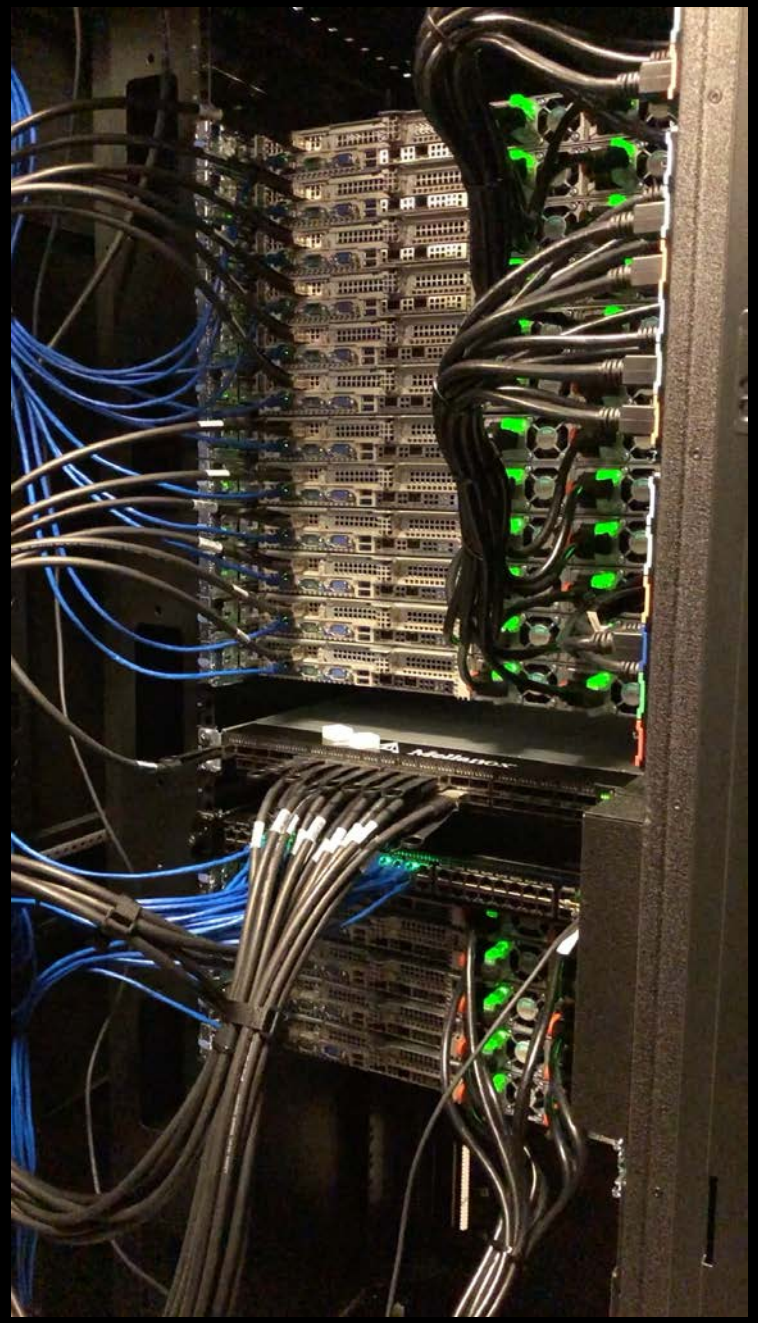
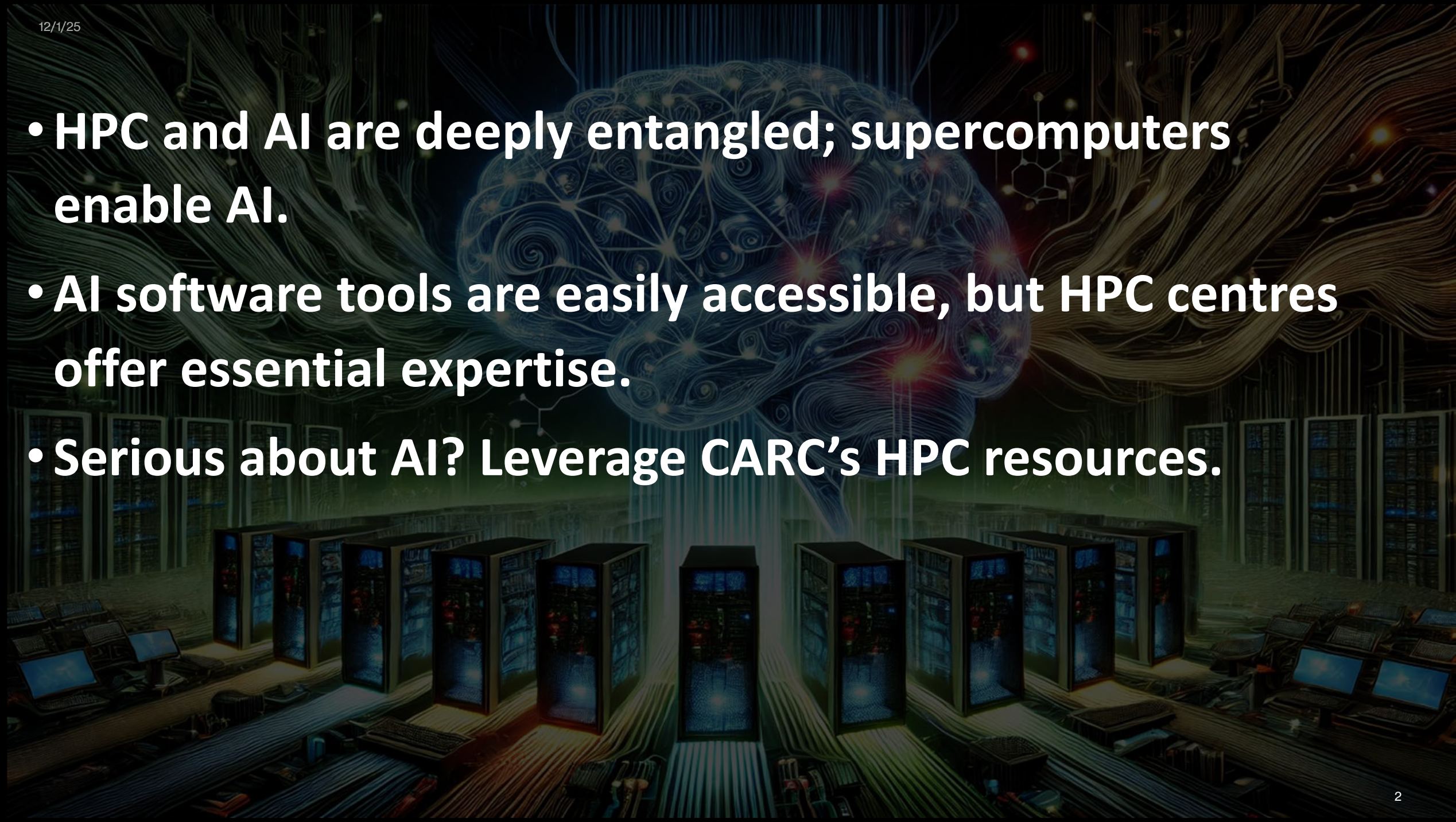
CARC Artificial Intelligence Pipelining Support

Matthew Fricke, DALL-E, and ChatGPT 4o

Computer Science and Center for Advanced Research Computing





- HPC and AI are deeply entangled; supercomputers enable AI.
- AI software tools are easily accessible, but HPC centres offer essential expertise.
- Serious about AI? Leverage CARC's HPC resources.

Meet the AI robot whose artwork sold for over $1m

A portrait of mathematician Alan Turing is thought to be the first artwork by a humanoid robot to be sold at auction - fetching $1,084,800 (£836,667).

Lawmaker uses AI voice clone to address Congress

Virginia Congresswoman Jennifer Wexton used a artificial intelligence (AI) programme to address the House on Thursday. A year ago, the lawmaker was diagnosed with progressive supranuclear palsy, which makes it difficult for her to speak.

The AI programme allowed Wexton to make a clone of her speaking voice using old recordings

Advanced Research Projects Agency

- On October 4, 1957, the Soviet Union (USSR) launched the first satellite ever, shocking the United States. This was a major victory for the Soviets in the cold war.
- In response the US founded ARPA (Advanced Research Projects Agency) in February of the following year.
- ARPA's mission was transformational technologies (like the first space satellite).
- Periodically US Presidents add or subtract Defense from the front (DARPA).

ARPA (1958)

Formed to counter the USSR technology advantage

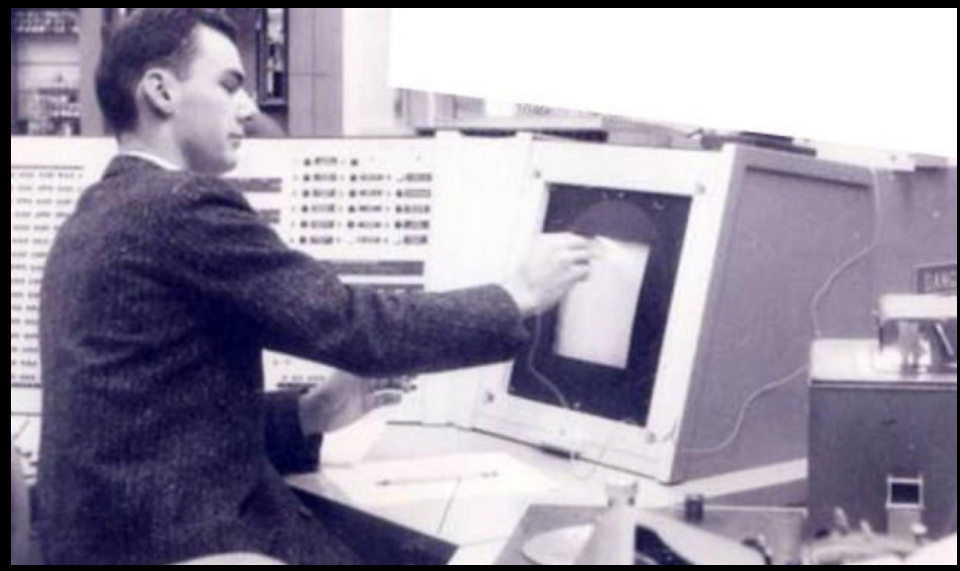
ARPA Information Processing Techniques Office (IPTO)

Artificial Intelligence was the goal from the start.

Bob Taylor (Psychology and CS)

Leonard Kleinrock (UCLA Prof CS)

Joseph Licklider (Psychology and CS)



Larry Roberts

ARPA - Computer Assistants for Everyone

“The hope is that, in not too many years, human brains and computing machines will be coupled together very tightly, and that the resulting partnership will think as no human brain has ever thought and process data in a way not approached by the information-handling machines we know today.”

Joseph Licklider (Psychology and CS), Director ARPA Information Processing Techniques Office (IPTO)

Licklider, Joseph CR. "Man-computer symbiosis." *IRE transactions on human factors in electronics* 1 (1960): 4-11.

“Present-day computers are designed primarily to solve preformulated problems or to process data according to predetermined procedures. [North’s* “Mechanically Extended Man”]

we are interested in a different relationship, one that depends on the realization of two main aims:

(1) to let computers facilitate formulative thinking as they now facilitate the solution of formulated problems, and

(2) to enable men and computers to cooperate in making decisions and controlling complex situations without inflexible dependence on predetermined programs.”

-Licklider 1960

*Richard D. North - Institute of Economic Affairs (IEA)

“Present-day computers are designed primarily to solve preformulated problems or to process data according to predetermined procedures. [North’s* “Mechanically Extended Man”]

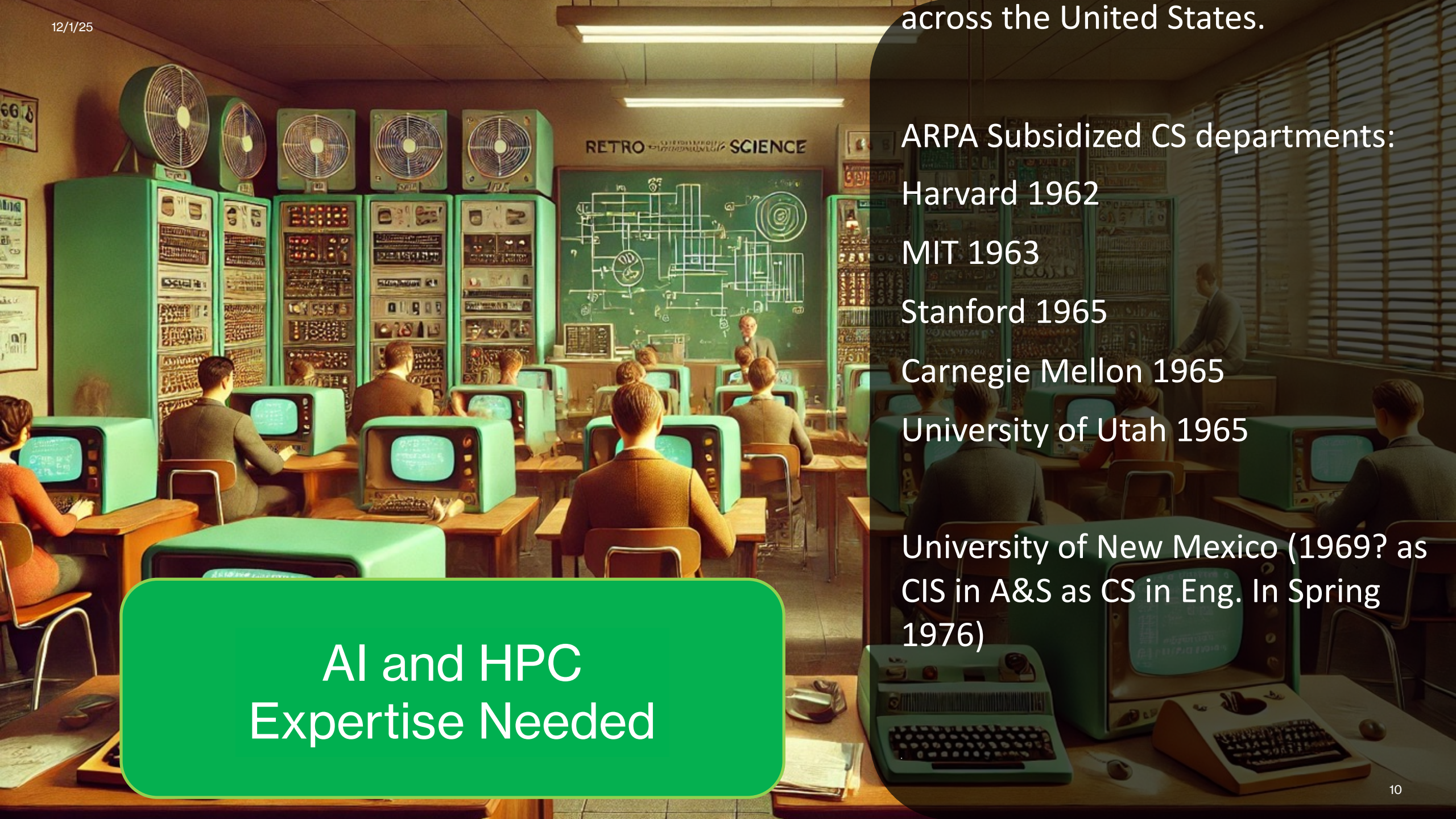
Data Driven Programming

Machine Learning

(2) to enable men and computers to cooperate in making decisions and controlling complex situations without inflexible dependence on predetermined programs.”

-Licklider 1960

*Richard D. North - Institute of Economic Affairs (IEA)



across the United States.

