Designing Buffer Capacity of Crosspoint-Queued Switch

Guo Chen, Dan Pei, Youjian Zhao, and Yongqian Sun

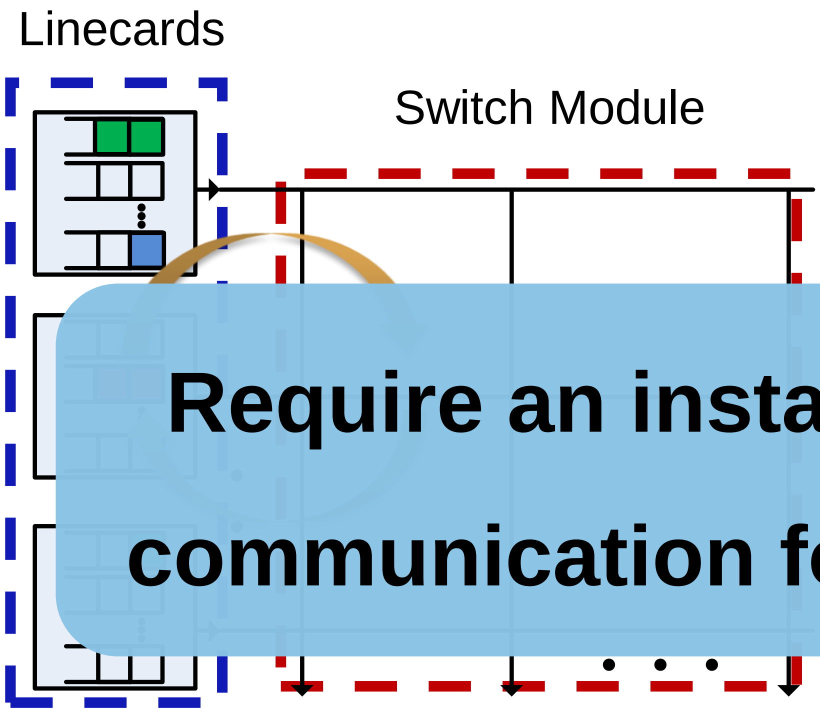
Tsinghua University

Designing Buffer Capacity of Crosspoint-Queued Switch

