IEEE NETSOFT

# SCLP: Segment-oriented Connection-less Protocol for High-Performance Software Tunneling in Datacenter Networks

Ryota Kawashima[†]

Shin Muramatsu[†]

Hiroki Nakayama[‡]

Tsunemasa Hayashi[‡]

Hiroshi Matsuo[†]

[†] Nagoya Institute of Technology, [‡] BOSCO Technologies, Inc.

## The Goal

Improving performance of overlay-based virtual networks

## Our Proposal

- VXLAN (SCLP)
- Geneve (SCLP)
- …

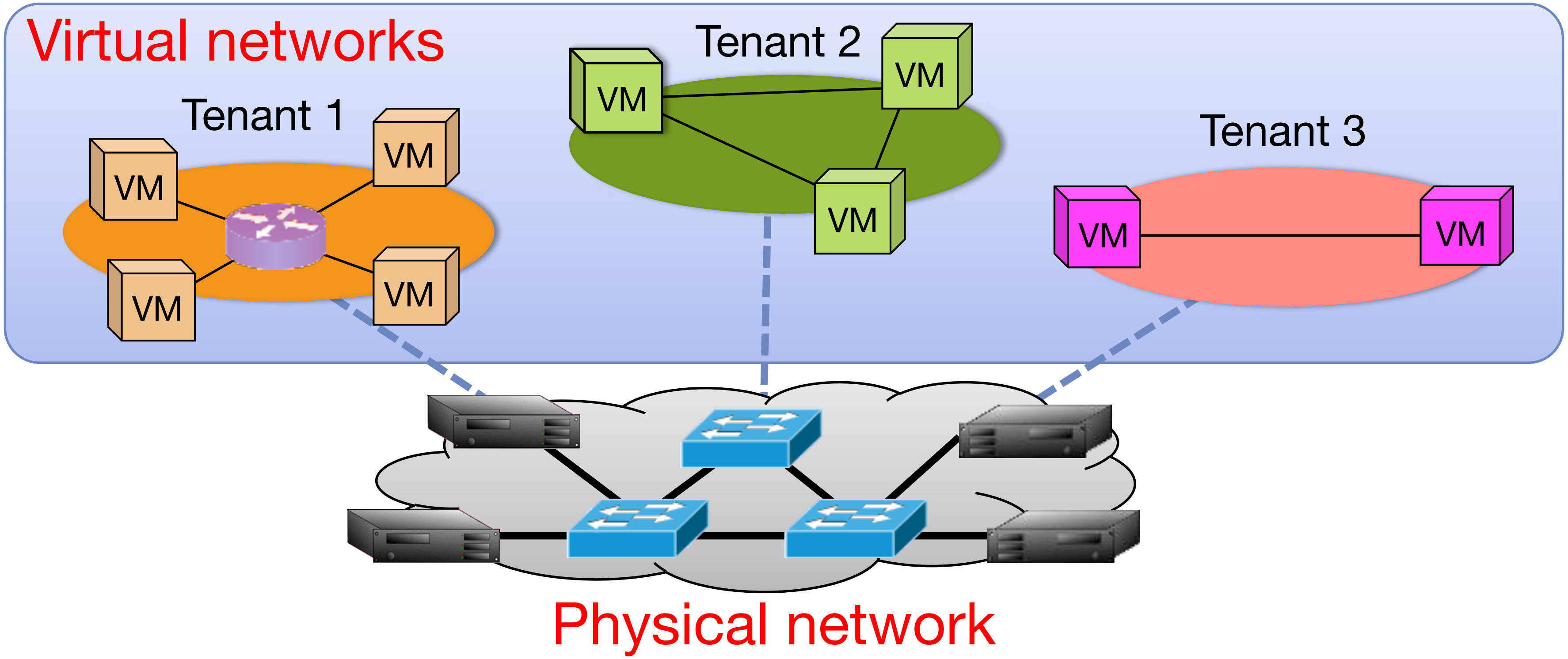
# Outline

- **Backgrounds**
  - Network Virtualization
  - Tunneling Protocols
  - L4 protocol characteristics
- **Proposal**
  - SCLP (Segment-oriented Connection-less Protocol)
- **Evaluation**
  - VM-to-VM communication using VXLAN over SCLP

# Network Virtualization

## ❖ Multi-tenant Datacenter Networks

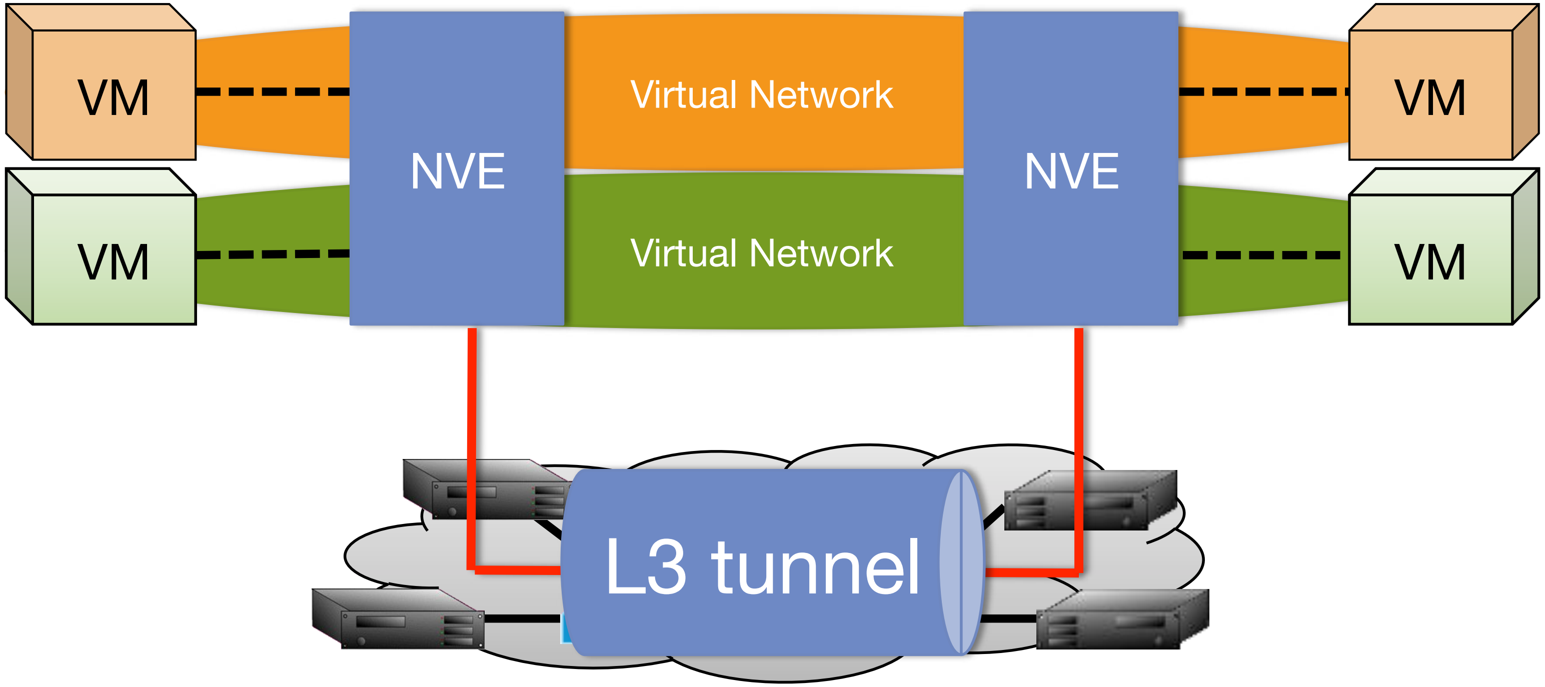
- Each tenant can have its own virtual networks
- Each virtual network shares the physical network resources



# The Overlay-based Approach

- **NVO3: Network Virtualization Overlays**
  - RFC 7364, 7365

NVE : Network Virtualization Edge



Physical network

# Outline

## ❖ Backgrounds

- ➢ Network Virtualization
- ➢ Tunneling Protocols
- ➢ L4 protocol characteristics

