4CeeD: Real-Time Operating Infrastructure for Capturing, Curating, Correlating, Coordinating and Distributing Materials-related Data

Klara Nahrstedt

University of Illinois at Urbana-Champaign

http://t2c2.csl.illinois.edu

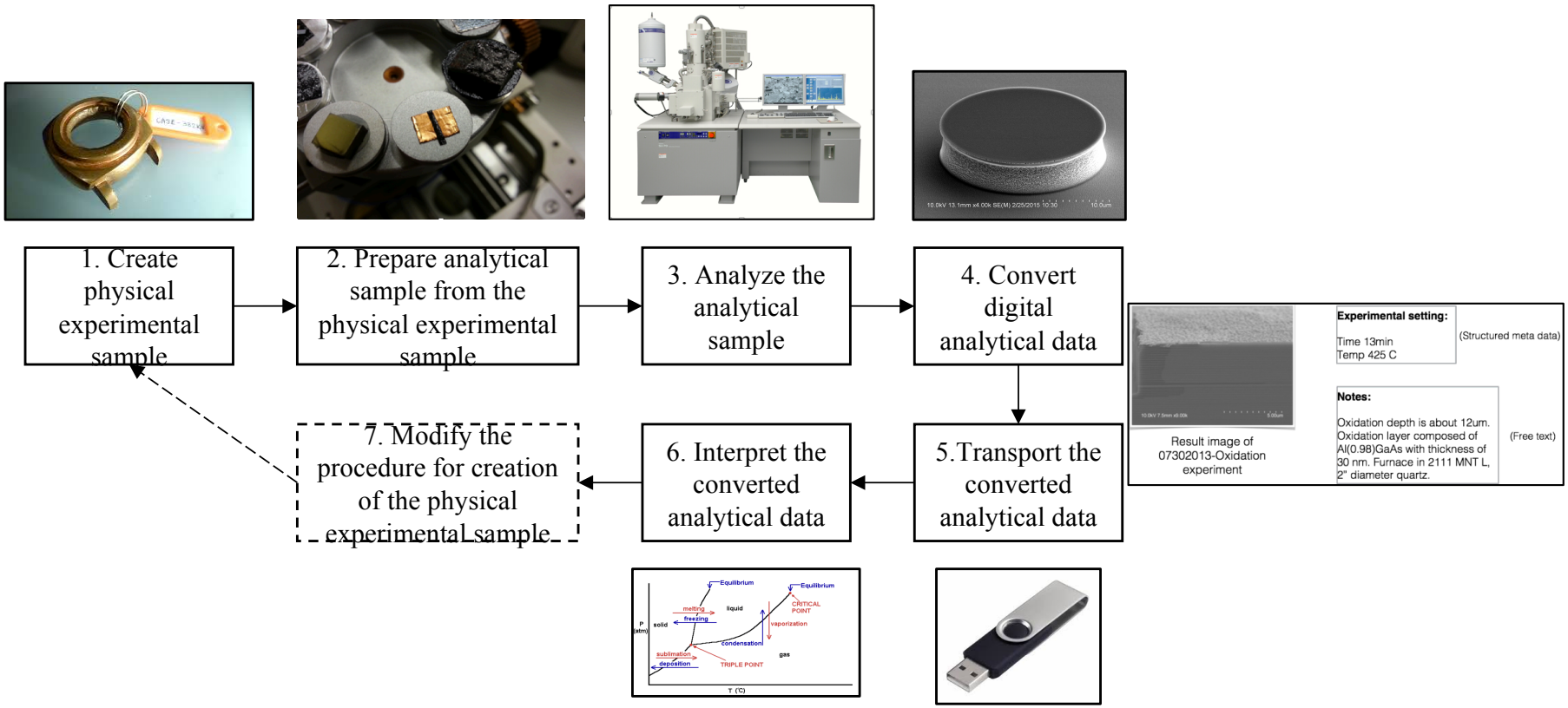
Co-Authors: Phuong Nguyen, Steven Konstanty, Todd Nicholson, Thomas O'brien, Patric Su, Aaron Schwartz-Duval, Tarek Elgamal, Tim Spila, Roy Campbell, Indy Gupta, Michael Chan, Kenton McHenry, Norman Paquin,

Outline

- Overall Problem Description
- 4CeeD Solution
  - System Perspective
  - User Perspective
- BRACELET in 4CeeD
  - System Perspective
- Conclusion

Background: Long process from material discovery to device fabrication

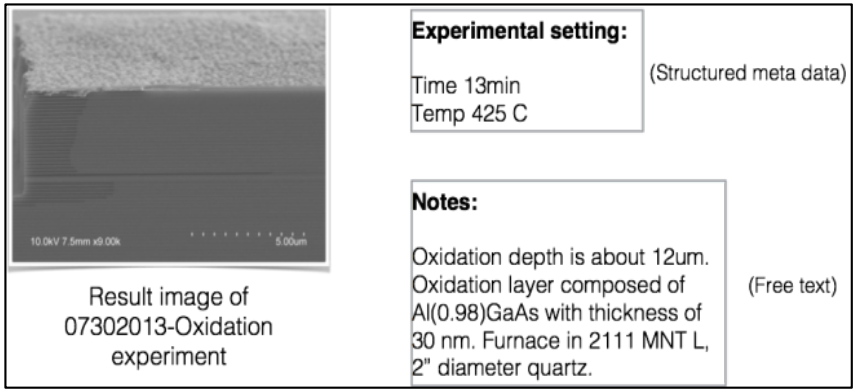
- It typically takes 20 years to go from the discovery of new materials to fabrication of new and next-generation devices*



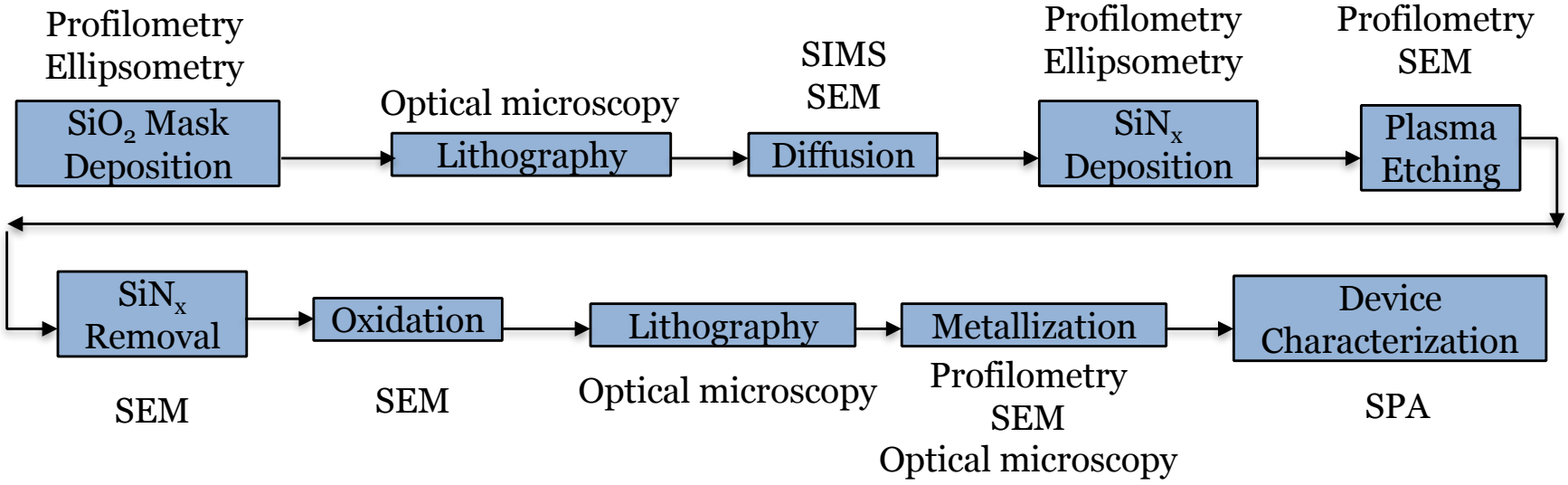
* Holdren, J.P. "*Materials genome initiative for global competitiveness*". National Science and Technology Council OSTP 2011.

Scientific Digital Data Acquisition and Workflow Challenges

- Data come of various types and multimodal formats
  - Each type requires a different data processing workflow



Sample output data from SEM microscopy



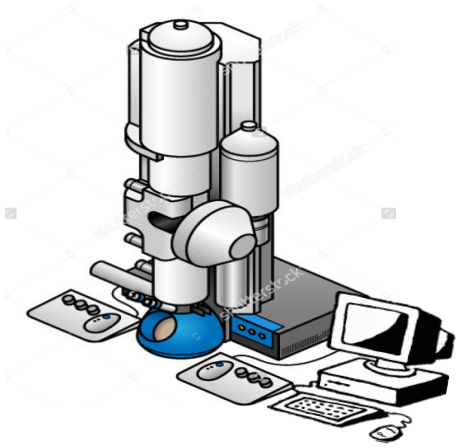
Current issues with scientific data capture, management and sharing

- Data acquisition & transfer
  - *"Sneaker-net"* data transfer
  - No data conversion is available during data acquisition

- Data management & processing
  - Using local file system, or general-purpose cloud storage services (e.g., DropBox, Google Drive, etc.)

- Data access & sharing
  - Materials science & semiconductor fabrication areas have never been digitally connected
  - Only *"best"* results and data are kept for publishing

Current State of Data Capture



Instrument
(MRL/MNTL)

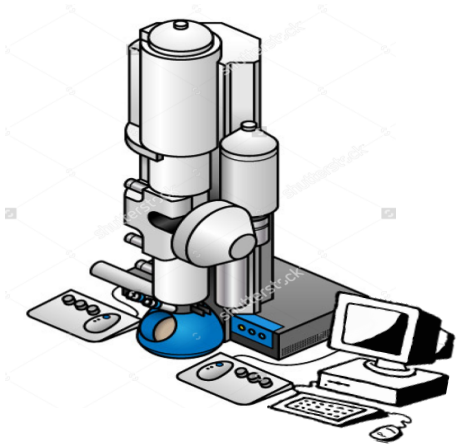
- Fabricate experimental sample
- Prepare analytical sample
- Bring sample to instrument for analysis
- **Extract data (File conversion)**
- **Transport data to office computer**
- **Extract data (File conversion)**
- **Analyze data**
- **Transport data to office computer**
- **Analyze data**
- Repeat per iteration

Flash Drives

Limits of home analysis
'Sneakernet'
Metadata Loss
-Only what you have
-Security risk at both ends
-Excessive file name schemes
-Connect notes to data
-Very limited transport space
-Manual notes
(after time delay)
-Lost or forgotten flash drives
-What metadata is important?
-Retroactive not reactive

Office

Our Goal for Data Capture



Instrument
(MRL/MNTL)

- Fabricate experimental sample
- Prepare analytical sample
- Bring sample to instrument for analysis
- **Extract data (NO FILE CONVERSION)**
- **Transport data to office computer (DIRECT)**
- **Analyze data (REAL TIME)**
- Repeat per iteration

Real time interface

- Lossless data
- More simple file names
- Automatic note taking
- Easy to search/annotate
- All metadata included

