DEVELOPING DISTRIBUTED HIGH-PERFORMANCE COMPUTING CAPABILITIES OF AN OPEN SCIENCE PLATFORM FOR ROBUST EPIDEMIC ANALYSIS (OSPREY)

Jonathan Ozik[1,2]
Valerie Hayot-Sasson[1,2]

ParslFest 2023 Oct 19, 2023 – Chicago, IL

Work with: Nick Collier[1,2], Justin M. Wozniak[1,2], Abby Stevens[1,2], Yadu Babuji[1,2], Mickaël Binois[3], Arindam Fadikar[1,2], Alexandra Würth[3], Sudershan Sridhar[2], Kyle Chard[1,2]

1. Argonne, 2. UChicago, 3. Inria

Argonne NATIONAL LABORATORY

OUTLINE

- Why OSPREY?
- Prototype OSPREY Architecture
- Data Streams as a Service (DSaaS)

FUNDING ACKNOWLEDGEMENTS

- NSF 2200234: PIPP Phase I: Robust Epidemic Surveillance and Modeling (RESUME)
- DOE Office of Science through the Bio-preparedness Research Virtual Environment (BRaVE) initiative
- France and Chicago Collaborating in the Sciences (FACCTS): Developing Next Generation Tools for Large-scale Computational Science
- NIH R01DA055502: Leveraging data synthesis to identify optimal and robust strategies for HIV elimination among substance-using MSM

WHY OSPREY?

67 Days to Lockdown

From Chicago's first known COVID-19 case to the governor's historic shelter-in-place order, here's how those critical, frenzied early weeks of the pandemic unfolded, in the words of those at the center of the crisis.

BY MIKE THOMAS
ILLUSTRATION BY KEITH NEGLEY
MARCH 2, 2021, 8:00 AM

https://www.chicagomag.com/chicago-magazine/march-2021/67-days-to-lockdown/

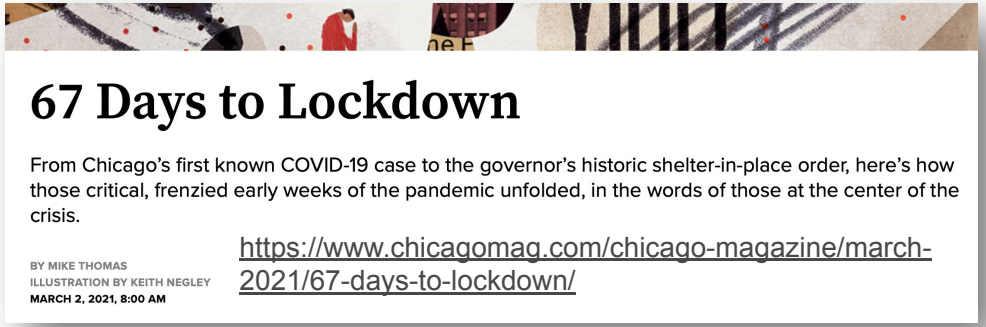
ARWADY
Commissioner of the Chicago Department of Public Health

Argonne National Lab built what's called an agent-based model specifically for Chicago that predated COVID and mimics the movement of its 2.7 million residents. All the gyms, schools, and businesses are built in, and it [takes into account] how people interact throughout the day. So pretty early on, we connected with them to start thinking about using that Chicago model for COVID.

MONDAY, MARCH 9

The State's Disaster Proclamation

One day after a Cook County man was announced as Illinois's seventh case and its first known instance of "community spread" — the spread of a disease to people who have had no known contact with other infected people — Pritzker issued a disaster proclamation that marshaled state resources to combat the virus. Four additional cases were revealed this day, all of them in Cook County. The next day saw eight more cases, including the first two outside Cook, bringing the state total to 19. The numbers were rising fast now, just as experts had predicted.



67 Days to Lockdown

From Chicago's first known COVID-19 case to the governor's historic shelter-in-place order, here's how those critical, frenzied early weeks of the pandemic unfolded, in the words of those at the center of the crisis.

