On the Design and Analysis of Data Center Network Architectures for Interconnecting Dual-Port Servers

Dawei Li and Prof. Jie Wu

Temple University

Outline

- Introduction
- Preliminaries
- Maximizing the number of dual-port servers given network diameter and switch port number
- SWCube
- SWKautz
- Related Existing architectures
- On the Comparison of Various Architectures
- Evaluation of SWCube and SWKautz
- Conclusion

Introduction

- The number of servers in modern and future data centers will tend to be very large.
- Challenge: how to design network architectures to interconnect large numbers of servers, and also to meet the requirements of Data Center Networks (DCN).
- Basic connections in a DCN:
  - Server-server
  - Server-switch (switch-server)
  - Switch-switch

Introduction

- DCN architecture classification: based on whether the interconnection intelligence is put on switches or servers.
  - Switch-centric
  - Server-centric
- **Server-centric**
  - More than two Network Interface Cards (NICs) are used: BCube, DCell
  - **No more than two NICs are used: FiConn, HCN&BCN, Dpillar.**

Introduction

- Main contributions:
  - We propose the concept of ***Normalized Switch Delay*** (NSD), denoted by c, to unify the design and analysis of DCNs for dual-port servers.
  - We ask the following fundamental question: what is the maximum number of dual-port servers that any architecture can accommodate at most, given network diameter d, and switch port number n? And give an upper bound on this maximum number.
  - We propose two novel DCN architectures that try to achieve this upper bound. We also show that the two proposed architectures have good properties for DCNs.

Preliminaries

- Some definitions:
  - A **hop** is a path, from one node to another node of the same kind, which consists of no other nodes of the same kind. Thus, we have **switch-to-switch** hops and **server-to-server** hops.
  - Server-to-server hops consist of **server-to-server-direct** hops and **server-to-server-via-a-switch** hops.
  - The **length** of a path between two servers is the number of server-to-server-direct hop(s), plus 1 + c times the number of server-to-server-via-switch hop(s) in the path.
  - The **distance** of two servers is the length of the shortest path between the two servers.
  - The **diameter** of a DCN architecture is the maximum distance among all pairs of servers.

Preliminaries

- A preview of the influence of NSD (c).

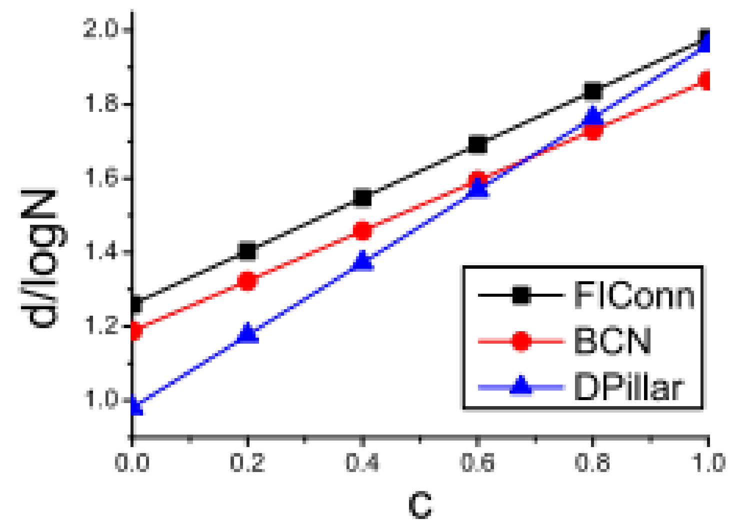


Fig. 1. Scaled diameter of FiConn, BCN, DPillar for different c values.

