

# Configuring a Load-Balanced Switch in Hardware



---

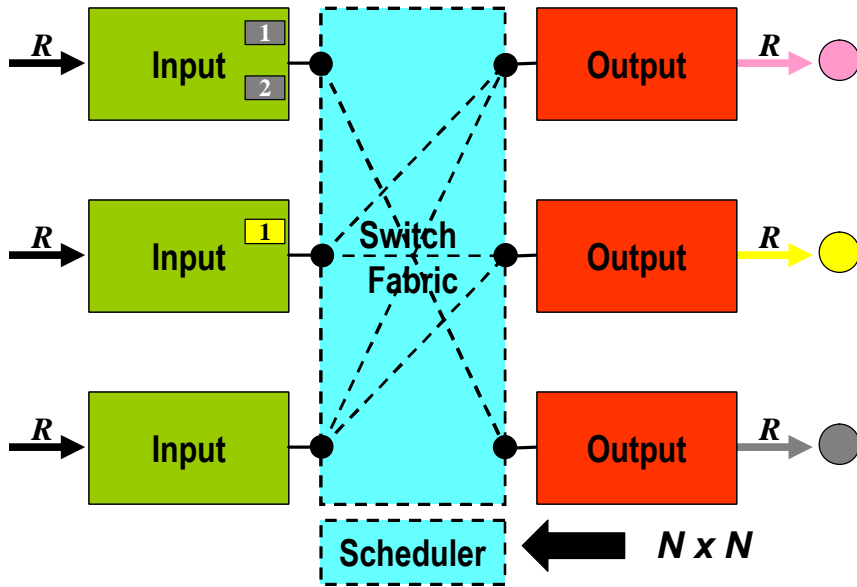
**Srikanth Arekapudi**, Shang-Tse (Da) Chuang,  
Isaac Keslassy, Nick McKeown

**Stanford University**

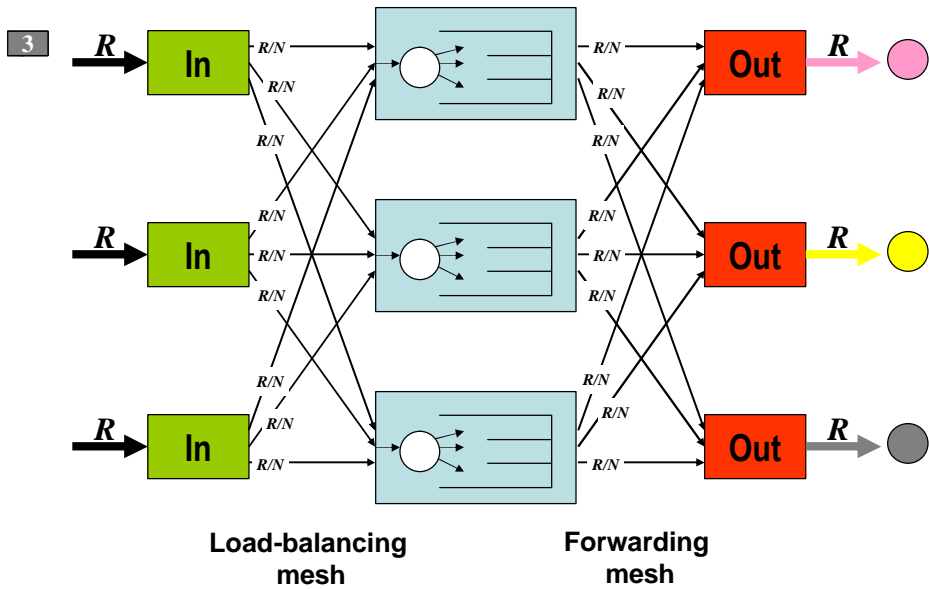
## Outline

- Load Balanced Switch
- Scalability
- Reconfiguration Algorithm
- Hardware Implementation

