100Gigabit and Beyond: Increasing Capacity in IP/MPLS Networks Today

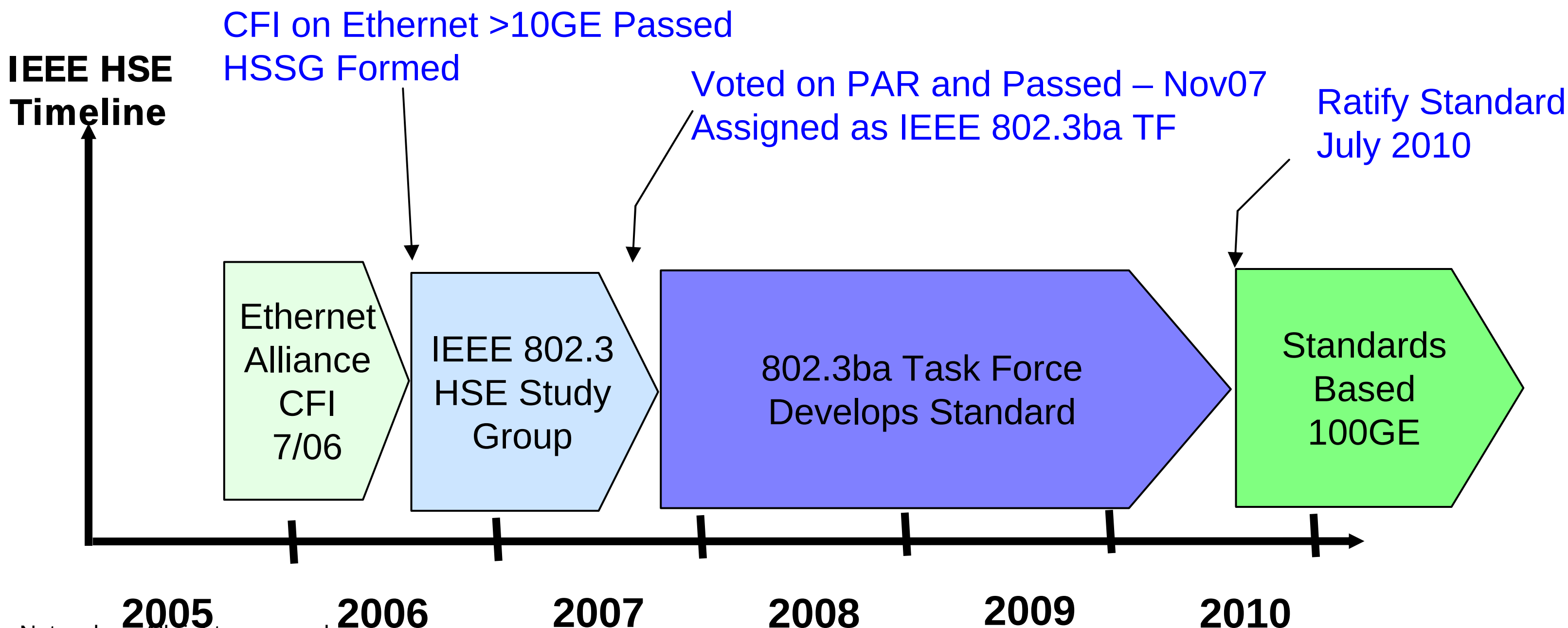
Rahul Vir
Product Line Manager
Foundry Networks
rvir@foundrynet.com

Agenda

- 40GE/100GE Timeline to Standardization
- Why We Need Load-Sharing
- Methods to Boost Capacity
- Methods for Efficient Utilization
- Summary

40GE/100GE Timeline to Standardization

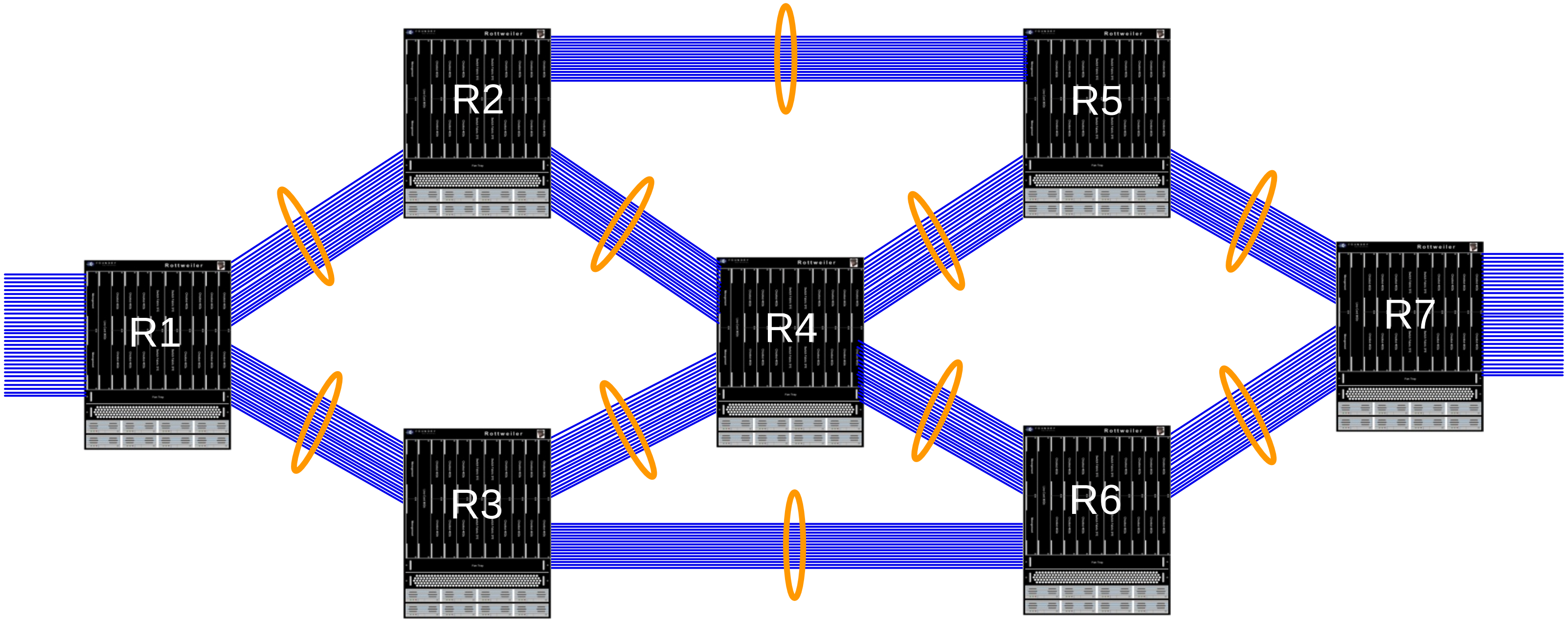
- The Ethernet Alliance sponsored the Call For Interest in July 2006
  - CFI approved and High Speed Study Group created
    - First Meeting held at September 2006 IEEE interim plenary
- High Speed Ethernet (HSE) Study Group given 6 months to develop PAR
- Process extended 6 Months due to push for 40GE addition
  - 100 GE for 100m MMF, 10km SMF and 40km SMF
  - 40 GE for Backplane, 100m MMF; 10km SMF added during March 08 Plenary