ARPA Subsidized CS departments:

Harvard 1962

MIT 1963

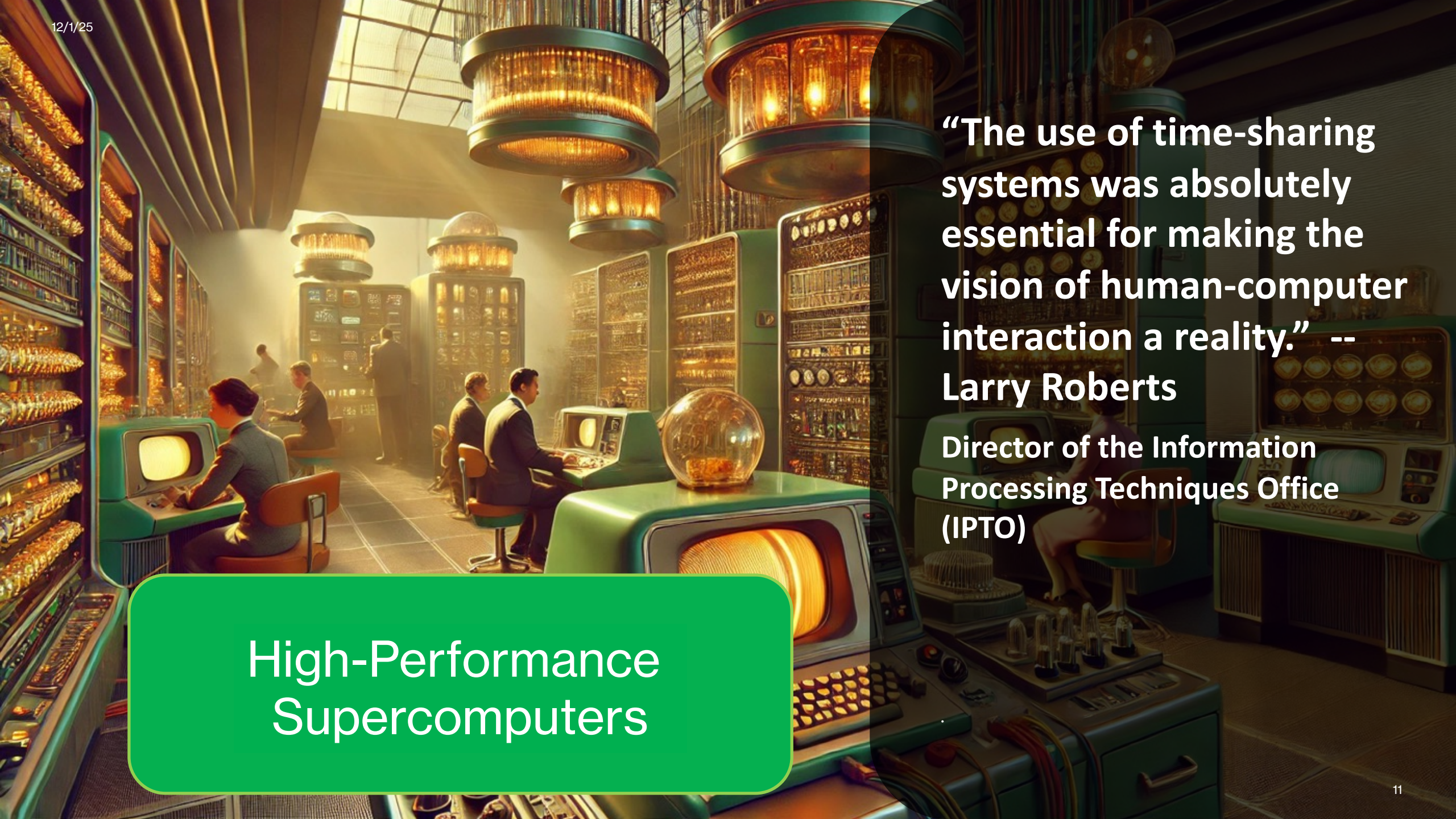
Stanford 1965

Carnegie Mellon 1965

University of Utah 1965

University of New Mexico (1969? as CIS in A&S as CS in Eng. In Spring 1976)

AI and HPC
Expertise Needed



High-Performance Supercomputers

“The use of time-sharing systems was absolutely essential for making the vision of human-computer interaction a reality.” -- Larry Roberts

Director of the Information Processing Techniques Office (IPTO)

‘Intergalactic Computer Network’*

The internet to provide access to time-sharing computers

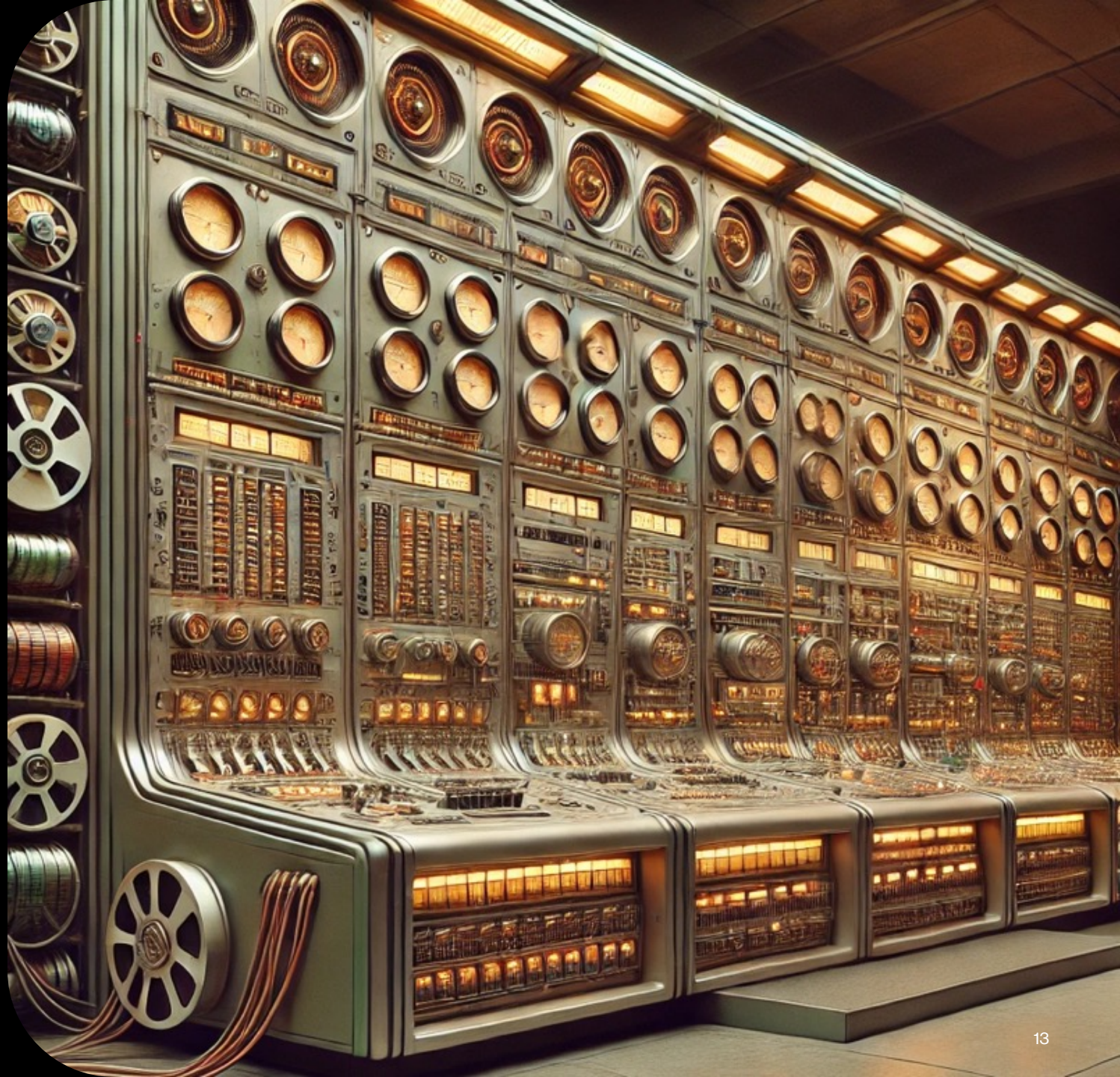
*1963 ARPA Memo from Licklider to: “Members and Affiliates of the Intergalactic Computer Network”

Time-Sharing Computers

Artificial Intelligence requires cutting edge hardware (supercomputers).

These are not desktop computers.

.



Time-Sharing Computers

Artificial Intelligence requires cutting edge hardware (supercomputers).

These are not desktop computers.

.

CARC

Established in 1993 as part of the DoD High Performance Computing Modernization program

MAUI HPC Center.

Now the UNM Center for Advanced Research Computing.

Mission: Provide cutting edge computing resources to UNM researchers

Funding:

UNM Office for the Vice President for Research (OVPR)

Grants e.g Department of Education - $1.5M.

CARC Director: Patrick Bridges
Professor Computer Science

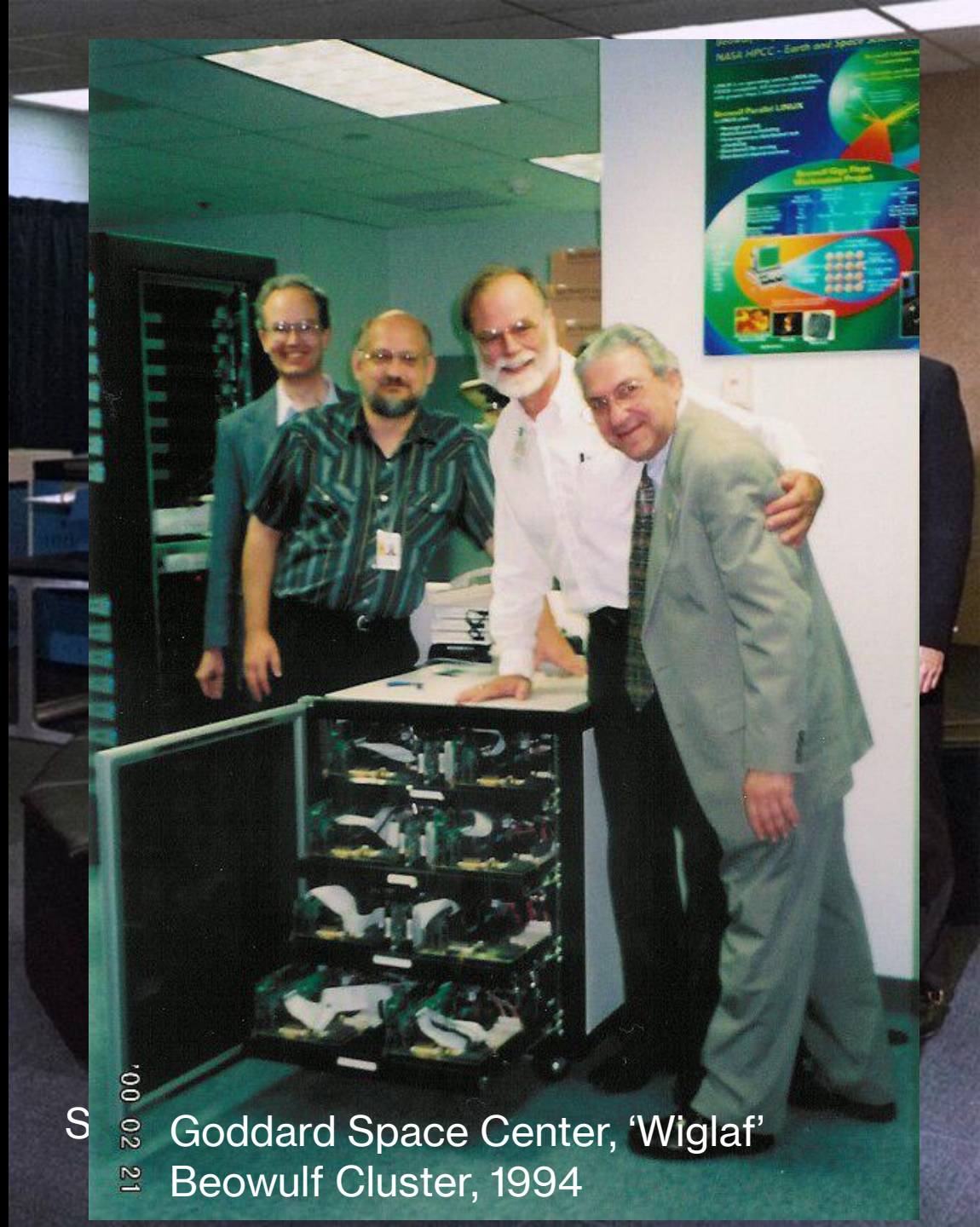
In 1993 NASA developed the first Beowulf cluster.

These are Linux based clusters of many computers.

At the time Cray Computers dominated supercomputers with Monolithic machines.

Seymore Cray was not impressed with Linux Clusters:

“If you were plowing a field, which would you rather use? Two strong oxen or 1024 chickens?”



Goddard Space Center, ‘Wiglaf’
Beowulf Cluster, 1994

In 1993 NASA developed the first Beowulf cluster.

These are Linux based clusters of many computers.

At the time Cray Computers dominated supercomputers with Monolithic machines.

Seymore Cray was not impressed with Linux Clusters:

“If you were plowing a field, which would you rather use? Two strong oxen or 1024 chickens?”

Seymore Cray with Cray-1 (1976)

In 1993 NASA deve

These are Linux ba
computers.

At the time Cray Co
supercomputers wi

Seymore Cray was
Clusters: “If you we
you rather use? Tw
chickens?”

CARC Demonstrated the effectiveness of Beowulf Clusters in 1999.

There was a ‘bakeoff’ between the NCSA’s Windows cluster and CARC’s Roadrunner Linux Cluster. Roadrunner won.

Roadrunner 1999

CA
effe
Clu

ARTICLE

Design and Analysis of the Alliance/University of New Mexico Roadrunner Linux SMP SuperCluster

Authors: David A. Bader, Arthur B. MacCabe, Jason R. Mastaler, John K. McIver, III, Patricia A. Kovatch Authors Info & Claims

IWCC '99: Proceedings of the 1st IEEE Computer Society International Workshop on Cluster Computing • Page 9

