Network Topology Aware Scheduling of Collective Communications

Emin Gabrielyan, Roger D. Hersch

Swiss Federal Institute of Technology Lausanne
Network Topology Aware Scheduling of Collective Communications

Emin Gabrielyan, Roger D. Hersch

Swiss Federal Institute of Technology Lausanne
25-transmission request
Round-robin schedule
Round-robin Throughput

\[ T_{round\text{-}robin} = \frac{25}{7} \cdot 1 \text{Gbps} = 3.57 \text{Gbps} \]
Liquid schedule

\[ T_{\text{liquid}} = \frac{25}{6} \cdot 1 \text{Gbps} = 4.16 \text{Gbps} \]
Transfers and Load of Links

\[ X = \{ R1, R2, R3, R4, R5 \} \]

The 25 transfer traffic

\[ \lambda(l_1, X) = 5, \ldots \lambda(l_{12}, X) = 6 \]

Transfers: \( \{ l_1, l_6 \}, \ldots \{ l_1, l_{12}, l_9 \}, \ldots \)
Duration of Traffic

\[ \lambda(l_1, X) = 5, \ldots \lambda(l_{10}, X) = 5 \]

\[ \lambda(l_{11}, X) = 5, \ldots \lambda(l_{12}, X) = 6 \]

\[ \Lambda(X) = 6 \]

\[ X = \left\{ \{l_1, l_6\}, \{l_1, l_7\}, \{l_1, l_8\}, \{l_1, l_{12}, l_9\}, \{l_1, l_{12}, l_{10}\}, \{l_2, l_6\}, \{l_2, l_7\}, \{l_2, l_8\}, \{l_2, l_{12}, l_9\}, \{l_2, l_{12}, l_{10}\}, \{l_3, l_6\}, \{l_3, l_7\}, \{l_3, l_8\}, \{l_3, l_{12}, l_9\}, \{l_3, l_{12}, l_{10}\}, \{l_4, l_{11}, l_6\}, \{l_4, l_{11}, l_7\}, \{l_4, l_{11}, l_8\}, \{l_4, l_9\}, \{l_4, l_{10}\}, \{l_5, l_{11}, l_6\}, \{l_5, l_{11}, l_7\}, \{l_5, l_{11}, l_8\}, \{l_5, l_9\}, \{l_5, l_{10}\} \right\} \]

Liquid Throughput

\[
X = \{ \{1, l_6\}, \{1, l_7\}, \{1, l_8\}, \{1, l_{12}, l_9\}, \{1, l_{12}, l_{10}\}, \\
\{2, l_6\}, \{2, l_7\}, \{2, l_8\}, \{2, l_{12}, l_9\}, \{2, l_{12}, l_{10}\}, \\
\{3, l_6\}, \{3, l_7\}, \{3, l_8\}, \{3, l_{12}, l_9\}, \{3, l_{12}, l_{10}\}, \\
\{4, l_{11}, l_6\}, \{4, l_{11}, l_7\}, \{4, l_{11}, l_8\}, \{4, l_9\}, \{4, l_{10}\}, \\
\{5, l_{11}, l_6\}, \{5, l_{11}, l_7\}, \{5, l_{11}, l_8\}, \{5, l_9\}, \{5, l_{10}\} \}
\]

the throughput of a single link

the total number of transfers

\[
T_{\text{liquid}} = \frac{\#(X)}{\Lambda(X)} \cdot T_{\text{link}} = \frac{25}{6} \cdot 1\, \text{Gb/s} = 4.17\, \text{Gb/s}
\]

traffic’s duration (the load of its bottlenecks)
Schedules yielding the liquid throughput

\[ X = \{ \{l_1, l_6\}, \{l_1, l_7\}, \{l_1, l_8\}, \{l_1, l_{12}, l_9\}, \{l_1, l_{12}, l_{10}\}, \{l_2, l_6\}, \{l_2, l_7\}, \{l_2, l_8\}, \{l_2, l_{12}, l_9\}, \{l_2, l_{12}, l_{10}\}, \{l_3, l_6\}, \{l_3, l_7\}, \{l_3, l_8\}, \{l_3, l_{12}, l_9\}, \{l_3, l_{12}, l_{10}\}, \{l_4, l_{11}, l_6\}, \{l_4, l_{11}, l_7\}, \{l_4, l_{11}, l_8\}, \{l_4, l_9\}, \{l_4, l_{10}\}, \{l_5, l_{11}, l_6\}, \{l_5, l_{11}, l_7\}, \{l_5, l_{11}, l_8\}, \{l_5, l_9\}, \{l_5, l_{10}\} \} \]

- Without a right schedule we may have intervals when the access to the bottleneck links is blocked by other transmissions.

