EARLY EVALUATION OF DIRECT LARGE-SCALE INFINIBAND NETWORKS WITH ADAPTIVE ROUTING

March 31, 2015

Alexander Daryin
Head of Advanced Research Lab
alexander.daryin@t-platforms.ru

Overview

- Introduction
- Reference Configuration
- Topologies
- Suggested IB Adaptive Routing Features
- Adaptive Routing Algorithms
- Performance Evaluation

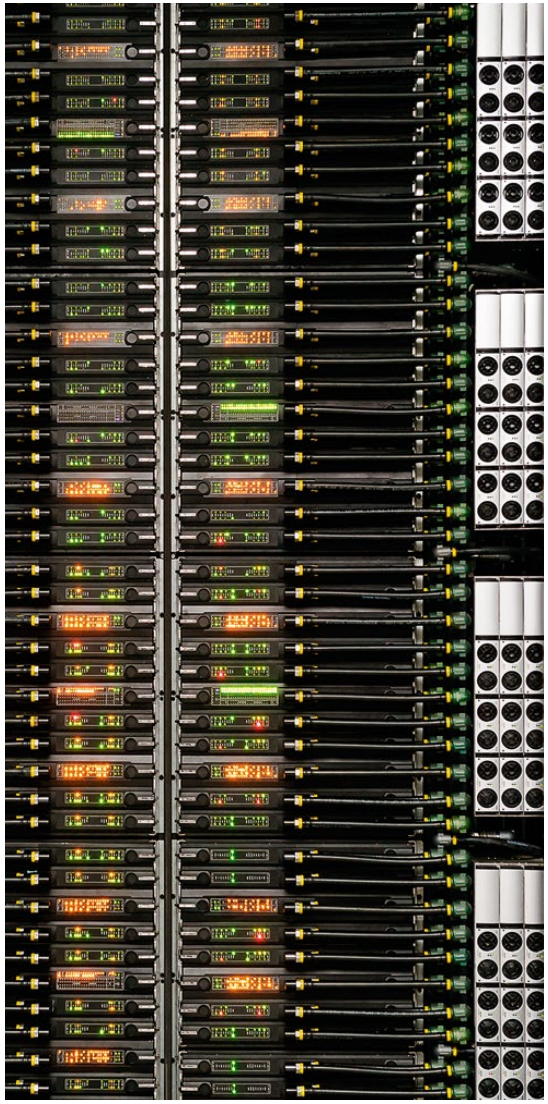
- Published paper:
  - A. Daryin, A. Korzh. Early evaluation of direct large-scale InfiniBand networks with adaptive routing // Supercomputing Frontiers and Innovations, Vol 1, No 3 (2014) pp 56-69.

Introduction

- **InfiniBand:** 45% of Nov 14 Top500 list
- Static (deterministic) routing
  - In-order packet delivery
- High-radix switches
  - Low-diameter topologies
  - High concentration
  - **Need for advanced routing algorithms**

Reference Configuration

- **T-Platforms A-Class**
  - Concentration 8 nodes/switch
  - 36 ports/switch
  - 32 switches/rack
- 32 twin racks (maximum 48):
  - 2048 switches, 16384 nodes
- Theoretical bounds:
  - Diameter: **3**
  - Relative bisection:
    - upper bound ≈ 163%
    - Ramanujan graphs: 100%
    - practical topologies: **50%**



Direct Topologies

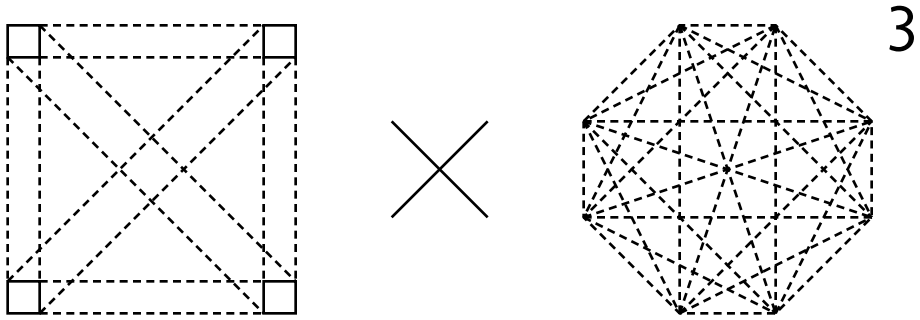
- **Flattened Butterfly**
- Dragonfly
- Slim Fly

