

# Real-time Time-variability Analysis of GB to TB Datasets: Experience from SuperMacho and Supernova projects at NOAO/CTIO\*

R. Chris Smith<sup>a</sup>, Armin Rest<sup>b</sup>, Rafael Hiriart<sup>a</sup>, Andrew Becker<sup>c</sup>,  
Christopher Stubbs<sup>b</sup>, Kem Cook<sup>e</sup>, Gabe Proctor<sup>e</sup>, Sergei Nikolaev<sup>e</sup>,  
Stefan Kellar<sup>e</sup>, Frank Valdes<sup>d</sup>, Nick Suntzeff<sup>a</sup>, Doug Welch<sup>f</sup>

<sup>a</sup>NOAO/CTIO, Casilla 603, La Serena, Chile

<sup>b</sup>University of Washington, Dept. of Astronomy, Seattle, WA, USA

<sup>c</sup>Lucent Technologies, Bell Labs, 600 Mountain Ave., Murray Hill, NJ

<sup>d</sup>NOAO/Tucson, P.O. Box 26732, Tucson, AZ 85726, USA

<sup>e</sup>Lawrence Livermore National Laboratory, Livermore, CA 94550

<sup>f</sup>McMaster University, Hamilton, ON L8S 4M1, Canada

## ABSTRACT

The era of large survey datasets has arrived, and the era of large survey telescope projects is upon us. Many of these new telescope projects will not only produce large datasets, they will produce datasets that require **real-time** astronomical analysis, including object detection, photometry, and classification. These datasets promise to open new horizons in the exploration of the time domain in astrophysical systems on large scales. But to fulfill this promise, the projects must design and develop data management systems on a much larger scale (many Terabytes per day *continuously*) than has previously been achieved in astronomy. Working together, NOAO<sup>†</sup> and the University of Washington are developing prototype pipeline systems to explore the issues involved in real-time time-variability analysis. These efforts are not simply theoretical exercises, but rather are driven by NOAO Survey programs which are generating large data flows. Our survey projects provide a science-driven testbed of data management strategies needed for future initiatives such as the Large Synoptic Survey Telescope and other large-scale astronomical data production systems.

**Keywords:** time-domain, variability, pipeline, supernovae, microlensing

## 1. INTRODUCTION

Astronomical research is undergoing a significant change in the methods through which scientists obtain and examine data. While still a valid approach, the days (and nights) when the majority of research being done used data taken over the course of a few nights on a single large, classically scheduled telescope are giving way to research done with data from archives and databases. The archives of space-based missions such as HEASARC and MAST have led this change, but these are now giving way to even larger, and more importantly, homogeneous, datasets from survey projects such as SDSS, 2MASS, and DPOSS. The next wave of survey telescope projects, including VISTA, VST, SNAP, Pan-STARRS, and LSST, is in the planning, design, and development stage. These projects promise to take data rates from gigabytes (GBs) per night well into the terabyte (TB) range, and will ultimately produce datasets in the petabytes (PBs).

---

\* Copyright 2002 Society of Photo-Optical Instrumentation Engineers. This paper will be published in *Survey and Other Telescope Technologies and Discoveries*, J. A. Tyson and S. Wolff, eds., SPIE Vol. **4836**, and is made available as an electronic preprint with permission of SPIE. One print or electronic copy may be made for personal use only. Systematic or multiple reproduction, distribution to multiple locations via electronic or other means, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper are prohibited.

Further author information: Contact R.C.S. – E-mail: csmith@noao.edu, Telephone: 56 51 205200

<sup>†</sup>NOAO, the National Optical Astronomy Observatories, is operated by the Association of Universities for Research in Astronomy, Inc. (AURA) under cooperative agreement with the National Science Foundation.

An interesting new focus of several of these survey telescope projects, the synoptic survey telescopes or SSTs, is on the time domain (e.g., SNAP, Pan-STARRS, and LSST). By revisiting the same fields over and over again, these projects provide datasets ideal for studies of time-variable phenomena. However, to effectively study many of these phenomena, especially those which are variable over only one cycle such as supernovae (SNe), one has to detect them not in the archive but as they occur in order to allow for follow-up and analysis. While many previous and planned surveys deal with large datasets, the SSTs are generally planning on analyzing even larger datasets in *near-real-time* to provide the opportunity for additional follow-up, usually on different telescopes. This added requirement puts significantly more pressure on the design and implementation of the end-to-end data management system for these projects.

Such near-real-time time-domain analysis is by no means a completely new idea in astronomy. Over the past decade, many projects have been handling moderate (MB to GB per night) data rates searching for transient phenomena. Several groups have been using NOAO facilities in such a survey mode, reducing and analyzing Mosaic datasets in near-real-time, albeit usually over only periods of several nights. These include the Deep Lens Survey,<sup>1</sup> the High-z SN Search,<sup>2</sup> and the Nearby Galaxies Supernova Search,<sup>3</sup> to name just a few. In particular, these three projects share a common pipeline software heritage, although each has developed independent solutions to the individual challenges of each survey.

