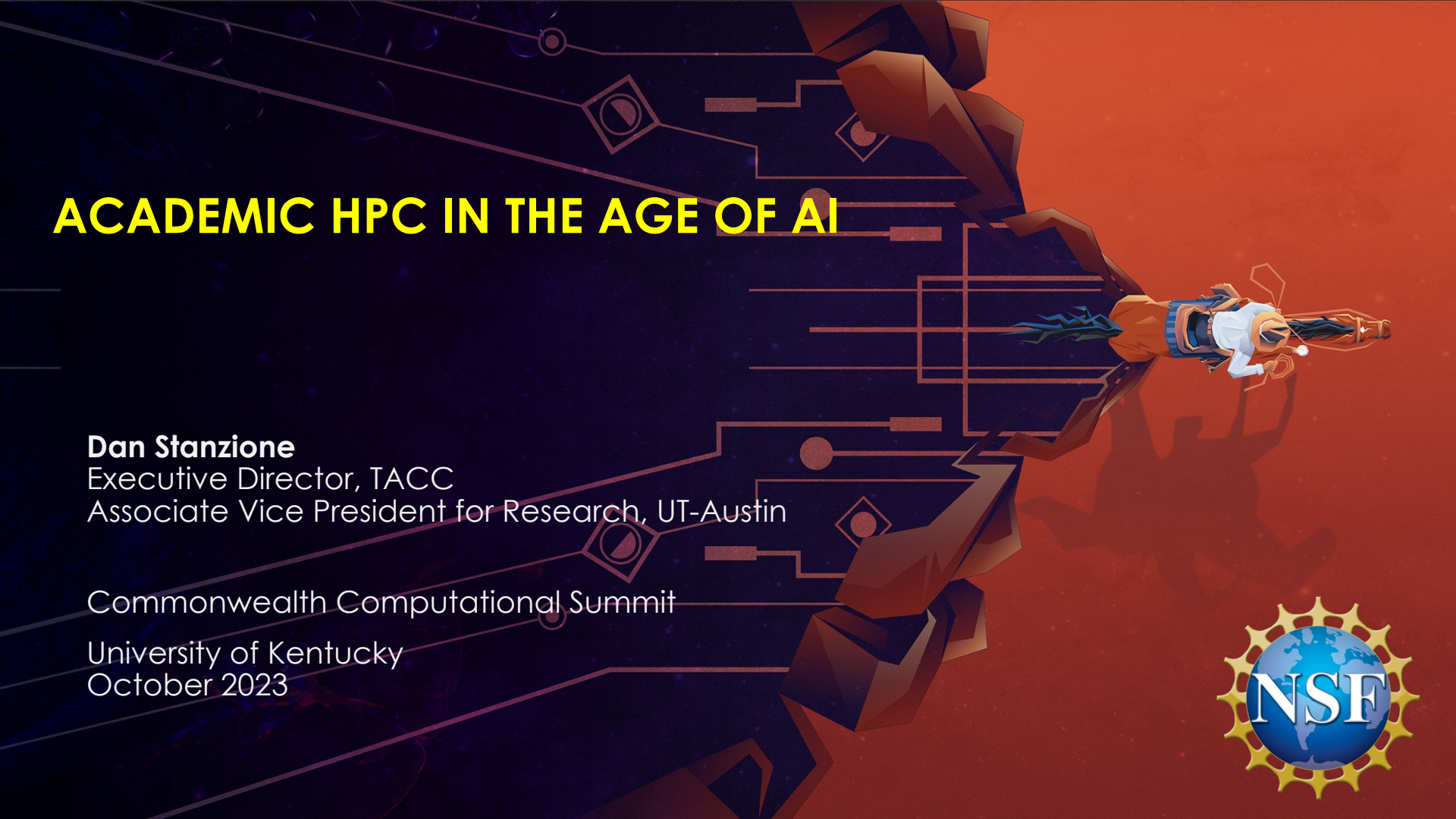
ACADEMIC HPC IN THE AGE OF AI



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

Commonwealth Computational Summit

University of Kentucky
October 2023

TACC - 2023

LEADERSHIP-CLASS
COMPUTING FACILITY

TACC
TEXAS ADVANCED COMPUTING CENTER

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.

