13 Web-Seiten im Proxy-Cache zwischenspeichern und filtern

Das World Wide Web wird oft lästerhaft World Wide Wait genannt, weil immer mehr Anwender immer mehr Seiten anfordern, als Netz-Anbieter Bandbreite für nicht bevorzugte Anwender schaffen.

Anwender können Web-Seiten schneller abrufen, wenn sie

- Verträge für schnellere Zugänge, Zugänge mit garantierter Bandbreite oder für Zusatzbandbreite über Satellit abschließen oder
- Seiten, die sie selbst oder andere Anwender der gleichen Gruppe wiederholt anfordern, nicht jedes mal neu laden, sondern aus einem Zwischenspeicher abrufen.

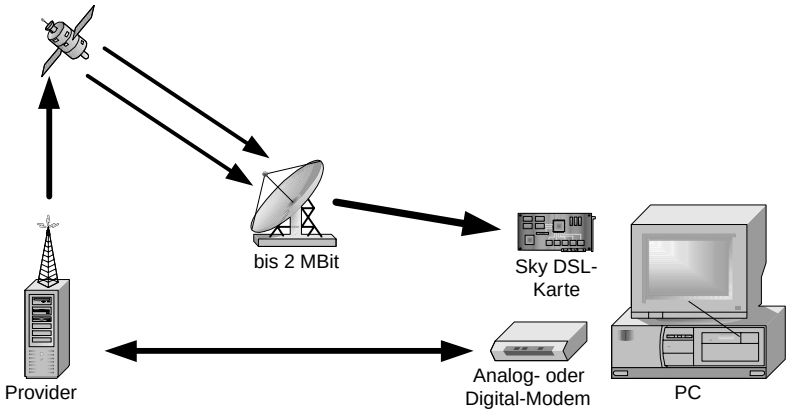

Abbildung 13.1: Mehr Bandbreite, z.B. durch Sky DSL

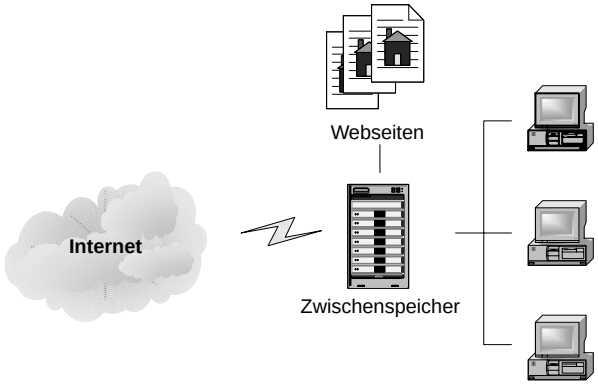

Abbildung 13.2: Web-Seiten im Proxy-Cache

Werden Internet-Seiten in geschützten Räumen wie Schulen, Betrieben oder Familien abgerufen, fordern einige System-Verantwortliche neben schnellem Seitenabruf auch Filterfunktionen und Inhalts-Kontrolle.

Web-Zugriffe lassen sich durch verschiedene Zwischenspeicher beschleunigen und filtern:

- Durch lokale Speicher und Filter beim Anwender oder
- zentrale Speicher und Filter zwischen Router und Hub.

Lokale Zwischenspeicher für Internetseiten, Cache genannt, benutzen fast alle Anwender, Anfänger sogar ohne es zu wissen, weil die Browser von Netscape und Microsoft diese Funktion schon in der Grundausstattung bieten. Um im Interesse des Jugendschutzes Seiten zu filtern, benötigt man Zusatzprogramme, die Seiten mit bestimmten Text- oder Grafikobjekten und von bestimmten Web-Sites ignorieren. Löscht man den Verlauf nicht nach einer Web-Sitzung, können Dritte ausspionieren, welche Web-Sites man besucht hat.

Diese lokalen Zwischenspeicher legen bereits einmal geladene Internetseiten im Hauptspeicher oder auf der Festplatte ab, so dass bei einem erneuten Zugriff auf die Seite kein weiteres Laden aus dem Internet erforderlich ist, es sei denn, die Seite hätte sich geändert.