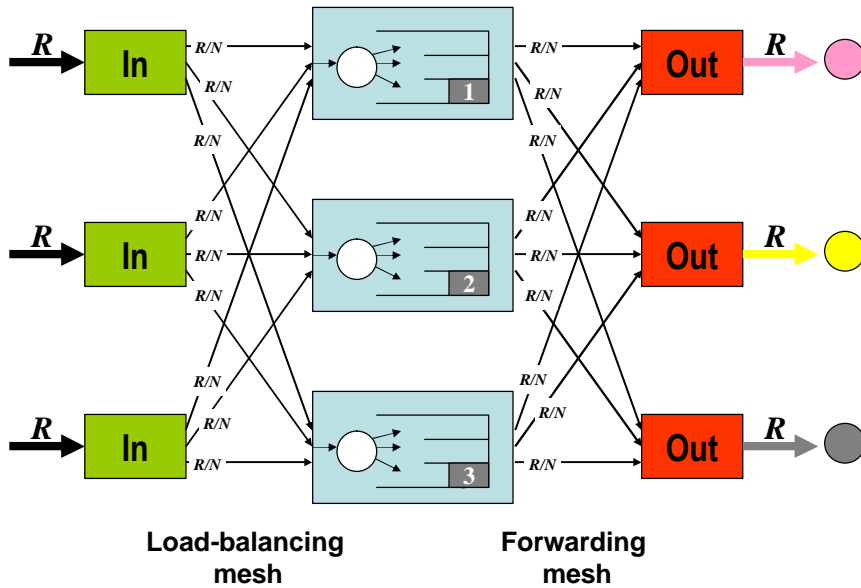
## Typical Router Architecture



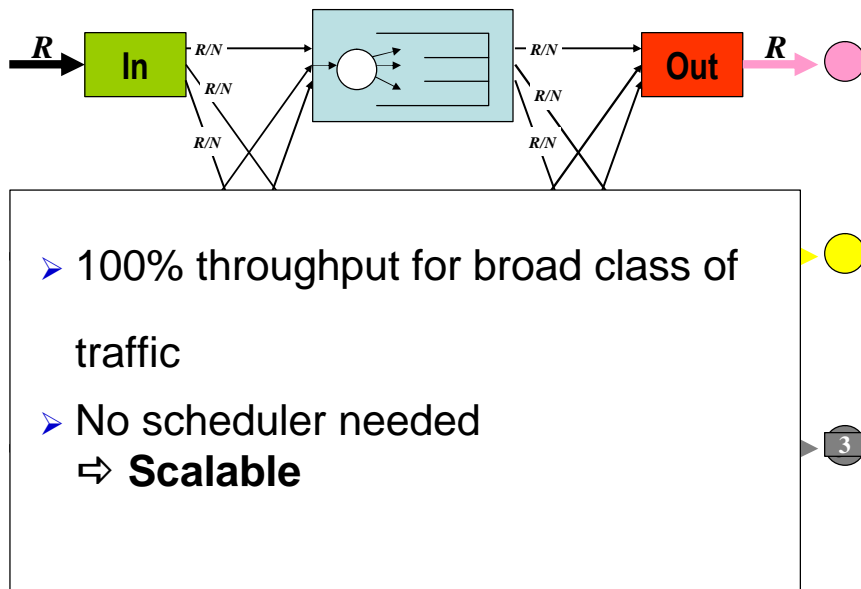
## Load-Balanced Switch



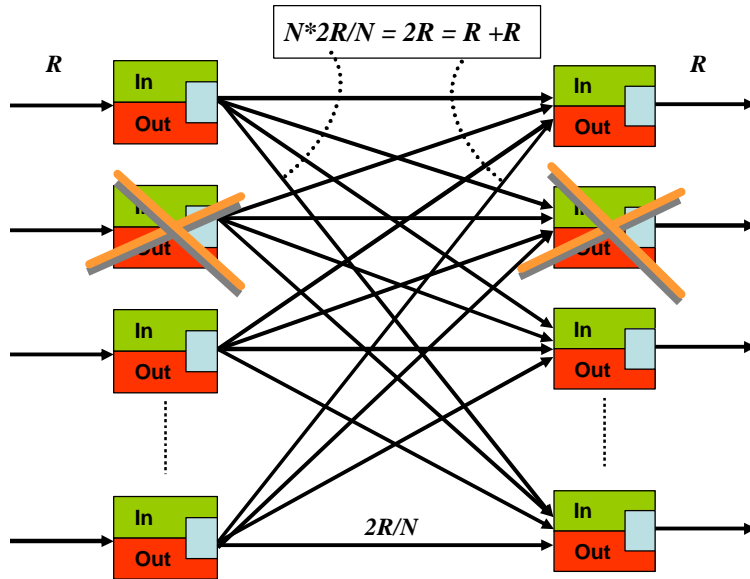
## Load-Balanced Switch



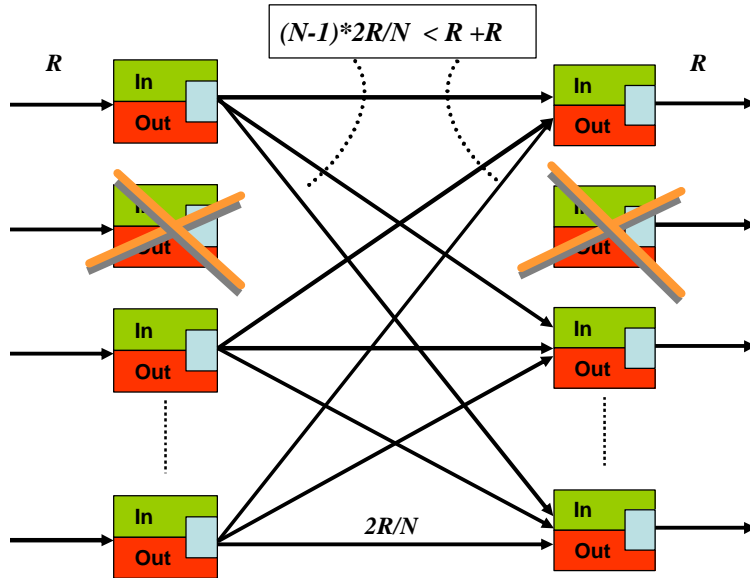
