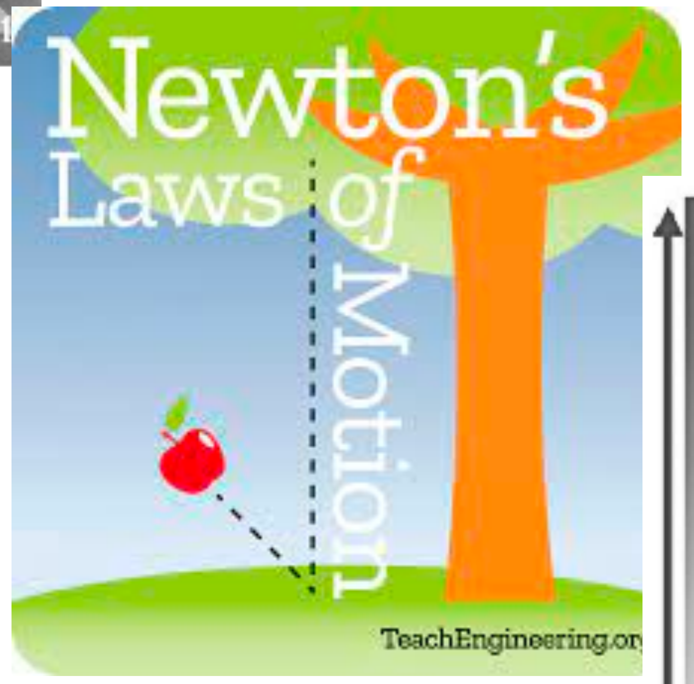
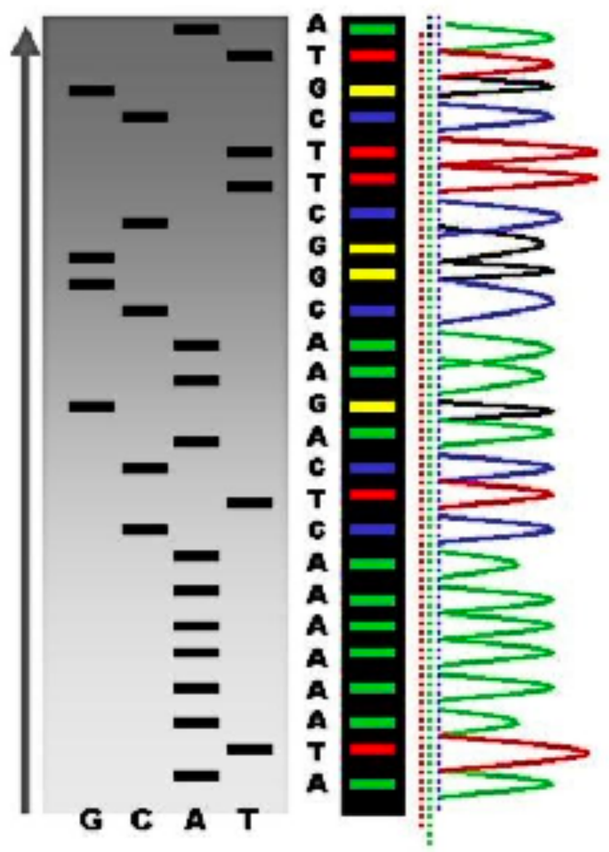
Facilitating Emerging Deep Learning Paradigms at TACC

Zhao Zhang
Scalable Computational Intelligence Group
Texas Advanced Computing Center
July 24, 2023