Maximizing the Number of Dual-Port Servers Given Network Diameter and Switch Port Number

- Background
- Moore bound: The maximum number of nodes in a graph, given diameter constraints, d ,and node degree δ is :
  - $N \le 1 + \delta + \delta(\delta-1) + \cdots + \delta(\delta-1)^{d-1} = 1 + \delta \sum_{i=0}^{d-1} (\delta-1)^{i}.$
- Illustration: any node can reach at most δ other nodes within distance 1. Each of the δ nodes can reach another δ-1 nodes within distance 2, because one degree has already been used for connecting to the original node. Extending to distance, the upper bound on the maximum number can be calculated.

Maximizing the Number of Dual-Port Servers Given Network Diameter and Switch Port Number

- Consider a DCN architecture with dual-port servers when c = 0.

Theorem 1: For $c = 0$, given switch port number n, $(n \geq 4)$, the maximum number of dual-port servers that any DCN architecture, with diameter less than or equal to d (d is a positive integer), can accommodate is: $N_v \leq N_v^{ub} = (2(n-1)^{d+1} - n)/(n-2)$.

Maximizing the Number of Dual-Port Servers Given Network Diameter and Switch Port Number

- Proof of Theorem 1

Proof: For c = 0, the lengths of a server-to-server-direct hop and a server-to-server-via-switch hop are equal. We consider the maximum number of other servers that a server S can reach within distance d. Within distance 1, S has two choices to reach other servers: the first one is to connect two other servers directly, and the second one is to connect two switches, each of which connects $n-1$ other servers, resulting in a total of $2(n-1)$ servers. Obviously, the second choice is better because S reaches more other servers, and more servers has one port remaining for further expansion. Within distance 2 of S, based on the second choice, the $2(n-1)$ servers connect to $2(n-1)$ switches, each of which connects $n-1$ other servers, resulting in another $2(n-1)^2$. Extending to distance d, S can reach at most $2(n-1)+2(n-1)^2+\cdots+2(n-1)^d$ other servers. Plus the original server S itself, the maximum number of dual-port servers that any network can accommodate is: $N_v \leq N_v^{ub} = 1 + 2(n-1) + 2(n-1)^2 + \cdots 2(n-1)^d = (2(n-1)^{d+1} - n)/(n-2)$. ■

Maximizing the Number of Dual-Port Servers Given Network Diameter and Switch Port Number

- when c != 0.

Theorem 2: Given switch port number n, $(n \geq 4)$, the maximum number of dual-port servers that any DCN architecture, with diameter less than or equal to d (d is an arbitrary positive number), can accommodate is: $N_v \leq N_v^{ub} = (2(n-1)^{\lceil d/(1+c) \rceil + 1} - n)/(n-2)$.

Maximizing the Number of Dual-Port Servers Given Network Diameter and Switch Port Number

- The upper bound may not be achievable

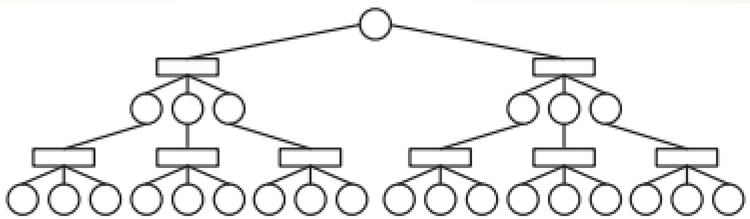


Fig. 2. The architecture greedily constructed to maximize the number of servers has a diameter $d = 2(1+c)l, l = 2$.

$$N_v \leq (2(n-1)^{\lceil d/(2(1+c)) \rceil + 1} - n)/(n-2)$$

Much less than $N_v^{ub} = (2(n-1)^{\lceil d/(1+c) \rceil + 1} - n)/(n-2)$.

Maximizing the Number of Dual-Port Servers Given Network Diameter and Switch Port Number

- In traditional graphs, a d-dimensional r-ary generalized hypercube has diameter d and network order (the number of nodes in a network) r^d.
- A Kautz graph with $r + 1$ symbols and diameter d has network order $r^d + r^{d-1}$.
- These facts motivate us to design large order DCN architectures, based on the generalized hypercube and the Kautz graph.

