Accelerating the experimental feedback loop: Data streams and the Advanced Photon Source

Ian Foster

Argonne National Lab & University of Chicago

Some of the many people involved

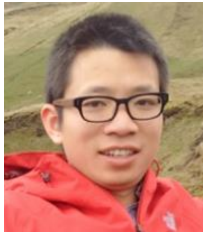

Rachana Ananthakrishnan


Tekin Bicer


Ben Blaiszik


Ryan Chard


Raj Kettimuthu


Zhengchun Liu


Sam Nickolay


Steve Tuecke


Dula Parkinson


Doga Gursoy


Francesco de Carlo


Todd Munson


Sven Leyffer


Jon Almer


Brian Toby


Justin Wozniak

Argonne National Laboratory, Lawrence Berkeley National Laboratory, and other institutions

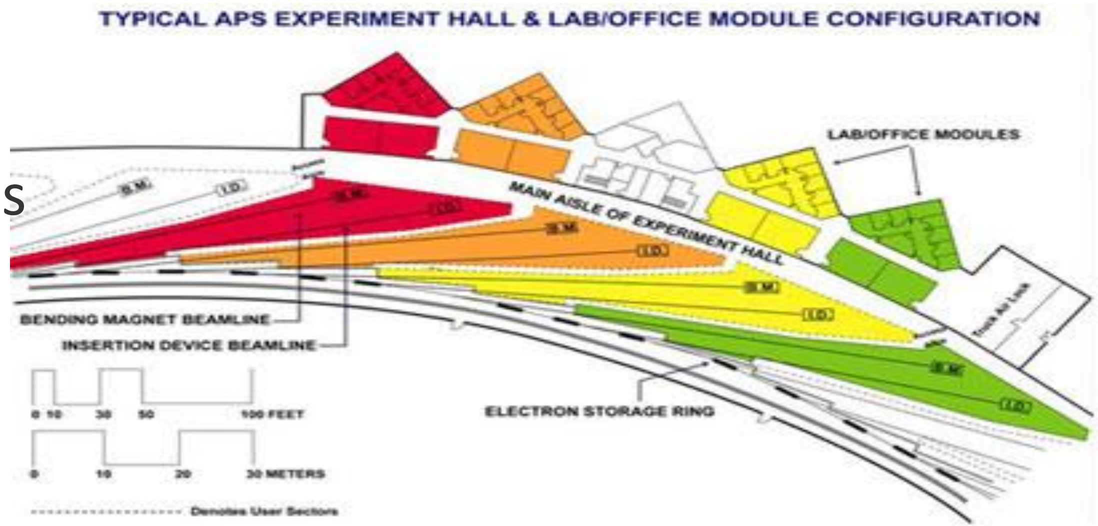

APS
1 km
5 μsec
ALCF
Madison
Milwaukee
Janesville
Kenosha
Rockford
Chicago
Argonne National Laboratory

Overview

- **Context**: What is a light source, why are the experimental feedback loop and data streaming important
- **Tomography**: Experimental data feedback loop in practice
- **Optimizing**: Modeling, analysis, and implementation methods to understand and improve performance
- **Automation**: Further steps towards accelerating end-to-end experimental data lifecycles
- **Publishing**: Collecting and organizing light source data
- **Futures**: Some of the many other things that need to be done

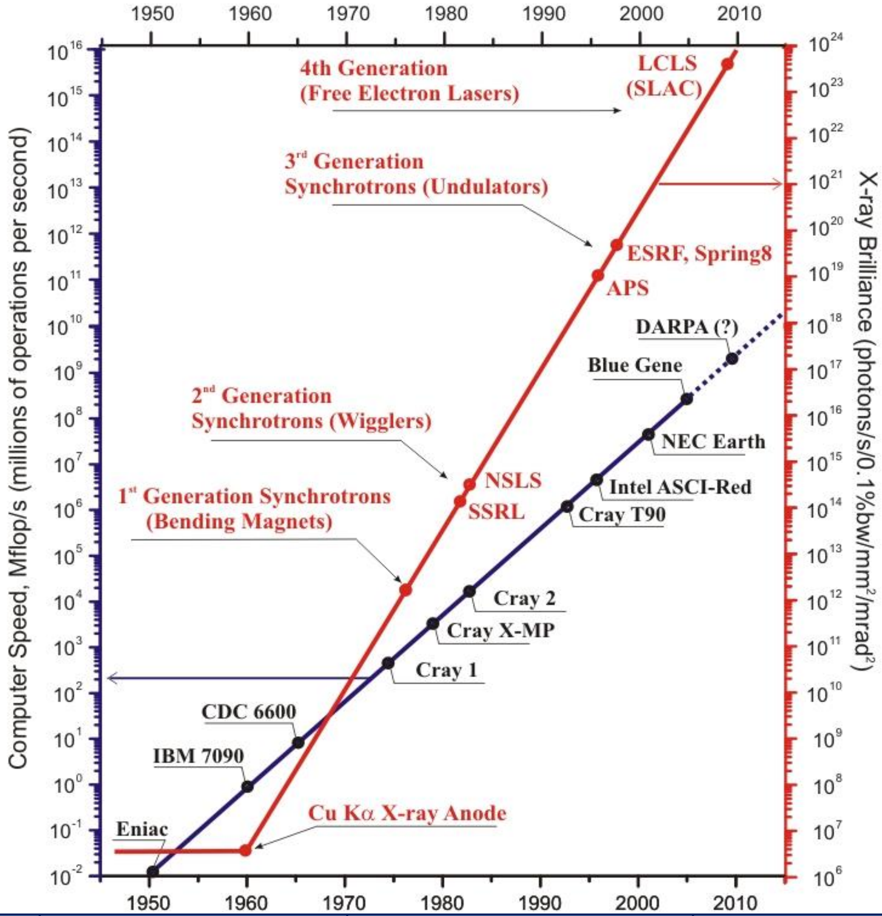
APS is one of four DOE synchrotron light source

- Moves electrons at >99.999999% of the speed of light
- Magnets bend electron trajectories, producing x-rays, highly focused onto a small area
- X-rays strike targets in 35 different sectors, with 70 beamlines
- Different types of optics and detectors → wide range of imaging modalities
- 2014: 22,000 visits, 5,000 unique users, 5,700 experiments



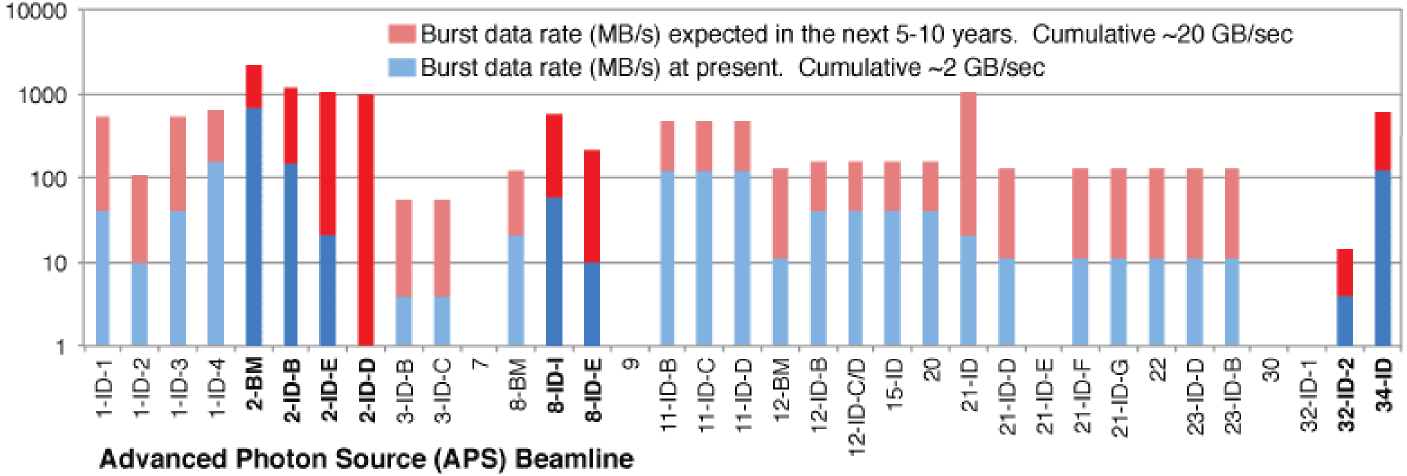
X-ray sources produce a lot of photons, which translates to a lot of data

Computer speed:
12 orders
of magnitude
in 6 decades



X-ray source brilliance:
18 orders
of magnitude
in 5 decades!

Light source data rates are growing dramatically



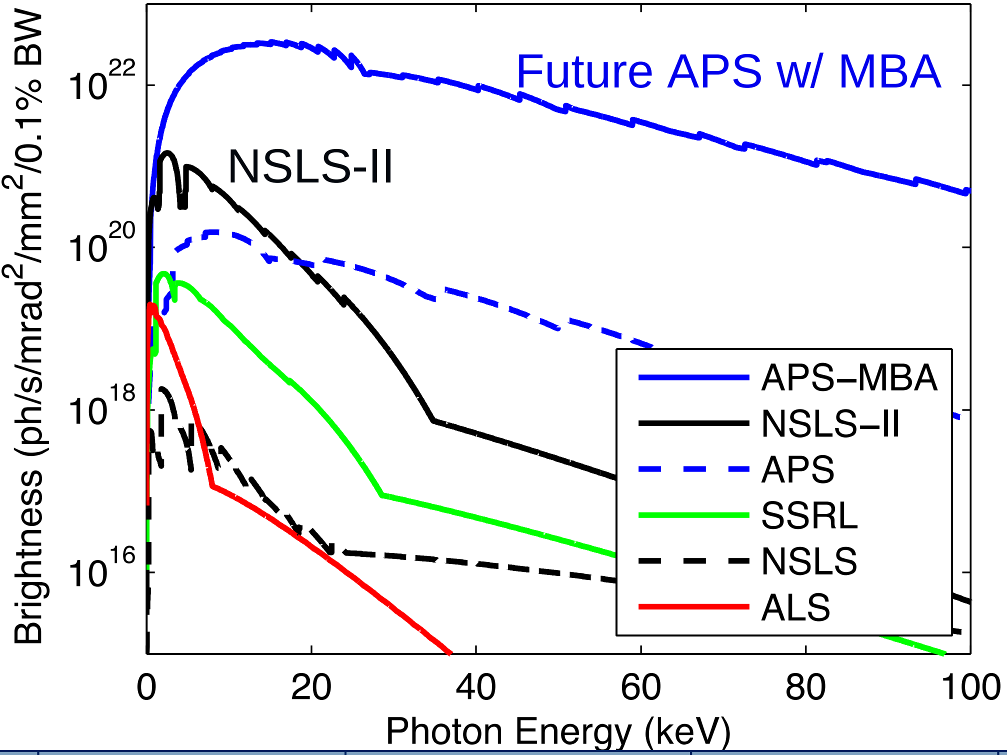
Source: Francesco de Carlo (Date: 2014)

Parameter	LCLS-I	LCLS-II 2020	LCLS-II 2025
Average throughput	0.1 - 1 GB/s	2 - 20 GB/s	2 GB/s - 1.2 TB/s
Peak throughput	5 GB/s	100 GB/s	4.8 TB/s
Disk storage	5 PB	100 PB	6 EB
Peak Processing	50 TFLOPS	1 PFLOPS	60 PFLOPS

