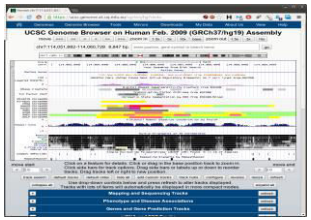
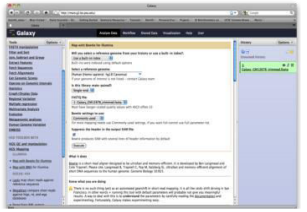
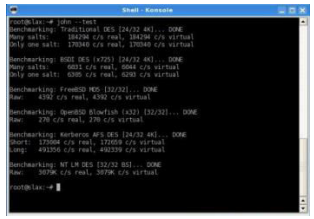
Building the Genomics Virtual Lab

Ron Horst, Uni QLD

Analysis and visualisation platform







UCSC

GENOMESPACE

genomics VIRTUAL LAB

Galaxy

CloudMan

Community Resources

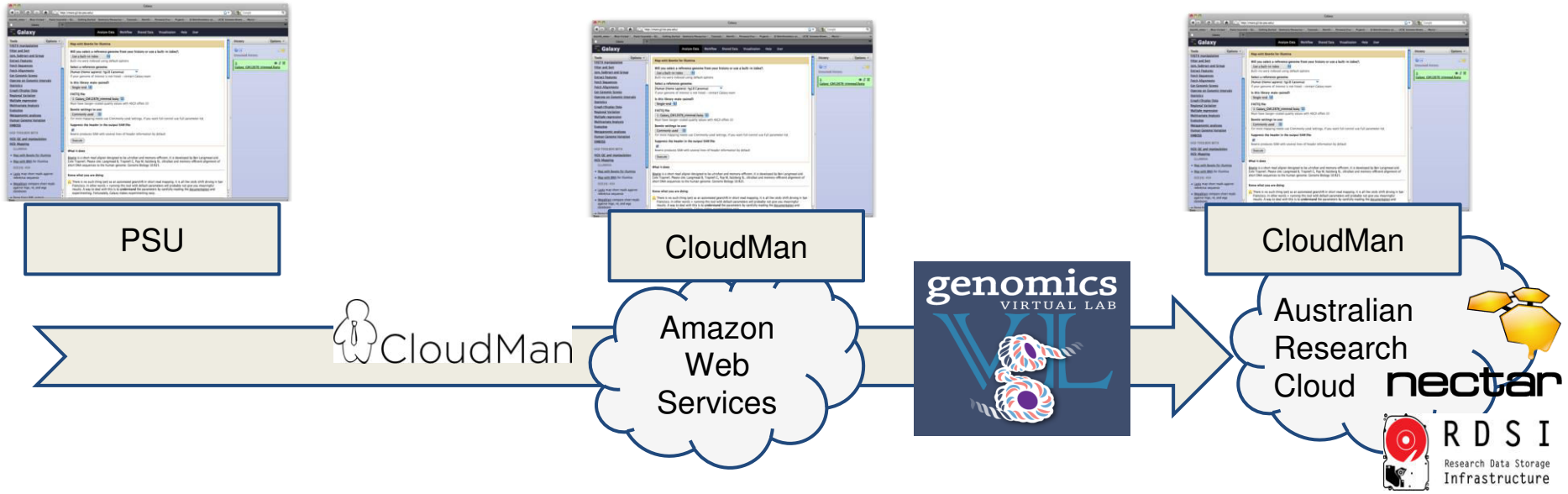
Australian Research Cloud

nectar

Agenda

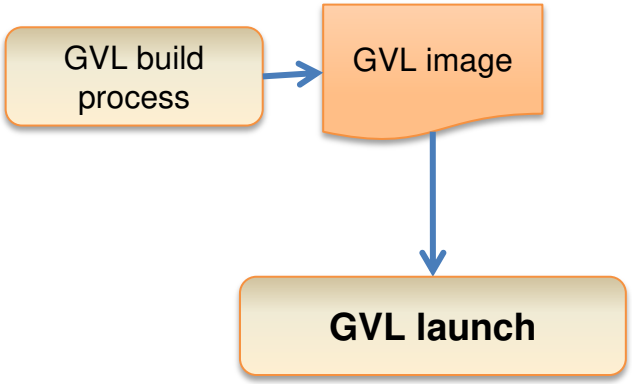
- GVL Objectives
  - –Scalable, on demand → Launch
  - –Latest tools, reproducible → Build Architecture
- Launch Challenges
- Build Challenges

