TECH ex 22

Globus: To Compute and Beyond

Vas Vasiliadis
vas@uchicago.edu

December 6, 2022

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

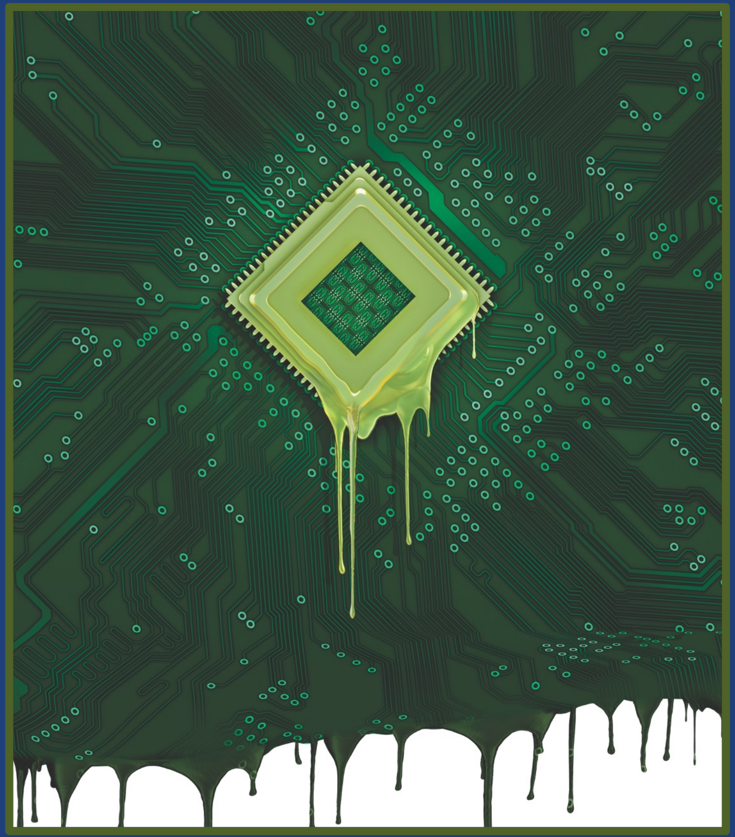
globus

Agenda

- Motivation
- Capabilities overview
- Current application examples
- Synergy with Globus data management
- → Discussion

General Purpose Computing is morphing…

“…the economic cycle that has led to the usage of a common computing platform, underpinned by rapidly improving universal processors, is giving way to **a fragmentary cycle, where economics push users toward divergent computing platforms driven by special purpose processors.”**

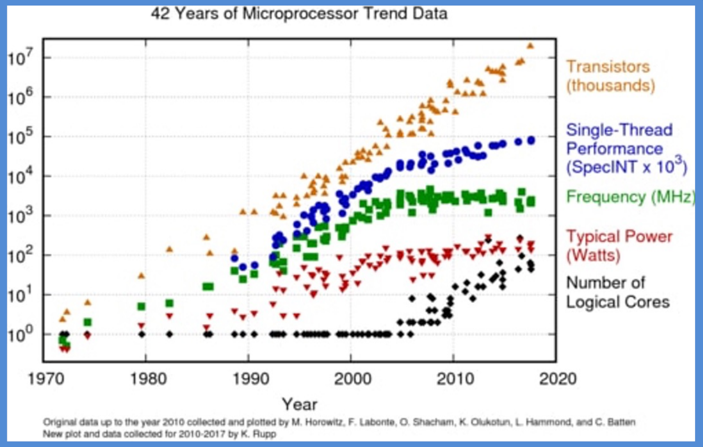


“The Decline of Computers as a General Purpose Technology”,
Thompson, N. & Spanuth, S., Communications of the ACM, March 2021

The research computing ecosystem is rapidly evolving

Resources

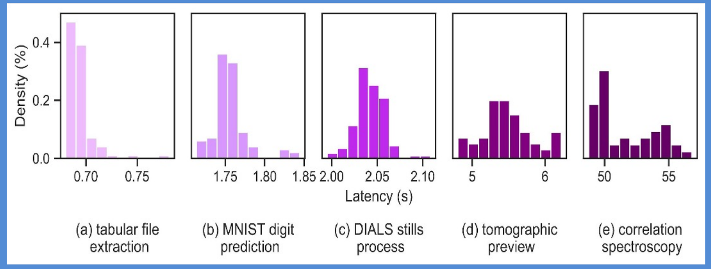
- Hardware specialization
- Specialization leads to distribution



