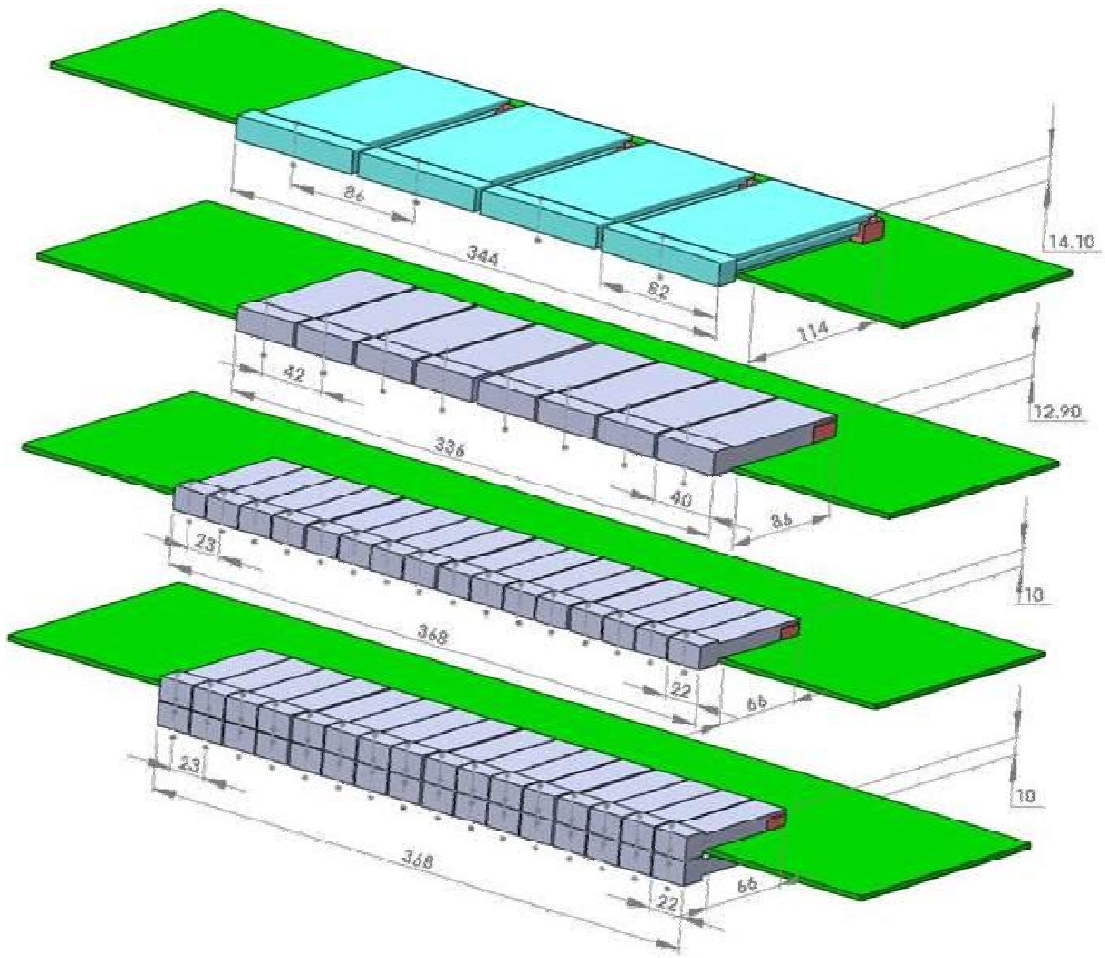
BROCADE

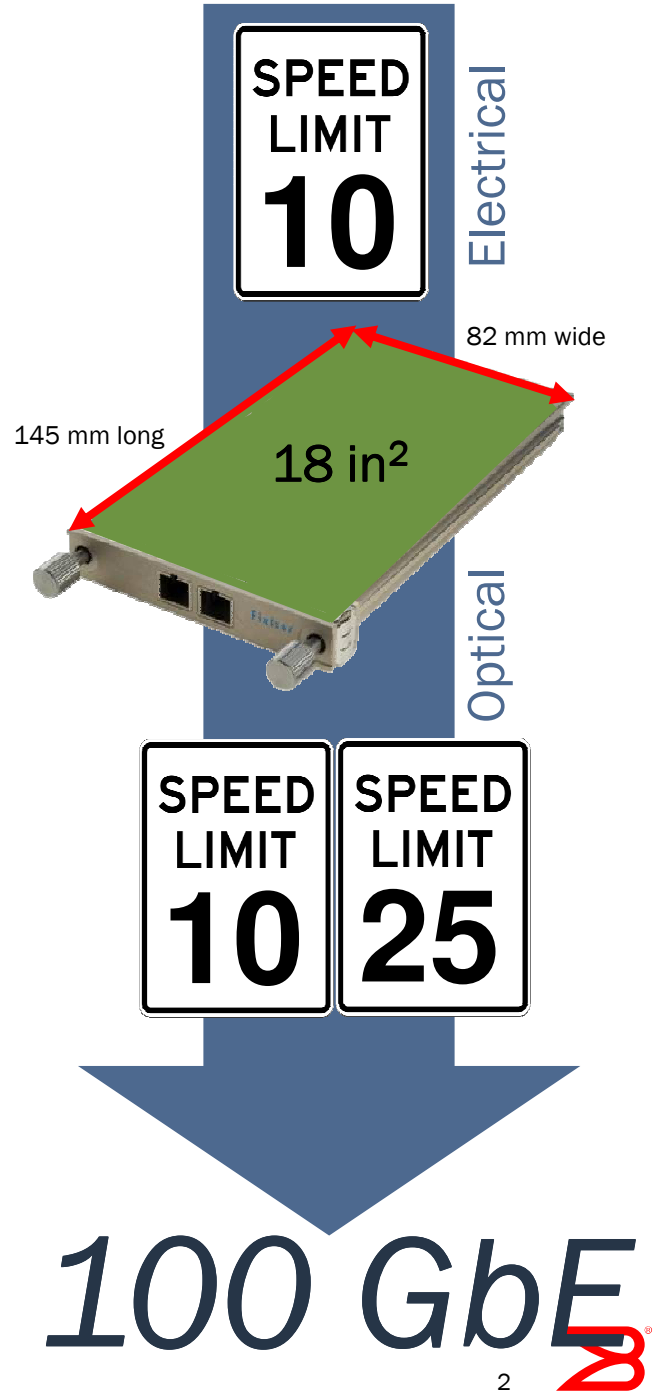


100 GBE AND BEYOND

Diagram courtesy of the CFP MSA.

