Fermilab

May 7, 2025

Using Apptainer in a Pilot-based Distributed Workload

Marco Mambelli
Senior Software Developer

U.S. DEPARTMENT of ENERGY
Fermi National Accelerator Laboratory is managed by FermiForward for the U.S. Department of Energy Office of Science

FERMILAB-SLIDES-25-0090-CSAID

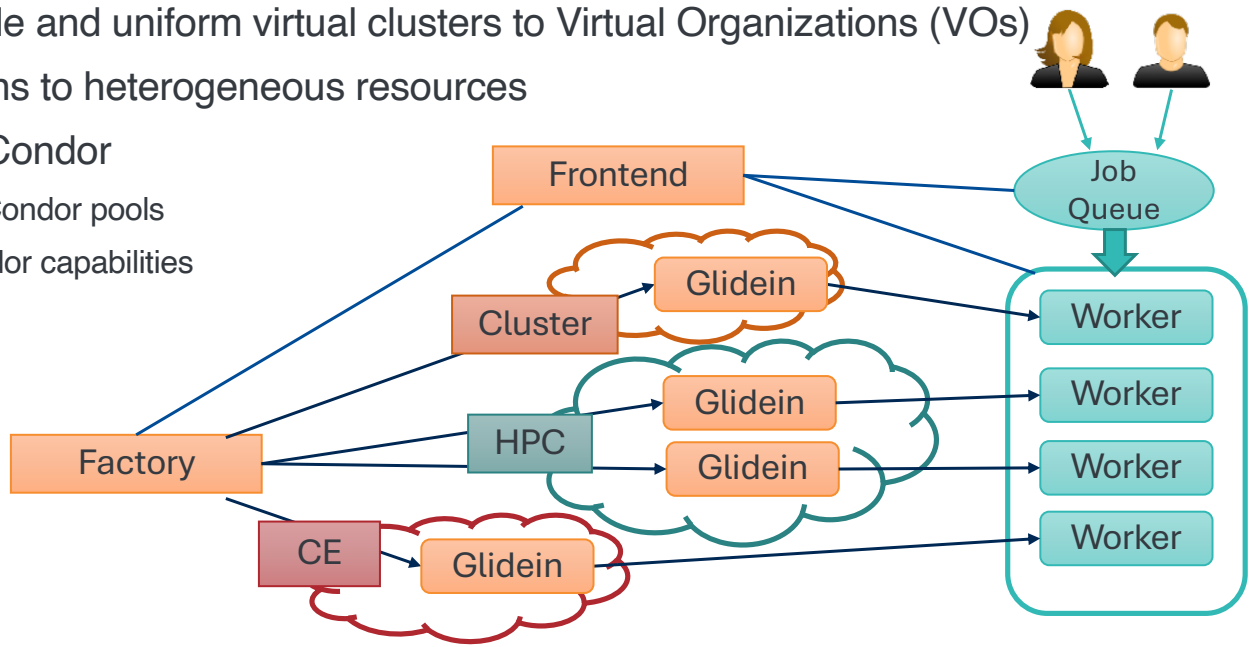
FERMI NATIONAL ACCELERATOR LABORATORY

GlideinWMS

HPSF CONFERENCE 2025

GlideinWMS is a pilot-based resource provisioning tool for distributed High Throughput Computing

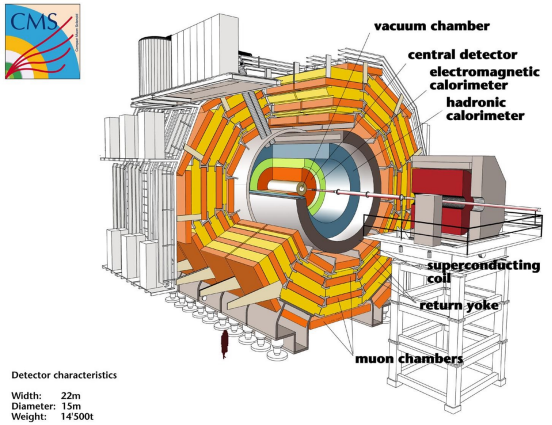
- Provides reliable and uniform virtual clusters to Virtual Organizations (VOs)
- Submits Glideins to heterogeneous resources
- Leverages HTCondor
  - Provides HTCondor pools
  - Uses HTCondor capabilities

