

CMS Glidein Factory - Bootstrapping a Condor Pool Spanning Computing Sites Around the Globe

Amjad Kotobi

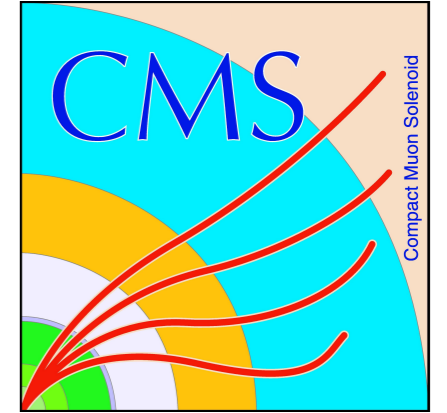
On behalf of the CMS Submission Infrastructure Group

and

OSG Factory Operations Team

June 2017

CMS Submission Infrastructure



Global Pool

There are two different types of jobs in pool:

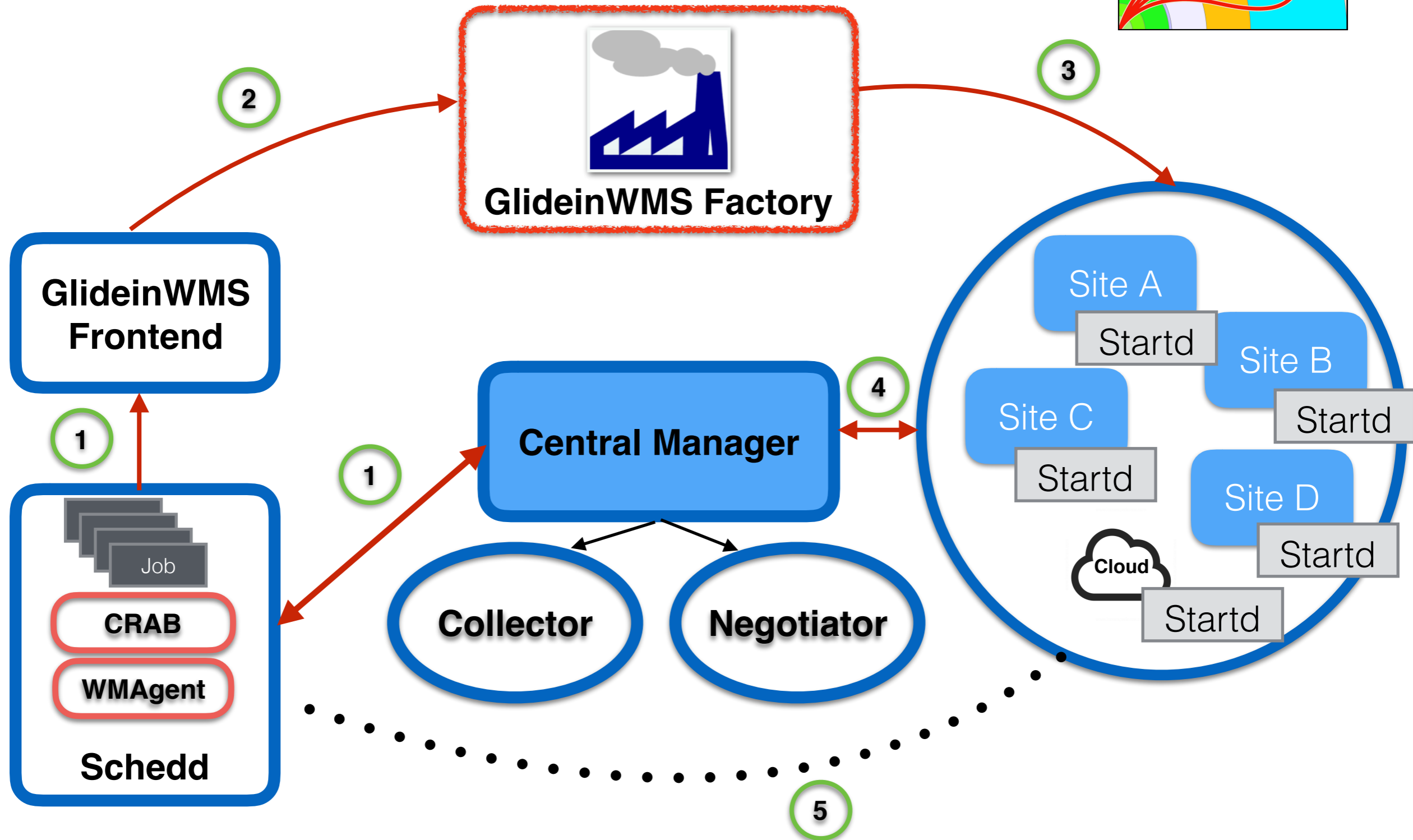
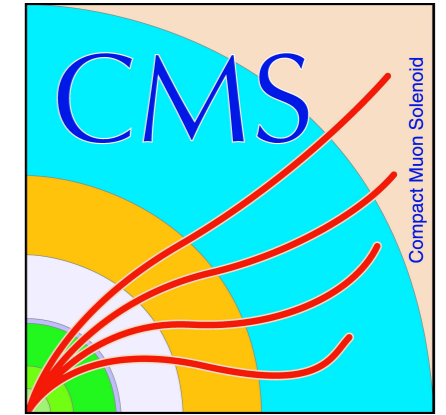
1. CMS Remote Analysis Builder (CRAB3)
2. Central production (WMAgent)

Global pool was created in 2014 for flexibility to use entire CMS resources for different kind of workflows.

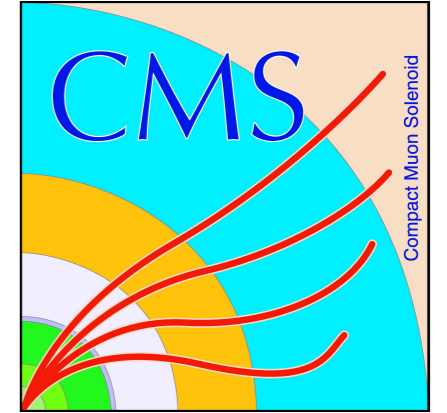
Submission infrastructure is pilot based with **two** main components:

1. HTCondor Pool
2. GlideinWMS (Glidein based workload management system, e.g Frontend, **Factory**)

CMS Submission Infrastructure

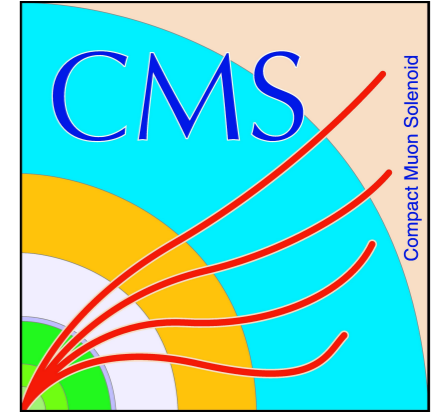


GlideinWMS Components



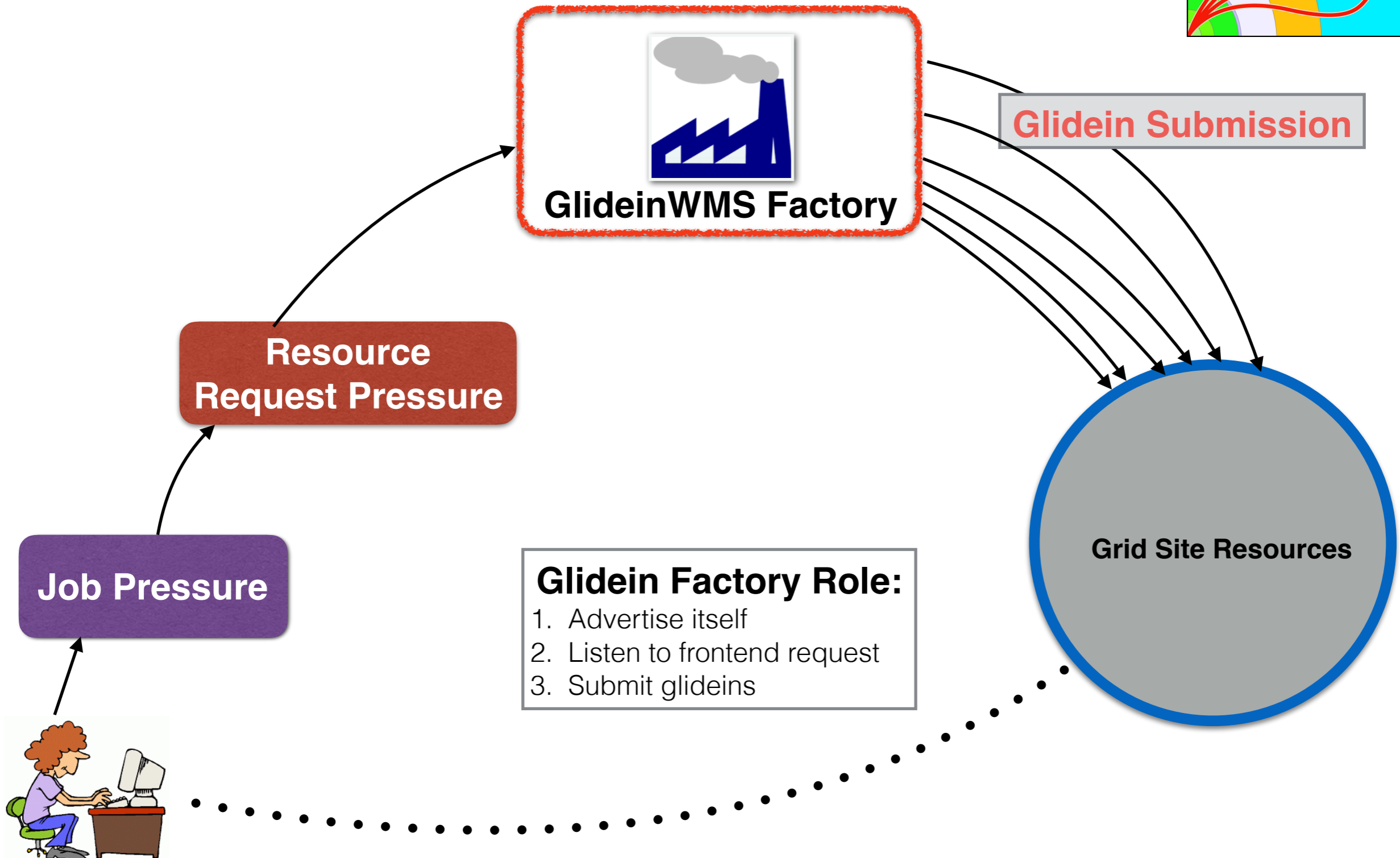
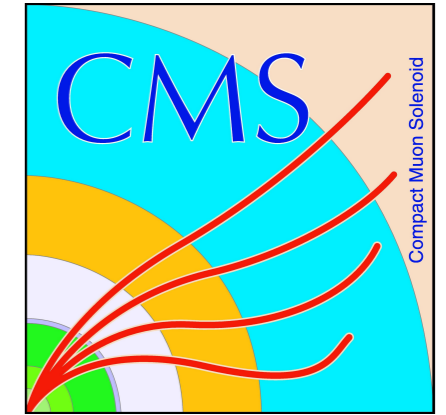
- **Glidein:** starts condor startd on the grid site.
- **Frontend:** Polls user jobs and make sure there are enough glideins for users job and make resource request to glidein factory.
- **Factory:** Receives request from frontend to submit glideins to grid sites.
- **WMS Collector:** A condor collector, keeps factory entry ClassAds and frontend request ClassAds.

