Proceedings of the UK e-Science All Hands Meeting 2005, ©EPSRC Sept 2005, ISBN 1-904425-53-4

*e*Science methods for the combinatorial chemistry problem of adsorption of pollutant organic molecules on mineral surfaces

Toby O White[1], Richard P Bruin[1], Jon Wakelin[1,2], Clovis Chapman[3], Daniel Osborn[4], Peter Murray-Rust[5], Emilio Artacho[1], Martin T Dove[1,6] and Mark Calleja[7]

[1] *Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ*
[2] *Present address: Centre for e-Research, University of Bristol, H.H. Wills Physics Laboratory, Royal Fort, Tyndall Avenue, Bristol BS8 1TL*
[3] *Department of Computer Science, University College London, Gower Street, London WC1E 6BT*
[4] *Centre for Ecology and Hydrology Monks Wood, Abbots Ripton, Huntingdon, Cambridgeshire PE28 2LS*
[5] *Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW*
[6] *National Institute for Environmental eScience, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA*
[7] *Cambridge eScience Centre, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA*

Abstract

We study a set of environmentally-relevant pollutant molecules adsorbing on a pyrophyllite surface. The set of calculations needed for this study exemplify a short and fairly simple workflow; we make use of the existing *e*Minerals minigrid, and produce our own very simple self-written workflow management tools to govern the process. We show that in the presence of a robust minigrid it is easy to set up and manipulate simple workflows on this scale, and produce scientifically interesting results; the interaction of the molecules with the surface is shown to be largely governed simply by the number of chlorines in the molecule.

SECTION I

Introduction

A significant environmental problem facing society is the presence of persistent organic pollutants and related molecules in the global ecosystem. These chemicals include halogenated substances used as pesticides or produced as by-products from industrial, commercial or social activities. Many are toxic, and there is concern because these substances enter human and wildlife food chains. They may also affect the quality of drinking water or the functionality of other environmental resources, such as soil or the atmosphere. There are no comprehensive remediation strategies. Risk management strategies are needed that enable stakeholders to make informed choices about use and disposal options. Our specific challenge is to learn how these molecules bind to mineral surfaces as this might affect environmental behaviour and influence exposure patterns.

The role of eScience methods can be seen by considering the example of various polychlorinated polycyclic aromatic series; as exemplified by the polychloro-biphenyls (PCBs), the polychloro-dibenzo-difurans (PCDFs), and the polychloro-dibenzo-*p*-dioxins (PCDDs). In each case, the series consists of an identical molecular backbone, with the members of the series (known as congeners) having different numbers of chlorines substituted for hydrogens at different positions. Combinatorics and symmetry result in there being 209 distinct congeners of the PCB family; 135 PCDFs, and 75 PCDDs. We need to gain information about each of these molecules, including obtaining binding energies for each molecule on different mineral surfaces. In addition to the mechanical problem of managing such a large number of calculations, we need also to face a "knowledge problem" of handling

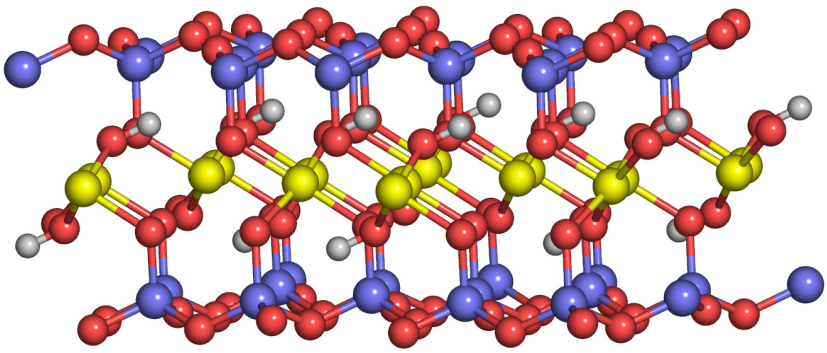

Figure 1: Representation of the surface of the layer silicate pyrophyllite. Colour scheme for atoms is red = oxygen, blue = silicon, yellow = aluminium, white = hydrogen.

data from such a large suite of calculations.

Science details

Although we intend to study all of these three families of molecules, here we show only results for the PCDDs, which are the smallest family, and for which symmetry reduces the problem space. However, the methods we have used to construct and manage the simulations and analyse the data will be easily extended to cover the other series.

