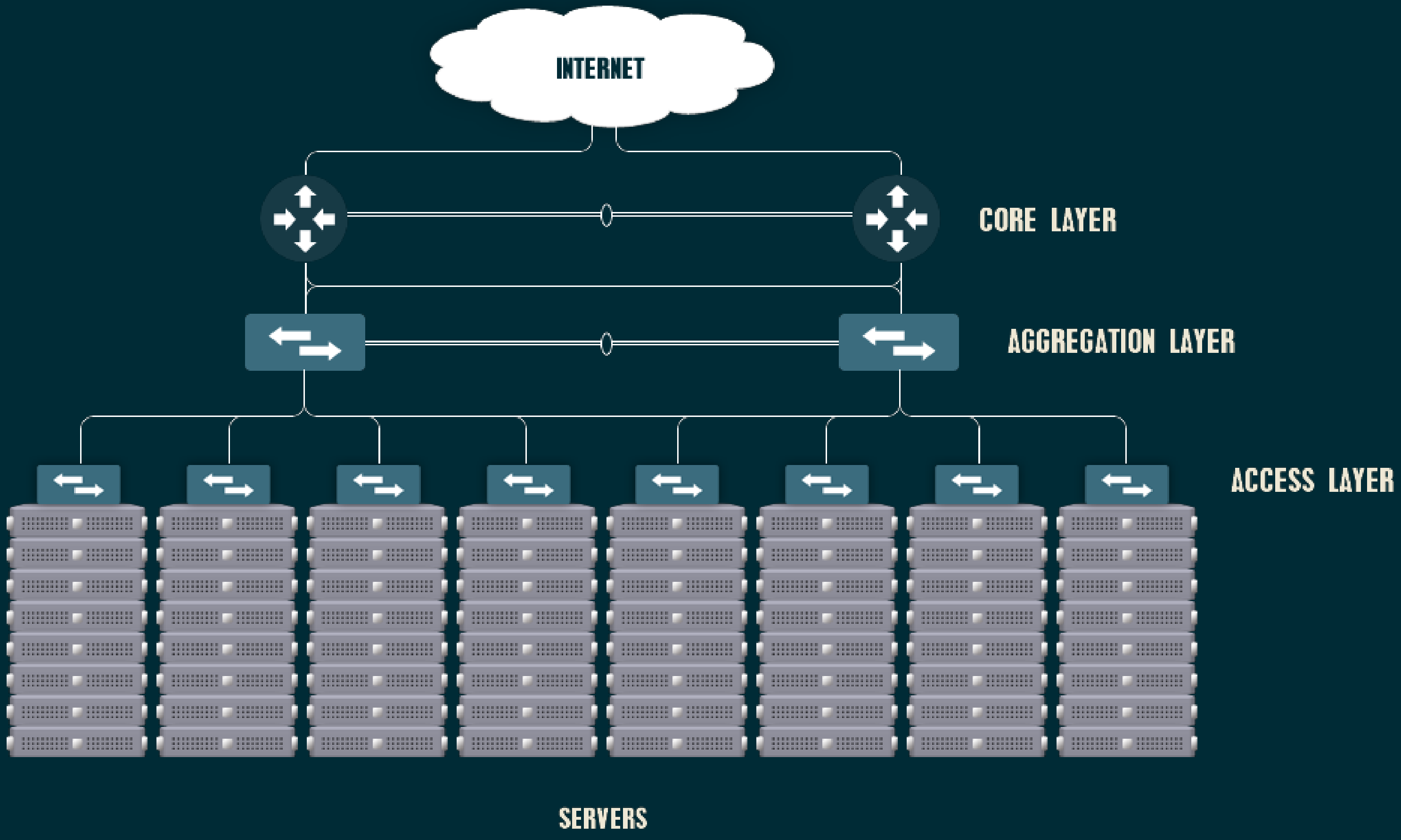
VIRTUAL NETWORK OVER TRILL

DESIGN, IMPLEMENTATION AND DEMONSTRATION

William Dauchy - Gandi.net

Kernel Recipes 2013

CONVENTIONAL DATA CENTER

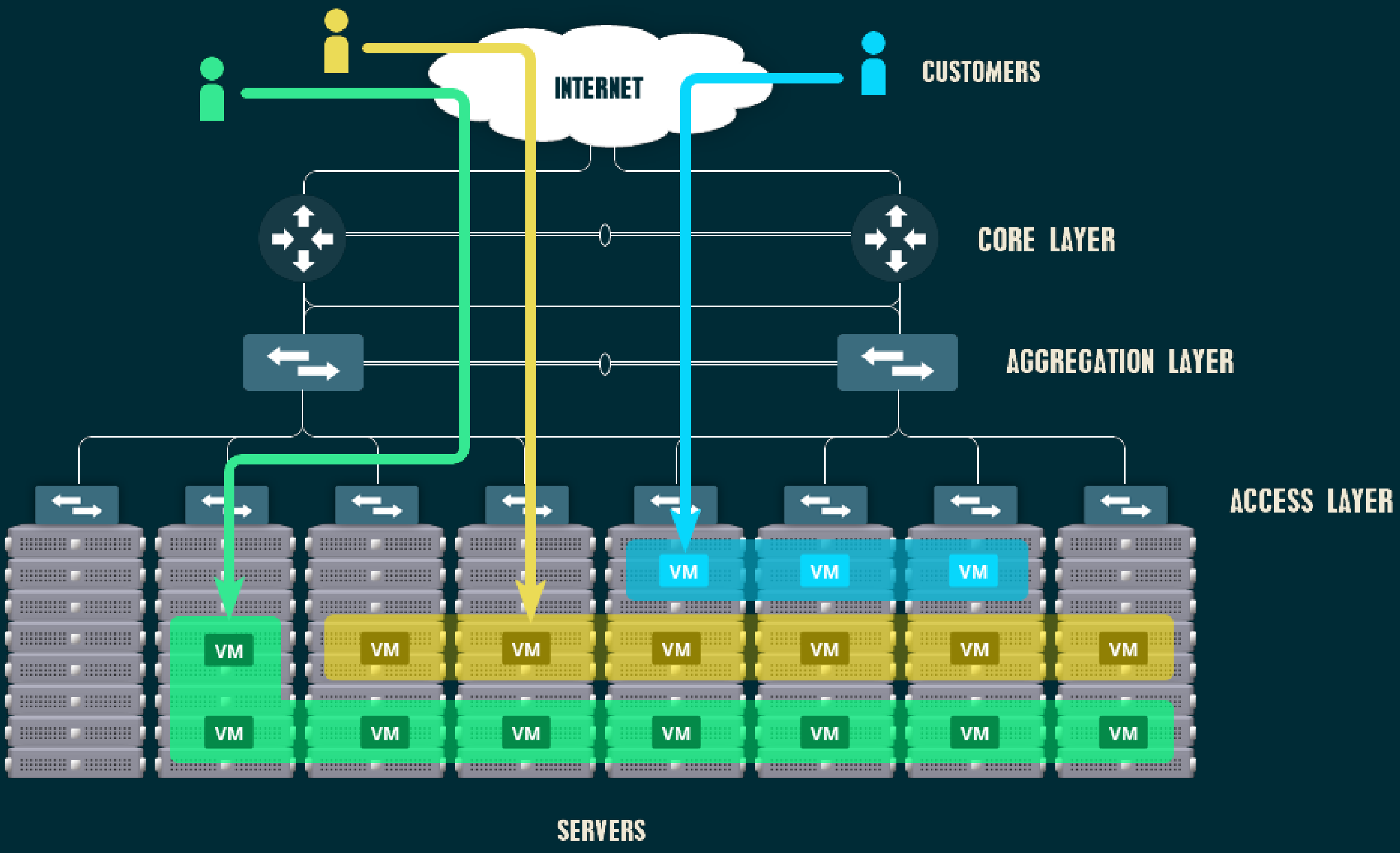


MAIN GOAL

- provide large scale multi-tenancy

LARGE SCALE MULTI-TENANCY

Multiples users using same resources



REQUIREMENTS

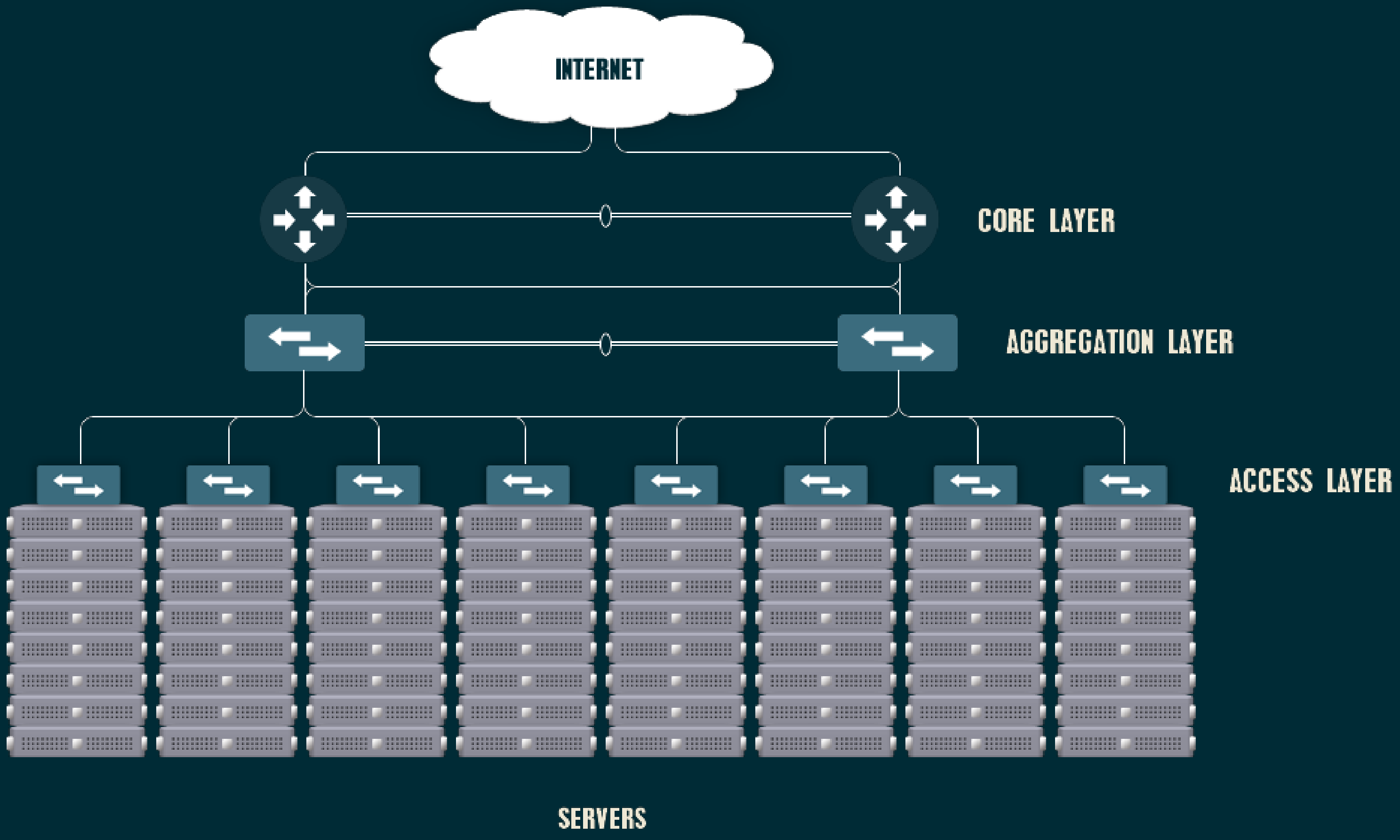
- Seamless VM mobility
- Easy management
- Layer 2 core scaling
- Fault resiliance
- VLAN scalability

