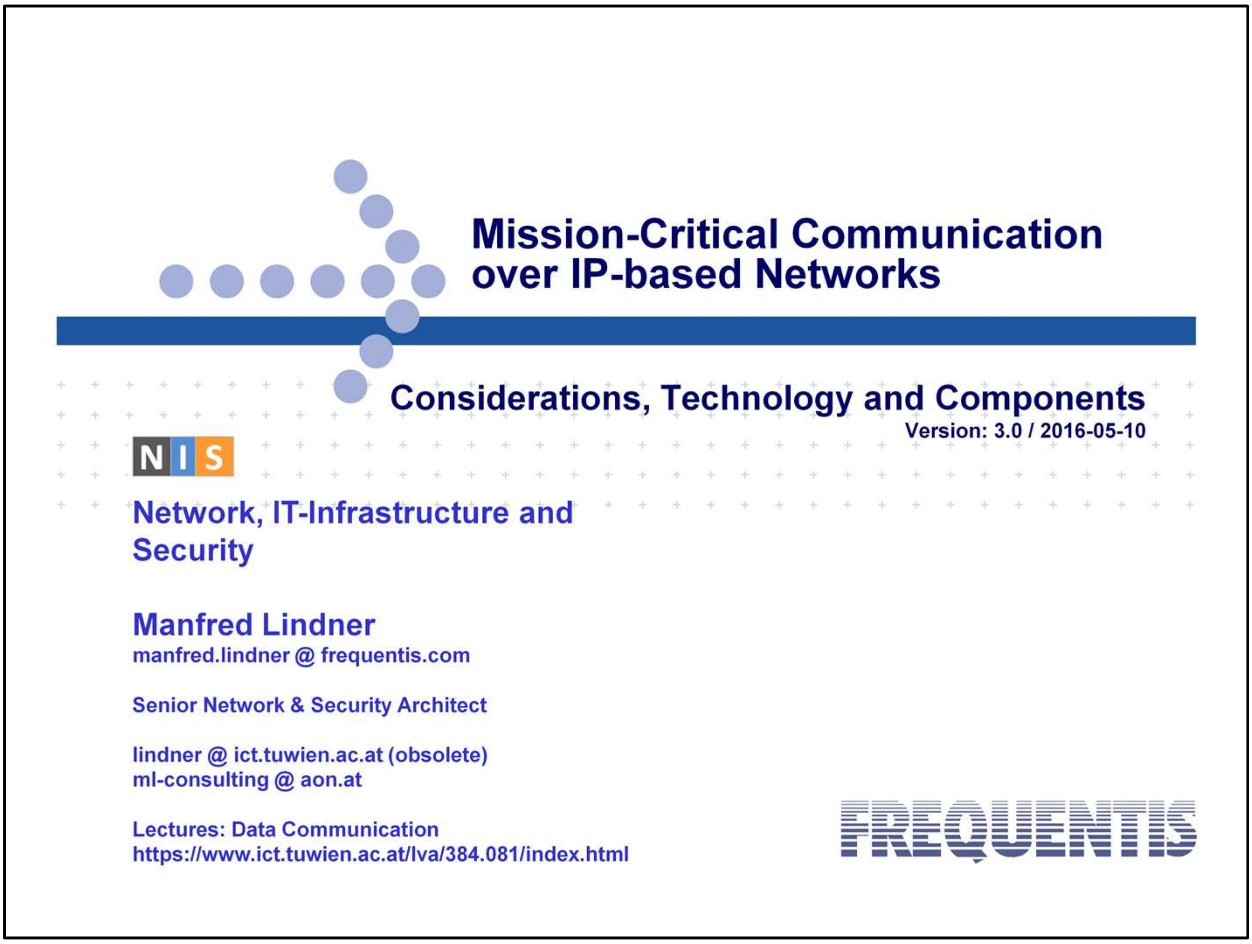


# Mission-Critical Communication over IP-based Networks

## Considerations, Technology and Components

Version: 3.0 / 2016-05-10

NIS

Network, IT-Infrastructure and Security

Manfred Lindner
manfred.lindner @ frequentis.com

Senior Network & Security Architect

lindner @ ict.tuwien.ac.at (obsolete)
ml-consulting @ aon.at

Lectures: Data Communication
https://www.ict.tuwien.ac.at/lva/384.081/index.html

FREQUENTIS

Changes to version 2.0:

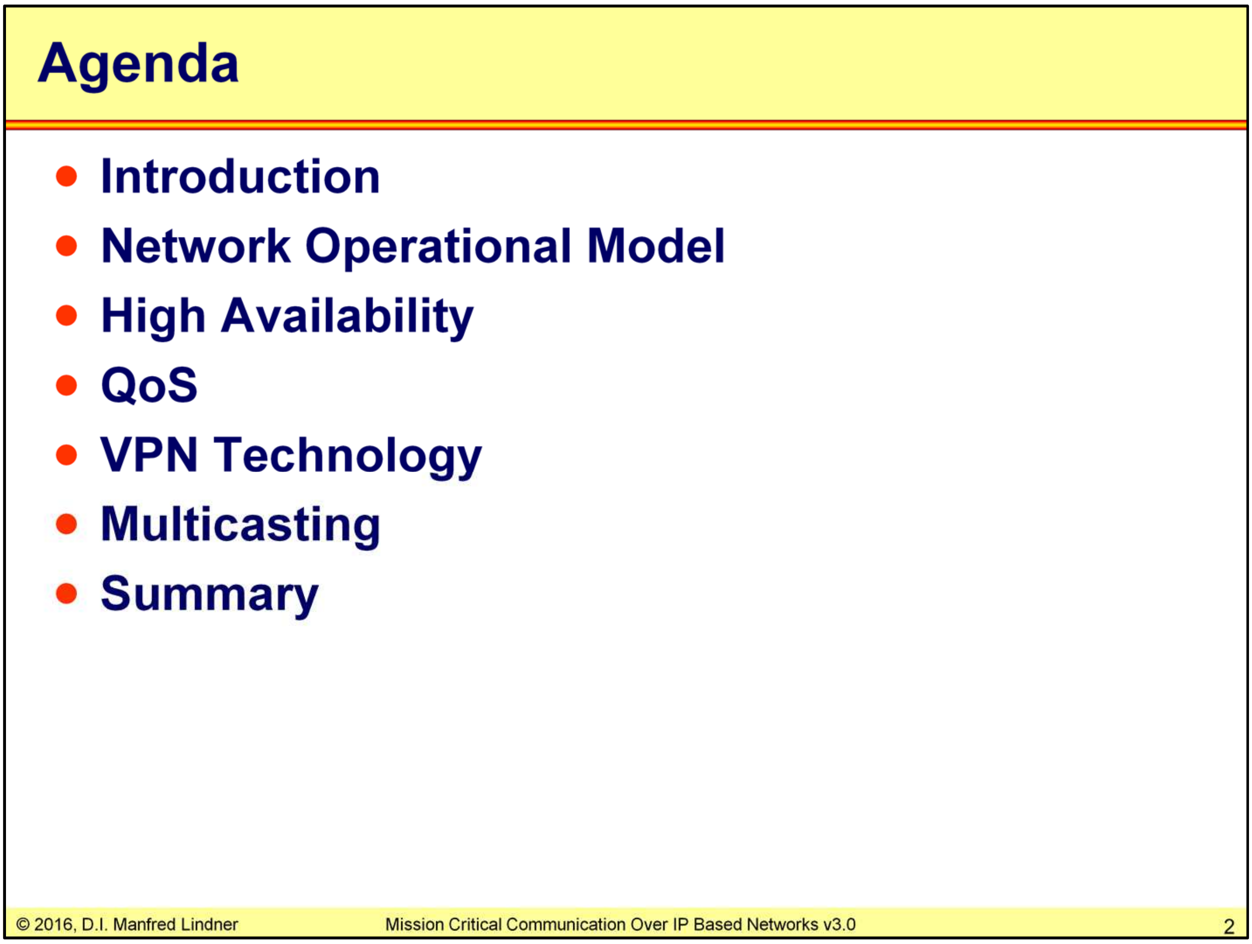
Chapter Multicasting was extended with new topics. IP multicast routing principles PIM-DM and PIM-SM are explained and multicast routing convergence is now covered.

Changes to version 2.1:

Pages 22, 23: additional content on slide

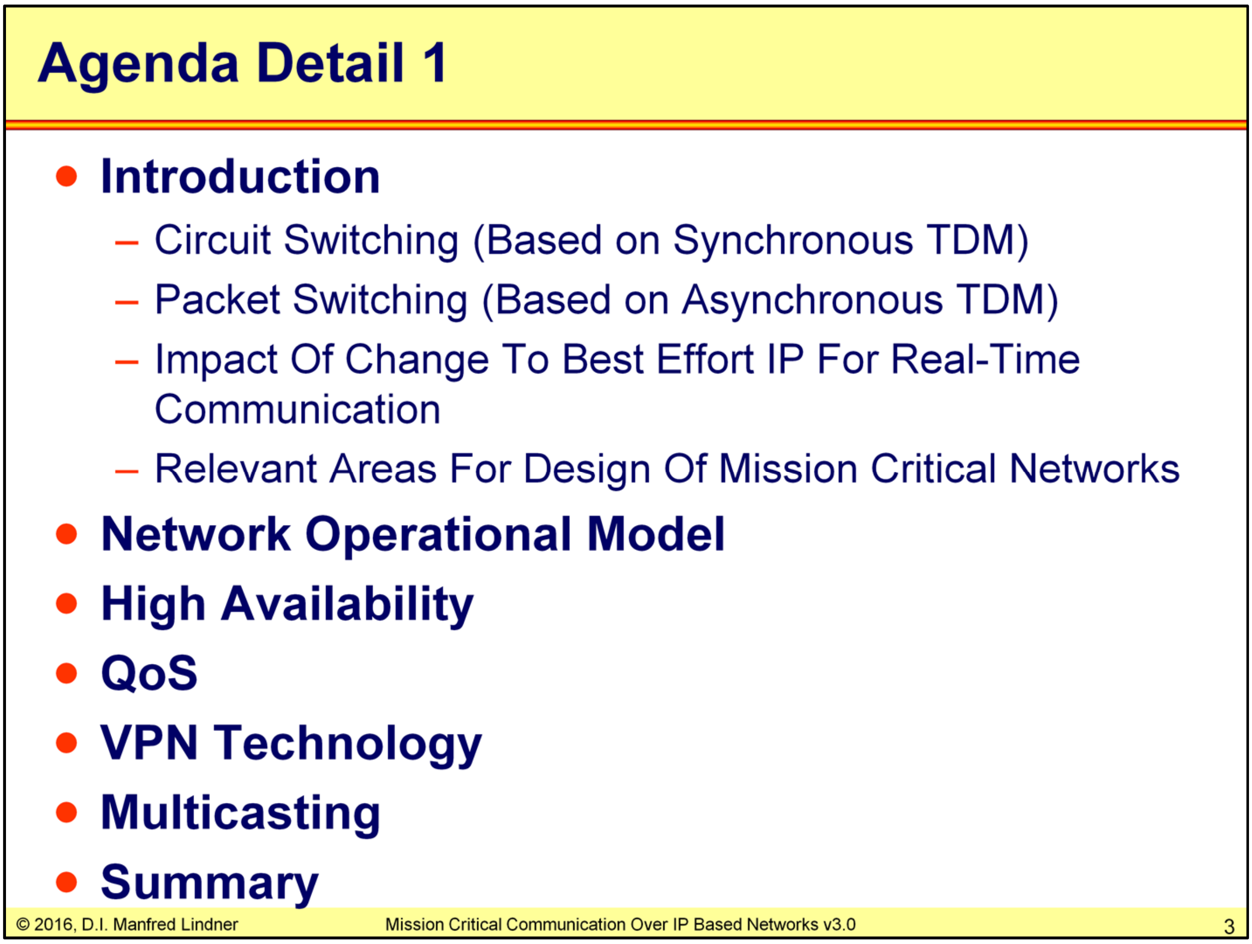
Changes to version 2.2:

Some enhancements in the QoS subchapter: new slides at the end of the chapter
Some typos removed and some text in the handout part improved.



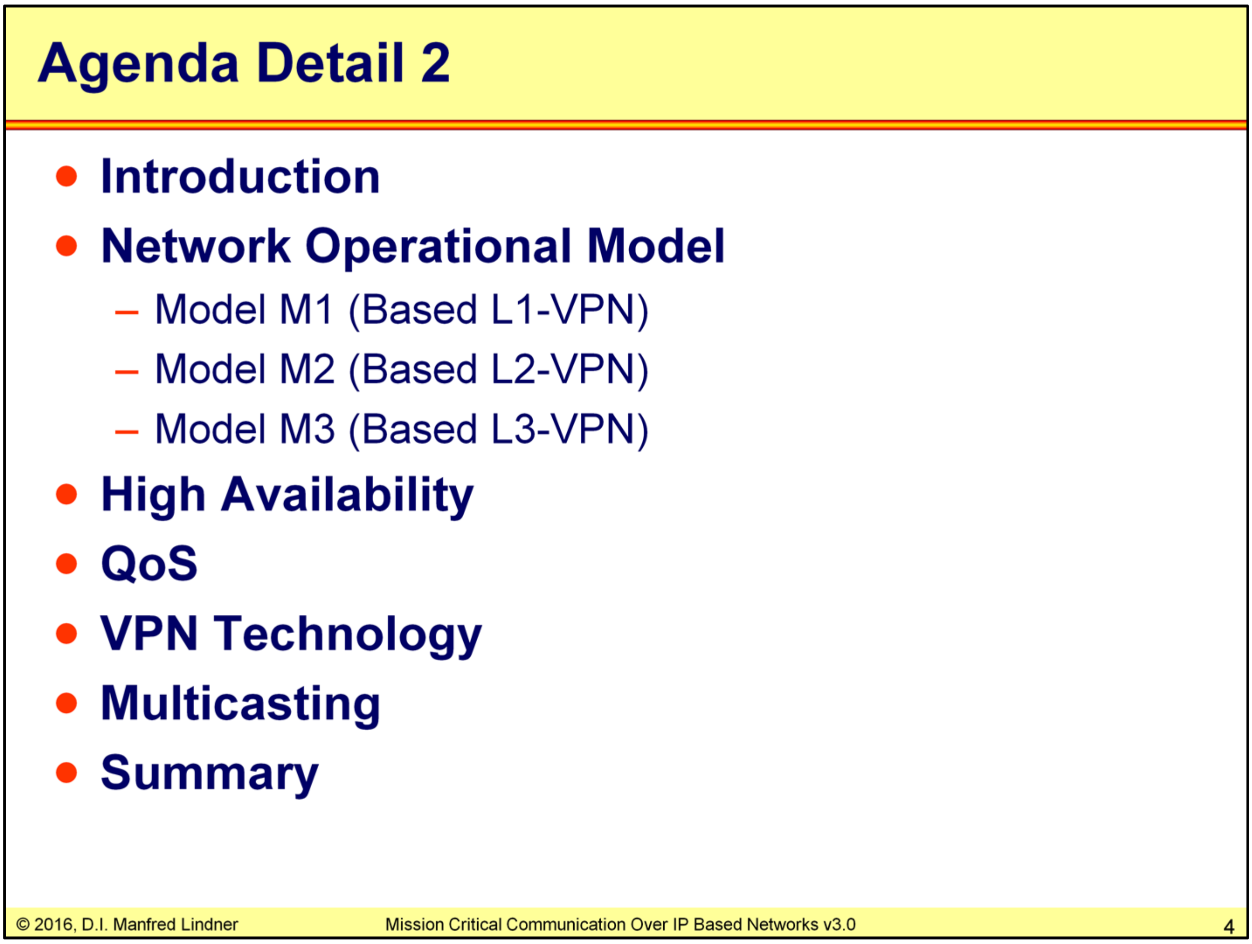
## Agenda

- Introduction
- Network Operational Model
- High Availability
- QoS
- VPN Technology
- Multicasting
- Summary



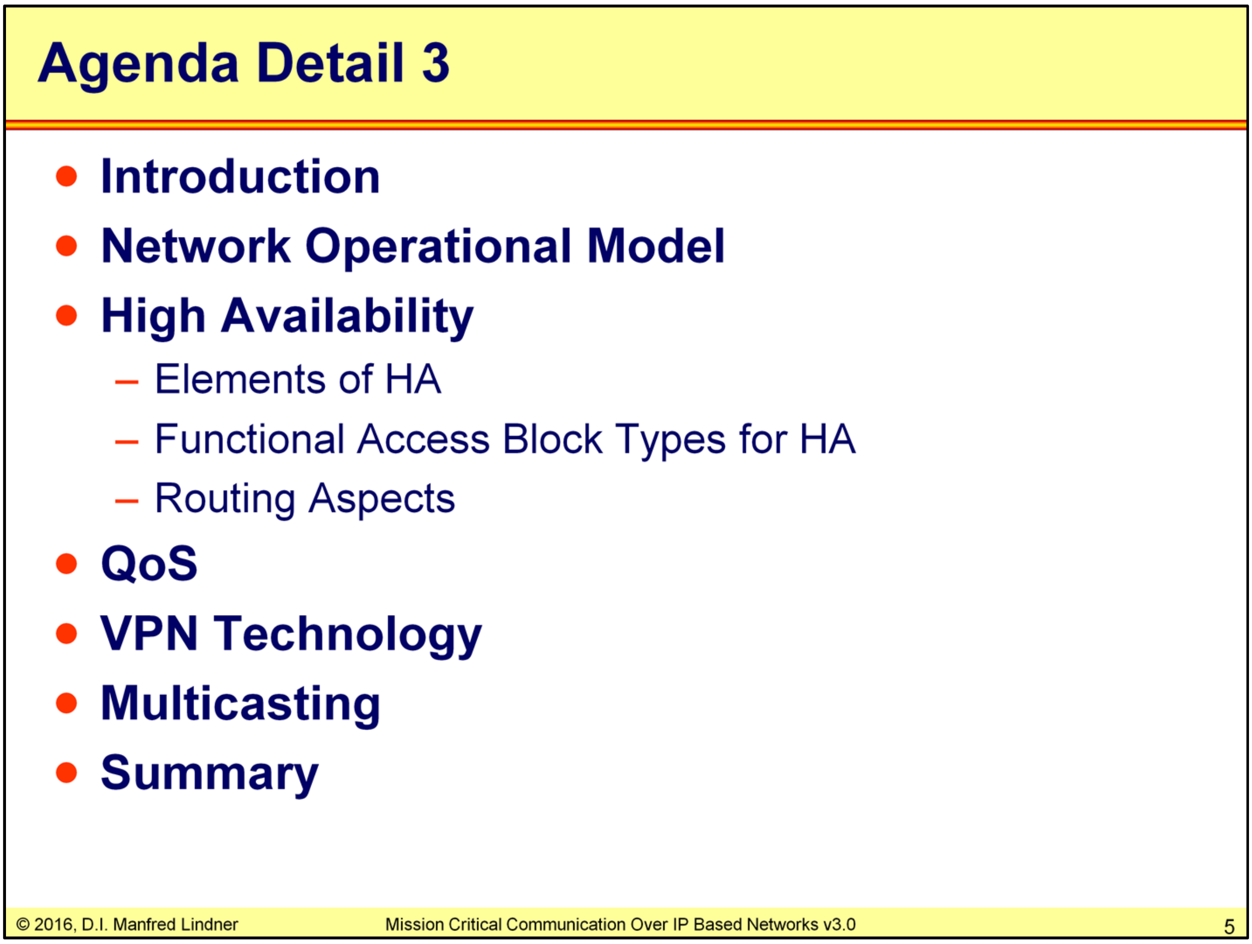
## Agenda Detail 1

- **Introduction**
  - Circuit Switching (Based on Synchronous TDM)
  - Packet Switching (Based on Asynchronous TDM)
  - Impact Of Change To Best Effort IP For Real-Time Communication
  - Relevant Areas For Design Of Mission Critical Networks
- **Network Operational Model**
- **High Availability**
- **QoS**
- **VPN Technology**
- **Multicasting**
- **Summary**



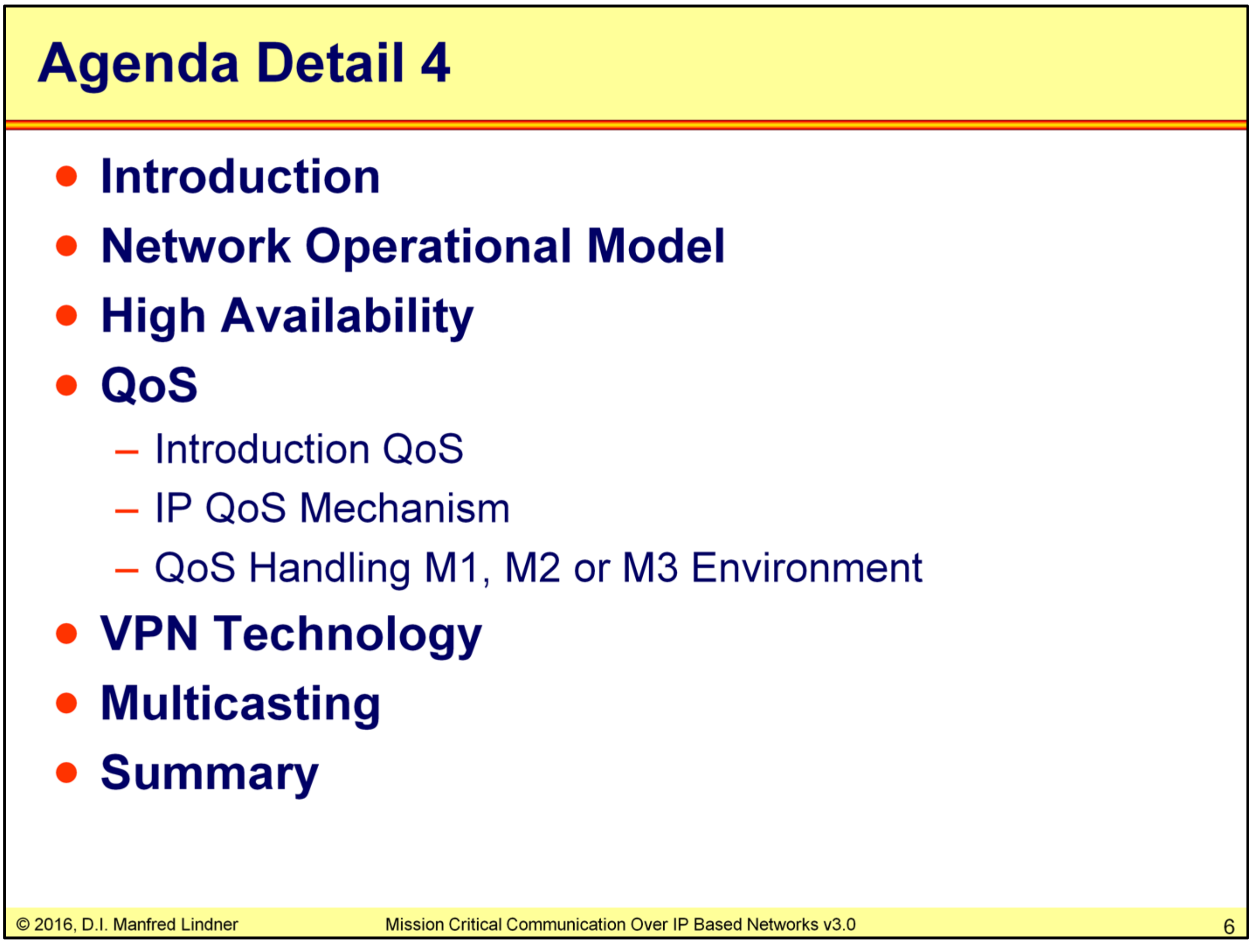
## Agenda Detail 2

- Introduction
- Network Operational Model
  - Model M1 (Based L1-VPN)
  - Model M2 (Based L2-VPN)
  - Model M3 (Based L3-VPN)
- High Availability
- QoS
- VPN Technology
- Multicasting
- Summary



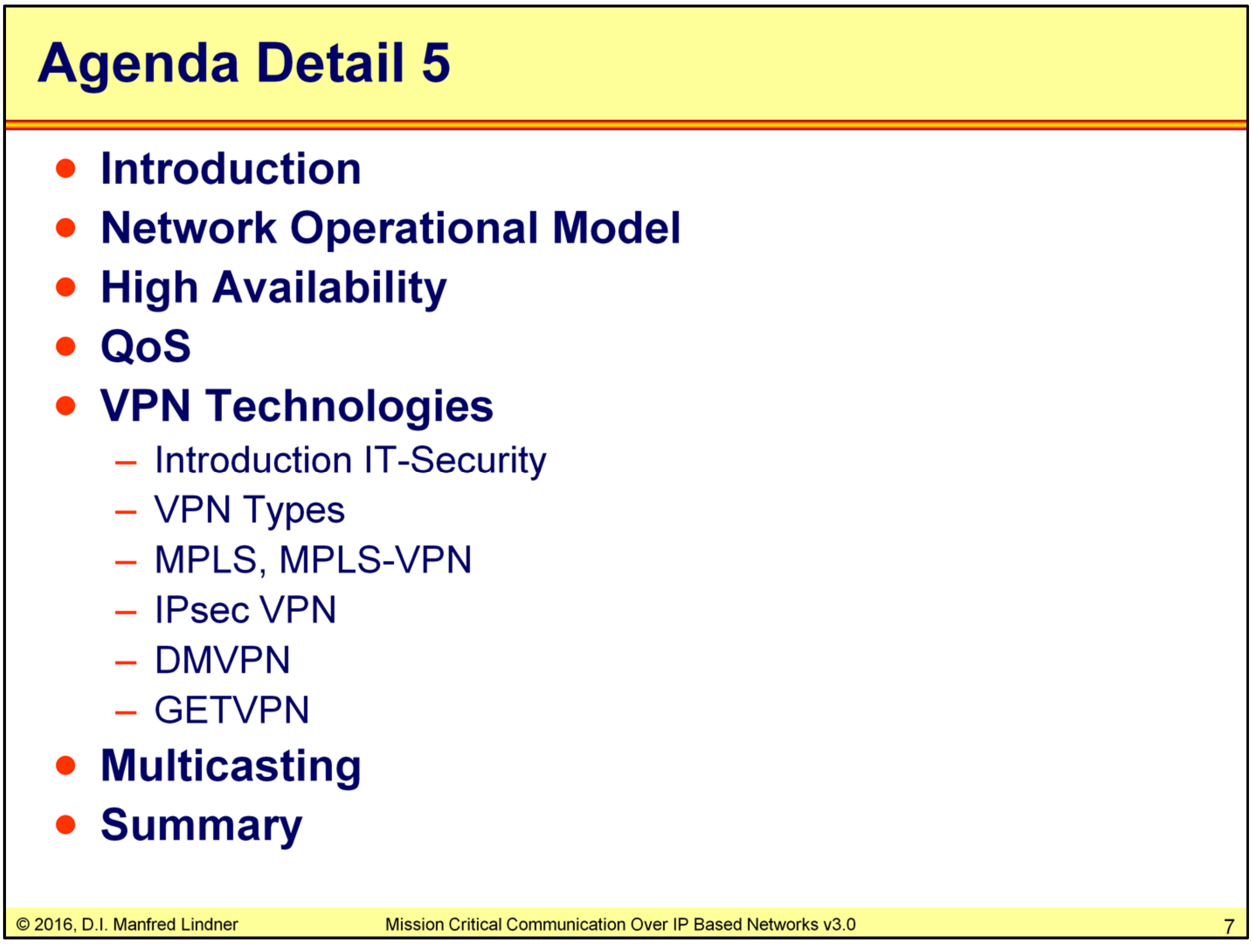
## Agenda Detail 3

- **Introduction**
- **Network Operational Model**
- **High Availability**
  - Elements of HA
  - Functional Access Block Types for HA
  - Routing Aspects
- **QoS**
- **VPN Technology**
- **Multicasting**
- **Summary**



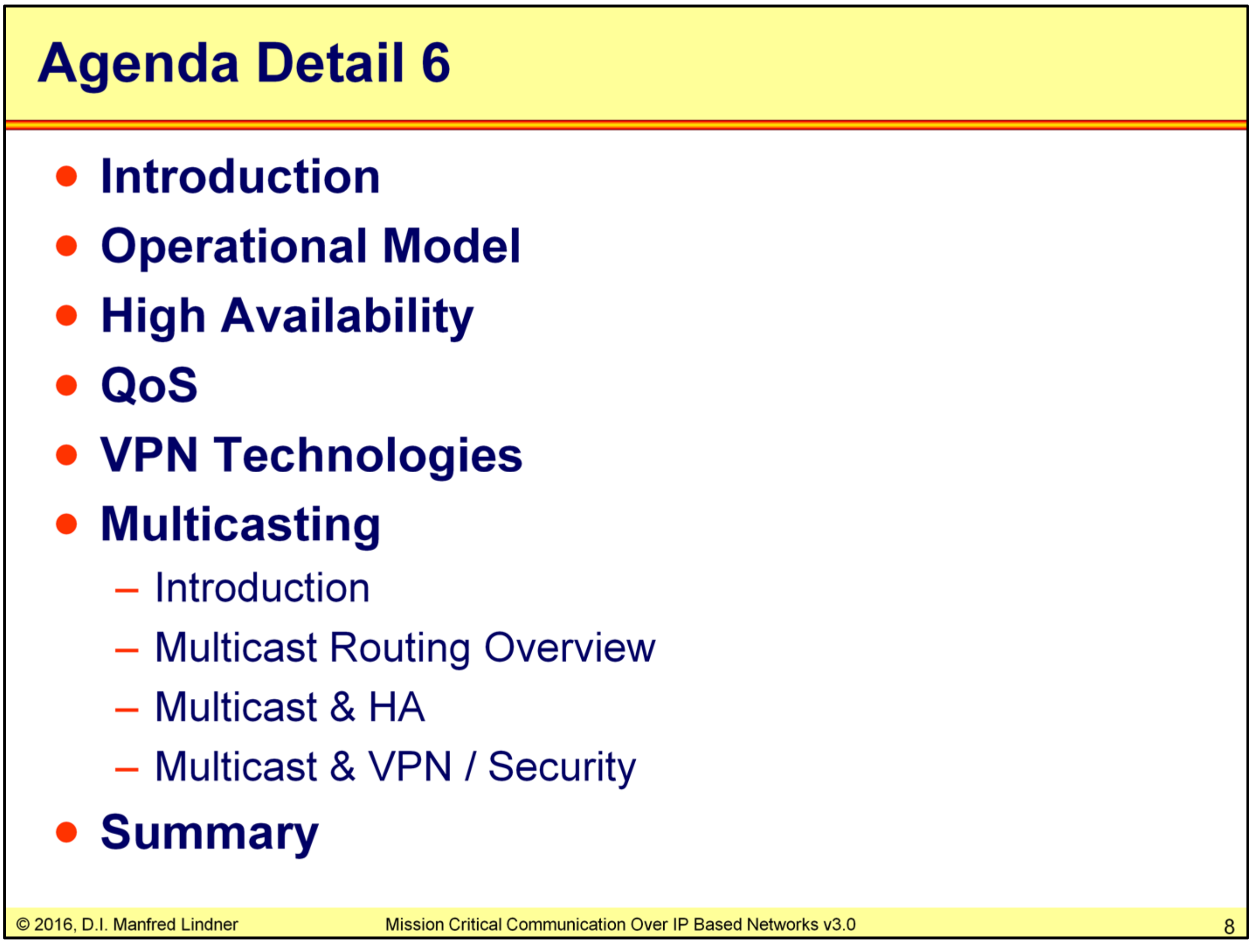
## Agenda Detail 4

- Introduction
- Network Operational Model
- High Availability
- QoS
  - Introduction QoS
  - IP QoS Mechanism
  - QoS Handling M1, M2 or M3 Environment
- VPN Technology
- Multicasting
- Summary



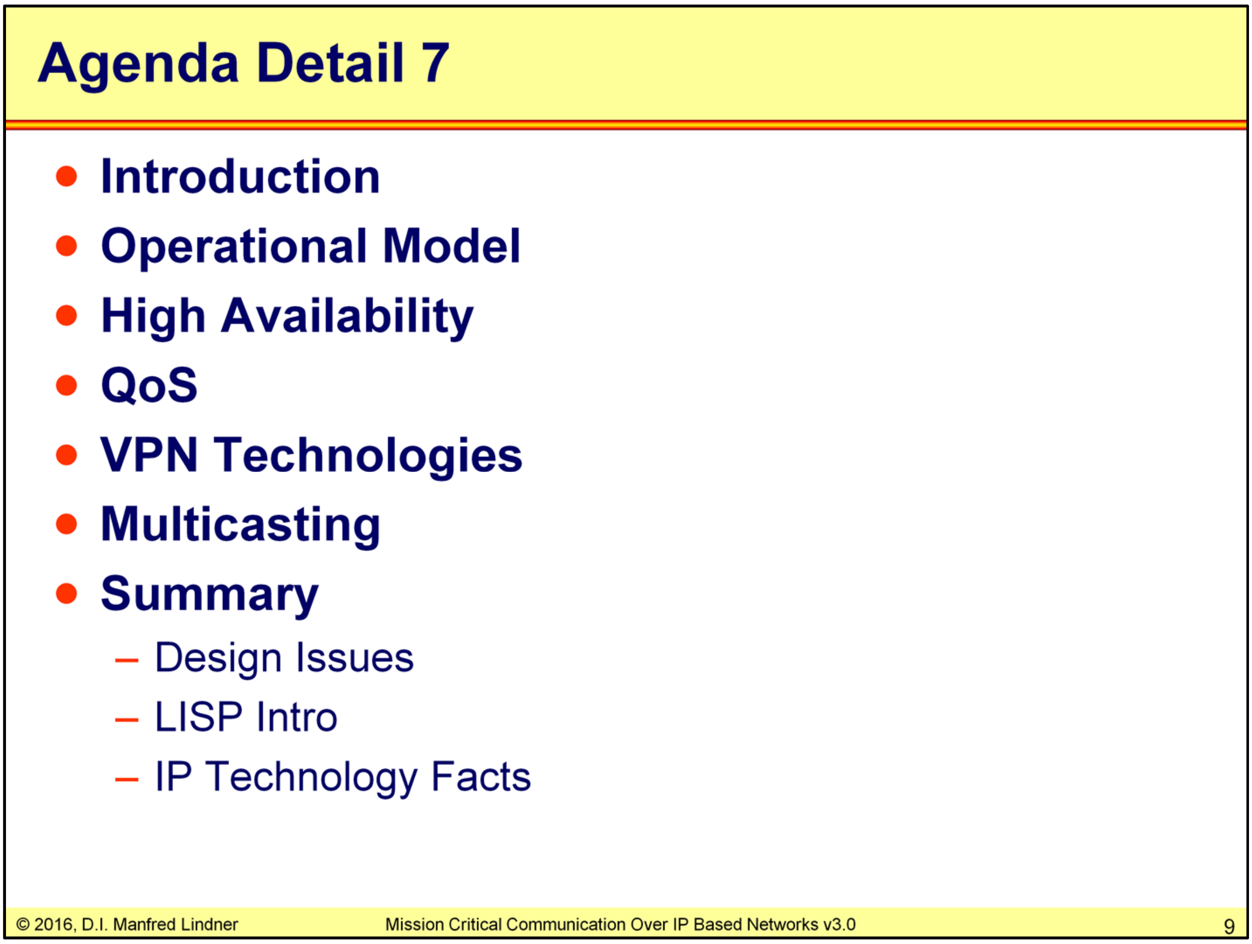
## Agenda Detail 5

- Introduction
- Network Operational Model
- High Availability
- QoS
- VPN Technologies
  - Introduction IT-Security
  - VPN Types
  - MPLS, MPLS-VPN
  - IPsec VPN
  - DMVPN
  - GETVPN
- Multicasting
- Summary



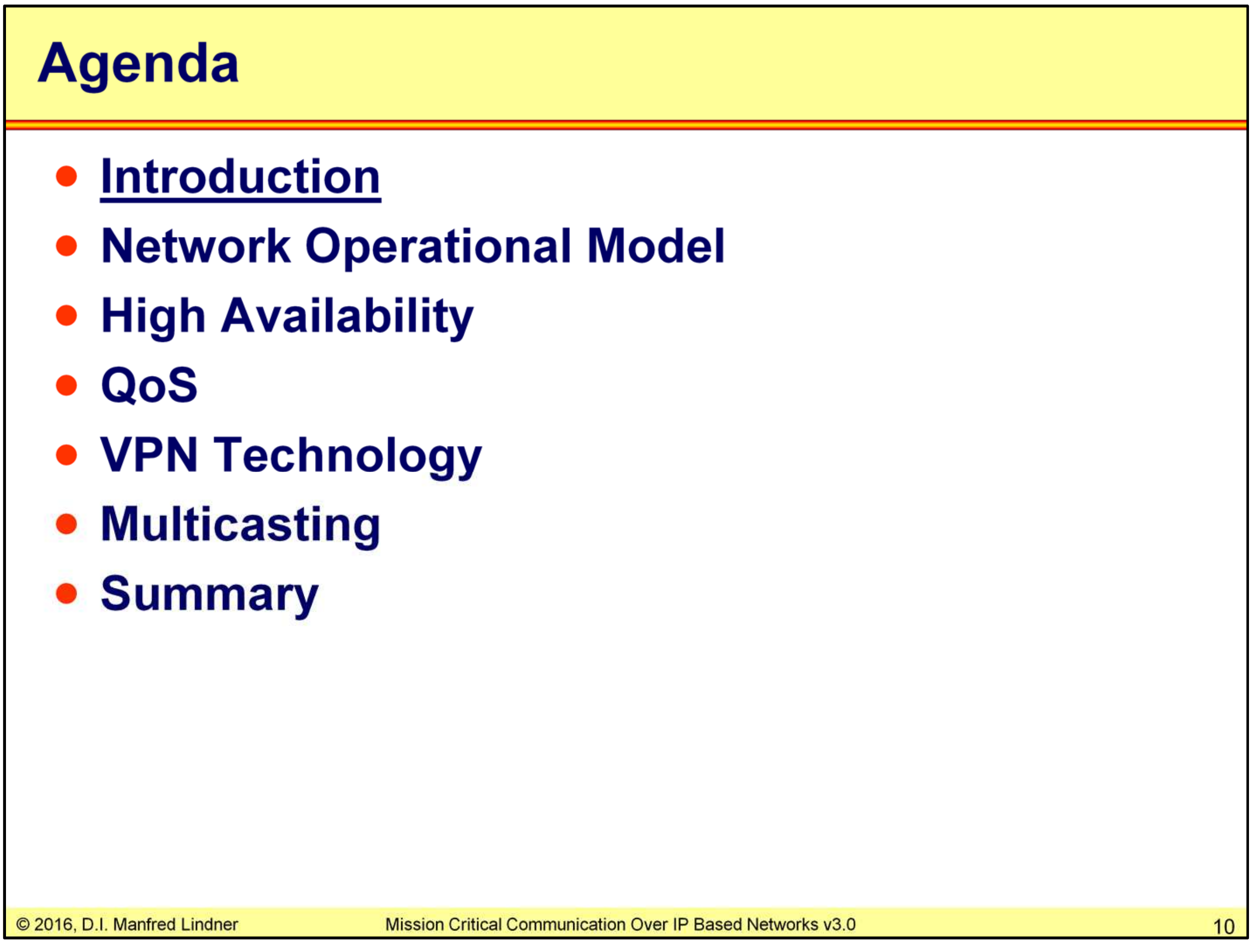
## Agenda Detail 6

- **Introduction**
- **Operational Model**
- **High Availability**
- **QoS**
- **VPN Technologies**
- **Multicasting**
  - Introduction
  - Multicast Routing Overview
  - Multicast & HA
  - Multicast & VPN / Security
- **Summary**