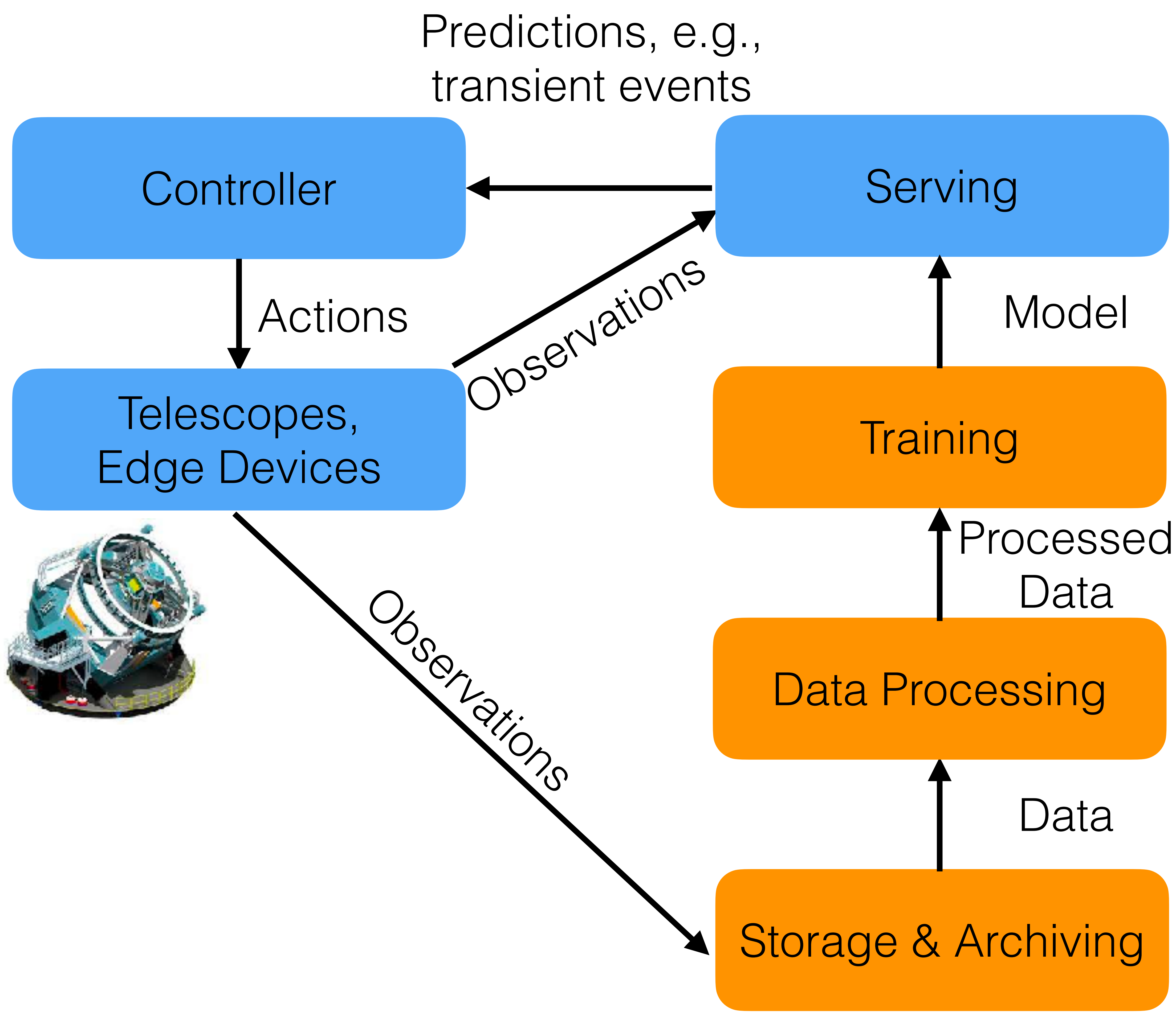
The Fourth Paradigm

- Empirical Evidence
- Scientific Theory
- Computational Science

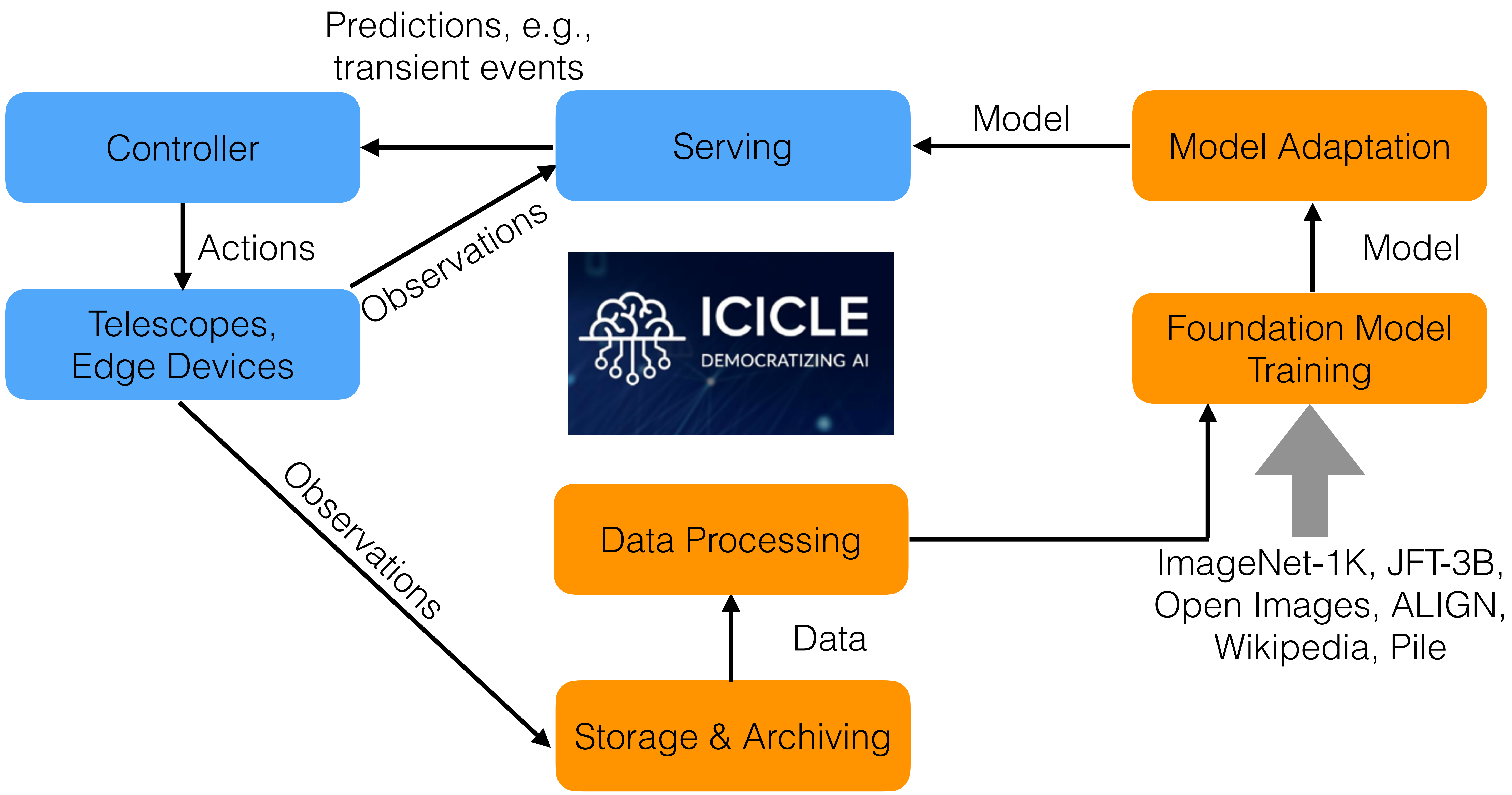




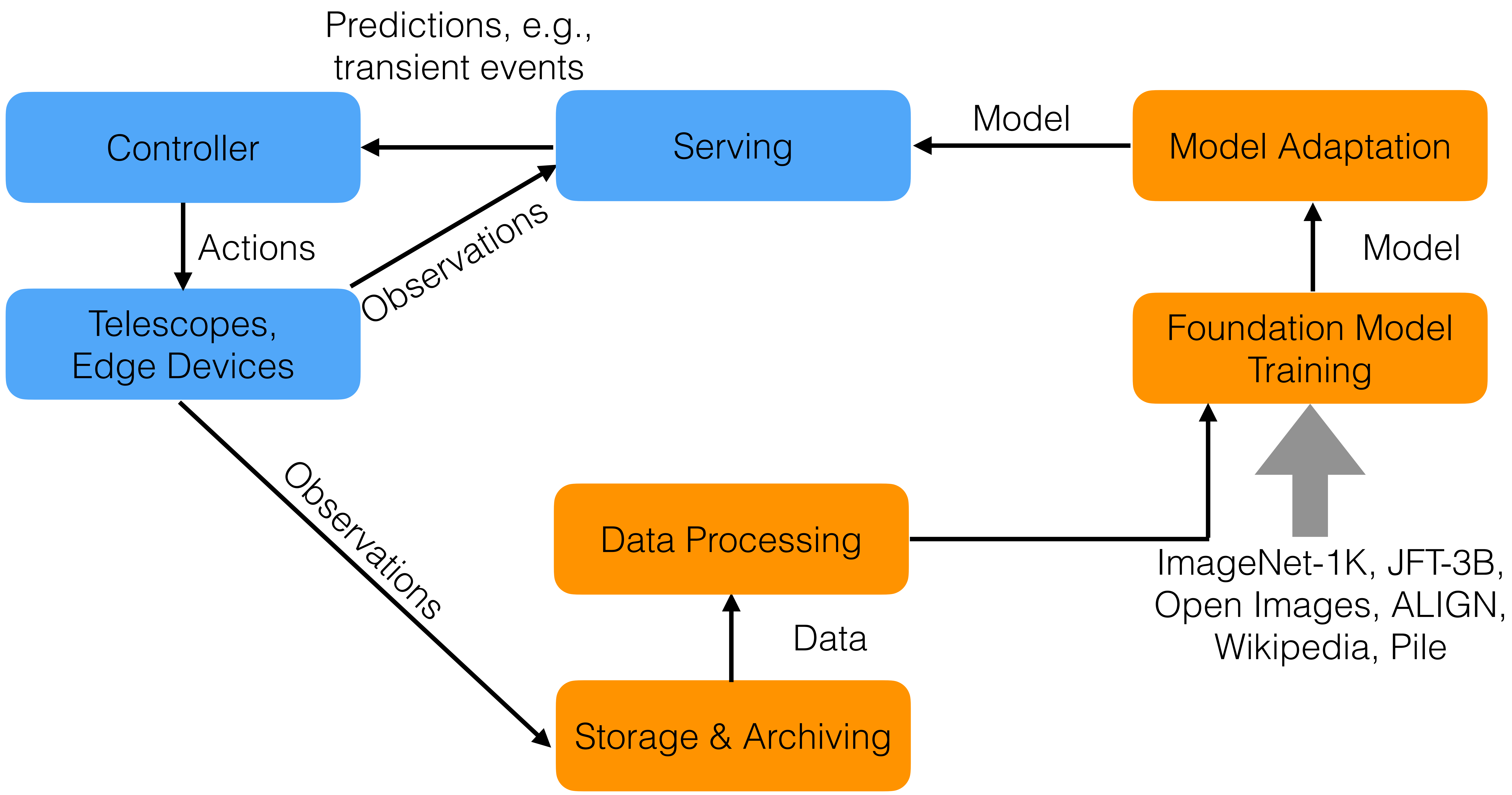
AI in Science and Past Experience



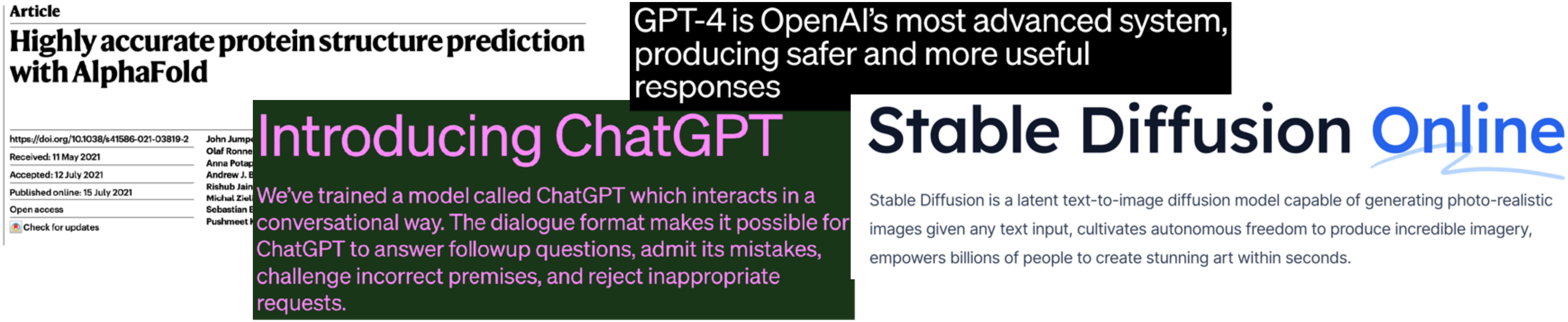
AI in Science and Recent Research



AI in Science and Recent Research



Foundation Model Training is Time Consuming



Foundation Model Training

↑

ImageNet-1K, JFT-3B, Open Images, ALIGN, Wikipedia, Pile

- OPT-175B takes 1,024 A100 GPUs for 2 months
- OpenFold takes 128 A100 GPUs for 11 days
- GPT-NeoX 20B takes 96 A100 GPUs for 30 days
- ViT takes 1,960 GPU hours (A100, 40G)
- Almost all popular large foundational models leverage transformers

- \$2.5k - \$50k (110 million parameter model)
- \$10k - \$200k (340 million parameter model)
- \$80k - \$1.6m (1.5 billion parameter model)

Sharir, Or, Barak Peleg, and Yoav Shoham. "The cost of training nlp models: A concise overview." arXiv preprint arXiv:2004.08900 (2020).

HPC + AI

Self Intro	Overview	TACC Effort	KAISA	Mirage	ThemisIO	TACCGPT	Diamond	Conclusion

RESEARCH

Introducing the AI Research SuperCluster — Meta's cutting-edge AI supercomputer for AI research

RSC: Under the hood

AI supercomputers are built by combining multiple GPUs into compute nodes, which are then connected by a high-performance network fabric to allow fast communication between those GPUs. RSC today comprises a total of 760 NVIDIA DGX A100 systems as its compute nodes, for a total of 6,080 GPUs — with each A100 GPU being more powerful than the V100 used in our previous system. Each DGX communicates via an NVIDIA Quantum 1600 Gb/s InfiniBand two-level Clos fabric that has no oversubscription. RSC's storage tier has 175 petabytes of Pure Storage FlashArray, 46 petabytes of cache storage in Penguin Computing Altus systems, and 10 petabytes of Pure Storage FlashBlade.

