# Toil

## Robust Pipeline Architecture for Genomic Workflows

John Vivian, CJ Ketchum, Hannes Schmidt, David Haussler, Benedict Paten

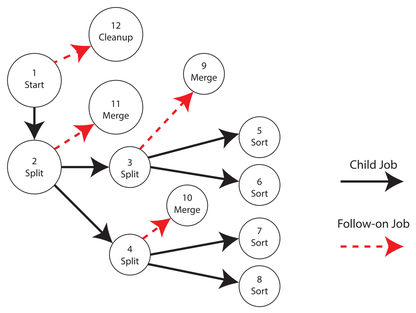
UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

TOIL
Massively Parallel Workflows
(and fire-breathing dragon slugs)

### Why Toil?

Toil is a massively scalable pipeline management system that is fault-tolerant, portable, simple, and open-source. Workflows can be easily developed on a laptop, then deployed to both cloud environments and standard clusters.

#### Dynamic DAG Generation




The dependency graph for the workflow can be described statically, dynamically, or using a combination of methods. Static descriptions follow a functional paradigm and can pass information between jobs using the concept of promises. Entire subgraphs can be encaspulated, greatly simplifying the "wiring" that needs to be declared. Toil also has early support for Common Workflow Language (CWL), an open-source project dedicated to workflow standards.

### Commercial Cloud Support

amazon web services™

Microsoft Azure

Google Cloud Platform

Also includes support for several high-performance batch systems:

- Grid Engine
- Mesos
- Parasol

### Simple

Workflows are written in Python, providing the power and convenience of a popular Turing-complete language without needing to learn another Domain Specific Langauge (DSL).

```
from toil.job import Job

def HelloWorld(message, cores=2, memory="2G", disk="3G")
      return "Hello world!, here's your message: %s" % message

j = Job.wrapFn(HelloWorld, "woot")

if __name__ == "__main__":
      options = Job.Runner.getDefaultOptions("./toilWorkFlow")
      print Job.Runner.startToil(j, options)
```

docker

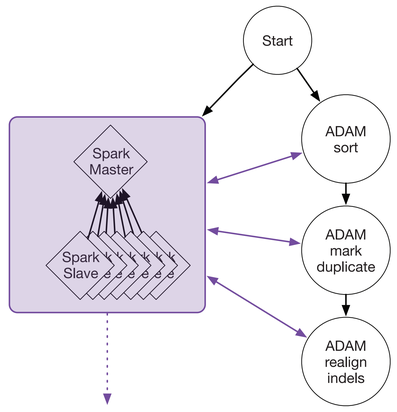
CoreOS

### Docker Integration

Standardized Docker containers are used in our workflows to maximize portability. These containers remove a significant number of dependencies, only requiring Python, Toil, and Docker be installed on the system. UCSC has joined the Global Alliance for Genomics and Health (GA4GH) in designing Docker standards for genomic tools.

### Spark & ADAM Integration

Spark is a computing platform that allows for primitive operations in memory (like map/reduce) that allow for drastic increases in speed by avoiding I/O bottleneck. ADAM is a genomics engine built on-top of Spark, aimed at reducing the compute time needed to run common genomic algorithms (e.g. BAM preprocessing).




Dockerized Spark containers allow spark clusters to be provisioned in seconds instead of minutes. In Toil, a special **service job** can be launched that provisions a long-running spark cluster that other jobs interact with.

### Who uses Toil?

Treehouse Project

STAND UP TO CANCER™

amplab

I-SPY 2 TRIAL

OREGON HEALTH & SCIENCE UNIVERSITY

The Children's Hospital *of* Philadelphia®