APAN/PRAGMA/Disaster Mitigation Workshop,
Manila, Philippine, January 27, 2016

Resilient Networking Solutions for Prompt Disaster Recoveries

*Shigeki Yamada[*1], Quang Tran Minh[*2], and Kien Nguyen[*3]*

*[*1]National Institute of Informatics (NII), Japan*

*[*2] Ho Chi Minh City University of Technology, Vietnam*

*[*3] National Institute of Information and Communications Technology (NICT), Japan*

大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

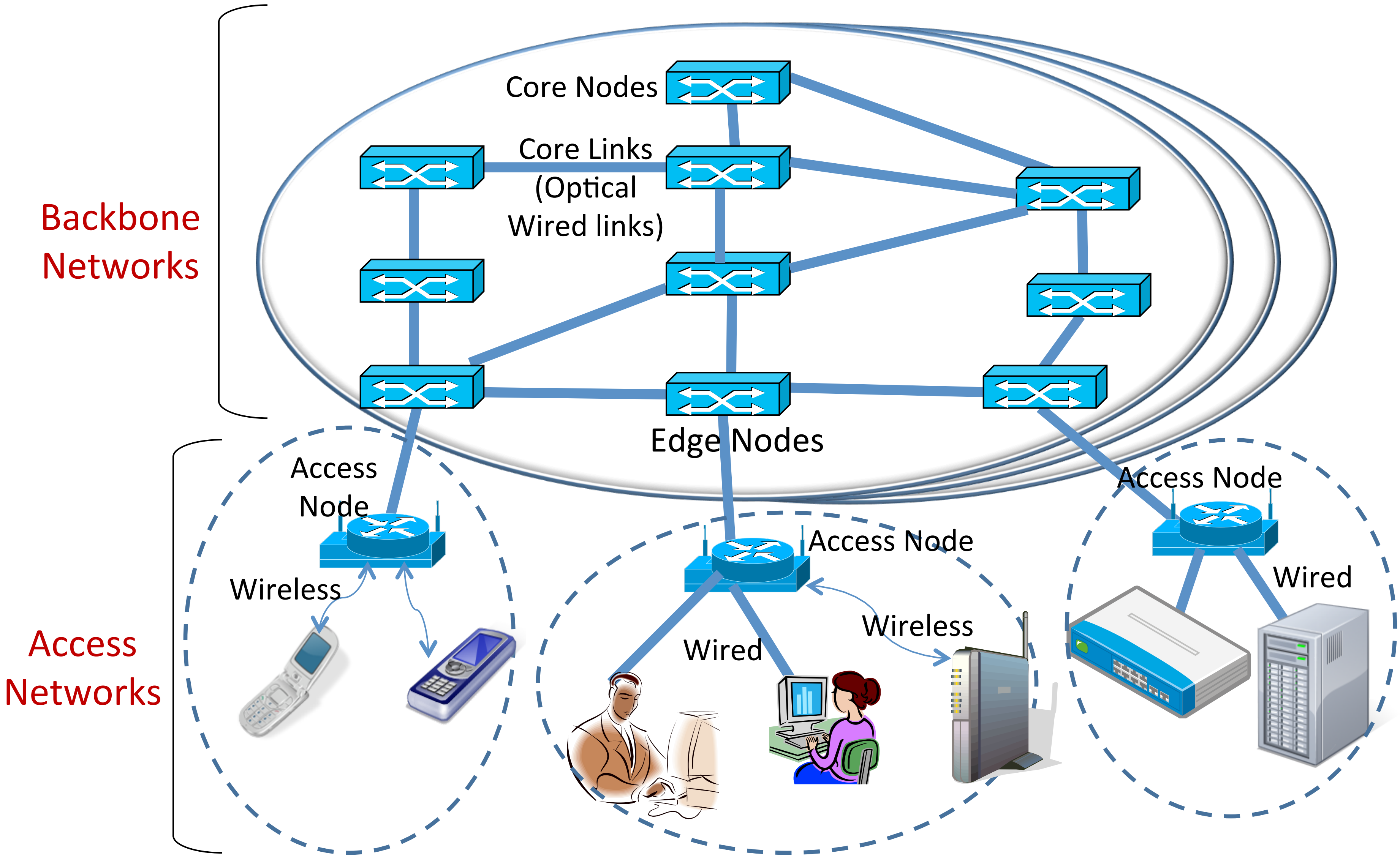
Introduction and Background (1)

- The Great East Japan Earthquake on March 11, 2011 with magnitude 9.0 (Mw) undersea off the coast of Japan
- Damage Situations (as of October 2015)
  - 15,893 deaths, 6,152 injured, 2,567 people missing,
  - 228,863 people living away from their homes in either temporary housing or due to permanent relocation.
  - 121,747 buildings totally collapsed, 277,679 buildings 'half collapsed', 725,858 buildings partially damaged.
- Many different types of large natural disasters ensue worldwide.
- Floods/landslides, earthquakes, cyclones, typhoons, storms, tornadoes, tsunamis, avalanches, blizzards, heat waves, volcanic eruptions, wildfires etc. They may never be gone.
- Academic communities should anyhow contribute to alleviating damages of natural disasters.

Introduction and Background (2)

- This presentation is a summary of three-year Resilient Network Research Project promoted under JSPS Resilient Life Space Umbrella Project.
- When 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



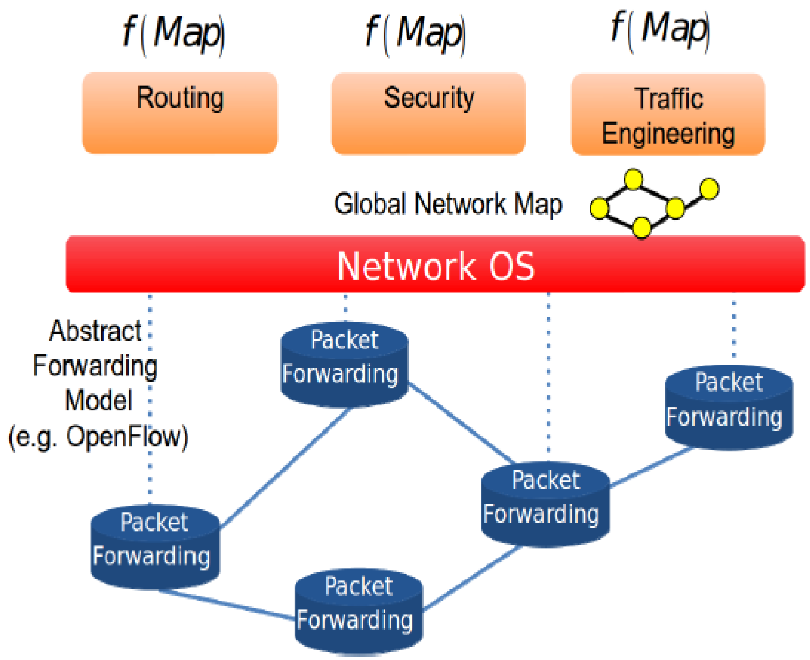
Basic Approaches to Make Backbone Networks & Access Networks More Resilient

- 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.
- Access networks
  - located close to users, 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, network users in the disaster area could construct their access network, using any available devices more quickly and more easily.

Requirements to Resilient Backbone Networks

- Network resilience could be measured by the Network Recovery Time with two major time components.
  1. Failure Detection Time: detection of alarms and alerts to locate network faults
  2. Switchover Time: disables a failed port, enables another, reroutes traffic around a failed switch or router
- For failure detection, existing detection technologies like BFD (Bidirectional Forwarding Detection) could be utilized.
- For switchover, we apply SDN (Software Defined Networking) / OpenFlow technology because SDN/OpenFlow has a potential capability to provide more programmability and flexibility to respond faster to network situational changes than existing technologies (like MPLS).
- For Network Recovery Time, at most 50 ms is considered tolerable to complete path restoration, in the provider networks.

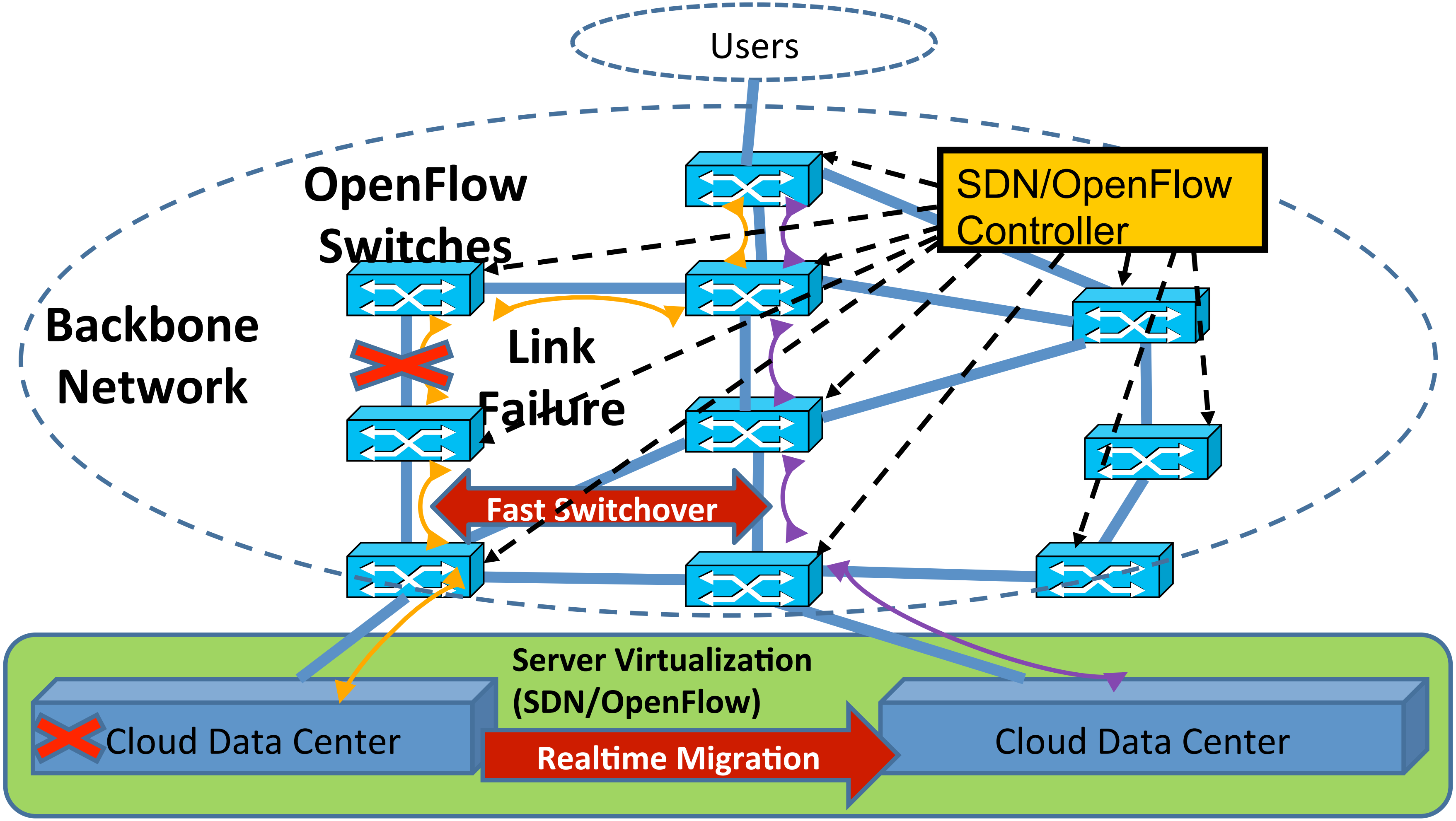
SDN/OpenFlow for Backbone Networks



- Network Operating System (NOS) with a global view of network controls forwarding hardwares via OpenFlow protocol
- Network intelligence is on top of NOS as applications.
- SDN/OpenFlow provides an easier way to manage and automate networks by separating the control plane and the data plane.

Goal of Resilient Backbone Network

- To provide non stoppable end-to-end services in the various critical environments, including link/path and node failures



Three Research Issues for Fast Network Recovery Using SDN/OpenFlow Technology (1)

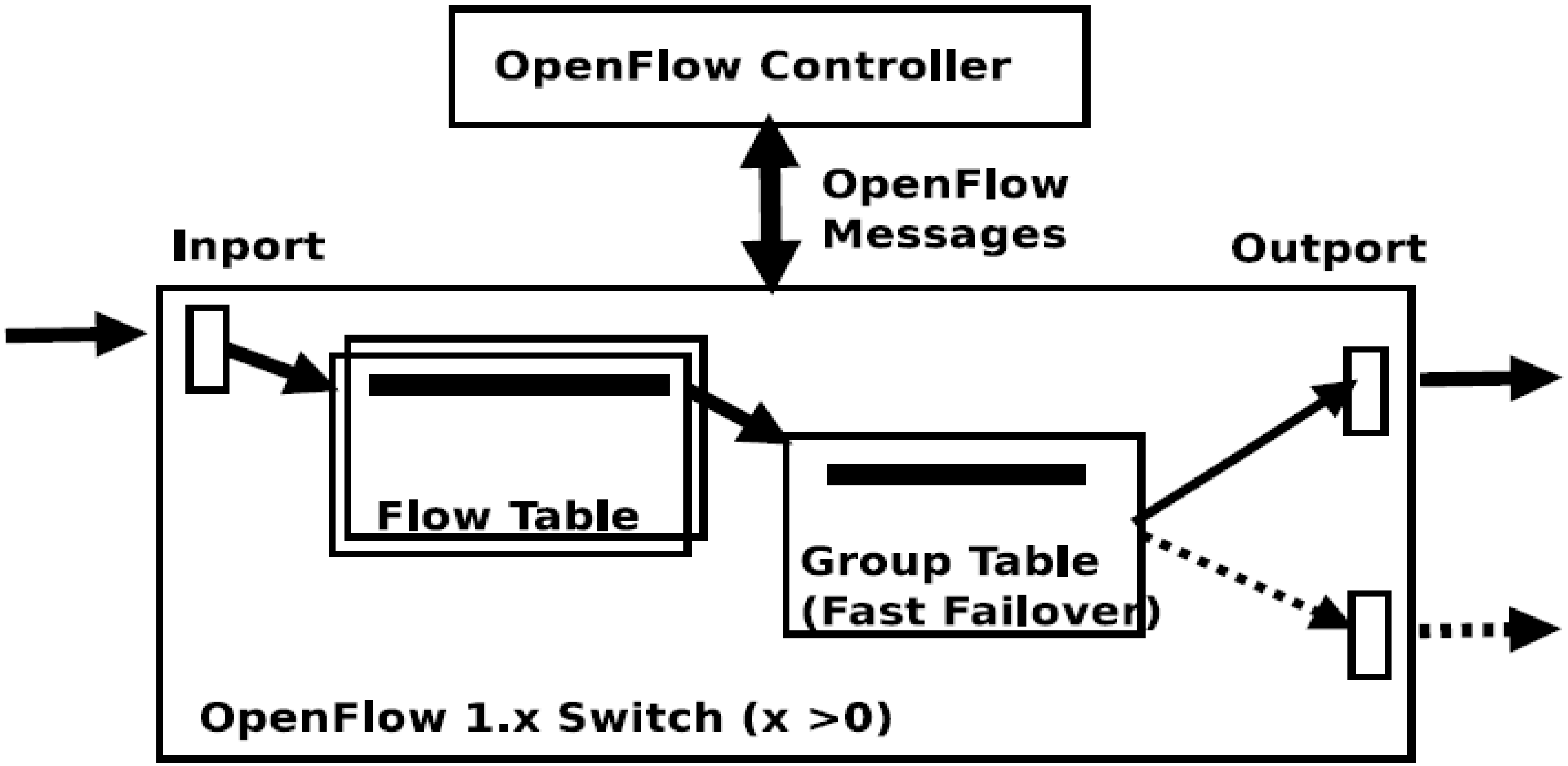
1. Switchover mechanism from a faulty link to a normal link
   - Switchover time: the time from failure detection to path restoration on an end-to-end basis.
   - OpenFlow: a wide variety of switchover mechanisms.
   - Implementation of several OpenFlow–specific and OpenFlow-integrated switchover mechanisms and evaluations of switchover performance.
     - OpenFlow–specific switchover mechanisms:
       - FAILOVER GROUP TABLE-based implementations
       - SELECT GROUP TABLE-based implementations
       - Both utilize local states of OpenFlow switches without direct involvement of remotely located controllers
     - OpenFlow-integrated switchover mechanisms: OpenFlow with Multipath TCP (MTCP) in the TCP layer
2. Communication delay (propagation delay) due to the separation of SDN controllers and switches
   - Switches and Controllers may be located far away with each other.
   - Analysis of the communication delays under a realistic network topology

Three Research Issues for Fast Network Recovery Using SDN/OpenFlow Technology (2)

3. Global view of the network: correct information on topology and link status
   - Necessary to find any available network resources (paths, links) to restore all the end-to-end paths.
   - IP network: a global view is maintained by the IP routing protocols like OSPF and BGP, but slow convergence time is a problem
   - SDN/OpenFlow network
     - a global view could be maintained among multiple SDN controllers, by existing IP routing protocols or any new routing protocols for SDN/OpenFlow) but there are very few standardized routing protocols especially designed for SDN/OpenFlow.
     - In our first experiment, under the assumption that a global view is always maintained with zero convergence time among the SDN controllers, we evaluate the end-to-end network recovery time.
     - In our second experiment, under the assumption that a global view is maintained by IP routing integrated with SDN (RouteFlow) among the SDN controllers, we evaluate the impact of the convergence time on the end-to-end network recovery time.

