CCIT REPORT #723 March 2009

Mitigating Congestion in High-Speed Interconnects for Computer Clusters

Vladimir Zdornov and Yitzhak Birk

March 2009

Contents

Abstract

Congestion arises in cluster-based supercomputers due to contention for links, spreads due to oversubscription of communication resources, and reduces performance. We mitigate it using efficient, scalable adaptive routing and explicit rate calculation. We use virtual circuits for in-order packet delivery; path setup is performed by switches locally with no blocking or backtracking. For random permutations in a slightly enriched fat-tree topology, maximum contention is reduced by up to 50% relative to static routing, but only rate control can translate this into actual gain. Unfortunately, TCP's window-based rate control fails because of the low bandwidth-delay product, and small buffers moreover cause congestion spreading even with a single-packet window. InfiniBand's CCA employs multiple parameters, which must apparently be tuned per topology and traffic pattern. Instead, we present distributed explicit rate-assignment algorithms, which are designed under three cluster-specific performance goals. Also, we propose a packet injection scheme, and show that desired rates can be realized even with very small buffers. Simulation results suggest that adaptive routing alone offers little benefit, rate control's effectiveness varies with traffic pattern, but together they boost performance by tens of percents. Finally, our explicit algorithms are faster than current reactive schemes.