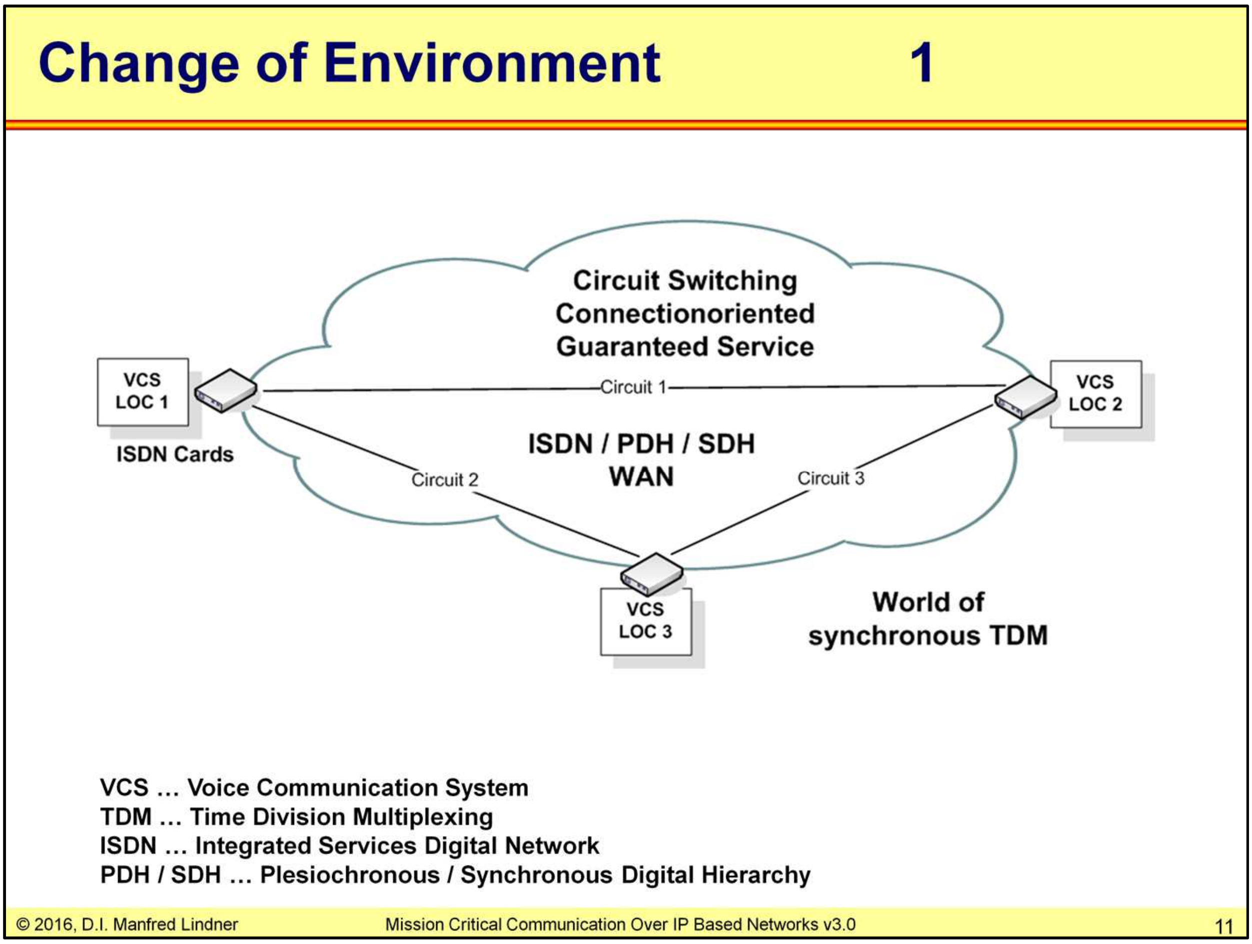
## Agenda Detail 7

- Introduction
- Operational Model
- High Availability
- QoS
- VPN Technologies
- Multicasting
- Summary
  - Design Issues
  - LISP Intro
  - IP Technology Facts



# Agenda

- <u>Introduction</u>
- Network Operational Model
- High Availability
- QoS
- VPN Technology
- Multicasting
- Summary



## Change of Environment 1

Circuit Switching
Connectionoriented
Guaranteed Service

VCS LOC 1

ISDN Cards

Circuit 1

VCS LOC 2

ISDN / PDH / SDH
WAN

Circuit 2

Circuit 3

VCS LOC 3

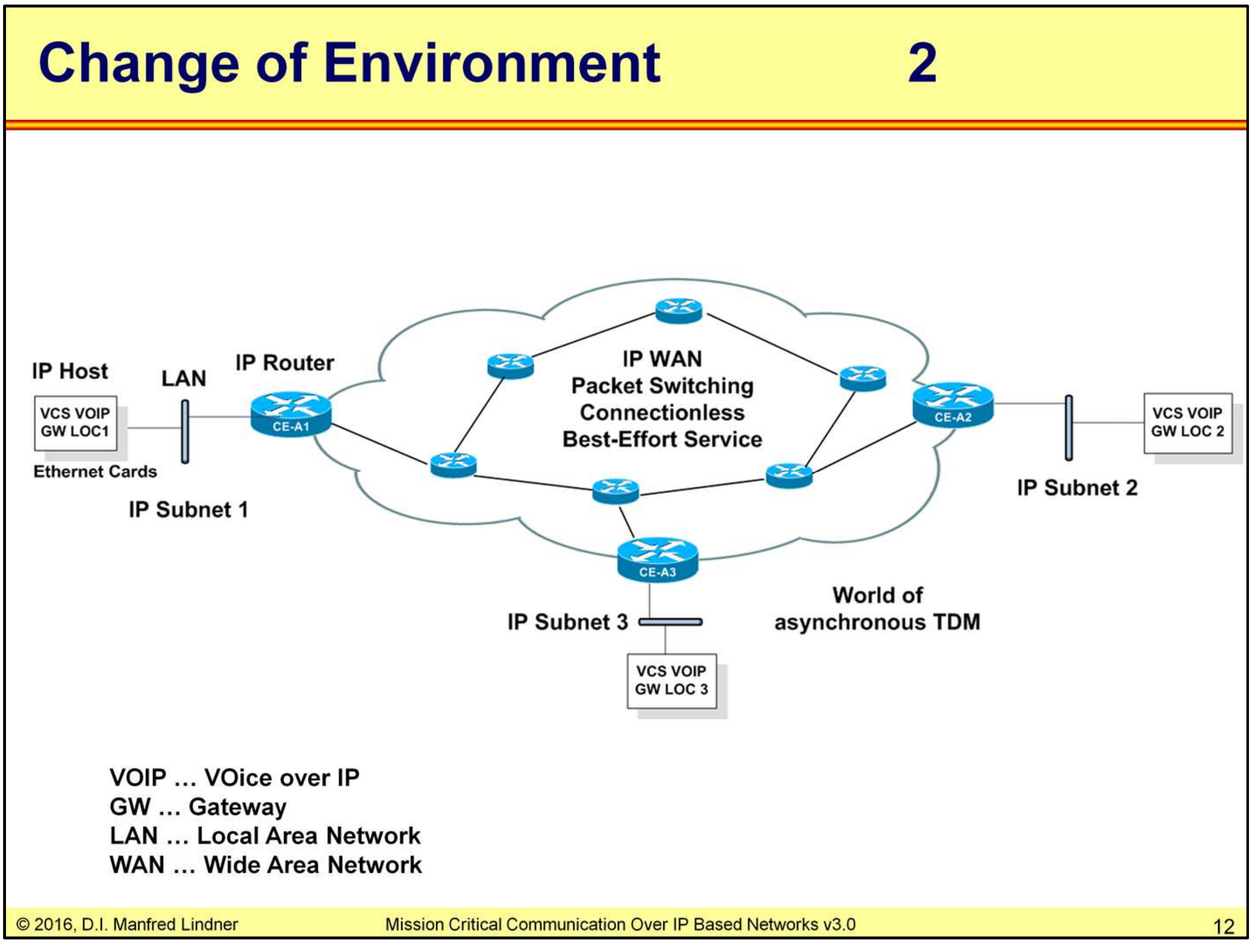
World of
synchronous TDM

VCS … Voice Communication System
TDM … Time Division Multiplexing
ISDN … Integrated Services Digital Network
PDH / SDH … Plesiochronous / Synchronous Digital Hierarchy

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 11

In the past and even nowadays mission critical communication very often bases on real-time voice transmission. Examples are voice communication in ATC (Air-Traffic-Control) area between controller in the tower and pilot of an airplane or in public safety area between dispatcher and rescue forces (fire brigade, police, ambulance). Special designed voice communication systems (VCS) based on real-time telephony switching in conjunction with radio communication were and are still used to satisfy the special needs of that community.

In the communication network part legacy circuit switching technology based on synchronous TDM is used to cover the traditional real-time aspects of voice communication.



## Change of Environment 2

IP Host

VCS VOIP GW LOC1

Ethernet Cards

LAN

IP Subnet 1

IP Router

CE-A1

IP WAN
Packet Switching
Connectionless
Best-Effort Service

CE-A2

VCS VOIP GW LOC 2

IP Subnet 2

CE-A3

IP Subnet 3

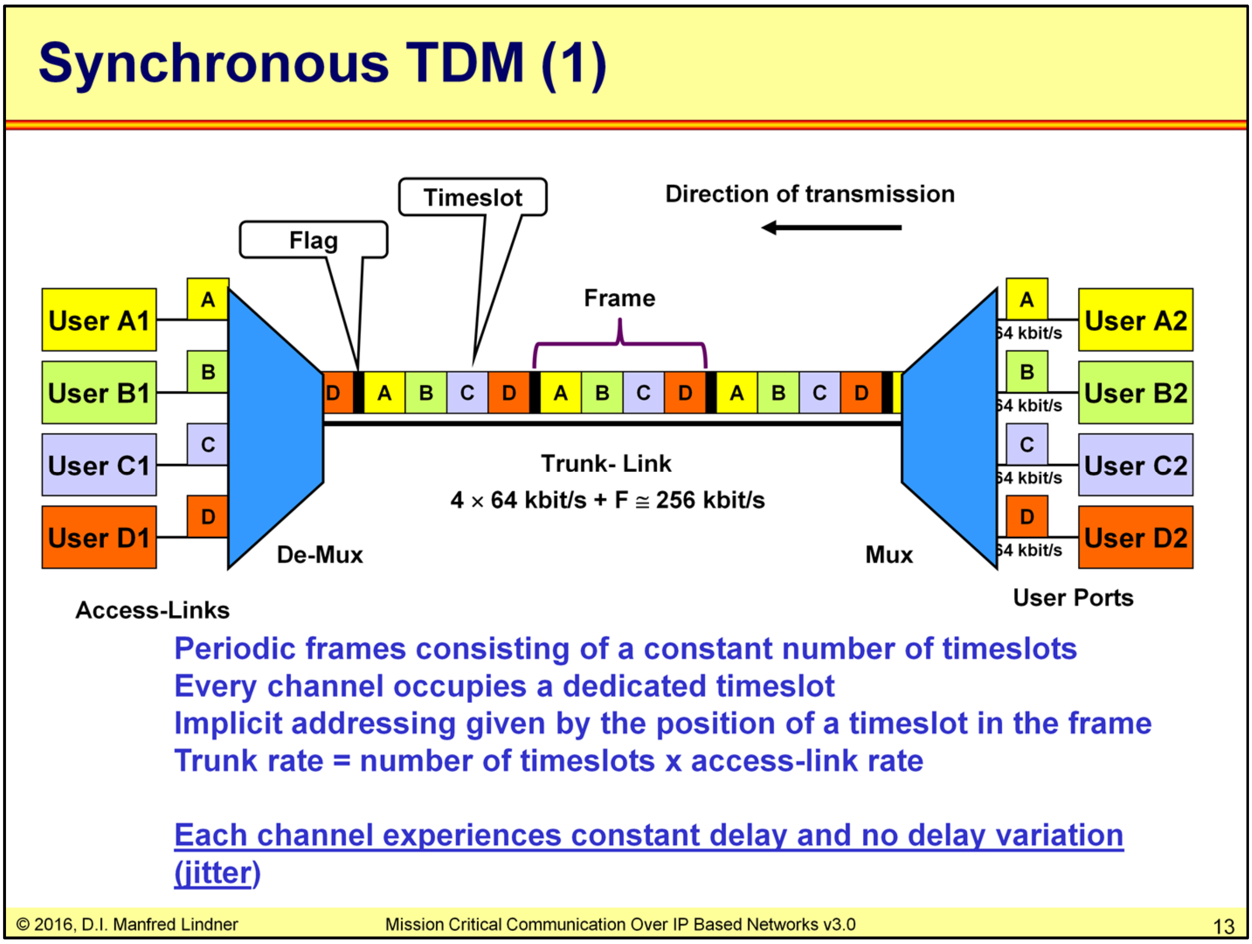
VCS VOIP GW LOC 3

World of
asynchronous TDM

VOIP … VOice over IP
GW … Gateway
LAN … Local Area Network
WAN … Wide Area Network

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 12

Nowadays a change of environment to IP networks has taken place which is a fundamental change for VCS systems. IP network uses packet-switching based on asynchronous TDM which – as best-effort system - was not built to cover real-time aspects of voice communication under all circumstances. The next pages show the basic features of synchronous and asynchronous TDM.



## Synchronous TDM (1)

Periodic frames consisting of a constant number of timeslots
Every channel occupies a dedicated timeslot
Implicit addressing given by the position of a timeslot in the frame
Trunk rate = number of timeslots x access-link rate

Each channel experiences constant delay and no delay variation (jitter)

© 2016, D.I. Manfred Lindner  Mission Critical Communication Over IP Based Networks v3.0  13

Time division multiplexer allocates each input channel a period of time (timeslot) and controls bandwidth of the trunk line among input channels. Individual time slots are assembled into frames to form a single high-speed digital data stream. The available transmission capacity of the trunk is time shared between various channels. At the destination a demultiplexer reconstructs individual channel data streams.

Synchronous TDM periodically generates a frame consisting of a constant number of timeslots each timeslot of constant length. A starting delimiter (Flag) is used for frame synchronization, which is needed to differentiate one frame from the next frame. Because of the “Flag” the individual timeslots can be identified by position within a frame (timeslot 1, timeslot 2, .... and so on). In our example we have four timeslots 1 – 4. Every input channel is assigned a dedicated timeslot e.g. data of port P1 will be carried in timeslot 1 for the A1 to A2 communication, data of port P2 will be carried in timeslot 2 for the B1 to B2 communication and son on. In our example we use “Byte-interleaving” that means a single timeslot carries 8 bit of the corresponding channel per frame.

Synchronous TDM framing on the trunk line can be vendor dependent which was used by proprietary TDM products or can be standard based.

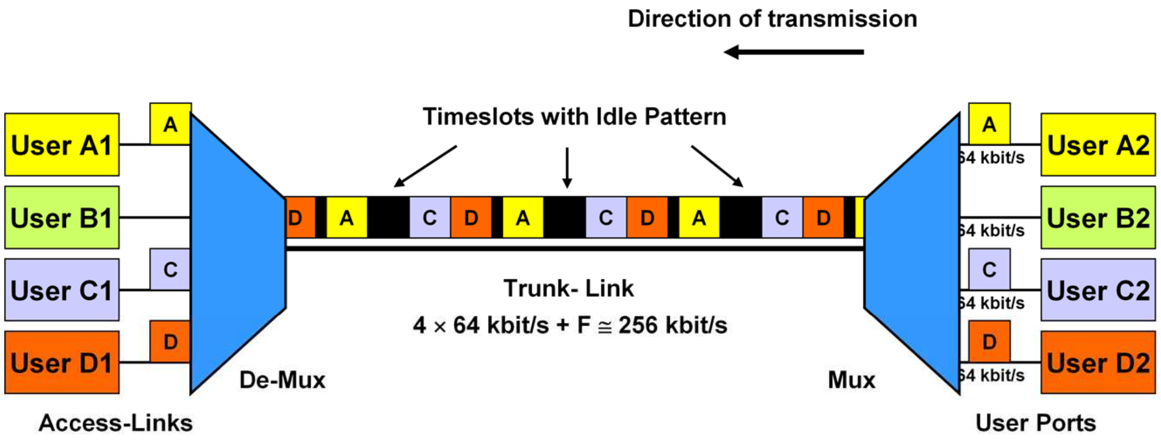
Two main architectures for standard based synchronous TDM on trunk lines for carrying PCM-coded digital telephony were established in the past:

**PDH - Plesiochronous Digital Hierarchy** developed in the 1960’s (e.g. E1 (2Mbit/s), E3 (34Mbit/s), E4, T1 (1,544Mbit/s), T3 .... )

**SDH - Synchronous Digital Hierarchy** developed in the 1980’s (e.g. STM-1 (155Mbit/s), STM-4 (622Mbit/s), STM-16 .... )

The bandwidth needed on a deterministic TDM trunk is always determined by the sum of all communication channels on the trunk plus some administrative overhead, because of the fixed correlation between communication channel and timeslot. In our example we find four communication channels with a capacity of 64Kbits/s each, so the transport capacity of the trunk needs to be 256 Kbits/s.

# Synchronous TDM (2)



Timeslot can be used for any kind of communication
-> protocol transparency
But empty timeslots are not useable by other communication channels
-> waste of bandwidth during times of inactivity

Lead to development of asynchronous/statistical multiplexing

© 2016, D.I. Manfred Lindner  Mission Critical Communication Over IP Based Networks v3.0  14

Compared to pure point-to-point physical link synchronous multiplexing adds only minimal delays: Time necessary to packetize and depacketize a byte and transmission/propagation delay on trunk

The end-to-end delay for transporting a byte is constant and the time between two bytes to be transported is constant, hence optimal for isochronous transmission requirements like traditional digital voice

Any line protocol could be used between devices, method is protocol-transparent. To endsystems such a channel looks like a single physical point-to-point line.

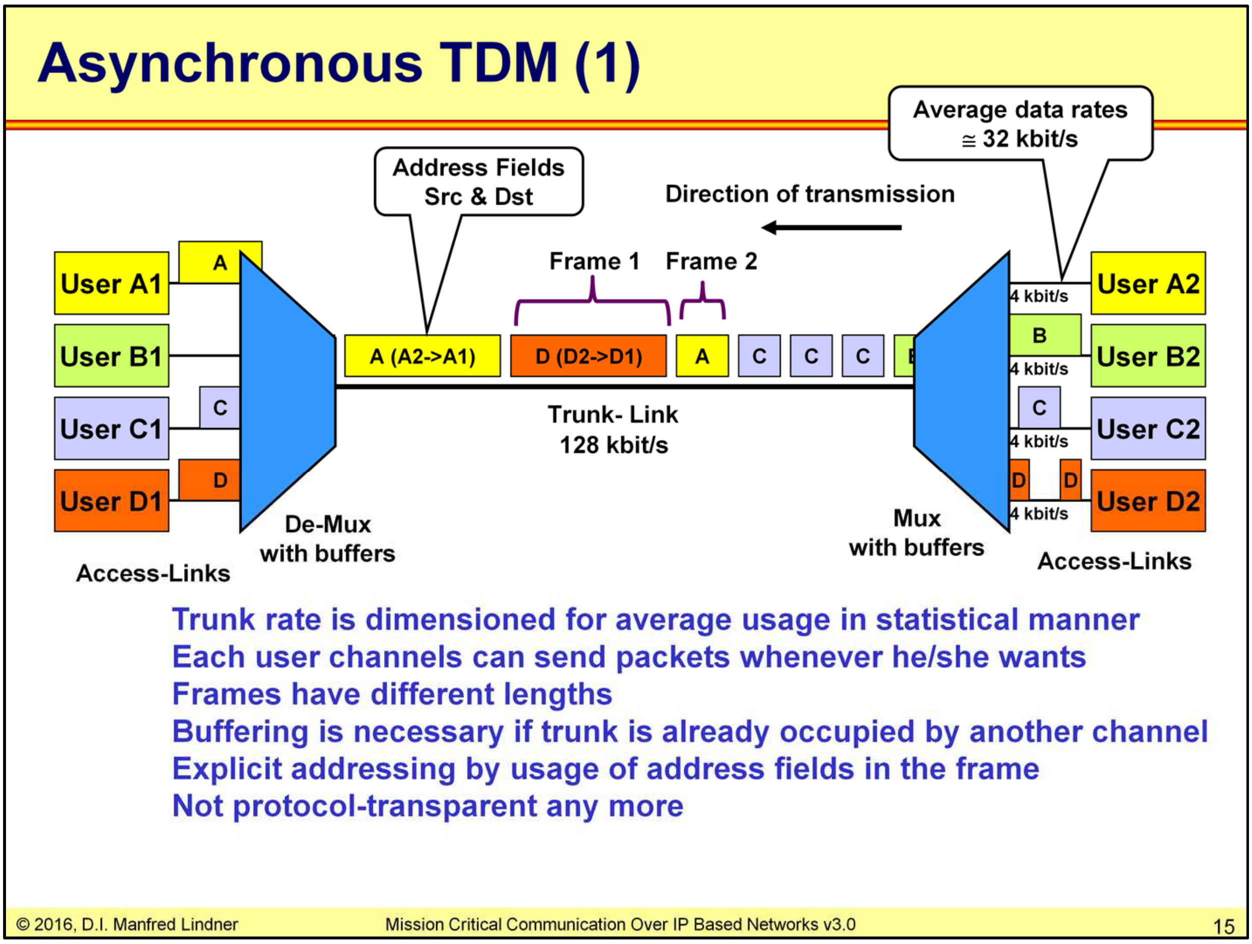
The major disadvantage of deterministic TDM systems is the fixed correlation between communication channel and time slot position. This means if one communication channel is not used it still occupies the time slot capacity by sending some kind of idle pattern. Bad trunk utilization could occur if only a few of the reserved timeslots are in use. That leads to development of asynchronous / statistical TDM.

In synchronous (deterministic) TDM systems the order of the data packets is maintained, no packet overtake or time slot position change is possible. The frames need to have always the same size because the timeslots in deterministic TDM systems have a constant length.

Address information is not required, because the destination is determined by the time slot position.

Deterministic TDM is connection-oriented because a point to point connection is typically setup by usage of SC (switched circuit) techniques like ISDN or permanently established by usage of PC (permanent circuit = leased line) techniques like PDH/SDH.

Buffers are not needed because the data stream is sent out with exactly the same speed as it is received.



# Asynchronous TDM (1)

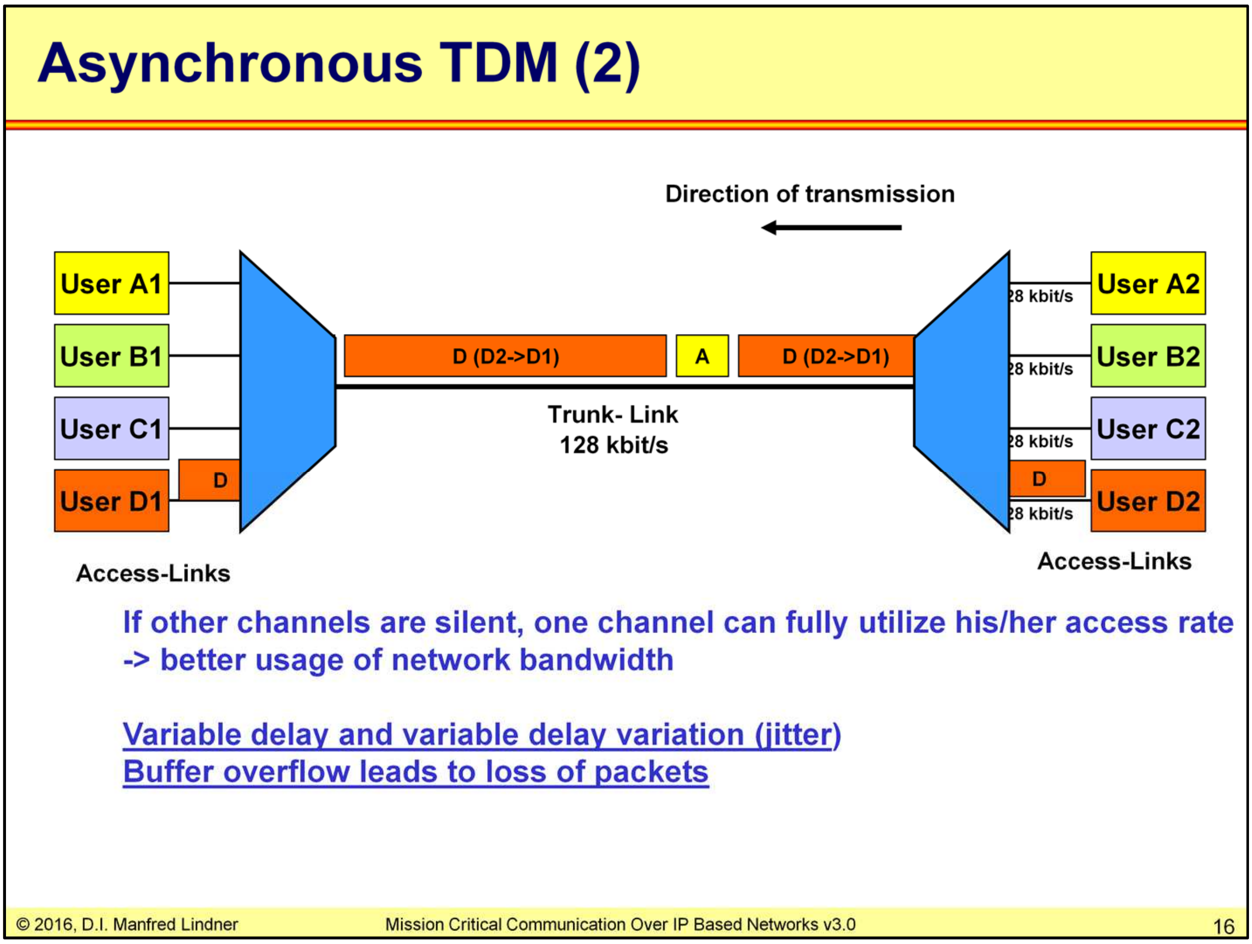
Trunk rate is dimensioned for average usage in statistical manner
Each user channels can send packets whenever he/she wants
Frames have different lengths
Buffering is necessary if trunk is already occupied by another channel
Explicit addressing by usage of address fields in the frame
Not protocol-transparent any more

Usually computer devices communicate in a statistical manner because not all devices have data to transmit at the same time. Therefore it is sufficient to calculate necessary bitrate of the multiplexer trunk line according to the average bitrates caused by device communication. The speed of the trunk could be chosen according to the average statistical transport needs of the users. Any user is allowed to send data at any time.

Now a  asynchronous multiplexer generates a transmission frame only if data bytes are present at input ports. The source of data must be explicitly identified in transmission frames so we need addressing because there is no fixed correlation between timeslot position and communication channel as it is with deterministic TDM systems.

But if devices transmit simultaneously only one channel can occupy trunk line at a given time, data of other channels must be buffered inside the multiplexer until trunk is available again (store and forward principle). Hopefully statistics is such that the trunk will not be monopolized by just a single channel. Otherwise a buffer overflow will occur in the multiplexer, leading to transmission errors seen by the individual channels.

In case of congestion buffering helps but causes additional delays compared to synchronous TDM. Delays are variable because of statistical behavior hence not optimal for synchronous transmission requirements like traditional digital voice but still sufficient for transmission requirements like bursty data transfer between computers.



## Asynchronous TDM (2)

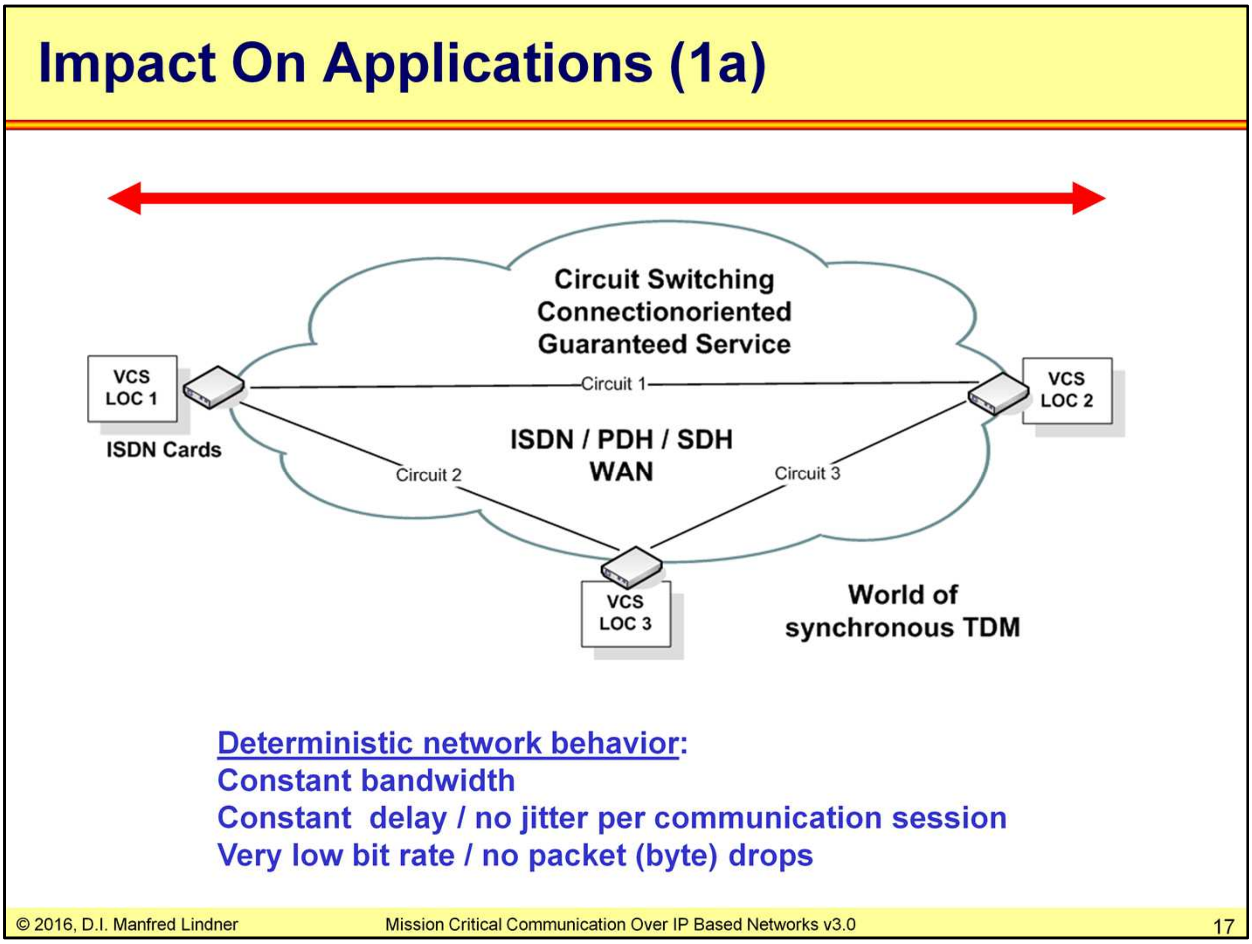
If other channels are silent, one channel can fully utilize his/her access rate -> better usage of network bandwidth

Variable delay and variable delay variation (jitter)
Buffer overflow leads to loss of packets

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 16

One of the major advantages of statistical TDM systems compared to deterministic TDM systems is the following fact: If the trunk is empty one user may use the complete transport capacity of the trunk. On the other hand it may occur that all users want to use the trunk at the same time. Because of the statistical dimensioning of the trunk capacity it may happen that more data is fed in by the users than the trunk capacity allows. For such cases buffers are needed by the statistical TDM devices to compensate the speed differences. In case of buffer overflow conditions it may even happen that data is lost.

Statistical TDM allows a good utilization of the trunk because there is no waste of bandwidth by the use of idle patterns and the capacity is determined by the average needs of the users. The frame size may vary depending on the need of the users. Buffering is required under trunk overload conditions. The delay is variable because of buffering. Statistical TDM is not protocol transparent because a separate packing as well as addresses are needed. End system and ADTM multiplexer have to speak the same protocol language and hence not anymore protocol transparent.



## Impact On Applications (1a)

Circuit Switching
Connectionoriented
Guaranteed Service

VCS LOC 1

ISDN Cards

VCS LOC 2

VCS LOC 3

Circuit 1

Circuit 2

Circuit 3

ISDN / PDH / SDH
WAN

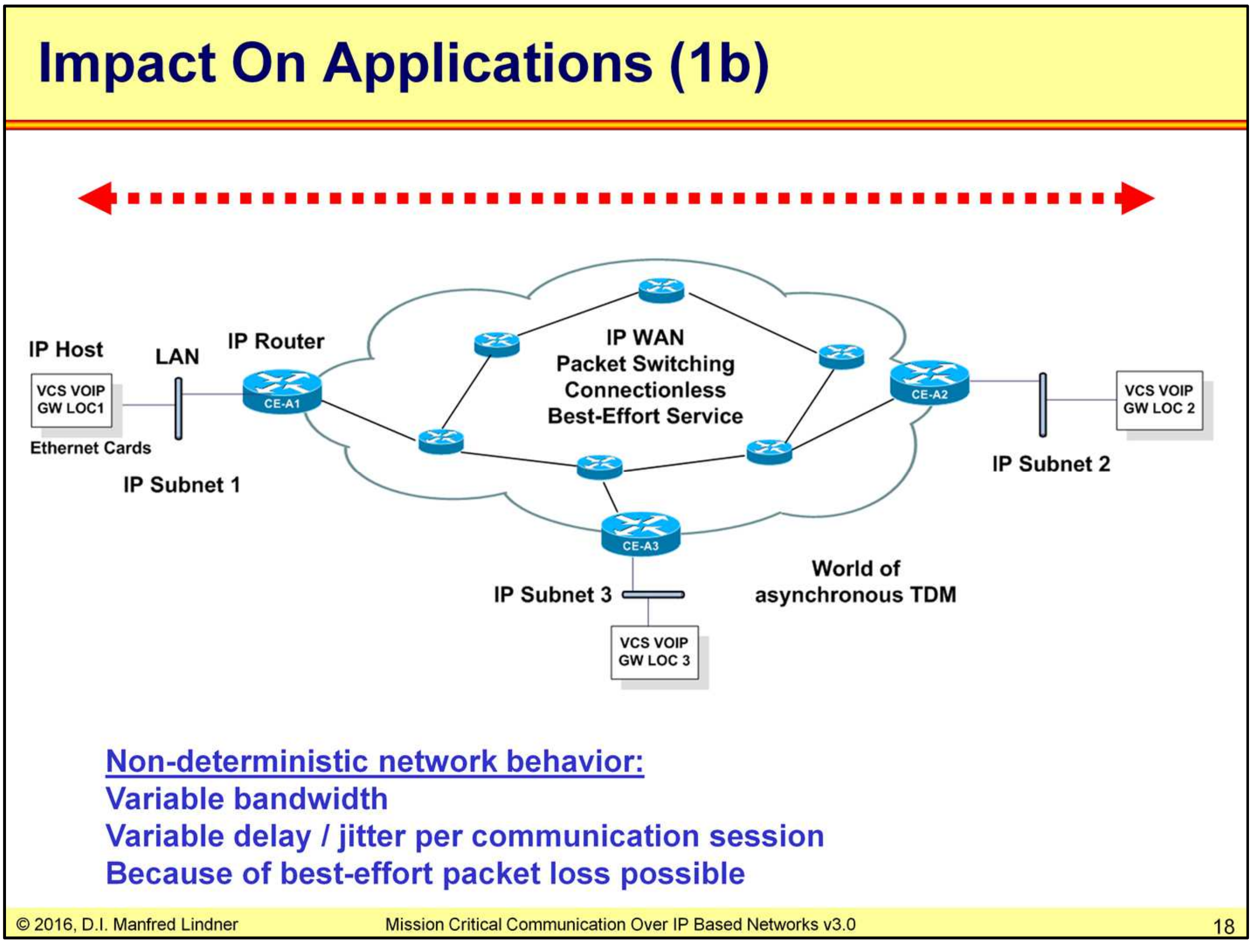
World of
synchronous TDM

**Deterministic network behavior:**
Constant bandwidth
Constant delay / no jitter per communication session
Very low bit rate / no packet (byte) drops

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 17

The change from synchronous TDM world to asynchronous TDM world has impacts on the applications implementing mission critical communication.

In the synchronous TDM world there is a deterministic behavior concerning delay, bandwidth. The red arrow in the slide should express that.

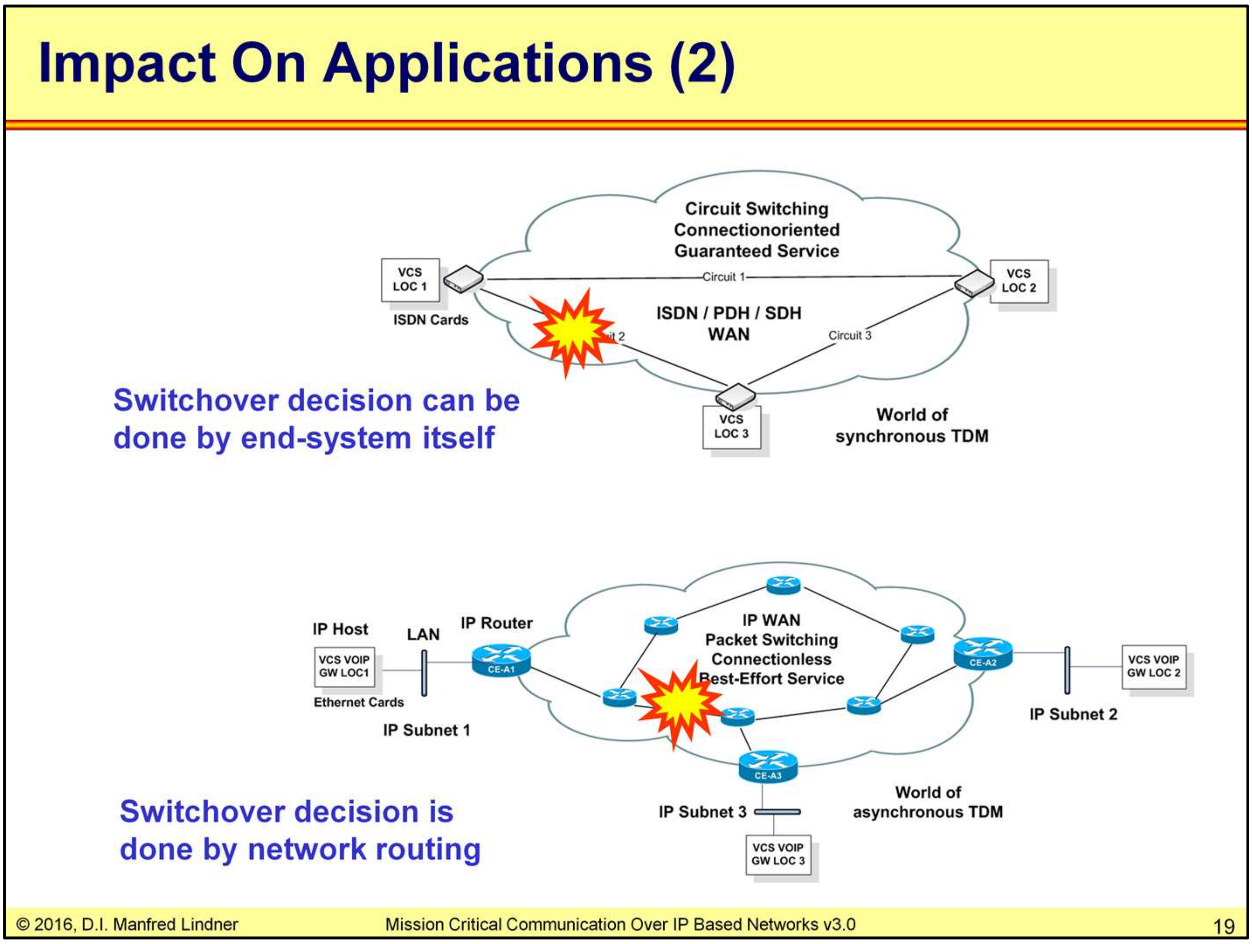


# Impact On Applications (1b)

**Non-deterministic network behavior:**
Variable bandwidth
Variable delay / jitter per communication session
Because of best-effort packet loss possible

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 18

In the asynchronous TDM world there is a statistic behavior concerning delay, bandwidth and loss. The dashed red arrow in the slide should express that.



