VIRT&L-COMM.8.2015.9

PRACE application form **12th PRACE Project Access Call**

Detailed Project Document

Project name	AMPEX-A (AB-INITIO MOLECULAR PROCESSES EFFICIENCY AS COLLABORATIVE SERVICE)- ASTROCHEMISTRY
Research field	Computational Chemistry HPC

Project leader

Title	Prof
Last name	Evangelisti
First name	Stefano
Organisation name*	University Paul Sabatier of Toulouse
Department*	Institut de Recherche sur les Systèmes Atomiques et Moléculaires Complexes
Group*	Laboratoire de Physique Quantique
Country	France

*Example
- *University of Amsterdam*
- *Department of Physics*
- *Group of Molecular Physics*

IMPORTANT NOTICE

Please upload a single document, using this template, in PDF format, which must not exceed 8 MB and addresses ALL of the points below. The minimum font size that you can use is 12 point and you must adhere to the page limits indicated below. The maximum number of pages you may submit, including references is 10 pages.

1. **Describe your research project. Include discussion of the scientific questions that you are planning to address and the overall scientific goals of the project. It is important that you describe the novelty, impact and timeliness of the proposal. (This section must be no longer than 2 pages).**

   The emphasis of PRACE on community proposals (for which in some recent calls a percentage of computing time was reserved) stimulated our Chemistry, Molecular and Material Sciences and Technologies (CMMST) Virtual Research Community (VRC) (see A. Laganà, G. Sipos, R. McLennan, "Towards a CMMST VRC, (VIRT&L-COMM.2.2012.2, http://www.hpc.unipg.it/ojs/index.php/virtlcomm/article/view/40/38)) to assemble a coordinated plan of submission under the name "AMPEX" (AB-INITIO MOLECULAR PROCESSES EFFICIENCY AS COLLABORATIVE SERVICE). The spirit of AMPEX leverages on a Virtual Community approach to complex computational simulations. In such approach highly demanding requests of both parallel computing runs and handling of large amounts of data are addressed to PRACE (from proper user interfaces built on interoperable and standardized components and services of distributed resources) for the part related to the use of highly parallelized packages. This enables effective collaboration between developers (f.e. application developers) and domain researchers based on the sharing of methodologies (models, algorithms and applications) and of data inside the same scientific application and across the different ones. In this way,

ISSN: 2279-8773

in fact, the most appropriate machines for different types of software solutions will be used when creating applications and transforming them into production tools for innovative technologies. This implies that huge amounts of data will be transferred and stored during initial, intermediate and final stages of the elaboration allowing the creation of large shared repositories of efficiency parameters for Molecular processes. Three community proposals for the following fields of applications are considered :

1. AMPEX-A: SMALL MOLECULES AB INITIO REACTIVE EFFICIENCY IN ASTROCHEMISTRY (Toulouse, Perugia, Chieti, Vitoria) Biennial. Packages: NWCHEM, MOLPRO, VENUS, DL_POLY, BERTHA, RWAVEPR (in-house quantum reactive TD scattering code for atom-diatom reactions), MIXQC (in-house mixed quantum-classical code for reactive and inelastic scattering in diatom-diatom collisions)
2. AMPEX-P: SMALL MOLECULES AB INITIO ENERGY TRANSFER IN PLASMAS (Bari, Milano) Biennial. Packages: QCT4M, QUANTUM ESPRESSO (see related proposal)
3. AMPEX-C: LARGE MOLECULES AB INITIO STRUCTURAL PROPERTIES IN CATALYST DESIGN (Aachen, Tubingen). Biennial. NWCHEM, GAUSSIAN (see related proposal).

Our community proposal is concerned with application N. 1 (AMPEX-A) that deals with the evaluation of *ab initio* efficiency parameters of chemical processes, in particular those occurring in interstellar space and producing bonded Carbon atoms. Astrochemistry represents, in fact, a multidisciplinary scientific field that involves large computations aimed at evaluating in an ab initio fashion the efficiency of formation and the rate of disappearance processes of several species.

