Report on the OCEVU LabEx computing needs

The OCEVU Computing Task Force

22nd October 2015

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)*

Contents

1 Introduction

The goal assigned to the task force is to define the need for computing resources of the OCEVU LabEx, in particular the ones linked to the two following projects:

- LUPM–*Cloud* project
- Dark Energy Centre (DEC)

Three items have to be explicitly clarified by the Task Force:

1. identify OCEVU scientific research items and their associated computing resources,
2. propose a technical solution including specific hardware and budget plans, and taking into account currently available resources,
3. express needs for man power support as an *Ingénieur d'Etude* hired for 3 years.

2 OCEVU Dark Energy Centre

2.1 Scientific case and computing power needs

The Dark Energy Centre is an initiative from the cosmology groups of the OCEVU LabEx to share their needs in term of computing. Laboratories involved in the DEC are the following:

- CPPM: 7 staff researchers, 3 postdocs and 3 students
- CPT: 7 staff researchers, 4 postdocs and 3 students
- LAM: 4 staff researchers, 2 postdocs and 3 students
- LUPM: 4 staff researchers
- IRAP: 2 staff researchers and 2 students

In this document, a computing unit named MARRSD is used as a reference. MARRSD is a DELL PowerEdge R420 equipped with 2 CPU Intel (R) E5-2470 v2 @ 2.4GHz, for a total of 20 cores with hyper threading (40 threads in total) and 384 GB of memory (about 10 GB/thread).

For more than a decade, cosmology has become an accurate science at the statistical level and at the systematic errors level as well. The discovery of dark energy and the confirmation of cold dark matter at cosmological scales convinced the scientific community to build new large ground–based telescopes (eBOSS / LSST/SUBARU) as well as space telescopes (Euclid/JWST), dedicated to sky surveys. These surveys will cover about 20000 square degree up to a redshift higher than 1 and will be operating in 2020. The total number of expected objects for Euclid or LSST is as huge as 10 billion, with about 1 billion of galaxies and 500000 well identified type 1A supernovæ.

Using full probes combinations of these new sets of data will help to break down degeneracies between cosmological and astrophysical parameters. Such a

high statistic will give the opportunity to explore accurately modification of gravity, general relativity, and cosmological principles. As a consequence, new developments on experimental analysis technique and theoretical models are needed to prepare the coming new data. The increase of statistic and the complexity of new theoretical models will not be feasible without a dedicated computing centre for dark energy. National centres like CC–IN2P3, local centres as the *Mesocentre* in Marseille or dedicated centres for LSST and Euclid will be necessary in the exploitations phases after 2020 but are not well adapted for development phases. In addition some analysis already performed within OCEVU are not directly included in collaborations and need a dedicated computing centre. The dynamism and the visibility of the dark energy group require reasonable computing power to face the extremely strong international competition in this field. Our needs are essentially dedicated to partial analysis on existing data (BOSS/eBOSS), preparation of future analysis, combinations, and phenomenology. The main science cases are:

- LSST: Study of type 1a supernovæ for cosmology.
- Euclid/BOSS/eBOSS/PFS–SuMIRe: Baryonic acoustic oscillation (BAO), redshift–space distortion (RSD), test Alcock-Paczynski, cosmic voids.
- Combinations, combined analysis of all available probes: SN, BAO, RSD, VOIDS, Weak Lensing, CMB.
- Phenomenology: Modification of gravity, extension of general relativity, cosmological principle.

2.1.1 Supernovae and LSST

Type 1A supernovæ constitutes a special probe. These objects have the property of being standardizable. The brightness at maximum of explosion is almost always the same. The measurement of this standardized brightness therefore provides a direct measurement of the distance and can constrain cosmological parameters. Although it has been shown that supernovæ can't break down the degeneracy between cosmological parameters, that probe remains inevitable because it's purely geometrical. Main activities on this subject are:

- Photometry: This part consists in extracting any transient object as supernovæ by comparing new image with a reference one. The first step is to create a reference image by co-addition. The main challenge is to bring each image at the same resolutions by using PSF de-convolution. Each new image is then compared to the reference one by using image subtraction to detect and extract the transient object. Both steps require an accurate calibration to get read of false detections and to measure magnitude of the object. Vector calculus, parallelism and memories are essential for image processing.

  A Dedicated machine like MARRSD will be necessary with at least 1.5 TB of internal memory to process images and to deal with databases

- Supernovae detection and light curves: Once a transient object is detected during a long period of time (light curve), it is necessary to precisely identify it. The proposed method is to use a neural network to identify type

1A supernovæ using light curve features, and environment properties as galaxy type. The definition of the neural network and especially its training is a process that requires memory, CPU and parallelism.

Analysis of supernovæ and cosmological parameters extraction: Light curves parameters and cosmological parameters are degenerated. Therefore, the fit of the Hubble diagram must be done simultaneously to the light curves analysis (stretch, color extinction, etc...). Correlations between parameters for each supernova imply large matrices inversion (at least 10000×10000 for 3000 supernovæ). Furthermore, the adjustment being made on a large number of parameters, it becomes necessary to use Markov chains. The minimum size of a chain for a simple estimate of the covariance matrix is about 1000000. Parallelism and memories are needed.

A MARRSD machine is envisaged for the 2 previous points with 10 TB of disk.

2.1.2 Baryonic accoustic oscillations: BOSS/eBOSS/Euclid/PFS–SuMIRe

Of all the methods for probing cosmic acceleration and thus constraining the characteristics of dark energy, the BAO technique is generally recognized as "the method least affected by systematic uncertainties'" (Albrecht et al. 2006). At low redshifts, the BAO approach is a powerful complement to supernova studies, in part because of its low systematic uncertainties, and in part because BAOs directly measure the cosmic expansion rate H(z) in addition to the distance-redshift relation DA(z). At high redshifts, the large comoving volume available to measure clustering allows the BAO method to obtain remarkably precise measurements of both distance and expansion rate.

Since 2010, we have been members of the SDSS collaboration, and in particular we have used the data from the BOSS and eBOSS surveys. In particular, the eBOSS surveys started in 2014, and will last for 6 years. It will produce spectroscopic redshift for about 300,000 luminous red galaxies (LRG), 180,000 emission line galaxies (ELG) and 500,000 quasars. In Marseille, our expertise lies in the combination of weak lensing and galaxy clustering, in order to test general relativity predictions. The computational problem lies in the construction of cross- and auto-correlation functions between these datasets. For about 500,000 galaxies, the computational load amounts to about 7 months on 1 thread of MARRSD.

