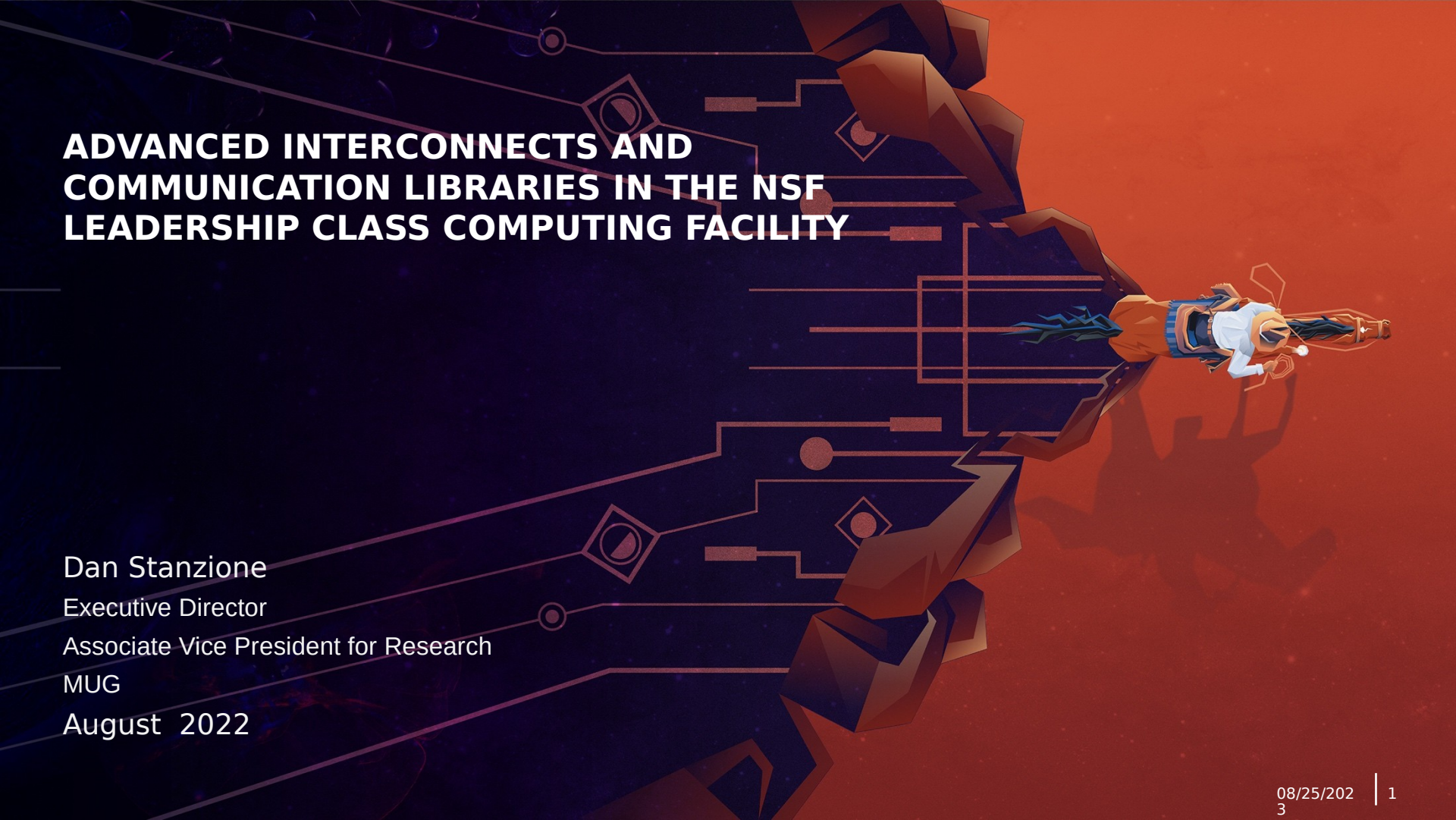
ADVANCED INTERCONNECTS AND COMMUNICATION LIBRARIES IN THE NSF LEADERSHIP CLASS COMPUTING FACILITY



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

THANKS FOR INVITING ME BACK!

First time back in Columbus since 2019

Pleased to continue what is now our *17 year* partnership with the MVAPICH team!

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, 1[illegible] GPUs
- About seven billion core hours over several million jobs per year.

TACC - 2023

LEADERSHIP-CLASS
COMPUTING FACILITY

TACC
TEXAS ADVANCED COMPUTING CENTER

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.