Current State of the Industry

- Fundamental 1st generation technology constraints limits higher 100 GbE density and lower cost
- Electrical signaling inside the box
  - 100 Gbps Attachment Unit Interface (CAUI) uses 10 x 10 Gbps
- Optical signaling outside the box
  - 10x10 MSA: 10 x 10 Gbps
  - 100GBASE-LR4 and 100GBASE-ER4: 4 x 25 Gbps
- CFP module size and power consumption

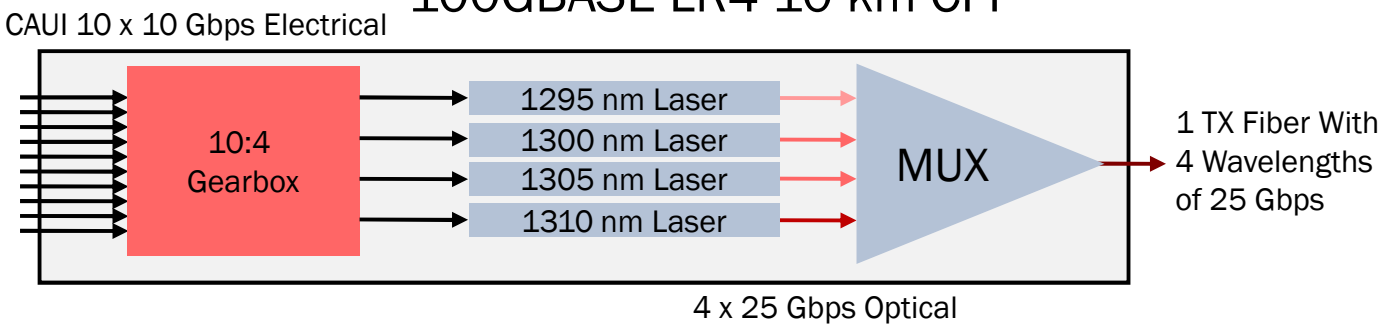


CFP image courtesy of Finisar.

100 GbE Module Technologies Compared

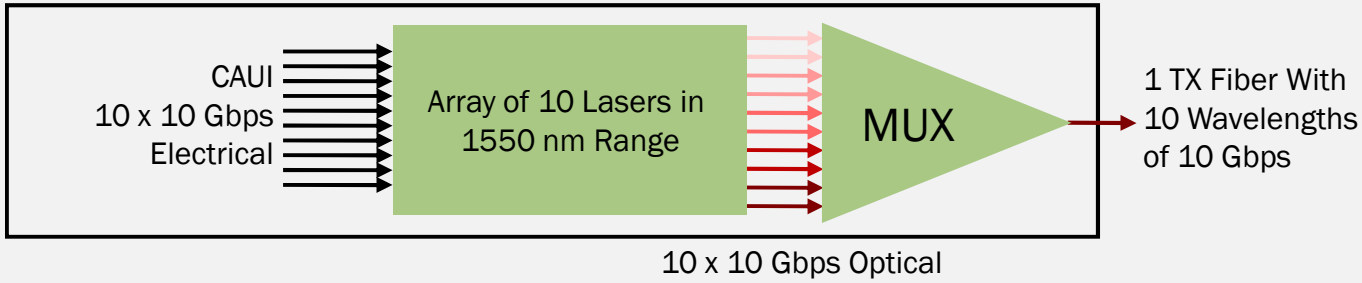
Transmit Side of Module

100GBASE-LR4 10 km CFP



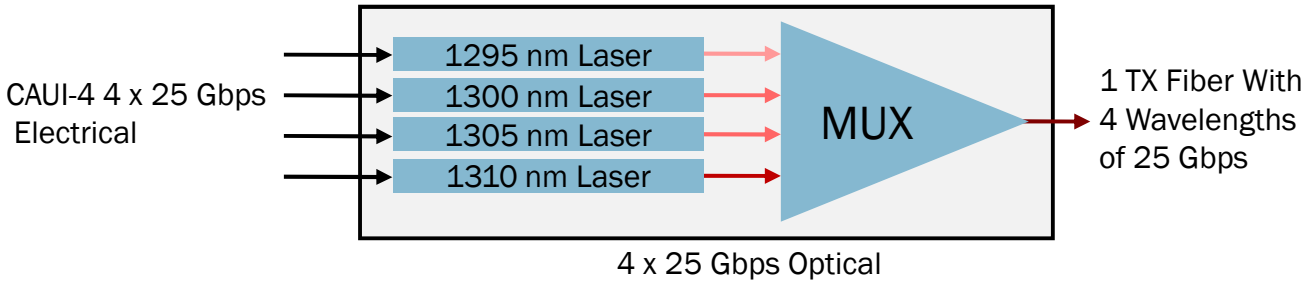
- Most expensive, complex and uses the most power
- Gearbox converts 10 x10 Gbps electrical signaling into 4 x 25 Gbps signaling

10x10 MSA 2 km, 10 km, 40 km CFP



- Less cost, complexity and power consumption
- Uses 10 x 10 Gbps electrical and optical signaling
- Doesn't need the gearbox

100GBASE-LR4 10 km CFP2



- Lower cost, complexity and power consumption
- Uses 25 Gbps electrical and optical signaling
- Doesn't need the gearbox

10X10 MSA

LOW COST 100Gb/s PLUGGABLE OPTICAL TRANSCEIVER

- Current IEEE standards have a gap
  - 100GBASE-SR10 supports up to 150 m on OM4 MMF
  - 100GBASE-LR4 supports up to 10 km on SMF
  - Missing a shorter SMF reach
- 100GBASE-LR4 100 GbE optics are complex and expensive
- 10x10 MSA bridges the gap
  - Support for 2 km, 10 km and 40 km on SMF
  - Considerably more economical
  - Eliminates expensive components
  - Consumes lesser power
- Network operator members!