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

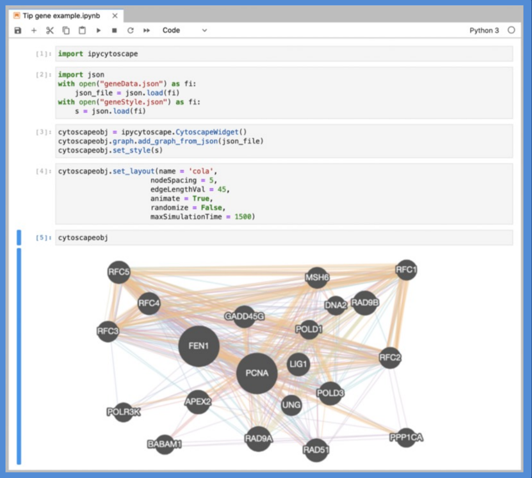
Workloads

- Interactive, real-time workloads
- Machine learning training and inference
- Components may best be executed in different places



Users

- Diverse backgrounds and expertise
- Different user interfaces (e.g., notebooks)



```
[1]: import ipycytoscape

[2]: import json
     with open("geneData.json") as fi:
         json_file = json.load(fi)
     with open("geneStyle.json") as fi:
         s = json.load(fi)

[3]: cytoscapeobj = ipycytoscape.CytoscapeWidget()
     cytoscapeobj.graph.add_graph_from_json(json_file)
     cytoscapeobj.set_style(s)

[4]: cytoscapeobj.set_layout(name = 'cola',
                             nodeSpacing = 5,
                             edgeLengthVal = 45,
                             animate = True,
                             randomize = False,
                             maxSimulationTime = 1500)

[5]: cytoscapeobj
```

Our data management legacy is also morphing…

- From fast, reliable, data transfer …
- … to secure data sharing …
- …and data management automation at scale
- But research flows inevitably include computation

→ Deliver the same “fire-and-forget” capabilities for computation as we do for data management

Why do we need managed computation?

- **Remote computing is notoriously complicated**
  - Authentication
  - Network connections
  - Configuring/managing jobs
  - Interacting with resources (waiting in queues, scaling nodes)
  - Configuring execution environment
  - Getting results back again
- **Researchers need to overcome the same obstacles every time they move to a new resource**