First Stage Matchmaking

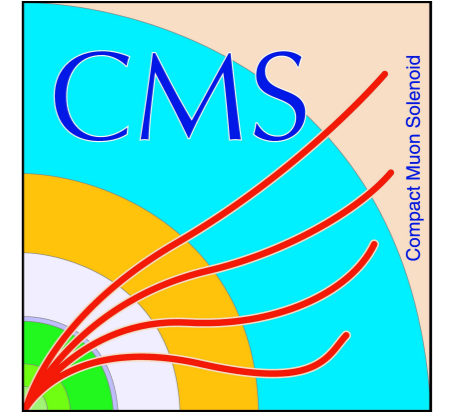


- Factory advertises entries ClassAds to the WMS collector.
- Factory has each entry description which advertises so glidein able to land on particular grid resources.
- Frontend check WMS collector to exert match expression against entry and user jobs.
- Match expression comes from frontend config and VO decides to where send user jobs to run.
- If there is no glidein so frontend will make request for more submission.

GlideinWMS Factory Role

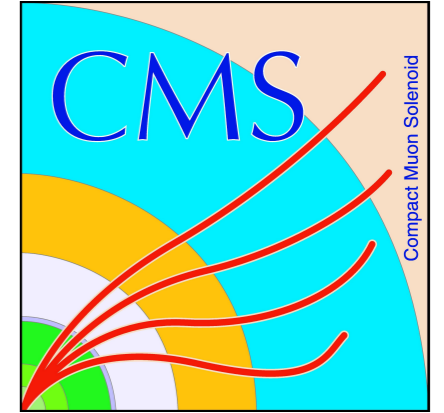


What is GlideinWMS Factory



- Glidein factory actually is a workload management system (**WMS**), uses pilot submission model to send jobs to grid resources.
- GlideinWMS factory works on top of HTCondor and heavily dependent on it, factory plays as schedds for pilot jobs.
- Glidens are actually pilot only glideinWMS calls it pilot.
- Pilot is an actual grid job and HTCondor job by itself that after reach to grid site calls real user jobs, in other words glideins are placeholder on remote resources.

Schematic View of GlideinWMS Factory



Collector

Frontend Req

Factory ClassAds

1

GlideinWMS Factory

Grid Sites

Site A

Gate Keeper

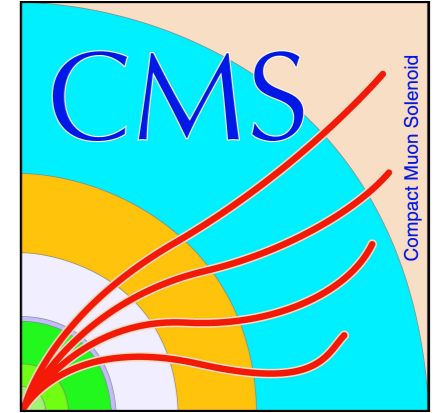
Site B

Gate Keeper

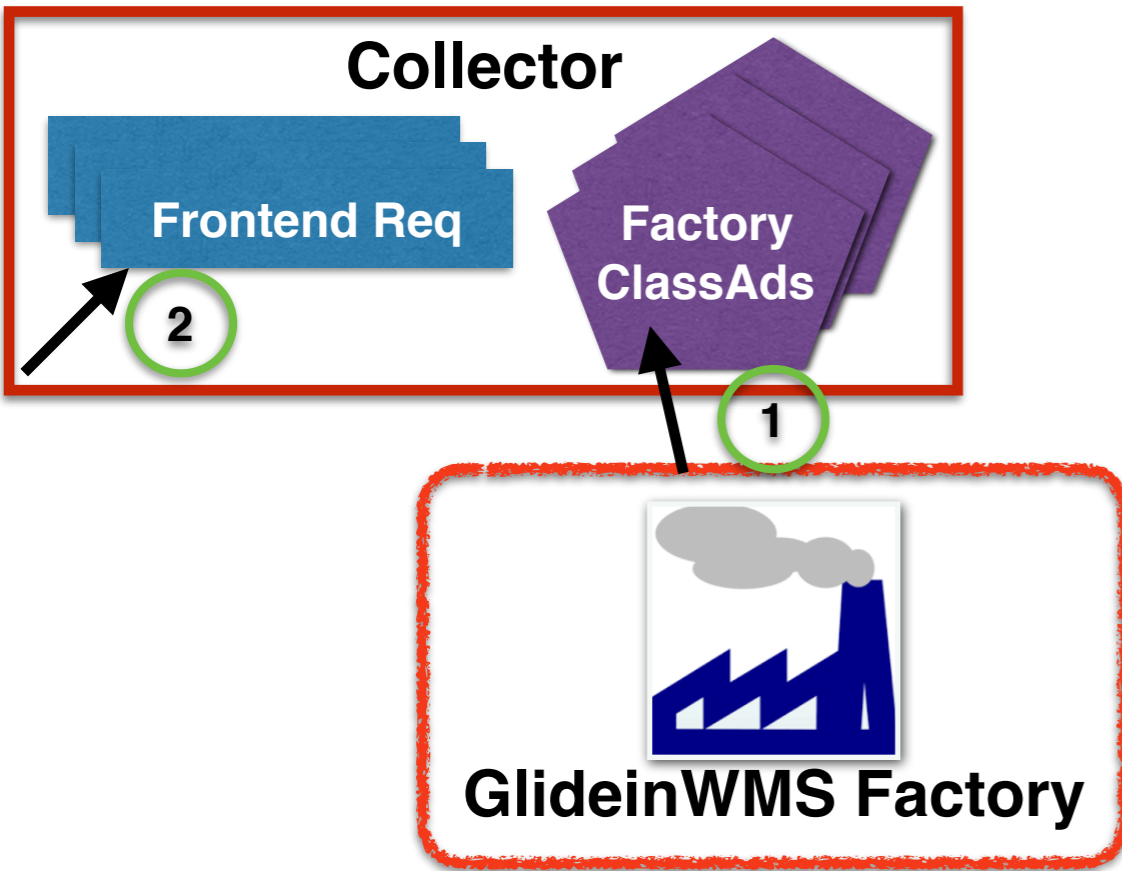
Glidein

Schedd

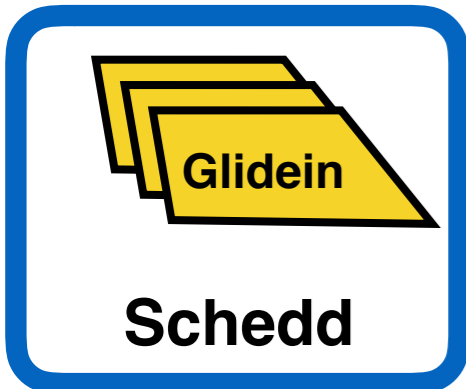
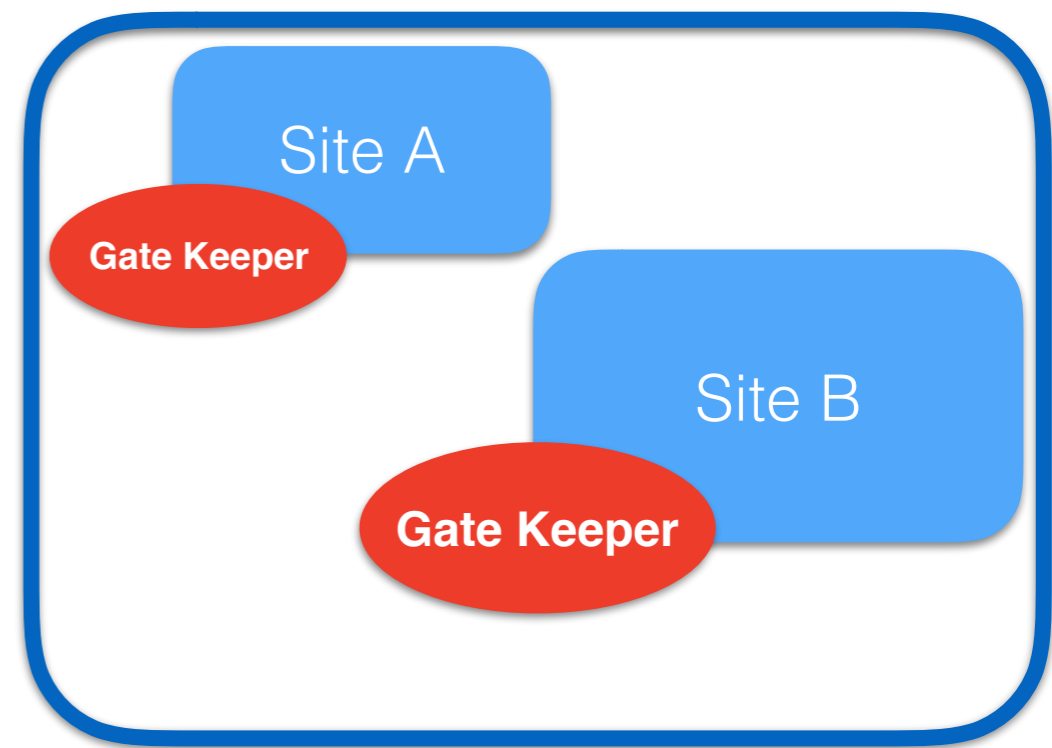
Schematic View of GlideinWMS Factory