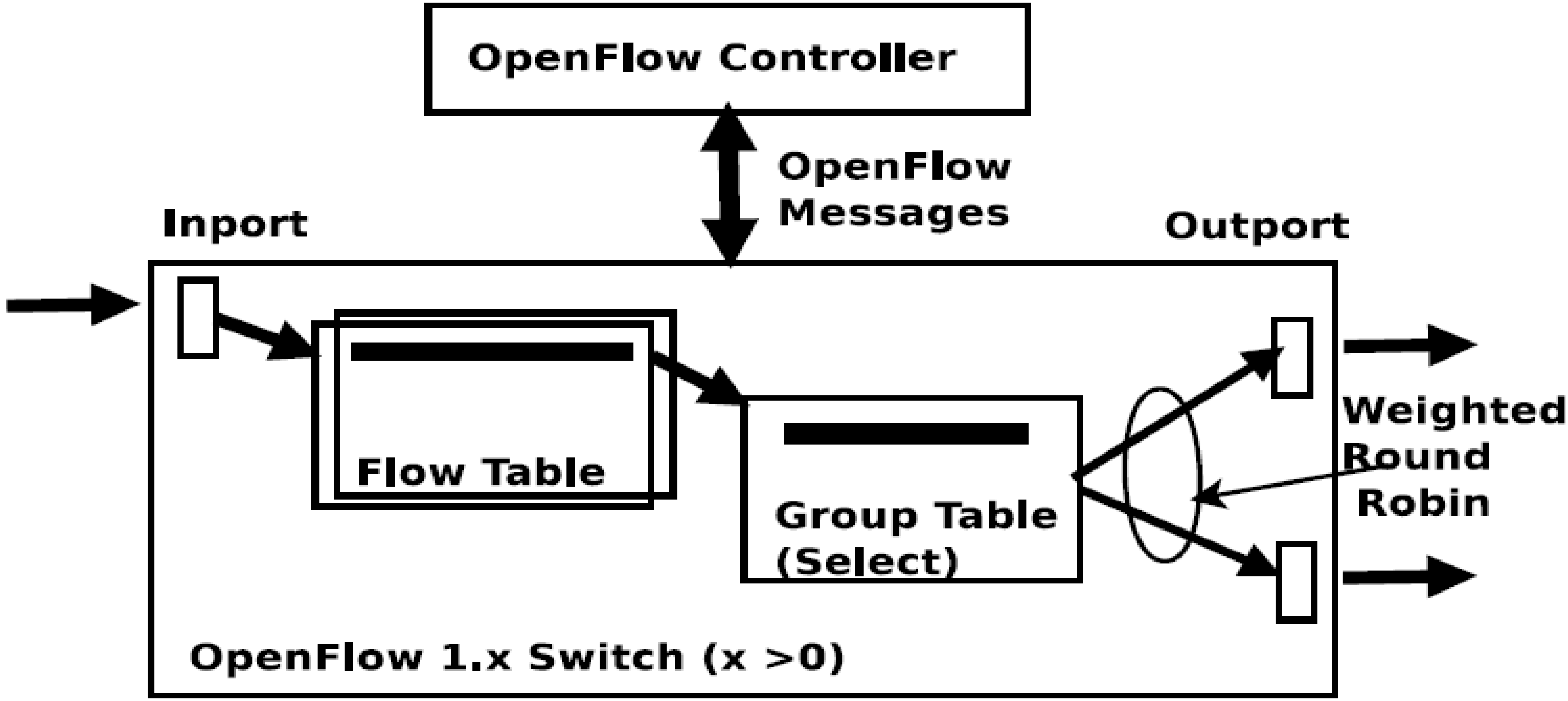
1^{st} Issue: Fast Local Switchover Mechanism (1): FAST FAILOVER GROUP TABLE

- FAST FAILOVER GROUP TABLE allows a fast switchover from the active output port to the standby output port. (in the active/standby mode) without direct involvement of controller.



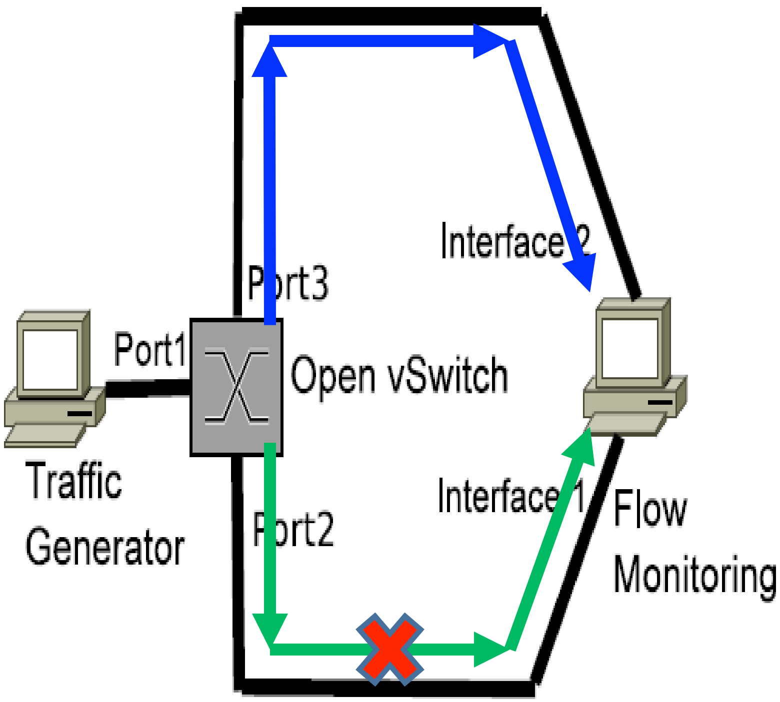
Fast Local Switchover Mechanism (2): SELECT GROUP TABLE

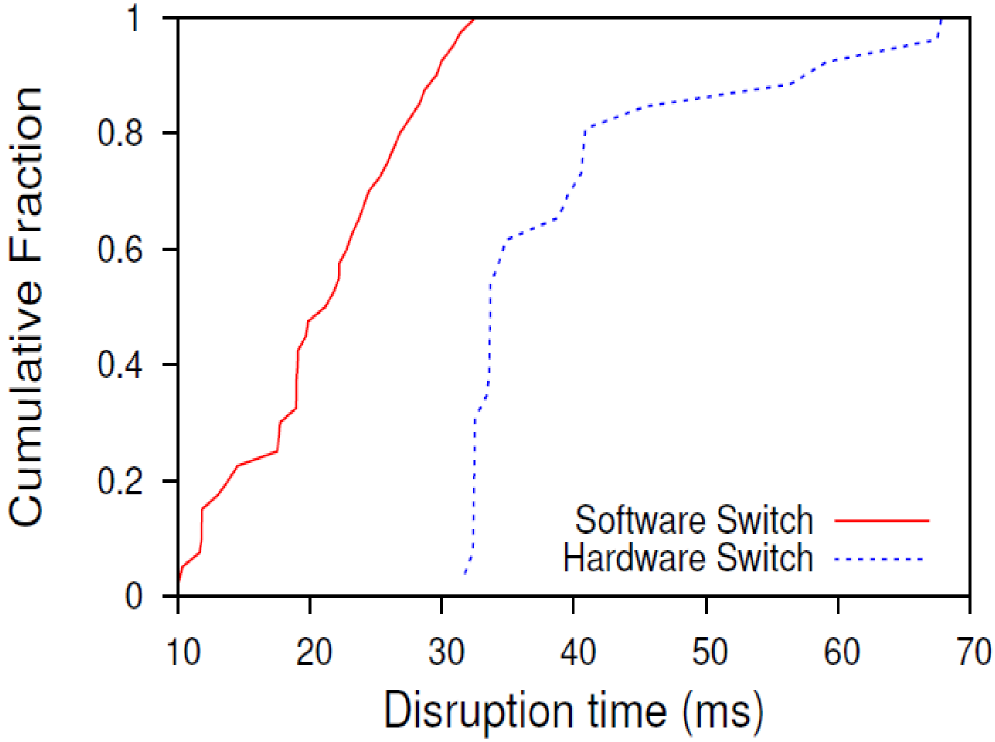
- SELECT GROUP TABLE allows a single data flow to be divided into multiple subflows, each with a different path (output port) in a weighted round robin manner (in the active/active mode)
- When link/port failures occur, the switch recalculates the weighted values, eliminates the failed ports and reallocates the traffic to active output ports.
- SELECT GROUP TABLE achieves a better resource allocation and less packet loss than FAST FAILOVER GROUP TABLE



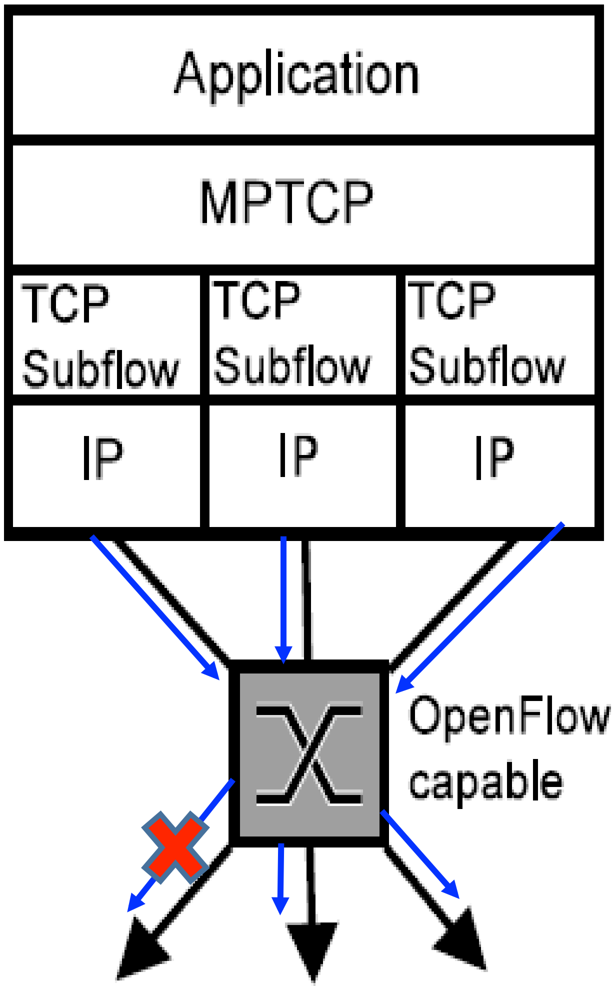
Implementation and Evaluation of Fast Local Switchover Mechanisms on Two Different Platforms

- 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 network recovery time (Disruption time) : 21.1 ms for software switch and 39.5 ms for hardware switch





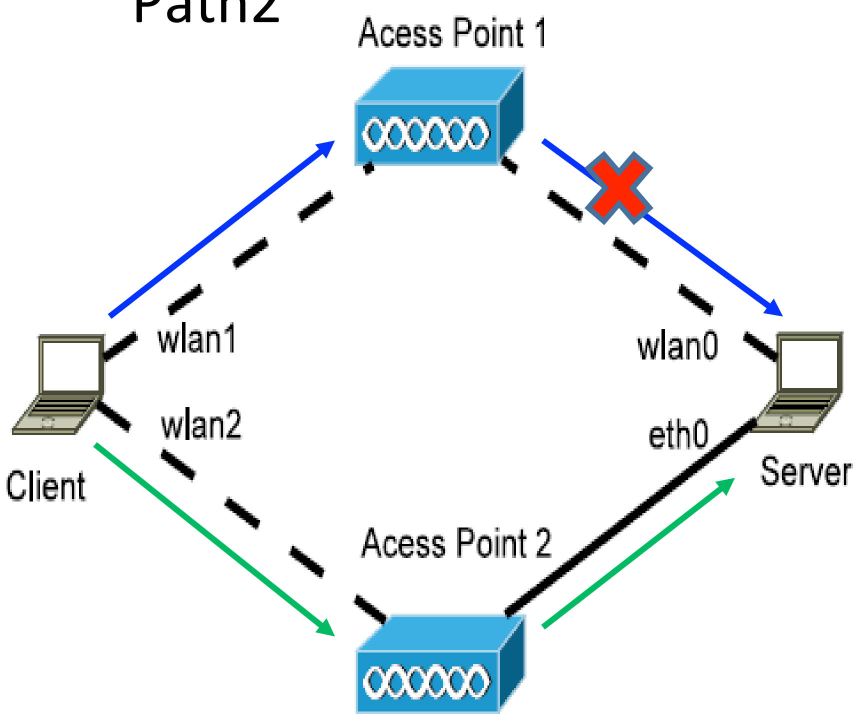
Fast Local Switchover Mechanism (3): Multipath TCP (MPTCP) Integrated with OpenFlow

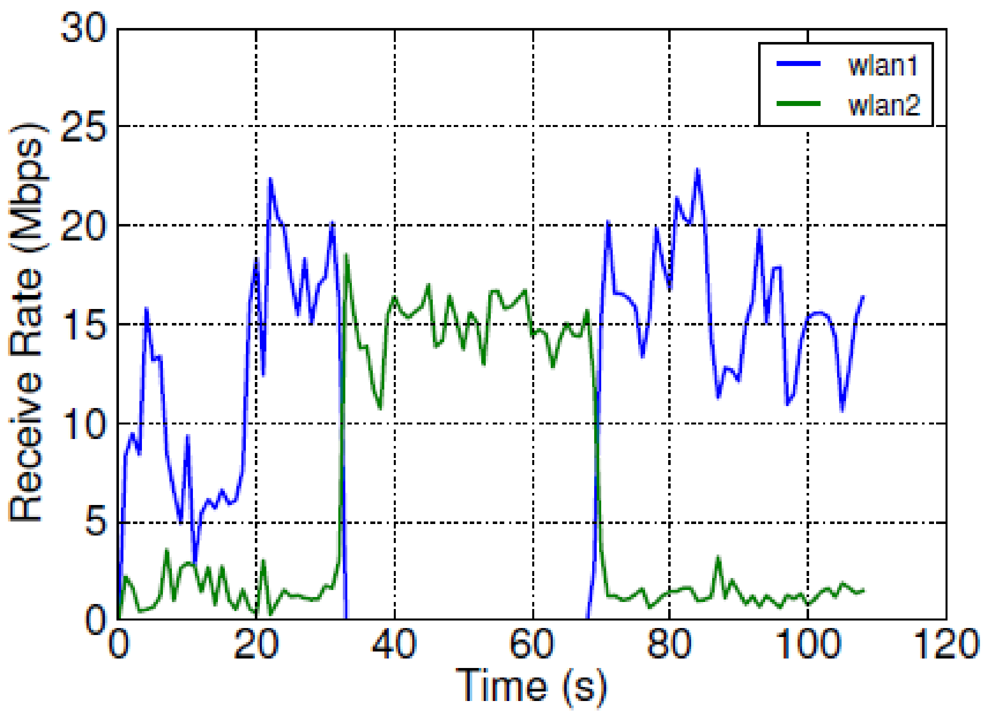


- MPTCP is standardized by IETF (Internet Engineering Task Force), does not need to modify existing applications
- MPTCP creates and maintains multiple active paths for an end–to-end connection
- Divides the TCP flow into multiple active TCP subflows
- Each subflow may go through a different path to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

