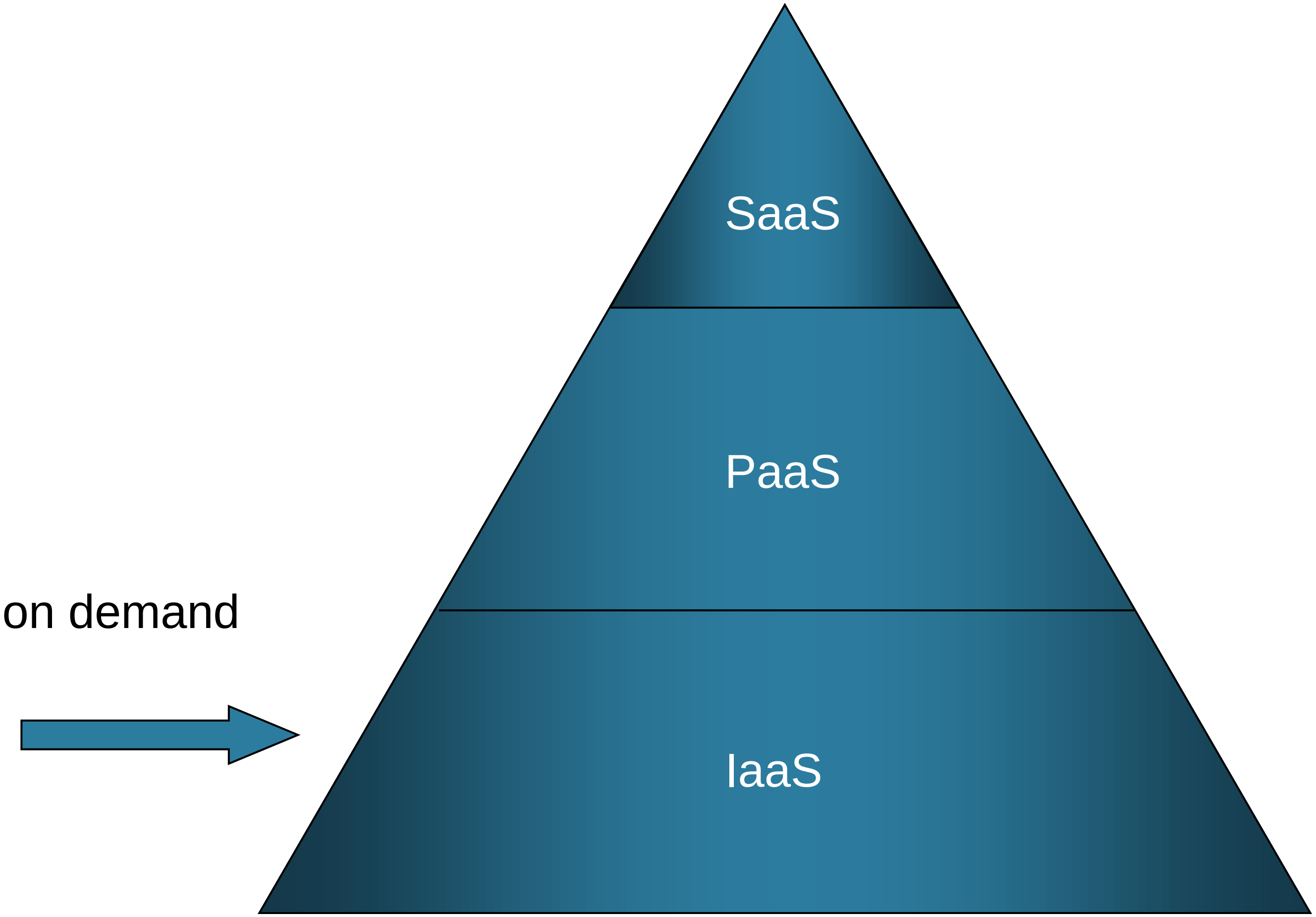
The Adoption of Cloud Technology within the LHC Experiments