BY MIKE THOMAS
ILLUSTRATION BY KEITH NEGLEY
MARCH 2, 2021, 8:00 AM

https://www.chicagomag.com/chicago-magazine/march-2021/67-days-to-lockdown/

ARWADY
Commissioner of the Chicago Department of Public Health

Argonne National Lab built what's called an agent-based model specifically for Chicago that predated COVID and mimics the movement of its 2.7 million residents. All the gyms, schools, and businesses are built in, and it [takes into account] how people interact throughout the day. So pretty early on, we connected with them to start thinking about using that Chicago model for COVID.

MONDAY, MARCH 9

Argonne National Lab built what's called an agent-based model specifically for Chicago that predated COVID and mimics the movement of its 2.7 million residents. All the gyms, schools, and businesses are built in, and it [takes into account] how people interact throughout the day. So pretty early on, we connected with them to start thinking about using that Chicago model for COVID.

WHY OSPREY?

- During COVID, individual research groups generally worked independently trying to exploit advances in HPC, data management, ML/AI, and automation methods to develop, calibrate, modify, verify, and validate their epidemiologic models.
  - Required large amounts of heroic, overlapping work
  - Risked lacking robustness, security, scalability, or efficiency
- Heterogeneous, changing, and incomplete data required complex integration across diverse and novel surveillance signals
  - Created significant challenges for use in epidemiologic workflows
- The Open Science Platform for Robust Epidemic analYsis (OSPREY) seeks to lower the barriers to and automate epidemiologic model analyses on HPC and cloud resources

WHY OSPREY?

- During COVID, individual research groups generally worked independently trying to exploit advances in HPC, ML/AI, and automation methods to develop, verify, and validate their epidemiologic models
  - Redundant, heroic, overlapping work
  - Risk of lacking robustness, security, scalability, or efficiency
- Heterogeneous, changing, and incomplete data required complex integration across diverse and surveillance signals
  - Created significant challenges to reuse in epidemiologic workflows
- The Open Science Platform for Robust Epidemic analYsis (OSPREY) seeks to lower the barriers to and automate epidemiologic model analyses on HPC and cloud resources

