Dual-Centric Data Center Network Architectures

JIE WU (TEMPLE UNIVERSITY)

PHD CANDIDATE: DAWEI LI

AUGUST 16, 2016

Agenda

Introduction

Unified Performance Model

Dual-Centric DCN Architectures

- FCell and FSquare

Comparisons of DCN Architectures

Simulations

Conclusion and Future Works

Introduction

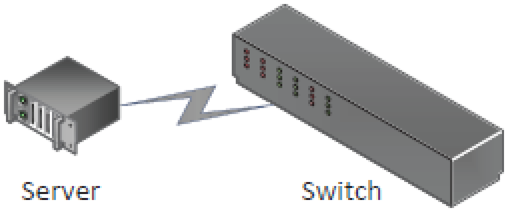
Data centers are important infrastructures to support various cloud computing services.

- Web search
- Email
- Video streaming
- Social networking
- Distributed file systems
- Distributed data processing

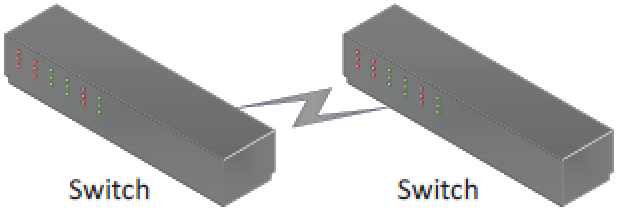
Introduction

Three types of connections:

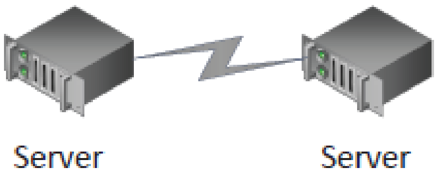
- Server-switch connection (a)
- Switch-switch connection (b)
- Server-server connection (c)



(a) Server-switch connection



(b) Switch-switch connection



(c) Server-server connection

Two classes of DCNs:

- Switch-centric
  - Only server-switch and switch-switch connections (a and b), no server-server
  - Eg, Fat-Tree , Flattened Butterfly
- Server-centric
  - Mostly, only server-switch and server-server connections (a and c), no switch-switch
  - Eg: BCube, FiConn, DCell

Introduction

Switch-centric vs. Server-centric

- Server-centric architectures
  - enjoy the high programmability of servers, but servers usually have larger processing delays than do switches.
- Switch-centric architectures
  - enjoy the fast switching capability of switches, but switches are less programmable than servers.
- Can we combine the advantages of both categories?

Introduction

Performance vs. Power Consumption

- To provide low end-to-end delays and high bisection bandwidth
  - Large numbers of networking devices are usually used in DCNs.
  - E.g, Fat-Tree: three levels of switches; BCube: three or more levels & extra Network Interface Card (NIC) ports.
- To achieve a low DCN power consumption
  - Other architectures use significantly fewer networking devices.
  - E.g, FiConn, DPillar etc.
- Can we achieve high performances and low power consumption at the same time?

Introduction

Overview

- Unified performance model
  - Path length (and hence diameter)
  - Power consumption
- A new category of DCN architectures: Dual-Centric
  - FCell and FSquare
  - Achieving tradeoffs in the design spectrum
- A range of DCN architectures
  - Comparison of existing architectures under our unified performance models

Agenda

Introduction

Unified Performance Model

Dual-Centric DCN Architectures

- FCell and FSquare

Comparisons of DCN Architectures

Simulations

Conclusion and Future Works

Unified Performance Model

- Unified Path Length Definition:

$$d_P = n_{P,w} d_w + (n_{P,v} + 1) d_v,$$

$n_{P,w}$: # of switches in a path
$n_{P,v}$: # of servers in a path (excluding *s* and *d*)
d_w : processing delay on a switch
d_v : processing delay on a server

- Unified Diameter in a DCN:

$$d = \max_{P \in \{\mathcal{P}\}} d_P,$$

Unified Performance Model

- DCN Power Consumption per Server:

$$p_V = p_{dcn}/N_v = p_w N_w/N_v + n_{nic}p_{nic} + \alpha p_{fwd}.$$

p_w : power consumption of a switch
N_w : # of switches in a DCN
N_v : # of servers in a DCN
n_{nic} : average # of NIC ports each server uses
p_{nic} : power consumption of a NIC port
α : whether the server is involved in packet relaying
p_{fwd} : power consumption of a server's packet forwarding

Unified Performance Model

- Bisection Bandwidth (B):
  - The *minimum number of links* to be removed, to partition all servers in the network into two “equal” halves.
    - When the total number of servers is odd, the sizes of the two halves should differ by 1.

Agenda

Introduction

Unified Performance Model

Dual-Centric DCN Architectures

- FCell and FSquare

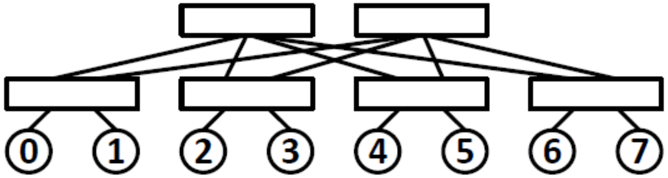
Comparisons of DCN Architectures

Simulations

Conclusion and Future Works

Dual-Centric DCNs: FCell

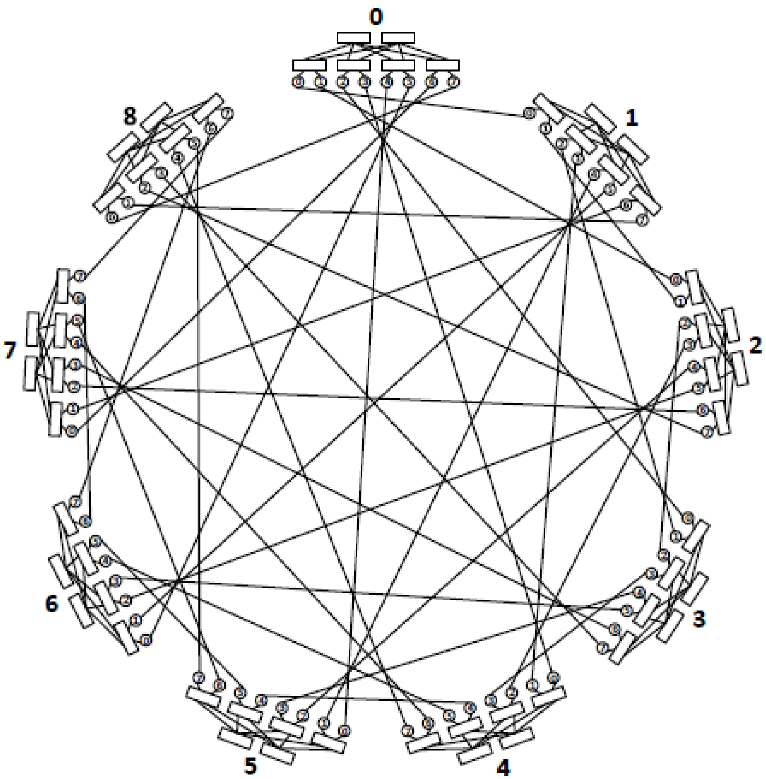
- Intra-cluster
  - The switches and servers form a simple instance of the folded Clos topology. We call it a cluster.
  - All switches are with **n ports**.
  - There are n level 1 switches, and n/2 level 2 switches.
  - Each level 1 switch uses n/2 ports to connect to n/2 servers, and n/2 ports to connect to n/2 level 2 switches.



