The 7th International Workshop on Information Systems for Social Innovation 2015 (ISSI2015) - Society and Big Data -, Feb. 17, 2015, NII, Tokyo, Japan

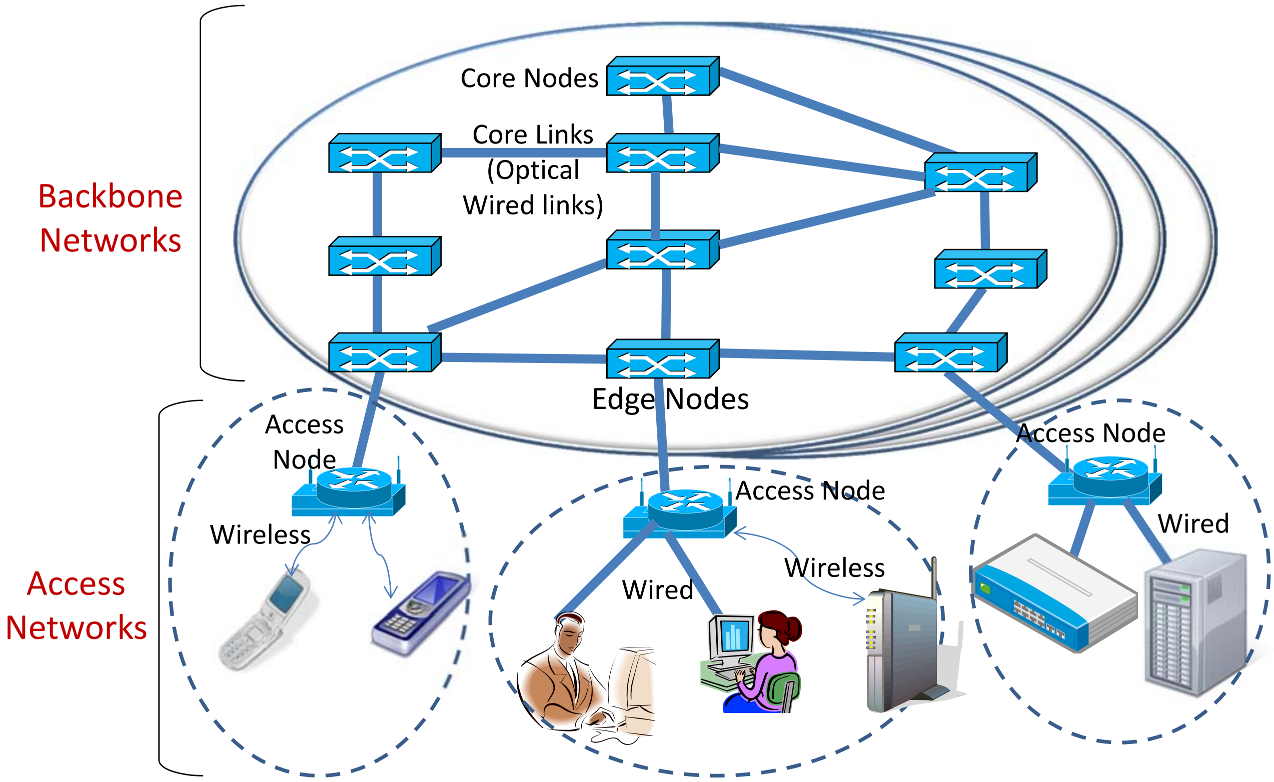
Disaster-Resilient Backbone and Access Networks

Shigeki Yamada
(shigeki@nii.ac.jp)
Principles of Informatics Research Division,
National Institute of Informatics (NII), Tokyo, Japan

Introduction and Background

- This presentation is a summary of two-year Resilient Network Research Project promoted under JSPS Resilient Life Space Umbrella Project.
- Once Natural disasters such as earthquakes, and tsunami occur, they may cause network breakdowns due to link and node failures, resulting in network service disruptions
- The network should quickly recover and keep operating after the disasters.
- Resilience: the ability of network to provide an acceptable level of service in the face of various faults and challenges to normal operations.
- Resilient technologies for two types of network (the backbone network and access network) are investigated to make networks more resilient.

Backbone Networks and Access Networks



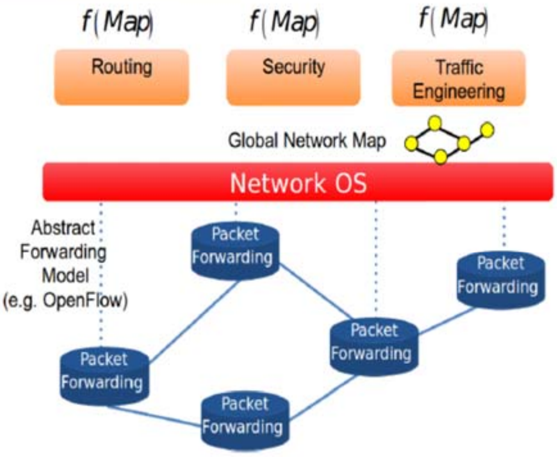
Proposed Approaches for Backbone Networks & Access Networks

- Backbone networks
  - Abundant and redundant network resources (links and routers/switches) for large bandwidth and high reliability.
  - A part of backbone networks may continue to survive even if a large scale link/node failure occurs due to a large disaster.
  - Utilizing still available network resources (links, nodes) could enable the network to continue providing acceptable services for quick recovery.
- Access networks
  - Closely located to users, diversified in technologies from wireless to wired, and usually not redundant: once a disaster breaks down access networks, it may be very difficult to quickly repair them.
  - Rather than repairing the destructed access network, volunteers in the disaster area could newly construct their access network on-site more quickly and more easily.

Requirements to Resilient Backbone Networks

- Network resilience must pass through two major phases
  1. Failure Detection Phase: detection of alarms and alerts to locate network faults
  2. Network Recovery Phase: disables a failed port, enables another, reroutes traffic around a failed switch or router
- Both phases should be fast completed in a scalable way even for a large-scale backbone network
- For fast failure detection, we can fully utilize the existing detection technologies like BFD (Bidirectional Forwarding Detection)
  - For fast network recovery, at most 50 ms is considered tolerable to complete path restoration, in the provider networks
  - We focus on applying SDN (Software Defined Networking) /OpenFlow technologies to the backbone network because SDN/OpenFlow have a potential capability to provide the programmability and flexibility, responding fast to network situational changes

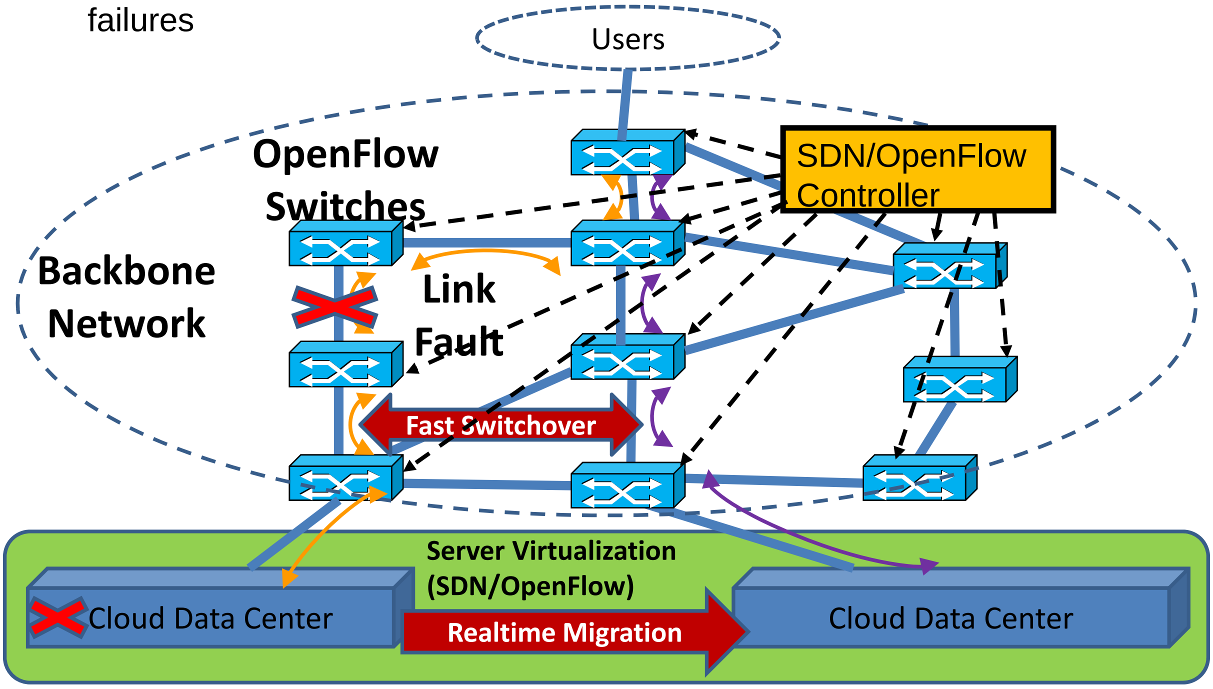
SDN/OpenFlow for Backbone Networks



- Network Operating System (NOS) with a global view of network controls forwarding hardwares via OpenFlow protocol
- Network intelligence is on top of NOS
- SDN/OpenFlow provides an easier way to manage and automate networks

Goal of Resilient Backbone Network

- The goal of resilient backbone network is to provide non stoppable end-to-end services in the various critical environments, including link/path and node failures



Users

OpenFlow Switches

SDN/OpenFlow Controller

Backbone Network

Link Fault

Fast Switchover

Server Virtualization (SDN/OpenFlow)

Cloud Data Center

Realtime Migration

Cloud Data Center

Influential Factors to Enable Fast Network Recovery Using SDN/OpenFlow Technologies (1)

