A case of GitHub for our software repo.

Jin Huang   Joe Osborn

BNL   ORNL

GitHub is used by many of the reputable collaborations

▸ **GitHub** , a subsidiary of Microsoft, largest source code host

▸ Who use GitHub as main software repo, many migrated from lab-hosting

  ◦ CMS: https://github.com/cms-sw

  ◦ ALICE: https://github.com/alisw

  ◦ DUNE: https://github.com/DUNE

  ◦ sPHENIX: https://github.com/sPHENIX-Collaboration

  ◦ STAR (recently migrated) https://github.com/star-bnl

  ◦ BELLE2: https://github.com/belle2

▸ Using GitLab server at host lab

  ◦ LHCb: https://gitlab.cern.ch/lhcb

  ◦ ATLAS: https://gitlab.cern.ch/atlas

The place for world-wide software ecosystem

- ▶ GitHub is widely used by software community, in academia and industry
- ▶ Foundational package forking – easier providing contribution via pull request
  - ◦ https://github.com/root-project
  - ◦ https://github.com/acts-project
  - ◦ https://github.com/AIDASoft
  - ◦ https://github.com/iris-hep
  - ◦ https://github.com/HSF
- ▶ Publish of our packages and recognized world-wide
- ▶ Many 3rd party site with built-in GitHub support:
  - ◦ GitHub authentication, minting DOI via Zenodo, Read-The-Doc, nbviewer
- ▶ Career building:
  - ◦ GitHub profile page can be a solid part of the resume

GitHub | Mapping collaborative software
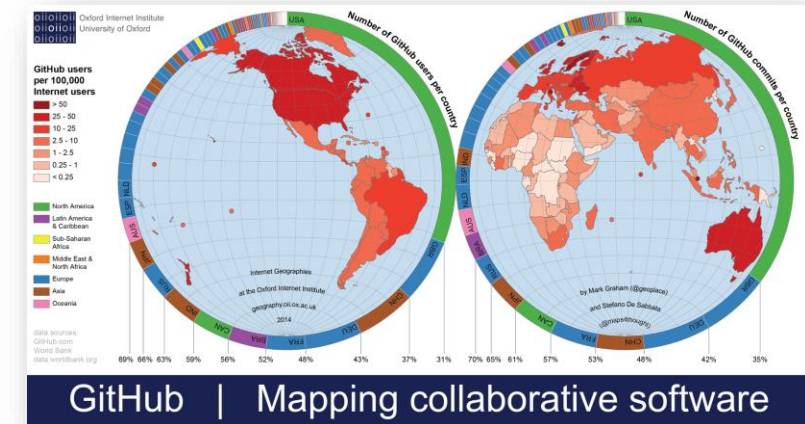
Continuous Integration and Continuous Deployment

▶ GitHub built-in CI/CD for computing intensive tasks is expensive [ref] (profit from cloud computing)

▶ But GitHub accumulated support from many CI/CD tools, including self-hosting Jenkins service provided by the host lab (topic of another meeting).   Example→

