CONTAINERIZING CONDA

Experiences Building 2000 Applications As Portable Docker And Singularity Images

Rion Dooley, John Fonner

@agaveapi

dooley@tacc.utexas.edu, jfonner@tacc.utexas.edu

HELLO, MY NAME IS…

- Rion Dooley
- Manager, Web and Cloud Services Group,
- Texas Advanced Computing Center (TACC)
- PI of the Agave project (https://agaveapi.co)

I LIKE LONG WALKS ON THE BEACH AND…

- That's taken me up and down the stack.
- Desktop -> Gateway -> Middleware -> SaaS -> PaaS
- TeraGrid File Manager -> GridChem/XUP -> XSEDE IIS/MPG -> GatewayDNA -> Agave

Helping people work smarter, collaborate better, and dream bigger


Agave
Platform

AGAVE IS A MULTI-TENANT PAAS

DELIVERING SCIENCE-AS-A-SERVICE SOLUTIONS

IN *HYBRID* CLOUD ENVIRONMENTS

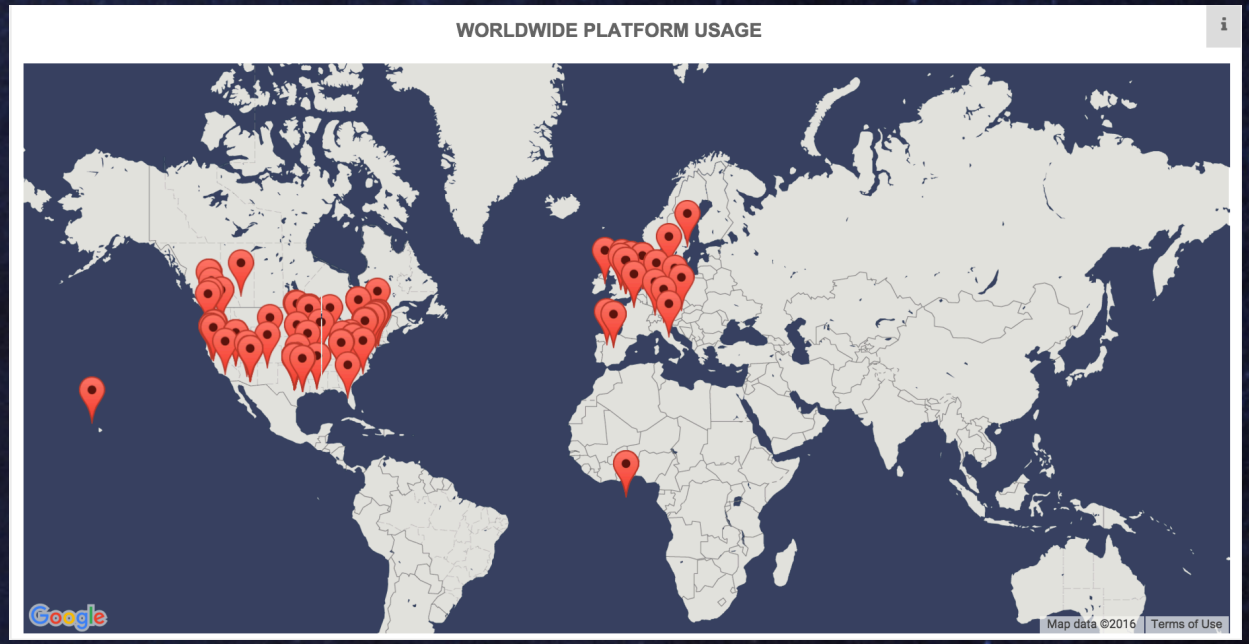
Think of it like

for Science

WHAT IS AGAVE?

- Manage it like a cloud platform
- Scale it like cloud infrastructure
- Use it like a cloud service
- Works with your new and legacy infrastructure
- FOSS - On premise, hybrid, or hosted deployment
- Multi-tenant
- Secure by default

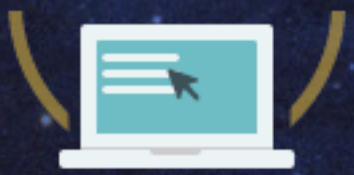
WORLDWIDE USAGE



WHAT DOES IT DO?

