SystemTap 5.1

SystemTap Beginners Guide

Introduction to SystemTap

Don Domingo

William Cohen
This guide provides basic instructions on how to use SystemTap to monitor different subsystems of a Linux system in finer detail.
Preface

1. Document Conventions

This manual uses several conventions to highlight certain words and phrases and draw attention to specific pieces of information.

In PDF and paper editions, this manual uses typefaces drawn from the Liberation Fonts set. The Liberation Fonts set is also used in HTML editions if the set is installed on your system. If not, alternative but equivalent typefaces are displayed. Note: Red Hat Enterprise Linux 5 and later include the Liberation Fonts set by default.

1.1. Typographic Conventions

Four typographic conventions are used to call attention to specific words and phrases. These conventions, and the circumstances they apply to, are as follows.

Mono-spaced Bold

Used to highlight system input, including shell commands, file names and paths. Also used to highlight keys and key combinations. For example:

To see the contents of the file my_next_bestselling_novel in your current working directory, enter the cat my_next_bestselling_novel command at the shell prompt and press Enter to execute the command.

The above includes a file name, a shell command and a key, all presented in mono-spaced bold and all distinguishable thanks to context.

Key combinations can be distinguished from an individual key by the plus sign that connects each part of a key combination. For example:

Press Enter to execute the command.

Press Ctrl+Alt+F2 to switch to a virtual terminal.

The first example highlights a particular key to press. The second example highlights a key combination: a set of three keys pressed simultaneously.

If source code is discussed, class names, methods, functions, variable names and returned values mentioned within a paragraph will be presented as above, in mono-spaced bold. For example:

File-related classes include filesystem for file systems, file for files, and dir for directories. Each class has its own associated set of permissions.

Proportional Bold

This denotes words or phrases encountered on a system, including application names; dialog-box text; labeled buttons; check-box and radio-button labels; menu titles and submenu titles. For example:

1 https://fedorahosted.org/liberation-fonts/
Choose **System → Preferences → Mouse** from the main menu bar to launch **Mouse Preferences**. In the **Buttons** tab, select the **Left-handed mouse** check box and click **Close** to switch the primary mouse button from the left to the right (making the mouse suitable for use in the left hand).

To insert a special character into a **gedit** file, choose **Applications → Accessories → Character Map** from the main menu bar. Next, choose **Search → Find…** from the **Character Map** menu bar, type the name of the character in the **Search** field and click **Next**. The character you sought will be highlighted in the **Character Table**. Double-click this highlighted character to place it in the **Text to copy** field and then click the **Copy** button. Now switch back to your document and choose **Edit → Paste** from the **gedit** menu bar.

The above text includes application names; system-wide menu names and items; application-specific menu names; and buttons and text found within a GUI interface, all presented in proportional bold and all distinguishable by context.

**Mono-spaced Bold Italic or Proportional Bold Italic**

Whether mono-spaced bold or proportional bold, the addition of italics indicates replaceable or variable text. Italics denotes text you do not input literally or displayed text that changes depending on circumstance. For example:

To connect to a remote machine using ssh, type `ssh username@domain.name` at a shell prompt. If the remote machine is `example.com` and your username on that machine is john, type `ssh john@example.com`.

The `mount -o remount file-system` command remounts the named file system. For example, to remount the `/home` file system, the command is `mount -o remount /home`.

To see the version of a currently installed package, use the `rpm -q package` command. It will return a result as follows: `package-version-release`.

Note the words in bold italics above: username, domain.name, file-system, package, version and release. Each word is a placeholder, either for text you enter when issuing a command or for text displayed by the system.

Aside from standard usage for presenting the title of a work, italics denotes the first use of a new and important term. For example:

**Publican is a DocBook publishing system.**

### 1.2. Pull-quote Conventions

Terminal output and source code listings are set off visually from the surrounding text.

Output sent to a terminal is set in **mono-spaced roman** and presented thus:
Source-code listings are also set in mono-spaced roman but add syntax highlighting as follows:

```java
package org.jboss.book.jca.ex1;
import javax.naming.InitialContext;
public class ExClient
{
    public static void main(String args[])
        throws Exception
    {
        InitialContext iniCtx = new InitialContext();
        Object         ref    = iniCtx.lookup("EchoBean");
        EchoHome       home   = (EchoHome) ref;
        Echo           echo   = home.create();
        System.out.println("Created Echo");
        System.out.println("Echo.echo('Hello') = "+ echo.echo("Hello"));
    }
}
```

### 1.3. Notes and Warnings

Finally, we use three visual styles to draw attention to information that might otherwise be overlooked.

---

**Note**

Notes are tips, shortcuts or alternative approaches to the task at hand. Ignoring a note should have no negative consequences, but you might miss out on a trick that makes your life easier.

---

**Important**

Important boxes detail things that are easily missed: configuration changes that only apply to the current session, or services that need restarting before an update will apply. Ignoring a box labeled “Important” will not cause data loss but may cause irritation and frustration.

---

**Warning**

Warnings should not be ignored. Ignoring warnings will most likely cause data loss.

---

### 2. We Need Feedback!
If you find a typographical error in this manual, or if you have thought of a way to make this manual better, we would love to hear from you! Please submit a report in Bugzilla: http://sourceware.org/bugzilla/ against the product systemtap.

When submitting a report, be sure to include the specific file or URL the report refers to and the manual's identifier: SystemTap_Beginners_Guide

If you have a suggestion for improving the documentation, try to be as specific as possible when describing it. If you have found an error, please include the section number and some of the surrounding text so we can find it easily.
Introduction

SystemTap is a tracing and probing tool that allows users to study and monitor the activities of the computer system (particularly, the kernel) in fine detail. It provides information similar to the output of tools like `netstat`, `ps`, `top`, and `iostat`, but is designed to provide more filtering and analysis options for collected information.

1.1. Documentation Goals

SystemTap provides the infrastructure to monitor the running Linux kernel for detailed analysis. This can assist administrators and developers in identifying the underlying cause of a bug or performance problem.

Without SystemTap, monitoring the activity of a running kernel would require a tedious instrument, recompile, install, and reboot sequence. SystemTap is designed to eliminate this and allows users to gather the same information by running user-written SystemTap scripts.

SystemTap was initially designed for users with intermediate to advanced knowledge of the kernel. As a consequence, it is less useful to administrators or developers with limited knowledge of and experience with the Linux kernel. Moreover, much of the existing SystemTap documentation is aimed at knowledgeable and experienced users, which makes learning the tool similarly difficult.

To lower these barriers, the SystemTap Beginners Guide was written with the following goals:

- to introduce users to SystemTap, familiarize them with its architecture, and provide setup instructions;
- to provide pre-written SystemTap scripts for monitoring detailed activity in different components of the system, along with instructions on how to run them and analyze their output.

1.2. SystemTap Capabilities

- **Flexibility**: SystemTap's framework allows users to develop simple scripts for investigating and monitoring a wide variety of kernel functions, system calls, and other events that occur in kernel space. As a result, SystemTap is not so much a tool as it is a system that allows you to develop your own kernel-specific forensic and monitoring tools.

- **Ease of use**: as mentioned earlier, SystemTap allows users to probe kernel-space events without having to resort to instrument, recompile, install, and reboot the kernel.

Most of the SystemTap scripts enumerated in Chapter 5, Useful SystemTap Scripts demonstrate system forensics and monitoring capabilities not natively available with other similar tools (such as `top`, `oprofile`, or `ps`). These scripts are provided to give readers extensive examples of the application of SystemTap and to educate them further on the capabilities they can employ when writing their own SystemTap scripts.

1.3. Limitations of SystemTap

The current iteration of SystemTap allows for a multitude of options when probing kernel-space events for a wide range of kernels. However, SystemTap's ability to probe user-space events depends on kernel
support (the Utrace mechanism) that is not available in many kernels. As a consequence, only some kernel versions support user-space probing.

At present, the developmental efforts of the SystemTap community are geared towards improving SystemTap's user-space probing capabilities.
Using SystemTap

This chapter documents how to install SystemTap in the system and explains how to use the `stap` utility to run SystemTap scripts.

2.1. Installation and Setup

To deploy SystemTap, install the SystemTap packages along with the corresponding set of `-devel`, `-debuginfo`, and `-debuginfo-common` packages for your kernel. If your system has multiple kernels installed and you intend to use SystemTap on more than one of them, also install the `-devel` and `-debuginfo` packages for each of those kernel versions.

The following sections discuss the installation procedures in greater detail.

<table>
<thead>
<tr>
<th>Important</th>
</tr>
</thead>
<tbody>
<tr>
<td>Many users confuse <code>-debuginfo</code> with <code>-debug</code>. Remember that the deployment of SystemTap requires the installation of the <code>-debuginfo</code> package of the kernel, not the <code>-debug</code> version of the kernel.</td>
</tr>
</tbody>
</table>

2.1.1. Installing SystemTap

To deploy SystemTap, install the following RPM packages:

- `systemtap`
- `systemtap-runtime`

To do so, run the following command as root:

```
yum install systemtap systemtap-runtime
```

Note that before using SystemTap, you still need to install the required kernel information packages. On modern systems, run the following command as root to install these packages:

```
stap-prep
```

If this command does not work, try manual installation as described below.

2.1.2. Installing Required Kernel Information Packages Manually

SystemTap needs information about the kernel in order to place instrumentation in it (in other words, probe it). This information also allows SystemTap to generate the code for the instrumentation.
Chapter 2. Using SystemTap

The required information is contained in the matching -devel, -debuginfo, and -debuginfo-common packages for your kernel. The necessary -devel and -debuginfo packages for the ordinary "vanilla" kernel are as follows:

- kernel-debuginfo
- kernel-debuginfo-common
- kernel-devel

Likewise, the necessary packages for the PAE kernel are kernel-PAE-debuginfo, kernel-PAE-debuginfo-common, and kernel-PAE-devel.

To determine what kernel your system is currently using, use:

```
uname -r
```

For example, if you intend to use SystemTap on kernel version 2.6.18-53.el5 on an i686 machine, download and install the following RPM packages:

- kernel-debuginfo-2.6.18-53.1.13.el5.i686.rpm
- kernel-debuginfo-common-2.6.18-53.1.13.el5.i686.rpm
- kernel-devel-2.6.18-53.1.13.el5.i686.rpm

**Important**

The version, variant, and architecture of the -devel, -debuginfo and -debuginfo-common packages must match the kernel you wish to probe with SystemTap exactly.

The easiest way to install the required kernel information packages is through `yum install` and `debuginfo-install` commands. The `debuginfo-install` command is included with later versions of the `yum-utils` package (for example, version 1.1.10) and also requires an appropriate `yum` repository from which to download and install -debuginfo and -debuginfo-common packages. You can install the required -devel, -debuginfo, and -debuginfo-common packages for your kernel.

When the appropriate software repositories are enabled, install the corresponding packages for a specific kernel with the following commands:

```
yum install kernelname-devel-version
depbuginfo-install kernelname-version
```

Replace `kernelname` with the appropriate kernel variant name (for example, `kernel-PAE`), and `version` with the target kernel's version. For example, to install the required kernel information packages for the kernel-PAE-2.6.18-53.1.13.el5 kernel, run:

```
yum install kernel-PAE-devel-2.6.18-53.1.13.el5
```
Once you have manually downloaded the required packages to the machine, run the following command as root to install them:

```
rpm --force -ivh package_names
```

## 2.1.3. Initial Testing

If you are currently using the kernel you intend to probe with SystemTap, you can immediately test whether the deployment was successful. If not, you restart the system and load the appropriate kernel.

To start the test, run the following command:

```
stap -v -e 'probe vfs.read {printf("read performed
"); exit()}'
```

This command instructs SystemTap to print **read performed** and then exit properly once a virtual file system read is detected. If the SystemTap deployment was successful, it prints output similar to the following:

```
Pass 1: parsed user script and 45 library script(s) in 340user/0system/358real ms.
Pass 2: analyzed script: 1 probe(s), 1 function(s), 0 embed(s), 0 global(s) in 290user/260system/568real ms.
Pass 3: translated to C into "/tmp/stapiArgLX/stap_e5886fa50499994e6a87aadcc43cd392_399.c" in 490user/430system/938real ms.
Pass 4: compiled C into "stap_e5886fa50499994e6a87aadcc43cd392_399.ko" in 3310user/430system/3714real ms.
Pass 5: starting run.
read performed
Pass 5: run completed in 10user/40system/73real ms.
```

The last three lines of the output (beginning with **Pass 5**) indicate that SystemTap was able to successfully create the instrumentation to probe the kernel, run the instrumentation, detect the event being probed (in this case, a virtual file system read), and execute a valid handler (print text then close it with no errors).

## 2.2. Generating Instrumentation for Other Computers

When users run a SystemTap script, SystemTap builds a kernel module out of that script. SystemTap then loads the module into the kernel, allowing it to extract the specified data directly from the kernel (refer to *Procedure 3.1, “SystemTap Session”* in *Section 3.1, “Architecture”* for more information).

Normally, however, SystemTap scripts can only be run on systems where SystemTap is deployed (as in *Section 2.1, “Installation and Setup”*). This could mean that if you want to run SystemTap on ten systems, you would need to deploy SystemTap on all those systems. In some cases, this may be neither feasible nor desired. For instance, corporate policy may prohibit an administrator from installing RPM packages that provide compilers or debug information on specific machines, and thus prevent the deployment of SystemTap. To work around this problem, SystemTap allows you to use **cross-instrumentation**.
Cross-instrumentation is the process of generating SystemTap instrumentation module from a SystemTap script on one computer to be used on another computer. This process offers the following benefits:

- The kernel information packages for various machines can be installed on a single host machine.
- Each target machine only needs one RPM package to be installed in order to use the generated SystemTap instrumentation module: the systemtap-runtime package.

For the sake of simplicity, the following terms are used throughout this section:

- **Instrumentation module** — the kernel module built from a SystemTap script. The SystemTap module is built on the host system, and will be loaded on the target kernel of target system.
- **Host system** — the system on which you compile the instrumentation modules from SystemTap scripts in order to load them on target systems.
- **Target system** — the system for which you are building the instrumentation modules from SystemTap scripts.
- **Target kernel** — the kernel of the target system. This is the kernel on which you intend to load or run the instrumentation module.

To configure a host system and target systems, complete the following steps:

1. Install the systemtap-runtime package on each target system.
2. Determine the kernel running on each target system by running the **uname -r** command on each of these systems.
3. Install SystemTap on the host system. You will be building the instrumentation module for the target systems on the host system. For instructions on how to install SystemTap, refer to Section 2.1.1, “Installing SystemTap”.
4. Using the target kernel version determined earlier, install the target kernel and related RPM packages on the host system as described in Section 2.1.2, “Installing Required Kernel Information Packages Manually”. If multiple target systems use different target kernels, repeat this step for each different kernel used on the target systems.

After completing these steps, you can now build the instrumentation module (for any target system) on the host system.

To build the instrumentation module, run the following command on the host system (be sure to specify the appropriate values):

```
stap -p4 -r kernel_version script -m module_name
```

Here, kernel_version refers to the version of the target kernel (the output of the **uname -r** command on the target machine), script refers to the script to be converted into the instrumentation module, and module_name is the desired name of the instrumentation module.
To determine the architecture notation of a running kernel, you can run the following command:

```
uname -m
```

Once the *instrumentation module* is compiled, copy it to the *target system* and then load it using:

```
staprun module_name.ko
```

For example, to create the *instrumentation module* `simple.ko` from a SystemTap script named `simple.stp` for the *target kernel* 2.6.18-92.1.10.el5 (on x86_64 architecture), use the following command:

```
stap -r 2.6.18-92.1.10.el5 -e 'probe vfs.read {exit()}' -m simple
```

This creates a module named `simple.ko`. To use this *instrumentation module*, copy it to the *target system* and run the following command (on the *target system*):

```
staprun simple.ko
```

**Important**

The *host system* must be the same architecture and running the same distribution of Linux as the *target system* in order for the built *instrumentation module* to work.

### 2.3. Running SystemTap Scripts

SystemTap is distributed with a number of command line tools that allow you to monitor the activities of the system. The `stap` command reads probing instructions from a SystemTap script, translates these instructions into C code, builds a kernel module, and loads it into the running Linux kernel. The `staprun` command runs SystemTap instrumentation, that is, a kernel module built from SystemTap scripts during a cross-instrumentation.

Running `stap` and `staprun` requires elevated privileges to the system. Because not all users can be granted root access just to run SystemTap, you can allow a non-privileged user to run SystemTap instrumentation on their machine by adding them to one of the following user groups:

```
stapdev
```
Chapter 2. Using SystemTap

Members of this group can use the `stap` command to run SystemTap scripts, or `staprun` to run SystemTap instrumentation modules.

Running the `stap` command involves compiling SystemTap scripts into kernel modules and loading them into the kernel. This operation requires elevated privileges to the system, which are granted to `stapdev` members. Unfortunately, such privileges also grant effective root access to `stapdev` members. As a consequence, only grant `stapdev` group membership to users whom you can trust with root access.

`stapusr`

Members of this group can only use the `staprun` command to run SystemTap instrumentation modules. In addition, they can only run modules from the `/lib/modules/kernel_version/systemtap/` directory. Note that this directory must be owned only by the root user, and must only be writable by the root user.

The `stap` command reads a SystemTap script either from a file, or from standard input. To tell `stap` to read a SystemTap script from a file, specify the file name on the command line:

```
stap file_name
```

To instruct `stap` to read a SystemTap script from standard input, use the `-` switch instead of the file name. Note that any command-line options you wish to use must be inserted before the `-` switch. For example, to make the output of the `stap` command more verbose, type:

```
echo "probe timer.s(1) {exit()}" | stap -v -
```

Below is a list of commonly used `stap` options:

- `-v`
  Makes the output of the SystemTap session more verbose. You can repeat this option multiple times to provide more details on the script's execution, for example:

  ```
stap -vvv script.stp
  ```

  This option is particularly useful if you encounter any errors in running the script. For more information about common SystemTap script errors, refer to *Chapter 6, Understanding SystemTap Errors*.

- `-o file_name`
  Sends the standard output to a file named `file_name`.

- `-S size,count`
  Limits the maximum size of output files to `size` megabytes and the maximum number of stored files to `count`. This option implements logrotate operations for SystemTap and the resulting file names have a sequence number suffix.

- `-x process_id`
  Sets the SystemTap handler function `target()` to the specified process ID. For more information about `target()`, refer to *SystemTap Functions*. 


-c 'command'
Sets the SystemTap handler function `target()` to the specified command and runs the SystemTap instrumentation for the duration of this command. For more information about `target()`, refer to `SystemTap Functions`.