In particular, our research project deals with the use of the distributed *ab initio* molecular simulator GEMS, the Grid Empowered Molecular Simulator developed for the purpose of supporting accurate studies of the reactive properties of chemical processes. GEMS flows first through INTERACTION (the module either performing ab initio calculations of the electronic structure of the molecules of interest for a massive number of geometries using both highly scalable parallel codes (like NWCHEM as specified later) and less scalable ones (like MOLPRO, BERTHA) or collecting related information from existing data) and FITTING (the module building analytical formulations of the molecular potential energy surface using inhouse non parallel codes by collecting related information from existing data). Next GEMS flows through DYNAMICS (the module utilizing stored potential energy surfaces to the end of performing classical dynamics calculations for billions of events using highly parallelized codes (like VENUS, DL_POLY as specified later) and the more difficult to parallelize quantum (like RWAVEPR) or quantum-classical (like MIXQC) ones) and OBSERVABLES (the module assembling out of stored computed data accurate evaluation of potentially measurable data).

The goal of AMPEX-A is to feed some popular Astrochemistry databases with accurate *ab initio* astrochemical data shareable among related research communities. For this purpose the work of the various partners of the proposal will be orchestrated in a way to maximise the productivity of the related heaviest computational steps thanks to the:

- use of PRACE machine based highly parallel procedures

ISSN: 2279-8773

- adoption of *de facto* format standards (like Q5COST and D5COST, see: J. Compu. Chem. 35, 611 (2014)) for the repositories of the intermediate results
- collaborative exploitation of auxiliary competences and platforms.

This will stimulate, for example, the coupled execution of electronic structure and dynamics codes (like NWCHEM and VENUS as suggested in Comp. Phys. Commun. 185(3), 1074–1080, 2014). Other examples will be the linking of BERTHA, with RWAVEPR and/or MIXQC as well as the porting of the RWAVEPR GPU enhanced scalability version to PRACE machines and the export of the scalability model of classical trajectories to time dependent quantum techniques.

The work that will be carried out for the proposed project will be of invaluable help to provide information on unfeasible cross sections and rate coefficients measurements as is often the case for interstellar medium processes and, more in general, of the planetary atmospheres ones.

1.1. The need for multi-year access

The need for a second year is motivated by the fact that the consolidated parallel procedures developed during the first year for classical mechanics based packages will be extended in the second year to quantum codes for a systematic use and accuracy checks.

2. Describe how you will manage the resources requested? Use a Gantt chart or equivalent to illustrate this. (1 page).

The design and implementation of procedures feeding open access repositories containing the intermediate data produced by the various modules of GEMS (in particular the electronic structure information generated by the INTERACTION module, the potential energy surface (PES) routines generated by the FITTING module, the detailed quantum, classical, semi-classical transition matrices generated by the DYNAMICS module, the measurable quantities generated from molecular dynamics and statistical treatments by the OBSERVABLES module in order to estimate cross sections, rate coefficients, spectra, etc. to be delivered as final data by taking as **test cases some astrochemistry processes forming C-C bond** relevant to the origin of life) is the goal of the present project.

The modules mostly involved in the parallel implementation of the test and execution, are INTERACTION and DYNAMICS. They rely, in fact on computationally intensive procedures at different levels of granularity in order to save time and memory. The use of Tier-0 system will be managed on the basis of the kind of calculation to be performed and, thus, by taking into account if the adopted code is more demanding in terms of time consumed or memory requested and if the implementation of parallel I/O is useful.

DETAILED WORK PLAN:

Months 1-4: implementation of INTERACTION and DYNAMICS modules including the link to externally generated input data and computational procedures of the (FITTING and OBSERVABLES modules) relevant to the test case chosen

Month 5-6: test of the overall workflow.

Months 7-16: production runs for chained NWCHEM, VENUS or DL_POLY calculations.

ISSN: 2279-8773

Months 17-20: calculation of OBSERVABLES from outcomes of the previous modules and export of the trajectory parallel model to RWAVEPR.

Month 21-24: parallel runs of RWAVEPR.