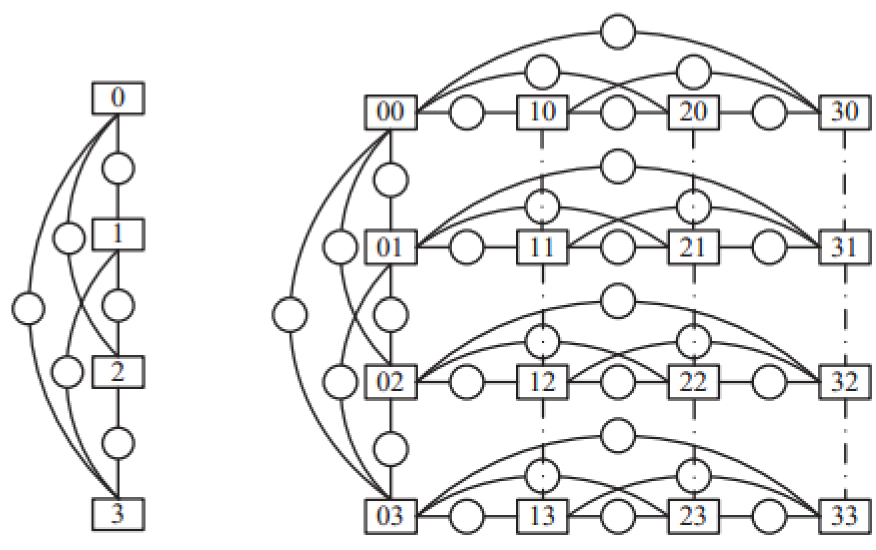
SWCube

- The Generalized Hypercube
  - A node W is represented by a k-tuple

$$w_1 w_2 \cdots w_k, \text{ where } 0 \leq w_i \leq r_i - 1, \forall i = 1, 2, \cdots, k.$$

  - Two nodes are connected directly by a link if and only if their addresses differ at one bit.
- SWCube Construction
  - 1.) replace the nodes in the original generalized hypercube with switches
  - 2.) insert one server into each link that connects two switches

SWCube



(a) 1D SWCube (b) 2D SWCube

SWCube

- Properties
  - Lemma 1: The distance of two servers that are along the same dimension is at most 2.
  - Lemma 2: The distance of two servers that are not along the same dimension is at most k + 1.
  - Theorem 3: The diameter of an SWCube(r, k) is d = k+1.
  - Theorem 4: In terms of network diameter and switch port number, the number of servers in an SWCube(r; k) is

$$N_v = \frac{kr^k(r-1)}{2} = \frac{n(\frac{n}{d-1}+1)^{d-1}}{2}.$$

TABLE I. CHOICES OF k GIVEN $n = 16$

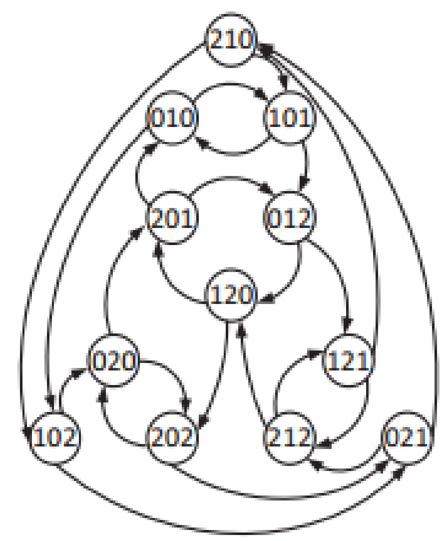
k	1	2	4	8	16
r	17	9	5	3	2
d	2	3	5	9	17
N_w	17	81	625	6561	65536
N_v	136	648	5000	52488	524288