In addition, the construction of covariance matrices requires the repetition of these calculations for several hundreds of random but realistic realisations of the observed datasets. With 280 mocks of 20,000 galaxies each, it would take us about 1 month on the 40 threads of a MARRSD machine, with CPUs working in parallel by group of 4. For a publication, we estimate our needs to 1000 mocks, i.e. 4 months of MARRSD. Our forthcoming analysis would require at least 3 MARRSD with 40 TB of storage to store the mock catalogues.

In the coming years, we will get involved in several other surveys, making use of the expertise we developed with the SDSS projects. We can cite the survey performed with the Prime Focus Spectrograph (PFS–SuMIRe) on the Subaru telescope, and Dark Energy Spectroscopic Instrument (DESI) on the Mayall telescope in the US, and the Euclid mission. The preparation of these programs requires large N-body simulations, and the development of algorithms and methodologies to process the large flows of data and investigate possible systematic errors. Typically, we estimate our needs to 100,000 hours per type of cosmological probe:

BAO, RSD. In order to get ready by 2020 (codes tested to run on larger data-center), at least 2 MARRSD are necessary with about 20 TB of storage.

2.1.3 Redshift Space Distorsions

This cosmological probe that uses the impact of galaxies peculiar speeds on the amplitude and the isotropy of the 2-points correlation function, still requires many improvements from a theoretical point of view. Before it can be applied to precision data as Euclid will provide, it is necessary to develop new models to be tested in simulations in order to verify if their ability to describe redshift distortions is sufficient to extract information on the galaxies speed field and in particular to be able to verify or disprove the laws of Einstein's gravitation. This theoretical development requires the study of simulations of the large-scale structure of the universe in models of standard and alternative gravity. The estimate of the essential element (the "pair-wise velocity distribution") to improve the theoretical description of the phenomenon of redshift distortion takes about 32 days (for all levels) on 1 thread on MARRSD. In order to develop new models by analysing many simulation volumes with different redshifts and for different models, we need to perform around 1000 analyses which can be obtained with a reasonable period of one year with 3 MARRSD (120 threads).

2.1.4 Cosmic voids

Modern surveys such as BOSS/eBOSS/Euclid allow us to access to high quality measurements, by sampling the galaxy distribution in detail also in the emptier regions, voids. Thus, cosmic voids present themselves as a new tool to constrain cosmology. In particular, by using void stacks as standard spheres, it is possible to perform an Alcock-Paczynski test and constrain cosmological models (Lavaux et al. 2012, Sutter et al. 2014). The test measures the ellipticity of void stacks to constrain cosmological parameters. Nevertheless peculiar velocities affect the way we observe cosmic voids, and thus their effect needs to be understood to optimally exploit this new cosmological probe. The investigation of systematics effects affecting the extraction of cosmological parameters from voids is important for their application with current and future data. The AP test on voids with current and future data promises to bring competitive constraints to cosmological models. Additionally, the number of voids itself can be used to constrain the properties of dark energy, such as its equation of state. The void abundance probe will also be investigated within eBOSS, to prepare its use for Euclid, that will provide a considerable number of cosmic voids. In the coming years we will focus on eBOSS data, in order to prepare the void analysis for Euclid. Such analysis will thus need the production of void catalogues from real data and mocks, corresponding to 1 MARRSD with 1 TB of disk space.

2.1.5 Lensing

The weak lensing signal over large areas is a powerful geometrical probe. Nevertheless, it can be significantly affected by systematics related to the intrinsic difficulty of measuring galaxy shapes. Simulations of the lensing effects over large field-of-view are therefore crucial to improve methods and control systematics to the lowest level, particularly for precision cosmology.

Preparing a lightcone area of 100 deg^2 of lensing requires about 15 days on 1 MARRSD thread. With an N-body simulation of size L = 2.5 h^{-1} Gpc (BigMultidark), we managed to simulate about 600 deg^2 between redshift z = 0 and z = 2.3. In total with this simulation, we created 6 lightcones of different sizes and geometries, in order to reproduce at best the reality of the observations and find out possible systematic errors. Hence, approximately $6 \times 600 = 3600$ deg^2 of lensing (non-independent) have been produced, representing 1.5 years of computation on 1 thread of MARRSD. In terms of storage, 400 deg^2 of lensing (catalogues and maps) occupies 100 GB on disk. Today, 750 deg^2 of imaging survey have just been made public by the American project DECam Legacy Survey (DECaLS). We expect to use a simulation of size L = 4 h^{-1} Gpc recently produced by our collaborators in Madrid to create 1600 deg^2 of lensing ligthcones, equivalent to 8 months on 1 thread of MARRSD, or 6 days on 1 MARRSD with 40 threads. Each snapshot occupies about 2.5 TB and it requires 24 snapshots to produce 1 lightcone, i.e. in total about 60 TB of storage.

2.1.6 Multi-probes analysis and MCMC

All analysis have to be done with a Markov chain. The number of parameters has become too large. The minimum size of the Markov chain for estimating the covariance matrix for a model that corresponds to a specific theory and a unique combination of probes is at least 10^5 to 10^6 steps. Everything therefore depends on the calculation time required for the likelihood estimation, variable between 10 seconds and 1 minute, depending on the type of analysis. The current time to calculate a single step of the chain, on an Intel machine (R) Xeon (R) 3.30GHz with 6.5 GB of RAM per core, is 15 s when you consider the model LCDM and only the likelihood of the CMB. In addition we have to take into account the time needed to write the output. In the future this time can only increase due to the higher complexity of the theoretical models to be tested, the precision required and the number of data. This means that with a single chain one would need between 17 and 68 days for the convergence of a model. In fact, thanks to parallelization that can be obtained with a multi-core machine you can decrease, using 8 parallel chains, between 2 and 8 days for the convergence of a model. However the ratio between the time of calculation and the number of chains used is not linear. For this type of analysis, it is much more useful to increase the RAM of each core which accelerates the computing of the likelihood, rather than on the number of cores of each machine, making MARRSD a good candidate for this type of analysis. With two MARRSD machines one can make a complete project, between 15 days and two months of computing time, where we consider a relatively small number of combinations for parameters and probes.

This estimate, however, is the minimum for a simple likelihood of CMB type, but if we want to combine the constraints coming from the clustering of galaxies (BAO and RSD) and weak lensing, the time calculation for the theoretical predictions is much more important, in particular due to the calculation of non-linear corrections. Thus the calculation time of a step from the Markov chain can be multiplied by a factor 2 to 4 depending on the cosmological model to be tested against the CMB case alone. In particular, in the case of modified gravity models, the computation time can become huge.