MONTH	2	4	6	8	10	12	14	16	18	20	22	24
INTERACTION												
FITTING												
DYNAMICS												
OBSERVABLES												

3. Describe the numerical methods and algorithms that you are planning to use, project, and how these will enable the research to be achieved. (1 page)

Algorithms and codes used in the four modules of GEMS are characterized by the adoption of both coarse and fine grained levels of parallelism, by the adoption of BLAS and LAPACK (as well ScaLAPACK) mathematical libraries, and some other toolkits such as Global Array or ARMCI.
The highly scalable suites of programs that will be used by us are:

NWCHEM (http://www.nwchem-sw.org/): an open source high-performance quantum chemistry package, implementing a huge variety of *ab initio* electronic structure calculation techniques, such as CASSCF, DFT, TDDFT, CCSD(T), post HF methods, QM/MM, etc. NWCHEM makes use of the BLAS library as well as of the ScaLAPACK one. Its parallel implementation is based upon the Global Arrays parallel environment and can benefit from the adoption of the ARMCI-MPI library, that provides also for MPI-2 and MPI-3 implementations.

DL_POLY (http://www.ccp5.ac.uk/DL_POLY_CLASSIC): a general purpose molecular dynamics simulation package with in-built parallel algorithms. It may be run on a wide selection of distributed memory parallel computers, from supercomputers with thousands of processors, to single processor workstations and can simulate systems whose size ranges from an order of 100 atoms to millions. In order to achieve high-level parallelism DL_POLY makes use the MPI based Domain Decomposition strategy which is based on a link-cells algorithm.

VENUS (http://hase-group.ttu.edu/venus.html): an open-source general chemical dynamics fortran computer program able to simulate collisions between two fragments (atoms, diatoms or polyatomic molecules) at fixed vibro-rotational states and translational energies. Its run needs a very small amount of memory. The in-house parallelization was performed using the MPI libraries with a master-slave paradigm where the master takes care of the input/output and the slaves carry out the integration of the classical equation of motions of a huge number of appropriately sampled orientations of fragments.

RWAVEPR: an in-house developed quantum reactive scattering time dependent Fortran code, integrating the time dependent Schroedinger equation starting from a given pair of v and j (vibrational and rotational) states of the reactant diatomic molecule (BC) colliding with

ISSN: 2279-8773

a generic atom (A). The code uses BLAS and LAPACK libraries. Both coarse grained (MPI) and fine grained (CUDA) implementations are available for this code.

The other GEMS programs implemented on other platforms are:

MOLPRO (https://www.molpro.net/): a suite of *ab initio* programs for molecular electronic structure calculations. It comprises all standard computational chemistry applications, such as DFT and coupled cluster methods. In order to calculate electronically excited states MOLPRO makes use of MCSCF/CASSCF, CASPT2, MRCI, or FCI methods. MOLPRO can run on multiprocessors computing platforms thanks to the use of the Global Arrays toolkit and or of the MPI-2 library.

MODULE FITTING: a set of small codes (like GFIT3C, see: Comput. Phys. Commun. 108, 259-266 (1998) and other in-house ones) carrying out the fitting of electronic structure energy values with the aim of assembling a PES to be used in the DYNAMICS module.

MIXQC: an in-house code computing both reactive and non reactive transition probabilities of four body processes by separating vibrations (treated quantally) from translation and rotations (treated classically).

4. **Explain why this project needs to run on a Tier-0 system, why the machine you have requested is suitable for the project and how the use of the system will enable the science proposed. You should describe the architecture, machine/system name and the problem sizes that have been used to test for scaling and provide supporting evidence. Provide both a table and scaling plot such as the ones shown below with example data to illustrate the information requested (2 pages).**

BLUE GENE/Q (Fermi hereinafter) of CINECA is a proper machine for carrying out AMPEX-A. In particular, for the purposes of our project, some GEMS applications will be run on Fermi (in particular, those exploiting the relevant feature of this the platform) some other ones will be run on other platforms, as is the case of the applications belonging to FITTING (GFIT3C) or to DYNAMICS (RWAVEPR). With respect to the applications to be run on Fermi, our key effort will spent for NWCHEM, DL_POLY and VENUS.

NWCHEM: scalability of the code is illustrated in the first Table + figure below by showing the results of the triples part of a completely renormalized equation-of-motion approach with singles, doubles and non iterative triples (CR-EOMCCSD(T)) calculation for the GFP (Green Fluorescent Protein) chromophore, made up by 238 aminoacids, represented by a cc-pVTZ basis set (for a total of 648 functions). The calculations refer to the Franklin Cray-XT4 computer at NERSC (see: J. Chem. Phys. 132, 154103 (2010)). The figure shows that the scalability of the triple part of the calculation is good up to 34008 cores. The values of the speedup in the table were calculated with respect to the time measured for 1024 cores.