## ❖ Proposal

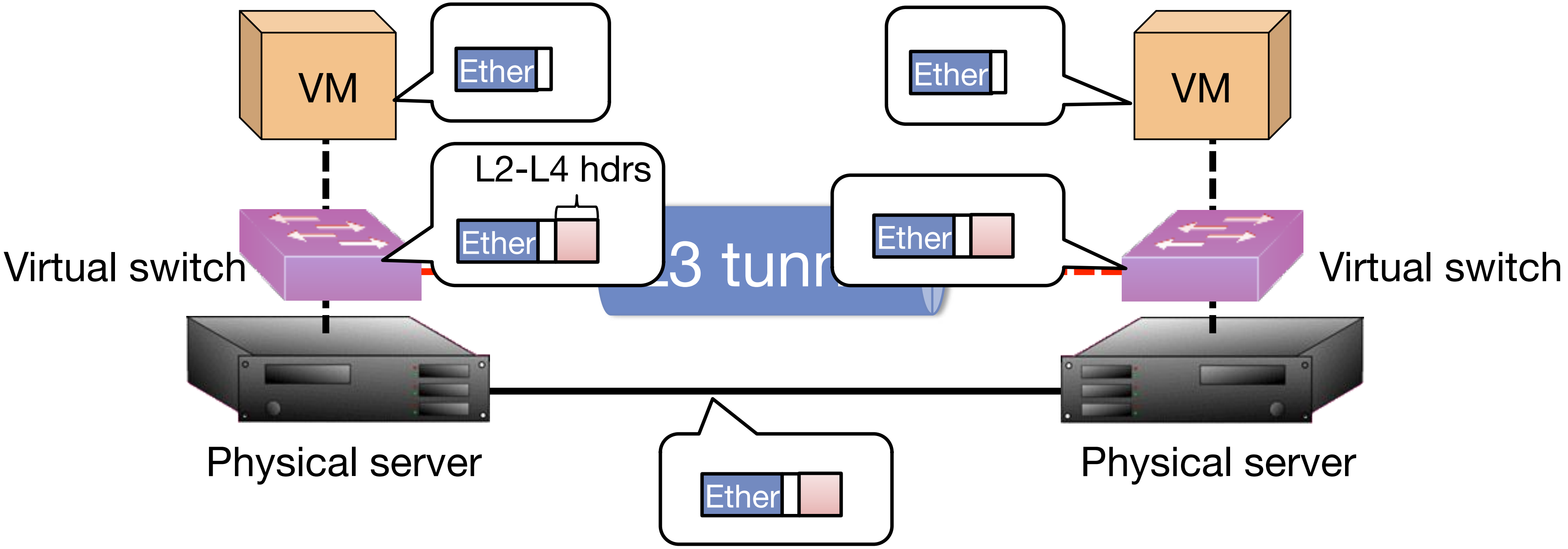
- ➢ SCLP (Segment-oriented Connection-less Protocol)

## ❖ Evaluation

- ➢ VM-to-VM communication using VXLAN over SCLP

# Tunneling Protocols

❖ **L2-in-L3 Tunneling**



# Major Tunneling Protocols

## ❖ VXLAN (RFC 7348)

- UDP based
- Linux kernel, OVS, VMware NSX, Cisco Nexus 1000V …

| | | | | | | |
|---|---|---|---|---|---|---|
| Ethernet<br>(Physical) | IP<br>(Physical) | UDP | VXLAN | Ethernet<br>(Virtual) | Payload | FCS |

## ❖ NVGRE (RFC draft)

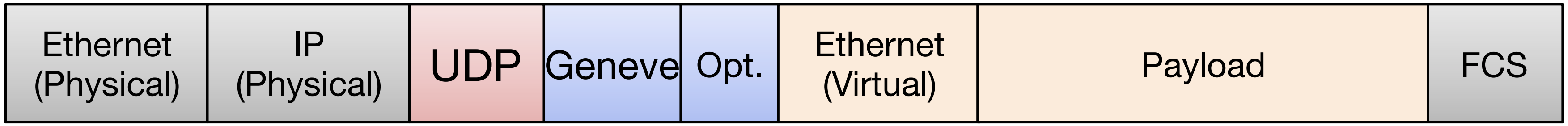
- GRE based (no L4 protocol)
- Microsoft Hyper-V

| | | | | | |
|---|---|---|---|---|---|
| Ethernet<br>(Physical) | IP<br>(Physical) | NVGRE | Ethernet<br>(Virtual) | Payload | FCS |

# Upcoming Tunneling Protocol

## ❖ Geneve (RFC draft)

- UDP based
- TLV option header
- H/W segmentation offload (future)

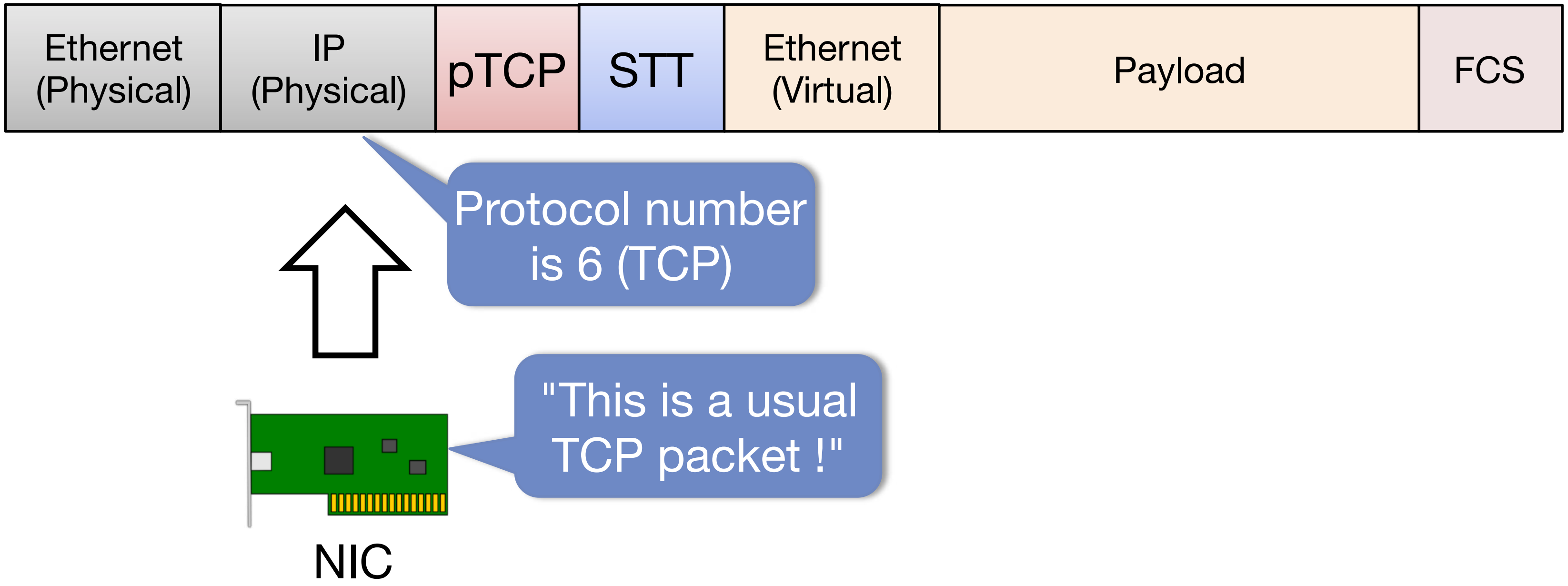


| Ethernet (Physical) | IP (Physical) | UDP | Geneve | Opt. | Ethernet (Virtual) | Payload | FCS |
|---|---|---|---|---|---|---|---|

# Yet Another Tunneling Protocol

## ❖ STT (Stateless Transport Tunneling, RFC draft)

- Pseudo-TCP header
  - ➢ Exploiting TSO (TCP Segmentation Offload) feature
  - ➢ Semantics of header fields are modified
- VMware NSX



# Problems of Existing Protocols

## ❖ Performance

- VXLAN, NVGRE, Geneve

Maximum throughput falls to one-half !

## ❖ Compatibility

- STT

Middleboxes can discard STT packets !

# Outline

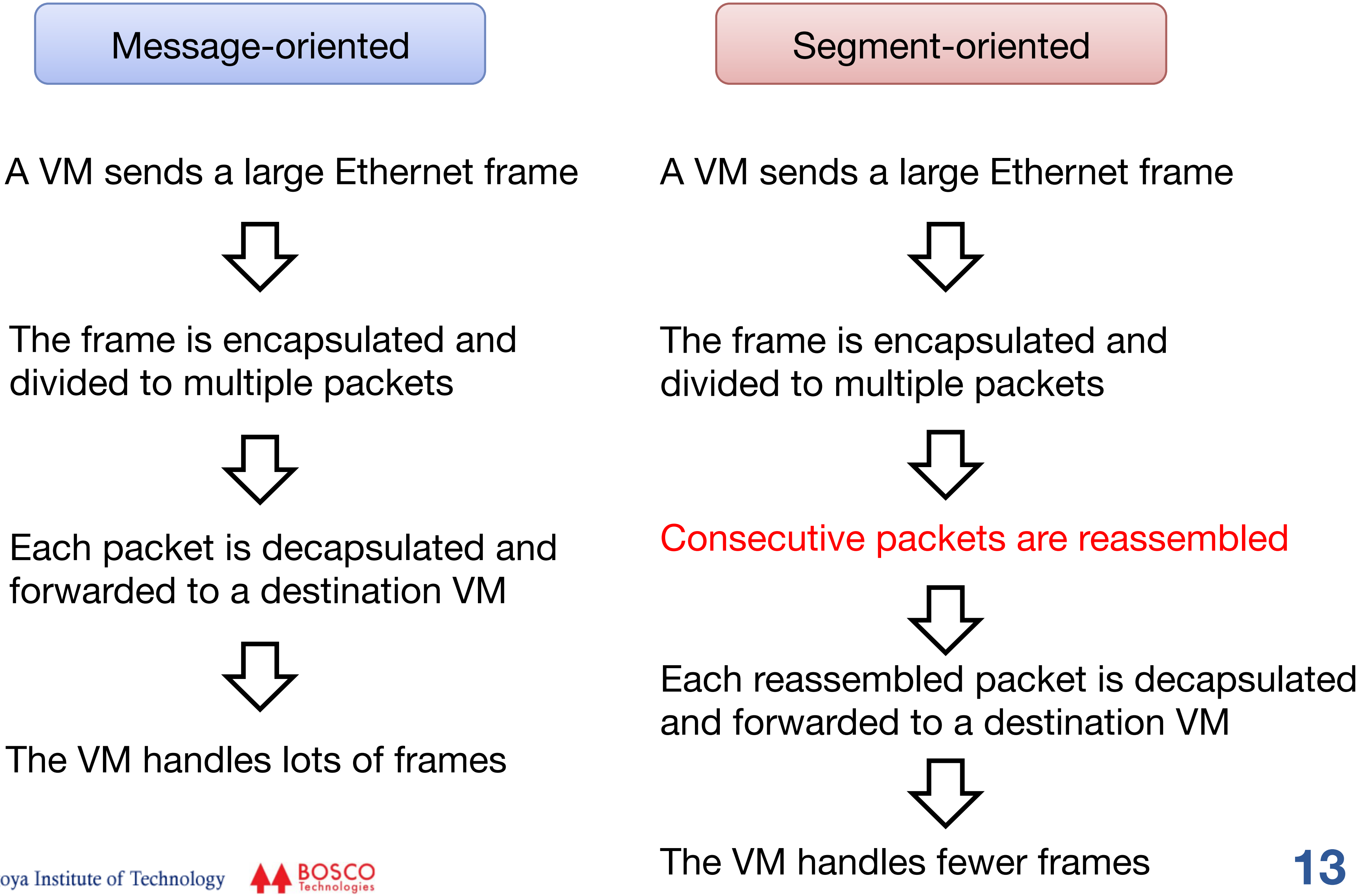
- **Backgrounds**
  - Network Virtualization
  - Tunneling Protocols
  - L4 protocol characteristics
- **Proposal**
  - SCLP (Segment-oriented Connection-less Protocol)
- **Evaluation**
  - VM-to-VM communication using VXLAN over SCLP

