RDMA Deadlock

Paper ID: x Pages: 12

1. ANALYSIS

1.1 Definitions

Link dependency. An input link l_j depends on an output link l_i of switch S if a flow is coming from l_j to l_i.

Pause-regulated link. A link is pause regulated if the destination of the link will send PAUSE frame with non-zero probability.

Pause-regulated flow. We say a flow f_j is pause regulated at link l_i if $l_i \in f_j$ and its upstream link is pause-regulated.

Deadlock necessary condition. There exists a circle of links in the network and all of them are pause-regulated.

1.2 How to identify pause-regulated links?

We denote r_j^i is the rate of f_j on link l_i. We denote c_i the capacity of link l_i.

The algorithm is executed iteratively.

Initial state

$\forall f_j, \forall l_i \in f_j = 1$

$\forall c_i = 1$

Stage 1

step 1.1 Identify bottlenecks.

A link l_i is bottleneck link if $\sum_{r_j^i} > c_i$

For all bottleneck links, do

step 1.2 Compute fair share.

Set r_j^i and r_j^k to the fair share on link l_i, for any l_k beyond l_i for f_j.

Loop

Stage 2

Do

step 2.1 Identify pause-regulated link.

Case 1 Actively induced pause: If l_k depends on l_i, and $\exists f_j$, $r_j^k > r_j^i$, then l_k is pause-regulated.

Case 2 Passively induced pause: If l_k and l_m depends on l_i, and l_m is pause-regulated, then l_i is pause-regulated. We call l_k is *pause-correlated* to l_m.

If no new links identified, break

For all pause-regulated links, do

step 2.2 Compute the pause probability.

If l_k is actively depends on l_i, $P^i(l_k) = 1 - \sum r_j^i / \sum r_j^k$, where $f_j \in l_i$.

if l_k is *pause-correlated* to l_m, $P^i(l_k) = \max_m P^i(l_m)$.

Let $P(l_k) = \max_i P^i(l_k)$.

step 2.3 Update flow rates on dependent links. For all pause regulated link, update $r_j^k = r_j^i$, if l_k depends on l_i.

Loop

Loop

Stage 3

step 3.1 Update effective capacity of all pause regulated link.

For all pause regulated links, update $c_k = 1 - P(l_k)$.

Repeat stage 1 and 2 until converge.

Remark At high-level, the algorithm tries to resolve a set of flow rates and pause probability, so that all flow rates fit into link capacity constraints. The stage 1 is to set target rates of each flow to satisfy the link capacity constraints. The second stage tries to compute the back-pressure, and get the pause probability. After this stage, all rates of all flows on all links are set. Finally, the stage 3 updates the effective link capacity according to the pause probability.

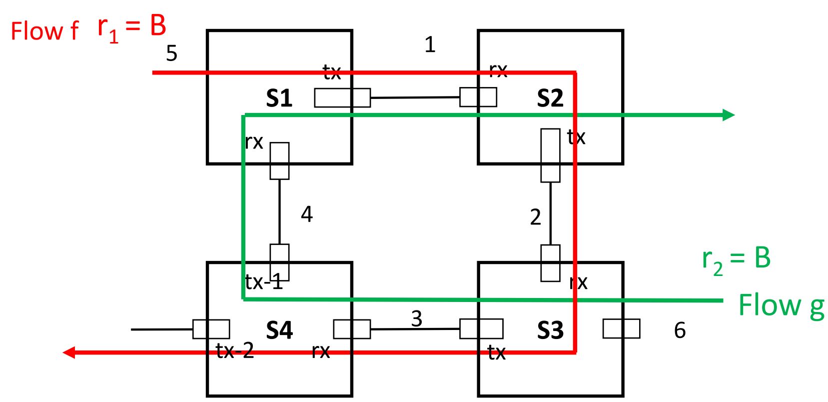
Hypothesis. For any undeadlock scenario, the algorithm should stop for just one iteration.

