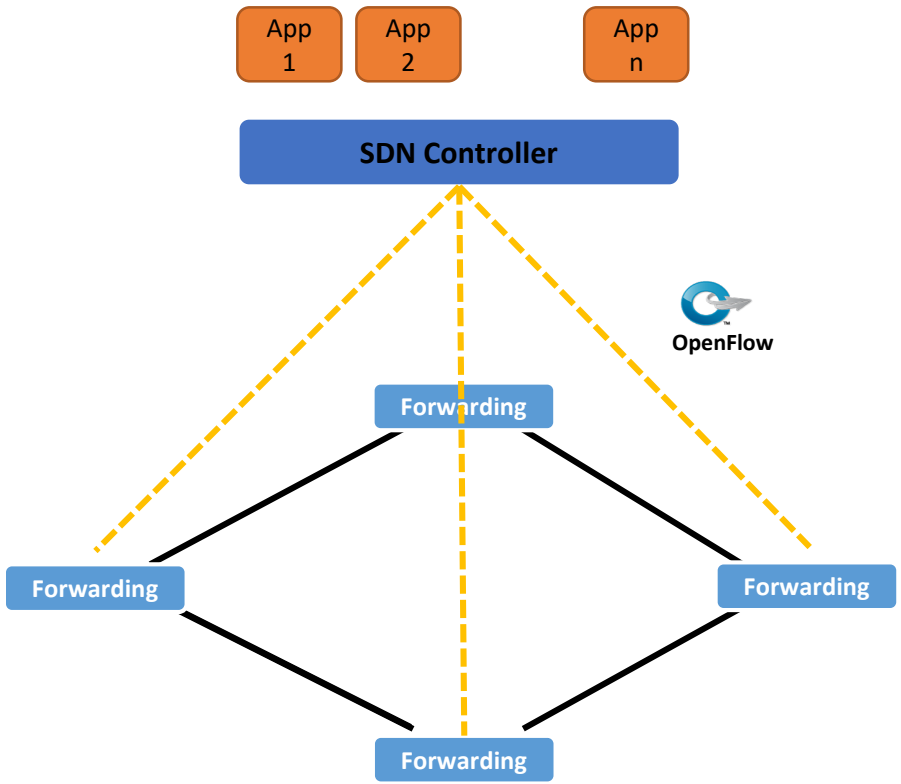
BNV : Enabling Scalable Experimentation through Bare-metal Network Virtualization

Pravein Govindan Kannan, Ahmad Soltani, Mun Choon Chan
Ee-Chien Chang

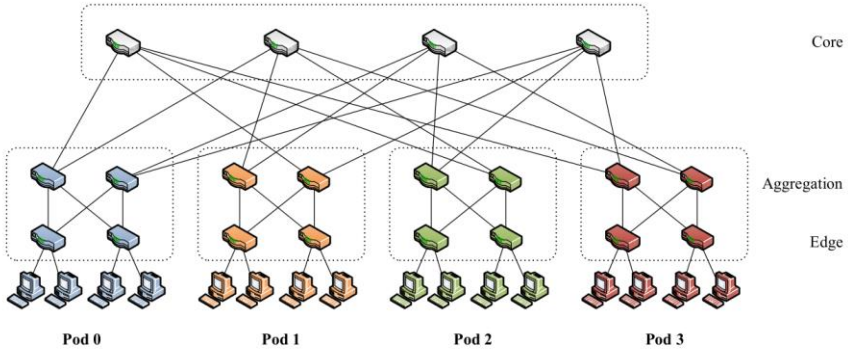
NUS
National University of Singapore



Optimizing Topologies

Congestion Control

Resource Allocation



No Single Topology is suitable for all use cases

How to validate ideas over multiple topologies?

Software Tools : Flexible

Mininet, Maxinet: Elegant tools enabling functional testing/ learning.

Container-based virtualization on a single machine or cluster of machines.

Cannot achieve the peak line-rate for performance testing.

Network Testbeds – High Fidelity

DeterLab, CloudLab, GENI, Staging Platforms:

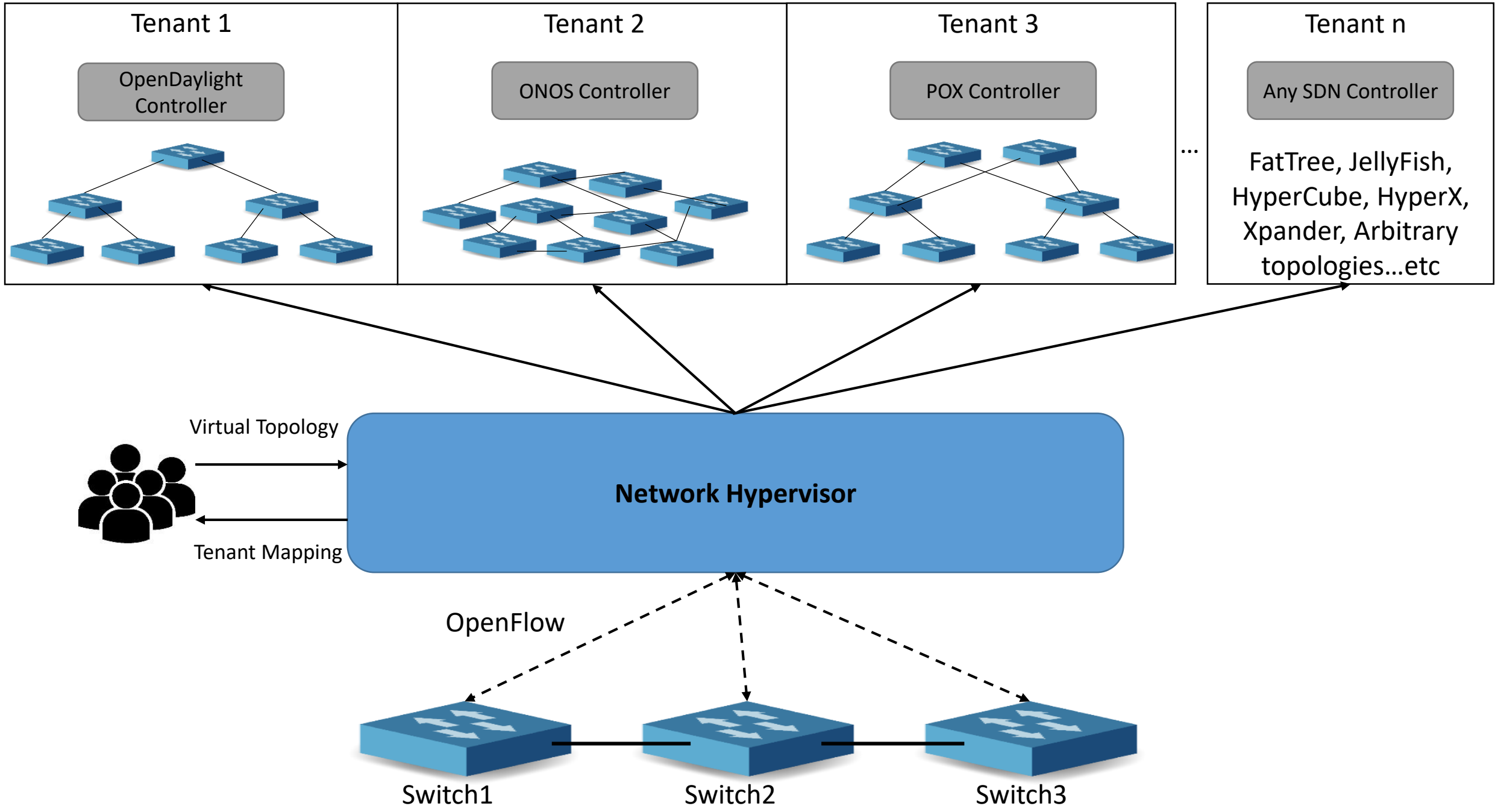
Can provide high fidelity emulation with Bare-Metal provisioning.

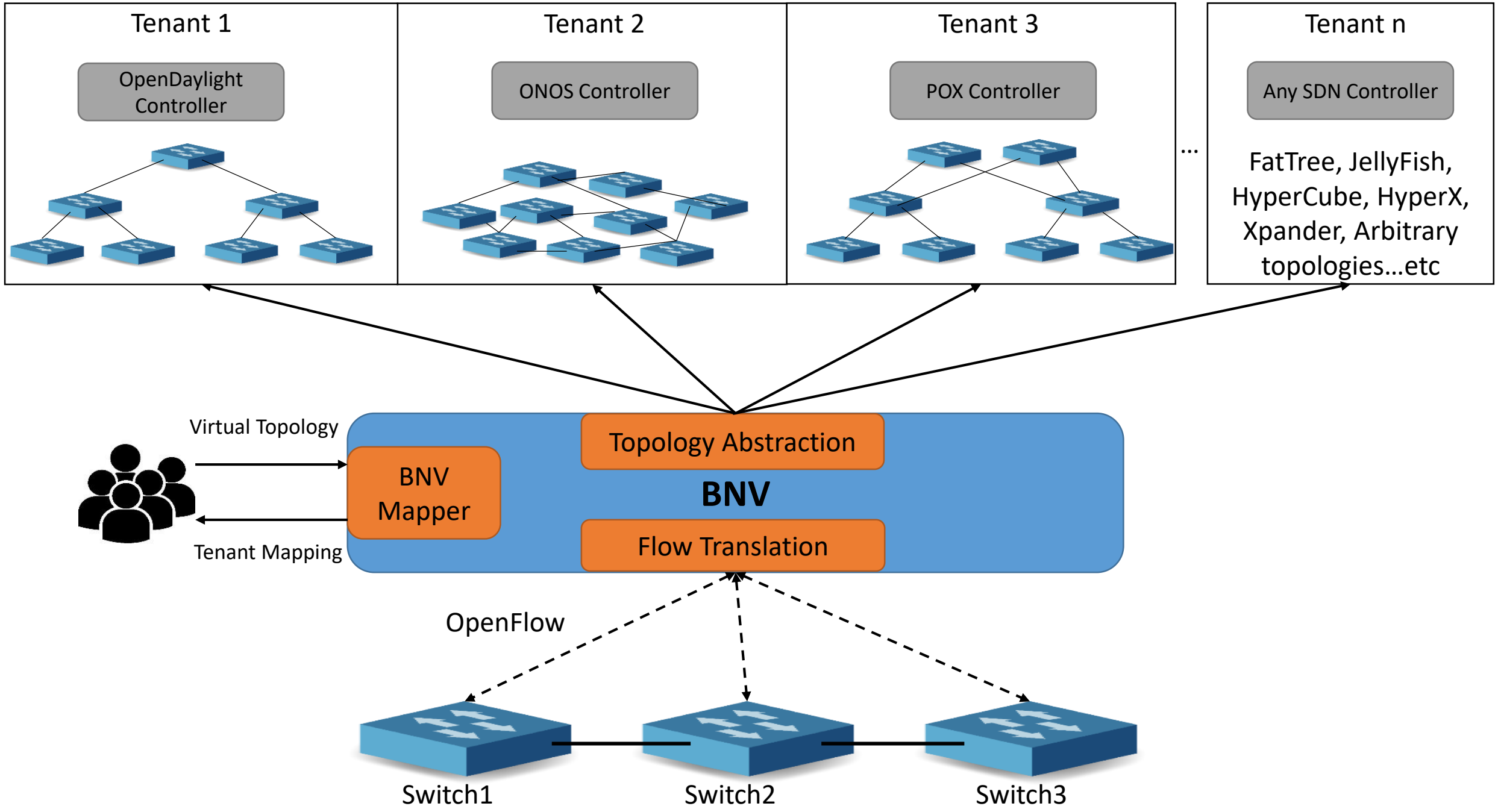
Can provide only a fixed topology or subset for Software Defined Networks.

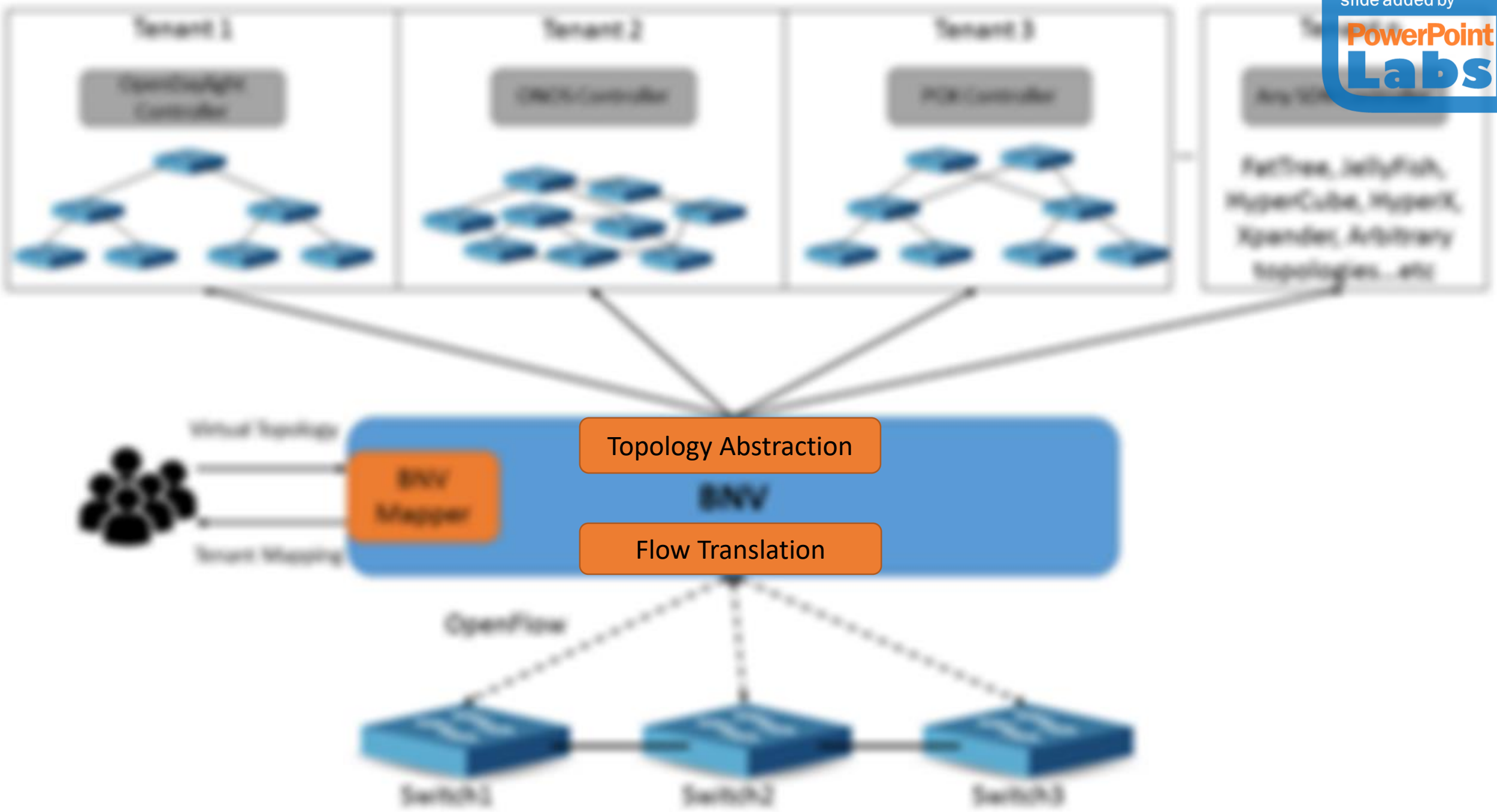
Change in Topology might require a long time (sometimes in order of weeks).

How do we achieve the flexibility of software and fidelity of hardware (bare-metal)?

Topology Scaling :
One-to-Many Switch Virtualization


Tenant 1
OpenDaylight Controller
Tenant 2
ONOS Controller
Tenant 3
POX Controller
...
Tenant n
Any SDN Controller
FatTree, JellyFish, HyperCube, HyperX, Xpander, Arbitrary topologies...etc
Virtual Topology
Tenant Mapping
Network Hypervisor
OpenFlow
Switch1
Switch2
Switch3

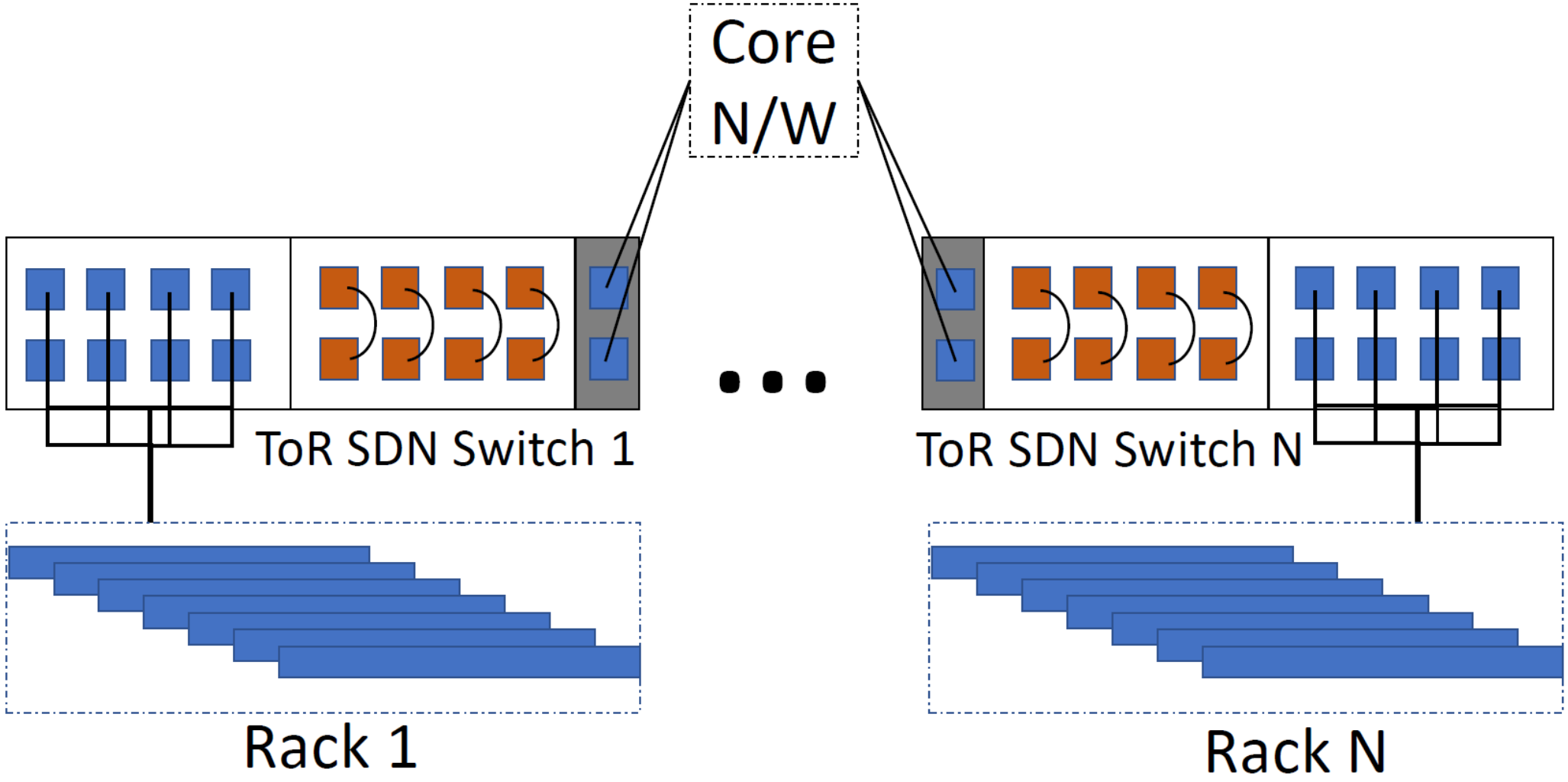

Tenant 1
OpenDaylight Controller
Tenant 2
ONOS Controller
Tenant 3
POX Controller
...
Tenant n
Any SDN Controller
FatTree, JellyFish, HyperCube, HyperX, Xpander, Arbitrary topologies...etc
Virtual Topology
Tenant Mapping
BNV Mapper
Topology Abstraction
BNV
Flow Translation
OpenFlow
Switch1
Switch2
Switch3



