

StarCluster - NumPy/SciPy Computing on Amazon's Elastic Compute Cloud (EC2)

Justin Riley

Software Tools for Academics and Researchers
Office of Educational Innovation and Technology
Massachusetts Institute of Technology

SciPy 2010

Outline

- 1 Introduction
 - STAR Group
 - Motivation behind StarCluster
- 2 Amazon EC2 Basics
- 3 StarCluster Overview
 - Features
 - Configuration
 - Quick Demo
 - Customizing StarCluster
 - Using Elastic Block Storage
 - Creating Plugins
- 4 Conclusions
 - Future Work
 - Where can I learn more?

Outline

- 1 Introduction**
 - STAR Group
 - Motivation behind StarCluster
- 2 Amazon EC2 Basics
- 3 StarCluster Overview
 - Features
 - Configuration
 - Quick Demo
 - Customizing StarCluster
 - Using Elastic Block Storage
 - Creating Plugins
- 4 Conclusions
 - Future Work
 - Where can I learn more?

Software Tools for Academics and Researchers

- 1 Work with faculty at MIT to develop software for classroom/research

Software Tools for Academics and Researchers

- 1 Work with faculty at MIT to develop software for classroom/research
- 2 StarBiochem - Protein Visualization Tool

Software Tools for Academics and Researchers

- 1 Work with faculty at MIT to develop software for classroom/research
- 2 StarBiochem - Protein Visualization Tool
- 3 StarGenetics - Genetic cross-simulator

Software Tools for Academics and Researchers

- 1 Work with faculty at MIT to develop software for classroom/research
- 2 StarBiochem - Protein Visualization Tool
- 3 StarGenetics - Genetic cross-simulator
- 4 StarMolsim - Web-based MD/Quantum simulations

Software Tools for Academics and Researchers

- 1 Work with faculty at MIT to develop software for classroom/research
- 2 StarBiochem - Protein Visualization Tool
- 3 StarGenetics - Genetic cross-simulator
- 4 StarMolsim - Web-based MD/Quantum simulations
- 5 ... and more (<http://web.mit.edu/star>)

Motivations for StarCluster...

Cluster Configuration is Hard



Cluster Configuration

Motivations for StarCluster...

Cluster Configuration is Hard



- Obtaining access to hardware can be a challenge

Motivations for StarCluster...

Cluster Configuration is Hard



- Obtaining access to hardware can be a challenge
- Configuring and maintaining cluster configurations is hard

Motivations for StarCluster...

Cluster Configuration is Hard



- Obtaining access to hardware can be a challenge
- Configuring and maintaining cluster configurations is hard
- Traditional resources = administrative overhead

StarHPC...

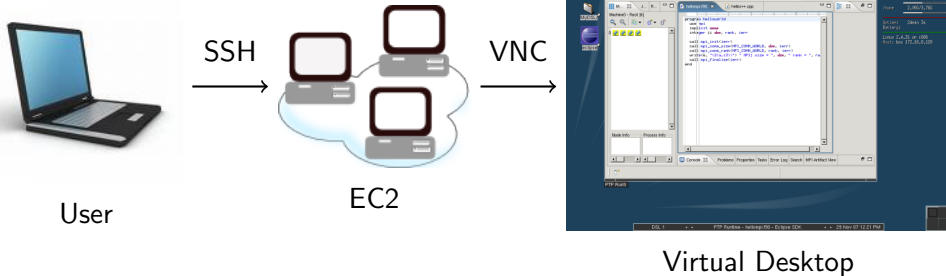
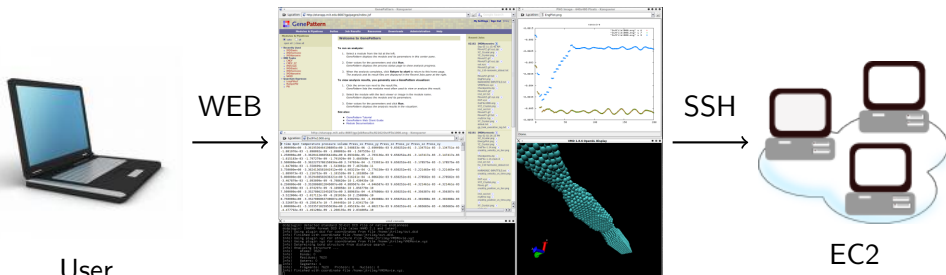


Figure: StarHPC Overview

StarMolSim...