1.3 Example

Scenario 4. Two flow, four switch

Scenario 5. Three flow, four switch

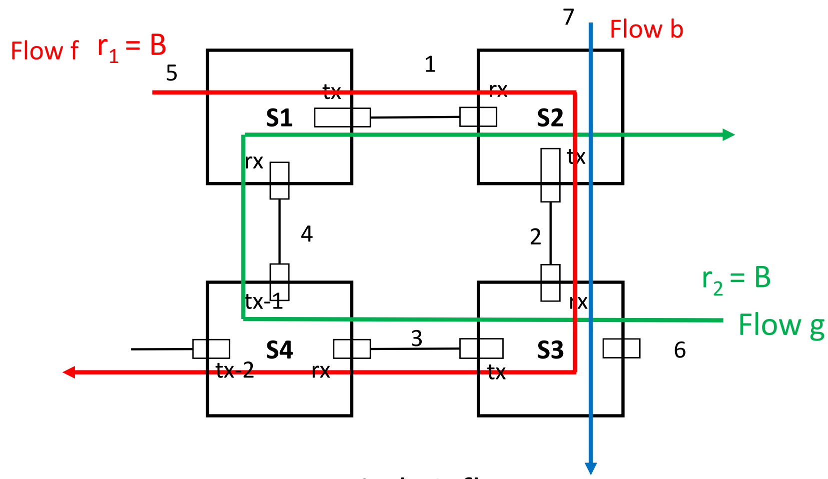
Experiment Scenario 4



4-switch 2-flow case

	f_1	f_2	f_3	f_5	g_1	g_3	g_4	g_6	$c_1(p_1)$	$c_2(p_2)$	$c_3(p_3)$	$c_4(p_4)$	$c_5(p_5)$	$c_6(p_6)$
init	1	1	1	1	1	1	1	1	1 (0)	1(0)	1(0)	1(0)	1(0)	1(0)
Iteration 1														
Stage 1	0.5	0.5	0.5	1	0.5	0.5	0.5	1	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)
Stage 2	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1(0)	1(0.5)	1(0)	1(0.5)	1(0.5)	1(0.5)
Stage 3	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1(0)	0.5(0.5)	1(0)	0.5(0.5)	0.5(0.5)	0.5(0.5)
Stop.														

Experiment Scenario 5



4-switch 3-flow case

	f_1	f_2	f_3	f_5	g_1	g_3	g_4	g_6	b_2	b_7
init	1	1	1	1	1	1	1	1	1	1
Iteration 1										
Stage 1	0.5	0.5	0.5	1	0.5	0.5	0.5	1	0.5	1
Stage 2	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Stage 3	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Iteration 2										
Stage 1	0.25	0.25	0.25	0.5	0.25	0.5	0.5	0.5	0.25	0.5
Stage 2	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
Stage 3	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25
drive to zero										

	$c_1(p_1)$	$c_2(p_2)$	$c_3(p_3)$	$c_4(p_4)$	$c_5(p_5)$	$c_6(p_6)$	$c_7(p_7)$
init	1 (0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)
Iteration 1							
Stage 1	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)
Stage 2	1(0.5)	1(0.5)	1(0)	1(0.5)	1(0.5)	1(0.5)	1(0.5)
Stage 3	0.5(0.5)	0.5(0.5)	1(0)	0.5(0.5)	0.5(0.5)	0.5(0.5)	0.5(0.5)
Iteration 2							
Stage 1	0.5(0.5)	0.5(0.5)	1(0)	0.5(0.5)	0.5(0.5)	0.5(0.5)	0.5(0.5)
Stage 2	0.5(0.75)	0.5(0.75)	1(0.5)	0.5(0.75)	0.5(0.75)	0.5(0.75)	0.5(0.75)
Stage 2	0.25(0.75)	0.25(0.75)	0.5(0.5)	0.25(0.75)	0.25(0.75)	0.25(0.75)	0.25(0.75)
drive to zero							