# L4 Protocol Types

- **Message-oriented  (e.g. UDP)**
  - Packets are independent of each other

- **Segment-oriented (e.g. TCP)**
  - Each packet has “byte-level sequence”
  - Consecutive packets can be reassembled

# Why is L4 Protocol Important ?



Message-oriented

A VM sends a large Ethernet frame

The frame is encapsulated and divided to multiple packets

Each packet is decapsulated and forwarded to a destination VM

The VM handles lots of frames

Segment-oriented

A VM sends a large Ethernet frame

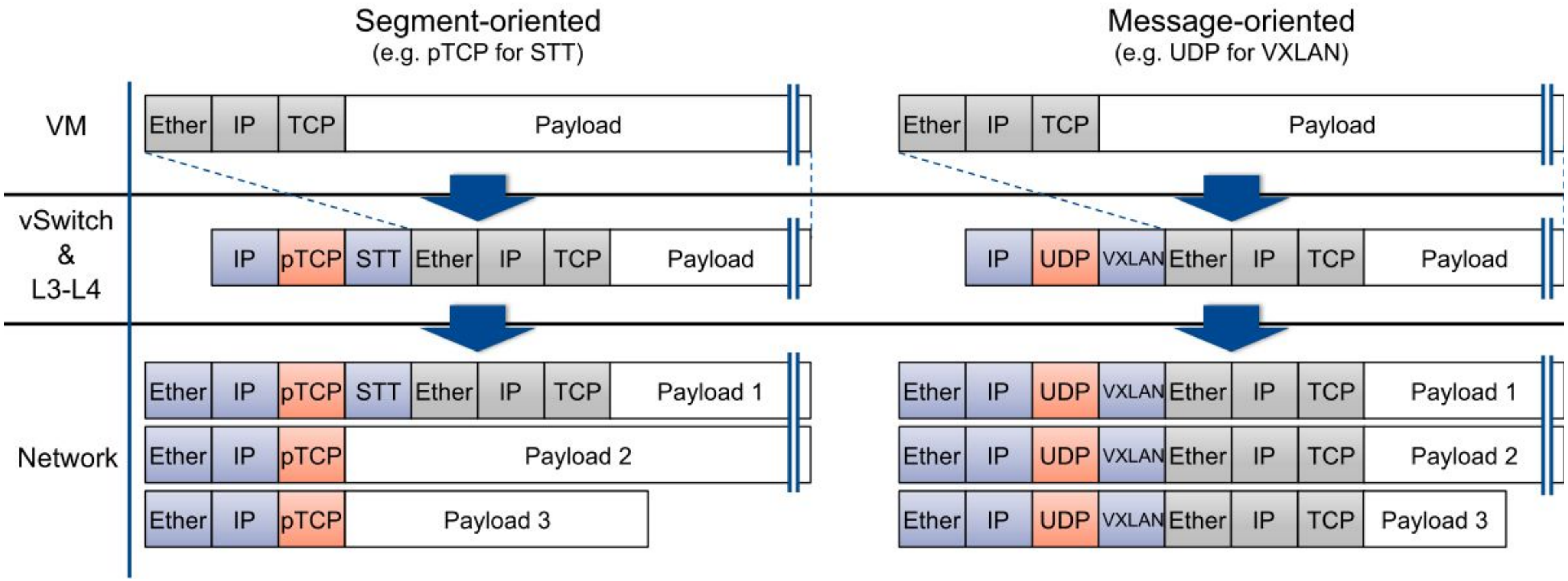
The frame is encapsulated and divided to multiple packets

Consecutive packets are reassembled

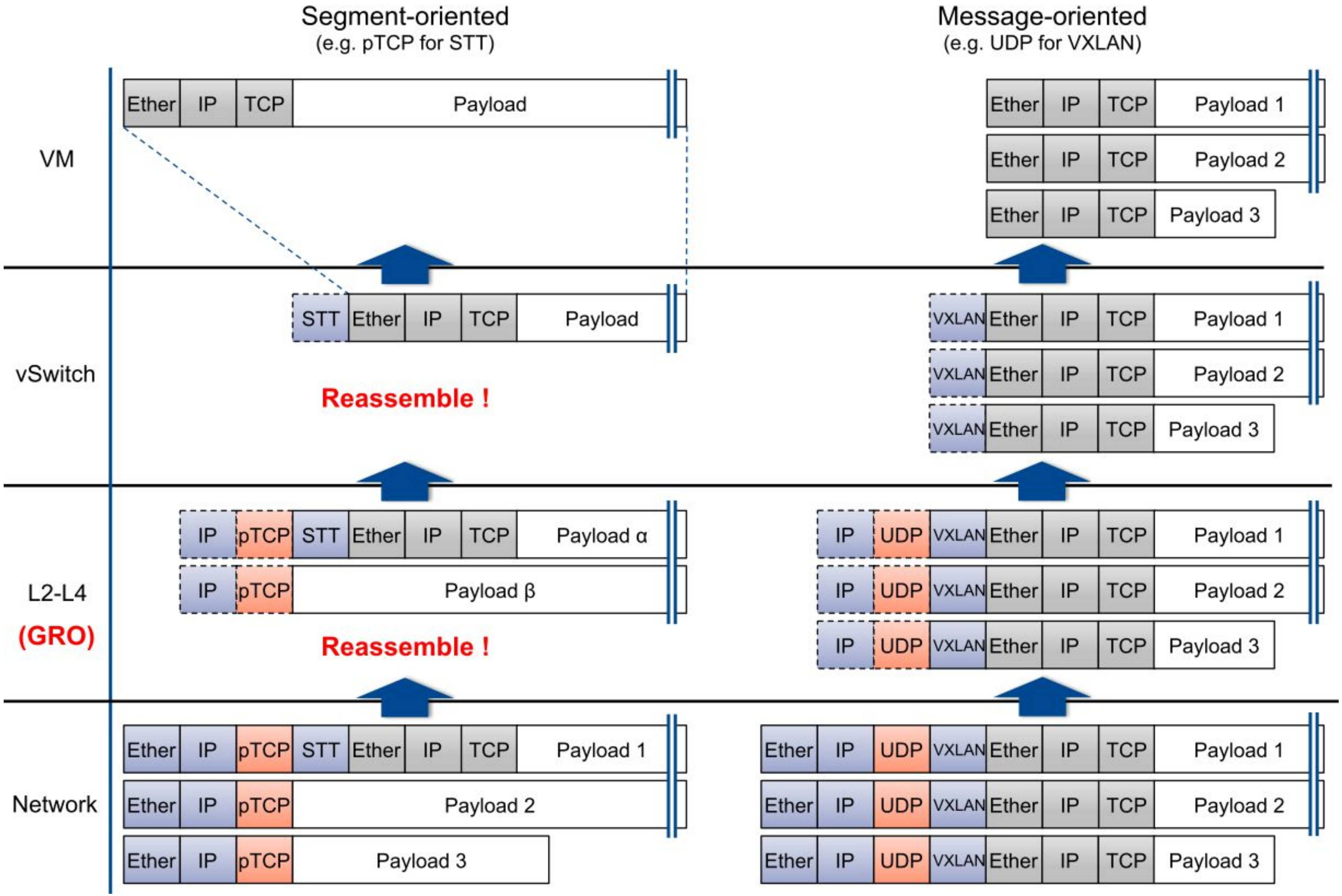
Each reassembled packet is decapsulated and forwarded to a destination VM

The VM handles fewer frames

# Packet Structure Example (Tx)



# Packet Structure Example (Rx)



Segment-oriented
(e.g. pTCP for STT)

Message-oriented
(e.g. UDP for VXLAN)

VM

Ether | IP | TCP | Payload

Ether | IP | TCP | Payload 1
Ether | IP | TCP | Payload 2
Ether | IP | TCP | Payload 3

vSwitch

STT | Ether | IP | TCP | Payload

Reassemble !

VXLAN | Ether | IP | TCP | Payload 1
VXLAN | Ether | IP | TCP | Payload 2
VXLAN | Ether | IP | TCP | Payload 3

L2-L4
(GRO)

IP | pTCP | STT | Ether | IP | TCP | Payload α
IP | pTCP | Payload β

Reassemble !

