ICON 2004 - IEEE International Conference On Networks

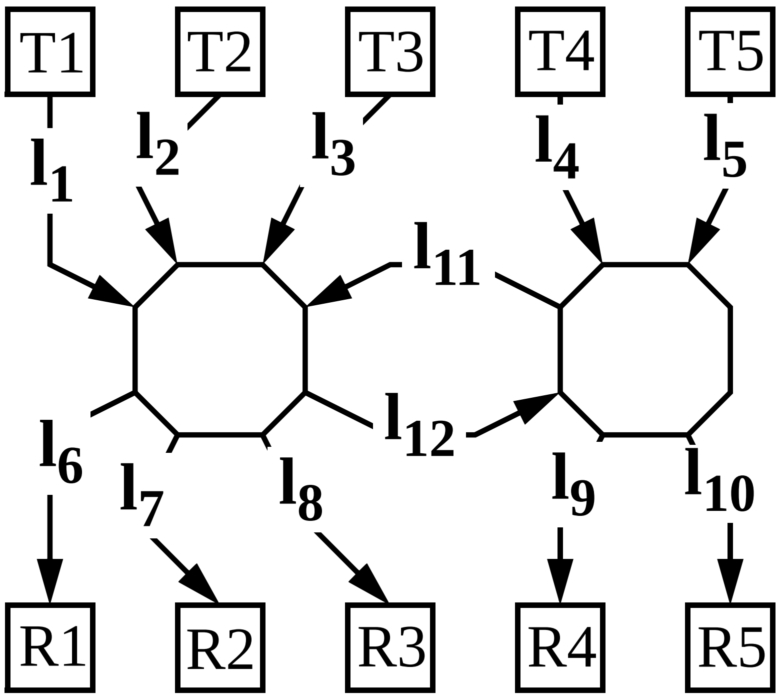
November 16-19, 2004, Singapore, Hilton

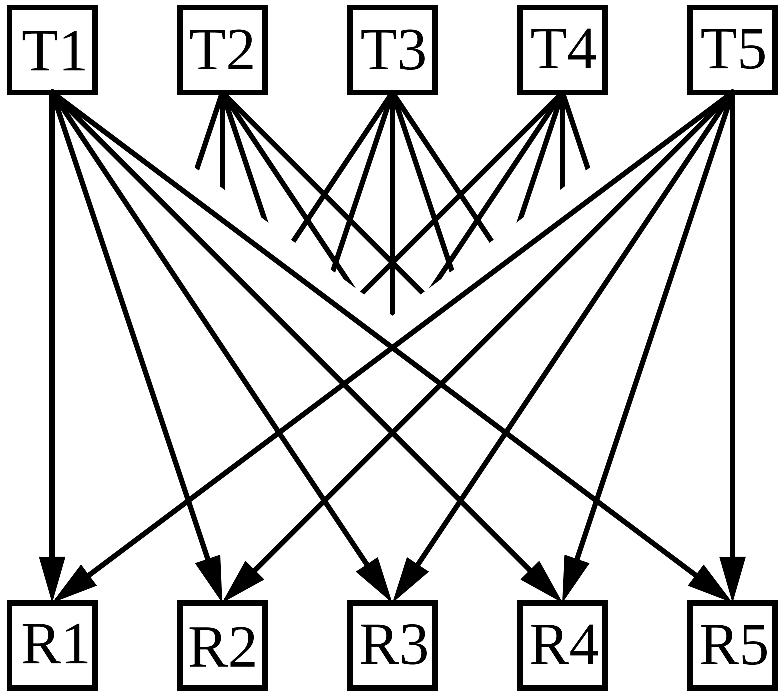
EFFICIENT LIQUID SCHEDULE SEARCH STRATEGIES FOR COLLECTIVE COMMUNICATIONS

Emin Gabrielyan, Roger D. Hersch

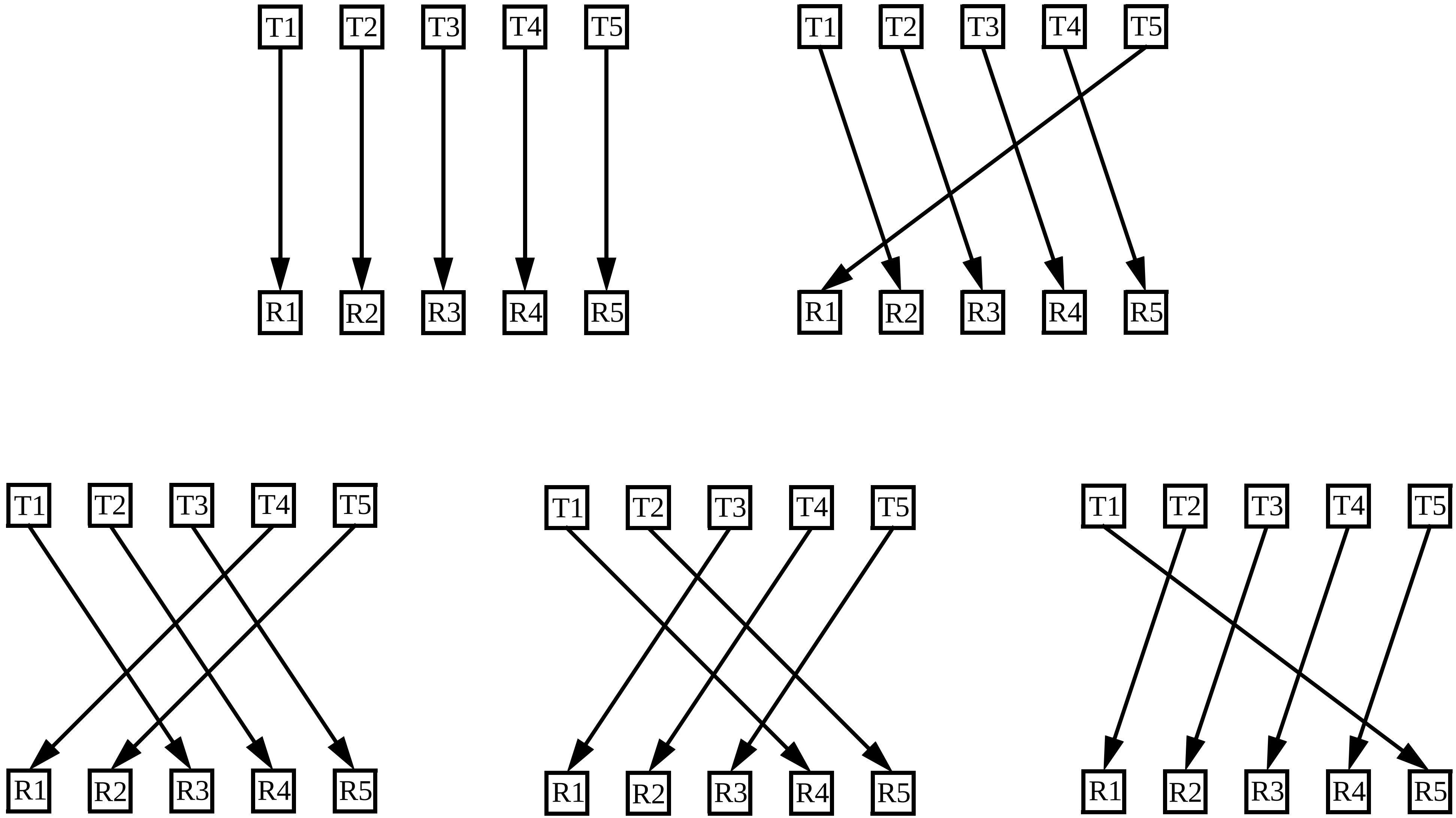
Swiss Federal Institute of Technology - Lausanne

