



UvA-DARE (Digital Academic Repository)

Transient and variable radio sources in the LOFAR sky: an architecture for a detection framework

Scheers, L.H.A.

Publication date
2011

[Link to publication](#)

Citation for published version (APA):

Scheers, L. H. A. (2011). *Transient and variable radio sources in the LOFAR sky: an architecture for a detection framework*. [Thesis, fully internal, Universiteit van Amsterdam].

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

Expected Data Rates and Volumes for the LOFAR Transients Key Project

2.1 Introduction

The International LOFAR Telescope is in many aspects a next-generation radio telescope. LOFAR Stations, consisting of dual-dipole antennas, are spread over the northern part of the Netherlands and into the European countries of Germany, France and the United Kingdom. The distribution of the LOFAR Stations allows for scalable baselines. Baselines of up to 3 km are from the core stations alone, up to 100 km when the remote (i.e. Dutch) stations join, and up to 1000 km when the European Stations are included. The large collecting area and high resolution, the large fields of view of the dipole antennas, and the high sensitivities due to the numerous stations and large frequency bandwidth, give unprecedented observation capabilities in the yet unexplored low-frequency domain of the electromagnetic spectrum below 240 MHz. Beam-formed data of a LOFAR station are transported over a 10 Gb/s ethernet link to an IBM BlueGene/P central processing supercomputer that correlates the signals from all the stations, producing raw data at 1 s time resolution and 1 kHz frequency resolution. Calibration and imaging pipelines run on a dedicated post-processing cluster, aiming at producing calibrated snapshots of large patches of the sky every second. At this point the Transients Key Science Project plugs in its automated software pipeline in the continuous data stream of calibrated images in order to detect transient and variable radio sources on various time scales for scientific analysis.

The number of sources likely to be detected and disk size to be allocated

for a given LOFAR observation mode, depend on the sensitivity, observing frequency and field of view of the observations, which in turn depend on the key configuration parameters of LOFAR: the frequency, bandwidth, resolution, integration time, number of beams, number of stations.

To enable searches for transients and wide-field or even full-sky monitoring modes, i.e., to see changes from one observation to another, we want to record all measurements of the observable sources. Comparison with known sources from the major catalogues at multiple frequencies, and in parallel with the running, gradually growing and updated LOFAR catalogue, enables us to classify sources based on their light-curve data at an early stage. The image size and input rate – the number of sources to be processed per unit time – then determine whether modes can reach the desired shortest time scales.

This Chapter focuses on the number of sources and measurements extracted from those images per second, and not on the number of floating point operations per second nor the raw image sizes that are produced by the calibration and imaging pipelines. Nor will we discuss the disk spaces needed for storing the (raw) image data. Instead, we will concentrate on the estimated data growth and disk space needed to build up the large catalogue database containing all LOFAR sources.

This Chapter is organised as follows. Section 2.2 briefly describes the design of the LOFAR Telescope and some of its characteristics. In Section 2.3 the number of sources likely to be detected is estimated for some frequently used observation modes of LOFAR by the Transients Key Project. The details down to the shortest time scales for the modes are reported in Appendix 2.A. Based on the number of measurements made per source, the size and growth of disk space to be allocated is determined. Section 2.4 describes the database schema and the design and performance of the implemented algorithms to enable fast transient and variability detection of which the main benchmark queries are shown in Appendix 2.B. Section 2.5 discusses the results.

2.2 The Design of the LOFAR Telescope

Field of view, resolution and sensitivity are the main players in the ongoing race of building more powerful telescopes. While the very expensive classical radio telescopes with steerable dishes have reached their construction limits, cheaper (but not less challenging) alternatives have come into view. During the 1990s, the concept of a software telescope developed and was getting more concrete by the end of that decade (Miley, 2010).

LOFAR's frequency range is divided into two bands bracketing the FM radio broadcasting band. Each band has its own antenna type, the Low

| LOFAR Station | L [km] | N_{station} | N_{LBA} | $N_{\text{HBA Tiles}}$ |
|---------------|----------|----------------------|------------------|------------------------|
| Core | 3 | 24 | 96 (48) | 2×24 |
| Remote | 100 | 16 | 96 (48) | 1×48 |
| European | 1000 | 8 | 96 | 1×96 |

Table 2.1: Three types of LOFAR stations exist. Core and remote stations are located in the Netherlands, whereas the European Stations are in Germany, France, Sweden and England. For each type the number of stations, N_{station} , the maximum baseline, L , the number of Low Band Antennas, N_{LBA} and the number of High Band Tiles, $N_{\text{HBA Tiles}}$, is given, where a single HBA Tile contains 16 High Band Antennas, arranged on a 4×4 grid. The number in parentheses in the column of LBAs represents the maximum number of available LBA antennas for dual polarisation observations.

Band Antennas (LBA) being sensitive from 30 to 80 MHz, and the High Band Antennas (HBA) in the range 120–240 MHz (see Fig. 1.1 in Chapter 1). LBAs and HBAs are most sensitive at 60 and 150 MHz, respectively (e.g. de Vos et al., 2009; Nijboer & Pandey-Pommier, 2009).

Dutch LOFAR stations are located either in the core at Exloo, where multiple stations are densely packed in an area of 3 km, and more sparsely remote, in the north-eastern Dutch provinces, with baselines up to about 100 km. European or international stations are sited in Germany, France, Sweden and the United Kingdom, giving baselines of 1000 km. The number of available LBA antennas and HBA tiles differ depending on the location of the station. All Dutch stations have 96 Low Band Antennas, but when observing in dual polarisation mode only 48 may be selected, whereas in single polarisation mode all 96 may be used. The EU Stations also have 96 Low Band Antennas, but do not have such a restriction. Furthermore, the core stations have two fields (also called "ears") of 24 HBA tiles, whereas a remote station has a single field of 48 tiles, and an EU Station a single field of 96 tiles. Table 2.1 gives the current status of the number of funded and planned LOFAR stations.

The LOFAR Telescope is configurable in many ways (we refer to de Vos et al. (2009) and Nijboer & Pandey-Pommier (2009) for more details), but here we will highlight the most characteristic properties and observation modes relevant to the Transients Key Project. Multiple beams may be constructed with the restriction that the total bandwidth does not exceed the maximum value of 48 MHz. Although LOFAR is able to observe at any frequency in the Low as well as in the High Band, simultaneously observing in both Bands is not possible. However, we will focus here on five frequencies per Band, evenly distributed across the Low and High Band, in accord with the findings of Law & Hessels (2009) of Standard Bands for the Radio Sky Monitor Mode.

Five different LBA antenna configurations are available of which we will

focus on the Inner configuration. This mode uses the innermost 48 antennas, thereby reducing the station size, and increasing the full width half maximum (FWHM) of the station beam

$$\theta_{\text{FWHM}} = k_1 \frac{\lambda}{D}, \quad (2.1)$$

where k_1 is of order unity, λ the observing wavelength and D the station diameter, or the distance between the two outermost antennas. And this consequently affects the station field of view (FoV) defined as

$$\Omega_{\text{FoV}} = \pi \left(\frac{1}{2} \theta_{\text{FWHM}} \right)^2. \quad (2.2)$$

The resolution of the LOFAR mode is then determined by

$$\theta_{\text{res}} = k_2 \frac{\lambda}{L}, \quad (2.3)$$

where k_2 is of order unity, λ the observing wavelength and L the distance between the two outermost stations. The sensitivity is then defined by

$$\Delta S = W \left[2(2\Delta\nu\tau) \left(\frac{N_C(N_C - 1)/2}{S_C^2} + \frac{N_C N_R}{S_C S_R} + \frac{N_R(N_R - 1)/2}{S_R^2} \right) \right]^{-1/2}, \quad (2.4)$$

where $\Delta\nu$ is the spectral bandwidth (in Hz), τ is the integration time (in seconds), W depends on the imaging weighting scheme used and is unity for core stations only and 1.3 for inclusion of remote stations, N_C and N_R are the number of core and remote stations, respectively, and S_C and S_R are the measured system equivalent flux densities (SEFDs in Jy; e.g., de Vos et al., 2009; Nijboer & Pandey-Pommier, 2009) for core and remote stations, respectively, which is an indication of the sensitivity of the antenna and receiving system (e.g., Thompson, Moran & Swenson, 2004; Wrobel & Walker, 1999). We adopted the effective bandwidth factor, $\eta = 0.89$, based on results from LOFAR test stations (Nijboer & Pandey-Pommier, 2009). If identical core stations are being used, as is the case in some observation modes, the sensitivity is reduced to ($W = 1$)

$$\Delta S = \frac{S_C}{\sqrt{2\Delta\nu\tau N_C(N_C - 1)}}. \quad (2.5)$$

From Eqs. 2.4 and 2.5 it can be seen that the sensitivity is inversely proportional to the square root of the bandwidth and the integration time.

Selecting a set of integration times that increases logarithmically enables detection of transient and variable sources on different time scales and flux densities. The Transients Key Project will use thirteen time scales: 1, 2, 5, 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, and 10000 seconds for this. Due to information on the broad frequency range the light-curve catalogue database contains the temporal as well as the spectral characteristics of all sources detected by LOFAR.

2.3 Expected Data Rates & Volumes for some TKP Observation Modes

The Transients Key Project designed and developed an automated software pipeline for detecting transient and variable sources in the input data streams on time scales as short as 1 second. It is clear that the millions of sources and their repeatedly measured properties on different time scales, cannot be simply stored in tuples or arrays of even files. Whether open source databases are capable of processing these large data volumes on those short time scales has to be investigated. Furthermore, we have to find out if we will gain significant processing time by shifting some of the algorithms into the database engine. We want to know whether the database is accessible and scalable to the expected rates and sizes with which the LOFAR catalogue will grow per year.

Estimates of the number of sources per second likely to be detected in typical LOFAR observation modes will serve as starting points in the choice of the database system, and the design of the database scheme. Benchmarking the most frequently used set of queries will eventually help in managing to process the large amounts of data entering the TKP database.

Miller-Jones (2008) calculated estimates on the number of source counts for the full array and the core configurations. We follow his calculations and investigate some LOFAR- and TKP-specific observing modes.

As will be shown in Section 2.4.1, the upper limit of disk space to be allocated for a source measurement is about 300 Bytes. A single measurement includes source properties like position, all Stokes parameters, timestamps, effective frequency, integration time, beam size, some auxiliary parameters for fast positional look-ups, and Gaussian fit parameters, plus all 1σ errors. All measurements of all sources are stored in an accessible and queryable database.

2.3.1 The Expected Number of Sources in the LOFAR Frequency Bands

Huynh et al. (2005) derived source counts from the 1.4 GHz Australia Telescope Hubble Deep Field–South (ATHDFS) Survey and compared these to other surveys carried out at the same frequency. They found a sixth-order polynomial function, describing the source counts down to the level of $50 \mu\text{Jy}$

$$\log\left(\frac{dN/dS}{S^{-2.5}}\right) = \sum_{i=0}^6 a_i \left[\log\left(\frac{S}{\text{mJy}}\right) \right]^i, \quad (2.6)$$

where $a_0 = 0.841$, $a_1 = 0.540$, $a_2 = 0.364$, $a_3 = -0.063$, $a_4 = -0.107$, $a_5 = 0.052$ and $a_6 = -0.007$. The differential number of source counts (dN/dS) is multiplied by $S^{5/2}$, as to have the ratio of observed numbers

to expected numbers in a simple Euclidean Universe, where the differential source counts should be constant.

Integration of this function will then predict the number of sources likely to be detected at 1.4 GHz, above a lower specified flux limit. However, if we assume that most of the detected sources, out of the Galactic Plane, are of extragalactic origin and obey a spectral flux scaling law described by $S_\nu \propto \nu^{-0.7}$, we can derive the expected source counts in the LOFAR frequency bands by scaling the predicted LOFAR rms sensitivities up to 1.4 GHz. Nijboer & Pandey-Pommier (2009) describe the general astronomical capabilities of the LOFAR Telescope. From these, and Eqs. 2.1–2.5, sensitivities at detection thresholds of five and thirty times the rms noise level can be computed for particular configurations of LOFAR. This is then used as the lower limit in the integral of the predicted source counts,

$$N = \int_{(5|30)\sigma_{\text{rms}}}^{S_{\text{up}}} S^{-2.5} 10^{\sum_{i=0}^6 a_i [\log(S/\text{mJy})]^i} dS, \quad (2.7)$$

where the upper limit is set high enough to not significantly contribute to the integral anymore; we set $S_{\text{up}} = 2 \text{ Jy}$.