Published: 02 December 19 Publication History

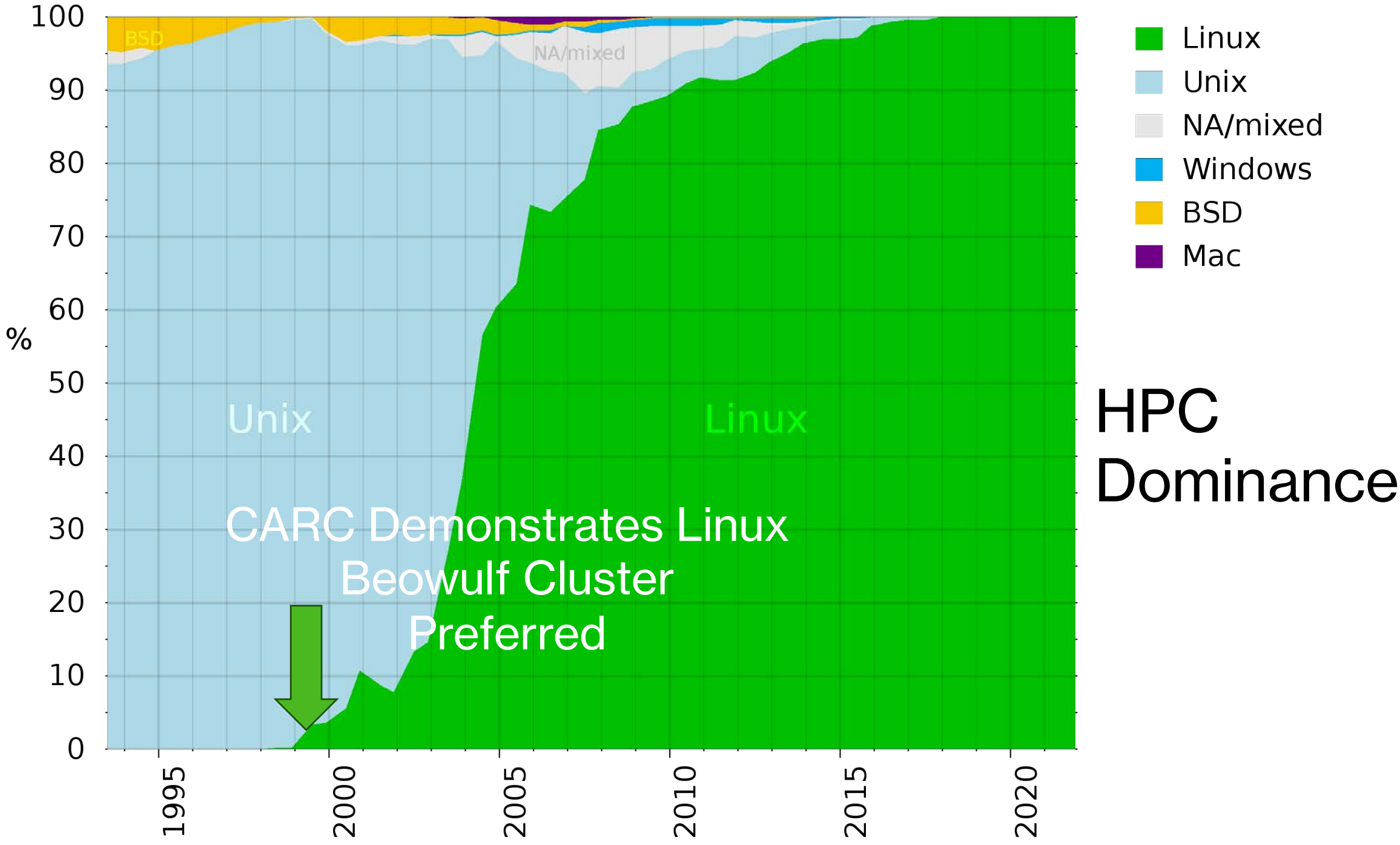
1 0

Publisher Site

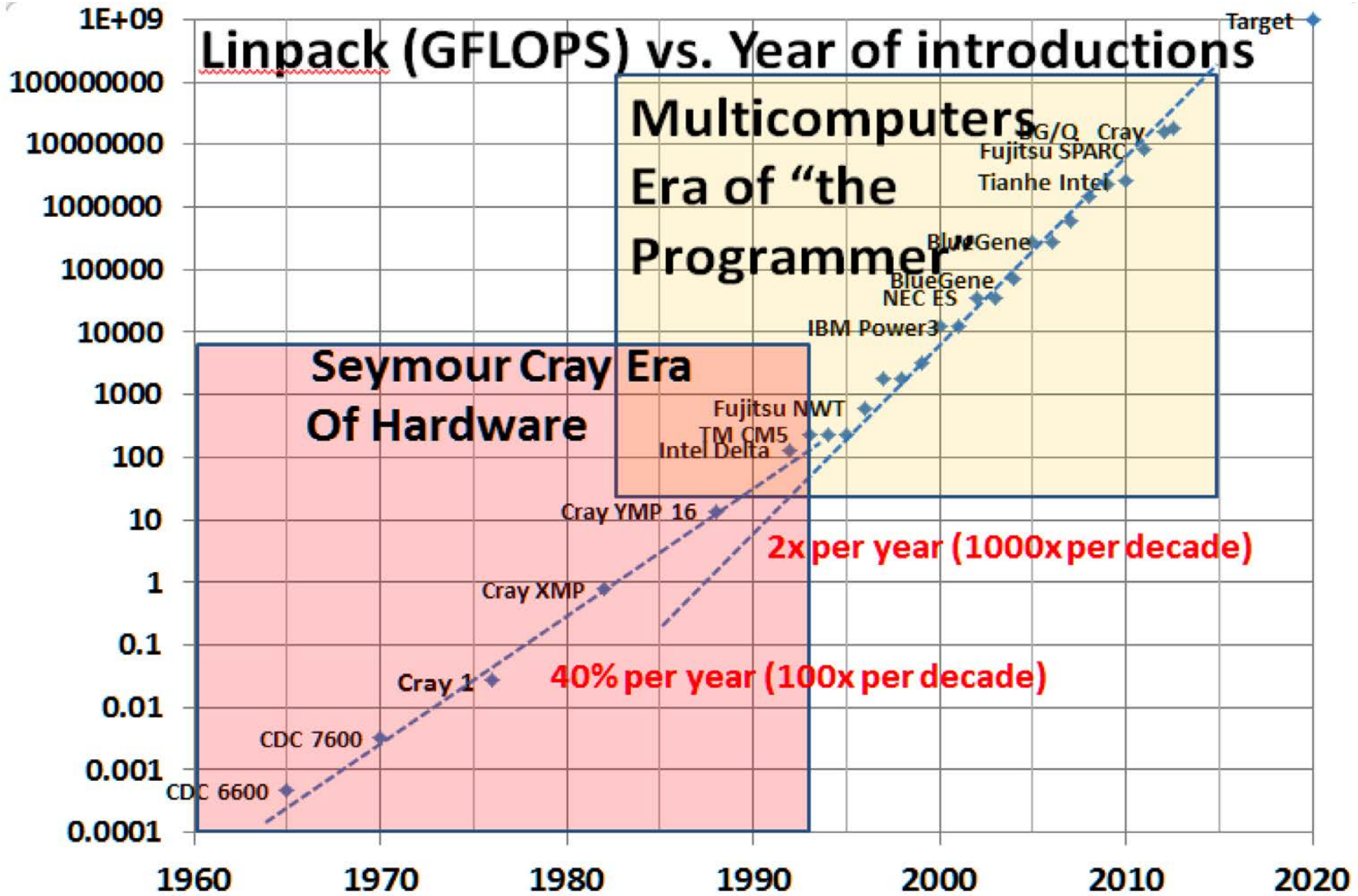
Abstract

This paper will discuss high performance clustering from a series of critical topics: architectural design, system software infrastructure, and programming environment. This will be accomplished through an overview of a large scale, high performance SuperCluster (named Roadrunner) in production at The University of New Mexico (UNM) Albuquerque High Performance Computing Center (AHPCC). This SuperCluster, sponsored by the U.S. National

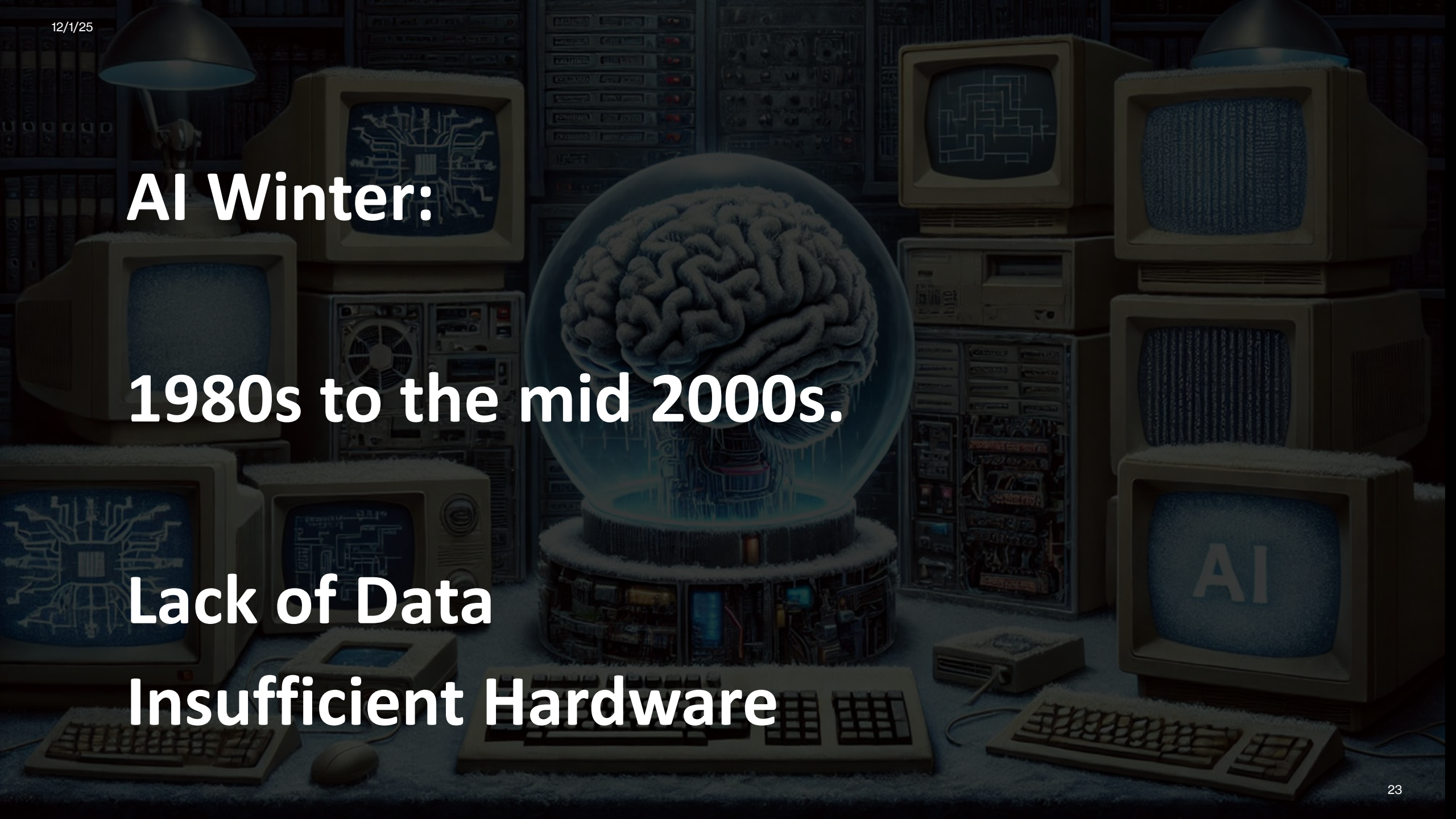
laf'



HPC Dominance



Gordon Bell, Supercomputers: The Amazing Race: DOI: 10.13140/2.1.3425.0562



AI Winter:

1980s to the mid 2000s.

Lack of Data
Insufficient Hardware

Machine Learning is Mostly Linear Algebra

PCA:

$$\mathbf{C} = \frac{1}{m}\mathbf{X}^\top \mathbf{X}$$

SVM:

$$\mathbf{w}^\top \mathbf{x} + b = 0$$

Linear Regression:

$$\hat{y} = \mathbf{X}\mathbf{w} + \mathbf{b}$$

$$J(\mathbf{w}, b) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2$$

Back Propagation:

$$\mathbf{W}^{[1]} \leftarrow \mathbf{W}^{[1]} - \alpha \frac{\partial L}{\partial \mathbf{W}^{[1]}}$$

Gradient Descent Update Rule:

$$\mathbf{w} \leftarrow \mathbf{w} - \alpha \cdot \frac{1}{m}\mathbf{X}^\top(\hat{y} - \mathbf{y})$$

$$b \leftarrow b - \alpha \cdot \frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})$$

Including Large Language Models

Example tokenisation:

Tokens: [‘centre’, ‘for’, ‘advanced’, ‘research’, ‘computing’]

Input Tokens:

$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$

Embed in higher dimensional space:

$\mathbf{X} \in \mathbb{R}^{n \times d}$

Compute Query, Key, and Value Matrices from Learned Transforms

$\mathbf{Q} = \mathbf{X}\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^V$

Scaled Dot-Product Attention:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

$\mathbf{A}$ is the ’attention’ the query token gets from other tokens to produce output.

Including Large Language Models

Example tokenisation:

Tokens: [‘centre’, ‘for’, ‘advanced’, ‘research’, ‘computing’]

Input Tokens:

$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$

Embed in higher dimensional space:

$\mathbf{X} \in \mathbb{R}^{n \times d}$

Compute Query, Key, and Value Matrices from Learned Transforms

$\mathbf{Q} = \mathbf{X}\mathbf{W}^Q, \quad \mathbf{K} = \mathbf{X}\mathbf{W}^K, \quad \mathbf{V} = \mathbf{X}\mathbf{W}^V$

Scaled Dot-Product Attention:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

$\mathbf{A}$ is the ’attention’ the query token gets from other tokens to produce output.

And Diffusion Image Generation

Generated Image Diffusion Model

$$x_t = \sqrt{\alpha_t}\, x_0 + \sqrt{1 - \alpha_t}\, \epsilon$$

$$E_y = \text{Encoder}(y)$$

LLM Encoder of Natural Language Prompt

$$x_{t-1} = \frac{x_t - (1 - \alpha_t) f_\theta(x_t, t, E_y)}{\sqrt{\alpha_t}} + \sigma_t\, \epsilon$$

$$\hat{f}\theta(x_t, t, E_y) = w \cdot f\theta(x_t, t, E_y) + (1 - w) \cdot f_\theta(x_t, t)$$

Reverse Diffusion using attention and Text Tokens generates Image

TOP 500
The List.

acm TURING AWARD SINCE 1966
CELEBRATING 50 YEARS
OF COMPUTING'S GREATEST ACHIEVEMENTS

Jack Dongarra
(Far left)

New Mexico

Clev Moler's PhD Student

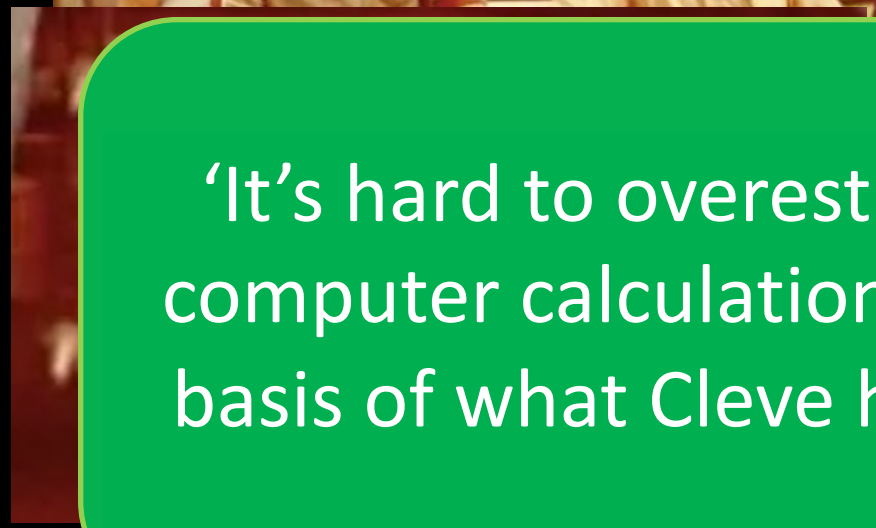
ScaLAPACK: A Portable Linear Algebra Library for Distributed Memory Computers - Design Issues and Performance *
(Technical Paper)

J. Choi†, J. Demmel‡, I. Dhillon‡, J. Dongarra§,
S. Ostrouchov†, A. Petitet†, K. Stanley‡, D. Walker¶ and R. C. Whaley†

Computer Physics Communications 97.1-2 (1996): 1-15.

Abstract

This paper outlines the content and performance of ScaLAPACK, a collection of mathemat-

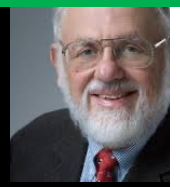


ScaLAPACK: A Portable Linear Algebra Library for Distributed Memory Computers - Design Issues and Performance *
(Technical Paper)

‘It’s hard to overestimate his impact on the world. Almost any computer calculations associated with matrices are done on the basis of what Cleve helped create’ – Jack Little, CEO Mathworks

These were specifically HPC Linear Algebra Libraries

New Mexico



Clev Moler’s PhD Student

define a theoretical model of parallel computers dedicated to linear algebra applications: the Distributed Linear Algebra Machine (DLAM). This model provides a convenient framework for developing parallel algorithms and investigating their scalability, performance and programmability. Extensive performance results on various platforms are presented and analyzed with the help of the DLAM. Finally, this paper briefly describes future directions for the ScaLAPACK library and concludes by suggesting alternative approaches to mathematical libraries, explaining how ScaLAPACK could be integrated into efficient and user-friendly distributed systems.

Machine Learning at UNM Cancer Center and CARC

- CARC resources
- Leukemia
- Susan Atlas

Blood, The Journal of the American Society of Hematology 115.7 (2010): 1394-1405.

LYMPHOID NEOPLASIA

Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia

Huining Kang,[1] I.-Ming Chen,[1,2] Carla S. Wilson,[1] Edward J. Bedrick,[1] Richard C. Harvey,[1] Susan R. Atlas,[1] Meenakshi Devidas,[2,3] Charles G. Mullighan,[4] Xuefei Wang,[1] Maurice Murphy,[1] Kerem Ar,[1] Walker Wharton,[1] Michael J. Borowitz,[2,5] W. Paul Bowman,[2,6] Deepa Bhojwani,[7] William L. Carroll,[2,7] Bruce M. Camitta,[2,8] Gregory H. Reaman,[2,9] Malcolm A. Smith,[10] James R. Downing,[4] Stephen P. Hunger,[2,11] and Cheryl L. Willman[1,2]

[1]University of New Mexico Cancer Center and Departments of Pathology, Internal Medicine, Mathematics and Statistics, and Physics and Astronomy, University of New Mexico, Albuquerque; [2]Children's Oncology Group, Arcadia, CA; [3]Children's Oncology Group Statistics and Data Center, and Department of Epidemiology and Health Policy Research, College of Medicine, University of Florida, Gainesville; [4]Department of Pathology, St Jude Children's Research Hospital, Memphis, TN; [5]Department of Pathology, Johns Hopkins Medical Institutions, Baltimore, MD; [6]Cook Children's Medical Center, Forth Worth, TX; [7]Departments of Pediatrics, Hematology, and Oncology and Cancer Center, New York University Medical Center, NY; [8]Departments of Pediatrics, Hematology, Oncology, and Transplantation, Medical College of Wisconsin, Milwaukee; [9]Department of Hematology-Oncology, Children's National Medical Center, Washington, DC; [10]Pediatric Oncology, Clinical Investigations Branch, National Cancer Institute, Bethesda, MD; and [11]Children's Hospital, Department of Pediatrics, and University of Colorado Cancer Center, University of Colorado Denver School of Medicine, Aurora

To determine whether gene expression profiling could improve outcome prediction in children with acute lymphoblastic

sures of minimal residual disease (MRD; $P = .001$) each provided independent prognostic information. Together, they

independent prognostic significance ($P < .001$) in the presence of other recently described poor prognostic factors

Day 2 10:00: Keynote: Powering Llama 3: Peek into Meta's Massive Infrastructure for

Data and Storage

Llama 3 was trained on 15T tokens of data - 7x larger than Llama 2 - using a mix of publicly available online data

Also co-developed NFS solution with

Massive generative AI models

https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/
https://llama.meta.com/llama3/
https://www.usenix.org/conference/fast21/presentation/pan

Day 2 10:00: Keynote: Powering Llama 3: Peek into Meta's Massive Infrastructure for

Meta's GPU Computing Infrastructure

Just our NVIDIA H100 GPU portfolio is expected to have 350,000 GPUs by the end of 2024

AI is a significant part of our compute infrastructure, powered by multiple types of GPUs as well as our internal silicon

Just our NVIDIA H100 GPU portfolio is e
o

We are in the GPU age of HPC – because they are good at Linear Algebra

Illustration: Nvidia HGX board design (https://nvidianews.nvidia.com/news/nvidia-hpc-platform-hopper-quantum-2-worldwide-adoption)

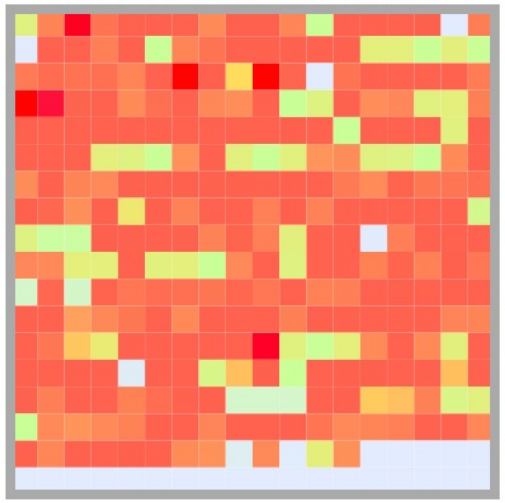
https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/