# Impact On Applications (2)

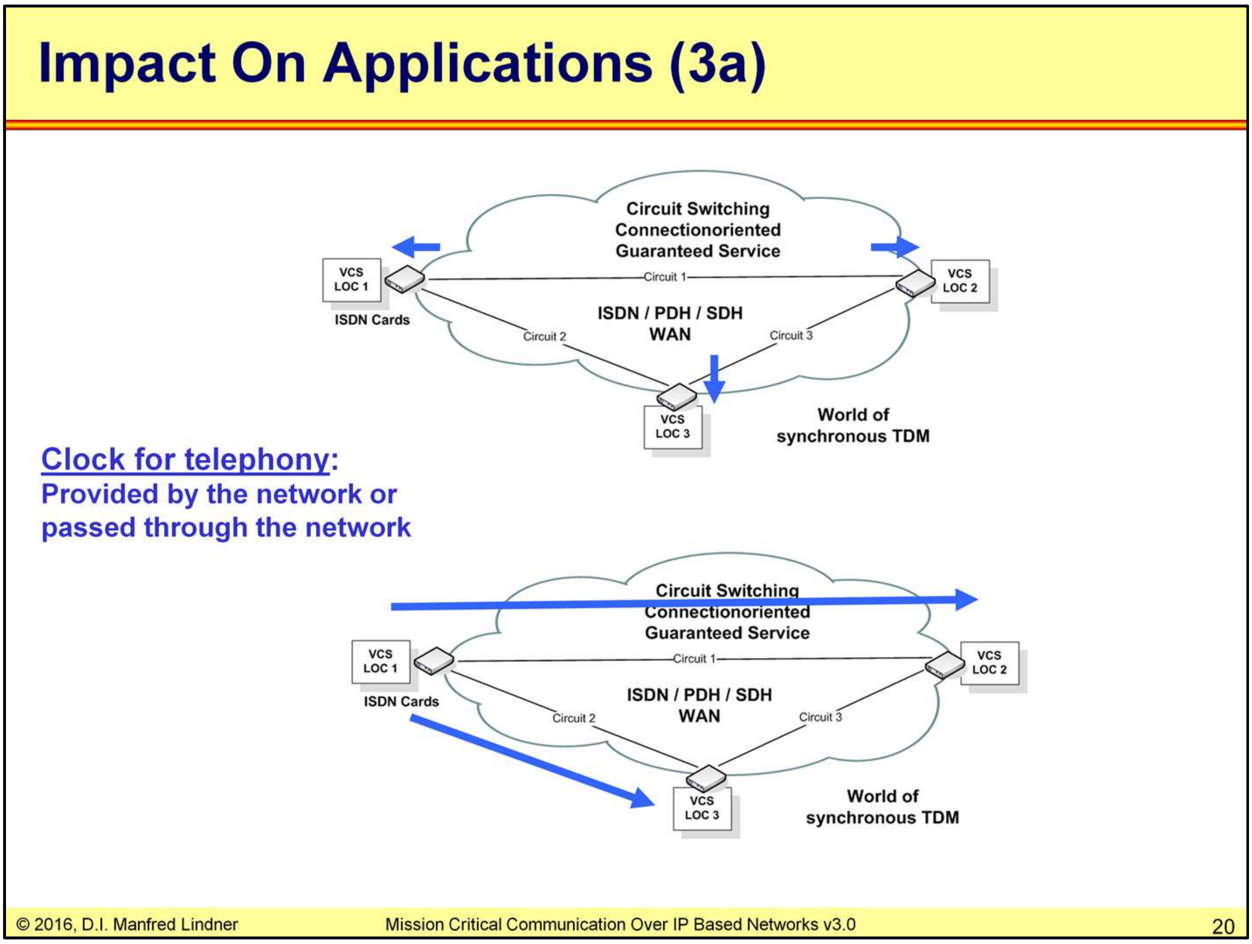
Switchover decision can be done by end-system itself

Switchover decision is done by network routing

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 19

In the synchronous TDM world - because of the inherent connection-oriented behavior - the application knows about the status of a circuit either by presence of physics or by signaling information from the network.

In the asynchronous TDM world the application can not concluded the status of a communication session from the presence of local physics. IP network with the connectionless behavior has no signaling information to the end-system about the status of the network or network parts.

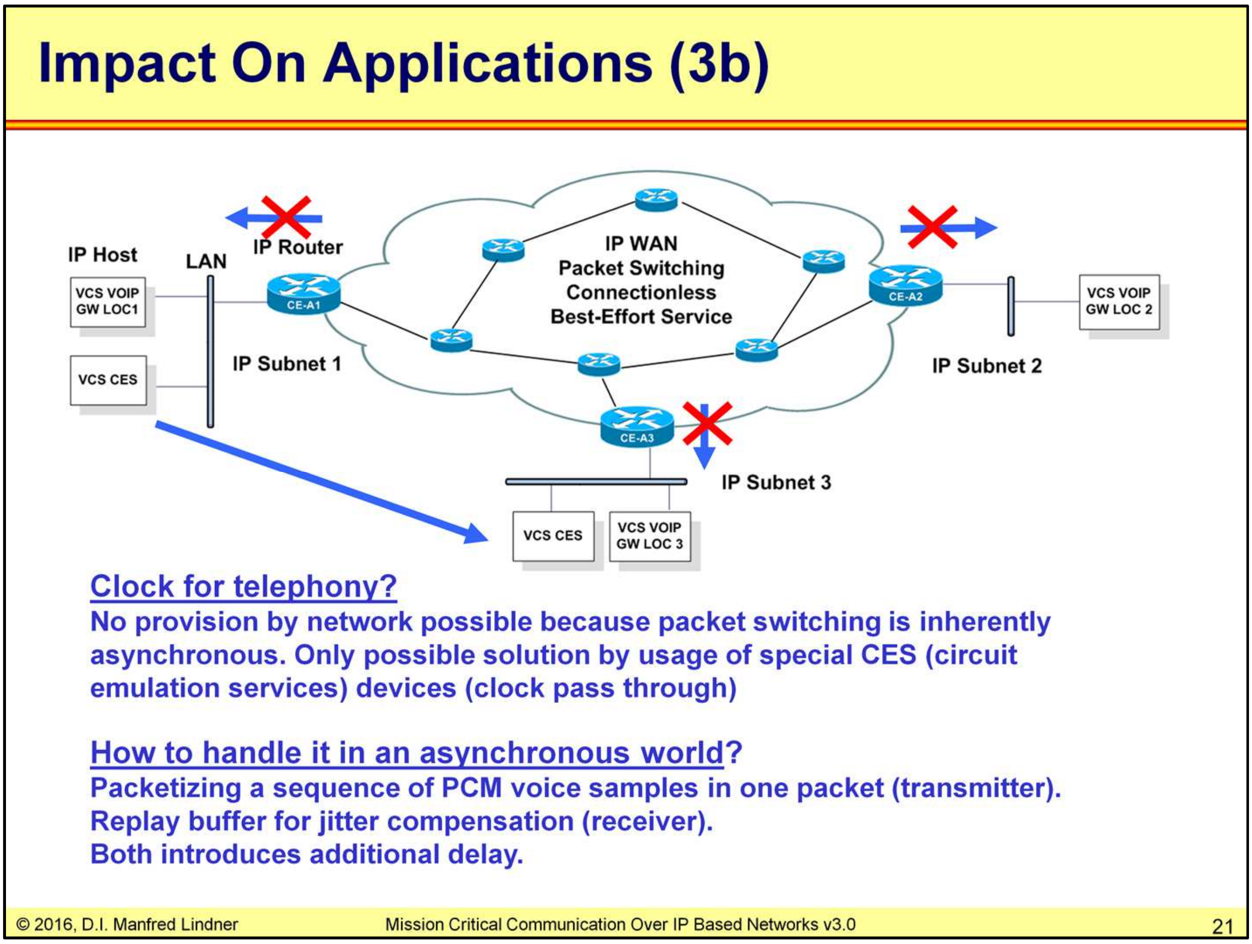


# Impact On Applications (3a)

Clock for telephony:
Provided by the network or
passed through the network

Mission Critical Communication Over IP Based Networks v3.0 20

In the synchronous TDM world clock can be given by the network or even transported over the network.



## Impact On Applications (3b)

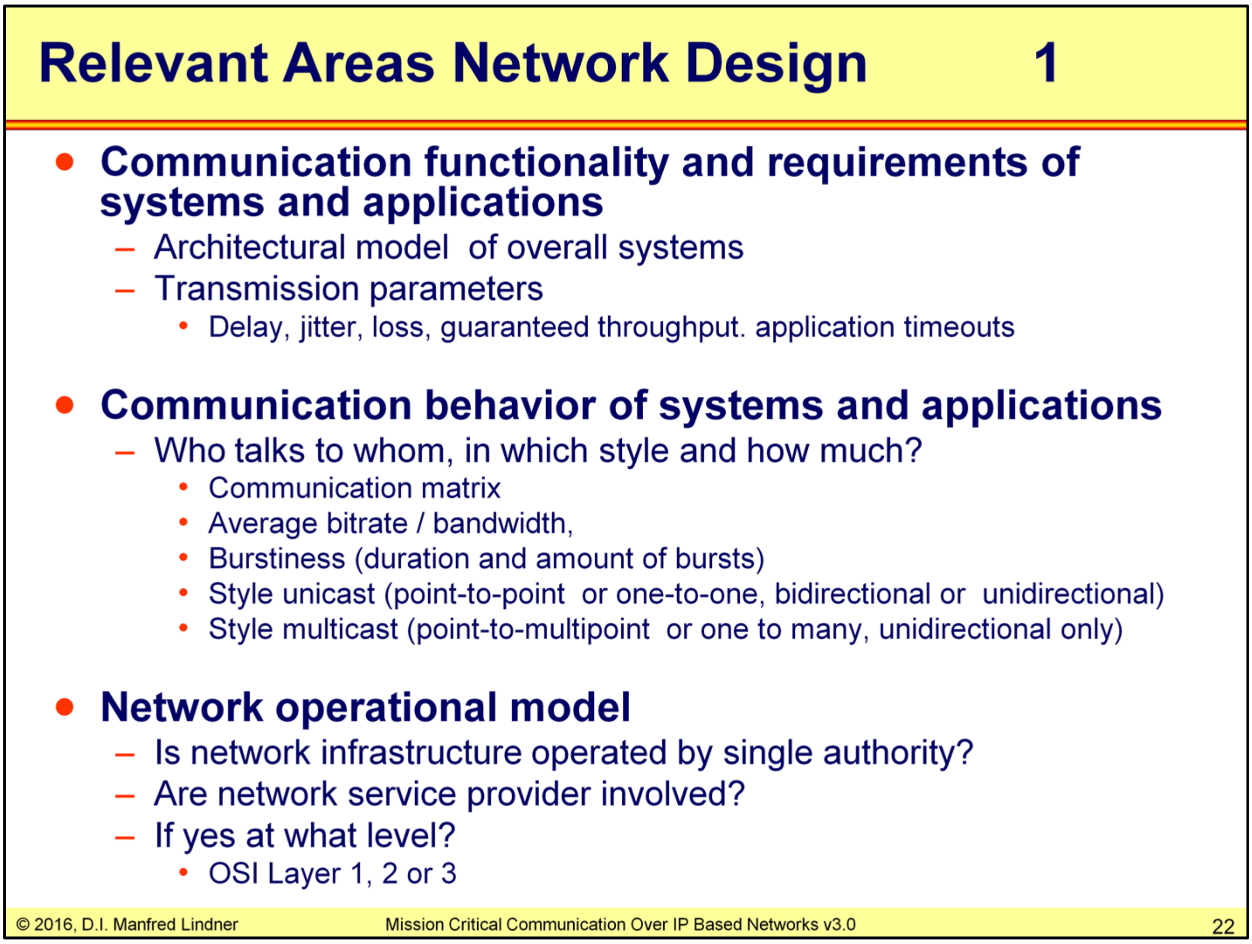
**Clock for telephony?**

No provision by network possible because packet switching is inherently asynchronous. Only possible solution by usage of special CES (circuit emulation services) devices (clock pass through)

**How to handle it in an asynchronous world?**

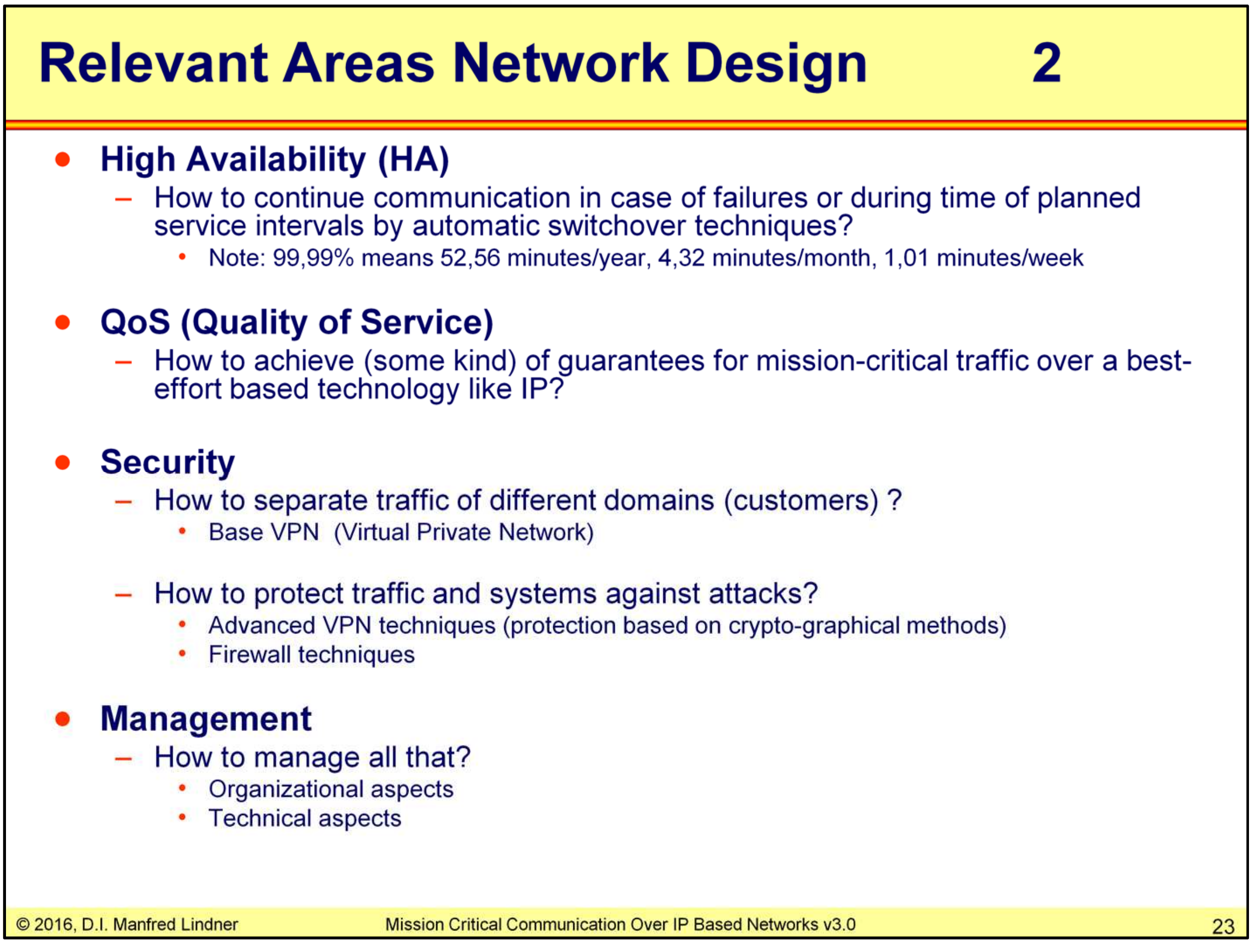
Packetizing a sequence of PCM voice samples in one packet (transmitter).
Replay buffer for jitter compensation (receiver).
Both introduces additional delay.

In the asynchronous TDM world clock there is no clock information available from the network.



## Relevant Areas Network Design 1

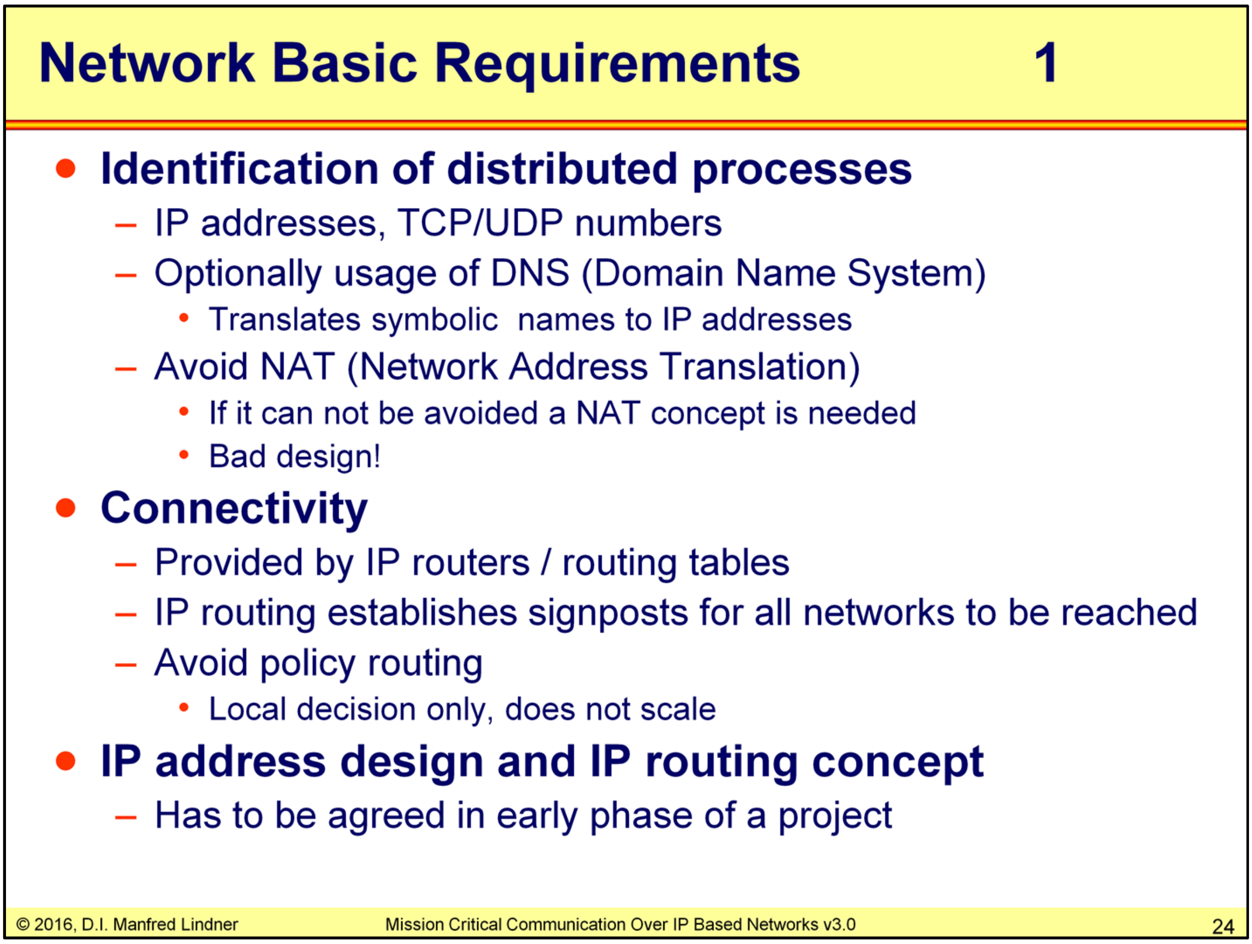
- **Communication functionality and requirements of systems and applications**
  - Architectural model of overall systems
  - Transmission parameters
    - Delay, jitter, loss, guaranteed throughput. application timeouts
- **Communication behavior of systems and applications**
  - Who talks to whom, in which style and how much?
    - Communication matrix
    - Average bitrate / bandwidth,
    - Burstiness (duration and amount of bursts)
    - Style unicast (point-to-point or one-to-one, bidirectional or unidirectional)
    - Style multicast (point-to-multipoint or one to many, unidirectional only)
- **Network operational model**
  - Is network infrastructure operated by single authority?
  - Are network service provider involved?
  - If yes at what level?
    - OSI Layer 1, 2 or 3



## Relevant Areas Network Design 2

- **High Availability (HA)**
  - How to continue communication in case of failures or during time of planned service intervals by automatic switchover techniques?
    - Note: 99,99% means 52,56 minutes/year, 4,32 minutes/month, 1,01 minutes/week
- **QoS (Quality of Service)**
  - How to achieve (some kind) of guarantees for mission-critical traffic over a best-effort based technology like IP?
- **Security**
  - How to separate traffic of different domains (customers) ?
    - Base VPN (Virtual Private Network)
  - How to protect traffic and systems against attacks?
    - Advanced VPN techniques (protection based on crypto-graphical methods)
    - Firewall techniques
- **Management**
  - How to manage all that?
    - Organizational aspects
    - Technical aspects

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 23



## Network Basic Requirements 1

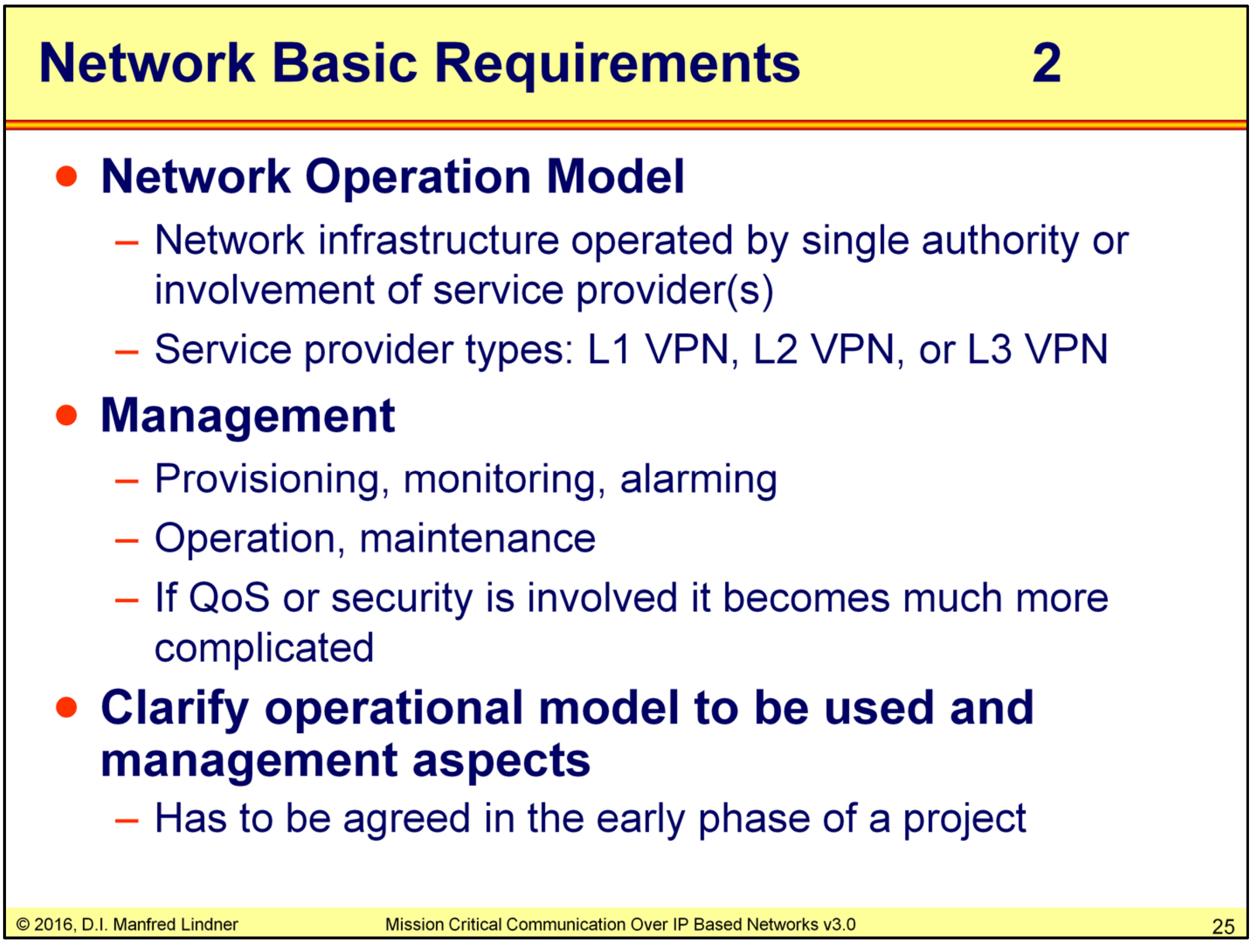
- Identification of distributed processes
  - IP addresses, TCP/UDP numbers
  - Optionally usage of DNS (Domain Name System)
    - Translates symbolic names to IP addresses
  - Avoid NAT (Network Address Translation)
    - If it can not be avoided a NAT concept is needed
    - Bad design!
- Connectivity
  - Provided by IP routers / routing tables
  - IP routing establishes signposts for all networks to be reached
  - Avoid policy routing
    - Local decision only, does not scale
- IP address design and IP routing concept
  - Has to be agreed in early phase of a project

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 24

For every IP network there are some basic requirements to be fulfilled in order to enable communication among systems.

First of all systems have to be identified in order to enable IP based communication. The identification is a combination of an IP address or symbolic DNS hostname and UDP|TCP port number. Additionally the IP address represents the location of the system in the network topology in a structured way (Net-ID, Host-ID). If systems are addressed by different authorities or administrative domains coordination between these authorities / domains is necessary because IP address have to be unique within an IP domain. If coordination is not done at the very first beginning the usual overcome is usage of NAT (Network Address Translation). A good address design will avoid NAT because a lot of troubles may arise by using NAT, which often are not foreseeable if NAT is introduced during the project. Typical NAT is implemented on firewalls and IP routers. Stateful NAT has quite some severe implications on redundancy which might be necessary because of request for high availability. One example for such implications is a tight and very close coupling of the components of a firewall cluster which forbids placing the components in different IP subnets (L2 only), hence physically distributing them becomes a problem

Net-IDs (in the following named IP routes) of involved systems have to be known by the IP routing system - especially in the routing table of an IP router - in order to forward IP datagrams to the destination network. The normal behavior of an IP router is “destination based routing” hence source address of an IP datagram is not involved in the forwarding decision. Involvement of IP source address in the forwarding decision is possible (IP policy routing) but leads to a local decision of such a policy-enabled IP router only. Therefore what will be happen by routers downstream cannot be enforced by a router applying IP policy routing (“hop-by-hop routing principle” of IP). IP policy routing maybe used in special situation to overcome a problem but should not be a rule of design because any change of network topology needs careful reconfiguration which may not scale in most topologies. Identification and connectivity both need IP address planning (IP address concept) and routing protocol coordination (IP routing concept).



## Network Basic Requirements 2

- **Network Operation Model**
  - Network infrastructure operated by single authority or involvement of service provider(s)
  - Service provider types: L1 VPN, L2 VPN, or L3 VPN
- **Management**
  - Provisioning, monitoring, alarming
  - Operation, maintenance
  - If QoS or security is involved it becomes much more complicated
- **Clarify operational model to be used and management aspects**
  - Has to be agreed in the early phase of a project

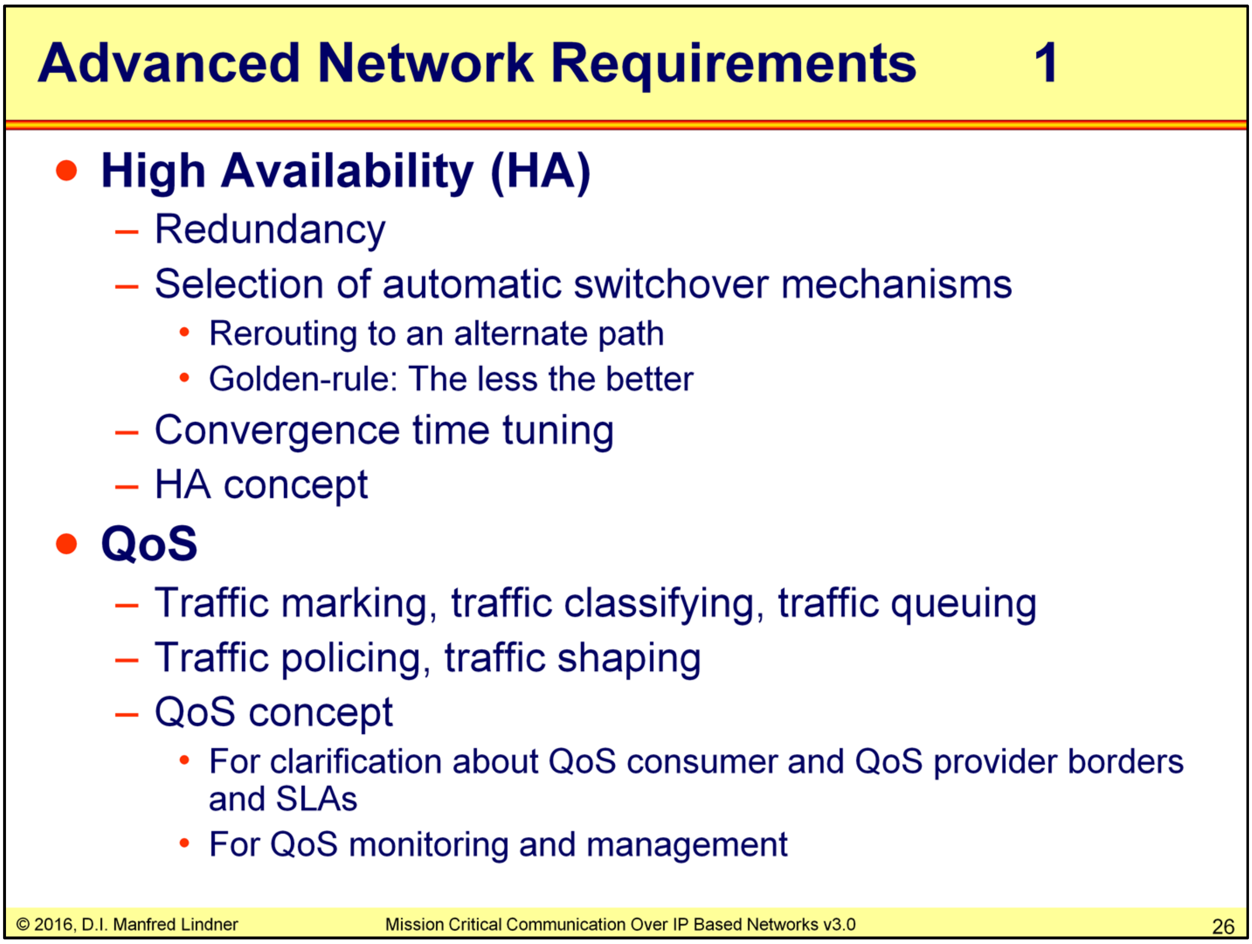
© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 25

Any IP network can be built based on different operational models. The complete network infrastructure might be owned and operated by one unique authority (customer) or parts of the network infrastructure might be operated by another authority (e.g. out-sourced to service provider).

In case of usage of a service provider the network layers covered by this service provider may vary. Usually the service provided is a kind of VPN (Virtual Private Network) which allows the provider to share a common infrastructure among several customers guaranteeing at least separation of the customers among each other. Optionally dedicated performance to the customers may be guaranteed by usage of either any kind of circuit-technology (PDH, SDH) or of IP QoS mechanism.

Service providers could be of type L1-VPN provider, L2-VPN provider or L3-VPN provider.

Network management includes activities, methods, procedures and tools that pertain to operation, administration, maintenance and provisioning of IP networked systems. Topics covered are FCAPS (Fault, Configuration, Accounting, Performance and Security). Depth of each individual FCAPS topic depends on the actual business requirements. It can start from the easy going into complexity. It is mentioned here in the basic requirements because often ignored or overseen judged as being from less importance. In reality it turns out to have a huge impact on the overall performance of your solution involving not only technology but also human operators.



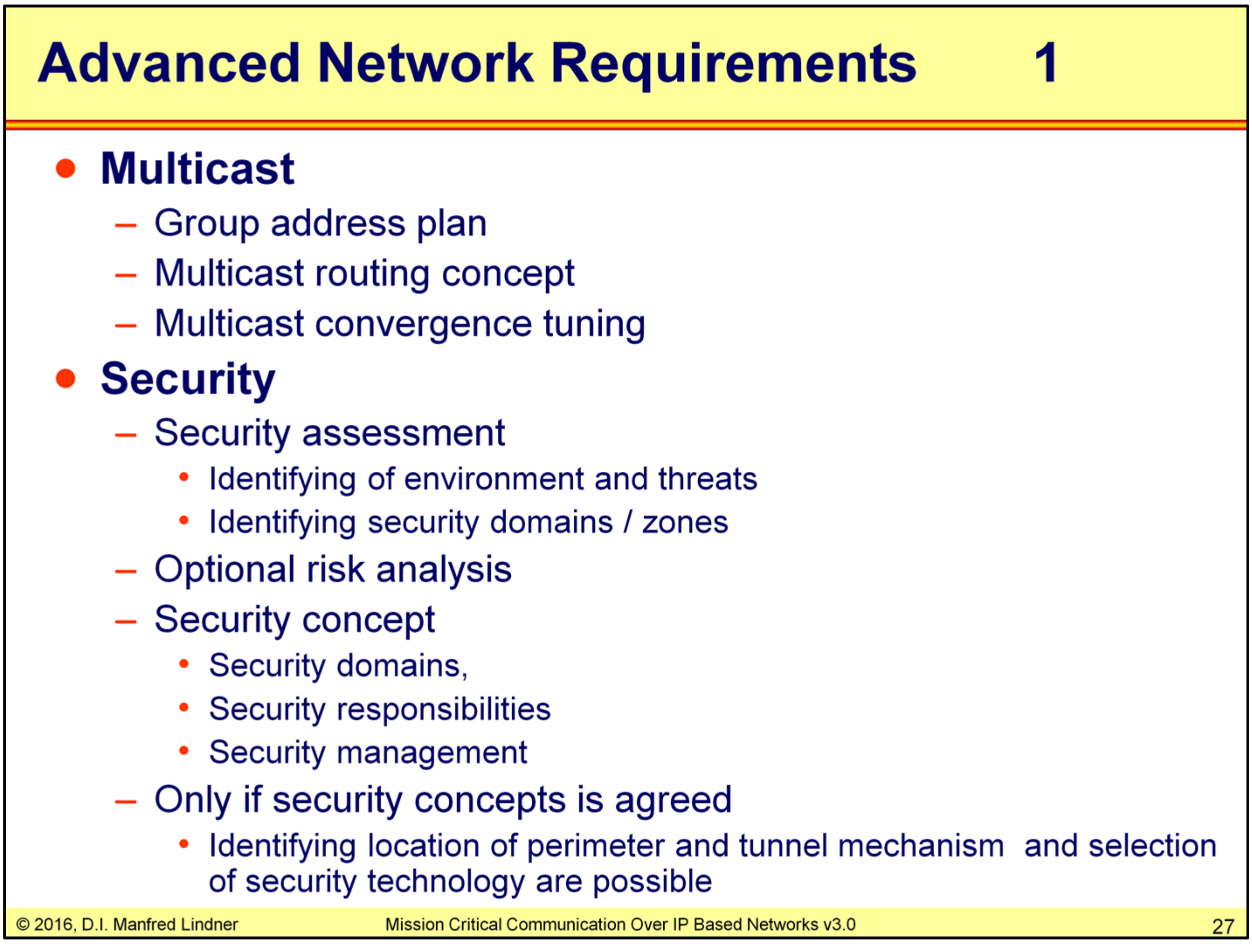
## Advanced Network Requirements 1

- **High Availability (HA)**
  - Redundancy
  - Selection of automatic switchover mechanisms
    - Rerouting to an alternate path
    - Golden-rule: The less the better
  - Convergence time tuning
  - HA concept
- **QoS**
  - Traffic marking, traffic classifying, traffic queuing
  - Traffic policing, traffic shaping
  - QoS concept
    - For clarification about QoS consumer and QoS provider borders and SLAs
    - For QoS monitoring and management

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 26

In case of mission critical communication advanced requirements have to be fulfilled by an IP network. These are high availability (HA), quality of service (QoS) , multicasting and security aspects.

HA and QoS issues will be discussed in the following sections of this lecture in sufficient detail.

