Dark Energy Center
A. Tilquin
8 Novembre 2018

Task force contributors:

J. Bregeon (LUPM, coordinator)

A. Tilquin (CPPM, DEC coordinator)

LUPM: N. Clémentin, M. Sanguillon

CPPM: T. Mouthuy, D. Fouchez, A. Ealet, A. Pisani

LAM: S. de la Torre, E. Jullo, C. Surace

CPT: J.M. Virey, J.R. Liebgott, V. Salvatelli

IRAP: A. Blanchard

Le LabEx OCEVU (ANR-11-LABX-0060) bénéficie d'une aide de l'Etat gérée par l'Agence Nationale de la Recherche au titre du programme d'investissements d'avenir portant la référence ANR-11-IDEX-0001-02 (A*Midex)

Outlook

- Why a dedicated regional cluster ?
- Hardware infrastructure
- Software tools and development
- DEC usage : which parallelism
- Summary

What about the DEC?

- Found by labex OCEVU to prepare large future survey >2020
  - LSST and Euclid on dark energy/dark matter and cosmology
  - 350 kEuros and 2 years of an IE
  - Shared by 5 lab: CPPM/CPT/LAM/LUPM/IRAP
- It has been design as:
  - a development machine (not a production one)
  - Running mainly interactive jobs (debugging).

    -> It's a sand box i.e no restriction, no quota etc….
- Main requirement was:
  - as many CPU as possible
  - Huge shared memory.

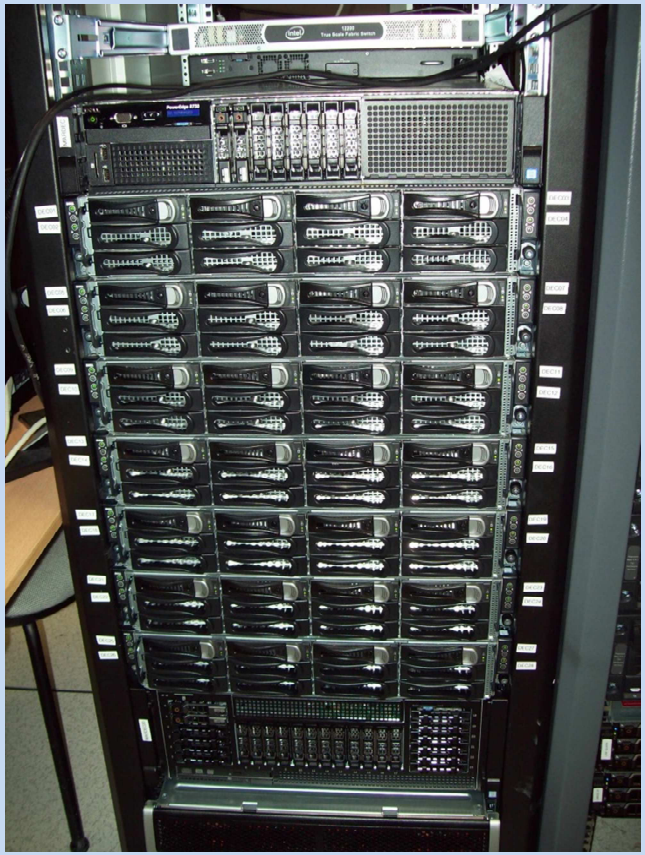
    ->HPC like system (DELL ou SGI ?)
- September 2016:
  - Installation at CPPM
- November 2016:
  - First light