IP | UDP | VXLAN | Ether | IP | TCP | Payload 1
IP | UDP | VXLAN | Ether | IP | TCP | Payload 2
IP | UDP | VXLAN | Ether | IP | TCP | Payload 3

Network

Ether | IP | pTCP | STT | Ether | IP | TCP | Payload 1
Ether | IP | pTCP | Payload 2
Ether | IP | pTCP | Payload 3

Ether | IP | UDP | VXLAN | Ether | IP | TCP | Payload 1
Ether | IP | UDP | VXLAN | Ether | IP | TCP | Payload 2
Ether | IP | UDP | VXLAN | Ether | IP | TCP | Payload 3

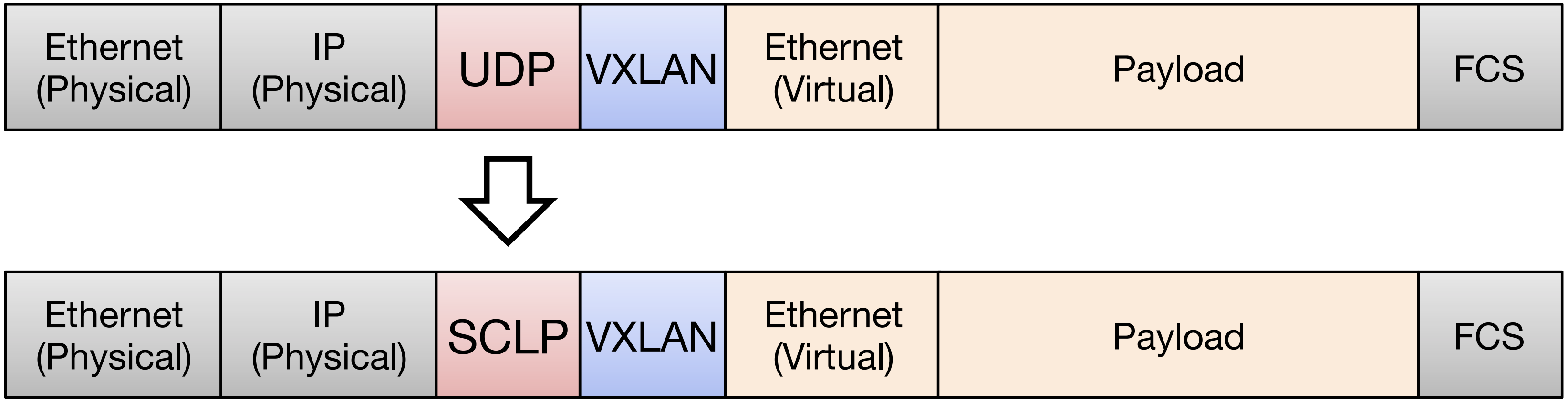
# Outline

- **Backgrounds**
  - Network Virtualization
  - Tunneling Protocols
  - L4 protocol characteristics
- **Proposal**
  - SCLP (Segment-oriented Connection-less Protocol)
- **Evaluation**
  - VM-to-VM communication using VXLAN over SCLP

# Our Proposal

❖ **SCLP: Segment-oriented Connection-less Protocol**

- L4 protocol
- Segment-oriented
- Connection-less
- Usage: Outer L4 protocol of existing tunneling protocols
    - ➢ e.g.) VXLAN over SCLP, Geneve over SCLP

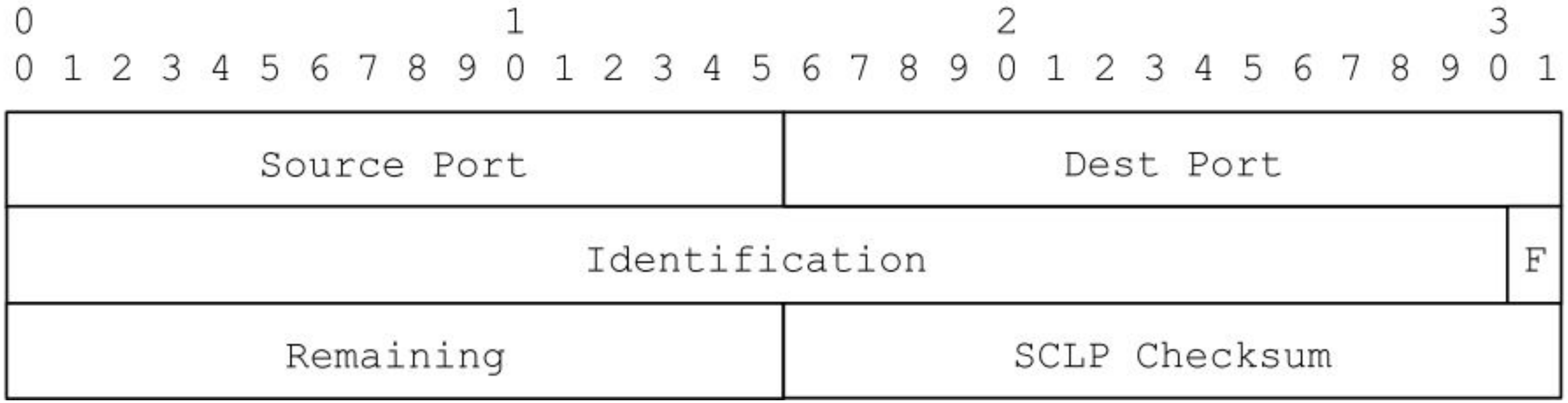


| | | | | | | |
|---|---|---|---|---|---|---|
| Ethernet (Physical) | IP (Physical) | UDP | VXLAN | Ethernet (Virtual) | Payload | FCS |

⇩

| | | | | | | |
|---|---|---|---|---|---|---|
| Ethernet (Physical) | IP (Physical) | SCLP | VXLAN | Ethernet (Virtual) | Payload | FCS |

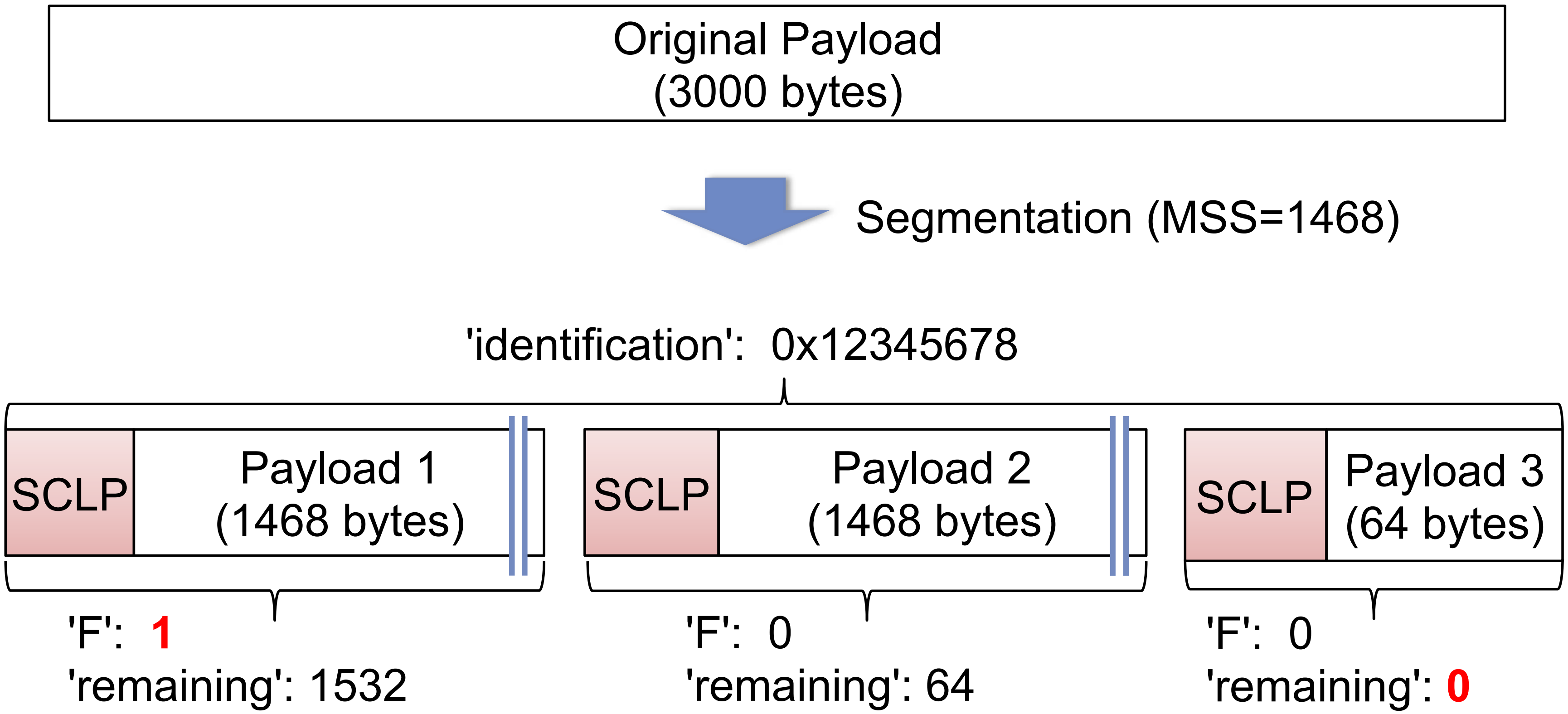
# Protocol Format



```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-------------------------------+-------------------------------+
|          Source Port          |           Dest Port           |
+-------------------------------+-----------------------------+-+
|                       Identification                        |F|
+-------------------------------+-----------------------------+-+
|           Remaining           |         SCLP Checksum         |
+-------------------------------+-------------------------------+
```