Speichert man in lokalen Netzen die von allen Anwendern besuchten Seiten zentral, kommt der Geschwindigkeitsvorteil für das erneute Laden allen zugute, da die Ladezeiten im lokalen Netz vergleichsweise kurz sind.

Für diesen Zweck setzt man auf Kommunikations-Servern einen Proxy-Server, bei Linux meist *Squid*, ein.

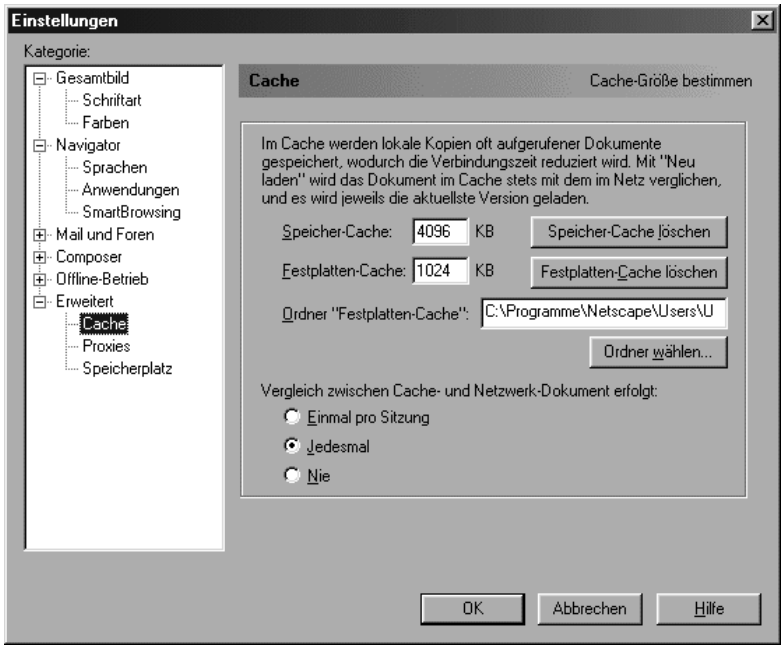

Abbildung 13.3: Cache-Einstellungen im Netscape Communicator

Abbildung 13.4: Cache-Einstellungen im Internet Explorer

Zusätzlich zu der Cache-Funktion verfügt Squid über eine Stellvertreter- (Proxy) Funktion. Bei der Einwahl ins Internet stellt der Provider nur eine einzige offizielle IP-Adresse zur Verfügung, die der Linux-Server bekommt. Die anderen Rechner im Netz verfügen nur über lokale IP-Adressen, an die Web-Server

keine Antworten schicken können. Diese lokalen Rechner fordern WWW-Seiten indirekt vom Squid an, welcher sie mit der IP-Adresse des Linux-Servers aus dem Internet abruft, sofern er sie nicht schon lokal gespeichert hat.

13.1 Wann lohnt sich ein Proxy-Cache?

Ein Proxy-Cache hat mehrere Vorteile bzw. Aufgaben:

- Er beschleunigt den Internet-Zugriff,
- hat eine Stellvertreterfunktion für die Rechner im Netz;
- er kann kontrollieren, welche Inhalte Benutzer im lokalen Netz anfordern dürfen und
- er dokumentiert, wer welche Web-Seiten tatsächlich geladen hat.

Wie sehr der Proxy-Server das Laden von Web-Seiten beschleunigt, indem er mehrfach angeforderte Seiten aus dem lokalen Netz statt aus dem Internet bereitstellt, hängt in der Praxis davon ab, wie viele Nutzer die gleichen Seiten anfordern und wie viele Nutzer sich eine vielleicht nur schmalbandige Internet-Anbindung teilen müssen.

Die Proxy- (Stellvertreter-) Funktion ist die einfachste Möglichkeit, beliebig vielen Rechnern im Intranet den Zugriff auf WWW-Seiten zu ermöglichen. Da dabei nur der Proxy Squid Anfragen ins Internet stellt, kommt man mit einer einzigen offiziellen IP-Adresse aus.