To be more precise, we need to perform different types of analyses. Within the COBESIX group (funded by the LabEx) we test various theoretical models

and different combinations of observables with present data. For COBESIX type analysis in the context of Euclid and LSST, we should consider an increase of the needs related to the preparation of accurate analysis of new cosmological models. These experiments will allow the measurements of various cosmological probes (like SN, BAO, RSD and weak lensing) at the percent level or even less. The analysis pipelines will be designed and run within the collaboration in order to get this level of statistical accuracy. But to adapt that in the context of a COBESIX type analysis, we anticipate that the increase of accuracy will allow to test more complex models (means more parameters) that need on one side to produce more complex covariant matrices (not standard in the collaboration) and, on other handle, with more complex likelihood and Markov chain Monte Carlo to take new degeneracies into account. At this level, no approximation or simplification of parameters will be implemented as we can do currently with the current precision. Also, new modelisation to take non linearity into account will probably require further developments of the tools. We think also that COBESIX will be prepared to add new 'non standard probes' that can be tested in a local level (as new cluster analyses, voids etc, and their comparison with standard analyses). For all these reasons, we anticipate an increase of at least factor 3 in the MARRSD like machine need (2 before Euclid/LSST and 6 after).

2.1.7 Phenomenology

For several years, experimentalists and theorists from OCEVU dark energy groups have developed joint analysis mainly with the current type 1A supernovæ data (about 1000 by combining JLA and Union 2 samples) and with simplified simulations of LSST (50000 s randomly distributed on the celestial sphere).

As an illustration, we have worked on the following subjects (almost all accepted and published) and we plan to apply these analysis on LSST supernovæ as soon as data are available (50k per year)

- Study of systematic bias induced by wrong assumptions on dark energy behavior: The exploration of parameters space required numerous simulations and minimizations, 100k hours of CPU expected for future analysis.
- Modification of the cosmological principle using a maximally symmetric universe on the light cone: This modification doesn't modify Friedmann equations for background, but perturbation theory should probably be modified and applied on other probes than supernovæ. No estimation of needed CPU is feasible at that point.
- Einstein-Cartan Universe: A torsion field related to spin density can replaces dark matter (Union 2 data). Cosmological constant is still needed. Four differential equations should be solved numerically. The analysis of the first year of LSST supernovæ will require more than 200k hours of CPU.
- Bianchi 1 metric: Supernovæ indicate a preferred direction (compatible with other measurements). The published analysis using JLA and Union 2 samples used 100k hours of CPU (Figure 1). A fast simulation of LSST shows that after 1 year an accuracy of 3 per a thousand on the relative variation of the Hubble parameter is possible. The main difficulty of this analysis is to explore the full galactic sphere and to perform a fit in each direction (by step of 1 degree). 1 million hours of CPU time is expected for one year of LSST

because of the global fitting procedure (cosmological and light curves parameters).

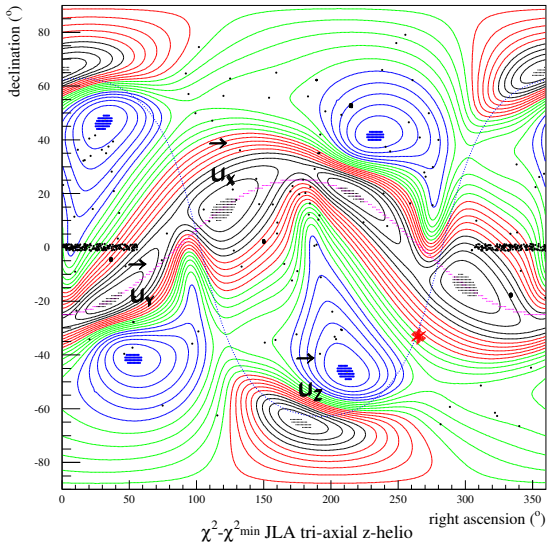

Figure 1: Confidence level contours of privileged directions in arbitrary color codes for the tri-axial Bianchi I metric. Black points represent supernovæ positions. Note the accumulation of supernovae in the equatorial plane. The blue line is the galactic plane and the purple line is the plane transverse to the main privileged direction $\vec{u}_z$ (gray specks). Blue specks correspond to regions where χ^2 is maximum. **The red star represents the galactic center direction**. 50k hours of CPU on a single thread is needed to construct this figure.

- Dark energy equation of state: Modifying the dark energy equation of state by adding a quadratic term has a drastically impact on the evolution of the universe. It leads to a universe without initial singularity replace by a hot sneeze and totally compatible with current supernova data. 100k hours of CPU will be enough for future analysis.

- Bianchi 1 metric coupled with Einstein–Cartan torsion: This work is still ongoing and already requires more than 100k hours of CPU. The main problem in this case comes from numerical instability. Future analysis will need new software development if we want to use it on 1 year of LSST data with less than 1 million hours of CPU.

Overall, two MARRSD machines will be necessary with 5 TB of disk.

2.1.8 Higher order statistics

The two-points correlation function is a key ingredient when analysing the large scale structure of the universe (matter and galaxies distributions). However, it is not sufficient for the complete characterisation of the cosmological fluids like the dark matter. To fulfil this aim, higher order statistics are mandatory. For instance, they provide many informations on the bi-spectrum of matter and galaxies distributions. At order three, the simplest extension is the 3-points correlation

function which contains all the informations needed for the reconstruction of the bi-spectrum. With optimized algorithms (which perform a smoothing of data on a scale of about 4h–1Mpc), its estimate on large scales (100h-1Mpc, i.e. the order of the scale of BAO) can take up to 46 days on one thread (MARRSD) for a volume corresponding to Euclid. The same execution speed is expected for the analysis of N-body simulations that correspond to the same volume as Euclid. Assuming that the analysis of covariance between measurements can be performed on a limited number of simulations (100), taking only 10 configurations of the 3-points correlation function on large scale (50 - 100 h-1Mpc), with 3 MARRSD (120 threads) a comprehensive analysis can stagger (data + simulations) for a period of one year.

Furthermore, the analysis of the galaxy field via higher order statistics can be carried out in a complementary manner using the 2-points cumulants of order (n + m), which contains integrated information about the N-points correlation function. This new statistic, mostly developed at CPT, is of major interest because by construction it is insensitive to certain forms of bias between galaxies and matter (such as non-linear and non-local biases). As such, it allows to study more precisely how galaxies are connected to the underlying dark matter field. Knowing that the study of the bias between galaxies and dark matter includes a large number of parameters, the minimum number of simulations to be analysed would be 1000. The calculation of correlators (up to the order n + m = 5) takes about 2 days on a single thread (MARRSD) for a Euclid type simulation. We can then test ten correlation scales (typical analysis) over a period of one year with 2 MARRSD (80 threads).