LAYER 2 - SWITCHING BENEFITS

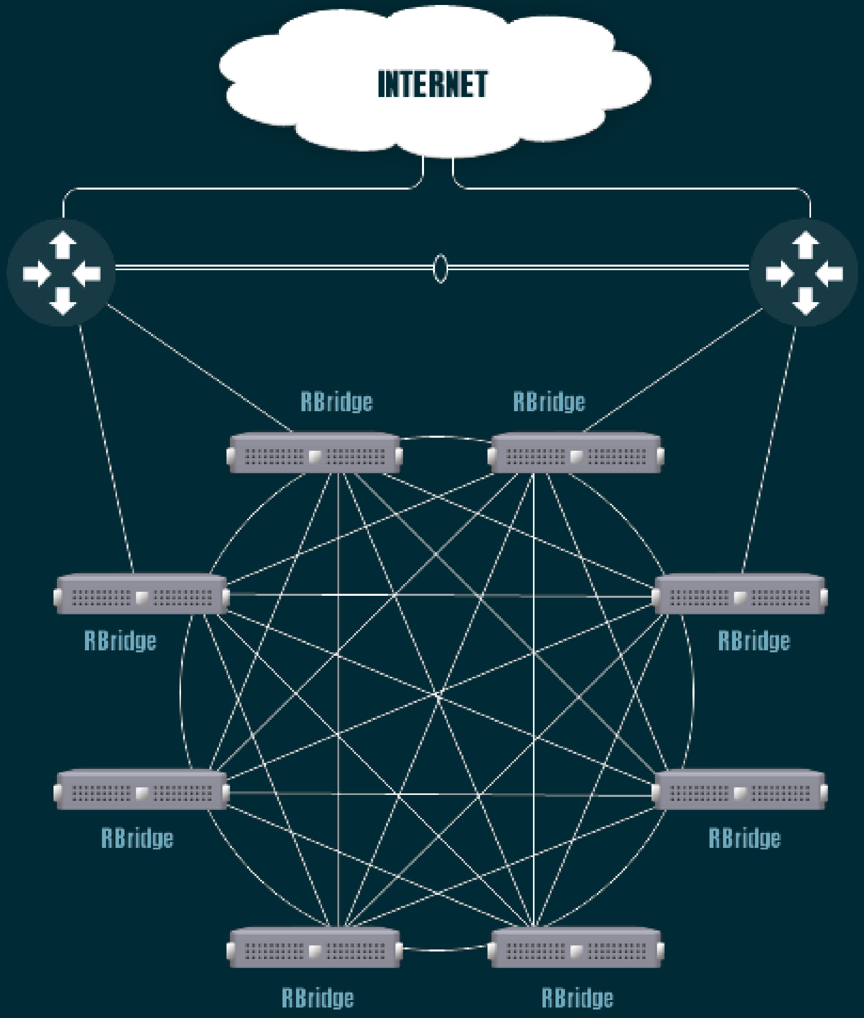
- Management simplified + Plug & play
- Seamless Virtual Machine mobility
- Auto learning + determistic failover

LAYER 2 - SWITCHING LIMITATION

- A large number of tenants implies
  - a huge number of MAC address in switch table (TCAM overflow)
  - ARP storm at nodes
- STP to ensure a loop free topology
  - blocking redundant paths
  - Core-computes required, recomputes when topology changes
- Number of VLANs is limited to 4096


INTERNET
CORE LAYER
AGGREGATION LAYER
ACCESS LAYER
SERVERS

TRILL BASED DATA CENTER

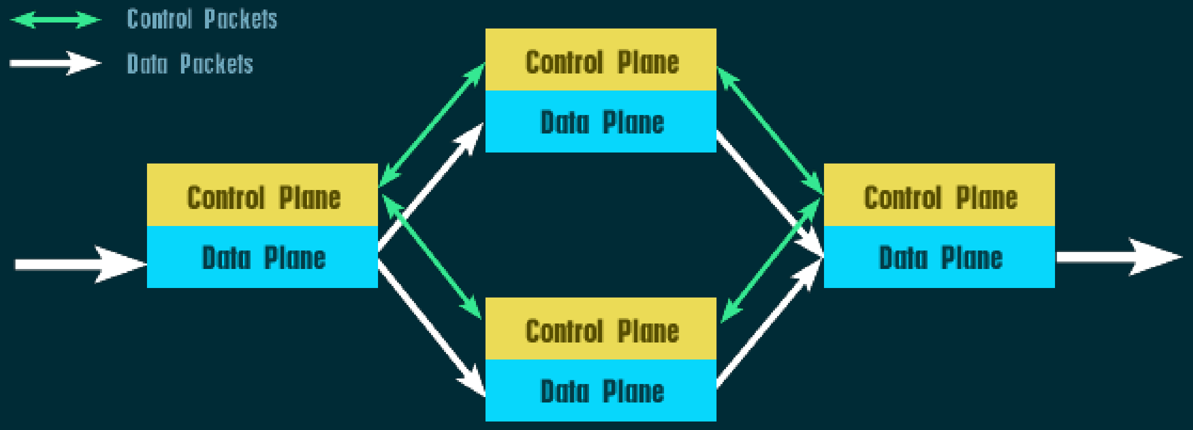


CORE LAYER

NON-BLOCKING
LAYER 2

WHAT IS TRILL

- New device: RBridge
  - Control plane
  - Data plane



- Encapsulate native frames in a transport header
- Providing a hop count and nickname
- Route the encapsulated frames using IS-IS
- Decapsulate native frames before delivery

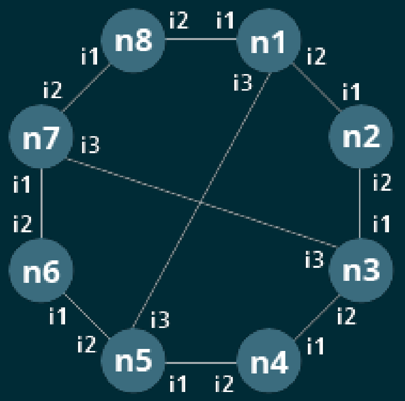
IETF STANDARD

- RFC 5556 Transparent Interconnection of Lots of Links (TRILL): Problem and Applicability Statement
- RFC 6325 Routing Bridges (RBridges): Base Protocol Specification
- RFC 6326 Transparent Interconnection of Lots of Links (TRILL) Use of IS-IS
- RFC 6327 Routing Bridges (RBridges): Adjacency
- RFC 6439 Routing Bridges (RBridges): Appointed Forwarders
- RFC 6361 PPP Transparent Interconnection of Lots of Links (TRILL) Protocol Control Protocol

DESIGN AND IMPLEMENTATION

CONTROL PLANE

unicast building



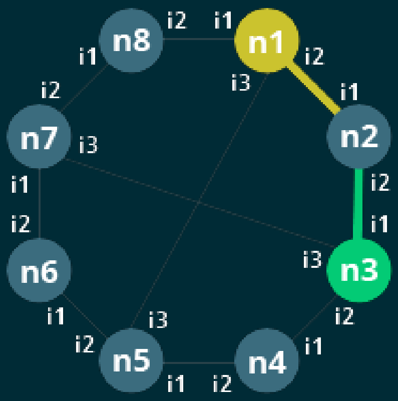
TRILL
FORWARDING
TABLE

Dest-Nickname	Port	Nexthop Mac

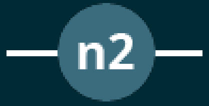
MULTICAST
TREE

CONTROL PLANE

unicast building - first iteration



TRILL FORWARDING TABLE

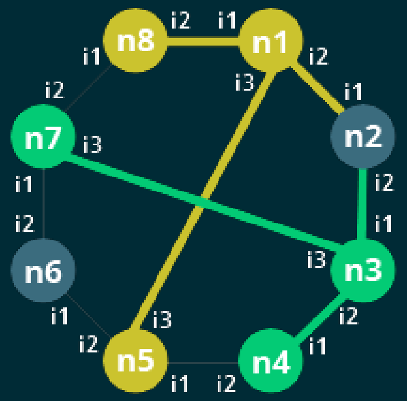


Dest-Nickname	Port	Nexthop Mac
n1	i1	-
n3	i2	-

MULTICAST TREE

CONTROL PLANE

unicast building - second iteration



TRILL FORWARDING TABLE

n2

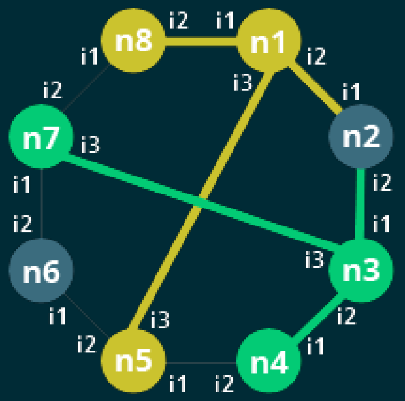
Dest-Nickname	Port	Nexthop Mac
n1	i1	-
n3	i2	-
n8	i1	MAC - n1
n5	i1	MAC - n1
n4	i2	MAC - n3
n7	i2	MAC - n3

