GenAP: A platform to provide Biomedical tools throughout Canadian HPCS

David Anderson de Lima Morais[1], Michel Barrette[1], David Bujold[2], Carol Gauthier[1], Kuang Chung Chen[2], Simon Nderitu[2], Maxime Levesque[1], Bryan Caron[2], Alain Veilleux[1], Pierre-Etienne Jacques[1], Guillaume Bourque[2]

1 Centre de Calcul Scientifique, Université de Sherbrooke, Quebec, Canada
2 McGill University and Genome Quebec Innovation Center, Montreal, Canada

OVERVIEW

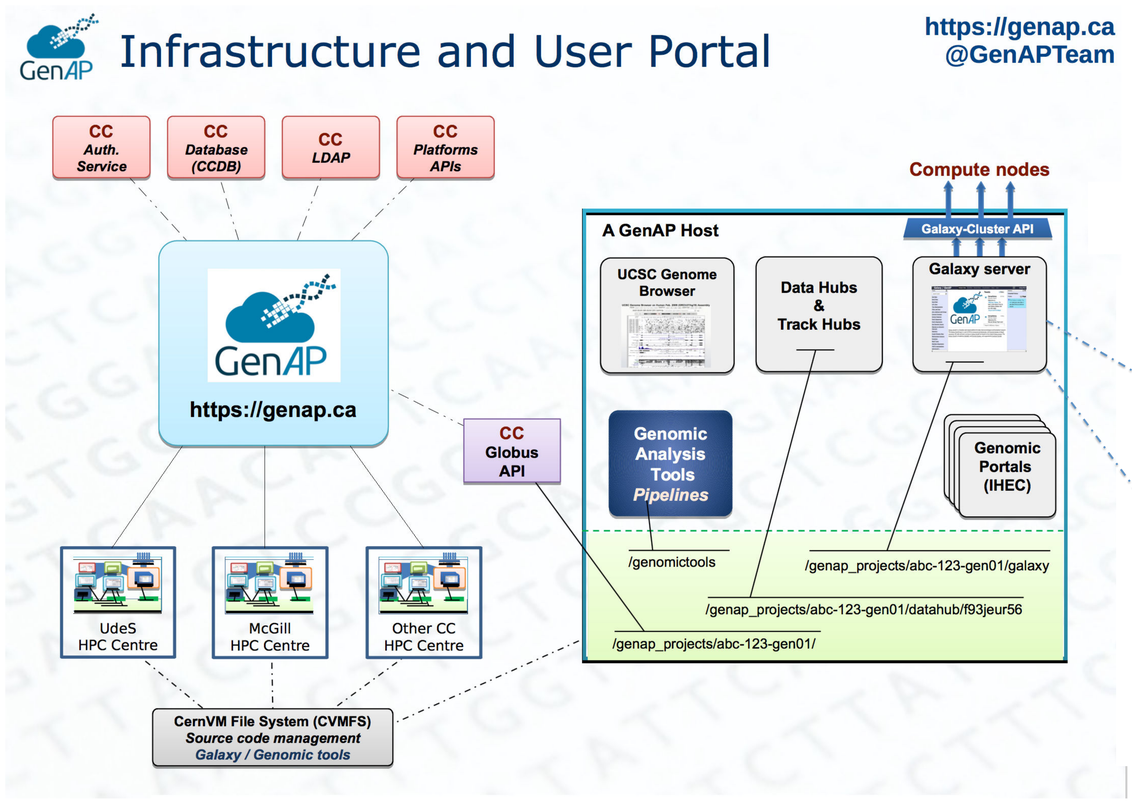
- GenAP is a computing platform for life sciences researchers that leverages both the CANARIE high-speed network and Compute Canada's High Performance Computing (HPC) resources.
- GenAP is targeting both life scientists through web portal platforms such as Galaxy as well as computational biologists through services such as state-of-the-art analysis pipelines and centralized code distribution.
- The platform also hosts data and resources for international projects such as the International Human Epigenome Consoritum (IHEC) Data Portal[1,2,3].
- We are currently integrating Galaxy and the CERN Virtual Machine File System (CVMFS) to facilitate the installation and maintenance of the platform across all GenAP hosts.

Compute / Calcul Canada

- 10+ Data Centers
- 200 000+ cores, 3+ PetaFlops of Compute Power
- 20+ PetaBytes of storage
- PIs and users register centrally using the CC portal

GenAP ENVIRONMENT

- The GenAP infrastructure combines traditional HPC resources with a cloud-like environment using Virtual Machines (VMs) and Access Control List (ACL) to manage applications, files and directories permissions.
- An Interactive server at one HPC site hosts VMs offering the genomic web-based services.
- GenAP is connected to Compute Canada (CC) database and authentication services. This allows simultaneous creation of CC and GenAP accounts.
- GenAP also provides state-of-art genomics pipelines (Chip-Seq, DNA-Seq, RNA-Seq, PacBio_assembly, RNA-denovo_assembly) and genomes indexes/annotations pre-installed in each GenAP host.
- Libraries and code distribution is done using CERN Virtual Machine File System (CVMFS).
- Once logged into the portal user can instantiate a GenAP-hosted application (i.e. data hub or Galaxy).
- In GenAP a private Galaxy can be instantiated from the portal and shared amongst users belonging to the same group or project.