Tesla Unveils Top AV Training Supercomputer Powered by NVIDIA A100 GPUs

'Incredible' GPU cluster powers AI development for Autopilot and full self-driving.

Stability AI, the startup behind Stable Diffusion, raises $101M

Kyle Wiggers @kyle_l_wiggers / 12:01 PM CDT • October 17, 2022

Stability AI has a cluster of more than 4,000 Nvidia A100 GPUs running in AWS, which it uses to train AI systems, including Stable Diffusion. It's quite costly to maintain — Business Insider reports that Stability AI's operations and cloud expenditures exceeded $50 million. But Mostaque has repeatedly asserted that the company's R&D will enable it to train models more efficiently going forward.

Nvidia and Microsoft team up to build 'massive' AI supercomputer

NVIDIA

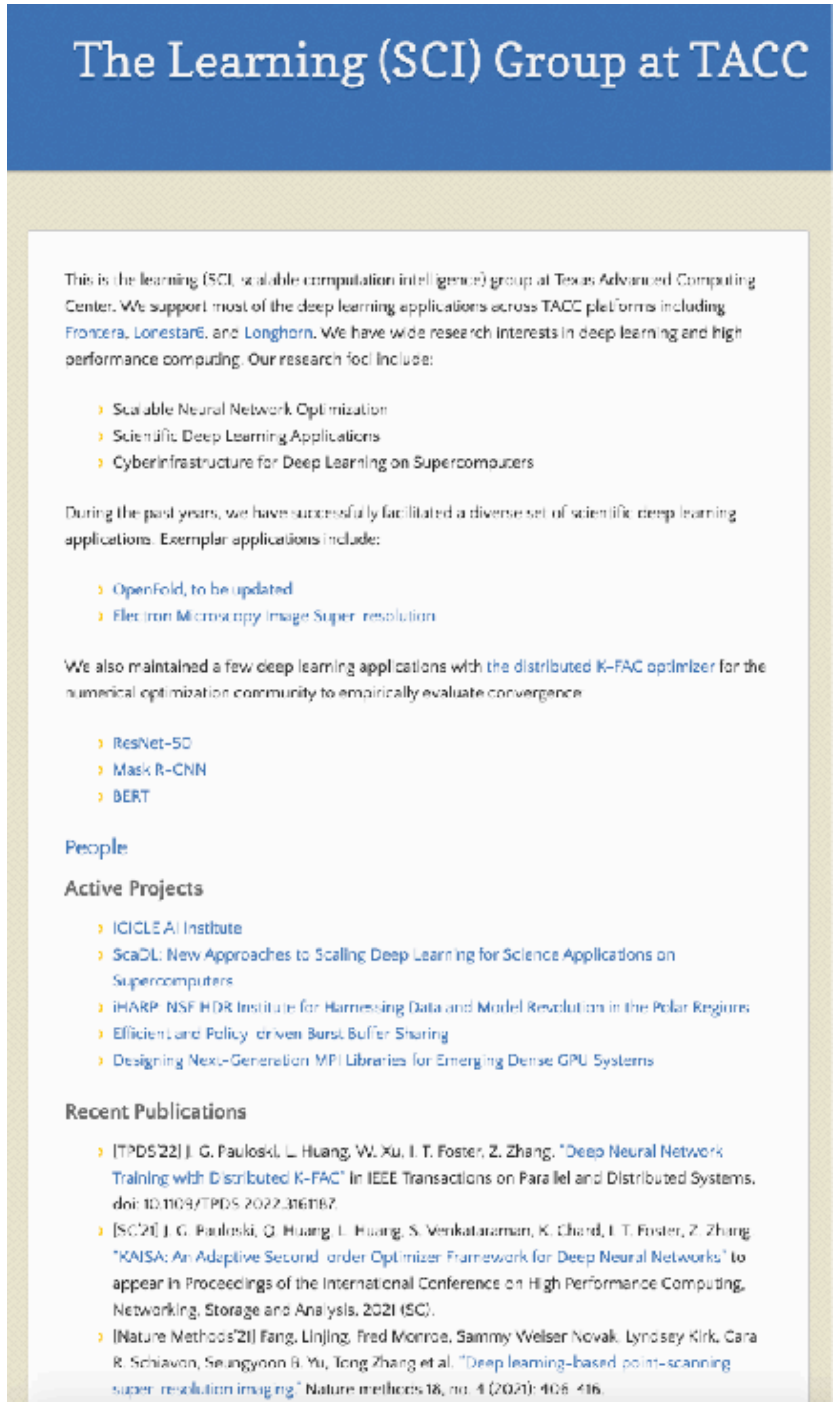
Microsoft

/ The companies hope to create 'one of the most powerful AI supercomputers in the world,' capable of handling the growing demand for generative AI.

"As envisioned, the impact of the NAIRR will be significant and far-reaching, enabling researchers to tackle problems that range from routine tasks to global challenges. In order to achieve its vision and goals, the Task Force estimates the budget for the NAIRR as $2.6 billion over an initial six-year period."

The Learning (SCI) Group

Self Intro	Overview	TACC Effort	KAISA	Mirage	ThemisIO	TACCGPT	Diamond	Conclusion



The Learning (SCI) Group at TACC

This is the learning (SCI, scalable computation intelligence) group at Texas Advanced Computing Center. We support most of the deep learning applications across TACC platforms including Frontera, Lonestar6, and Longhorn. We have wide research interests in deep learning and high performance computing. Our research foci include:

- Scalable Neural Network Optimization
- Scientific Deep Learning Applications
- Cyberinfrastructure for Deep Learning on Supercomputers

During the past years, we have successfully facilitated a diverse set of scientific deep learning applications. Exemplar applications include:

- OpenFold, to be updated
- Electron Microscopy Image Super-resolution

We also maintained a few deep learning applications with the distributed K-FAC optimizer for the numerical optimization community to empirically evaluate convergence.

- ResNet-50
- Mask R-CNN
- BERT

People

Active Projects

- ICICLE AI Institute
- ScaDL: New Approaches to Scaling Deep Learning for Science Applications on Supercomputers
- iHARP: NSF HDR Institute for Harnessing Data and Model Revolution in the Polar Regions
- Efficient and Policy-driven Burst Buffer Sharing
- Designing Next-Generation MPI Libraries for Emerging Dense GPU Systems

Recent Publications

- [TPDS'22] J. G. Pauloski, L. Huang, W. Xu, I. T. Foster, Z. Zhang. "Deep Neural Network Training with Distributed K-FAC" In IEEE Transactions on Parallel and Distributed Systems, doi: 10.1109/TPDS.2022.3161187.
- [SC'21] J. G. Pauloski, Q. Huang, L. Huang, S. Venkataraman, K. Chard, I. T. Foster, Z. Zhang "KAISA: An Adaptive Second-order Optimizer Framework for Deep Neural Networks" to appear in Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis, 2021 (SC).
- [Nature Methods'21] Fang, Linjing, Fred Monroe, Sammy Weiser Novak, Lyndsey Kirk, Cara R. Schiavon, Seungyoon B. Yu, Tong Zhang et al. "Deep learning-based point-scanning super-resolution imaging." Nature methods 18, no. 4 (2021): 406-416.

