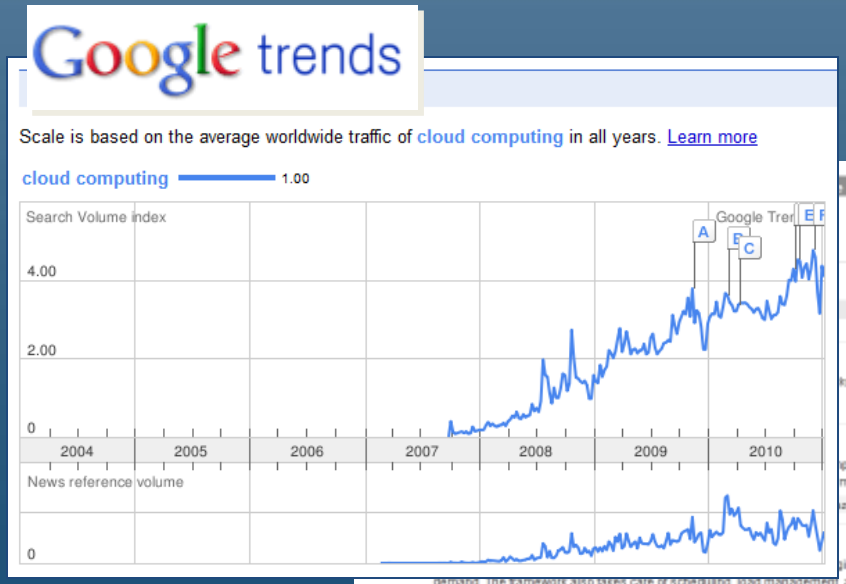
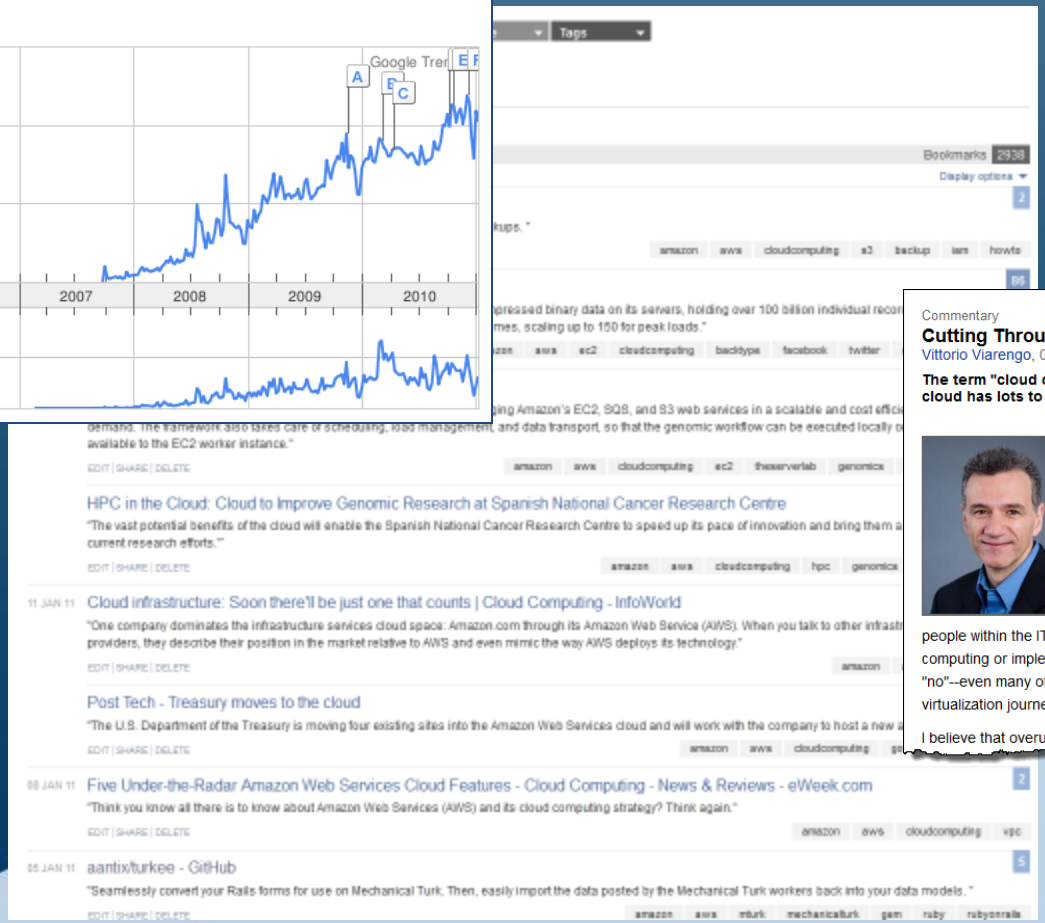