- Torus
- Hypercube

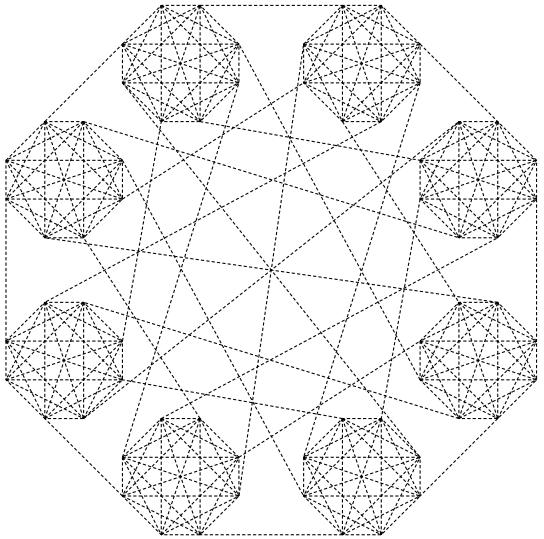
Flattened Butterfly

- Cartesian product of full graphs
- Alternative names:
  - Generalized Hypercube
  - HyperX

- Configuration:
  - Dimensions: 4×8×8×8
  - Link widths: 2, 1, 1, 1
  - Diameter: 4
  - Radix: 35
  - Relative bisection: 50%



Low Diameter Topologies



- **Dragonfly** (Kim, Dally, Scott, Abts 2008)
- Configuration:
  - 128 Groups of 16 switches
  - 8 global links/switch
  - Diameter: 3
  - Radix: 30
  - Relative bisection: 50%

- **SlimFly** (Besta, Hoefler 2014)
- MMS graph, diameter 2
- Configuration:

	Slim Fly q	L	FlatFly K_n	L_n	Degree	Diameter	Bisection
SF×FF-1	4	2	8×8	1, 1	34	4	50%
SF×FF-2	8	1	16	1	35	3	50%

Classical Topologies

- **Torus:** Maximum dimension: 4D (3D supported in OpenSM)

Dimension	Size	Link Widths	Degree	Diameter	Bisection
3D	$8 \times 16 \times 16$	3, 5, 5	34	20	15,6%
4D	$4 \times 8 \times 8 \times 8$	2, 4, 4, 4	36	14	25,0%

- **Hypercube:** Particular case of both Torus and Flattened Butterfly
  - Dimension (=diameter): 11D
  - Relative bisection: 25%
  - Radix: 30

Standard IB Routing

- LID (Local ID): 16-bit network address (48K addresses)
- SL (Service Level): 4-bit packet tag
- VL (Virtual Lane): 8 lanes per port

- Routing mechanisms:
  - PathSL: select packet SL by source and destination
  - LFT (Linear Forwarding Table): select output port by DLID
  - SL2VL: select output VL by SL, input and output port
  - LMC (LID Mask Control): assign several LIDs per endpoint

- Routing function:

