ARISTA

Scalable Network Designs for HBASE

- Scalable Network Design
- Some Measurements
- How the Network Can Help

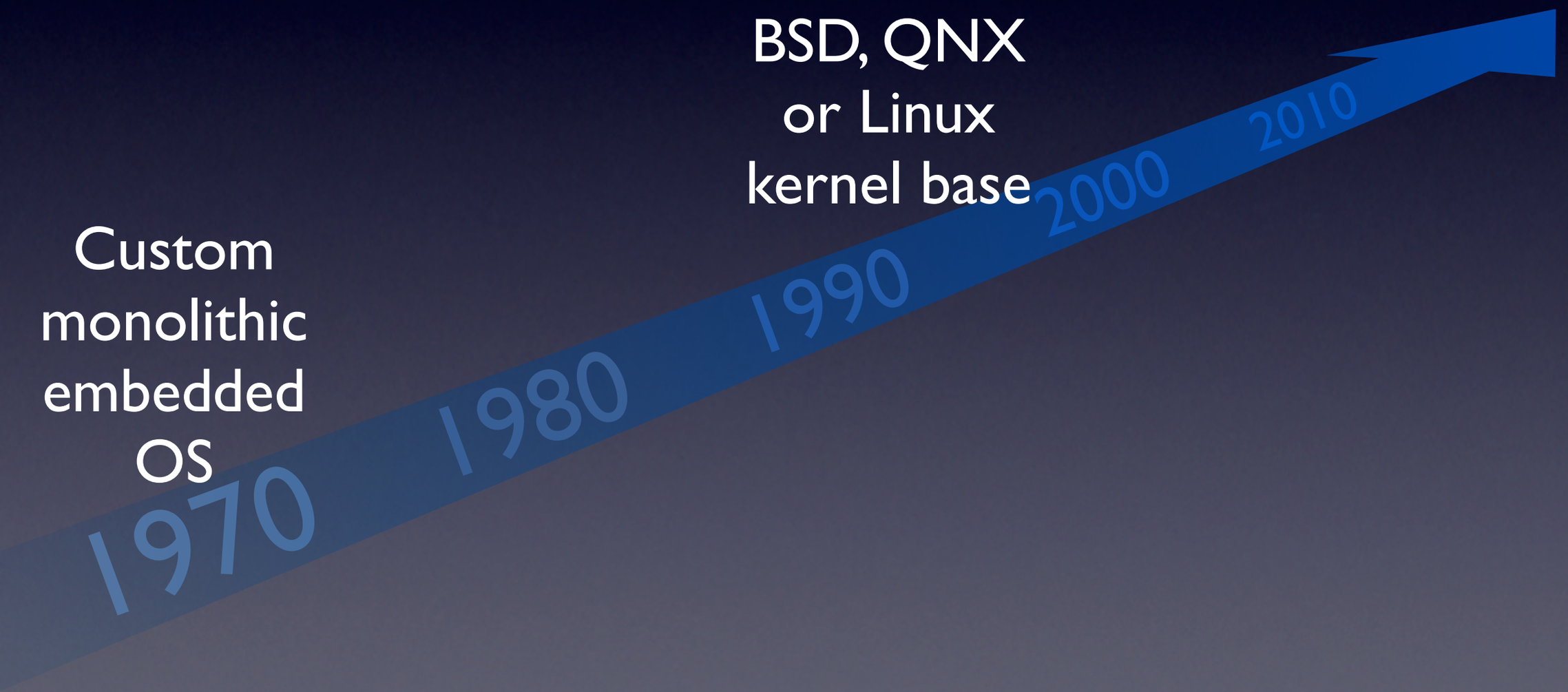
Benoît “tsuna” Sigoure
Member of the Yak Shaving Staff
tsuna@aristanetworks.com

@tsunanet

A Brief History of Network Software

Modified
BSD, QNX
or Linux
kernel base

Custom
monolithic
embedded
OS

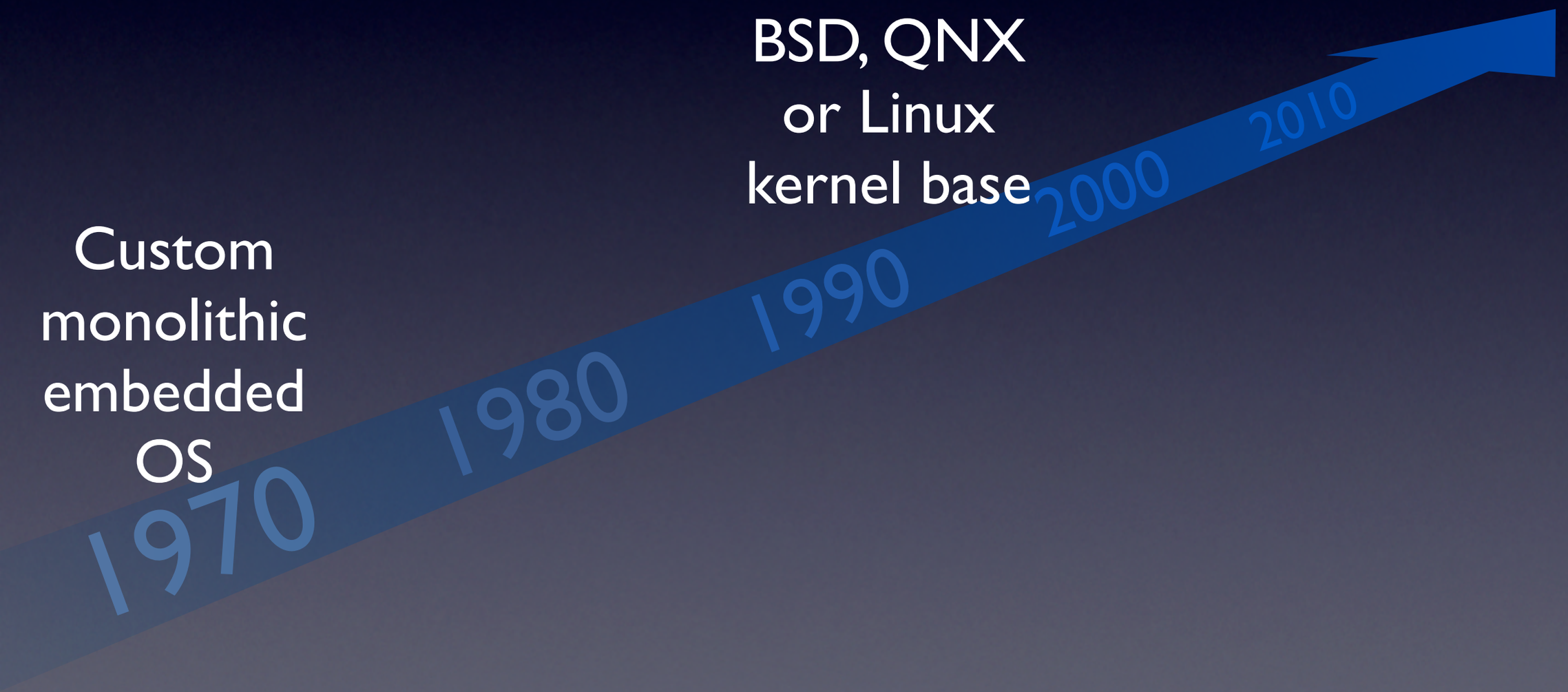


A Brief History of Network Software

Custom
monolithic
embedded
OS

Modified
BSD, QNX
or Linux
kernel base

"The switch
is just a
Linux box"



A Brief History of Network Software

1970

Custom monolithic embedded OS

1980

1990

Modified BSD, QNX or Linux kernel base

2000

2010

"The switch is just a Linux box"

EOS = Linux

A Brief History of Network Software

1970

Custom monolithic embedded OS

1980

Modified BSD, QNX or Linux kernel base

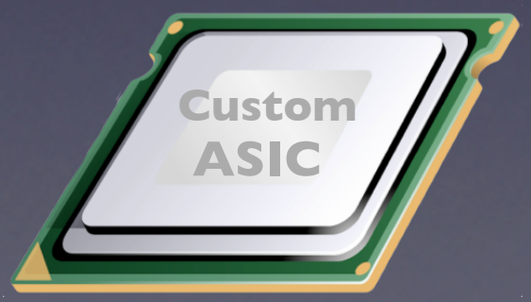
1990

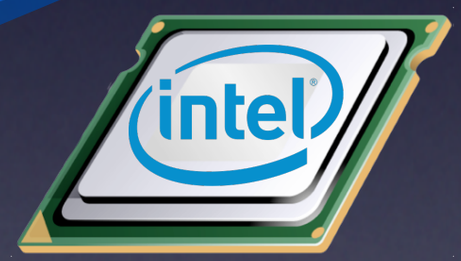
2000

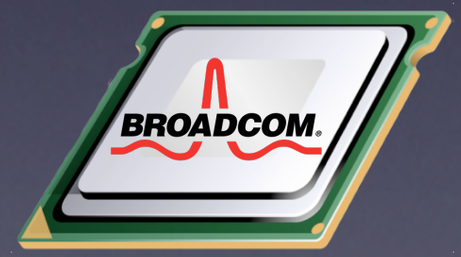
2010

"The switch is just a Linux box"