Current Expectations on Higher Capacity Ethernet

- Transition from 10GE to 40GE/100GE is intended mainly for the Core networks, Transit Networks, Data Center and IXP initially
- 40GE may be technologically achievable today but may not offer sufficient performance or cost benefits to warrant deployment versus 100GE
- The timeline to standardize is the same for both 40GE or 100GE
- Predict 45 nm to make 100GE solutions feasible by 2009
- IEEE 40GE/100GE expected to be ratified July 2010

Scale Beyond 10G/100G Ethernet, NOW



- Ever increasing demand for bandwidth in backbones, transit links, Internet peering points, data centers, …
- 100 Gigabit Ethernet is still ~24 months away
- OC-768 POS for many providers is an unaffordable alternative
- **Equal Cost Multi-Path (ECMP) with nx10GE Link Aggregation Groups (LAG) is a far more affordable way of scaling capacity**

Load Sharing Benefits

- Need more bandwidth
  - Utilize investment in existing infrastructure
  - Ability to add bandwidth in small increments
  - Cost-effectively add bandwidth
- Increased protection
  - End to End protection with diverse paths
  - 1+N link protection
  - Avoid idling of backup paths
- Allow scaling beyond 100G today
- Continued benefit after 40GE/100GE standardization
  - Many core/transit networks carrying over 100Gbps between critical nodes today
  - These bandwidth requirements expected to grow

Factors Affecting Load Sharing

- **Protocols:** Determine multiple paths for ECMP
  - Routing Protocols: IGP, BGP
    - Provide path diversity
- **Link Aggregation:** Offer multiple links for load-sharing
  - Link Aggregation/bundling/trunks
    - Provide link diversity
- **Data Forwarding:** Decision on how packets are load-shared
  - Load Balancing Algorithm
    - Provide efficient utilization
  - Fields in the packet used for load balancing
    - Ability to tune to various traffic types

Methods to boost capacity

Routing Protocols ECMP

- Routing Protocols determine multiple equal cost paths to a destination
  - IGP (ISIS/OSPF) ECMP:
    - Affects paths taken by IP traffic
    - Affects paths taken by MPLS LSPs
      - LDP paths follow IGP topology
      - RSVP-TE LSPs follow IGP and IGP-TE topologies
  - BGP ECMP:
    - Affects paths taken by IP traffic
    - Affects paths taken by IP & IP-VPN traffic in MPLS networks
      - Multiple equal cost BGP next-hops reachable by diverse LSPs
      - Multiple LSP paths to a BGP next-hop

Routing Protocols ECMP Considerations

- Number of ECMP paths per prefix supported by a router
  - More paths give better path diversity

- Support of ECMP with link aggregation
  - Very common that each path can contain LAG groups
  - LAG bandwidth changes should optionally be automatically reflected in Layer 3 interface metrics allowing routing protocols to choose better paths

- Does the router support even distribution over any number of paths?
  - For better utilization of network resources, must support even distribution for any number of paths (2, 3, 4, 5, 6,…..)

