Expeditions in Science on Federation DCI

Submitted by: Shantenu Jha, Andre Merzky, Mark Santcroos and Matteo Turilli, RADICAL, Rutgers University

http://radical.rutgers.edu

1. Identify the collaborating teams in Europe and the US.

The RADICAL (Research in Advanced Distributed Cyberinfrastructure and Applications Laboratory) is involved in the theory, practice and deployment of distributed computing to advance scientific research. The RADICAL group has several partners in Europe – spanning the spectrum from domain science to fundamental computer science research, that are either already using both EGI and XSEDE resources or are planning to in the near future.

These partners include: (i) Amsterdam Medical Center (AMC) and their work in bioinformatics, (ii) ExTASY – a EPSRC/NSF Joint funded project, (iii) Standards-based CI for Hydrometeorolgical Modeling (SCIHM Project), which collaborates with the DRIHM/DRIHM2US project (involving LMZ, CIMA Foundation etc.) (iv) Climate Modeling (Amsterdam, National eScience Center), (v) Rutgers-UCL collaboration on biomedical computing (including but not limited to the VPH projects), and (vi) Extending ExTENCI (NSF Funded project to bring XSEDE and OSG "together") to EGI, possibly also involving EMI/MEDIA.

2. The scientific justification for the collaborative activity.

What is common to all of the above scientific problems (with the exception of Use Case vi, which is essentially a computer science driven project), is that each of the use cases require more computing than is available locally or accessible easily. The more computation resources that can be utilized (without any increase in the complexity), better the solution is likely to be.

3. The justification for access to resources in both the US and Europe, where applicable

The projects described above have the following common characteristics: (i) There exists naturally distributed teams; the teams are either cyberinfrastructure teams and science teams that collaborate to provide deep integration of RADICAL cybertools for end-to-end application solutions, or teams with shared expertise on science problems (e.g., in the context of project v), (ii) Also common to all projects is the need for more computational and data resources than are locally available and thus the need to use federated resources.

4. The resource levels required to support the collaborative work.

(i) Technical Support from both the XSEDE and EGI teams, to help integrate RADICAL solutions into the XSEDE and EGI environments. We would propose

something akin to an XSEDE ECSS, and/or the assignment of EGI Champions. (ii) We have XSEDE resources and we make them available to our European partners via XRAC allocation/proposals. We need reciprocal resources on EGI.

5. A description of the patterns of usage of EGI and XSEDE resources and services.

Here is a brief description of the intellectual drivers on a per project basis. We identify the European collaborating partner team.

Project I (Partner: Amsterdam Medical Center): Bioinformatics, data-intensive workflows.

Determine when the IO bottleneck on an HPC machine (e.g., XSEDE) machine becomes the barrier? Determine *when* to off-load tasks to a HTC resource such as OSG/EGI, i.e., when is the cost of distribution worth it? Having determined when to distribute, determine *how* best to utilize federated HTC and HPC resources for data-intensive workflows/simulations. In the current phase, before exposing the end-scientist, we are getting both the *reasoning* and the *plumbing* right.

Project II ExTASY (Partner Edinburgh, Nottingham): Biomoelcular simlations.

The ExTASY project is concerned about the efficient and flexible execution of a large ensembles of Molecular Dynamic simulations. There can be a range of variation in the coupling between the ensembles; sometimes there is no coupling, sometimes large exists "tighter" coupling. Furthermore, there is a variation in the duration of the simulation: at times multiple very short runs; at times long running simulation., as well as variation in the size of the systems being simulated (i.e., sometimes small systems, thus limited inherent fine-grained parallelism, but large levels of tasks-level parallelism).

This work is already funded by the UK EPSRC and is about to receive funding from the NSF under the SI2 project.

Existing collaboration between Rutgers-UCL (Coveney) has some of the same intellectual drivers as the ExTASY project, but with a focus on HIV-modeling (drug resistance) and Replica-Exchange simulations.

Project III Standards-based CI for Hydrometerogical Modeling (SCIHM), (Partners: LMZ/Munich, CIMA Foundation) and Climate Modeling with National e-Science Center, Amsterdam.

Some of the questions are similar to ExTASY (ie efficient execution of "many simulations" with varying degrees of coupling), but the underlying kernels are very different (mostly WRF and WRF-Hydro based) and with different characteristics. This work is jointly funded by NSF and the EC, mandates the use of standards -- both for data and for jobs/computes.

Project IV Extending ExTENCI to EGI (EGI, possibly MEDIA)

RADICAL's contribution to the ExTENCI project entails both the theory and practice of providing interoperable middleware, tools and services and using them to federate distributed infrastructure. We propose to extend these capabilities to include EGI, thus making it a XSEDE-OSG-EGI investigation which will also include Cloud Federation. Of relevance is the fact that the RADICAL group is responsible for the SAGA Project – an OGF standard, and has signed a MoU directly with EGI to support SAGA on EGI; SAGA along with a SAGA-based Pilot-Job is a common underlying fabric for many of the projects listed above.

This is not just a matter of "clubbing together" infrastructure; there are deeper principles waiting to be discovered and expostulated, for example: when to distribute as well as how to distribute, rather than just doing it in any way that works. Furthermore this should lead to an understanding of the role of Standards based approach to distributed infrastructure and Cloud Federation.

The plumbing is important and we have specific solutions. For example, SAGA can be used to submit jobs to a federated infrastructure: SAGA-OCCI to virtualized resource, and SAGA to submit jobs to non-virtualized endpoints. (This work can feed directly into extending the upcoming MEDIA project to the US)

6. Identification of the most significant limiting factors to using these resources and services as part of your collaborative work. These factors could include barriers that are currently hindering your work or issues that need to be addressed to make the collaborative use of resources easier

(i) Absence of a common runtime environment (one component of which is a common single API for job submission, file transfer, access to information service etc.), execution model and even a deployment model. Currently the two systems (XSEDE and EGI) look different, are used differently, have different assumptions and services; all this provides a very high barrier to learning new environment, (ii) Lack of clarity of available services on resources, (iii) Absence of a common resources allocation (For example the XSEDE-PRACE solicitation for joint computer time, has been in the pipelines for the past 6 months and there has not even been movement towards a formal call. Will this have the same fate?)