Furthermore, we have restricted ourselves to one mineral surface, the (001) surface of pyrophyllite. Pyrophyllite is an aluminosilicate mineral prevalent in clay soil and representative of the class of clay and mica minerals; its (001) surface predominates in nature.

Computational details

The simulations required, which will be described below, form a fairly simple, but highly parallelizable and repeatable workflow when considered for the whole system of molecules. It is anticipated that this workflow represents a common work paradigm and so each step of the workflow has been written to be as generic as possible without compromising workflow performance. Due to the simplicity of the workflow the decision was made to not use popular workflow editors / generators such as Taverna, but rather to create our system from scratch. The compute resources used were a campus-wide Condor grid, and a back-end for data storage based on the Storage Resource Broker (SRB) [1].

SECTION II

Scientific overview of simulation programme

The focus of this study was to calculate binding energies for molecules on a mineral surface; in such cases the accuracy of the calculations requires quantum mechanical methods, and our tool of choice is Density Functional Theory (DFT). Due to the relatively large size of our system (182 atoms when studying PCDDs), and the surface nature of the calculation, we require a DFT method which scales linearly with the number of atoms, and is not hindered by the simulation cell containing empty space. For this reason, we used the SIESTA [2] code.

Our simulated systems consisted of a portion of pyrophyllite surface combined with one of the molecular congeners of interest. The simulated surface, of (001) pyrophyllite, is illustrated in figure 1. The exposed layer of atoms is primarily composed of silicon and oxygen atoms, and it is the interaction of these surface atoms with the molecular atoms that will largely determine the adsorption energy. The illustrated portion of the surface consists of four crystal unit cells, in a 2 × 2 configuration, with the surface being one unit cell deep. SIESTA imposes periodic boundary conditions in all directions, so our system is in fact an infinite plane, with a separation between surfaces that we may tune to remove surface–surface interaction effects. For the accuracy of calculations

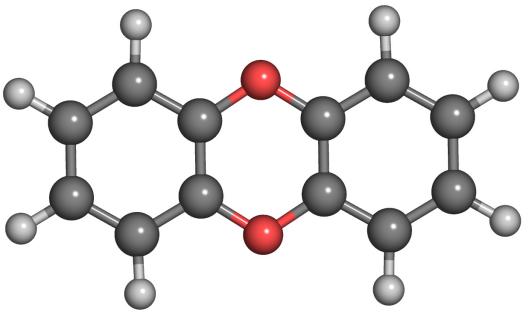

*Figure 2: Representation of the non-chlorinated dibenzo-*p*-dioxin. Colour key is gray = carbon, red = oxygen, white = hydrogen.*

performed in this study, a surface depth of 1 unit cell was found to reasonably represent the bulk mineral. When a molecule is placed in the system, it will be periodically imaged in the same way, and we found a 2 × 2 cell to be sufficient for removing intermolecular effects.

At this stage of the work, we focused only on the PCDD family of molecules. The non-chlorinated dibenzo-p-dioxin is illustrated in figure 2; as can be seen, it has eight substitution sites. The molecule in its ground state is planar, and its symmetry in this configuration is such that repeated substitutions of chlorine atoms onto these eight sites gives rise to the total of 75 chlorinated congeners. In addition to the unchlorinated base molecule, we therefore required 76 sets of calculations in this study.

The first task was to construct and parameterize a simulation cell containing the surface. This was achieved by simulating a cell of bulk pyrophyllite, and cleaving the surface in the appropriate plane. We tested our system with respect to various input parameters for energy convergence consistent with the level of accuracy required. For the interest of DFT affectionadoes, we found sufficient the use of the GGA functional, operating on an autogenerated DZP basis set with energy shift of 50 meV, combined with non-relativistic PBE pseudopotentials, using the default cutoffs recommended by Troullier and Martins [3], and a grid cutoff of 100 Rydbergs. Although representing a high level of approximation, the salient features of the system emerged. This gave us the energy and structure of the bare surface. Although this stage of the work did not require some of the eScience tools which became invaluable in later stages, nevertheless it was facilitated by the existence of the *e*Minerals minigrid [4].