Extrapolation of the LOFAR sensitivities in the High Band at the longer integration times to the frequency of Huynh’s model are below the $50 \mu\text{Jy}$ limit and may introduce uncertainties into the source count numbers. Only the full array mode at the longest logarithmic integration times at 120 and 150 MHz will be affected by this.

2.3.2 Confusion Limited Images

At the longer integration times, especially at the lower frequencies in core modes, the LOFAR images will be crowded with sources and the restoring beam can no longer distinguish between sources, which will affect the astrometry and photometry. Confusion becomes a problem when the source densities in the synthesised beam are larger than 1/50 to 1/15 (Hogg, 2001) depending on background noise. A standard rule of thumb is to use 1/30 as the confusion limit. Theoretically, differencing of properly calibrated images can reach the thermal noise levels. In our estimates we therefore use the criterion that an image is *classically* confusion limited, when the source counts times the ratio of the beam area and the field of view is larger than 1/30

$$\langle N \rangle \frac{\pi (\theta_{\text{res}}/2)^2}{\text{FoV}} > \frac{1}{30}, \quad (2.8)$$

where $\langle N \rangle$, θ_{res} , and FoV are given for several configuration modes in the next Section.

2.3.3 The Radio Sky Monitor

The Radio Sky Monitor (RSM) enables key observational modes of the LOFAR Telescope for the Transients Key Project. Initial strategies are described by Fender et al. (2007). The RSM has three observation modes, of which we will discuss the Zenith Monitor and the Rapid All-Sky Monitor, and not the Galactic Plane scans. Both modes exploit LOFAR’s capabilities of simultaneously observing large patches of the sky with multiple beams. With LOFAR’s large collecting area these survey modes are fast. The goal of these modes is to detect transients and variable sources at the different time scales and flux densities.

The Zenith Monitor Mode

The Zenith Monitor mode aims to stare at the zenith and map out the entire field of view that passes by. By using a hexagonal pattern of 7 beams formed by 24 core stations (LBA in inner configuration), an instantaneous field of view of 475 deg^2 at 60 MHz and 82.8 deg^2 at 150 MHz is constructed. This hexagonal pattern will scan a declination strip of about 20° wide, centered at $\delta = +54^\circ$, with a total area of 4211 deg^2 or about 10% of the entire sky. The number of pointings needed to scan the whole strip is about 12 at 60 MHz and 60 at 150 MHz. This means that an integration time for each field of 2 h in the Low Band and 30 min in the High Band is allowed to carry it out within a day. Sensitivities for single 4 MHz beams at these integration times reach the milliJansky level. Table 2.2 gives an overview of the characteristics of the Zenith Monitor (ZM) and the expected number of unique sources likely to be detected, data rates as measurements per second and the storage capacities needed.

| Freq. | θ_{res} | FoV | ΔS | N | m | d |
|-------|-----------------------|--------------------|------------|-------------------|---------------------|----------|
| [MHz] | [arcsec] | [deg^2] | [mJy] | [$\times 10^6$] | [s^{-1}] | [GB/day] |
| 60 | 413 | 475 | 6.0–510 | 0.16 | 1785 | 46 |
| 150 | 165 | 82.8 | 0.5–22 | 0.60 | 2996 | 78 |

Table 2.2: Characteristics of the RSM Zenith Monitor. The instantaneous field of view (FoV) of the hexagonal pattern of seven 4 MHz beams is given for the two observing frequencies. Note that it is not possible to observe at both frequencies simultaneously. The sensitivities listed range from the longest (7.2 ks at 60 MHz and 1.8 ks at 150 MHz) to the shortest (1 s) integration times. At the 5σ level, N is the number of distinct sources likely to be detected in this mode and m is the number of (source) measurements to be stored every second. d is the growth of disk size per day assuming a time allocation of 100%.

A more detailed description of the expected numbers specified per logarithmic integration time, the way the TKP will search for transients, is given in Table 2.7 in Appendix 2.A. Table 2.7 gives the number of sources likely to be detected, $\langle N \rangle$, for the given integration time τ_{int} at 5 and 30 times the rms noise, assuming a single 4 MHz beam for the observing frequency. The last column of Table 2.7, m , gives the total number of (source) measurements per second to be stored in the database. A grand total, for a *single* ZM 4 MHz beam, is given in the last row per frequency, which is simply the sum of the measurements made over all the integration times. Fig. 2.1 shows the number of measurements per second, m , to be stored at every integration time, τ_{int} for the ZM hexagonal pattern of seven beams. As noticed in Section 2.3.2, from Eq. 2.8 we deduce that the Low Band images are *classically* confused when $\langle N \rangle > 338$, which is at $\tau_{\text{int}} > 10$ s and $\tau_{\text{int}} > 500$ s for the 5σ and 30σ detection levels, respectively. In the High Band, the *classical* confusion limits arise when the integration times exceed 5 and 200 seconds for the two detection levels.

Tables 2.2 and 2.7 show that the RSM Zenith Monitor in LBA mode collects about 1800 measurements per second at the 5σ level, corresponding to about 150×10^6 measurements per day to be stored. For a single measurement a maximum storage size of 300 Bytes is anticipated (see Section 2.4.1), thus storing about 46 GB/day of 5σ measurements at 60 MHz. For the HBA frequencies, at the 5σ level, the numbers are nearly twice as large. At the

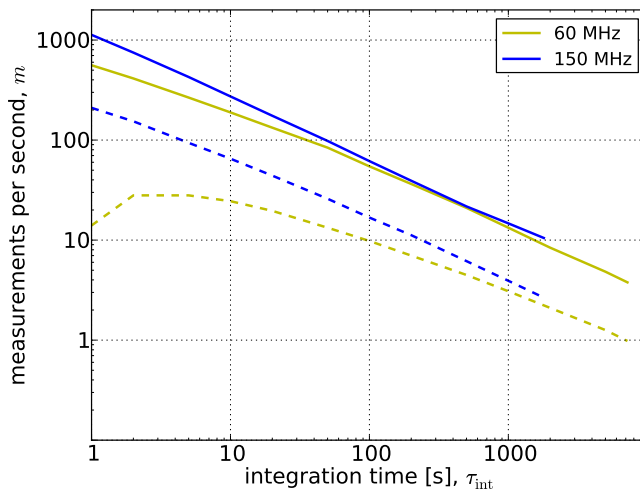


Figure 2.1: Expected number of measurements per second (m) to be stored in the RSM Zenith Monitor as related to the integration time (τ_{int}) at the two observational frequencies. For the seven-beam pattern, the thick lines represent the 5σ detection level measurements, whereas the dashed lines represent the 30σ level. A measurement of a source is about 300 Bytes of data.

30σ level this is about 13×10^6 measurements per day or 4.0 GB/day in the Low Band and 16 GB/day in the High Band. As the LB and HB cannot operate at the same time the average growth including both frequencies is then about 62 GB/day at the 5σ and 10 GB/day at the 30σ level.

Assuming the noise in an image to be Gaussian, the probability of having a 5σ pixel is 10^{-6} . The number of pixels for the full hexagonal pattern per day is about 3×10^5 pixels per FoV per second (see Table 2.2) which is approximately 3×10^{10} pixels per day, meaning that there will be 3×10^4 false noise spikes catalogued as a measurement. This corresponds to a fraction of about 3×10^{-4} of the total number of 5σ measurements per day. At the 30σ threshold the numbers of false positives can be neglected. More tests with confusion limited images and False Discovery Ratio (FDR) algorithm (Spreeuw, 2010) need to be carried out to choose a proper threshold as starting point for these observations.

Important for the database load, is the number of (source) measurements entering the database every second, $m \approx 1100$ at 150 MHz and $\tau_{\text{int}} = 1$ s, or $m_{\text{max}} \approx 3000$ at 150 MHz, that needs to be processed within 1 second. It is however not specified yet how the pattern of seven beams will be processed in the calibration and imaging pipelines. In Section 2.4.1 and further we use these numbers to benchmark the processing and detection algorithms.

The Rapid All-Sky Monitoring Mode

The Rapid All-Sky Monitor is a survey mode of LOFAR to search for the rare second-timescale transient events in the whole sky that is visible by LOFAR. Fender et al. (2007) describe the initial strategy for this survey. To have the largest field of view only core stations will be used with the LBAs in the inner configuration mode (Nijboer & Pandey-Pommier, 2009) and the core stations HBAs. In the calculations here, we will adopt a total of 24 core stations, and a bandwidth of 4 MHz per beam at the most sensitive observing frequencies of 60 and 150 MHz. According to Fender et al. (2007), the number of pointings needed to tile out the hemisphere is achieved when the beam pointings are offset by $\theta_{\text{FWHM}}/\sqrt{2}$, which corresponds to about 100 pointings at 60 MHz and 600 pointings at 150 MHz. The corresponding times to track a field are then 14 minutes and 140 seconds for the LB and HB, respectively. By spacing the integration times logarithmically a set of images is created in order to detect transients at different time scales. Table 2.3 gives an overview of the expected source counts, measurements per second to be stored, and the data growth per day for the RASM mode.

Table 2.8 in Appendix 2.A gives an overview of the configuration parameters and the expected source counts and measurements for the different integration times. From Eq. 2.8 and the numbers in Table 2.8 it can be seen that the *classical* confusion limit is reached at integration times longer than

2. Expected Data Rates and Volumes for the LOFAR Transients Key Project

| Freq. | θ_{res} | FoV | ΔS | N | m | d |
|-------|-----------------------|---------------------|------------|-------------------|--------------------|----------|
| [MHz] | [arcsec] | [deg ²] | [mJy] | [$\times 10^6$] | [s ⁻¹] | [GB/day] |
| 60 | 413 | 105 | 17.6–510 | 0.36 | 255 | 6.6 |
| 150 | 165 | 18.4 | 1.9–22 | 1.1 | 419 | 11 |

Table 2.3: Characteristics of the RSM Rapid All-Sky Monitor. The instantaneous field of view of a single 4 MHz beam is given for the observing frequencies. The sensitivities listed range from the longest to the shortest integration times. N is the number of distinct sources likely to be detected and m is the number of measurements per second to be stored. d is the growth of disk size per day assuming a time allocation of 100%.

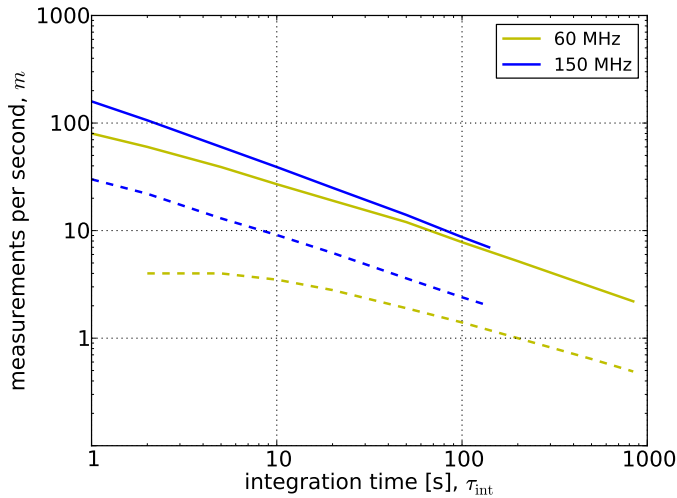


Figure 2.2: Expected measurements per second (m) to be stored in the RSM Rapid All-Sky Monitor as related to the integration time (τ_{int}) at the two observing frequencies. For a single 4 MHz beam (see text), the thick lines represent the 5σ detection level measurements, whereas the dashed lines represent the 30σ level. A measurement of a source is about 300 Bytes of data.

10 and 500 seconds in the Low Band for the 5σ and 30σ detection levels, respectively, whereas 5 second integration time is just below the limit in the High Band for the 5σ measurements. At 30σ no confusion limits are reached. Fig. 2.2 shows the measurements per second, m , made at the consecutive integration times for the Rapid All-Sky Monitor mode.