-e 'script'
Uses `script` rather than a file as input for the SystemTap translator.

-F
Uses SystemTap's flight recorder mode and makes the script a background process. For more information about flight recorder mode, refer to `Section 2.3.1, “SystemTap Flight Recorder Mode”`.

For more information about the `stap` command, refer to the stap(1) man page. For more information about the `staprun` command and cross-instrumentation, refer to `Section 2.2, “Generating Instrumentation for Other Computers”` or the staprun(8) man page.

### 2.3.1. SystemTap Flight Recorder Mode

SystemTap's flight recorder mode allows you to run a SystemTap script for long periods of time and just focus on recent output. The flight recorder mode limits the amount of output generated.

There are two variations of the flight recorder mode: `in-memory` and `file` mode. In both cases, the SystemTap script runs as a background process.

#### 2.3.1.1. In-memory Flight Recorder

When flight recorder mode is used without a file name, SystemTap uses a buffer in kernel memory to store the output of the script. Once the SystemTap instrumentation module is loaded and the probes start running, the instrumentation detaches and is put in the background. When the interesting event occurs, you can reattach to the instrumentation to see the recent output in the memory buffer and any continuing output.

To run a SystemTap script by using the flight recorder in-memory mode, run the `stap` command with the `-F` command line option:

```
stap -F iotime.stp
```

Once the script starts, `stap` prints a message similar to the following to provide you with the command to reconnect to the running script:

```
Disconnecting from systemtap module.
To reconnect, type “staprun -A stap_5dd0073edcb1f13f7565d8c343063e68_19556”
```

When the interesting event occurs, run the following command to connect to the currently running script, output the recent data in the memory buffer, and get continuing output:

```
staprun -A stap_5dd0073edcb1f13f7565d8c343063e68_19556
```
Chapter 2. Using SystemTap

By default, the kernel buffer is 1MB in size. You can increase this value by using the \texttt{-s} option with the size in megabytes (rounded up to the next power over 2) for the buffer. For example, \texttt{-s2} on the SystemTap command line would specify 2MB for the buffer.

2.3.1.2. File Flight Recorder

The flight recorder mode can also store data to files. You can control the number and size of the files kept by using the \texttt{-S} option followed by two numerical arguments separated by a comma: the first argument is the maximum size in megabytes for each output file, the second argument is the number of recent files to keep. To specify the file name, use the \texttt{-o} option followed by the name. SystemTap automatically adds a number suffix to the file name to indicate the order of the files.

The following command starts SystemTap in file flight recorder mode with the output going to files named \texttt{/tmp/pfaults.log.[0-9]+}, each file 1MB or smaller, and keeping latest two files:

\begin{verbatim}
stap -F -o /tmp/pfaults.log -S 1,2 pfaults.stp
\end{verbatim}

The command prints the process ID to standard output. Sending a SIGTERM to the process terminates the SystemTap script and stops the data collection. For example, if the previous command listed 7590 as the process ID, the following command would stop the SystemTap script:

\begin{verbatim}
kill -s SIGTERM 7590
\end{verbatim}

In this example, only the most recent two files generated by the script are kept: SystemTap automatically removes older files. As a result, the \texttt{ls -sh /tmp/pfaults.log.*} command lists two files:

\begin{verbatim}
1020K /tmp/pfaults.log.5   44K /tmp/pfaults.log.6
\end{verbatim}

To examine the latest data, read the file with the highest number, in this case \texttt{/tmp/pfaults.log.6}.
Chapter 3.

Understanding How SystemTap Works

SystemTap allows users to write and reuse simple scripts to deeply examine the activities of a running Linux system. These scripts can be designed to extract data, filter it, and summarize it quickly (and safely), enabling the diagnosis of complex performance (or even functional) problems.

The essential idea behind a SystemTap script is to name events, and to give them handlers. When SystemTap runs the script, SystemTap monitors for the event; once the event occurs, the Linux kernel then runs the handler as a quick sub-routine, then resumes.

There are several kind of events; entering/exiting a function, timer expiration, session termination, etc. A handler is a series of script language statements that specify the work to be done whenever the event occurs. This work normally includes extracting data from the event context, storing them into internal variables, and printing results.

3.1. Architecture

A SystemTap session begins when you run a SystemTap script. This session occurs in the following fashion:

Procedure 3.1. SystemTap Session

1. First, SystemTap checks the script against the existing tapset library (normally in /usr/share/systemtap/tapset/ for any tapsets used. SystemTap will then substitute any located tapsets with their corresponding definitions in the tapset library.

2. SystemTap then translates the script to C, running the system C compiler to create a kernel module from it. The tools that perform this step are contained in the systemtap package (refer to Section 2.1.1, “Installing SystemTap” for more information).

3. SystemTap loads the module, then enables all the probes (events and handlers) in the script. The staprun in the systemtap-runtime package (refer to Section 2.1.1, “Installing SystemTap” for more information) provides this functionality.

4. As the events occur, their corresponding handlers are executed.

5. Once the SystemTap session is terminated, the probes are disabled, and the kernel module is unloaded.

This sequence is driven from a single command-line program: stap. This program is SystemTap's main front-end tool. For more information about stap, refer to man stap (once SystemTap is properly installed on your machine).

3.2. SystemTap Scripts

For the most part, SystemTap scripts are the foundation of each SystemTap session. SystemTap scripts instruct SystemTap on what type of information to collect, and what to do once that information is collected.
Chapter 3. Understanding How SystemTap Works

As stated in *Chapter 3, Understanding How SystemTap Works*, SystemTap scripts are made up of two components: *events* and *handlers*. Once a SystemTap session is underway, SystemTap monitors the operating system for the specified events and executes the handlers as they occur.

### Note

An event and its corresponding handler is collectively called a **probe**. A SystemTap script can have multiple probes.

A probe's handler is commonly referred to as a **probe body**.

In terms of application development, using events and handlers is similar to instrumenting the code by inserting diagnostic print statements in a program's sequence of commands. These diagnostic print statements allow you to view a history of commands executed once the program is run.

SystemTap scripts allow insertion of the instrumentation code without recompilation of the code and allows more flexibility with regard to handlers. Events serve as the triggers for handlers to run; handlers can be specified to record specified data and print it in a certain manner.

### Format

SystemTap scripts use the file extension `.stp`, and contains probes written in the following format:

```plaintext
probe event { statements }
```

SystemTap supports multiple events per probe; multiple events are delimited by a comma (`,`). If multiple events are specified in a single probe, SystemTap will execute the handler when any of the specified events occur.

Each probe has a corresponding **statement block**. This statement block is enclosed in braces (`{ `) and contains the statements to be executed per event. SystemTap executes these statements in sequence; special separators or terminators are generally not necessary between multiple statements.

### Note

Statement blocks in SystemTap scripts follow the same syntax and semantics as the C programming language. A statement block can be nested within another statement block.
SystemTap allows you to write functions to factor out code to be used by a number of probes. Thus, rather than repeatedly writing the same series of statements in multiple probes, you can just place the instructions in a function, as in:

```plaintext
function function_name(arguments) { statements }
probe event { function_name(arguments) }
```

The statements in function_name are executed when the probe for event executes. The arguments are optional values passed into the function.

### Important

Section 3.2, “SystemTap Scripts” is designed to introduce readers to the basics of SystemTap scripts. To understand SystemTap scripts better, it is advisable that you refer to Chapter 5, Useful SystemTap Scripts; each section therein provides a detailed explanation of the script, its events, handlers, and expected output.

### 3.2.1. Event

SystemTap events can be broadly classified into two types: synchronous and asynchronous.

#### Synchronous Events

A synchronous event occurs when any process executes an instruction at a particular location in kernel code. This gives other events a reference point from which more contextual data may be available.

Examples of synchronous events include:

- `syscall.system_call`
  
  The entry to the system call `system_call`. If the exit from a syscall is desired, appending a `return` to the event monitors the exit of the system call instead. For example, to specify the entry and exit of the system call `close`, use `syscall.close` and `syscall.close.return` respectively.

- `vfs.file_operation`
  
  The entry to the `file_operation` event for Virtual File System (VFS). Similar to `syscall` event, appending a `return` to the event monitors the exit of the `file_operation` operation.

- `kernel.function("function")`
  
  The entry to the kernel function `function`. For example, `kernel.function("sys_open")` refers to the "event" that occurs when the kernel function `sys_open` is called by any thread in the system. To specify the `return` of the kernel function `sys_open`, append the `return` string to the event statement; that is, `kernel.function("sys_open").return`. 

Chapter 3. Understanding How SystemTap Works

When defining probe events, you can use asterisk (*) for wildcards. You can also trace the entry or exit of a function in a kernel source file. Consider the following example:

Example 3.1. wildcards.stp

```plaintext
probe kernel.function("@net/socket.c") { }
probe kernel.function("@net/socket.c").return { }
```

In the previous example, the first probe's event specifies the entry of ALL functions in the kernel source file net/socket.c. The second probe specifies the exit of all those functions. Note that in this example, there are no statements in the handler; as such, no information will be collected or displayed.

kernel.trace("tracepoint")

The static probe for tracepoint. Recent kernels (2.6.30 and newer) include instrumentation for specific events in the kernel. These events are statically marked with tracepoints. One example of a tracepoint available in systemtap is `kernel.trace("kfree_skb")` which indicates each time a network buffer is freed in the kernel.

module("module").function("function")

Allows you to probe functions within modules. For example:

Example 3.2. moduleprobe.stp

```plaintext
probe module("ext3").function("*") { }
probe module("ext3").function("*").return { }
```

The first probe in Example 3.2, “moduleprobe.stp” points to the entry of all functions for the ext3 module. The second probe points to the exits of all functions for that same module; the use of the .return suffix is similar to kernel.function(). Note that the probes in Example 3.2, “moduleprobe.stp” do not contain statements in the probe handlers, and as such will not print any useful data (as in Example 3.1, “wildcards.stp”).

A system's kernel modules are typically located in /lib/modules/kernel_version, where kernel_version refers to the currently loaded kernel version. Modules use the file name extension .ko.

Asynchronous Events

Asynchronous events are not tied to a particular instruction or location in code. This family of probe points consists mainly of counters, timers, and similar constructs.

Examples of asynchronous events include:

begin
The startup of a SystemTap session; that is, as soon as the SystemTap script is run.

end

The end of a SystemTap session.

timer events

An event that specifies a handler to be executed periodically. For example:

Example 3.3. timer-s.stp

```
probe timer.s(4)
{
    printf("hello world\n")
}
```

*Example 3.3, “timer-s.stp”* is an example of a probe that prints *hello world* every 4 seconds. It is also possible to use the following timer events:

- `timer.ms` *(milliseconds)*
- `timer.us` *(microseconds)*
- `timer.ns` *(nanoseconds)*
- `timer.hz` *(hertz)*
- `timer.jiffies` *(jiffies)*

When used in conjunction with other probes that collect information, timer events allows you to print out get periodic updates and see how that information changes over time.

**Important**

SystemTap supports the use of a large collection of probe events. For more information about supported events, refer to *man stapprobes*. The SEE ALSO section of *man stapprobes* also contains links to other *man* pages that discuss supported events for specific subsystems and components.

### 3.2.2. SystemTap Handler/Body

Consider the following sample script:

**Example 3.4. helloworld.stp**

```
probe begin
{

}
```
In Example 3.4, “helloworld.stp”, the event `begin` (that is, the start of the session) triggers the handler enclosed in `{ }`, which simply prints `hello world` followed by a new-line, then exits.

**Note**

SystemTap scripts continue to run until the `exit()` function executes. If the users wants to stop the execution of the script, it can interrupted manually with `Ctrl+C`.

**printf () Statements**

The `printf ()` statement is one of the simplest functions for printing data. `printf ()` can also be used to display data using a wide variety of SystemTap functions in the following format:

```
printf ("format string
", arguments)
```

The *format string* specifies how *arguments* should be printed. The format string of Example 3.4, "helloworld.stp" instructs SystemTap to print `hello world`, and contains no format specifiers.

You can use the format specifiers `%s` (for strings) and `%d` (for numbers) in format strings, depending on your list of arguments. Format strings can have multiple format specifiers, each matching a corresponding argument; multiple arguments are delimited by a comma (,).

**Note**

Semantically, the SystemTap `printf` function is very similar to its C language counterpart. The aforementioned syntax and format for SystemTap’s `printf` function is identical to that of the C-style `printf`.

To illustrate this, consider the following probe example:

```
Example 3.5. variables-in-printf-statements.stp

probe syscall.open {
    printf ("%s(%d) open
", execname(), pid())
}
```
Example 3.5, “variables-in-printf-statements.stp” instructs SystemTap to probe all entries to the system call `open`; for each event, it prints the current `execname()` (a string with the executable name) and `pid()` (the current process ID number), followed by the word `open`. A snippet of this probe’s output would look like:

```
vmware-guestosd(2206) open
hald(2360) open
hald(2360) open
hald(2360) open
df(3433) open
df(3433) open
df(3433) open
hald(2360) open
```

SystemTap Functions

SystemTap supports a wide variety of functions that can be used as `printf()` arguments. *Example 3.5, “variables-in-printf-statements.stp”* uses the SystemTap functions `execname()` (name of the process that called a kernel function/performed a system call) and `pid()` (current process ID).

The following is a list of commonly-used SystemTap functions:

- **tid()**
  
  The ID of the current thread.

- **uid()**
  
  The ID of the current user.

- **cpu()**
  
  The current CPU number.

- **gettimeofday_s()**
  
  The number of seconds since UNIX epoch (January 1, 1970).

- **ctime()**
  
  Convert number of seconds since UNIX epoch to date.

- **pp()**
  
  A string describing the probe point currently being handled.

- **thread_indent()**
Chapter 3. Understanding How SystemTap Works

This particular function is quite useful in providing you with a way to better organize your print results. The function takes one argument, an indentation delta, which indicates how many spaces to add or remove from a thread’s "indentation counter". It then returns a string with some generic trace data along with an appropriate number of indentation spaces.

The generic data included in the returned string includes a timestamp (number of microseconds since the first call to `thread_indent()` by the thread), a process name, and the thread ID. This allows you to identify what functions were called, who called them, and the duration of each function call.

If call entries and exits immediately precede each other, it is easy to match them. However, in most cases, after a first function call entry is made several other call entries and exits may be made before the first call exits. The indentation counter helps you match an entry with its corresponding exit by indenting the next function call if it is not the exit of the previous one.

Consider the following example on the use of `thread_indent()`:

**Example 3.6. thread_indent.stp**

```plaintext
probe kernel.function("*@net/socket.c").call
{
    printf("%s -> %s\n", thread_indent(1), probefunc())
}
probe kernel.function("*@net/socket.c").return
{
    printf("%s <- %s\n", thread_indent(-1), probefunc())
}
```

**Example 3.6, “thread_indent.stp”** prints out the `thread_indent()` and probe functions at each event in the following format:

```
0 ftp(7223): -> sys_socketcall
1159 ftp(7223): -> sys_socket
2173 ftp(7223): -> __sock_create
2286 ftp(7223): -> sock_alloc_inode
2737 ftp(7223): <- sock_alloc_inode
3349 ftp(7223): -> sock_alloc
3389 ftp(7223): <- sock_alloc
3417 ftp(7223): <- __sock_create
4117 ftp(7223): -> sock_create
4160 ftp(7223): <- sock_create
4301 ftp(7223): -> sock_map_fd
4644 ftp(7223): -> sock_map_file
4699 ftp(7223): <- sock_map_file
4715 ftp(7223): <- sock_map_fd
4775 ftp(7223): <- sys_socket
```

This sample output contains the following information:

- The time (in microseconds) since the initial `thread_indent()` call for the thread (included in the string from `thread_indent()`).
- The process name (and its corresponding ID) that made the function call (included in the string from `thread_indent()`).
Basic SystemTap Handler Constructs

- An arrow signifying whether the call was an entry (←) or an exit (→); the indentations help you match specific function call entries with their corresponding exits.
- The name of the function called by the process.

name

Identifies the name of a specific system call. This variable can only be used in probes that use the event syscall.system_call.

target()

Used in conjunction with stap script -x process ID or stap script -c command. If you want to specify a script to take an argument of a process ID or command, use target() as the variable in the script to refer to it. For example:

Example 3.7. targetexample.stp

```plaintext
probe syscall.* {
    if (pid() == target())
        printf("%s\n", name)
}
```

When Example 3.7, “targetexample.stp” is run with the argument -x process ID, it watches all system calls (as specified by the event syscall.*) and prints out the name of all system calls made by the specified process.

This has the same effect as specifying if (pid() == process ID) each time you wish to target a specific process. However, using target() makes it easier for you to re-use the script, giving you the ability to pass a process ID as an argument each time you wish to run the script (that is, stap targetexample.stp -x process ID).

For more information about supported SystemTap functions, refer to man stapfuncs.

3.3. Basic SystemTap Handler Constructs

SystemTap supports the use of several basic constructs in handlers. The syntax for most of these handler constructs are mostly based on C and awk syntax. This section describes several of the most useful SystemTap handler constructs, which should provide you with enough information to write simple yet useful SystemTap scripts.

3.3.1. Variables

Variables can be used freely throughout a handler; simply choose a name, assign a value from a function or expression to it, and use it in an expression. SystemTap automatically identifies whether a variable should be typed as a string or integer, based on the type of the values assigned to it. For instance, if you use set the variable foo to gettimeofday_s() (as in foo = gettimeofday_s()), then foo is typed as a number and can be printed in a printf() with the integer format specifier (%d).
Note, however, that by default variables are only local to the probe they are used in. This means that variables are initialized, used and disposed at each probe handler invocation. To share a variable between probes, declare the variable name using `global` outside of the probes. Consider the following example:

```stp
Example 3.8. timer-jiffies.stp

global count_jiffies, count_ms
probe timer.jiffies(100) { count_jiffies ++ }
probe timer.ms(100) { count_ms ++ }
probe timer.ms(12345)
{  
  hz=(1000*count_jiffies) / count_ms
  printf("jiffies:ms ratio %d:%d \n", count_jiffies, count_ms, hz)
  exit()
}
```

Example 3.8, “timer-jiffies.stp” computes the `CONFIG_HZ` setting of the kernel using timers that count jiffies and milliseconds, then computing accordingly. The `global` statement allows the script to use the variables `count_jiffies` and `count_ms` (set in their own respective probes) to be shared with `probe timer.ms(12345)`.