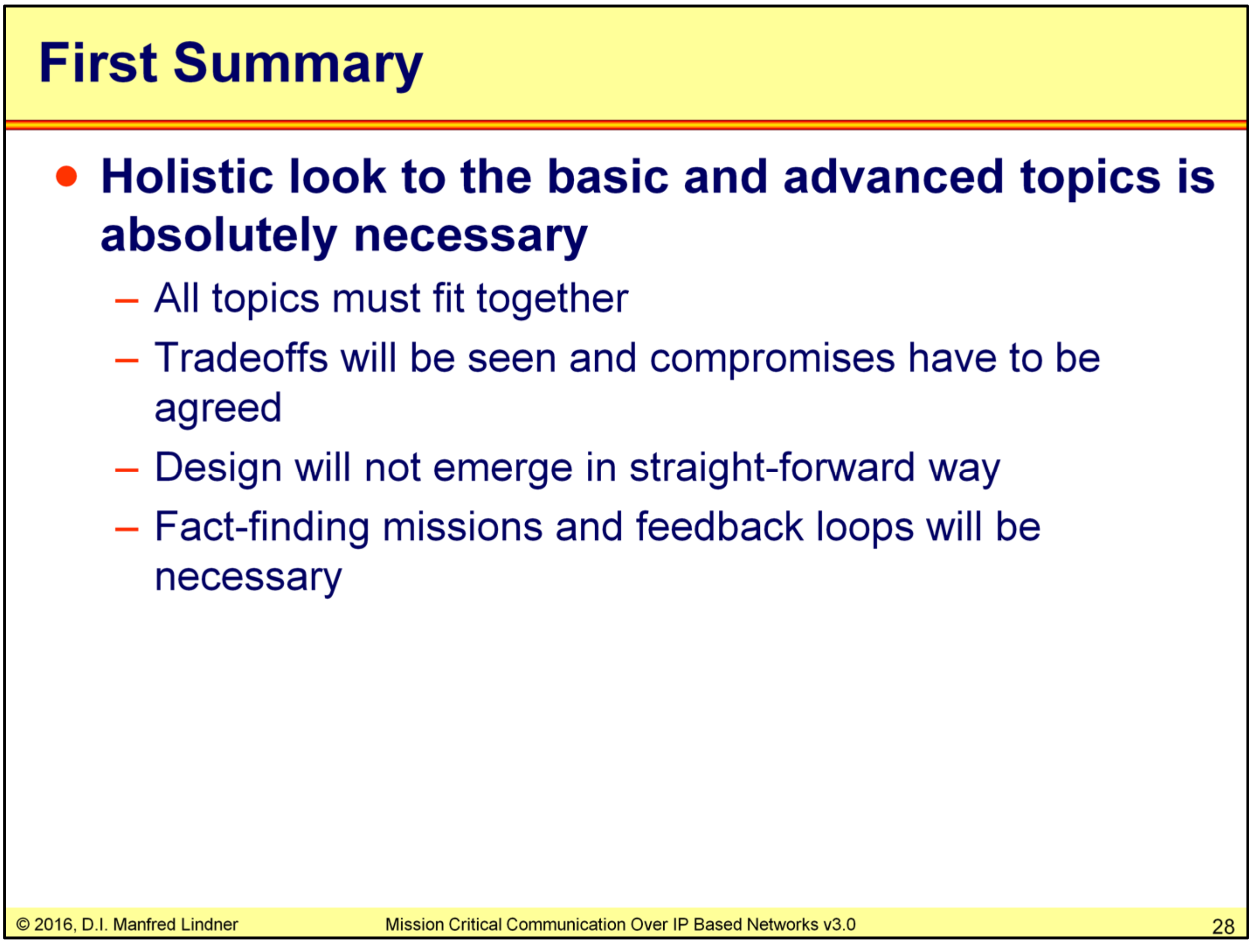


## Advanced Network Requirements 1

- **Multicast**
  - Group address plan
  - Multicast routing concept
  - Multicast convergence tuning
- **Security**
  - Security assessment
    - Identifying of environment and threats
    - Identifying security domains / zones
  - Optional risk analysis
  - Security concept
    - Security domains,
    - Security responsibilities
    - Security management
  - Only if security concepts is agreed
    - Identifying location of perimeter and tunnel mechanism and selection of security technology are possible

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 27

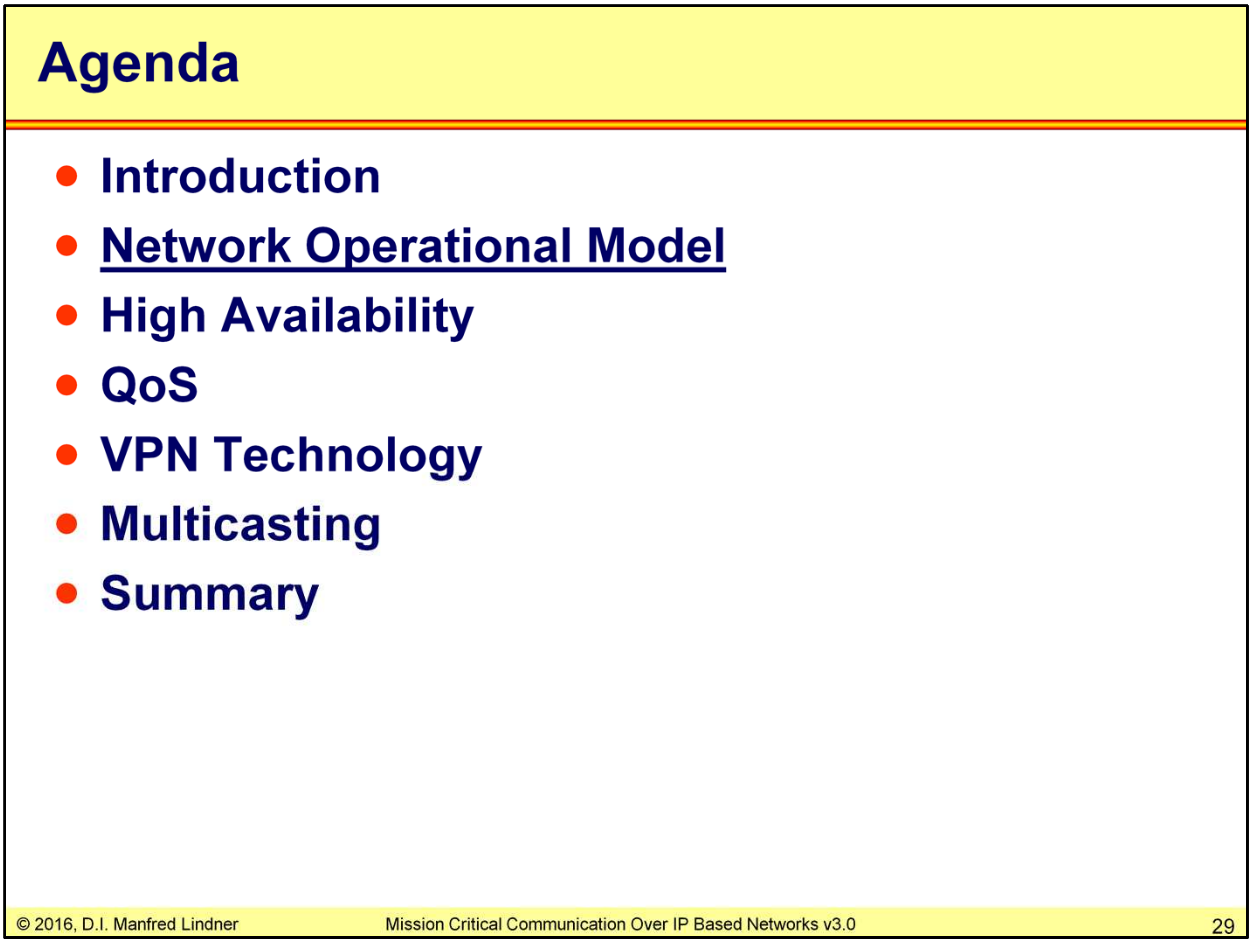
Multicast and security issues will be discussed in the following sections of this lecture in sufficient detail.



## First Summary

- **Holistic look to the basic and advanced topics is absolutely necessary**
  - All topics must fit together
  - Tradeoffs will be seen and compromises have to be agreed
  - Design will not emerge in straight-forward way
  - Fact-finding missions and feedback loops will be necessary

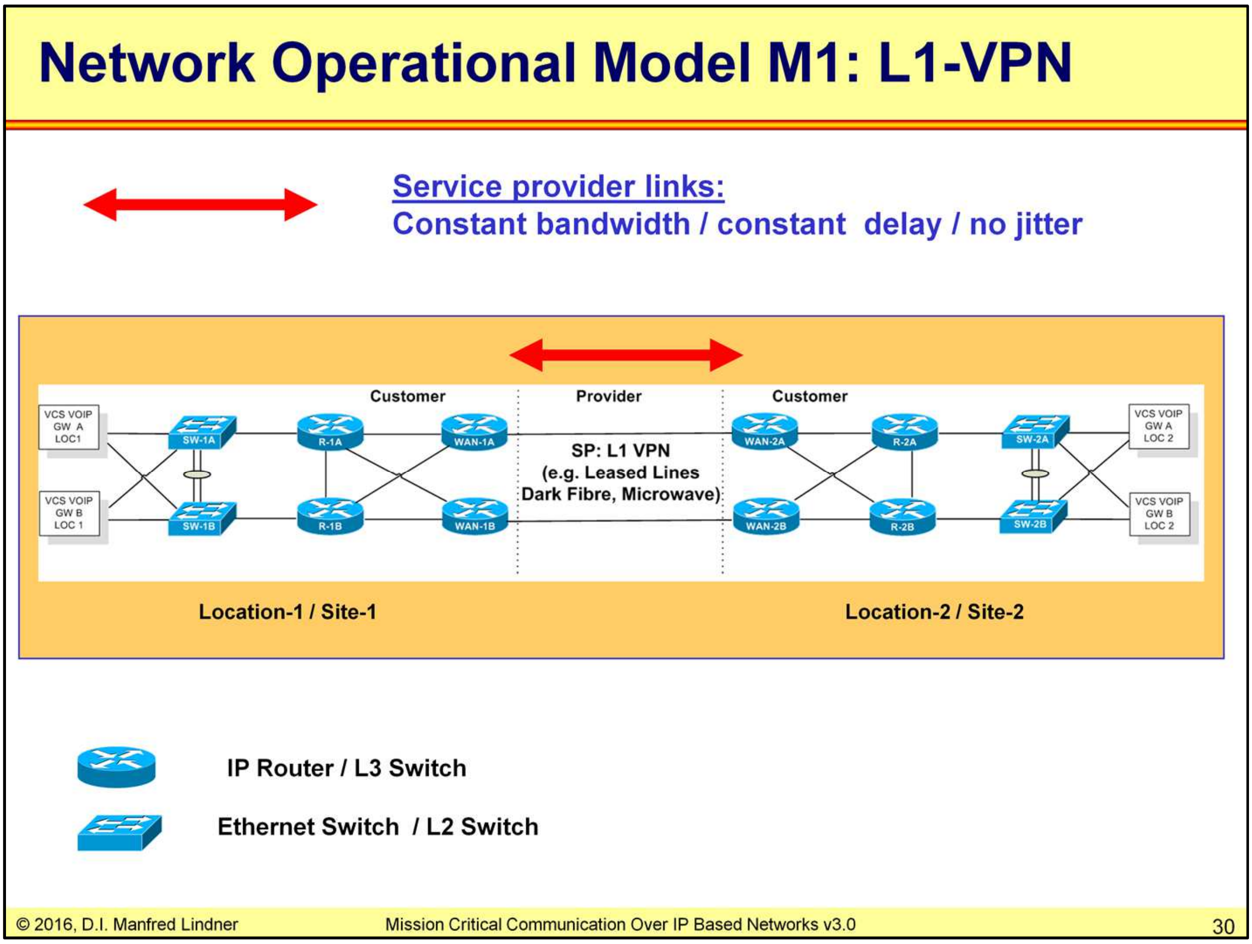
© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 28



# Agenda

- Introduction
- Network Operational Model
- High Availability
- QoS
- VPN Technology
- Multicasting
- Summary

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 29



## Network Operational Model M1: L1-VPN

Service provider links:
Constant bandwidth / constant delay / no jitter

IP Router / L3 Switch

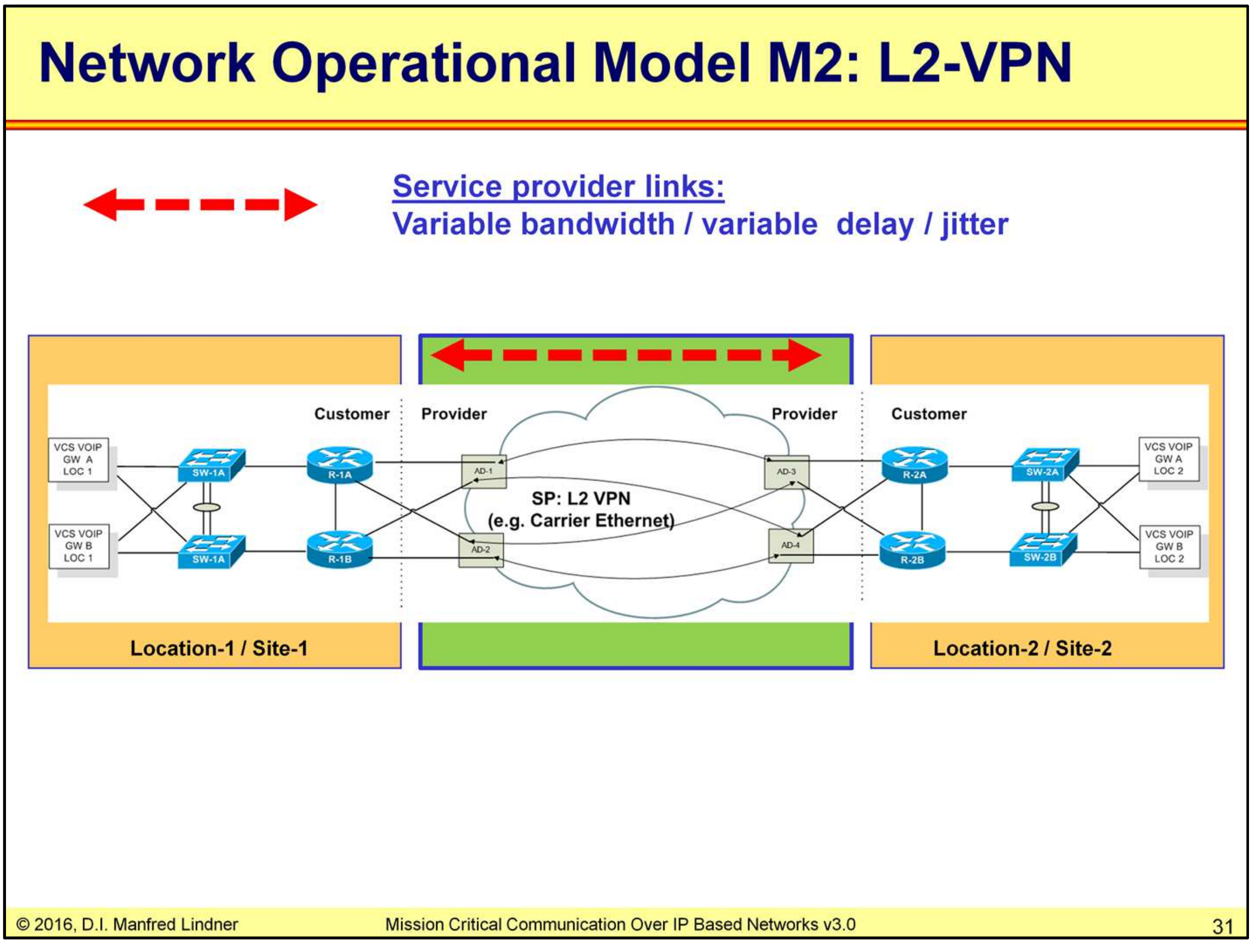
Ethernet Switch / L2 Switch

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 30

Model M1 (L1-VPN) shows that all IP routers are under control of the customer. L1-VPN SP provides link with dedicated bandwidth and constant delay, which is not shared by other customers. Customer has full control regarding IP addressing, IP routing and QoS implementation.

Examples for L1-VPN are leased line (based on circuit switching PDH/SDH infrastructure), dark fiber or microwave radio. In case of microwave the bandwidth may vary based on environment influences such as weather but the network operator can calculate at least with a minimum (fallback) bandwidth available all the time as long the microwave modems are in operation and synchronized.

Even if it looks like that the end-systems of a location need six elements (e.g. Ethernet switches SW-1A, SW-1B and IP routers R-1A, R-1B, WAN-1A, WAN-1B) in order to connect to the IP WAN, depending on the actual situation concerning the number of end-systems and the kind of provider equipment the elements may be reduced to such two elements incorporating all the necessary functionality (Ethernet switch, IP router and WAN access) in just one single component. The reason for the six (logical) elements will be explained in chapter about HA (High Availability).



## Network Operational Model M2: L2-VPN

Service provider links:
Variable bandwidth / variable  delay / jitter

Location-1 / Site-1

Location-2 / Site-2

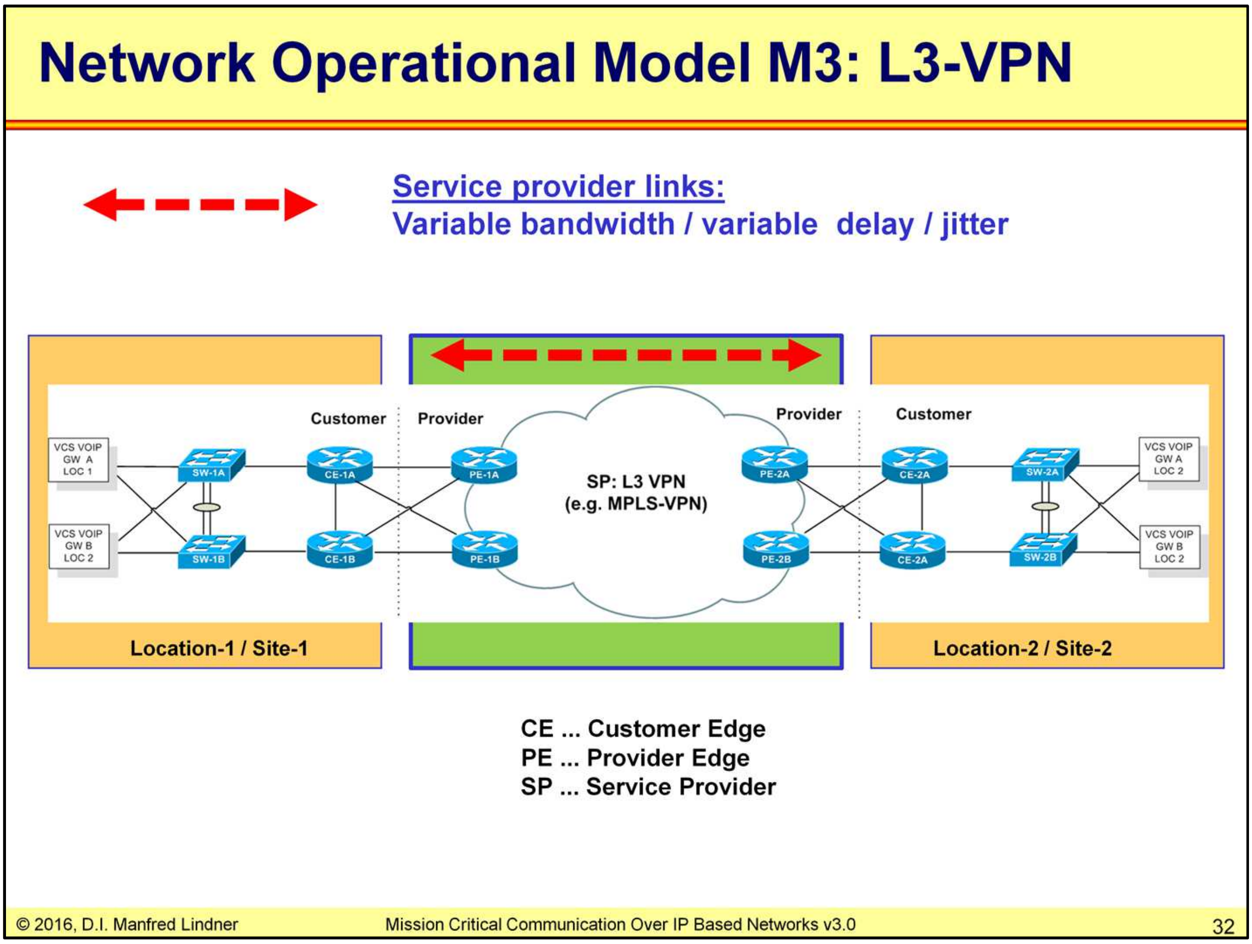
© 2016, D.I. Manfred Lindner  Mission Critical Communication Over IP Based Networks v3.0  31

Model M2 (L2-VPN) shows that all IP routers are under control of the customer but the connectivity between customer routers is based on just L2 link technology. In the past examples for L2-VPN were legacy X.25 virtual circuits  or FR (Frame Relay) virtual circuits both based on a connection-oriented packet switching infrastructure.

Nowadays VLANs (= multiplexing of LANs across a shared L2 Ethernet switching infrastructure), Carrier-Ethernet or VPLS (Virtual Private LAN Service) are typical examples. In such a case the customer sees just (virtual) Ethernet links. Nevertheless these Ethernet links are provided by the SP on top of another - for the customer not visible - connectionless packet switching technology. For example both Carrier-Ethernet and VPLS are based on pseudo-wires implemented on top of IP-MPLS packet switching operated by the L2-VPN provider.

A Carrier-Ethernet or VPLS service will support different customers based on the common infrastructure of the service provider. Therefore even with dedicated physical bandwidth for every Ethernet link a variable delay will be experienced on such a virtual Ethernet link end-to-end depending on the traffic statistics of all the customers sharing the sane SP. Customer has full control regarding IP addressing, IP routing but QoS is still inherited from the characteristics/statistics of the service provider.

Other examples for such L2-VPNs are microwave/SAT modems or SDH-AD (Add/Drop) multiplexers offering their service on a corresponding Ethernet link to the customer routers. Usually there is speed mismatch between the Ethernet link and the microwave-link or SDH circuit (e.g. 100 Mbps Ethernet should use a 34 Mbps circuit). In such a case the provider equipment acts as (transparent aka hidden to the customer) Ethernet switch performing packet switching between the different physical technologies. In case of packet bursts it can happen that buffering is not possible anymore and some packets get lost. So it looks like a Ethernet link, but some weird things can get on in such a constellation.



## Network Operational Model M3: L3-VPN

Service provider links:
Variable bandwidth / variable delay / jitter

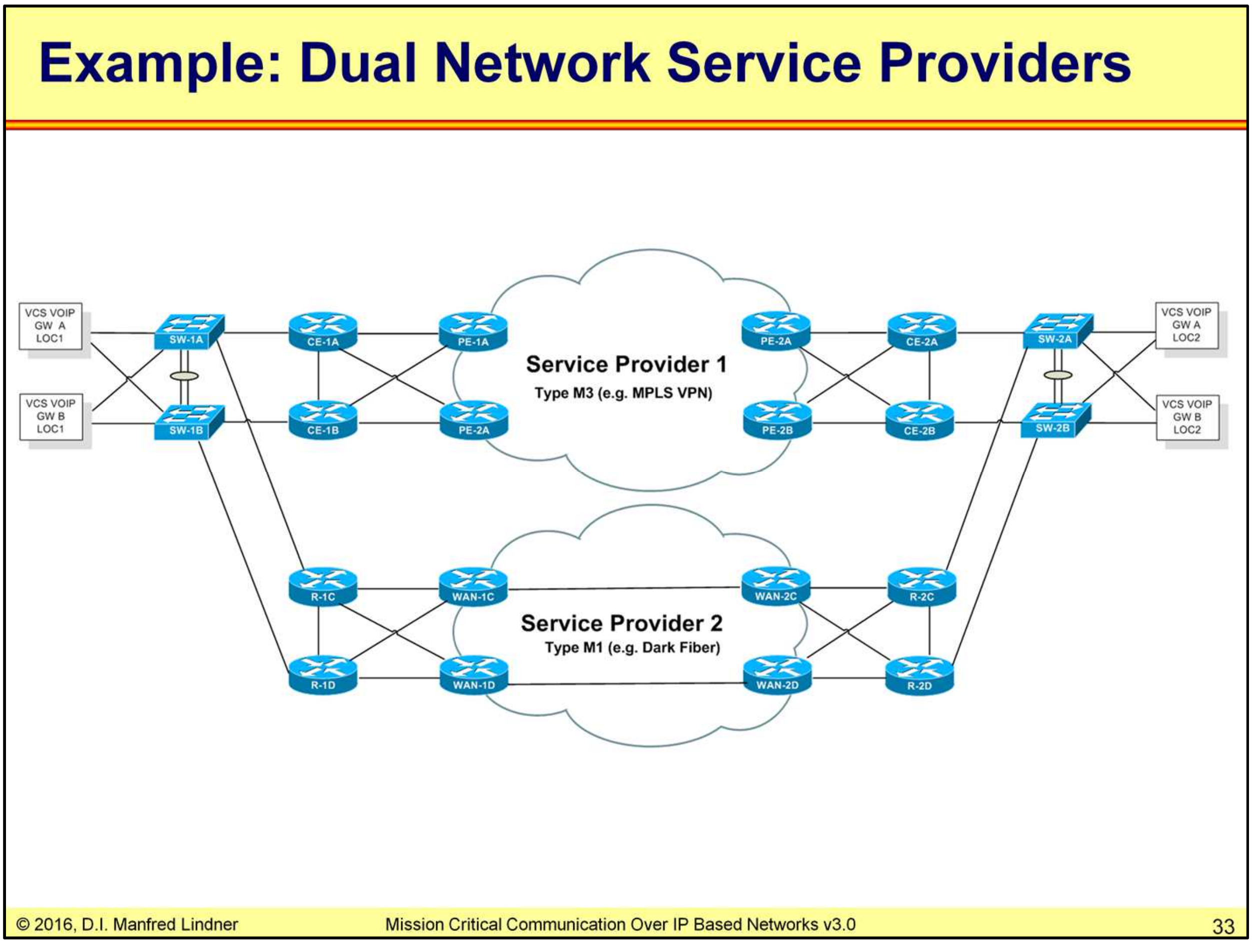
CE ... Customer Edge
PE ... Provider Edge
SP ... Service Provider

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 32

Model M3 L3-VPN shows that only CE (Customer Edge) IP routers are under control of the customer. The PE (Provider Edge) IP router and the WAN links are under the control of the service provider. L3 VPN SP provides separated IP connectivity for each customer but all customers shares the same infrastructure. That results in variable bandwidth and variable delay depending on the traffic statistics of all the customers sharing the same SP. Customer and provider only control IP addressing, IP routing and QoS in their own domain. Interaction, coordination and supervision are necessary at the border in between. IP routing convergence and QoS characteristics are inherited from the service provider.

Examples for L3-VPNs are MPLS-VPNs (=multiplexing of IP nets across a shared L3 IP/MPLS infrastructure) or IPsec Site-Site-VPNs (=tunneling of over a given network infrastructure with inherent crypto-graphical support for encryption and integrity). L3-VPNs are also called overlay VPNs to express that routing within a service provider network is separated from the routing of the customer.

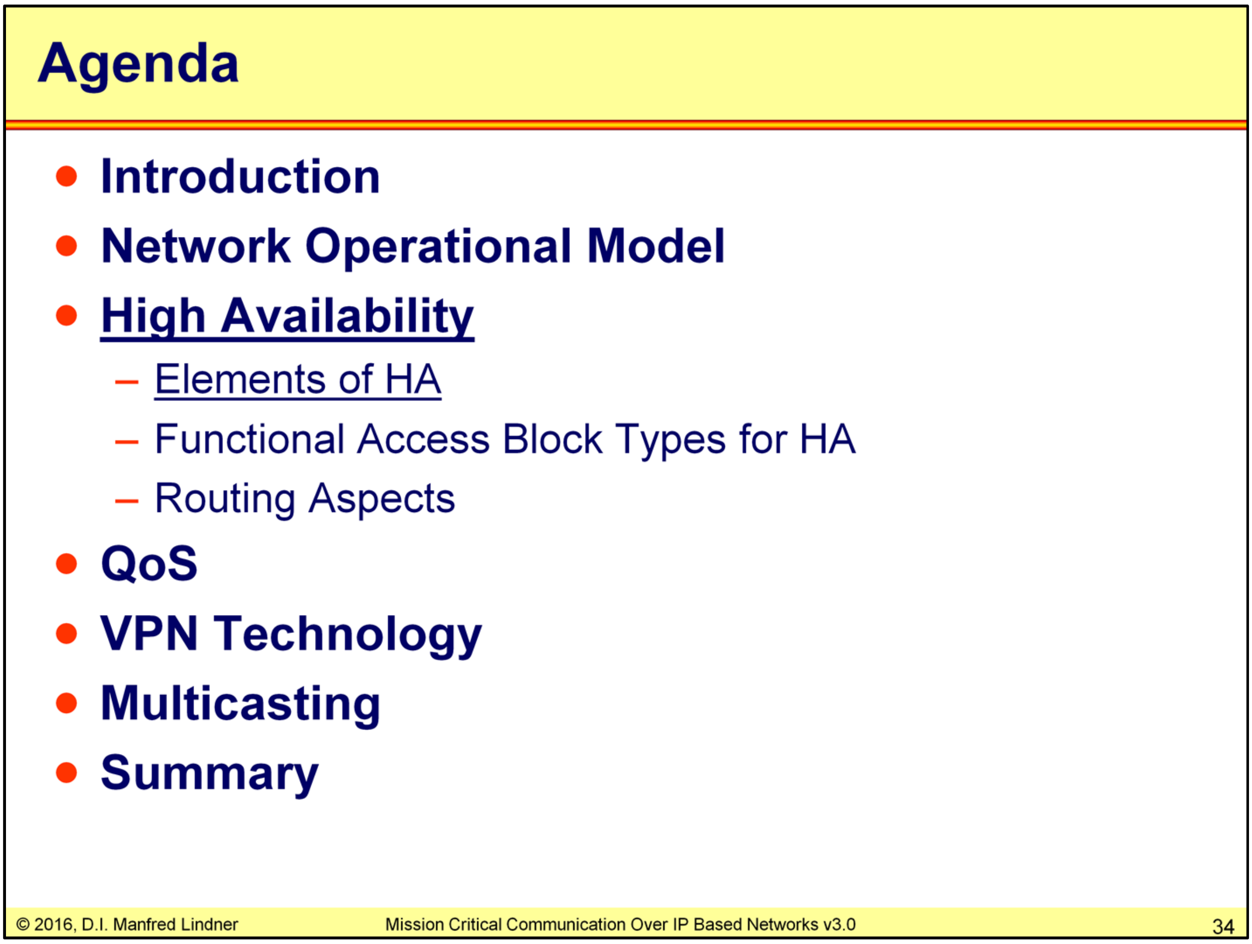
]



# Example: Dual Network Service Providers

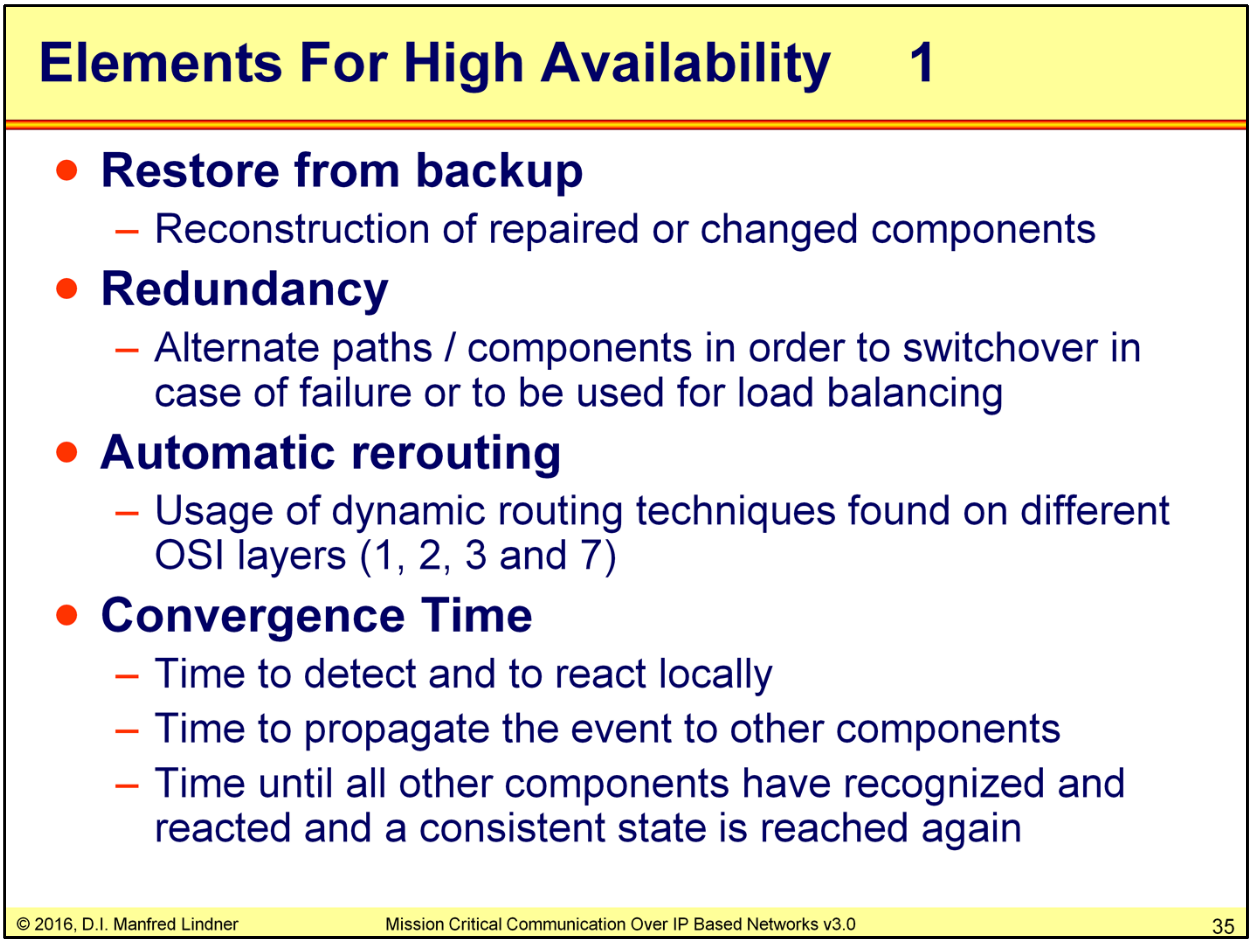
© 2016, D.I. Manfred Lindner    Mission Critical Communication Over IP Based Networks v3.0    33

A single service provider can give you a  certain kind of availability (amount of “Nines”). For achieving an overall higher availability sending duplicated data on parallel paths is an option. Selection of “better” stream can be done by the receiving application.



## Agenda

- Introduction
- Network Operational Model
- High Availability
  - Elements of HA
  - Functional Access Block Types for HA
  - Routing Aspects
- QoS
- VPN Technology
- Multicasting
- Summary



## Elements For High Availability 1

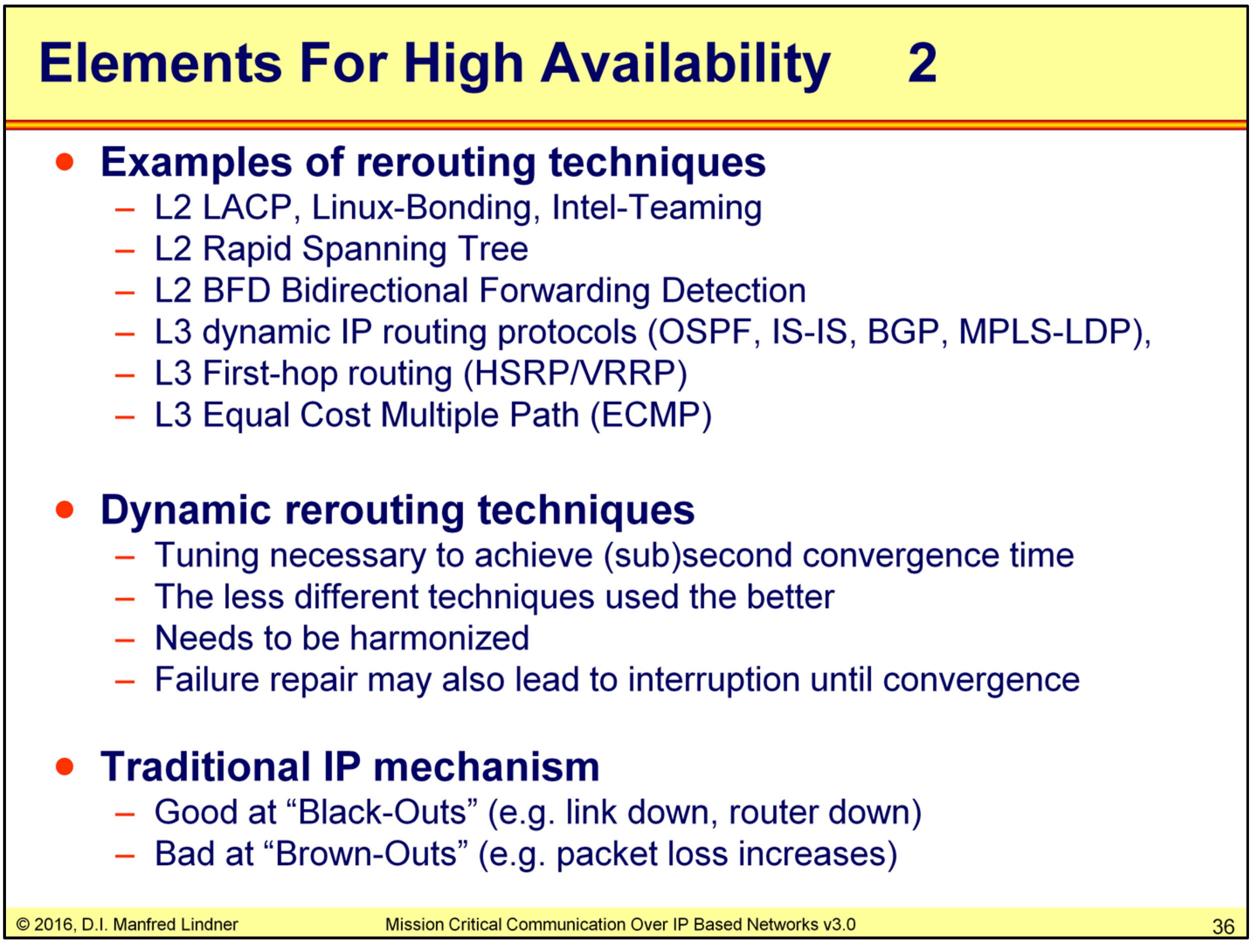
- **Restore from backup**
  - Reconstruction of repaired or changed components
- **Redundancy**
  - Alternate paths / components in order to switchover in case of failure or to be used for load balancing
- **Automatic rerouting**
  - Usage of dynamic routing techniques found on different OSI layers (1, 2, 3 and 7)
- **Convergence Time**
  - Time to detect and to react locally
  - Time to propagate the event to other components
  - Time until all other components have recognized and reacted and a consistent state is reached again

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 35

HA mechanisms ensure communication in case of failures or during time of planed service intervals by usage of automatic switchover mechanism in case of single point of failures. Typical elements are redundancy (alternate paths, alternate components), automatic rerouting and restore from backup.

Dynamic IP routing protocols like will consistently create routing table entries by exchanging information about IP routes among routers. These dynamic routing protocols can detect and overcome a network failure in a certain time - convergence time. Network failures detectable are local link down (fastest possible convergence) or neighbor down (timeout based convergence). Dynamic IP routing protocols will not include traffic statistics like usage of link bandwidth, delay of buffering or bit-errors on links to initiate a routing convergence.

Convergence time involves the following elements which sum up: First time to detect a failure and to react locally, second time to propagate the failure to other components and third time until all other components have recognized the failure and react on it to achieve a consistent state again. During convergence time inconsistent states are possible. Such inconsistency should not lead to an overall system crash or deadlock hence complexity shall be reduced to a minimum as far as possible. Unfortunately routing convergence (selection and switchover to an alternate path) can and will happen on different levels.



