FUJITSU

shaping tomorrow with you

First-Come First-Served Routing for the Data Center Network

March, 2012

Keiji Miyazaki
Fujitsu Laboratories Ltd.

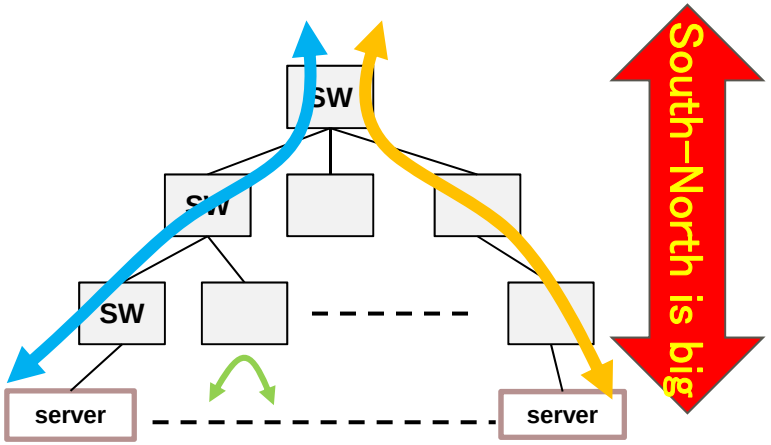
Outline

FUJITSU

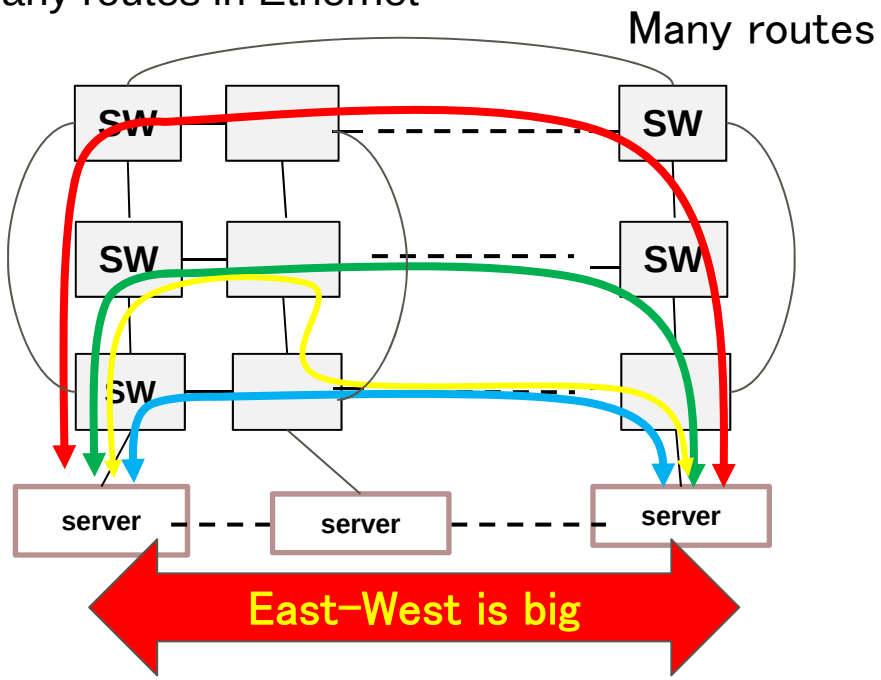
- Introduction
- First-Come First-Served Routing
- Evaluation results
- Conclusion

Data center network

- Legacy Datacenter
  - Ethernet using traditional tree topology, root switch becomes bottleneck
  - South-North traffic is big
- Current Datacenter
  - East-West traffic is big for the distributed data processing such as Hadoop
    - Low latency and high reliability is needed for such service
  - Fat tree or Cube topology to cancel bottleneck of the tree topology
  - Loop-free is important because there are many routes in Ethernet



Legacy datacenter
Traditional Tree topology

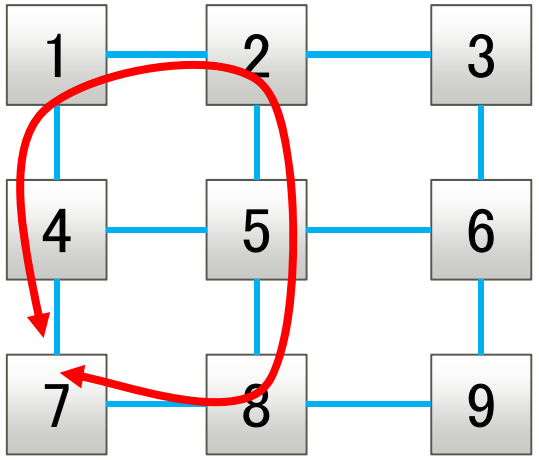


Large-scale datacenter

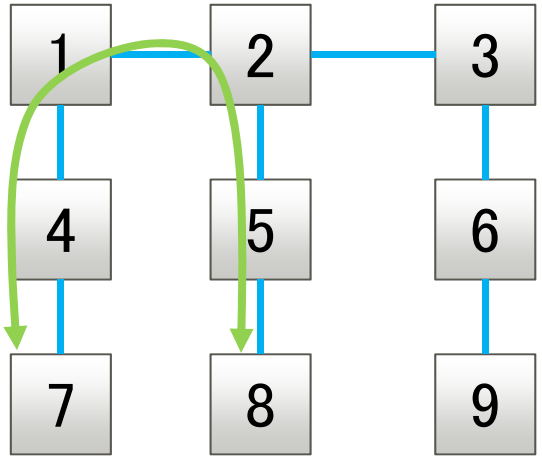
Loop free

FUJITSU

- To ensure loop-free in Ethernet
- STP
  - Un-used link is exist for loop-free
  - Not shortest routes
- SPB(IEEE), TRILL(IETF)
  - Use IS-IS routing protocol
  - Additional overhead（20~22byte）is needed
- Solve these issues in Ethernet
  - Realize loop-free based on packet forwarding without Control Plane
  - First-Come First-Served routing



