" project exactly 30 years ago.
  - In Thomas' exact words, those of us doing scientific computing needed to be "bottom feeding scumsuckers" - words I've built me career around ;-).

ADAPTING TO THE MARKET

- The gist – silicon is expensive, use the commodity parts.
  - Step 1 – Don wrote network drivers for this thing called "Linux". First time it talked via Ethernet. That worked out.
  - Step 2 – Come up with ways to use commodity processors.
  - Almost all Top 500 machines since have used this.
  - Even the addition of GPUs to HPC was about riding the commodity (gaming) markets.
- Universities led, agencies followed kicking and screaming (DOE still makes NRE investments with vendors).
- **WE CAN DO THIS AGAIN – and this time we have more to offer in the other directions.**

AI HARDWARE FOR SCIENCE

- There have been lots of initiatives around "AI for Science" and "Science of AI".
- We need to focus – again – on how to exploit commodity hardware for scientific computation.
- This is the next Beowulf project – what if we built a cluster of *AI* chips for our next gen of scientific computing?

A BIT ON SUSTAINABILITY

- "Green" Computing has been largely considered a datacenter problem.
- And there is stuff we can do in the datacenter… but I would argue that though those investments are good, they are not even where *most* green computing will happen.

COOLING TECH HAS NOT *ONLY* BEEN ABOUT IMPROVING PUE

- It's about density.
  - At the chip level, we need something that can dissipate heat in the given area – increasingly, that's not going to be air.
  - At the rack/datacenter level, it's about cable length/latency – ~1 ns/foot of fiber/cable.
  - Low latency matters not just for HPC but for AI now.
- Chip power is increasing fast:
  - Intel CPU : 130W (2012), 145W (2017), 210W (2019), 350W (2024)
  - NVIDIA GPU: 300W (PCI ,~2019): 600W (SXM,2023), >1,000W (2025?)
- So rack power goes up too:
  - At TACC: 33KW/rack (2012), 60KW/rack (2019), 70KW/rack (2021), forecast 135KW/rack (2025).
- PUE is a happy side effect, but we can't keep doing air, or servers would look like:

EVOLUTION OF TACC COOLING STRATEGIES

- Ranger (2008) Stampede 1 and 2 – In-row Chillers enclosed hot aisles (2012 build out).
- Frontera (2019) Stampede 3 (2023) - Direct Liquid Cooling of processors (CoolIT, CoolTerra, Vertiv) .
- Frontera RTX (2019), Lonestar-6 (2021) – Immersion cooling (GRC).
- We also employ chilled water storage to offload the power grid at peak demand.
- We employ roughly 200kw of direct solar, and by wind credits for about 20% of the remainder.
  - New datacenter will be 100% wind offsets.
- Next datacenter – we will definitely have (probably warmer) water to each rack location, the rest is somewhat TBD

COOLING WILL KEEP IMPROVING

- New heat spreaders to take immersion (high viscocity fluid) past 2KW/socket.
- For DLC, new innovators will improve density and reduce leaks:
  - E.g. Zuttacore (multi-phase cooling), Chilldyne (negative pressure DLC).
- Warm water supplies will reduce the need for chillers most of the times, in most (non-Texas) climates.
- We can expect continued improvements in PUE.   But. . .

PUE IMPROVEMENTS HAVE DIMINISHING RETURNS

- The "average" datacenter hit about 1.67PUE in 2018, probably below 1.5 now.
- Almost all new build, dense, large scale datacenters are 1.2-1.3 or better.
- Like in every other part of HPC, Amdahl's Law eventually becomes a big problem.
  - Getting PUE from 2 to 1.2 reduced power by 40%.
  - Getting from 1.2 to 1.05 will reduce power by ~10%.
  - Only 5% left from there to theoretically perfect.
- Against hundreds of GW of datacenters consuming thousands of TW/hours, this won't make much difference.
  - At any value of X in a 1.X PUE, we still have the 1.