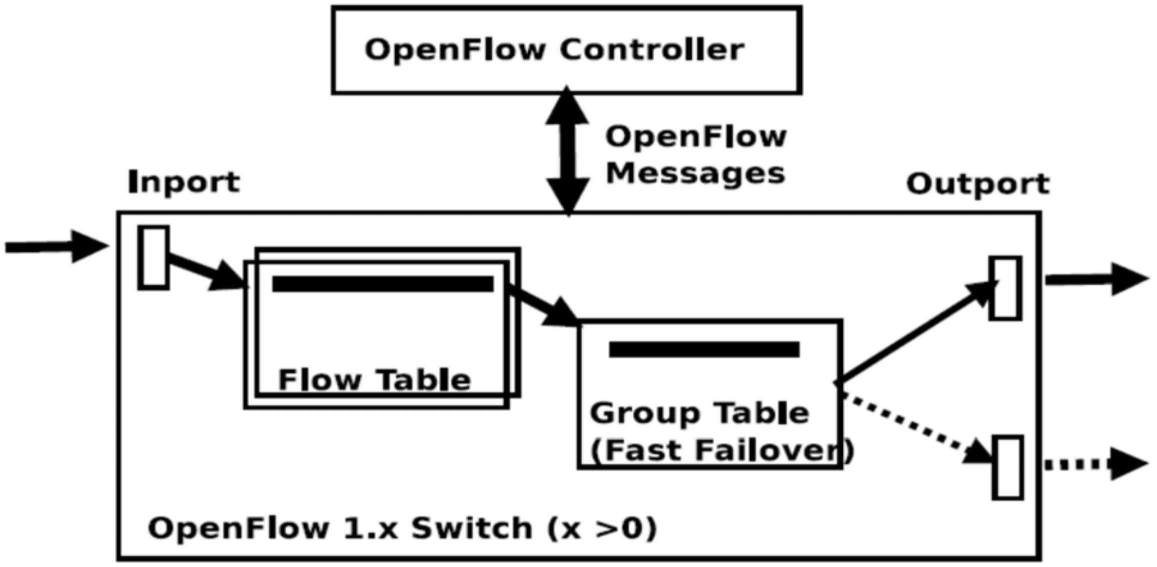
1. Switchover mechanisms from a faulty link to a normal link
   - ◆ Switchover mechanism in a single OpenFlow switch is the essential component to implement resilient backbone networks.
   - ◆ Switchover time is a part of the network recovery time that includes all the time from failure detection to path restoration on a end-to-end basis.
   - ◆ The network recovery time should be less than 50ms for provider networks
   - ◆ OpenFlow supports a wide variety of switchover mechanisms not available in existing network recovery mechanism.
   - ◆ We investigate some of the OpenFlow–specific and OpenFlow-integrated switchover mechanisms and evaluate their switchover performance Implementation.
     - ■ OpenFlow–specific switchover mechanisms: FAILOVER GROUP TABLE and SELECT GROUP TABLE- based implementations, using local states of OpenFlow switches without involvement of remotely located controllers
     - ■ OpenFlow-integrated switchover mechanisms: OpenFlow with Multipath TCP (MTCP) in the TCP layer

Influential Factors to Enable Fast Network Recovery Using SDN/OpenFlow Technologies (2)

2. Communication delay between SDN controllers and switches
   - Communication delay directly affects the network recovery time
   - We analyze the communication delays under a realistic network topology
3. Global view of the network
   - Global view of the network is necessary to find any available network resources (paths, links) to restore all the end-to-end paths.
   - Maintaining a global view is achieved in the conventional IP network by the IP routing protocols like OSPF and BGP that suffer from slow convergence time
   - Maintaining and updating a global view is complicated especially for multiple SDN controllers to keep the network scalable, but several solutions have been proposed.
   - In our implementation, we assume that a global view is maintained among the SDN controllers either by the existing IP routing protocols or any new routing protocols for SDN/OpenFlow.
   - We evaluate the overall network recovery performance under the above assumption.

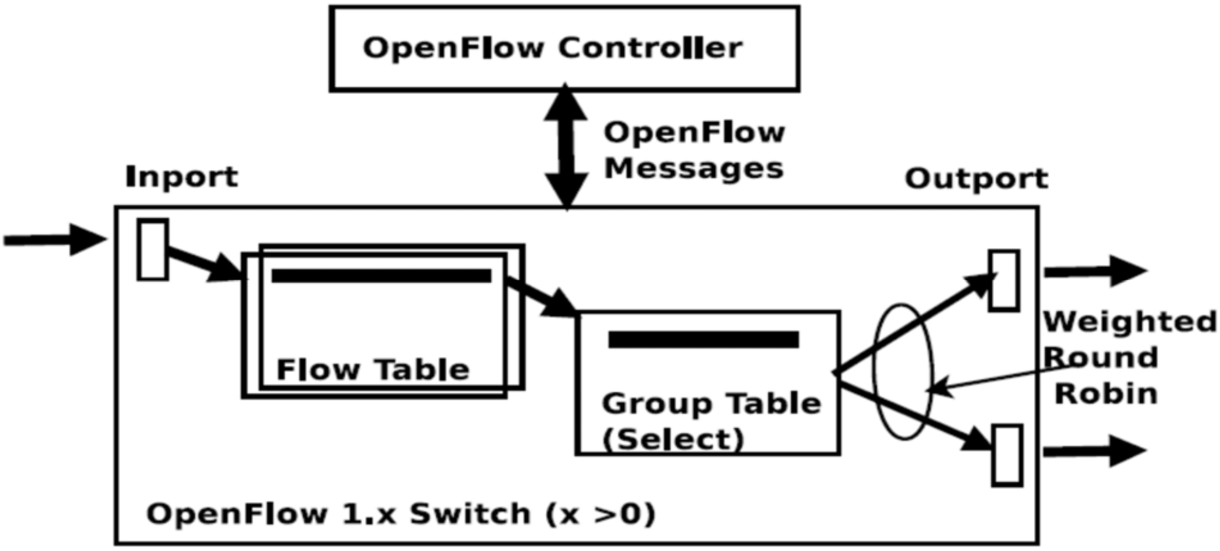
Fast Local Switchover Mechanism (1): FAST FAILOVER GROUP TABLE

- Two Types of Group Table in OpenFlow:
  - FAST FAILOVER GROUP TABLE for active/standby mode
  - SELECT GROUP TABLE for active/active mode



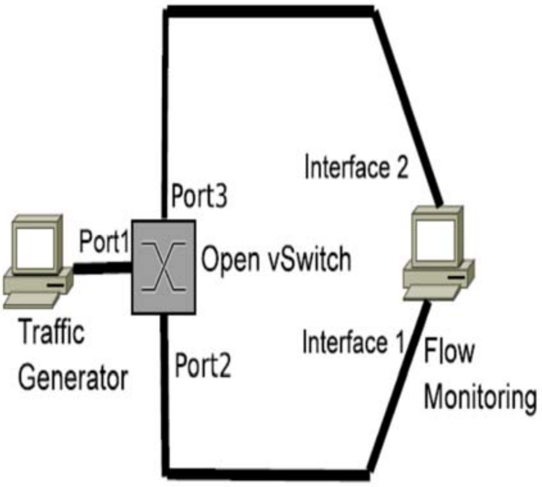
Fast Local Switchover Mechanism (2): SELECT GROUP TABLE

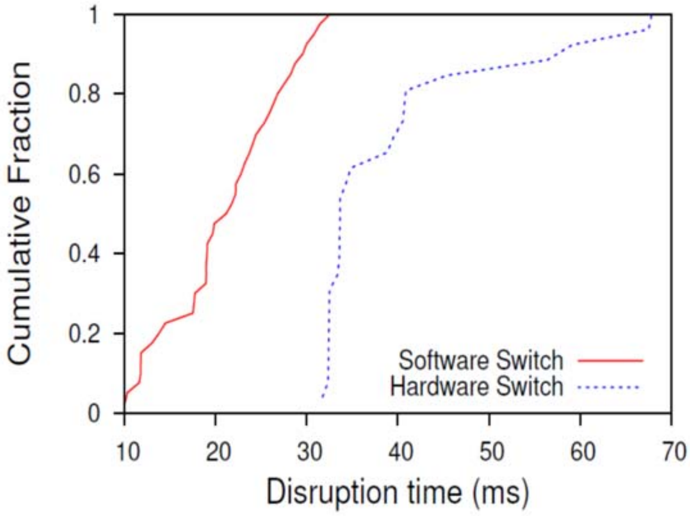
- SELECT GROUP TABLE allows a single data flow to be divided into multiple subflows, each with a different path (outport)
- SELECT GROUP TABLE achieves a better resource allocation and less packet loss than FAST FAILOVER GROUP



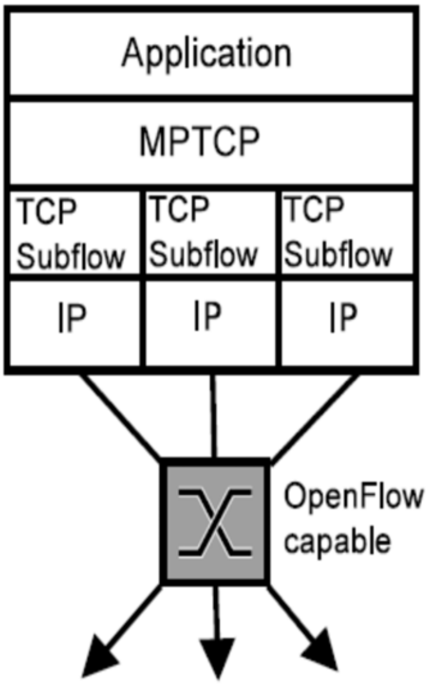
Implementation and Evaluation of Fast Local Switchover Mechanisms on Two Different Platforms

- Software switch: Open vSwitch (OVS) on a Linux PC to support the FAST FAILOVER GROUP TABLE
- Hardware switch: Open vSwitch (OVS) mode on the hardware switch (Pica8 P3295) to support the FAST FAILOVER TABLE
- Average switchover time (Disruption time) : 21.1 ms for software switch and 39.5 ms for hardware switch





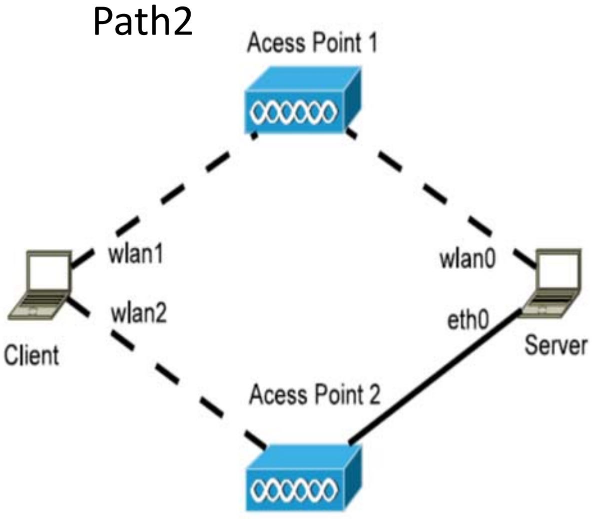
Fast Local Switchover Mechanism (3): Multipath TCP (MPTCP) Integrated with OpenFlow

