Data analysis progress report
For the STAC and EGO Council

VIR-039A-08
May 29th, 2008

Summary

This report describes the Virgo data analysis activities and progress for the November 2008 to May 2008 period.
It reports in separate sections the activity of each search group, then comments about the developments of services and tools for data cataloging and data transfer, then provides up to date estimates about the computing needs for the present year, and preliminary estimates for the following year.
The last section is dedicated to answer to the Nov. '07 STAC recommendations.

Table of contents

1.1 The analysis activity of physics groups

We recall that Virgo runs four physics groups, closely interfaced and actually joint with the homologous LSC groups, with different scientific targets:

Burst signal search	*(Burst group)*
Coalescent Binary signal search	*(CBC group)*
Continuous Waves signal search	*(CW group)*
Stochastic Background signal search	*(SBGW group)*

The activity of the joint LSC-Virgo search groups is reviewed by joint Review Committees.

To these groups we need to add the *h-Reconstruction* which takes care of calibrating the data and removing some of the known disturbances, the *Detector Noise* study group, and a *Data Quality* study group which presently gathers members of Burst and CBC.

1.2 Calibration and h-Reconstruction

The Calibration group over the last months has analyzed data taken during calibration sessions held during and after VSR1, with the goal to improve the model of the interferometer actuation and of the detector response to mirror motions, and to provide a correct model for h-Reconstruction.

Essentially, the calibration requires to know accurately how the dark fringe signal can be converted into a differential arm signal, and how the measured control signals affect the same differential mirror motion. To this end, a number of mechanical and optical transfer functions have to be characterized accurately in modulus, phase and delay.

A major effort has been made to improve the accuracy at frequencies in the range 10 – 50 Hz; this requires a good knowledge of the mechanical response of the payload at those frequencies. Since in that frequency range the payload drive is split between different suspension stages, the mirror and the marionette, the response of the latter had to be included in the model.

An alternative calibration method, based on the use of an auxiliary laser pushing a mirror by means of radiation pressure modulation, has been compared with the standard drive and found consistent at low frequencies. At higher frequencies instead differences are present, that are consistent with what has been observed in GEO, namely that the elasticity of the mirror leads to different responses depending on where and how forces are applied; since the photon calibrator acts on a small spot close to the mirror center, while the magnetic drive acts on magnets at the edges of the coated surface, the difference is expected and roughly consistent with a finite element modeling.
The important thing though is that at low frequencies the measured gains are consistent to within 5%, thus giving confidence in the overall model.

Other major efforts have been the measurement of the time delays accumulated both in sensing and in actuation, and the measurement of the absolute timing, which is also crucial for the LSC-Virgo joint analysis.

The h-Reconstruction in turn has the purpose to translate the dark fringe signals and the correction signals into an equivalent h(t) signal. The first step is to subtract from the dark fringe the effect of the control signal, to reproduce (under a linearity assumption) the behavior of "free" mirrors. Then the optical transfer function is used to translate the resulting signal into an equivalent h(t).

A new (V2) preliminary reconstruction has been released to the search groups for checks, and is currently being reviewed. Its major advancement, in parallel with the calibration results, is the inclusion of the marionette and beam-splitter control signals, with the aim to be accurate within a few % down to 10 Hz.

The activity of the calibration group is being reviewed by a Virgo review committee, to which a LIGO member has also been recently added. The methods and results are also regularly discussed with the LIGO calibration group.

The h(t) reconstruction currently takes place in Lyon and contributes for small amounts, of the order of 100 kSI2k.day, to the computing energy spent there.

More information about the group activity can be found at:
https://workarea.ego-gw.it/ego2/virgo/data-analysis/calibration-reconstruction

1.3 Burst Group

The Burst Group activities are as usual manifold : analysis of the C7 and S5/VSR1run data, definition of vetoes for VSR1, network analysis with LSC/LIGO for the common part of S5 and VSR1 …

1.3.1 Analysis of the C7 data : status of the papers.

The analysis of the C7 data has been completed beginning of 2008. A first paper concerning the joint analysis of one day of C7 data in coincidence with resonant bar detectors, mainly with a methodological purpose, was already submitted in 2007, has undergone extensive modifications to answer referees concerns and was recently resubmitted.
A second paper concerning a search in coincidence with the gamma ray burst GRB050915a has been submitted after completion of the review process at the end of 2007.
A third one is an "all-sky" search in the frequency band [150Hz, 2kHz], with astrophysical interpretation of the search results (upper limits) in particular in the context of recently predicted supernova signals and recent black hole –black hole merger waveforms as well. It is still under review but we expect a submission this summer.

1.3.2 VSR1 data quality and vetoes studies

Many members of the Virgo burst group are deeply involved in a Virgo burst-CBC joint effort on VSR1 data characterization. The main output of this activity is to define Data Quality (DQ) flags, to study and propose interpretation of the loudest events, and to set up the event by event vetoes for subsequent burst analyses. Weekly teleconferences are the place where activity is reported and discussed. A first list of several dozen DQ flags is already available and stored in the Virgo Data Base (VDB). They are available for our LSC colleagues too thanks to VDB. It allows suppressing the main source of high SNR glitch in the VSR1 data. Nevertheless, some outlier events still remain and we need to identify their origin.
The novelty is that we have now a new glitch finder tool (developed in the LSC) at our disposal, namely the Kleine-Welle pipeline which has been run (and is still running) on all the sensitive channels (such as the ones concerning acoustic or seismic probes). The main

advantage of this tool is that it runs rather quickly. It is too early to write down definitive conclusions but we hope much from this tool for helping identifying critical auxiliary channels and understanding the origins of the VSR1 outlier events distribution.

