



European Organization for Nuclear Research

Cloud Science

or..

Astrometric Data Processing in Amazon EC2

CERN, Wednesday October 28th, 2009

Paul Parsons, CTO, The Server Labs



October 09





Paul Parsons

- Founder & CTO of The Server Labs
- > Over 20 years experience in many fields of IT





- Two main objectives
 - Evaluate the Tecnical Feasibility of using Amazon EC2 to run scientific data processing applications
 - Evaluate the Financial Viability of using pay on demand compute power vs. traditional in-house data processing



A Stereoscopic Census of our Galaxy

(based on slides from Jos de Bruijne and William O'Mullane)



10/29/2009

© The Server Labs S.L. 2009





- Primary goal of the Gaia mission is to create an astrometric catalogue of 1 billion stars (approx 1% of our Galaxy) with micro arc second precision.
- Gaia satellite to be launched in 2011.
- Observations done until 2017.
- Catalogue ready around 2019.







10/29/2009

© The Server Labs S.L. 2009



Satellite – Obligatory Reminder



☐ Mission:

- Stereoscopic Census of Galaxy
- µarcsec Astrometry G<20 (10^{^9} sources)
- Radial Velocities G<16</p>
- Photometry G < 20</p>
- Discover structure and unravel formation history of Galaxy.
- Status:
 - ESA Corner Stone 6
 - ESA provide the hardware and launch
 - Launch: Spring 2012.
 - Satellite In development
 - EADS/Astrium







Using Cebreros, Spain (35M)

- > 3-8Mb/s downlink
 - * depends on encoding
 - * which depends on weather !
- > ~ 30GB/day -> ~100TB
- Occasionally New Norcia, Australia
 - > during Galactic plane scans



- > data accumulated onboard downlinked later
- Data is compressed encoded and requires a lot of processing (~10^{^21} FLOP)





If it took 1 millisecond to process one image, the processing time for just one pass through the data (on a single processor) would take 30 years.

Obviously the adopted solution is much faster distributed/parallel processing.



10/29/2009







Block-iterative least-squares solution of the over-determined system of equations

O = G + S + A + C + n









Datatrains drive through AGIS Database passing observations to algorithms.
There can be as many **Datatrains** in parallel as we wish



10/29/2009

© The Server Labs S.L. 2009











- Very simple ..
 - Keep all machines busy all the time!
 - ★ Busy = CPU ~90%
 - Post jobs on whiteboard

Trains/Workers Mark Jobs – and do them Mark finished – repeat until done



Previous attempt had much more general scheduling It was also ~1000 times slower.





Data Train load





AGIS runs like this High load, low network, high CPU

10/29/2009



Whiteboard/Data train concept







Whiteboard/Data train concept







Whiteboard/Data train concept



DataTrain. A DataTrain represents a worker. Normally a DataTrain consists of a Reader that reads data from a store (database or file) and one or more Takers. The Takers are interface that are implemented by the algorithms. Takers can also be used to write computed data back to a store. The DataTrain uses the Reader to read data in blocks and then passes these data blocks to the takers.







Taker Architecture









Data centre cost

AGIS run times decrease as more processors are added. Note that the data volume increased from 2005 to 2006 from 18 months to 5 years, the processor power also increased but the run time went up. This was dramatically improved in 2007. The normalised column shows throughput per processor in the system (total observations/processors/hours) e.g. an indication of the real performance.

Date	Dataset	Procs	Time	Normalised	
2005	18 mnt 10^6 src	12	3h	$0.9 * 10^6$ obs/hour	
2006	60 mnt 10^6 src	36	5h	$0.5 * 10^6 \text{obs/hour}$	
2007	60 mnt 10^6 src	24	3h	$1.3 * 10^6$ obs/hour	
2008	60 mnt 10^6 src	31	1.5h	$2.1 * 10^{6} \text{ obs/hour}$	
Feb 2008	60 mnt 10^6 src	31	1h	$3.2 * 10^{6} \text{ obs/hour}$	

n







- Iterative processing 6 month Data Reduction Cycles
- At current estimates AGIS will run 2 weeks every 6 months
- Amount of data increases over the 5 year mission





The Study: Running AGIS in Amazon EC2



- Technical Feasibility:
 - Can AGIS run in the cloud?
 - > What are the restrictions?
 - > What modifications do we have to make?
- Financial Viability
 - What would be the cost of using EC2 for AGIS?
 - Can we do a hybrid solution using a local data centre followed by a mix of local/EC2?