Will man den lokalen Rechnern erlauben, selbst direkt auf Web-Server zuzugreifen, muss der Server die lokalen IP-Adressen jeweils durch seine eigene ersetzen (IP-Masquerading). Um auch dann noch Sperr- und Kontrollmöglichkeiten zu garantieren, muss man jedoch einen Firewall einrichten und betreiben (s. Kapitel 14).

Proxies können gezielt einzelne Seiten oder ganze Internet-Domains sperren, damit kein darauf zugreifender Browser diese überhaupt sehen oder laden kann.

Da ein Proxy alle Zugriffe protokollieren kann, lässt sich überwachen, wer welche Seiten aufgerufen hat.

13.2 So funktioniert ein Proxy-Cache

Anfragen von Client-Browsern gehen nicht mehr direkt ins Internet, sondern zum Proxy-Server. Dieser prüft, ob er eine aktuelle Version der angeforderten Seite gespeichert hat. Wenn die Seite vorliegt und noch aktuell ist, liefert er sie direkt aus dem lokalen System heraus an den Browser.

Hat er die Seite nicht im Speicher oder ist sie nicht mehr aktuell, so lädt der Proxy sie aus dem Internet, speichert sie bei sich und stellt sie dann den Browsern der Clients zur Verfügung.

13.3 Squid installieren und konfigurieren

Da alle Linux-Distributionen Squid enthalten, lässt er sich einfach durch Auswahl des zugehörigen Pakets einrichten. Bei SuSE befindet sich die bewährte Version des Squid in der Serie `n` im Paket `squid2` bzw. in der Datei `squid2.rpm` im Verzeichnis `n1`.

Datei	*Bedeutung*
`/usr/sbin/squid`	Binärdatei, die den eigentlichen Squid-Server bildet.
`/sbin/init.d/squid`	Start/Stop-Script für Squid.
`/etc/squid.conf`	Squid-Konfigurationsdatei.

Tabelle 13.1: Die Dateien zu Squid

Nach der Installation muss man dafür sorgen, dass Squid automatisch startet. Dazu ruft man YaST auf und geht in *Administration des Systems* in das Menü *Konfigurationsdatei verändern*. Dort sucht man aus der Liste die Variable *START_SQUID* und setzt ihren Wert auf *yes*. Anschließend kann man YaST beenden.

Diese Änderung wird erst beim nächsten Neustart des Netzwerks wirksam. Von Hand starten Sie Squid mit

```
/sbin/init.d/squid start
```

Die für den laufenden Betrieb benötigten Ordner und Dateien legt Squid beim ersten Start selbstständig an.

Squid konfiguriert man über die 1.900 Zeilen große Datei `squid.conf`, deren größter Teil aus Kommentaren und Dokumentation besteht.

`/etc/squid.conf` (Auszug):

```
#  TAG: emulate_httpd_log      on|off
# The Cache can emulate the log file format which many 'httpd'
# programs use.  To disable/enable this emulation, set
# emulate_httpd_log to 'off' or 'on'.  The default
# is to use the native log format since it includes useful
#  information that Squid-specific log analysers use.
#
#emulate_httpd_log off
```

Die ersten sieben Zeilen sind Kommentartext, erkennbar an dem einleitenden # Zeichen. Der Kommentar erklärt die Schalter. Der Schalter selber ist hier durch ein # deaktiviert, wodurch die Vorgabe `emulate_httpd_log off` gilt. Will man die Vorgabe ändern, so muss man den Schalter durch Entfernen des Kommentarzeichens aktivieren und `off` durch `on` ersetzen.

Um die Vorgaben individuell einzustellen, sollte man die Konfigurationsdatei sorgfältig bearbeiten. Insbesondere sollte man

- die Größe des Cache im laufenden Betrieb beobachten (s. Logdateien des Squid) und
- den tatsächlichen Bedürfnissen anpassen (s.u.).