Source: Amedeo Perazzo

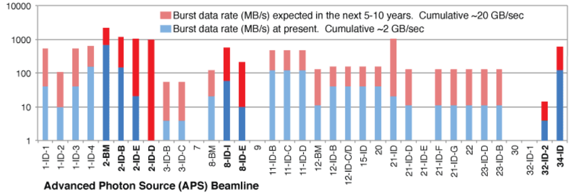
APS upgrade (APS-U): multi-bend achromat (MBA) lattice will yield unprecedented brightness and coherence up to high energies

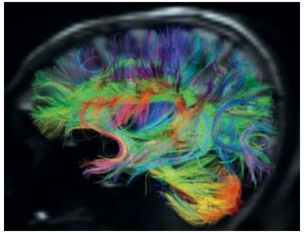
Brightness vs. x-ray energy at top beamlines among DOE synchrotron facilities

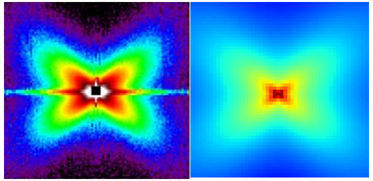


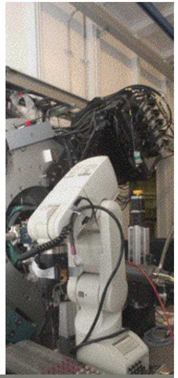
Major data and computation challenges arise across APS; Explode with APS-U

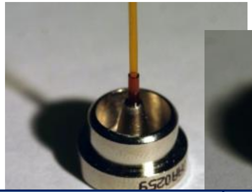
- **Huge data** from new detectors and from APS-U
  - E.g., XPCS: <u>Today</u>: 2MB images @ 100 Hz; <u>Soon</u>: 1MB images @ 2000 Hz (x 10!); <u>Eiger</u>: 2Mbyte @ 3000 Hz (x 3!); APS-U another 2-3 orders of magn.
- **Complex, multi-modal data** needs advanced computation for interpretation
  - E.g., Ptychography+elemental mapping+visual images as a function of reaction conditions
- Advanced modeling and theory enable **fitting and co-optimization** of model and experiment
  - Goal: Fit one model to all measurements
- New user demographics → **automation**
  - Scale to more and different users, many with limited/no experience
- New usage modalities requiring **computer-in-the-loop control**
  - E.g., detect errors or interesting features in data as they are collected





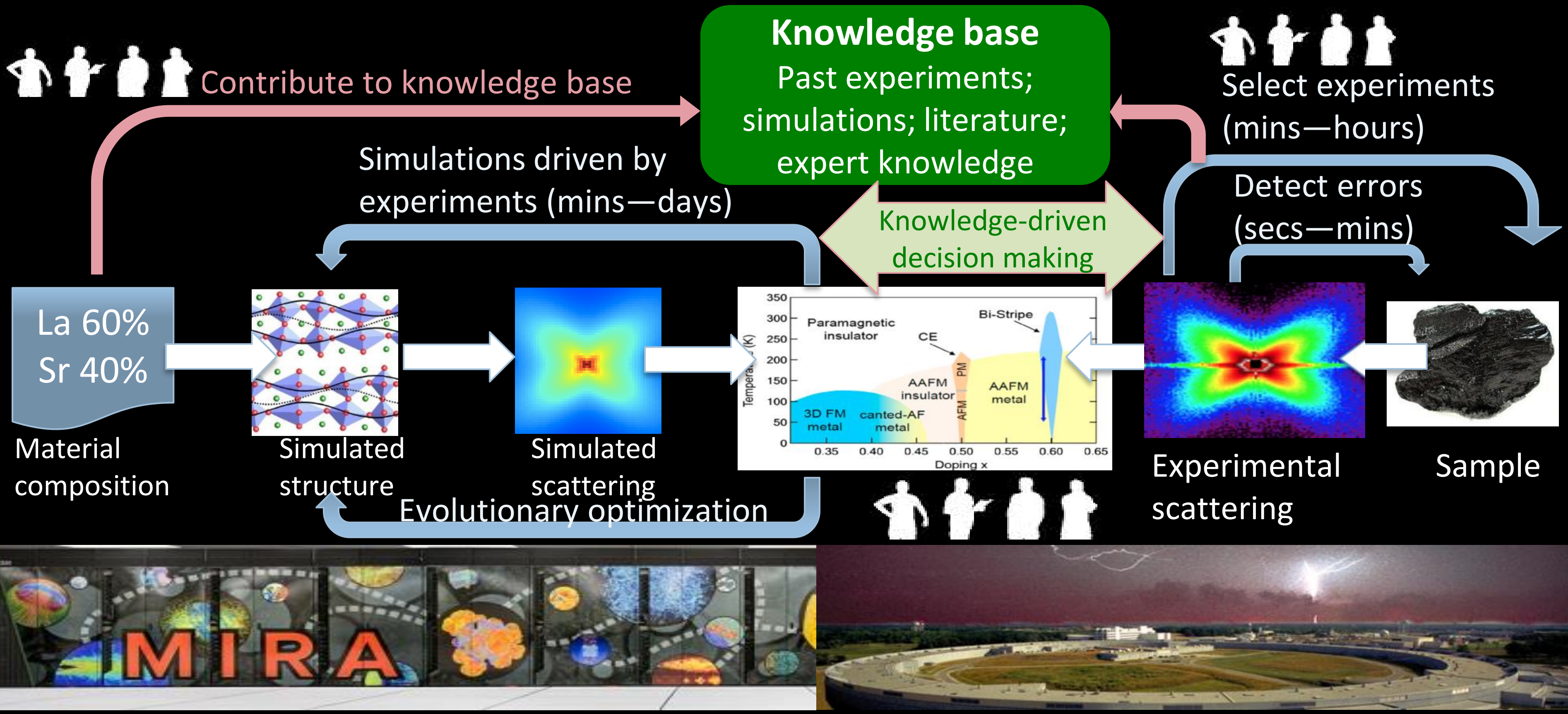






Context | Tomography | Optimizing | Automation | Publishing | Futures

A discovery engine for the study of materials



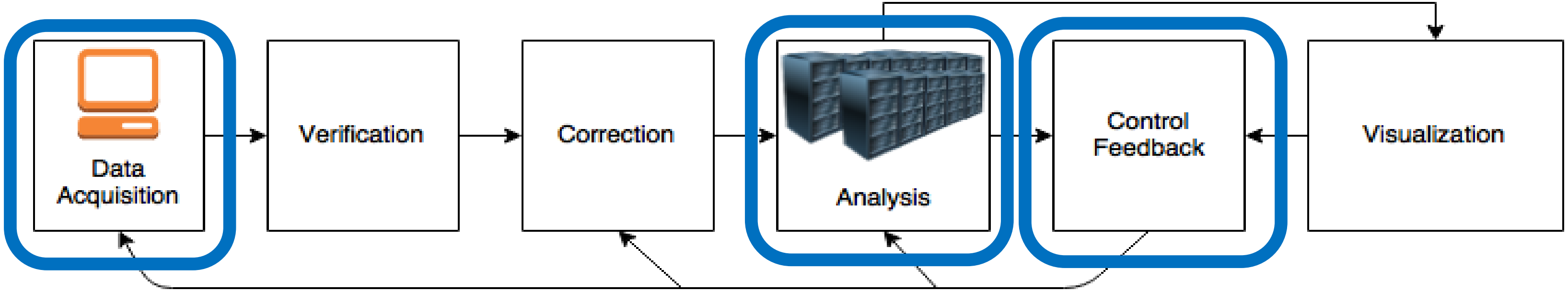
Diffuse scattering images from Ray Osborn et al., Argonne

Experimental steering using HPC


Tekin Bicer


Doga Gursoy


Raj Kettimuthu

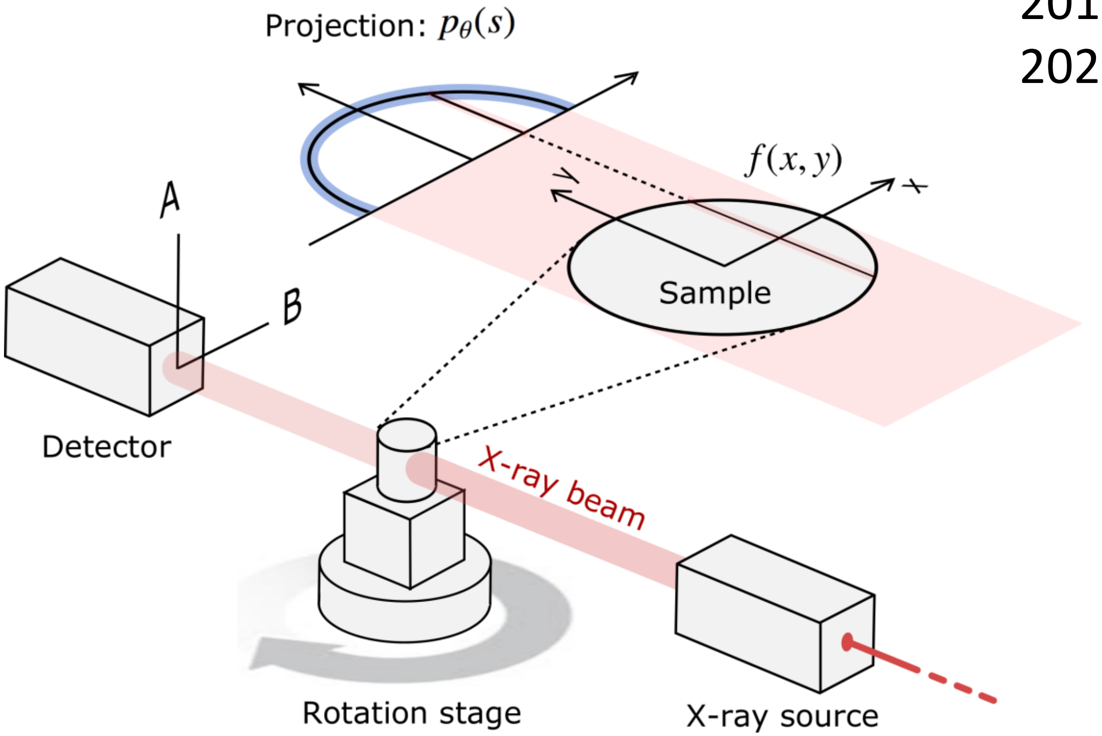


- "Real-time" analysis of streaming experimental data
  - Enables *smart experimentation*
  - Requires HPC resources
- Examples
  - Detect features in hierarchical structures
  - Change data acquisition for dynamic systems
  - Minimize damage to dose-sensitive specimens
  - Adjust experimental parameters on the fly
  - Detect errors early in experiments

Use case: Acquire only enough data to meet quality goals

- Adaptive data acquisition
- Incremental reconstruction
- Image quality check (MS-SSIM similarity scores)
- Finalize data acquisition based on image quality

Example: Computing microtomography (CMT)