1.3.3 Gravitational burst searches in the VSR1 / S5 data

After some effort on a test week selected for prototyping the joint searches (JW1), with studies of triggers from different pipelines (Peak Correlator or coherent WaveBurst from instance), we have turned our attention to the analysis of the full VSR1 data set. We can note two all-sky searches mainly carried on by Virgo people.

One is using the EGC pipeline, focusing on the [300 - 5000] Hz bandwidth where Virgo and LIGO detectors had similar sensitivity during VSR1 and S5. The unique characteristic of the pipeline is the coincidence implementation where instead of making 3 fold or 4 fold coincidence between the two Hanford, Livingston and Virgo detectors, we consider all coincidence between two interferometers (OR combination of 2 fold coincidence). It has been shown using simulated data that this the best way to add Virgo to the LIGO network coping with the fact that Virgo and LIGO detectors are not aligned.

The other is based on coherent WaveBurst focusing on the high frequency [2 -- 6] kHz. This is a pipeline developed in the Florida group of the LSC, whose functionality has been extended by Virgo in order to be used efficiently in the higher frequency range.

It is believed that the extensions of searches in the higher frequency range will have some impact also on our DQ studies. Indeed, despite we expect the instruments noise to be more Gaussian and more stationary at high frequency, all the DQ studies made up to now concern rather low frequency regions (up to 2 kHz). We expect that new categories of artifacts will be revealed, that we will need to be studied, flagged and vetoed.

Let us note that these two searches are not isolated and are in competition with other LSC searches.

Concerning the searches triggered by external astrophysical signals, we are participating to the combined study of coincidences with gamma ray bursts (GRB). Two GRBs, namely GRB070520b and GRB070729, have been identified for a first joint study (all the interferometers were in science mode around these GRB events). The main effort is currently put on GRB070729.

1.3.4 Computing resource utilization plan in 2008

The burst group uses Virgo computing resources essentially at the Lyon computing center, where until May 18th, the consumption amounted to about 17'000 kSI2k.day[1], out of a total of 60'000 kSI2k.day units requested for bursts studies in the 2008 computing plans.

The most part of these resources are used for the EGC all-sky search over VSR1-S5 coincident data[2], which is now ramping up its CPU utilization very rapidly. A recent and precise analysis of the needs for the rest of the year 2008 shows that in order to process the VSR1 and S5 data, including the estimate of the detection efficiency by means of software injections at different scales, about 400'000 kSI2k.day are likely to be spent. This computing energy will be needed to carry out the event search proper, plus 4 rounds of simulated injections in real data for efficiency estimation.

This figure alone exceeds the total 2008 Virgo request for Lyon by a factor > 3, and to put it in perspective, it corresponds to about 1/8th of the total computing energy requested at Lyon by all experiments.

---

1 We recall that the kSI2k is a unit of computing power corresponding to 1 Pentium 3 class CPU running at 1 GHz. The kSI2k.day is the computing energy provided by one such CPU over 1 day of running. Typical dual-core CPUs installed in computing centers, like an AMD Opteron, provide about 1.2 kSI2k / core, corresponding over one year approximately to 1'000 kSI2k.day for two cores.

2 cWB analysis is performed using the Caltech cluster in collaboration with the Florida LSC group.

Other searches for the burst group sum up to about 12'000 kSI2k.day estimated as needed in 2008, and are therefore substantial but of lesser impact.
The big discrepancy between the needs initially foreseen and those resulting from this more recent assessment has been a major source of concern for the group. In consideration of the budget limitations, the group has agreed to aim this year at running only 2 rounds of simulated event injections, corresponding to an estimated total energy needed for burst searches of 250'000 kSI2k.day, and postponing the rest of the analysis to 2009. Such a plan is still compatible with the goal to produce science results in time with the other competing LSC analyses, and before VSR2/S6 starts in mid 2009.
The burst group is currently tuning their pipeline parameters in order to reduce the computing costs while keeping the same efficiency.
Anyway, we will see later, when reviewing the overall needs and comparing with LSC, that these requests are realistic in the LSC-Virgo joint analysis context.

1.3.5 Preliminary needs for 2009

Depending on how much of the VSR1 – S5 analysis is carried out in 2008, the overall CPU request for 2009 for the burst group are in the ballpark of 500'000 kSI2k.day. This figure still needs to be defined within factors of 2, but is anyway pretty substantial and expected to follow the trend of growth registered since the beginning of the LSC-Virgo collaboration, including that other burst searches will be conducted providing substantial scientific results .

1.3.6 Comparison with LSC computing costs

The LSC quotes that the analysis of 1 year of S5 data (three detectors), restricted to the [0 – 2048] Hz frequency band, had a cost of 8 million hours CPU, spent on Caltech and UWM clusters, paying for three independent searches: using the Q-pipeline, the coherent Wave Burst (cWB), and AstroBurst.
Considering just one search, the cWB analysis costed 3.9 million hours: to translate into kSI2k.day, we compared jobs running on LSC clusters and in Lyon, measuring a conversion factor 1 LSC CPU hour = 33/480 kSI2k.day. Therefore a cWB analysis for 1 year of S5, restricted to [0 – 2kHz] has an estimated cost of 270'000 kSI2k.day.
The coincident VSR1 – S5 analysis will be extended to a wider bandwidth, say by a factor 4; will run on 4.5 months instead of 12, hence a factor 0.375; will use 4 instead of 3 detectors, hence a 4/3 factor.
Overall, this will lead the analysis to cost at least a factor 2 more; a cWB analysis would cost about 540'000 kSI2k.day.
This figure can be directly compared with the Virgo request for an EGC-based all-sky analysis; it is fair to state that it is pretty consistent with Virgo cost estimates.