MULTICAST TREE

n1

CONTROL PLANE

unicast building - third iteration



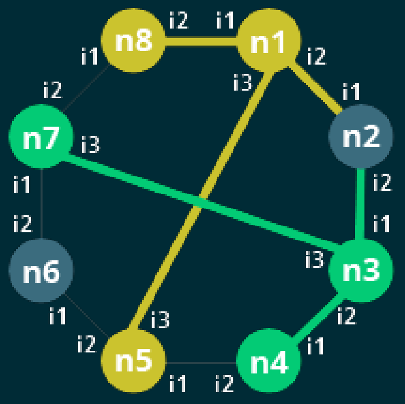
TRILL FORWARDING TABLE

MULTICAST TREE

Dest-Nickname	Port	Nexthop Mac
n1	i1	-
n3	i2	-
n8	i1	MAC - n1
n5	i1	MAC - n1
n4	i2	MAC - n3
n7	i2	MAC - n3
n6	i1	MAC - n1
n6	i2	MAC - n3

CONTROL PLANE

unicast building - final result



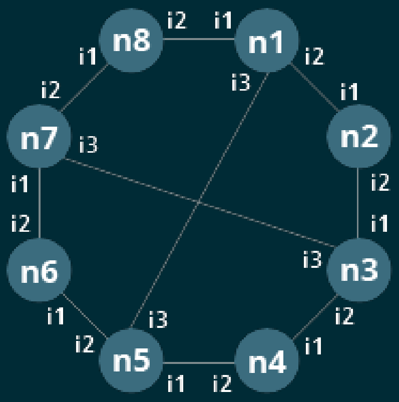
TRILL FORWARDING TABLE

Dest-Nickname	Port	Nexthop Mac
n1	i1	-
n3	i2	-
n8	i1	MAC - n1
n5	i1	MAC - n1
n4	i2	MAC - n3
n7	i2	MAC - n3
n6	i1	MAC - n1
n6	i2	MAC - n3

ECMP:
Equal Cost
Multipath

MULTICAST TREE

CONTROL PLANE



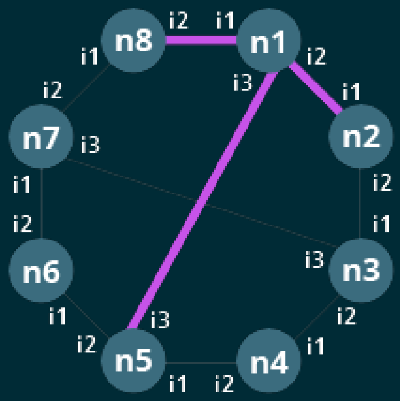
TRILL FORWARDING TABLE

Dest-Nickname	Port	Nexthop Mac
n1	i1	-
n3	i2	-
n8	i1	MAC - n1
n5	i1	MAC - n1
n4	i2	MAC - n3
n7	i2	MAC - n3
n6	i1	MAC - n1
n6	i2	MAC - n3

MULTICAST TREE

CONTROL PLANE

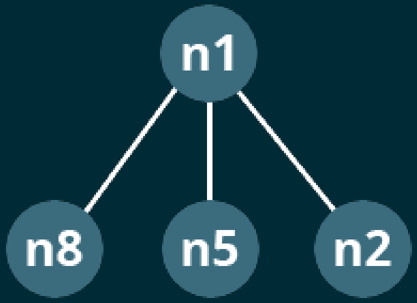
multicast building - first iteration



TRILL FORWARDING TABLE

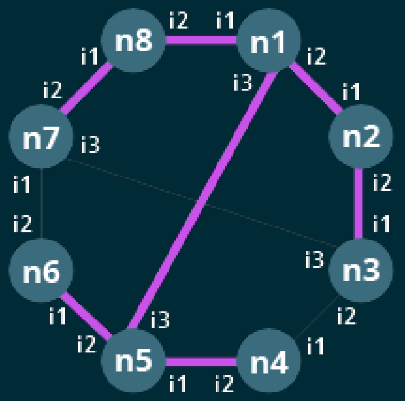
Dest-Nickname	Port	Nexthop Mac
n1	i1	-
n3	i2	-
n8	i1	MAC - n1
n5	i1	MAC - n1
n4	i2	MAC - n3
n7	i2	MAC - n3
n6	i1	MAC - n1
n6	i2	MAC - n3

MULTICAST TREE



CONTROL PLANE

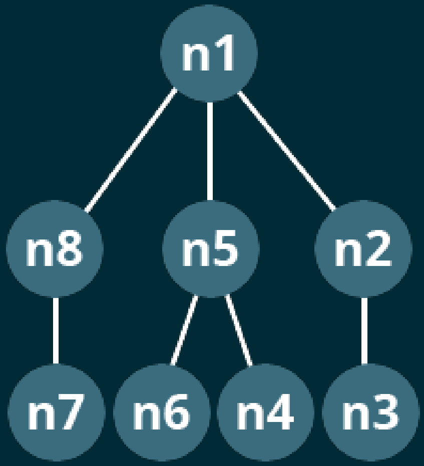
multicast building - final iteration



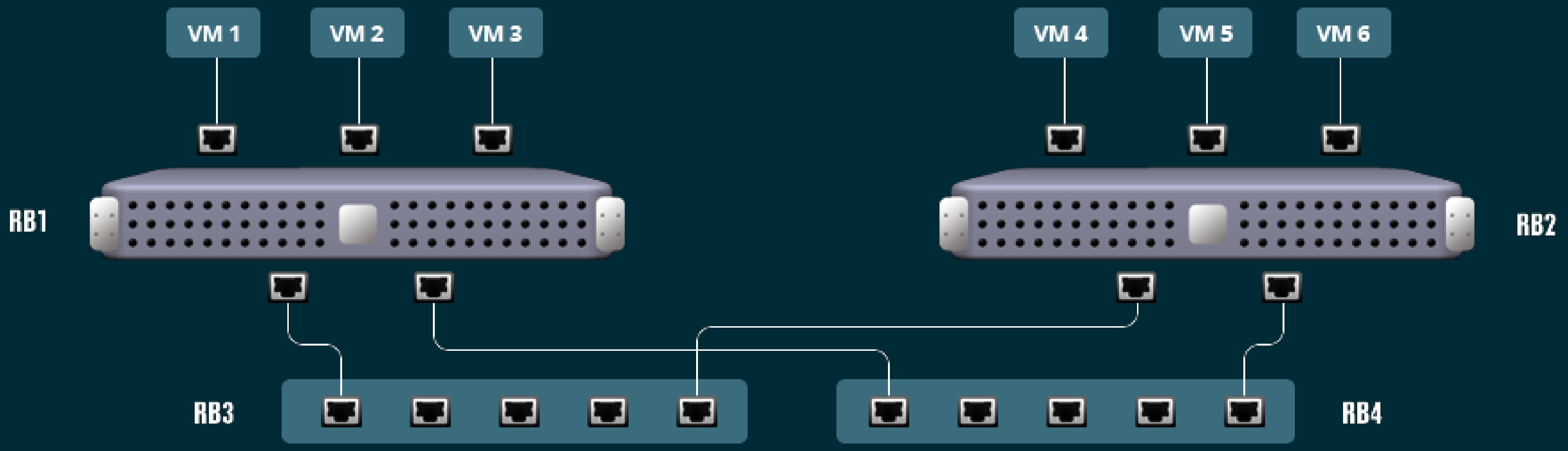
TRILL FORWARDING TABLE

Dest-Nickname	Port	Nexthop Mac
n1	i1	-
n3	i2	-
n8	i1	MAC - n1
n5	i1	MAC - n1
n4	i2	MAC - n3
n7	i2	MAC - n3
n6	i1	MAC - n1
n6	i2	MAC - n3

MULTICAST TREE

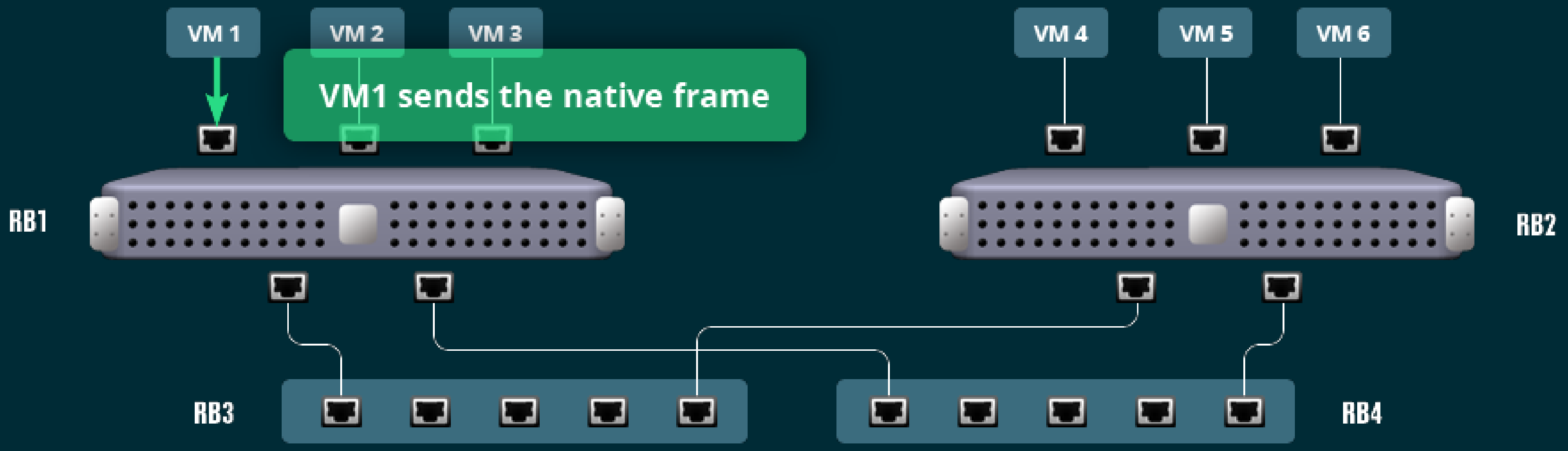


DATA PLANE



	MAC - RB3	MAC - RB2	
	MAC - RB1	MAC - RB3	
	NickName RB2	NickName RB2	
	Nickname RB1	Nickname RB1	
MAC - VM6	MAC - VM6	MAC - VM6	MAC - VM6
MAC - VM1	MAC - VM1	MAC - VM1	MAC - VM1
IP - VM6	IP - VM6	IP - VM6	IP - VM6
IP - VM1	IP - VM1	IP - VM1	IP - VM1
DATA	DATA	DATA	DATA

