Exploiting Parallelism to Boost Data-Path Rate in High-Speed IP/MPLS Networking

Indra Widjaja and Anwar I. Elwalid

IEEE INFOCOM 2003

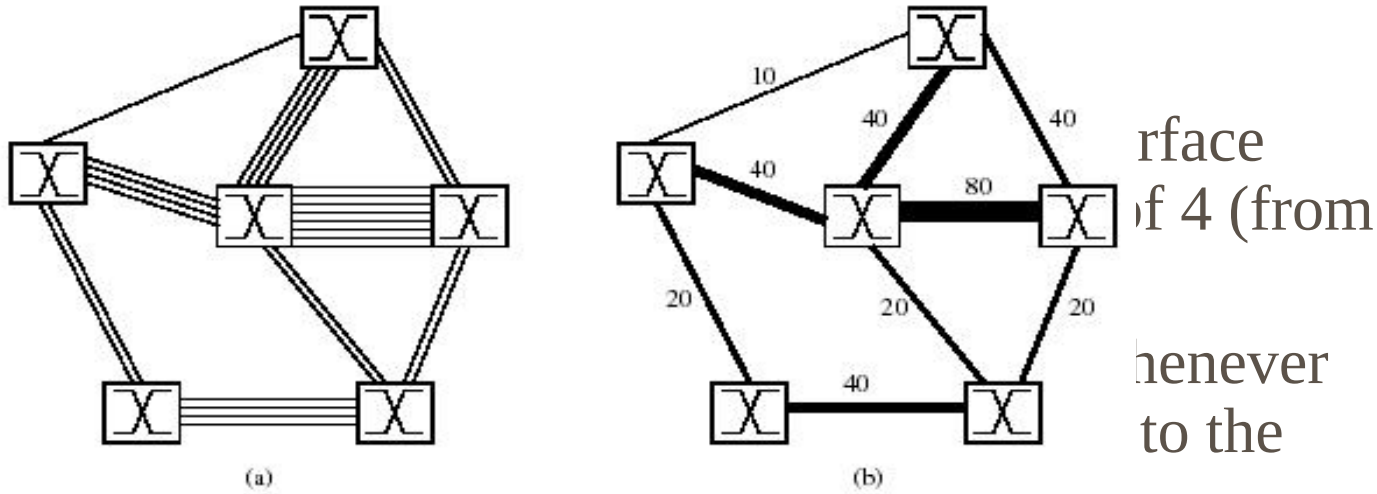
報告人：唐崇實

Outline

- Introduction
- LB versus LB/DA
- Switch Architecture and Design
- Switch Performance
- Switch Extensions
- Conclusions

Introduction

- Th
  - rface
  - f 4 (from
  - henever
    to the



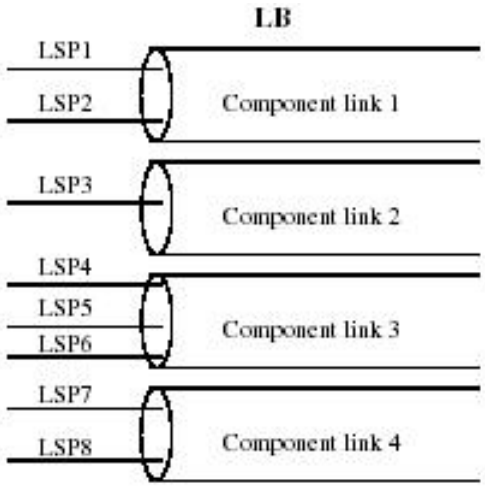
next-generation speed.

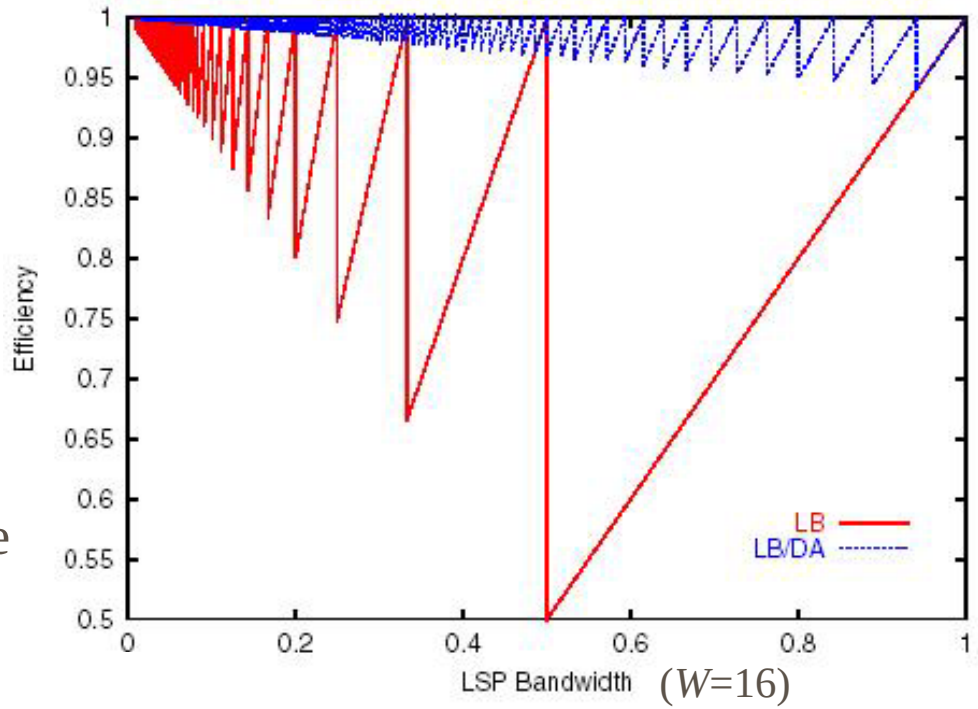
- Link bundling
  - a way to increase routing scalability
  - done by advertising the parallel links as a single link into IGP

