



WRENCH

Workflow Management System Simulation Workbench

- *Accurate, scalable, and reproducible simulations*

Motivation

Scientific Workflows are key to advances in science and engineering

Their executions are complex:

Workflow structures are large and can be configured in various ways

Workflow Management Systems (WMS) are **large multi-component software systems** that employ ranges of decision making algorithms

Workflow execution platforms are **heterogeneous** and diverse

We need a strong “**experimental science**” approach to study these complex systems in a view to optimizing workflow executions

Yet real-world experiments are inherently limited

- Time- and resource-intensive

- Limited to existing configurations

- Require full-fledged implementations

- ...

Objectives

Realize a **workflow execution simulation methodology** that has high simulation accuracy, low execution time, and low memory footprint

This framework is to be used:

By **workflow users** to study workflow executions

By **WMS developers** to inform system and algorithm design decisions

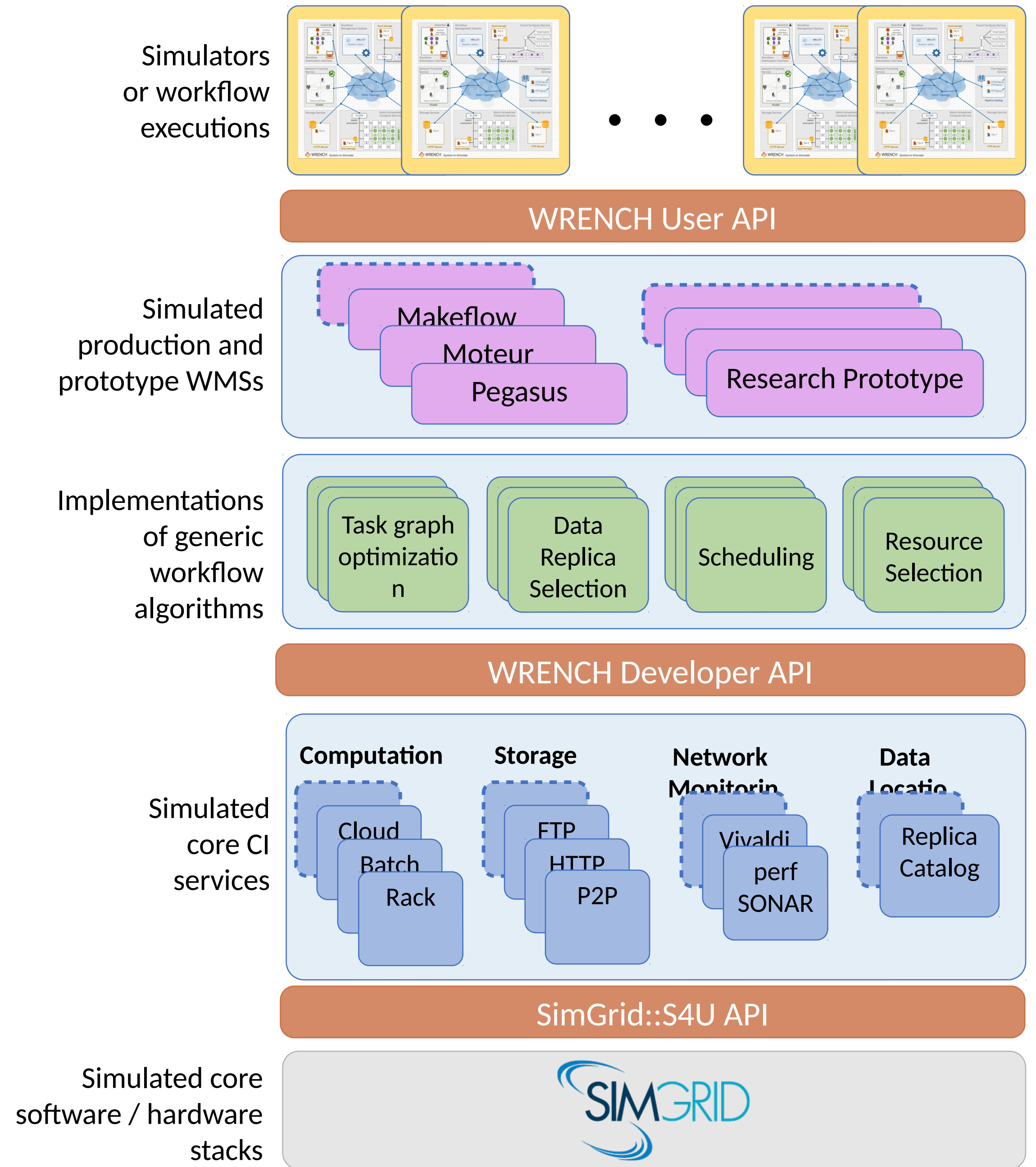
By **educators** to teach distributed computing in the context of workflows

What is WRENCH?

WRENCH enables novel avenues for scientific workflow use, **research**, **development**, and **education** in the context of large-scale scientific computations and data analyses

WRENCH is an **open-source library** for developing simulators

WRENCH exposes several high-level simulation abstractions to provide high-level **building blocks** for developing custom simulators



Why SimGrid?