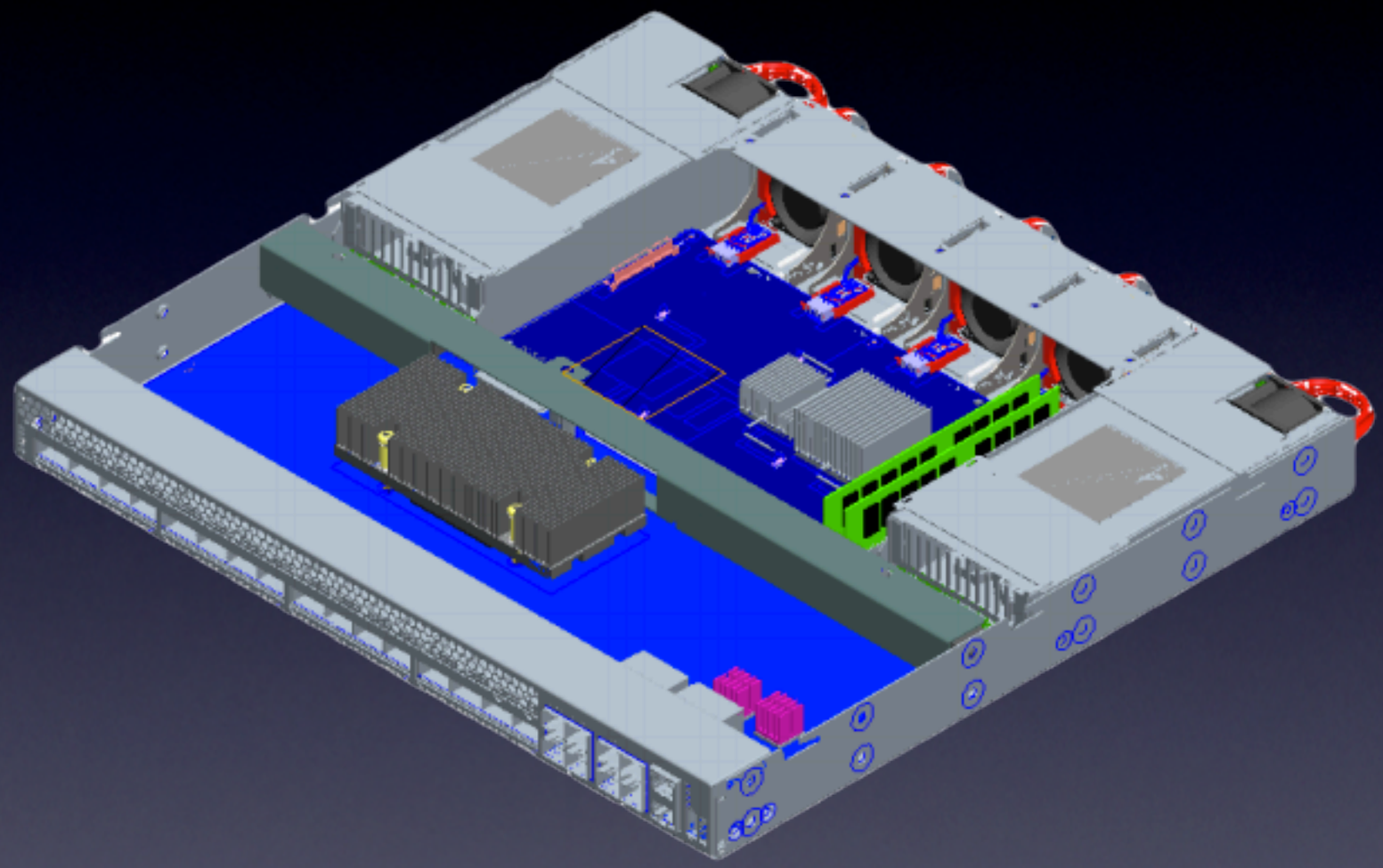
EOS = Linux





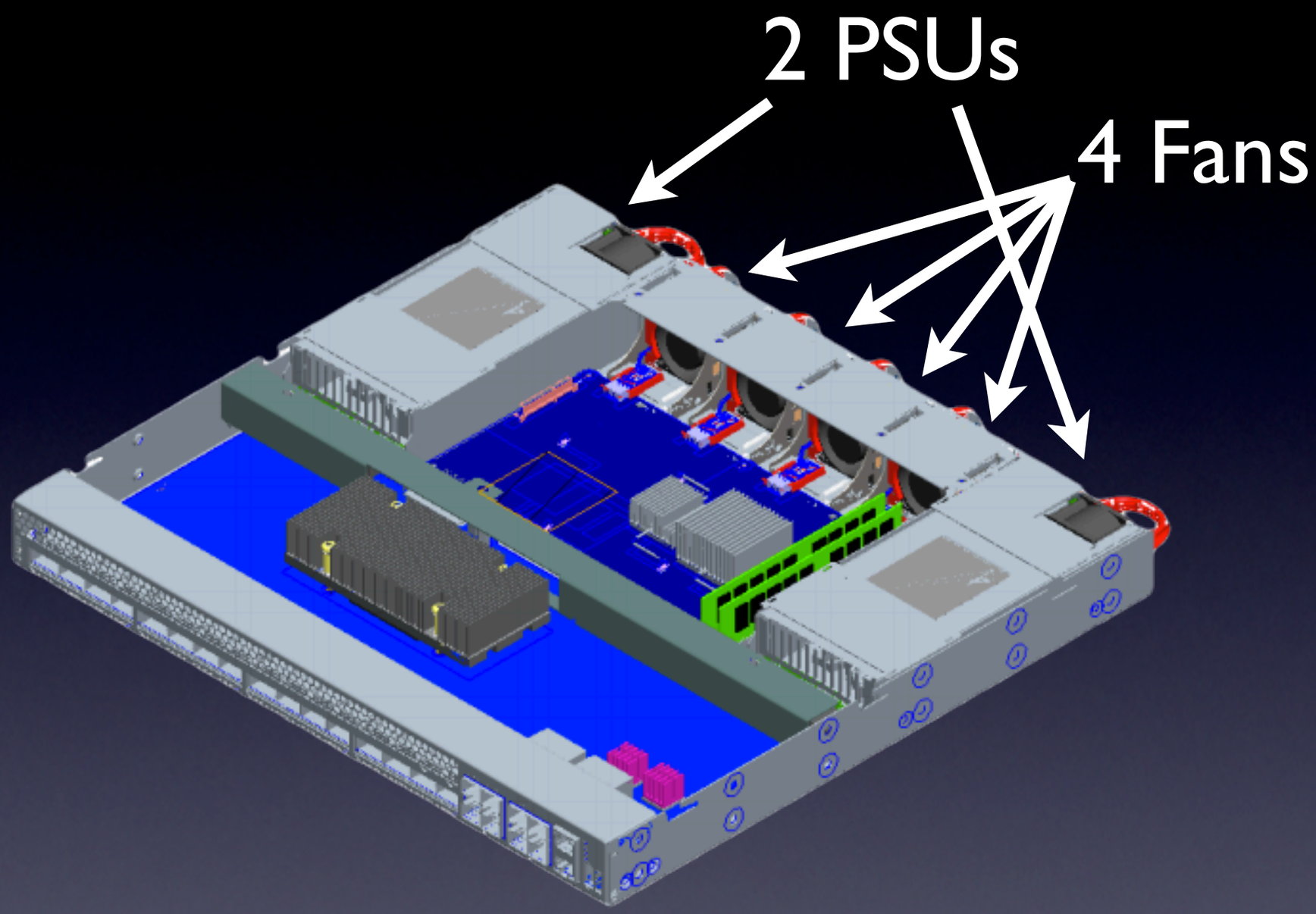


Inside a Modern Switch



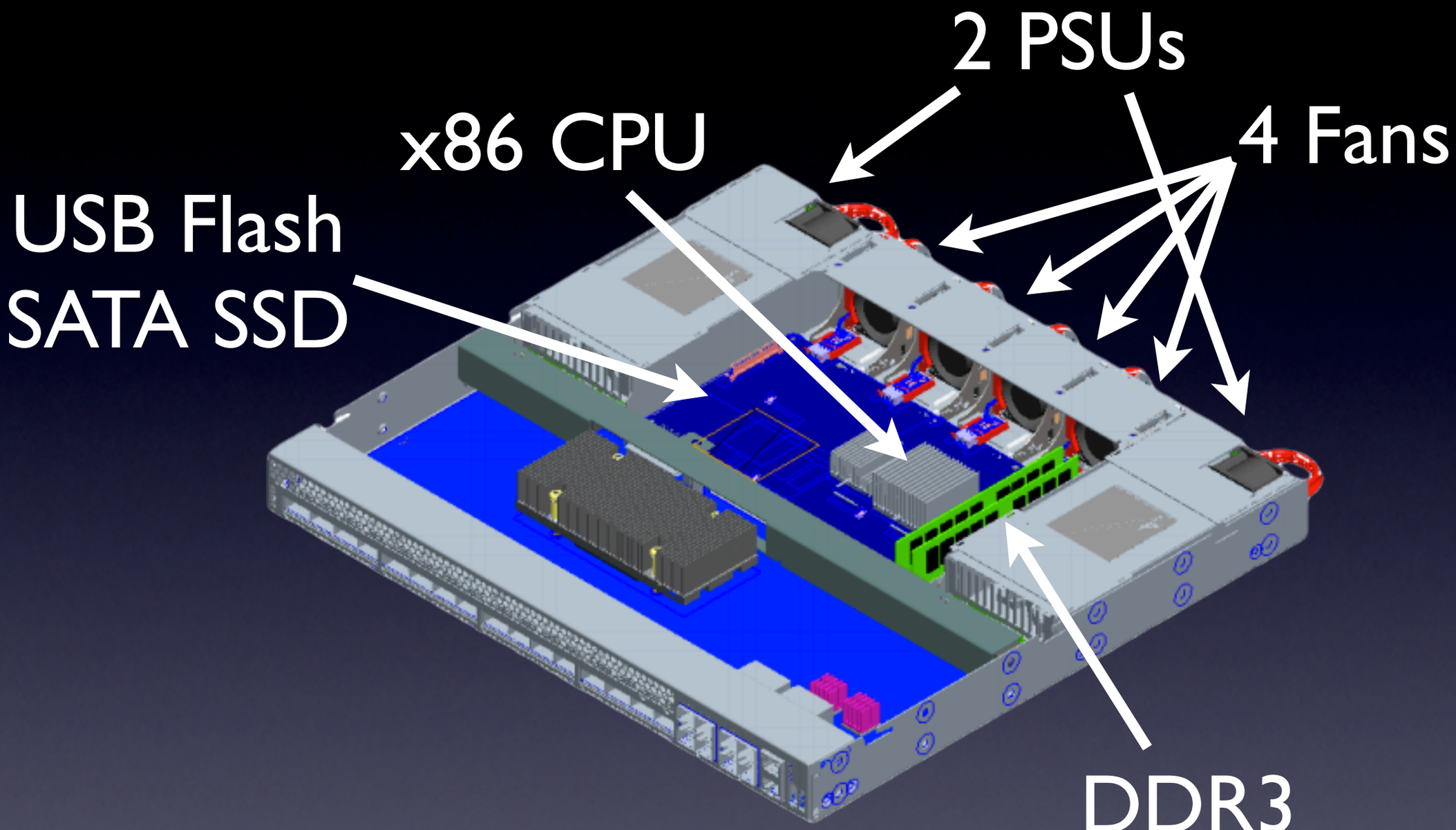
Arista 7050Q-64

Inside a Modern Switch



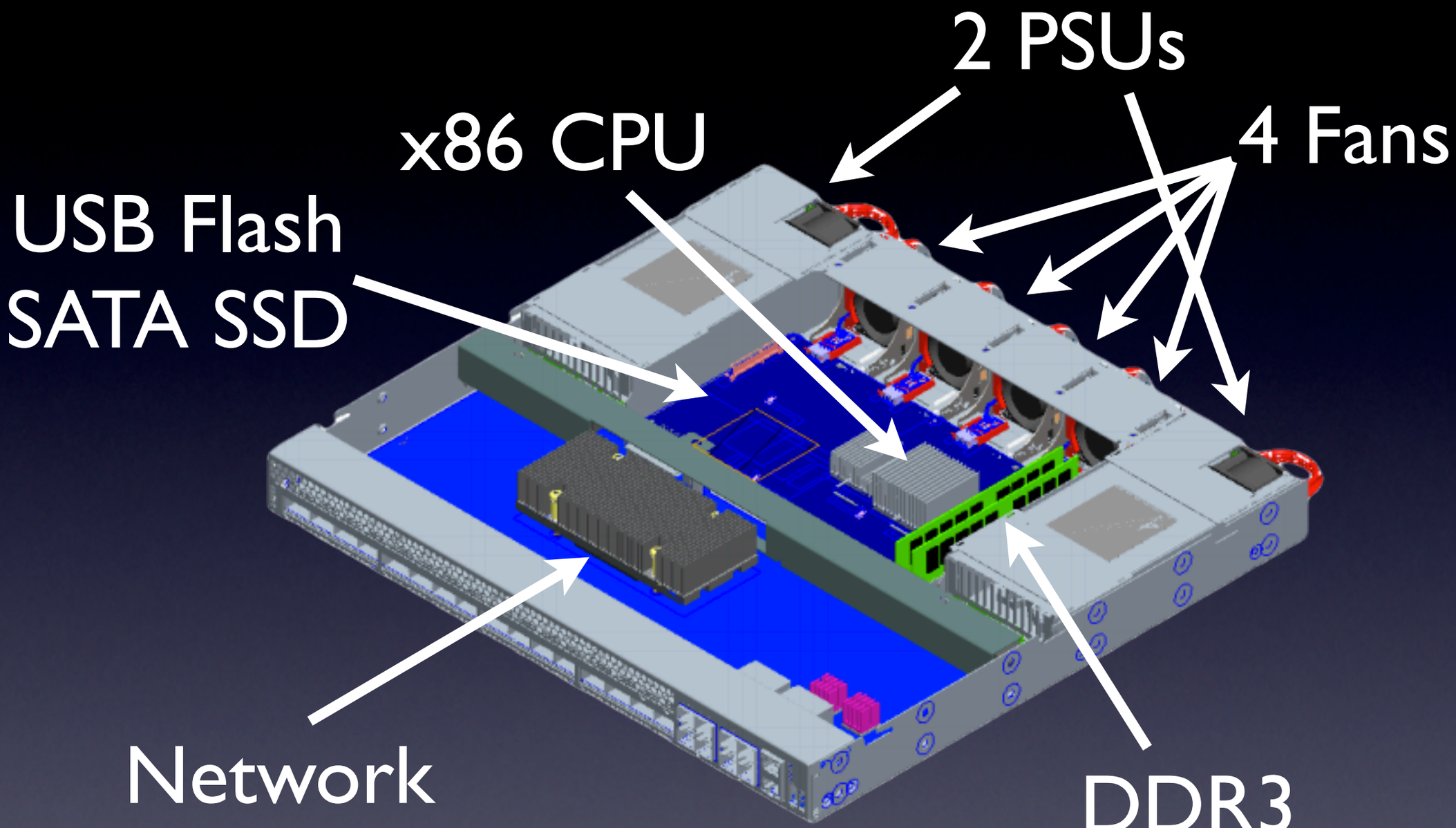
Arista 7050Q-64

Inside a Modern Switch



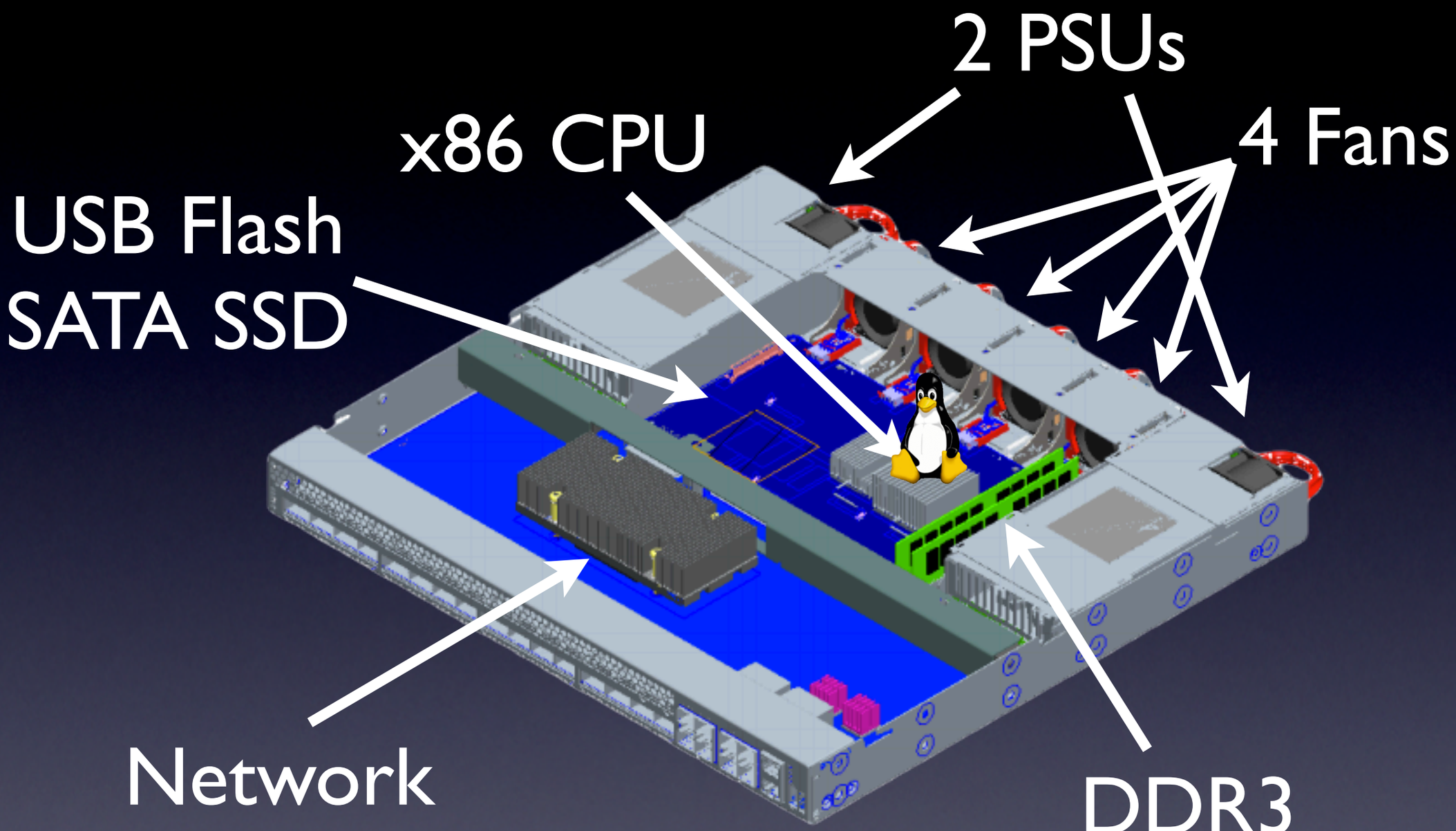
Arista 7050Q-64

Inside a Modern Switch



Arista 7050Q-64

Inside a Modern Switch

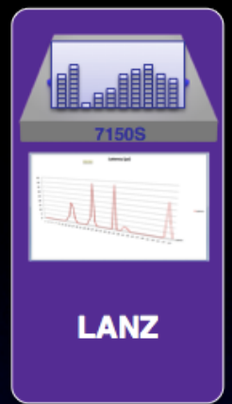


Arista 7050Q-64

It’s All About The Software

It's All About The Software

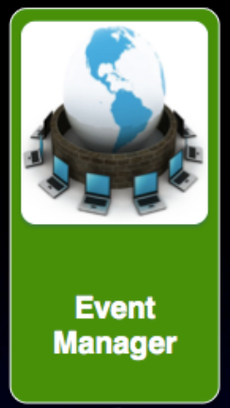
Time Stamping



LANZ

Advanced Mirroring

CLI Scheduler

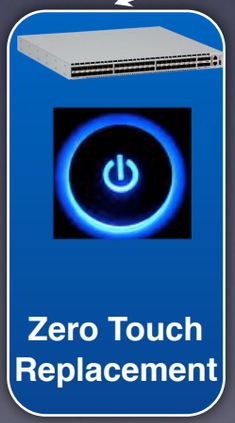


Event Manager

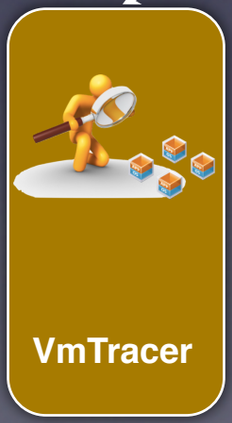
Command API

Event Monitor

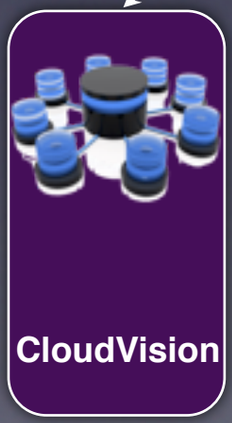
OpenTSDB



Zero Touch Replacement



VmTracer

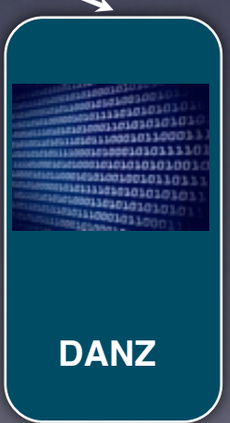


