Basic Course on Bioinformatics tools for Next Generation Sequencing data mining

IT infrastructure and user interface: The Galaxy architecture and ARIES cluster

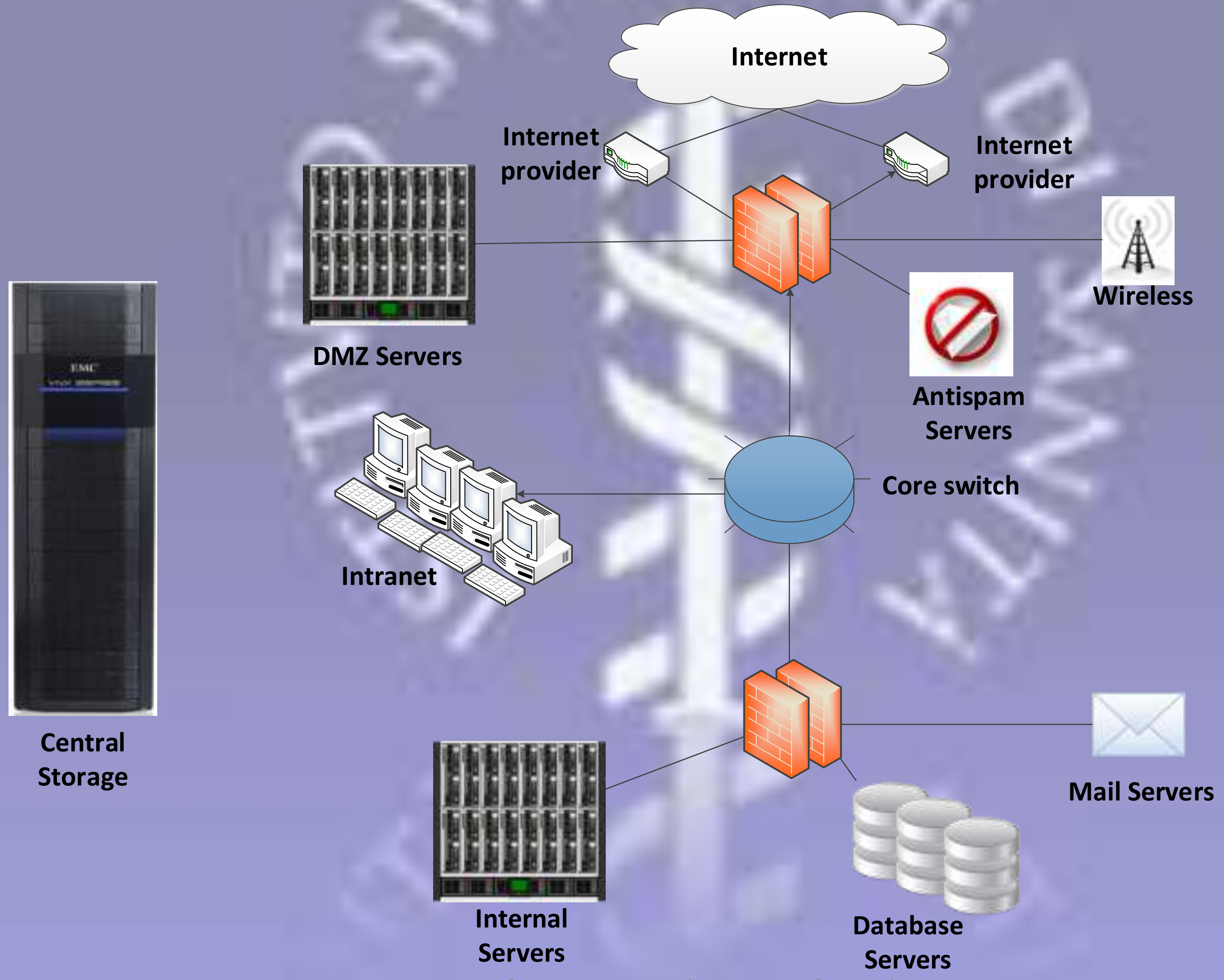
Arnold Knijn

IT Sector - ISS

Istituto Superiore di Sanità

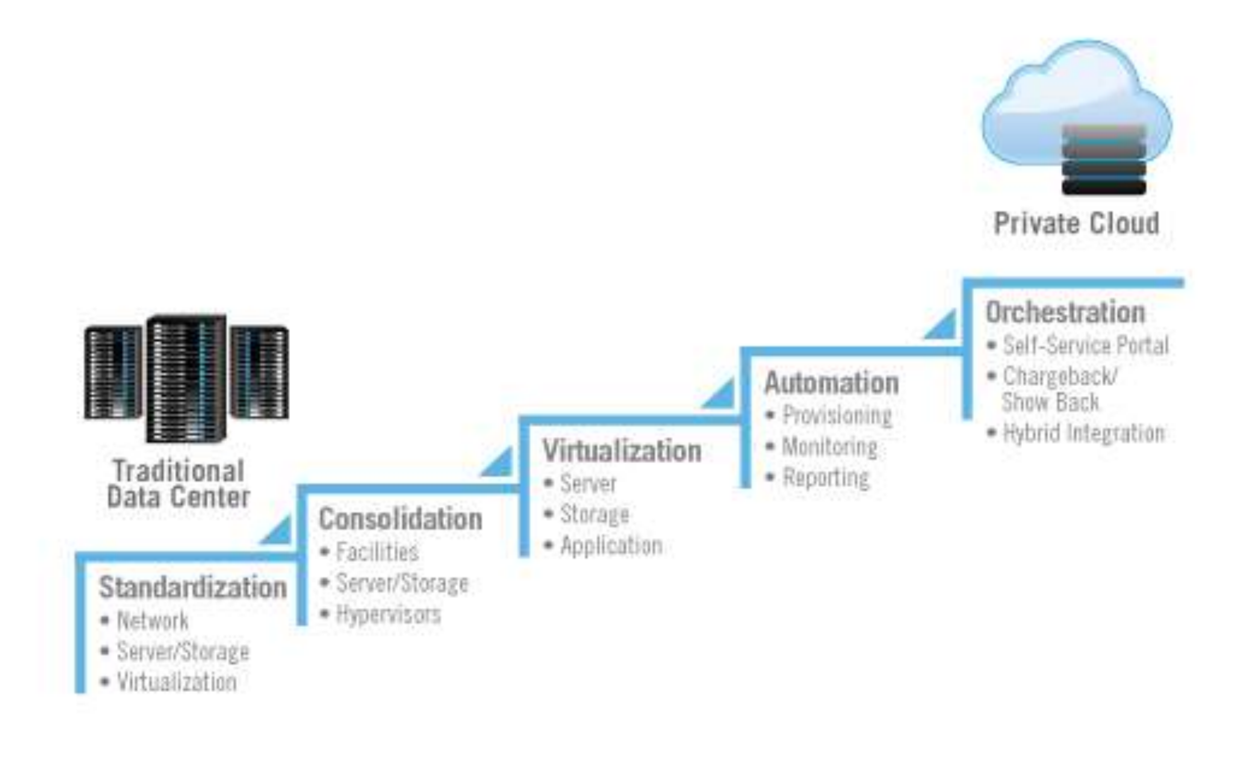
- Personel: ~ 2400
- Wired end-points: ~ 3000
- LAN: 1
- Buildings: 37
- Core switches: 2
- Distribution switches: 87
- Appliances: ~ 20 (firewalls, IPS, etc.)
- 2 Blade systems: 16 hosts (16 logical CPUs, 32/36 GB RAM)
- Servers/virtual machines: > 130 (60% Windows, 40% Linux)
- Databases: > 100 (~ 450 GB)
- Mailboxes: > 3500 (> 6 TB)
- Central Storage: > 50 TB high-level, > 75 TB low-level

ISS infrastructure



Data Center Paradigm Evolution

- Mainframe
- One-application, one-server model



Standardisation/consolidation

Simplify, through reduction of system number and types.