2017: 4K×4K×1500×12 bits/s = 36 GB/s
2025: 20K×20K×8000×20 bits/s = 8 TB/s

Acquisition

2017: 10 Gbps
2022: 1 Tbps

Reconstruction

Smart online data acquisition strategies to minimize time to useful information

Naïve: Collect a continuous set of angles

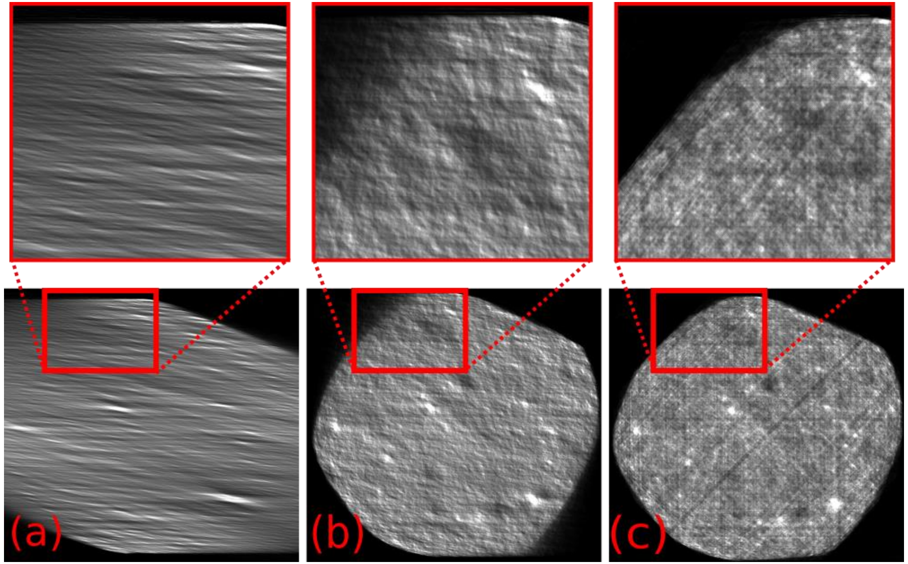
- E.g.: Offset = 1; θs = (0, 1, 2, …, 179) °

Interleaved:

- E.g.: Offset = 5;
  θs = (0, 5, 10, …, 175, 1, 6, …, 174, 179, …)

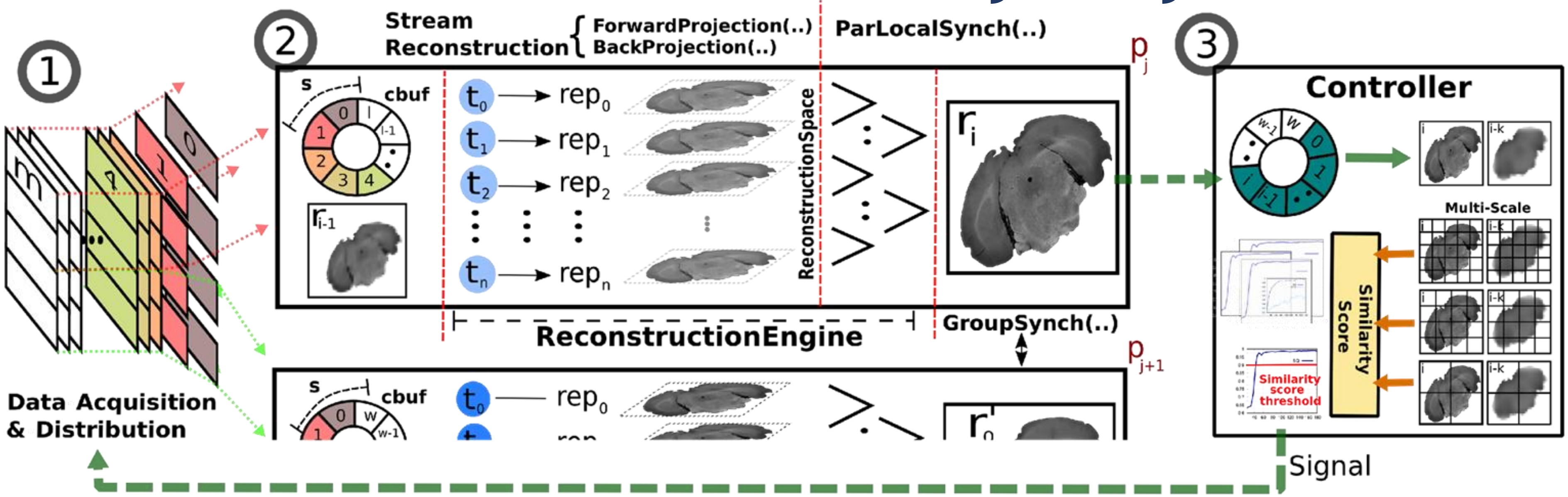
Optimized interleaved: Halve collected projection angles after each round

- E.g.: θs = (0, 90, 45, 135, 22, 67, …, 179)



Reconstructed image of a shale sample with only 30 streamed projections: (a) fixed angle, offset=1°; (b) inter- leaved, offset=5°; (c) optimized interleaved. The range of angles is [0,180)°.

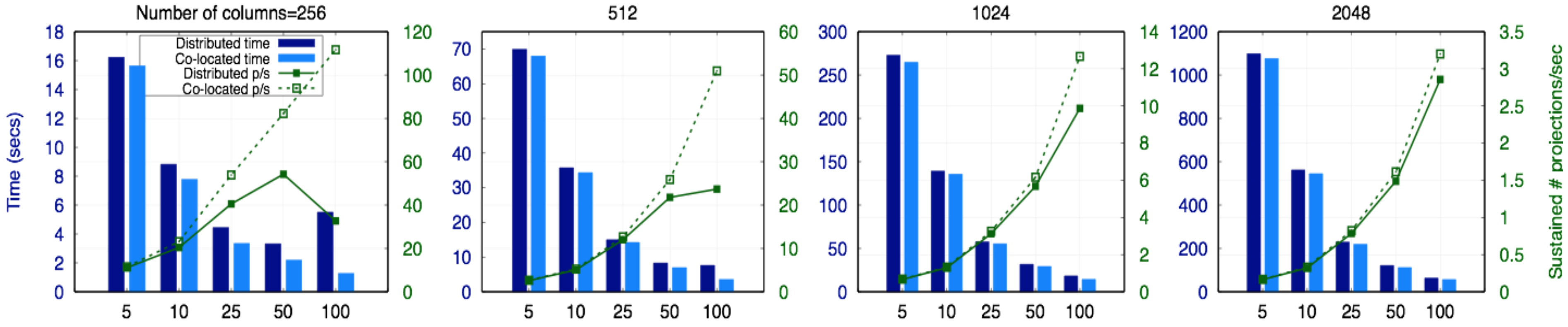
Automated stream analysis system



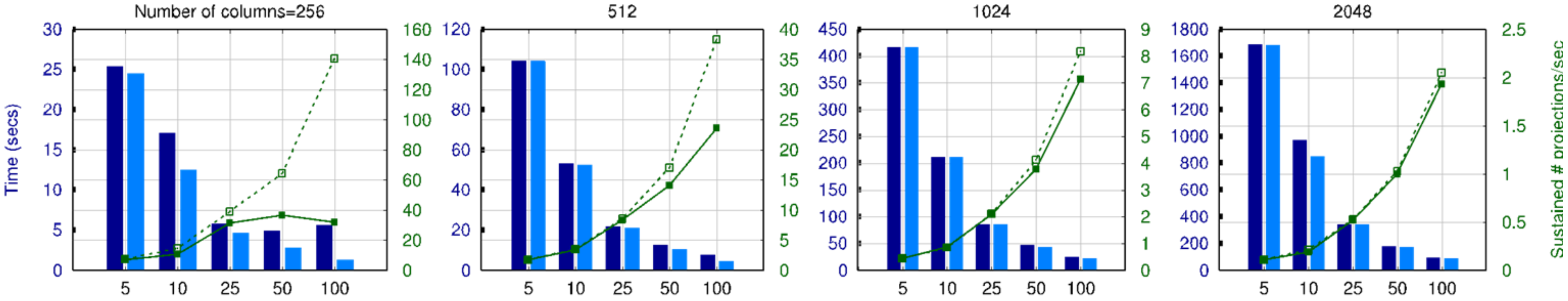
System parameters:

- **Analysis**: window_length, step_size, window_iteration, reconstruction_algorithm
- **Computational resource**: # nodes, # threads
- **Controller**: scale, back-check (i-k), threshold
- **Data acquisition strategy**: naïve, interleaved, optimized interleaved

Stream reconstruction performance on 12-core nodes

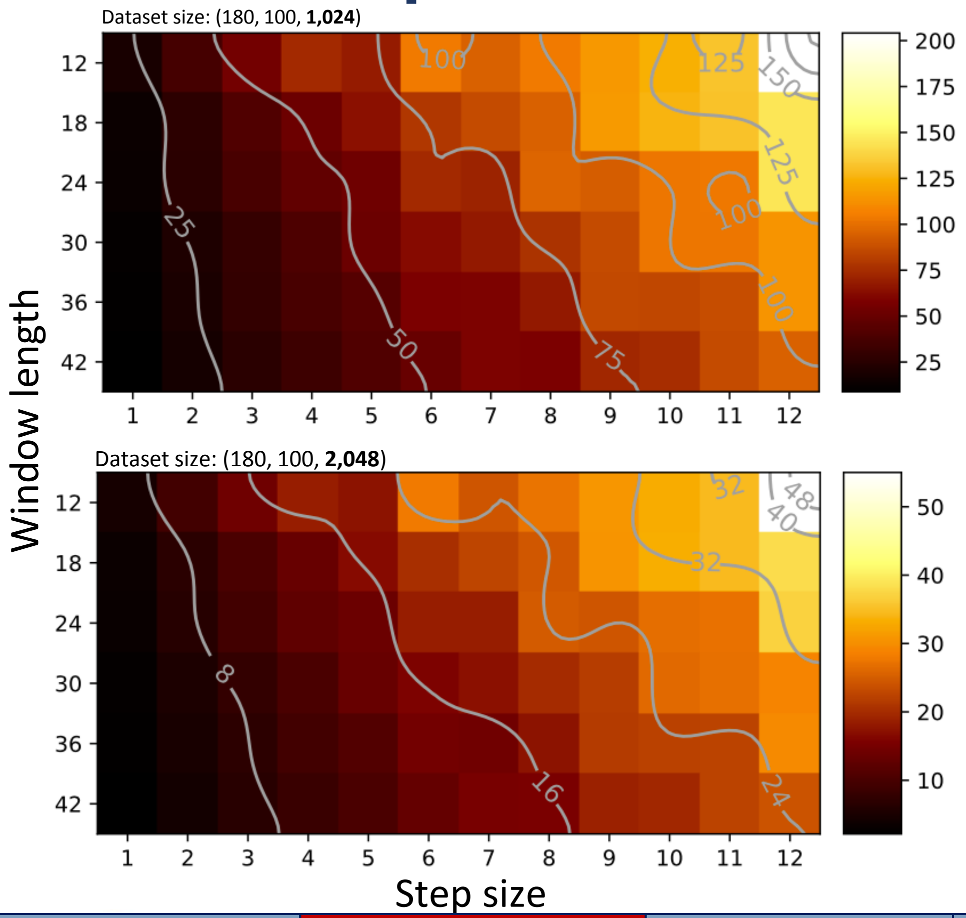


Maximum Likelihood Expectation Maximization (MLEM)



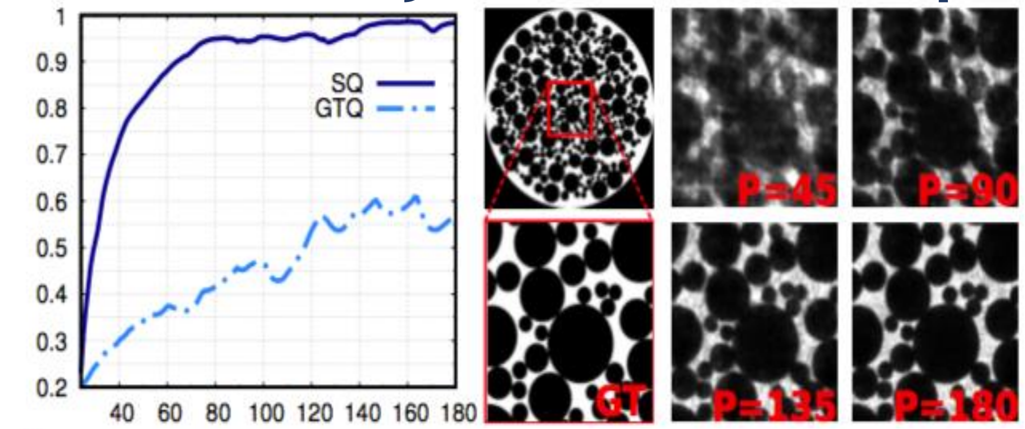
Penalized Maximum Likelihood (PML)