(a) The interconnections in one cluster. (n = 4)

Dual-Centric DCNs: FCell

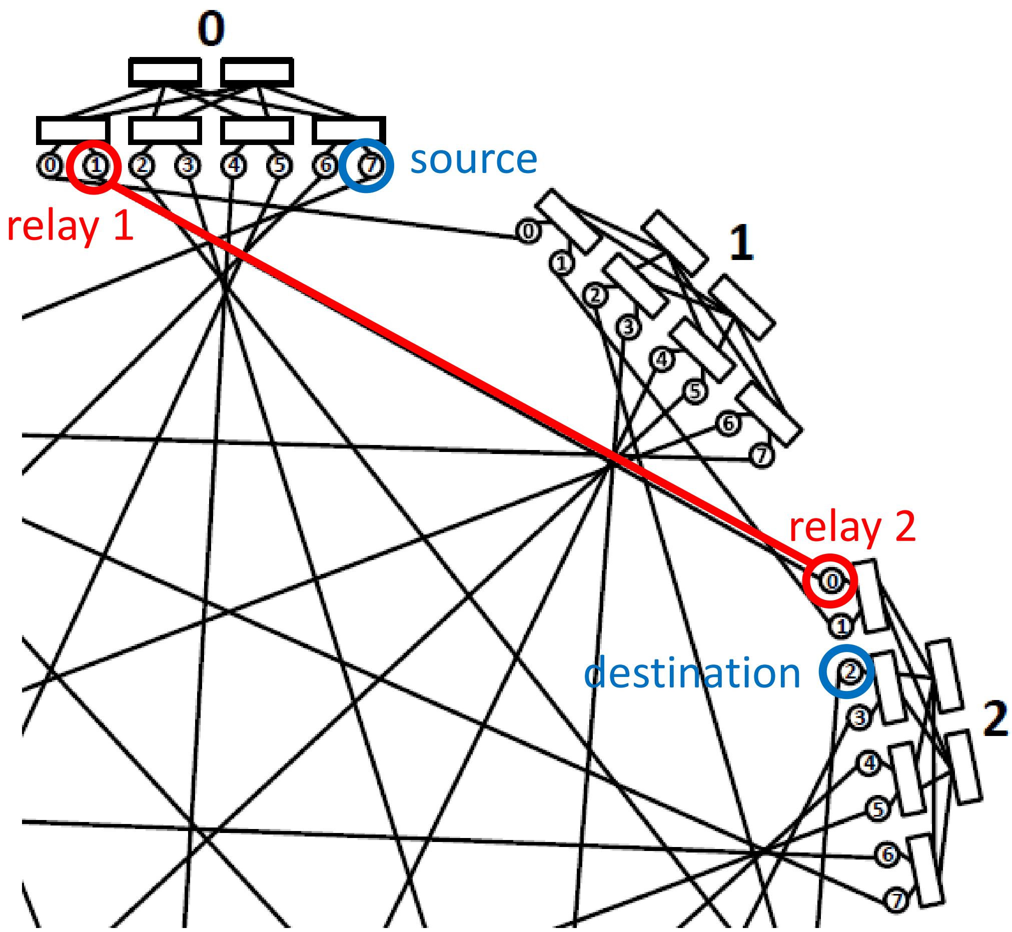
- Inter-cluster
  - Each of the servers in a cluster is directly connected to another server in each of the other clusters.
- Each server has 2 NIC ports and each switch has n ports
  - (n/2)n servers in each cluster.
  - Total (n/2)n+1 clusters.



(b) Final interconnections of FCell(4).

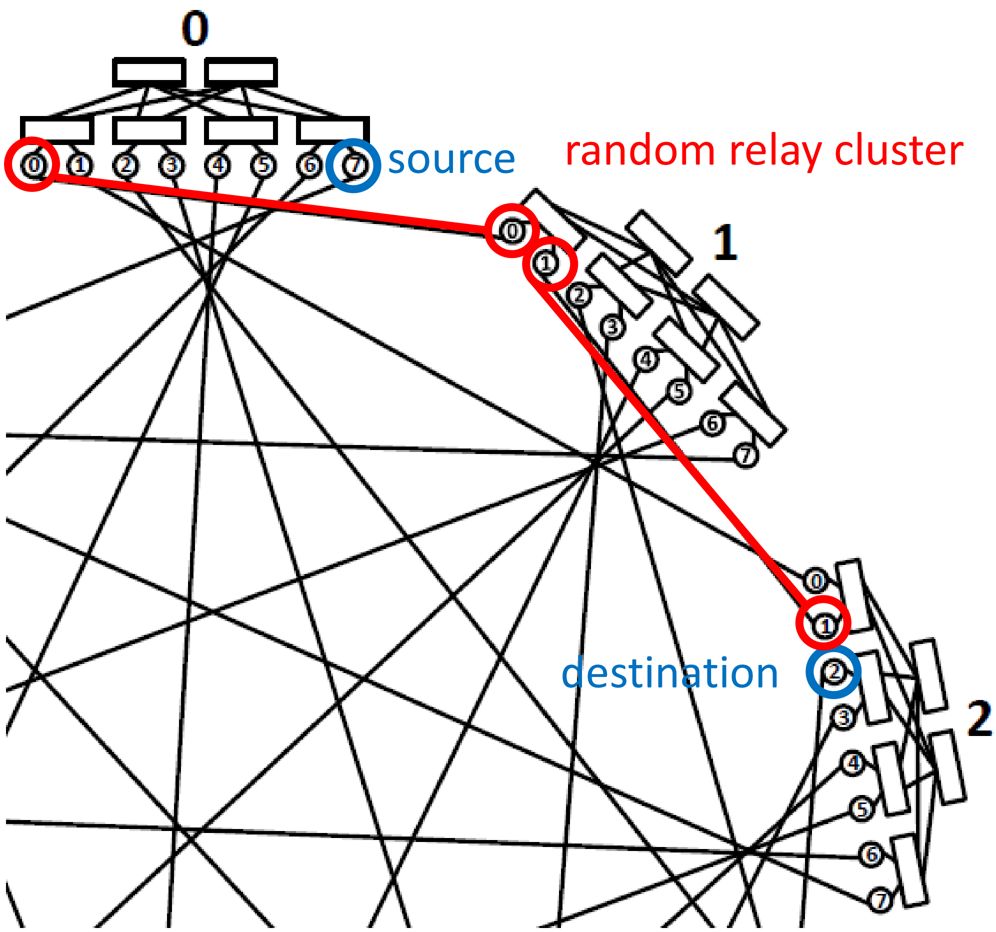
Dual-Centric DCNs: FCell

- Routing in FCell
  - Shortest Path Routing:
    - Determines the relay servers.
    - Source to relay 1 in the source cluster.
    - Relay 1 to relay 2.
    - Relay 2 to destination in the destination cluster.



Dual-Centric DCNs: FCell

- Detour Routing:
  - Randomly select a relay cluster.
  - Conduct shortest path routing from the source cluster to the relay cluster.
  - And then, from relay cluster to destination cluster.