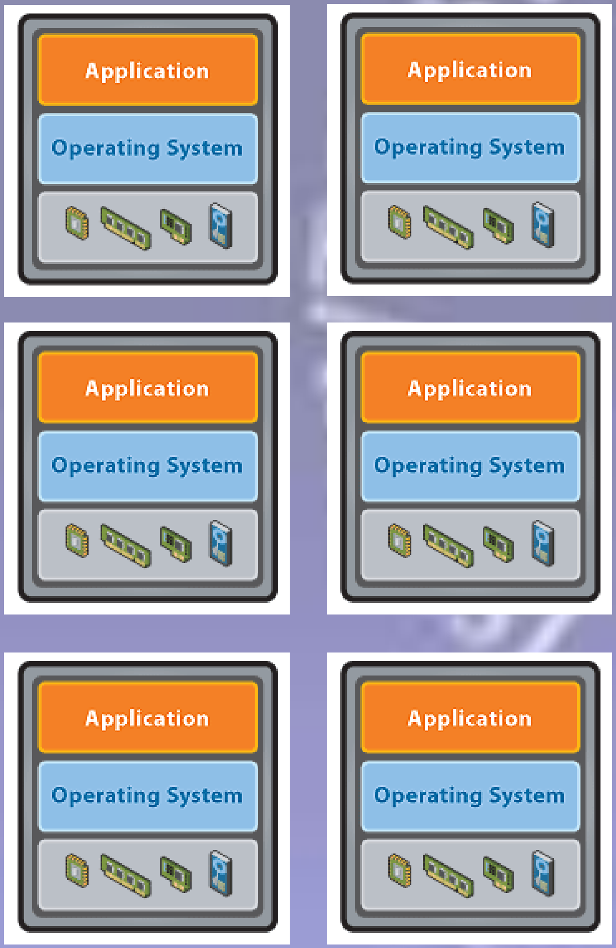
Simplification

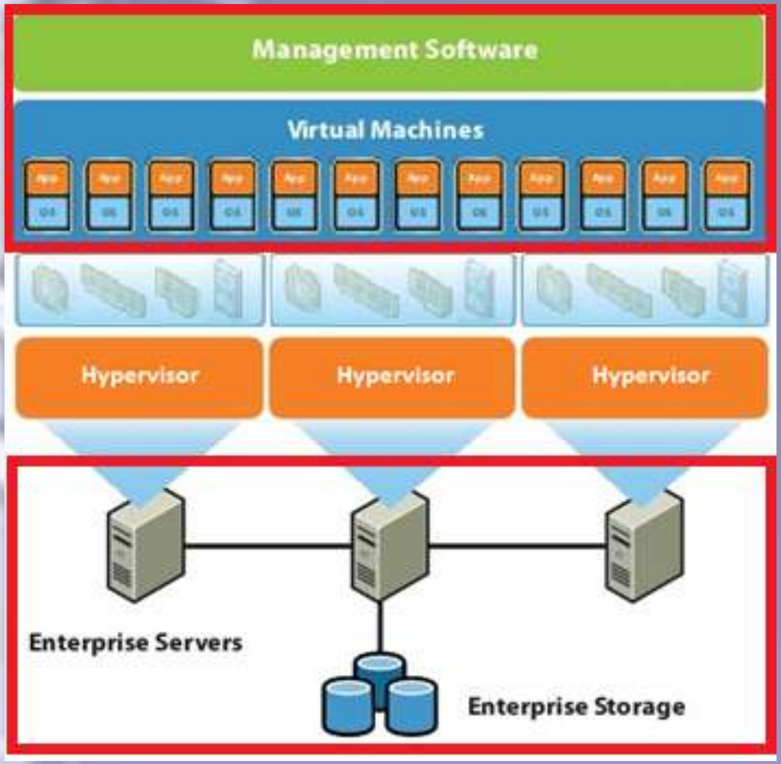
Storage

CPU/RAM

Networking

Server virtualisation



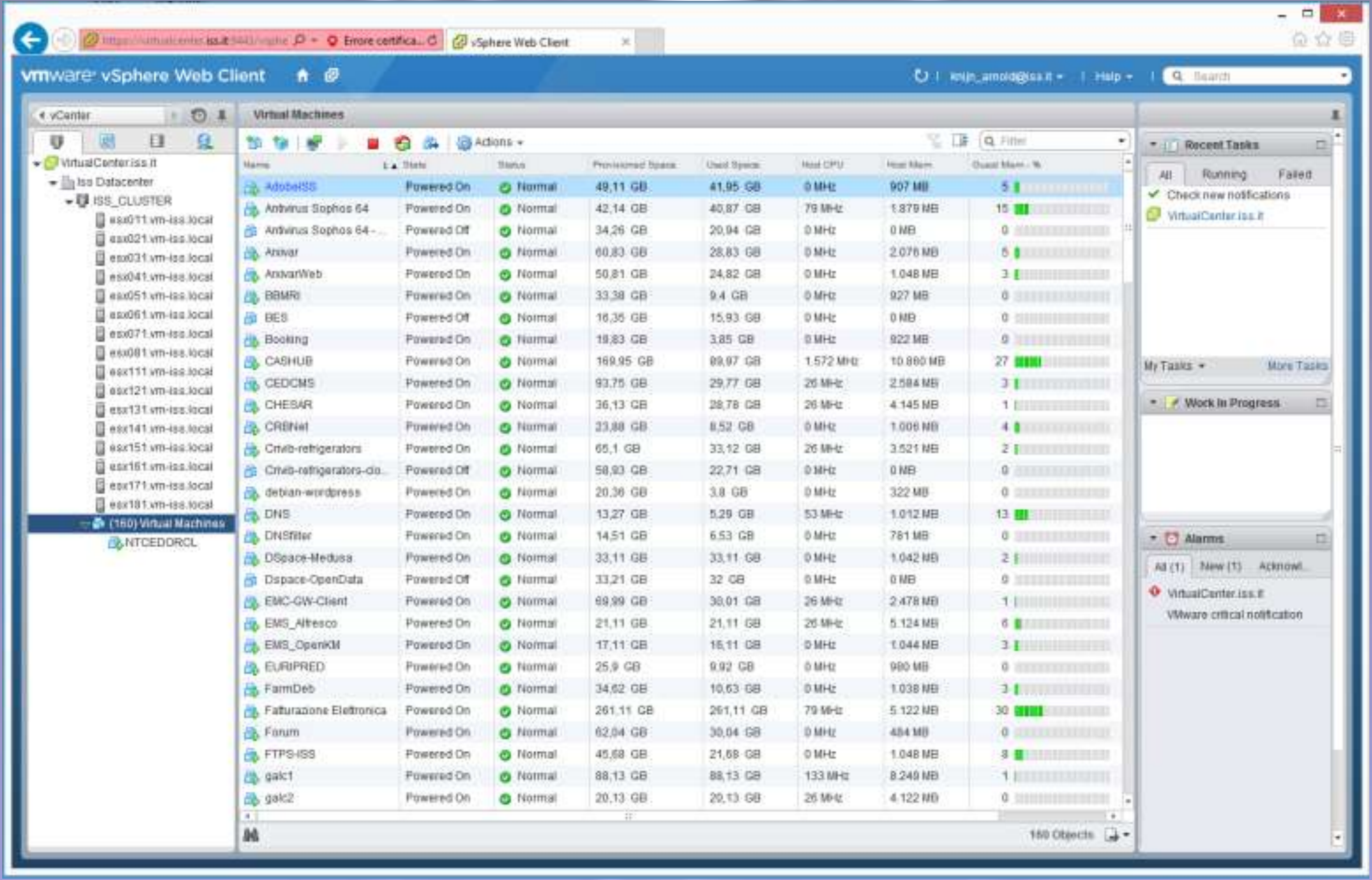


Virtualisation characteristics

- ✓ Physical server reduction (1:10 – 1:4)
- ✓ Decoupling from hardware
- ✓ Simpler management (installation, backup)
- ✓ Resource optimisation
- ✓ Adaptability/ Scalability
- ✓ Availability
- ✓ Test environment

