# Hey! Presto: Edge-based Load Balancing for Fast Datacenter Networks

Jon Crowcroft
The Computer Laboratory, University of Cambridge
Cambridge, UK
jon.crowcroft@cl.cam.ac.uk

## 1. PUBLIC REVIEW BY JON CROWCROFT

Presto is a host-based system design and implementation to reduce the negative impact of congestion arising from traffic imbalances on throughput, round trip times, and therefore flow completion times. Key design features are that it entails low, per-hypervisor overhead at sender, where load-balancing is achieved by having the vSwitch send traffic along different paths using different labels.

The approach combines known techniques with some new tricks, including

- the use of shadow-MACs to implement label switching, to cause packets belonging to different application layer data units (flowcells) to take different paths,
- an application layer data unit (called a flowcell in the paper), and
- Generic Receiver Offload to deal with packet reordering, and distinguish loss from reordering.
- quick recovery from outages.

The reviewers and PC agreed that the evaluation in this paper was very well executed including detailed comparison for all the metrics (overheads and gains) with conventional use of Equal Cost Multipath routing (ECMP), and with MPTCP (another largely edge-based approach).

The paper contributes further to our knowledge about techniques for moving control out of the switches and routers, and into end-systems, leaving minimal work remaining in the switches/routers in the datacenter network (layer 2 switching, and MAC addresses based forwarding). The approaches for offload (re-ordering) and a per-hypervisor approach also help with scaling CPU costs. The paper offers (but doesn't evaluate) alternative techniques if some of the approaches don't deploy (e.g. tunnels instead of Shadow-MACs). The stated assumptions about the workload and datacenter network topology are clear. Further work could extend the evaluation over more irregular topologies, and importantly, to more complex (and possibly trace-driven) node failure/recovery evaluation, for which they present initial results here for now.

Note for trend watchers. At the time of publication, there are two opposing trends: one moving more functions into the network (Network Function Virtualisation) and one moving more functions into end/edge systems, like this. We will see which come to dominate.