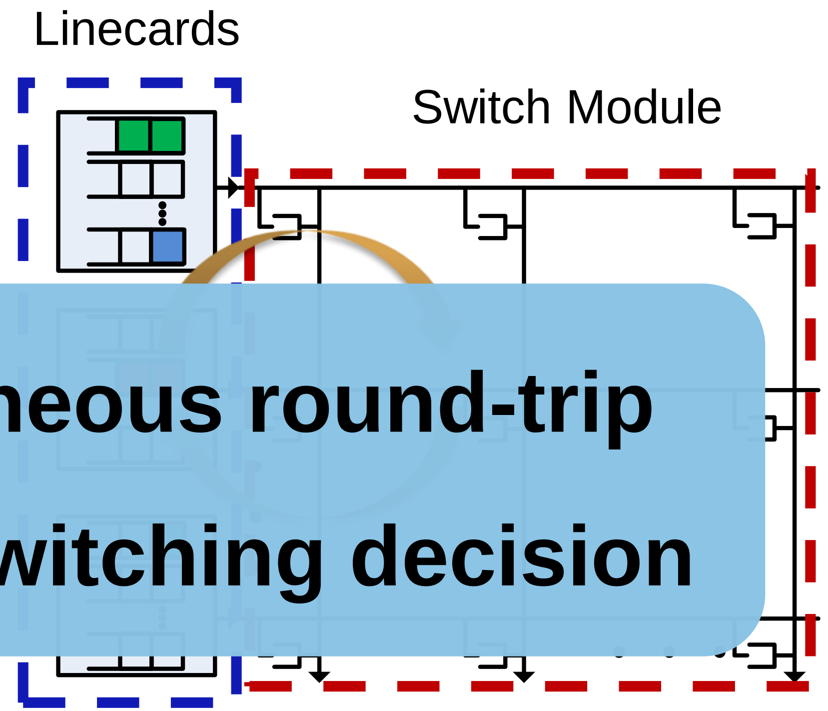
Outline

CQ Switch — Buffer Capacity V.S. Performance — Evaluation

Typical Switch Fabric Architectures



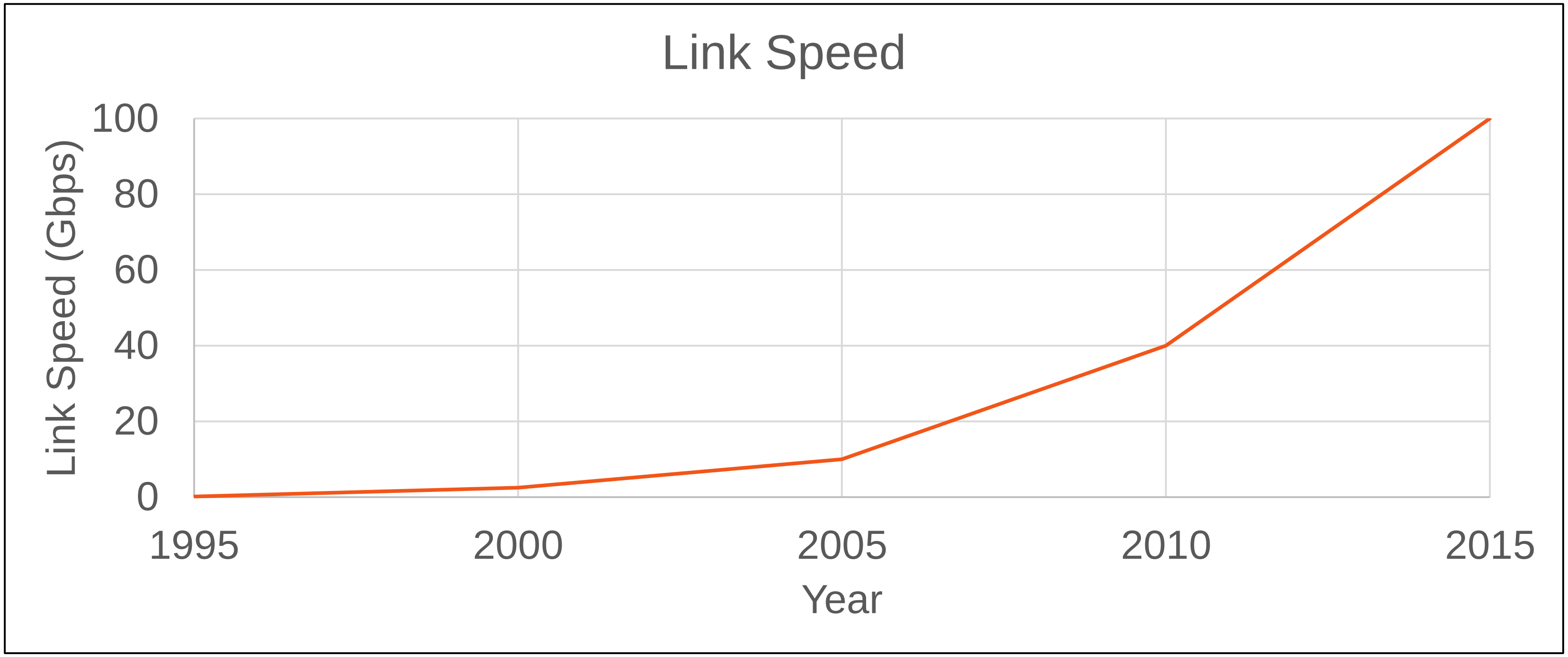
Input Queued (IQ)



Combined Input and Crosspoint Queued (CICQ)