In der aktuellen Distribution hat SuSE den Squid so weit auf Sicherheit getrimmt, dass er auch Zugriffe aus dem lokalen Netz ablehnt.

`/etc/squid.conf` (Auszug ab Zeile 992):

```
#Defaults:

acl all src 0.0.0.0/0.0.0.0
acl manager proto cache_object
acl localhost src 127.0.0.1/255.255.255.255
acl SSL_ports port 443 563
acl Safe_ports port 80 21 443 563 70 210 1025-65535
acl CONNECT method CONNECT

#  TAG: http_access
#     Allowing or Denying access based on defined access lists
#
#     Access to the HTTP port:
#     http_access allow|deny [!]aclname ...
#
#     Access to the ICP port:
#     icp_access  allow|deny [!]aclname ...
#
#     NOTE on default values:
#
#     If there are no "access" lines
#     present, the default is to allow
#     the request.
#
#     If none of the "access" lines cause a match,
#     the default is the
```

```
#     opposite of the last line in
#     the list.  If the last line was
#     deny, then the default is allow.
#     Conversely, if the last line
#     is allow, the default will be deny.
#     For these reasons, it is a
#     good idea to have an "deny all"
#     or "allow all" entry at the end
#     of your access lists to avoid potential confusion.
#
#Default configuration:
http_access allow manager localhost
http_access deny manager
http_access deny !Safe_ports
http_access deny CONNECT !SSL_ports
#
# INSERT YOUR OWN RULE(S) HERE TO
# ALLOW ACCESS FROM YOUR CLIENTS
#
http_access deny all
```

Die Regel in der letzten Zeile aus diesem Ausschnitt verbietet jeglichen Zugriff per HTTP, wenn ihn bis dahin nicht eine andere Regel erlaubt hat.

Ersetzen Sie diese Zeile durch

```
http_access allow all
```

und veranlassen Sie den Squid, seine Konfigurationsdatei neu einzulesen:

```
/sbin/init.d/squid reload
```

Nach dieser Änderung stellt der Squid seine Dienste im lokalen Netz zur Verfügung.

13.4 Zugriffskontrolle durch den Proxy-Cache

Squid kann jeglichen Zugriff auf Internetadressen ausschließen, die Systembetreiber als unerwünscht einstufen:

Um einzelne Server, hier die fiktiven Rechner `www.chat-server.de` und `www.chat-dienst.de` vollständig zu sperren, richtet man in `squid.conf` eine Zugriffsregel (`Access List=acl`) ein:

```
acl chat dstdomain www.chat-server.de  www.chat-dienst.de
```

Hinter dem Schlüsselwort `acl` folgt erst ein frei definierbarer Name für diese Regel, dann deren Gültigkeitstyp und danach eine Aufzählung der zu sperrenden Adressen.

Den in `squid.conf` bereits voreingestellten `acl`-Zeilen, fügt man eigene einfach hinzu.

Die so definiert Regel muss man noch aktivieren:

```
http_access deny chat
```

Dadurch verweigert Squid Zugriff auf alle Seiten, auf die die Regel zutrifft. Diese Zeile muss vor der Zeile

```
http_access allow all
```

stehen.

Nach diesen Änderungen muss der Squid mit

```
/sbin/init.d/squid reload
```

seine Konfigurationsdatei neu einlesen.

Sobald die Sperren aktiv sind, zeigt der Browser des Clients beim Versuch, gesperrte Seiten aufzurufen, eine Fehlermeldung.