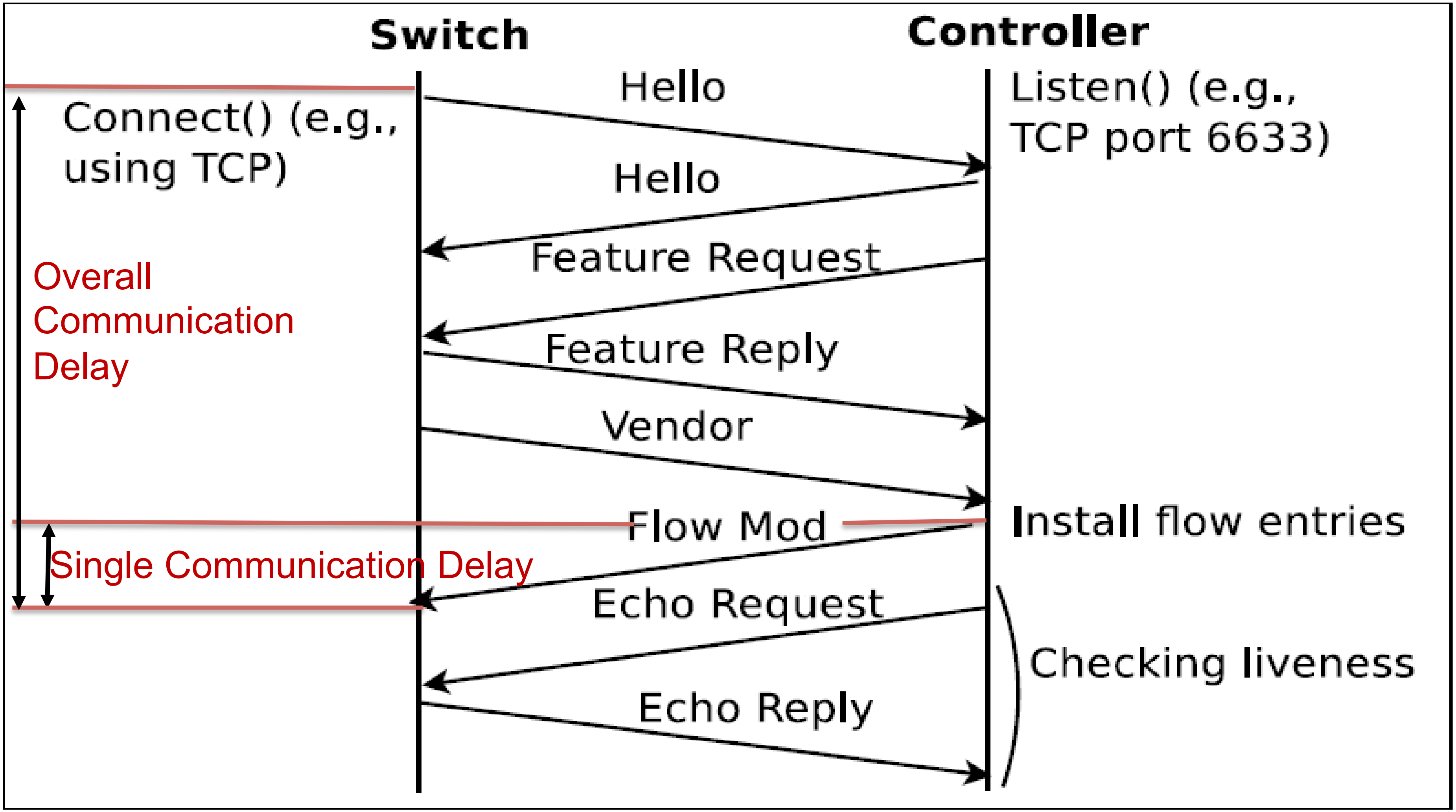
Implementation and Evaluation of MPTCP on WiFi Network Environment

- 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

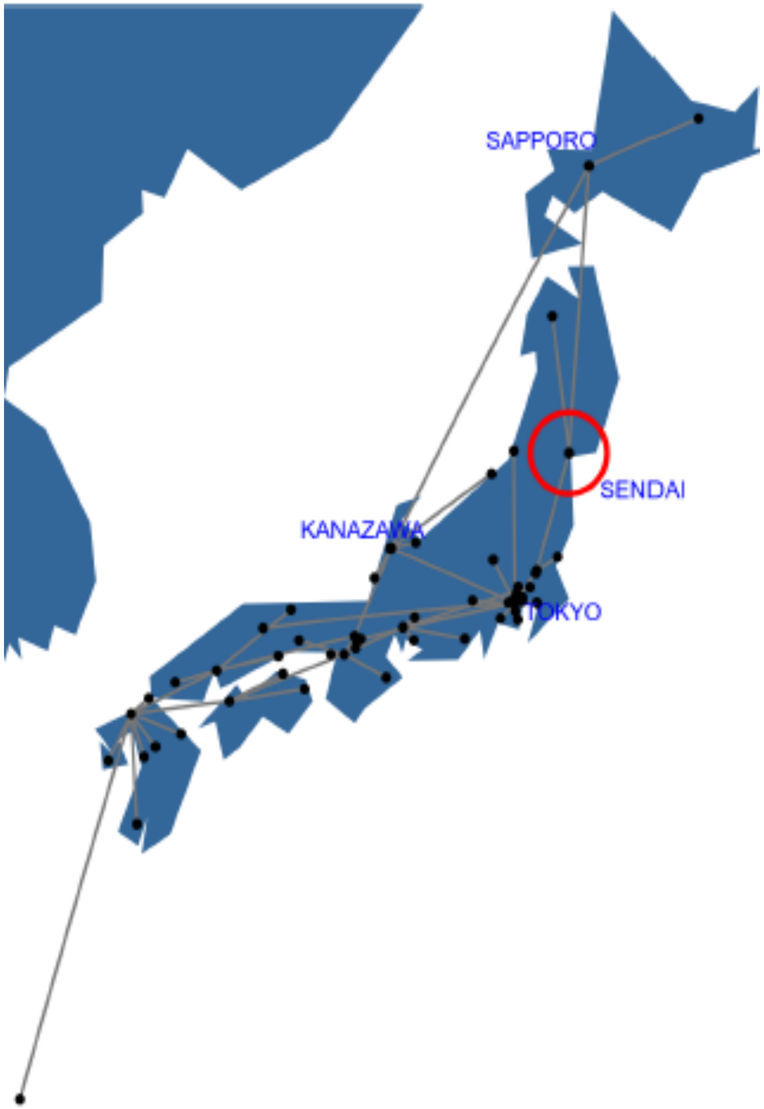




2nd Issue: Communication Delay (Latency) between Controllers and Switches



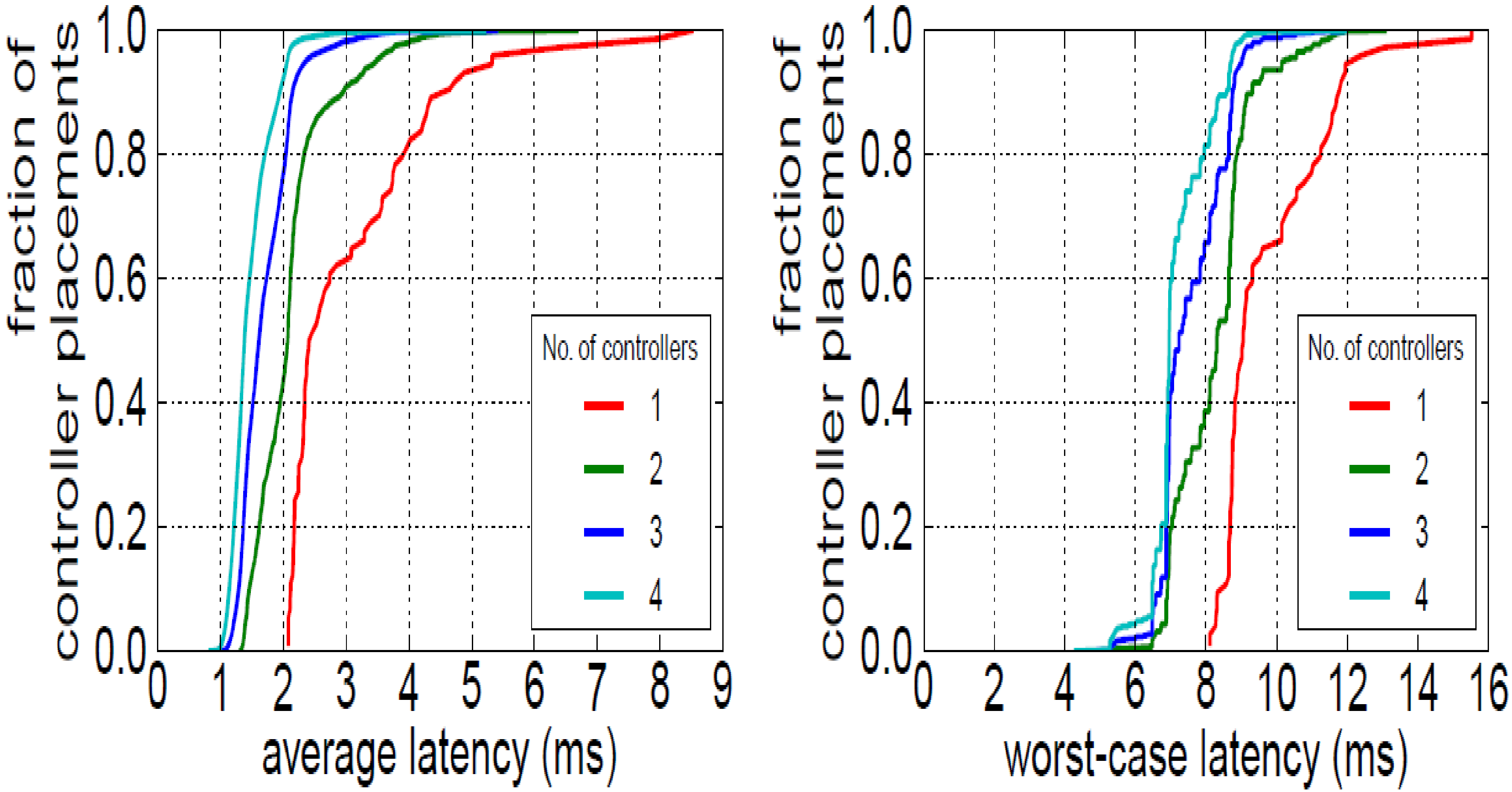
Analysis of Communication Delay between Controllers and Switches



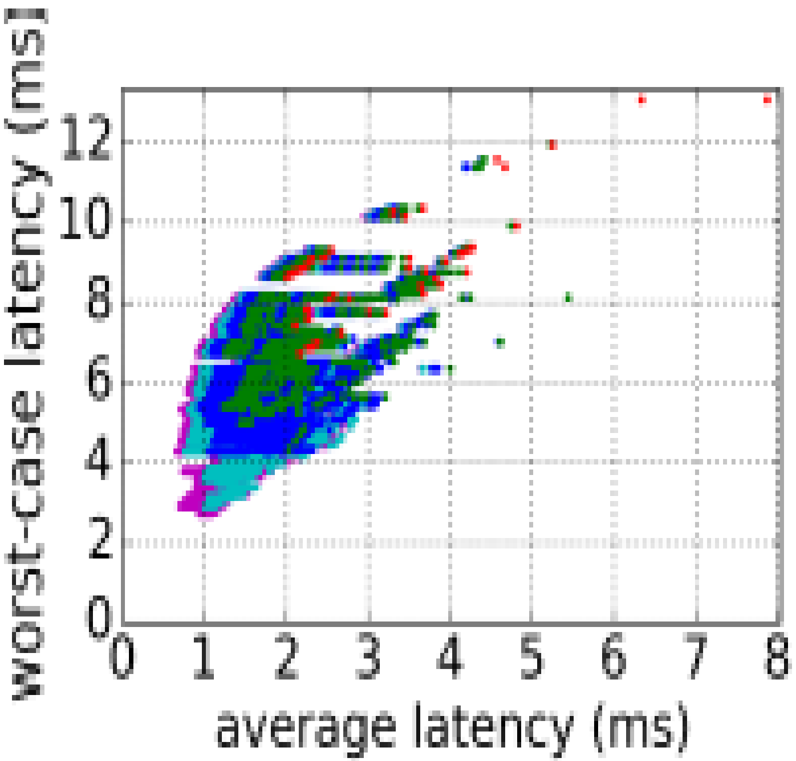
- SINET3 topology is used to evaluate communication latencies between controllers and switches.
- SINET3: the previous version of current SINET4, a Japanese national research and education network.
- Two latency metrics under 2/3 propagation delay of light speed :
  - Average Latency for all the possible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

Evaluation of Average and Worst-Case Latencies

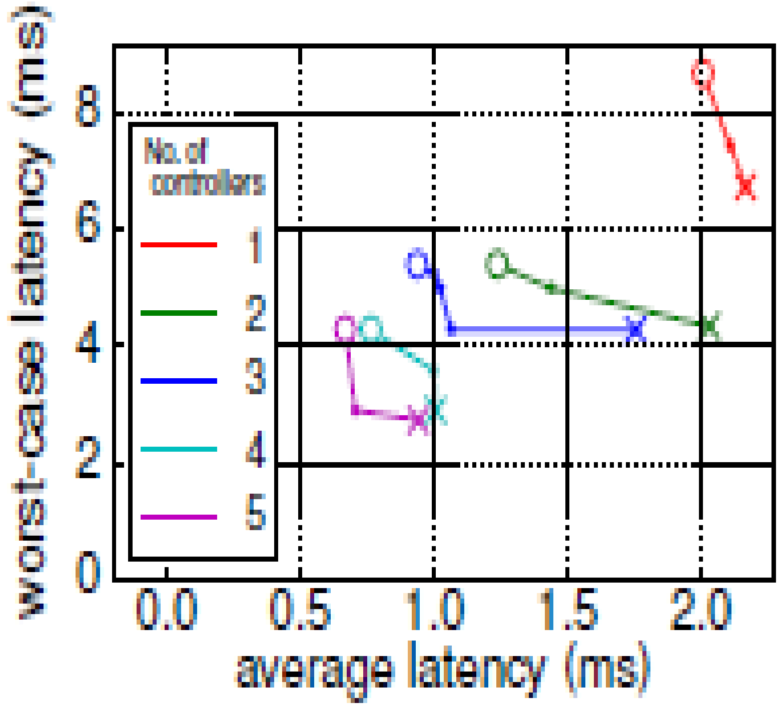
- The more controllers, the lower latencies
- Should carefully choose the location of controller