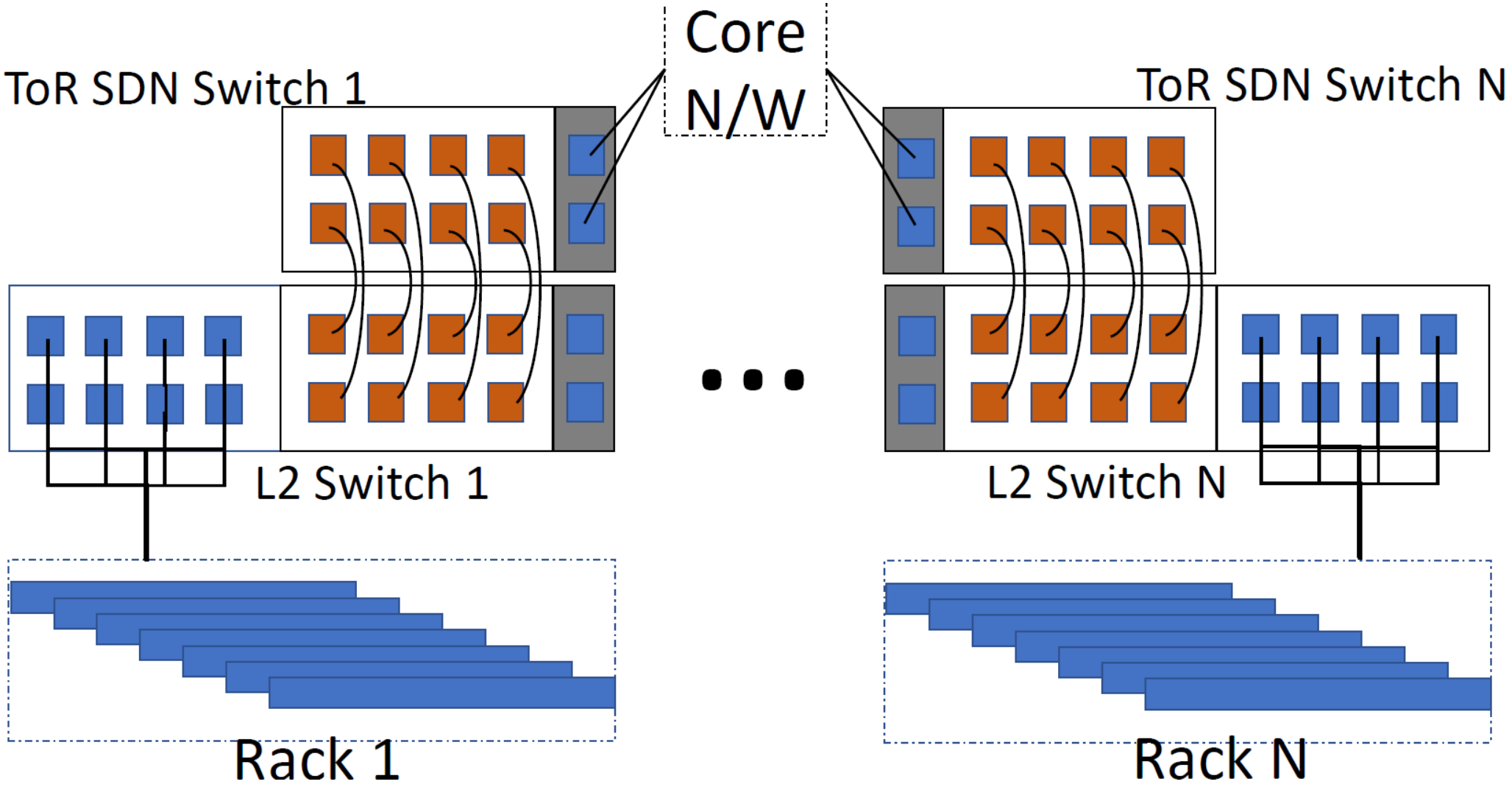
Idea : Loopback links in substrate topology

Idea : Loopback links in substrate topology

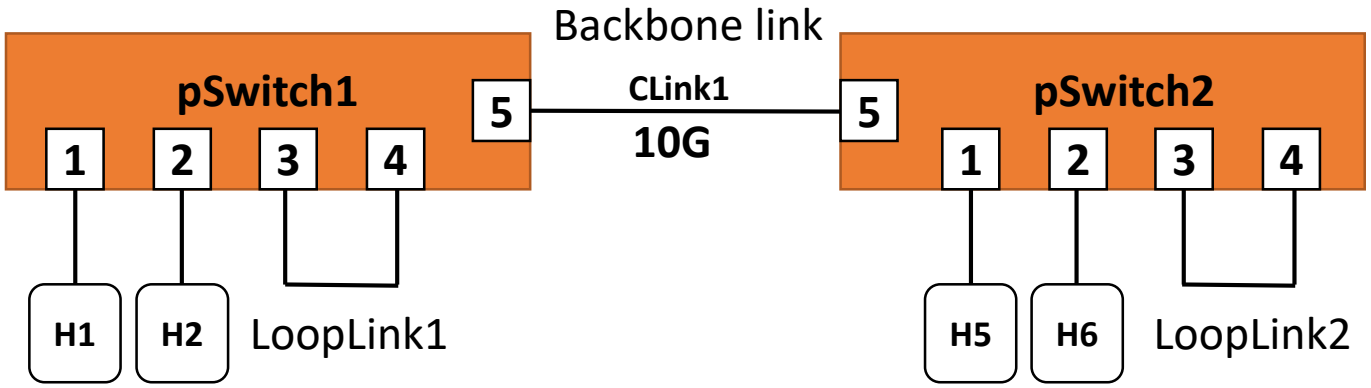
Idea : Loopback links in substrate topology