- Some hypervisor overhead
- Non-virtualisable hardware (server non x86, etc.)
- More restrictive hardware requirements
- The infrastructure has to be solid
- Virtual machine proliferation

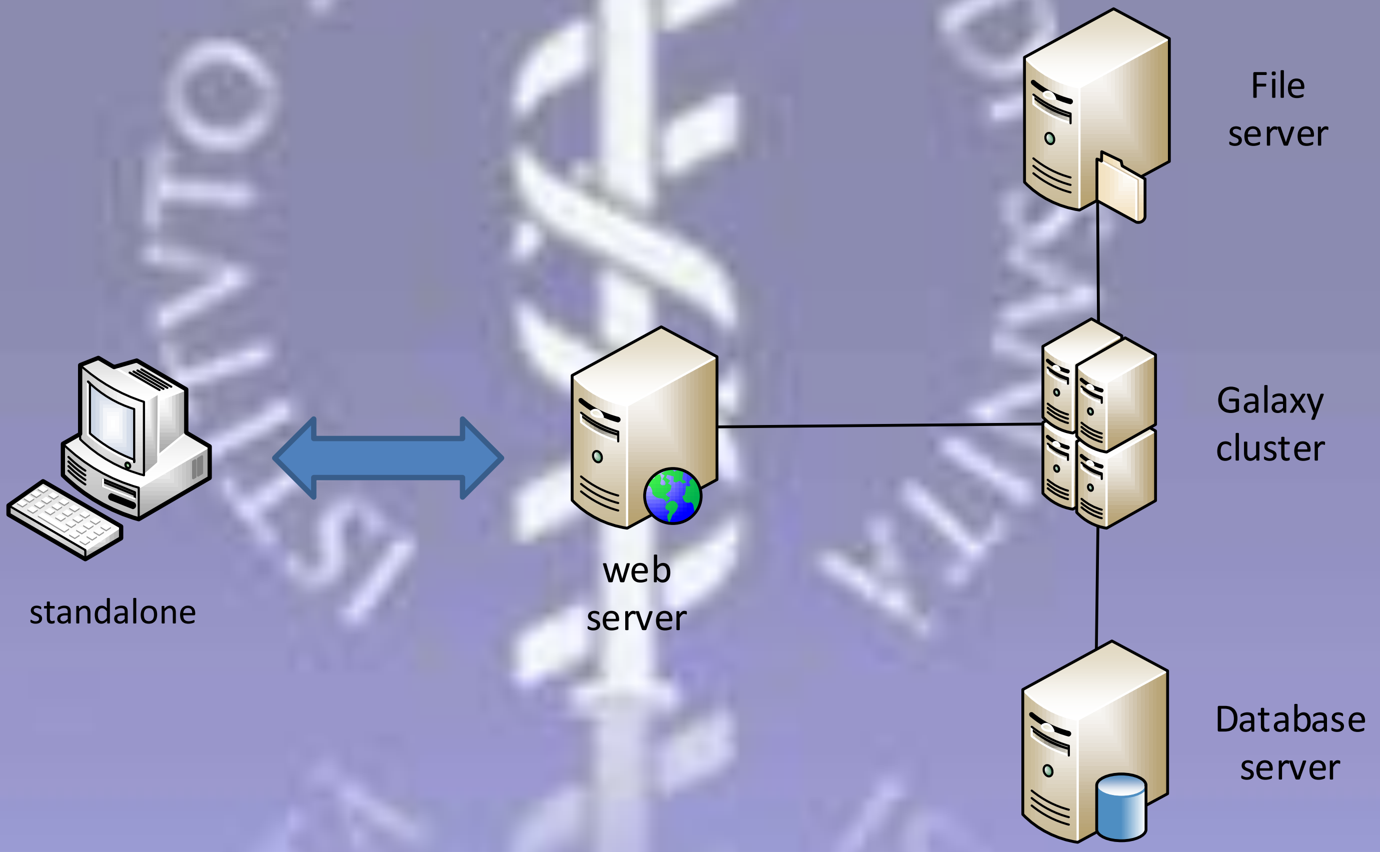
Centralised management



Name	State	Status	Provisioned Space	Used Space	Host CPU	Host Mem	Guest Mem - %
AdobeISS	Powered On	Normal	49,11 GB	41,95 GB	0 MHz	907 MB	5
Antivirus Sophos 64	Powered On	Normal	42,14 GB	40,87 GB	79 MHz	1.879 MB	15
Antivirus Sophos 64 - ...	Powered Off	Normal	34,26 GB	20,94 GB	0 MHz	0 MB	0
Anivar	Powered On	Normal	60,83 GB	28,83 GB	0 MHz	2.076 MB	5
AnivarWeb	Powered On	Normal	50,81 GB	24,82 GB	0 MHz	1.048 MB	3
BBMRI	Powered On	Normal	33,38 GB	9,4 GB	0 MHz	927 MB	0
BES	Powered Off	Normal	16,35 GB	15,93 GB	0 MHz	0 MB	0
Booking	Powered On	Normal	19,83 GB	3,85 GB	0 MHz	922 MB	0
CASHUB	Powered On	Normal	169,95 GB	89,97 GB	1.572 MHz	10.880 MB	27
CEDCMS	Powered On	Normal	93,75 GB	29,77 GB	26 MHz	2.584 MB	3
CHESAR	Powered On	Normal	36,13 GB	28,78 GB	26 MHz	4.145 MB	1
CRBNet	Powered On	Normal	23,88 GB	8,52 GB	0 MHz	1.006 MB	4
Crivib-refrigerators	Powered On	Normal	65,1 GB	33,12 GB	26 MHz	3.521 MB	2