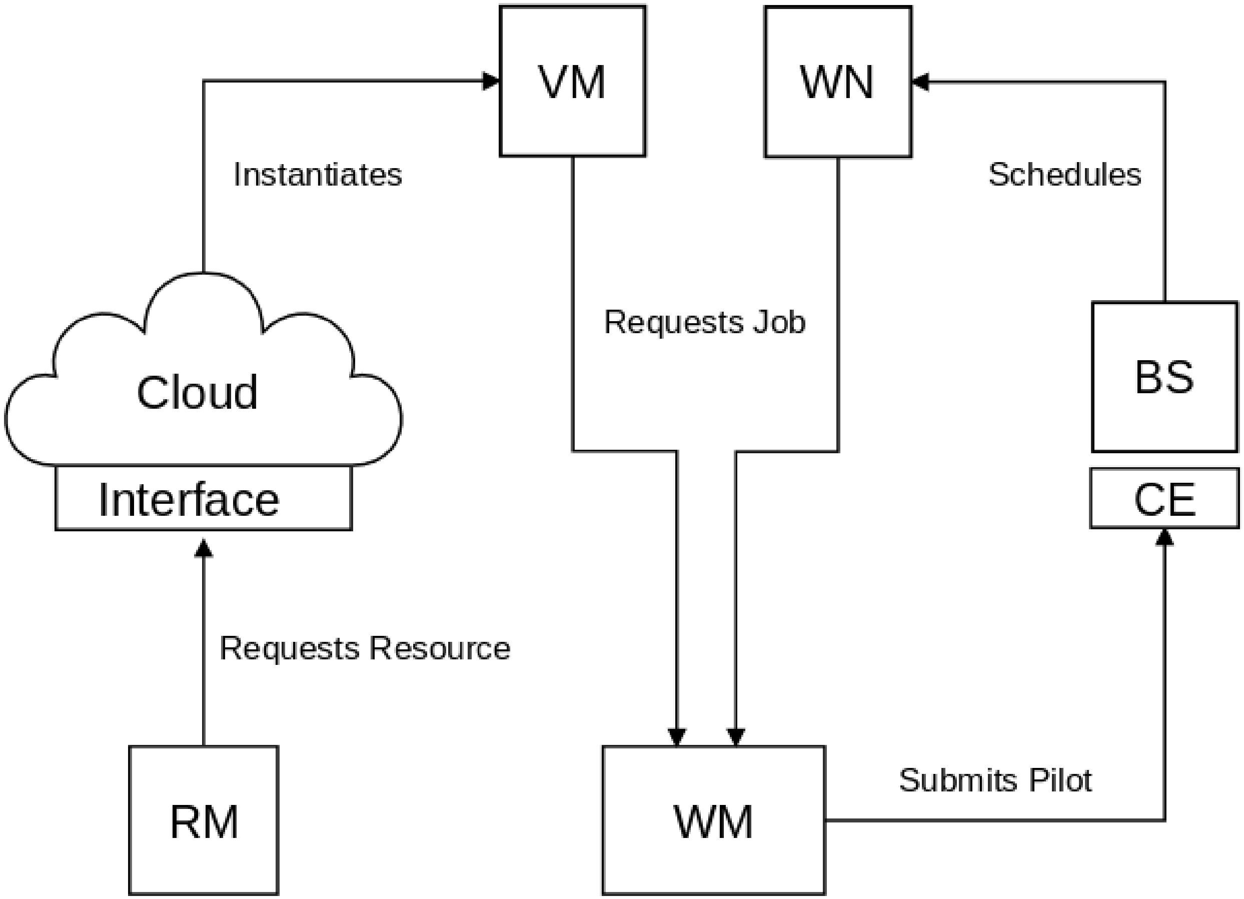
Laurence Field
IT/SDC

17/10/2014

Cloud



High Level View



Areas

- Image Management
- Capacity Management
- Monitoring
- Accounting
- Pilot Job Framework
- Data Access and Networking
- Quota Management
- Supporting Services

