Designing a CAPI Case Management and Synchronization Infrastructure

Steven Bochte, Brendan Day

UW Survey Center
University of Wisconsin Madison
sbochte@ssc.wisc.edu

Paper presented at IFD&TC May 2013

UWSC
UNIVERSITY of WISCONSIN
SURVEY CENTER

Goals

- Design a way to allow dozens of interviewers across the world to simultaneously synchronize CAPI interview data back to UWSC servers
- Reload the data into our custom survey software and MySQL tables for analysis and case assignment

Legacy solution (circa 2002)

- We designed our own tools and case management system
- We discovered issues over the years
  - Case assignment and ownership
  - Importing of data into MySQL tables on the server
  - Performance issues
  - Issues maintaining legacy Linux virtual machines

Cases 6 Experiment

- Advantages
  - No need to run a separate synchronization server- included in CASES 6
  - Encryption is included in CASES 6
  - Data natively stored in MySQL tables
- Modifying our existing case management tools and infrastructure to work with the new CASES 6 data structure was a substantial undertaking.
- In the end, CASES 6 didn’t meet our needs.
- We took the stricter case ownership model of CASES 6 and implemented it in our infrastructure.

New CAPI Case Management and Synchronization Infrastructure

Server Platforms

- Windows Server 2012
- CwRsync
  - Pro: Rsync implementation for Windows
  - Pro: $3 per seat when 100 seats ordered
  - Pro: License costs not tied to quantity of data transferred

Scalability

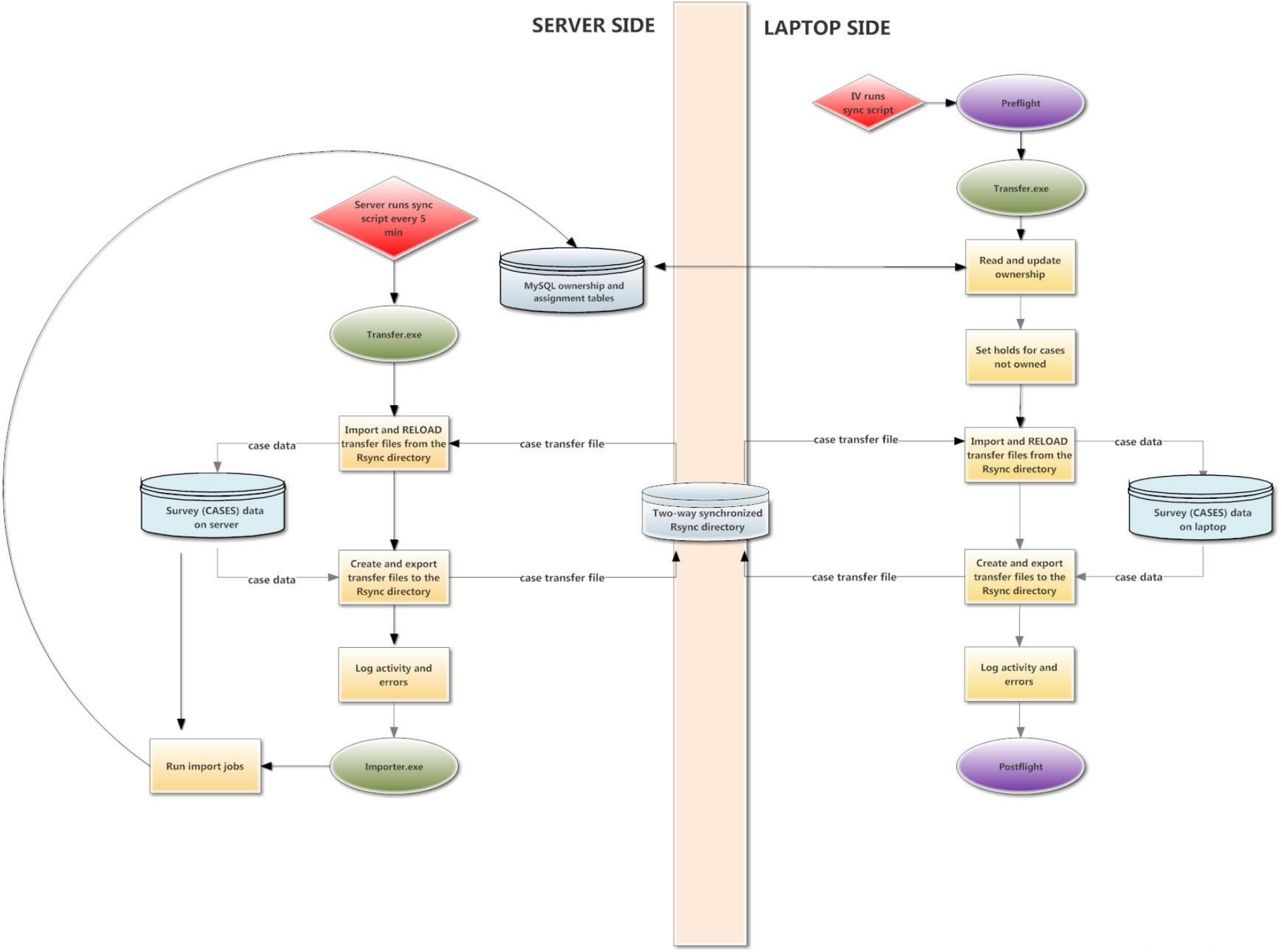
- Synchronization solution must be able to scale to at least 20,000 cases with at least 10 accesses and 100+ megabytes of audio recordings each.
- Synchronization and data loading must both scale acceptably beyond 10 interviewers syncing 100 files each, simultaneously.