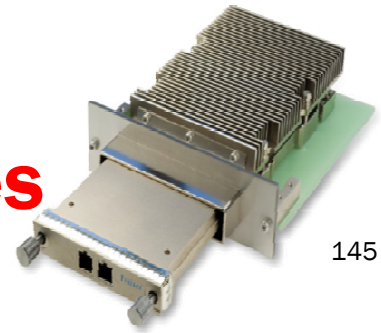
Members

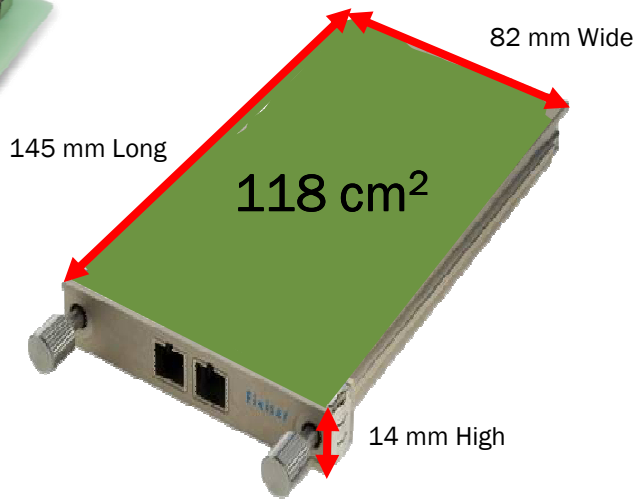
BROCADE, EXFO, KAIAM, Google, HUAWEI, INNOLIGHT, JDSU, KOTURA, Verdant Networks, Santur, facebook, VITESSE, Enablence, EFFDON, MAXIM, CYOPTICS, amsix, Viscore, AFOP, sti systems, MRV, CORTINA, OPLINK, HitachiCable, source photonics

www.10x10msa.org

100 GbE CFP Modules

C (100) Form-factor Pluggable

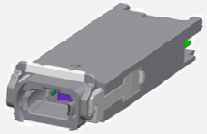
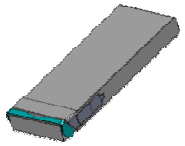
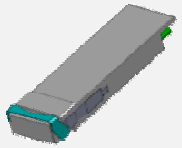




- New module optimized for 100 GbE long reach applications
- Used for 40GBASE-SR4, 40GBASE-LR4, 100GBASE-SR10, 100GBASE-LR4, 100GBASE-ER4, and 10x10 MSA
- Complex electrical and optical components need a large module
- Large module form factor and power consumption limits front panel density (larger than an iPhone)

100 Gbps Module Evolution

Two Generations of 100 GbE Expected to Take 5 Years

	1st Generation		2nd Generation		
Module Name (Images not to Scale)	 CFP	 CXP	 25 Gbps QSFP	 CFP2	 CFP4
Approximate Module Dimensions (Length x Width to Scale)					
Front Panel Density	4	16	22 - 44	8	16 - 32
Electrical Interface	CAUI	CPPI	CPPI-4	CAUI-4	CPPI-4
Electrical Signaling (Gbps)	10 x 10	10 x 10	4 x 25	4 x 25	4 x 25
Media Type	SMF	Twinax, MMF	MMF/SMF?	SMF	SMF
Advantages	Long Reach, High Power Dissipation	Small Size, Designed for Passive Cabling	Highest Density, Established Form Factor	Long Reach, Higher Density	Highest Density, Smaller Size,
Disadvantages	Too Big	Short Reach, Too Small	Limited Power Dissipation and Reach	Bigger Size	Unproven Form Factor (vs. QSFP)
Availability *(Subject to Change)*	2010	2010	2011 (InfiniBand) 2013+ (Ethernet)	2013+	2014+

CFP and CXP images courtesy of Finisar, QSFP image courtesy of the SFF Committee, SFP2 and SFP4 images courtesy of the CFP MSA.

Future 100 GbE Projects

- In the short term, 4 x 25 Gbps electrical and optical interfaces will keep the IEEE 802.3 Working Group busy for 2+ years
- 100 GbE serial is still not feasible in the near future
  - 25 Gbps signaling is challenging
  - We'll get a better idea of what is possible as 25 Gbps technology matures
- 3rd generation 100 GbE is likely to be developed several years from now