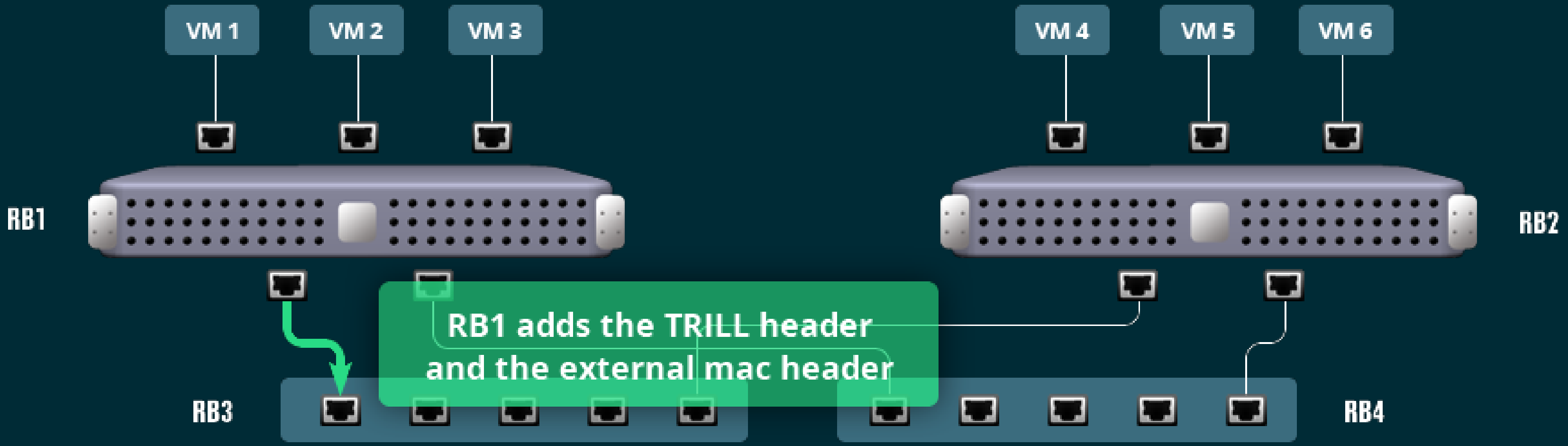
DATA PLANE



VM1 sends the native frame

	MAC - RB3	MAC - RB2	
	MAC - RB1	MAC - RB3	
	NickName RB2	NickName RB2	
	Nickname RB1	Nickname RB1	
MAC - VM6	MAC - VM6	MAC - VM6	MAC - VM6
MAC - VM1	MAC - VM1	MAC - VM1	MAC - VM1
IP - VM6	IP - VM6	IP - VM6	IP - VM6
IP - VM1	IP - VM1	IP - VM1	IP - VM1
DATA	DATA	DATA	DATA

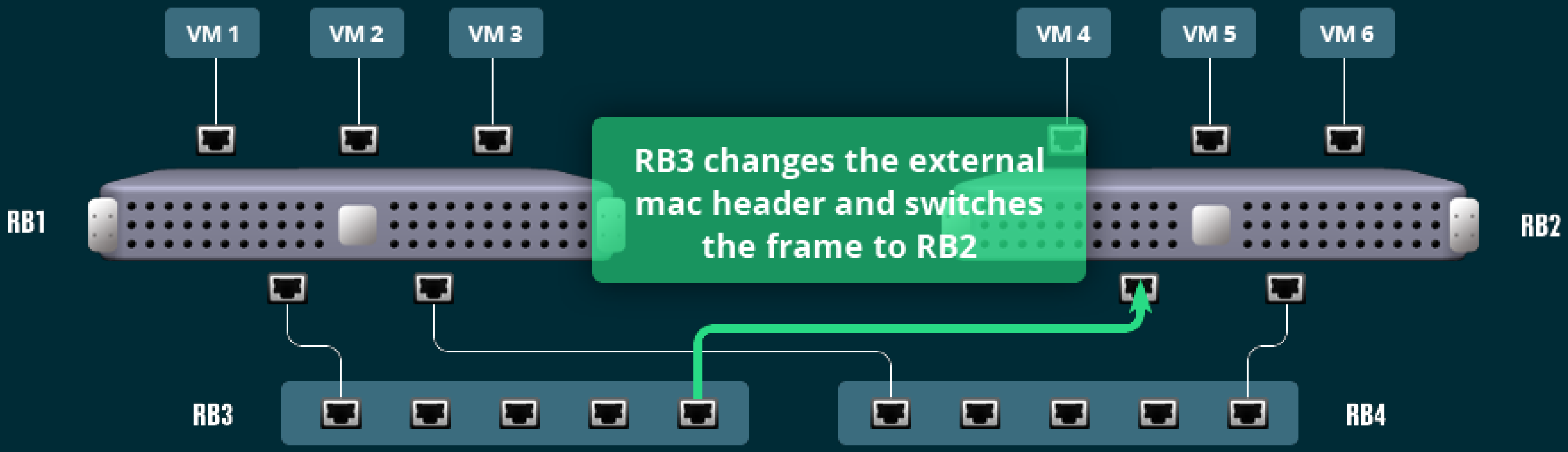
DATA PLANE



RB1 adds the TRILL header and the external mac header

	MAC - RB3	MAC - RB2	
	MAC - RB1	MAC - RB3	
	NickName RB2	NickName RB2	
	Nickname RB1	Nickname RB1	
MAC - VM6	MAC - VM6	MAC - VM6	MAC - VM6
MAC - VM1	MAC - VM1	MAC - VM1	MAC - VM1
IP - VM6	IP - VM6	IP - VM6	IP - VM6
IP - VM1	IP - VM1	IP - VM1	IP - VM1
DATA	DATA	DATA	DATA

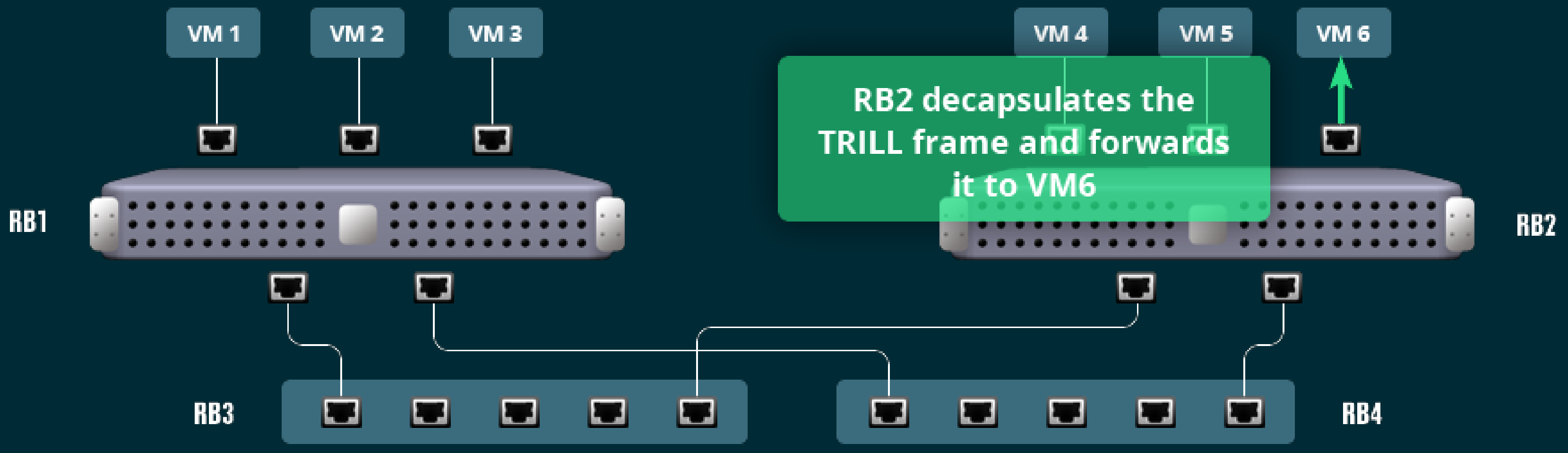
DATA PLANE



RB3 changes the external mac header and switches the frame to RB2

	MAC - RB3	MAC - RB2	
	MAC - RB1	MAC - RB3	
	NickName RB2	NickName RB2	
	Nickname RB1	Nickname RB1	
MAC - VM6	MAC - VM6	MAC - VM6	MAC - VM6
MAC - VM1	MAC - VM1	MAC - VM1	MAC - VM1
IP - VM6	IP - VM6	IP - VM6	IP - VM6
IP - VM1	IP - VM1	IP - VM1	IP - VM1
DATA	DATA	DATA	DATA