SWKautz

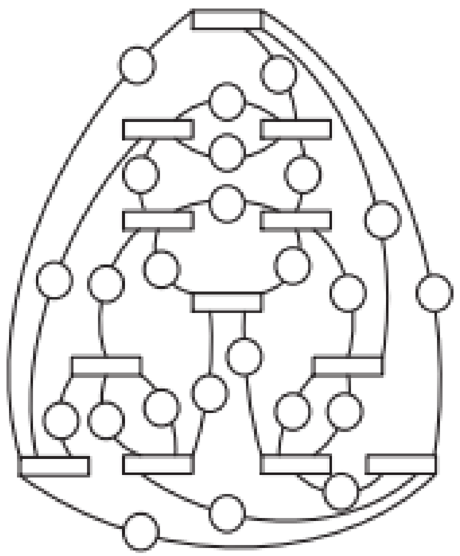
- Kautz graph: A k-dimensional Kautz directed graph with $r+1$ symbols is denoted by $KA(r,k)$.
  - The node set of $KA(r,k)$ is given by all possible strings of length k where each symbol of the string is from the set $Z = \{0,1,2,\cdots,r\}$.
  - Restriction: two consecutive symbols of the string are always different.
  - There exists a directed edge from node $W^1 = w_1^1 w_2^1 \ldots w_k^1$ to node $W^2 = w_1^2 w_2^2 \ldots w_k^2$ if and only if W^2 is a left shifted version of W^1, i.e., $w_2^1 w_3^1 \ldots w_k^1 = w_1^2 w_2^2 \ldots w_{k-1}^2$, and $w_k^2 \neq w_{k-1}^2$.

SWKautz

- SWKautz construction
  - 1.) replace each node in the original KA(n/2, k) graph with an n-port switch.
  - 2.) remove the direction of all the edges and insert a server into each edge.



(a) KA(2, 3)



(b) SWKautz(2, 3)

SWKautz

- Properties
  - Theorem 5: The diameter of an SWKautz$(n/2, k)$ is $d = k + 1$.
  - Theorem 6: In terms of network diameter and switch port number n, the number of servers in an SWKautz$(n/2, k)$ is

$$N_v = (\frac{n}{2})^d + (\frac{n}{2})^{d-1}.$$

Existing architectures

- FiConn

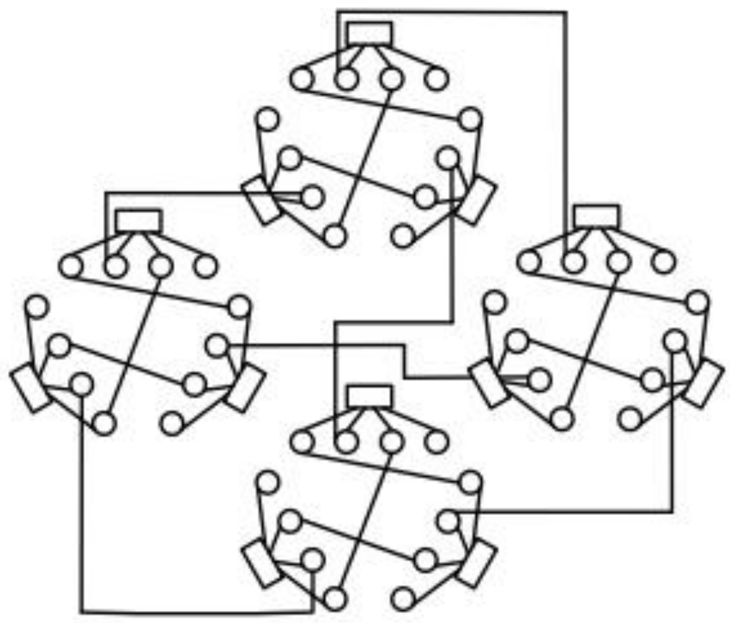
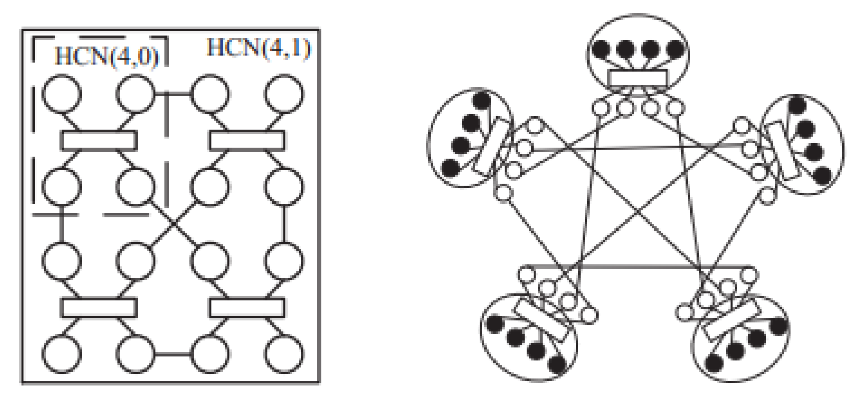