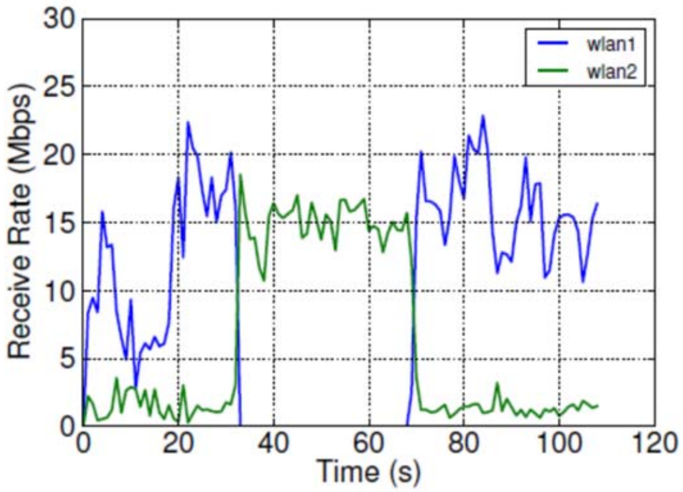


- MPTCP creates and maintains multiple active paths for an end–to-end connection
- Divides the TCP flow into multiple active TCP subflows
- Each subflow may go through a different path to achieve better resilience
- OpenFlow achieves fast switchover among multiple active paths when some of the paths fails
- MPTCP is standardized by IETF (Internet Engineering Task Force), does not need to modify existing applications

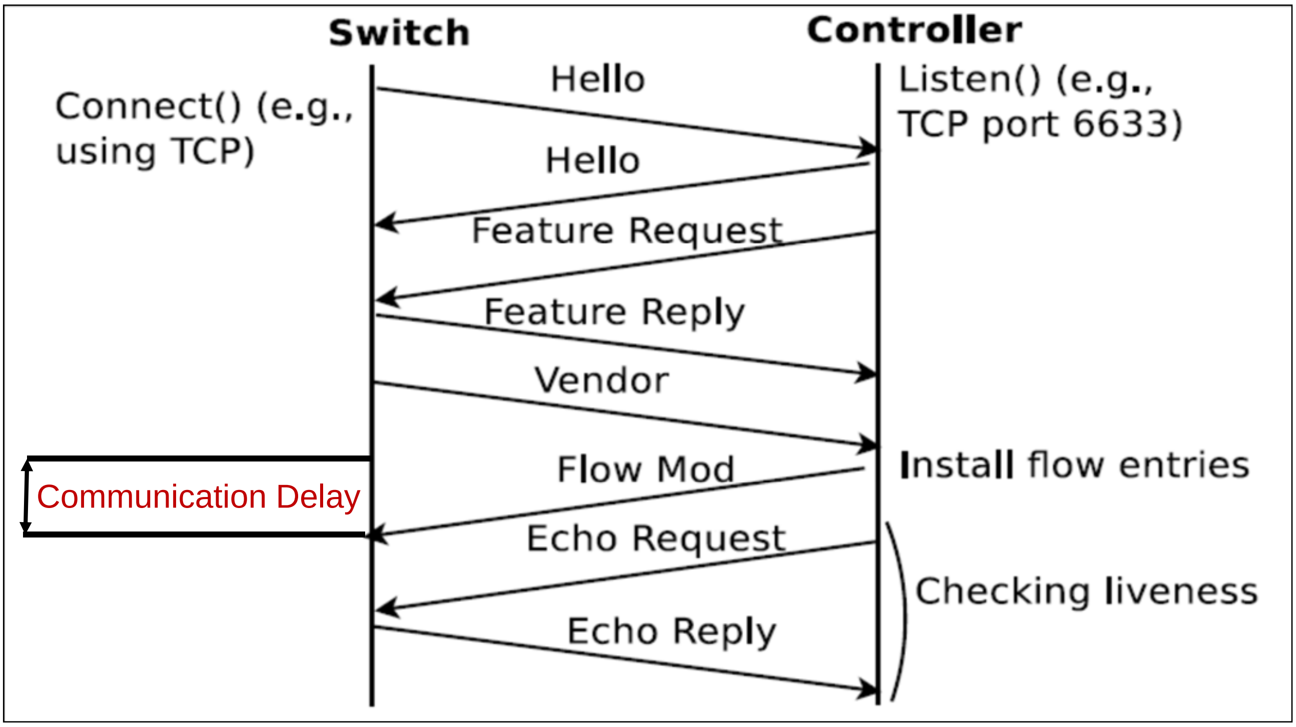
Implementation and Evaluation of MPTCP on WiFi Network Environment

- Two subflows each with a different path through a different WiFi access point
- When the path1 fails, path2 keeps transferring the added path1 traffic to achieve seamless handover from Path1 to Path2

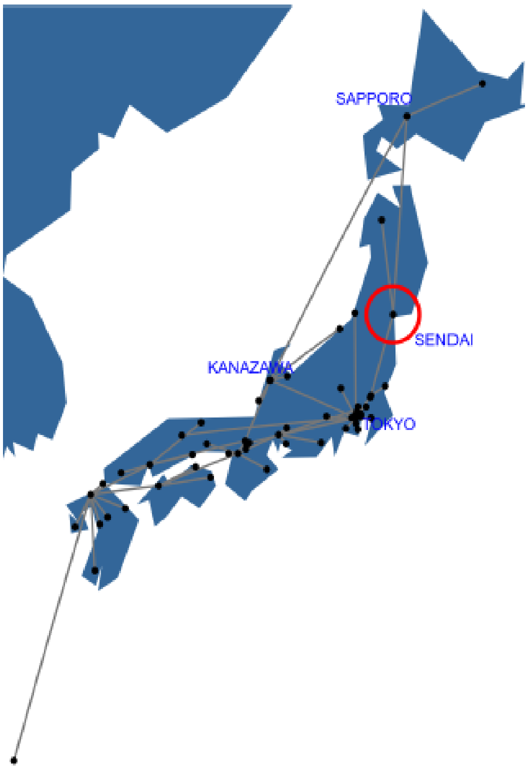




Communication Delay between Controllers and Switches



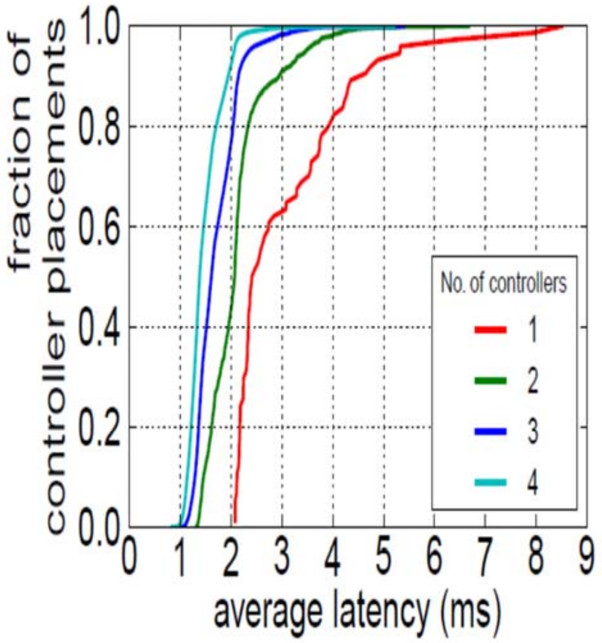
Analysis of Communication Delay between Controllers and Switches

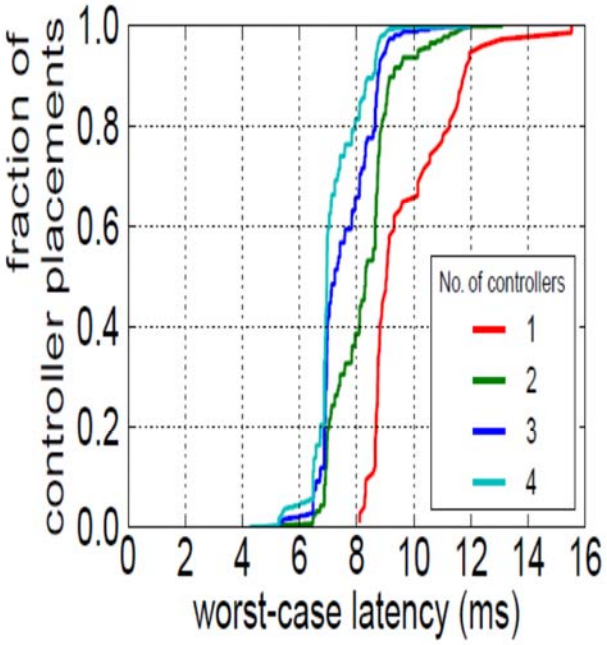


- SINET3 topology is used to evaluate communication latencies between controllers and switches.
- SINET3: the previous version of current SINET4, a Japanese national research and education network.
- Two latency metrics under 2/3 propagation delay of light speed :
  - Average Latency for allocation of controllers
  - Worst-Case latency: the maximum propagation delay between nodes and controllers

