Plenary Presentation at KSTAR Conference 2019
February 20-22, 2019, COEX, Seoul Korea

Federated Data-Science for Accelerated KSTAR (& ITER) Collaboration: a new paradigm

CS Chang[1]

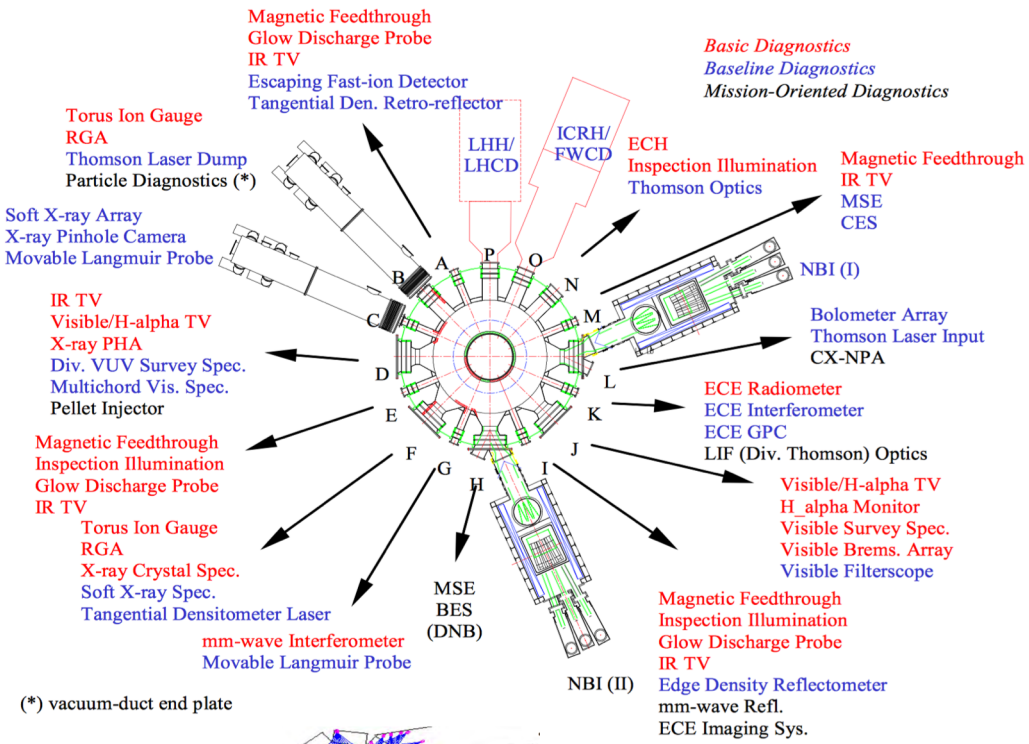
R.M. Churchill[1], R. Hager[1], S. Ku[1], (Ralph Kube[1]), JS Park[2], MJ Choi[2], S. Klasky[3], J. Choi[3], M. Wolf[3], J. Wong[3], R. Gerber[4], K. Antypas[4], Prabhat[4], D. Bard[4], L. Stephey[4], R.S. Canon[4], R. Thomas[4], W. Bhimji[4], H. Park[5], E. Dart[6], BS Cho[7]

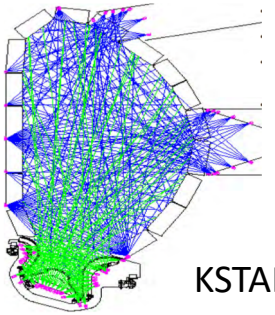
[1]PPPL, KSTAR[2], ORNL[3], NERSC[4], UNIST[5], ESnet[6], KISTI[7]

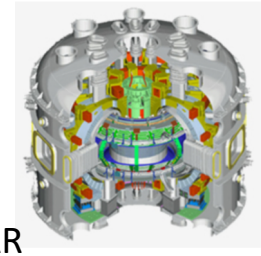
Outline

- What is the federated data-science and why does fusion need this new paradigm?
- Data federation workflow utilizing
  - computer science, applied math and artificial intelligence tools
- Some basics of machine learning techniques
- Present status of the federated workflow research in the KSTAR-NERSC-PPPL (HBPS-Group)-ORNL-ESNet-KREONET Consortium
- Short-term plans
- Discussions

A tokamak reactor is packed with thousands of sensors to be federated together with many users and many computers







KSTAR

World Fusion Facilities

Fusion data from experiment and simulation possess all the big V properties

5 Vs of big ITER data

Volume
- ~1PB/day
- Subsequent simulations can greatly enhance data volume

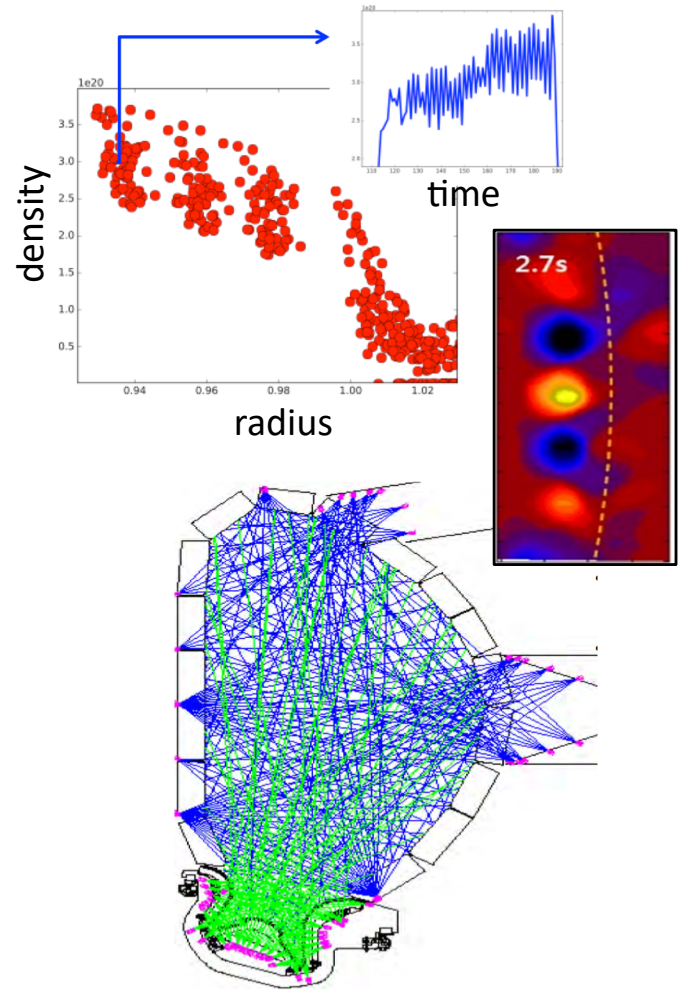
Velocity
- 50GB/s for up to 1000s, each shot
- Hierarchy of analysis for near-real-time feedback

Variety
- Wide variety data from thousands of sensors
- Many-scale many-physics

Value
- All the expensive sensors are there because they produce valuable data

Veracity
- All the data have different quality: e.g., validity range, error, and noise