As can be seen from Tables 2.3 and 2.8, the average input rate per day is about 250 and 400 measurements per second at the 5σ level at the Low Band and High Band frequencies, respectively. The larger ratio at the 30σ level is caused by the relative higher sensitivity in the High Band at the shorter integration times, where more sources contribute in the sum. A 24 h RASM run collects about 22×10^6 measurements in the Low and 36×10^6

in the High Band at the 5σ level, resulting in storing 6.6 and 11 GB/day, respectively. At 30σ the storage capacity needed is 500 MB/day for the LB data and 2.3 GB/day for the HB data, assuming a 100% time allocation in a single Band. Again, the Low and High Bands cannot observe simultaneously. The dominating short time scales determine the rate at which the source measurements enter the database, where at the shortest time scale at 5σ , the storage rate is about 160 source measurements per second, with an average of about 350 source measurements per second, assuming that the time allocation is evenly distributed between Bands.

2.3.4 The Full Dutch Array

LOFAR's flexibility admits several configurational set-ups, all leading to slightly different resolutions, sensitivities and fields of view and thus source counts and data rates. We cannot treat them all, but here we will elaborate the full Dutch array mode (Full-NL), where all 24 core and 16 remote stations are used. The longest baseline is about 100 km and multiple beams from the same Band might observe at several frequencies at the same time. The only restrictions are that the LBA and HBA cannot operate at the same time and the total bandwidth of the beams sums to 48 MHz. This 40-station interferometer with its long baselines and large fields of view, achieves unprecedented resolutions and sensitivities at a wide range of time scales. In the calculations here we assume a 4 MHz beam for both Bands (LBA in inner configuration) and deduce the sensitivities (Eq. 2.4) with the logarithmically spaced integration times ranging from 1 to 10,000 seconds. Longer integration times are not treated here, since we focus on the transient detection strategy. Table 2.4 summarises for some selected frequencies the expected numbers, where we take into account using the maximum allowable bandwidth of 48 MHz, meaning we can construct twelve similar beams.

We refer to Tables 2.9 and 2.10 in Appendix 2.A for a more detailed overview of the Low and High Band frequencies of the Full-NL mode, respectively. Fig. 2.3 shows the measurements per second for the range of integration times for the Full-NL mode at the selected frequencies for a single (left) and twelve (right) 4 MHz beams.

From Tables 2.4, 2.9 and 2.10 it can be seen that at the shortest time scales in the lower end of both Bands the source counts are the highest, whereas the Low Band source counts outnumber the ones of the High Band, due to the applied spectral scaling law, $S_\nu \propto \nu^{-0.7}$, the field of view and the high sensitivity. The images will, however, not be confusion limited.

Similarly as in the previous sections, the summed data rates, m , for the whole range of integration times in twelve 4 MHz beams at 30 MHz at the 5σ level give an average of about 9000 source measurements per second that will be detected and have to be stored (see Table 2.4). Consequently, the disk space needed after a full day observing in the Full-NL mode at 30 MHz

2. Expected Data Rates and Volumes for the LOFAR Transients Key Project

| Freq. [MHz] | θ_{res} [arcsec] | FoV [deg ²] | ΔS [mJy] | N [$\times 10^6$] | | m [s ⁻¹] | | d [GB/day] | |
|----------------|-----------------------------------|----------------------------|---------------------|--------------------------|------------|---------------------------|------------|-----------------|------------|
| | | | | 5σ | 30σ | 5σ | 30σ | 5σ | 30σ |
| 30 | 20.6 | 419 | 11–1100 | 0.71 | 0.19 | 9000 | 600 | 230 | 15 |
| 60 | 10.3 | 105 | 3.9–390 | 1.0 | 0.29 | 4000 | 450 | 105 | 11 |
| 150 | 4.1 | 10.3 | 0.17–17 | 9.2 | 1.7 | 3500 | 780 | 91 | 20 |
| 210 | 3.0 | 5.3 | 0.23–23 | 5.3 | 1.2 | 1200 | 250 | 31 | 6.5 |

Table 2.4: Characteristics of the full Dutch Array Mode, assuming 24 core and 16 remote stations, and a beam spectral bandwidth of 4 MHz. Resolutions, θ_{res} , and fields of view, FoV_{Beam} , of single 4 MHz beams are given for the selected frequencies. The sensitivities, ΔS , range from the longest (10,000 s) to the shortest (1 s) logarithmic integration times. The number of distinct sources likely to be detected in 2π steradians, assuming observing only at the specified frequency, is given by N , for 5σ and 30σ detection levels and an integration time of 10,000 seconds. Using the maximum total bandwidth of 48 MHz (i.e. twelve of such 4 MHz beams), the corresponding 5σ and 30σ data rates, i.e. measurements to be stored per second, are given by m . The data growth per day, assuming a 100% time allocation, is given by d for the two detection levels.

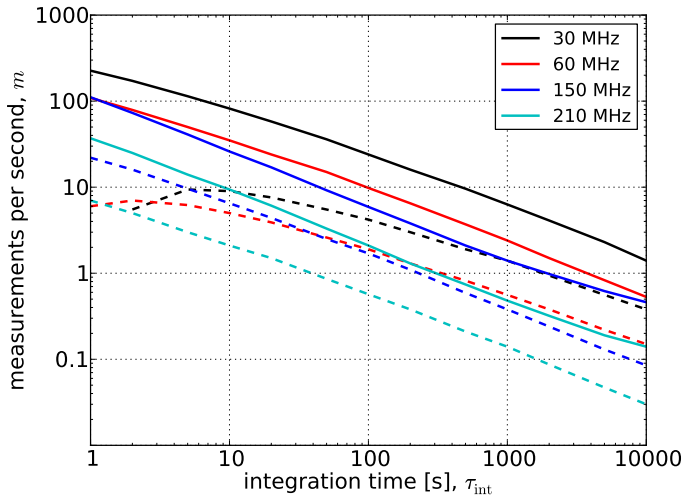


Figure 2.3: Expected measurements per second (m) to be stored in the Full-NL mode as related to the integration time (τ_{int}) for some selected frequencies, assuming a *single* beam of spectral beamwidth of 4 MHz, although 12 of these are allowed, giving rise to 12 times higher values. Thick lines represent the 5σ detection level measurements, whereas dashed lines represent the 30σ level. A measurement of a source is about 300 Bytes of data.

is $\lesssim 230$ GB. The numbers at the other frequencies yield lower data rates and less storage capacity, as can be seen from Tables 2.4, 2.9 and 2.10.

Furthermore, from the daily averages in Table 2.4, annual estimates can be made based on the data accumulation. Here we take into account that the TKP will be given as much allocation time as the other Science Key Projects, about 20%, and the assumption that the LBA and HBA observe evenly in time. Then we can conclude, if LOFAR is in the Full-NL mode, the TKP pipeline will process and store about 8 TB/yr at the 5σ level and about 0.8 TB/yr at the 30σ level. Of course, when commensal processing, *piggy-backing*, is enabled and the TKP is granted full data access, these numbers will be 5 times as high, with a maximum of 40 TB/yr.

2.3.5 Million Sources Sky Survey – Commensal Mode

The primary goal of the Million Sources Sky Survey¹ (MSSS) is to build a Global Sky Model that serves the calibration of LOFAR. The Global Sky Model (GSM) contains the spectral information of $10^5 - 10^6$ sources in the Northern Hemisphere in the frequency range of roughly 30–200 MHz. It is envisaged that because of the excellent *uv*-coverage of 24 LOFAR core stations, a 10 min. observation around transit gives sufficient *uv*-coverage to carry out MSSS. The observing frequencies will lie roughly in the middle of the Low and High Band, at 60 and 150 MHz, respectively. About 600 pointings for the LBA and 3500 for the HBA array are needed to Nyquist sample the Northern Hemisphere. The achievable rms sensitivities depend

¹http://www.astron.nl/mssswiki/lib/exe/fetch.php?media=-lofar_tsm_msss_debruyen.pdf

| Freq. [MHz] | θ_{res} [arcsec] | FoV [deg ²] | ΔS [mJy] | N [$\times 10^6$] | | M [$\times 10^6$] | | D [GB] | |
|----------------|-----------------------------------|----------------------------|---------------------|--------------------------|------------|--------------------------|------------|-------------|------------|
| | | | | 5σ | 30σ | 5σ | 30σ | 5σ | 30σ |
| 60 | 413 | 105 | 10–255 | 0.52 | 0.13 | 188 | 26 | 56 | 7.8 |
| 150 | 165 | 18.4 | 0.45–11 | 3.4 | 0.86 | 1525 | 363 | 456 | 109 |
| Totals | | | | | | 1713 | 389 | 512 | 117 |

Table 2.5: Characteristic parameters of the Million Sources Sky Survey (MSSS) once it has been carried out by using 20 core stations at the LB and HB frequencies of 60 and 150 MHz, respectively, and assuming a 16 MHz beam. The number of pointings needed to Nyquist sample the Northern Hemisphere for the LB and HB frequencies is $P_{\text{LB}} = 619$ and $P_{\text{HB}} = 3515$, respectively. ΔS is the theoretical 1σ rms noise level of the longest (600 s) and shortest (1 s) integration times. The total number of distinct sources likely to be detected by the Survey is given by N for 5σ and 30σ detection levels. In TKP commensal mode M is the number of total source measurements made and stored in the TKP database after survey completion and the storage disk size needed is given by D in the last two columns for 5σ and 30σ detections (see text for the derivations).

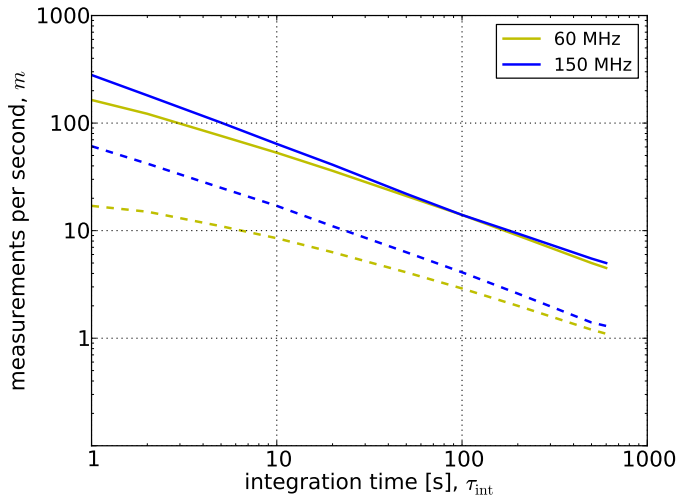


Figure 2.4: Expected measurements per second (m) to be stored in the MSSS commensal mode as related to the integration time (τ_{int}) for the observational frequencies, assuming a beam of spectral beamwidth of 16 MHz. Thick lines represent the 5σ detection level measurements, whereas dashed lines represent the 30σ level. A measurement of a source is about 300 Bytes of data.

on whether the final strategy is to use a single beam of 48 MHz bandwidth (i.e. covering the whole Low Band at once) or three beams of 16 MHz bandwidth each. In our calculations we adopt the latter.

Table 2.5 reports some resulting parameters of MSSS once the survey has been completed in the assumed mode. The number of distinct sources, N , likely to be detected is more than three million at the 5σ level, and about a million at the 30σ level, if no cross-frequency band source association is taken into account.

However, commensal detection of transients in the MSSS Survey allows the Transients Key Project to inspect the MSSS data at logarithmically spaced time intervals at which the calibrated images will be produced. For this *piggybacking* mode, the expected number of sources likely to be detected and the number of measurements made in a single 16 MHz MSSS beam at the frequencies 60 and 150 MHz at different integration times are presented in Table 2.11 in Appendix 2.A. As can be seen from Eq. 2.8 and Table 2.11 images are *classically* confused in the Low Band at integration times longer than 2 and 100 seconds at the 5σ and 30σ detection levels, respectively, whereas in the High Band integration times as short as seconds cause the images to be confusion limited. Fig. 2.4 shows the measurements per second, m , plotted against the integration times for the MSSS commensal mode.