CloudVision

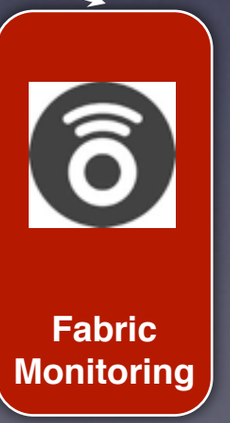
Email

Wireshark

Fast Failover



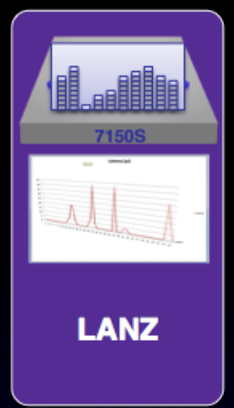
DANZ



Fabric Monitoring

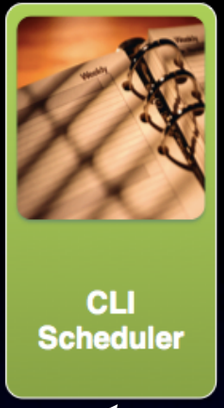
It's All About The Software

Time Stamping

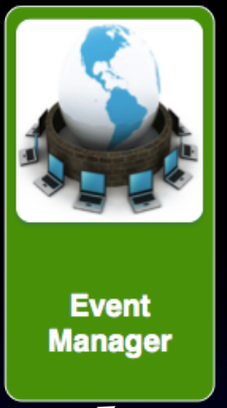


LANZ

Advanced Mirroring



CLI Scheduler



Event Manager

Command API

Event Monitor

python

#!/bin/bash

puppet labs

Chef

NGINX

Ganglia

OpenTSDB

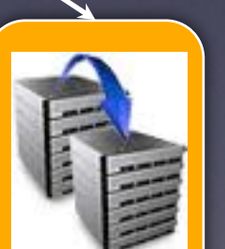
Zero Touch Replacement

VmTracer

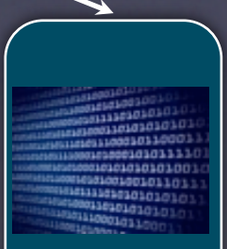
CloudVision

Email

Wireshark



Fast Failover

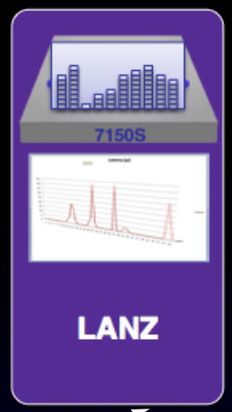


DANZ

Fabric Monitoring

It’s All About The Software

Time Stamping



LANZ

Advanced Mirroring

CLI Scheduler

Event Manager

Command API

Event Monitor

python

#!/bin/bash

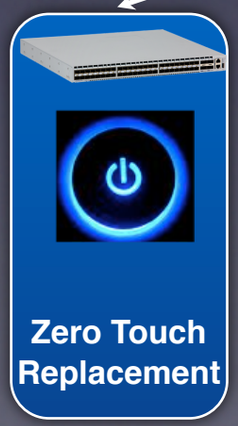
puppet labs

Chef

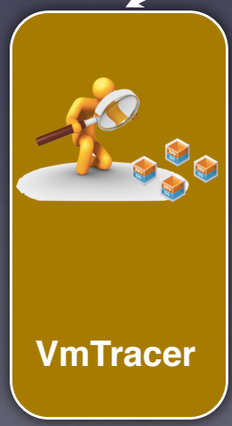