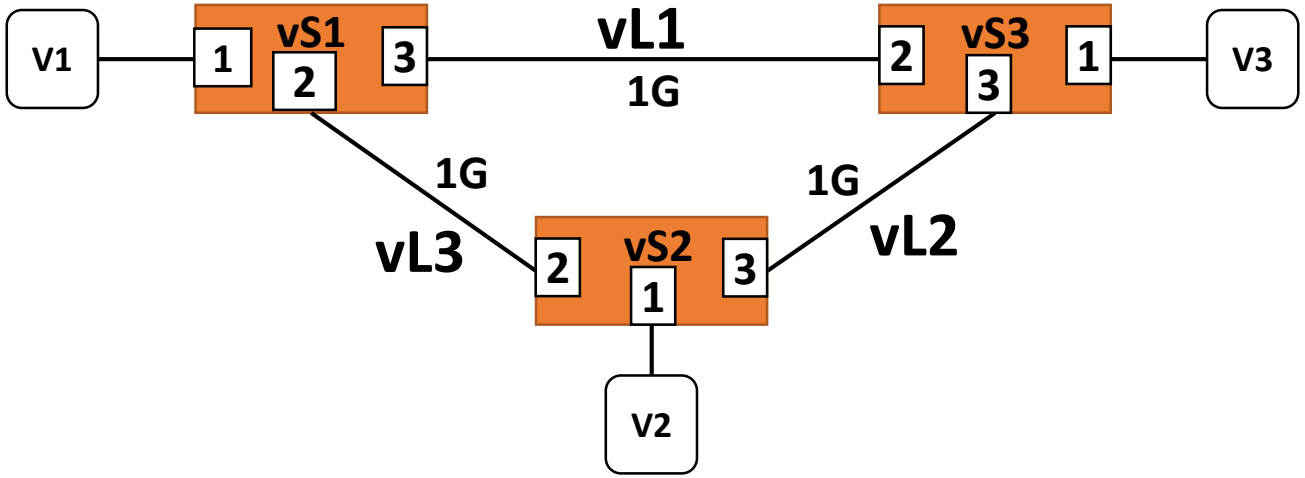


Software-Configurable Loopbacks



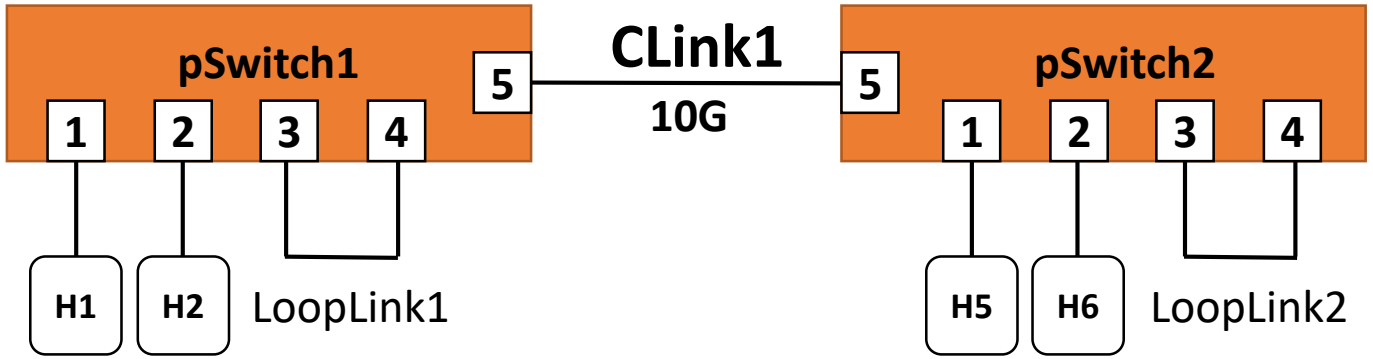
Topology + Loopback link

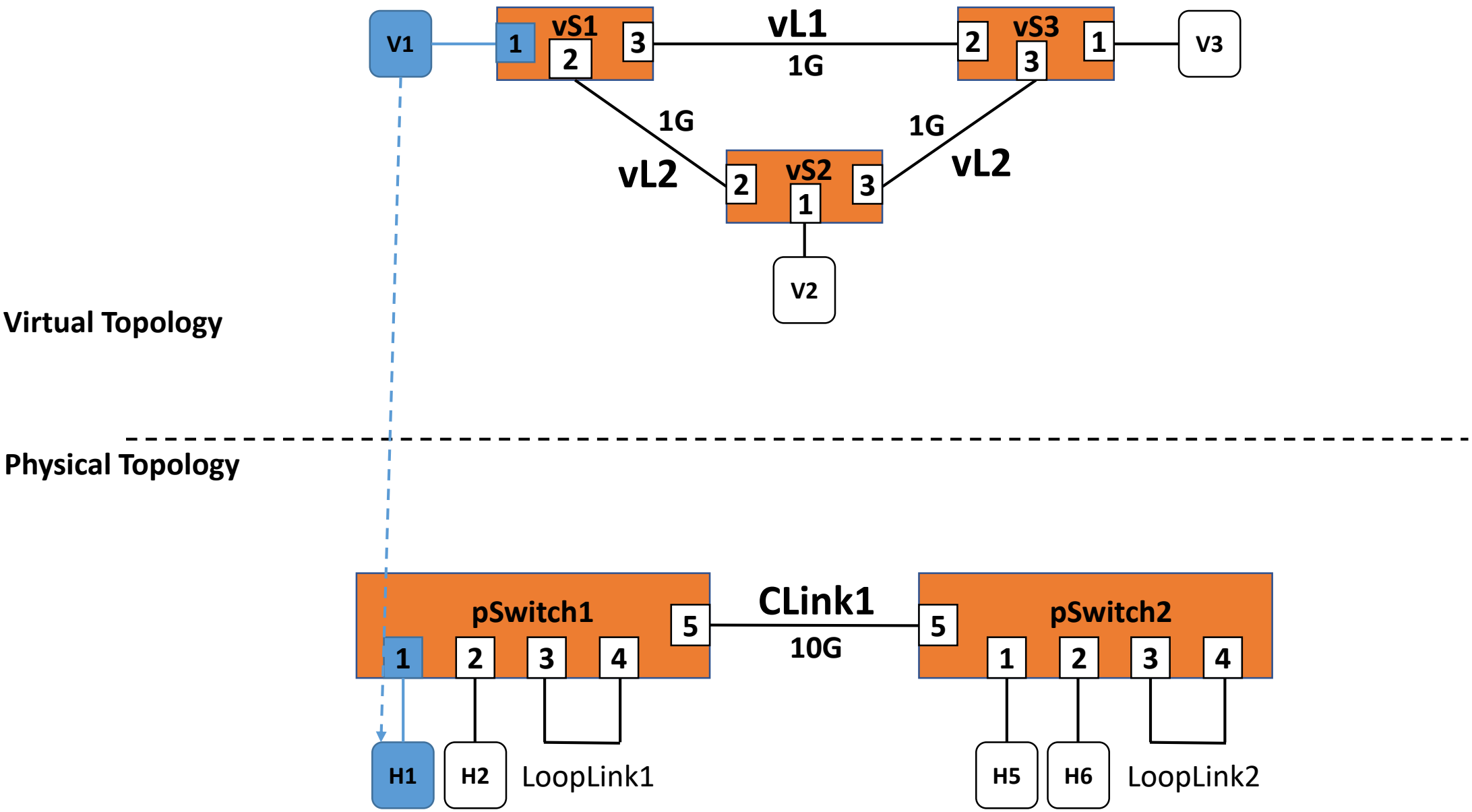


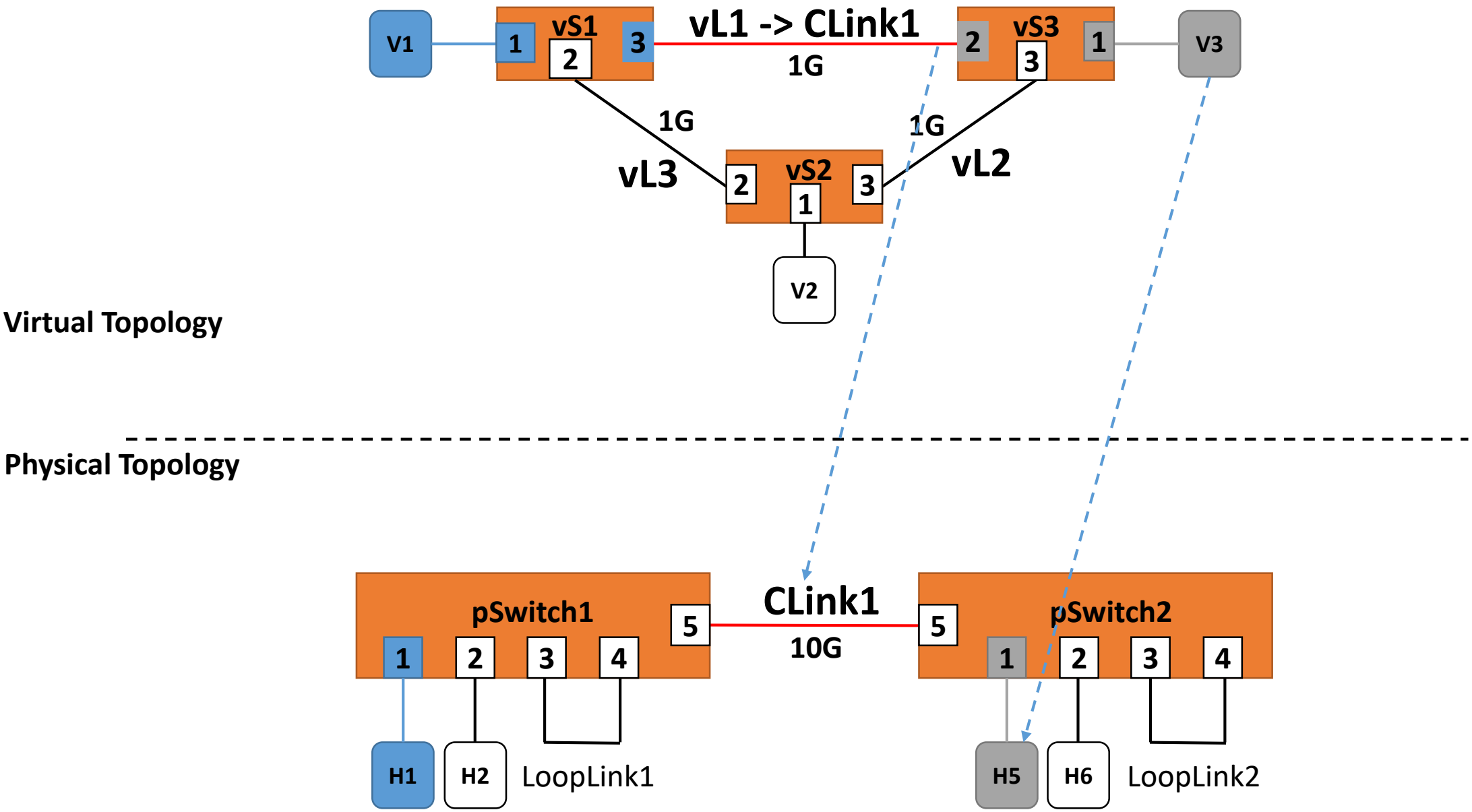


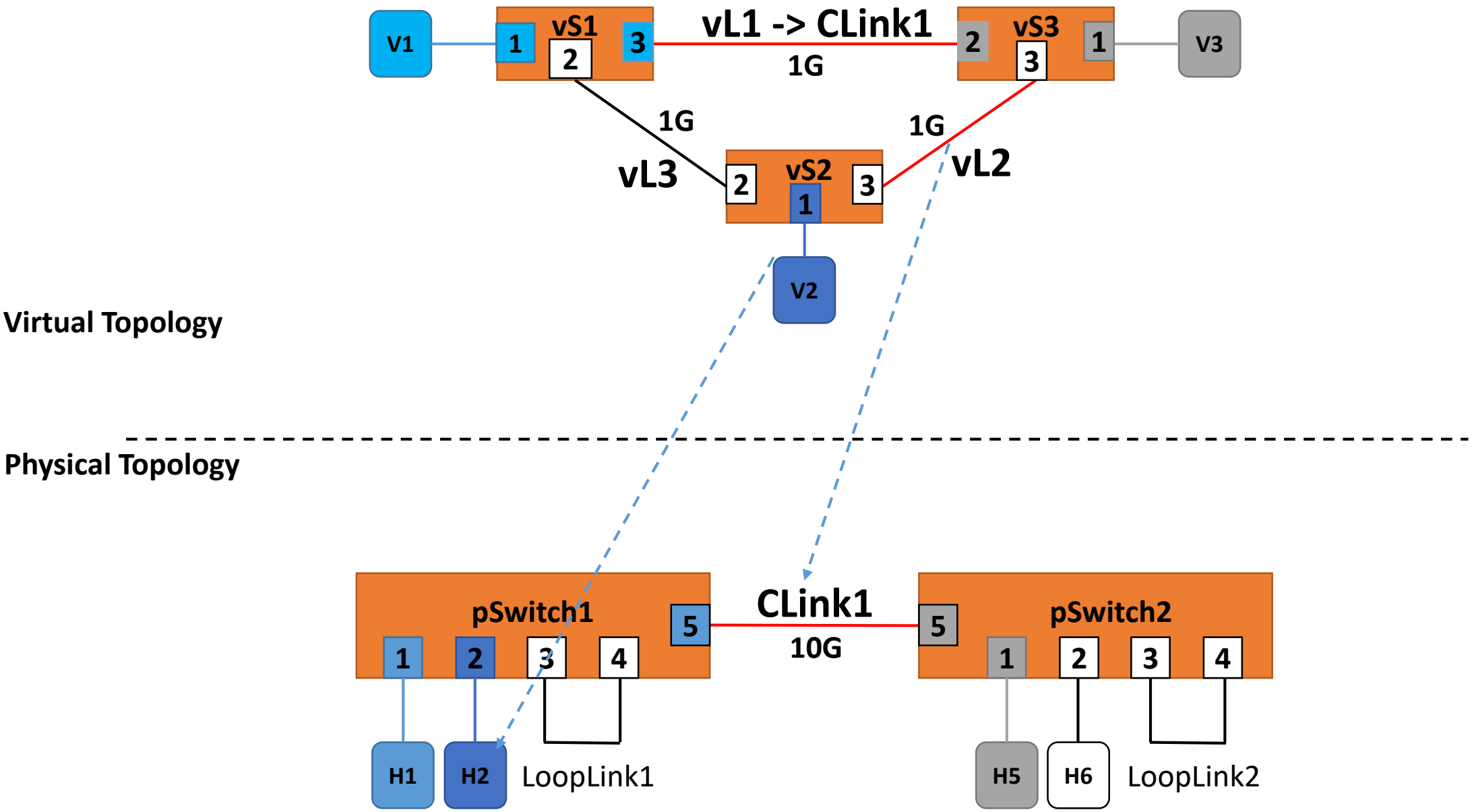
Virtual Topology

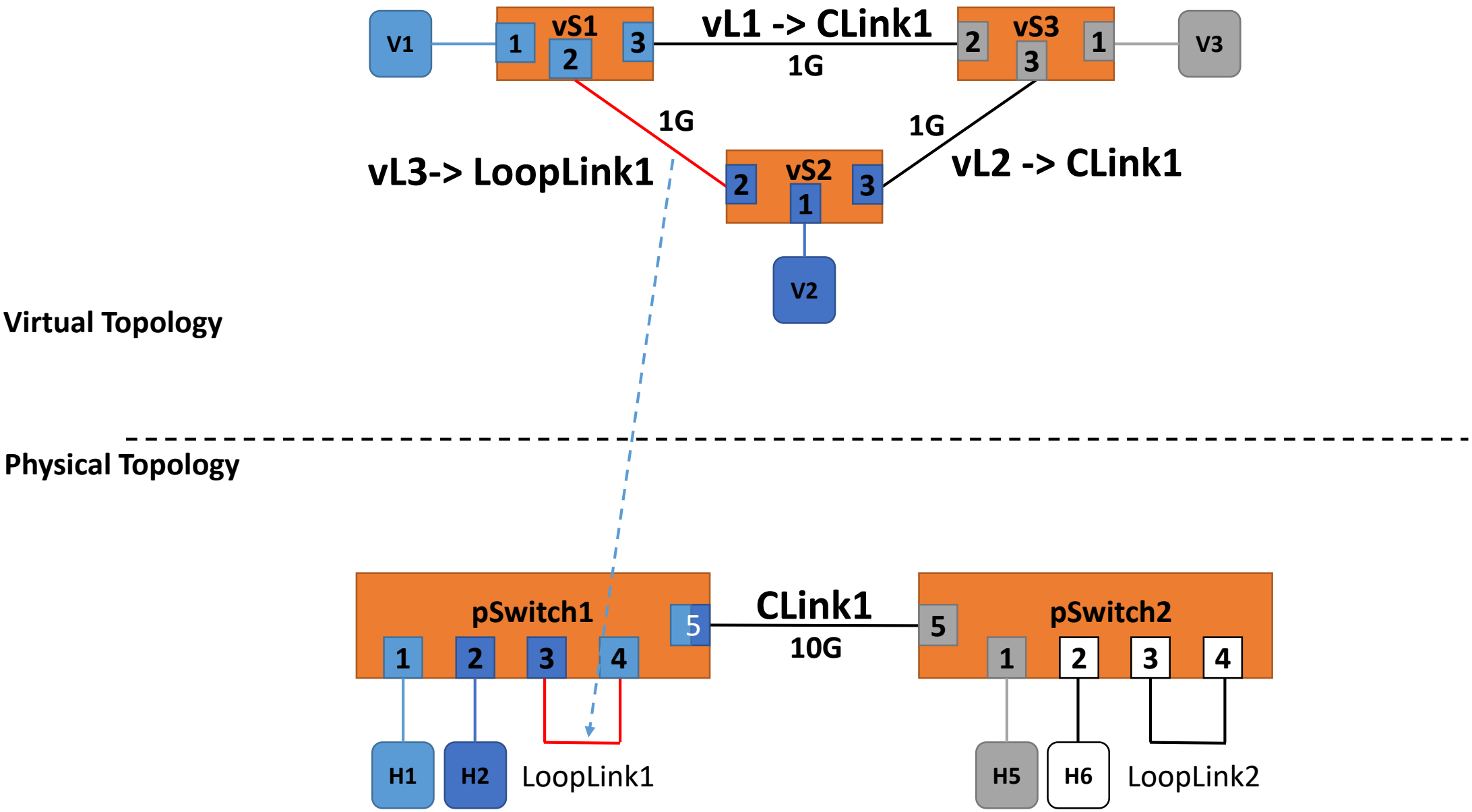
Physical Topology

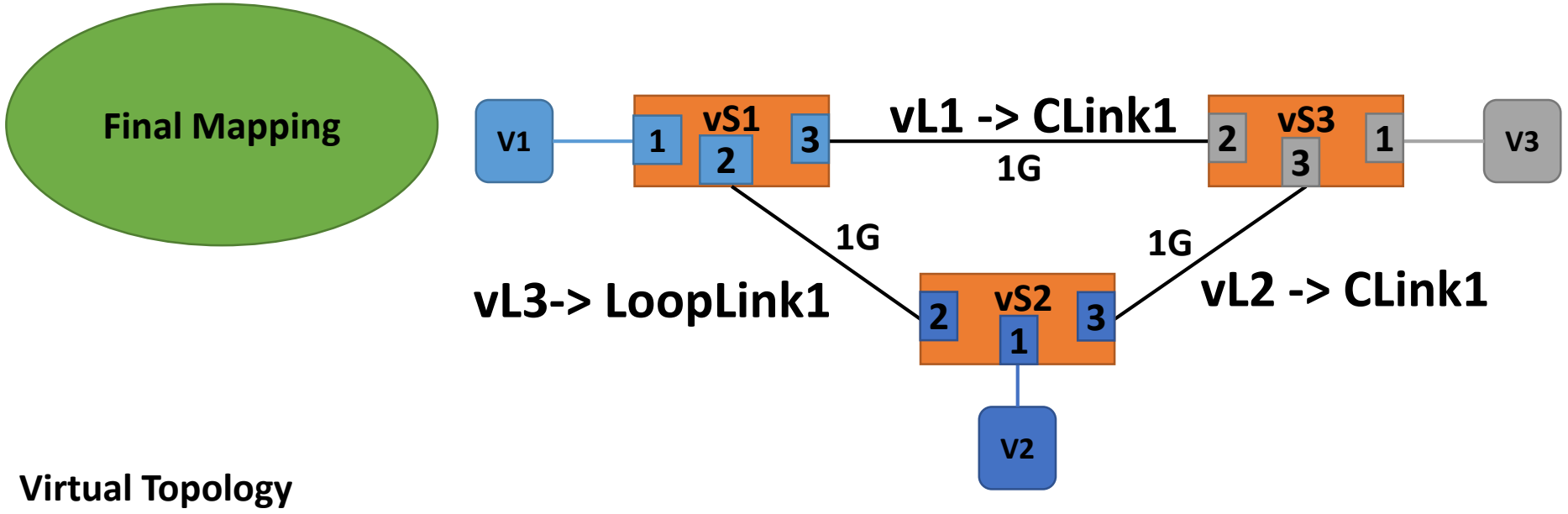



Virtual Topology
V1
vS1
1
2
3
vL1
1G
vS3
2
3
1
V3
1G
vL2
1G
vL2
vS2
2
1
3
V2
Physical Topology
pSwitch1
1
2
3
4
5
CLink1
10G
pSwitch2
5
1
2
3
4
H1
H2
LoopLink1
H5
H6
LoopLink2

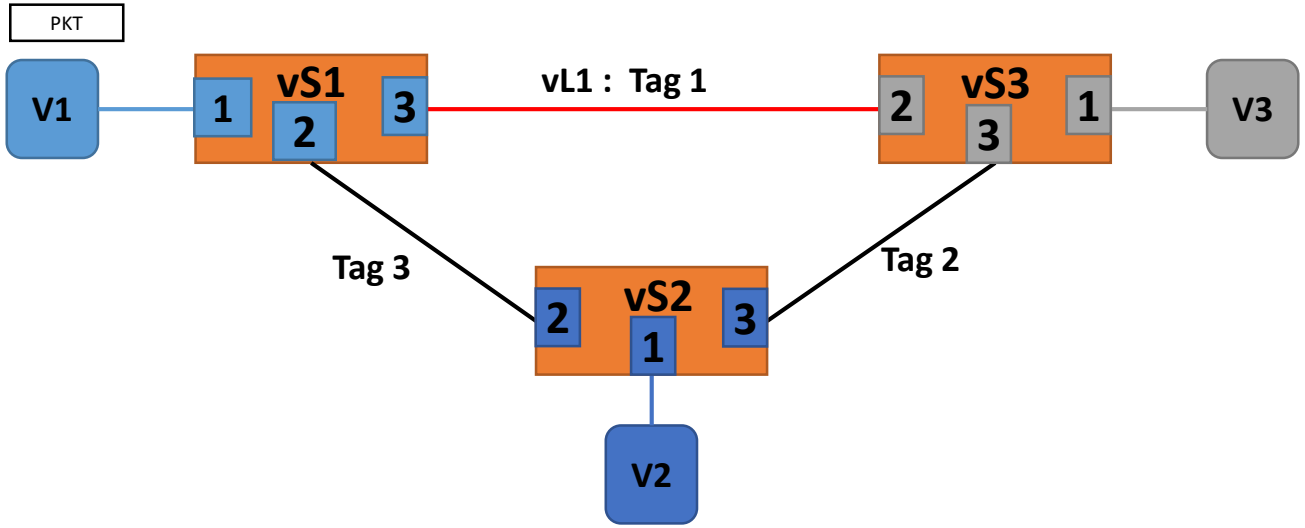

V1
1
vS1
2
3
vL1 -> CLink1
1G
2
vS3
3
1
V3
1G
1G
vL3
2
vS2
1
3
vL2
V2
Virtual Topology
Physical Topology
pSwitch1
1
2
3
4
5
CLink1
10G
5
pSwitch2
1
2
3
4
H1
H2
LoopLink1
H5
H6
LoopLink2


V1
vS1
1
2
3
vL1 -> CLink1
1G
2
vS3
3
1
V3
1G
vL3
1G
vL2
2
vS2
1
3
V2
Virtual Topology
Physical Topology
pSwitch1
1
2
3
4
5
CLink1
10G
5
pSwitch2
1
2
3
4
H1
H2
LoopLink1
H5
H6
LoopLink2


V1
vS1
1
2
3
vL1 -> CLink1
1G
2
vS3
3
1
V3
1G
1G
vL3-> LoopLink1
2
vS2
1
3
vL2 -> CLink1
V2
Virtual Topology
Physical Topology
pSwitch1
1
2
3
4
5
CLink1
10G
5
pSwitch2
1
2
3
4
H1
H2
LoopLink1
H5
H6
LoopLink2

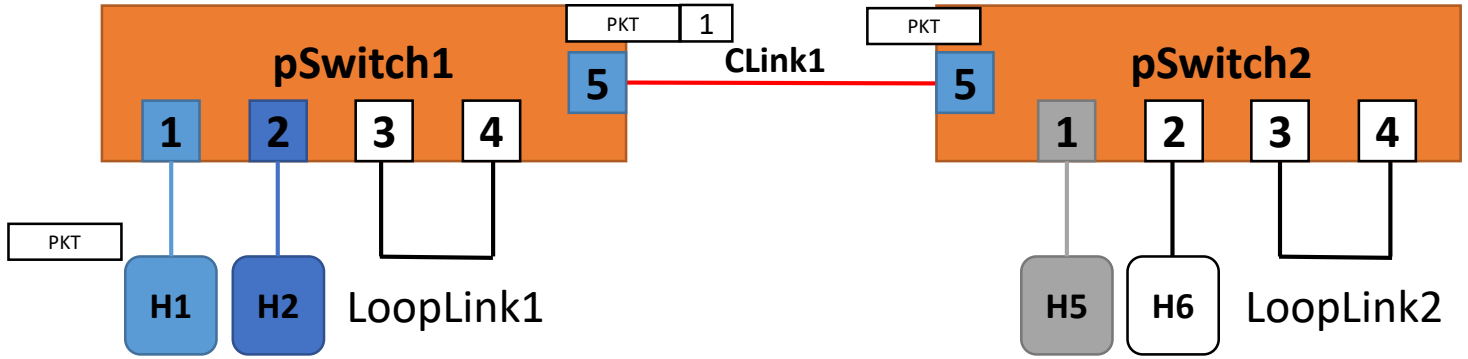

Final Mapping
V1
vS1
1
2
3
vL1 -> CLink1
1G
2
vS3
3
1
V3
1G
1G
vL3-> LoopLink1
2
vS2
1
3
vL2 -> CLink1
V2
Virtual Topology
Physical Topology
pSwitch1
1
2
3
4
5
CLink1
10G
5
pSwitch2
1
2
3
4
H1
H2
LoopLink1
H5
H6
LoopLink2