Next Higher Speed Ethernet

250 GbE, 300 GbE, 400 GbE, or TbE?

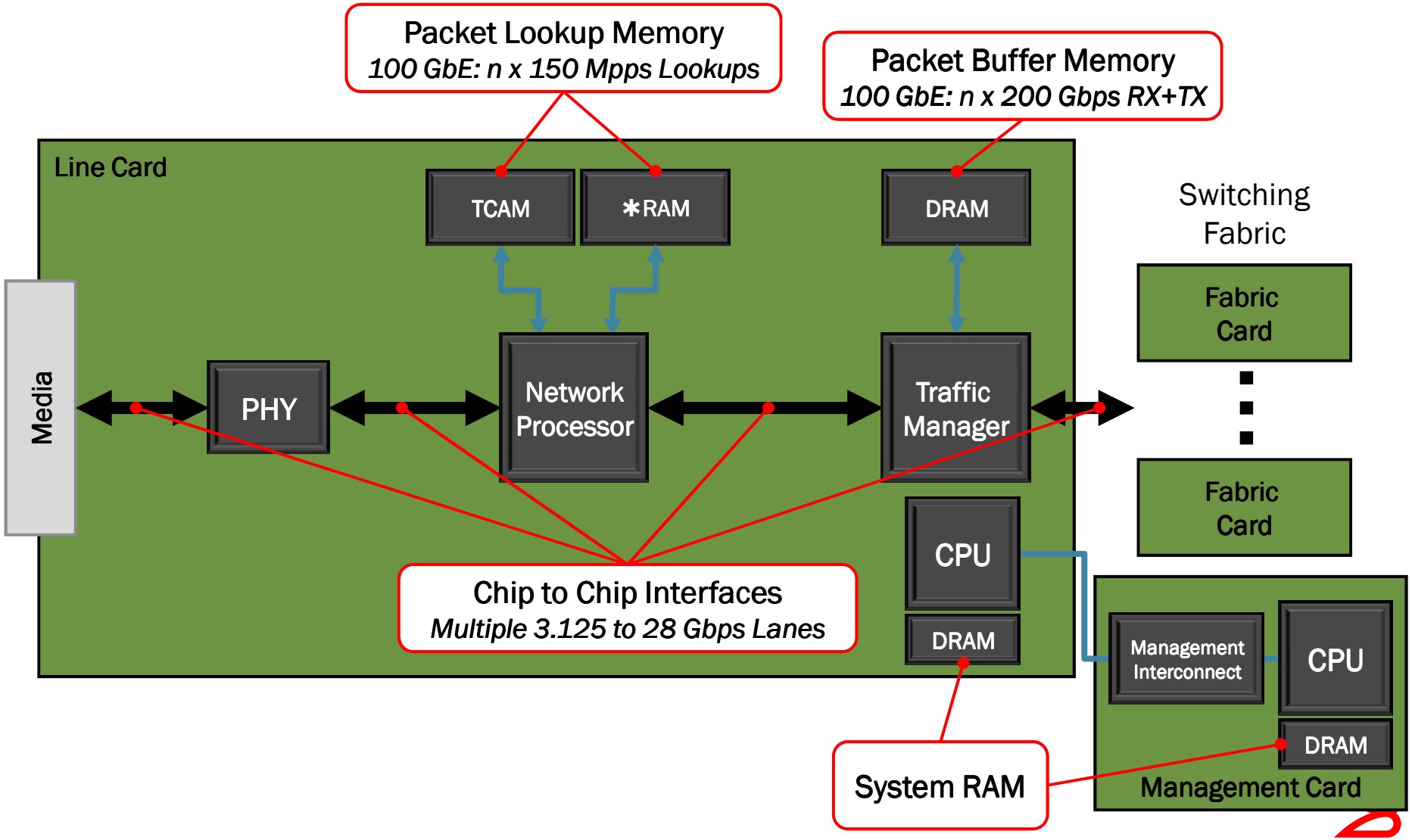
- Using 10 x 25 Gbps signaling the next speed could be 250 GbE
  - The industry wants a larger jump
- 12 x 25 Gbps signaling matches the number of fibers in a high density MMF cable for 300 GbE
  - Unpopular too
- The likely candidate for the next speed is 400 GbE using 16 x 25 Gbps signaling
  - 16 x 25 Gbps wavelengths can be easily muxed/demuxed onto one SMF
  - MMF solutions would need 32 fibers in a high density cable MPO/MTP assembly
  - Evolution to 10 x 40 Gbps signaling
- TbE is simply impractical in the near future
  - 40 x 25 Gbps lanes in and 40 x 25 Gbps lanes out would make a gigantic media module
  - 40 Gbps serial lanes aren't expected to be economical until after 2016, and will take considerable work as electrical losses grow exponentially with super high frequency signaling

BROCADE

PUSHING THE LIMITS, A PERSPECTIVE ON ROUTER ARCHITECTURE CHALLENGES

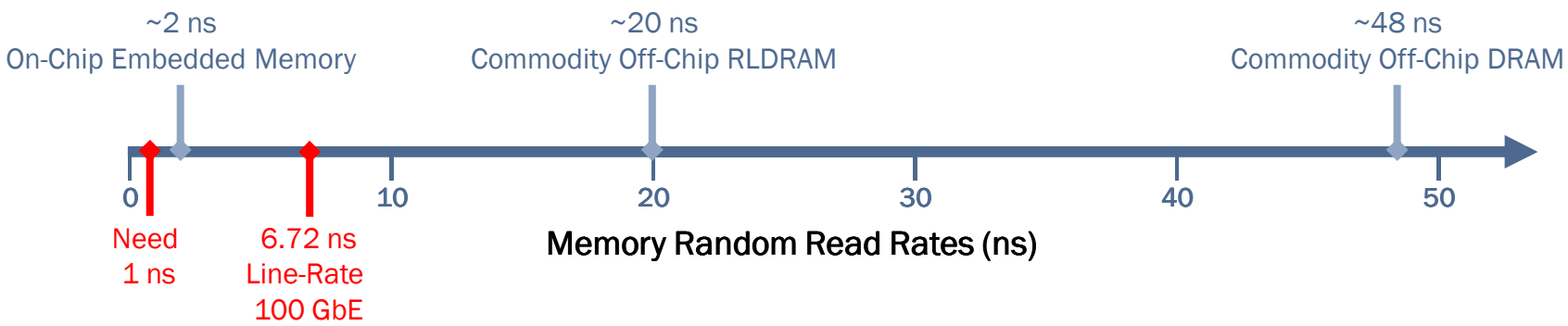
2012/02/28

Basic Router Forwarding Architecture

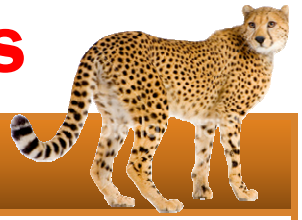


Key Memory Challenges

- Packet rates gave greatly exceeded memory random read rates
  - Lookup: 1 ns random read rates needed yesterday
  - Buffer: 1 ns random read and writes rates needed yesterday
- Dynamic memory technology characteristics impose significant constraints on lookup and buffering architectures
  - Inherent restrictions and non-random access
  - Bank blocking due to previous read/write events
  - Applies to both on-chip and off-chip solutions
  - Adds forwarding latency