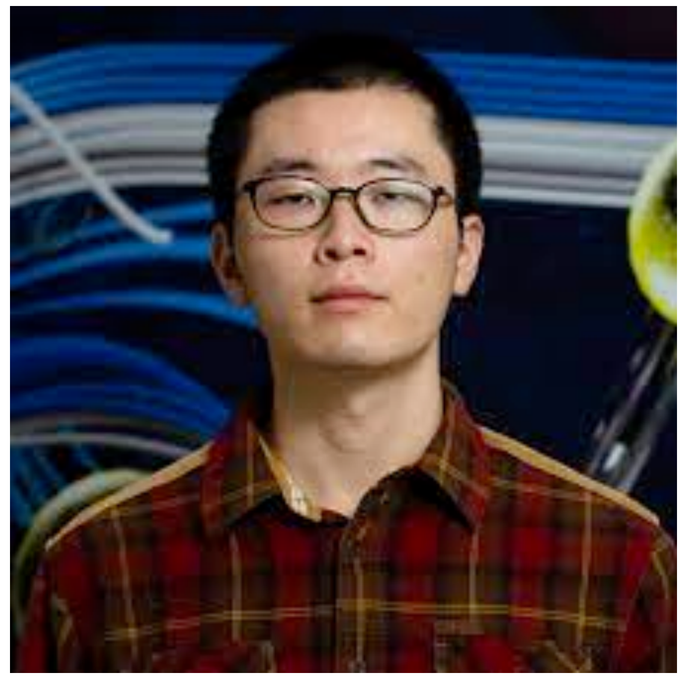
- Staff Members

Zhao Zhang

Juliana Duncan

Sikan Li

Amit Gupta



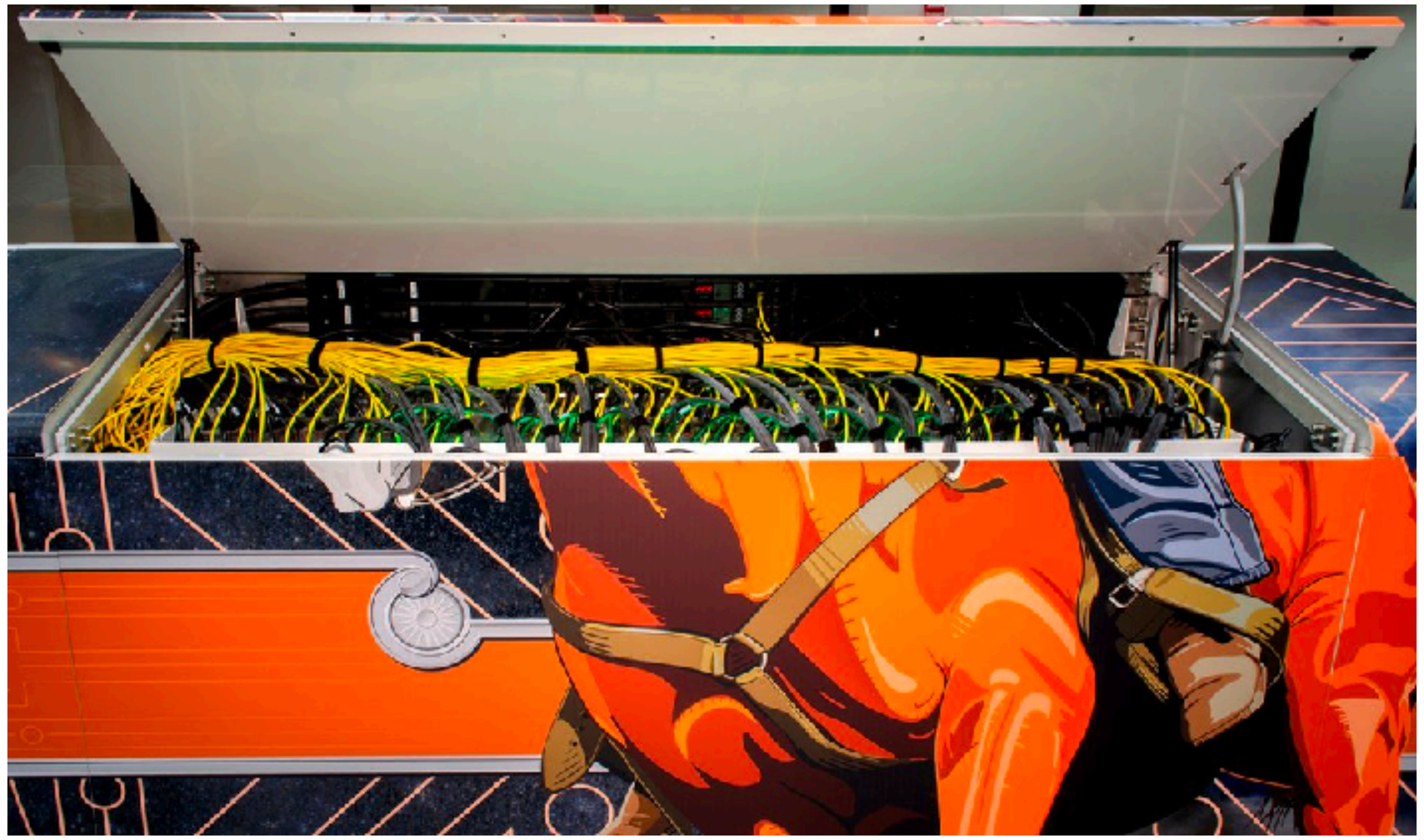
Mingkai Zheng

Deep Learning Hardware at TACC

- Frontera
  - Primary compute system:
    - 39PF PetaFlops Peak Performance -- 8,368 nodes of Intel Cascade Lake
    - 16 Large memory nodes — 2.1TB NVDIMM and 3.2 TB local storage
  - GPU Subsystem:
    - 90 node with four RTX5000 GPU each

Deep Learning Hardware at TACC

- Lonestar6
  - 560 compute nodes, each with two AMD EPYC 7764 processors (Milan)
  - 72 GPU nodes, each with three Nvidia A100 GPUs



Deep Learning Software

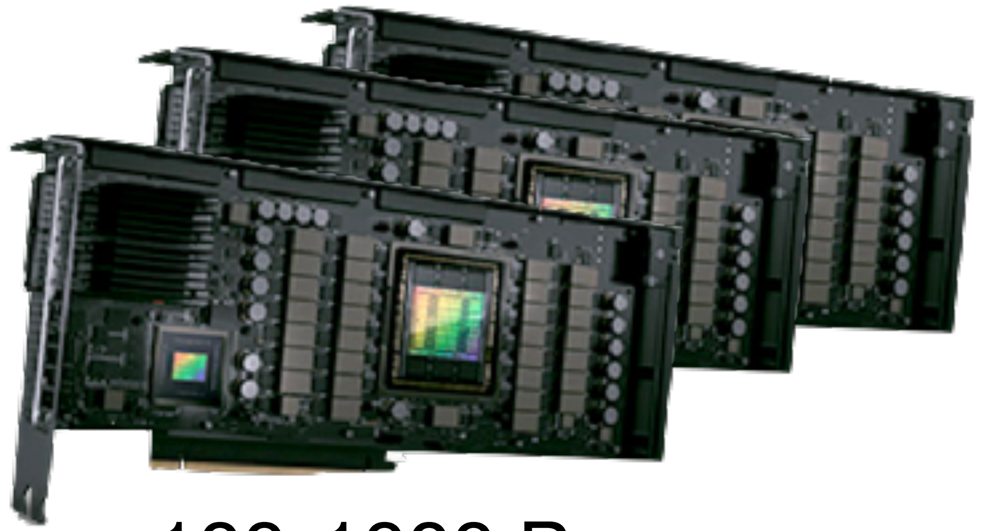
- DL Frameworks
  - PyTorch (primary focus), TensorFlow
- Distributed DL Frameworks
  - torch.distributed, DeepSpeed, Horovod, mpi4py
- Front-end Interface
  - Jupyter Notebooks via TACC Analysis Portal
- Applications
  - BERT, GPT-NeoX, ResNet, Mask R-CNN, OpenFold, DeepSpeed-chat, HuggingFace

Deep Learning Support Focus

- Efficient and scalable large neural network optimization on supercomputers
- User-friendly AI cyberinfrastructure
  - Plug-and-play AI with ICICLE software stack
  - User-friendly GPU cluster interface with embedded HPC and DL knowledge

Research Landscape: HPC+AI

Self Intro	Overview	TACC Effort	KAISA	Mirage	ThemisIO	TACCGPT	Diamond	Conclusion



100-1000 Processors

Supercomputer

- Non-convex Optimization
  - [TPDS’19, ICPP’18] Large-batch Training
  - [TPDS’22, SC’21, SC’20]2nd-order Optimization *
  - [In progress] Gradient Sparsification *
  - [In progress] Lossy Compression on 2nd-order Info *
- I/O System
  - [IPDPS’20] Efficient I/O for Neural Network Training with Compressed Data
  - [SC’23] Fine-grained Policy-driven I/O Sharing for Burst Buffers
- HPC System
  - [SC’23] Mirage: Towards Low-interruption Services on Batch GPU Clusters with Reinforcement Learning”

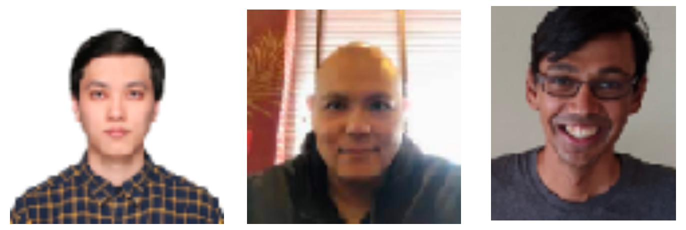
Berkeley | EECS
Electrical Engineering and Computer Sciences
UCLA Samueli Computer Science

THE UNIVERSITY OF CHICAGO DEPARTMENT OF COMPUTER SCIENCE
Argonne NATIONAL LABORATORY
Computer Sciences

THE OHIO STATE UNIVERSITY COLLEGE OF ENGINEERING
ICICLE DEMOCRATIZING AI

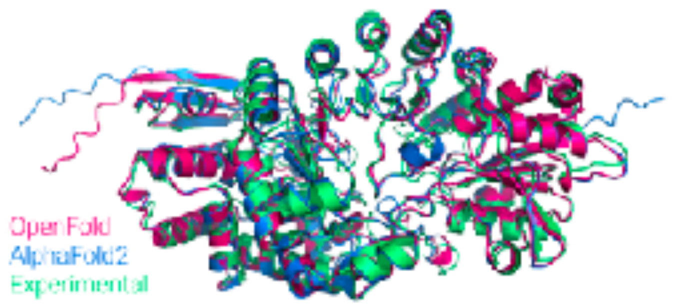
Indiana University Bloomington

The Grainger College of Engineering Computer Science
NCSA

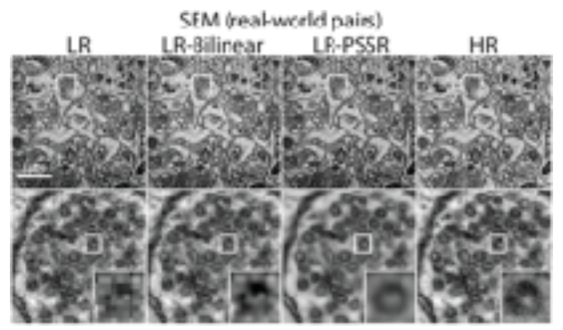


Computer Sciences

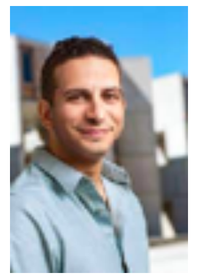
Research Landscape: Science



- [In submission to Science] OpenFold, an open source implementation of AlphaFold