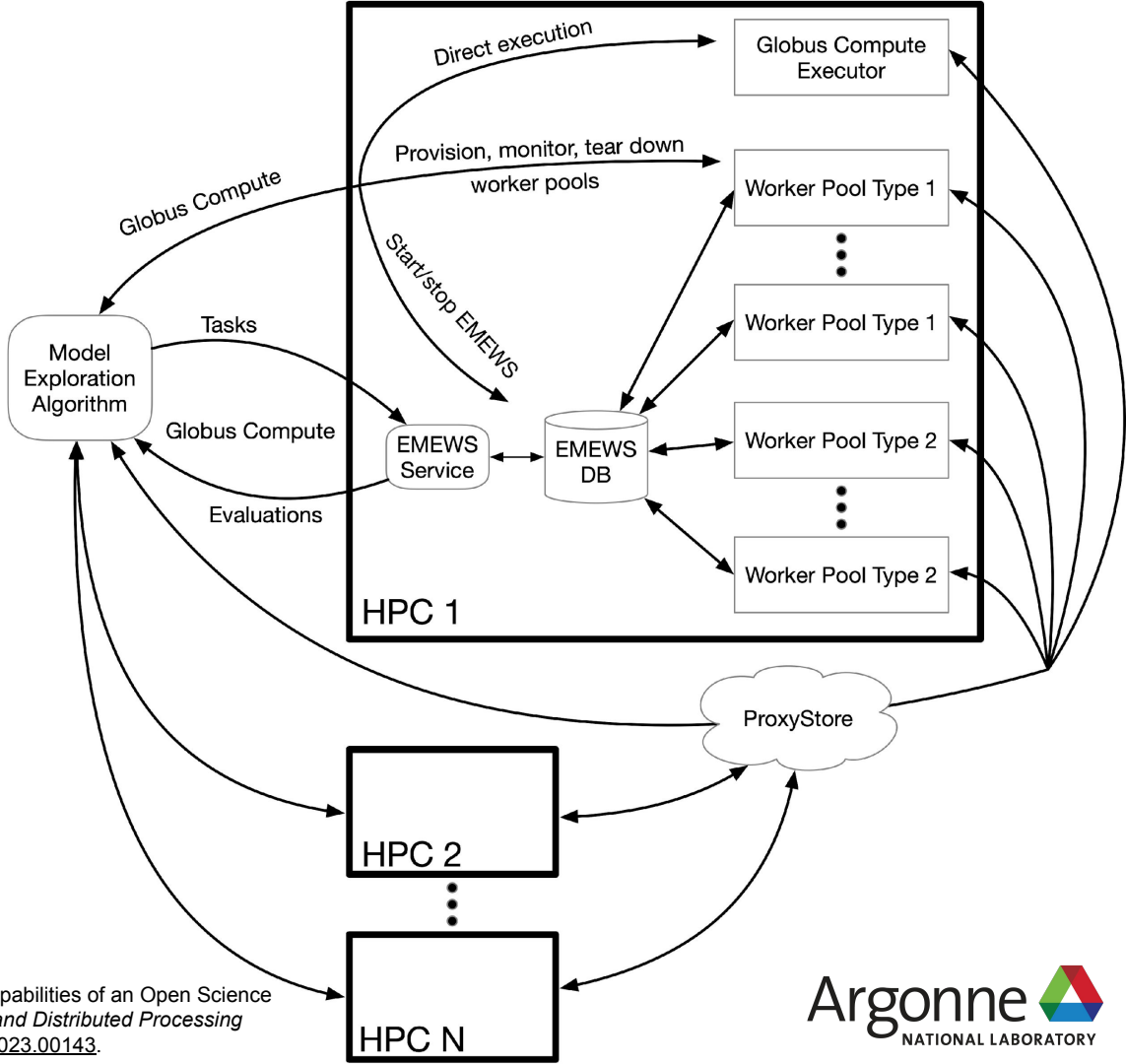
Prototype OSPREY Architecture

DSaaS

PROTOTYPE OSPREY ARCHITECTURE

PROTOTYPE OSPREY ARCHITECTURE

- The OSPREY architecture includes
  - Model exploration module
  - FaaS task distribution
  - EMEWS task DB
  - Scalable worker pools
  - Data sharing
- These collectively provide the ability to robustly and securely coordinate (epidemic) workflows across a distributed ecosystem of heterogeneous remote resources