Optimal Values of Average and Worst-Case Latencies



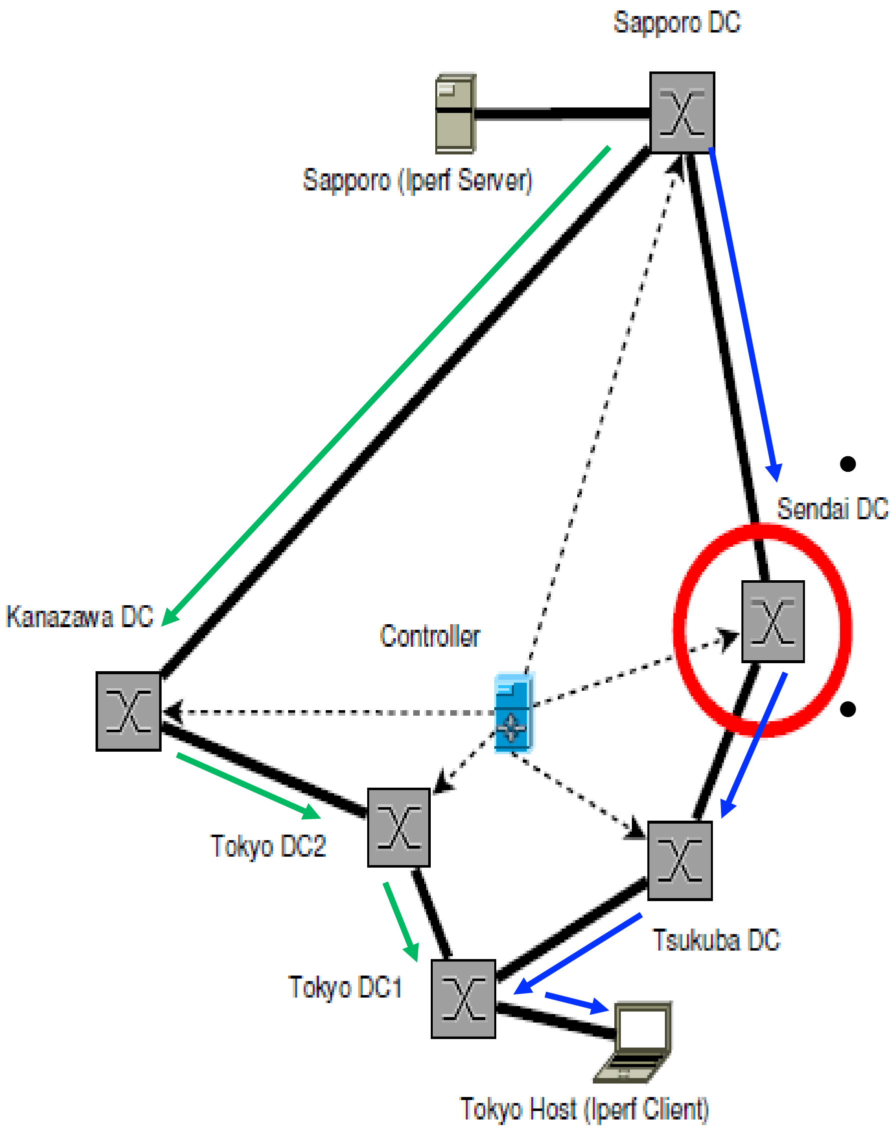
(a) All the combination



(b) Pareto optimal

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

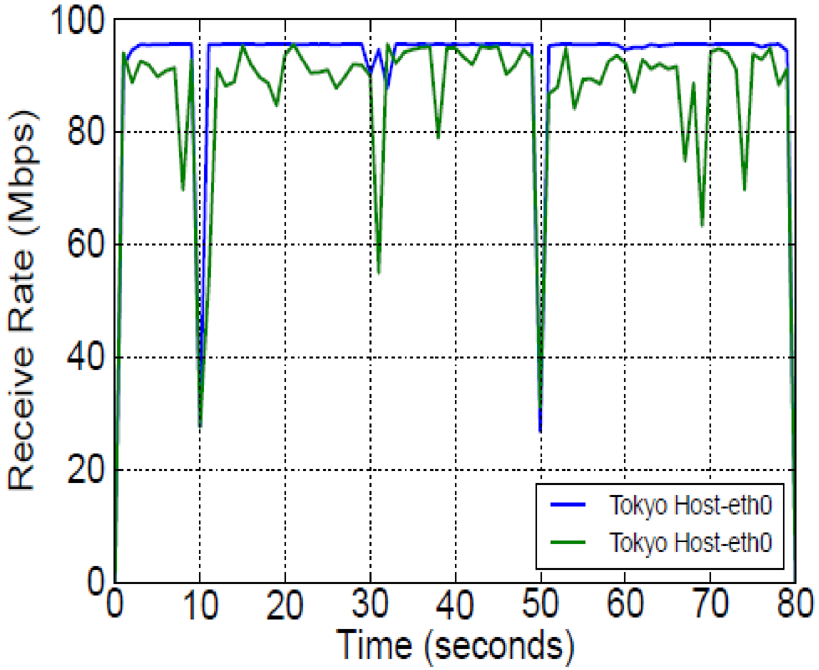
3rd Issue: Global View of the Network



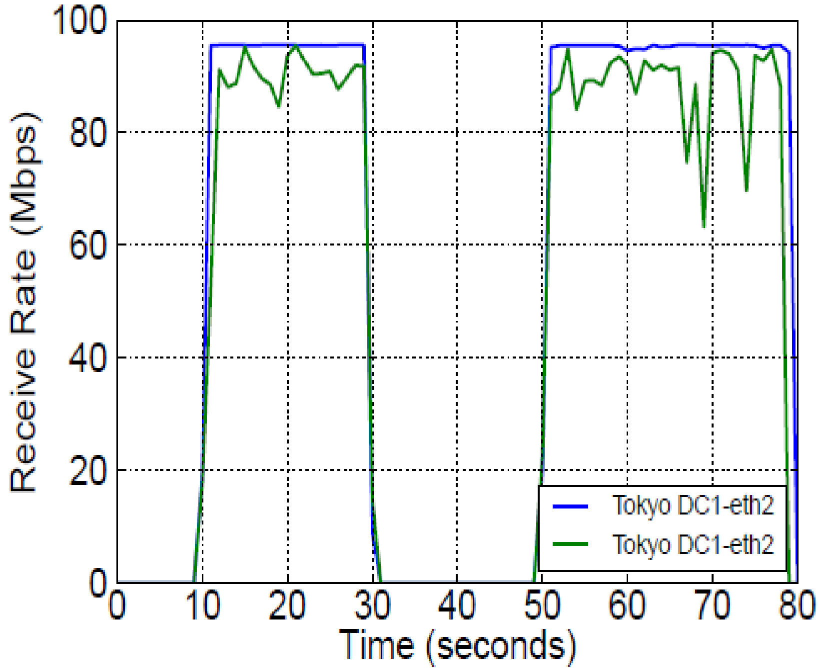
- Mininet 2.0, the virtual network simulator is used to investigate the overall behaviors of link network recovery under the Great East Japan Earthquake scenario for SINET3 topology.
- The worst case latencies between the controller and switches are assumed
- When a link failure occurs on the main path, the controller software (POX) with the global view 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 50th seconds, we confirmed that 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



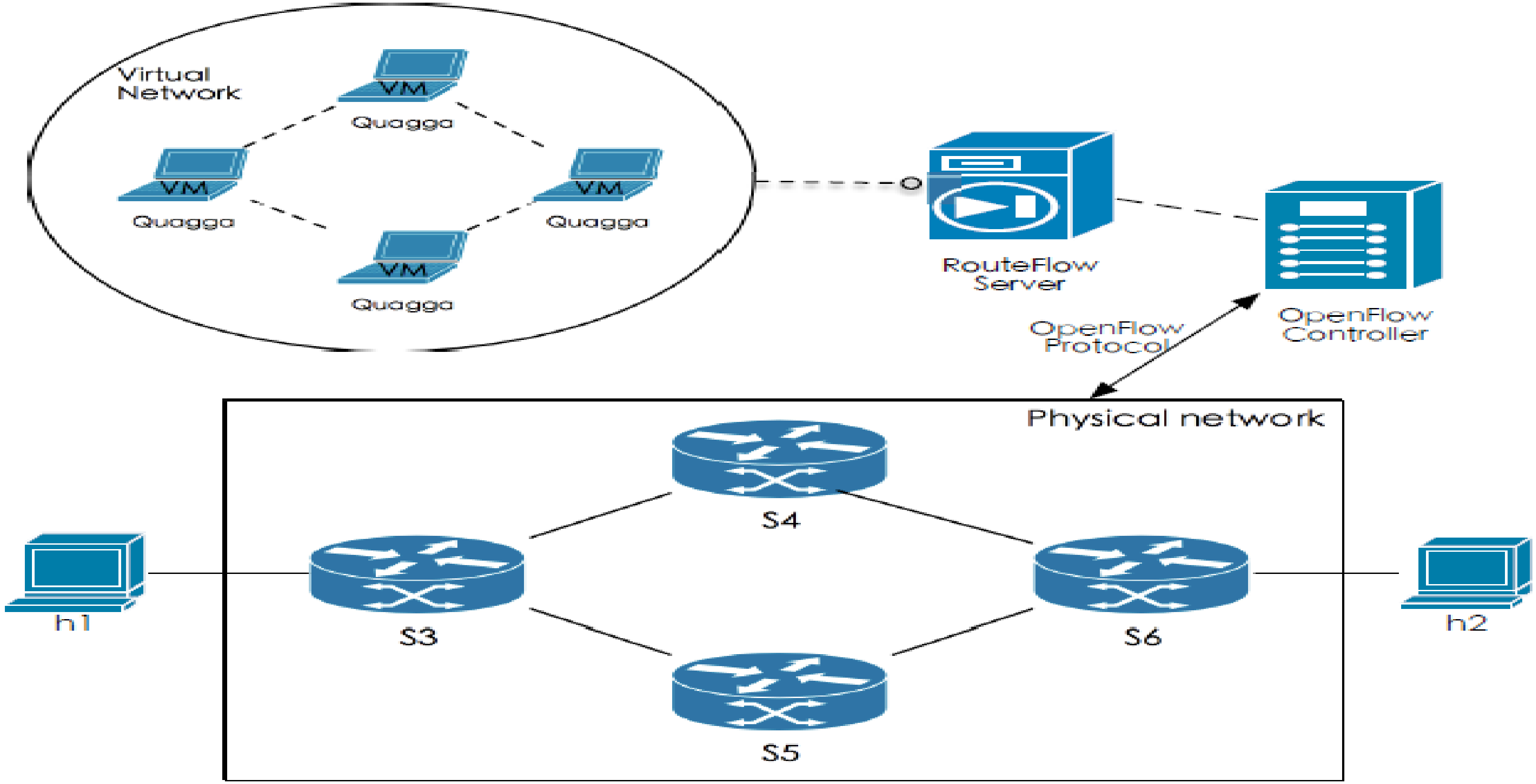
Receive traffic (i.e., goodput) at Tokyo-Host



Goodput on the alternative path

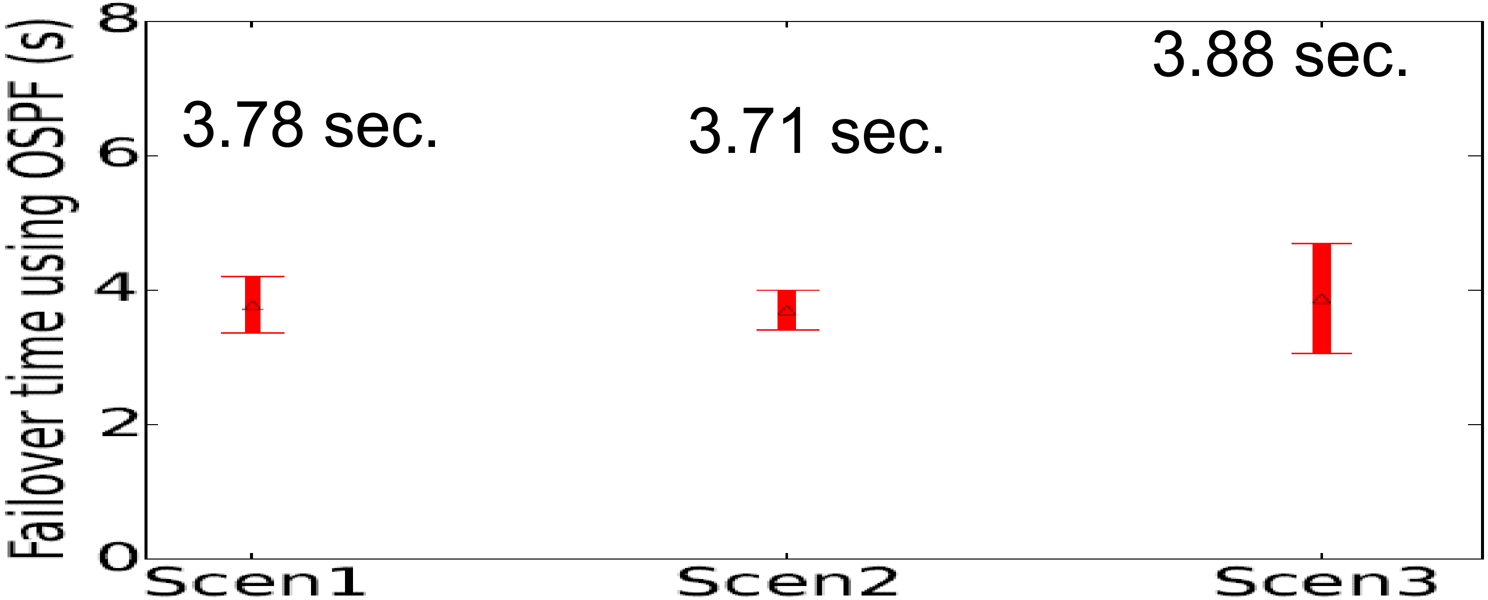
Implementation and Evaluation of Network Recovery by SDN integrated with IP Routing for Global View

- Network recovery under the global view that integrates SDN with conventional IP routing (Route Flow) is implemented and evaluated on both the Mininet simulator and a real testbed with physical OpenFlow switches.



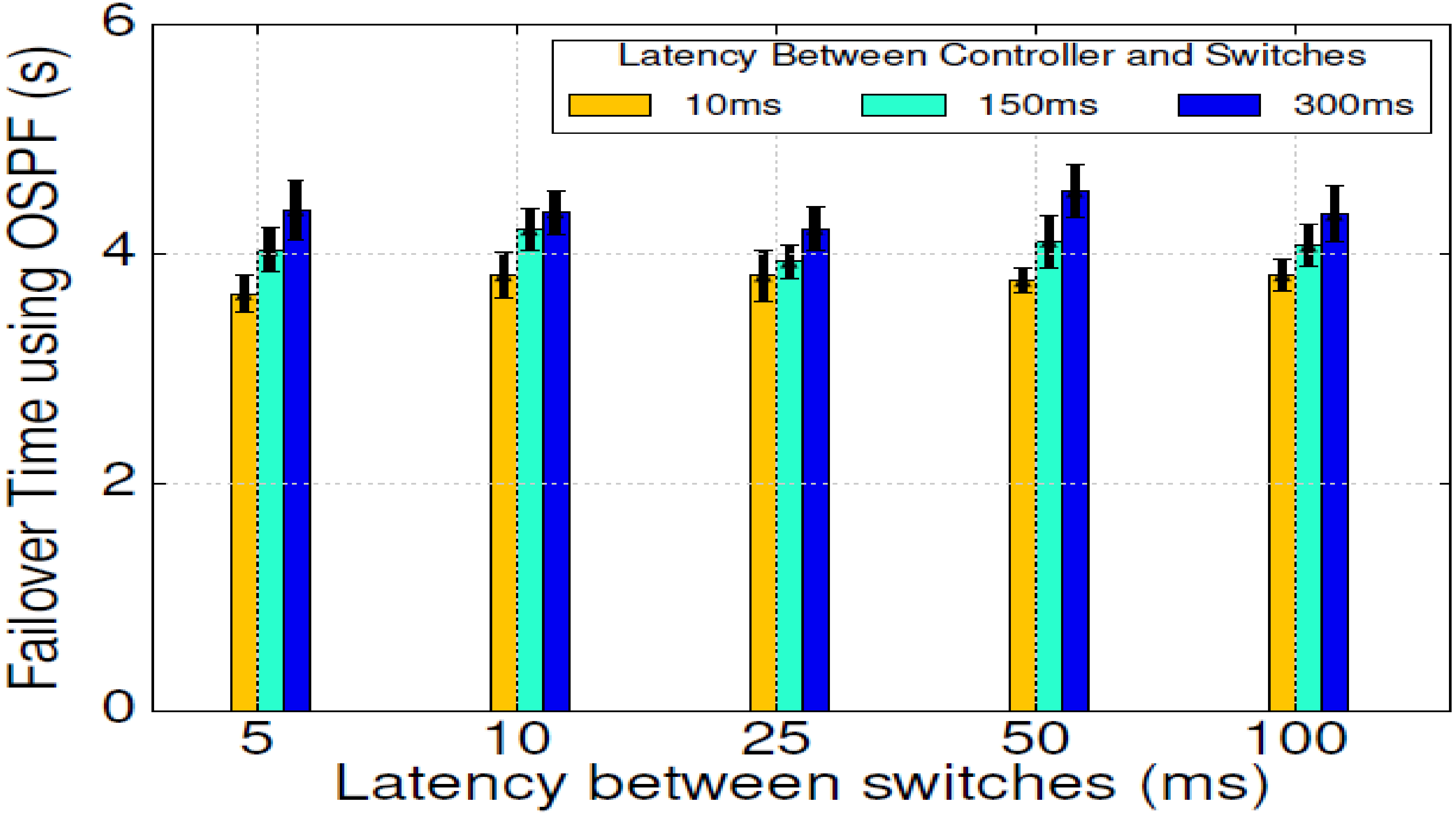
Network Recovery Time Comparisons under OSPF Protocol