1.3.7 Other concerns

The data transfer of the S5 data from Caltech to Europe has been a major source of frustration. The full transfer of the LIGO instruments strain data from Caltech to the Computing Centers (CCs) has been completed only in spring. The situation must be improved for S6/VSR2: we cannot be penalized by so long delays (months) any longer.

1.3.8 Links

All the burst activities are detailed in the virgo working area pages :
https://workarea.ego-gw.it/ego2/virgo/data-analysis/burst/burst-working-area/

Vigo data quality web page:
http://wwwcascina.virgo.infn.it/DataAnalysis/VDBdoc/index.html

1.4 CBC group

The coalescing binaries group is concerned with the search of events from Binary Neutron Stars (BNS) and Black Holes (BH) collisions, with or without the simultaneous presence of other messengers like GRBs.

1.4.1 LIGO-Virgo analysis

The joint LSC-Virgo CBC group has made the decision to include Virgo data in the low mass binary search, retaining the Virgo data for a focused search for binary neutron stars. The analysis of the joint VSR1/S5 data has been split in monthly runs. So far the efforts have concentrated on the second month of VSR1, used to develop the joint analysis framework. Some work has been done (primarily at UWM) to include Virgo data in the LSC *ihope* pipeline. The initial frequency for the analysis of the Virgo data has been tuned. The loudest triggers produced in the first stage of the playground data analysis have been followed-up. The pipeline is used in its latest version, which includes many new features but still needs some debugging, which has taken some effort in the past months. The next step will be the tuning of the analysis parameters, before running the analysis on the second month, and then proceed with the other monthly runs.
The CBC group also considers including Virgo data in the triggered search around some of the GRBs that occurred during VSR1/S5. This work has not started yet, but is expected to gear up from July 2008, with the start of a post-doc position in the Annecy group. The other DA post-doc position assigned to the CB group – the one in Urbino – is expected to be filled in summer and be operative next fall.

1.4.2 Data quality and vetoes

A lot of effort in the past months has been devoted to characterizing the VSR1 data in order to define data quality flags and build up vetoes for the CB analysis (and other analyses). This work takes place within a transversal group where members from the CB group play an active role. We are close to reaching a first milestone regarding this task, with the expected release in a few weeks of a first list of data quality flags and veto prescriptions.

1.4.3 Follow-ups

Some effort has been going on to make available the tools necessary to follow-up the Virgo triggers which will come out of the joint LIGO-Virgo analysis. Further work is still needed, especially for automation.

1.4.4 Tests for h(t) validation

Some work is also going on in the CB group to validate the new (v2) h(t) for VSR1, which is in the process of being produced. Some checks on CB triggers and CB hardware injections are being done to assess the quality of the new h(t).

1.4.5 Computing plans

For the longer term, the group is starting to explore the possibility of using the Lyon and Bologna computing centers to run the ihope pipeline for VSR2/S6 analysis. The perspective here is to use grid facilities to run the pipeline in a transparent way on multiple centers. The work needed to make this a reality needs to be assessed.

1.4.6 Computing resource utilization in 2008

The CBC group expects to use 1000 kSI2k.day for the remaining of 2008, which are expected to be devoted to run Virgo pipelines on the VSR1 (version 2) reconstructed data (a.k.a. h(t)), but not to contribute yet to the joint S5/VSR1 LIGO-Virgo analyses, which uses the LSC « ihope » pipeline and runs at LSC dedicated computing centers. Some little more is also expected to be spent for DQ studies.

1.4.7 Preliminary computing needs for 2009

For 2009, with VSR2/S6 bringing new data, the Virgo CBC component would want to contribute to the computing effort for the joint CBC analyses. This requires being able to run the ihope pipeline in Lyon/Bologna, and offer substantial computing power.
Currently, ihope runs at dedicated LSC clusters and relies on the "Condor" scheduling system. However, developments are going on within the LSC to be able to run ihope across the GRID, which make it a realistic goal to run *ihope* pipeline in French – Italian CCs next year.
Under the assumption that the *ihope* pipeline can be run on LSC-Virgo data in 2009, a reasonable goal is to contribute to the CBC searches with about 200'000 kSI2k.day, corresponding to running full time over a 200 nodes cluster.

1.4.8 Comparison with LSC computing costs

Typical figures quoted by the LSC, when considering the cost of a CB analysis indicate that 100 processes running full time are required to carry out an analysis in time with a data taking. Assuming each process runs at 1.2 kSI2k power, a single reprocessing of the full 2nd year of S5 in coincidence with VSR1 would take about 54'000 kSI2k*day.

This figure does not take into account the need of doing reprocessing, for fixing bugs, and tuning up the codes, but does include the calculation of the detection efficiency by means of software injections.

The reprocessing typically increase the costs by factors of about 5; to this cost for S5/VSR1 one should add the prospective costs for starting the analysis of VSR2 data in 2009, which are expected to be larger because of a wider bandwidth. Therefore a request for 2009 of about 200'000 kSI2k.day, within a factor 2, seems consistent with LSC computing costs.

1.5 CW Group

The CW groups works on the search of GW signals from deformed rotating neutron stars (isolated or in binary systems).
These are the main activities of the CW group in the last 6 months:

1 analysis of VSR1 data with a hierarchical procedure;
2 development and tests of a new implementation of the Hough transform for the hierarchical blind search;
3 procedures for targeted searches;
4 development and tests of a resampling technique for Doppler correction;
5 start of the work for the search of signals from pulsars in binary systems;
6 refined cleaning procedure for the construction of the short FFT database;
7 comparison of the PSS_Astro software with LIGO LaLsoftware.

