Akamai® Site Snapshot Tool

User Guide

Akamai Confidential
For Customer Use Under NDA Only

September 6, 2006

Akamai Technologies, Inc

Akamai Customer Care: **1-877-425-2832** or, for routine requests, e-mail **ccare@akamai.com**

The EdgeControl® Management Center, for customers and resellers: **http://control.akamai.com**

US Headquarters
8 Cambridge Center
Cambridge, MA 02142

Tel: 617.444.3000
Fax: 617.444.3001

US Toll free 877.4AKAMAI (877.425.2624)

For a list of offices around the world, see:
http://www.akamai.com/en/html/about/locations.html

Akamai® Site Snapshot Tool User Guide

Contents

Using the Site Snapshot Tool (SST)

This guide discusses the use of Akamai's Site Snapshot Tool (SST), an advanced, robust, and flexible failover (continuity) solution that automates the process of downloading content from an enterprise onto the EdgePlatform by pulling files from your origin to your failover site.

Major features include:

- A Web interface on the Akamai EdgeControl® Management Center, as well as a command line interface, that provide the ability to create any number of simple or complex periodic Snapshot configurations.
- File transfer under HTTP and FTP. For FTP, Site Snapshot also provides the "passive retrieval" option. For HTTP, the ability to specify HTTP headers and load cookies, and to ignore or respect robots files.
- Specifying the URLs to download manually or by listing in an *input* file (a flat text file listing URLs one per line).
- Ability to set recursive behavior and the levels of recursive behavior; that is, you can set the tool to follow and download the links in the URLs to any number of levels. You can choose to download page requisites such as images and style sheet links, and you can convert links to reference the failover site.
- Ability to limit downloads to files that have changed since the previous snapshot, limit the total size of the download, set a time-out period and error logging.

Before You Begin

Setting Up with Net Storage

Before you can use SST, it must be set up using the Net Storage management tool on the Akamai EdgeControl Management Center at **https://control.akamai.com**. After initial configuration, Akamai completes the provisioning, and then you can use SST. For more information on setup, see *Managing Akamai Net Storage Accounts*.

This setup should include the upload of a security certificate public key to Akamai for secure command line access. This can be done as part of the Net Storage setup.

SST customers have access to the functionality provided with Akamai Net Storage. To use certain SST options—for example, uploading a cookie file or a flat file containing a list of URLs to download—you need to upload files in the same manner you would upload to Net Storage. For more information, see the documents *Managing Akamai Net Storage Accounts* and *Akamai Net Storage User Guide*.

Understanding How Site Snapshot Works: Considerations

Recursion and SST

SST downloads URLs and the files related to URLs. In the SST, *recursion* refers to following links, not following a directory tree. SST can download the HTML and FTP symbolic links it finds in the URLs, as well as the links it finds in *those* URLs; it can also download images and stylesheets referenced in the URLs.

Recursion Limits, Input Files, and Cookies

SST cannot recursively find and download links embedded in JavaScript such as pop-ups or image links, and it cannot follow links that generate a pull-down menu or mouse-overs. You can, however, specify objects to download via *input* files (text files containing lists of URLs) or cookies.

Ignore Robots?

Also, SST can ignore robots—origin files meant to prevent spiders from downloading objects you do not want them to download. SST will obey the robot and potentially not download files you want, unless you tell SST to ignore the robot.

Testing

You will probably find that in order to take the "right snapshot"—a download scheme that gets all the files you want and disregards the ones you do not—you will need to test your configuration, and you may need to use more than one.

As part of integration with the Akamai network, your site will be given a host name, such as failover.example.com, that you can use as a URL for browser tests.

Multiple Configurations

Using the command line interface, "more than one configuration" simply means multiple SST command lines.

Using the SST interface in the EdgeControl Management Center, you can set up a number of different configurations at the same frequency. For example, you could set up three configurations that all download weekly on Sundays at 2:00 a.m.

To illustrate using pseudo-code, you might create three different commands or configurations which, taken together, download the entire site.

```
-- get the host www.example.com and its page requisites and links
-- get the menu objects specified in menus.txt
-- get the art and image objects needed for dynamically created pages
```

Using Net Storage Content Management Server (CMS) sst to Test Your Snapshots

The best practice in most situations is to test your configurations as command lines using the Net Storage CMS **sst** command and options (see "SST Command Options" on page 8).