- [Nature Method'21] SRGAN, super-resolution of low-dose electromagnetic brain images



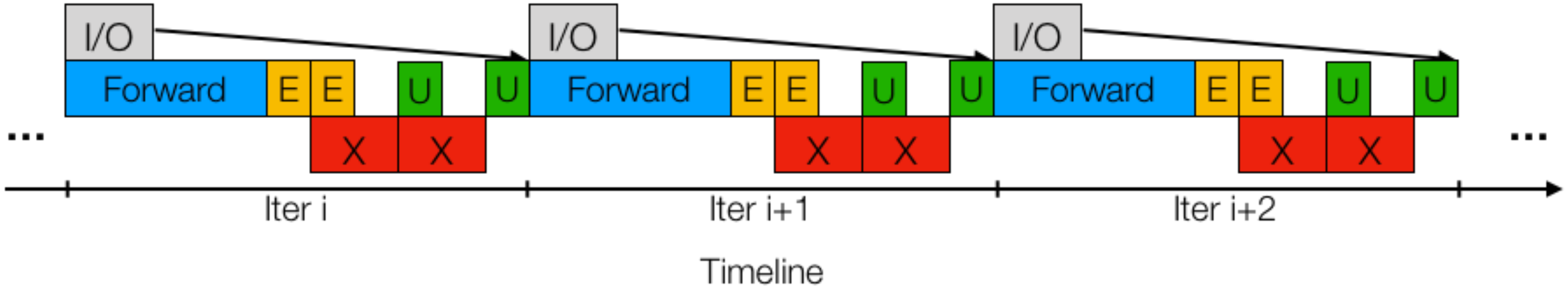
- [In progress] Animal Ecology

- [In progress] Digital Agriculture

Deep Learning Support Focus

- TACC Machine Learning Summer Institute
- TACC Machine Learning Tutorial

Distributed Deep Learning



- Opportunities for faster training
  - Reduce forward time cost, e.g., flash attention, fused attention
  - Reduce gradient exchange time cost, e.g., sparsification, gradient compression
  - Reduce number of iterations —> second-order optimization

Evaluation: Time-to-Convergence w/ Fixed Batch Size

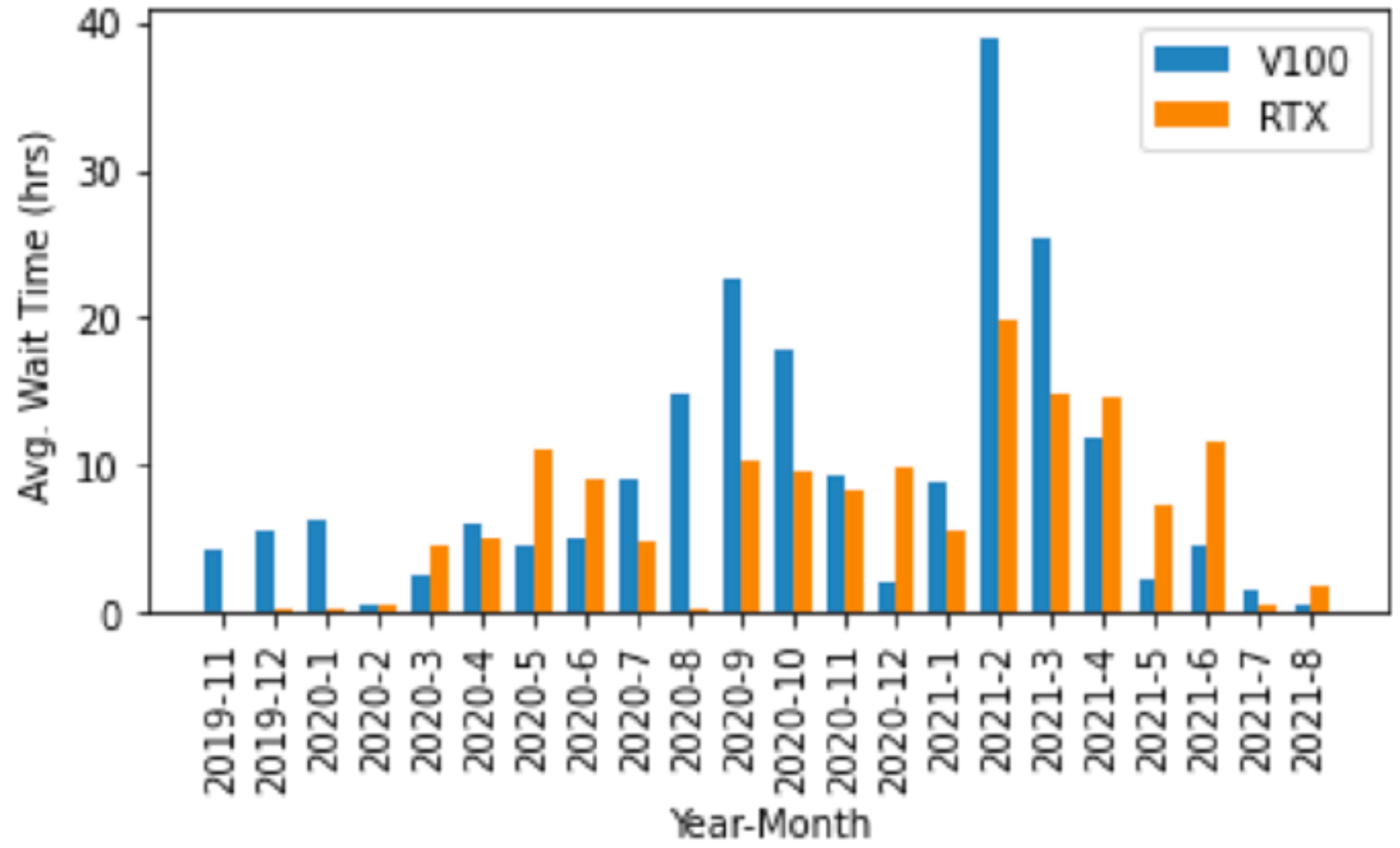
App	Default Optimizer	Baseline	# GPUs	Global Batch Size	Precision	KAISA Time-to-Convergence
ResNet-50	SGD	75.9% Val. Acc.	8 A100	2,048	FP16	**24.3%**
Mask R-CNN	SGD	0.377 bbox mAP 0.342 segm mAP	64 V100	64	FP32	**18.1%**
U-Net	ADAM	91.0% Val. DSC	4 A100	64	FP32	**25.4%**
BERT-Large (Phase 2)	LAMB	90.8% SQuAD v1.1 F1	8 A100	65,536	FP16	**36.3%**

Evaluation: Time-to-Convergence w/ Fixed Memory Budget

App	Optimizer	# GPUs	Grad. Worker Frac.	Local Batch Size	Time-to-Convergence
ResNet-50	SGD	64 V100	--	128	**123**
	KAISA		1/64	80	**96**
	KAISA		1/2	80	**83**
BERT-Large (Phase 2)	LAMB	8 A100	--	12	**2918**
	KAISA		1/2	8	**1703**
	KAISA		1	8	**1704**

Mirage: Intelligent Resource Provisioning with Reinforcement Learning

- Training large models on batch GPU clusters experiencing long interruptions between jobs
  - Training the 345-million parameter BERT takes 5 days on 8 A100 GPUs
  - Training the 20-billion parameter GPT-Neox model takes 30 days on 96 A100 GPUs
  - Computing centers usually enforce a fixed wall clock time
    - TACC has 48-hour limit
    - NERSC Perlmutter has 6-hour limit
    - ALCF Theta GPU has 12-hour limit