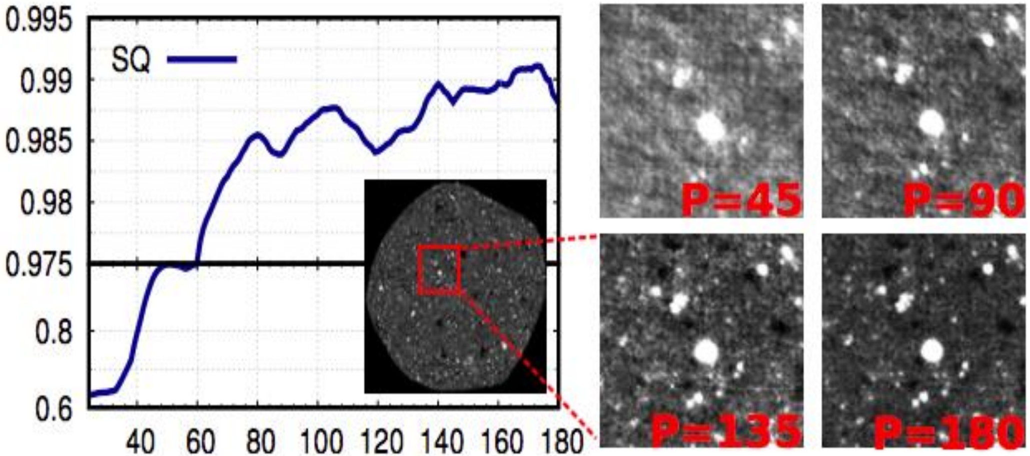
Runtime parameters determine processing rate



- Detectors have different data generation rates
- Runtime parameters can be adjusted to meet data generation rates
- MLEM reconstruction performance w.r.t. different window length and step size values
  - # Nodes = 100 nodes (1,200 cores)
- Color represents the projection consumption rates
  - Max: 204 projections per second (top fig.)
    - Dataset: (180, 100, **1,024**)
  - Max: 55 projections per second (bottom fig.)
    - Dataset: (180, 100, **2,048**)

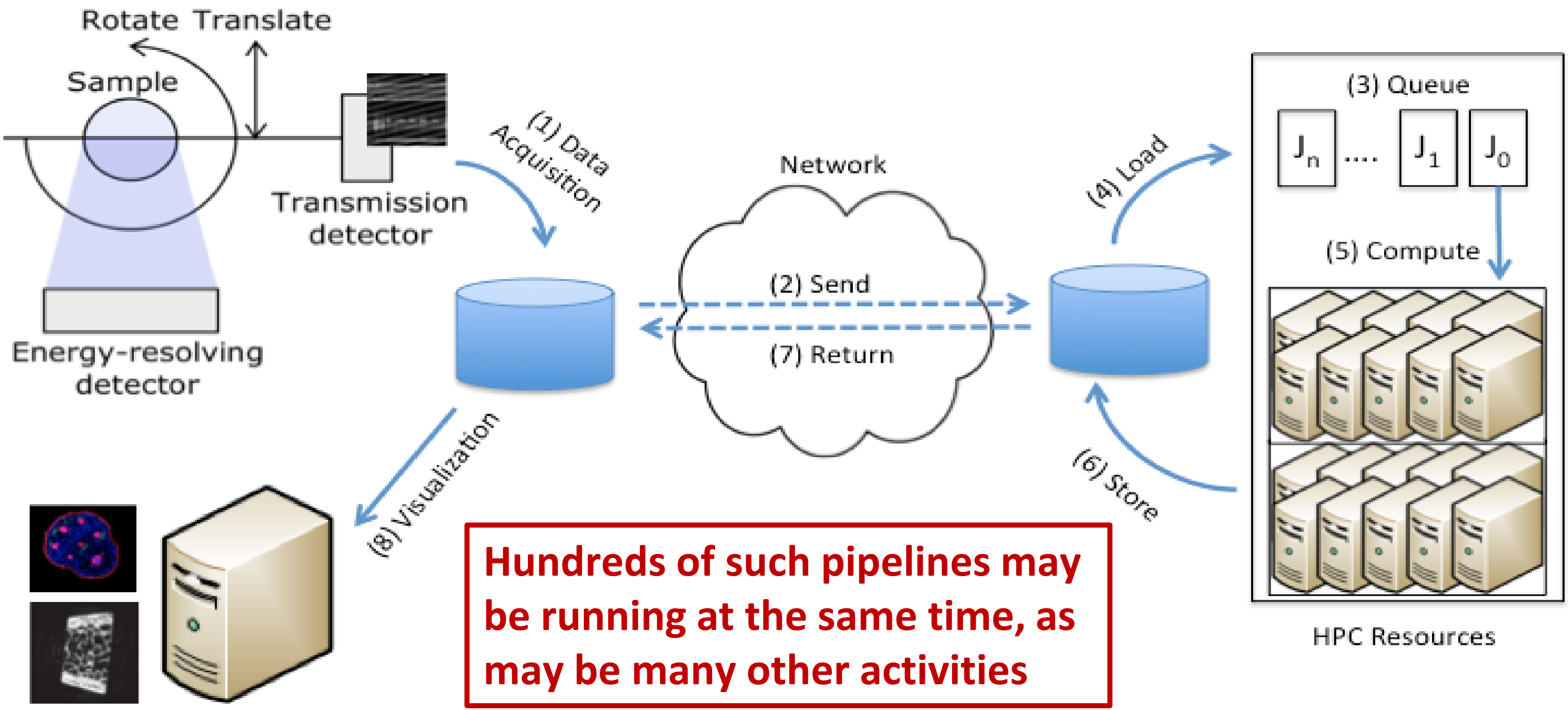
Quality cutoffs for experimental steering





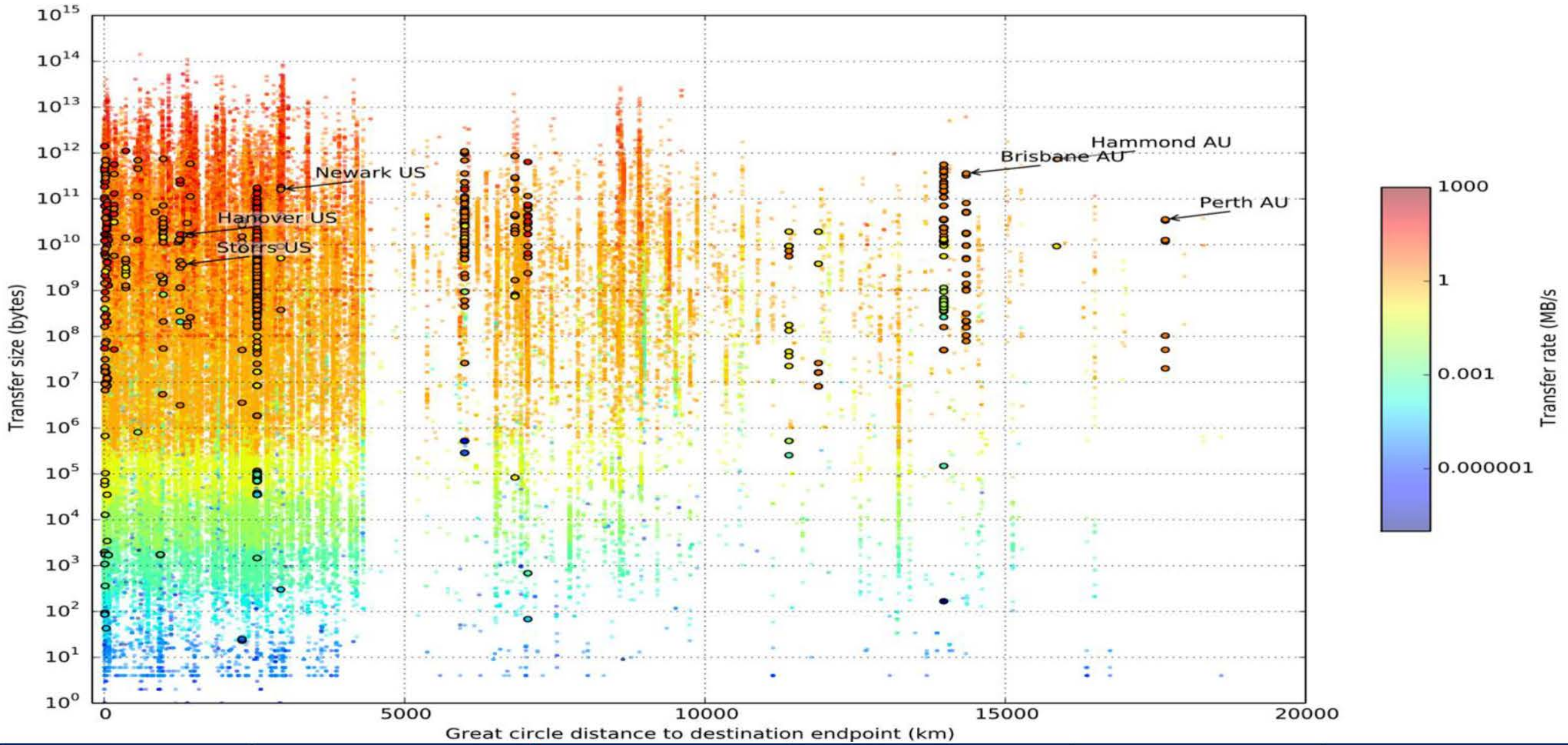
- Quality check between reconstructed images
  - Small changes in similarity score indicate convergence of values
- Real-time stream analysis and steering can minimize data acquisition while meeting data quality constraints
  - 22-44% reduction in # collected projections
  - Less dose effect, shorter data acquisition and analysis, better utilization of instruments …

Understanding and optimizing the end-to-end pipeline



Hundreds of such pipelines may be running at the same time, as may be many other activities

Globus transfers, showing rate (via color) as a function of distance and volume.
The 1921 transfers from aps#clutch are highlighted.