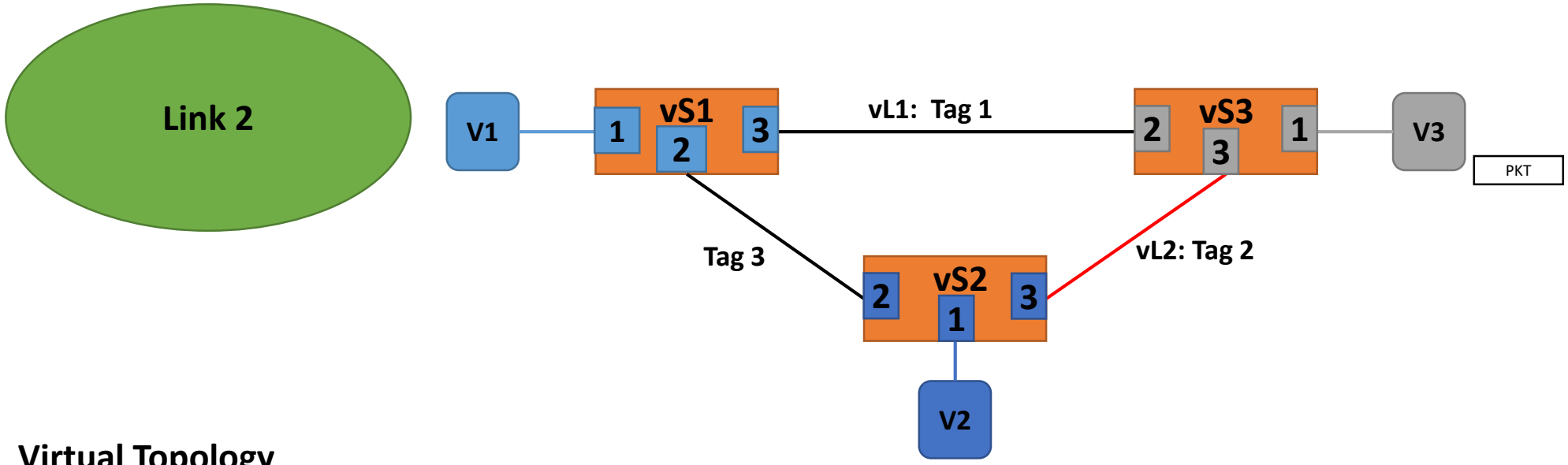
Link 1 using Packet Tagging



Virtual Topology

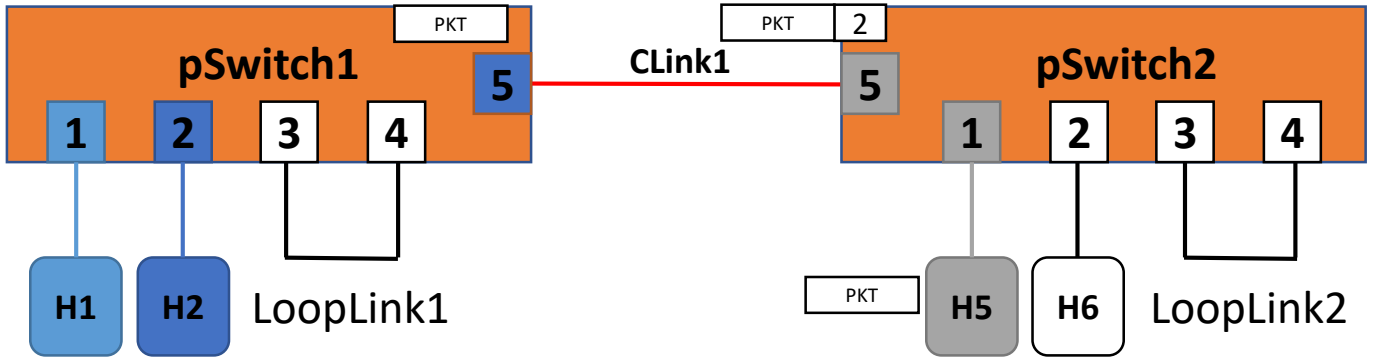
Physical Topology

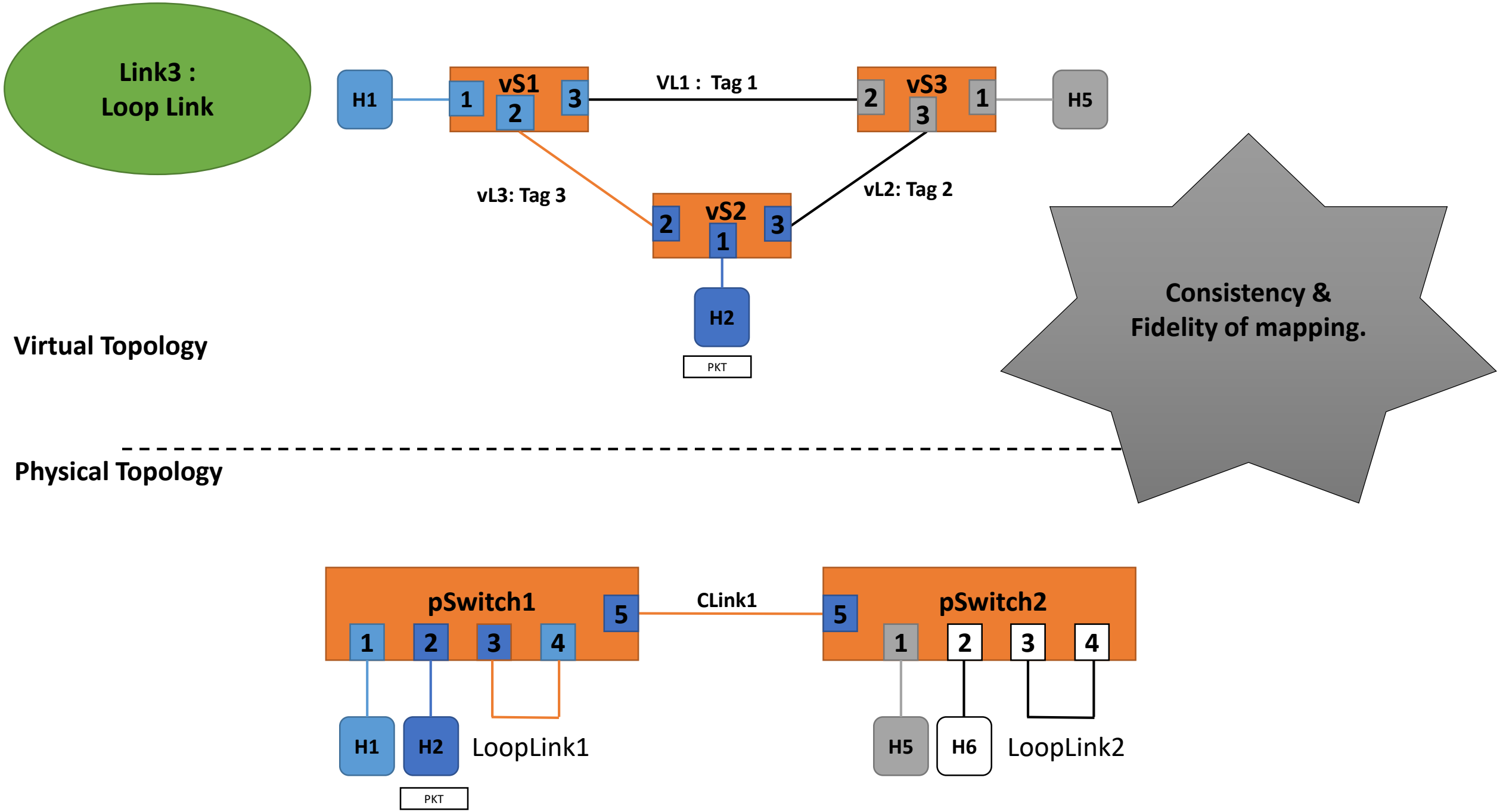




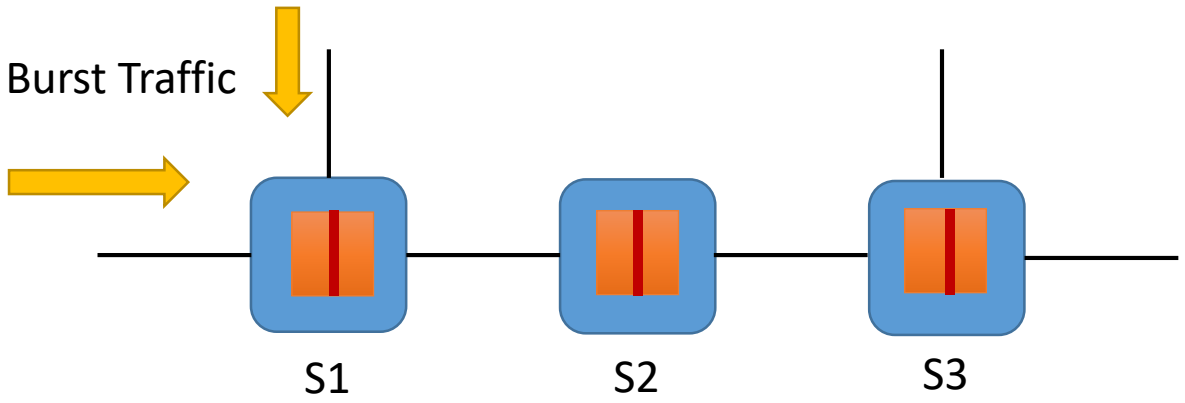
Virtual Topology

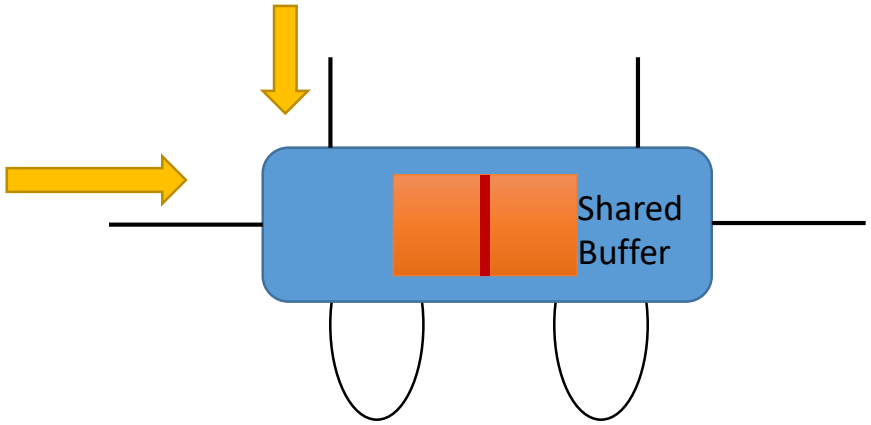
Physical Topology




Link3 :
Loop Link
H1
1
vS1
2
3
VL1 : Tag 1
2
vS3
3
1
H5
vL3: Tag 3
vL2: Tag 2
2
vS2
1
3
H2
PKT
Virtual Topology
Consistency &
Fidelity of mapping.
Physical Topology
pSwitch1
1
2
3
4
5
CLink1
5
pSwitch2
1
2
3
4
H1
H2
LoopLink1
PKT
H5
H6
LoopLink2

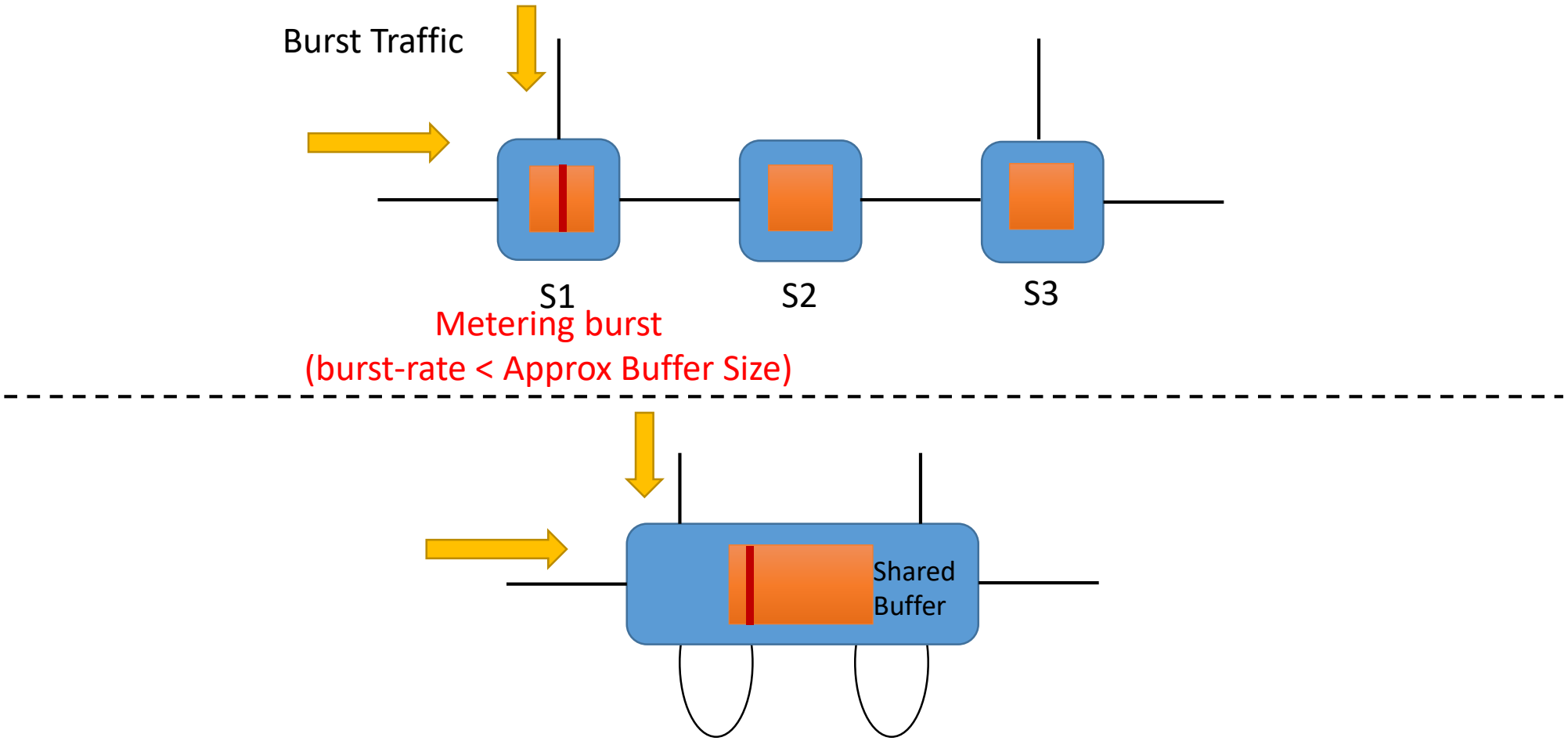
How to achieve buffer isolation between multiple virtual switches ?



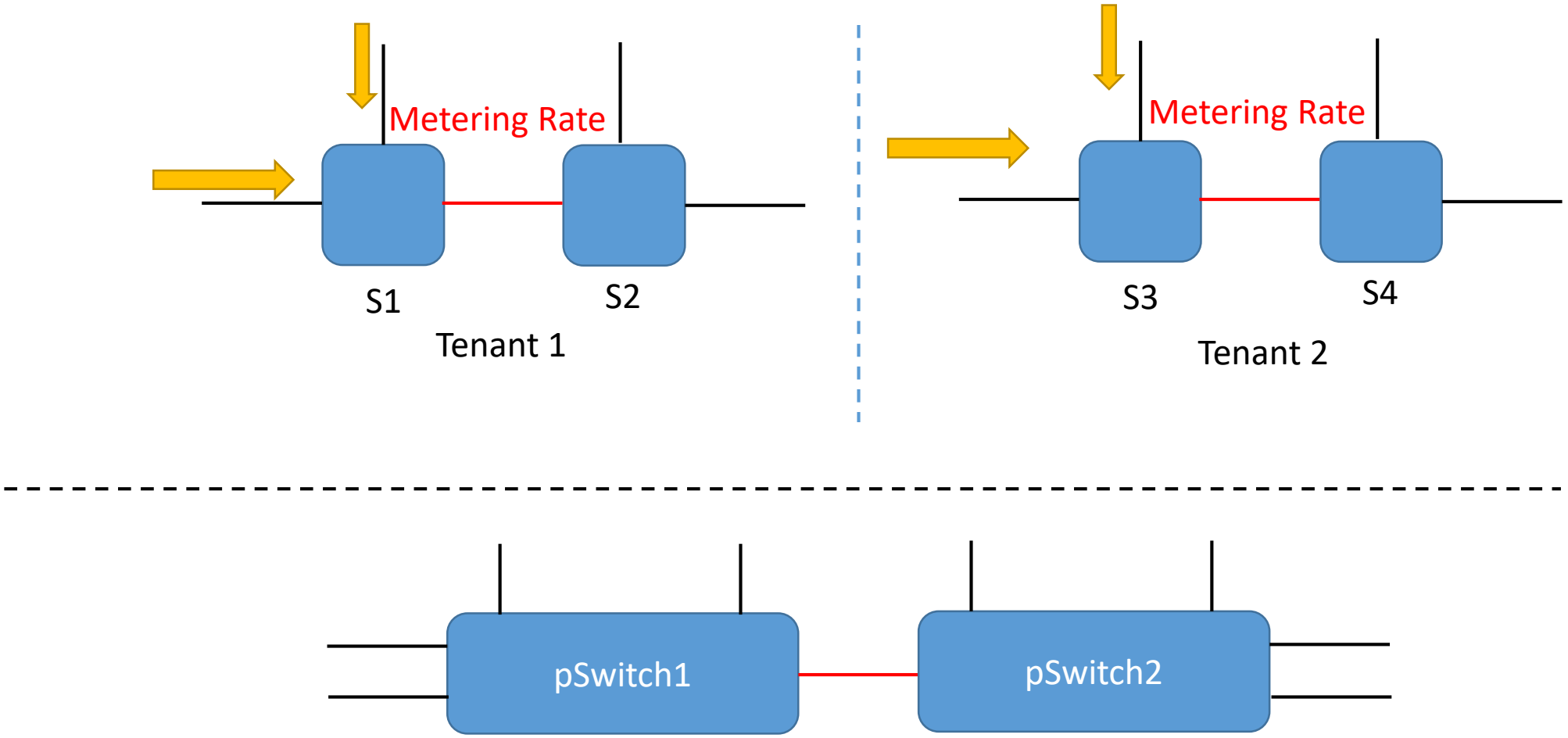


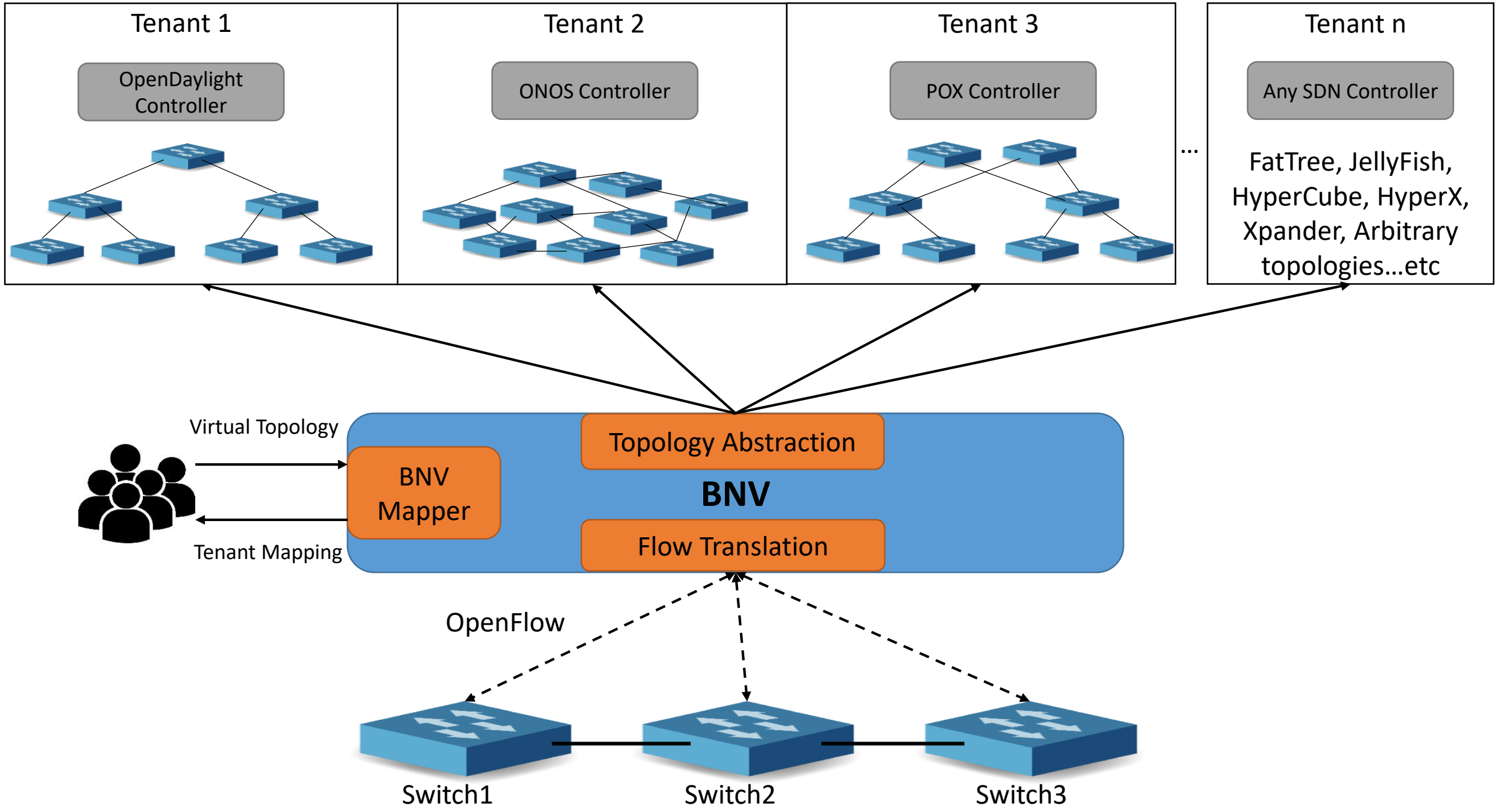
A single virtual switch can fill up the entire shared Buffer !!

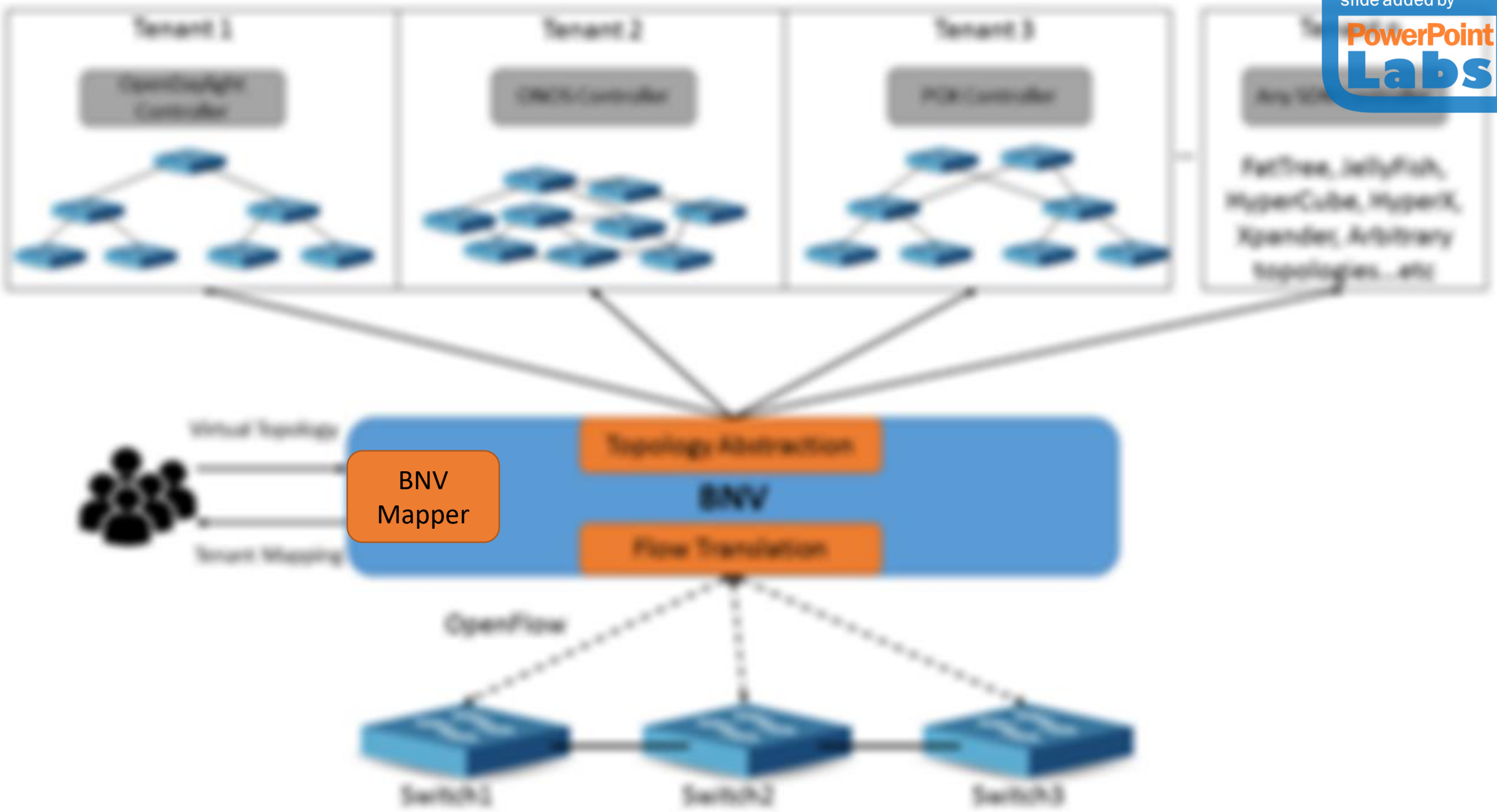
How to achieve buffer isolation between multiple virtual switches ?