amazon web services™
Cloud Computing for Research
Jeff Barr - January 2011
Amazon EC2
Amazon Elastic MapReduce
Amazon S3
Amazon RDS
Elastic Load Blancing
Amazon Mechanical Turk
Amazon Simple
AWS Import/Export
Amazon SimpleDB
Amazon Flexible Payment Service
AWS Premium Support
Amazon Virtual Private Cloud
Amazon Simple Notification Service
Auto Scaling
Amazon Elastic Block Storage

The Cloud is Suddenly Everywhere

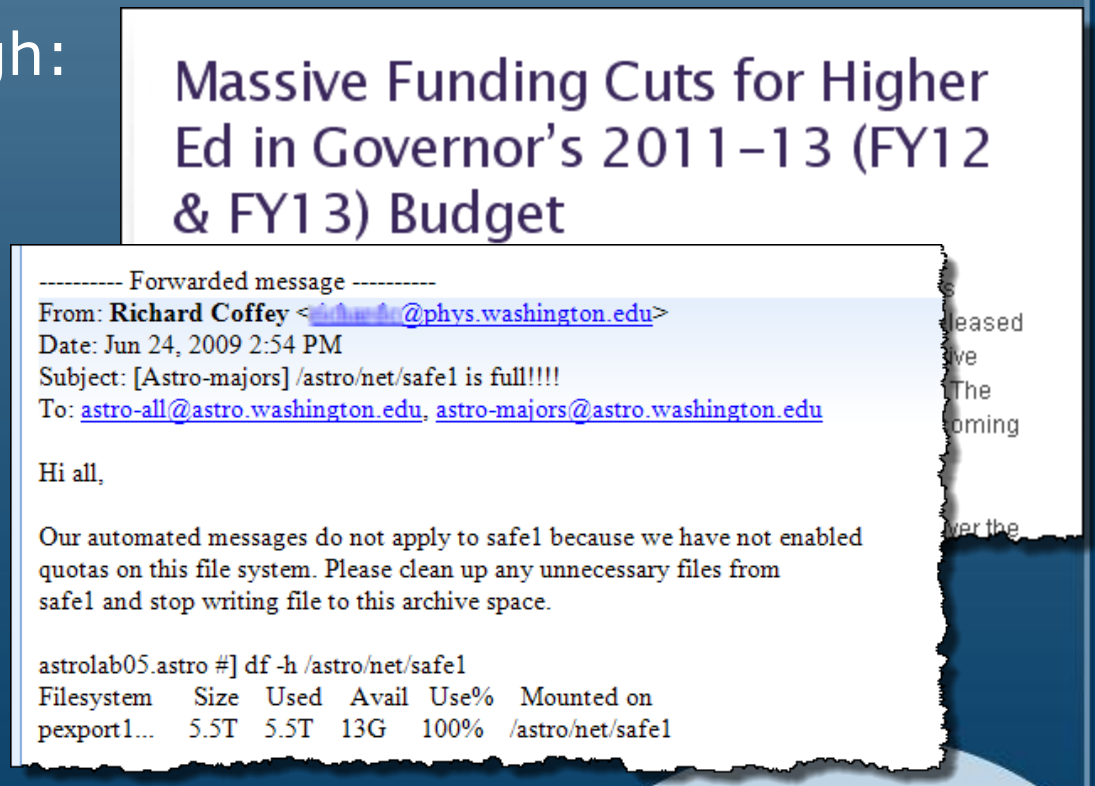




amazon web services™

Current Research Challenges

- There is never enough:
  - Time
  - Money
  - CPU power
  - Storage
  - Physical space
  - Power or cooling



