VIRGO off-line activity:data flow, data distribution and computational problem

Fulvio Ricci

Dipartimento di Fisica Università di Roma La Sapienza

&

I.N.F.N. Sezione di Roma

Outline

- Virgo: detector complexity and constraints for data acquisition and analysis
- The Virgo on-line software, the data flow and data distribution
- Off-line computing: the data Analysis
  - **The general strategy for G.W. data analysis**
  - **Main examples of data analysis**
    - **Search of signals from Coalescent binary stars**
    - **Full sky neutron stars search**
    - **Stochastic background**
- Computing necessities and Conclusion

Detector Complexity Constraints

- VIRGO has many different sub-systems

  (vacuum, laser, suspensions, DAQ,...)

  ⇒separate R&D test, commissioning -> modularity

- VIRGO is a set of feedback loops

  ⇒main feedback loop (mirror pos.): sampling @ 10 kHz.

  ⇒main physics signal (h): sampling @ 20 kHz

  ⇒the controls and the ‘physics’ are strongly coupled

  ⇒the data acquisition spies the detector

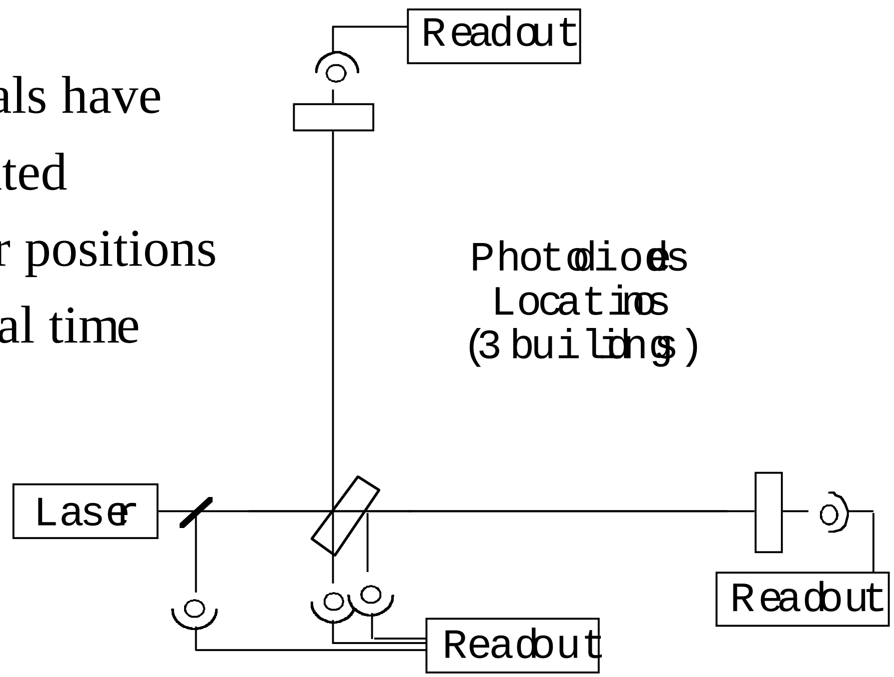
  ⇒on-line/off line software should follow the same design principle and obey the same quality criteria

Detector Size Constraints

- Experiment distributed over 4 buildings up to 3km apart

example:

The Photodiode signals have
to be globally correlated
to compute the mirror positions
and act on them in real time



⇒data transmission issues (timing, synchronisation, protocol,...)

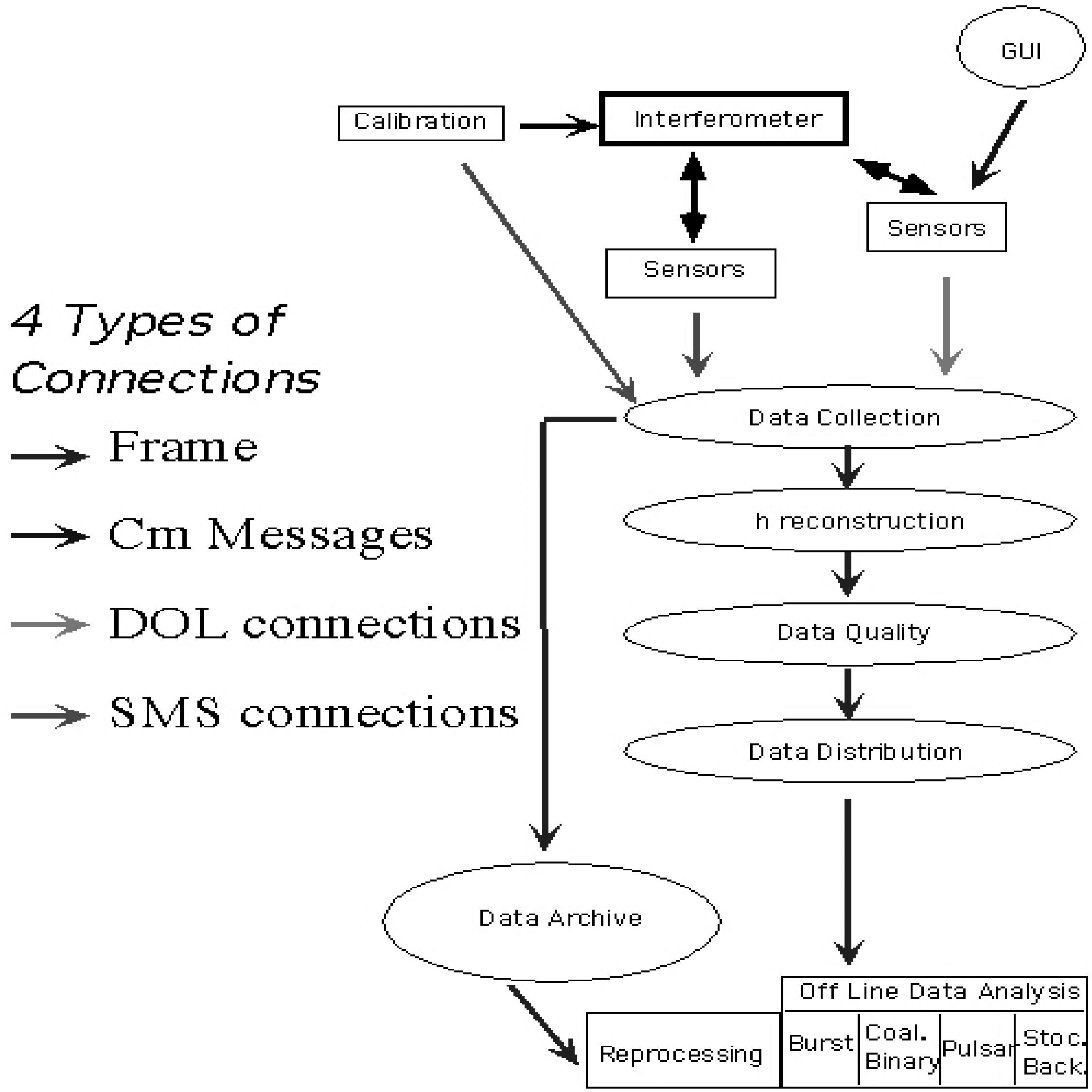
⇒networking issues

⇒digital feed back loop

Constraints Summary

- Distributed system: Communications issues
- Modularity
  - **Modularity, portability**
  - **Interfaces, (protocol, format)**
  - **The software has to run on many different platform**
- From Real Time to Offline Software
- Reliability (Quality)
  - **Use long term standard**

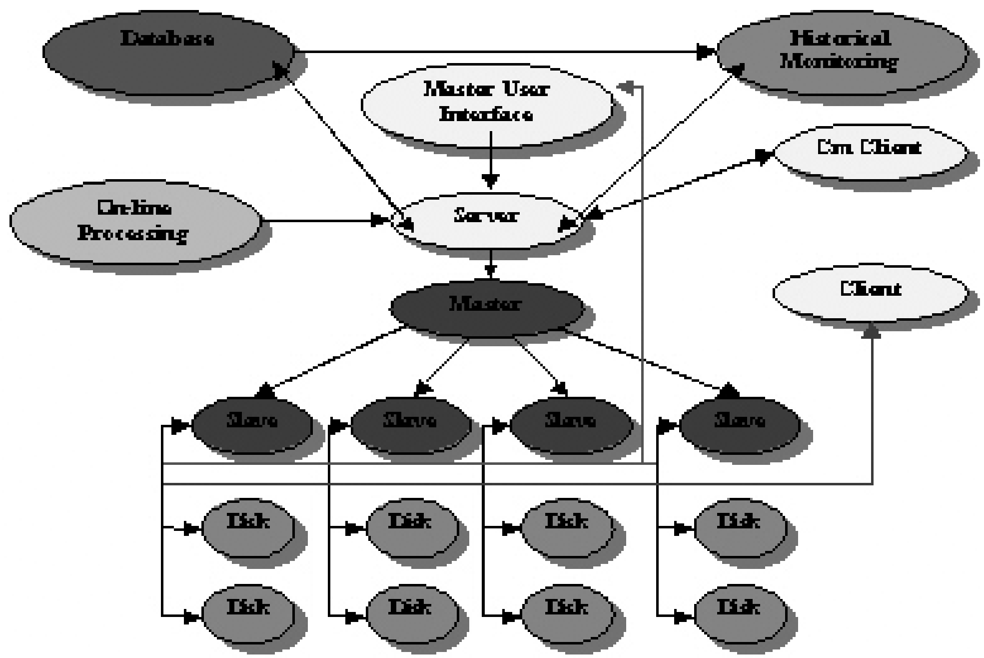
Data Collection and data flow

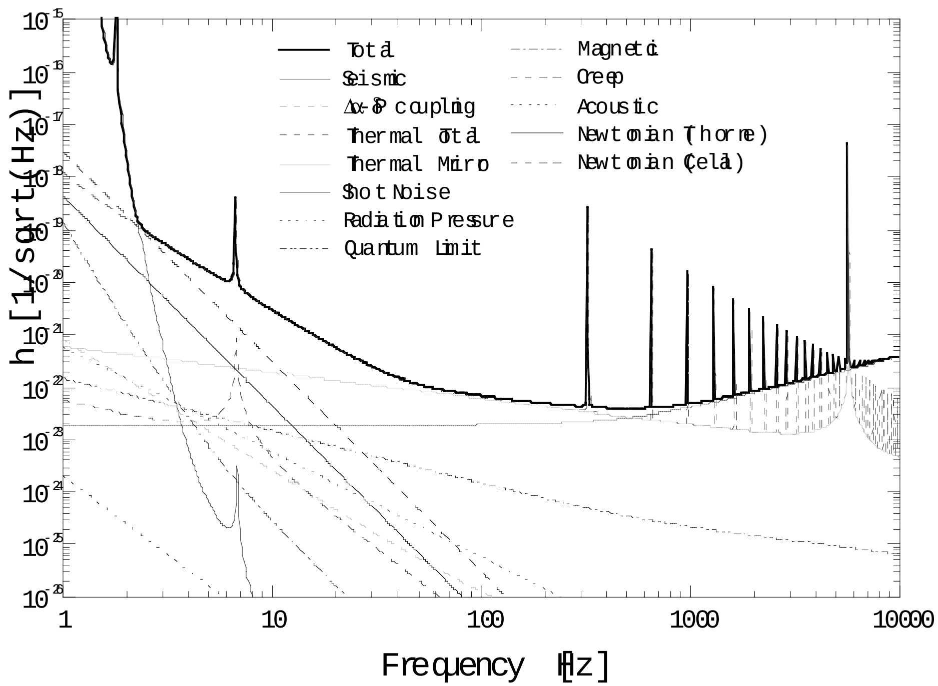


	Raw data	Processed data	Selected data
Flow/s	4.0 MB	.3 MB	.34 MB (1%) .7 MB (10%)
Flow/day	350 GB	26 GB	30 G B(1%) 60 GB (10%)
Flow/year	126 TB	9.5 TB	11 TB (1%) 22 TB (10%)
Tapes/year (50 GB DLT)	2500	190	220 (1%) 440 (10%)