Crivib-refrigerators-clo...	Powered Off	Normal	58,93 GB	22,71 GB	0 MHz	0 MB	0
debian-wordpress	Powered On	Normal	20,36 GB	3,8 GB	0 MHz	322 MB	0
DNS	Powered On	Normal	13,27 GB	5,29 GB	53 MHz	1.012 MB	13
DNSfilter	Powered On	Normal	14,51 GB	6,53 GB	0 MHz	781 MB	0
DSpace-Medusa	Powered On	Normal	33,11 GB	33,11 GB	0 MHz	1.042 MB	2
Dspace-OpenData	Powered Off	Normal	33,21 GB	32 GB	0 MHz	0 MB	0
EMC-GW-Client	Powered On	Normal	69,99 GB	30,01 GB	26 MHz	2.478 MB	1
EMS_Alfresco	Powered On	Normal	21,11 GB	21,11 GB	26 MHz	5.124 MB	6
EMS_OpenKM	Powered On	Normal	17,11 GB	16,11 GB	0 MHz	1.044 MB	3
EURIPRED	Powered On	Normal	25,9 GB	9,92 GB	0 MHz	980 MB	0
FarmDeb	Powered On	Normal	34,62 GB	10,63 GB	0 MHz	1.038 MB	3
Fatturazione Elettronica	Powered On	Normal	261,11 GB	261,11 GB	79 MHz	5.122 MB	30
Forum	Powered On	Normal	62,04 GB	30,04 GB	0 MHz	484 MB	0
FTPS-ISS	Powered On	Normal	45,68 GB	21,68 GB	0 MHz	1.048 MB	8
galc1	Powered On	Normal	88,13 GB	88,13 GB	133 MHz	8.249 MB	1
galc2	Powered On	Normal	20,13 GB	20,13 GB	26 MHz	4.122 MB	0

160 Objects

Galaxy architecture



Differences

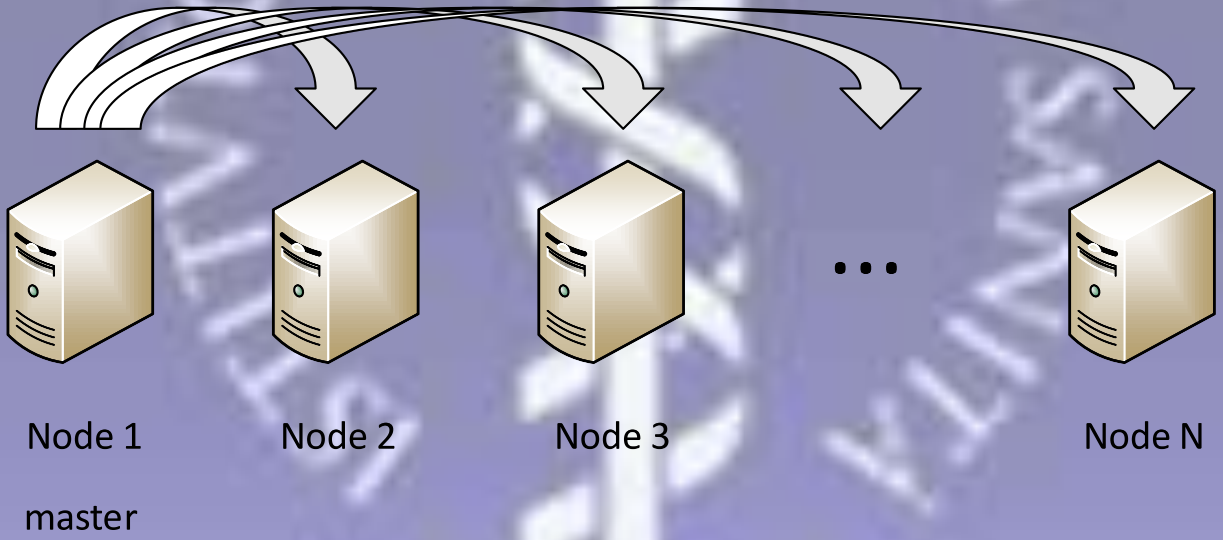
By default, Galaxy uses:

- SQLite
- Built-in HTTP server for all tasks
- Local job runner
- Single process
- Simplest error-proof configuration

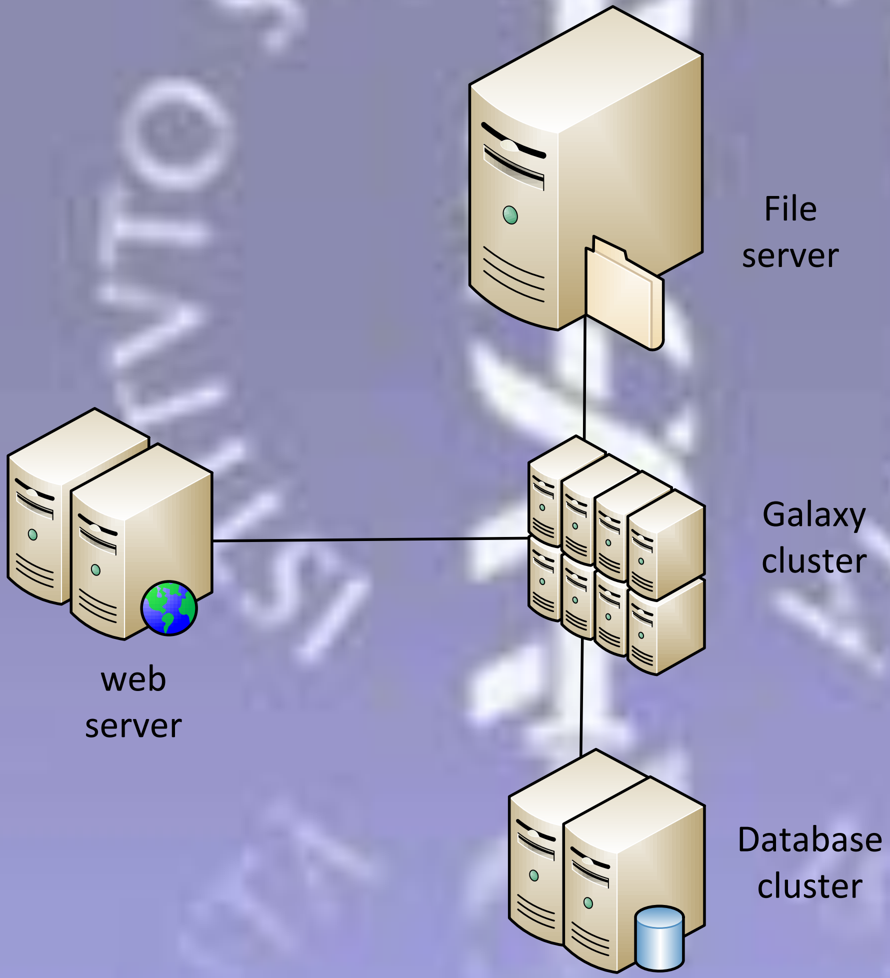
In production:

- Real database
- Real HTTP server for many tasks
- Cluster job runner
- Multi process
- More complex configuration