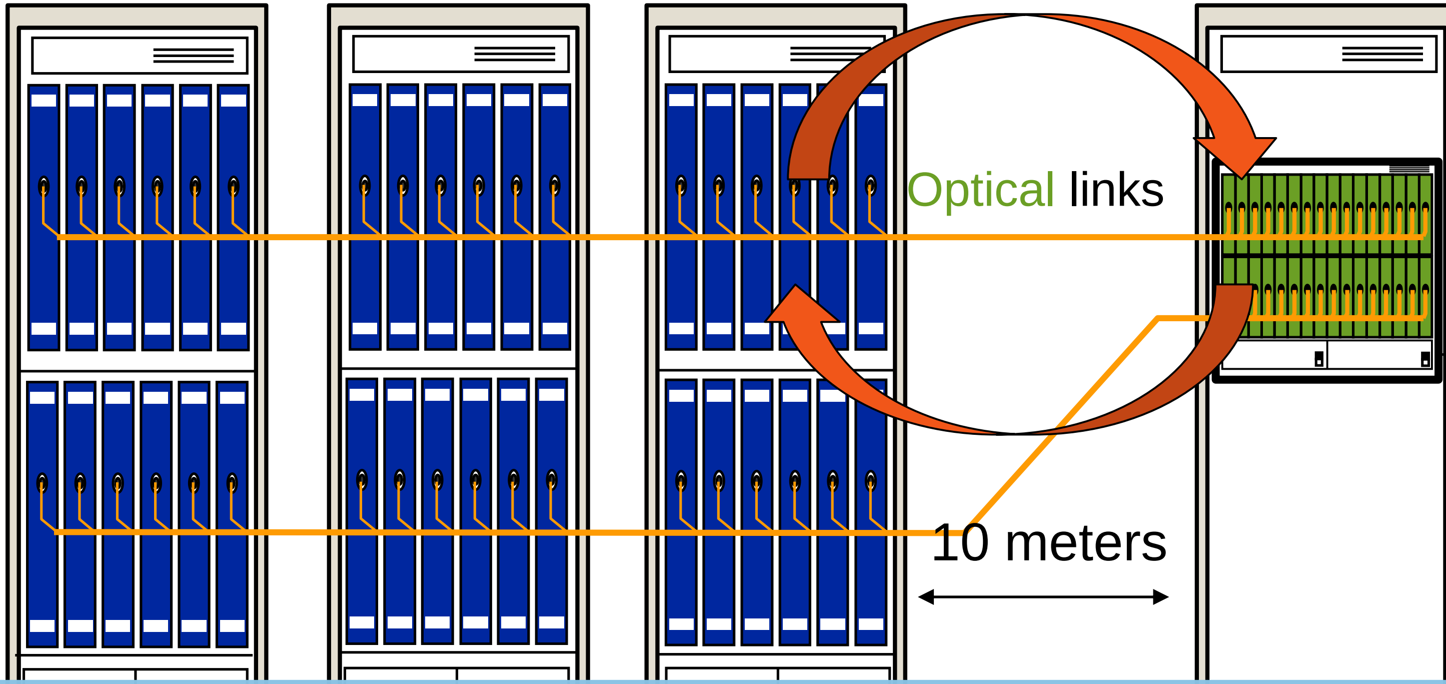
Require an instantaneous round-trip communication for switching decision