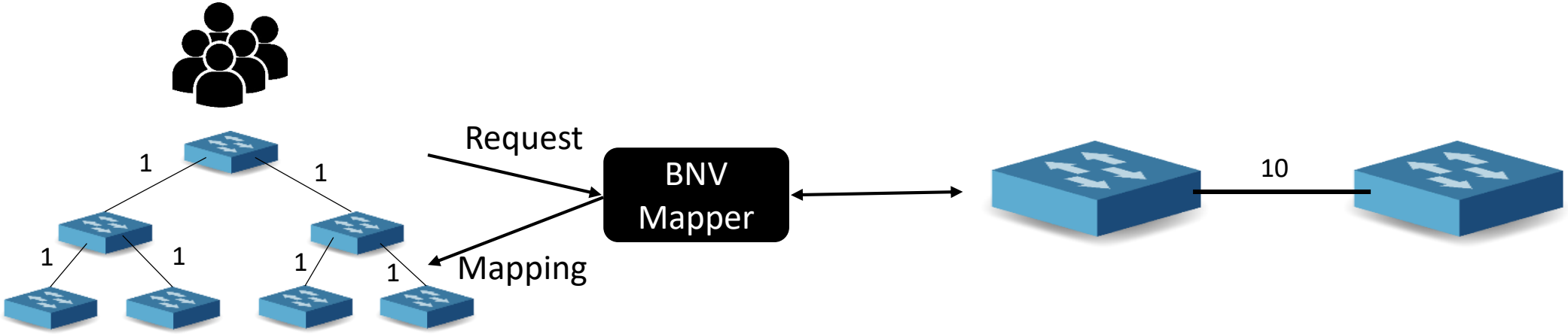
How to achieve link isolation between multiple virtual links?




Tenant 1
OpenDaylight Controller
Tenant 2
ONOS Controller
Tenant 3
POX Controller
...
Tenant n
Any SDN Controller
FatTree, JellyFish, HyperCube, HyperX, Xpander, Arbitrary topologies...etc
Virtual Topology
Tenant Mapping
BNV Mapper
Topology Abstraction
BNV
Flow Translation
OpenFlow
Switch1
Switch2
Switch3

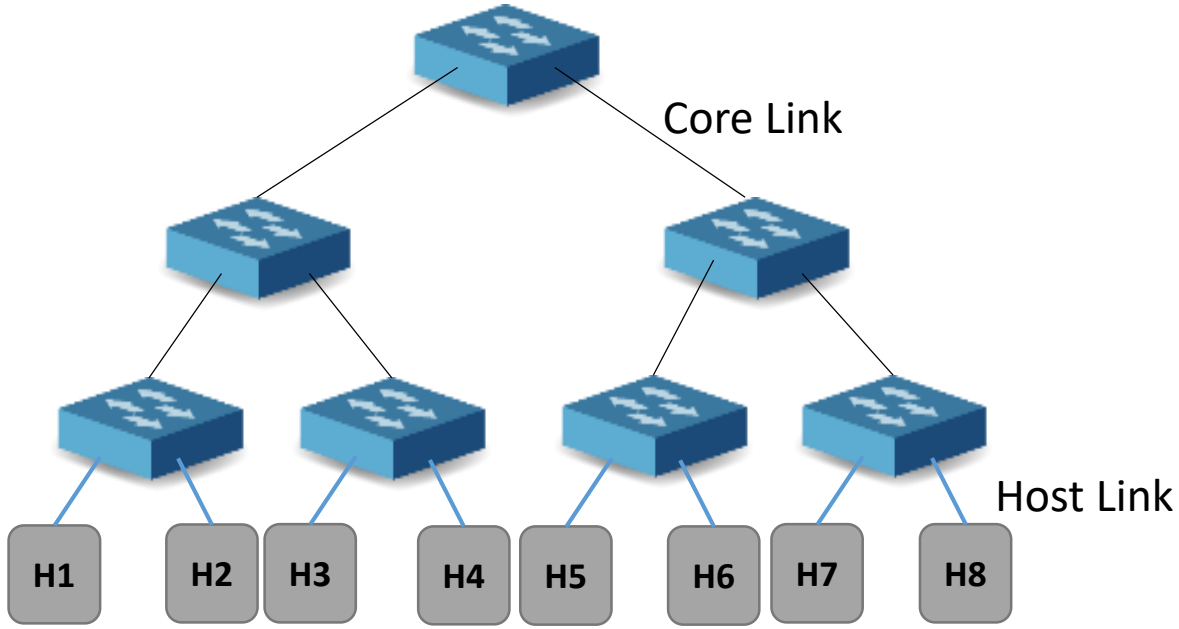


BNV Mapper



BNV Mapper

- A Virtual Topology Consists of :
  - A set of hosts V,
  - A set of (virtual) switches S, (feature: TCAM capacity, meter capacity, etc)
  - A set of (virtual) links L (feature : Bandwidth)



BNV Mapper

- Solved using ILP
- Variables :

Core Link: $$X_{i,v} = \begin{cases} \mathbf{1}\ ; \boldsymbol{if\ i\ is\ mapped\ to\ v} \\ \mathbf{0}\ ; \quad \boldsymbol{otherwise} \end{cases}$$

Host link: $$Y_{j,w} = \begin{cases} \mathbf{1}\ ; \boldsymbol{if\ j\ is\ mapped\ to\ w} \\ \mathbf{0}\ ; \quad \boldsymbol{otherwise} \end{cases}$$

BNV Mapper

Objective is to Minimize the usage of Core-links:

a) Maximizes isolation to the tenant, keeping links dedicated.

b) Provides scope for dynamic expansions. (Adding/upgrading tenants)

$$\mathbf{min}: \underbrace{\sum_{\substack{i \in Lc \\ v \in Y}} x_{iv} b_i}_{\text{Inter-Switch Links}} + \underbrace{\sum_{\substack{m \in S \\ \{p,q\} \in R}} M^{m}_{pq}}_{\text{Big-Switch abstraction}}$$

More details in paper!

Implementation

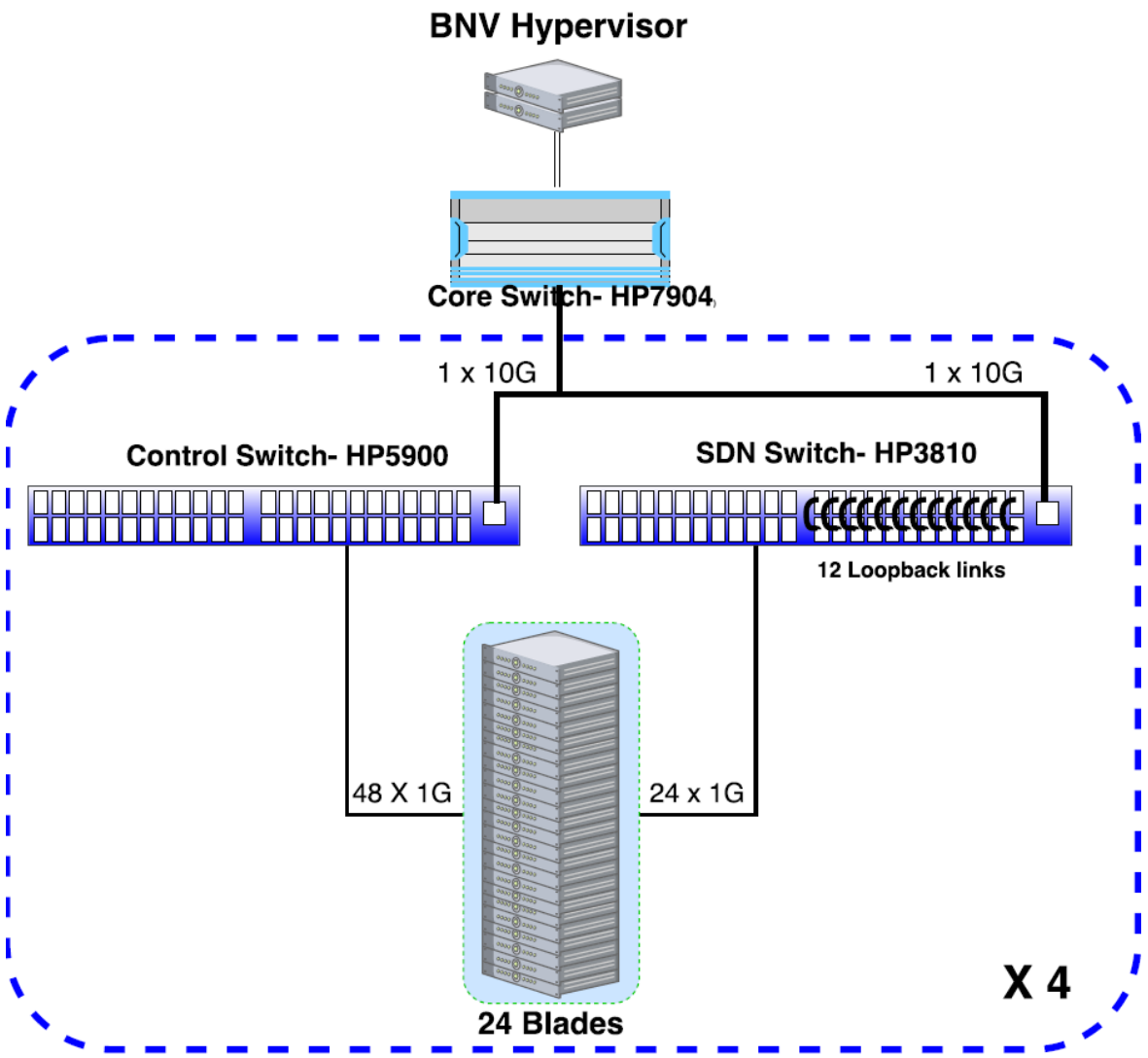
Implemented on top of OpenVirteX

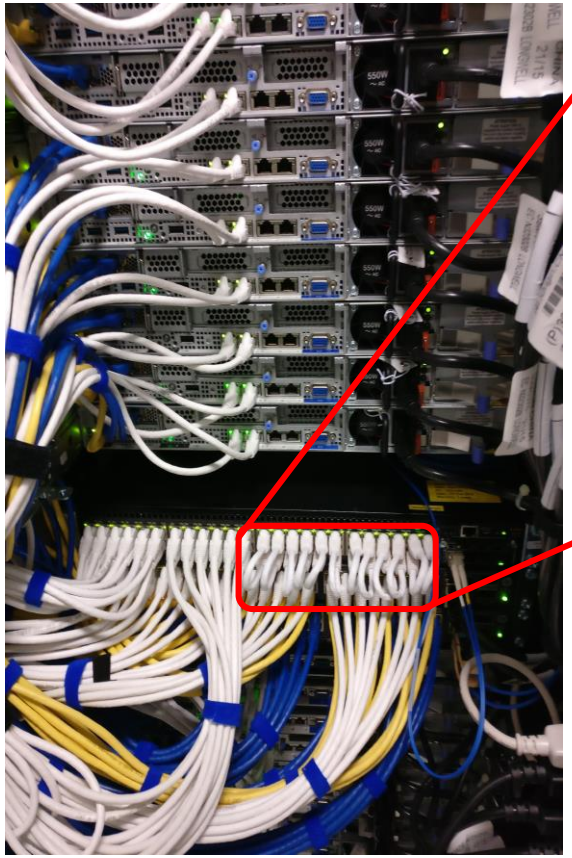
Integrated with a Bare-metal provisioning service (Testbed). DeterLab

Deployed on National Cybersecurity Lab(NCL) Testbed.

Support OpenFlow 1.0/ 1.3

Implementation





Static loopback wiring
24-ports of SDN Switch

Dynamic Loopback Wiring
using L2 intermediate Switch

Evaluation

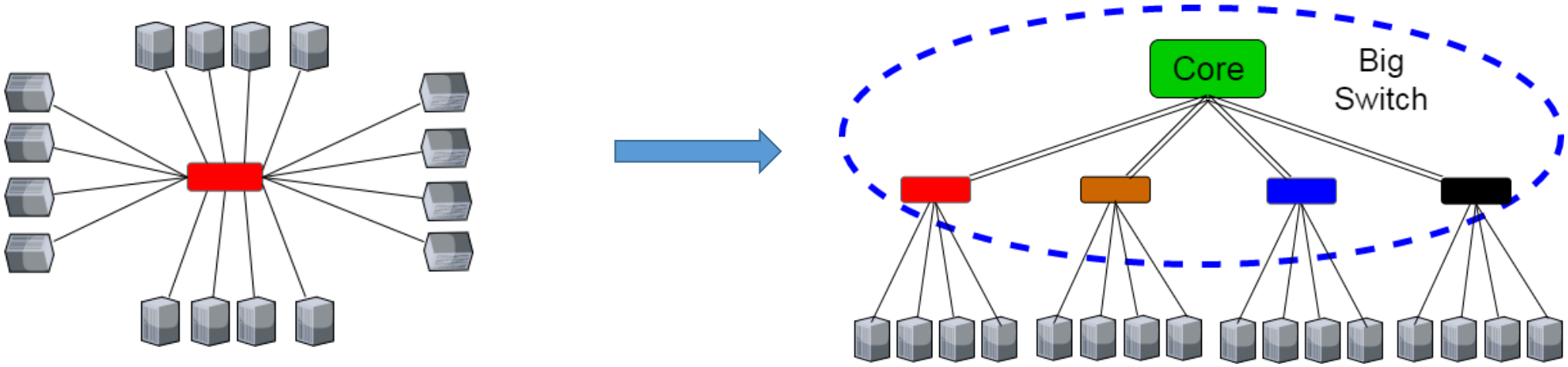
Can BNV :

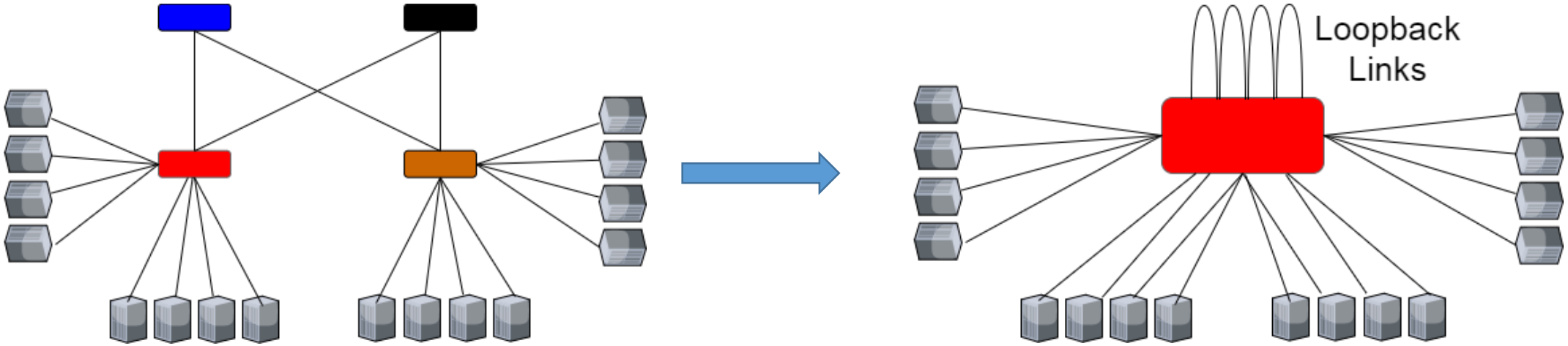
1) faithfully replicate behaviour of topologies?

2) Provide control/data plane isolation to multiple tenants?

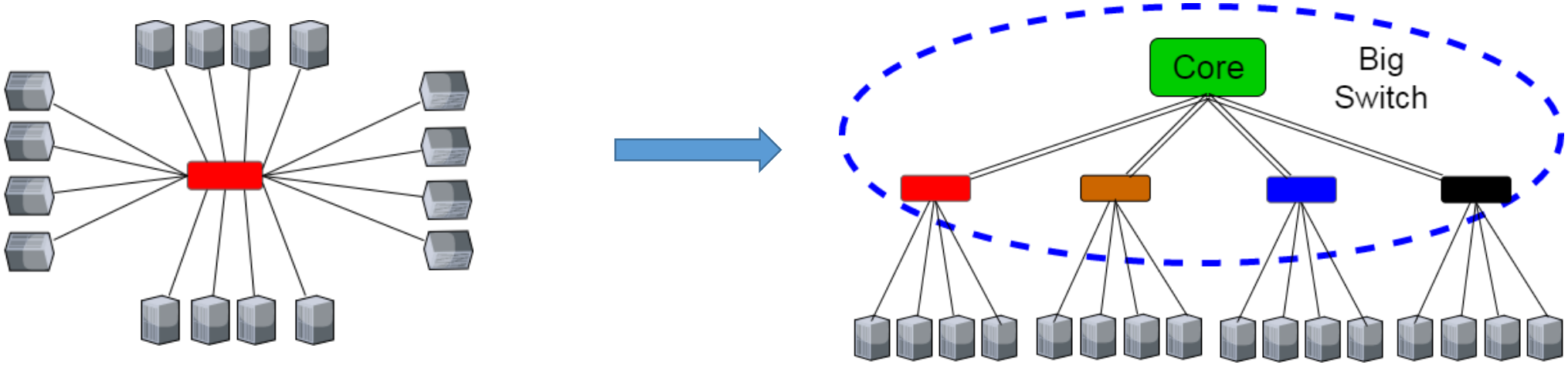
3) Scale of complex topologies embedded?