Speed

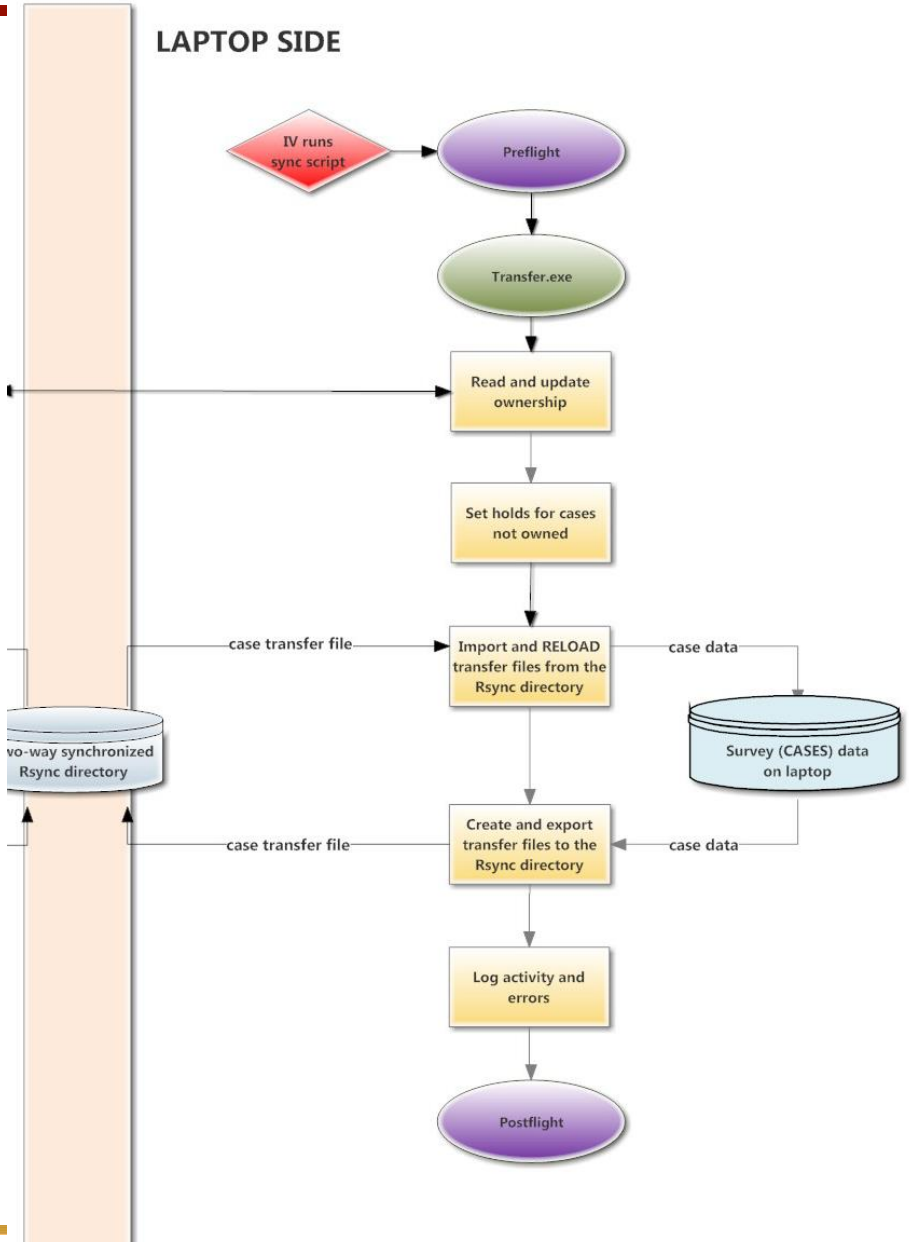
- Windows Server 2012 Rsync server
  - Ideal scenario: store all data on the Rsync server itself. Performance gains and reduction in network traffic.
  - Our existing hardware didn’t permit such a large Rsync server, so an alternate solution was needed.
  - Solution: one shared Windows account on a locked down machine, always logged in, with the Windows server-hosted network drives mapped.
  - Continuously run Windows scheduled tasks

