Kapitel 8

Ausfallschutz

Switches und ihre Verbindungen untereinander fallen manchmal aus. Und dann erfüllen sie ihre fundamentalste Aufgabe nicht mehr, die darin besteht, Pakete zu transportieren.
Und Switches fallen genauso gerne aus wie andere elektronische Bauteile. Das ist eine akzeptierte Tatsache und aus diesem Grund haben High-End–Geräte mehrere Netzteile, Lüfter, CPUs oder Uplinks. Zusätzlich hilft man sich meist damit, dass mehrere Switches als Gruppe (Cluster) auftreten. Dann entsteht ein Cluster für Hochverfügbarkeit und Ausfallschutz. Sehr beliebt ist auch die mehrfache Verkabelung zwischen zwei Geräten, um den Defekt einer einzelnen Verbindung abzufangen.

Link Aggregation

Wenn zwei Switches über mehrere Kabel miteinander verbunden sind, wird das *Spanning-Tree Protokoll* (STP, vgl. Kap. 12) aufmerksam und sperrt alle bis auf eine Verbindung. Das ist kein böswilliges Verhalten von STP, sondern die Strategie zum Vermeiden von Schleifen im Netz. Und sobald die einzige genutzte Verbindung ausfällt, wird STP eine der anderen Netzadapter entsperren und die Daten können wieder fließen.
Das Prinzip ist ganz brauchbar, aber *alle* redundanten Leitungen sind inaktiv. Das geht besser, wenn auch bei STP nur mit Tricks. Die vorteilhaftere Methode ist die Bündelung von mehreren physikalischen Leitungen zu einer logischen Portgruppe, wobei jede Leitung aktiv ist. Neben dem Ausfallschutz

steht auch noch zusätzliche Bandbreite zur Verfügung. Für STP gibt es nur noch die eine logische Verbindung und keinen Grund diese zu blockieren. Die verschiedenen Hersteller waren bei der Namensgebung kreativ und die Bezeichnung der Kanalbündelung reicht vom standardisierten *Link Aggregation* über *Bonding* im Linux-Umfeld, *EtherChannel* bei Cisco, *Port Trunk* bei HPE und *Teaming* bei Microsoft Windows.

Grundlagen

Sobald mehrere Leitungen zwischen zwei Geräten als gemeinsamer Kanal arbeiten, hat der Sender die Aufgabe, die ausgehenden Pakete auf die verschiedenen Leitungen zu verteilen. Die parallele Nutzung kann eine höhere Bandbreite erreichen; im Maximum die Summe aller einzelnen Leitungen. Die beiden Endpunkte eines Kanals müssen nicht unbedingt Switches sein. Üblich ist auch die mehrfache Anbindung eines Servers oder Routers an einen Switch.
Für die Bündelung gibt es den allgemein anerkannten Standard *Link Aggregation Control Protocol* (LACP nach IEEE 802.3ad) und häufig noch herstellerspezifische Erweiterungen. OpenSwitch setzt auf LACP ohne weitere Zusätze.
Beide LACP-Partner verhandeln über ihre physikalischen Ports und bilden daraus den logischen Kanal. Im laufenden Betrieb tauschen die Partner kontinuierlich LACP-Pakete aus, um defekte Leitungen zu erkennen oder Änderungen zu propagieren.

Die Voraussetzungen für eine Kanalbündelung sind:

- Alle Leitungen müssen dieselbe Bandbreite haben.
- Alle Leitungen müssen im Vollduplexmodus arbeiten.
- Die Leitungen verbinden exakt zwei Geräte.

Interessanterweise gehört die Kenntnis von LACP nicht zu der Liste, denn ein OpenSwitch-Gerät bündelt auch Verbindungen zu unwissenden Partnern. Der Trick liegt darin, dass beide Geräte die konfigurierten Netzadapter bedingungslos zum Bündel hinzufügen und darauf vertrauen, dass die Gegenstelle dasselbe macht.