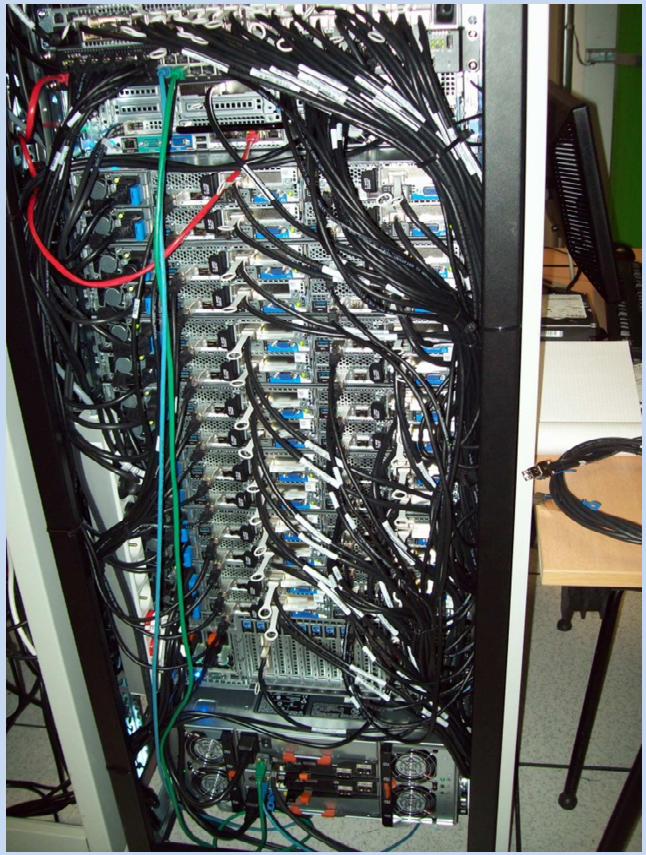
    ->Running smootly up to now (>99% efficiency)

Dark Energy Center: Hardware

- Hardware infrastructure: Cluster HPC/DELL
  - 29 nodes:
    - 56 threads/node: total of 1624 threads : ~1 million hours/month
    - 1.5 TB+ 28*0.5 TB: total de 15 TBytes of memories
    - Same amount of swap memories.
  - 1 main server:
    - 40 threads et 256 GBytes of memories
    - 330 TBytes of hard disk (raid 5). No backup !
  - 3 networks:
    - 2 Ethernet network at 1 Gb/s and 10 Gb/s
      - Disk (nfs) + ssh connexion
    - 1 infini band network at 40 Gb/s
      - Shared memories + parallelism (MPI)

