So You Want to Do AI Today? Here's How to Do it at Scale in a Hurry!

Spring 2024

It’s Not Just About the Hardware

Research Computing at OSU

- Currently 95% distributed, with enclaves in various colleges.
- Preparing for new supercomputer in two years with Research Computing Taskforce 2030 and AI Advisory Group and AI Coordinating team.
- Datacenter will house the supercomputer and is built to hold Version 2 several years down the road.
- Networking to most campus buildings, including the CIC, will be 100Gb, with 100Gb connectivity from the campus to the Internet.
- Purchased several servers to prepare researchers on the new platform.
- Working with human resources office to develop a classification for Research Software Engineers / AI Engineers and plan on hiring up to 6 within the next year.

Research Computing at OSU

- Digital Research Infrastructure will provide research computing to the campus, with plans to deploy centralized services to the campus.
- Key to this will be both faculty lead steering committee and governance.
- DRI looks to have at least 50% of their personnel as researcher-facing, with the other 50% responsible for the day-to-day operations of the supercomputer, clusters, and storage systems.
- The idea is to offer a white-glove service for research computing, going beyond the HPC and storage.

- Travelled to Southern Methodist University and the University of Florida to better understand heir research computing and academic programs.

University Data Center

University Data Center Design Goals

1. Anchor a high-speed fiber network to all buildings and the Internet.
2. Provide space, power, and cooling to house expansion of High-Performance Computing (HPC).
3. Provide long term ability to grow power/cooling capacity.
4. Flexible design to meet current as well as anticipated future technology needs of the University.
5. Durable building construction to last more than 50 years.

Switchgear-A and Switchgear-B

- Dual underground feeds from Campus Power Grid
- Steps power down from 13,800 volts to 480 volts
- Distributes power to Mechanical Equipment and UPS Units for IT Gear

Notes:

- High voltage - Power lines running across country: 345,000 volts
- Medium voltage - City power lines: 13,800 volts
- Low voltage – Commercial Building distribution: 480 volts

UPS Room - 3 Uninterruptible Power Supply Units

- Receives Power Distribution from Low Voltage Switchgear
- UPS Batteries Provide Continuous Power to Critical IT Gear When Utility Power Fails.
- Allows Time For Generator to Start & Take Over The Load.
- Battery Run Time Design For 7-Minutes At Full Load.

