This talk presents ScaleBricks, a new design for building scalable, clustered network appliances that must pin flow state to a specific handling node without being able to choose which node that should be. ScaleBricks applies a new, compact lookup structure to route packets directly to the appropriate handling node, without incurring the cost of multiple hops across the internal interconnect. Its lookup structure is many times smaller than the alternative approach of fully replicating a forwarding table onto all nodes. As a result, Scale is able to improve throughput and latency while simultaneously increasing the total number of flows that can be handled by such a cluster. This architecture is effective in practice: Used to optimize packet forwarding in an existing commercial LTE-to-Internet gateway, it increases the throughput of a four-node cluster by 23%, reduces latency by up to 10%, saves memory, and provides continued (though sub-linear) scaling through 32 nodes.