PAUL SCHERRER INSTITUT

PSI

Nina Loktionova

CMS Tier3 Report

ETH-CMS-user-group Face to Face Meeting, Zürich, 12.12.2019

PAUL SCHERRER INSTITUT

Outline

- T3 overview
- SL6 -> RH7 migration
- Plans
- AOB

T3 Overview

- Login nodes t3ui01-03 (one node per institute)
- Storage
  - home
  - work
  - dCache
  - scratch
  - EOS

- Batch:
  - Slurm
  - SGE (legacy)

- Infrastructure:
  - Hardware (and Virtual Machines), Network and Management
  - OS provisioning and Configurations
  - User Management
  - Monitoring
  - GRID/CMS services (example: Phedex->Rucio)

Storages I

available on all UIs and WNs

home

- PSI nfs server
- 10GB/user with snapshots (daily, weekly, monthly)

work

- total: 20TB; 2*10Gb/s network ;
- full quota 400 GB includes snapshots; effective quota ~ 150GB/user;

dCache

- total: 1.2 PB; free: 200TB
- 5 servers *250 TB and 2*10 Gb/s network
- type of data: not only root; protocols: dcap, xrootd,...
- no user quota
- distribution of data:
  - /mc - 58TB; /data - 78TB (written by Phedex)
  - **/user - 800TB**
  - /t3groups - 80TB

Storages II

scratch

- UI for job submission and short tests
  - ~ 5TB/node
  - no automatic cleaning
- WN
  - ~ 200-600GB/node
  - only for time of job (cleaned automatically after all user job ended on WN)
  - could be beneficial to use for job output (and afterwards copy on dCache/work) but depends on job

EOS

- what is the purpose to use this storage at Tier3? Cannot be used as data storage, only feasible “as buffer” for small files to exchange with CERN
- remote storage; slow because of network latency (“dd” to EOS 5 times slower in compare to /work)
- kerberos authentication
- mounted only to UIs (t3ui07 as a test)

Data Types and Workflows

Standard data analysis:

ROOT files (data/mc/user) are analysed using ROOT/pyROOT

understands root/dcap protocols, can work efficiently with dCache

New methods/format of data:

binary python data (numpy arrays) and analysis by user code.

Does software understands grid/HEP protocols?

If not, then a way to use dCache storage:

- copy data from dCache to POSIX storage (work/, scratch/) by dccp, xrdcp...
- run analysis
- copy results to dCache (with certificate)

Question: ratio of "Standard" vs "New" for next years?

SL6 -> RH7 and Slurm -> SGE migration

Reasons for change:

- End of SL6 support in 2020
- lxplus/CERN migration to CC7
- Request from users for compatibility with CC7 software
- Test period at T3 May-September with positive feedback
- Demand to support SL6/SGE till the end of November from 2 T3 users
- Slurm - modern powerful, flexible, well documented batch system widely used (among others: CSCS and PSI)

Batch Systems

Slurm

- Examples how to use:
  https://wiki.chipp.ch/twiki/bin/view/CmsTier3/SlurmUsage
- 1200 cores
- MaxJobCount=100000 (number of waiting+running jobs)
- priority depends on AGE and SIZE
- “quick” partition = < 1 hour
- current limit on number of running jobs/user : 400
- GPU policies:
  - 2 GPU machines, each with 8 GPU’s
  - opened for all T3 users
  - low usage last weeks
  - what kind of limits to introduce (# allocated GPU’s/user)?

SGE

- 150 cores on ~12 years old hardware (SUN blades)
- no support from Jan 2020

Slurm usage

Slurm utilisation in December is 67.31%:

```
[root@t3ui07 ~]# sreport cluster AccountUtilizationByUser format=Accounts,Login,Proper,Used Start=2019-12-01 -t percent
--------------------------------------------------------------------------------------------------------------------------------
      Account     Login     Proper Name      Used
--------------- --------- --------------- ---------
       cn-test                                67.31%
      cn-test  acalandr Alessandro Cal+      0.62%
      cn-test    anlyon Anne-Mazarine +      0.61%
      cn-test berger_p2 Pirmin Berger +      0.55%
      cn-test   erdmann Wolfram Erd+         0.08%
      cn-test  koschwei Korbinian Schw+      1.27%
      cn-test    mratti Maria Giulia R+      3.06%
      cn-test  oozcelik Ozlem Ozcelik +      0.00%
      cn-test  pbaertsc Pascal Baertsc+     14.18%
      cn-test    swertz Sebastien Wert+      0.09%
      cn-test      ursl Urs Langenegge+     16.97%
      cn-test   vmikuni Vinicius Mikun+      2.77%
      cn-test   vstampf Vinzenz Stampf+     27.10%
     gpu_gres                                0.01%
     gpu_gres berger_p2 Pirmin Berger +      0.00%
     gpu_gres  creissel Christina Reis+      0.00%
```