	Small	Large	Extra Large	High CPU Medium	High CPU Large	
Bits	32	64	64	32	64	
RAM	1.7 GB	7.5 GB	15 GB	1.7 GB	7 GB	
Disk	160 GB	850 GB	1690 GB	350 GB	1690 GB	
EC2 Compute Units	1	4	8	5	20	
I/O Performance	Medium	High	High	High	High	
Firewall	Yes	Yes	Yes	Yes	Yes	





• 64 bit images

Large, Extra Large and High CPU Large

- Oracle ASM Image based on Oracle Database 11g Release 1 Enterprise Edition - 64 Bit (Large instance) -<u>ami-7ecb2f17</u>
- AGIS Self configuring Image based on Ubuntu 8.04 LTS Hardy Server 64-Bit (Large, Extra Large and High CPU Large Instances) - <u>ami-e257b08b</u>



Architecture in the Cloud





© The Server Labs S.L. 2009





- EC2 Large instance (m1.large)
- Oracle Enterprise Edition 11g 64 bit (11.0.6)
- Oracle ASM
- Elastic Block Storage
 - > 5 x 100GB disks /dev/sdg /dev/sdk







- Launch an m1.large instance of ami-7ecb2f17
- Attach the /mnt partition properly so it has enough space
- Create 5 EBS vols of 100GB each and attach them to the instance
- Set up Oracle ASMLib
 - Install drivers
 - Run oracleasm_debug_link
 - Run oracleasm configure, createdisk
- Copy a pre-recorded Oracle response file up to create an ASM instance
- Run Oracle installer to create the ASM instance
- Copy a pre-recorded Oracle response file up to create the AGIS DB instance
- Run Oracle installer to create the AGIS DB instance





- Create an Elastic IP and associate it with the instance
- Change the hostname to be the new public DNS name
 - hostname ec2-174-129-223-59.compute-1.amazonaws.com
- Run localconfig remove followed by localconfig add
- This will run for ever unless you edit /etc/inittab and change the following line

h1:35:respawn:/etc/init.d/init.cssd run >/dev/null 2>&1 </dev/null

to

h1:345:respawn:/etc/init.d/init.cssd run >/dev/null 2>&1 </dev/null

- Start the ASM instance
- Start the AGIS DB instance

Don't forget to make a new image!!!





- Instances (m1.large and c1.xlarge).
- Java version 1.6.0_13
 - Java(TM) SE Runtime Environment (build 1.6.0_13-b03)
 - Java HotSpot(TM) 64-Bit Server VM (build 11.3-b02, mixed mode)
- Apache Tomcat 6.0.14
- AGIS software.
- Creation of agis user account
- rc.local script modified to run the AGIS process
 - Self configuring using user-data





- Launch an m1.large instance of ami-e257b08b
- Checkout the source code from the svn server.
- Create an agis user to run the process.
- Set up /etc/rc.local to execute the runAgis.sh
- Create a generic runAgis.sh script that reads data passed to the ami during boot time (using the AWS).
 - Data contains JVM parameters and especific application parameters (Depending on the DataTrain that will be executed at that node).





- Launch the RunManager
 - User data is passed to the AMI at boot time.
 - Process type (RunManager, DataTrain, Server)
 - JVM args
 - User Data args

```
export RUNMANAGER_DATA="\"RunManager|-Xmx6g\${IFS}-
XX:MaxPermSize=512M\${IFS}-Xnoclassgc\${IFS}-
Dgaia.cu3.agis.algo.gis.attitude.AttitudeUpdateServer.numServers=${NODES}\${IFS}
-D\${HOSTNAME}=gaia.cu3.agis.progs.run.RunManager\""
export NODES=1
ec2-run-instances ami-9ea442f7 -n 1 -k gaia-keypair -g default -z us-east-1c -t m1.large -d
${RUNMANAGER_DATA} > instance_id.txt 2>&1
```

ec2-associate-address \$IP -i \$INSTANCE_ID





- Script executed at boot time
 - Get the data from an Amazon WebService
 - Run the Launcher with the JVM args

export USER_DATA=`GET <u>http://169.254.169.254/2007-03-01/user-data</u>` export NODE_TYPE=`echo \${USER_DATA} | awk -F\| '{print(\$1)}'` export VMARGS=`echo \${USER_DATA} | awk -F\| '{print(\$2)}'`

nohup java \${VMARGS} -cp \$AGIS_HOME/conf:\$AGIS_HOME/dist/GaiaAGIS.jar Dgaia.cu1.tools.hostid=1 -Dgaia.cu3.agis.mgr.CycleDesc=\$1
gaia.cu3.agis.progs.run.Launcher >> \${LOG_FILE} 2>&1 &