4CeeD Data Capture

Laptop

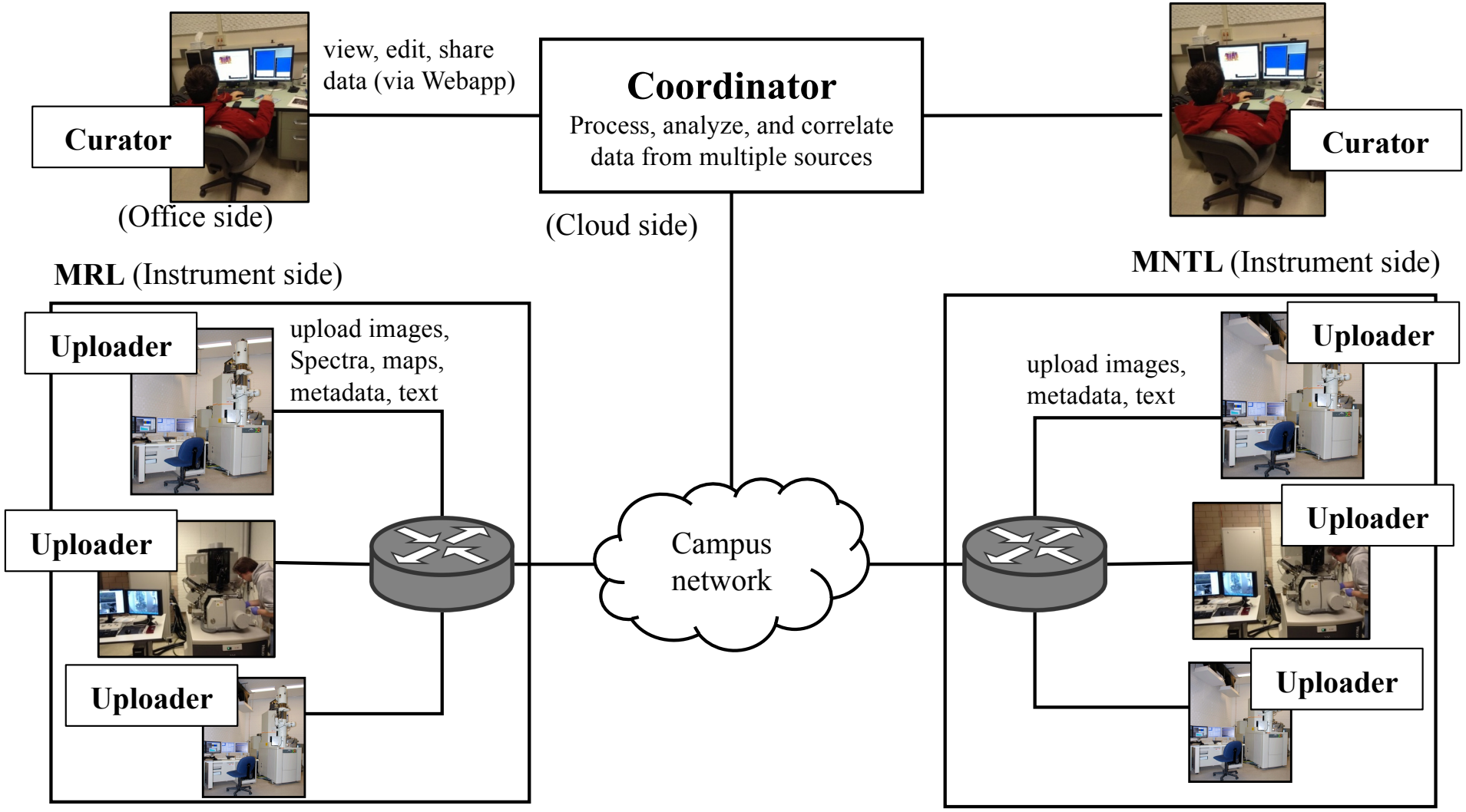
Campus PC

Collaborators

Office

Our Approach: 4CeeD System Perspective

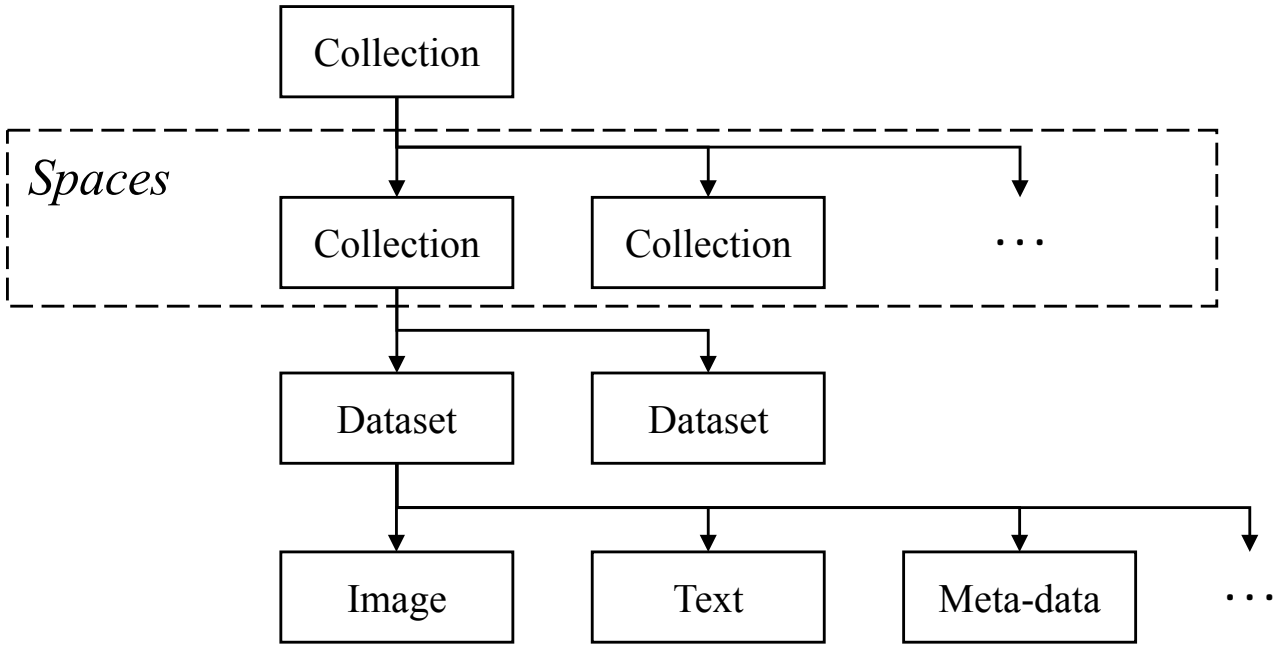
Our approach: 4CeeD Private Cloud Compute and Storage Service



* **4CeeD** stands for **C**apture, **C**urate, **C**oordinate, **C**orrelate, and **D**istribute material-related experimental data.

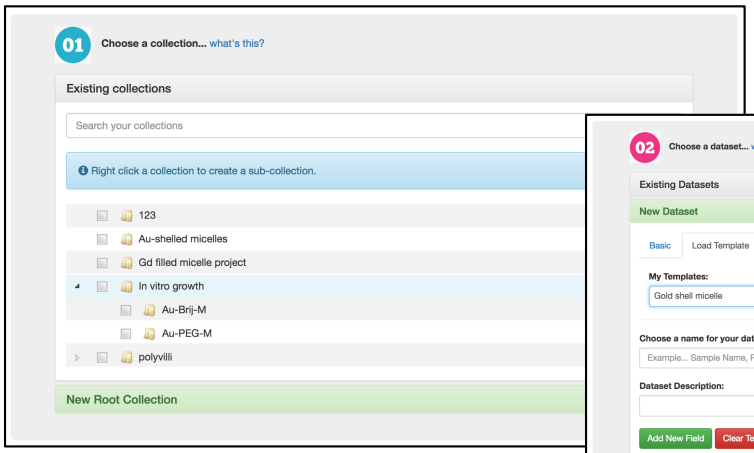
4CeeD Data Model

- Extendable data model
  - ***Vertical*** organization: Using concepts of *collections*, *datasets*, and *files* to provide flexible data organization
  - ***Horizontal*** organization: Using *spaces* to support sharing of data

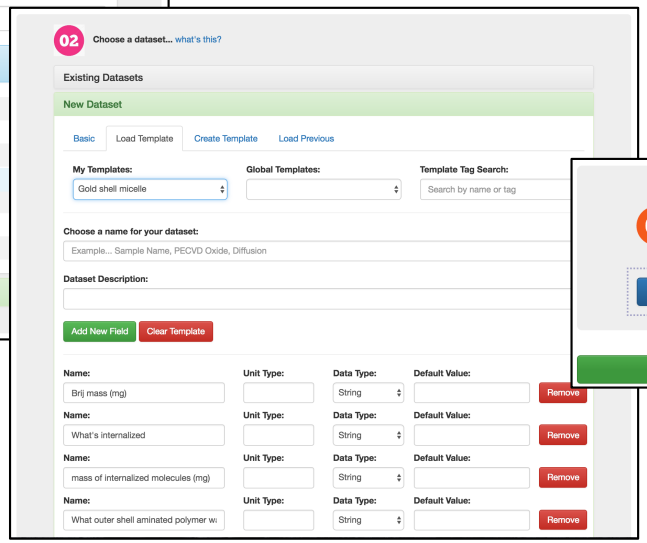


4CeeD Data Uploader Service

- Uploader offers a simple 3-step interface, following the data model



Create collection

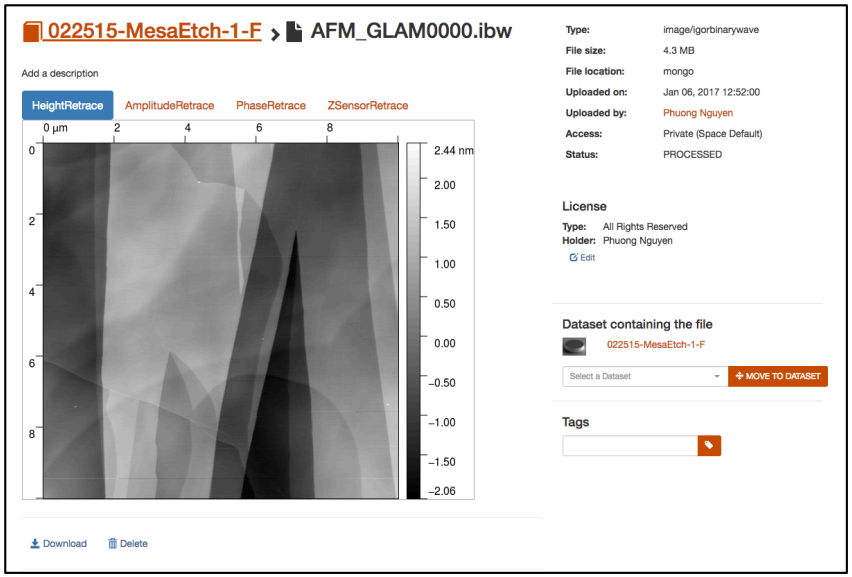


Create dataset & associated meta-data

Select raw data files to upload

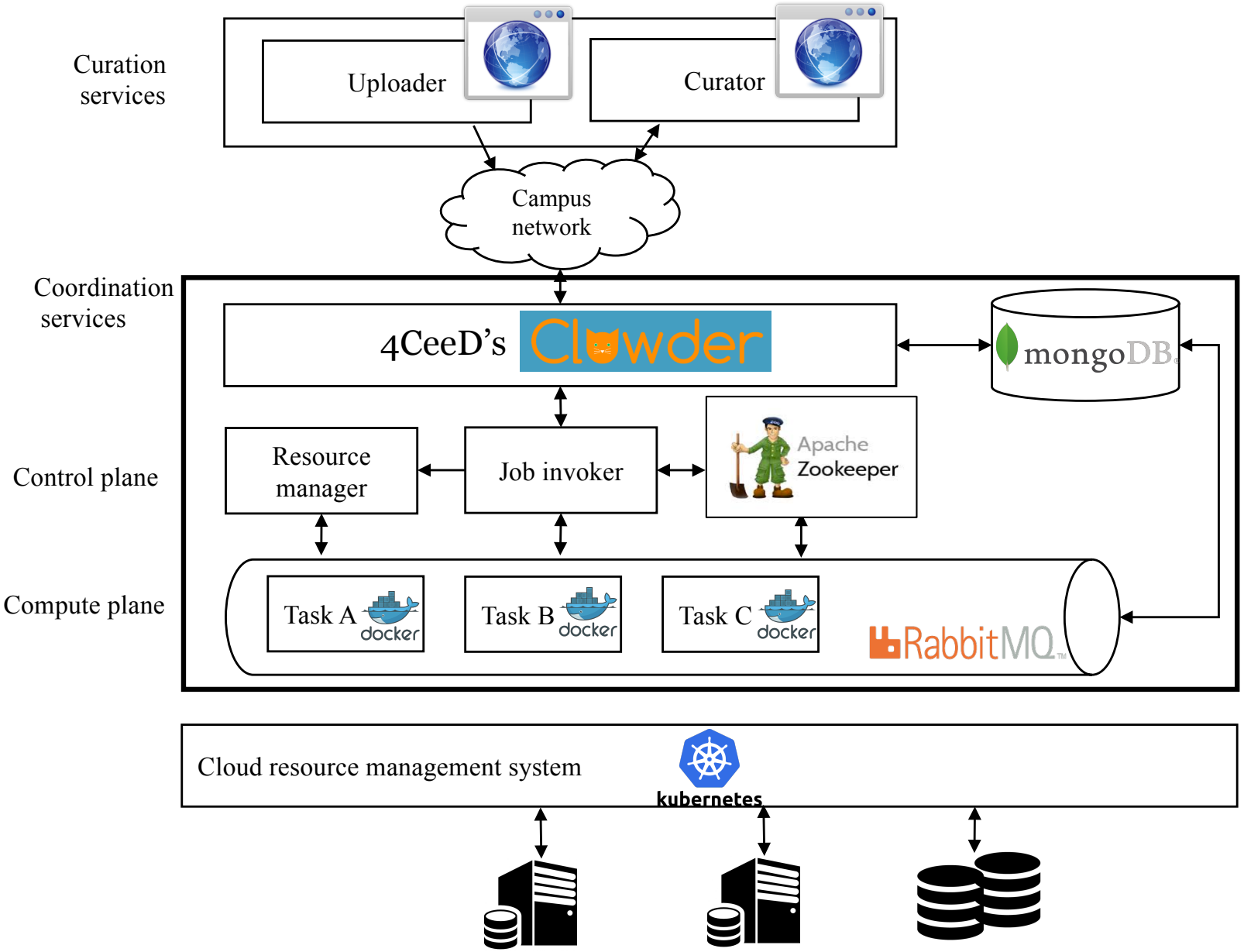
4CeeD Data Curator Service

- Uploaded raw data is processed and available for curation (tagging, annotation) and sharing
- Processed data is indexed and available for search