Die Anzahl der Ports im Bündel folgt keiner festen Regel. Der Algorithmus zum Verteilen der Last streut die ausgehenden Pakete brav über alle konfigurierten Interfaces, auch wenn die Anzahl ungerade ist.
Was ist mit der alten Daumenregel, dass die Anzahl der Ports immer auf der Basis von Zwei sein muss, um die Last optimal verteilen zu können? Diese Weisheit galt für Switches mit älteren Netzwerkprozessoren, die bei Bündeln aus 3, 5 oder 7 Netzadaptern relativ schief verteilt haben. OpenSwitch hält sich beim Austeilen strikt an den Algorithmus, ohne einen bestimmten Ausgang zu bevorzugen. Auf moderner Hardware sind alle Ports im Bündel gleichmäßig beteiligt.

Laboraufbau

Die Kanalbündelung kann zwischen beliebigen Teilnehmern stattfinden, solange mindestens zwei Kabel von derselben Quelle zum selben Ziel verlaufen. Für den praktischen Anfang bilden die Switches sw01 und sw02 über ihre jeweiligen Netzadapter *e101-005-0* und *e101-006-0* eine gebündelte Leitung (Abbildung 8.1).

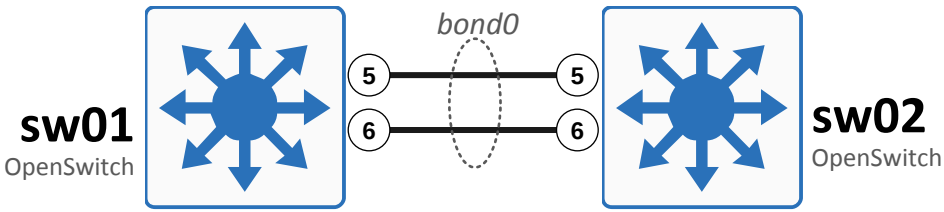

Abbildung 8.1: Die beiden Switches formen einen gemeinsamen Kanal per LACP

Das fertige Bündel erhält einen eigenen Netzadapter mit einem passenden Namen. Die Vorgabe von Linux ist *bond0*, wobei der Name auch den Zweck beschreiben kann, z. B. *bond_sw01_sw02*.

> Die Konfiguration erwartet auf beiden Enden den Namen des Multi-Link–Adapters und seine Teilnehmer.

```
1 ip link del bond0
2 ip link add bond0 type bond mode 802.3ad
```

```
3
4 ip link set e101-005-0 master bond0
5 ip link set e101-006-0 master bond0
6
7 ip link set dev e101-005-0 up
8 ip link set dev e101-006-0 up
9 ip link set dev bond0 up
```

Die letzten drei Zeilen stellt lediglich sicher, dass die Interfaces auch angeschaltet sind. Davor bündelt das passsende opx-Kommando die einzelnen Adapter. Der Verbund ist sofort einsatzbereit:

```
root@sw01:~# opx-show-lag --summary
Ifindex | Name  | Member ports            | Admin state | Oper state
-----------------------------------------------------------------
15      | bond0 | e101-005-0 e101-006-0   | up          | up
```

Ausfallschutz

Die Switches sw01 und sw02 sind nun mehrpfadig verbunden und gewappnet, falls eine einzelne Leitung versagt. Dabei muss es sich nicht um einen physikalischen Defekt handeln. Geplante Umverkabelung im Serverschrank ohne Wartungsfenster ist ebenfalls eine mögliche Ursache für einen unvollständigen Leitungsverbund.
In beiden Fällen bemerkt LACP den Wegfall eines Netzadapters und schickt die Datenpakete über eine andere Leitung. Das LACP-Bündel bleibt im Status *UP*, aber nicht alle Teilnehmer sind bereit für die Arbeit. Die nutzbare Bandbreite reduziert sich um die Bandbreite des havarierten Netzadapters.

```
root@sw01:~# opx-show-interface --port e101-005-0,e101-006-0 --summary
Port       | Enabled | Operational status | Supported speed
-----------------------------------------------------------
e101-005-0 | yes     | down               | 1G 10G 40G
e101-006-0 | yes     | up                 | 1G 10G 40G
```

OpenSwitch protokolliert den Ausfall mit einer knappen Meldung im eigenen Logbuch.

```
Aug 16 14:26:46 sw01 kernel: bond0: link status definitely down \
  for interface e101-005-0, disabling it
```

Im Fehlerfall wünscht sich das Monitoring-Team bestimmt eine Alarmierung und so kann OpenSwitch seinen Besitzer per Syslog und/oder SNMP-Trap benachrichtigen (vgl. Kap. 5).