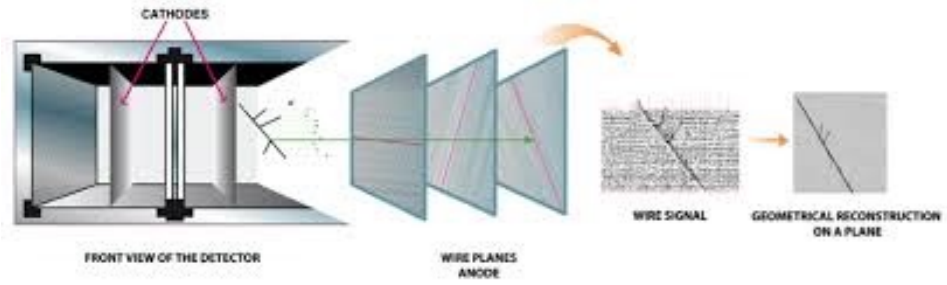


High Throughput Computing

- Split the task in many independent parts
- Resolve them on separate resources

- Even HPC resources are used node by node
  - Multi-node submission
  - One Glidein per node





Glidein: node testing and customization

- Scouts for resources and validates the Worker node
  - Cores, memory, disk, GPU, …
  - OS, software installed
  - CVMFS and available File Systems
  - VO specific tests
- Customizes the Worker node
  - Environment, GPU libraries, …
  - Starting containers (Apptainer/Singularity)
  - VO specific setup
- Provides a reliable and customized execute node to HTCondor

Factory

- A Glidein Factory knows how to submit to sites
  - Sites are described in a local configuration
  - Only trusted and tested sites are included
- Each site entry in the configuration contains
  - Contact info (hostname, resource type, queue name)
  - Site configuration (startup dir, OS type, …)
  - VOs authorized/supported
  - Other attributes (Site name, core count, max memory, ...)
  - Glideins can also auto-detect resources
- Configuration can be auto-generated (e.g. from CRIC), admin curated, stored in VCS (e.g. GitHub)
- HTCondor does the heavy lifting of submissions.

Factory: Supported resources

- Remote or local clusters:
  - Can have batch systems other than HTCondor: PBS, SGE, Slurm, all supported.
- Grid sites (CREAM, ARC, HTCondor-CE)
- Hosted CEs
- Commercial Clouds (AWS, Google)
- Open Source Clouds (OpenStack, OpenNebula)
- HPC sites
  - Uses an ssh-based system to ssh into HPC sites and submit directly from their login nodes
  - One node - one Glidein

FERMI NATIONAL ACCELERATOR LABORATORY

Frontend

- Monitors jobs to see how many Glideins are needed
- Compares requests with available entries (computing resources)
  - Knows about policies and expected resources
- Requests Glideins from the Factory
- Requests to a Factory to kill Glideins if there are too many
- Pressure-based system
  - Works keeping a certain number of Glideins running or idle on the resources
  - Glidein requests are gradual to avoid spikes and overloads
- Manages credentials and delegates them to the Factory.

Why Apptainer/Singularity

- Unprivileged
- Low overhead
- Single-file images and expanded images
- Nested containers