In sub-sections 8-11 we also comment about the utilization of computing resources in 2008 and the current, tentative requests for 2009.

1.5.1 VSR1 blind analysis

We have continued to work on the analysis of the first half of VSR1, applying a new, more robust cleaning of the peak maps, needed to remove spectral disturbances with a complex structure, like sidebands of some spectral lines (that emerge only over long times) or features of the violin modes that are not caught by the on-line line monitor (due, e.g., to thermal fluctuations).
The parameter space that has been explored is: all-sky, frequency in [20,1100]Hz, minimum spin-down decay time 20kyr.
The incoherent step of the analysis (based on the Hough transform) has been done on the grid, submitting ~6500 jobs each covering a variable frequency band between 5Hz (at the low frequency end) and 0.1 Hz (at the high frequency end) so that the duration of each job is similar.

First order candidates have been produced, setting a threshold of 3.8 on the critical ratio of the hough maps and their analysis is underway.

1.5.2 New frequency Hough transform

It has been implemented and tested. It is based on the transformation between the time-frequency and the frequency-spin down plane (instead of sky) and results based on simulated peak maps suggest some important advantages respect to the standard one. In particular, the possibility to increase the resolution in frequency only slightly affecting the computation time. This new procedure has been present at the GWDAW12 (Dec 2007).
A study of the behavior using VSR1 data is now underway, in particular concerning the possibility to efficiently reject spurious lines (we are evaluating the possibility to use it also as a tool to clean peak map).

1.5.3 Search for known pulsars

We have developed a procedure to extract frequency bandwidths from the SFDB data, construct the time data subsampled sequence, cleaning the data from huge time peaks in the subsampled series. It has been tested on VSR1 data. The next step, consisting of a spectral matched filter, is being developed.
The procedure to setup upper limits, using software injected signals, is also being developed.

1.5.4 Resampling technique

The work on the resampling technique to correct for the Doppler effect is continued. The core of the procedure consists in producing a “mask” of corrections for each direction in the sky and each spin-down value. This method seems to be particularly suitable for semi-targeted searches in which the source position is known but the frequency is uncertain. Tests has been done and we are now ready for producing masks for the whole VSR1 data set.

1.5.5 Blind search for pulsars in binary systems

The method is based on a bank of matched filters in the frequency domain. The data are divided into short Fourier transforms (SFTs).
For each of these SFTs the full bank is applied in an efficient way and a threshold is set. This threshold is defined by estimating the average noise spectrum of a particular data stretch. The filters which have a response higher than the threshold are called a "hit". After all the data has been analyzed in this way, the hits are plotted in a time-frequency diagram called a "hit-map". From this hit-map it is possible to select the most probable waveform by looking at where the hits have the most overlap. We have done some tests with this analysis method on simulated waveforms from binary pulsars using the SIESTA package and added VIRGO data to the simulation.
After code validation, we will apply the method to the VSR1 data.

1.5.6 A more refined data cleaning, to construct the SFDB data.

We have noticed that our cleaning procedure fails in situations in which the presence of very huge peaks spoils the data even for minutes, tens of minutes in some cases. To face this problem we have refined the data cleaning by applying a threshold, over which we consider the data as "saturated" and we limit them to a fixed constant value. In this way, the presence of highly disturbed data does not spoil the cleaning of the data by removal of time peaks.
If the number of saturated data exceed a given fraction of the length of one basic SFDB, the whole SFDB is vetoed. We are now testing this procedure on the VSR1 data, to fix the value of the parameters of the cleaning. This procedure will be applied to VSR1 data, we will thus procedure a new SFDB data set and new peakmaps.

1.5.7 PSS_astro

We have performed new and very refined checks of the PSS_astro software, comparing outputs of the software with the outputs of the LIGO LaLsoftware.
This comparison was requested by people of the LIGO pulsar group. Results have confirmed the very good accuracy of our software. Reports on this have been put on web.

1.5.8 Computing needs: introduction

The two main search areas, the targeted and the blind searches, differ strongly in terms of the computing power needed: targeted and blind searches.
Targeted searches are directed toward sources for which parameters are known (position, rotation frequency, spin-down), like known pulsars, and are relatively less expensive from the computational point of view.
Blind searches assume the source parameters are unknown and try to explore a parameter space as large as possible. These kinds of searches are computationally very expensive.
The motivations for an expensive blind search are manifold.

- The expected number of neutron stars in the Galaxy, which could be emitting in the sensitive band of Virgo, is much larger than the number of (electromagnetically) observed objects.
- Although neutron stars are expected to be distributed mainly on the galactic disk, at least for the target sensitivity of Virgo/Virgo+, the detection probability is larger for near objects, a few hundreds of parsec from us at most, where the sky distribution is pretty uniform.
- Unless neutron stars can sustain deformations much larger than reasonably expected, it is rather unlikely that known objects emit detectable signals.

It can be easily shown that blind searches over a large parameter space and with a long observation time cannot be done using optimal analysis methods (matched filter) due to the huge number of points in the source parameter space that should be explored. Sub-optimal methods have been developed that reduce the computing requirements to still high, but reasonable, values with a small sensitivity loss.
It is important to stress that **such kind of analysis are computationally bound**: the **higher is the available computing power and the bigger is the portion of parameter space that can be covered**. This would allow, for instance to search for higher frequency sources (which emit stronger signals) and sources with higher spin-down rate (likely to be younger and then, possibly, more deformed).