VENUS: the scalability of the code was calculated for the H_2 + H_2 reaction on the MareNostrum supercomputer at BSC (Spain). Time measured for serial and parallel runs are shown in the second Table + Figure below showing that results obtained with 1024 cores are satisfactory. The decay of the performances after 512 cores is due to the fact that the work

ISSN: 2279-8773

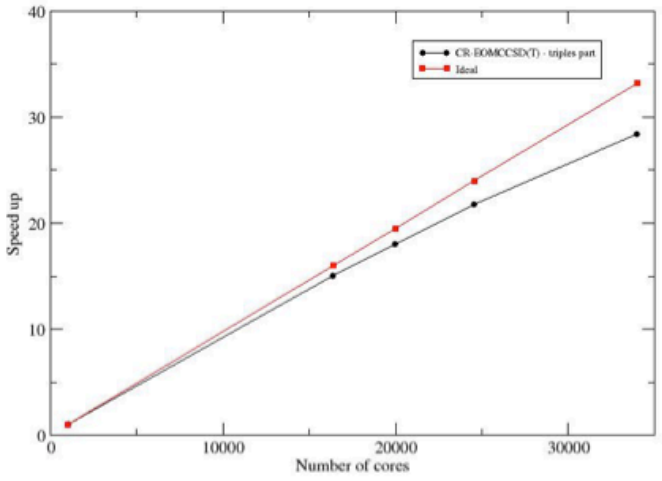
load distribution could not benefit from an input large enough. The use of a larger input is expected to give results close to the ideal speedup up to 2048 cores.

DL_POLY: scalability results were obtained for systems with 6000 and 10000 atoms and were excellent (they are not shown here for space limitation reasons).

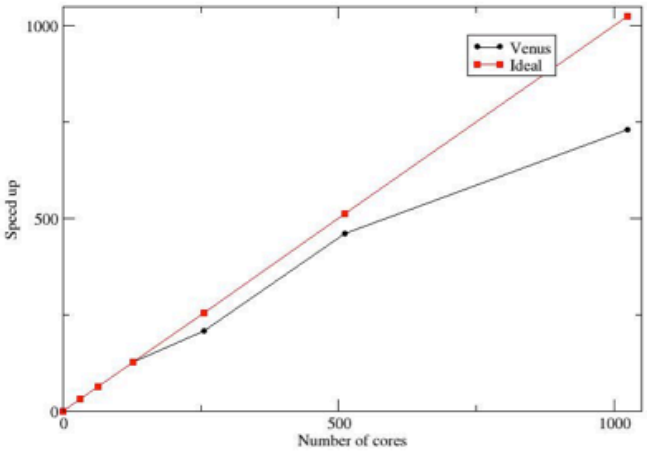
RWAVEPR: the scalability of the code was calculated on NVIDIA GPUs. In particular, performances were obtained on both GTX and (the more powerful) Fermi architectures and are an interesting study case. The third plot below shows the increment of performances with the number of adopted CUDA threads (solid line GTX results, dashed line Fermi results). Results are very satisfactory and our goal is to reproduce them with the use of MPI on a TIER-0 system.

For the cases considered by AMPEX-A no explicit use of parallel I/O is foreseen though some trials will be attempted, especially with DL_POLY.

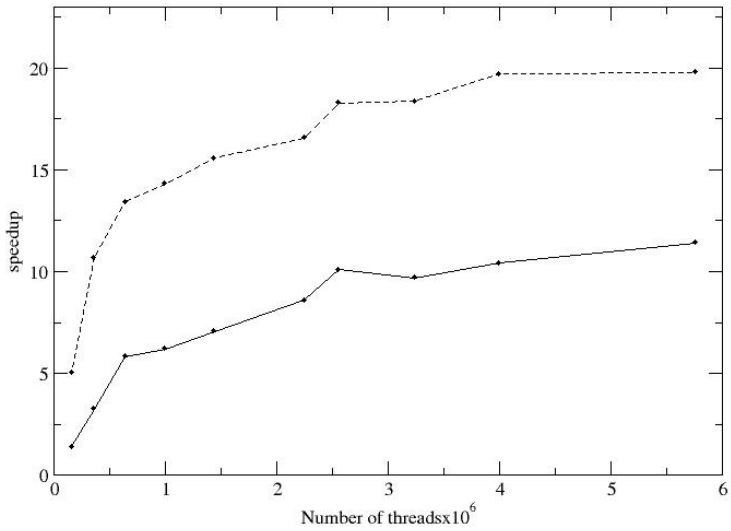
# cores	absolute time (s)	speedup
1024	43124	1.00
16384	2872	15.02
20000	2392	18.03
24572	1982	21.76
34008	1518	28.41