Abbildung 13.5: Zugriffsverweigerung bei gesperrten Seiten

Nach diesen Änderungen hat der besprochene Abschnitt der Konfigurationsdatei folgendes Aussehen:

`/etc/squid.conf` (Auszug ab Zeile 992 nach Veränderungen):

```
#Defaults:
acl all src 0.0.0.0/0.0.0.0
acl manager proto cache_object
acl localhost src 127.0.0.1/255.255.255.255
acl SSL_ports port 443 563
acl Safe_ports port 80 21 443 563 70 210 1025-65535
acl CONNECT method CONNECT
acl chat dstdomain www.chat-server.de  www.chat-dienst.de

#  TAG: http_access
#     Allowing or Denying access based on defined access lists
#
#     Access to the HTTP port:
#     http_access allow|deny [!]aclname ...
#
#     Access to the ICP port:
#     icp_access  allow|deny [!]aclname ...
#
#     NOTE on default values:
#
#     If there are no "access" lines
#     present, the default is to allow
#     the request.
#
#     If none of the "access" lines
#     cause a match, the default is the
#     opposite of the last line in the
#     list.  If the last line was
#     deny, then the default is allow.
#     Conversely, if the last line
#     is allow, the default will be deny.
#     For these reasons, it is a
#     good idea to have an "deny all" or
#     "allow all" entry at the end
#     of your access lists to avoid potential confusion.
#
# Default configuration:
http_access allow manager localhost
http_access deny manager
```

```
http_access deny !Safe_ports
http_access deny CONNECT !SSL_ports
#
# INSERT YOUR OWN RULE(S) HERE TO
# ALLOW ACCESS FROM YOUR CLIENTS
#
http_access deny chat
http_access allow all
```

Bei einer Festverbindung ins Internet sollten Sie aus Sicherheitsgründen die Default-Regel nicht wie hier beschrieben auf

```
http_access allow all
```

setzen, denn dann können alle Internetnutzer auf Ihren Squid zugreifen. In diesem Fall ist es sicherer, eine neue Zugriffsregel für Ihr lokales Netz zu erstellen und über diese Regel den Zugriff zu erlauben.

Wenn Ihr lokales Netz den Adressbereich 192.168.1.xx benutzt, dann müssen Sie nur Rechner, deren IP-Adresse mit 192.168.1 beginnt, den Zugriff erlauben. Die Netzwerkmaske ist also 255.255.255.0 bzw. 24.

```
acl lokal src 192.168.1.0/255.255.255.0
➥ 127.0.0.1/255.255.255.255
```

Diese Zeile können Sie nach der `acl chat` einfügen.

Die letzten Zeilen des Ausschnittes aus der Konfigurationsdatei lauten dann:

```
http_access deny chat
http_access allow lokal
http_access deny all
```

Damit haben Sie Ihren Squid gegen Zugriffe von Rechnern aus fremden Netzen geschützt.

13.5 Browser der (Windows)-Clients einstellen

Clients müssen in ihren Browsern den Proxy-Cache aktivieren, damit sie ihn nutzen können. Dazu muss man im jeweiligen Browser die IP-Adresse des Proxy-Servers und seine Portnummer (voreingestellt 3128) eintragen.

Den Netscape Communicator konfiguriert man mit:

Bearbeiten • Einstellungen • Erweitert • Proxies • Manuelle Proxy-Konfiguration

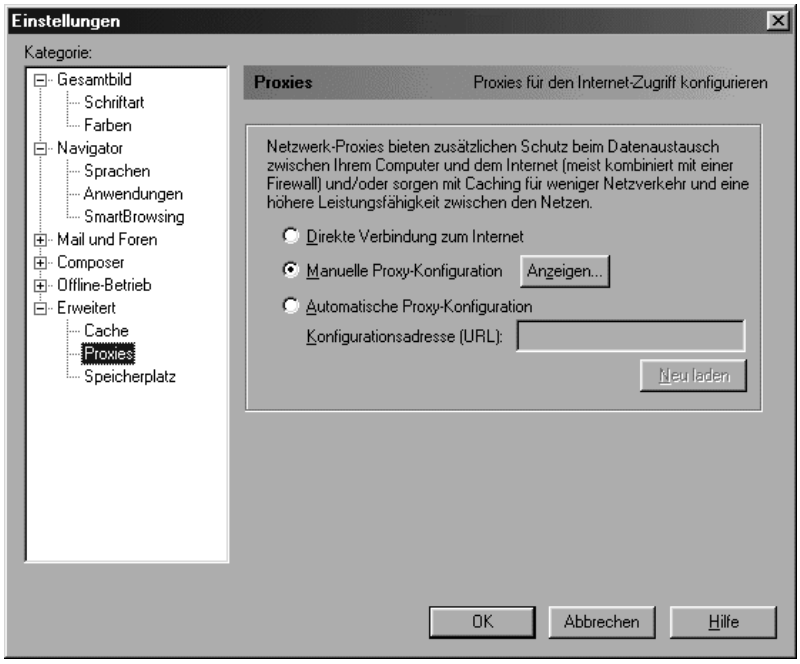

Abbildung 13.6: Einstellungen im Netscape Communicator

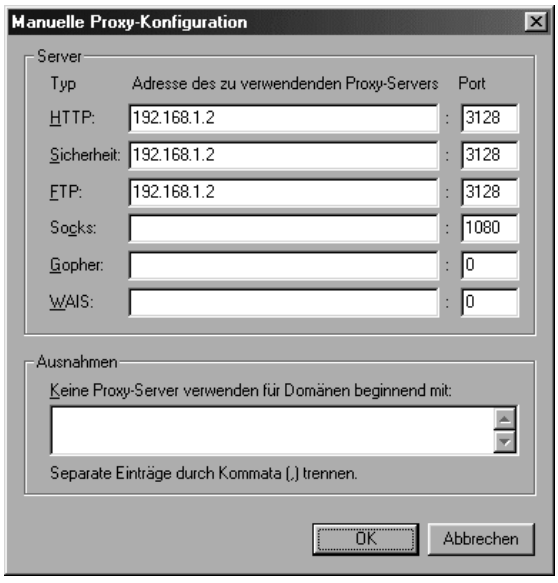

Abbildung 13.7: Manuelle Proxy-Konfiguration im Netscape Communicator

Für HTTP, HTTPS (Sicherheit) und FTP gibt man die IP-Nummer oder den Namen des Kommunikations-Servers und den *Port 3128*, die Voreinstellung von Squid an.

Die restlichen Zeilen bleiben wie voreingestellt. In dem großen Eingabefeld kann man Adressen (im lokalen Netz) angeben, für die der Browser den Proxy nicht benutzten soll.

Beim Microsoft Internet Explorer finden sich die gleichen Einstellmöglichkeiten unter

Ansicht • Internetoptionen • Verbindung

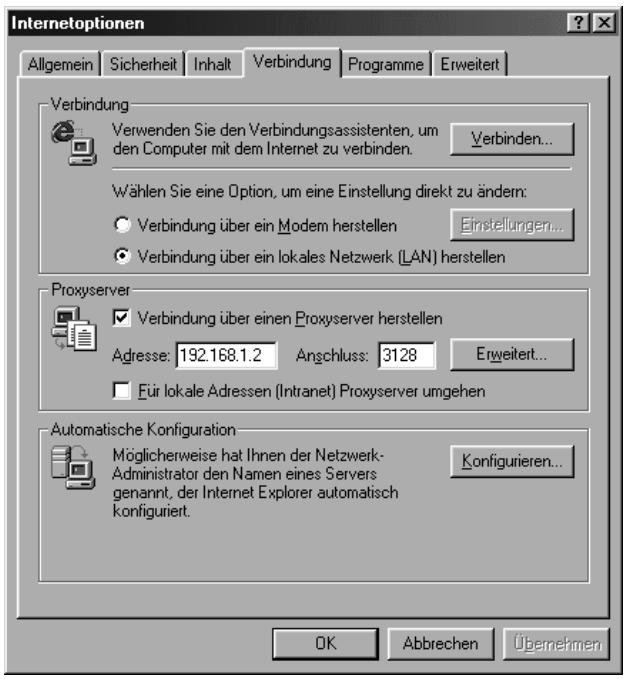

Abbildung 13.8: Menü Verbindung im Internet Explorer

Geht man hier auf *Proxy-Server • Erweitert*, so öffnet der Explorer einen weiteren Dialog mit der praktischen Einstellmöglichkeit *Für alle Protokolle denselben Server verwenden*.

