COMPUTER SCIENCE
DATAVETENSKAP

Statefull Layer 4 Load balancing with P4

Load balancing is widely deployed in large web and cloud service infrastructure to map an application with a virtual IP address (VIP) to a server pool with multiple direct IP addresses (DIPs).

Building load balancing function faces two major challenges:
(1) With rapid traffic growth in data centers, how to support full bisection traffic with low latency;
(2) With constantly changing data centers, how to provide the per-connection consistency that always map a connection to the same DIP?

Many data centers today use software load balancers (SLBs), which cannot scale to full bisection traffic with low cost. In this project, we propose to maintain per-connection state at the Netronome card and processes every packet of a VIP connection in the data plane. The design of using Netronome card inherits all the benefits of high-speed low-cost data plane packet processing such as high throughput, low latency and latency variation, and better performance isolation, while ensuring per-connection consistency during DIP pool changes. We build a prototype on Netronome card using P4 and evaluate the performance over different cluster size.

Contact Point: Andreas Kassler, andreas.kassler@kau.se

Company involved: Ericsson Research