SUSTAINABILITY AND DATACENTERS

- Obviously, sustainability is a priority.
- But the mission - providing the best computational resources – is the highest priority.
  - We are both the cause of and solution to many of these problems ☺.
- Datacenters are still a tiny fraction of usage compared to, say, transportation.
  - And our datacenters help design batteries, carbon capture and storage, better photovoltaic materials, remediation for plastics and chemicals, etc, etc.
  - A better use of power than the much larger datacenters for X/Twitter, Cat Videos, and generating targeted ads.
- *If we had a green power grid, not only would our datacenters not be a problem, a lot of other stuff wouldn't be either* – but we can't change that unilaterally.

A FEW BITS OF OUR SUSTAINABILITY PLANS:

- We continue to run experiments to improve the efficiency of our datacenter operations:
  - We are working with several startups on novel cooling technologies.
  - We continue to work with our vendors to be able to raise inlet temperatures for water – while maintaining a high enough delta-T to keep chillers running efficiently.
    - We are in Texas, we are probably going to still need chillers, even if water temps reach 35C.
  - Going to 100% wind credits for a 7% markup – willing to pay that.
- Storage technologies will help us incorporate renewables more efficiently.
  - We have an experimental Hydrogen fuel cell in our current datacenter power loop.
  - Various other storage technologies being explored.
- Similarly, we are working to improve how power is managed:
  - Capping power at modules (e.g. Grace-Hopper cards, and future versions with potentially more components) rather than at the server level will reduce the datacenter build out for "max power".
  - We will be below 9MW in our current projected design for Horizon, the "10x" replacement for the Frontera system in 2025.
- Still. . .

INCORPORATING RENEWABLES HELPS. . .

- But the whole grid will not move swiftly, and there is still only so much available power using it all in datacenters means less green power somewhere else.
  - **Maybe a little more swiftly than some think – In April, more power came from wind than coal in the US.**
- But if projections are to be believed, GenAI demand alone will add approximately one Texas (75GW) to the power grid when current construction is completed.

TO GET SERIOUS IMPROVEMENTS IN EFFICIENCY:

- We have to move past the discussion of just pushing on the datacenter facility systems.
  - These are great, but the returns will be a small fraction of total power.
- Serious improvements will come from the hard problems – better hardware and software.

SOFTWARE AND SUSTAINABILITY

- We know, for instance, that per "peak" FLOP, we get a 5-6x multiple moving to GPUs.
  - But outside of AI, a large fraction of codes don't run on GPUs.
  - (And arguments can be made on yield of peak flops across architectures).
  - 5x is more than 15%.
- We also know, but don't really talk about, that most actual app runs get a single digit percentage of peak performance.
  - Which means code efficiency offers the potential for an order of magnitude improvement.
    - Yes, more efficient code uses somewhat more instantaneous power – but shorter runtimes help a lot.
- The problems is software is hard, diverse, and often beyond our reach. . .
  - ***But a crappy job on software, with 1,000% potential, is probably better than a great job on datacenter, with 10% potential.***

IS HARDWARE POWER EFFICIENCY IMPROVEMENT POSSIBLE? YES.

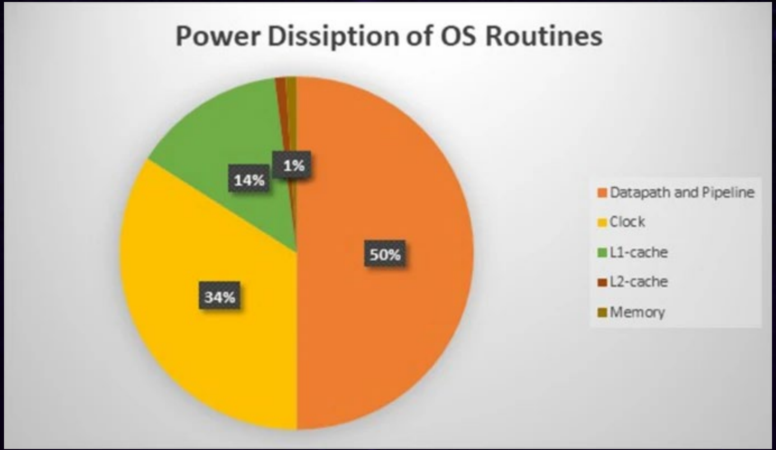
	TFlops	Watts	**Gflops/Watt**	BW	**Flops/Byte**
Intel ICX (Dual-Socket)	5.9	540	**10.93**	300	**20**
AMD Milan (Dual-Socket)	5.1	560	**9.11**	300	**17**
AMD MI250x	47.9	560	**85.54**	3277	**15**
NVIDIA A100	9.7	400	**24.25**	1600	**6**
NVIDIA A100 (Tensor)	19.5	400	**48.75**	1600	**12**