Oregon State University

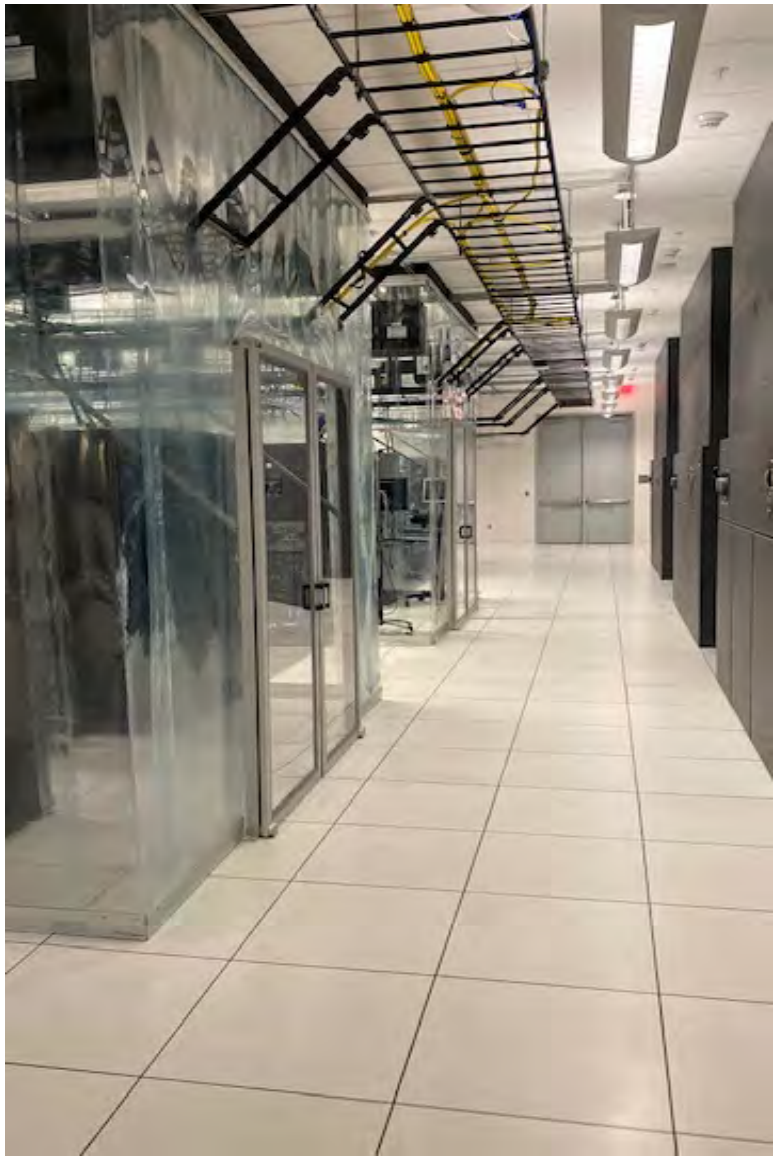
SMU

Information Technology Data Hall

Safe Home for IT Gear

- Cold Air Plenum under 3-Foot Raised Floor
- Hot Air Return Above Drop Ceiling

- Fire Suppression System: Double Action Water
- VESDA: Very Early Smoke Detection



Jen-Hsun and Lori Huang Collaborative Innovation Complex

Jen-Hsun and Lori Huang Collaborative Innovation Complex

Jen-Hsun and Lori Huang Collaborative Innovation Complex

ManeFrame I

March 19, 2014

ManeFrame II

» Dramatically increases the computational capability and performance available for data science and research

» Features state-of-the-art CPUs, accelerators, advanced networking technologies, and future-forward GPU-accelerated nodes and remote desktop capabilities

» Dedicated cluster helps to recruit research top faculty and students

Survey – Research Technology Services Satisfaction - 2018

RESEARCH SUPPORT SATISFACTION SPECIFICS

How satisfied are you with the following SMU research support service(s)?

PI Faculty

RESULTS	
High Performance Computing (HPC)	3.67
Technology consulting for research	3.44
Access to data scientists, data analysts, and data visualization specialists to help with research	2.54
Institutional repository of intellectual output (e.g. publications, prints, posters, etc.)	2.71
Digital preservation and curation of research data	2.64
Support for finding and using open content (course materials, texts, data sets, etc.)	2.79

SCALE Very Satisfied (5)—Satisfied (4)—Neutral (3)—Dissatisfied (2)—Very Dissatisfied (1)

SMU

OIT Research Technology Services