amazon web services

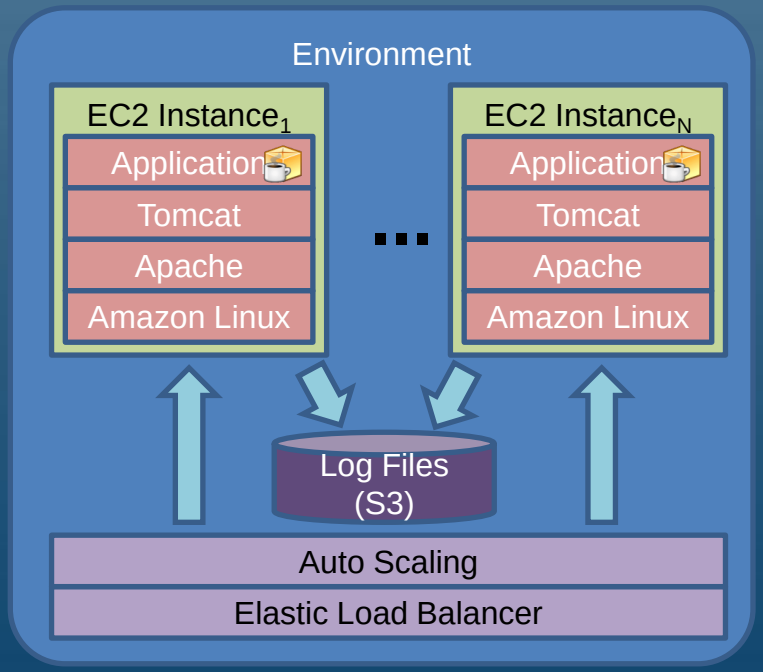
How the Cloud Can Help

- Time
  - Less time spent on system & infrastructure issues.
  - More time for research.
- Money
  - Economical, pay-as-you-go access to resources.
  - AWS in Education program.
- CPU power
  - As little or as much as you need (0 – 10^4 cores and beyond).
- Storage
  - GB, TB, PB on demand.
- Physical space
  - No longer "the final frontier."
- Power or cooling
  - Our problem, not yours.