Overview of entire synchronization and management infrastructure

(abstracted ~10x)



Interviewer (laptop) side processes



UWSC
UNIVERSITY of WISCONSIN
SURVEY CENTER

Interviewer side-Rsync wrapper

- Interviewer runs the sync script.
- We need to be able to run yet-undefined processes and executables throughout the life of the project.
- Solution:
  1. Use Rsync to pull down pre- and post-synchronization files to run
  2. Run pre-sync scripts in preflight.bat. Examples:
     - Grant NTFS rights to C:\Program Files\Subdirectory if we need to modify it.
     - Silently compress photos and audio files on the laptop to save space.
     - Grab all audio and picture files from the laptop and upload them.
     - Silently set Windows registry keys.
     - Silently install software.

Interviewer side: TRANSFER.EXE

- We need a way to package up and transfer CASES data back and forth from the server to the field for reloading into the main CASES server and the CASES instrument on each laptop.
- Solution: TRANSFER, a UWSC-designed program
  - Transfers case data (individual data records) between interviewer laptops and the server, by creating, copying, and reloading individual “transfer files.”
  - Copies transfer files to and from the Rsync folder on the server, reloading only those cases assigned to and owned by the interviewer.
  - Each transfer file contains a backup of a single case, with the file extension indicating highest session number stored. Example: file 1001.5 contains all data for case 1001, up to and including session 5.
  - Logs all transfer activity and errors in MySQL tables on the server. Tracks when assignment and ownership changes.
  - Suspends transfer of cases with unresolved errors, to prevent data corruption by repeated attempts to reload or backup case.

Case assignment and ownership

- Both our old and new systems track case assignment, and put cases on hold when they are no longer assigned to an interviewer.
- However, merely tracking assignment is not enough, because IVs do not sync regularly.
- Case Ownership
  - When interviewers synchronize, they “take ownership” of any cases assigned to them which are not already owned, and “relinquish ownership” of cases which are no longer assigned to them.
  - Any change of ownership is logged in the “transfer_history” table.

Client synchronization summary

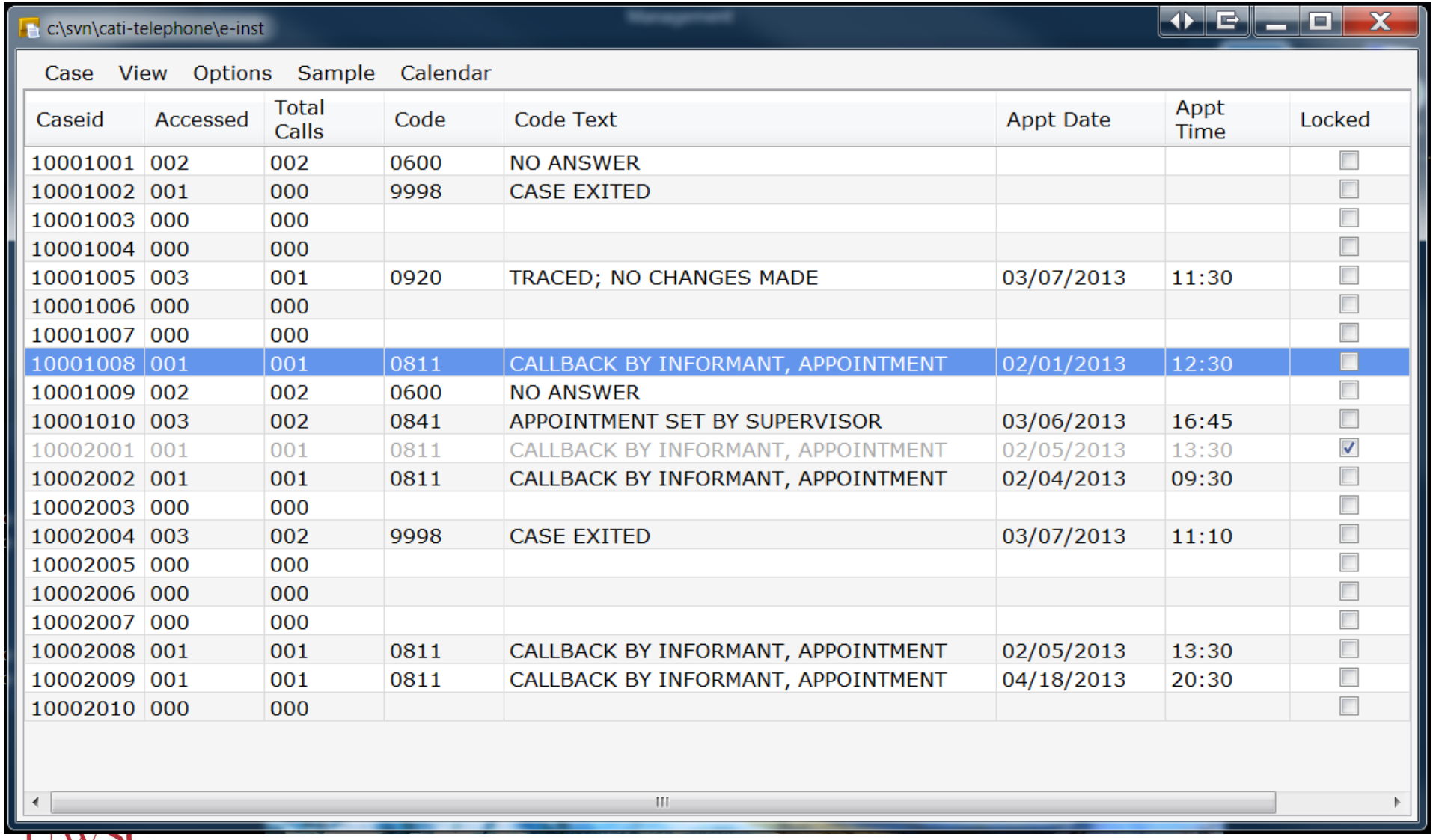
- Two-way synchronized Rsync directory
  - Download transfer files and preflight.bat from server
- Execute preflight.bat
- TRANSFER.EXE
  - Relinquish ownership of cases no longer assigned to interviewer, and set “hold” to prevent further access on laptop
  - Take ownership of cases newly assigned to interviewer
  - Create transfer files for cases accessed on laptop
  - Reload transfer files received from server
- Rsync
    - Upload new transfer files to server
- Log activity and errors