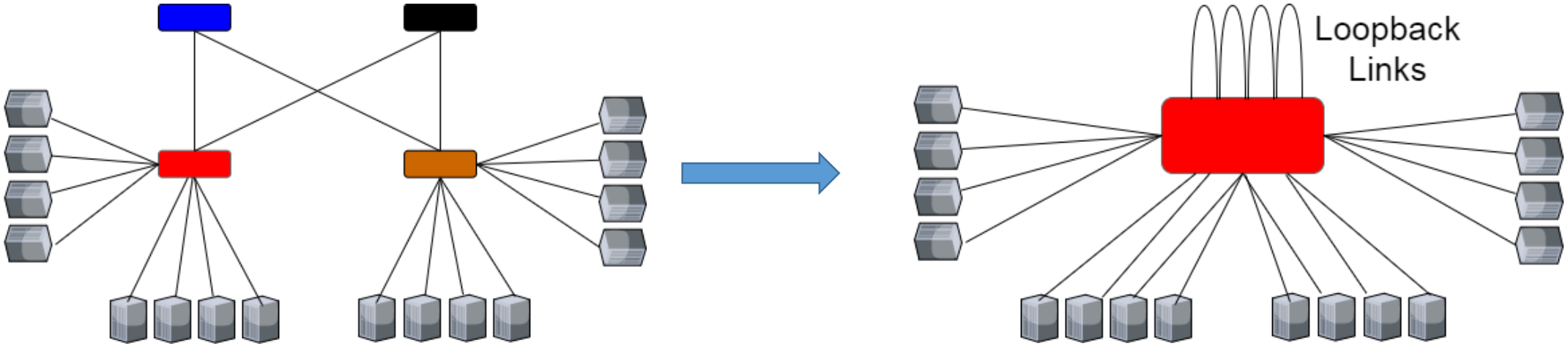
Evaluation : loopbacks & Big Switch Abstraction



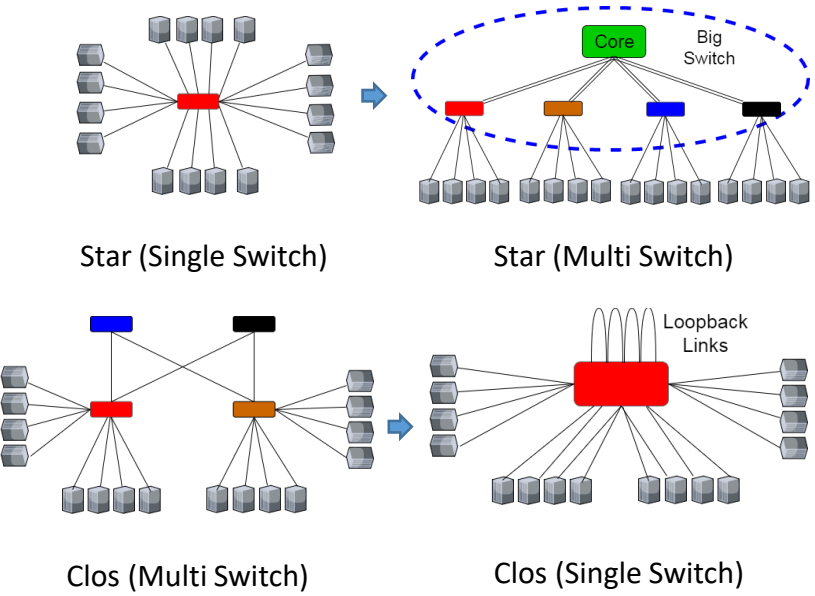


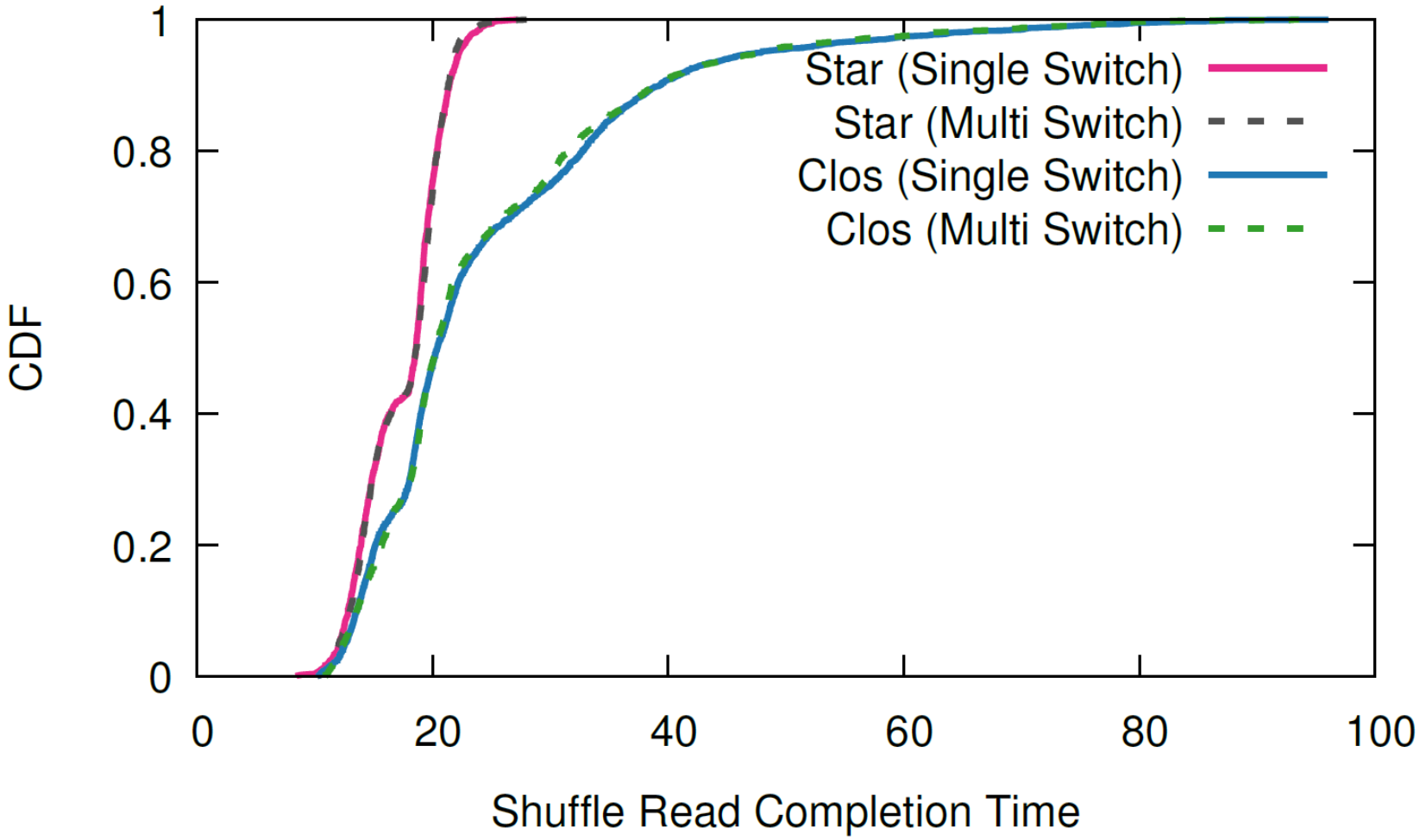
Evaluation : loopbacks & Big Switch Abstraction



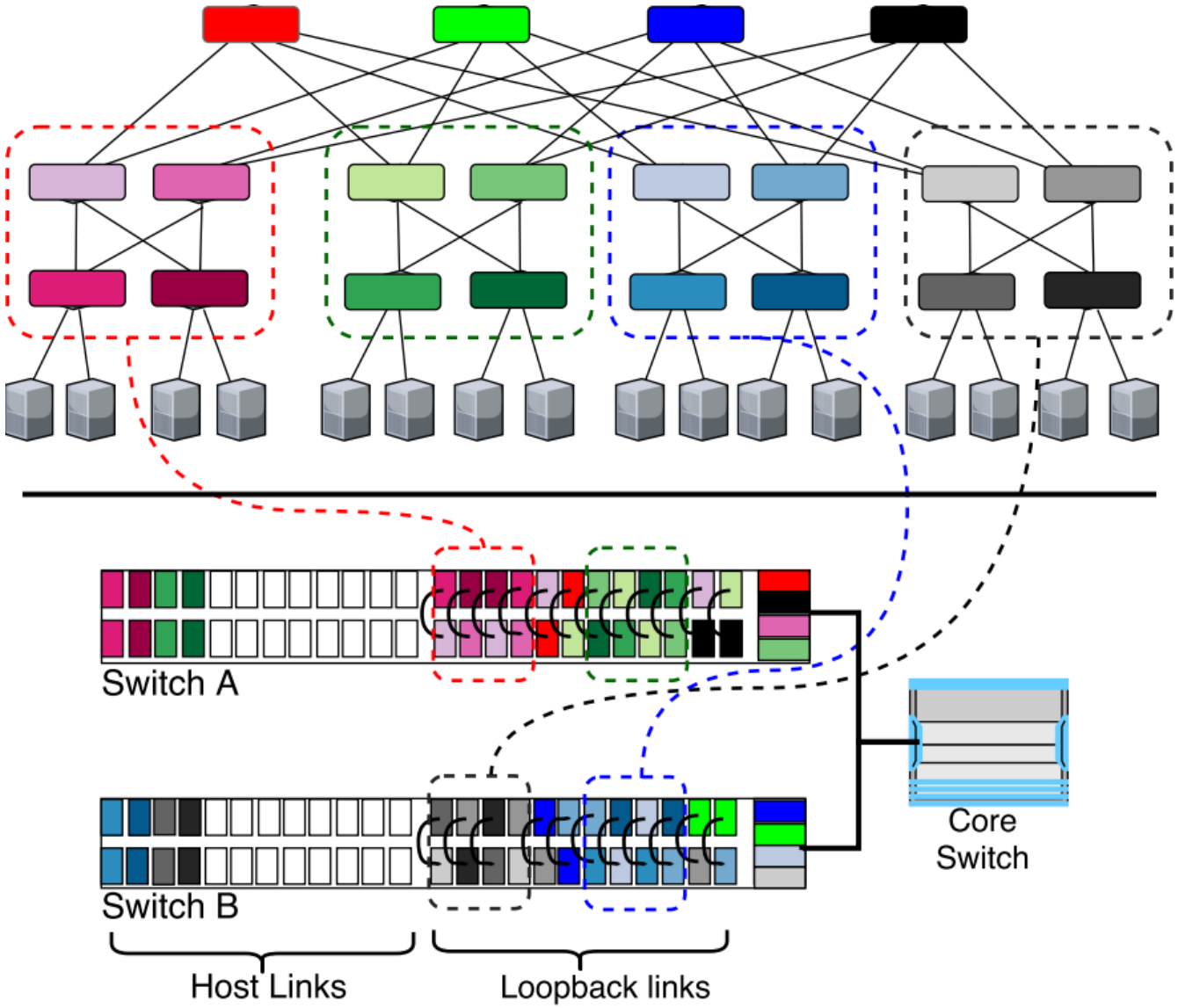


Evaluation : loopbacks & Big Switch Abstraction

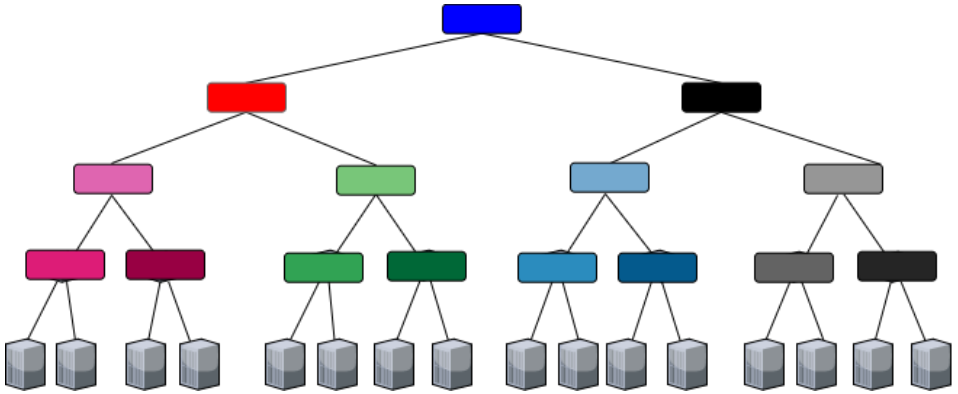




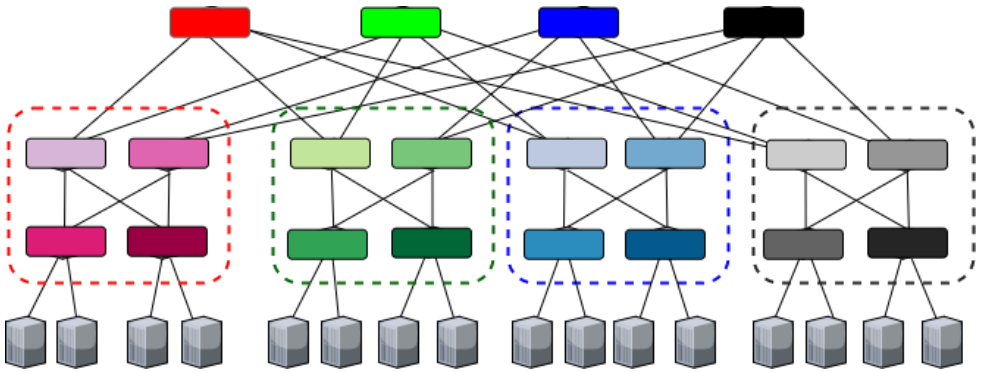
Mapping Complex Topologies



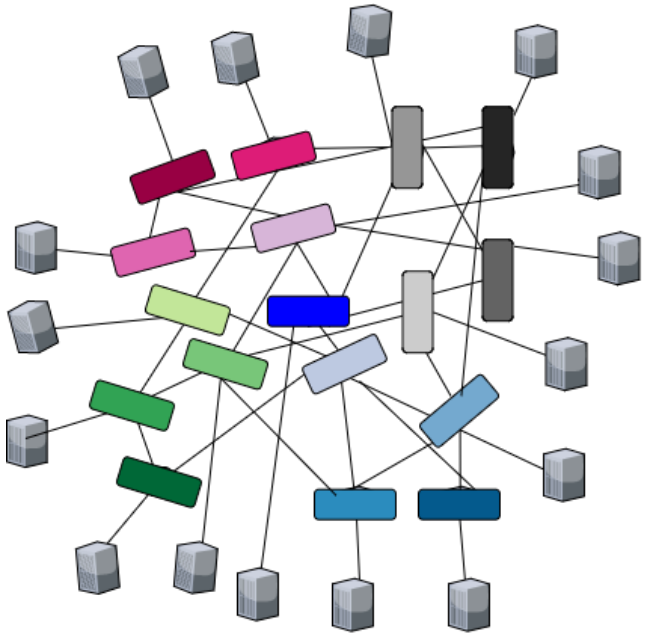
Evaluation : Fidelity of Complex Topologies