MolSim using GenePattern

Figure: StarMolSim Overview

Outline

- 1 Introduction
 - STAR Group
 - Motivation behind StarCluster
- 2 Amazon EC2 Basics
- 3 StarCluster Overview
 - Features
 - Configuration
 - Quick Demo
 - Customizing StarCluster
 - Using Elastic Block Storage
 - Creating Plugins
- 4 Conclusions
 - Future Work
 - Where can I learn more?

Elastic Compute Cloud Overview

- Infrastructure as a Service (IaaS) Cloud Computing Model

Elastic Compute Cloud Overview

- Infrastructure as a Service (IaaS) Cloud Computing Model
- Request up to 20 virtual machines by default

Elastic Compute Cloud Overview

- Infrastructure as a Service (IaaS) Cloud Computing Model
- Request up to 20 virtual machines by default
- Full root access via SSH

Elastic Compute Cloud Overview

- Infrastructure as a Service (IaaS) Cloud Computing Model
- Request up to 20 virtual machines by default
- Full root access via SSH
- Only pay for what you use

Elastic Block Storage

- Analogous to a "Virtual USB pendrive"

Elastic Block Storage

- Analagous to a "Virtual USB pendrive"
- Size can be anywhere from 1GB-1TB per volume

Elastic Block Storage

- Analagous to a "Virtual USB pendrive"
- Size can be anywhere from 1GB-1TB per volume
- Supports snapshotting volumes to create backups

Elastic Block Storage

- Analagous to a "Virtual USB pendrive"
- Size can be anywhere from 1GB-1TB per volume
- Supports snapshotting volumes to create backups
- Ability to create new volumes based on snapshots

Standard Instances

Definition

1 Compute Unit (CU) = 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor.

Instance	Arch	CPU (CU)	RAM	Storage	I/O	Cost/hr
Small	32bit	1 (x1)	1.7GB	160GB	Moderate	\$0.085
Large	64bit	2 (x2)	7.5GB	860GB	High	\$0.34
X-Large	64bit	2 (x4)	15GB	1.69TB	High	\$0.68

High-Memory Instances

Definition

1 Compute Unit (CU) = 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor.

Instance	Arch	CPU (CU)	RAM	Storage	I/O	Cost/hr
X-Large	64bit	3.25 (x2)	17.1GB	420GB	Moderate	\$0.50
2X-Large	64bit	3.25 (x4)	34.2GB	850GB	High	\$1.20
4X-Large	64bit	3.25 (x8)	68.4GB	1.69TB	High	\$2.40

High-CPU Instances

Definition

1 Compute Unit (CU) = 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor.

Instance	Arch	CPU (CU)	RAM	Storage	I/O	Cost/hr
Medium	32bit	2.5 (x2)	1.7GB	160GB	Moderate	\$0.17
X-Large	64bit	2.5 (x8)	15GB	1.69TB	High	\$0.68

AWS Funding Opportunities

AWS In Education

<http://aws.amazon.com/education/>

AWS Funding Opportunities

AWS In Education

<http://aws.amazon.com/education/>

- Teaching Grants for educators using AWS in courses

AWS Funding Opportunities

AWS In Education

<http://aws.amazon.com/education/>

- Teaching Grants for educators using AWS in courses
- Research Grants for academic researchers using AWS in their work

AWS Funding Opportunities

AWS In Education

<http://aws.amazon.com/education/>

- Teaching Grants for educators using AWS in courses
- Research Grants for academic researchers using AWS in their work
- Project Grants for student organizations pursuing entrepreneurial endeavors

Questions?

?

Outline

- 1 Introduction
 - STAR Group
 - Motivation behind StarCluster
- 2 Amazon EC2 Basics
- 3 StarCluster Overview
 - Features
 - Configuration
 - Quick Demo
 - Customizing StarCluster
 - Using Elastic Block Storage
 - Creating Plugins
- 4 Conclusions
 - Future Work
 - Where can I learn more?