THIS PAST 12 MONTHS HAS BEEN A WATERSHED FOR AI

- ChatGPT was the "Sputnik moment" in an already building wave.
  - AI has been capturing headlines for the last 7-8 years.
  - The release of Transformers (from Google) unleashed the ability to scale to enormous sizes.
  - But ChatGPT changed everything, especially public perception.
- There is now a global "AI Arms Race", leading to a scramble (in both public and private sectors) for:
  - Funding
  - Expertise
  - Regulation/Policy
- I'm regularly hearing about billion-dollar machine orders paid for entirely by venture money to train products that don't yet exist.
- AI and HPC are deeply intertwined – so academic HPC can't pretend this is business as usual.
  - **Modern AI would not exist without scientific supercomputing**

THREE MAIN THEMES

- How does this change the hardware/software we deploy?
  - Or what we can get?
- What does this mean for our workloads and user base?
- What do we need to do about our operations?
  - Funding, people, day to day ops

IN MANY WAYS, AI VINDICATES THE "HPC WAY"

- AI needs fast interconnects. We had them, the cloud and the enterprise did not.
- AI needs message passing; MPI was built for HPC, but is now the standard library for transformer-based generative AI wave (e.g. ChatGPT, DeepSpeed, etc.).
- AI needs heterogeneity – GPUs for general purpose compute came out of the HPC world.

- This means AI needs HPC hardware (probably good) and HPC programmers (good if you are one, bad if you need to hire one).

AI HARDWARE WILL DOMINATE

- Per Hyperion:
  - The market for AI-driven hardware will be $300B/year in 2025.
  - The market for “pure” HPC hardware will be $10B/year in 2025
  - Guess which will get more vendor attention?

AI HARDWARE WILL DOMINATE

- Interconnects, filesystems likely to be the *same* for AI. (More on that in a few minutes). So, AI momentum will be good for Academic HPC.
- Processors – will be similar, but not the same (lower precision, for instance).
  - We are unlikely to be able to deeply influence what gets built (maybe some nudges around the edges, e.g. memory controllers).
  - We are more likely to need to adapt.
  - In general, *if the cloud folks won't buy it, it probably won't succeed* – so we should buy that.
- Another downside for us, in the short term, is that GPU prices are through the roof.
  - **It is cheaper per ounce to buy gold bars than GPU sockets**.