- Our goal is to schedule the transfers such that all bottlenecks are always kept occupied ensuring that the liquid throughput is obtained.

- A schedule yielding the liquid throughput we call as a liquid schedule and our objective is to find a liquid schedule whenever it exists.
Swiss-T1 Cluster

- N00 Node
- PR01 Rx Proc
- PR00 Tx Proc
- Routing
- Link

363 Test Traffics

Number of contributing nodes

Liquid throughput (MB/s)
363 Topology Test-bed

Aggregate throughput (MB/s)

Topology (contributing nodes)

Team: a set of mutually non-congesting transfers using all bottlenecks

\[ X = \{ \{l_1, l_6\}, \{l_1, l_7\}, \{l_1, l_8\}, \{l_1, l_{12}, l_9\}, \{l_2, l_{12}, l_9\}, \{l_2, l_{12}, l_{10}\}, \{l_3, l_6\}, \{l_3, l_7\}, \{l_3, l_8\}, \{l_3, l_{12}, l_9\}, \{l_3, l_{12}, l_{10}\}, \{l_4, l_{11}, l_6\}, \{l_4, l_{11}, l_{17}\}, \{l_4, l_{11}, l_8\}, \{l_4, l_9\}, \{l_4, l_{10}\}, \{l_5, l_{11}, l_6\}, \{l_5, l_{11}, l_{17}\}, \{l_5, l_{11}, l_8\}, \{l_5, l_9\}, \{l_5, l_{10}\} \} \]

\[ \alpha = \{ \{l_1, l_{12}, l_9\}, \{l_2, l_7\}, \{l_3, l_8\}, \{l_4, l_{11}, l_6\}, \{l_5, l_{10}\} \}, \{ \{l_1, l_{12}, l_{10}\}, \{l_2, l_6\}, \{l_4, l_{11}, l_7\}, \{l_5, l_9\} \}, \{ \{l_1, l_8\}, \{l_2, l_{12}, l_9\}, \{l_3, l_6\}, \{l_4, l_{10}\}, \{l_5, l_{11}, l_7\} \} \}

\[ \text{schedule } \alpha \text{ is liquid } \iff \text{load of the bottlenecks} \iff \#(\alpha) = \Lambda(X) \iff \forall (A \in \alpha) A \text{ is a team of } X \]
\[ \mathcal{I}(X), \text{ all teams of the traffic } X \]

- transfer \( x \)
- transfers congesting with \( x \)
- transfers non-congesting with \( x \)

- To cover the full solution space when constructing a liquid schedule an efficient technique obtaining the whole set of possible teams of a traffic is required.

- We designed an efficient algorithm enumerating all teams of a traffic traversing each team once and only once.

- This algorithm obtains each team by subsequent partitioning of the set of all teams.

- We introduced triplets consisting of subsets of the traffic, representing one-by-one partitions of the set of all teams.
Liquid schedule search tree

\[ X \rightarrow \emptyset(X) = \{ A_1, A_2, A_3 \ldots A_n \} \]

\[ X_1 = X - A_1 \rightarrow \emptyset(X_1) = \{ A_1, 1, A_1, 2 \ldots \} \]

\[ X_{1, 1} = X_1 - A_1, 1 \]

\[ X_{1, 2} = X_1 - A_1, 2 \]

\[ \ldots \]

\[ X_2 = X - A_2 \rightarrow \emptyset(X_2) = \{ A_2, 1, A_2, 2 \ldots \} \]

\[ X_{2, 1} = X_2 - A_2, 1 \]

\[ X_{2, 2} = X_2 - A_2, 2 \]

\[ \emptyset(Y) = \{ A \in \mathcal{F}(X) | A \subset Y \} \]

all teams of \( X \)

possible steps to the next layer

Additional bottlenecks

\[ A_1 \]
\[ A_{1,1} \]
\[ A_{1,1,1} \]
\[ A_{1,...} \]
\[ A_{1,...} \]
\[ A_{1,...} \]
\[ X \]