<http://simgrid.gforge.inria.fr>

SimGrid is a **research project**

Development of **simulation models** of hardware/software stacks

Models are **accurate** (validated/invalidated) and **scalable** (low computational complexity, low memory footprint)

SimGrid is **open source usable software**

Provides different APIs for a range of simulation needs, e.g.:

S4U: General simulation of Concurrent Sequential Processes

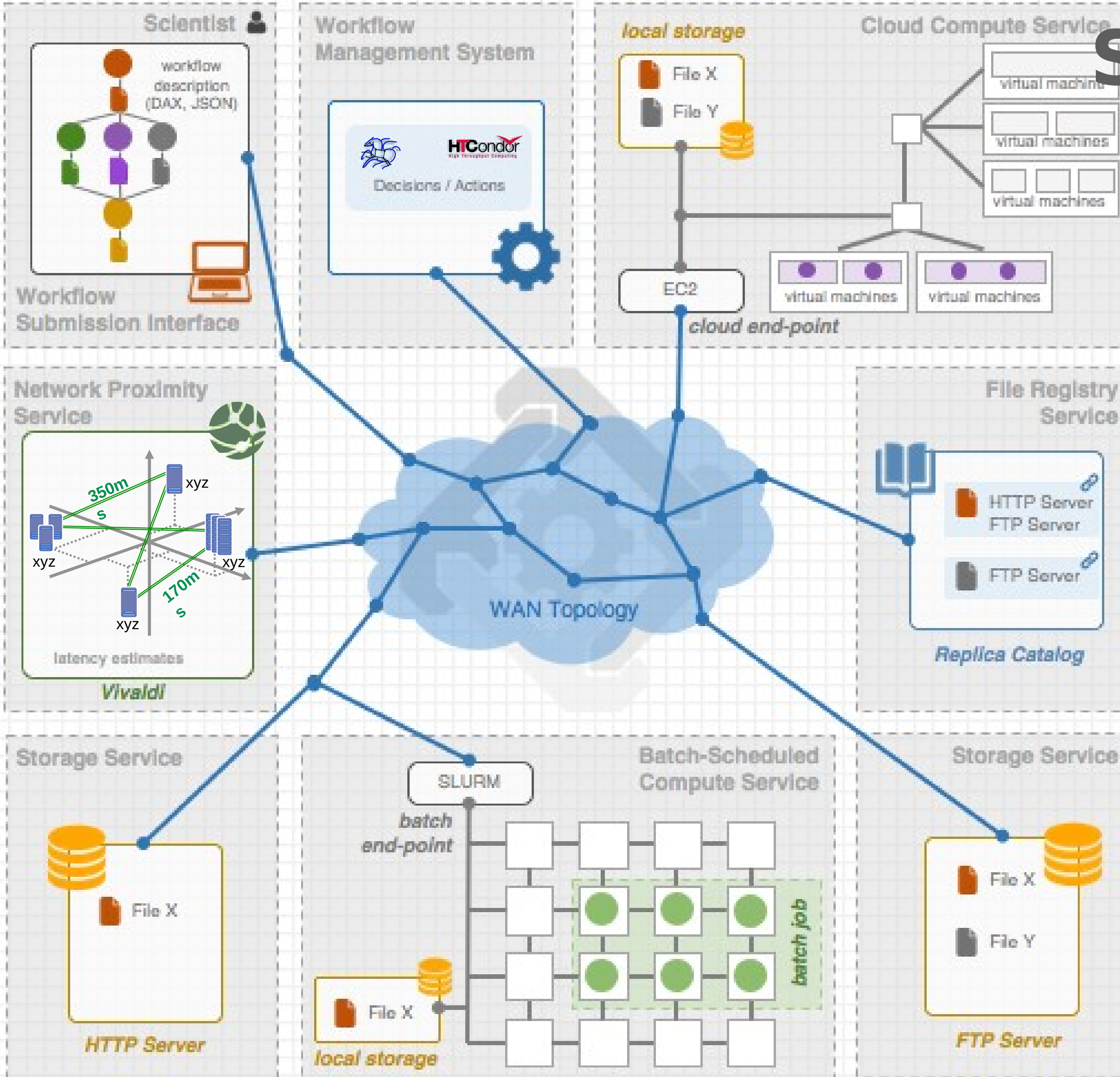
SMPI: Fine-grained simulation of MPI applications

SimGrid is **versatile scientific instrument**

Used for (combinations of) **Grid, HPC, Peer-to-Peer, Cloud, Fog** simulation projects

First developed in 2000, latest release: v3.21 (September 2018)

System to Simulate



Compute Services

- Bare-metal servers
- Cloud platforms
- Virtualized Cluster platforms
- Batch-scheduled clusters

Storage Services

- Including scratch spaces

File Registry Services

- Replica catalog (key-value pairs)

Network Proximity Services

- Database of host-to-host network distances (Vivaldi)

Workflow Management Systems

- Decision-making for optimizing various objectives (static and dynamic)
- Pilot Jobs

Building a Simulator

Blueprint for a WRENCH-based simulator

Create and initialize a simulation

Instantiate a simulated platform

Instantiate services on the platform

Create at least one workflow

Instantiate at least one WMS per workflow

Launch the simulation

Process simulation output



WRENCH + SimGrid internals

Agent: some code, some private data, running on a given host

Task: amount of work to do and of data to exchange

Host: location on which agents execute

Mailbox: Rendez-vous points between agents

You can send 'data' to a mailbox; you receive 'data' from a mailbox

Communication time between sender/receiver is accounted (**payload**) and depends on the network traffic