### Note

The `++` notation in Example 3.8, “timer-jiffies.stp” (that is, `count_jiffies ++` and `count_ms ++`) is used to increment the value of a variable by 1. In the following probe, `count_jiffies` is incremented by 1 every 100 jiffies:

```stp
probe timer.jiffies(100) { count_jiffies ++ }
```

In this instance, SystemTap understands that `count_jiffies` is an integer. Because no initial value was assigned to `count_jiffies`, its initial value is zero by default.

### 3.3.2. Target Variables

The probe events that map to actual locations in the code (for example `kernel.function("function")` and `kernel.statement("statement")`) allow the use of `target variables` to obtain the value of variables visible at that location in the code. You can use the `-L` option to list the target variable available at a probe point. If the debug information is installed for the running kernel, you can run the following command to find out what target variables are available for the `vfs_read` function:

```bash
stap -L 'kernel.function("vfs_read")'
```

This will yield something similar to the following:
Each target variable is proceeded by a "$" and the type of the target variable follows the "":. The kernel's `vfs_read` function has $file (pointer to structure describing the file), $buf (pointer to the user-space memory to store the read data), $count (number of bytes to read), and $pos (position to start reading from in the file) target variables at the entry to the function.

When a target variable is not local to the probe point, like a global external variable or a file local static variable defined in another file then it can be referenced through `@var("varname@src/file.c")`.

It is also supported to specify an executable or library file path as the second argument, as in `@var("varname", "/path/to/exe/or/lib")`

SystemTap tracks the typing information of the target variable and can examine the fields of a structure with the `->` operator. The `->` operator can be chained to look at data structures contained within data structures and follow pointers to other data structures. The `->` operator will obtain the value in the field of the structure. The `->` operator is used regardless whether accessing a field in a substructure or accessing another structure through a pointer.

For example to access a field of the static files_stat target variable defined in fs/file_table.c (which holds some of the current file system sysctl tunables), one could write:

```bash
stap -e 'probe kernel.function("vfs_read") {
    printf("current files_stat max_files: %d\n", 
        @var("files_stat@fs/file_table.c")->max_files);
    exit(); }'
```

Which will yield something similar to the following:

```
current files_stat max_files: 386070
```

For pointers to base types such as integers and strings there are a number of functions listed below to access kernel-space data. The first argument for each functions is the pointer to the data item. Similar functions are described in Section 4.2, “Accessing User-Space Target Variables” for accessing target variables in user-space code.

- **kernel_char(address)**
  Obtain the character at `address` from kernel memory.

- **kernel_short(address)**
  Obtain the short at `address` from kernel memory.

- **kernel_int(address)**
  Obtain the int at `address` from kernel memory.

- **kernel_long(address)**
  Obtain the long at `address` from kernel memory.

- **kernel_string(address)**
  Obtain the string at `address` from kernel memory.
Chapter 3. Understanding How SystemTap Works

kernel_string_n(address, n)
Obtain the string at address from the kernel memory and limits the string to n bytes.

3.3.2.1. Pretty Printing Target Variables

SystemTap scripts are often used to observe what is happening within the code. In many cases just printing the values of the various context variables is sufficient. SystemTap makes a number operations available that can generate printable strings for target variables:

$$vars$$
Expands to a character string that is equivalent to sprintf("parm1=%x ... parmN=%x var1= %x ... varN=%x", parm1, ..., parmN, var1, ..., varN) for each variable in scope at the probe point. Some values may be printed as “=?” if their run-time location cannot be found.

$$locals$$
Expands to a subset of $$vars$$ containing only the local variables.

$$parms$$
Expands to a subset of $$vars$$ containing only the function parameters.

$$return$$
Is available in return probes only. It expands to a string that is equivalent to sprintf("return= %x", $return) if the probed function has a return value, or else an empty string.

Below is a command-line script that prints the values of the parameters passed into the function vfs_read:

```bash
stap -e 'probe kernel.function("vfs_read") {printf("%s
", $$parms); exit(); }'
```

There are four parameters passed into vfs_read: file, buf, count, and pos. The $$parms$$ generates a string for the parameters passed into the function. In this case all but the count parameter are pointers. The following is an example of the output from the previous command-line script:

```
file=0xffff8800b40d4c80 buf=0x7fff634403e0 count=0x2004 pos=0xffff8800af96df48
```

Having the address a pointer points to may not be useful. Instead the fields of the data structure the pointer points to may be of more use. Use the “$” suffix to pretty print the data structure. The following command-line example uses the pretty printing suffix to print more details about the data structures passed into the function vfs_read:

```bash
stap -e 'probe kernel.function("vfs_read") {printf("%s
", $$parms$); exit(); }'
```

The previous command line will generate something similar to the following with the fields of the data structure included in the output:

```
file={.f_u={...}, .f_path={...}, .f_op=0xffffffffa06e1d80, .f_lock={...}, .f_count={...}, .f_flags=34818, .f_mode=31, ... .f_version=0, .f_security=0xffff8800b8dce560, .private_data=0x0, .f_ep_links={...}, .f_mapping=0xffff880037f8fdf8} buf="" count=8196 pos=-131938753921208
```

With the “$” suffix fields that are composed of data structures are not expanded. The “$$” suffix will print the values contained within the nested data structures. Below is an example using the “$$” suffix:
The "$\$" suffix, like all strings, is limited to the maximum string size. Below is a representative output from the previous command-line script, which is truncated because of the string size limit:

```
file={.u={.list={.next=0xffff8801336ca0e8, .prev=0xffff88012ded0840}, .rcuhead={.next=0xffff8801336ca0e8, ... .uid=0, .euid=0, .signum=0}, .cred=0xffff880130129a80, .ra={.start=0, .size=0, .async_size=0, .pages=32, ....}
```

### 3.3.2.2. Typecasting

In most cases SystemTap can determine a variable’s type from the debug information. However, code may use void pointers for variables (for example memory allocation routines) and typing information is not available. Also the typing information available within a probe handler is not available within a function; SystemTap functions arguments use a long in place of a typed pointer. SystemTap’s `@cast` operator (first available in SystemTap 0.9) can be used to indicate the correct type of the object.

The Example 3.9, “Casting Example” is from the `task.stp` tapset. The function returns the value of the `state` field from a `task_struct` pointed to by the long `task`. The first argument of the `@cast` operator, `task`, is the pointer to the object. The second argument is the type to cast the object to, `task_struct`. The third argument lists what file that the type definition information comes from and is optional. With the `@cast` operator the various fields of this particular `task_struct` can be accessed; in this example the `state` field is obtained.

**Example 3.9. Casting Example**

```c
function task_state:long (task:long) {
    return @cast(task, "task_struct", "kernel<linux/sched.h>")->state
}
```

### 3.3.2.3. Checking Target Variable Availability

As code evolves the target variables available may change. The `@defined` makes it easier to handle those variations in the available target variables. The `@defined` provides a test to see if a particular target variable is available. The result of this test can be used to select the appropriate expression.

The Example 3.10, “Testing target variable available Example” from the `memory.stp` tapset provides an probe event alias. Some version of the kernel functions being probed have an argument `$flags`. When available, the `$flags` argument is used to generate the local variable `write_access`. The versions of the probed functions that do not have the `$flags` argument have a `$write` argument and that is used instead for the local variable `write_access`.

**Example 3.10. Testing target variable available Example**

```c
probe vm.pagefault = kernel.function("__handle_mm_fault@mm/memory.c") ?,
    kernel.function("handle_mm_fault@mm/memory.c") ? {
    name = "pagefault"
    write_access = (@defined($flags) ? $flags & FAULT_FLAG_WRITE : $write_access)
}
3.3.3. Conditional Statements

In some cases, the output of a SystemTap script may be too large. To address this, you need to further refine the script's logic in order to delimit the output into something more relevant or useful to your probe.

Do this by using conditionals in handlers. SystemTap accepts the following types of conditional statements:

If/Else Statements

Format:

```
if (condition)
  statement1
else
  statement2
```

The `statement1` is executed if the `condition` expression is non-zero. The `statement2` is executed if the `condition` expression is zero. The `else` clause (`else statement2`) is optional. Both `statement1` and `statement2` can be statement blocks.

Example 3.11. ifelse.stp

```
global countread, countnonread
probe kernel.function("vfs_read"), kernel.function("vfs_write")
{
  if (probefunc() == "vfs_read")
    countread ++
  else
    countnonread ++
}
probe timer.s(5) { exit() }
probe end
{
  printf("VFS reads total %d\n VFS writes total %d\n", countread, countnonread)
}
```

Example 3.11, “ifelse.stp” is a script that counts how many virtual file system reads (`vfs_read`) and writes (`vfs_write`) the system performs within a 5-second span. When run, the script increments the value of the variable `countread` by 1 if the name of the function it probed matches `vfs_read` (as noted by the condition `if (probefunc() == "vfs_read")`); otherwise, it increments `countnonread` (`else { countnonread ++}`).

While Loops

Format:

```
while (condition)
```
So long as \texttt{condition} is non-zero the block of statements in \texttt{statement} are executed. The \texttt{statement} is often a statement block and it must change a value so \texttt{condition} will eventually be zero.

For Loops

Format:

\begin{verbatim}
for (initialization; conditional; increment) statement
\end{verbatim}

The \texttt{for} loop is shorthand for a while loop. The following is the equivalent \texttt{while} loop:

\begin{verbatim}
initialization
while (conditional) {
    statement
    increment
}
\end{verbatim}

\section*{Conditional Operators}

Aside from $==$ ("is equal to"), following operators can also be used in conditional statements:

\begin{itemize}
  \item $\geq$  \hspace{1cm} Greater than or equal to
  \item $\leq$  \hspace{1cm} Less than or equal to
  \item $\neq$  \hspace{1cm} Is not equal to
\end{itemize}

\subsection*{3.3.4. Command-Line Arguments}

A SystemTap script can also accept simple command-line arguments using a $\$ or @ immediately followed by the number of the argument on the command line. Use $\$ if you are expecting the user to enter an integer as a command-line argument, and @ if you are expecting a string.

\begin{verbatim}
Example 3.12. commandlineargs.stp
probe kernel.function(@1) { }
probe kernel.function(@1).return { }
\end{verbatim}

\textit{Example 3.12.} "commandlineargs.stp" is similar to \textit{Example 3.1.} "wildcards.stp", except that it allows you to pass the kernel function to be probed as a command-line argument (as in \texttt{stap commandlineargs.stp kernel function}). You can also specify the script to accept multiple command-line arguments, noting them as @1, @2, and so on, in the order they are entered by the user.
### 3.4. Associative Arrays

SystemTap also supports the use of associative arrays. While an ordinary variable represents a single value, associative arrays can represent a collection of values. Simply put, an associative array is a collection of unique keys; each key in the array has a value associated with it.

Since associative arrays are normally processed in multiple probes (as we will demonstrate later), they should be declared as `global` variables in the SystemTap script. The syntax for accessing an element in an associative array is similar to that of `awk`, and is as follows:

```
array_name[index_expression]
```

Here, the `array_name` is any arbitrary name the array uses. The `index_expression` is used to refer to a specific unique key in the array. To illustrate, let us try to build an array named `foo` that specifies the ages of three people `tom`, `dick`, and `harry` (which are unique keys). To assign them the ages (associated values) of 23, 24, and 25 respectively, we'd use the following array statements:

#### Example 3.13. Basic Array Statements

```plaintext
foo["tom"] = 23
foo["dick"] = 24
foo["harry"] = 25
```

You can specify up to nine index expressions in an array statement, each one delimited by a comma (,). This is useful if you wish to have a key that contains multiple pieces of information. The following line from `disktop.stp` uses 5 elements for the key: process ID, executable name, user ID, parent process ID, and string "W". It associates the value of `devname` with that key.

```
device[pid(),execname(),uid(),ppid(),"W"] = devname
```

#### Important

All associate arrays must be declared as `global`, regardless of whether the associate array is used in one or multiple probes.

### 3.5. Array Operations in SystemTap
Assigning an Associated Value

This section enumerates some of the most commonly used array operations in SystemTap.

### 3.5.1. Assigning an Associated Value

Use `=` to set an associated value to indexed unique pairs, as in:

```
array_name[index_expression] = value
```

*Example 3.13, “Basic Array Statements”* shows a very basic example of how to set an explicit associated value to a unique key. You can also use a handler function as both your `index_expression` and `value`. For example, you can use arrays to set a timestamp as the associated value to a process name (which you wish to use as your unique key), as in:

*Example 3.14. Associating Timestamps to Process Names*

```
foo[tid()] = gettimeofday_s()
```

Whenever an event invokes the statement in *Example 3.14, “Associating Timestamps to Process Names”*, SystemTap returns the appropriate `tid()` value (that is, the ID of a thread, which is then used as the unique key). At the same time, SystemTap also uses the function `gettimeofday_s()` to set the corresponding timestamp as the associated value to the unique key defined by the function `tid()`. This creates an array composed of key pairs containing thread IDs and timestamps.

In this same example, if `tid()` returns a value that is already defined in the array `foo`, the operator will discard the original associated value to it, and replace it with the current timestamp from `gettimeofday_s()`.

### 3.5.2. Reading Values From Arrays

You can also read values from an array the same way you would read the value of a variable. To do so, include the `array_name[index_expression]` statement as an element in a mathematical expression. For example:

*Example 3.15. Using Array Values in Simple Computations*

```
delta = gettimeofday_s() - foo[tid()]
```

This example assumes that the array `foo` was built using the construct in *Example 3.14, “Associating Timestamps to Process Names”* (from *Section 3.5.1, “Assigning an Associated Value”). This sets a timestamp that will serve as a reference point, to be used in computing for `delta`.

The construct in *Example 3.15, “Using Array Values in Simple Computations”* computes a value for the variable `delta` by subtracting the associated value of the key `tid()` from the current
Chapter 3. Understanding How SystemTap Works

gettimeofday_s(). The construct does this by reading the value of tid() from the array. This particular construct is useful for determining the time between two events, such as the start and completion of a read operation.

### Note

If the index_expression cannot find the unique key, it returns a value of 0 (for numerical operations, such as Example 3.15, “Using Array Values in Simple Computations”) or a null/empty string value (for string operations) by default.

#### 3.5.3. Incrementing Associated Values

Use ++ to increment the associated value of a unique key in an array, as in:

```plaintext
array_name[index_expression] ++
```

Again, you can also use a handler function for your index_expression. For example, if you wanted to tally how many times a specific process performed a read to the virtual file system (using the event vfs.read), you can use the following probe:

### Example 3.16. vfsreads.stp

```plaintext
probe vfs.read {
  reads[execname()] ++
}
```

In Example 3.16, “vfsreads.stp”, the first time that the probe returns the process name gnome-terminal (that is, the first time gnome-terminal performs a VFS read), that process name is set as the unique key gnome-terminal with an associated value of 1. The next time that the probe returns the process name gnome-terminal, SystemTap increments the associated value of gnome-terminal by 1. SystemTap performs this operation for all process names as the probe returns them.

#### 3.5.4. Processing Multiple Elements in an Array

Once you’ve collected enough information in an array, you will need to retrieve and process all elements in that array to make it useful. Consider Example 3.16, “vfsreads.stp”: the script collects information about how many VFS reads each process performs, but does not specify what to do with it. The obvious means for making Example 3.16, “vfsreads.stp” useful is to print the key pairs in the array reads, but how?
The best way to process all key pairs in an array (as an iteration) is to use the `foreach` statement. Consider the following example:

**Example 3.17. cumulative-vfsreads.stp**

```plaintext
global reads
probe vfs.read
{
  reads[execname()] ++
}
probe timer.s(3)
{
  foreach (count in reads)
    printf("%s : %d \n", count, reads[count])
}
```

In the second probe of Example 3.17, “cumulative-vfsreads.stp”, the `foreach` statement uses the variable `count` to reference each iteration of a unique key in the array `reads`. The `reads[count]` array statement in the same probe retrieves the associated value of each unique key.

Given what we know about the first probe in Example 3.17, “cumulative-vfsreads.stp”, the script prints VFS-read statistics every 3 seconds, displaying names of processes that performed a VFS-read along with a corresponding VFS-read count.

Now, remember that the `foreach` statement in Example 3.17, “cumulative-vfsreads.stp” prints all iterations of process names in the array, and in no particular order. You can instruct the script to process the iterations in a particular order by using `+` (ascending) or `-` (descending). In addition, you can also limit the number of iterations the script needs to process with the `limit` value option.

For example, consider the following replacement probe:

```plaintext
probe timer.s(3)
{
  foreach (count in reads- limit 10)
    printf("%s : %d \n", count, reads[count])
}
```

This `foreach` statement instructs the script to process the elements in the array `reads` in descending order (of associated value). The `limit 10` option instructs the `foreach` to only process the first ten iterations (that is, print the first 10, starting with the highest value).

### 3.5.5. Clearing/Deleting Arrays and Array Elements

Sometimes, you may need to clear the associated values in array elements, or reset an entire array for re-use in another probe. Example 3.17, “cumulative-vfsreads.stp” in Section 3.5.4, “Processing Multiple Elements in an Array” allows you to track how the number of VFS reads per process grows over time, but it does not show you the number of VFS reads each process makes per 3-second period.
To do that, you will need to clear the values accumulated by the array. You can accomplish this using the `delete` operator to delete elements in an array, or an entire array. Consider the following example:

```
Example 3.18. noncumulative-vfsreads.stp

    global reads
    probe vfs.read
    {
      reads[execname()] ++
    }
    probe timer.s(3)
    {
      foreach (count in reads)
        printf("%s : %d 
", count, reads[count])
      delete reads
    }
```

In Example 3.18, “noncumulative-vfsreads.stp”, the second probe prints the number of VFS reads each process made within the probed 3-second period only. The `delete reads` statement clears the `reads` array within the probe.
You can have multiple array operations within the same probe. Using the examples from Section 3.5.4, “Processing Multiple Elements in an Array” and Section 3.5.5, “Clearing/Deleteing Arrays and Array Elements”, you can track the number of VFS reads each process makes per 3-second period and tally the cumulative VFS reads of those same processes. Consider the following example:

```
global reads, totalreads

probe vfs.read
{
    reads[execname()] ++
    totalreads[execname()] ++
}

probe timer.s(3)
{
    printf("-------\n")
    foreach (count in reads-)
        printf("%s : %d \n", count, reads[count])
    delete reads
}

probe end
{
    printf("TOTALS\n")
    foreach (total in totalreads-)
        printf("%s : %d \n", total, totalreads[total])
}
```

In this example, the arrays `reads` and `totalreads` track the same information, and are printed out in a similar fashion. The only difference here is that `reads` is cleared every 3-second period, whereas `totalreads` keeps growing.

### 3.5.6. Using Arrays in Conditional Statements

You can also use associative arrays in `if` statements. This is useful if you want to execute a subroutine once a value in the array matches a certain condition. Consider the following example:

```
Example 3.19. vfsreads-print-if-1kb.stp

global reads
probe vfs.read
{
    reads[execname()] ++
}

probe timer.s(3)
```
Chapter 3. Understanding How SystemTap Works

Every three seconds, *Example 3.19, “vfsreads-print-if-1kb.stp”* prints out a list of all processes, along with how many times each process performed a VFS read. If the associated value of a process name is equal or greater than 1024, the *if* statement in the script converts and prints it out in kB.

Testing for Membership

You can also test whether a specific unique key is a member of an array. Further, membership in an array can be used in *if* statements, as in:

```
if([index_expression] in array_name) statement
```

To illustrate this, consider the following example:

*Example 3.20. vfsreads-stop-on-stapio2.stp*

```
global reads

probe vfs.read
{
    reads[execname()] ++
}

probe timer.s(3)
{
    printf("-------\n")
    foreach (count in reads+)
    {
        printf("%s : %d \n", count, reads[count])
    }
    if(["stapio"] in reads) {
        printf("stapio read detected, exiting\n")
        exit()
    }
}
```

The *if(["stapio"] in reads)* statement instructs the script to print *stapio read detected, exiting* once the unique key *stapio* is added to the array *reads*.

3.5.7. Computing for Statistical Aggregates

Statistical aggregates are used to collect statistics on numerical values where it is important to accumulate new data quickly and in large volume (that is, storing only aggregated stream statistics). Statistical aggregates can be used in global variables or as elements in an array.
To add value to a statistical aggregate, use the operator $\ll< value$.

**Example 3.21. stat-aggregates.stp**

```plaintext
global reads
probe vfs.read
{
    reads[execname()] $\ll< $count
}
```

In *Example 3.21, “stat-aggregates.stp”*, the operator $\ll< $count stores the amount returned by $\texttt{count}$ to the associated value of the corresponding `execname()` in the `reads` array. Remember, these values are *stored*; they are not added to the associated values of each unique key, nor are they used to replace the current associated values. In a manner of speaking, think of it as having each unique key (`execname()`) having multiple associated values, accumulating with each probe handler run.

**Note**

In the context of *Example 3.21, “stat-aggregates.stp”*, `count` returns the amount of data read by the returned `execname()` to the virtual file system.

To extract data collected by statistical aggregates, use the syntax format `@extractor(variable/array index expression)`. *extractor* can be any of the following integer extractors:

**count**

- Returns the number of all values stored into the variable/array index expression. Given the sample probe in *Example 3.21, “stat-aggregates.stp”*, the expression `@count(reads[execname()])` will return *how many values are stored* in each unique key in array `reads`.

**sum**

- Returns the sum of all values stored into the variable/array index expression. Again, given sample probe in *Example 3.21, “stat-aggregates.stp”*, the expression `@sum(reads[execname()])` will return *the total of all values stored* in each unique key in array `reads`.

**min**

- Returns the smallest among all the values stored in the variable/array index expression.

**max**

- Returns the largest among all the values stored in the variable/array index expression.
avg

Returns the average of all values stored in the variable/array index expression.

When using statistical aggregates, you can also build array constructs that use multiple index expressions (to a maximum of 5). This is helpful in capturing additional contextual information during a probe. For example:

Example 3.22. Multiple Array Indexes

```
global reads
probe vfs.read
{
    reads[execname(),pid()] <<< 1
}
probe timer.s(3)
{
    foreach([var1,var2] in reads)
        printf("%s (%d) : %d \n", var1, var2, @count(reads[var1,var2]))
}
```

In Example 3.22, “Multiple Array Indexes”, the first probe tracks how many times each process performs a VFS read. What makes this different from earlier examples is that this array associates a performed read to both a process name and its corresponding process ID.

The second probe in Example 3.22, “Multiple Array Indexes” demonstrates how to process and print the information collected by the array reads. Note how the foreach statement uses the same number of variables (that is, var1 and var2) contained in the first instance of the array reads from the first probe.

### 3.6. Tapsets

**Tapsets** are scripts that form a library of pre-written probes and functions to be used in SystemTap scripts. When a user runs a SystemTap script, SystemTap checks the script's probe events and handlers against the tapset library; SystemTap then loads the corresponding probes and functions before translating the script to C (refer to Section 3.1, “Architecture” for information on what transpires in a SystemTap session).

Like SystemTap scripts, tapsets use the file name extension .stp. The standard library of tapsets is located in /usr/share/systemtap/tapset/ by default. However, unlike SystemTap scripts, tapsets are not meant for direct execution; rather, they constitute the library from which other scripts can pull definitions.

The tapset library is an abstraction layer designed to make it easier for users to define events and functions. Tapsets provide useful aliases for functions that users may want to specify as an event; knowing the proper alias to use is, for the most part, easier than remembering specific kernel functions that might vary between kernel versions.

Several handlers and functions in Section 3.2.1, “Event” and SystemTap Functions are defined in tapsets. For example, thread_indent() is defined in indent.stp.
User-space Probing

SystemTap initially focused on kernel-space probing. Because there are many instances where user-space probing can help diagnose a problem, SystemTap 0.6 added support to allow probing user-space processes. SystemTap can probe the entry into and return from a function in user-space processes, probe predefined markers in user-space code, and monitor user-process events.

SystemTap requires the uprobes module to perform user-space probing. If your Linux kernel is version 3.5 or higher, it already includes uprobes. To verify that the current kernel supports uprobes natively, run the following command:

```
grep CONFIG_UPROBES /boot/config-`uname -r`
```

If uprobes is integrated, the output of this command is as follows:

```
CONFIG_UPROBES=y
```

4.1. User-Space Events

All user-space event probes begin with `process`. You can limit the process events to a specific running process by specifying the process ID. You can also limit the process events to monitor a particular executable by specifying the path to the executable (PATH). SystemTap makes use of the `PATH` environment variable, which allows you to use both the name used on the command-line to start the executable and the absolute path to the executable.

Several of the user-space probe events limit their scope to a particular executable name (PATH), because SystemTap must use debug information to statically analyze where to place the probes. But for many user-space probe events, the process ID and executable name are optional. Any `process` event in the list below that include process ID or the path to the executable must include those arguments. The process ID and path to the executable are optional for the `process` events that do not list them:

- `process("PATH").function("function")`
  
  The entry to the user-space function `function` for the executable `PATH`. This event is the user-space analogue of the `kernel.function("function")` event. It allows wildcards for the function `function` and `.return` suffix.

- `process("PATH").statement("statement")`
  
  The earliest instruction in the code for `statement`. This is the user-space analogue of `kernel.statement("statement")`.

- `process("PATH").mark("marker")`
  
  The static probe point `marker` defined in `PATH`. You can use wildcards for `marker` to specify multiple marks with a single probe. The static probe points may also have numbered arguments ($1, $2, and so on) available to the probe.

  A variety of user-space packages such as Java include these static probe points. Most packages that provide static probe points also provide aliases for the raw user-space mark events. Below is one such alias for the x86_64 Java hotspot JVM:
**Chapter 4. User-space Probing**

```plaintext
probe hotspot.gc_begin =
    process("/usr/lib/jvm/java-1.6.0-openjdk-1.6.0.0.x86_64/jre/lib/amd64/server/libjvm.so").mark("gc__begin")
```

process.begin
A user-space process is created. You can limit this to a particular process ID or a full path to the executable.

process.thread.begin
A user-space thread is created. You can limit this to a particular process ID or a full path to the executable.

process.end
A user-space process dies. You can limit this to a particular process ID or a full path to the executable.

process.thread.end
A user-space thread is destroyed. You can limit this to a particular process ID or a full path to the executable.

process.syscall
A user-space process makes a system call. The system call number is available in the $syscall context variable, and the first six arguments are available in $arg1 through $arg6. The .return suffix places the probe at the return from the system call. For syscall.return, the return value is available through the $return context variable.

You can limit this to a particular process ID or a full path to the executable.

### 4.2. Accessing User-Space Target Variables

You can access user-space target variables in the same manner as described in Section 3.3.2, "Target Variables". In Linux, however, there are separate address spaces for the user and kernel code. When using the -> operator, SystemTap accesses the appropriate address space.

For pointers to base types such as integers and strings, there are a number of functions listed below to access user-space data. The first argument for each functions is the pointer to the data item.

**user_char(address)**
Obtains the character at address for the current user process.

**user_short(address)**
Obtains the short integer at address for the current user process.

**user_int(address)**
Obtains the integer at address for the current user process.

**user_long(address)**
Obtains the long integer at address for the current user process.
user_string(address)
Obtains the string at address for the current user process.

user_string_n(address, n)
Obtains the string at address for the current user process and limits the string to n bytes.

4.3. User-Space Stack Backtraces

The probe point (pp) function indicates which particular event triggered the SystemTap event handler. A probe on the entry into a function would list the function name. However, in many cases the same probe point event may be triggered by many different modules in the program; this is particularly true for functions in shared libraries. A SystemTap backtrace of the user-space stack can provide additional context on how the probe point event is triggered.

The user-space stack backtrace generation is complicated by the compiler producing code optimized to eliminate stack frame pointers. However, the compiler also includes information in the debug information section to allow debugging tools to produce stack backtraces. SystemTap user-space stack backtrace mechanism makes use of that debug information to walk the stack to generate stack traces for 32-bit and 64-bit x86 processors; other processor architectures do not yet support the use of debug information to unwind the user-space stack. To ensure that the needed debug information is used to produce the user-space stack backtraces, use the -d executable option for executables and --ldd for shared libraries.

For example, you can use the user-space backtrace functions to see how the xmalloc function is being called by the ls command. With the debuginfo for the ls command installed, the following SystemTap command provides a backtrace each time the xmalloc function is called:

```
stap -d /bin/ls --ldd "\n-e 'probe process("ls") .function("xmalloc") {print_usyms(ubacktrace())}'; \n-c "ls /""
```

When executed, this command produces output similar to the following:

```
bin dev lib media net proc sbin sys var
cgroup home lost+found mnt opt root srv usr
0x4116c0 : xmalloc+0x0/0x20 [/bin/ls]
0x4116fc : xmemdup+0x1c/0x40 [/bin/ls]
0x40e68b : clone_quoting_options+0x3b/0x50 [/bin/ls]
0x4087e4 : main+0x3b4/0x1900 [/bin/ls]
0x3fa441ec5d : __libc_start_main+0xfd/0x1d0 [/lib64/libc-2.12.so]
0x402799 : _start+0x29/0x2c [/bin/ls]
0x4116c0 : xmalloc+0x0/0x20 [/bin/ls]
0x4116fc : xmemdup+0x1c/0x40 [/bin/ls]
0x40e68b : clone_quoting_options+0x3b/0x50 [/bin/ls]
0x40884a : main+0x41a/0x1900 [/bin/ls]
0x3fa441ec5d : __libc_start_main+0xfd/0x1d0 [/lib64/libc-2.12.so]
...```

For more details on the functions available for user-space stack backtraces, refer to ucontext-symbols.stp and ucontext-unwind.stp tapsets. You can also find the description of the functions in the aforementioned tapsets in the SystemTap Tapset Reference Manual.
Useful SystemTap Scripts

This chapter enumerates several SystemTap scripts you can use to monitor and investigate different subsystems. All of these scripts are available at `/usr/share/systemtap/testsuite/systemtap.examples/` once you install the `systemtap-testsuite` RPM.

5.1. Network

The following sections showcase scripts that trace network-related functions and build a profile of network activity.

5.1.1. Network Profiling

This section describes how to profile network activity. `nettop.stp` provides a glimpse into how much network traffic each process is generating on a machine.

```
nettop.stp

#!/usr/bin/env stap

global ifxmit, ifrecv
global ifmerged

probe netdev.transmit
{
    ifxmit[pid(), dev_name, execname(), uid()] <<< length
    ifmerged[pid(), dev_name, execname(), uid()] <<< 1
}

probe netdev.receive
{
    ifrecv[pid(), dev_name, execname(), uid()] <<< length
    ifmerged[pid(), dev_name, execname(), uid()] <<< 1
}

function print_activity()
{
    printf("%5s %5s %-12s %7s %7s %7s %7s %-15s\n", 
           "PID", "UID", "DEV", "XMIT_PK", "RECV_PK", 
           "XMIT_KB", "RECV_KB", "COMMAND")

    foreach ([pid, dev, exec, uid] in ifmerged-)
    {
        n_xmit = @count(ifxmit[pid, dev, exec, uid])
        n_recv = @count(ifrecv[pid, dev, exec, uid])
        printf("%5d %5d %-12s %7d %7d %7d %7d %-15s\n", 
               pid, uid, dev, n_xmit, n_recv, 
               @sum(ifxmit[pid, dev, exec, uid])/1024, 
               @sum(ifrecv[pid, dev, exec, uid])/1024, 
               exec)
    }

    print("\n")

    delete ifxmit
    delete ifrecv
```
Chapter 5. Useful SystemTap Scripts

```plaintext
delete ifmerged
}

probe timer.ms(5000), end, error
{
    print_activity()
}
```

Note that function `print_activity()` uses the following expressions:

```plaintext
n_xmit ? @sum(ifxmit[pid, dev, exec, uid])/1024 : 0
n_recv ? @sum(ifrecv[pid, dev, exec, uid])/1024 : 0
```

These expressions are if/else conditionals. The first statement is a more concise way of writing the following pseudo code:

```plaintext
if n_recv != 0 then
    @sum(ifrecv[pid, dev, exec, uid])/1024
else
    0
```

`nettop.stp` tracks which processes are generating network traffic on the system, and provides the following information about each process:

- **PID** — the ID of the listed process.
- **UID** — user ID. A user ID of 0 refers to the root user.
- **DEV** — which ethernet device the process used to send / receive data (for example, eth0, eth1)
- **XMIT_PK** — number of packets transmitted by the process
- **RECV_PK** — number of packets received by the process
- **XMIT_KB** — amount of data sent by the process, in kilobytes
- **RECV_KB** — amount of data received by the service, in kilobytes

`nettop.stp` provides network profile sampling every 5 seconds. You can change this setting by editing `probe timer.ms(5000)` accordingly. *Example 5.1, “nettop.stp Sample Output”* contains an excerpt of the output from `nettop.stp` over a 20-second period:

### Example 5.1. nettop.stp Sample Output

```
[...]
<table>
<thead>
<tr>
<th>PID</th>
<th>UID</th>
<th>DEV</th>
<th>XMIT_PK</th>
<th>RECV_PK</th>
<th>XMIT_KB</th>
<th>RECV_KB</th>
<th>COMMAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>eth0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>swapper</td>
</tr>
<tr>
<td>11178</td>
<td>0</td>
<td>eth0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>synergyc</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PID</th>
<th>UID</th>
<th>DEV</th>
<th>XMIT_PK</th>
<th>RECV_PK</th>
<th>XMIT_KB</th>
<th>RECV_KB</th>
<th>COMMAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>2886</td>
<td>4</td>
<td>eth0</td>
<td>79</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>cups-polld</td>
</tr>
<tr>
<td>11362</td>
<td>0</td>
<td>eth0</td>
<td>0</td>
<td>61</td>
<td>0</td>
<td>5</td>
<td>firefox</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>eth0</td>
<td>3</td>
<td>32</td>
<td>0</td>
<td>3</td>
<td>swapper</td>
</tr>
</tbody>
</table>
```
5.1.2. Tracing Functions Called in Network Socket Code

This section describes how to trace functions called from the kernel's `net/socket.c` file. This task helps you identify, in finer detail, how each process interacts with the network at the kernel level.

socket-trace.stp

```stap
#if /usr/bin/env stap

probe kernel.function("@net/socket.c").call {
    printf ("%s -> %s\n", thread_indent(1), ppfunc());
}

probe kernel.function("@net/socket.c").return {
    printf ("%s <- %s\n", thread_indent(-1), ppfunc());
}

socket-trace.stp
```

socket-trace.stp is identical to Example 3.6, "thread_indent.stp", which was earlier used in SystemTap Functions to illustrate how thread_indent() works.

Example 5.2. socket-trace.stp Sample Output

```plaintext
[...]
0 Xorg(3611): -> sock_poll
3 Xorg(3611): <- sock_poll
0 Xorg(3611): -> sock_poll
3 Xorg(3611): <- sock_poll
0 gnome-terminal(11106): -> sock_poll
5 gnome-terminal(11106): <- sock_poll
0 scim-bridge(3883): -> sock_poll
3 scim-bridge(3883): <- sock_poll
0 scim-bridge(3883): -> sys_socketcall
4 scim-bridge(3883): -> sys_recv
8 scim-bridge(3883): -> sys_recvfrom
12 scim-bridge(3883):-> sock_from_file
16 scim-bridge(3883):< sock_from_file
20 scim-bridge(3883):-> sock_recvmsg
24 scim-bridge(3883):< sock_recvmsg
28 scim-bridge(3883): <- sys_recvfrom
31 scim-bridge(3883): <- sys_recv
```
Chapter 5. Useful SystemTap Scripts

Example 5.2, “socket-trace.stp Sample Output” contains a 3-second excerpt of the output for socket-trace.stp. For more information about the output of this script as provided by thread_indent(), refer to SystemTap Functions Example 3.6, “thread-indent.stp”.

5.1.3. Monitoring Incoming TCP Connections

This section illustrates how to monitor incoming TCP connections. This task is useful in identifying any unauthorized, suspicious, or otherwise unwanted network access requests in real time.

tcp_connections.stp

```
#!/usr/bin/env stap

probe begin {
    printf("%6s %16s %6s %6s %16s\n",
           "UID", "CMD", "PID", "PORT", "IP_SOURCE")
}

probe kernel.{function("tcp_accept"),function("inet_csk_accept")}.return? {
    sock = $return
    if (sock != 0)
        printf("%6d %16s %6d %6d %16s\n", uid(), execname(), pid(),
               inet_get_local_port(sock), inet_get_ip_source(sock))
}
```

While tcp_connections.stp is running, it will print out the following information about any incoming TCP connections accepted by the system in real time:

- Current **UID**
- **CMD** - the command accepting the connection
- **PID** of the command
- Port used by the connection
- IP address from which the TCP connection originated

Example 5.3. tcp_connections.stp Sample Output

<table>
<thead>
<tr>
<th>UID</th>
<th>CMD</th>
<th>PID</th>
<th>PORT</th>
<th>IP_SOURCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>sshd</td>
<td>3165</td>
<td>22</td>
<td>10.64.0.227</td>
</tr>
<tr>
<td>0</td>
<td>sshd</td>
<td>3165</td>
<td>22</td>
<td>10.64.0.227</td>
</tr>
</tbody>
</table>

5.1.4. Monitoring TCP Packets
This section illustrates how to monitor TCP packets received by the system. This is useful in analyzing network traffic generated by applications running on the system.