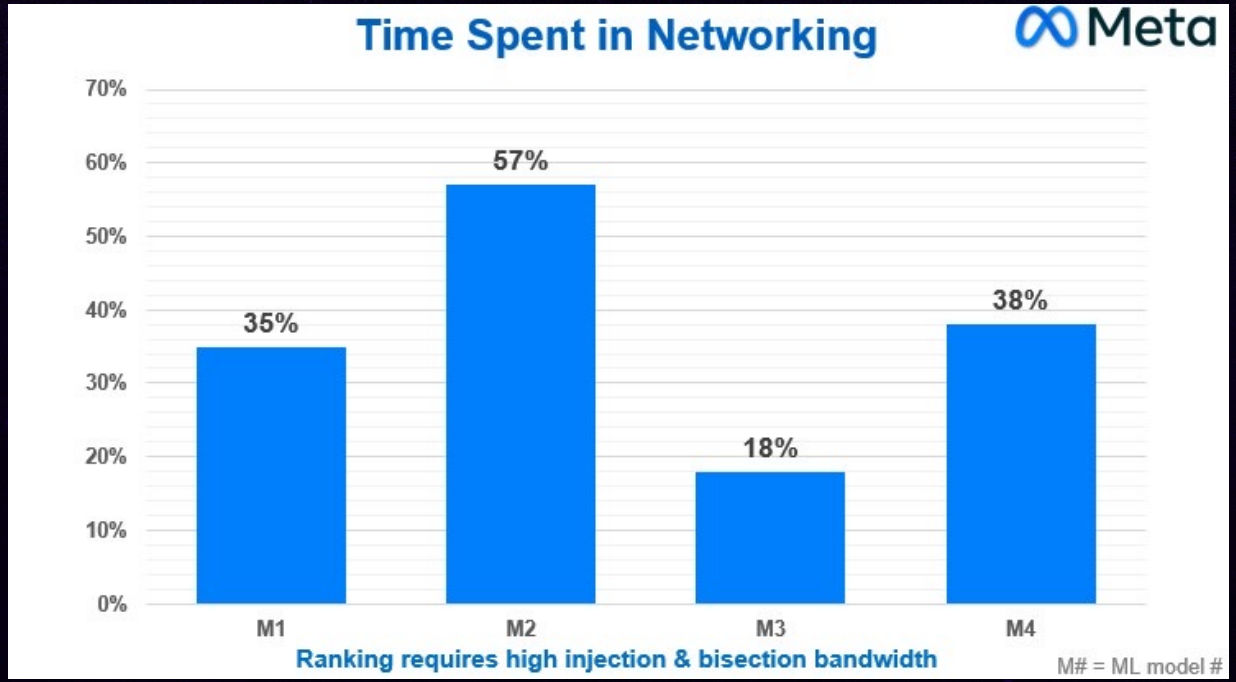
ADAPTING TO THE MARKET

- 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 ;-).
- 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.**

INTERCONNECTS ARE ONLY GROWING IN IMPORTANCE

- Interconnects have *always* been critical for HPC.
  - Mostly latency, but also bandwidth.
- The long time cloud rallying cry was “you don’t need all that expensive interconnect bandwidth if it’s not HPC”.
- Then AI came along. . .

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.

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? ←- several vendors here, from the traditional, to, well, Amazon.

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?

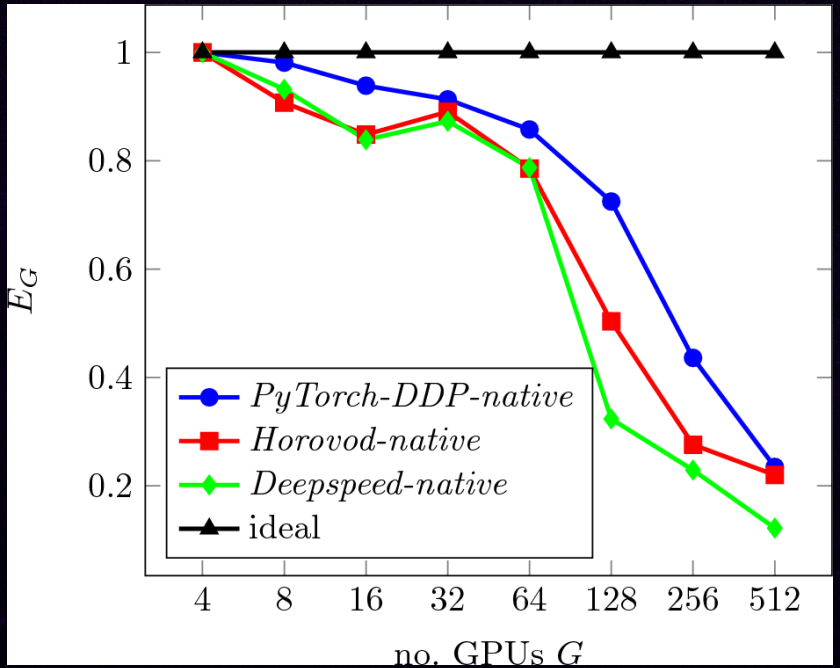
BUT SHOULD THEY EAT A LARGER AMOUNT OF SYSTEM BUDGET?

