6th SOS Workshop on Distributed Supercomputing: Data Intensive Computing
March 4-6, 2002, Badehotel Bristol, Leukerbad, Valais, Switzerland

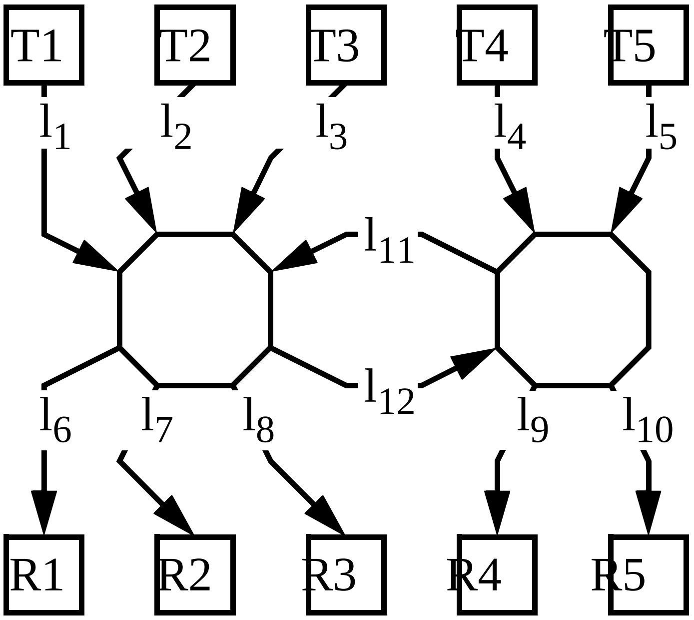
Network Topology-aware Traffic Scheduling