AI/DL in high demand



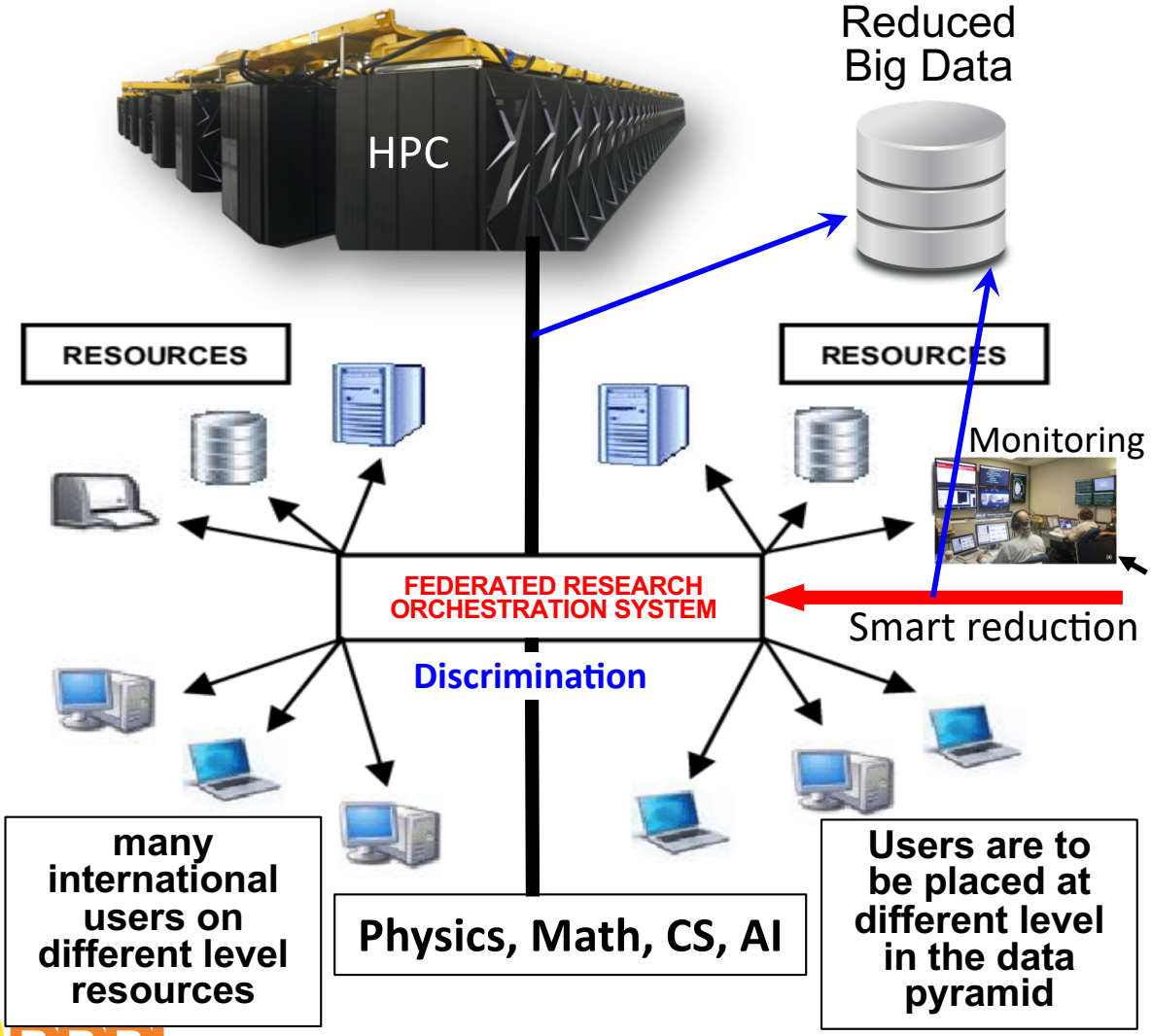
Why an end-to-end federated framework?

- Fusion plasma dynamics is **multi-scale, multi-physics** phenomena.
  - Nonlinear interactions (self-organization) determines physics outcome
  - Physics should be determined jointly from many instruments or scales
  - Each instrument has its own quality, error, and uncertainty (veracity),
  - Error from each instrument propagates to error in other physics estimates
- Different instruments are managed by **different scientists**
  → human integration → not efficient
- **Future fusion devices** may be in **different physics** regime
- **Experts and computing capabilities** are **scattered** all over the world
  - **Experimental steering takes too long** for fast scientific progress
  - Combining-in the more complete physics (from 1st principles simulations) will be difficult.

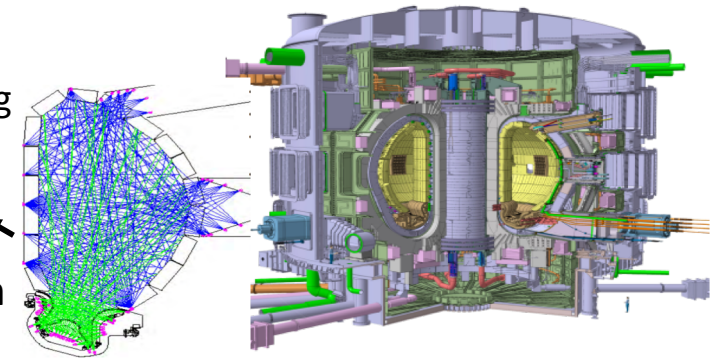
Utilize computer science, applied math and AI tools.

Simplify analysis setup: end-to-end abstraction for parallelization and libraries

System structure in the federated fusion research



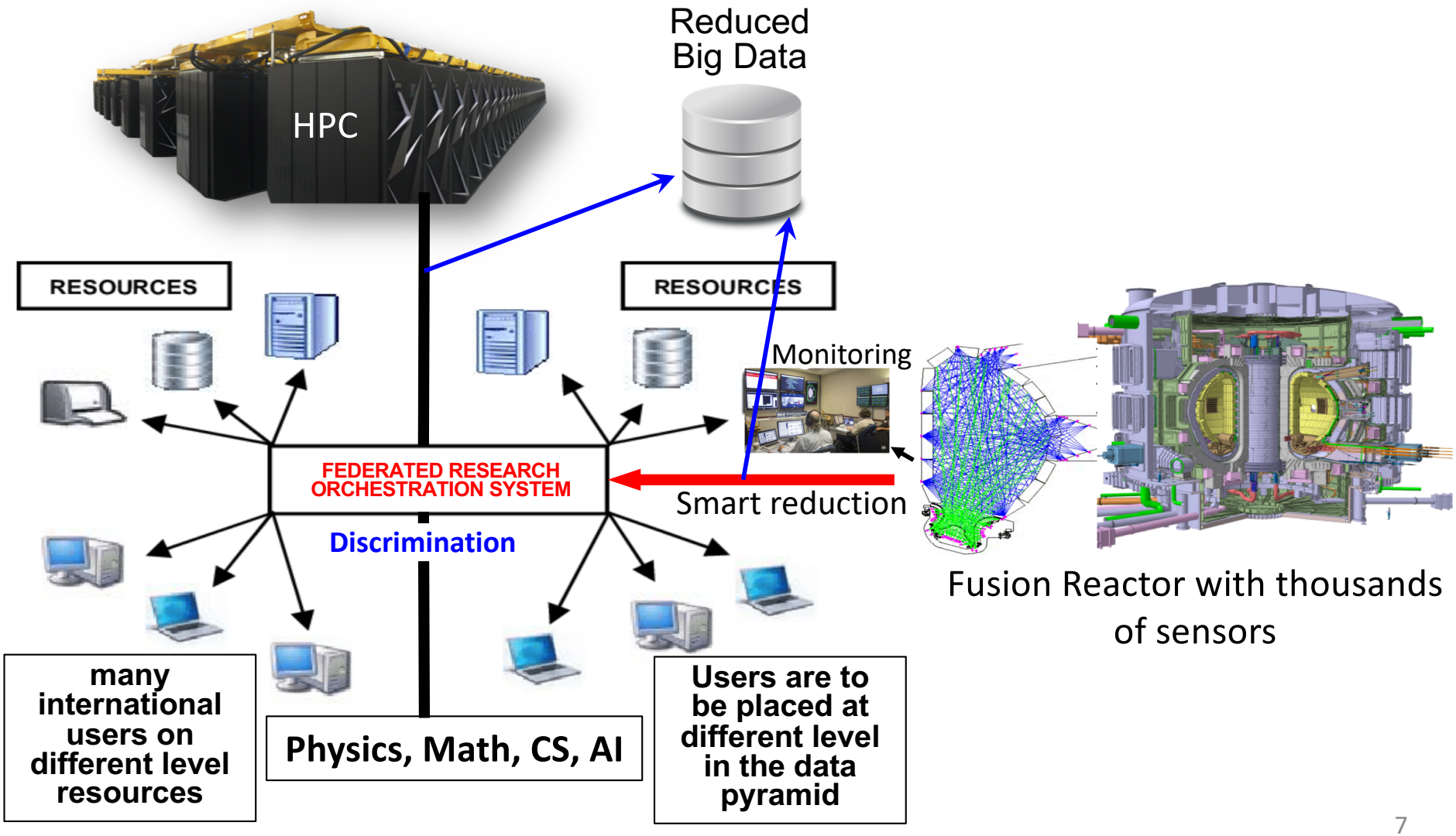
(In future DEMO reactors in "nuclear" environment, sensors can be limited, and **HPC**s will become a more important part of the federated research.)