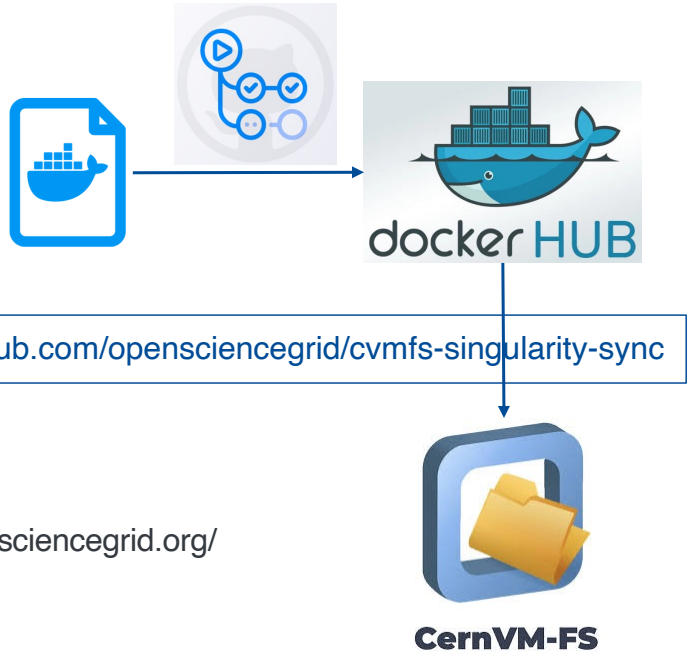
- Started using Singularity in 2017

How a Glidein uses Apptainer

- Site and experiment negotiation
  - Required, optional, never
  - Provisioning only on compatible resources
- Find the runtime executable
  - HTCondor distribution, CVMFS package, system install
- Determine the image
  - Image dictionary, local file, CVMFS expanded image, registry
  - Site or experiment constrains
- Glidein shell scripts with support for the services above and Apptainer invocation
  - Services, test scripts, user jobs

CernVM-FS expanded images

- CernVM-FS (CVMFS): Write once, read everywhere HTTP-based distributed file system
  - Provides deduplication
  - Multi-tiered structure
  - Uses HTTP and HTTP caching to replicate
- SIF images can be unpacked on a File System
  - Access only part of the image
- Automatic deployment of images on CVMFS
  - Dockerfile on GitHub
  - Register on the OSG repository
  - Workflow to build and push to Docker Hub
  - Pulling from Docker Hub and publishing on /cvmfs/singularity.opensciencegrid.org/



Distributing Fermilab Worker Node images

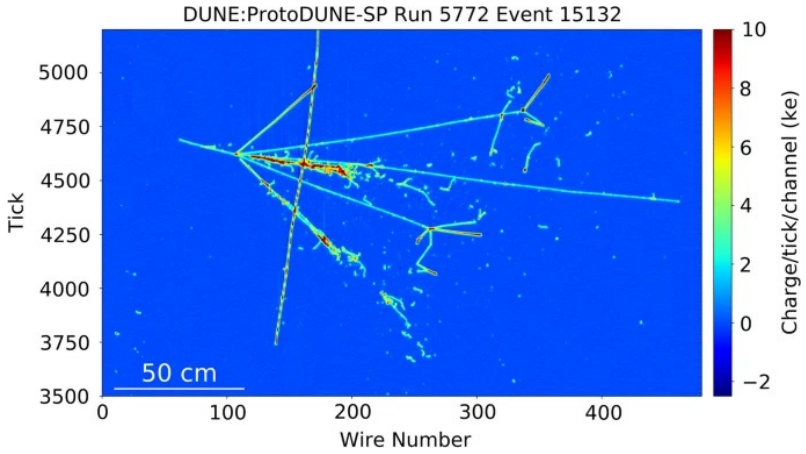
- Fermilab experiments like to run in an environments similar to the local cluster
  - Local cluster running Docker images on the worker nodes
- Experiments need also platforms that are obsolete
- Worker node images available on CVMFS:
  - Scientific Linux 6
  - Scientific Linux 7
  - Alma Linux 8
  - Alma Linux 9

NVIDIA Triton Server on Perlmutter nodes

HPSF CONFERENCE 2025

- Based on DUNE GPUaaS studies
- ProtoDUNE tests using local GPUs on Perlmutter nodes
  - The Glidein starts the Triton inference server on a container
  - All the jobs running on that node use the client to accelerate processing
  - No latency since running on the same node
  - 5x to 10x speedup compared to CPU only