Data Distribution

Process Link





The Spectral sensitivity curve of Virgo: $\{S_h(f)\}^{1/2}$ vs. f

Data analysis: a basic concept

Basic Concept in the G. W. data analysis: <u>the Linear Optimum Filter</u>

The detector output ***o(t)*** is the sum of the signal ***h(t)*** and the noise ***n(t)*** *:*

$$o(t)=h(t)+n(t)$$

The “matched filter” is a linear technique of “patter-matching” that permits to enhance the signal to the respect of the noise by optimizing the Signal to Noise ratio (**SNR**).

To apply this technique we need as filter input

-An hypothesis concerning the form of the signal (or its Fourier transform) ***H(f)***

-The spectral properties of the noise ***S(f)***

Thus, it can be demonstrated that the optimum filter output ***c(t)*** is

$$c(t)=k \int exp(i2\pi\ f\ t)\ H^*(f)\ [\ O(f)\ /\ S(f)\ ]\ d\ f$$

In the time domain the filter corresponds to the convolution of the output with the waveform function. In a discrete domain

$$c(t) = \Sigma_j \; w(t-j)\, o(j)$$

where the index *j* spans from the past to the future (no causality) and *w* is the filter weight defined by the detector noise and signal model.

The filter acts on both signal and noise so that the filter variable ***c*** is the sum of the filtered noise ***N*** and the filtered signal ***S***. Thus, the analytical expression of ***c(t)*** is derived by imposing the maximisation condition to the *SNR* filtering function

$$SNR = \langle c,h\rangle^2 / \langle c,c\rangle =$$

$$= \left\{\int C^*(f)\,[H(f)/S(f)]df\right\}^2 \quad \left\{\int [(C^*(f)C(f))/S(f)]\,df\right\}$$

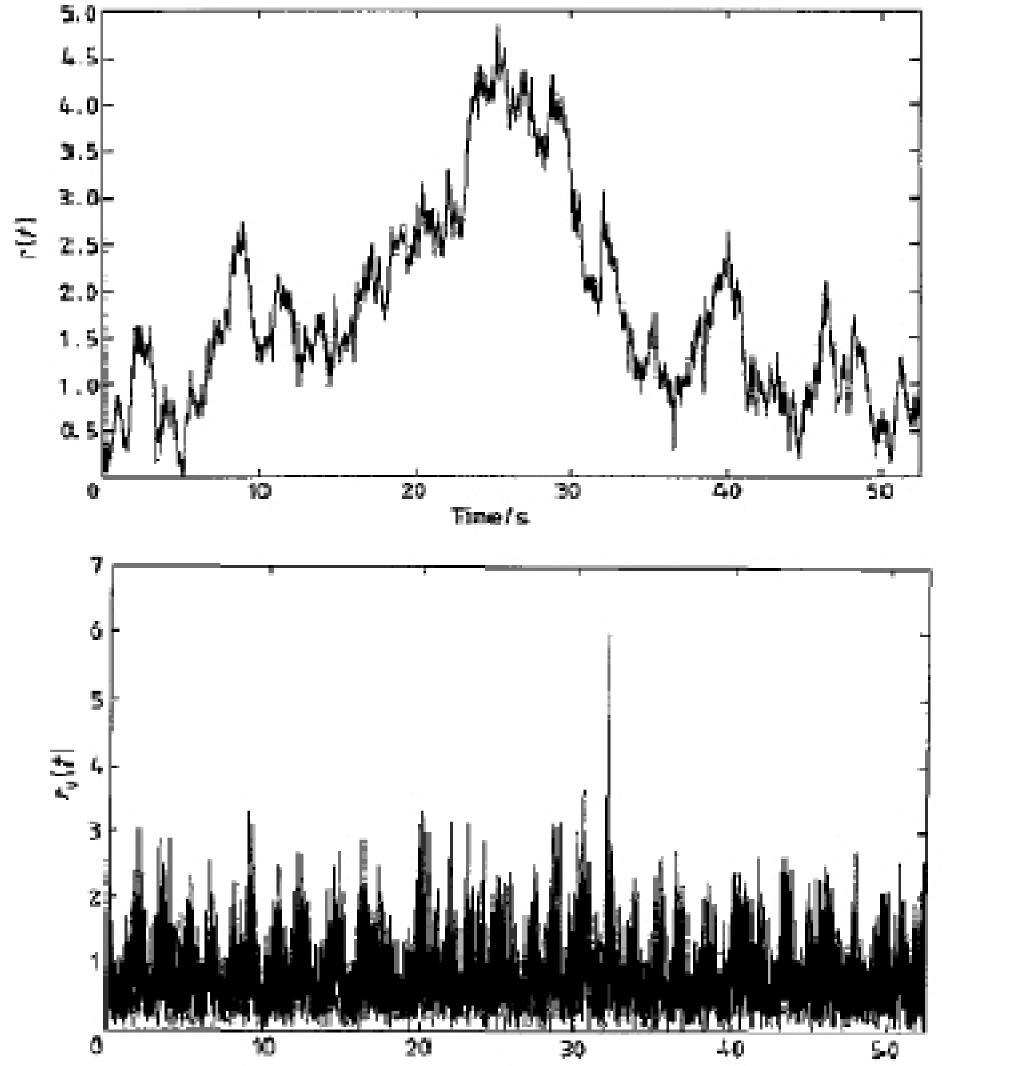
$$= O(c,h)\,\langle h,h\rangle$$

where

$$O(c,h) = \langle c,h\rangle^2 / (\langle h,h\rangle \langle c,c\rangle)^{1/2}$$

Data Analysis

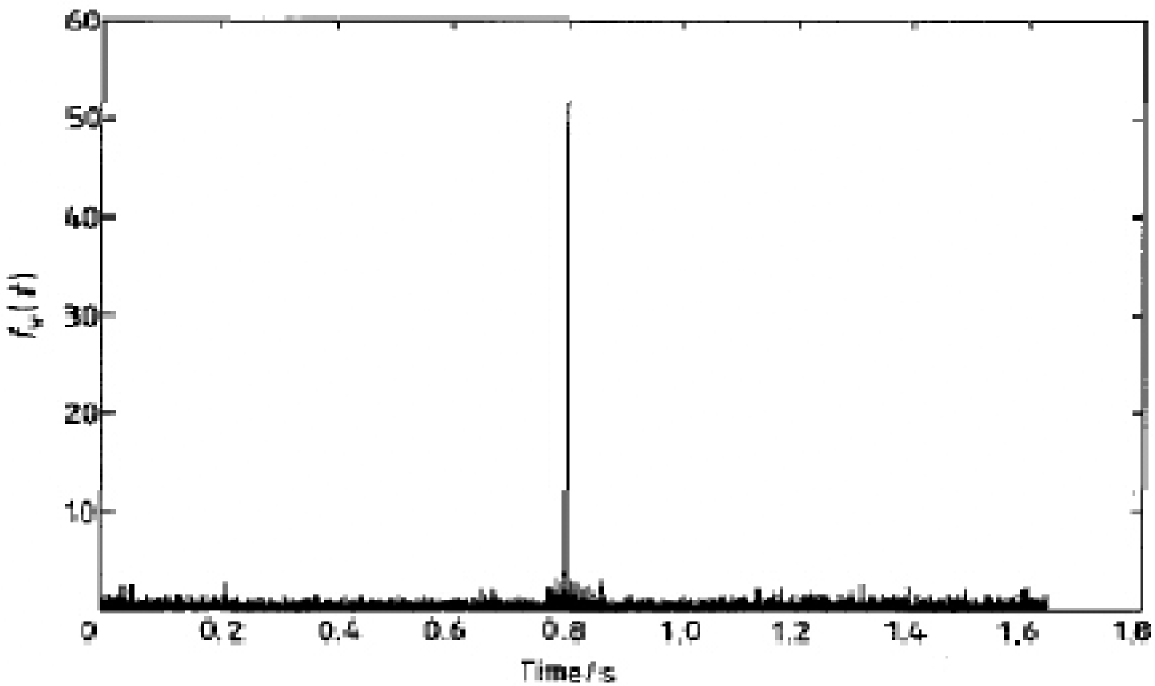
Application of the optimum filter to the data of the Garching (Germany) interferometer. Here a short burst signal has been added to the data and the filter is designed for a δ-like function



Band limited output

Filter variable derived from the above output signal

Filter variable derived from the original output signal (bandwidth unlimited)



Data Analysis Items

-simulation
-noise studies
-h reconstruction
-search for bursts from Supernovae explosions
-Pulsars in binary systems
-stochastic background signal
- ………….
-coincidences among different detectors and G.W. network analysis