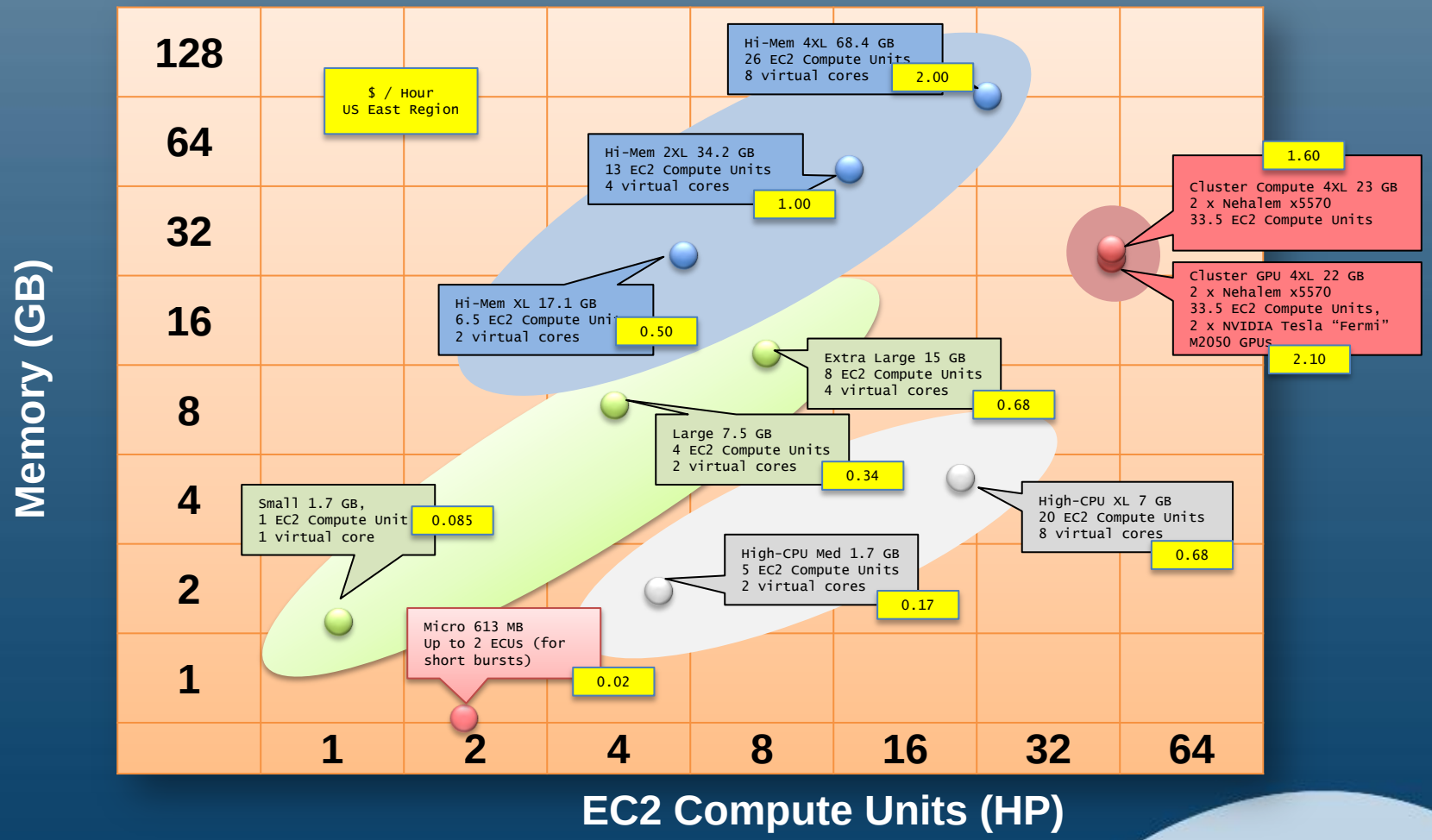
amazon web services™

Amazon Web Services (AWS)

- Compute
  - Elastic Compute Cloud (EC2)
  - Auto Scaling + Elastic Load Balancing
  - AWS Elastic Beanstalk
- Storage
  - Simple Storage Service (S3)
  - Elastic Block Storage (EBS)
- Database
  - Relational Database Service (RDS)
  - Amazon SimpleDB
  - Third-Party offerings
- Messaging
  - Simple Queue Service
  - Simple Notification Service