- **Identification: Original payload ID**
- **F: First Segment Flag**
- **Remaining: Remaining payload size**

# How SCLP Works (Tx)



Original Payload
(3000 bytes)

Segmentation (MSS=1468)

'identification': 0x12345678

SCLP | Payload 1 (1468 bytes)

SCLP | Payload 2 (1468 bytes)

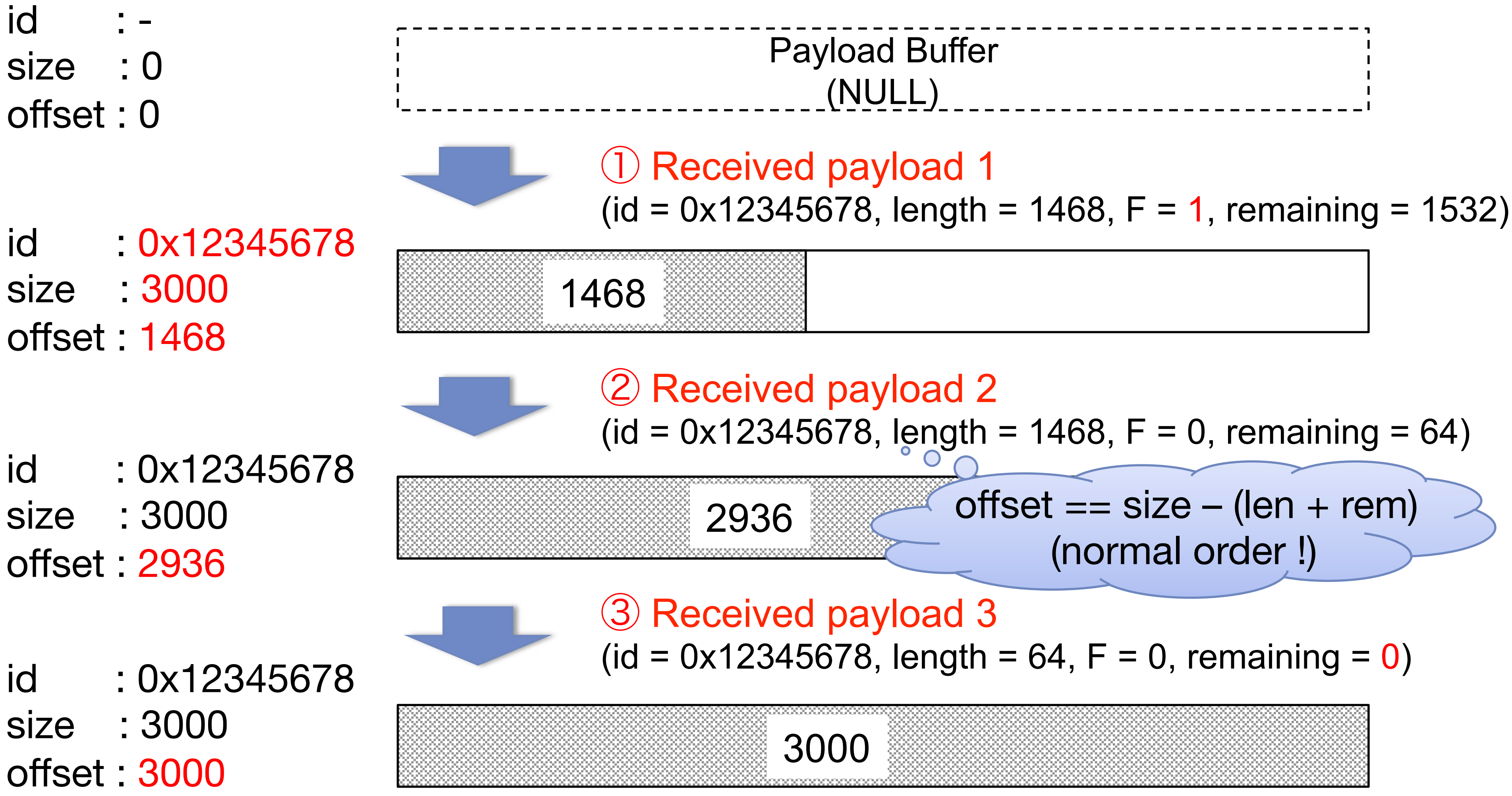
SCLP | Payload 3 (64 bytes)

'F': 1
'remaining': 1532

'F': 0
'remaining': 64

'F': 0
'remaining': 0

# How SCLP Works (Rx)



id : -
size : 0
offset : 0

Payload Buffer
(NULL)

① Received payload 1
(id = 0x12345678, length = 1468, F = 1, remaining = 1532)

id : 0x12345678
size : 3000
offset : 1468

1468

② Received payload 2
(id = 0x12345678, length = 1468, F = 0, remaining = 64)

id : 0x12345678
size : 3000
offset : 2936

2936

offset == size – (len + rem)
(normal order !)

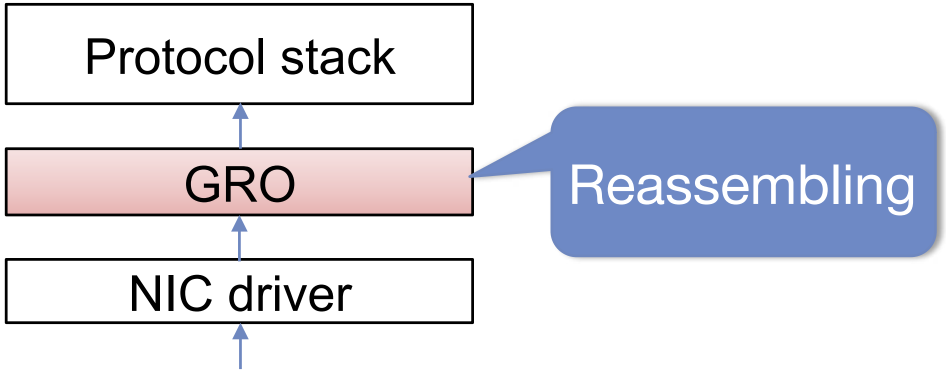
③ Received payload 3
(id = 0x12345678, length = 64, F = 0, remaining = 0)

id : 0x12345678
size : 3000
offset : 3000

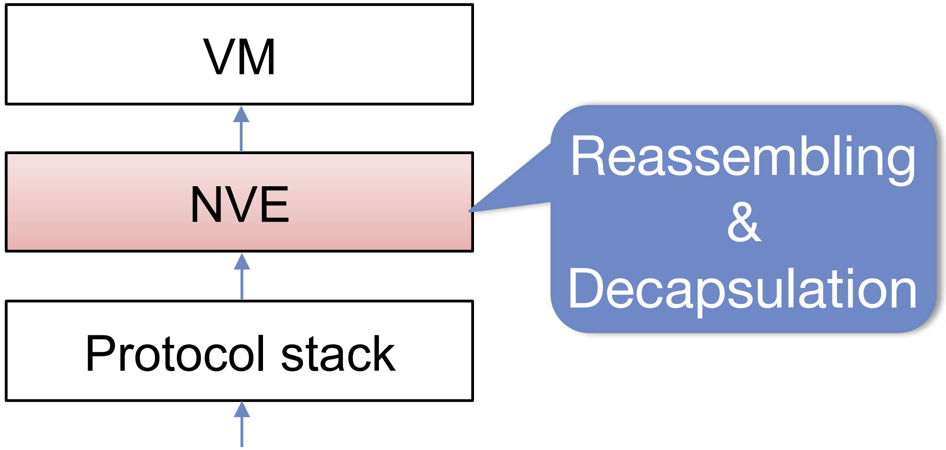
3000

# 2-Level Pre-reassembling

- 1st: GRO (Generic Receive Offload)