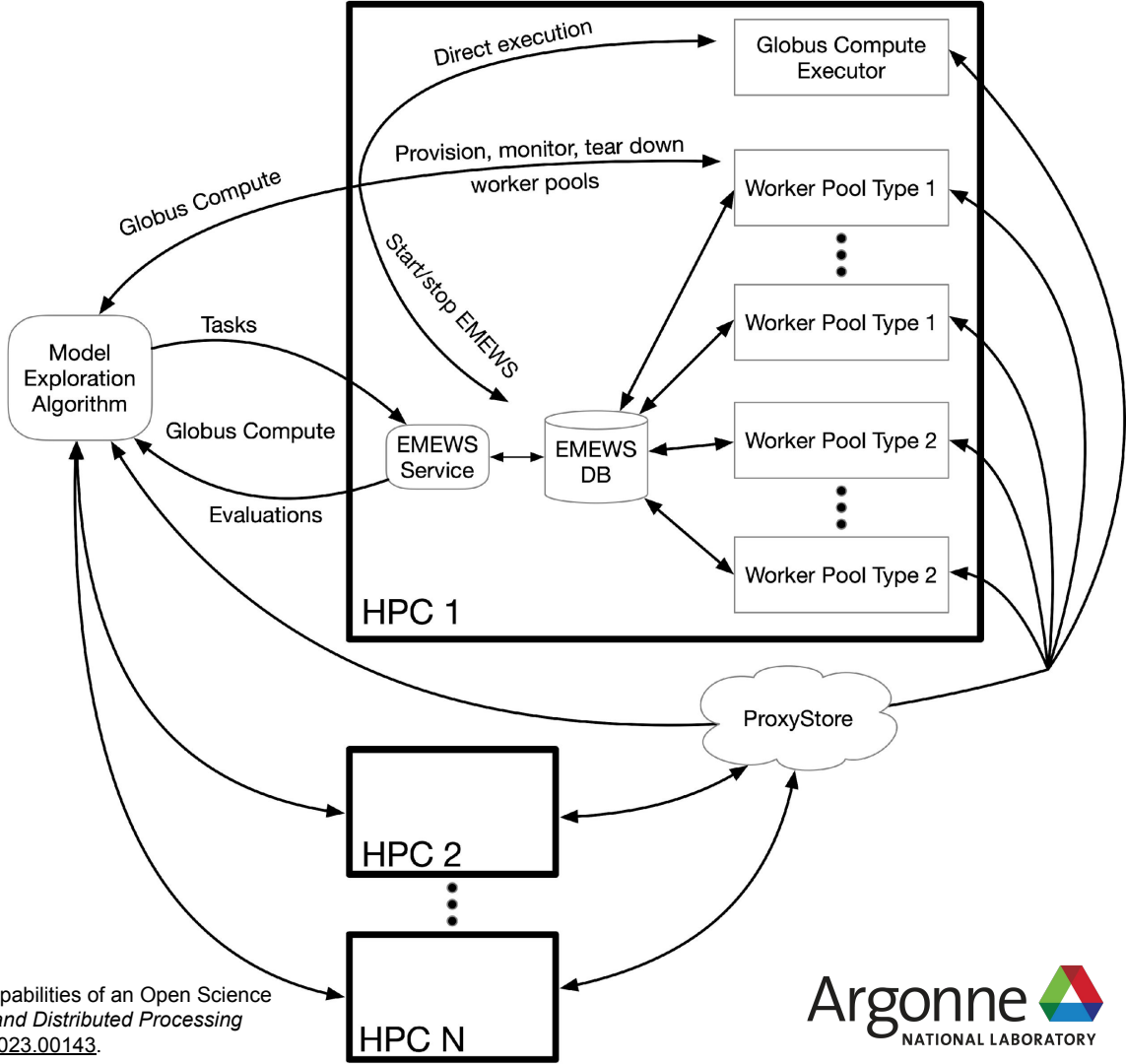


Collier et al. 2023. "Developing Distributed High-Performance Computing Capabilities of an Open Science Platform for Robust Epidemic Analysis." In *2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. https://doi.org/10.1109/IPDPSW59300.2023.00143.