GPUs have a serious advantage in GF/Watt.

The silicon process is the same. Why? Architectural choices.

WHY ARE GPUS MORE EFFICIENT?

- Simpler circuits – push the work back to the programmer.
  - Complex branch prediction, fetch-decode-execute cycles are expensive in power.
  - Hardware and Software are inevitably interrelated.
- Moving data 2MM across the chip takes more power than floating point operations to produce it.
- The push to AI-specific chips is taking this trend much further.
  - Lots of upside, but SW price to be paid.
- Once we are willing to open up the software, even current chips give us lots of opportunities. . .



From Katal, et al, "Energy Efficiency in cloud computing datacenters"

H100 PERFORMANCE ACROSS PRECISIONS

- *Source: NVIDIA*
- For Vector units, SP is unsurprisingly 2x DP.
- For Matrix units, it.s 15-1!!!
- At FP16, 2PF *Per socket*
- Maybe we need to spend a bit more time on using mixed precision Matrix ops, given **the 30X advantage**

FP64	34 teraFLOPS
FP64 Tensor Core	67 teraFLOPS
FP32	67 teraFLOPS
TF32 Tensor Core	989 teraFLOPS*
BFLOAT16 Tensor Core	1,979 teraFLOPS*
FP16 Tensor Core	1,979 teraFLOPS*
FP8 Tensor Core	3,958 teraFLOPS*

NOT JUST LOW/MIXED PRECISION OPPORTUNITIES

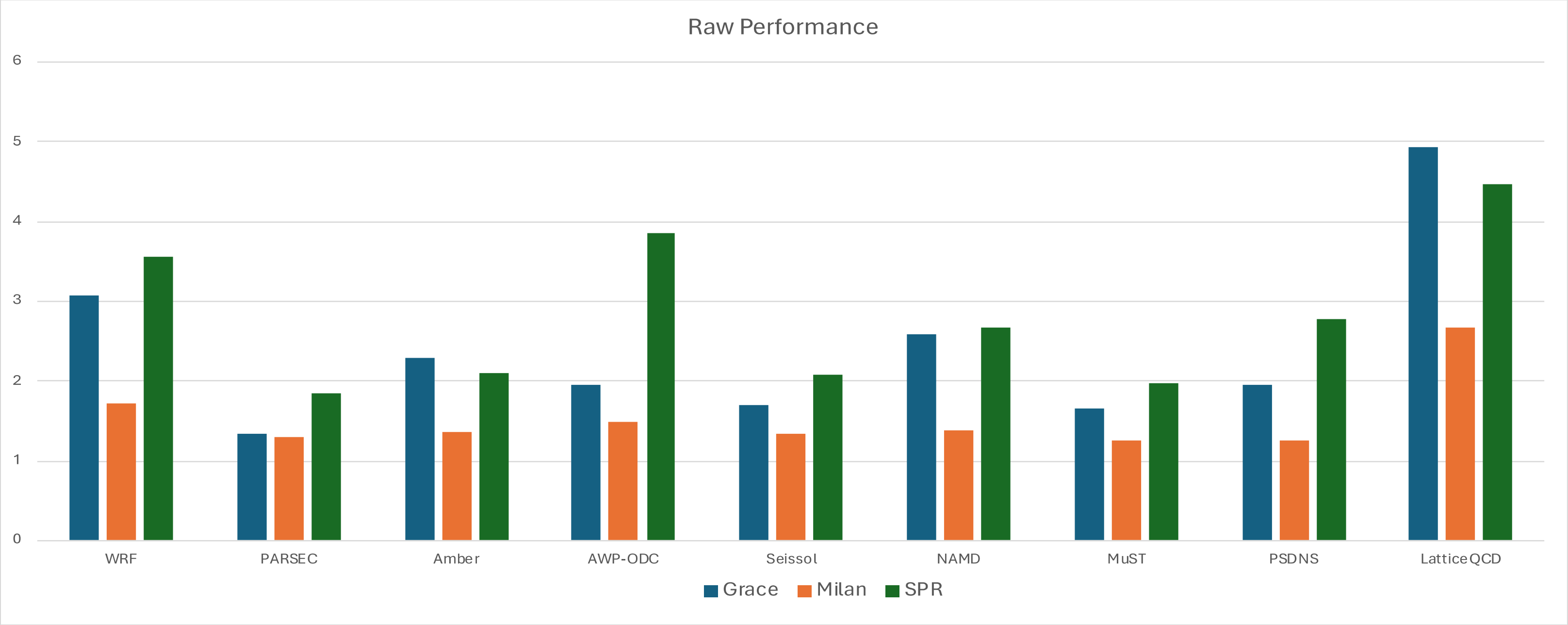
- There are plenty of other architectural things that can happen, even without radical change.
- For instance, change the balance in our CPUs by improving memory bandwidth.
  - Our benchmarking shows typically ~1.7x improvement, with outliers up to 4x, for adding HBM to CPUs (comparing two Intel SPR chips at 350W each).
  - This improvement happens at the same power per socket, and the same peak flops! It's just re-balancing the architecture to raise efficiency.
- Other configurations are possible.