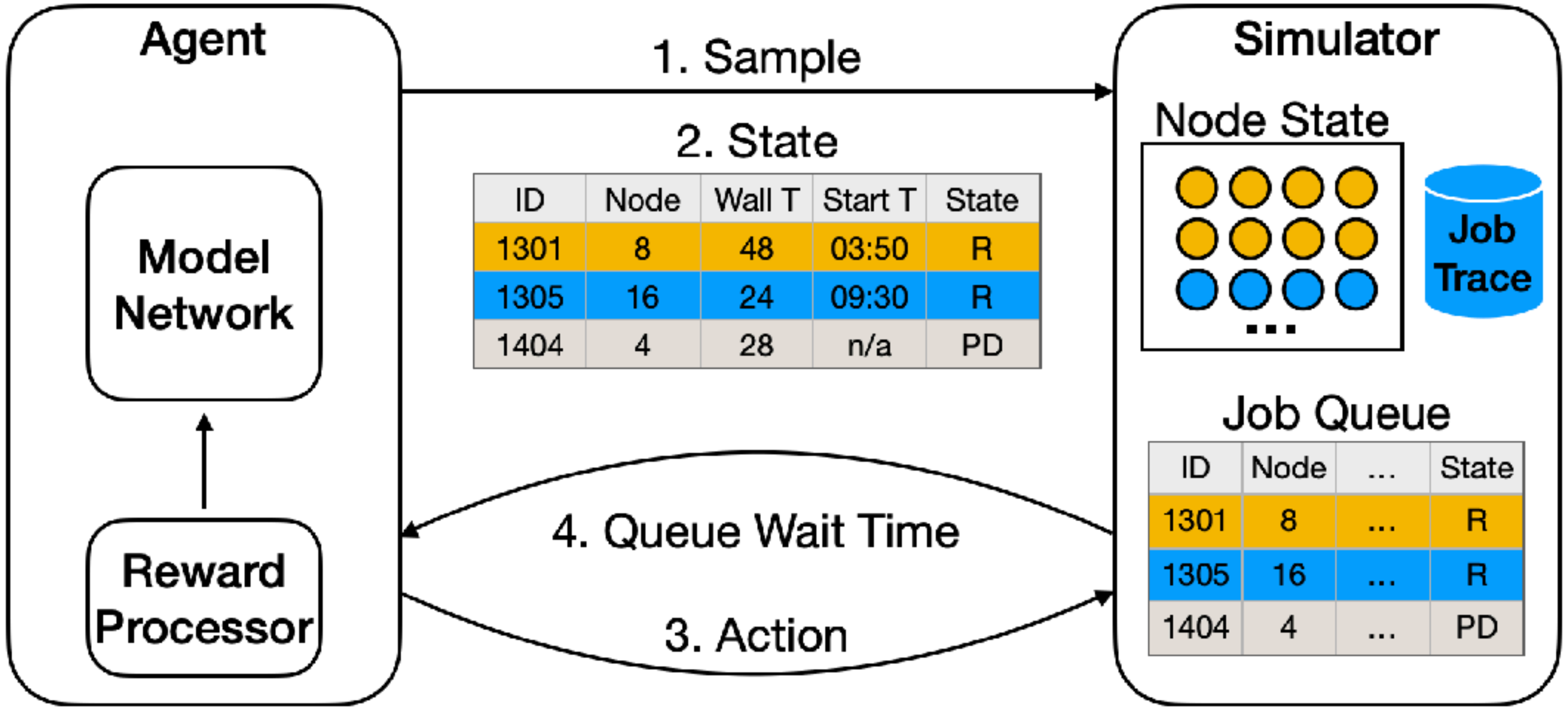


Mirage: Intelligent Resource Provisioning with Reinforcement Learning

- The Reactive Baseline
  - Using SLURM job arrays
    - “The age factor represents the length of time a job has been sitting in the queue and eligible to run. In general, the longer a job waits in the queue, the larger its age factor grows. However, **the age factor for a dependent job will not change while it waits for the job it depends on to complete**. Also, the age factor will not change when scheduling is withheld for a job whose node or time limits exceed the cluster's current limits.”
  - Equivalent to submit the subsequent job upon the completion of the current one
- The Average Baseline
  - monitoring the average queue wait time T_{avg} and submitting the second job T_{avg} time units before the first job finishes

Mirage: Intelligent Resource Provisioning with Reinforcement Learning

- When to submit the next job forms a sequence of actions, which only depends on the current machine state (running jobs, queuing jobs) —> Markov Decision Process



Mirage: Intelligent Resource Provisioning with Reinforcement Learning

- Deep Q-Learning (DQN)
  - Evaluating the State-Action Value Function

$$L(\theta_{k+1}) = E[(R + \gamma max_{a'} Q(s', a'; \theta_k) - Q(s, a; \theta_{k+1}))^2]$$

- Policy Gradient (PG)

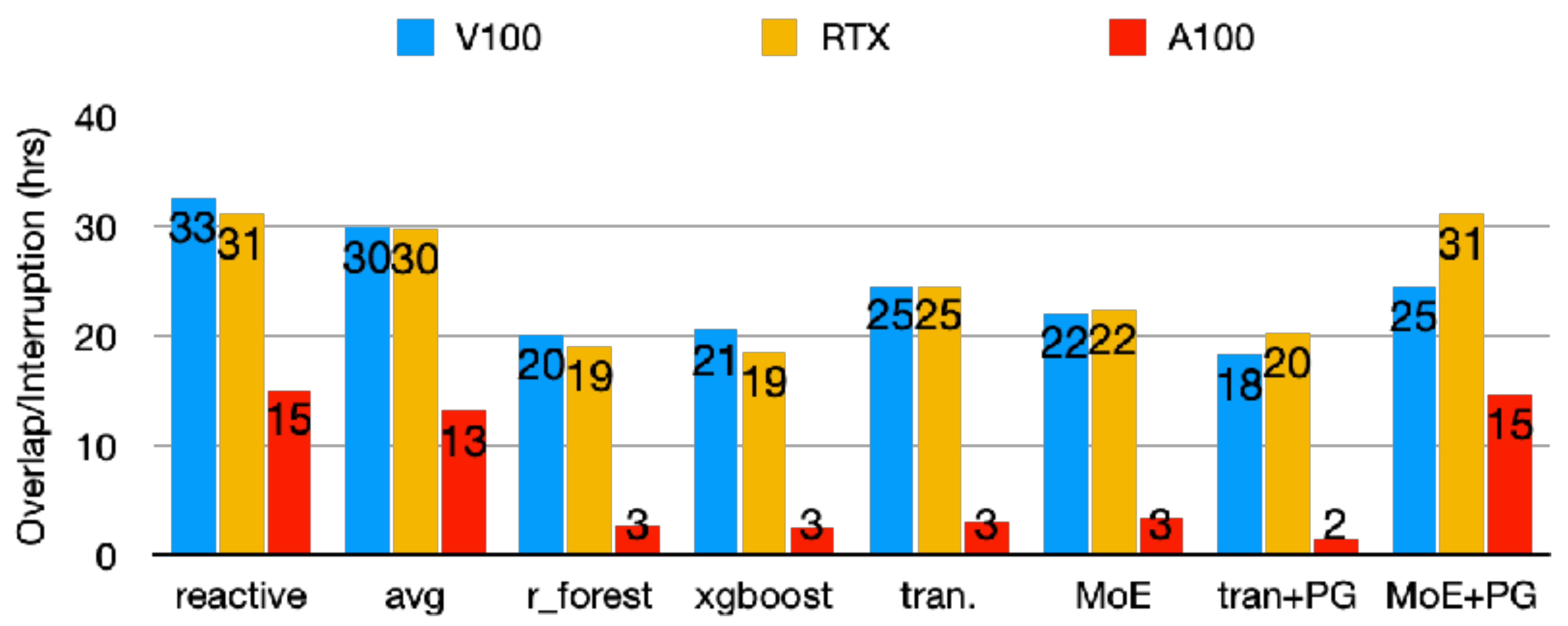
$$J(\pi_\theta) = \underset{\tau \sim \pi_\theta}{\mathrm{E}} [R(\tau)]$$

- Mixture-of-Experts (MoE)

$$Q(s, a) = \sum_{e=1}^{E} G_\theta(e) Q_e(s, a) \quad , G_\theta(\cdot) = softmax(x \cdot W_\theta)$$

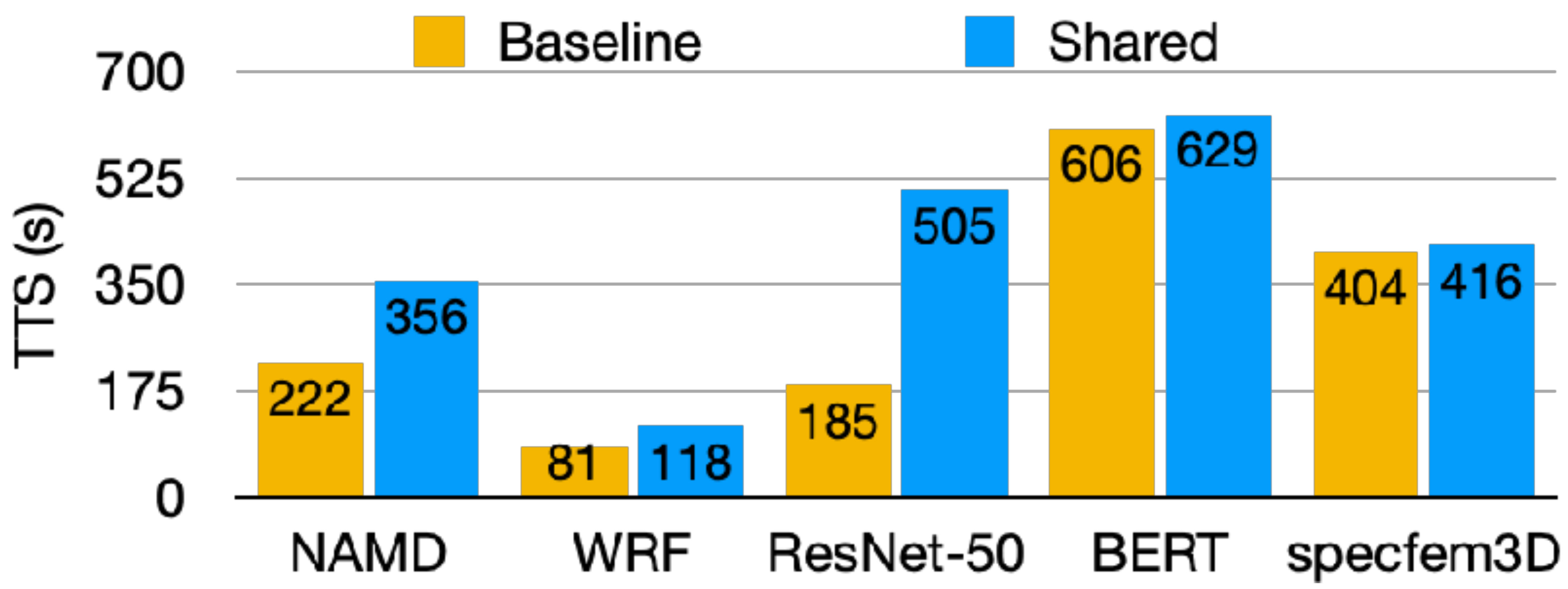
Mirage: Intelligent Resource Provisioning with Reinforcement Learning