Introduction

- Problem
  - Link bundling requires an LSP to be associated with a particular component link.
  - A new LSP request with bandwidth reservation *B* can be established only if at least one of the component links has unreserved bandwidth greater than *B*
- A more efficient bundling scheme is proposed
  - LB/DA (link bundling with distributed traffic assignment): allow LSP traffic to be distributed among the component links within a bundled link.
  - An IP/MPLS switch architecture supporting LB/DA

LB versus LB/DA





(W=16)

Assume all LSP requests have the same bandwidth requirement of x

$$E_{LB}(x) = \begin{cases} x\lfloor 1/x \rfloor & 0 \le x \le 1, \\ 0 & x > 1, \end{cases}$$

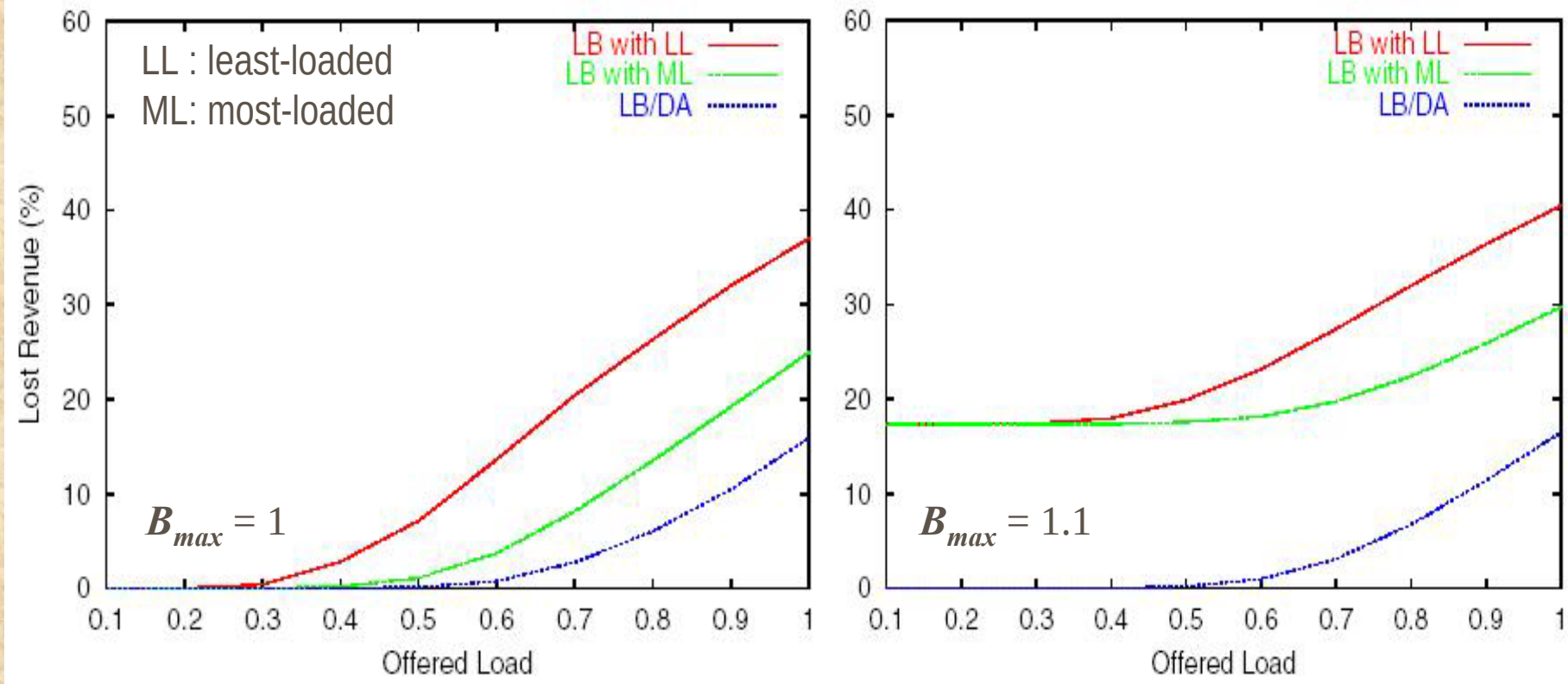
$$E_{LB/DA}(x) = \begin{cases} (x/W)\lfloor W/x \rfloor & 0 \le x \le W, \\ 0 & x > W, \end{cases}$$

where W is bundle size.