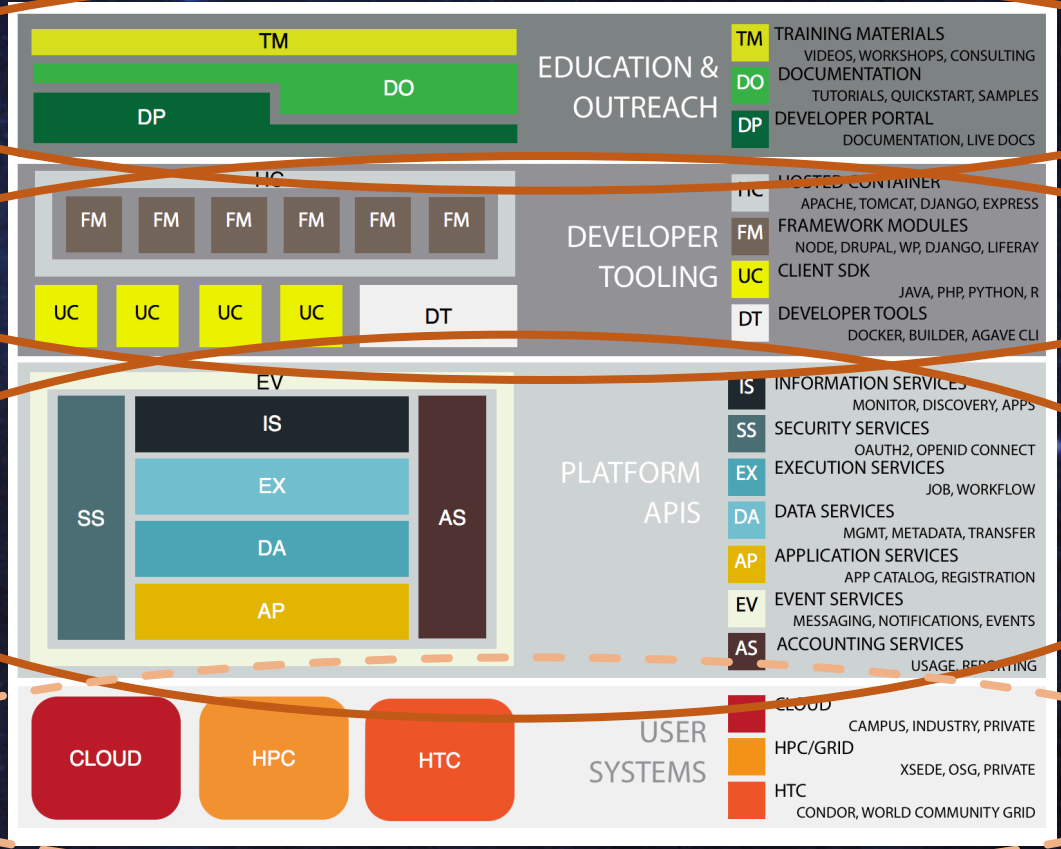
MANAGE DATA

RUN CODE



COLLABORATE ANYWHERE

CONNECT ANYTHING


TM
DO
DP
EDUCATION & OUTREACH
TM TRAINING MATERIALS
VIDEOS, WORKSHOPS, CONSULTING
DO DOCUMENTATION
TUTORIALS, QUICKSTART, SAMPLES
DP DEVELOPER PORTAL
DOCUMENTATION, LIVE DOCS
HC
FM FM FM FM FM FM
UC UC UC UC
DT
DEVELOPER TOOLING
HC HOSTED CONTAINER
APACHE, TOMCAT, DJANGO, EXPRESS
FM FRAMEWORK MODULES
NODE, DRUPAL, WP, DJANGO, LIFERAY
UC CLIENT SDK
JAVA, PHP, PYTHON, R
DT DEVELOPER TOOLS
DOCKER, BUILDER, AGAVE CLI
EV
SS
IS
EX
DA
AP
AS
PLATFORM APIS
IS INFORMATION SERVICES
MONITOR, DISCOVERY, APPS
SS SECURITY SERVICES
OAUTH2, OPENID CONNECT
EX EXECUTION SERVICES
JOB, WORKFLOW
DA DATA SERVICES
MGMT, METADATA, TRANSFER
AP APPLICATION SERVICES
APP CATALOG, REGISTRATION
EV EVENT SERVICES
MESSAGING, NOTIFICATIONS, EVENTS
AS ACCOUNTING SERVICES
USAGE, REPORTING
CLOUD
HPC
HTC
USER SYSTEMS
CLOUD
CAMPUS, INDUSTRY, PRIVATE
HPC/GRID
XSEDE, OSG, PRIVATE
HTC
CONDOR, WORLD COMMUNITY GRID

WHAT ABOUT CONTAINERS AND CONDA?