Fusion Reactor with thousands of sensors

A preliminary workflow was created in the ICEE project by ANL-ORNL-PPPL-KSTAR-ESnet.

System structure in the federated fusion research



Physics workflow in the federated fusion research

Pre & post computing at large scale

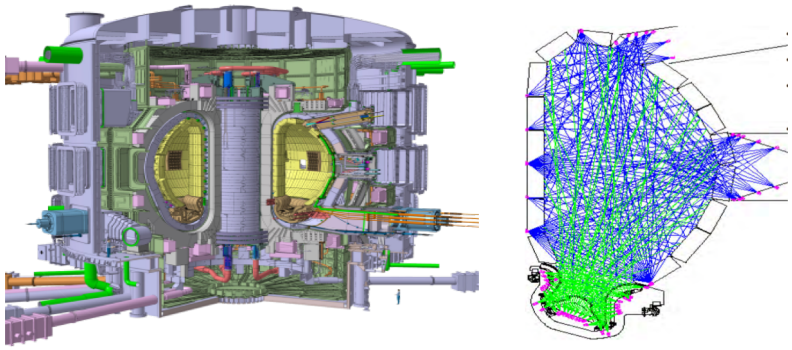
- Guide "next-day" experiments
- Understand fundamental physics
- Simulation data too big for filesystem
- Many collaborators → Federated
- In-memory on-the-fly analysis/vis, & reduction: train AI tools

Small-scale, reduced model, empirical computing

- For the quick empirical analysis and experimental support

understand

analyze



Experiments: high cost

- Fusion data is big: ITER~400PB/yr
- Thousands of sensors: PB per day
  - Data analized & reduced near instruments
  - Physics too complicated for onsite study
  - Near-realtime federated data flow needed
  - **"Streaming" data-science with AI**

Experiment → analysis/understanding → steer present/next experiment

Goal: Accelerated & cheaper realization of commercial reactors

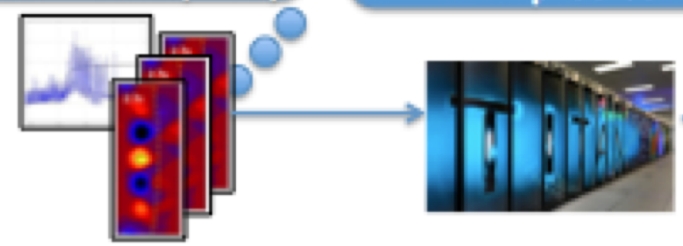
Example KSTAR workflow:

Machine learning for online ECEi frame classification

Offline machine learning model training

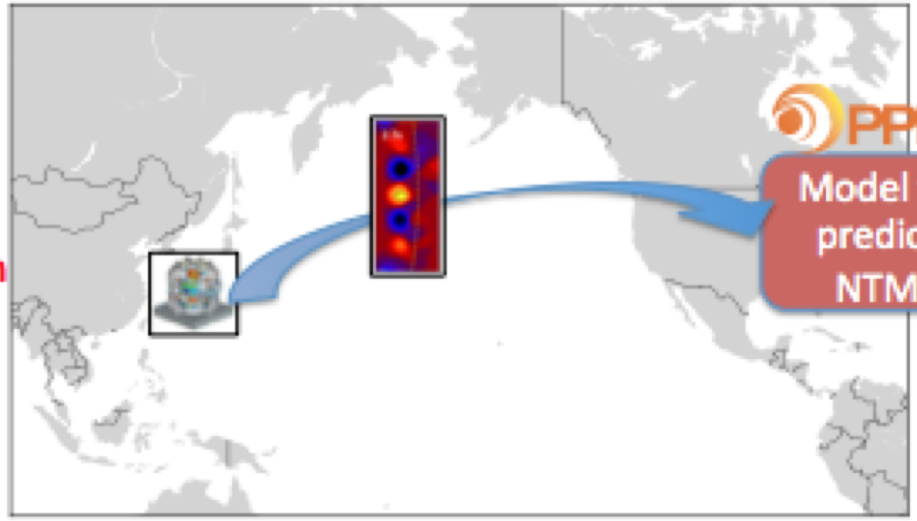
Large KSTAR multi-shot dataset (ECEi + other, ~TB)

TRAIN on HPC resources Machine learning model to predict NTM



Model to predict NTM

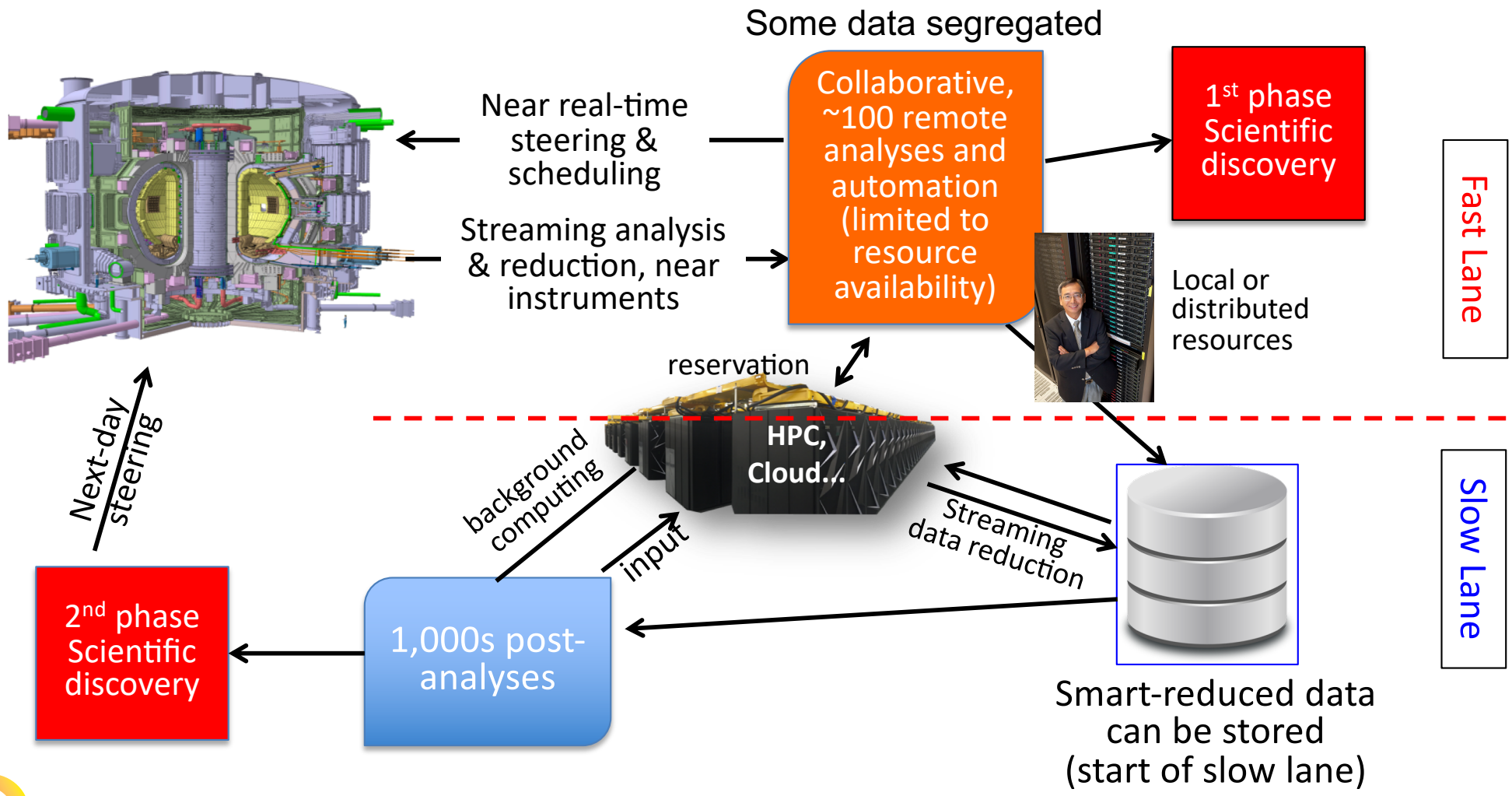
Online NTM prediction and ECEi frame classification with trained machine learning model



Model to predict NTM

Allows scientists to quickly find physics of interest in large datasets like ECEi

Slow and Fast Data lanes in federated fusion research



- NERSC has strong ML support programs for DOE Offices.
- Several other areas are already in the NERSC program.
- The KSTAR-PPPL initiative will get fusion into the NERSC ML program

Deep Learning Software

NERSC supports several software frameworks for machine learning and deep learning. If there is a framework you would like to see evaluated and supported at NERSC, please **let us know**.