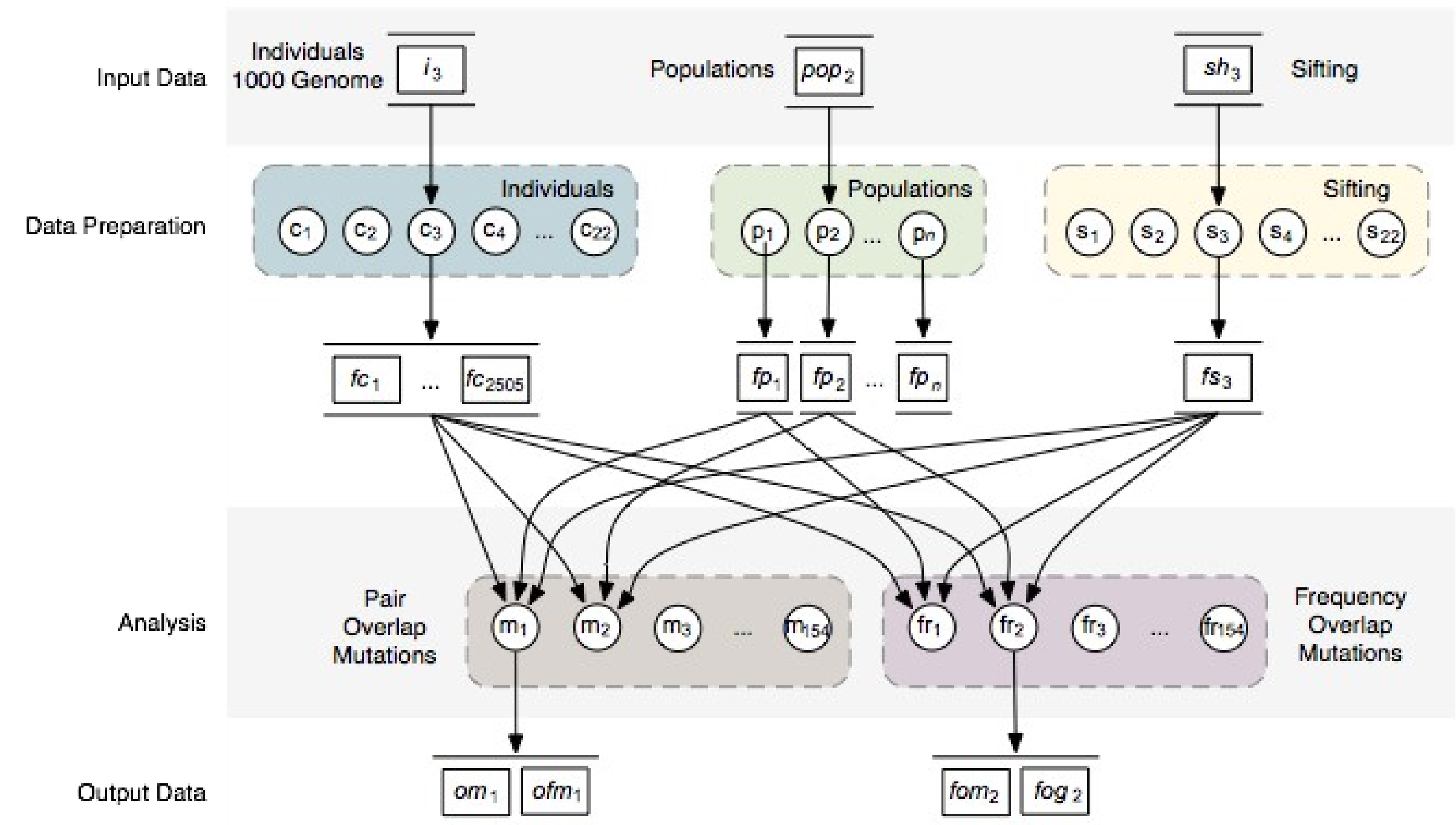
Preliminary Results

Scientific Workflow Application

1000 Genome Sequencing Analysis Workflow

Identifies mutational overlaps using data from the 1000 genomes project

22 Individual tasks, 7 Population tasks, 22 Sifting tasks, 154 Pair Overlap Mutations tasks, and 154 Frequency Overlap Mutations tasks (**Total 359 tasks**)



