Odessa State Environmental University, Ukraine
Sigma-T NVO, LTD, Dnipro, Ukraine

Switching vs Routing within Multidimensional Torus Interconnect

Dmitry A. Zaitsev
Serhii I. Tymchenko,
Nataliia Z. Shtefan

http://daze.ho.ua

Torus interconnect

- Three-dimensional torus network: IBM Blue Gene/L and Blue Gene/P, and the Cray XT3
- Five-dimensional torus network: IBM Blue Gene/Q
- Six-dimensional torus network: Fujitsu K computer, PRIMEHPC FX10 – three-dimensional torus 3D mesh interconnect Tofu
- Fugaku, ~0.5 exaflops – TOFU interconnect D

Fugaku (Fujitsu, RIKEN)

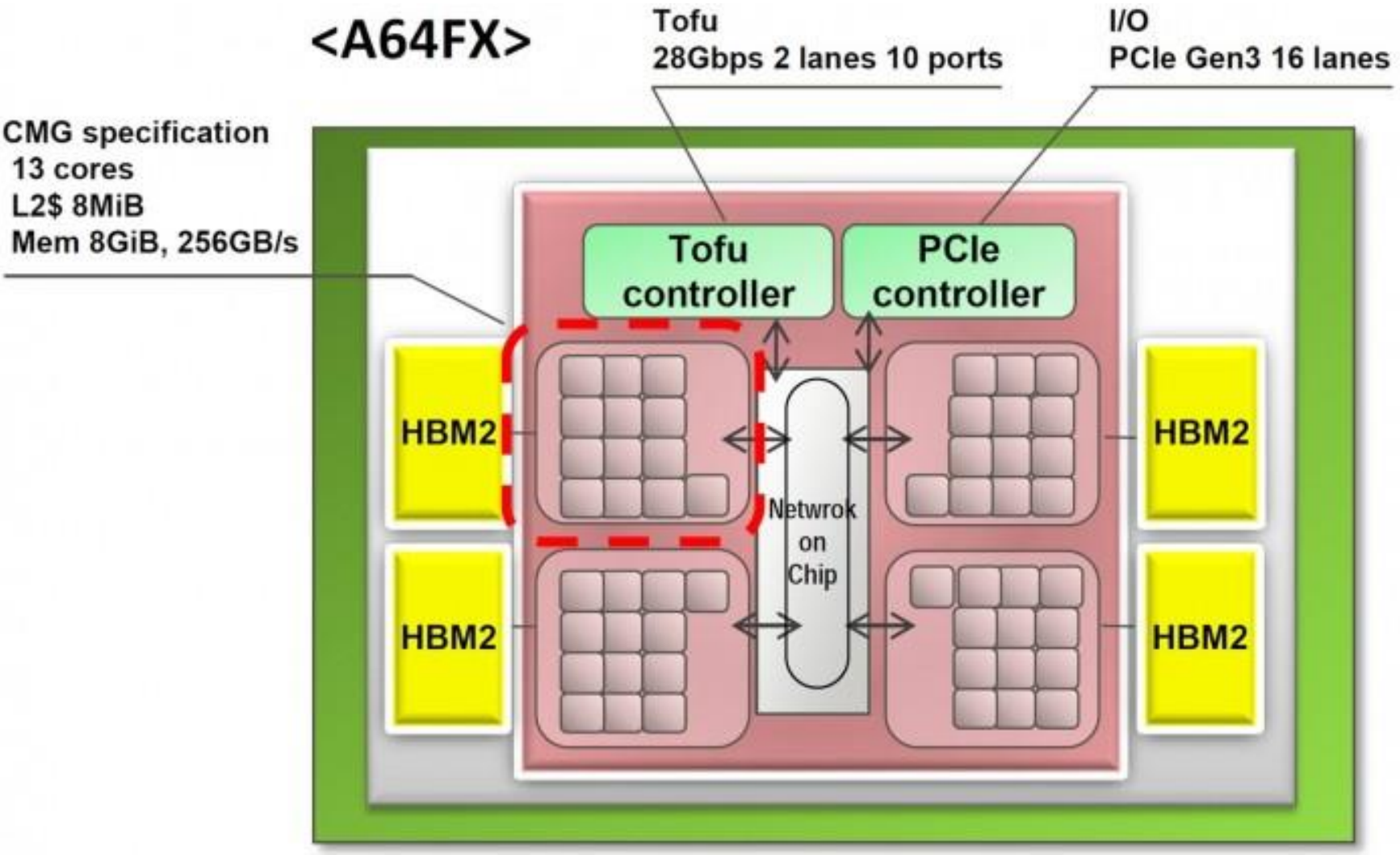
http://top500.org

A64FX

A64FX® Microarchitecture Manual

http://github.com/fujitsu/A64FX

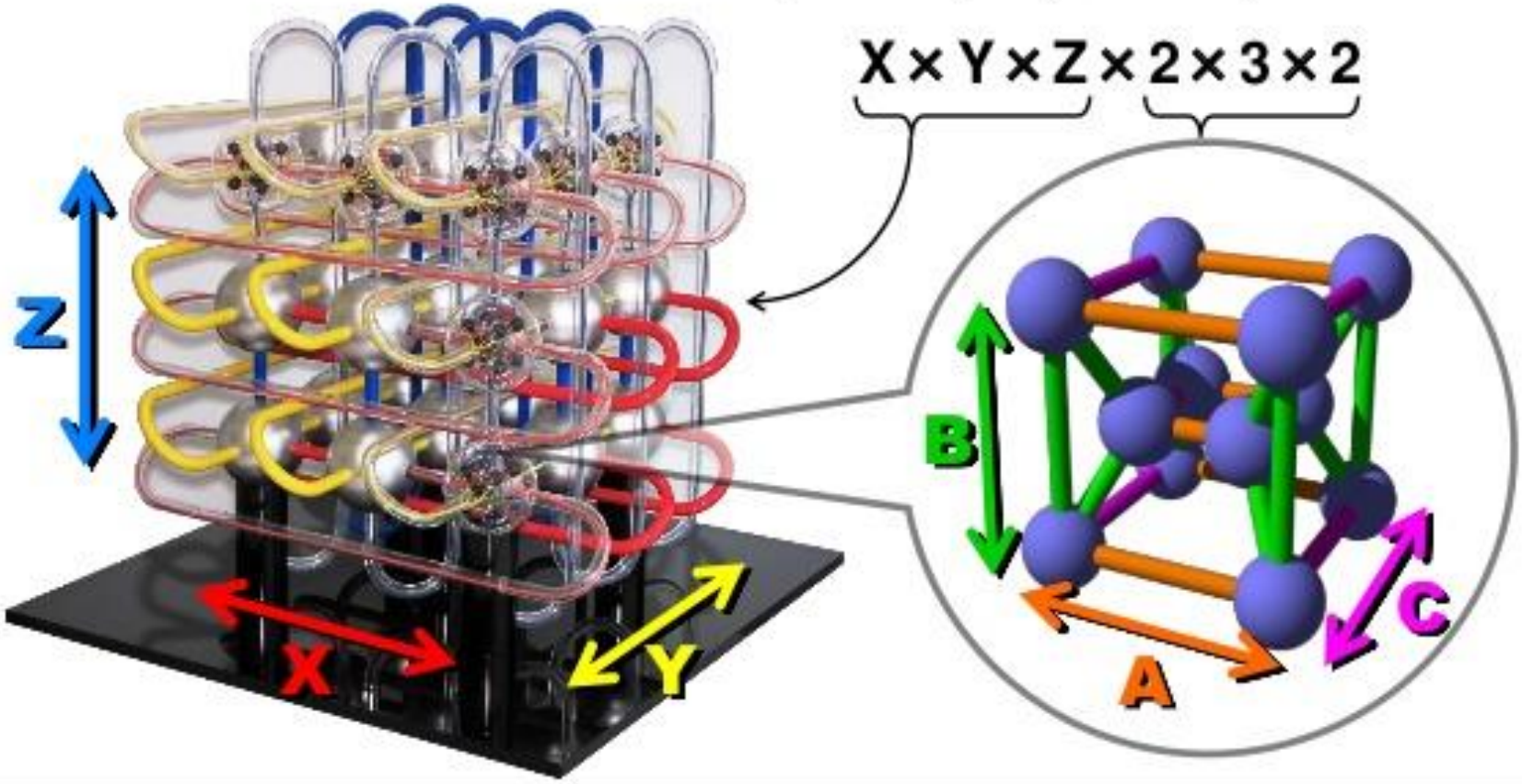
A64FX block diagram



6D Mesh/Torus Network

FUJITSU

- Six coordinate axes: X, Y, Z, A, B, C
  - X, Y, Z: the size varies according to the system configuration
  - A, B, C: the size is fixed to 2 × 3 × 2