## Load-Balanced Switch



## A Single Combined Mesh

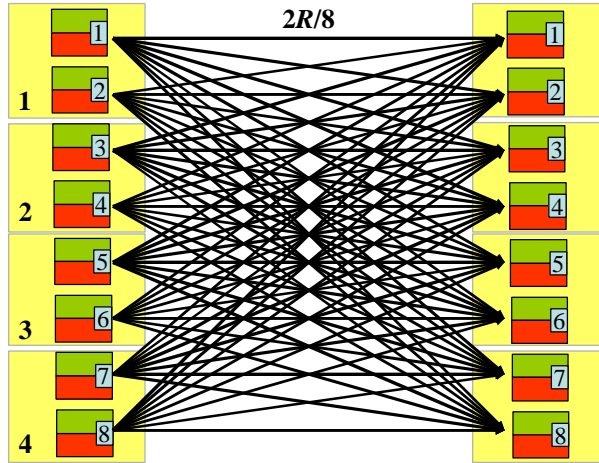


## A Single Combined Mesh



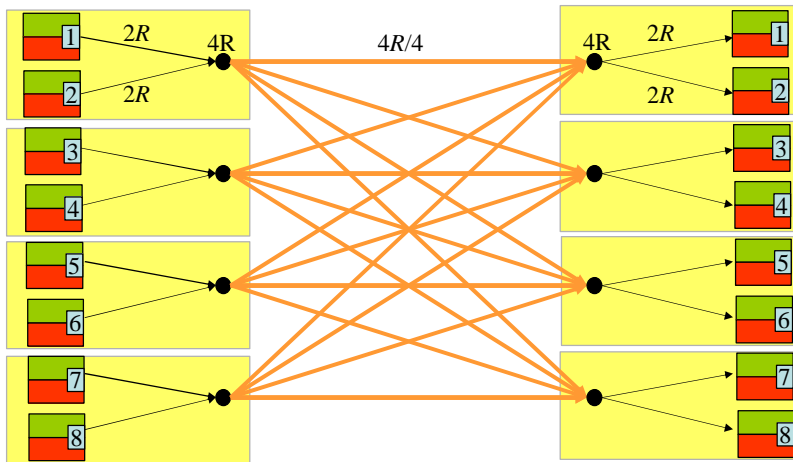
# Scalability

$N=8$



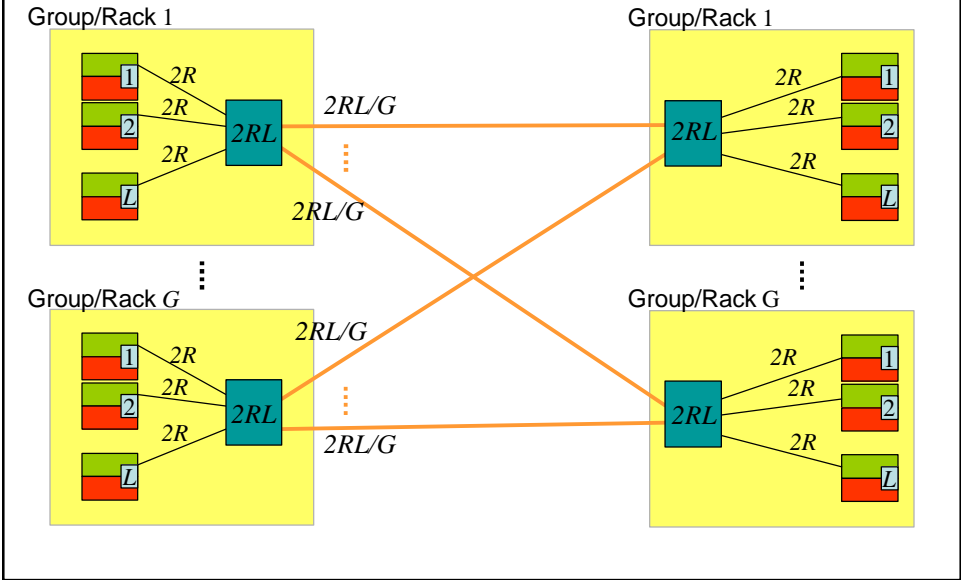