Eric Godat, PhD
Director
Data Scientist

Rob Kalescky, PhD
HPC Application
Scientist

John LaGrone,
PhD
HPC Application
Scientist

Tue Vu, PhD
AI/ML Research
Scientist

Guillermo Vasquez
Internet of Things
Developer

Mateo Langston-Smith
Human Trafficking
Data Analyst

OIT Research Technology Services

- Eric Godat, Ph.D.
- Director – Research Technology Services
- Ph.D. Theoretical Particle Physics – SMU ’18
- Data Scientist
- Signature Projects:
  - Data Science for Social Good
  - Digital Humanities Research Institute
  - LASSO Scheduling Algorithm
  - Quantifying Urban Resilience with Cell Phone GPS

OIT Research Technology Services

- Rob Kalescky, Ph.D.
- Research Technology Scientist
- Ph.D. Computational and Theoretical Chemistry – SMU ’14
- HPC Applications Scientist
- Signature Projects:
  - ChemGen
  - PyHOPs Memory Optimization and Parallelization
  - Gaussian, Local Modes, URVA Porting and Optimization
  - LASSO Scheduling Algorithm

OIT Research Technology Services

- John LaGrone, Ph.D.
- Research Technology Scientist
- Ph.D. Applied Mathematics – SMU ‘16
- HPC Applications Scientist
- Signature Projects:
  - NonInvasive Dance Scanning
  - Class HPC Containerization
  - BioSignal Analysis

OIT Research Technology Services

- Tue Vu, Ph.D.
- Research Technology Scientist
- Ph.D. Computational Hydro-Climatology
- AI/ML Applications Scientist
- Signature Projects:
  - Quantifying Urban Resilience with Cell Phone GPS
  - Personality Trait Characterization with LLMs
  - African Parliament Data Extraction
  - VR Training for Post-Partum Hemorrhage