Example: 25 transmissions to be carried out

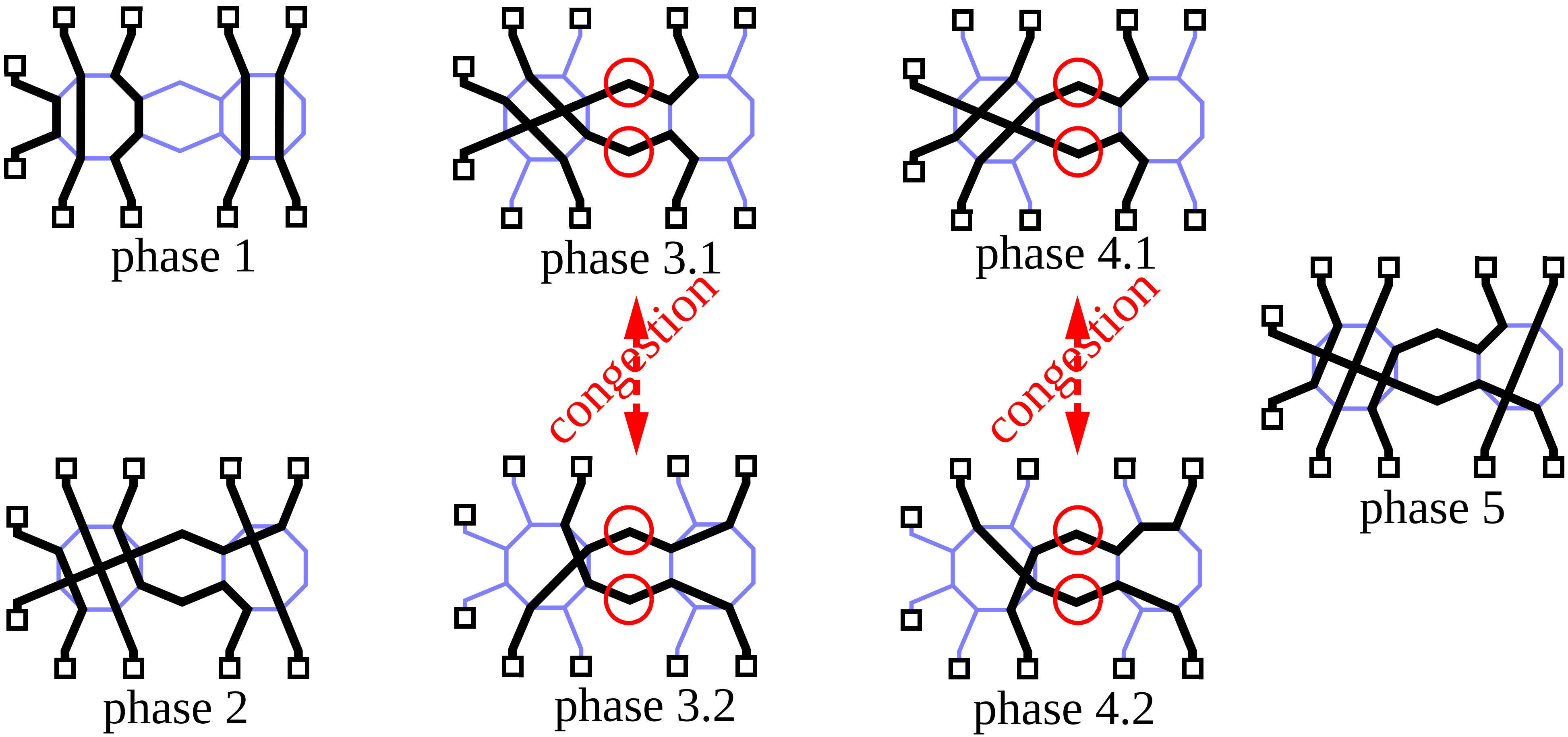




Round-robin schedule

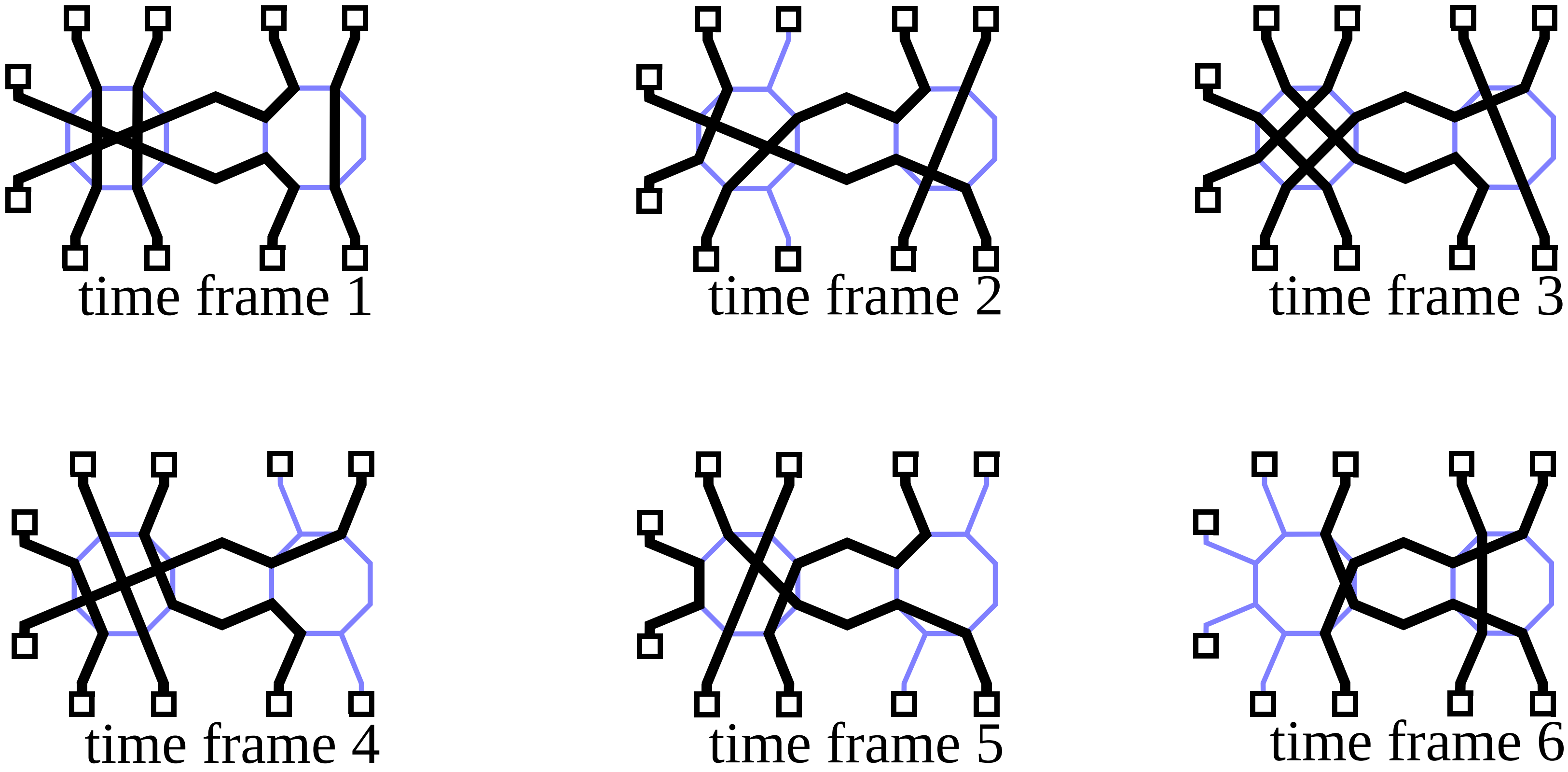


Round-robin Throughput



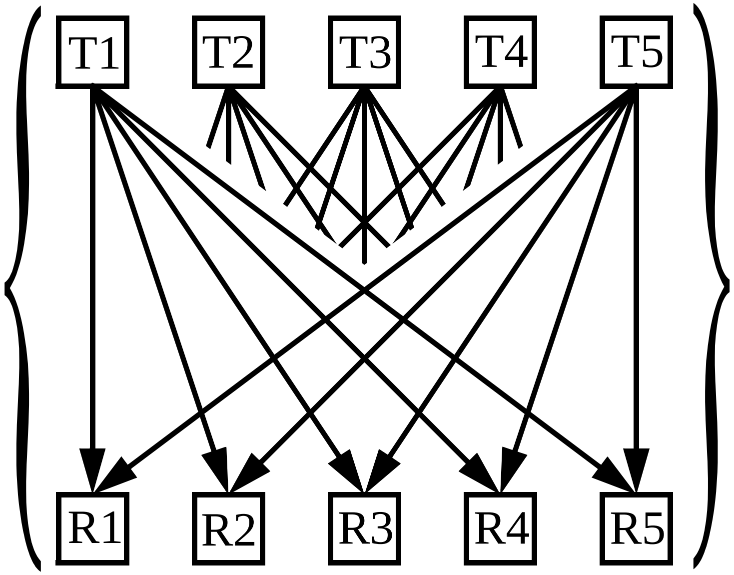
$$T_{roundrobin} = 25/7 \cdot 1Gbps = 3.57Gbps$$