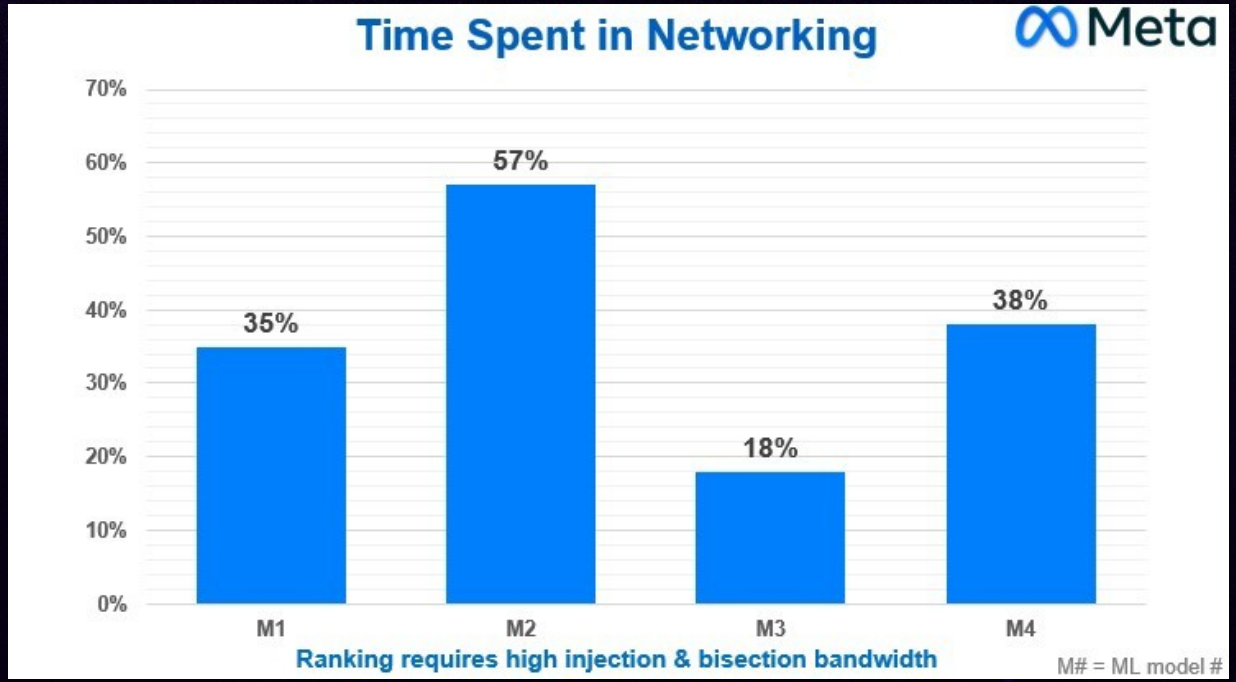
PHASE 2

LEADERSHIP-CLASS
COMPUTING FACILITY

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. . .

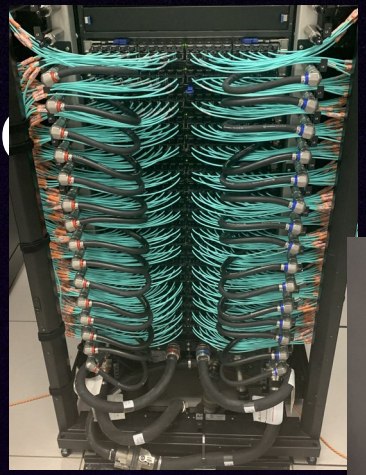
INTERCONNECTS ARE ONLY GROWING IN IMPORTANCE – AI



- 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.

HOW WE SEE SYSTEMS T

- 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? - 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?

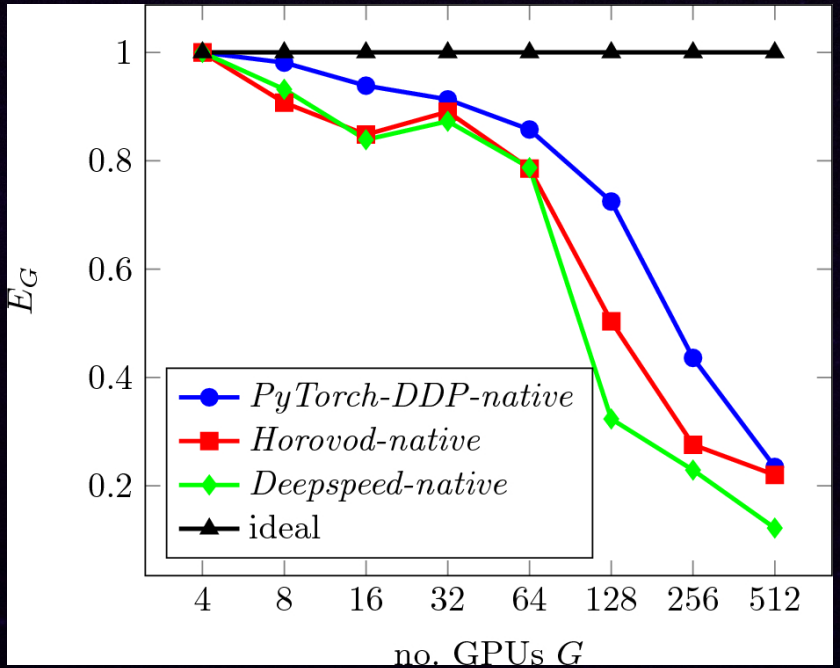
BUT SHOULD THEY EAT A LARGER AMOUNT OF SYSTEM BUDGET?