The CMS **sst** command is largely the same as the **sst** command used on the EdgeControl Management Center, and it takes the same options. The only difference is that the CMS **sst** is command line only, and you do not name a configuration and set up a schedule; instead, you can simply run the commands. Further, CMS **sst** also provides commands that aid in testing: for example, commands to remove files or directories, and upload individual files.

You can test input files, recursive and download options, and you can test to see whether a download is more effective using a single command or a multiple configuration setup. After you are satisfied, you can set up a named, scheduled configuration using the EdgeControl Management Center's SST interface.

Snapshot the First Time

Downloading an entire site the first time can require a significant amount of time—as much as a day or possibly more. Subsequent downloads can take much less time if you download only those files that have been modified since the previous download.

About the Failover Site

Setting up the failover site for use in the Akamai environment is, for you, the customer, a largely transparent process.

Akamai sets a flag on your normal configuration so that if there is a failure, the failover site content is used automatically. The failover content used as your origin is checked periodically at a configurable frequency, such as every 30 or 60 secs. Once the origin comes back up, normal service is restored.

For the end user, the switch from a failed origin to the failover site is also seamless, but there may be some latency as timeouts run and it is confirmed the origin is down.

Failover and Redirects

One notable difference between your origin and the failover site is that the failover site does not follow HTTP redirects. Redirects must be written into your failover configuration; check with your Akamai representative if you need to do this.

Accessing the Site Snapshot Tool

To access the SST:

1. Log in to the EdgeControl Management Center at **https://control.akamai.com**.
2. In the left-hand navigation menu, click **Net Storage** to display the **All Storage Groups** page.
3. Click the View Details link of the storage group for which you wish to configure Site Snapshot.
4. On the **Storage Group Details** page, click Site Snapshot Configuration next to the CP code you want to use.

You will now see the **Site Snapshot Tool** page, an example of which is shown here.

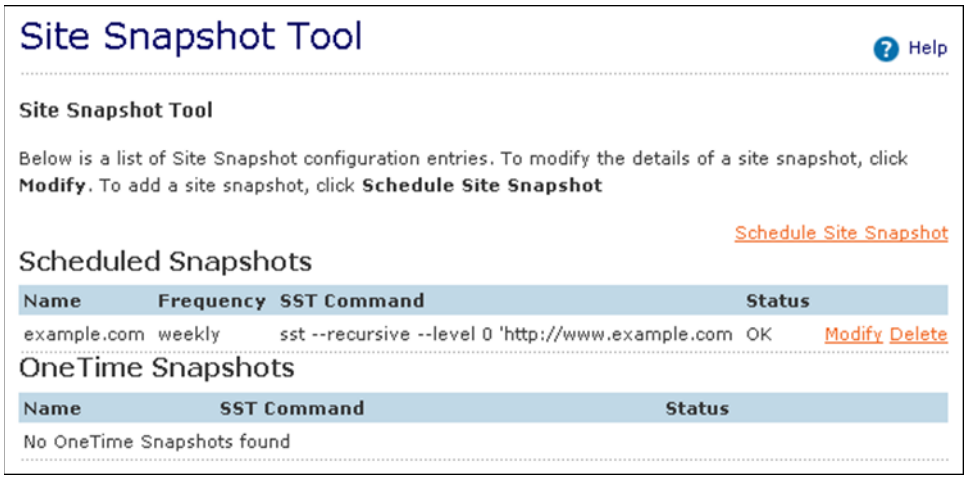

Figure 1. Site Snapshot Tool main page

The **Status** column shows whether the snapshot is in process, waiting, is completed, etc. If it displays "Error" check to see whether your options and URLs are set correctly, and if you cannot find a problem, check with Akamai Customer Care.

The **SST Command** column shows the command line results of using the interface to set the options. The command options are described beginning on page 8.

Creating Snapshots

To create and configure a snapshot:

1. On the **Site Snapshot Tool** page, click the Schedule Site Snapshot link to open the **Schedule Site Snapshot** page, parts of which are shown here.

