IoA, Cambridge Expression of Interest:

LSST:UK Phase B DAC WPs A,B,D,E

Proposing team: Richard McMahon (IoA), Nicholas Walton (IoA), Paul Alexander (SKA, Cavendish), Manda Banerji (IoA), Rosie Bolton (SKA, Cavendish), Vasily Belokurov (IoA) Paul Calleja (HPCS), Jeremy Coles (SKA/GridPP, Cavendish), Sergey Koposov (IoA /CMU)

Staff Request: Total 61sm (split: WP-A 4sm; WP-B 21sm; WP-D 12sm; WP-E 24sm). One software developer (yr1: 50%, yr2: 50%; yr3 75%; yr4 100% FTE) to deliver WP-B and WP-D tasks. One application scientist will deliver WP-E tasks (yr1, 2, 3, 4 at 50% FTE). Management provided by Walton and McMahon (4sm in WP-A).

Rationale: ***LSST:UK Access to L1/L2 and L3 data products via a distributed DAC:EU+L3 DAC model.***

We propose a centrally managed distributed LSST:UK DAC facility, modelled on the highly successful STFC DiRAC facility, that (a) leverages the technical and scientific expertise in UK institutes and (b) leverages an emerging pan-European L3DAC model that benefits from the LSST:France investment in a full L1/L2 DAC. As envisaged prior to the LSST:UK Phase A submission, a recent game changing development is that LSST:France will, in addition to processing 50% of the LSST Level-2 data[1], host a complete copy of all LSST data, deploying a full DAC serving both the French and the wider European LSST community. The scale of the CC-IN2P3 Lyon (DAC:EU) would be similar to NCSA's DAC which provides access to the entire USA community.

In order to provide cost effective delivery of L3 data products, and a mechanism to ensure quality assured production of those L3 data, an elegant model linking a small number of Level 3 Data Analysis Centre's (L3DACs) to the CC-IN2P3 DAC (DAC:EU) is being developed in conjunction with LSST:FR and CC-IN2P3[2]. The L3DACs located across Europe, will be constituted to optimally run the L3 processing chains developed through science focused L3 algorithm projects. In the UK, this would entail one or more L3DACs[3] deploying the L3 chains developed through UK DEV activity. L3DAC:Cam would deploy both the Cambridge L3 chains (e.g. Star-Galaxy separation, Stream Finder, Forced Photometry), and other UK L3 algorithms requiring access to ancillary data held within the Cambridge data infrastructure (e.g. VISTA, Gaia and many catalogues through the WSDB[4]), for instance 'Crowded Field X-Match' proposed by Exeter).

Each UK L3DAC will both generate L3 data products, and provide data access to those L3 data products that are not released through the main DAC:EU L2/L3 repository (e.g. the L3 data product does not have a sufficiently wide range of potential end users to warrant ingression into the main LSST data system). Access will be via deployment of the main DAC

---

[1] Recent discussions with Dominique Boutigny (LSST:FR PI) confirm that the Lyon DAC is also likely to hold the LSST L1 alert stream.

[2] Additional funding of €6M for the DAC:EU facility (to allow expansion of DAC:FR to support European LSST access) will be provided via H2020 funding. An initial bid is being developed, led by LSST:FR and including UK, DE, ES and IT partners, for submission in March 2018, for funds from early 2019, hence in line with LSST:UK Phase B timelines.

[3] We anticipate that a second UK L3DAC would focus on other L3 science chains e.g. real time alerts, taking a smaller volume of data could be a single use case deployed standalone or at either UK L3DAC, directly linked to the DAC:EU real time L1 alert server.

[4] WSDB = Whole Sky Database: high performance system providing SQL access to most large astronomy UVOIR catalogues e.g. SDSS, Pan-STARRS, VISTA, DES

data access stack, this LSST Science Platform providing for example Jupyter client/server and programmatic API access to the data.

Preparatory throughput testing ('perfsonar') shows sustained transfer speeds today in excess of 3 Gbps being achieved between the IoA, Cambridge and CC-IN2P3, Lyon (demonstrating the feasibility of L2/L3 staged processing: so can already transfer one complete night of LSST raw data (~15TB) in ~10 hours). A number of science pilots are underway, testing the performance of image based and catalogue based data processing between CC-IN2P3 and IoA (this pilots the DAC:EU - L3DAC:Cam model) e.g. interfacing to the Lyon hosted Qserv development system with 400 cores, 800 GB memory and 500 TB storage hosting LSST stack reprocessed products such as HSC. In addition tests leveraging the SKA-SDP ALaSKA openstack testbed (which includes a high performance NVME/SSD accelerated disk array) are being investigated.