Elastic Ips for Oracle and the Whiteboard

nstances Images KeyPairs Security Groups ^{Elastic IPs} Volumes and Snapshots Bundle Tasks Availability Zones

/ Your Elastic IPs			
📀 😳 📦 🥥	😮 🕡		
Address	⊽ Instance ID	Tag	
174.129.209.236	i-7b82f712	Run Manager IP	
174.129.223.59	i-6f096106	Oracle IP	

- Monitoring
 - Plan to use Hyperic or Nagios
 - Amazon CloudWatch now available
- Security
 - Be very careful with keys and passwords





- Creating new images takes a long time so make sure you get it right ⁽ⁱ⁾
- Oracle is very fussy about it's IP address, hence the Elastic IP
 - > Oracle instance hostname changed to be the public DNS name
- Some work still needed on the startup script to make sure the ASM boots first time
 - Fixed with new Oracle AMIs
- The Attitude Servers took a long time to start up (20 mins)
 - This was due to a race condition caused by spin locks in the type of java Thread Pool we were using.





 Took a team of 2 less than 20 man days to get running (Parsons,Olias).





 To avoid the Spin Locks in the Thread Pool we had to change these lines of code:

<pre>this.stack = new ConcurrentStack<runnable>();</runnable></pre>	 Non blocking
<pre>this.threadPoolExecutor = new ThreadPoolExecutor(numberOfWorkingThreads, numberOfWorkingThreads, 100, TimeUnit.SECONDS, this.stack);</pre>	stack - CAS
this .threadPoolExecutor.prestartAllCoreThreads();	

To this



EC2 Instances



Instances Images KeyPairs Security Groups Elastic IPs Volumes and Snapshots Bundle Tasks Availability Zones

Your Instances

	🥥 🚳) 🥥 🕥 🥥						
Owner	State	Public DNS	Private DNS	Idx	Туре	Local Launch Time 🗸	Tag	₽.
	running	ec2-174-129-223-59.compute-1	amazon domU-12-31-39-03-69-71.compute-	1.internal 0	m1.large	2009-04-21 10:56:20	Oracle	
	running	ec2-174-129-209-236.compute-	1.amazo domU-12-31-39-00-29-51.compute-	1.internal 0	m1.large	2009-05-07 14:27:58	RunMan	ager
	running	ec2-75-101-208-129.compute-1	amazon domU-12-31-39-00-F0-B1.compute-	1.internal 0	c1.xlarge	2009-05-07 14:35:56		
	running	ec2-174-129-187-101.compute-	1.amazo domU-12-31-39-01-85-B1.compute-	1.internal 1	c1.xlarge	2009-05-07 14:35:56		
	running	ec2-75-101-248-164.compute-1	amazon domU-12-31-39-00-90-A1.compute-	1.internal 2	c1.xlarge	2009-05-07 14:35:56		
	running	ec2-174-129-175-175.compute-	1.amazo domU-12-31-39-00-F4-81.compute-	1.internal 0	c1.xlarge	2009-05-07 14:40:38		
	running	ec2-67-202-33-107.compute-1.a	mazona domU-12-31-39-00-F5-D1.compute-	1.internal 1	c1.xlarge	2009-05-07 14:40:38		
	running	ec2-67-202-42-56.compute-1.ar	nazonaw domU-12-31-39-01-70-E1.compute-	1.internal 2	c1.xlarge	2009-05-07 14:40:38		
	running	ec2-174-129-184-8.compute-1.a	mazona domU-12-31-39-01-70-A1.compute-	1.internal 3	c1.xlarge	2009-05-07 14:40:38		
	running	ec2-67-202-11-61.compute-1.ar	nazonaw domU-12-31-39-00-BD-B1.compute	-1.internal 4	c1.xlarge	2009-05-07 14:40:38		
	running	ec2-75-101-206-68.compute-1.a	mazona domU-12-31-39-01-B9-01.compute-	1.internal 5	c1.xlarge	2009-05-07 14:40:38		
	running	ec2-75-101-204-217.compute-1	amazon domU-12-31-39-00-D4-71.compute-	1.internal 6	c1.xlarge	2009-05-07 14:40:39		
	running	ec2-67-202-2-184.compute-1.ar	nazonaw domU-12-31-39-00-B8-01.compute-	1.internal 7	c1.xlarge	2009-05-07 14:40:39		
	running	ec2-75-101-176-225.compute-1	amazon domU-12-31-39-00-F9-61.compute-	1.internal 8	c1.xlarge	2009-05-07 14:40:39		
	running	ec2-174-129-142-128.compute-	1.amazo domU-12-31-39-00-8C-B1.compute-	1.internal 9	c1.xlarge	2009-05-07 14:40:39		
	running	ec2-174-129-98-165.compute-1	amazon domU-12-31-39-00-98-61.compute-	1.internal 10	c1.xlarge	2009-05-07 14:40:39		
	running	ec2-67-202-10-82.compute-1.ar	nazonaw domU-12-31-39-00-60-11.compute-	1.internal 11	c1.xlarge	2009-05-07 14:40:39		
	running	ec2-72-44-60-224.compute-1.ar	nazonaw domU-12-31-39-00-71-21.compute-	1.internal 12	c1.xlarge	2009-05-07 14:40:39		
	running	ec2-174-129-184-53.compute-1	amazon domU-12-31-39-01-91-91.compute-	1.internal 13	c1.xlarge	2009-05-07 14:40:39		
	running	ec2-174-129-188-181.compute-	1.amazo domU-12-31-39-01-88-01.compute-	1.internal 14	c1.xlarge	2009-05-07 14:40:39		