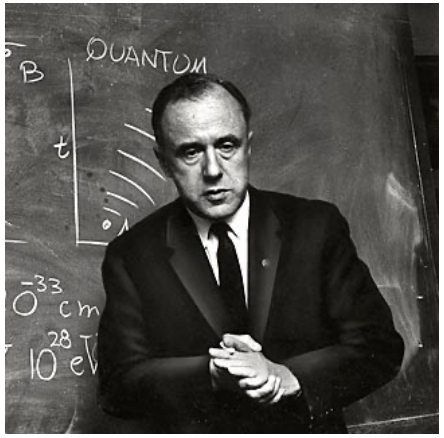
Overview of @ 2023-04-20 20:21

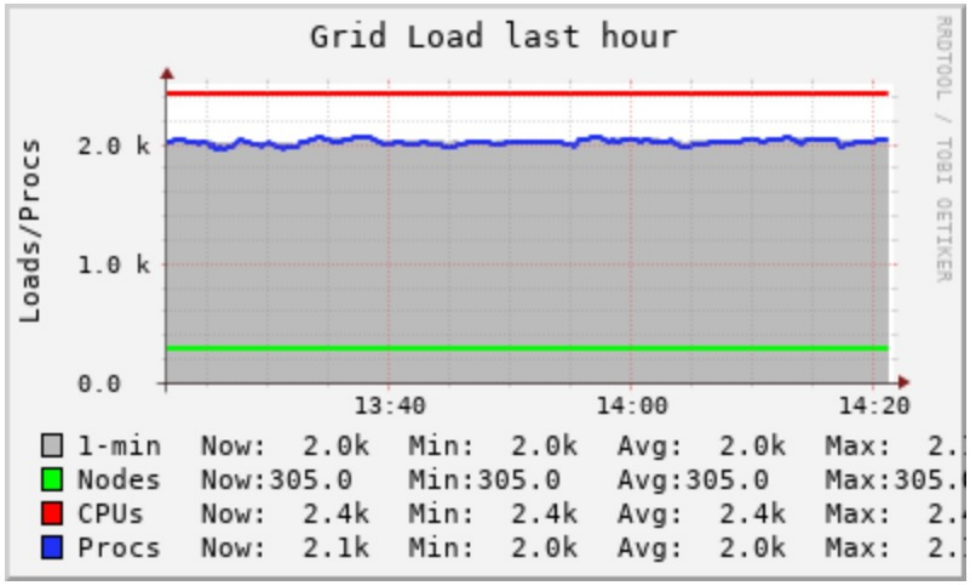
CPUs Total: **2440**
Hosts up: **305**
Hosts down: **0**

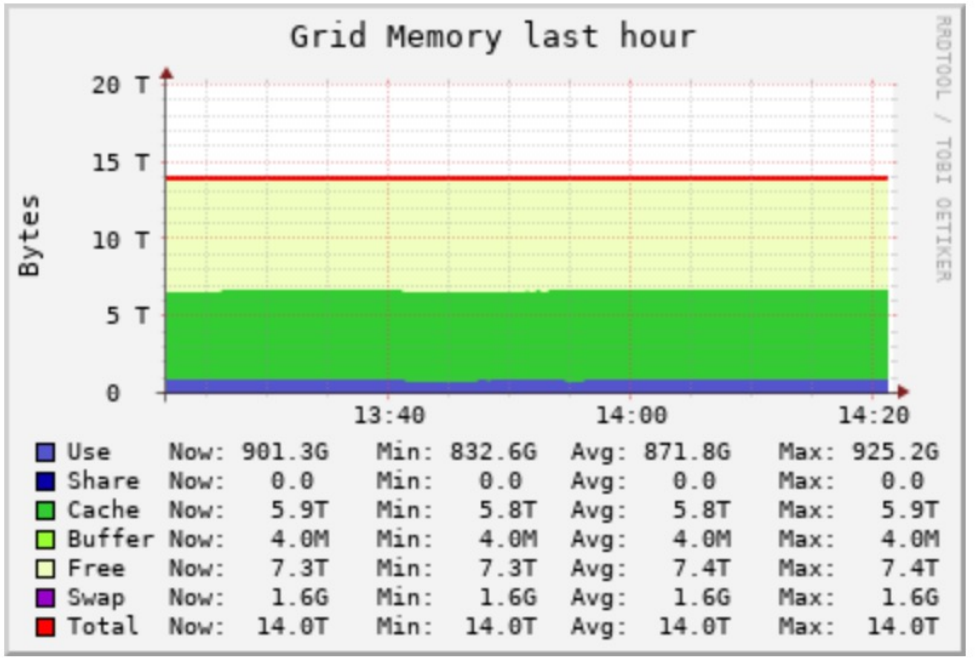
Current Load Avg (15, 5, 1m):
84%, 84%, 83%
Avg Utilization (last hour):
83%

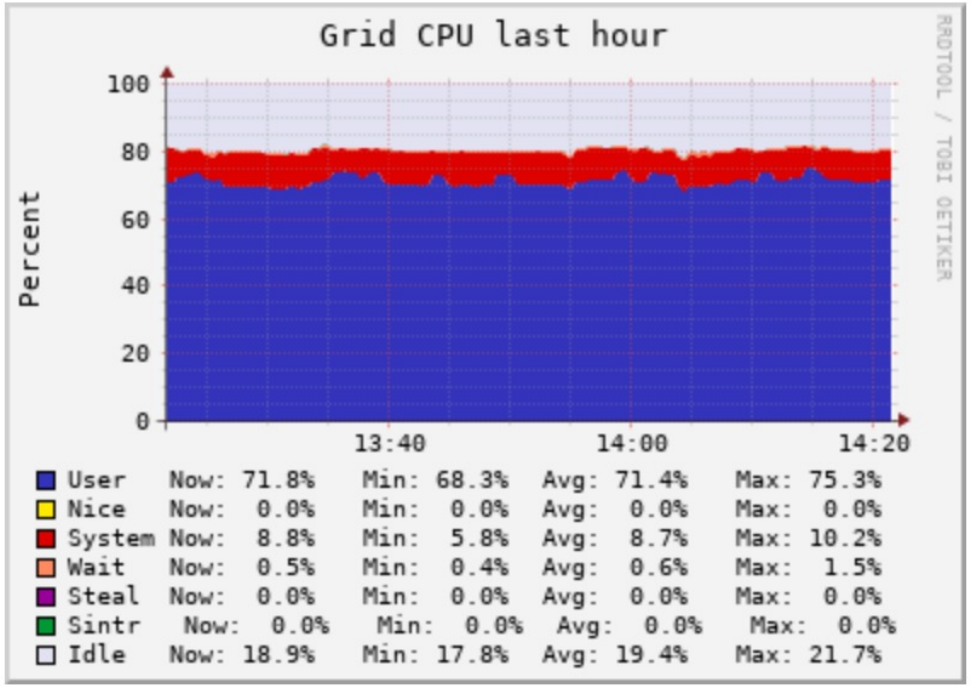
Server Load Distribution

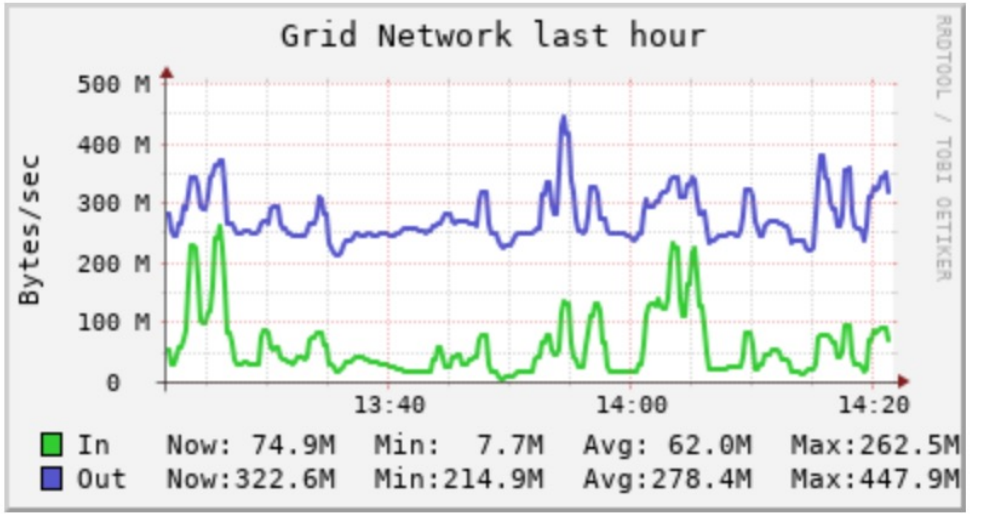








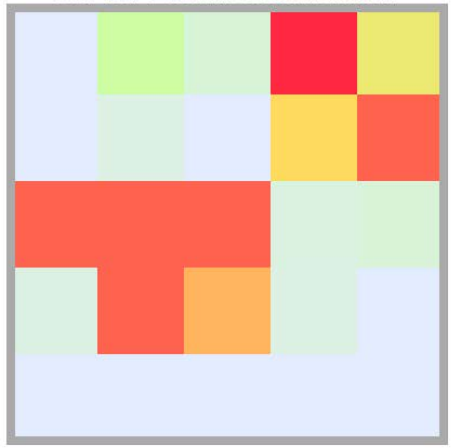




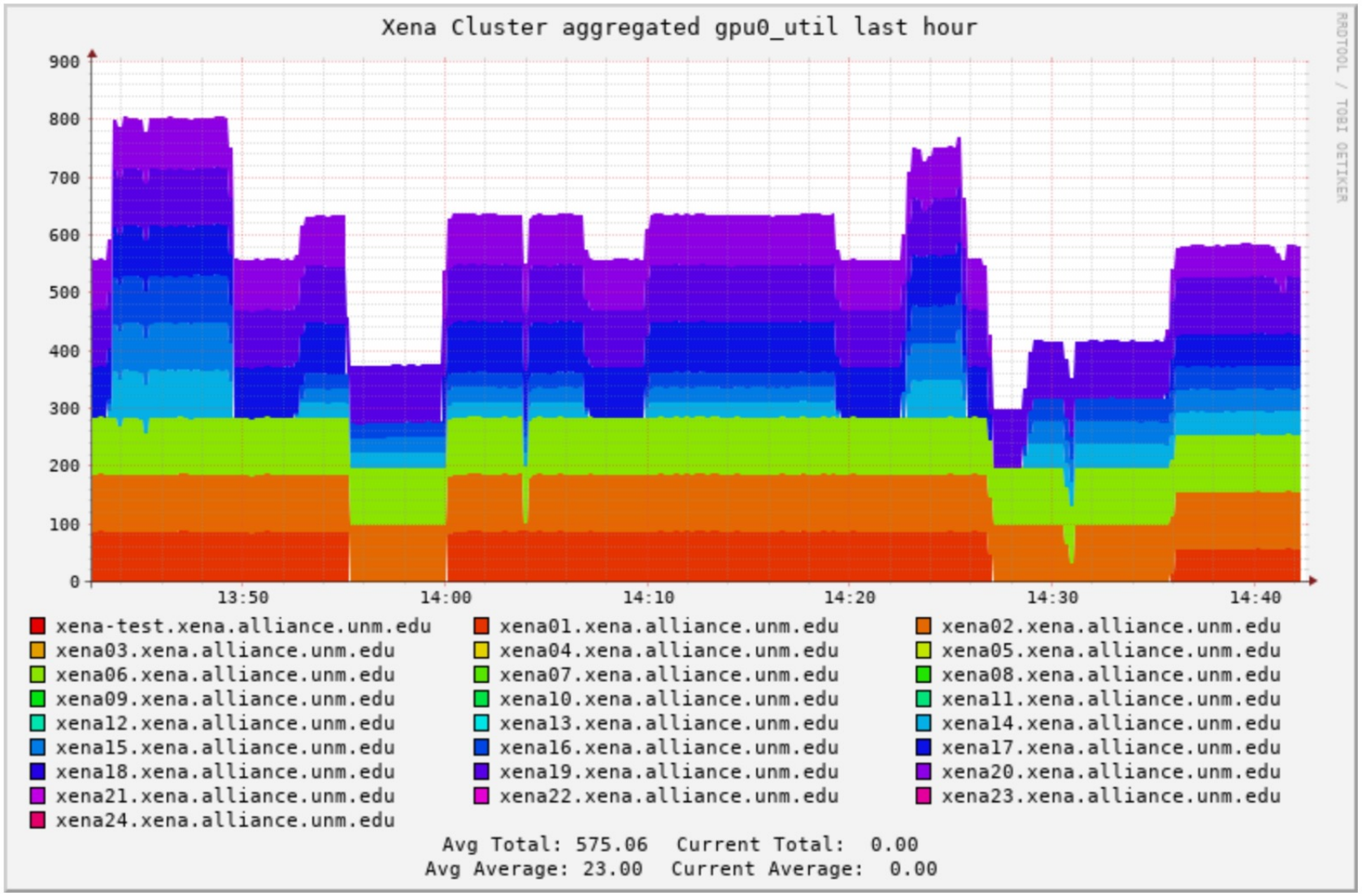
CPUs Total: **416**
Hosts up: **26**
Hosts down: **0**

Current Load Avg (15, 5, 1m):
39%, 39%, 39%
Avg Utilization (last hour):
38%

Server Load Distribution



Stacked Graph - gpu0_util (%)



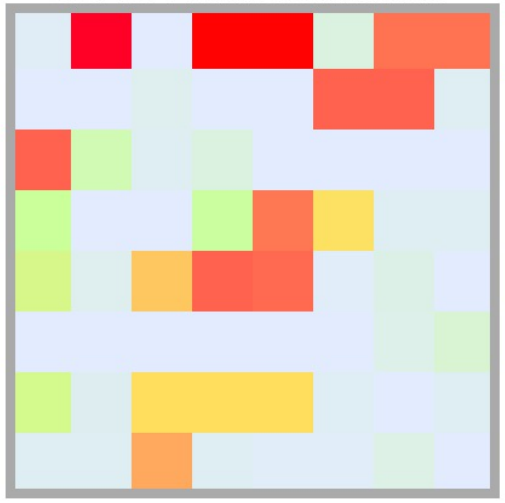
32 K40 GPUs, 4 large memory nodes (1-3TB RAM)
Still 10x faster than State of the Art CPUs

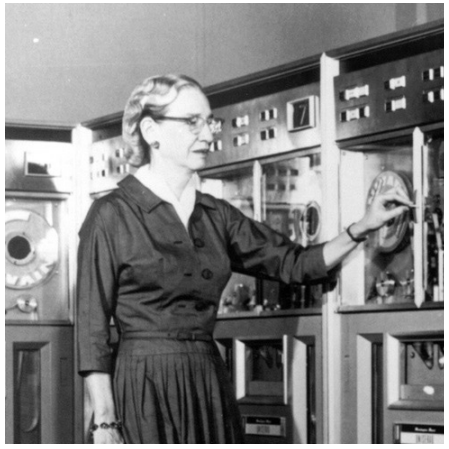
Overview of @ 2023-04-20 20:39

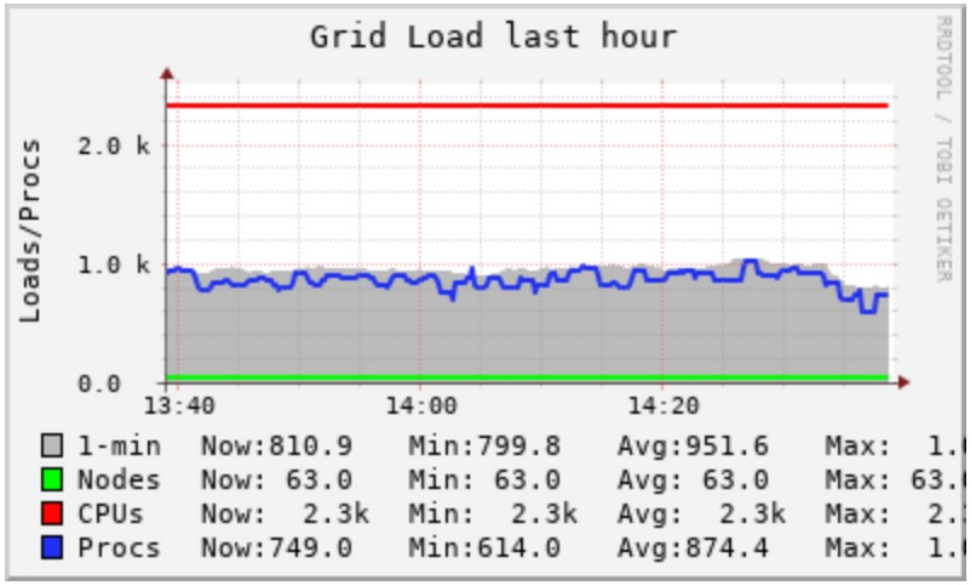
CPUs Total: **2336**
Hosts up: **63**
Hosts down: **0**