GVL Project Objectives

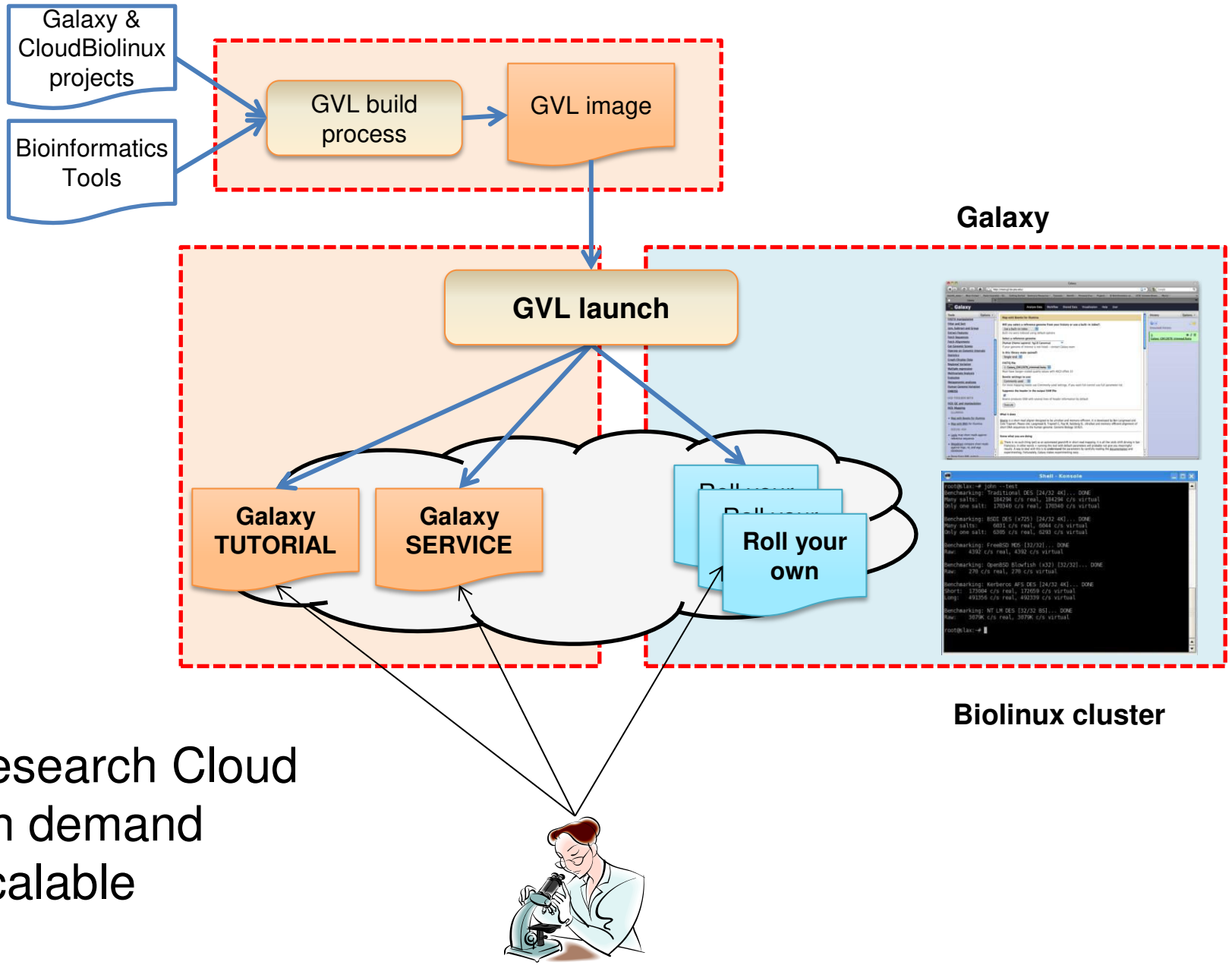


- Leverage Australian National Infrastructure
- On demand
- Scalable

GVL LAUNCH

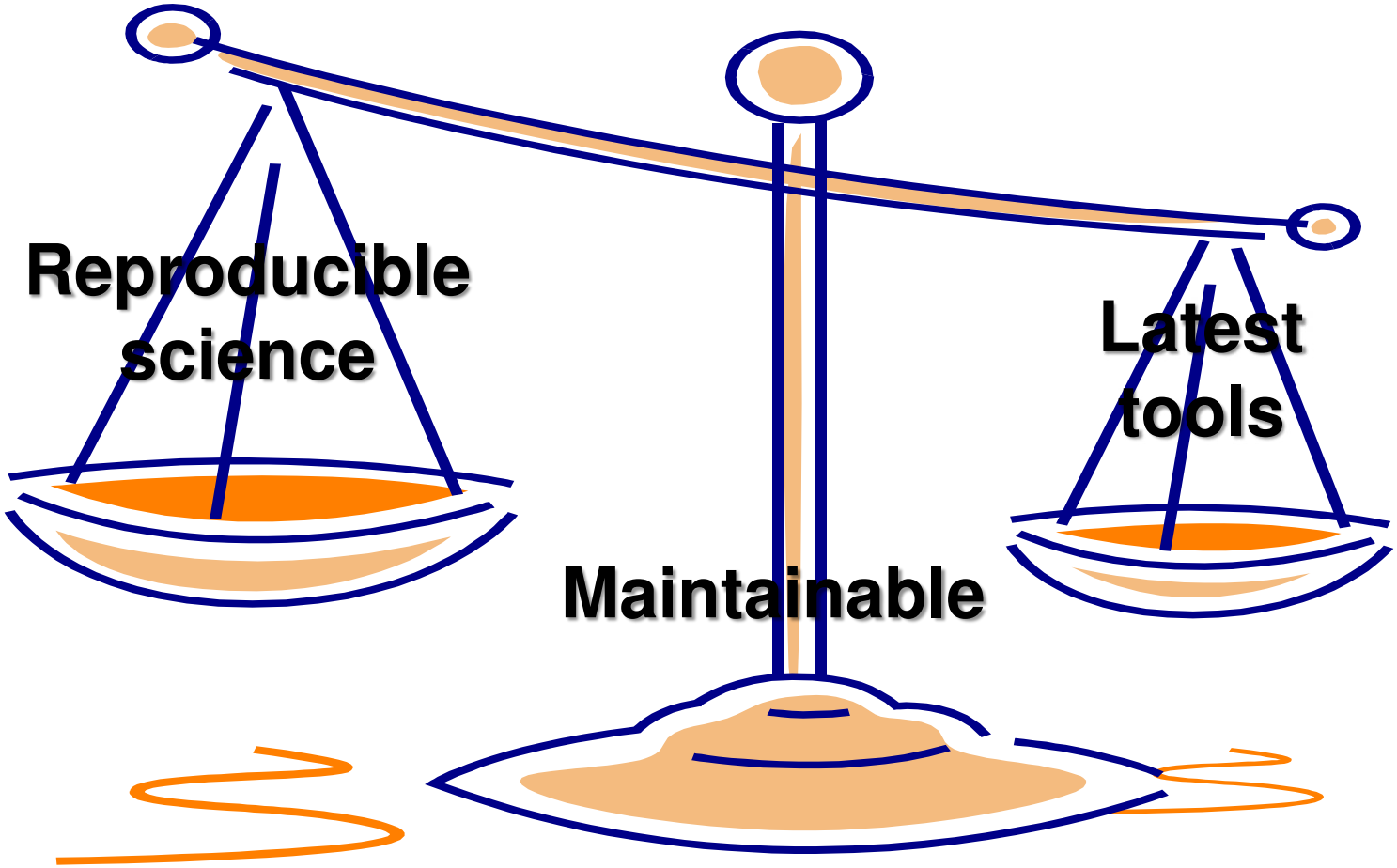


GVL LAUNCH

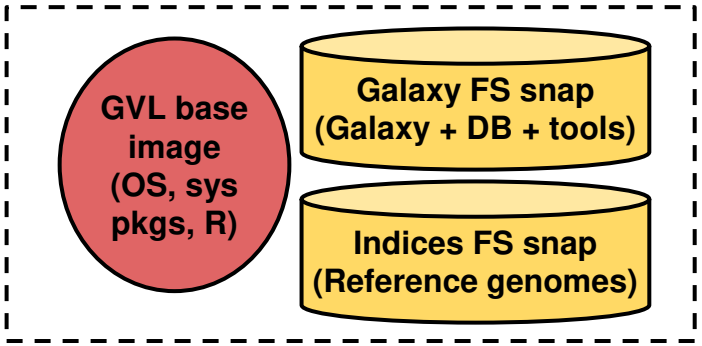


- ✓ Research Cloud
- ✓ On demand
- ✓ Scalable

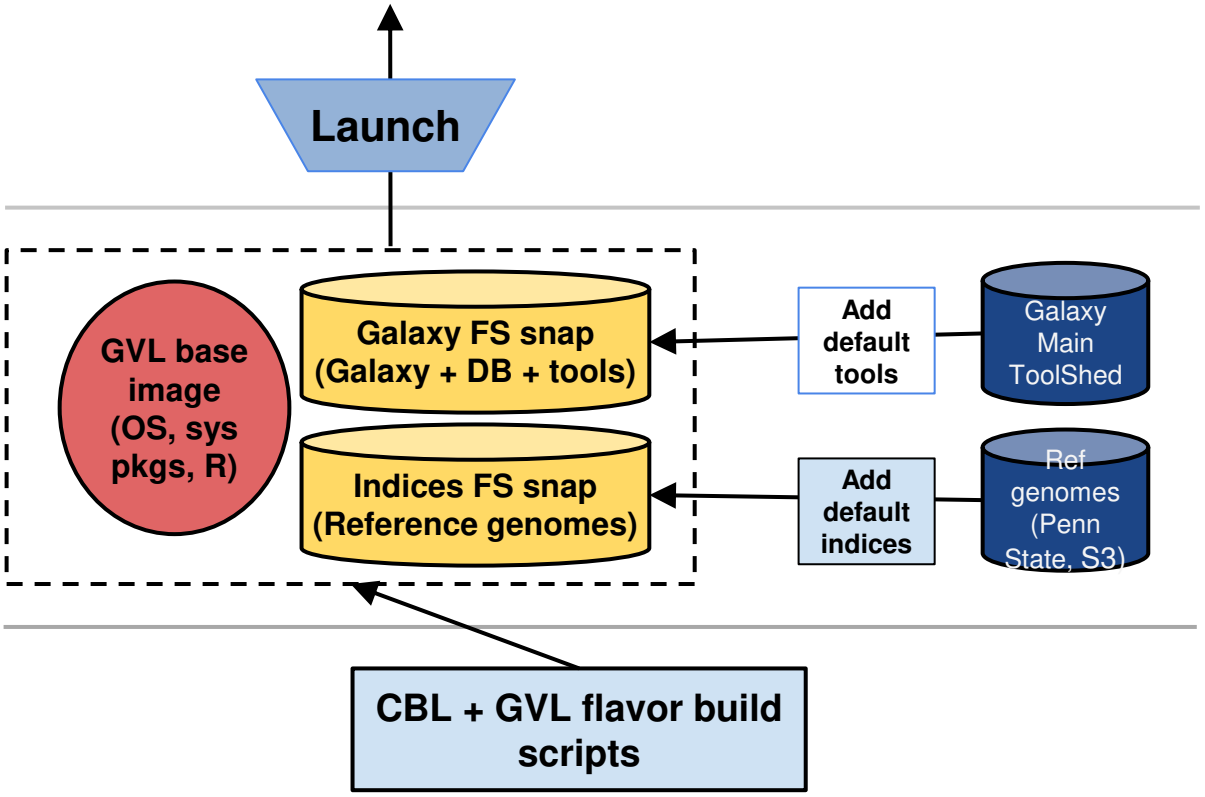
More GVL Objectives



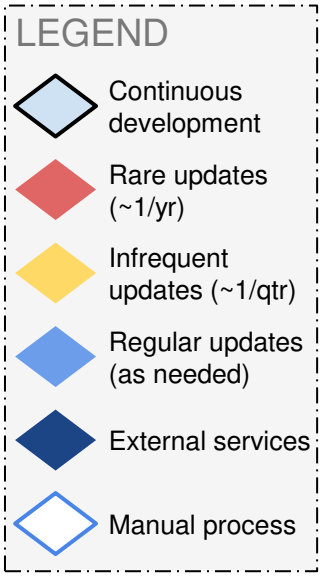
GVL Architecture



GVL Architecture