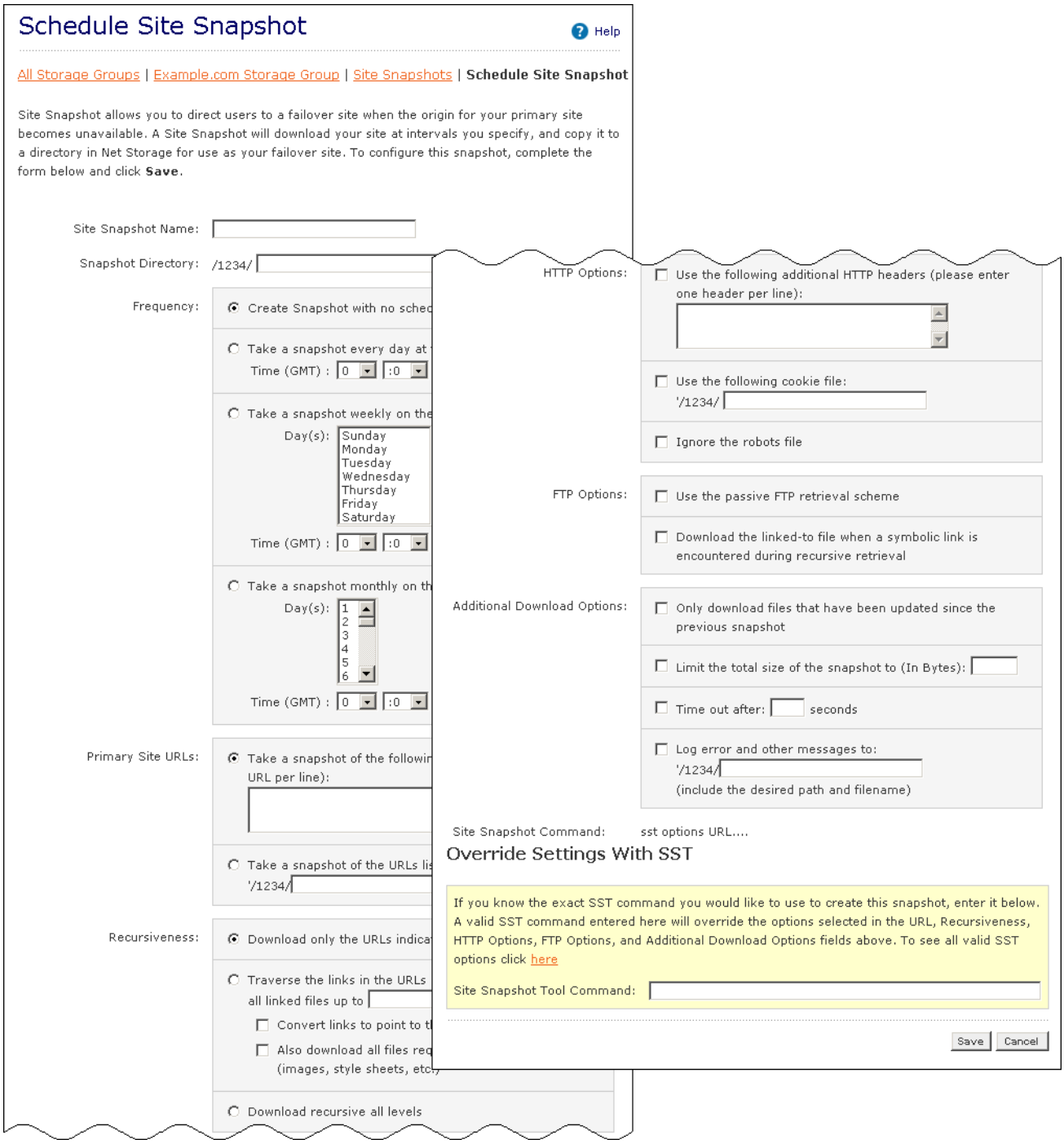

Figure 2. Parts of the Schedule Site Snapshot page

2. Complete the configuration options:

Site Snapshot Name

- The name is for your convenience and use in distinguishing between different configurations.

Snapshot Directory

- The directory to which to download as the root.

Frequency

- Set a daily, weekly, or monthly schedule, and for any frequency, you can set the time at which the download should begin.

Primary Site URLs

- Specify one or more URLs in the text box, or name a file that contains a list of URLs to download. The file needs to be a text file that has been uploaded to your Net Storage directory. Each URL should be on one line. Note that the URLs can be a combination of FTP (ftp://example.com) and HTTP (http://www.example.com), and SST will download appropriately.

Recursiveness

The Site Snapshot tool can identify and follow HTML links such as a/href and img/src. For a list of tags/attributes followed, see the **`sst`** command option "`-r --recursive`" on page 9. Note also that the command interface has additional options.