1.5.9 Plan for computing resource utilization in 2008

The target for 2008 is the analysis of VSR1 data. The two main analysis activities (in terms of needed computing power) that are being be performed at the Computing Centers are:
Hierarchical wide area blind search
We recall that the minimum target is to analyze the full VSR1 data set with the following parameter ranges: all-sky, frequency between 20Hz and 1.5kHz, minimum decay time 10kyr. In practice, the data will be divided into two chunks and coincidences among candidates found in each data set will be done. This will require an energy of 40.000kSI2k*day. This analysis is mainly done on the grid.
The extension of the frequency band up to 2kHz would push the requirement to ~90.000kSI2k*day.
Coherent search over short stretches of data
The target is to analyze the full VSR1 data dividing it in 2-day long chunks, making an optimal search for each and making coincidences among candidates of all the subsets. This will require an energy of ~60.000kSI2k*day. This analysis requires to run MPI jobs.

Then, the MINIMUM total computing power needed for 2008 is ~100'000 kSI2k*day.

The search for CW is not very demanding from the storage and data movement point of view. A few TB (say, <5TB) of data are involved, among input, intermediate and output files, in a typical analysis run for several months of observation time.

Most of the analysis is being carried at Bologna and Lyon Tier-1 (accessible both via grid and through the batch system), but also other computing resource will be used. In particular, the use of grid allows to transparently access other resources both completely or partially dedicated to Virgo. Among these we mention:

- a new Virgo grid farm being deployed in Rome, with more than 400 cores fully dedicated to Virgo analysis;
- the Infn-Pisa farm with ~1000 processors (of which ~100 will be fully dedicated to Virgo);
- the APC grid farm in Paris which is being habilitated to the Virgo VO

In all, Virgo VO can access via grid to ~9000 processors, of which ~1000 fully dedicated. The farm in Lyon and Bologna can be used also via standard batch system. About 100 processors will become available from the Poland Virgo group. The extension of the Virgo VO to other important EGEE sites where Virgo laboratories are present, like Nikhef, and the deployment of new Virgo Tier-2 sites, would be very welcome as it would allow to make wider area and more sensitive searches, thus increasing the scientific impact of this activity especially with regard to the relation/collaboration with LSC.
To this respect, the interoperability of Virgo and LSC computing facilities would be a very important step toward the realization of the "single machine" concept. The emerging of common standards between EU and US grids (e.g. EGEE and OSG) will make this possible in the future. At present, an hybrid solution in which EGEE grid jobs can be submitted to a Condor pool and vice versa should be pursued.

1.5.10 Computing energy used so far in 2008 at the CCs

Until the end of April, 2008 we have used ~6500kSI2k*day (nearly all via grid), mainly for the hierarchical blind analysis of ~1/2 of the VSR1 data set over a reduced parameter space (f<1100Hz, decay time>20kyr). Of these, 55% has been used in Lyon (~17% of the total CP used by Virgo in Lyon) and the rest in Bologna (~85% of the total CP used by Virgo in Bologna).
It is to be noted that the production analysis has been started only recently, after a series of tests, and this explains why the utilization is still far from the requests; with the production running, we expect that the computing energy allocated will be rapidly used.

1.5.11 Requirements for 2009

For 2009 we foresee to extend the VSR1 analysis to the whole frequency band, if this cannot be accomplished in 2008 and analyze data of the second Scientific Run (VSR2) which should start at half of the year. The coincidence analysis with the S6 LIGO run is also foreseen. Moreover, also the pipeline for the search of CW signals from binary systems should start "production" activity on both VSR2 and S6 data.
In all we can estimate a MINIMUM computing energy for 2009 of 240'000 kSI2k*day (that could increase to 300'000 kSI2k*day).

1.6 Stochastic Background (SBGW)

The activity of the group is totally dedicated at the LSC/Virgo collaboration.

1.6.1 Software injections

An extensive set of software injections is planned, using the first week of VSR1/S5 real data. The injections will be at several different SNR and several different spectra, with the main aim of a exhaustive check of the Virgo software pipeline.

1.6.2 Targeted search

It is planned to include VSR1/S5 in this kind of analysis. Currently the analysis code is developed in LSC, and it is not a priority to implement it in Virgo, as there is not enough manpower to do that.

1.6.3 VSR1 data analysis

The search for an isotropic stochastic background is currently in progress on the Caltech cluster. A complete analysis with an artificial time shift has been performed, and the results will be presented at the next LSC/Virgo meeting in June.
Concerning the Virgo side, the possibility of using the GRID farm in Pisa is confirmed. The data analysis software has been ported to the GRID environment. A limited number of modifications has been introduced in the SB/NAP library in order to do that, mainly introducing the possibility of directly and transparently access the data published in the GRID catalog.
In particular, a patched version of the Frame library has been produced and tested with success, and a set of specialized input/output classes has been introduced in NAP.
The ported code has been tested with success using a limited amount of data. The next step is the acquirement of a large enough storage volume to allow the publication of the GRID catalog of the entire set of reconstructed h for the VSR1/S5 run. The completion of this task is foreseen for the mid of June, and after that the Virgo pipeline will be applied to the full set of data, with the following priorities:

- Software injections on the first week of data
- Noise characterization of the full dataset
- Time shifted analysis on the full dataset, isotropic model with several power law spectra
- Analysis of the full dataset (without time shift), isotropic model with several power law spectra
- Bayesian determination of the power spectra, analysis with time shift
- Bayesian determination of the power spectra, analysis without time shift

1.6.4 Non Gaussian backgrounds

There is an activity in the Nice group aimed at the study of non Gaussian stochastic background, in the context of astrophysically motivated models. The main investigator in this activity is Tania Regimbau, that will be helped in the near future by a dedicated postdoc.
This subject is currently discussed in the LSC/Virgo group, and there is the concrete possibility that a dedicated data analysis pipeline will be implemented. If this will be the case, it will be very natural to apply it not only to the VSR1/S5 data but also to the full S5 dataset.