Galaxy cluster



Architecture scalability



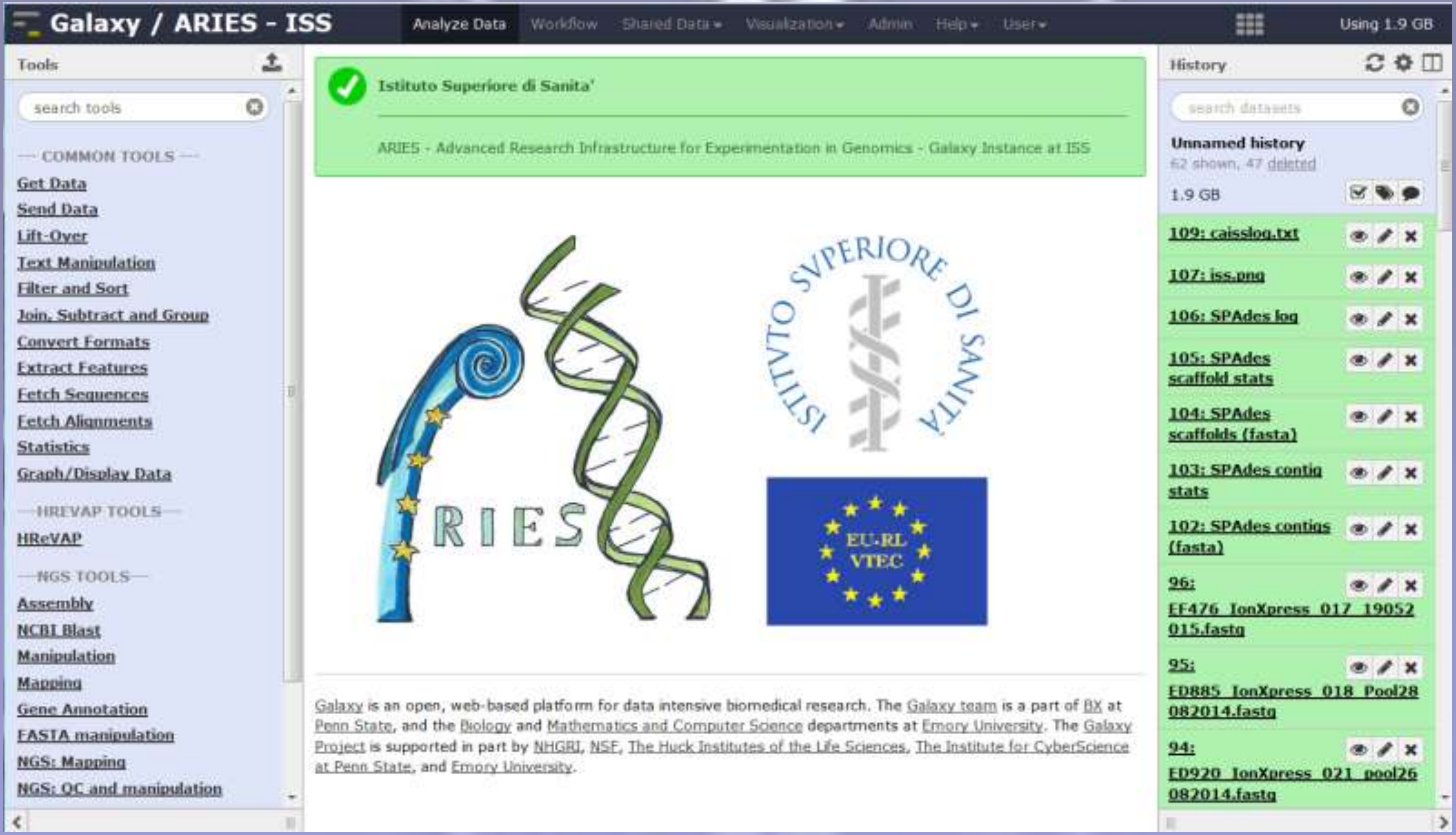
File server

Galaxy cluster

web server

Database cluster

Galaxy user interface

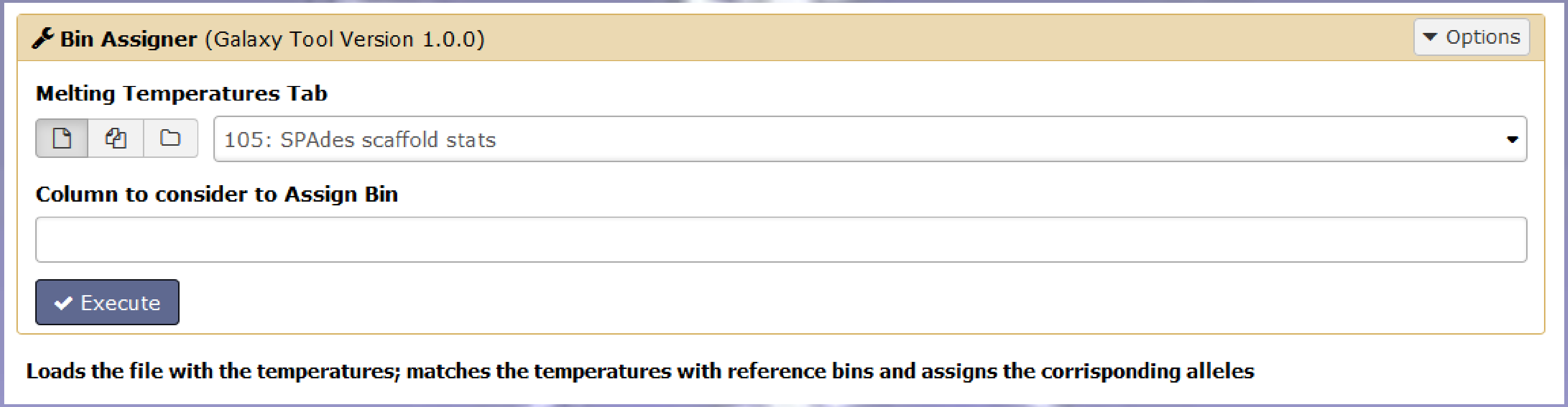


Istituto Superiore di Sanita'

ARIES - Advanced Research Infrastructure for Experimentation in Genomics - Galaxy Instance at ISS

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Intuitive and self-documenting



Bin Assigner (Galaxy Tool Version 1.0.0)