Preliminary Results



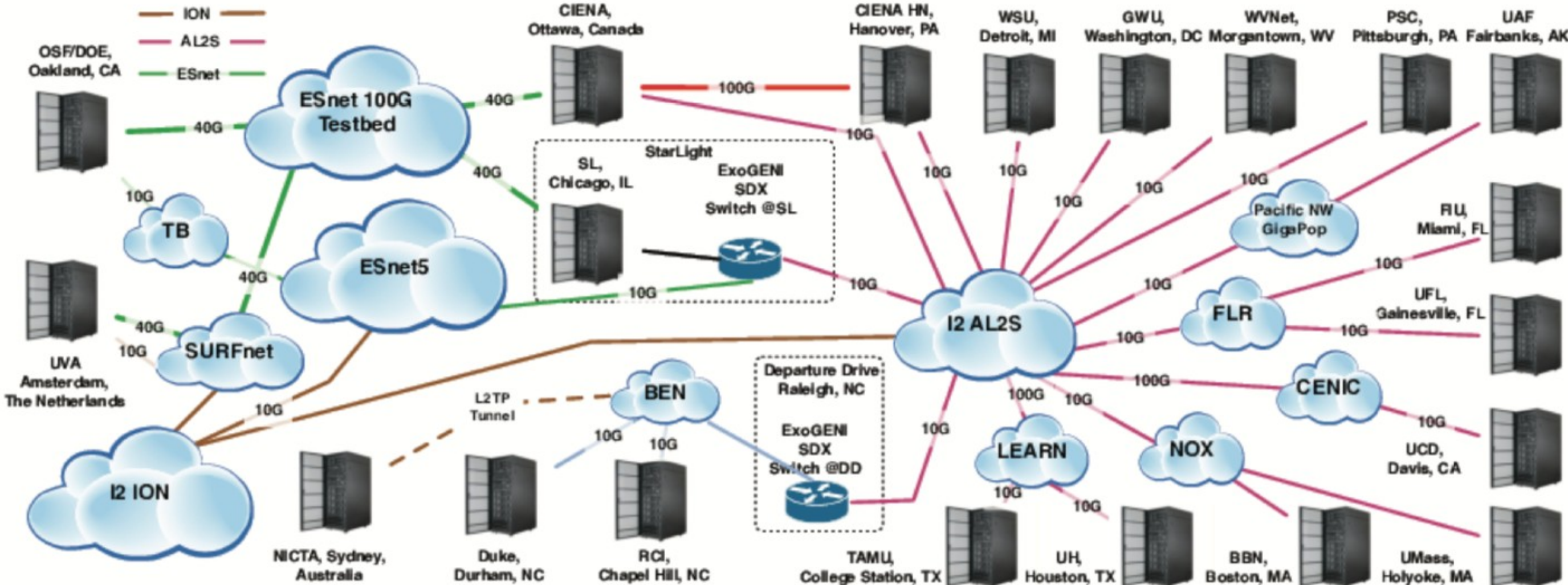
Experiment Configuration

Simulation based on a real Computing Infrastructure

ExoGENI testbed

Network IaaS national testbed powered by the ORCA (Open Resource Control Architecture) control software used for GENI

ORCA allows users to create mutually **isolated slices** of interconnected infrastructure from multiple independent providers (**compute, network, and storage**) and commodity infrastructure



Preliminary Results

Simulated Platform

Data node



Master server
(submit host)



Worker nodes
(4 cores each)

Modeled as a bare metal system

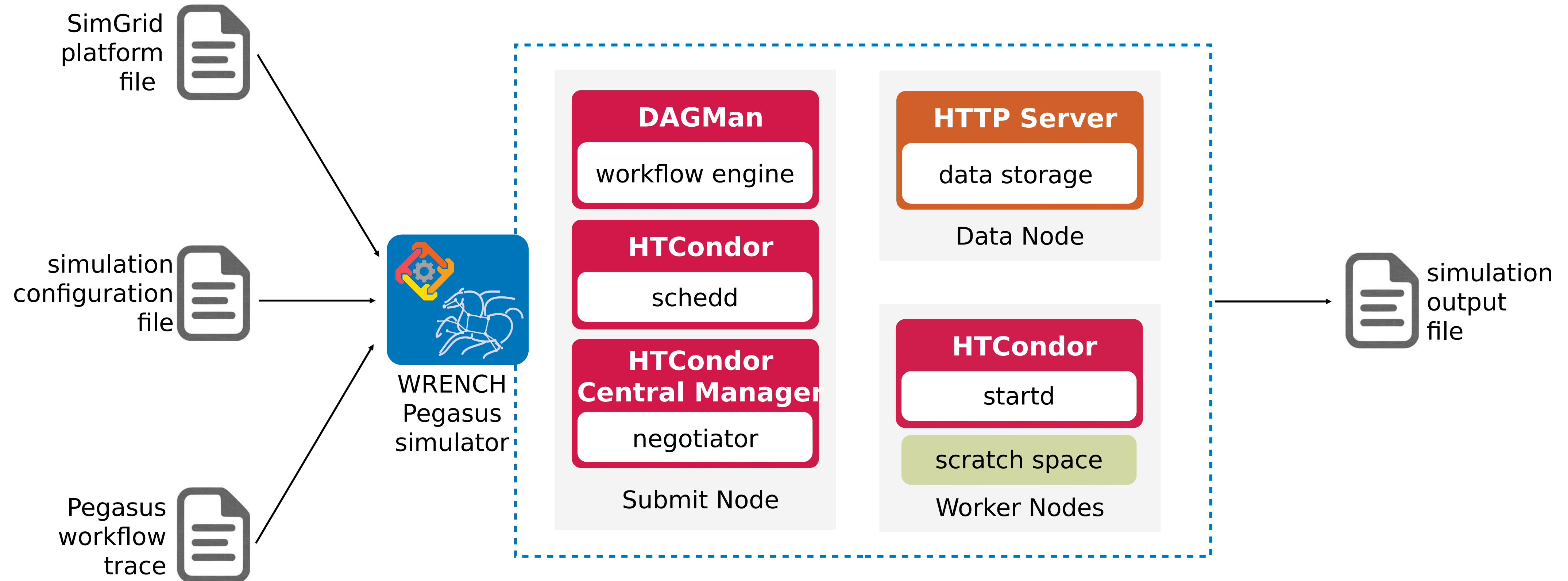
```
<?xml version='1.0'?>
<!DOCTYPE platform SYSTEM "http://simgrid.gforge.inria.fr/simgrid/simgrid.dtd">
<platform version="4.1">
  <zone id="AS0" routing="Full">
    <host id="master" speed="1f" core="4"/>
    <host id="data" speed="1f" core="1"/>
    <host id="workers1-2" speed="1f" core="4"/>
    <host id="workers1-0" speed="1f" core="4"/>
    <host id="workers1-3" speed="1f" core="4"/>
    <host id="workers1-1" speed="1f" core="4"/>
    <host id="workers1-4" speed="1f" core="4"/>
    <link id="1" bandwidth="125MBps" latency="100us"/>
    <link id="2" bandwidth="55MBps" latency="100us"/>
    <route src="master" dst="workers1-2">
      <link_ctn id="1"/>
    </route>
    <route src="master" dst="workers1-0">
      <link_ctn id="1"/>
    </route>
    <route src="master" dst="workers1-3">
      <link_ctn id="1"/>
    </route>
    <route src="master" dst="workers1-1">
      <link_ctn id="1"/>
    </route>
    <route src="master" dst="workers1-4">
      <link_ctn id="1"/>
    </route>
    <route src="data" dst="master">
      <link_ctn id="2"/>
    </route>
  </zone>
</platform>
```