Ultra-high link speed



Only 5.12ns to make switching decision for a 64B packet in 100G routers

Linecards and switch module in different racks



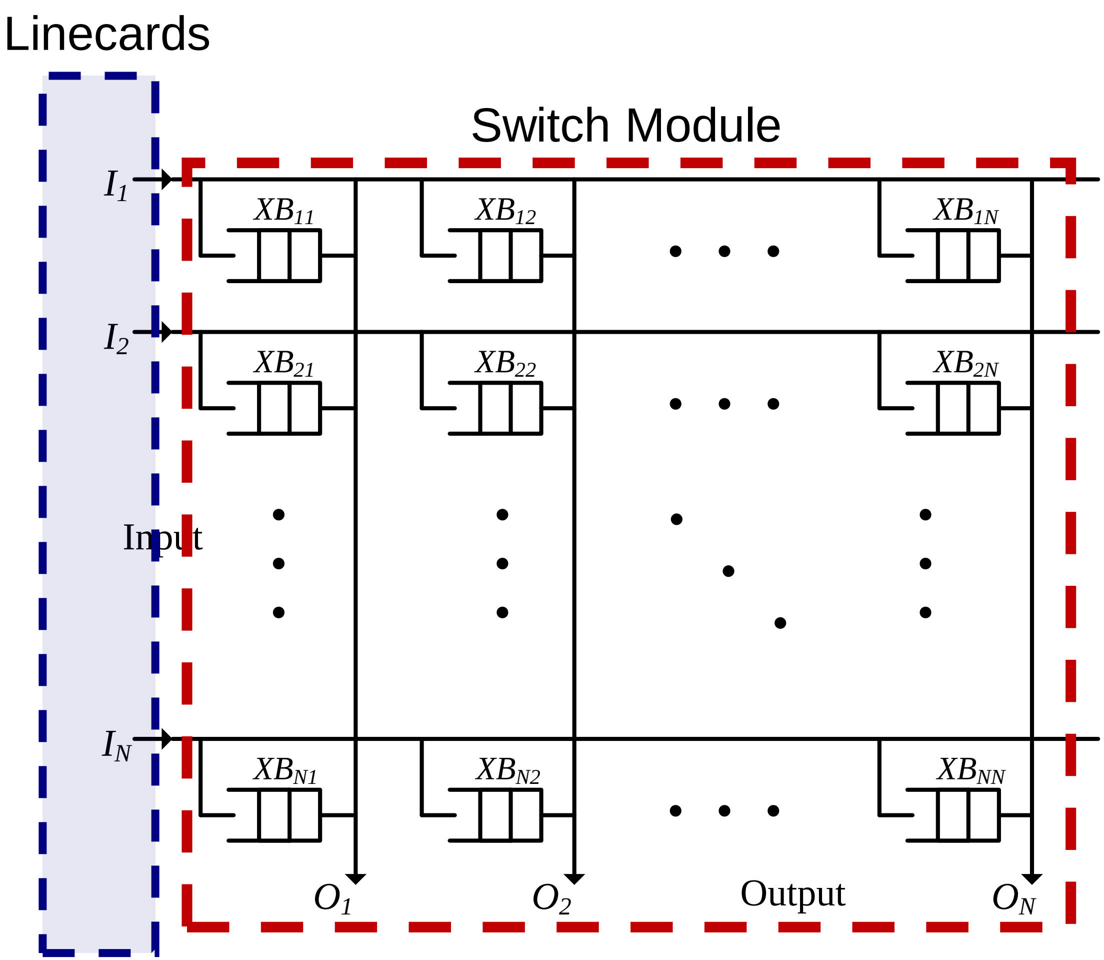
~100ns for round-trip communication in 10m link **(~2×10^8m/s propagation speed)**

Solution:

Self-sufficient switch fabric with no need of instantaneous communication between linecards and switch module

Crosspoint-Queued (CQ) Switch

- **No buffer at linecards**
- **Buffering only inside the switch Module**
- **Independent output schedulers**
- **Drops with full buffers**



- **But how to design the crosspoint buffers' size to meet performance requirement?**

- Our contribution
  - Study the different **buffer capacity's influence** to the CQ switch fabric's **performance** (throughput & delay)

Designing Buffer Capacity of Crosspoint-Queued Switch

Outline

CQ Switch — Buffer Capacity V.S. Performance — Evaluation

- Discrete-time Quasi-birth-death process

a_{ij}^{k}: Probability of cells arrived at in a given time slot. k=0,1.

