R2C2: A Network Stack for Rack-scale Computers – Public Review

George Porter
UC San Diego
La Jolla, CA
gmporter@cs.ucsd.edu

For modern Internet data centers, a key challenge to meeting the immense compute, storage, and networking needs of next-generation applications is the ability of the underlying infrastructure to scale. An important new trend in this direction is the introduction of so-called "rack scale" computers, which are large numbers of tightly integrated systems-on-chip (SoC) processors, interconnected with a network fabric. The benefits of rack-scale computing include denser integration, lower cost and power, and larger scale. However, those benefits can only be achieved if the network fabric interconnecting the SoCs is capable of delivering low-latency, high-bandwidth, low congestion and loss packet delivery. This paper presents R2C2, which is a network stack for rack-scale computers.

R2C2 builds upon a large body of work in the design of high-performance interconnects for data centers. However the rack-scale environment that R2C2 targets differs from data center networks in two key ways: (1) the physical network is contained within a single rack, meaning that the one-way propagation delay could theoretically be as small as a few nanoseconds, and (2) the dense integration means that a large number of SoCs are co-located into a single rack, placing enormous cost, power, and cooling pressures on any network fabric that connects those SoCs. This makes FatTree-based switching relying on ToR switches or merchant silicon switch chips impractical. Instead, R2C2 proposes a distributed switch architecture in which SoCs are directly interconnected, and packets are forwarded by multi-hop routing with indirection. Orchestrating the intertwined processes of route selection and congestion control are the primary contribution of R2C2. What makes R2C2's design of interest to the networking community is its use of the low latency present in rack-scale architectures, and its ability to mix routing protocols on a fine-grained basis, even at the per-flow level.

In the first case, R2C2 proposes a low-overhead broadcast primitive based on overlapping trees. The key intuition is that with near zero propagation delay, the cost of "global" control within a rack-scale architecture is minimal. As a result, flow-level events such as the start and end of flows are broadcasted among all nodes in the network, permitting them to recalculate route and rate limiting decisions. This is the key to R2C2's approach to rate selection, which enables each SoC to calculate the rates of each of its flows in a fully distributed way, based on broadcasted flow events. Each SoC then implements a rate-based congestion control algorithm, without the need for probing the network. R2C2 permits a wide diversity of congestion control approaches with different optimization goals, such as max-min fair sharing, proportional sharing, or unfair sharing to implement shortest completion time first scheduling.

In the second case, a number of recent proposals for the data center have outlined different route selection algorithms. R2C2 is unique in permitting independent decision-making of routes at each SoC. By encoding those routes into each packet, R2C2 implements a fully distributed routing approach based on a common view of network conditions, both of topology changes as well as traffic conditions. A major challenge to this type of distribution computation is the combinatorial explosion of all potential paths across each flow. The authors propose a genetic algorithm to address the route selection problem, where each route is scored based on its contribution to the rack's aggregate throughput, and then ranked to prefer those selections that deliver high bandwidth.

R2C2 is situated on a controversial premise, namely that global broadcast is a core feature of the network, and systems built on top of global knowledge of all endpoint events can scale. While this premise is anathema to the assumption of decentralized control that has underpinned network design for decades, its presence, at least in cluster computers and data center networks, is increasingly becoming a reality. The truth is that data center networks are becoming increasingly synchronized for a variety of reasons, and R2C2 explores one extreme in this research space. While somewhat controversial, we feel that the lessons presented in this work might not only be applicable within rack-scale architectures, but could perhaps benefit larger-scale data center networks under the right conditions.