Several new NOAO Survey programs promise to push the data rates from total GBs per year to TBs per year. The NOAO Data Products Program (DPP),<sup>4</sup> together with colleagues in the University of Washington (UW) Department of Astronomy, has identified this challenge as an **opportunity** to begin exploring and developing the data management infrastructure necessary to support the TB per night data flows of the SSTs. In support of, at least initially, the SuperMacho and the “*w*” Supernova NOAO survey projects (SM+SN), the combined DPP/UW team is building upon the experiences of the above mentioned surveys, as well as experiences from the MACHO project, to produce a robust data reduction and transient analysis data management system. This pipeline system produces reduced data products and also measures time variability of millions of objects per night in near real time. In addition, it provides alerts on events for follow-up on other major facilities throughout the world, including Keck, Magellan, Gemini, and others. Although the data flow is currently only at the milli-LSST level, this system provides a *science-driven* sandbox to gain experience in the real-time analysis required for the data systems of the SSTs. Below we discuss the current state of our near-real-time data analysis system and some of our plans for future development.

## 2. SCIENCE DRIVERS

The requirements and design of the planned survey telescopes will be (or at least *should be*) driven by the science which astronomers hope to extract from the vast datasets which will be produced. This axiom applies even more specifically to the data management systems of these projects, as well as to any precursor or prototype projects. The DPP/UW collaboration is motivated by the scientific needs of the NOAO Survey programs, in particular two which have been allocated significant blocks of 4m telescope time for transient detections: the SuperMacho project which will probe the dark matter of the universe and the “*w*” Supernova project designed to probe the dark energy’s equation of state.

### 2.1. Probing Dark Matter with SuperMacho

One of the foremost outstanding problems in the physical sciences is the nature and distribution of the “dark matter” that is the gravitationally dominant component of the mass in all galaxies, including our own Milky Way. One way to search for a class of astrophysical dark matter objects called MAssive Compact Halo Objects, or MACHOs, is to search for the transient brightening of background stars due to the gravitational lensing by foreground MACHOs. This “microlensing” signature has been detected by several projects,<sup>5-7</sup> but the nature of the lensing objects remains a mystery. The SuperMacho project is designed to detect an order of magnitude more microlensing events toward the Large Magellanic Cloud than previous surveys ( $\sim 12$  per year over five years), which will both (a) move the analysis out of the realm of small number statistics and (b) provide a greater number of “exotic” events (e.g., binary lenses<sup>8</sup>) which can lift the degeneracy between the mass, location, and velocity of the lens.

## 2.2. Probing Dark Energy with Supernovae

Perhaps the most surprising cosmological result in the last five years has been the credible evidence for cosmic acceleration<sup>9,10</sup> based on observations of high-redshift Type Ia supernovae. These observations indicated that the SNe were about 0.25 magnitudes fainter than they would have been in a matter-dominated universe with  $\Omega_M = 0.3$  and no cosmological constant (see Leibundgut 2001<sup>11</sup> for review). Faint SNe imply that the expansion of the universe from the time of the Big Bang to the present is larger than in any decelerating model. This requires *acceleration* which could have its origin in the negative pressure associated with “dark energy.” In order to determine the properties of this dark energy, and in particular how its effects have varied with cosmic epoch, the “*w*” Supernova project aims to discover and follow  $\sim 200$  Type Ia SNe distributed evenly over the redshift range [0.15, 0.75] during the five year lifetime of the project.

## 2.3. Science-driven Requirements

While these two projects seek widely differing scientific goals, the observing requirements for both are extremely similar due to the fact that both seek to detect and follow faint transient sources with significant variability over timescales of days. To achieve this, both are designed around an observing cadence of a half night every other night. In addition, both projects have selected the NOAO/CTIO Blanco 4m telescope combined with the Mosaic 8Kx8K camera to achieve the photometric depth and sky coverage necessary for each project. These commonalities have allowed us to merge the data management efforts into one SM+SN reduction and transient detection pipeline system. Indeed, it is this sort of synergy of differing science drivers satisfied by common observational parameters which provides the foundation upon which the SSTs can build their science cases.

The design of data management system for the SM+SN projects is constrained by the data rates and the need for rapid transient detection. During one half-night’s observations, each project produces roughly 10 GB of raw data, giving a total of  $\sim 20$  GB per combined SM+SN night. Both projects observe every other night during dark time ( $\sim \pm 10$  nights from new moon) for three consecutive months, creating a total of 0.6 TBs of raw data per year. Both projects must process this data and automatically detect faint transient sources on complicated backgrounds. These detections must be cataloged, matched against previously known variable sources, and if new, announcements must be generated.

The timescale for these announcements is one of the driving issues of the data management system design. Although the typical LMC microlensing event to be studied by SuperMacho lasts of order 80 days, the system must be prepared to alert on much shorter timescales, within a day or two of an event detection, to trigger studies of the important exotic events. Supernova have a much shorter rise time (usually  $\sim 10$  days from detection), and spectroscopy of these objects during the rise can be critical to the classification of relatively uncommon types of SNe Ia such as the 1991T-like events.<sup>12</sup> Given these needs, the transient analysis system should produce announcements, or at least events to be reviewed, within at most 12 hours of the observations to allow for the possibility of follow-up observations on other telescopes the subsequent night.