AGIS Works!









I/O Breakdown 💿 I/O Function 🔘 I/O Type 🔘 Consumer Group











• Let's see it in action!





- AGIS Primary star update
 - > 2 million stars 5 years observations
- 1x Oracle Instance (m1.large)
- 1x AGIS RunManager/Monitor Instance (m1.large)
- 3x AGIS AttitudeServer Instances (m1.large)
- 15x AGIS DataTrain Instances (c1.xlarge)
- Results
 - Iteration time 100 mins
 - Performance slightly slower than in-house cluster





- AGIS Primary star update
 - > 2 million stars 5 years observations
- 1x Oracle Instance (m1.large)
- 1x AGIS RunManager/Monitor Instance (m1.large)
- 4x AGIS AttitudeServer Instances (c1.xlarge)
- 45x AGIS DataTrain Instances (c1.xlarge)
- Results
 - Iteration time 50 mins
 - Performance comparable to in-house cluster





- AGIS Secondary star update
 - > 20 million stars 5 years observations
- 1x Oracle Instance (m1.large)
- 1x AGIS RunManager/Monitor Instance (m1.large)
- 30x AGIS DataTrain Instances (c1.xlarge)
- Results
 - Run Time 43 minutes





- AGIS Secondary star update
 - > 20 million stars 5 years observations
- 1x Oracle Instance (m1.large)
- 1x AGIS RunManager/Monitor Instance (m1.large)
- 48x AGIS DataTrain Instances (c1.xlarge)
- Results
 - Run Time 48 minutes (Slightly Slower)
 - Oracle Row lock contention problems

Test run 4 continued! Loading.



Done

ec2-174-129-223-59.compute-1.amazonaws.com:1158 🔒 🗞 Profile1 🙎 S3 Fox

96916509

ord11g.us.oracle.com

ord11g.us.oracle.com

3QL Details: crdxzbwq7qm5k

SELECT *

Statistics.