# When $N$ is Too Large

*Decompose into groups (or racks)*



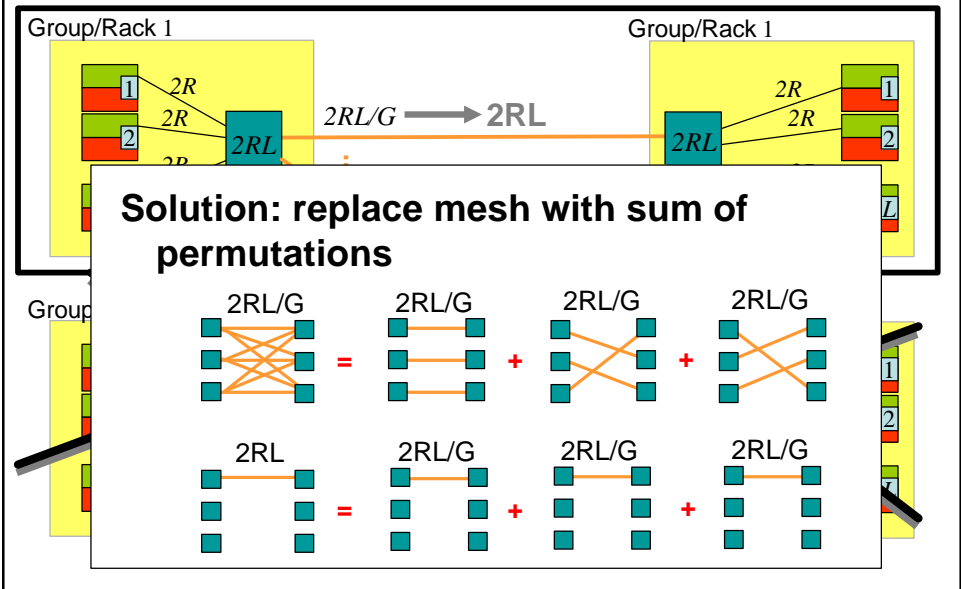
# When $N$ is Too Large

*Decompose into groups (or racks)*

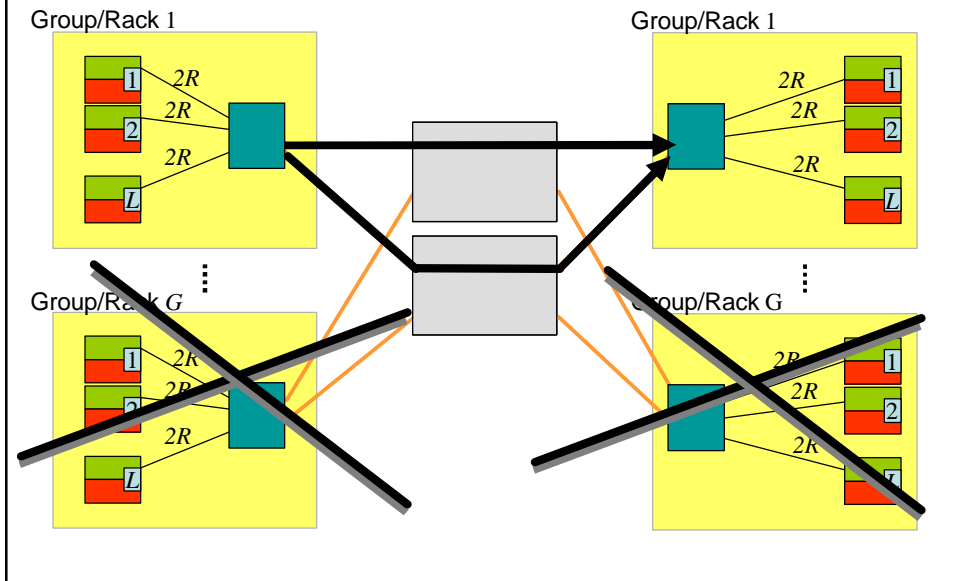