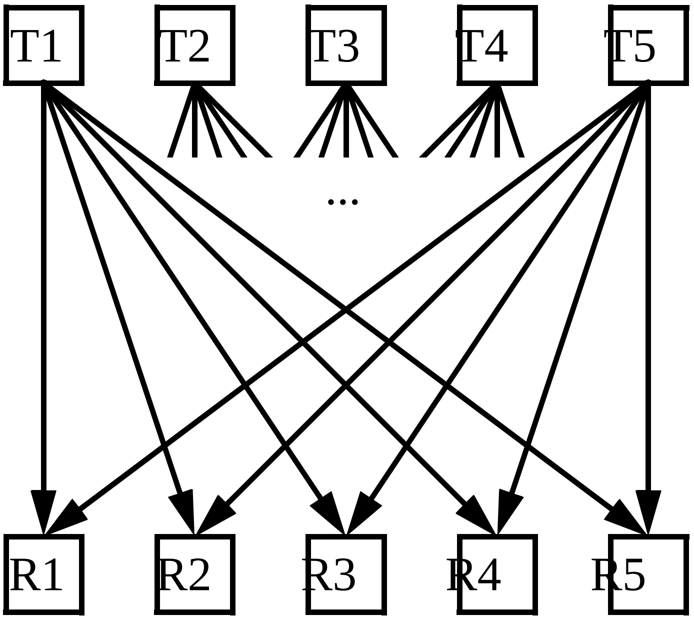
Emin Gabrielyan

École Polytechnique Fédérale de Lausanne, Switzerland

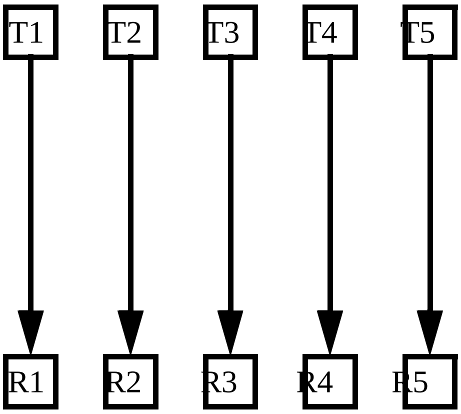
Emin.Gabrielyan@epfl.ch

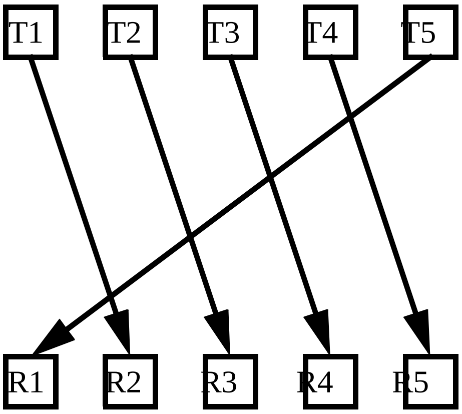
25-transfer data exchange

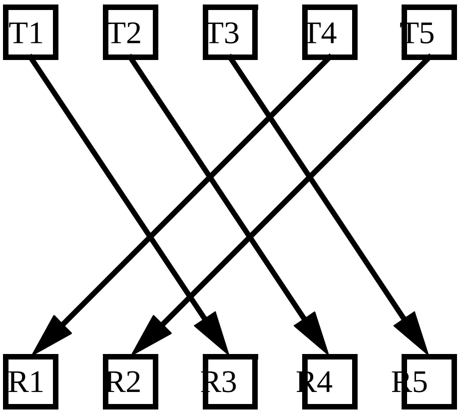


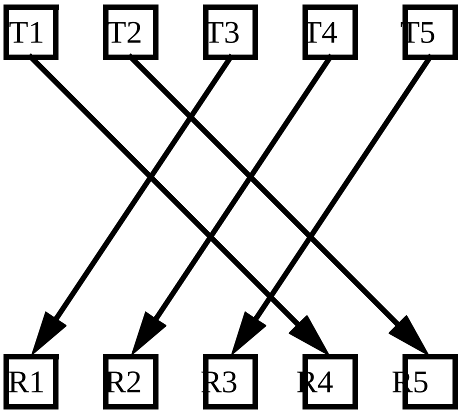


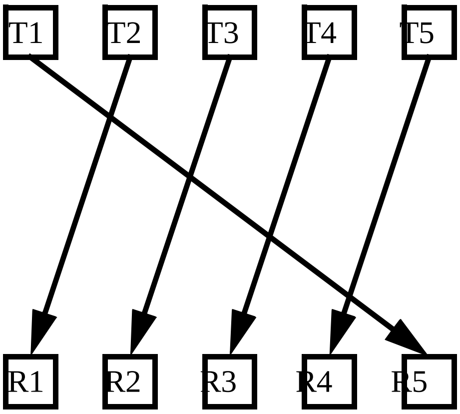
Round-robin schedule



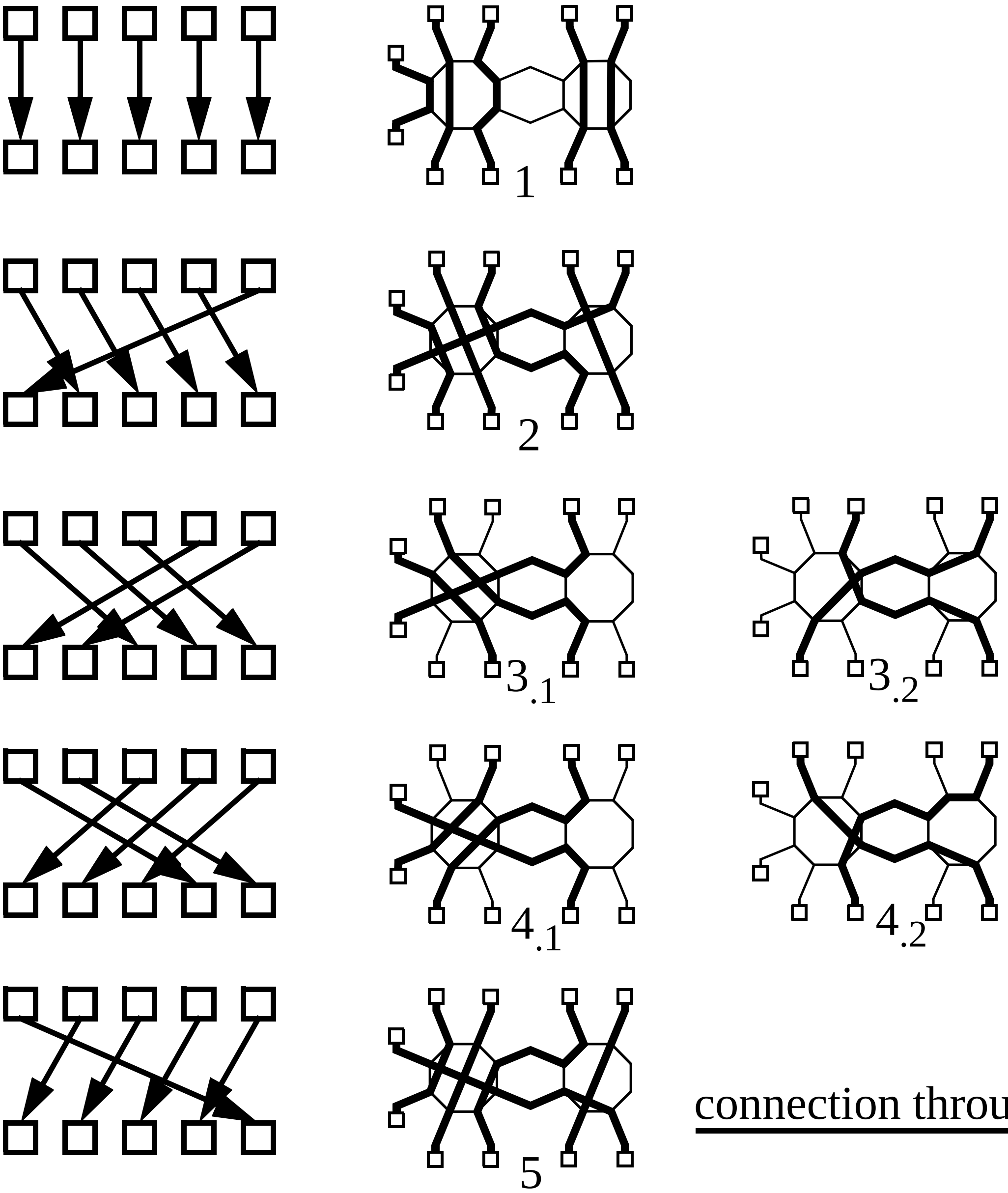








Round-robin Throughput



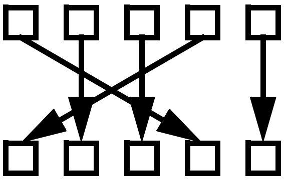
$$T_{roundrobin} = \overbrace{25/7}^{\text{mean number of connections per timeframe}} \cdot \overbrace{100MB/s}^{\text{connection throughput}} = 357MB/s$$