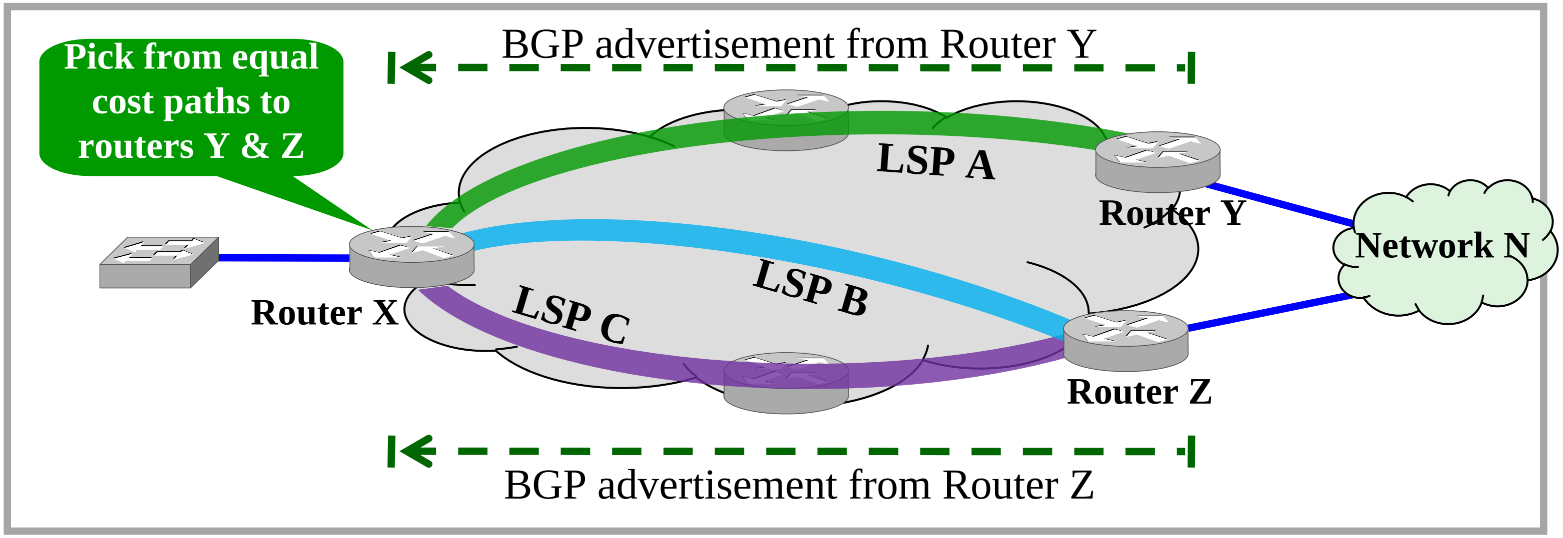
MPLS Signaling Protocols ECMP

- MPLS signaling allows multiple LSPs to the same destination

- RSVP-TE: Selects a path for a LSP from multiple equal cost paths that satisfy the LSP constraints, as determined through CSPF
  - Typical criteria used:
    - Hops: Pick the path with least number of hops
      - Less probability of failure
    - Least-fill: Pick the path with highest available bandwidth
      - Even spread of traffic
    - Most-fill: Pick the path with lowest available bandwidth
      - Leave room for higher bandwidth LSPs

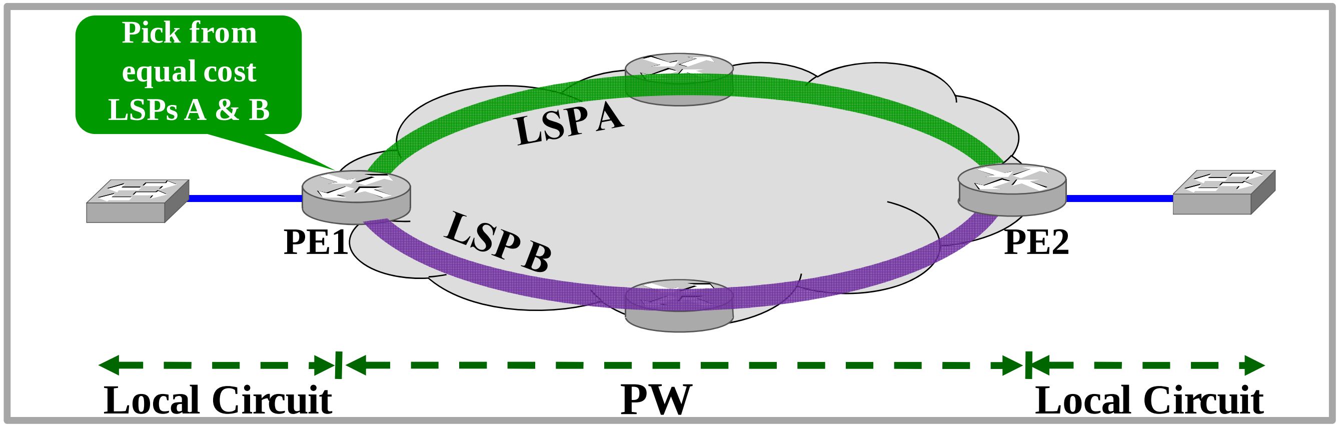
- LDP: Allows a prefix to be reachable through multiple equal cost label paths

IP Mapping to LSPs
For IPv4/v6 Routing and BGP/MPLS-VPNs



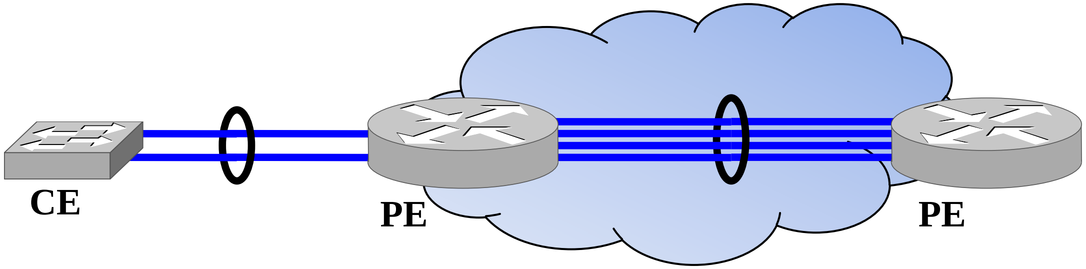
- Typical mapping criteria used:
  - Assign a prefix to single LSP
    - Better predictability
  - Map prefixes within a VRF to single LSP
    - Better operator control
  - Load-share on per flow basis
    - Better traffic distribution

PW Mapping to LSPs
For VPWS and VPLS



- Typical mapping criteria used:
  - Bind PW to least used LSP (LSP with lowest number of PWs)
    - Good distribution of traffic
  - Bind PW to LSP with most available bandwidth or same class of service
    - Useful for services with dedicated bandwidth requirements
  - Explicitly bind PW to LSP
    - Better operator control
  - PW traffic split across multiple LSPs
    - Better distribution of traffic based on flows