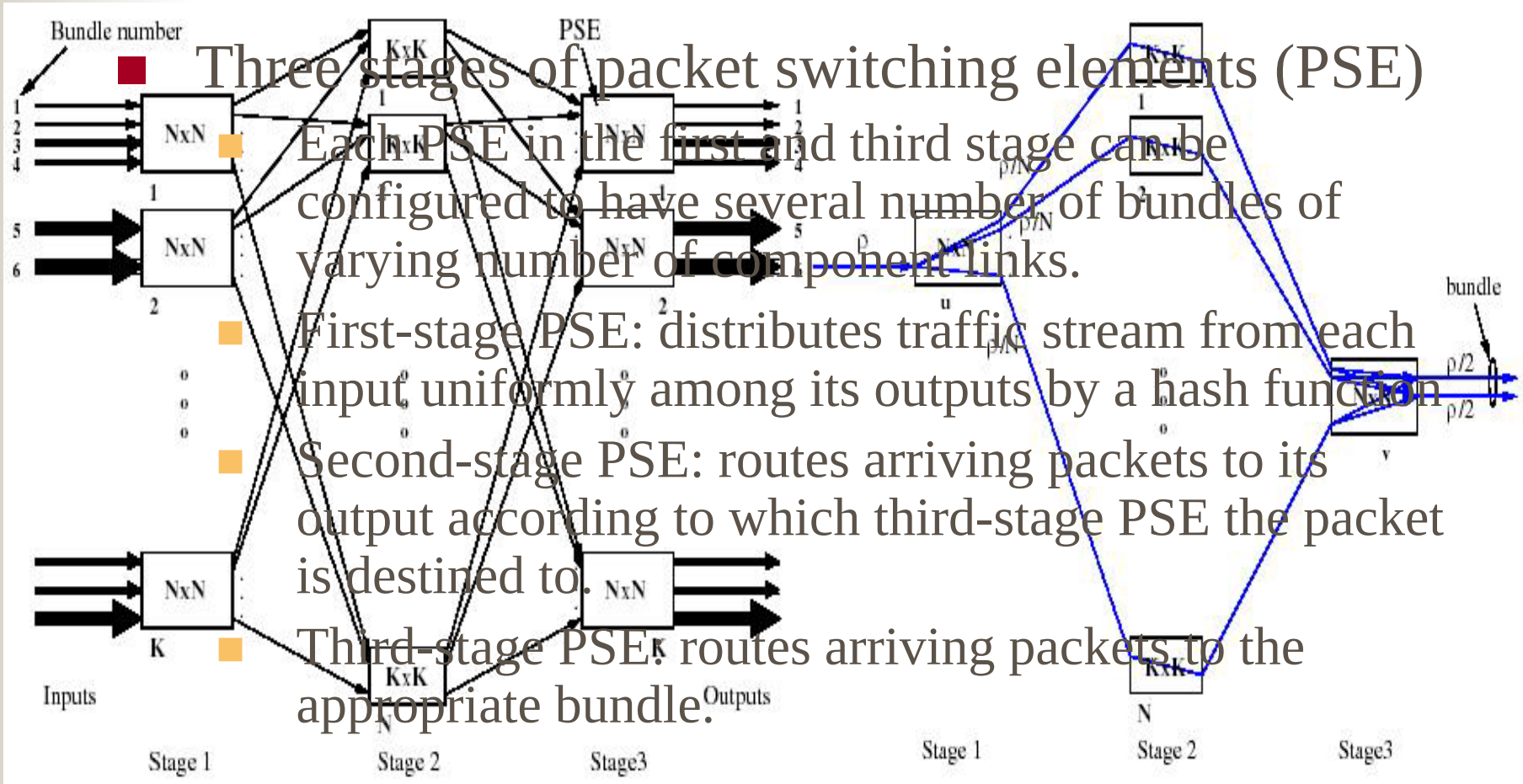
Comparison of LB and LB/DA

R_r (realized revenue) = total bandwidth provided / total bandwidth requested

Lost revenue = 1 - R_r



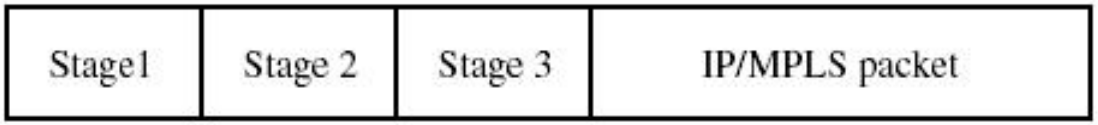
Switch Architecture



- Three stages of packet switching elements (PSE)
- Each PSE in the first and third stage can be configured to have several number of bundles of varying number of component links.
- First-stage PSE: distributes traffic stream from each input uniformly among its outputs by a hash function
- Second-stage PSE: routes arriving packets to its output according to which third-stage PSE the packet is destined to.
- Third-stage PSE: routes arriving packets to the appropriate bundle.