Anaconda

Most standard machine learning and deep learning packages are available via the anaconda installation on both Edison and Cori. Anaconda is the easiest way to access and use these frameworks, and is the way we recommend most users get started with machine learning at NERSC. The packages include:

- Tensorflow
- XGBoost
- Scikit-Learn
- Theano
- Torch

Spark

We also provide **Spark** as a **separate module**.

Caffe

Caffe is also available in a separate module - simply use `module load caffe/master`

Back to Top

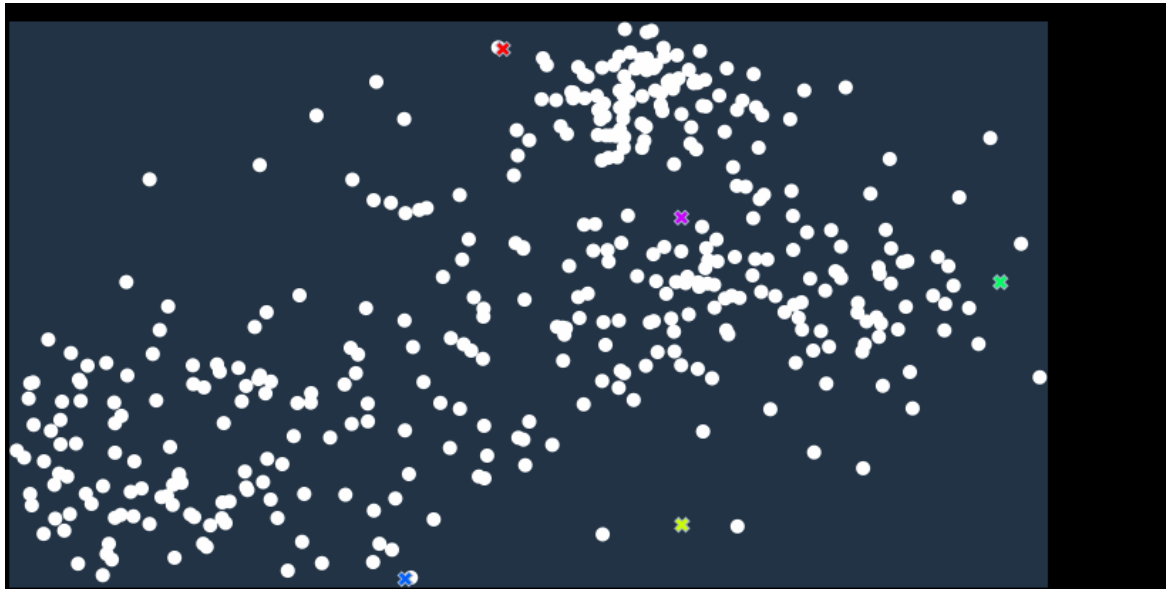
Tensorflow

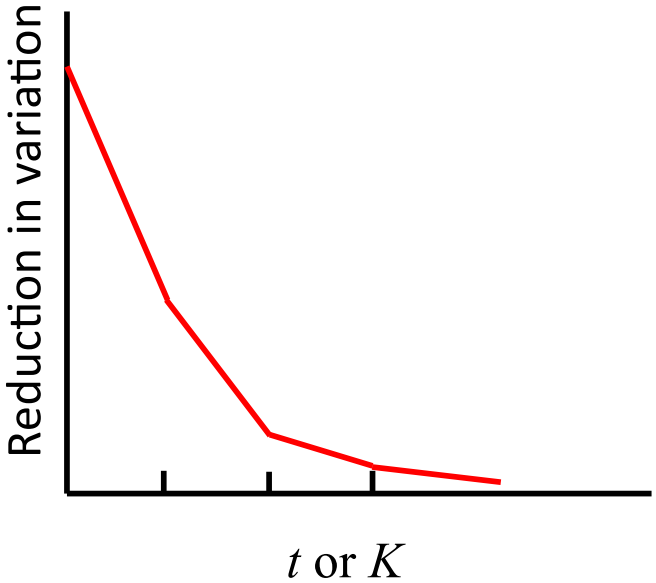
As well as the version available under the anaconda distribution, we also provide a version that has been optimized to run very efficiently on Intel architectures. Please see **Using TensorFlow at NERSC** for more details.

Some basics of ML techniques

K-Means Clustering

(movie by R. Churchill, PPPL)





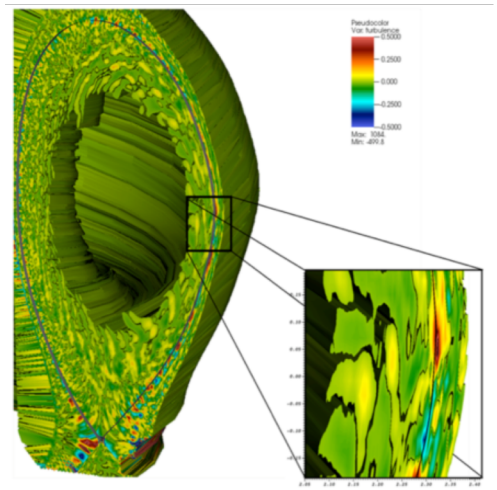
$\mathbf{m} = (m_1, m_2, \ldots m_K)$: mean distance vector

If R [= $\|m(t) - m(t-1)\|$] < ζ
Then, use $\mathbf{m}(t)$ as solution

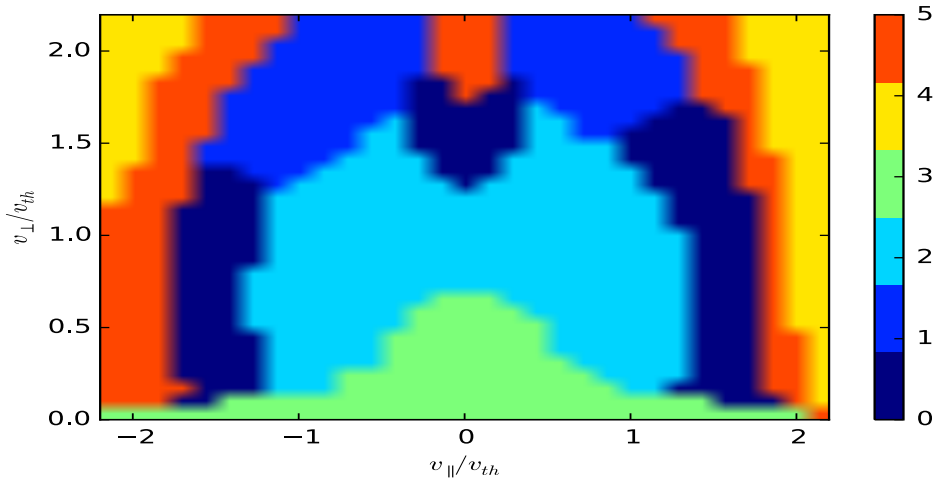
Example for the slow lane usage: K-Means Clustering reveals interesting v-space structure of electron dynamics in blobby edge turbulence

(R. Churchill, PPPL,)

Apache Spark is used here to generate the example physics shown here.