ARM VS. X86

- So, we've done a ton of x86, and those have largely been predictable.
- But, new CPUs obviously fill us with trepidation.
- That said, things have gone remarkably smoothly on the software side.
  - Our 20 major benchmark codes all built from source with relative ease.
    - Despite a much younger tool chain.
  - Performance is predictable, and pretty good.
- Let's look at some pure CPU numbers where we can do comparisons.
  - Note, for us, Frontera (Intel Cascade Lake, Platinum, 8280, dual-socket) is "1" for speedup purposes).

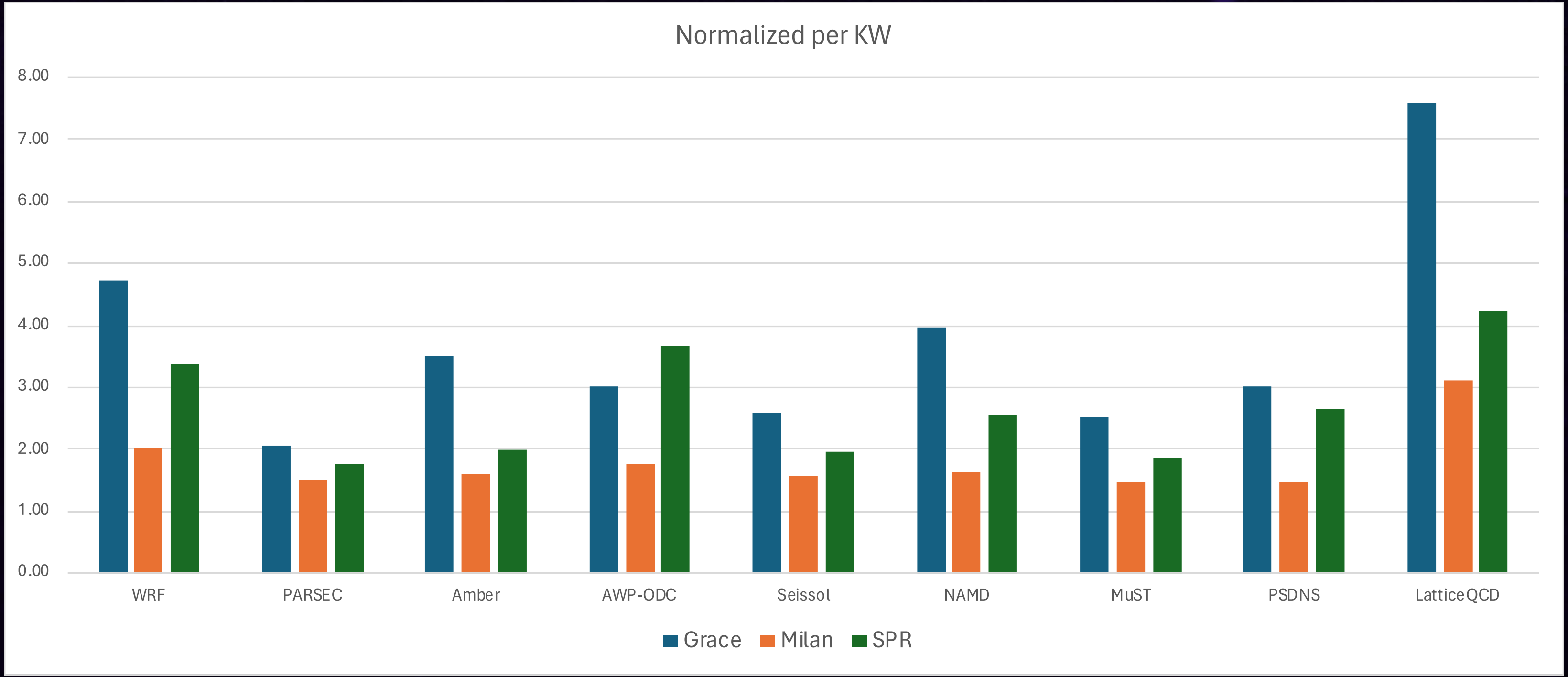
BENCHMARKS
(WITH THE USUAL CAVEATS)

- 8 application codes, single node benchmark cases.
- Grace – Vista; AMD Milan - Lonestar-6 (one gen old); Intel –SPR with HBM (Stampede-3)



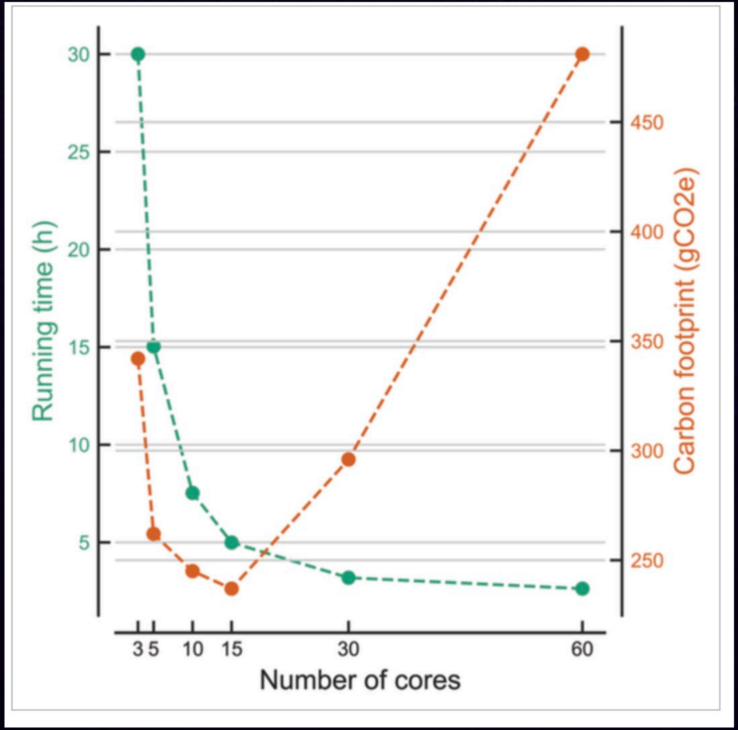
BENCHMARKS
(WITH THE USUAL CAVEATS)

Grace is top performer on 8 out of 9 apps When power is considered.