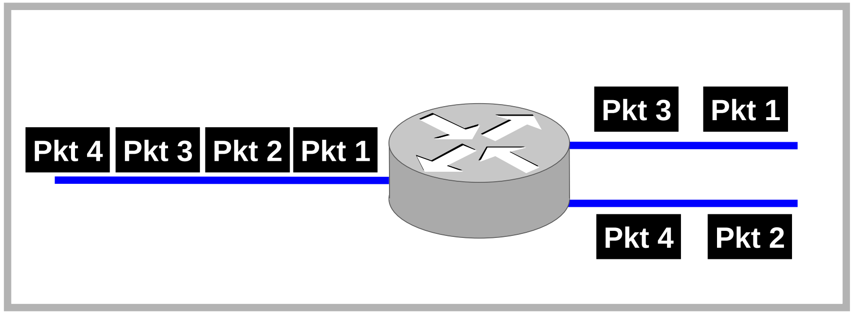
Link Aggregation
Options and Considerations

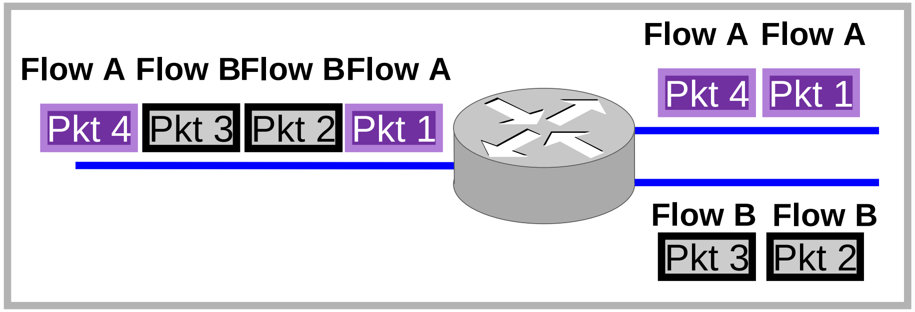


- Provides bundling multiple physical links between 2 devices
- Typically, higher layer protocols unaware of the link bundling
- IEEE 802.3 LAG (LACP) support
  - Dynamic configuration, provides increased availability
- Static Link Aggregation Groups (LAG) support
  - No need for control protocol, and works in multi-vendor scenario
- LAG capacity
  - Number of links in a LAG
    - Provide 10G bundling to scale beyond 100G bandwidth today
  - Number of LAG groups

Methods for efficient utilization

Load-Sharing in the Forwarding Plane
Common Schemes





Packet Based Forwarding

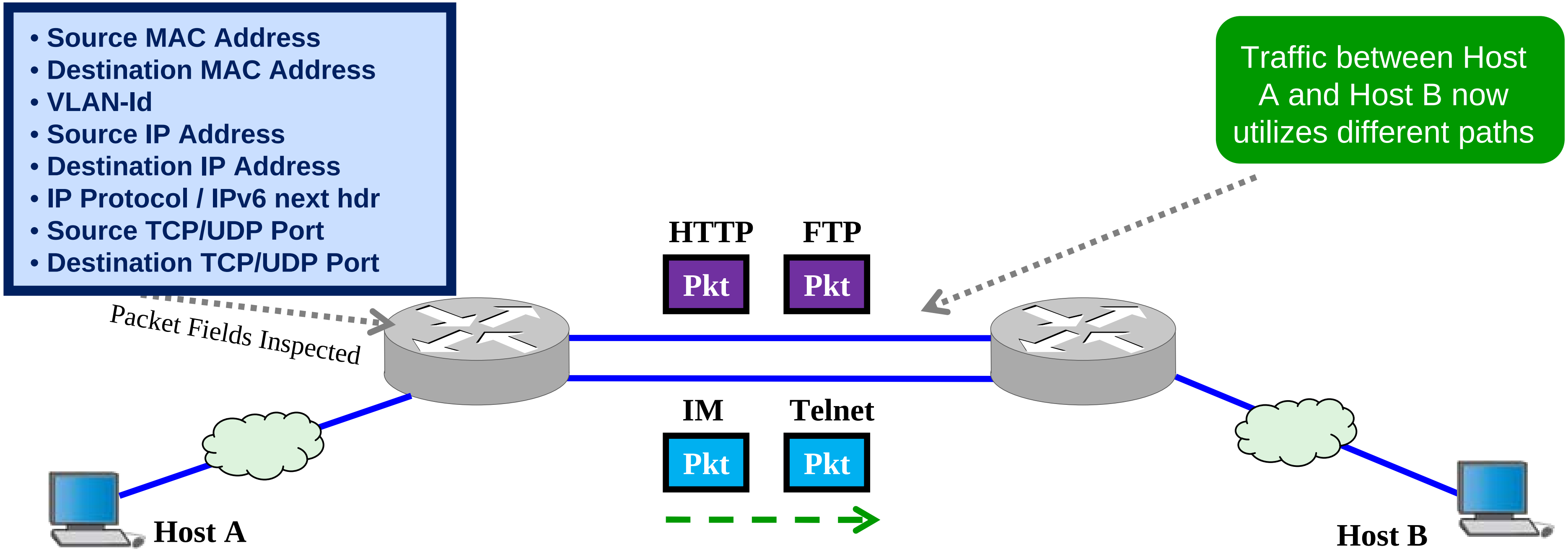
- Each packet sent on the next link
- Perfect load balancing
- Potential packet reordering issues
- Possible increase in latency and jitter for some flows