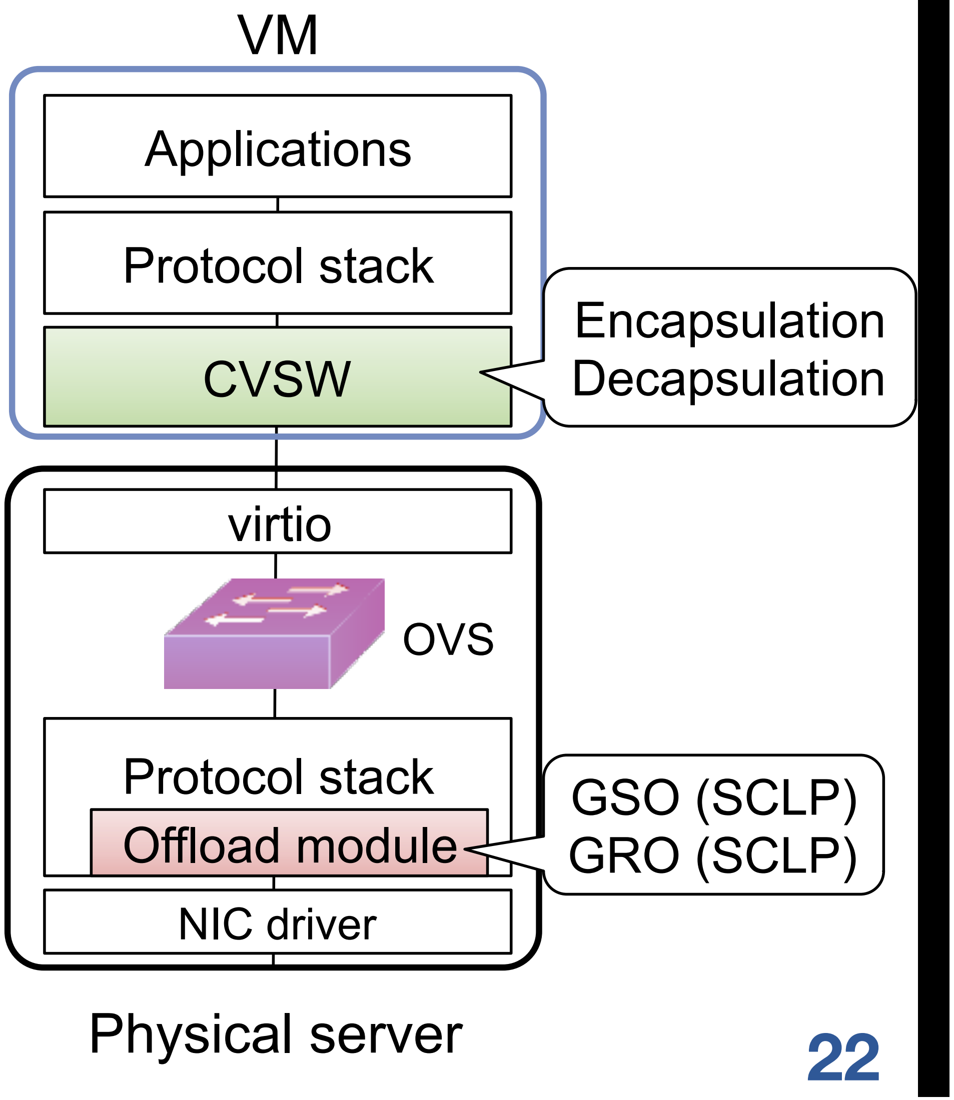


- 2nd: NVE's decapsulation processing



# Implementation

- **VXLAN over SCLP**
  - CVSW component †
  - Virtual NIC implementation of NVE
- **GSO/GRO offloading**
  - Linux kernel module



† https://github.com/sdnnit/cvsw_net

# Outline

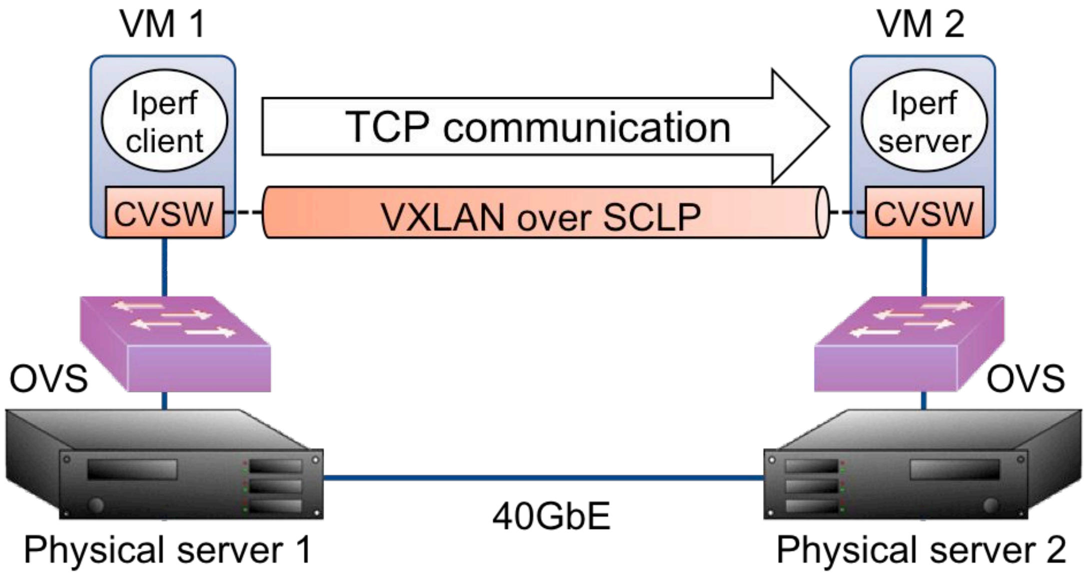
- **Backgrounds**
  - Network Virtualization
  - Tunneling Protocols
  - L4 protocol characteristics
- **Proposal**
  - SCLP (Segment-oriented Connection-less Protocol)
- **Evaluation**
  - VM-to-VM communication using VXLAN over SCLP

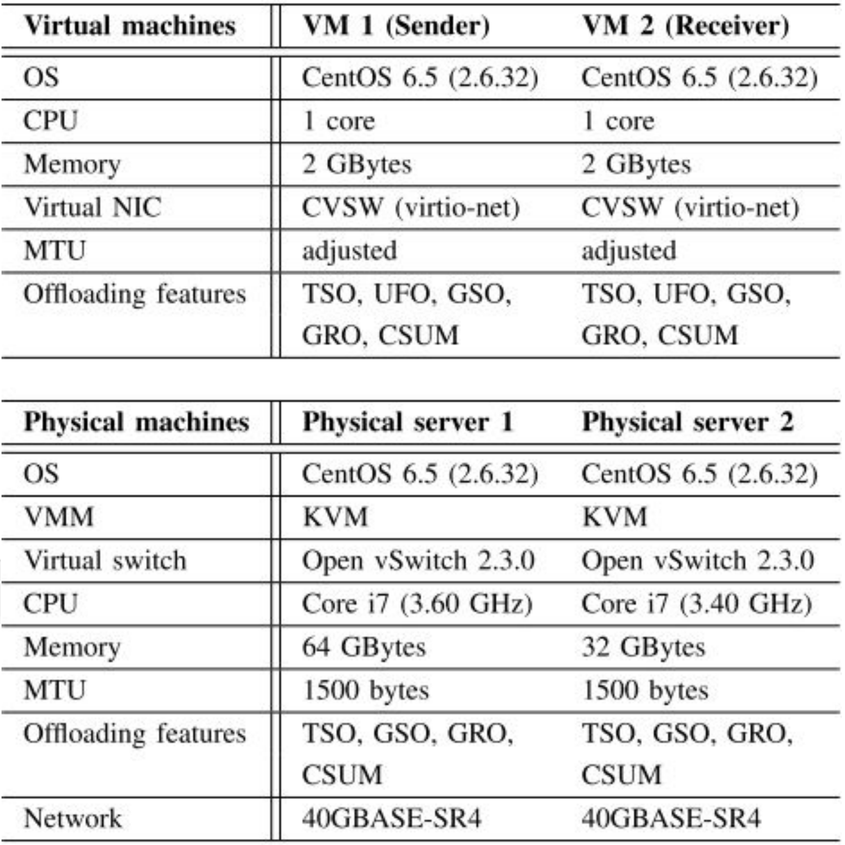
# Evaluation

- **Throughput of VM-to-VM communication with Iperf**
  1. TCP communication
  2. Effect of 2-level pre-reassembling

- **Competitors**
  - VXLAN (over UDP)
  - NVGRE
  - STT
  - Geneve (w/o HW offloading)