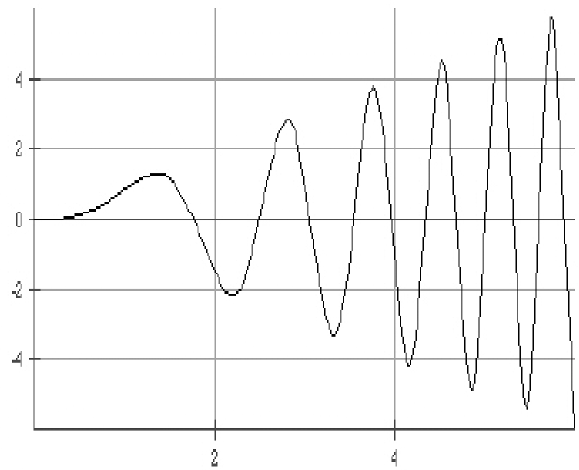
Signal search based on high demanding computing task:

-*Signal from Coalescent binary systems*

-*Continuos signals from Neutron Stars*

Coalescent binary system

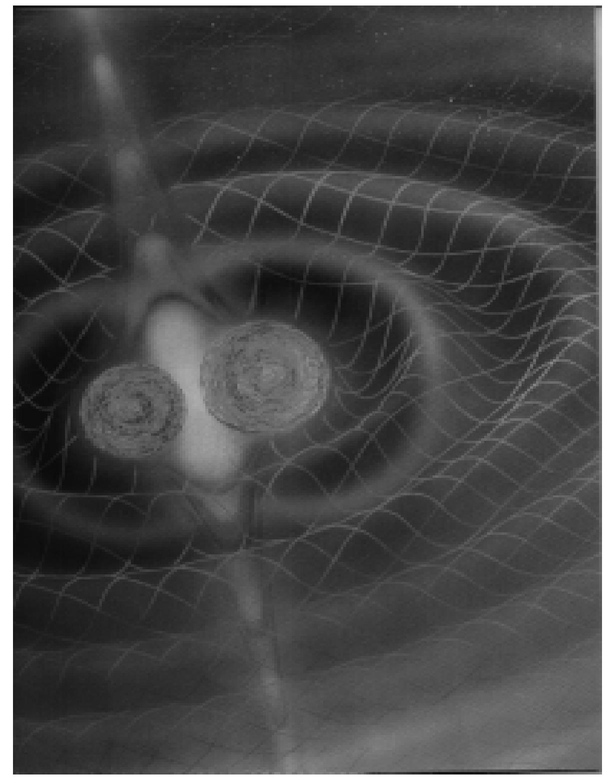
h(t)



time

Two compact stars are rotating one to respect of the other while the orbital radius is decreasing progressively up to the final stable orbit.

Coalescent binary system



The CHIRP

(expected signal from a coalescent binary system)

$$h(t) = (A/r)\, M\, (\pi M F)^{2/3}\, e^{-i\Phi(t)} \qquad \text{with}$$

$$\Phi(t) = \Phi_c - (1/16)\, [256\, (t_c - t)/(5M)]^{5/8}$$

where

$$M = [(m_1\, m_2)^{3/5} / (m_1 + m_2)^{1/5}]$$

$$F(t) = (1/\pi M)\, [5M / 256\, (t_c - t)]^{3/8}$$

Φ_c t_c phase and time of coalescence Ë $F(t_c) = \infty$

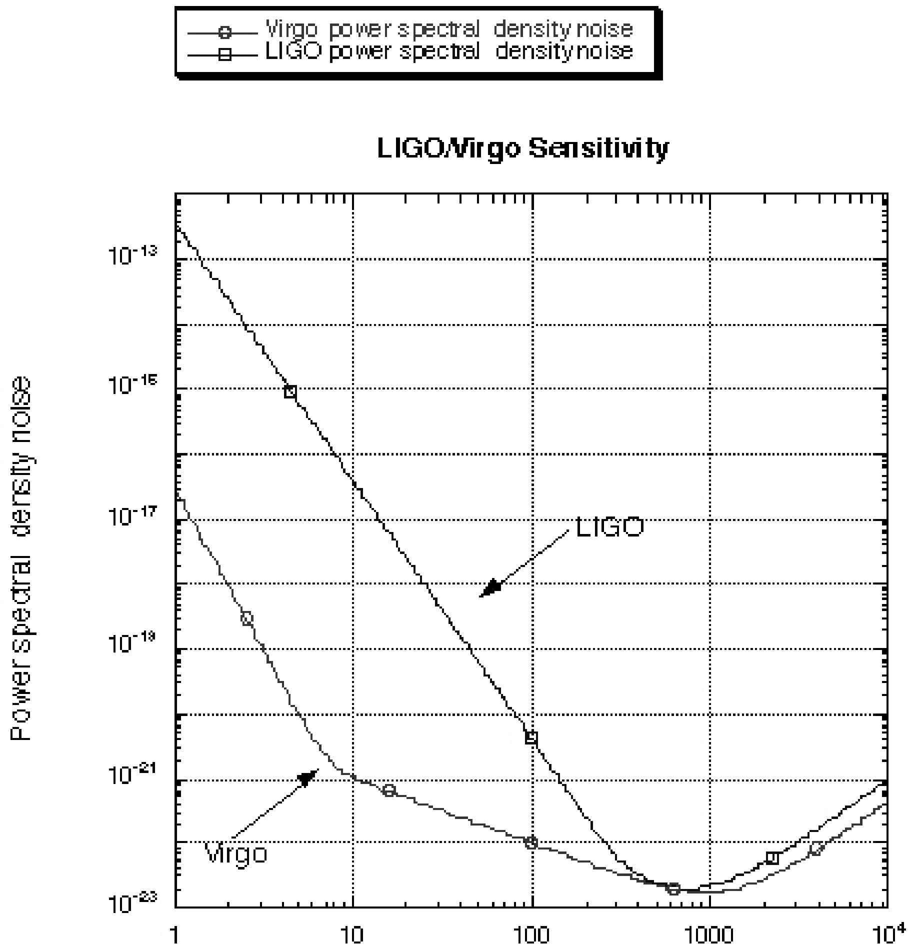
$$\Phi(t) = 2\pi \int F(t')\, dt'$$

The Fourier transform of $h(t)$

$$H(f) = [(30\pi)^{1/2} / 24]\; [A M^2/r]\; (\pi M f)^{-7/6}\; e^{i\psi}$$

with

$$\psi = 2\pi f t_c - \Phi c - \pi/4 + (3/128)(\pi M f)^{-5/3}$$



SNR $\rightarrow$ $f_{min}^{-7/6}$

The chirp is tracked during its evolution for a time T_{chirp}

$$T_{chirp} = 2.9\,(100/\,f_{min}\,)^{8/3}$$

LIGO I $f_{min} \sim 100$ Hz **$T_{chirp} \sim 3$ s**

VIRGO $f_{min} \sim 10$ Hz **$T_{chirp} \sim 1400$ s**

$$SNR_{VIRGO}/SNR_{LIGO\ I} \sim 15^{2}$$

Longer waveform Template, Higher Computing Power !

HOW MANY TEMPLATES?

This number depends on the dimensions of the parameter space and its discretization. The choice of the metric of the parameter space is based on the optimization of the event detection probability.

To compute the values we assumed that the event are uniformly distributed in the parameter space and the event number is proportional to the ratio $(SNR/ SNR_{max})^3$

A bank of templates with 40 Hz cut off costs 450 Gflops and 1.1 Tword with 15% of event loss

A factor 3 in computing power and a factor 30 in memory is required to lose less than 10 % in events

Results

f_{min}(Hz)	M/Mo	P (flops)	S (words)	Events
20	0.25	1.5 [1.8] $*10^{12}$	3.2 [24] $*10^{13}$	90.7 %
20	0.5	3.0 [3.6] $*10^{11}$	2.2 [16] $*10^{12}$	90.7 %
20	1.0	5.6 [6.7] $*10^{10}$	1.4 [3.6] $*10^{11}$	90.7%
30	0.25	8.5 [10] $*10^{11}$	6.4 [49] $*10^{11}$	88.7%
30	0.5	1.6 [2.0] $*10^{11}$	4.4 [32] $*10^{11}$	88.7%
30	1.0	3.1 [3.7] $*10^{10}$	3.0 [21] $*10^{10}$	88.7%
40	0.25	4.5 [6.0] $*10^{11}$	1.1 [14] $*10^{12}$	85.1%
40	0.5	8.6 [1.1] $*10^{10}$	7.4 [88] $*10^{10}$	85.1%
40	1.0	1.7 [2.1] $*10^{10}$	5.0 [57] $*10^{9}$	85.1 %

Results in [ ] are for a standard one step search strategy

CONTINUOUS SIGNAL

(G.W. emission from asymmetric Neutron stars)

Periodic gravitational waves are emitted in many processes involving single stars or binary systems. For a single rotating deformed star the gravitational signal emitted is:

$$h(t) = h_o \cos(2\,\Omega\, t)$$

Ω angular velocity of the star

$$h_o = (4\,G / c^4 \delta)\; I\; \Omega^2\, \varepsilon$$

I momentum of inertia; ε ellipticity; δ distance

We are interested in periodic sources radiating in the Virgo sensitivity band. These involve compact stars, namely NEUTRON STARS.

About 10^9 NEUTRON STARS are expected to exist in the Galaxy, but on ~ 1000 have been detected, most as PULSARS.

Pulsar spin-down

It can be due to many causes, among which emission of electromagnetic radiation and, hopefully, gravitational radiation.