# cores	absolute time (s)	speedup
1	2267.2	1.00
32	70.85	32.0
64	35.08	64.6
128	17.67	128.31
256	10.09	208.0
512	4.91	461.8
1024	3.1	731.35



ISSN: 2279-8773



5. Describe your experience of using HPC resources in the past and how you will manage using a Tier-0 system. What other experience do you and your team bring to this project? (1 page).

The experience of the members of AMPEX-A with HPC resources moves from the development of efficient parallel computational codes (making use of both MPI and OpenMP libraries as well as of innovative technologies such as those needed for running on modern GPUs), till the massive use of HP facilities for solving large scale problems. For illustrative purposes we describe here some of them concerning the use of long range potentials and relevant to the cold processes considered for the present Astrochemistry proposal.

- **ICTS GESGA (2010) project** (Acceso a Infraestructuras Científicas y Tecnológicas Singulares). Title of the project: *$(N_2)_2$ ab initio calculations.* The project was developed under the activities pertaining to the Scientific Research Project ESTEC N. 21790/08/NL/HE called "Fundamental issues in the aerothermodynamics of planetary atmosphere re-entry (see: ESTEC 21790/08/NL/HE Technical Proposal, M. Capitelli, A. Laganà, A. Varandas, C. Petrongolo, P. Jensen, B. Sutcliffe, M. Linoda Silva, I. Schneider, J. Tennyson (2008, April)). In particular, the ICTS CESGA project focused on the *ab initio* calculation of the potential energy values of the $(N_2)_2$ dimer, with the aim of assembling an accurate 4-body global PES for the $N_2 + N_2$ reaction, based on a many expansion approach. Calculations were carried out in parallel using the GAMESS-US *ab initio* package (http://www.msg.ameslab.gov/gamess/).
- **CASPUR-IT (2010) project.** Title of the project: *The effect of long range tails on the quantum reactivity of heavy systems*. Using semiempirical formulations of the interaction (metti un paper con Pirani) the effect of adding long range tails to strong interaction was investigated. Calculation were performed on ……using ….

ISSN: 2279-8773

- **CINECA ISCRA B (2011) project**. Title of the project: *Coupling High Performance with High Throughput calculations to study the N_2 + N_2 reaction* (acronym: HPHTN4). The goal of the project was to assemble a PES for reactive scattering of the nitrogen molecules, by means of the Collins fitting method (see: J. Chem. Phys., 108, 564 (1998) more suited for massive sets of data points and for parallel computing) and the consequent gas-phase reactive scattering calculations in order to obtain dynamics information on the N_2 + N_2 system. Also in this case, the GAMESS-US suite of codes was adopted for carrying out in parallel electronic structure calculations.
- Other efforts were spent on recent computational projects targeted to different scientific objectives and run on CESCA, Oklahoma University Computing center and CINECA, related joint usage of the US XSEDE supercomputer systems and the EGI Grid (see the specific BOF session at the 2014 XSEDE conference held in Atlanta https://www.xsede.org/documents/527334/748982/EGI_introduction_XSEDE_2014-07-14-17.pdf)

6. Justify the number of core hours requested. This should include information such as: run type, wall clock time per step, number of jobs per run type, the number of CPU cores and the total core hours per run type. This information should take the form of a table like the one shown below with example data. Explain how the core hours requested will be used (1 page).

Run type	# Runs	# Steps/Run	Walltime/Step	# CPU cores	Total core hours/Type Run
Nwchem	parallel	s1	w1	p1	R1*s1*w1*p1
Dl_poly	parallel	s2	w2	p2	R2*s2*w2*p2
Venus	parallel				
					TOTAL

ISSN: 2279-8773