**tcpdumplike.stp**

```bash
#! /usr/bin/env stap

// A TCP dump like example

probe begin, timer.s(1) {
    printf("-----------------------------------------------------------------
")
    printf("       Source IP         Dest IP  SPort  DPort  U  A  P  R  S  F \n")
    printf("-----------------------------------------------------------------
")
}

probe udp.recvmsg /* ,udp.sendmsg */ {
    printf("  %15s %15s  %5d  %5d  %d  %d  %d  %d  %d  %d\n",
        saddr, daddr, sport, dport)
}

probe tcp.receive {
    printf("  %15s %15s  %5d  %5d  %d  %d  %d  %d  %d  %d\n",
        saddr, daddr, sport, dport, urg, ack, psh, rst, syn, fin)
}
```

While *tcpdumplike.stp* is running, it will print out the following information about any received TCP packets in real time:

- Source and destination IP address (saddr, daddr, respectively)
- Source and destination ports (sport, dport, respectively)
- Packet flags

To determine the flags used by the packet, *tcpdumplike.stp* uses the following functions:

- **urg** - urgent
- **ack** - acknowledgement
- **psh** - push
- **rst** - reset
- **syn** - synchronize
- **fin** - finished

The aforementioned functions return 1 or 0 to specify whether the packet uses the corresponding flag.

**Example 5.4. tcpdumplike.stp Sample Output**

<table>
<thead>
<tr>
<th>Source IP</th>
<th>Dest IP</th>
<th>SPort</th>
<th>DPort</th>
<th>U</th>
<th>A</th>
<th>P</th>
<th>R</th>
<th>S</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>209.85.229.147</td>
<td>10.0.2.15</td>
<td>80</td>
<td>20373</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

43
5.1.5. Monitoring Network Packets Drops in Kernel

The network stack in Linux can discard packets for various reasons. Some Linux kernels include a tracepoint, `kernel.trace("kfree_skb")`, which easily tracks where packets are discarded. `dropwatch.stp` uses `kernel.trace("kfree_skb")` to trace packet discards; the script summarizes which locations discard packets every five-second interval.

`dropwatch.stp`

```
#!/usr/bin/env stap

# A script to mimic the behavior of the dropwatch utility
# http://fedorahosted.org/dropwatch

# Array to hold the list of drop points we find
global locations

# Note when we turn the monitor on and off
probe begin { printf("Monitoring for dropped packets\n") }
probe end { printf("Stopping dropped packet monitor\n") }

# Increment a drop counter for every location we drop at
probe kernel.trace("kfree_skb") { locations[$location] <<< 1 }

# Every 5 seconds report our drop locations
probe timer.sec(5)
{
    printf("%s %s
", ctime(gettimeofday_s()),
        "\n    foreach (l in locations-)
        printf("%d packets dropped at %s\",
            @count(locations[l]), symdata(l))
    }
    delete locations
}
```

The `kernel.trace("kfree_skb")` traces which places in the kernel drop network packets. The `kernel.trace("kfree_skb")` has two arguments: a pointer to the buffer being freed (`$skb`) and the location in kernel code the buffer is being freed (`$location`). The `dropwatch.stp` script provides the function containing `$location` where possible. The information to map `$location` back to the function...
is not in the instrumentation by default. On SystemTap 1.4 the `--all-modules` option will include the required mapping information and the following command can be used to run the script:

```
stap --all-modules dropwatch.stp
```

On older versions of SystemTap you can use the following command to emulate the `--all-modules` option:

```
stap -dkernel \
`cat /proc/modules | awk 'BEGIN { ORS = " " } {print "-d"$1}''` dropwatch.stp
```

Running the dropwatch.stp script 15 seconds would result in output similar in Example 5.5, “dropwatch.stp Sample Output”. The output lists the number of misses for each tracepoint location with either the function name or the address.

**Example 5.5. dropwatch.stp Sample Output**

```
Monitoring for dropped packets

Tue Nov 17 00:26:51 2020  
1762 packets dropped at unix_stream_recvmsg
4 packets dropped at tun_do_read 
2 packets dropped at nf_hook_slow

Tue Nov 17 00:26:56 2020  
467 packets dropped at unix_stream_recvmsg
20 packets dropped at nf_hook_slow 
6 packets dropped at tun_do_read

Tue Nov 17 00:27:01 2020  
446 packets dropped at unix_stream_recvmsg
4 packets dropped at tun_do_read 
4 packets dropped at nf_hook_slow
Stopping dropped packet monitor
```

When the script is being compiled on one machine and run on another the `--all-modules` and `/proc/modules` directory are not available; the `symname` function will just print out the raw address. To make the raw address of packet drops more meaningful, refer to the `/boot/System.map-`uname -r` file. This file lists the starting addresses for each function, allowing you to map the addresses in the output of Example 5.5, “dropwatch.stp Sample Output” to a specific function name. Given the following snippet of the `/boot/System.map-`uname -r` file, the address 0xffffffff8149a8ed maps to the function `unix_stream_recvmsg`:

```
[...]
0xffffffff8149a420 t unix_dgram_poll
0xffffffff8149a5e0 t unix_stream_recvmsg
0xffffffff8149ad00 t unix_find_other
[...]
```

### 5.2. Disk

The following sections showcase scripts that monitor disk and I/O activity.
5.2.1. Summarizing Disk Read/Write Traffic

This section describes how to identify which processes are performing the heaviest disk reads/writes to the system.

**disktop.stp**

```bash
#!/usr/bin/env stap
#
# Copyright (C) 2007 Oracle Corp.
#
# Get the status of reading/writing disk every 5 seconds,
# output top ten entries
#
# This is free software, GNU General Public License (GPL);
# either version 2, or (at your option) any later version.
#
# Usage:
#  ./disktop.stp
#

global io_stat, device
global read_bytes, write_bytes

probe vfs.read.return {
    if (returnval()>0) {
        if (devname!="N/A") {/*skip read from cache*/
            io_stat[pid(),execname(),uid(),ppid(),"R"] += returnval()
            device[pid(),execname(),uid(),ppid(),"R"] = devname
            read_bytes += returnval()
        }
    }
}

probe vfs.write.return {
    if (returnval()>0) {
        if (devname!="N/A") { /*skip update cache*/
            io_stat[pid(),execname(),uid(),ppid(),"W"] += returnval()
            device[pid(),execname(),uid(),ppid(),"W"] = devname
            write_bytes += returnval()
        }
    }
}

probe timer.ms(5000) {
    /* skip non-read/write disk */
    if (read_bytes+write_bytes) {
        printf("%-25s, %-8s%4dKb/sec, %-7s%6dKb, %-7s%6dKb

", ctime(gettimeofday_s()),
        "Average:", ((read_bytes+write_bytes)/1024)/5,
        "Read: ",read_bytes/1024,
        "Write: ",write_bytes/1024)
    }
    /* print header */
    printf("%8s %8s %8s %25s %8s %4s %12s
",
        "UID", "PID", "PPID", "CMD", "DEVICE", "T", "BYTES")
}
/* print top ten I/O */
```

Summarizing Disk Read/Write Traffic

```c
foreach ([process, cmd, userid, parent, action] in io_stat | limit 10)
    printf("%8d %8d %8d %25s %8s %4s %12d\n", 
         userid, process, parent, cmd, 
         device[process, cmd, userid, parent, action], 
         action, io_stat[process, cmd, userid, parent, action])

/* clear data */
delete io_stat
delete device
read_bytes = 0
write_bytes = 0
}

probe end{
    delete io_stat
    delete device
    delete read_bytes
    delete write_bytes
}
```

disktop.stp outputs the top ten processes responsible for the heaviest reads/writes to disk. Example 5.6, “disktop.stp Sample Output” displays a sample output for this script, and includes the following data per listed process:

- **UID** — user ID. A user ID of 0 refers to the root user.
- **PID** — the ID of the listed process.
- **PPID** — the process ID of the listed process’s *parent process*.
- **CMD** — the name of the listed process.
- **DEVICE** — which storage device the listed process is reading from or writing to.
- **T** — the type of action performed by the listed process; W refers to write, while R refers to read.
- **BYTES** — the amount of data read to or written from disk.

The time and date in the output of disktop.stp is returned by the functions `ctime()` and `gettimeofday_s()`. `ctime()` derives calendar time in terms of seconds passed since the Unix epoch (January 1, 1970). `gettimeofday_s()` counts the actual number of seconds since Unix epoch, which gives a fairly accurate human-readable timestamp for the output.

In this script, the `$return` is a local variable that stores the actual number of bytes each process reads or writes from the virtual file system. `$return` can only be used in return probes (for example, `vfs.read.return` and `vfs.read.return`).

Example 5.6. disktop.stp Sample Output

```
[...
Mon Sep 29 03:38:28 2008 , Average: 19Kb/sec, Read: 7Kb, Write: 89Kb
UID    PID    PPID    CMD    DEVICE    T    BYTES
0    26319    26294    firefox    sda5    W    90229
```
5.2.2. Tracking I/O Time For Each File Read or Write

This section describes how to monitor the amount of time it takes for each process to read from or write to any file. This is useful to determine what files are slow to load on a given system.

**iotime.stp**

```bash
#!/usr/bin/env stap

/*
* Copyright (C) 2006-2018 Red Hat Inc.
*
* This copyrighted material is made available to anyone wishing to use,
* modify, copy, or redistribute it subject to the terms and conditions
* of the GNU General Public License v.2.
*
* You should have received a copy of the GNU General Public License
* along with this program.  If not, see <http://www.gnu.org/licenses/>.
*
* Print out the amount of time spent in the read and write systemcall
* when each file opened by the process is closed. Note that the systemtap
* script needs to be running before the open operations occur for
* the script to record data.
*
* This script could be used to to find out which files are slow to load
* on a machine. e.g.
*
* stap iotime.stp -c 'firefox'
*
* Output format is:
* timestamp pid (executabale) info_type path ...
*
* 200283135 2573 (cupsd) access /etc/printcap read: 0 write: 7063
* 200283143 2573 (cupsd) iotime /etc/printcap time: 69
*
*/

global start
global time_io

function timestamp:long() { return gettimeofday_us() - start }

function proc:string() { return sprintf("%d (%s)", pid, execname) }

probe begin { start = gettimeofday_us() }

global possible_filename, filehandles, fileread, filewrite
```

```bash
Mon Sep 29 03:38:38 2008 , Average: 1Kb/sec, Read: 7Kb, Write: 1Kb

<table>
<thead>
<tr>
<th>UID</th>
<th>PID</th>
<th>PPID</th>
<th>CMD</th>
<th>DEVICE</th>
<th>T</th>
<th>BYTES</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2758</td>
<td>2757</td>
<td>pam_timestamp_c</td>
<td>sda5</td>
<td>R</td>
<td>8064</td>
</tr>
<tr>
<td>0</td>
<td>2885</td>
<td>1</td>
<td>cupsd</td>
<td>sda5</td>
<td>W</td>
<td>1678</td>
</tr>
</tbody>
</table>
```

```bash
0     2758     2757           pam_timestamp_c     sda5    R         8064
0     2885        1                     cupsd     sda5    W         1678
```
Tracking I/O Time For Each File Read or Write

```c
probe syscall.open, syscall.openat {
    possible_filename[tid()] = filename
}

probe syscall.open.return, syscall.openat.return {
    // Get filename even if non-dwarf syscall return probe are used.
    filename = possible_filename[tid()]
    delete possible_filename[tid()]
    if (retval != -1) {
        filehandles[pid(), retval] = filename
    } else {
        printf("%d %s access %s fail\n", timestamp(), proc(), filename)
    }
}

global read_fds, write_fds

probe syscall.read { read_fds[tid()] = fd }

probe syscall.read.return {
    p = pid()
    // Get fd even if non-dwarf syscall return probe.
    fd = read_fds[tid()]
    delete read_fds[tid()]
    bytes = retval
    time = gettimeofday_us() - @entry(gettimeofday_us())
    if (bytes > 0)
        fileread[p, fd] <<< bytes
time_io[p, fd] <<< time
}

probe syscall.write { write_fds[tid()] = fd }

probe syscall.write.return {
    p = pid()
    // Get fd even if non-dwarf syscall return probe.
    fd = write_fds[tid()]
    delete write_fds[tid()]
    bytes = retval
    time = gettimeofday_us() - @entry(gettimeofday_us())
    if (bytes > 0)
        filewrite[p, fd] <<< bytes
time_io[p, fd] <<< time
}

probe syscall.close {
    if ([pid(), fd] in filehandles) {
        printf("%d %s access %s read: %d write: %d\n",
             timestamp(), proc(), filehandles[pid(), fd],
             @sum(fileread[pid(), fd]), @sum(filewrite[pid(), fd]))
        if (@count(time_io[pid(), fd]))
            printf("%d %s iotime %s time: %d\n", timestamp(), proc(),
                   filehandles[pid(), fd], @sum(time_io[pid(), fd]))
    }
    delete fileread[pid(), fd]
    delete filewrite[pid(), fd]
    delete filehandles[pid(), fd]
    delete time_io[pid(), fd]
}
```
### Chapter 5. Useful SystemTap Scripts

*iotime.stp* tracks each time a system call opens, closes, reads from, and writes to a file. For each file any system call accesses, *iotime.stp* counts the number of microseconds it takes for any reads or writes to finish and tracks the amount of data (in bytes) read from or written to the file.

*iotime.stp* also uses the local variable $count to track the amount of data (in bytes) that any system call attempts to read or write. Note that $return (as used in *disktop.stp* from Section 5.2.1, “Summarizing Disk Read/Write Traffic”) stores the actual amount of data read/written. $count can only be used on probes that track data reads or writes (that is, syscall.read and syscall.write).

---

**Example 5.7. *iotime.stp* Sample Output**

```plaintext
[...]
825946 3364 (NetworkManager) access /sys/class/net/eth0/carrier read: 8190 write: 0
825955 3364 (NetworkManager) iotime /sys/class/net/eth0/carrier time: 9
[...]
117061 2460 (pcscd) access /dev/bus/usb/003/001 read: 43 write: 0
117065 2460 (pcscd) iotime /dev/bus/usb/003/001 time: 7
[...]
397373 2886 (sendmail) access /proc/loadavg read: 4096 write: 0
397374 2886 (sendmail) iotime /proc/loadavg time: 11
[...]
```

---

**Example 5.7, “*iotime.stp* Sample Output”** prints out the following data:

- A timestamp, in microseconds.
- Process ID and process name.
- An access or iotime flag.
- The file accessed.

If a process was able to read or write any data, a pair of access and iotime lines should appear together. The access line's timestamp refers to the time that a given process started accessing a file; at the end of the line, it will show the amount of data read/written (in bytes). The iotime line will show the amount of time (in microseconds) that the process took in order to perform the read or write.

If an access line is not followed by an iotime line, it means that the process did not read or write any data.

---

### 5.2.3. Track Cumulative IO

This section describes how to track the cumulative amount of I/O to the system.

**traceio.stp**

```plaintext
#!/usr/bin/env stap
# traceio.stp
# Copyright (C) 2007-2018 Red Hat, Inc., Eugene Teo <eteo@redhat.com>
```

---

50
traceio.stp prints the top ten executables generating I/O traffic over time. In addition, it also tracks the cumulative amount of I/O reads and writes done by those ten executables. This information is tracked and printed out in 1-second intervals, and in descending order.

Note that traceio.stp also uses the local variable $return, which is also used by desktop.stp from Section 5.2.1, “Summarizing Disk Read/Write Traffic”.

Example 5.8. traceio.stp Sample Output

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Xorg r:</td>
<td>583401</td>
<td>KiB</td>
</tr>
<tr>
<td>w:</td>
<td>0</td>
<td>KiB</td>
</tr>
</tbody>
</table>
5.2.4. I/O Monitoring (By Device)

This section describes how to monitor I/O activity on a specific device.

```
traceio2.stp

#! /usr/bin/env stap

global device_of_interest

probe begin {
   /* The following is not the most efficient way to do this.
      One could directly put the result of usrdev2kerndev()
      into device_of_interest. However, want to test out
      the other device functions */
   dev = usrdev2kerndev($1)
   device_of_interest = MKDEV(MAJOR(dev), MINOR(dev))
}

probe vfs.{write,read} {
   if (dev == device_of_interest)
      printf("%s(%d) %s 0x%x
",
         execname(), pid(), ppfunc(), dev)
}
```

`traceio2.stp` takes 1 argument: the whole device number. To get this number, use `stat -c "0x%D" directory`, where `directory` is located in the device to be monitored.

The `usrdev2kerndev()` function converts the whole device number into the format understood by the kernel. The output produced by `usrdev2kerndev()` is used in conjunction with the `MKDEV()`, `MINOR()`, and `MAJOR()` functions to determine the major and minor numbers of a specific device.
The output of `traceio2.stp` includes the name and ID of any process performing a read/write, the function it is performing (that is, `vfs_read` or `vfs_write`), and the kernel device number.

The following example is an excerpt from the full output of `stap traceio2.stp 0x805`, where `0x805` is the whole device number of /home. /home resides in /dev/sda5, which is the device we wish to monitor.

```
Example 5.9. traceio2.stp Sample Output

[...]  
synergyc(3722) vfs_read 0x800005  
synergyc(3722) vfs_read 0x800005  
cupsd(2889) vfs_write 0x800005  
cupsd(2889) vfs_write 0x800005  
cupsd(2889) vfs_write 0x800005  
[...]  
```

5.2.5. Monitoring Reads and Writes to a File

This section describes how to monitor reads from and writes to a file in real time.

inodewatch.stp

```bash
#!/usr/bin/env stap
probe vfs.{write,read}
{
    # dev and ino are defined by vfs.write and vfs.read
    if (dev == MKDEV($1,$2) # major/minor device
        && ino == $3)
        printf (%s(%d) %s 0x%x/%u

            execname(), pid(), ppfunc(), dev, ino)

        }
```

`inodewatch.stp` takes the following information about the file as arguments on the command line:

- The file's major device number.
- The file's minor device number.
- The file's `inode` number.

To get this information, use `stat -c '%D %i' filename`, where `filename` is an absolute path.

For instance: to monitor /etc/crontab, run `stat -c '%D %i' /etc/crontab` first. This gives the following output:

```
805 1078319
```
805 is the base-16 (hexadecimal) device number. The lower two digits are the minor device number and the upper digits are the major number. 1078319 is the inode number. To start monitoring /etc/crontab, run stap inodewatch.stp 0x8 0x05 1078319 (The 0x prefixes indicate base-16 values.

