Accelerating Inference at SDSC

Presented at:
Accelerating Physics with ML@MIT

Frank Würthwein
Director SDSC
January 31st, 2023

SDSC's Mission

Translating Innovation into Practice

- **SDSC adopts and partners on innovations from industry and academia in the areas of software, hardware, computational & data sciences, and related areas, and translates them into cyberinfrastructure that solves practical problems across any and all scientific domains and societal endeavors.**
- **We are globally renowned practitioners in translating innovation into practice.**

Intro of SDSC by Numbers

250++
Employees

~3,000
Training & Event
Participants/year

~10,000
Active Unix
Accounts
on HPC systems

1M++
users on
Science gateways

1M++
Students took
our Big Data courses

4 HPC Systems
~200,000 x86 cores
~1,500 GPUs

AI/ML Supercomputing
Habana/Intel hardware
SDSC Expertise

Universal Scale Storage
Open for business from
200TB to 10's of PB

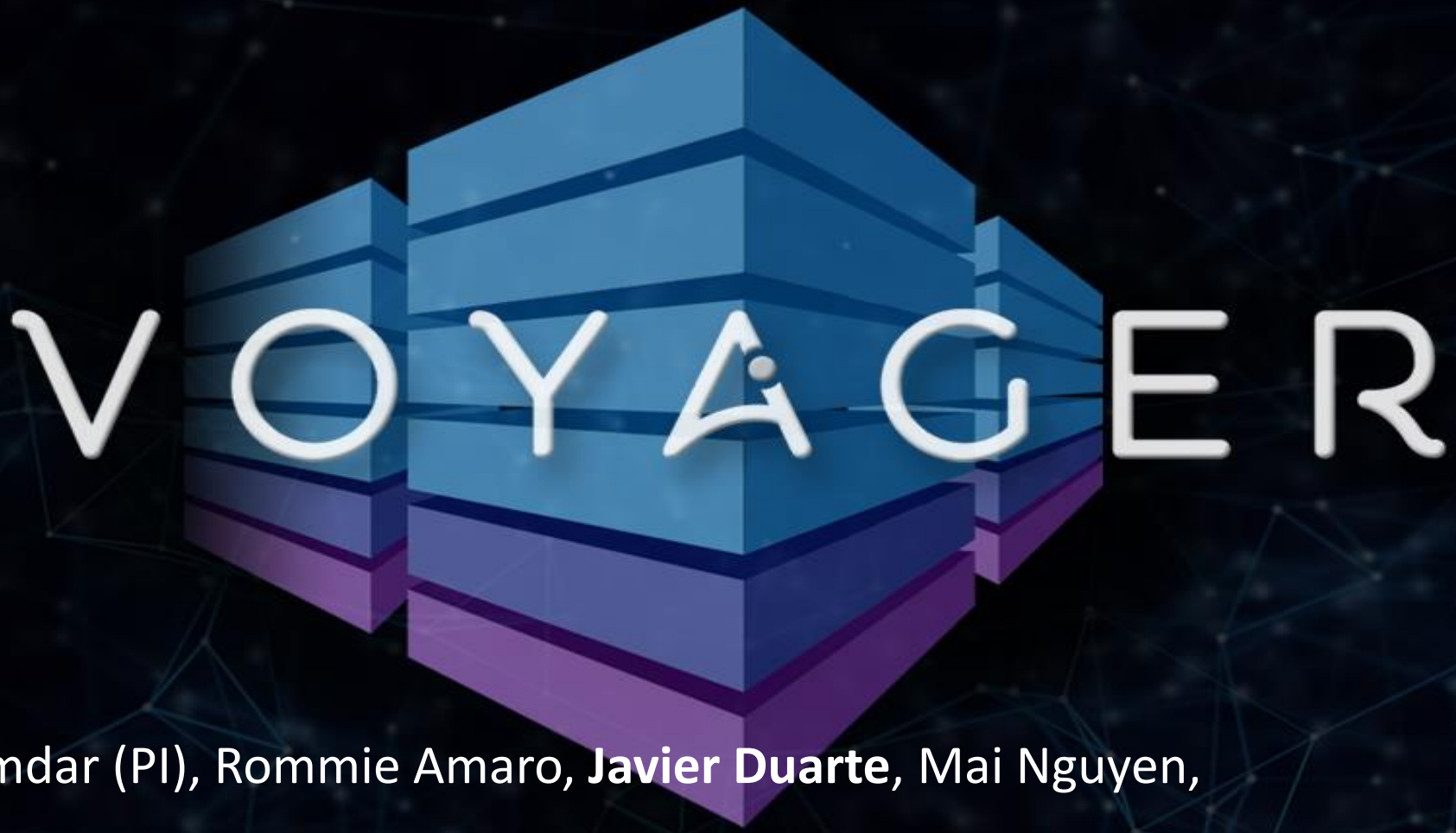
Globally Federated
Cyberinfrastructure
100++ institutions
on 5 continents