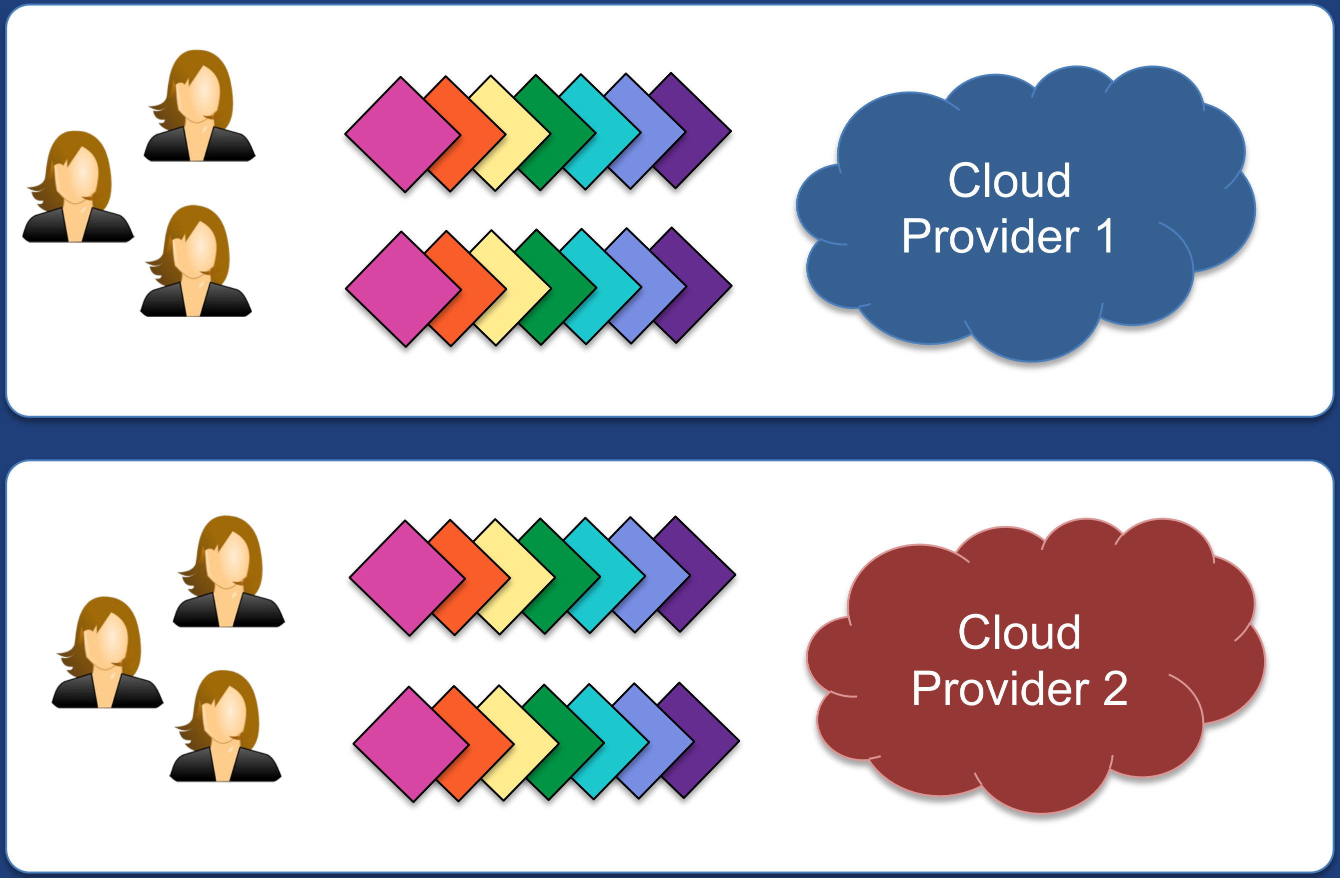
Does FaaS make sense for research?

- **Support new workloads by decomposing applications into functions**
  - Simplify development, maintenance, testing
- **Facilitate use of diverse compute resources**
  - Abstract heterogeneous compute infrastructure
- **Enable fluid function execution across the computing continuum**
  - Enable portability and sandboxing

Goal: Move closer to researchers’ environments

- **Researchers primarily work in high level languages**
- **Functions are a natural unit of computation**
- **FaaS allow researchers to work in a familiar language (e.g., Python) using familiar interfaces (e.g., Jupyter)**

FaaS as offered by cloud providers



- Single provider, single location to submit and manage tasks
- Homogenous execution environment
- Transparent and elastic execution (of even very small tasks)
- Integrated with cloud provider data management

FaaS as the interface to the research computing ecosystem?

NVIDIA
Argonne
ENERGY
intel
CRAY
Aurora

Still need…

- **Single interface**
- **Homogenous execution environment**
- **Transparent and elastic execution**
- **Integrated with data management**

funcX: Globus for compute

Managed, federated Functions-as-a-Service for reliably, scaleably and securely executing functions on remote endpoints from laptops to supercomputers