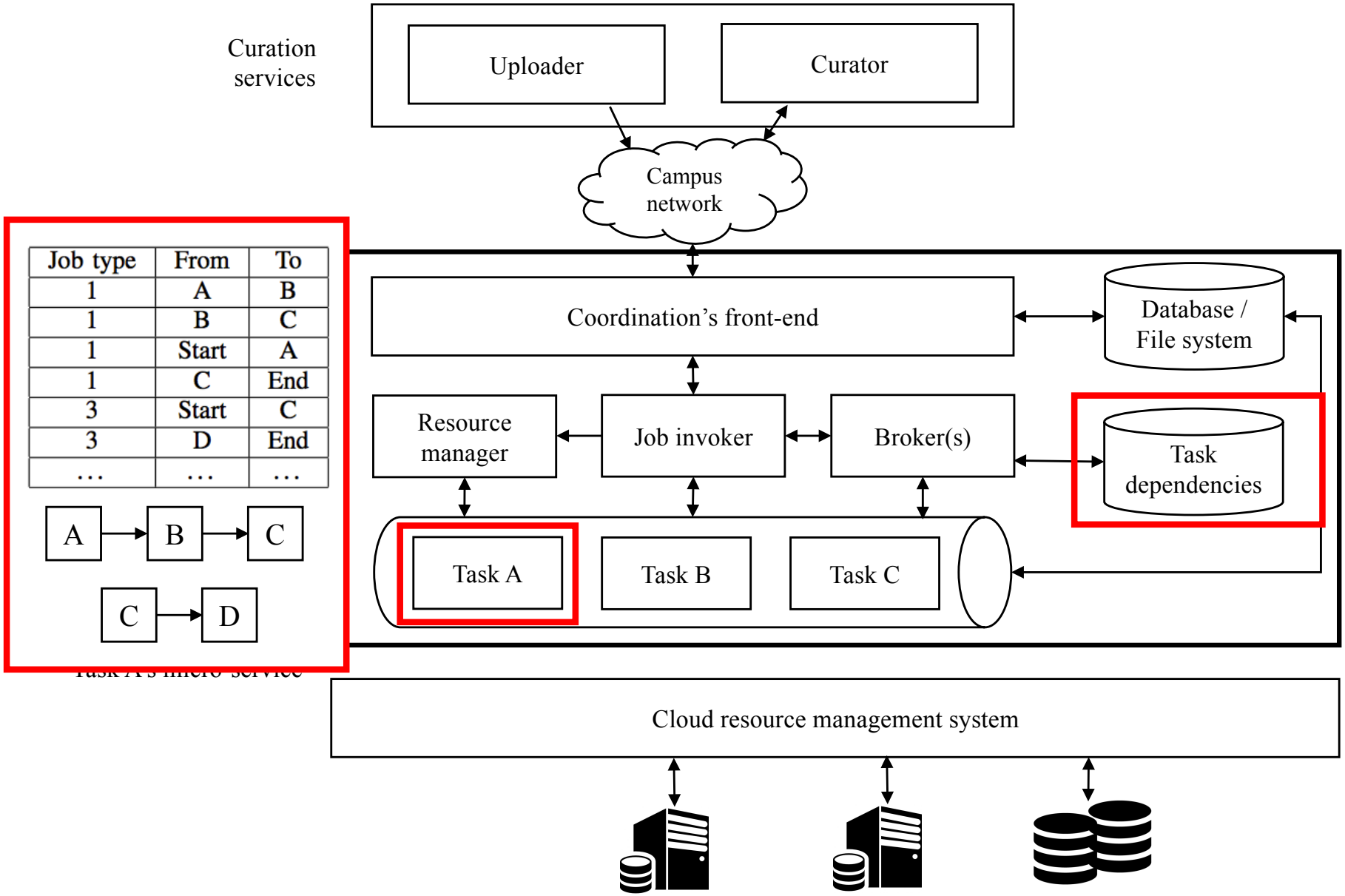


Example of processed AFM experiment data with gray-shade preview & meta-data

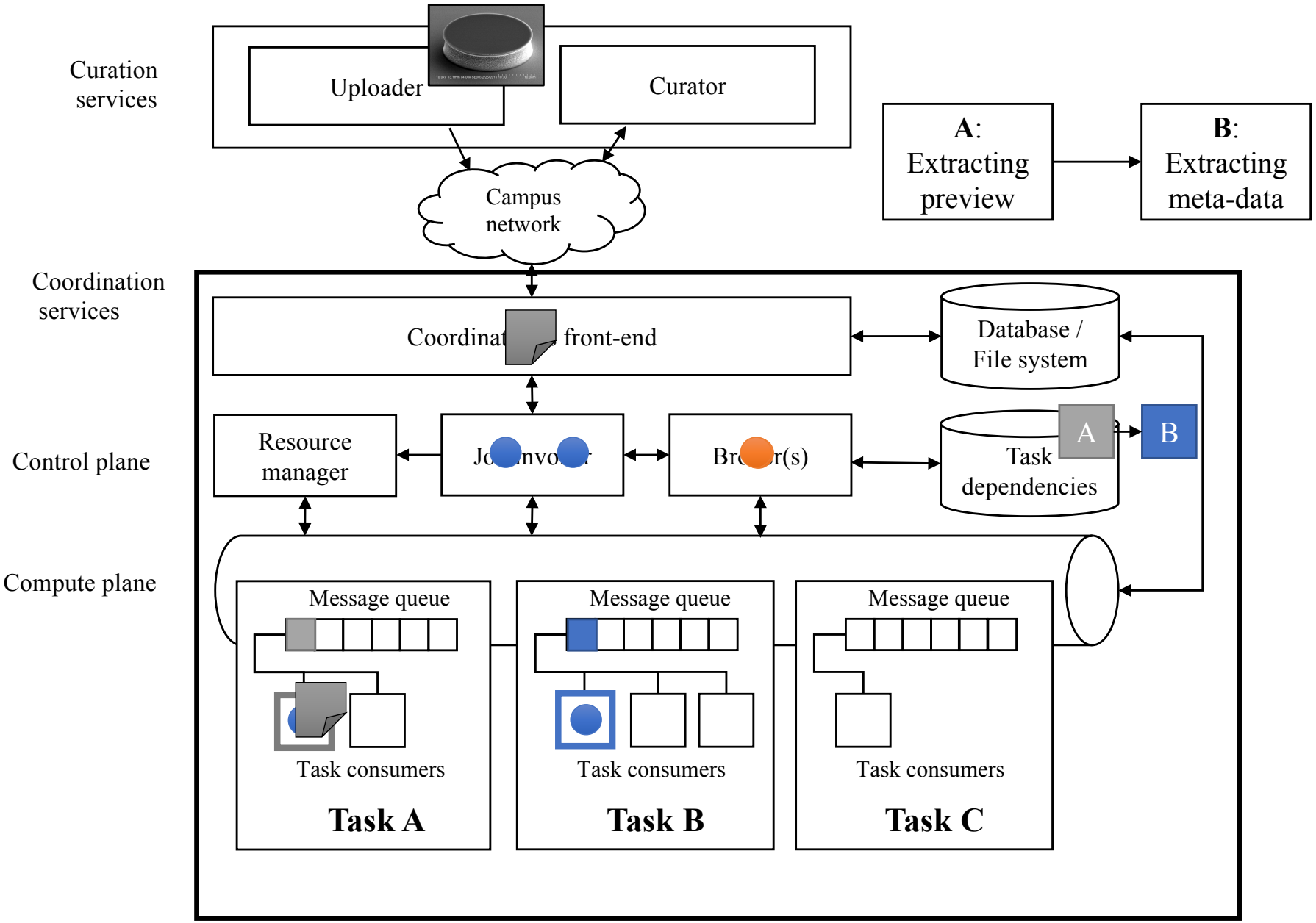
4CeeD's Coordination Service



4CeeD Coordination Service - Architecture



4CeeD's Coordination Service



Coordination Service - Consumer Scheduling (White Box Approach)

- Objective:
  - To **decide how many consumers should be allocated to each task's micro-service** to optimize the performance, subjected to resource constraint
- (a) white box approach: We use ***work-in-progress*** as system performance metric:
  - $WIP = \sum_j WIP_j$ (WIP is proportional to response time, via Little's law)
  - Leverage our previous work* on performance model of elastic pub/sub system to represent WIP_j as a function of number of consumers of task j-th: $WIP_j(m_j)$

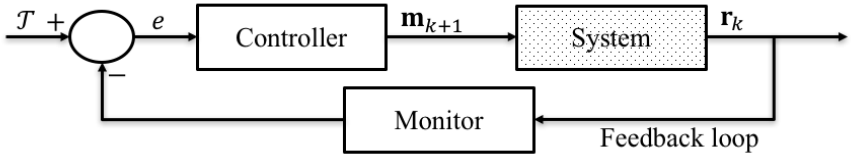
$$argmin_{\{m_j\}} \sum_j WIP_j(m_j)$$

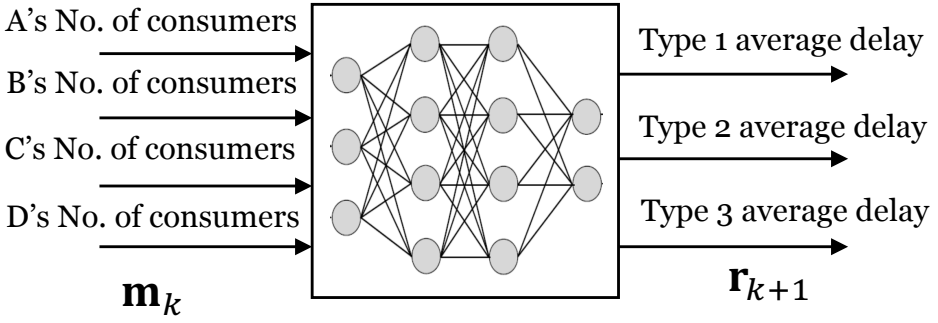
$$\sum_j m_j \leq \mathcal{C}$$

- (b) black box approach: We use model-predictive control

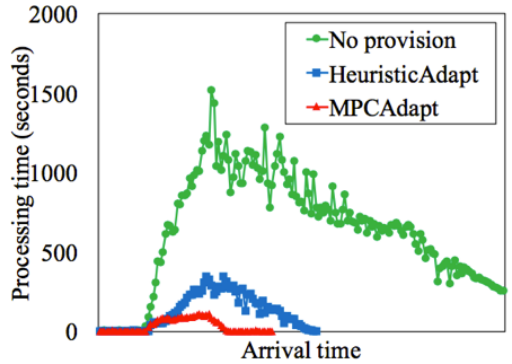
Resource Management (Black Box Approach)

- Robust resource management using model predictive control (from control theory) and neural network-based performance model

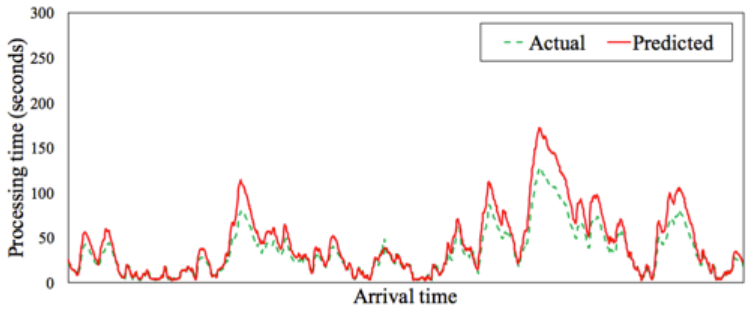




$$\mathbf{r}_{k+1} = f(\mathbf{r}_k, \mathbf{m}_k)$$



Effectiveness of feedback control-based adaptation algorithms when dealing with bursty workloads