If the loss mechanisms of the star is just the G.W. emission, the neutron star ellipticity is

$$\varepsilon_{max} = 1.9\ 10^{5}\ [(df / dt) / f^{5}]^{1/2}$$

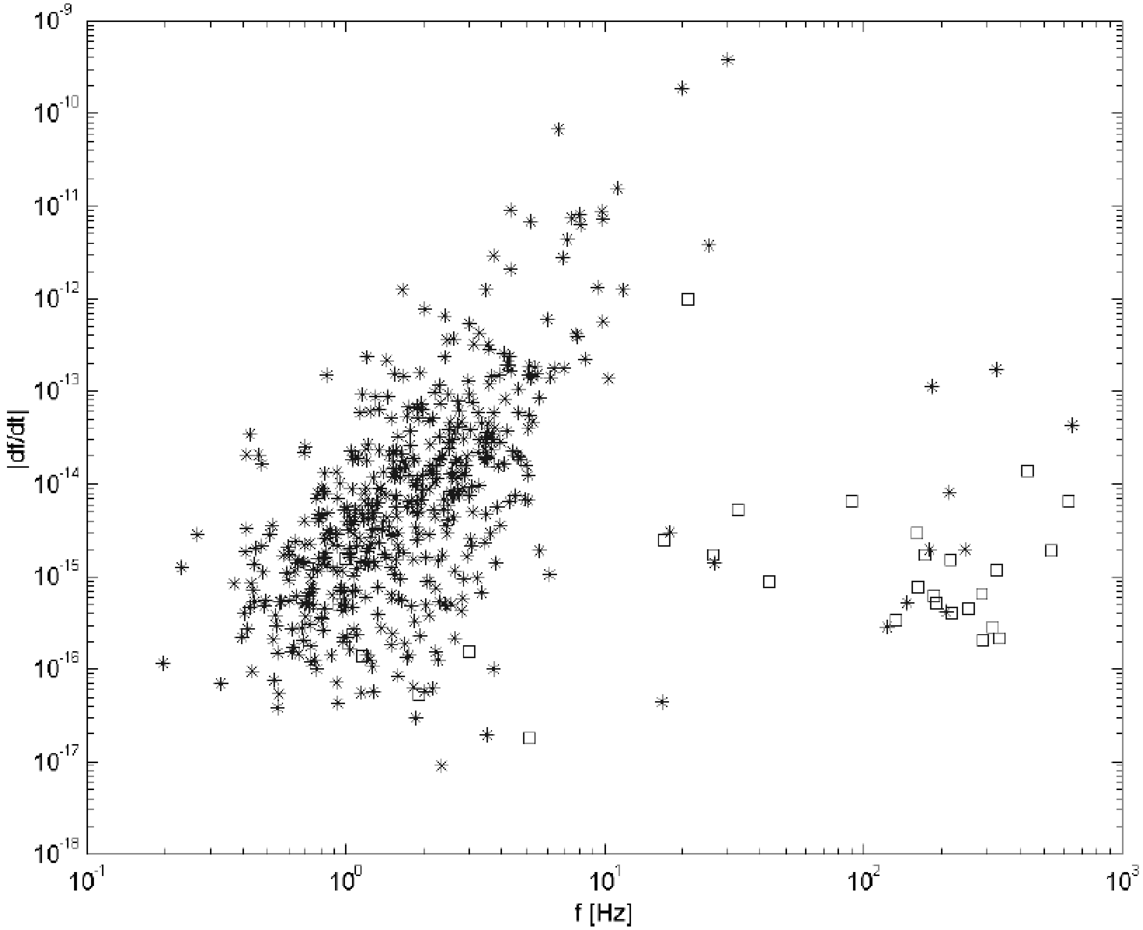
where, f is the rotation frequency and $I \sim 10^{38}\ kg\ m^{2}$ the momentum of inertia of the star

$$h_o \sim 1\ 10^{-27} (I\ /10^{38}\ \mathrm{kg\ m}^{2})\ (f\ /100\ \mathrm{Hz})^{2}\ (10\ \mathrm{kpc} / \delta)\ (\varepsilon_{max} / 10^{-6})$$

LESSON: search at lower distances (full sky ~ 4 π solid angle) and ii

the high frequency region

$$h \sim 10^{-26}$$

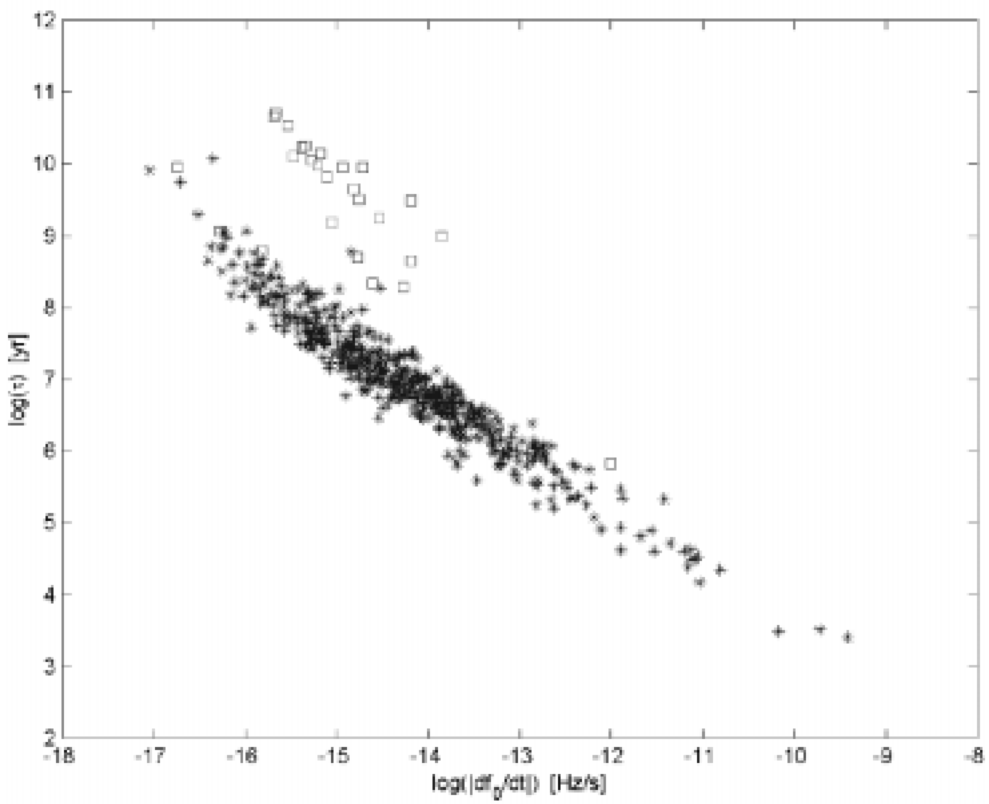


10^-9
10^-10
10^-11
10^-12
10^-13
10^-14
10^-15
10^-16
10^-17
10^-18
|df/dt|
10^-1
10^0
10^1
10^2
10^3
f [Hz]

[df/dt] vs. f

$\tau = f / (df/dt)$

Spin-down age



τ vs. [df/dt]

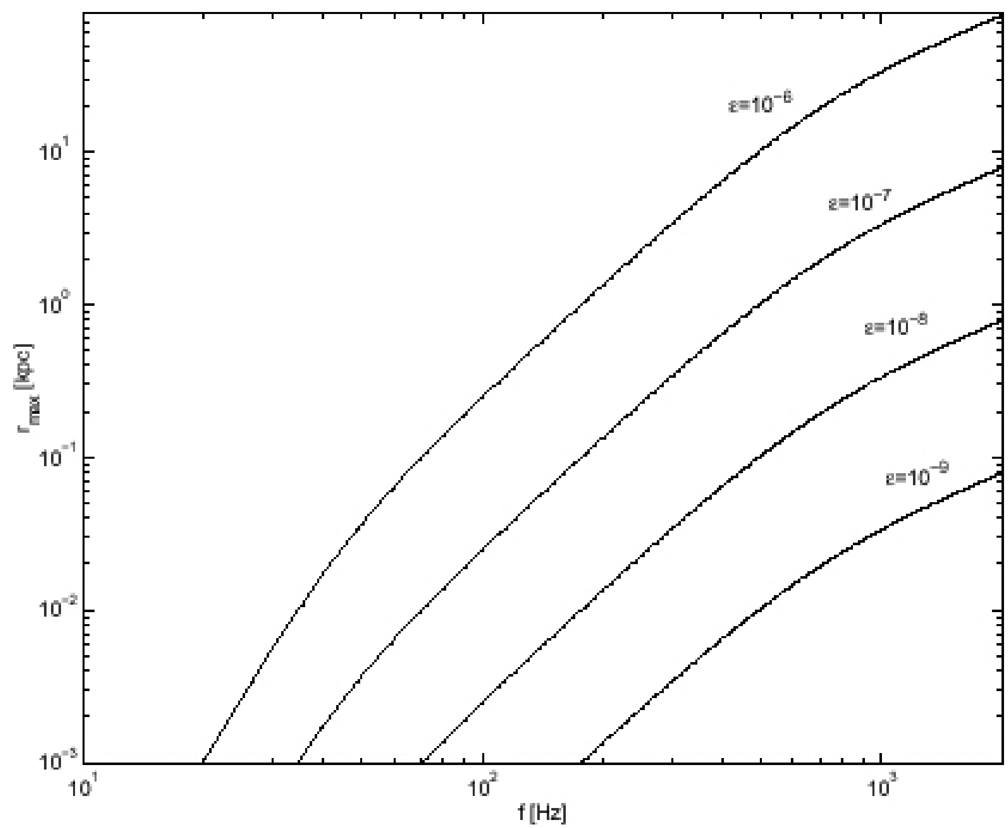


Figure 3: Maximum distance within which a pulsar could be detected with $SNR = 1$, as a function of the gravitational wave frequency, for different values of the ellipticity ϵ.