Liquid schedule

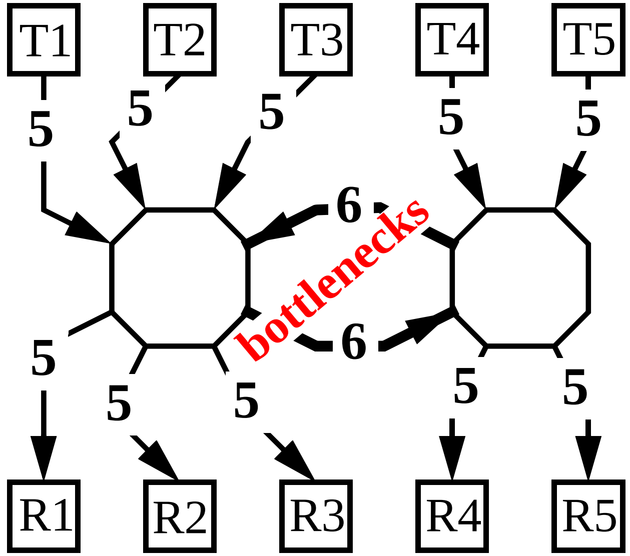


$$T_{liquid} = 25/6 \cdot 1Gbps = 4.16Gbps$$

Transfers and Load of Links

X = 

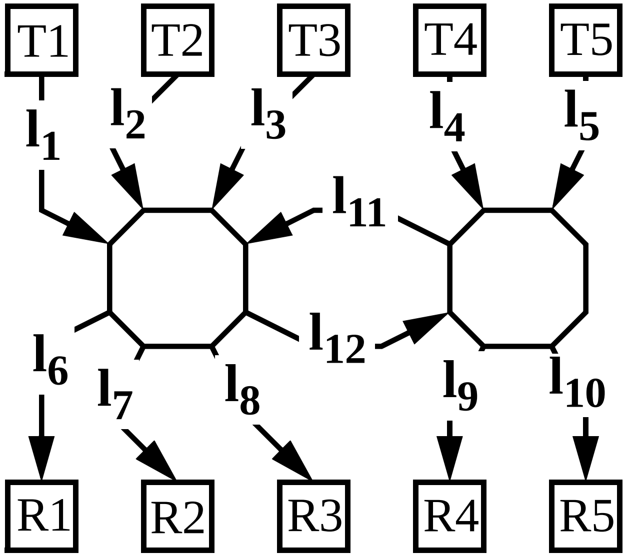
The 25 transfer traffic



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

Transfers: $\{l_1, l_6\}, \ldots \{l_1, l_{12}, l_9\}, \ldots$

Duration of Traffic



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

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

$$\Lambda(X) = 6$$

$$X = \left\{ \begin{array}{c} \{l_1, l_6\}, \{l_1, l_7\}, \{l_1, l_8\}, \{l_1, l_{12}, l_9\}, \{l_1, l_{12}, l_{10}\}, \\ \{l_2, l_6\}, \{l_2, l_7\}, \{l_2, l_8\}, \{l_2, l_{12}, l_9\}, \{l_2, l_{12}, l_{10}\}, \\ \{l_3, l_6\}, \{l_3, l_7\}, \{l_3, l_8\}, \{l_3, l_{12}, l_9\}, \{l_3, l_{12}, l_{10}\}, \\ \{l_4, l_{11}, l_6\}, \{l_4, l_{11}, l_7\}, \{l_4, l_{11}, l_8\}, \{l_4, l_9\}, \{l_4, l_{10}\}, \\ \{l_5, l_{11}, l_6\}, \{l_5, l_{11}, l_7\}, \{l_5, l_{11}, l_8\}, \{l_5, l_9\}, \{l_5, l_{10}\} \end{array} \right\}$$

Liquid Throughput

$$X = \left\{ \begin{array}{c} \{l_1, l_6\}, \{l_1, l_7\}, \{l_1, l_8\}, \{l_1, l_{12}, l_9\}, \{l_1, l_{12}, l_{10}\}, \\ \{l_2, l_6\}, \{l_2, l_7\}, \{l_2, l_8\}, \{l_2, l_{12}, l_9\}, \{l_2, l_{12}, l_{10}\}, \\ \{l_3, l_6\}, \{l_3, l_7\}, \{l_3, l_8\}, \{l_3, l_{12}, l_9\}, \{l_3, l_{12}, l_{10}\}, \\ \{l_4, l_{11}, l_6\}, \{l_4, l_{11}, l_7\}, \{l_4, l_{11}, l_8\}, \{l_4, l_9\}, \{l_4, l_{10}\}, \\ \{l_5, l_{11}, l_6\}, \{l_5, l_{11}, l_7\}, \{l_5, l_{11}, l_8\}, \{l_5, l_9\}, \{l_5, l_{10}\} \end{array} \right\}$$

the throughput of a single link

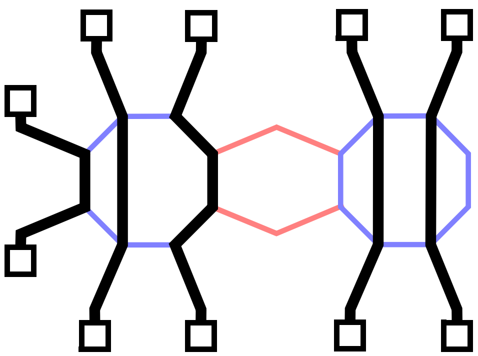
total number of transfers

$$T_{liquid} = \frac{\#(X)}{\Lambda(X)} \cdot T_{link} = \frac{25}{6} \cdot 1Gbps = 4.17Gbps$$

traffic's duration (the load of its bottlenecks)

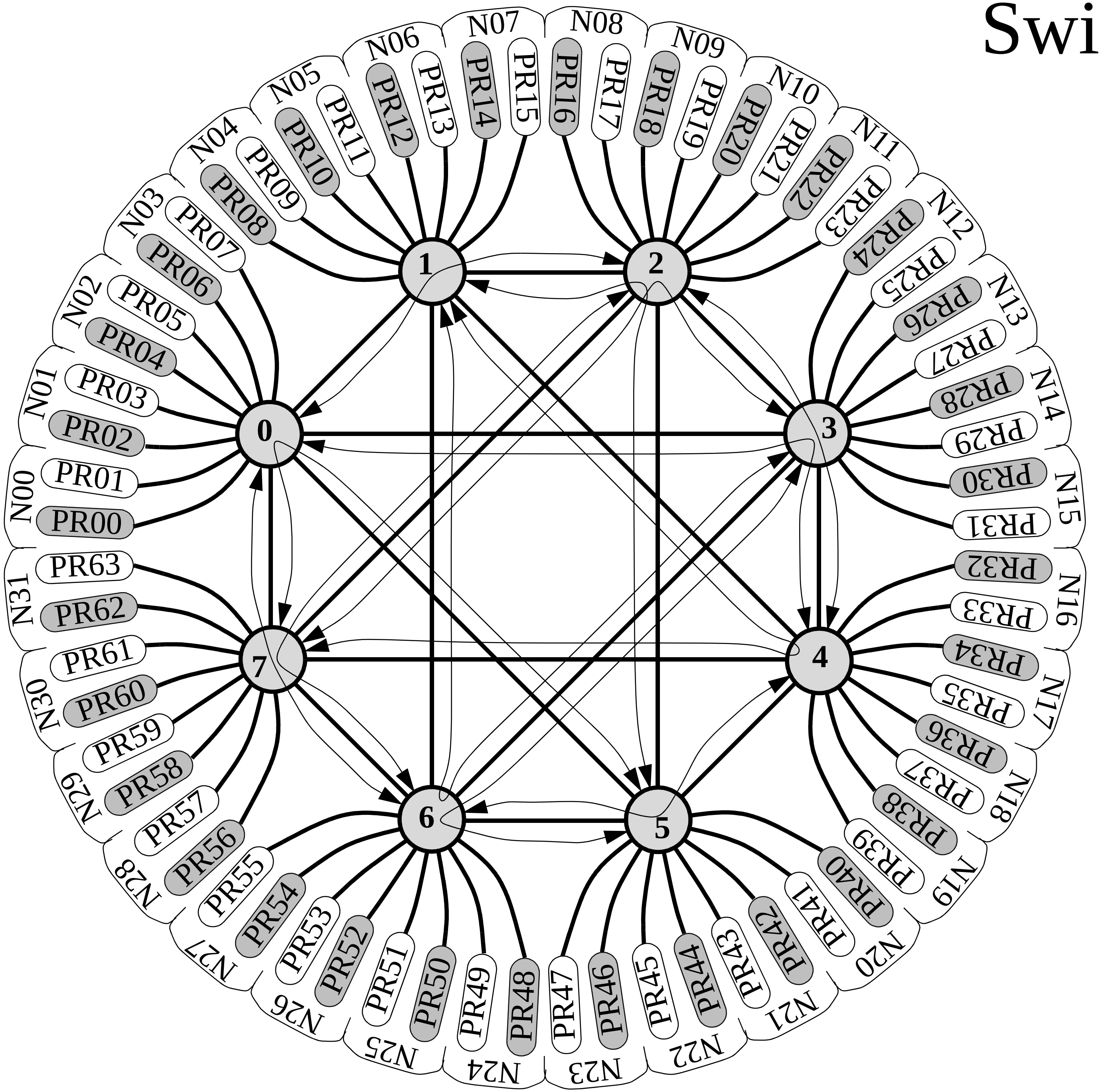
Schedules yielding the liquid throughput