- Tofu stands for "torus fusion": (X, Y, Z) × (A, B, C)



Related Work

- W.J. Dally, C.L. Seitz, “The torus routing chip,” *Distrib. Comput.*, no. 1, 1986, pp. 187–196.
- A. Singh, W.J. Dally, B. Towles, and A.K. Gupta, "Globally Adaptive Load-Balanced Routing on Tori," *IEEE Computer Architecture Letters*, vol. 3, no. 1, pp. 2-2, Jan. 2004.
- P. Ren, M. A. Kinsy and N. Zheng, "Fault-Aware Load-Balancing Routing for 2D-Mesh and Torus On-Chip Network Topologies," *IEEE Transactions on Computers*, vol. 65, no. 3, pp. 873-887, 1 March 2016.
- P. Xie, H. Gu, K. Wang, X. Yu, S. Ma, "Mesh-of-Torus: a new topology for server-centric data center networks", *The Journal of Supercomputing*, vol. 75, pp. 255, 2019.

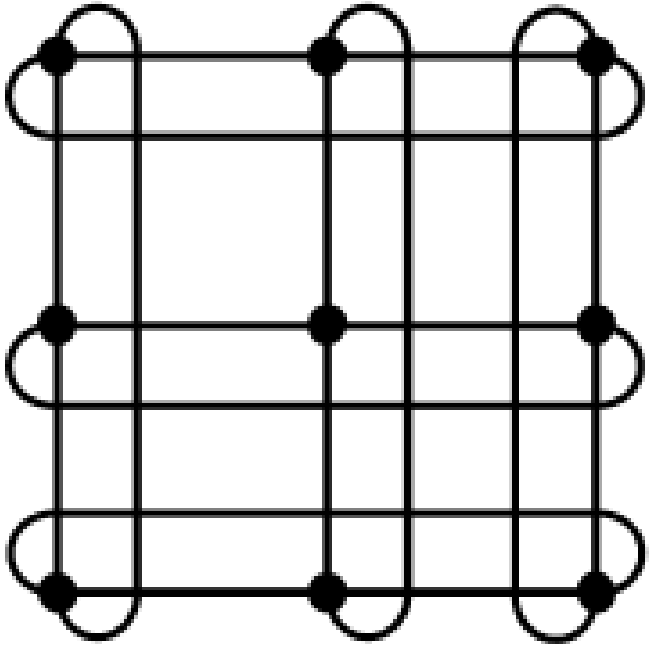
Torus Interconnect in Trend

- Network-on-chip
- Interconnect of supercomputers and cluster
- Prospects for networks of service providers
- Prospects for campus and metropolitan networks
- Prospect for connecting Internet autonomous systems