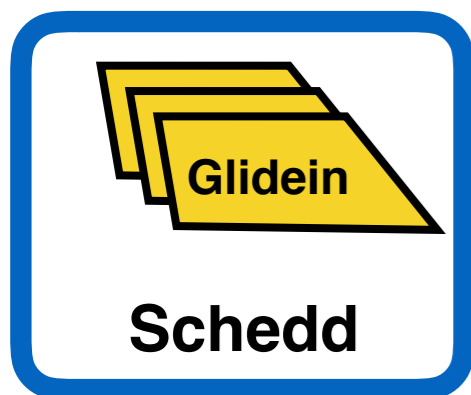
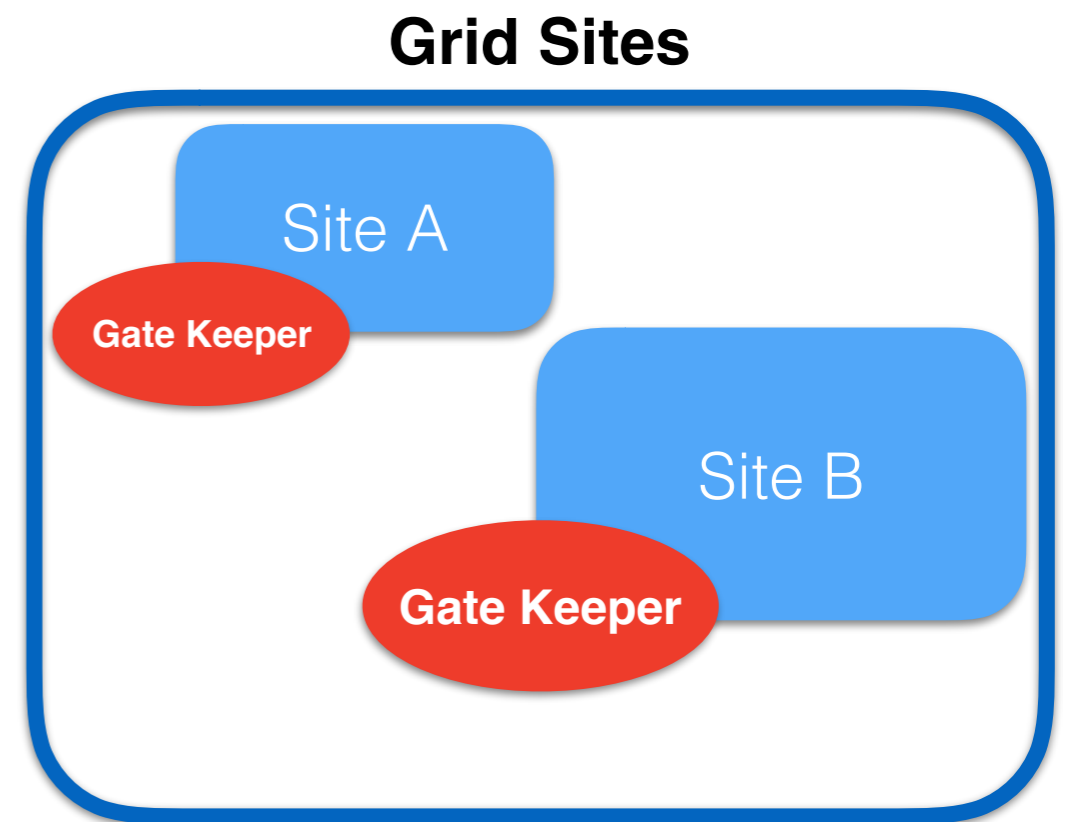
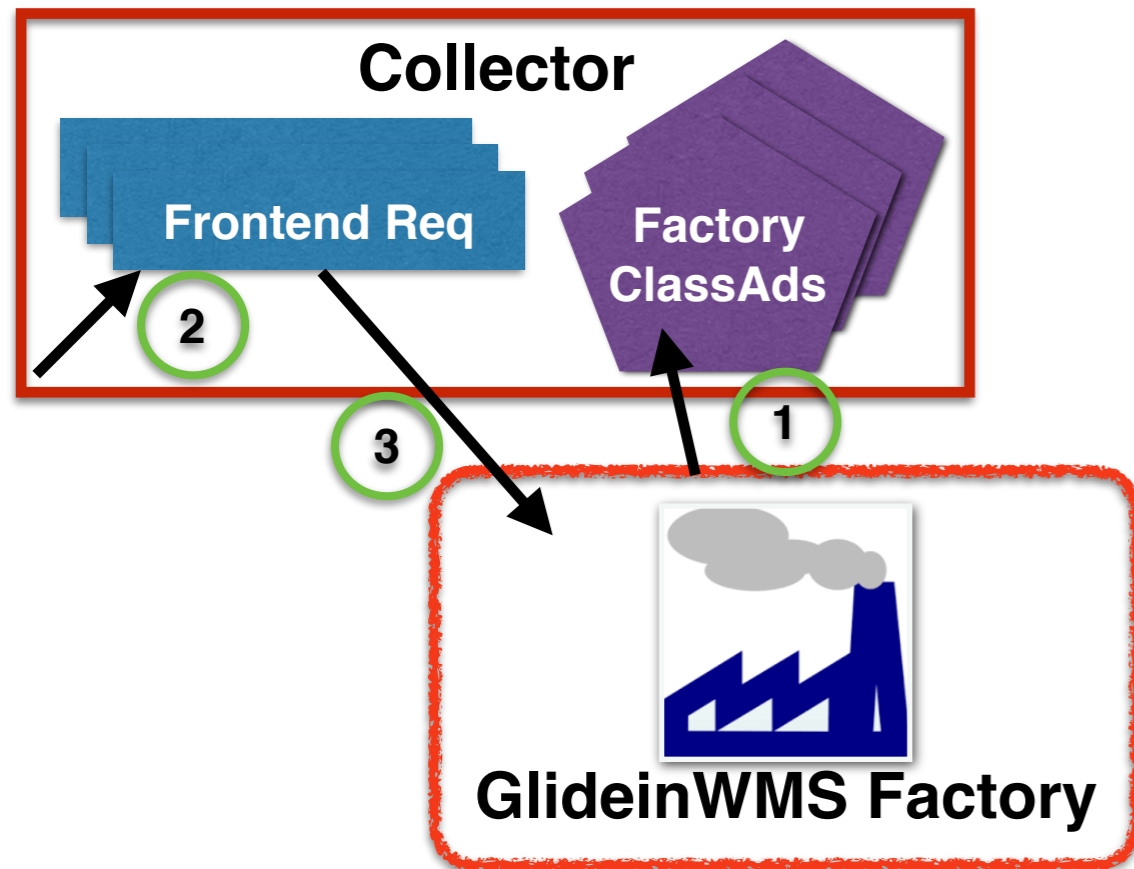
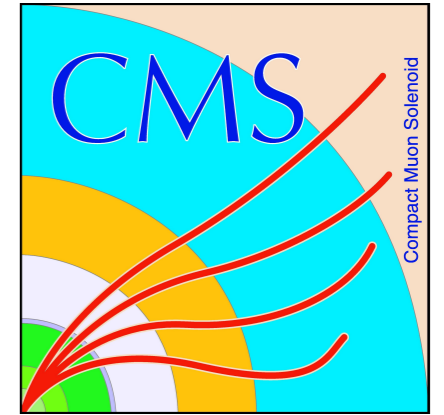
Collector