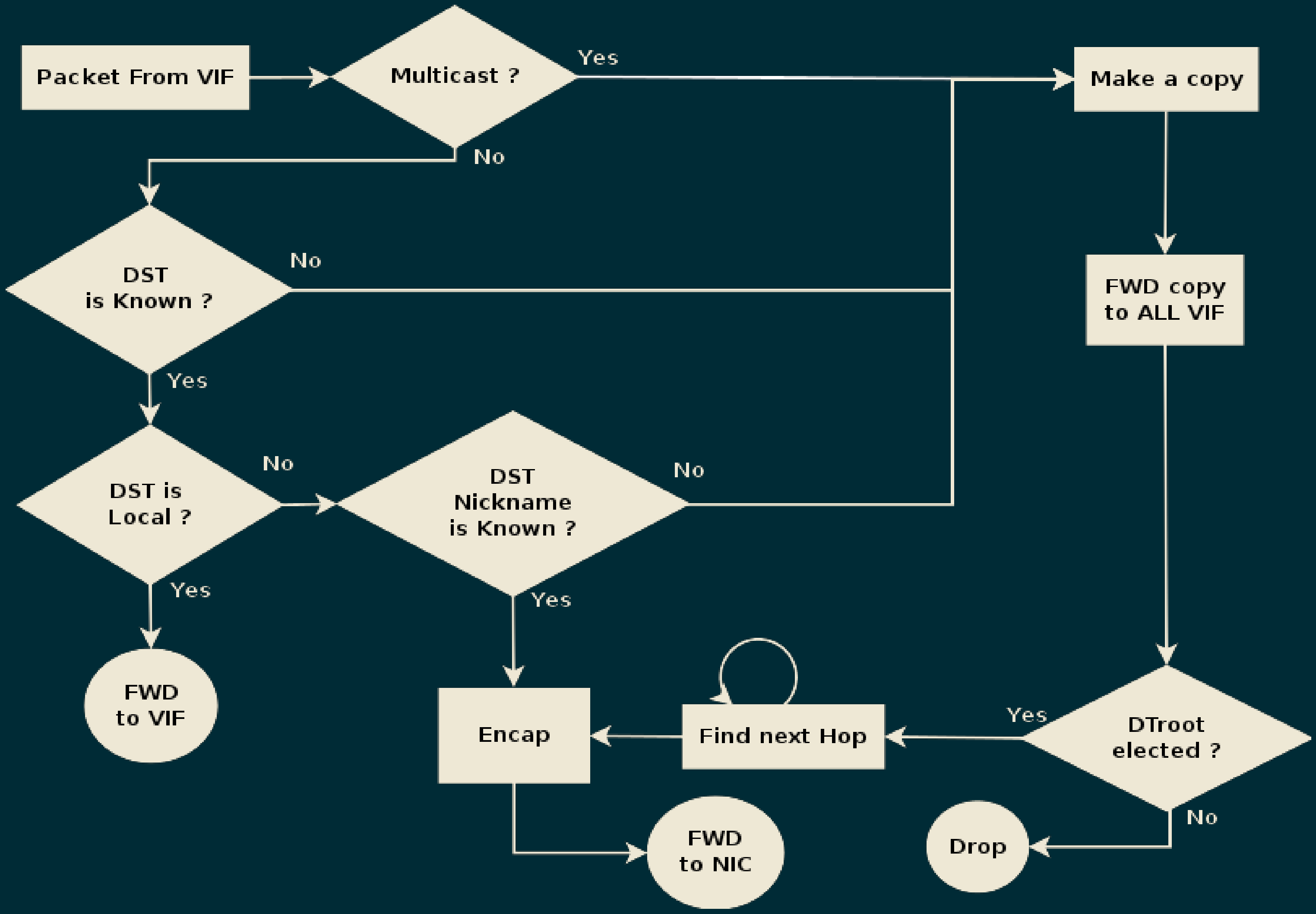
DATA PLANE



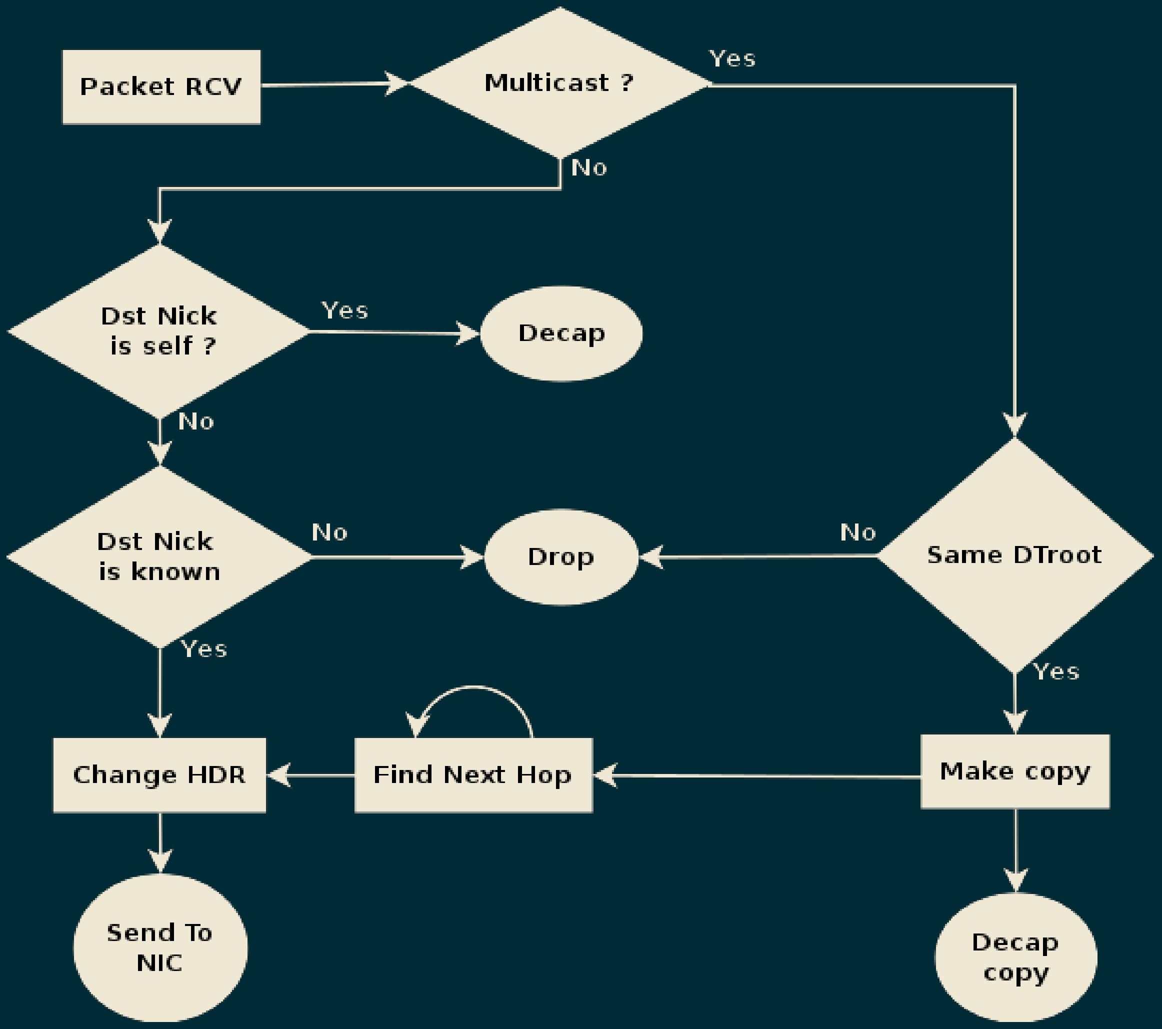
RB2 decapsulates the TRILL frame and forwards it to VM6

	MAC - RB3	MAC - RB2	
	MAC - RB1	MAC - RB3	
	NickName RB2	NickName RB2	
	Nickname RB1	Nickname RB1	
MAC - VM6	MAC - VM6	MAC - VM6	MAC - VM6
MAC - VM1	MAC - VM1	MAC - VM1	MAC - VM1
IP - VM6	IP - VM6	IP - VM6	IP - VM6
IP - VM1	IP - VM1	IP - VM1	IP - VM1
DATA	DATA	DATA	DATA