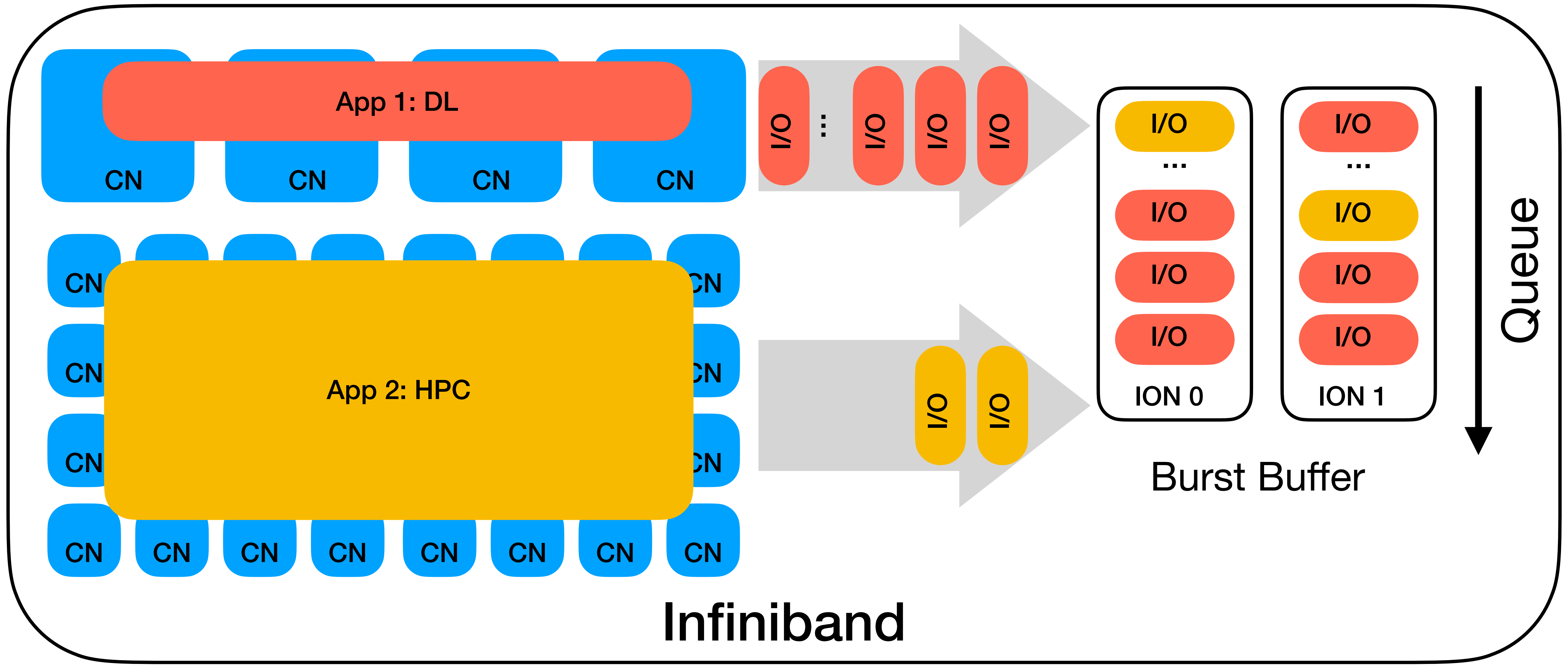
- Transformer+PG has the best overall interruption elimination performance compared to all methods, followed by ensemble methods
- They introduce 2x longer overlap (~4 hours with 48 hour jobs) compared to MoE+DQN.
- Mirage uses the MoE+DQN as its default model. We leave transformer+PG as an option for users as an aggressive provisioner, which will be more effective when the machine load is high.



ThemisIO: Fine-grained Policy-driven I/O Sharing for Burst Buffers

- I/O interference on HPC



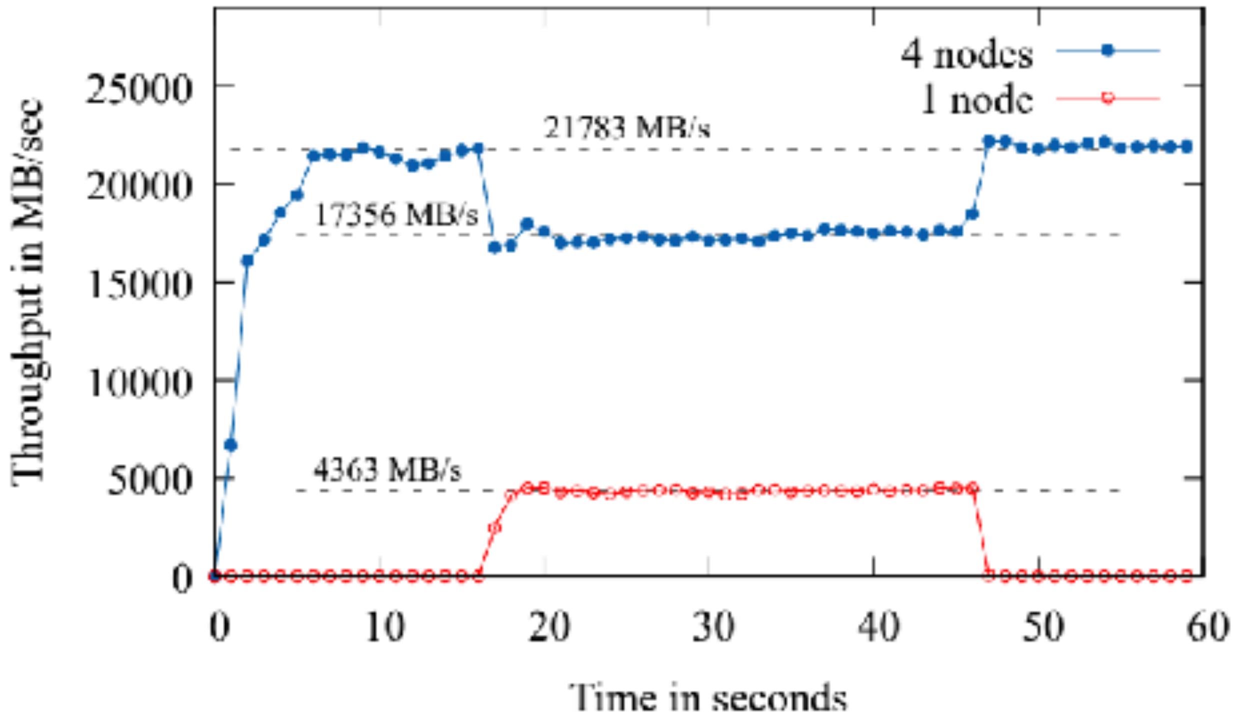


- Embedding job meta info (job size, user id, job id) in I/O request
- Size-fair, job-fair, user-fair
- User-then-size-fair, group-user-size-fair

