

CaRCC Researcher-Facing Call, 2024-11-14

Title/Topic

Research Process Automation with Globus

Speakers

Vas Vasiliadis, Chief Customer Officer, Globus

Abstract

The Globus platform offers reliable and secure data management and remote computation, as well as a managed task orchestration system for automation. In this session we will provide an overview of Globus Compute, a federated FaaS platform designed for research computing, and Globus Flows, which provides secure, managed automation of complex workflows. Together, Globus Compute and Flows enable users to construct data processing pipelines that are reliably managed and executed by Globus.

Please Note:

- ***We will record this call and post shortly thereafter on CaRCC's YouTube channel.***
- ***We expect all persons on the call to adhere to [CaRCC's Code of Conduct](#).***

[Connection Details](#)

[Agenda](#)

[Announcements](#)

[Sign-In \(Name / Affiliation /Email\)](#)

[Notes from the call](#)

[Slides](#)

[Notes](#)

[Questions](#)

[Chat Comments](#)

Connection Details

<https://utah.zoom.us/my/carcc?pwd=TjFuR3VVM2d5eE5zWnEvWWxDTFBCUT09>

Meeting ID: 824 051 8198

Password: 31415926

One tap mobile

+13462487799,,8240518198#,,#31415926# US (Houston)

+16699006833,,8240518198#,,#31415926# US (San Jose)

Dial by [your location](#)

+1 346 248 7799 US (Houston)

+1 669 900 6833 US (San Jose)

Join by Skype for Business: <https://utah.zoom.us/skype/8240518198>

Agenda

Time	Item
12:55-1:05 pm EST	<p>Intro slides: https://docs.google.com/presentation/d/19kG20KMFOsbN4gs5AiqU41jhglBLXKWvNLLN9fAhbMs/edit?usp=sharing</p> <p>Welcome to the Researcher Facing Track!</p> <ul style="list-style-type: none">• Researcher-Facing Track description• CaRCC is your community for research computing and data professionals. Please see this brief overview; more information on activities on our Groups web pages.• If you have questions about CaRCC or are interested in becoming more involved, please contact:<ul style="list-style-type: none">○ rf-coordinators@carcc.org for R-F-related activities, or○ getstarted@carcc.org or getinvolved@carcc.org for other CaRCC work• We expect all persons on the call to adhere to CaRCC's Code of Conduct.
1:05 -1:30 pm EST	Presentation from Globus
1:30-1:55 pm EST	Q&A
1:55 - 2:00 pm EST	Session close

Talk Slides:

Announcements

- **CaRCC People Network [Events](#)**
- **SC24 - Atlanta, GA Nov 17-22**
- **Internet2 2024 Technology Exchange - Dec 9th**
- **December R-F Call in CaRCC Party - Dec 12th**

Sign-In (Name / Affiliation /Email)

Note: please follow the suggested sign-in format so our evolving data science intelligence routines won't trip up and forget to enter you in our \$1m sweepstakes.

1. Justin Booth / Michigan State / boothj@msu.edu
2. Richard Lawrence / Texas A&M / rlawrence@tamu.edu
3. Jacalyn Huband / University of Virginia / jmh5ad@virginia.edu
4. Andrew Strumpf / University of Virginia / as9qn@virginia.edu
5. James Kelly / Chapman University / jakelly@chapman.edu
6. Vas Vasiliadis / University of Chicago / vas@uchicago.edu
7. Zane Gray / University of Oklahoma / zgray@ou.edu
8. Doug Dodson / Penn State / sxd123@psu.edu
9. Anna Alber/ Chapman University/ alber@chapman.edu
10. Aaron Wemhoff / Villanova University / aaron.wemhoff@villanova.edu
11. Grigory Shamov / University of Manitoba / grigory.shamov@umanitoba.ca
12. Bobby Roybal / University of Nevada, Las Vegas / bobby.roybal@unlv.edu
13. Sarah Ghazanfari/ Virginia Tech/ sarahghazanfari@vt.edu
14. Elizabeth Kwon / Columbia University / ek2864@columbia.edu
15. Alberto Cano / Virginia Tech / acano@vt.edu
16. Robben Migacz / University of Utah / robben.migacz@utah.edu
17. Paul DiBello/ Harvard Business School / pdibello@hbs.edu
18. Jan Day / Amazon Web Services / janday@amazon.com
19. Bryan Raney / Rice University / bryan.raney@rice.edu
20. Ken Taylor / University of Illinois / ktaylo@illinois.edu
21. Jessica Eaton / Columbia University / j.eaton@columbia.edu
22. Calvin D. Cox / Harvard Medical School / calvin_cox@hms.harvard.edu
23. Arun Seetharam/ Purdue University / aseethar@purude.edu
24. Jessica Pierce / Harvard Medical School / jessica_pierce@hms.harvard.edu
25. Patrick Schmitz / Semper Cogito / patrick@sempercogito.com
26. Michael Gutteridge / Fred Hutchinson Cancer Ctr. mrg@fredhutch.org
27. Martin Cuma / University of Utah / m.cuma@utah.edu
28. Matt Gregas, Boston College, gregas@bc.edu
29. Jim Leous / consultant / jim.leous@gmail.com
30. Ahmad Sheikhzada/University of Virginia/jus2yw@virginia.edu
31. Maze Ndukum/Washington University in St Louis/ndukummaze@wustl.edu