## Elements For High Availability 2

- **Examples of rerouting techniques**
  - L2 LACP, Linux-Bonding, Intel-Teaming
  - L2 Rapid Spanning Tree
  - L2 BFD Bidirectional Forwarding Detection
  - L3 dynamic IP routing protocols (OSPF, IS-IS, BGP, MPLS-LDP),
  - L3 First-hop routing (HSRP/VRRP)
  - L3 Equal Cost Multiple Path (ECMP)
- **Dynamic rerouting techniques**
  - Tuning necessary to achieve (sub)second convergence time
  - The less different techniques used the better
  - Needs to be harmonized
  - Failure repair may also lead to interruption until convergence
- **Traditional IP mechanism**
  - Good at “Black-Outs” (e.g. link down, router down)
  - Bad at “Brown-Outs” (e.g. packet loss increases)

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 36

For example some standardized mechanism are:

On Layer 1 (e.g. SDH): Link Fast Reroute

On Layer 2 (e.g. Ethernet): BFD (Bidirectional Forwarding Detection), LACP (Link Aggregation Control Protocol), Teaming, Bonding

On Layer 2 (e.g. Ethernet Switching): STP (Spanning Tree Protocol), Rapid-Spanning Tree

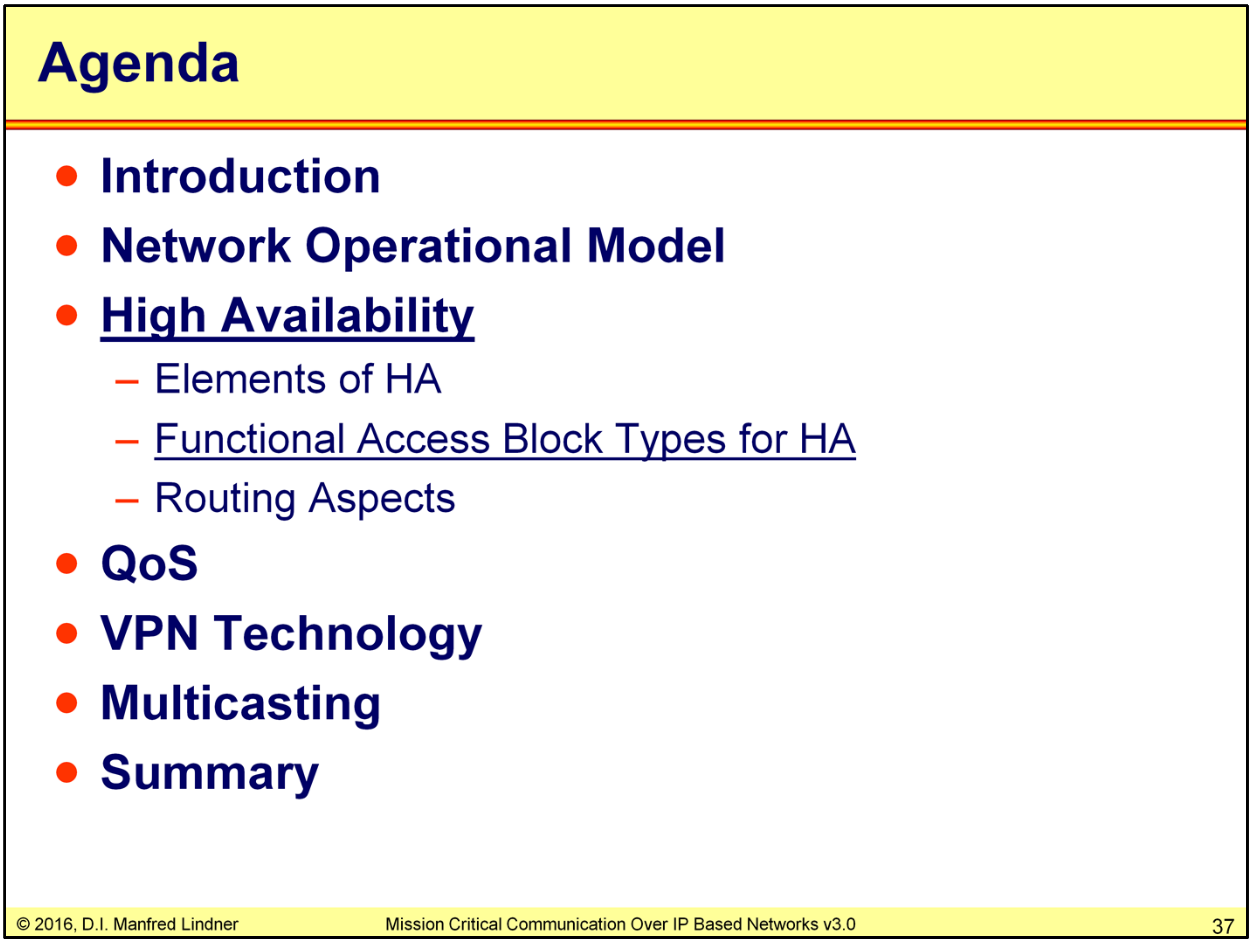
On Layer 3 (e.g. IP routing): HSRP (Hot Standby Routing Protocol) and OSPF Fast (subsecond) Hello

On Layer 7 (e.g. Application): Switchover triggered by supervision done by the application it selves.

Timeouts of these mechanisms need to be harmonized and their standard behavior has to be tuned.

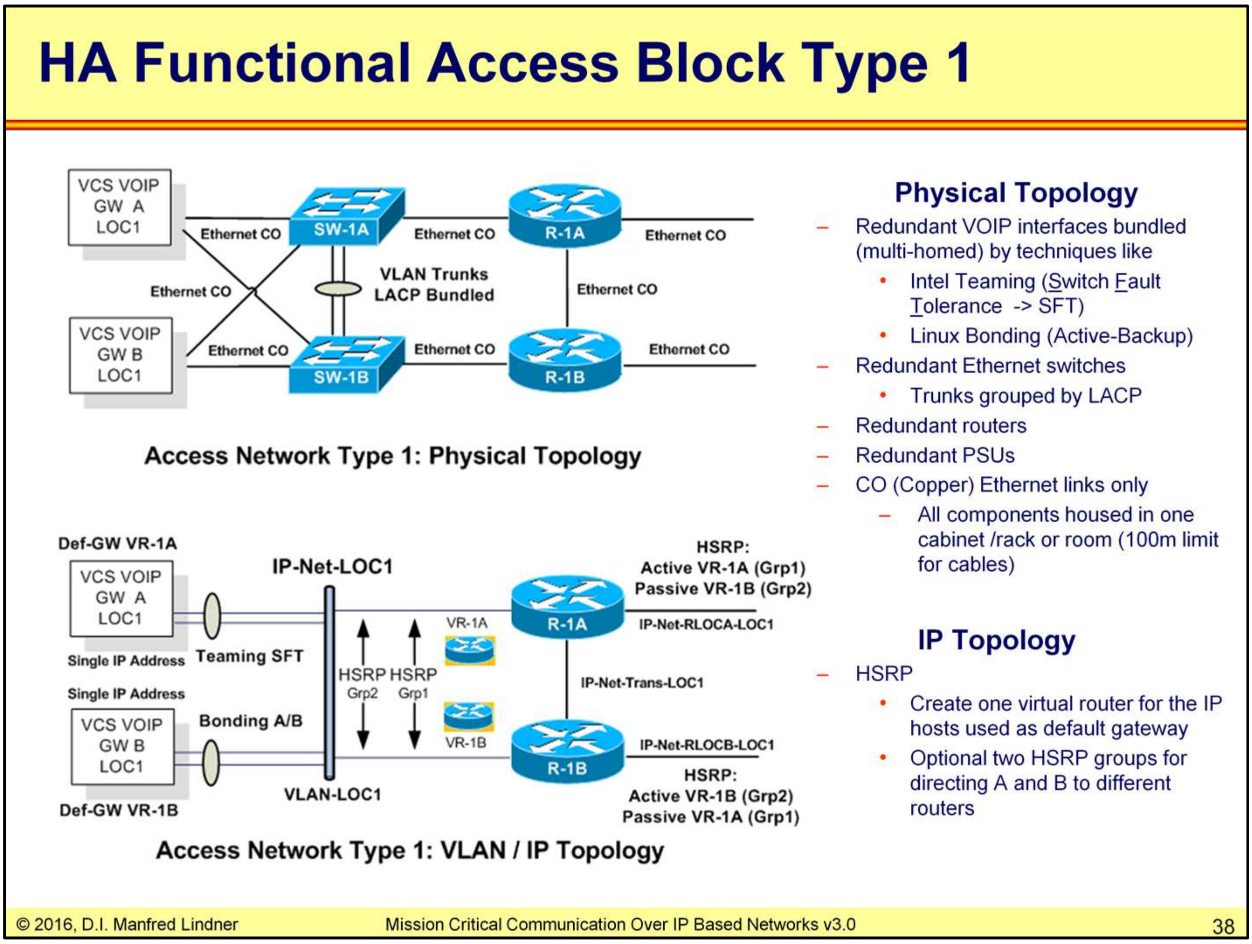
Golden rules: The less convergence mechanisms are used the better. Fast reaction is the enemy of stability. Burst of failures need a dampening mechanism in order to avoid a network meltdown. Summarization of IP routes following the physical topology of the network relieves administration/operation, keeps changes of routing table entries in case of network topology events local, minimize amount of routing table entries in the core network and may accelerate routing convergence in case of a network failure Last but not least effects on failure repair (rerouting back to failure free operation) should be analyzed and tuned to an acceptable level (e.g. done during service time interval) because this may lead to connectivity outage for some time, too.

Keep in mind: IP routing protocols are excellent on “black-outs” (something get lost) but bad on “brown-outs” (something becomes weaker and weaker).



# Agenda

- Introduction
- Network Operational Model
- High Availability
  - Elements of HA
  - Functional Access Block Types for HA
  - Routing Aspects
- QoS
- VPN Technology
- Multicasting
- Summary



## HA Functional Access Block Type 1

Access Network Type 1: Physical Topology

Access Network Type 1: VLAN / IP Topology

### Physical Topology

- Redundant VOIP interfaces bundled (multi-homed) by techniques like
  - Intel Teaming (Switch Fault Tolerance -> SFT)
  - Linux Bonding (Active-Backup)
- Redundant Ethernet switches
  - Trunks grouped by LACP
- Redundant routers
- Redundant PSUs
- CO (Copper) Ethernet links only
  - All components housed in one cabinet /rack or room (100m limit for cables)

### IP Topology

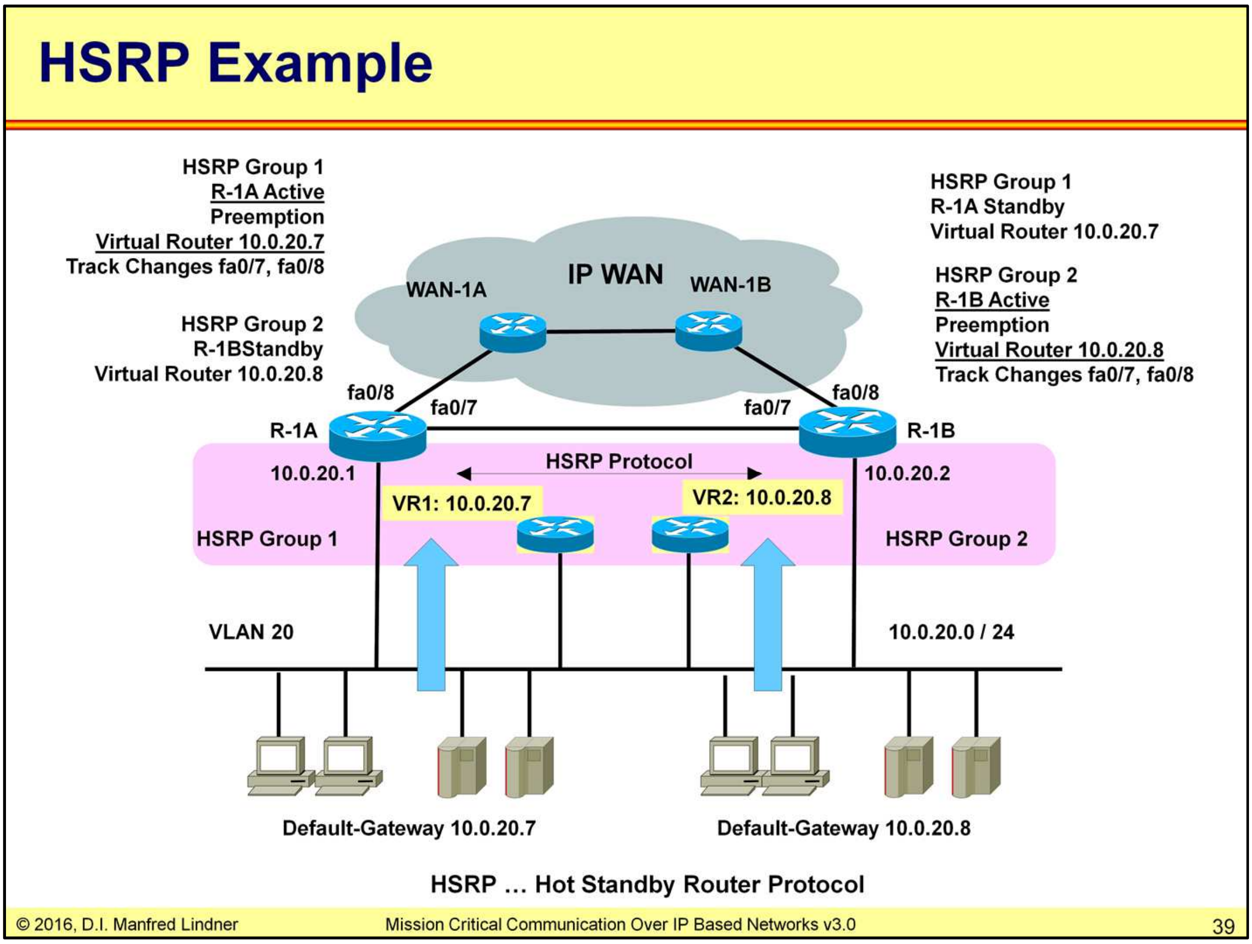
- HSRP
  - Create one virtual router for the IP hosts used as default gateway
  - Optional two HSRP groups for directing A and B to different routers

First let us have a look into more detail how end-systems (VCS - Voice communication systems) can be connected to the Ethernet / IP infrastructure.

HA Access type 1 shows a basic access method based on two Ethernet switches and two IP routers. On the left side there is the physical topology representing the cable plan. On the right side there is the (logical) VLAN/IP topology representing the VLAN-IDs, IP Subnets and HSRP aspects.

Physical part of network access type 1: Each VCS is connected via copper Ethernet to two different Ethernet switches. Switches are trunked together by two copper Ethernet links. Each Ethernet switch is connected with a single copper Ethernet link to an IP router. The routers are connected together and to the IP WAN using one copper Ethernet link per connection. The Ethernet link between the two routers will keep a failure local in some failure situations, which will be explained a little bit later in this chapter. So a failure of one physical link or component will not create a problem because of overall redundancy.

Logical part of network access type 1: The Ethernet switches are using LACP (Link Aggregation Control Protocol) for bundling the trunks in between to create just one logical trunk seen from the Spanning-Tree protocol, hence only one STP path is in between and no STP convergence has to take place in case of a failure. Both VCS shares the same VLAN and IP subnet. The two physical interfaces of a VCS are bundled to one logical interface representing a single IP address and single Ethernet (MAC) address. Following bundling methods are possible: Either Intel teaming in SFT (Switch Fault Tolerance) mode in case of Microsoft environment or LINUX/UNIX bonding in A/B (Active/Backup) mode. In both cases just one link (the active) is used and a switch over to the other link (the standby) is performed in case of a failure. Direct failures (link cable broken) are easy to detect. Indirect failures (like Ethernet switch failure but link to Ethernet switch still active) need additional methods to be implemented. These methods are periodical ARP checks of the default-gateway performed by the VCS and periodical HSRP hellos performed by the routers. VCS are configured with just one single default-gateway IP address, which is represented by the corresponding virtual IP address of the HSRP group. HSRP is performed among the routers to establish one virtual router for the corresponding HSRP group. Two different HSRP groups are used to direct traffic from the VCSs to different physical routers during normal operation and to disturb only one VCS by the HSRP convergence time in case of a failure. The other VCS will not experience any disturbance. During normal operation one router (e.g. R1-A) will be active for a virtual IP address VR-1A and the other router (R1-B) will be standby. In case of a failure the other will take over. After failure repair R1-A again will become the active router. By applying the same schema for VR-1B but with opposite logic the overall decoupling of VCS-A and VCS-B concerning single-point of failures can be achieved.



## HSRP Example

HSRP Group 1
R-1A Active
Preemption
Virtual Router 10.0.20.7
Track Changes fa0/7, fa0/8

HSRP Group 2
R-1BStandby
Virtual Router 10.0.20.8

HSRP Group 1
R-1A Standby
Virtual Router 10.0.20.7

HSRP Group 2
R-1B Active
Preemption
Virtual Router 10.0.20.8
Track Changes fa0/7, fa0/8

HSRP … Hot Standby Router Protocol

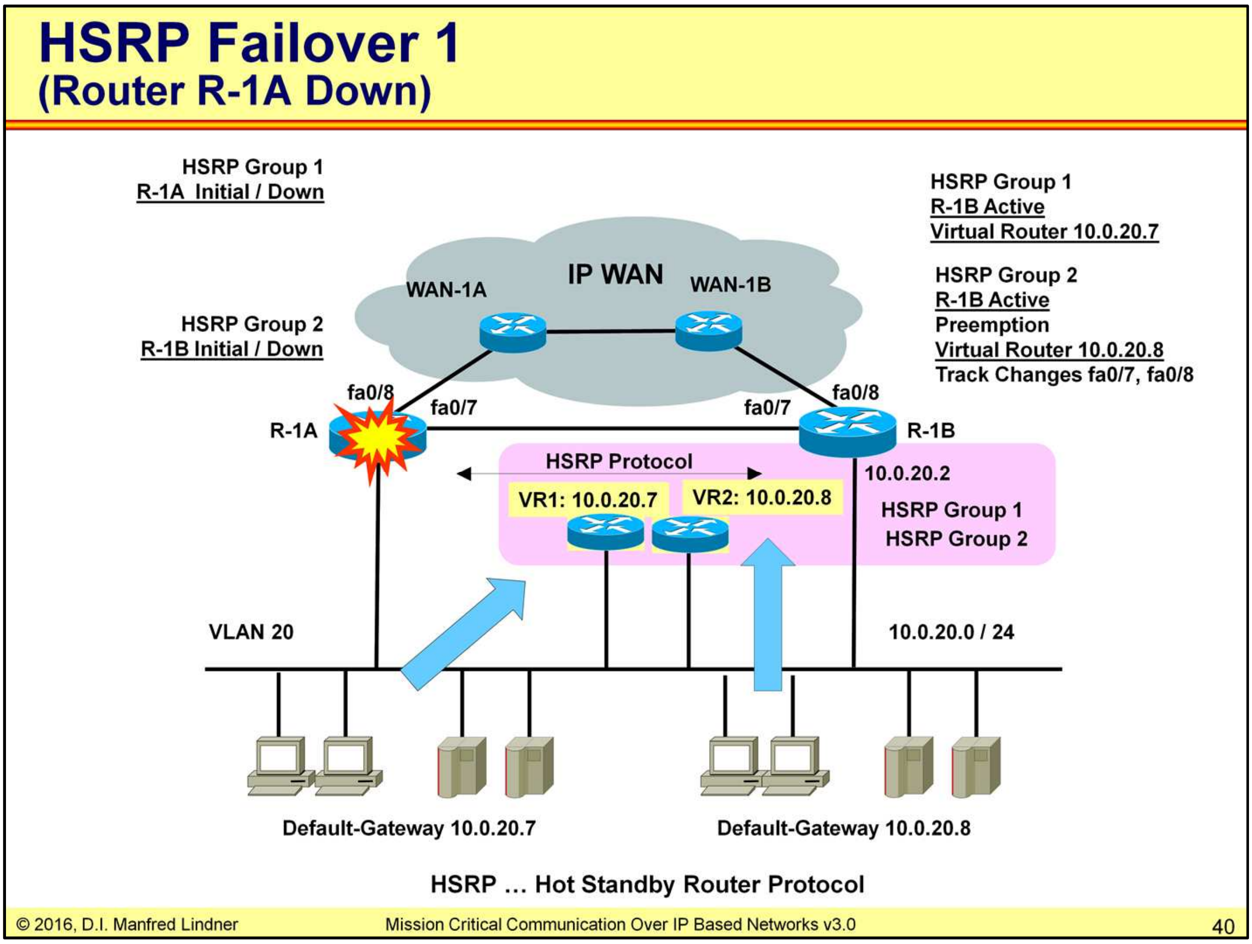
© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 39

Basics of HSRP:

A group of routers forms a HSRP group. The group is represented by a virtual router with a virtual IP address and virtual MAC address for that group. IP hosts are configured with the virtual IP address as default gateway. One router is elected by HSRP as the active router, one router is elected as the standby router of that group.

HSRP messages are UDP messages to port 1985, addressed to IP multicast 224.0.0.2 using Ethernet multicast frames (HSRP version 1; IP multicast 224.0.0.102 is used for HSRP version 2). Active router responds to ARP request directed to the virtual IP address with the virtual MAC address. Standby router supervises if the active router is alive by listening to HSRP messages sent by the active.

Active and standby is defined by HSRP priority. Priority value can be configured. The higher the better. Preemption allows to give up the role of the active router when a router with higher priority is reported by HSRP messages. Preemption happens either when the failed router comes back, a better router has been activated or tracking has changed priority.



## HSRP Failover 1
## (Router R-1A Down)

HSRP Group 1
R-1A Initial / Down

HSRP Group 2
R-1B Initial / Down

HSRP Group 1
R-1B Active
Virtual Router 10.0.20.7

HSRP Group 2
R-1B Active
Preemption
Virtual Router 10.0.20.8
Track Changes fa0/7, fa0/8

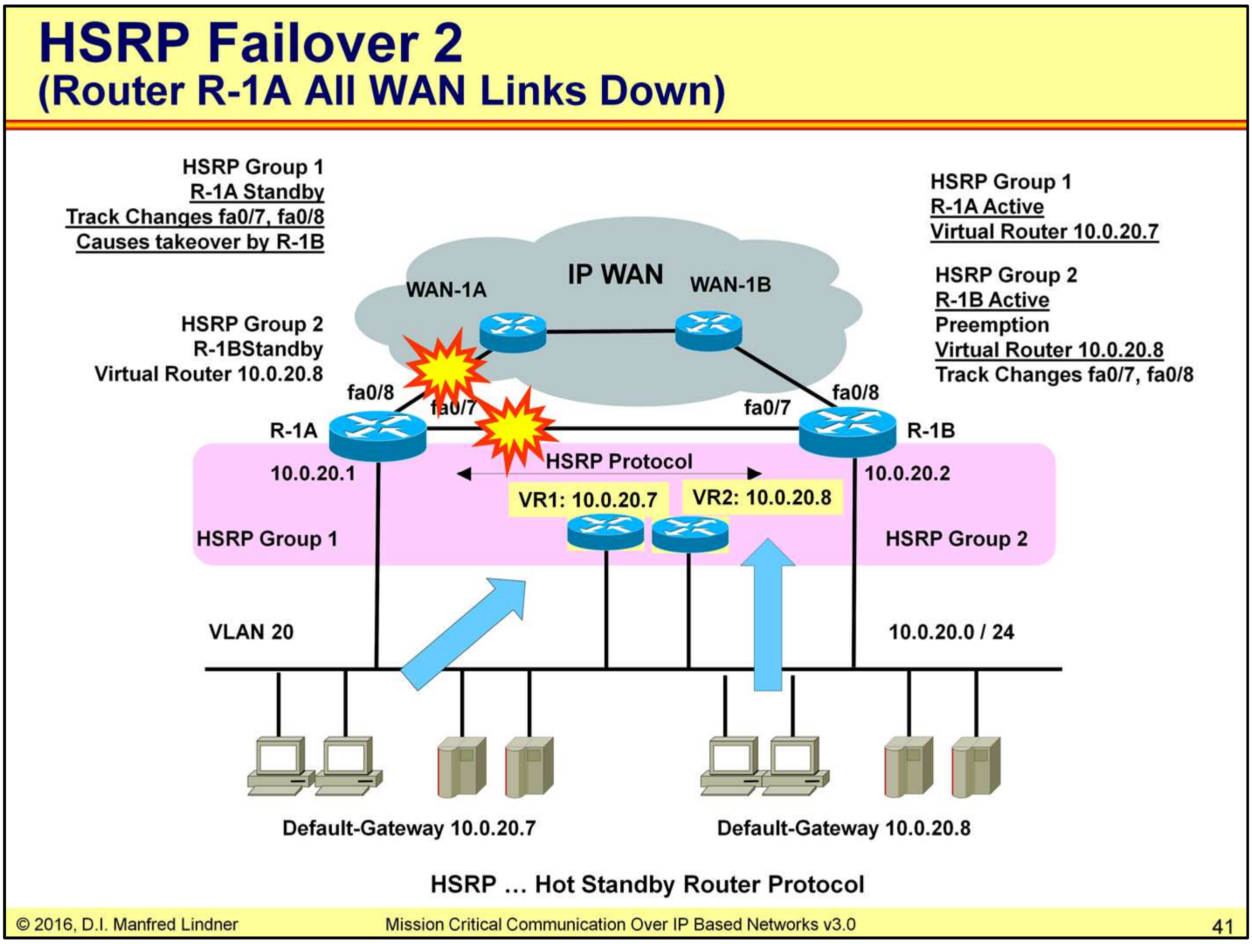
HSRP … Hot Standby Router Protocol

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 40

Failover scenario 1:

Active router is not reachable via LAN. Standby router will take over active role. Timing depends on HSRP hello message interval and hold-time (default hello-time = 3 seconds, default hold-time = 10 seconds, HSRP version 1; for  HSRP version 2 hello-time is configurable between 15 - 999 milliseconds and hold-time is configurable up to 3000 milliseconds).

.



## HSRP Failover 2
## (Router R-1A All WAN Links Down)

HSRP Group 1
R-1A Standby
Track Changes fa0/7, fa0/8
Causes takeover by R-1B

HSRP Group 2
R-1BStandby
Virtual Router 10.0.20.8

HSRP Group 1
R-1A Active
Virtual Router 10.0.20.7

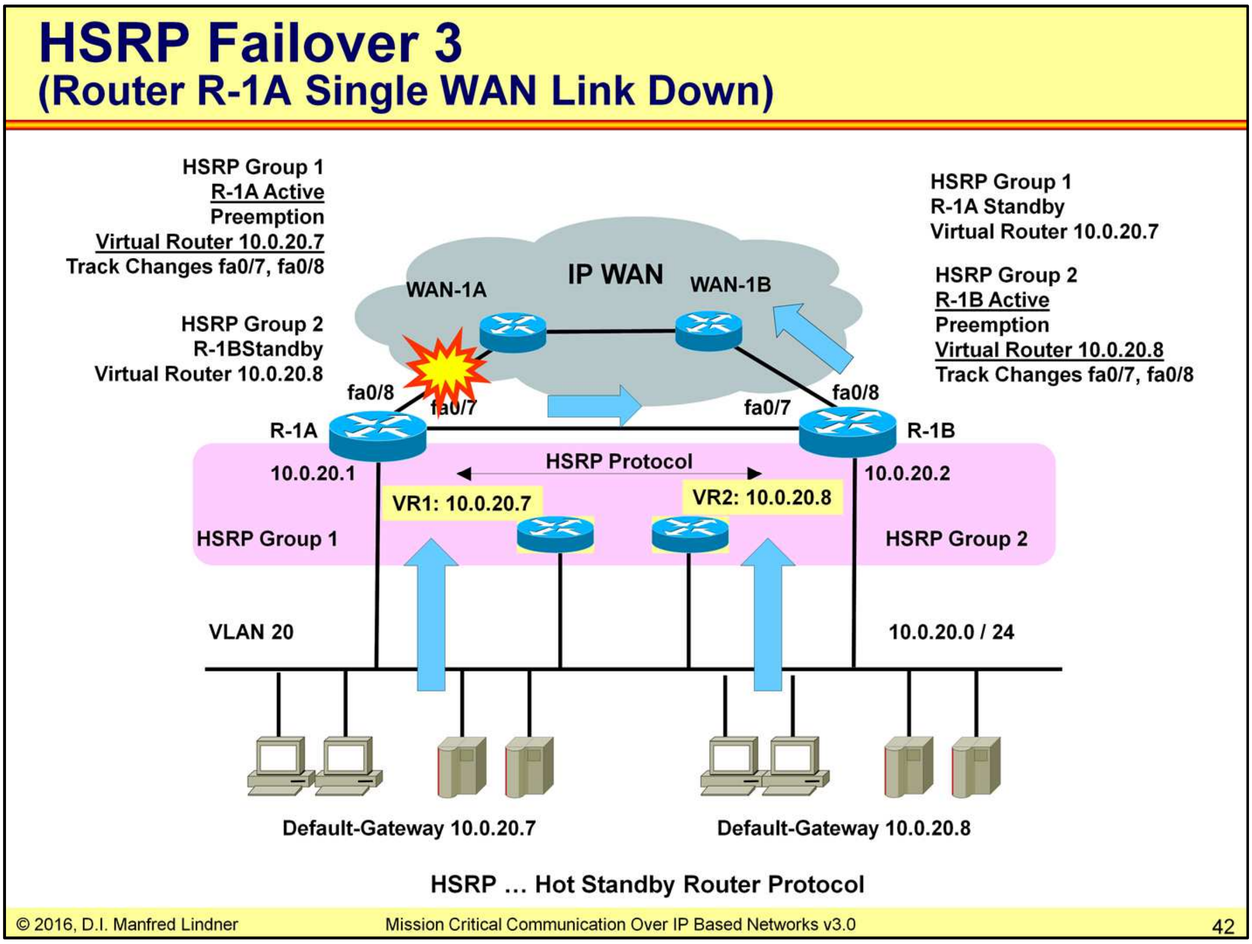
HSRP Group 2
R-1B Active
Preemption
Virtual Router 10.0.20.8
Track Changes fa0/7, fa0/8

HSRP … Hot Standby Router Protocol

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 41

Failover scenario 2:

Active router losses connectivity to both WAN interfaces. Tracking will lower the priority of the active router. The standby router will take over even the current active router is still reachable via LAN.



# HSRP Failover 3
## (Router R-1A Single WAN Link Down)

HSRP Group 1
R-1A Active
Preemption
Virtual Router 10.0.20.7
Track Changes fa0/7, fa0/8

HSRP Group 2
R-1BStandby
Virtual Router 10.0.20.8

HSRP Group 1
R-1A Standby
Virtual Router 10.0.20.7

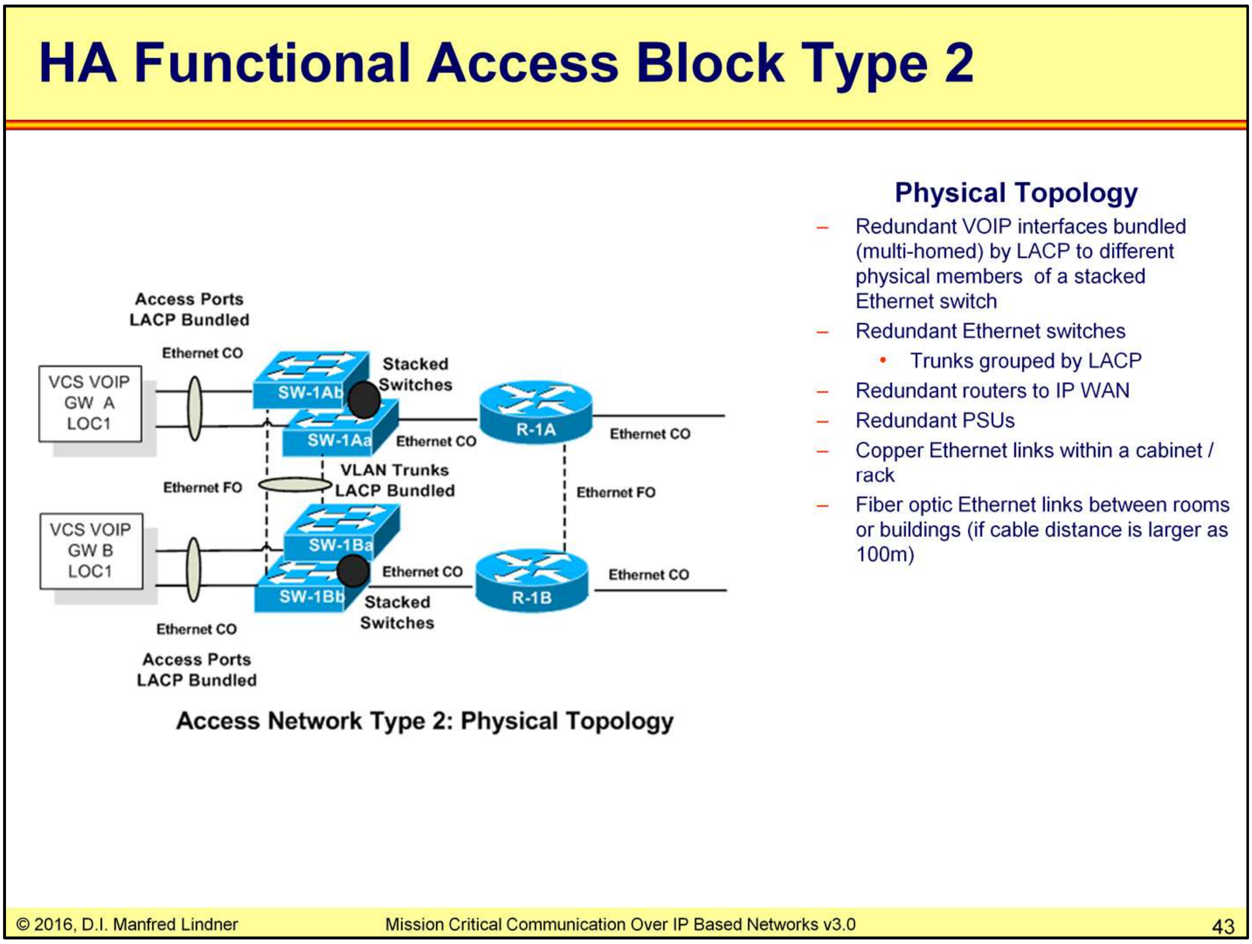
HSRP Group 2
R-1B Active
Preemption
Virtual Router 10.0.20.8
Track Changes fa0/7, fa0/8

HSRP … Hot Standby Router Protocol

© 2016, D.I. Manfred Lindner   Mission Critical Communication Over IP Based Networks v3.0   42

Failover scenario 3:

Active router losses connectivity to single WAN interface. Tracking will lower the priority of the active router. But resulting priority still higher than priority of standby router. Therefore no switchover will take place. Routing will forward packets using the cross link to R-1B (of course an appropriate routing have to be established meaning running a dynamic routing protocol across the link R-1A to R1-B with a lower routing metric compared to vlan 20.)



## HA Functional Access Block Type 2

Access Network Type 2: Physical Topology

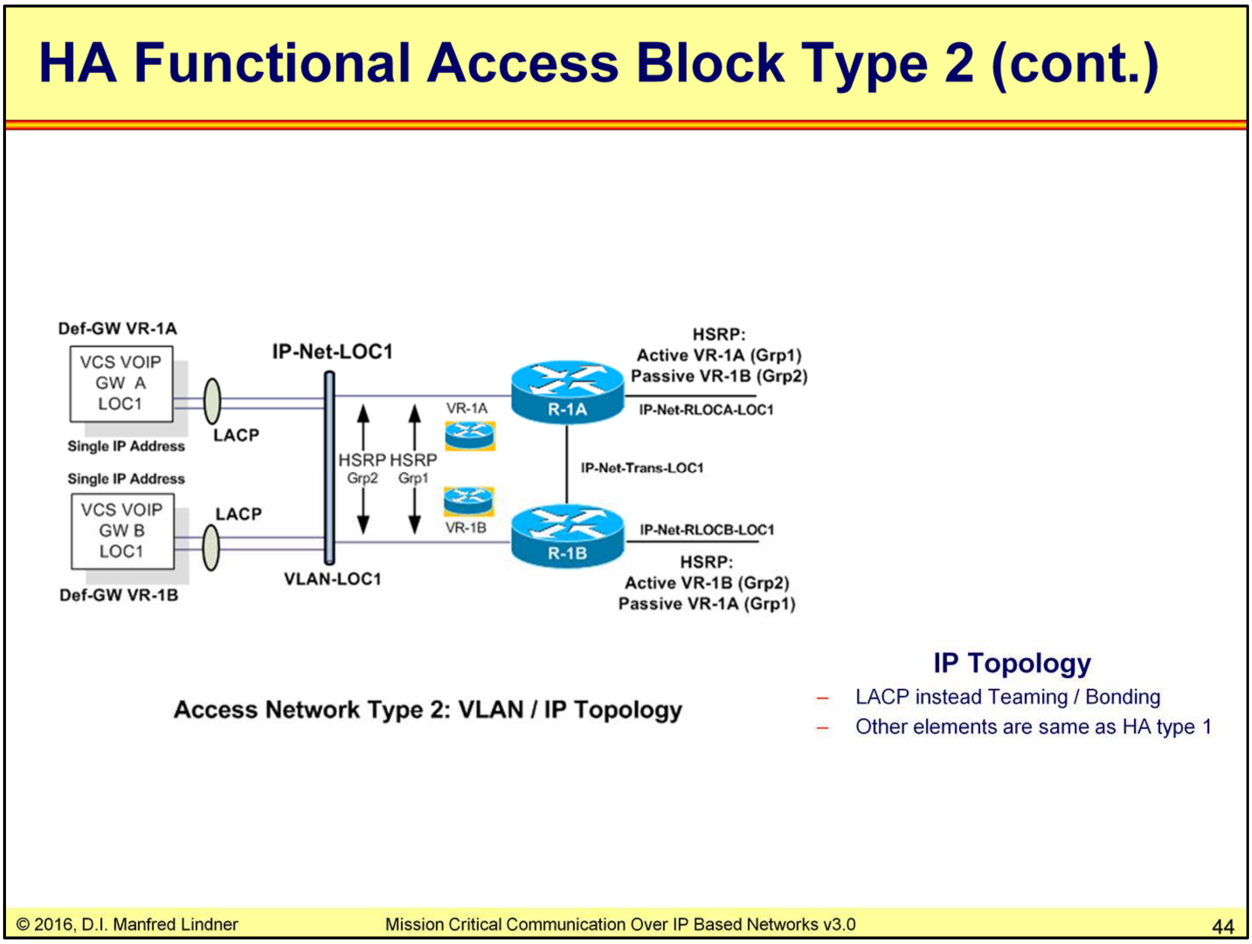
### Physical Topology

- Redundant VOIP interfaces bundled (multi-homed) by LACP to different physical members of a stacked Ethernet switch
- Redundant Ethernet switches
  - Trunks grouped by LACP
- Redundant routers to IP WAN
- Redundant PSUs
- Copper Ethernet links within a cabinet / rack
- Fiber optic Ethernet links between rooms or buildings (if cable distance is larger as 100m)

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 43

HA Access type 2 shows a modified access method following the same principle as access type-1 which is eligible for in two cases: Either more Ethernet ports for end-systems are necessary than a single Ethernet switch can provide. Or all components of part A should be separated from all components of pars B by usage of fiber optic Ethernet links because equipment should be housed in different cabinets or buildings.

Physical part of network access type 2: Ethernet switches are stacked in order to provide just one single logical Ethernet switch regarding Spanning-Tree protocol and management. Typical stacks allow forming a single logical switch out of up to 8 physical switches. Each VCS is connected via copper Ethernet to two different Ethernet switches of the same stack by using LACP instead of Teaming/SFT or Linux Bonding Active/Backup. Switches are trunked together using two fiber optic Ethernet links with LACP as in block type 1. Also the Ethernet link between the two routers is based on fiber optic Ethernet.