The output of this command contains the name and ID of any process performing a read/write, the function it is performing (that is, vfs_read or vfs_write), the device number (in hex format), and the inode number. Example 5.10, “inodewatch.stp Sample Output” contains the output of stap inodewatch.stp 0x8 0x05 1078319 (when cat /etc/crontab is executed while the script is running):

Example 5.10. inodewatch.stp Sample Output

```
cat(16437) vfs_read 0x800005/1078319
cat(16437) vfs_read 0x800005/1078319
```

### 5.2.6. Monitoring Changes to File Attributes

This section describes how to monitor if any processes are changing the attributes of a targeted file, in real time.

**inodewatch2.stp**

```
#!/usr/bin/env stap

global ATTR_MODE = 1

probe kernel.function("notify_change") {
    dev_nr = $dentry->d_inode->i_sb->s_dev
    inode_nr = $dentry->d_inode->i_ino

    if (dev_nr == MKDEV($1,$2) # major/minor device
        && inode_nr == $3
        && $attr->ia_valid & ATTR_MODE)
        printf("%s(%d) %s 0x%x/%u %o %d\n",
            execname(), pid(), ppfunc(), dev_nr, inode_nr, $attr->ia_mode, uid())
}
```

Like inodewatch.stp from Section 5.2.5, “Monitoring Reads and Writes to a File”, inodewatch2.stp takes the targeted file’s device number (in integer format) and inode number as arguments. For more information on how to retrieve this information, refer to Section 5.2.5, “Monitoring Reads and Writes to a File”.

The output for inodewatch2.stp is similar to that of inodewatch.stp, except that inodewatch2.stp also contains the attribute changes to the monitored file, as well as the ID of the user responsible (uid()). Example 5.11, “inodewatch2.stp Sample Output” shows the output of inodewatch2.stp while monitoring /home/joe/bigfile when user joe executes chmod 777 /home/joe/bigfile and chmod 666 /home/joe/bigfile.

Example 5.11. inodewatch2.stp Sample Output

```
chmod(17448) inode_setattr 0x800005/6011835 100777 500
```
5.2.7. Periodically Print I/O Block Time

This section describes how to track the amount of time each block I/O requests spends waiting for completion. This is useful in determining whether there are too many outstanding block I/O operations at any given time.

`ioblktime.stp`

```bash
#!/usr/bin/env stap

global req_time[25000], etimes

probe ioblock.request
{
    req_time[$bio] = gettimeofday_us()
}

probe ioblock.end
{
    t = gettimeofday_us()
    s = req_time[$bio]
    delete req_time[$bio]
    if (s) {
        etimes[devname, bio_rw_str(rw)] <<< t - s
    }
}

/* for time being delete things that get merged with others */
probe kernel.(trace("block_bio_frontmerge"),
    trace("block_bio_backmerge"))
{
    delete req_time[$bio]
}

probe timer.s(10), end {
    ansi_clear_screen()
    printf("%10s %3s %10s %10s %10s
",
        "device", "rw", "total (us)", "count", "avg (us)"
    foreach ([dev,rw] in etimes - limit 20) {
        printf("%10s %3s %10d %10d %10d
", dev, rw,
            @sum(etimes[dev,rw]), @count(etimes[dev,rw]), @avg(etimes[dev,rw]))
    }
    delete etimes
}
```

`ioblktime.stp` computes the average waiting time for block I/O per device, and prints a list every 10 seconds. As always, you can revise this refresh rate by editing the specified value in `probe timer.s(10), end {`.

In some cases, there can be too many outstanding block I/O operations, at which point the script can exceed the default number of `MAXMAPENTRIES`. `MAXMAPENTRIES` is the maximum number of rows in an array if the array size is not specified explicitly when declared. If the script exceeds the default `MAXMAPENTRIES` value of 2048, run the script again with the `stap` option `--DMAXMAPENTRIES=10000`.

```bash
chmod(17449) inode_setattr 0x800005/6011835 100666 500
```
Example 5.12. *ioblktime.stp* Sample Output

<table>
<thead>
<tr>
<th>device</th>
<th>rw</th>
<th>total (us)</th>
<th>count</th>
<th>avg (us)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sda</td>
<td>W</td>
<td>9659</td>
<td>6</td>
<td>1609</td>
</tr>
<tr>
<td>dm-0</td>
<td>W</td>
<td>20278</td>
<td>6</td>
<td>3379</td>
</tr>
<tr>
<td>dm-0</td>
<td>R</td>
<td>20524</td>
<td>5</td>
<td>4104</td>
</tr>
<tr>
<td>sda</td>
<td>R</td>
<td>19277</td>
<td>5</td>
<td>3855</td>
</tr>
</tbody>
</table>

*Example 5.12, “ioblktime.stp Sample Output”* displays the device name, operations performed (*rw*), total wait time of all operations (*total (us)*), number of operations (*count*), and average wait time for all those operations (*avg (us)*). The times tallied by the script are in microseconds.

## 5.3. Profiling

The following sections showcase scripts that profile kernel activity by monitoring function calls.

### 5.3.1. Counting Function Calls Made

This section describes how to identify how many times the system called a specific kernel function in a 30-second sample. Depending on the use of wildcards, you can also use this script to target multiple kernel functions.

*functioncallcount.stp*

```bash
#!/usr/bin/env stap
# The following line command will probe all the functions
# in kernel's memory management code:
#
# stap  functioncallcount.stp "*@mm/*.c"

probe kernel.function($1).call {  # probe functions listed on commandline
    called[ppfunc()] <<< 1  # add a count efficiently
}

global called

probe end {
    foreach (fn in called-)  # Sort by call count (in decreasing order)
        (fn+ in called)  # Sort by function name
        printf("%s %d\n", fn, @count(called[fn]))
    exit()
}
```

*functioncallcount.stp* takes the targeted kernel function as an argument. The argument supports wildcards, which enables you to target multiple kernel functions up to a certain extent.

The output of *functioncallcount.stp* contains the name of the function called and how many times it was called during the sample time (in alphabetical order). *Example 5.13, “functioncallcount.stp Sample Output”* contains an excerpt from the output of `stap functioncallcount.stp "*@mm/*.c"`:  

```
Example 5.13. functioncallcount.stp Sample Output

[...]
__vma_link 97
__vma_link_file 66
__vma_link_list 97
__vma_link_rb 97
__xchg 103
add_page_to_active_list 102
add_page_to_inactive_list 19
add_to_page_cache 19
add_to_page_cache_lru 7
all_vm_events 6
alloc_pages_node 4630
alloc_slabmgmt 67
anon_vma_alloc 62
anon_vma_free 62
anon_vma_lock 66
anon_vma_prepare 98
anon_vma_unlink 97
anon_vma_unlock 66
arch_get_unmapped_area_topdown 94
arch_get_unmapped_exec_area 3
arch_unmap_area_topdown 97
atomic_add 2
atomic_add_negative 97
atomic_dec_and_test 5153
atomic_inc 470
atomic_inc_and_test 1
[...]

5.3.2. Call Graph Tracing

This section describes how to trace incoming and outgoing function calls.

para-callgraph.stp

```bash
#!/usr/bin/env stap

function trace(entry_p, extra) {
  #( $# > 1 ? if (tid() in trace) %)
  printf("%s\n", thread_indent (entry_p),
         (entry_p>0?"->":*"<"),
    ppfunc (),
    extra)
}

%( $# > 1 ?
global trace
probe $2.call {
  trace[tid()] = 1
}
probe $2.return {
  delete trace[tid()]
}```
Chapter 5. Useful SystemTap Scripts

```plaintext
para-callgraph.stp takes two command-line arguments:

- The function/s whose entry/exit call you’d like to trace ($1).
- A second optional trigger function ($2), which enables or disables tracing on a per-thread basis. Tracing in each thread will continue as long as the trigger function has not exited yet.

para-callgraph.stp uses `thread_indent()`; as such, its output contains the timestamp, process name, and thread ID of $1 (that is, the probe function you are tracing). For more information about `thread_indent()`, refer to its entry in SystemTap Functions.

The following example contains an excerpt from the output for stap para-callgraph.stp 'kernel.function("*@fs/*.c")' 'kernel.function("sys_read")':

```
Example 5.14. para-callgraph.stp Sample Output
```

```
[...]
267 gnome-terminal(2921): <-do_sync_read return=0xfffffffffffffff5
269 gnome-terminal(2921): <-vfs_read return=0xfffffffffffffff5
  0 gnome-terminal(2921):->fput file=0xffff880111eebbc0
  2 gnome-terminal(2921):<-fput
  0 gnome-terminal(2921):->fget_light fd=0x3 fput_needed=0xfffff88010544df54
  3 gnome-terminal(2921):<-fget_light return=0xffff8801116ce980
  0 gnome-terminal(2921):->vfs_read file=0xffff8801116ce980 buf=0xc86504 count=0x1000
  pos=0xfffff88010544df48
  4 gnome-terminal(2921): ->rw_verify_area read_write=0x0 file=0xfffff8801116ce980
  pp=os=0xfffff88010544df48 count=0x1000
  7 gnome-terminal(2921): <-rw_verify_area return=0x1000
  12 gnome-terminal(2921): <-do_sync_read filep=0xfffff8801116ce980 buf=0xc86504 len=0x1000
  pp=os=0xfffff88010544df48
  15 gnome-terminal(2921): <-do_sync_read return=0xfffffffffffffff5
  18 gnome-terminal(2921):<-vfs_read return=0xfffffffffffffff5
  0 gnome-terminal(2921):->fput file=0xfffff8801116ce980
```

5.3.3. Determining Time Spent in Kernel and User Space

This section illustrates how to determine the amount of time any given thread is spending in either kernel or user-space.

```
thread-times.stp
```

```plaintext
#!/usr/bin/env stap

```
Determining Time Spent in Kernel and User Space

thread-times.stp lists the top 20 processes currently taking up CPU time within a 5-second sample, along with the total number of CPU ticks made during the sample. The output of this script also notes the percentage of CPU time each process used, as well as whether that time was spent in kernel space or user space.

Example 5.15, “thread-times.stp Sample Output” contains a 5-second sample of the output for thread-times.stp:

```
Example 5.15. thread-times.stp Sample Output

<table>
<thead>
<tr>
<th>tid</th>
<th>%user</th>
<th>%kernel (of 20002 ticks)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.00%</td>
<td>87.88%</td>
</tr>
<tr>
<td>32169</td>
<td>5.24%</td>
<td>0.03%</td>
</tr>
<tr>
<td>9815</td>
<td>3.33%</td>
<td>0.36%</td>
</tr>
<tr>
<td>9859</td>
<td>0.95%</td>
<td>0.00%</td>
</tr>
<tr>
<td>3611</td>
<td>0.56%</td>
<td>0.12%</td>
</tr>
<tr>
<td>9861</td>
<td>0.62%</td>
<td>0.01%</td>
</tr>
<tr>
<td>11106</td>
<td>0.37%</td>
<td>0.02%</td>
</tr>
</tbody>
</table>
```
### 5.3.4. Monitoring Polling Applications

This section describes how to identify and monitor which applications are polling. Doing so allows you to track unnecessary or excessive polling, which can help you pinpoint areas for improvement in terms of CPU usage and power savings.

#### timeout.stp

```plaintext
#!/usr/bin/env stap
#
# Copyright (C) 2009-2018 Red Hat, Inc.
# Written by Ulrich Drepper <drepper@redhat.com>
# Modified by William Cohen <wcohen@redhat.com>

global process, timeout_count, to
global poll_timeout, epoll_timeout, select_timeout, itimer_timeout
global nanosleep_timeout, futex_timeout, signal_timeout

probe syscall.{poll, epoll_wait} {
    if (timeout) to[pid()]=timeout
}

probe syscall.poll.return {
    if (retval == 0 && to[pid()] > 0 ) {
        poll_timeout[pid()]++
        timeout_count[pid()]++
        process[pid()] = execname()
        delete to[pid()]
    }
}

probe syscall.epoll_wait.return {
    if (retval == 0 && to[pid()] > 0 ) {
        epoll_timeout[pid()]++
        timeout_count[pid()]++
        process[pid()] = execname()
        delete to[pid()]
    }
}

probe syscall.select.return {
    if (retval == 0) {
        select_timeout[pid()]++
        timeout_count[pid()]++
    }
}
```
process[pid()] = execname()
}
}

probe syscall.futex.return {
    if (errno_str(retval) == "ETIMEDOUT") {
        futex_timeout[pid()]++
        timeout_count[pid()]++
        process[pid()] = execname()
    }
}

probe syscall.nanosleep.return {
    if (retval == 0) {
        nanosleep_timeout[pid()]++
        timeout_count[pid()]++
        process[pid()] = execname()
    }
}

probe kernel.function("it_real_fn") {
    itimer_timeout[pid()]++
    timeout_count[pid()]++
    process[pid()] = execname()
}

probe syscall.rt_sigtimedwait.return {
    if (errno_str(retval) == "EAGAIN") {
        signal_timeout[pid()]++
        timeout_count[pid()]++
        process[pid()] = execname()
    }
}

probe syscall.exit {
    if (pid() in process) {
        delete process[pid()]
        delete timeout_count[pid()]
        delete poll_timeout[pid()]
        delete epoll_timeout[pid()]
        delete select_timeout[pid()]
        delete itimer_timeout[pid()]
        delete futex_timeout[pid()]
        delete nanosleep_timeout[pid()]
        delete signal_timeout[pid()]
    }
}

probe timer.s(1) {
    ansi_clear_screen()
    printf(" pid | poll select epoll itimer futex nanosleep signal| process\n")
    foreach (p in timeout_count- limit 20) {
        printf("%5d |%7d %7d %7d %7d %7d %7d | %.38s\n", p,
            poll_timeout[p], select_timeout[p],
            epoll_timeout[p], itimer_timeout[p],
            futex_timeout[p], nanosleep_timeout[p],
            signal_timeout[p], process[p])
    }
}

global prom_arr
Chapter 5. Useful SystemTap Scripts

probe prometheus {
    foreach (p in timeout_count limit 20) {
        prom_arr[poll_timeout[p], select_timeout[p],
        epoll_timeout[p], itimer_timeout[p],
        futex_timeout[p], nanosleep_timeout[p],
        signal_timeout[p]] = process[p]
    }
    @prometheus_dump_array7(prom_arr, "process_timeouts", "poll_timeout", "select_timeout",
        "epoll_timeout", "itimer_timeout",
        "futex_timeout", "nanosleep_timeout",
        "signal_timeout")
    delete prom_arr
}

timeout.stp tracks how many times each of the following system calls completed due to time expiring rather than due to an actual event occurring:

- poll
- select
- epoll
- itimer
- futex
- nanosleep
- signal

Example 5.16. timeout.stp Sample Output

<table>
<thead>
<tr>
<th>uid</th>
<th>poll</th>
<th>select</th>
<th>epoll</th>
<th>itimer</th>
<th>futex</th>
<th>nanosleep</th>
<th>signal</th>
<th>process</th>
</tr>
</thead>
<tbody>
<tr>
<td>28937</td>
<td>148793</td>
<td>0</td>
<td>0</td>
<td>4727</td>
<td>37288</td>
<td>0</td>
<td>0</td>
<td>firefox</td>
</tr>
<tr>
<td>22945</td>
<td>0</td>
<td>56949</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>scim-bridge</td>
</tr>
<tr>
<td>4275</td>
<td>23140</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>mixer_applet2</td>
</tr>
<tr>
<td>4191</td>
<td>0</td>
<td>14405</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>scim-launcher</td>
</tr>
<tr>
<td>22941</td>
<td>7908</td>
<td>1</td>
<td>0</td>
<td>62</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>gnome-terminal</td>
</tr>
<tr>
<td>4261</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>7622</td>
<td>0</td>
<td>0</td>
<td>escd</td>
</tr>
<tr>
<td>3695</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7622</td>
<td>0</td>
<td>0</td>
<td>gdm-binary</td>
</tr>
<tr>
<td>3483</td>
<td>0</td>
<td>7206</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>dhcdbus</td>
</tr>
<tr>
<td>4189</td>
<td>6916</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>scim-panel-gtk</td>
</tr>
<tr>
<td>1863</td>
<td>5767</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>iscsid</td>
</tr>
<tr>
<td>2562</td>
<td>0</td>
<td>2881</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1438</td>
<td>0</td>
<td>pcscd</td>
</tr>
<tr>
<td>4257</td>
<td>4255</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>gnome-power-man</td>
</tr>
<tr>
<td>4278</td>
<td>3876</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>multiload-apple</td>
</tr>
<tr>
<td>4083</td>
<td>0</td>
<td>1331</td>
<td>0</td>
<td>1728</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Xorg</td>
</tr>
<tr>
<td>3921</td>
<td>1603</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>gam_server</td>
</tr>
<tr>
<td>4248</td>
<td>1591</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>nm-applet</td>
</tr>
<tr>
<td>3165</td>
<td>0</td>
<td>1441</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>xterm</td>
</tr>
<tr>
<td>29548</td>
<td>0</td>
<td>1440</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>httpd</td>
</tr>
<tr>
<td>1862</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1438</td>
<td>0</td>
<td>0</td>
<td>iscsid</td>
</tr>
</tbody>
</table>
You can increase the sample time by editing the second probe (\texttt{timer.s(1)}). The output of \texttt{timeout.stp} contains the name and UID of the top 20 polling applications, along with how many times each application performed each polling system call (over time). \textit{Example 5.16, "timeout.stp Sample Output"} contains an excerpt of the script. In this particular example firefox is doing an excessive amount of polling due to a plugin module.

### 5.3.5. Tracking Most Frequently Used System Calls

\texttt{timeout.stp} from \textit{Section 5.3.4, "Monitoring Polling Applications"} helps you identify which applications are polling by examining a small subset of system calls (\texttt{poll}, \texttt{select}, \texttt{epoll}, \texttt{itime}, \texttt{futex}, \texttt{nanosleep}, and \texttt{signal}). However, in some systems, an excessive number of system calls outside that small subset might be responsible for time spent in the kernel. If you suspect that an application is using system calls excessively, you need to identify the most frequently used system calls on the system. To do this, use \texttt{topsys.stp}.

\texttt{topsys.stp}

```bash
#!/usr/bin/env stap
#
# This script continuously lists the top 20 system calls in the interval
# 5 seconds
#
global syscalls_count

probe syscall_any {
    syscalls_count[name] <<< 1
}

function print_systop () {
    printf("%25s %10s\n", "SYSCALL", "COUNT")
    foreach (syscall in syscalls_count- limit 20) {
        printf("%25s %10d\n", syscall, @count(syscalls_count[syscall]))
    }
    delete syscalls_count
}

probe timer.s(5) {
    print_systop()
    printf("---------------------------------------------------------------\n")
}

global prom_arr

probe prometheus {
    foreach (syscall in syscalls_count- limit 20)
        prom_arr[syscall] = @count(syscalls_count[syscall])

    @prometheus_dump_array1(prom_arr, "top_syscall_count", "name")
    delete prom_arr
}
```

\texttt{topsys.stp} lists the top 20 system calls used by the system per 5-second interval. It also lists how many times each system call was used during that period. Refer to \textit{Example 5.17, "topsys.stp Sample Output"} for a sample output.
Example 5.17. `topsys.stp` Sample Output

```
--------------------------------------------------------------
SYSCALL      COUNT
gettimeofday       1857
read       1821
ioctl       1568
poll       1033
close       638
open        503
select        455
write        391
writev        335
futex        303
recvmsg       251
socket        137
clock_gettime     124
rt_sigprocmask    121
sendto         120
setitimer       106
stat          90
time          81
sigreturn       72
fstat          66
--------------------------------------------------------------
```

5.3.6. Tracking System Call Volume Per Process

This section illustrates how to determine which processes are performing the highest volume of system calls. In previous sections, we've described how to monitor the top system calls used by the system over time (Section 5.3.5, "Tracking Most Frequently Used System Calls"). We've also described how to identify which applications use a specific set of "polling suspect" system calls the most (Section 5.3.4, "Monitoring Polling Applications"). Monitoring the volume of system calls made by each process provides more data in investigating your system for polling processes and other resource hogs.