Dual-Centric DCNs: FCell

- FCell basic properties:

Property 1. *In an FCell*(n)*, the number of switches is* $N_w = 3n(n^2+2)/4$*, and the number of servers is* $N_v = n^2(n^2+2)/4$*.*

Proof. There are $n^2/2+1$ clusters, each with $3n/2$ switches and $n^2/2$ servers. $\square$

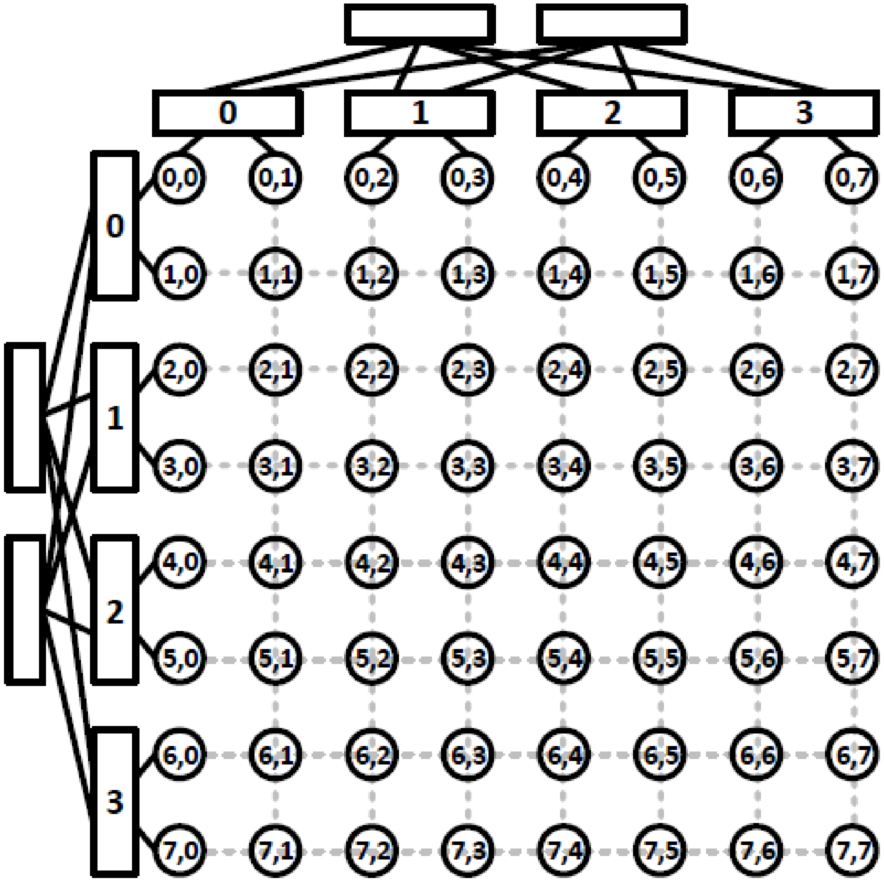
Property 2. *The diameter of an FCell*(n) *is* $d = 6d_w + 3d_v$*.*

Property 3. *The bisection bandwidth of an FCell*(n) *is* $B \approx N_v/4$*.*

Property 4. *The DCN power consumption per server of an FCell*(n) *is* $p_V = 3p_w/n + 2p_{nic} + p_{fwd}$*.*

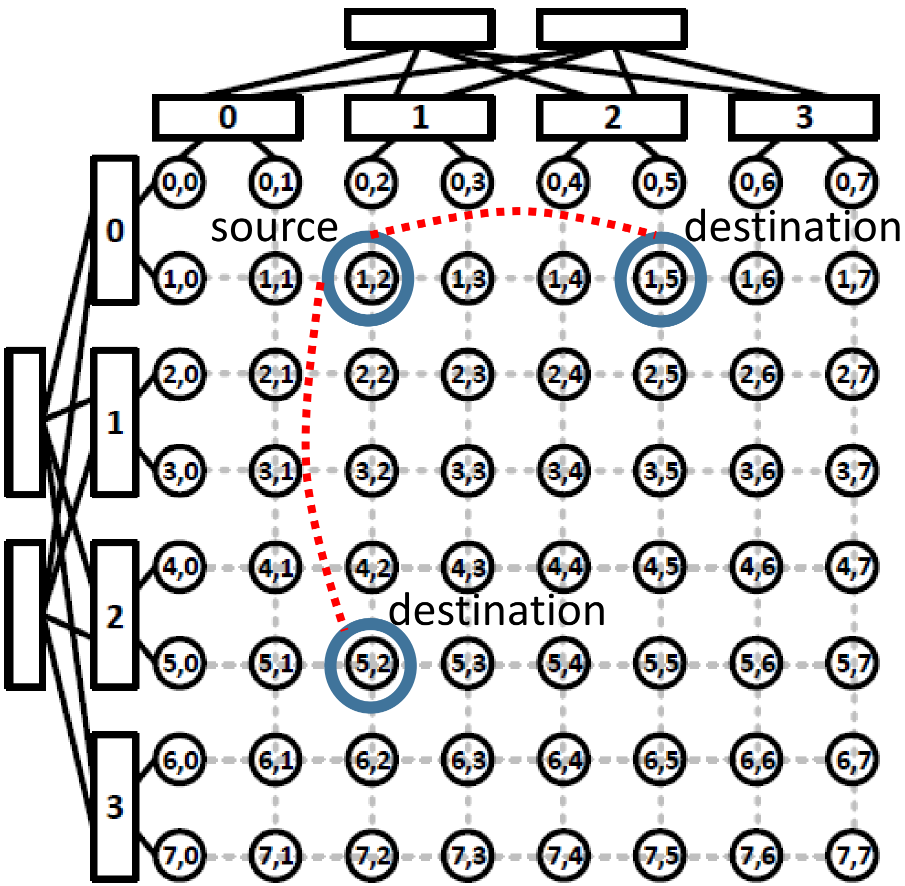
Dual-Centric DCNs: FSquare

- FSquare(n)
- Each column and each row form the same cluster as in FCell.
- i.e., in each cluster, the set of n/2 level 2 switches and the set of n level 1 switches form a complete bipartite graph.



Dual-Centric DCNs: FSquare

- Routing in FSquare(n):
  - If source and destination are in **the same row** (or the same column).
  - Routing only need to go through the switches in the row cluster (or the switches in the column cluster).