NGINX

OpenTSDB

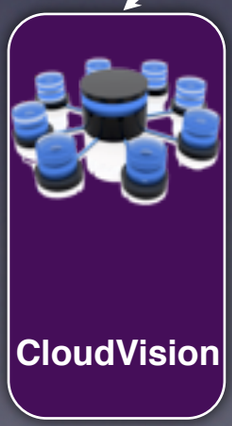
$ sudo yum install <pkg>



Zero Touch Replacement

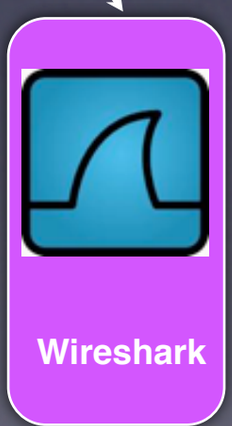


VmTracer



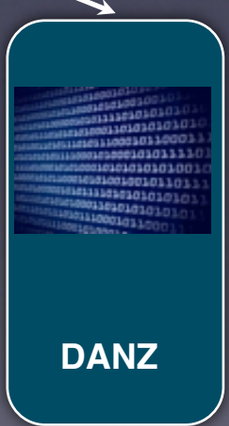
CloudVision

Email

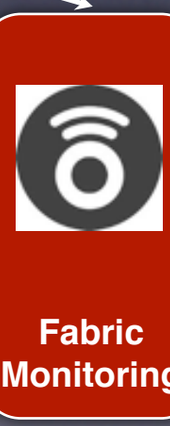


Wireshark

Fast Failover

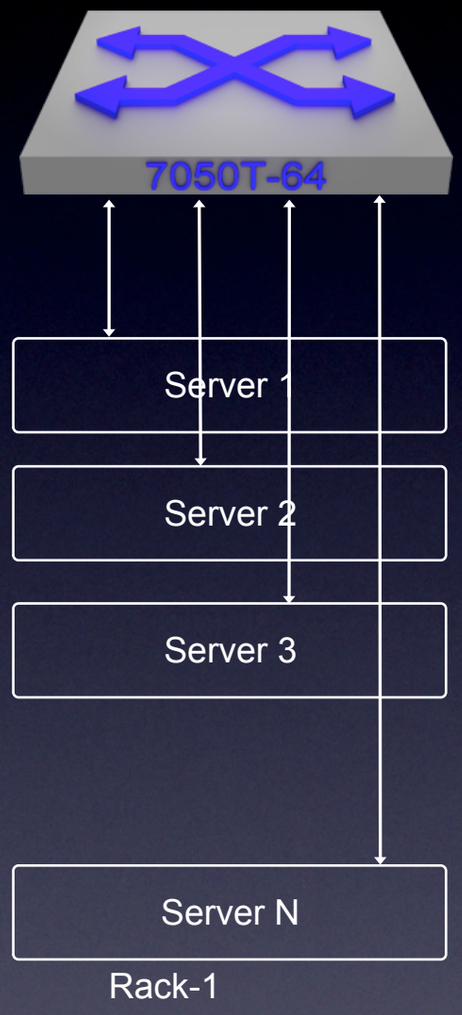


DANZ



Fabric Monitoring

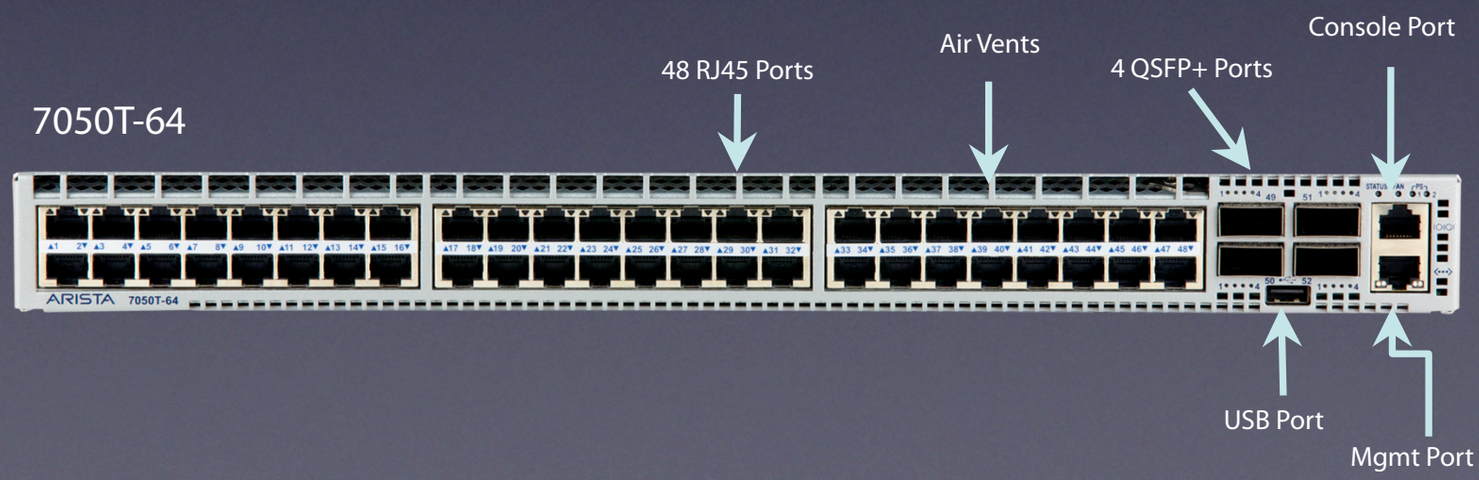
Typical Design: Keep It Simple



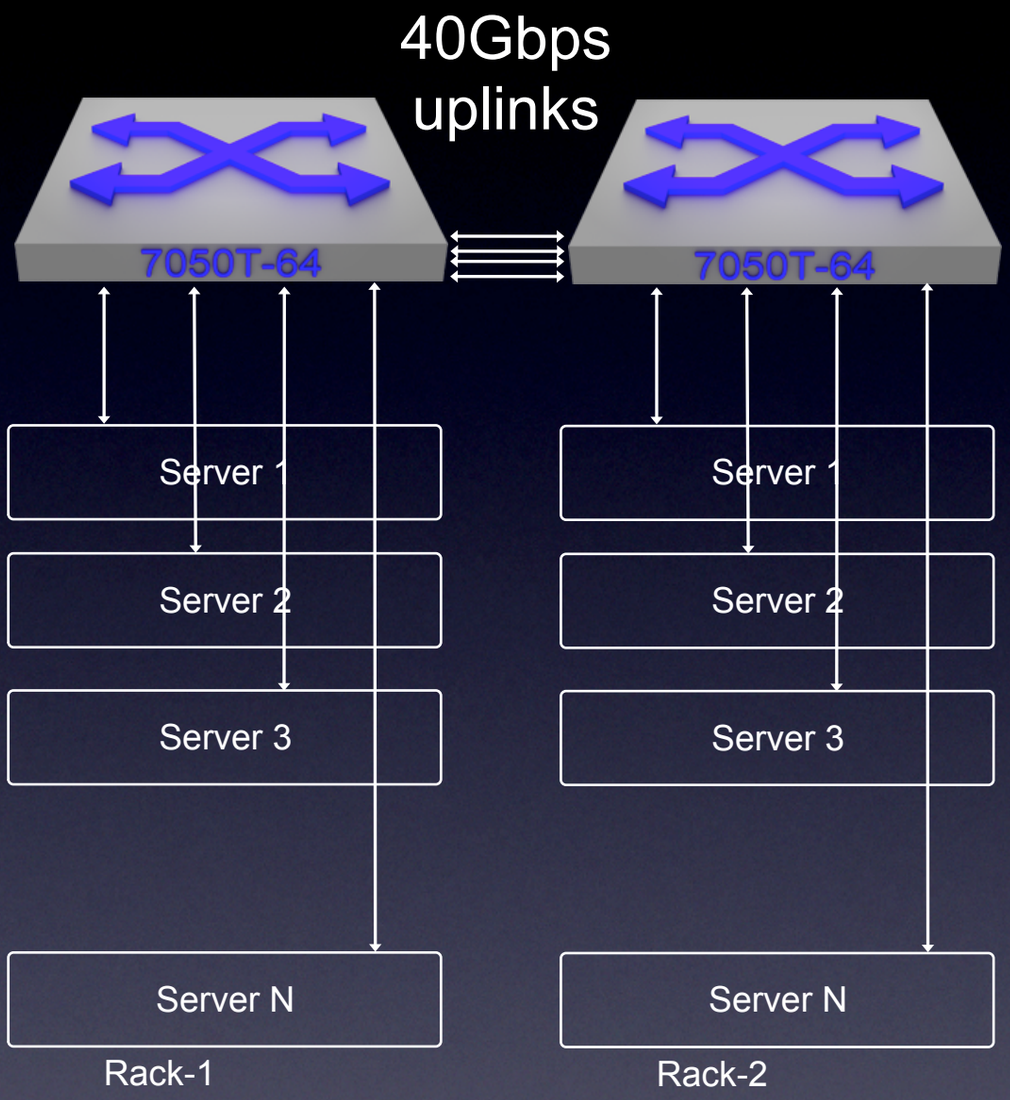
Start small: single rack, single switch

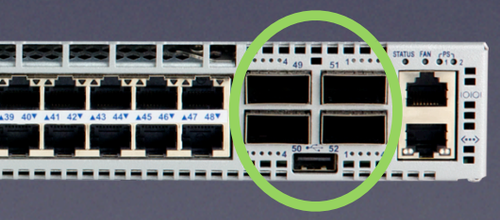
Scale: 64 nodes max

10Gbps network, line-rate



Typical Design: Keep It Simple



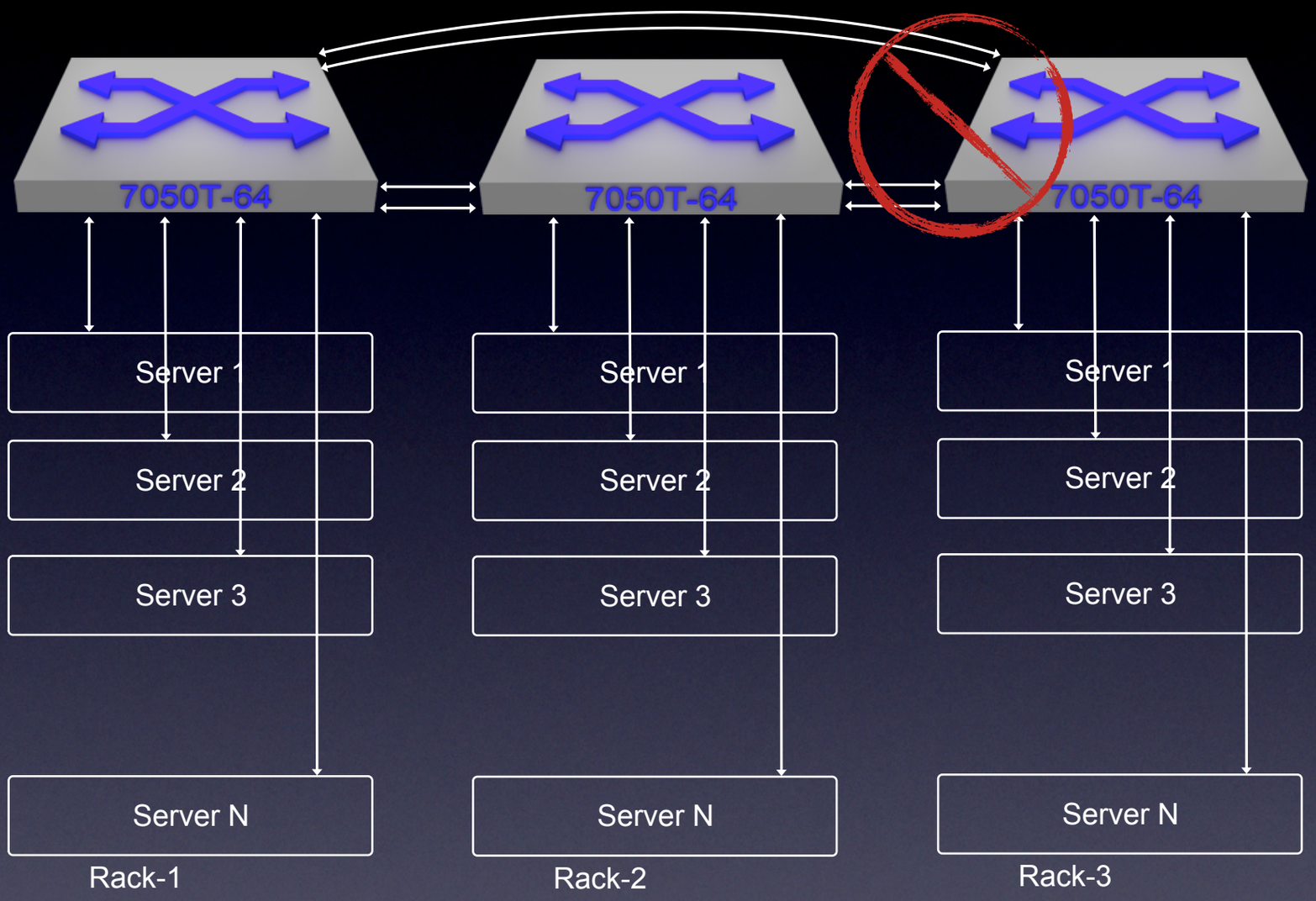


Start small: 2 racks, 2 switches, no core

Scale: 96 nodes max with 4 uplinks

Oversubscription ratio 1:3

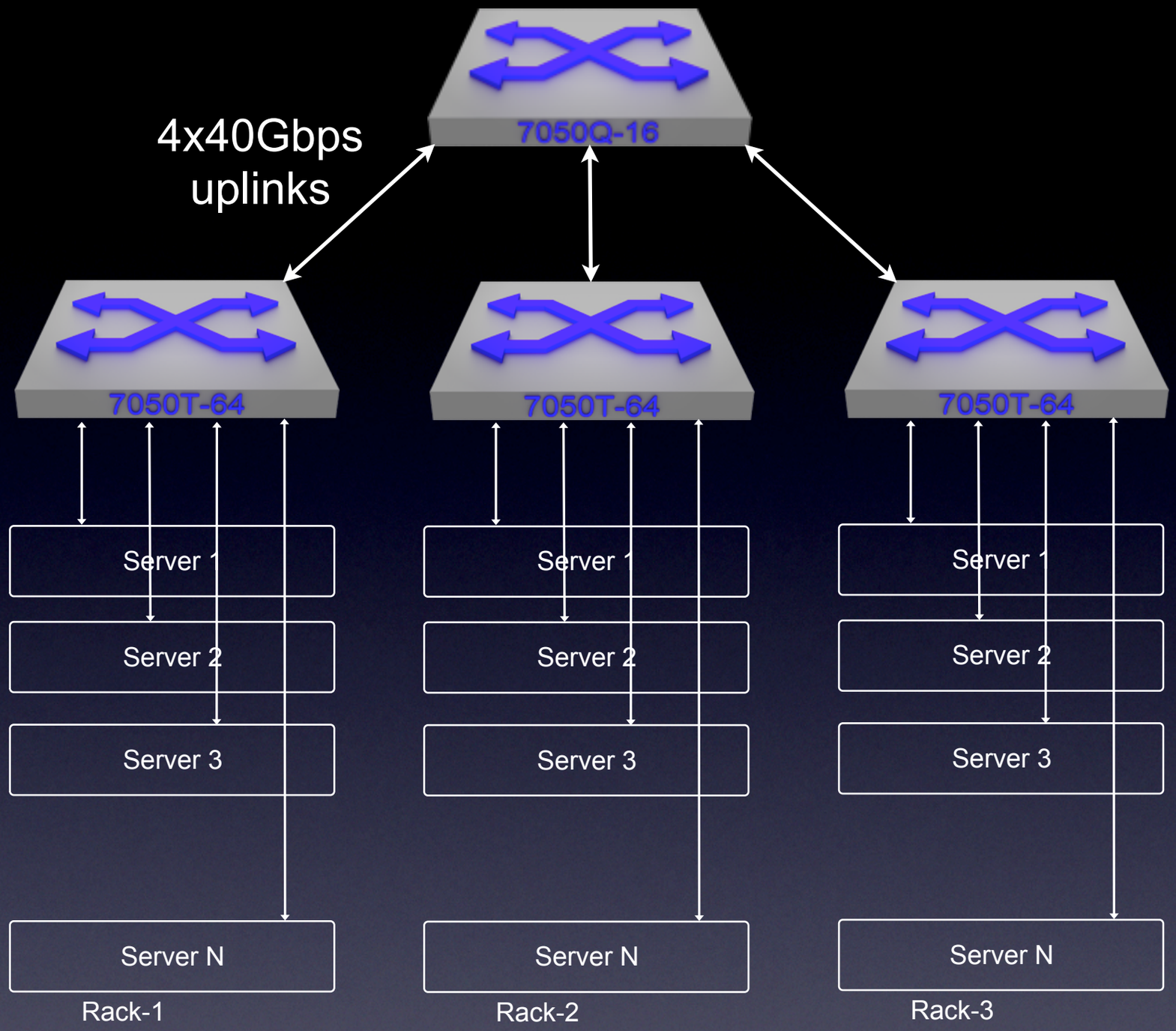
Typical Design: Keep It Simple