total throughput

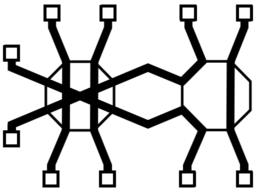
number of transfers

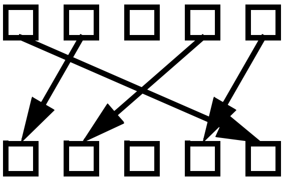
number of timeframes

Liquid Schedule

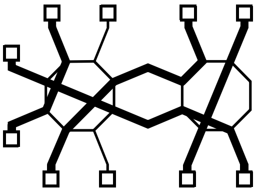


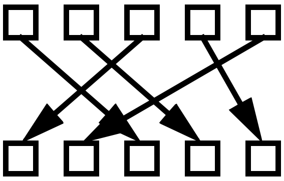
step 1



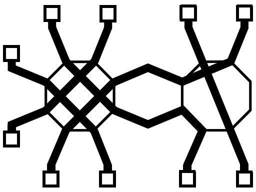


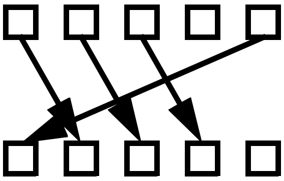
step 2



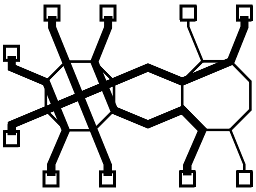


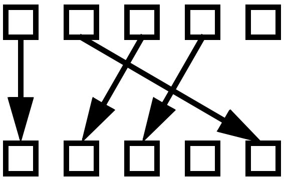
step 3



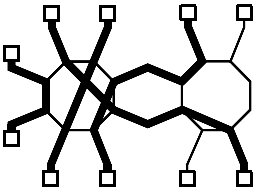


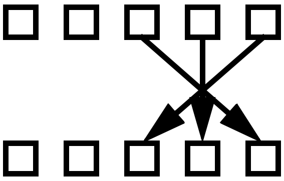
step 4



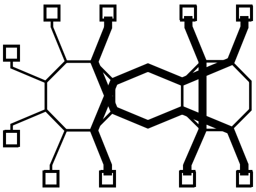


step 5





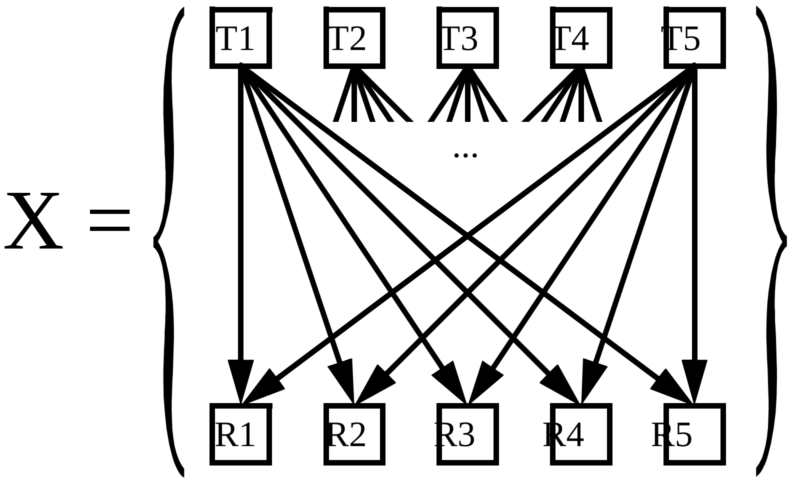
step 6



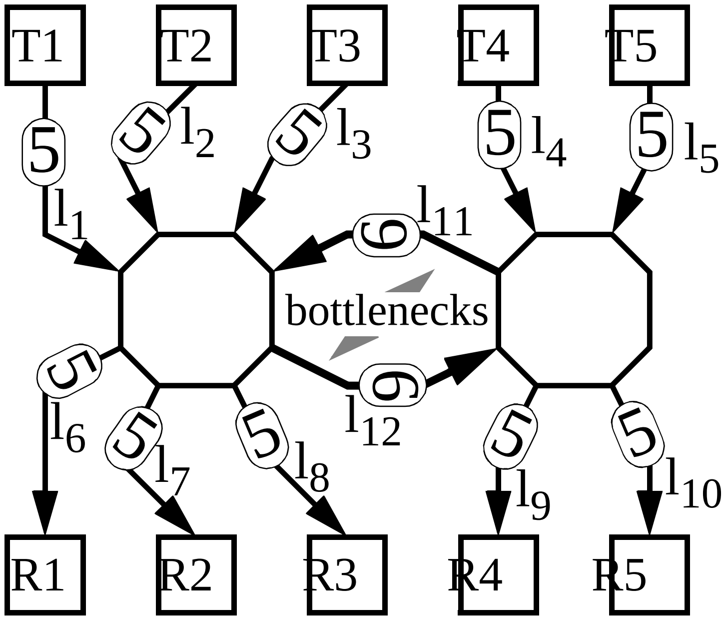
$$T_{liquid} = \underbrace{25/6} \cdot 100MB/s = 416MB/s$$

mean number of connections per step

Load of Links and Transfers



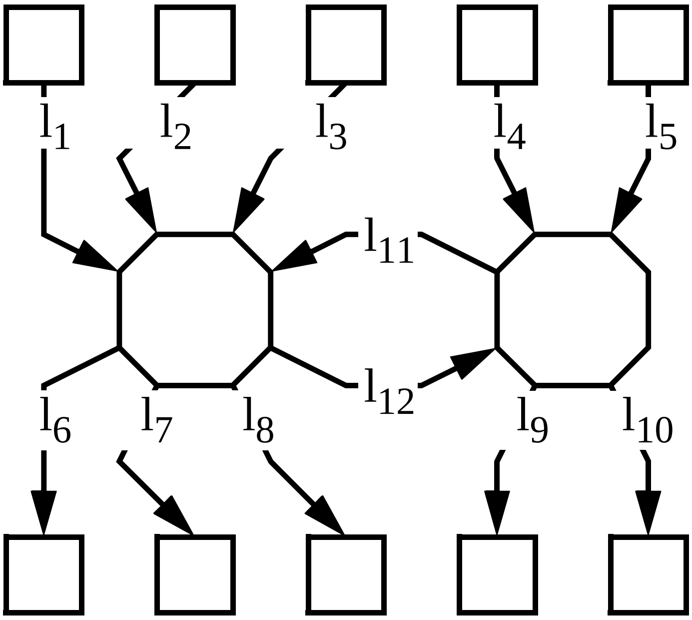
The 25 transfer traffic



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

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

Duration of the Traffic



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

$$\lambda(l_1, X) = 5, \ \lambda(l_2, X) = 5, \ldots$$

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

$$\Lambda(X) = 6$$

Liquid Throughput

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

the throughput of a single link

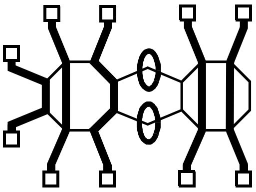
total number of transfers

$$T_{liquid} = \frac{\#(X)}{\Lambda(X)} \cdot T_{link} =$$

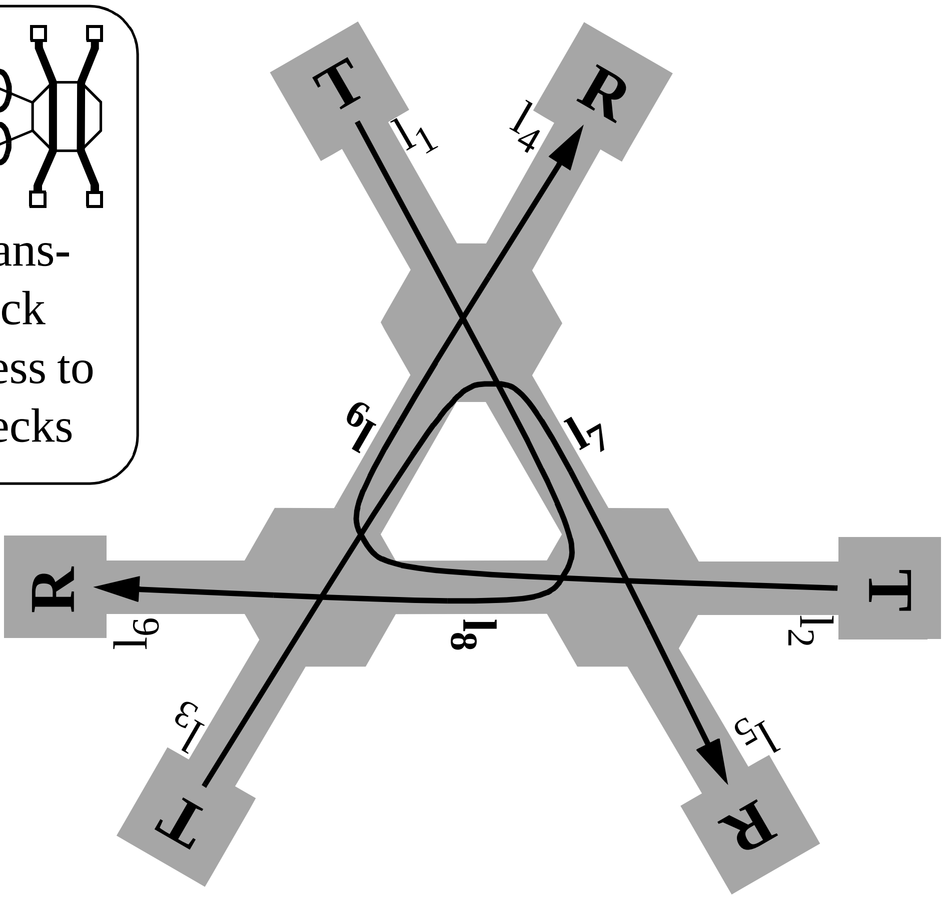
the duration of the traffic (the load of its bottlenecks)