OIT Research Technology Services

- Guillermo Vasquez
- Research Technology Scientist
- Internet of Things Developer
- Signature Projects:
  - NonInvasive Dance Sensors
  - Water Quality Helmet
  - Newton’s eApple
  - Biometric sensors for Mice

OIT Research Technology Services

- Mateo Langston Smith
- Human Trafficking Data Analyst
- Signature Projects:
  - Human Trafficking Data Warehouse
  - 12 Hills Interactive Dashboard
  - Dialogic Classroom Data Visualization
  - IoT GPS Tracking Module Dashboard

SuperPOD

ManeFrame III and SuperPOD

	ManeFrame III (2023)	NVIDIA DGX SuperPOD (2022)
Computational Ability	1000 TFLOPS	1,644 TFLOPS
Number of Nodes	200	20
CPU Cores (AVX2)	25,600	2,560
Total Accelerator Cores	0	1,392,640
Total Memory	112 TB	52.5 TB
Node Interconnect Bandwidth	200 Gb/s	10x200 Gb/s
Work Storage	3.4 PB	768 TB (Shared)
Scratch Space	3.4 PB	750 TB
Archive Capabilities	No	No
Operating System	Ubuntu 22.04	Ubuntu 20.04
GPUs	None	160 Nvidia A100s

HPC Software Stack

Applications:

- » R
- » aimall
- » amber
- » ansys/electronics
- » ansys/fluidstructures
- » cfour/mpi
- » cfour/nompi
- » charmm
- » cplex
- » demon
- » emacs
- » eog
- » evince
- » gamess
- » gedit
- » matlab
- » molpro
- » q-chem
- » quantum_atk
- » synopsys/photonicsolutions
- » tcad
- » texlive

Compilers:

- gcc/11.2.0
- intel/2023.1
- intel/oneapi/2023.2
- nvidia/21.3
- nvidia/23.5

Applications available via Spack:

- amd-aocl
- arpack-ng
- arrayfire
- bedtools2
- boost
- cuda
- cudnn
- cvs
- dalton
- ddd
- dos2unix
- freeglut
- g2c
- gcc
- gdb
- ghostscript
- git-lfs
- gnuplot
- graphviz
- gromacs
- imagemagick
- intel-mkl
- intel-oneapi-mkl
- intel-tbb
- jags
- kokkos
- krb5
- lammps
- llvm
- mesa
- molden
- mumps
- namd
- nccl
- ncurses
- netlib-lapack
- netlib-scalapack
- nvtop
- openblas
- opencv
- openfoam
- openfoam-org
- openmm
- openmolcas
- openmpi
- paraview
- patchelf
- psi4
- py-deepdiff
- py-pint
- py-pydantic
- py-reportseff
- py-scipy
- quantum-espresso
- root
- ruby
- samtools
- sextractor
- sratoolkit
- star
- tcl
- tcsh
- unixodbc
- valgrind
- yaml-cpp

Users can also install software themselves or request software be added by OIT RTS

HPC Software Stack – AI/ML

Applications:

- R
- aimall
- amber
- ansys/electronics
- ansys/fluidstructures
- cfour/mpi
- cfour/nompi
- charmm
- cplex
- demon
- emacs
- eog
- evince
- gamess
- gedit
- matlab
- molpro
- q-chem
- quantum_atk
- synopsys/photonicsolutions
- tcad
- texlive

Compilers:

- gcc/11.2.0
- intel/2023.1
- intel/oneapi/2023.2
- nvidia/21.3
- nvidia/23.5

Applications available via Spack:

- amd-aocl
- arpack-ng
- arrayfire
- bedtools2
- boost
- cuda
- cudnn
- cvs
- dalton
- ddd
- dos2unix
- freeglut
- g2c
- gcc
- gdb
- ghostscript
- git-lfs
- gnuplot
- graphviz
- gromacs
- imagemagick
- intel-mkl
- intel-oneapi-mkl
- intel-tbb
- jags
- kokkos
- krb5
- lammps
- llvm
- mesa
- molden
- mumps
- namd
- nccl
- ncurses
- netlib-lapack
- netlib-scalapack
- nvtop
- openblas
- opencv
- openfoam
- openfoam-org
- openmm
- openmolcas
- openmpi
- paraview
- patchelf
- psi4
- py-deepdiff
- py-pint
- py-pydantic
- py-reportseff
- py-scipy
- quantum-espresso
- root
- ruby
- samtools
- sextractor
- sratoolkit
- star
- tcl
- tcsh
- unixodbc
- valgrind
- yaml-cpp