Flow Based Forwarding

- Identifies packets as flows
  - Based on packet content such as IP header
- Keeps flows on the same path
  - Maintains packet ordering
- Hashing is one of the most popular load sharing scheme for flow based forwarding

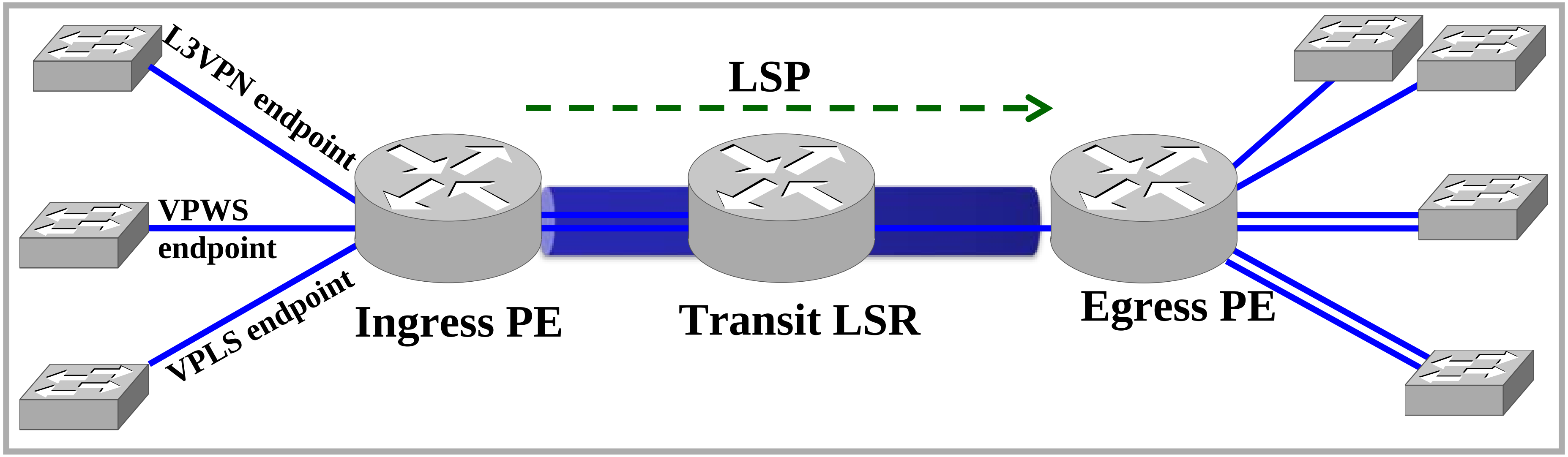
Load Sharing for Layer 3 Flows
IPv4 and IPv6

- Flows based on Source IP & Destination IP addresses
  - Works in most scenarios
  - Issue: Traffic between 2 hosts gets relegated to one path
    - Can lead to over-utilization of one path
- Flows based on L2, L3 and L4 information
  - Better traffic distribution for applications between 2 hosts



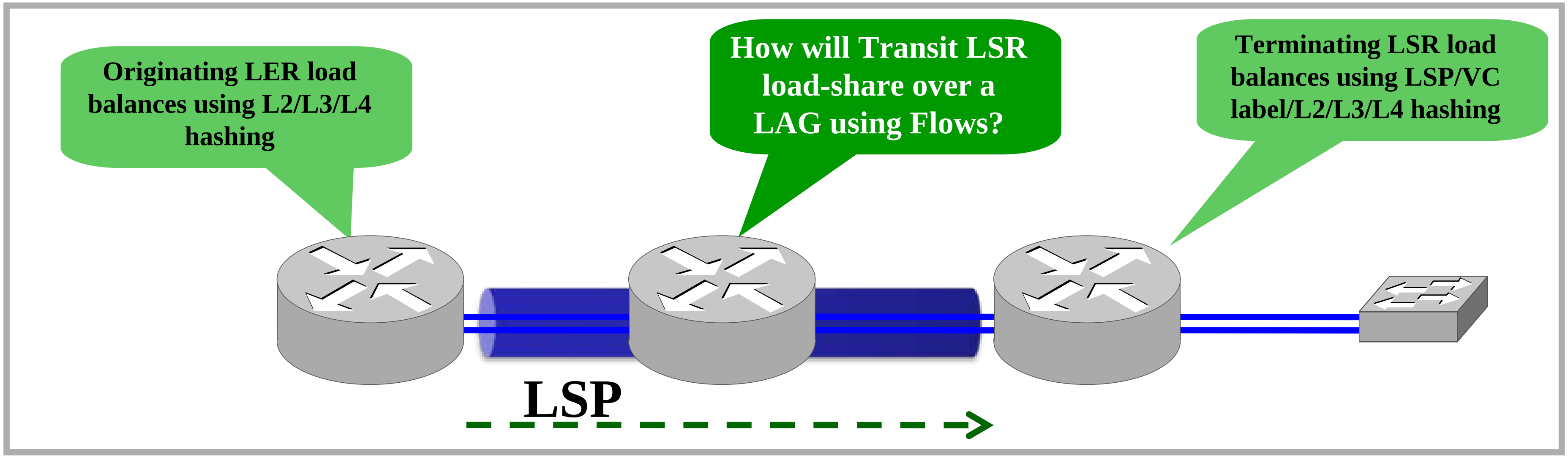
Load Sharing on MPLS PE router

Ingress and Egress PE



- At Ingress PE (packets entering a MPLS LSP):
  - Can load share across multiple LSPs and multiple links in a LAG
    - Apply load sharing principles of L2 & L3 flows
- At Egress PE (packets exiting a MPLS LSP):
  - Can load share per LSP/VC label:
    - High usage PWs/VPN labels will over-utilize one path
  - Per flow: Better distribution of traffic
    - Using LSP/VC label and load sharing principles of L2 & L3 flows

