SETI@home: Internet Distributed Computing for SETI

D. P. Anderson, D. Werthimer, J. Cobb, E. Korpela, M. Lebofsky

Space Sciences Laboratory, University of California, Berkeley, California

D. Gedye

Vulcan Northwest, Inc.

W. Sullivan

Department of Astronomy, University of Washington

Abstract. SETI@home is a radio SETI project that does its primary signal analysis using Internet-connected computers. In the first five months of operation of SETI@home, a million people have participated, and have contributed 100,000 years of computer time. The use of distributed computing limits frequency coverage but allows greater sensitivity and generality in the signal analysis.

1. Introduction

SETI@home is a radio SETI sky survey which, like SERENDIP IV (Werthimer, et al. 1997), gets data from a "piggyback" receiver at the Arecibo radio telescope. Whereas SERENDIP analyzes this data primarily using a special-purpose supercomputer located at the telescope, SETI@home distributes the data through the Internet to hundreds of thousands of personal computers. This approach provides a tremendous amount of computing power but limits the amount of data that can be handled. Hence SETI@home covers a relatively narrow frequency range (2.5 MHz) but searches for a wider range of signal types, and with better sensitivity, than other SETI sky surveys to date.

SETI@home was launched on May 17, 1999. In its first five months, it has attracted over a million participants. Together they have contributed over 100,000 years of computer time, making SETI@home the largest computation ever performed.

2. Science Design

SETI@home is a SETI sky survey at the National Astronomy and Ionospheric Center's 305 meter radio telescope in Arecibo, Puerto Rico. It shares the piggyback receiver used by SERENDIP IV (Werthimer et al. 1997), but its search space is roughly orthogonal to that of SERENDIP IV; although SETI@home has

1/40 the frequency coverage of SERENDIP IV, its sensitivity is ten times better. The SETI@home search also covers a richer variety of signal bandwidths, drift rates, and time scales than SERENDIP IV or any other SETI program to date.

Primary data analysis, done using distributed computing, computes power spectra and searches for "candidate" signals such as spikes, Gaussians, and other signal types. Secondary analysis, done on the project's own computers, rejects RFI and searches for repeated events within the database of candidate signals.

2.1. Receiver and Data Recording

SETI@home uses a dedicated flat feed and cryogenic receiver mounted on the carriage house of the Arecibo telescope. The feed provides a single linear polarization with a gain of 3K/Jy and a 0.1 degree beam width. System temperature is 45K. The SETI@home survey covers 28% of the sky (declinations ranging from +1 to +35 degrees) with a sensitivity of 3E-25 W/m2. SETI@home observations will span a total of two years, during which most of the sky will be observed two or three times. Observations began in October 1998.

SETI@home covers a 2.5 MHz bandwidth centered at the 1420 MHz Hydrogen line. The receiver output is down-converted with quadrature analog mixers and filters, then digitized and converted to baseband by a digital quadrature mixer and a pair of 256 tap finite impulse response low pass filters. The resulting 2.5 MHz band is recorded continuously on 35 Gbyte DLT IV tapes with two bit complex sampling, along with data on telescope coordinates, time and engineering monitors. Tapes are mailed to UC Berkeley for analysis; the complete sky survey requires 1100 tapes to record a total of 39 terabytes of data.

We expect to record high quality data 65% of the time, observing each of the one million beams two or three times during the two year program. It is important to observe each beam several times because sources may scintillate (Cordes, 1991) or have short duty cycles, and most of our robust detection algorithms require multiple detections. SETI@home is able to collect useful data whenever the telescope is stationary or the Gregorian feed is tracking a source. When the Gregorian system tracks a source, the SETI@home feed is moving at 1 to 2 times sidereal rate on the sky and a source remains in the beam for 12 to 24 seconds. When the telescope is stationary, a source is in the beam for 24 seconds.

2.2. Primary Data Analysis

SETI@home data tapes from the Arecibo telescope are divided into small "work units" as follows: the 2.5 MHz bandwidth data is first divided into 256 subbands by means of a 2048 point fast Fourier transform (FFT) and 256 eight point inverse transforms. Each work unit consists of 107 seconds of data from a given 9,765 Hz sub-band. Work units are then sent over the Internet to the client programs for the primary data analysis.

Because an extraterrestrial civilization's signal has unknown bandwidth and time scale, the client software searches for signals at 15 octave spaced bandwidths ranging from 0.075 Hz to 1220 Hz, and time scales from 0.8 mSec to 13.4 seconds. The rest frame of the transmitter is also unknown (it may be on a planet that is rotating and revolving), so extraterrestrial signals are likely to be drifting in frequency with respect to the observatory's topocentric reference frame.

