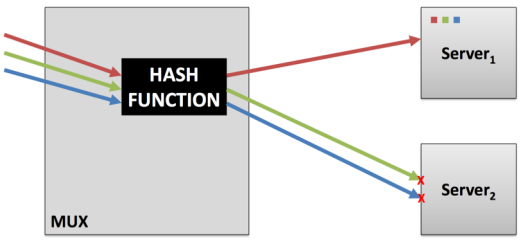
KTH Royal Institute of Technology
Communication Systems Department
Network Systems Lab (NSLab)

Contacts: Prof. Marco Chiesa <mchiesa@kth.se>
Dr. Tom Barbette <barbette@kth.se>

Fully stateless load-balancing

When we access a service online, our requests often end up being served by one server located in a datacenter network.

For scalability reasons, services are deployed over thousands of servers.



Picture taken from [3]

In order to balance the multitude of incoming requests among the servers, load balancers (also called MUXes) are deployed at the edge of a datacenter.

Today's load balancers use simple hash functions that map each single connection request to a specific server. Increasing or decreasing the number of servers is problematic. Hash functions do not keep any information about the active connections. Therefore, connections may be remapped to a different server whenever the number of servers changes. In this case, a connection crashes because the new server does not have any state to handle it.

To mitigate this issue, Google's and Microsoft's load balancers keep state of all active connections [1,2]. This state negatively affect the performance and security of the MUXes [3].

Inspired by the recent work in [3], we aim at exploring fully stateless approaches to load-balancing. The project can be tailored to the student's interests, whether more system-oriented or algorithmic.

[1] D. E. Eisenbud et al. "Maglev: A Fast and Reliable Software Network Load Balancer". In NSDI 2016. https://research.google.com/pubs/archive/44824.pdf

[2] P. Patel et al. "Ananta: cloud scale load balancing". In SIGCOMM 2013. https://dl.acm.org/citation.cfm?id=2486026

[3] V. Olteanu et al. "Stateless Datacenter Load-balancing with Beamer". In NSDI 2018. https://www.usenix.org/conference/nsdi18/presentation/olteanu