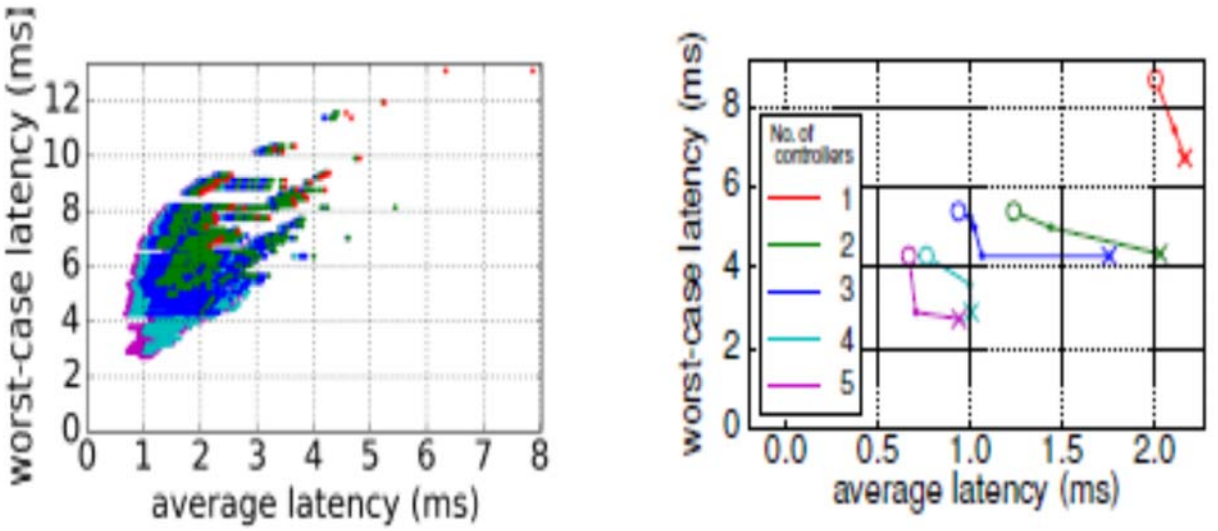
Evaluation of Average and Worst-Case Latencies

- The more controllers, the lower latencies
- Should carefully choose the location of controller





Optimal Values of Average and Worst-Case Latencies

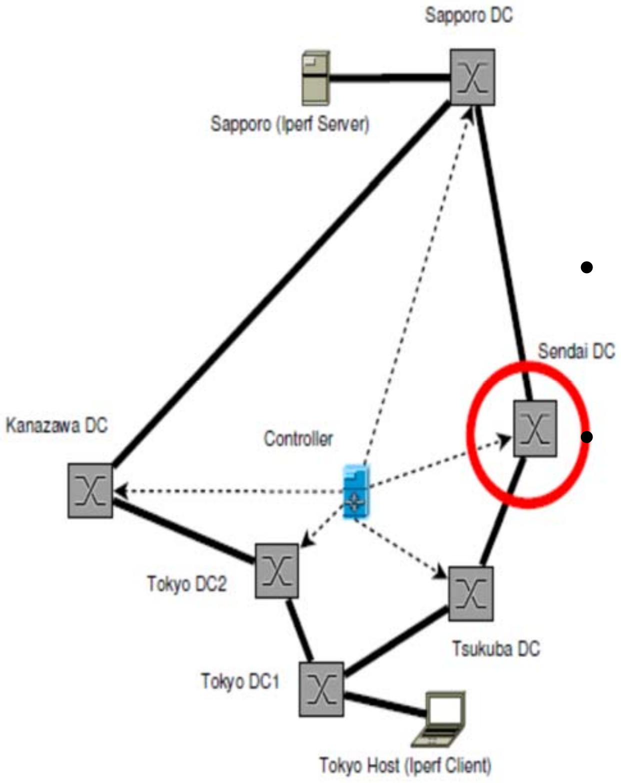


(a) All the combination

(b) Pareto optimal

- These latencies are much smaller, compared with the general requirement of 50ms failover time

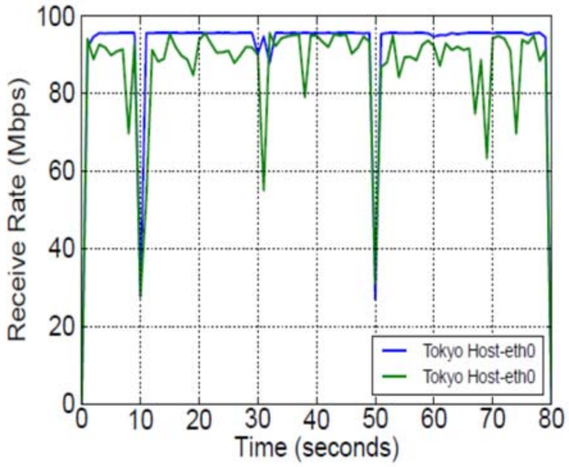
Network Recovery Simulation



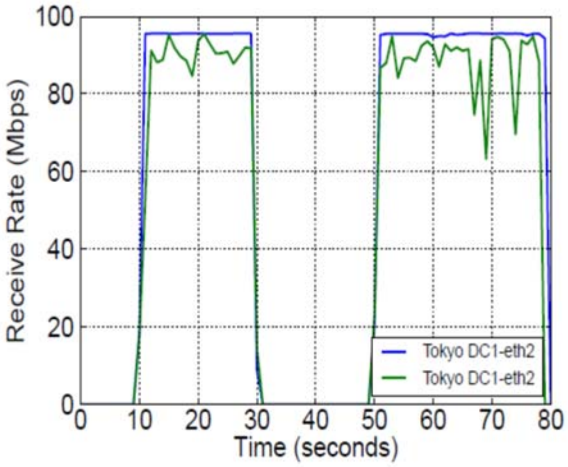
- Mininet 2.0, the virtual network simulator is used to investigate the overall behaviors of link failure recovery under a realistic scenario and SINET3 topology.
- The worst case latencies between the controller and switches are assumed
- When a link failure occurs on the main path, the controller software (POX) should install new rules to the switches to switchover the traffic flow from the faulty path to the backup path.

Network Recovery Simulation Result

- When a link failure occurs at 10th and 30th seconds, the traffic flow is effectively turned from the faulty path to a new path thanks to the POX controller’s global view of the network



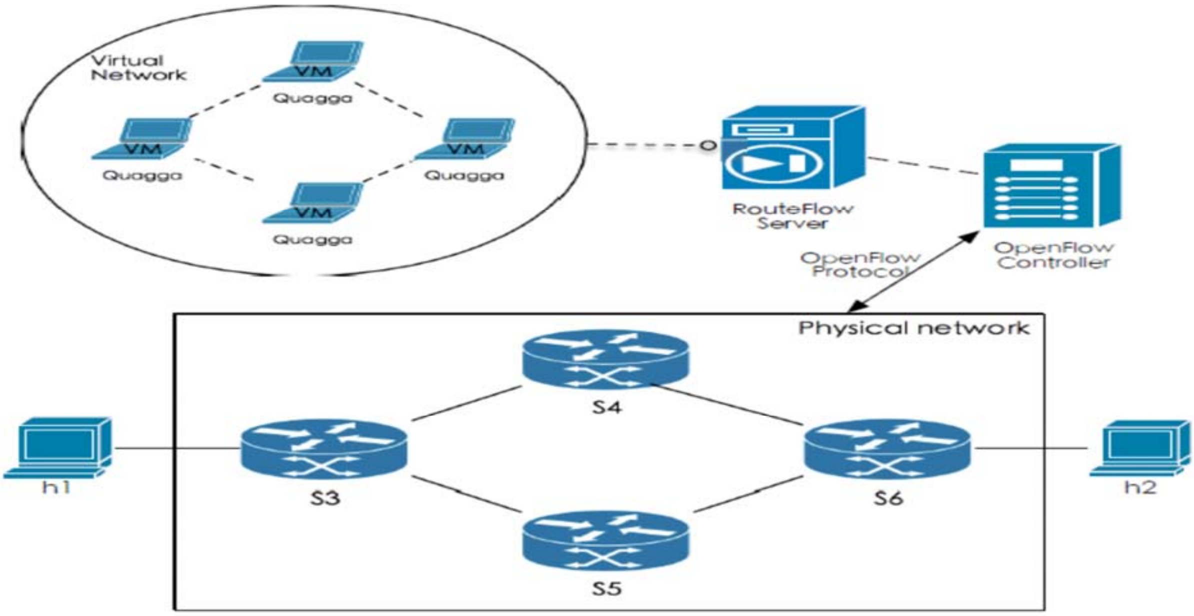
Receive traffic (i.e., goodput) at Tokyo-Host



Goodput on the alternative path

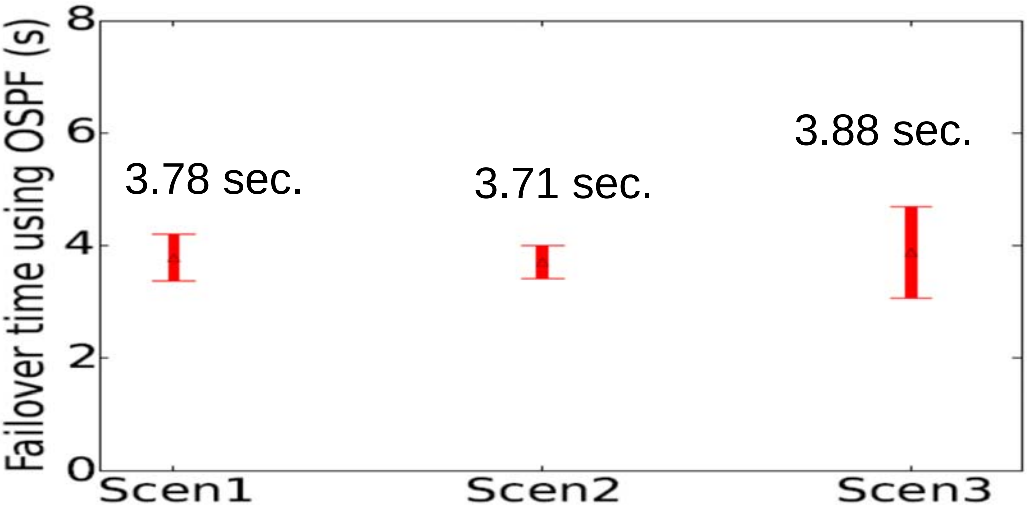
Implementation and Evaluation of Network Recovery by SDN integrated with IP Routing

- Network recovery that integrates SDN with conventional IP routing is implemented and evaluated on the Mininet simulator and a real testbed with physical OpenFlow switches.