Because the reference frame is unknown, the client software examines 6761 different Doppler acceleration frames of rest (dubbed "chirp rates"), ranging from -10 Hz/sec to +10 Hz/sec.

At each chirp rate, peak searching is done by computing non-overlapping FFTs and their resulting power spectra. FFT lengths range from 8 to 131,072 in 15 octave steps. Peaks greater than 22 times the mean power are recorded and sent back to the SETI@home server for further analysis.

Besides searching for peaks in the multi-spectral-resolution data, SETI@home also searches for signals that match the telescope's Gaussian beam pattern. Gaussian beam fitting is computed at every frequency and every chirp rate at spectral resolutions ranging from 0.6 to 1220 Hz (temporal resolutions from 0.8 mS to 1.7 seconds). The beam fitting algorithm attempts to fit a Gaussian curve at each time and frequency in the multi-resolution spectral data, of the form:

$$P = B + Ae^{(-(\frac{(t-t_0)}{b})^2)}$$

where:

P = predicted power
B = baseline power
A = peak power
t = time
t_0 = time of Gaussian peak
b = half power beamwidth

B, A, and t_0 are free parameters in the fit, but the beamwidth is known, calculated from the slew rate of the telescope beam for each work unit. Gaussian fits whose A/B exceeds 3.2 and whose chi-squared < 10 are reported by the client software to the server for secondary analysis.

We plan to extend the primary analysis to search for pulsed signals using the Fast Folding Algorithm (FFA) (Staelin 1969) and to search for regularly-spaced triplet peaks.

2.3. Secondary Data Analysis

Most of the signals found by the client programs turn out to be terrestrial based radio frequency interference (RFI). We employ a substantial number of algorithms to reject the several types of RFI (see Cobb et al, these proceedings).

After the RFI is rejected, we search the remaining data set for multiple detections in any reference frame, giving higher weights to drifting or pulsed signals, those that repeat in the barycentric frame, that match the antenna beam pattern, or detections coincident with newly detected planets, nearby stars (from the Gliese catalog) or globular clusters (again, details in Cobb, et al). We compare candidates signals with SERENDIP IV data, and will follow up interesting candidates with dedicated observations.

3. Software Design

The SETI@home software can be divided into two parts: the "client", the program that runs on volunteer computers, and the "server", which runs on computers at UC Berkeley.

3.1. Server

The indexing and searching capabilities of a relational database are critical for the huge volumes of information handled by SETI@home. The SETI@home server uses a database containing tables for:

- Users (name, email address, work completed).
- Accounting records storing the amount of work done.
- Country, CPU type, Internet domain, and so on.
- Tapes processed.
- Work units, whether on disk or not.
- Results (per work unit).
- Platforms (types of client computers).
- Versions of the client software.

For performance, the database is divided between two servers, each running Informix, a commercial database server.

The functions of the server are divided among several programs (see Figure 1):

- Data Server: This program communicates with clients. It sends work units, accepts results, and handles requests to create new user accounts. It must handle about 10-15 requests per second, some of which can take up to a minute to complete, so several hundred copies of the program are run concurrently.
- Splitter: This program converts raw data into work units, as described in Section 2.2. Splitting is slower than real time on a Sparc Ultra 10, so we run the splitter on several machines.
- Disk Cleaner: This program deletes work units, making room for new work units. It deletes work units for which a result has been received, and if disk space is low it also deletes work units that have been sent several times.
- Accountant: This program scans flat files describing results returned, and updates accounting records. It maintains a memory cache of frequently-accessed records, minimizing database traffic.
- CGI program: This program, invoked from an Apache web server, provides database-driven features on the web site, such as Groups and Polls.
- Web page generator: This program generates frequently-accessed dynamic web pages such as Totals and Country Totals (generated every hour) and Group pages (generated every 24 hours).

All these programs are written in C++ using ESQL (embedded SQL) for database access.