- 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?

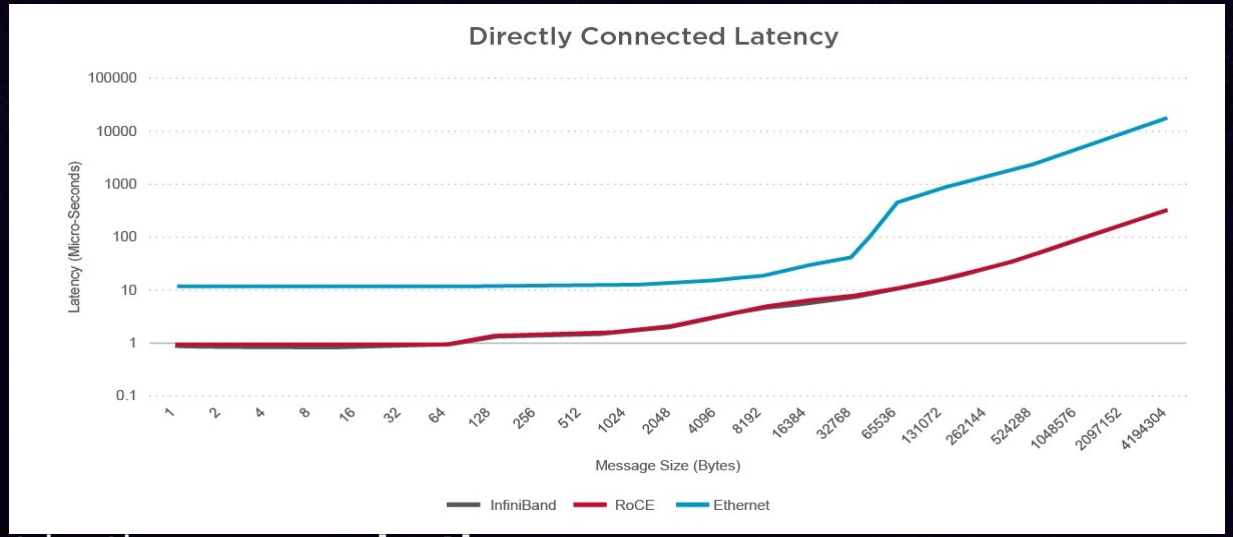
IT IS *NOT* THE APPLICATION FRAMEWORKS

- 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. . .