We design, deploy, and operate end-to-end solutions for our partners from academia, government, industry & non-profits

Machine Learning at SDSC

Voyager, NRP, and Expanse

Voyager: Exploring AI Processors in Science and Engineering



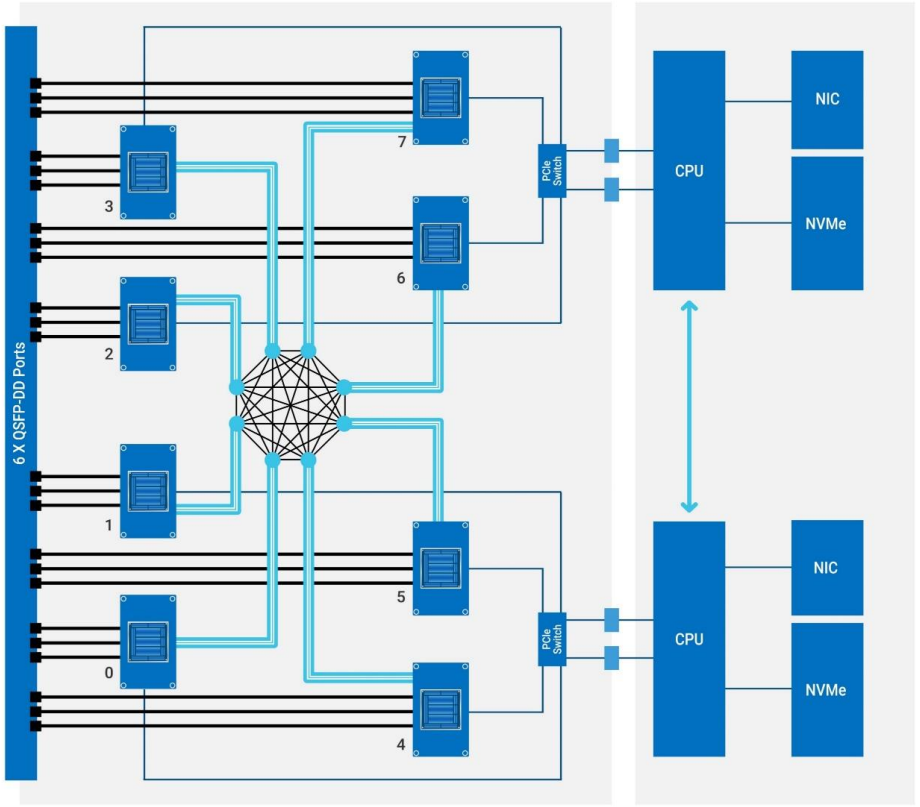
PIs: Amit Majumdar (PI), Rommie Amaro, **Javier Duarte**, Mai Nguyen, Bob Sinkovits
SDSC, UCSD