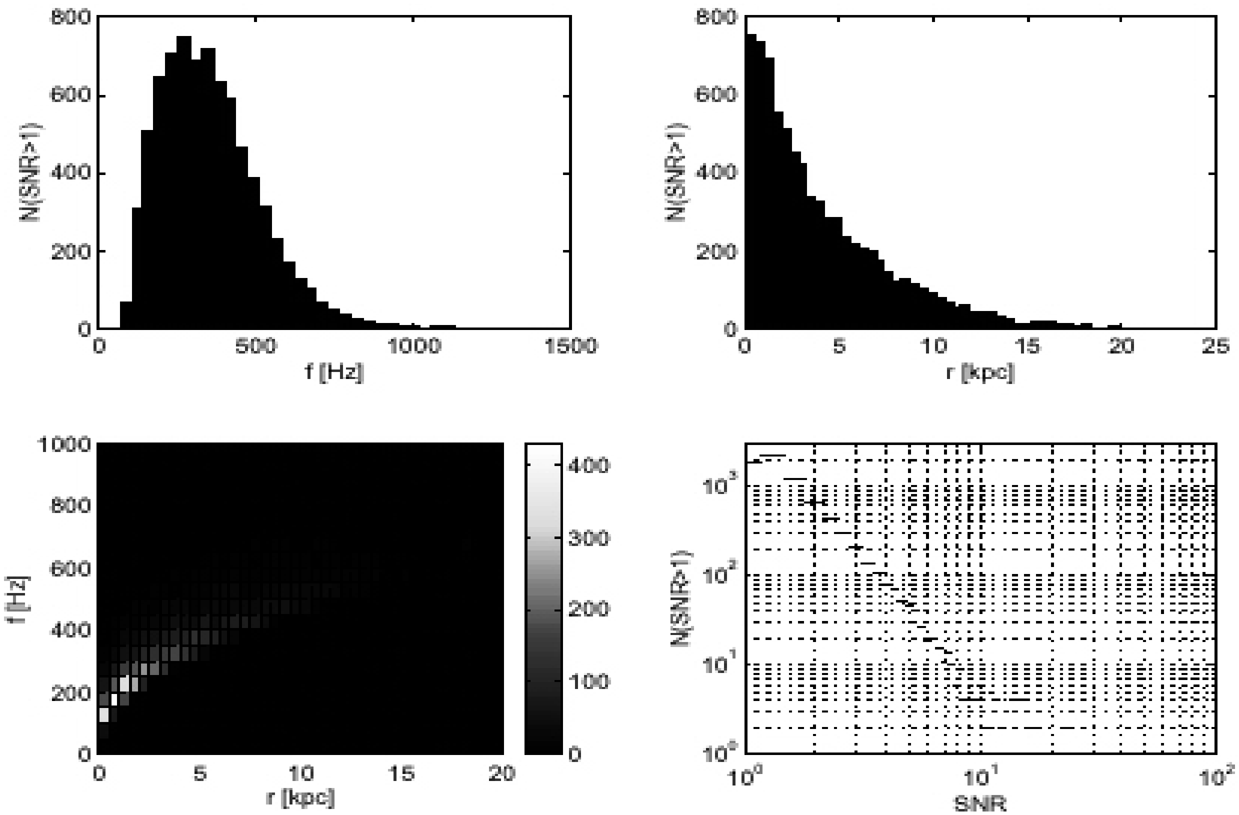


Figure 5: Simulated gw-driven population with $N = 2 \cdot 10^6$. Distribution of detectable population parameters: gravitational frequency, distance, SNR. Ellipticity $\epsilon = 10^{-6}$. Standard case file $standard - GW - 6 - 2M.eps$.

SN1987A. It appears clear that such kind of sources could be very promising but the final answer will depend on the total number of objects existing in the Galaxy.

…Neutron Stars

Basically the analysis is developed by using the coordinates of the SOLAR SYSTEM REFERENCE frame ("SSB frame")
Hypothesis about the signal wave form: the signal is <u>quasi</u> monochromatic because of

-the slowing down of the source ("spin down" parameters)

-the relative motion of the source and the detector (the Doppler correction depends on the position of the source in the sky)

The time evolution of the intrinsic frequency of the signal is

$$f(t) = f_o \left(1 + \{\mathbf{v} \cdot \mathbf{n}\}/c \right) \left(1 + \Sigma f_k \, t^k \right)$$

$\mathbf{v}$ detector velocity , $\mathbf{n}$ unity vector of the detector - source direction, f_k spin down parameters, f_o source intrinsic frequency, $f(t)$ observed frequency

$\mathbf{v}(t) = \mathbf{v}_{rev}(t) + \mathbf{v}_{rot}(t)$ $\quad\quad$ $< v_{rev} > \sim 32$ km/s $\sim 100 < v_{rev} >$

<u>Maximum Doppler shift</u> $\quad\quad$ $\Delta f_{max} \sim 10^{-4} f_o$

The full sky signal search based on <u>the optimum linear strategy</u>

- The expected signal on the Earth is quasi monochromatic.
In order to reduce the search problem to that of a pure monochromatic signal, we have to impose to the data the Doppler correction. Thus, we need divide the sky in patches and for each frequency we apply the Doppler correction.

- Then we have just to perform FFT on strecths of data as long as possible (1/ T_{obs}). This is equivalent in the time domain to apply to the data an time integration equal to the observation time T_{obs}.

In this way we have for the SNR

$$\boldsymbol{SNR} \sim 1\ 10^{-21} (I/10^{38}\ kg\ m^2)\ f^2 \varepsilon\ (1\ kpc/\delta)\ (\boldsymbol{T_{obs}}\ /1yr)^{1/2} (1/S_h)^{1/2}$$

$$\varepsilon \sim 10^{-6}\ ;\quad S_h^{1/2} \sim 10^{-23}\ [1/Hz]\ ;\quad f \sim 500\ Hz\ :\quad SNR \sim 25$$

We need to divide the sky in patches . How many of them ?
This depends on the angular resolution of the detector $\Delta\Theta$ that is a function of the observed frequency and it improves with the observation time T_{obs}.

$$N_{patch} = 1.3\ 10^{13} (f / 1\ kHz)^{2} (T_{obs} / 10^{7})^{4}$$

For a given frequency f and bandwidth Δf the dimension of the data set to be considered is is immediately derived in function of T_{obs}

<u>We assume that the computation time is equal to T_{obs} ,</u> and we derive the number of floating point operations N_{flop} to be performed for the linear optimum strategy and the needed computing power P_c.

$$N_{flop} = 2\ 10^{10} (f / 1\ kHz) (\mathbf{T_{obs}} / 10^{10})$$

$$P_c\ [Mflop] = 5\ 10^{20} (f / 1\ kHz)^{2} (\mathbf{T_{obs}})^{4}$$

$\mathbf{T_{obs}}$ = *72 hours ==>> 200 Mflop*

$\mathbf{T_{obs}}$ = *1 year ==>> 5 10^{10} Tflop*

Conclusion: we must abandon the linear optimum strategy

	1 month	4 months	1 year
FFT length	2.6E+09	1.0E+10	3.1E+10
Sky points	2.1E+11	3.4E+12	3.1E+13
Spin-down points (1st)	2.2E+04	3.5E+05	3.1E+06
Spin-down points (2nd)	1.0E+00	1.1E+01	3.2E+02
Freq. points (500 Hz)	1.3E+09	5.0E+09	1.6E+10
Total points	6.0E+24	6.5E+28	4.8E+32
Comp. power (Tflops)	1.0E+12	1.5E+16	3.6E+19
Sensitivity (nominal)	1.2E-26	6.0E-27	3.6E-27
(background 10^-23*Hz^-0.5)			
Sens. for 10^9 candidates	7.4E-26	4.1E-26	2.7E-26

The sensitivity limit of the detector for this kind of signal is determined by the available computing power P_c

We have to introduce a sub-optimal method <u>with limited SNR loss</u> designed for realistic values of

P_c ~ 1 - 10 Tflop

Hierarchical search

The idea is to alternate incoherent search step to coherent ones.

Incoherent search step: no phase information is used

Coherent search step: in practice it concerns frequency zooming and Complex FFT

The starting point is the filling of a specific Data Base where short FFTs of the VIRGO data are stored (low frequency resolution)

Incoherent Step: the main porpuose is to select a large number of possible source candidates that we validate by the following coherent step

At present there are two possible approaches for the first step

1) STACKING SLIDE technique
We divide in small fequency interval (slide) the periodograms acquired in time sequence.
The corresponding slides are summed incoherently and the candidate peaks are selected by choosing a suitable threshold.

Advantage**: easier implementation on a parallel computing architecture**

Disadvantage**: weak method to the respect of non-stationary noise behavior of the detector because the selection criteria depends on the amplitude of the signal into the single bin of the periodogram.**

2) HOUGH TRANSFORM

The method has been proposed by the Virgo Rome group in order to have realy in a more robust algorithm.

The length of the single short FFT is chosen to have a bin larger than the maximum possible Doppler shift. However, looking at the various FFT taken in time sequence the signal peak moves out of the original bin because of the Doppler effect.

Thus, we look how the frequency peaks move in function of the time that define the FFT sequence. If it is a real signal it has to follow the deterministic Doppler law

$$f(t) = f_o \,(1+ \{\mathbf{v} \cdot \mathbf{n}\} /c) \,(1 + \Sigma \, f_k \; t^{\,k})$$

The formula tell us that for each observed frequency $f(t_i)$ ***the locus of the points of the celestial sphere where we can locate a source generating a g.w. signal is a circle. If the signal is still there, at the time*** t_j ***the new circle related to*** $f(t_j)$ ***will intersect the old one in the source location of the sky.***

Advantage: **robust algorithm and high computational speed (most part of the computation is performed with integers)**

Disadvantage: **it requires a large software development and accurate tuning**

Number of points in the parameter space

Number of frequency bins $N_{\nu} = T_{FFT} / 2\ \Delta T$

Freq. bins in the Doppler band $N_{DB} = N_n\ 10^{-4}$

Sky points $N_{sky} = 4\ p\ N_n^2$

Spin-down points $N_{SD}^{(j)} = 2\ N_{\nu}\ (T_{FFT} / \tau_{\mu in})^{(j)}$

Total number of points $N_{tot} = N_{sky}\ \Pi_j\ N_{SD}^{(j)}$

Minimum detectable h (*CR= 1*)

Optimal detection nominal sensitivity

$$h^{(OD)} = (4\, S_h / T_{obs})^{1/2}$$

Hierarchical method nominal sensitivity

$$h^{(HM)} = h^{(OD)} (T_{obs} / T_{FFT})^{1/4}$$

¥ divide the data in (interlaced) chunks; the length is such that the signal remains inside one frequency bin

• do the FFT of the chunks; this is the SFDB

• do the first “incoherent step” (Hough or Radon transform) and take candidates to follow

• do the first “coherent step”, following up candidates with longer “corrected” FFTs, obtaining a refined SFDB (on the fly)

• repeat the preceding two step, until we arrive at the full resolution

The SFDB is a collection of (interlaced) FFTs, with such a length that the varying frequency signal power would be collected all in a single bin.

The FFTs will be windowed.

We plan to do 4 different SFDBs, with different sampling times, in order to optimise the detection in different bands.