SST does not parse JavaScript. Therefore, links embedded in JavaScript such as pop-ups or image links are not downloaded under the recursiveness option. Also, SST doesn't follow links that generate a pull-down menu or mouse-overs. For these types of pages, you can create a list of URLs to download as primary site URLs. Also, SST does not download pages that are generated in response to a user interaction such as filling out a form field.

Choose one of the following recursion options:

- Download *only* the URLs specified as primary site URLs.
- Follow the links in the URLs and download linked files up to *N* levels deep.

  If you use this option, you can also convert the links to point to the failover site. Further, you can request download of all files required to view the page—images, stylesheets, and so forth.
- Recurse "all levels"—follow all links and download all possible files.

HTTP Options

- **Use the following additional HTTP headers....** That is, you can enter request headers, one per line, that will be sent from the requestor (the failover site), to your origin when the files are requested.

- Use a cookie file. The cookie file must be on your failover site, and the cookie will be passed to the HTTP server on your origin site.
- Ignore the robots file and any restrictions it contains about what may be downloaded.

FTP Options

- Use the "passive FTP retrieval" scheme to circumvent firewall issues. You can read a good explanation of the differences between active and passive FTP retrieval at **http://www.slacksite.com/other/ftp.html**.
- Follow and download files linked from symbolic links.

Additional Download Options

- Download only those files that have been modified since the last time this Snapshot was taken.
- Limit the total size of the download to some number of bytes.
- Specify a time-out in seconds. If you leave this blank, the default time-out of 900 seconds is used.
- You can turn on logging by specifying a file to which to log errors and other messages.

Override Settings with SST

You can manually override the SST command options by entering the command you want in the **Site Snapshot Tool Command** text box.

*Note: Just below the options groups, the **Site Snapshot Command** is displayed, and the command is updated as you choose options. The easiest way to create a new command is to copy and paste the command into the text box, then modify it as desired.*

The complete set of SST command options you can use is described beginning on page 8.

3. Once you are satisfied with your configuration, click **Save**.

The new Snapshot configuration is now displayed on the **Site Snapshot Tool** page.

Modifying or Deleting Snapshots

To modify a Snapshot configuration:

1. Access the **Site Snapshot Tool** page as described on page 3.
2. Click the Modify link for the Snapshot you want to change.

   This takes you to the **Schedule Site Snapshot** page.
3. Change the options as desired, as described in the discussion on page 6.
4. Click **Save**.

To delete a Snapshot configuration:

1. Access the **Site Snapshot Tool** page as described on page 3.
2. Click the Delete link for the Snapshot you want to delete and confirm the action.

Taking a Snapshot "Right Now"

This feature allows you to initiate a scheduled Snapshot now, regardless of its scheduling. When you use this feature, you clone an existing Snapshot except for the frequency, and you give the clone a name in the process.

1. Access the **Site Snapshot Tool** page as described on page 3.
2. Click the Take a Snapshot Now link to open the **Take Snapshot Now** page.
3. Choose the Snapshot you want to clone, and give the clone a name.
4. Click **Save** to initiate the Snapshot.

SST Command Options

Note: By default, SST is limited to 50,000 files per operation. If necessary, however, you may override this limit with the `--upload-quota` *option.*

Examples:

```
sst -r -N -nH http://www.example.com
```

Download www.example.com, and recurse through links to download those files as well. Download only those files that are newer than the ones already on the failover site, and do not create a host directory www.example.com/.

```
sst -p --load-cookies=cookies.txt --header="Referer: http://
      www.example.com/index.jsp" http://www.example.com
```

Load the cookies file and download based on its contents; include page requisites, and add this Referer header.

```
sst -p -N --input-file=clipart.txt
```

Download the files listed in clipart.txt and the page requisites needed to display the pages, but download files only if they are newer than the ones on the failover site.

Following is the full list of command options

-A --accept=*LIST*	Download only files with the extensions or patterns specified in ***LIST***
-B --base=*URL*	Prepend any relative links in an input file with ***URL***
-D --domains=*LIST*	Follow only domains specified in ***LIST***
-F --force-html	Regard an input file to be HTML
-g --glob=*on/off*	Enable or disable globbing to allow or disallow the special wildcard characters