DEC in picture



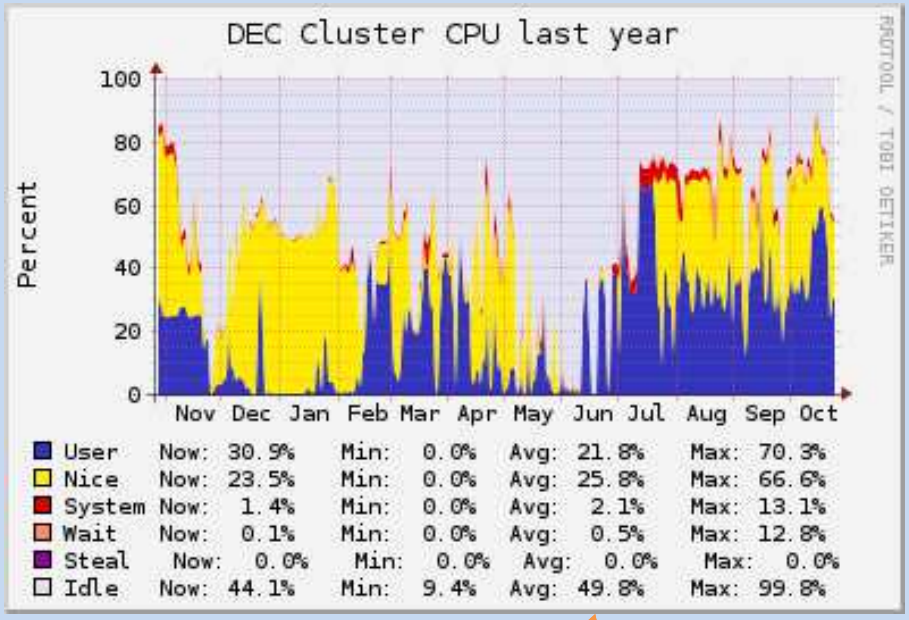


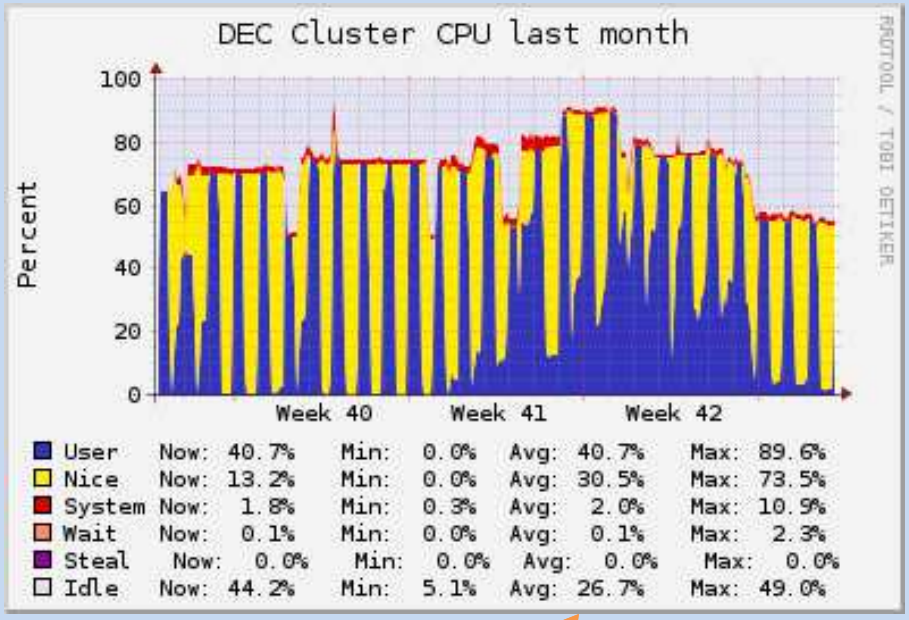
Thanks to Thierry Mouthuy and Adrien Riviere

Software

- System software:
  - Scientific Linux on all nodes
  - C/C++/Fortran/GDL/… and Python !!!
  - Open-MPI (mpich)
- Users monitoring command (first 6 months, no IE):
  - mardec_load : classify nodes according to loading factor (mpi)
  - mardec_renice : renice all jobs or jobs name for a given user (long jobs)
  - mardec_cpu : instantaneous cpu and memories usage for a given user/all
  - mardec_clear: kill almost all processes for a user i.e logout
- System monitoring (second year):
  - Ganglia2 (global monitoring)
  - Queue batch: Torque/MAUI: 1 queue and no limit
    - Use in a non standard way (bypass the scheduler) to take into account of interactive jobs.
  - Automatic renice: priority to interactive jobs (day/night)
  - mardec_stat : users cpu monitoring

DEC cpu usage.