- using the SFDB, create the periodograms and then a time-frequency map

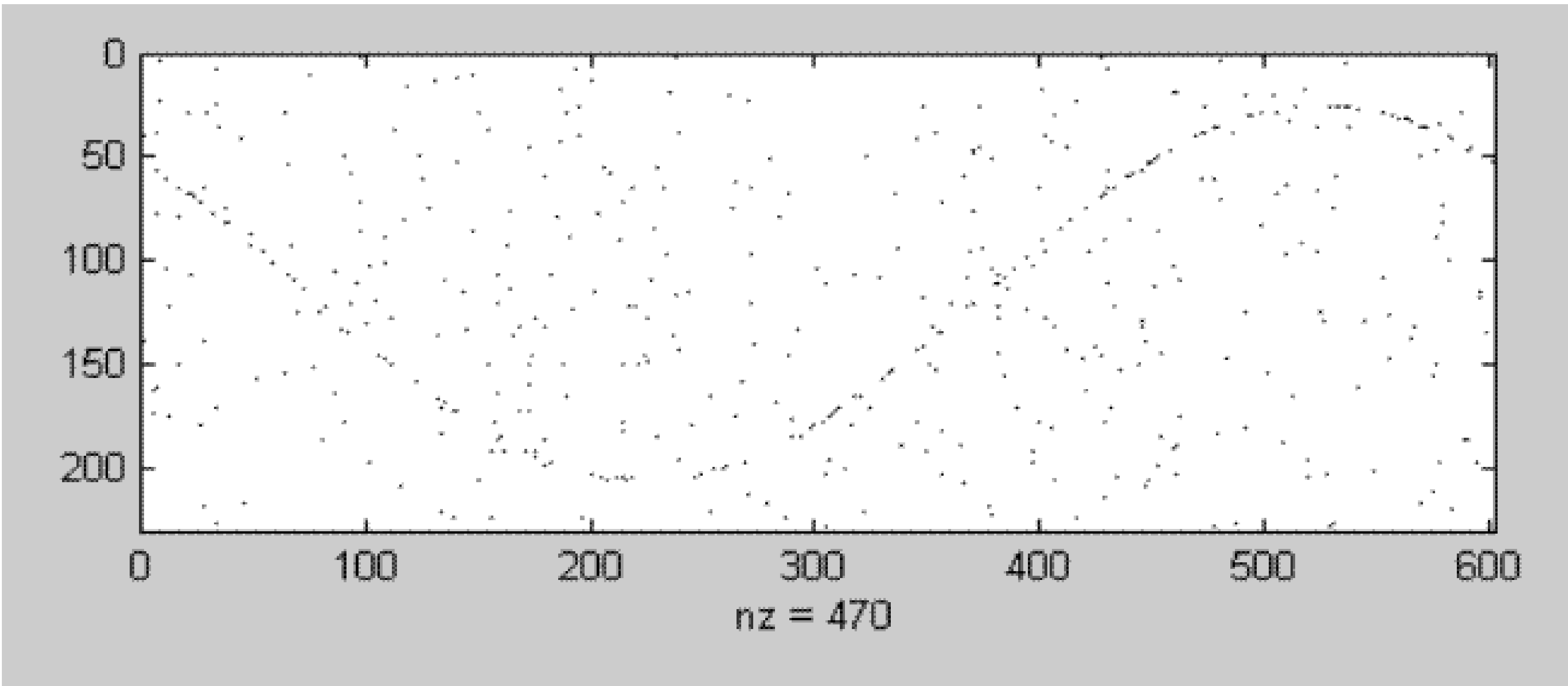
- for each point in the parameter space, shift and add the periodograms, in order to all the bins with the signal are added together

Incoherent steps: Hough transform

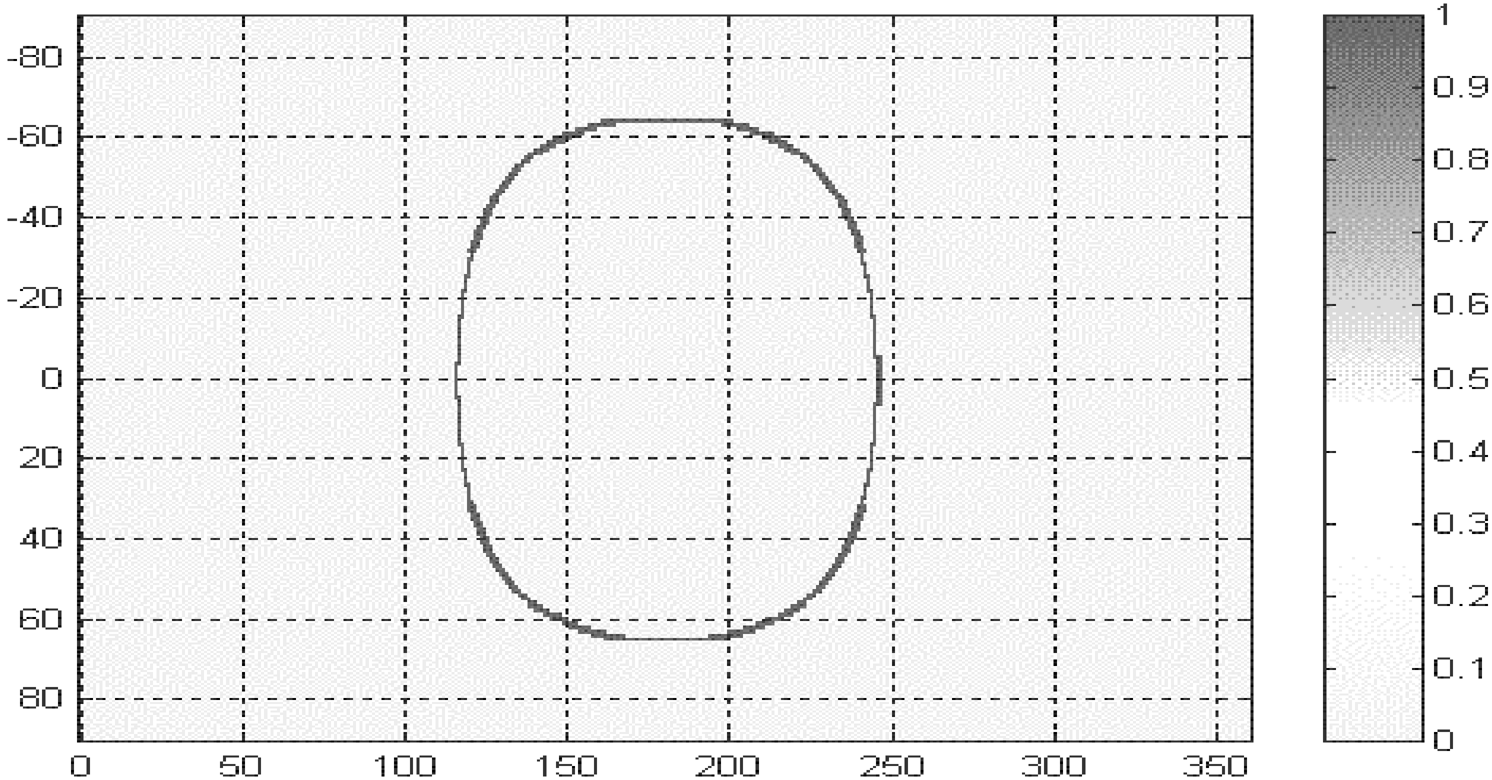
- using the SFDB, create the periodograms and then a time-frequency map of the peaks above a threshold

- for each spin-down parameters point and each frequency value, create a sky map (“Hough map”); to create a H map, sum an annulus of “1” for each peak; an histogram is then created, that must have a prominent peak at the “source”

