Low Cost, Scalable Proteomics Data Analysis Using Amazon's Cloud Computing Services and Open Source Search Algorithms

ViPDAC

Brian D. Halligan, Ph.D.
Medical College of Wisconsin

http://proteomics.mcw.edu/vipdac

Halligan@mcw.edu
What is ViPDAC?

• ViPDAC => **Virtual Proteomics Data Analysis Cluster**

• One of the slowest parts of proteomics is data analysis.

• Single CPU machines analyze data much slower than instruments can generate it.

• Computer Clusters offer increased speed, but have high costs to implement and maintain.
Cloud Computing

• Distributed or Cloud computing allows for the use of virtual computers to perform computer intensive tasks without having to own the computer.
• Amazon has built large scale computing facilities that they offer for use on an hourly basis.
• The cost of analysis using this system is very low and the size of the cluster can expand, contract or even disappear based on need.
Amazon Web Services (AWS)

- **EC2 – Amazon Elastic Compute Cloud**
  
  “a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers.”

- **S3 - Amazon Simple Storage Service**
  
  “provides a simple web services interface that can be used to store and retrieve any amount of data, at any time, from anywhere on the web.”
Workflow

Turn on ViPDAC cluster via web

Job 1
- Pick parameters and database
- Upload mass spec data file
- Start analysis

Job 2
- Pick parameters
- Upload mass
- Start analysis

Job 3
- Pick parameters
- Upload mass
- Start analysis

Job 4
- Pick parameters and database
- Upload mass spec data file
- Start analysis

Monitor progress
- Download results when complete
- Turn Off Cluster

Add servers on the fly to go faster

Multiple analyses can be started and run at the same time
Time vs. Nodes
## ViPDAC Costs per Run

<table>
<thead>
<tr>
<th>Charge</th>
<th>Amount Used</th>
<th>Unit Size</th>
<th>Units</th>
<th>Cost / Unit</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EC2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EC2 - Data Transfer In</td>
<td>156 MB</td>
<td>1 GB</td>
<td>1</td>
<td>$0.10</td>
<td>$0.10</td>
</tr>
<tr>
<td>EC2 - Data Transfer Out</td>
<td>3.3 MB</td>
<td>1 GB</td>
<td>1</td>
<td>$0.17</td>
<td>$0.17</td>
</tr>
<tr>
<td>High CPU Instance (Medium)</td>
<td>2 instance-hr</td>
<td>1 instance-hr</td>
<td>2</td>
<td>$0.20</td>
<td>$0.40</td>
</tr>
<tr>
<td><strong>S3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Request - Tier 1</td>
<td>227</td>
<td>1,000</td>
<td>1</td>
<td>$0.01</td>
<td>$0.01</td>
</tr>
<tr>
<td>Request - Tier 2</td>
<td>394</td>
<td>10,000</td>
<td>1</td>
<td>$0.01</td>
<td>$0.01</td>
</tr>
<tr>
<td>C3 Data Transfer In</td>
<td>191 MB</td>
<td></td>
<td>No charge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3 Data Transfer Out</td>
<td>798 MB</td>
<td></td>
<td>No charge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Storage</td>
<td>36.6 MB</td>
<td>1 GB</td>
<td>1</td>
<td>$0.15</td>
<td>$0.15</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$0.84</td>
</tr>
</tbody>
</table>

Amount for requests is the number of messages and as indicated for data and computation.
Unit is the metric that Amazon uses to assess charges. Charges are assessed for any partial unit usage.
Advantages of ViPDAC

• Low cost
  – No startup costs
  – Low hourly usage costs
  – No cost when not in use

• Scalable
  – Everyone is first in line
  – Launch as few or as many worker nodes as needed
  – Fast costs the same as slow
  – 1 instance for 20 hrs = 20 instances for 1 hr
Advantages of ViPDAC

• Secure
  – Data is stored and transferred in a secure system
  – Your data/database does not leave your control and is not seen or shared with others

• Stable
  – AMI can be cloned and saved
  – Consistent data analysis for long term projects
  – SOP across laboratories
Advantages of ViPDAC

• Cost Accounting
  – Very easy to determine cost of a single run with ViPDAC compared to physical cluster

• Freedom to experiment
  – Can perform complex analysis on a dataset without blocking routine analysis
  – Custom interface or analysis
Amazon.com Sign In

You may sign in using your existing Amazon account or you can create a new account by selecting "I am a new user."