Interviewer tools

Interviewer tools

- CASEREPORT
  - Another UWSC-designed tool
  - Runs CASES utilities to output case data into a user-friendly format
  - Interviewers launch cases (data records) to conduct interviews from within Casereport

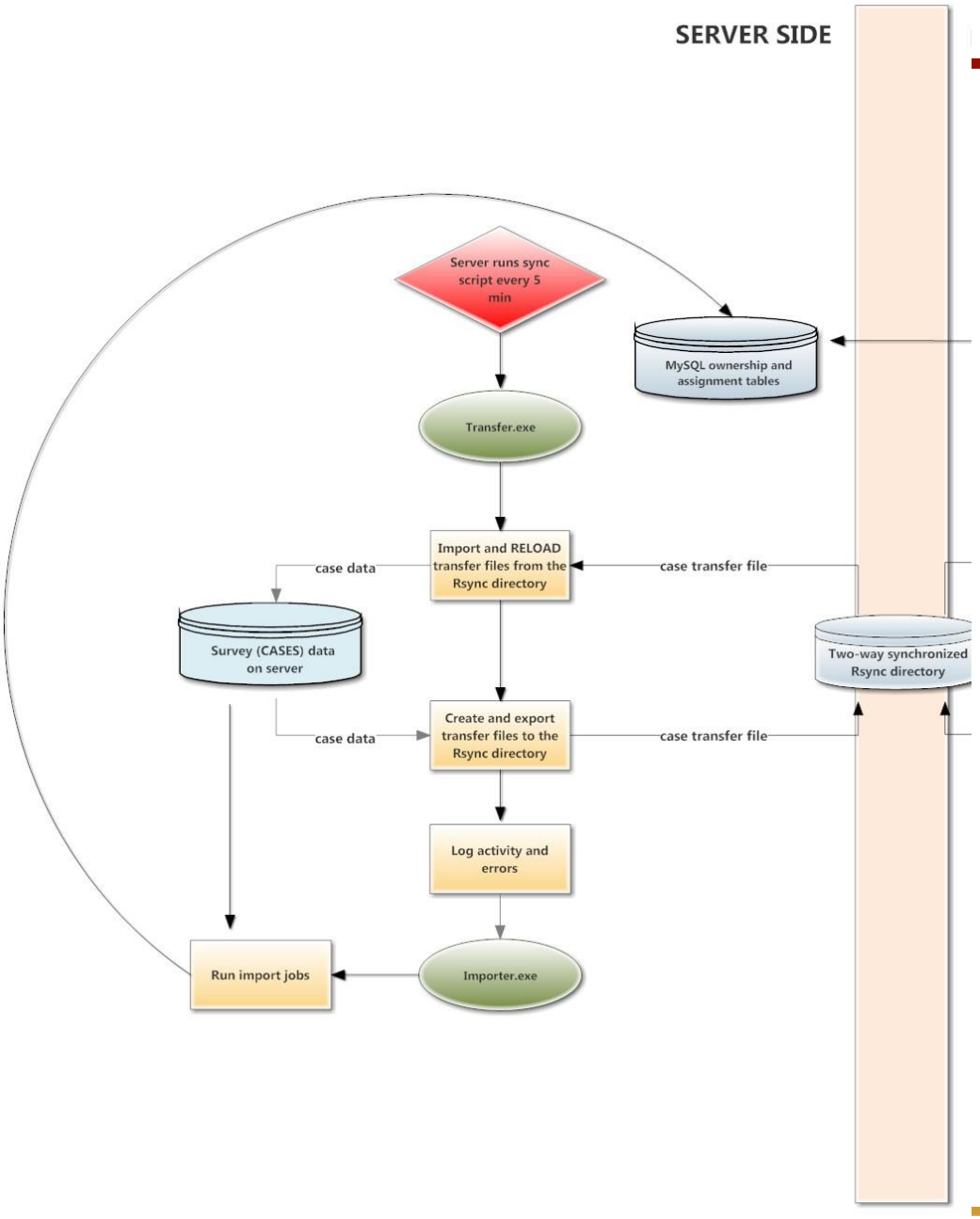
Casereport Screenshot-interviewer side



Caseid	Accessed	Total Calls	Code	Code Text	Appt Date	Appt Time	Locked
10001001	002	002	0600	NO ANSWER			☐
10001002	001	000	9998	CASE EXITED			☐
10001003	000	000					☐
10001004	000	000					☐
10001005	003	001	0920	TRACED; NO CHANGES MADE	03/07/2013	11:30	☐
10001006	000	000					☐
10001007	000	000					☐
10001008	001	001	0811	CALLBACK BY INFORMANT, APPOINTMENT	02/01/2013	12:30	☐
10001009	002	002	0600	NO ANSWER			☐
10001010	003	002	0841	APPOINTMENT SET BY SUPERVISOR	03/06/2013	16:45	☐
10002001	001	001	0811	CALLBACK BY INFORMANT, APPOINTMENT	02/05/2013	13:30	☑
10002002	001	001	0811	CALLBACK BY INFORMANT, APPOINTMENT	02/04/2013	09:30	☐
10002003	000	000					☐
10002004	003	002	9998	CASE EXITED	03/07/2013	11:10	☐
10002005	000	000					☐
10002006	000	000					☐
10002007	000	000					☐
10002008	001	001	0811	CALLBACK BY INFORMANT, APPOINTMENT	02/05/2013	13:30	☐
10002009	001	001	0811	CALLBACK BY INFORMANT, APPOINTMENT	04/18/2013	20:30	☐
10002010	000	000					☐

Server-side processes

Server-side processes



Server-side processes

- A Windows scheduled task runs the server processes every 5 minutes.
- First, the Windows task launches the same transfer process on the server, using TRANSFER.EXE, with the difference that when TRANSFER is run on the server, it knows to reload ALL cases.
  - Transfer, when run on interviewer machines (clients), only reloads cases assigned to them.

Logging Case Assignment and Ownership

- MySQL tables
  - Case ownership and assignment is tracked in the “transfer_cases” MySQL table, along with information about the transfer files most recently created or reloaded on the server and the owner’s laptop.
  - All transfer activity is logged to the “transfer_history” table on the server.