Loop



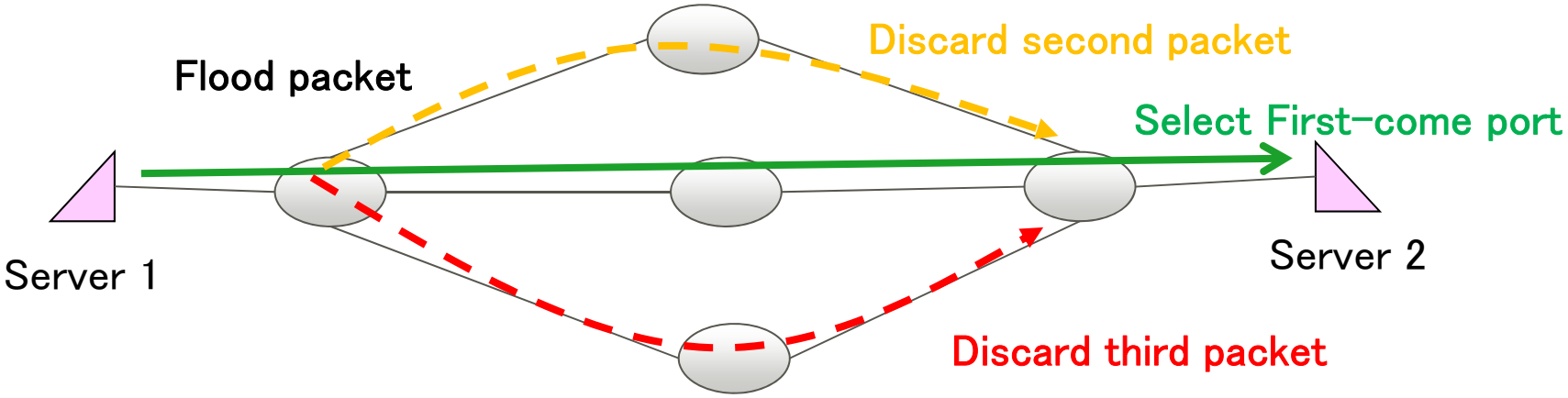
Loop-free

Objective of First-Come First-Served Routing

FUJITSU

- Objective
  - Loop-Free
  - Low latency
  - High utilization（multi-path）
  - High reliability(rerouting)

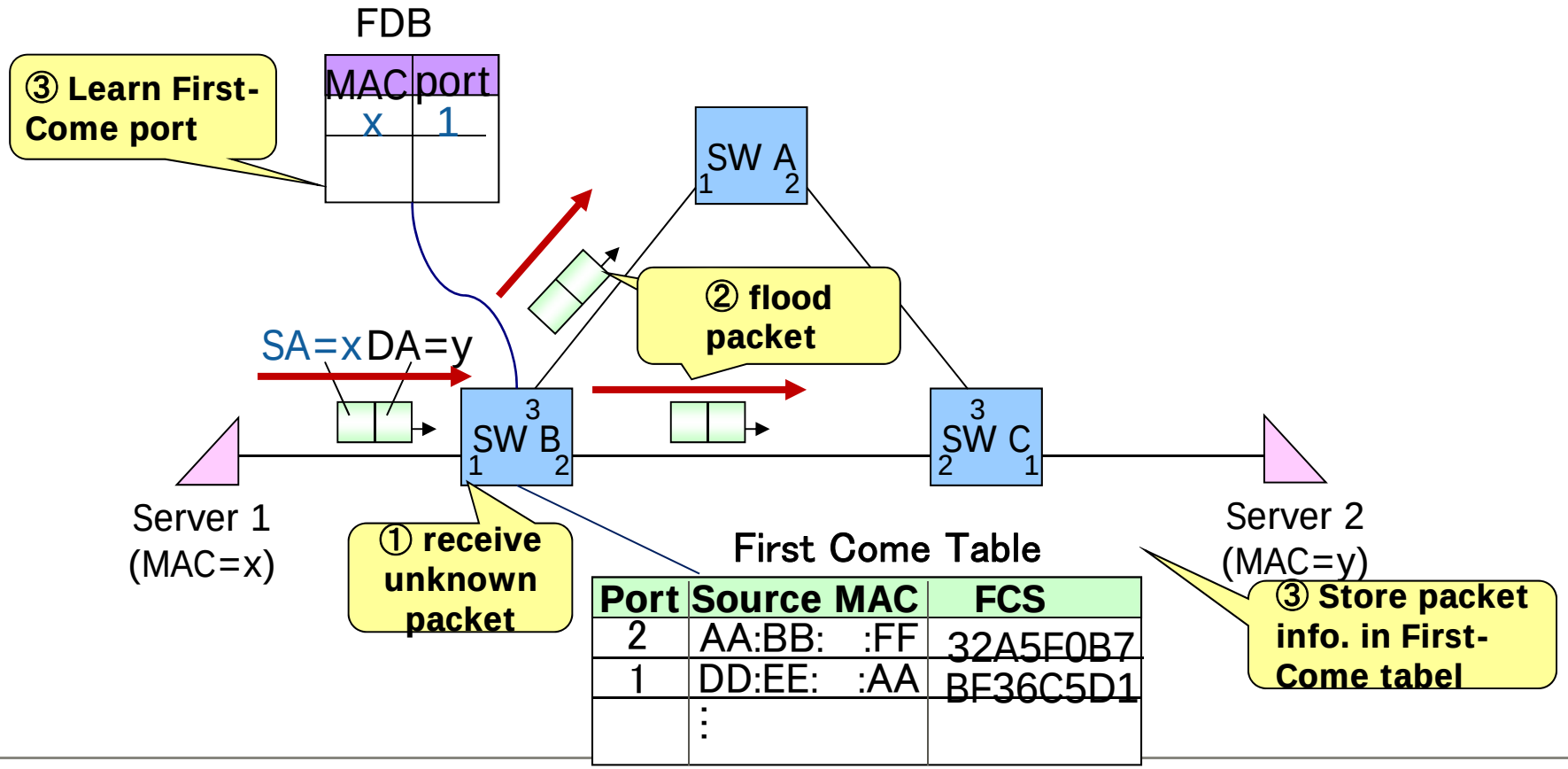
- Mechanism
  - Select port that received packet first
  - Discard not-first packet
  - No control plane