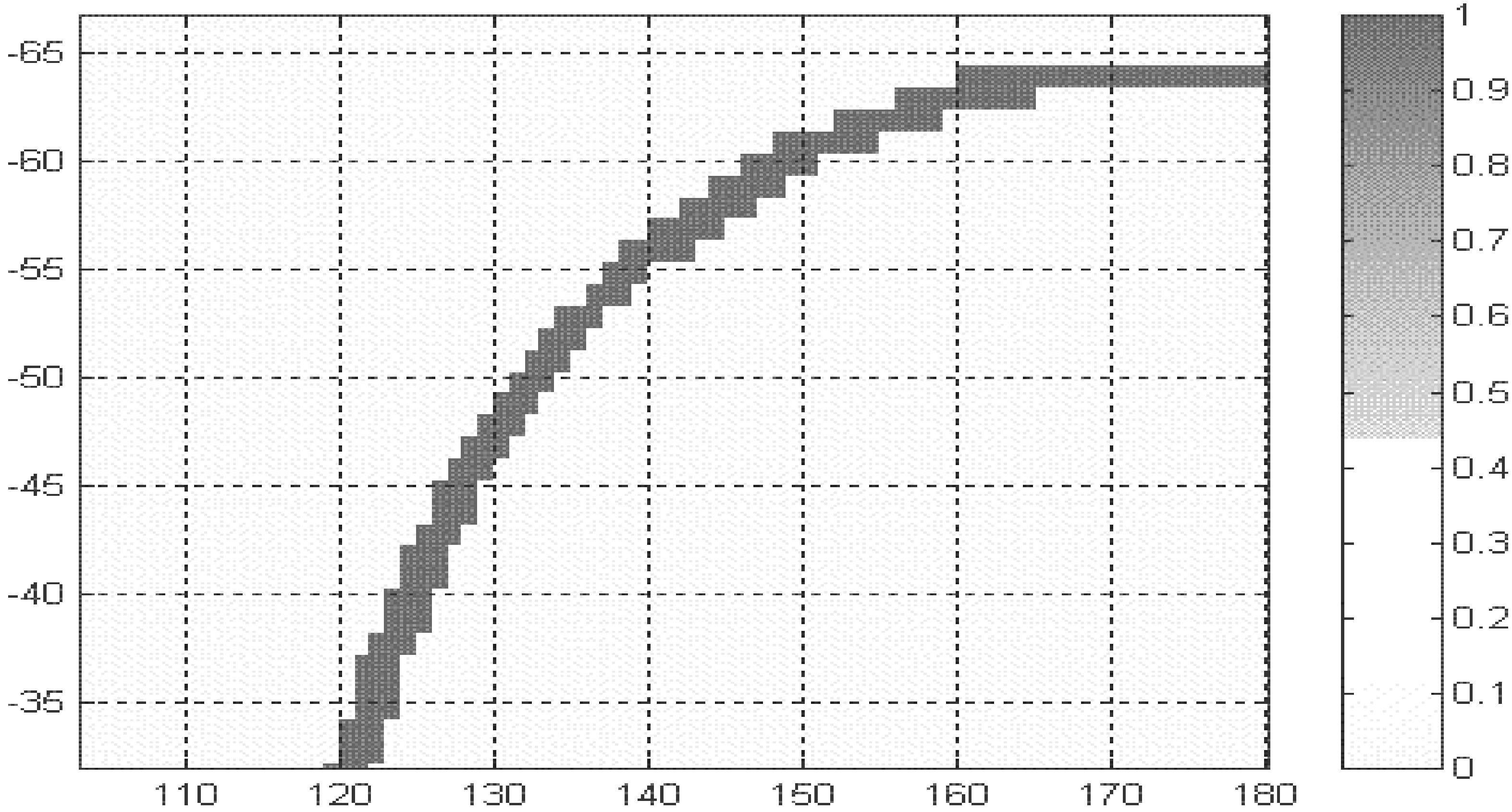
Time-frequency peak map



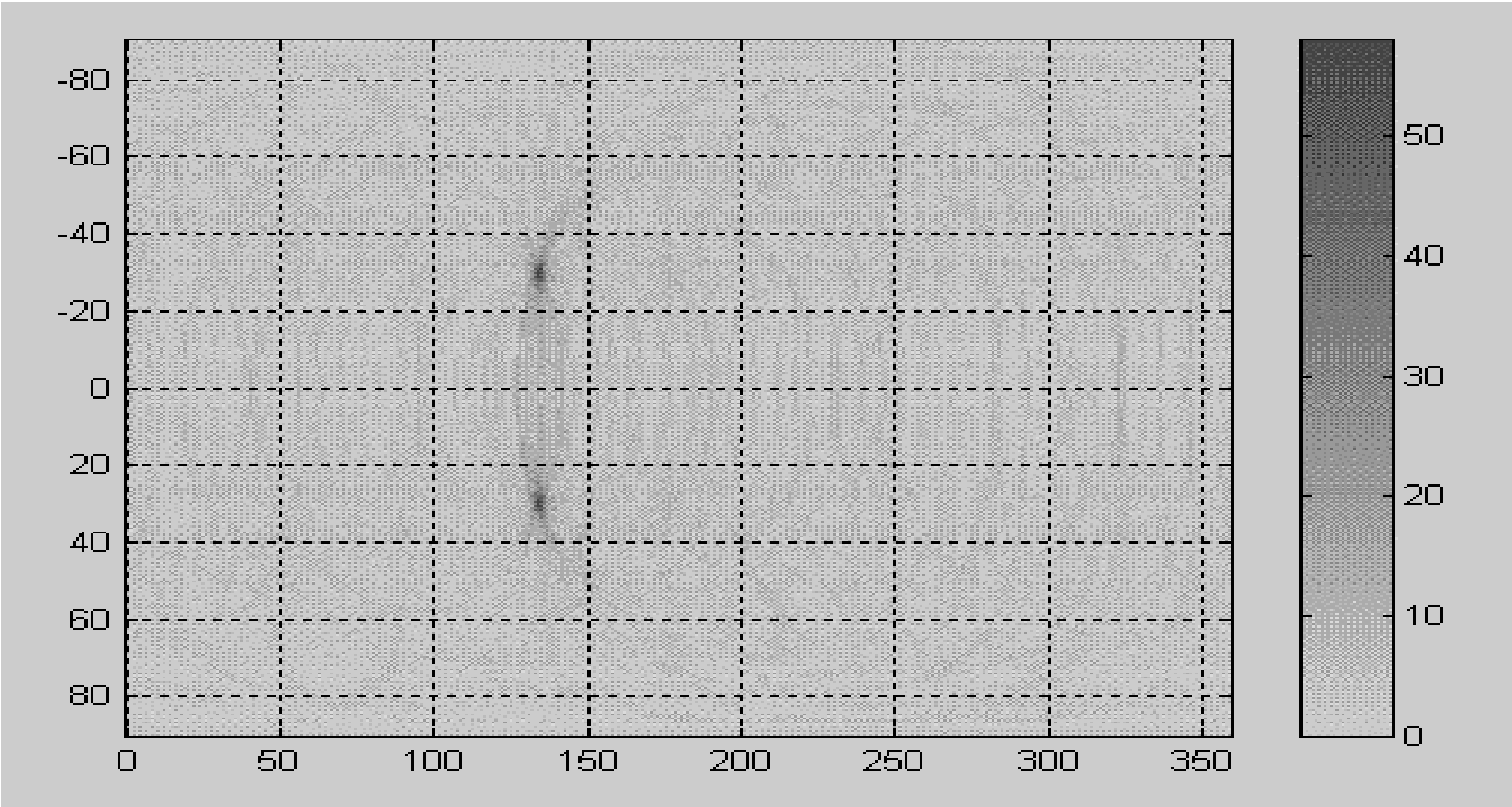
Hough map – single annulus



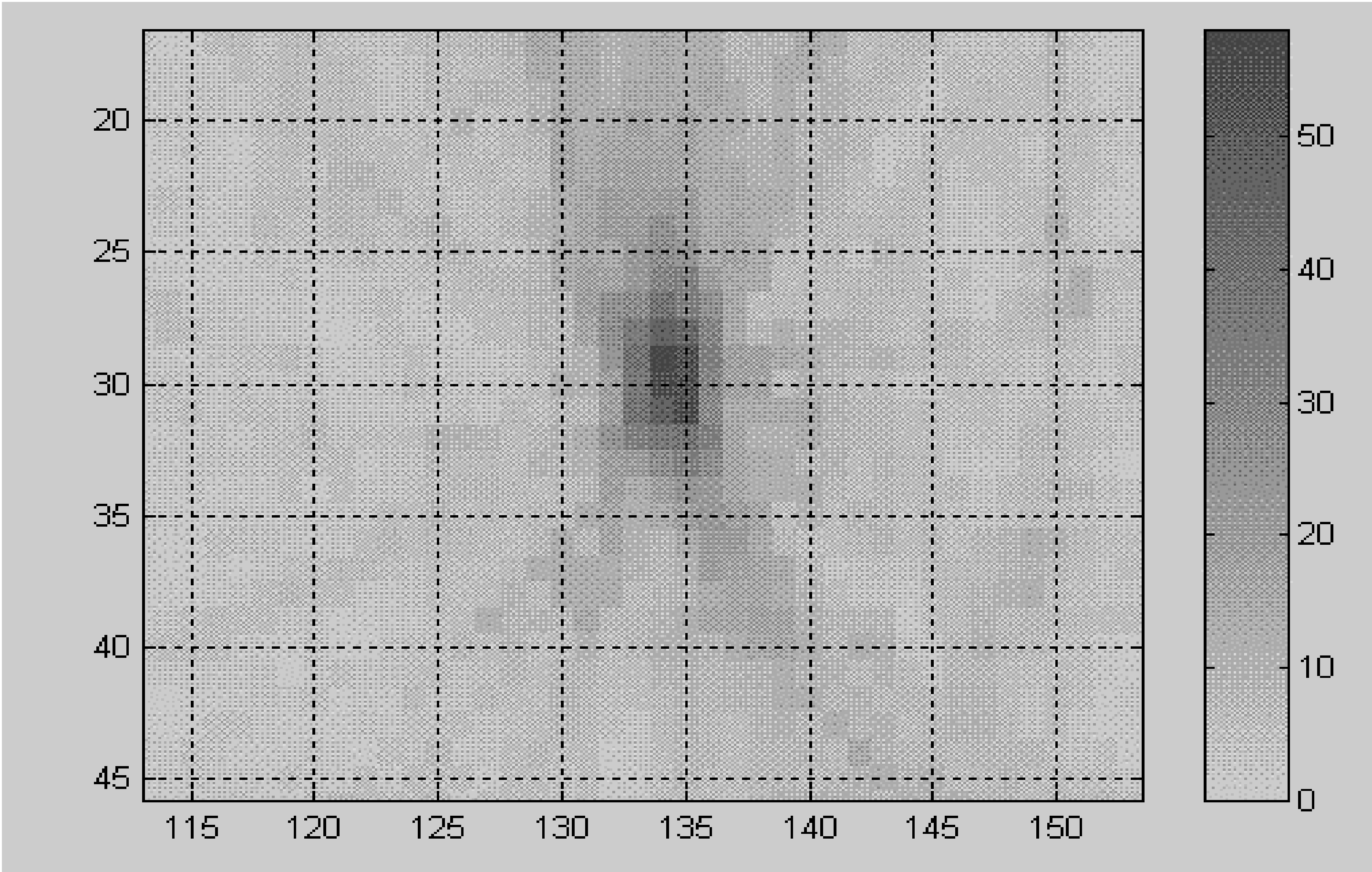
Hough map – single annulus (detail)



Hough map – source reconstruction



Source reconstruction - detail



What we gain with Hough ?

• about 10 times less in computing power
• robustness respect to non-stationarity
• operation with 2-bytes integers

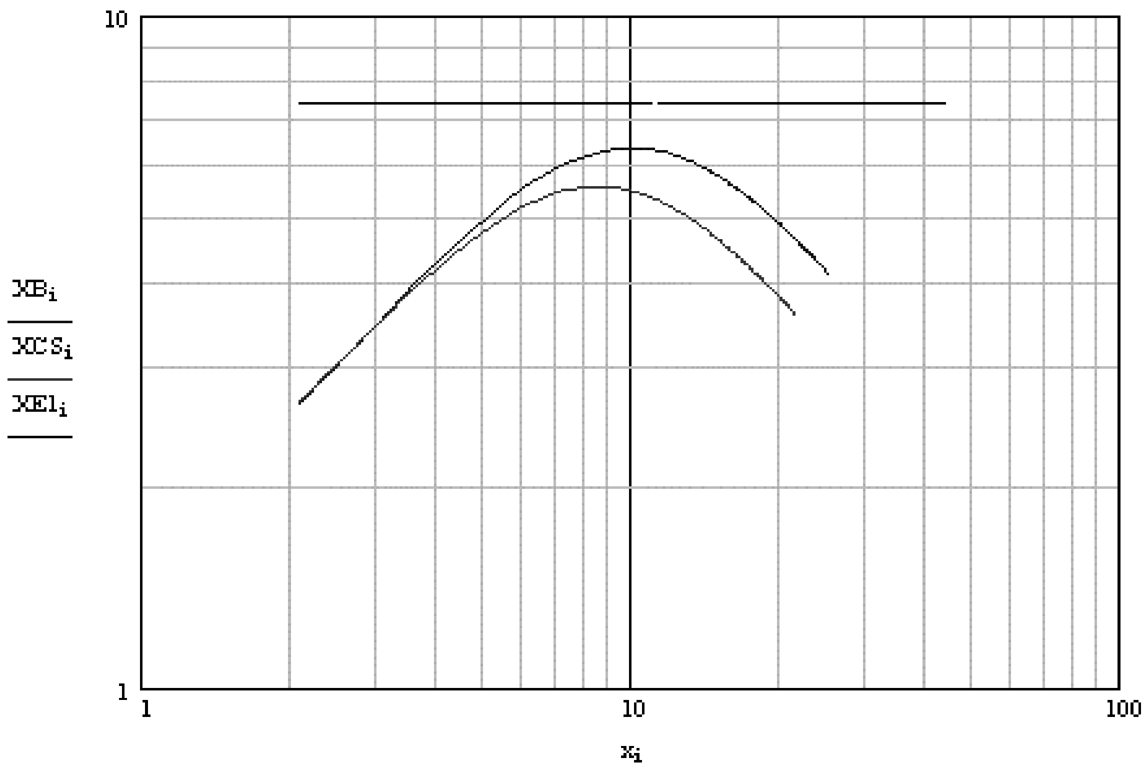
What we lose ?