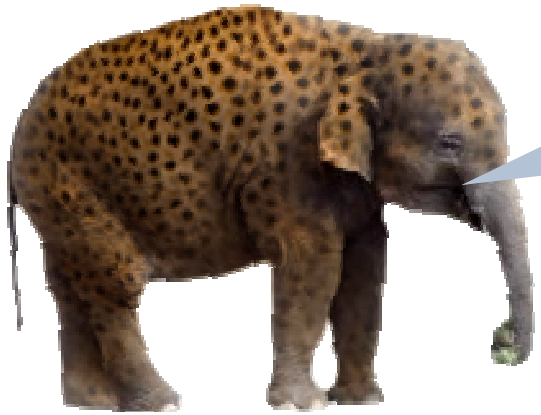


Lookup and Buffering Memory Requirements

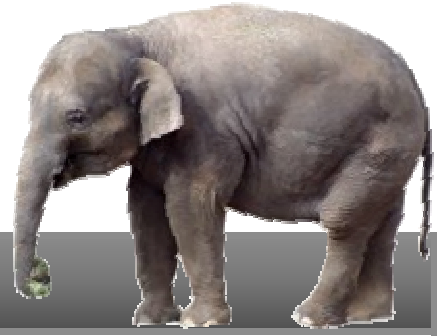


Fast!

- Have to store everything needed for packet lookups in hardware to forward at line rate with all features enabled
  - 10 GbE: 15 Mpps or one packet every 67 ns
  - 100 GbE: 150 Mpps or one packet every 6.72 ns
  - Multiple ports on a network processor
  - Multiple lookups are needed per packet



I'm Fast and Big, But I Don't Exist



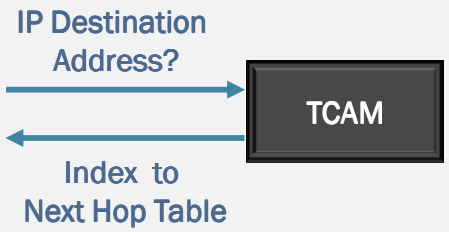
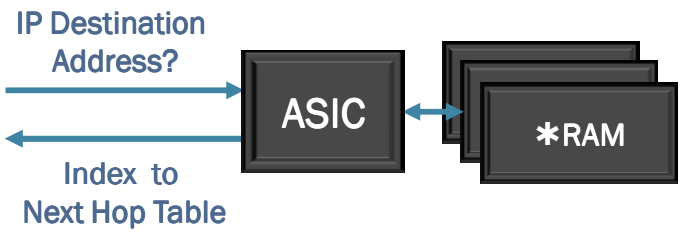
Big!

- Packet lookup tables hold
  - MAC address table (L2, VPN)
  - IPv4 FIB (unicast and multicast, VPN)
  - IPv6 FIB (unicast and multicast, VPN)
  - VLAN tags, MPLS labels
  - ACLs (L2, IPv4, IPv6, ingress and egress)
  - QoS policies (PHB, rewrite, rate limiting/shaping)
- Deep buffering, queuing and shaping
  - Multiple 100 Gbps of sustained throughput into buffer memory
  - 1 GB buffer is only 80 ms at 100 GbE rates
  - 1000s of queues per port

Lookup and Buffering Memory Technology Overview

	TCAM	SRAM	DRAM		
			DDR	RLDRAM	GDDR
What is it?	Ternary Content Addressable Memory	Static RAM	Double Data Rate Dynamic RAM	Reduced-Latency Dynamic RAM	Graphics Double Data Rate Dynamic RAM
Primary Function	Wildcard Search	High-Speed Storage	High-Capacity Storage	High-Speed Storage	High-Bandwidth Storage
Industry Usage	Specialty	Commodity	Commodity	Specialty	Specialty
Access Speed	Lower	Lowest	Highest	Higher	Higher
Density	Lower	Lower	Much Higher	Higher	Much Higher
Cost per Bit	Much Higher	High	Much Lower	Lower	Lower
Power Consumption	Highest	Lower	Higher	Higher	Higher
Mass Production Capacities	40 Mbit Today, 80 Mbit Soon	72 Mbit Today, 144 Mbit Soon	2 Gbit DDR3 Today, DDR4 Soon	576 Mbit Today, 1 Gbit Soon	2 Gbit GDDR5 Today