Load Sharing on MPLS LSRs
Packet Speculation



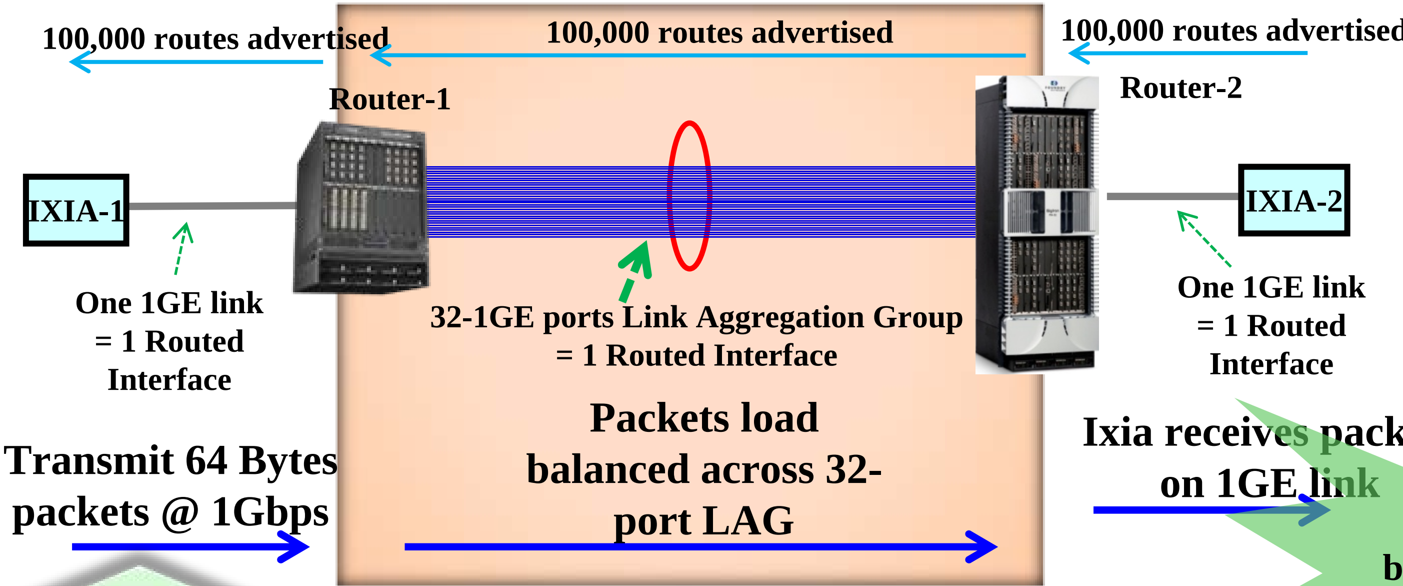
- Transit LSRs (and PHP nodes) have no information on packet payload
- Transit LSR speculates on the packet type
  - Checks first nibble after bottommost label
    - If 4/6, speculates on packet as IPv4/IPv6
    - Else (optionally) speculates on packet as Ethernet
  - Can now load-share using “LSP Label/VC label/L2/L3/L4 headers”

Load Balancing Algorithm Considerations for flow based forwarding

- A good load balancing algorithm is essential for efficiently utilizing the increased capacity of LAG/ECMP paths
  - Must Distribute Traffic Evenly
    - For example, a good algorithm needs to ensure that effective capacity of a 32-port 10GE LAG should be close to 320Gbps
- Other Considerations:
  - Number of fields in packet header that can be used for load balancing
    - More the fields, better the distribution
  - Number of hash buckets
    - More hash buckets result in better distribution
  - Minimal correlation of ECMP with LAG
    - Correlation will lead to over-utilization of some paths/links
  - Can treat each packet type differently
    - For example, L2 & L3 flows have to be treated differently

Use Case: Load Sharing across a 32-port LAG Group

IPv4 Traffic Distribution Test



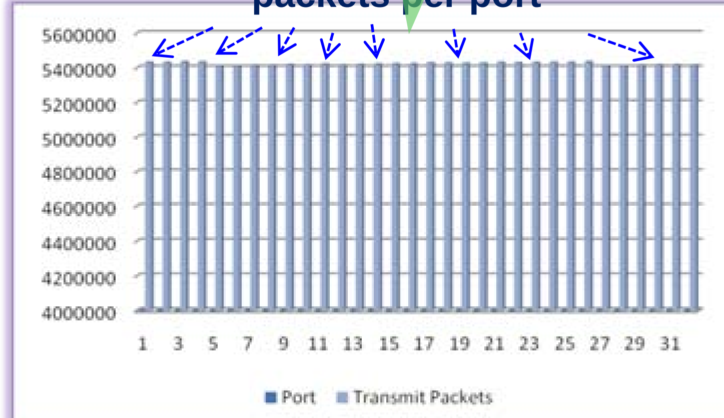
Very small difference between packet rates across links

Random Distribution
127,000 Source IP addr.
16,645,890 Destination IP addr.

