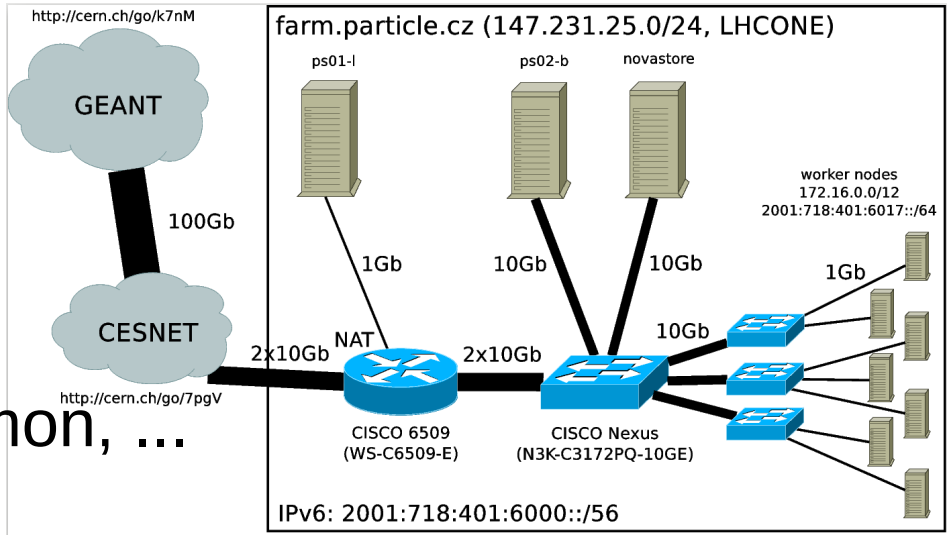
Containers for OSG NOνA jobs

Petr Vokáč
Institute of Physics ASCR

NEC2017, Montenegro, Budva
25-29 September 2017

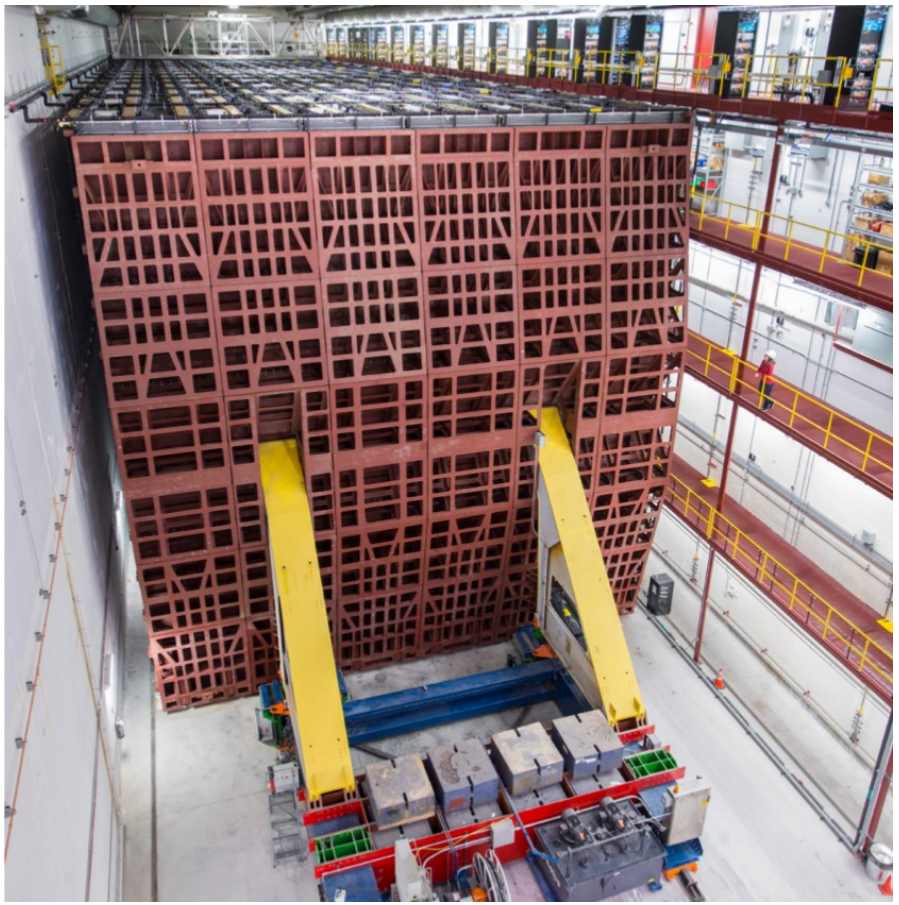
FZU cluster (farm.particle.cz)

- HEP Regional Computing Center
  - ATLAS (T2), ALICE, AUGER, CTA NOνA, DUNE, (DZero) users
- WN – 300 nodes ≈ 6200 cores
- Network
  - very simple (LHCONE)
  - IPv4 (NAT) + IPv6
- Storage (16 servers / no free BEER)
  - Nucleus for ATLAS, 2.8PB on DPM
  - xrootd with 1.2PB for ALICE
  - StashCache + SE for OSG
  - NFS for local users
- Services – argus, VO auger, …
- Infrastructure – provisioning, conf, mon, …
- 2.2 FTE