Fig. 5. FiConn$(4, 2)$.

Existing architectures

- HCN & BCN



(a) HCN (b) BCN

Fig. 6. HCN and BCN.

Existing architectures

- DPillar

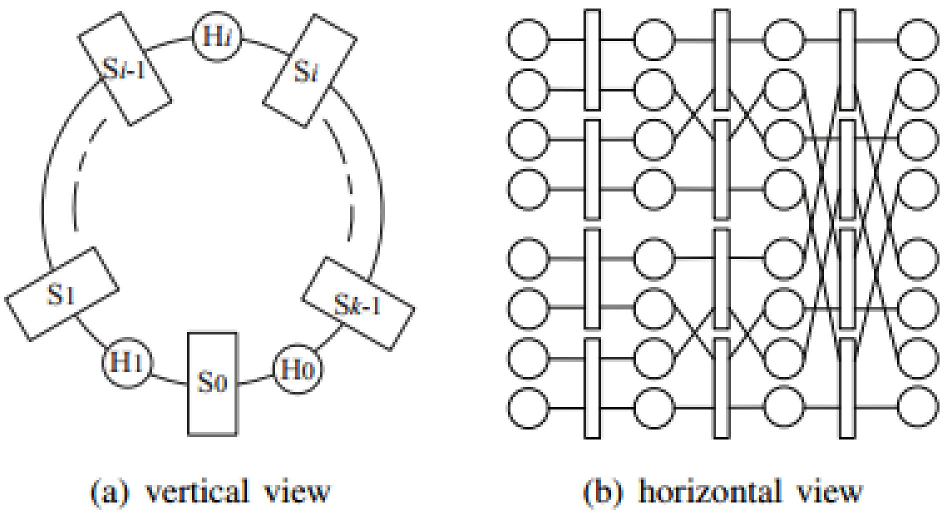


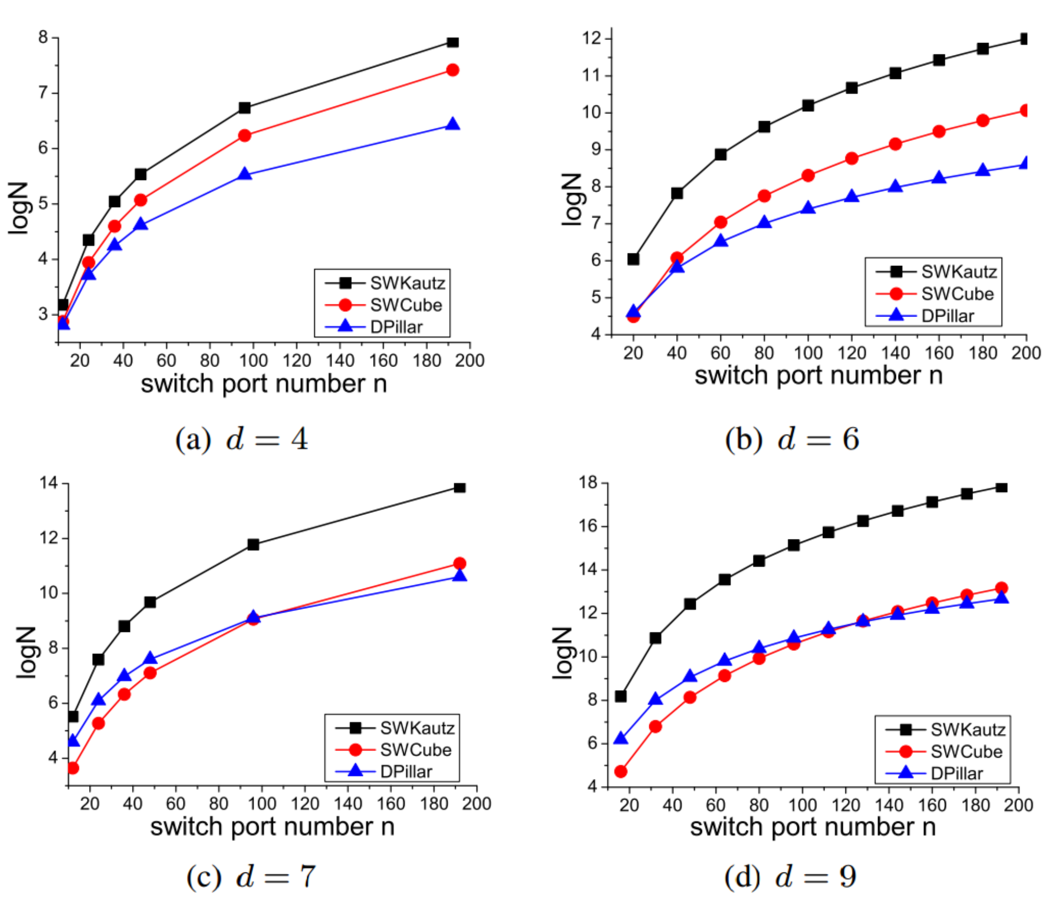
Fig. 7. DPillar.

On the Comparison of Various Architectures

- Design Flexibility
  - FiConn: $d = 2^k - 1 (k \geq 0)$
  - BCN: $d = 2^{h+1} + 2^{\gamma+1} - 1$
  - Dpillar: d = $k + \lfloor k/2 \rfloor$.
  - SWCube: $n = k(r-1) = (d-1)(r-1)$, it is only required that $d-1$ is a divisor of n.
  - SWKautz: allows the most flexible choice of network diameters because it can choose arbitrary positive integers independent of the switch port number.