1.6.5 Hardware injections

The hardware injections performed during the VSR1/S5 run are affected by several problems, mainly because of the not well tested injection code in Virgo. The code has been fixed now and their problems understood, but we were not able to run a final completely successful synchronized injection.
The option of doing a Virgo only test during the foreseen next engineering run, mainly with the purpose of checking calibration and synchronization issues, is no more available owing to the anticipated Virgo shutdown.

We reconstructed a limited amount of synchronized injections trying to fix the known issues, and some information will be obtained using the results of hardware injections of other groups. But it remains to be understood if this will be enough to validate the synchronization.

1.6.6 Computational and storage requirements

The computational requirements are quite modest for the foreseen analysis. The computation model is a simple pipeline, which require sequential access to a limited amount (4 in the worst case) of time ordered streams of data.
The floating point operations required for the analysis of 1 second of data with 4 streams and no resampling is given by 40 KFLOP. Multiplying for the duration of the VSR1/S5 dataset we get 400 GFLOP, which grows to 4 TFLOP introducing an expected efficiency of 10%.
Here "complete analysis" means the search for a particular kind of stochastic background (a given spectra) in the full common dataset. Several of these analysis are planned, of the order of 10^2. The expected total number of floating point operations needed is given so by 400 TFLOP. If we want to be able to complete the full analysis in a week this means 600 MFLOPS for the VSR1/S5 dataset. One should keep in mind that we will be largely dominated by the storage access time in any case, so computational power is not an issue.
The analysis produces a very limited amount of processed data, which can be stored and post processed with a negligible computational and storage cost.
The main requirement for the storage comes from the necessity of having all the VSR1/S5 data set available. This is possible both in the CNAF and IN2P3 computational centers, and as discussed before the appropriate amount of storage (6 TB) will soon be available on the Pisa GRID farm.
Concerning this last computational resource, it must be underlined that

- The use of this resource is at this level "parasitic", in the sense that no resource has been spent by the collaboration until now in it. In the periods of intensive data analysis activity of other collaborations (mainly CMS) its availability will be strongly reduced.
- The storage will be financed using resources of the Pisa group, which is interested in having a test bed for GRID applications.
- As a byproduct, the full VSR1/S5 h reconstructed dataset will be published on the GRID catalogs, and available to other groups in the collaboration which will want to do data analysis activities in this environment.

1.6.7 Manpower

The stable members of the group are currently 3 (two in Pisa and one in Nice). There will be soon a couple of postdoctoral physicist, one in Pisa (starting in September) and one in Nice.

1.7 Data services and tools

Since the last STAC meeting, the production of data of scientific interest for the Virgo Collaboration was limited to a fraction of nights and weekends (approximately summing up to about 20 days of science data) and have not constituted a challenge for data replication.
The activity has focused, instead, on fixing some residual issue with the transfer of S5/VSR1 data, and on starting the preparation for data transfer activity to be run during the S6/VSR2 run in 2009.

1.7.1 Data Replica

As reported to the previous STAC, the transfer of LIGO data to VIRGO during the S5 run was affected by several problems, leading to a delay of several months in publishing LIGO data at the CNRS and INFN computing centers, for offline usage.

The solution adopted during the run for the transfer, namely to run at the Caltech computing center a set of scripts and tools to copy data towards Virgo, failed because of a combination of technical limitations and misunderstandings in interfacing the catalog system at Caltech with the Virgo data transfer tools.

This problem has now been circumvented by adopting a LIGO tool (Lightweight Data Replicator, LDR) to perform the data transfer. The LDR has been deployed in Cascina because of the need to further process the files received, and pack them in larger files more suitable to be stored efficiently on the Lyon HPSS system.

The LDR has allowed to replicate 4.5 months of h-reconstructed LIGO data to Cascina in about 10 days; these data were then transferred to Lyon by means of SRB at the beginning of March 2008. We had however another trouble due to a configuration mistake on the LIGO side, which caused a fraction of the files, corresponding to 100 hrs of data of the H1 detector, not to be copied to VIRGO. This additional issue was fixed in the second half of March, and it is fair to say that only at the end of March we could have all the LIGO data available, a delay which caused several problems to the Virgo components of the analysis groups.

Clearly LIGO and VIRGO need to work together to ensure that this situation does not occur again in the future. To this end, the overall data transfer strategy has been reviewed and a baseline has been defined which should ensure a working solution.

The present baseline for handling LSC <-> Virgo data transfer is the following:

- The transfer of Virgo reduced data to LIGO is done using Virgo tools, which are able to track the creation of files on the Cascina disks and schedule their transfer rapidly, as demonstrated by the few minutes required for publishing Virgo data at Caltech cluster.
- The transfer of LIGO reduced data to Virgo on the Caltech -> Cascina leg is instead performed using LDR. The further replica from Cascina to the Lyon and Bologna Ccs is performed independently, to simplify the architecture and minimize service interruptions, at the price of an increase in bandwidth utilization. The possibility to eliminate the Cascina stop for the files depends, among other things, on the evolution of the LIGO format towards larger files.

Concerning the transfer of raw and reduced Virgo data to the computing centers, no changes are expected in the present architecture, except that Virgo Collaboration and EGO have agreed that the responsibility for the transfer is fully entrusted to the EGO Computing Department, which is also responsible of choosing the most appropriate technical means and architectures, including the choice among star or chain architectures for replicating the raw data to the computing centers.