Importer details

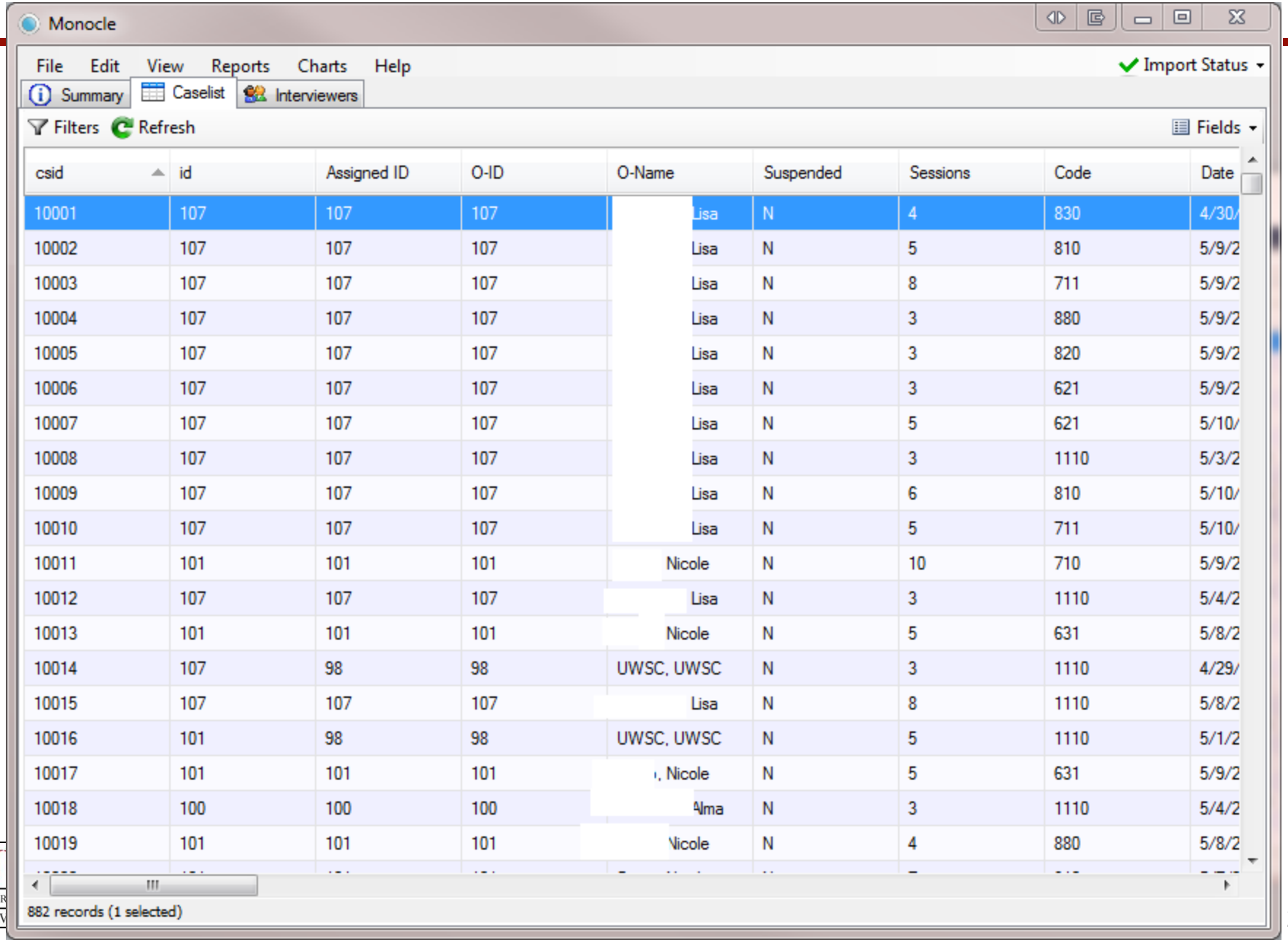
Then, the server runs UWSC-designed IMPORTER.EXE

- We have an importer processes table that defines the importer jobs to run.
- An IMPORTER records table tracks when the date, time, and size of a trace file, which is then compared to the current trace file.
- The IMPORTER compares the IMPORTER records table to the trace file time stamps to determine if a case has changed.
- If a case has changed, the server imports data for it.