Lastverteilung

In der Voreinstellung hält sich OpenSwitch brav an den vorgegebenen Algorithmus von Linux. Dieser berücksichtigt Quell- und Ziel-MAC-Adresse. Diese Informationen wandern in eine XOR-Operation und das Ergebnis ist die Nummer des ausgehenden Netzadapters. Folglich werden alle Verbindungen eines Clients zu seinem Server immer über denselben Adapter versendet, denn während einer Verbindung ändern sich die Ethernetadressen nicht.
Der Bonding-Treiber von Linux bietet alternative Methoden für die Lastverteilung, von denen nicht alle kompatibel mit LACP sind. Die *Transmit Hash Policy* `layer2+3` verwendet MAC- und IP-Adressen für die XOR-Operation und harmoniert mit 802.3ad. Abseits des Standards bewegt sich die Policy `layer3+4`, welche Quell- und Ziel-IP-Adresse und – falls vorhanden – die TCP/UDP-Portnummer berücksichtigt.
Die Hash-Policy wird beim Anlegen des neuen Adapters festgelegt. Wenn das nächste Bündel eine erweiterte Lastverteilung benutzen soll, dann lautet das beispielhafte Kommando:

```
ip link add bond1 type bond xmit_hash_policy layer2+3
```

Interoperabilität

Sobald die ersten Switches mit OpenSwitch im eigenen Datacenter an die Tür klopfen, beginnen die Prüfungen zur Verträglichkeit mit der Hausmarke. In der Theorie ist das kein Problem, denn LACP stellt eine gemeinsame Sprache für alle Hersteller dar. Die Praxis bringt kleinere Hürden, die einer der beiden Partner angleichen muss.

Die LACP-Implementierung von OpenSwitch nutzt den Linux-Kernel und hat damit seine Flexibilität. Wenn es beim Zusammenspiel mit Switches

anderer Hersteller zu Problemen kommt, sind zwei Ursachen häufig anzutreffen:

- Die Rate der LACP-Statuspakete ist unterschiedlich. Es gibt zwei Raten: Schnell (jede Sekunde) oder langsam (alle 30 Sekunden) und beide Partner müssen dieselbe Rate nutzen. Empfohlen ist die schnelle Rate.
- Die Konfiguration und Liste der VLANs sind unterschiedlich. Die transportierten VLANs auf beiden Enden des Bündels müssen identisch sein. Das gilt auch für das *Native VLAN*.

Bei LACP gibt es noch den passiven und aktiven Modus, der bestimmt, ob die Aushandlung selbstständig begonnen werden darf, oder nur auf Rückfrage der Gegenstelle. Falls beide Partner passiv bleiben, beginnt keine Verhandlung und die Kanalbündelung bleibt aus. OpenSwitch verzichtet auf den passiven Modus, sodass beide Teilnehmer immer aktiv werden und keine Fehlerquelle darstellen.

Wenn sich beide Partner gar nicht einigen wollen, bietet OpenSwitch eine Alternative: Mit der Option *balance-xor* handelt der Switch nicht mehr nach Standard, sondern deaktiviert LACP und aktiviert alle Netzadapter im Bündel. Die Entscheidung für eine bedingungslose Lastverteilung in einem neuen Bond trifft das Kommando:

```
ip link add bond0 type bond mode balance-xor
```

Technischer Hintergrund

Die IEEE-Norm 802.3ad *Link aggregation* definiert, wie sich zwei Switches verhalten, um Datenpakete über mehrere aktive Leitungen auszutauschen. Der Standard ist seit dem Jahr 2000 verfügbar und alle namhaften Hersteller und Betriebssysteme haben gute Unterstützung dafür. 2008 strukturiert die IEEE um und führt den Standard unter der Bezeichnung 802.1AX fort. Für die Implementierung von LACP setzt OpenSwitch auf den vorhandenen `bonding`-Treiber des Linux-Kernels. Damit übernimmt OpenSwitch alle Fähigkeiten eines Linux-Bonds ohne in zusätzliche Programmierarbeit zu investieren. Der Ansatz ist legitim, da der Treiber unter der Lizenz GPL

steht und die Weiterverwendung gestattet.