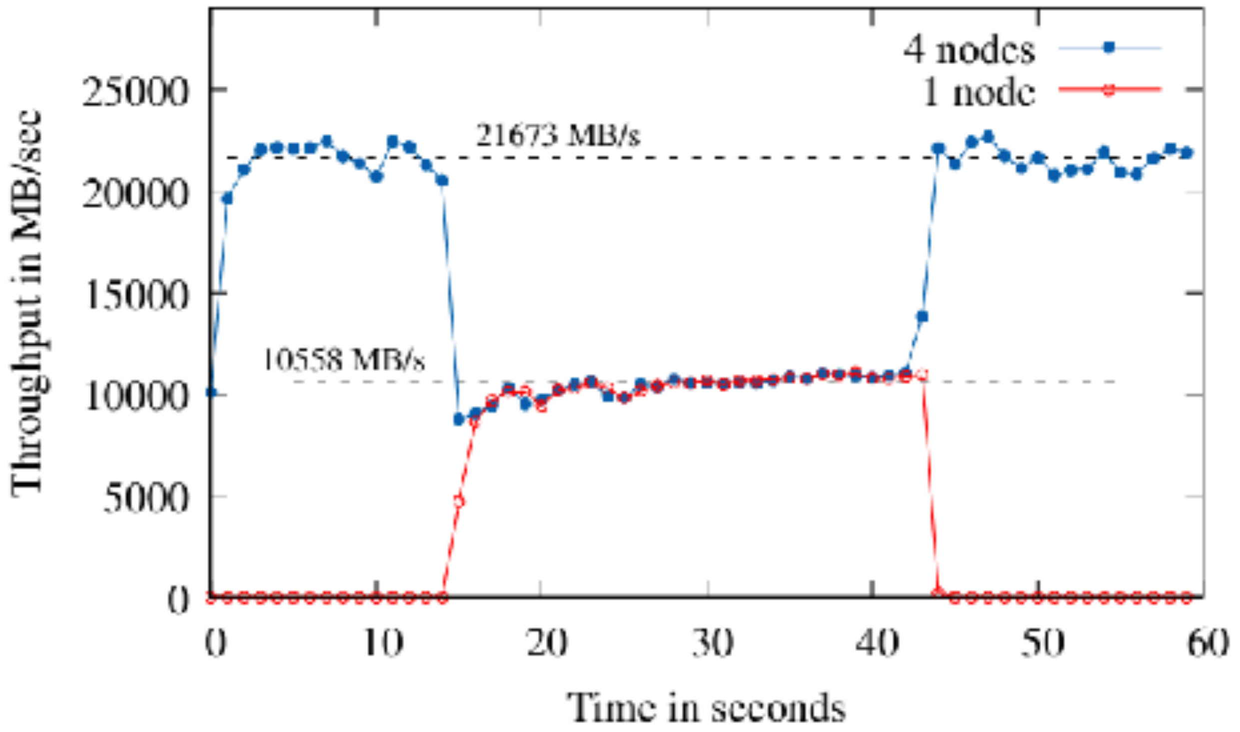
ThemisIO: Fine-grained Policy-driven I/O Sharing for Burst Buffers

Self Intro	Overview	TACC Effort	KAISA	Mirage	ThemisIO	TACCGPT	Diamond	Conclusion

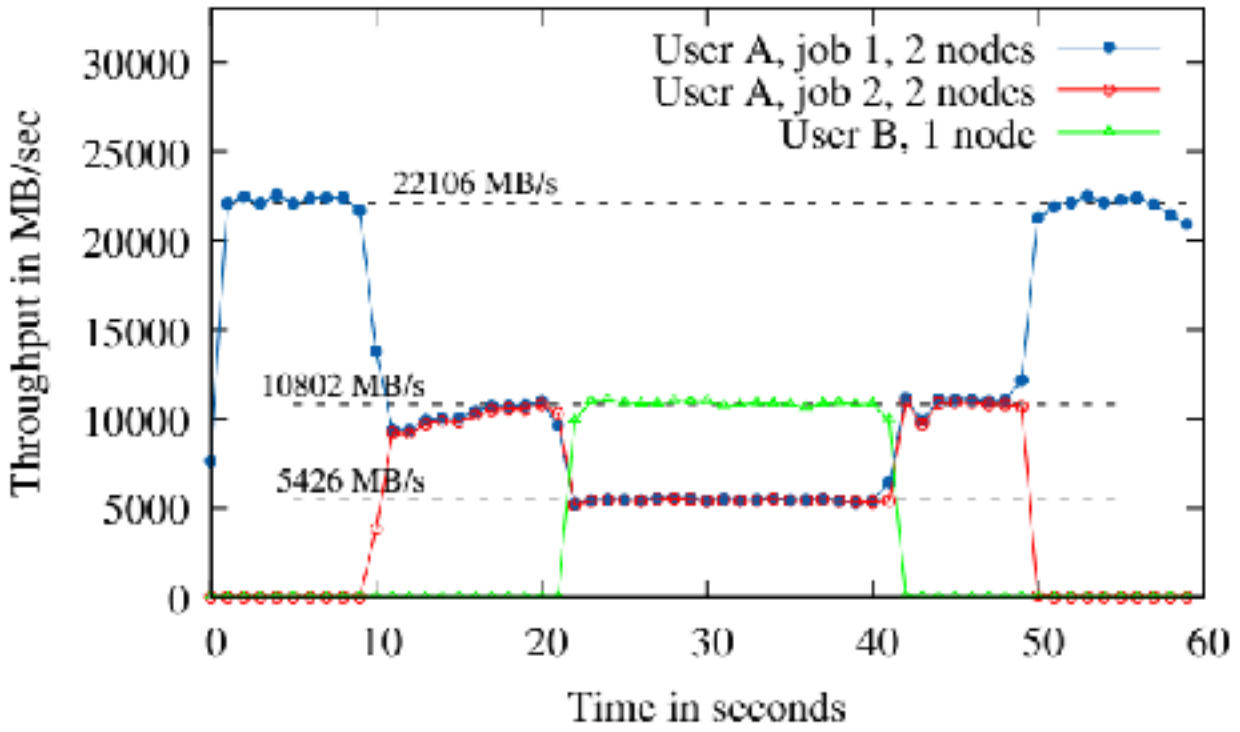
- Benchmark Sharing



(a) Size-fair, 4-node job competing with 1-node job

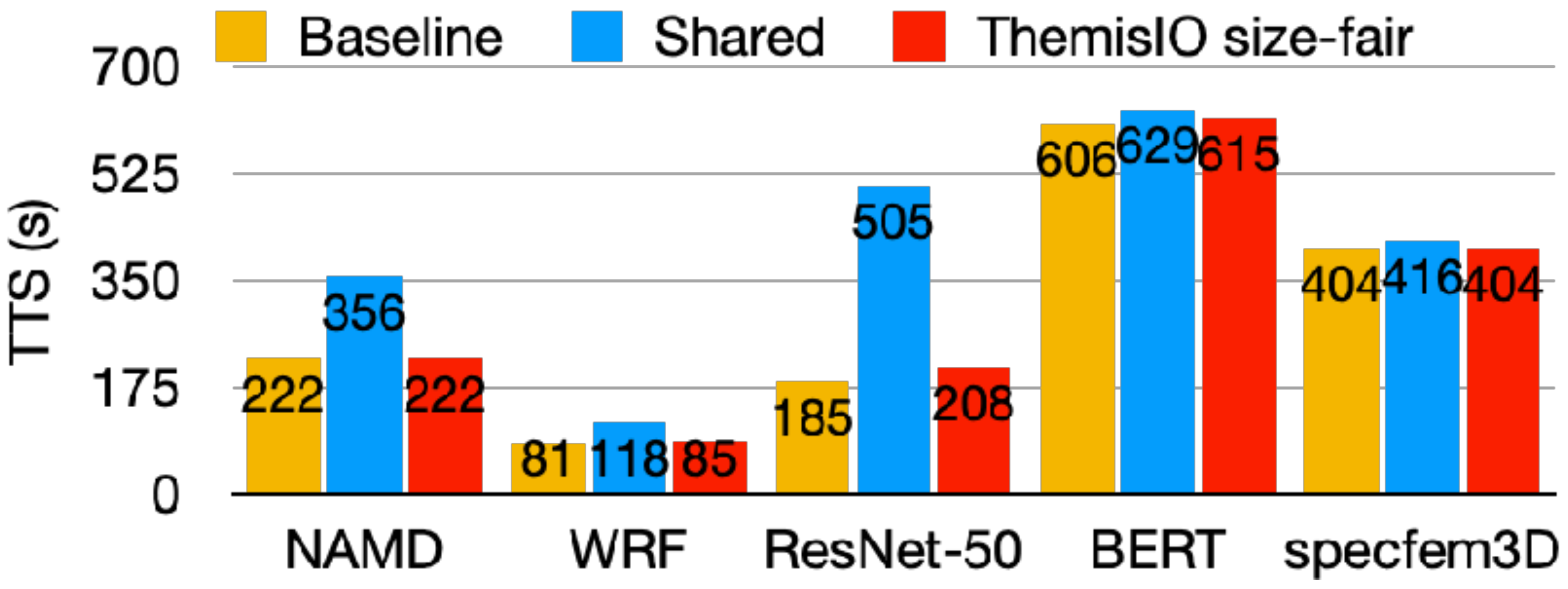


(b) Job-fair, 4-node job competing with 1-node job



(c) User-fair, Two 2-node jobs competing with a 1-node job

- Application Sharing



Funding Source

- NSF OAC-2106661 “Collaborative Research: OAC Core: ScaDL: New Approaches to Scaling Deep Learning for Science Applications on Supercomputers” (10/1/21-9/30/24)

- NSF OAC-2112606 “AI Institute for Intelligent CyberInfrastructure with Computational Learning in the Environment (ICICLE)” (11/1/21- 10/31/26)

- NSF OAC-2008388 “Collaborative Research: OAC Core: Small: Efficient and Policy-driven Burst Buffer Sharing” (10/1/20-9/30/22)

- NSF OAC-1931354 “Collaborative Research: Frameworks: Designing Next-Generation MPI Libraries for Emerging Dense GPU Systems” (11/1/19- 10/31/22)

Questions?

zzhang@tacc.utexas.edu