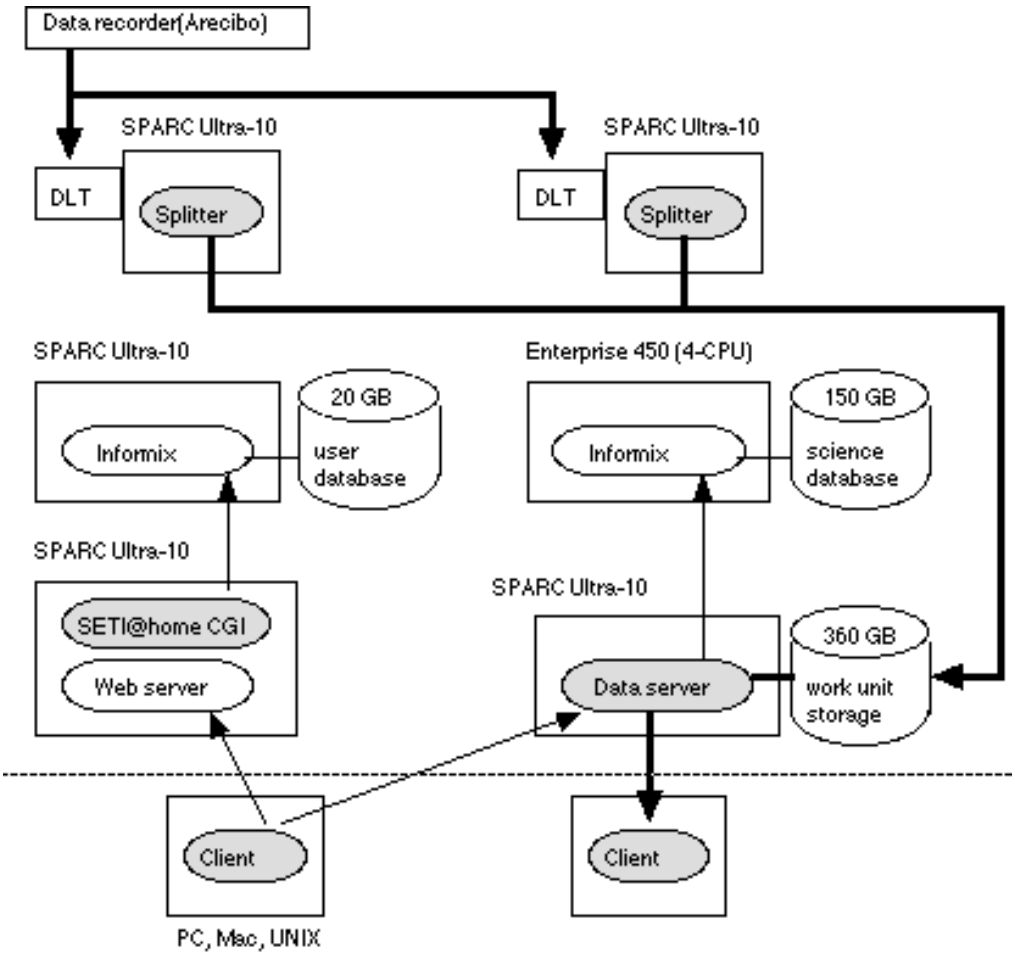

Figure 1. The SETI@home server system. Each rectangle represents a computer. Shaded ovals represent programs developed by SETI@home. The flow of radio telescope data is shown with heavy lines.

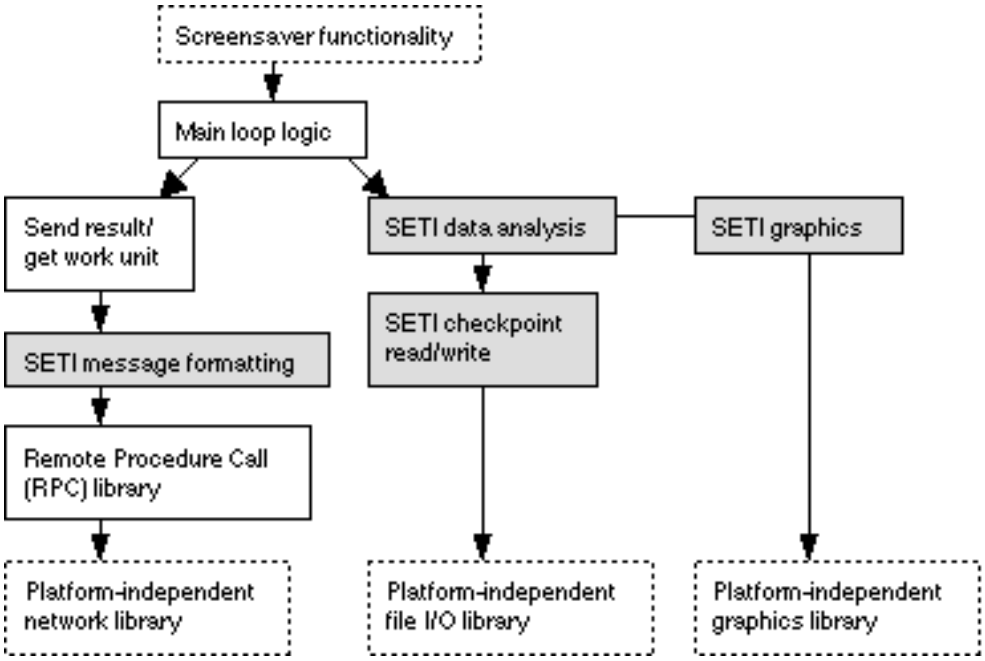

Figure 2. The client software architecture. Platform-dependent parts are in dotted boxes. SETI-specific parts are in shaded boxes.

