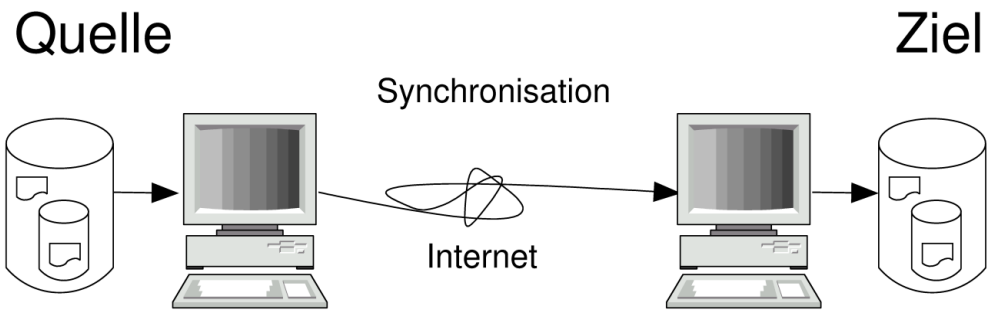
Johannes' mirroring-Seite

Johannes Franken
<jfranken@jfranken.de>

Auf dieser Seite stelle ich verschiedene Möglichkeiten vor, Kopien einer Verzeichnisstruktur unter Verwendung von Internetprotokollen zu erstellen und aktualisieren.

Inhalt

Übersicht



Anforderungen an Kopierprogramme

- Der **Transport** muß in der gegebenen Infrastruktur durchführbar sein:
  - als **lokalen** Transport bezeichne ich Operationen auf Datenträgern, die in das Filesystem gemounted werden können. Das betrifft
    - alle Block-Devices (z.B. über PCI, IDE, SCSI, USB, PCMCIA angeschlossene Festplatten, Disketten, CD/DVD-Laufwerke, Speicherkarten), sowie
    - die Freigaben von Fileservern (z.B. über nfs, smbfs, ncpfs, gfs, coda)
    - Filesysteme auf Loop-Devices
  - Über **http(s),ftp** greift man auf Web- und FTP-Server zu.
  - Eine **pipe** bietet die Möglichkeit, die Kommunikation an ein anderes Programm zu delegieren (z.B. ssh, rsh, rexec, netcat).
- Der **Umfang** einer Kopie bewertet ihre Übereinstimmung mit dem Original. Neben Dateinamen und -inhalt sind folgende Kriterien von Interesse:
  - **delete:** werden bei einer Synchronisation alle Objekte aus der Kopie gelöscht, die in der Quelle nicht mehr vorhanden sind?
  - **filedate:** wird der Zeitpunkt der letzten Änderung der Datei übernommen?
  - **permissions:** werden Benutzer, Gruppe und Berechtigungen übernommen?
- Die **Performance** eines Synchronisationsvorgangs kann durch folgende Maßnahmen wesentlich verbessert werden:
  - Bei einer **differentiellen** Kopie werden nur die Dateien übertragen, die seit der letzten Synchronisation geändert wurden.
  - Bei einer **partiellen** Kopie werden nicht ganze Dateien, sondern nur die Bytes oder Blocks

übertragen, die seit der letzten Synchronisation geändert wurden.
  - Die Dateien vor der Übertragung **komprimieren** und auf der empfangenden Seite dekomprimieren.

Bewertung einiger Kopierprogramme

Programm	Transport				Umfang			Performance		
	lokal	http(s)	ftp	pipe	delete	filedate	permissions	differentiell	partiell	Kompression
wget	-	X	X	-	-	X	-	X	X	-
cp -duRp	X	-	-	-	-	X	X	X	-	-
tar	X	-	-	X	-	X	X	-	-	X
rsync	X	-	-	X	X	X	X	X	X	X

Zusammenfassung: rsync deckt die meisten Anforderungen ab. Wer mit dem Server über http oder ftp kommunizieren muss, sollte wget verwenden.

Details zu den Kopierprogrammen

wget

Beschreibung

wget kopiert Dateien über http, https oder ftp von entsprechenden Servern. Zusätzlich kann es

- (bei http(s):) die in HTML-Dateien enthaltenen Verweise verfolgen und
- (bei ftp:) alle Unterverzeichnisse mitnehmen.