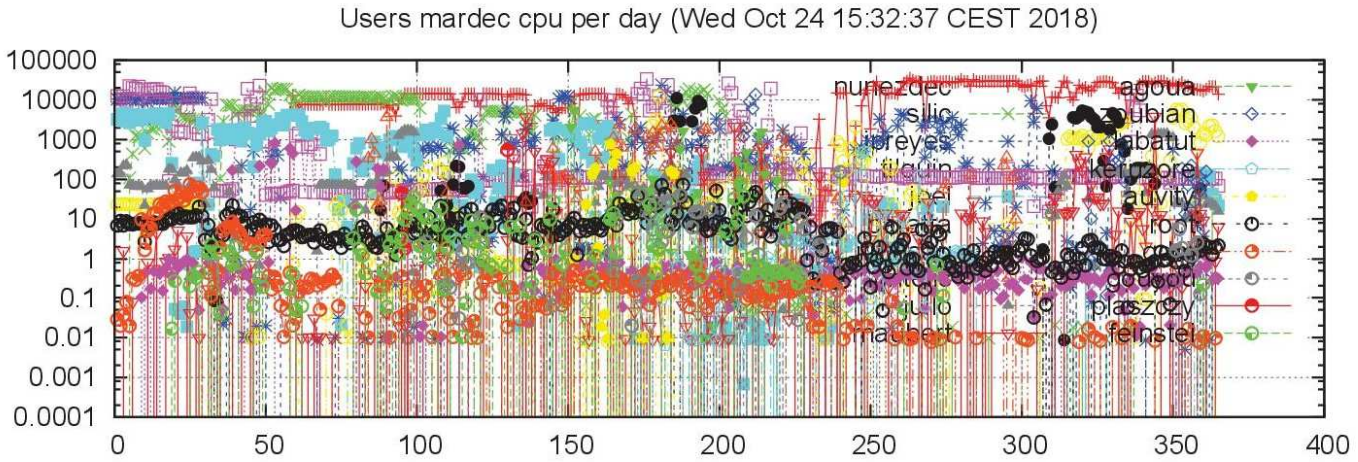


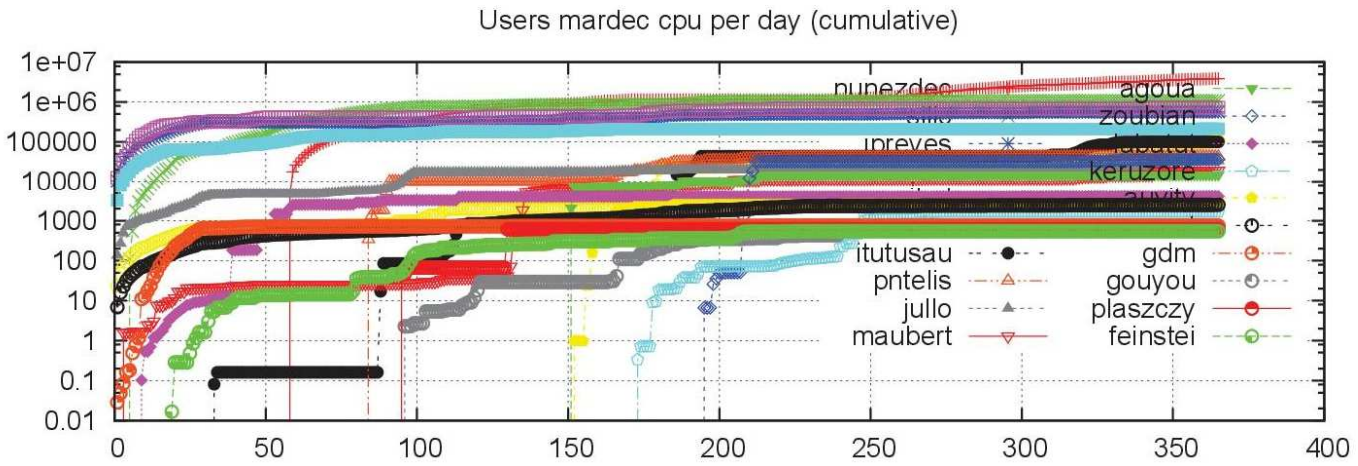


50 % free CPU

27 % free CPU

DEC users cpu monitoring





nunezdec	4044551.954170
silic	1113830.475906
jpreyes	618371.330571
tilquin	483692.073360
jbel	226542.304456
baratta	109337.169007
itutusau	103216.118805
pntelis	46854.976000
jullo	34587.542332
maubert	19436.329416
agoua	15005.900000
zoubian	4557.516676
labatut	4142.974165
keruzore	2666.530000
auvity	2598.550000
root	2489.940729
gdm	636.508931
gouyou	619.542463
plaszczy	616.580110
feinstei	615.003512
jlambert	366.643500
mcc	296.108993
pasquet	30.060850
lalloue	9.620350
aubergie	9.129550
binome02	6.769540
fouchez	5.115910
binome07	4.601780
binome06	1.076580
binome03	1.035840
nobody	0.008851

Which kind of parallelism ?

3 kinds of parallelism:

- Embarrassing: one job -> one thread
  - As in particle physics
- Vectorial or multi-threads: one job->many threads
  - Vector algebra (python,numpy)
  - Linear algebra/matrix inversion: lapack
  - FFT : fast Fourier transform
- HPC: one master -> many slaves on many nodes
  - Mainly for Nbody simulation on big space volume
  - Using mpi and infiniband for shared memory