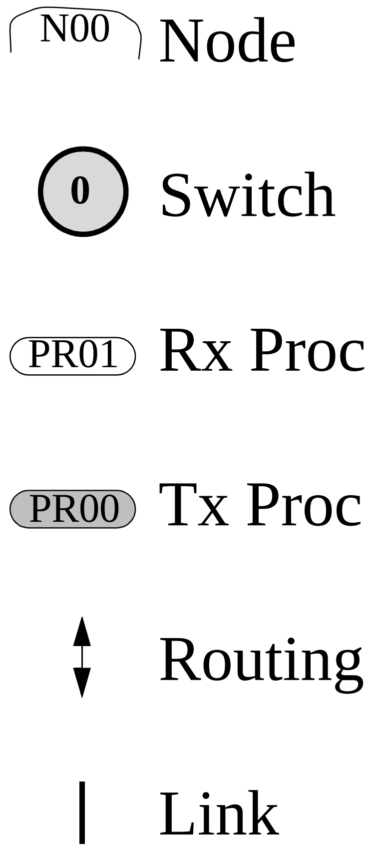
$$X = \left\{ \begin{array}{c} \{l_1, l_6\}, \{l_1, l_7\}, \{l_1, l_8\}, \{l_1, \mathbf{l_{12}}, l_9\}, \{l_1, \mathbf{l_{12}}, l_{10}\}, \\ \{l_2, l_6\}, \{l_2, l_7\}, \{l_2, l_8\}, \{l_2, \mathbf{l_{12}}, l_9\}, \{l_2, \mathbf{l_{12}}, l_{10}\}, \\ \{l_3, l_6\}, \{l_3, l_7\}, \{l_3, l_8\}, \{l_3, \mathbf{l_{12}}, l_9\}, \{l_3, \mathbf{l_{12}}, l_{10}\}, \\ \{l_4, \mathbf{l_{11}}, l_6\}, \{l_4, \mathbf{l_{11}}, l_7\}, \{l_4, \mathbf{l_{11}}, l_8\}, \{l_4, l_9\}, \{l_4, l_{10}\}, \\ \{l_5, \mathbf{l_{11}}, l_6\}, \{l_5, \mathbf{l_{11}}, l_7\}, \{l_5, \mathbf{l_{11}}, l_8\}, \{l_5, l_9\}, \{l_5, l_{10}\} \end{array} \right\}$$



- 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.