ON THE SOFTWARE SIDE, IT'S NOT JUST PORTING TO THE NEW CHIPS

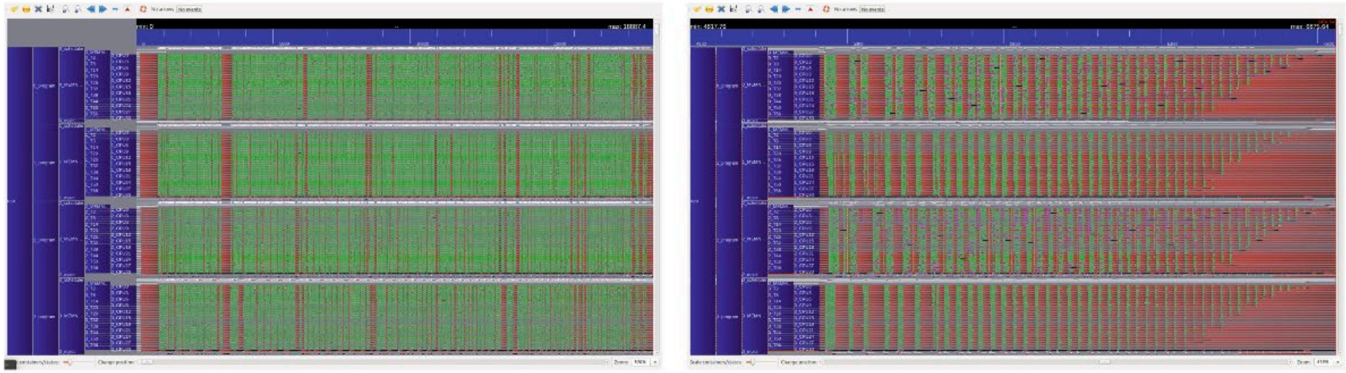
- Mixed/Low precision
- Reduced Rank
- Take advantage of sparsity
- Higher order methods
- All sorts of other algorithmic cleverness
- Even just **picking the right number of cores**



Geant4 Particle Physics code, from Lannelonge, Grealey, and Inouye
Green Algorithms: Quantifying the Carbon Footprint of Computation,

ALGORITHMS CAN HAVE A HUGE IMPACT. . .

Exploit Lower
Rank Algorithms



(a) Dense **dpotrf** time=18.120s (b) Data-sparse **dpotrf** time=1.761s

Cholesky factorization times on 4 nodes of Shaheen-3, Matrix size 54k
Akbudak et al, "Exploiting Data Sparsity for Large Scale Matrix Computations"

INCENTIVES FOR SUSTAINABLE SOFTWARE

- We are sampling performance data every few minutes on every job to keep a profile of efficiency
  - This is one of the ways we target consultants.
- Pushing the user base (somewhat) towards increasing GPU usage.
  - Just added GPU monitoring; anecdotally, there is massive inefficiency there.
- A problem we have is *incentives* -- users just want the fastest answer – no incentive to get a slower answer that uses less power (we saw this a lot on Stampede 2).
- **Perhaps we change our charging units from wall clock hours to total Joules consumed??**
- We hope to start reporting energy usage to users next year – not sure when/if we will go to energy-based charging.
  - Incentivize more efficient codes.
  - Maybe incentivize moving loads to optimal power cost times? (West Texas wind power can be somewhere between free and negative a fair number of hours per year).

AI HARDWARE FOR SCIENCE *AND* SUSTAINABILITY

- There have been lots of initiatives around “AI for Science” and “Science of AI”.
- We need to focus – again – on how to exploit commodity hardware for scientific computation.
- We also need to focus on actual optimization of software for AI.
- With an estimated spend of $300B on AI hardware this year, and proposed plans for $30B/yr in US Gov AI spending (that won’t happen, but still), can’t we find ~1% to make the software exploit the hardware a little more efficiently?
    - What if it “only” got us a 10% improvement in average efficiency?

THANKS!

dan@tacc.utexas.edu

Acknowledgements:

NSF, the source of the money!

The University of Texas at Austin

Dell, and lots of other vendors

The whole team at TACC and our partners!