From Table 2.11 it can be seen that for both Bands the expected source counts, $\langle N \rangle$, are of the same order. The HB, however, needs about 5.5

times more pointings to sample the whole sky, because of its smaller field of view. Furthermore, from the values of $\langle m \rangle$ at the shortest integration times it can be deduced that when we take into account *three* similar beams operating at the same time, the average rate is of the order of 1800 sources per second at the 5σ level. The last two columns in Table 2.11 show the total number of measurements made per second in this configuration at the two significance levels. Multiplying the summed value from all integration times by the number of pointings to be made and the time spent on a field (i.e. 600 s), will give the total number of measurements made in the Survey, M , and is given in Table 2.5 for the 5σ and 30σ detection levels. The summed measurements made during the MSSS Survey at the 5σ level has a total of approximately 1.7×10^9 , which corresponds to about 0.5 TB of disk space needed, whereas at 30σ this is about 400×10^6 measurements and 120 GB of data. It should be noted that these numbers are TKP specific, assuming that the data can be processed commensally, and do not reflect the raw data formats to be stored on the cluster nodes. Furthermore, if the MSSS Survey would produce a real-time data stream of calibrated images, the TKP pipeline should be able to store about 2000 (source) measurements per second, assuming three simultaneously operating 16 MHz beams.

2.3.6 Summary

The TKP pipeline faces a wide range of data rates to process and volumes to store, depending strongly on the observation mode of the LOFAR Telescope. Table 2.6 gives a comparable overview of some of the key observation modes of LOFAR relevant to the TKP. Averaged values take into account that the allocation time of the LBA and HBA is evenly distributed during an observation run.

| Observation Mode | m [s ⁻¹] | | $\langle d \rangle$ [GB/day] | |
|------------------|------------------------|------------|------------------------------|------------|
| | 5σ | 30σ | 5σ | 30σ |
| ZM | 2400 | 400 | 62 | 10 |
| RASM | 350 | 50 | 9.1 | 1.3 |
| Full-NL | 4400 | 500 | 114 | 13 |
| MSSS* | 1800 | 360 | 47 | 9.3 |

Table 2.6: Averaged data rates and volumes expected for some key observational modes of LOFAR. (*) refers to the fact that MSSS will be processed in *piggybacking* mode by the TKP and is carried out once, while the others are repetitive modes. MSSS will store in total about 0.5 TB of data at the 5σ and about 120 GB at the 30σ level. The Zenith Monitor mode has seven beams scanning the zenith during operation, giving rise to larger rates than the RASM mode where we assumed a single beam operating during observation. Note that for the Full-NL mode the measurement peak values may be as high as 9000 (see Tables 2.4, 2.9 and 2.10).

The full Dutch array mode (Full-NL) is the most data intensive configuration, and in Section 2.3.4 the estimated annual growth of the data size was $\lesssim 8$ TB/yr, if LOFAR were in this mode 20% per year. Other modes of LOFAR are less intensive, and commensal modes will increase this percentage probably by a factor of 5 to ≈ 40 TB/yr. Note that the presented numbers do not take into account any back-up storage or processing facilities.

Going from *"working to working"*, starting at the 30σ detection threshold will already offer a wealth of information for the various modes, detecting hundreds of thousands of sources, making millions of measurements per day.

2.4 TKP Databases

The data rate and cadence of LOFAR put stringent demands on the TKP pipeline. The time between two subsequent images should not exceed the combined source extraction and database processing time of an image. Image processing times depend strongly on properties like size, integration time, number of frequency bands, resolution and sensitivity. From a database perspective, these may be condensed in the expected number of sources to be detected in an image, which determines the disk space to be allocated, and the query execution times. Spreeuw (2010) gives a rigorous description of the source extraction modules that are implemented in the TKP pipeline, while the previous section of this chapter focuses on the expected database storage sizes of the detected sources. Disk space needed for the storage of the raw data and images is taken care of by the LOFAR Storage Group and will not be discussed here. In this section we will describe the database system and schema used in the TKP pipeline to meet the data flow and retrieval in an optimal way.

The Sloan Digital Sky Survey is by far the most successful astronomical project that uses a database-centric computing approach for their large-scale scientific datasets. Furthermore, it serves as a pathfinder for current and future planned surveys at multiwavelengths, that will be taking enormous amounts of data, e.g. the Panoramic Survey Telescope and Rapid Response System (Pan-STARRS; Hodapp et al., 2004) and the Large Synoptic Survey Telescope (LSST; Abell et al., 2009) in the optical band and the Allen Telescope Array (ATA; Welch et al., 2009) and Murchison Wide-field Array (MWA; Lonsdale et al., 2009) in the radio band. Their database schema, functions, procedures and algorithms, which has been vastly extended and evaluated over the years, is a good example how the architecture of the environment should be built. The way to approach and start is formulated by one of Gray's laws: *"Bring computations to the data, rather than data to the computations"* (Szalay & Blakeley, 2009). This is why procedures and functions run inside the database. The Structured Query Language (SQL)

can access the data directly, giving back summarised or full result sets, instead of transporting the data to other nodes in- or outside the network and process the results, expensively, iteratively. Two other rules from Gray are to start the design with the *"top 20 queries"* and go from *"working to working"*. Because we operate in a pipeline framework, in which data are processed in a structured way, we know what our top 20 queries have to do, enabling us to optimise their execution plans.

2.4.1 The TKP pipeline Database Schema

Overviews of the TKP pipeline framework in which the calibrated images are processed are given by, e.g., Swinbank et al. (2007) and Swinbank (2010). Spreeuw (2010) developed the Source Finding modules, while here we will focus on the database part. A diagram of the data flow centred at the database interactions in the TKP pipeline is shown in Fig. 2.5. From Fig. 2.5 it can be seen that two separate databases are implemented. One is used in the nearly real-time pipeline during observations and the other is used as a permanent offline catalogue database, collecting all the LOFAR source measurements over time.

At the start of an observation the `pipeline` database is initialised and loaded with known sources from the `catalog` database that are in the fields of view during the running observation. These will include sources from the major catalogues as well as the known LOFAR sources once the MSSS Survey has been carried out. The Source Finding modules extract the Gaussian fitted sources and fluxes at positions of particular interest on the sky from the images and pass them on to the database. In this `pipeline` database sources are then associated to previously detected sources in the observation and to the already known sources, preloaded from the catalogues. After this the transient and variability detection queries run that may send triggers for further actions. Finally, typically on a daily basis, the data will be flushed to the `catalog` database where the measurements will be appended to the LOFAR catalogue sources. Once LOFAR is fully capable of observing in several different modes, LOFAR's TKP catalogue is expected to grow with roughly 8 TB/yr, as was worked out in the previous section.

The `pipeline` database is at the heart of the TKP pipeline system. It should be kept clean and fast to be able to detect differences on the shortest time scales in the images. Therefore, the `pipeline` database has a temporary character because its content will be flushed at certain periods in times to the permanent `catalog` database. Furthermore, due to its temporary and limited size, the `pipeline` database can reside on a single node, close to the input data to follow Gray's Law.

The goal of the `pipeline` database is to serve as a tool for the TKP pipeline for detecting transient and variable sources. The properties and measured quantities of all the detected sources will be stored in the database.

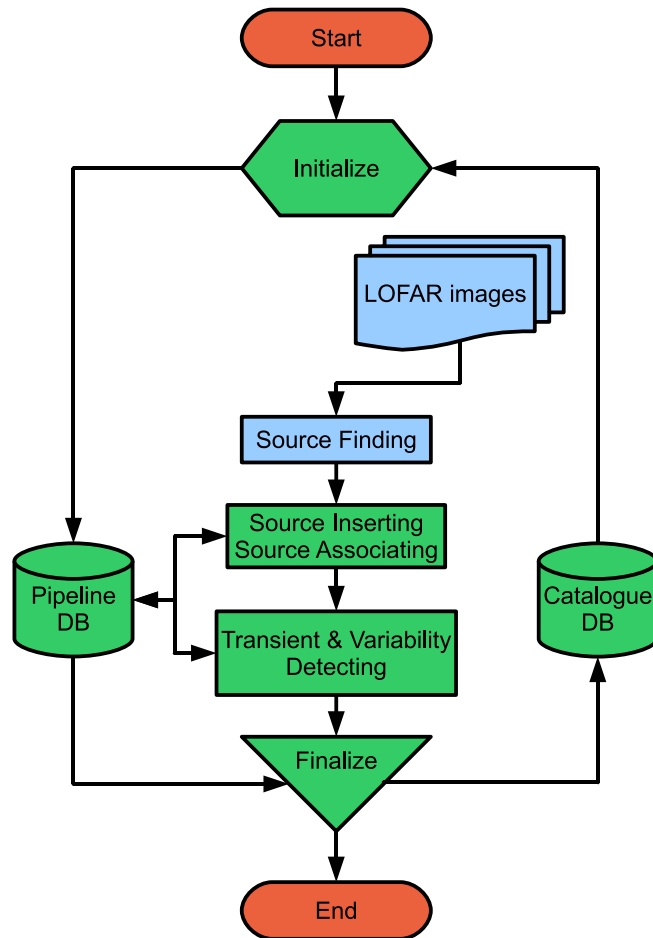


Figure 2.5: Schematic view of the data flow in the TKP pipeline.

In order to respond quickly when necessary, the design should be optimised to these needs. The full and up-to-date data definition and manipulation SQL statements of the database, tables, functions and procedures are maintained in the TKP svn repository². A schematic overview of the most relevant tables in the TKP pipeline database is given in Fig. 2.6.

In the design of the pipeline table definitions we adopt Gray’s Laws and incorporate the aspects of the lessons learned from the SDSS SkyServer (Thakar, 2008). Central in the pipeline database are the `extracted-sources` and `catalogedsources` tables, together with the `basesources` and `assocatsources`, containing the association information. Extracted

²<http://svn.transientskp.org/code>

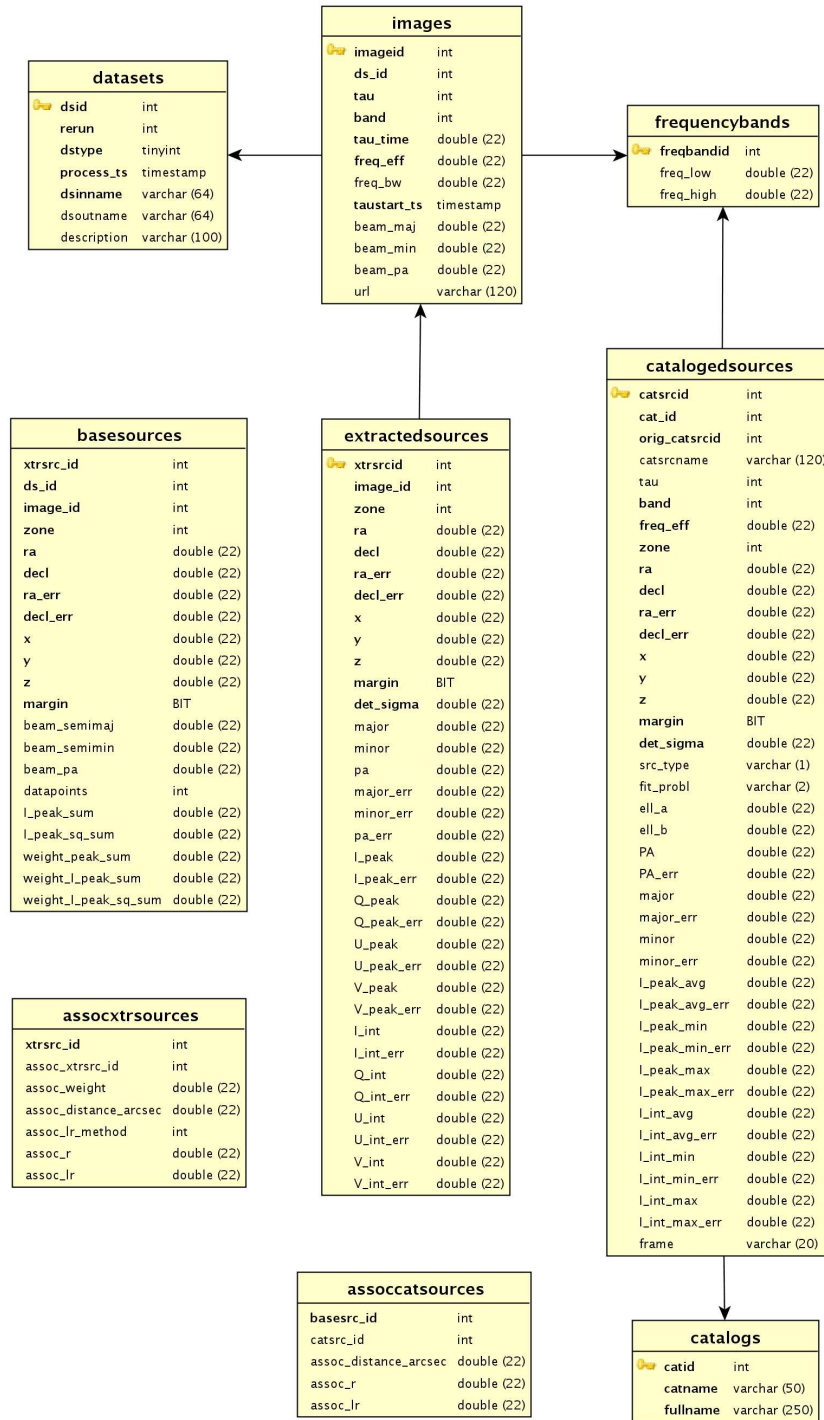


Figure 2.6: Schema of the most relevant tables in the TKP pipeline database. Arrows indicate column references between tables.

sources from the Source Finding procedures will be appended to the `extractedsources` table, containing the measured source properties and their 1σ errors, like position, all Stokes parameters of peak and integrated fluxes, the Gaussian fitting parameters and via `images` the observational frequency and timestamp information, giving rise to about 300 Bytes of data. The `catalogedsources` table contains the preloaded sources from the major catalogues.