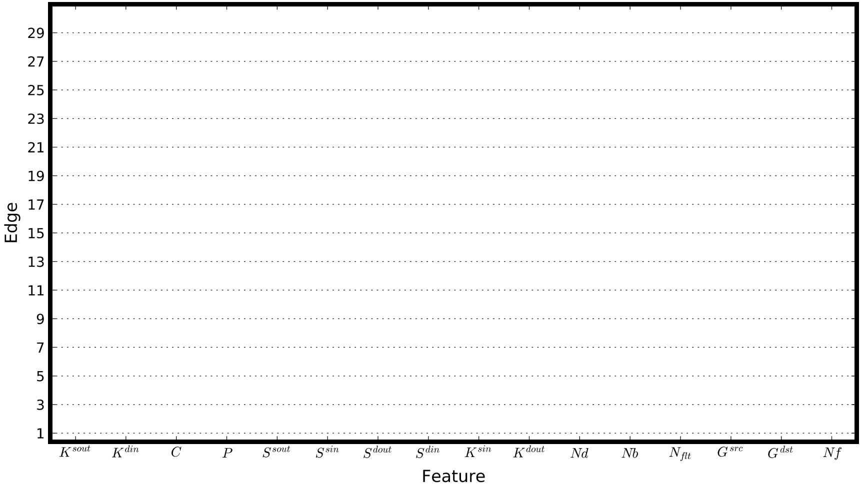
Data-driven models yield new insights into wide area data transfer performance

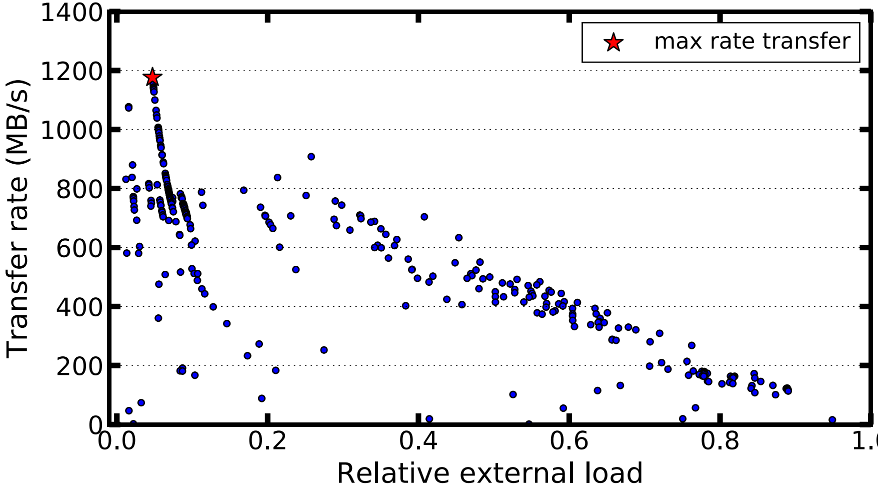
Zhengchun Liu

Prasanna Balaprakash

Raj Kettimuthu

- What factors determine wide area data transfer performance? Can we predict performance? How can we improve performance?
- We use Globus transfer records to develop machine learning models. A model of heavily used network links has median average % error of just 7.8%
- Evidence of the negative impact of high endpoint load is driving new optimizations





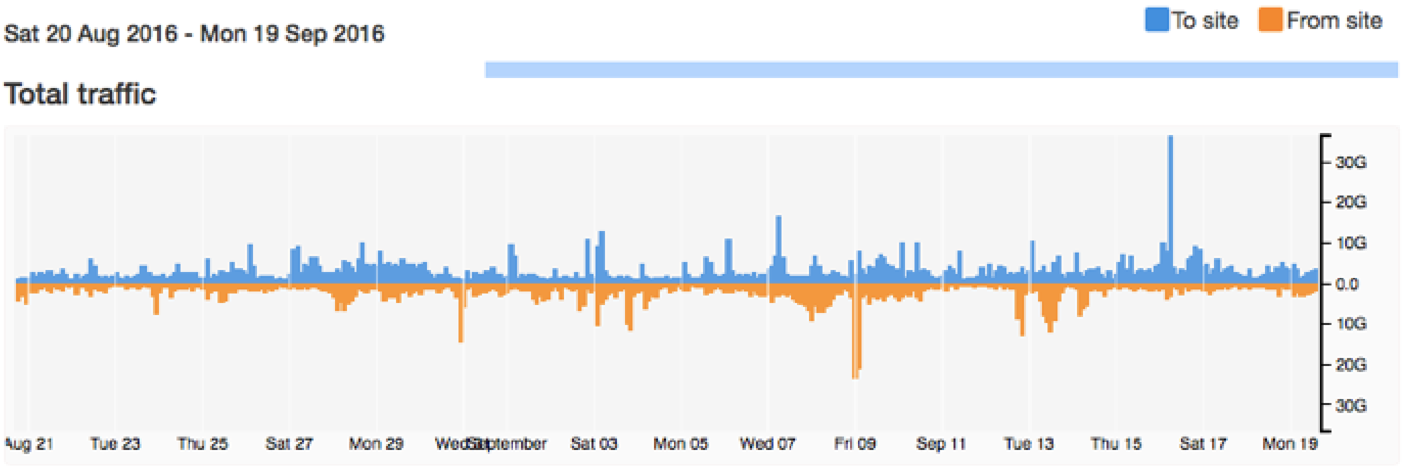
Z. Liu, P. Balaprakash, R. Kettimuthu, I. Foster. Explaining Wide Area Data Transfer Performance, HPDC’2017.

Sam Nickolay

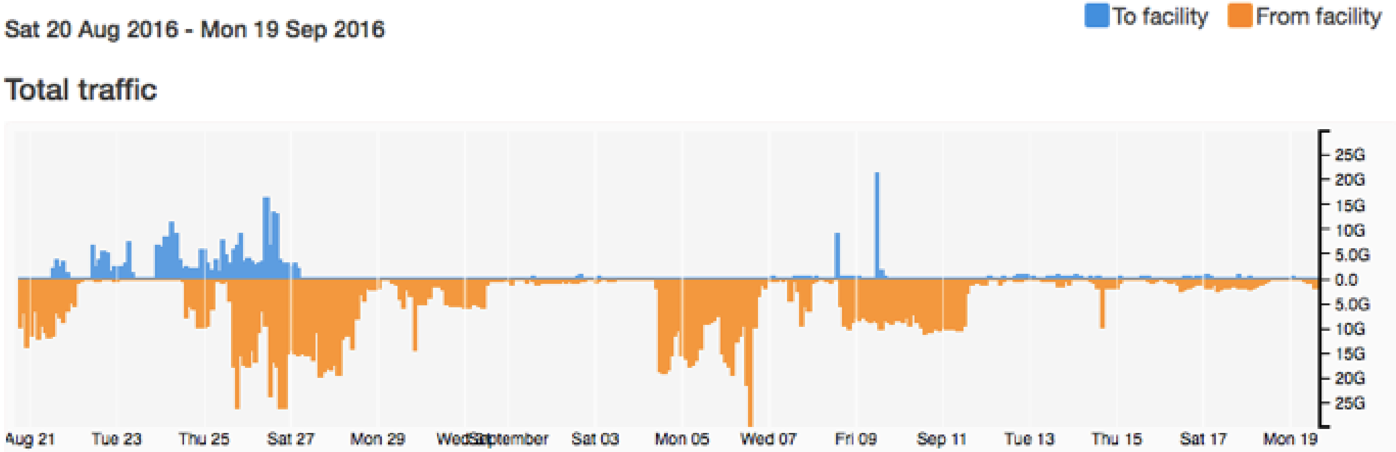
Raj Kettimuthu

Eun-Sung Jung

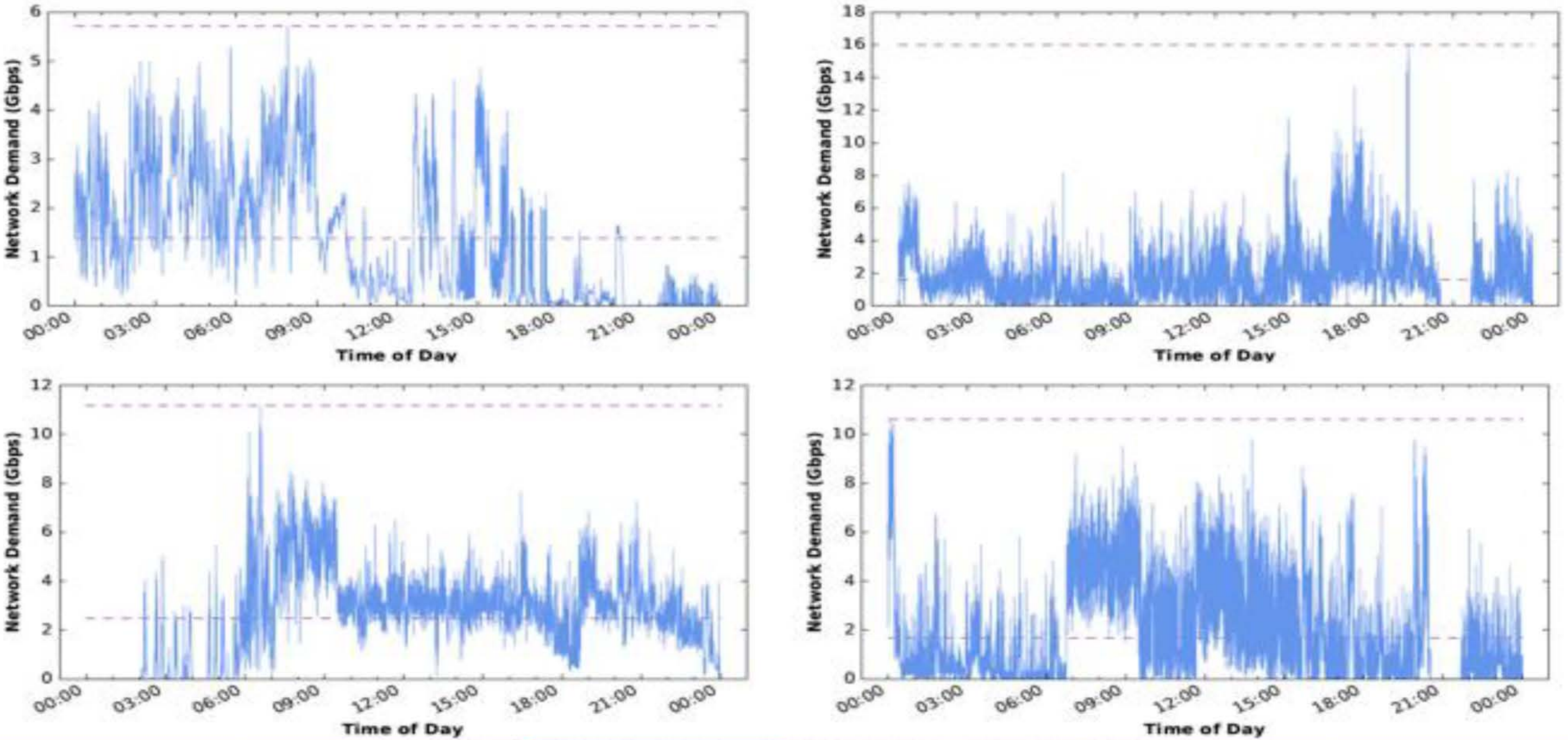
Source: my.es.net



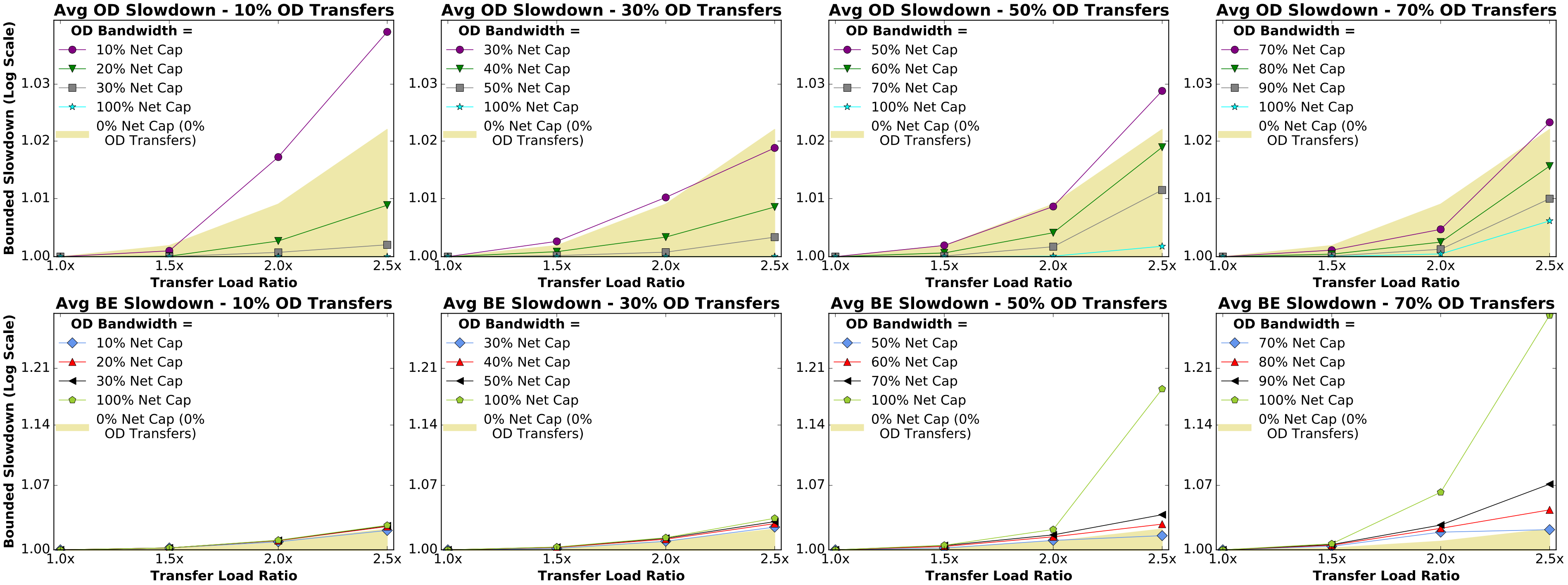
Gap between peak and average network load