The next step was the calculation of the structures & energies of all of the molecules. The crystal structures for a small number of them were available from the Cambridge Structural Database (CSD), and approximate initial structures for the rest could be estimated by appropriate substitution of Cl atoms for H, and accompanying changes in bond-length. A small program was written to generate all congeners, discard symmetry-related duplicates, and calculate appropriate initial structures. The structures for the molecules in the gas phase were then optimized. The simulations for the molecules were performed under the same approximations as those for the surface.

In order to gain assurance of the reasonableness of these structures, we compared our calculated geometry with that obtained by experiment, for one particular congener for which crystallographic data is abundant: the 2,3,6,7-tetrachloride [5]. It should be noted that exact reproduction of the geometry is not to be expected since the molecular structure will not be unchanged between gas-phase and crystalline structure; nevertheless, all structural properties calculated differed by less than 2.9% from experiment.

The final stage in the calculation was to calculate the energy of the surface with adsorbed molecule. An important question to ask concerned the position and orientation of the adsorbed molecule relative to the structure. In each case, we placed the relevant molecule parallel to, and a short distance from, the surface; in every case, we placed it above the same point. We then allowed the system to relax while holding constant the positions of the surface atoms, the relative positions of the

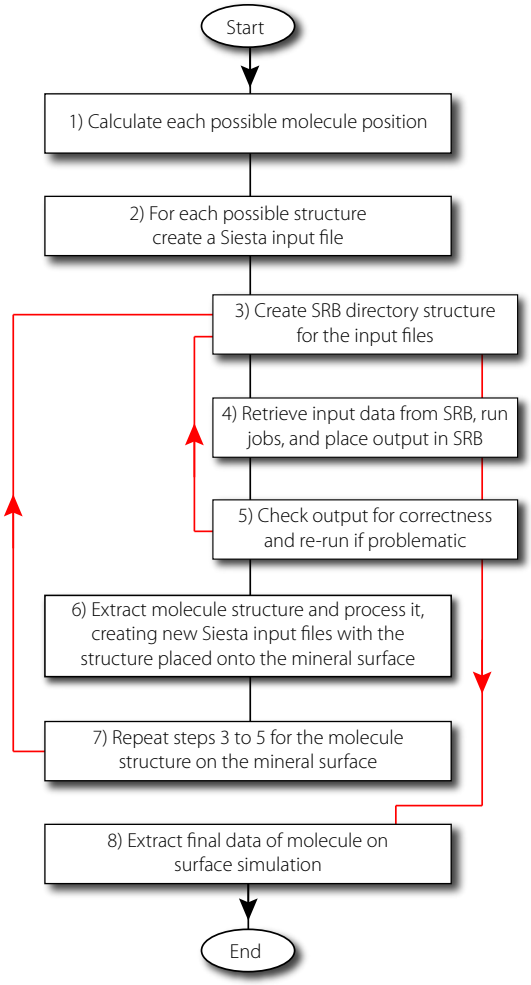

Figure 3. Representation of the workflow for the simulations described in the text. This is a two-pass process, with the arrows showing the directions of the workflow, and the red lines showing how calculations are repeated either as the second pass or to correct failures.

molecular atoms, and the relative orientation of the surface and molecule. Thus the only degrees of freedom allowed to vary were the distance from the surface to the molecule, and the position of the molecule in the plane of the surface. Given the results below, we feel this restriction on the geometry of the adsorption was justified. Therefore, under these conditions, and using the same level of approximation as earlier, the 76 relaxations were then performed.

Computational overview of simulation programme

The created system workflow is illustrated in figure 3. In brief, the workflow consists of two major parts; the simulations of each of the molecules alone, and the simulation of the molecules on the surfaces. The challenge for eScience was to facilitate the management of both large sets of simulations, and to provide ways to use the output of the first set as input to the second.

All of the individual job submissions were performed using a locally-developed tool called my_condor_submit [4]. This code wraps Condor-G job submission, adding support for automatic retrieval of data from and storage back to the SRB. This allows for automated use of the SRB and Condor/Globus technologies and forms the basis for the *e*Minerals integrated compute and data grids, removing the burden of data transfer from the user, and ensuring that a complete electronic log-book can be kept of the user's work.