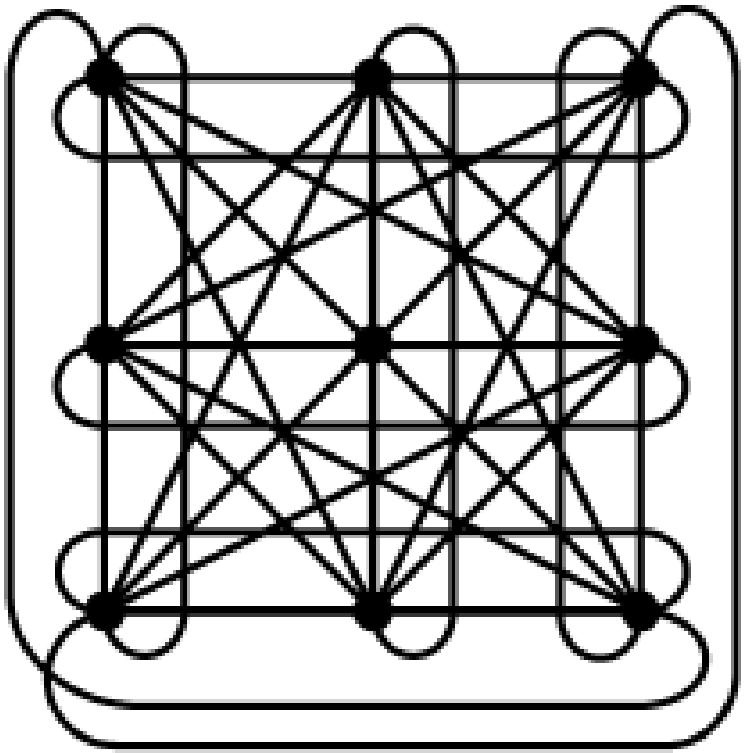
Advantages of Torus Interconnect

- Regular graph structure
- Short distance between nodes
- Many alternative (shortest) routes
- Possibility for packet delivery based on predefined switching rules
- Possibility of load balancing based on (random) alternative choice

Neighborhoods in 2D torus



Von Neumann



Moore

Neighborhood of torus interconnect

- Von Neumann neighborhood
- Mixed mesh and von Neumann interconnect
- Moore neighborhood is too dense in multidimensional space
- Generalized neighborhood is flexible, density is adjusted using a parameter
- Cross-By-Pass-Torus can by implemented as a generalized neighborhood with radius > 1

Basic Notions

- Node address:

$$\mathrm{i} = (i_0, i_1, \ldots, i_{d-1}),$$

- Node ports:

$$p = (m, r)$$

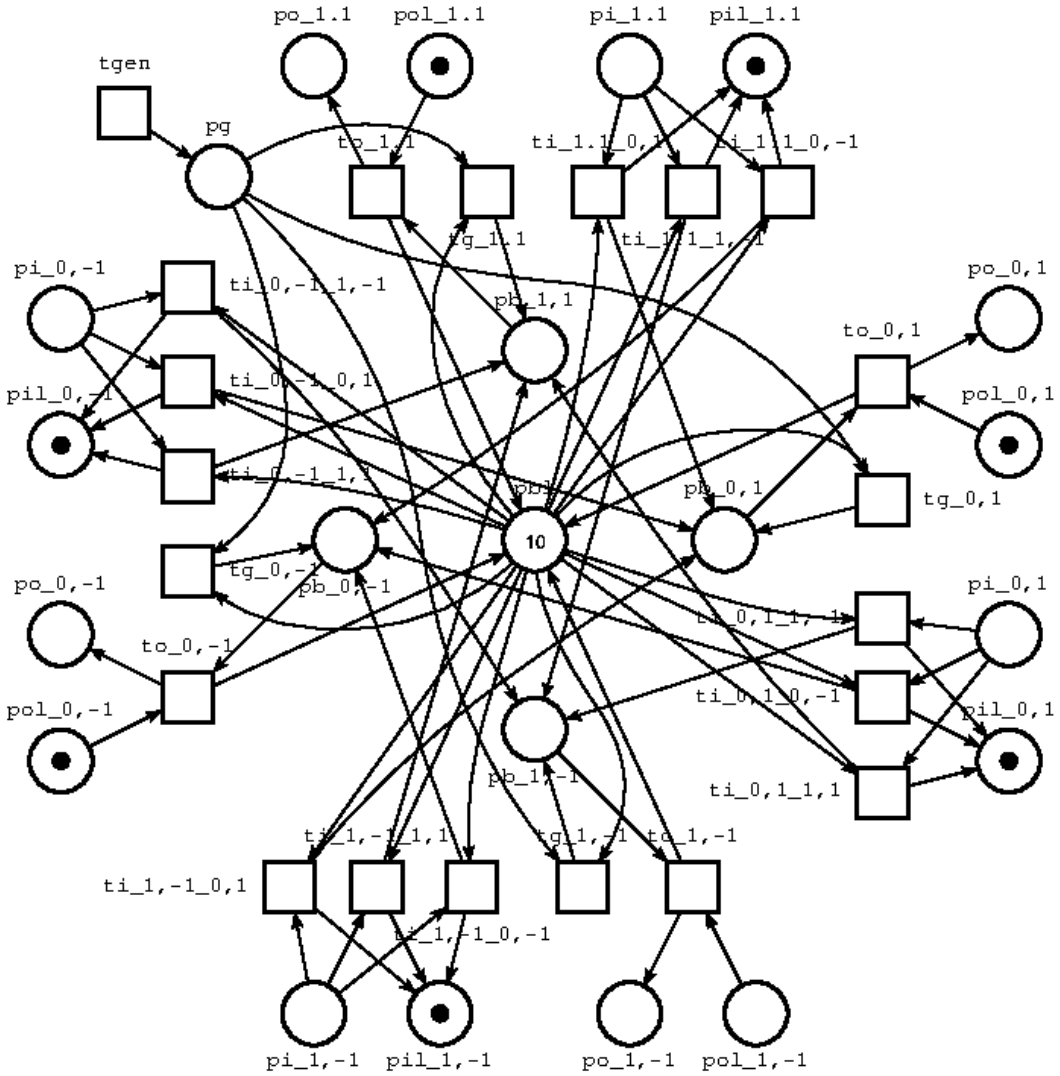
- Neighbor node:

$$i_j \pm 1 \; mod \; k$$

- Distance between nodes – Manhattan (taxicab) norm

$$D(\mathrm{i}, \mathrm{i}') = \sum_{j=0}^{d-1} \min(\left|i'_j - i_j\right|, k - \left|i'_j - i_j\right|)$$