Argonne NATIONAL LABORATORY

PROTOTYPE OSPREY ARCHITECTURE

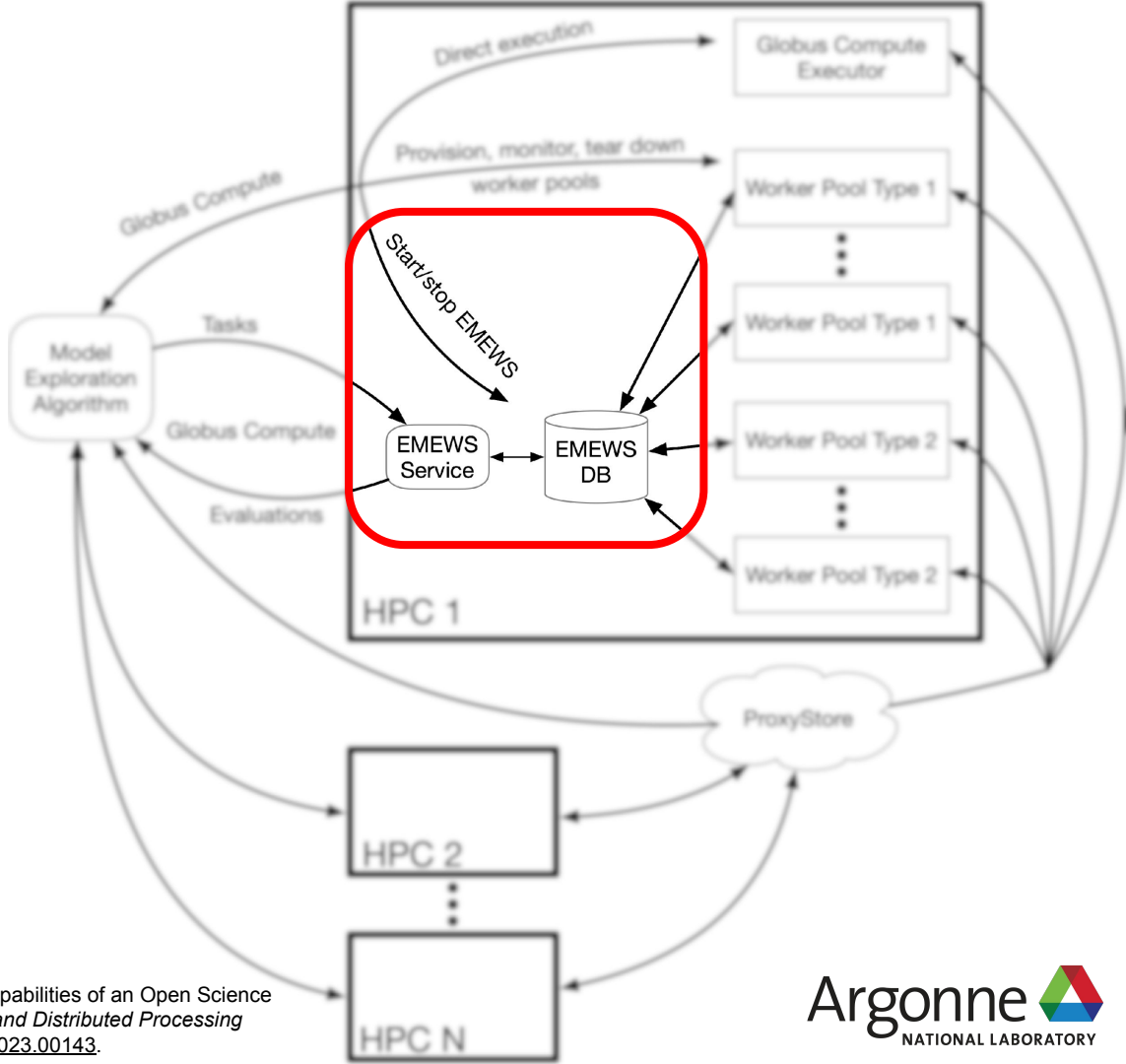
- We developed a task API, including remote worker and service monitoring and coordination
- Globus Compute provides:
  - Ability to Start/stop EMEWS Service/DB on remote machine
  - (with PSI/J) Ability to provision, monitor, tear down worker pools
  - Secure, long running access to EMEWS Service/DB



Collier et al. 2023. "Developing Distributed High-Performance Computing Capabilities of an Open Science Platform for Robust Epidemic Analysis." In *2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. https://doi.org/10.1109/IPDPSW59300.2023.00143.