IMPLEMENTATION - SENDING



IMPLEMENTATION - RECEIVING



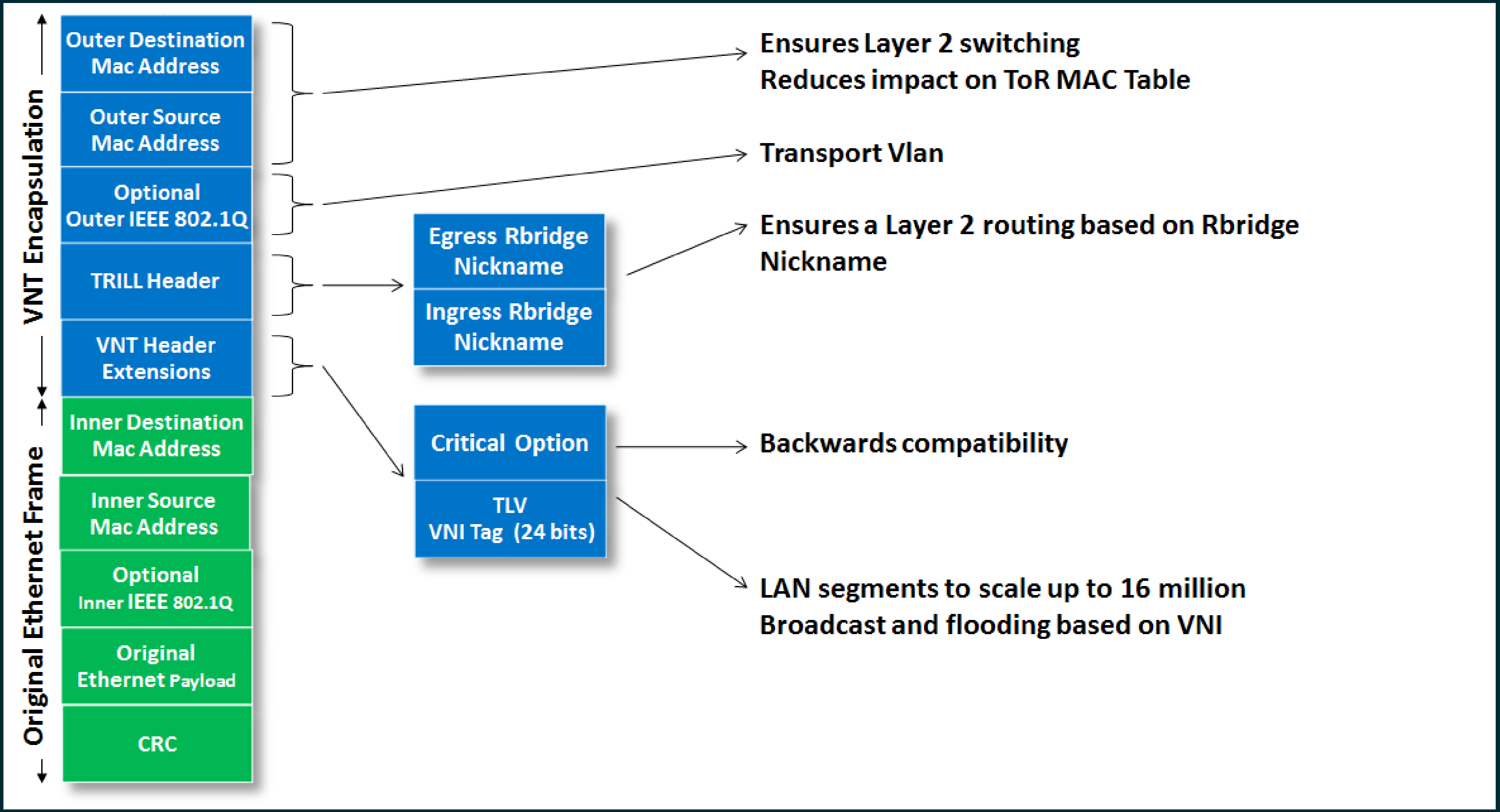
LAYER 2 - SWITCHING LIMITATION

- ~~A large number of tenants implies~~
  - ~~a huge number of MAC address in switch table~~
  - ~~ARP storm at nodes~~
- ~~STP to ensure a loop free topology~~
  - ~~blocking redundant paths~~
  - ~~Core-computes required, recomputes when topology changes~~
- Number of VLANs is limited to 4096

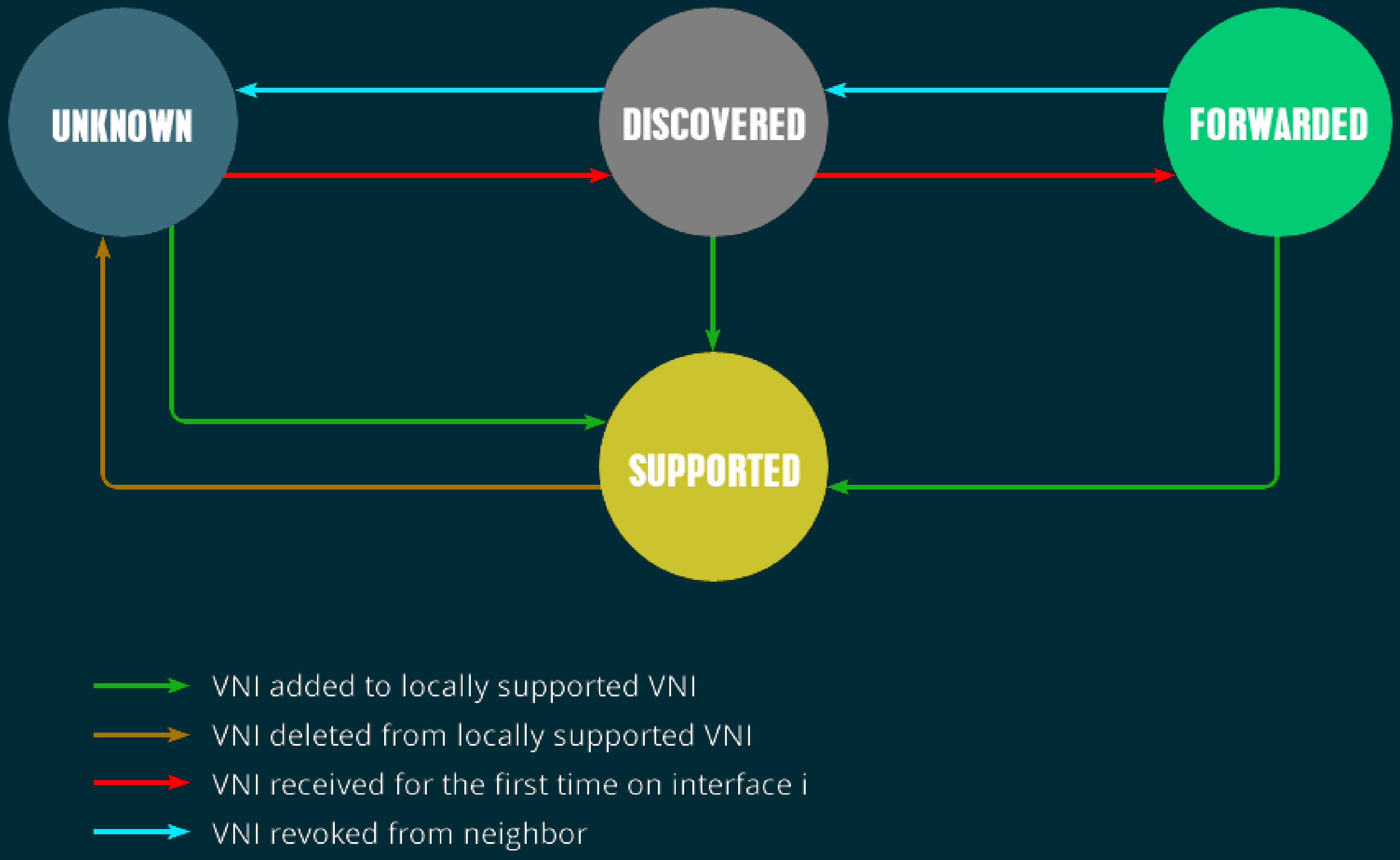
TRILL + VNI = VNT

Virtual Network over TRILL

VNT FRAME FORMAT


VNT Encapsulation
Original Ethernet Frame
Outer Destination Mac Address
Outer Source Mac Address
Optional Outer IEEE 802.1Q
TRILL Header
VNT Header Extensions
Inner Destination Mac Address
Inner Source Mac Address
Optional Inner IEEE 802.1Q
Original Ethernet Payload
CRC
Egress Rbridge Nickname
Ingress Rbridge Nickname
Critical Option
TLV
VNI Tag (24 bits)
Ensures Layer 2 switching
Reduces impact on ToR MAC Table
Transport Vlan
Ensures a Layer 2 routing based on Rbridge Nickname
Backwards compatibility
LAN segments to scale up to 16 million
Broadcast and flooding based on VNI

VNI LIFE



🟩 → VNI added to locally supported VNI

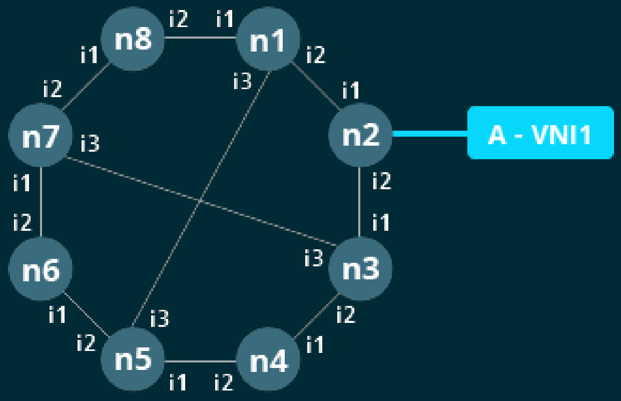
🟫 → VNI deleted from locally supported VNI