Abbildung 13.9: Menü Proxy-Einstellungen im Internet Explorer

Auch hier kann man wieder die lokalen Adressen ausnehmen.

Achtung: Wenn auf dem Kommunikations-Server IP-Masquerading aktiviert ist, können Anwender den Proxy umgehen, indem sie im Browser die Proxy-Einstellungen deaktivieren.

13.6 Die Logdateien des Squid

Die folgenden Logdateien helfen Systembetreuern, Squid zu überwachen. Die angegebenen Pfade beziehen sich auf SuSE 7.0 und können bei anderen Distributionen abweichen.

Datei	*Bedeutung*
`/var/squid/squid.out`	Startmeldungen.
`/var/run/squid.pid`	Prozess-ID.
`/var/squid/logs/cache.log`	Sehr ausführliche Meldungen und Statistik-Informationen des Squid.
`/var/squid/logs/access.log`	Hier wird jeder einzelne Zugriff auf den Proxy protokolliert. Das Format der Datei ähnelt dem der HTTP-Logdatei.

Datei	*Bedeutung*
`/var/squid/logs/store.log`	Verzeichnis der gespeicherten Dateien mit Speicherort und Web-Quelle.
`/var/squid/cache/*`	Vielzahl von durchnummerierten Verzeichnissen, die den eigentlichen Cache bilden.

Tabelle 13.2: Logdateien des Squid

Normalerweise interessiert es weniger, wo der Squid welche Datei abgelegt hat. Interessant kann es sein, festzustellen, wer welche Internetseiten aufgerufen hat. Dazu muss man sich die Datei `accesss.log` anschauen. Eine typische Zeile sieht folgendermaßen aus:

```
946847924.369   119 192.168.1.40 TCP_HIT/200 490 GET
  http://www.linuxbu.ch/ - NONE/- text/html
```

In der ersten Spalte stecken Datum und Uhrzeit, leider nicht in einem menschenlesbaren Format, sondern als UNIX-Zeit, d.h. als Sekunden seit der Geburt der Programmiersprache C, (dem 1.1.1970). In der dritten Spalte steckt die Information, von welchem Rechner aus die Seite aufgerufen wurde und in der siebten Spalte die URL. Will man diese Datei häufiger kontrollieren, sollte man das Logfile-Format für die Zeitangabe ändern. Aktiviert man in der `squid.conf` den Schalter

```
emulate_httpd_log on,
```

so legt er die Zeitangaben lesbar ab:

```
192.168.1.40 - -[02/Jan/2000:12:54:21 +0100] "GET
  http://www.linuxbu.ch/" 200 487 TCP_MISS:DIRECT
```

Über die Datei `accesss.log` können Sie alle WWW-Zugriffe aus Ihrem Netz heraus nachvollziehen. In der ersten Spalte eines Eintrages steht immer die IP-Adresse des Rechners, der eine Seite aufgerufen hat. Danach folgen Datum und Uhrzeit, sowie die URL des angeforderten Dokumentes. Zuletzt kommen dann noch der Statuscode des Web-Servers, die Dateigröße und ob Squid das Dokument bereits im Cache vorgefunden hat oder nicht.

Bei den umfangreichen Möglichkeiten der Überwachung darf man die geltenden Gesetze und Vorschriften nicht aus dem Auge verlieren. Dazu gehören:

- Bundesdatenschutzgesetz,
- Landesdatenschutzgesetz des jeweiligen Bundeslandes und
- Telekommunikationsgesetz.

Sinnvoll ist es in diesem Zusammenhang mit den Benutzern genaue Regelungen für die Internet-Nutzung und die mögliche Überwachung dieser Regeln zu vereinbaren.

13.7 Cache-Dateien überwachen

In sehr aktiven Umgebungen kann es gelegentlich zu Problemen mit dem Cache kommen. Voreingestellt sind 8 MB Hauptspeicher und 100 MB Festplattenspeicher für den Squid. Wird der Festplattenplatz wirklich ausgenutzt, dann kommt der Squid gelegentlich mit der maximalen Zahl gleichzeitig offener Dateien in Schwierigkeiten. Bei überlasteten Verbindungen ins Internet kann es auch dazu kommen, dass Squid unvollständig geladene Dateien im Cache speichert.