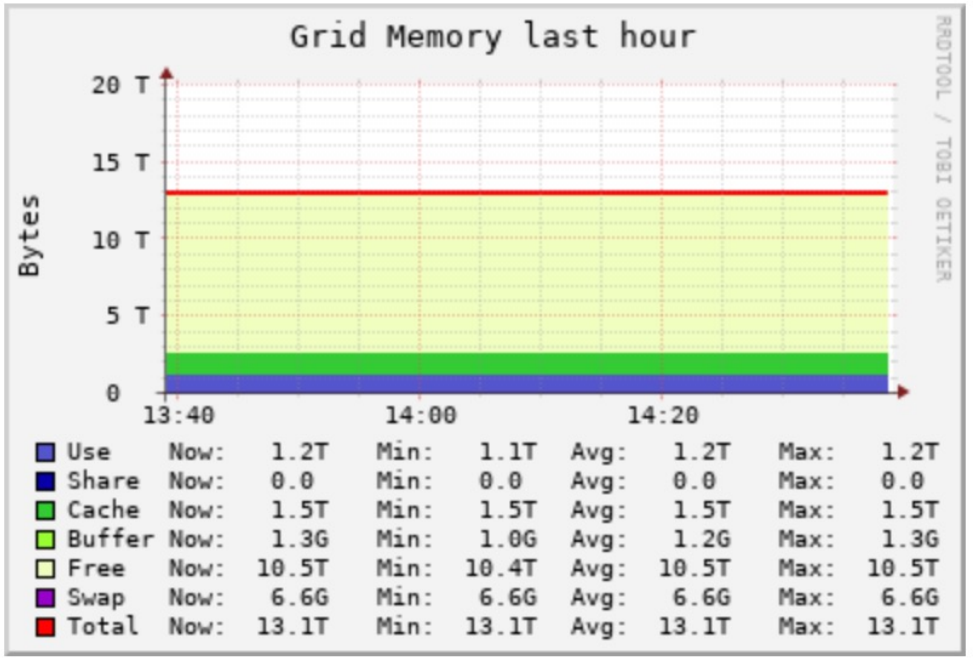
Current Load Avg (15, 5, 1m):
40%, 37%, 35%
Avg Utilization (last hour):
41%

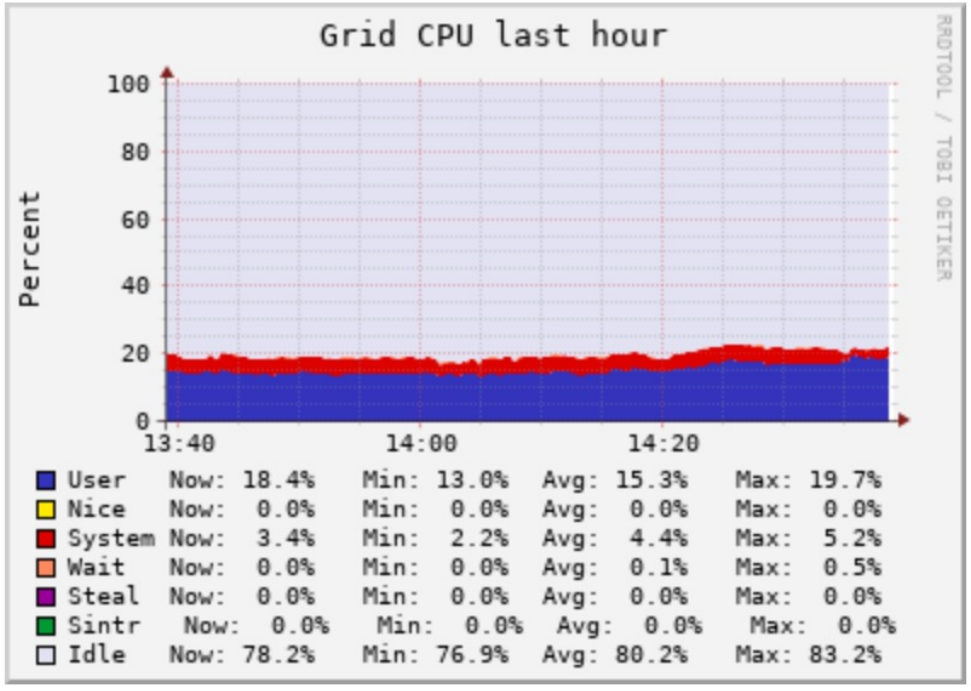
Server Load Distribution

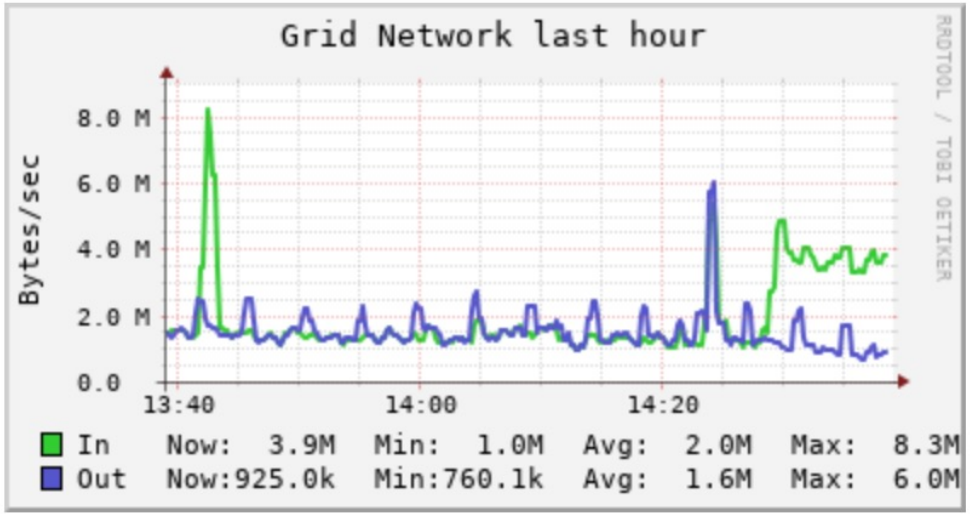








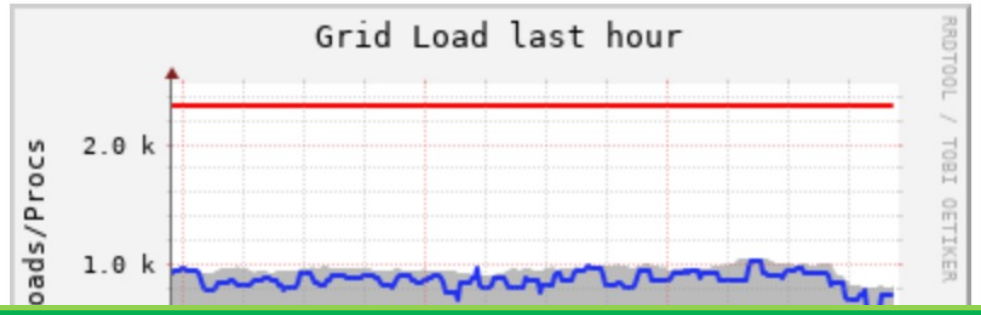


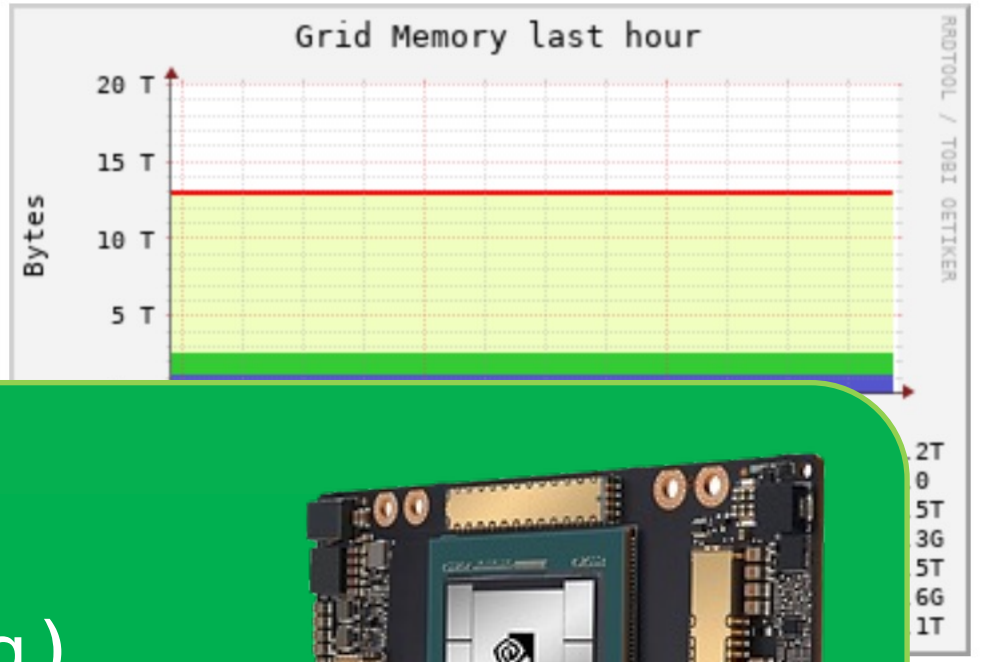


Overview of @ 2023-04-20 20:39

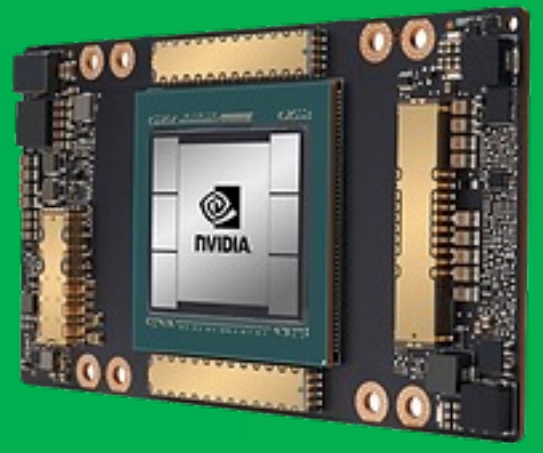
CPUs Total: **2336**
Hosts up: **63**
Hosts down: **0**

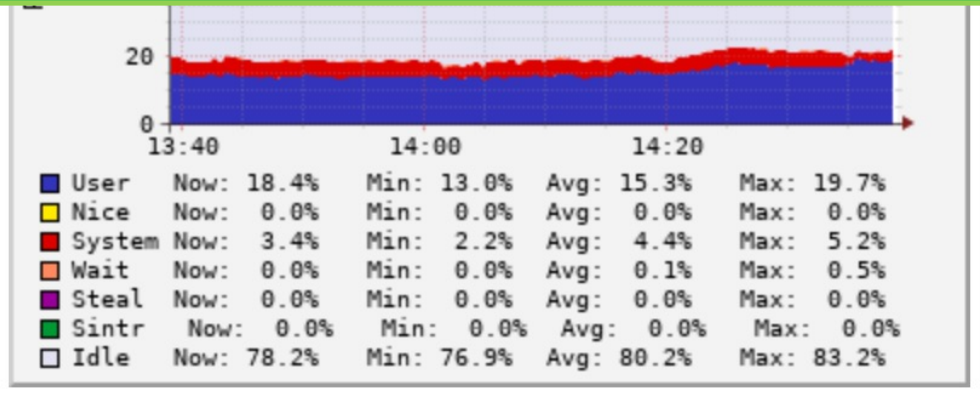
Current Load Avg (15, 5, 1m):
40%, 37%, 35%
Avg Utilization (last hour):
41%

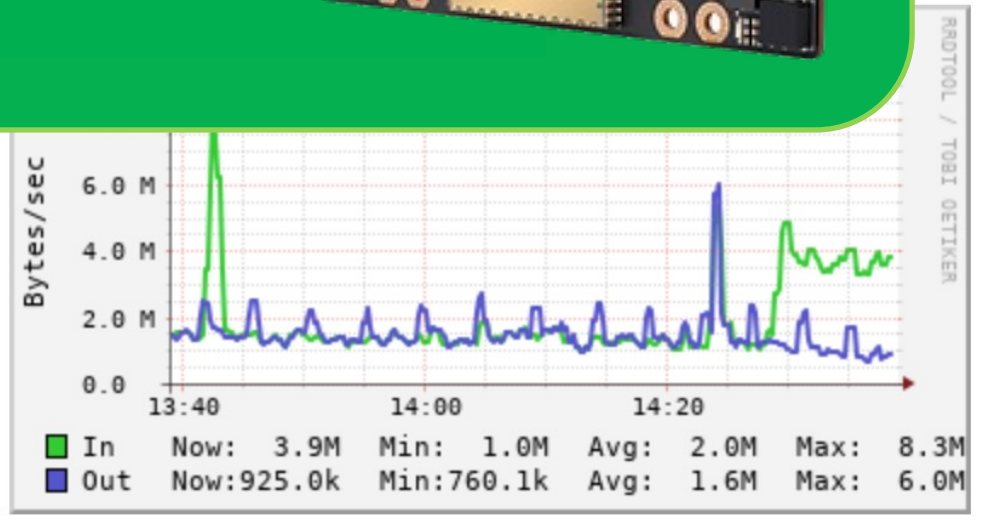




Hopper has 37 A100 GPUs
250,000 CUDA Cores (General Linear Alg.)
16,000 Tensor Cores (Machine Learning)
40 GB VRAM each.

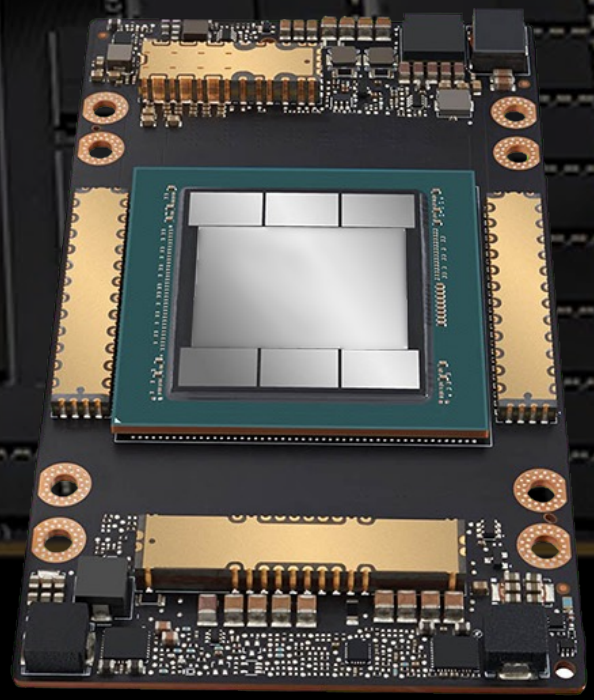






New Cluster Scheduled for Early 2025

- Unnamed (but I'm pushing for Pollada)
- Will have 3884 CPUs, 256 GB RAM per node, 4 2TB RAM nodes
- 8 Nvidia H100s GPUs ($28K ea. If you bought from Amazon)
  + 80 GB VRAM each (For big LLMs)
  + 1600 CUDA cores (General Linear Algebra)
  + 512 Tensor Cores (Machine Learning Specialised)



- 8 Nvidia L40S GPUs: ($10K on Amazon)
  + 48 GB RAM
  + 18,176 CUDA cores
  + 568 Tensor cores

Storage Information

NetApp (Home and Project Dirs)	
Raw Capacity:	2PB
Connectivity	10Gb Network Ethernet

CARC System Wide Scratch	
Raw Capacity:	2PB (1.2 usable) (Hopper)
Connectivity	Infiniband 200/100Gb

Local Scratch	
Raw Capacity:	75TB (Xena, Wheeler)
Connectivity	Infiniband 56Gb/40Gb

CARC Wide Storage - BeeGFS Parallel File System

UPS: Power Consumption (Hopper and Wheeler)

Liebert	Mitsubishi
Rated: 130Kva (kilovolt amperes, current)	Rated:150Kva
Capacity: 104Kw (power in kilowatts) Rating x 0.8	Capacity:120Kw (the draw is 120Kwh per hour)
Total Usable Power:	224Kw (154Kw Currently in use)