## 3. THE SM+SN PIPELINE

These requirements have been translated into a data management system which had its first beta testing in October 2001 at CTIO during the first run of the SuperMacho project. It is currently implemented in a combination of IRAF<sup>‡</sup> routines, C code, Perl and Python scripting tied together to provide an integrated but modular environment. The design can be broken down into three major sections: data reduction, transient detection, and data products. The actual data processing is broken down into basic units of “actions” (e.g., flat fielding), which are organized into “stages” (e.g., aligning images). Actions are generally stand alone Perl scripts or C-executables which are called from the pipeline in a standardized way. This modular structure allows the actions to be replaced or upgraded without changing the functional stages of the pipeline, and the functional stages can be changed to accommodate different uses of the pipeline (e.g. direct photometry instead of image differencing).

---

<sup>‡</sup>IRAF, the Image Reduction and Analysis Facility, is distributed by the National Optical Astronomy Observatories

To meet the specification of producing detections within 12 hours, we have emphasized not only the speed of the system (which can, to some extent, be solved by hardware), but also the degree to which the system is robust to failures. The design includes thorough bookkeeping at the stage level, allowing it to recover where it left off in the event of computer malfunction and/or allowing the user to re-run any given stage (and all subsequent stages, if desired) in the event of problems in the results of processing. All of this bookkeeping is currently file based, with input lists and output lists driving the system. These lists are automatically generated by action modules which look for new data as it is written to disk from either the data acquisition system (in the case of the first stage) or previous stages.

### **3.1. Data Reduction and Quality Assessment**

Our initial stage of processing involves taking images as they are produced by the Mosaic 8Kx8K camera and removing the instrumental signatures. The first steps rely upon the tools developed at NOAO for reduction of Mosaic data in IRAF, including the correction for crosstalk (xtalkcor) and astrometric WCS calibration (msccmatch). These are called from Perl wrapper scripts (these scripts being the “action” components) to provide integration into the pipeline system. Additional image reduction action components are written in C for speed. These include overscan, trim, and bias subtraction to remove bias levels and structures, and flat-field calibration. The necessary calibration files (e.g., combined bias frames and flat fields) are also automatically generated from the incoming data stream based on FITS header keywords. During object frame processing the pipeline looks for a set of previously made standard calibrations (made, for example, by a previously run process) or initiates an action to generate the needed calibrations.

While the basic components necessary to deliver data to the following transient detection portion of the pipeline have been integrated as described, there are still several areas in which this initial portion of system is under active development. At UW we are actively pursuing the propagation of noise arrays through the pipeline as a whole, starting with the initial reduction stages and carrying them through transient detection to provide true statistical measures of the significance of our detections (as well as better filtering of false positive detections). At NOAO we are pursuing adding corrections for other instrumental artifacts, such as fringing (not an issue in the VR-band SuperMacho data, but definitely a problem for the I-band supernova data) and removal of pupil ghosts.

An additional area of joint development is the process of automated data quality assessment. Our pipeline currently has very little integrated “data evaluation” to automatically detect problematic or bad data. This is not only an issue for the pipeline system, but also for the larger observing system, in that the sooner problems in the data flow are detected, the sooner they can be rectified at the telescope or instrument level if necessary. We are developing strategies to provide automated data quality assessment, including detector signatures such as noise characteristics, PSF variations (e.g., due to focus), transparency, etc. which will provide feedback to the observer in near-real-time.

### **3.2. Transient Detection and Analysis**

#### **3.2.1. Design Drivers**

Standard photometry of transient or variable objects becomes very inefficient in images with variable background light (e.g., SNe in galaxies) or crowded fields (e.g., variables in the bar of the LMC). Moreover, the crowding makes it very difficult to obtain accurate photometry since the standard technique of fitting multiple profiles only works sufficiently well if the point-spread function is very well known and all other close objects are modelled very carefully.

The MACHO project used a customized version of DoPHOT as its object detection software in an attempt to increase the detection efficiency of microlensing in crowded fields. In their approach, the objects in a given field are fixed by defining a list of objects based on a deep, good quality image. In subsequent images, only those objects are monitored by forcing DoPHOT to fit only the known objects. The centroid positions are also not allowed to move. The downside to this approach is that objects which are intrinsically too faint to be detected are not monitored directly even if they could be detected after they are magnified by microlensing. In this case, the added flux introduces a deviation into the flux profile of neighboring object(s). So even though

the microlensed star is not monitored directly, the microlensing will leave some signature in the lightcurves of neighboring objects. However, even in these cases the efficiency is limited since the S/N of the signal will be diminished, especially if the flux of the microlensed star is significantly smaller than the flux of the neighboring objects (which might be expected in most cases).

An alternative solution to the detection of transient or variable objects is a method called difference image analysis (DIA) which has rapidly evolved in the last few years. The first implementation was done by Phillips & Davis,<sup>13</sup> who introduced a method that registered images, matched the point spread function (PSF), and matched the flux of objects in order to detect transients. Tomaney & Crotts<sup>14</sup> expanded the method by calculating a convolution kernel in Fourier space in order to match the point spread function (PSF) of the images. A more efficient way of finding the kernel which utilizes a simple least-square analysis using many pixels from both images, termed Optimal Image Subtraction (OIS), was described by Alard & Lupton.<sup>15</sup> Derivatives of DIA and more recently OIS have been widely applied in various projects (MACHO<sup>16</sup>; M31 microlensing<sup>17</sup>; OGLE<sup>18</sup>)

Precisely matching the PSFs in different images is easy in principle but difficult in practice. The biggest intrinsic problem is that the signal of genuine variability may be swamped in frames for which the PSF is not well matched. In such frames, the residuals of badly subtracted bright stars can mimic the variability signal of faint stars. This can be due to improperly aligned images, atmospheric diffraction effects or simply that the fitted kernel is not perfect. Also, saturated stars and bad pixels can cause residuals in the image. The challenge is to identify and reject the false positives without diminishing the efficiency of detecting genuine variability.