Longest Prefix Matching (LPM) Mechanisms

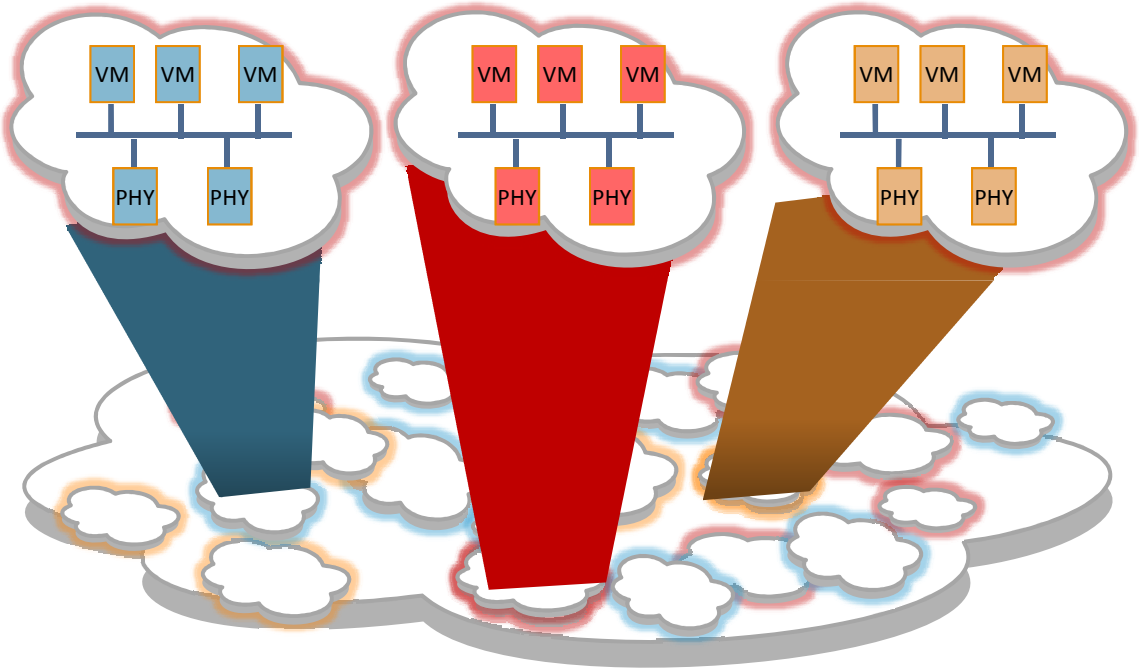
	TCAM	*RAM
What is it?	 Wildcard Hardware Data Search	 Ordered Tree Data Structure Search in Hardware
Cost	Much Higher	Lower
Power Consumption	Higher	Lower
Search Latency	Fixed	Variable
Throughput	Higher	Lower, Must Parallelize
Add/Delete Time	Fixed	Variable
Prefix Capacity	Fixed, Lower	Variable, Higher
Search Algorithm	Ordering of Data in TCAM	Lots of Different (Patented) Algorithms with Tradeoffs
Software Complexity	Versatile Data Storage No Need to Worry About How You Write the Data	Need to Implement Lookup and Data Structure Mechanisms Separately Must Optimize Layout for Data Structure

LPM Memory Architecture Solutions

- Divide and conquer parallel architecture
  - Deterministic search using a combination of SRAM/DRAM
  - Large number of banks allows parallel searching in reasonable time
- Integrate lookup memory into packet processing ASICs
  - Combine embedded memories for higher performance and reasonable density
- Component technology must be available for 5+ years at a minimum
  - Many specialized memory technologies have a window of production that is too small

OpenFlow Basics

A SDN Application: Network Virtualization

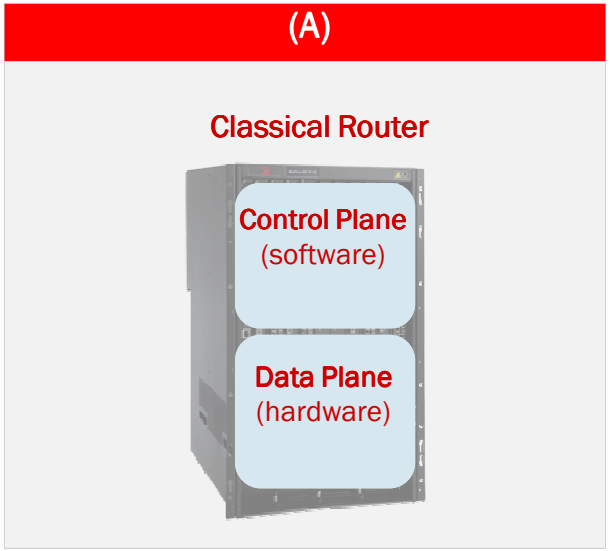


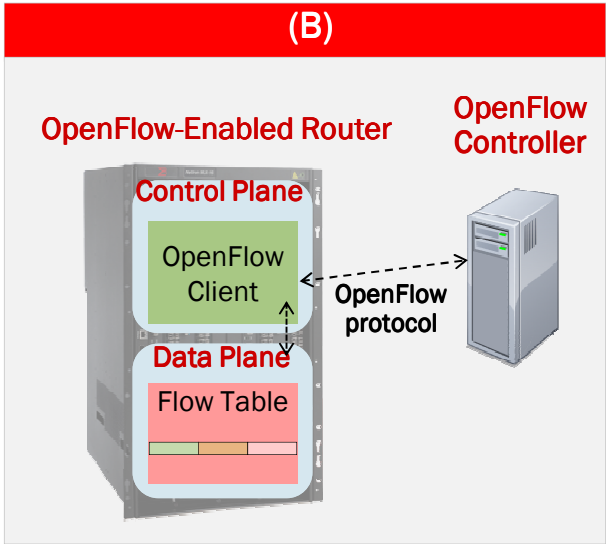
SP Physical Infrastructure

- SDN enables creation of logical networks (multi-tenancy) over a common physical network
- Enables seamless control of network resources regardless of location
- Logical networks can be used to bridge private and public clouds

OpenFlow Introduction

- In a classical router, the data plane (hardware) and control plane (software) are on the same device
- Part of the control plane functionality supported outside the router
  - "Flow table" in a router manipulated by controller
  - Router and controller communicate via OpenFlow protocol
- Originally developed by the OpenFlow Consortium
  - http://www.openflow.org
- OpenFlow is now being developed at the ONF
  - http://www.opennetworkingfoundation.org/