Enter your e-mail address: bdhalligan@earthlink.net

- I am a new user.
- I am a returning user, and my password is:

Forgot your password? Click here.
Change your name, e-mail address, or password for your Amazon account.

About Amazon.com Sign In

Amazon.com Sign In allows you to log in to applications that use Amazon technology using your Amazon.com account. To protect your information, you should only enter your Amazon.com e-mail address and password into a web site if the address of the site starts with https://www.amazon.com (see your browser address bar to check this). You can find out about how your personal information is protected in our Privacy Notice.
To start using Amazon EC2 you will want to launch a virtual server, known as an Amazon EC2 instance.

Launch Instances
### Amazon Machine Images

<table>
<thead>
<tr>
<th>AMI ID</th>
<th>Manifest</th>
<th>Visibility</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>ami-0022c769</td>
<td>level22-ec2-images/ubuntu-7.04-feisty-base-20071225a.manifest.xml</td>
<td>Public</td>
<td></td>
</tr>
<tr>
<td>ami-005db969</td>
<td>aelastic-64/ubuntu-8.04-hardy-base-64-20081222.manifest.xml</td>
<td>Public</td>
<td></td>
</tr>
<tr>
<td>ami-005bb069</td>
<td>keevo-dnt-db/imdnter-32bit-FC-DB.manifest.xml</td>
<td>Public</td>
<td></td>
</tr>
<tr>
<td>ami-014da868</td>
<td>sml.urium.com/image_bundles/Fedora6_Jetty_OpenBD/image.manifest.xml</td>
<td>Public</td>
<td></td>
</tr>
<tr>
<td>ami-015db968</td>
<td>aelastic-64/ubuntu-8.04-hardy-rocket-64-20081222.manifest.xml</td>
<td>Public</td>
<td></td>
</tr>
<tr>
<td>ami-01789c68</td>
<td>thamtech-ami-images/gsampsync/image.manifest.xml</td>
<td>Public</td>
<td></td>
</tr>
<tr>
<td>ami-018e6b68</td>
<td>rbuilder-online/phione/home-1.5.5-x86_64-10132.img.manifest.xml</td>
<td>Public</td>
<td></td>
</tr>
<tr>
<td>ami-01c92c68</td>
<td>rbuilder-online/bt-0.01-x86_17328.img.manifest.xml</td>
<td>Public</td>
<td></td>
</tr>
<tr>
<td>ami-01cs2c68</td>
<td>scoll.images/www.x84-2008-02-19-0921/image.manifest.xml</td>
<td>Public</td>
<td></td>
</tr>
</tbody>
</table>

Select an AMI above to view information about it here.
Launch Instances

Please review the information below, then click Launch.

**AMI Name:** Other Linux (ami-52f5123b, i386)

**Number of Instances**: 1

**Instance Type (32 bit):**
- Small (m1.small)
- High-CPU Medium (c1.medium)

**Key Pair Name**: vpddac_bh

**Security Groups**: default

**Kernel ID**: Select One

**Ramdisk ID**: Select One

**Availability Zone**: No Preference

You will be charged the hourly rate for any instances you launch until you successfully shut them down.

Launch
Launch Instances

Your instances are now launching.
Note: Your instances may take a few minutes to launch, depending on the software you are running.

> View your instances on the Instances page

---

Other EC2 Features

**Volumes**
EBS Volumes provide off-instance storage that persists independently of the life of an instance. Add a persistent storage device to an instance using the Elastic Block Store (EBS) Volumes page.