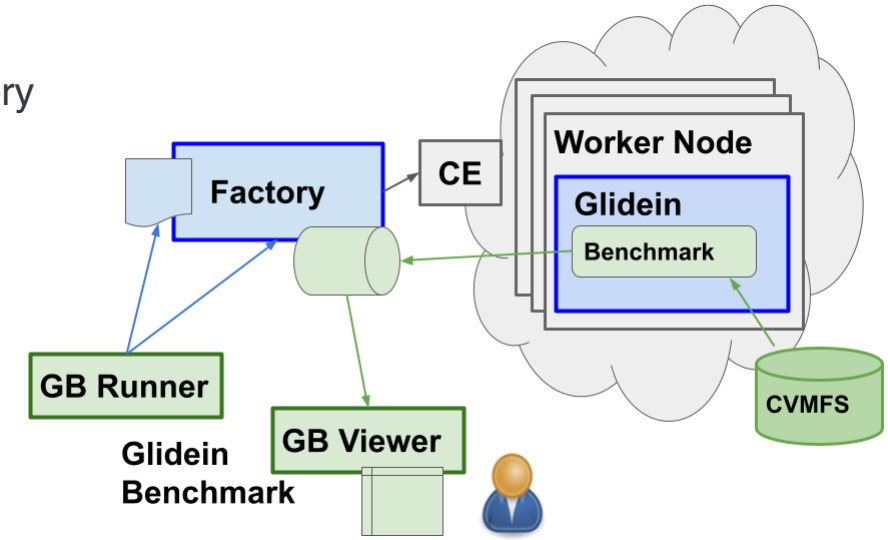
K.Herner et al, CCE-IOS Jun 2023



Resource Nodes Benchmarks

HPSF CONFERENCE 2025

- Benchmark results allow resource evaluation and better provisioning decisions
- Glideins run benchmarks and collect results
- Benchmarks running in containers
- Benchmarks images on CernVM-FS
- Benchmarks results stored at the Factory and available to clients

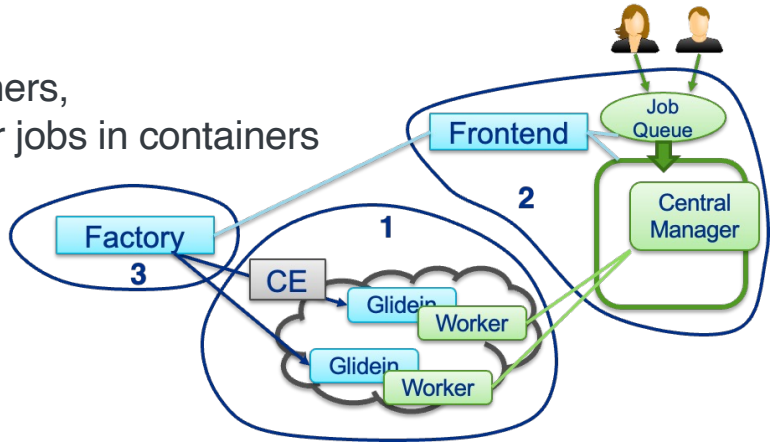


FERMI NATIONAL ACCELERATOR LABORATORY

Testing GlideinWMS with nested containers

HPSF CONFERENCE 2025

- A GlideinWMS is emulated running all resources in Podman containers
  - A one-node cluster, the CE
  - A Factory
  - A Frontend and submit node and central manager of the User Pool
- Used for GlideinWMS development and for Integration Tests
  - Deployment from Git, controlled by IDE like VSCode
  - Quasi-automatic test of new releases
- Thanks to the ability to run nested Apptainer containers, can run the regular workflow, including running User jobs in containers



GlideinWMS and Apptainer

HPSF CONFERENCE 2025

- GlideinWMS is a pilot-based resource provisioner for HTC workloads
- Uses Apptainer to run resources validation and jobs
  - Unprivileged
  - Low overhead
  - Nested containers
  - Single-file images and expanded images
- Notable use cases
  - Providing on distributed resources familiar or unsupported platforms
  - GPUaaS on a node using a containerized Triton server
  - Leverage expanded images to provide benchmarks

FERMI NATIONAL ACCELERATOR LABORATORY

Acknowledgement

This manuscript has been authored by FermiForward Discovery Group, LLC under Contract No. 89243024CSC000002 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.