Polyglot, Event Driven Computational Science Using the Actor Model

Joe Stubbs
Texas Advanced Computing Center
University of Texas, Austin
What is TACC?

**Mission:** To enable discoveries that advance science and society through the application of advanced computing technologies.

- High performance computing (HPC)
- Cloud & high throughput computing
- Data intensive computing
- Visualization
- Scientific software development & optimization
- **APIs and tools - Agave Platform**
- Web and mobile applications
- Life sciences
- Training & outreach
What is Agave?

- Register storage and compute systems
- Ingest, move and transform data files and folders
- Register applications (binaries) on large systems
- Launch jobs to invoke applications
What is Agave?

/sysystems  /files  /apps  /jobs

- Register storage and compute systems
- Ingest, move and transform data files and folders
- Register applications (binaries) on large systems
- Launch jobs to invoke applications

/notifications

*All activities are events that can be subscribed to*
Agave Powers web & mobile apps
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach
Computational Science: Traditional Approach

Average Wait Time To Run A Simulation Over Time

Wait Time (Hours)

7/1/2012 1/1/2013 7/1/2013 1/1/2014 7/1/2014 1/1/2015

Timeframe

Trestles
Keeneland
Blacklight
Ranger
Lonestar
Stampede
Meanwhile, Data accumulating...

Cumulative growth in INSDC. (A) Base pairs (black, 2365.5 trillion) and sequence reads (blue, 17.8 trillion) for INSDC raw data.

Agave alone moves 1PB+ data/month

Aggressive purge policy of 2 weeks on TACC’s global /SCRATCH
Event Based Processing

- File or folder appears or is modified on a server...
  - Run a checksum
  - Launch a job to do some analysis
  - Compress the file
  - Move the file to archive storage
- Job completes
  - Job was successful, launch another job
  - Job failed, check inputs and launch again?
- Execution system goes offline for maintenance
  - Submit to a secondary system
- Storage system goes offline for maintenance
  - Submit jobs using data from a different system
- New user signs up for portal/project
  - Bootstrap storage and compute
Actor Model

Message Arrives

State Computation

Send messages to other actors

Create new actors

{ "uuid": "000141157089814", "event": "UPDATED", "updateTime": "2016-03-22T17:39:30.6:00", "owner": "jdoe" }
Actor Model: Inherently Concurrent
Actor Model: Inherently Concurrent
Actor Model: Inherently Concurrent
Actor Model: Inherently Concurrent
Actor Model: Inherently Concurrent
Actor Model: Inherently Concurrent
User-Defined Actors Via Docker

- Associate an actor with a Docker image.
- Assign the actor’s inbox to a unique URI.
- Launch a container from the image in response to a message.
Containers: Reproducible Environments

Isolated Userland Processes

Virtualized:
- Network
- I/O
- CPU and MEM

Containers:
- Include all dependencies
- Ease installation
- Start up in milliseconds

App 1
Libs

App 2
Libs

App 3
Libs

Container Runtime

Host Kernel/OS

Infrastructure
Abaco: Actor Based Containers

POST
https://api.tacc.utexas.edu/actors/184326

{ "uuid": "000141157089814", "event": "UPDATED", "updateTime": "2016-03-22T17:39:30.6:00", "owner": "jdoe" }

POST
https://api.tacc.utexas.edu/actors/184326/messages
Abaco: Agave Event Processors

- Notifications API in Agave allows users to subscribe to events
- Event subscriptions can be in of several forms:
  - Email
  - SMS
  - **Web callback** - in this case, Agave sends details of event in message payload.
- Events API coming this summer makes registering subscriptions even easier.
Abaco: Agave Event Processors

- Create a Docker image to “process” an event.
- Register the image as an actor in abaco.
- Register the actor’s inbox URI as the callback to a notification for the event.
Abaco Architecture
Challenges

Abaco in beta, available to select “friendly” users.

Potential issues:
- Accidentally subscribing to “way too many” events.
- Bugy containers “hanging” during execution.
Early Use Case: GSAF

Genome Sequencing and Analysis Facility: from sequencer to SNPs

- Raw genetic material sequenced by Illumina Sequencer.
- "short reads" files dumped to server - kicks off a chain of events.
  - Initial quality checks
  - Alignment routines.
  - Basic analyses: Single Nucleotide Polymorphism (SNP) calls, etc.
- Data moved to scientist’s storage system.
Conclusion

- Massive data collections make real-time processing more and more of a necessity.
- The actor model provides a simple yet robust paradigm for concurrent, event-driven programming.
- Containers can be used to provide portable, reproducible environments.
Thanks!

Questions?

Email: jstubbs@tacc.utexas.edu
Agave: http://agaveapi.co/
abaco: https://github.com/TACC/abaco