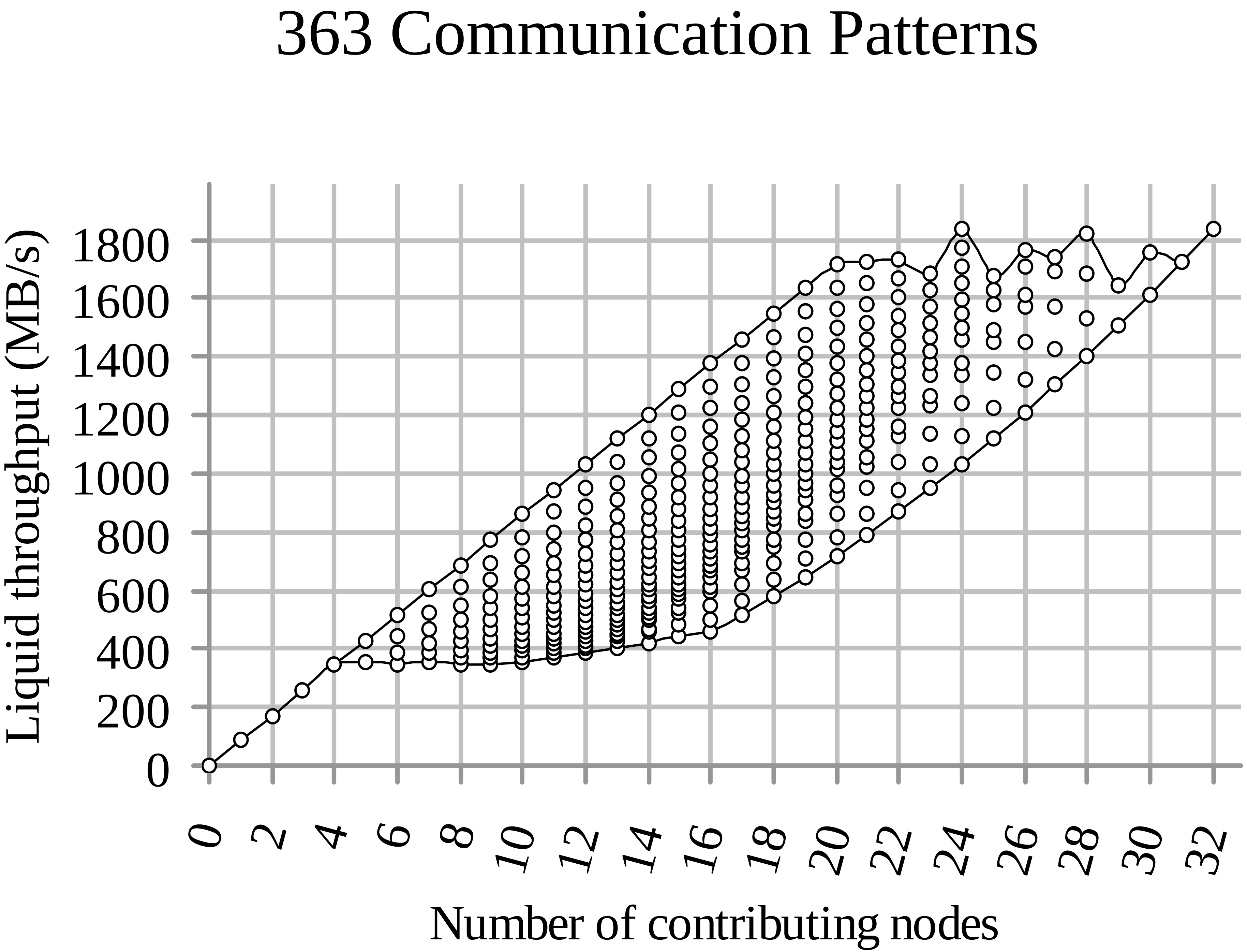
Swiss-T1 Cluster



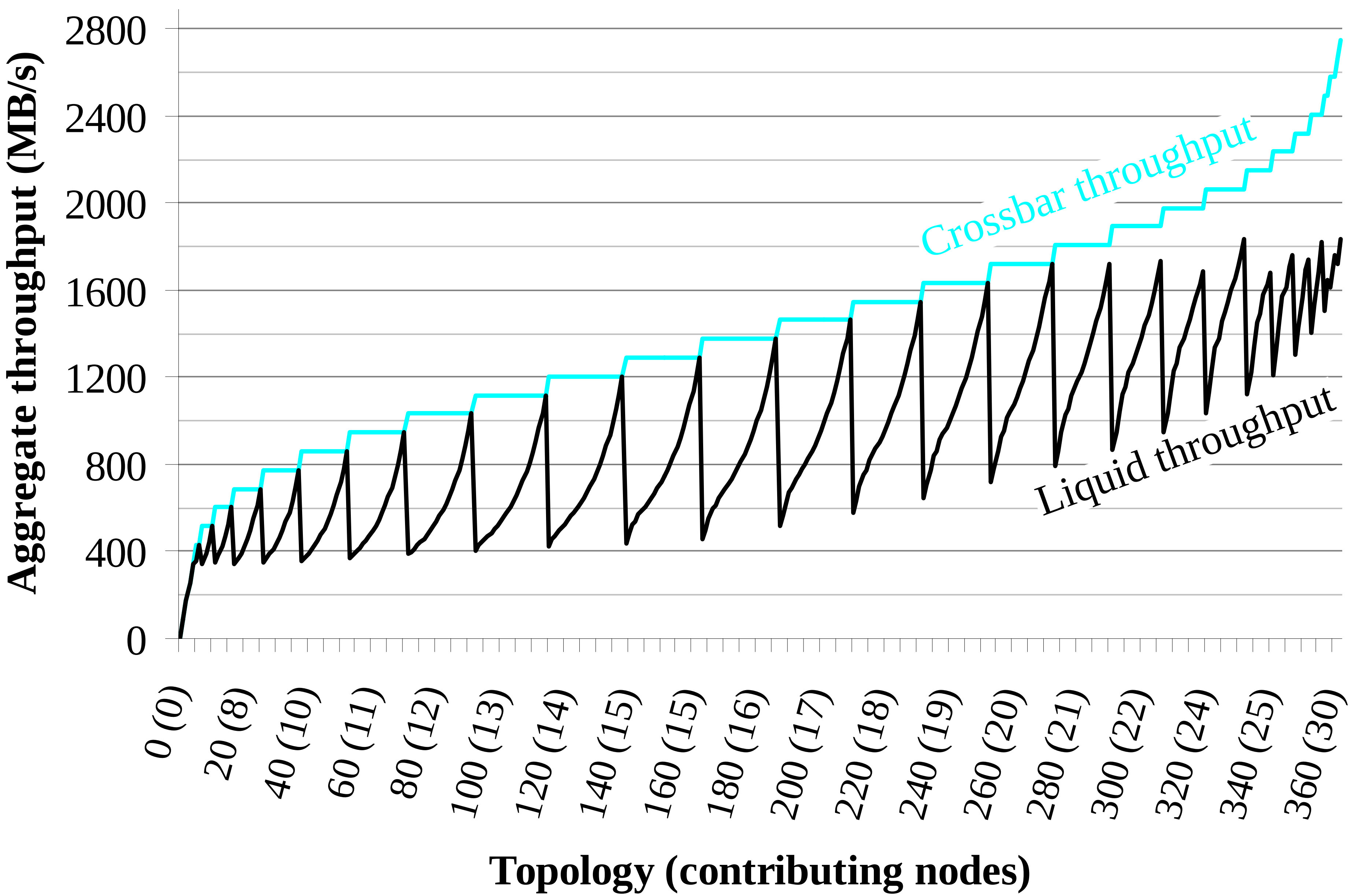


-- ICON 2004, IEEE International Conference On Networks, November 16-19, 2004, Singapore, Hilton --

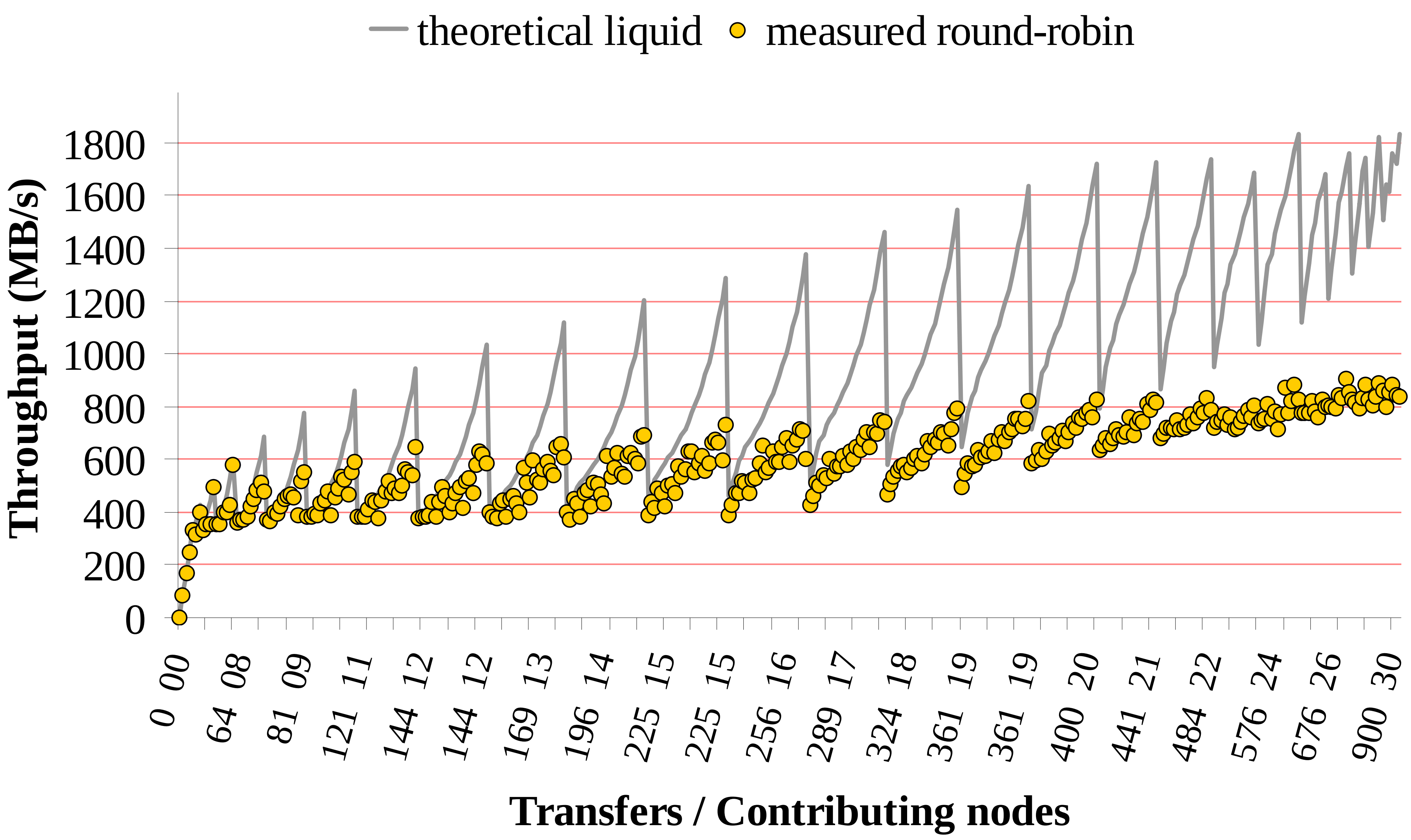
-- EFFICIENT LIQUID SCHEDULE SEARCH STRATEGIES FOR COLLECTIVE COMMUNICATIONS --


363 Communication Patterns
Liquid throughput (MB/s)
0
200
400
600
800
1000
1200
1400
1600
1800
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
Number of contributing nodes