Server-side process: Case Assignment

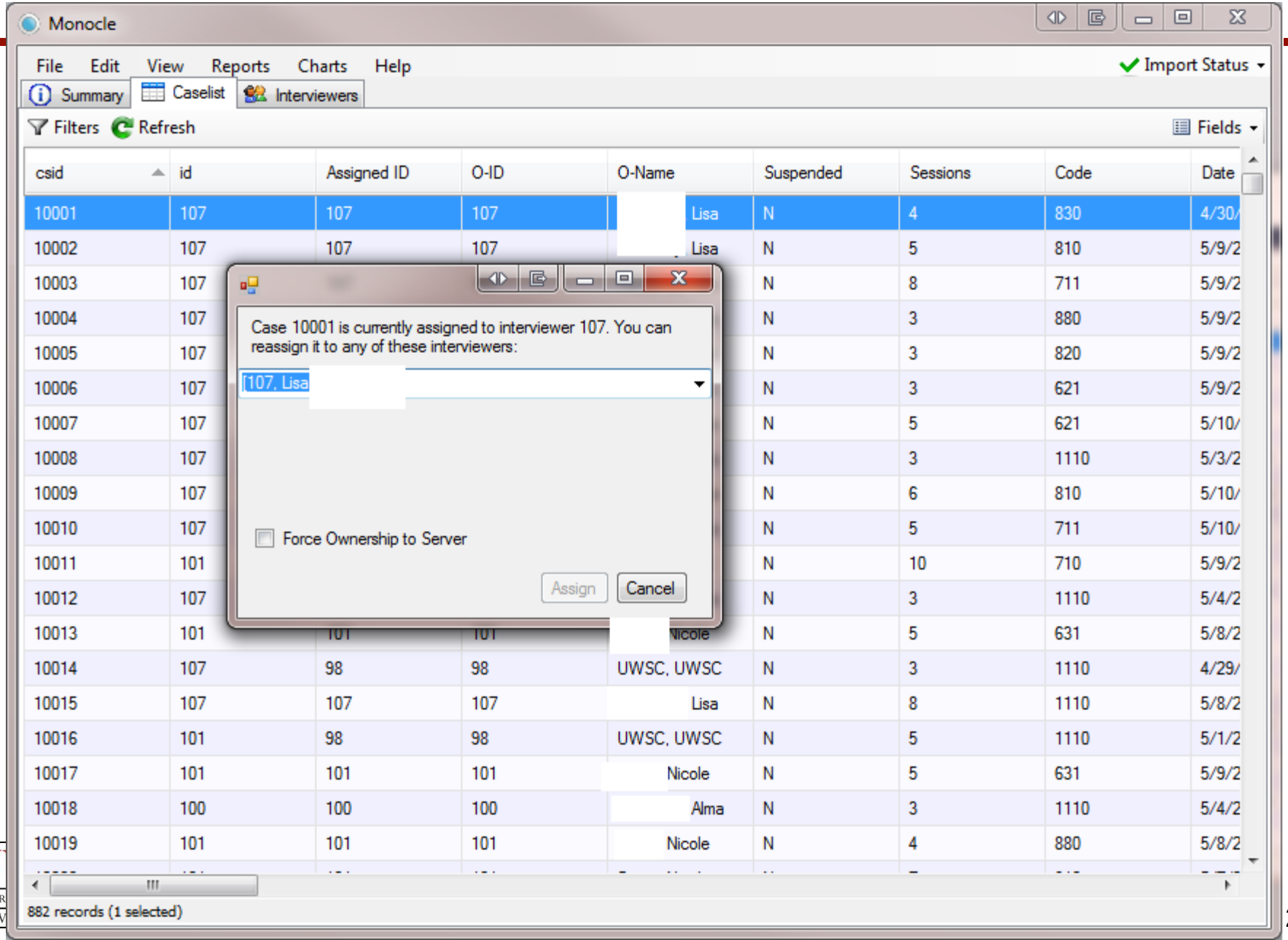
- Case Assignment
  - Field supervisors assign cases to interviewers on the server using UWSC-designed MONOCLE.EXE

Monocle screenshots-case management side



csid	id	Assigned ID	O-ID	O-Name	Suspended	Sessions	Code	Date
10001	107	107	107	Lisa	N	4	830	4/30/
10002	107	107	107	Lisa	N	5	810	5/9/2
10003	107	107	107	Lisa	N	8	711	5/9/2
10004	107	107	107	Lisa	N	3	880	5/9/2
10005	107	107	107	Lisa	N	3	820	5/9/2
10006	107	107	107	Lisa	N	3	621	5/9/2
10007	107	107	107	Lisa	N	5	621	5/10/
10008	107	107	107	Lisa	N	3	1110	5/3/2
10009	107	107	107	Lisa	N	6	810	5/10/
10010	107	107	107	Lisa	N	5	711	5/10/
10011	101	101	101	Nicole	N	10	710	5/9/2
10012	107	107	107	Lisa	N	3	1110	5/4/2
10013	101	101	101	Nicole	N	5	631	5/8/2
10014	107	98	98	UWSC, UWSC	N	3	1110	4/29/
10015	107	107	107	Lisa	N	8	1110	5/8/2
10016	101	98	98	UWSC, UWSC	N	5	1110	5/1/2
10017	101	101	101	, Nicole	N	5	631	5/9/2
10018	100	100	100	Alma	N	3	1110	5/4/2
10019	101	101	101	Nicole	N	4	880	5/8/2