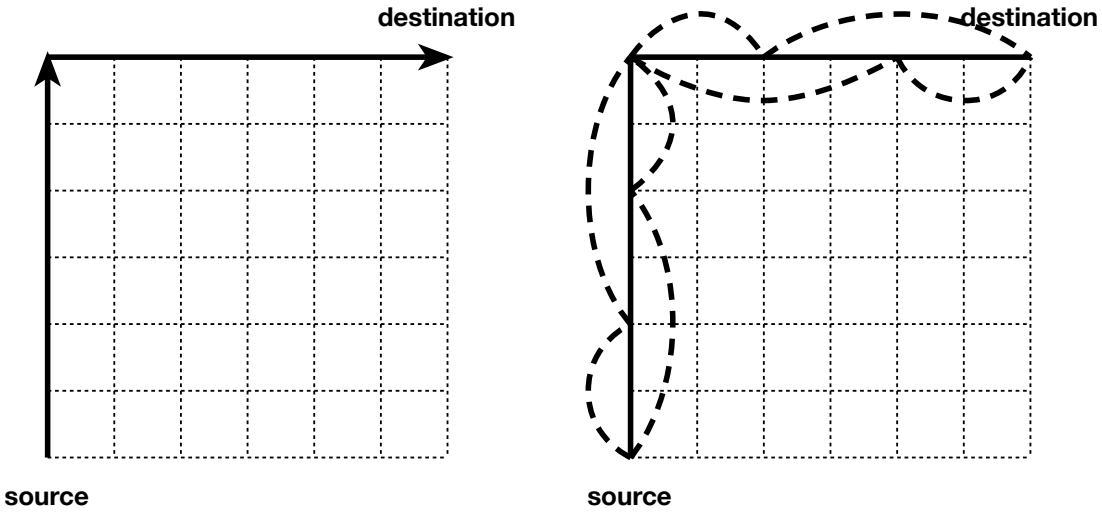
$$\begin{cases} (\text{Source}, \text{Destination}) \rightarrow (\text{DLID}, \text{SL}), \\ (\text{Switch}, \text{DLID}) \rightarrow \text{Out Port}, \\ (\text{Switch}, \text{SL}, \text{In Port}, \text{Out Port}) \rightarrow \text{Out VL}. \end{cases}$$

Adaptive Routing: Suggested Features

- Minimal AR:
  - Each LFT entry contains a set of ports
- Non-minimal AR
  - Need to track number of hops so far
    - VL is the only changing header field
  - SL2VL replaced with VL2VL
  - Multiple copies of LFT, one per input VL
  - 2 or more priority levels
- Routing Function:

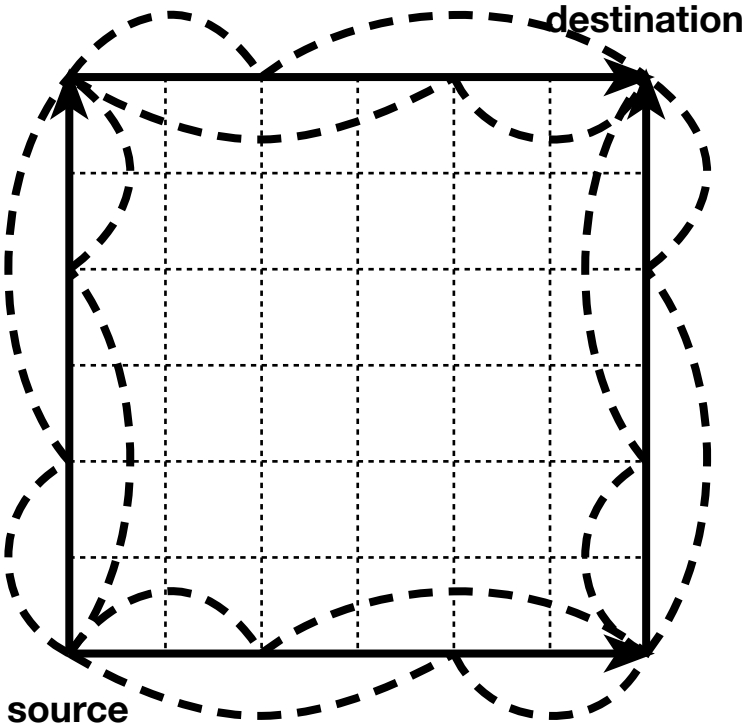
$$\begin{cases} (\text{Source}, \text{Destination}) \rightarrow (\text{DLID}, \text{VL}), \\ (\text{Switch}, \text{In VL}, \text{DLID}) \rightarrow \{\text{Out Port}_1, \ldots, \text{Out Port}_k\}, \\ (\text{Switch}, \text{In VL}, \text{In Port}, \text{Out Port}) \rightarrow \text{Out VL}. \end{cases}$$