Many types of sciences on the DEC

- Embarassing: (10%)
  - Photon spin in cosmology with SN : chi2 statistic :(CPPM)
  - Bouncing universe at Planck temperature with SN : chi2 (CPPM)
  - Cosmological probes combination using MCMC statistic (IRAP)
- Embarrasing + Vectorial (40%)
  - Image processing for LSST (needs more than 1 TB of memory) (CPPM)
  - Image simulation for EUCLID (stray light)
  - Analysis of EUCLID infrared detectors characterizations. (CPPM)
  - Strong lensing: ray tracing (LAM)
  - Cosmic void with galaxies (Voronoi tessellation) using different cosmologies. (CPPM)
  - Fast fourrier transform on two points galaxies correlation function on a cube (4096)^3 (CPT)
  - Algebraic simulation of large scale structure of galaxies (CPPM/CPT)
- HPC: (50%)
  - Two or 3 points correlation function on millions of galaxies: combinatorial (CPPM/CPT/IRAP)
  - Nbody simulation for dark matter galaxies (LAM)

Summary

- In cosmology, high parallelism is necessary
  - Universe is huge and contains many objects !
- HPC cluster are suitable
  - Memories required : at least 20 GBytes/core

HPC cluster is not incompatible with interactive at least for development software

Ganglia DEC Cluster Report overloading.htm (Command Line)

Examples.

mardec_load

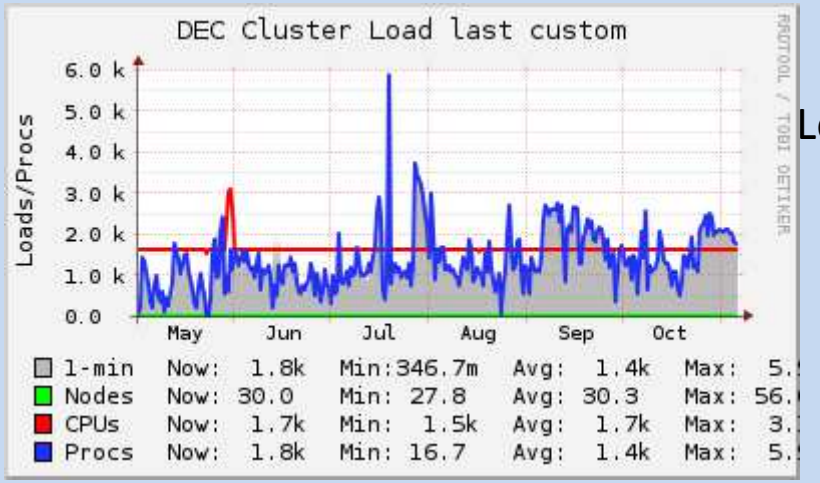
```
mardec00, load average: 1 %
mardec01, load average: 1 %
……
mardec20, load average: 86 %
mardec21, load average: 86 %
mardec22, load average: 84 %
mardec23, load average: 84 %
mardec24, load average: 88 %
mardec25, load average: 84 %
mardec26, load average: 81 %
mardec27, load average: 84 %
mardec28, load average: 82 %
----------------------------
--->Total dec load: 26 %
--->You can run on mardec02
----------------------------
--->machinefile has been updated
```

mardec_cpu all

User	Total DEC cpu(%)	Total DEC memory (%)
nunezdec	55.0193	10.5429
baratta	3.90842	3.03214
jpreyes	3.8604	0.271429
pasquet	0.661542	1.00357
maubert	0.5706	0.0714286
jullo	0.375701	0
tilquin	0.261543	0
jbel	0.0609057	0.557143
avahi	0	0