Our proposal implements a more cost effective alternative to the current baseline DAC:UK model, which assumes UK hosting all L1/L2 data, as outlined in the EoI call documentation and Phase A proposal. In our model, the UK will leverage the significant investment in LSST Level 2 processing and a European based copy of all L1/L2 at CC-IN2P3 in Lyon. The UK will not need to set up a full L1/L2 DAC along with its fixed costs for associated infrastructure (including significant power and cooling costs), to support a relatively small user community, at a correspondingly high DAC cost per user. Instead, UK resource can be diverted to delivering UK priority L3 data products (which require tailored access to complimentary data resources served by the UK L3DACs), whilst providing high performance UK community access to LSST L1/L2 data products through the DAC:EU[5]. ***Because effort is not required in developing and deploying the L1/L2 UK DAC, 36sm in Phase B alone can be diverted from the UK Phase B DAC to DEV area in our model***. Phase B scoping will include L3DAC hardware scoping for Phase C & D, and deployment of an operational L3DAC to support early Phase 1 ComCam and Phase 2 LSST full camera commissioning and science verification data product releases from 2020/21.

Work package description and justification of resources

WP-A DAC Management: [4sm] Managerial direction of L3DAC development, and interaction with the wider LSST:UK project. Will liaise with National e-Infrastructure also via Cam HPCS. Will provide a point of contact to the DAC:FR EU HPC CoE and European Open Science Cloud (EOSC) programmes. Will ensure delivery of a suite of L3 data products, together with access to those products and associated ancillary products, and through DAC:EU UK access to all LSST L1/L2 data. Preparation for DAC operations in Phase C/D.

WP-B Data ingestion and preparation: [21sm] All Level 2 data product ingestion is handled by DAC:EU at CC-IN2P3 in Lyon, hence only limited WP-B resource is required to support data ingestion and preparation of UK L3 data products. This includes ingestion of ancillary data required by the L3DAC:Cam suite of L3 applications, including e.g. VISTA, DES, Gaia etc. catalogues to facilitate the L3 processing. The Cambridge WSDB will be utilised to support ancillary data access and will incorporate an extended range of multi-wavelength data from the radio (e.g. SKA pathfinders: ASKAP, MeerKAT) and X-ray (e.g. eRosita) to support L3 processing requirements. All L3 data products will be suitable annotated to enable effective science utilisation by the UK community.

WP-D Infrastructure: [12sm] All L1/L2 DAC infrastructure is provided by DAC:EU in Lyon. The support requested here is to allow for the procurement, installation and maintenance of L3DAC hardware hosted at the HPCS and co-located or integrated with the STFC/EPSRC

[5] CC-IN2P3 locally host a number of key data (Euclid) and provides high throughput links to a range of ESA and ESO data products.

Tier 2 facility as part of the National e-Infrastructure and related LSST stack software. L3DAC:Cam will include operational support of the WSDB, a high performance database installation hosting multiple high value sky survey catalogs, which will be integrated into the processing chain for a number of the L3 analysis chains to be implemented. Phase B effort will develop a technical roadmap to scale the provision of the L3DAC infrastructure for support of future LSST software releases. The L3DAC infrastructure will benefit from existing expertise in Cambridge (Gaia; CASU for VISTA, WEAVE, 4MOST, PLATO; Cambridge HPCS; Cambridge CSD3 UKT2 facility, Cambridge Exascale HPC/Data processing design work for the SKA within the SDP consortium)

WP E Science Support: [24sm] Supporting design of L3 data analysis chain for deployment in the L3DAC. This will include interfacing with the L3 DEV teams defining the algorithms from both Phase A and Phase B. We will leverage IoA expertise in delivery of high performance pipelines for Gaia and VISTA (operations) and PLATO and WEAVE/4MOST (under development). Training and user support for effective use of wider UK (e.g. UKT0) and EU (e.g. PRACE and EOSC) or LSST data analysis. Liaison with LSST Project, LSSTC, LSST:FR and partners for provision of documentation and training materials, implemented via training networks[6], workshops (e.g. the and conferences (e.g. extension of the LSST Data Fusion and LSST@Europe[7] series).

The development of the L3DAC:Cambridge will leverage existing close industrial links with Dell, Intel, Mellanox, NVIDIA, Xyratex, DDN, IBM, Microsoft, ARM, Samsung. Training will benefit from the Cambridge STFC Centre for Doctoral Training, which links with our industrial partners.

This Expression of Interest has understandably focussed on the Cambridge part of the distributed DAC model to underline our commitment to this distributed UK model that is aimed to maximise the science return to the LSST:UK community. If our proposal is supported we welcome the opportunity to develop this concept with other partners who may be interested in contributing to this distributed LSST:UK DAC model.

Version: 20171130

[6] A UK led H2020 ITN is under development linking key groups in UK, FR, DE, IT, ES etc to provide PhD training in key LSST science areas.

[7] LSST@Europe 3; Lyon 11-15 June 2018 - see http://europe2018.lsst.fr/