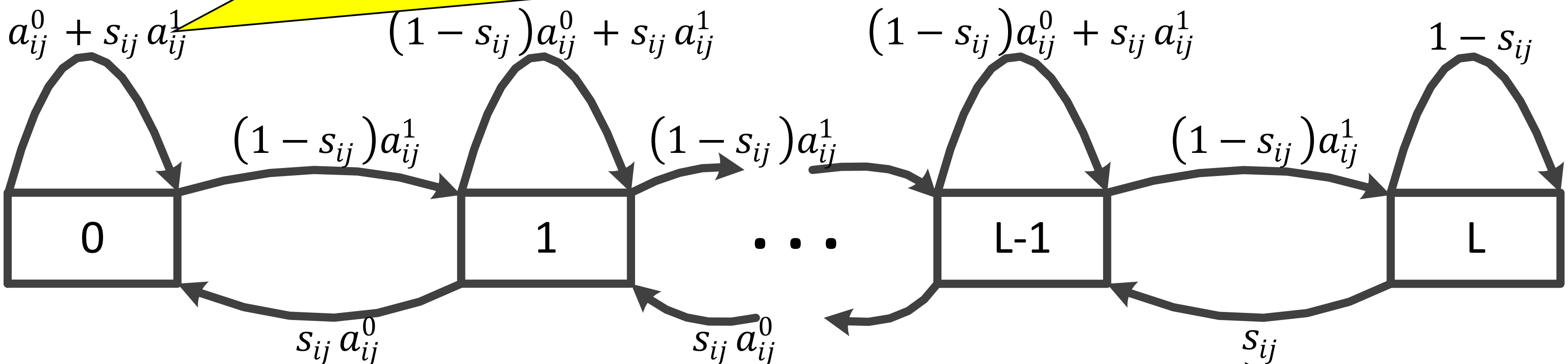


Fig. The Quasi-birth-death state transition diagram for XB_{ij}'s queue length

S_{ij}: Probability of crosspoint buffer XB_{ij} being selected by output O_j.

- Assumption
  - Independent Bernoulli traffic (Bernoulli parameters & destination distribution known)
  - Static non-work-conserving random scheduling algorithm (known)
  - NxN switch

- Closed-form throughput calculation formula

$$TP = 1 - LR = 1 - \frac{\sum_{i=1}^{N} \rho_i \left(\sum_{j=1}^{N} d_{ij} \eta_{ij}^{L} \right)}{\sum^{N} \rho_i}$$

ρ_i: Bernoulli parameter of the cell arri

…ll arrived … ut O_j.

η_{ij}^{L}: Steady-state probability of XB_{ij}'s length equals l.

$$\eta_{ij}^{0} = \frac{1}{1 + \sum_{l=1}^{L-1} \left(\frac{(1-s_{ij})a_{ij}^{1}}{s_{ij}a_{ij}^{0}} \right)^{l} + a_{ij}^{0} \left(\frac{(1-s_{ij})a_{ij}^{1}}{s_{ij}a_{ij}^{0}} \right)^{L}}$$

$$\eta_{ij}^{l} = \eta_{ij}^{0} \left(\frac{(1-s_{ij})a_{ij}^{1}}{s_{ij}a_{ij}^{0}} \right)^{l}, \quad l = 1, \ldots, L-1$$

$$\eta_{ij}^{L} = \eta_{ij}^{0} a_{ij}^{0} \left(\frac{(1-s_{ij})a_{ij}^{1}}{s_{ij}a_{ij}^{0}} \right)^{L}$$

Delay Analysis

- Non-closed-form but convergent average delay calculation formula

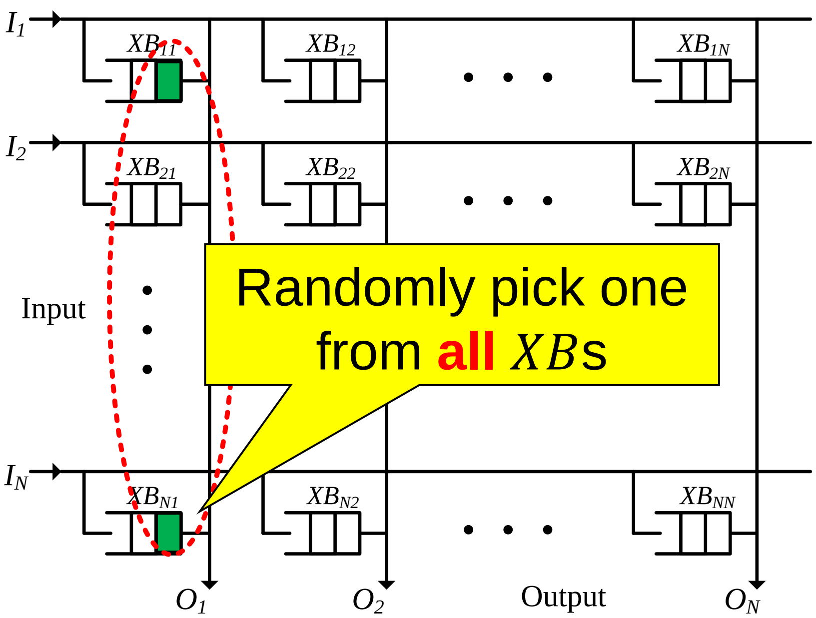
$$DL = \frac{\sum_{i=1}^{N} \rho_i E\{W_i\}}{\left(\sum_{i=1}^{N} \rho_i\right)}$$

W_i: Time slots a cell spent in input I_i.