2.1.9 N-body simulations

In order to prepare the analysis of future high redshift surveys it appears necessary to be able, independently, to generate a large number of simulations of the large scale structure of the universe with characteristics tailored to our needs (that is implementing prescriptions for massive neutrinos and modified gravity). In this perspective it is necessary to obtain a total of 800 threads (10 GB of RAM per thread). To allow the production of competitive simulations in terms of resolution, it takes a number of dark matter particle of the order of 2048^3, which must be simultaneously stored in the RAM. In addition, one should identify dark matter halos and save them continuously during the running of the simulation. This requires a total of 8000 GB of RAM, then with 800 threads (and 10 GB of RAM per thread), one can obtain in 21 days a reasonable resolution for a single cosmological simulation. For comparison, such simulations require only 2 days to run on the machines involving many cores (8000), like for example the machines of CINECA (Bologna), which are among the most powerful in Europe. One can thus expect a dozen simulations over a year. To save 8 redshifts per simulation one requires a total of 24 TB of disk storage. This large amount of disk space implies that the generation of our own simulations is a net advantage to save time on data transfer. We estimate for this item that at least 3 MARRSD are necessary (when taking into account the request of items 3-6).

2.2 Hardware configuration, budget and deployment plans

2.2.1 Overview of actual computing resources

The main computing resources for the dark energy group are summarized as follow:

- LAM has a cluster with 338 threads distributed in 31 nodes (8 or 12 cores each) Each core has 4 GB of RAM, i.e. between 48 and 32 GB per node. The LAM cosmology group uses about 1/3 the overall with a low access priority. The disk storage is about 10 TB.
- CPPM dark energy group has 128 threads distributed in 5 nodes. The available memory per thread is in between 1 GB to 10 GB. The disk space is about 45 TB.
- CPT has 128 threads on 4 nodes with 1 GB per thread and 300 GB of disk space.

The total amount of computing power then amounts to at most 600 threads and 60 TB of disk space, spread over 3 laboratories. Other laboratories involved, namely IRAP and LUPM, have now very limited resources associated to the dark energy groups.

2.2.2 Requested ressources

The overview of needed resources is shown in table 1. The total number of MARRSD is 28 corresponding to 1120 threads.

	MARRSD	Disk space (TB)
Supernovae	2	13
BAO	5	60
Cosmic voids	1	1
Lensing	1	60
RSD	3	Shared
MCMC	6	Shared
Phenomenology	2	5
Higher order	5	Shared
N-body	3	24
Total	28	163

Table 1: Overview of resources needs

The proposed hardware configuration answers to many specific needs. The main one is reactivity to fight against the international competition. Each time a set of new public data is available, new analysis and potentially new accurate results can be obtained and should be done in a short time. A good example is combined analysis with Markov chain Monte Carlo. In such a case parallelism with many processors should be possible and requires high interconnectivity between the different nodes. The second one is memory per thread. Power spectrum reconstruction needs to work with huge catalogues that should be uploaded to the internal memory to avoid input/output saturation.

For all these reasons we propose to use a flexible hardware configuration: At least 1000 threads with 10 GB of memory per thread, an InfiniBand bus to connect nodes for parallelism and at least 200 TB of disk space. Details can be found in table 2.

In cosmology the object of our studies is the whole Universe. As a consequence, we can't use standard parallelism as is for particle physic experiments. Software architectures like grid or cloud are not suitable in development or debugging phases because even in these phases high parallelism is required. Because of this specificity we propose to use mixed software architecture. A part of the DEC will be dedicated to interactive jobs, while the remaining machines will be used via a batch system. This batch system should be able to share parallelism on many different machines. The part dedicated to interactive jobs will be progressively adapted to the need of users.

Resource	Description	Reference	Units	Price (k Euros) P.U.	Total
CPU	560 cores HT	DELL C6300 + 4× C6320 (28 cores) †	5	50200	251000
Storage	head node	DELL D630	1	2	2
	disk server (>200 TB)	DELL Powervault MD3460 ‡	1	26	26
RAM Server	1.5 TB RAM	DELL R930 (24 cores) ⋆	1	26	26
InfiniBand	switches	Intel True Scale Fabric Edge Switch 12200BS01			30
Extra	UPS or rack		1	20.5	20.5
Total					355.5

† 2 × Intel Xeon E5-2695 v3 @2.4 GHz (14 cores, 28 threads) with 512 GB of RAM
‡ disk server filled up 60 × 4 TB disks used in RAID 6
⋆ 2 × Intel Xeon E7-4830 v3 @ 2.1GHz (12 cores, 24 threads)

Table 2: Hardware for the Dark Energy Centre.

2.2.3 Installation planning

The physical installation of the Dark Energy Centre is possible in 3 different institutes. By order of preference:

- CPPM: The current infrastructure of the CPPM can host the Dark Energy Centre without extra cost. However in case the FEDER project is accepted no more room will be available.

- CPT: In case physical space is not enough in CPPM, CPT can host this new centre. An additional UPS is necessary corresponding to an extra cost of ∼10 kEuros.

- LAM: This last possibility requires a new electrical power line (20 kEuros).

It must be noted that for each of the three possible laboratories, the task force has been in touch with the local farm system administrator to discuss with them the different opportunities. In addition, all of them have mentioned that they would be able to ensure a minimum part of the maintenance of the DEC machines, but that they would obviously need part time man power support to provide a really reliable service to the DEC users.

In case funding should be spread over several years, we propose the following schedule, also summarized in table 3. In 2016, we first install the infrastructure, 1/3 of CPU and disk storage. The full computing power and the RAM server will be installed in 2017. Appendix A presents the budget plans for the LUPM–*Cloud* and DEC projects.

Ressources	% in 2016	% in 2017	kE in 2016	kE in 2017
CPU	30%	70%	100	151
Storage	100%	0%	28	0
RAM Server	0%	100%	0	26
Network	100%	0%	30	0
Divers	100%	0%	20.5	0
Total			178.5	177

Table 3: DEC budget planning

3 LUPM–*Cloud* project

The LUPM–*Cloud* project aims at providing the Labex research teams with a new cloud platform built on top of reliable and powerful new hardware. The LUPM grid computing farm has been reoriented toward cloud computing a couple of years ago and now actively participates (with limited resources though) to the Federated Cloud project lead by France–Grille. Virtualisation has already lead to an increasing use of computing resources by LUPM research teams, in particular by the EMA (*Expérience et Modélisation en Astroparticules*) and Theory groups who are much involved in OCEVU. With a more reliable and more powerful infrastructure, it will be possible to propose an access to the LUPM–*Cloud* platform as a service (PaaS) to all OCEVU physicists.