3 racks, 3 switches, **no core**

Scale: 144 nodes max with 4 uplinks

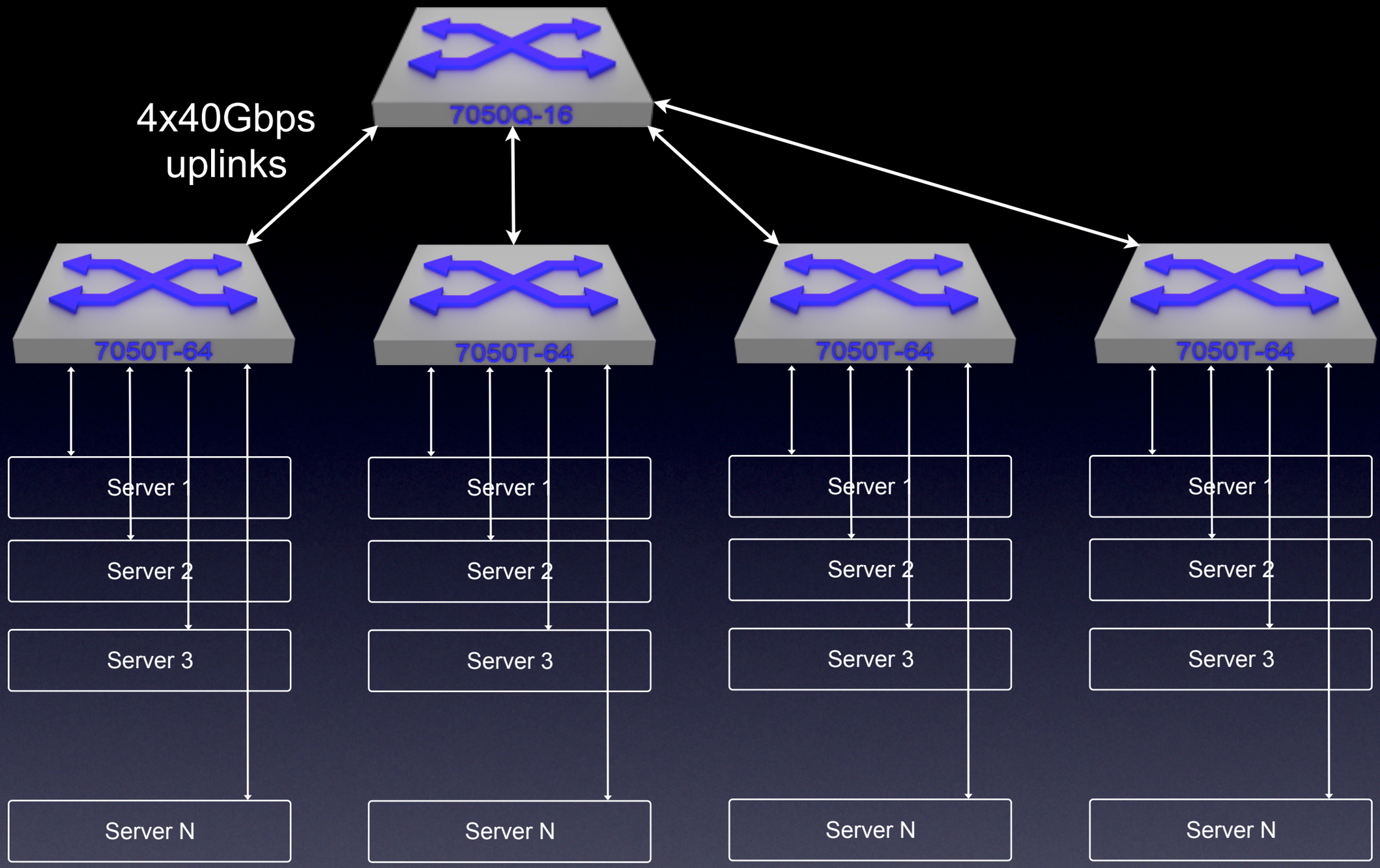
Oversubscription ratio 1:6



3 racks, 4 switches

Scale: 144 nodes max with 4 uplinks/rack

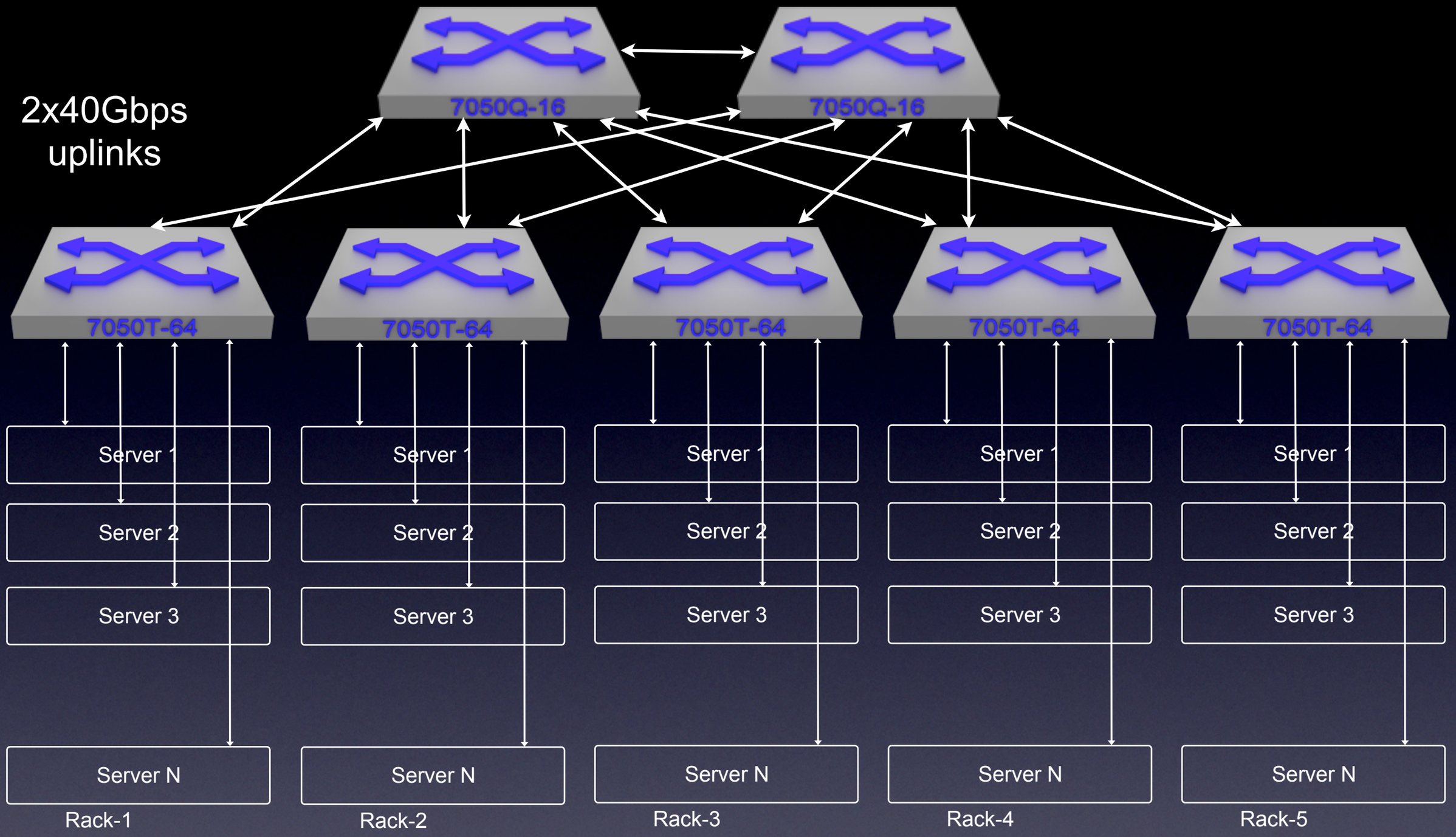
Oversubscription ratio 1:3



4 racks, 5 switches

Scale: 192 nodes max

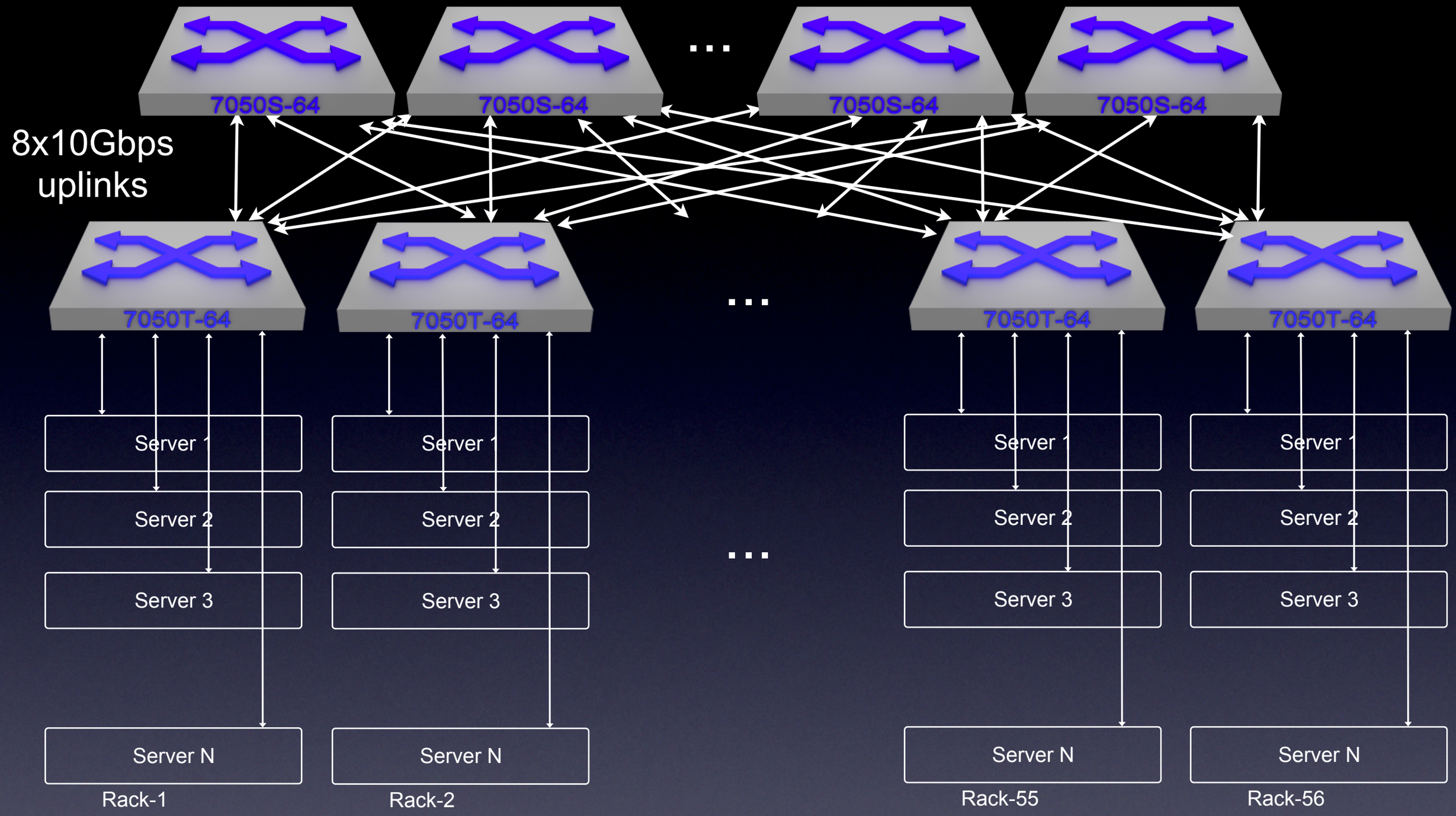
Oversubscription ratio 1:3



5 to 7 racks, 7 to 9 switches

Scale: 240 to 336 nodes

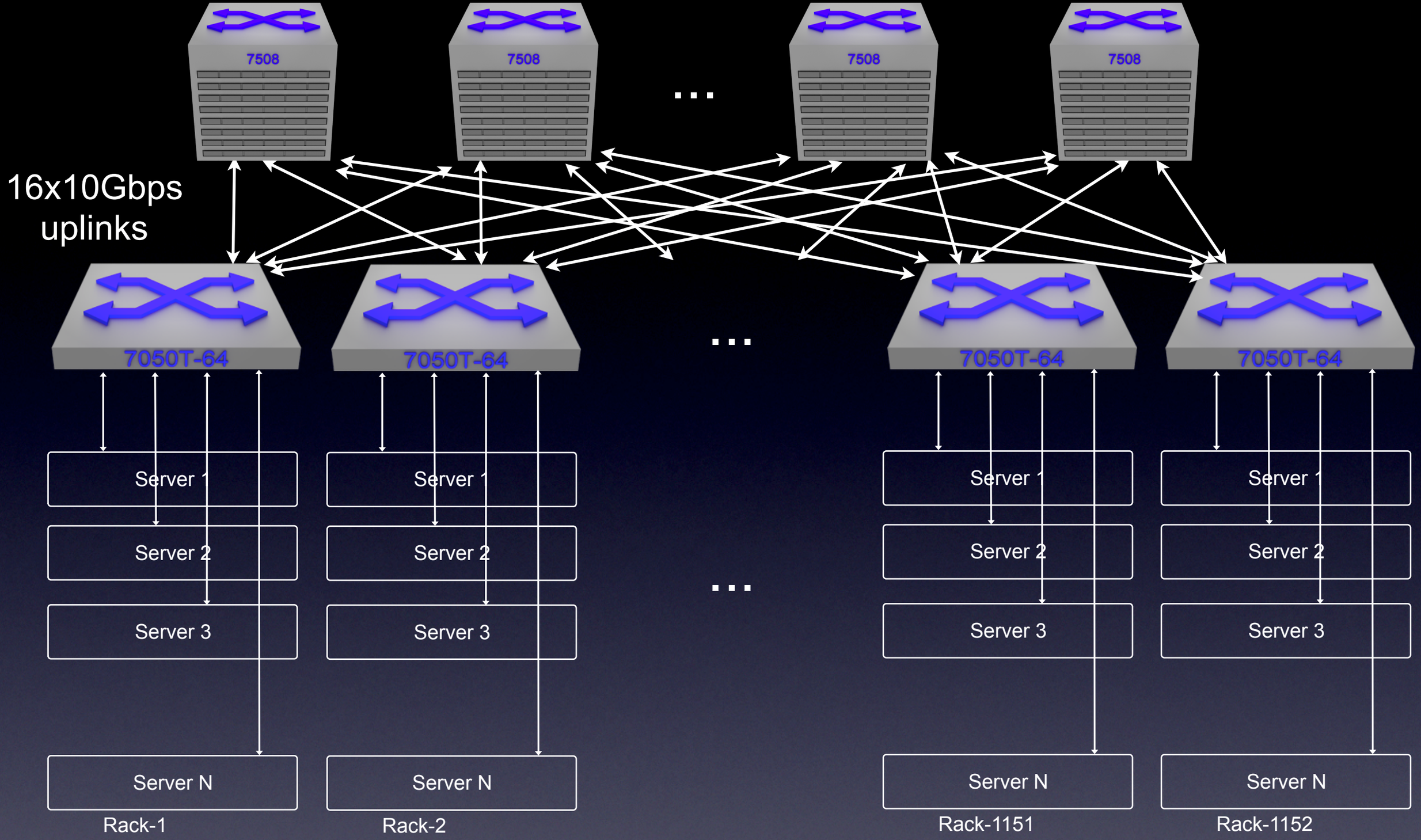
Oversubscription ratio 1:3



56 racks, 8 core switches

Scale: 3136 nodes

Oversubscription ratio 1:3



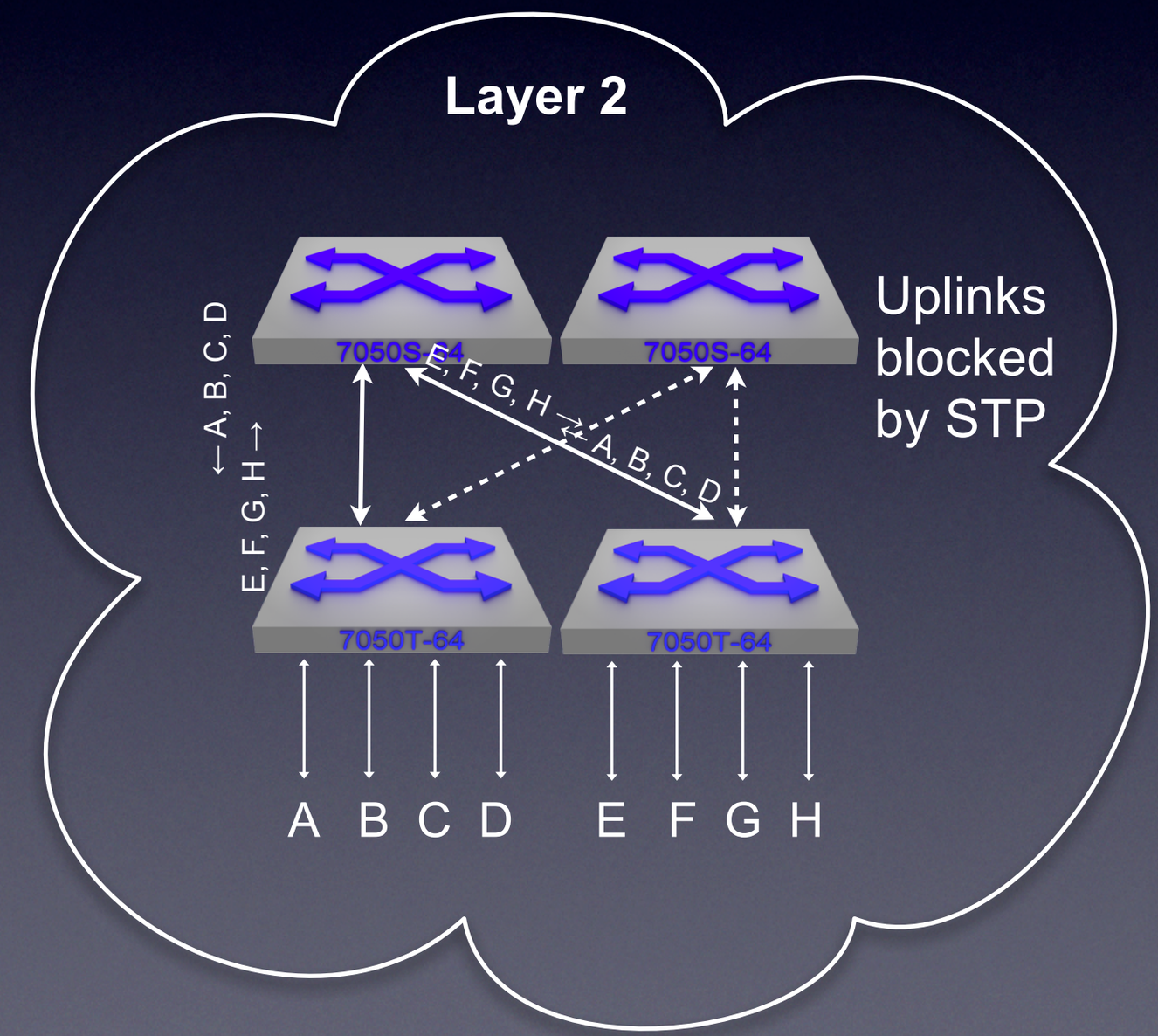
1152 racks, 16 core modular switches

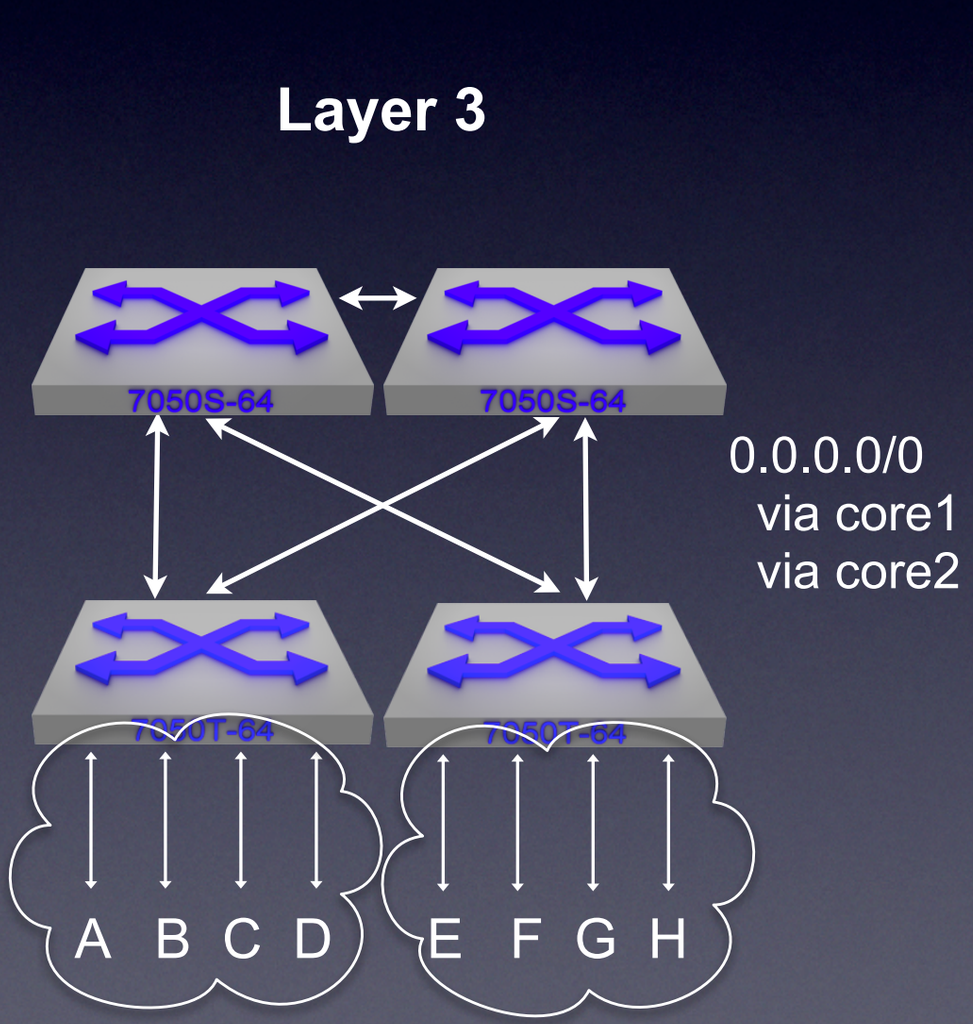
Scale: 55296 nodes

Oversubscription ratio still 1:3

Layer 2 vs Layer 3

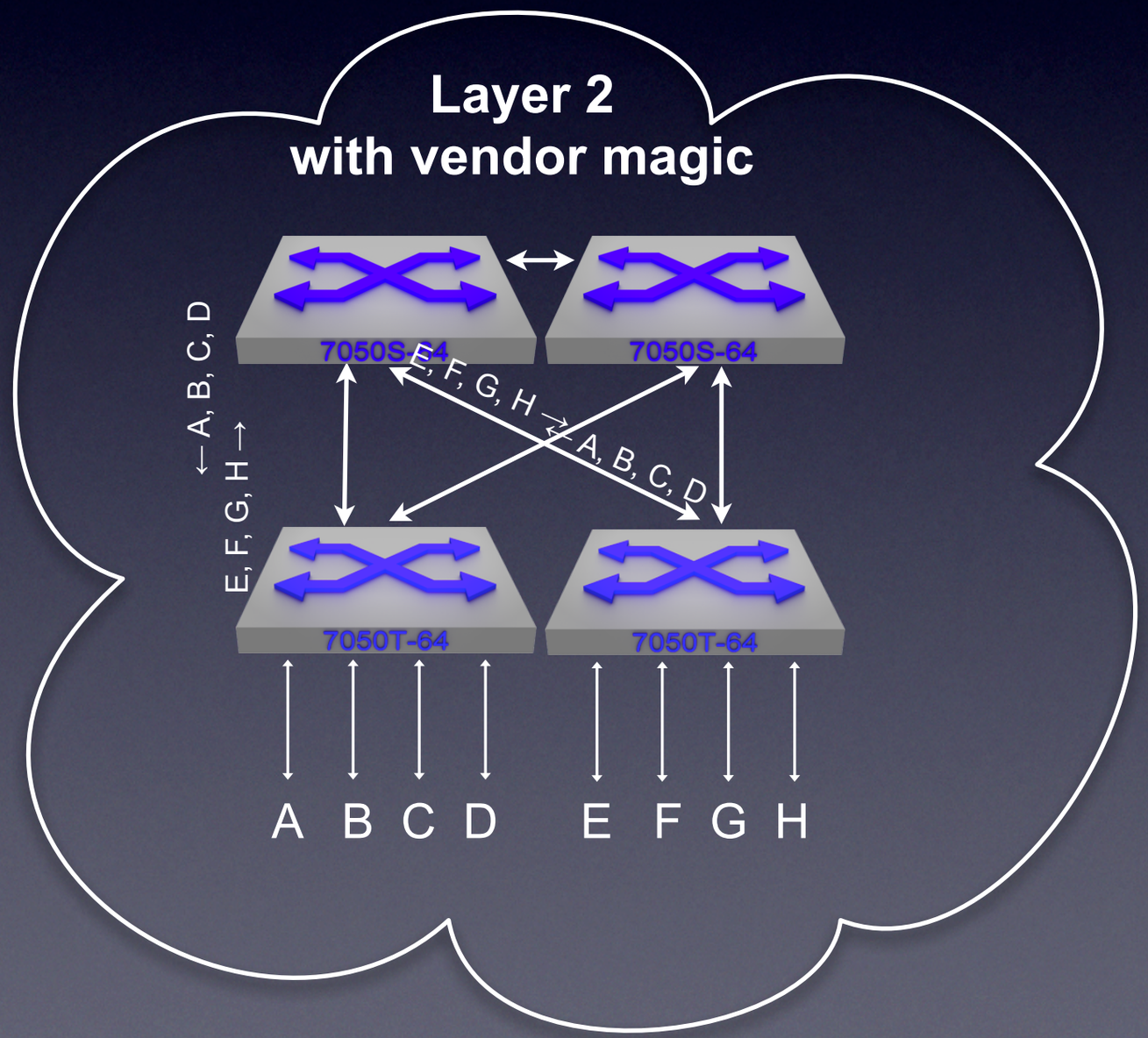
- Layer 2: “data link layer” – Ethernet “bridge”
- Layer 3: “network layer” – IP “router”