About



StarCluster allows anyone to create their own scientific computing cluster on Amazon's Elastic Compute Cloud (EC2)

Dependencies:

- Registered and fully configured EC2 account

About



StarCluster allows anyone to create their own scientific computing cluster on Amazon's Elastic Compute Cloud (EC2)

Dependencies:

- Registered and fully configured EC2 account
- Python 2.4+

About



StarCluster allows anyone to create their own scientific computing cluster on Amazon's Elastic Compute Cloud (EC2)

Dependencies:

- Registered and fully configured EC2 account
- Python 2.4+
- Boto (AWS library for Python)

About



StarCluster allows anyone to create their own scientific computing cluster on Amazon's Elastic Compute Cloud (EC2)

Dependencies:

- Registered and fully configured EC2 account
- Python 2.4+
- Boto (AWS library for Python)
- Paramiko (SSH library for Python)

StarCluster Features

- Simple configuration file for defining cluster settings

StarCluster Features

- Simple configuration file for defining cluster settings
- Single "start" command to create a cluster

StarCluster Features

- Simple configuration file for defining cluster settings
- Single "start" command to create a cluster
- "stop" command to terminate a cluster and stop paying for it

StarCluster Features

- Simple configuration file for defining cluster settings
- Single "start" command to create a cluster
- "stop" command to terminate a cluster and stop paying for it
- Automatic configuration of:

StarCluster Features

- Simple configuration file for defining cluster settings
- Single "start" command to create a cluster
- "stop" command to terminate a cluster and stop paying for it
- Automatic configuration of:
 - Network File System (/home and all EBS volumes)

StarCluster Features

- Simple configuration file for defining cluster settings
- Single "start" command to create a cluster
- "stop" command to terminate a cluster and stop paying for it
- Automatic configuration of:
 - Network File System (/home and all EBS volumes)
 - Sun Grid Engine

StarCluster Features

- Simple configuration file for defining cluster settings
- Single "start" command to create a cluster
- "stop" command to terminate a cluster and stop paying for it
- Automatic configuration of:
 - Network File System (/home and all EBS volumes)
 - Sun Grid Engine
 - Passwordless-ssh

StarCluster Features

- Simple configuration file for defining cluster settings
- Single "start" command to create a cluster
- "stop" command to terminate a cluster and stop paying for it
- Automatic configuration of:
 - Network File System (/home and all EBS volumes)
 - Sun Grid Engine
 - Passwordless-ssh
 - OpenMPI, etc

NumPy/SciPy on StarCluster

- Custom compiled Atlas/NumPy/SciPy for 8-core instance types

NumPy/SciPy on StarCluster

- Custom compiled Atlas/NumPy/SciPy for 8-core instance types
- Custom NumPy/SciPy Cookbook:
<http://starcluster.scripts.mit.edu/starcluster/wiki/index.php>

Configuration

- INI-based configuration file

Configuration

- INI-based configuration file
- "cluster templates" define cluster configuration

Example Config

```
1 [aws info]
2 aws_access_key_id = #your_aws_access_key_id
3 aws_secret_access_key = #your_secret_access_key
4 aws_user_id = #your_userid
5
6 [key mykeypair]
7 key_location = /home/myuser/.ssh/mykeypair.rsa
8
9 [ cluster smallcluster ]
10 cluster_size = 2
11 keyname = gsg-keypair
12 cluster_user = sgeadmin
13 cluster_shell = bash
14 node_image_id = ami-d1c42db8
15 node_instance_type = m1.small
```

Extending Cluster Templates

Re-using cluster template settings:

```
1 ....  
2  
3 [ cluster largecluster ]  
4 extends=smallcluster  
5   cluster_size =16  
6   node_image_id = ami-a5c42dcc  
7   node_instance_type = c1.xlarge
```

Brief Demo

```
1 $ starcluster start physicscluster
2 >>> Starting cluster ...
3 >>> Launching a 2-node cluster...
4 >>> Launching master node...
5 >>> Launching worker nodes...
6 >>> Waiting for cluster to start ...
7 >>> The master node is ec2-123-12-12-123.compute-1.amazonaws.com
8 >>> Attaching volume vol-99999999 to master node on /dev/sdz ...
9 >>> Setting up the cluster ...
10 >>> Mounting EBS volume vol-99999999 on /home...
11 >>> Creating cluster user: myuser
12 >>> Configuring scratch space for user: myuser
13 >>> Configuring /etc/hosts on each node
14 >>> Configuring NFS...
15 >>> Configuring passwordless ssh for user: myuser
16 >>> Installing Sun Grid Engine ...
```

Customizing StarCluster

How do I install my own software?

- Launch a single instance using either 32/64bit StarCluster AMI

Customizing StarCluster

How do I install my own software?

- Launch a single instance using either 32/64bit StarCluster AMI
- Login via ssh and install software

Customizing StarCluster

How do I install my own software?