# Evaluation Environment



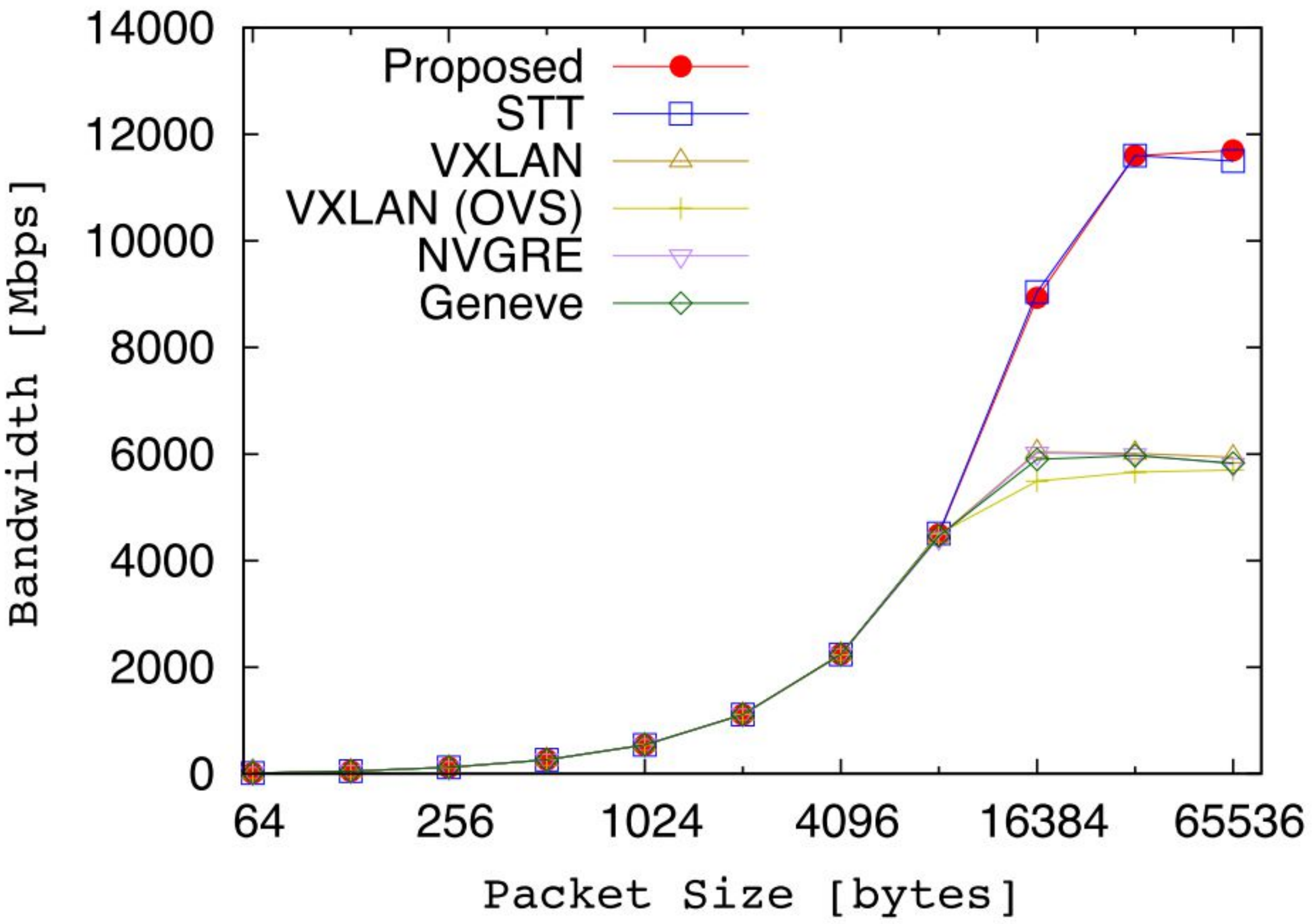


| Virtual machines | VM 1 (Sender) | VM 2 (Receiver) |
|---|---|---|
| OS | CentOS 6.5 (2.6.32) | CentOS 6.5 (2.6.32) |
| CPU | 1 core | 1 core |
| Memory | 2 GBytes | 2 GBytes |
| Virtual NIC | CVSW (virtio-net) | CVSW (virtio-net) |
| MTU | adjusted | adjusted |
| Offloading features | TSO, UFO, GSO, GRO, CSUM | TSO, UFO, GSO, GRO, CSUM |

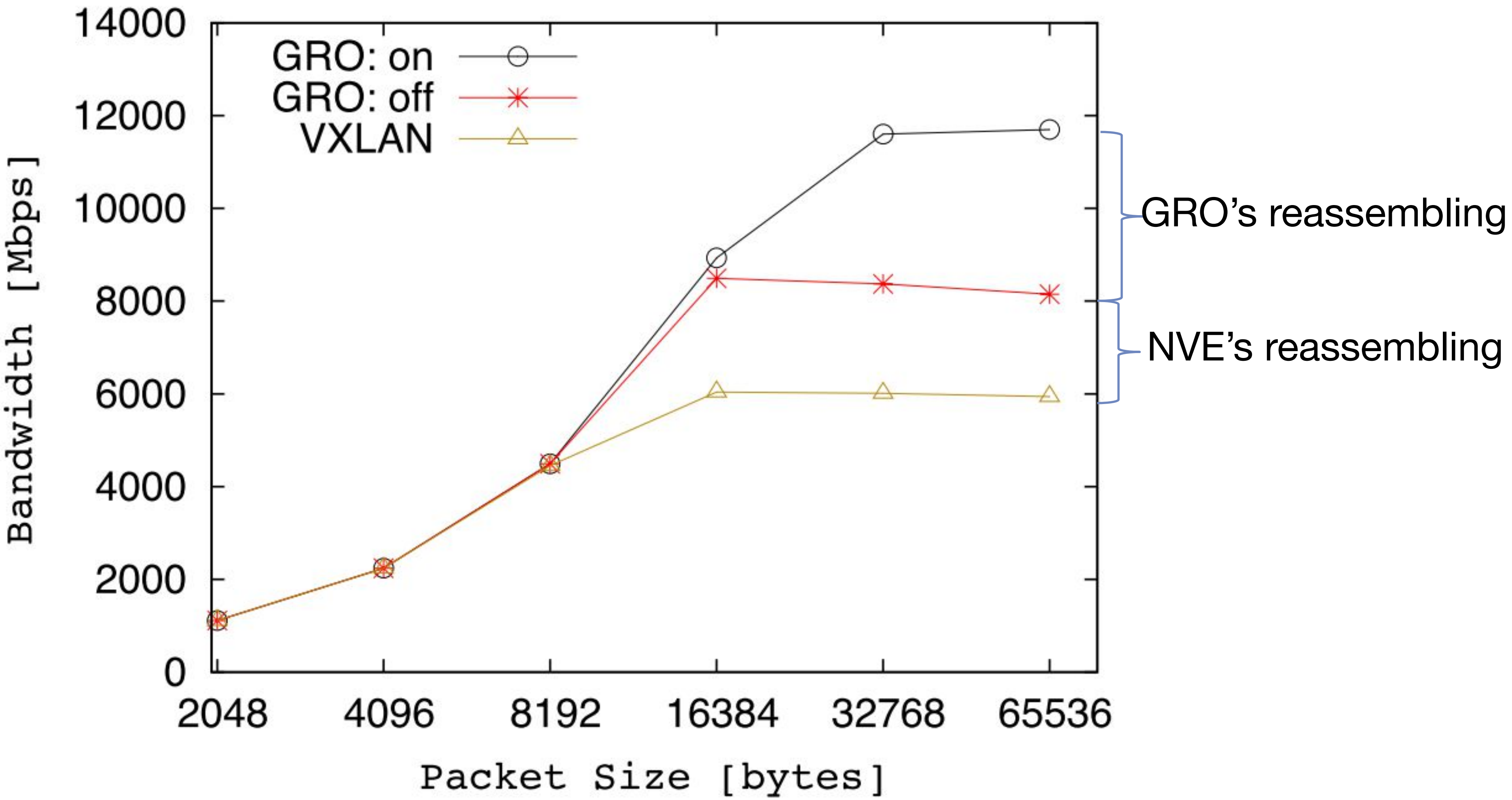
| Physical machines | Physical server 1 | Physical server 2 |
|---|---|---|
| OS | CentOS 6.5 (2.6.32) | CentOS 6.5 (2.6.32) |
| VMM | KVM | KVM |
| Virtual switch | Open vSwitch 2.3.0 | Open vSwitch 2.3.0 |
| CPU | Core i7 (3.60 GHz) | Core i7 (3.40 GHz) |
| Memory | 64 GBytes | 32 GBytes |
| MTU | 1500 bytes | 1500 bytes |
| Offloading features | TSO, GSO, GRO, CSUM | TSO, GSO, GRO, CSUM |
| Network | 40GBASE-SR4 | 40GBASE-SR4 |

# Evaluation Results (TCP)

- **TCP communication**



# Evaluation Results (Pre-reassembling)



GRO's reassembling

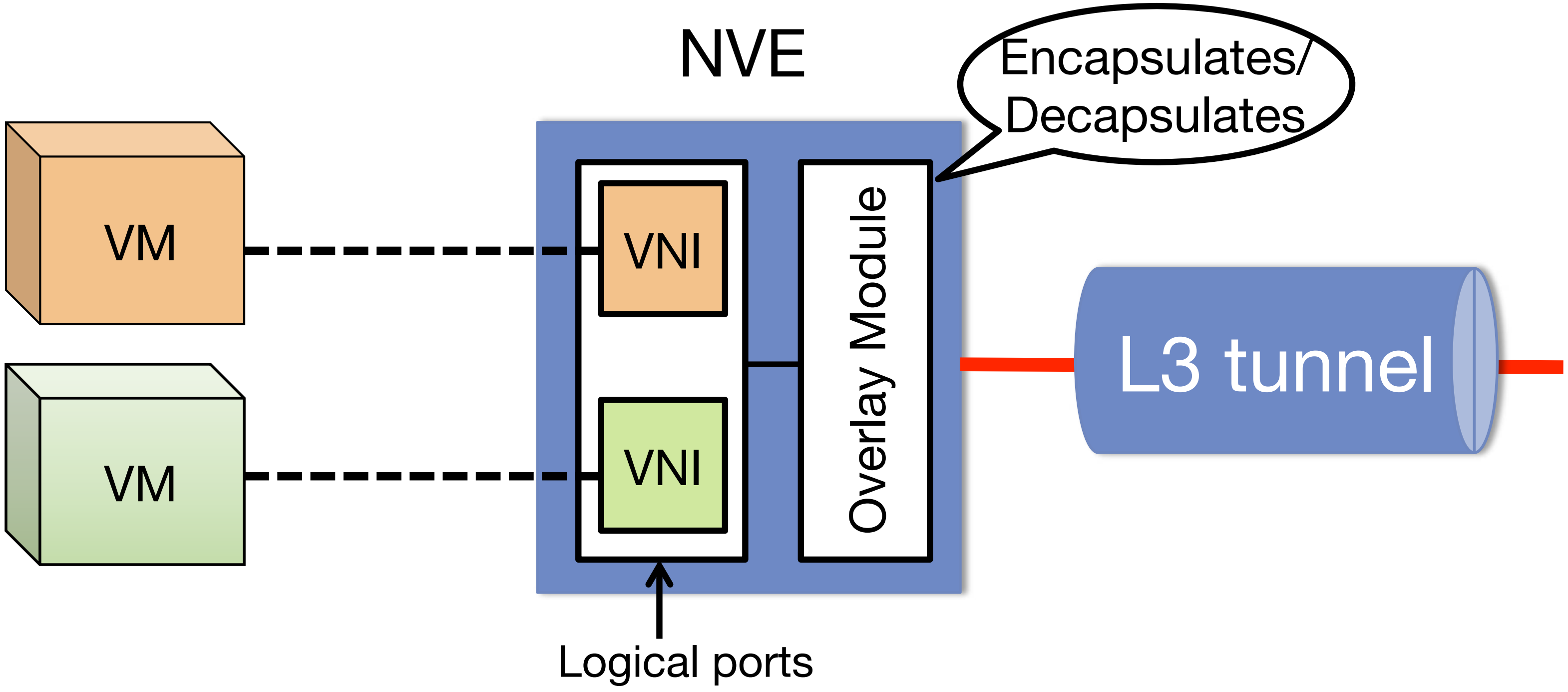
NVE's reassembling

# Conclusion

- **Network virtualization**
  - Overlay-based approach has become popular
  - VXLAN is a de-facto tunneling protocol
  - UDP-based tunneling has performance problems
- **Proposal: SCLP**
  - Segment-oriented and connection-less L4 protocol
  - 2-level pre-reassembling before decapsulation
  - STT-comparable performance
- **Future work**
  - Implementation of OVS-based SCLP
  - Open source