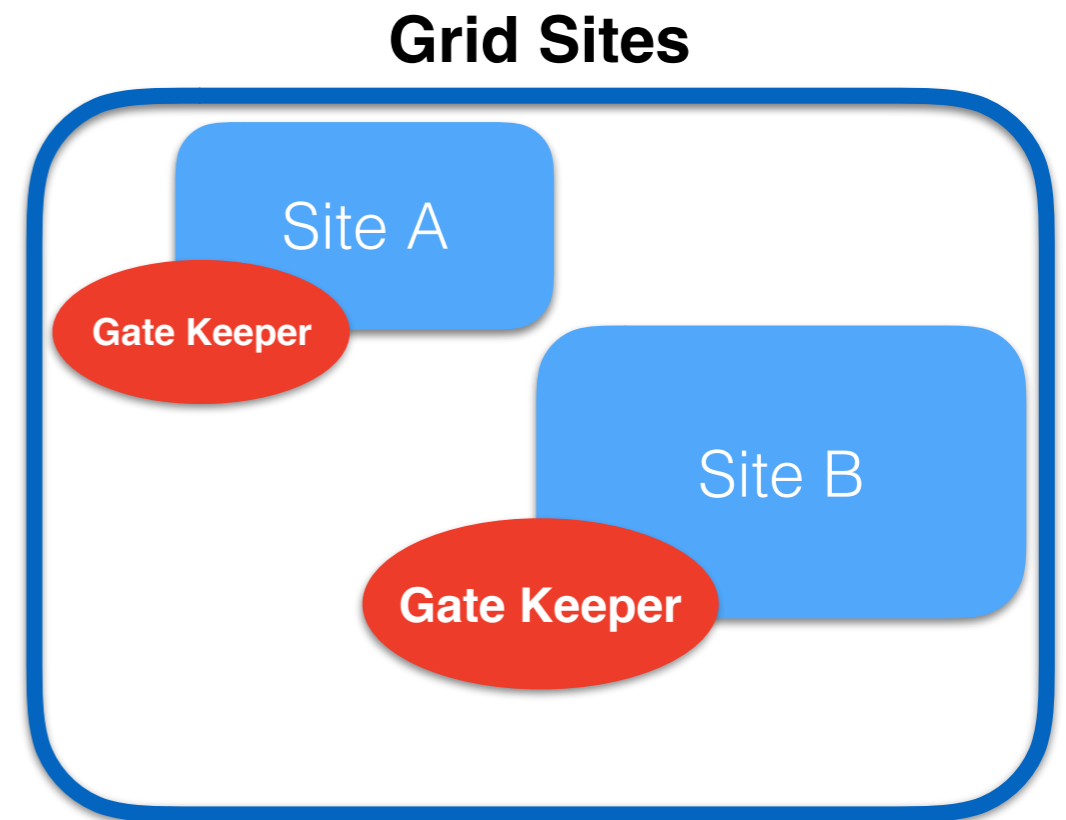
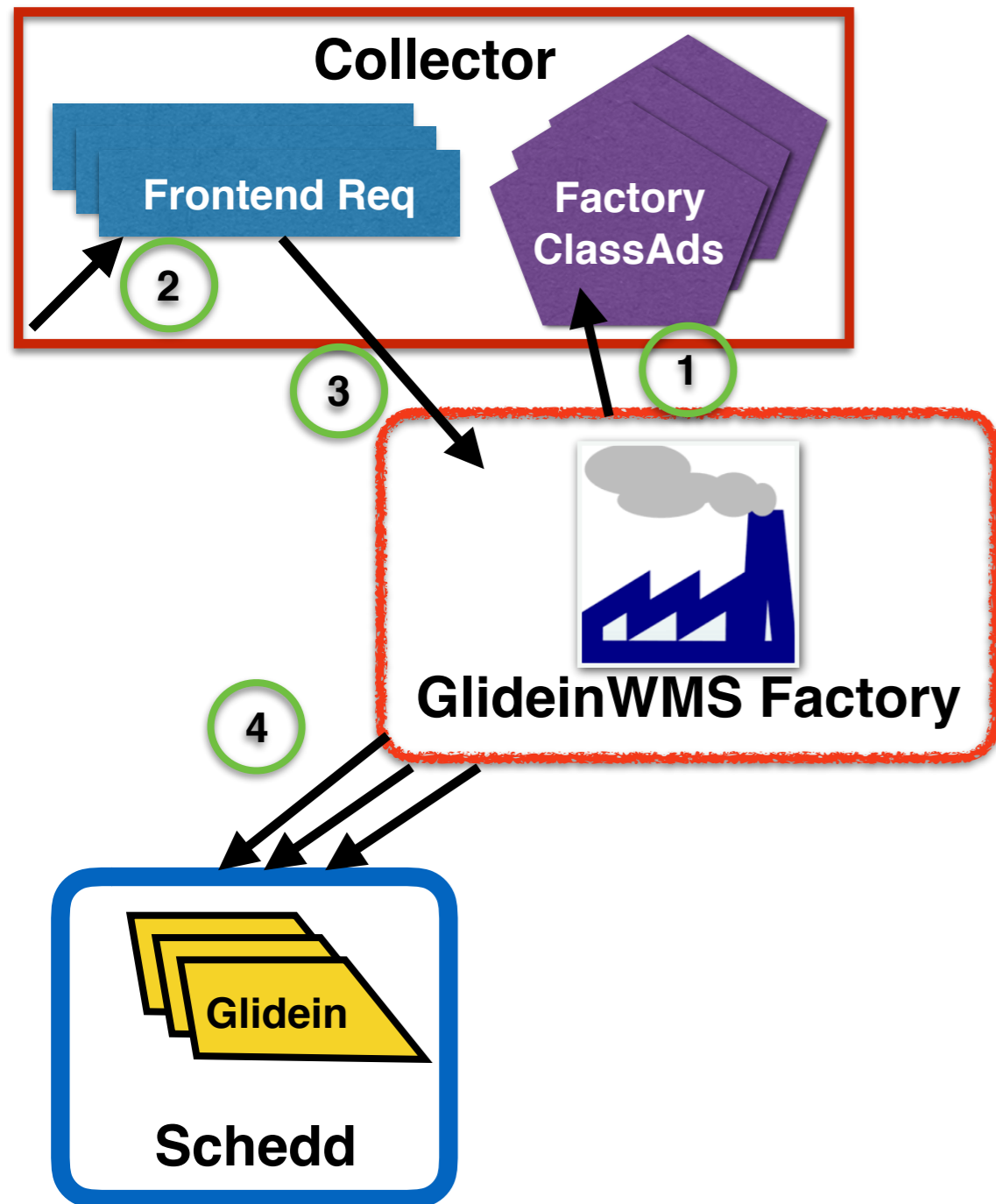
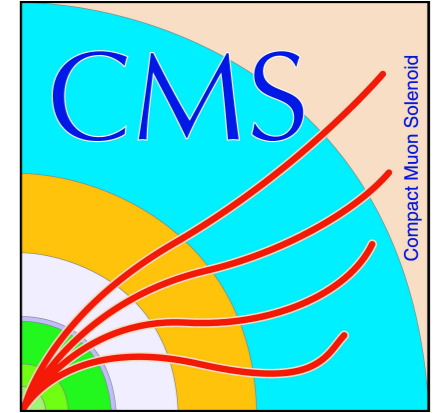
Grid Sites



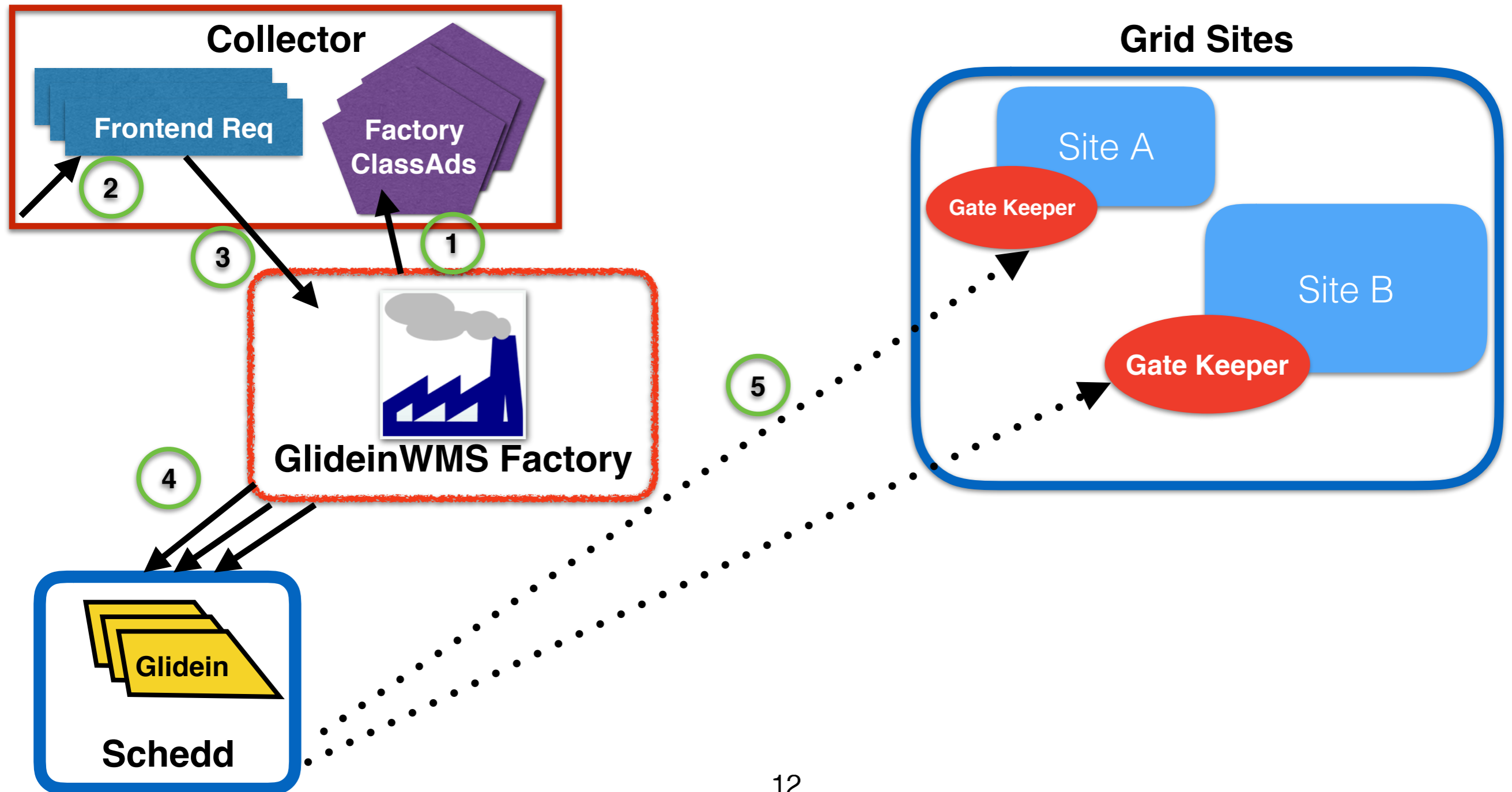
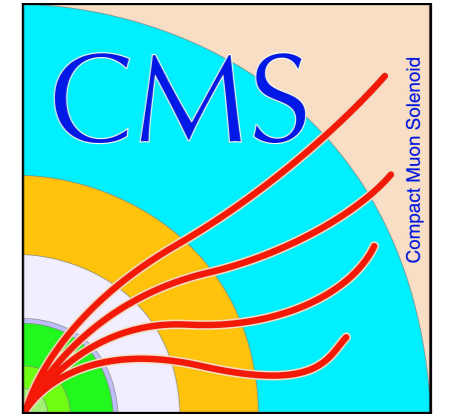
Schematic View of GlideinWMS Factory