### 3.2.2. Implementation

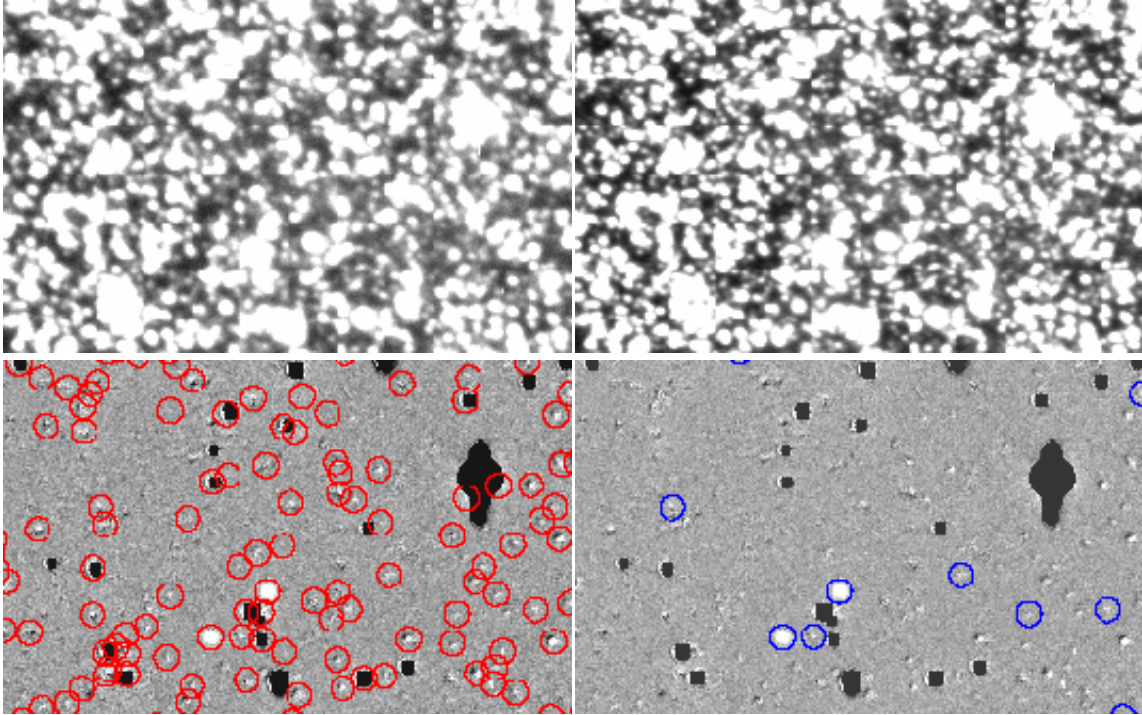
The primary path in the SM+SN transient detection system is defined to be through difference imaging. Optional stages have been implemented to perform the MACHO-style “absolute object” detection, possibly to be run in parallel with the difference imaging stages but at lower priority.

One of the main problems with the image differencing approach is that there are more residuals than genuinely variable objects in the difference image. Therefore standard profile-fitting software like DoPHOT has problems determining the proper PSF to use in the difference image. We use the fact that the original, flattened image has already been analyzed with DoPHOT before subtraction to save the PSF and carry that information forward to the difference image. When the difference image is analyzed with our customized version of DoPHOT, we force the PSF to be the one determined beforehand. Applying this a priori knowledge of the PSF helps to guard against bright false positives such as cosmic rays.

The majority of false positives, however, are the faint residuals of cores of stars. In general, they look like “dipoles”, i.e. they have positive and negative flux, concentrated into two neighbouring areas (see Figure 1). We reject such false positives with the following statistical method: For a given candidate object, we select all pixels within an aperture (determined by the seeing) that have flux either bigger than  $F_{min} = N\sigma_{bkg}$  or smaller than  $-F_{min}$ , where  $\sigma_{bkg}$  is the average deviation from mean in the difference image and  $N$  is a user-defined number. By comparing the number of pixels with positive flux and their cumulative flux to the ones with negative flux, one can identify and reject the residuals with dipolar properties.

In the lower panel of Figure 1 objects initially identified with our DoPHOT version are marked with red circles. Clearly, we have configured DoPHOT so that it triggers on nearly everything in the image since we don’t want to miss any objects with genuine variability. In this particular image of our most crowded field, we find nearly 8000 possible detections, for which the vast majority are false positives (see solid line histogram in left panel Figure 2). Only about 1000 objects pass our cuts (see red circles in lower panel of Figure 1, and dotted histogram in left panel Figure 2). The biggest fraction of these are still false detections, but it should be noted that most of the false positives have a small S/N ratio between 2 and 5. Any object that passes the cuts with a S/N larger than 5 has a high probability of genuine variability. However, with coincidence of detections in images taken at other epochs, we can even detect variability down to the lowest S/N.