Neural network-based system identification can accurately predict system performance

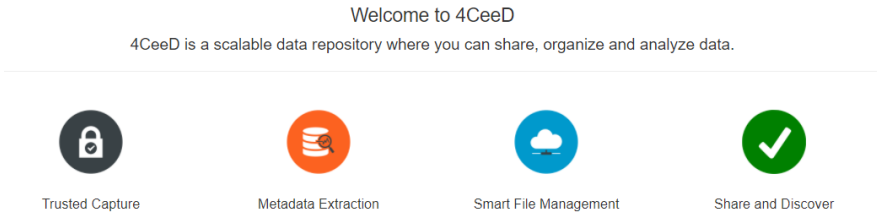
Our Approach: 4CeeD User Perspective (Example from MNTL)

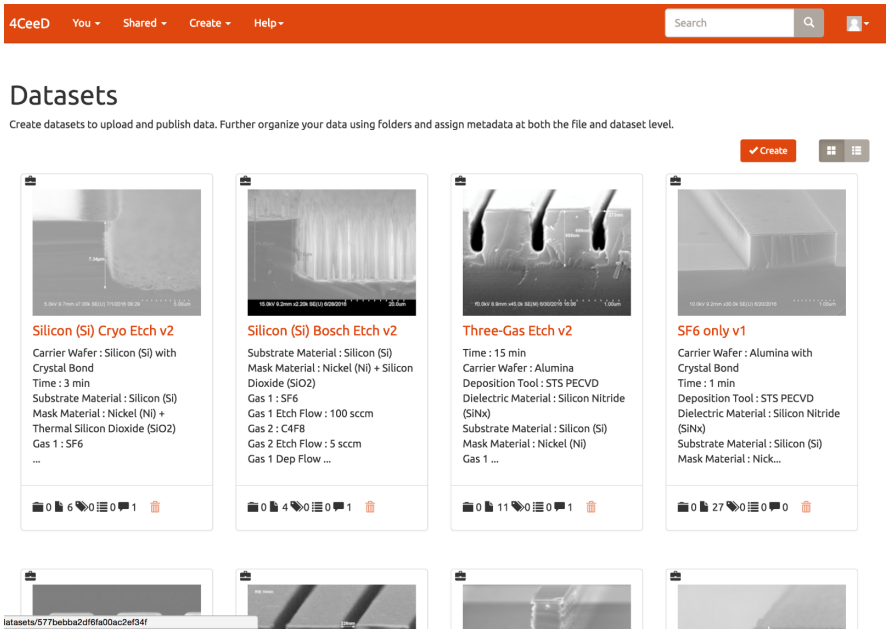
http://t2c2.csl.illinois.edu/demo/

4CeeD Capabilities

- Key Features

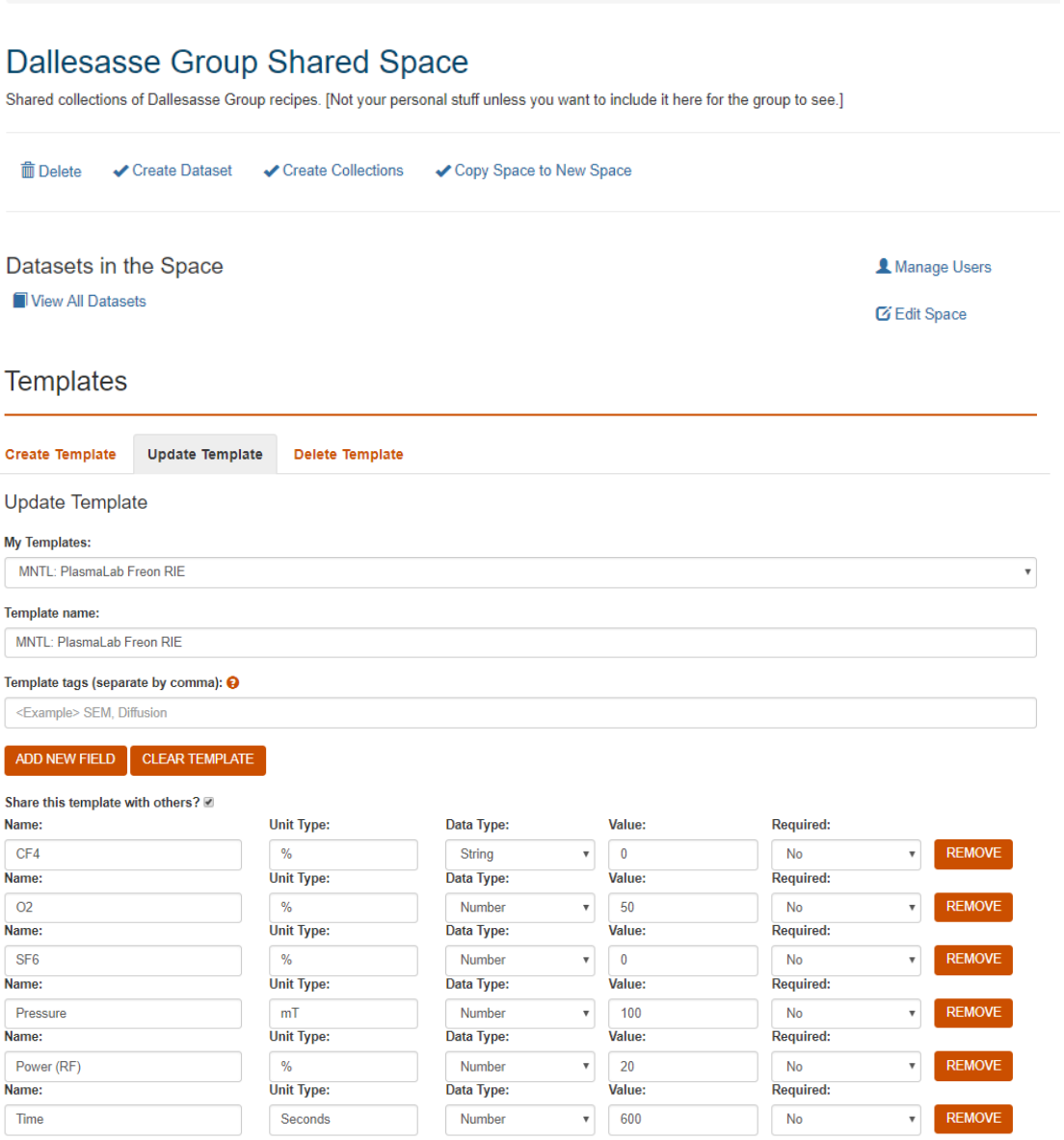
- Usage Cases



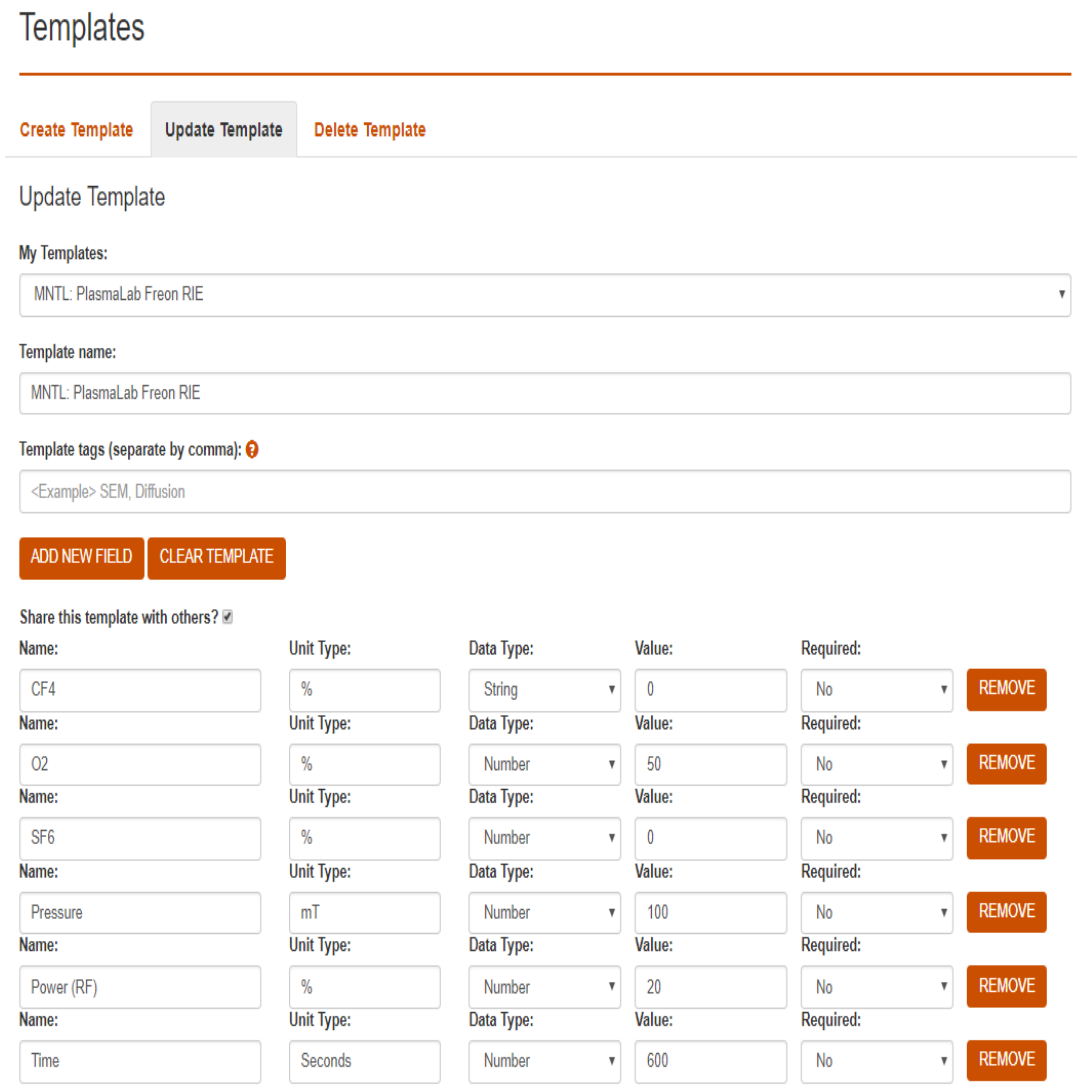


4CeeD Capabilities

- Key Selected Features
  - Metadata
  - Template
  - File Sharing/Spaces
  - File Extraction
- Usage Cases