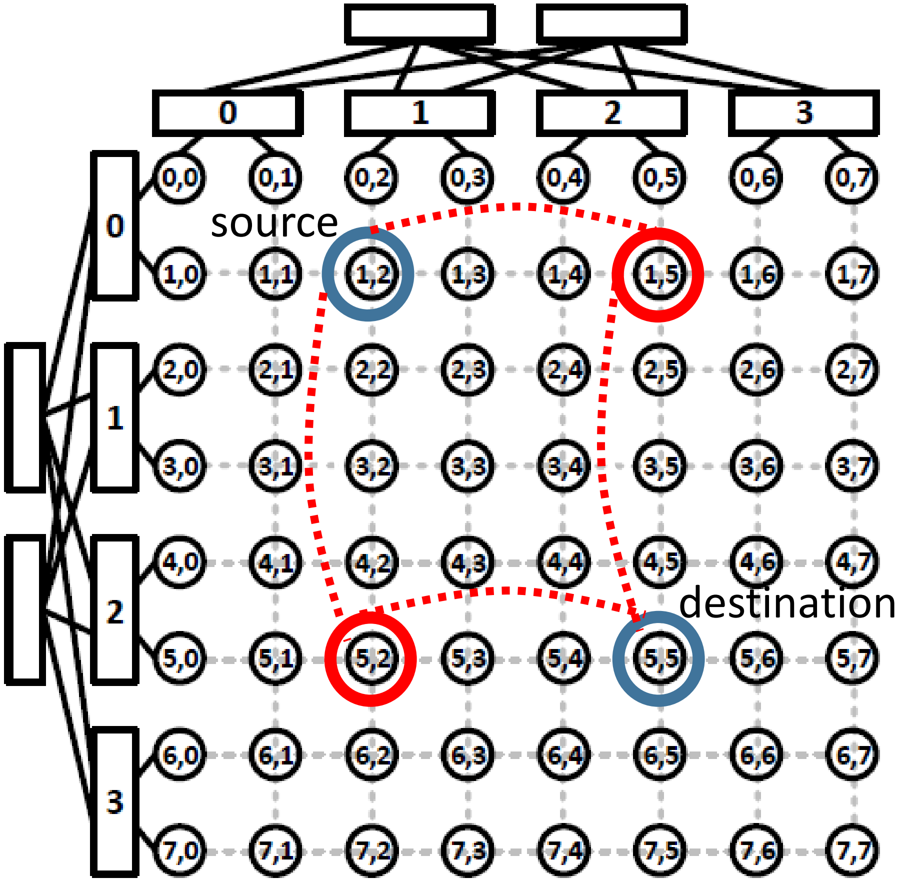


Dual-Centric DCNs: FSquare

- Routing in FSquare(n):

If source and destination are **not in the same row and not in the same column**.

Row first or column first, or based on traffic condition within the row or column.



Dual-Centric DCNs: FSquare

- FSquare Basic Properties

Property 1. *In an FSquare*(n)*, the number of servers is* $N_v = n^4/4$*, and the number of switches is* $N_w = 3n^3/2$.

Property 2. *FSquare*(n) *has a diameter of* $d{=}6d_w{+}2d_v$.

Property 3. *The bisection bandwidth of an FSquare*(n) *is* $B{=}N_v/2$.

Property 4. *The DCN power consumption per server of an FSquare*(n) *is* $p_V{=}6p_w/n + 2p_{nic} + p_{fwd}$.

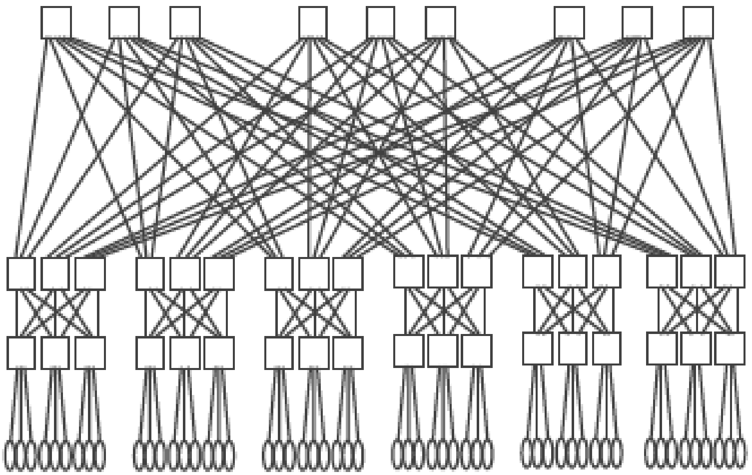
Agenda

- Introduction
- Unified Performance Model
- Dual-Centric DCN Architectures
  - FCell and FSquare
- Comparisons of DCN Architectures
- Simulations
- Conclusion and Future Works

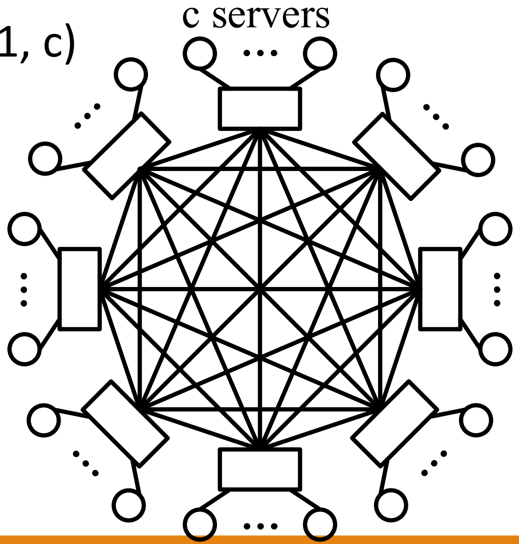
Comparisons of DCN Architectures

- Existing architectures:
  - Switch-centric
    - **Folded Clos** with k levels of n-port switches (**FDCL(n, k)**).
    - **Flattened Butterfly** (**FBFLY(r, k, c)**): switches form a generalized hypercube; then each switch connects to **c** servers. **r**: the number of switches in each dimension; **k**: the number of dimensions.

FDCL(6, 3)

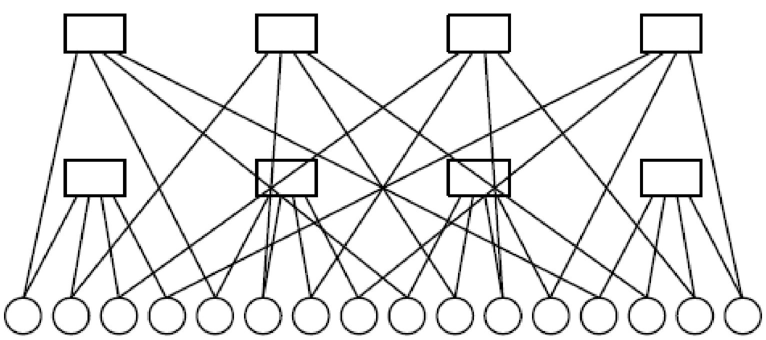


FBFLY(8, 1, c)

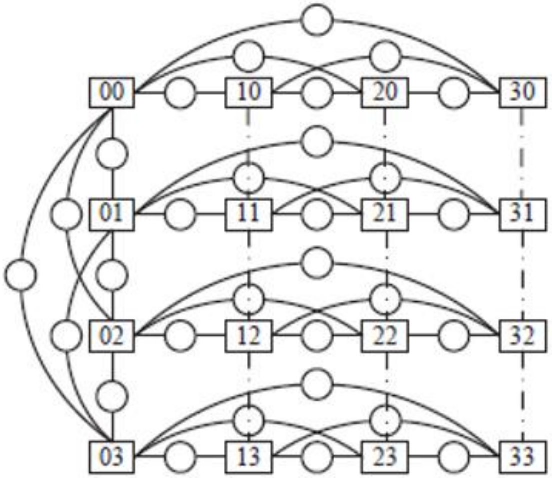


Comparisons of DCN Architectures

- Existing architectures:
  - Server-centric:
    - **BCube(n,k)**: with **n**-port switches and **k** levels.
    - **SWCube(r, k)**: switches form a generalized hypercube; then, servers are inserted into the links between switches. **k** is the number of dimensions. Each dimension has **r** switches.



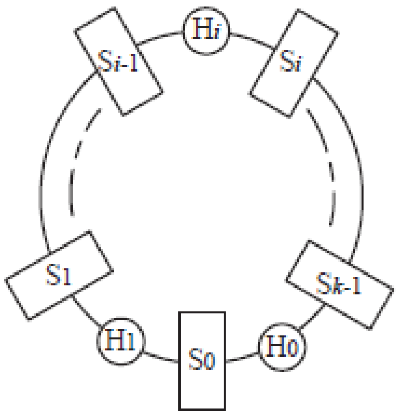
BCube(4,2)



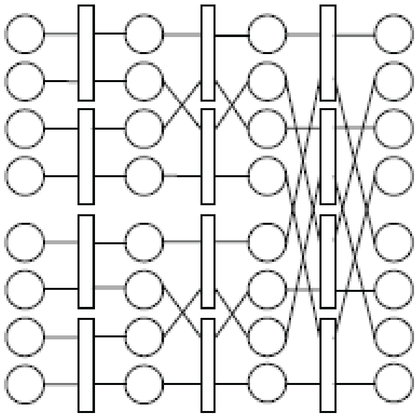
SWCube(4, 2), n = 6.