Rules based on the current and destination node addresses only



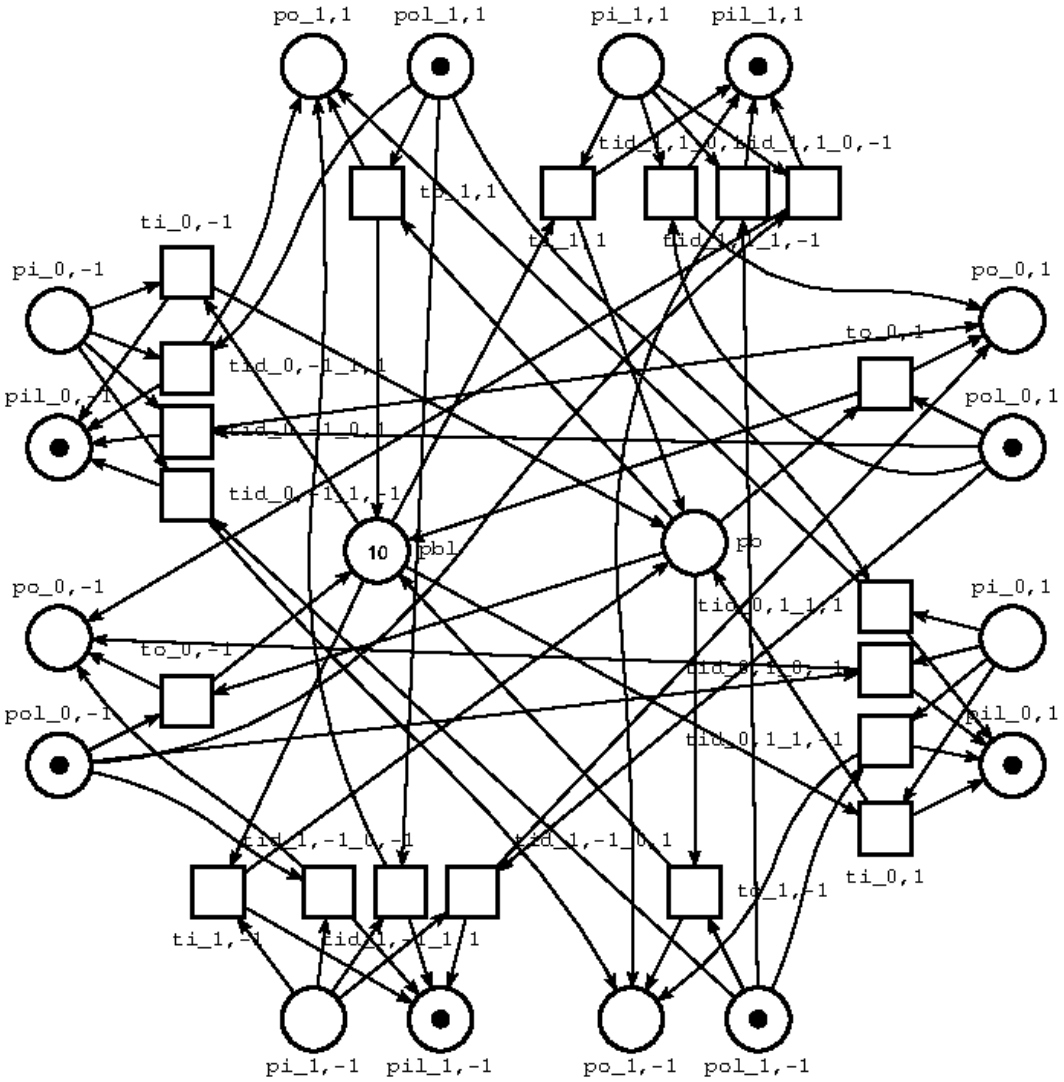
Local switching rule

- *Statement 1.* Repeated forwarding a packet from a current node to the next node decreasing the distance (by unit), starting from the source node, provides finally the packet delivery to the destination node using one of the shortest paths.

Modifications of local switching rule

- A) the first coordinate with nonzero difference (that corresponds to the deterministic dimension-order routing);
- B) random choice of coordinate among coordinates with nonzero difference;
- C) a random coordinate among coordinates with nonzero difference with the coordinate choice proportional to the coordinate differences.

Rules taking into consideration the current node state



Modifications of local switching rule

- D) the first coordinate with nonzero difference for *a free port*;
- E) random choice of coordinate among coordinates of *free ports* with nonzero difference;
- F) a random coordinate among coordinates of *free ports* with nonzero difference with the coordinate choice proportional to the coordinate differences.