# When Linecards are Missing

*Failures, Incremental Additions, and Removals...*



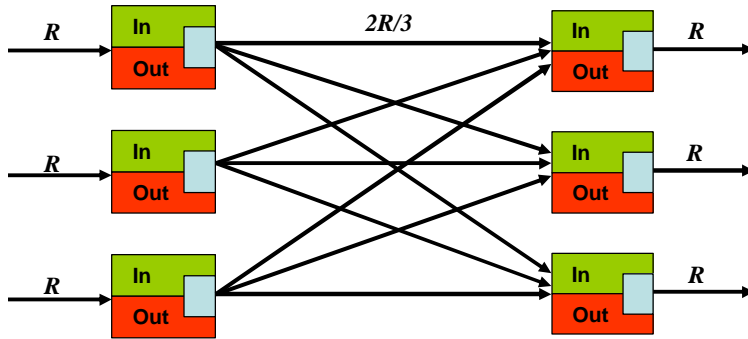
## When Linecards Fail



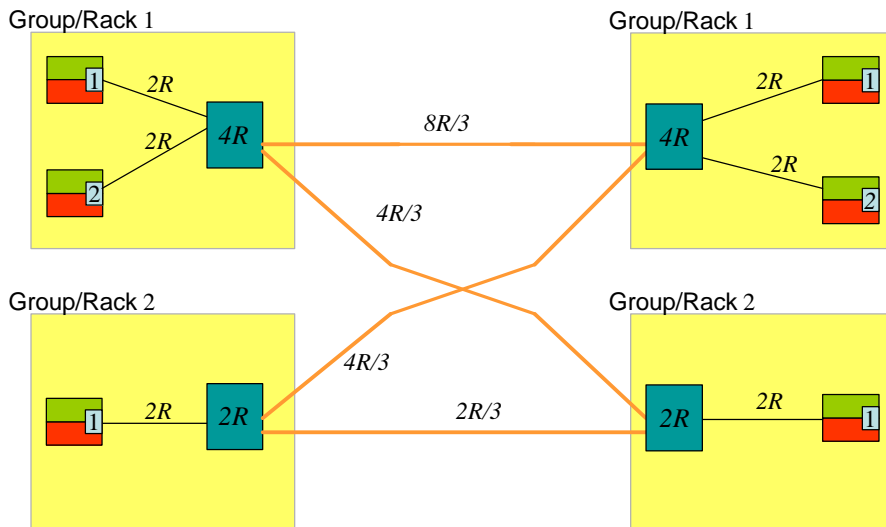
## Questions

- Number of MEMS Switches?
- TDM Schedule?