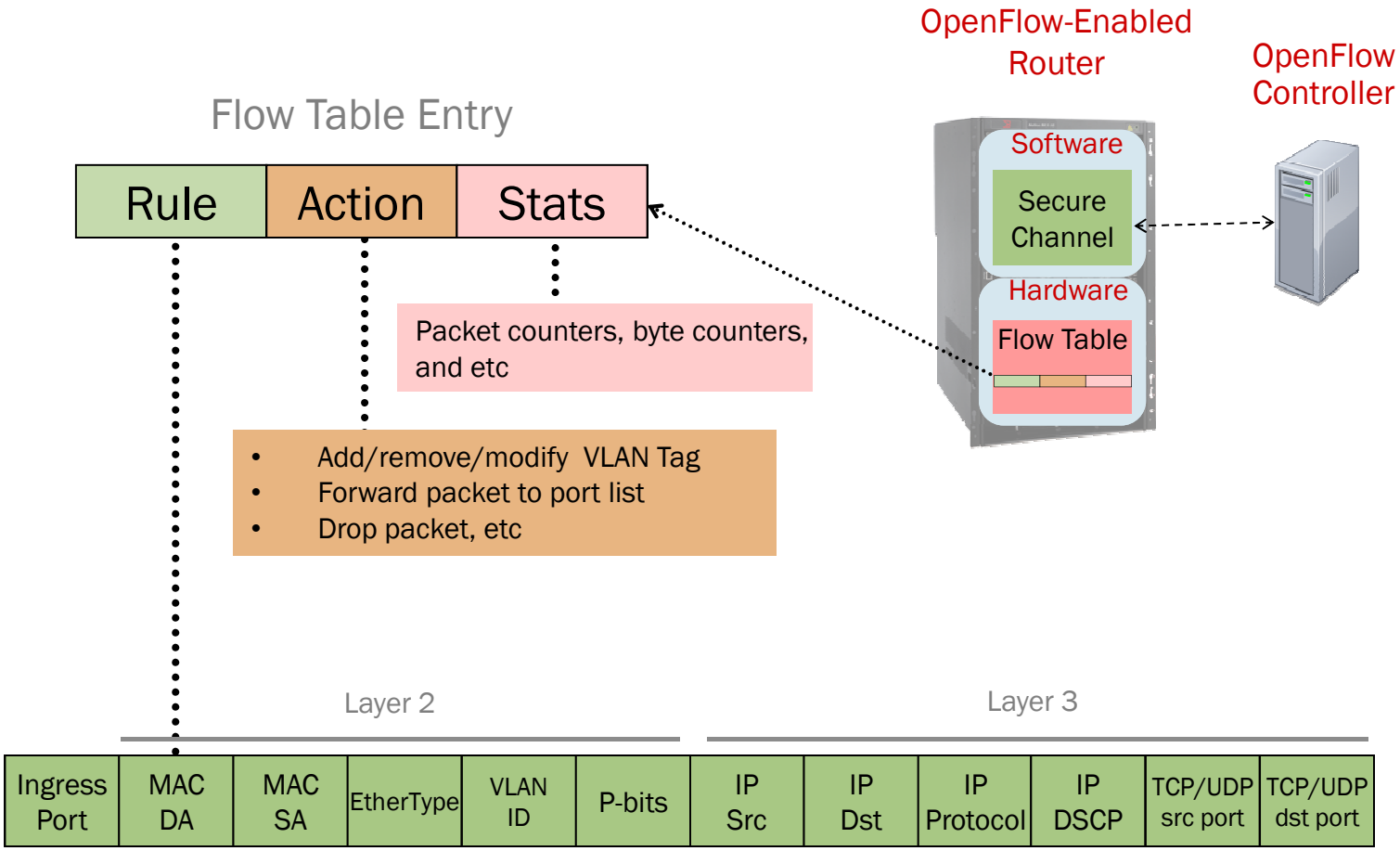


OpenFlow Enabled Router Operation

- Incoming packets are matched against the flow entries (in order)
- If there is match, the set of actions for that flow entry are performed
- Packets that don't match any flow entry are typically dropped (default)
  - Optionally, packets can be encapsulated and sent to the OpenFlow Controller
  - The OpenFlow Controller can decide how to process such a packet flow, and then (optionally) add a Flow Table entry for that flow
- OpenFlow is backward compatible with legacy networks
  - An OpenFlow-enabled router can support classic router functionality on some ports along with OpenFlow-enabled functionality on other ports