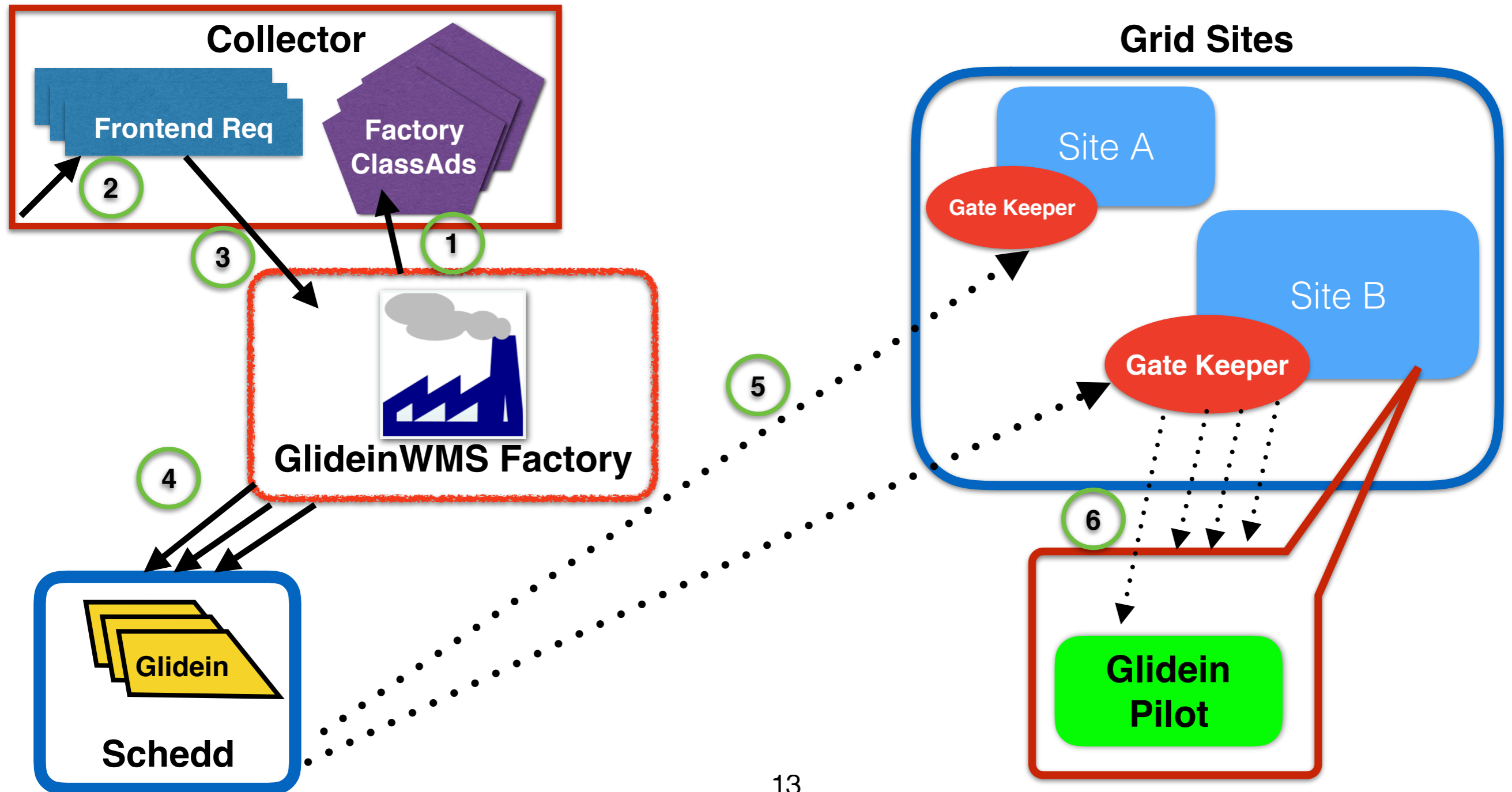
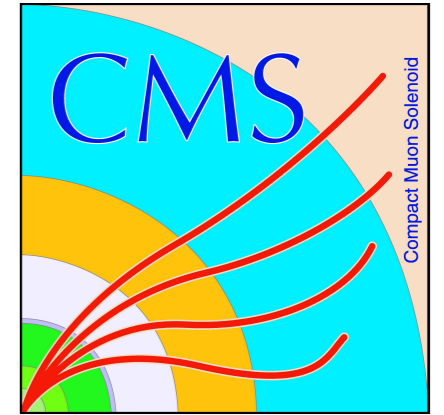
Schematic View of GlideinWMS Factory



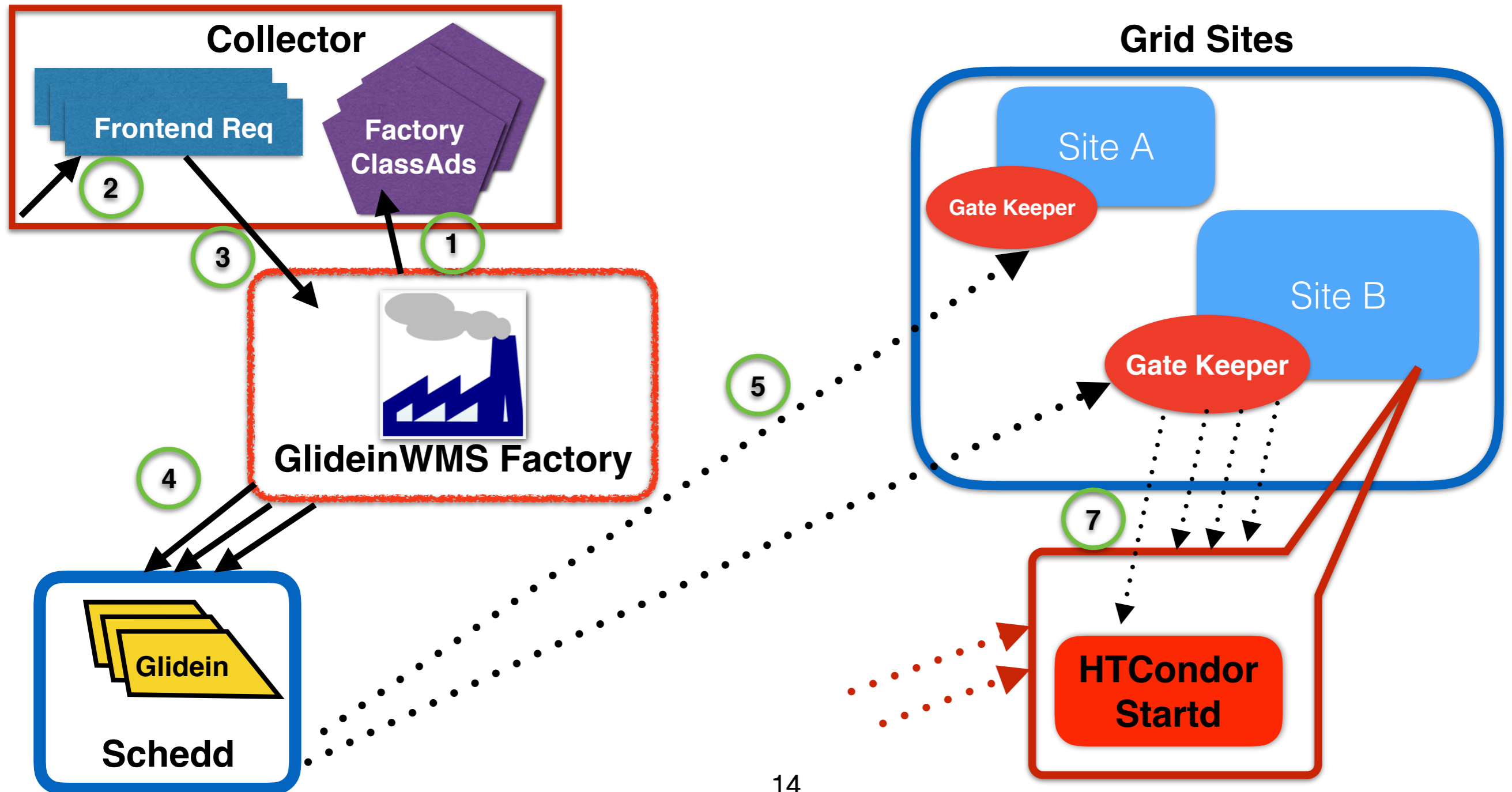
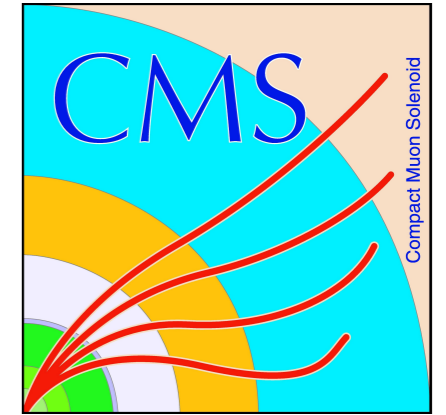
Schematic View of GlideinWMS Factory