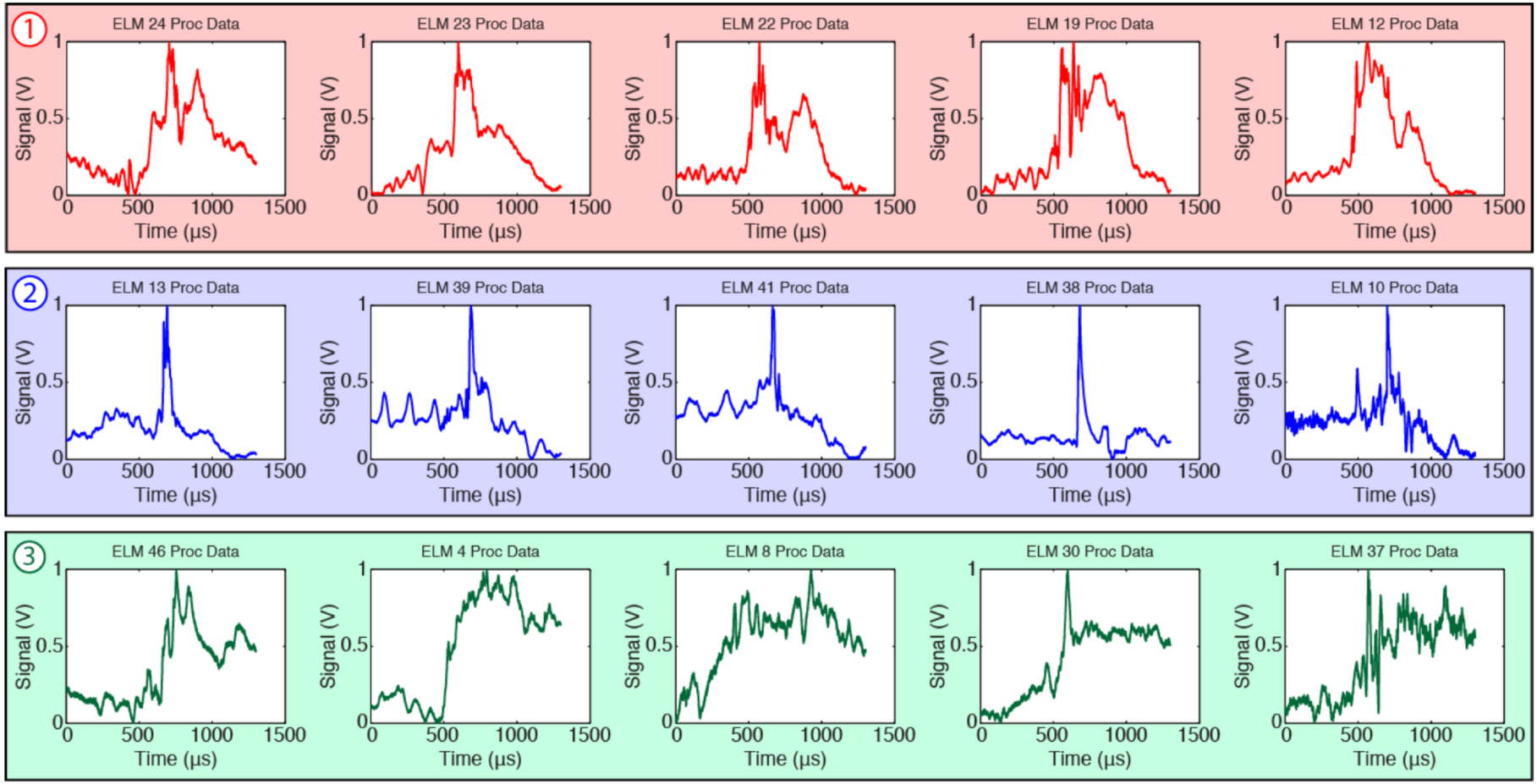


- K=5
- At a higher energy band relative to local T_e, trapped particle's response to blobs is mutually correlated and decorrelated from passing particles
  - A sign of TEM driven turbulence



ELM evolution patterns identified with machine learning techniques

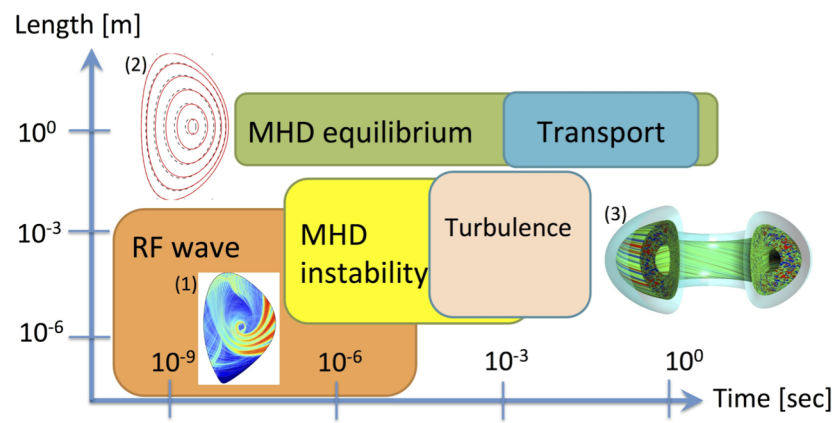
Utilized clustering methods, D. Smith (U. Wisconsin, PPPL)



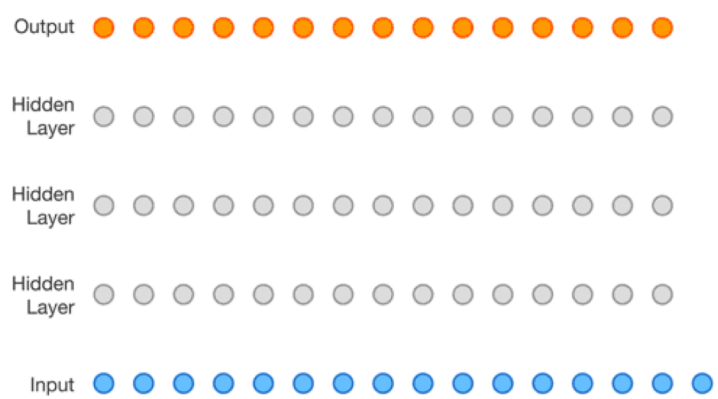
Next step: automate discovery and tagging of ELMs in the NSTX/NSTX-U data archive

Deep convolutional neural networks for long-time series analysis (Churchill, PPPL)

- Fusion plasmas exhibit a range of timescales and physics
- Fusion experimental diagnostics are disparate, and highly time resolved
- PPPL is developing deep convolutional neural networks to automatically analyze and identify physics of interest
- This will allow end-to-end training on raw data from multiple sources, in an efficient manner



$$\sigma\left(\begin{bmatrix} w_{0,0} & w_{0,1} & \cdots & w_{0,n} \\ w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{k,0} & w_{k,1} & \cdots & w_{k,n} \end{bmatrix} \begin{bmatrix} a_0^{(0)} \\ a_1^{(0)} \\ \vdots \\ a_n^{(0)} \end{bmatrix} + \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix}\right)$$



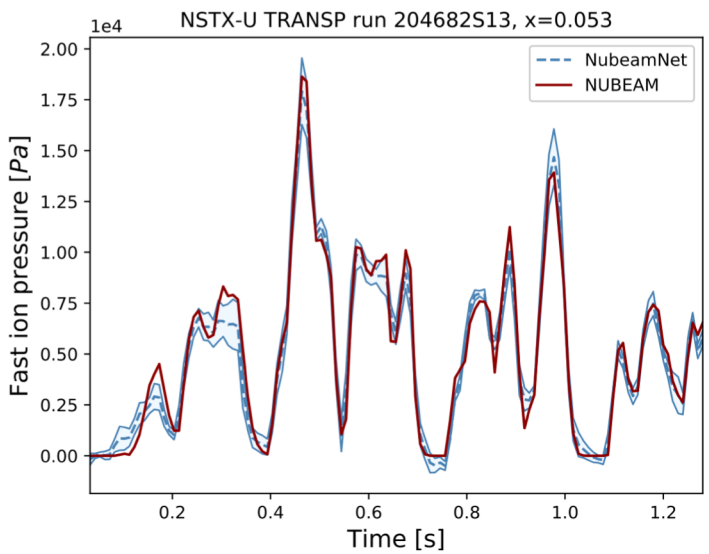
$$\sigma(x) = \frac{1}{1+e^{-x}}$$ Sigmoid function