Downtime Plans

1. Jan 10-13 due to PSI shutdown
2. 1-2 days in February for dCache upgrade
3. 1 day in March/April for User Management system migration

Plans for Q1 2020: migration of user management system, local T3 LDAP to central PSI AD

LDAP phase out reasons:

- reach limit of reserved unix UIDs range
- legacy OS (SL5)
- current central PSI solution is good enough to use and facilitate integration with IT services

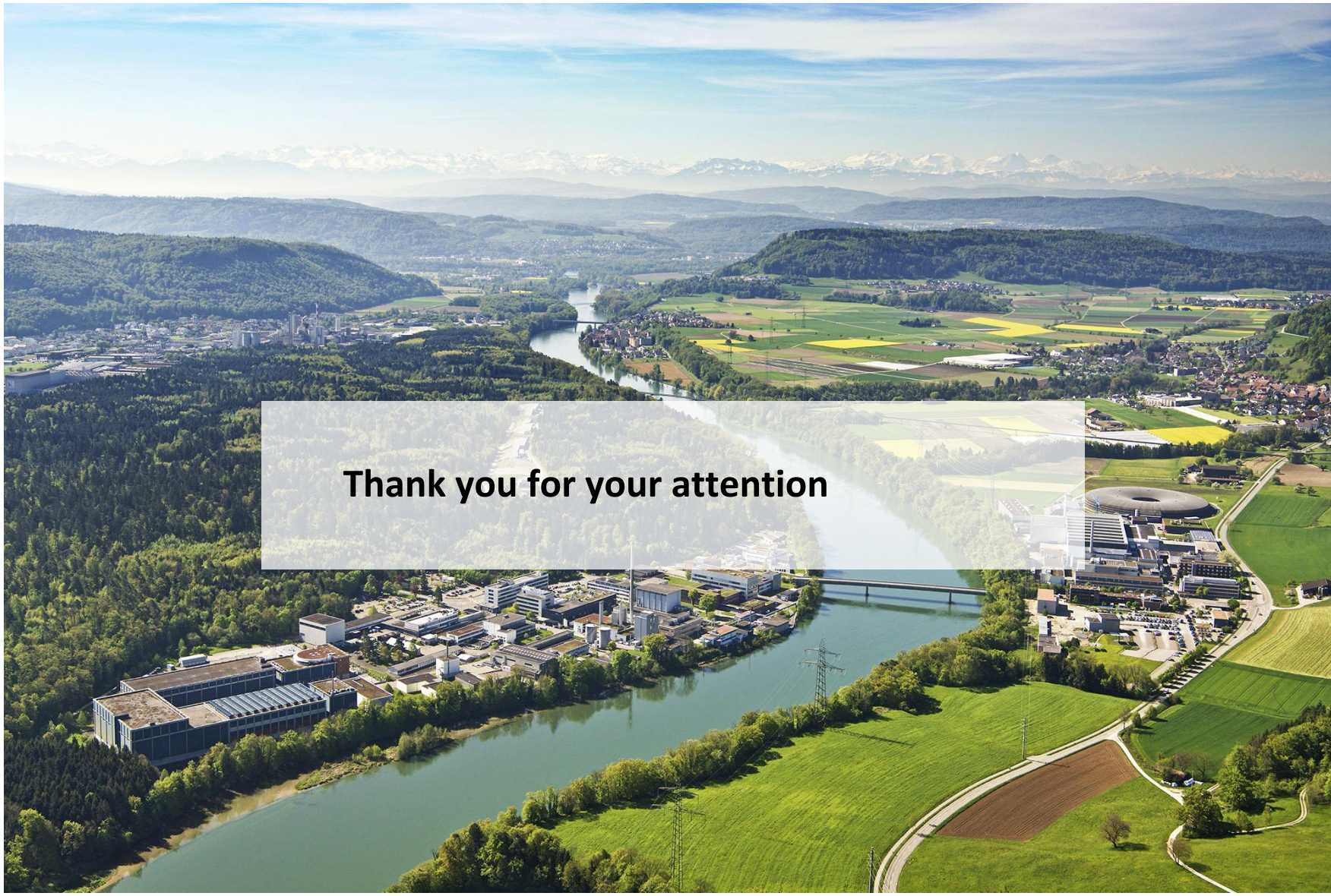
User side:

- change of user names and password policies
  - names : (vstampf -> ext-stampf_v)
  - password change every 6 months
- account expiry policy: max. 12 months, no automatic prolongation
- password change on a dedicated machine (cpw.psi.ch)
- migration from LDAP is after shutdown of SL6 (SGE)

Directions of T3 development next year(s)

budget is limited and most will be spent to renew/extend current aging hardware

- More CPUs? GPU's?
- More storage like /work (~ 2TB/user) ?
- Larger storage like dCache, but POSIX-like?


Thank you for your attention