Wenn der Server das Dateidatum übermittelt, übernimmt wget dieses für die empfangengen Dateien und kann so ein erneutes Downloaden bereits vorhandener Dateien vermeiden.
Die erstellte Kopie weicht in folgenden Punkten vom Original ab:

- Dateien, die zwischen zwei Kopiervorgängen auf dem Server gelöscht werden, bleiben in der Kopie erhalten.
- Dateien, auf die bei kein Verweis zeigt, fehlen (bei http).
- Die Permissions (Owner, Gruppe, Rechte) werden nicht mit übertragen.

Aufruf:

```
wget Optionen URL
```

Interessante Optionen sind:

Option	Auswirkung
-N	die Datei nicht downloaden, wenn sie bereits lokal vorliegt und das Dateidatum übereinstimmt.
-nH --cut-dirs=2	Im rekursiven Modus erstellt wget normalerweise Verzeichnisse für den Hostnamen und alle in der URL genannten Unterverzeichnisse. Die Option **-nH** verhindert das Anlegen der Hostverzeichnisse und **--cut-dirs=2** das Anlegen der ersten beiden Verzeichnisse der übergebenen URL. Beispiel: **wget -r -nH --cut-dirs=2 http://www.jfranken.de/homepages/johannes/vortraege** legt als erstes das Verzeichnis **vortraege** an.
-k	ersetzt in HTML-Dateien enthaltene, absolute URLs durch relative. Vorsicht, das funktioniert nicht in allen Situationen.
-r -np	(rekursiv, no-parent): Wenn die übergebene URL eine HTML-Datei liefert, alle von ihr referenzierten Elemente (insb. Verweise und Grafiken) ebenfalls holen und das Verfahren für diese wiederholen. Die Option **-np** verhindert, dass dabei das übergebene Verzeichnis nach oben hin verlassen wird. Referenzen auf andere Hosts werden ignoriert, es sei denn, man gibt den Parameter **-H** an.
-p -l 10	Der Parameter **-l 10** beschränkt die Rekursionstiefe für **-r** auf 10 Ebenen. Ohne Angabe der Rekursionstiefe werden maximal 5 Ebenen verfolgt. Die Angabe **-l 0** bedeutet unendliche Tiefe, und ebensolche Probleme im Filesystem bei zyklischen Verweisen. Der Parameter **-p** verhindert, dass die Grafiken der HTML-Dateien der letzten Ebene fehlen.
-H -Djfranken.de,our-isp.org	Auch Verweisen auf andere Server folgen, allerdings nur in den Domains **jfranken.de** und **our-isp.org**.
-nv	Verhindert die Ausgabe von Debugging-Meldungen.

wget wird seine ftp- und http-Zugriffe automatisch an einen Proxyserver richten. wenn die Environmentvariable **http_proxy** oder **ftp_proxy** gesetzt sind, z.B. mittels

```
$ export http_proxy=http://jfranken:secret@proxy.jfranken.de:3128/
$ export ftp_proxy=$http_proxy
```

Verweise

- wget project page
- wget(1) manpage

rsync

Beschreibung

rsync kann Dateien oder Verzeichnisse synchronisieren über

- das lokale Filesystem (Festplatten, Disketten, CD/DVD-Laufwerke, Speicherkarten, ...)
- Fileserver (z.B. NFS, Windows, Novell), die in das Filesystem gemountet sind.
- eine remote shell (rsh, ssh)
- eine tcp-Verbindung an den rsyncd

Bei Verwendung mit einer remote shell oder eines rsyncd muss die andere Seite lokal sein.
Mit den passenden Parametern aufgerufen, überträgt rsync nur die Veränderungen innerhalb der Dateien, bewahrt Dateiattribute und löscht alle Dateien aus dem Zielverzeichnis, die in der Quelle gelöscht worden sind.
Aufruf:

```
rsync Optionen Quelle(n) Ziel
```

Interessante Optionen sind:

Option	Auswirkung
-a	(archive mode): Kopiert alle Unterverzeichnisse, allerhand Attribute (Symlinks, Rechte, Dateidatum, Gruppe, Devices) und (wenn man root ist) den Eigentümer der Dateien.
-v --progress	(verbose): **-v** gibt während der Übertragung eine Liste der übertragenen Dateien aus. Wenn man zusätzlich **--progress** setzt, zeigt rsync dabei laufend die Zahl der übertragenen Bytes und den Fortschritt in Prozent an.
-n	(dry-run): Nichts schreiben, sondern den Vorgang nur simulieren.
-z -e Programm	Wenn in der Quelle oder dem Ziel ein Doppelpunkt enthalten ist, interpretiert rsync den Teil vor dem Doppelpunkt als Hostnamen und kommuniziert über das mit **-e** spezifizierte Programm, dem es folgende Parameter übergibt: • als Quelle: **hostname rsync --server --sender . Quellpfad** • als Ziel: **hostname rsync --server . Zielpfad** Als Programm bietet sich insbesondere **ssh** an. Wenn man dem Programm weitere Parameter voranstellen möchte, sind diese mit dem Programm in Anführungszeichen zu fassen. Der Parameter **-z** bewirkt, dass rsync die Daten komprimiert überträgt.
--delete --force --delete-excluded	löscht alle Einträge aus dem Zielverzeichnis, die in der Quelle nicht (mehr) vorhanden sind.
--partial	Nach einem Verbindungsabbruch die unvollständig empfangenen Dateien nicht löschen. So kann die Übertragung der Datei bei einem folgenden rsync-Aufruf fortgesetzt werden.
--exclude=Pattern	Die Dateien ignorieren, die dem übergebenen Pattern entsprechen, z.B. **--exclude *~** . Den umfangreichen Exclude-Möglichkeiten widmet sich ein ganzes Kapitel in der rsync(1) manpage.
-x	Schließt alle Dateien auf Filesystemen aus, die in das Quellverzeichnis hineingemountet sind.

Zur Notation von Quelle und Ziel:

- Wenn die Quelle als letztes Zeichen einen Slash (**/**) enthält, kopiert rsync dieses Verzeichnis selbst nicht mit, sondern nur die darin enthaltenen Objekte.
- Ohne Doppelpunkt interpretiert rsync sie als Pfadangaben im Filesystem.
- Wenn sie einen Doppelpunkt enthalten, interpretiert rsync den Teil vor dem Doppelpunkt als Hostnamen und kommuniziert über das mit **-e** spezifizierte Programm.
- Wenn in der Quelle oder dem Ziel zwei aufeinander folgende Doppelpunkte enthalten sind, interpretiert rsync den Teil vor ihnen als Hostnamen und kommuniziert über Port 873 mit einem anderen rsync, den der dortige inetd mit dem Parameter **--daemon** aufruft. Wenn dabei die Environmentvariable **RSYNC_PROXY** den Namen und Port eines Proxyservers enthält, tunnelt rsync seine Kommunikation als https-CONNECT-Anweisungen über diesen Proxy.

Verweise

- rsync project page
- rsync(1) manpage
- rsyncd.conf(5) manpage
- The rsync algorithm

tar

tar war usprünglich zur Datensicherung auf Magnetbänder entwickelt worden. Es wandelt ganze Verzeichnisse in einen Strom um und zurück. Indem man diesen Strom durch pipes auf andere Rechner leitet, kann man Kopien ganzer Verzeichnisse durch ein Internet übertragen.
Die erstellte Kopie weicht in folgendem Punkt vom Original ab:

- Dateien, die zwischen zwei Kopiervorgängen auf dem Server gelöscht werden, bleiben in der Kopie erhalten.

Die folgende Befehlsfolge kopiert das Verzeichnis **`mydir`** in das **`/tmp`**-Verzeichnis des Rechners **`gate`**

```
$ tar cf - mydir/ | ssh gate 'cd /tmp && tar xpvf -'
```

In meinem Vortrag über netcat zeige ich, wie man tar ohne remote shell über beliebige TCP-Verbindungen tunnelt.

cp

`cp` ist das Standardwerkzeug zum Kopieren von Dateien. Mit den Parametern **`-duRp`** kopiert es auch Verweise, Dateirechte, Zeitstempel sowie Unterverzeichnisse und zwar nur für die Dateien, die seit der letzten Kopie hinzugekommen oder verändert worden sind.
Mehr dazu findet sich in der cp(1) manpage