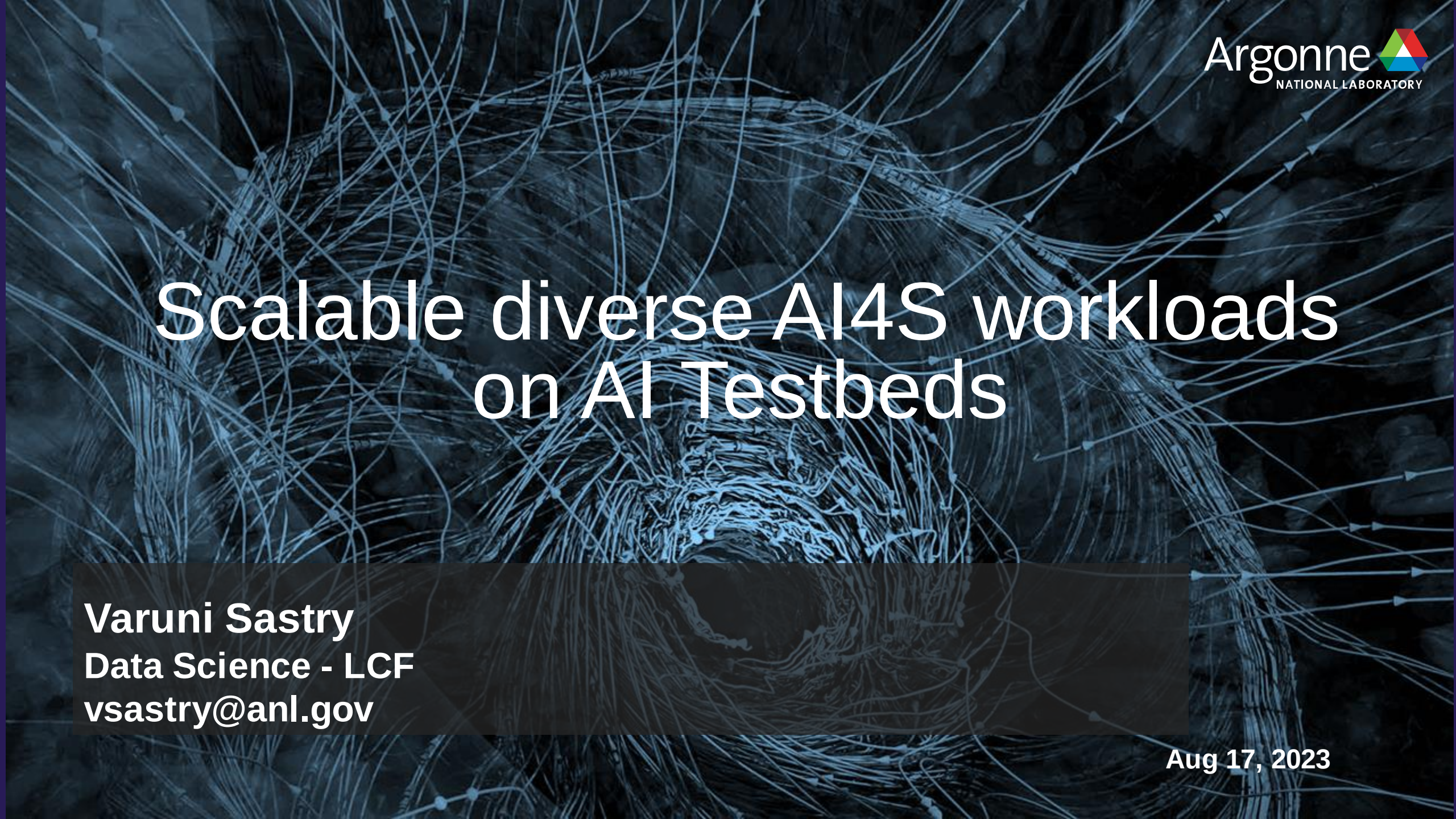

Scalable diverse AI4S workloads on AI Testbeds

Varuni Sastry
Data Science - LCF
vsastry@anl.gov

Aug 17, 2023


ALCF AI Testbeds



Evaluation and Benchmarking of DL workloads

- DL primitives – GEMMs, Conv2d, ReLu, Recurrent, Transformer block (multi-head attention + FFN)
- CV models - Unet-2D on Brain MRI data, ResNet on ImageNet
- NLP models - Bert, GPT-xl, GPT-13B.
- Sparsity Study – 2x speed up with extremely sparse with 15% degradation in accuracy

Science Applications

- Uno - Drug response to cancer cells (I/O bound problems)
- GenSLMs - Large language Models for genomic data. (ACM Gordon Bell Special Prize SC'22)
- BraggNN, PtychoNN, AutoPhaseNN - Bragg Coherent Diffraction Imaging (BCDI) with 3D Convolutions and 3D FFT. High resolution image data with APS upgrade.
- Scaling the “Memory Wall” for Multi-dimensional Seismic Processing with Algebraic Compression on CS-2 (Gordon Bell Finalist 2023.)


ALCF AI Testbeds



Key Insights

- Ability to handle high resolution data and better scaling efficiency compared to GPUs.
- Works for larger batch sizes, larger sequence lengths.

HPC focus

- Cerebras Software Language (CSL) - (Massively scalable stencil algorithm).
- C++ SDK on SambaNova (experimental version).

Collaboration Opportunities

- LLMs with Large context length window (Ashton Wells).
- Unet2d based models for MRI images. (Marta Garcia).
- Cosmological simulations (surrogates) and other HPC kernels.
- Multimodal models.