PROTOTYPE OSPREY ARCHITECTURE

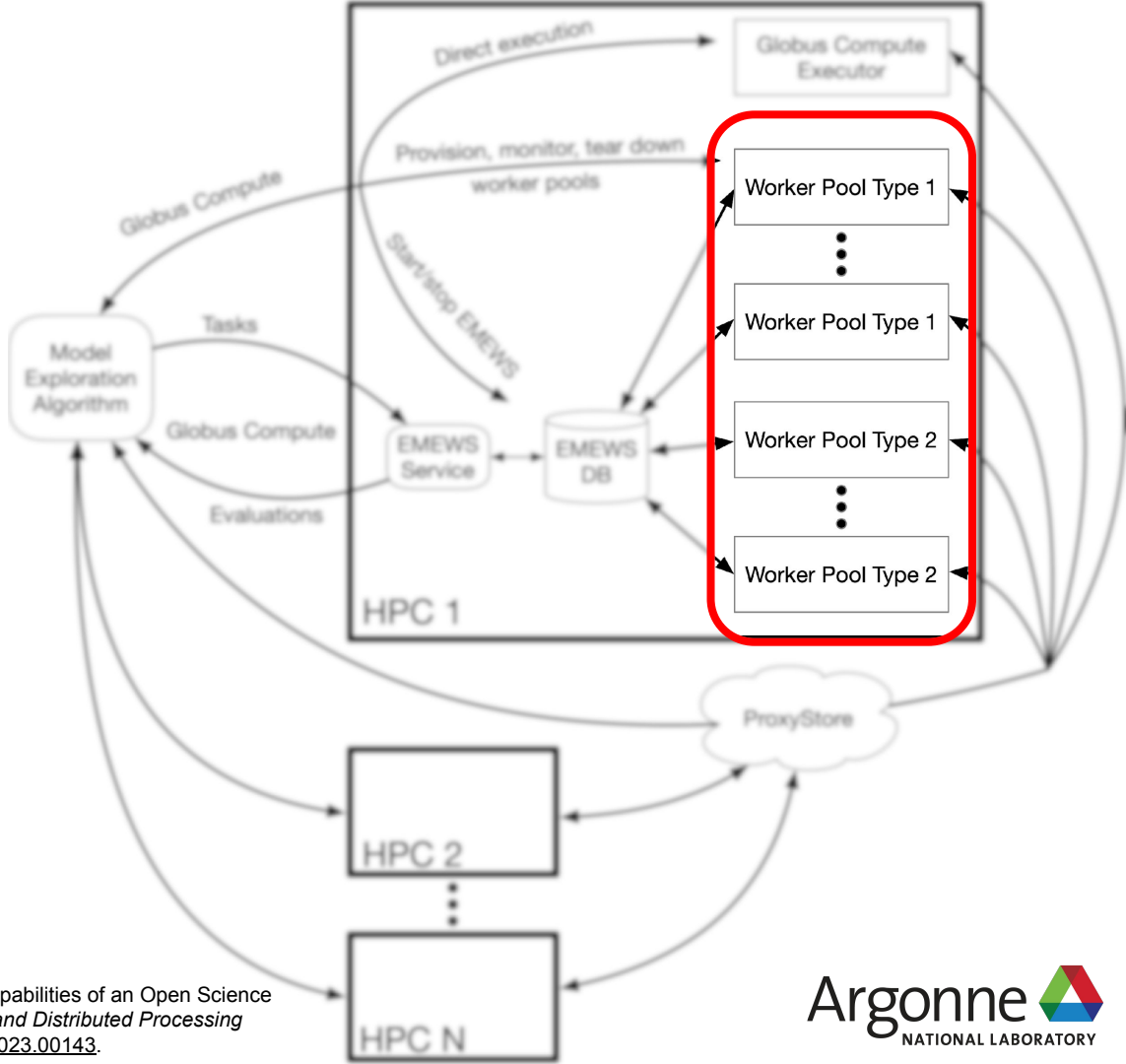
- We developed a task API, including remote worker and service monitoring and coordination
- Globus Compute provides:
  - Ability to Start/stop EMEWS Service/DB on remote machine
  - (with PSI/J) Ability to provision, monitor, tear down worker pools
  - Secure, long running access to EMEWS Service/DB



Collier et al. 2023. "Developing Distributed High-Performance Computing Capabilities of an Open Science Platform for Robust Epidemic Analysis." In *2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. https://doi.org/10.1109/IPDPSW59300.2023.00143.

PROTOTYPE OSPREY ARCHITECTURE

- We developed a task API, including remote worker and service monitoring and coordination
- Globus Compute provides:
  - Ability to Start/stop EMEWS Service/DB on remote machine
  - (with PSI/J) Ability to provision, monitor, tear down worker pools
  - Secure, long running access to EMEWS Service/DB



Collier et al. 2023. "Developing Distributed High-Performance Computing Capabilities of an Open Science Platform for Robust Epidemic Analysis." In *2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. https://doi.org/10.1109/IPDPSW59300.2023.00143.