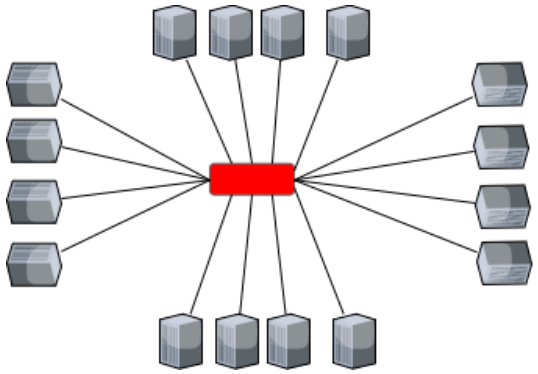
Single-Rooted Tree



Fat Tree

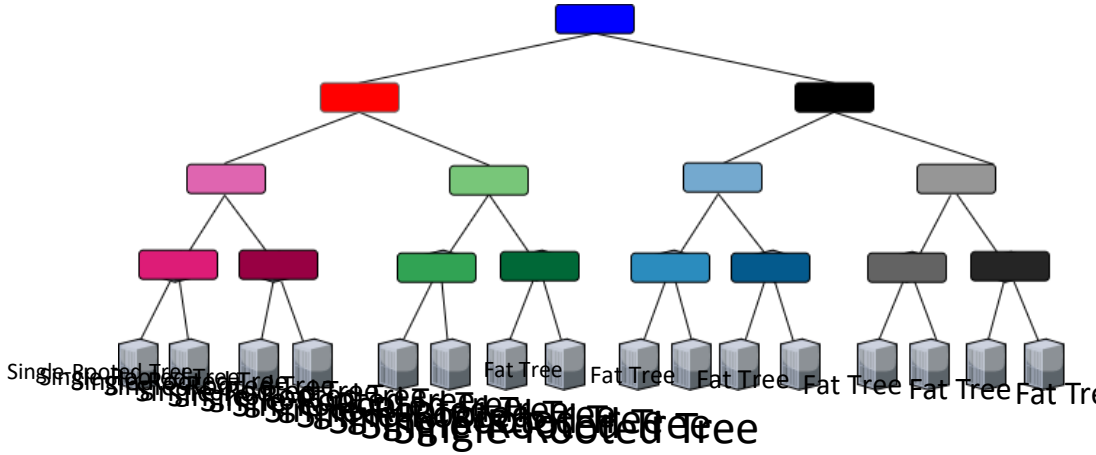


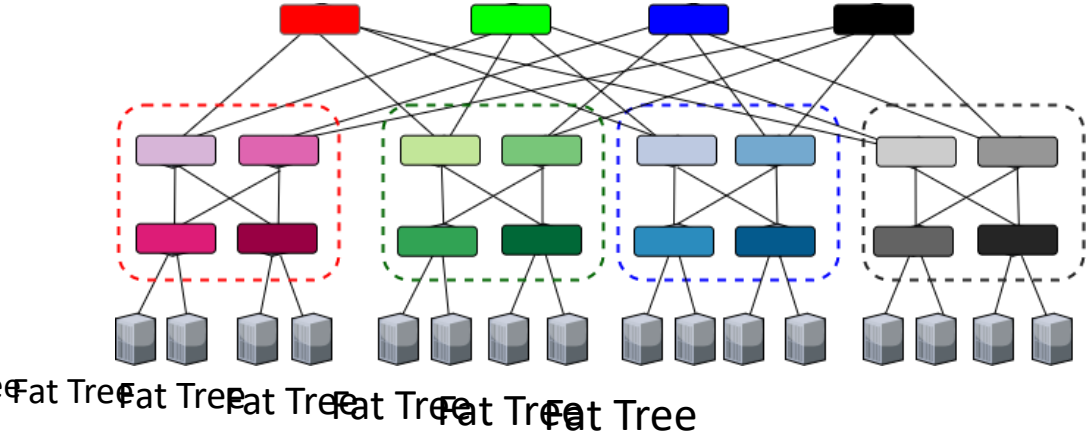
Jellyfish

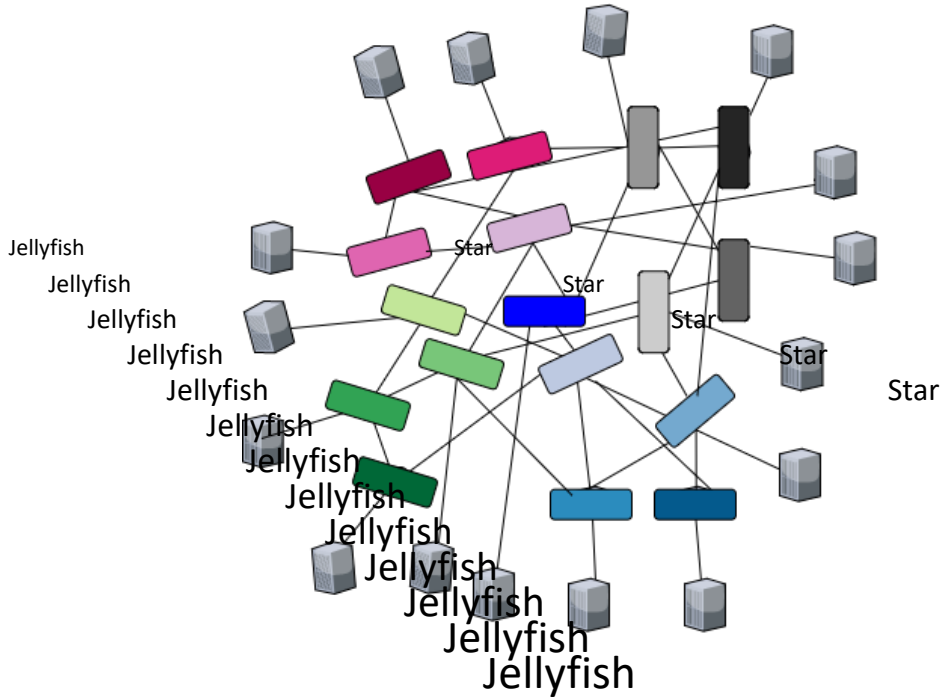


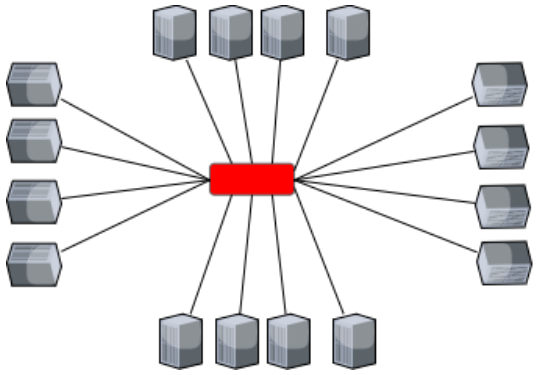
Star

Evaluation : Fidelity of Complex Topologies

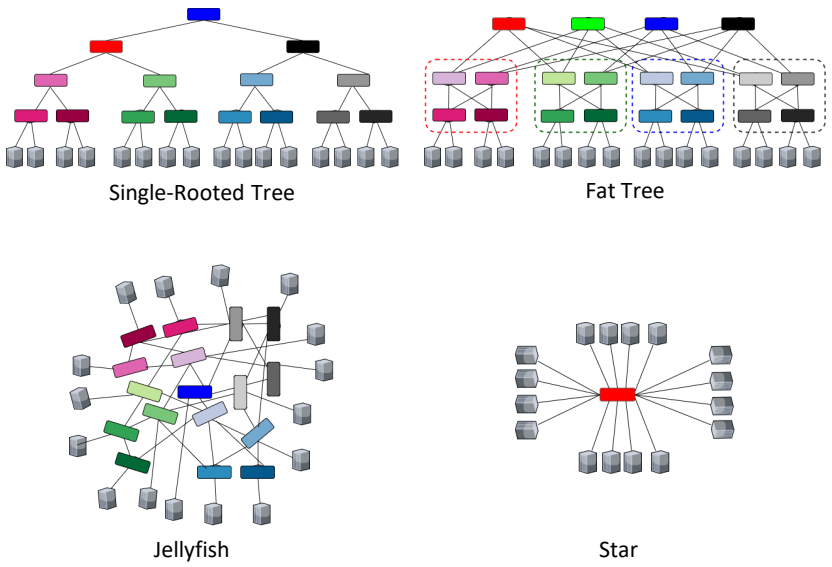


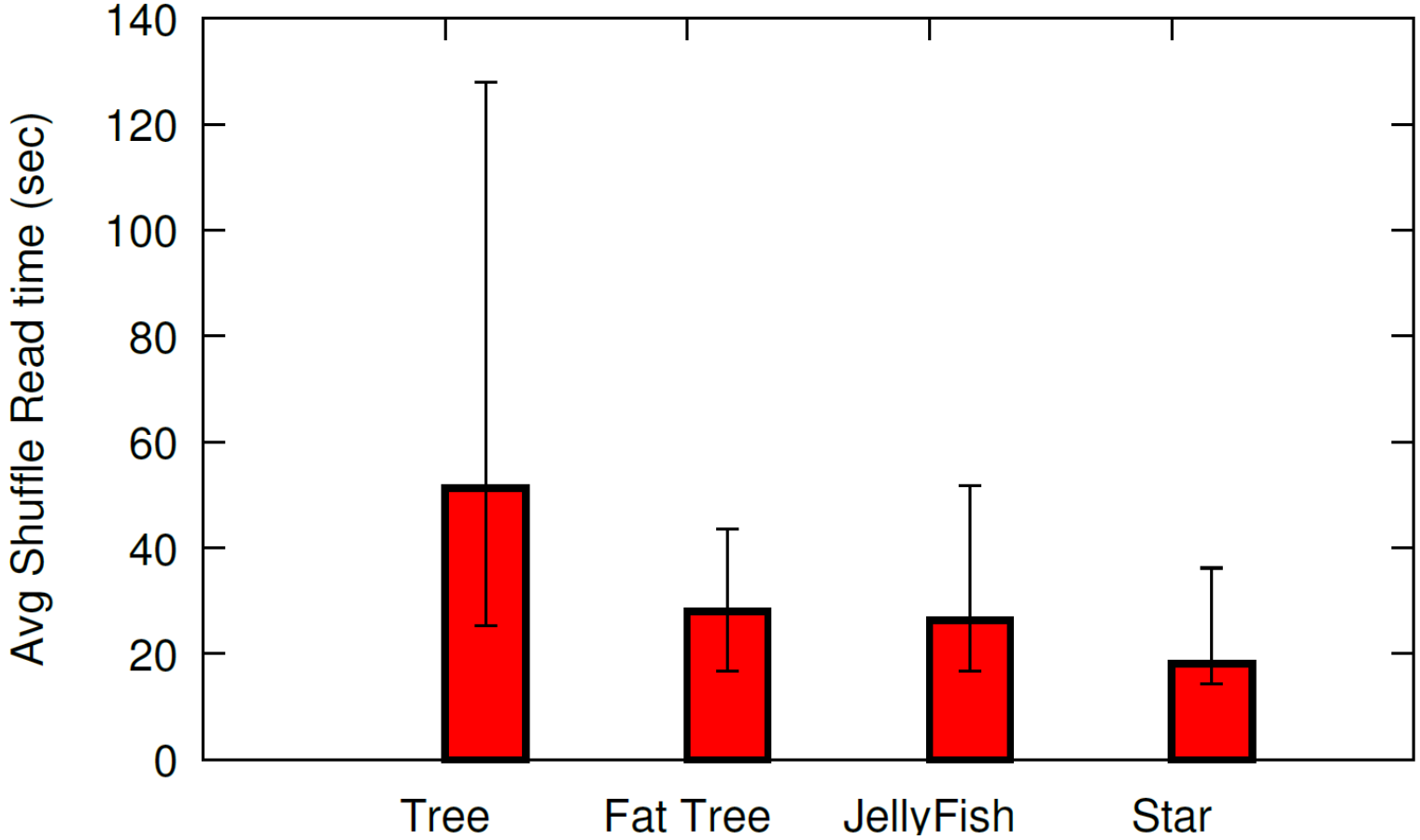




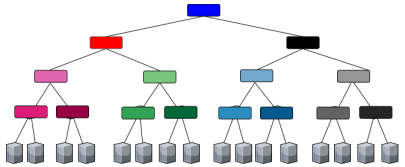


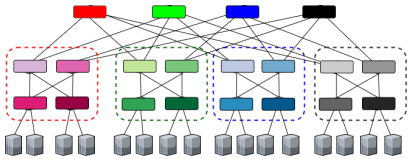
Evaluation : Fidelity of Complex Topologies

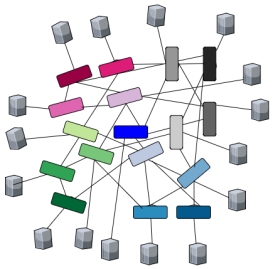


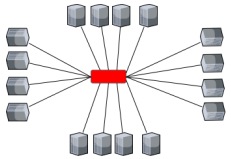


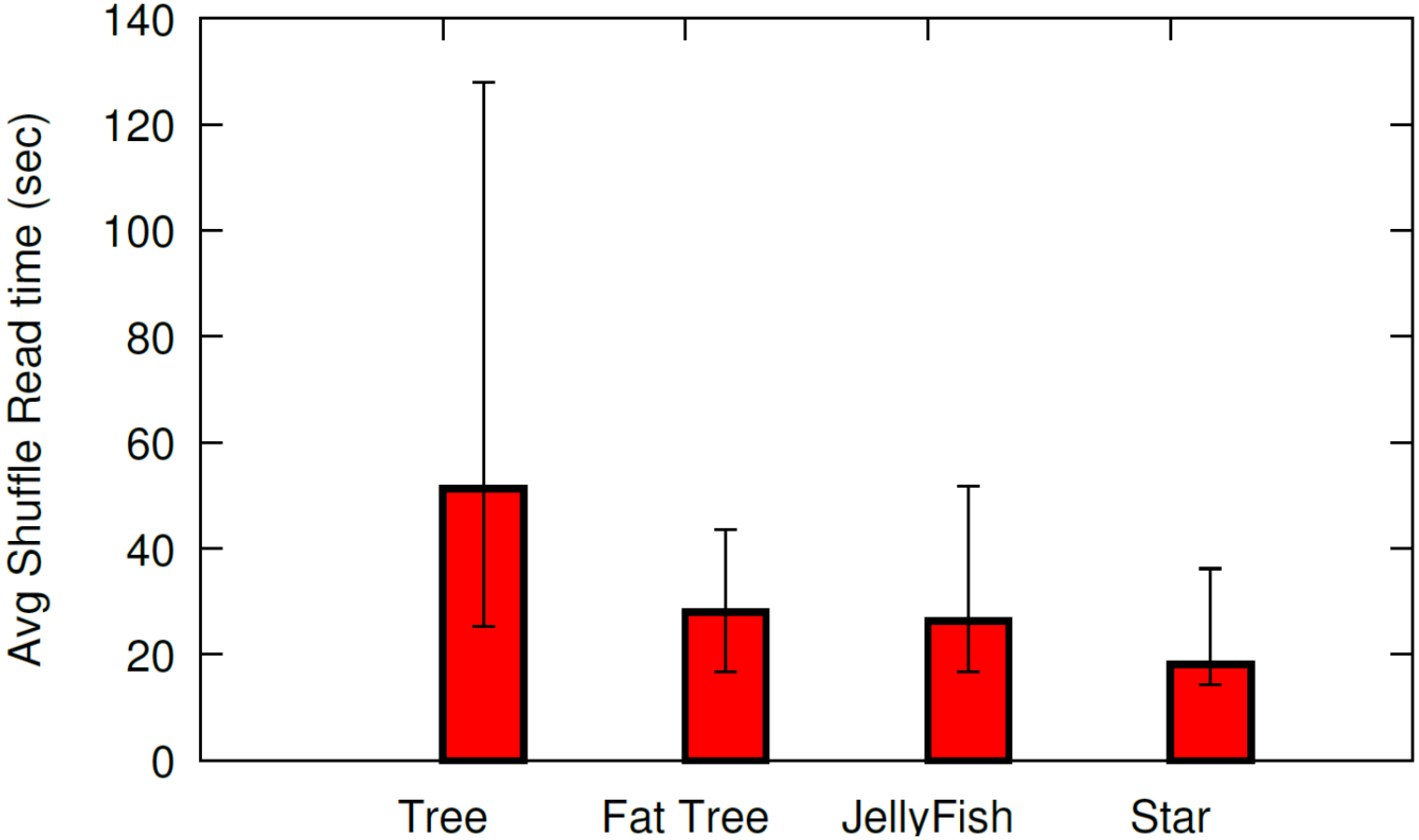
Evaluation : Fidelity of Complex Topologies


Single-Rooted Tree

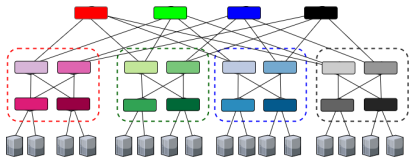

Fat Tree


Jellyfish

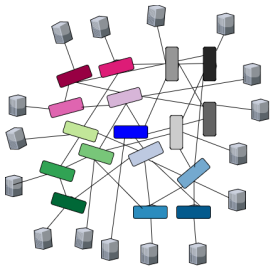

Star



Evaluation : Fidelity of Complex Topologies

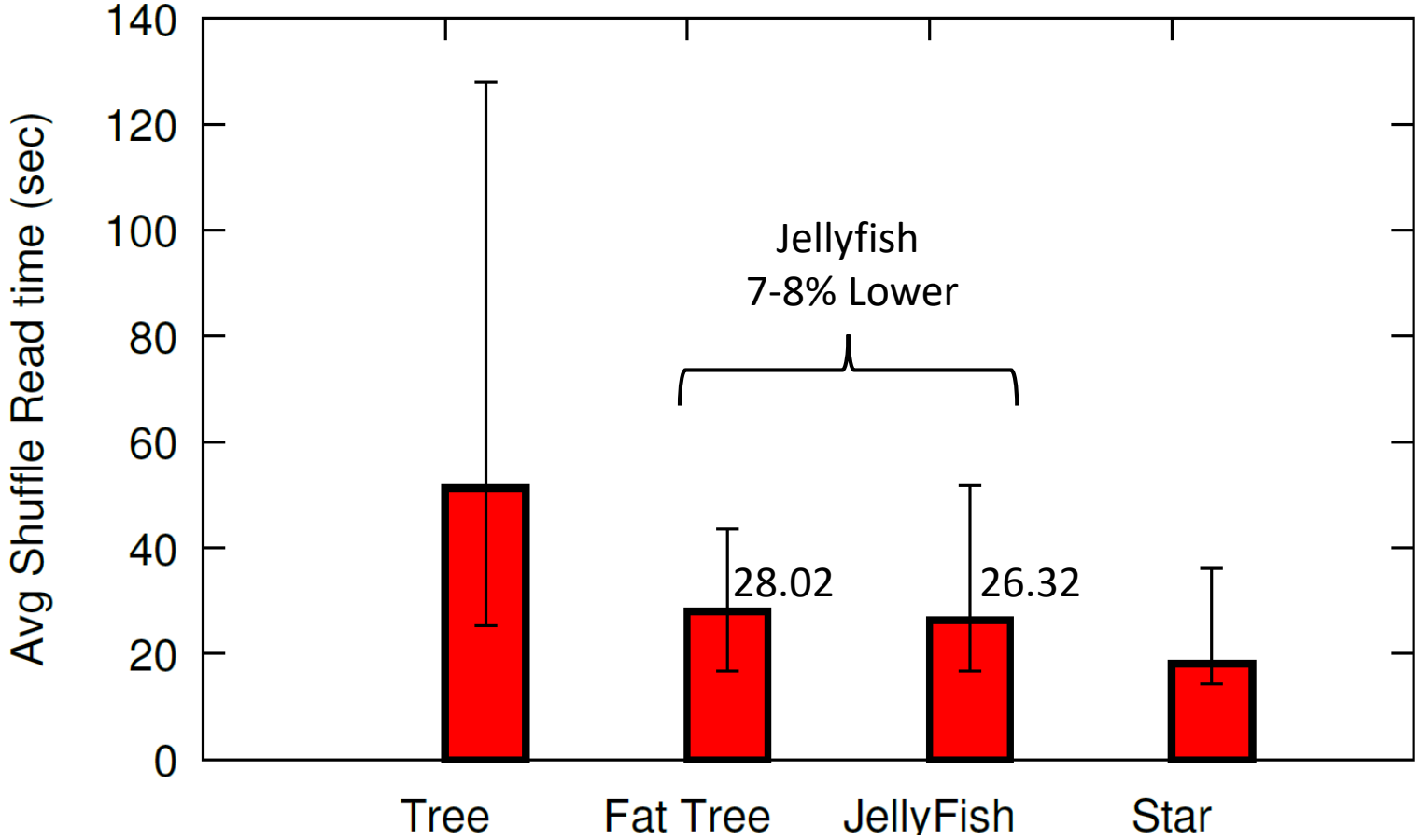


Fat Tree

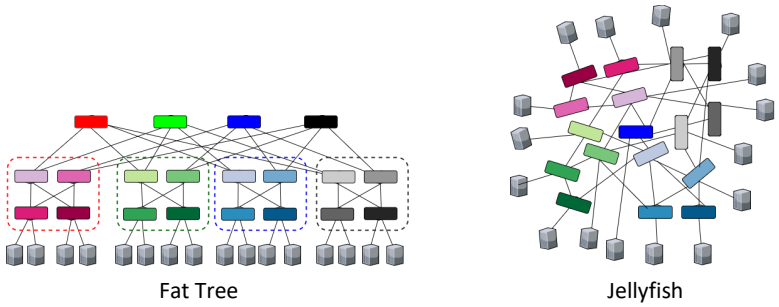


Jellyfish

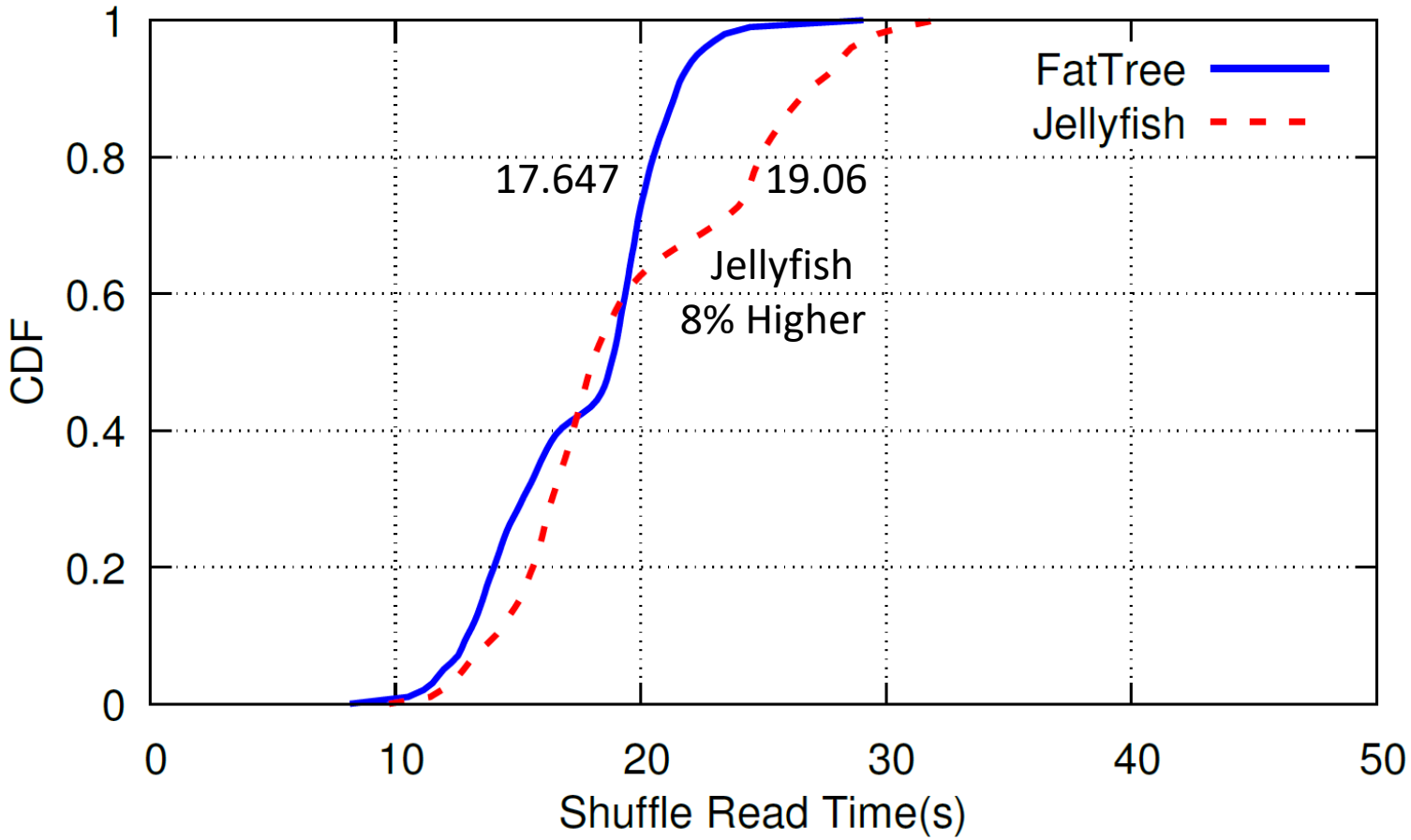
Low Intra-pod locality
Heavy cross rack traffic!!