$$= \frac{25}{6} \cdot 100 MB/s = 417 MB/s$$

No liquid schedule



the 5 transfers block the access to bottlenecks

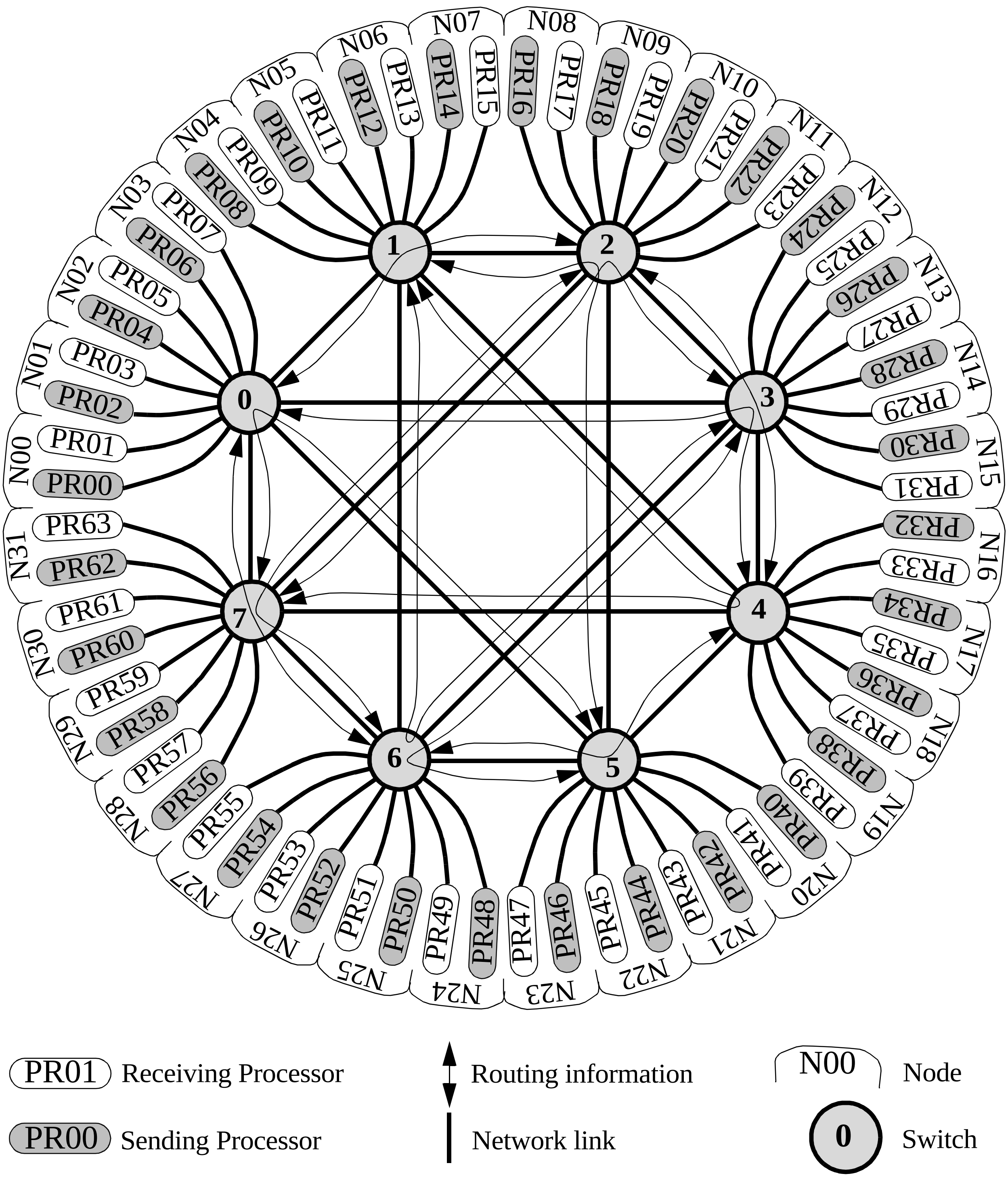


$$X = \left\{ \begin{array}{c} \{l_1, \mathbf{l_7}, \mathbf{l_8}, l_6\}, \\ \{l_2, \mathbf{l_8}, \mathbf{l_9}, l_4\}, \\ \{l_3, \mathbf{l_9}, \mathbf{l_7}, l_5\} \end{array} \right\} \quad \begin{array}{l} \#(X) = 3 \\ \Lambda(X) = 2 \end{array}$$

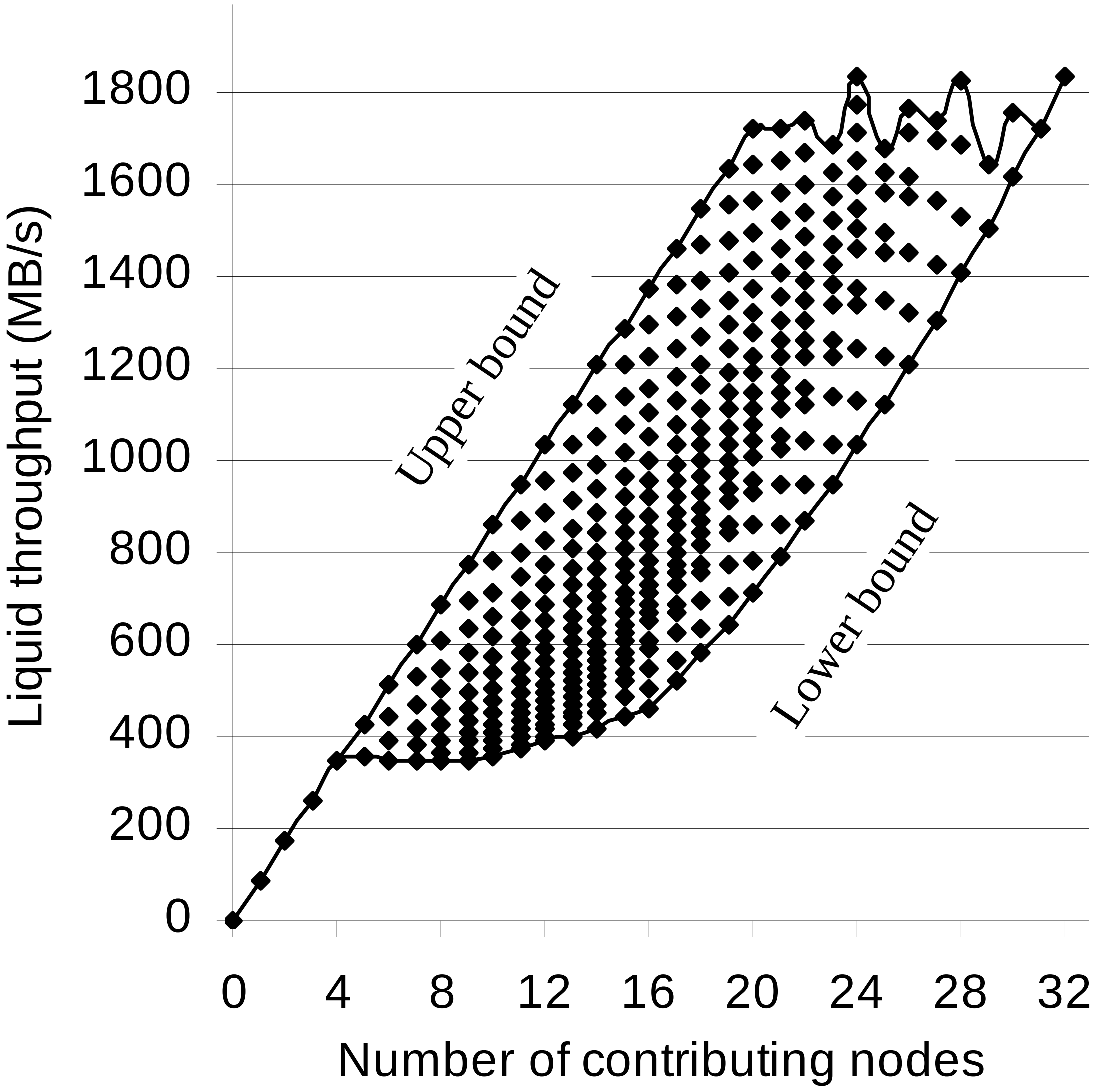
$$T_{liquid} = \frac{\#(X)}{\Lambda(X)} \cdot T_{link} =$$