363 Topology Test-bed



Round-robin throughput



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

$$X = \left\{ \begin{array}{c} \{l_1, l_6\}, \{l_1, l_7\}, \{l_1, l_8\}, \{l_1, \mathbf{l_{12}}, l_9\}, \{l_1, \mathbf{l_{12}}, l_{10}\}, \\ \{l_2, l_6\}, \{l_2, l_7\}, \{l_2, l_8\}, \{l_2, \mathbf{l_{12}}, l_9\}, \{l_2, \mathbf{l_{12}}, l_{10}\}, \\ \{l_3, l_6\}, \{l_3, l_7\}, \{l_3, l_8\}, \{l_3, \mathbf{l_{12}}, l_9\}, \{l_3, \mathbf{l_{12}}, l_{10}\}, \\ \{l_4, \mathbf{l_{11}}, l_6\}, \{l_4, \mathbf{l_{11}}, l_7\}, \{l_4, l_{11}, l_8\}, \{l_4, l_9\}, \{l_4, l_{10}\}, \\ \{l_5, \mathbf{l_{11}}, l_6\}, \{l_5, \mathbf{l_{11}}, l_7\}, \{l_5, l_{11}, l_8\}, \{l_5, l_9\}, \{l_5, l_{10}\} \end{array} \right\}$$

$$\alpha = \left\{ \begin{array}{c} \left\{ \begin{array}{c} \{l_1, \mathbf{l_{12}}, l_9\}, \\ \{l_2, l_7\}, \\ \{l_3, l_8\}, \\ \{l_4, \mathbf{l_{11}}, l_6\}, \\ \{l_5, l_{10}\} \end{array} \right\}, \left\{ \begin{array}{c} \{l_1, \mathbf{l_{12}}, l_{10}\}, \\ \{l_2, l_6\}, \\ \{l_4, \mathbf{l_{11}}, l_7\}, \\ \{l_5, l_9\} \end{array} \right\}, \left\{ \begin{array}{c} \{l_1, \mathbf{l_8}\}, \\ \{l_2, \mathbf{l_{12}}, l_9\}, \\ \{\mathbf{l_3}, l_6\}, \\ \{l_4, l_{10}\}, \\ \{l_5, \mathbf{l_{11}}, l_7\} \end{array} \right\}, \\ \left\{ \begin{array}{c} \{l_1, l_7\}, \\ \{l_2, \mathbf{l_8}\}, \\ \{\mathbf{l_3}, \mathbf{l_{12}}, l_9\}, \\ \{l_5, \mathbf{l_{11}}, l_6\} \end{array} \right\}, \left\{ \begin{array}{c} \{l_1, l_6\}, \\ \{l_2, \mathbf{l_{12}}, \mathbf{l_{10}}\}, \\ \{\mathbf{l_3}, l_7\}, \\ \{\mathbf{l_4}, \mathbf{l_{11}}, \mathbf{l_8}\} \end{array} \right\}, \left\{ \begin{array}{c} \{\mathbf{l_3}, \mathbf{l_{12}}, \mathbf{l_{10}}\}, \\ \{\mathbf{l_4}, \mathbf{l_9}\}, \\ \{\mathbf{l_5}, \mathbf{l_{11}}, \mathbf{l_8}\} \end{array} \right\} \end{array} \right\}$$