Example of Routes in (6,8)-torus

- Destination address: $\mathbf{i}' = (0,6,5,7,2,4)$
- Current node address: $\mathbf{i} = (7,3,5,1,2,2)$
- Port availability vector:
  $$\mathbf{a} = \begin{pmatrix} 1 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix}$$
- Address difference:
  $\Delta = \mathbf{i}' - \mathbf{i} = (1,3,0,-2,0,2)$
- Distance: $D(\mathbf{i}, \mathbf{i}') = 1 + 3 + 2 + 2 = 8$

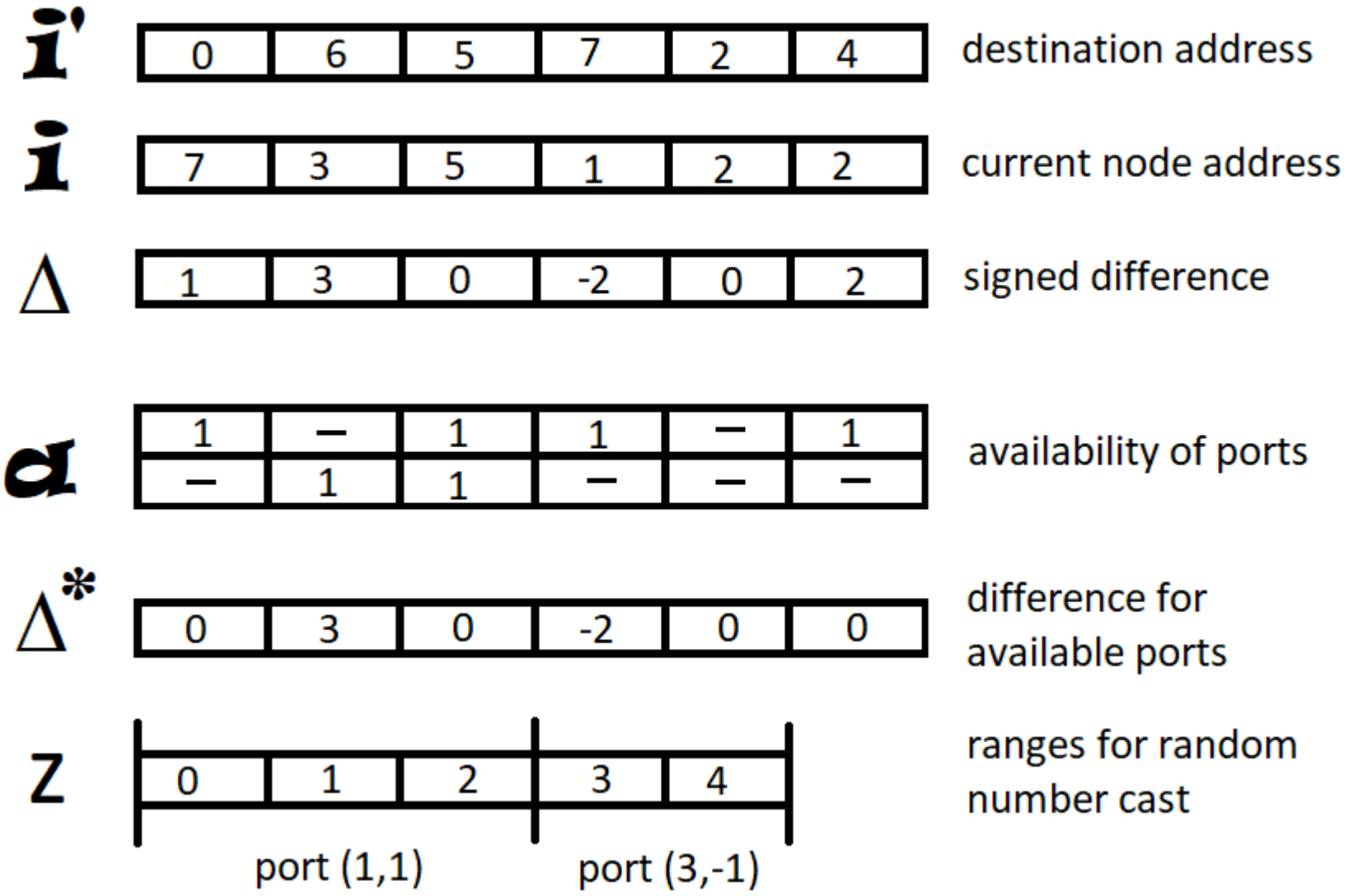
Direction Preference for $\Delta = (1, 3, 0, -2, 0, 2)$

- $7 \to 0$ before $7 \to 6 \to 5 \to 4 \to 3 \to 2 \to 1 \to 0$
- $3 \to 4 \to 5 \to 6$ before $3 \to 2 \to 1 \to 0 \to 7 \to 6$
- 5
- $1 \to 0 \to 7$ before $1 \to 2 \to 3 \to 4 \to 5 \to 6 \to 7$
- 2
- $2 \to 3 \to 4$ before $2 \to 1 \to 0 \to 7 \to 6 \to 5 \to 4$