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



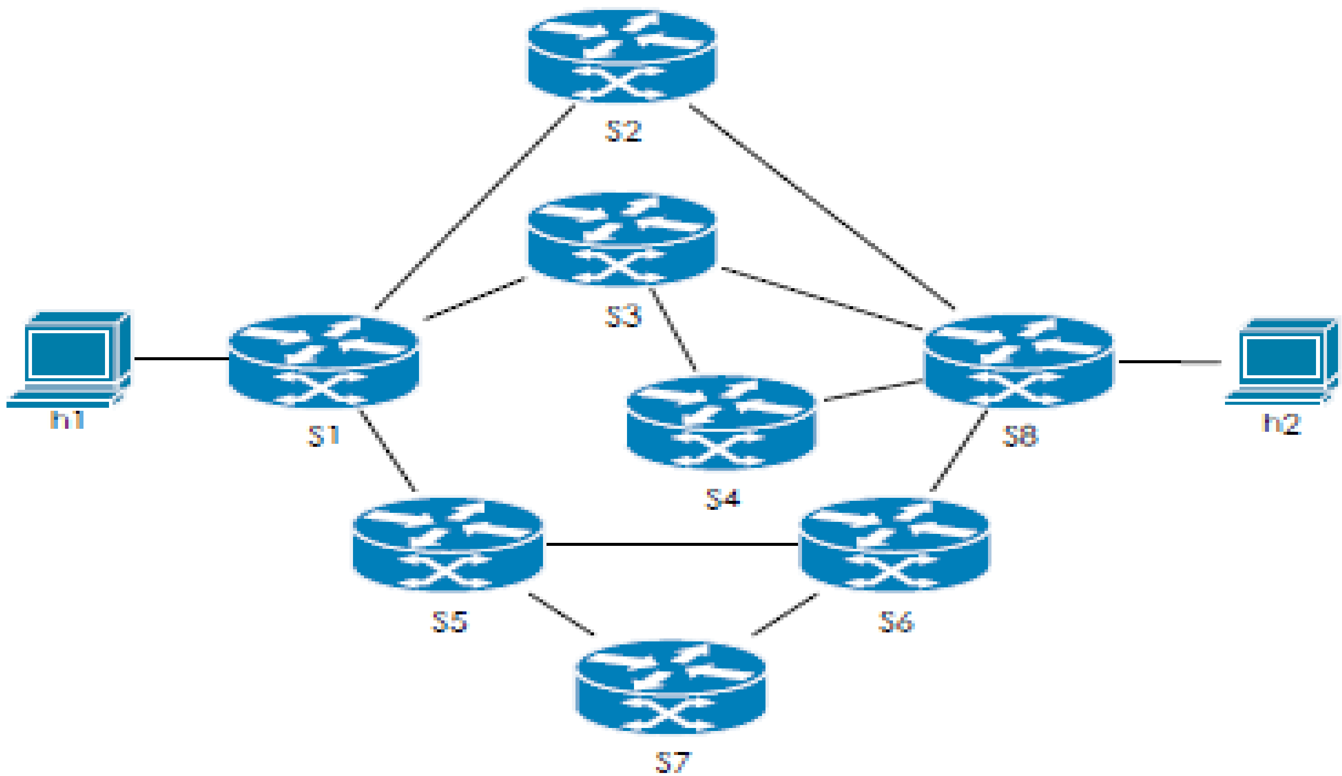
Effect of Communication Delays on Network Recovery Time, under OSPF Protocol

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



Evaluation of Network Recovery Time under Multiple Link Failures, under OSPF Protocol

- A more complex network topology with 8 switches, assuming multiple link failures
- 5 redundant paths from source (h1) to destination (h2)
- Mininet simulation running Route Flow (for OSPF)



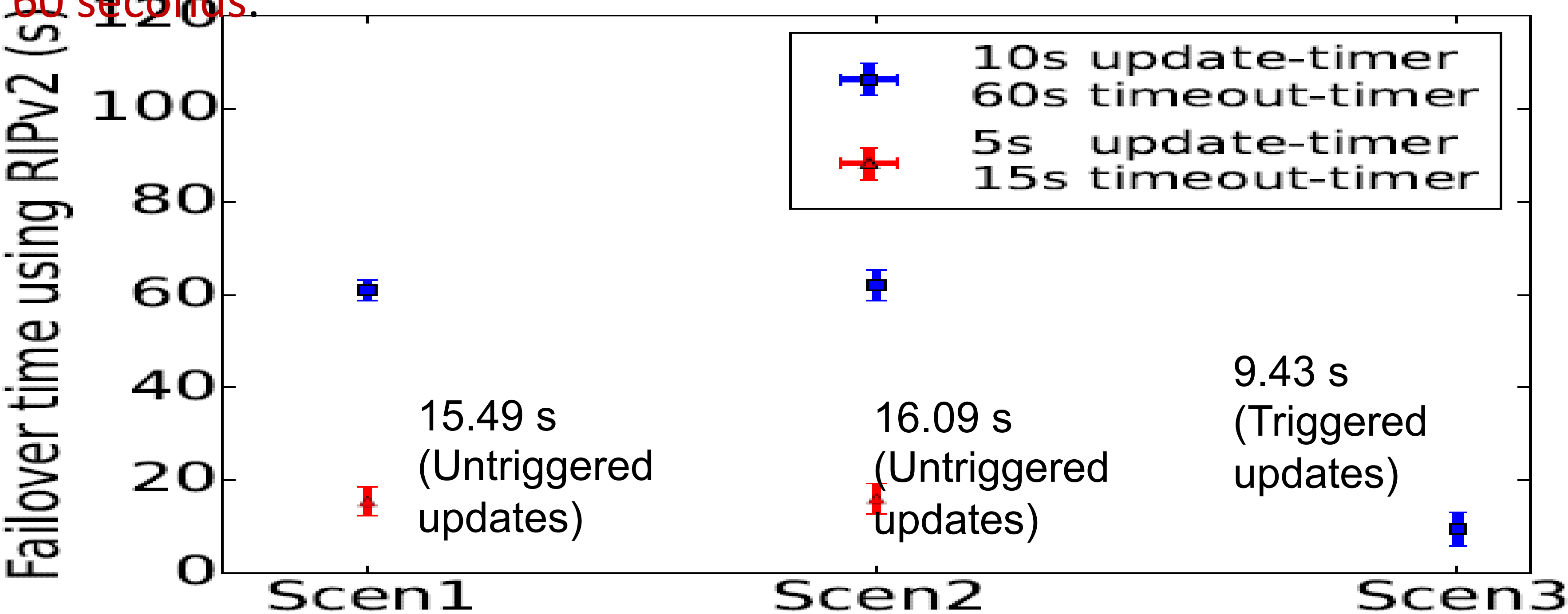
Evaluation Result of Network Recovery Time under Multiple Link Failures, under OSPF Protocol

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

Number of Link Failures	Link Down	Mean (s)
1	S2-S8 Down	4.131 ± 0.378
	S3-S8 Down	4.229 ± 0.441
	S4-S8 Down	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 (SDN/OpenFlow): Mininet simulation running Route Flow (for RIPv2)
- Scen2 (SDN/OpenFlow): Pica8 switches running Route Flow (for RIPv2)
- Scen3 (Conventional) : Pica8 switches running IP routing protocol (L2/L3 RIPv2)
- All Scen1 and Scen2 results are close to the timeout-timer of 15 seconds or 60 seconds.



Summary of Resilient Backbone Network Evaluation

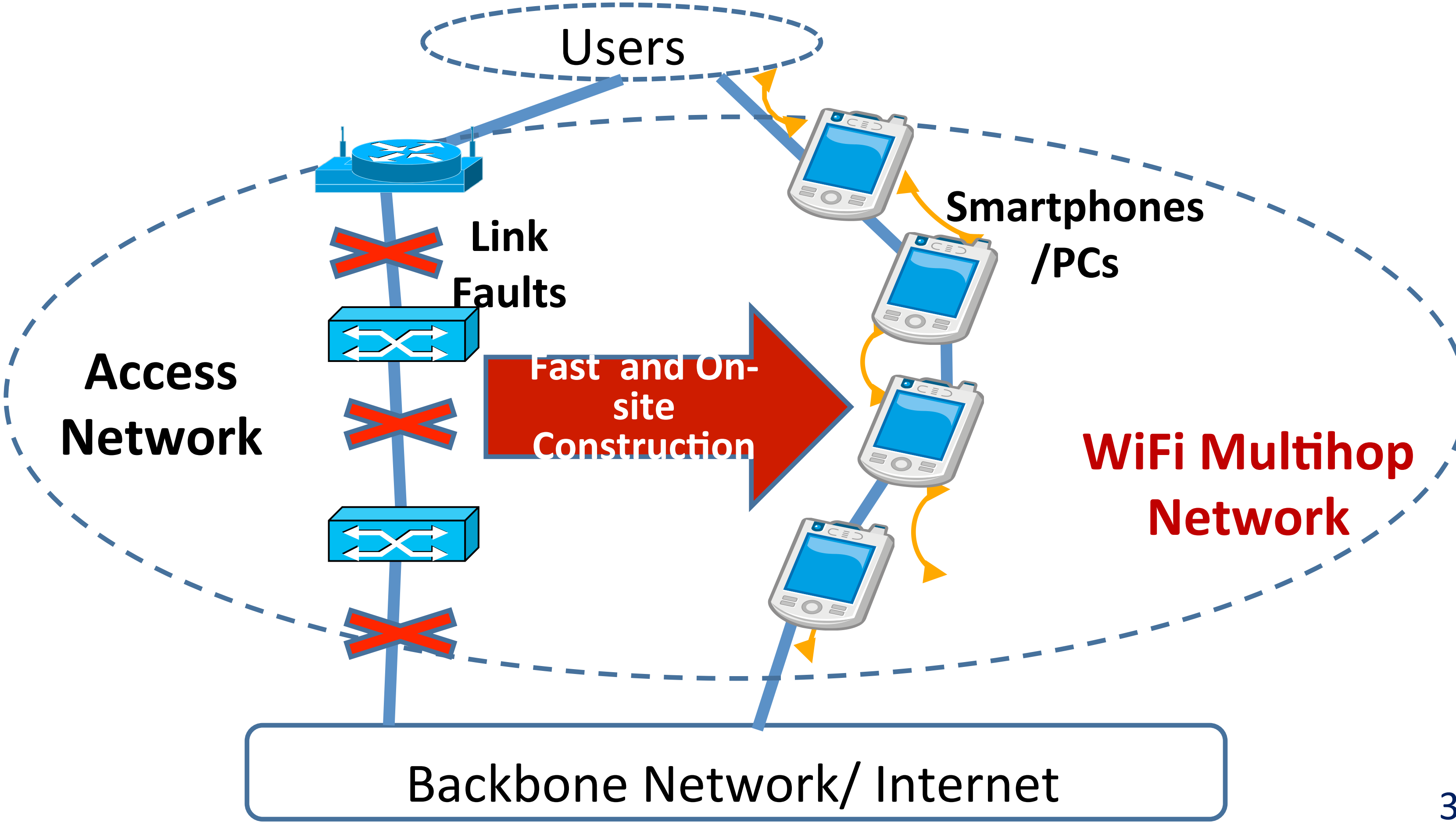
- SDN/OpenFlow technologies are technically feasible.
- It can offer a wide variety of switchover mechanism with fast switchover time of 20 to 40 ms under current implementation technologies
- The overall network recovery time of our SDN/OpenFlow approach ranges from 20 ms to 60 seconds, largely depending on the convergence time of the employed routing protocol.
- The network recovery time is a little bit better than or comparable to conventional IP network approach.
- New SDN/OpenFlow friendly routing protocol may be necessary to greatly reduce the network recovery time.
- Real deployment is still a challenge.

Resilient Access Network

- Requirements
  - (R1) Internet connectivity is required for survivors to contact their families and relatives and get help from external areas.
  - (R2) Construction, configuration and use by Survivors just after a disaster in stable environments (e.g. evacuation centers). Supporting complete mobility (handoff) is not always necessary.
  - (R3) Use of available commodity mobile devices available in evacuation centers.
- Key Ideas
  - WiFi-based Technologies: for ubiquitous internet access.
  - Commodity Mobiles Devices: smart phones, and laptop PCs.
  - Wireless Virtualization : a single WiFi interface of a device to be used for two wireless channels to allow both a Virtual Access Point (VAP) and Wireless Station（STA） capabilities
  - Multihop Communication Abstraction to allow users to recognize a multihop network as if it looks an ordinary single hop WiFi network
  - Tree Topology to make the routing overhead to the minimum.

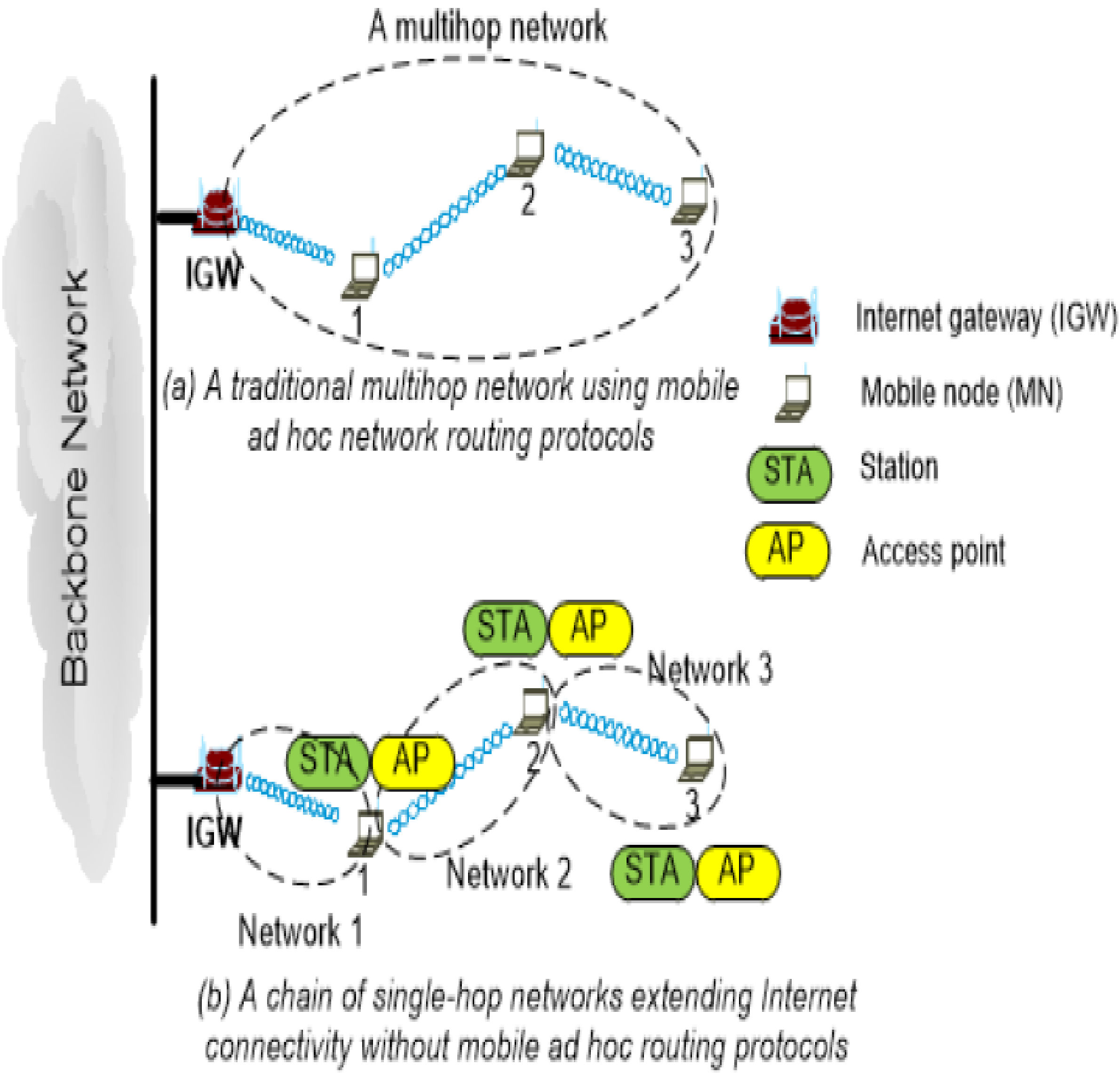
General Concept of Resilient Access Network

- Survivors (network users) construct a WiFi multihop access network on site, using commodity mobile devices to provide internet access services.