3.1 Scientific case and associated computing resources

3.1.1 Scientific projects

This section describes a few concrete examples of cloud computing at LUPM by different research groups.

The Fermi Gamma-ray Space Telescope is operational since June 2008, and has provided so far a wealth of data publicly available, almost a billion candidate photons. Fermi science is now very much oriented towards the production of catalogues of sources (all sources, gamma-ray bursts, AGNs or supernova remnants…), or towards getting a deep in-sight on some peculiar region of the sky by using more than 7 years of data. Algorithms and analysis techniques are each time more powerful, this particularly true to fully exploit the new set of photon event classes distributed by the Fermi–LAT collaboration in the so–called *Pass 8* framework. Computing resources required to achieve the analysis of these data has increased consequently, and it's mandatory to be able to get quick turn–around in order to develop new techniques and tune analysis parameters for the most convincing results and estimates of systematic uncertainties. In parallel, the Fermi ScienceTools development has slow down significantly, and the software that was initially developed more than ten years ago, is starting to get quite old, with many hard to maintain dependencies. For all these reasons, the Fermi group at LUPM has proposed to develop a new cloud based computing model for Fermi high level data analysis done with the official ScienceTools. The goal is to design one or more virtual machines on which the ScienceTools software would be installed and tested : a first version would allow physicists to develop their analysis and test their tools, even providing a graphical user interface if needed, and a second minimalist version, but entirely compatible with the first one, would be used for intensive cloud computing (dozens of these could be instantiated in parallel). This project is well underway, and the first cloud image is being tested right now.

The Cherenkov Telescope Array (CTA) has had a pretty intense year in 2015, with a couple of major events: The first one was the successful handling of the Critical Design Review in late June, and the second one was the choice of the building sites, at Paranal (ESO–Chile) for the southern hemisphere and La Palma (Canary Island, Spain) for the northern hemisphere. The LUPM is very much involved in CTA computing, leading the Monte Carlo simulation production on the EGI grid with more than 8000 cores (peak value) available on 20 sites, and almost

2 PB of storage distributed on 6 sites. Indeed, L. Arrabito, IR at LUPM, is the technical coordinator of the CTACG project for the management of CTA grid ressource, and also leads the development of a computing resource management solution for CTA, taking advantage of the DIRAC framework. The EMA team also contributes to the analysis of simulated data, and had a Master 2 student working on extracting the instrument angular and energy responses from different arrays simulated for the so–called PROD2 in 2015. This work is done in coordination with the CPPM CTA group within the CTASci Labex project. The CPPM group has so far mostly been involved in the fine tuning of the small telescope electronic simulation, but with a Labex postdoc starting in October 2015, the idea is to put more work into understanding the array response, and also participating to the data analysis pipelines development. In 2013, the grid node MSFG at LUPM was decommissioned because of a combination of hardware getting too old, and not reliable enough for the limited man power available. The LUPM farm is hence unfortunately out of the production scheme for CTA, however with brand new hardware and cloud based solutions this could easily change. Indeed, resources used for the CTA Monte Carlo simulation productions are standard grid resources, but the DIRAC framework is flexible enough to interface via a plug-in system to any kind of computing site, from a simple batch farm to a computing cloud. The aim of the LUPM cloud computing project for CTA is to anticipate the global transition from standard grid nodes toward a federated cloud infrastructure, relying on local competences in cloud computing, DIRAC and a brand new computing infrastructure. Virtualisation will be useful to solve some CTA specific requirements, like the need for at least 8 GB of RAM per core (more like 16 GB for software testing), or scratch disk space for around 200 GB: indeed, virtual machines can be easily configured with *ad-hoc* characteristics. With refurbished hardware, the LUPM farm will get over a critical threshold of 300 cores and 100 TB of redundant disk so as to be integrated in the regular pool of computing resources used in production by CTA. Given the intense I/O throughput required for the analysis of Monte Carlo data, it's much more efficient to have computing resources and storage available on the same node. This project will benefit primarily to the CTASci labex team that will have an easy access to CTA compatible virtual machines, computing resources and simulated data. This is particularly true for the CTASci postdoc, C. Trichard, who started to work on October 2015, at CPPM with H. Constantini on the reconstruction and data analysis pipeline: a project in which J. Bregeon at LUPM is also involved.

Besides, on a medium term, the CTASci team could also benefit from the Fermi science tools virtualisation experience, by using a similar computing model to run the CTA science tools on the LUPM computing cloud. This is particularly true for the IRAP members of the CTASci project who lead the development of the CTA science tools. But, in a wider context, the HESS experiment also now has internally produced high level fits files format for all HESS-I data and a complete scientific analysis chain that runs on these fits files, in a very similar approach to the Fermi and CTA science tools. A foreseen interesting work is to build up a virtual machine dedicated to combined multi–instruments analysis (HESS/Fermi/CTA), with all scientific software readily available. This idea is to be followed up by HESS members of the CTASci team, and should be part of the PhD work of J. Devin who just started at LUPM with M. Renaud, but should eventually benefit to a wider range of people in the OCEVU LabEx interested in high energy and very high energy gamma-ray data.

Theorists of the Interactions *Fondamentales, Astroparticules et Cosmologie* (IFAC) group are also getting into cloud computing at LUPM, in order to achieve specific calculus. C. Hugonie, in particular, has been using up to 3 virtual machines in parallel, each setup with 6 cores and 12 GB of memory in order to run NMSSM computations over several days. Two paper have already been published based on these results (B. Allanach et al 2015, and U. Ellwanger and C. Hugonie 2014), and a third one is in now pre-print. This type of activity is expected to take more and more importance in the future, thanks to the flexibility of virtualisation that can provide the best hardware configuration corresponding to one needs.

Overall, this last statement also clearly shows that man power is needed to get along with the LUPM–*Cloud* hardware upgrade, in order to fully exploit the potential of the new available platform, and answer quickly and correctly to the scientists requests. Indeed, most users do not want to handle by themselves the configuration and administration of the virtual machines by any mean, so that user requests should be taken care of in a PaaS computing model: this means providing a service in the form of one or more virtual hosts, with correct amount of computing power, memory and storage, and with pre-installed readily available software. Hence, the work to be done will be to promote the use of cloud computing, get in touch with physicists to understand their needs, design and instantiate the most appropriate virtual machine configuration, install and configure the software needed on the machine. Several iterations between the physicist and the engineer will be needed to achieve this kind of task, and we can also foresee that when the work is done, feedback to improve both the cloud services and the virtual machine setup procedure will be of great importance.