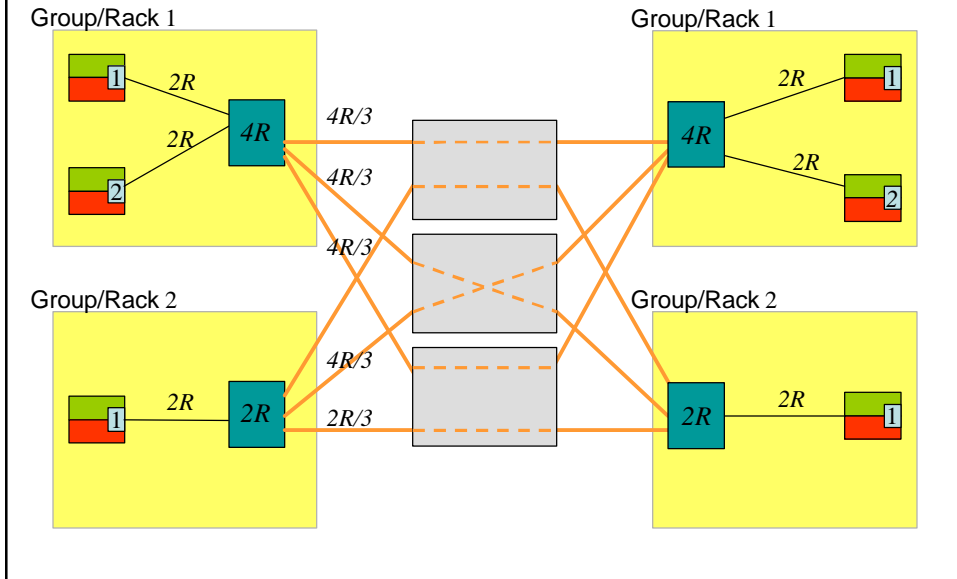
## Example – 3 Linecards



## Example 2 Groups



## Example 2 Groups



## Number of MEMS Switches

- **MEMS switches between groups  $i$  and  $j$**

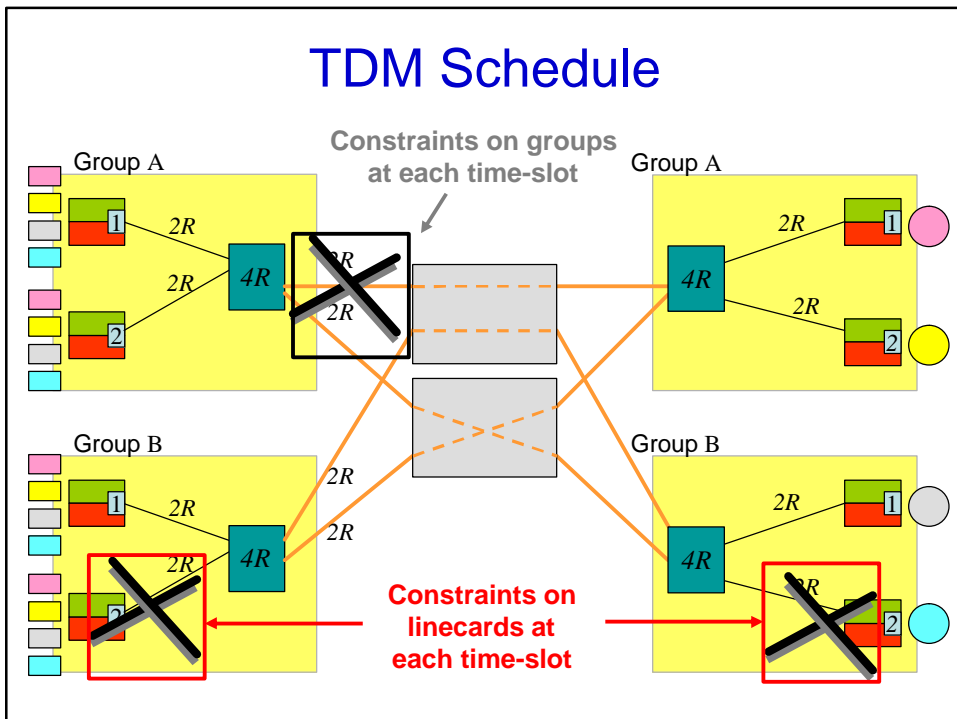
$$\left\lceil \frac{L_i L_j}{N} \right\rceil$$