SimGrid Platform description file

Preliminary Results

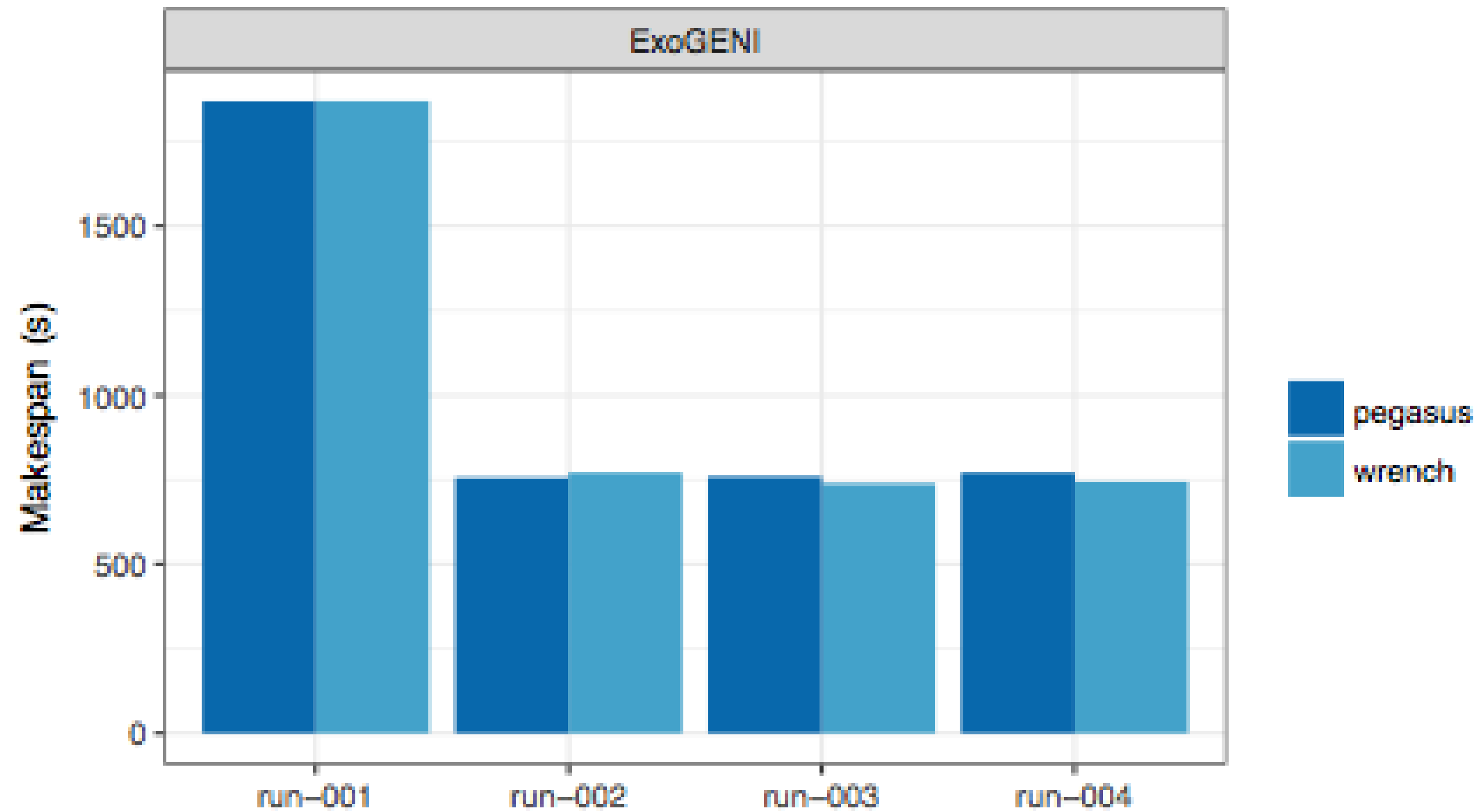
<https://github.com/wrench-project/pegasus>

Simulated Workflow Management System



Preliminary Results

Simulation Results and Accuracy



>98% workflow makespan accuracy

Simulated **compute** and **data transfer** tasks includes simulation of auxiliary tasks (e.g., create_dir, cleanup, and registration), and PRE and POST script jobs