First-Come First-Served Routing

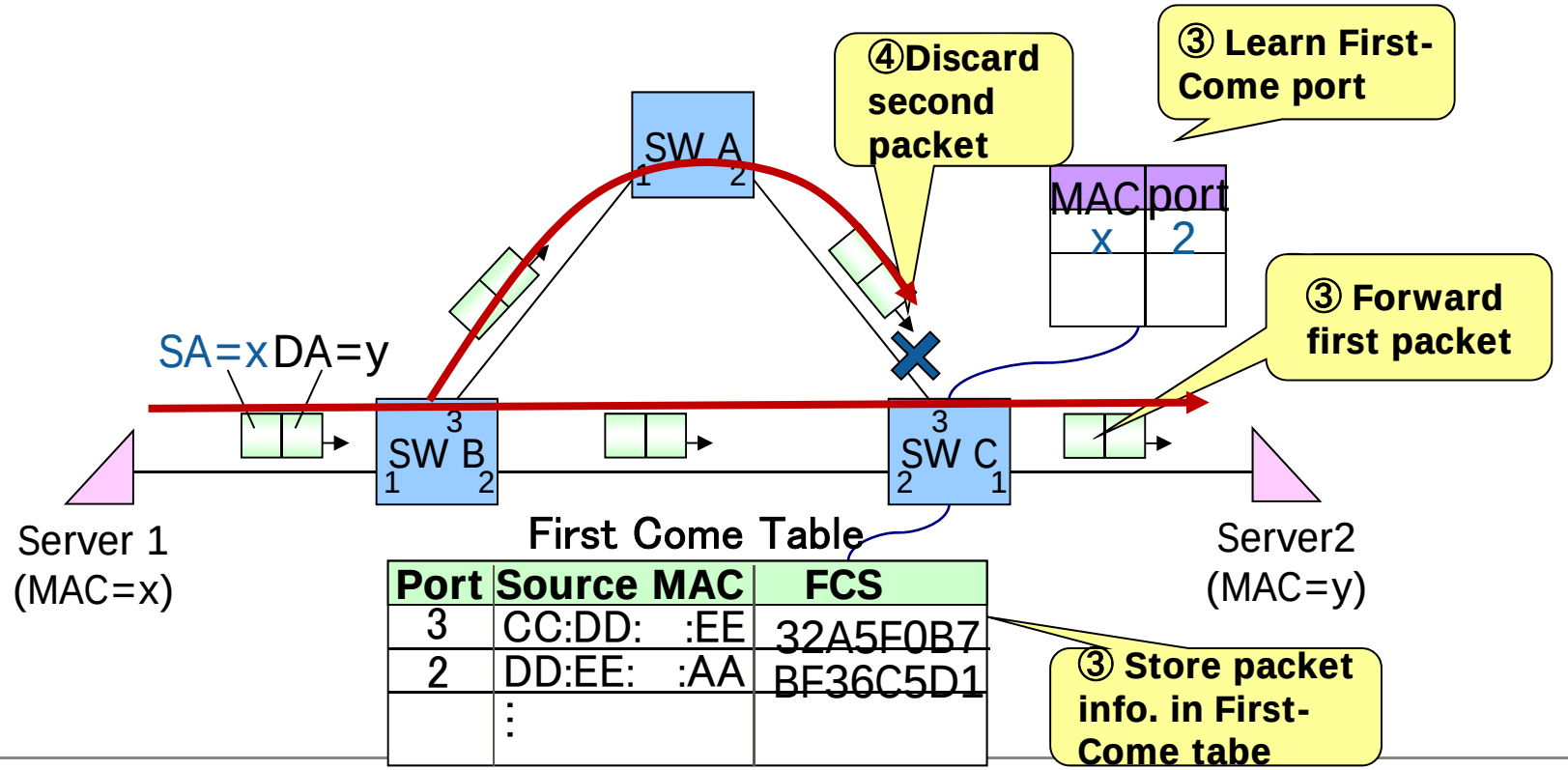
FUJITSU

- Learn port that received first packet as first-come port, source MAC address and FCS
  - When SW received unknown packet, SW stores the received port, source MAC address and FCS in First-Come table.
  - SW floods to the all port except received port