To study our completeness and examine how the cuts affect the detection of genuine objects, we have added fake stars to the image and run it through exactly the same reduction and analysis process. In the middle panel of Figure 2, the completeness of detected fake stars is shown for no cuts (solid blue line) and our cuts



**Figure 1.** The upper two panels show example stamps of an image (left) and the associated template (right). The lower left panel shows the corresponding difference image, with the objects identified with DoPHOT marked with red circles. Objects that passed our cuts are indicated with blue circles in the lower right panel. Note the two genuine variable objects close to the center of the stamps. The cores of saturated stars are masked out (black spots).

(dotted blue line) applied. A few real objects are indeed rejected, but only very small fraction considering the advantage of cutting down the number of false detection by nearly an order of magnitude. We are in the process of adjusting the various cut parameters to optimize this rejection algorithm.

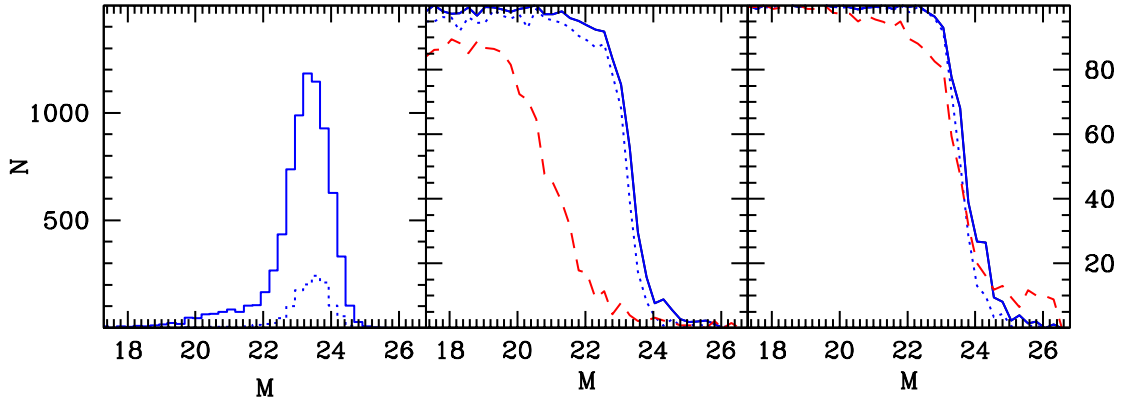
The middle panel of Figure 2 also shows how the DIA completeness fares compared to the completeness of standard photometry (red, dashed line). This figure implies that it is possible to detect significantly fainter variability with difference imaging in a crowded field like this, although a full MACHO-style analysis of the neighboring objects would recover some fraction of the test objects. For images that are not as crowded (see right panel of Figure 2), the completenesses are comparable.

### 3.3. Database

The open source database PostgreSQL has been selected to support the needs of the data management of the SM+SN projects. These needs can be divided in three broad categories: support for the survey operations, analysis of the detections found by the pipeline, and communication of the results to the project users and the astronomical community.

As the survey is performed, a considerable amount of information is generated as the observations take place. This information includes: observations or pointings, area of sky covered, images produced, related calibration frames, reduced images, and template images for image subtraction. The attributes that define these entities, as well as the relationships that these entities have with each other, are stored in the database using the structures provided by current relational database technologies.

As a result of the transient analysis, a set of detections is generated for each processed image. These detections, including positions and all additional information produced by the analysis, are stored in the database,



**Figure 2.** All x-axis are in approximate magnitudes  $M_{VR}$  (since SuperMacho uses a specialized broad band “VR” filter, the magnitude is approximately between V and R). The right panel shows a histogram of all objects detected with DoPHOT on the difference image (solid line) and objects that passed the cuts (dotted blue line). The middle (crowded field) and right (sparse field) panel compare the completeness of difference image analysis (DIA) without cuts (solid blue line) and with cuts (dotted blue line). The difference between these is our incompleteness. The red dashed line in both of the middle and right panels indicates the completeness of standard DoPHOT photometry methods in detection of new variable sources, although no attempt to measure the effect of the new source on adjacent objects was made.

together with the relationships between the detections and the images and observations they come from. Because of the inherent precision of the observations and numerical analysis performed by the pipeline, two detections which were presumably generated by the same source at different epochs don’t necessarily share the same exact spatial coordinates, although they are very close. As a result, the detections need to be aggregated into *clusters* around the same spatial coordinates before the extraction of lightcurves and object classification. A state of the art clustering algorithm, OPTICS,<sup>19</sup> has been coded to be applied to the detections, creating additional entries and relationships in the database. This algorithm was selected because of its scalability with the number of detections as well as its ability to find subclusters in a given cluster of detections. These clusters or aggregations of detections are then classified as different types of astronomical objects, currently through human interaction.

A number of additional attributes and relations are associated with the object tables in the database, including follow-up photometric data (both from re-analysis of the survey data and from follow-up imaging on other telescopes), spectroscopic data, and relationships with external catalogs. Modern database systems such as PostgreSQL allow databases to grow, so the SM+SN database can expand as additional attributes and relations are defined to support the analysis of the dataset, both for the goals of the projects and possibly future, non-SM or SN applications.