## HA Functional Access Block Type 2 (cont.)

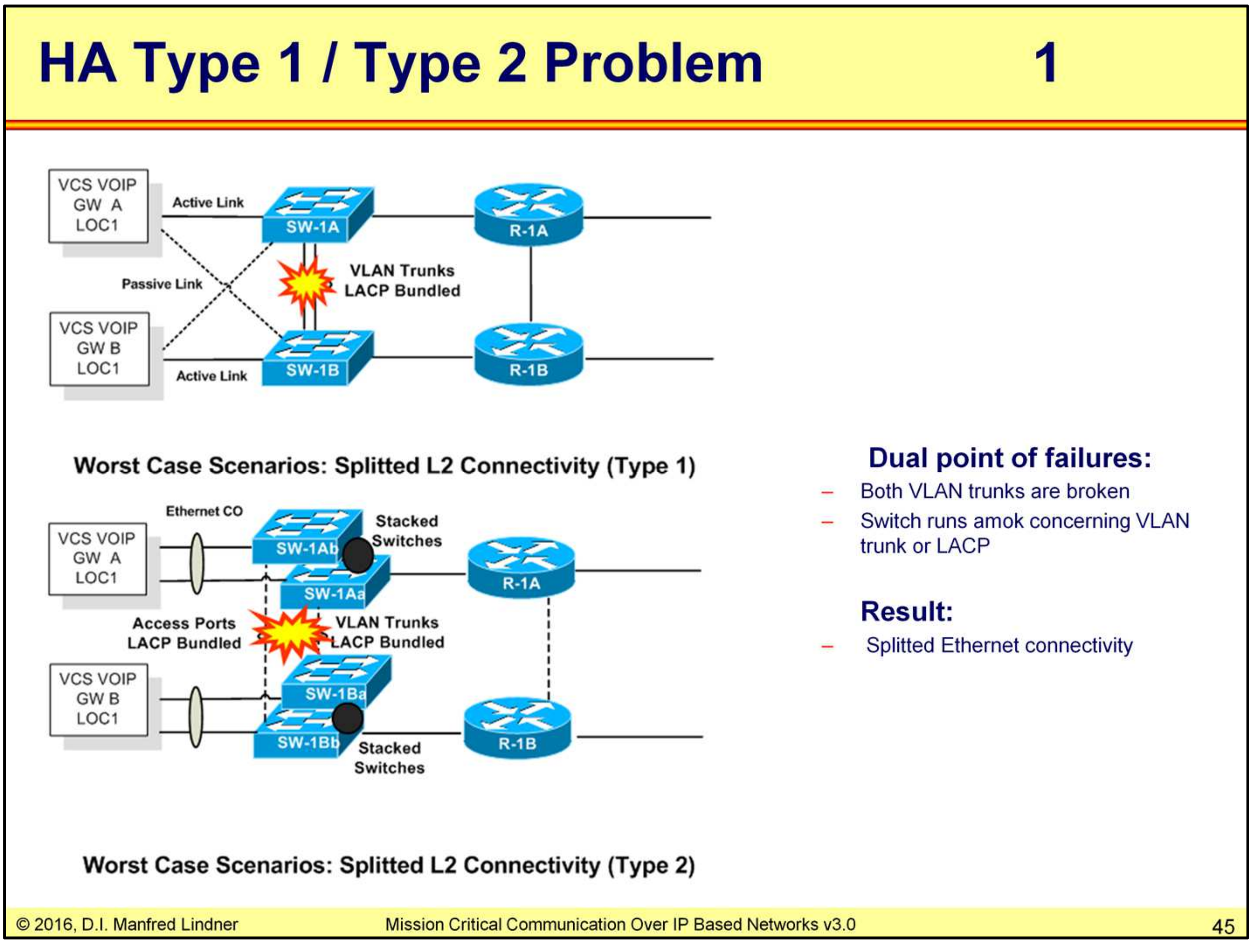
Access Network Type 2: VLAN / IP Topology

**IP Topology**

- LACP instead Teaming / Bonding
- Other elements are same as HA type 1

© 2016, D.I. Manfred Lindner    Mission Critical Communication Over IP Based Networks v3.0    44

Logical part of network access type 2: It is pretty much the same as for type 1. Both VCS shares the same VLAN and IP subnet and HSRP works in the same style. The only difference is the usage of LACP to bundle the two physical interfaces of a VCS together to just one logical link which now allows load balancing and avoids any teaming-SFT or bonding-A/B convergence to take place in case of a failure.



## HA Type 1 / Type 2 Problem 1

Worst Case Scenarios: Splitted L2 Connectivity (Type 1)

Worst Case Scenarios: Splitted L2 Connectivity (Type 2)

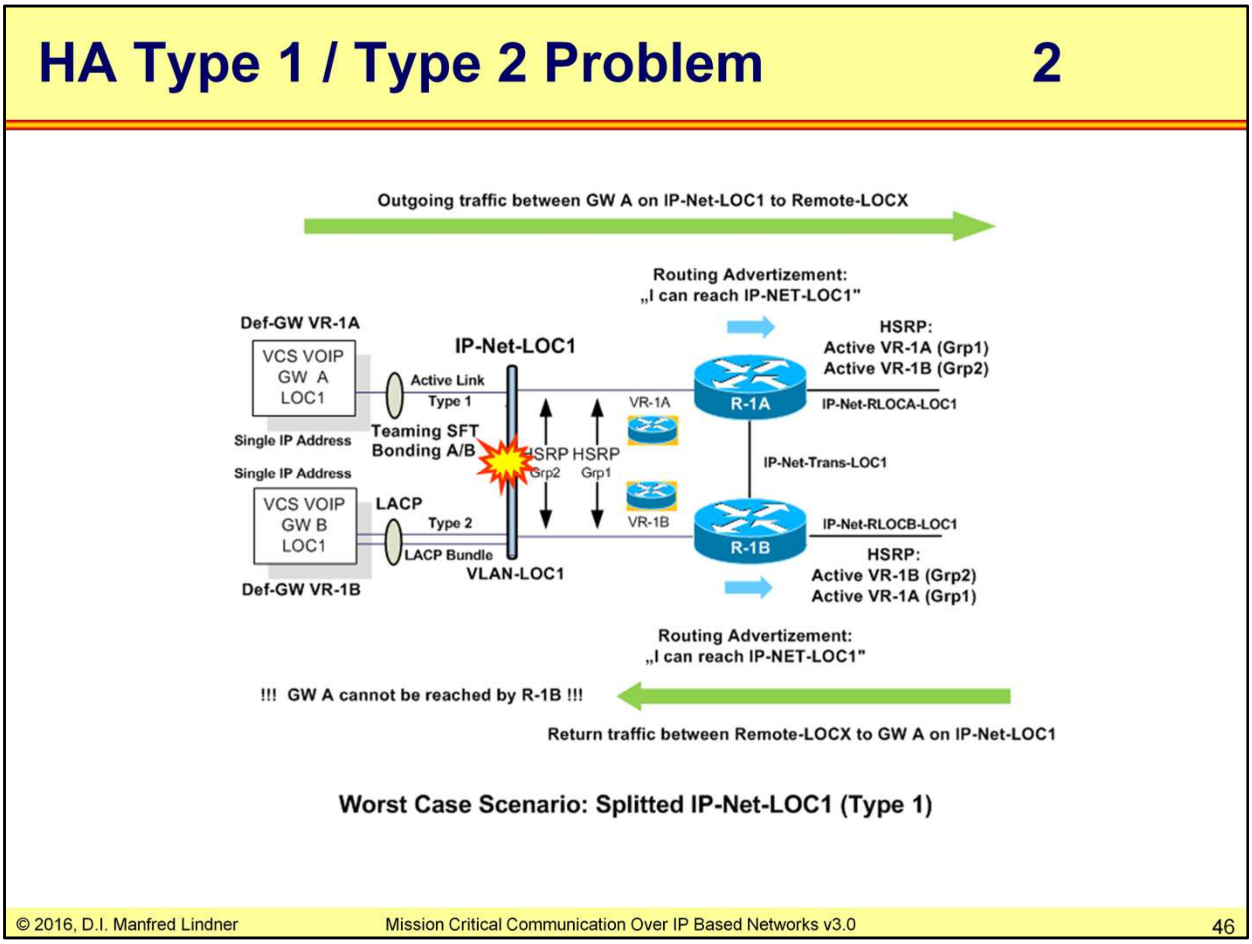
**Dual point of failures:**

- Both VLAN trunks are broken
- Switch runs amok concerning VLAN trunk or LACP

**Result:**

- Splitted Ethernet connectivity

Now let us look what can happen to access type 1 or 2 in a worst case scenario where parts are getting separated by applying dual point of failures. Splitted L2 connectivity shows such a worst case scenario. The interconnection of the Ethernet switches - although robust against single point of failures by usage of two separated Ethernet links – can lead to splitting the L2 into separated parts. Either if both links are broken or the LACP logic in the switches has a problem the result is a catastrophe for the IP view of the network.  Next slide explains why.



## HA Type 1 / Type 2 Problem 2

Outgoing traffic between GW A on IP-Net-LOC1 to Remote-LOCX

Routing Advertizement:
„I can reach IP-NET-LOC1"

Def-GW VR-1A

VCS VOIP
GW A
LOC1

IP-Net-LOC1

Active Link
Type 1

HSRP:
Active VR-1A (Grp1)
Active VR-1B (Grp2)

R-1A

IP-Net-RLOCA-LOC1

VR-1A

Single IP Address

Teaming SFT
Bonding A/B

HSRP Grp2

HSRP Grp1

IP-Net-Trans-LOC1

Single IP Address

VCS VOIP
GW B
LOC1

LACP

Type 2

LACP Bundle

VR-1B

R-1B

IP-Net-RLOCB-LOC1

HSRP:
Active VR-1B (Grp2)
Active VR-1A (Grp1)

VLAN-LOC1

Def-GW VR-1B

Routing Advertizement:
„I can reach IP-NET-LOC1"

!!! GW A cannot be reached by R-1B !!!

Return traffic between Remote-LOCX to GW A on IP-Net-LOC1

Worst Case Scenario: Splitted IP-Net-LOC1 (Type 1)

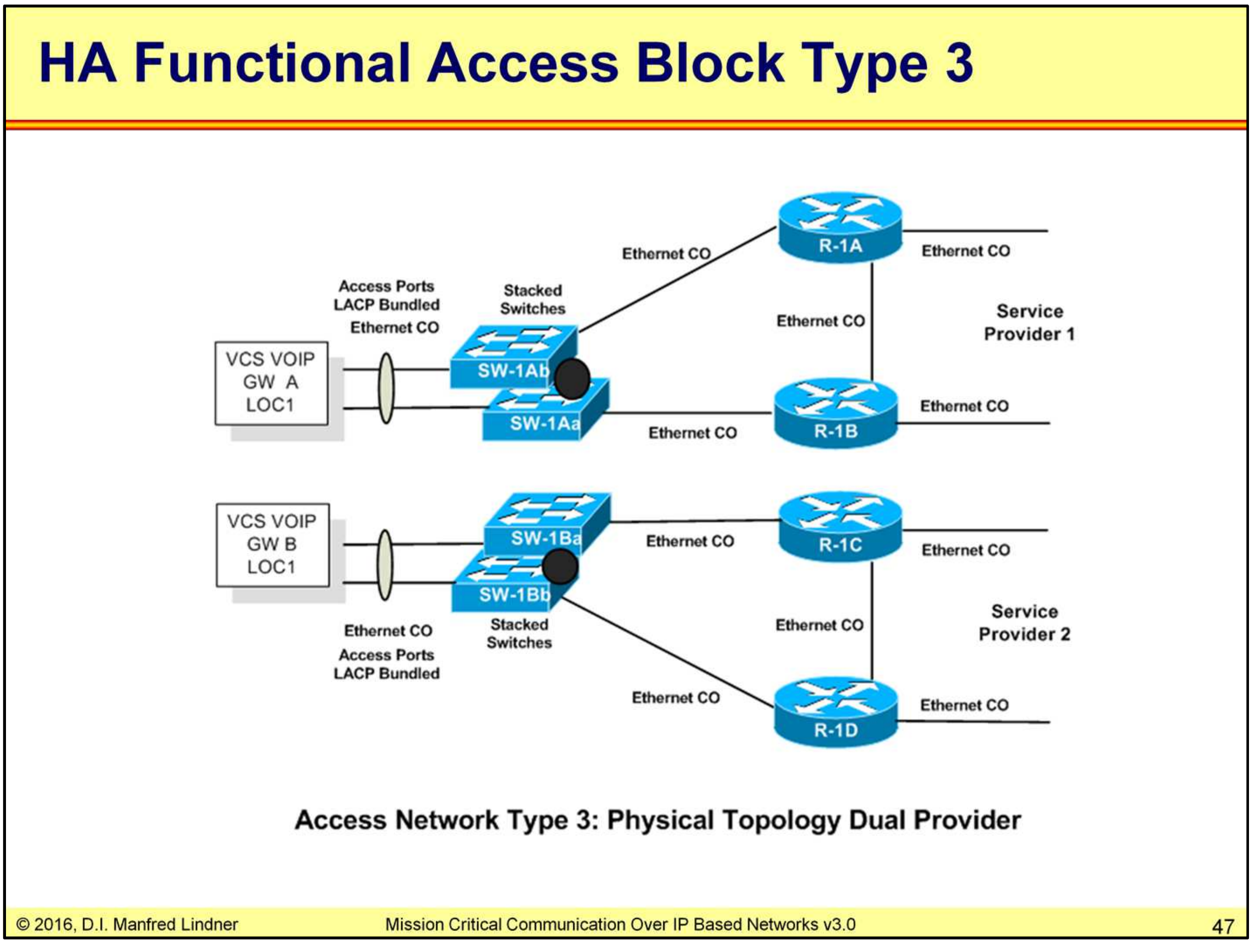
© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 46

An IP subnet never ever is allowed to be splitted in two separated parts because separated parts must have different IP Net-ID. Both routers still announce reachability of IP-Net-LOC1 in such a failure scenario. From the outside world both routers are valid entry points to reach IP-Net-LOC1. The summary routing metric seen from foreign routers in order to reach IP-NET-LOC1 will decide which entry should be used.

For example traffic still leaving VCS-A can reach a foreign location but return traffic for VCS-A may be received on the wrong entry point. So bidirectional communication will die. More badly it could happen that for some communication it still works. Or what works and what not depends on other operational parameters like routing metric changes. So fixing the problem based on symptoms is not so easy.

Another effect is that on both routers now both HSRP groups are active which is not a problem in the moment but will cause a longer convergence time in case of failure repair. Why? Because consistency in any active/active scenario of cluster needs some additional care in case the two systems see each other again after failure repair.

Unfortunately there are no automatic mechanisms available to overcome such situation. One possible approach is to perform OSPF locally between the two routers on IP-NET-LOC1 and to detect if HSRP is active for a group on both routers (SYSLOG). By correlation with a missing OSPF peer but still seeing the IP-NET-LOC1 in the OSPF topology database twice (router-LSA of R-1A and router-LSA of R-1B both shows the same IP network) a configuration script may be written to deactivate one of the router but that needs careful design and testing in order to automate.



## HA Functional Access Block Type 3

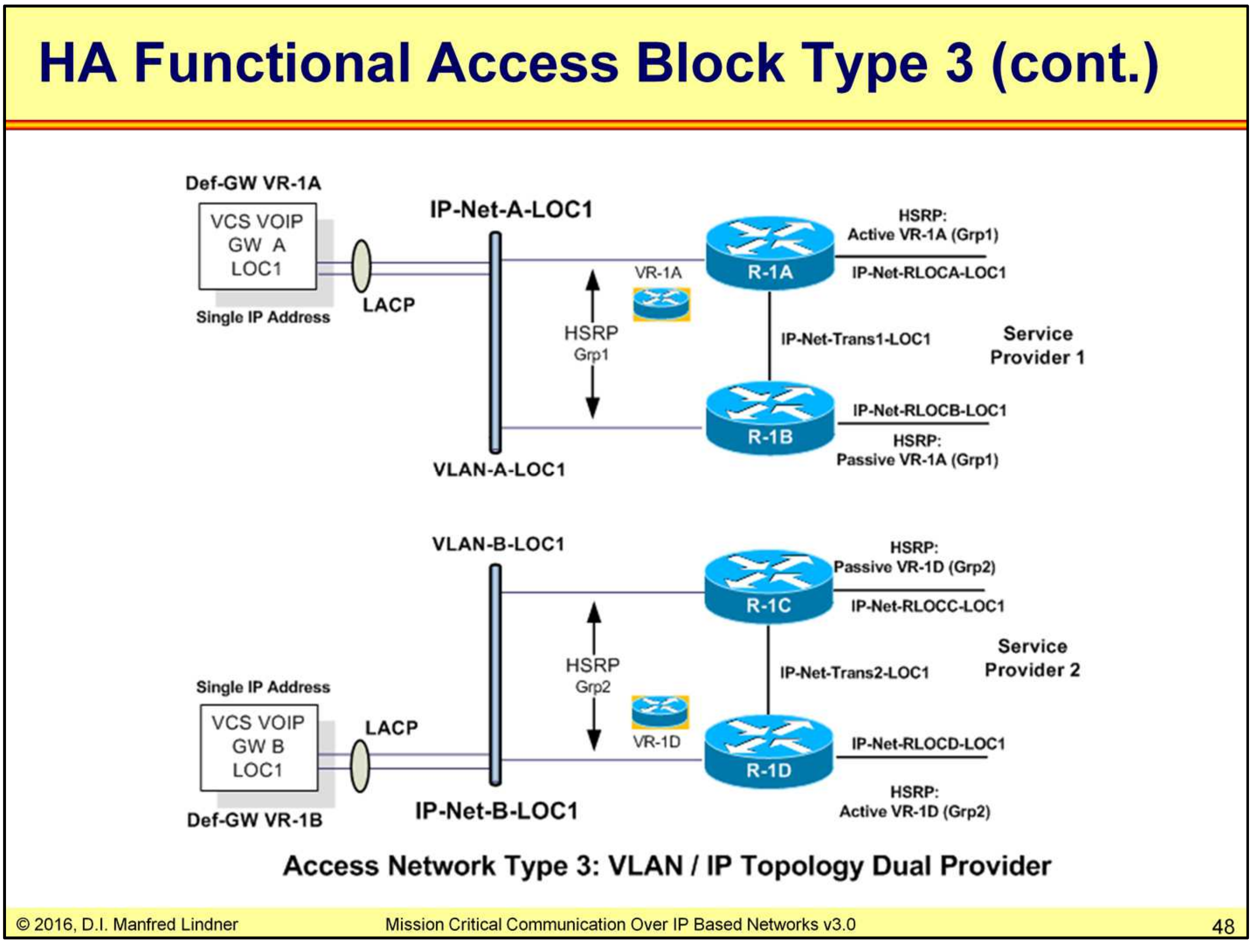
Access Network Type 3: Physical Topology Dual Provider

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 47

Now let us show some possibilities to overcome the splitted L2 connectivity problem.

One possible approach is usage of separate L2 domains for VCS-A and VCS-B resulting in different IP Subnet -> IP-Net-A-LOC1 and IP-Net-B-LOC1. VCS-A and VCS-B have no L2 connectivity (VLAN) anymore. Together by usage of separated IP WAN service providers for each of the VCSs the problem is solved.

HA access network type 3 shows such an approach. Type 3 access takes the principle of type 2 access (stacked switches, LACP from VCS to stacked switches) concerning physical topology.



## HA Functional Access Block Type 3 (cont.)

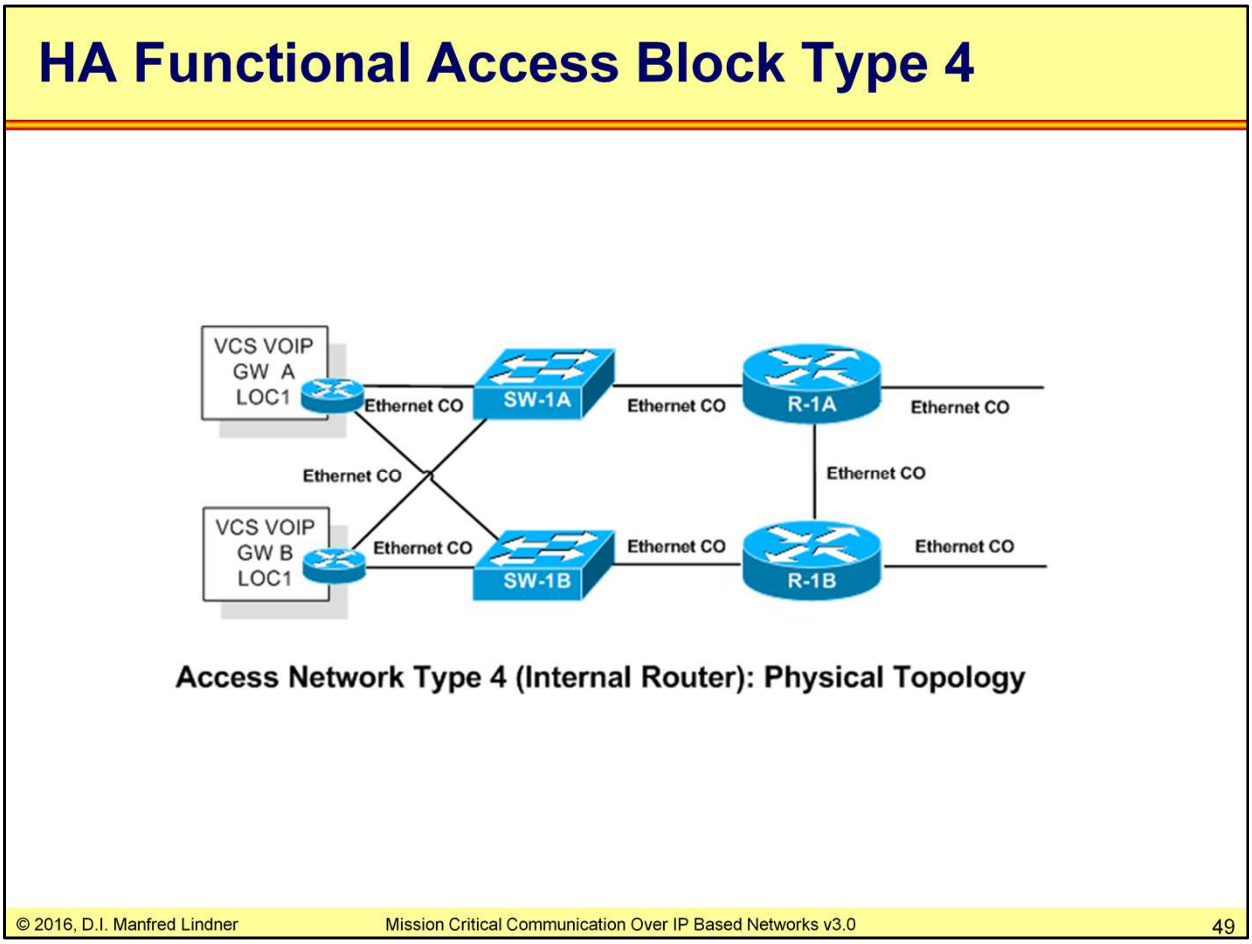
Access Network Type 3: VLAN / IP Topology Dual Provider

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 48

Type 3 access uses HSRP with only one HSRP group per VCS, feed the IP-WAN with dual IP routers per VCS and use a dedicated service provider per VCS. Service providers 1 and 2 have no visibility of each other and perform independently.

Alternative approach is usage of just one Ethernet link from the VCS to just one Ethernet switch connected to just one router leading to dedicated service provider in order to minimize number of components.

This follows either the philosophy of dual data transport over the IP-WAN and selected the better signal at the receiving end or just reduces the distance of any single point of failure to just one VCS system.



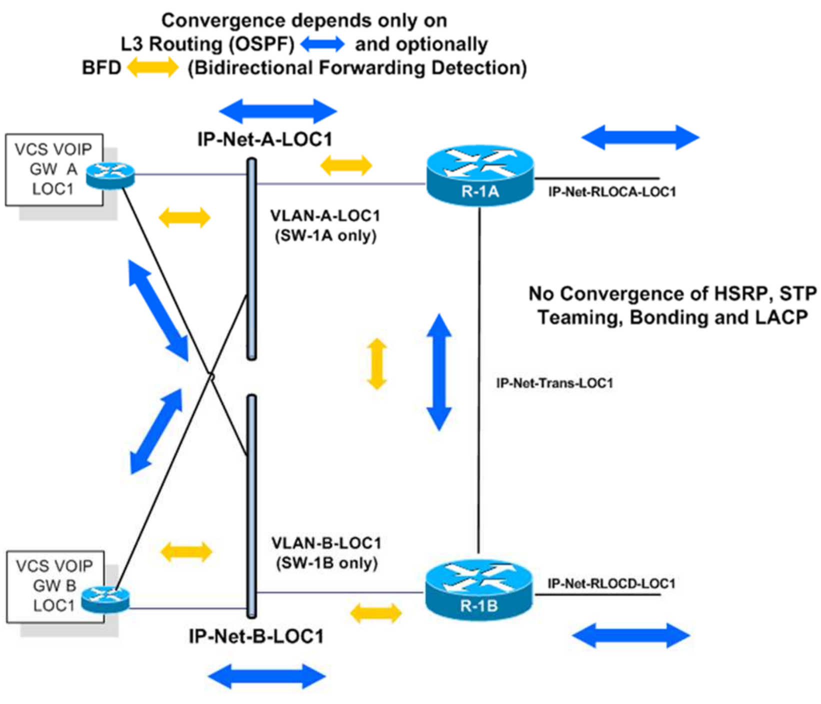
## HA Functional Access Block Type 4

Access Network Type 4 (Internal Router): Physical Topology

The next possible approach is introduction of L3 routing functionality in the end-systems. The big advantage of such an approach is that knowledge about the network topology is brought to the end-system and number of convergence mechanism is reduced to a minimum. If you are afraid that end-systems are burdened in case of a huge network topology, the view of the network topology for the end-systems might be suppressed to the local environment only. For example running two routing processes on R-1A (one for WAN, one for LAN) and redistributing local IP network via static routes into WAN and installing default-routes into LAN are a possible way of decoupling.

HA Access type 4 shows end-systems with internal IP router connected to two Ethernet switches. Every Ethernet switch just represents a single VLAN and is used to connect the three corresponding routers together of an IP subnet together (e.g. VLAN-A-LOC1 connecting VCS A, VCS-B and R-1A in IP-NET-A-LOC1 -> see next slide).

# HA Functional Access Block Type 4 (cont.)



Convergence depends only on L3 Routing (OSPF) and optionally BFD (Bidirectional Forwarding Detection)

No Convergence of HSRP, STP Teaming, Bonding and LACP

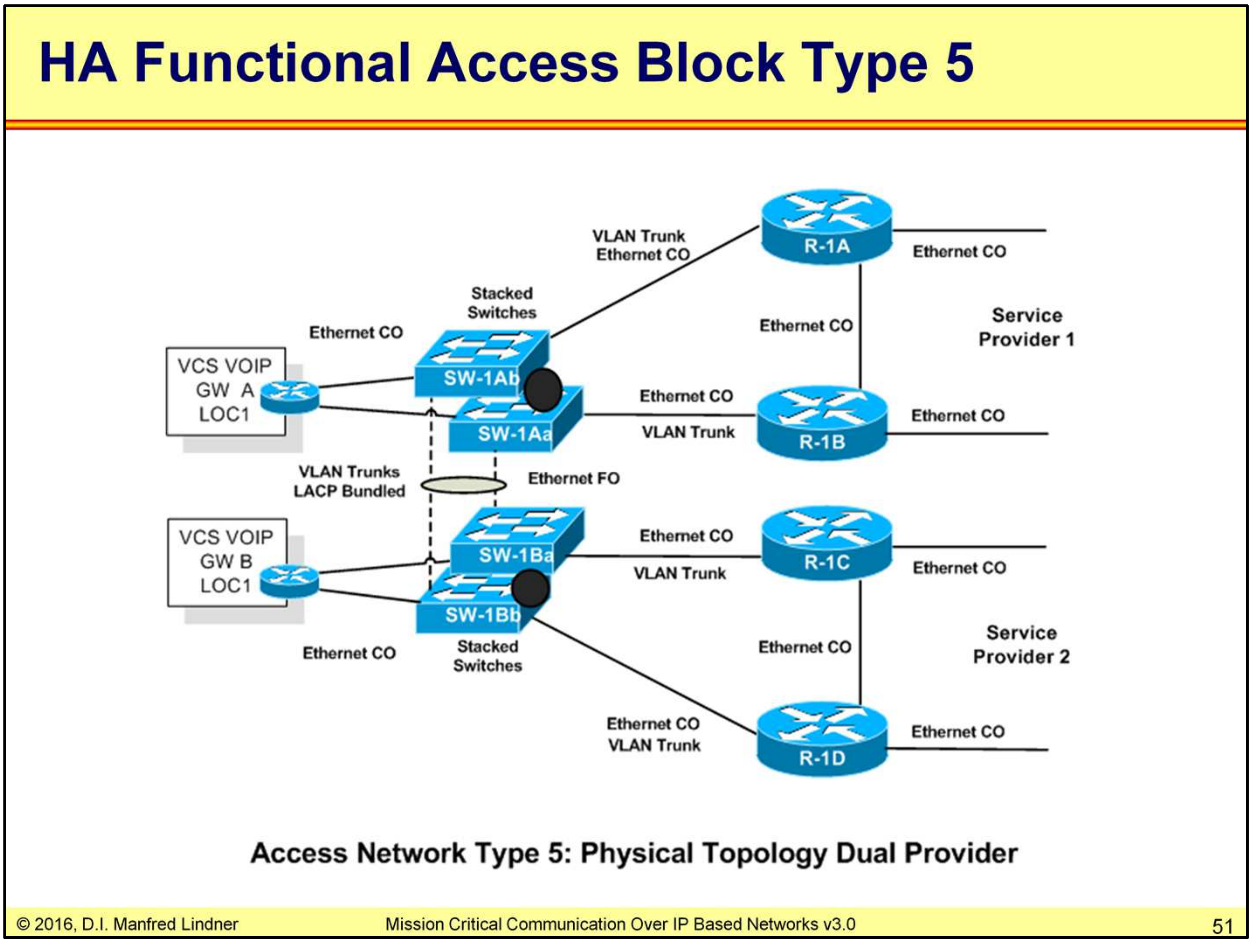
Access Network Type 4: VLAN / IP Topology

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 50

No more Spanning-Tree, no more teaming/bonding/LACP, and no more HSRP are necessary. Convergence time in case of failure depends on detection of direct failures (link down) or indirect failures (OSPF or BFD) timeouts.

An end-system is identified either by the two IP interface addresses or by an internal loopback address which is independent from the IP interface addresses. Hence the internal router advertize reachability of the loopback network address to the IP routers by usage of a routing protocol.

If the number of VCS interfaces (systems) increases on a location, routing scalability might become a problem. In order to keep failures local and separate routing at the location from the routing towards IP-WAN, OSPF areas can be used as alternate method to that mentioned on last page. So the IP WAN would become OSPF area 0 (OSPF backbone) and location 1 represents OSPF area 1. R-1A and R-1B become OSPF area border router between area 0 and area 1. VCS routers are just internal OSPF routers in area 1, hence any changes in the internal topology will not results in performance-greedy OSPF recalculations in OSPF area 0.



## HA Functional Access Block Type 5

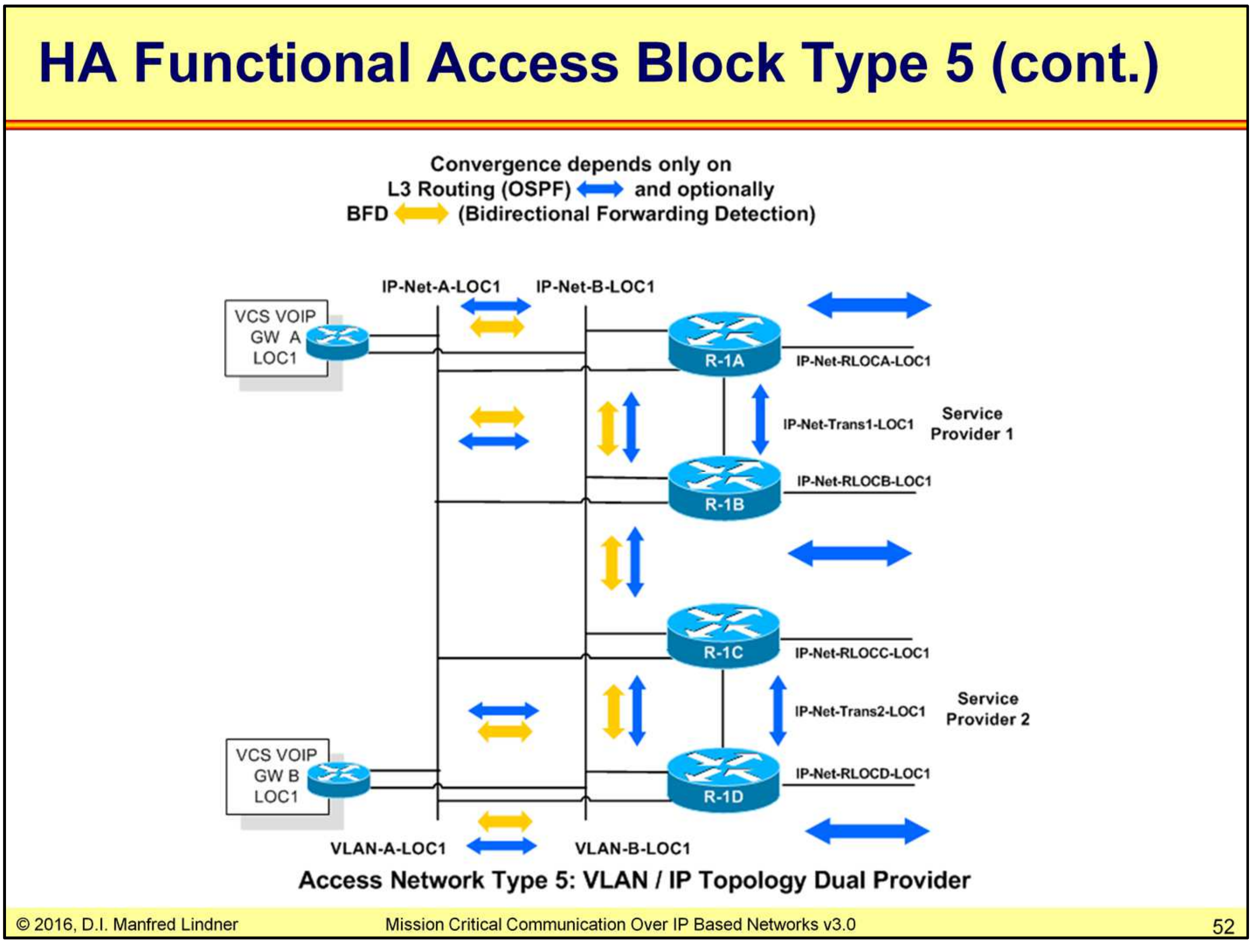
Access Network Type 5: Physical Topology Dual Provider

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 51

In case of dual service providers access type 5 may be used instead of access-type 4. HA access network type 5 shows usage of stacked switches to increase either Ethernet port number or minimize numbers of FO Ethernet links.

Further the service providers are interconnected at the customer location.

Now an end-system can select among four possible exit points to the IP- WAN depending either on IP routing only (best or available paths) or on a local policy decision.



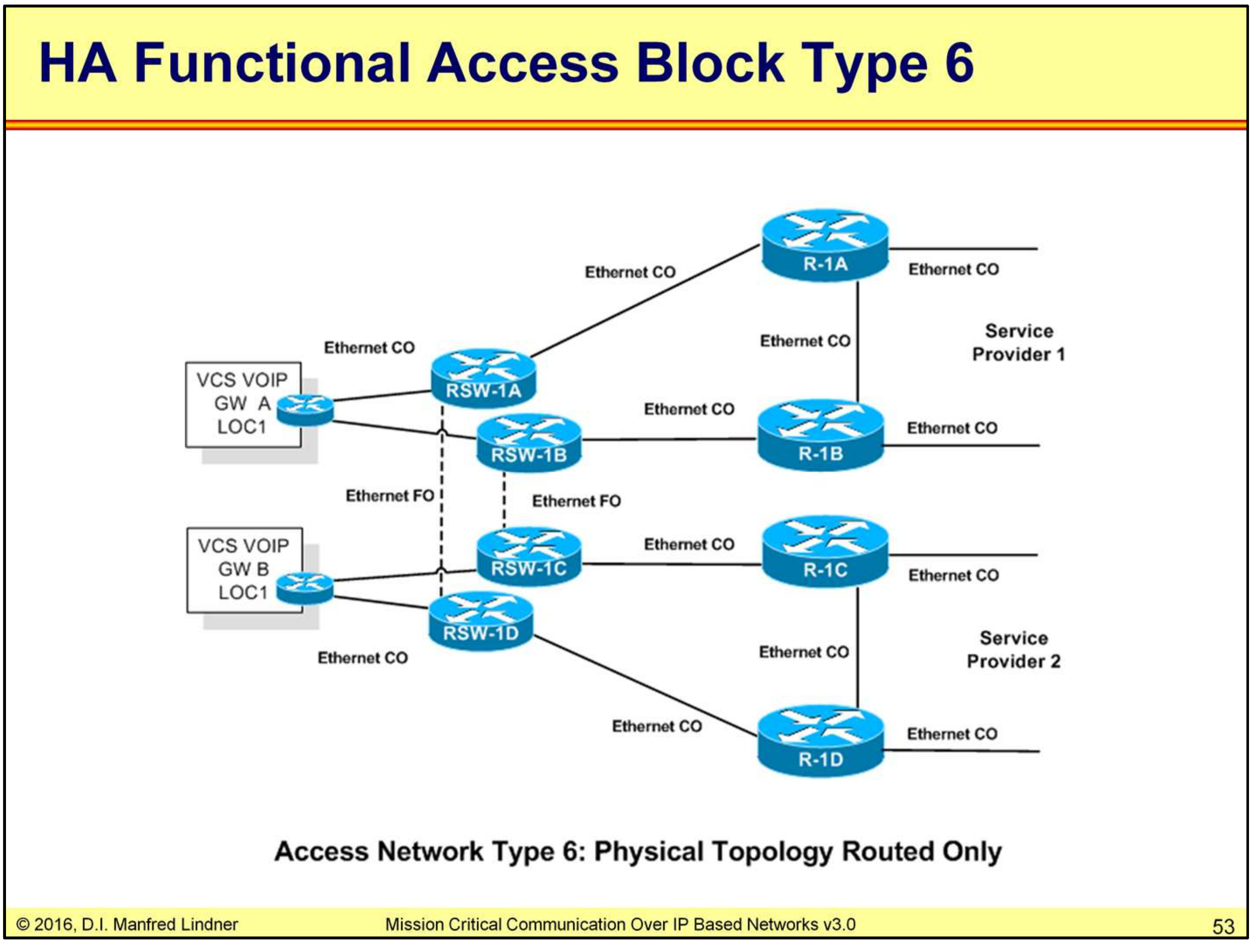
## HA Functional Access Block Type 5 (cont.)

Convergence depends only on L3 Routing (OSPF) ⟷ and optionally BFD ⟷ (Bidirectional Forwarding Detection)

Access Network Type 5: VLAN / IP Topology Dual Provider

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 52

Of course this approach suffers from a splitted L2 connectivity in a similar way as already described in the worst case scenario.