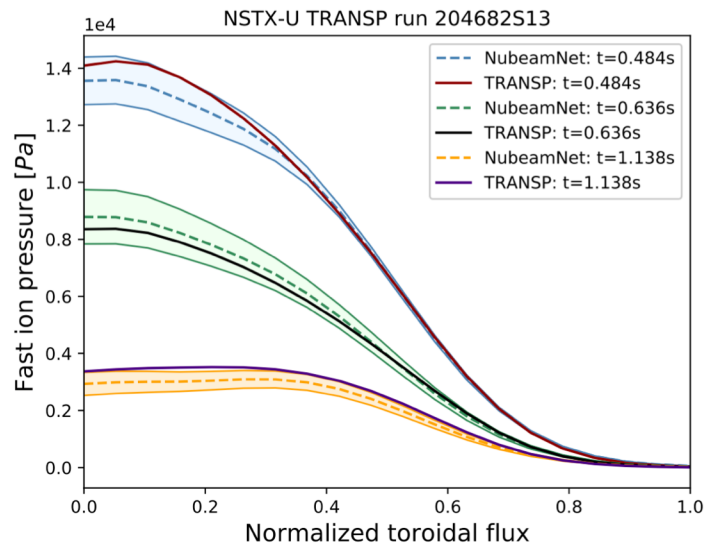
Example for the fast lane usage using neural network, for near real-time NUBEAM simulation in TRANSP

Dan Boyer (PPPL)

- Predictive TRANSP simulations can take hours per simulation second
- **NUBEAM** is a Monte Carlo code that calculates the effect of neutral beams on the plasma (heating, current drive, torque)
  - Often takes >70% of TRANSP calculation time for high quality NUBEAM
- Basic machine learning approaches enable the development of **NubeamNet**

Calculation of beam effects (at 5ms intervals) took **less than 50ms for the entire shot**, compared to hours for high quality NUBEAM





Mathematical Tools for Lossy & Faithful Data Reduction

Mark Ainsworth Brown U. & ORNL

NIMROD – MGARD Reduction (s=1)

Magnetic island detection

Trajectories depend on quantity

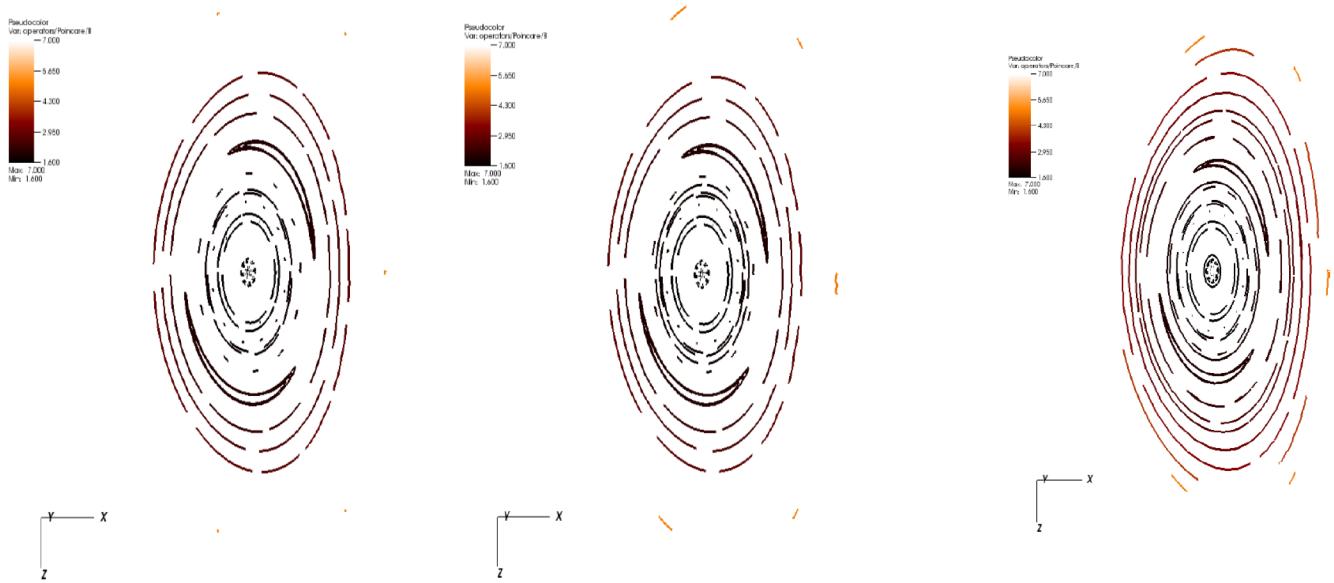
$$\mathcal{Q}_k(v) = \int_{t_k}^{t_{k+1}} v(\tilde{x}(t))\,\mathrm{d}t$$

Continuous linear functional

$$|\mathcal{Q}(u-\tilde{u})| \le \Upsilon_s(\mathcal{Q})\|u-\tilde{u}\|_s$$

for s = (d-1)/2 in d-dimensions

Here d=3, so take s=1

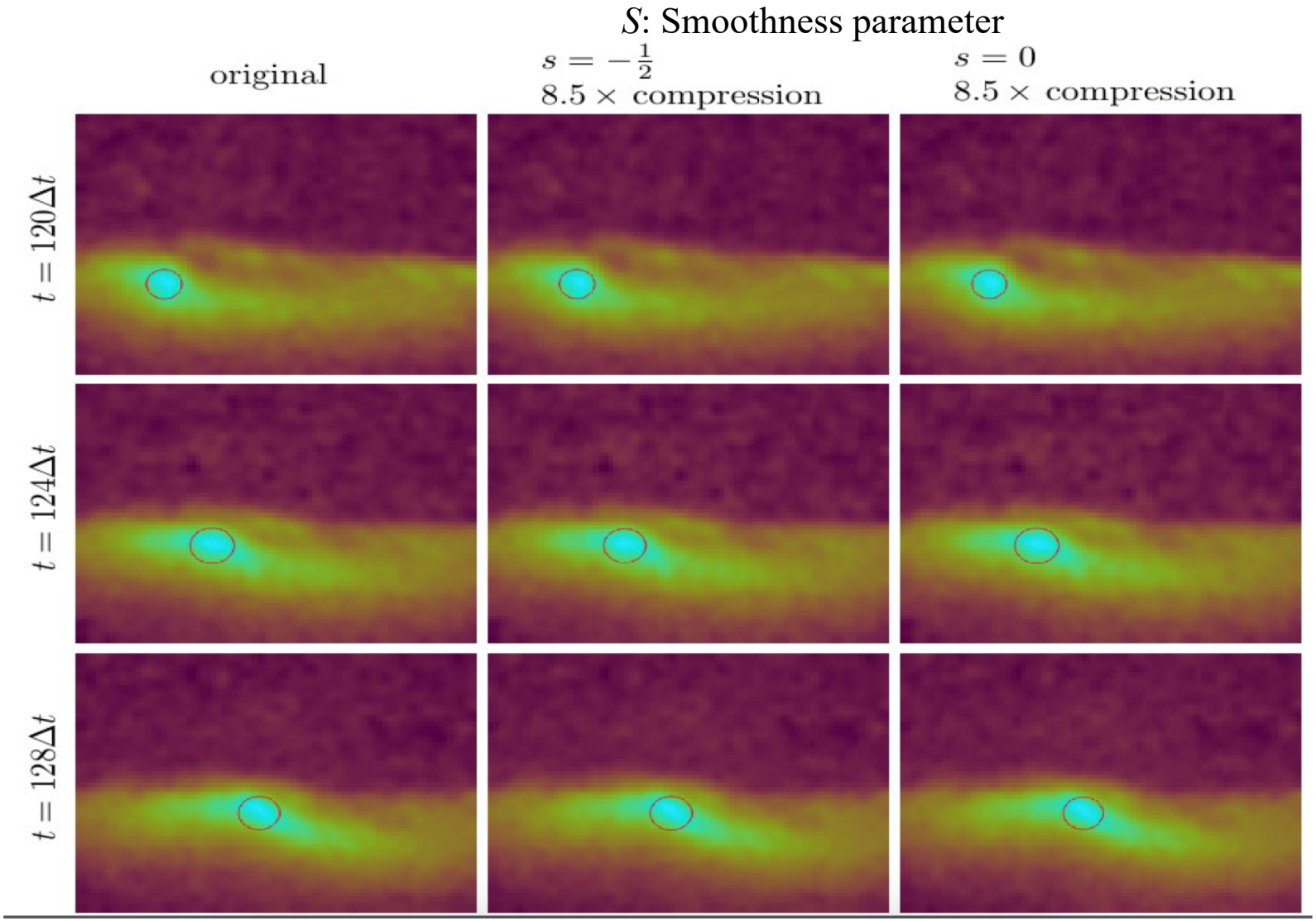


True | MGARD (s=1,10X) | MGARD (s=1,100X)