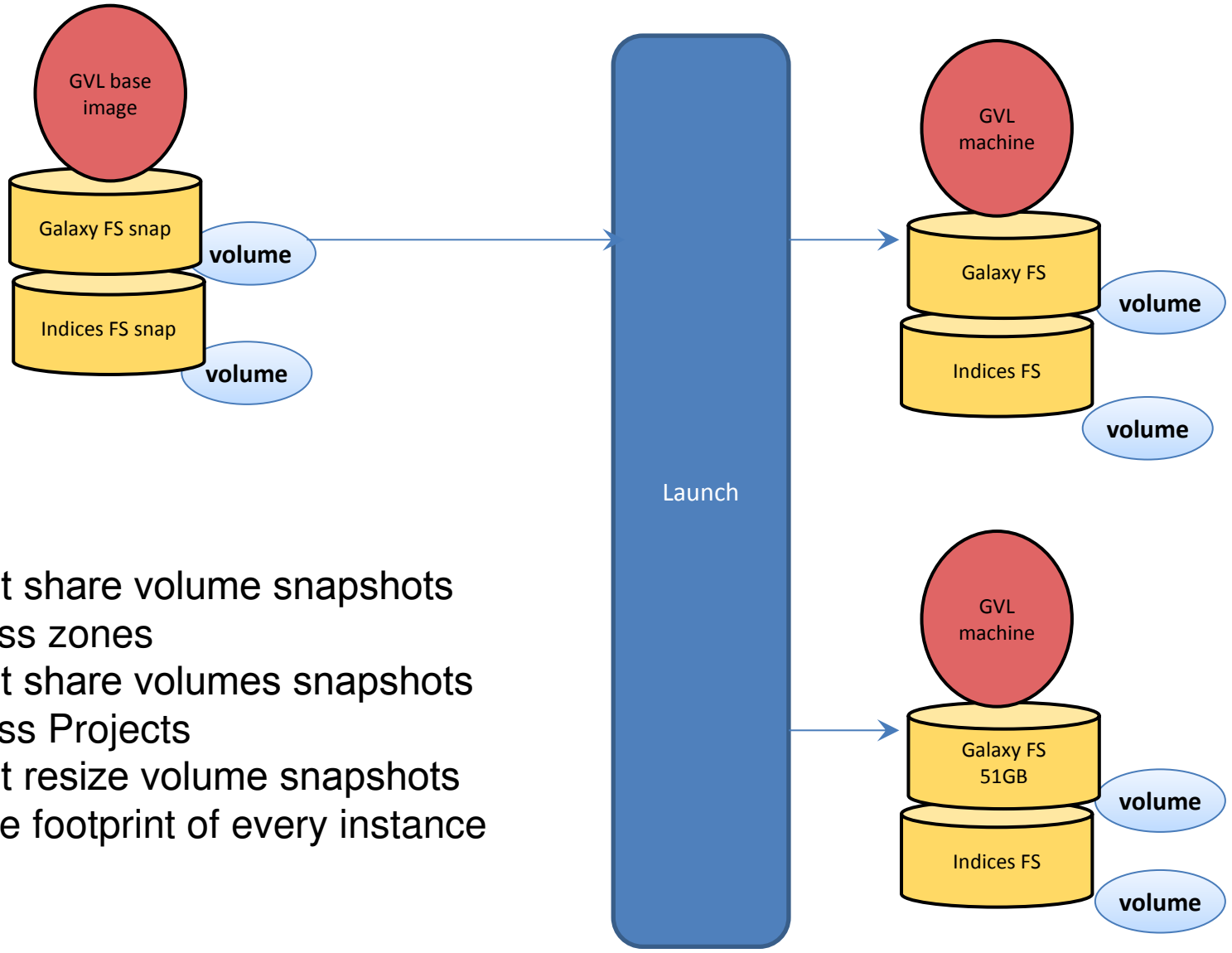


1 per cloud



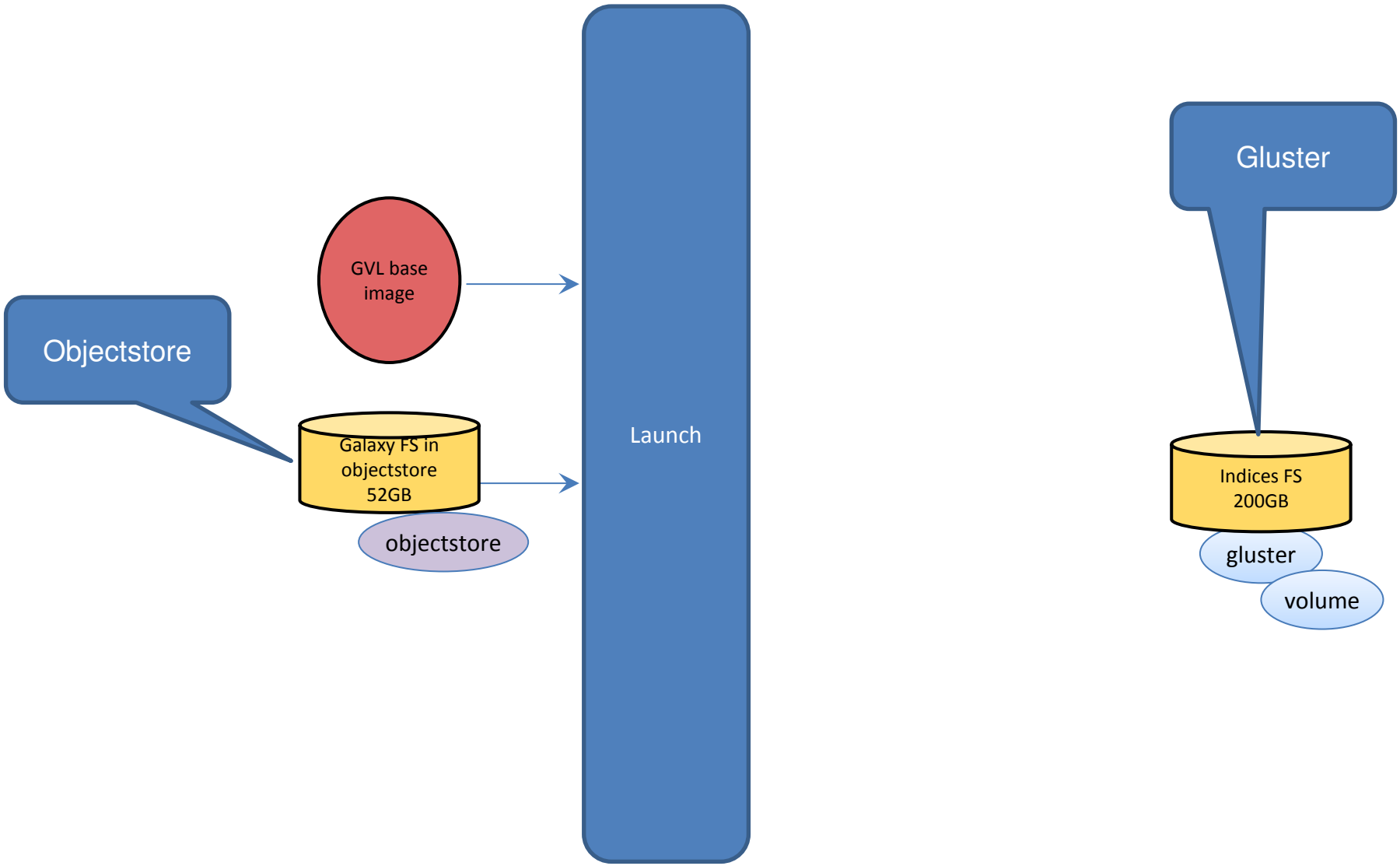
- ✓ Galaxy → Reproducible Science
- ✓ Toolshed → Latest Tools
- ✓ Architecture → Maintainable

GVL Launch process – v1

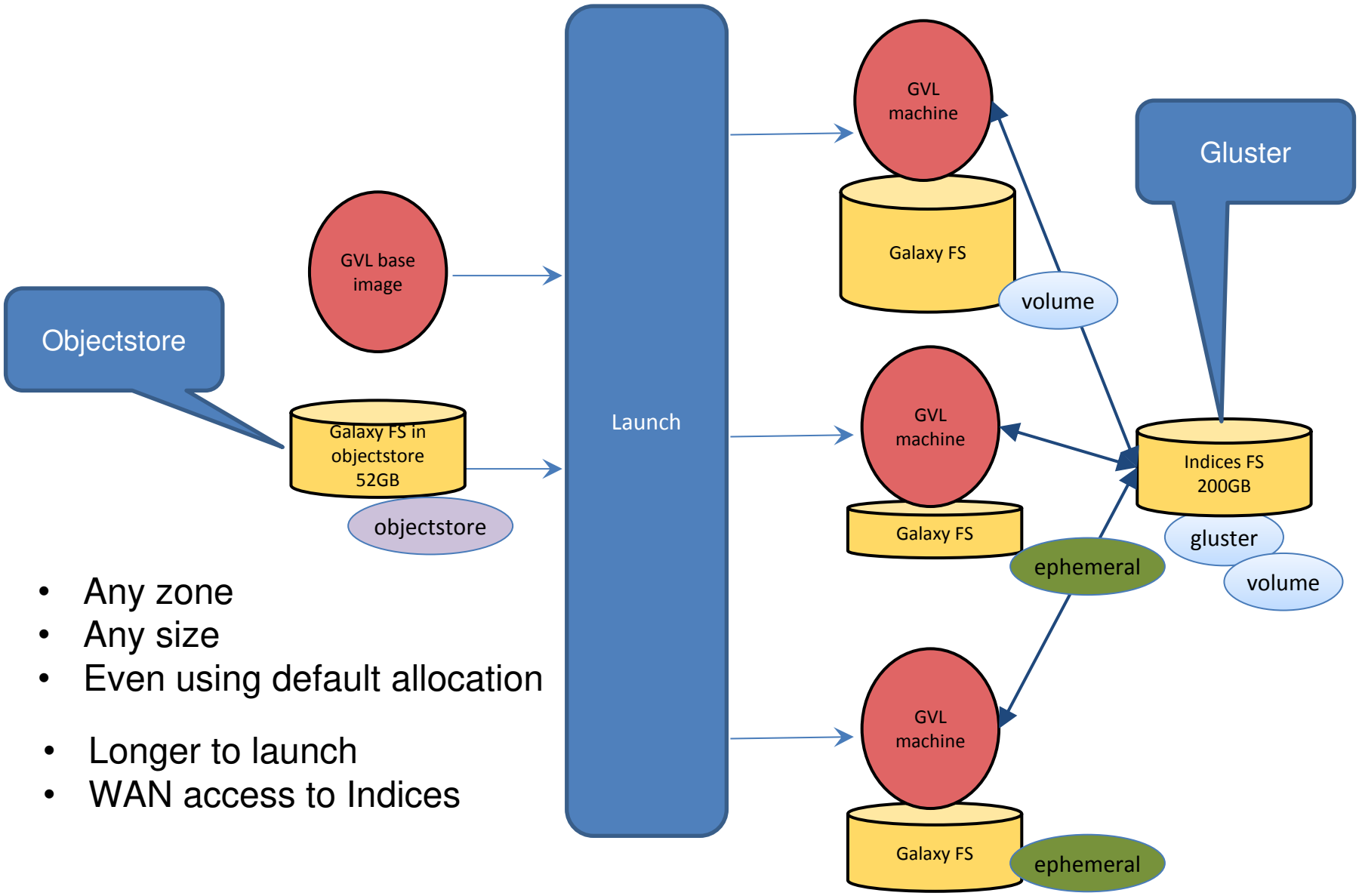


- Can’t share volume snapshots across zones
- Can’t share volumes snapshots across Projects
- Can’t resize volume snapshots
- Large footprint of every instance