Sollte einer dieser Effekte auftreten oder finden sich in der Datei `/var/log/warn` vermehrt Fehlermeldungen des Squid, so kann man einfach den kompletten Cache löschen. Dazu geht man folgendermaßen vor:

```
/sbin/init.d/squid stop
```

beendet den Squid. Man sollte ihm aber zum Beenden etwas Zeit lassen (mindestens 30 Sekunden), bevor man weitermacht. Die Zeile

```
rm -r /var/squid/cache/*
```

löscht einfach vollständig alle Cache-Ordner.

Vom root-Account aus richtet man die Cache-Ordner neu ein mit:

```
su squid -c "/usr/sbin/squid -z"
```

Danach kann man Squid wieder starten

```
/sbin/init.d/squid start
```

13.8 Auswertung mit Webalizer

Im Abschnitt 13.6 haben Sie gelesen, wie die Logdatei des Squid aufgebaut ist und wie Sie sie analysieren können. Manchmal ist man aber an statistischen Aussagen über die Squid-Nutzung interessiert. Es kann z.B. interessant sein festzustellen, welche Seiten aus dem Internet die Nutzer am häufigsten aufrufen. Eine Auswertung analog zur Auswertung des Webservers Apache macht also Sinn.

Die Datei `/var/squid/logs/access.log` ähnelt in ihrem Aufbau der Logdatei des Webservers Apache, vor allem wenn Sie wie beschrieben die http-Emulation aktivieren.

Daher können Sie auch diese Datei mit dem Programm Webalizer auswerten. Webalizer haben Sie bereits in Kapitel 6 kennen gelernt und vermutlich auch installiert.

Die folgende Beschreibung geht davon aus, dass Sie die Squid-Auswertung zusätzlich zu einer eventuell vorhandenen Webserver-Auswertung nutzen wollen.

Sie müssen ein Verzeichnis einrichten, in das Webalizer die FTP-Statistik ablegen kann, z.B. `/usr/local/httpd/htdocs/squidalizer`:

```
mkdir  /usr/local/httpd/htdocs/squidalizer
```

Die Lage der Logdateien unterscheidet sich zwischen Squid und Apache. Daher müssen Sie für die Squid-Auswertung auch eine spezielle Konfigurationsdatei erstellen.

Zum Erzeugen dieser zweiten Konfigurationsdatei sollten Sie einfach die vorhandene Datei kopieren, z.B. als `squidalizer.conf`:

```
cp  /etc/webalizer.conf  /etc/squidalizer.conf
```

Nun müssen Sie diese Datei für die Pfade des Squid anpassen, damit der Webalizer die richtige Logdatei bearbeitet.

`/etc/squidalizer.conf` (Auszug ab Zeile 29)

```
# LogFile defines the web server log file to use.
# If not specified
# here or on on the command line, input will default to STDIN.

LogFile        /var/squid/logs/access.log

# LogType defines the log type being processed.
# Normally, the Webalizer
# expects a CLF or Combined web server log as input.
# Using this option,
# you can process ftp logs as well (xferlog as produced
# by wu-ftp and others).
# Values can be 'web' or 'ftp', with 'web' the default.

#LogType web

# OutputDir is where you want to put the output files. should
```

```
# This should be a full path name, however relative ones
# might work as well.
# If no output directory is specified, the current directory
will be used.

OutputDir       /usr/local/httpd/htdocs/squidalizer
```

Damit ist die Konfiguration bereits funktionsfähig und Sie können den Webalizer mit dieser Konfigurationsdatei starten:

```
webalizer  -c  /etc/squidalizer.conf
```

Natürlich können Sie auch diese Squid-Daten automatisch auswerten, indem Sie den Programmaufruf in die Cron-Tab von root aufnehmen.