## HA Functional Access Block Type 6

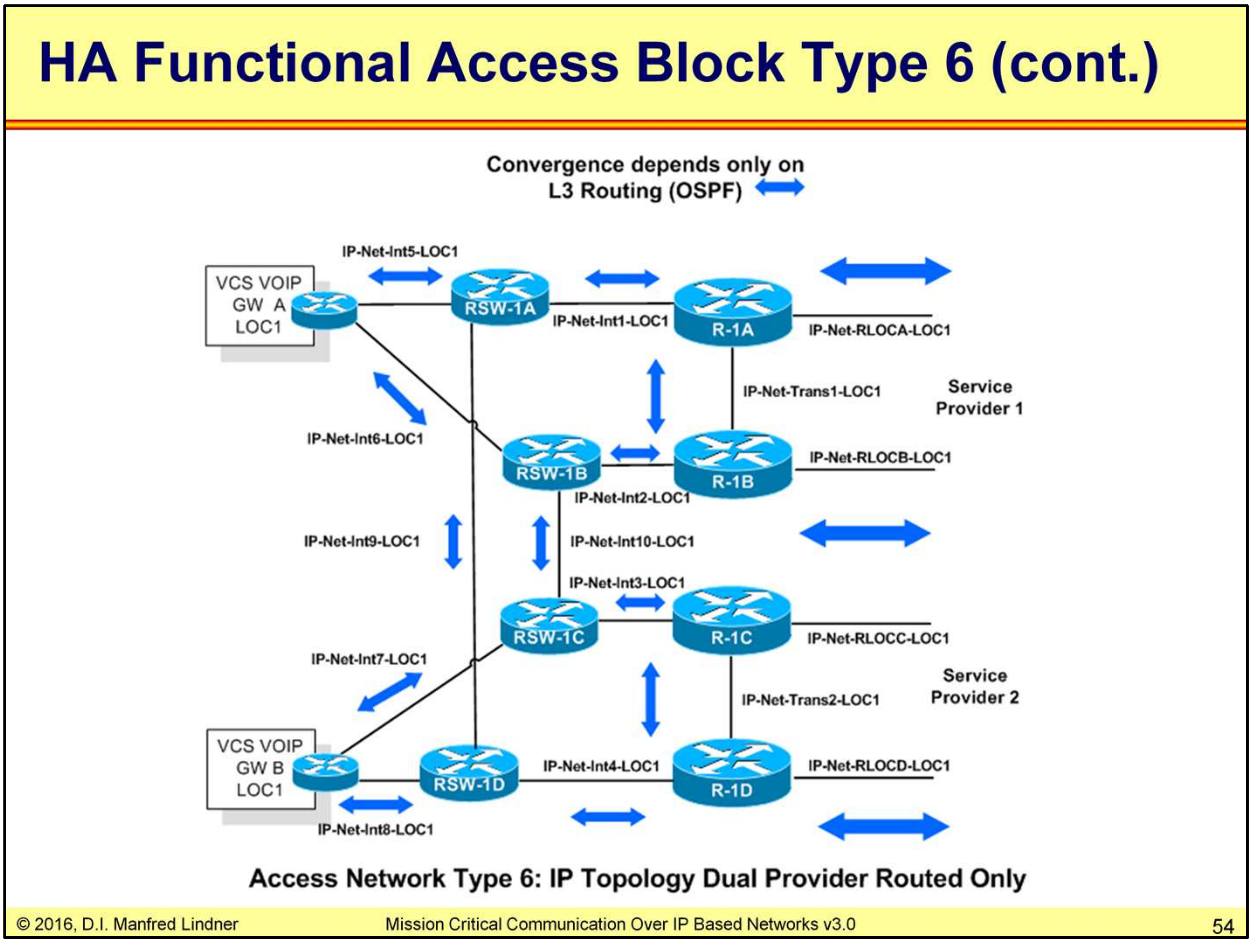
Access Network Type 6: Physical Topology Routed Only

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 53

The ultimate approach is an IP router only approach. HA access network type 6 shows usage of L3 switches instead of L2 Ethernet switches, which is a fundamental trend seen in future datacenter infrastructure network solutions (routing to the access layer).

No VLANs and no Ethernet switching anymore. The physical topology and IP topology are completely aligned. Only point-to-point Ethernet links are used and routing convergence depends on OSPF only.

Again in order to keep failures local and separate routing at the location from the routing towards IP-WAN, OSPF areas can be used as already described in elaborations about access type 4.



## HA Functional Access Block Type 6 (cont.)

Convergence depends only on L3 Routing (OSPF)

Access Network Type 6: IP Topology Dual Provider Routed Only

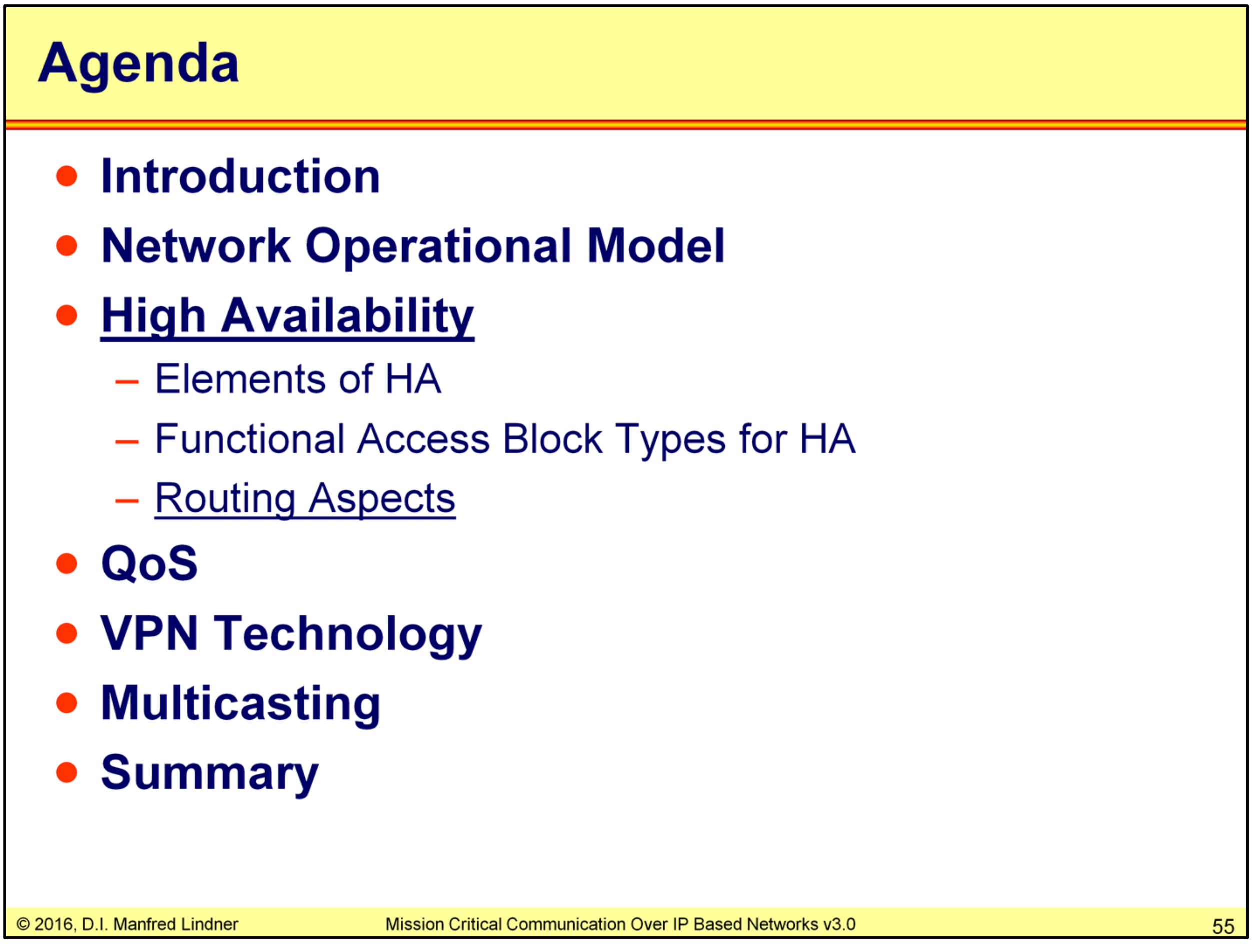
© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 54

Now the IP topology and physical topology are the same. Only number of necessary IP subnets per location has increased to address the point-to-point links.

We have introduced several advanced HA mechanisms (type 3, 4, 5 and 6) to overcome situations where more than a single point of failure happens and network is splitted in separate domains. Often overseen or ignored this can lead to real weird situations with often unpredictable consequences.

Therefore if methods types 1 or 2 are applied they should be analyzed thoroughly and systems designed such as that they get into a stable state even if consistency cannot be guaranteed during all times of separation.

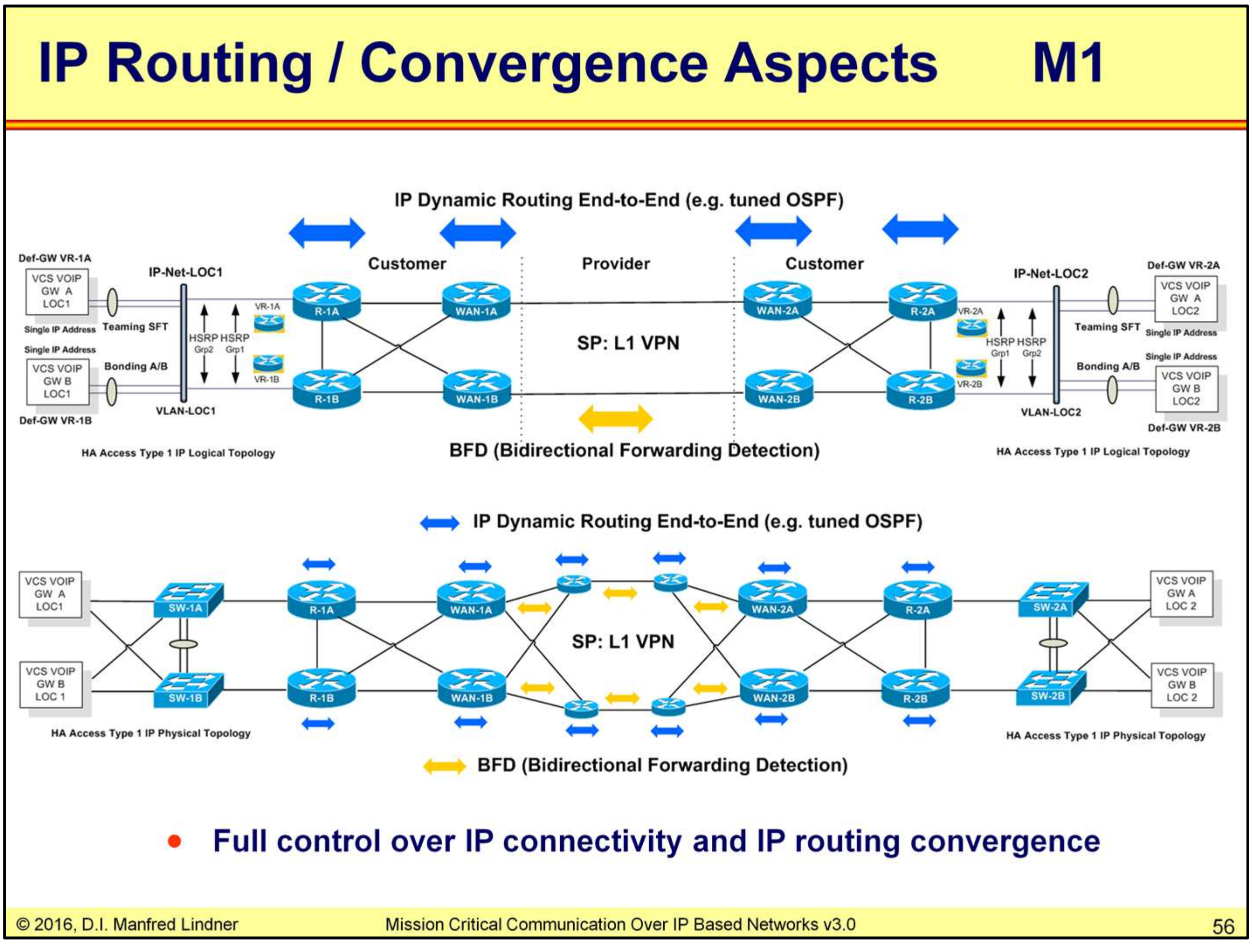
Last but not least effects on failure repair (rerouting back to failure free operation) should be analyzed and tuned to an acceptable level (e.g. postponed to service time interval) because this may lead to connectivity outage for some time, too.



## Agenda

- Introduction
- Network Operational Model
- High Availability
  - Elements of HA
  - Functional Access Block Types for HA
  - Routing Aspects
- QoS
- VPN Technology
- Multicasting
- Summary

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 55



## IP Routing / Convergence Aspects M1

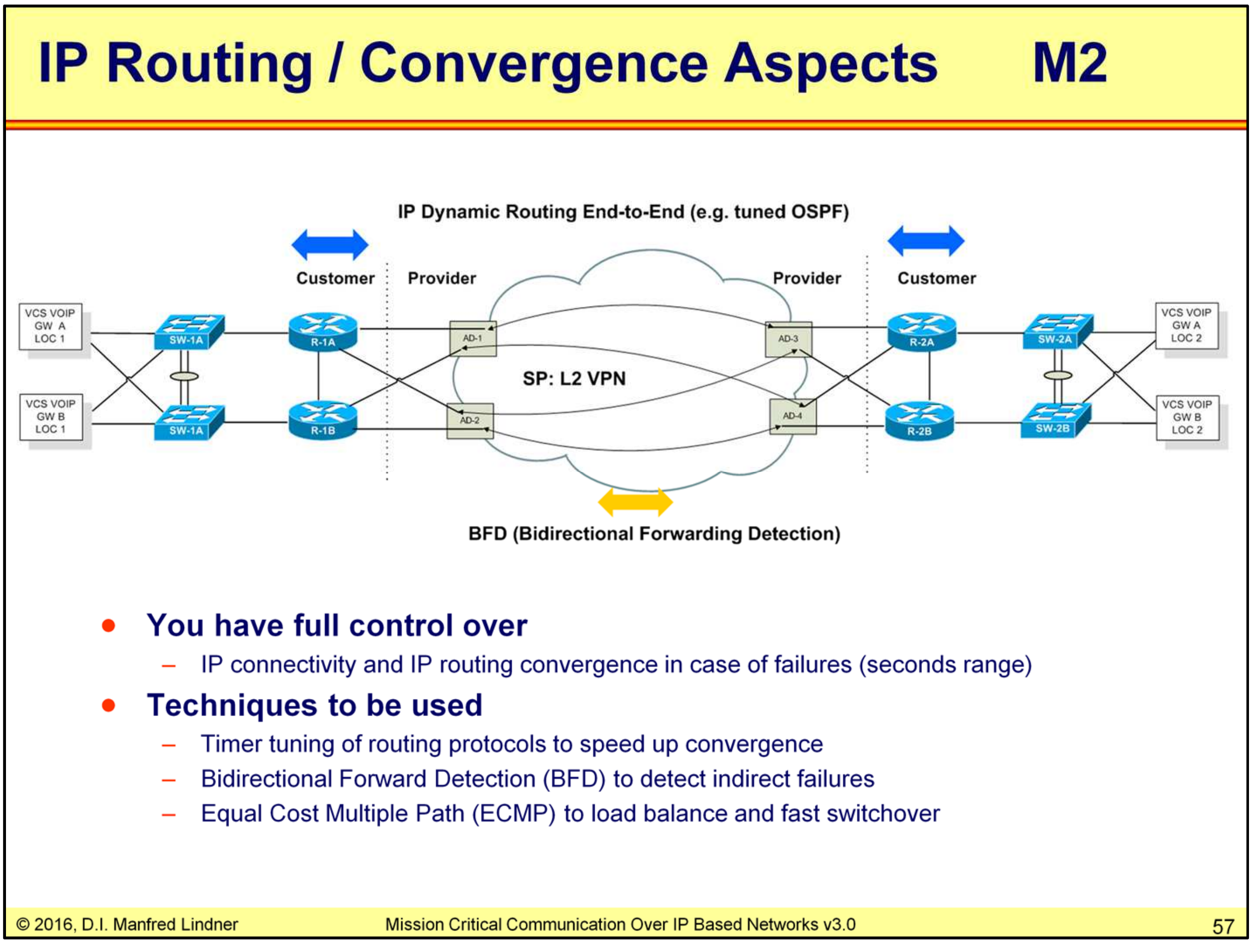
- Full control over IP connectivity and IP routing convergence

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 56

Now after introduction of possible HA mechanism useable in the access part we can deal with the IP WAN concerning routing and routing convergence. Of course the network operational model has an important impact on what can be achieved and what problems may arise and need to been solved. For the following discussions we use just HA access type 1 for reasons of simplicity but the conclusions are valid for all other access types too.

Routing aspects for operational model M1:

In this case all IP routers are under control of the customer. L1 VPN provider just provides L1 links which are used to interconnect the IP routers. Hence dynamic OSPF routing (tuned OSPF to reach convergence with in a second) and BFD (to detect any indirect failures caused by the L1 VPN provider) can be applied straight forward. The first picture just shows a redundant topology between two locations. The second picture expands this principle to an overall redundant topology supporting any amount of locations by usage of partial meshed topologies or ring structures. The separation of the WAN access from location access LAN by usage of two routers (e.g. R-1A and WAN-1A) may be caused by operational aspects (e.g. WAN team to operate the WAN infrastructure and LAN team responsible for security) or by interface aspects of the to be used routers (e.g. R-1A has Ethernet only interfaces and WAN-1A is capable of supporting special serial interfaces (E1/PRI, SDH or microwave/SAT).



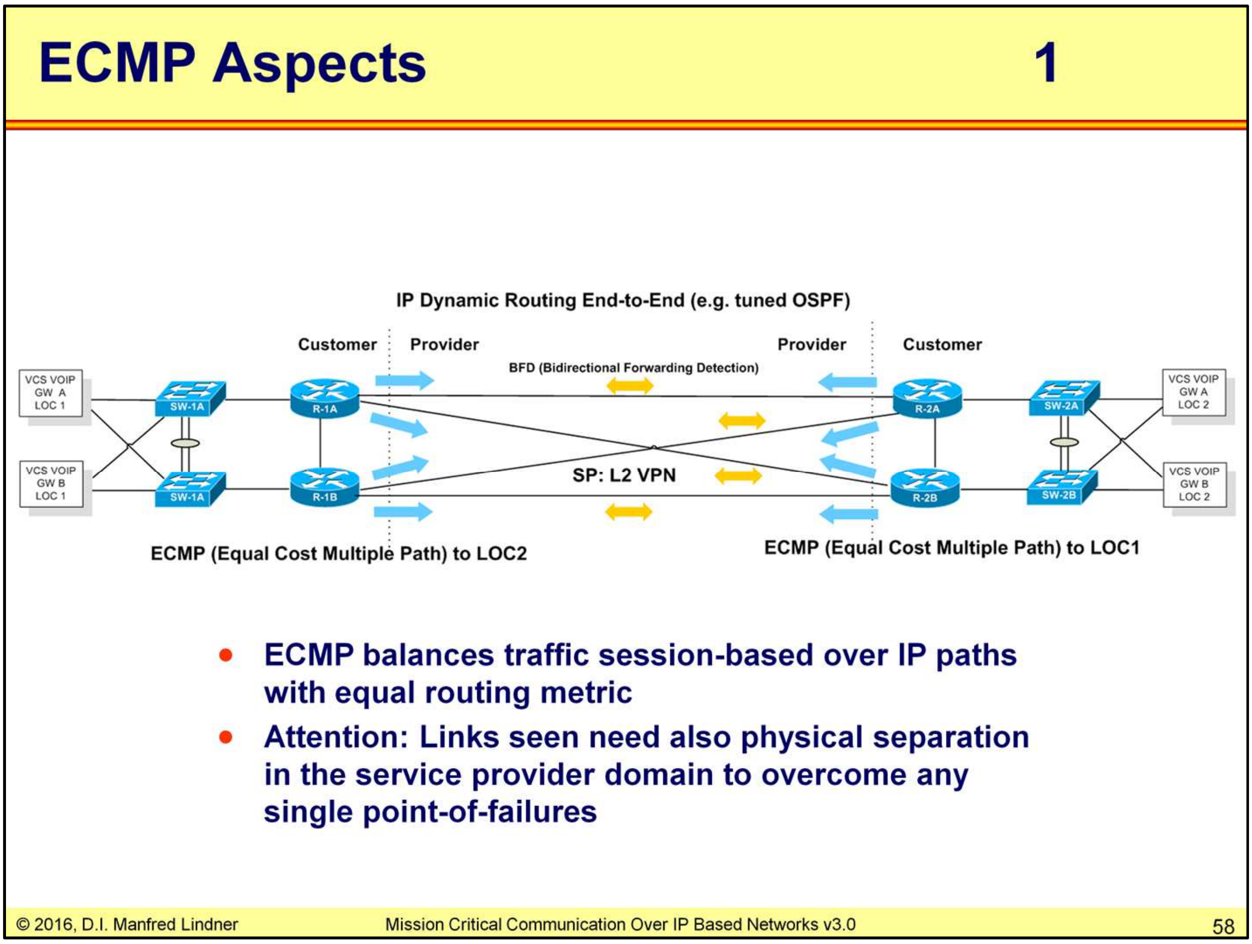
## IP Routing / Convergence Aspects M2

- **You have full control over**
  - IP connectivity and IP routing convergence in case of failures (seconds range)
- **Techniques to be used**
  - Timer tuning of routing protocols to speed up convergence
  - Bidirectional Forward Detection (BFD) to detect indirect failures
  - Equal Cost Multiple Path (ECMP) to load balance and fast switchover

Routing aspects for operational model M2:

In this case again all IP routers are under control of the customer. L2 VPN provider just provides L2 (nowadays Ethernet) links which are used to interconnect the IP routers. Remember that these Ethernet links are just a kind of virtual links because L2-VPN providers put their service on top of a packet-switching infrastructure (e.g. IP-MPLS) shared among different customers, hence statistics concerning delay and throughput will be experienced on such links. Again dynamic OSPF routing (tuned OSPF to reach convergence with in a second) and BFD (to detect any indirect failures caused by the L2 VPN provider) can be applied straight forward.

The advantage of such a model is that a customer can build any point-to-point topology between locations by using just L2 links of the provider. The picture shows a redundant topology between two locations.



## ECMP Aspects 1

IP Dynamic Routing End-to-End (e.g. tuned OSPF)
Customer
Provider
Provider
Customer
BFD (Bidirectional Forwarding Detection)
VCS VOIP GW A LOC 1
VCS VOIP GW B LOC 1
SW-1A
SW-1A
R-1A
R-1B
SP: L2 VPN
R-2A
R-2B
SW-2A
SW-2B
VCS VOIP GW A LOC 2
VCS VOIP GW B LOC 2
ECMP (Equal Cost Multiple Path) to LOC2
ECMP (Equal Cost Multiple Path) to LOC1

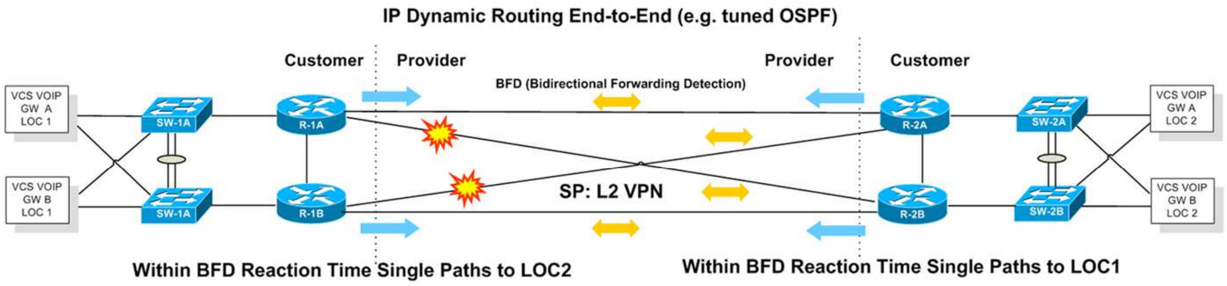
- ECMP balances traffic session-based over IP paths with equal routing metric
- Attention: Links seen need also physical separation in the service provider domain to overcome any single point-of-failures

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 58

The following picture shows usage of ECMP (Equal Cost Multiple Paths) which provides the fastest convergence time which can be achieved with IP network technology. With ECMP traffic is load balanced session/flow-based across all available paths showing the same routing metric. For example R-1A will sent traffic of one session to R-2A and traffic of another session to R-2B. In Cisco routers this is done by hardware friendly CEF switching per default. Even if the two Ethernet links are coming from different service providers using different provider-internal base technologies from the point of IP routing there are equal. By forcing a session to one of the links the session will experience only the statistics of one provider over time which is in any case better than load-balancing on an individual packet base (without any knowledge of sessions).

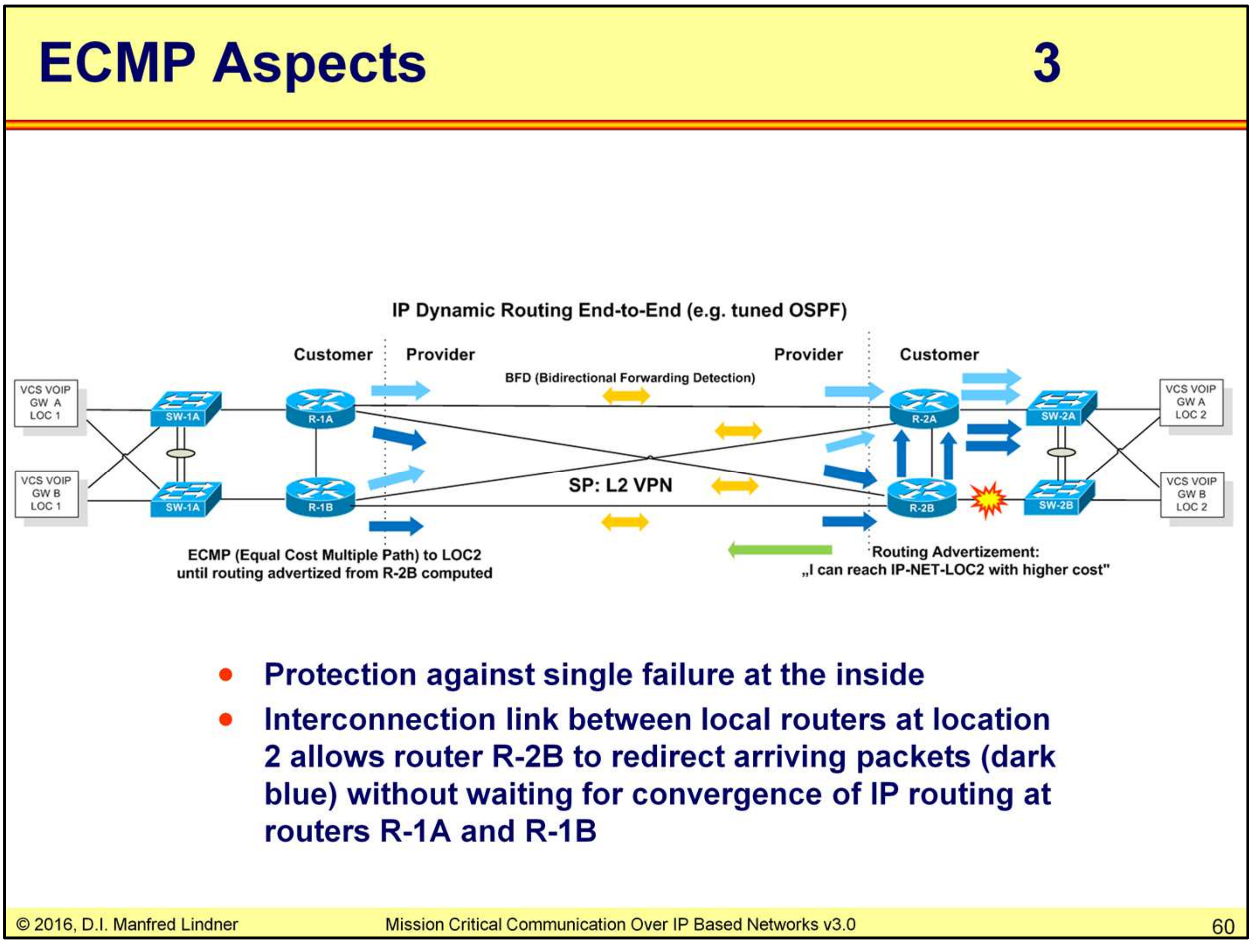
Attention: Now we see just point-to-point virtual links between our IP routers on the last picture pretending to have four separated links between the two locations. It is always a challenge to clarify or to agree with the L2 VPN provider that these four links should be using separated physical paths and separated provider internal components (AD1, AD2, AD3, AD4 -> e.g. Add/Drop SDH multiplexers).

## ECMP Aspects 2


IP Dynamic Routing End-to-End (e.g. tuned OSPF)
Customer
Provider
Provider
Customer
BFD (Bidirectional Forwarding Detection)
VCS VOIP GW A LOC 1
VCS VOIP GW B LOC 1
SW-1A
SW-1A
R-1A
R-1B
SP: L2 VPN
R-2A
R-2B
SW-2A
SW-2B
VCS VOIP GW A LOC 2
VCS VOIP GW B LOC 2
Within BFD Reaction Time Single Paths to LOC2
Within BFD Reaction Time Single Paths to LOC1

- Convergence time depends only on BFD timeout
- Fastest way to direct traffic to remaining links
- Routing updates will inform routers about new topology but not necessary for rerouting

In case of failure on a WAN Ethernet link the following will occur: Packet in transmit will of course be destroyed, but just at the moment when the routers recognize the link failure (directly by disappearance of physics or indirectly by timeout of BFD) the next and all following packets will be forced to use the other path without any need to wait for convergence time of an IP routing protocol. The picture shows a double failure which forces the traffic to flow on the remaining links. Of course the routing protocol will produce an update sent to all other routers to inform about break of links, but the affected routers need not to wait for such routing updates in order to react (reroute the traffic).



# ECMP Aspects 3

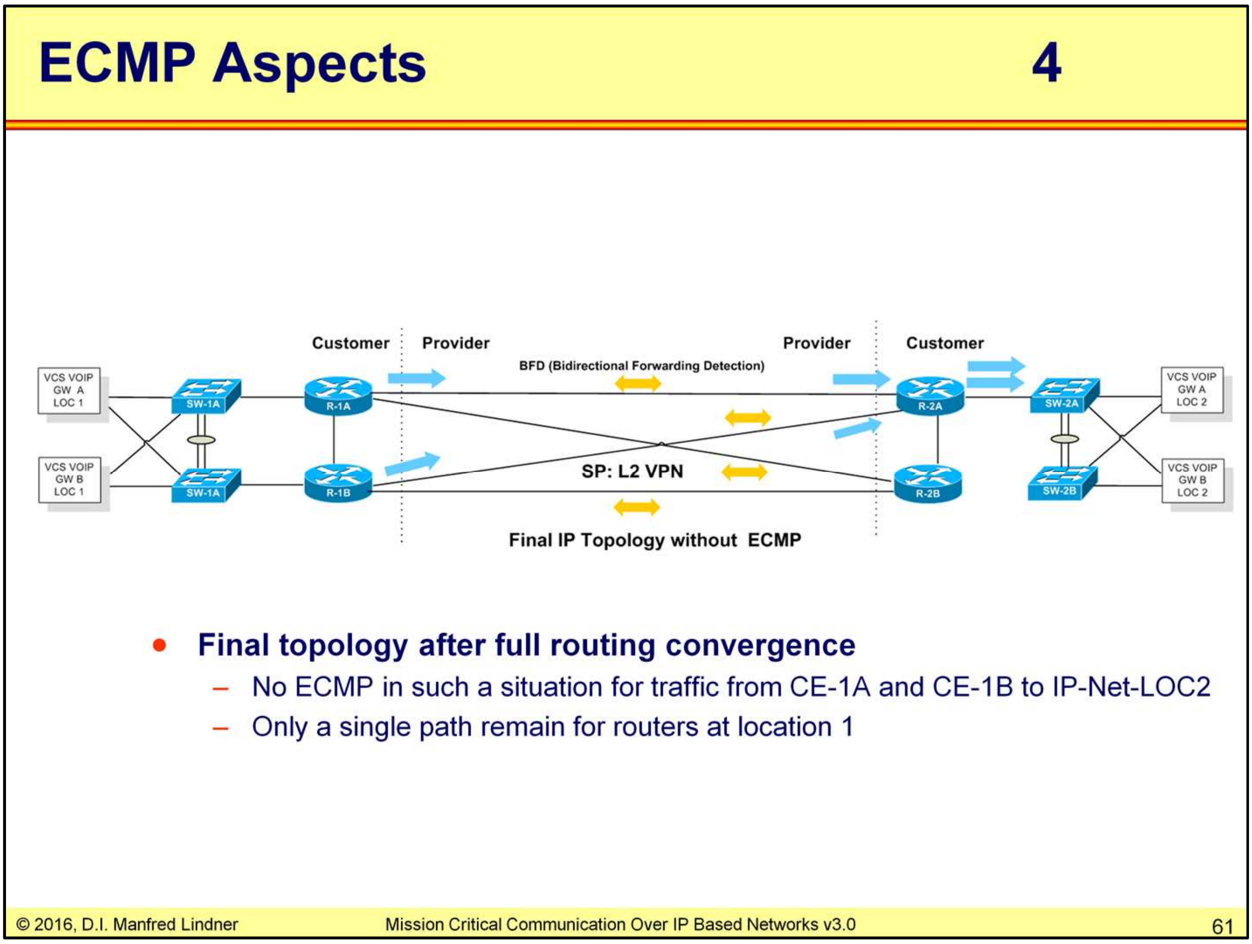
IP Dynamic Routing End-to-End (e.g. tuned OSPF)
Customer
Provider
Provider
Customer
BFD (Bidirectional Forwarding Detection)
VCS VOIP GW A LOC 1
VCS VOIP GW B LOC 1
SW-1A
SW-1A
R-1A
R-1B
SP: L2 VPN
R-2A
R-2B
SW-2A
SW-2B
VCS VOIP GW A LOC 2
VCS VOIP GW B LOC 2
ECMP (Equal Cost Multiple Path) to LOC2
until routing advertized from R-2B computed
Routing Advertizement:
„I can reach IP-NET-LOC2 with higher cost"

- Protection against single failure at the inside
- Interconnection link between local routers at location 2 allows router R-2B to redirect arriving packets (dark blue) without waiting for convergence of IP routing at routers R-1A and R-1B

© 2016, D.I. Manfred Lindner    Mission Critical Communication Over IP Based Networks v3.0    60

Finally this picture shows one reason for interconnection of routers on one location.

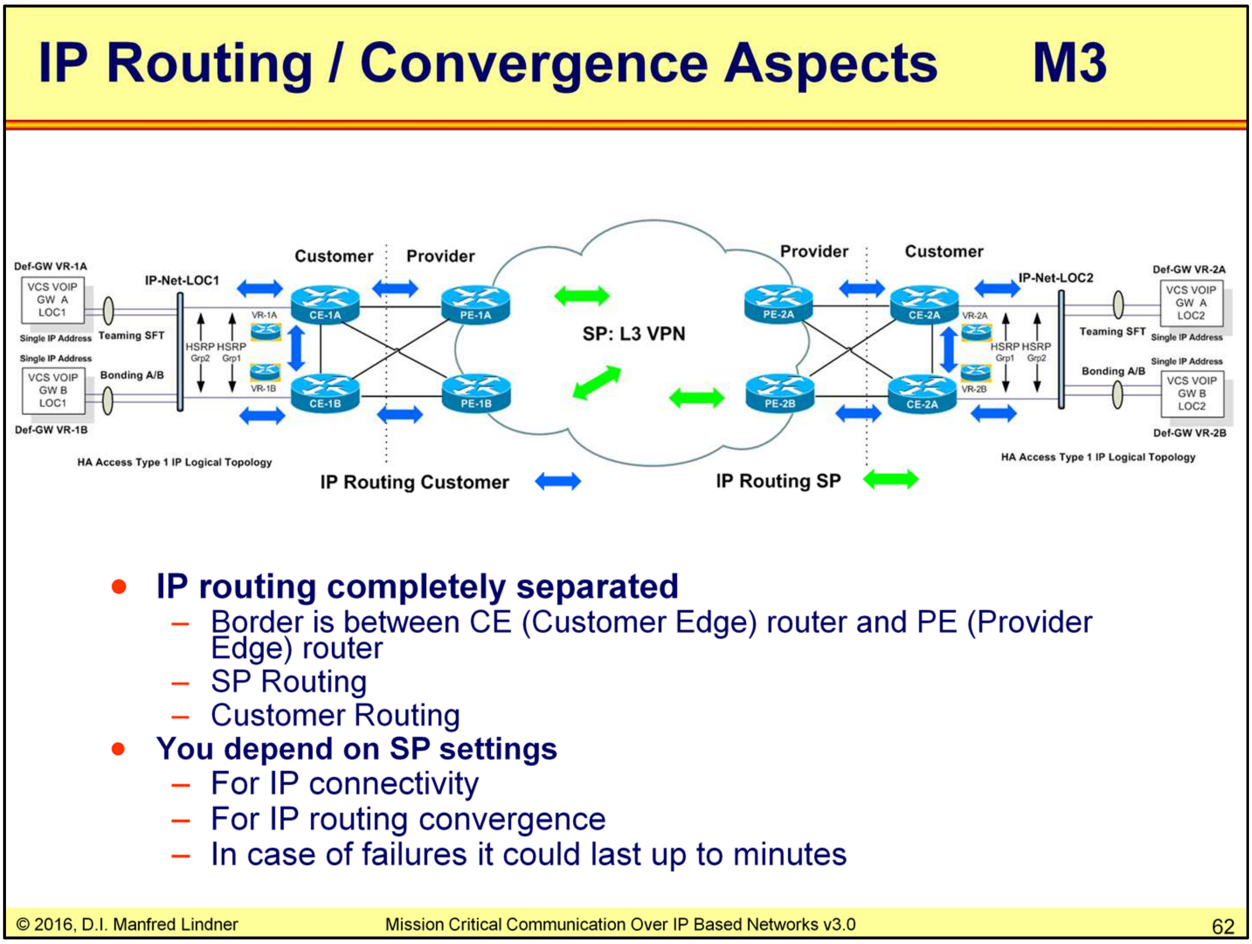
Assume R-2B loses the Ethernet connection to the systems of location 2. For convergence of outgoing traffic HSRP will do the job. For redirecting incoming traffic to router R-2A a routing update to the other site will force all traffic to router R-2A because we have no equal cost paths anymore. But this takes some time. Still packets for VCSA/B-LOC2 will arrive at R-2B and of course R-2B will forward them towards the alternate router R-2A using the interconnection link (which is a local fast decision based on the physical failure). So again only the packet in transmit across the erroneous link will be destroyed. All the next packets can flow without any disturbance. Traffic successfully transmitted over the WAN needs not to be discarded because one link at the location goes down. The interconnection link keeps the failure concerning convergence time local and sessions will experience minimal disturbance only for the time to discover the failure which is not avoidable per se in any case.