GVL Launch process – v2



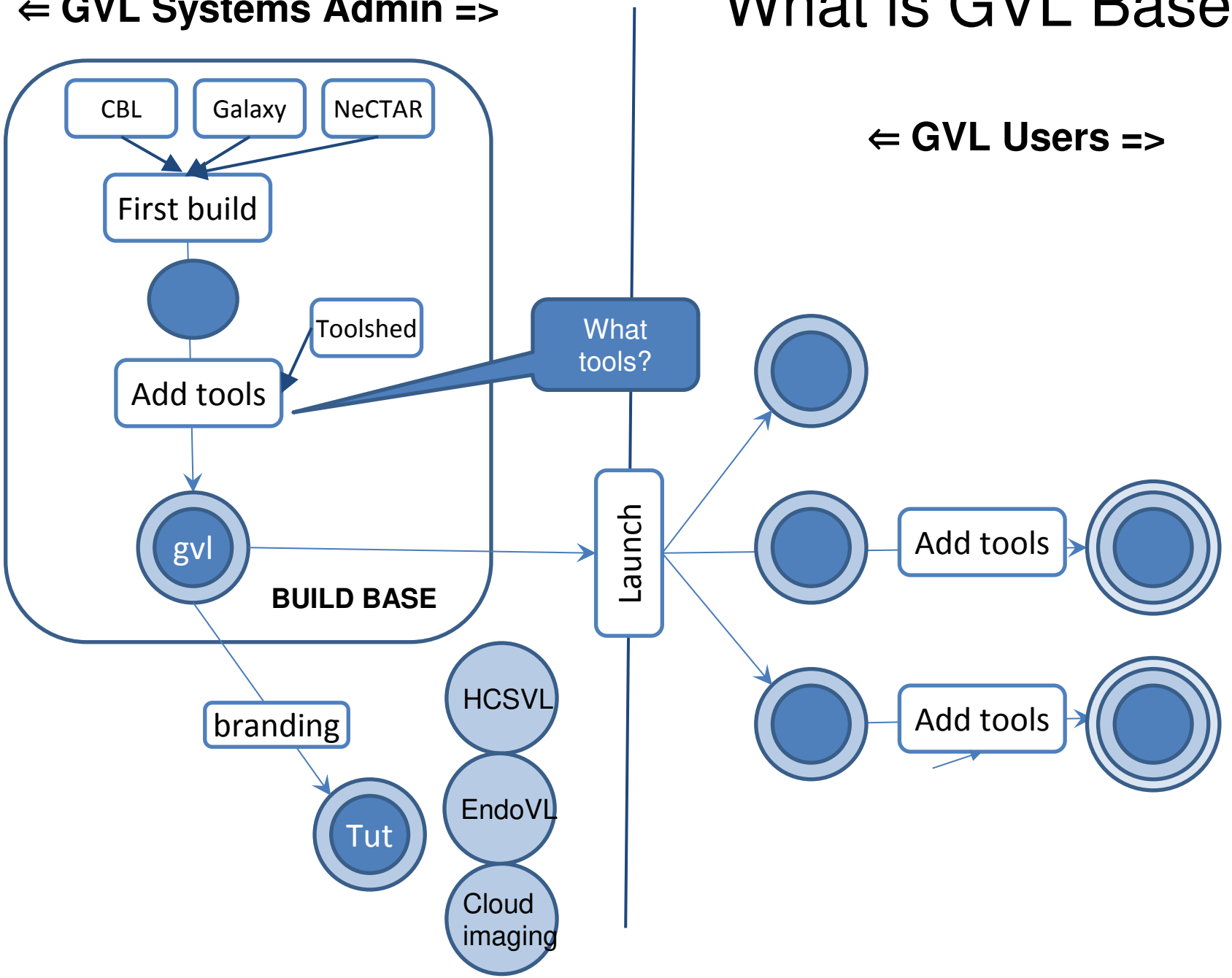
GVL Launch process – v2

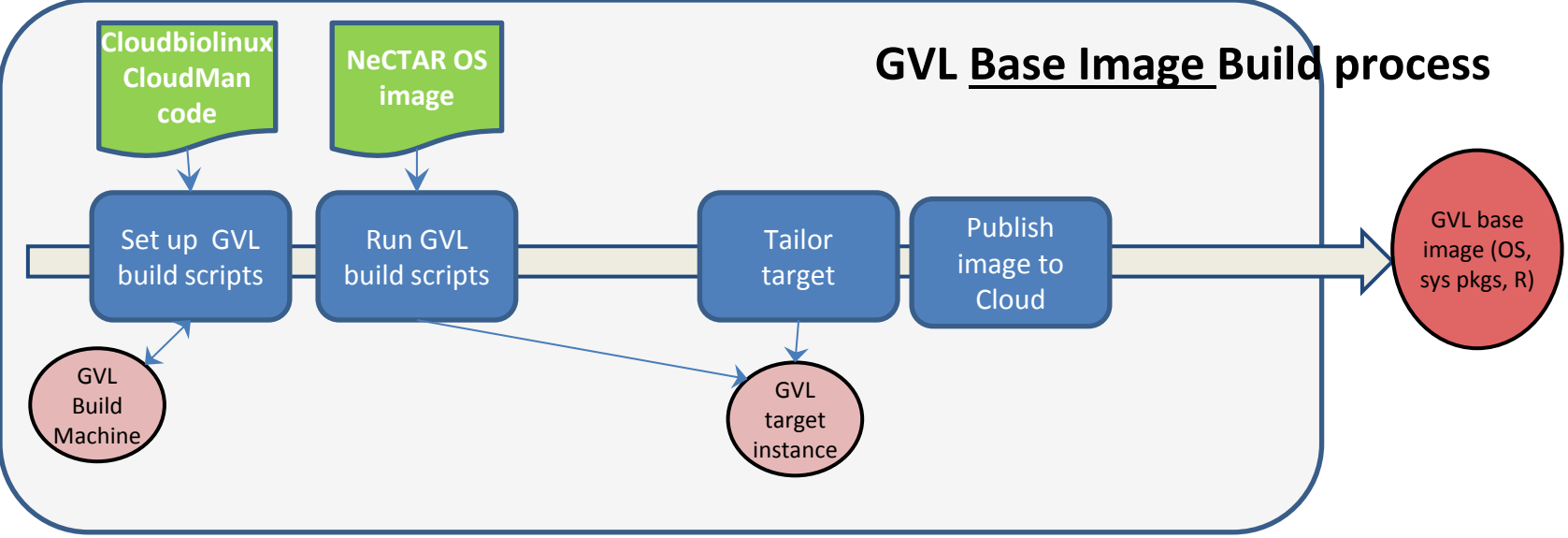


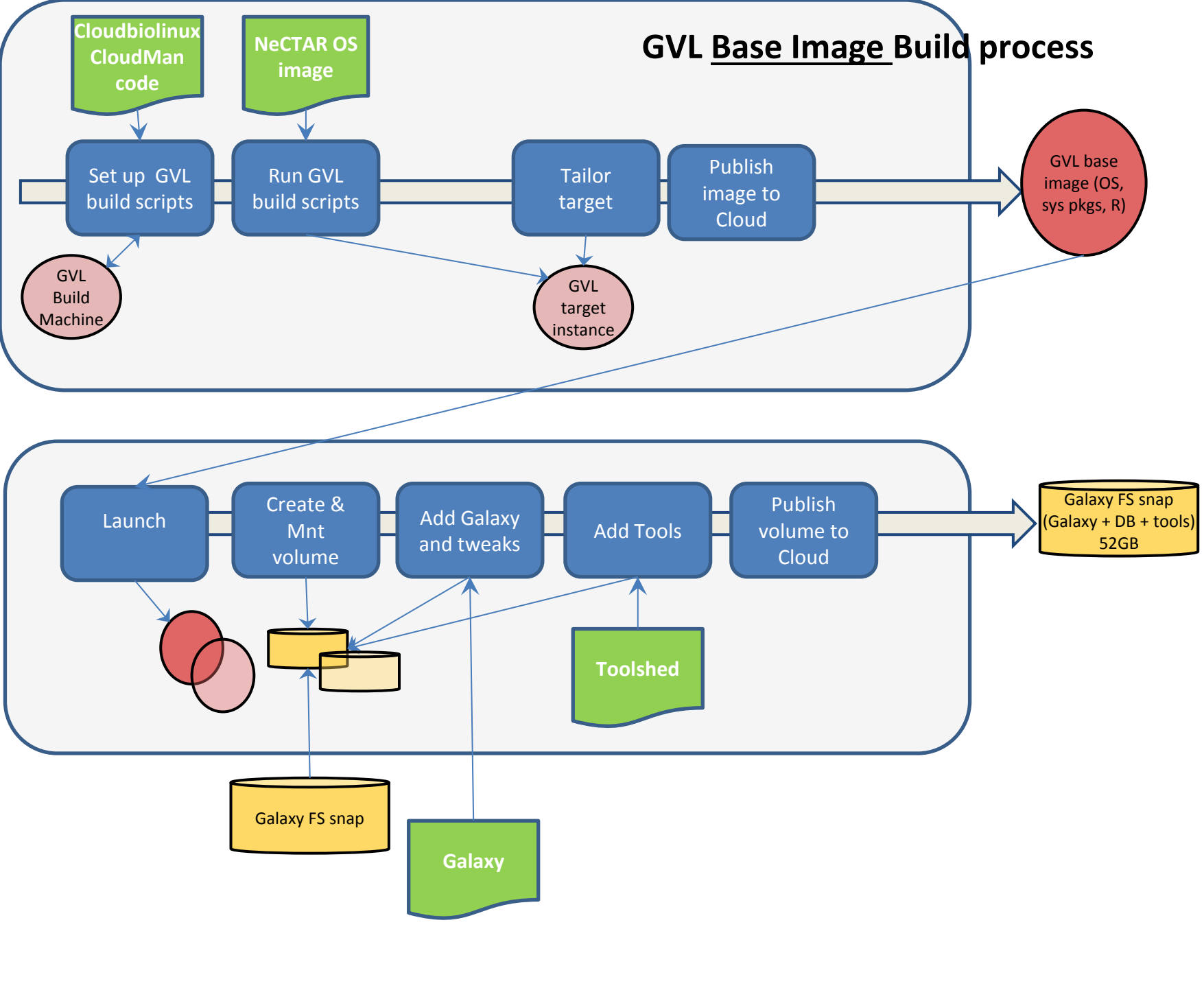
- Any zone
- Any size
- Even using default allocation