3.2. Client

The client program is architected so that it can be easily ported to many platforms, and so that it can be retargeted for computations other than SETI@home (see Figure 2).

On Windows and Macintosh the program acts as a screensaver: it runs only when the user is not active. When the program runs, its main loop repeatedly attemps to connect to the server, get a work unit, process it, and return the result. While analyzing data, the program periodically writes a "checkpoint" file to disk, so that it will resume from the same place if the computer is stopped and restarted.

The client program has been ported to about 50 platforms. Two of these (Windows and Macintosh) required platform-specific programming. The other platforms (Linux and other forms of UNIX, BeOS, OS/2, VMS) are POSIX compliant and support the Gnu development environment; they all compile from the same source code.

To ensure that each version is numerically correct, we test each one on a "reference work unit" and validate its result before making it available for download.

4. User Involvement

The SETI@home project was announced early in 1998 and launched on May 17, 1999. During the interim our web site allowed people to sign up for notification.

We received over 400,000 such requests. Within 2 weeks of the launch there were 200,000 users, and the number has steadily increased, to a current total of 1,300,000. Users come from 224 countries, and about 50% of the users are from outside the U.S.

4.1. Web site

The SETI@home web site (http://setiathome.berkeley.edu) serves to attract SETI@home users, to educate them, and to maintain their interest and involvement in the project. The web site has many functions:

- It allows users to download the client program.
- It has educational material about SETI in general and SETI@home in particular. Separate versions are aimed at a general and scientific audiences.
- It has a Frequently Asked Questions (FAQ) section for common user problems, and a bug report submission form.
- It has News sections, updated every few days, for current general and technical information.
- It shows current usage statistics (work units completed, CPU time) both in total and broken down by various criteria (Internet domains, CPU types, countries, top 100 users, etc.).
- It shows current science results: a map of the sky showing where data has been analyzed, graphs of the distributions of spikes and Gaussians with respect to frequency and chirp rate, and so on.

4.2. Groups

The web site allows users to form "groups", typically of users within a company or school. These groups are divided into nine categories: companies (large, medium, small) schools (primary, secondary, 2-year, university), government agencies, and clubs. Users can see the top 100 (ordered by usage) groups within a category, and can search for groups by name.

The group mechanism has been very popular. Over 25,000 groups have been formed. Group leaders, in order to increase the standing of their group, often actively recruit new SETI@home users; this has expanded our user base.

4.3. Polls

In an effort to learn more about our users, we conducted a poll on the web site. This poll has questions in several areas. Some examples:

- Demographics: 92.7% of users are male.
- Attitudes about SETI: 95.6% of users think that life exists outside Earth; only 9.2% think that humans will detect an extraterrestrial signal within 2 years.
- Attitudes about distributed computing: 38% of users leave their computer on 24 hours a day because of SETI@home. 34% run SETI@home on more than one computer.

5. Conclusion

In its first five months, SETI@home has performed the largest computation in history. While it is not clear if other research projects will have the same mass appeal as does SETI, this clearly shows the viability of distributed computing for other scientific problems. We are investigating the adaptation of the SETI@home infrastructure for handling other problems.

Applied to radio SETI, distributed computing allows greater sensitivity and generality than dedicated supercomputers. However, limitations in the rates of recording and sending data limit the frequency range that can be handled.

References

Cobb, J., Lebofsky, M., Werthimer, D., Bowyer, S., & Lampton, M. 2000, this volume

Cordes. J., Lazio, T., Joseph, W., & Sagan, C. 1997, ApJ, 487, 782

Staelin, D. H. 1969, Proc. IEEE, 57, 724

Sullivan, W., Werthimer, D., Bowyer, S., Cobb, J., Gedye, D., & Anderson, D. 1997, in Astronomical and Biochemical Origins and the Search for Life in the Universe, ed: Cosmovici, Bowyer and Werthimer

Werthimer, D., Bowyer, S., Ng, D., Donnelly, C., Cobb, J., Lampton, M., & Airieau, S. 1997, in Astronomical and Biochemical Origins and the Search for Life in the Universe, ed: Cosmovici, Bowyer and Werthimer