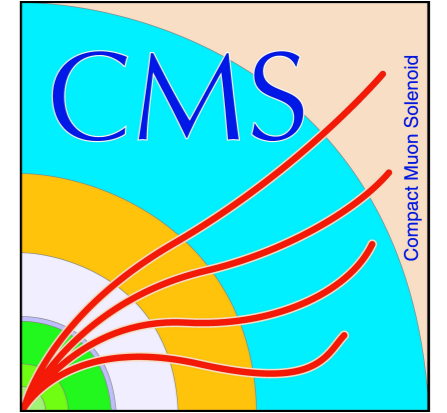
Schematic View of GlideinWMS Factory



Schematic View of GlideinWMS Factory

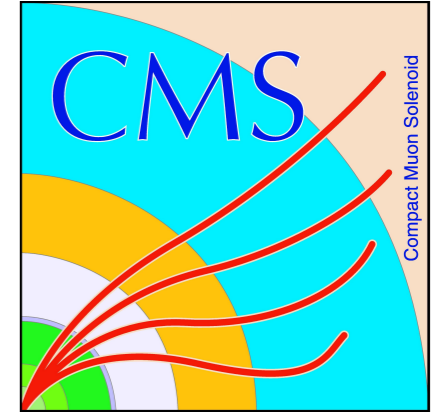


Second Stage Matchmaking



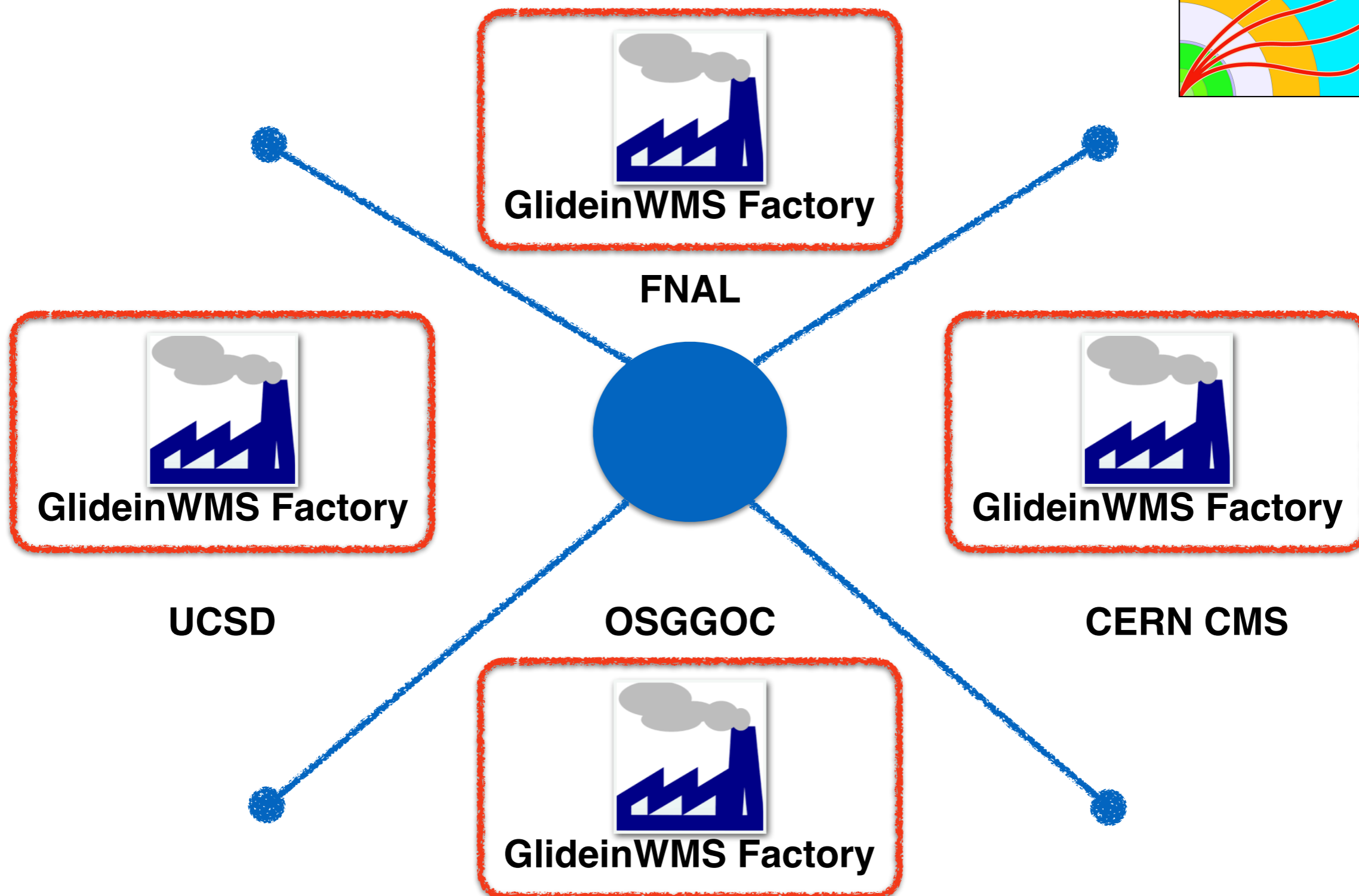
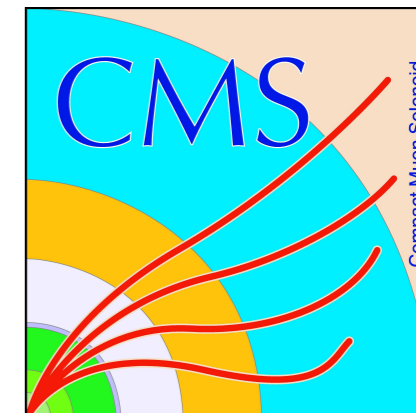
- **After** pilots/glideins start running on grid sites resources
- **Then** launched startds get connected back to the global pool HTCondor
- **Finally,** the negotiator matches job requests to the advertised available resources.

Importance of Factory to CMS

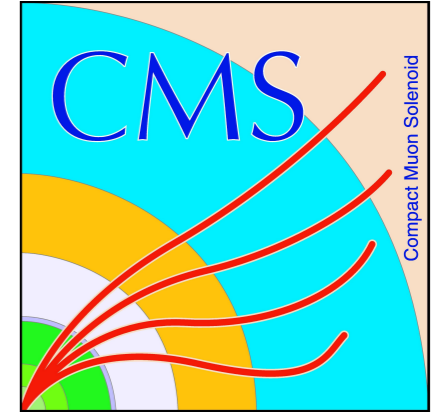


- Glideins run validation scripts to make sure environment suits users job
- Users job never starts in broken worker node and will find another match glideins
- Glideins play a role of placeholder to reserve resources for users job
- Glideins do not tie to a single user job during its lifetime
- Factory holds grid sites configurations and less trouble on site admin side.
- Factory can serve several frontends e.g ITB and Global
- Submitting jobs get simpler and resource shows as condor pool

CMS Support Factories

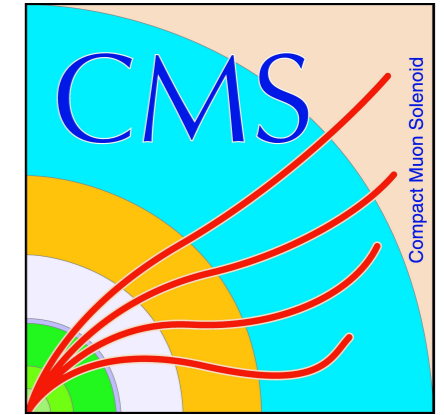


GlideinWMS Factory

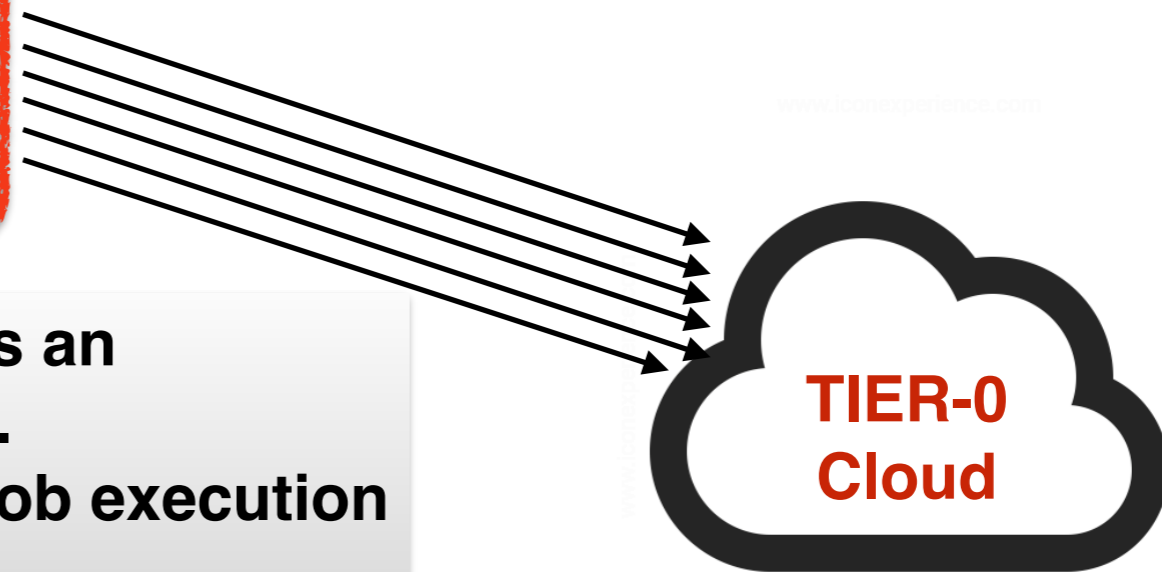


- Each factory has CMS global frontend to get resource request pressure
- Each factory has **nine** schedds
- Factory has collector
- Factory can send single and multi-core glideins to grid site
- Redundancy of factories
- Factory submit jobs to grid sites by **Condor-G**
- Each factory can support many frontends VOs

Factory Submission to Non-Traditional Resources

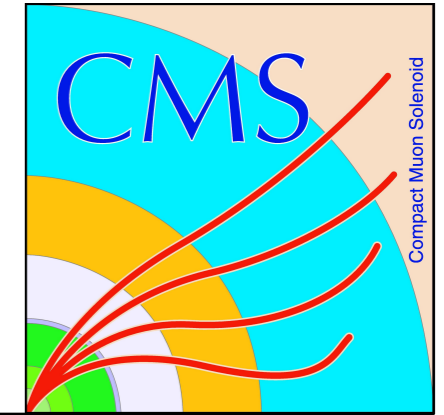


- Factory covers glideins submission interface to non-traditional resource e.g Tier-0 cloud and NERSC



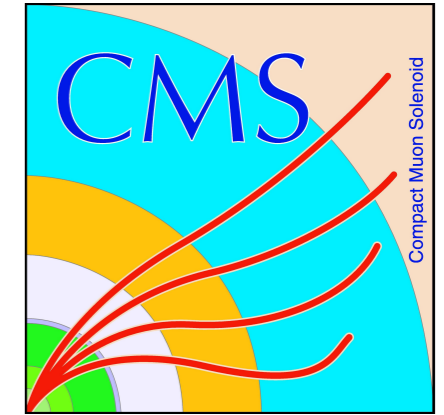
- CERN agile infrastructure (CERN AI) is an Openstack CERN cloud for resources.
- HTCondor and GlideinWMS used for job execution backbone.
- EC2/GlideinWMS interface make it feasible the pilot itself starts the VM and integrated in the image.