• about 12 % in sensitivity (can be cured)
• more complicate analysis

With the coherent step we partially correct the frequency shift due to the Doppler effect and to the spin-down. Then we can do longer FFTs, and so we can have a more refined time-frequency map.

This steps is done only on “candidate sources”, survived to the preceding incoherent step.

Loss respect to the optimum



The fundamental points are:

• the sensitivity is proportional to $T_{FFT}^{1/4}$

• the computing power for the incoherent step is proportional to T_{FFT}^{3}

• the computing power for the coherent step is proportional to $log\ T_{FFT}$, but it is also proportional to the number of candidates that we let to survive.

The result of an analysis is a list of candidates (for example, 10^6 candidates).

Each candidate has a set of parameters:

- the frequency at a certain epoch
- the position in the sky
- 2~3 spin-down parameters

	Banda 1	Banda 2	Banda 3	Banda 4
Frequenza Max della banda (Hz)	2500	625	156.25	39.0625
Banda Doppler (Hz)	0.25	0.0625	0.0156	0.0039
Numero di Picchi nella banda Doppler	1643684	410921	102730	25683
Risoluzione angolare nel cielo (rad)	1.9073E-03	3.8147E-03	7.6294E-03	1.5259E-02
Numero di pixels nel cielo	3.4542E+06	8.6355E+05	2.1589E+05	5.3972E+04
Numero di frequenze indipendenti	3.1457E+06	1.5729E+06	7.8643E+05	3.9322E+05
Prametri di Spin down (solo l'ordine 1)	839	419	210	105
Numero Totale di parametri	2.8976E+09	3.6220E+08	4.5275E+07	5.6594E+06
Numero di operazioni per ogni picco	6.5884E+03	3.2942E+03	1.6471E+03	8.2355E+02
Numero totale di operazioni	2.8577E+19	8.9302E+17	2.7907E+16	8.7209E+14
Potenza di calcolo per passo (GFlops)	1.8109E+03	5.6592E+01	1.7685E+00	5.5265E-02
Potenza di calcolo complessiva (GFlops)	**3.6219E+03**	**1.1318E+02**	**3.5370E+00**	**1.1053E-01**

The detection of stochastic background of Gravitational Waves is an other ambitious goal of the gravitational wave experiments.

Two main sources have to be considered :
a) Stochastic background of Cosnmological Origin
Relic G.W. carry information on the State of the Universe at the Planck era when gravitons decoupled from the primordial plasma *(Very High Energy Physics).*

b) Stochastic background of AstrophysicalOrigin generated by the Supernovae explosions at high red shift (Z~ 4)

The detection strategy is basically a cross-correlation of the output of two antennas at least.

Stochastic background

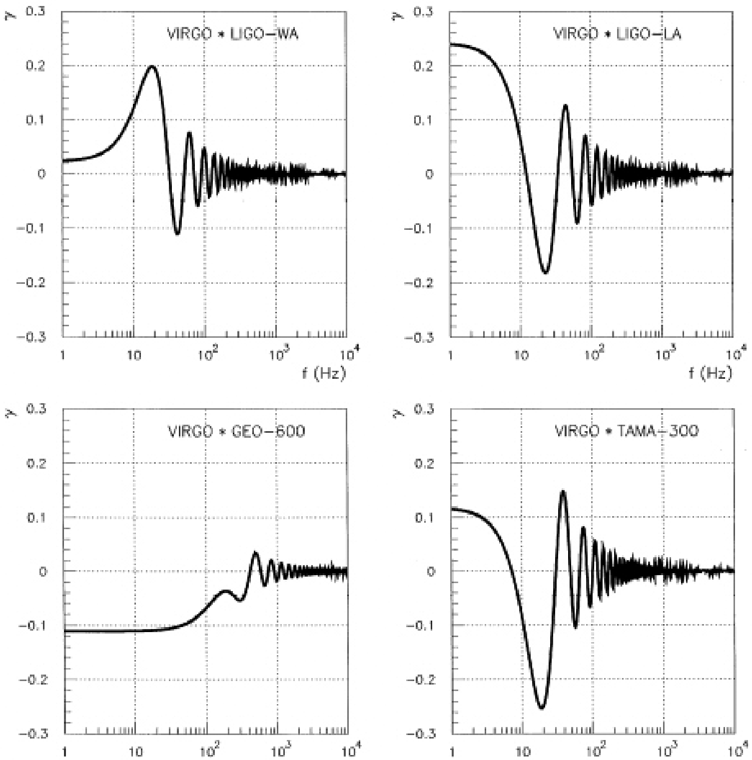


Fig. 9. The overlap reduction functions $\gamma(f)$ for the correlation of VIRGO with the other major interferometers (from Ref. [27]).

Table 1
The sensitivity of various two-interferometers correlation

Correlation	$h_0^2\Omega_{gw}$
LIGO–WA * LIGO–LA	5×10^{-6}
VIRGO * LIGO–LA	4×10^{-6}
VIRGO * LIGO–WA	5×10^{-6}
VIRGO * GEO600	5.6×10^{-6}
VIRGO * TAMA300	1×10^{-4}

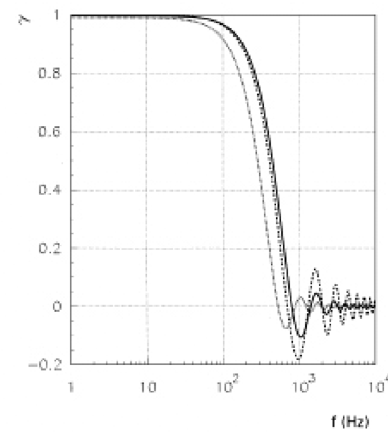


Fig. 10. The overlap reduction function $\gamma(f)$ for the correlation of VIRGO with NAUTILUS (solid line), AURIGA (dashed line) and EXPLORER (dotted line) in the case in which the bars have been reoriented so to be aligned with VIRGO (from Ref. [27]).

Table 2
The sensitivity of the correlation of VIRGO with resonant bars, and of two resonant bars (from Ref. [27]), but setting the confidence level to 90%)

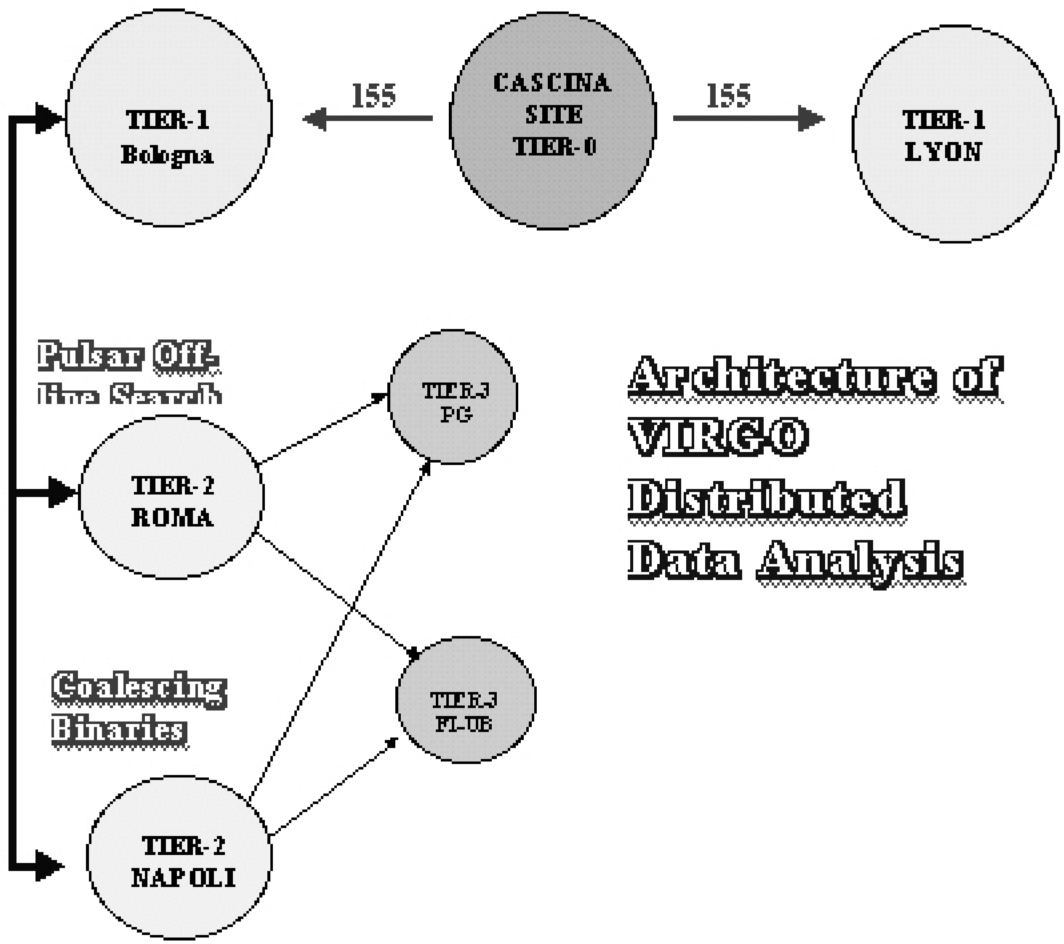
Correlation	$h_0^2\Omega_{gw}$
VIRGO * AURIGA	4×10^{-4}
VIRGO * NAUTILUS	7×10^{-4}
AURIGA * NAUTILUS	5×10^{-4}

Quadro complessivo delle risorse di calcolo off-line di Virgo – Italia

units		end 2001	end 2003
CPU capacity*		80 kSI2000 (350 Gflops)	800kSI2000(3.5Tflops)
disk capacity	TBytes	10	100
disk I/O rate	GBytes/sec	5	5

Quadro di sviluppo del Tier 1 di Virgo Italia

units		end 2001	end 2003
CPU capacity*		46 kSI2000 (200 Gflops)	300 kSI2000 (1.3Tflops)
disk capacity	TBytes	8	100
disk I/O rate	GBytes/sec	5	5


TIER-1
Bologna
155
CASCINA
SITE
TIER-0
155
TIER-1
LYON
Pulsar Off-
line Search
TIER-3
PG
Architecture of
VIRGO
Distributed
Data Analysis
TIER-2
ROMA
Coalescing
Binaries
TIER-3
FI-UB
TIER-2
NAPOLI