Image Management

- Provides the job environment
  - Software
    - CVMFS
    - PilotJob
  - Configuration
  - Contextualization
- Balance pre- and post-instantiation operations
  - Simplicity, Complexity, Data Transfer, Frequency of Updates
- Transient
  - No updates of running machines
    - Destroy (gracefully) and create new instance

CernVM

- The OS via CVMFS
  - Replica via HTTP a reference file system
    - Stratum 0
- Why?
  - Because CVMFS is already a requirement
    - Removes the overhead of distributed image management
      - Manage version control centrally
- CernVM as a common requirement
  - Availability becomes and infrastructure issue
  - Recipe to contextualize
    - Responsibility of the VO
- The goal is to start a CernVM-based instance
  - Which needs minimal contextualization

Capacity Management

- Managing the VM life cycle isn't the focus
  - It is about ensuring the is enough resources (capacity)
- Requires a specific component with some intelligence
  - Do I need to start of VM and if so where?
  - Do I need to stop a VM and if so where?
  - Are the VMs that I started OK?
- Existing solutions focus on deploying applications in the cloud
  - Difference components, one cloud
  - May managed load balancing and failover
    - Is this a load balancing problem?
  - One configuration, many places, enough instances?
- Developing our own solutions
  - Site centric
    - The VAC model
  - VO centric

Monitoring

- Fabric management
  - The responsibility of the VO
  - Basic monitoring is required
- The objective is to triage the machines
  - Invoke a restart operation if it not ok
    - Detection of the not ok state maybe non-trivial
- Other metrics may be of interest
- Spotting dark resources
  - Deployed but not usable
- Can help to identify issues in other systems
  - Discovering inconsistent information through cross-checks
- A Common for all VOs
  - Pilot jobs monitoring in VO specific

Provider Accounting

- Helix Nebula
  - Pathfinder project
    - Development and exploitation
      - Cloud Computing Infrastructure
  - Divided into supply and demand
  - Three flagship applications
    - CERN (ATLAS simulation)
    - EMBL
    - ESA
- **FW: New Invoice!**
  - *Can you please confirm that these are legit?*
  - Need to method to ***record*** usage to cross-check i
  - Dark resources
    - Billed for x machines but not delivered (controllable)

Consumer-Side Accounting

- Monitor resource usage
  - Course granularity acceptable
    - No need to accurately measure
- What, where, when for resources
  - Basic infrastructure level
    - VM instances and whatever else is billed for
- Report generation
  - Mirror invoices
    - Use same metrics as charged for
- Needs a uniform approach
  - Should work for all VOs
    - Deliver same information to the budget holder

Cloud Accounting in WLCG

- Sites are the suppliers
  - Site accounting generates invoices
    - For resources used
- Need to monitor the resource usage
  - We trust sites and hence their invoices
    - Comparison can detect issues and inefficiencies
- Job activities in the domain of the VO
  - Measurement of work done
    - i.e. value for money
  - Information not included in cloud accounting
    - Need a common approach to provide information
      - Dashboard?

Comparison to Grid

- Grid accounting = supply side accounting
- No CE or batch system
  - Different information source
    - No per job information available
- Only concerned about resources used
  - Mainly time-based
    - For a flavour
      - A specific composition of CPU, memory, disk

Core Metrics

- Time
  - Billed by time per flavour
- Capacity
  - How many were used?
- Power
  - Performance will differ by flavour
    - And potentially over time
- Total computing done
  - power x capacity x time
- Efficiency
  - How much computing did something useful

Measuring Computing

- Resources provided by flavour
    - How can we compare?
- What benchmarking metrics?
    - And how we obtain them
- Flavour = SLA
- SLA monitoring
    - How do we do this?
- Rating sites
    - Against the SLA
        - Variance

Data Mining Approach

- VOs already have metrics on completed jobs
  - Can they be used to define relative performance?
- Avoids dicussion on bechmarking
  - And how the benchmark compares to the job
- May work for within the VO
  - But what about WLCG?
- Specification for procurement?
  - May not accept a VO specific metric
- Approach currently being investigated

The Other Areas

- Data access and networking
    - Have so far focus on non-data intensive workloads
- Quota Management
    - Currently have fixed limits
        - Leading the partitioning of resources between VOs
            - How can the sharing of resources be implemented?
- Supporting Services
    - What else is required?
        - Eg squid caches in the provider
    - How are these managed and by who?