Users can also install software themselves or request software be added by OIT RTS

AI/ML Projects

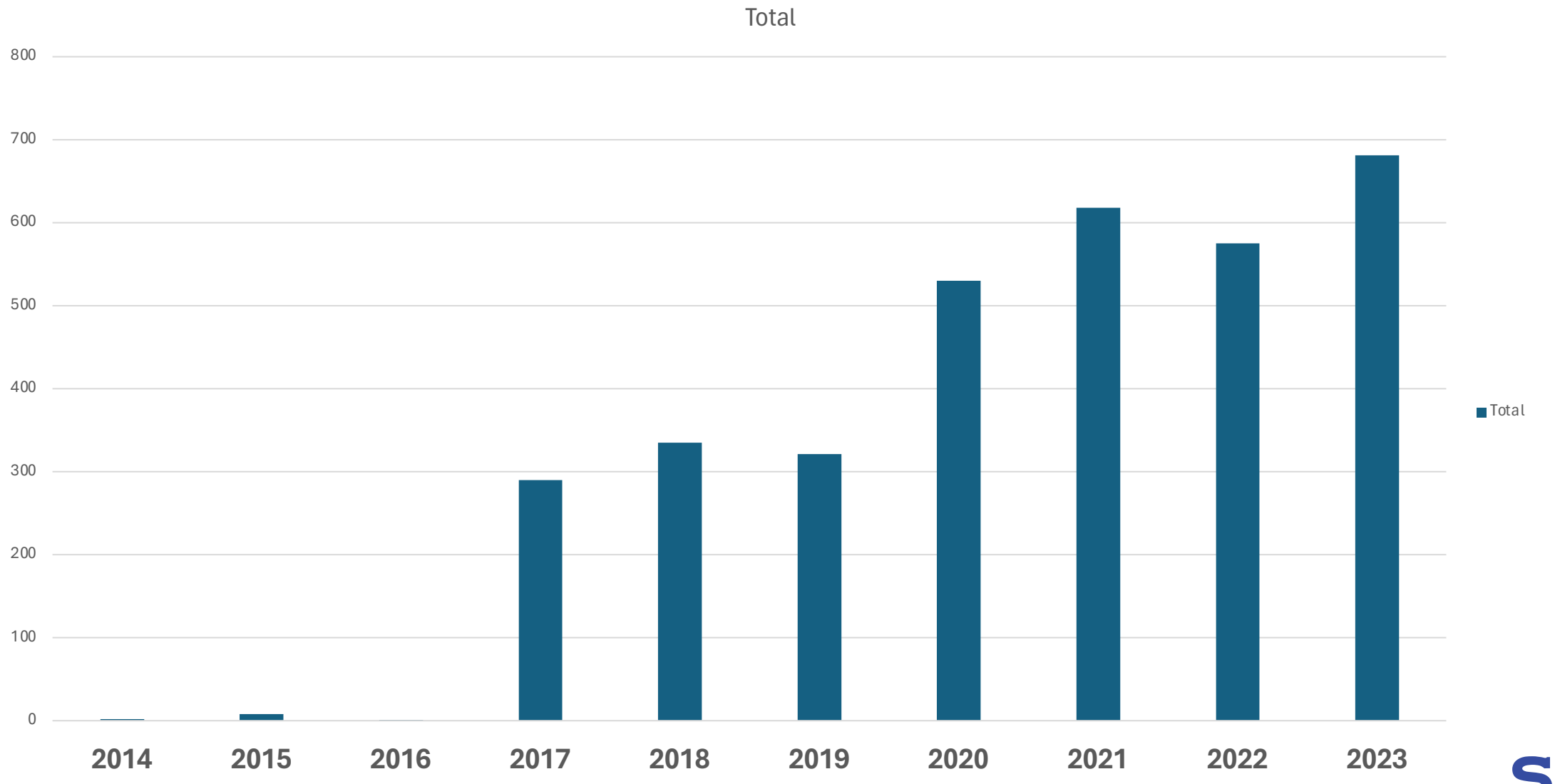
- Quantifying Urban Resilience with respected to natural hazard for multi-cities using Cellphone GPS
  - Big data project involving parallel processing and GPU utilization
  - Dask, cuDask, RAPIDS
- Characterize personality traits from life narrative interviews using Large Language Modeling
  - NLP with Transformer models
  - Pytorch, Transformers
- Web-scraping and Text mining to extract 14 African countries' parliament members for Political science research
  - Text mining with Name Entity Recognition
  - NLTK, spacy, NER, scikit-learn
- Benchmarking different GPU architectures for CIFAR100 image detection
  - Computer vision modeling
  - Tensorflow, Convolution Neural Network
- Finetuning chatbot model using Gutenberg's sacred library
  - Large Language Model application
  - Pytorch, Transformers, OpenAI, HuggingFace
- Route to school identification using cellphone data
  - Big data project involving parallel processing and GPU utilization
  - Dask, cuDask, RAPIDS
- Bus route to hospital optimization using cellphone GPS
  - Big data project involving parallel processing and GPU utilization
  - Dask, cuDask, RAPIDS
- Community traveling pattern using cellphone GPS
  - Big data project involving parallel processing and GPU utilization
  - Dask, cuDask, RAPIDS
- Using LASSO to identify the most sensitive parameters for bike share companies
  - Machine Learning project
  - scikit-learn, LASSO
- Using YOLO to detect zebra crossing at infrastructure deserted neighbourhood
  - Computer vision project
  - tensorflow, yolo