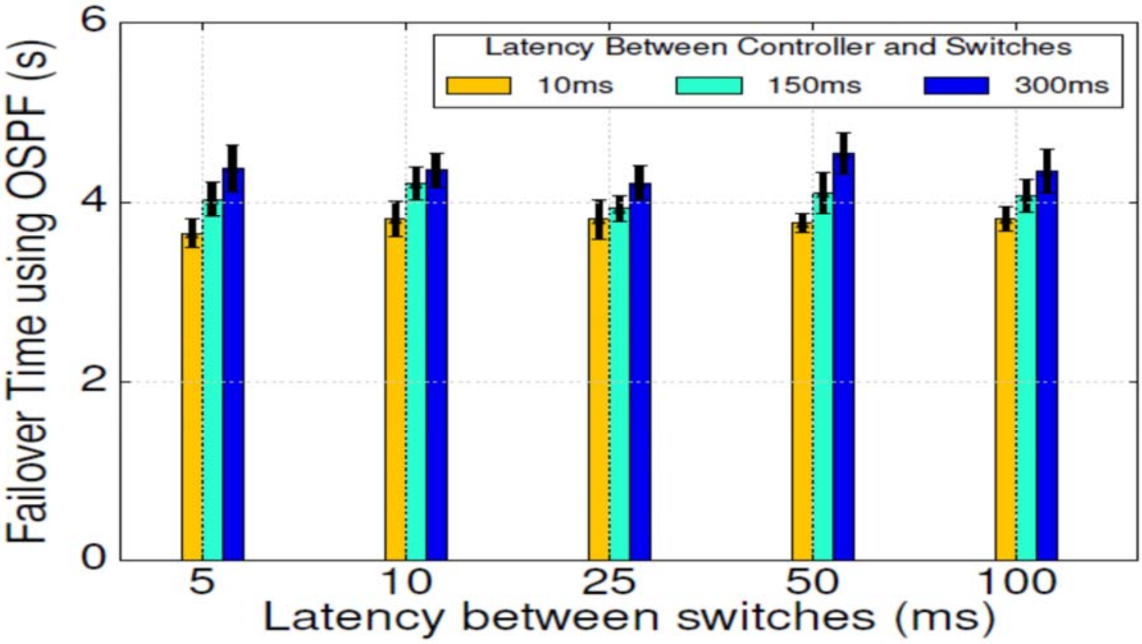
Network Recovery Time Comparison under OSPF protocol

- Scen1: Mininet simulation running Route Flow (for OSPF)
- Scen2: Pica8 switches running Route Flow (for OSPF)
- Scen3: Pica8 switches running conventional IP routing protocol (L2/L3 OSPF)
- All results are close to the dead-interval of 4 seconds.



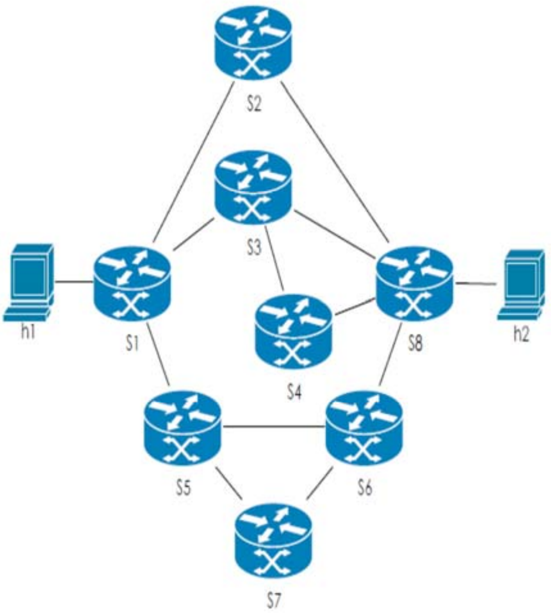
Effect of Communication Delays on Network Recovery Time

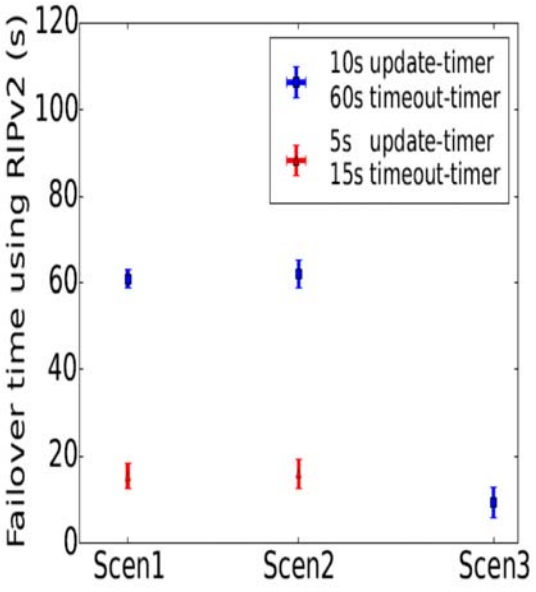
- The larger the communication delay, the longer the network recovery time



Evaluation of Network Recovery Time under Multiple Link Failures

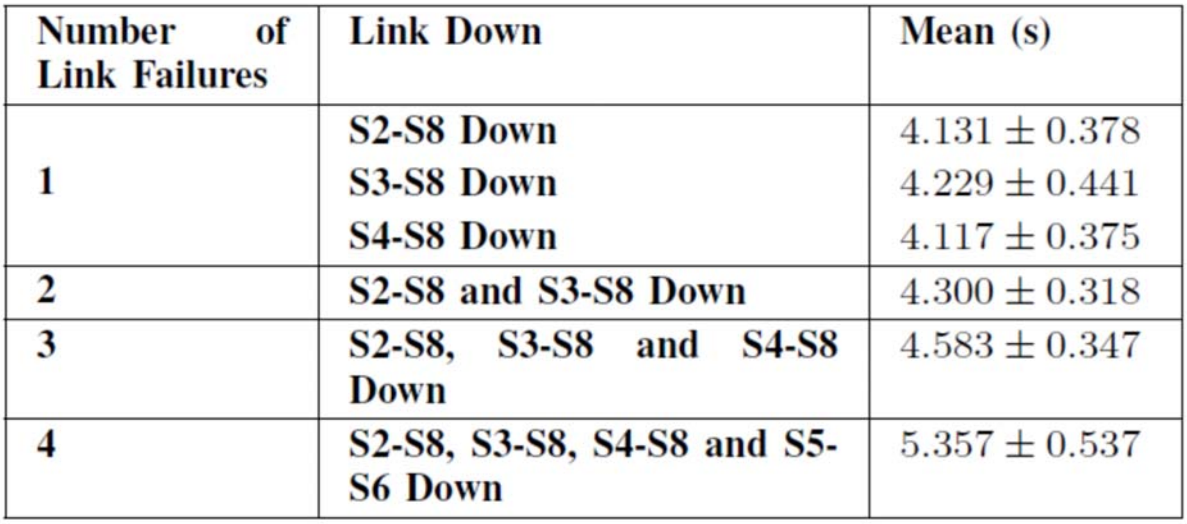
- A more complex network topology with 8 switches, assuming multiple link failures
- 5 redundant paths from source to destination





Evaluation Result of Network Recovery Time under Multiple Link Failures

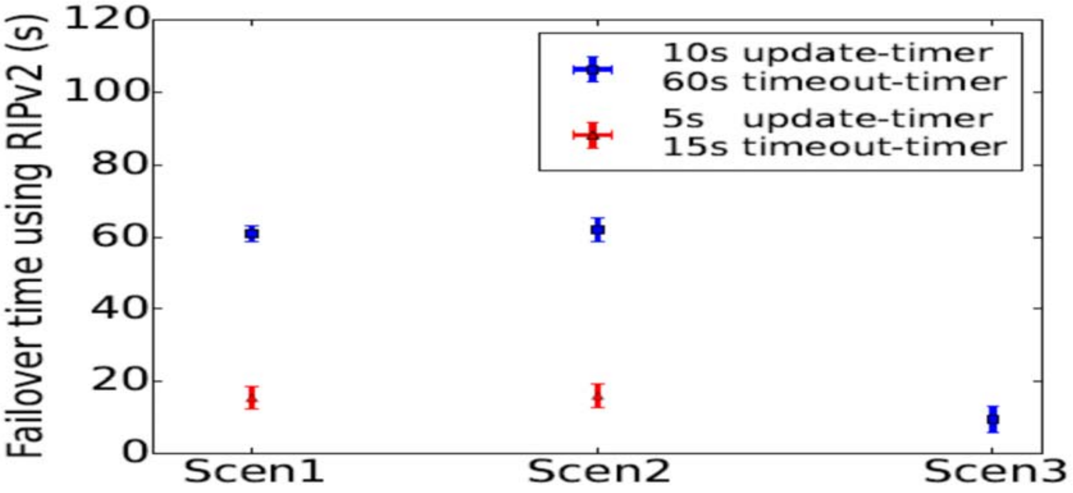
- The network recovery time of multiple link failures is roughly 10% larger than that of a single link failure



Number of Link Failures	Link Down	Mean (s)
1	S2-S8 Down S3-S8 Down S4-S8 Down	4.131 ± 0.378 4.229 ± 0.441 4.117 ± 0.375
2	S2-S8 and S3-S8 Down	4.300 ± 0.318
3	S2-S8, S3-S8 and S4-S8 Down	4.583 ± 0.347
4	S2-S8, S3-S8, S4-S8 and S5-S6 Down	5.357 ± 0.537

Network Recovery Time Comparison under RIPv2 protocol

- Scen1: Mininet simulation running Route Flow (for RIPv2)
- Scen2: Pica8 switches running Route Flow (for RIPv2)
- Scen3: Pica8 switches running conventional IP routing protocol (L2/L3 RIPv2)
- All Scen1 and Scen2 results are close to the timeout-timer of 15 seconds or 60 seconds.
- Scen3 result comes from a different failure detection mechanism from Scen1 and Scen2