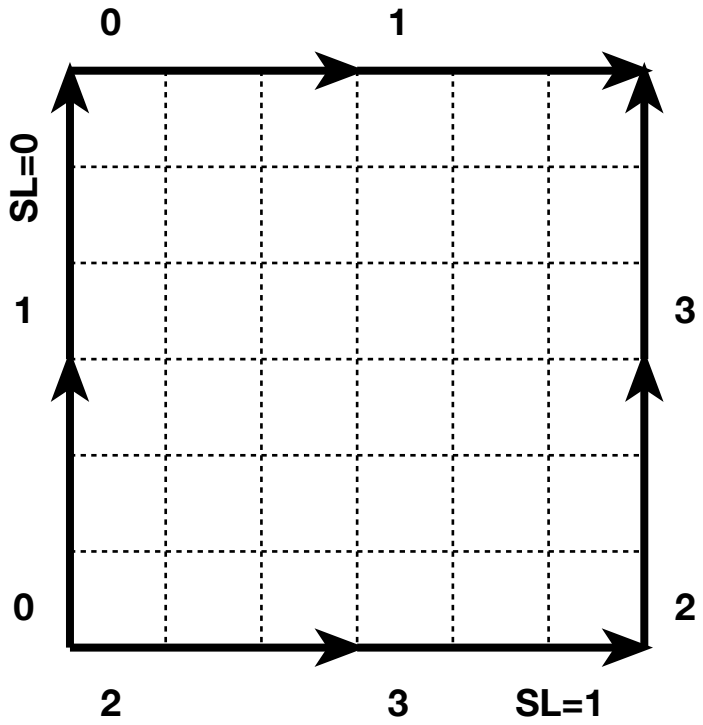
FlatFly Routing: Adaptive DOR



FlatFly Routing: Mixed DOR

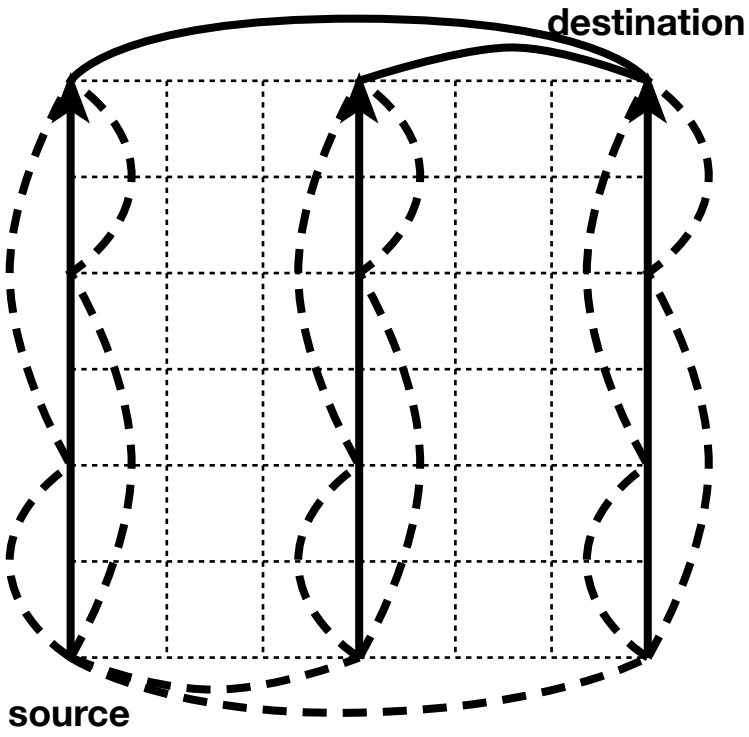
- 4 dimension orders based on destination, encoded in SL:
  - 1234, 2413, 3142, 4321