Research Technology Services Help Desk Tickets (2014-2023)

- Research Technology Services: those requests aligning to research technology support software, hardware, infrastructure, connectivity programming, consultation, etc.)
- Research Technology Services tickets include support and system administration for:
    1. HPC account management
    2. client computer management
    3. software support
    4. basic technical troubleshooting
    5. consultation/other
- 815 unique faculty, students, staff submitted 3361 research services help tickets
- 3361 research technology support tickets closed by 58 different OIT staff members
- Top 5 OIT research technology services team members:
    - Richard England (2413 research support tickets closed)
    - Amit Kumar (313 research support tickets closed)
    - John LaGrone (264 research support tickets closed)
    - Rob Kalescky (215 research support tickets closed)
    - Other OIT Personnel (156 research support tickets closed)

SMU

Research Technology Services Help Desk Tickets (2014-2023)



Research Technology Services: Projects (work requests requiring 10 or more hours)

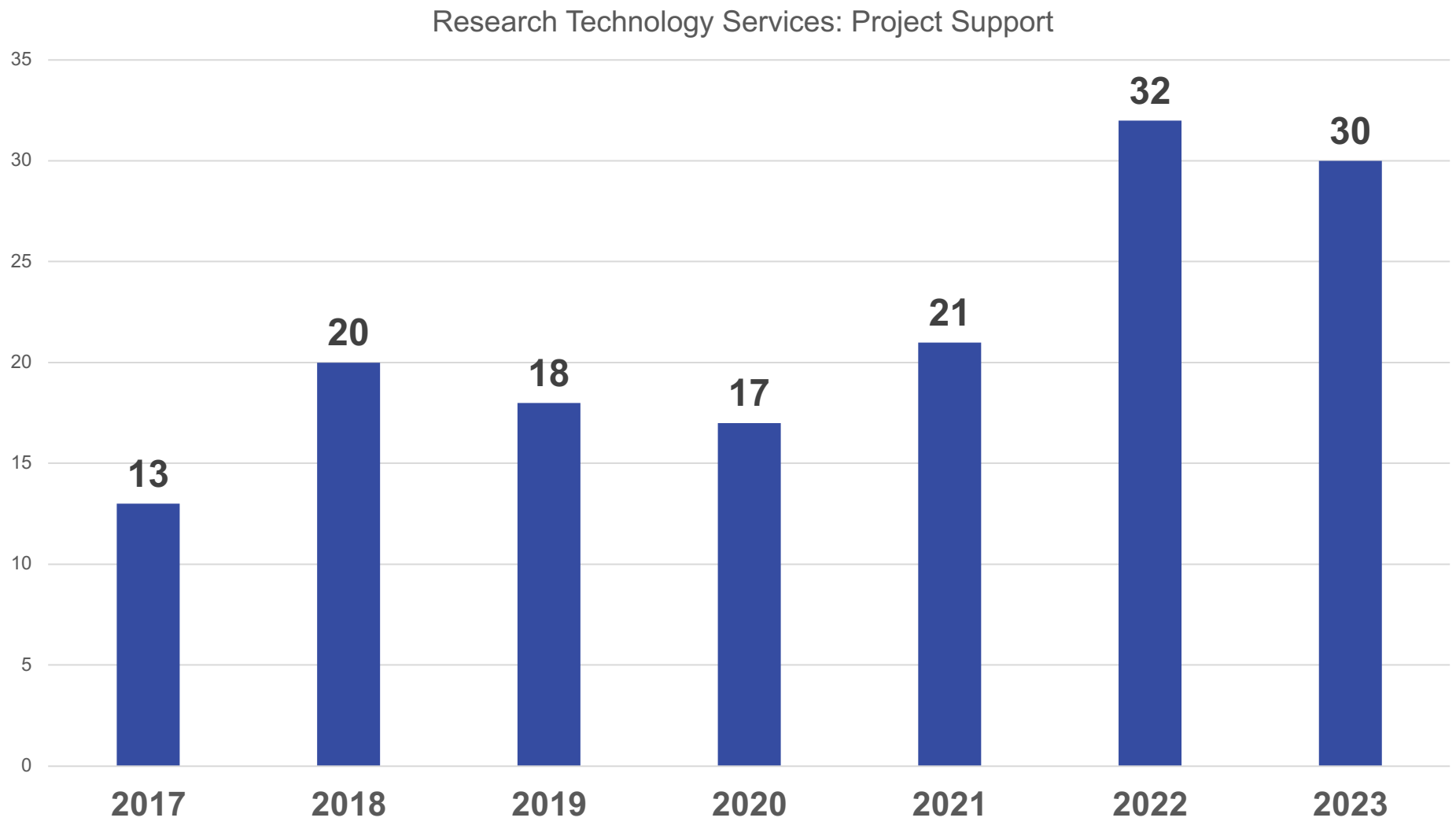
- 151 projects completed between 2017-2023
  - 56% projects relate to supporting other research technology & researchers
  - 44% projects include some support for HPC technologies & researchers
- 25% of projects relate to general research technology system administration & support
- 32% of project requests come from Dedman College
- 43% of project requests come from other schools/areas