```
#syscalls_by_proc.stp

#!/usr/bin/env stap

# Copyright (C) 2006 IBM Corp.
# Copyright (C) 2018-2019 Red Hat, Inc.

# This file is part of systemtap, and is free software. You can
# redistribute it and/or modify it under the terms of the GNU General
# Public License (GPL); either version 2, or (at your option) any
# later version.

# Print the system call count by process name in descending order.

#global syscalls
```
Tracking System Call Volume Per Process

```c
probe begin {
    print ("Collecting data... Type Ctrl-C to exit and display results\n")
}

probe syscall_any {
    syscalls[execname()] <<< 1
}

probe end {
    printf("%-10s %-s\n", "#SysCalls", "Process Name")
    foreach (proc in syscalls-)
        printf("%-10d %-s\n", @sum(syscalls[proc]), proc)
}
```

`syscalls_by_proc.stp` lists the top 20 processes performing the highest number of system calls. It also lists how many system calls each process performed during the time period. Refer to Example 5.18, “topsys.stp Sample Output” for a sample output.

**Example 5.18. topsys.stp Sample Output**

```
Collecting data... Type Ctrl-C to exit and display results
#SysCalls Process Name
1577    multiload-apple
692     synergyc
408     pcscd
376     mixer_applet2
299     gnome-terminal
293     Xorg
206     scim-panel-gtk
95      gnome-power-man
90      artsd
85      dhcdbus
84      scim-bridge
78      gnome-screensav
66      scim-launcher
[...]
```

To display the process IDs instead of the process names, use the following script instead.

**syscalls_by_pid.stp**

```c
#!/usr/bin/env stap

# Copyright (C) 2006 IBM Corp.
#
# This file is part of systemtap, and is free software. You can
# redistribute it and/or modify it under the terms of the GNU General
# Public License (GPL); either version 2, or (at your option) any
# later version.
#
# Print the system call count by process ID in descending order.
#
global syscalls
probe begin {
```
5.4. Identifying Contended User-Space Locks

This section describes how to identify contended user-space locks throughout the system within a specific time period. The ability to identify contended user-space locks can help you investigate poor program performance that you suspect may be caused by futex contentions.

Simply put, futex contention occurs when multiple processes are trying to access the same lock variable at the same time. This can result in a poor performance because the lock serializes execution; one process obtains the lock while the other processes must wait for the lock variable to become available again.

The futexes.stp script probes the futex system call to show lock contention.

futexes.stp

```bash
#!/usr/bin/env stap

# This script tries to identify contended user-space locks by hooking
# into the futex system call.

global FUTEX_WAIT = 0 */, FUTEX_WAKE = 1 */
global FUTEX_PRIVATE_FLAG = 128 /* linux 2.6.22+ */
global FUTEX_CLOCK_REALTIME = 256 /* linux 2.6.29+ */

global lock_waits # long-lived stats on (tid,lock) blockage elapsed time
global process_names # long-lived pid-to-execname mapping
global entry_times%, uaddrs%
```
Identifying Contended User-Space Locks

```c
probe syscall.futex {
    if ((op & ~(FUTEX_PRIVATE_FLAG|FUTEX_CLOCK_REALTIME)) != FUTEX_WAIT) next
    entry_times[tid()] = gettimeofday_us()
    uaddrs[tid()] = futex_uaddr
}

probe syscall.futex.return {
    if (!(entry_times[tid()])) next
    elapsed = gettimeofday_us() - entry_times[tid()]
    lock_waits[pid()], uaddrs[tid()] <<< elapsed
    delete entry_times[tid()]
    delete uaddrs[tid()]
    if (!(pid() in process_names))
        process_names[pid()] = execname()
}

probe end {
    foreach ([pid+, lock] in lock_waits)
        printf("%s[%d] lock %p contended %d times, %d avg us\n", process_names[pid], pid, lock, @count(lock_waits[pid,lock]), @avg(lock_waits[pid,lock]))
}
```

`futexes.stp` needs to be manually stopped; upon exit, it prints the following information:

- Name and ID of the process responsible for a contention
- The location of the contested lock variable
- How many times the lock variable was contended
- Average time of contention throughout the probe

*Example 5.19, “futexes.stp Sample Output”* contains an excerpt from the output of `futexes.stp` upon exiting the script (after approximately 20 seconds).

*Example 5.19. futexes.stp Sample Output*

```
[...]
automount[2825] lock 0x00bc7784 contended 18 times, 999931 avg us
synergyc[3686] lock 0x0861e96c contended 192 times, 101991 avg us
synergyc[3758] lock 0x08d98744 contended 192 times, 101990 avg us
synergyc[3938] lock 0x0982a8b4 contended 192 times, 101997 avg us
[...]
```
Understanding SystemTap Errors

This chapter explains the most common errors you may encounter while using SystemTap.

6.1. Parse and Semantic Errors

Parse and semantic errors occur while SystemTap attempts to parse and translate the script into C, before converting it into a kernel module. For example, type errors result from operations that assign invalid values to variables or arrays.

**parse error: expected foo, saw bar**

The script contains a grammatical or typographical error. SystemTap detected the type of the construct that is incorrect, given the context of the probe.

For example, the following invalid SystemTap script is missing its probe handlers:

```plaintext
probe vfs.read
probe vfs.write
```

An attempt to run this SystemTap script fails with the following error message showing that the parser expects something other than the `probe` keyword in column 1 of line 2:

```
parse error: expected one of '.', ',', '{', '?', '!', '=', '+', '-'  
saw: keyword at perror.stp:2:1
1 parse error(s).
```

**parse error: embedded code in unprivileged script**

The script contains unsafe embedded C code, that is, blocks of code surrounded by ` %{` and `%}`. SystemTap allows you to embed C code in a script, which is useful if there are no tapsets to suit your purposes. However, embedded C constructs are not safe and SystemTap reports this error to warn you if such constructs appear in the script.

If you are sure that any similar constructs in the script are safe and you are a member of the `stapdev` group (or have root privileges), run the script in "guru" mode by using the `-g` option:

```
stap -g script
```

**semantic error: type mismatch for identifier 'foo' ... string vs. long**

The function `foo` in the script used the wrong type (such as `%s` or `%d`). In the following example, the format specifier should be `%s` and not `%d`, because the `execname()` function returns a string:

```plaintext
probe syscall.open
```
Chapter 6. Understanding SystemTap Errors

{  
  printf ("%d(%d) open\n", execname(), pid())
}

**semantic error: unresolved type for identifier 'foo'**

The identifier (variable) was used, but no type (integer or string) could be determined. This occurs, for instance, if you use a variable in a `printf` statement while the script never assigns a value to the variable.

**semantic error: Expecting symbol or array index expression**

SystemTap could not assign a value to a variable or to a location in an array. The destination for the assignment is not a valid destination. The following example code would generate this error:

```
probe begin { printf("x") = 1 }
```

**while searching for arity N function, semantic error: unresolved function call**

A function call or array index expression in the script used an invalid number of arguments or parameters. In SystemTap, *arity* can either refer to the number of indices for an array, or the number of parameters to a function.

**semantic error: array locals not supported, missing global declaration?**

The script used an array operation without declaring the array as a global variable (global variables can be declared after their use in SystemTap scripts). Similar messages appear if an array is used, but with inconsistent arities.

**semantic error: variable 'foo' modified during 'foreach' iteration**

The array *foo* is being modified (being assigned to or deleted from) within an active `foreach` loop. This error also displays if an operation within the script performs a function call within the `foreach` loop.

**semantic error: probe point mismatch at position N, while resolving probe point foo**

SystemTap did not understand what the event or SystemTap function *foo* refers to. This usually means that SystemTap could not find a match for *foo* in the tapset library. *N* refers to the line and column of the error.
semantic error: no match for probe point, while resolving probe point foo

SystemTap could not resolve the events or handler function foo for a variety of reasons. This error occurs when the script contains the event \texttt{kernel.function("something")}, and \texttt{something} does not exist. In some cases, the error could also mean the script contains an invalid kernel file name or source line number.

semantic error: unresolved target-symbol expression

A handler in the script references a target variable, but the value of the variable could not be resolved. This error could also mean that a handler is referencing a target variable that is not valid in the context when it was referenced. This may be a result of compiler optimization of the generated code.

semantic error: libdwfl failure

There was a problem processing the debugging information. In most cases, this error results from the installation of a \texttt{kernel-debuginfo} package whose version does not match the probed kernel exactly. The installed \texttt{kernel-debuginfo} package itself may have some consistency or correctness problems.

semantic error: cannot find foo debuginfo

SystemTap could not find a suitable \texttt{kernel-debuginfo} package.

6.2. Runtime Errors and Warnings

Runtime errors and warnings occur when the SystemTap instrumentation has been installed and is collecting data on the system.

\textbf{WARNING: Number of errors: }$N$, \textbf{skipped probes: }$M$

Errors and/or skipped probes occurred during this run. Both $N$ and $M$ are the counts of the number of probes that were not executed due to conditions such as too much time required to execute event handlers over an interval of time.

division by 0

The script code performed an invalid division.

aggregate element not found

A statistics extractor function other than \texttt{@count} was invoked on an aggregate that has not had any values accumulated yet. This is similar to a division by zero.
aggregation overflow

An array containing aggregate values contains too many distinct key pairs at this time.

MAXNESTING exceeded

Too many levels of function call nesting were attempted. The default nesting of function calls allowed is 10.

MAXACTION exceeded

The probe handler attempted to execute too many statements in the probe handler. The default number of actions allowed in a probe handler is 1000.

kernel/user string copy fault at ADDR

The probe handler attempted to copy a string from kernel or user space at an invalid address (ADDR).

pointer dereference fault

There was a fault encountered during a pointer dereference operation such as a target variable evaluation.
Chapter 7.

References

This chapter enumerates other references for more information about SystemTap. Refer to these sources in the course of writing advanced probes and tapsets.

SystemTap Wiki

The SystemTap Wiki is a collection of links and articles related to the deployment, usage, and development of SystemTap. You can find it at http://sourceware.org/systemtap/wiki/HomePage.

SystemTap Tutorial

Much of the content in this book comes from the SystemTap Tutorial. The SystemTap Tutorial is a more appropriate reference for users with intermediate to advanced knowledge of C++ and kernel development, and you can find it at http://sourceware.org/systemtap/tutorial/.

man stapprobes

The stapprobes(3stap) man page enumerates a variety of probe points supported by SystemTap, along with additional aliases defined by the SystemTap tapset library. The bottom part of the man page includes a list of other man pages enumerating similar probe points for specific system components, such as tapset::scsi, tapset::kprocess, tapset::signal, and so on.

man stapfuncs

The stapfuncs(3stap) man page enumerates numerous functions supported by the SystemTap tapset library, along with the prescribed syntax for each of them. Note, however, that it does not provide a complete list of all supported functions; there are more undocumented functions available.

SystemTap Tapset Reference Manual

The SystemTap Tapset Reference Manual describes the individual predefined functions and probe points of the tapsets in greater detail. You can find it at http://sourceware.org/systemtap/tapsets/.

SystemTap Language Reference

The SystemTap Language Reference is a comprehensive reference of SystemTap's language constructs and syntax. It is recommended for users with a rudimentary to intermediate knowledge of C++ and other similar programming languages, and is available to all users at http://sourceware.org/systemtap/langref/.

Tapset Developers Guide

Once you have sufficient proficiency in writing SystemTap scripts, you can try to write your own tapsets. The Tapset Developers Guide describes how to add functions to your tapset library.

Test Suite

The systemtap-testsuite package allows you to test the entire SystemTap toolchain without having to build it from source code. In addition, it also contains numerous examples of SystemTap scripts to study and test; some of these scripts are also documented in Chapter 5, Useful SystemTap Scripts.

By default, the example scripts included in systemtap-testsuite are located in the /usr/share/systemtap/testsuite/systemtap.examples/ directory.
Appendix A. Revision History

Revision 2.0-1  Mon Jul 20 2009  Don Domingo ddomingo@redhat.com
includes 5.4 minor updates and additional script "dropwatch.stp"

Revision 1.0-1  Wed Jun 17 2009  Don Domingo ddomingo@redhat.com
Building+pushing to RHEL
Index

Symbols
$\text{count}$
  sample usage
  local variables, 50
$\text{return}$
  sample usage
  local variables, 47, 51
$\text{avg}$ (integer extractor)
  computing for statistical aggregates
  array operations, 34
$\text{count}$ (integer extractor)
  computing for statistical aggregates
  array operations, 33
$\text{max}$ (integer extractor)
  computing for statistical aggregates
  array operations, 33
$\text{min}$ (integer extractor)
  computing for statistical aggregates
  array operations, 33
$\text{sum}$ (integer extractor)
  computing for statistical aggregates
  array operations, 33

A
adding values to statistical aggregates
  computing for statistical aggregates
  array operations, 33
advantages of cross-instrumentation, 6
aggregate element not found
  runtime errors/warnings
    understanding SystemTap errors, 71
aggregates (statistical)
  array operations, 32
aggregation overflow
  runtime errors/warnings
    understanding SystemTap errors, 72
algebraic formulas using arrays
  reading values from arrays
    array operations, 27
architecture notation, determining, 7
architecture of SystemTap, 11
array locals not supported
  parse/semantics error
    understanding SystemTap errors, 70
array operations
  assigning associated values, 27
  associating timestamps to process names, 27
associative arrays, 26
  clearing arrays/array elements, 29
  delete operator, 30
  multiple array operations within the same
    probe, 31
  virtual file system reads (non-cumulative),
    tallying, 30
  computing for statistical aggregates, 32
    $\text{avg}$ (integer extractor), 34
    $\text{count}$ (integer extractor), 33
    $\text{max}$ (integer extractor), 33
    $\text{min}$ (integer extractor), 33
    $\text{sum}$ (integer extractor), 33
    adding values to statistical aggregates, 33
    count (operator), 33
    extracting data collected by statistical
      aggregates, 33
  conditional statements, using arrays in, 31
    testing for array membership, 32
  deleting arrays and array elements, 29
  incrementing associated values, 28
    tallying virtual file system reads (VFS reads), 28
  multiple elements in an array, 28
  processing multiple elements in an array, 28
    cumulative virtual file system reads, tallying,
      29
    foreach, 28
    iterations, processing elements in an array as,
      28
    limiting the output of foreach, 29
    ordering the output of foreach, 29
  reading values from arrays, 27
    computing for timestamp deltas, 27
    empty unique keys, 28
    using arrays in simple computations, 27
arrays, 26
  (see also associative arrays)
assigning associated values
  array operations, 27
    associating timestamps to process names, 27
    associating timestamps to process names
      array operations, 27
associated values
  introduction
    arrays, 26
assigning timestamps to process names
  assigning associated values
    array operations, 27
    associative arrays
introduction, 26
associated values, 26
example, 26
index expression, 26
key pairs, 26
syntax, 26
unique keys, 26
asynchronous events
Events, 14

B
begin
Events, 14
building instrumentation modules from SystemTap
scripts, 5
building kernel modules from SystemTap scripts, 5

C
call graph tracing
examples of SystemTap scripts, 57
capabilities of SystemTap
Introduction, 1
changes to file attributes, monitoring
elements of SystemTap scripts, 54
clearing arrays/array elements
array operations, 29
delete operator, 30
multiple array operations within the same
probe, 31
virtual file system reads (non-cumulative),
tallying, 30
command-line arguments
SystemTap handler constructs
handlers, 25
compiling instrumentation/kernel modules from
SystemTap scripts, 5
components
SystemTap scripts
introduction, 12
computing for statistical aggregates
array operations, 32
@avg (integer extractor), 34
@count (integer extractor), 33
@max (integer extractor), 33
@min (integer extractor), 33
@sum (integer extractor), 33
adding values to statistical aggregates, 33
count (operator), 33
extracting data collected by statistical
aggregates, 33
computing for timestamp deltas
reading values from arrays
array operations, 27
conditional operators
conditional statements
handlers, 25
conditional statements, using arrays in
array operations, 31
testing for array membership, 32
CONFIG_HZ, computing for, 20
contended user-space locks (futex contentions),
identifying
examples of SystemTap scripts, 66
copy fault
runtime errors/warnings
understanding SystemTap errors, 72
count operator
computing for statistical aggregates
array (operator), 33
counting function calls
examples of SystemTap scripts, 56
CPU ticks
examples of SystemTap scripts, 59
cpu()
functions, 17
cross-compiling, 5
cross-instrumentation
advantages of, 6
building kernel modules from SystemTap scripts,
5
configuration
host system and target system, 6
generating instrumentation from SystemTap
scripts, 5
host system, 6
instrumentation module, 6
target kernel, 6
target system, 6
cftime()
functions, 17
cftime(), example of usage
script examples, 47
cumulative I/O, tracking
examples of SystemTap scripts, 50
cumulative virtual file system reads, tallying
processing multiple elements in an array
array operations, 29

D
delete operator
clearing arrays/array elements
  array operations, 30
determining architecture notation, 7
determining the kernel version, 4
determining time spent in kernel and user space
  examples of SystemTap scripts, 58
device I/O, monitoring
  examples of SystemTap scripts, 52
device number of a file (integer format)
  examples of SystemTap scripts, 53
disk I/O traffic, summarizing
  script examples, 46
division by 0
  runtime errors/warnings
    understanding SystemTap errors, 71
documentation goals
  Introduction, 1

E
embedded code in unprivileged script
  parse/semantics error
    understanding SystemTap errors, 69
empty unique keys
  reading values from arrays
    array operations, 28
end
  Events, 15
division by 0
  runtime errors/warnings
    understanding SystemTap errors, 71

E
embedded code in unprivileged script
  parse/semantics error
    understanding SystemTap errors, 69
number of errors: N, skipped probes: M, 71
pointer dereference fault, 72
event types
  Understanding How SystemTap Works, 11
Events
asynchronous events, 14
begin, 14
end, 15
events
  examples of synchronous and asynchronous events, 13
  introduction, 13
  kernel.function("function"), 13
  kernel.trace("tracepoint"), 14
  module("module"), 14
  synchronous events, 13
  syscall.system_call, 13
timer events, 15
user-space, 35
  vfs.file_operation, 13
  wildcards, 14
events and handlers, 11
events wildcards, 14
example
  introduction
    arrays, 26
example of multiple command-line arguments
  examples of SystemTap scripts, 58
examples of synchronous and asynchronous events
  Events, 13
timeout, 15
examples of SystemTap scripts, 39
call graph tracing, 57
CPU ticks, 59
time(), example of usage, 47
determining time spent in kernel and user space, 58
file device number (integer format), 53
futex (lock) contentions, 66
futex system call, 66
identifying contended user-space locks (futex contentions), 66
if/else conditionals, alternative syntax, 40
inode number, 53
monitoring changes to file attributes, 54
monitoring device I/O, 52
monitoring I/O block time, 55
monitoring I/O time, 48
monitoring incoming TCP connections, 42
monitoring polling applications, 60
monitoring reads and writes to a file, 53
monitoring system calls, 63
monitoring system calls (volume per process), 64
monitoring TCP packets, 42
multiple command-line arguments, example of, 58
net/socket.c, tracing functions from, 41
network profiling, 39, 44
stat -c, determining file device number (integer format), 53
stat -c, determining whole device number, 52
summarizing disk I/O traffic, 46
tallying function calls, 56
thread_indent(), sample usage, 58
timer.ms(), sample usage, 56
timer.s(), sample usage, 62, 64
tracing functions called in network socket code, 41
tracking cumulative I/O, 50
trigger function, 58
usrdev2kerndev(), 52
whole device number (usage as a command-line argument), 52
exceeded MAXACTION
runtime errors/warnings
understanding SystemTap errors, 72
exceeded MAXNESTING
runtime errors/warnings
understanding SystemTap errors, 72
exit()
functions, 16
expected symbol/array index expression
parse/semantics error
understanding SystemTap errors, 70
extracting data collected by statistical aggregates
computing for statistical aggregates
array operations, 33

F
feedback
contact information for this manual, vii
file attributes, monitoring changes to
examples of SystemTap scripts, 54
file device number (integer format)
examples of SystemTap scripts, 53
file reads/writes, monitoring
examples of SystemTap scripts, 53
flight recorder mode, 9
file mode, 10
in-memory mode, 9
gettimeofday_s(), functions, 17

global
  SystemTap handler constructs
    handlers, 20
goals, documentation
  Introduction, 1
grammatical/typographical script error
  parse/semantics error
    understanding SystemTap errors, 69
guru mode
  parse/semantics error
    understanding SystemTap errors, 69

handler functions, 17

handlers
  conditional statements, 24
    conditional operators, 25
    for loops, 25
    if/else, 24
    while loops, 24
  introduction, 15
  SystemTap handler constructs, 19
    command-line arguments, 25
    global, 20
    syntax and format, 19
    variable notations, 26
    variables, 19
  target variables, 20

handlers and events, 11

SystemTap scripts
  introduction, 12
heaviest disk reads/writes, identifying
  script examples, 46
host system
  cross-instrumentation, 6
host system and target system
cross-instrumentation
  configuration, 6

I/O block time, monitoring
  examples of SystemTap scripts, 55
I/O monitoring (by device)
  examples of SystemTap scripts, 52
I/O time, monitoring
  examples of SystemTap scripts, 48
I/O traffic, summarizing
  script examples, 46
identifier type mismatch
  parse/semantics error
    understanding SystemTap errors, 69
identifying contended user-space locks (futex contentions)
  examples of SystemTap scripts, 66
identifying heaviest disk reads/writes
  script examples, 46
if/else
  conditional statements
    handlers, 24
  if/else conditionals, alternative syntax
    examples of SystemTap scripts, 40
if/else statements, using arrays in
  array operations, 31
incoming TCP connections, monitoring
  examples of SystemTap scripts, 42
incoming/outgoing function calls, tracing
  examples of SystemTap scripts, 57
incrementing associated values
  array operations, 28
    tallying virtual file system reads (VFS reads), 28
index expression
  introduction
    arrays, 26
initial testing, 5
inode number
  examples of SystemTap scripts, 53
Installation
  initial testing, 5
  kernel information packages, 3
  kernel version, determining the, 4
  required packages, 3
  Setup and Installation, 3
  systemtap package, 3
  systemtap-runtime package, 3
instrumentation module
  cross-instrumentation, 6
instrumentation modules from SystemTap scripts,
  building, 5
integer extractors
  computing for statistical aggregates
    array operations, 33
Introduction
  capabilities of SystemTap, 1
  documentation goals, 1
  goals, documentation, 1
limitations of SystemTap, 1
performance monitoring, 1
invalid division
runtime errors/warnings
understanding SystemTap errors, 71
invalid values to variables/arrays
parse/semantics error
understanding SystemTap errors, 69
iterations, processing elements in an array as
processing multiple elements in an array
array operations, 28
K
kernel and user space, determining time spent in
examples of SystemTap scripts, 58
kernel information packages, 3
kernel modules from SystemTap scripts, building, 5
kernel version, determining the, 4
kernel.function("function")
Events, 13
kernel.trace("tracepoint")
Events, 14
key pairs
introduction
arrays, 26
L
libdwfl failure
parse/semantics error
understanding SystemTap errors, 71
limitations of SystemTap
Introduction, 1
limiting the output of foreach
processing multiple elements in an array
array operations, 29
local variables
name, 19
sample usage
$count, 50
$return, 47, 51
M
MAXACTION exceeded
runtime errors/warnings
understanding SystemTap errors, 72
MAXNESTING exceeded
runtime errors/warnings
understanding SystemTap errors, 72
membership (in array), testing for
conditional statements, using arrays in
array operations, 32
module("module")
Events, 14
monitoring changes to file attributes
examples of SystemTap scripts, 54
monitoring cumulative I/O
examples of SystemTap scripts, 50
monitoring device I/O
examples of SystemTap scripts, 52
monitoring I/O block time
examples of SystemTap scripts, 55
monitoring I/O time
examples of SystemTap scripts, 48
monitoring incoming TCP connections
examples of SystemTap scripts, 42
monitoring polling applications
examples of SystemTap scripts, 60
monitoring reads and writes to a file
examples of SystemTap scripts, 53
monitoring system calls
examples of SystemTap scripts, 63
monitoring system calls (volume per process)
examples of SystemTap scripts, 64
monitoring TCP packets
examples of SystemTap scripts, 42
multiple array operations within the same probe
clearing arrays/array elements
array operations, 31
multiple command-line arguments, example of
examples of SystemTap scripts, 58
multiple elements in an array
array operations, 28
N
name
local variables, 19
net/socket.c, tracing functions from
examples of SystemTap scripts, 41
network profiling
examples of SystemTap scripts, 39, 44
network socket code, tracing functions called in
examples of SystemTap scripts, 41
network traffic, monitoring
examples of SystemTap scripts, 39, 44
no match for probe point
parse/semantics error
understanding SystemTap errors, 71
non-global arrays
parse/semantics error
understanding SystemTap errors, 70
number of errors: N, skipped probes: M
runtime errors/warnings
understanding SystemTap errors, 71

O
operations
assigning associated values
assigning timestamps to process names, 27
associative arrays, 26
clearing arrays/array elements, 29
delete operator, 30
multiple array operations within the same
probe, 31
virtual file system reads (non-cumulative),
tallying, 30
computing for statistical aggregates, 32
@avg (integer extractor), 34
@count (integer extractor), 33
@max (integer extractor), 33
@min (integer extractor), 33
@sum (integer extractor), 33
adding values to statistical aggregates, 33
count (operator), 33
extracting data collected by statistical
aggregates, 33
conditional statements, using arrays in, 31
testing for array membership, 32
deleting arrays and array elements, 29
incrementing associated values, 28
tallying virtual file system reads (VFS reads),
28
multiple elements in an array, 28
processing multiple elements in an array, 28
cumulative virtual file system reads, tallying,
29
foreach, 28
iterations, processing elements in an array as,
28
limiting the output of foreach, 29
ordering the output of foreach, 29
reading values from arrays, 27
computing for timestamp deltas, 27
empty unique keys, 28
using arrays in simple computations, 27
options, stap
Usage, 8
ordering the output of foreach
processing multiple elements in an array
array operations, 29
overflow of aggregation

P
packages required to run SystemTap, 3
parse/semantics error
understanding SystemTap errors, 69
embedded code in unprivileged script, 69
expected symbol/array index expression, 70
grammatical/typographical script error, 69
guru mode, 69
invalid values to variables/arrays, 69
libdwfl failure, 71
no match for probe point, 71
non-global arrays, 70
probe mismatch, 70
type mismatch for identifier, 69
unresolved function call, 70
unresolved target-symbol expression, 71
unresolved type for identifier, 70
variable modified during ‘foreach’, 70
performance monitoring
Introduction, 1
pointer dereference fault
runtime errors/warnings
understanding SystemTap errors, 72
polling applications, monitoring
examples of SystemTap scripts, 60
pp()
functions, 17
printf()
format specifiers, 16
format strings, 16, 16
syntax and format, 16
printing I/O activity (cumulative)
examples of SystemTap scripts, 50
printing I/O block time (periodically)
examples of SystemTap scripts, 55
probe mismatch
parse/semantics error
understanding SystemTap errors, 70
probe point (no match for)
parsing/semantics error
understanding SystemTap errors, 71
probes
SystemTap scripts
introduction, 12
processing multiple elements in an array
array operations, 28
cumulative virtual file system reads, tallying
Index

array operations, 29
foreach
array operations, 28
limiting the output of foreach
array operations, 29
ordering the output of foreach
array operations, 29
profiling the network
examples of SystemTap scripts, 39, 44

R
reading values from arrays
array operations, 27
empty unique keys, 28
using arrays in simple computations, 27
computing for timestamp deltas
array operations, 27
reads/writes to a file, monitoring
examples of SystemTap scripts, 53
required packages, 3
RPMs required to run SystemTap, 3
running scripts from standard input, 8
running SystemTap scripts
Usage, 7
runtime errors/warnings
understanding SystemTap errors, 71
aggregate element not found, 71
aggregation overflow, 72
copy fault, 72
division by 0, 71
MAXACTION exceeded, 72
MAXNESTING exceeded, 72
number of errors: N, skipped probes: M, 71
pointer dereference fault, 72

S
script examples
call graph tracing, 57
CPU ticks, 59
time(), example of usage, 47
determining time spent in kernel and user space, 58
file device number (integer format), 53
futex (lock) contentions, 66
futex system call, 66
identifying contended user-space locks (futex contentions), 66
if/else conditionals, alternative syntax, 40
inode number, 53
monitoring changes to file attributes, 54
monitoring device I/O, 52
monitoring I/O block time, 55
monitoring I/O time, 48
monitoring incoming TCP connections, 42
monitoring polling applications, 60
monitoring reads and writes to a file, 53
monitoring system calls, 63
monitoring system calls (volume per process), 64
monitoring TCP packets, 42
multiple command-line arguments, example of, 58
net/socket.c, tracing functions from, 41
network profiling, 39, 44
stat -c, determining file device number (integer format), 53
stat -c, determining whole device number, 52
summarizing disk I/O traffic, 46
tallying function calls, 56
timer.ms(), sample usage, 56
timer.s(), sample usage, 62, 64
tracing functions called in network socket code, 41
tracking cumulative I/O, 50
trigger function, 58
usrdev2kerndev(), 52
whole device number (usage as a command-line argument), 52
scripts
introduction, 11
components, 12
events and handlers, 12
format and syntax, 12
functions, 12
probes, 12
statement blocks, 12
sessions, SystemTap, 11
Setup and Installation, 3
Stack backtrace
user-space, 37
standard input, running scripts from
Usage, 8
stap
Usage, 7
stap options, 8
stapdev
Usage, 7
staprun
Usage, 7
stapusr
Usage, 8
stat -c, determining file device number (integer format)
examples of SystemTap scripts, 53
stat -c, determining whole device number
total examples of SystemTap scripts, 52
statement blocks
SystemTap scripts
introduction, 12
statistical aggregates
array operations, 32
summarizing disk I/O traffic
script examples, 46, 46
synchronous events
Events, 13
syntax
introduction
arrays, 26
syntax and format
printf(), 16
SystemTap handler constructs
handlers, 19
SystemTap scripts
introduction, 12
syscall.system_call
Events, 13
system calls volume (per process), monitoring
examples of SystemTap scripts, 64
system calls, monitoring
total examples of SystemTap scripts, 63
SystemTap architecture, 11
SystemTap handlers
SystemTap handler constructs, 19
syntax and format, 19
systemtap package, 3
SystemTap script functions, 17
SystemTap scripts
introduction, 11
components, 12
events and handlers, 12
format and syntax, 12
functions, 12
probes, 12
statement blocks, 12
useful examples, 39
SystemTap scripts, how to run, 7
SystemTap sessions, 11
SystemTap statements
conditional statements, 24
total conditional operators, 25
for loops, 25
if/else, 24
while loops, 24
SystemTap handler constructs
cross-instrumentation, 6
global, 20
variable notations, 26
variables, 19
systemtap-runtime package, 3
systemtap-testsuite package
sample scripts, 39
T
tallying function calls
total examples of SystemTap scripts, 56
tallying virtual file system reads (VFS reads)
incrementing associated values
array operations, 28
Tapsets
definition of, 34
target kernel
cross-instrumentation, 6
target system
cross-instrumentation, 6
target system and host system
cross-instrumentation, 6
target variables, 20
pretty printing, 22
typecasting, 23
user-space, 36
variable availability, 23
target()
funsctons, 19
target-symbol expression, unresolved
parse/semantics error
understanding SystemTap errors, 71
TCP connections (incoming), monitoring
total examples of SystemTap scripts, 42
TCP packets, monitoring
total examples of SystemTap scripts, 42, 42
testing for array membership
conditional statements, using arrays in
array operations, 32
testing, initial, 5
thread_indent()
funsctons, 17
thread_indent(), sample usage
total examples of SystemTap scripts, 58
tid()
Index

functions, 17

time of I/O
  examples of SystemTap scripts, 48
time spent in kernel/user space, determining
  examples of SystemTap scripts, 58
timer events
  Events, 15
timer.ms(), sample usage
  examples of SystemTap scripts, 56
timer.s(), sample usage
  examples of SystemTap scripts, 62, 64
timestamp deltas, computing for
  reading values from arrays
    array operations, 27
timestamps, association thereof to process names
  assigning associated values
    array operations, 27
tracepoint, 14, 44
tracing call graph
  examples of SystemTap scripts, 57
tracing functions called in network socket code
  examples of SystemTap scripts, 41
tracing incoming/outgoing function calls
  examples of SystemTap scripts, 57
tracking cumulative I/O
  examples of SystemTap scripts, 50
trigger function
  examples of SystemTap scripts, 58
type mismatch for identifier
  parse/semantics error
  understanding SystemTap errors, 69
typographical script error
  parse/semantics error
  understanding SystemTap errors, 69

U

uid()
  functions, 17
uname -m, 7
uname -r, 4
understanding SystemTap errors
runtime errors/warnings, 71
  aggregate element not found, 71
  aggregation overflow, 72
  copy fault, 72
division by 0, 71
  MAXACTION exceeded, 72
  MAXNESTING exceeded, 72
  number of errors: N, skipped probes: M, 71
  pointer dereference fault, 72
Understanding How SystemTap Works, 11
  architecture, 11
  event types, 11
  events and handlers, 11
SystemTap sessions, 11
understanding SystemTap errors
  parse/semantics error, 69
  expected symbol/array index expression, 70
  grammatical/typographical script error, 69
  guru mode, 69
  invalid values to variables/arrays, 69
  libdwfl failure, 71
  no match for probe point, 71
  non-global arrays, 70
  probe mismatch, 70
  type mismatch for identifier, 69
  unresolved function call, 70
  unresolved target-symbol expression, 71
  unresolved type for identifier, 70
  variable modified during ‘foreach’, 70
unique keys
  introduction
    arrays, 26
unprivileged script, embedded code in
  parse/semantics error
  understanding SystemTap errors, 69
unresolved function call
  parse/semantics error
  understanding SystemTap errors, 70
unresolved target-symbol expression
  parse/semantics error
  understanding SystemTap errors, 71
unresolved type for identifier
  parse/semantics error
  understanding SystemTap errors, 70
unsafe embedded code in unprivileged script
  parse/semantics error
  understanding SystemTap errors, 69
Usage
  options, stap, 8
  running SystemTap scripts, 7
  standard input, running scripts from, 8
stap, 7
stapdev, 7
staprun, 7
stapusr, 8
useful examples of SystemTap scripts, 39
user and kernel space, determining time spent in
examples of SystemTap scripts, 58
using arrays in simple computations
   reading values from arrays
      array operations, 27
Using SystemTap, 3
usrdev2kerndev()
   examples of SystemTap scripts, 52

V
values, assignment of
   array operations, 27
variable modified during 'foreach'
   parse/semantics error
      understanding SystemTap errors, 70
variable notations
   SystemTap handler constructs
      handlers, 26
variables
   SystemTap handler constructs
      handlers, 19
variables (local)
   name, 19
   sample usage
      $count, 50
      $return, 47, 51
VFS reads, tallying of
   incrementing associated values
      array operations, 28
vfs.file_operation
   Events, 13
virtual file system reads (cumulative), tallying
   processing multiple elements in an array
      array operations, 29
virtual file system reads (non-cumulative), tallying
   clearing arrays/array elements
      array operations, 30

W
while loops
   conditional statements
      handlers, 24
whole device number (usage as a command-line argument)
   examples of SystemTap scripts, 52
wildcards in events, 14
writes/reads to a file, monitoring
   examples of SystemTap scripts, 53