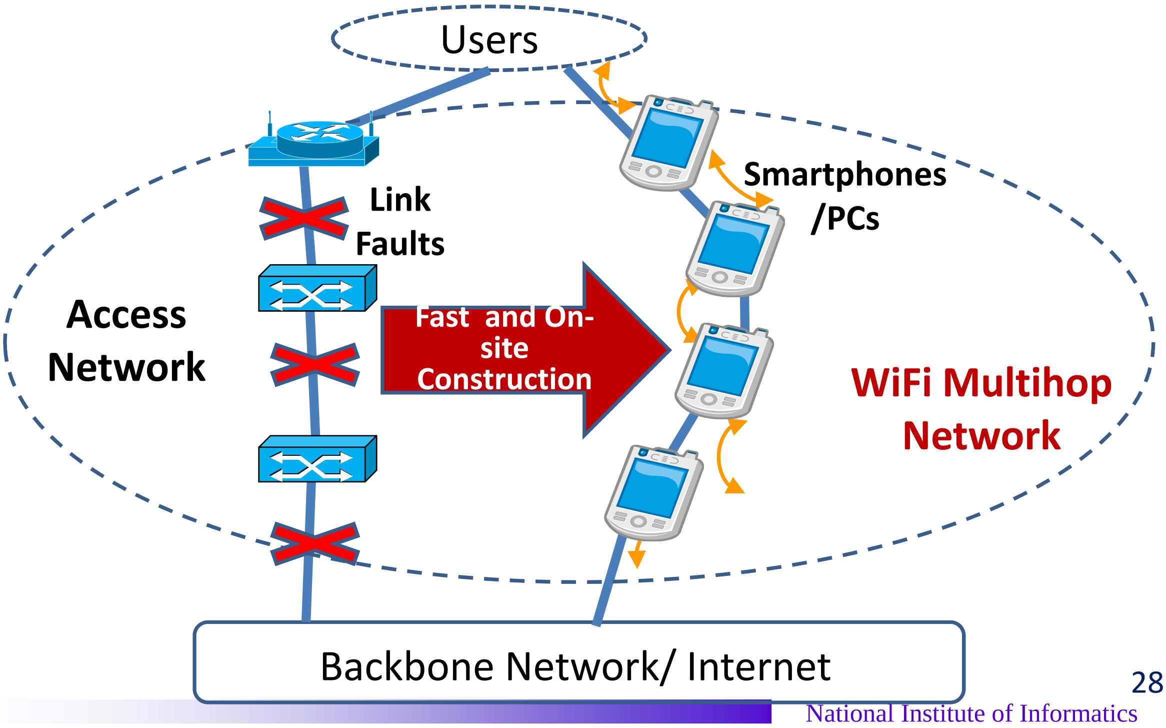


Summary of Resilient Backbone Network Evaluation

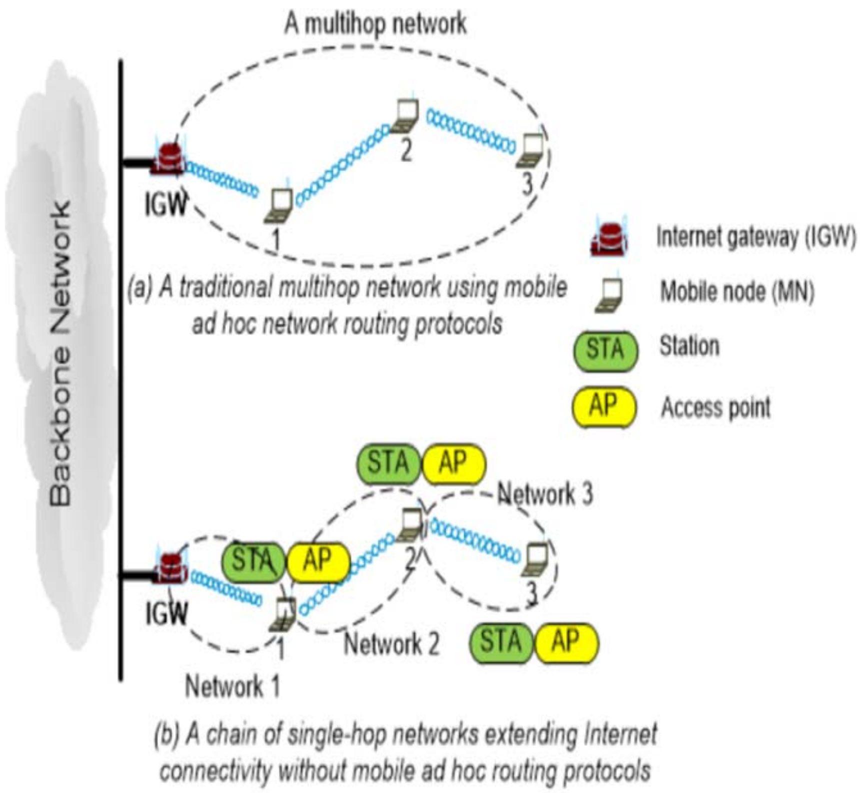
- SDN/OpenFlow technology is technically feasible.
  - It can offer a wide variety of switchover mechanism with fast switchover time of 20 to 40 miliseconds under current implementation technologies
  - The overall network recovery time ranges from 20 ms to 60 seconds, largely depending on the employed routing protocol and its timer values to update the global view of network

Resilient Access Network

- We propose to apply WiFi multihop network technologies for access networks to provide internet access services.



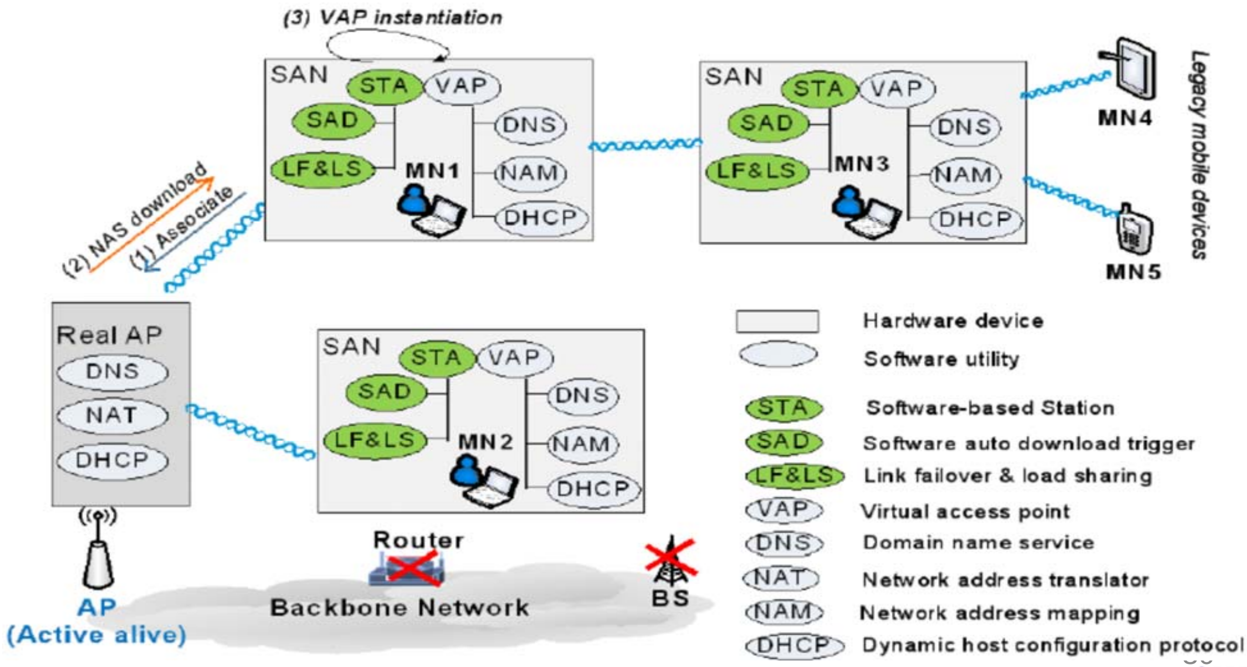
Multihop Communication Abstraction


A multihop network
Backbone Network
IGW
1
2
3
(a) A traditional multihop network using mobile ad hoc network routing protocols
Internet gateway (IGW)
Mobile node (MN)
STA Station
AP Access point
STA AP
Network 3
Network 2
Network 1
(b) A chain of single-hop networks extending Internet connectivity without mobile ad hoc routing protocols

- A conventional multihop access network requires each node to implement a traditional ad hoc routing protocol and maintain the routing information for all nodes.
- The proposed multihop communication abstraction allows a chain of single hop WiFi network and does not maintain multihop routing tables

Tree Structure of Multihop Wireless Access Network

- A network auto-configuration software (NAS) is downloaded to transform each node into the WiFi virtual access point (VAP) and the WiFi station (STA), finally forming a tree-structured multihop network

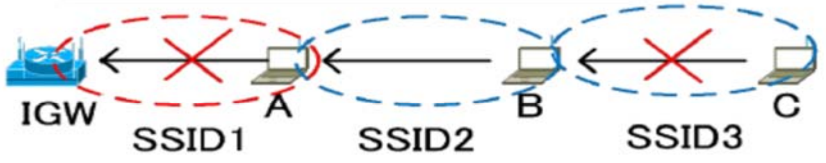


Network Reconfiguration Support: Connectivity Status Table (CST)

- Each node manages the CST containing
  - the status ("Connected"/"Disconnected") of all the upward links over the path from the Internet gateway (IGW) and its own node.
  - the Hop_count that represent the distance from IGW
- Each CST is automatically updated and propagated downward when the link status changes