Evaluation : Fidelity of Complex Topologies



High Intra-pod locality
Low cross rack traffic!!

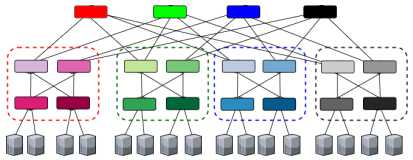


This observation is backed by Tale of Two Topologies [SIGCOMM '17]

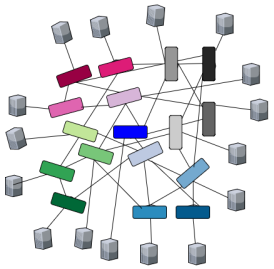
Take-away: BNV emulates the intrinsic characteristics of the topologies, otherwise observed with heavy wiring & building costs

Evaluation : Isolation of multiple tenants

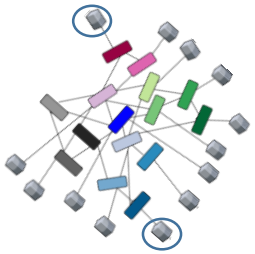
Take-away: BNV can support multiple arbitrary tenants providing isolated & repeatable environments.



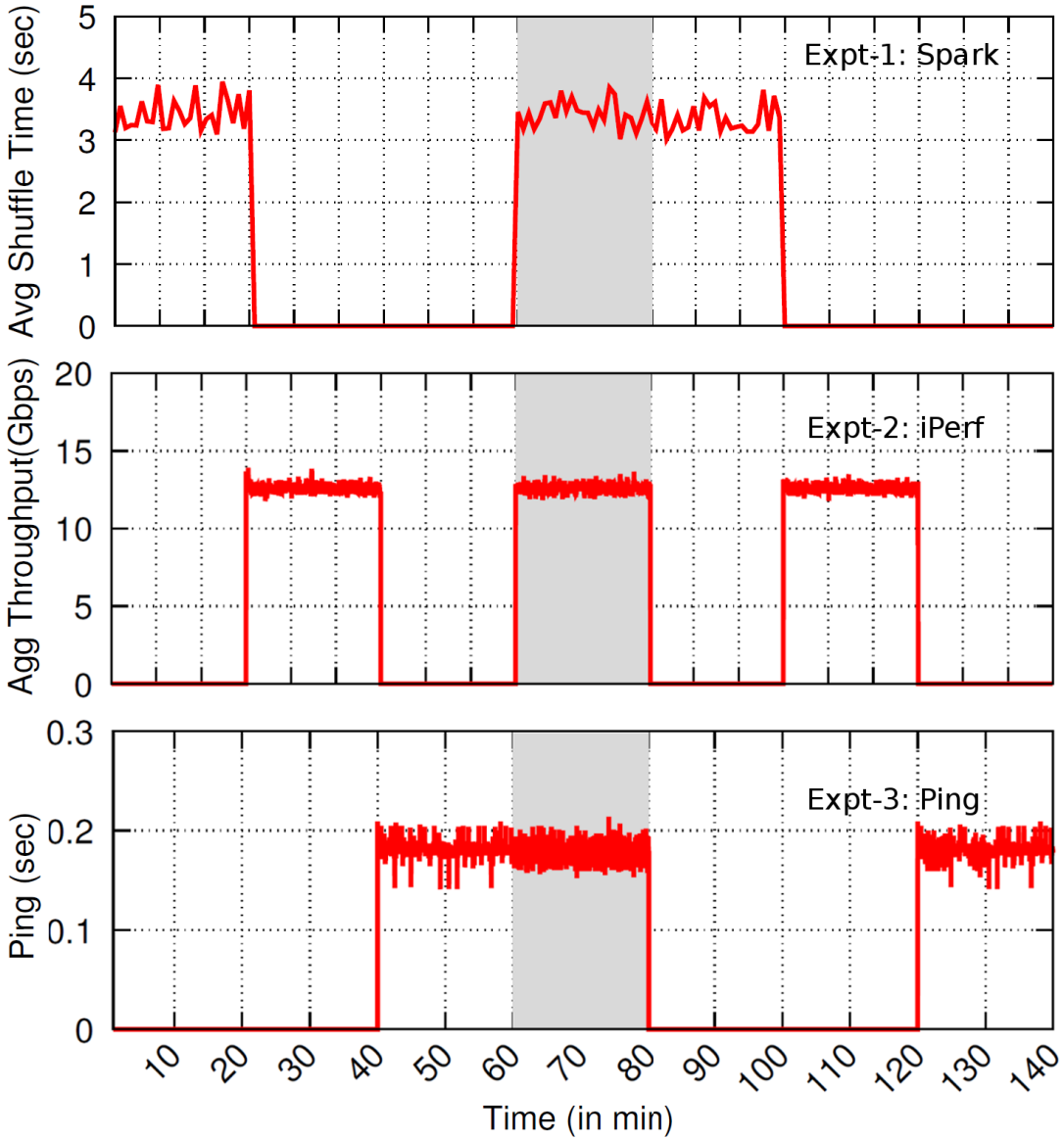
Fat Tree



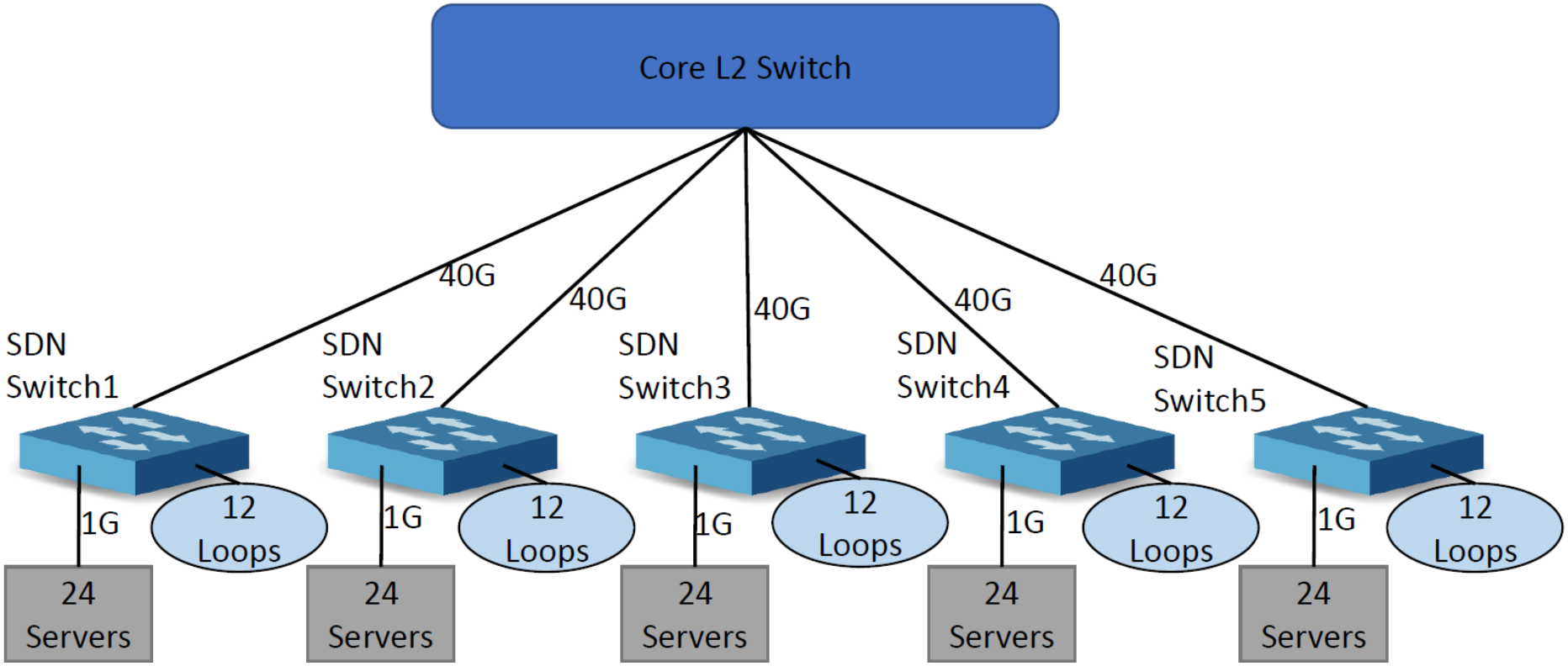
Jellyfish



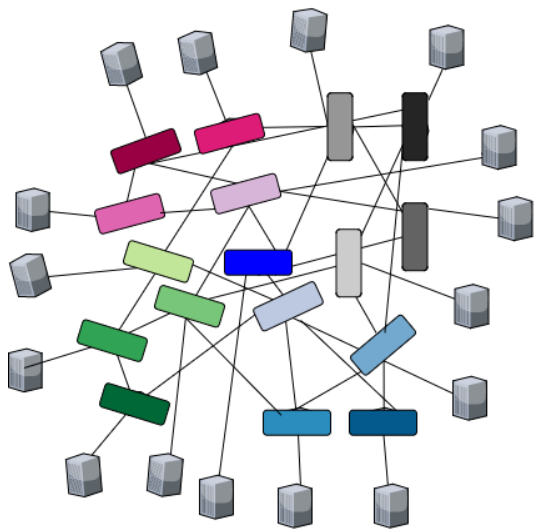
Random (Rest of the nodes)

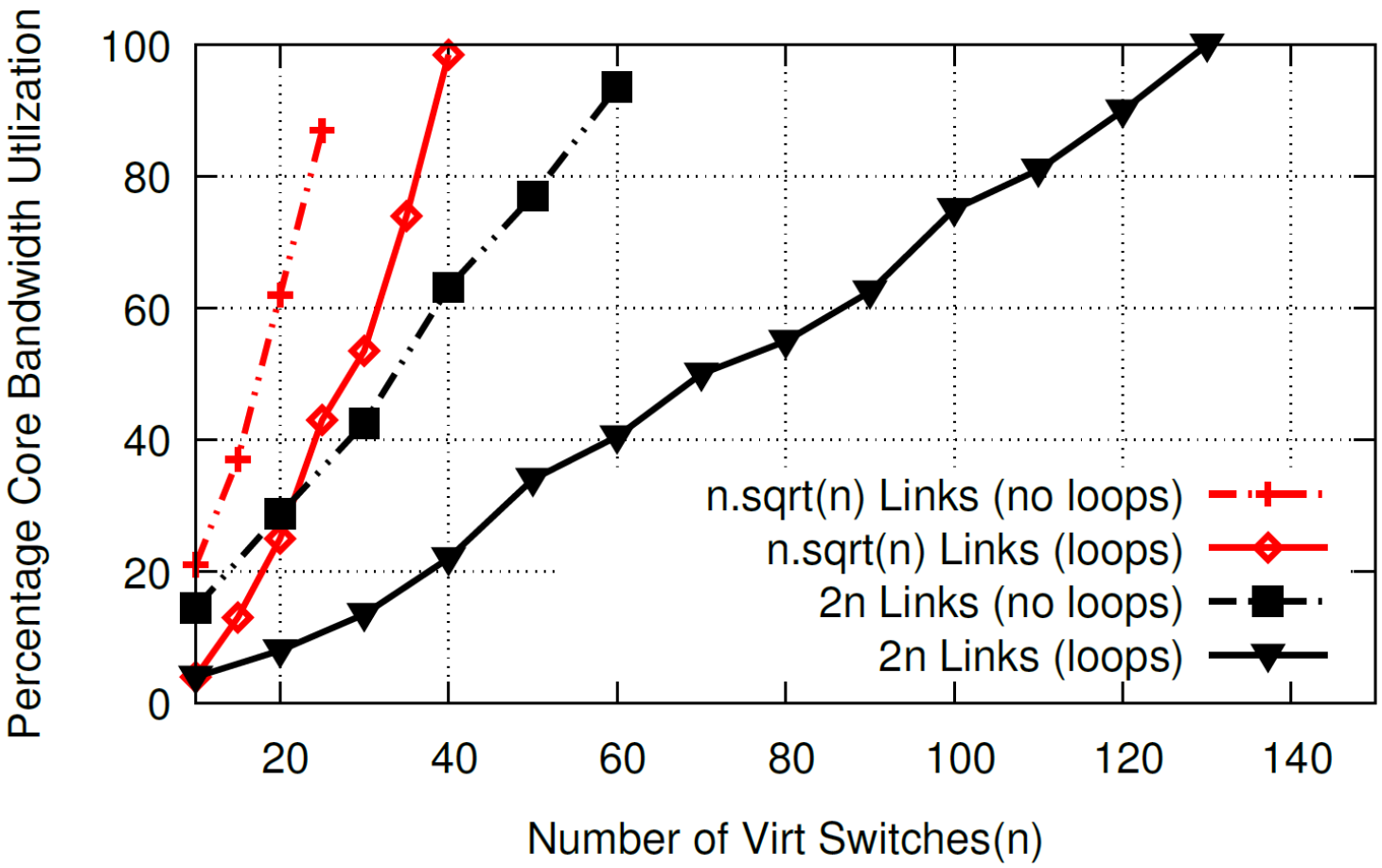


Evaluation : Scale of Topologies

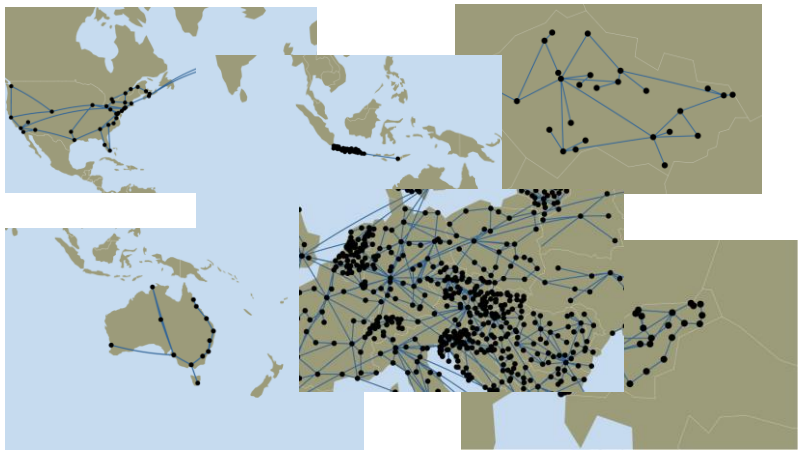


Evaluation : Scale of Topologies

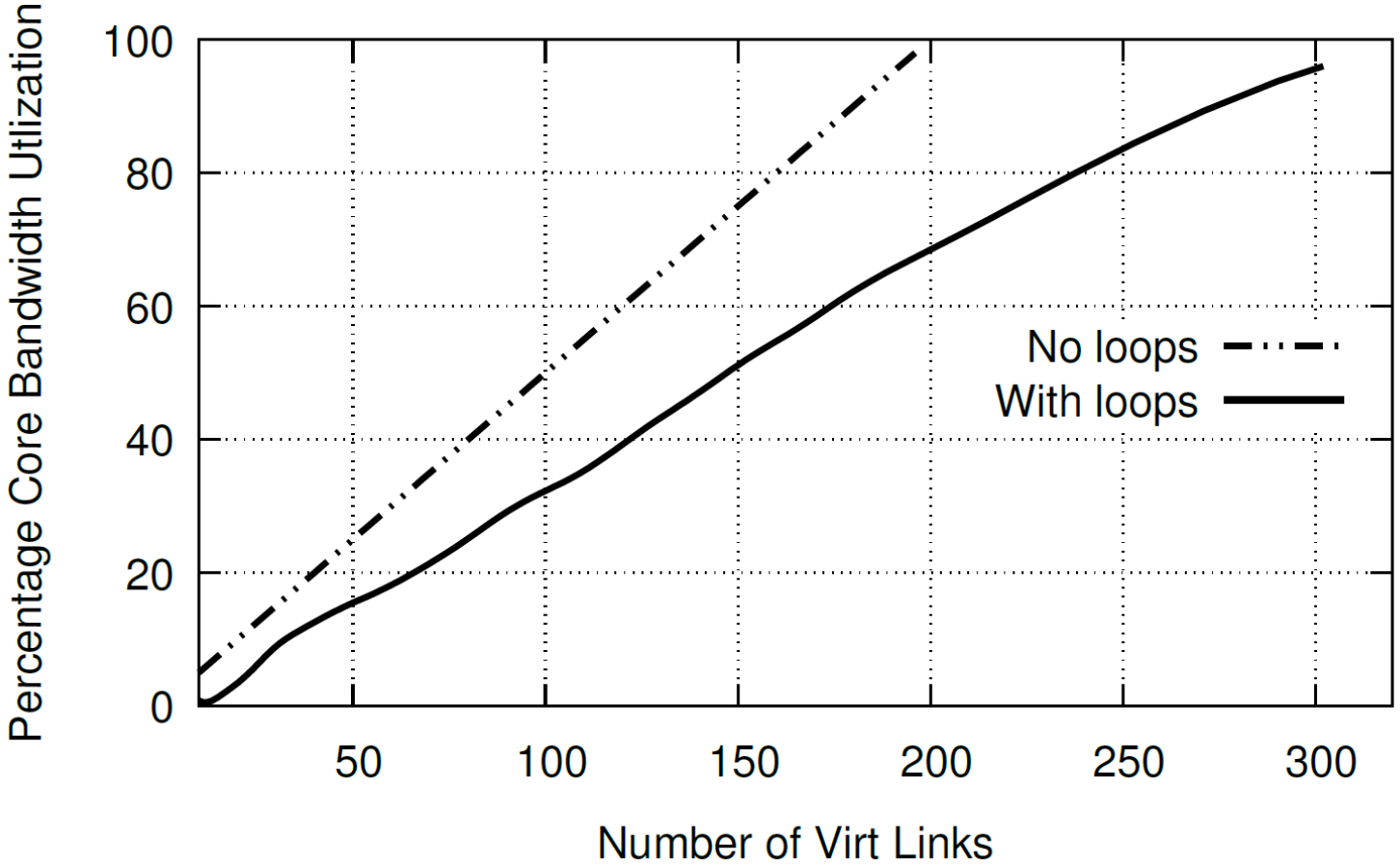




Evaluation : Scale of Topologies



Internet Topology Zoo dataset



Take-away: Addition of loopback links increases the scalability significantly.
Can emulate topologies with 130 switches and 300 links with just 5 ToR switches

Summary

BNV provides High-fidelity network experimentation/cloud platform.

BNV can scale a network to support arbitrary topologies using the existing network.

Can emulate networks of more than 100 network devices with just 5 ToR Switches.

Deployed in NCL testbed https://ncl.sg

Code Available : https://github.com/praveingk/bnvirt