- Or should we be more clever?
- Compression seems to have serious benefits with large messages (often in AI), and is almost free (particularly if you put processing in the network path – e.g. DPU – or you have like 192 cores on a node).
- But since we are here to talk about network *libraries*, how much is the physical network vs. library vs. application?

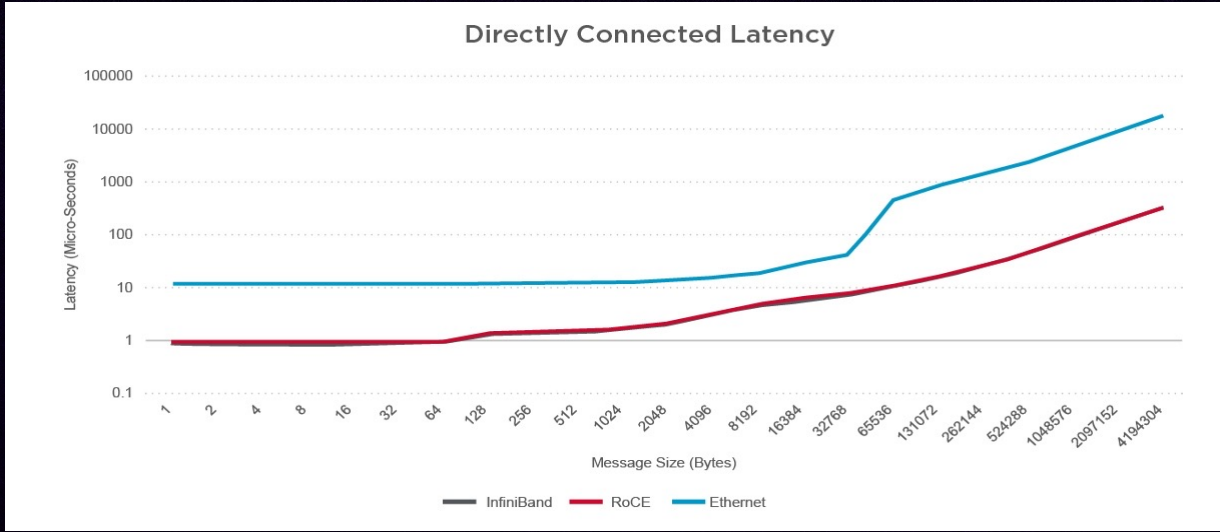
IT IS *NOT* THE APPLICATION FRAMEWORKS

- Pytorch vs. Deepspeed vs. Horovod – not much significant difference there (for AI apps).
- Note – all of these rely on MPI under the covers to scale.

- Aach et al, "Large scale performance analysis of distributed deep learning frameworks for convolutional neural networks", June 2023



IT MAY NOT SO MUCH BE THE NETWORK HARDWARE. . .



- It might be the **communications software**.
- "Regular" ethernet sucks – but add RoCE at same BW as IB...
- (highly biased source: Broadcom)

USERS AND WORKLOADS

- First of all, we are all seeing lots and lots more AI users.
  - We need to adapt (systems, policies and support) to meet the needs of these highly dynamic workloads.
    - Maybe less shared filesystem?
    - Container support a 100% must – but we should do that anyway.
    - Staff need to support this now too – but performance tuning is performance tuning.
- We also need to *protect* the role of traditional modeling and simulation in scientific computing. . .

PROTECTING TRADITIONAL MOD/SIM

- On the one hand – we need to push users (gently) to modernize code, and exploit GPUs/heterogeneity.
  - There is increasing evidence we can *theoretically* get almost all algorithms to work at least OK on GPUs, and some have huge advantages (see, for instance, the Exascale Computing Project at DOE).
- On the other hand – a lot of the *actual code* in existence – probably 90% of the code and 50% of the workload – still won't work on GPUs today.
  - So, giving them an AI-only machine is a serious problem. But lots of places are doing it anyway, which works as long as there are other places to go.
  - At TACC, we have committed to our users who need CPUs:

We will have >1M CPU cores on Horizon

BLENDING AND POSITIVE FEEDBACK

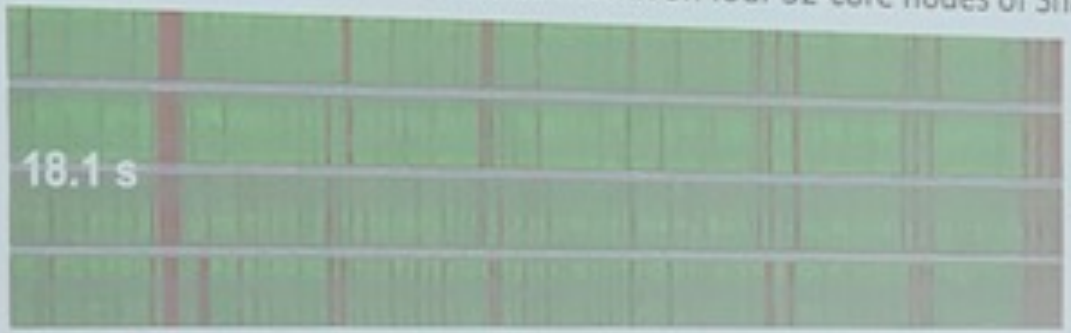
- The notion of "AI Users" and "HPC Users" won't hold up for long.
- There are a diversity of ways to use "AI for Science", and we need to help our *entire* user bases get there.
- The converse is also true, and perhaps more strongly true.
  - Given that the vast number of weights in neural nets are effectively zero, perhaps knowing something about sparse matrix methods could provide an order-of-magnitude improvement in their *HUNDRED BILLION DOLLAR TRAINING BUDGETS*.
    - Scientific computing has a 60 year head start on this.
  - As they scale to many nodes (ChatGPT trains on 9,000 GPUs), squeezing more GPUs in a box is unsustainable. They might want to learn about distributed memory algorithms (you know, the thing we had to do to make Beowulf clusters work starting 30 years ago.
- HPC is not only necessary for AI, we have the algorithms to move AI forward (and the obligation to do it). ..

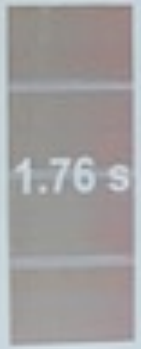
HPC, AI HARDWARE, AND SUSTAINABILITY

- To borrow from my friend David Keyes:
- *As computational infrastructure demands a growing sector of research budgets and global energy expenditure, we must enhance utilization efficiency.*
- As a community, we have excelled at this historically in three aspects:
  - architectures
  - applications
  - algorithms
- Among other opportunities, algorithmic opportunities abound:
  - reduced rank representations/ reduced precision representations

Our journey in tuned approximation began in 2018 with these time traces…

… for factorization of a dense 54K covariance matrix on four 32-core nodes of Shaheen-2





- TLR scores a lower percentage of peak (after squeezing out flops)
- TLR has poorer load balance (a higher percentage of idle time (red) vs. computation (green))

Akbudak, Ltaief, Mikhalev, Charara & K., *Exploiting Data Sparsity for Large-scale Matrix Computations*, Euro-Par 2018

Exploit Lower Rank Algorithms

AI *HARDWARE* FOR SCIENCE

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*

GPU ADVANTAGE – NAÏVE FIRST CUT

	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**

In terms of FLOPS/Watt, GPUs clearly win right now!