Monitor Traffic Distribution on 32-port LAG on Router-1

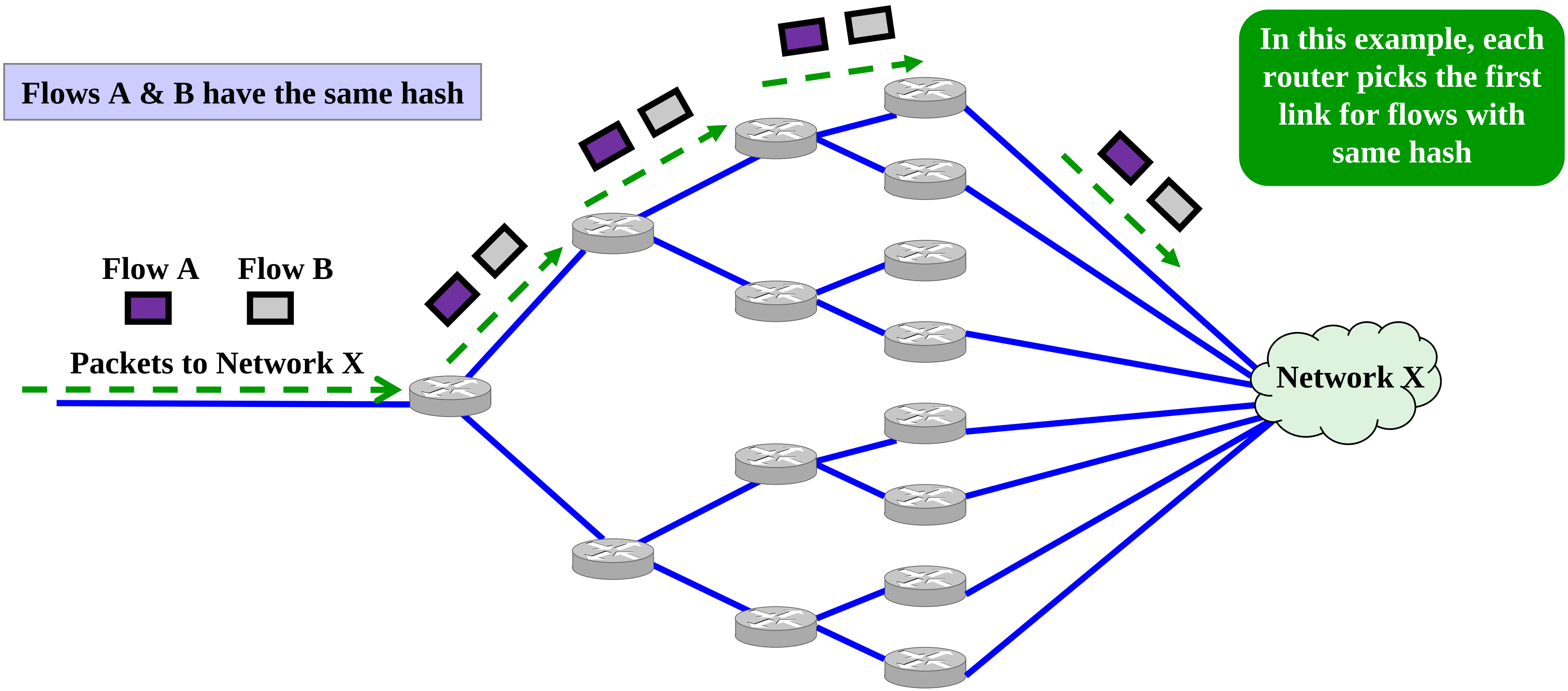
Traffic distributed evenly across 32-port LAG group

Number of transmitted packets per port



Hash based forwarding issues and solutions
Polarization Effect

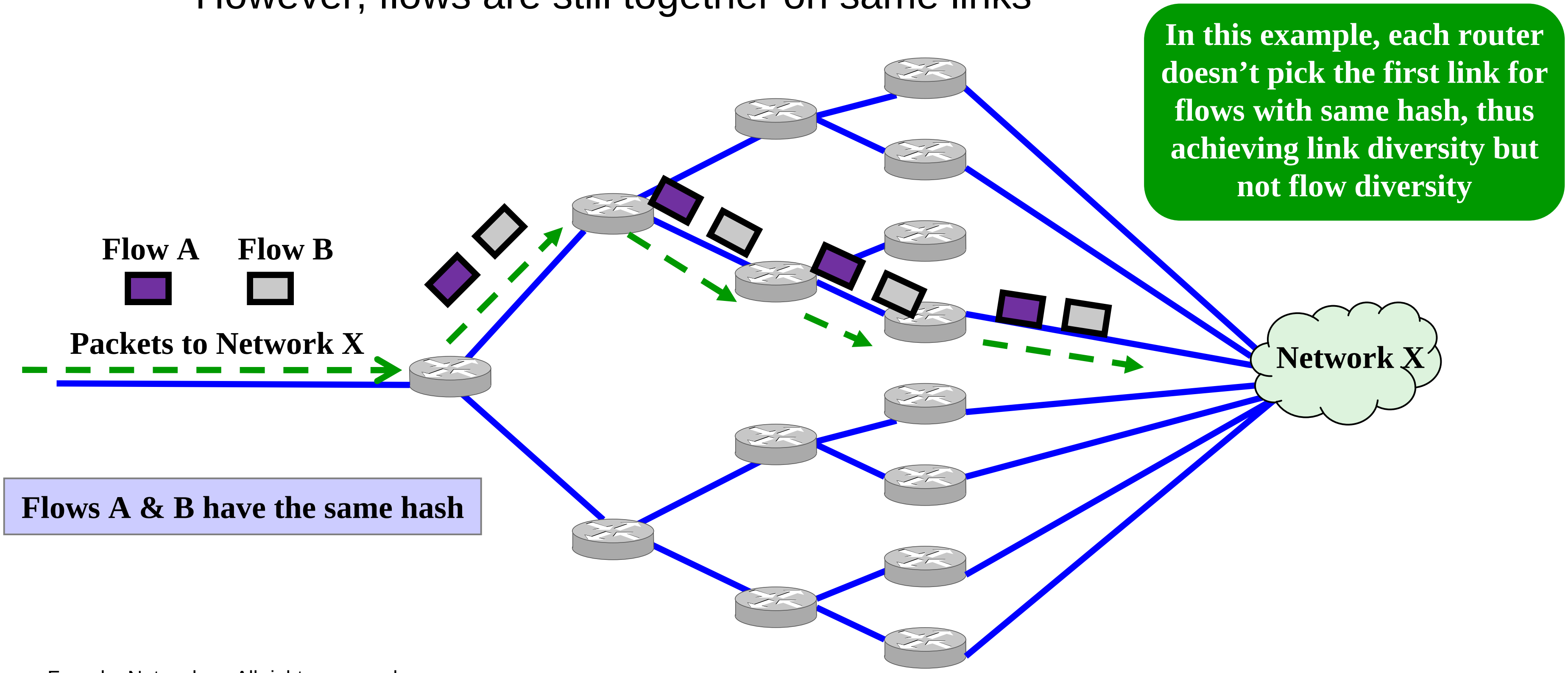
- In a multi-stage network, similar routers pick the same path for flows with identical hash
  - Leads to over-utilization of some parts of the network



