Enabling Continuous Large Scale Software Engineering Experimentation in the Cloud

Jonathan Bell (NEU)
Christopher Timperley, Michael Hilton and Lauren Herckis (CMU)

More resources: https://classee.cloud
State of the practice for building and evaluating software tools

What if we tried a new mutator to improve fuzzing?

Novel research idea

Implement prototype

Perform manual effort to find dependencies needed to run evaluation

Manually run evaluation

Study results, form new hypotheses to improve prototype

Use Example from previous project which may or may not currently run

Publish approach and results in a research paper

After months, try to remember all manual steps to build artifact

Manual Large-Scale Experiment Execution
Automating Large-Scale Experiment Execution

A new workflow for building and evaluating software tools

What if we tried a new mutator to improve fuzzing?

Novel research idea

Use Starter Template for Fuzzing/Test Generation

Implement prototype

Push work-in-progress to GitHub

Cloud Evaluation Runs Automatically

Study results, form new hypotheses to improve prototype

Publish approach and results in a research paper

Publish artifact for reusable tool + evaluation

Use Starter Template for Fuzzing/Test Generation
Prototype application: Evaluating Fuzzing in CI

Fully automated performance tests of our fuzzer, in CI, using on-prem HPC cluster

- Our ICSE ’22 “CONFETTI” project needs 100 x 4-core VMs for 24 hrs each for evaluation
- Now: fully automated in CI, 20x4-core VM evaluation runs for 10 mins each commit, full evaluation manually triggered
- Greatly simplifies testing and makes pull-requests far easier

Sample graph generated by the CI workflow - automatically compares triggering commit to baseline

We used this workflow to find and fix a significant performance bug in the upstream tool
Our Research & Infrastructure Agenda

Infrastructure to write evaluations
- Workflow templates and reusable actions
- Community Research
- Documentation and resources

Infrastructure to run evaluations
- CLASSEE.cloud launches; GitHub Actions configures and manages
- Evaluation reports
- Dashboard with plugins
- Shared Utility Infrastructure

Evaluation scripts utilize
- Hosted CLASSEE.cloud Service
- Compute Service
- CI Runner
- CLASSEE.cloud launches; GitHub Actions configures and manages

Virtual & In-Person Training

Community Research
Open Research Questions

• What are the best practices for running and monitoring large evaluation campaigns in CI?

• How to target specific projects/evaluation targets, while still providing general resources?

• How to design flexible interfaces for integrating tools and datasets?

• Where to store and how to manage artifacts and dependencies?

• How to encourage adoption of best practices for evaluation design?

• …What else haven't we thought of?