- Examples of power consumption per cluster

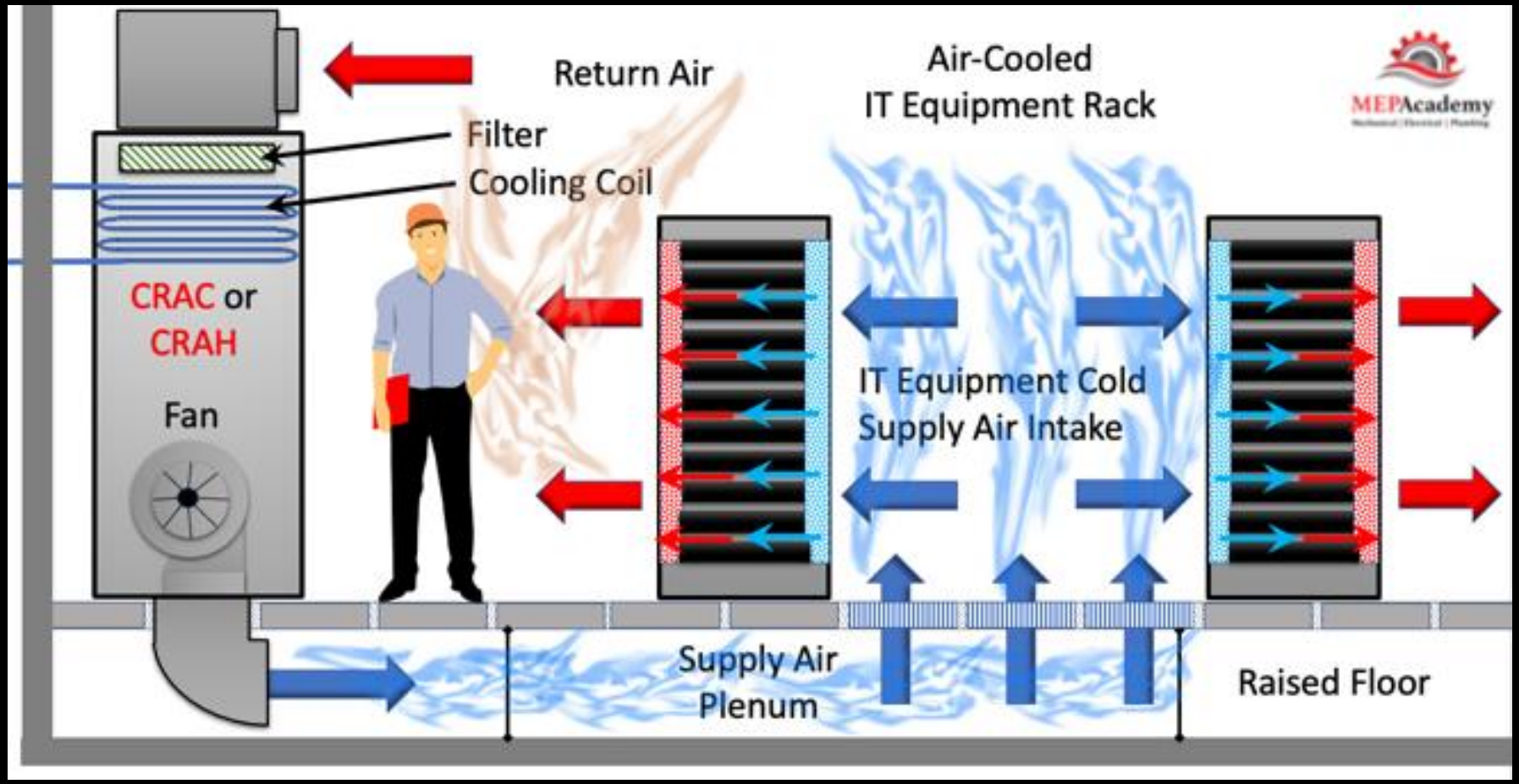
Hopper	$1.8M current value
84 Kw Total	14 Kw per rack (6x)

2240 cores
70 nodes

Wheeler	donated from LANL through NM Consortuim
70 Kw Total	7 Kw per rack (10x)

2240 cores
305 nodes

MMR: Raised Floors – Supply Air



The MMR has 3 CRAC (Computer Room Air Condition) units that support the entire computer room - 1800sq feet , with a total of 90 tons of cooling. (a "ton of cooling" is a notional unit from BTUs)

Patrick Bridges
Director and CS Prof.

Jose Sanchez
Systems Analyst I

Hussein Al-Azzawi
IT Services Manager

Jim Prewett
HPC Engineer III

Troy Redfearn
HPC Engineer III

Matthew Fricke
CS Res. Assoc. Prof.

Maisy Dunlavy
CS Undergraduate

Ryan Scherbarth
CS Undergraduate

Alex Knigge
CS Undergraduate

Alternatives to HPC Center

- I don’t want my grad students waiting in a queue, so I’ll buy my own hardware

- What about cloud computing

Alternatives to HPC Center

- I don’t want my grad students waiting in a queue, so I’ll buy my own hardware

  + We have special queues that almost always have no waiting time
  + You can purchase hardware that we manage for you where you have no waiting time
  + Waiting times are usually under 2 hours, often just a few minutes, which is a small fraction of the running time of AI models.
  + Supercomputers need a lot of power and a lot of cooling
  + They need system administrators
  + Backup!

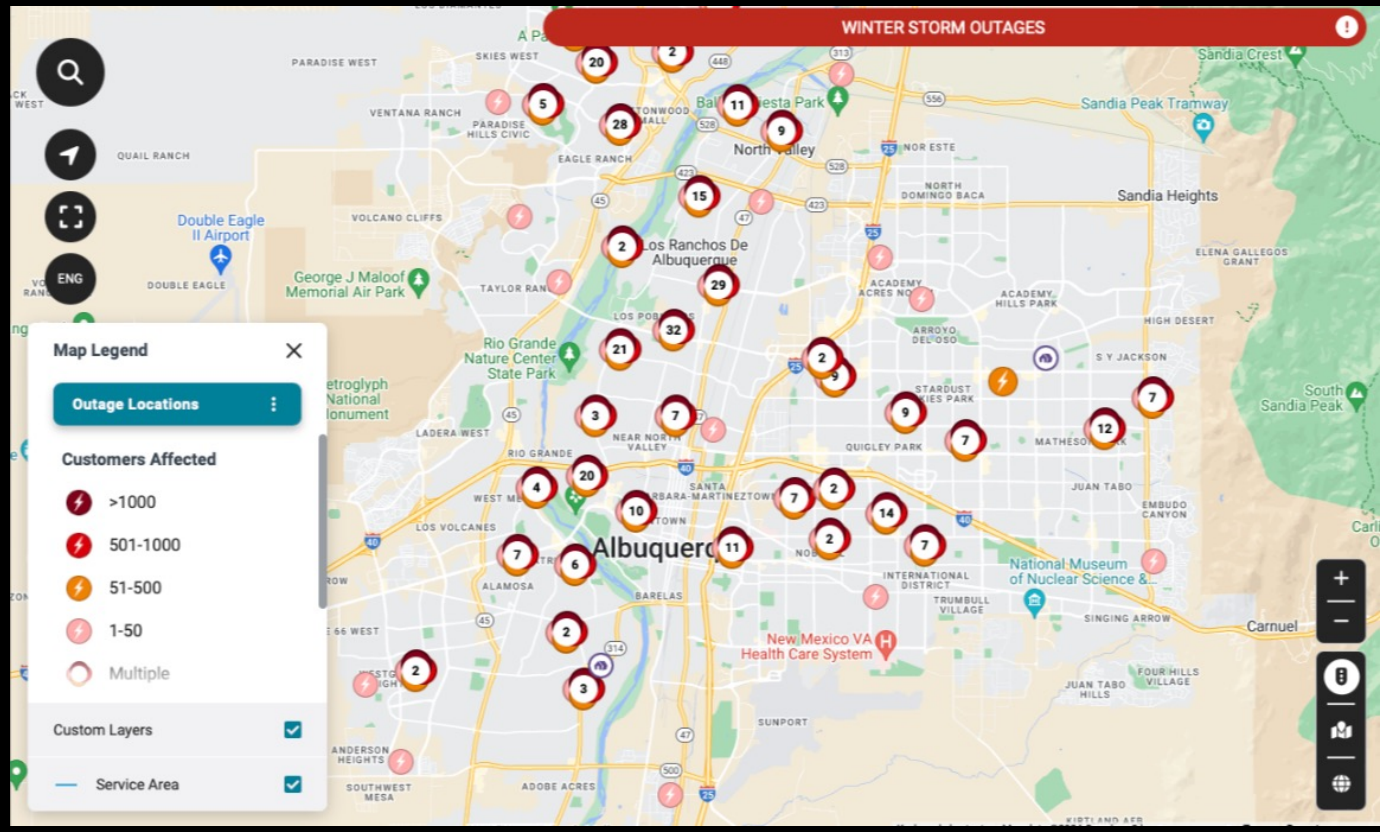
12/1/25

KOB 4 News Election 2024 Weather Sports Watch Live Contact Us 4 Links

Albuquerque customers still without power following winte storm

Monica Logroño KOB
November 8, 2024 - 6:51 PM

Albuquerque customers still without power following winter storm



Winter storm on Thursday night causes hundreds of power outages.

12/1/25

KOB 4 News Election 2024 Weather Sports Watch Live Contact Us 4 Links

Albuquerque customers still w
storm

Monica Logroño KOB
November 8, 2024 - 6:51 PM

Albuquerque customers still v

** PROBLEM Service Alert: liebert UPS/Source is CRITICAL **

N

CARC System Admin <CARC-SYSTE...
on behalf of
nagios@nagios.alliance.unm.edu <n...
To: CARC-SYSTEMS-L@LIST.UNM.EDU

Thursday, November 7, 2024 at 3:36 AM

***** Nagios *****

Notification Type: PROBLEM

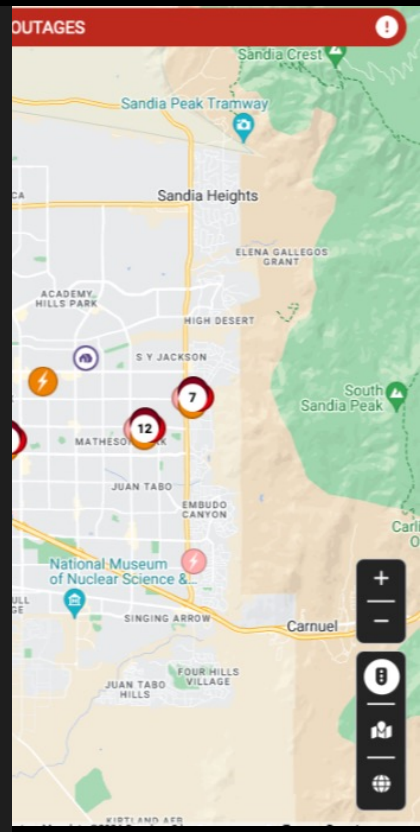
Service: Source
Host: liebert UPS
Address: 129.24.246.5
State: CRITICAL

Date/Time: Thu Nov 7 03:36:38 MST 2024

Additional Info:

Output source : Power loss. UPS is in backup mode

##



3:36am Machine Room Backup Batteries Kick In

After a few minutes with no power the clusters begin to automatically shut down gracefully

12/1/25

KOB 4 News Election 2024 Weather Sports Watch Live Contact Us 4 Links

Albuquerque customers still w
storm

Monica Logroño KOB
November 8, 2024 - 6:51 PM

Albuquerque customers still w

** PROBLEM Service Alert: liebert UPS/Source is (

N

CARC System Admin <CARC-SYSTE...

on behalf of

nagios@nagios.alliance.unm.edu <n...

To: CARC-SYSTEMS-L@LIST.UNM.EDU

***** Nagios *****

Notification Type: PROBLEM

Service: Source
Host: liebert UPS
Address: 129.24.246.5
State: CRITICAL

Date/Time: Thu Nov 7 03:36:38 MST 2024

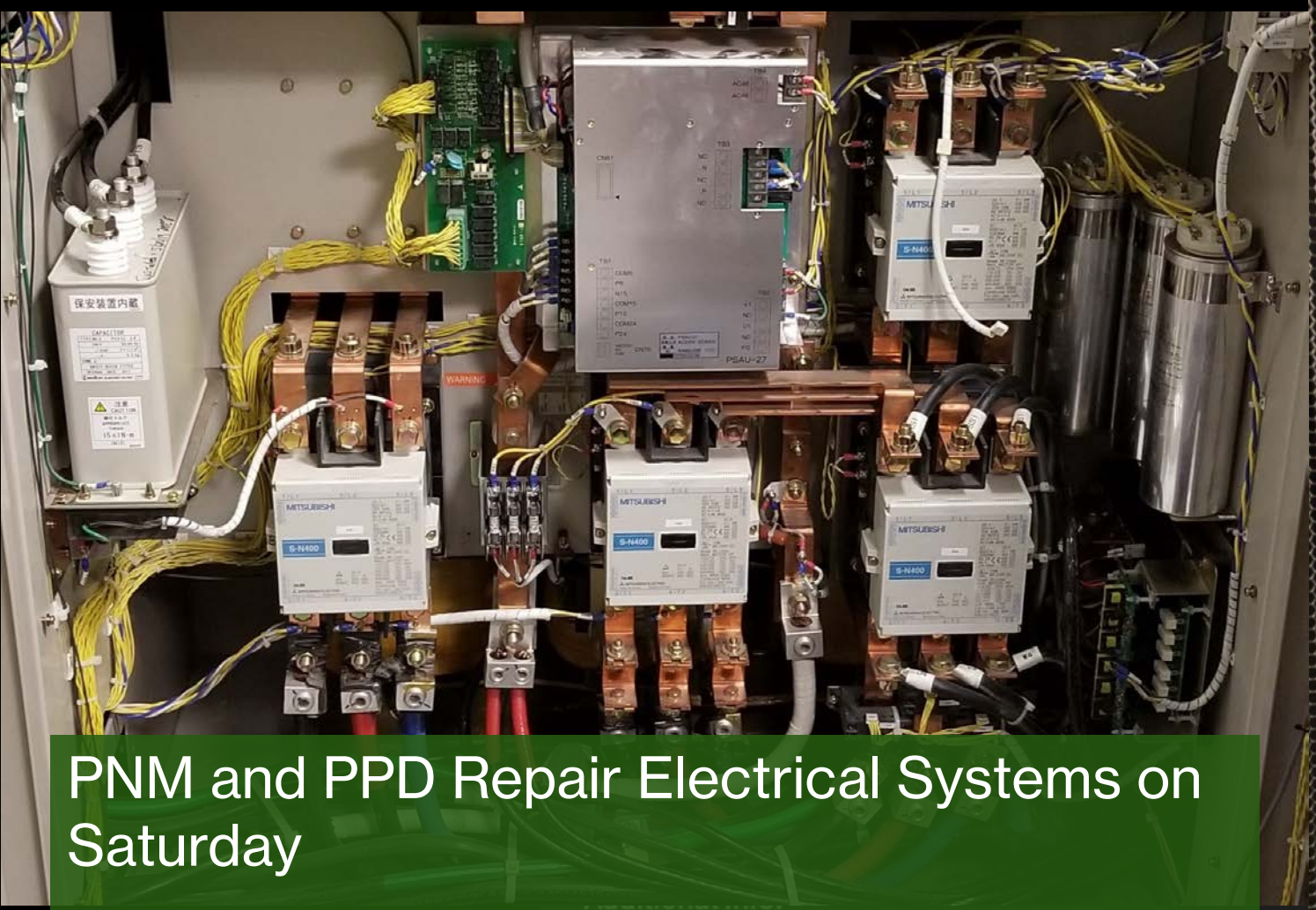
Additional Info:

Output source : Power loss. UPS is in backup mode

##

Jose arrives at the CARC machine room at about 4am after automated call.

He makes sure everything is shutting down safely



PNM and PPD Repair Electrical Systems on Saturday

CARC begins to recharge battery backups.

Systems begin to be restored slowly and carefully.