- Launch a single instance using either 32/64bit StarCluster AMI
- Login via ssh and install software
- Use starcluster's "createimage" command to create a new custom AMI

Customizing StarCluster

How do I install my own software?

- Launch a single instance using either 32/64bit StarCluster AMI
- Login via ssh and install software
- Use starcluster's "createimage" command to create a new custom AMI
- Specify your new AMI id in the StarCluster configuration file

Elastic Block Storage

- Attached (mounted) to the master node

Elastic Block Storage

- Attached (mounted) to the master node
- NFS-shared to all nodes

Elastic Block Storage

- Attached (mounted) to the master node
- NFS-shared to all nodes
- All data written to EBS is persisted automatically

Creating New EBS Volume with StarCluster

How do we create new EBS volumes?

```
1 $ starcluster createvolume 20 us-east-1d
```

- This command automatically handles:

Creating New EBS Volume with StarCluster

How do we create new EBS volumes?

```
1 $ starcluster createvolume 20 us-east-1d
```

- This command automatically handles:
- Launching a "host" instance

Creating New EBS Volume with StarCluster

How do we create new EBS volumes?

```
1 $ starcluster createvolume 20 us-east-1d
```

- This command automatically handles:
- Launching a "host" instance
- Attaching the volume to the instance

Creating New EBS Volume with StarCluster

How do we create new EBS volumes?

```
1 $ starcluster createvolume 20 us-east-1d
```

- This command automatically handles:
- Launching a "host" instance
- Attaching the volume to the instance
- Partitioning the entire volume into a single partition

Creating New EBS Volume with StarCluster

How do we create new EBS volumes?

```
1 $ starcluster createvolume 20 us-east-1d
```

- This command automatically handles:
- Launching a "host" instance
- Attaching the volume to the instance
- Partitioning the entire volume into a single partition
- Formatting the volume with ext3 filesystem

Creating New EBS Volume with StarCluster

How do we create new EBS volumes?

```
1 $ starcluster createvolume 20 us-east-1d
```

- This command automatically handles:
- Launching a "host" instance
- Attaching the volume to the instance
- Partitioning the entire volume into a single partition
- Formatting the volume with ext3 filesystem
- Terminating the host instance

Plugin System

ubuntu.py

```
1 from starcluster .logger import log
2 from starcluster . clustersetup import ClusterSetup
3
4 class PackageInstaller ( ClusterSetup ):
5
6     def __init__ ( self , pkg_to_install ):
7         self . pkg_to_install = pkg_to_install
8
9     def run( self , nodes, master, user , user_shell , volumes):
10         for node in nodes:
11             log . info ( " Installing %s on node: %s" % \
12                 ( self . pkg_to_install , node . alias ) )
13             node . ssh . execute ( ' apt-get -y install %s ' % \
14                 self . pkg_to_install )
```

Plugin Config

Enabling the ubuntu plugin in the config

```
1 [plugin pkginstaller ]
2 setup_class = ubuntu.PackageInstaller
3 pkg_to_install = htop
```

Questions?

?

Outline

- 1 Introduction
 - STAR Group
 - Motivation behind StarCluster
- 2 Amazon EC2 Basics
- 3 StarCluster Overview
 - Features
 - Configuration
 - Quick Demo
 - Customizing StarCluster
 - Using Elastic Block Storage
 - Creating Plugins
- 4 Conclusions
 - Future Work
 - Where can I learn more?

Future Work

- Dynamic Load Balancing via Sun Grid Engine

Future Work

- Dynamic Load Balancing via Sun Grid Engine
- Supported plugins (ipcluster, mpi implementations, etc)

Where can I learn more?

Questions and Answers

Want to know more?

- Homepage: `http://web.mit.edu/starcluster`

Questions and Answers

Want to know more?

- Homepage: <http://web.mit.edu/starcluster>
- Code: <http://github.com/jtriley/StarCluster>

Questions and Answers

Want to know more?

- Homepage: <http://web.mit.edu/starcluster>
- Code: <http://github.com/jtriley/StarCluster>
- Mailing list:
<http://web.mit.edu/stardev/cluster/maillinglist.html>

Questions and Answers

Want to know more?

- Homepage: <http://web.mit.edu/starcluster>
- Code: <http://github.com/jtriley/StarCluster>
- Mailing list:
<http://web.mit.edu/stardev/cluster/maillinglist.html>
- Software Tools for Academics and Researchers:
<http://web.mit.edu/star>