THE UNIVERSITY OF CHICAGO

ILLINOIS

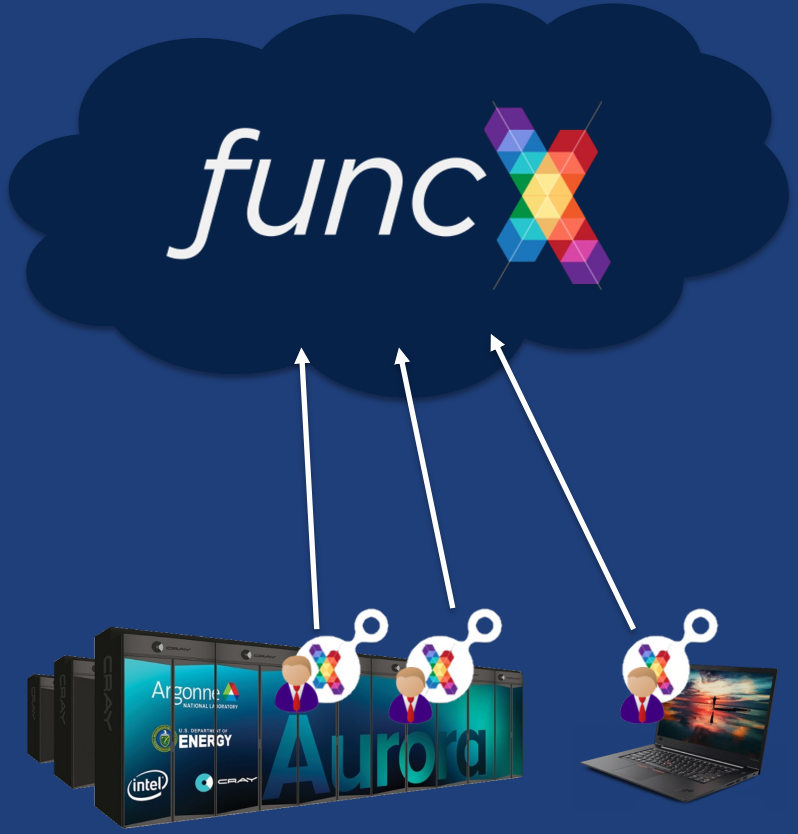
Argonne NATIONAL LABORATORY

The ***funcX*** model

- funcX service — Highly available cloud-hosted service; provides managed fire-and-forget function execution
- funcX endpoint — Abstracts access to compute resources (edge device to supercomputer)
- SDK — Python interface for interacting with the service; familiar Globus look and feel
- Security — Leverages Globus Auth; funcX endpoints are resource servers, authN and access via tokens

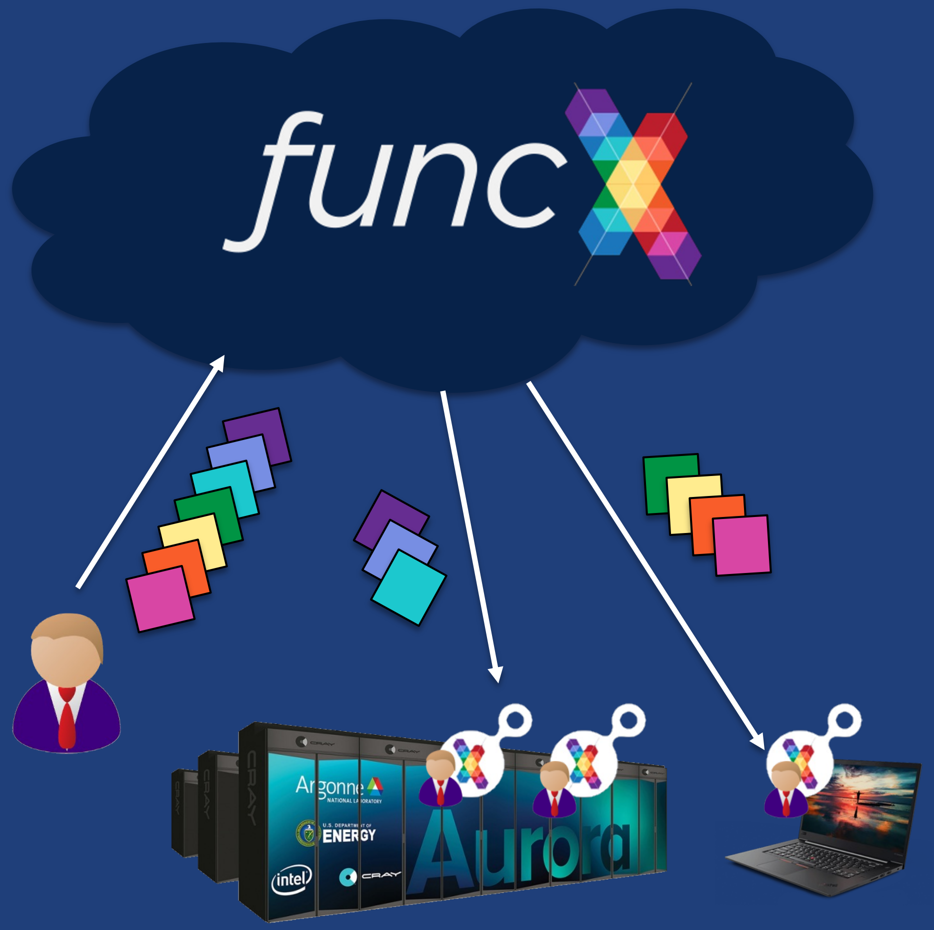
Transform laptops, clusters, clouds into function serving endpoints