NSF Award 2005369

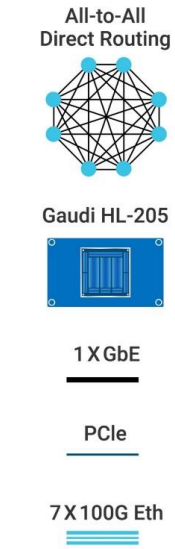
NSF Award 2005369

Voyager: Architected for ML

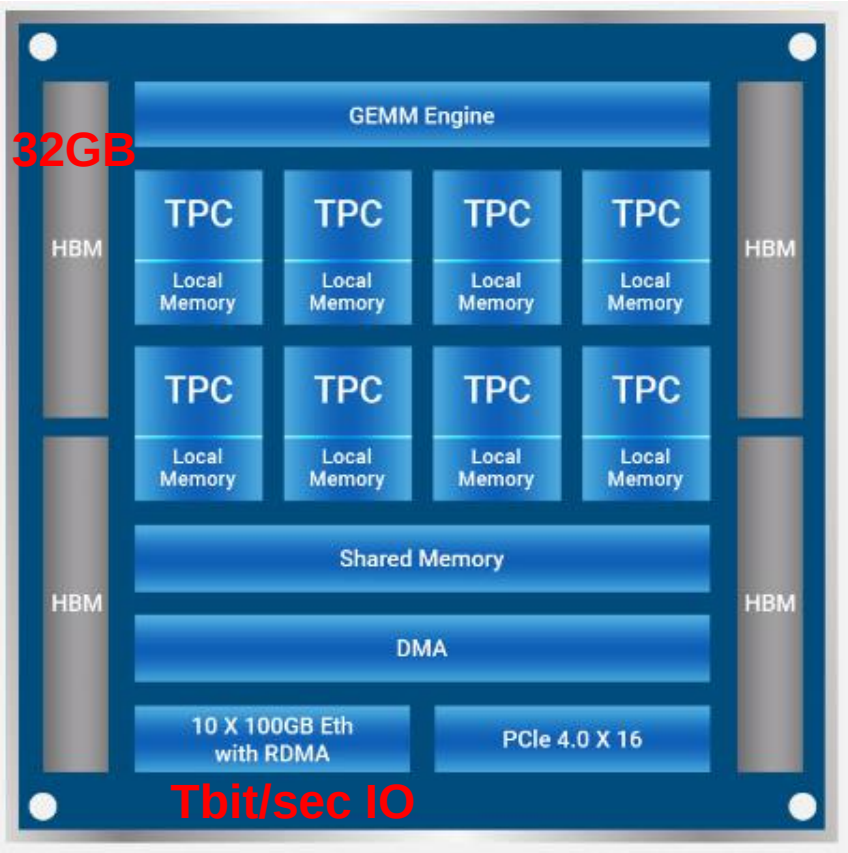
6x400G to node



8x Gaudi per node



7x100G all-2-all Gaudi inside a node



Gaudi Architecture

42 Nodes in Voyager all-2-all at 2.4 Tbit/sec

Habana on Training Language Models

- https://developer.habana.ai/blog/training-causal-language-models-on-sdscs-gaudi-based-voyager-supercomputing-cluster/
- The Habana developer team trained GPT-2 XL and GPT-3 XL on Voyager, showing scalability vs # of Gaudi devices used during the training.
  - Used up to 128 Gaudi devices
  - Voyager actually has 336 such devices, so scaling to 256 ought to be possible

Petabyte Inference on Voyager

- Together with the Habana team, we want to explore the use of Gaudi also for inference.
- We are actively looking for interesting science cases for “petabyte scale” inference on Voyager