Monocle screenshots-case management side



Basic Server Process Summary

Server Synchronization

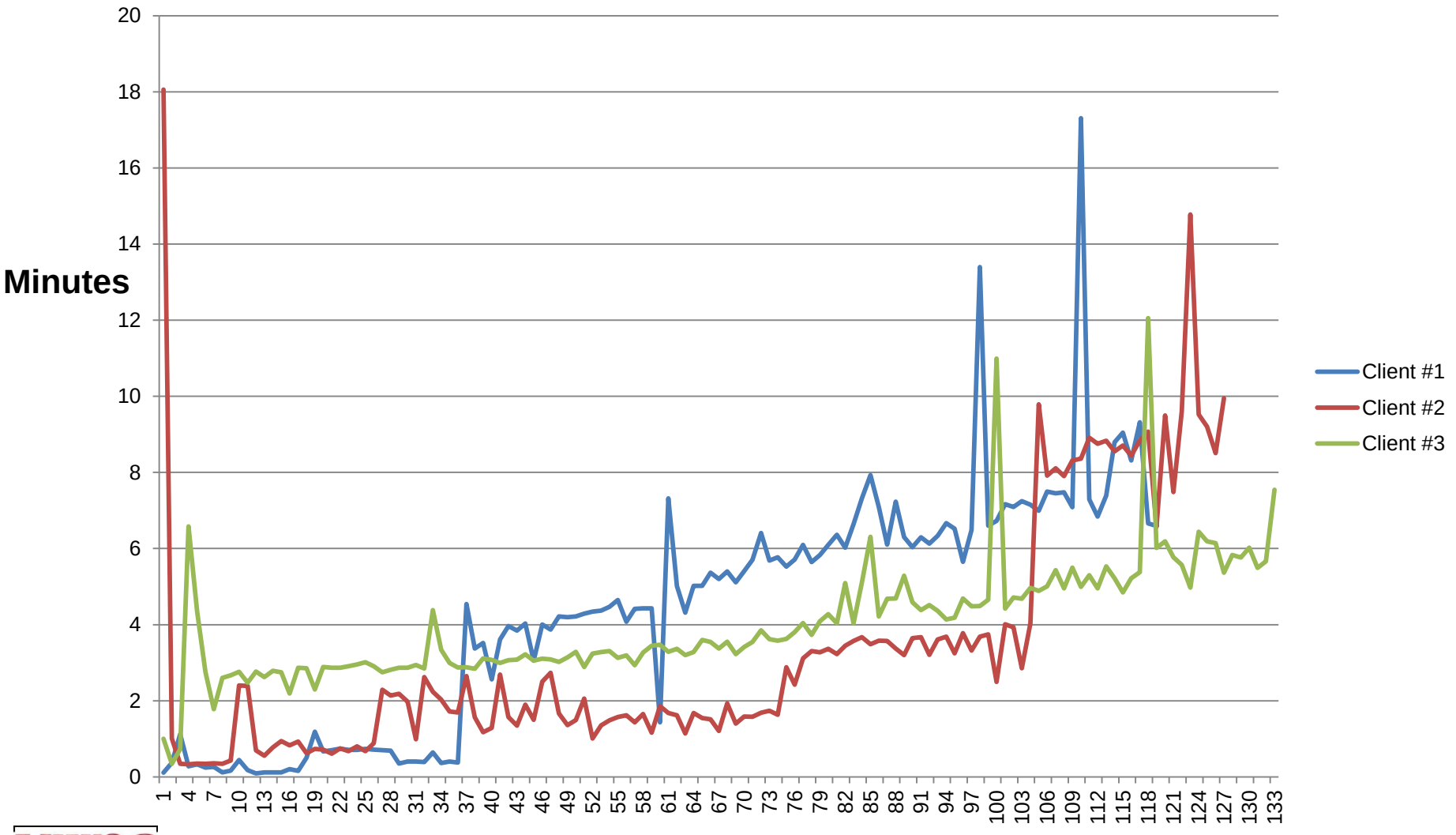
- Rsync
  - Synchronize the two-way synchronized Rsync directory
- Transfer
  - Create transfer files for cases accessed on server
  - Reload transfer files received from field
- Importer
  - Import fixed-width and open-text data to MySQL tables on server, used by project directors and supervisors
- Rsync
  - Upload transfer files to field interviewers

Speed analysis of our new case management and synchronization system

Speed

- Our shared single-user approach to the Rsync server scales well. Timings for 3 clients syncing 100 <2 MB files each every 5 minutes. We also tested 10 clients with one hundred 2 MB files each with similar results.
- The 5 minute interval between runs of the server-side synchronization scripts allows for rapid updating of information in supervisor tools.
- Audio files (often the slowest part of the sync) are always transferred last during the Rsync process, ensuring the CASES data is updated as soon as possible.

Timings



UWSC
UNIVERSITY of WISCONSIN
SURVEY CENTER

Securing the data in transit via Rsync

Security

- A VPN connection with a static IP address is required to communicate with the Rsync server. All traffic over the VPN connection is encrypted.
- We use a password file. The password is >40 characters long, randomly generated, only stored on AES full disk encrypted machines.
- A third layer of security and authentication is to run an OpenSSH server, but doing so adds overhead.
- A fourth layer of security which used in the old infrastructure is Axcrypt file level AES encryption.

Thoughts after fielding three projects with the new infrastructure

Looking back/future plans

- Looking back
  - Speed-our new infrastructure is able to handle hundreds of megabytes of audio files per case, with >15 interviewers hammering the server. The interviewer internet connections are now the bottleneck.
  - The stricter management model saves staff time, vastly reducing the need to sort out Case data conflicts
- Future plans
  - OpenSSH encryption-an extra layer of security
  - Mobile data collection-Microsoft Surface tablets/3g laplets?

Questions?

Thank You!

For copies of this presentation or more information, contact:

Steve Bochte sbochte@ssc.wisc.edu

Brendan Day bday@ssc.wisc.edu

Please visit us at:
www.uwsc.wisc.edu

UWSC
UNIVERSITY of WISCONSIN
SURVEY CENTER