Flow Table Entry

OpenFlow v1.0.0

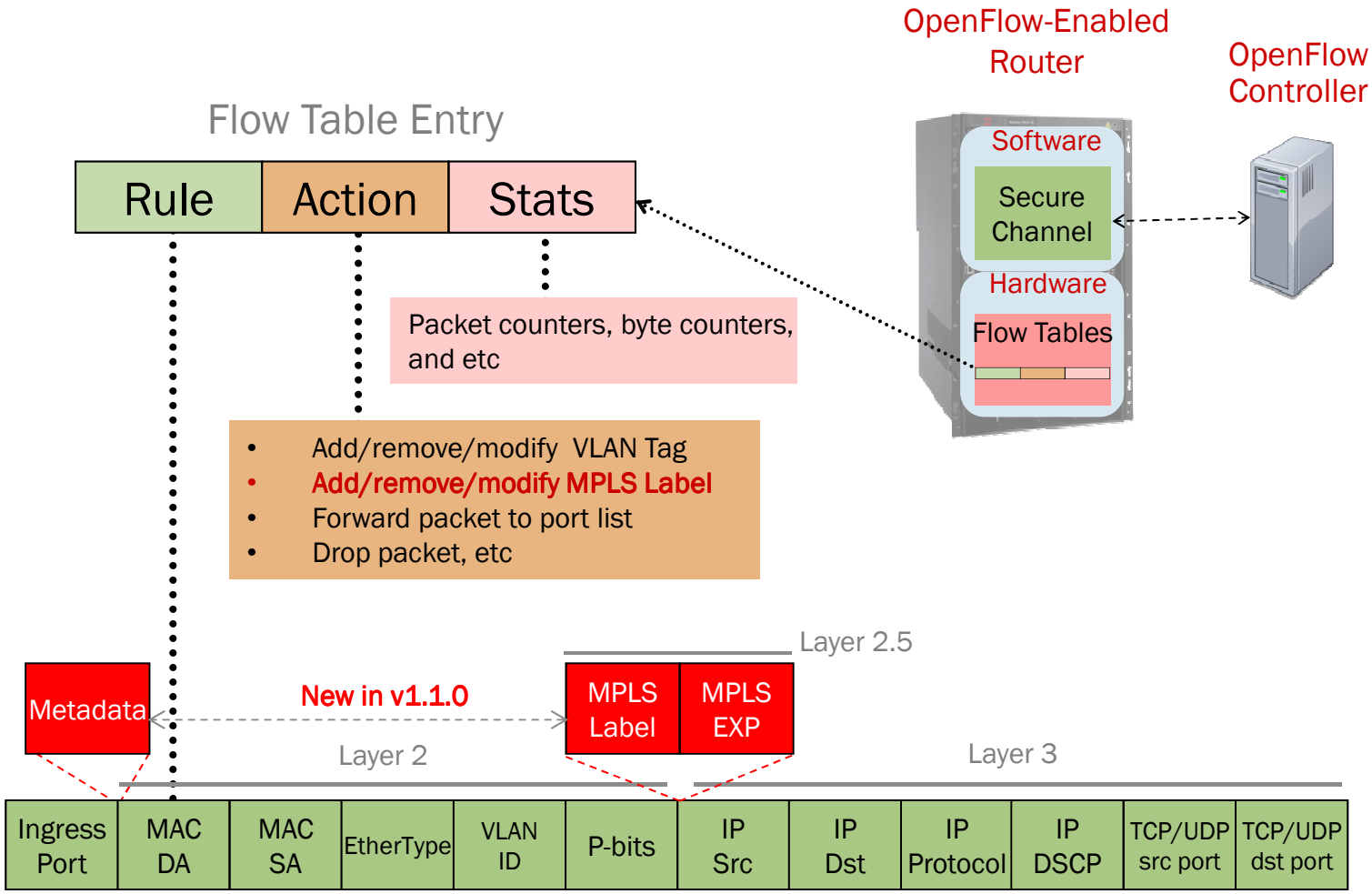


- Each flow table entry contains a set of rules to match (e.g., IP src) and an action list to be executed in case of a match (e.g., forward to port list)

OpenFlow v1.1.0

- Published in December 2010
- Main additions to v1.1
  - Multiple tables
  - VLAN stacking
  - MPLS labels
  - Logical ports (e.g., GRE)

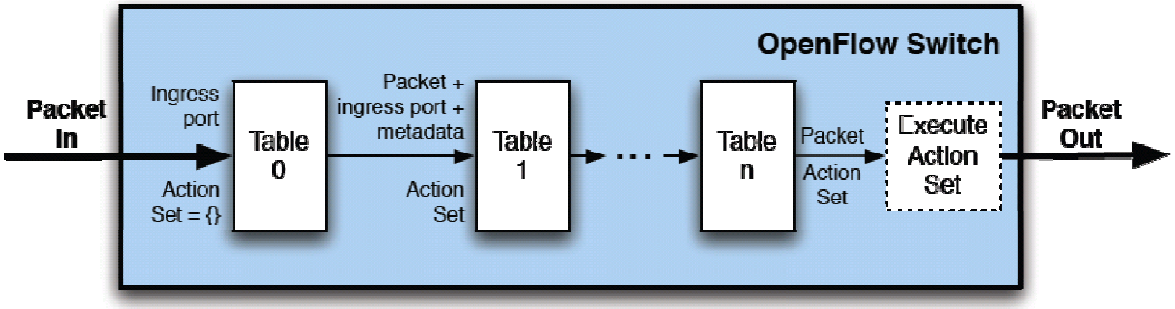
OpenFlow v1.1 Flow Table Entry Format

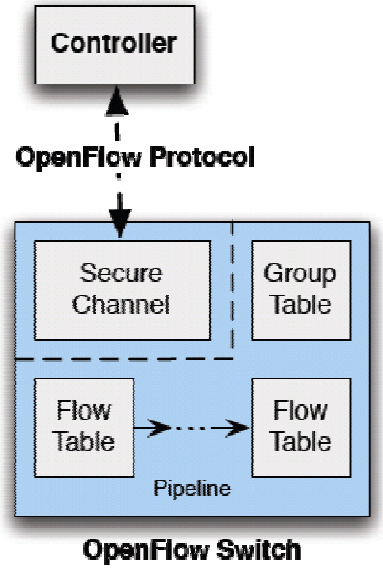


- OpenFlow v1.1 adds the capability to manipulate MPLS labels and supports multiple Flow Tables

Multiple Flow Tables and Group Table

OpenFlow v1.1





- An OpenFlow router supports one or more flow tables and a group table
- An OpenFlow controller can add, update, and delete flow entries
- Each flow entry consists of match fields, counters, and a set of instructions to apply to matching packets
- Matching starts at the first flow table and may continue to additional flow tables
  - If a matching entry is found, the instructions associated with the specific flow entry are executed
  - If no match is found in a flow table, the outcome depends on the configuration: the packet may be forwarded to the controller, dropped, or may continue to the next flow table
- Instructions associated with each flow entry describe packet forwarding, packet modification (e.g., push/pop VLAN or MPLS Labels), group table processing, and pipeline processing (packets to be sent to subsequent tables)

Group Table

OpenFlow v1.1

- The ability for a flow to point to a group enables OpenFlow to represent additional methods of forwarding
  - Group entry contains action buckets
  - Each action bucket contains a set of actions to execute associated parameters
- Group types
  - "All"
    - Process all buckets (e.g., multicast)
  - "Select"
    - Process one of the buckets (e.g., load balancing – similar to LAG)
  - "Indirect"
    - A group with one bucket
  - "Fast failover"
    - Process the first "live" bucket (e.g., active/standby LAG/LSP)