Comparisons of DCN Architectures

- Existing architectures:
  - Server-centric:
    - **DPillar(n, k)**: n-port switches and k levels (k columns/pods).



(a) vertical view

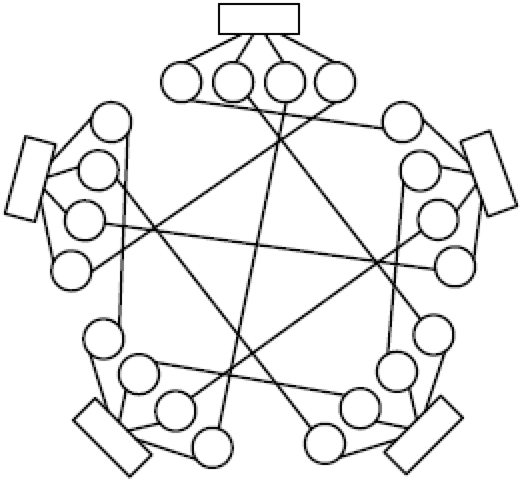


(b) horizontal view

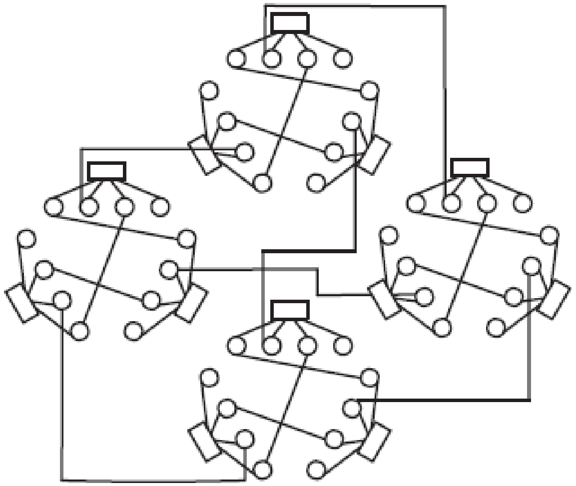
DPillar(4, 3) (the first column and the last column are overlapped.)

Comparisons of DCN Architectures

- Existing architectures:
  - Server-centric:
    - **DCell(n, k)**: n-port switches and k levels.
    - **FiConn(n, k):** n-port switches and k levels.



DCell(4, 2)

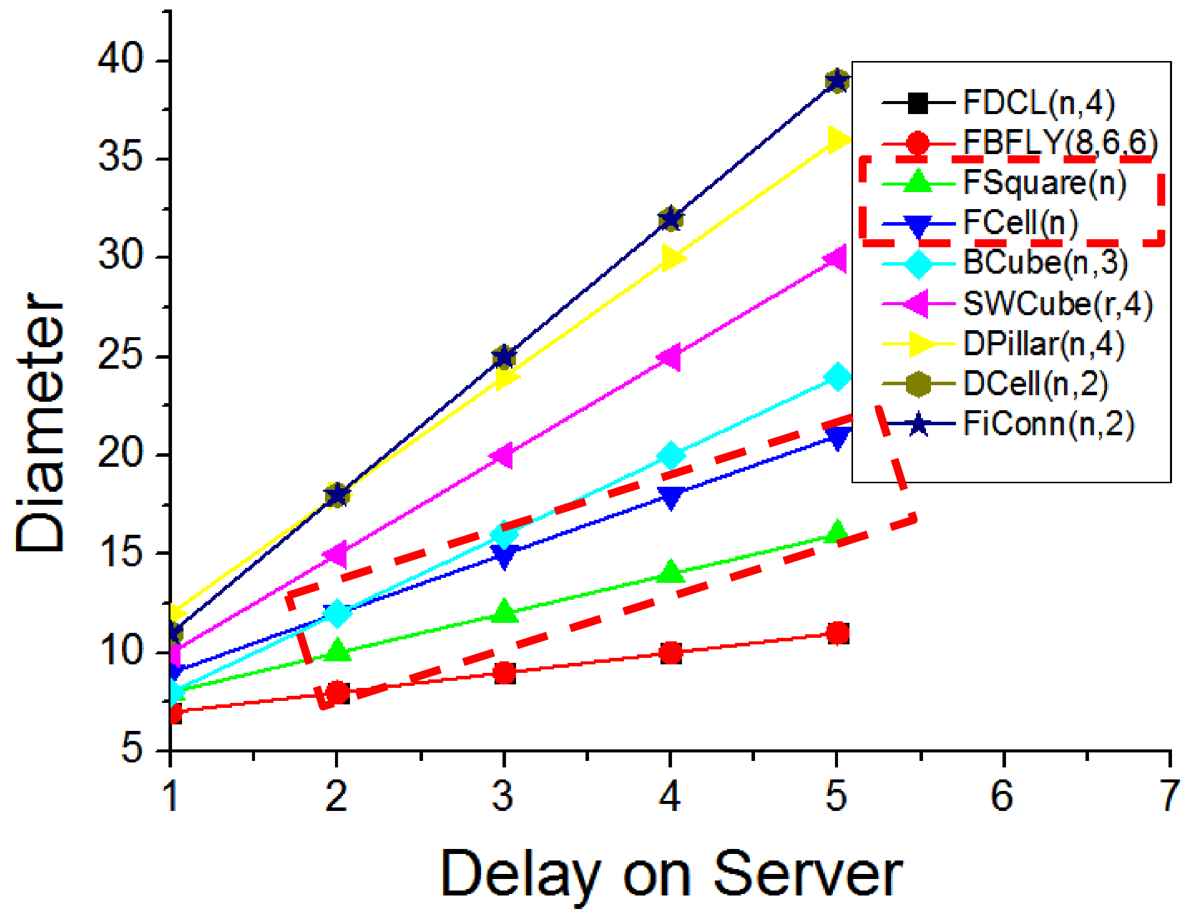


FiConn(4, 3)