  9% Simmons | 7% Lyle | 7% Libraries | 6% Meadows

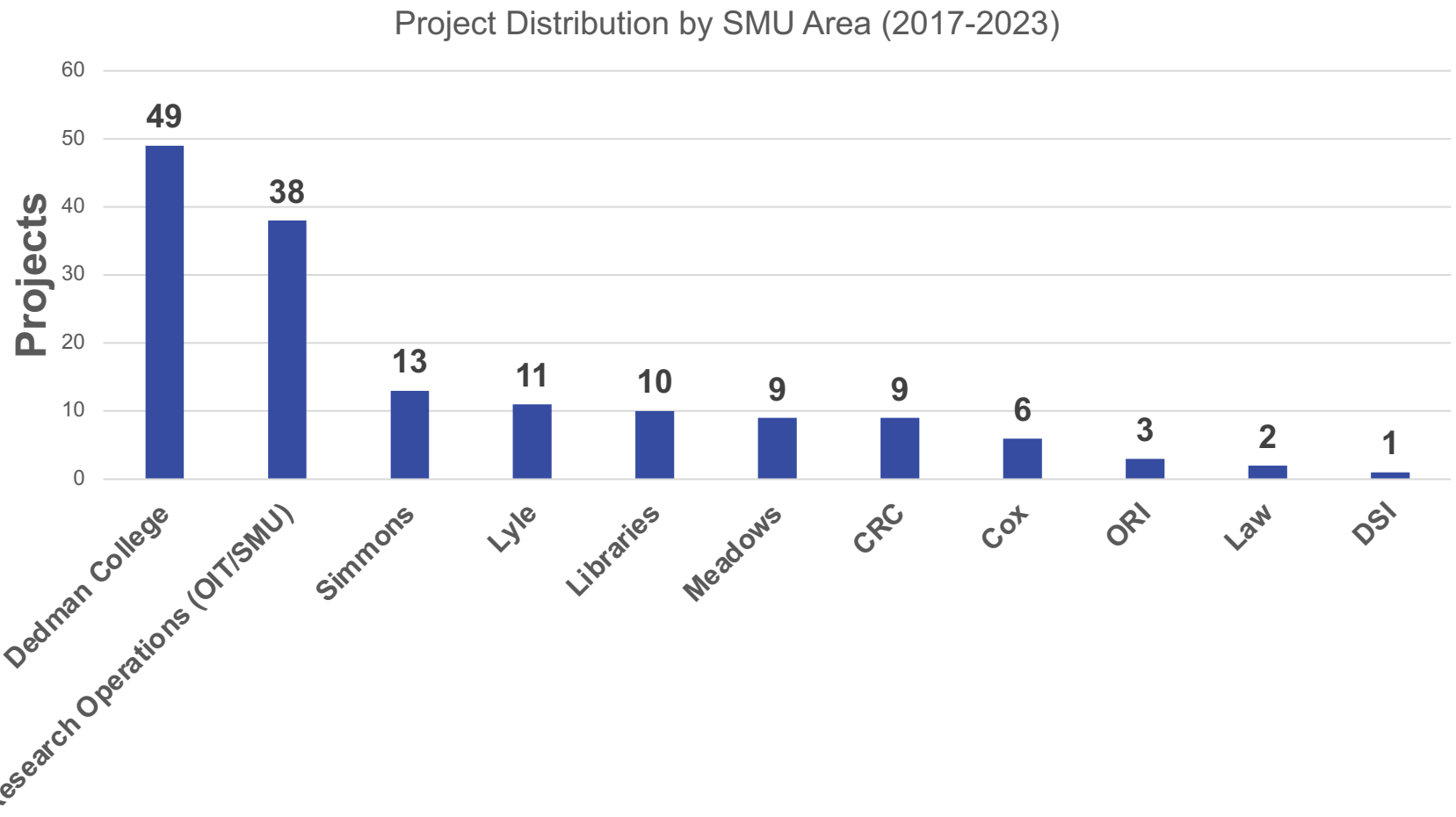
  6% CRC | 4% Cox | 2% ORI | 1% Law | 1 % DSI
- The OIT Research Technology Services team is involved in most projects and many other areas of OIT also contribute to research projects and provide technical SME as necessary
- Some PI's use grant funding to hire dedicated project-based OIT people

SMU

Research Technology Services: Projects (work requests requiring 10 or more hours)



OIT Research Support Project Distribution (areas)



- 32% of project work is for Dedman College
- 25% of project work is spent maintaining research systems, HPC, and other infrastructure for SMU
- 43% project work in:
  - 9% Simmons
  - 7% Lyle
  - 6% Meadows
  - 6% CRC
  - 4% Cox
  - 2% ORI
  - 1% Law
  - .7% DSI

How can OIT improve services and experiences for researchers?

1. Do a better job communicating availability of and how faculty can get help from OIT research support team
2. Develop closer personal relationships and deeper understanding of research needs
3. Reduce layers/barriers between researchers and relevant OIT SME's
4. Create comprehensive website "menu" of all OIT research services and solutions (with examples!)
5. More OIT-hosed lunches, seminars, workshops, road-shows, demonstrations, etc.—specifically for research service
6. Create a "researcher-pairing" or "research-community" portal where faculty can find research collaborators or information on specific projects, topics, software, partnerships, etc.