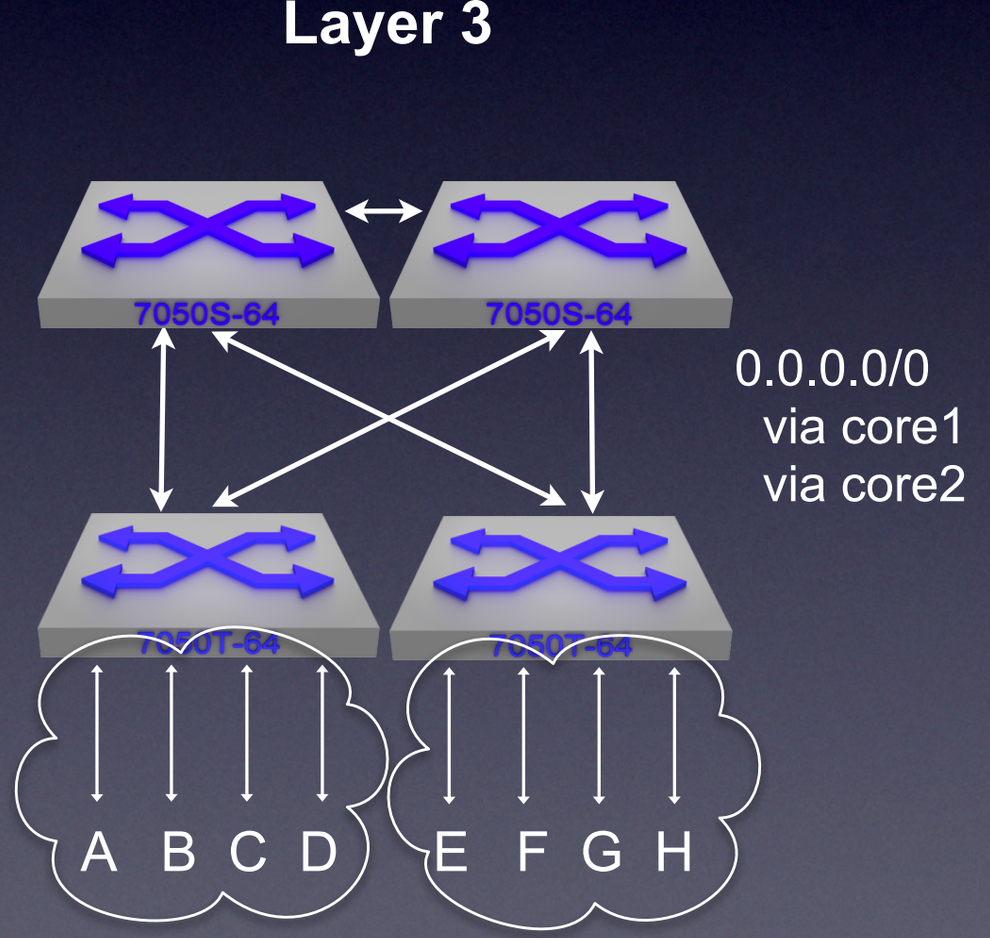




Layer 2 vs Layer 3

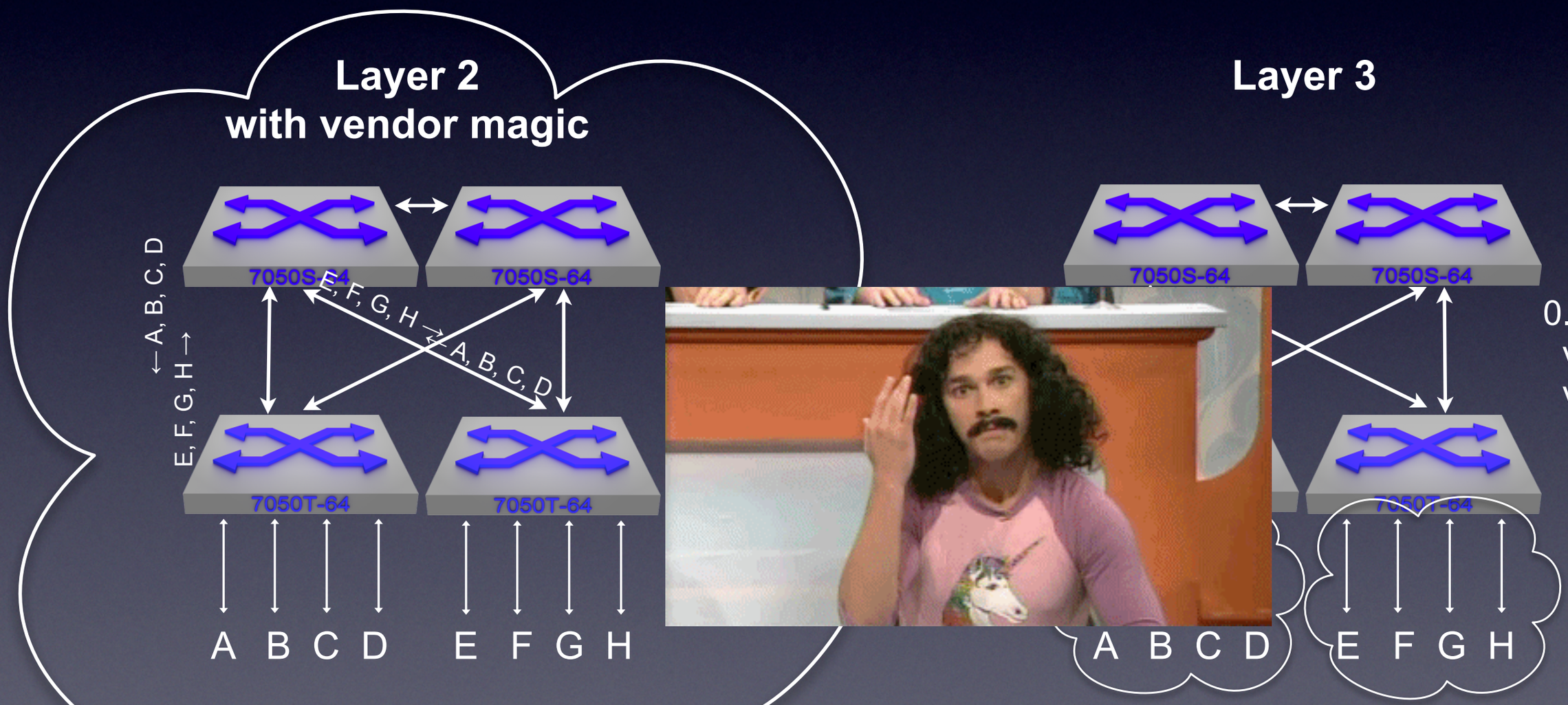
- Layer 2: “data link layer” – Ethernet “bridge”
- Layer 3: “network layer” – IP “router”





Layer 2 vs Layer 3

- Layer 2: “data link layer” – Ethernet “bridge”
- Layer 3: “network layer” – IP “router”



Layer 2

Pros:

- Plug'n'play

Cons:

- Everybody needs to be aware of everybody
  ⇒ watch your MAC & ARP table sizes
- No visibility of individual hops
- Vendor magic or spanning tree
- If you break STP and cause a loop
  ⇒ game over
- Broadcast packets hit everybody

Layer 3

Pros:

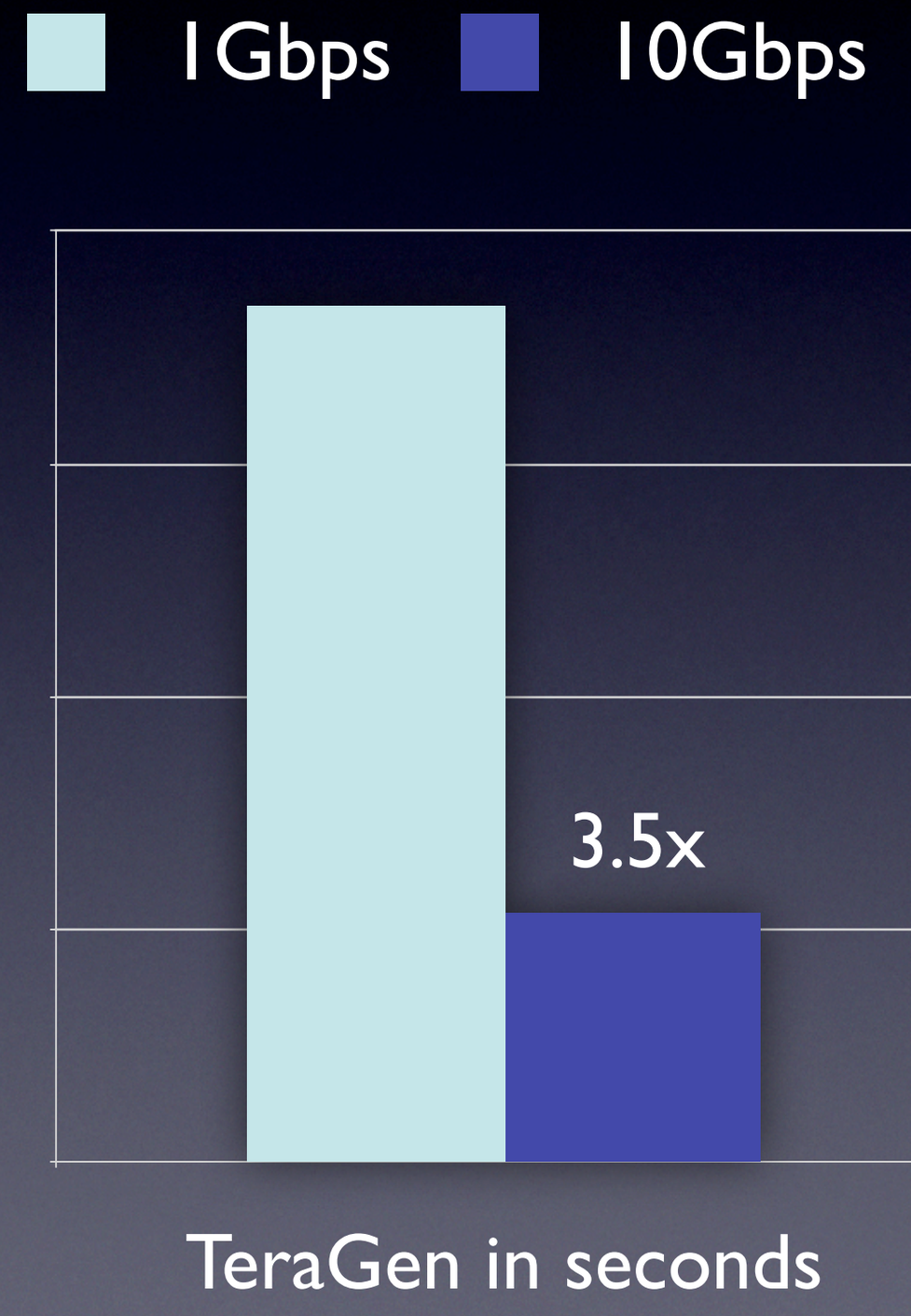
- Horizontally scalable
- Deterministic IP assignment scheme
- 100% based on open standards & protocols
- Can debug with ping / traceroute
- Degrades better under fault or misconfig

Cons:

- Need to think up your IP assignment scheme ahead of time

1G vs 10G

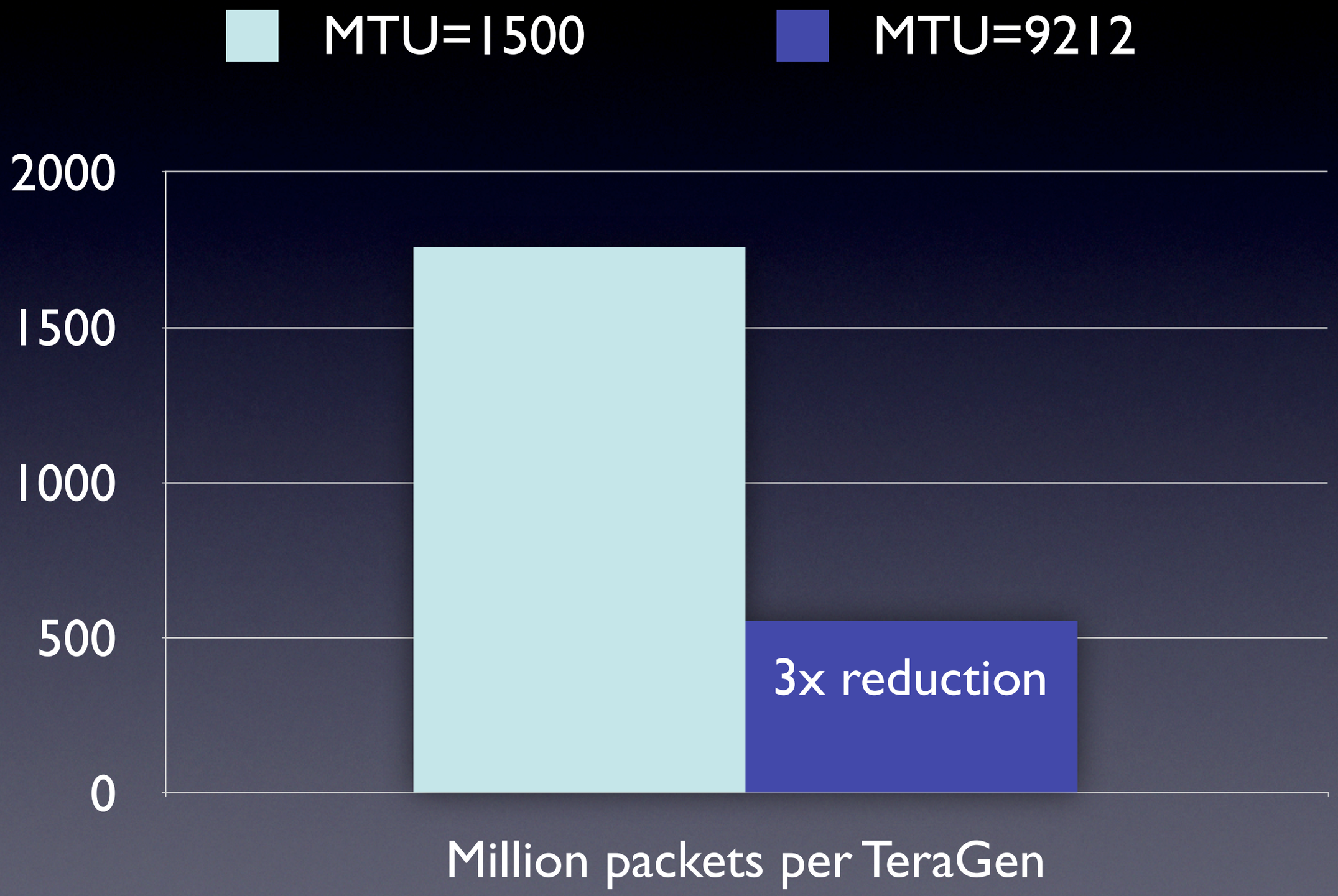
- 1Gbps is today what 100Mbps was yesterday
- New generation servers come with 10G LAN on board
- At 1Gbps, the network is one of the first bottlenecks
- A *single* HBase client can very easily push over 4Gbps to HBase