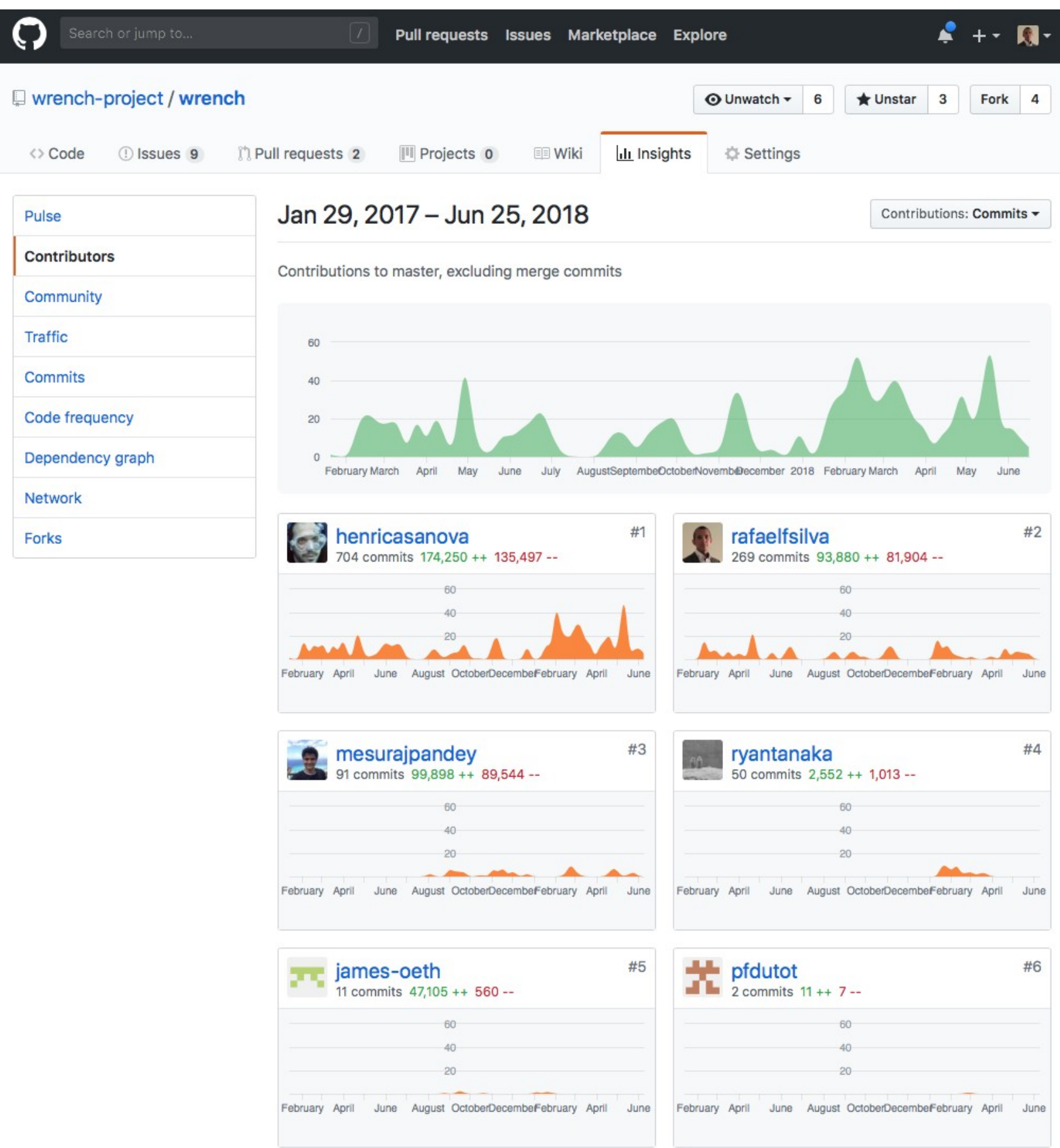
Simulates delays on both DAGMan and HTCondor daemons

Software Availability

Code Repository, Releases, Software Engineering Process

Open-source repository

<https://github.com/wrench-project/wrench>



Releases

1.0 (June 16, 2018)

1.0-beta (April 15, 2018)

1.0-alpha (December 1, 2017)

Upcoming releases (estimated)

1.0.1 (August 2018)

1.1 (September 2018)