GridFTP usage data for top servers



Increase average usage by differentiating traffic



S. Nickolay, E.-S. Jung, R. Kettimuthu, I. Foster, Bridging the Gap between Peak and Average Loads on Science Networks, FGCS 2017.

RIPPLE: A prototype responsive storage solution


Ryan Chard

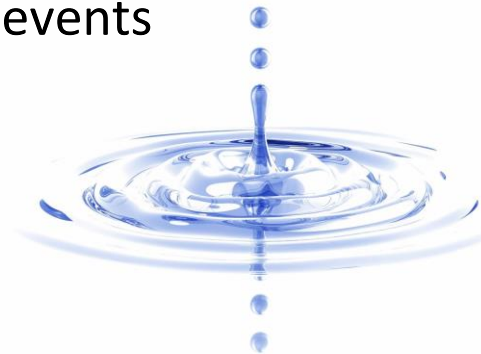

Dula Parkinson


Steve Tuecke

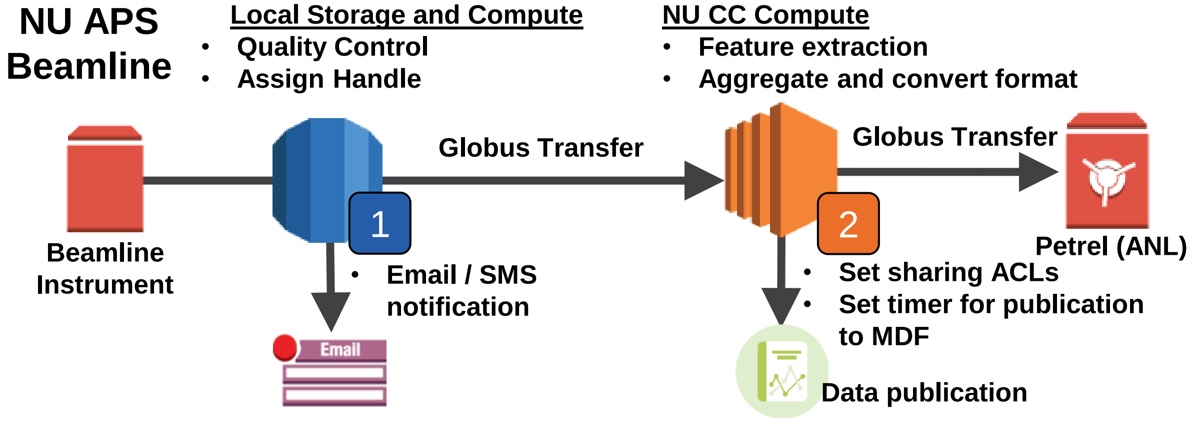

Kyle Chard

Transform static data graveyards into active, responsive storage devices

- Automate data management processes and enforce best practices
- **Event-driven**: actions are performed in response to data events
- Users define simple **if-trigger-then-action recipes**
- Combine recipes into **flows** that control end-to-end data transformations
- Passively wait for filesystem events (little overhead)
- Filesystem agnostic – works on both edge and leadership platforms



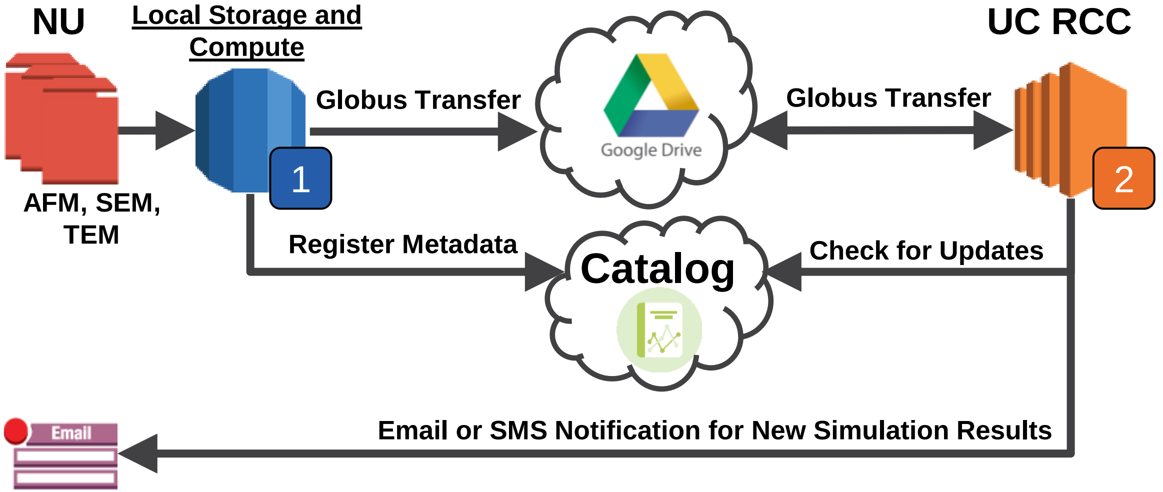
R. Chard, K. Chard, J. Alt, D. Parkinson, S. Tuecke, I. Foster, RIPPLE: Home Automation for Research Data Management, WOSC, 2017.



Rules

1
- IF new files THEN run quality control scripts
- IF quality is good THEN send email and transfer data to NU

2
- IF new files THEN run feature extraction
- IF feature detected THEN transfer data to archival storage (Petrel)
- IF time since ingest > 6 months THEN publish dataset to MDF



Rules

1
- IF new files THEN map elastic modulus
- IF new elastic modulus map THEN register in catalog and extract image metadata and move raw and derived data to shared Google Drive folder

2
- IF new data in Google Drive THEN fetch data from Google Drive and associated metadata in catalog
- IF files and metadata represent an elastic modulus map THEN re-run simulations and email notification to NU PIs

RIPPLE recipes

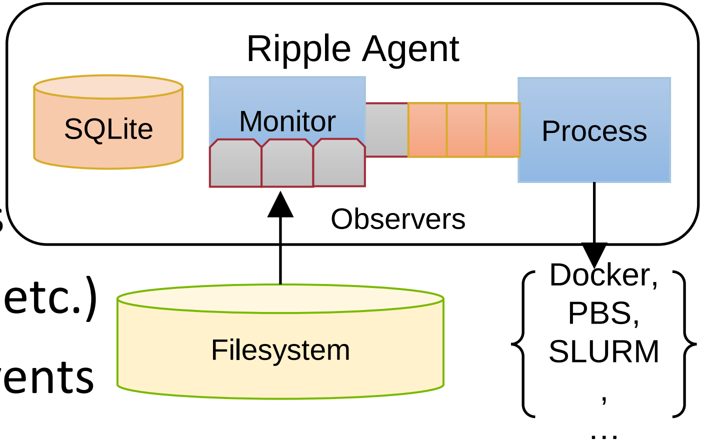
IFTTT-inspired programming model:

Triggers describe the event source (filesystem create events) and the conditions to match (/path/to/monitor/.*.h5)

Actions describe what service to use (e.g., Globus transfer) and arguments for processing (source/dest endpoints).

```
"recipe":{
  "trigger": {
    "username": "ryan",
    "monitor": "filesystem",
    "event": "FileCreatedEvent",
    "directory": "/path/to/monitor/",
    "filename": ".*.h5$"
  },
  "action": {
    "service": "globus",
    "type": "transfer"
    "source_ep": "endpoint1",
    "dest_ep": "endpoint2",
    "target_name": "$filename",
    "target_match": "",
    "target_replace": "",
    "target_path": "/~/$filename.h5",
    "task": "",
    "access_token": "<access token>"
  }
}
```

RIPPLE Agent



Triggers: Python Watchdog observers listen for events

- inotify, etc., for filesystem events (create, delete, etc.)
- Globus Transfer API for transfer, create, delete events

Rule evaluation: Performed by cloud-based service

- Recipes are stored locally in a SQLite database
- Local filtering then dispatched to AWS Lambda for evaluation

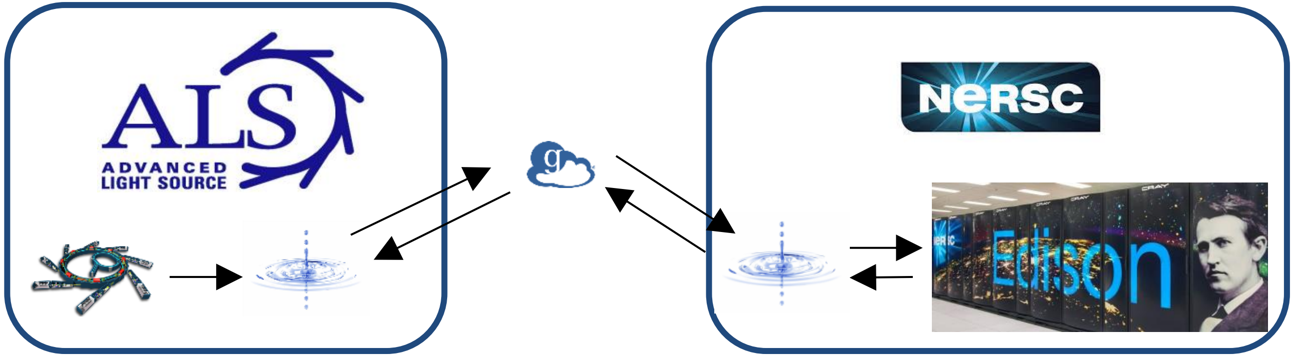
Actions: Local and cloud-based

- Docker containers act on local files (metadata extraction, dispatch jobs, etc.)
- Other tasks on cloud (Globus transfers, create shared endpoints, send emails, invoke other Lambda functions etc.)