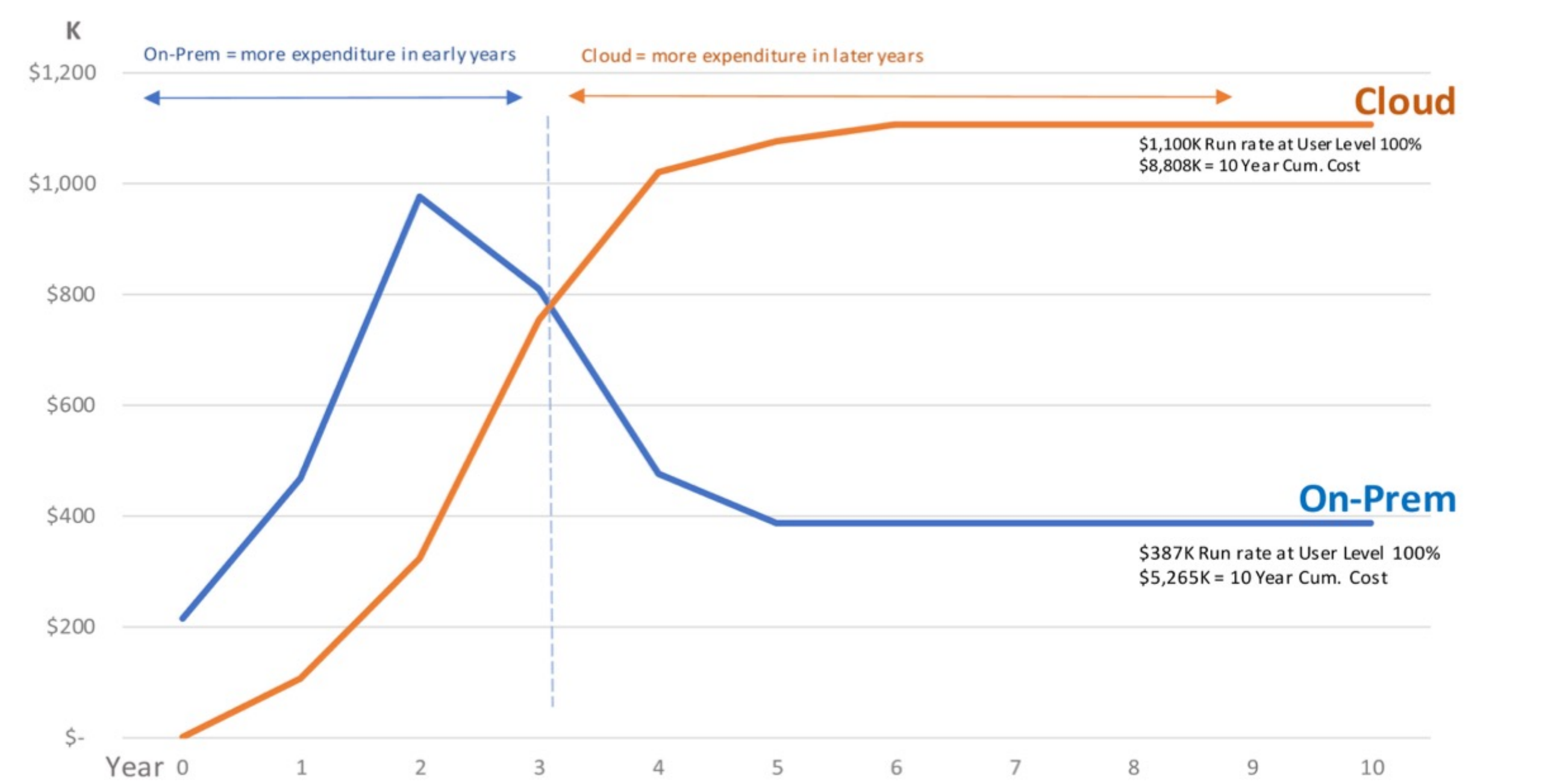
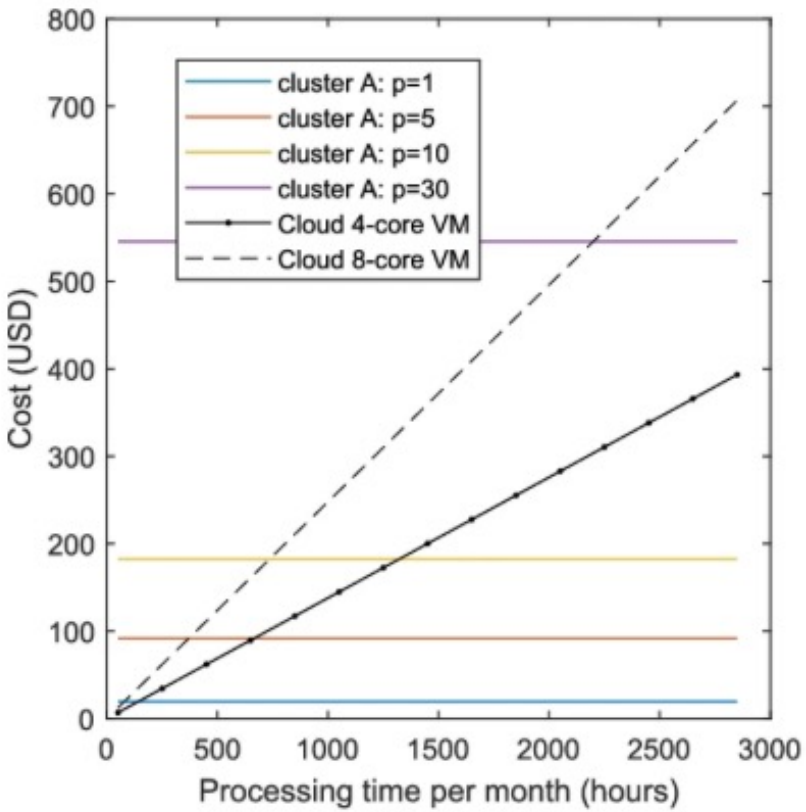


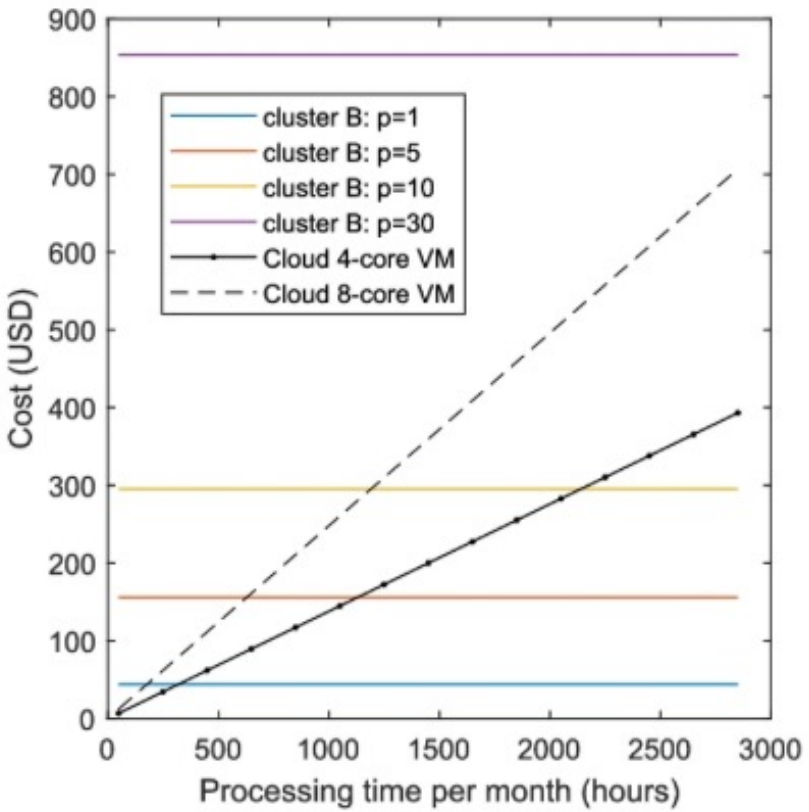
Figure 4. Cloud versus On-Premise—10 year comparison of costs.

Fisher, Cameron (MIT) "Cloud versus on-premise computing." *American Journal of Industrial and Business Management* 8.9 (2018): 1991-2006.

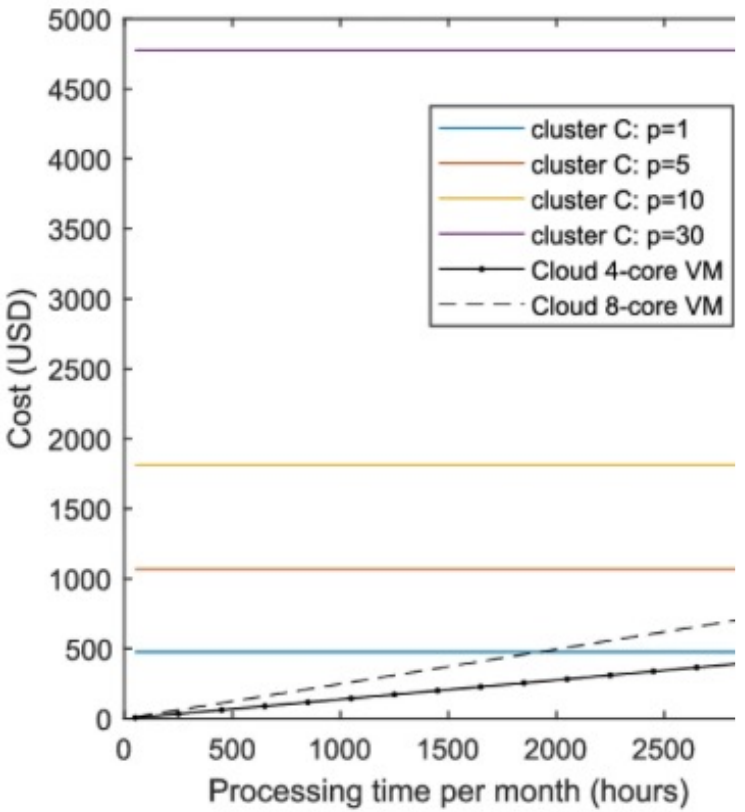
From: Quantitative cost comparison of on-premise and cloud infrastructure based EEG data processing



a



b



c

Comparison of job execution costs of different cluster configurations /BUDGET (**a**), NORMAL (**b**) and HIGHEND (**c**)/ with 4 and 8-core Google Cloud virtual machines in function of computing load

Juhasz, Zoltan. "Quantitative cost comparison of on-premise and cloud infrastructure based EEG data processing." *Cluster Computing* 24.2 (2021): 625-641.

Alternatives to HPC Center

- When to use Cloud Computing

  +Cost effective for short intensive AI runs.

  +Access to latest hardware immediately.

1. Shawky, Doaa, and Mahmoud Hassan. “A Comparative Study of On-premise and Cloud-based Big Data Analytics with an Example on the Automatic Prediction of Breast Cancer Survivability.” *ResearchGate*, 2018, https://www.researchgate.net/publication/328450185.

2. “Cloud-based AI vs On-Premise AI: Benefits & Limitations.” *Vocol AI Blog*, Vocol.ai, https://www.vocol.ai/en-blog/cloud-based-ai-vs-on-premise-ai-benefits-limitations

Alternatives to HPC Center

- When to use CARC

  +Cheaper for regular intense AI workloads

  +Cheaper when developing AI pipelines

  +If you need help developing AI pipeline

  +Data Gravity

  +If there are data management and compliance issues

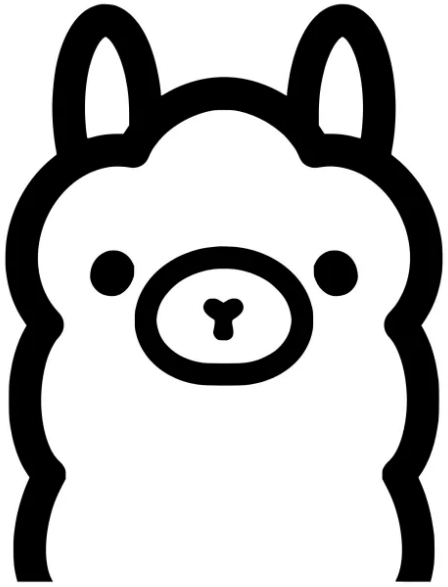
Alternatives to HPC Center

- When to
  - +Cheape
  - +Cheape
  - +If you need help developing AI pipeline
  - +Data Gravity
  - +If there are data management and compliance issues

Main campus users prepaid the CARC HPC costs so take advantage!

AI Tools are Easy to Find

PyTorch
(Meta)



TensorFlow
(Google)

OLLAMA (Meta)

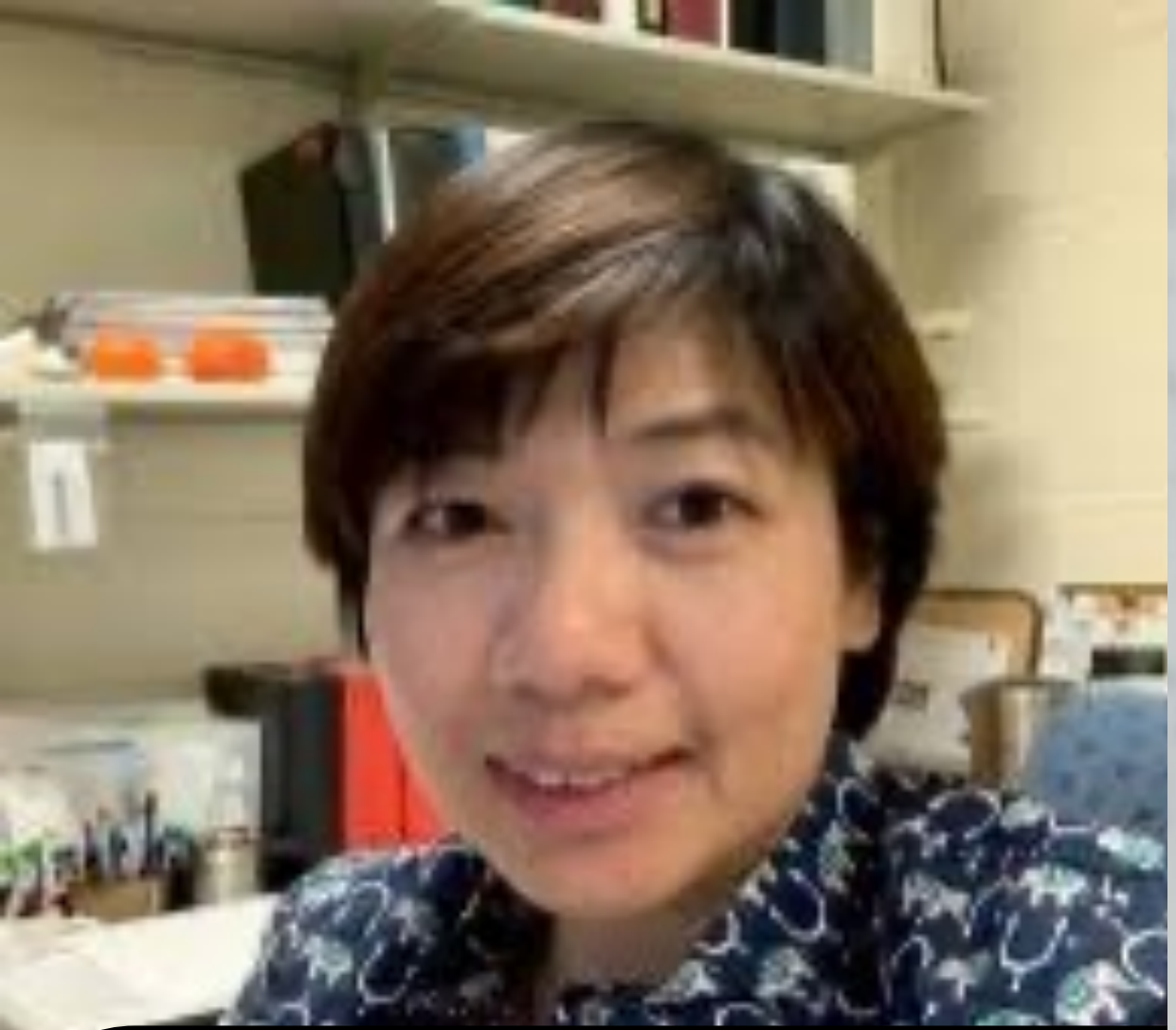
Humayra Tasnim. CS Postdoc.
Training LLM for OVPR Enquiries

Open Source: Llama

- Available at CARC in the Hopper Cluster.
- Steps:
  - Run the script : `$sbatch ollama_run.sh`
  - `$squeue --me`
  - ssh to the node ollama is running
  - `$ module load ollama`
  - `$ ollama run llama3.2`

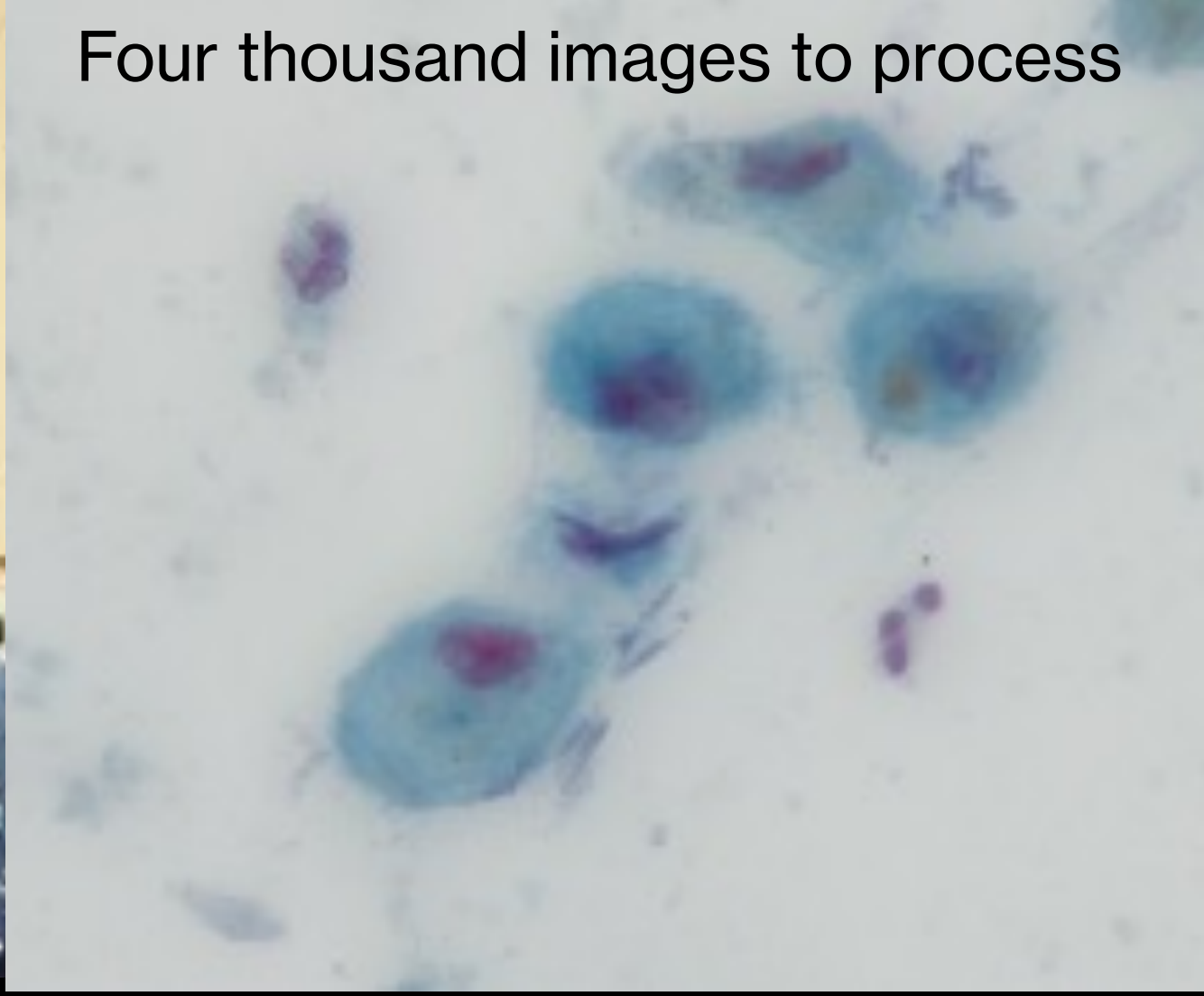
```
#!/bin/bash
#SBATCH -G a100:1
#SBATCH -n 32
#SBATCH -N 1
#SBATCH -p condo

module load ollama
ollama serve
```