- **`pip` installable endpoint**
  - Globus Auth for registration
- **Elastic resource provisioning from local, cluster, or cloud system (via Parsl)**
- **Parallel execution using local fork or via common schedulers**
  - Slurm, PBS, LSF, Cobalt, K8s
- **Optional managed execution in Docker, Singularity, Shifter containers**
- **Endpoint sharing with collaborators**



Executing functions with funcX

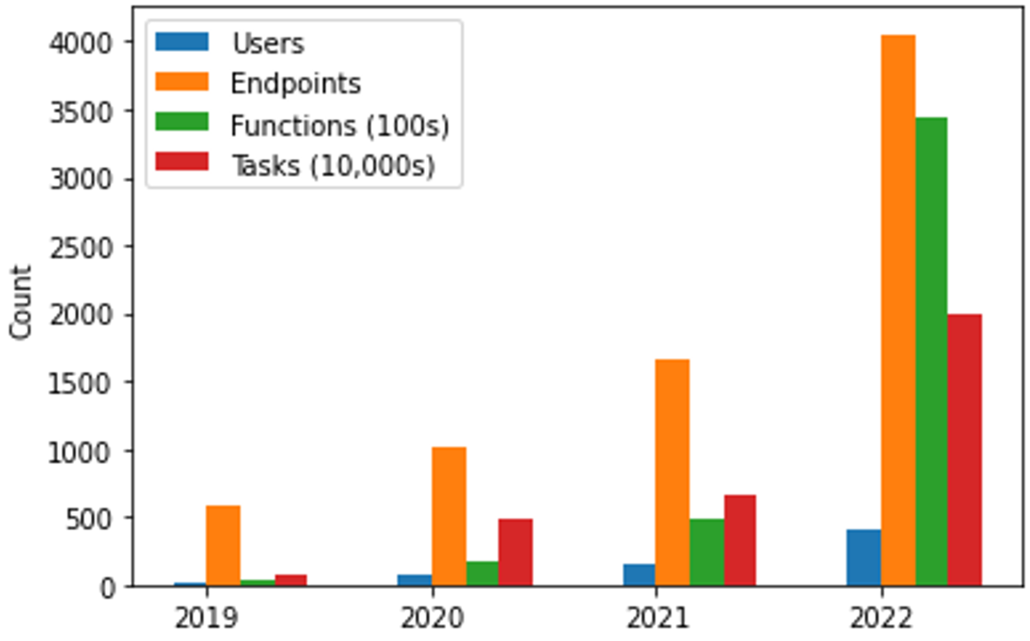
- **Users invoke functions as tasks**
  - Register Python function body
  - Pass input arguments
  - Select endpoint(s)
- **funcX stores tasks in the cloud**
- **Endpoints fetch waiting tasks (when online), run the task, and return the results (or errors)**
- **Results stored in the cloud; users retrieve results asynchronously**
- **Functions shareable with collaborators**