The `basesources` table keeps track of the unique sources per band detected in the current observation, and may be regarded as a running catalogue. The band is a standard frequency band as specified in `frequencybands`. For every new detection of a source, the association parameters (columns) will be updated with the new averaged values. The determination of a source association is treated in Section 2.4.3 and Chapter 3.

Sources in the running LOFAR catalogue, i.e. the `basesources` table, will be mapped to the known sources from the catalogues and stored in the `assoccatsources` table. Similarly associations between current and previously observed sources are maintained in the `assocxtrsources` table as light-curve datapoints. In this way we keep the status of all the (unique) sources detected in the current observation.

The main tables are defined in such a way that the data points from the same source are retrieved in a simple SQL statement, and might therefore be regarded as the light-curve tables. To optimise source association, transient and variability detection, and the data selections for the dumps to facilitate data transport between the two databases, we will create the tables with several columns that also exist in the SkyServer data model as described by Stoughton et al. (2002) and enhanced by Gray et al. (2006).

Ivanova et al. (2007) analysed a typical SkyServer query log of 1.2 Mqueries and found that 83% contained spatial data searches. Gray et al. (2006) replaced the recursive hierarchical triangular mesh (HTM) algorithm by the faster zone algorithm. The latter divides the celestial sphere into declination strips of equal width, a so-called `zone`, defined as an INT in the clustered primary key (`zone`, `ra`, `id`) of the table. Combined with the Cartesian coordinates, (`x`, `y`, `z`), the dot product is used to calculate distances between sources. To compare the processing times in a MySQL database of both algorithms, we generated 1000 images of 1000^2 pixels with tens of sources in each image and processed these in the TKP pipeline. From the results shown in Fig. 2.7 it can be seen that the zone algorithm is nearly two orders of magnitude faster, although both increase with the number of images being processed, due to the increasing number of sources needing to be searched.

Because the source association algorithms and related queries rely heavily on spatial searches and are the most intense processing tasks inside the `pipeline` database, we decided to use the zone algorithm.

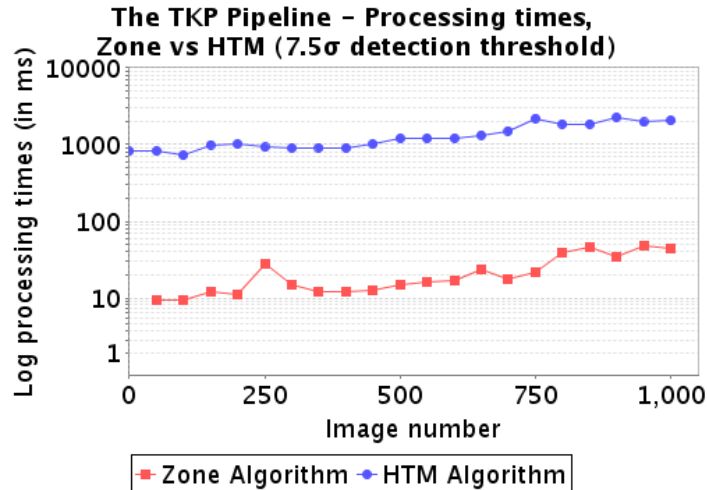


Figure 2.7: Processing times (vertically) of the HTM (blue circles) and Zone (red squares) algorithms for a series of 1000 simulated images being processed (horizontally). The zone algorithm is nearly two orders of magnitude faster. It was implemented on an Intel(R) Pentium(R) 4 CPU 3.00 GHz with 1 GB RAM, running Fedora 3 (Linux kernel 2.6.12), using MySQL 5.0.22.

2.4.2 MonetDB

The TKP pipeline has implemented the open source column-store database MonetDB³ (see e.g., Boncz, 2002), developed at the Centre of Mathematics & Informatics (CWI). The column-store model was formerly known as the Decomposition Storage Model (see e.g., Copeland & Khoshafian, 1985; Khoshafian et al., 1987); it splits up a (relational) table vertically into c_n binary tables, where c_n is the number of columns. This database system is of a fundamentally different design than the classical relational database systems (RDBMS), such as the open source MySQL and PostgreSQL, or the commercial products Oracle or DB2, but all can be interfaced with the same Structured Query Language (SQL). At CWI, TPC-H⁴ benchmark performance tests were executed on the alternative open source database systems of MySQL and PostgreSQL. A series of queries is executed on a predefined and preloaded database. Scaling factors of up to 20 times the initial size of 1 GB reveal that it is not uncommon for the tested classical RDBMSs to give erroneous (empty) result sets and/or that the processing times are extremely long for some of the queries⁵.

Full-sized SDSS SkyServer data releases were successfully ported into MonetDB. Although the SkyServer data management system is a tuned

³<http://monetdb.cwi.nl>

⁴<http://www.tpc.org/tpch/>

⁵<http://monetdb.cwi.nl/SQL/Benchmark/TPCH>

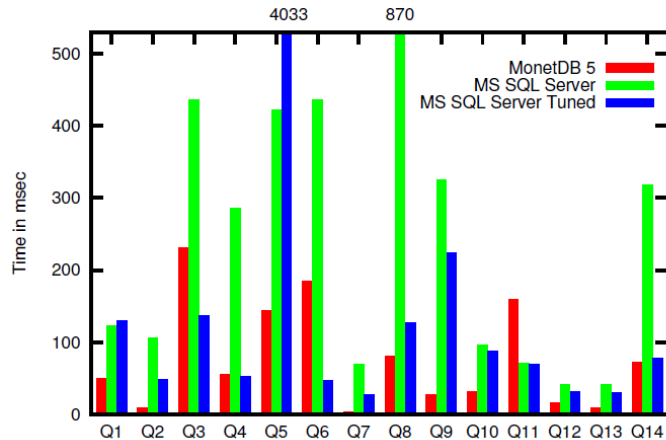


Figure 2.8: A subset of 2 GB of Data Release DR4 was ported into MonetDB. The elapsed times for a test set of 14 queries were compared between MonetDB and a tuned and non-tuned Microsoft SQL Server. 12 Queries are executed faster in MonetDB. Figure adapted from Ivanova et al. (2007).

Microsoft SQL Server, Ivanova et al. (2007) showed that on a subset of the data, the performance of 12 of the 14 most executed queries ran faster in MonetDB, see Fig. 2.8.

In MonetDB every relational table is represented by a group of binary relations, consisting of Binary Association Tables (BATs). A BAT represents a mapping from a unique object id to a single attribute. When the object ids form a dense ascending sequence highly efficient positional lookups are enabled. Direct consequences are that queries only touch the relevant columns, and when in contiguous memory it allows compression and good cache-hit ratios. Furthermore, MonetDB’s kernel is a programmable relational algebra machine operating on ”array”-like structures, exactly what CPUs are good at.

To speed up query processing further, MonetDB/SQL implements a query processing architecture based on cracking, in which a column is sorted according to the subsequent insert statements that touch the column. In this scheme the first query pays the price, but all the others benefit from previous queries. Because our database starts small the drawback from this initial start-up is small. Idreos, Kersten & Manegold (2002) showed that a simple `count(*)` query with a range predicate of which the two boundaries were randomly chosen, and was fired a 1000 times, each time with a new random range, on a single column table populated with 10^7 random values between 0 and 9999 achieved response times that were two orders of magnitude faster than those from PostgreSQL, MySQL and MonetDB with cracking disabled. Fig. 2.9 is adopted from Idreos, Kersten & Manegold (2002) and shows clearly that cracking makes sense at an early stage,

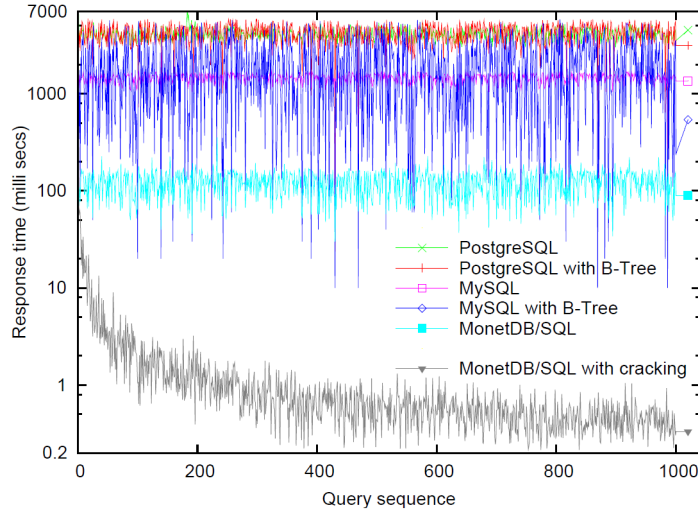


Figure 2.9: A simple count(*) query. Figure adapted from Idreos, Kersten & Manegold (2002).

and even more cracking will decrease the response times significantly. This technique fits the TKP databases that start small and grow gradually by appending data, since that is the optimal case for applying the cracking algorithms.

2.4.3 Associating LOFAR Sources in the Database

Although the format of the input data is not fully specified yet, we can make some assumptions that will hold, because from the database perspective the number of sources detected per image is relevant. Besides LOFAR’s configuration mode, the characteristic properties of an image are the resolution, synthesised beam parameters, integration time, frequency band, and timestamp of observation, all stored in the `images` table. We envisage a dataset as streams of image cubes arriving at subsequent timestamps. Streams are divided according to their logarithmic integration times ($\tau_1, \tau_2, \dots, \tau_{13}$), and each image cube has the same observational timestamp, whereas the individual image planes in this cube fall in different frequency (sub)bands. Fig. 2.10 depicts this, and in this view a dataset might also be regarded as an observation producing the image cubes. From Fig. 2.10 it can be seen that we can search for transient and variability behaviour in the time and frequency domains.

All measured properties of a source, e.g. position, frequency and all Stokes parameters, plus errors, are stored in the `extractedsources` table, which is essentially the table containing all the measurements made during an observation. The corresponding properties of the image, in which the source

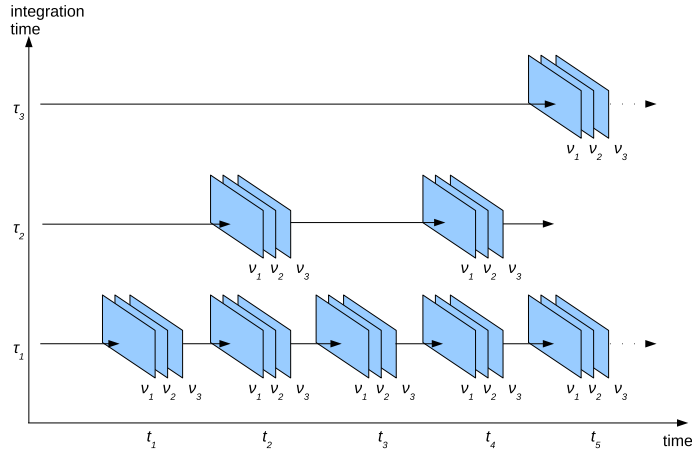


Figure 2.10: Schematic view of the TKP pipeline input streams of image cubes, all belonging to the same dataset. The blue planes represent images taken at certain frequencies ν_1, ν_2, \dots , integration times $\tau_1, \tau_2, \dots, \tau_{13}$, and at observation timestamps t_1, t_2, \dots, t_n . Image planes of same integration time and timestamp, but with different frequencies, are grouped into image cubes. A dataset is considered as the collection of all the image cubes.

was detected, are retrievable via ID referencing. A list of unique sources in the current observation is maintained in the `basesources` table, and a list of known sources, from major catalogues and eventually classified LOFAR sources, is kept in the `cataloguedsources` table. Positions of sources detected by LOFAR are checked for uniqueness and reoccurrence against both lists. Finding a positional match – either a genuine or background association – is done by the source association procedure, which is carried out inside the database by a couple of SQL commands. Construction of light curves is moderated by joining the `basesources` and `extractedsources` tables.