32 Xilinx U55C
for ML Inference
@ SDSC

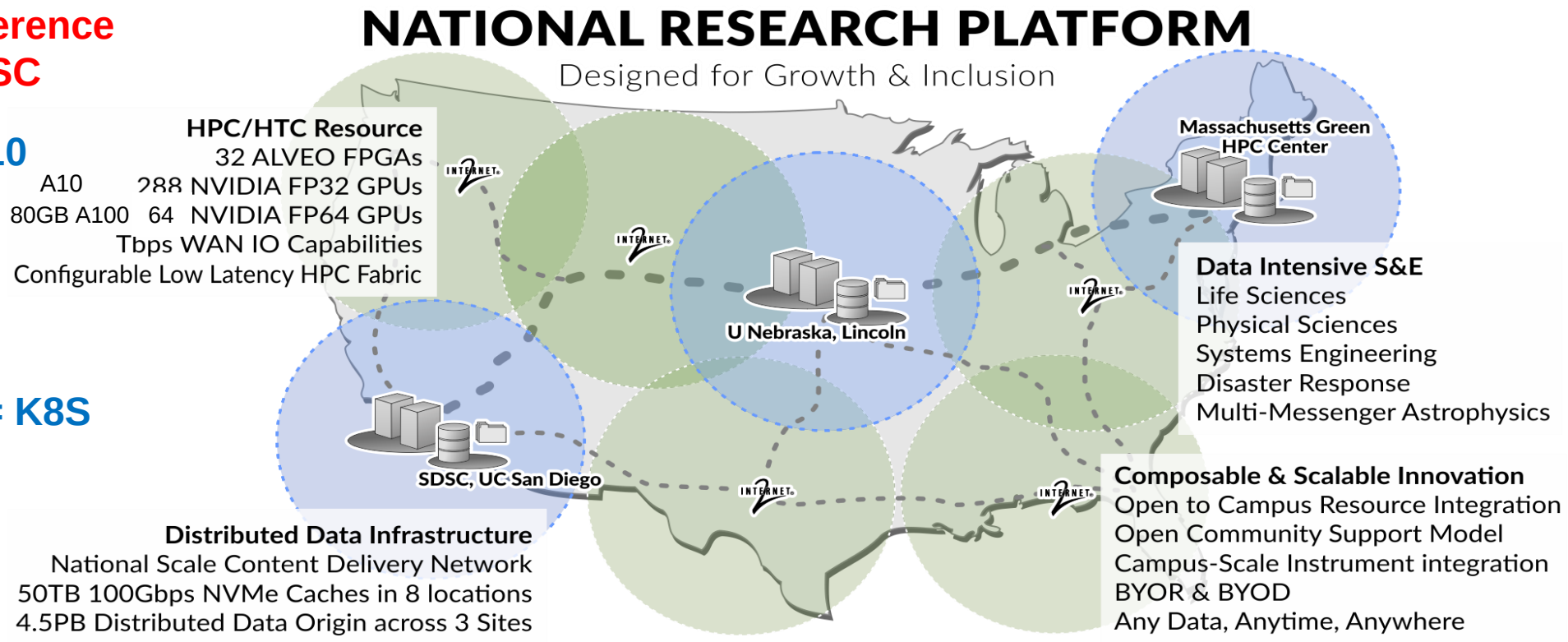
Funded as NSF 2112167

288 NVIDIA A10

Default UX = K8S

8 NVIDIA DGX

350TB NVMe



NRP is both an NSF funded project and a community supported infrastructure.

Growth goals of community infrastructure

1,000++ GPUs end of 2022

50 PB storage end of 2024

Lots of gaming GPUs
1080Ti to 3090

Expanse HPC Workhorse

- 13 Racks w 56 CPU nodes & 2x 4xV100 GPU nodes
  - 90k CPU cores allocated entirely via ACCESS program at NSF
  - Roughly 10,000 active unix accounts
- 1 industry rack
  - industry & academia collaborations on platform academics know and love
  - De-identified Human Genomics data allowed
- 2 PATh racks to support HTC using HTCondor … think OSG
  - More than 100M core hours a year to be had “for free” with nothing more than a 1 pager to your program manager at the NSF => PATh facility

Tbit/sec Network Architecture at SDSC

- NRP has its own 32x400G TOR switch
  - 2x100G for each of the 32 Xilinx U55C
  - 2x100G for each of the 8 DGX
  - 2x100G for each of the CPU nodes
- Nx400G from TOR to core HPC switch that connects to all other systems and the WAN
  - We will build out as neccessary
- 400G to the WAN
  - Caltech T2 reachable via “400G”
  - UCSD T2 reachable via 2x100G (build out as needed)

Managed Network Bandwidth

- Esnet, Caltech, UCSD are engaged in R&D towards ”Accountable” network bandwidth
- Imagine:
  - You have a production workflow running across Expanse/Caltech/UCSD at a scale of 12,000 cores, 4,000 cores at each location.
  - You know the network bandwidth this requires for SONIC, because you have benchmarked the IO per CPU.
  - You know the inference compute required to keep all 12,000 cores busy
  - **We manage the entire system of CPUs, WAN/LAN IO, and inference capacity as a composed system, avoiding any bottlenecks.**
    - **For 2023, a reasonable goal would be SONIC at up to**
      - **300G Caltech to NRP-FPGAs/A100**
      - **100G Expanse to NRP-FPGAs/A100**
      - **100G UCSD T2 to NRP-FPGAs/A100**

SDSC is the perfect playground for scaling out SONIC and alike

Standard compute jobs running on Expanse PATh racks (14k cores)
Or on one of the two US CMS T2s at Caltech or UCSD (~20k cores)

Inference running on 32 Xilinx U55C in NRP racks
And/or the various GPUs in SoCal (A100, V100, …)

Looking for collaboration in ML training, Inference, & composable systems R&D