3.1.2 Common needs

Virtualisation brings a large amount of flexibility in the usage of computing resources, and that is essential for the LUPM computing services group to give a quick answer to users request in terms of computing power, storage or services.

Students of the "Cosmos, Champs et Particules" master at the Montpellier University, have been on the front line to benefit from this flexibility. For a couple of years now, students advisors and professors are invited to send in advance their needs to the computing services group, in particular for M1 and M2 internship. Requests are then translated into a number of virtual hosts instantiated from standard images, or even directly from images given by the advisors (e.g. Fermi science tools). Virtual machine images may be stored to be re-used and/or improved the year after. If more than one student work on a given project, identical virtual hosts can be easily instantiated providing an homogeneous working environment to all students involved.

Here is a non-exhaustive list of works realised on virtual hosts by master students in the past couple of years:

- astronomical modelling with MESA, 4 virtual machines for 10 students in 2014
- Fermi gamma-ray burst analysis using the *official* Fermi virtual machine, M1 internship, in 2014
- low level data analysis of CTA Monte–Carlo simulations, 1 virtual machine and access to the grid user interface, M2 internship in 2015

Improving the computing cloud infrastructure, and having man power as an *Ingénieur d'Etude* partly devoted to interactions with physicists will vastly improve the level of service. This will make it possible to be very responsive to very specific requests such as testing software framework in development or production phase, what may be of particular interest to scientists who have recently been involved in new projects (SVOM, LSST, ...).

The LUPM computing services group also provides a virtual host that is configured as a grid user interface, used a lot in particular for the CTA project. The host image is based on work done during a grid workshop in 2014, and is regularly updated. A few TB of disk have also been made available on that machine so that user can download large amount of data from the grid storage elements when needed. In fact, it must be noted that as the grid user interface, most of the internal services provided by the computing group are virtualized, including owncloud, redmine, svn, web server, application and file servers etc...

3.2 Hardware configuration, budget and deployment plans

3.2.1 Overview of HPC computing in Montpellier and historical view of HTC computing at LUPM

Montpellier hosts 2 HPC (high performance computing) centres: the CINES is a national wise service, while HPC@LR operates a the regional level. Both centres are focused on highly parallel computing, providing access to very specialised machines, of different power though. HPC@LR hosts around a thousand computing cores, where the CINES has as many as 15000 cores.

In 2008, the LUPM computing service got involved in grid computing being part of the EGEE grid, with an infrastructure based on 256 cores and 10 TB of disk storage. At the end of 2013 though, the grid node has been decommissioned, and LUPM entered the France–Grilles Federated Cloud project, as the LUPM–*Cloud*. Today, LUPM–*Cloud* is operating a dozen of hypervisors with 2 CPU for 4 to 6 physical cores with hyper–threading, for a total of around 180 computing cores available for around 30 virtual machines available. All the pieces of old grid hardware have not been recycled for cloud computing purpose, for performance and maintenance issues.

3.2.2 Hardware, deployment and budget

Current hardware needs concerns an upgrade of the internal network, the installation of a real disk storage capacity, and a significant increase of computing power in order to be more in line with the nominal grid node capacities while integrating the new requirements due to the developed cloud technology.

For the computing power, the goal is to bring back the farm to at least it's nominal capacity, including a reasonable margin to insure a stable capacity for the next 5 years, so that new users can be sure of the perennity of high quality services. As shown in table 4, two-thirds of the budget will be allocated to computing servers, as an example a DELL C6300 crate filled with 4 C6320 blades with low consumption CPU, can provide 96 cores HT for 24 k Euros: 3 of these provide 188 cores HT for 72 k Euros.

For storage, the issue is much more drastic since the available disk space in the LUPM–*Cloud* is today less than 10 TB. In a coherent way with the increase in computing power, the goal is to rise the amount of available disk by an order of

magnitude while still providing reliable redundant storage. The solution could be based on just one DELL Powervault MD3060e server with 240 TB of raw disk, corresponding to roughly 200 TB of available space when considering RAID 5/6 technology, for a cost of 24 k Euros.

Eventually, the infrastructure will greatly benefit from a number of high bandwidth switches (10 Gb/s) in order to provide fast Ethernet connection between the different machines, in particular between the computing and storage nodes.

The total estimated budget of the operation amounts to roughly 100 k Euros: table 4 gives the details of the needed hardware, based on DELL tenders dated October 10th 2015.

Resource	Description	Reference	Units	Price (k Euros)	
				P.U.	Total
CPU	288 cores HT	DELL C6300 + 4× C6320 (24 cores) †	3	24	72
Storage	head node	DELL D630	1	2	2
	disk server (>200 TB)	DELL Powervault MD3060e ‡	1	24	24
Network	10 Gb/s switch	DELL M8024-k (24 ports)	2	2.2	4.4
Total					102.4

† 2 × Intel Xeon E5-2650l v3 @1.8 GHz (24 cores, 48 threads) with 256 GB of RAM
‡ disk server filled up 60 × 4 TB disks used in RAID 6

Table 4: Hardware request for the LUPM–*Cloud* project

Ideally, the LUPM–*Cloud* project would prefer to have the money available as soon as possible in 2016, in order to deploy quickly the new hardware, and finalize the transition to the `OpenStack` cloud framework in good conditions. The CDD IE would then be of great help to assist the LUPM computing administrator from the beginning of the deployment. In case, the budget has to be spent over 2 years, 2016 and 2017, ∼75 k Euros would be needed in 2016 to buy 200 TB worth of disk storage and 192 cores, and the 25 k Euros left in 2017 would be used for further computing power. This strategy is summarized in the following table 5.

Ressources	% in 2016	% in 2017	kE in 2016	kE in 2017
CPU	30%	70%	48	24
Storage	100%	0%	26	0
Network	100%	0%	4.4	0
Total			78.4	24

Table 5: LUPM–*Cloud* budget planning

Appendix A presents the budget plans for the LUPM–*Cloud* and DEC projects.

From the point of view of deployment and availability, the idea is first to transfer all existent services to the new platform, and then make it available to new users at LUPM, and to a restricted number of OCEVU users who work closely with LUPM physicists (e.g. members of the `CTASci` team, but not only). Then,

provided that the man power is available, and once the new LUPM–*Cloud* platform will be running smoothly, the PaaS will be opened to all OCEVU members.