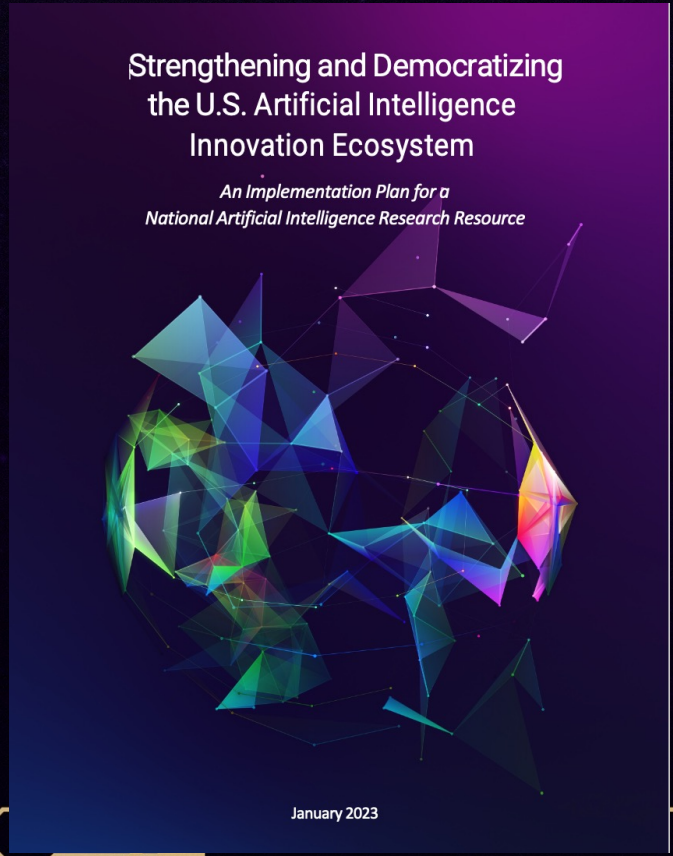
Even at this level, the GPU cost/TF advantage isn't that clear cut (Assume a node with two A100 cards cost 3x a node with no GPUs).

DON'T FORGET OPERATIONS

- Don't forget, AI impacts us the way it does other organizations as well.
- I haven't made the users face AI chatbots directly yet, but ---
  - They actually do write tickets better than staff. . .
    - …when not completely lying
  - We do 9k tickets per year – we have a system now that can write answers, trained on our docs.
  - Soon, we will auto-generate a draft of every ticket that will go to the staff member assigned for review.
- Infinite possibilities in scheduling, performance monitoring, fault prediction, etc.
- And everybody doing coding should be getting help from AI.
  - Jupyter-enabled code assistant being deployed trained on the TACC API docs.
    - E.g. "generate some code to copy this data to Frontera". "Generate code to start WRF on Stampede-3".
  - Note, it is help… you still have to know how to code!!!

THE POLICY SIDE IS CATCHING UP

- NAIRR, CREATE-AI Act, etc. will push the funding opportunities forward (as discussed earlier today).



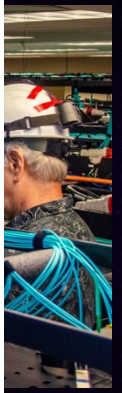
OH YEAH, AND NEW STUFF AT TACC:

- Stampede-3 was announced this summer **(Intel)**
  - 560 nodes - Sapphire Rapids with High Bandwidth memory
  - Hang on to some Ice Lake and Skylake Xeon nodes from S2 (~1,300 nodes).
  - A little bit of Intel Ponte Vechio GPU (80 GPUs, 20 nodes)
  - New storage and interconnect (OPA 400Gbps) , ~2k nodes total
- Vista – Pre-Horizon bridge system **(NVIDIA)**
  - Grace-Grace and Grace-Hopper (later 23/early 2024) 500-600 nodes and Infinband.
- Lonestar-6 will continue to expand **(AMD)**
  - ~600 Milan Nodes
  - VM-queues for smaller throughput jobs
  - (Also 85 GPU nodes with A100)
  - APUs to be added.

CONCLUSIONS

- AI is here to stay, and it impacts virtually everything Academic HPC does. . .
- "AI for Science" for our scientific users
  - But don't forget the traditional stuff!
- "CI for AI" to translate what we need from our knowledge base to the AI community.
- "AI *Hardware* for Science" to exploit AI hardware advantages in more sustainable scientific computing.
- Don't forget AI in Operations – we need to modernize like everyone else.
- Future funding is going to pivot – be ready.

TH




FRONTERA
TACC
NSF
TEXAS
NSF