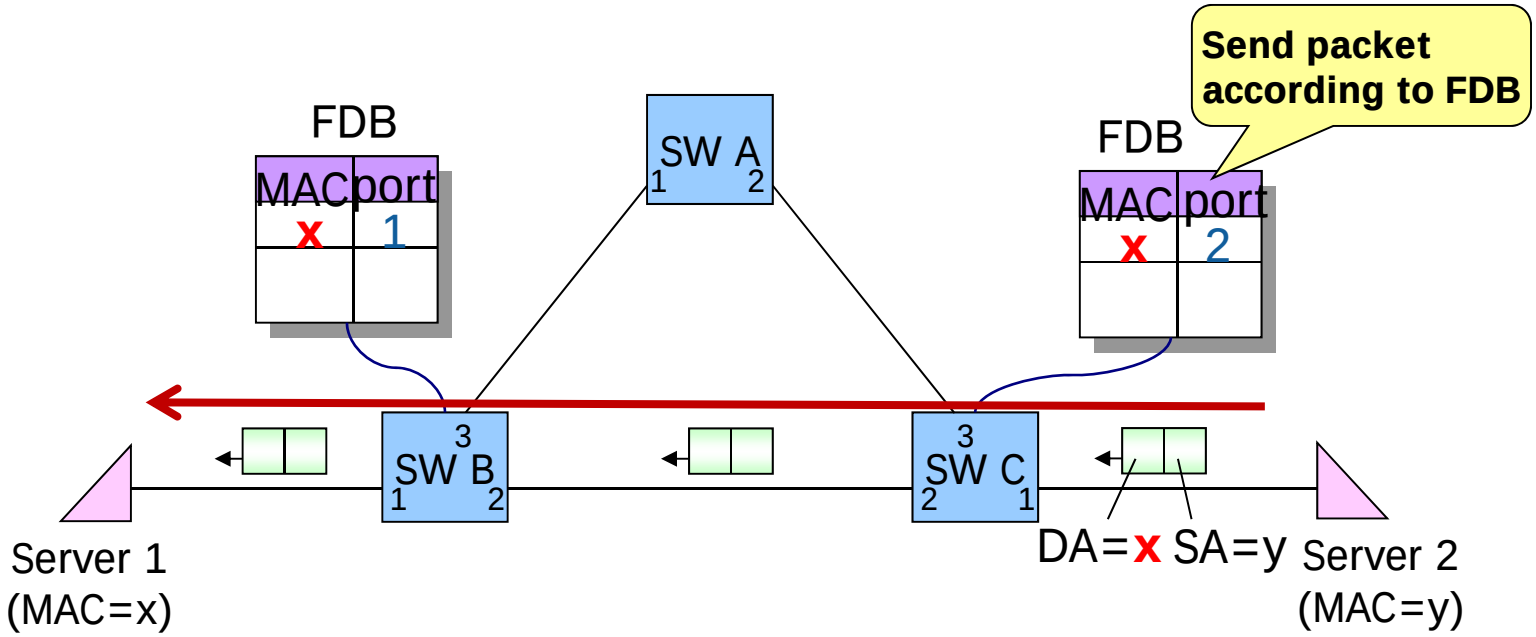
First-Come First-Served Routing（Cont.）

- Learn the port that received packet firstly
  - When SW received packet, it check if the same information is in the First-Come table
  - When the same information is exist in First-Come table, received packet is discard



First-Come First-Served Routing (Cont.)