S: Smoothness parameter

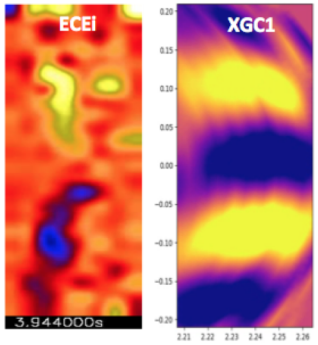
The same data reduction technique can be applied to ELM and turblent blobs

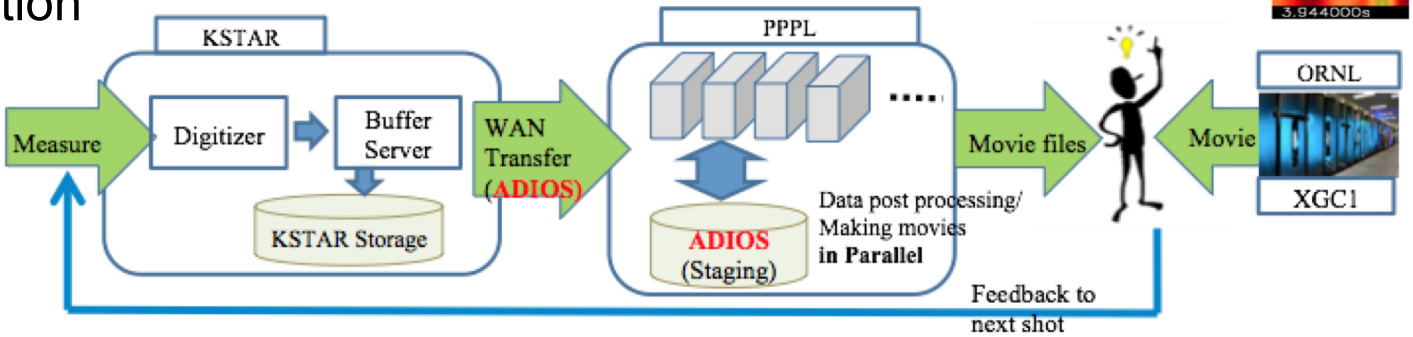
Mark Ainsworth
Brown U. &
ORNL



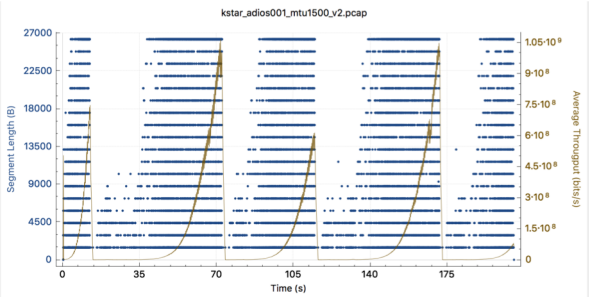
Where are we in the process?

- Remote streaming, with data reduction, implemented in data I/O framework ADIOS [1]
- KSTAR - PPPL streaming demo on 8/2017 of near real-time processing of ECEi data, with comparison to XGC simulation



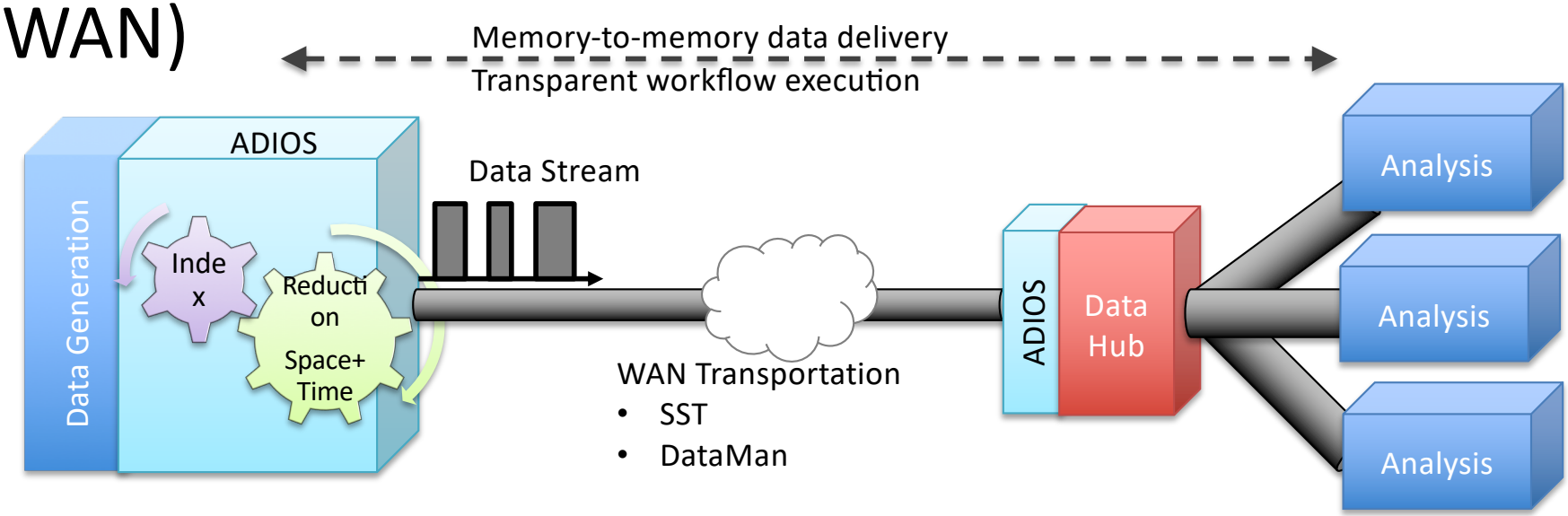


- Demo used special one-time PPPL network configuration (a la science DMZ), which gave ~7Gbps > ITER-Japan test speed
  - Normal PPPL firewall setup gives ~0.2Gbps
  - Ongoing discussion if dedicated science DMZ can be established at PPPL
- Initial KSTAR-NERSC test performed in 10/2018



[1] Choi *IEEE Tran. NYSDS 2016*

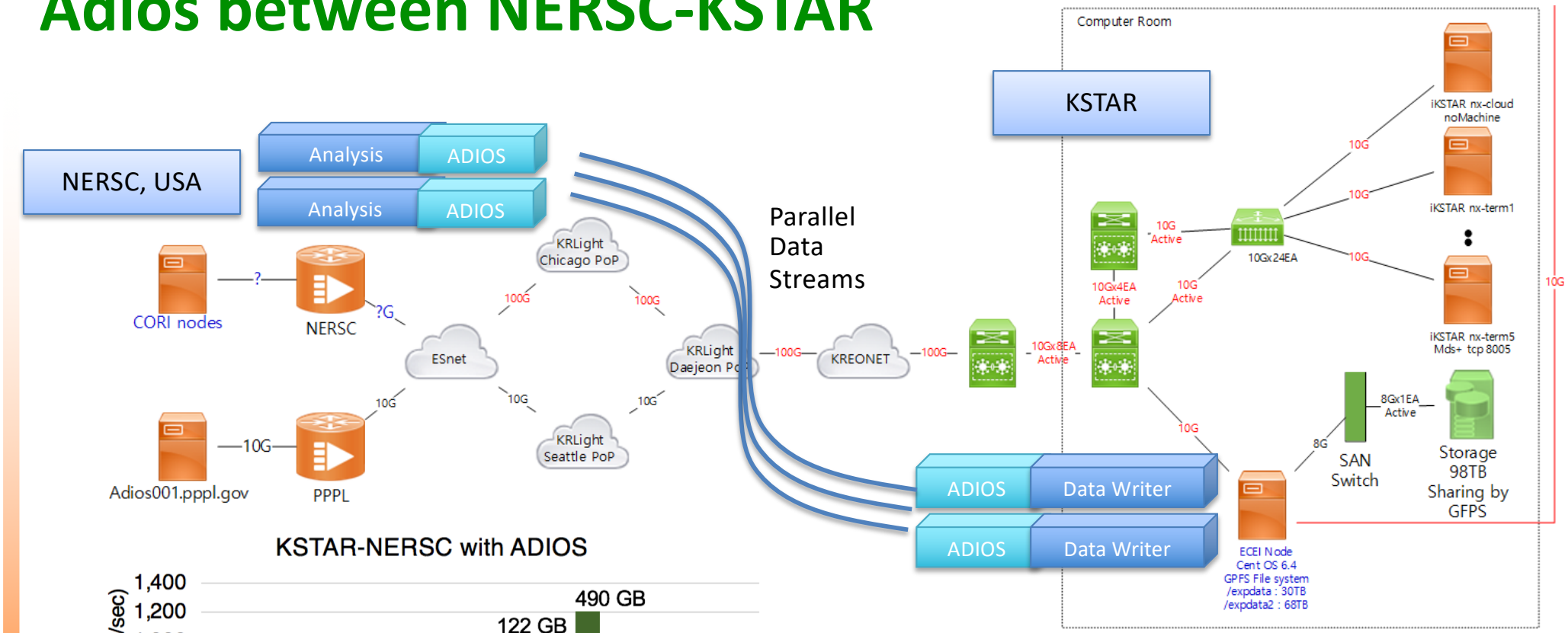
Adios Data Streams Over Wide-Area-Network (WAN)