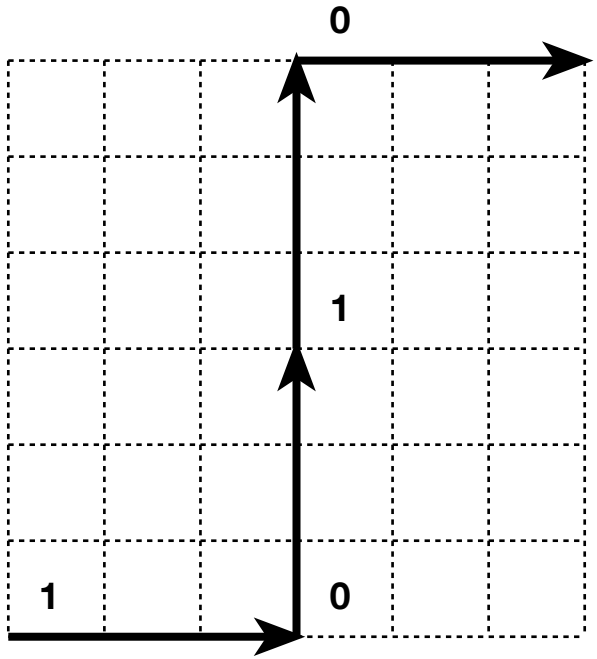




FlatFly Routing: Twisted DOR

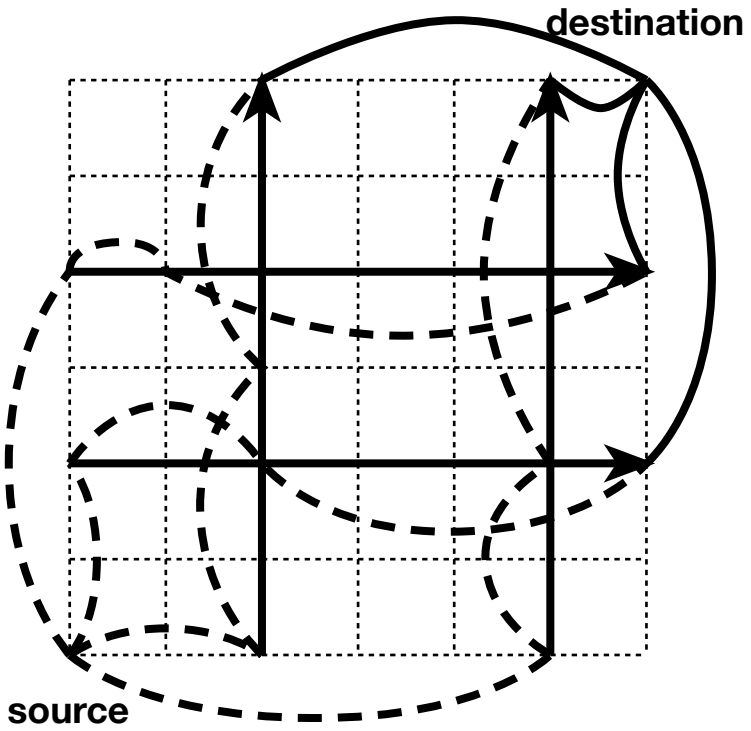
- Compulsory initial extra hop in the last dimension