Scenario: Advanced Light Source

Deployed Ripple on an ALS and NERSC machine to automate data analysis

- **At ALS:** Detect new heartbeat beamline data and initiate transfer to NERSC
- **At NERSC:** Extract metadata, create sbatch file, dispatch analysis job to Edison queue, detect result and transfer back to ALS
- **At ALS:** Create a shared endpoint, notify collaborators of result via email



Materials Data Facility aggregates and enables analysis of materials data and metadata

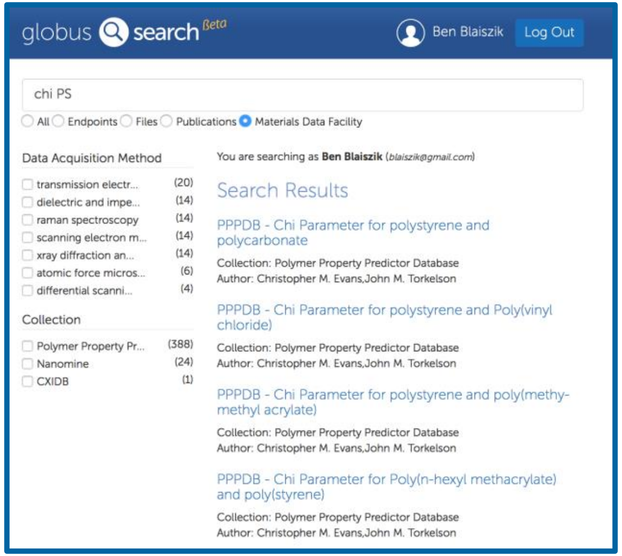

Ben Blaiszik

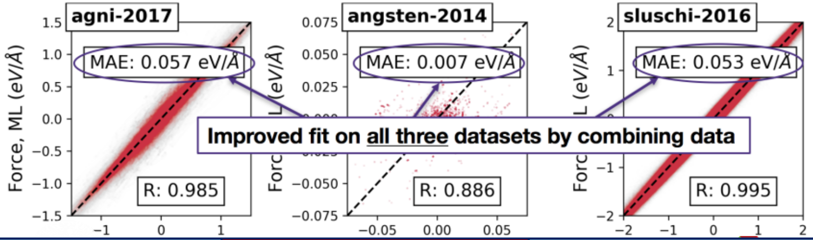

Logan Ward


Jim Pruyne


Kyle Chard

- Large quantities of materials data can enable new data-driven approaches to discovery, but are largely inaccessible
- MDF provides locus for both automated publication of new data and aggregation of metadata from existing collections
- 200 datasets, 270TB, 1M records aggregated to date; 10x more data in the pipeline
- Integrated schema, APIs, and machine learning methods enable programmatic discovery and access
- Early successes include improved force field potentials based on integration of data from multiple sources



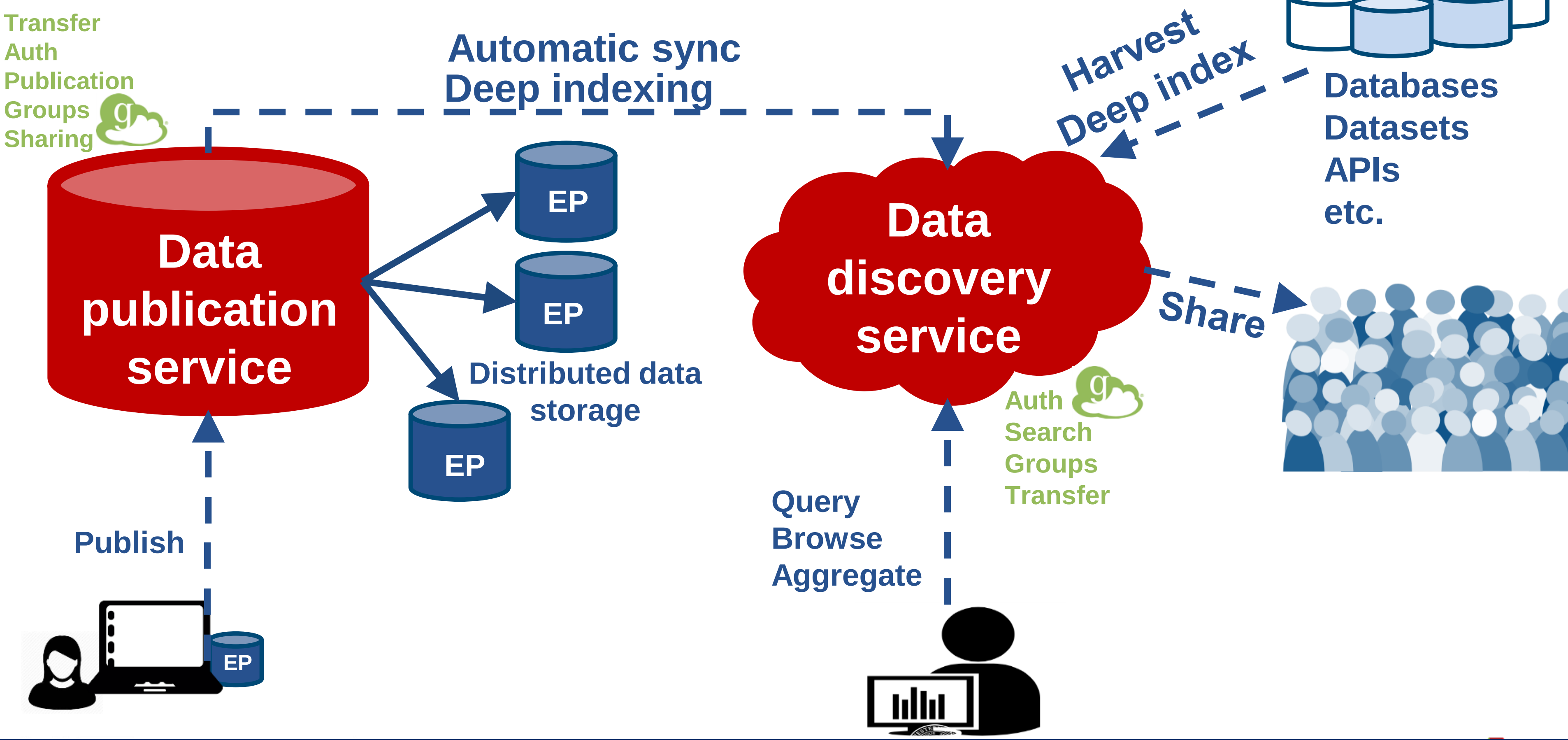


MATERIALS DATA FACILITY

Materialsdatafacility.org

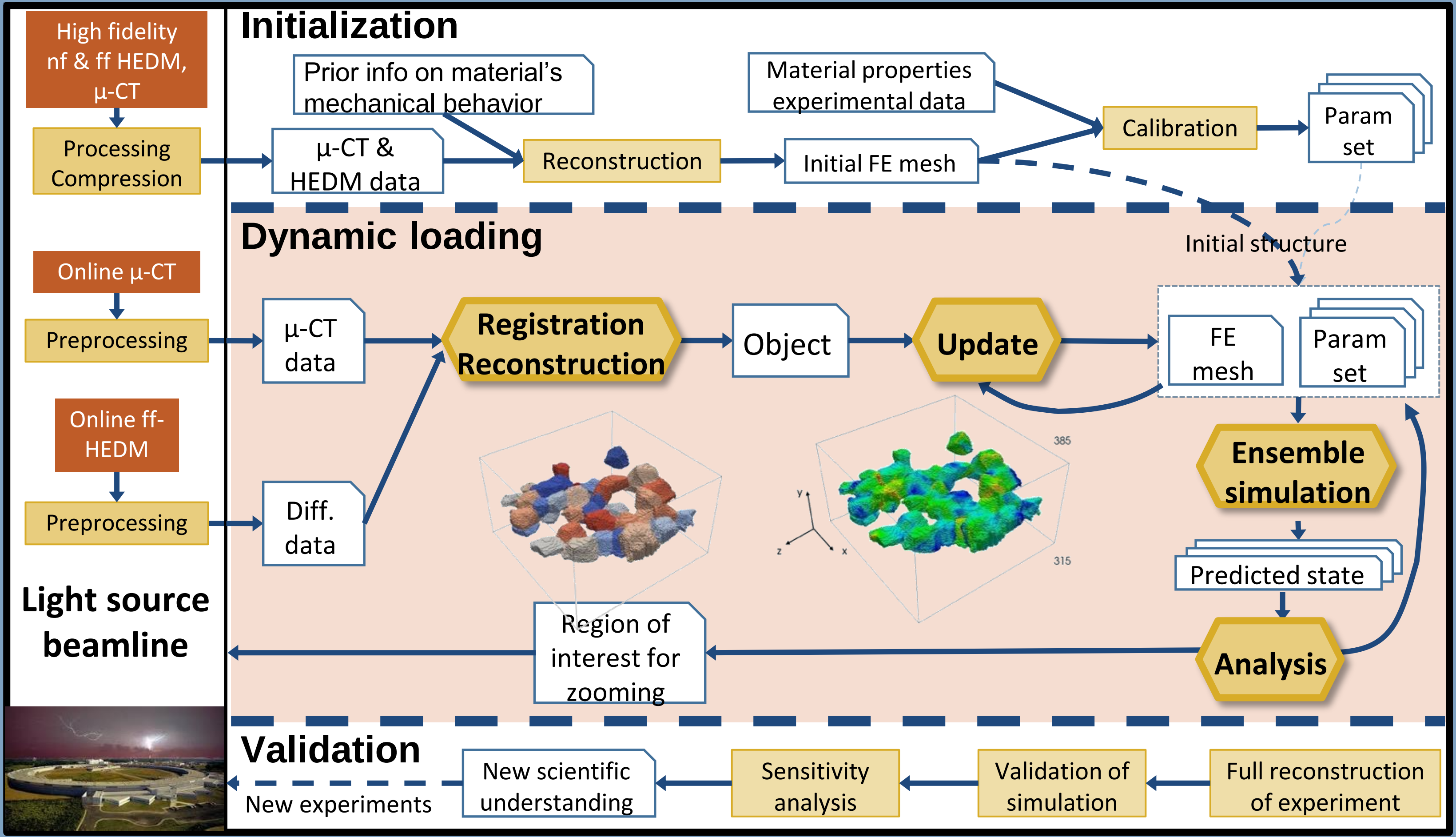
B. Blaiszik, K. Chard, J. Pruyne, I. Foster, The Materials Data Facility: Data Services to Advance Materials Science Research, Journal of Materials, 2016.