Furthermore, we have (for this and other projects) added to SIESTA an XML-compliant output, in the form of CML [6]. This facilitated much of the data manipulation that was required. We did not, though it might seem equally logical for workflow architecture, add CML input to SIESTA. Though it was considered, it was felt to be a much more challenging task, and one which offered fewer rewards. The ease of manipulation of XML data means that it is relatively trivial for transformations of the XML to produce a non-XML format.

We now give a brief overview of the problems to be solved at each step, and describe our solutions. We also comment on the more general usefulness of our solutions.

1. Generation of all molecular structures. This was described in the previous section. The tools used for this were written locally, using the FROWNS library [7] and may be considered generally applicable. They have been used separately for other series of chlorinated molecules.
2. SIESTA uses its own input file format, which is not generic, but is well-documented. It is also modular, which

enabled us to separate out parameters relating to the generic features of the simulation, and create individual input files containing only those parameters relevant to each simulation; ie atomic positions, atomic identities, and a system identifier. Clearly our solution here was specific to SIESTA, due to the nature of the file format; however it was not a difficult step, and the same effect may be achieved in a few lines of any scripting language, given sufficient knowledge of the input format for a particular program.

3. A directory structure was then created on the SRB, with a separate directory for each simulation. The structure was chosen to allow easy traversal by both scripts and humans; this was important for ease of use when writing the tools to submit & monitor the jobs, while allowing for some manual overview of the process. Into each simulation directory was placed the input file relating directly to that simulation. In addition, however, each simulation needed an input file for generic parameters, pseudopotential files for each element, and of course, access to a SIESTA executable. These last are very large when compared to the input files; and it is also important to ensure that every simulation is using precisely the same versions of the generic input files, pseudopotentials, and executable. Furthermore, the pseudopotentials and executables will be of use across many sets of simulations, and it makes little sense to replicate these for each set of simulations, let alone for each individual job. They were therefore kept in separate directories, to be retrieved when necessary. This enabled both problems to be circumvented - space was saved by not needlessly duplicating pseudopotentials and executables; and consistency in their use was ensured.

   Similar issues will arise in many sets of simulations, where the individual simulations require consistent, shared data, as well as individual parameters which vary across the individual jobs. The particular choice of directory structure may well vary according to the system under consideration, but such structures are easily created with appropriate scripting, and as long as they are kept consistent, tools may be reused with little effort between studies.

   Each simulation also required a job submission file. The job submission tool in use, as mentioned above, is my_condor_submit, which abstracts both data storage and compute resource allocation behind the SRB and a condor/globus grid. This requires a submission script in the same format as those of Condor (with additional keywords to govern interactions with the SRB), and was easily generated. By dint of careful choice of filenames, we were able to ensure that the only difference in each script was in which SRB directory it should look for its input files. Again, 76 such input files may be easily generated, given a sensible choice in directory structure.
4. Given the existence of the necessary tools, job submission was now a simple matter of "my_condor_submit"ting the 76 submission scripts, which may be trivially automated. Naturally, this hides the complexity within my_condor_submit, which, for each job, generates a) a script for SRB interaction, to retrieve the necessary files from multiple directories in the SRB, b) a normal condor submission script for the simulation, and c) a second SRB interaction script to upload the simulation results into (potentially multiple directories in) the SRB; and finally ties all of these up into a small DAGman workflow submission script, which is then submitted to condor. The feature of retrieval from multiple directories was necessary to ensure that every job was using the same pseudopotentials and executable.