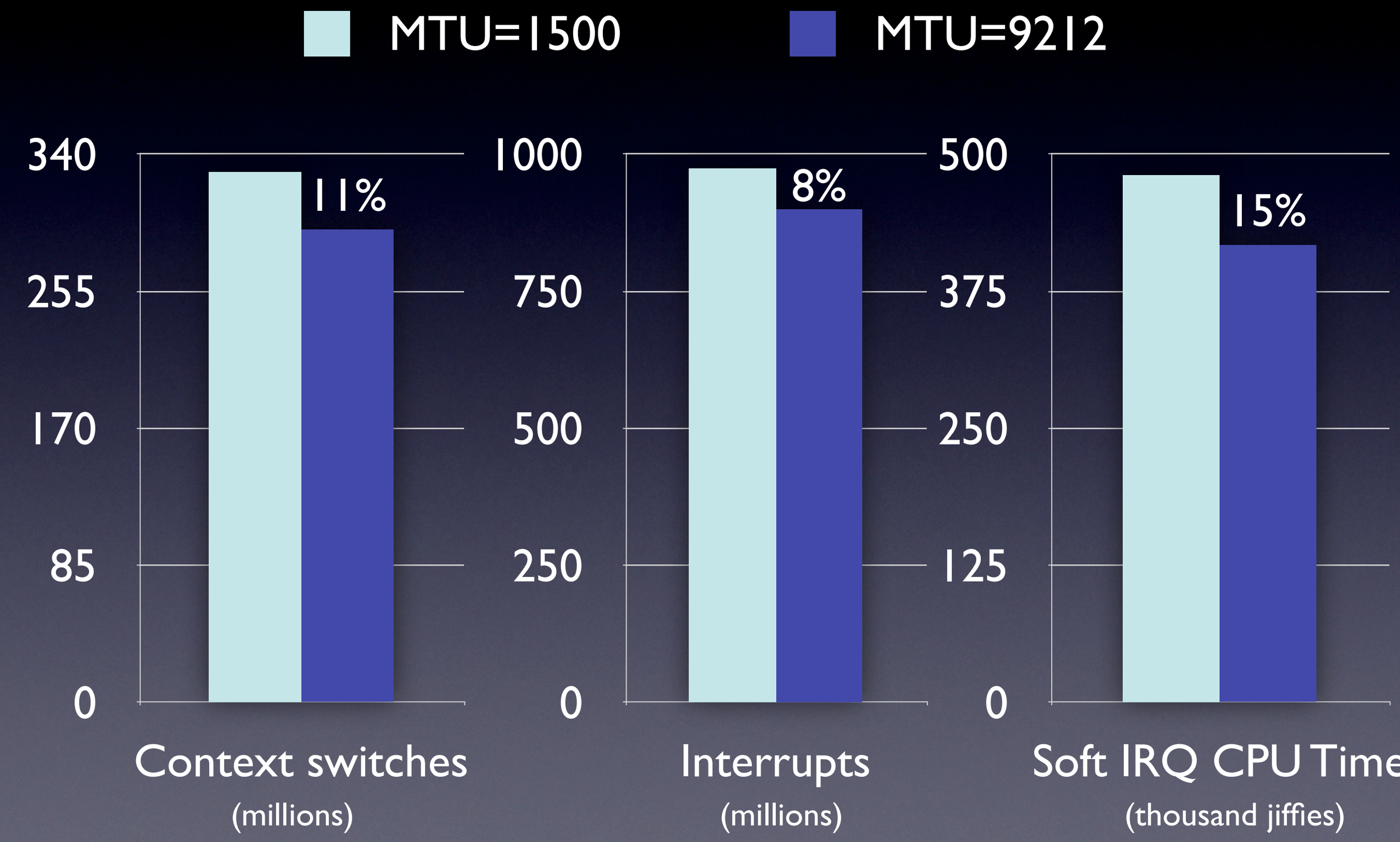
Jumbo Frames

Elephant Framed! by Dhinakaran Gajavarathan

Jumbo Frames

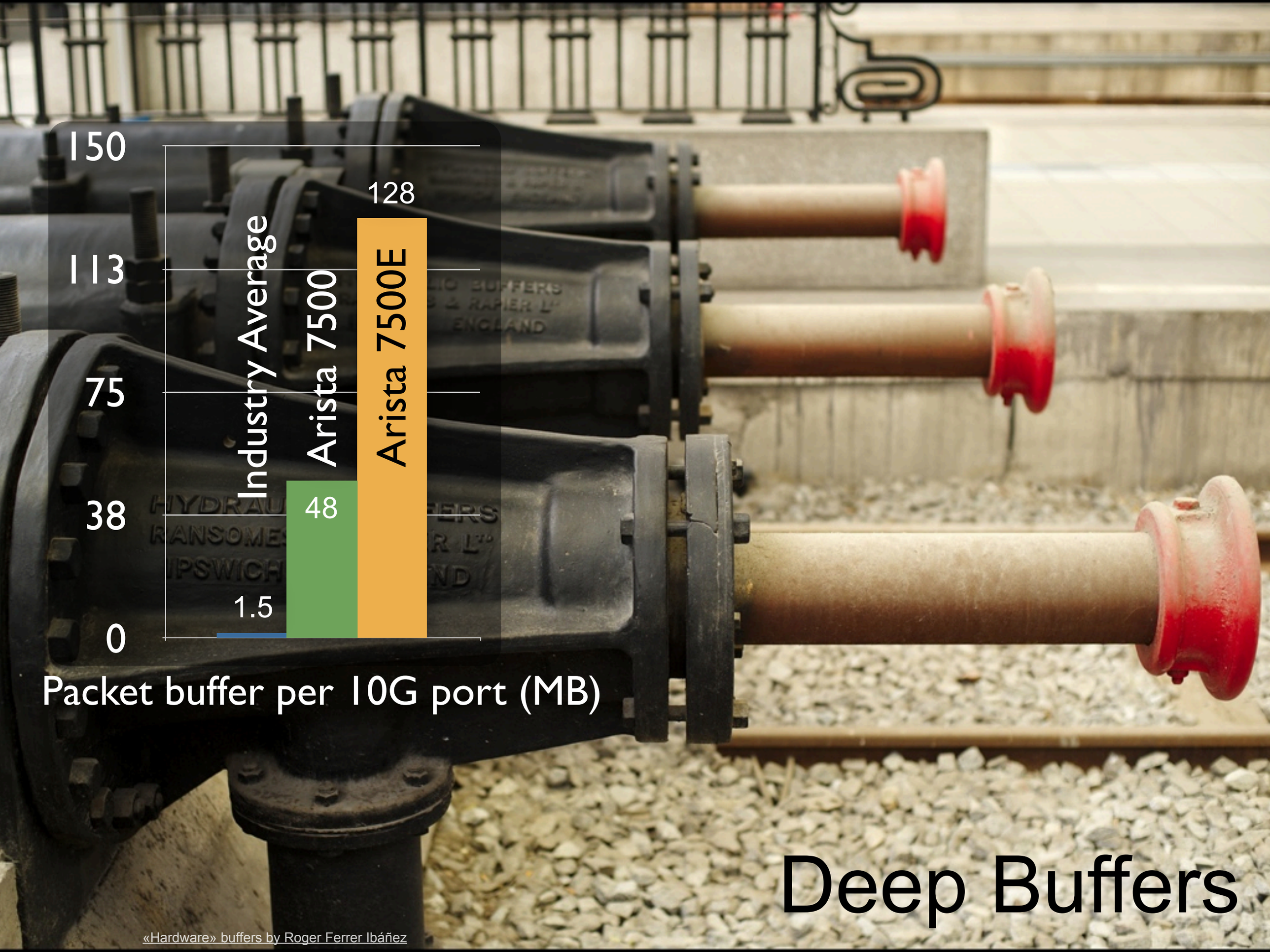


Jumbo Frames

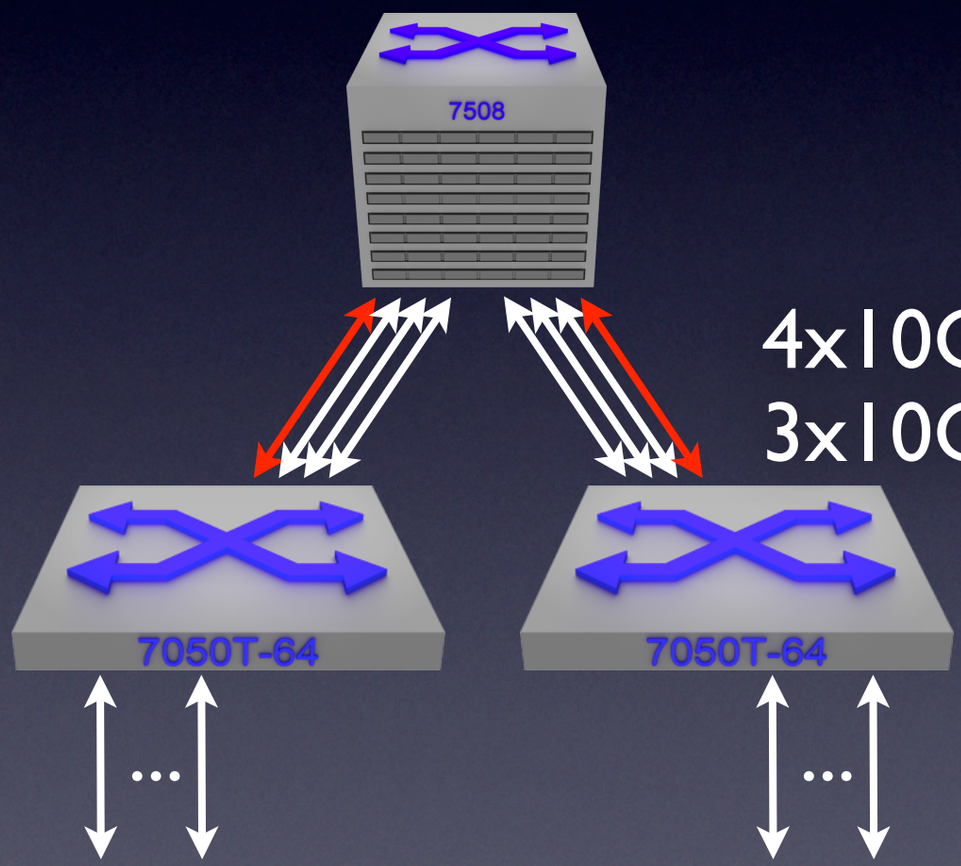


Deep Buffers

«Hardware» buffers by Roger Ferrer Ibáñez


150
113
75
38
0
Industry Average
128
Arista 7500
48
Arista 7500E
1.5
Packet buffer per 10G port (MB)
Deep Buffers
«Hardware» buffers by Roger Ferrer Ibáñez