> Go to the Volumes Page

**Elastic IPs**
Elastic IP addresses allow you to remap a public IP address to any instance in your account. Elastic IPs also enable you to engineer around problems by quickly remapping your Elastic IP address to a replacement instance.

> Go to the Elastic IPs Page

Close
AWS Management Console - Mozilla Firefox

Navigation
- EC2 Dashboard
- Images & Instances
  - Instances
  - AMIs
- Bundle Tasks
- Elastic Block Store
  - Volumes
  - Snapshots
- Configuration
  - Elastic IPs
  - Key Pairs
  - Security Groups

My Instances

<table>
<thead>
<tr>
<th>Instance</th>
<th>AMI ID</th>
<th>Security Groups</th>
<th>Type</th>
<th>Status</th>
<th>Public DNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>i-9825b1f1</td>
<td>ami-52f5123b</td>
<td>default</td>
<td>c1 medium</td>
<td>starting</td>
<td></td>
</tr>
</tbody>
</table>

0 EC2 Instances selected

Select an instance above to view information about it here
### My Instances

<table>
<thead>
<tr>
<th>Instance</th>
<th>AMI ID</th>
<th>Security Groups</th>
<th>Type</th>
<th>Status</th>
<th>Public DNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>i-9825b1f1</td>
<td>ami-52f5123b</td>
<td>default</td>
<td>c1.medium</td>
<td>running</td>
<td>ec2-174-129-95-51.compute-1.amazonaws.com</td>
</tr>
</tbody>
</table>

0 EC2 Instances selected

Select an instance above to view information about it here.
Welcome to the Virtual Proteomics Data Analysis Cluster

VIPDAC tutorials and help files are available at http://proteomics.mcw.edu/vipdac

Information about Visualize is available at http://proteomics.mcw.edu/visualize

Blast version: 2.2.19
OMSSA version: 2.1.4
X Tandem version: 08-12-01-1

VIPDAC is a product of the Medical College of Wisconsin.
Nodes

All AWS Nodes: 1
Active AWS Nodes: 1
DB Nodes: 1

<table>
<thead>
<tr>
<th>ID</th>
<th>Size</th>
<th>Age</th>
<th>State</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>-9825b1f1</td>
<td>Medium</td>
<td>3 Minutes 1 Second</td>
<td>running</td>
<td></td>
</tr>
</tbody>
</table>
New Node

<table>
<thead>
<tr>
<th>Instance Type:</th>
<th>Choose an instance type...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantity:</td>
<td>1</td>
</tr>
</tbody>
</table>

Save  Cancel
### New Node

<table>
<thead>
<tr>
<th>Instance Type</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium</td>
<td>1</td>
</tr>
</tbody>
</table>

- [Cancel](#)
1 Node(s) were successfully launched.

## Nodes

All AWS Nodes: 2  
Active AWS Nodes: 2  
DB Nodes: 2

<table>
<thead>
<tr>
<th>ID</th>
<th>Size</th>
<th>Age</th>
<th>State</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>0825b1f1</td>
<td>Medium</td>
<td>5 Minutes 14 Seconds</td>
<td>running</td>
<td></td>
</tr>
<tr>
<td>3c26b255</td>
<td>Medium</td>
<td>0.25 Seconds</td>
<td>pending</td>
<td>Destroy</td>
</tr>
</tbody>
</table>
# Nodes

All AWS Nodes: 2  
Active AWS Nodes: 2  
DB Nodes: 2

<table>
<thead>
<tr>
<th>ID</th>
<th>Size</th>
<th>Age</th>
<th>State</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>i-9225b111</td>
<td>Medium</td>
<td>6 Minutes 37 Seconds</td>
<td>running</td>
<td></td>
</tr>
<tr>
<td>i-3c26b256</td>
<td>Medium</td>
<td>1 Minute 23 Seconds</td>
<td>running</td>
<td>Destroy</td>
</tr>
</tbody>
</table>
### New Job

<table>
<thead>
<tr>
<th>Name:</th>
<th>My Demo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority:</td>
<td>Low</td>
</tr>
<tr>
<td>Spectra count per chunk:</td>
<td>200</td>
</tr>
<tr>
<td>Data File:</td>
<td>short_test</td>
</tr>
<tr>
<td>Search Application:</td>
<td>OMSSA</td>
</tr>
<tr>
<td>Parameter File:</td>
<td>ipi_test</td>
</tr>
</tbody>
</table>

[Save] [Cancel]
Job was successfully launched.