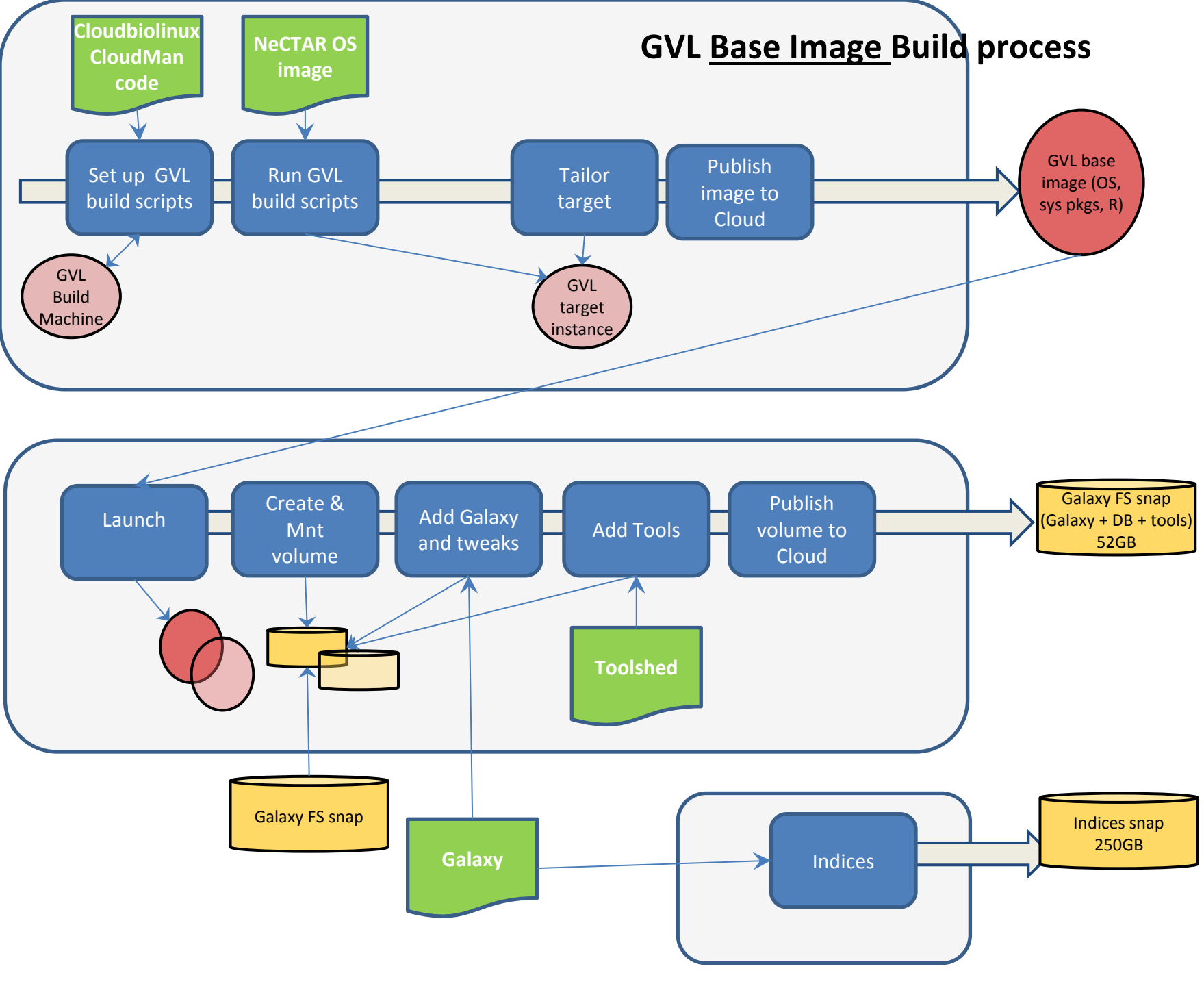
- Longer to launch
- WAN access to Indices

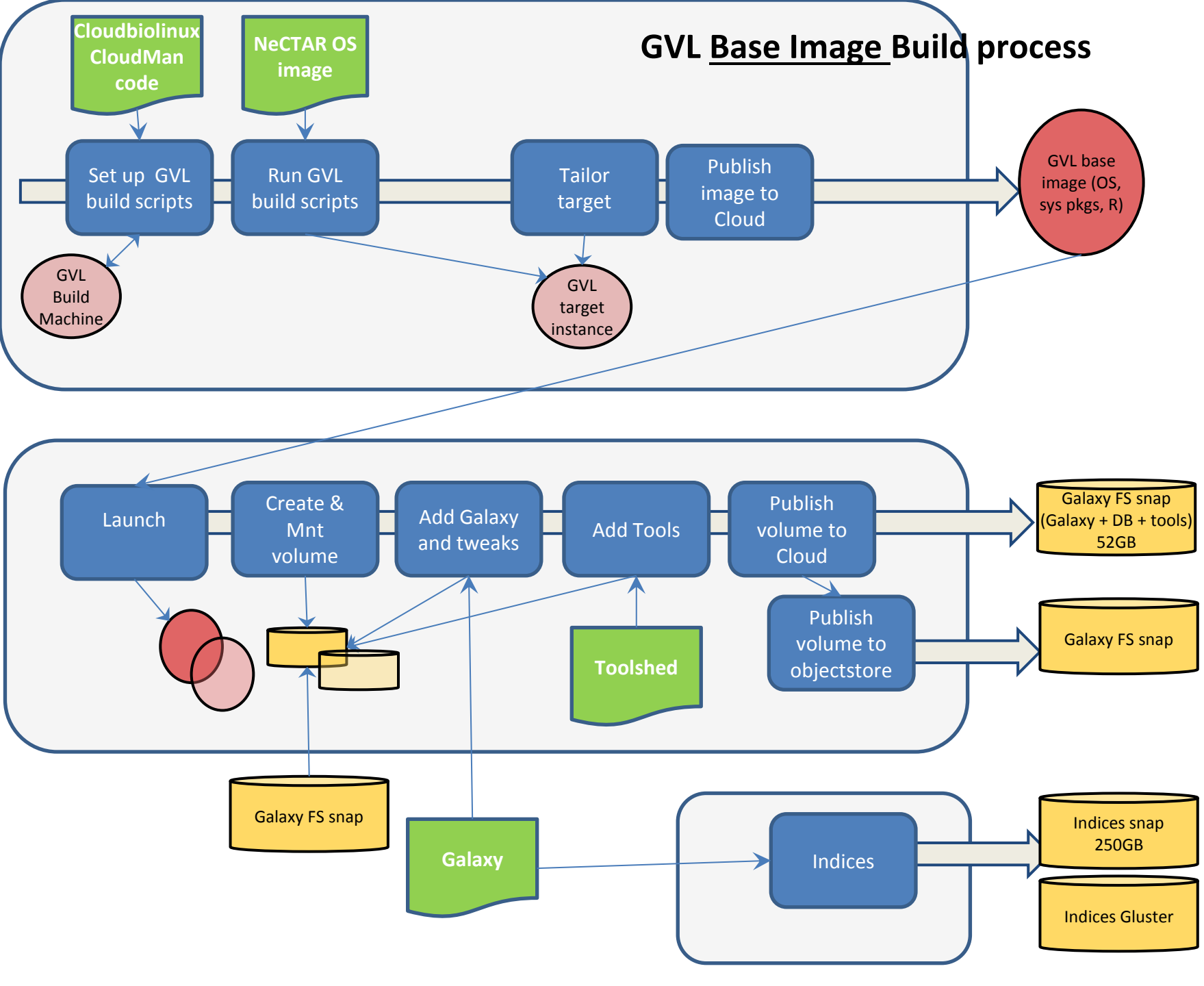
What is GVL Base?



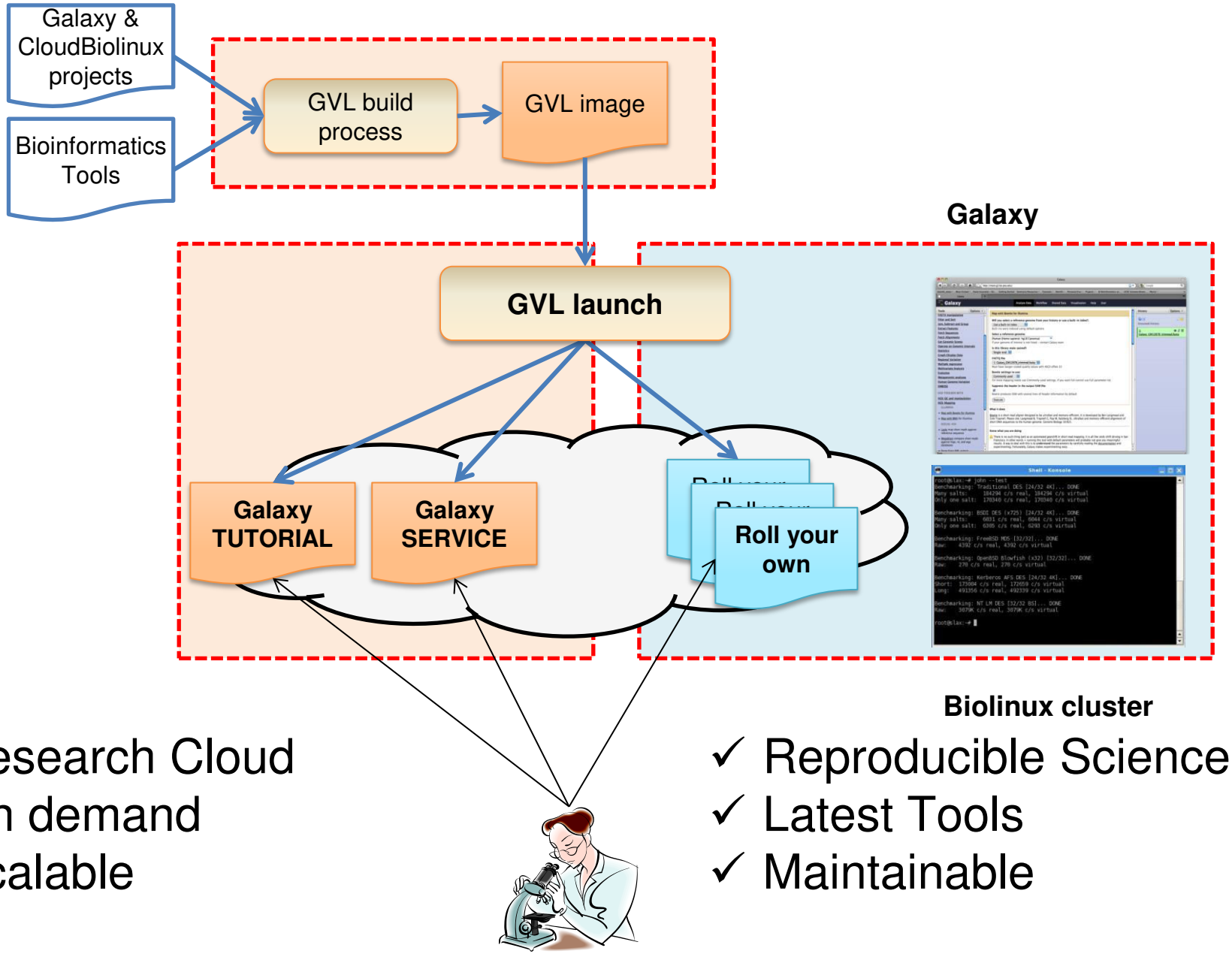

GVL Base Image Build process
Cloudbiolinux CloudMan code
NeCTAR OS image
Set up GVL build scripts
Run GVL build scripts
Tailor target
Publish image to Cloud
GVL base image (OS, sys pkgs, R)
GVL Build Machine
GVL target instance