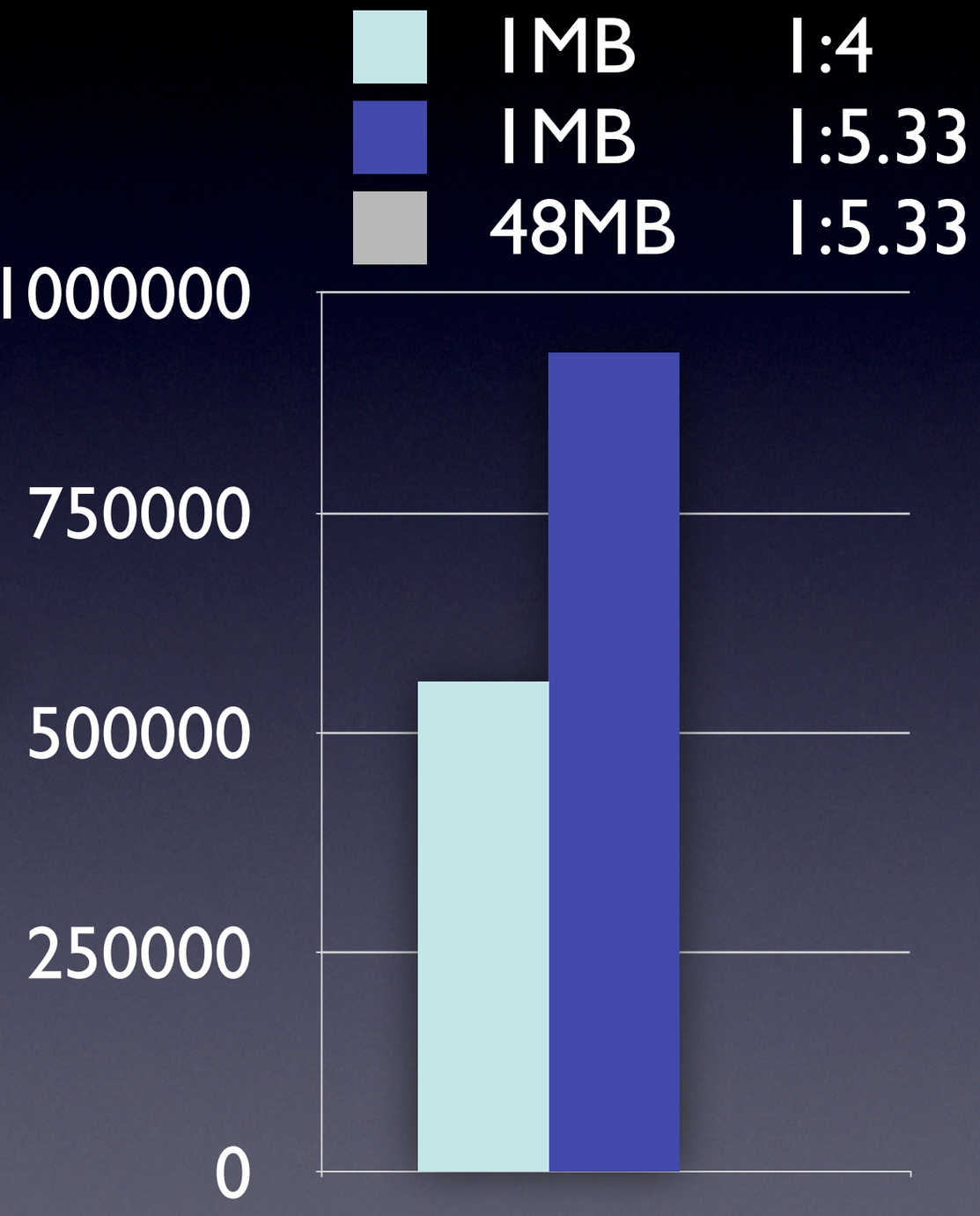
Deep Buffers



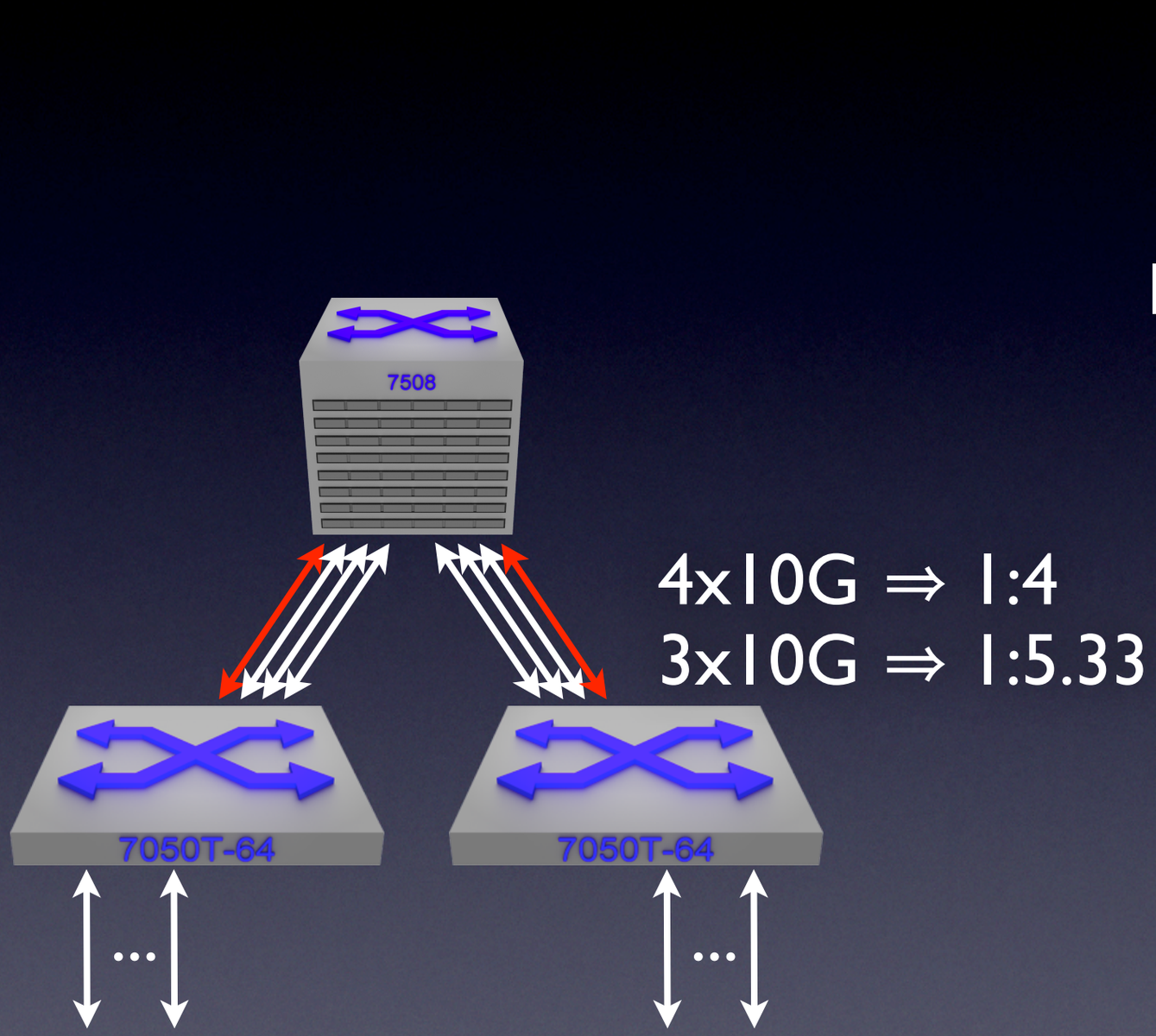
4x10G ⇒ 1:4
3x10G ⇒ 1:5.33

16 hosts
x 10G

16 hosts
x 10G



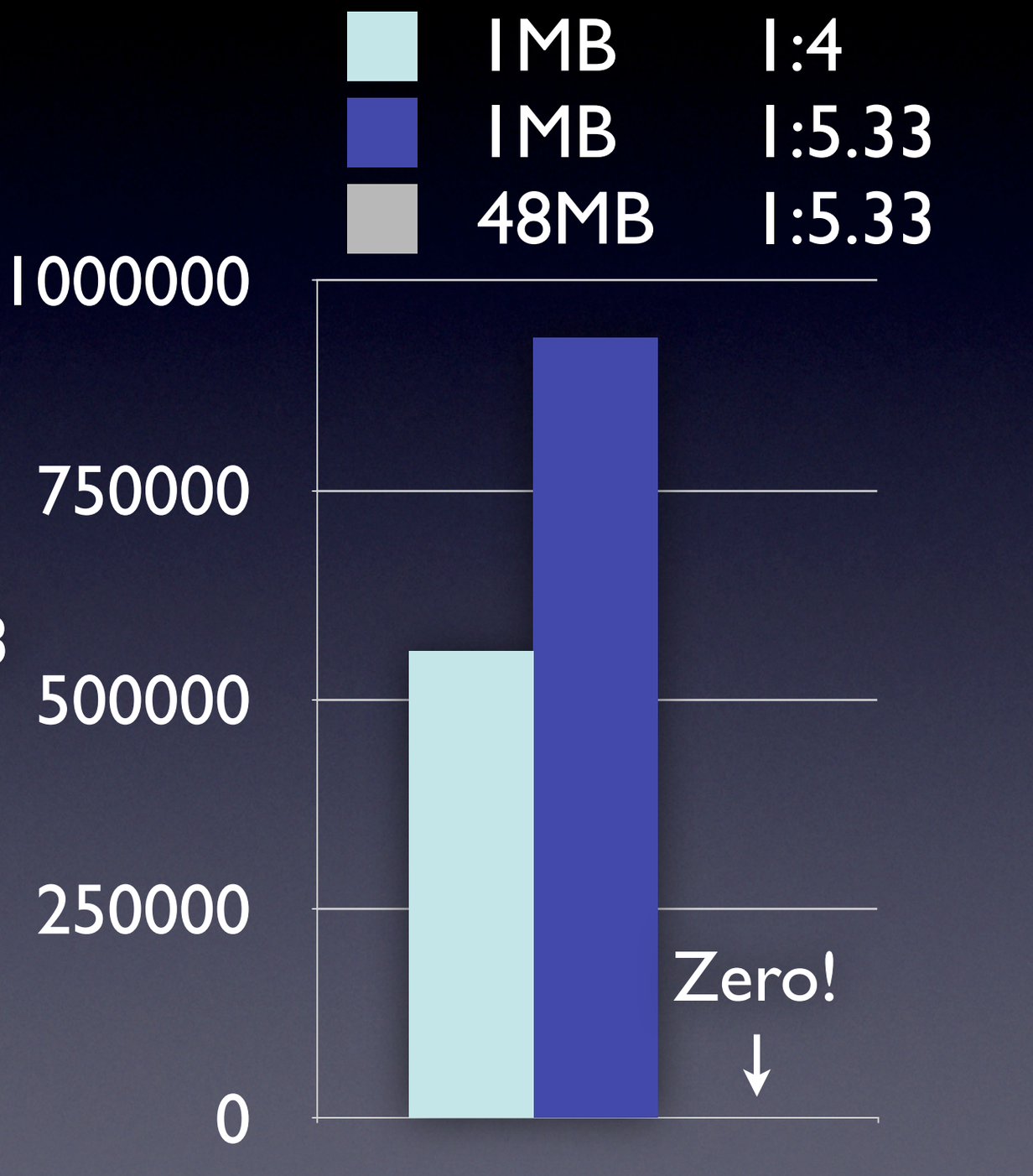
Deep Buffers



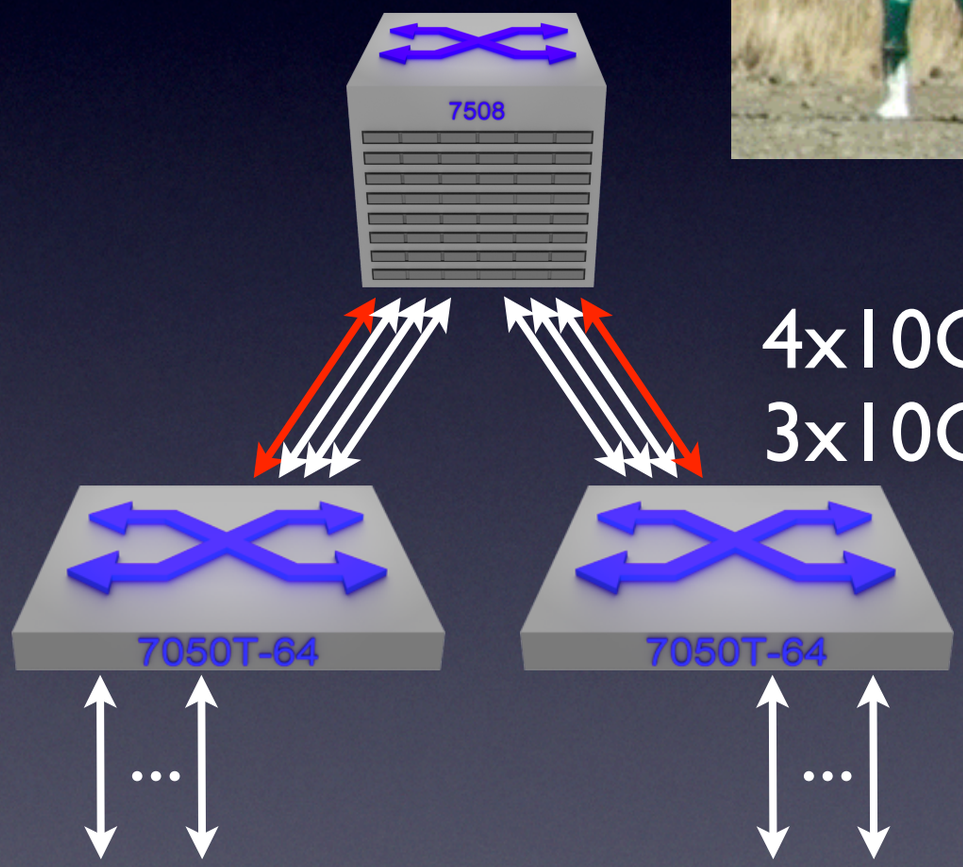
4x10G ⇒ 1:4
3x10G ⇒ 1:5.33

16 hosts
x 10G

16 hosts
x 10G



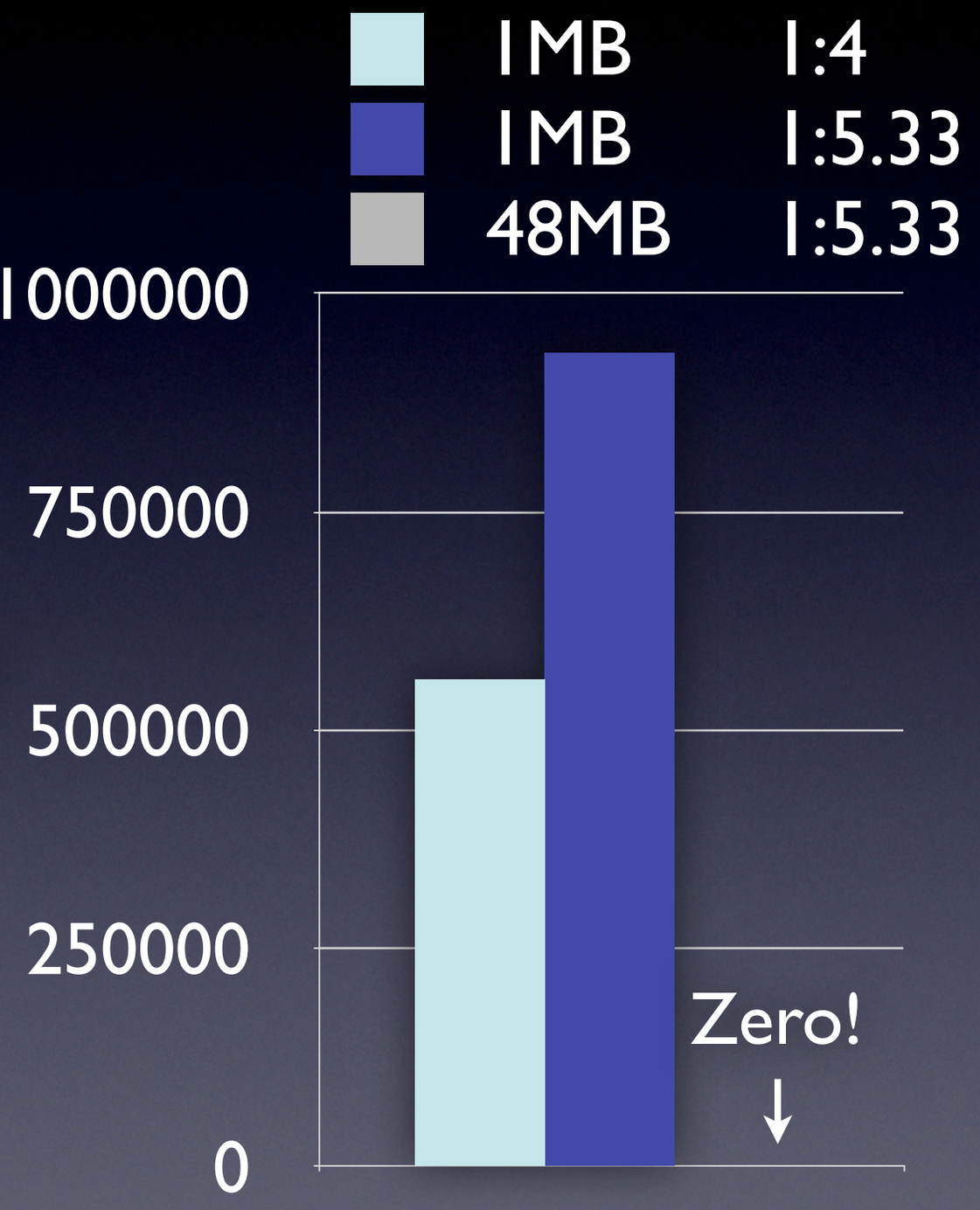
Deep Buffers



4x10G ⇒ 1:4
3x10G ⇒ 1:5.33

16 hosts
x 10G

16 hosts
x 10G



Machine Crashes

Machine Crashes

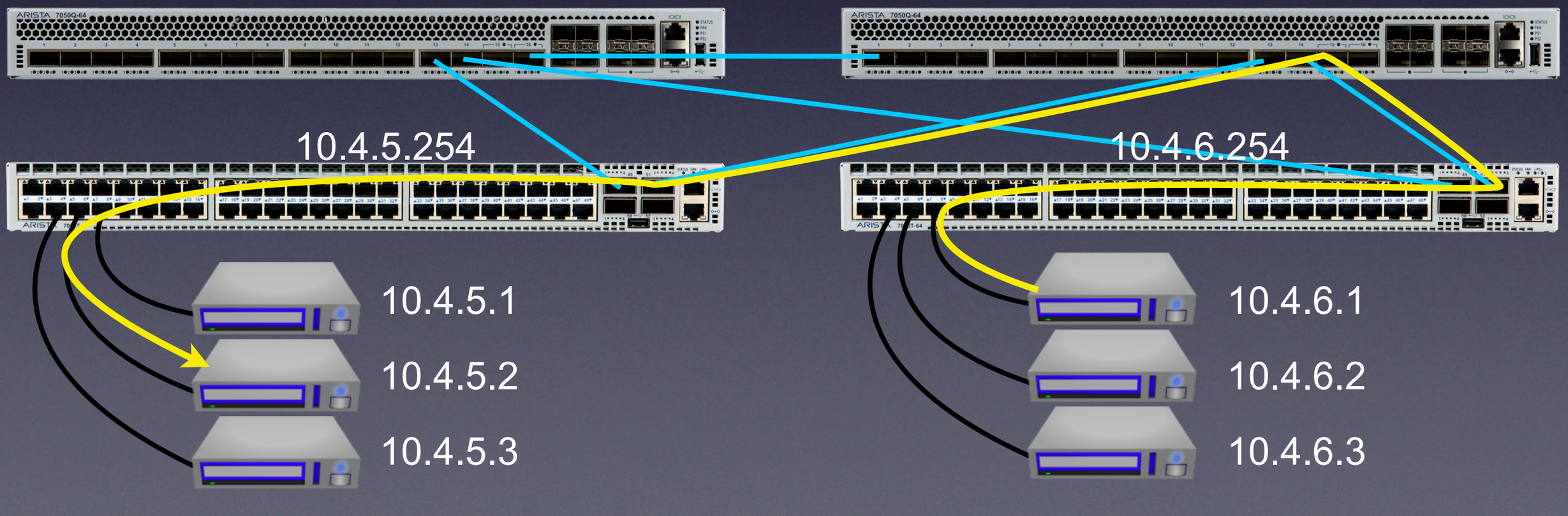
Machine Crashes

Machine Crashes

Machine Crashes

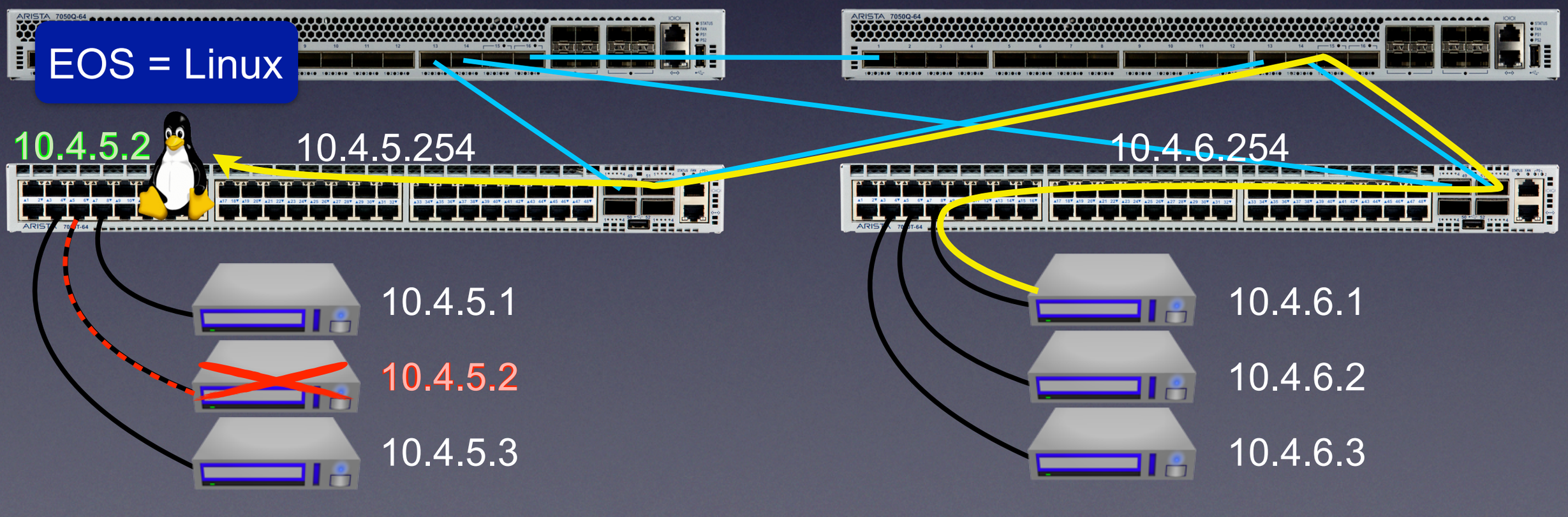
How Can a Modern Network Help?

- Detect machine failures
- Redirect traffic to the switch
- Have the switch's kernel send TCP resets
- Immediately kills all existing and new flows



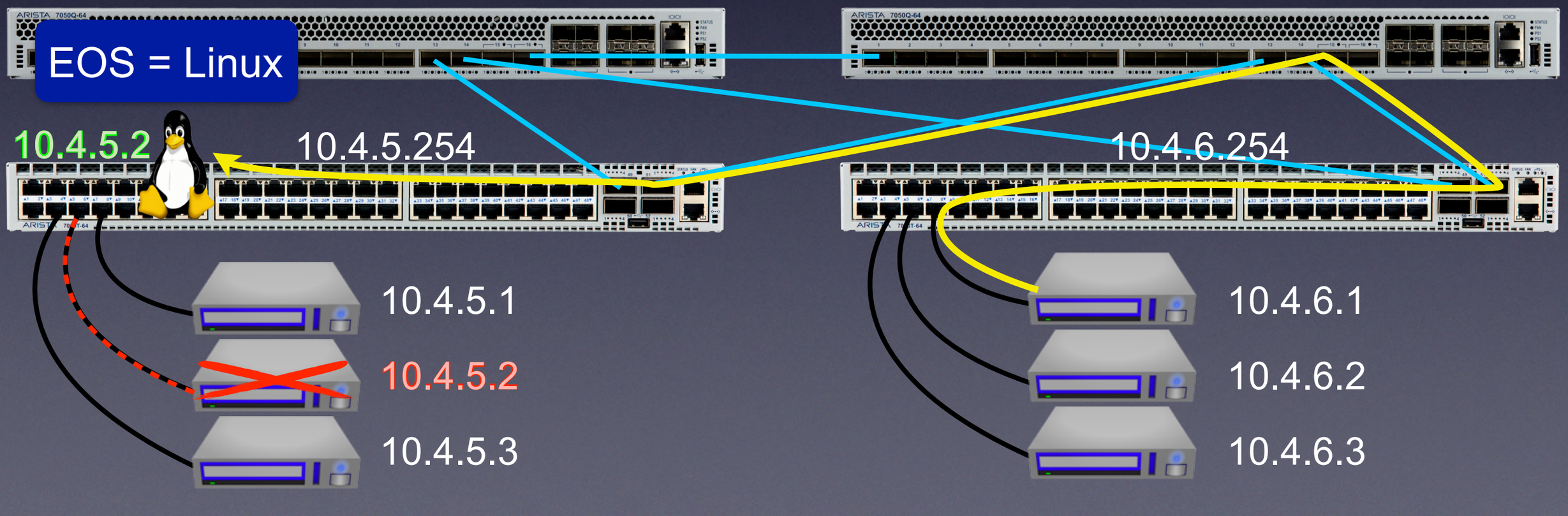
How Can a Modern Network Help?

- Detect machine failures
- Redirect traffic to the switch
- Have the switch's kernel send TCP resets
- Immediately kills all existing and new flows



Fast Server Failover

- Switch learns & tracks IP ↔ port mapping
- Port down → take over IP *and MAC* addresses
- Kicks in as soon as hardware notifies software of the port going down, within a few *milliseconds*
- Port back up → resume normal operation
- Ideal to prevent RegionServers stalling on WAL



Conclusion

- Start thinking of the network as yet another service that runs on Linux boxes
- Build layer 3 networks based on standards
- Use jumbo frames
- Deep buffers prevent packet loss on oversubscribed uplinks
- The network sits in the middle of everything; it can help. Expect to see more in that area.

Thank You

We’re hiring in SF, Santa Clara, Vancouver, and Bangalore

ARISTA

Benoît “tsuna” Sigoure
Member of the Yak Shaving Staff
tsuna@aristanetworks.com

@tsunanet