### 3.4. Data Products

The end products of the SM+SN pipeline system, to zeroth order, are the new transient sources discovered and the measurements of their characteristics. The database is the principal repository for this information, linking multiple detections into objects and storing ancillary information such as classification, photometric information, spectral information, and links to external references. Much of the analysis of both the SuperMacho and Supernova datasets will be enabled simply through database queries accessing this information.

Detailed analysis of any given event, however, will almost certainly require access to the images in some form. In order to support analysis at various levels, many different intermediate data products must be retained. Indeed, the data processing through transient detection is not “data reduction” but instead “data expansion”! In addition to the noise arrays generated to allow detailed analysis of the uncertainties in the detections and

photometric measurements, the data management system generates *and keeps* reduced data (with instrumental signatures removed), remapped aligned images, convolved images, and difference images, along with all of the associated log information that describes these images. Eventually our data management system may include methods of regenerating some of these data products “on demand” to decrease the storage requirements.

The accumulated dataset also provides the astronomical community with the opportunity to use the data for science other than microlensing and supernova detections. Both projects have waived all proprietary rights to the data, so the raw and reduced images become public as soon as they are written out, either to the disks at the end of the SM+SN pipeline system or directly out to the NOAO Science Archive after image reduction. The public data products additionally include the near-real-time announcement of transient sources and finder charts for these sources.

To effectively communicate the results of the transient analysis, we are developing a web site which is integrated with the PostgreSQL database through PHP, Perl, and Python. The delivery of data products to the community will eventually make use of XML-based technologies as these evolve to meet the needs of the surveys and the astronomical community.

#### 4. HARDWARE

Although we’ve optimized the software described above somewhat for speed considerations, as pointed out in section 3, many of the speed concerns can be addressed by computing hardware and the use of parallel processing. The resulting data management hardware is summarized in Figure 3, in which the data management system hardware components are labeled with their general functions.

We have standardized on commodity PC hardware running the Linux operating system (currently Red Hat 7.x). Based partially on our own experience and partially on anecdotal information, we have favored Athlon CPUs for the data reduction nodes and Pentium CPUs for the disk servers. Similarly, in areas where access speed and reliability is a major issue, we’ve used SCSI disks, while we’ve used less expensive IDE disks for on-line storage areas (which are usually cumulative).

The raw images are initially written out in MEF (multiple-extension FITS) format on the disks of the data acquisition system, which is a dual-CPU Linux PC. These images are automatically written out to Exabyte tapes via the standard NOAO Save-the-Bits system, and will soon also be written out to DVD. From the data acquisition machine the raw images are pulled over to the first stage of the SM+SN pipeline, which is currently a single dual-CPU 1.2GHz machine. The link between the data acquisition machine and this first SM+SN pipeline machine has in the past been a standard 100Mbps link. However starting in late 2002 the whole pipeline system will be moved off the mountain to ease computer and user support, so the connection will be limited to CTIO’s mountain-top to downtown link speed of 77Mbps. With some compression, this limited bandwidth does not pose a significant bottleneck to the data management system. Once the data arrive on the initial SM+SN machine, all networking through to the final server is done over a dedicated 1Gbps ethernet backbone.

As soon as the images are reduced by the single node reduction stage, they are passed to the heart of the SM+SN pipeline system, the transient detection and analysis cluster. This is composed of a high-throughput 2.2GHz Pentium III SCSI disk server with a one TB SCSI disk array for “scratch space” connected to eight dual-CPU 1.2GHz Athlon PCs. In the current implementation, with the file-based bookkeeping, all nodes use the SCSI disk array as their primary workspace. This has the advantage of providing CPU independence (any cluster CPU can operate on any data) but necessitates a high-throughput NFS server to prevent any bottleneck in disk access.

After the processing, and results of data expansion discussed above, all of the images and associated information are transferred to the final component of the SM+SN data management system, the database and interactive server (a hardware duplicate of the SCSI disk server above), and stored on IDE RAID arrays. This system hosts both the SM+SN SQL database server, the database(s), and multiple 2TB IDE RAID disk arrays. Due to the data expansion, each year’s worth of data fills roughly 2TB of disk space, so we plan for one 2TB disk array per year over the lifetime of the projects.