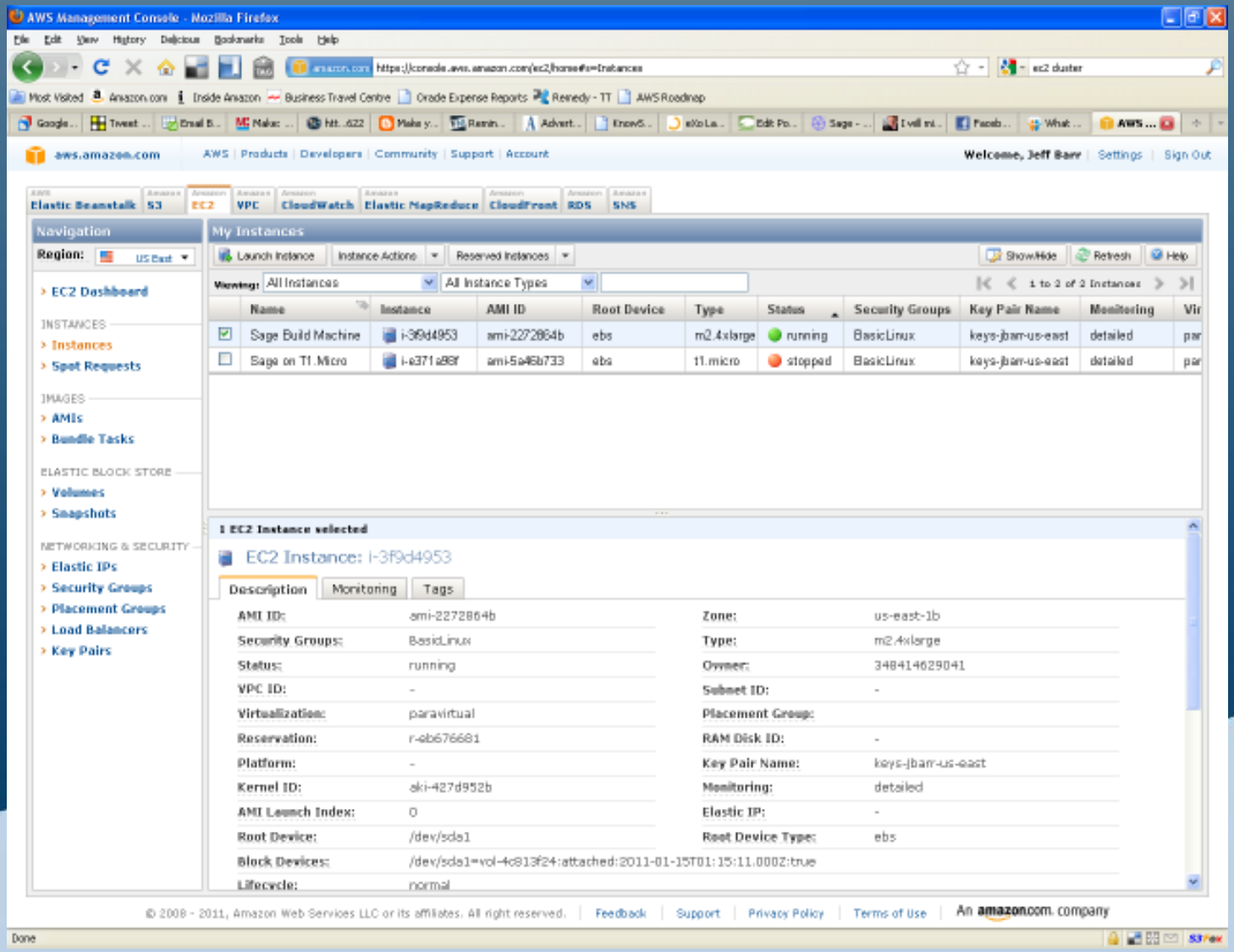


amazon web services™



amazon web services™

AWS Management Console



amazon web services

AWS Security

- Certifications and Validations:
  - SAS 70 Type II
  - PCI DSS
  - FISMA Low
- Security White Paper
- HIPAA White Paper
- Physical Security
  - Military-grade perimeters
  - Non-descript facilities
  - 3+ levels of two-factor auth
- Data Security
  - Redundant data storage
  - SSH keys for EC2 access
  - Stateful firewall / security groups
  - Identity and Access Management (IAM)
  - Multifactor Authentication

amazon web services™

AWS Sample Pricing

- Micro server 24x7, 10 GB disk, 15 GB in, 15 GB out
  - $0 (AWS Free Tier for new customers)
- Small server, 24x7, 100 GB Disk, 5 GB in, 5 GB out
  - $97.12 / month
- Large server, 24x7, 250 GB Disk, 20 GB in, 20 GB out
  - $338.88 / month *On-Demand*
  - $177.84 / month + $1400 3 Year *Reserved Instance*
- 8 Cluster Compute for 24 hours, 1 GB in, 1 GB out
  - $153.60 *On-Demand*
- All prices in US East as of 2011-02-01

amazon web services™

AWS in Education

Enable the worldwide academic community to easily leverage the benefits of Amazon Web Services for teaching and research.

- **Teaching Grants** for educators using AWS in courses (plus access to selected course content resources).
- **Research Grants** for academic researchers using AWS in their work.
- **Project Grants** for student organizations pursuing entrepreneurial endeavors; Tutorials for students that want to use AWS for self-directed learning.
- **Solutions** for university administrators looking to use cloud computing to be more efficient and cost-effective in the university's IT Infrastructure.

amazon web services™

AWS in Education Success Stories

University of Maryland, College Park

"Thanks to the generous support of Amazon Web Services, students were provided AWS access through my teaching grant which they applied to coursework. The ability to provision Hadoop clusters on-demand gave students hands-on experience with utility computing and provided a vehicle for completing coursework and a final project."

—Jimmy Lin, Associate Professor, University of Maryland, College Park

University of California at Berkeley

"Using AWS for our Web 2.0 Application Development courses has been a phenomenal resource. Administration was so easy that students were able to get their projects deployed quickly, and venture capitalists attending the final project demos were impressed at the level of polish and creativity that a small student team could produce in just a few weeks."

—Armando Fox, Adjunct Associate Professor, University of California at Berkeley

University of California, San Diego

"The generous grant we have received from the AWS in Education program will dramatically increase the reach of my student group's collaboration with the National Federation of the Blind (NFB) to develop distributed assistive computer vision technology for the visually impaired."

—Serge Belongie, Associate Professor, University of California, San Diego

amazon web services™

Use of Cloud Computing for Bioinformatics

A collaboration between the San Francisco State University CS Department, SFSU Center for Computing and Life Sciences and Stanford Helix Group

Jinesh Lalan, Mike Wong and Dr. Dragutin Petkovic

Use of cloud computing for bioinformatics: case study of Amazon Elastic Compute Cloud EC2

Supported in part by NIH grant LM05652, SFSU Center for Computing for Life Sciences and Amazon Educational Grant

Background

- Genomic and proteomic sequencing data have grown tremendously in size because of advancing in sequencing technology
- Consequently, analysis is intractably computationally intensive for common PC architectures and building and operating even a relatively small cluster can be a formidable undertaking.
- Recently, virtual computing clouds are drawing attentions for its flexibility since users can create a high performance cluster of any number of virtual computing units

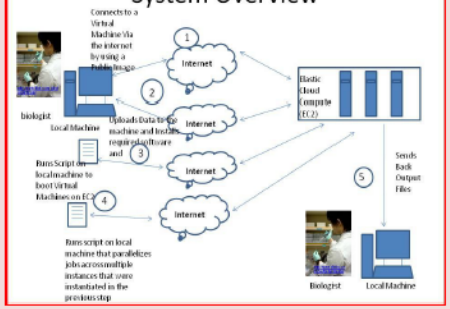
Purpose

- Investigate applicability of Amazon Cloud Computing EC2 for typical bioinformatics applications, specifically for performance and ease of use
- Provide tutorials and setup SW for non-expert users

