Network Topology-aware Traffic Scheduling

Emin Gabrielyan

École Polytechnique Fédérale de Lausanne, Switzerland

Emin.Gabrielyan@epfl.ch
25-transfer data exchange
Round-robin schedule
Round-robin Throughput

\[ \frac{257}{100} \text{MBs} / 357 \text{MBs} = 1.25 \text{MBs} / \text{s} \]

$T_{\text{roundrobin}} = \frac{25}{7} \cdot 100 \text{MB/s} = 357 \text{MB/s}$

connection throughput

mean number of connections per timeframe

total throughput

number of transfers

number of timeframes
Liquid Schedule

\[ T_{liquid} = \frac{25}{6} \cdot 100\text{MB/s} = 416\text{MB/s} \]

mean number of connections per step
Load of Links and Transfers

\[ X = \{ T_1, T_2, T_3, T_4, T_5 \} \]

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_6\}, \ldots
Duration of the Traffic

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

\[ \lambda(l_1, X) = 5, \ \lambda(l_2, X) = 5, \ldots \]

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

\[ \Lambda(X) = 6 \]
Liquid Throughput

\[ X = \{ \{1, 1\_6\}, \{1, 1\_7\}, \{1, 1\_8\}, \{1, 1\_12, 1\_9\}, \{1, 1\_12, 1\_10\}, \{1\_2, 1\_6\}, \{1\_2, 1\_7\}, \{1\_2, 1\_8\}, \{1\_2, 1\_12, 1\_9\}, \{1\_2, 1\_12, 1\_10\}, \{1\_3, 1\_6\}, \{1\_3, 1\_7\}, \{1\_3, 1\_8\}, \{1\_3, 1\_12, 1\_9\}, \{1\_3, 1\_12, 1\_10\}, \{1\_4, 1\_11, 1\_6\}, \{1\_4, 1\_11, 1\_7\}, \{1\_4, 1\_11, 1\_8\}, \{1\_4, 1\_9\}, \{1\_4, 1\_10\}, \{1\_5, 1\_11, 1\_6\}, \{1\_5, 1\_11, 1\_7\}, \{1\_5, 1\_11, 1\_8\}, \{1\_5, 1\_9\}, \{1\_5, 1\_10\} \} \]

the throughput of a single link

total number of transfers

\[ T_{\text{liquid}} = \frac{\#(X)}{\Lambda(X)} \cdot T_{\text{link}} = \]

the duration of the traffic (the load of its bottlenecks)

\[ = \frac{25}{6} \cdot 100\text{MB/s} = 417\text{MB/s} \]
No liquid schedule

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

\[ #(X) = 3 \quad \Lambda(X) = 2 \]

\[ T_{\text{liquid}} = \frac{#(X)}{\Lambda(X)} \cdot T_{\text{link}} = \]

\[ = \frac{3}{2} \cdot 100MB/s = 150MB/s \]
363 Test Traffics

Number of contributing nodes vs. Liquid throughput (MB/s)
363-Topology Test-bed

Aggregate throughput (MB/s)

Crossbar throughput

Liquid throughput

0 (0)
30 (9)
60 (11)
90 (12)
120 (14)
150 (15)
180 (16)
210 (18)
240 (19)
270 (20)
300 (22)
330 (24)
360 (30)
Numbers of nodes for the 363 sub-topologies

Round-robin throughput

- T1 Cluster
- Liquid throughput
- Measured round-robin
Team: set of non-congesting transfers using all bottlenecks

\[ X = \{ \{1, l_6\}, \{1, l_7\}, \{1, l_8\}, \{1, l_{12}, l_9\}, \{1, l_{12}, l_{10}\}, \{1, l_6\}, \{1, l_7\}, \{1, l_8\}, \{1, l_{12}, l_9\}, \{1, l_{12}, l_{10}\}, \{1, l_6\}, \{1, l_7\}, \{1, l_8\}, \{1, l_{12}, l_9\}, \{1, l_{12}, l_{10}\}, \{1, l_6\}, \{1, l_{12}, l_{10}\}, \{1, l_6\}, \{1, l_{12}, l_{10}\} \} \]