- In learned state, SW send packet according to FDB

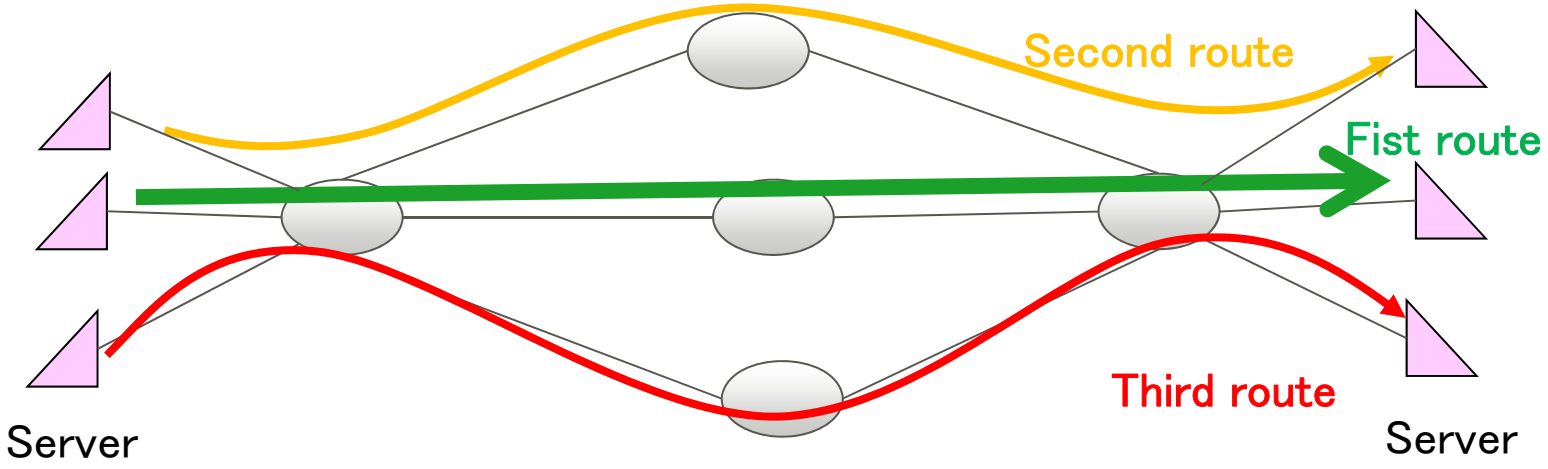


- In this way, loop-free and low latency route is realized

High Utilization

FUJITSU

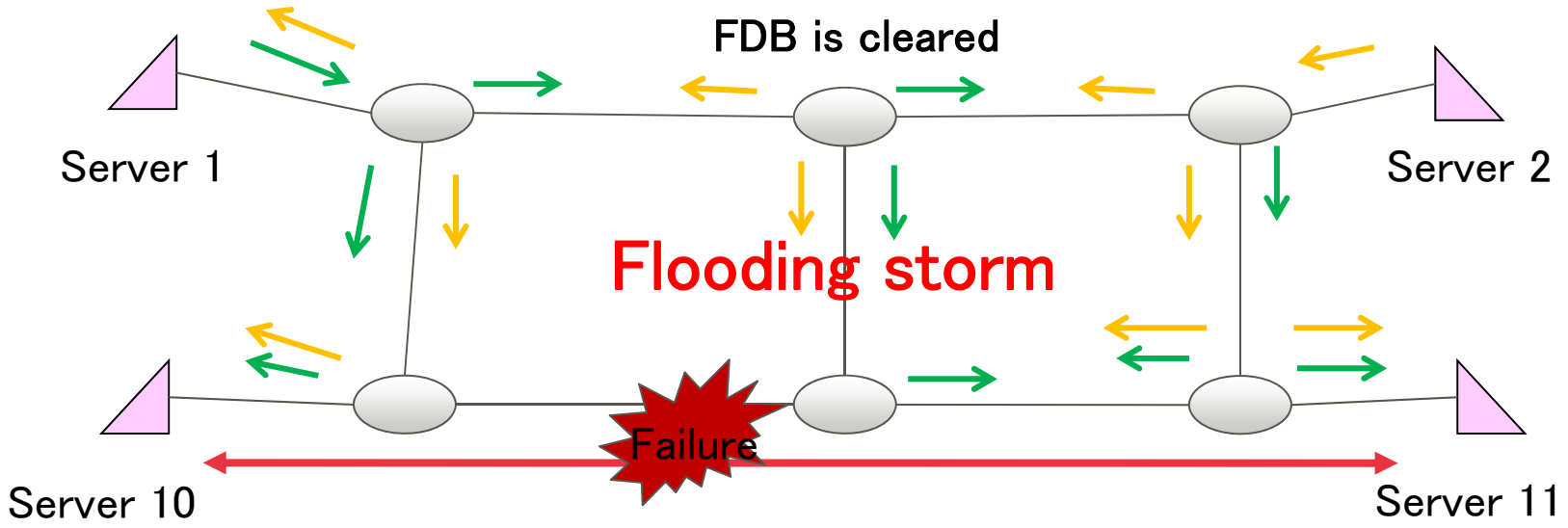
- When first route traffic is increased, latency of first route is increased
- When latency of second route is smaller than first route, second route is chosen