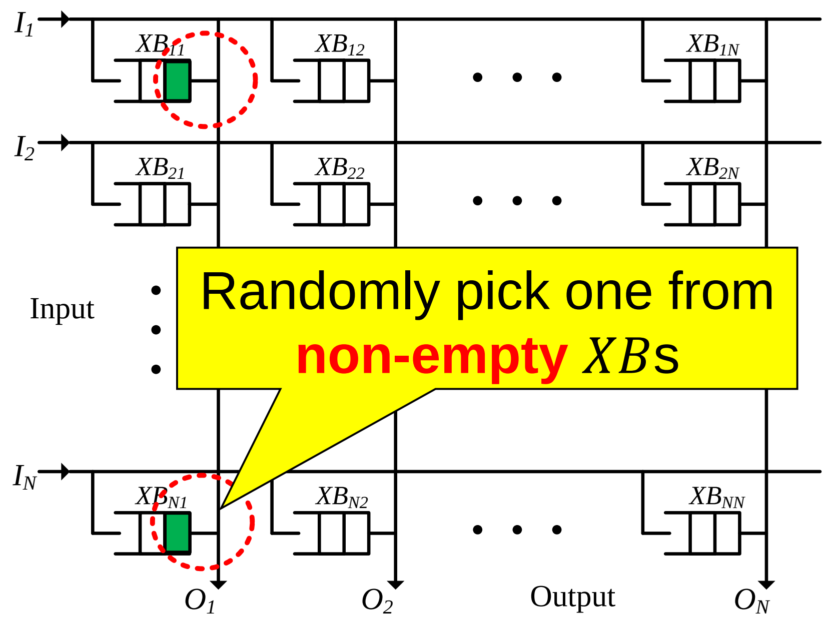
$$E\{W_i\} = \sum_{j=1}^{N} d_{ij} E\{W_{ij}\} \qquad E\{W_{ij}\} = \sum_{n=0}^{\infty} n P\{W_{ij} = n\}$$

$$P\{W_{ij} = n\} = \begin{cases} \eta_{ij}^{0} s_{ij} \left(\frac{1-s_{ij}}{a_{ij}^{0}}\right)^{n}, & 0 \le n \le L-1 \\ \eta_{ij}^{0} s_{ij} (1-s_{ij})^{n} \sum_{l=0}^{L-1} \left[C_{n}^{n-l} \left(\frac{a_{ij}^{1}}{a_{ij}^{0}}\right)^{l}\right], & n > L-1 \end{cases}$$

Lower Bound for Work-conserving Scheduling Algorithms



nWCRand



WCRand

- **Theorem.** *Under same independent Bernoulli traffic, a CQ switch using work-conserving random (WCRand) scheduling algorithm has a higher throughput and lower average delay than using non-work-conserving (nWCRand) fair random scheduling algorithm.*