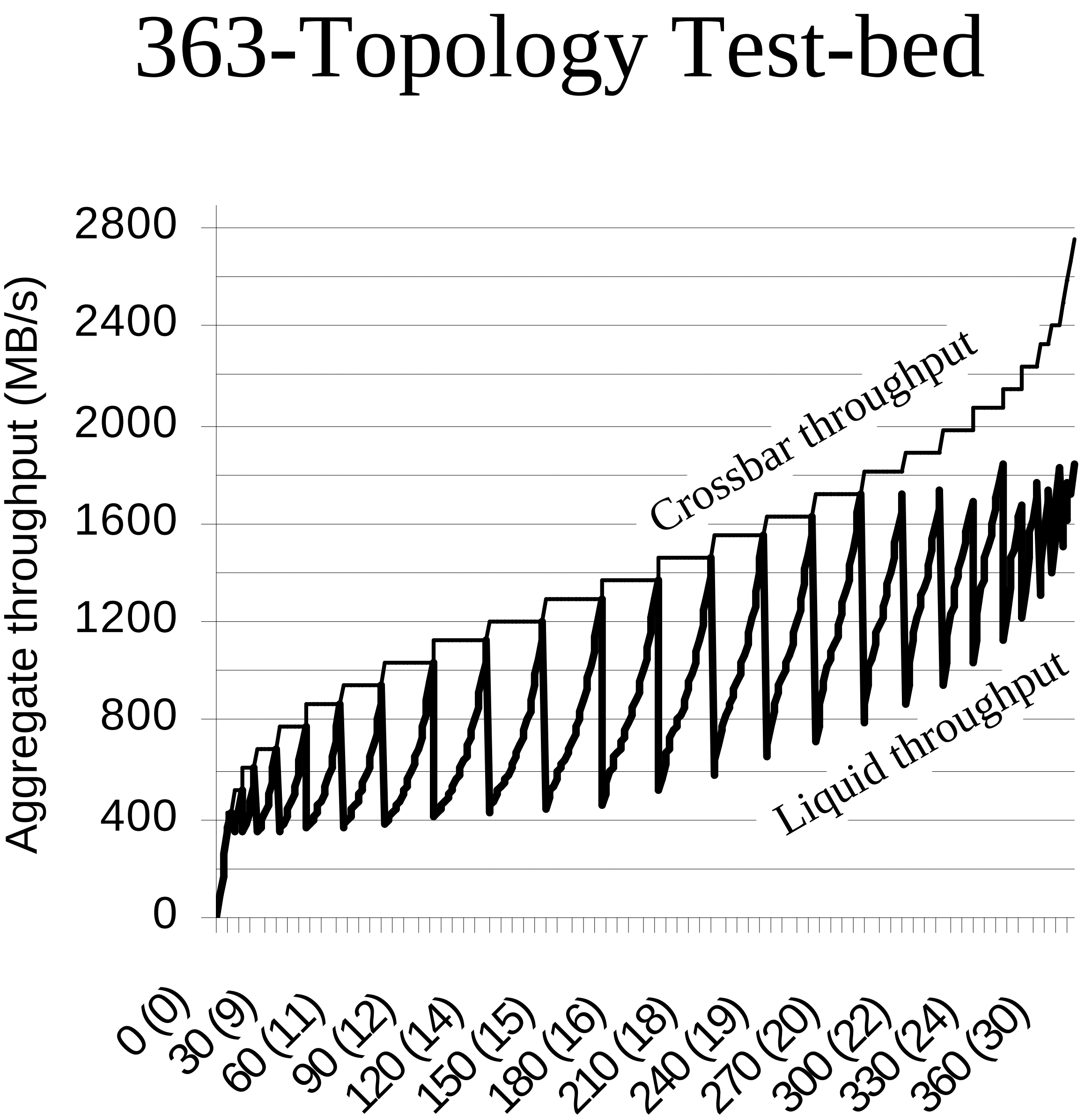
$$= 3 / 2 \cdot 100 MB / s = 150 MB / s$$

Swiss-T1 Cluster

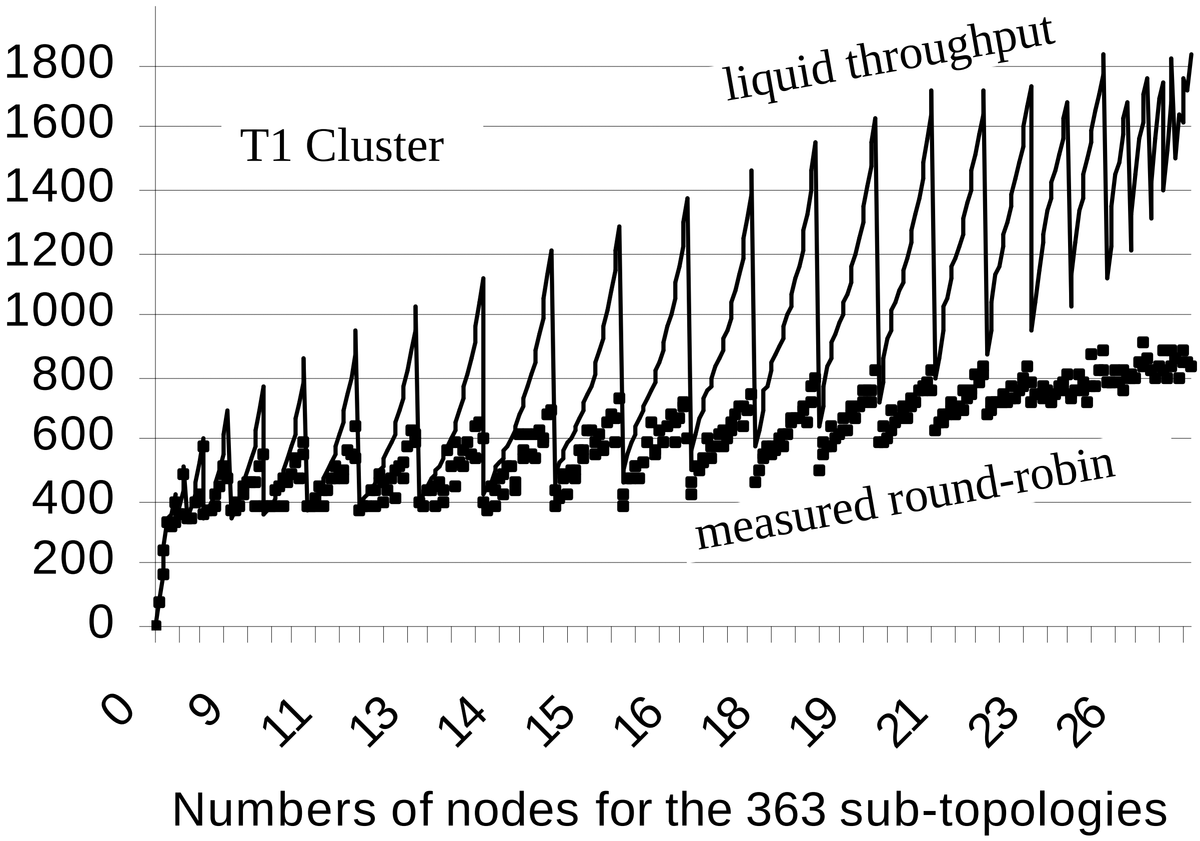


363 Test Traffics




363-Topology Test-bed
Aggregate throughput (MB/s)
2800
2400
2000
1600
1200
800
400
0
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)