The goal of the source association is to find for every source detected by LOFAR all its measurements, current and archived, in order to construct light curves that will aid the source classification. The criteria for which an association pair is considered as real or by chance is done by evaluating three association parameters, as described in Chapter 3. One of the association parameters that is very useful is the normalised distance between the two sources i and j , weighted by their positional uncertainties:

$$r_{ij}^2 = \frac{(\alpha_i - \alpha_j)^2}{\sigma_{\alpha,i}^2 + \sigma_{\alpha,j}^2} + \frac{(\delta_i - \delta_j)^2}{\sigma_{\delta,i}^2 + \sigma_{\delta,j}^2}, \quad (2.9)$$

which follows a Rayleigh probability distribution. Cutoff values for this dimensionless positional difference were determined by the simulation runs

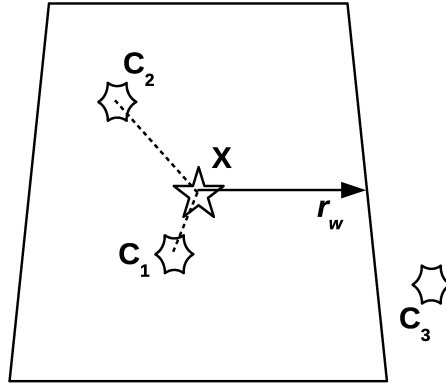


Figure 2.11: Extracted source X has two association candidates C_1 and C_2 , that were found in the box with width and height of $2r_w$ at the location of source X . Source C_3 falls outside the search area and is not considered as a candidate. The measured properties of all the sources, X, C_1, C_2 , and C_3 are stored in the database. Depending on the association parameters of the candidate(s), $X - C_i$, the pair may be classified as genuine or chance. If genuine, it is recorded as a related measurement to the source. If it could not be associated, it will be recorded as the first measurement of a source. See text for other cases.

discussed in Chapter 3.

The position of an extracted source is placed at the centre of a box that is searched for counterparts. The distance of the source to the edges of the box, r_w , is set fixed in the north- and westward directions, to a value of order of 1 arcmin, but will be determined dynamically depending on the image resolution and local source density in future versions. All sources found in the area are considered as candidate associations. Fig. 2.11 gives a sketch of the case where two candidates were found within the search area of an extracted source.

Then, for every pair the association parameters are calculated and based on the criteria the association is considered as real or by chance. The database implementation of the source association algorithm takes care of the situation where extracted sources from an image are matched to previous detections, as depicted in Fig. 2.12 and is as follows:

- (i) For associations found to be genuine, the measurements are appended to the corresponding (unique) source in the running source list, i.e. we **update** the `basesources` table. This table maintains the source position and frequency dependent averaged values for flux, $\overline{I_\nu}$, flux squared, $\overline{I_\nu^2}$, weight of flux, $\overline{w} \equiv 1/\sigma_{I_\nu}^2$ weighted flux, $\overline{wI_\nu}$, weighted squared flux, $\overline{wI_\nu^2}$, and the number of data points N . These values are used in the variability monitoring indices, which are the primary tools for detecting transient and variability behaviour in the LOFAR

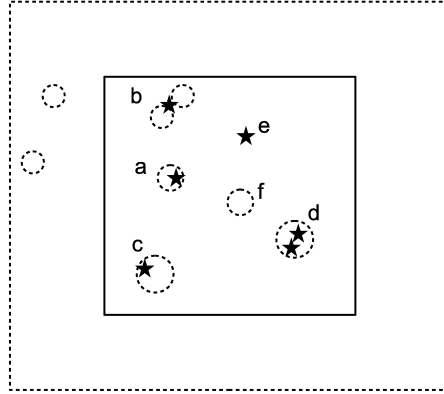


Figure 2.12: Associating extracted sources (star symbols) from an image (square, continuous line) to previous detections (dashed circles) in the same field of view (dashed square). The cases that are taken care of by the association procedure are labeled alphabetically.

sources. The indices are defined in Chapter 3 and are given by

$$V_\nu \equiv \frac{s_\nu}{\bar{I}_\nu} = \frac{1}{\bar{I}_\nu} \sqrt{\frac{N}{N-1} (\overline{I_\nu^2} - \bar{I}_\nu^2)} \quad (2.10)$$

and

$$\eta_\nu = \frac{N}{N-1} \left(\frac{\overline{wI_\nu^2}}{w} - \frac{\overline{wI_\nu^2}}{\bar{w}} \right). \quad (2.11)$$

From Fig. 2.12 it can be seen that case a follows the above described procedures, whereas case b is similar, except that both association candidates are considered as genuine and no discrimination can be made between the two. Both sources in the source list will be updated with the new values from the extracted sources.

- (ii) Sources extracted from a higher-resolution image might be resolved compared to the association candidate sources. In these cases, c and d in Fig. 2.12, we replace the lower-resolution source position in the running source list with the higher-resolution source(s). As a consequence, the averaged values needed for the variability indices should be recalculated for this higher-resolution source.
- (iii) Other situations are the first detection of a source, case e in Fig. 2.12, in which case the newly detected source will be added to the running source list. Internal triggers should take further analysis actions, in order to classify this new source. Case f in Fig. 2.12 represents the case

where a previously detected source is not detected again, in which case the local noise levels in the image should be reported.

The sources in the image will also be looked up for occurrences in the static `catalogedsources` table, following analogous procedures as described above. Associations between sources in the running source list (i.e. sources stored in `basesources`) and the known sources from the catalog source list (i.e. the sources stored in `catalogedsources`) are maintained in the `assoccatsources` table.

We define a group of SQL statements that execute the above described procedures, trying to associate all the sources detected in a LOFAR image. Parallelised processing of the images is taken care of by atomic transaction-safe storage of all sources detected in an image. Therefore, the association procedure should run on consecutive, but not necessarily chronological images, and not in parallel.

Spatial searches are the most used predicates in queries that select one or (many) more sources, and they should therefore have fast response times. A location of interest, `(@ra,@decl)`, is set at the centre of a searchable area with radius `@r_search`. All sources that are within a distance `@r_search` of this location are found when using the dot product and Cartesian instead of celestial coordinates. They should obey the clause $\mathbf{x} \cdot \mathbf{c} > \cos r_s$, where \mathbf{x} and \mathbf{c} represent the Cartesian vectors of the location of interest and the sources found, respectively, and r_s is search radius in radians. Refinement of the search is done by excluding the candidates that fall outside the box surrounding the search area, see Fig. 2.11. Candidates having declinations between the minimum and maximum values of the box are included. An extra clause `zone BETWEEN @zone_min AND @zone_max` is used for the candidates falling in the declination strips of the box. From Gray et al. (2006) we adopted the `alpha(@r_search, @decl)` function that inflates the RA of the search radius, with increasing declination towards the celestial poles. It is applied in the RA box boundaries as `ra BETWEEN @ra - alpha() AND @ra + alpha()`. These four clauses make up the spatial search predicates.

Database Processing Times

Code snippets of the most intensive queries that make up the source association and detection procedures are shown in Appendix 2.B, where the first query collects the just extracted sources and their association counterparts from the running catalogue and the second query updates the running catalogue with the new values. These queries were timed for the benchmark tests and were executed in a MySQL and MonetDB database. Identical database tables and queries were created, except for a few minor SQL syntax differences, and both were installed on the same machine, a dual-core

2. Expected Data Rates and Volumes for the LOFAR Transients Key Project

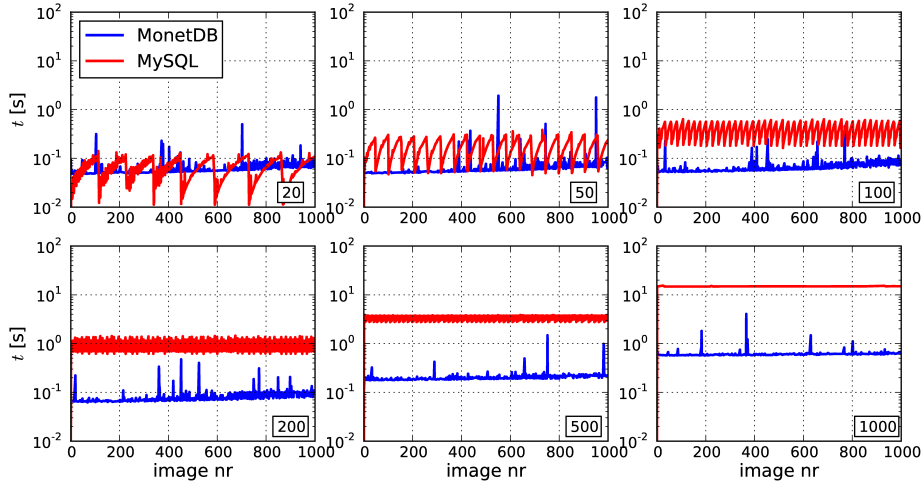


Figure 2.13: Comparison of performance tests of the source association procedures in a MySQL 5.0.45 (red line) and MonetDB v5.20.4 Jun2010-SP1 (blue line) database, carried out on a dual-core 64 bit Intel(R) Pentium(R) 4 CPU 3.00 GHz with 1 GB of RAM, running Fedora 8 (Linux kernel 2.6.26.8-57) desk-top computer. We processed a series of 1000 images (horizontal axes), each containing the number of sources as labeled in the bottom right of the subplots. The response times are shown on the vertical axes.

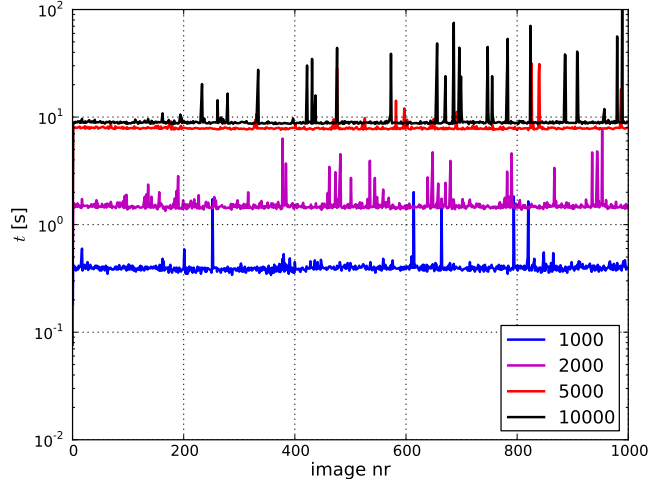


Figure 2.14: Performance tests of the source association procedures on the data server node in the LOFAR computing cluster. The installed database is MonetDB (Feb2010-SP2), and runs on an eight-core 64 bit Intel(R) Xeon(R) CPU L5420 2.50 GHz with 16 GB of RAM. We processed a series of 1000 images (horizontal axes), each containing the 1000 sources per images. The response time is shown on the vertical axes.

64 bit Intel(R) Pentium(R) 4 CPU 3.00GHz with 1 GB of RAM, running Fedora 8 (Linux kernel 2.6.26.8-57). The series of queries were executed by a Python script that interfaced either with MonetDB or MySQL. For this, 1000 images, each containing n_s sources, were processed in the TKP pipeline, where n_s was set to 20, 50, 100, 200, 500 and 1000. For $n_s > 1000$ the processing took unacceptably long, which might have been caused by unavailable memory resources. Fig. 2.13 shows the processing times of the association and detection procedure in the two database systems.