schedule α is liquid $\Leftrightarrow$

load of the bottlenecks

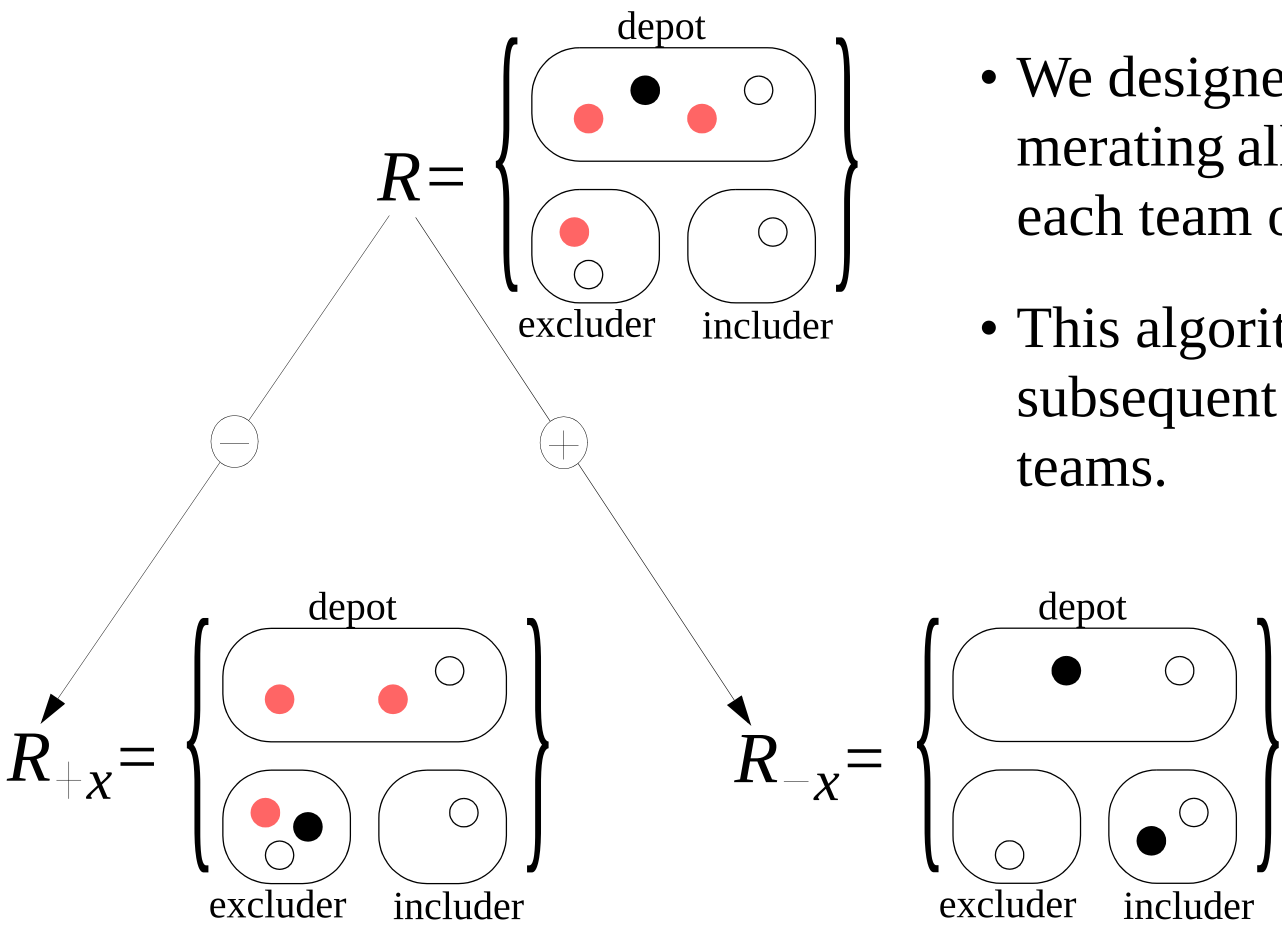
number of timeframes

$$\Leftrightarrow \#(\alpha) = \Lambda(X) \Leftrightarrow$$

$$\Leftrightarrow \forall(A \in \alpha)$$
A is a team of X

$\Im(X)$, 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

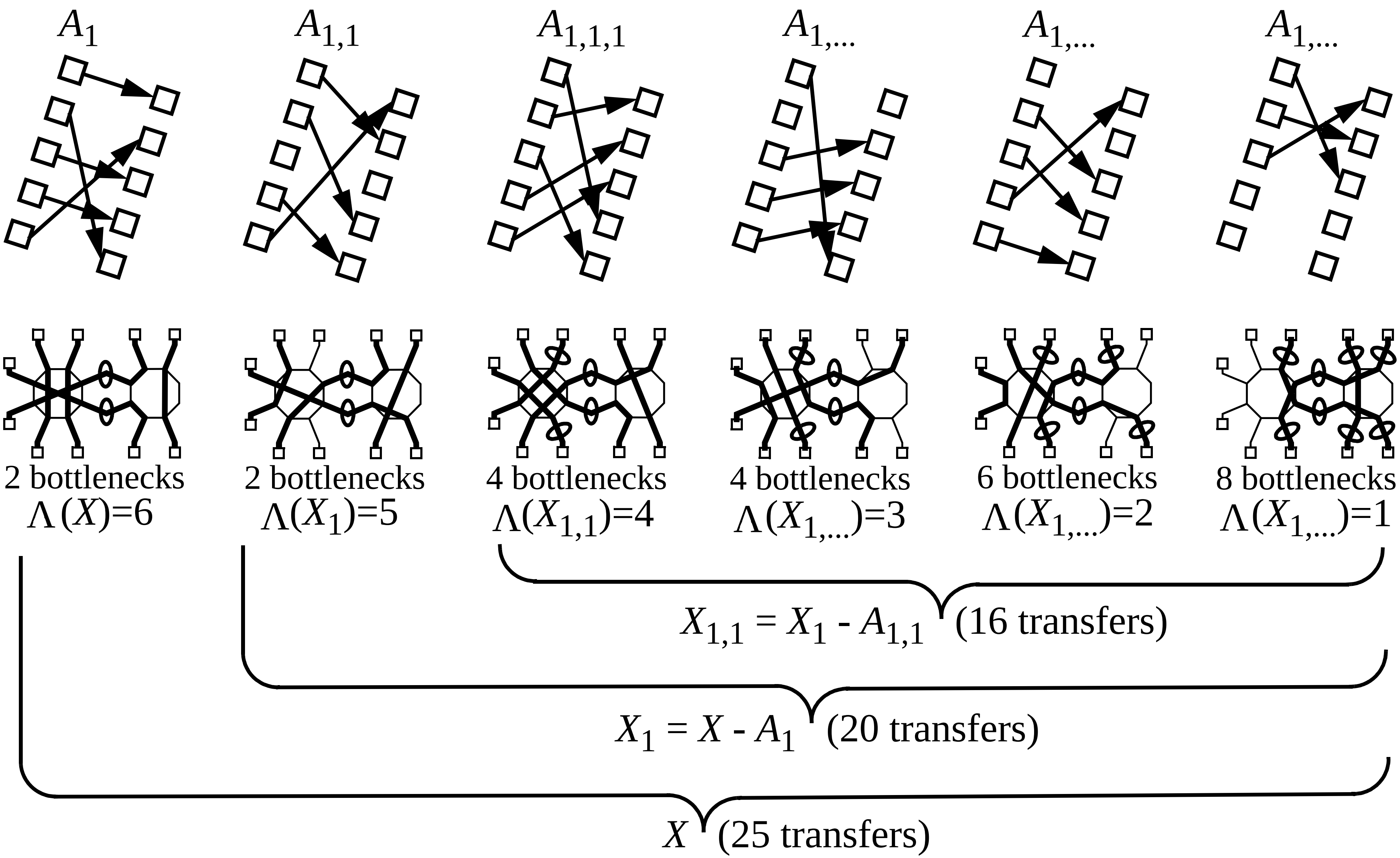
- $X \rightarrow \wp(X) = \{A_1, A_2, A_3 \ldots A_n\}$
  - $X_1 = X - A_1 \rightarrow \wp(X_1) = \{A_{1,1}, A_{1,2} \ldots\}$
    - $X_{1,1} = X_1 - A_{1,1}$
    - $X_{1,2} = X_1 - A_{1,2}$
    - ...
  - $X_2 = X - A_2 \rightarrow \wp(X_2) = \{A_{2,1}, A_{2,2} \ldots\}$
    - $X_{2,1} = X_2 - A_{2,1}$
    - $X_{2,2} = X_2 - A_{2,2}$

all teams of X

$$\wp(Y) = \{A \in \Im(X) | A \subset Y\}$$

possible steps to the next layer

Additional bottlenecks



Prediction of dead-ends and search optimization

- When a team of transfers is carried out - for the remaining traffic we have the same bottleneck links as before - with possibly new additionally emerged bottleneck links.

- Considering new bottleneck links (at every step of construction) in the choice of the further teams substantially reduces the search space.

- Team is a collection of simultaneous transmissions using all bottlenecks of the network. Teams are full if they congest with all other transmissions of the traffic.

- Limiting our choice with only full teams reduces the search space without affecting the solution space.

Liquid schedules construction

teams of the reduced traffic

$$\Im(Y) \subset \{A \in \Im(X) | A \subset Y\}$$

original traffic's teams formed from the reduced traffic

$$\underbrace{\Im^{full}}(Y) \subset \Im(Y)$$

full teams of the reduced traffic

$$Choice = \wp(Y) = \{A \in \Im(X) | A \subset Y\} \rightarrow \wp(Y) = \Im(Y)$$

$$\downarrow$$

$$Choice = \wp(Y) = \Im(Y)$$

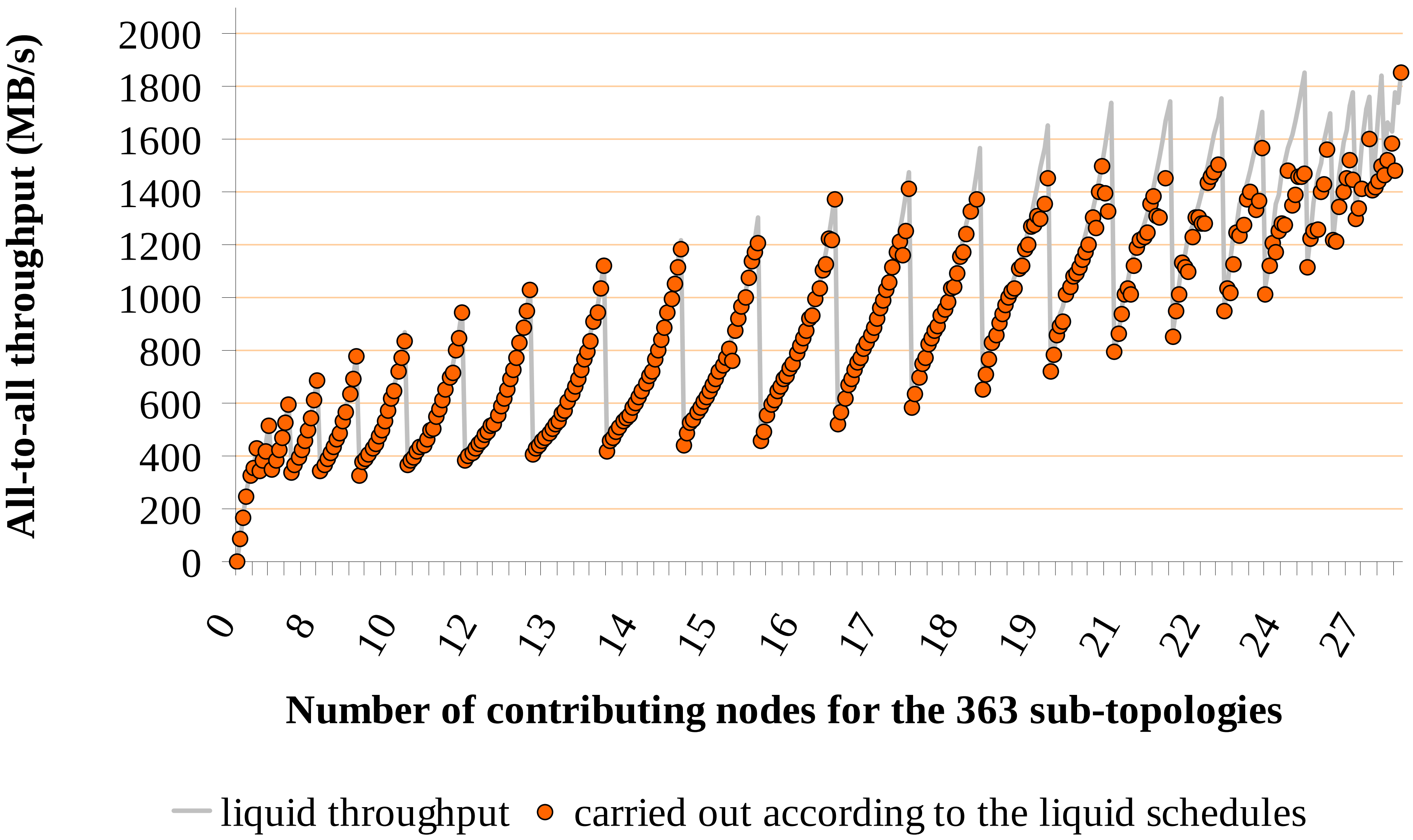
$$\downarrow$$

$$Choice = \wp(Y) = \Im^{full}(Y)$$

additionally decreasing the search space without affecting the solution space

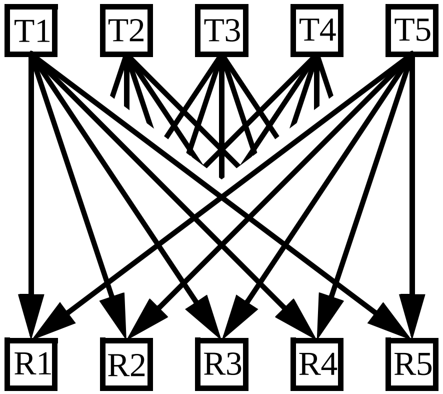
For more than 90% of the test-bed topologies construction of a global liquid schedule is completed in a fraction of a second (less than 0.1s).