GVL Base Image Build process
Cloudbiolinux CloudMan code
NeCTAR OS image
Set up GVL build scripts
Run GVL build scripts
Tailor target
Publish image to Cloud
GVL base image (OS, sys pkgs, R)
GVL Build Machine
GVL target instance
Launch
Create & Mnt volume
Add Galaxy and tweaks
Add Tools
Publish volume to Cloud
Galaxy FS snap (Galaxy + DB + tools) 52GB
Toolshed
Galaxy FS snap
Galaxy


GVL Base Image Build process
Cloudbiolinux CloudMan code
NeCTAR OS image
Set up GVL build scripts
Run GVL build scripts
Tailor target
Publish image to Cloud
GVL base image (OS, sys pkgs, R)
GVL Build Machine
GVL target instance
Launch
Create & Mnt volume
Add Galaxy and tweaks
Add Tools
Publish volume to Cloud
Galaxy FS snap (Galaxy + DB + tools) 52GB
Toolshed
Galaxy FS snap
Galaxy
Indices
Indices snap 250GB

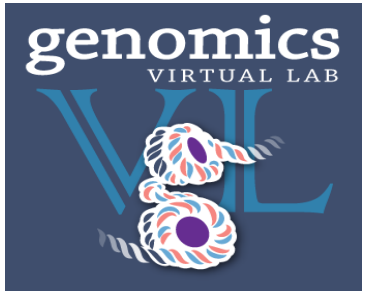

GVL Base Image Build process
Cloudbiolinux CloudMan code
NeCTAR OS image
Set up GVL build scripts
Run GVL build scripts
Tailor target
Publish image to Cloud
GVL base image (OS, sys pkgs, R)
GVL Build Machine
GVL target instance
Launch
Create & Mnt volume
Add Galaxy and tweaks
Add Tools
Publish volume to Cloud
Galaxy FS snap (Galaxy + DB + tools) 52GB
Publish volume to objectstore
Galaxy FS snap
Toolshed
Galaxy FS snap
Galaxy
Indices
Indices snap 250GB
Indices Gluster