2 bottlenecks
\( \Lambda(X) = 6 \)
\( \Lambda(X_1) = 5 \)
4 bottlenecks
\( \Lambda(X_{1,1}) = 4 \)
4 bottlenecks
\( \Lambda(X_{1,...}) = 3 \)
6 bottlenecks
\( \Lambda(X_{1,...}) = 2 \)
8 bottlenecks
\( \Lambda(X_{1,...}) = 1 \)

\[ X_{1,1} = X_1 - A_{1,1} \]
(16 transfers)

\[ X_1 = X - A_1 \]
(20 transfers)

\[ X \]
(25 transfers)
Prediction of dead-ends

\[ A_1 \]
\[ A_{1,1} \]
\[ A_{1,1,1} \]

2 bottlenecks \( \Lambda(X) = 6 \)
2 bottlenecks \( \Lambda(X_1) = 5 \)
4 bottlenecks \( \Lambda(X_{1,1}) = 4 \)

\[ X_{1,1} = X_1 - A_{1,1} \] (16 transfers)
\[ X_1 = X - A_1 \] (20 transfers)
\[ X \] (25 transfers)

load is 4
16-transfer traffic
load is 4
Liquid schedule search optimization

teams of the reduced traffic \( \mathfrak{Z}(Y) \subset \{ A \in \mathfrak{Z}(X) \mid A \subset Y \} \)

original traffic’s teams formed from the reduced traffic

\[
X \rightarrow \varnothing(X) = \{ A_1, A_2, A_3 \ldots A_n \}
\]

\[
X_1 = X - A_1 \rightarrow \varnothing(X_1) = \{ A_1, 1, A_1, 2 \ldots \}
\]

\[
X_{1,1} = X_1 - A_{1,1}
\]

\[
X_{1,2} = X_1 - A_{1,2}
\]

\[
\ldots
\]

\[
X_2 = X - A_2 \rightarrow \varnothing(X_2) = \{ A_2, 1, A_2, 2 \ldots \}
\]

decreasing the search space without affecting the solution space

\[
\varnothing(Y) = \{ A \in \mathfrak{Z}(X) \mid A \subset Y \} \rightarrow \varnothing(Y) = \mathfrak{Z}(Y)
\]
Liquid schedules construction

\[ \mathcal{S}^{\text{full}}(Y) \subset \mathcal{S}(Y) \]

full teams of the reduced traffic

Choice = \( \emptyset(Y) = \mathcal{S}(Y) \)

additionally decreasing the search space without affecting the solution space

Choice = \( \emptyset(Y) = \mathcal{S}^{\text{full}}(Y) \)

For more than 90% of the test-bed topologies construction of a global liquid schedule is completed in a fraction of a second (less than 0.1s).
Results

Number of contributing nodes for the 363 sub-topologies

liquid throughput  •  carried out according to the liquid schedules
The 25 vertices of the graph represent the 25 transfers transfers. The edges represent congestion relations between transfers, i.e. each edge represents one or more communication links shared by two transfers.

**Congestion Graph**

Bold edges represent all congestions due to bottleneck links.
Loss of performance induced by schedules computed with a graph colouring heuristic algorithm.

- For 74% of the topologies Dsatur algorithm does not induce a loss of performance.
- For 18% of topologies, the performance loss is bellow 10%.
- For 8% of topologies, the loss of performance is between 10% and 20%.
Conclusion

• Data exchanges relying on the liquid schedules may be carried out several times faster compared with topology-unaware schedules.

• Thanks to introduced theoretical model we considerably reduce the liquid schedule search space without affecting the solution space.

• Our method may be applied to applications requiring efficiency in concurrent continuous transmissions, such as video and voice traffic management, high energy physics data acquisition and reassembling.

• Liquid scheduling is applicable in wormhole, cut-through networks and can be useful in wavelength assignment problem in WDM optical networks.

Thank You!

Contact: Emin.Gabrielyan@epfl.ch