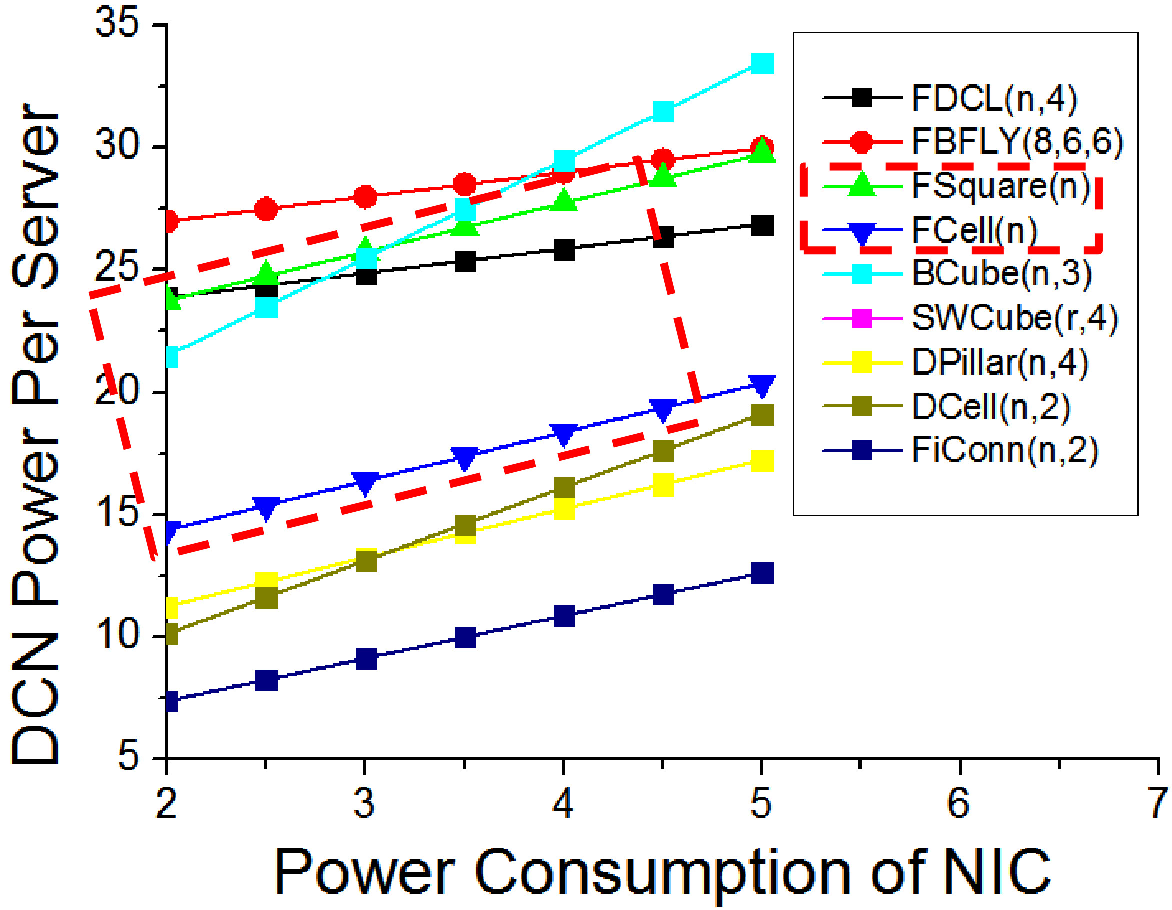
Comparisons of DCN Architectures

	$N_v(n{=}24)$	$N_v(n{=}48)$	N_w/N_v	d	B	p_V
FDCL$(n,4)$	41,472	663,552	$7/n$	$7d_w{+}d_v$	$N_v/2$	$7p_w/n+p_{nic}$
FBFLY$(4,7,3)$	49,125	—	$8/24$	$8d_w{+}d_v$	$N_v/3$	$8p_w/24+p_{nic}$
FBFLY$(8,6,6)$	—	1,572,864	$8/48$	$7d_w{+}d_v$	$N_v/3$	$8p_w/48+p_{nic}$
FSquare(n)	82,944	1,327,104	$6/n$	$6d_w{+}2d_v$	$N_v/2$	$6p_w/n+2p_{nic}+p_{fwd}$
FCell(n)	83,232	1,328,256	$3/n$	$6d_w{+}3d_v$	$N_v/4$	$3p_w/n+2p_{nic}+p_{fwd}$
BCube$(n,3)$	331,776	5,308,416	$4/n$	$4d_w{+}4d_v$	$N_v/2$	$4p_w/n+4p_{nic}+p_{fwd}$
SWCube$(r,4)$	28,812	685,464	$2/n$	$5d_w{+}5d_v$	$(N_v/8)\times r/(r-1)$	$2p_w/n+2p_{nic}+p_{fwd}$
DPillar$(n,4)$	82,944	1,327,104	$2/n$	$6d_w{+}6d_v$	$N_v/4$	$2p_w/n+2p_{nic}+p_{fwd}$
DCell$(n,2)$	360,600	5,534,256	$1/n$	$4d_w{+}7d_v$	$> N_v/(4\log_n N_v)$	$p_w/n+3p_{nic}+p_{fwd}$
FiConn$(n,2)$	24,648	361,200	$1/n$	$4d_w{+}7d_v$	$> N_v/16$	$p_w/n+7p_{nic}/4+3p_{fwd}/4$

Comparisons of DCN Architectures



Comparisons of DCN Architectures



Agenda

Introduction

Unified Performance Model

Dual-Centric DCN Architectures

- FCell and FSquare

Comparisons of DCN Architecture

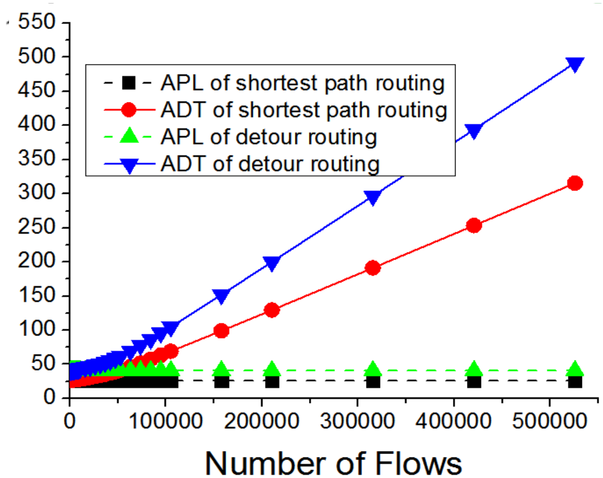
Simulations

Conclusion and Future Works

Simulations

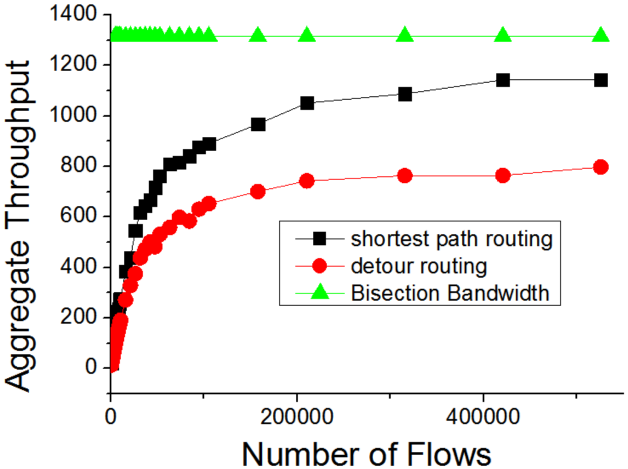
- Simulations for FCell
  - Random traffic and bursty traffic.
  - Metrics: Average Path Length (APL), Average Delivery Time (ADT), and Aggregate Throughput (amount of flow delivered in a unit time).

Simulations for random traffic: the performances of shortest path routing and detour routing demonstrate graceful degradation.



(a) APL and ADT.

ADT increases linearly with the number of flows.