## Jobs

<table>
<thead>
<tr>
<th>Name</th>
<th>Status</th>
<th>Progress</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>My Demo</td>
<td>Launching</td>
<td>N/A</td>
<td>Delete</td>
</tr>
</tbody>
</table>
**My Demo**

Datafile name: Mukut_7445_81.mgf
Status: Complete
Priority: Low
Searcher: OMSSA
Spectra Count: 200
Processing Time (in seconds): 35.454
Progress: 100%
Max chunk time: 8.330
Min chunk time: 5.103
Avg chunk time: 6.106
Total chunks: 5
Pending chunks: 0
Working chunks: 0
Complete chunks: 5

**Last 10 working chunks**

<table>
<thead>
<tr>
<th>Status</th>
<th>MGF File</th>
<th>Filesize</th>
<th>Node - PID</th>
<th>Job ID</th>
<th>Searcher</th>
<th>Sent at (UTC)</th>
<th>Processing Time (in seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Last 10 recently completed chunks**

<table>
<thead>
<tr>
<th>Status</th>
<th>MGF File</th>
<th>Filesize</th>
<th>Node - PID</th>
<th>Job ID</th>
<th>Searcher</th>
<th>Sent at (UTC)</th>
<th>Processing Time (in seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>Mukut_7445_81.00000002.mgf</td>
<td>1049316</td>
<td>i-9825bf1f-1362</td>
<td>1</td>
<td>OMSSA</td>
<td>02-19-2009 10:33pm</td>
<td>5.429</td>
</tr>
</tbody>
</table>
### Result files

<table>
<thead>
<tr>
<th>Name</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>ipinoenzyme_Mukut_7445_81_iPi.HUMAN.v3.54_omssa.ipinoenzyme_50</td>
<td>Download</td>
</tr>
<tr>
<td>mydemo_Mukut_7445_81_iPi.RAT.v3.54_omssa.ipitest_200</td>
<td>Download</td>
</tr>
<tr>
<td>myjob_Mukut_7445_81_iPi.RAT.v3.54_omssa.ipitest_100</td>
<td>Download</td>
</tr>
<tr>
<td>shortest_Mukut_7445_81_iPi.RAT.v3.54_omssa.ipitest_50</td>
<td>Download</td>
</tr>
<tr>
<td>usp1job_USP1_DS_19jul08_25.H_sapiens_omssa.usp_200</td>
<td>Download</td>
</tr>
</tbody>
</table>

Displaying all 5 result files
Nodes

Create new node

All AWS Nodes: 2
Active AWS Nodes: 2
DB Nodes: 2

<table>
<thead>
<tr>
<th>ID</th>
<th>Size</th>
<th>Age</th>
<th>State</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>b5625b1f1</td>
<td>Medium</td>
<td>13 Minutes 34 Seconds</td>
<td>running</td>
<td></td>
</tr>
<tr>
<td>b3c26b255</td>
<td>Medium</td>
<td>8 Minutes 20 Seconds</td>
<td>running</td>
<td>Destroy</td>
</tr>
</tbody>
</table>
# Nodes

All AWS Nodes: 2  
Active AWS Nodes: 1  
DB Nodes: 1

<table>
<thead>
<tr>
<th>ID</th>
<th>Size</th>
<th>Age</th>
<th>State</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>-8825b11</td>
<td>Medium</td>
<td>14 Minutes 55 Seconds</td>
<td>running</td>
<td></td>
</tr>
</tbody>
</table>
### My Instances