Results

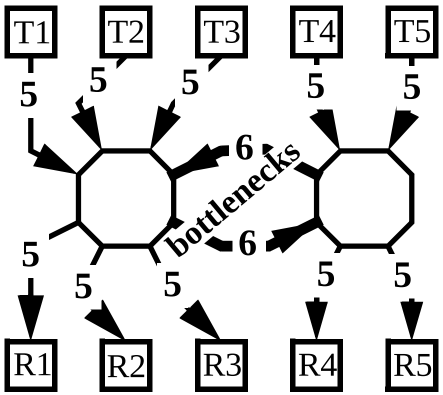


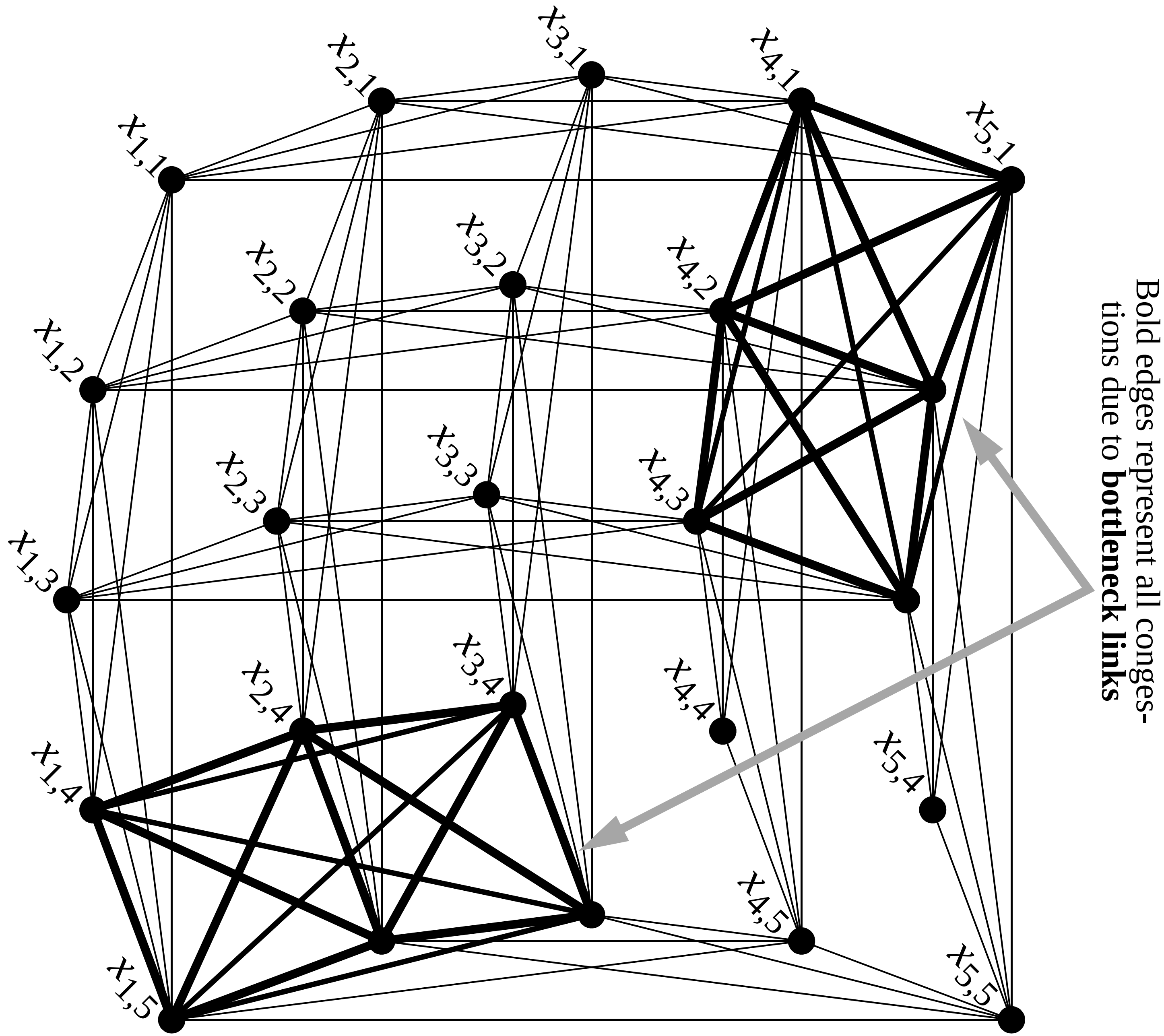
2000
1800
1600
1400
1200
1000
800
600
400
200
0
All-to-all throughput (MB/s)
0
8
10
12
13
14
15
16
17
18
19
21
22
24
27
Number of contributing nodes for the 363 sub-topologies
liquid throughput
carried out according to the liquid schedules

Congestion Graph



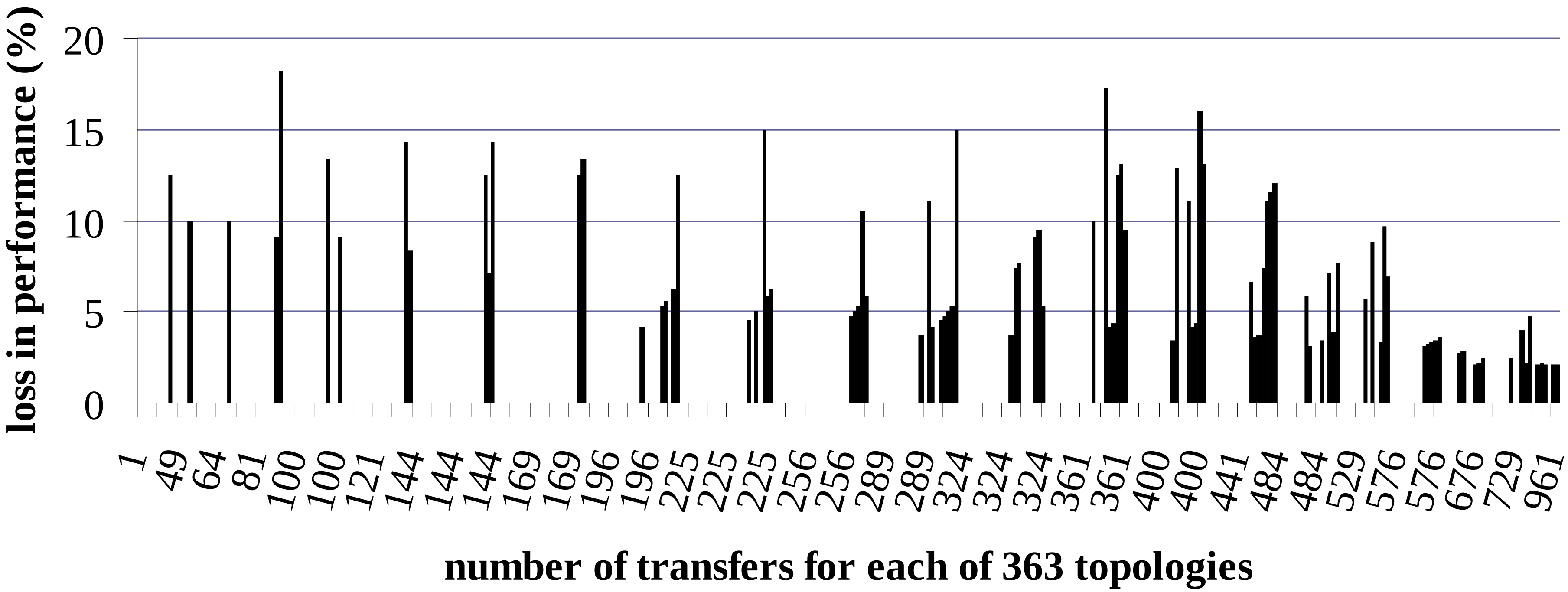
The 25 vertices of the graph represent the 25 transfers transfers. The edges represent congestion relations between transfers, i.e. each edge represents one or more communication links shared by two transfers.





-- ICON 2004, IEEE International Conference On Networks, November 16-19, 2004, Singapore, Hilton --
-- EFFICIENT LIQUID SCHEDULE SEARCH STRATEGIES FOR COLLECTIVE COMMUNICATIONS --

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.
- Liquid scheduling is applicable in wormhole, cut-through and WDM optical networks.

Thank You!

Contact: *Emin.Gabrielyan@epfl.ch*