Round-robin throughput



Team: set of non-congesting transfers using all bottlenecks

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

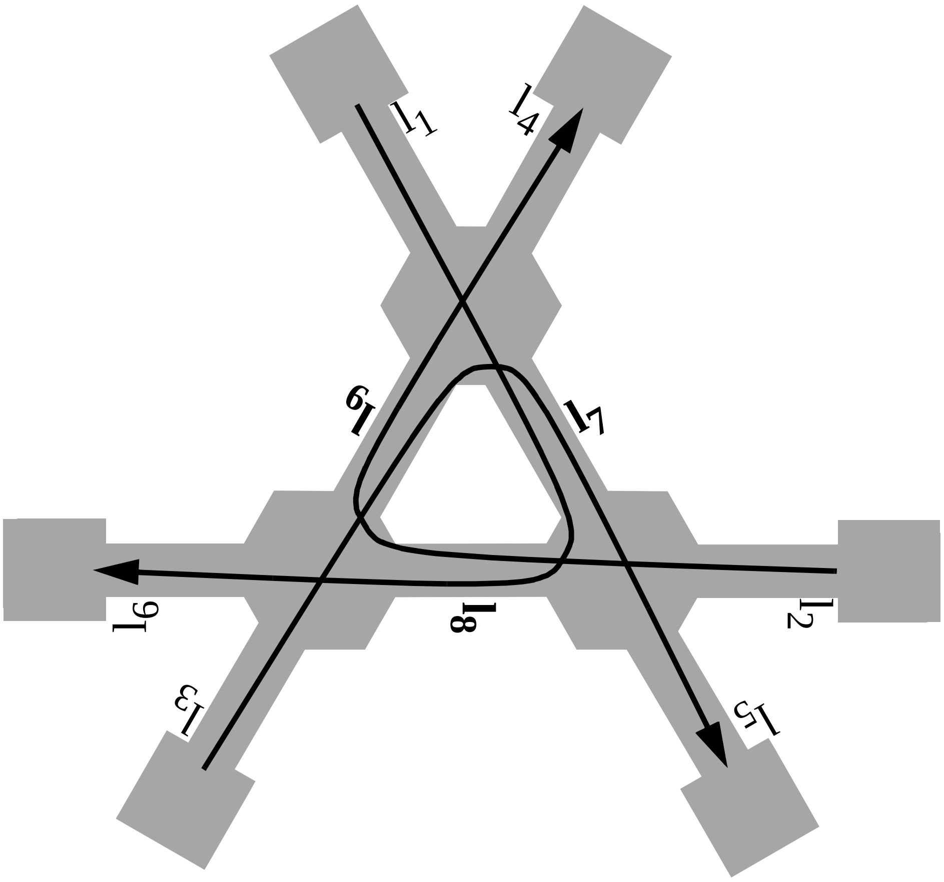
$$\alpha = \left\{ \begin{array}{c} \left\{ \begin{array}{c} \{l_1, \mathbf{l_{12}}, l_9\}, \\ \{l_2, l_7\}, \\ \{l_3, l_8\}, \\ \{l_4, \mathbf{l_{11}}, l_6\}, \\ \{l_5, l_{10}\} \end{array} \right\}, \left\{ \begin{array}{c} \{l_1, \mathbf{l_{12}}, l_{10}\}, \\ \{l_2, l_6\}, \\ \{l_4, \mathbf{l_{11}},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

schedule α is liquid $\Leftrightarrow$

$$\Leftrightarrow \overset{\text{number of steps}}{\#(\alpha)} = \overset{\text{load of the bottlenecks}}{\Lambda(X)} \Leftrightarrow$$

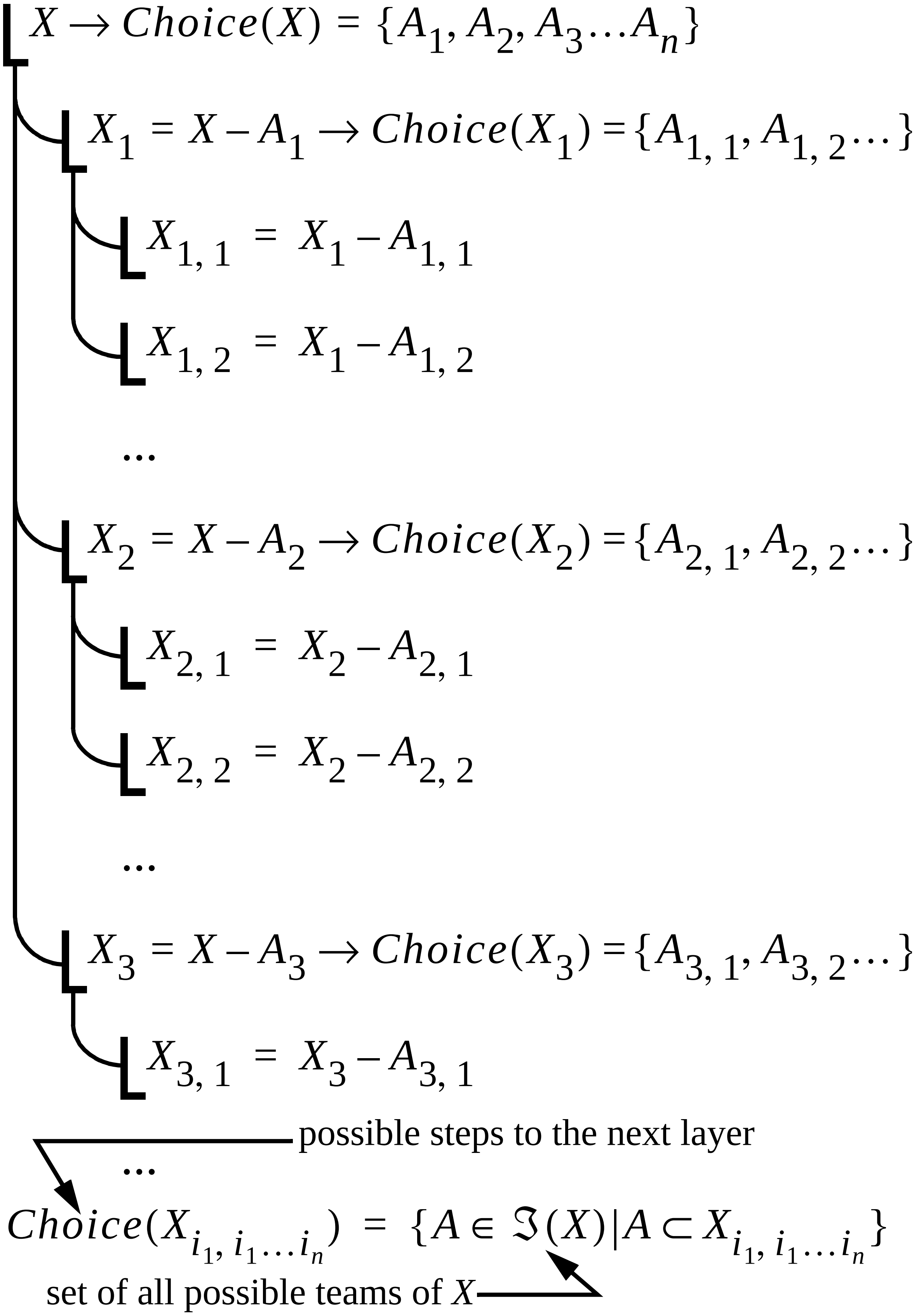
$$\Leftrightarrow \forall (A \in \alpha)\ A \text{ is a team of } X$$