▶ Wide support allows for distributed CI jobs

  ○ e.g. GitHub-Actions for doc releases, Jenkins for PR check, HPC jobs for auto AI model build, FELIX test stand for FPGA code testing
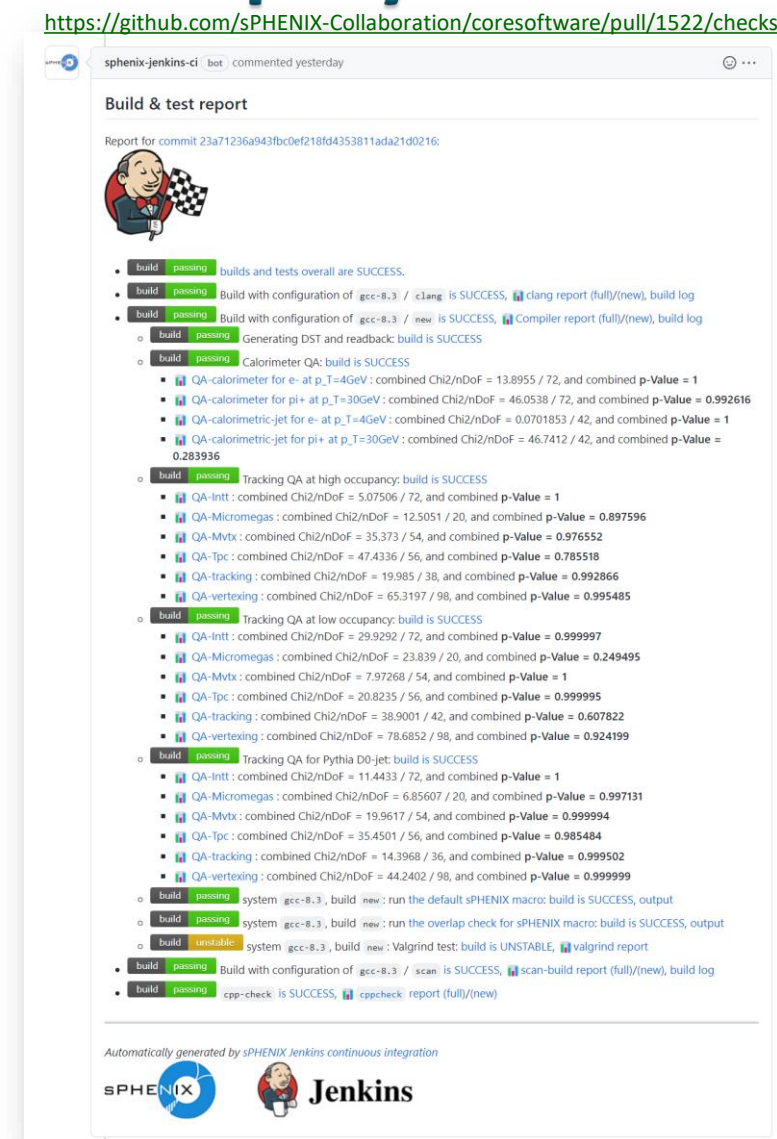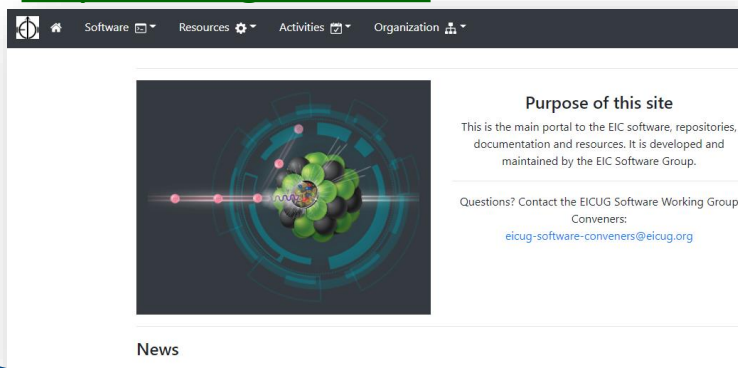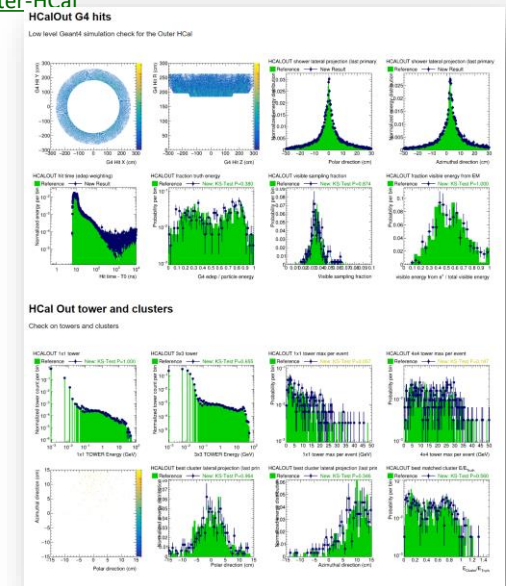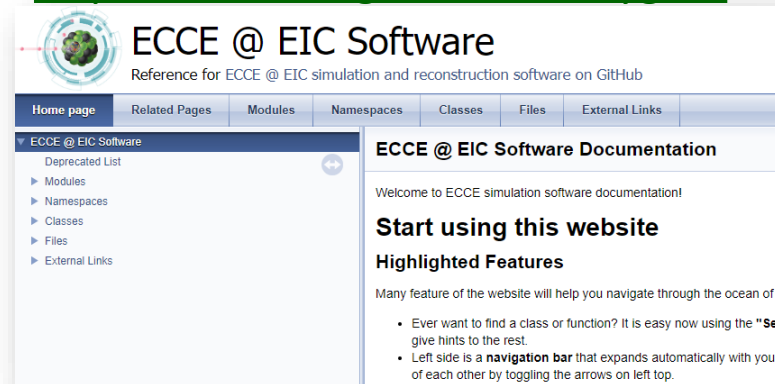
Documentation with code