🟥 → VNI received for the first time on interface i

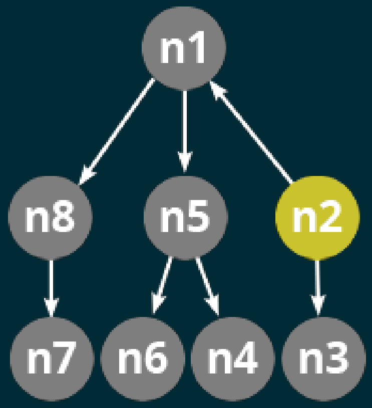
🟦 → VNI revoked from neighbor

VNI TOPOLOGY BUILDING

TOPOLOGY



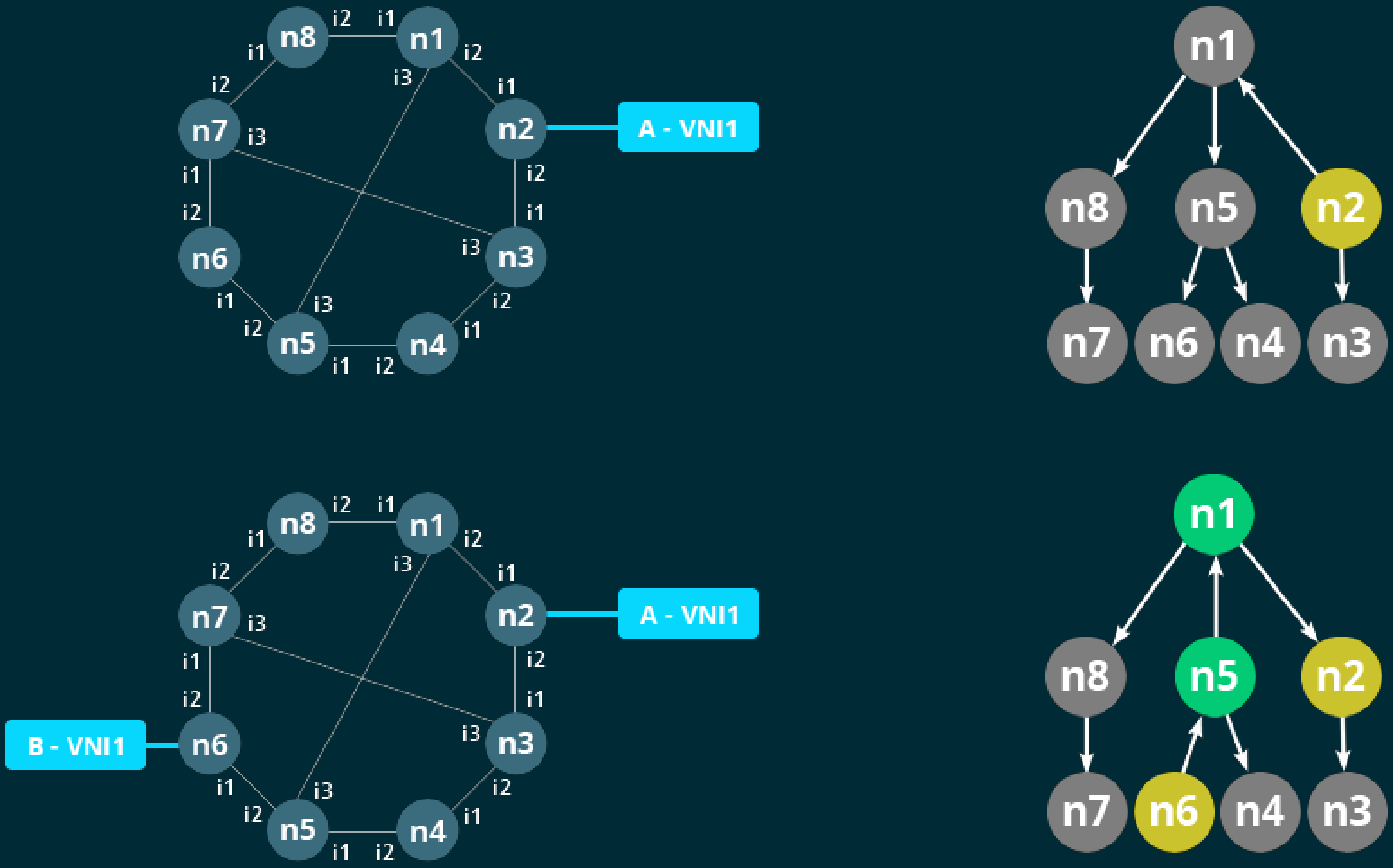
MULTICAST TREE



VNI TOPOLOGY BUILDING

TOPOLOGY

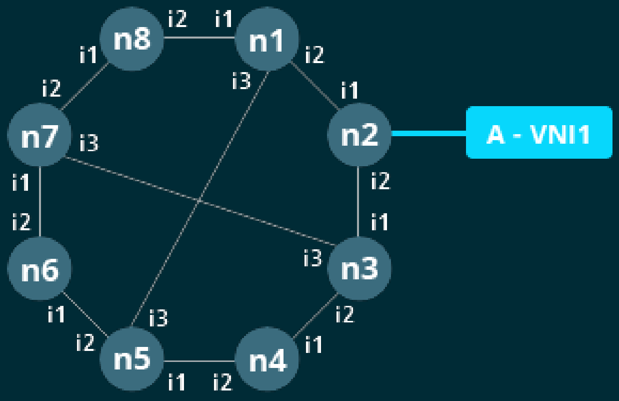
MULTICAST TREE

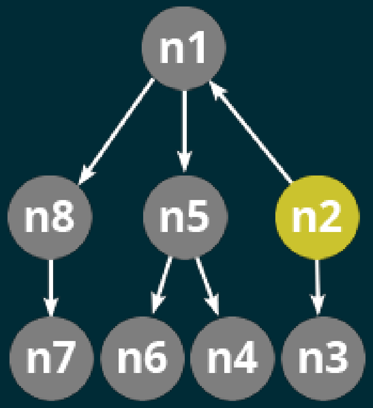


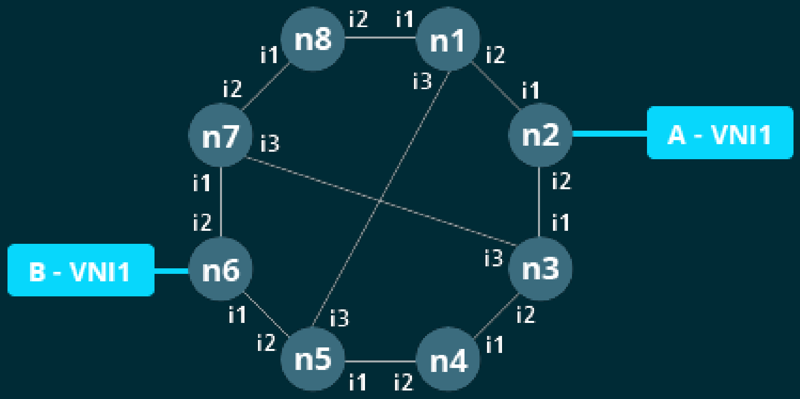
VNI TOPOLOGY BUILDING

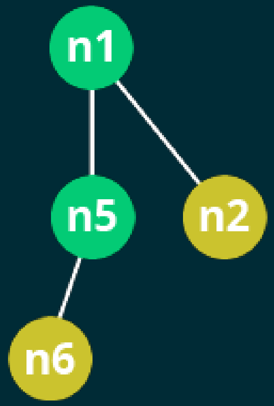
TOPOLOGY

MULTICAST TREE

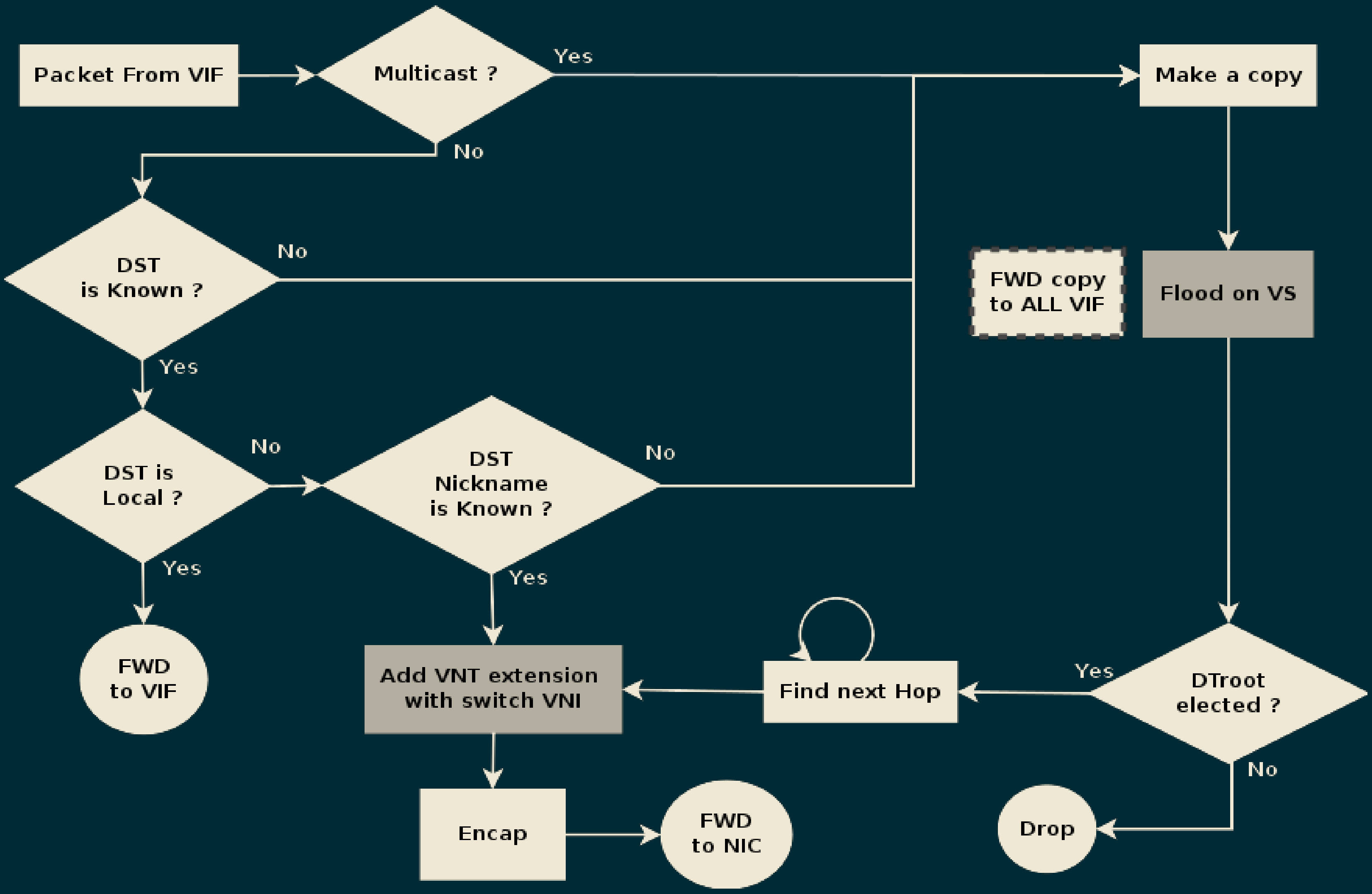




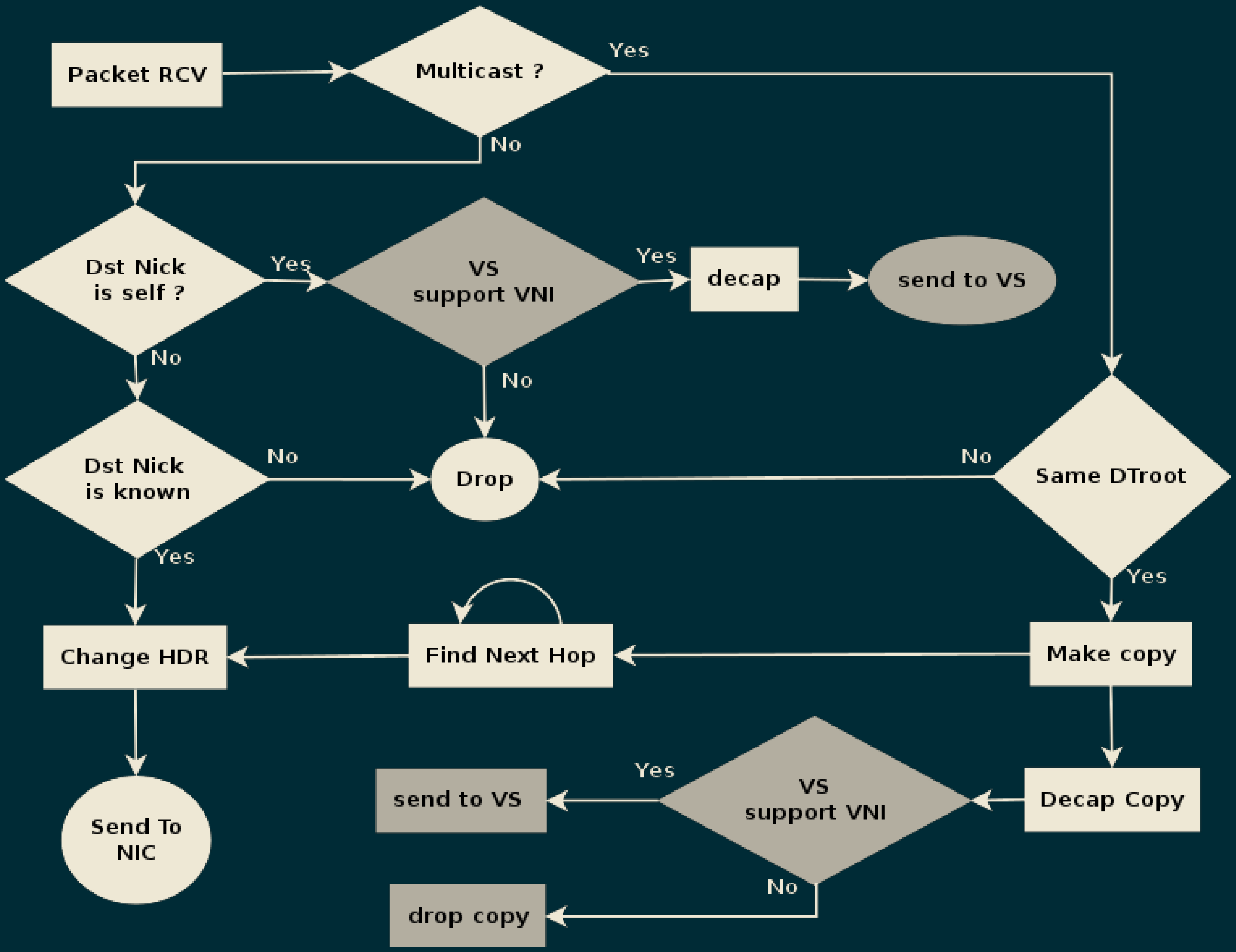




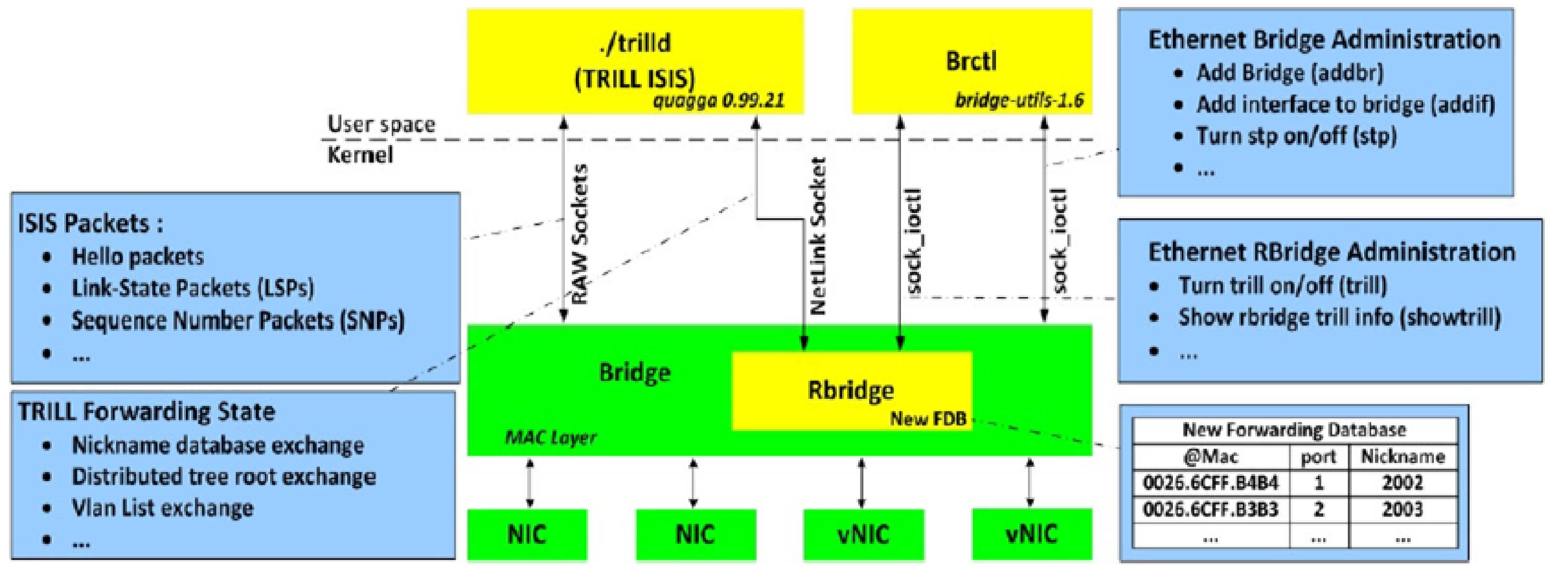
IMPLEMENTATION WITH VNI - SENDING



IMPLEMENTATION WITH VNI - RECEIVING



LINUX BIG PICTURE



ISIS Packets :

- Hello packets
- Link-State Packets (LSPs)
- Sequence Number Packets (SNPs)
- ...

TRILL Forwarding State

- Nickname database exchange
- Distributed tree root exchange