Traffic without a team



$$X = \left\{ \begin{array}{c} \{\mathrm{l}_1, \mathbf{l_7}, \mathbf{l_8}, \mathrm{l}_6\}, \\ \{\mathrm{l}_2, \mathbf{l_8}, \mathbf{l_9}, \mathrm{l}_4\}, \\ \{\mathrm{l}_3, \mathbf{l_9}, \mathbf{l_7}, \mathrm{l}_5\} \end{array} \right\}$$

Liquid schedule search tree



$$X \rightarrow Choice(X) = \{A_1, A_2, A_3 \ldots A_n\}$$

$$X_1 = X - A_1 \rightarrow Choice(X_1) = \{A_{1,\,1}, A_{1,\,2} \cdots\}$$

$$X_{1,\,1} = X_1 - A_{1,\,1}$$

$$X_{1,\,2} = X_1 - A_{1,\,2}$$

...

$$X_2 = X - A_2 \rightarrow Choice(X_2) = \{A_{2,\,1}, A_{2,\,2} \cdots\}$$

$$X_{2,\,1} = X_2 - A_{2,\,1}$$

$$X_{2,\,2} = X_2 - A_{2,\,2}$$

...

$$X_3 = X - A_3 \rightarrow Choice(X_3) = \{A_{3,\,1}, A_{3,\,2} \cdots\}$$

$$X_{3,\,1} = X_3 - A_{3,\,1}$$

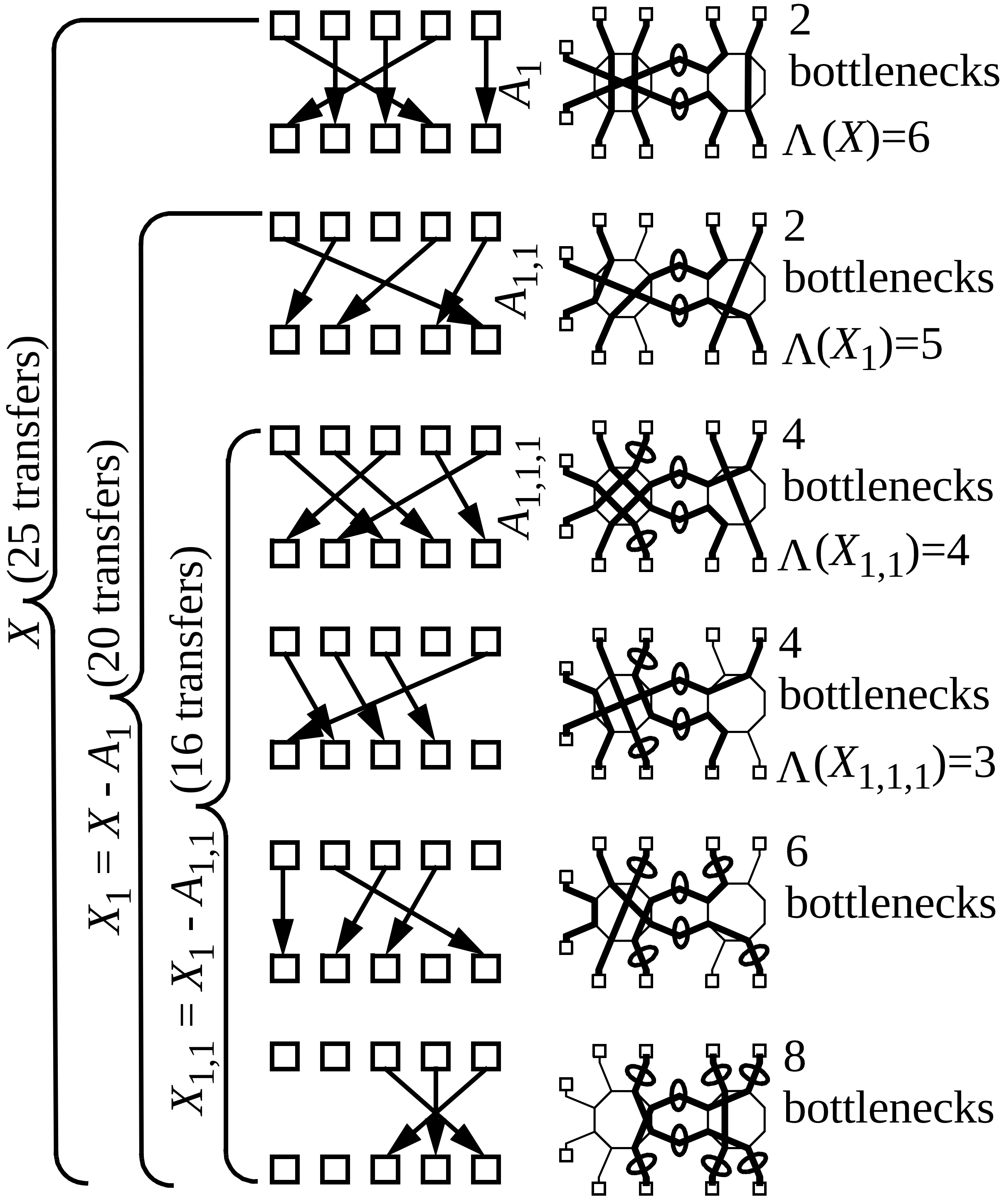
possible steps to the next layer

...