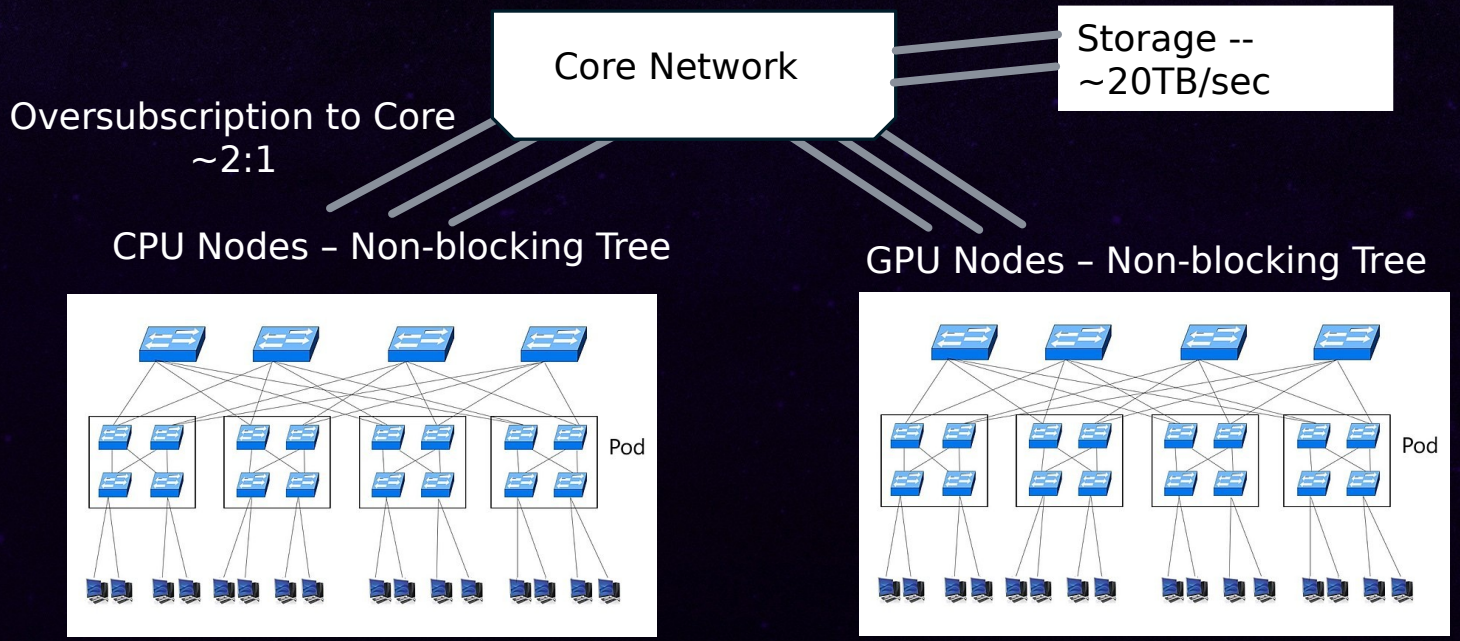


- It might be the **communications software**.
- “Regular” ethernet sucks – but add RoCE at same BW as IB…
- (highly biased source: Broadcom)

A FEW WORDS ON TOPOLOGY

- At TACC, we have typically built fat trees (though occasionally with small amounts of oversubscription at the top level).
- Conventional wisdom says this network is the most expensive, and other topologies can deliver *most* of the performance for a smaller price.
- But that savings never materializes for us. . .

TOPOLOGY FOR LCCF SYSTEM



WHAT'S GOING ON WITH THE LCCF SYSTEM

- Right now, we have submitted a final plan, but are in budget limbo.
  - Without a start date, it's tough to have final choices on technology.
- So, we are using the *planned* start date, but all system details rely on us coming within six months of this date – if that changes, nothing on the next slides is true anymore!!!

TENTATIVE LCCF SYSTEM PLANS

- Based on a March 2024 start date, and a July 2025 delivery date (or everything changes!).
- Primary system: NVIDIA Grace and Grace-Hopper nodes.
  - Approximately 20/80 split in performance, but 60/40 split in investment between CPU/GPU nodes.
  - Infiniband, one rail per node (GPU nodes will have *one* Grace-Hopper per node).
- Still 1M cores of CPU
- ~400PF peak DP64 performance -- ~10 Exaflops at Bfloat16 for AI.
- 400PB of (solid state) storage to match.
- MVAPICH as primary communication library.
- Vs. today's top "exascale" systems:
  - Faster on AI
  - Faster on I/O
  - Faster on CPU-only

HORIZON - TENTATIVE

- Assuming budgets happen on time (a big assumption) and vendor roadmaps hold (another big assumption):
- We will build Horizon around NVIDIA Grace-Next and Hopper-Next modules, with summer 2025 delivery.
  - Significant “Grace only” (ARM) CPU capability, with LP-DDR memory.
  - Multi-GPU nodes tightly coupled with Grace, with interesting power properties.
- Peak power ***below* 9MW**, including storage.
- Is ARM a risk? Yes – but it’s not just NVIDIA, it’s also Apple (this Mac), Amazon, and the whole Mobile space.

DISTRIBUTED CENTERS

- The LCCF Hardware (and staffing) will not only be at TACC, but also at four other sites around the country. (Through construction and operations).

- NCSA --

  Focus on accelerating applications

- SDSC --

  High throughput, and HT Inference systems

  for large scale scientific Instruments

PSC –

Focus on storage systems (and data rep site)

AUCC --

Accessibility, Workforce, interactive

WHY ALL THESE GPUS?

- For starters, progress continues to be made on GPU codes. . .
  - All Deep Learning codes are “GPU-native”
  - About 40% of the scientific apps have moved successfully.
  - (But 60% haven’t – hence we will still have 1M cores of CPU).
- We also feel some need to keep pushing the community on this - though not as hard as DOE - but for the same reasons as DOE.
  - The power/performance ratio is compelling in GPU’s favor right now.

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).

IN THE INTERIM AT TACC

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

GPUS MEAN MORE THAN PORTING TO A NEW LANGUAGE, OR TIGHTLY COUPLING COMMUNICATION LIBRARIES.

- While we look at the impact of MemBW on our workloads, and continue to look at the impacts of porting to GPU.
- A somewhat underappreciated factor is the non-linearity in performance of new devices as precision decreases. . .
- Let’s take the NVIDIA Hopper H100, as that is public. . .

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*

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