- **Total Number of MEMS switches:  $M = L + G - 1$**

# Questions

➤ Number of MEMS Switches?

➔ TDM Schedule?



## Rules for TDM Schedule

At each time-slot:

- Each transmitting linecard sends one packet
- Each receiving linecard receives one packet
- (MEMS constraint) Each transmitting group  $i$  sends at most one packet to each receiving group  $j$  through each MEMS connecting them

In a schedule of  $N$  time-slots:

- Each transmitting linecard sends exactly one packet to each receiving linecard

## TDM Schedule

		$T+1$	$T+2$	$T+3$	$T+4$
Tx Group A	Tx LC A1	?	?	?	?
	Tx LC A2	?	?	?	?
Tx Group B	Tx LC B1	?	?	?	?
	Tx LC B2	?	?	?	?

## TDM Schedule

	$T+1$	$T+2$	$T+3$	$T+4$	
Tx Group A	Tx LC A1	A1	A2	B1	B2
	Tx LC A2	B2	A1	A2	B1
Tx Group B	Tx LC B1	B1	B2	A1	A2
	Tx LC B2	A2	B1	B2	A1

## Bad TDM Schedule

	$T+1$	$T+2$	$T+3$	$T+4$	
Tx Group A	Tx LC A1	A1	<del>A2</del>	B1	B2
	Tx LC A2	B2	<del>A1</del>	A2	B1
Tx Group B	Tx LC B1	B1	B2	A1	A2
	Tx LC B2	A2	B1	B2	A1

## TDM Schedule Algorithm

- The algorithm constructs three consecutive schedules.
  1. **Sending Groups to Receiving Groups**
    - Connection Assignment Problem
  2. **Sending Linecards to Receiving Groups.**
    - Matrix Decomposition Problem
  3. **Sending Linecards to Receiving Linecards**
    - Matrix Decomposition Problem

## Group to Group Schedule

	$T+1$	$T+2$	$T+3$	$T+4$
Tx Group A	AB	AB	AB	AB
Tx Group B	AB	AB	AB	AB

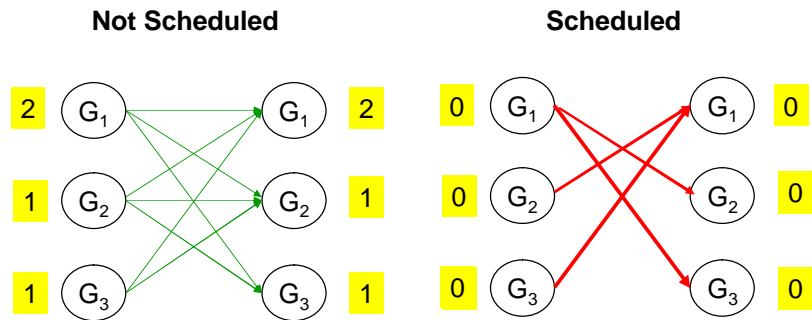
## Linecard to Group Schedule

		$T+1$	$T+2$	$T+3$	$T+4$
Tx Group A	Tx LC A1	A	A	B	B
	Tx LC A2	B	B	A	A
Tx Group B	Tx LC B1	B	B	A	A
	Tx LC B2	A	A	B	B