GVL LAUNCH



- ✓ Research Cloud
- ✓ On demand
- ✓ Scalable

- ✓ Reproducible Science
- ✓ Latest Tools
- ✓ Maintainable


genomics
VIRTUAL LAB
VL


nectar

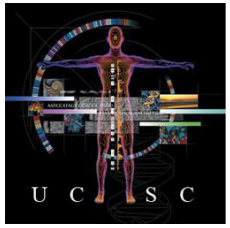

vlsci
LIFE SCIENCES COMPUTATION
QFAB


THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA


GENOMESPACE


Galaxy


CloudMan


U C S C


Victorian
Bioinformatics
Consortium

Thank you


Bioplatforms Australia


CSIRO


qcif


POSTERA CRESCAM LAUDE
THE UNIVERSITY OF
MELBOURNE


Baker IDI
HEART & DIABETES INSTITUTE


GARVAN
INSTITUTE

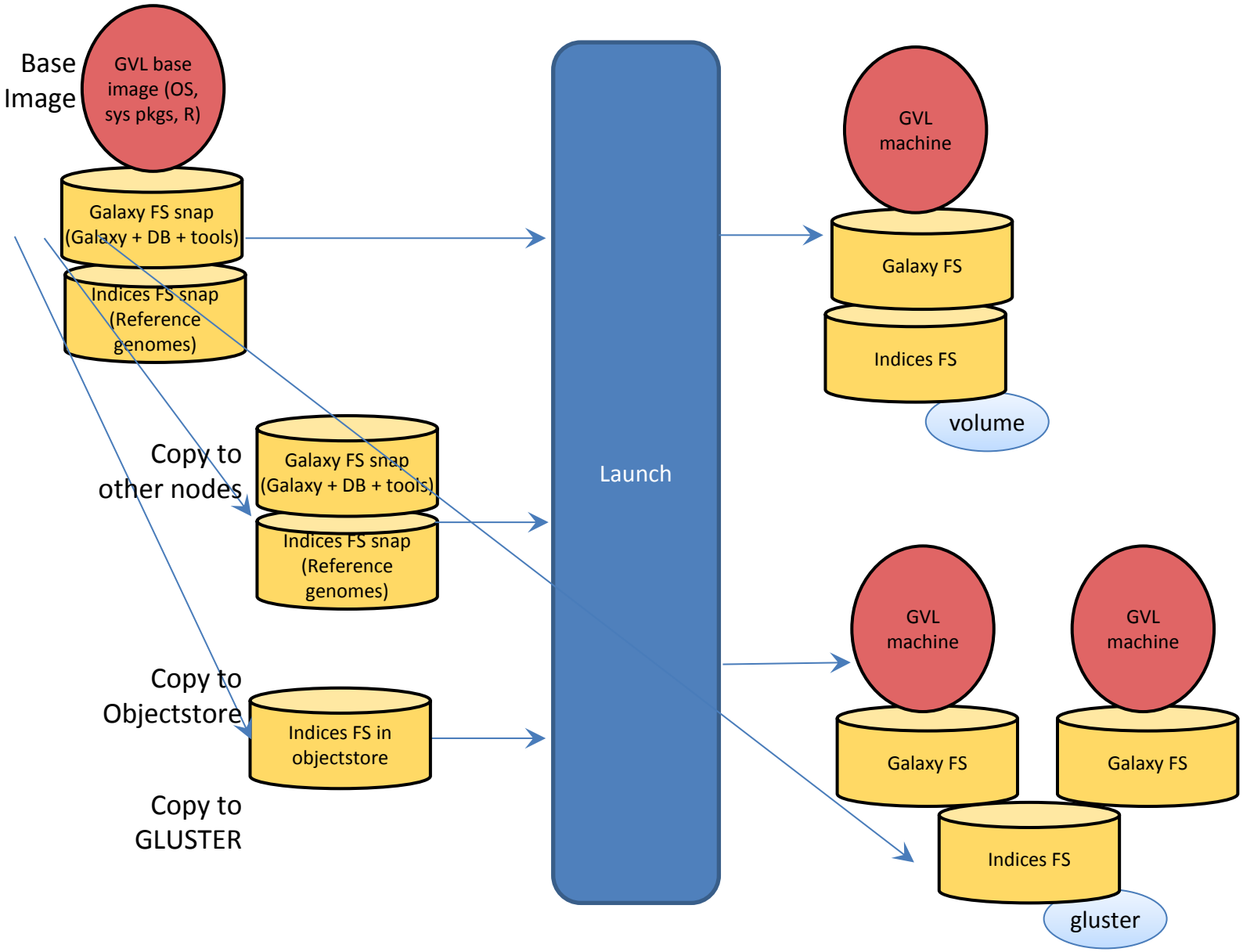

RDSI
Research Data Storage
Infrastructure


VeRSI


Peter Mac

The end

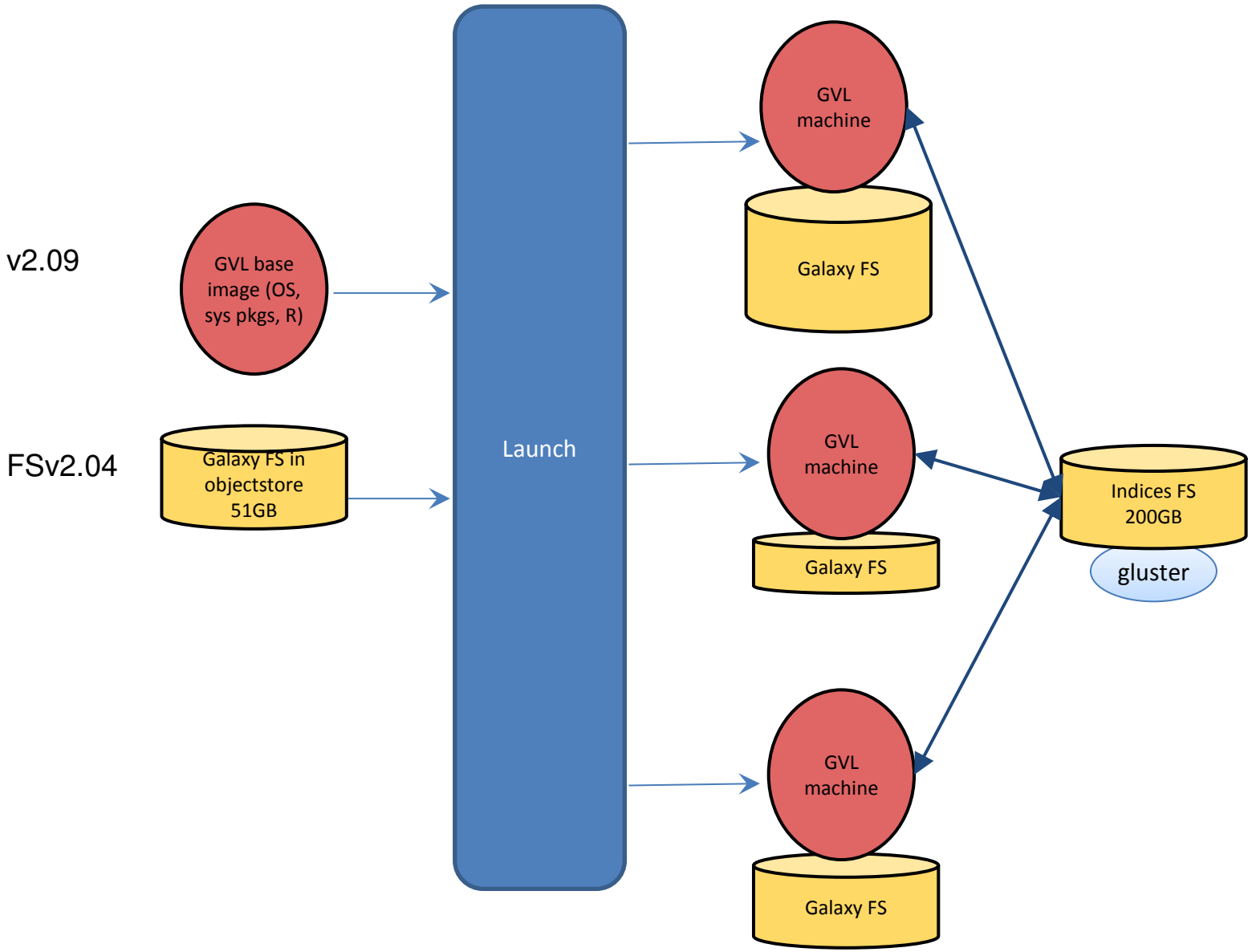
GVL Launch process



GVL Launch process - “NeCTAR (openstack)” option

Any NeCTAR project, any resizing of GalaxyFS

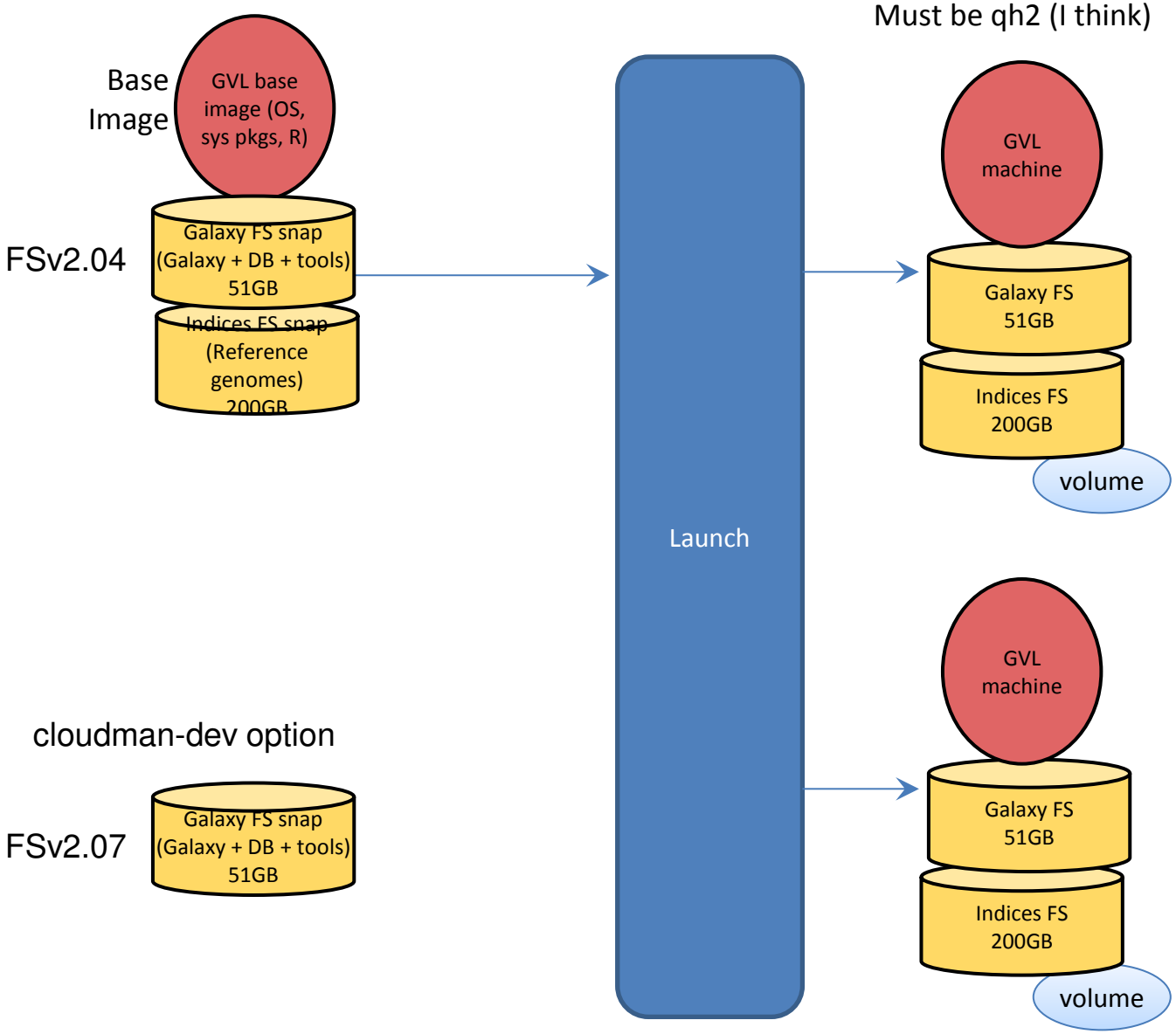
Any placement zone with Volumes

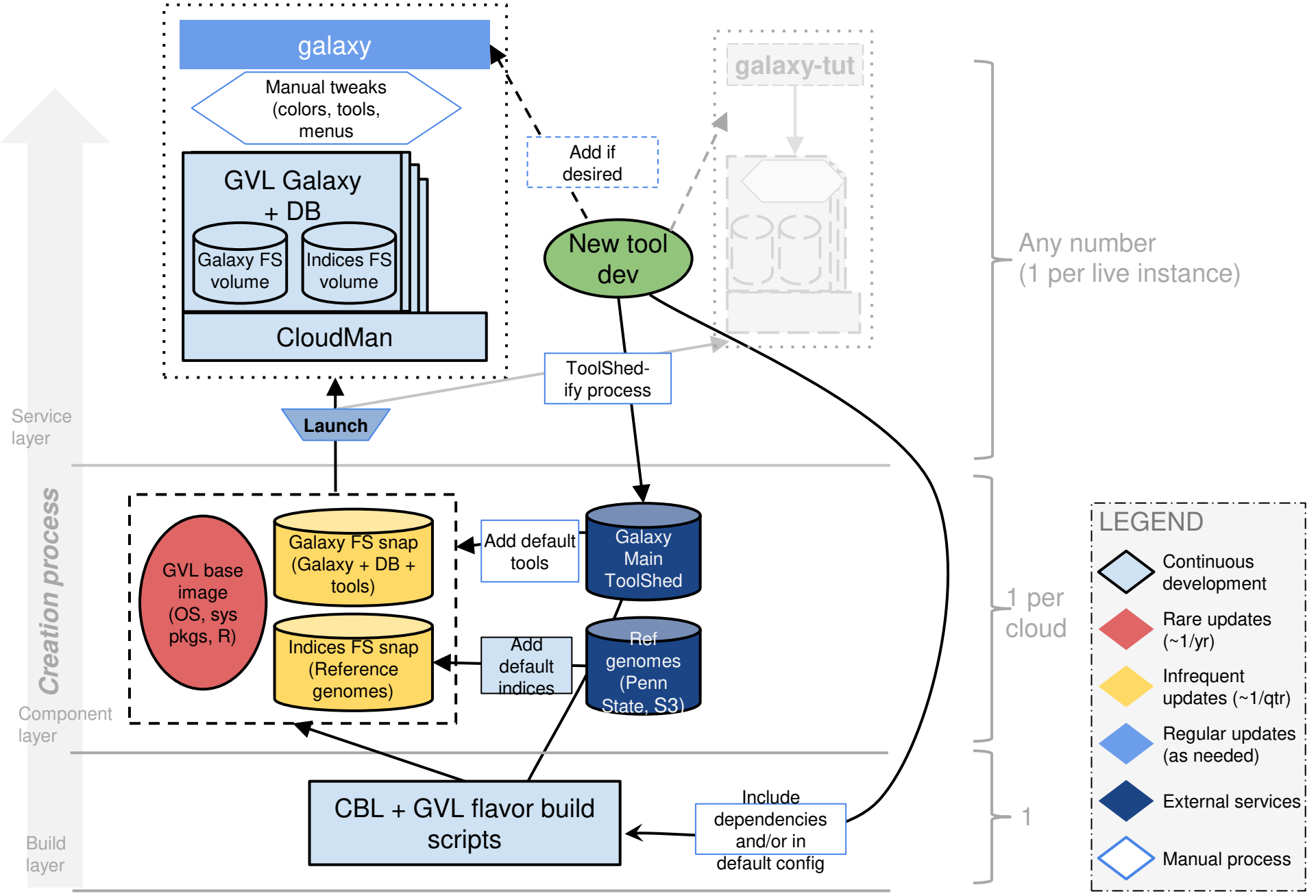


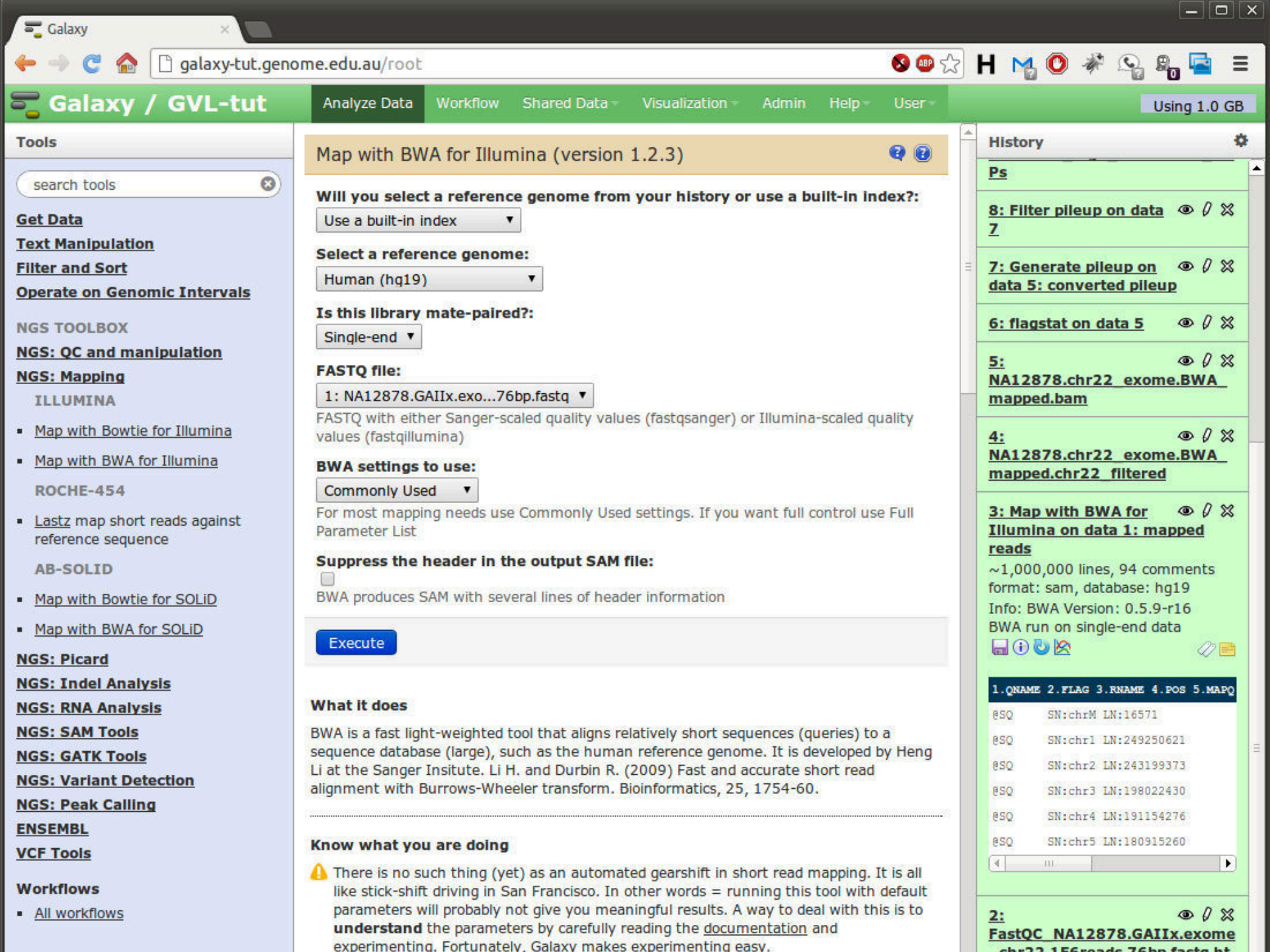
GVL Launch process - “NeCTAR melbourne dev” option

Must be GenomicsVL project, cannot resize GalaxyFS

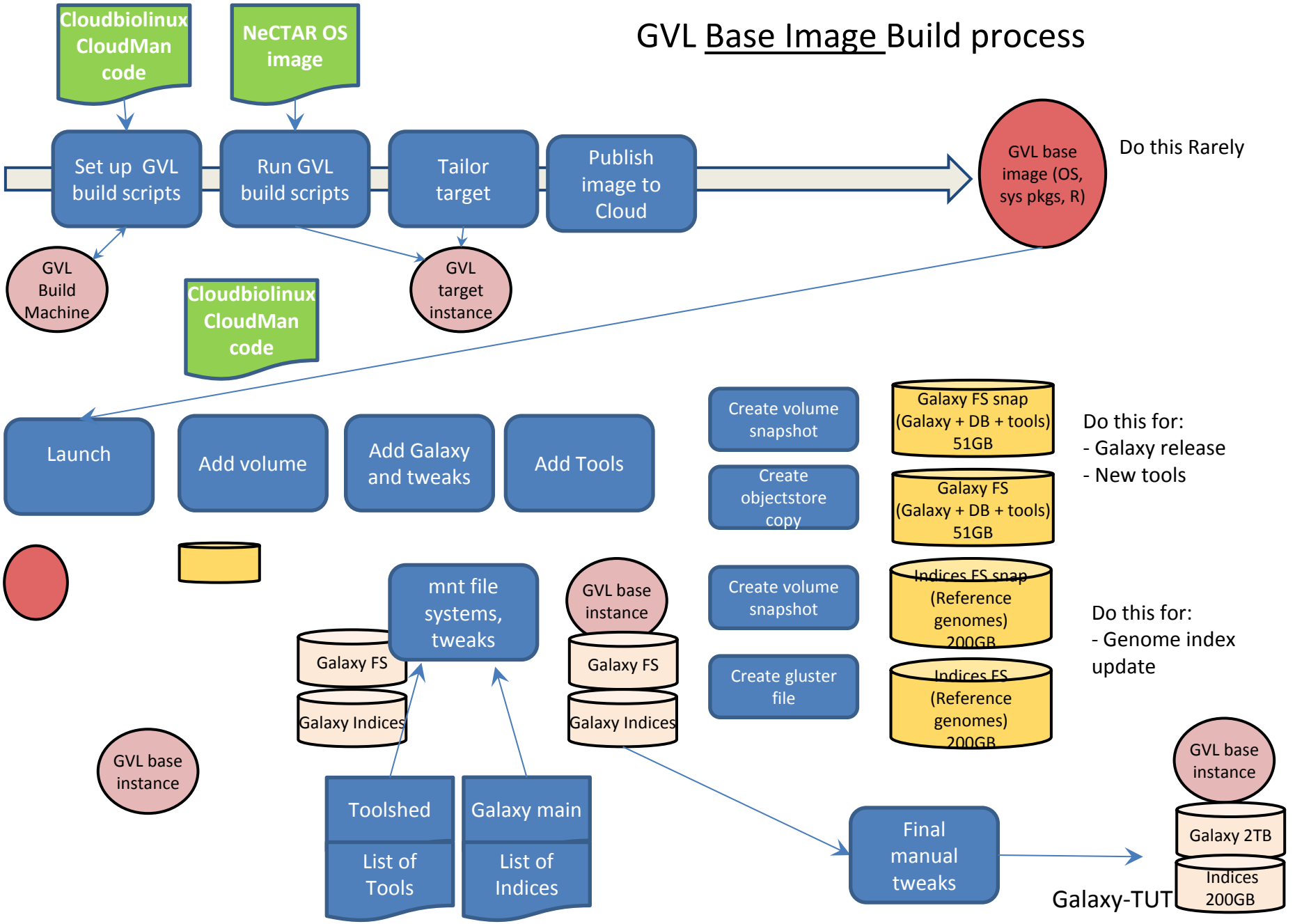
Must be qh2 (I think)




galaxy
Manual tweaks (colors, tools, menus
GVL Galaxy + DB
Galaxy FS volume
Indices FS volume
CloudMan
Add if desired
New tool dev
galaxy-tut
ToolShed-ify process
Launch
Service layer
Creation process
Component layer
Build layer
GVL base image (OS, sys pkgs, R)
Galaxy FS snap (Galaxy + DB + tools)
Indices FS snap (Reference genomes)
Add default tools
Add default indices
Galaxy Main ToolShed
Ref genomes (Penn State, S3)
CBL + GVL flavor build scripts
Include dependencies and/or in default config
Any number (1 per live instance)
1 per cloud
1
LEGEND
Continuous development
Rare updates (~1/yr)
Infrequent updates (~1/qtr)
Regular updates (as needed)
External services
Manual process


Galaxy
galaxy-tut.genome.edu.au/root
Galaxy / GVL-tut
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 1.0 GB
Tools
search tools
Get Data
Text Manipulation
Filter and Sort
Operate on Genomic Intervals
NGS TOOLBOX
NGS: QC and manipulation
NGS: Mapping
ILLUMINA
Map with Bowtie for Illumina