On the Comparison of Various Architectures

- The Number of Servers Given d and n



(a) $d = 4$ (b) $d = 6$ (c) $d = 7$ (d) $d = 9$

On the Comparison of Various Architectures

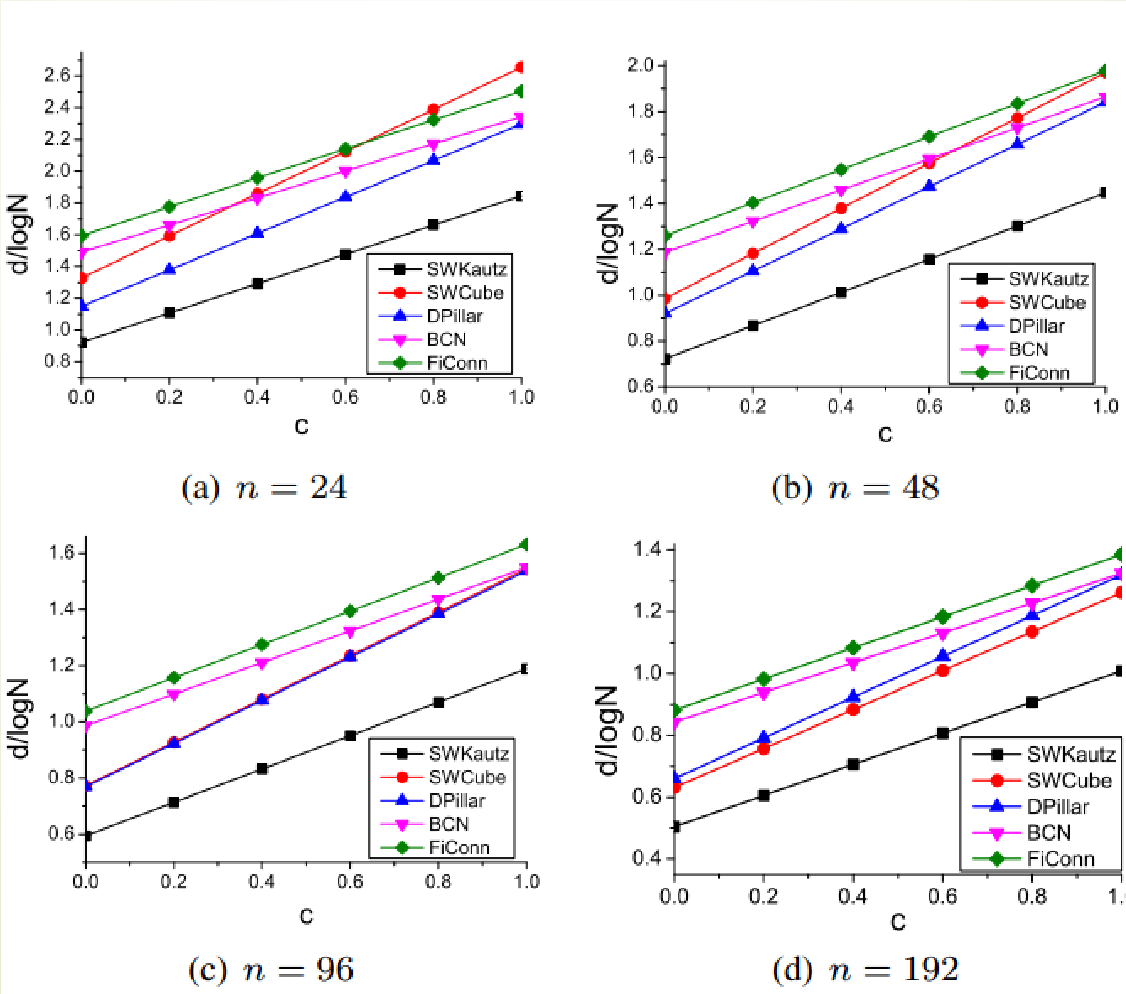
- Hardware Interconnection Cost per Server

TABLE V. HARDWARE INTERCONNECTION COST COMPARISON

	FiConn(n, k)	BCN$(\alpha, \beta, h, \gamma)$	DPillar(n, k)	SWCube(r, k)	SWKautz$(n/2, k)$
N_w/N_v	1/n	1/n	2/n	2/n	2/n
average server degree	$2 - 1/2^k$	$2 - 1/(\alpha^{h-1} n)$	2	2	2
cost per server	$P_w/n + P_l(2 - 1/2^k)$	$P_w/n + P_l(2 - 1/(\alpha^{h-1} n))$	$2P_w/n + 2P_l$	$2P_w/n + 2P_l$	$2P_w/n + 2P_l$

On the Comparison of Various Architectures

- Influence of c on Various Architectures



Evaluation of SWCube and SWKautz

- Routing Properties of SWCube and SWKautz
  - SWCube

Lemma 3: The shortest path length between two servers, $S = (S^1, S^2)$ and $D = (D^1, D^2)$, in an SWCube can be calculated by: $1 + \min\{hd(S^1, D^1), hd(S^1, D^2), hd(S^2, D^1), hd(S^2, D^2)\}$, where $hd()$ is the Hamming distance between two switches.