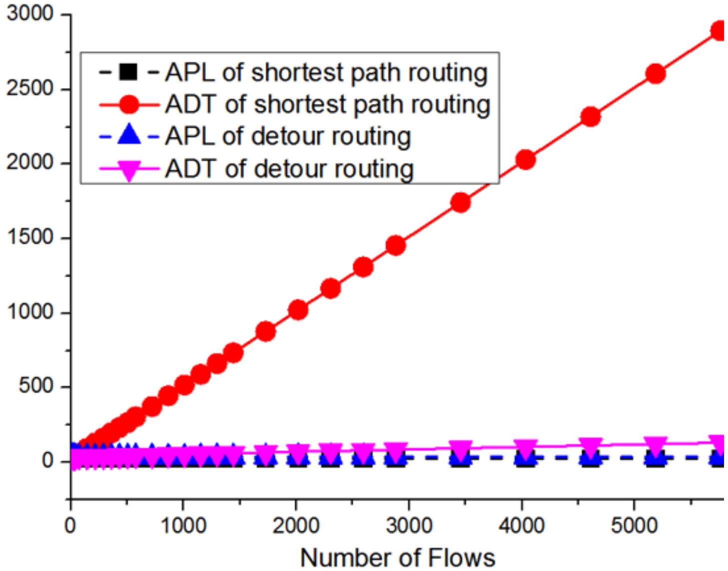
(b) Aggregate throughput.

Throughput tends to saturated when # of flows is large

Simulations

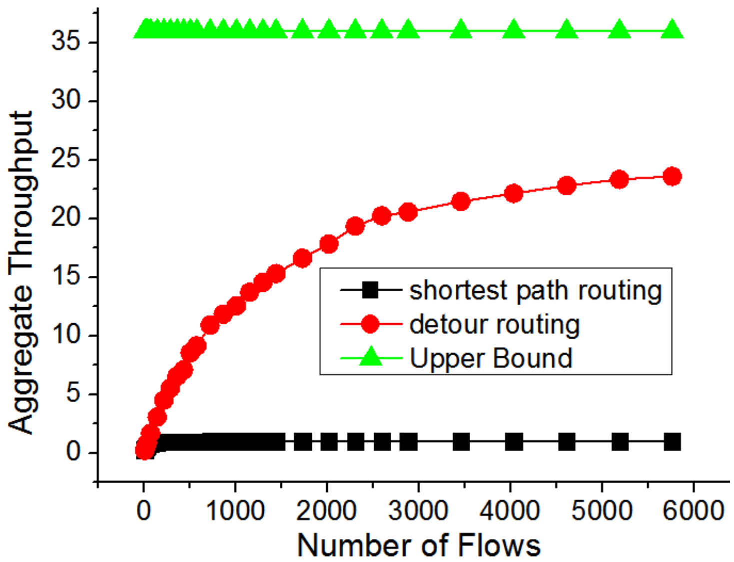
◦ Simulations for FCell

Simulations for bursty traffic: the performances of shortest path routing is poor; detour routing significantly improves the performances.



(a) APL and ADT.

ADT increases linearly with the number of flows.



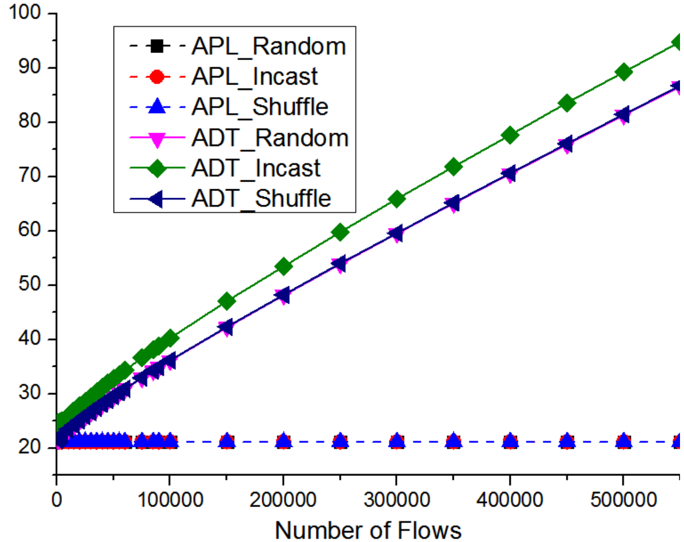
(b) Aggregate throughput.

Throughput is upper bounded by the servers' sending and receiving capability.

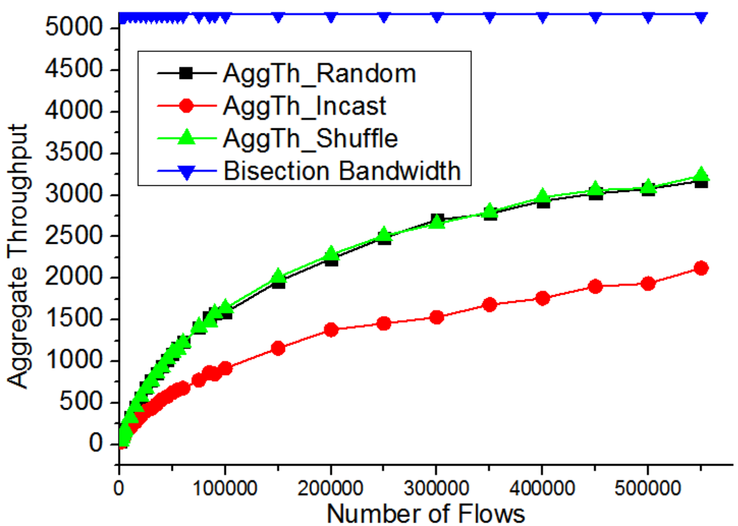
Simulations

- Simulations for FSquare

The shortest path routing demonstrates good performances under various traffic conditions.



(a) APL and ADT.



(b) Aggregate throughput.

Randon, Incast (same destination), Shuffle
ADT increases linearly with the number of flows.

Throughput tends to saturated when # of flows is large.
Incast throughput is lower because high congestion increases delivery time and thus reduces throughput.

Conclusion and Future Works

Conclusion

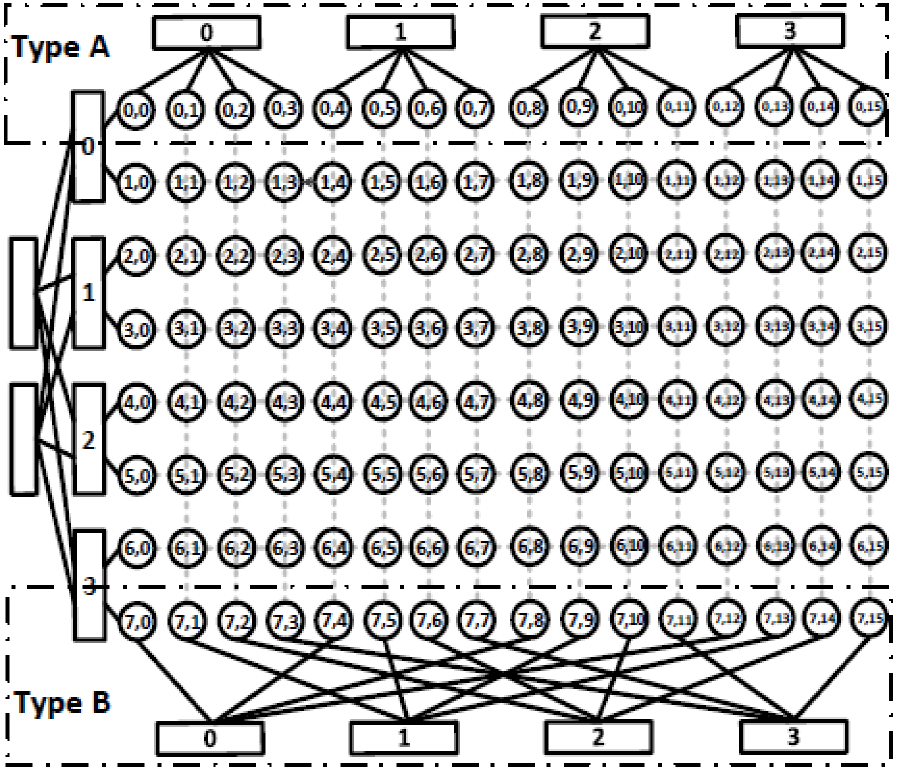
- A **unified path length definition** and a **unified power consumption model** for general DCNs
  - Enabling **fair** and **meaningful** comparisons
- A new class of DCNs, that can be regarded as **dual-centric**, with **FCell** and **FSquare** as examples.
  - Basic routing schemes
  - Performance under different traffic conditions
- **Tradeoff designs** for DCN architectures
  - Performance and power, switch-centric and server-centric designs