High Reliability

FUJITSU

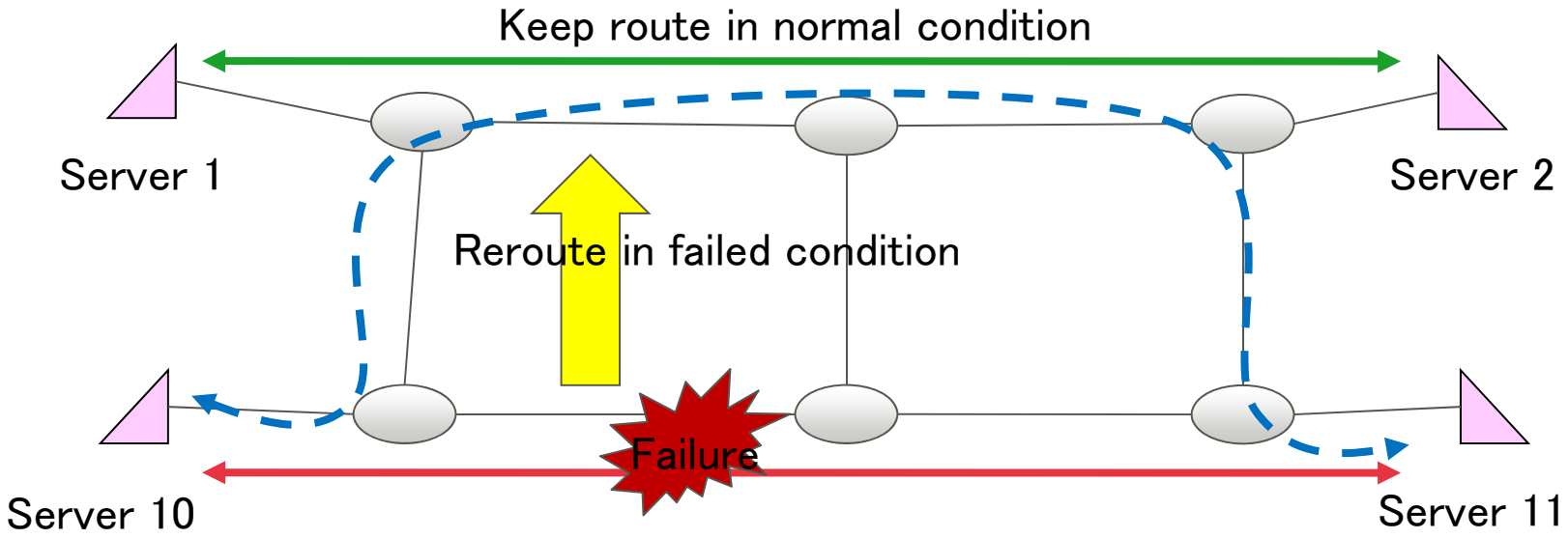
- When failure occurs in the datacenter, failed route must be reroute for high reliability
- Routing base on control plane has rerouting function
- We developed rerouting based on MAC flash
  - Existing MAC flash deletes all MAC address in network
  - flooding storm occurs



Selective MAC Flash

FUJITSU

- Objective
  - Delete only MAC addresses influenced by failure and reroute rapidly
  - Keep MAC addresses not-influenced by failure



Comparison with other routing

FUJITSU

- First-Come First-Served routing is simpler than SPD and TRILL because no routing protocols and no additional tags.

	SPB	**TRILL**	**First-Come First Served**
Additional Tag	24 bytes: 802.1ah tag and B-VLAN tag	20 bytes: TRILL header and Outer Mac header	0 byte: Not required
Control plane	IS-IS	IS-IS	Not required
Loop prevention	RPFC(Reverse Path Forwarding Check)	TTL base and RPFC	First-Come base
Multi-path support	Yes	Yes	Yes
Blocked Link	none	none	none

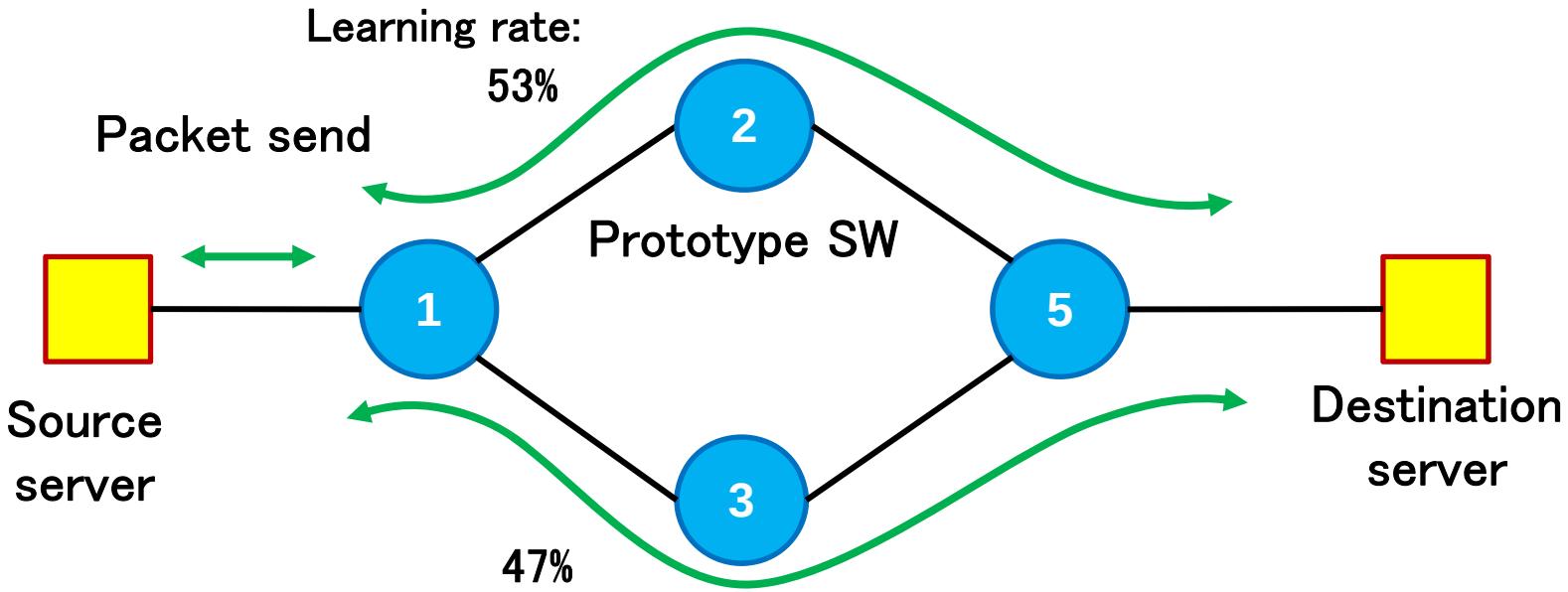
Evaluation

FUJITSU

- Developed prototype software on evaluation board
- Check behaviour of First-Come First Served routing
  - Link utilization
  - MAC flash