\[ \alpha = \{ \{1, l_{12}, l_9\}, \{1, l_7\}, \{1, l_8\}, \{1, l_{12}, l_{10}\}, \{1, l_6\}, \{1, l_{12}, l_7\}, \{1, l_{12}, l_9\}, \{1, l_{12}, l_{10}\}, \{1, l_{12}, l_7\}, \{1, l_{12}, l_9\}, \{1, l_{12}, l_{10}\}, \{1, l_{12}, l_7\}, \{1, l_{12}, l_9\}, \{1, l_{12}, l_{10}\} \} \]

schedule \( \alpha \) is liquid \iff

number of steps \( \#(\alpha) = \Lambda(X) \iff \)

\( \forall (A \in \alpha) A \) is a team of \( X \)
Traffic without a team

\[ X = \left\{ \{l_1, l_7, l_8, l_6\}, \{l_2, l_8, l_9, l_4\}, \{l_3, l_9, l_7, l_5\} \right\} \]
Liquid schedule search tree

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

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

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

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

\[ \ldots \]

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

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

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

\[ \ldots \]

\[ X_3 = X - A_3 \rightarrow \text{Choice}(X_3) = \{A_3, 1, A_3, 2\ldots\} \]

\[ X_{3,1} = X_3 - A_{3,1} \]

\[ \ldots \]

\[ \text{Choice}(X_{i_1, i_1 \ldots i_n}) = \{A \in \mathcal{S}(X) | A \subset X_{i_1, i_1 \ldots i_n}\} \]

set of all possible teams of \(X\)
Additional bottlenecks

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

\[ X_1 = X - A_1,1 \] (16 transfers)

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

\[ 2 \text{ bottlenecks} \]
\[ \Lambda(X) = 6 \]

\[ 2 \text{ bottlenecks} \]
\[ \Lambda(X_1) = 5 \]

\[ 4 \text{ bottlenecks} \]
\[ \Lambda(X_{1,1}) = 4 \]

\[ 4 \text{ bottlenecks} \]
\[ \Lambda(X_{1,1,1}) = 3 \]

\[ 6 \text{ bottlenecks} \]

\[ 8 \text{ bottlenecks} \]
Prediction of Dead-ends

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

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

\[
A_1, 2 \text{ bottlenecks } \Lambda(X) = 6
\]

\[
A_{1,1}, 2 \text{ bottlenecks } \Lambda(X_1) = 5
\]

\[
A_{1,1,1}, 4 \text{ bottlenecks } \Lambda(X_{1,1}) = 4
\]

16-transfer traffic

load is 4
Liquid schedule search optimization

Y: reduced traffic

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

teams of the reduced traffic

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

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

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

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

\[ \ldots \]

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

\[ \ldots \]

\[ \text{Choice}(Y) = \{ A \in \mathcal{Y}(X) | A \subset Y \} \]

Choice space decreased without affecting the solution space
Liquid schedules construction

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

full teams of the reduced traffic

\[ Choice(Y) = \mathcal{Y}(Y) \]

\[ Choice(Y) = \mathcal{Y}^{full}(Y) \]

decrease of the search space without affecting the solution space

• For more than 90% of the test-bed topologies the search of liquid schedules took less than 0.1s on a single 500MHz processor.

• For 8 topologies out of 363 solution was not found within 24 hours.
Results

Number of nodes for the 363 sub-topologies

All-to-all throughput (MB/s)
Conclusion

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

- Our method may be applied to applications requiring high network efficiency, such as video or voice traffic management, high energy physics data acquisition and event assembling.

- At the present we consider only static routing scheme. Dynamic routing could possibly be also combined in the algorithms.

- Fixed packet size transfers are considered.

- The network latency are neglected in comparison with the transfer times.

Thank You!

Contact: Emin.Gabrielyan@epfl.ch