Status of Adoption

Alice

- Will gradually increase adoption
- Using CERN’s AI for release validation
  - Dynamic provisioning of an elastic virtual cluster
- Offline simulation jobs for the HLT farm
  - Set up as hypervisors for VMs
    - Suspend or even terminate when DAQ is required
  - Possibly will enable more I/O-intensive jobs
    - During LHC shutdown periods

Alice

- CERN Analysis Facility
  - Another cluster in CERN’s AI
    - Allows elastic scaling
    - Disk space needs served directly by EOS
- Using CernVM via the WebAPI portal
  - An upcoming outreach activity
  - Extend as volunteer computing platform
    - ALICE@home to support simulation activities

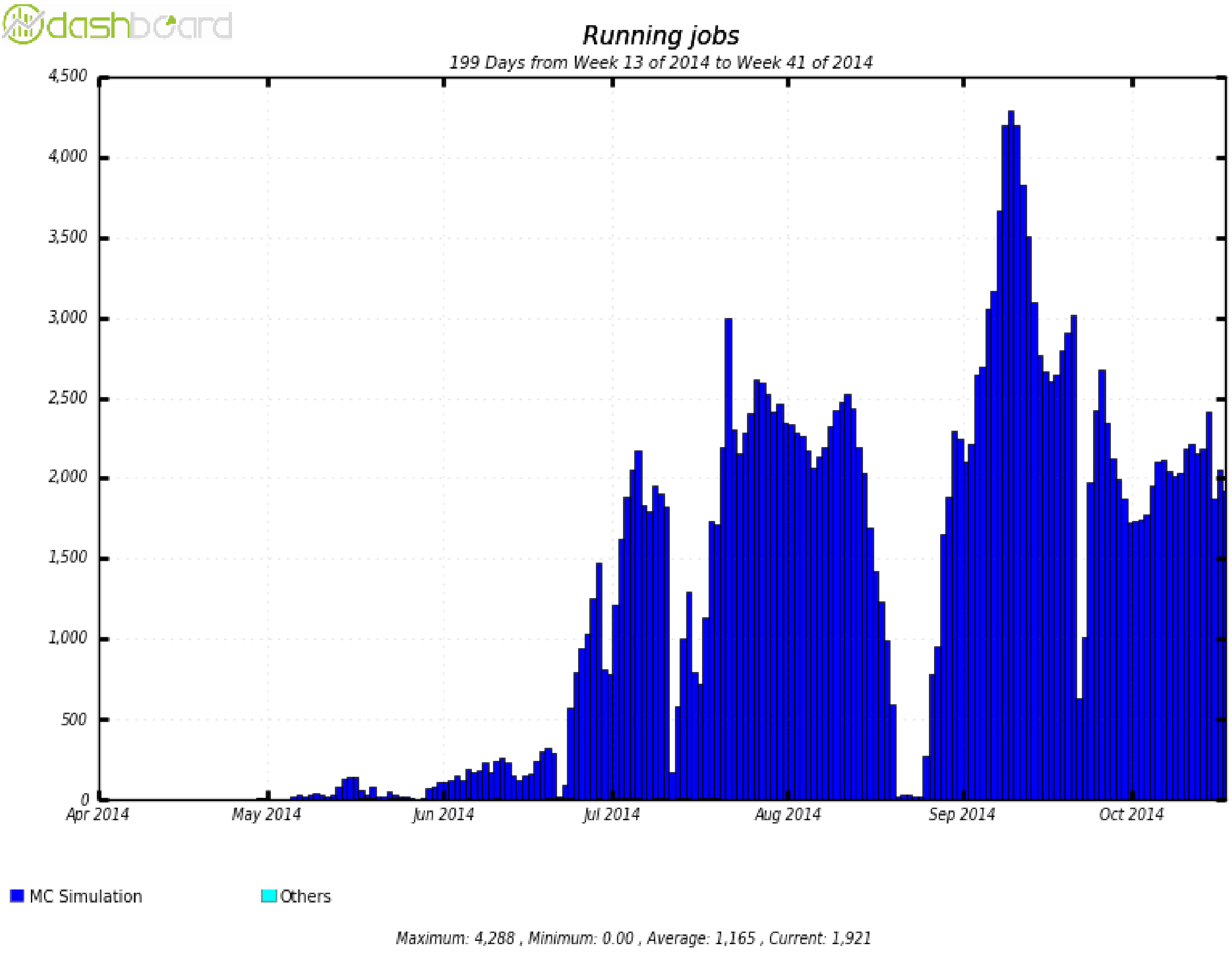
ATLAS

- HLT Farm
  - Virtualization is used to ensure isolation
    - For tasks that require external network access
  - OpenStack is used to provide an IaaS layer
- Distributed Computing
  - Existing Grid sites can also provide resource using their IaaS
  - Cloud Scheduler used for capacity management
  - VAC is also used by some sites in the UK
  - Also experimenting with VCycle
  - VM instances are monitored Ganglia service
    - Repurpose for the consumer-side accounting

ATLAS@home

- Volunteer computing is supported by the BOINC
- Uses a pre-loaded CernVM image
- Run on the volunteer's machine using VirtualBox
- Job files injected via a share directory
- An ARC-CE used as a gateway interface
  - Between PanDA and the BOINC server
  - PandDA submits the job to the ARC-CE as normal
  - ARC-CE uses a specific BOINC backend plugin