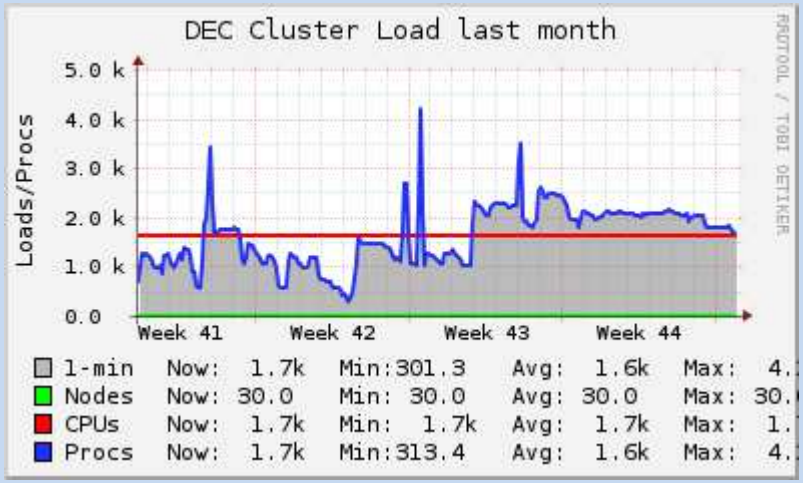
Global DEC monitoring : Ganglia

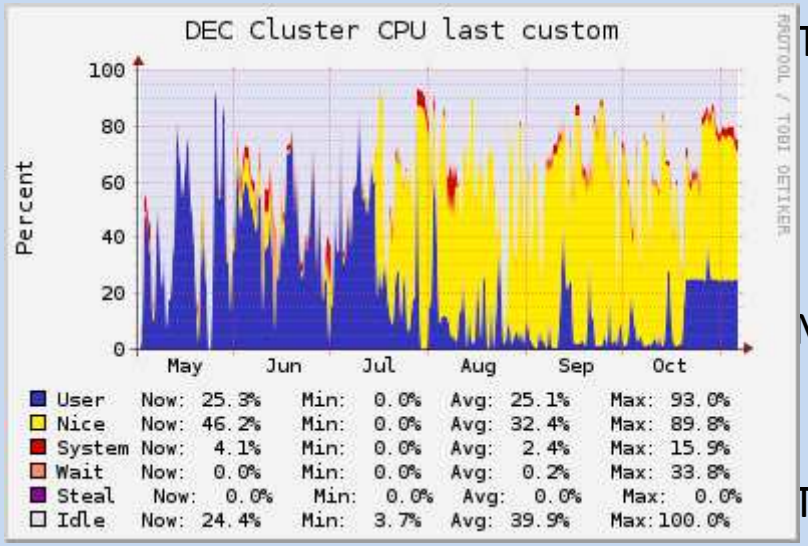
DEC usage last 6 months



Load:
Average: 81 %
Max : 400 %
Now : 105 %

DEC usage last month

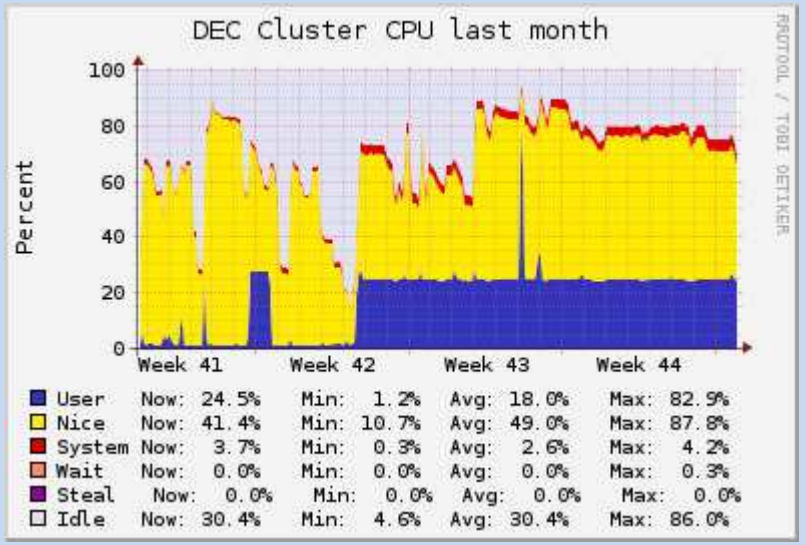




Total cpu:
Average: 60 %
Max : 100 %
Now : 70 %

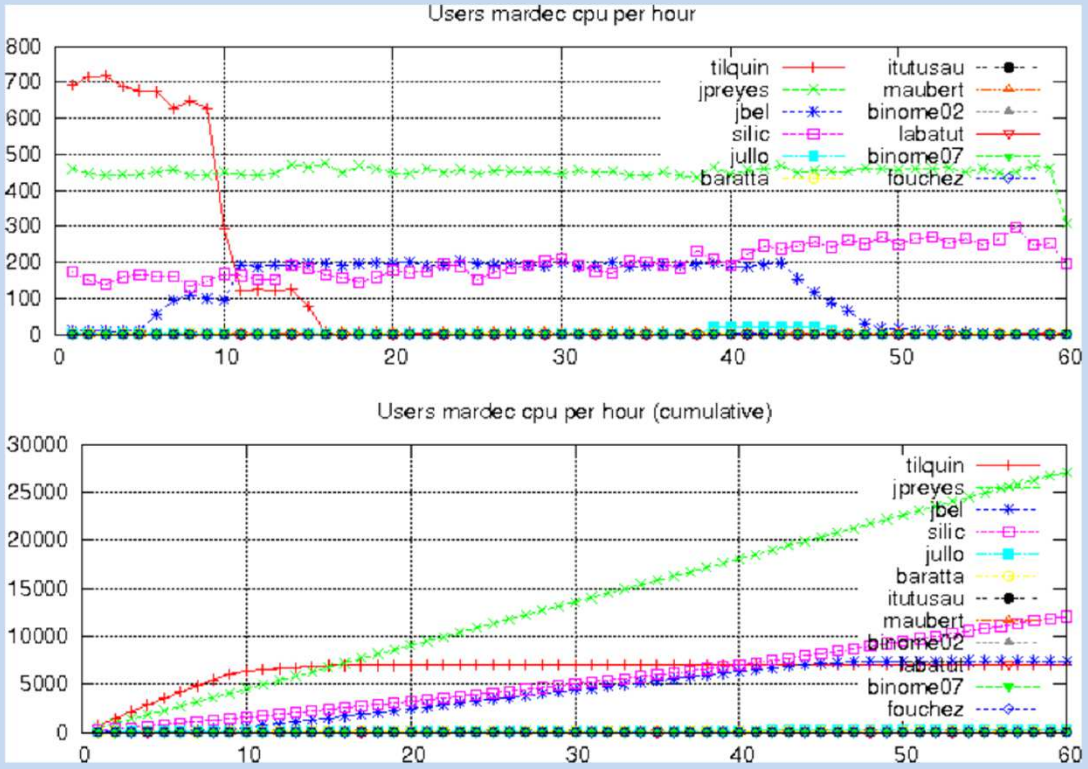
Memory used:
Now : 52%

Total disk used:
120 TB (40%)



Monitoring total user CPU for Batch/interactif Under development

- Need to create a new tool (no public software)
  - Create a personal database
  - Collect information (ascii or graphical)



- Still do be done:
  - Use previous plots to automatically shared users priority

Regular users: 15

- CPPM: 7 users
  - LSST: Images processing, calibration and simulation
  - Euclid : Stray light
  - Phenomenology and analysis:
    - BOSS:Void/Clustering/
    - SNLS:Light curve fitting/Photon spin/
- CPT : 2 users
  - MCMC/FFT BAO
- IRAP : 2 users
  - Montepython: Probes Combinations
- LAM : 4 users
  - Strong lensing
  - BAO/Galaxy clustering
- Euclid school (July 2017): About 30 users

Papers(submitted/in preparation): 3/4

Summary

- DEC run smoothly (80-90% occupancy)
  - 30% of free CPU (new users !)
- In fully interactive mode
  - Only few spike in loading factor (debugging)
- Users share their own priority
  - High/Low (Blue/Yellow)
- Batch system working in modify version
  - Take care about interactive jobs

Questions from OSTC (DEC+cloud)

- **OSTC requests the setting up of a Technical Advisory Committee (name to be defined), essentially composed of members external to the project, to manage the strategic technical choices on the OCEVU computing resources, as well as the policy of access to them. It requests that a report be made at the next OSTC meeting on this topic.**
  - All technical choice already done one year ago for DEC and Cloud.
  - The DEC is essentially a “sandbox” and not yet ready to manage interactive and batch users.
  - The DEC main policy (priority) is shared by users and almost respected by them. No conflict.
  - Right now no need of a computing resources committee.

- **OSTC insists on having the engineers linked to the OCEVU computing resources spend a significant fraction of their time on user training and software development.**
  - Some training sessions already done for LUPM-Cloud. One plane for beginning of 2018.
  - Most DEC users already have some expertise on HPC cluster and I wrote a small documentation on MPI and DEC usage.
  - Some human problem for the DEC.

But no critical point !