Over the remaining months of 2008 and the first quarter of 2009 the data transfer activity is expected to be limited, except the replica of new versions of LIGO and Virgo reduced data sets, and possibly the replica of data from commissioning runs.

In view of VSR2/S6, the baseline solution appears sufficient to ensure an efficient and smooth transfer; however, it appears sensible to investigate whether the LDR system could be adopted to perform the transfer directly to the computing centers.
This appears technically possible but needs still to be checked for compatibility with the CCs rules, and then implemented and tested in practice.

1.7.2 Data distribution

The VSR1 raw data have been transferred to the computing centers and except a few interesting data segments the copy stored in Cascina has been deleted to recover space for the circular buffer used in the Commissioning activity.

The access of raw data is now possible from the computing centers using a variety of means, which allow the commissioners to use remote data as they were local with their visual tools.

In Bologna, a frame server allows to access data remotely on the basis of GPS time.

In Lyon, the middleware SRB provides a frame access library which enables remote listing and access to frame files stored on the HPSS system.

While some groups are presently able to access data over the GRID, Virgo as a whole does not have a general, unified GRID based approach to data access and distribution; the NIKHEF group has been entrusted the responsibility to study the problem and propose solutions.

1.7.3 Bookkeeping database

The Virgo Bookkeeping Database, online at the address http://vdb.virgo.infn.it, is now routinely used by the Burst and CBC groups to store information about the data location, the definition of run segments, as well as tables of status conditions and vetoes.
The database stores now both Virgo and LIGO tables, thus allowing to easily cross the information and to build complex queries using a simple graphical interface, or directly using mySQL commands.
The Data Quality group is in fact active since a few months in defining and filling the tables of Virgo veto and quality information, checking the content and publishing on the database.
Contact persons in LIGO provide the guidance to upload similar information about LIGO data.
In fact, the VDB tool is far more efficient than the corresponding LIGO tools, and appears to be a very good contribution by Virgo to the joint effort.

1.8 Offline computing in Bologna and Lyon

We recall that the off line computing for Virgo is performed mainly in the two computing centers of INFN and IN2P3, located in Bologna and Lyon. Some of the groups also access GRID resources at Tier 2 sites.

1.8.1 Resources used to date at INFN Tier1-Bologna

The Bologna CNAF center is used both to store data, mostly on spinning media for immediate use, and to perform offline analysis on the Linux computing cluster.

Data taken during older commissioning runs are stored on CASTOR plus, for a total of 12 TB.
Data collected during the VSR1 and previous Weekly Science Runs are stored on spinning media, under the GPFS file system.

Status of the storage

At present in Bologna the Virgo collaboration has available 108 Terabytes, of which 101 TB are dedicated to data and 7 TB to processed data and user space.
Of the assigned space, 94 TB are used, and 14 TB are available.

CPU utilization

As of May 2008 Virgo has used at CNAF 3450 kSI2k.day, mainly for CW searches.

1.8.2 Resources used to date at CCIN2p3 - Lyon

Lyon computing center (CCIN2P3) is used to store permanently the Virgo data. That includes:

1. all streams recorded during data taking periods since the first commissioning run (E0);
2. all 50Hz files since 2002;
3. all trend data files since 2001.

Since the beginning of the joint LSC-Virgo S5 SR1 data taking, Lyon stores the 4 additional processed data streams sent from Caltech by LIGO.

Beginning 2008, the S5 LIGO data files have been successful transferred from Cascina to Lyon into HPSS thanks to the SRB (Storage Resource Broker) tools developed for inter-site data replica and access. SRB provides especially the facility to merge on the fly set of files before storing them in HPSS. This facility is mandatory as HPSS cannot accept short files. SRB has been also used to transfer the Monte Carlo data sets from Caltech to Lyon HPSS, merging files on the fly.

CCIN2P3 is also intensively used for off-line data analysis. The main use of the computational resources is done by submitting jobs via standard batch queue (BQS). Since the beginning of 2008, the CPU consumption has increased compared to previous years due mainly to the burst search analyzing VSR1 and S5 data. CW search jobs are also submitted through Grid. We summarize the Virgo resources used in Lyon, as of May 18th:

Storage:
128 TB used in HPSS for all data taking periods since 2001
2TB of disk has been recently added to the existing 300 GB to allow the h(t) VSR reprocessing.

CPU:
Use of the CPUs since 2008 January 1st: 21820 kSI2000.day. Virgo consumption represents for the moment 3.5% of the total CPU consumed by all the other experiments performing data analysis in 2008.

1.8.3 Year 2008 computing cost estimates

The computing costs forecast computed in April 2008, on the basis of requests placed at Bologna and Lyon in 2007, can be summarized as follows:

CC-IN2P3 (unit costs: 0.533E/ KSI2K.day, 0.633E/GB disk, 0.15E/GB tape)
CPU: 120000 KSI2K.days → 64 k€
Disk storage: new 60TB Xrootd cache → 38 kE
Tape storage: 50TB more in HPSS → 7.5 kE
Virgo/User space : 5TB → 7 k€

CNAF (unit costs set equal to those of CC-IN2P3)
CPU 100000 KSI2K.days → ~51k€
Disk storage: new 30TB on the farm
+ new 10TB for users → ~35k€
Tape storage: 200TB in Castor → ~30k€

The total cost of the computing is therefore estimated at 240k€, and starts to be dominated by the computing, no more just by the storage. This is a good sign since it shows the progress of the analysis activity.