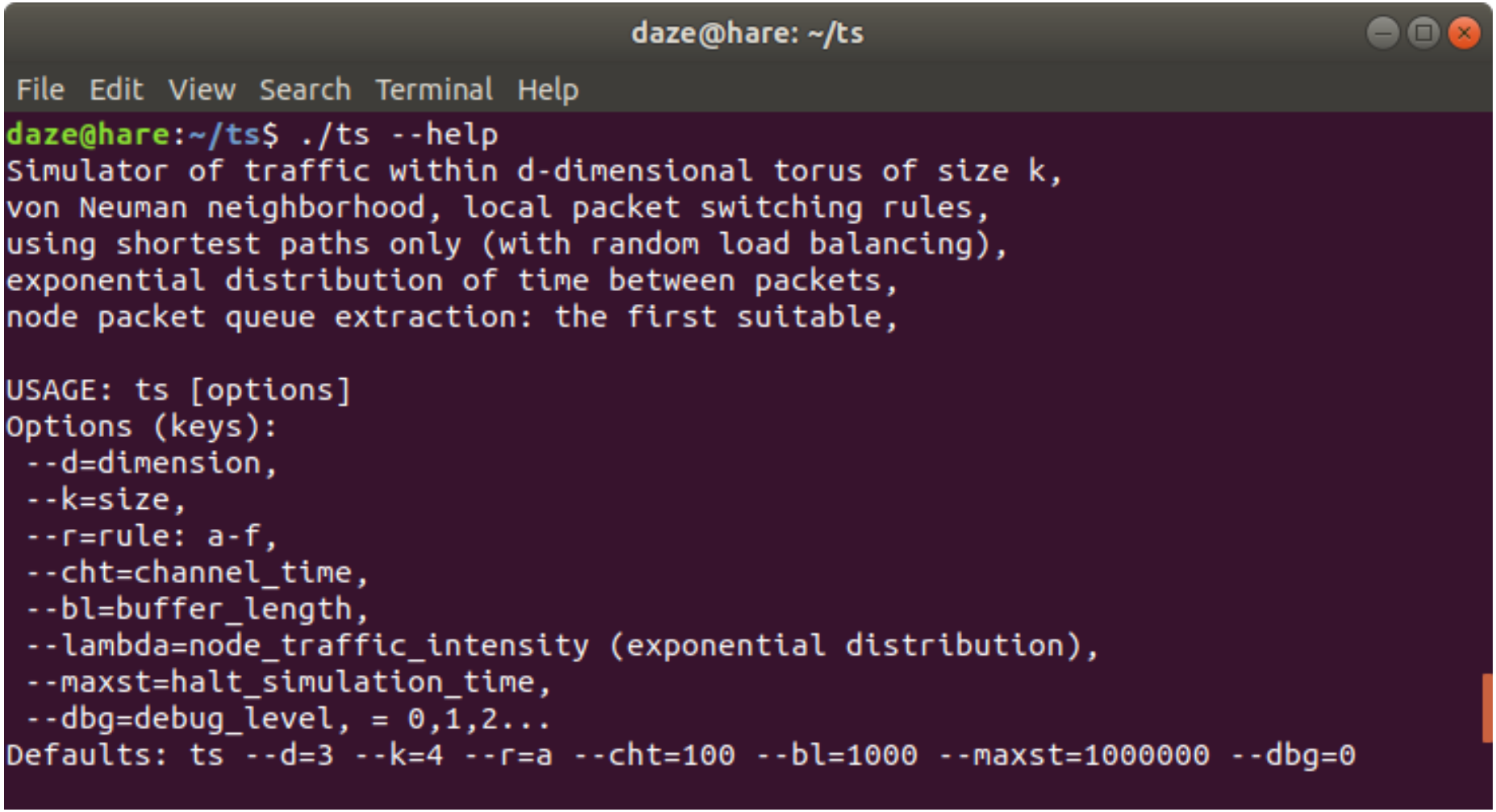
Port for Packet Forwarding

- A) Port $(0,1)$
- B) Random choice between ports $(0,1)$, $(1,1)$, $(3,-1)$, $(5,1)$
- C) Random choice between ports $(0,1)$, $(1,1)$, $(3,-1)$, $(5,1)$ with probabilities $(\frac{1}{8},\frac{3}{8},\frac{1}{4},\frac{1}{4})$
- D) Port (1,1)
- E) Random choice between ports $(1,1)$, $(3,-1)$

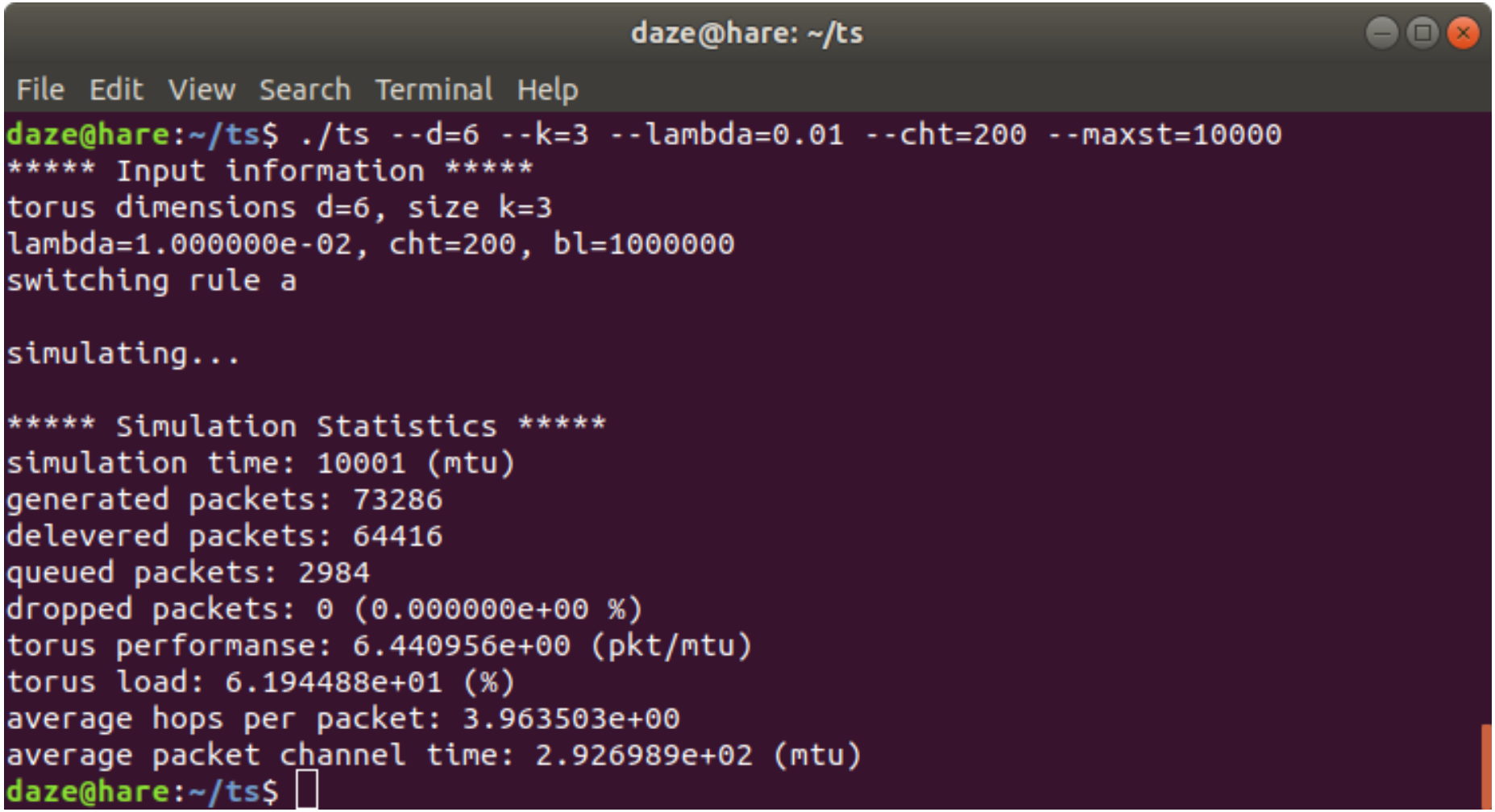
F)