A's CST

SSID	Status	Hop_count
SSID1	"D"	1



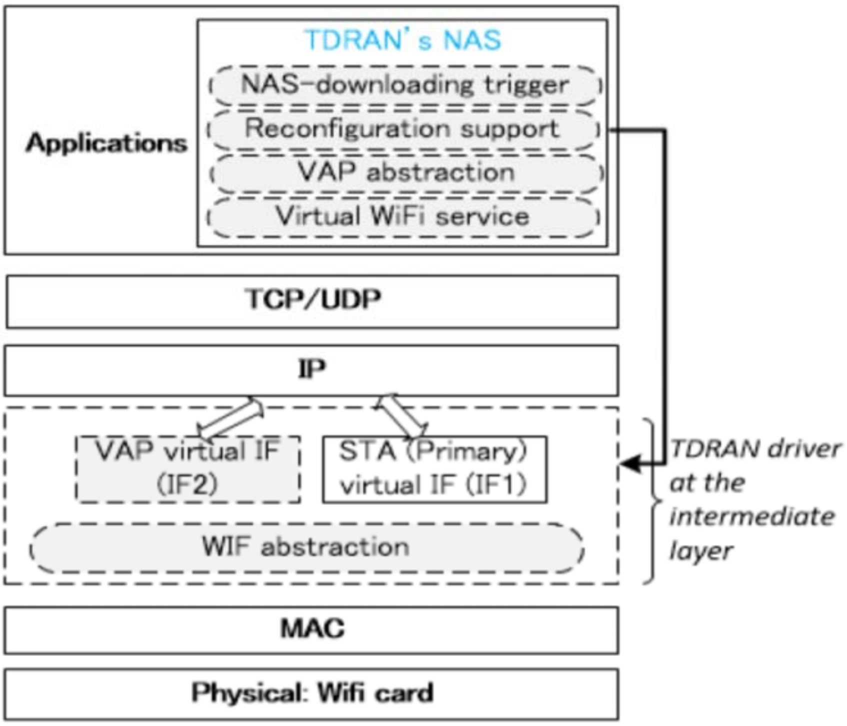
B's CST

SSID	Status	Hop_count
SSID1	"D"	1
SSID2	"C"	2

C's CST

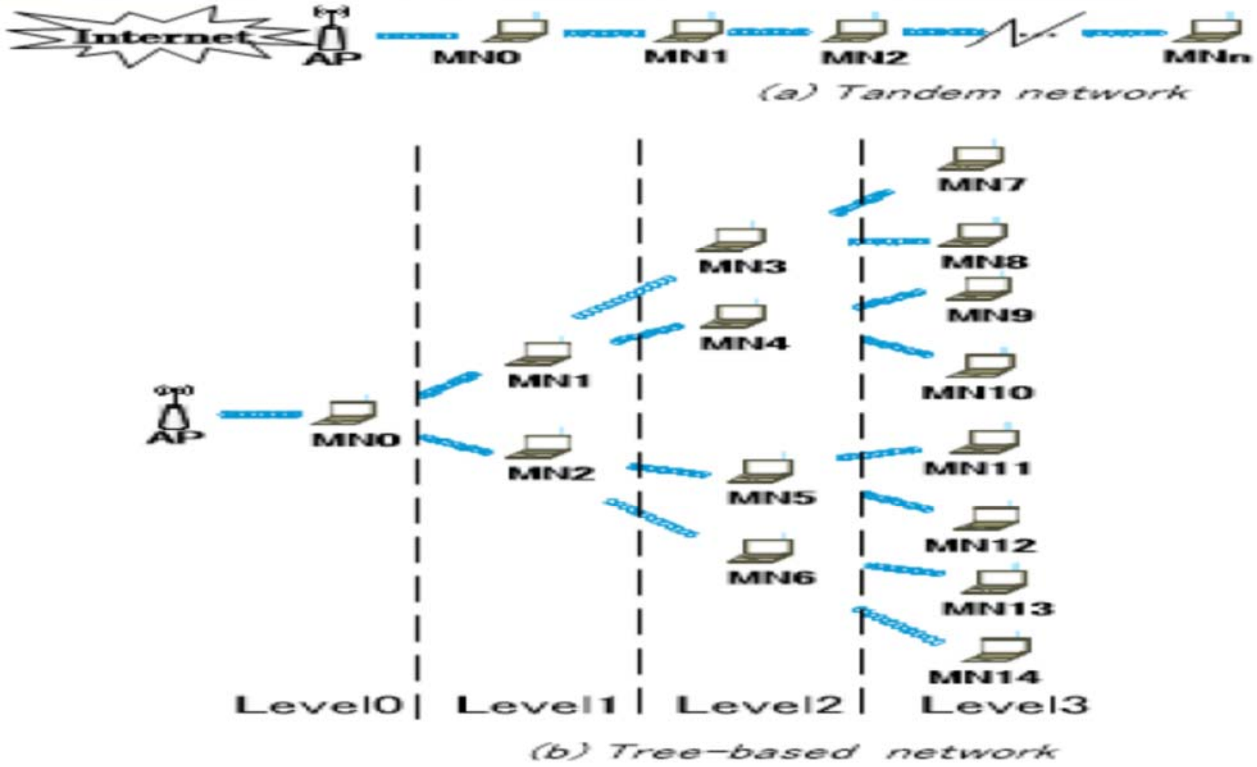
SSID	Status	Hop_count
SSID1	"D"	1
SSID2	"C"	2
SSID3	"D"	3

Network Auto-Configuration Software (NAS) Components in Each Node



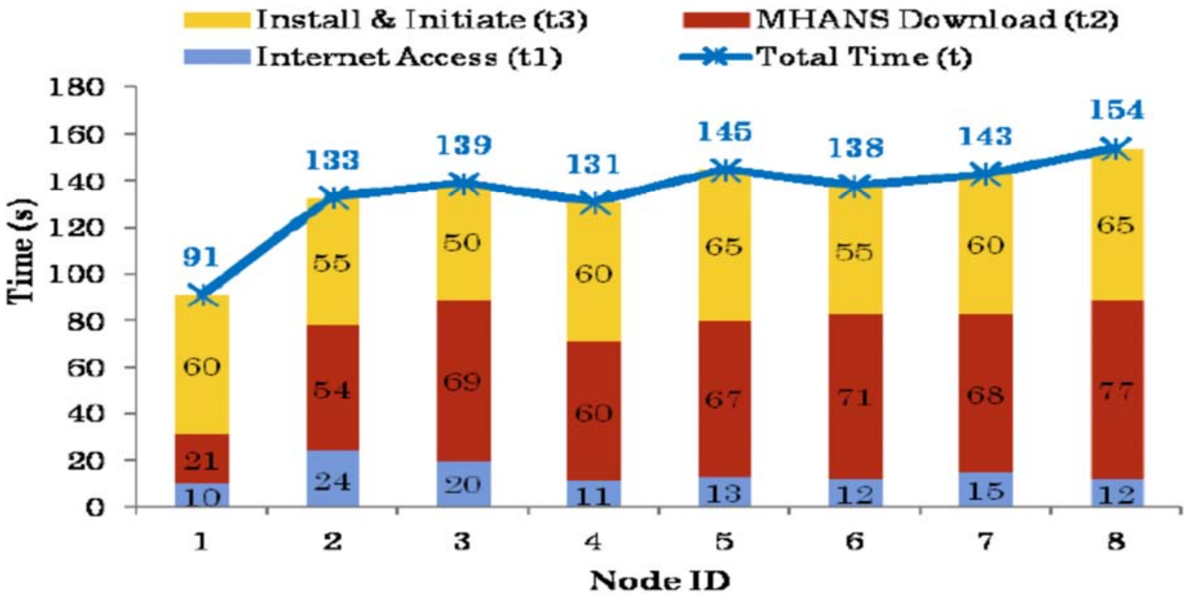
- WiFi Abstraction for multiple logical WiFi interfaces
- VAP abstraction to act as Virtual access point (VAP)
- Reconfiguration support for Connectivity Status Table (CST)
- NAS-downloading trigger to download the NAS
- Only NAS is necessary to construct the WiFi multihop network

Field Experiments at Iwate Prefectual University and Ishinomaki Senshu University



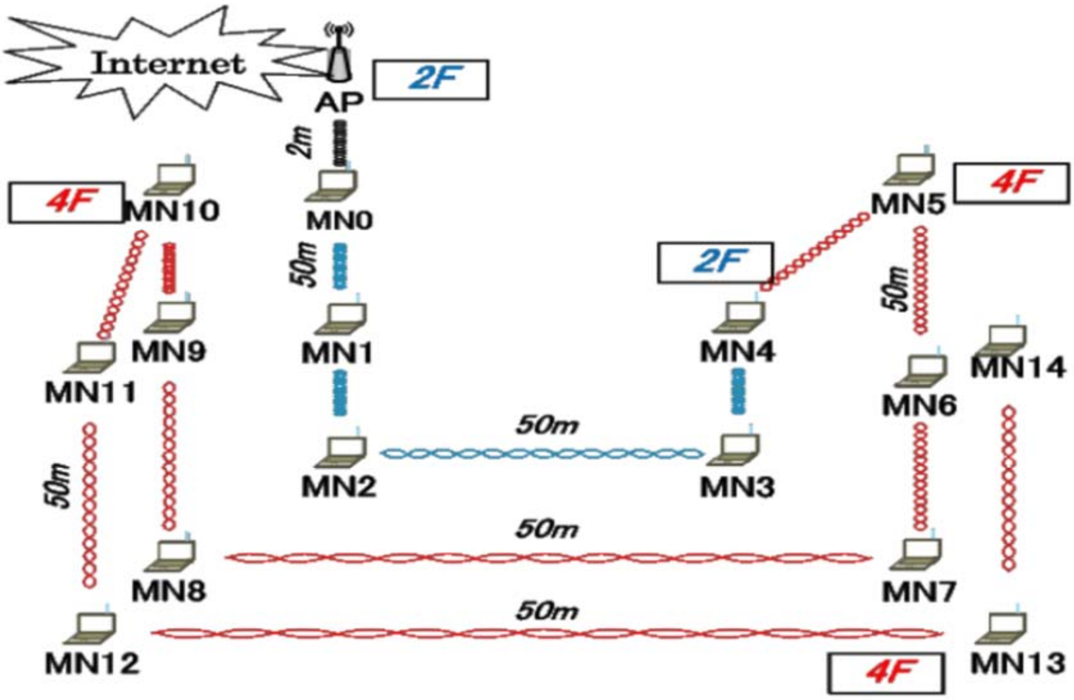
Network Set-up Time

- All the tandem connected nodes were set up in parallel by the university students
- The network setup time is less than 154 seconds in 8 hop network: quick Enough for emergency response



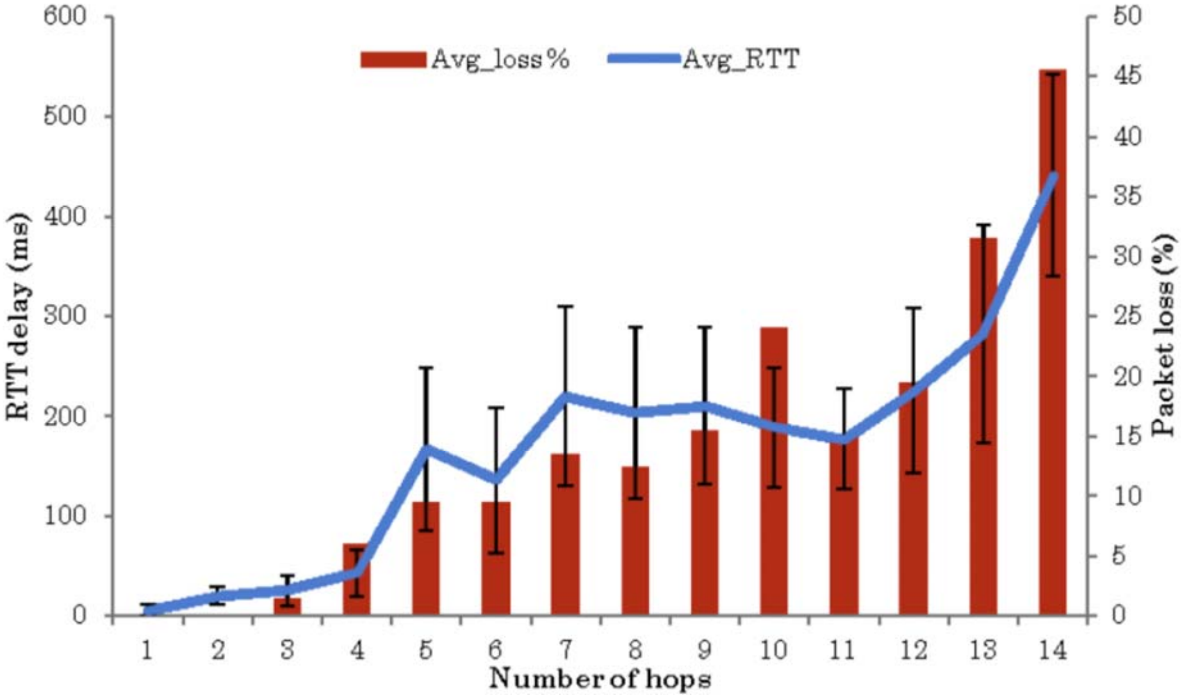
Experiment of Indoor Tandem-Connected Network with 50m Hop Distance