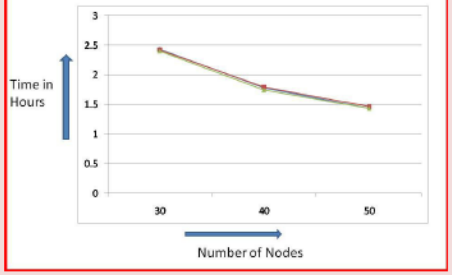
Methods

- We downloaded the Anopheles Gambiae Gene data from Vector Base.
- Installed BLAST [1] Software on the Amazon Machine Image. Already installed on the Cluster
- We ran BLAST on EC2 with 30, 40 and 50 instances and measured the run time of each. At least three runs of each number of instances were carried out. Instance type used for the research was "Small" (1.7 GB of Memory, 1 -1.2 GHz Opteron or Xeon Processor, 1 virtual core)
- We also carried out BLAST runs on a local cluster of 36 cores, 152 GB of Memory and 72 GB of Disks

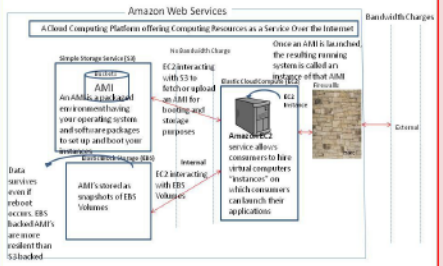
System Overview



Results



Amazon Web Services



Discussion

- Looking at the graphs, the **performance** seems to be stable and **scales linearly**.
- **Noisy Neighbor Problem**. A noisy neighbor here means a Virtual Machine that uses the shared resources disproportionately. Sadly, you cannot choose your neighbors
- Network Bandwidth issues (Internal as well as external)
- Provides Simple and Easy tutorials to get started using AWS services
- Various alternative Instance types. Instance types specially designed for bioinformatics. Various instance types have different compute capacity
- **Public Datasets** on AWS can be accessed with no charge

Cloud Computing Vs. Local Cluster

- Cheap→ Zero cost of equipments, physical space, Cooling, Power, maintenance
- Highly Scalable and Elastic → Can scale Resources up and down as an when you want
- Your Jobs don't have to wait in the Queue
- Root access to your Nodes
- Researcher with no access to a local cluster can have their results cheaply within hours.
- Expensive→ Heavy cost of equipments, physical space, cooling, power, highly paid system admin
- Fixed Number of Resources → Cluster Owners have to take care that the Cluster is not over or under utilized
- Your Jobs have to share the resources with other cluster users
- No root Access

Genome Analysis Resources on AWS

- Publicly available Amazon Machine Images and Volumes for Bioinformatics
  a. *ami-4e57a227* – 64 bit image
  b. *ami-0af91263* – 32 bit image
  c. *snap-84e771ef* – Data volume [5]
- Some of the tools installed on the above AMI's : PRIMER3, MUSCLE, BLAST2, EMBOSS, HMMER, READSEQ, BOWTIE, NOVOALIGN [5]
- Databases available on the above mentioned volume
  a. Genome sequences pre-indexed for search with next-gen aligners like Bowtie, Novoalign
  b. UniRef protein databases, indexed for searching with BLAST+ [5]

References

1. NCBI/ BLAST [http://blast.ncbi.nlm.nih.gov/Blast.cgi]
2. Amazon Web Services. [http://aws.amazon.com]
3. EC2. [http://aws.amazon.com/ec2/]
4. S3. [http://aws.amazon.com/s3/]
5. CloudBioLinux [http://www.cloudbiolinux.com/]