Map with BWA for Illumina
ROCHE-454
Lastz map short reads against reference sequence
AB-SOLID
Map with Bowtie for SOLiD
Map with BWA for SOLiD
NGS: Picard
NGS: Indel Analysis
NGS: RNA Analysis
NGS: SAM Tools
NGS: GATK Tools
NGS: Variant Detection
NGS: Peak Calling
ENSEMBL
VCF Tools
Workflows
All workflows
Map with BWA for Illumina (version 1.2.3)
Will you select a reference genome from your history or use a built-in index?:
Use a built-in index
Select a reference genome:
Human (hg19)
Is this library mate-paired?:
Single-end
FASTQ file:
1: NA12878.GAIIx.exo...76bp.fastq
FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)
BWA settings to use:
Commonly Used
For most mapping needs use Commonly Used settings. If you want full control use Full Parameter List
Suppress the header in the output SAM file:
BWA produces SAM with several lines of header information
Execute
What it does
BWA is a fast light-weighted tool that aligns relatively short sequences (queries) to a sequence database (large), such as the human reference genome. It is developed by Heng Li at the Sanger Insitute. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-60.
Know what you are doing
There is no such thing (yet) as an automated gearshift in short read mapping. It is all like stick-shift driving in San Francisco. In other words = running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to understand the parameters by carefully reading the documentation and experimenting. Fortunately, Galaxy makes experimenting easy.
History
Ps
8: Filter pileup on data 7
7: Generate pileup on data 5: converted pileup
6: flagstat on data 5
5: NA12878.chr22_exome.BWA_mapped.bam
4: NA12878.chr22_exome.BWA_mapped.chr22_filtered
3: Map with BWA for Illumina on data 1: mapped reads
~1,000,000 lines, 94 comments
format: sam, database: hg19
Info: BWA Version: 0.5.9-r16
BWA run on single-end data
1.QNAME 2.FLAG 3.RNAME 4.POS 5.MAPQ
@SQ SN:chrM LN:16571
@SQ SN:chr1 LN:249250621
@SQ SN:chr2 LN:243199373
@SQ SN:chr3 LN:198022430
@SQ SN:chr4 LN:191154276
@SQ SN:chr5 LN:180915260
2: FastQC_NA12878.GAIIx.exome

GVL Base Image Build process



What is GVL Base?