Continuous Integration

<https://travis-ci.org/wrench-project/wrench>



Travis CI

build passing

Tests Coverage

<https://coveralls.io/github/wrench-project/wrench>

COVERALLS

coverage 89%

Code Review

<https://sonarcloud.io/dashboard?id=wrench>

sonarcloud

lines of code 15k

code quality A

Education

WRENCH Stand-alone Pedagogical Module

It is crucial to teach undergraduate students parallel and distributed computing

But it is not easy

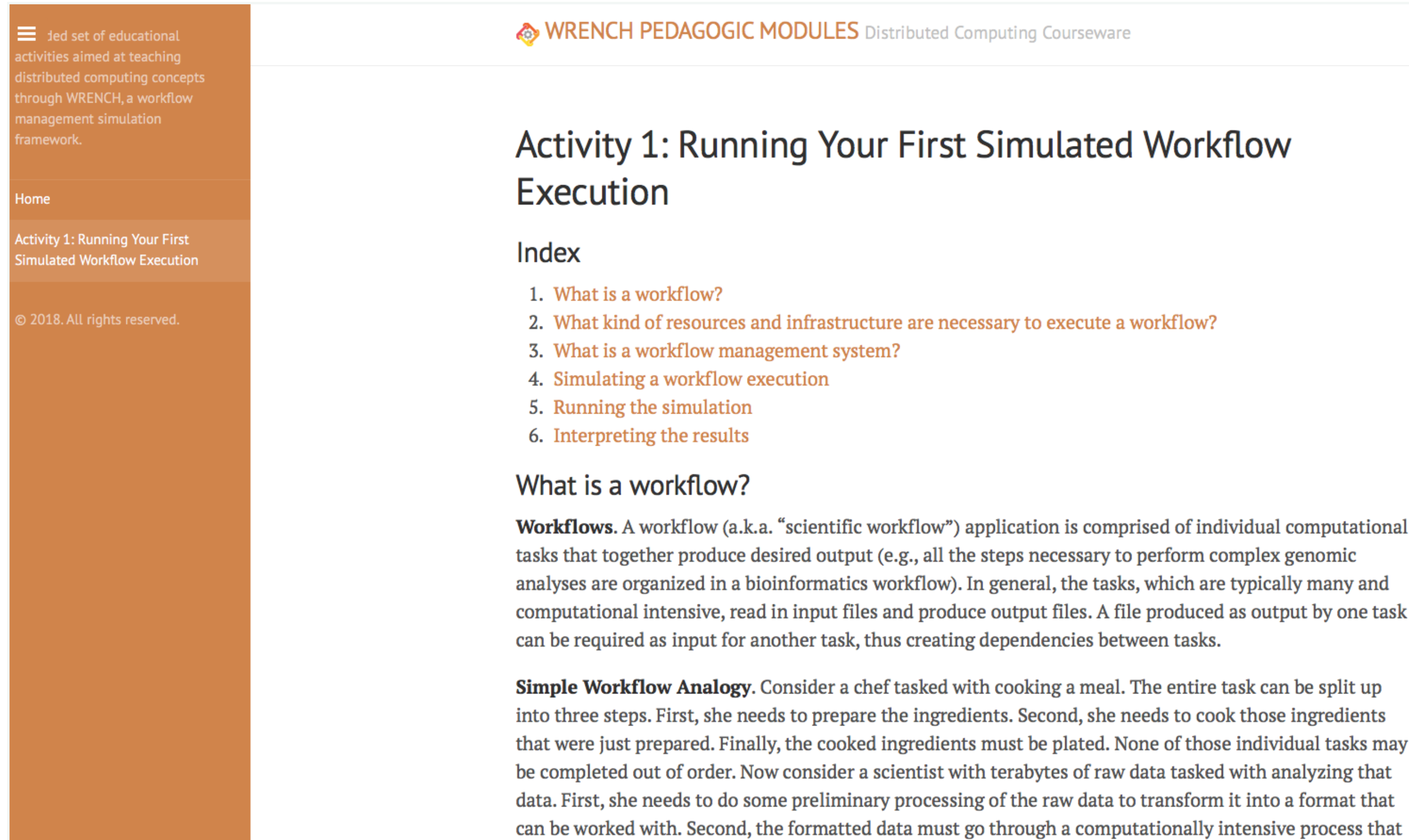
giving students access to sufficiently diverse and **realistic** software/hardware platforms

dealing with platform **down-times** and **instabilities**

dealing with time-consuming and possible **costly executions**

Simulation resolves these difficulties and WRENCH provides the foundation for **pedagogic modules on parallel and distributed computing** that use workflows as a motivating context

<http://wrench-project.org/wrench-pedagogic-modules>



The screenshot shows the WRENCH Pedagogic Modules website. The left sidebar is orange and contains a menu with the following items: a hamburger icon followed by the text 'A curated set of educational activities aimed at teaching distributed computing concepts through WRENCH, a workflow management simulation framework.', 'Home', 'Activity 1: Running Your First Simulated Workflow Execution', and '© 2018. All rights reserved.'. The main content area is white and features the WRENCH logo and the text 'WRENCH PEDAGOGIC MODULES Distributed Computing Courseware'. Below this is the title 'Activity 1: Running Your First Simulated Workflow Execution' and an 'Index' section with a numbered list of six items: 1. What is a workflow?, 2. What kind of resources and infrastructure are necessary to execute a workflow?, 3. What is a workflow management system?, 4. Simulating a workflow execution, 5. Running the simulation, and 6. Interpreting the results. Below the index is the section 'What is a workflow?' which includes a definition of workflows and a 'Simple Workflow Analogy' comparing a chef's tasks to a scientist's data processing tasks.

WRENCH PEDAGOGIC MODULES Distributed Computing Courseware

Activity 1: Running Your First Simulated Workflow Execution

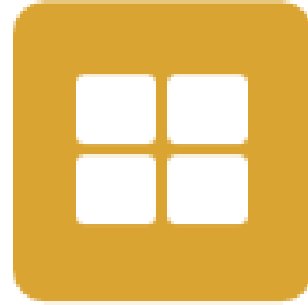
Index

1. What is a workflow?
2. What kind of resources and infrastructure are necessary to execute a workflow?
3. What is a workflow management system?
4. Simulating a workflow execution
5. Running the simulation
6. Interpreting the results

What is a workflow?

Workflows. A workflow (a.k.a. “scientific workflow”) application is comprised of individual computational tasks that together produce desired output (e.g., all the steps necessary to perform complex genomic analyses are organized in a bioinformatics workflow). In general, the tasks, which are typically many and computational intensive, read in input files and produce output files. A file produced as output by one task can be required as input for another task, thus creating dependencies between tasks.