Usage is growing rapidly

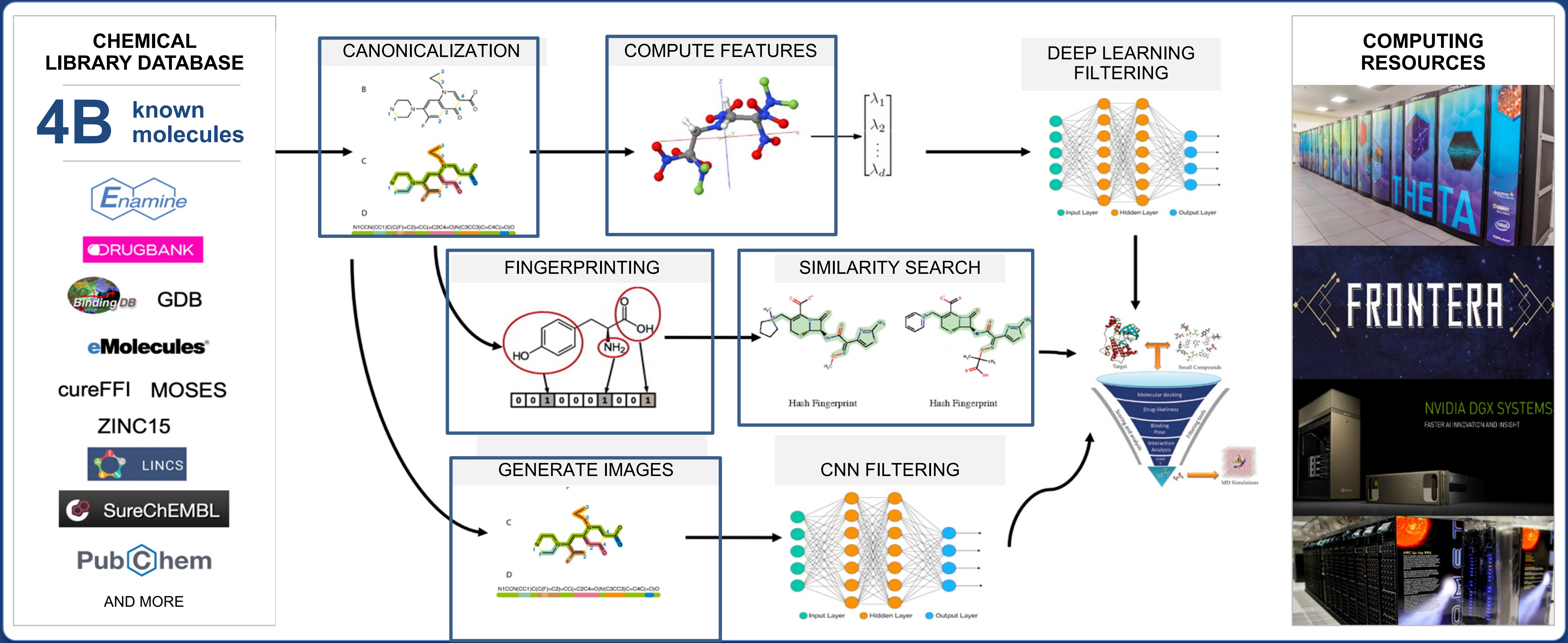
Early adopters generally in one of three categories:

- Remote (bag-of-tasks) execution
- Research automation
- Platform for building other services