Key Feature: Metadata/Template

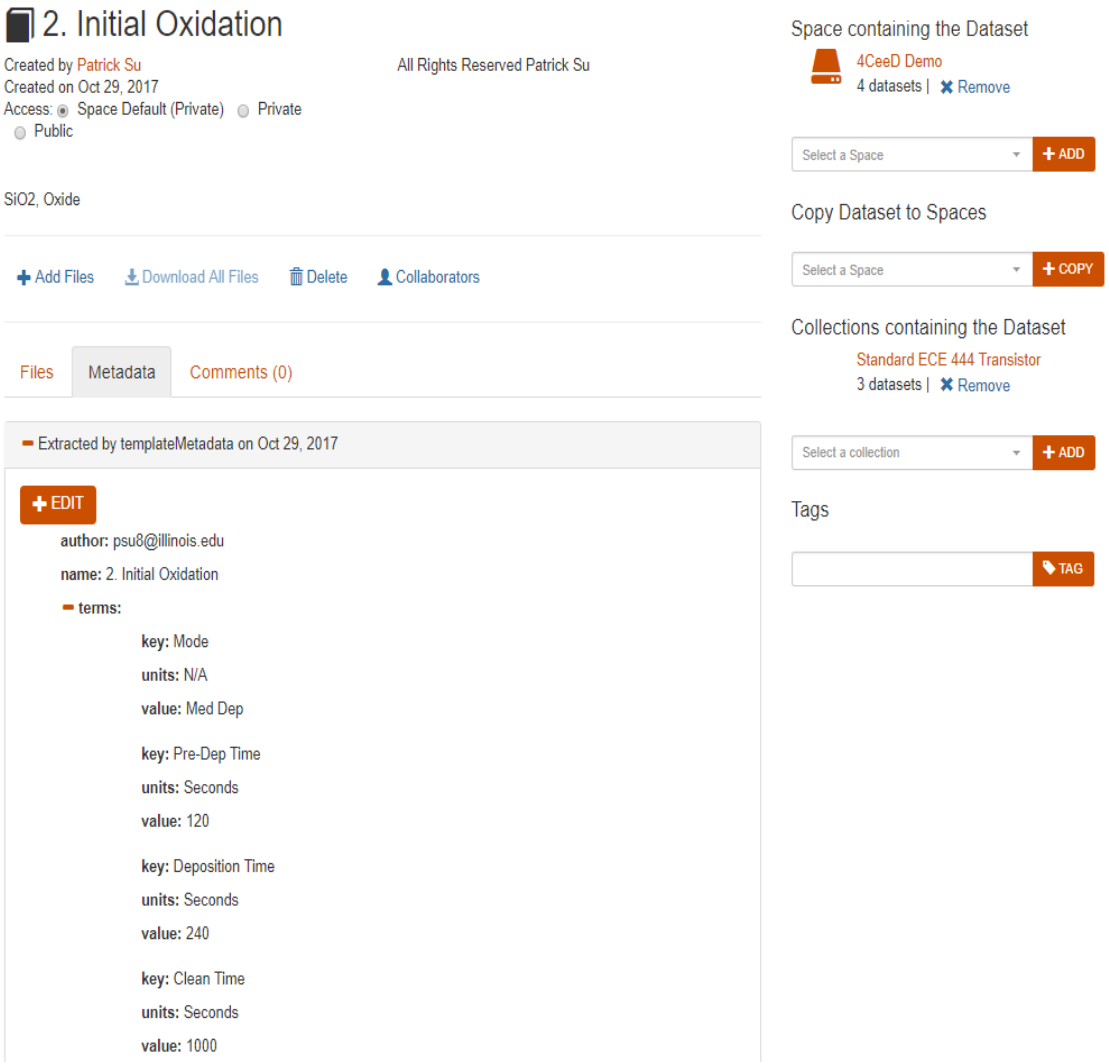


MNTL Template for Plasmalab Freon RIE

- Tracking and keeping fabrication recipes
- Easy to enter in all of your data at once
- MNTL Tool templates already made for users:

  Trion Minilock PECVD,
  STS PECVD,
  Oxford Freon ICP RIE
  Oxford ICP RIE
  Lithography (Standard, Image Reversal)
  Ellipsometer
  and more!

Key Feature: Metadata/Template

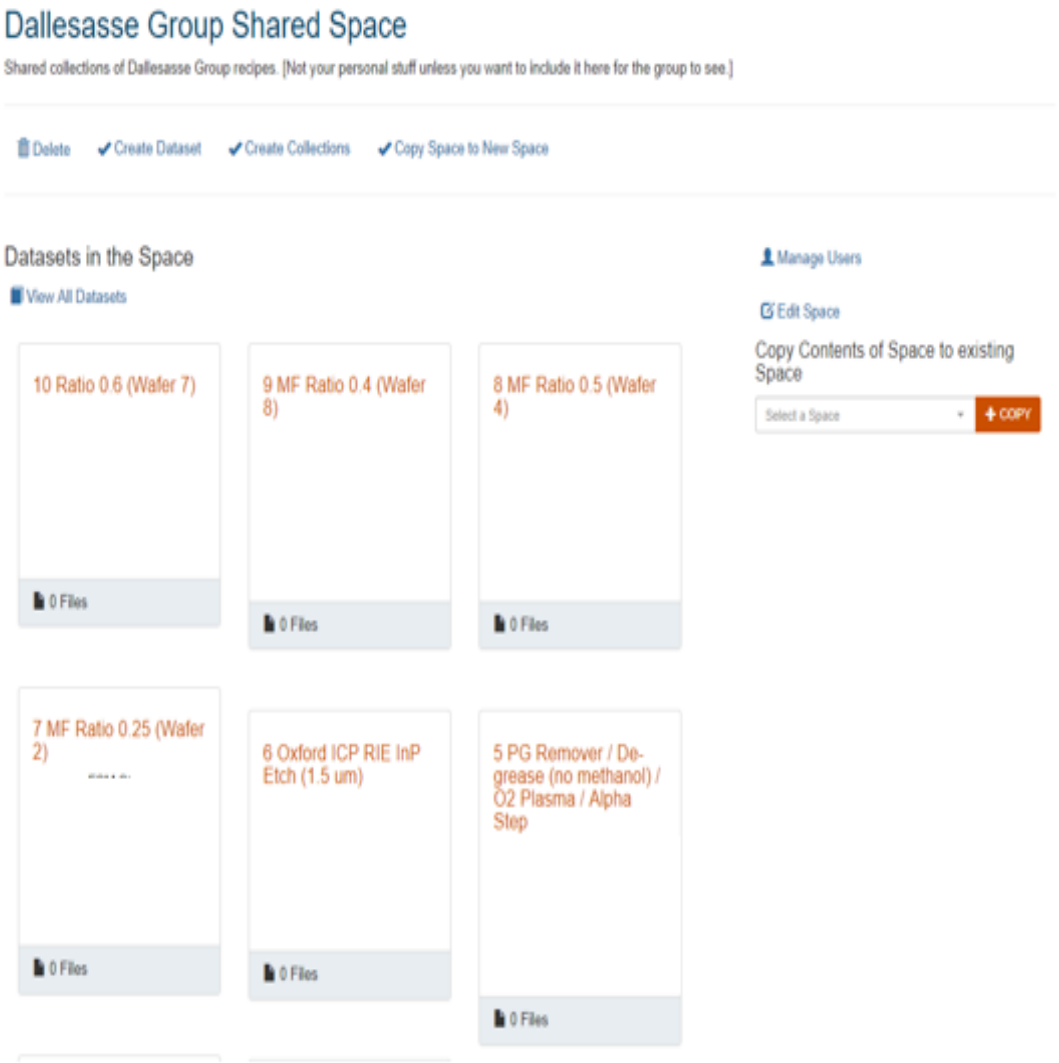


Metadata for SiO2 Deposition

- Tracking and keeping fabrication recipes
- Easy to enter in all of your data at once
- MNTL Tool templates already made for users:

  Trion Minilock PECVD,
  STS PECVD,
  Oxford Freon ICP RIE
  Oxford ICP RIE
  Lithography (Standard, Image Reversal)
  Ellipsometer
  and more!

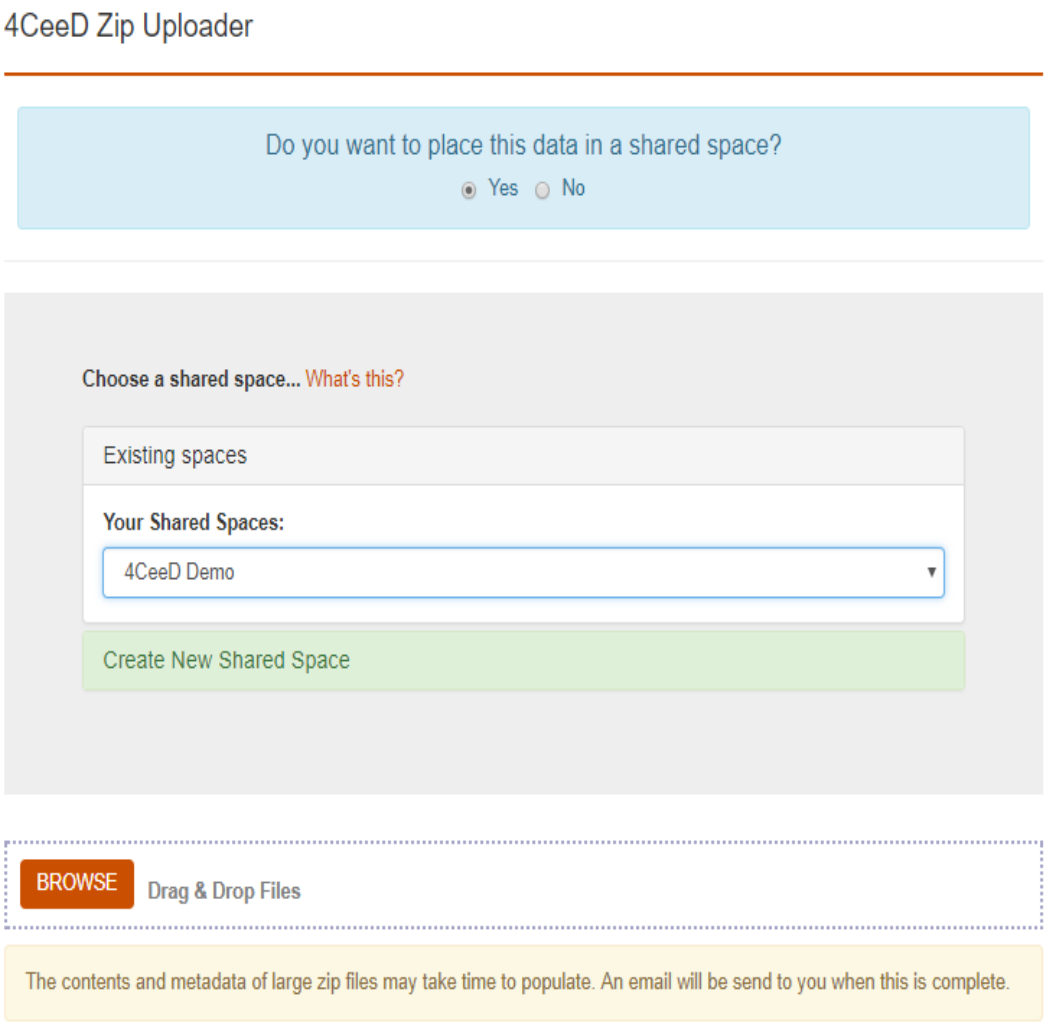
Key Feature: Group Spaces/File Extraction



Shared Spaces

- Group spaces facilitate updated coordination amongst collaborators
- Fabrication flows and experiments are always up to date
- Easy visual analysis of all updated data

Key Feature: Group Spaces/File Extraction



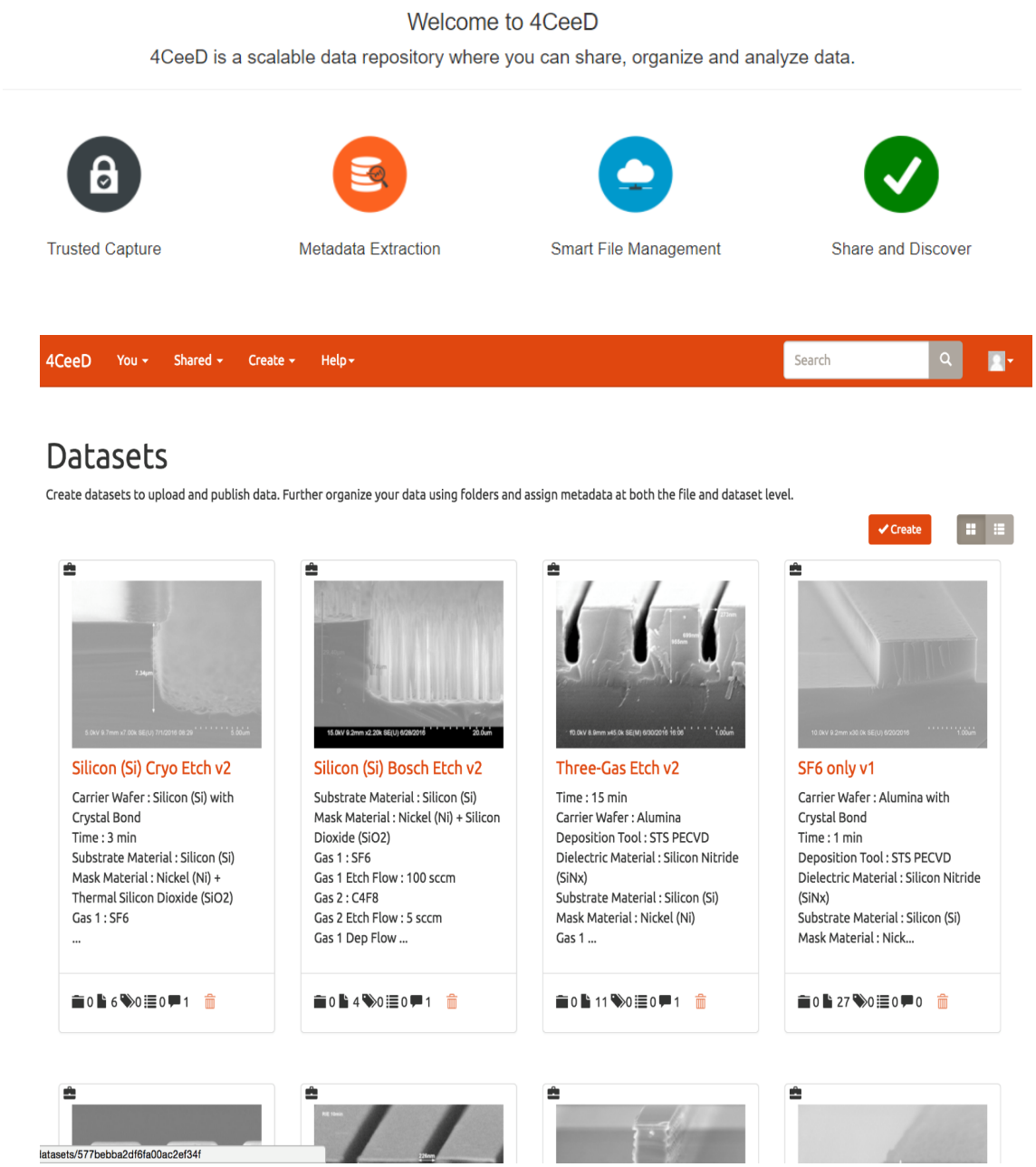
4CeeD Zip Uploader

- Zip Uploader and File Extractor allows easy transition to 4CeeD
- Capable of extracting large files (record is 30 GB from single user)