4 Synthesis of man power needs for support

Creating a new dedicated computing centre like the DEC, or providing a new service platform like the LUPM–*Cloud*, first relies on a significant amount of raw computing power, but it's obviously not enough when such a large number of machines are concerned. Indeed both projects will also require a certain amount of man power to first setup the hardware, and then make sure that users are indeed able to use it smoothly. We will now review the details of man power need for support for both projects.

As mentioned in section 2.2.3, all system administrators of the local computing farm of the laboratories that could host the DEC, have agreed to provide minimum help for the installation and maintenance of the machines. However, they have also all mentioned, that they would probably not have time to do much more, i.e. to take care of the users. The primary responsibility of the engineer (IE) will be to contribute to the hardware installation of the DEC, under the supervision of the local system administrator. It involves organization of the computing room, verification of the power supply and cooling and integration of machines and disk. This being done, the engineer will install and configure the network in particular the InfiniBand bus that will require particular attention. After verification of the hardware configuration, the engineer will have to install and configure all the needed software to make the operating system working. In interaction with researchers, he will install and maintain scientific libraries, compiler and other software facilities. He will be responsible to create users account and manage user disk space and backup as well as data disk space. In a second phase, when the system is stable, he will have to control the overall resources of the DEC in a flexible way to allow specific computing for each potential user. For this, he will have to investigate with users the best strategy and software. Some software development can be necessary. In the ideal case, the engineer should be able to interact with researcher to help them to optimize specific software, in particular high parallelism algorithm. Finally, he will have the full responsibility to make the Dark Energy Centre to work safely and efficiently in agreement with users and with the local system administrator.

The LUPM–*Cloud* project would also benefit a lot from the support of an IE from the beginning of the deployment. The first task will be to assess the quality and performance of the available hardware, retire old outdated machines and reorganize the computing room to make space for the new machines. The IE will then contribute to the installation of the hardware, and help with the design of the network: both the hardware network to optimize connections between the different elements (computing, storage and head nodes), and the (many) virtual networks used within the OpenStack cloud framework. Once the new infrastructure will be in place, work will be needed to move all the existing cloud services to the new OpenStack platform: this will be a key work for the IE who will have to interact with the scientists in order to re-assess their needs, and potentially take the opportunity to propose an upgrade of the virtual hosts. To finalize the transition, the IE will also participate to the integration of the new LUPM computing cloud to the France–Grille Federated Cloud project. This first phase will probably last around half a year, and will let the IE to complete his cloud computing knowledge and get acquainted with the working environment at LUPM. Once all

already existing services will run smoothly in the new infrastructure, the IE will be in charge of promoting the new platform within the LUPM. New services will then be designed and instantiated by a close interaction between the IE and the group or persons concerned, showing again that the IE profile has to be technically very qualified in terms of service virtualisation, but also good at team work. In routine operation, following-up with the different projects will then become the main activity of the IE. He or she will have to make sure that all virtual hosts are running smoothly, provide when needed high availability services and keep on a continuous interaction with research groups in order to improve the service quality through virtual hosts upgrades or software setup. Once the PaaS is proved to be stable and is able to satisfy most users at LUPM (and closely related groups), services shall be opened to all OCEVU research teams. This last phase transition may hopefully happen within one year of the project start. The IE will then also have to work in close contact with other OCEVU institutes involved in cloud computing in order to understand how to distribute services among the different clouds: at this point in particular, the CPPM cloud shall be just starting its deployment based on new hardware funded by the CPER (and FEDER) plans.

As a summary, the IE will first have to assist both projects with hardware and software setup, and second to provide user support in order to get the most out of the available computing power. In routine operations, the two projects will require different level of support though. On the one hand, the LUPM–*Cloud* PaaS should be a very dynamic service that will require many interactions with users, and continuous developments of new virtual hosts. On the other hand, the DEC will have to be much more static as the main challenge will be to keep machines as stable as possible in a given configuration in order to run calculus for several days, weeks, months. Even though, we understand that the natural physical location of the IE is as close as possible to the biggest hardware architecture (DEC and/or Meso-centre in CPPM), we believe the IE might be much more useful if physically located in Montpellier to support the complex architecture of a cloud project.

Although remote interventions on the DEC and on the LUPM–*Cloud* will be the norm, the IE will very likely have to do a certain number of missions between Marseille and Montpellier each year, in particular to be able to interact directly with the physicists he will be working with. It would then be very welcome if the position could come with a reasonable amount of money for travel.

Technical details of the IE profile are available in the *fiche de poste* in appendix B.

5 Conclusions

Both projects presented here are decidedly ambitious both in terms of computing techniques and power, and of scientific impact for OCEVU. By defining computing needs, and establishing the best balanced budget, it appeared that both projects would clearly benefit of being granted the totality of their funding to maximize their impact. Additional help as a 3 years fixed term contract for an IE also appears quite mandatory to get the full benefit out of the hardware investment, and also avoid local system administrators to become completely overloaded.

The two projects defined each their specific requirements in terms of computing power and organization, but there is no doubt that we may eventually look for dark energy in the cloud… or for gamma rays in the DEC.

A Hardware and budget plans summary

The following table 6 and bar chart 2 present two different possible scenarii for the projects funding. The first scenario just considers everything funded in 2016, and the second scenario is based on funding split half in 2016 and half in 2017. Note however, the LUPM–*Cloud* project would much rather get the full funding on the first year in order to maximize the impact of the new proposed platform, in particular by avoiding new cloud users to be limited by computing resources.

Furthermore, as is always the case for this kind of hardware, costs from now to 2016/2017, may vary by 10% to 30%, very likely, but not necessarily, in the right direction depending upon the international context.