i'
0 6 5 7 2 4
destination address
i
7 3 5 1 2 2
current node address
Δ
1 3 0 -2 0 2
signed difference
a
1 – 1 1 – 1
– 1 1 – – –
availability of ports
Δ*
0 3 0 -2 0 0
difference for available ports
z
0 1 2 3 4
port (1,1)
port (3,-1)
ranges for random number cast

http://github/dazeorgacm/ts

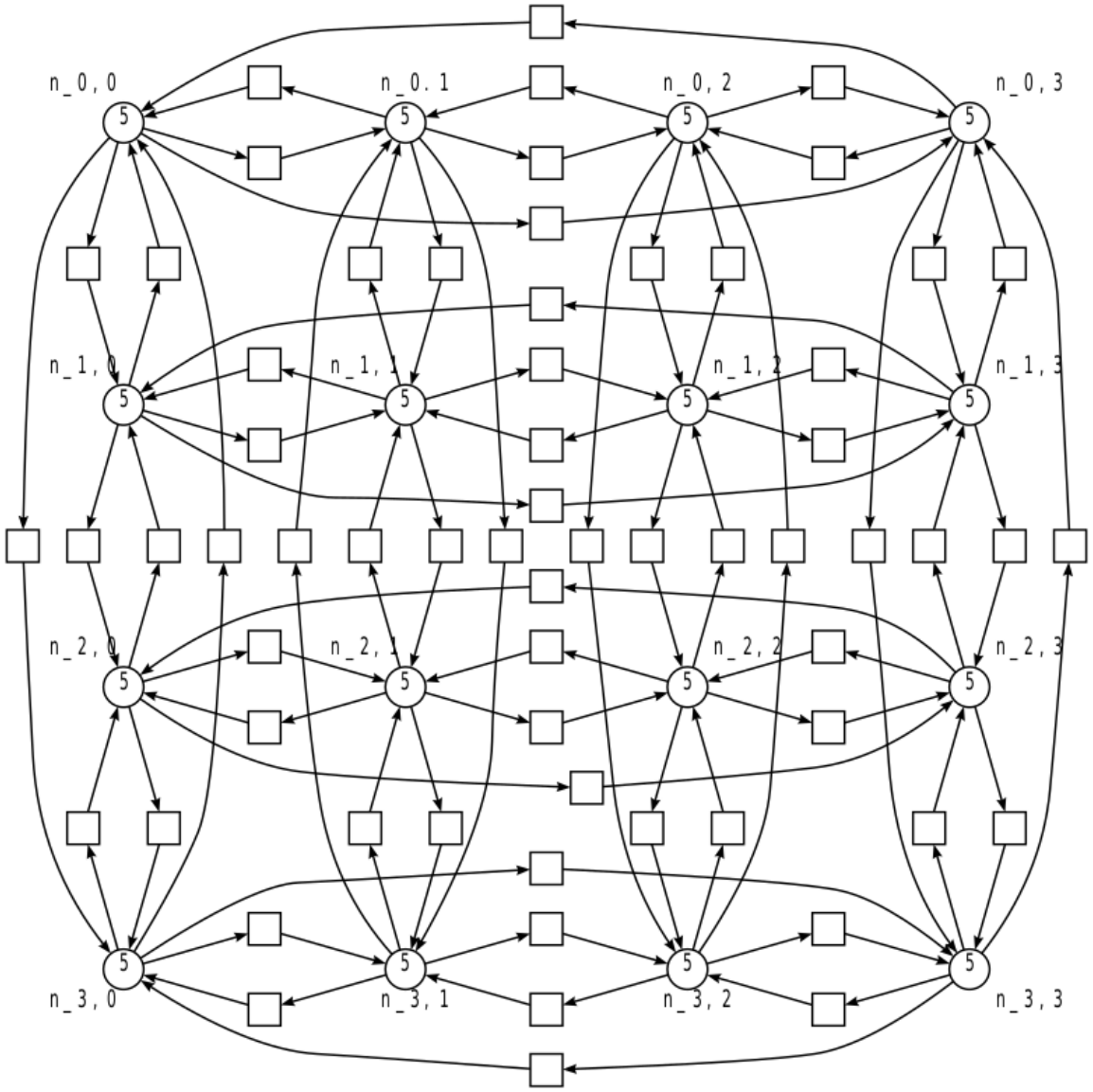


Torus Simulator



Create Packet Forwarding Canvas

2D Torus
4x4
Example



Event Queue Records

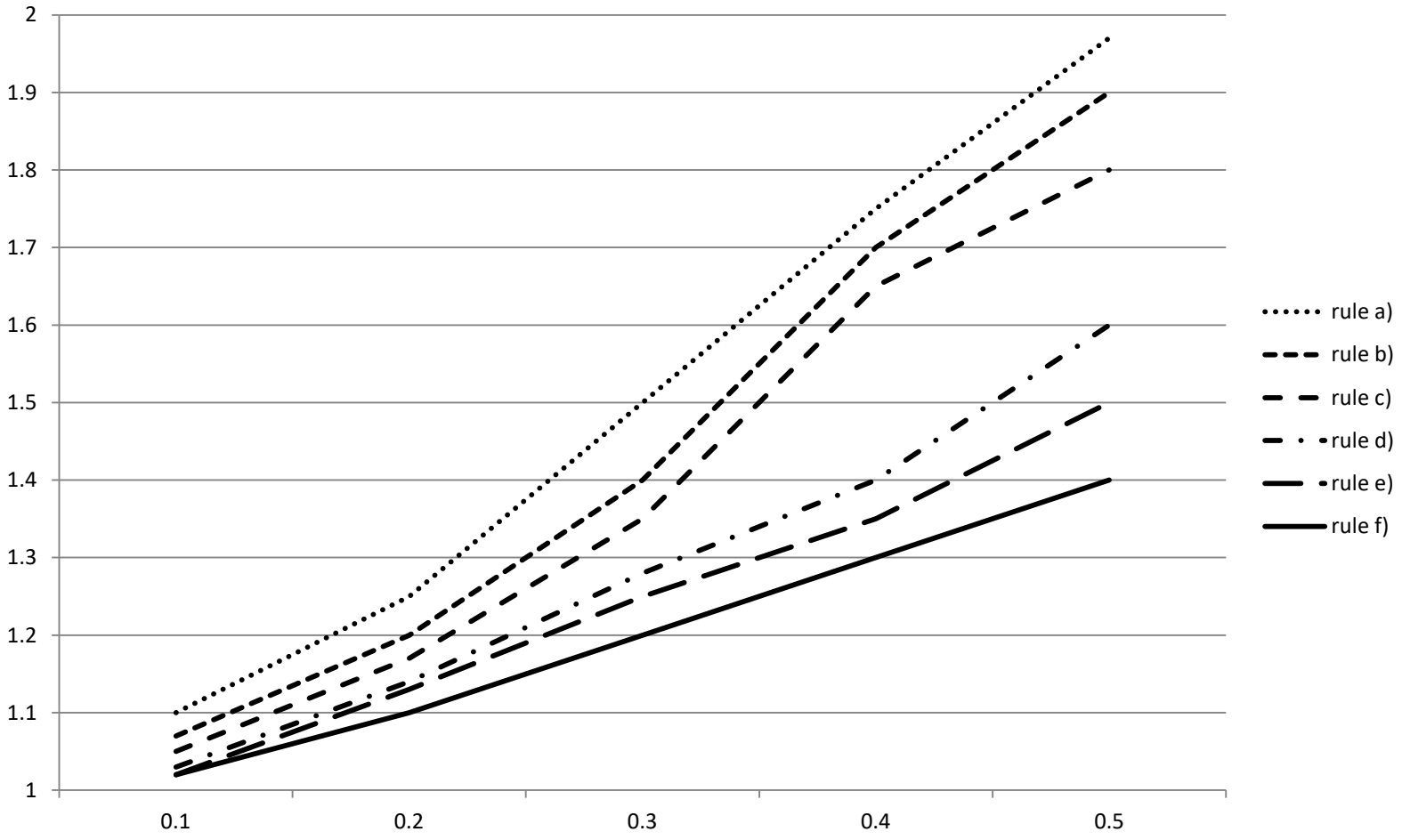
i. Insertion of a packet into a node:
   - packet generation in the source node;
   - packet arrival to an intermediate or destination node.

ii. Finishing of a packet transmission via a channel:
   - checking node queue for a packet waiting the corresponding port;
   - moving a packet from port to the next node inducing event (i) for the next node.

Basic Loop of Simulator

1) termination conditions checked and if one of them is true, the simulation process halts;
2) the nearest time value among future events is chosen and the simulator current time is advanced to this moment of time;
3) all the events having activation time equal to (or less than) the simulator current time are processed.

Evaluation of Actual Channel Transmission Time



Conclusions