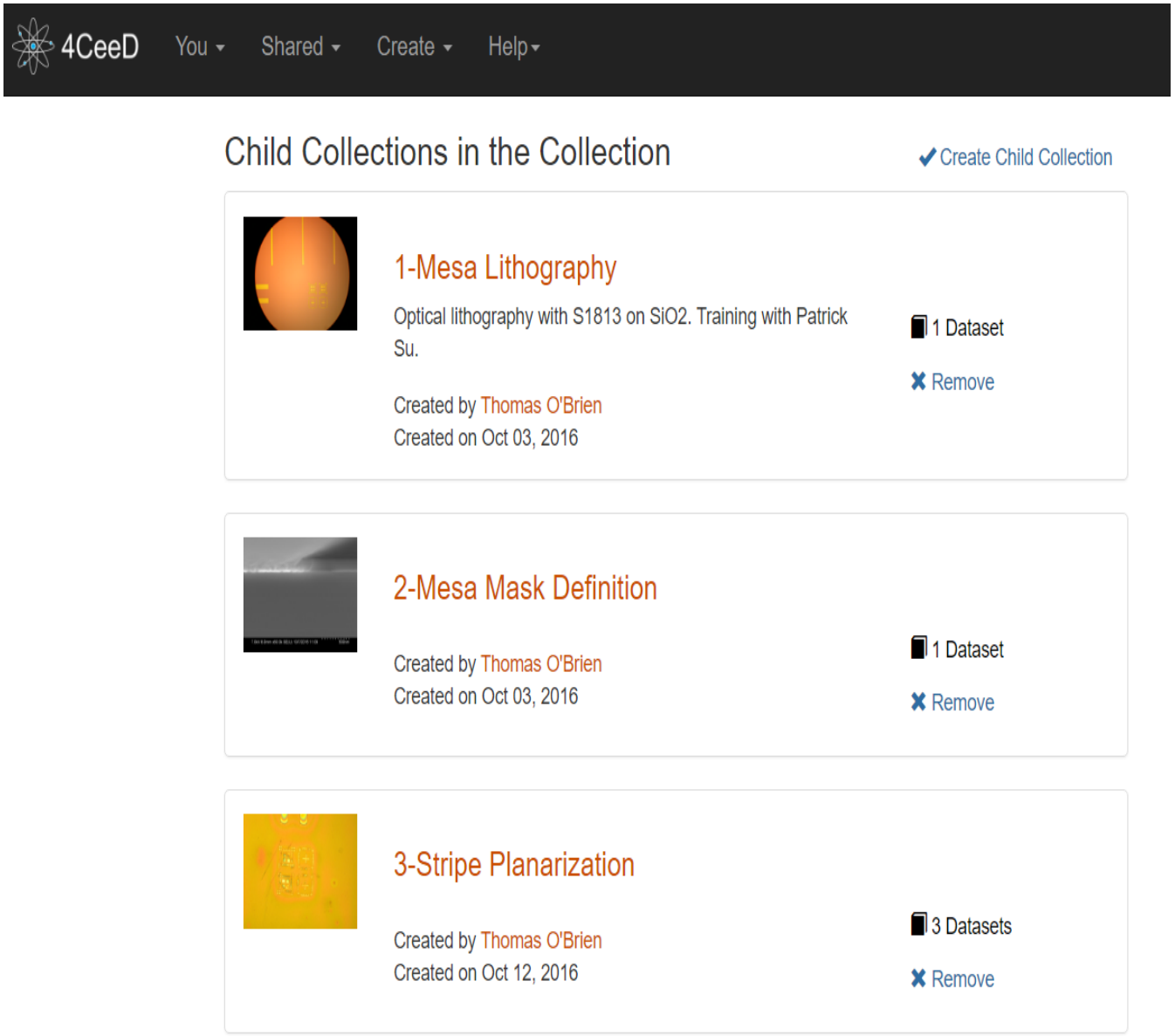
4CeeD Capabilities

- Key Features

- Usage Cases
  - Fabrication Flow
  - Process Analysis
  - Device Archiving

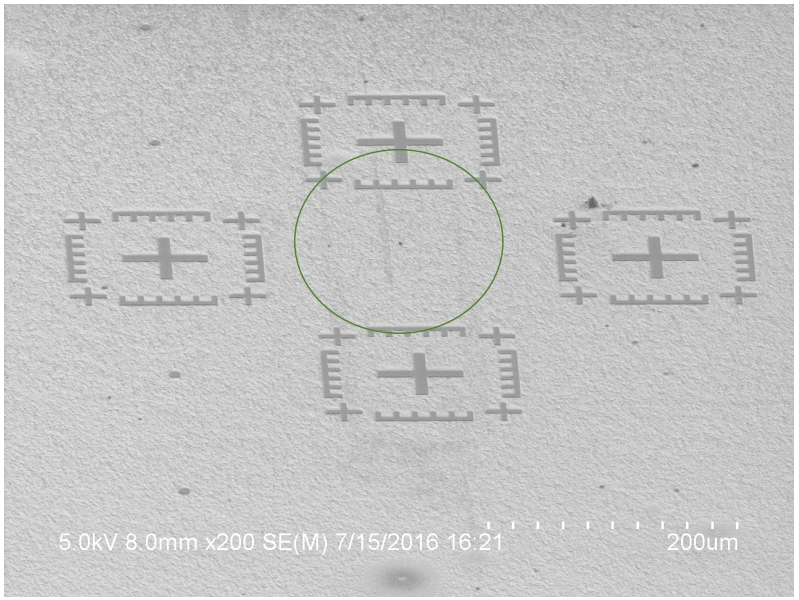


Usage Case: Fabrication Flow

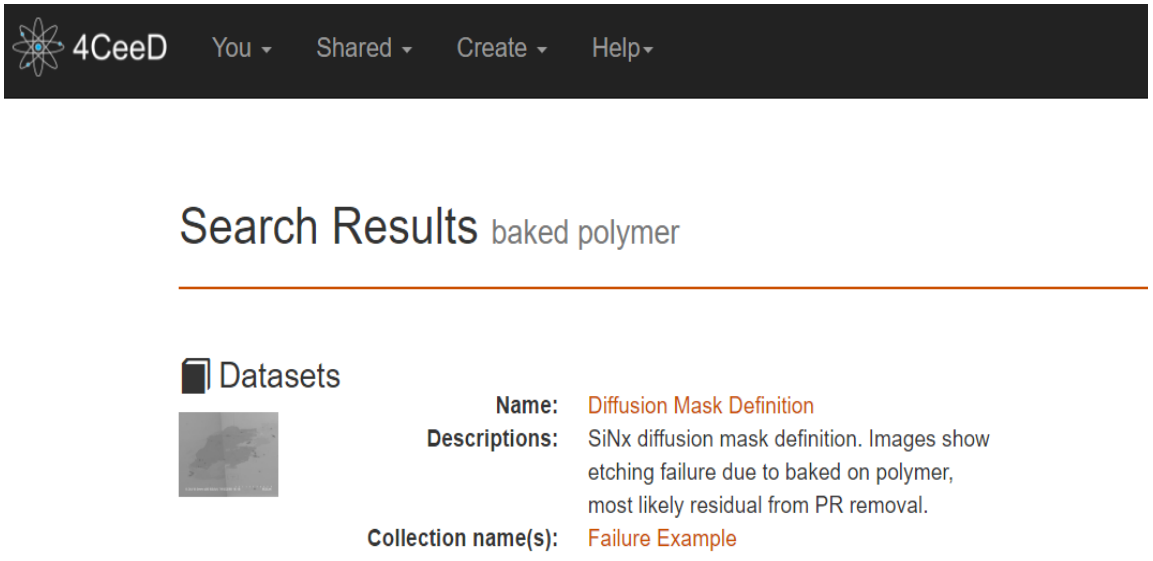


- Allows easy track of fabrication process flows
- Facilitates ramp-up for new members on projects/group
- Visual thumbnail image is an easy way to verify if their sample is on track

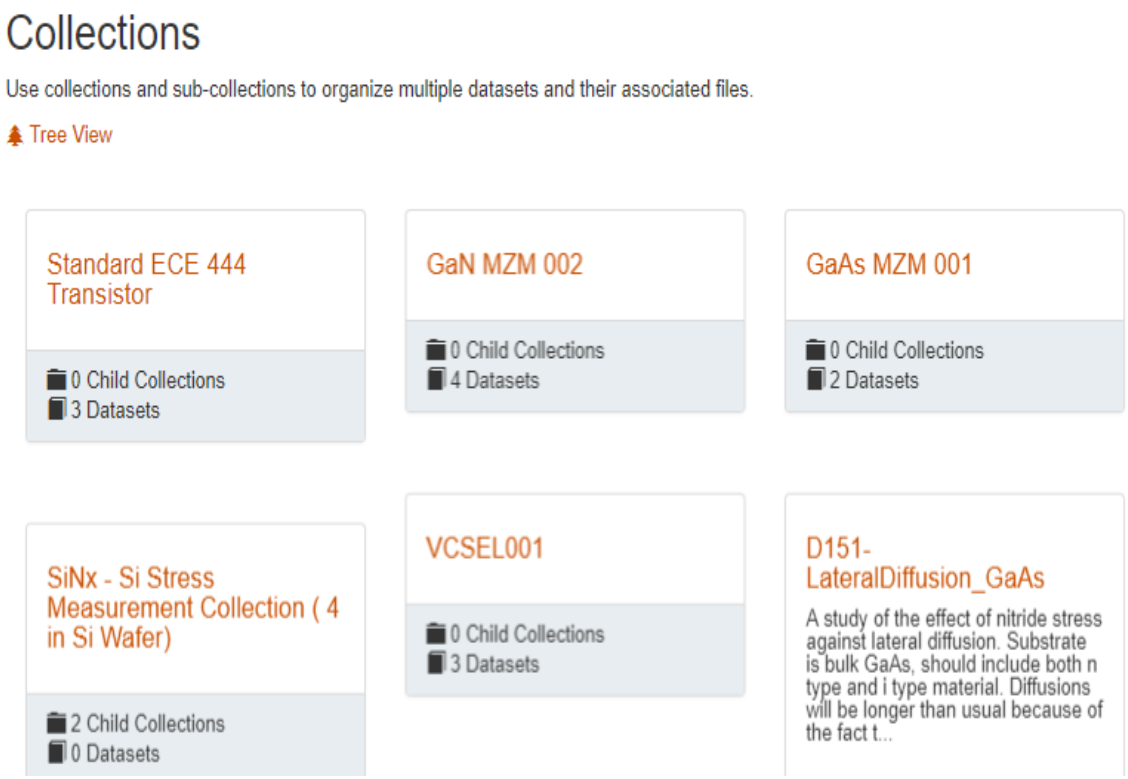
Usage Case: Process Analysis



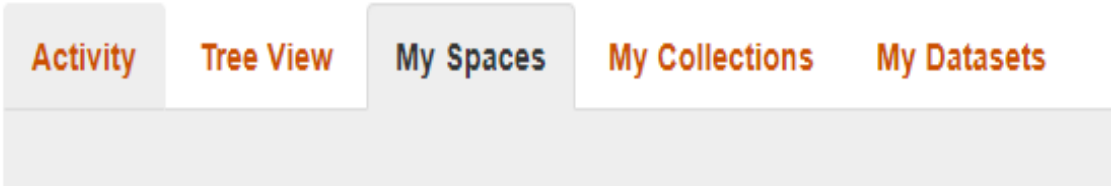
- 4CeeD digitalizes data that can be searched
- Facilitates failure diagnostics
- Data that extends back to the first user of 4CeeD
- Baked on polymer example