- About 4 years ago, we were in ops hell
- Build an test, deployments, scaling, configuration management, polyglot services, networking, multi-cloud, stateful everything, app installs, versioning, dependency management
- Enter a tech talk about Docker…

WE THOUGHT CONTAINERS WERE COOL

- Docker was still at 0.4, but we were hooked
- Encapsulation and the *ecosystem* solved so many of the problems we had
- But we weren't just looking at our hosting, we were looking at a bigger picture.

The juice was worth the squeeze!

DOCKER FOR COMPUTATION

Compute containers are the Magic 8 Ball of science...

- Compartmentalize code
- Eliminate build and run complexities
- Introduce portability, reuse, & versioning
- Widgetize the creation of a scientific pipeline

...but better because results are reproducible.

Compute containers enable reproducible science by composition.

DOCKER FOR DATA

Data containers serve as universal adapters between compute containers

- Transform data
- Bridge file systems
- Enable distributed data access
- Virtualize interfaces

Data containers enable clean integration between containers and ***standardize*** how we interact with ***distributed data***.

WHAT DO THESE HAVE IN COMMON?

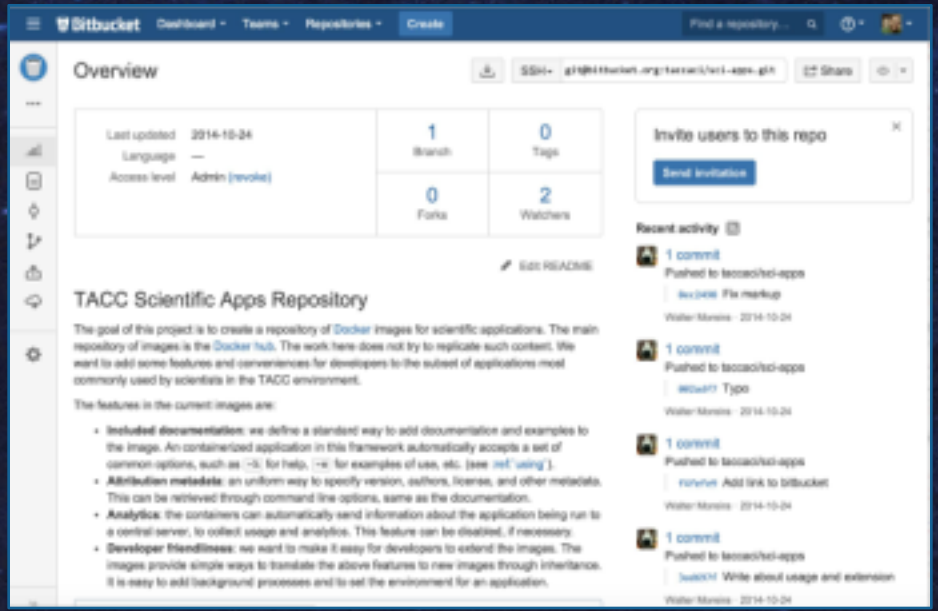
KONAMI CODE

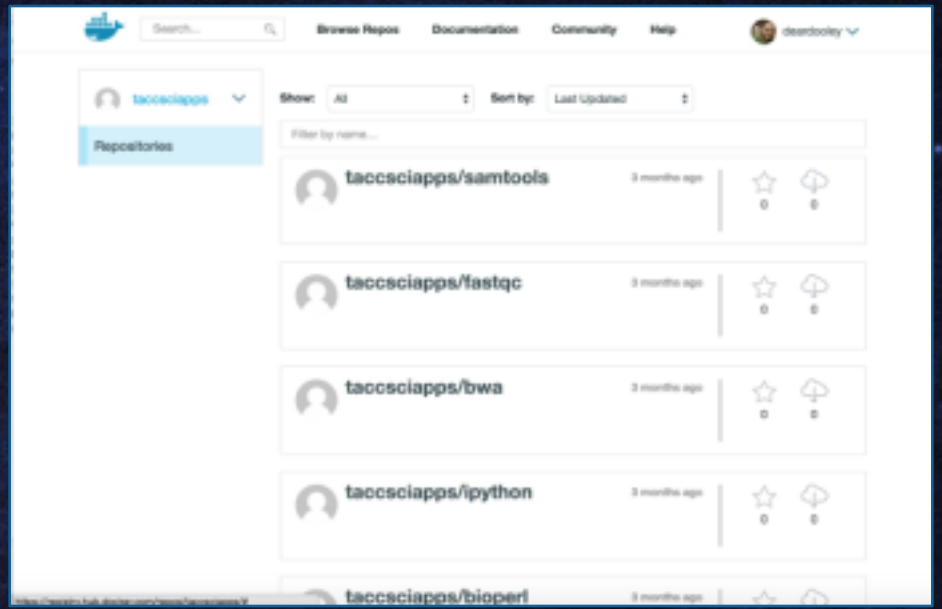
DOCKER FOR RESEARCH

- Compute and data containers *are* cheat codes when it comes to science
- Building a library of "Science Apps" in the Docker Hub
- Built from a few value-added trusted base images
  - Attribution
  - Extended metadata
  - Documentation
  - Extension points
  - Common CLI

Science apps are ***cheat codes*** when it comes to discovery.

DOCKER FOR RESEARCH





https://bitbucket.org/taccaci/sci-apps/overview

AND THEN NOTHING HAPPENED

- Well, we got tired
- And frustrated
- And lost a lot of sleep
- And wrote a bunch of code that will never see the light of day…except the previous slide.

IT'S THE ECOSYSTEM STUPID

- We forgot the very reason that made us love containers in the first place, the *ECOSYSTEM!*
- We returned to our roots and started looking for places we could fill in the gap between IT and research
- Better base images and Dockerfiles
- Integration with autobuilds
- Private repos as part of your personal storage