   The simulations were performed on a variety of resources; some on Cambridge

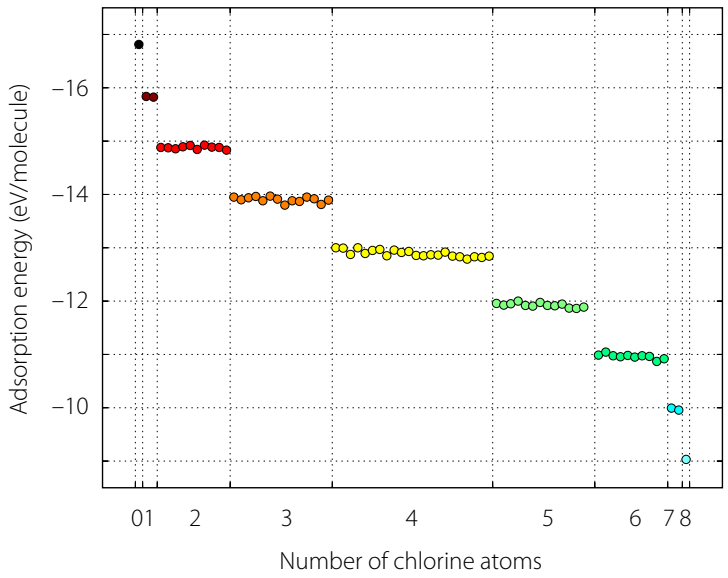

Figure 4. Adsorption energy of PCDD molecules on the (001) surface of pyrophyllite, plotted for each cogener and indexed by the number of chlorine atoms.

University's campus grid, CamGrid [8], which is a flocking-together of several regular condor pools across the university; and some on *e*Minerals clusters both in Cambridge and Bath, which are globus-enabled resources. Once submitted, the jobs could be monitored using standard condor and globus tools. In addition, those jobs running on CamGrid benefited from a set of bespoke tools which enabled output to be monitored during execution, which enabled the authors to check that convergence was being correctly achieved.

The my_condor_submit tool is clearly of potentially wide utility, and has been well-tested within the *e*Minerals project. The CamGrid monitoring tools may also be of interest, since the use of condor generally precludes monitoring of output during execution. Interested parties are encouraged to approach the authors for further information.

5. It is necessary to confirm that each of the jobs completed successfully. This may be done partly mechanically by testing that the SIESTA CML output is fully XML compliant, but it is also necessary to judge some of the output for reasonableness with a scientific eye. This perhaps may be seen as impeding the natural workflow, but was unavoidable at this stage. It was, though, in a relatively small workflow such as this, no great hindrance.
6. The files were then transformed by XSLT in order to extract the final, optimized, position of the molecule, which position could then be manipulated to be placed within the simulation cell of the (previously calculated) surface, and rotated and translated such that every molecule was placed in a comparable position above the surface. A new SIESTA input file was then generated for each such structure, and uploaded to the SRB within the appropriate directory structure. The tools used to perform this processing relate fairly closely to the nature of the system under investigation, but the principles involved are generic. The fact that the task was made much easier by the output of well-formed XML documents – and the fact that the tools now written are independent of the precise syntactical form of SIESTA's normal output – provides anecdotal support for the wider use of CML output in simulations. It provides an easily manipulable data format which is suitable for scientific simulations and is resilient to small changes being made to data output without having to completely recode any post-processing parsing applications.
7. All of the combined surface/molecule

input files having been created and uploaded to the SRB, the submission and data-collection for this, second, set of simulations proceeds again according to steps 3–5. Again, the successful completion of the jobs may be checked partly mechanically, but a practised eye must run over some of the output.

8. Finally, a second set of XSLT-based post-processing tools may be brought into play to extract and manipulate the data from the 76 simulations, and explore it to find which, if any, results are of interest. These tools are, by and large, particular to the project at hand, but again SIESTA's XML output massively facilitates this task, avoiding the regex hell so often associated with postprocessing. The results of this stage are explored in the next section.

SECTION III

Scientific results

The workflow above having been completed, the results collated and manipulated as described in the previous section, the following observations were found to be of interest.

The adsorption energy for a molecule on a surface may be calculated as the difference in total energies;

$$E_{\text{ads}} = E_{\text{sys}} - (E_{\text{surf}} + E_{\text{mol}})$$

where E_{surf} is the total energy of the bare surface, calculated in generating the structure of its structure; E_{mol} is the energy of the molecule alone, calculated in the first half of the workflow; and E_{sys} is the energy of the combined molecule/surface structure, calculated in the second half of the workflow. E_{ads}, therefore, was calculated for each molecule, and the results are shown in figure 4. This clearly shows that the largest factor in the variation of adsorption energy across the series of molecules is the number of chlorine atoms involved; the differences in E_{ads} between isomers of the same chlorination level is dwarfed by the difference in adsorption energy between different chlorination levels.