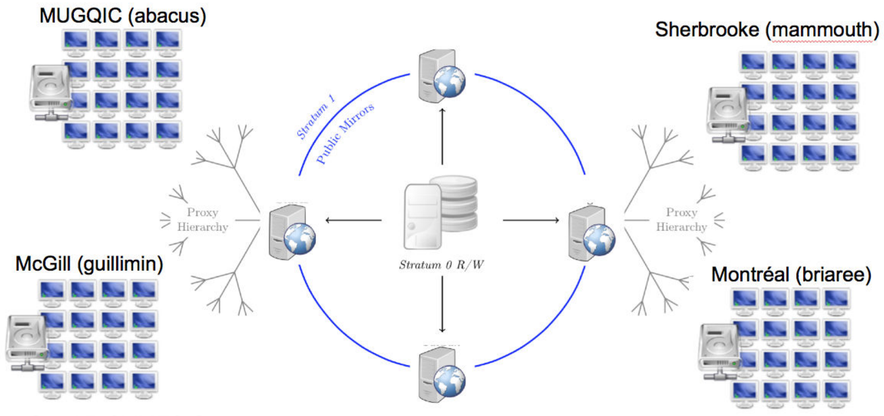


GALAXY INSTANCES

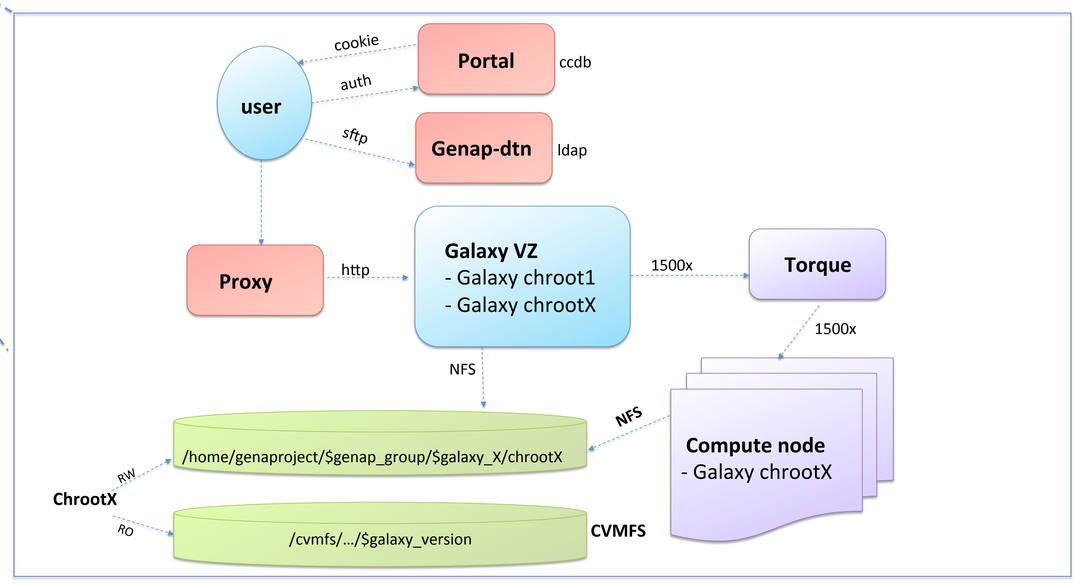
- All Galaxies are installed in a VZ hosted by one of CC HPC centers. Once a user instantiate a Galaxy a chroot of the Galaxy image is created inside the VZ.
- The chroot is created from a two-components template, a read/write NFS, and a read-only CVMFS.
- To run Galaxy the HPC center only needs a CVMFS access in CC. Galaxy runs in userspace with fakechroot utility.
- All software being part of fake(chroot) are distributed read-only by CVMFS. Any software update/patch is automatically propagated on every site.
- The read/write part of the chroot and user data are stored in the local distributed cluster filesystem (NFS, Lustre, GPFS).
- The GenAP portal and proxy handles authentication and safely passes all credential to the Galaxy instances.
- Users can instantiate their own Galaxy as well as add users from other projects. All Galaxy jobs run on the HPC compute node and are computed toward the PI's CC resource allocation.

GALAXY ON CVMFS

- CVMFS is a file system hosted on standard web servers and accessed by clients using FUSE (file system in userspace).
- Software can be installed in one location and cached on demand anywhere using CVMFS.
- How does it work:
  - Changes are done by a software maintainer/librarian on a single location, repository node;
  - Files are committed and replicated from stratum-0 to stratum-1;
  - When file is requested on CVMFS client, local CVMFS client cache is checked;
  - If local CVMFS client cache file is not valid or cannot be found, a new copy is fetched from the local Squid;
  - If local Squid cache cannot find the file or the cache is not valid, a new copy is fetched from Stratum-1 and stored locally on Squid;
- Advantages:
  - Software can be installed and maintained in one location and propagated to multiple sites.
  - Software versioning is reinforced across sites.
  - Software need not be embedded in the original image or VM. VM images and software can be maintained separately.
  - The CVMFS structure is highly scalable and redundant.
  - Software and prerequisites can be installed in CVMFS in order to reduce remote software administration.



GenAP CVMFS Structure



Galaxy and GenAP architecture

Galaxy Instantiation

REFERENCES

1. IHEC http://epigenomesportal.ca/ihec/
2. Phenovar (http://phenovar-dev.udes.genap.ca/)
3. PCRTiler (http://pcrtiler.genap.ca/PCRTiler/).

ACKNOWLEDGMENTS

Jean-François Landry, Universite de Sherbrooke.
Marc-Étienne Rousseau, System Architect, CBRAIN Project.