Summary

Active Sessions 23

46

9:34

Activity (%)

Activity

9:40

Start Time 15-May-2009 10:29:41

15-May-2009

9:45

2.17

2.17

2.17

2.17

2.17

Electricy for ayricle anous alligic block reads

<u>34</u>

72

<u>Plan</u>

 G_0

Switch to SQL ID

Text

Details

JDBC Thin Client

AGISWB5_0

AGISWB5_0





- 2 million stars 5 years data
 - > 24 iterations of 100 minutes = 40 hrs x 20 EC2 instances
 - Secondary update 30mins * 10 = 5hrs x 20 EC2 instances
- For the Full 1 billion data set we will have 100 million primary stars plus 6 years of data
 - > 40 x 50 x 6/5 = 2,400 hrs x 20 EC2 inst.
 - Sec = 300 hrs x 20 EC2 inst.
- Average over the whole mission will be

> 2,700 x 12 / 2 = 16,200 hrs x 20 EC2 inst.





• For the 2 million star run the cost will be

- > 45 x 15 x 0.8\$ (c1.xlarge) = 540\$
- > 45 x 5 x 0.4\$ (m1.large) = 90\$
- > Data transfer in 50GB x .1 = 5\$
- > EBS 471GB Storage x .1\$ = 47\$
- ▷ EBS I/O = 9\$
- Elastic IP = 2.29\$
- > S3 Storage = 5.91\$
- Total 699.2\$ = 525 EUR (at current exchange rates)





- Extrapolating to 1 billion stars 6 year mission
 - > 2700 x 6 x 15 x 0.8\$ (c1.xlarge) = 194,400\$
 - > 2700 x 6 x 5 x 0.4\$ (m1.large) = 32,400\$
 - Data transfer in 500 x 50GB x .1\$ = 2,500\$
 - EBS 500 x 250GB Storage x 6 x .1\$ = 75,000\$
 - > EBS I/O = 500 x 9\$ = 4500\$
 - Elastic IP = 48180 x 0.01\$ = 481\$
 - S3 Storage = 5.91\$
- Total 309,286.\$ = 229,066 EUR
- Allowing for mistakes (wrong configs etc.)
- Estimated Total 343,599 EUR at current exchange rates





- AGIS runs intermittently with growing Data volume.
- Estimate 2015 ~1.1MEuro (machine) + 1Meuro (energy bill less
 ?) = ~2Meuro
 - In fact staggered spending for machines
 - buy machines as data volume increase
- Estimate on Amazon at today prices with 10 intermittent runs ~400Keuro
 - Possibility to use more nodes and finish faster !
- Reckon you still need in house machine to avoid wasting time testing on E2C
- Old nut, Vendor lock-in ? Need standards





- AGIS can be run in the cloud!
- Running with 48 Data train nodes gave us new scalability problems we hadn't seen before!
- The Economics work out:
 - To do the same amount of processing in the same time, EC2 works out cheaper for AGIS!
 - With the knowledge that EC2 is an option we can delay buying more machines until the middle of the mission (2014) and decide then.





- Currently running a second feasibility study
 - New data set, 60 million primary stars (1/3 final primaries). Approx 1.5 TB
 - Idea is to see how scalable the Gaia DataTrain grid is. Try and scale to 1000 8CPU EC2 instances
 - Compare S3 performance with Oracle
 - Using Rightscale technology



Thank you for listening





Thank you for listening

http://blog.theserverlabs.com http://www.theserverlabs.com

mailto:pparsons@theserverlabs.com

10/29/2009

© The Server Labs S.L. 2009



ESA Copyright Notice



This presentation contains images and videos which have been released publicly from ESA. You may use ESA images or videos for educational or informational purposes. The publicly released ESA images may be reproduced without fee, on the following conditions:

* Credit ESA as the source of the images:

Examples: Photo: ESA; Photo: ESA/Cluster; Image: ESA/NASA - SOHO/LASCO

* ESA images may not be used to state or imply the endorsement by ESA or any ESA employee of a commercial product, process or service, or used in any other manner that might mislead.

* If an image includes an identifiable person, using that image for commercial purposes may infringe that person's right of privacy, and separate permission should be obtained from the individual.

If these images are to be used in advertising or any commercial promotion, layout and copy must be submitted to ESA beforehand for approval to:

ESA Multimedia multimedia@esa.int

Some images contained in this presentation have come from other sources, and this is indicated in the Copyright notice. For reuse of non-ESA images contact the designated authority.

Use of ESA videos

The use of ESA video images in streaming and downloadable format is limited to direct viewing and/or file storage on a single computer per stream and/or download. Forwarding of files or streams to other computers, or use on any non-ESA Web is prohibited. For the authorisation of any such use, please contact:

ESA Multimedia multimedia@esa.int