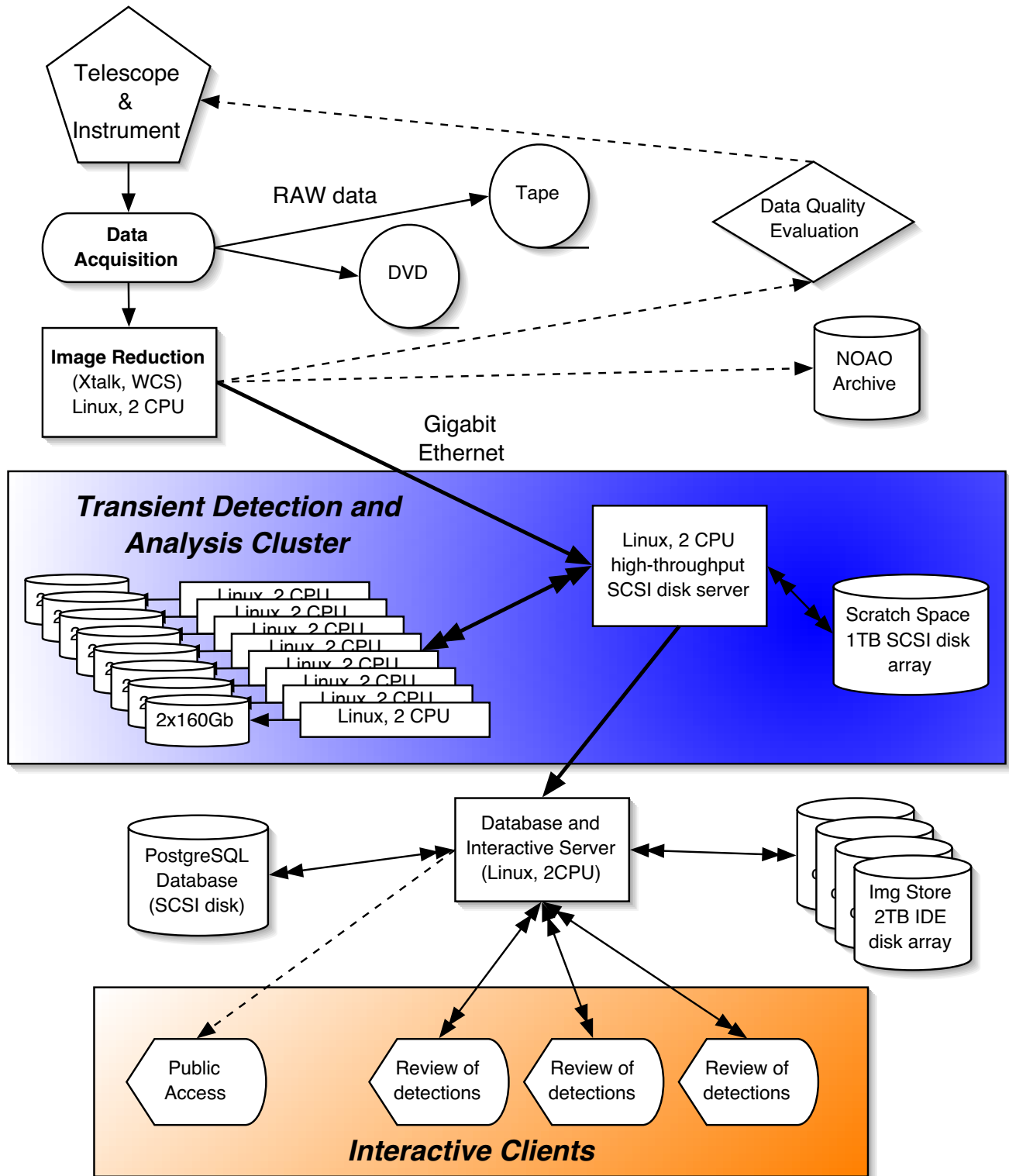


Figure 3. The SM+SN data management system from the hardware perspective.

## 5. PARALLIZATION

The processing of data from almost any mosaic CCD camera is a naturally parallelizable computational task, since the data are broken down into separate image “units” corresponding to the individual CCDs or amplifiers used. In the case of the 8Kx8K NOAO Mosaic camera at CTIO, the focal plane is covered by eight CCDs, each of which is read through two amplifiers, giving 16 semi-independent units which make up each image (or observation).

There are, however, certain aspects of the data processing which are better done or must be done on the whole image because the units are not completely independent. The correction for crosstalk between the amplifiers in the NOAO Mosaic camera is one example which involves interaction between multiple units of the larger image. Correction for the pupil ghost, which is a large-scale structure crossing many units, is another example. Depending on the density of reference stars for astrometric calibration, the (re)mapping of the WCS may also be a task which is better done at the whole image level.

In the SM+SN pipeline implementation, we have currently separated these multi-unit tasks out to be performed on the initial SM+SN system machine. We currently do only crosstalk and astrometric calibration at this stage, which takes  $\sim 200$ s per Mosaic image. Since our average observing cadence is greater than 200s, this does not result in a bottleneck in the data management system. It is however a point of concern as we add more features to the pre-transient analysis system such as pupil ghost correction and data quality analysis. Additional CPUs and careful planning will be necessary to prevent this stage from limiting the throughput of the system below acceptable levels.

The rest of the image reduction, as well as all of the transient analysis, does break down naturally into the 16 now independent units, and can therefore be efficiently handled in parallel. The eight dual-CPU units are available for this parallel processing. Using all of the processors, a complete Mosaic image can be processed through the transient portion of the system in  $\sim 225$ s, fast enough to keep up with the observing cadence of the SM+SN programs.

While the transient analysis of the 16 image units can logically be matched to 16 individual CPUs in the cluster, such a static mapping is not necessarily ideal. In particular, it is not robust against computer failures. In order to provide a greater degree of system reliability, we have used Condor to manage and optimize the parallel processing. The modular design of the pipeline system made the transition from the original linear processing strategy, where each image unit was processed independently but serially, to a Condor-controlled architecture relatively easy. The pipeline system management task now submits each set of 16 parallel transient analysis tasks to the Condor server, which distributes the processes across the available CPUs.