<table>
<thead>
<tr>
<th>Instance</th>
<th>AMI ID</th>
<th>Security Groups</th>
<th>Type</th>
<th>Status</th>
<th>Public DNS</th>
<th>Key Pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>i-5825b1f1</td>
<td>ami-52f5123b</td>
<td>default</td>
<td>c1.medium</td>
<td>running</td>
<td>ec2-174-129-95-51.compute-1</td>
<td></td>
</tr>
<tr>
<td>i-3c26b255</td>
<td>ami-52f5123b</td>
<td>default</td>
<td>c1.medium</td>
<td>terminated</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**0 EC2 Instances selected**

*Select an instance above to view information about it here*
Acknowledgments

- Joey F. Geiger
- Andrew K. Vallejos
- Simon N. Twigger
- Andrew S. Greene
  – MCW NHLBI Proteomics Center

http://proteomics.mcw.edu/vipdac
<table>
<thead>
<tr>
<th>Protein Information</th>
<th>Scan Count</th>
<th>Protein Probability</th>
<th>Peptide Count</th>
<th>Peptide Filter</th>
<th>Scan Count</th>
<th>Reference</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>tau Homo sapiens</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**OMSSA Data**

- **Scan number**: USP1_DS_19jul08.8545.8545.3.dta
- **Matched Protein**: TAU_HUMAN
- **Matched sequence**: HLSNVSSTGSDMVDSPQLATLDEVASLAKQ
- **Peptide Probability**: 1.0000
- **OMSSA e value**: 2.075E-011
- **OMSSA Probability**: 3.474E-016
- **Peptide Start**: 378
- **Peptide Stop**: 409
- **Observed Mass**: 3259.5600
- **Calculated Mass**: 3259.5900
- **Delta mass**: 0.0300
- **Charge**: 3
- **Ions total**: 3
<table>
<thead>
<tr>
<th>Protein</th>
<th>Database</th>
<th>Peptide</th>
<th>Scan Count</th>
<th>Peptide Count</th>
<th>Peptide Filer</th>
<th>Accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAU_HUMAN</td>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DATA_HUMAN</td>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PROX1_HUMAN</td>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TXK27_HUMAN</td>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALBU_HUMAN</td>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIRI_HUMAN</td>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BID_HUMAN</td>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BICALH1_HUMAN</td>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYB5F_HUMAN</td>
<td>Homo sapiens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**OMSSA Data**

- **Scan number**: USP1_DS_19Jul08.8546.8546.3.ota
- **Matched Protein**: TAU_HUMAN
- **Matched sequence**: RHLNVSSTGSDMVPQLATLDEVASLAKQ
- **Peptide Probability**: 0.0000
- **OMSSA q value**: 2.075E-011
- **OMSSA Probability**: 3.474E-016
- **Peptide Start**: 378
- **Peptide Stop**: 409
- **Observed Mass**: 3255.5600
- **Calculated Mass**: 3255.5900
- **Delta mass**: 0.0300
- **Charge**: 3

**Ions total**
Visualize version 1.15

There are 81 proteins, 81 above threshold 100.0% of total scans.

Export | Search | Help

WinScp is an Open Source FTP client for Windows.

OMSSA Data
- Scan number: USP1_DS_19jul08.8545.8545.3.dta
- Matched Protein: TAU_HUMAN
- Matched sequence: HLSNVSSTGSIDMVDSPQLATLAEVSASLAK.Q
- Pecptide Probability: 1.0000
- OMSSA evalue: 2.075E-011
- OMSSA Probability: 3.474E-016
- Peptide Start: 378
- Peptide Stop: 409
- Observed Mass: 3258.5600
- Calculated Mass: 3258.5900
- Delta mass: 0.0300
- Charge: 3

Ions total
Acknowledgments

- Joey F. Geiger
- Andrew K. Vallejos
- Simon N. Twigger
- Andrew S. Greene
  - MCW NHLBI Proteomics Center

http://proteomics.mcw.edu/vipdac