SMU

Igniting Artificial Intelligence Supercomputing in Dallas


SMU

SMU

NVIDIA

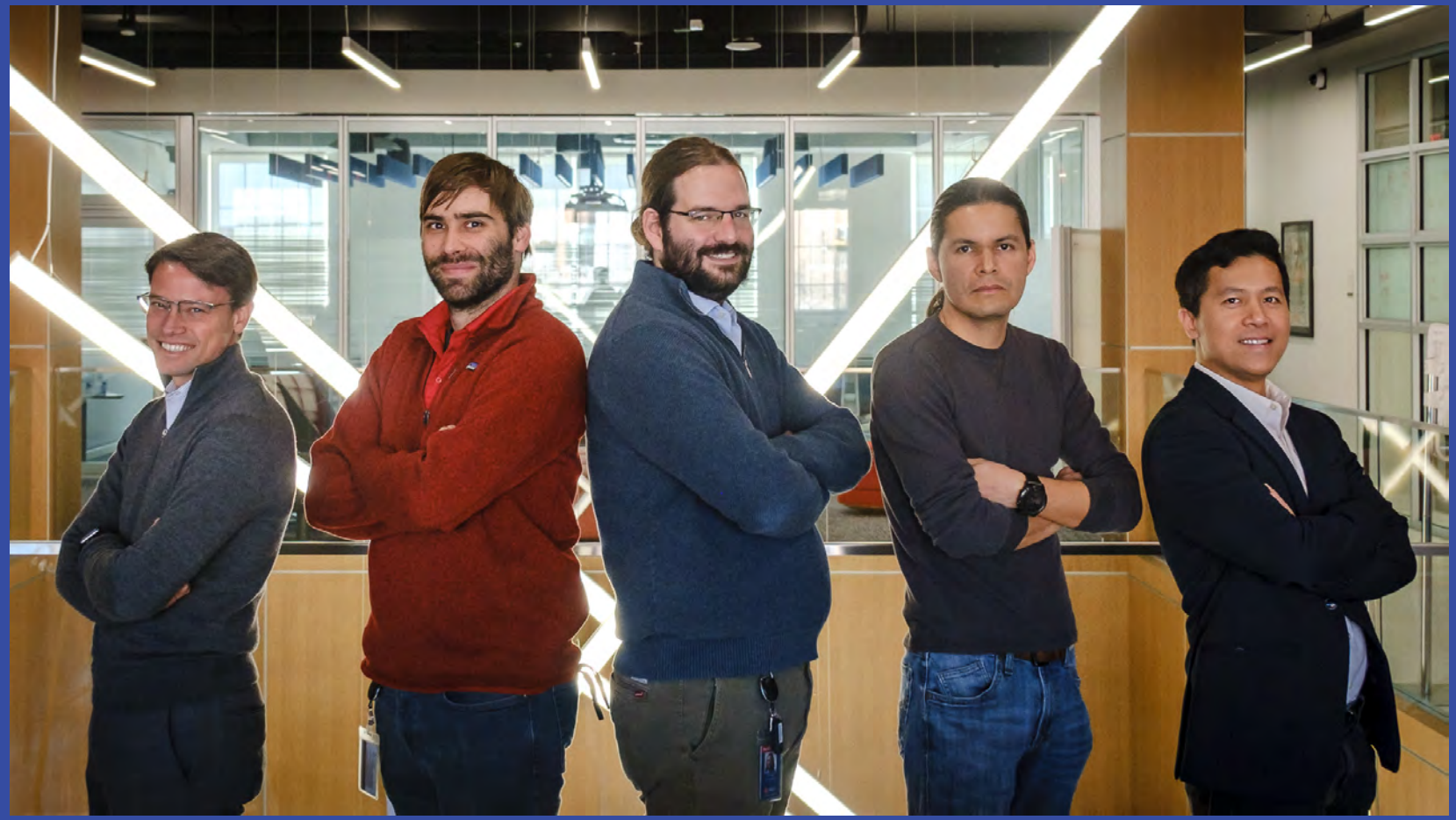
Home AI Data Center Driving Gaming Pro Graphics Robotics Healthcare Startups A

Tiny Computer, Huge Learnings: Students at SMU Build Baby Supercomputer With NVIDIA Jetson Edge AI Platform

The mini cluster will be on display at the SC22 supercomputing conference in Dallas, running Nov. 13-18.

November 7, 2022 by Angie Lee

Contact Us



egodat@smu.edu
smu.edu/OIT/research
Data Science Hub - Ford Hall 119

Prosperity Widely Shared

- Our training for the supercomputer started by leveraging existing relationship with NVIDIA for classes and workshops.
- We are able to share these classes and workshops with regional universities and colleges.
- We plan to share the supercomputer itself with regional universities, colleges, and K-12 for academic purposes.

Harvard Business Review 13 Principles

- Informed Consent
- Aligned Interests
- Opt In and Easy Exits
- Conversational Transparency
- Debiased and Explainable
- AI Training and Development
- Health and Well-Being
- Data Collection
- Data Sharing
- Privacy and Security
- Third Party Disclosure
- Communication
- Laws and Regulations

Thank You