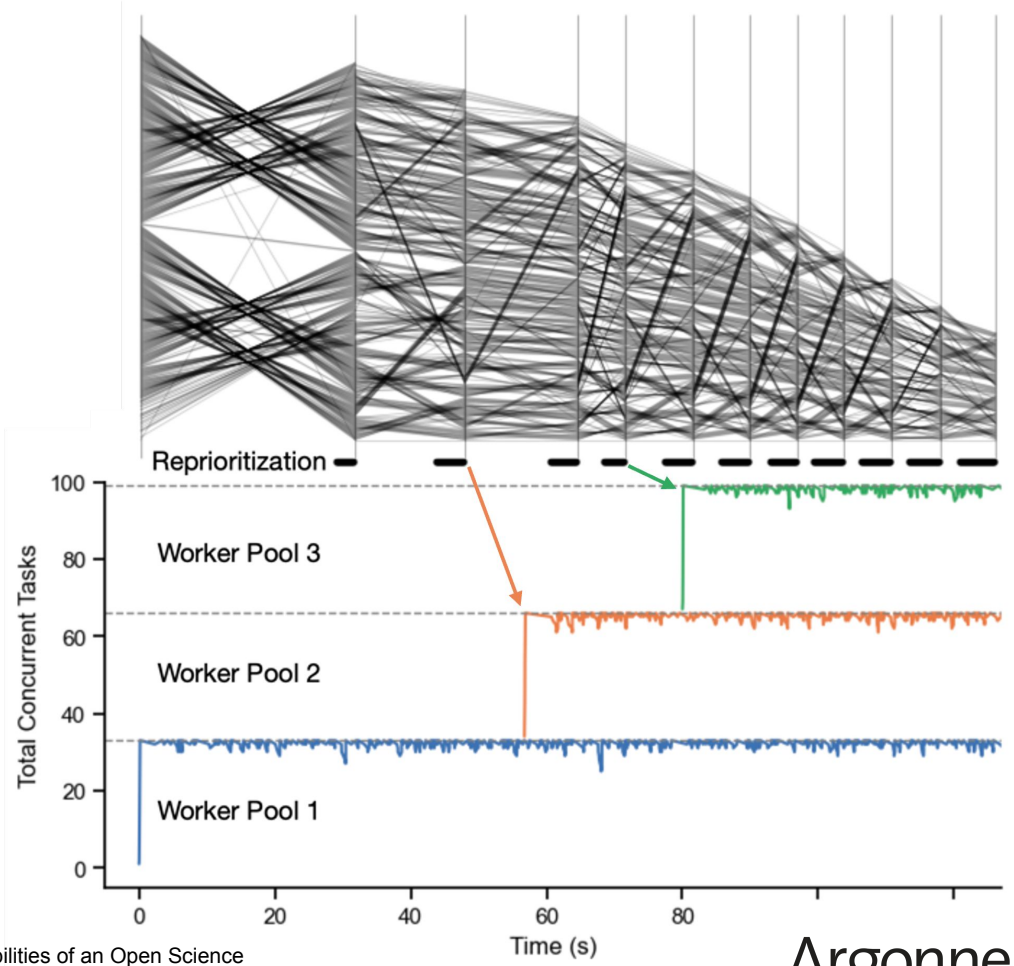
PROTOTYPE OSPREY ARCHITECTURE

- We developed a task API, including remote worker and service monitoring and coordination
- Globus Compute provides:
  - Ability to Start/stop EMEWS Service/DB on remote machine
  - (with PSI/J) Ability to provision, monitor, tear down worker pools
  - Secure, long running access to EMEWS Service/DB

Collier et al. 2023. "Developing Distributed High-Performance Computing Capabilities of an Open Science Platform for Robust Epidemic Analysis." In *2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. https://doi.org/10.1109/IPDPSW59300.2023.00143.