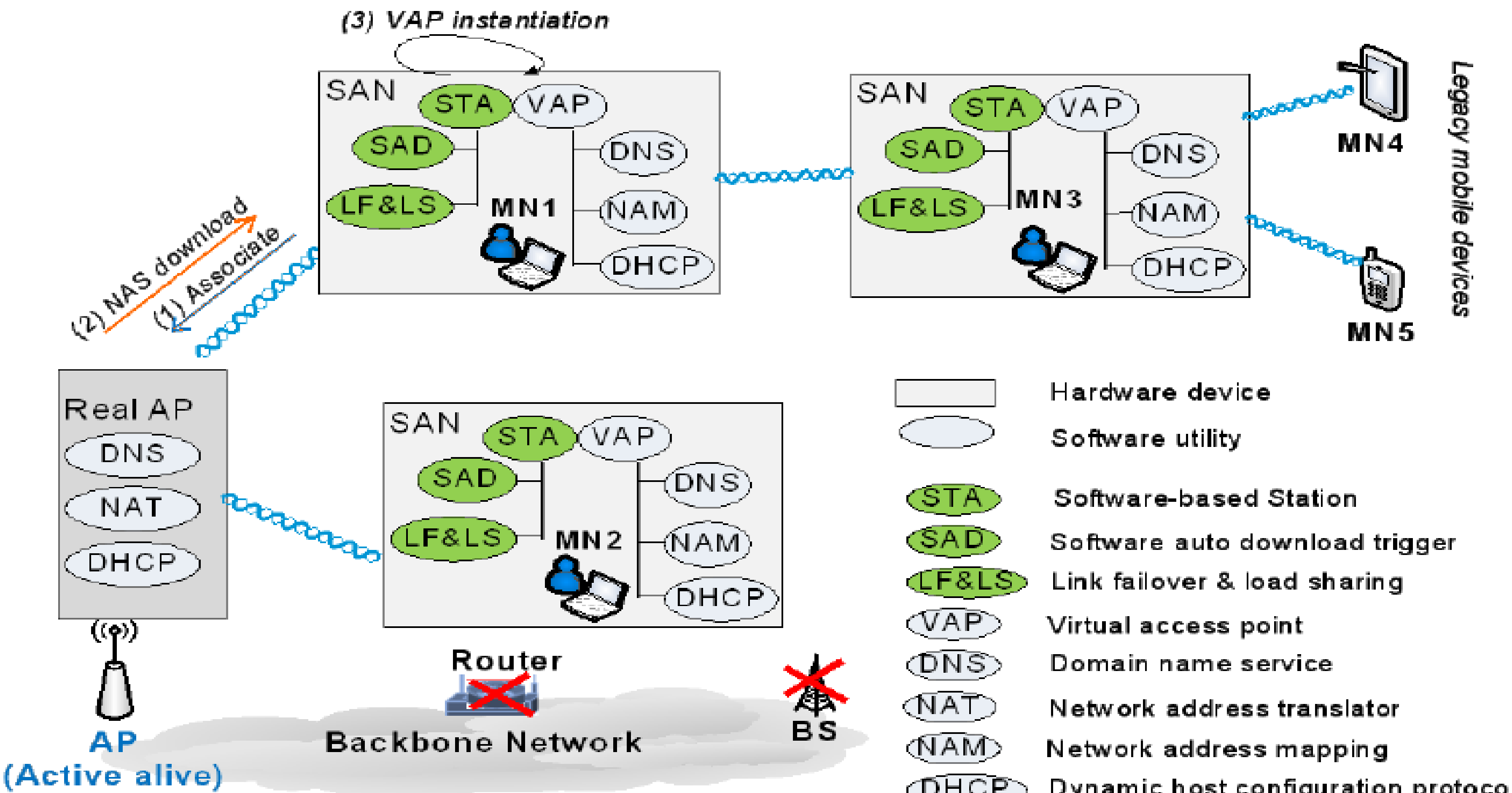
Multihop Communication Abstraction

- Ad hoc network (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



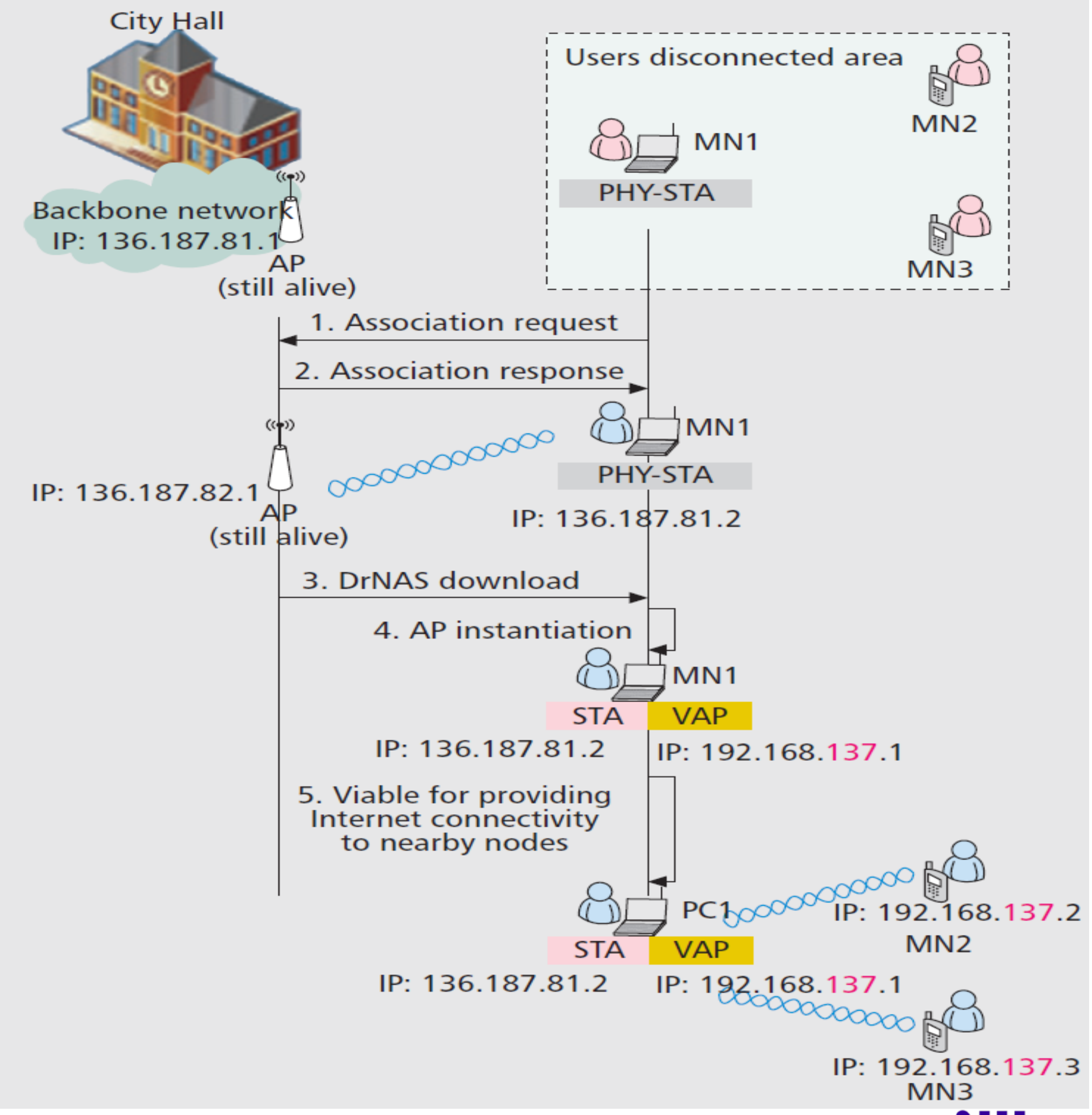
Network Auto-Configuration Software (NAS) for WiFi Virtual Access Point (VAP) and WiFi Station (STA)

- 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



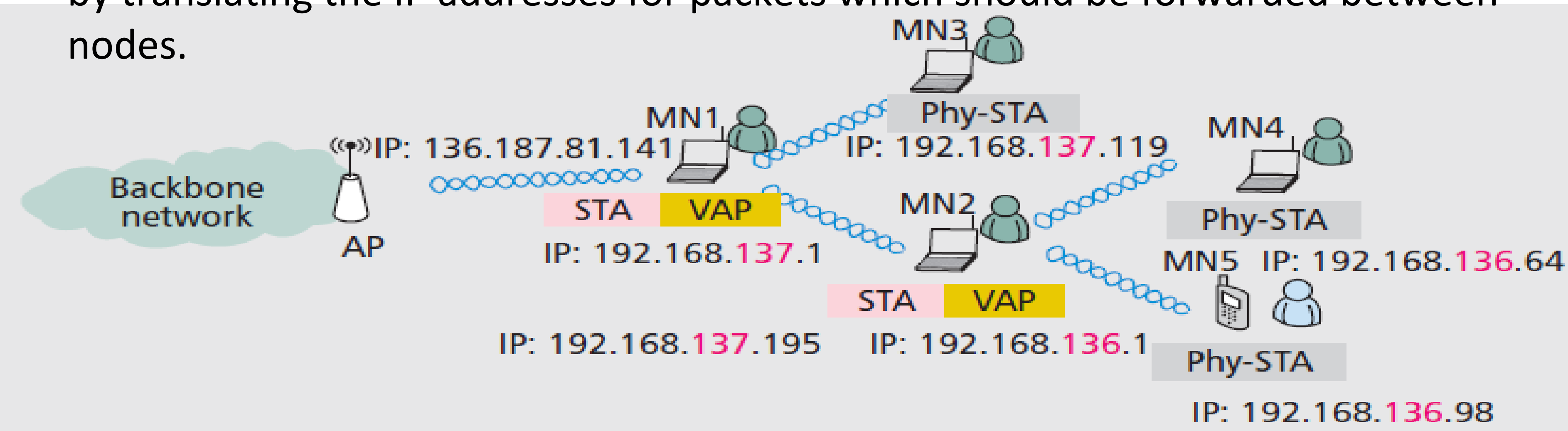
Procedure for NAS Download and Network Configuration

- Steps 1 & 2: Association between the AP and MN1. MN1 downloads NAS from the AP.
- Steps 3 and 4: NAS transforms MN1 into VAP, using its second logical wireless interface, while the first logical interface works as a wireless station (STA).
- Step 5: MN1's VAP offers Internet connectivity to MN2 and MN3. Their IP addresses are assigned by the DHCP server on MN1's VAP.
- The above process is repeated to form a multihop network



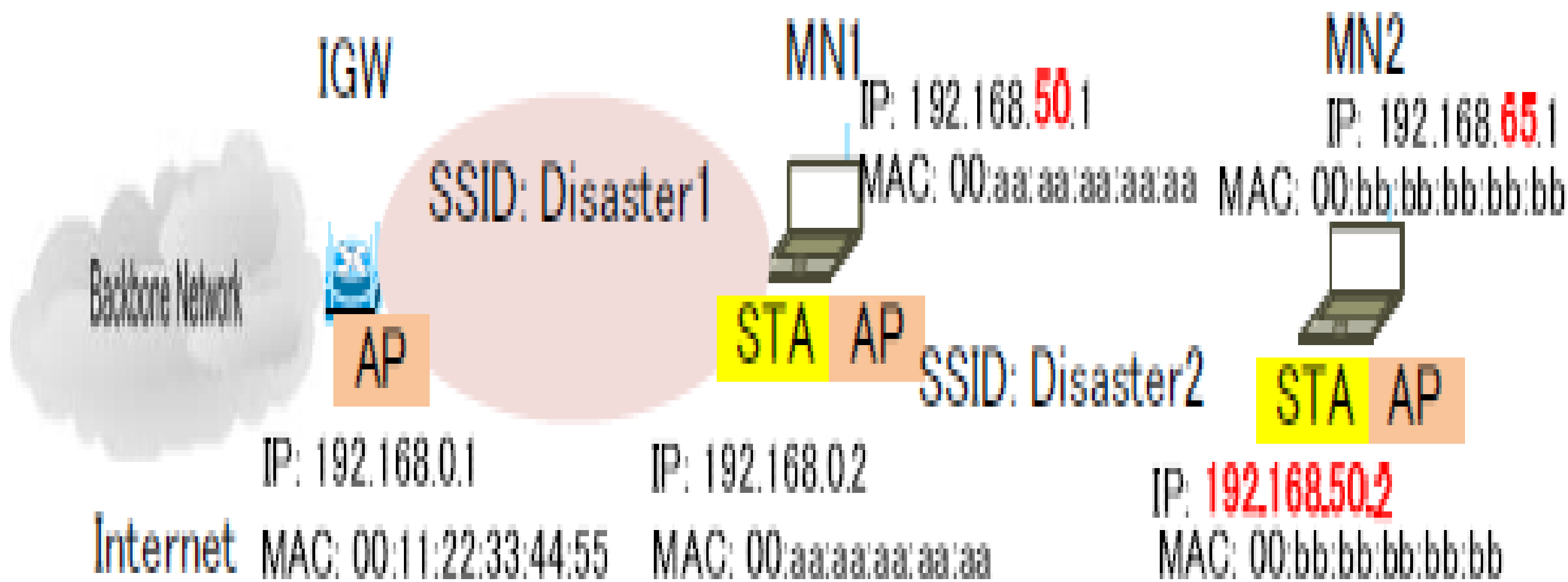
Simple Routing under Tree Topology

- Each mobile node has two logical wireless interfaces (for VAP & STA), each with a different private IP address.
- The STA's private IP address is assigned by a DHCP server in the upstream (parent) node's VAP.
- The VAP's private IP address is assigned by the DHCP server in its own node's VAP.
- For upstream flows, the VAP forwards packets via its STA to the parent VAP.
- For down stream flows, the NAT (Network Address Translation) at each node identifies to which STA the packet should be forwarded and serves as an IP router by translating the IP addresses for packets which should be forwarded between nodes.



DNS Resolution Mechanism to Allow Internet Access in Multihop Network

- MN2 submits the DNS query to MN1.
- MN1 delegates this request to its default gateway (IGW).
- IGW provides the DNS response which propagates via MN1 to MN2.
- Getting the actual (global) IP address of the destination, MN2 can issue an HTTP request to that host.

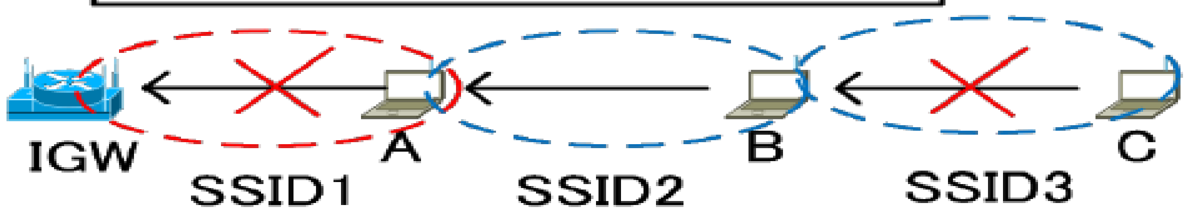


Network Reconfiguration Support: Connectivity Status Table (CST)

- Each node manages the CST containing
  - the status ("Connected"/"Disconnected") of all the upward links over the path from the Internet gateway (IGW) to its own node.
  - the Hop_count that represents the hop 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

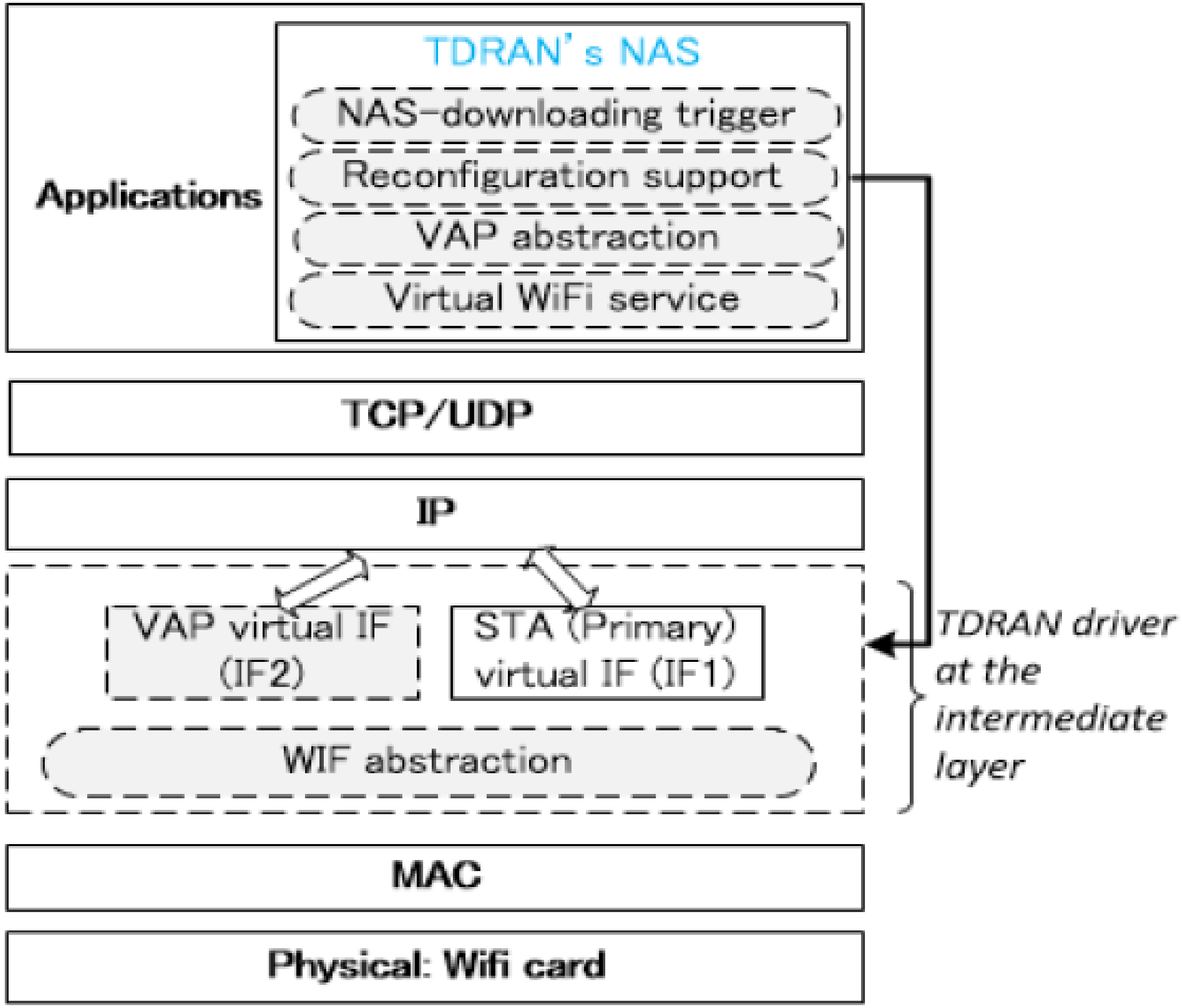
SSID	Status	Hop_count
SSID1	"D"	1
SSID2	"C"	2