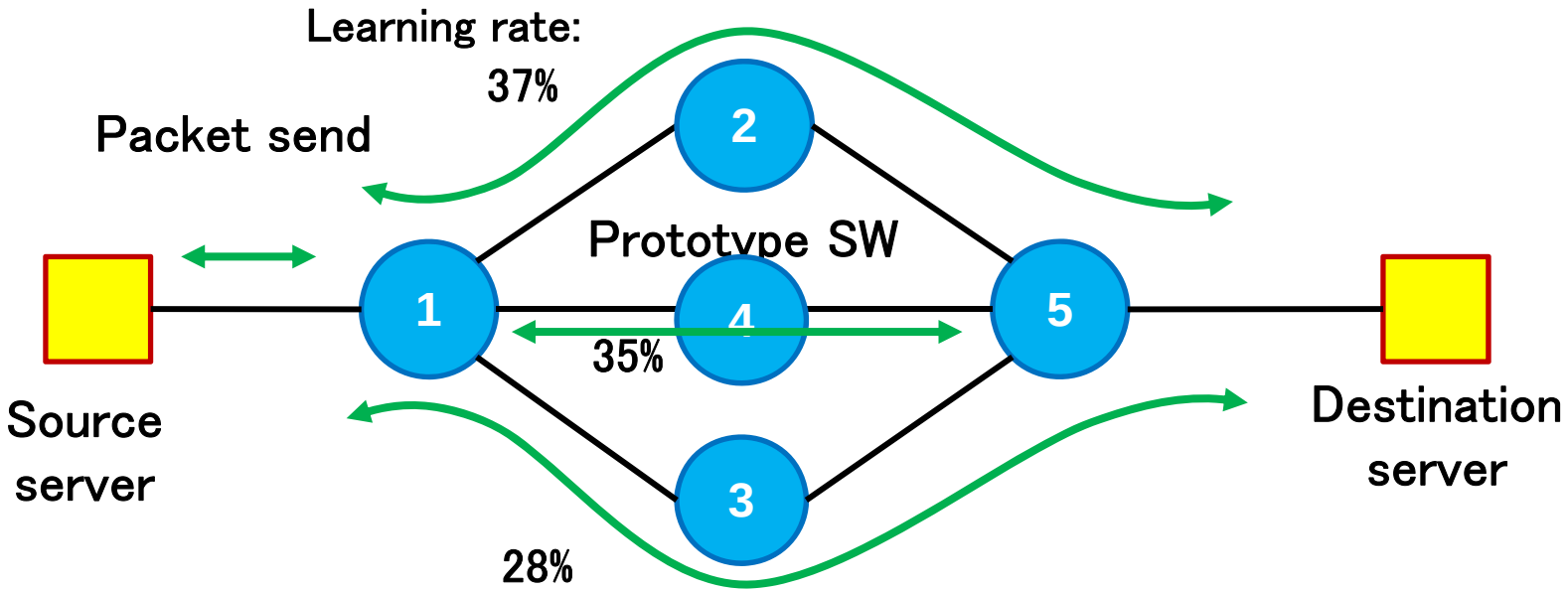
Link utilization（2 routes）

- Transmit 1000 packets with different source and destination from source server to destination server
- Data Flow was divided 1/2



Link utilization（3 routes）

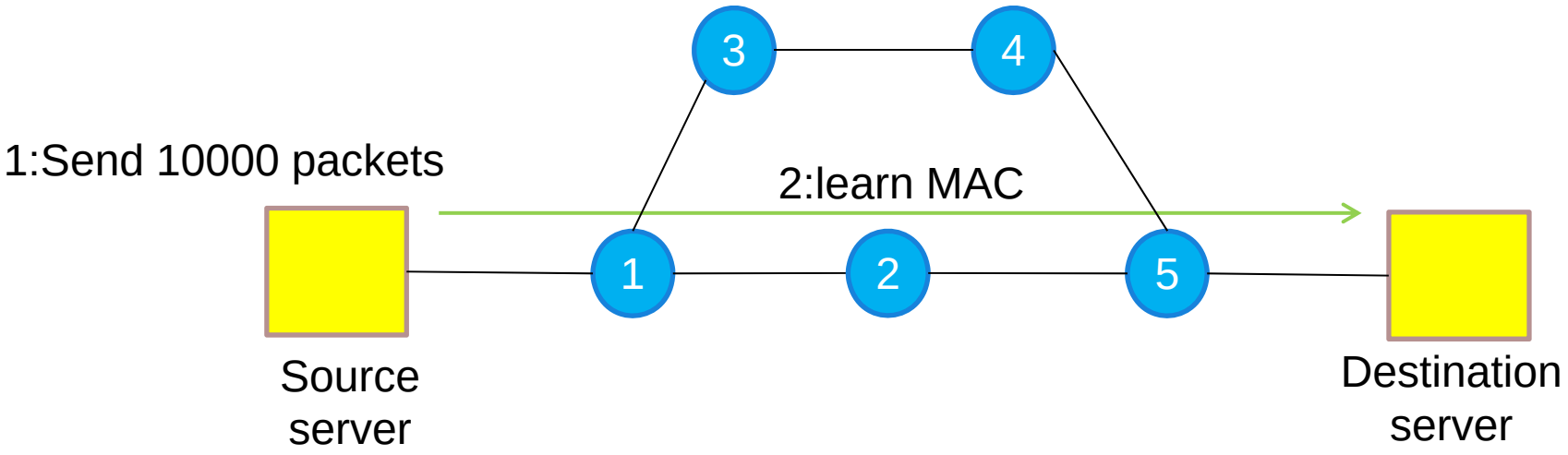
- Transmit 1000 packets with different source and destination from source server to destination server
- Data flow was divided about 1/3



