Questions we would like to be addressed in a future session between CNAF and CERN.

Legenda:

- VWN = Virtual Machine Worker Node
- CWN = Container Worker Node (Docker)
- HV = OpenStack Hypervisor
- BM = Ironic Bare Metal host
- RH = Raw Host

Networking

1. Do you virtualize the network somehow?
2. How are all the WNs connected to the network?
    a. Public IP/private address/NAT/IPv6 in production
    b. Which is the throughput you are able to achieve from a single wn?
3. How are you connecting the storage elements to the network?
4. Can you run concurrently on a production physical WN several VMs or CWN belonging to different networks (VLANs or Vxlans) or different projects/context?
5. How are you segregating traffic of WLCG WNs from other projects?
6. Are you using any orchestration tools for the configuration of the network?
7. Can you briefly describe the network topology of the production farm?

Computing

1. Can you describe the software setup of a generic worker node @CERN?
    a. Do you run jobs on bare metal?
    b. Have you got Openstack deployed on all the nodes?
2. Have you evaluated the impact of virtualization on production WN? According to our investigations 2 to 5% performance is lost. Is this acceptable?
3. how many Virtual WNs per hypervisor? (or: how many cores and RAM per VWN?)
    a. do you reserve a core for OpenStack (and/or other) core services?
    b. is there an optimal VWN size (how many VWN per HV)
4. How dynamic the life of a RH can be?
    c. the same RH can work as HV hosting (only) VWNs (Once installed, the RH is expected to work only as HV for VWNs)
    d. the same RH can work as HV hosting (only) VMs (VWN and a mix of other VMs too)
    e. Once installed as HV, it is expected to host both VWNs and generic VMs
    f. the same RH can work as HV or BM (mutually exclusive)
    g. the same RH can host both VMs and Container on bare metal at the same time
5. Access to shared posix filesystems (gpfs, nfs, cvmfs, ...) happens via:
    h. A mountpoint on the RH, then the VM accesses it through its physical host
    i. mountpoint on the VM
6. How do you compute HS06 computing power on nodes?

   j. run hs06 benchmark in the Raw Host, then the power of a VWN comes as VWN_hs06 ~= hs06_rh / rh_cores * VWN_cores (when running on *that* host)
   k. run hs06 benchmark in the VWN
   l. other?

7. do you host multi-node jobs in the HTC (Batch) farm?
8. do you run HPC batch systems nodes on OpenStack (e.g., Slurm)? If so, are there special tunings to interface with infiniband or similar high-performance networking?
9. Container WN: what size do you allow for WN containers on Bare Metal?
   m. one Container as a "whole node"
   n. one container per CPU
   o. other?
   (Note: This choice determines the overall number of STARTDs in the HTCondor pool: are there observable effects on the reactivity / readiness of HTCondor the pool?)
10. Do you use a specific docker container orchestrator? How do you handle common batch system admin task such as “security upgrade of kernel and other packages”? More precisely the process should go through these steps:
   p. CWNs are drained (at HTCondor level)
   q. A new docker image is set up, with the upgraded sw package (at “puppet” level)
   r. When a CWN is drained, the container can be shut off (the container orchestrator level does that, however, the information “drain done” which is needed to trigger the shutdown comes from the HTCondor level)
   s. When a BM does not run any active container, it can be rebooted with the new kernel; this information should be known to the Openstack/Ironic layer.
   t. When rebooted, a Bare Metal should re-instantiate its CWNs
   u. Somehow one must coordinate activities at different “layers” (HTCondor to drain WNs, docker to poweroff/start containers, Ironic to reboot physical nodes). Do you have specific tools to handle this?
11. Do you provision batch (HTCondor) access to GPU resources?

Cloud

- Can you describe some use-cases for cloud adoption?
- Can you briefly describe the main cloud setup at CERN?
- How do you use ironic for bare metal provisioning? Can you explain the interaction with network? How many physical interfaces are connected? How do you manage network in openstack if bare metal needs more than one network?
- How do you contextualize a bare metal node after the Ironic provisioning? Cloud-init, Puppet-Foreman, other...
- Have you ever configured Openstack to manage vGPUs, like V100?

References:

Topic 1 “Large scale infrastructures management, including KPIs and monitoring”

- Subtopic T1.1 “Kubernetes & Virtualization”
- Subtopic T1.2 “Large Scale Deployments” -
- Subtopic T1.3 “GPUs”
- Subtopic T1.4 “SDN/Network”

- INFN Visit on Agile Infrastructure