Switch Design

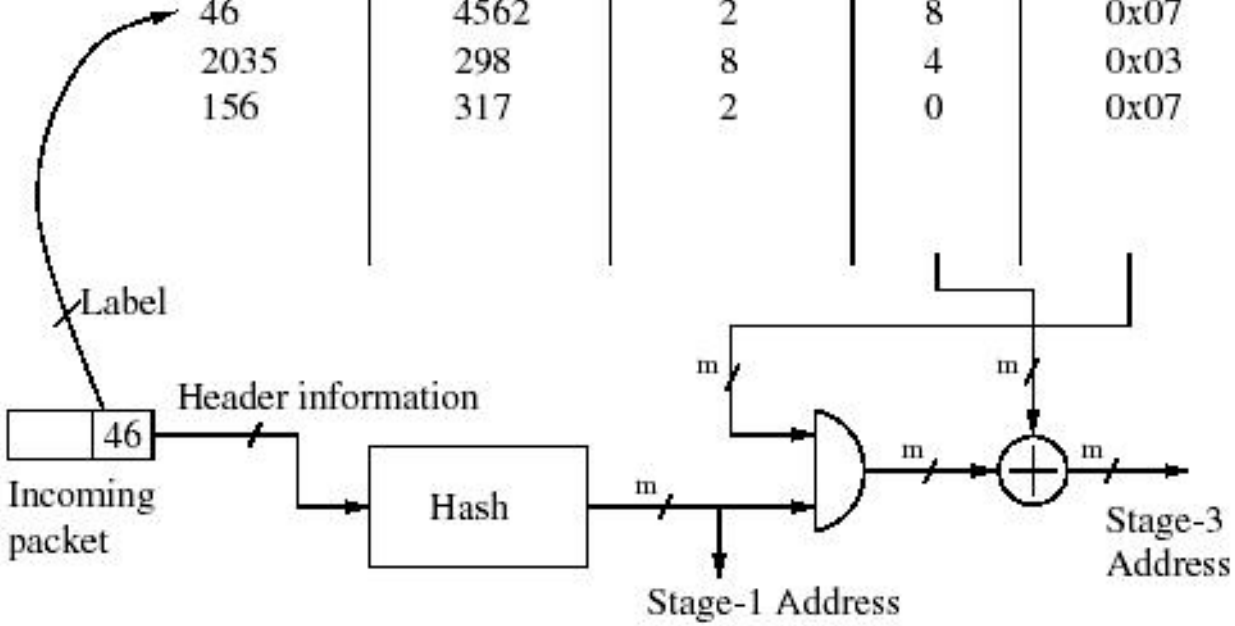
- Inte
- A



dress id

hen an

Incoming Label	NHLFE Label	Stage-2 Address	Bundle Offset	Bundle Mask
101	58	1	0	0x00
46	4562	2	8	0x07
2035	298	8	4	0x03
156	317	2	0	0x07

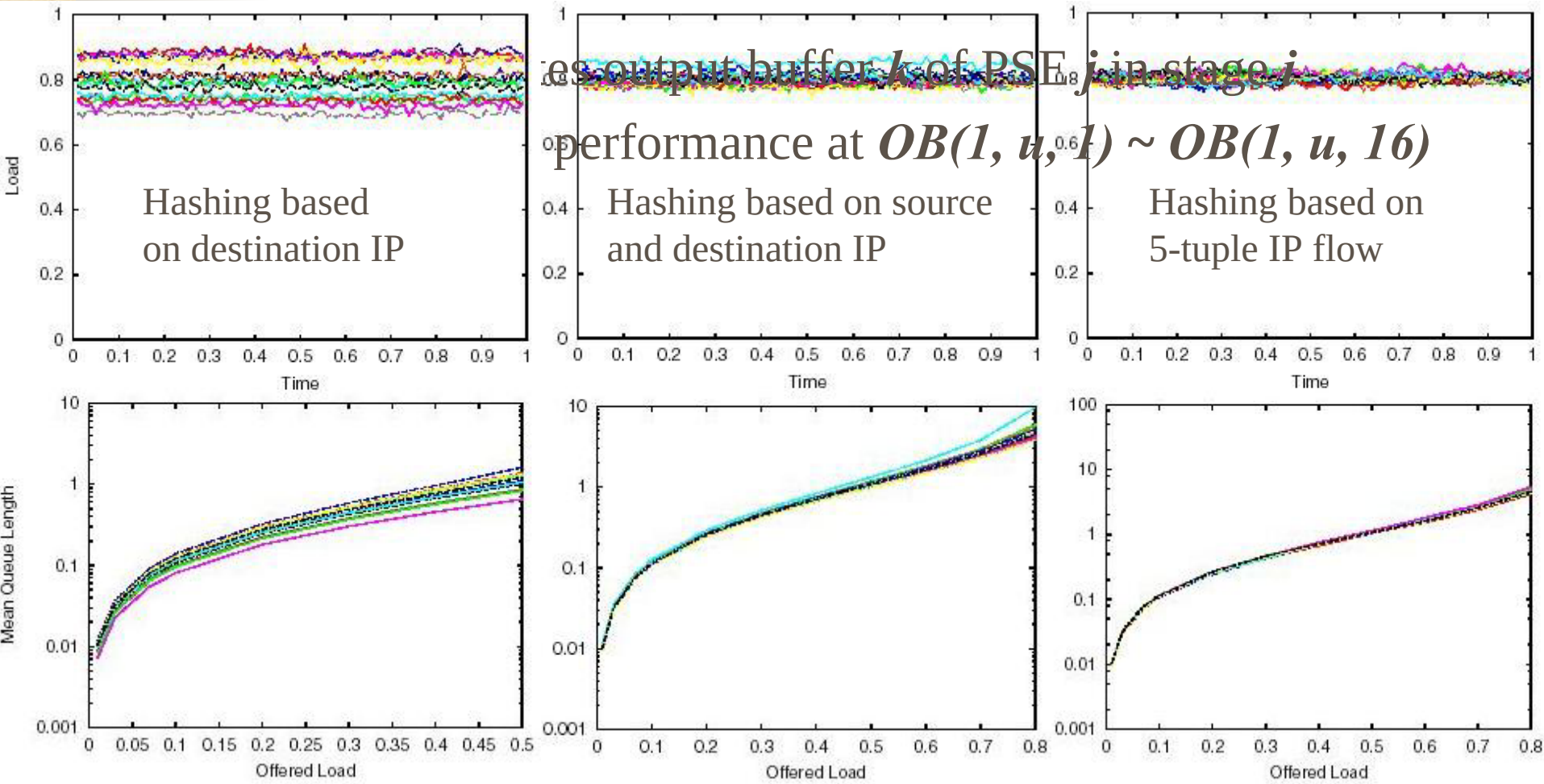


Switch Performance

- Use IP traffic traces from NLANR as the source model
- Each PSE is a 16x16 output buffered packet switch
- Each first/third-stage PSE has five bundles
  - Two small bundles containing two component links each
  - Three large bundles containing four component links each
- Simulation scenario
  - Each large input bundle has LSPs to all output bundles where incoming traffic to the large input bundle is split to each output bundle in proportion to output bundle's capacity
  - Incoming traffic to a small input bundle is entirely delivered to a unique large output bundle.