- Vlan List exchange
- ...

Ethernet Bridge Administration

- Add Bridge (addbr)
- Add interface to bridge (addif)
- Turn stp on/off (stp)
- ...

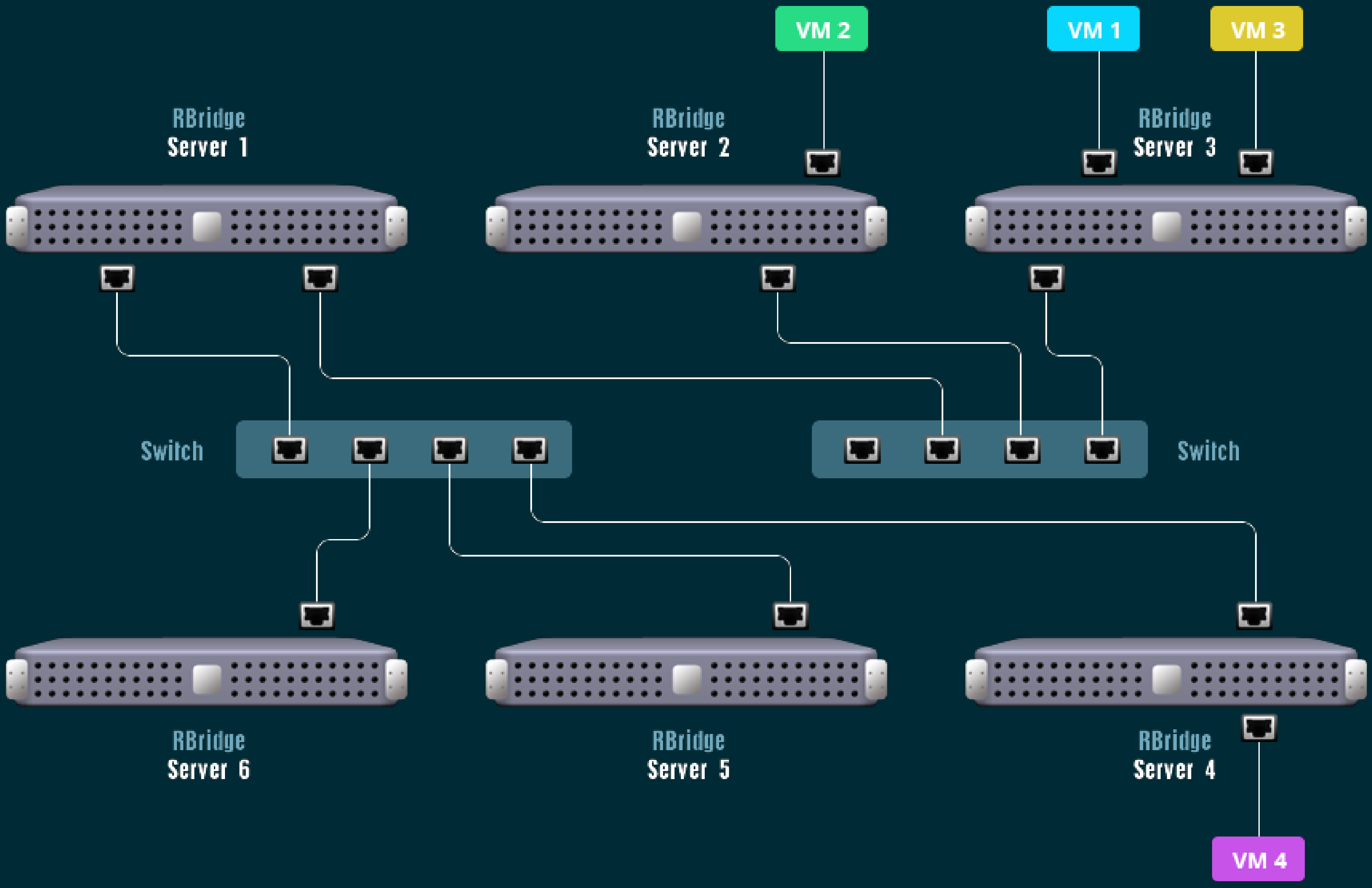
Ethernet RBridge Administration

- Turn trill on/off (trill)
- Show rbridge trill info (showtrill)
- ...

New Forwarding Database

@Mac	port	Nickname
0026.6CFF.B4B4	1	2002
0026.6CFF.B3B3	2	2003
...	...	...

DEMONSTRATION



SCREENCAST

screencast

(live explanation to understand what's going on)

PH.D. STUDY

Ahmed Amamou - **ahmed@gandi.net**

"Network isolation for Virtualized Datacenters"

University Pierre & Marie Curie - GANDI SAS

project still in development and cleaning

TRILL sources: **github.com/Gandi/ktrill**

VNT: still two research projects working on it - drafts

GANDI.NET

Gandi Hosting - gandi.net/hosting

William Dauchy - william@gandi.net

slides pres.gandi.net/kr2013