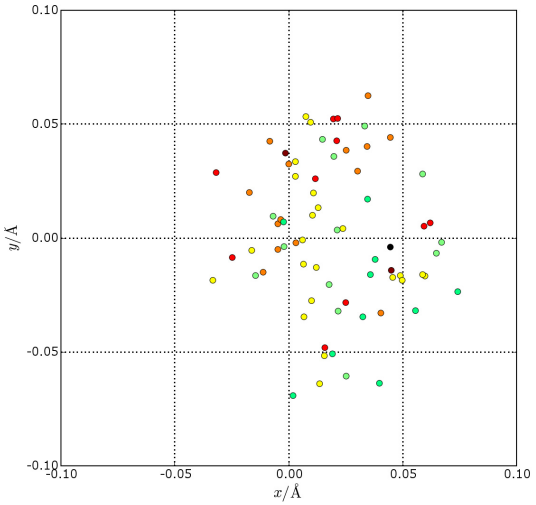

Figure 5. Minimum energy position of each PCDD congener in the plane of the (001) surface of pyrophyllite.

Indeed, it is possible to quantify this energy; each time we replace a hydrogen atom by a Chlorine, we lose the energy of a surface-H interaction, and gain the energy of a surface-Cl interaction; and this difference in interaction energies is approximately 1 eV, regardless of the nature of the surface directly beneath the atom in question.

This indifference to the nature of the surface is confirmed by figure 5, which shows the total molecular motion in the plane of the surface between its starting position and its fully relaxed position; as can be seen, there is no pattern, and the magnitudes for each simulation are tiny on the scale of the simulation, which is entirely consistent with the surface-molecule interaction being relatively unaffected by the nature of the surface.

The surface–molecule distance shows a similar result; the average is approximately 3.65 Å, which is sufficiently far that electrostatically, the charge variation of the surface will be largely "smeared out", rendering the surface more or less uniform from the molecule's point of view. The energy of the relaxed adsorbed system also seems to be relatively insensitive to this surface-molecule distance, within a range of about 0.5 Å around the average.

Both of these observations of the relaxed

position are consonant with the observed variation in E_{ads} – since the adsorption energy is more or less unrelated to position on the surface, any energy minima related to the *x*–*y* position are very shallow, which is reflected in their being little or no *x*–*y* movement in the relaxations. Similarly, since E_{ads} is insensitive to the nature of the surface beneath it, only to the number of interactions, we saw that the relaxed surface-molecule distance was more or less constant, and there was very little motion parallel to the surface.

eScience results

The workflow used for this project was explored and explained in the previous section. Of course, the idealized view described therein exemplifies, but does not define, the processes involved in this study. In fact, we investigated various levels of approximation initially, and the first half of the workflow was performed several times to generate several collections of molecules with structures optimized for these various approximation levels. Similarly, the precise spatial relationship of the molecule and surface was varied, so steps 6 and onwards were also repeated using different algorithms for generating the joint surface/molecule input positions. The fact that it was possible for these variations to be easily performed illustrates the versatility of the relatively-low technology tools in use.

We found ourselves in the position where robust computational (Condor) and data-handling (SRB) abstractions were in place, and were well-used and tested. This allowed us to, with remarkably little additional effort, write a family of small helper tools for manipulation of XML output, and of small shell scripts for directory creation and traversal. From the solid base of the *e*Minerals minigrid, and from the previous work that had gone in to ensuring SIESTA's CML output was well-defined, and well-formed, a small workflow was easily created and its processes exploited to perform sets of simulations which highlighted an interesting scientific result. Without these underlying infrastructures, it is very unlikely that anything of the sort could have been successfully completed.

Acknowledgements

We are grateful to NERC for funding the *e*Minerals project, to the Centre for Ecology and Hydrology for CASE student support for RPB, and to Unilever for support for the work of PMR.

References

1. Soler JM et al, *J Phys Cond Matter* **14**, 2745, 2002 (www.uam.es/siesta)
2. Rajasker A et al, in *11th IEEE International Symposium on High Performance Distributed Computing*, p 301, 2002
3. Troullier N and Martins JL, *Phys Rev B* **43**, 1993, 1991
4. Calleja M et al, *Mol Sim* **31**, 303, 2005
5. Boer FP et al, *Adv Chem Series*, **120**, 14, 1972
6. Wakelin J et al, *Mol Sim* **31**, 315, 2005
7. http://frowns.sourceforge.net/
8. Calleja M et al, *All Hands 2004*, p 173