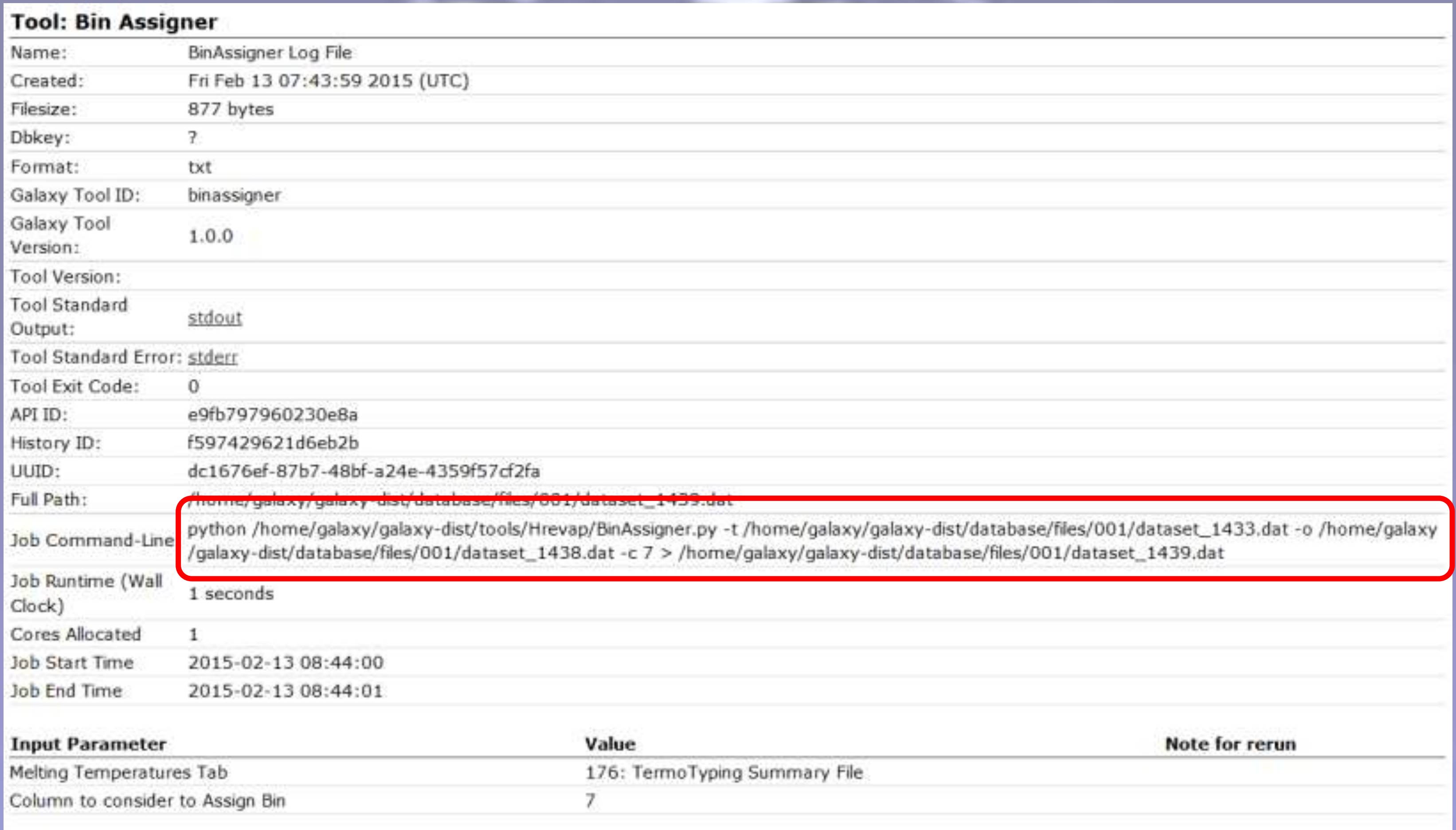
Melting Temperatures Tab

105: SPAdes scaffold stats

Column to consider to Assign Bin

Loads the file with the temperatures; matches the temperatures with reference bins and assigns the corrisponding alleles

UI vs Command-Line



Tool: Bin Assigner

Name:	BinAssigner Log File
Created:	Fri Feb 13 07:43:59 2015 (UTC)
Filesize:	877 bytes
Dbkey:	?
Format:	txt
Galaxy Tool ID:	binassigner
Galaxy Tool Version:	1.0.0
Tool Version:	
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0
API ID:	e9fb797960230e8a
History ID:	f597429621d6eb2b
UUID:	dc1676ef-87b7-48bf-a24e-4359f57cf2fa
Full Path:	/home/galaxy/galaxy-dist/database/files/001/dataset_1439.dat
Job Command-Line	python /home/galaxy/galaxy-dist/tools/Hrevap/BinAssigner.py -t /home/galaxy/galaxy-dist/database/files/001/dataset_1433.dat -o /home/galaxy/galaxy-dist/database/files/001/dataset_1438.dat -c 7 > /home/galaxy/galaxy-dist/database/files/001/dataset_1439.dat
Job Runtime (Wall Clock)	1 seconds
Cores Allocated	1
Job Start Time	2015-02-13 08:44:00
Job End Time	2015-02-13 08:44:01

Input Parameter	Value	Note for rerun
Melting Temperatures Tab	176: TermoTyping Summary File	
Column to consider to Assign Bin	7	

Home –made tools

```
<tool id="binassigner" name="Bin Assigner">
    <description>Bin Assigner tool</description>
    <command interpreter="python">
        BinAssigner.py -t $tmstab -o $output -c $columntab > $logfile
    </command>
    <inputs>
         <param name="tmstab" type="data" format="tabular" label="Melting Temperatures Tab"/>
         <param name="columntab" type="text" format="integer" label="Column to consider" />
     </inputs>
     <outputs>
         <data format="tabular" name="output" label="Allele Table"/>
         <data format="txt" name="logfile" label="BinAssigner Log File" />
     </outputs>
     <help>
          **Loads the file with the temperatures; matches the temperatures with reference bins
             and assigns the corrisponding alleles**
     </help>
</tool>
```