-G --ignore-tags=*LIST*	Follow all of the default HTML tags that are normally following during recursion (see the "`-r --recursive`" option below) except those specified in ***LIST***
-h --help	Display help information
-H --span-hosts	Allow recursion to move to other hosts (must be used with the **--domains=*LIST*** option)
-i --input-file=*FILE*	Get the list of URLs to download from ***FILE***
-I --include-directories=*LIST*	Follow only directories specified in ***LIST***
-k --convert-links	Change absolute hyperlinks to relative
-l --level=*NUMBER*	Limit recursion depth to ***NUMBER*** levels
-L --relative	Do not follow any links but relative ones
-m --mirror	Enable options necessary to perform mirroring
-nd --no-directories	When performing recursive downloads, do not re-create the site's directory hierarchy structure; instead, copy all files to the working directory
-nH --no-host-directories	Do not include a hostname directory in the hierarchy
-N --timestamping	Only download files if they are newer than the existing ones
-o --output-file=*FILE*	Send operation information to ***FILE*** instead of the standard output
-O --output-document=*FILE*	Do not download files, but concatenate their contents and write them to ***FILE***
-q --quiet	Do not display the operation's step-by-step execution
-Q --quota=*NUMBER*	Set a byte-count limit for downloading multiple files recursively or from an input file (suffix with "k" for kilobytes or "m" for megabytes)
-p --page-requisites	In addition to the specified HTML page, also download any other files required to display the page correctly
-P --directory-prefix=*PREFIX*	Download all files and subdirectories to a directory called ***PREFIX***
-r --recursive	Download with recursion (**use this option with caution**)

By default, if you use recursion the following tags/attributes will be followed:

a/href	frame/src	script/src
applet/code	iframe/src	table/background
area/href	img/href, lowsrc, src	td/background
bgsound/src	input/src	th/background
body/background	layer/src	base/href
embed/href, src	overlay/src	link/href
fig/src		

-R --reject=*LIST*	Download all files except those with the extensions or patterns specified in ***LIST***
-S --server-response	Display sent HTTP server headers and FTP server responses
-t --tries=*NUMBER*	Make ***NUMBER*** attempts to download each URL (20 is the default; use 0 to make unlimited retries)
-T --timeout=*SECONDS*	Do not allow DNS lookups, connections attempts, and read idle times to exceed ***SECONDS***
-v --verbose	Display the operation's execution step by step (this is implied when using the sst command)
-w --wait=*SECONDS*	At the end of a file retrieval, wait ***SECONDS*** before retrieving the next file
-x --force-directories	Re-create the directory hierarchy, regardless of whether one normally would be created
-X --exclude-directories=*LIST*	Follow all directories except those specified in ***LIST***
-z --convert-absolute	Change relative hyperlinks to absolute
--exclude-domains=*LIST*	Follow all domains except those specified in ***LIST***
--follow-ftp	Do not ignore FTP links within HTML pages
--follow-tags=LIST	Follow only a subset ***LIST*** of the default HTML tags that are followed when recursing (see the "`-r --recursive`" option above).
--header=*STRING*	Include ***STRING*** with HTTP requests' headers
--http-passwd=*PASS*	Specify the HTTP server's password
--http-user=*USER*	Specify the HTTP server's user
--ignore-robots	Do not honor the robot.txt file or the robots metatag
--limit-rate=*RATE*	Do not download faster than ***RATE*** (suffix with "k" for kilobytes/second or "m" for megabytes/second)
--load-cookies=*FILE*	Prior to the first download, load the cookies contained in ***FILE*** *The cookie file format is:* *domain ignore path secure expires name*
--no-clobber	Do not download a file if it already exists in the working directory
--no-http-keep-alive	Disable the persistent connection feature
--no-parent	When using recursion, never ascend to the starting point's parent directory
--passive-ftp	Use passive mode to require the client to start communications with the server

`--random-wait`	If a file fails to download, wait either 0xWAIT, 1xWAIT, and 2xWAIT, determined randomly, before reattempting the download (WAIT is the ***SECONDS*** value set with the **`--wait=`*`SECONDS`*** option)
`--retr-symlinks`	Ignore symbolic links when performing recursive download, and download the link targets instead, unless the target is a directory
`--save-cookies=`*`FILE`*	Before quitting the session, save all of the valid cookies to ***FILE*** *The cookie file format is:* *domain ignore path secure expires name*
`--spider`	Check for the presence of files without actually downloading them
`--upload-quota=`*`QUOTA`*	Override the default 50,000 file limit for this operation only, and set the new limit to ***QUOTA*** (e.g., **`--upload-quota=100000`**)
`--waitretry=`*`SECONDS`*	If a file fails to download, reattempt after 1 second; if it again fails to download, wait 2 seconds and try again, and so on until ***SECONDS*** between attempts is reached and then stop