Acknowledgements

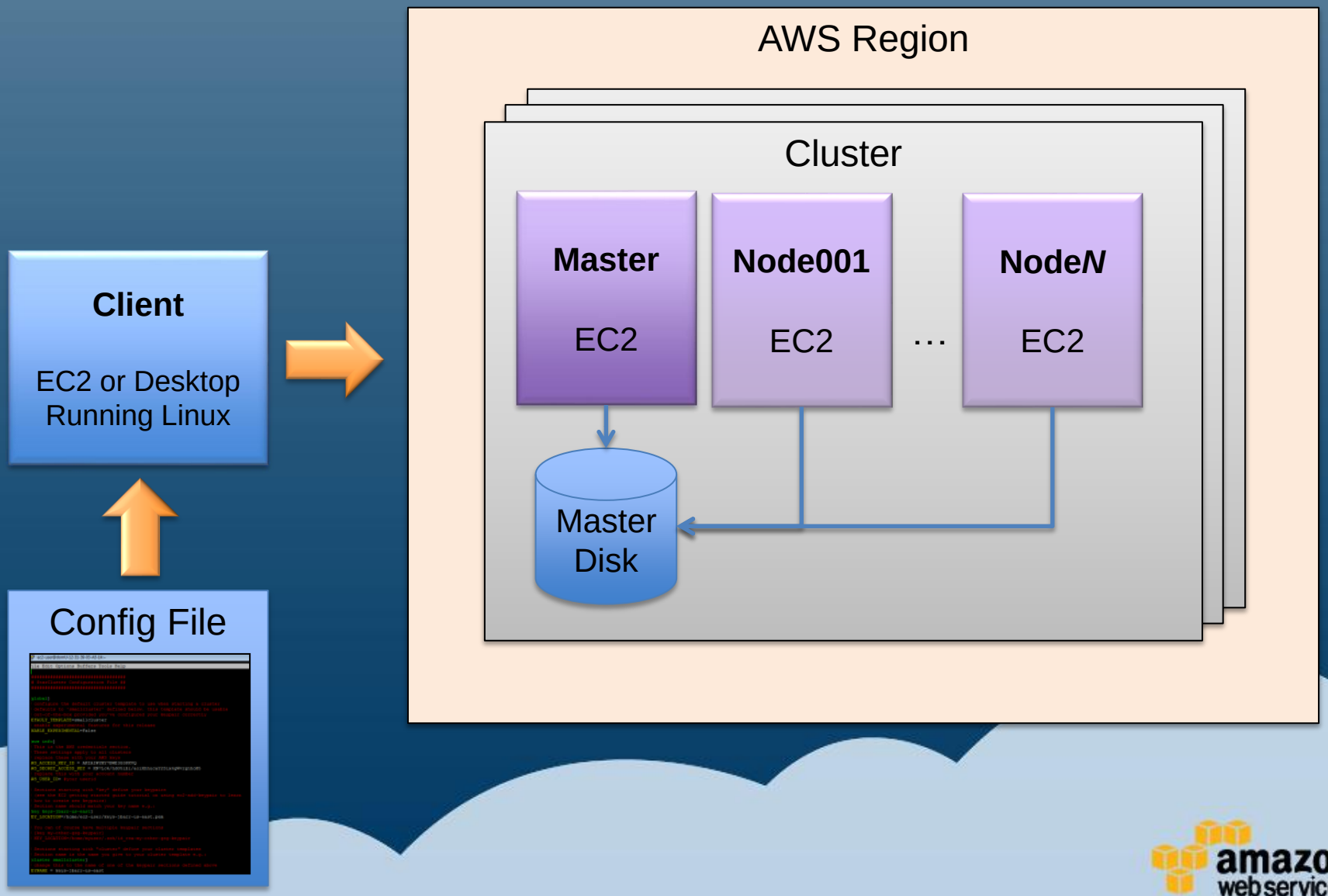
I would like to thank the gracious support and guidance of Dr. Russ Altman, Dr. Dragutin Petkovic, Mike Wong, Ljubomir Buturovic

amazon web services™

Introducing MIT StarCluster

- Open source cluster
- Simplifies creation and management of EC2 clusters
  - OpenMPI
  - Oracle Grid Engine
- Launch a cluster of EC2 instances:
  - One command (“starcluster”) to rule them all
  - Passwordless SSH pre-configured
  - Security group for SSH access
  - Shared disk volume (NFS)
  - Preinstalled libraries (NumPy, SciPy, etc.)