Load Balancing (Stage 1)

- Uses output buffer *k* of PSE *j* in stage *i*
- performance at ***OB(1, u,1) ~ OB(1, u, 16)***

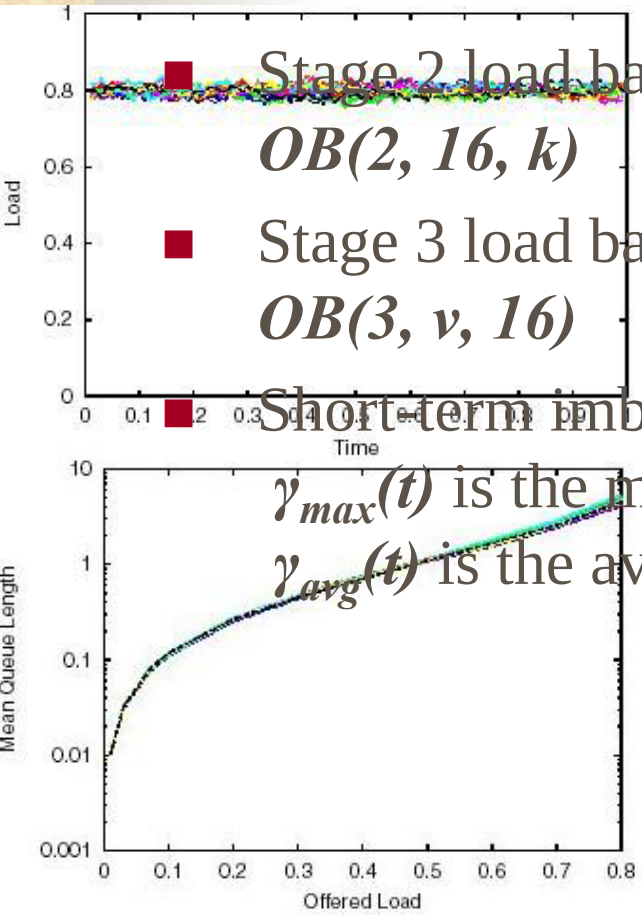


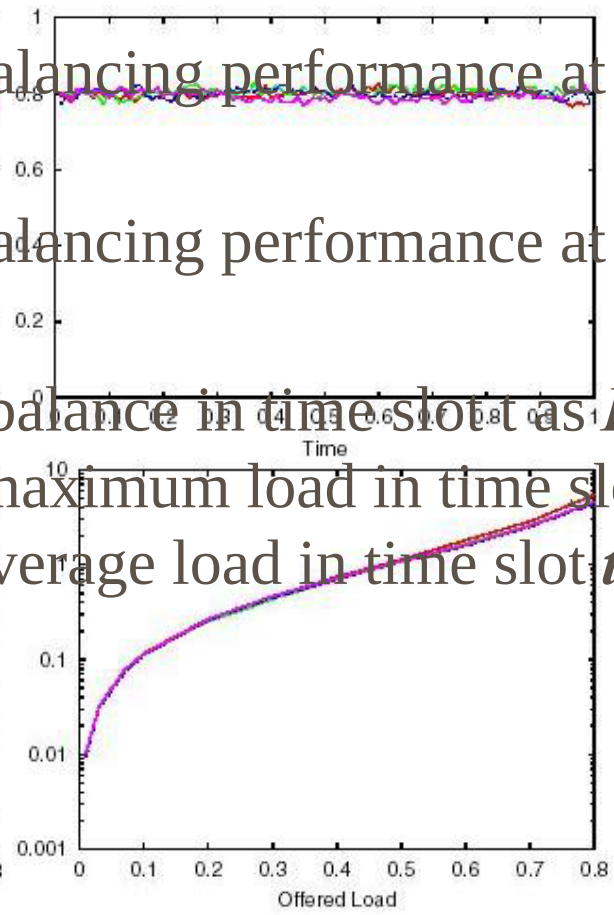
Stage 2&3 Load Balancing and Short-term Imbalance

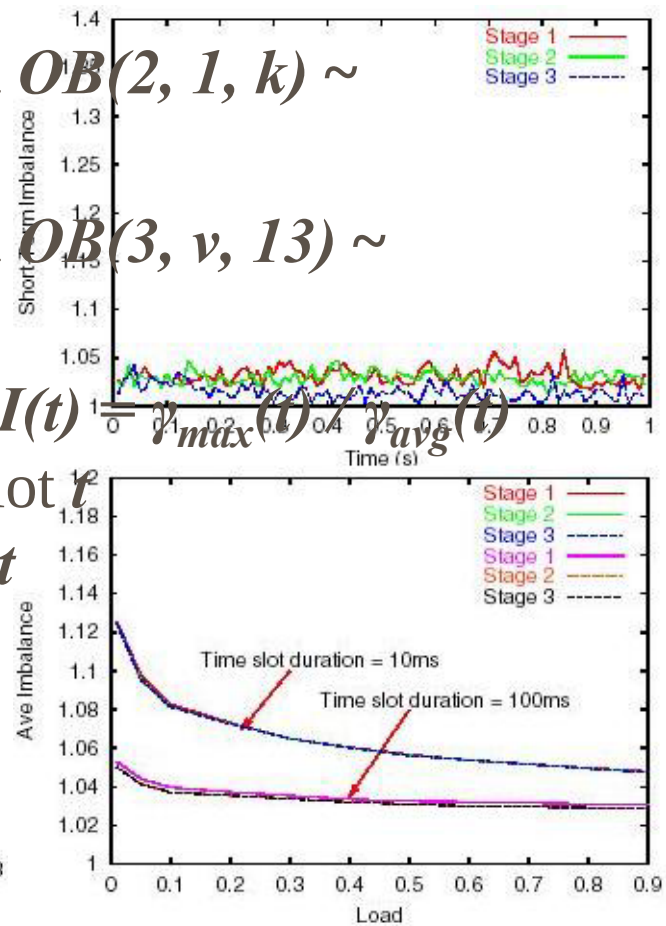
- Stage 2 load balancing performance at ***OB(2, 1, k) ~ OB(2, 16, k)***
- Stage 3 load balancing performance at ***OB(3, v, 13) ~ OB(3, v, 16)***
- Short-term imbalance in time slot as $I(t) = \gamma_{max}(t) / \gamma_{avg}(t)$