## 6. SUMMARY

The coming generation of large-scale synoptic survey telescope projects will produce data flows in the range of many TBs per night, requiring a significant leap in astronomical data management systems beyond anything currently in use. Smaller scale programs, based on scientifically motivated observations, can begin to explore the issues involved in near-real-time data management today. The SuperMacho and “*w*” Supernova projects together provide an excellent test-bed for such exploration.

We have designed and are implementing an end-to-end data management system which will efficiently handle the data flow from the SM+SN projects. The modular design facilitates incrementally upgrading the system to accommodate different datasets (for example, different Mosaic cameras) and/or different scientific analysis. While we have identified a subset of operations which are more difficult to run in parallel, the majority of the data processing is executed in parallel, providing for a great deal of scalability in the data management system.

In this context, we are learning that the data rates of the SST projects are not necessarily overwhelming. They are manageable, even given today’s technology. This is not to say that the data management at TB/night rates is a solved problem. Based on our experience with our “sand-box” data management system, we have identified from several outstanding areas for further investigation, including:

- the break down or optimization of processing stages which require information from multiple image units in a way such that those stages do not become a bottleneck
- automated object classification
- strategies for visualization of time-domain data products, not limited to light curves of individual objects

We are turning our efforts to these and other issues of data management as we execute the SM+SN program in pursuit of the better understanding of the dark matter and dark energy in our universe.

## REFERENCES

1. D. Wittman, J. A. Tyson, I. P. Dell'Antonio, A. C. Becker, V. Margoniner, and G. Wilson, "The Deep Lens Survey," in *Astronomy with Large Telescopes: Survey and Other Telescope Technologies and Discoveries*, J. A. Tyson and S. Wolff, eds., *Proc. SPIE* **4836**, in press, 2002.
2. B. Schmidt and et al., "The High-Z Supernova Search: Measuring Cosmic Deceleration and Global Curvature of the Universe Using Type IA Supernovae," *Astrophysical Journal* **507**, p. 46, 1998.
3. L. Strolger, C. Smith, et al., "The Nearby Galaxies Supernova Search program," *in preparation*, 2002.
4. D. Shaw, T. Boroson, and C. Smith, "The NOAO Data Products Program," in *Observatory Operations to Optimize Scientific Return*, P. J. Quinn et al., eds., *Proc. SPIE* **4844**, in press, 2002.
5. C. Alcock et al., "The MACHO Project: Microlensing Results from 5.7 Years of Large Magellanic Cloud Observations," *Astrophysical Journal* **542**, p. 281, 2000.
6. A. Udalski, M. Kubiak, and M. Szymanski, "Optical Gravitational Lensing Experiment. OGLE-2 – the Second Phase of the OGLE Project," *Acta Astronomica* **47**, p. 319, 1997.
7. C. Renault et al., "Observational limits on MACHOS in the Galactic Halo," *Astronomy and Astrophysics* **324**, pp. L69, 1997.
8. C. Alcock et al., "Candidate Binary Microlensing Events from the MACHO Project," *Bulletin of the American Astronomical Society* **30**, p. 1415, 1998.
9. A. G. Riess et al., "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant," *Astronomical Journal* **116**, p. 1009, 1998.
10. S. Perlmutter et al., "Measurements of Omega and Lambda from 42 High-Redshift Supernovae," *Astrophysical Journal* **517**, p. 565, 1999.
11. B. Leibundgut, "Cosmological Implications from Observations of Type Ia Supernovae," *Annual Reviews of Astronomy and Astrophysics* **39**, p. 67, 2001.
12. L. Strolger, C. Smith, et al., "Template Lightcurve Comparisons of 91T<sub>aa</sub> and Normal Type Ia Supernovae," *in preparation*, 2003.
13. A. C. Phillips and L. E. Davis, "Registering, PSF-Matching and Intensity-Matching Images in IRAF," in *ASP Conf. Series: Astronomical Data Analysis Software and Systems IV*, R. A. Shaw, H. E. Payne, and J. J. E. Hayes, eds., **77** **4**, p. 297, 1995.
14. A. B. Tomaney and A. P. S. Crotts, "Expanding the Realm of Microlensing Surveys with Difference Image Photometry," *Astronomical Journal* **112**, p. 2872, 1996.
15. C. Alard and R. H. Lupton, "A Method for Optimal Image Subtraction," *Astrophysical Journal* **503**, p. 325, 1998.
16. C. Alcock et al., "Difference Image Analysis of Galactic Microlensing. II. Microlensing Events," *Astrophysical Journal Supplement* **124**, p. 171, 1999.
17. A. P. S. Crotts, R. Ugesich, G. Gyuk, and A. B. Tomaney, "MEGA: Mapping Halo and Bulge Microlensing in M31," in *ASP Conference Series vol. 182*, D. R. Merritt, M. Valluri, and J. A. Sellwood, eds., **182** **1**, p. 409, 1999.
18. P. R. Wozniak, "Difference Image Analysis of the OGLE-II Bulge Data. I. The Method," *Acta Astronomica* **50**, p. 451, 2000.
19. M. Ankerst, M. M. Breuning, H.-P. Kriegel, and J. Sander, "OPTICS: Ordering Points To Identify the Clustering Structure," in *Proc. 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99)*, 1999.