Example Factory Entry of Grid Site



<i>Attributes:</i>		<i>Descript:</i>	
FactoryType	production	AllowedVOs	
GCB_ORDER	NONE	AuthMethod	grid_proxy
GLEEXEC_BIN	glite	DefaultPerFrontendMaxGlideins	50
GLIDEIN_CMSSite	T1_DE_KIT	DefaultPerFrontendMaxHeld	20
GLIDEIN_CPUS	8	DefaultPerFrontendMaxIdle	20
GLIDEIN_Country	DE	Gatekeeper	arc-6-kit.gridka.de
GLIDEIN_Gatekeeper	arc-6-kit.gridka.de	GlobusRSL	(queue=grid)(count=8)(memory=2500)(runtimeenvironment=ENV/GLITE)
GLIDEIN_GlobusRSL	(queue=grid)(count=8)(memory=2500)(runtimeenvironment=ENV/GLITE)	GridType	nordugrid
GLIDEIN_GridType	nordugrid	MaxReleaseRate	20
GLIDEIN_MaxMemMBs	20240	MaxRemoveRate	5
GLIDEIN_Max_Walltime	216000	MaxSubmitRate	10
GLIDEIN_REQUIRED_OS	rhel6	PerEntryMaxGlideins	50
GLIDEIN_REQUIRE_GLEEXEC_USE	False	PerEntryMaxHeld	20
GLIDEIN_REQUIRE_VOMS	False	PerEntryMaxIdle	20
GLIDEIN_Req_MUPJ_gLExec	False	PerFrontendMaxGlideins	
GLIDEIN_ResourceName	FZK-LCG2	PerFrontendMaxHeld	
GLIDEIN_Retire_Time	108000	PerFrontendMaxIdle	
GLIDEIN_Retire_Time_Spread	7200	ReleaseSleep	0.2
GLIDEIN_SEs	cmssrm-kit.gridka.de	RemoveSleep	0.2
GLIDEIN_Site	KIT	RequireGlideinGlexecUse	False
GLIDEIN_SlotsLayout	fixed	RequireVomsProxy	False
GLIDEIN_SupportedAuthenticationMethod	grid_proxy	StartupDir	TMPDIR
GLIDEIN_Supported_VO	CMS	SubmitCluster	10
GLIDEIN_TrustDomain	grid	SubmitSleep	2
GLIDEIN_Verbosity	std	SubmitSlotsLayout	fixed
GLIDEIN_WorkDir	TMPDIR	TrustDomain	grid

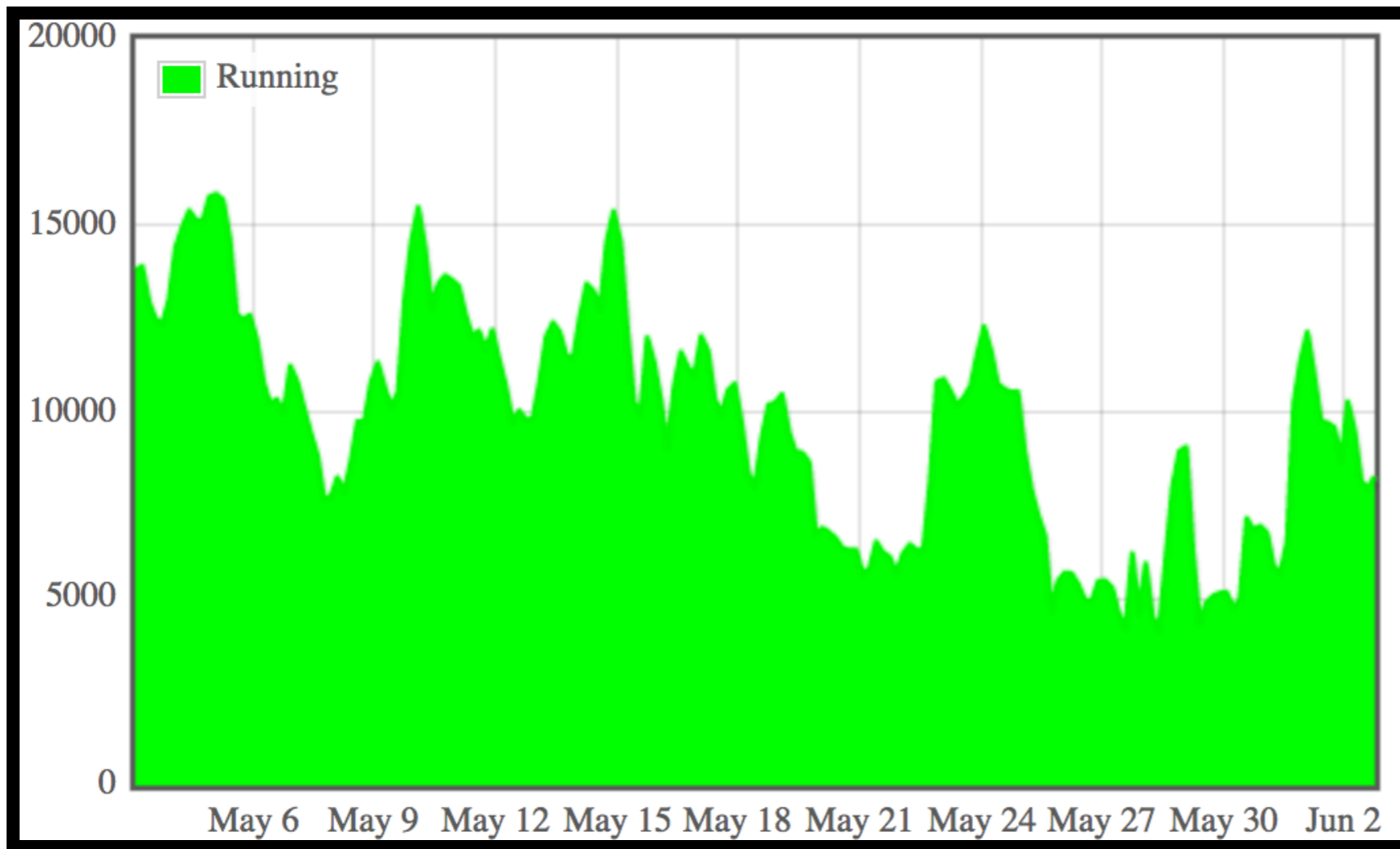
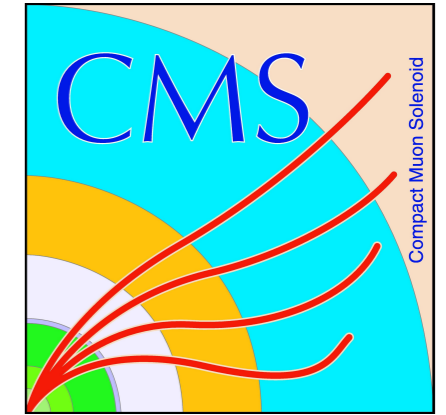
Grid Site Entries