- Transparent build and run on demand

BUILDING TOWARDS SOMETHING BIGGER

- That went pretty well. We added several hundred docker images to the Cyverse infrastructure.
- Ran 100k or containers
- Taught users how to contribute their own images
- Then we hit the Peter Covney problem
- Just not enough cycles and/or hours in the day for the demand

THEN THE WHEELS CAME OFF

- Containers are portable. TACC has lots of cores. Problem solved
- Except Docker won't or can't run on many HPC systems
- Security issues
- Scheduler issues
- Data issues
- Repository issues
- Networking issues
- Accounting issues

ENTER SINGULARITY

- Singularity is "Containers for HPC"
- Singularity is NOT Docker
- Technology aside, there is no ecosystem, which made onboarding users nearly impossible
- But the potential is significantly larger to serve the research community since it's a portable, secure runtime that can swap out 1 for 1 with existing system calls.

CONDA GETTING THE HANG OF IT

- Built our own automated build farm to convert Docker image to Sinularity images
- 1 for 1 on published images
- Published the entire BioConda repository as Singularity images
- Added transparent support for one click execution of any image.
- Published images as publicly available downloads

STRENGTH IN NUMBERS

- We're not alone…not even close
- Currently working with Vanessa Sochat from the Singularity team to make images available on the new singularity-hub.org and public registry.
- Working with Björn Grüning on the BioContainers side to upstream generating the Singularity images, something they already want to do.
- Working with Common Workflow Language team to ensure we have a standard for invoking containers, regardless of the runtime.

DOING OUR PART

- Added first-class support for running Singularity containers in the Agave Platform.
- Working on REPL support and additions to the Agave CLI tools to help people go from daydream to discovery faster than ever.
- Incorporating the container ecosystem into all TACC resources.
- Clearing a smooth migration path for computation between local, cloud, htc, and HPC resources.
- Creating additional content to raise awareness, train, an support users as they begin looking at these technologies.

THANK YOU!

FOLLOW US

@agaveapi

agaveapi.co