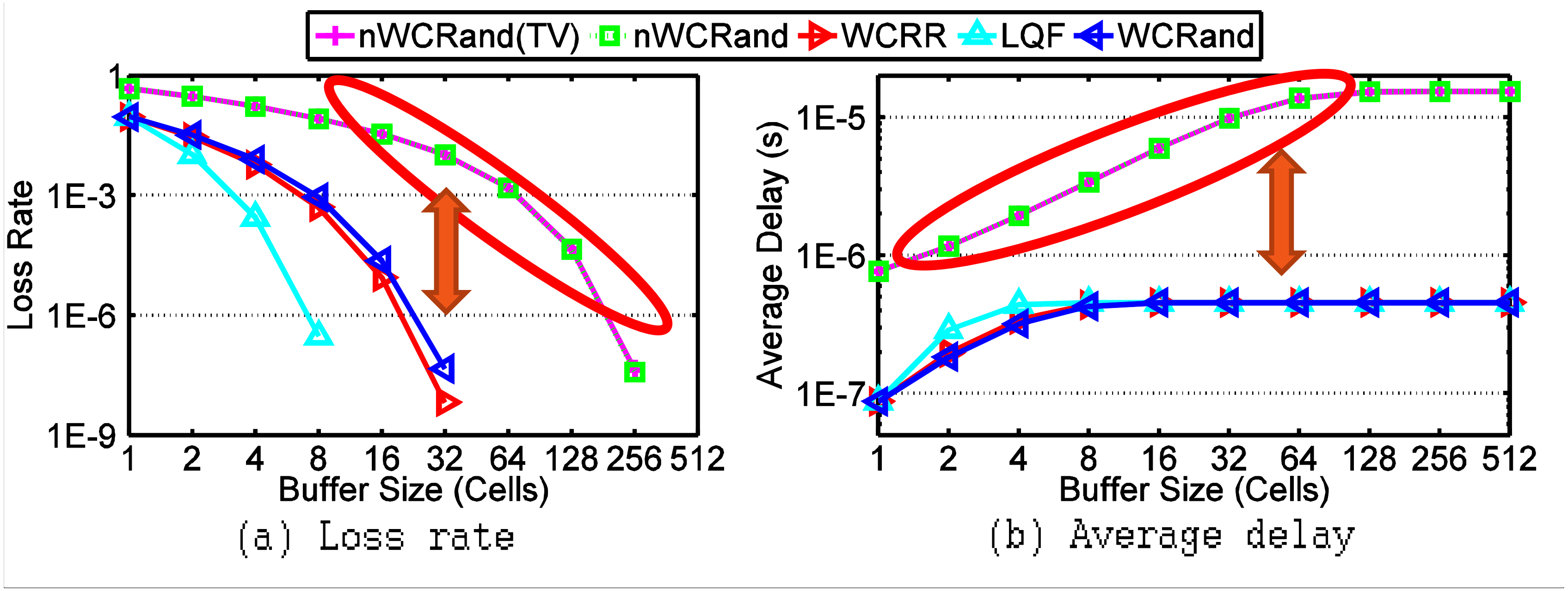
Designing Buffer Capacity of Crosspoint-Queued Switch