Conclusion and Future Works

- Future Works
  - Designing **efficient and/or adaptive** routing schemes for Fcell, Fsquare, and their extensions.
  - Exploring other possible dual-centric architectures that also have appealing properties.
  - Designing dual-centric architectures where each server uses **more than 2 NIC ports**.
  - Exploring the **limitations** of the dual-centric design philosophy, and how to **control and apply** them in practical DCN designs

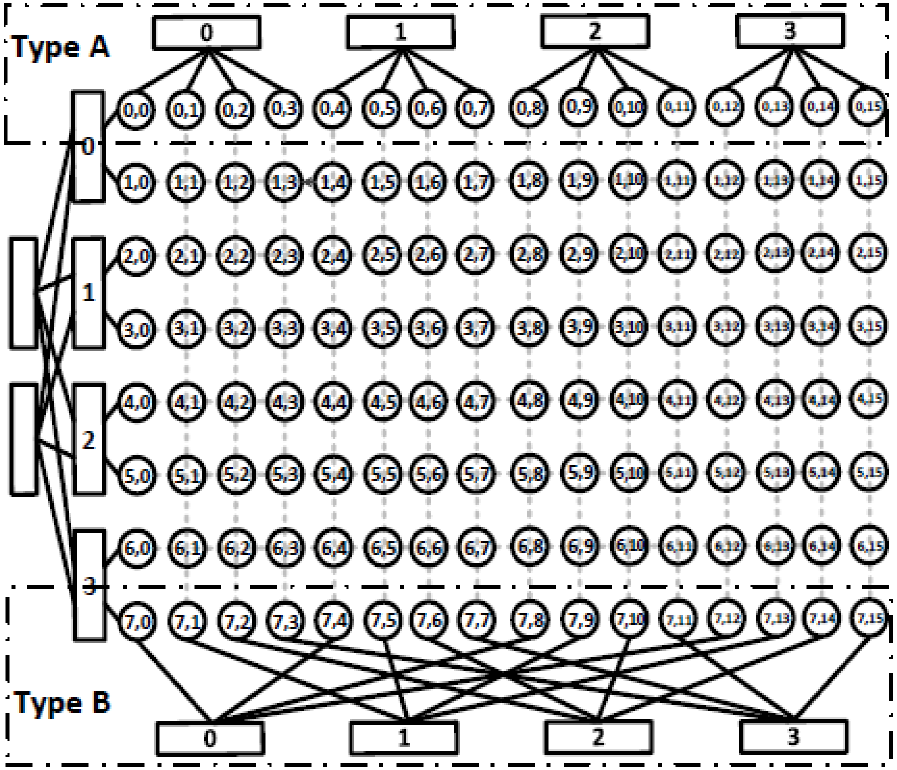
Dual-Centric DCNs: FRectangle

- Frectangle (an extension of Fsquare)
  - The switches and servers in each column form the same cluster as in FCell.
  - Switches and servers in each row can adopt Type A or Type B connections.



Dual-Centric DCNs: FRectangle

- FRectangle
  - FRectangle chooses from the 2 types of interconnections in an interleaved fashion.
  - Denote $a_{i,j}$ as the server in the *i*th row and *j*th column.
  - If i % 2 = 0, type A row.
  - If i % 2 = 1, type B row.



Dual-Centric DCNs: FRectangle

- FRectangle Basic Properties

Property 5. *In an FRectangle*(n), *the number of servers is* $N_v = n^4/2$, *and the number of switches is* $N_w = 2n^3$.

Property 6. *FRectangle*(n) *has a diameter of* $d = 6d_w + 4d_v$.

Property 7. *The bisection bandwidth of an FRectangle*(n) *is* $B = N_v/4$.

Property 8. *The DCN power consumption per server of an FRectangle*(n) *is* $p_V = 4p_w/n + 2p_{nic} + p_{fwd}$.

Comparison of DCNs with FRectangle

	$N_v(n{=}24)$	$N_v(n{=}48)$	N_w/N_v	d	B	p_V
FDCL$(n,4)$	41,472	663,552	$7/n$	$7d_w{+}d_v$	$N_v/2$	$7p_w/n + p_{nic}$
FBFLY$(4,7,3)$	49,125	—	$8/24$	$8d_w{+}d_v$	$N_v/3$	$8p_w/24 + p_{nic}$
FBFLY$(8,6,6)$	—	1,572,864	$8/48$	$7d_w{+}d_v$	$N_v/3$	$8p_w/48 + p_{nic}$
FSquare(n)	82,944	1,327,104	$6/n$	$6d_w{+}2d_v$	$N_v/2$	$6p_w/n + 2p_{nic} + p_{fwd}$
FRectangle(n)	165,888	2,654,208	$4/n$	$6d_w{+}4d_v$	$N_v/4$	$4p_w/n + 2p_{nic} + p_{fwd}$
FCell(n)	83,232	1,328,256	$3/n$	$6d_w{+}3d_v$	$N_v/4$	$3p_w/n + 2p_{nic} + p_{fwd}$
BCube$(n,3)$	331,776	5,308,416	$4/n$	$4d_w{+}4d_v$	$N_v/2$	$4p_w/n + 4p_{nic} + p_{fwd}$
SWCube$(r,4)$	28,812	685,464	$2/n$	$5d_w{+}5d_v$	$(N_v/8) \times r/(r-1)$	$2p_w/n + 2p_{nic} + p_{fwd}$
DPillar$(n,4)$	82,944	1,327,104	$2/n$	$6d_w{+}6d_v$	$N_v/4$	$2p_w/n + 2p_{nic} + p_{fwd}$
DCell$(n,2)$	360,600	5,534,256	$1/n$	$4d_w{+}7d_v$	$> N_v/(4\log_n N_v)$	$p_w/n + 3p_{nic} + p_{fwd}$
FiConn$(n,2)$	24,648	361,200	$1/n$	$4d_w{+}7d_v$	$> N_v/16$	$p_w/n + 7p_{nic}/4 + 3p_{fwd}/4$

References

Dawei Li and Jie Wu, "FCell: Towards the Tradeoffs in Designing Data Center Network Architectures, " *IEEE ICCCN*, August 3 - August 6, 2015.

Dawei Li, Jie Wu, Zhiyong Liu and Fa Zhang, "Dual-Centric Data Center Network Architectures," *ICPP*, September 1 - September 4, 2015.

Dawei Li and Jie Wu, "On the Design and Analysis of Data Center Network architectures for Interconnecting Dual-Port Servers," *IEEE INFOCOM*, April 27 - May 2, 2014.

Thank you!

Additional questions can be sent to:
jiewu@temple.edu