>20M tasks, >340K functions, >4000 endpoints

Application: Using AI and supercomputers to accelerate drug development



CHEMICAL LIBRARY DATABASE

4B known molecules

Enamine

DRUGBANK

BindingDB GDB

eMolecules

cureFFI MOSES

ZINC15

LINCS

SureChEMBL

PubChem

AND MORE

CANONICALIZATION

COMPUTE FEATURES

DEEP LEARNING FILTERING

FINGERPRINTING

SIMILARITY SEARCH

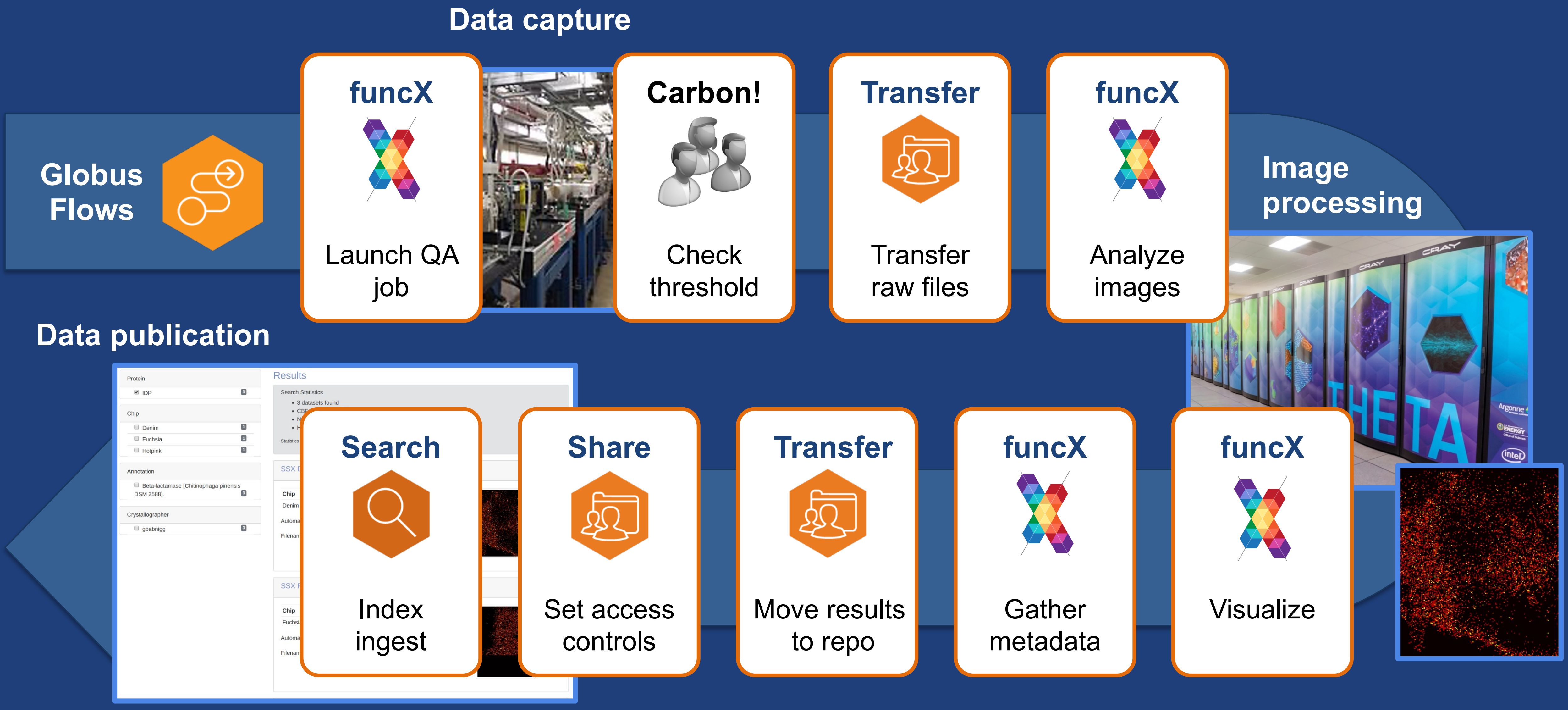
GENERATE IMAGES

CNN FILTERING

COMPUTING RESOURCES

Babuji, et al. Targeting SARS-CoV2 with AI- and HPC-enabled Lead Generation: A First Data Release.

Automation: Serial crystallography



Data capture

Globus Flows

funcX

Launch QA job

Carbon!