Evaluation of MAC flash

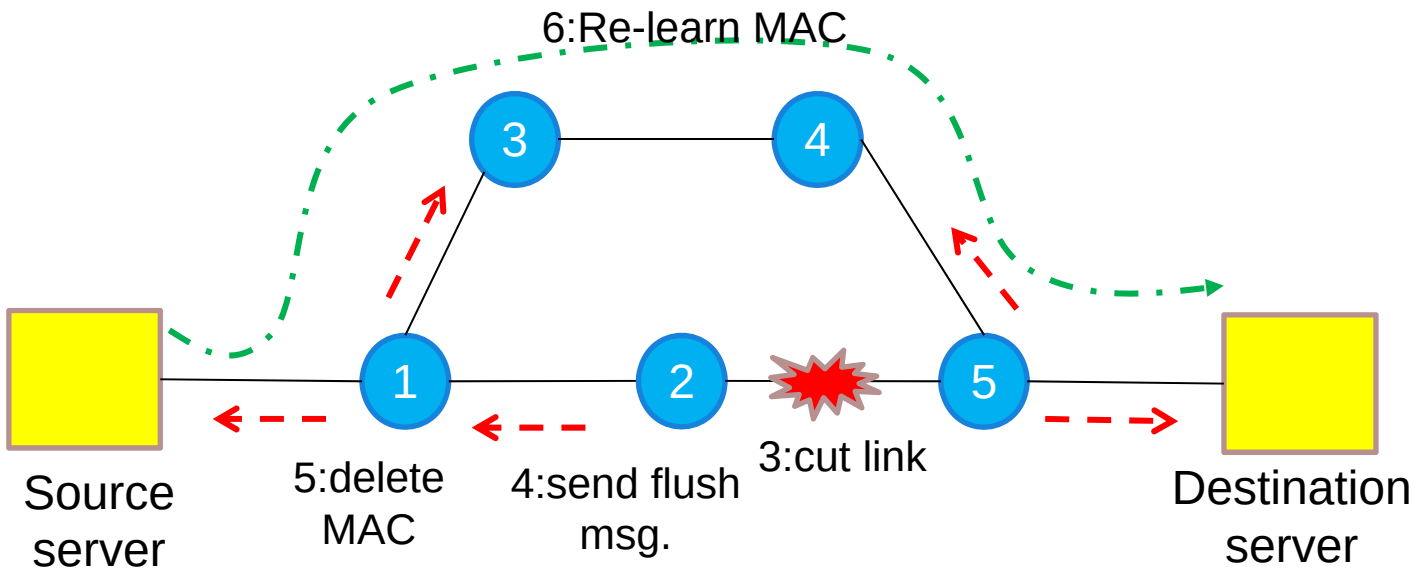
FUJITSU

- Transmit 10000 packets with different source and destination from source server to destination server
  - All routes become same route via SW1, SW2 and SW5



Evaluation of MAC flash(Cont.)

- Cut link between SW2 and SW5
- SW2 and SW5 flood MAC flash message.
  - SW2 and SW5 delete failed MAC in FDB
- Only failed MAC address was deleted in FDB
- 10000 failed routes was reroute in 31 ms

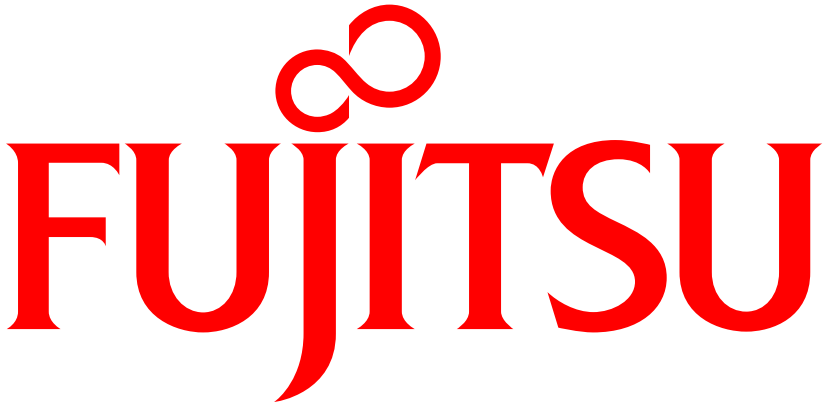


Conclusion

FUJITSU

- We proposed new routing based on packet forwarding
  - Loop-free
  - Low latency
  - High utilization（multi-path）
  - High reliability(rerouting)

- We developed prototype software and evaluate our routing



shaping tomorrow with you