Simple Workflow Analogy. Consider a chef tasked with cooking a meal. The entire task can be split up into three steps. First, she needs to prepare the ingredients. Second, she needs to cook those ingredients that were just prepared. Finally, the cooked ingredients must be plated. None of those individual tasks may be completed out of order. Now consider a scientist with terabytes of raw data tasked with analyzing that data. First, she needs to do some preliminary processing of the raw data to transform it into a format that can be worked with. Second, the formatted data must go through a computationally intensive process that



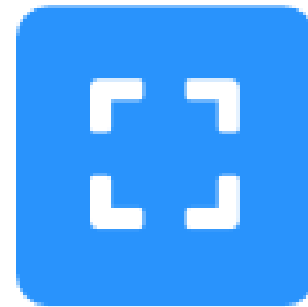
Simulation Building Blocks

Prototype implementations of Workflow Management System (WMS) components and underlying algorithms



Simulation Accuracy

Captures the behavior of a real-world system with as little bias as possible via validated simulation models



Scalability

Low ratio of simulation time to simulated time, ability to run large simulations on a single computer with low compute, memory, and energy footprints



Reproducible Results

Enable the reproduction or repetition of published results by a party working independently using the same or different simulation models



Our Team



UNIVERSITY
of HAWAII®
MĀNOA

USC Viterbi

School of Engineering
Information Sciences Institute



WRENCH is funded by the National Science Foundation (NSF) under grants number 1642369 and 1642335, and the National Center for Scientific Research (CNRS) under grant number PICS07239.



Workflow Management System Simulation Workbench

Accurate, scalable, and reproducible simulations

Download wrench-1.0

Get Started!

Documentation

WRENCH 101

GitHub Repository

Docker Containers



Thank You



Simulation Building Blocks

Prototype implementations of Workflow Management System (WMS) components and underlying algorithms



Simulation Accuracy

Captures the behavior of a real-world system with as little bias as possible via validated simulation models



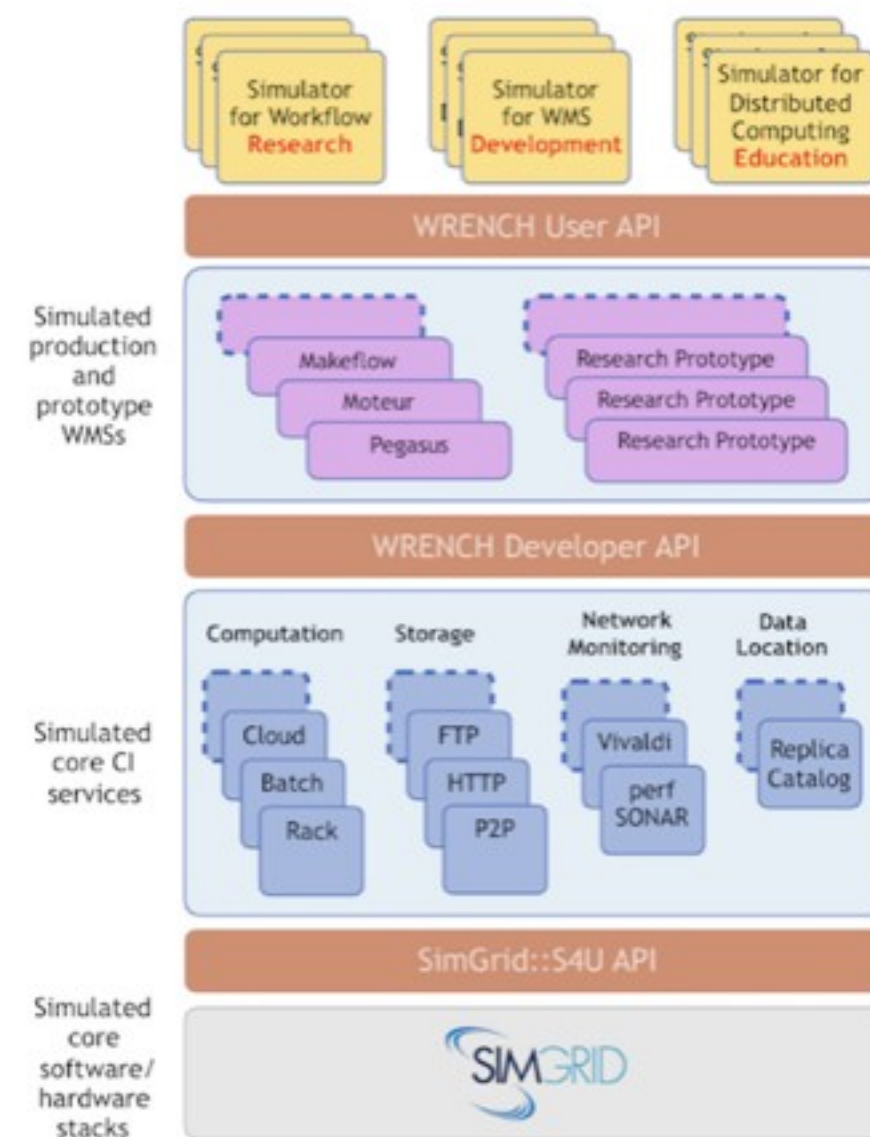
Scalability

Low ratio of simulation time to simulated time, ability to run large simulations on a single computer with low compute, memory, and energy footprints



Reproducible Results

Enable the reproduction or repetition of published results by a party working independently using the same or different simulation models.



Get Started:

<http://wrench-project.org>