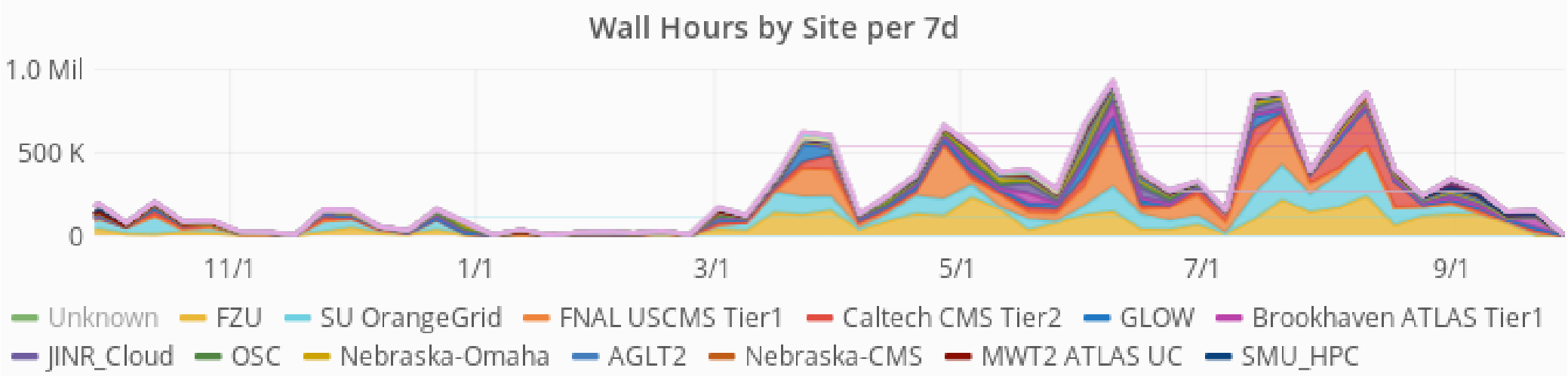
NOνA experiment

- ν experiment in Fermilab
- Off-axis ν_e Appearance
  - precise measurement ν_μ to ν_e oscillation
  - determine mass hierarchy for ν
  - measure CP-violating parameter δ
- two detectors 810km far away
  - ν detector huge
  - right distance & position
- remote operation



NOνA computing

- use tools provided by FIFE
  - started with local resources and later added support for grid
  - job submission with Glidein factory
  - OSG middle-ware
- mostly use OSG opportunistic resources
  - FZU have dedicated CPU & storage
  - 8% of total production done at FZU
- NOνA main task is ν physics
  - use stable production/analysis environment
  - can’t afford to spend to much effort with computing



FZU cluster environment

- one infrastructure that supports all VOs and local users
  - difficult to add new features
  - different requirements, sometimes contradictionary
  - changes driven mostly by biggest VO
- software installation base is getting rusty
  - WN SL6 with EGI / UMD packages (no OSG on WN)
  - no longer supported CREAM-CE for ATLAS, ALICE, AUGER, CTA
  - ancient local batch system (very old torque 2.4.16 + maui)
    - no support for “modern” features (e.g. cgroups)
    - sometimes unpredictable behavior (draining, fairshare, ...)
    - fragile
- started to evaluate more modern alternatives
  - ARC-CE for LHC
  - HTCondor as local batch system
  - CentOS7 as base system for WN
    - new hardware

FZU cluster & OS update

- survey for FZU supported VOs about CentOS7
  - NOνA did not even started to test their analysis / production software
  - difficult to synchronize update between all experiments
    - WLCG & OSG & other small experiments & users
- HTCondor 8.6.x – direct support (vs. starting containers in pilot)
  - chroot + cgroups + ...
  - virtual machines
  - containers
    - docker
    - singularity
- Containers
  - “lightweight virtualization” – no VM performance penalty
  - different/minimal OS on WN and for jobs (no update coordination)
  - allow experiment full control of job environment – reproducible
  - job isolation (PID, UID, net, mount), cgroup resource management
  - image deployment / management

FZU cluster & containers

- Evaluated Singularity & Docker
  - currently Singularity seem to be better option
    - designed for batch job execution
    - simple and minimal configuration
    - no daemon, can run without SUID
    - images stored on CVMFS (file, docker, tar, …)
      - FIFE GPGrid Docker image → CVMFS
      - “standard” images available for OSG & WLCG
        - /cvmfs/singularity.opensciencegrid.or/g/opensciencegrid
        - /cvmfs/atlas.cern.ch/repo/images/singularity
        - /cvmfs/farm.particle.cz/images/gpgrid
  - bind mount CVMFS & site disks
  - CVMFS cache for OS & application distribution
- NOνA jobs and containers
  - only simple manually submitted jobsub tests
  - production jobs not yet submitted
    - bind mounts
    - UPS/UPD

Summary

- FZU site in transition to more modern / supported OS & CE & batch
  - HTCondor is great batch system
    - very flexible with a lot of features (and configuration options)
    - takes time to get familiar with all details
  - ARC-CE
  - HTCondor-CE
- Containers makes our site maintenance more flexible
  - gives users and VO opportunity to use customized OS
  - there are still use-cases where full VM is better option
    - applications in container runs on different kernel
    - running really old OS environment on newest kernel
      - I would not try to sell containers as Data and Software preservation tool
  - still early stage of using singularity at FZU
    - needs some additional work to run real NOνA production
    - test also other VO jobs in containers

BACKUP