Usage Case: Device Archiving



- Easily store old device fabrication flows
- Can be referred to by newer generation of group members
- Personal File Hierarchy:
  Spaces: Group/Project
  Collections: Devices
  Dataset: Process Flow



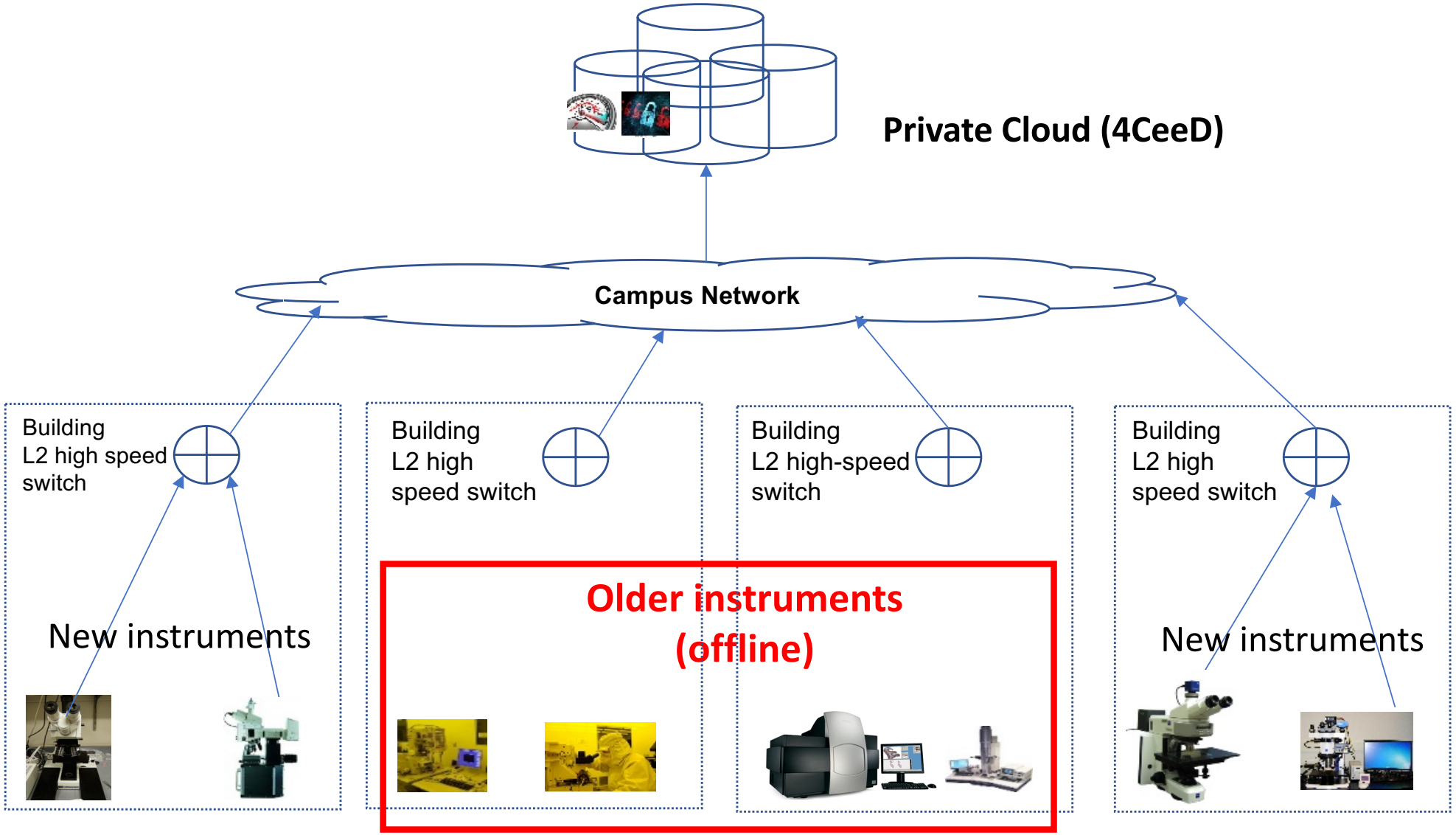
Group/Projects

Devices/Experiment

Process Flow/Experiment Step

BRACELET in 4CeeD

Current situation in campus cyber-infrastructure

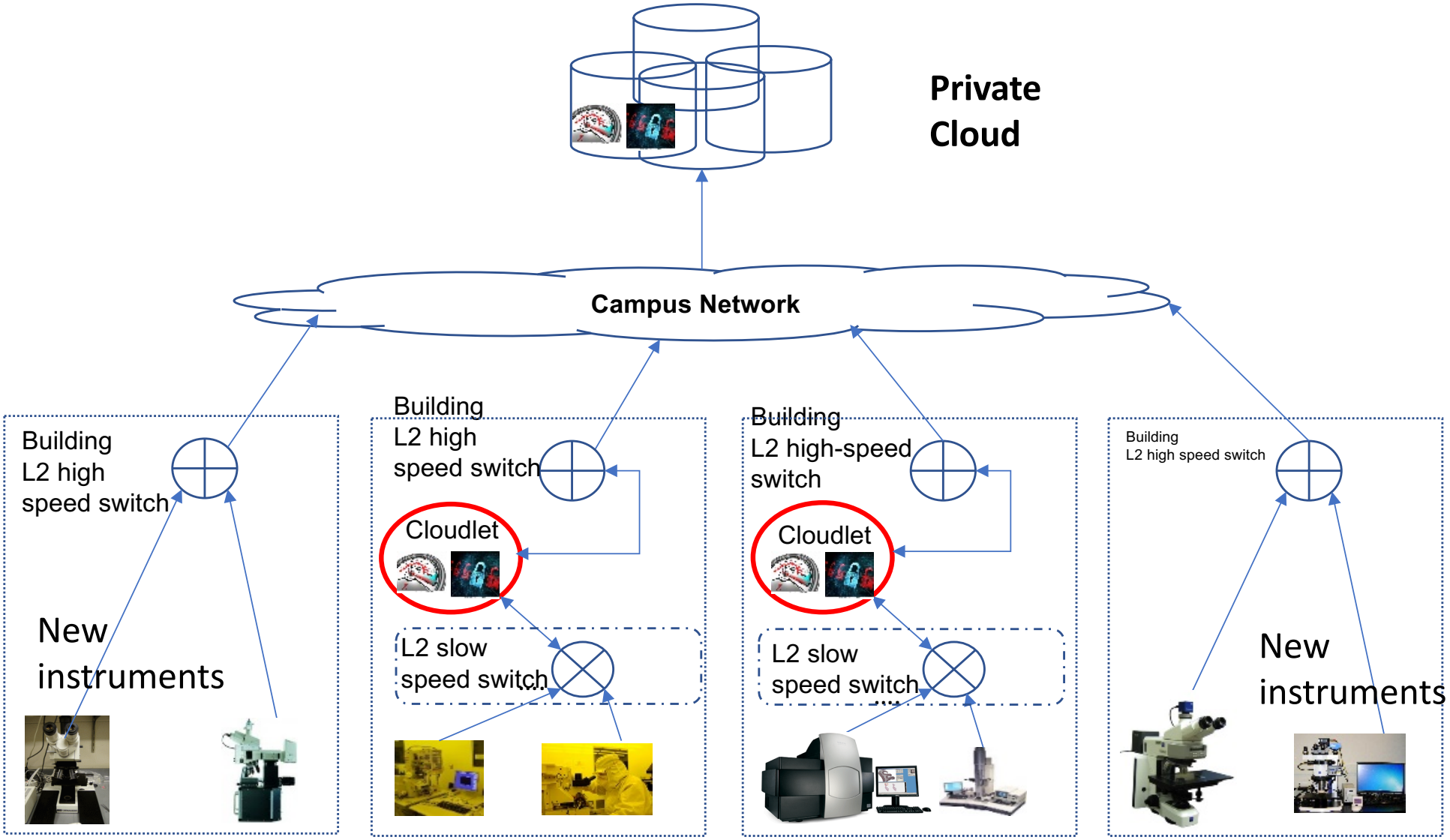


Current challenges with old instruments

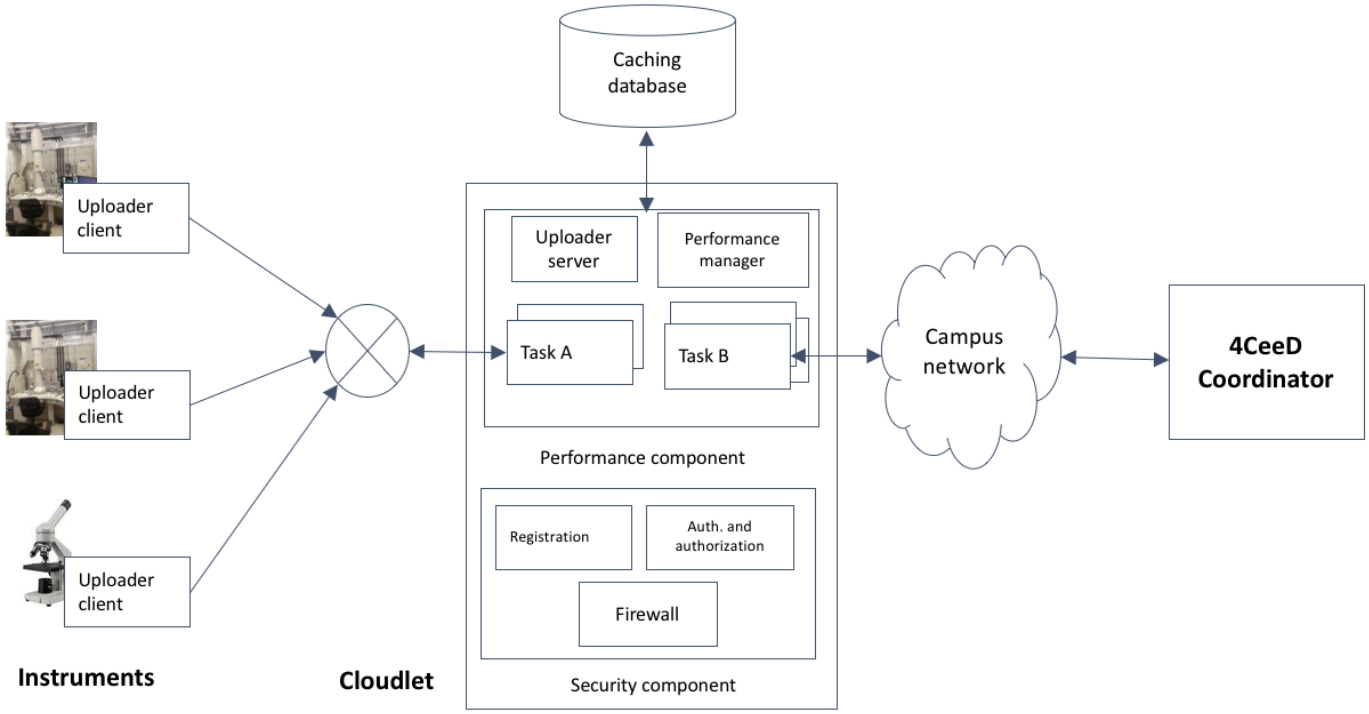
- **Performance mismatch**: Older instruments' Windows NT or XP runs network protocols at lower bandwidth speeds (10Mbps or 100Mbps)

- **Obsolete security**: Older devices and their OS systems cannot be patched, hence being vulnerable & taken offline