Check threshold

Transfer

Transfer raw files

funcX

Analyze images

Image processing

Data publication

Protein

☑ IDP 3

Chip

□ Denim 1

□ Fuchsia 1

□ Hotpink 1

Annotation

□ Beta-lactamase [Chitinophaga pinensis DSM 2588]. 3

Crystallographer

□ gbabnigg 3

Results

Search Statistics

- 3 datasets found
- CBF
- N
- H

Statistics

SSX D

Chip

Denim

Automa

Filenam

SSX F

Chip

Fuchsi

Automa

Filenam

Search

Index ingest

Share

Set access controls

Transfer

Move results to repo

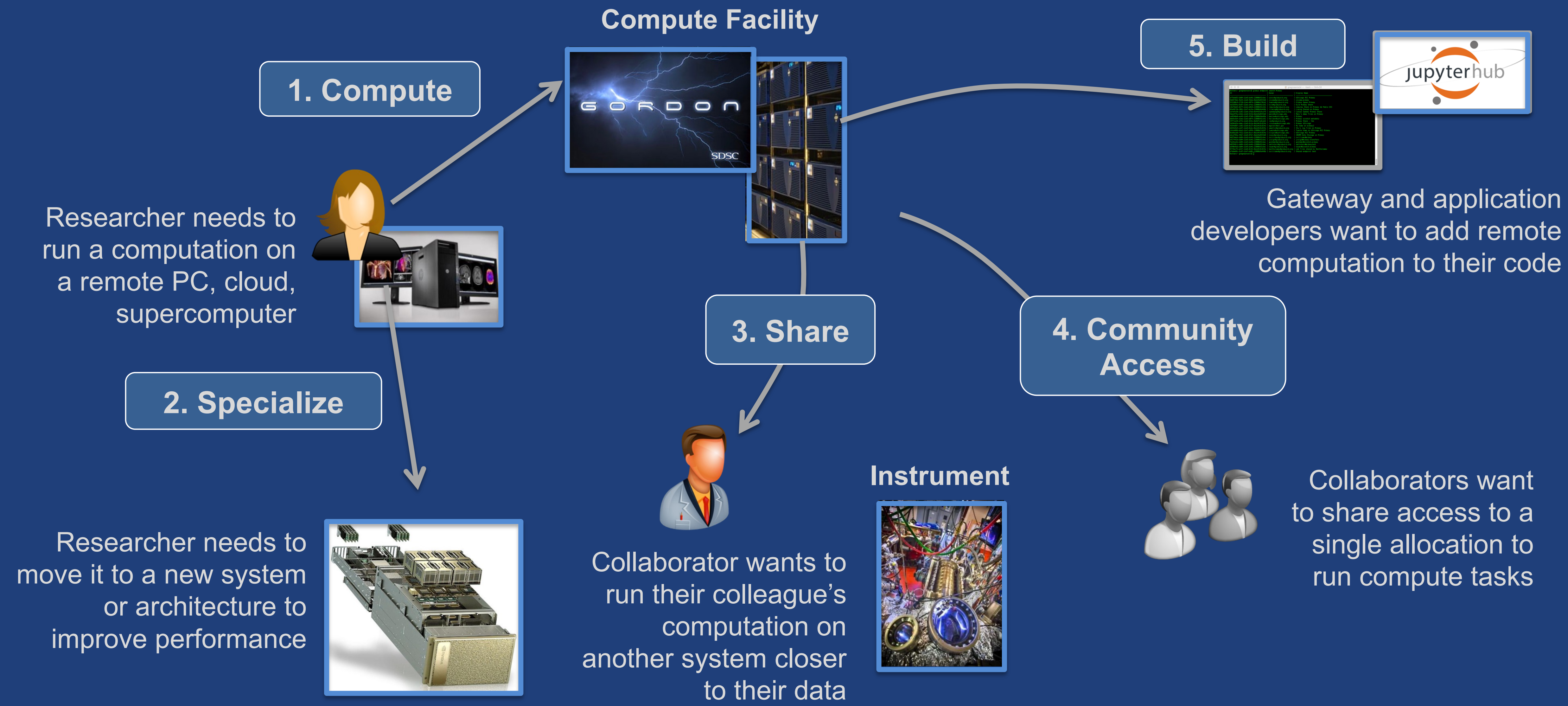
funcX

Gather metadata

funcX

Visualize

funcX as a platform for research computing



Compute Facility

1. Compute

Researcher needs to run a computation on a remote PC, cloud, supercomputer

2. Specialize

Researcher needs to move it to a new system or architecture to improve performance

3. Share

Collaborator wants to run their colleague's computation on another system closer to their data

Instrument

4. Community Access

Collaborators want to share access to a single allocation to run compute tasks

5. Build

Gateway and application developers want to add remote computation to their code

Discussion

- What use cases do you think funcX will be useful for?
- What barriers do you see to adoption?
- What questions would you want to ask before deploying?
- Imagine a world in which all computing resources had a funcX endpoint, what new use cases would be enabled?

Acknowledgements

- Kyle Chard
- Ryan Chard
- Yadu Babuji
- Zhuozhao Li
- Tyler Skluzacek
- Anna Woodard
- Ben Blaiszik
- Ben Galewsky,
- Josh Bryan
- Daniel S. Katz
- Ian Foster

Thank you, funders...

U.S. DEPARTMENT OF ENERGY

NSF

THE UNIVERSITY OF CHICAGO

NIH

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

ALFRED P. SLOAN FOUNDATION
1934

Argonne NATIONAL LABORATORY

powered by amazon web services