Theorem 7: For two servers $S = (S^1, S^2)$ and $D = (D^1, D^2)$, if their shortest path length is $l \geq 2$, their exist at least $l - 1$ server-disjoint shortest paths between them.

Evaluation of SWCube and SWKautz

- Routing Properties of SWCube and SWKautz
  - SWKautz

Theorem 8: There exist at least $n/2$ server-disjoint paths between any pair of servers in an SWKautz$(n/2, k)$, and their lengths are no greater than $k + 3$.

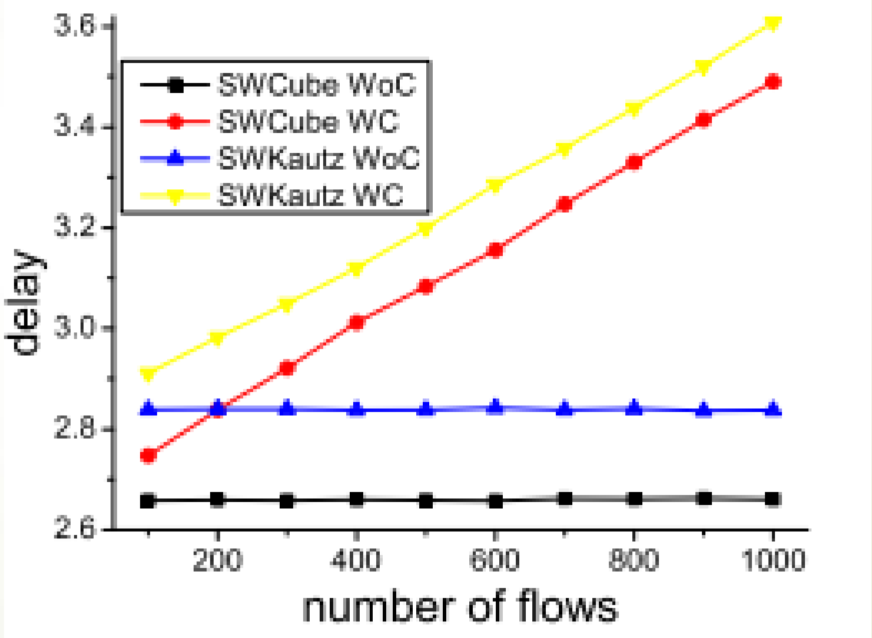
- Conclusion: both SWCube and SWKautz have good fault-tolerance properties.

Evaluation of SWCube and SWKautz

- Routing Simulation With Congestion
  - A time-step based simulation
    - All randomly generated flows are imposed on the network at the same time step, $ts = 0$.
    - each server can send at most one packet at each time step;
    - If more than one packet needs to be sent out, the packages will be queued by the First-In-First-Out (FIFO) scheme.
  - For SWCube, we adopt the shortest path for SWCube.
  - For SWKautz, we choose the long path routing algorithm.

Evaluation of SWCube and SWKautz

- Routing Simulation With Congestion
  - SWCube(13,2)
    - 2028 servers
  - SWKautz(12,2)
    - 1872 servers



- Conclusion: Both SWCube and SWKautz have the capability of efficiently handling network congestion.

Conclusion

- We propose the concept of Normalized Switch Delay (NSD), denoted by ***c***, to unify the design and analysis of DCNs for dual-port servers.
- We ask the following fundamental question: what is the maximum number of dual-port servers that any architecture can accommodate at most, given network diameter ***d***, and switch port number ***n***? And give an upper bound on this maximum number.
- We propose two novel DCN architectures that try to achieve this upper bound. Comparisons with the existing one demonstrate various advantages. Evaluations on themselves show they have good properties for DCNs.

The End!
Thank you for your attention!

Questions?

dawei.li@temple.edu