Approach: Putting edge device between older instruments and private cloud



Approach: Putting edge device between older instruments and private cloud



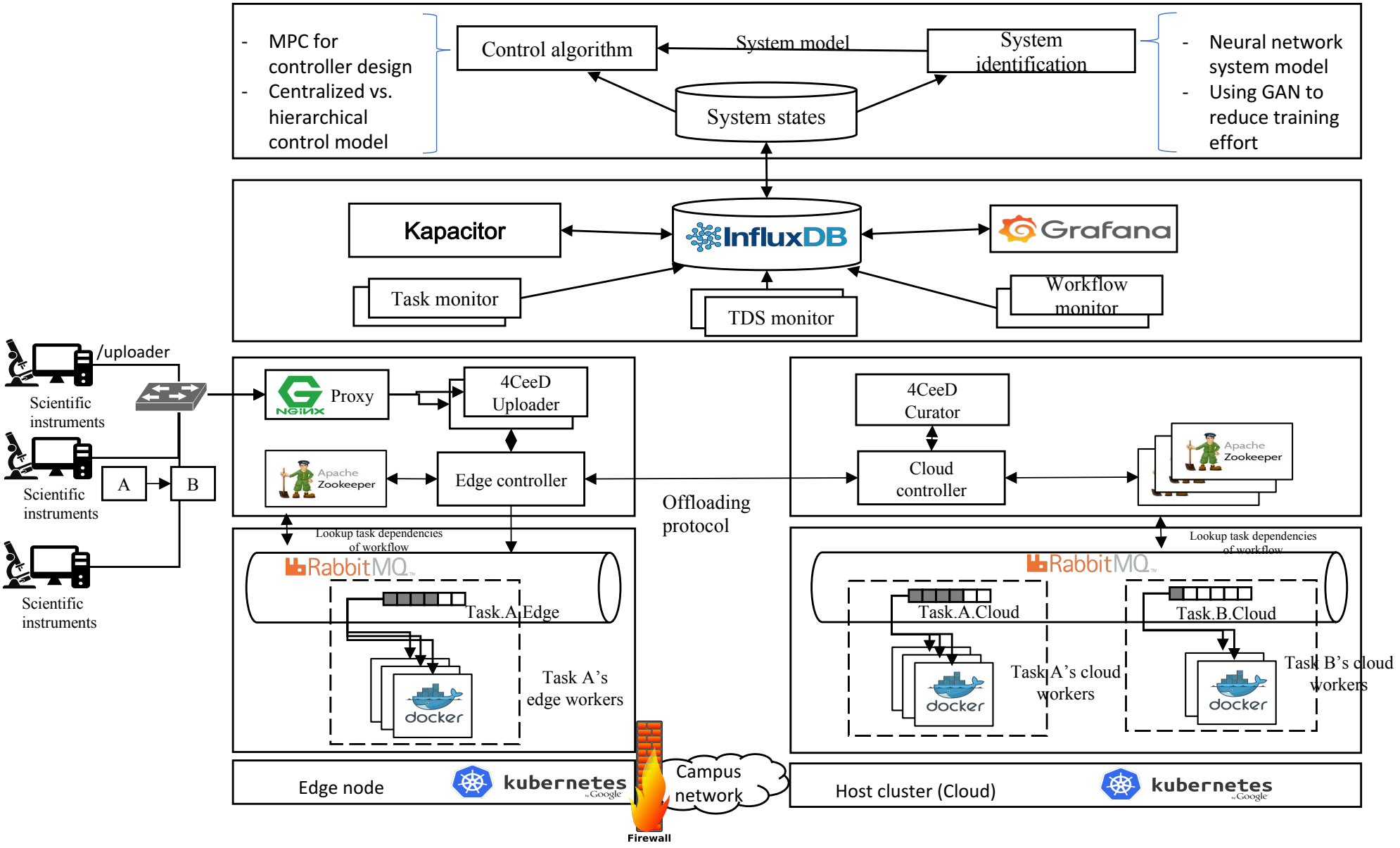
Performance:

- Have two network interfaces configured at different speeds
- Perform traffic shaping from slower to faster network

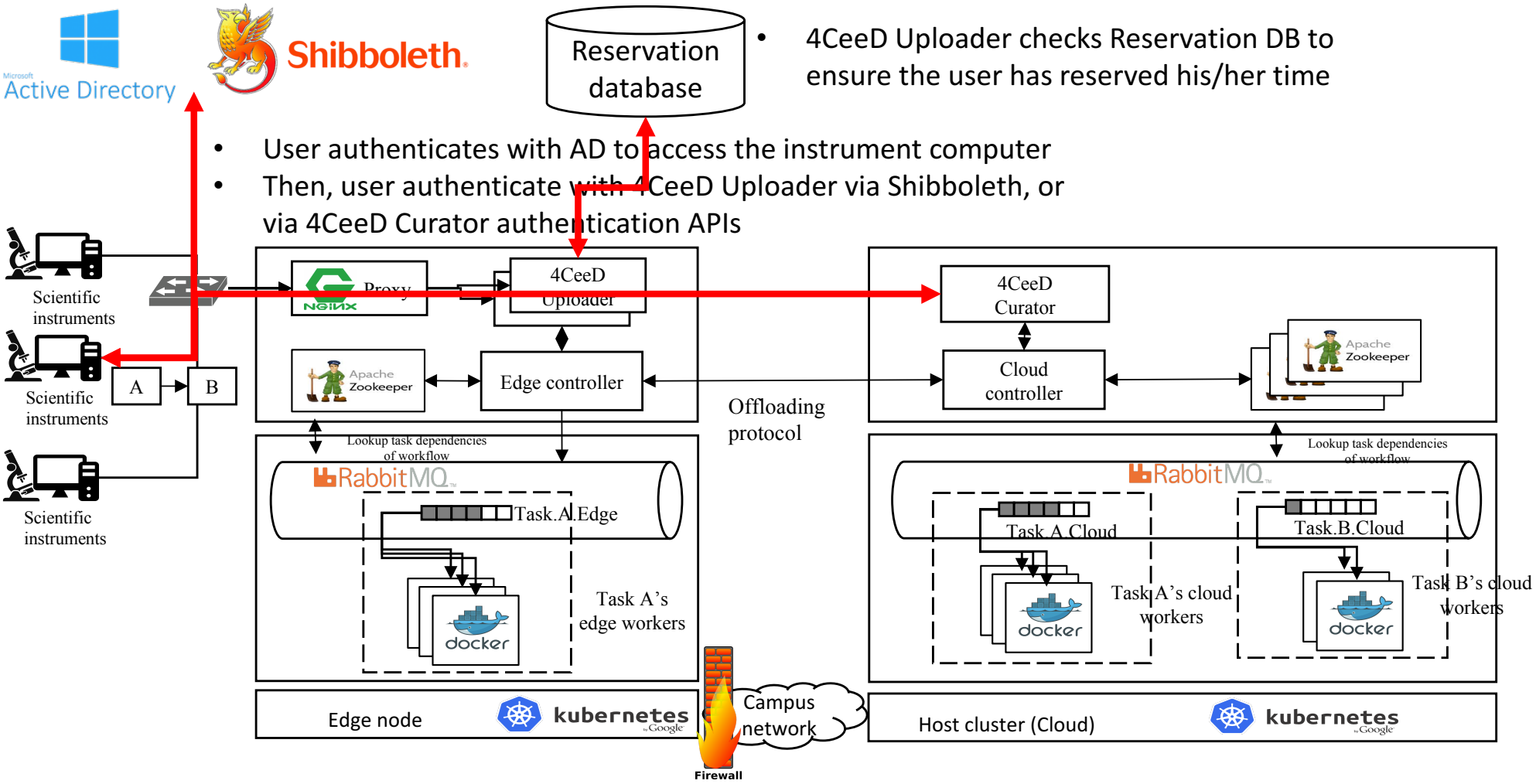
Security:

- User & instrument registration
- Data encryption during upload
- Firewall to protect against external threats

Cloudlet & private cloud integration



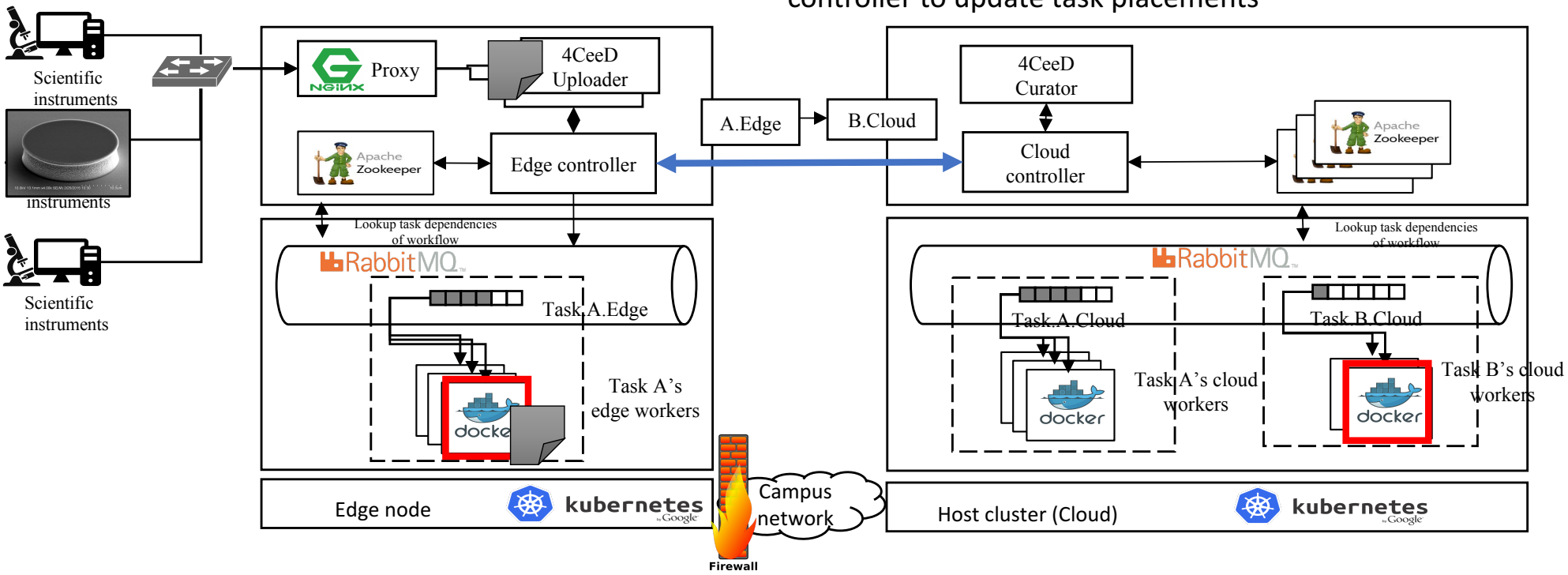
Cloudlet & private cloud integration (authentication)



Cloudlet & private cloud integration (data upload and processing)

- 4CeeD Uploader communicates with local Edge controller to learn about where to send request to
- Edge controller periodically communicates with cloud controller to update task placements

- After learning about the next task (following current placement)
- After processing the request, the task consumer forwards processing request to the next task



Conclusion – 4CeeD Advantages

- One stop location for research data
- Sharable between scientists, and her research partners
- Visual interpretation of data without having to open specialty software
- Cloud based → access data anywhere
- Available machine-specific meta-data on measurement for future replication of measurements
- Available templates which allow for the entry of metadata specific to scientist's project

Conclusions and impact

- *First* real-time data acquisition and cloud coordination framework for materials-related and semiconductor fabrication areas
- Major Savings:
  - **30% time saving** during MICROSCOPE experimental sessions (e.g., file transfer, save metadata, export previews, etc.)
  - At least **$30000 SAVINGS PER GRADUATE STUDENT** (initial estimates)
    - The time saving translates into **saving $25 to $30 each hour of lab session cost** and each grad student spends at **least 1000 hours in lab during his/her PhD time**.
  - Shorten time from digital capture to curation, interpretation & insights
    - Traditionally took **2-3 hours** to capture, upload and curate data via memory stick, notebook, sneaker net to upload data, now it takes **few seconds.**
  - Other benefits include better **data preservation, exploration, and security !**

References

- **“MONAD: Self-adaptive Micro-service Infrastructure for Heterogeneous Scientific Workflows”**
  Phuong Nguyen and Klara Nahrstedt
  The 14th IEEE International Conference on Autonomic Computing (ICAC) 2017, Columbus, OH, USA.
- **“4CeeD: Real-time Acquisition and Analysis Framework for Materials-related Cyber-Physical Environments”**
  Phuong Nguyen, Steven Konstanty, Todd Nicholson, Thomas O’brien, Aaron Schwartz-Duval, Timothy Spila, Klara Nahrstedt, Roy H. Campbell, Indranil Gupta, Michael Chan, Kenton McHenry, and Normand Paquin
  The 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid) 2017, Madrid, Spain.
- **“Resource Management for Elastic Publish Subscribe Systems: A Performance Modeling-based Approach”**
  Phuong Nguyen, Klara Nahrstedt
  The 9th IEEE International Conference on Cloud Computing (CLOUD) 2016, San Francisco, CA, USA.