Outline

CQ Switch — Buffer Capacity V.S. Performance — Evaluation

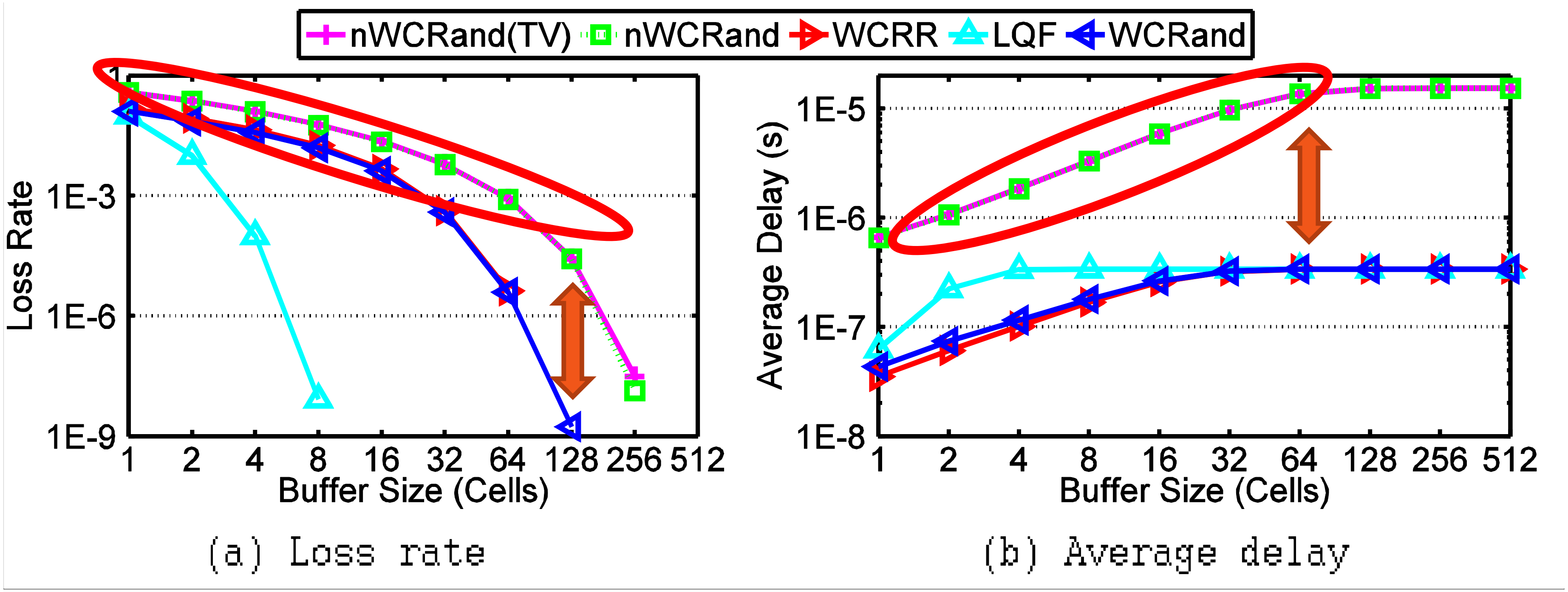
Verification of Theoretical Analysis

uniform Bernoulli traffic



Loss rate and average delay of a 16×16 CQ switch under uniform Bernoulli traffic with ρ=0.95

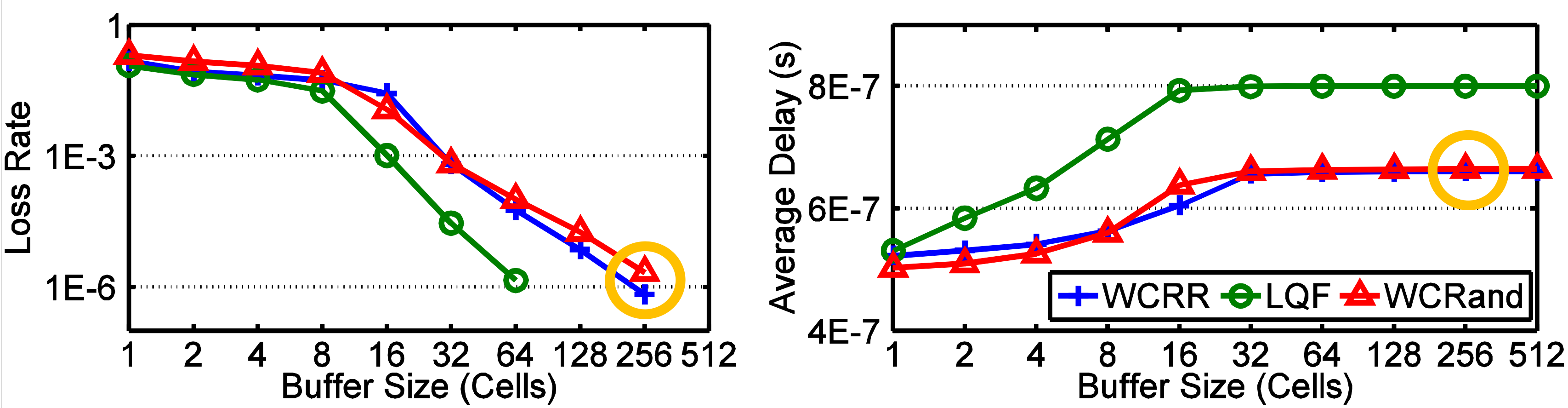
Non-uniform Bernoulli traffic



Loss rate and average delay of a 16×16 CQ switch under non-uniform Bernoulli traffic with ρ=0.95 and ω=0.5

Simulations under Real-Trace

- Data sets
  - From CAIDA
  - 1-minute traces from 10Gbps links in San Jose



(a) Loss rate (b) Average delay

A simple Round-robin or Random scheduling is able to reach a very good performance with feasible buffer size

Conclusions

- Reveals the impact of buffer size on CQ switches performance
- Provides a theoretical guidance on designing the buffer size
- CQ shows good performance under real traces

Thanks!