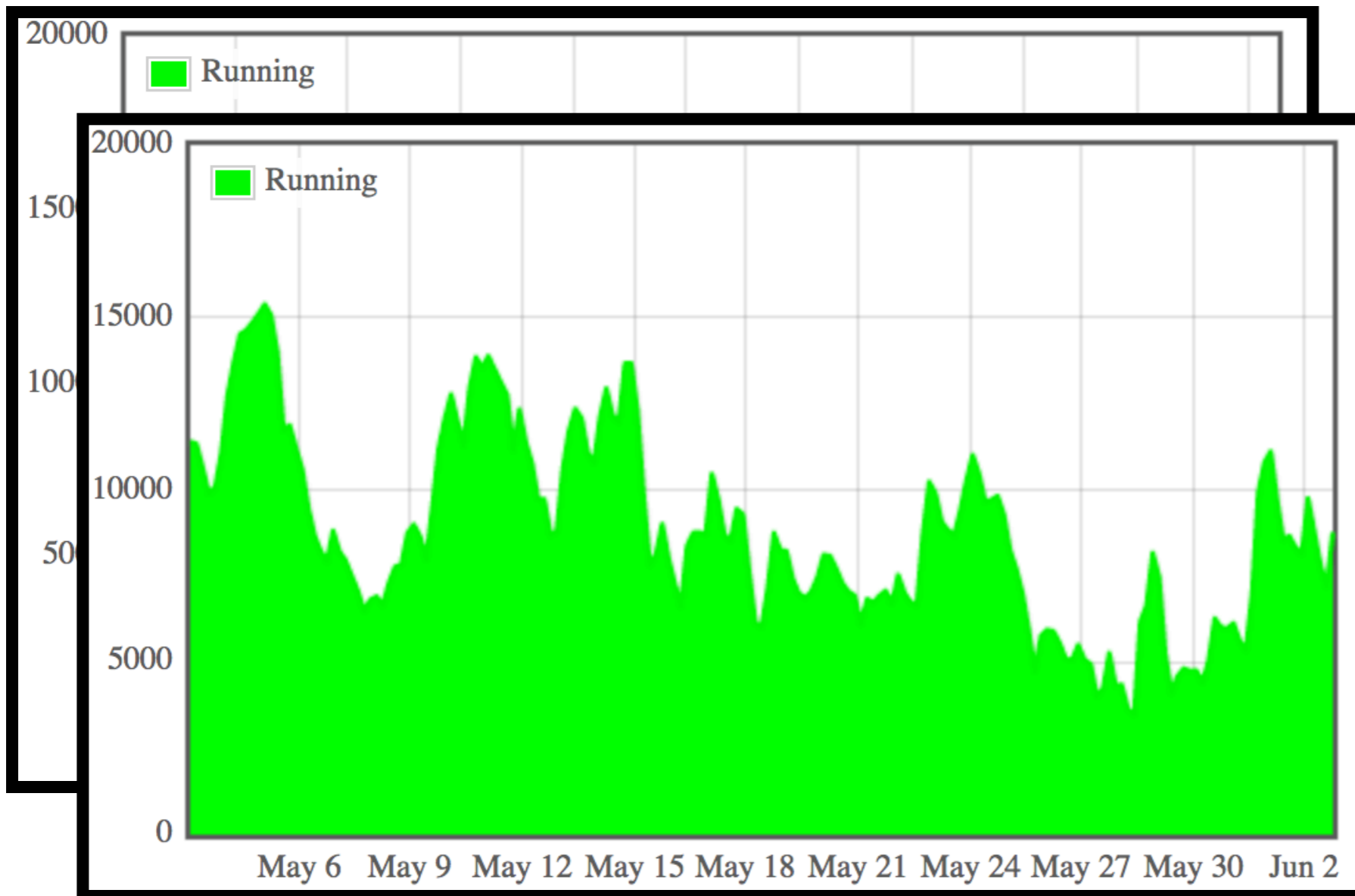
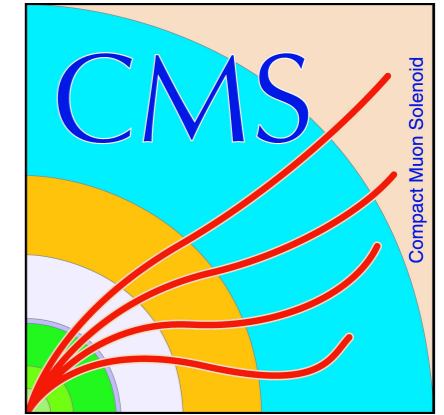
Grid sites may have many entries in factory with different gatekeepers or queues that connected to diverse resources

CMSHTPC_T1_DE_KIT_arc-1	↑	6	0	0	0	0	0	0	0
CMSHTPC_T1_DE_KIT_arc-2	↑	4	0	0	0	0	0	0	0
CMSHTPC_T1_DE_KIT_arc-3	↑	0	9	9	0	0	0	0	0
CMSHTPC_T1_DE_KIT_arc-4	↑	0	4	4	0	0	0	0	0
CMSHTPC_T1_DE_KIT_arc-5	↑	5	0	0	0	0	0	0	0
CMSHTPC_T1_DE_KIT_arc-6	↑	4	1	1	0	0	0	0	0

CMS Support Factory Usage



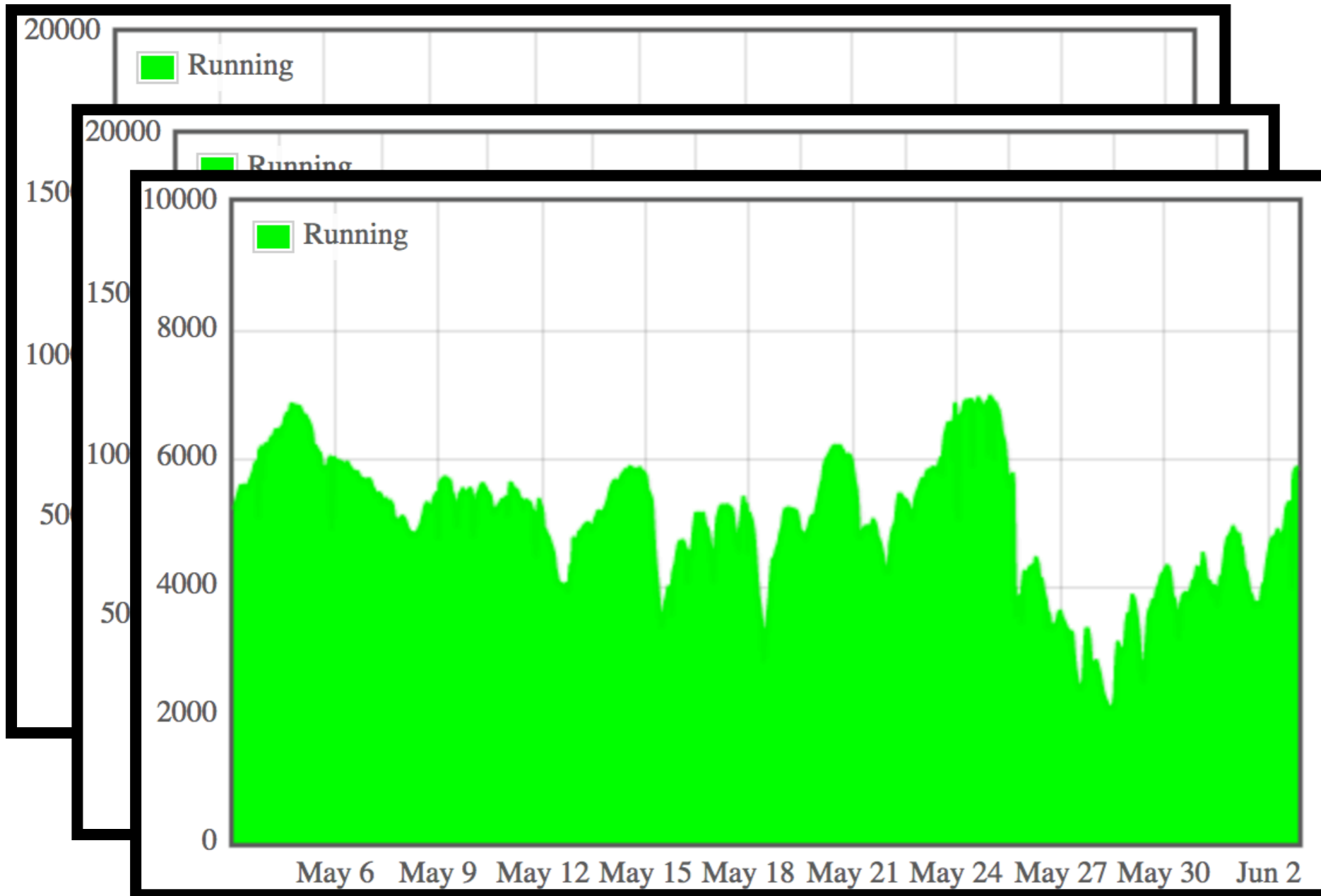
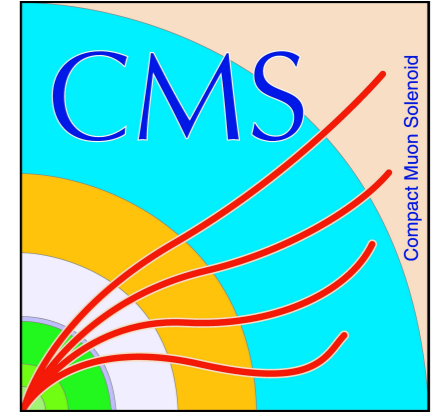
CMS Support Factory Usage



CERN

OSGGOC

CMS Support Factory Usage

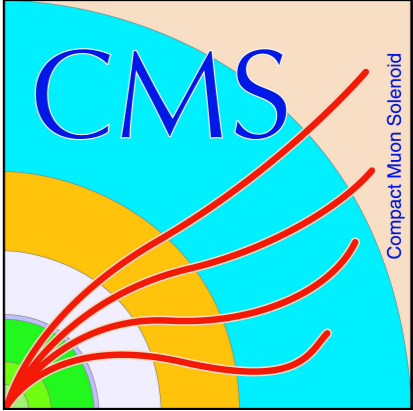


CERN

OSGGOC

UCSD

CMS Support Factory Usage



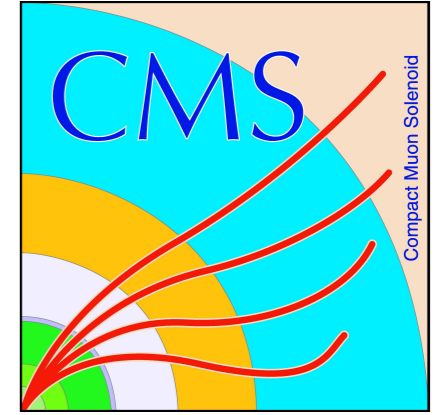
CERN

OSGGOC

UCSD

FNAL

Acknowledgements



- Special thanks to **HTCondor** developers for their support and suggestions
- **Factory Operations team**
 - Jeff Dost, Krista Larson, Marian Zvada, Marty Kandes
- **Submission infrastructure team leaders**
 - James Letts, Antonio Perez-Calero Yzquierdo, David Mason
- **GlideinWMS developers**
- **CMS Global Pool**
 - Diego Davila
- **Brian Bockelman, Farrukh Aftab Khan**