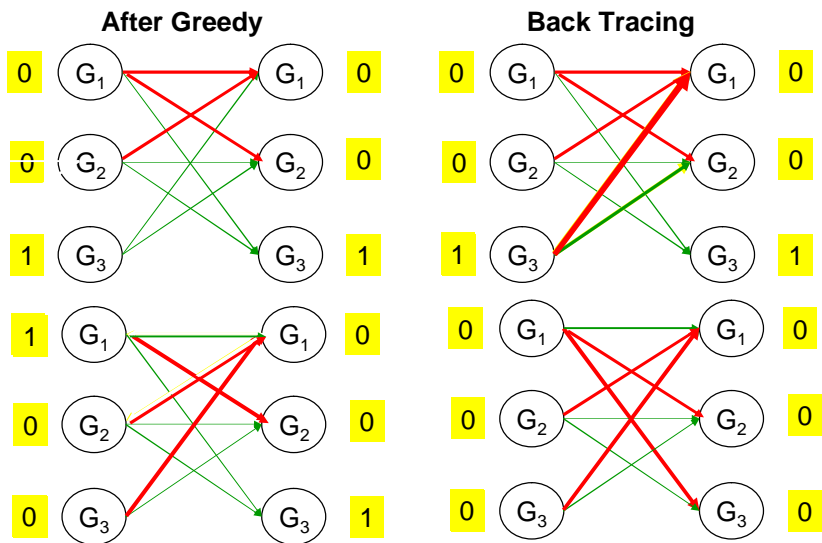
## Linecard to Linecard Schedule

		$T+1$	$T+2$	$T+3$	$T+4$
Tx Group A	Tx LC A1	A1	A2	B1	B2
	Tx LC A2	B2	B1	A2	A1
Tx Group B	Tx LC B1	B1	B2	A1	A2
	Tx LC B2	A2	A1	B2	B1

# Connection Assignment Problem



# Connection Assignment Problem



## Matrix Decomposition Problem

$$\begin{pmatrix} 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

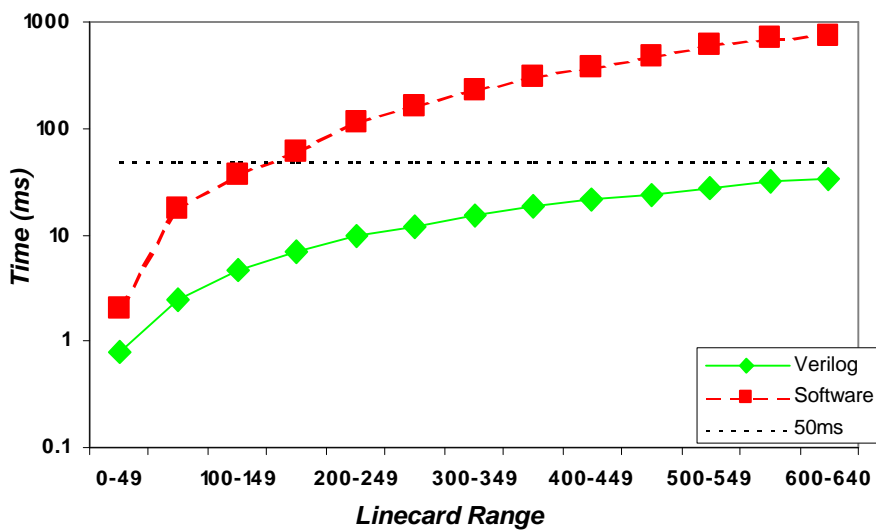
## Matrix Decomposition Problem

- Use of sparsity of matrices to represent the ones as a row-column pair
- Consists of two stages
  - Greedy Algorithm
  - Slepian-Duguid Algorithm
    1. Decomposes all the permutation matrices at once
    2. Uses the row-column pair list structure

# Synthesis

- 40 Groups and 640 Linecards
- 0.13u process
- Cycle time within 4ns
- Connection Assignment Problem
  1. 10K gates
  2. 24Kbits memory
- Matrix Decomposition Problem
  1. 25K gates
  2. 230Kbits of memory

# Reconfiguration Time



Thank you.