C's CST

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

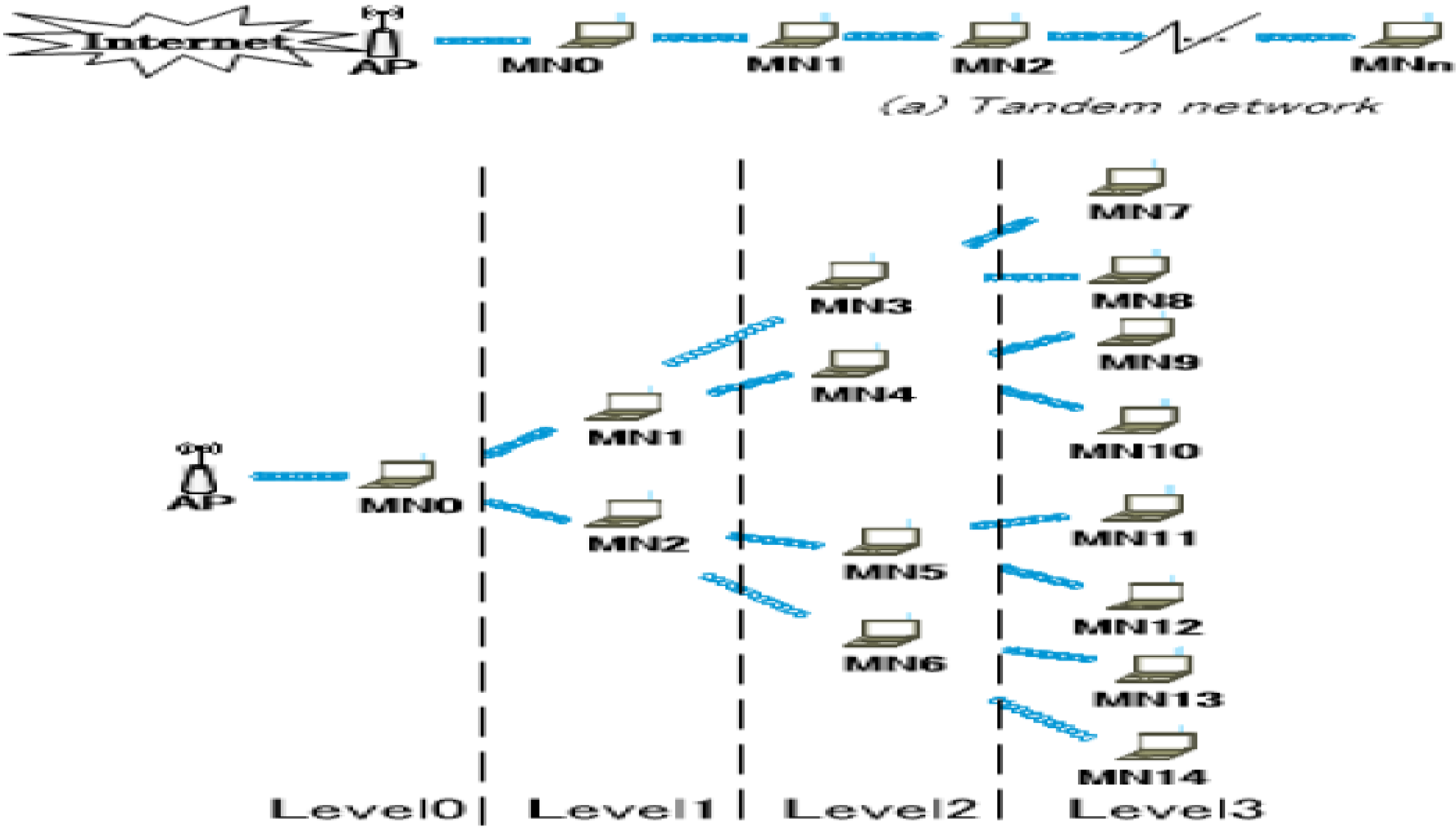
Network Auto-Configuration Software (NAS) Components in Each Node

- 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

The areas were hit by Great East Japan Earthquake



(a) Tandem network

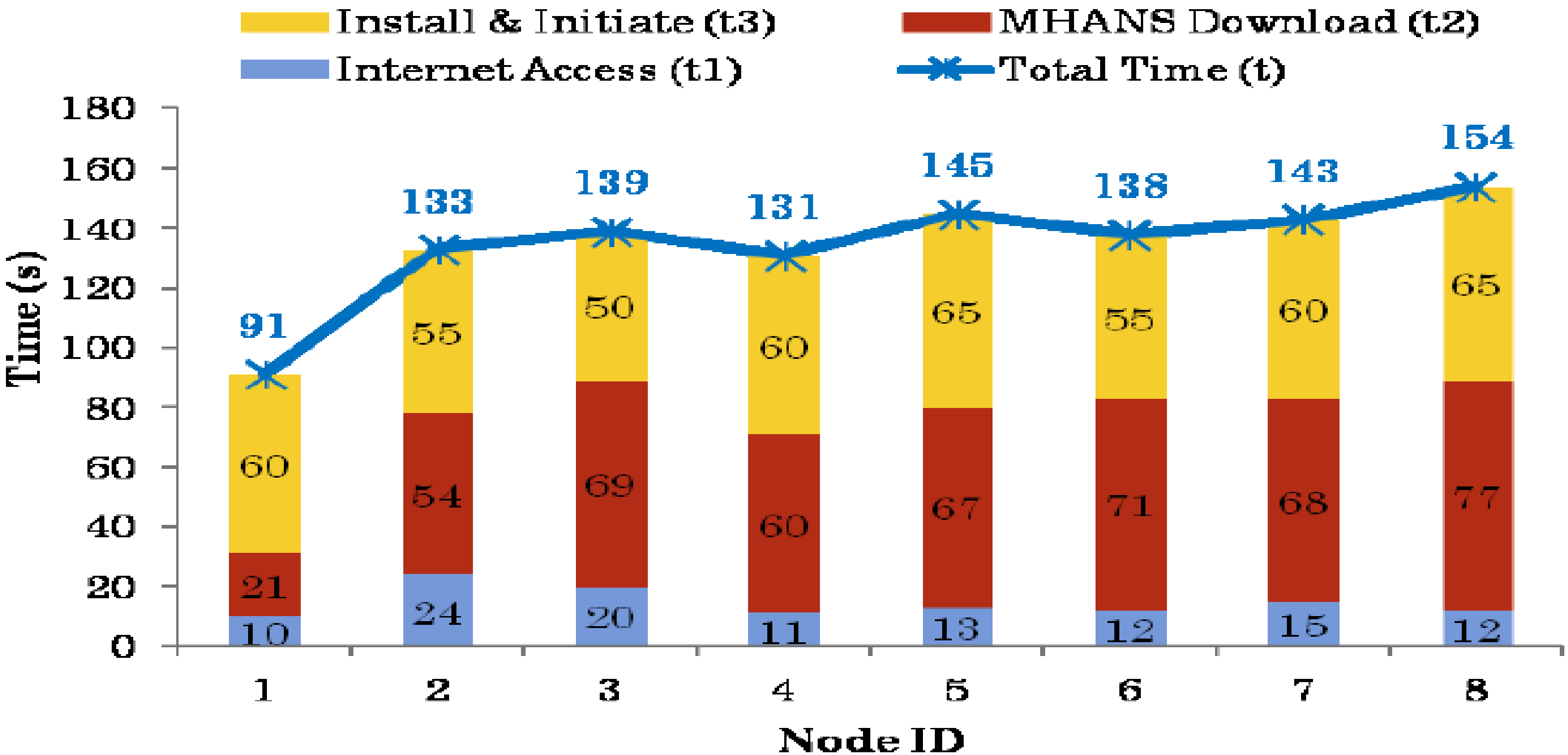
(b) Tree-based network

Parameters used in the Field Experiments

Parameter	Value/Description
Topology	Tree-based and Tandem networks
Environment	Indoor and outdoor settings
Hop distance	-Tree-based: 30m between levels -Tandem indoor (at IPU): 50m -Tandem outdoor (at ISU): 15m and 30m
Network size	-Tree-based outdoor: 3 levels (hops) and 15 nodes -Tandem indoor (at IPU): 14 hops (area: 700m in radius) -Tandem outdoor (at ISU): 20 hops and 16 hops (area: 300m and 480m in radius) for 15m-distance and 30m-distance networks, respectively
Mobile node (MN)	ASUS U24A-PX3210 laptop with 4GB memory, Core-i5 2.5Ghz CPU, Atheros AR9002WB-1NG WiFi, and Windows 7 OS
TCP window size	64KB
Buffer length (in Iperf)	8KB: Iperf works by writing an array of 8KB continuously
Maximum Transmission Unit (MTU)	1500 Bytes
Evaluation duration	100s
Wireless link	IEEE 802.11g
Packet size (FPing)	1470 Bytes

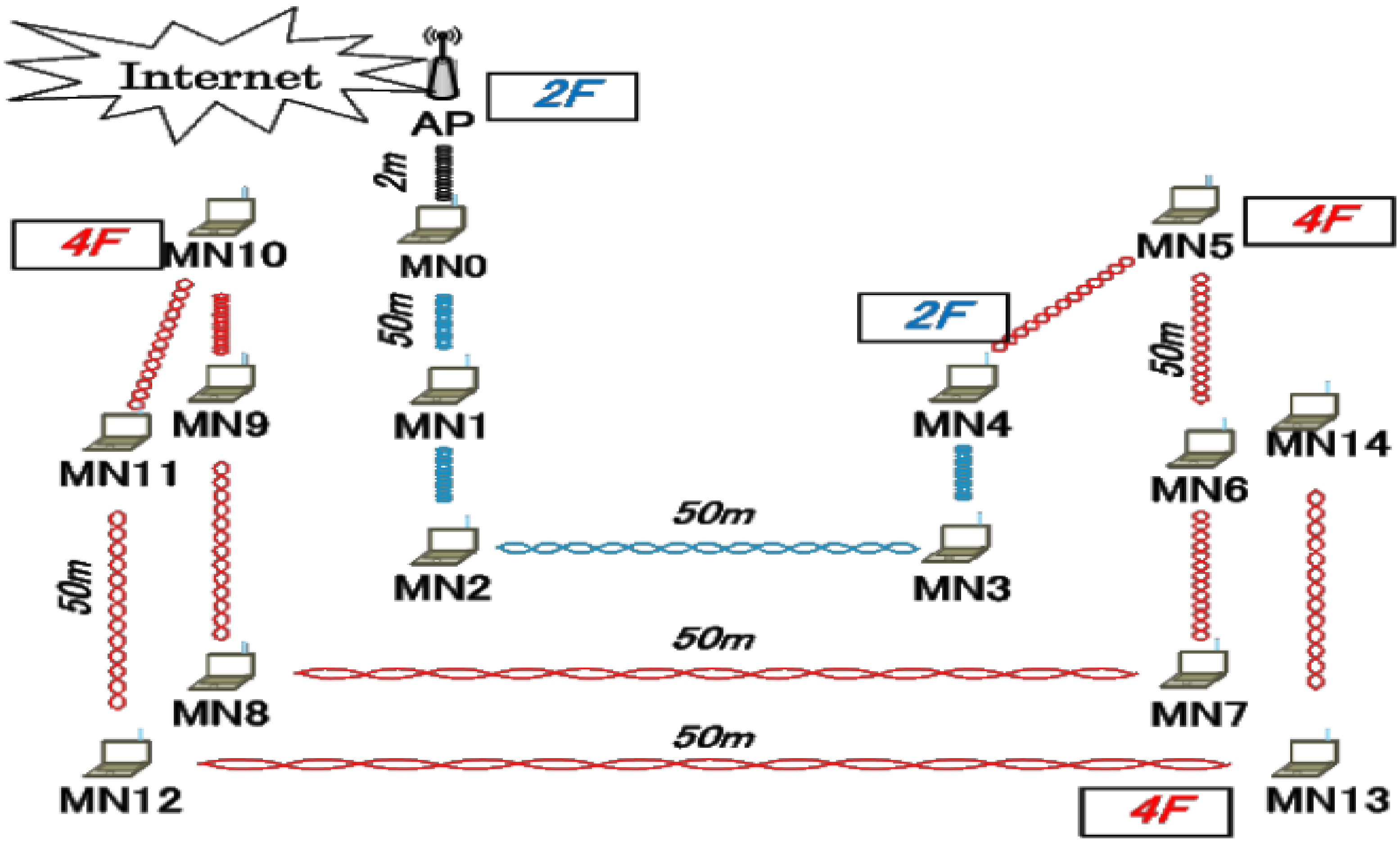
Network Set-up Time

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



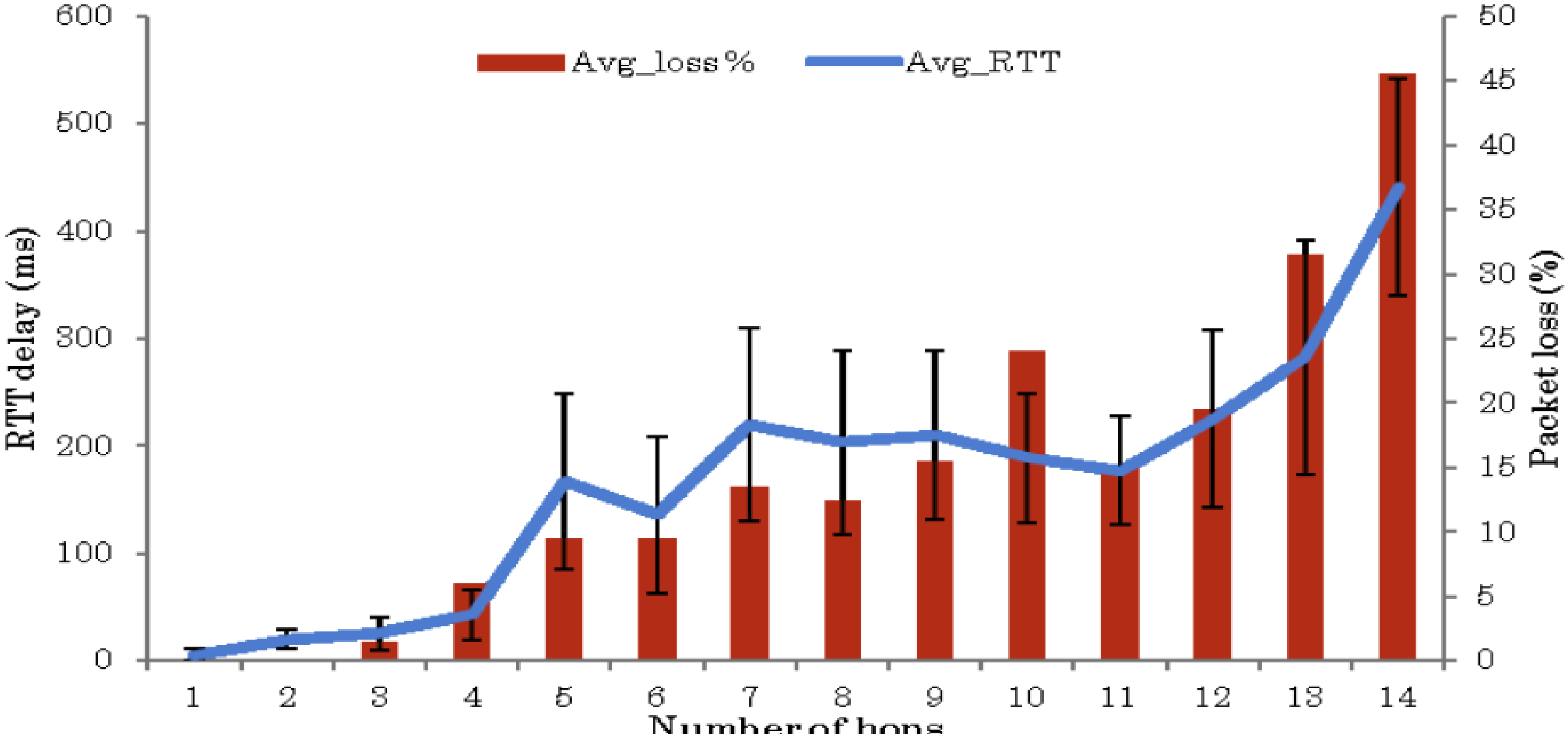
Experiment of Indoor Tandem-Connected Network with 50m Hop Distance

