

Mitigating Congestion in High-Speed Interconnects for Computer Clusters

Vladimir Zdornov and Yitzhak Birk

March 2009

Contents

Abstract	1
1 Introduction	1
1.1 Modern Clusters and Interconnects	1
1.2 Network Model and Terminology	2
1.3 The Congestion Problem	4
1.4 Performance Measures and Goals	6
1.5 Contributions of this Work	7
2 Related Work	10
2.1 Adaptive Routing	10
2.2 Rate Control	12
2.3 Packet Injection	16
3 Adaptive Flow Routing	18
3.1 Routing Scheme	18
3.2 Adaptation Policy	23
3.2.1 The Choice of Heuristic	23
3.2.2 Adaptation Flexibility	24
4 Explicit Rate Calculation	27
4.1 A Single Phase-Based Application	27
4.1.1 The Optimal Completion Time	28
4.1.2 SAA Algorithm	30
4.1.3 SAA-M Algorithm	33

Contents

4.2	Independent Flows	42
4.2.1	Weighted Max-Min Fairness	42
4.2.2	Centralized Algorithm	43
4.2.3	Distributed Algorithm	44
4.3	Multiple Phase-Based Applications	59
4.3.1	Application Rates	59
4.3.2	Centralized Algorithm	60
4.3.3	Distributed Algorithm	64
4.4	Theoretical Comparison	68
4.4.1	Single Application	69
4.4.2	Independent Flows	69
4.4.3	Multiple Applications	71
5	Practical Issues of Rate Control	73
5.1	Mixed Traffic	73
5.2	Packet Injection Scheme	74
5.2.1	The Scheme	75
6	Empirical Results	79
6.1	Fat Trees	79
6.2	Adaptive Routing	81
6.3	A Single Phase-Based Application	87
6.4	Independent Flows	88
6.5	Multiple Phase-Based Applications	92
6.6	Realization of Calculated Rates	96
7	Conclusions	98
	Bibliography	100

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.