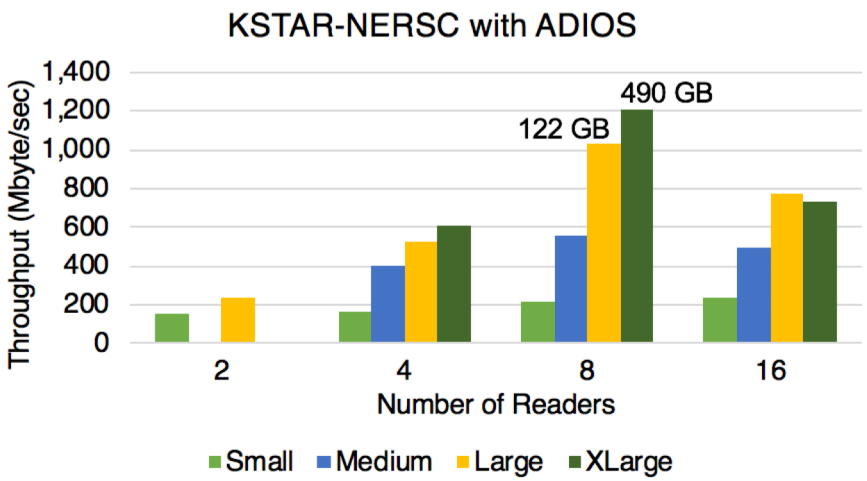


Research on stream-based WAN data process

- In-transit processing (supporting data-in-memory)
- Data indexing & reduction to reduce network payload
- Plug-in methods: SST (EVPath-based), DataMan (ZMQ-based control)

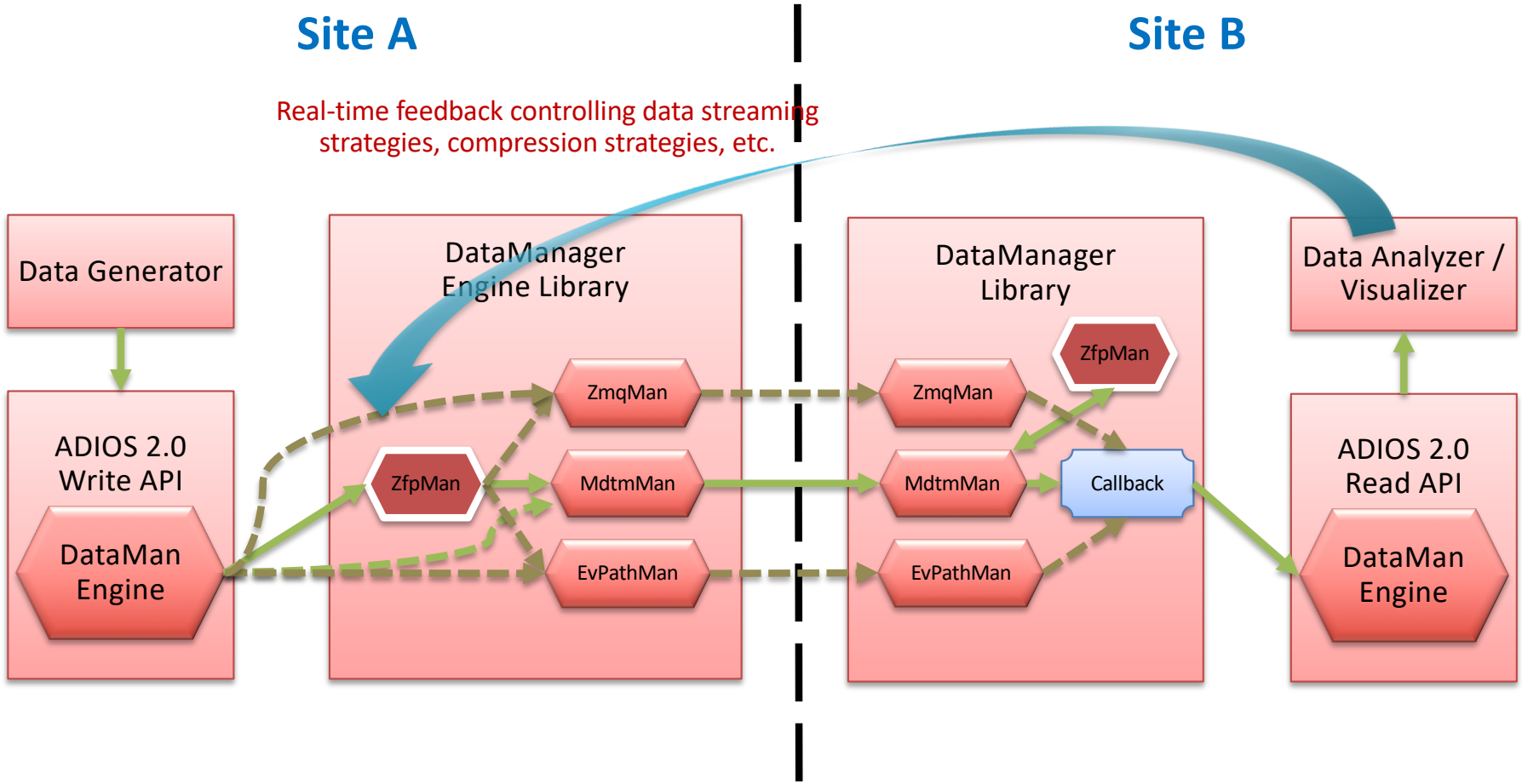
Adios between NERSC-KSTAR





- Measured in September 2018
- Between KSTAR and NERSC
- Run a simulation code with writers and readers
- 8 writers at KSTAR side
- 2-16 readers on a NERSC DTN node

Software Infrastructure to Enable the Workflow



Short-Term Plans (to be discussed with KSTAR researchers)

Low hanging fruits

- Off-line data reduction and ML-model training on KSTAR-ECEI data for ELM physics, based on the existing post-processing routines (POSTECH, UNIST)
  - ELM precursor/onset, interaction with turbulence and RMPs
- On-line near real-time data reduction & ML analysis of the KSTAR ECEI ELM
  - Quick feed back to next-day experiments

Extension

- Initiate simulation and provide real-time physics results on federated dashboard

Further extension

- Federated decision using ML on the experimentally observed plasma profiles, with EFIT reconstruction on multiple time slices (highly parallel): Sabbagh
- Near real-time analysis of NTM onset and growth → feedback to experiment
  - Accelerated research on NTM control
- Real-time observation/control of disruption precursor events: Sabbagh
- Near real-time analysis of the divertor heat-load footprint → feedback
  - Accelerated research on divertor heat-load width control