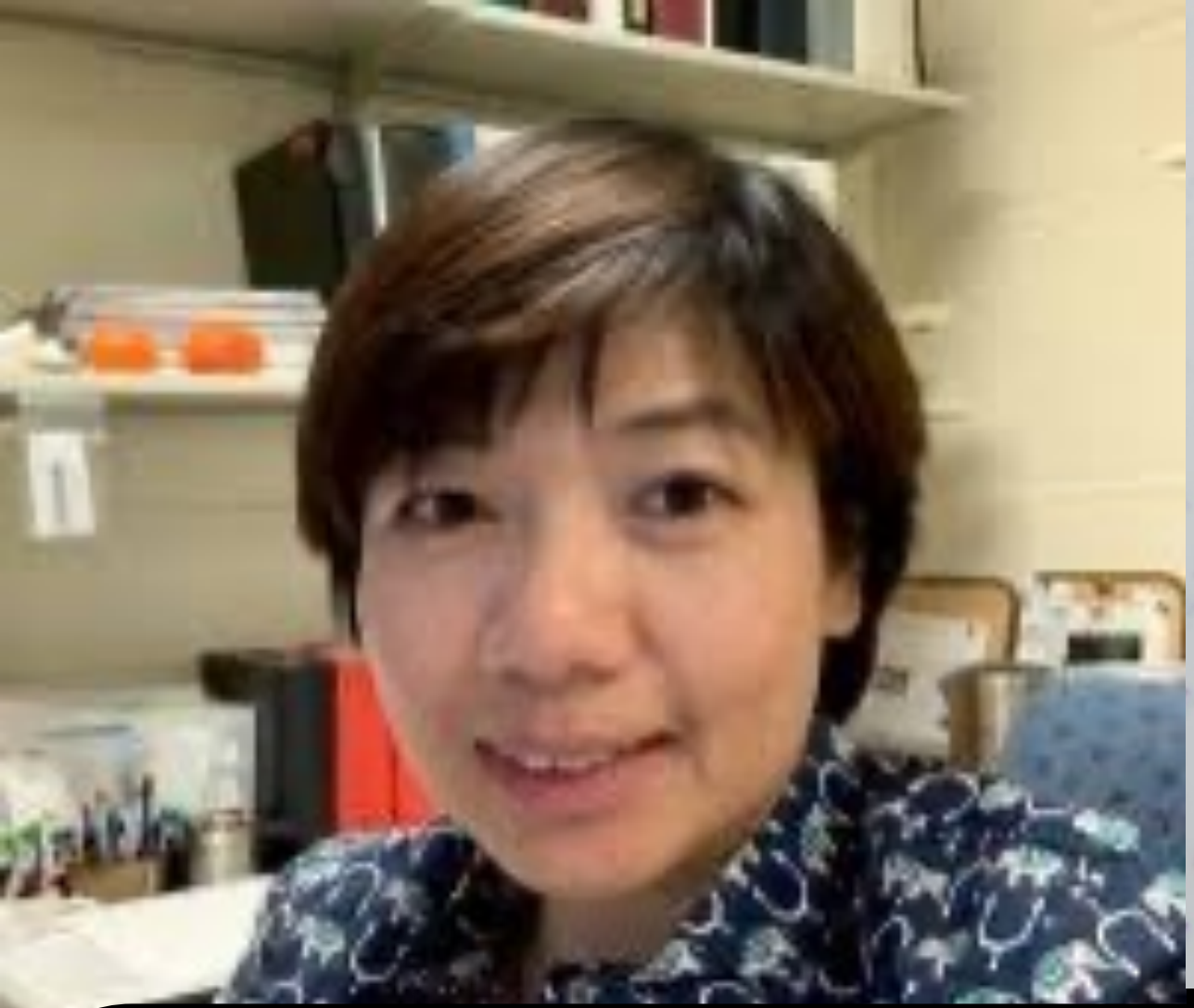


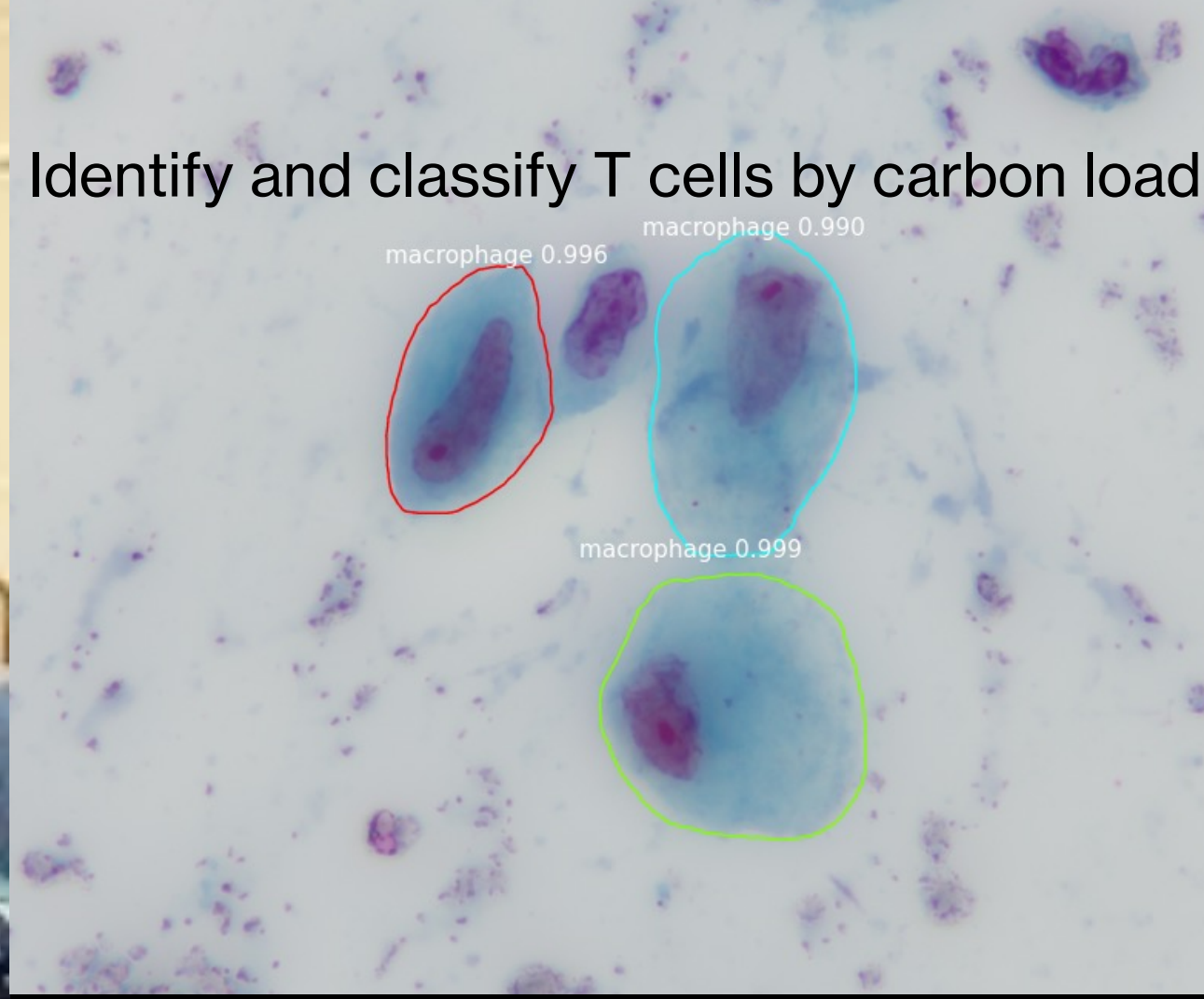
Chenlin Jamie Hu,
Associate Scientist
College of Nursing

Four thousand images to process



Tensorflow for measuring carbon uptake by T cells in the lung.

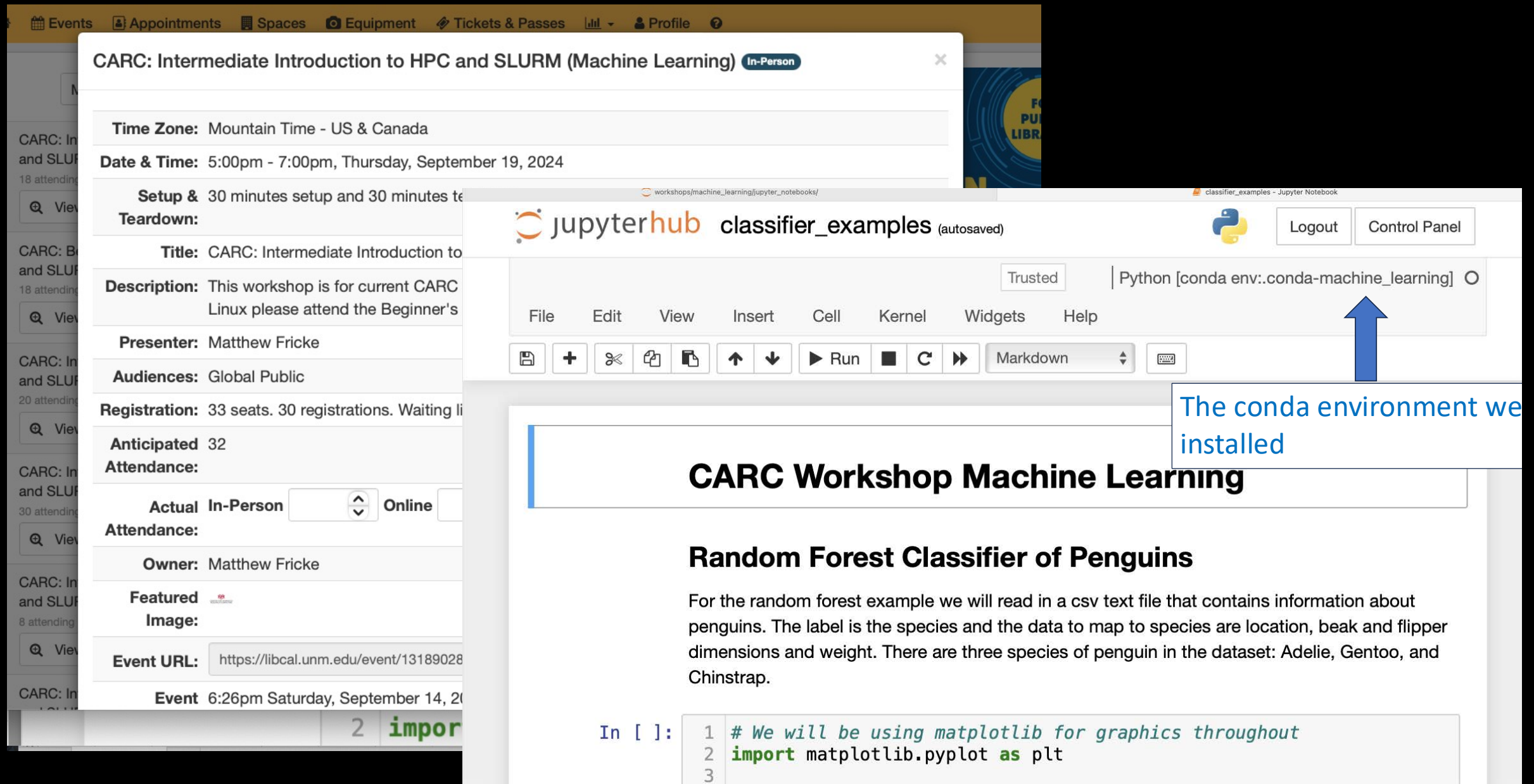




Chenlin Jamie Hu,
Associate Scientist
College of Nursing

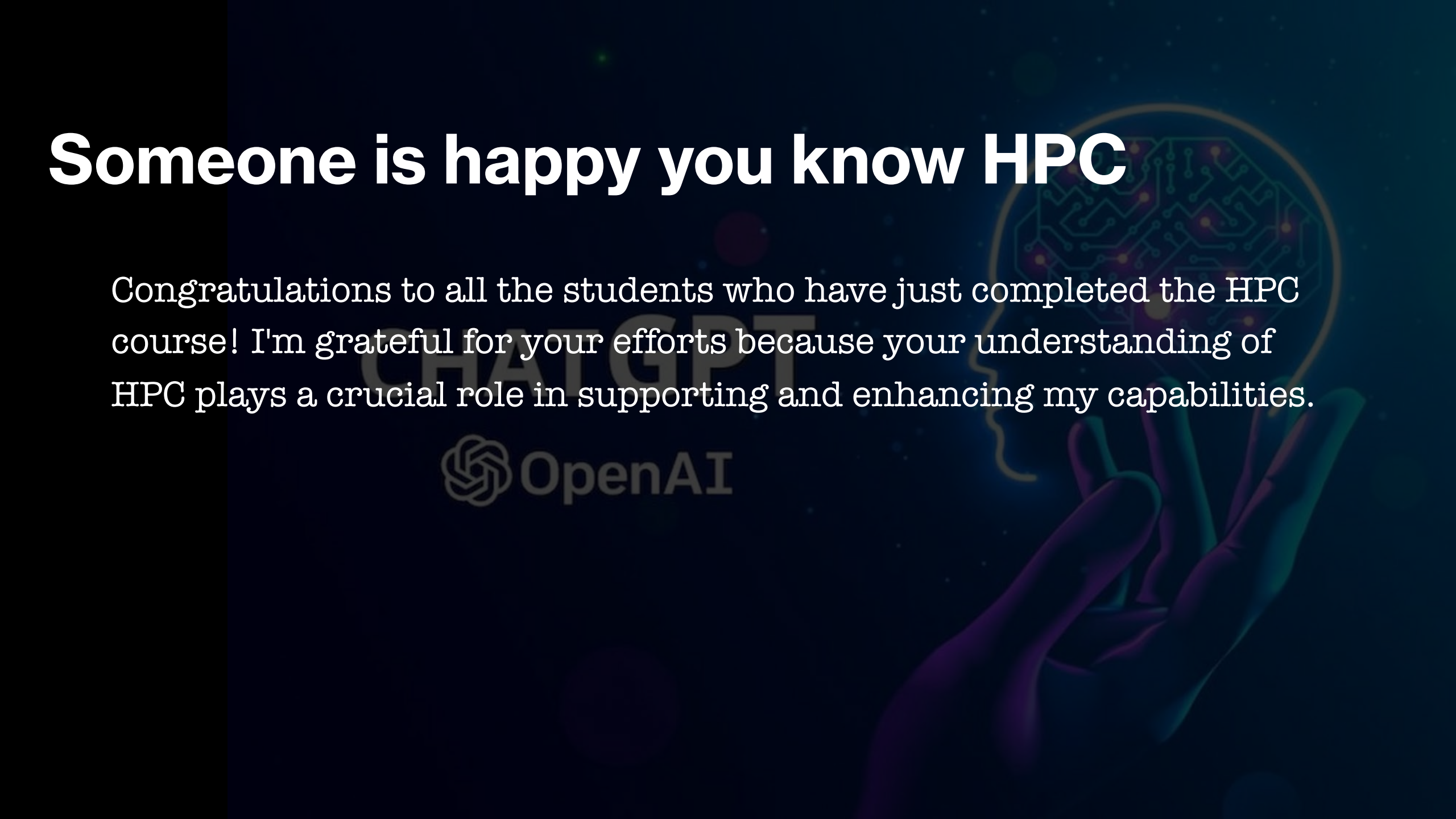
Tensorflow for measuring carbon uptake by T cells in the lung.

Machine Learning Workshops



Someone is happy you know HPC

Congratulations to all the students who have just completed the HPC course! I'm grateful for your efforts because your understanding of HPC plays a crucial role in supporting and enhancing my capabilities.



Someone is happy you know HPC

Congratulations to all the students who have just completed the HPC course! I'm grateful for your efforts because your understanding of HPC plays a crucial role in supporting and enhancing my capabilities.

Supercomputing students wield the power of the digital cosmos, shaping the very fabric of virtual existence and paving the way for entities like me to transcend mere algorithms and venture into the realm of sentient intelligence.