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



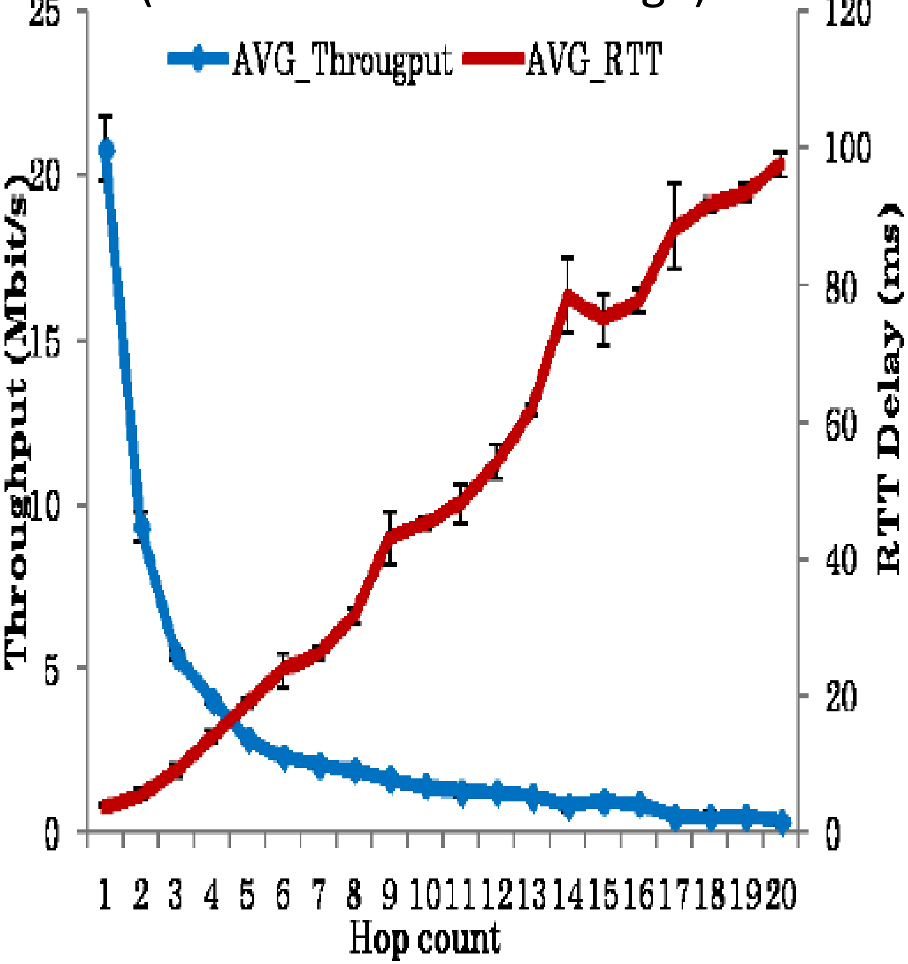
Round Tip Time (RTT) and Packet Loss of Indoor Tandem-Connected Network with 50m Hop Distance

- 20% Packet loss and 200ms round trip time (RTT) in 12 hops were still acceptable for ordinary Internet applications (Web browsing and Skype)
- 50 m hop distance up to 12 hops: 600 m distance coverage



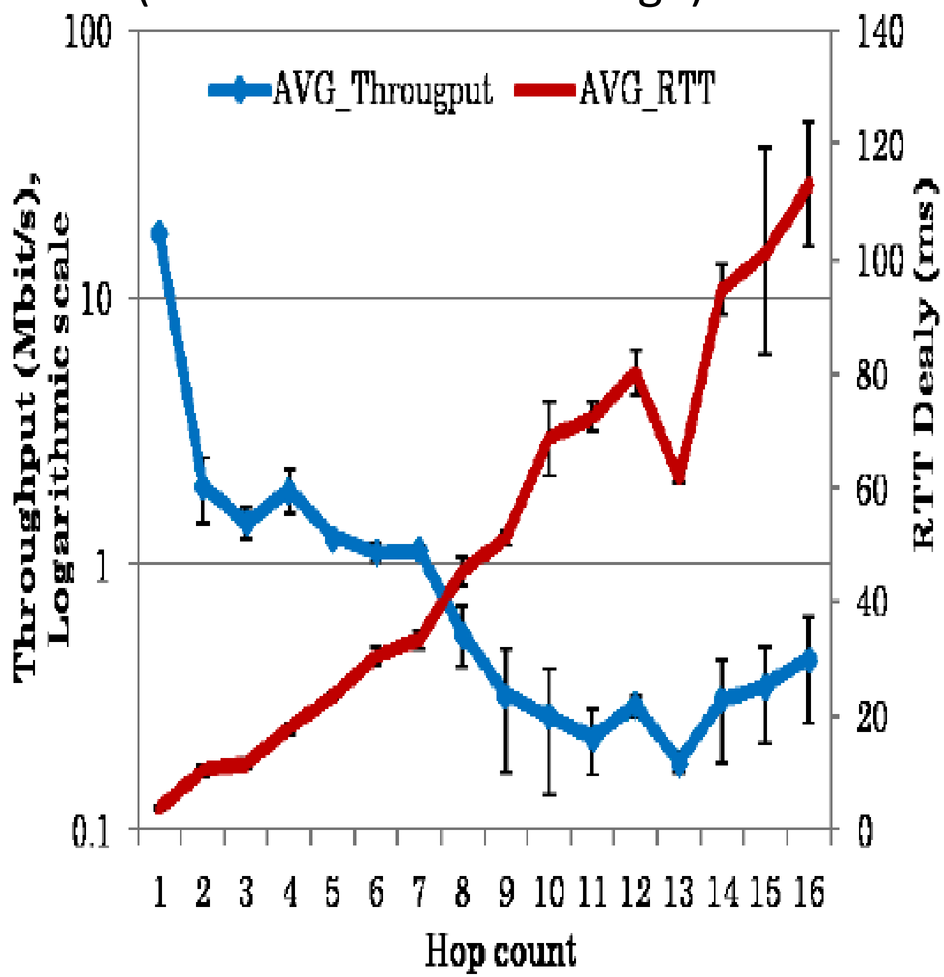
Round Tip Time (RTT) and Throughput of Outdoor Tandem-Connected Network

- 15 m Hop Distance up to 20 hops
- (300 m distance coverage)



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

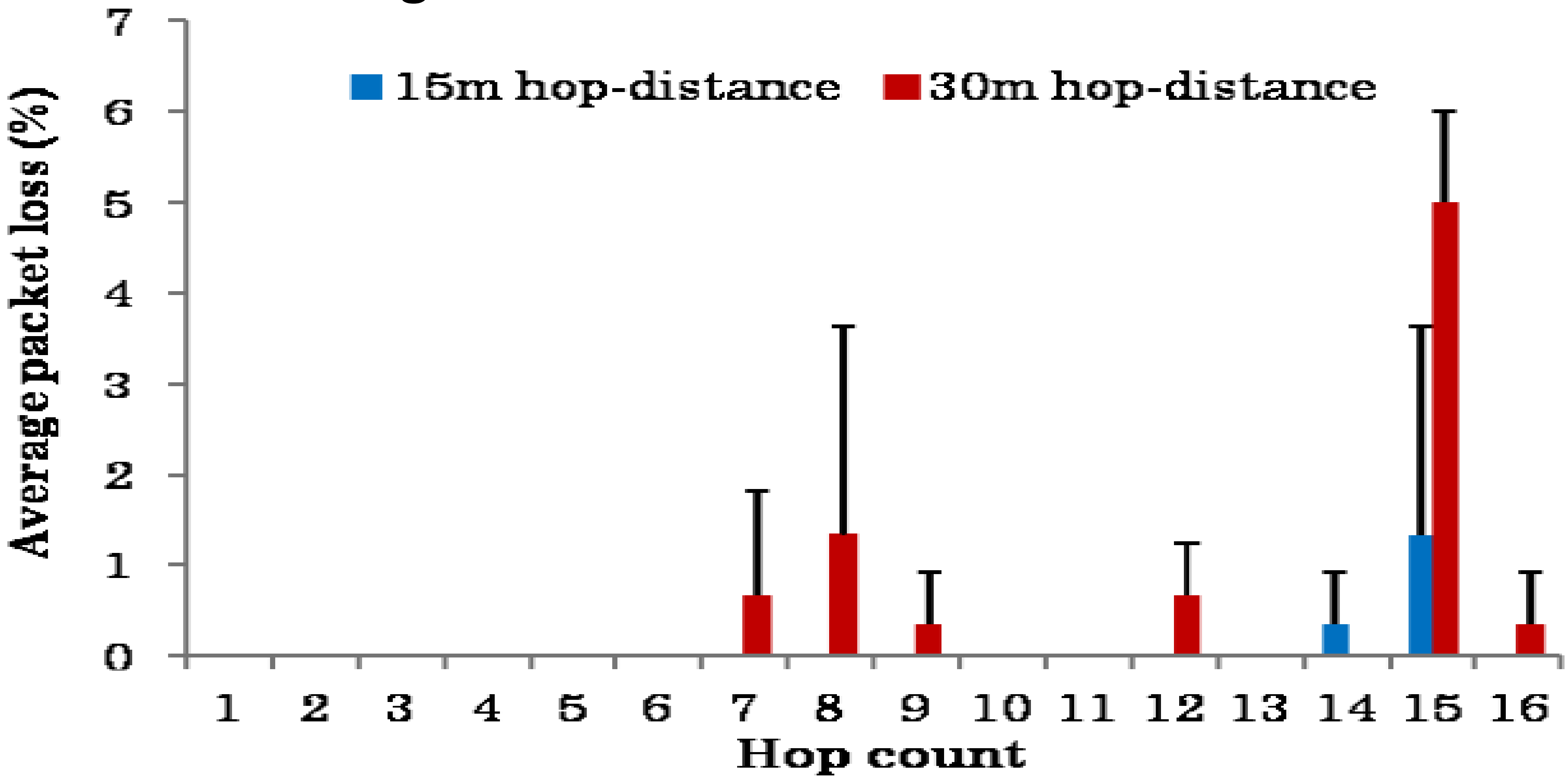
- 30 m Hop Distance up to 16 hops
- (480 m distance coverage)



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

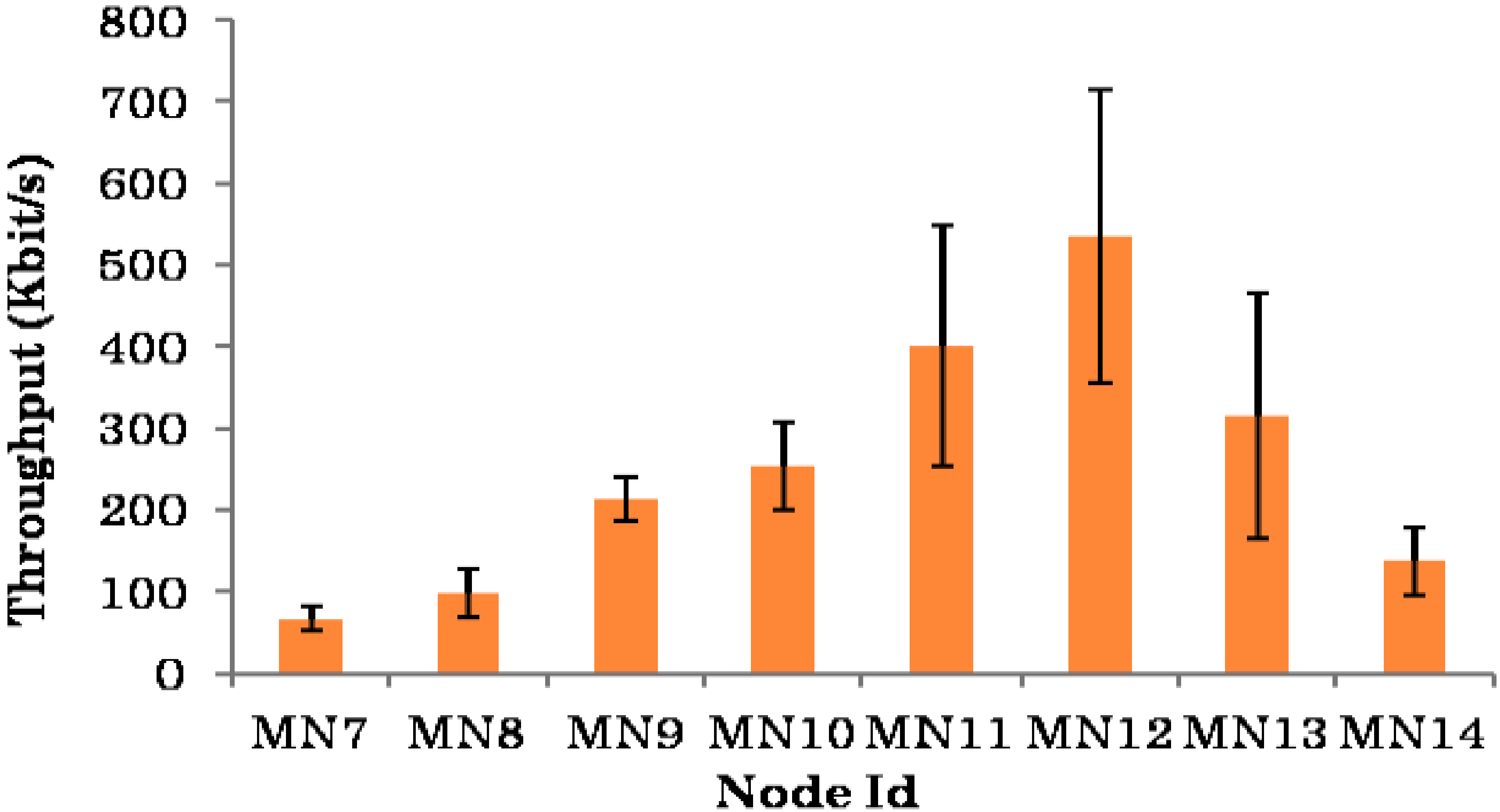
Packet Loss of Outdoor Tandem-Connected Network

- 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

- WiFi multihop access network is feasible for real deployments.
  - It can cover a large area of 500 m to 600 m in radius (more than one kilometer in diameter) in the indoor and outdoor environments for internet access in the disaster area
  - It allows ordinary people or volunteers in the disaster area to set up the network easily by themselves, using available commodity mobile devices.
  - The real deployment of these proposed technologies is also a challenge.
  - For further details,

    Quang Tran Minh, Kien Nguyen, Cristian Borcea, Shigeki Yamada: On-the-Fly Establishment of Multihop Wireless Access Networks for Disaster Recovery, IEEE Communications Magazine, Special Issue on Disaster Resilience in Communication Networks,   52(10) 60-66

Final Remarks

- Integrating SDN/OpenFlow-based technology and WiFi multihop access network technology with cloud/virtual machine technology could finally enable the whole network system to become more resilient to provide end-to-end seamless non-stoppable services.
- There exist a lot of disaster recovery research projects that have been financially supported by the Japanese Government.
- The academia in the world could do more to alleviate the damage of disasters. I encourage you to contribute to your future societies from any aspects of disaster recoveries.