- The experiment was made from the 1st floor to the 4th floor inside the building of Iwate Prefectural University



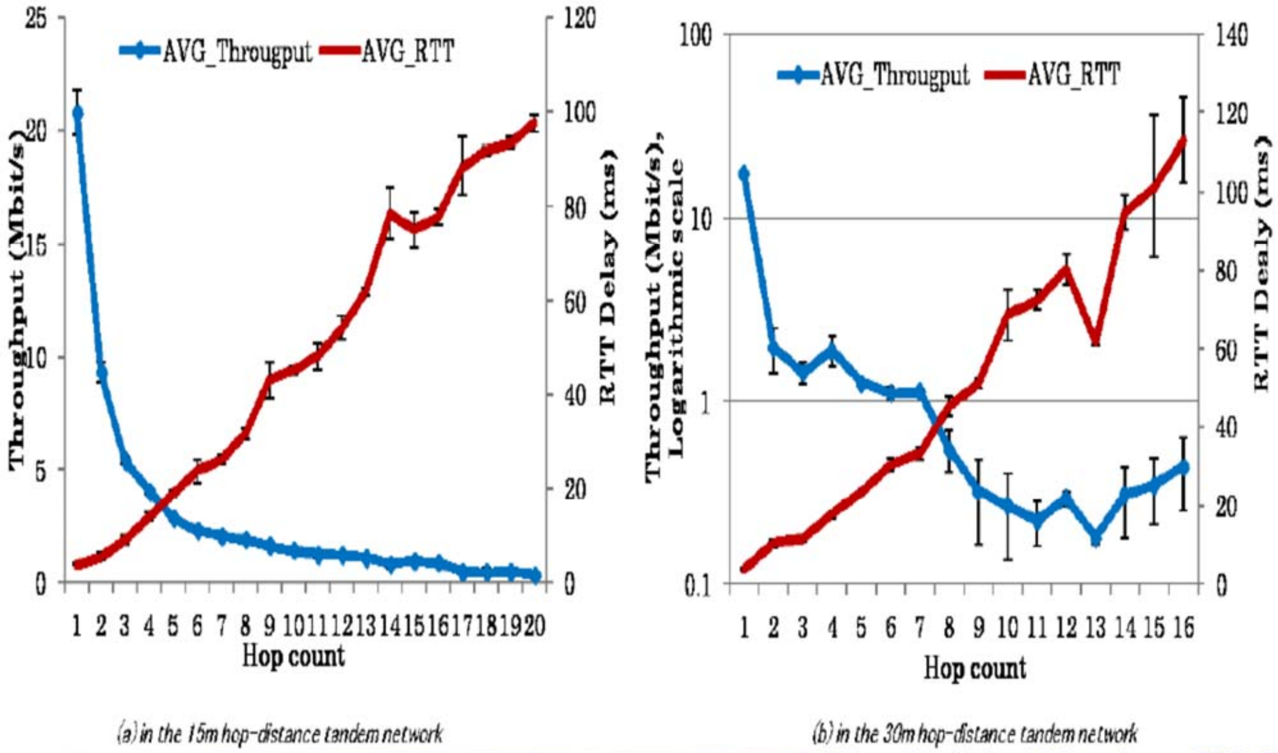
Round Tip Time (RTT) and Packet Loss of Indoor Tandem-Connected Network with 50m Hop Distance

- 20% Packet loss and 200ms round trip time (RTT) in 12 hops were still acceptable for ordinary Internet applications (Web browsing and Skype)



Round Tip Time (RTT) and Throughput of Outdoor Tandem-Connected Network

- 15-m Hop Distance
- 30-m Hop Distance

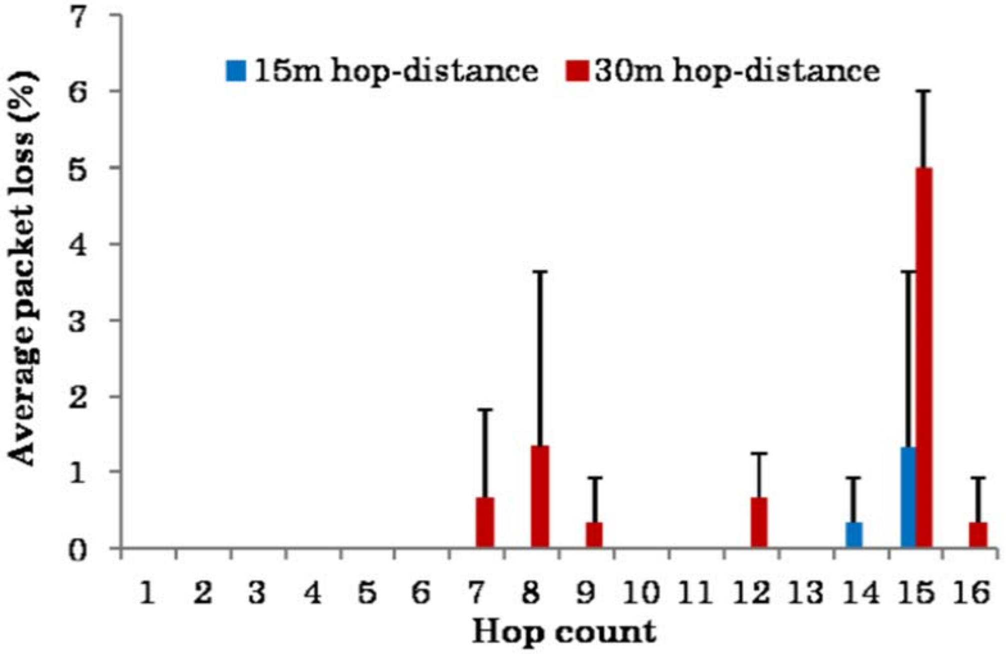


(a) in the 15m hop-distance tandem network

(b) in the 30m hop-distance tandem network

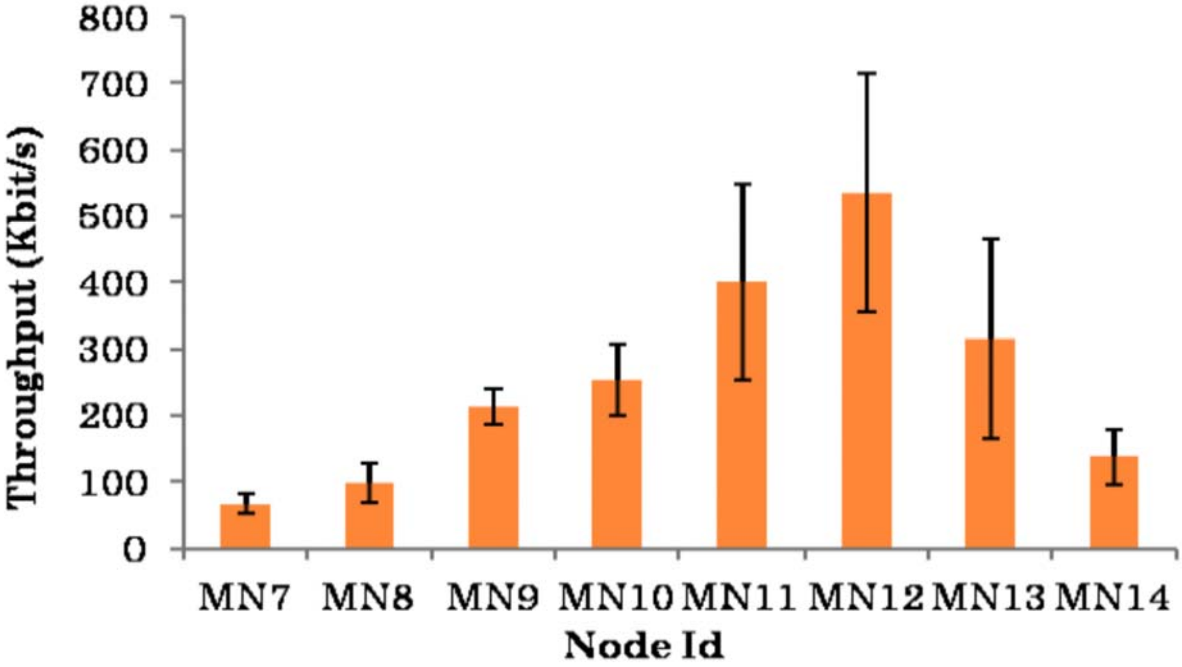
Packet Loss of Outdoor Tandem-Connected Network

- The largest packet loss is only 5% at Node 15 with 30m hop distance: acceptable enough for VoIP services and web browsing



Throughput of Outdoor Tree-Structured Network when Leaf Nodes concurrently transmit the packets

- The throughput are different at different nodes
- The lowest throughput of around 100Kbps was acceptable for Web browsing



Summary of Resilient Access Network Evaluation and Conclusion

- Resilient Access Networks: WiFi multihop access network is feasible for real deployments.
  - It can cover a large area of one kilometer in distance for internet access with up to 20 hops in the disaster area
  - It allows ordinary people or volunteers in the disaster area to set up the network easily by themselves.
- Integrating SDN/OpenFlow based backbone network with WiFi multihop access network could enable the network to become more resilient to provide end-to-end seamless non-stoppable services