StarCluster Architecture / Terminology



amazon web services™

StarCluster Commands

$ starcluster help

$ starcluster start jb1

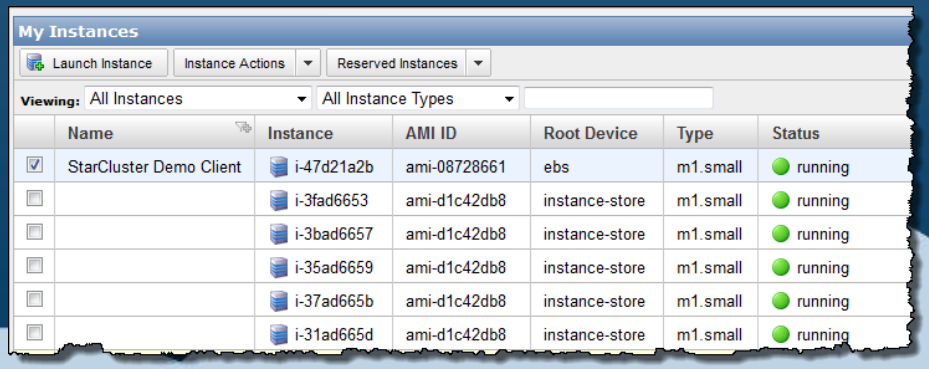
$ starcluster sshmaster jb1

$ starcluster listclusters

$ starcluster stop jb1

```
<client>: starcluster listclusters
StarCluster - (http://web.mit.edu/starcluster)
Software Tools for Academics and Researchers (STAR)
Please submit bug reports to starcluster@mit.edu


-----------------------------
jb1 (security group: @sc-jb1)
-----------------------------
Launch time: 2011-01-14T05:43:44.000Z
Zone: us-east-1c
Keypair: keys-jbarr-us-east
Cluster nodes:
     master running i-3fad6653 ec2-50-16-41-160.compute-1.amazonaws.com
    node001 running i-3bad6657 ec2-184-73-107-91.compute-1.amazonaws.com
    node002 running i-35ad6659 ec2-174-129-124-218.compute-1.amazonaws.com
    node003 running i-37ad665b ec2-50-16-32-211.compute-1.amazonaws.com
    node004 running i-31ad665d ec2-50-16-31-114.compute-1.amazonaws.com
```

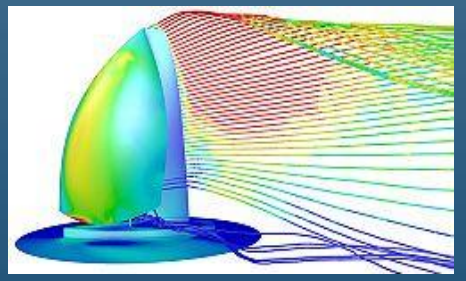


amazon web services

AWS - HPC Use Cases

- CFD – Computational Fluid Dynamics
  - OpenFOAM on EC2
  - CloudFlu
- Molecular Modeling
  - Eli Lilly, Pfizer
- Sequence Analysis
  - CloudBioLinux
- Engineering Design
- Energy Trading & Financial Modeling
- I/O-intensive Applications
- Graphics / 3D Rendering

OpenFOAM

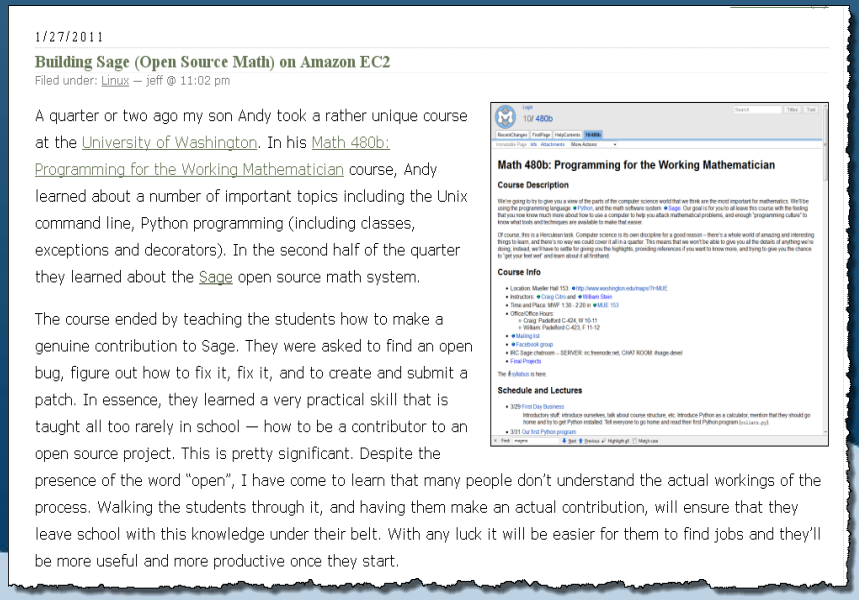


Cloud BioLinux

amazon web services

AWS at the UW

- I'm available for on-campus meetings on Wednesdays
- 10 Gbit connection to AWS via CENIC/PNW Gigapop



amazon web services™

Getting Started

- AWS Home page: http://aws.amazon.com
- AWS in Education: http://aws.amazon.com/education/
- AWS Security: http://aws.amazon.com/security
- AWS Blog: http://aws.typepad.com
- AWS Calculator: http://calculator.s3.amazonaws.com/calc5.html
- StarCluster: http://web.mit.edu/stardev/cluster/
- Me:
  - jbarr@amazon.com
  - @jeffbarr on Twitter
  - http://www.jeff-barr.com

amazon web services™