Although this test was carried out on a relatively modest machine, it shows the capabilities of the column-oriented database systems as compared to the classical systems. These results made us decide to use MonetDB in the Transients pipeline. The same tests ran on the data server in the LOFAR computing cluster environment and the results of the processing times of 1000 images containing $n_s \geq 1000$ sources each are shown in Fig. 2.14. It has to be noted, however, that we did not have full control over the other processes running on the data server, where we experienced about 20% differences between some runs.

2.5 Discussion

As was shown in Section 2.3, a fully operational LOFAR Telescope will provide the Transients Key Project with millions of sources that will be detected millions of times, giving peak data rates that may well exceed 200 GB/day and with storage capacities needed per year of the order of 100 TB if we include commensal, *piggybacking*, processing and data backups. A dedicated fast-responding database system should be available and accessible from within the pipeline framework during ongoing observation runs, whereas the database query and procedure response times should be obviously less than the input rates.

The performance tests reported in Section 2.4.3 show the relationship between the number of sources per image and the processing response time for a series of 1000 consecutive images. If the processing time is shorter than the time between two consecutive images, the observation mode is executable from a database point of view. The same is valid for the time to carry out surveys, where the data need to be processed within the survey time. In these figures we have to take into account the processing times of the calibration and imaging pipelines and the source extraction procedures as well. These depend on many parameters and are not in the scope of the investigations here. However, the tests pointed out that MonetDB is able to process 1000 source measurements within 1 second, whereas MySQL processing times exceed ten seconds. This means that monitor modes can be carried out from the database perspective. At the shortest integration

time scales and when using the full 48 MHz bandwidth, the full-array mode (of which an example was discussed in Section 2.3.4), exceeds the limits of the number of source measurements that are still processable at the shortest time scales. Assuming the ultimate image input rate is one per second, we cannot process these at similar time scales yet. Simple data server hardware extensions of the RAM memory and running it in a more dedicated mode in which MonetDB is installed, will improve the performance (Boncz, Kersten & Manegold, 2008), and allow us to process a larger number of sources per image. Minor improvements of the response times are to be expected from converting data type definitions to a smaller number of bits, e.g., conversion of INT to TINYINT gains a storage factor of four.

Acknowledgement

I would like to thank Ronald Nijboer for helpful discussions about the LOFAR configurations and observation modes.

2.A Source Counts & Data Rates of LOFAR Modes

| Freq. [MHz] | θ_{res} [arcsec] | FoV [deg ²] | Image [pixels] | τ_{int} [s] | ΔS [mJy] | $\langle N \rangle$ | | $\langle m \rangle$ | |
|----------------|-----------------------------------|----------------------------|-------------------|----------------------------|---------------------|---|-----|---------------------|------|
| | | | | | | [FoV ⁻¹ τ_{int}^{-1}] | | [s ⁻¹] | |
| 60 | 413 | 105 | 512 ² | 1 | 510 | 80 | 2 | 80 | 2.0 |
| | | | | 2 | 361 | 118 | 8 | 59 | 4.0 |
| | | | | 5 | 228 | 192 | 20 | 38 | 4.0 |
| | | | | 10 | 162 | 272 | 35 | 27 | 3.5 |
| | | | | 20 | 114 | 379 | 56 | 19 | 2.8 |
| | | | | 50 | 72 | 577 | 97 | 12 | 1.9 |
| | | | | 100 | 51 | 780 | 141 | 7.8 | 1.4 |
| | | | | 200 | 36 | 1041 | 203 | 5.2 | 1.0 |
| | | | | 500 | 23 | 1497 | 319 | 3.0 | 0.64 |
| | | | | 1000 | 16 | 1943 | 442 | 1.9 | 0.44 |
| | | | | 2000 | 11 | 2498 | 604 | 1.2 | 0.30 |
| | | | | 5000 | 7.2 | 3444 | 896 | 0.69 | 0.18 |
| 7200 | 6.0 | 3905 | 1041 | 0.54 | 0.14 | | | | |
| 60 | | | | | | | 255 | 22.3 | |
| 150 | 165 | 18.4 | 512 ² | 1 | 22 | 161 | 30 | 161 | 30 |
| | | | | 2 | 16 | 214 | 43 | 107 | 22 |
| | | | | 5 | 9.9 | 304 | 67 | 61 | 13.4 |
| | | | | 10 | 7.0 | 393 | 93 | 39 | 9.3 |
| | | | | 20 | 4.9 | 503 | 126 | 25 | 6.3 |
| | | | | 50 | 3.1 | 691 | 185 | 14 | 3.7 |
| | | | | 100 | 2.2 | 877 | 243 | 8.8 | 2.4 |
| | | | | 200 | 1.6 | 1118 | 317 | 5.6 | 1.6 |
| | | | | 500 | 0.99 | 1564 | 442 | 3.1 | 0.88 |
| | | | | 1000 | 0.70 | 2055 | 564 | 2.1 | 0.56 |
| 1800 | 0.52 | 2638 | 691 | 1.5 | 0.38 | | | | |
| 150 | | | | | | | 428 | 90.5 | |

Table 2.7: This table gives the expected source counts, $\langle N \rangle$, per integration time τ_{int} and number of measurements per second, $\langle m \rangle$, at the 5σ and 30σ detection levels, in a *single* Zenith Monitoring beam of assumed bandwidth of 4 MHz. This beam is one of the *seven* beams that make up the hexagonal pattern that scans the declination strip at the zenith. The last row per observing frequency gives the total number of measurements per second, for a *single* beam, to be stored during the time spent at a single field at both detection levels. Classical confusion limits arise in the Low and High Band when $\langle N \rangle$ is larger than 338 and 371, respectively.

2. Expected Data Rates and Volumes for the LOFAR Transients Key Project

| Freq. [MHz] | θ_{res} [arcsec] | FoV [deg ²] | Image [pixels] | τ_{int} [s] | ΔS [mJy] | $\langle N \rangle$ [FoV ⁻¹ τ_{int}^{-1}] | | $\langle m \rangle$ [s ⁻¹] | |
|----------------|-----------------------------------|----------------------------|-------------------|----------------------------|---------------------|--|-------------|---|-------------|
| | | | | | | 5 σ | 30 σ | 5 σ | 30 σ |
| 60 | 413 | 105 | 1024 ² | 1 | 510 | 80 | – | 80 | – |
| | | | | 2 | 361 | 119 | 8 | 60 | 4 |
| | | | | 5 | 228 | 193 | 20 | 39 | 4 |
| | | | | 10 | 161 | 272 | 35 | 27 | 3.5 |
| | | | | 20 | 114 | 380 | 56 | 19 | 2.8 |
| | | | | 50 | 72 | 577 | 97 | 12 | 1.9 |
| | | | | 100 | 51 | 781 | 142 | 7.8 | 1.4 |
| | | | | 200 | 36 | 1042 | 203 | 5.2 | 1.0 |
| | | | | 500 | 23 | 1498 | 319 | 3.0 | 0.64 |
| | | | 840 | 17.6 | 1823 | 408 | 2.2 | 0.49 | |
| 60 | | | | | | | 255 | 19.7 | |
| 150 | 165 | 18.4 | 1024 ² | 1 | 22 | 159 | 30 | 159 | 30 |
| | | | | 2 | 16 | 211 | 43 | 106 | 22 |
| | | | | 5 | 10 | 301 | 66 | 60 | 13 |
| | | | | 10 | 7.1 | 388 | 91 | 39 | 9.1 |
| | | | | 20 | 5.0 | 497 | 124 | 25 | 6.2 |
| | | | | 50 | 3.2 | 683 | 182 | 14 | 3.6 |
| | | | | 100 | 2.2 | 867 | 240 | 8.7 | 2.4 |
| | | | | 140 | 1.9 | 975 | 273 | 7.0 | 2.0 |
| | | | | 150 | | | | | |

Table 2.8: The RSM Rapid All-Sky Monitor mode. This table gives per integration time τ_{int} the expected source counts, $\langle N \rangle$, and measurements per second, $\langle m \rangle$ at the 5 σ and 30 σ detection levels, in a Rapid All-Sky Monitor beam of assumed bandwidth of 4 MHz. The last row per observing frequency gives the total number of measurements to be stored per second during the time spent at a single field at both detection levels. Classical confusion limits arise in the Low and High Band when $\langle N \rangle$ is larger than 338 and 371, respectively.

2.A Source Counts & Data Rates of LOFAR Modes

| Freq. [MHz] | θ_{res} [arcsec] | FoV [deg ²] | Image [pixels] | τ_{int} [s] | ΔS [mJy] | $\langle N \rangle$ | | $\langle m \rangle$ | |
|----------------|-----------------------------------|----------------------------|--------------------|----------------------------|---------------------|--|-------------|---------------------|-------------|
| | | | | | | $[\text{FoV}^{-1} \tau_{\text{int}}^{-1}]$ 5 σ | 30 σ | 5 σ | 30 σ |
| 15 | 41.3 | 1676 | 16384 ² | 1 | 5900 | 112 | – | 112 | – |
| | | | | 10 | 1900 | 835 | – | 84 | – |
| | | | | 100 | 590 | 3138 | 334 | 31 | 3.3 |
| | | | | 1000 | 190 | 9232 | 1547 | 9.2 | 1.5 |
| | | | | 10000 | 59 | 24287 | 5198 | 2.4 | 0.52 |
| 30 | 20.6 | 419 | 16384 ² | 1 | 1100 | 226 | – | 226 | – |
| | | | | 2 | 778 | 344 | 11 | 172 | 5.5 |
| | | | | 5 | 492 | 572 | 47 | 114 | 9.4 |
| | | | | 10 | 350 | 815 | 89 | 82 | 8.9 |
| | | | | 20 | 246 | 1156 | 152 | 58 | 7.6 |
| | | | | 50 | 156 | 1783 | 277 | 36 | 5.5 |
| | | | | 100 | 110 | 2437 | 415 | 24 | 4.2 |
| | | | | 200 | 78 | 3288 | 605 | 16 | 3.0 |
| | | | | 500 | 49 | 4789 | 967 | 9.6 | 1.9 |
| | | | | 1000 | 35 | 6246 | 1346 | 6.2 | 1.3 |
| | | | | 2000 | 25 | 8131 | 1871 | 4.1 | 0.94 |
| | | | | 5000 | 16 | 11292 | 2813 | 2.3 | 0.56 |
| | | | | 10000 | 11 | 14371 | 3772 | 1.4 | 0.38 |
| 45 | 13.8 | 186 | 16384 ² | 1 | 590 | 151 | 5 | 151 | 5 |
| | | | | 10 | 190 | 501 | 65 | 50 | 6.5 |
| | | | | 100 | 59 | 1450 | 266 | 15 | 2.7 |
| | | | | 1000 | 19 | 3542 | 811 | 3.5 | 0.81 |
| | | | | 10000 | 5.9 | 8057 | 2204 | 0.81 | 0.22 |
| 60 | 10.3 | 105 | 16384 ² | 1 | 390 | 109 | 6 | 109 | 6 |
| | | | | 2 | 276 | 158 | 14 | 79 | 7 |
| | | | | 5 | 174 | 252 | 31 | 50 | 6.2 |
| | | | | 10 | 120 | 362 | 52 | 36 | 5.2 |
| | | | | 20 | 87 | 487 | 78 | 24 | 3.9 |
| | | | | 50 | 55 | 730 | 130 | 15 | 2.6 |
| | | | | 100 | 39 | 978 | 188 | 9.8 | 1.9 |
| | | | | 200 | 28 | 1293 | 266 | 6.5 | 1.3 |
| | | | | 500 | 17 | 1836 | 412 | 3.7 | 0.82 |
| | | | | 1000 | 12 | 2411 | 578 | 2.4 | 0.58 |
| | | | | 2000 | 8.7 | 3021 | 764 | 1.5 | 0.38 |
| | | | | 5000 | 5.5 | 4148 | 1118 | 0.83 | 0.22 |
| | | | | 10000 | 3.9 | 5269 | 1469 | 0.53 | 0.15 |
| 75 | 8.25 | 67.0 | 16384 ² | 1 | 630 | 32 | – | 32 | – |
| | | | | 10 | 200 | 120 | 12 | 12 | 1.2 |
| | | | | 100 | 63 | 362 | 60 | 3.6 | 0.60 |
| | | | | 1000 | 20 | 938 | 199 | 0.94 | 0.20 |
| | | | | 10000 | 6.3 | 2168 | 562 | 0.22 | 0.056 |

Table 2.9: The LOFAR full Dutch array mode (Full-NL), 24 core and 16 remote stations, assuming a 4 MHz beam at each frequency. Beams from Low and High Band cannot operate together, but multiple beams in either the Low or High Band are possible. The rms sensitivities, ΔS , for a few integration times are shown. The number of sources, $\langle N \rangle$, likely to be detected in the corresponding 4 MHz beam is listed for 5 and 30 times the rms noise. The number of measurements to be stored per second, m , is reported in the last column, again specified for the two levels of significance.