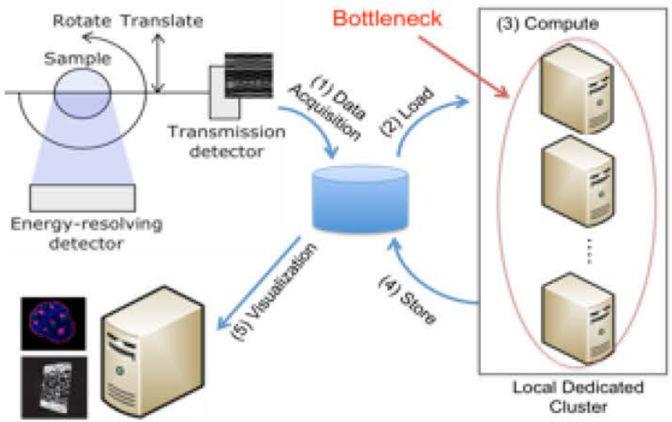
MDF supports publication & discovery

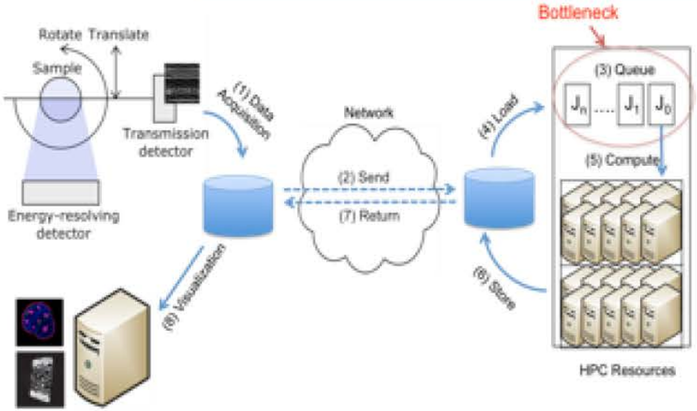


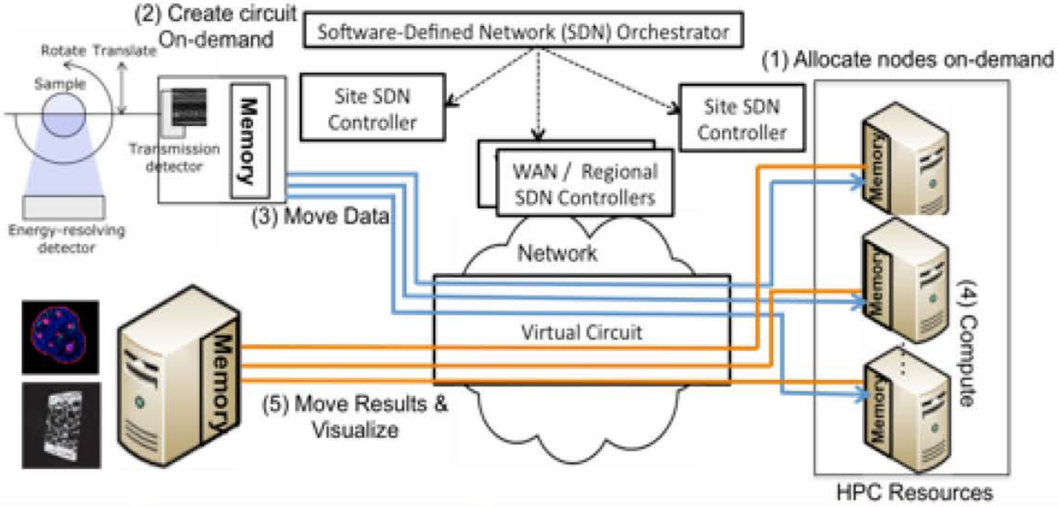
Challenges and opportunities

- Create new scientific instruments that link data acquisition and computation to measure the previously unmeasurable & increase utility of, and access to, expensive resources
- Enable reliable end-to-end streaming applications that span from instruments to networks to parallel computer memories
- Integrate pre-experiment and post-experiment activities
- Automation at all levels for throughput, reliability, and economy
- Architect and operate distributed computing systems to support varied, often demanding and mission-critical, workloads


Initialization
High fidelity nf & ff HEDM, μ-CT
Processing Compression
Prior info on material's mechanical behavior
μ-CT & HEDM data
Reconstruction
Material properties experimental data
Initial FE mesh
Calibration
Param set
Dynamic loading
Initial structure
Online μ-CT
Preprocessing
μ-CT data
Registration Reconstruction
Object
Update
FE mesh
Param set
Online ff-HEDM
Preprocessing
Diff. data
Ensemble simulation
Predicted state
Analysis
Region of interest for zooming
Light source beamline
Validation
New experiments
New scientific understanding
Sensitivity analysis
Validation of simulation
Full reconstruction of experiment

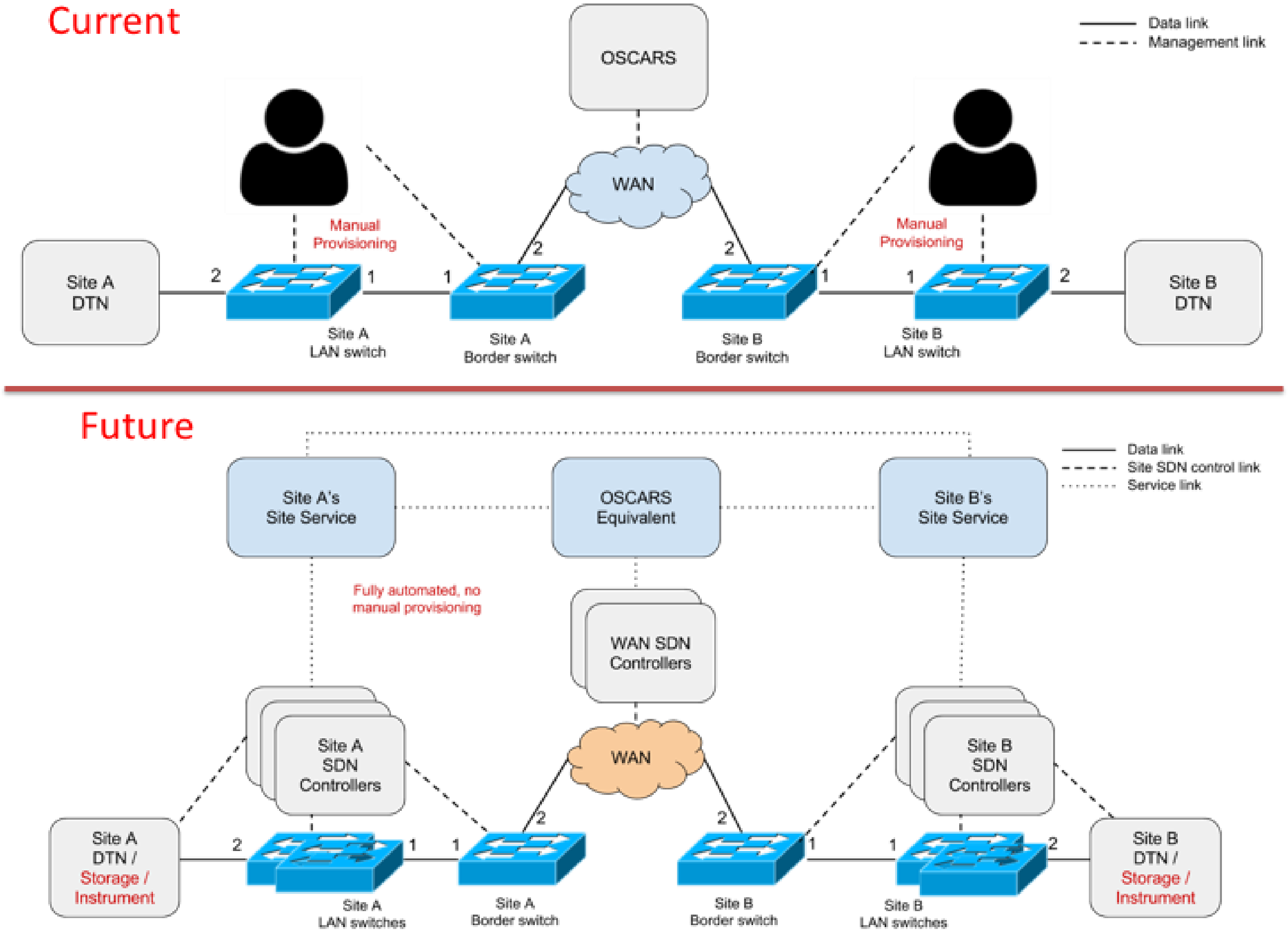




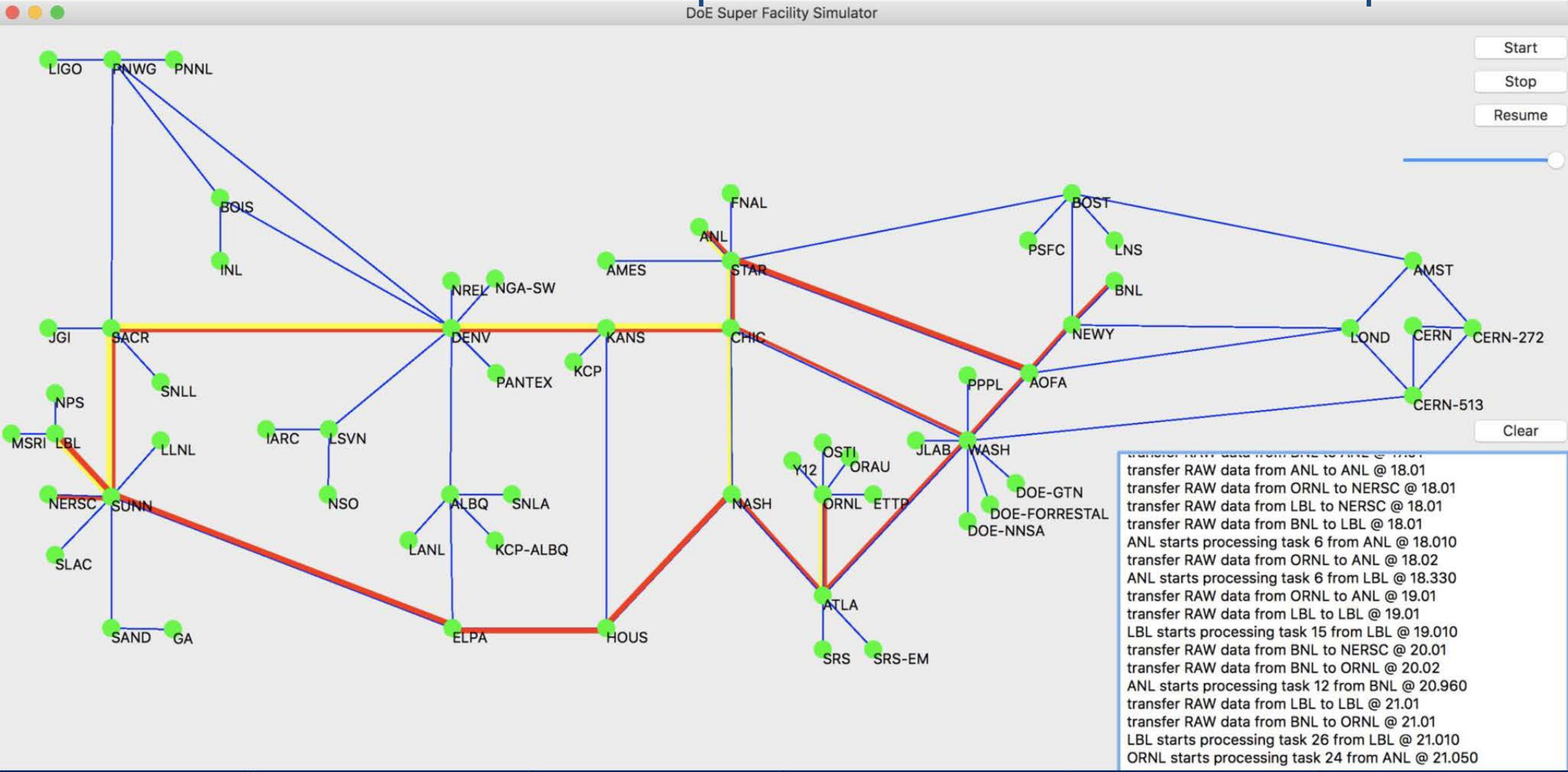


- Eliminate queue wait-time for online analysis - incentivize batch jobs
- Eliminate network contention - automated provisioning of network
- Eliminate disk I/O - stream data directly from detector to compute memory

Software defined networks science flows: Automated provisioning of end-to-end network paths



Simulation to understand and optimize the entire science complex



Thank you to our sponsors

U.S. DEPARTMENT OF ENERGY

NIST National Institute of Standards and Technology U.S. Department of Commerce

NSF

NATIONAL INSTITUTES OF HEALTH

THE UNIVERSITY OF CHICAGO

ALFRED P. SLOAN FOUNDATION 1934

Argonne NATIONAL LABORATORY

For more information

www.globus.org

labs.globus.org

foster@anl.gov