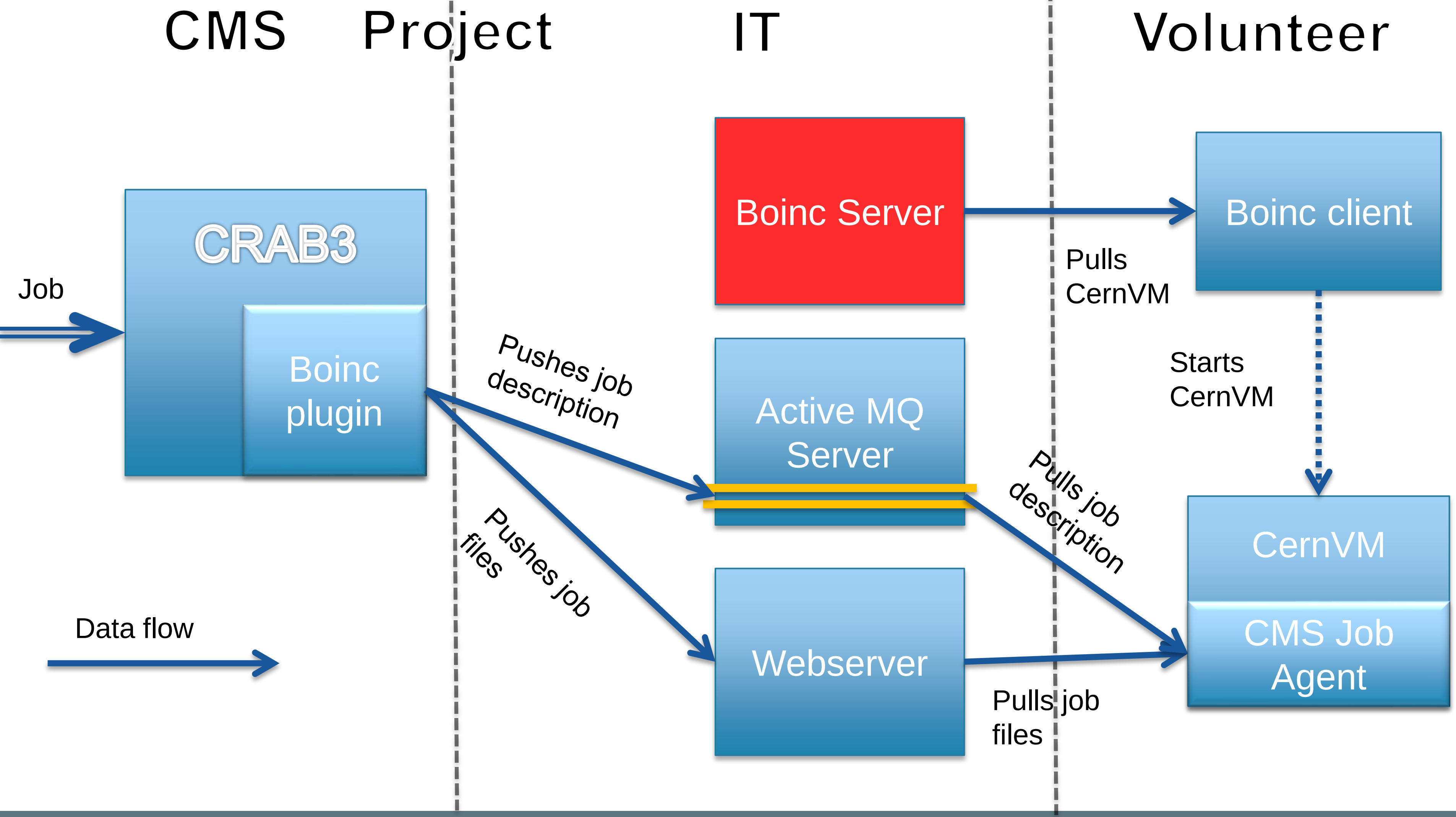
ATLAS@home


dashboard
Running jobs
199 Days from Week 13 of 2014 to Week 41 of 2014
4,500
4,000
3,500
3,000
2,500
2,000
1,500
1,000
500
0
Apr 2014
May 2014
Jun 2014
Jul 2014
Aug 2014
Sep 2014
Oct 2014
MC Simulation
Others
Maximum: 4,288 , Minimum: 0.00 , Average: 1,165 , Current: 1,921

CMS

- HLT Farm
  - Focus of the majority of work with respect to the adoption of cloud
  - Overlaying IaaS provisioning based on Openstack
    - Open vSwitch to virtualize the network
  - CMSooooCloud
    - CMS Openstack, OpenSwitch-ed, Opportunistic, Overlay, Online-cluster Cloud
  - GlideinWMS v3 is used to manage the job submission
- CERN’s Agile Infrastructure
  - Two different projects based at CERN and Wigner
  - Plan to consolidate the resources into just one Tier-0 resource
- Volunteer Computing
  - A prototype for CMS@home has recently been developed
  - Further development required for production

CMS@home prototype



LHCb

- CernVM is used for the VM image
- Monitoring is done via Ganglia
- The pilot system used is the same as for bare-metal worker node execution
  - A pilot is injected into the VM via CVMFS
  - The DIRAC job agent is started
  - It contacts the central task queue
  - Retrieves a matching a payload for execution
- VAC for hypervisor only based sites
  - Used in production on several WLCG sites
    - Mainly T2 sites for simulation payloads
- VCycle for IaaS controlled sites
  - Currently using the CERN AI
  - Planned to be expanded to more sites
- Developing a LHCb@home project using BOINC
- Not using virtualization on the HLT farm
  - Running offline workloads directly on the physical machines

Summary

- It is all about starting a CernVM image
  - And running a job agent
    - Similar to a pilot job
- Different options are being explored
  - To manage the VM life cycle
  - To deliver elastic capacity
- Already a great deal of commonality exists between the VOs
  - Should be exploited and built upon to provide common solutions
- How resources are going to be accounted?
  - Counting cores is easy
    - Normalizing for power is hard
- Investigating using job metrics
  - To discover relative performance
- Many open questions remain