PROTOTYPE OSPREY ARCHITECTURE

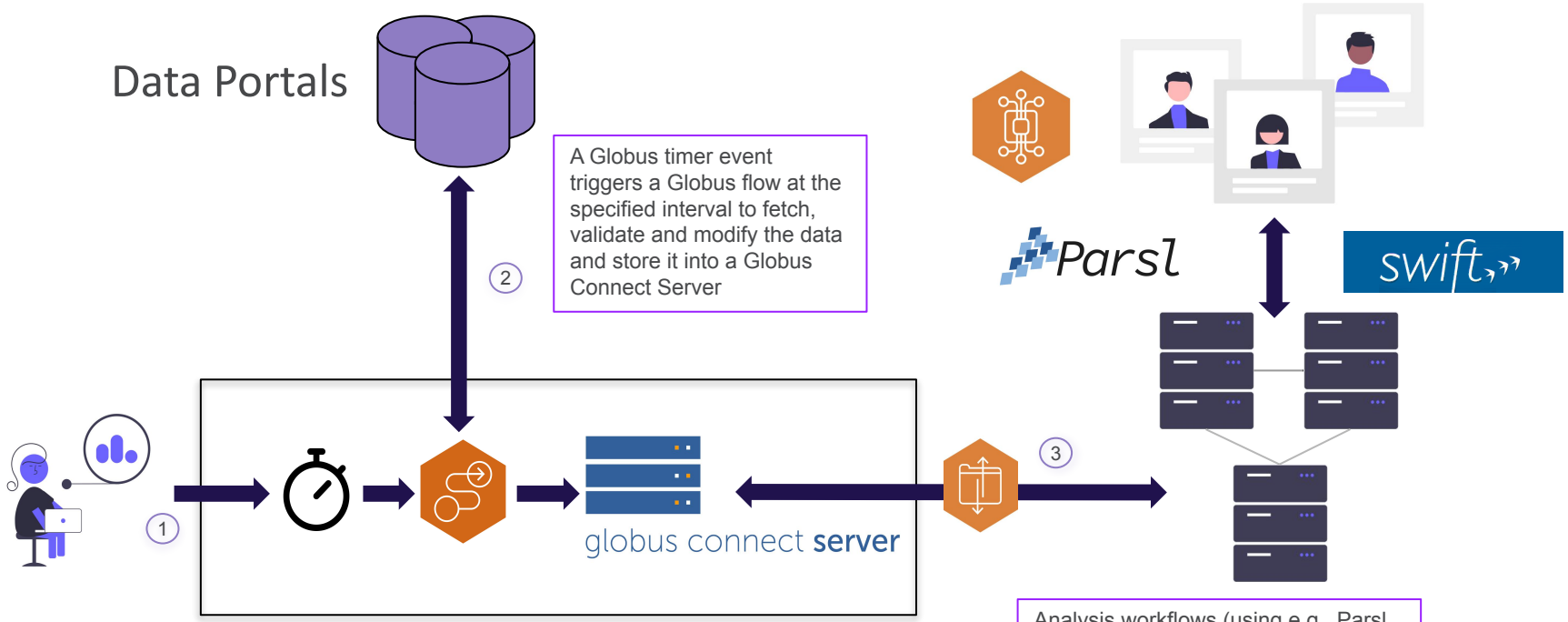
- We developed a task API, including remote worker and service monitoring and coordination
- Globus Compute provides:
  - Ability to Start/stop EMEWS Service/DB on remote machine
  - (with PSI/J) Ability to provision, monitor, tear down worker pools
  - Secure, long running access to EMEWS Service/DB



Collier et al. 2023. "Developing Distributed High-Performance Computing Capabilities of an Open Science Platform for Robust Epidemic Analysis." In *2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. https://doi.org/10.1109/IPDPSW59300.2023.00143.

DATA STREAM AS A SERVICE (DSaaS)

DATA STREAM AS A SERVICE (DSaaS) OVERVIEW



USING DSAAS IN PARSL FUNCTIONS

- Source data should be added to the service outside of Parsl via the CLI

```
dsaas -create_source -n 'covid19' -u
https://data.cityofchicago.org/resource/naz8-j4nc.json -t 86400 -v
'df2ab7fc-dad7-440f-a9cb-e02dbeac4141'
```

- Data can be accessed within a function using the DSaaS client API (WIP)

```
@python_app
def analysis(source_id=1):
    from dsaas.client import source_file

    df = source_file(source_id=1) # returns pandas dataframe. Transfer via Globus Transfer
    …
```

- Future work: analysis outputs can be stored within DSaaS

```
from dsaas.client import save_analysis

save_analysis(df, name='parslfest-analysis.json', version=2)
```

THANK YOU FOR YOU ATTENTION!
QUESTIONS?

JOZIK@ANL.GOV
VHAYOT@UCHICAGO.EDU