Hash based forwarding issues and solutions

Basic Hash Diversification (Neutralizes Polarization Effect)

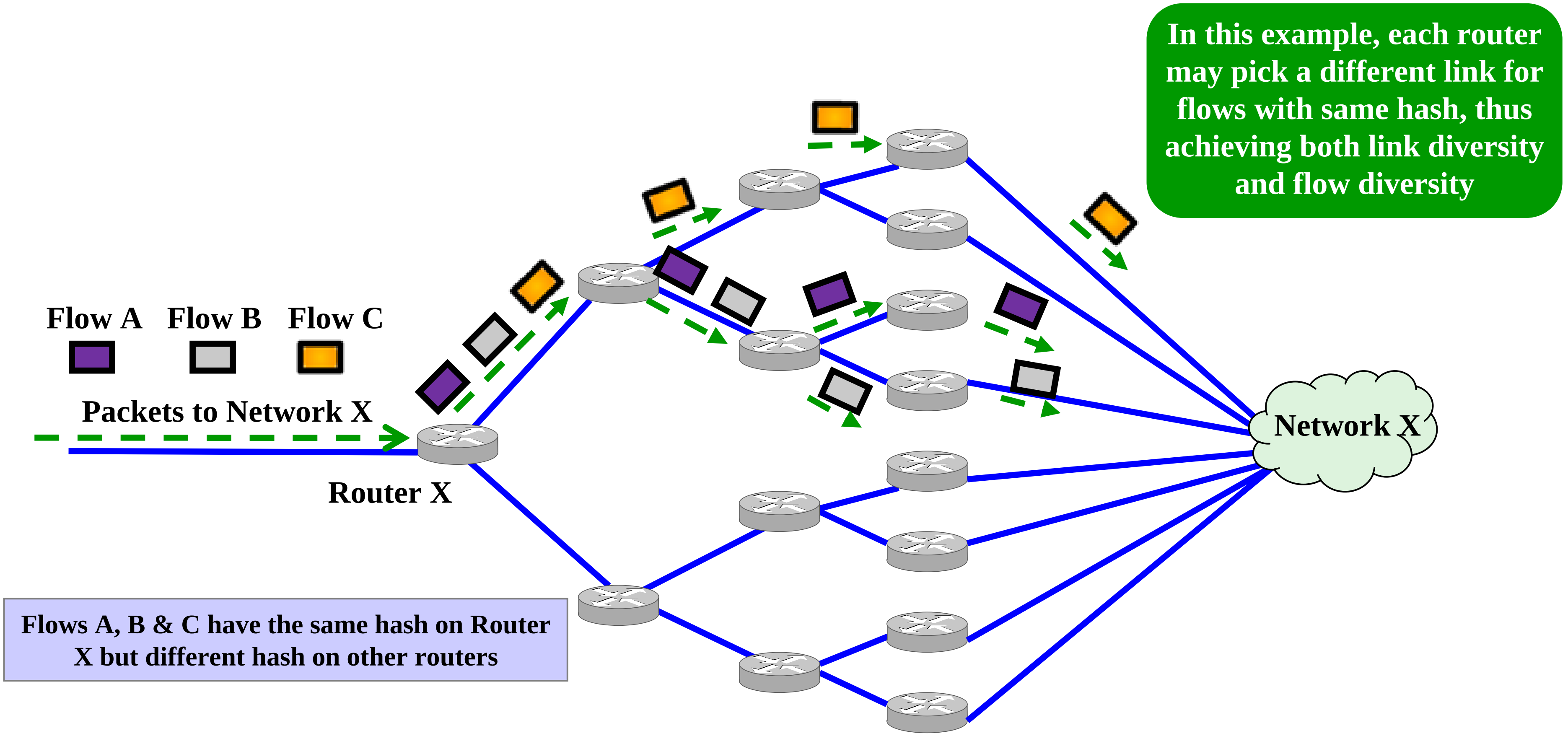
- Each router uses a unique-id per router in hash calculations
  - Alternatively, hashing using Source and Destination MACs may give comparable results in most scenarios
  - Similar routers now pick different links
  - However, flows are still together on same links



Hash based forwarding issues and solutions

Advanced Hash Diversification (Neutralizes Polarization Effect)

- Routers in each stage of the network run a different variant of the hash algorithm and neutralize polarization effect
  - Flows can now be distributed



Summary

- Multiple load balancing options to boost capacity at various layers
  - Increase throughput beyond the current limits of physical link capacity
  - Useful up to and even after 40GE/100GE standardization
  - Cost effective and efficient
- Load-Sharing improves network utilization
  - Efficient hashing algorithm determines the efficiency
  - Works over multiple paths and links
- Flow based forwarding offers many advantages for efficient utilization of the increased capacity
  - Watch out for polarization effects in multi-stage networks
  - Options are available to neutralize them
- Not a one size fits all approach
  - Choose optimal schemes based on traffic types and operator policy

Thank You!