However the figures for CPU utilization appear outdated; the latest 2008 forecast for the use of CPU resources, resulting from the previous sections of this document, is

Burst: 250'000 kSI2k.day
CW: 100'000 kSI2k.day
CBC: 1'000 kSI2k.day
SBGW: 0 (not running at the CC)

for a total of about 350'000 kSI2k.day.
It is to be remarked that the CW request is believed to be a minimum needed, while the main increase is due to the Burst group forecast, partially compensated by a review of the CBC forecast towards smaller figures.

The extra computing energy of 130'000 kSI2k.day would correspond to an extra cost of 70k€, for a total cost of the offline computing of 310k€ in 2008.

1.8.4 Comparison with LSC resources

The LSC has available about 6'000 nodes full time, not including the new 5'000 nodes cluster in Hannover. Assuming each of these nodes is a dual core capable of delivering 1'000 kSI2k.day over one year utilization, the total computing energy available to LSC is of about 6'000'000 kSI2k.day, roughly twenty times the 2008 Virgo forecast. This figure might easily be wrong by a small factor depending on the class of the CPU considered, but would be more than compensated by the new, recent Hannover cluster.
If we consider the relative size of LSC and Virgo, and estimate at 5 to 1 the ratio of people actually running searches in the respective collaboration, we see that the Virgo forecast of computing energy / researcher is still short by at least a factor 4 with respect to the LSC one.

1.8.5 Year 2009 trends and preliminary estimates

The preliminary computing estimates for 2009 by the different search groups are the following
Burst: 500'000 kSI2k.day
CW: 300'000 kSI2k.day
CBC: 200'000 kSI2k.day
while the SBGW group plans not to use the resources at the CCs.

This gives a total of 1'000'000 kSI2k.day which should be taken with great care since numbers could still vary by small factors. Particularly the CW might turn to be an underestimate.

Such an energy roughly correspond to the computing power delivered by a 1'000 nodes computing center, to be compared to the 11'000+ nodes available in 2009 to the LSC.
We believe that these figures are therefore realistic and probably underestimating the real needs, when considering a prudent 5 to 1 ratio of researchers active in analysis.

At the 2008 costs, the CPU alone would therefore cost about 500 k€, to which one may expect to add O(100 k€) for storage and user disks.

1.9 Answers to November 2007 STAC recommendations

Quoting the November 2007 STAC report (EGO-PRE-STAC-94)

1. The Data Analysis activities are proceeding very well. The team is well organized, the methods and plans are clearly stated. **The STAC considers that the collaboration should avoid concentrating on data processing only and should be more active in data analysis and interpretation if it wishes to play a leading role in this field. The STAC thinks that such a role is an essential part of the Virgo success. Of course the collaboration is undermanned compared to the LSC and cannot compete with it in quantitative way but leadership is not about quantity but quality.**

The Virgo components of the search groups have undertaken several initiatives in order to have a visible impact on the LSC-Virgo collaboration, by means also of *flagship* searches:

- a coherent high frequency (above 1-2kHz) search for burst events has been proposed, implementing modifications to the LSC algorithm "Coherent Wave Burst" which have enabled it to run efficiently over a wider frequency band. The first analysis tests are encouraging and the LSC-Virgo collaboration is considering to extend the application of the modified method to the analysis of data prior to the start of the data sharing agreement, with an obvious impact on the publication plans.
- An all-sky search for burst events based on a Virgo algorithm (EGC), targeting the intermediate and high frequency range (300Hz - 5 kHz), has been proposed as an original contribution to the analysis of coincident VSR1 - S5 data. The method is based on the coincidence of pairs of detectors, which allows to optimize the sky coverage and is expected to lead to an increase of detection efficiency up to 50% if compared with a search requiring coincidence among all detectors; preliminary results are encouraging. The method requires a large computing power, which motivates an increase of our requests for computing budget, that we have tried to justify in the burst search section of this report.
- The development of data quality (DQ) and veto information for Virgo data is being carried out by Virgo people. New DQ are regularly developed and released for the use by the search groups.
- The Virgo Data Base system has been proposed by Virgo as the solution to the complex bookkeeping problem represented by the many detector and veto conditions on the different instruments, and is being considered as a possible reference tool.
- In the CBC group, the searches around GRB for which Virgo has a favorable orientation will be led by Virgo folks.
- The CBC group made also an effort to contribute to coherent searches, and to techniques exploiting higher harmonics content in the signal; these may not lead to a leadership in the short term but should keep the group up to date with promising developments.
- The Virgo CW group is pursuing a strategy for targeting low frequency pulsars, leveraging on the better Virgo sensitivity, and the calibration team has supported this initiative by a large effort to provide accurate data down to 10Hz.
- In parallel, the CW group keeps developing original methods for noise removal and for targeting pulsars in the frequency – spin down parameter space more efficiently than with conventional methods. These techniques are likely to become reference techniques for LSC-Virgo.
- The SB group is investigating the detection prospects of non-Gaussian stochastic background, as resulting from astrophysical sources. This could lead to the

development of a pipeline, which could be applied to the entire S5 data and not only the portion coincident with VSR1.

It is to be underlined again that acquiring and keeping a leadership in some search has a price in terms of computing costs. In order to carry out a credible search, to win a consensus about its sensitivity and correctness in front of the LSC colleagues, it is necessary to invest a computing power consistent with what the LSC is investing.

One reasonable criterion is to spend a similar amount in computing for each FTE in data analysis; our requests for 2008 and 2009 aim precisely at this objective, which would procure an equal status to Virgo DA scientists in the LSC-Virgo collaboration.

For this reason, we are asking the STAC to review the computing requests and endorse them.