- Torus structure: short distance between nodes, alternative shortest paths, possibility of packets delivery based on local rules
- 6 local rules offered including randomized choice for load balancing
- Torus Simulator *ts* has been implemented
- Simulation results show rather good performance and QoS

Basic References

- D.A. Zaitsev, T.R. Shmeleva, and J.F. Groote, “Verification of Hyper-torus Communication Grids by Infinite Petri Nets and Process Algebra,” *IEEE/CAA Journal of Automatica Sinica*, 6(3), 2019, 733-742.
- T.R. Shmeleva, “Analysis of a Hypertorus Grid,” In *Proc. of IEEE 38th International Conference Electronics and Nanotechnology ELNANO-2018*, Kyiv, Ukraine, April 24-26, 2018. NTUU "Igor Sikorsky Kyiv Polytechnic Institute", 2018. P. 56-59.
- D.A. Zaitsev, T.R. Shmeleva, W. Retschitzegger, B. Proll, “Security of grid structures under disguised traffic attacks,” *Cluster Computing*, vol. 19, no. 3, 2016, pp. 1183-1200.
- D.A. Zaitsev, “Verification of Computing Grids with Special Edge Conditions by Infinite Petri Nets,” *Automatic Control and Computer Sciences*, 2013, vol. 47, no. 7, pp. 403-412.
- D.A. Zaitsev, “A generalized neighborhood for cellular automata,” *Theoretical Computer Science*, vol. 666, 2017, pp. 21-35,