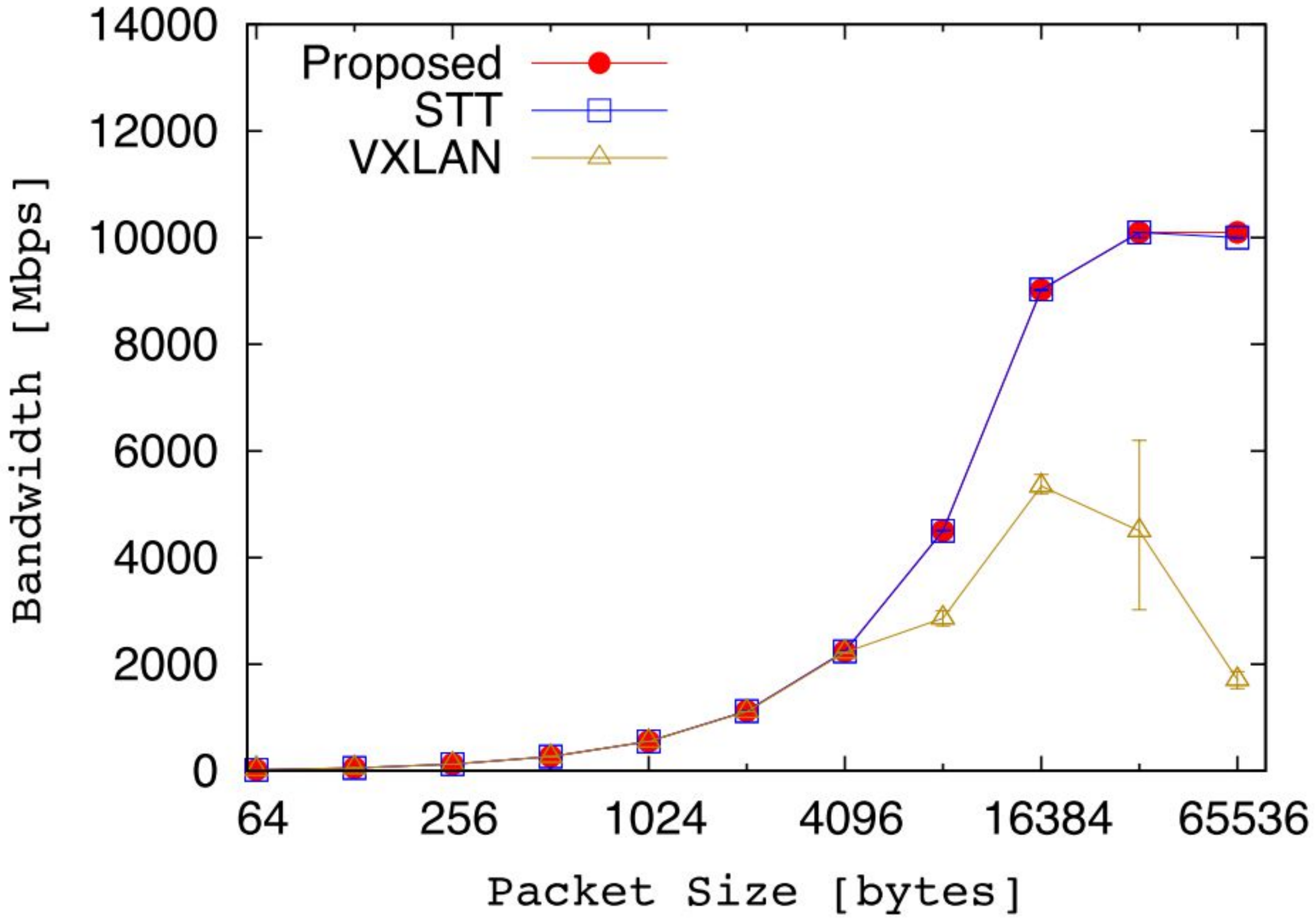
# NVE: Network Virtualization Edge

- **Tunnel End-Point**
  - Physical switches
  - Virtual switches
    - Open vSwitch (OVS), NSX switch, Hyper-V virtual switch



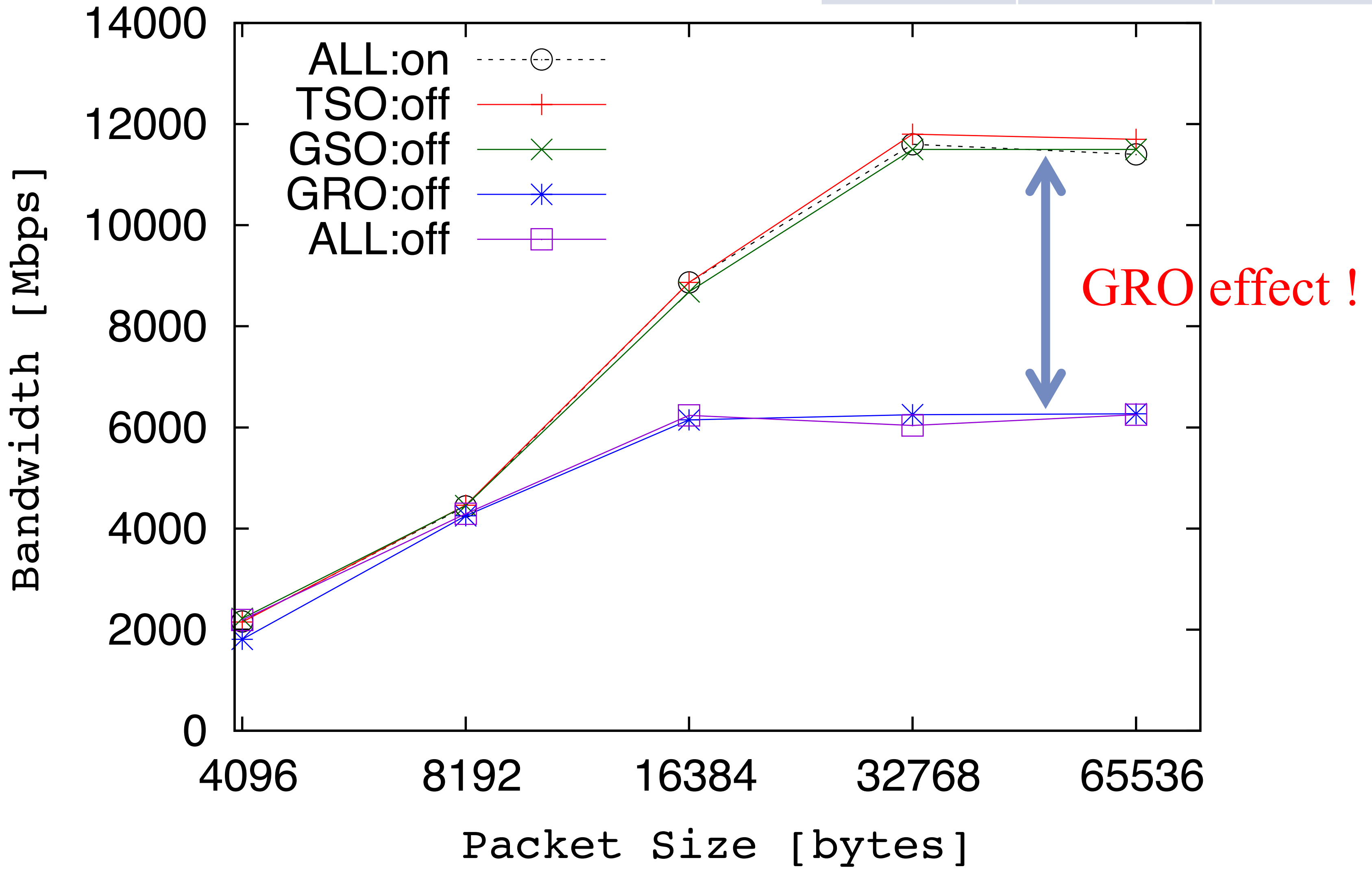
# Evaluation Results (UDP)

- **UDP communication**



# Offloading Effects (STT)

| Offload | Tx / Rx | NIC / Kernel |
|---|---|---|
| TSO | Tx | NIC |
| GSO | Tx | Kernel |
| GRO | Rx | Kernel |



GRO effect !