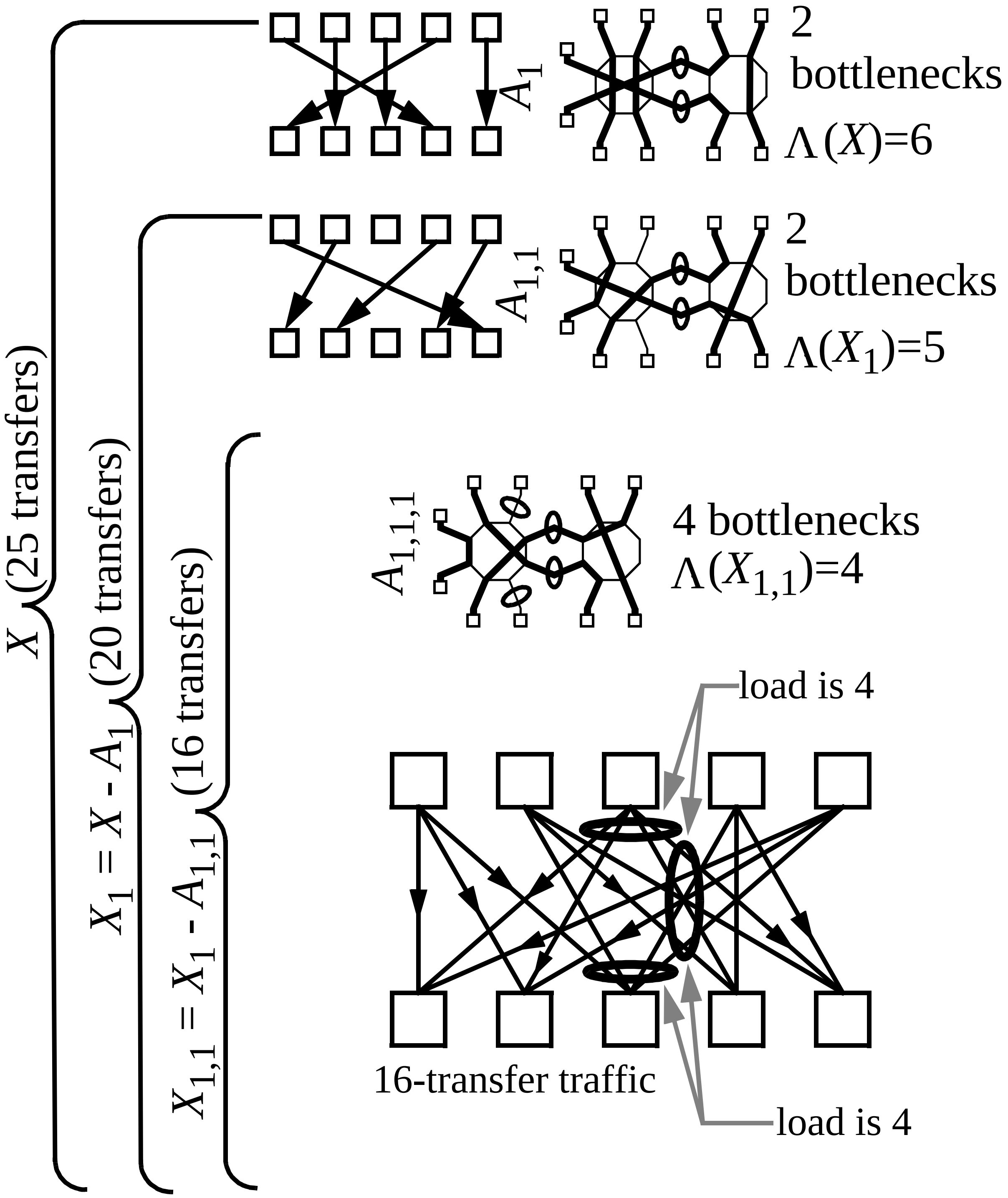
$$Choice(X_{i_1,\, i_1 \ldots i_n}) = \{A \in \Im(X) | A \subset X_{i_1,\, i_1 \ldots i_n}\}$$

set of all possible teams of X

Additional bottlenecks



Prediction of Dead-ends



Liquid schedule search optimization

Y: reduced traffic

$$\underbrace{\Im(Y)}_{\text{teams of the reduced traffic}} \subset \underbrace{\{A \in \Im(X) | A \subset Y\}}_{\text{original traffic's teams formed from the reduced traffic}}$$

original traffic's teams formed from the reduced traffic

teams of the reduced traffic

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

$$\begin{tikzcd}
Choice(Y) = \{A \in \Im(X) | A \subset Y\} \arrow[d] \\
Choice(Y) = \Im(Y)
\end{tikzcd}$$

decrease of the search space without affecting the solution space

Liquid schedules construction

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

full teams of the reduced traffic

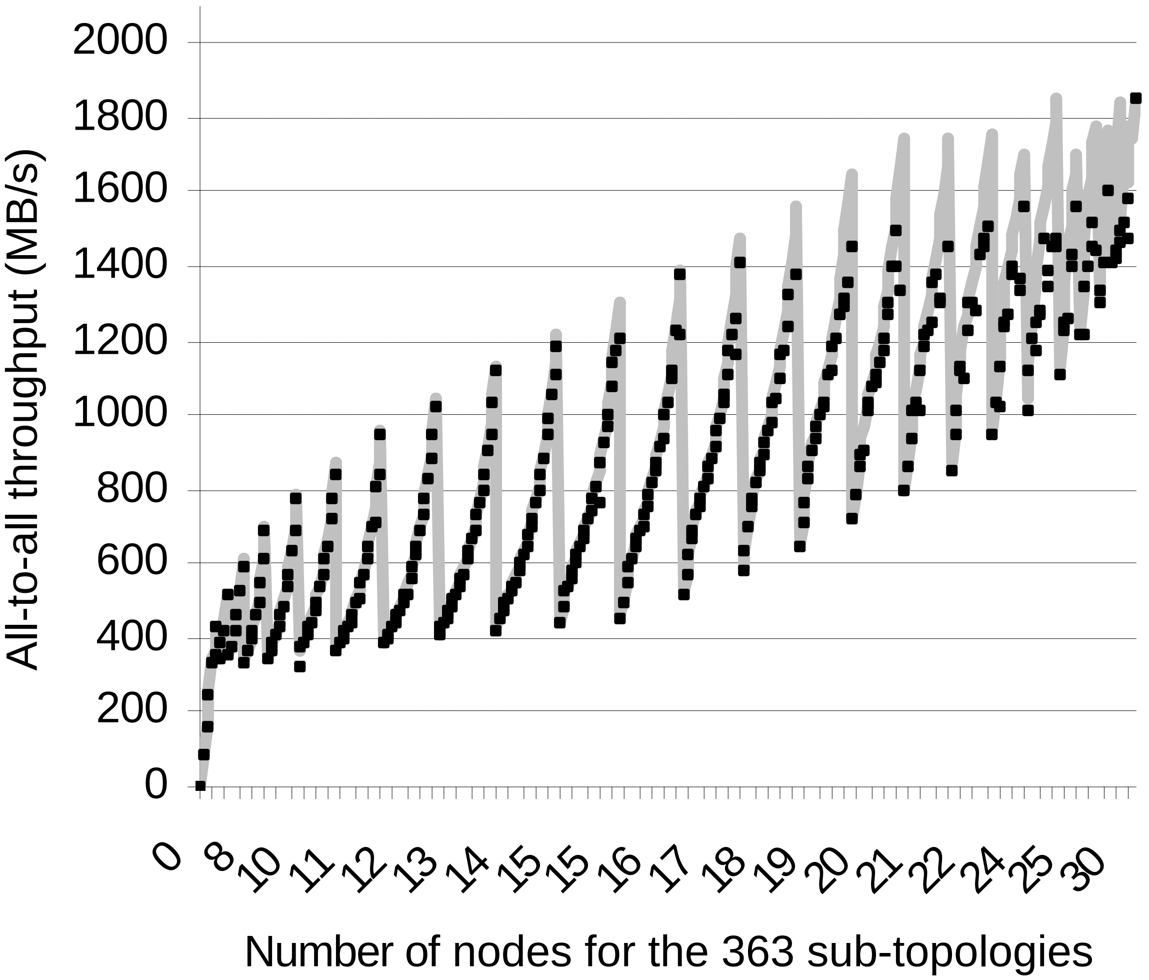
$$\begin{tikzcd}
Choice(Y) = \Im(Y) \arrow[d] \\
Choice(Y) = \Im^{full}(Y)
\end{tikzcd}$$

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



All-to-all throughput (MB/s)
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
8
10
11
12
13
14
15
15
16
17
18
19
20
21
22
24
25
30
Number of nodes for the 363 sub-topologies

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*