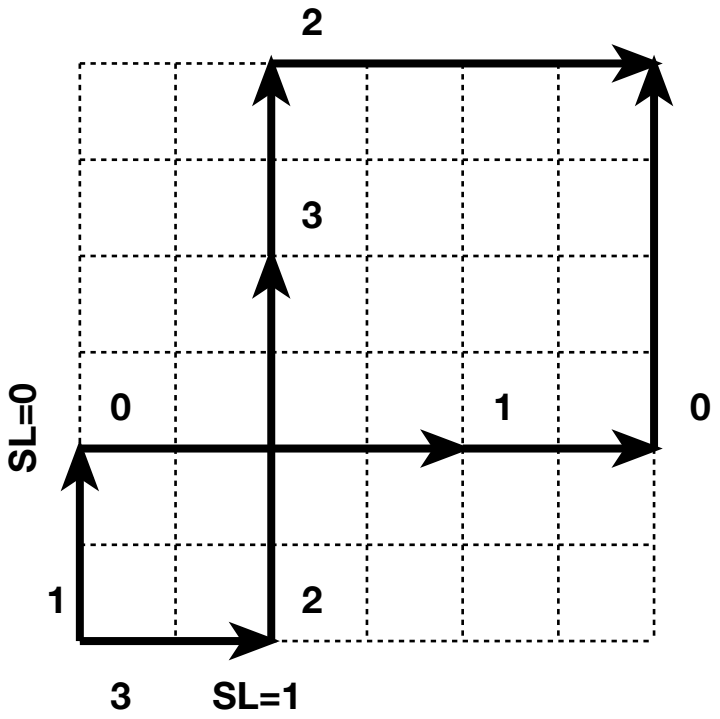




FlatFly Routing: Twisted Mixed DOR

- Combination of two previous routings





Topology Agnostic Adaptive Routing

- **Distance Based** *+k*
  - Set VL = d(*src*, *dest*) + *k* - 1
  - Route from *curr* do *dest* via *next*
    - with priority 1 if d(*next*, *dest*) = d(*curr*, *dest*) - 1
    - with priority 2 if d(*next*, *dest*) = d(*curr*, *dest*) ≤ VL
  - Decrease VL at each hop

- **Deflection** *+k*
  - Same as above, plus
  - Route from *curr* do *dest* via *next*
    - with priority 3 if d(*next*, *dest*) = d(*curr*, *dest*) + 1 ≤ VL

Simulation Results

- Saturation levels in % of maximum bandwidth

Routing	All2All	Bit				
		Cmpl	Rvrs	Rotn	Shuf	Trns
		Torus				
DOR 3D	26,5	15,6	7,6	7,0	7,6	7,1
DOR 4D	48,2	25,0	2,9	17,0	12,2	2,9
		Hypercube				
DOR	24,2	24,2	0,8	24,2	11,7	0,8
Mixed DOR	24,2	24,2	2,3	11,7	11,7	0,8
		Dragonfly				
Static	24,2	5,5	5,5	5,5	5,5	5,5
GSID	23,4	24,2	5,5	24,2	24,2	5,5
Distance	82	0,0	5,5	0,8	0,8	5,5
Distance +2	59,4	0,8	11,7	2,3	2,3	9,4
Deflection +3	48,4	2,3	11,7	2,3	5,5	11,7

Simulation Results

Routing	All2All	Bit				
		Cmpl	Rvrs	Rotn	Shuf	Trns
	SlimFly × FlatFly 1					
Distance	80,5	11,7	11,7	10,9	5,5	9,4
Distance +2	80,5	19,5	30,5	11,7	11,7	11,7
Deflection +4	80,5	21,9	24,2	24,2	24,2	32,8
	SlimFly × FlatFly 2					
Distance	61,7	5,5	5,5	5,5	5,5	5,5
Distance +2	61,7	11,7	14,8	7,0	7,0	8,6
Deflection +4	74,2	11,7	18,8	11,7	10,9	11,7

Simulation Results

Routing	All2All	Bit				
		Cmpl	Rvrs	Rotn	Shuf	Trns
	FlatFly					
DOR	81,3	11,7	1,6	12,5	7,8	2,3
ADOR	52,3	50	3,1	25,0	24,2	2,3
Mixed ADOR	56,3	50,0	10,2	24,2	24,2	5,5
Twisted ADOR	24,2	50,0	21,1	24,2	24,2	11,7
Twisted Mixed	53,1	28,1	22,7	24,2	24,2	5,5
Distance	93,8	11,7	9,4	11,7	9,4	10,2
Distance +4	96,1	50,8	14,8	39,8	37,5	37,5
Deflection +4	94,5	50,0	24,2	50,0	24,2	24,2