	Scenario 1		Scenario 2	
	2016	2017	2016	2017
LUPM–*Cloud*	100	0	78.4	24
DEC	350	0	177.5	172.5
Total	450	0	250.9	196.5

Table 6: Table of the proposed budget plans

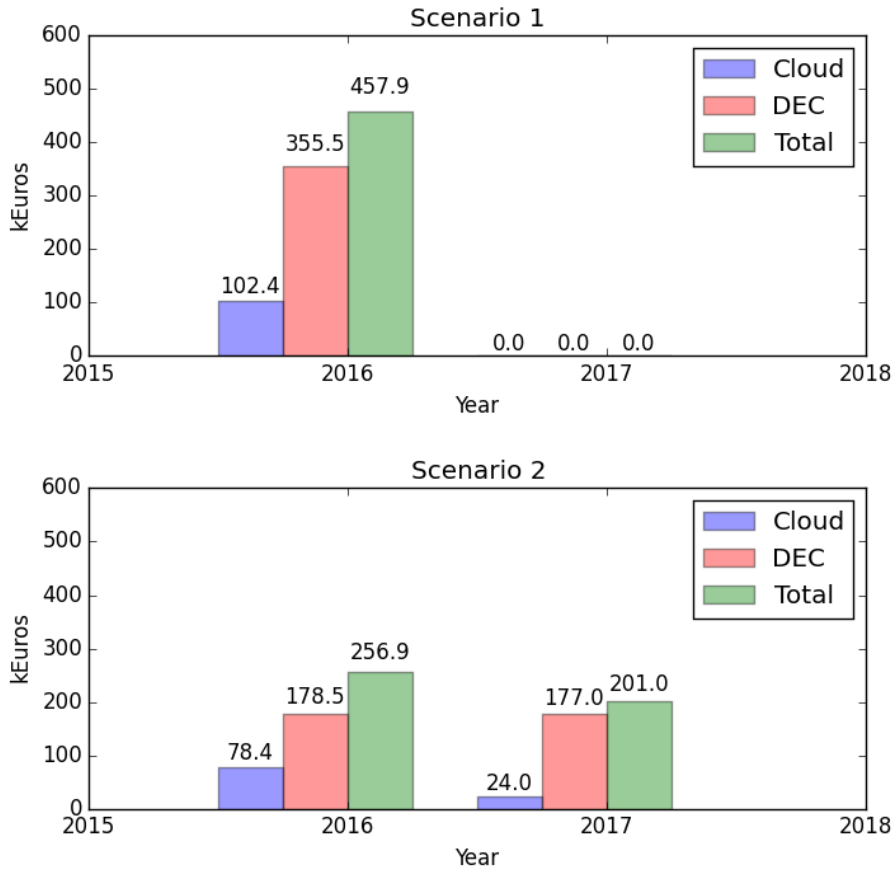

Figure 2: Bar chart for the proposed budget plans

B Fiche de poste IE

Destinataire de la demande :
Labex OCEVU

Description de la demande:
Motif : Accompagnement des projets OCEVU LUPM–*Cloud* et *Dark Energy Centre*
Corps : Ingénieur d'Etudes
BAP : E
Emploi-type : Ingénieur des systèmes informatiques, réseaux et télécommunications
Quotité : 100

Description des missions :
L'ingénieur d'étude (IE) aura pour mission d'accompagner les projets du LabEx OCEVU LUPM–*Cloud* et *Dark Energy Centre* (DEC), qui prévoient la mise à disposition de nouvelles ressources de calcul et services aux membres d'OCEVU. Pour les deux projets, l'IE aura pour mission première d'assister les administrateurs locaux pour l'installation, la configuration et la mise en réseau des machines dans les fermes de calcul, suivie de l'installation des logiciels adéquats.

Pour le *Dark Energy Centre*, l'IE aura pour mission la gestion administrative des machines (gestion des comptes, système de *batch*, sauvegarde des données, ...), ainsi que la mise à jour des logiciels et bibliothèques, sur demande des utilisateurs. Le DEC sera physiquement installé à Marseille dans l'un des trois laboratoires suivants: CPPM, CPT, LAM.

Pour le projet LUPM–*Cloud*, l'IE aura pour mission d'assister le service informatique du LUPM à Montpellier pour le déploiement et le développement du cloud OpenStack dans le contexte du LabEx OCEVU et dans le projet de Cloud fédéré France-grilles. L'IE sera le contact privilégié des scientifiques du LabEx OCEVU pour l'utilisation des services du cloud, chargé de promouvoir les services disponibles, de comprendre les besoins des utilisateurs et de leur fournir la meilleure solution possible.

Description des compétences :

- maîtrise de l'environnement Linux,
- connaissance approfondie des concepts et techniques d'architecture des systèmes et réseaux,
- connaissance approfondie des techniques de virtualisation,
- connaissance des problématiques liées au stockage et à l'accès aux données,
- connaissance des problématiques liées au calcul distribué
- bonne connaissance des langages de scripts,
- des connaissances en programmation pour la parallélisation des algorithmes seraient appréciées,
- pratique courante parlée et écrite de l'anglais,
- goût du travail en équipe.

Description du contexte :
L'activité DEC s'exercera à Marseille, *a priori* au sein du laboratoire dans lequel seront installées les machines : CPPM, CPT ou LAM.

L'activité LUPM–*Cloud* s'exercera au sein du Service Informatique du laboratoire Univers et Particules de Montpellier (UMR 5299). Le laboratoire LUPM est une unité mixte de recherche dans le domaine astrophysique, astronomie et physique des particules (http://www.lupm.univ-montp2.fr). Regroupant environ 60 chercheurs, enseignants-chercheurs, ingénieurs, techniciens, administratifs, le LUPM est structuré en 3 groupes de recherche. Il a pour tutelles le CNRS (IN2P3) et l'Université de Montpellier. Il est également rattaché à l'OSU OREME. Le Service Informatique du LUPM est composé de 5 ingénieurs. En tant que service technique, il a en charge l'appui à la recherche dans les domaines suivants : activités scientifiques, calcul distribué, administration système et réseau et support utilisateurs.

La séparation des activités sur deux sites différents implique d'une part qu'une grande partie du travail devra être réalisé à distance (via connexion réseau), et d'autre part que l'IE aura a effectuer des missions fréquentes entre Marseille et Montpellier.

Profil recherché :
Le profil recherché est celui d'un ingénieur système et réseaux, avec une expérience significative dans le domaine du calcul distribué. Idéalement, le candidat aurait aussi des connaissances en programmation lui permettant d'assister les chercheurs souhaitant paralléliser leurs algorithmes.

Informations complémentaires :
Equipe(s) concernée(s) :

- Nicolas Clémentin - Service Informatique du LUPM
- Thierry Mouthuy - Service Informatique du CPPM
- Thomas Fenouillet - Services informatiques du LAM
- Jean-Roch Liebgott - Services informatiques du CPT

Commentaire/Justification :
Le recrutement de l'ingénieur sur un contrat de 3 ans permettra de rentabiliser au maximum l'investissement financier significatif du LabEx OCEVU pour la création du *Dark Energy Centre* et de la PaaS LUPM–*Cloud*. Son travail permettra en outre d'éviter une surcharge de travail pour les administrateurs systèmes des laboratoires concernés (LUPM et CPPM/CPT/LAM).