2. Expected Data Rates and Volumes for the LOFAR Transients Key Project

| Freq. [MHz] | θ_{res} [arcsec] | FoV [deg ²] | Image [pixels] | τ_{int} [s] | ΔS [mJy] | $\langle N \rangle$ | | $\langle m \rangle$ | |
|----------------|-----------------------------------|----------------------------|--------------------|----------------------------|---------------------|---|-------------|---------------------|-------------|
| | | | | | | [FoV ⁻¹ τ_{int}^{-1}] | | [s ⁻¹] | |
| | | | | | | 5 σ | 30 σ | 5 σ | 30 σ |
| 120 | 5.16 | 16.2 | 16384 ² | 1 | 22 | 161 | 31 | 161 | 31 |
| | | | | 10 | 7.0 | 387 | 94 | 39 | 9.4 |
| | | | | 100 | 2.2 | 861 | 242 | 8.6 | 2.4 |
| | | | | 1000 | 0.70 | 2061 | 553 | 2.1 | 0.55 |
| | | | | 10000 | 0.22 | 6506 | 1242 | 0.65 | 0.12 |
| 150 | 4.13 | 10.3 | 16384 ² | 1 | 17 | 111 | 22 | 111 | 22 |
| | | | | 2 | 12 | 146 | 31 | 73 | 16 |
| | | | | 5 | 7.7 | 205 | 48 | 41 | 9.6 |
| | | | | 10 | 5.4 | 263 | 65 | 26.3 | 6.5 |
| | | | | 20 | 3.8 | 335 | 87 | 17 | 4.4 |
| | | | | 50 | 2.4 | 460 | 126 | 9.2 | 2.5 |
| | | | | 100 | 1.7 | 585 | 165 | 5.9 | 1.7 |
| | | | | 200 | 1.2 | 750 | 213 | 3.8 | 1.1 |
| | | | | 500 | 0.77 | 1064 | 295 | 2.1 | 0.59 |
| | | | | 1000 | 0.54 | 1429 | 375 | 1.4 | 0.38 |
| | | | | 2000 | 0.38 | 1953 | 477 | 0.98 | 0.24 |
| | | | | 5000 | 0.24 | 3121 | 657 | 0.62 | 0.13 |
| | | | | 10000 | 0.17 | 4615 | 847 | 0.46 | 0.085 |
| 180 | 3.44 | 7.18 | 16384 ² | 1 | 20 | 63 | 12 | 63 | 12 |
| | | | | 10 | 6.2 | 153 | 36 | 15 | 3.6 |
| | | | | 100 | 2.0 | 341 | 94 | 3.4 | 0.94 |
| | | | | 1000 | 0.62 | 797 | 219 | 0.80 | 0.22 |
| | | | | 10000 | 0.20 | 2397 | 488 | 0.24 | 0.049 |
| 210 | 2.95 | 5.28 | 16384 ² | 1 | 23 | 37 | 7 | 37 | 7 |
| | | | | 2 | 16 | 50 | 10 | 25 | 5 |
| | | | | 5 | 10 | 72 | 15 | 14 | 3 |
| | | | | 10 | 7.1 | 93 | 21 | 9.3 | 2.1 |
| | | | | 20 | 5.0 | 121 | 29 | 6.1 | 1.5 |
| | | | | 50 | 3.2 | 166 | 43 | 3.3 | 0.86 |
| | | | | 100 | 2.3 | 211 | 57 | 2.1 | 0.57 |
| | | | | 200 | 1.6 | 269 | 75 | 1.3 | 0.38 |
| | | | | 500 | 1.0 | 372 | 106 | 0.74 | 0.21 |
| | | | | 1000 | 0.71 | 482 | 136 | 0.48 | 0.14 |
| | | | | 2000 | 0.50 | 637 | 173 | 0.32 | 0.087 |
| | | | | 5000 | 0.32 | 958 | 237 | 0.19 | 0.047 |
| | | | | 10000 | 0.23 | 1353 | 301 | 0.135 | 0.030 |
| 240 | 2.58 | 4.04 | 16384 ² | 1 | 25 | 24 | 4 | 24 | 4 |
| | | | | 10 | 7.9 | 62 | 13 | 6.2 | 1.3 |
| | | | | 100 | 2.5 | 141 | 37 | 1.4 | 0.37 |
| | | | | 1000 | 0.79 | 318 | 90 | 0.32 | 0.090 |
| | | | | 10000 | 0.25 | 849 | 201 | 0.085 | 0.020 |

Table 2.10: Same as Table 2.9, but here specified for some of the High Band frequencies. The source count numbers at 120 and 150 MHz for the longest integration times have a higher uncertainty due to the extrapolations from the model from Huynh et al. (2005).

2.A Source Counts & Data Rates of LOFAR Modes

| Freq. [MHz] | θ_{res} [arcsec] | FoV [deg ²] | N_{point} | Image [pixels] | τ_{int} [s] | ΔS [mJy] | $\langle N \rangle$ | | $\langle m \rangle$ | |
|----------------|-----------------------------------|----------------------------|--------------------|-------------------|----------------------------|---------------------|---|------------|---------------------|------------|
| | | | | | | | [FoV ⁻¹ τ_{int}^{-1}] | | [s ⁻¹] | |
| | | | | | | | 5σ | 30σ | 5σ | 30σ |
| 60 | 413 | 105 | 619 | 512 ² | 1 | 255 | 164 | 17 | 164 | 17 |
| | | | | | 2 | 180 | 244 | 30 | 122 | 15 |
| | | | | | 5 | 114 | 380 | 56 | 76 | 11 |
| | | | | | 10 | 81 | 523 | 85 | 53 | 8.5 |
| | | | | | 20 | 57 | 710 | 126 | 36 | 6.3 |
| | | | | | 50 | 36 | 1043 | 203 | 21 | 4.1 |
| | | | | | 100 | 25 | 1375 | 287 | 14 | 2.9 |
| | | | | | 200 | 18 | 1792 | 399 | 9.0 | 2.0 |
| | | | | | 500 | 11 | 2502 | 606 | 5.0 | 1.2 |
| | | | | | 600 | 10 | 2670 | 656 | 4.5 | 1.1 |
| <hr/> | | | | | | | | 505 | 69.1 | |
| 150 | 165 | 18.4 | 3515 | 512 ² | 1 | 11 | 280 | 61 | 280 | 61 |
| | | | | | 2 | 7.8 | 362 | 84 | 181 | 42 |
| | | | | | 5 | 4.9 | 503 | 126 | 101 | 25 |
| | | | | | 10 | 3.5 | 640 | 168 | 64 | 17 |
| | | | | | 20 | 2.5 | 813 | 223 | 41 | 11 |
| | | | | | 50 | 1.6 | 1118 | 317 | 22 | 6.3 |
| | | | | | 100 | 1.1 | 1439 | 408 | 14 | 4.1 |
| | | | | | 200 | 0.78 | 1878 | 522 | 9.4 | 2.6 |
| | | | | | 500 | 0.49 | 2765 | 717 | 5.5 | 1.4 |
| | | | | | 600 | 0.45 | 3004 | 763 | 5.0 | 1.3 |
| <hr/> | | | | | | | | 723 | 172 | |

Table 2.11: The MSSS commensal mode. Characteristic parameters for the Million Sources Sky Survey carried out by 24 LOFAR core stations at the frequencies of 60 MHz and 150 MHz, assuming a single 16 MHz beam. Eventually, MSSS might be carried out with three simultaneous 16 MHz beams. Logarithmically spaced integration times τ_{int} are taken into account. N_{point} is the number of pointings needed to Nyquist sample the Northern Hemisphere for the given field of view. ΔS is the theoretical 1σ rms noise level. $\langle N \rangle$ is the number of sources likely to be detected in the MSSS beam per integration time τ_{int} , specified for 5σ and 30σ detections. The total number of measurements made per second, $\langle m \rangle$, is given in the last two columns per detection level. The last row per frequency gives the grand total, which is simply the sum of the measurements made at every integration time. To sample the whole sky with 619 pointings about 190×10^6 source measurements at 5σ are made during the MSSS Survey at 60 MHz, whereas at 150 MHz with ~ 3500 pointings about 1.5×10^9 measurements are made. For the 30σ levels, we expect 400 million measurements in total. Classical confusion limits arise in the Low and High Band when $\langle N \rangle$ is larger than 338 and 371, respectively.

2.B Benchmark Queries

```

INSERT INTO tempbasesources
(xtrsrc_id
,datapoints
,I_peak_sum
,I_peak_sq_sum
,weight_peak_sum
,weight_I_peak_sum
,weight_I_peak_sq_sum
)
SELECT b0.xtrsrc_id
,b0.datapoints
+ 1 AS datapoints
,b0.I_peak_sum
+ x0.I_peak AS i_peak_sum
,b0.I_peak_sq_sum
+ x0.I_peak * x0.I_peak AS i_peak_sq_sum
,b0.weight_peak_sum
+ 1 / (x0.I_peak_err * x0.I_peak_err) AS weight_peak_sum
,b0.weight_I_peak_sum
+ x0.I_peak / (x0.I_peak_err * x0.I_peak_err)
AS weight_i_peak_sum
,b0.weight_I_peak_sq_sum
+ x0.I_peak * x0.I_peak / (x0.I_peak_err * x0.I_peak_err)
AS weight_i_peak_sq_sum
FROM basesources b0
,extractedsources x0
WHERE x0.image_id = @imageid
AND b0.zone BETWEEN CAST(FLOOR((x0.decl - @theta) / x0.zoneheight
) AS INTEGER)
AND CAST(FLOOR((x0.decl + @theta) / x0.zoneheight
) AS INTEGER)
AND ASIN(SQRT((x0.x - b0.x)*(x0.x - b0.x)
+(x0.y - b0.y)*(x0.y - b0.y)
+(x0.z - b0.z)*(x0.z - b0.z)
) / 2
)
/
SQRT(x0.ra_err * x0.ra_err + b0.ra_err * b0.ra_err
+x0.decl_err * x0.decl_err + b0.decl_err * b0.decl_err)
< @assoc_r
;

UPDATE basesources
SET datapoints =
(SELECT datapoints
FROM tempbasesources
WHERE tempbasesources.xtrsrc_id = basesources.xtrsrc_id
)
,i_peak_sum =
(SELECT i_peak_sum

```

```
        FROM tempbasesources
        WHERE tempbasesources.xtrsrc_id = basesources.xtrsrc_id
    )
    ,i_peak_sq_sum =
    (SELECT i_peak_sq_sum
     FROM tempbasesources
     WHERE tempbasesources.xtrsrc_id = basesources.xtrsrc_id
    )
    ,weight_peak_sum =
    (SELECT weight_peak_sum
     FROM tempbasesources
     WHERE tempbasesources.xtrsrc_id = basesources.xtrsrc_id
    )
    ,weight_i_peak_sum =
    (SELECT weight_i_peak_sum
     FROM tempbasesources
     WHERE tempbasesources.xtrsrc_id = basesources.xtrsrc_id
    )
    ,weight_i_peak_sq_sum =
    (SELECT weight_i_peak_sq_sum
     FROM tempbasesources
     WHERE tempbasesources.xtrsrc_id = basesources.xtrsrc_id
    )
WHERE EXISTS (SELECT xtrsrc_id
             FROM tempbasesources
             WHERE tempbasesources.xtrsrc_id = basesources.xtrsrc_id
            )
;
```