32. Paula Sanematsu / Harvard University, FASRC / paula_sanematsu@g.harvard.edu
33. Sarah Marchese / Harvard University / sarah_marchese@fas.harvard.edu
34. Prentice Bisbal / NSF NCAR / prentice@ucar.edu
35. Robert McDermott, Fred Hutch Cancer Center, rmcdermo@fredhutch.org
36. Maggie McFee / Harvard FASRC - mmcfee@g.harvard.edu
37. Jeff Dusenberry / UMass Boston / jeff.dusenberry@umb.edu
38. Nandan Tandon/ University of Central Florida/ Nandan.tandon@ucf.edu
39. Sean Smith / Rice University / mrsmith@rice.edu
40. Rachel Lombardi / Georgia Tech / rlombardi6@gatech.edu
41. Ben Eisenbraun / Harvard Medical School / bene@hkl.hms.harvard.edu
42. Yun Shen / Boston University / yshen16@bu.edu
43. Shawn Sivy / The College of New Jersey / ssivy@tcnj.edu
44. Zhiyu Li / University of Utah / zhiyu.li@utah.edu
45. Susan Tussy / University of Chicago, Globus / stussy@uchicago.edu
46. Lev Gorenstein / Globus - University of Chicago / lev@globus.org
47. Brad Spitzbart / Harvard University / bradley_spitzbart@g.harvard.edu
48. Suzanna Gardner / Purdue University - Rosen Center for Advanced Computing
49. Daniel Lucio / NC State University / dalucio@ncsu.edu
50. Katherine Holcomb / University of Virginia / kah3f@virginia.edu
51. Amir Karger / Harvard Medical School
52. Ed Hall / University of Virginia / edhall@virginia.edu
53. Carl Baribault / Tulane University / cbaribault@tulane.edu
54. Sean Bethune / Fred Hutch / rbethune@fredhutch.org

Max attendee Zoom count 65 - with 55 sign in

Notes from the call

Notes

Globus is a non-profit organization within U of Chicago, though they do some marketing to sustain their services.

Data processing pipelines needs (reliability, data access, compute across storage, data management best practices)

Globus flow example - from instrument to storage to compute to store elsewhere, share with others, do more processing, add metadata, publish

Demo of Windows VM connected to “microscope”, taking picture starts a flow to process and copy the data to a shared repository

Globus compute - adding computation to the data workflow

- Needs to rethink computing to make it easier to access
- Globus compute service can have different compute resources
- Register code with Globus compute service, then request computation through this service

- Multi-user endpoints interface the compute resources

Globus flows manage sequence of tasks, outsourcing the tasks to Globus services.

Flow lifecycle - define it in json format, deploy flow, set access policy, run (debug) and monitor, and run over and over

Flow examples are available and Globus is willing to help in writing the json flows.

Can trigger workflows automatically e.g. from instruments

Slide with details of the example “microscope” workflow steps

The service allows “human in the loop” for manual steps - important for CryoEM

Examples of deployments

- APS at Argonne, variation of this pipeline for many different steps
- Vescovi et al paper referenced
[https://www.cell.com/patterns/pdf/S2666-3899\(22\)00231-8.pdf](https://www.cell.com/patterns/pdf/S2666-3899(22)00231-8.pdf)
<https://doi.org/10.1016/j.patter.2022.100606>
- Public repo [accdc.alcf.anl.gov](https://github.com/accdc/accdc)

Freemium sustainability model

- Basic capabilities free, enhanced features subscription, ~260 subscribers

Questions

Where does the computation run and is it suitable for sensitive data?

- Globus compute service is not certified for sensitive data, but they’re working on it
-

Does Globus compute support standard Bioinformatics pipelines in Nextflow (WDL) format?

- No native support, but, more generic workflow managers like Nextflow should work.
-

Can you speak to Globus integration with public cloud services related to serverless functions, compute provisioning, and budget / quota management?

- Globus is “serverless” from user’s perspective, but with support from the background infrastructure (e.g. local Globus endpoints), compute is defined via configuration (e.g. provision compute on cloud based on the definition?, or perhaps access to local compute resource/cluster?)
-

Is Flows on the roadmap for inclusion in the high-assurance subscription?

- Yes, same thing with compute, on the roadmap.

Can you comment on best practice for an institution’s onboarding one or more entire research teams for institutional access to globus - must each team member apply for access separately?

- Self service within institution with subscription

What are new features planned?

- Compute - multi-user capabilities, admin console
- Flows - HA availability
- Compliance with (evolving) security standards

- Further connectors (cloud storages,...)

how complicated can the analysis part be involved?

- Can be very complex - run on cluster

Can I synchronize a globus group with an ACCESS project? So if I add an access id to the access project, the globus id corresponding to that access id can be added to the globus group? If that's automatable, I bet a lot of PIs would want it.

- Can programmatically using Globus API parse the group members and map their identities to those of Globus

What are the hurdles for implementing the Globus workflows? I can see that this would be a good approach for some of our University's core facilities, but don't see GFs used.

Can you comment on management of DOS denial of service attacks, if appropriate?

- Have some capabilities - more question for operations group. Haven't had DoS but did have a couple users that put extensive load by mistake