> Die Einrichtung eines Bonds übernehmen wahlweise die Kommandos `ip` oder `opx-config-lag`. Falls die Aussage von `opx-show-lag` über Status und Statistik mal zu knapp ist, liefert `/proc/net/bonding/bond0` die volle Palette an Informationen über den Bond (hier `bond0`).

Ausblick

Der Verbund von mehreren Netzadaptern zu einem starken Multi-Gigabit–Bündel ist klasse, hilft aber nicht, wenn der gesamte Switch die Arbeit einstellt. Alternativ könnten ein paar Leitungen des Bündels zu einem weiteren Switch führen, um den Ausfall eines Chassis abzufangen, aber das macht LACP nicht mit.
Der Weg führt zur *Multi-Chassis Link Aggregation* (MLAG), die genau diesen Ansatz erlaubt. Zwei Switches stellen ein *Multi-Chassis*–System dar, welches eine gemeinsame Kanalbündelung zum Partner ermöglicht. Für den Partner sieht das Multi-Chassis-Gerät wie ein einzelner Switch aus, der sogar LACP spricht.
Der Trick bei MLAG ist, dass sich beide Teile des Chassis nach außen als *ein* Switch verkleiden und sich an den LACP-Standard halten. Nach innen gibt es zwischen den Geräten intensive Kommunikation, um den Schein nach außen zu wahren.

MLAG hat sich nie als Standard etabliert, sodass jeder Hersteller seine eigene Implementierung zusammenstrickt.

> Leider gehört OpenSwitch nicht dazu. Im Open-Source–Bereich sind nur wenige Implementierungen bekannt, von denen noch keine ins Linux-Repository eingezogen ist. Wenn der eigene OPX-Switch fit für MLAG werden soll, muss sich eine passende Software finden und im Anschluss kompilieren, einrichten und testen.
>
> - *MC-LAG* aus dem SONiC-Projekt. Die Bemühungen von Microsoft für einen soliden Netzwerkswitch mit Features für Ausfallschutz ver-

langen nach MLAG. Sowohl OpenSwitch als auch SONiC basieren auf Debian Linux, sodass eine Portierung gute Chancen auf Erfolg hat.

- *MLAG* von Cumulus Network. Der Vorreiter von offener Software für Switches im Rechenzentrum hat bereits eine stabile Implementierung für MLAG im eigenen Betriebssystem *Cumulus Linux* eingebaut. Die Software basiert auf Python und ist damit portierbar. Aber Vorsicht: Teile des Programmcodes stehen unter einer proprietären Lizenz, die einen finanziellen Invest erwartet.
- *MLAG* von Open-Ethernet. Endlich eine fertige *und* kostenlose Implementierung. Allerdings ist der Quellcode auf Red Hat Enterprise Linux abgestimmt und liegt bereits seit mehreren Jahren unverändert auf GitHub rum.

Zusammenfassung

OpenSwitch kann mehrere Netzadapter zusammenschalten und alle involvierten Leitungen aktiv benutzen. Damit erreicht der Switch höhere Bandbreiten und gleichzeitig noch Ausfallschutz. Denn wenn eine einzelne Leitung in den Status DOWN wechselt, bleibt das Bündel aktiv und benutzt die verbliebenen Verbindungen für den Datentransport. Bei der Auswahl der Gegenstelle zeigt sich OpenSwitch offen, denn es unterstützt den marktüblichen Standard LACP, der Kanalbündelung herstellerunabhängig macht.