$\gamma_{max}(t)$ is the maximum load in time slot t

$\gamma_{avg}(t)$ is the average load in time slot t

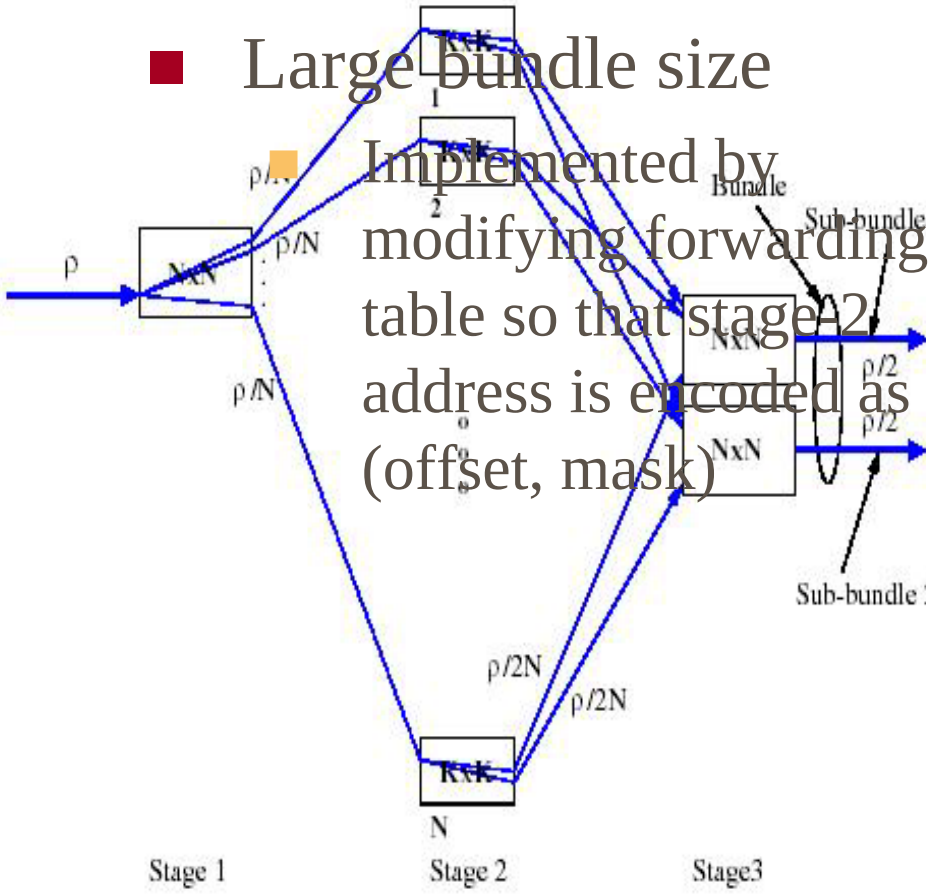




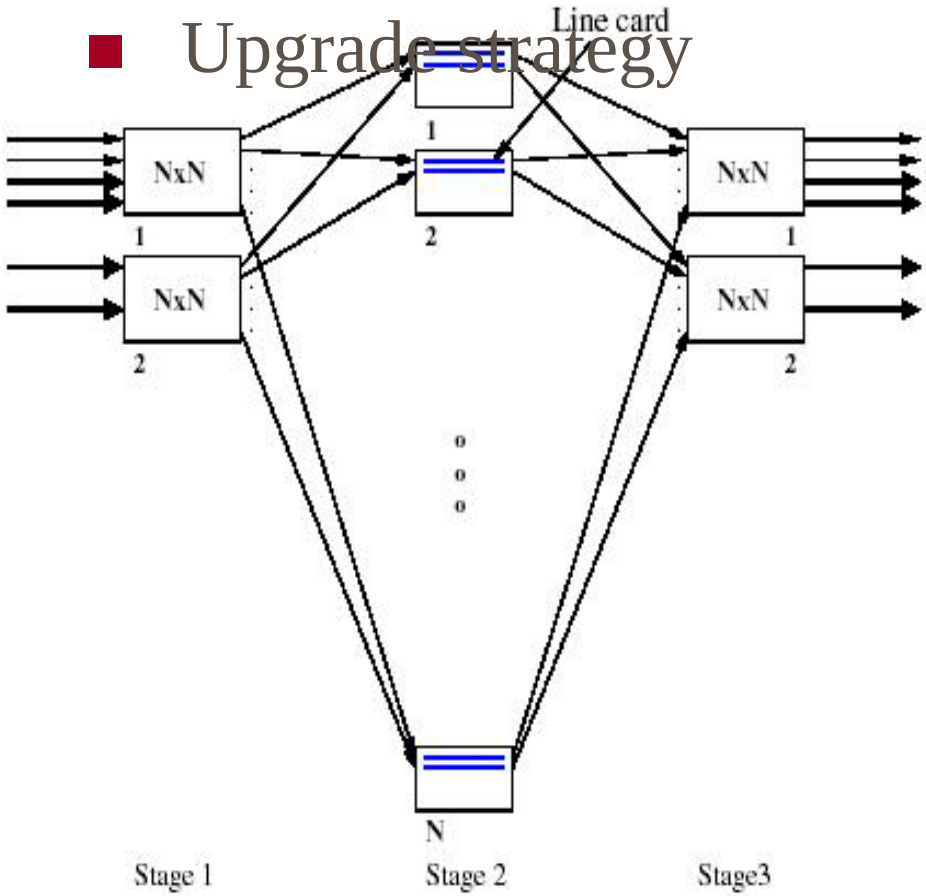


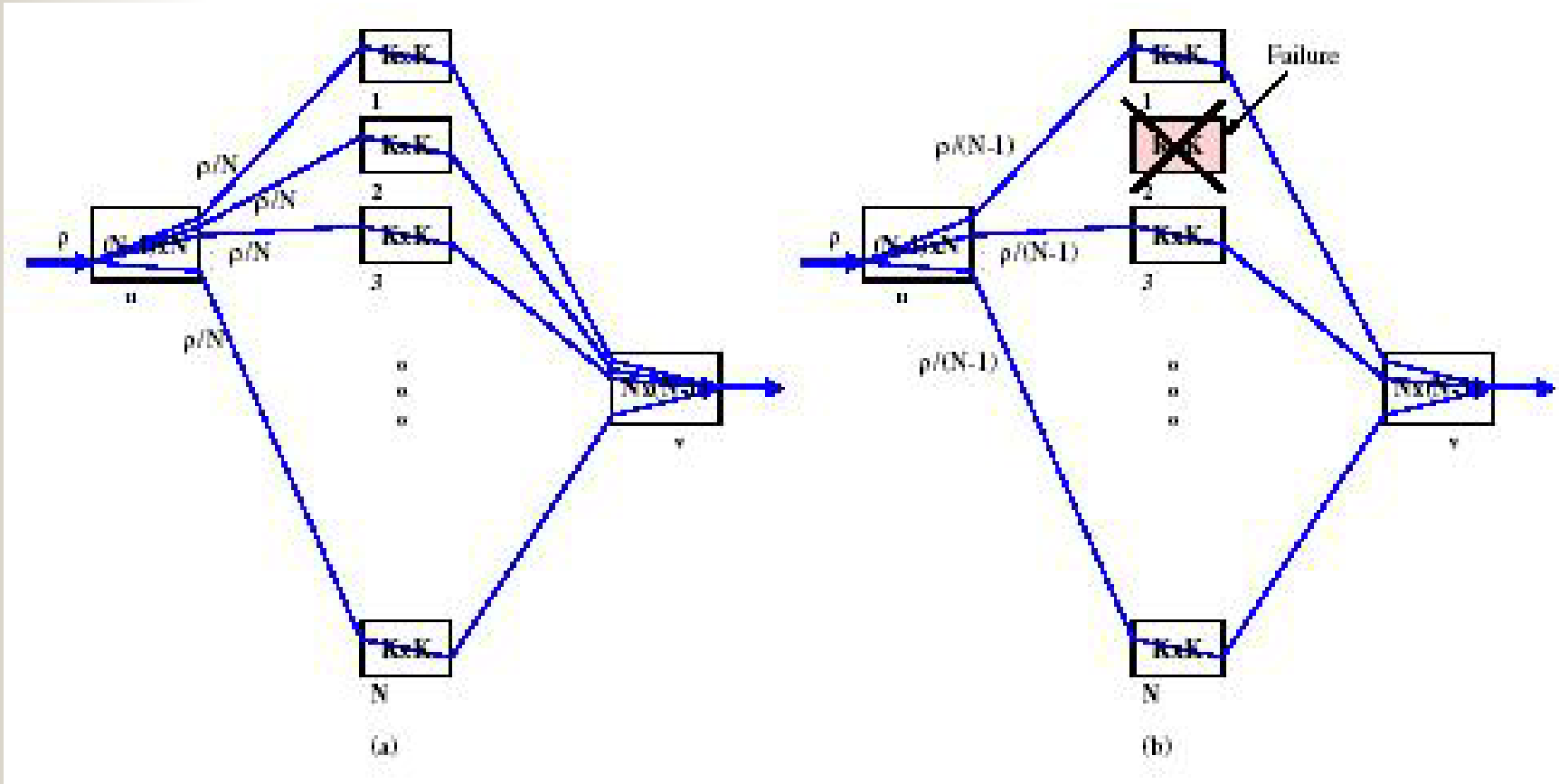
Switch Extensions

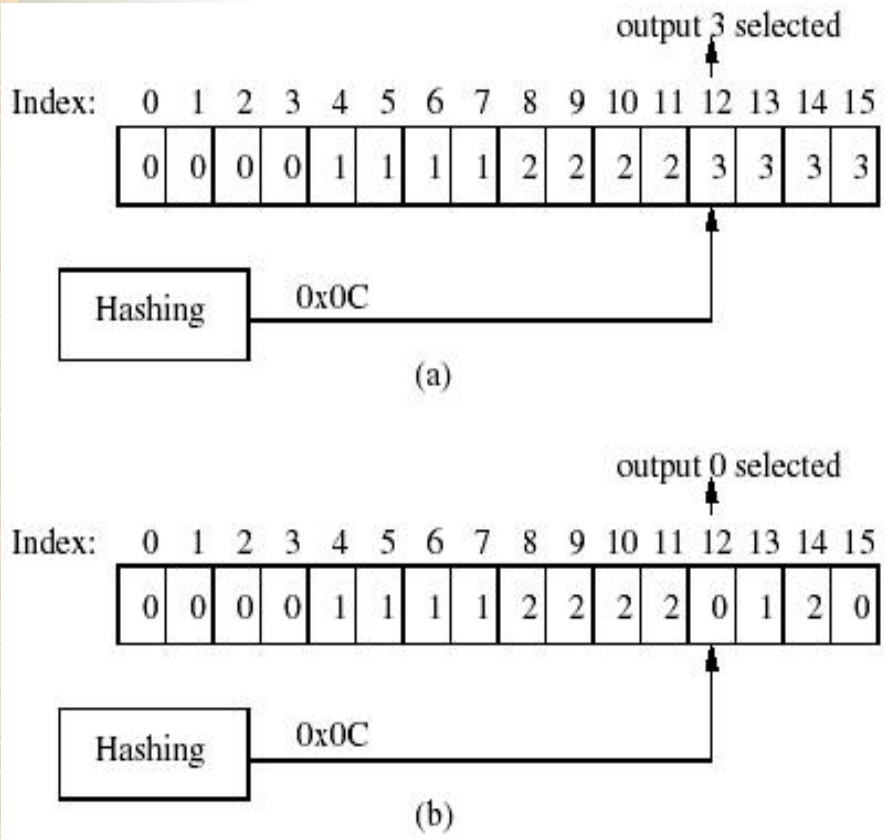
- Large bundle size
  - Implemented by modifying forwarding table so that stage-2 address is encoded as (offset, mask)

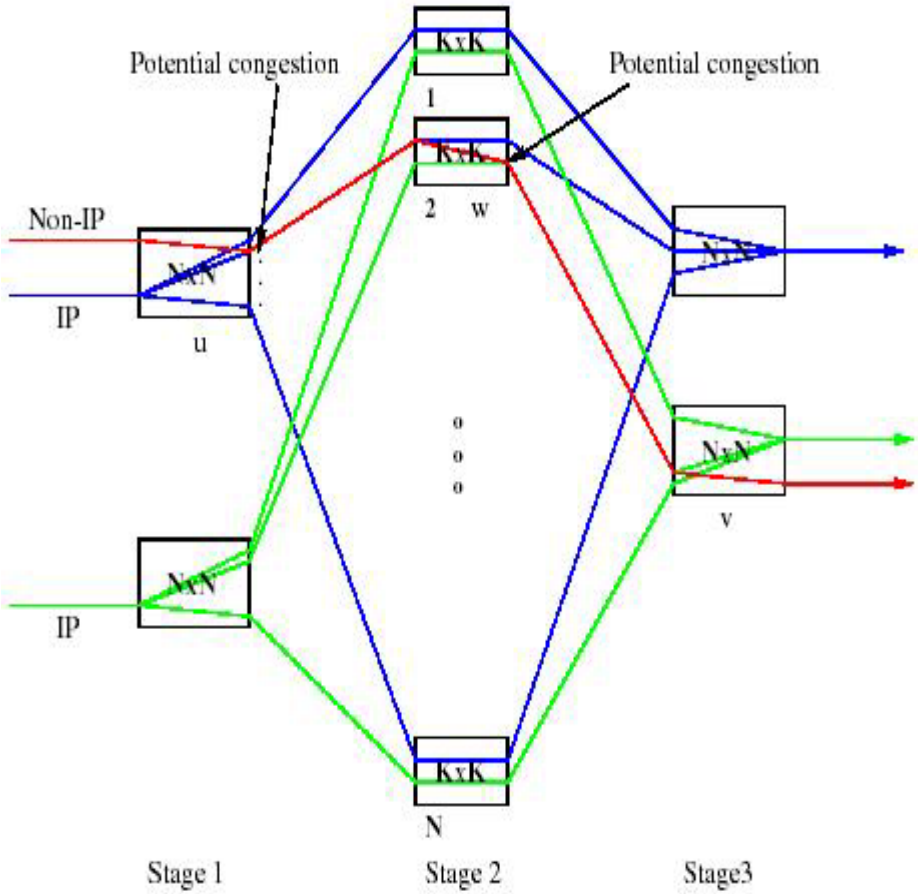


- Upgrade strategy




ρ/N
ρ/N
ρ/N
ρ/N
ρ
1
2
3
N
(a)
Failure
ρ/(N-1)
ρ/(N-1)
ρ/(N-1)
ρ
1
2
3
N
(b)


output 3 selected
Index: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0 0 0 0 1 1 1 1 2 2 2 2 3 3 3 3
Hashing
0x0C
(a)
output 0 selected
Index: 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0 0 0 0 1 1 1 1 2 2 2 2 0 1 2 0
Hashing
0x0C
(b)


Potential congestion
Potential congestion
KxK
1
KxK
2 w
Non-IP
IP
NxN
u
o
o
o
NxN
v
NxN
IP
KxK
N
Stage 1
Stage 2
Stage3

Conclusions

- A novel HDP for the creation of MPLS path is proposed.
- HDP reduces the setup time at the expense of an increased number of signaling messages.
- Discussion
  - Although BB's are separate from the physical nodes, it still needs to provide a "physical path" for signaling messages.
  - It is a question that if the hierarchy of more than two levels is really necessary.
  - Is is worthy to reduce the setup time at the expense of an increased number of signaling messages?
  - Other applications?