MultiPath

Multiple Paths, Multi-Path, Equal-Cost Multipath (ECMP)

Networks often provide multiple paths from a given source (host) to a given destination (host). Typically, only a single "best" path is used between a given pair of hosts - even when routing metrics ("costs") are equal for two paths towards a given set of destinations, some "tie-breaking" rule is normally used to reduce the number of paths or "next hops" to one. In these setups, multiple paths mostly serve to increase the resilience of the network to failures by providing redundancy.

However, many forwarding devices such as switches and routers now support multi-path forwarding. That is, they can store multiple next-hops or "adjacencies" in their forwarding tables, and distribute traffic across them. Often this is called equal-cost multipath.

Methods for Load Balancing across Multiple Paths

Per-Packet Round-Robin

When distributing traffic across multiple paths, a simple method is per-packet round-robin. In this scheme, the next paths are ordered into a ring, and each subsequent packet is forwarded across the next path. This method can give near-perfect traffic distribution, but has a big drawback: Packets from a single conversation are likely to be distributed across different paths. That leads to packet reordering, which has detrimental influence on performance for many protocols such as TCP.

Hashing Header Fields

In order to avoid reordering of packets belonging to the same conversation, hashing schemes are usually employed: A set of properties - often header fields - are hashed into a numerical value; the hash value is then divided by the number of next-hops, and the remainder (modulus) is used as an index to determine the path to be used.

Early router implementations of this used the destination (IPv4) address as the only basis for the hashing. This implied that all traffic to a single destination would use the same path. When traffic over a path group is dominated by a low number of destination addresses, this can lead to uneven distribution of load.

Later router implementations (such as Cisco's "CEF" or Cisco Express Forwarding) adopted a combination of source and destination IP address as the basis for the hash computation. This lead to a big improvement in traffic distribution.

New devices can include the source and destination port numbers of transport protocols such as TCP or UDP to smoothen traffic distribution even more.

Limitations

ECMP Group Size

The load balancing decision has to be made for each packet, so it is typically implemented with hardware support and a limited time and logic budget. Also, the next-hop information is often tightly coupled with the forwarding table, where space is precious. Therefore there is usually a hard limit on the number of next-hops that a forwarding device can support.

"Uneven" ECMP Groups

Some devices also had issues with next-hop groups of sizes that weren't powers of two: For such "odd" groups, traffic distribution would often be skewed, presumably because they didn't implement a full modulus operation on the hash, but just some kind of bit-masking operation.

Unequal-Cost Multipath

Significant complications occur when traffic should be spread across multiple paths unequally, for example because a link bundle consists of links of different speeds. Routing protocols have to be able to represent capacities, use them properly in route calculations, and expose them to the forwarding engine. A few routing protocols such as EIGRP support this, but other common protocols such as OSPF don't, at least not in their standard version. Also, the forwarding logic needs to be able to use "weights" to distribute traffic unevenly, which requires additional logic in a highly constrained part of the device. Therefore, unequal-cost multipath must be considered a "high-end" feature today.
No Load-Spreading Within Individual Connections

Because all hash-based methods try to keep packets of a single connection on the same path, traditional transport protocols cannot utilize the full bandwidth for a single connection when hashing is used. Specialized protocols such as Multipath TCP (MP-TCP) are required to achieve this.

Encapsulated Traffic

Traffic that is encapsulated through tunneling mechanisms often doesn't distribute well over multiple paths, because its header signature is defined by the outer (encapsulation) header. This issue can be addressed from two angles:

Devices could be made smart enough to recognize encapsulated traffic, and extract fields from the inner header to be used as the basis of the hash. This is only possible when (a) the encapsulation method is known, (b) the headers aren't encrypted and (c) the possible reordering doesn't affect the decapsulation process at the receiving end of the tunnel.

Alternatively, the encapsulation method can encode some "entropy" in the outer headers that can be used for ECMP traffic distribution. In IPv4, the most suitable headers are the port numbers. In IPv6, the flow label can also be used.

Applications

In the past, ECMP has been mostly used with link bundles, combining several lower-speed links to a higher-speed link aggregate. For example, multiple T-1 (1.5 Mb/s) or E-1 (2 Mb/s) links were a common setup for transatlantic connections of some NRENs in the mid/late 1990s.

More recently, multi-path has become a foundation of modern data-center network topologies such as leaf-spine/fat tree.

References

- RFC 7424, Mechanisms for Optimizing Link Aggregation Group (LAG) and Equal-Cost Multipath (ECMP) Component Link Utilization in Networks, R. Krishnan, L. Yong, A. Ghanwani, N. So, B. Khasnabish, January 2015

– Main.SimonLeinen - 2014-12-29 - 2015-01-17