▸ Search a piece of code in our organization or over the entire GitHub

▸ Built-in GitHub page (or external spt e.g. Readthedoc): https://eic.github.io/

▸ Code referencing hosting, e.g. https://ecce-eic.github.io/doxygen/

▸ QA history hosting,
  e.g. https://nbviewer.org/github/sPHENIX-Collaboration/QA-gallery/blob/jenkins-sPHENIX-test-calo-single-qa-3237-pi+_pT30_Sum8/QA-calorimeter.ipynb#Outer-HCal

https://eic.github.io/

https://ecce-eic.github.io/doxygen/

Free accounts/repos, come with size limitations

https://github.com/pricing#compare-features

https://docs.github.com/en/repositories/working-with-files/managing-large-files/about-git-large-file-storage

GitHub Free
The basics for organizations and developers

∞ Unlimited public/private repos

∞ Unlimited collaborators

✓ 2,000 Actions minutes/month

✓ 500MB of Packages storage

✓ Community support

About Git Large File Storage

Git LFS handles large files by storing references to the file in the repository, but not the actual file itself. To work around Git's architecture, Git LFS creates a pointer file which acts as a reference to the actual file (which is stored somewhere else). GitHub manages this pointer file in your repository. When you clone the repository down, GitHub uses the pointer file as a map to go and find the large file for you.

Using Git LFS, you can store files up to:

Product	Maximum file size
GitHub Free	2 GB
GitHub Pro	2 GB
GitHub Team	4 GB
GitHub Enterprise Cloud	5 GB

About billing for Git Large File Storage

Every account using Git Large File Storage receives 1 GB of free storage and 1 GB a month of free bandwidth. If the bandwidth and storage quotas are not enough, you can choose to purchase an additional quota for Git LFS. Unused bandwidth doesn't roll over month-to-month.

Brookhaven
National Laboratory

Comments (in our view)

- Keep our software open sourced
  - Before data taking, it is probably best practice to keep all scientific software open sourced
  - Build wider collaboration and build our credibility
- Large file storage
  - I would suggest not to use *git* for large file storage
  - EIC has solution for large files: CVMFS, S3, XROOTD
    - Successfully used by ATHENA and ECCE in past year
  - In operation:
    - Calibration database (will be addressed by another discussion)

Brookhaven
National Laboratory

Check list

- Cloud service accessible from anywhere in the world
  - Sure. GitHub's sell point
- Does not require a paid account for each user
  - Unlimited users, account recognized elsewhere via OAuth
- Top-level repositories can be configured with different access policies ranging from world-readable to private (with access only for select users)
  - Unlimited public/private repos [ref]
- Supports Continuous Integration (CI)
  - In operation for many collaborations
- Non-restrictive limits:
  - >=1000 repositories
    - Unlimited [ref]
  - >=1TB (w/ ability to increase as needed w/o significant additional cost)
  - >=10TB/mo (w/ ability to increase as needed w/o significant additional cost)
    - These two is tied to number of repos (100GB/repo) and number of files (~100MB/file). But… don't just use git for large file storage, we have better solution for that

Brookhaven
National Laboratory