# ECMP Aspects 4

Final IP Topology without ECMP

- **Final topology after full routing convergence**
  - No ECMP in such a situation for traffic from CE-1A and CE-1B to IP-Net-LOC2
  - Only a single path remain for routers at location 1



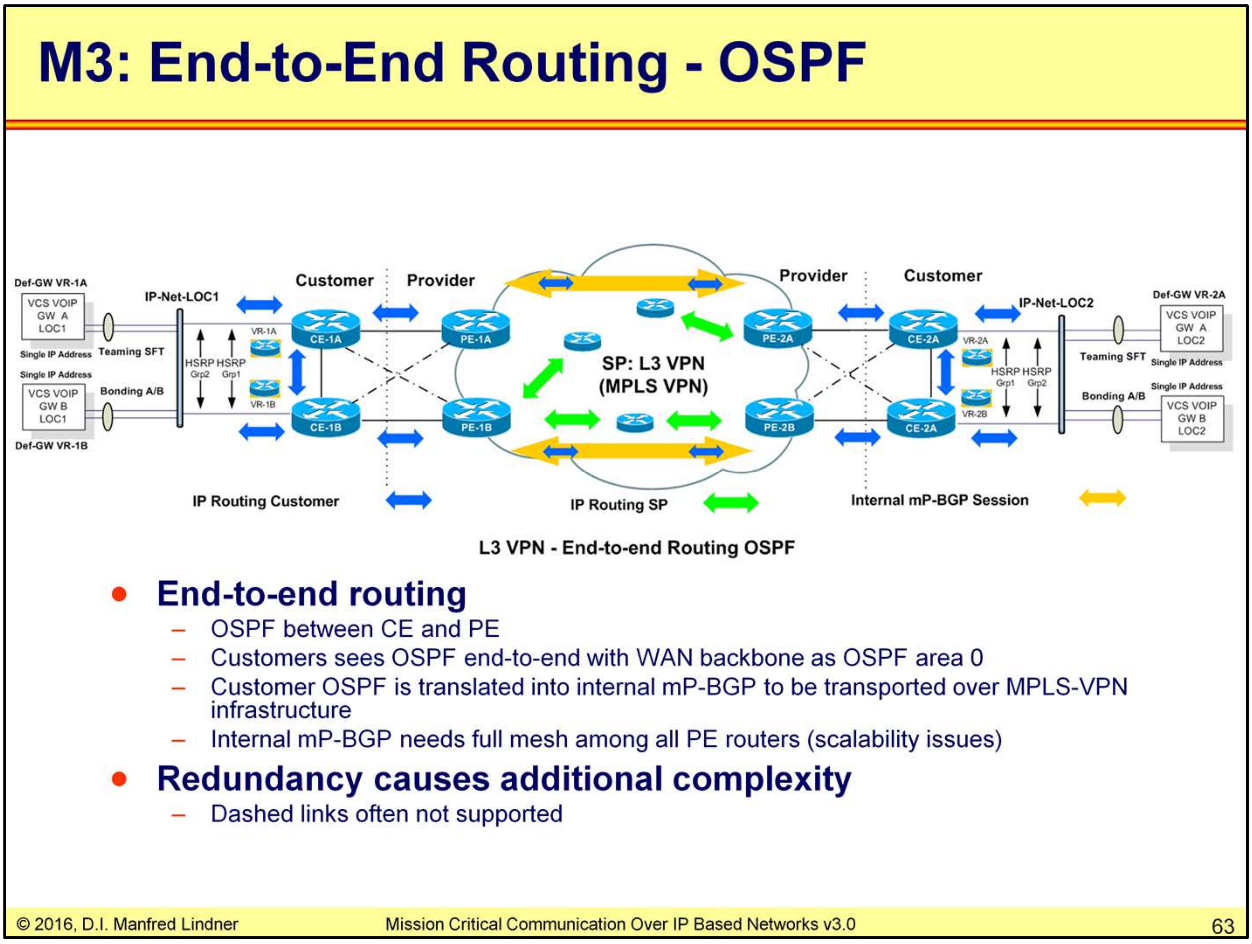
# IP Routing / Convergence Aspects M3

- **IP routing completely separated**
  - Border is between CE (Customer Edge) router and PE (Provider Edge) router
  - SP Routing
  - Customer Routing
- **You depend on SP settings**
  - For IP connectivity
  - For IP routing convergence
  - In case of failures it could last up to minutes

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 62

Routing aspects for operational model M3:

In this case IP routing within the service provider is completely separated from the IP routing in the customer domain. The separation can be seen by differentiation of CE (customer edge) routers and PE (provider edge) routers. That clarifies who is responsible for operation of the router devices.



# M3: End-to-End Routing - OSPF

L3 VPN - End-to-end Routing OSPF

- **End-to-end routing**
  - OSPF between CE and PE
  - Customers sees OSPF end-to-end with WAN backbone as OSPF area 0
  - Customer OSPF is translated into internal mP-BGP to be transported over MPLS-VPN infrastructure
  - Internal mP-BGP needs full mesh among all PE routers (scalability issues)
- **Redundancy causes additional complexity**
  - Dashed links often not supported

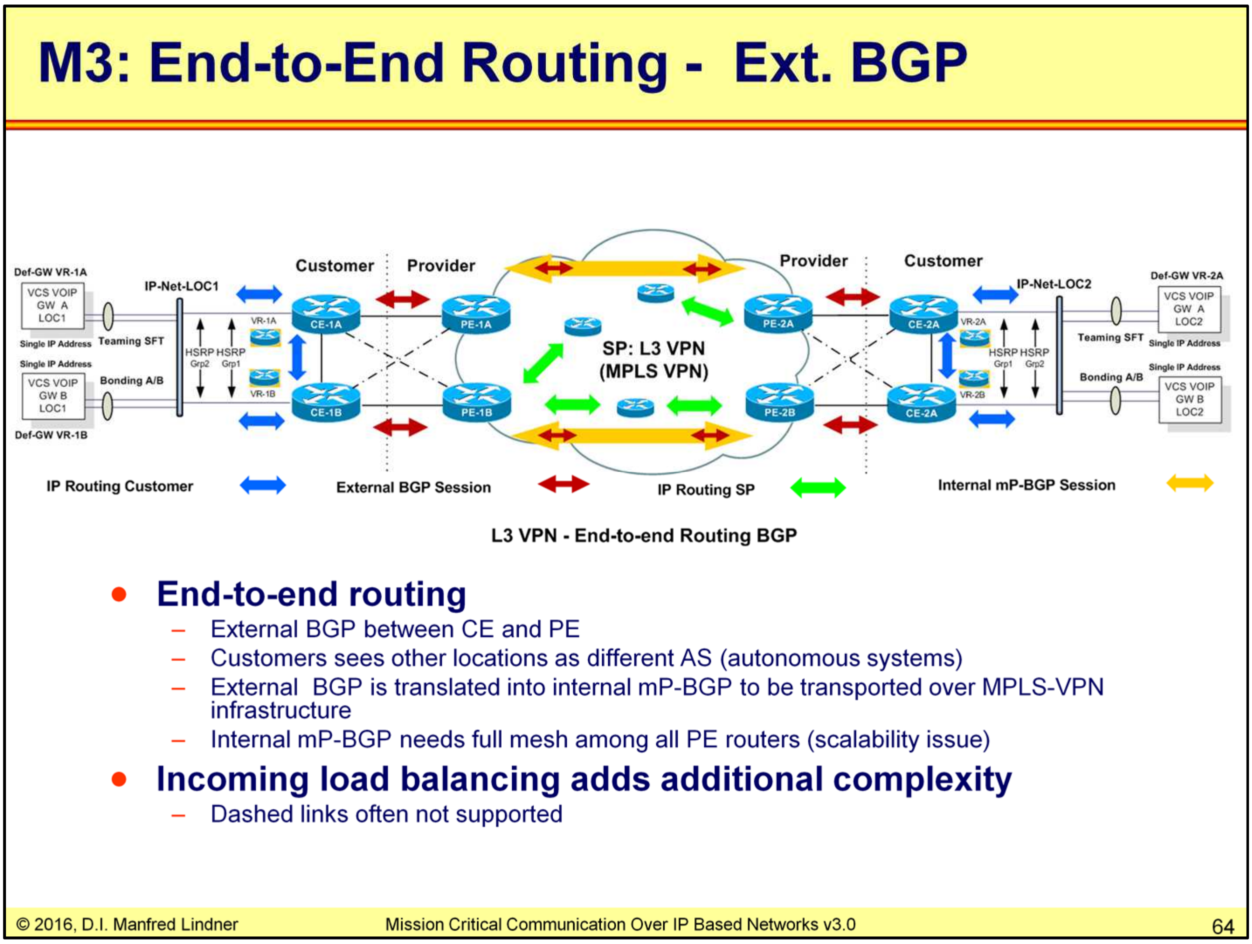
© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 63

One of the basic questions to be answered is, if customer needs end-to-end IP routing in order to select different paths/providers or if customer can use default routing to all remote locations.

End-to-end IP routing can be achieved by L3-VPN service providers by usage of MPLS-VPN technology. For example IP OSPF customer routing can be transported end-to-end over provider internal mP-BGP sessions of MPLS-VPN technology.

In case of redundancy that is a real challenge for the service provider. So the cross-links between CE and PE routers are drawn as dashed lines or even omitted in the following pictures to indicate that MPLS VPN providers usually do not support such topologies.

A single link between a CE and PE router is the preferred way from the viewpoint of the provider. You can see that such a redundancy makes sense from the viewpoint of physical topology but it will not supported by the providers with their routing techniques. So how many links from CE to PE are really useful depends on the operational model you choose.



# M3: End-to-End Routing - Ext. BGP

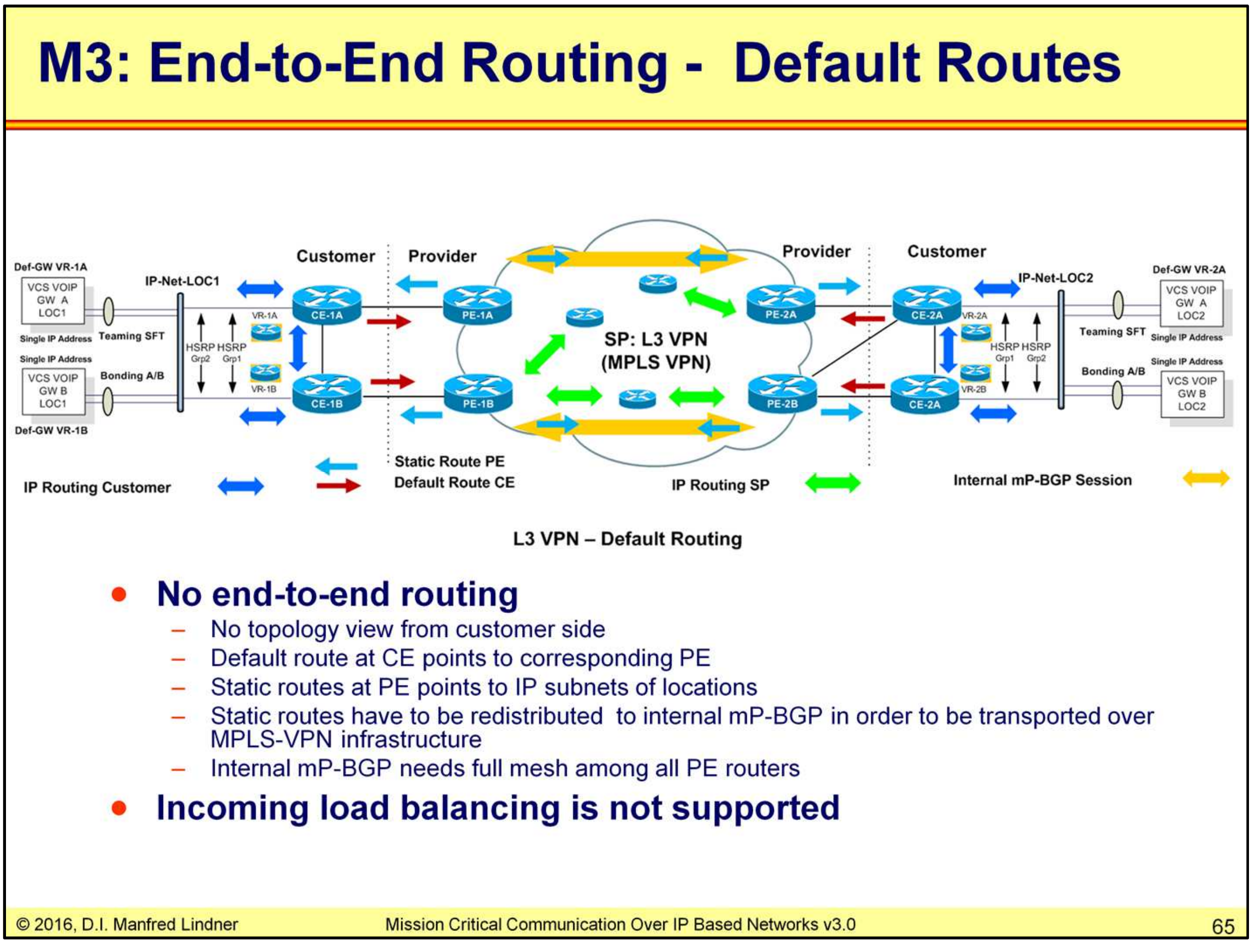
L3 VPN - End-to-end Routing BGP

- **End-to-end routing**
  - External BGP between CE and PE
  - Customers sees other locations as different AS (autonomous systems)
  - External BGP is translated into internal mP-BGP to be transported over MPLS-VPN infrastructure
  - Internal mP-BGP needs full mesh among all PE routers (scalability issue)
- **Incoming load balancing adds additional complexity**
  - Dashed links often not supported

© 2016, D.I. Manfred Lindner    Mission Critical Communication Over IP Based Networks v3.0    64

Another approach for end-to-end IP customer routing is to terminate OSPF at the CE router of a location and interact with the PE router via external BGP.

You need to design end-to-end routing, configure all the corresponding routing protocols, try to tune them for low convergence time and test it for all failure scenarios.



## M3: End-to-End Routing - Default Routes

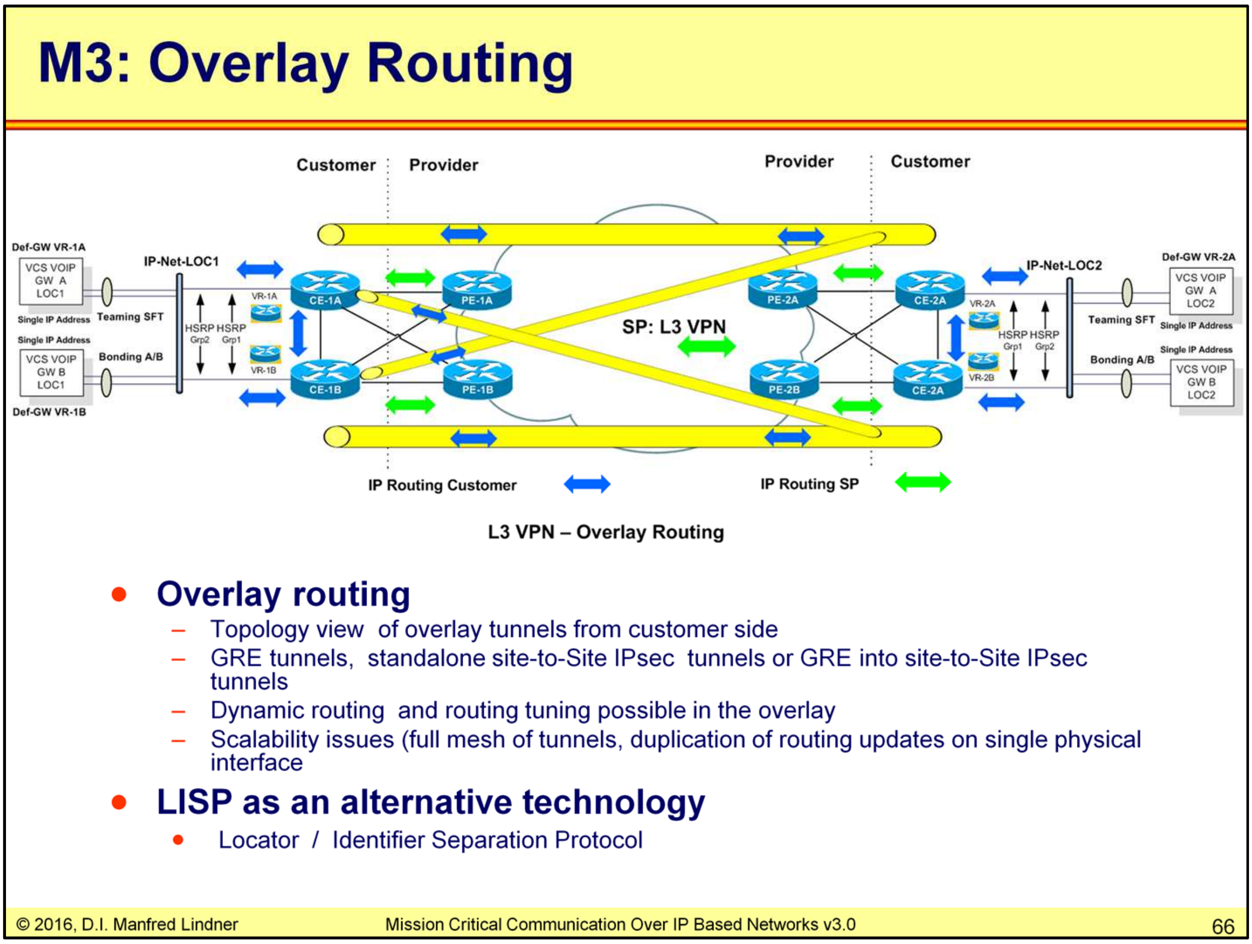
L3 VPN – Default Routing

- **No end-to-end routing**
  - No topology view from customer side
  - Default route at CE points to corresponding PE
  - Static routes at PE points to IP subnets of locations
  - Static routes have to be redistributed to internal mP-BGP in order to be transported over MPLS-VPN infrastructure
  - Internal mP-BGP needs full mesh among all PE routers
- **Incoming load balancing is not supported**

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 65

Default routing at the CE routers together with static routes configured at the PE routers is much simpler, but because of the static nature not so flexible in case changes have to be implemented fast.

In any case of MPLS-VPN IP addressing and IP routing has to be harmonized between the customer and the service provider. Also SLAs have to be defined on the borders to settle down responsibilities and procedures for interaction of involved parties.



## M3: Overlay Routing

L3 VPN – Overlay Routing

- **Overlay routing**
  – Topology view of overlay tunnels from customer side
  – GRE tunnels, standalone site-to-Site IPsec tunnels or GRE into site-to-Site IPsec tunnels
  – Dynamic routing and routing tuning possible in the overlay
  – Scalability issues (full mesh of tunnels, duplication of routing updates on single physical interface
- **LISP as an alternative technology**
  - Locator / Identifier Separation Protocol

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 66

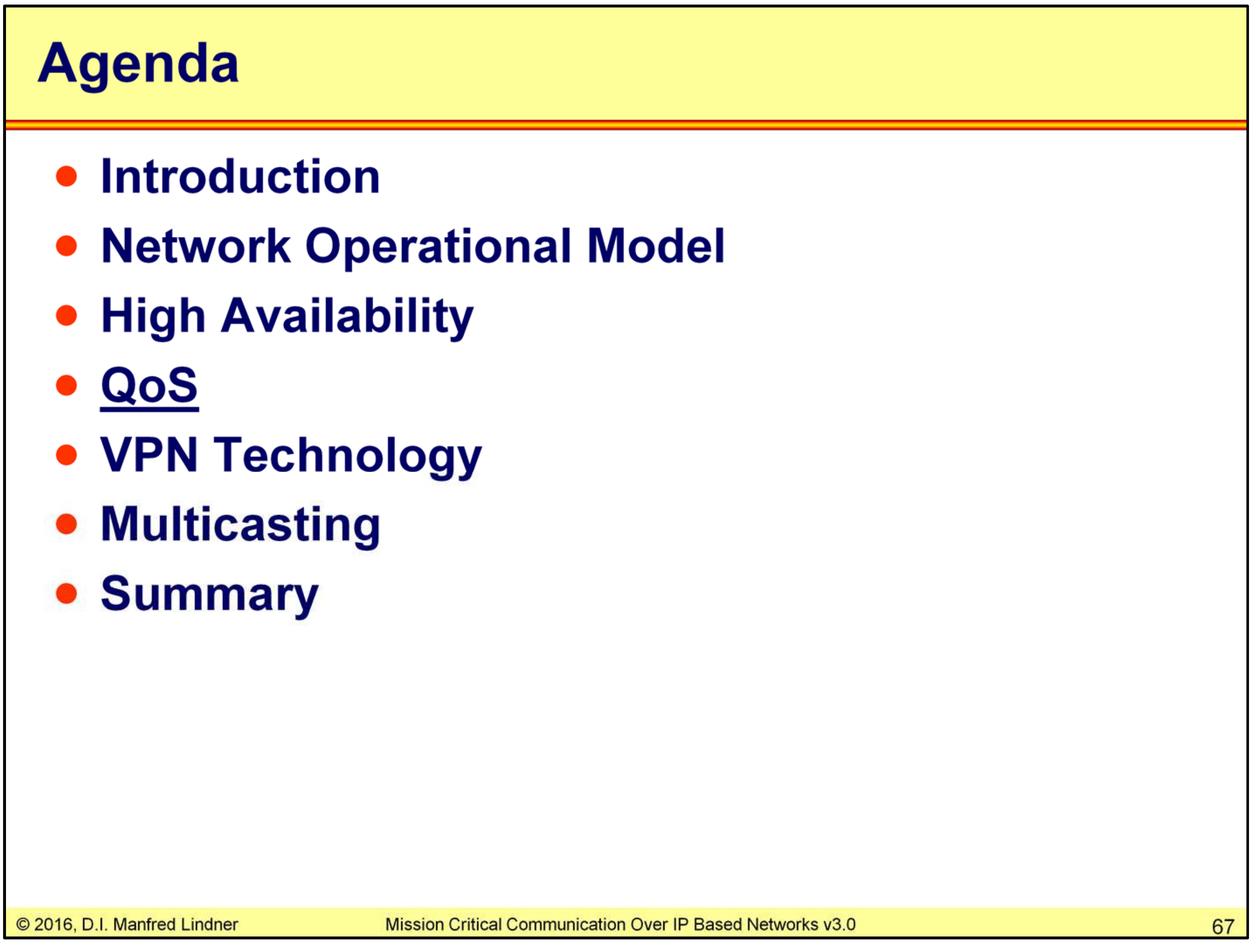
Another approach when dealing with L3 VPN service providers or even with the Internet as IP WAN is to implement an overlay network starting at the CE routers on top of any given L3 connectivity offered by service providers or ISPs. Now the challenge is at the customer side to implement IP connectivity and tuning routing convergence across the overlay.

The benefit for the customer is to get control over these topics back from the provider. Examples for such overlay networks are GRE, LISP as separation techniques controlled by the customer or site-to-site VPNs based on IPsec or SSL if security issues for integrity and privacy are involved. In all these cases the provider acts just as simple IP connectivity enabler transporting encapsulated IP messages between PEs in the provider controlled IP address space. Hence no interference with IP addresses of customers or complex NAT solutions are necessary.

Attention: There are issues concerning scalability with L3 VPN based on MPLS-VPN or legacy overlay techniques (site-to-site IPsec VPN, SSL VPN). MPLS-VPN suffers from internal mP-BGP (fully meshed or usage of route reflectors) and number of customer injected into MPLS-VPN for customer end-to-end routing. Site-to-site IPsec/SSL VPN suffers from number of point-to-point security associations (configuration and performance) and from duplication of routing broadcast (hello, updates) into each logical tunnel to which the CE router is connected (performance and congestion point on the physical interface to the PE router).

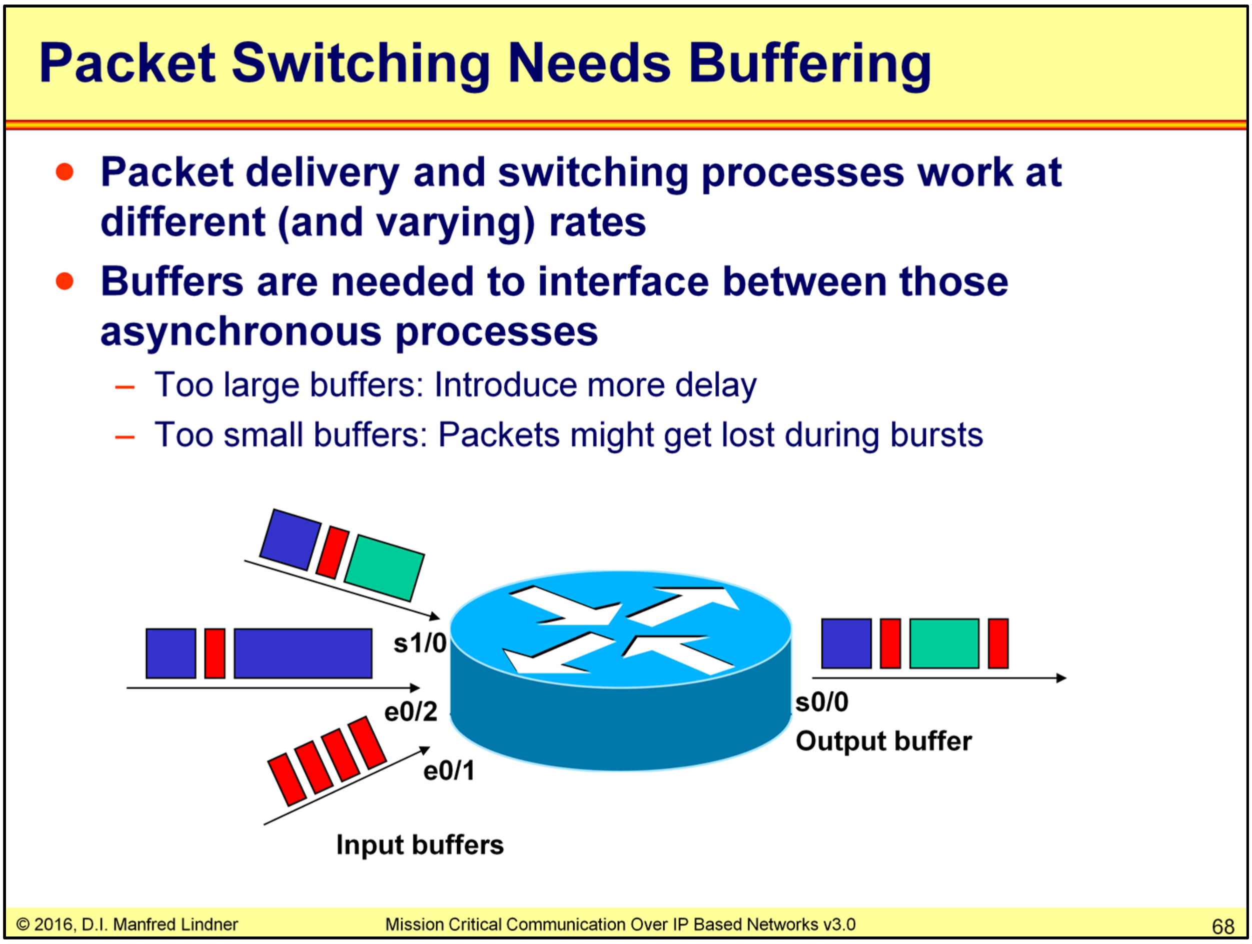
LISP (Locator / Identifier Separation Protocol) - as possible future provider independent network technology - seems to be interesting candidate for VPN networks. LISP does not suffer from scalability issues because of a tunnel-free, configuration friendly and self-deploying approach to build the overlay. LISP enables customer to build a basic VPN over any service provider infrastructure, which separates the customer traffic from all other traffic in the service provider in the same way as MPLS-VPN does. LISP together with GETVPN (Group Encrypted Transport VPN) enhance this basic VPN with security (integrity, privacy) maintained by GETVPN. GETVPN is IPsec based on group keys distributed by key servers, needs no point-to-point security associations and hence scales much better than traditional site-to-site IPsec VPN.

If you want to get more information on LISP consult the correspondent lecture part.



## Agenda

- Introduction
- Network Operational Model
- High Availability
- QoS
- VPN Technology
- Multicasting
- Summary



## Packet Switching Needs Buffering

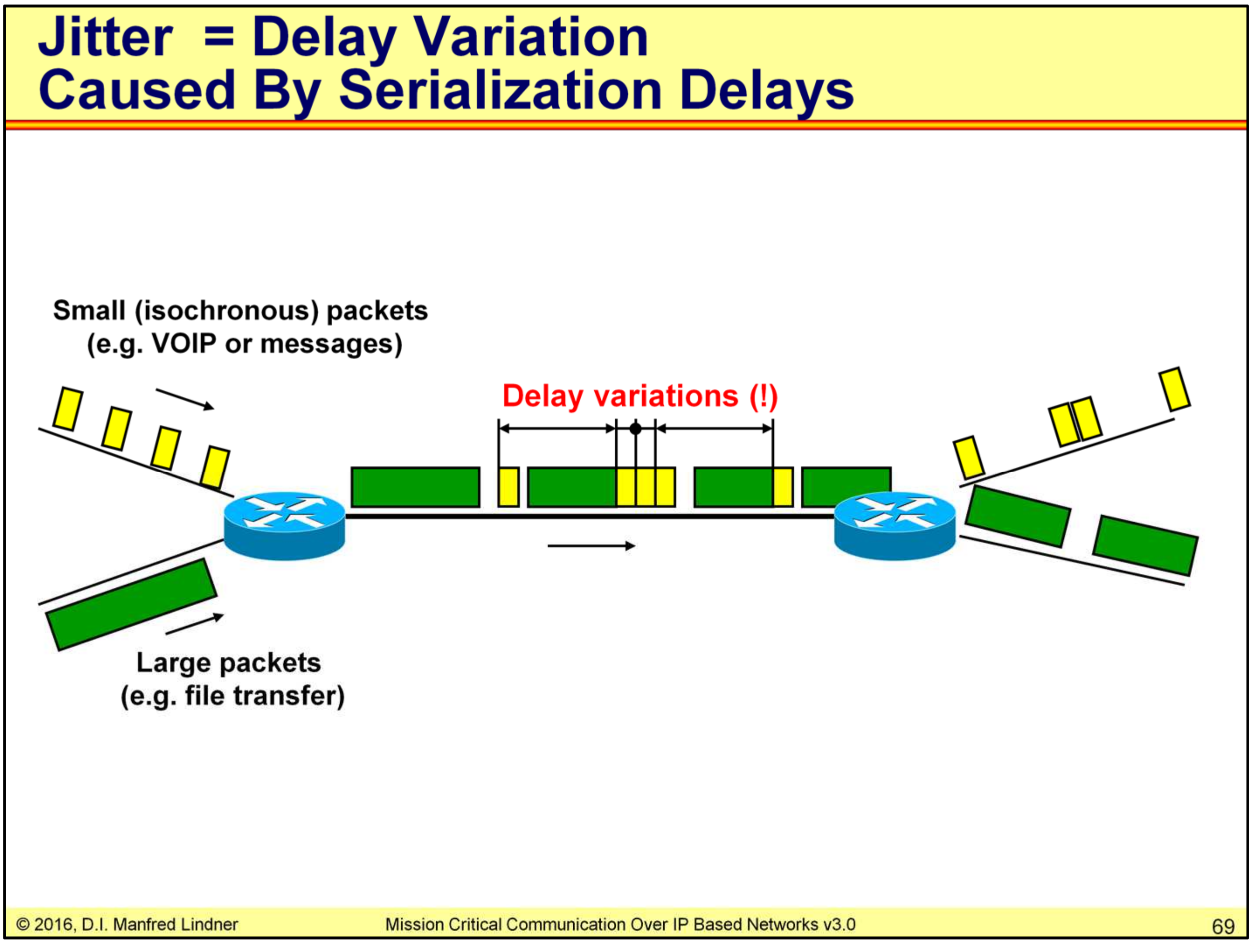
- **Packet delivery and switching processes work at different (and varying) rates**
- **Buffers are needed to interface between those asynchronous processes**
  - Too large buffers: Introduce more delay
  - Too small buffers: Packets might get lost during bursts

s1/0
e0/2
e0/1
Input buffers
s0/0
Output buffer

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 68

It is a common misunderstanding to assume that queues are always filled at a certain percentage. If this would be the case then traffic bursts would occur too frequent and the probability of congestion is very high.

Note that queues should only smooth bursts! On average the traffic rate should be easily handled by the network devices (packet switches in general) without filling the queues considerably.



## Jitter = Delay Variation Caused By Serialization Delays

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 69

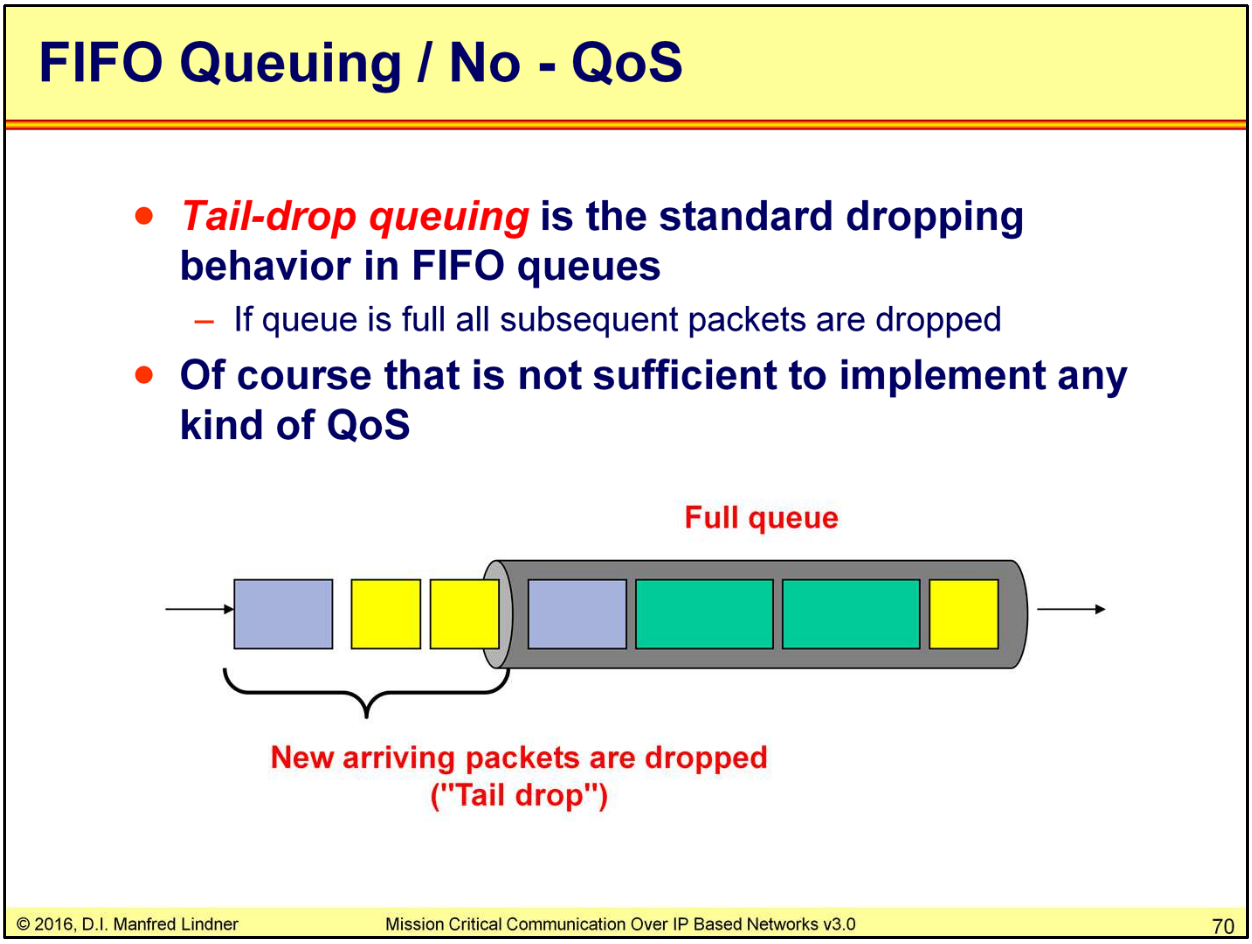
Delay in networks:

Serialization delay is the time which is necessary to put a block of bits on a serial line with a given bitrate.

Propagation delay is the time which is needed for a electrical signal to propagate along a given length of line / transmission path. The upper limit for velocity is of course the speed of light.

Switching delay additionally occurs in case of the presence of an active component in the transmission path.

If small packets can be forwarded without waiting for the link to be ready then delay variation will be small. If small packets cannot immediately be forwarded because link to be used is occupied by another ongoing transmission (especially if long packets are on the way) then delay variation will be large. Also you can see in the picture that isochronous traffic can lead to packet burst which happens if for a certain time a sequence of such small isochronous packets was buffered and then send out one packet after the other without any gap.



## FIFO Queuing / No - QoS

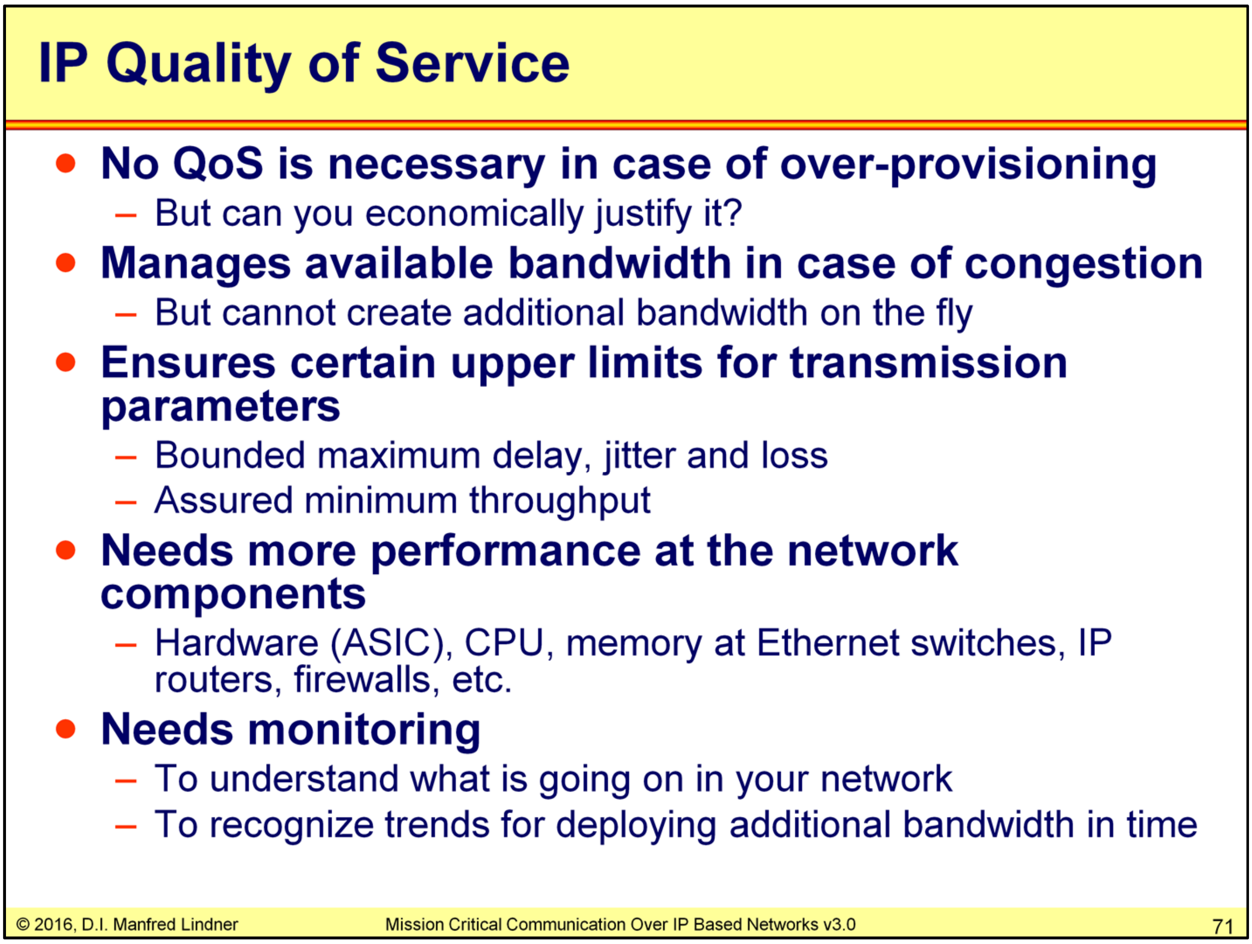
- ***Tail-drop queuing*** **is the standard dropping behavior in FIFO queues**
  - If queue is full all subsequent packets are dropped
- **Of course that is not sufficient to implement any kind of QoS**

**Full queue**

**New arriving packets are dropped ("Tail drop")**

© 2016, D.I. Manfred Lindner  Mission Critical Communication Over IP Based Networks v3.0  70

“First In First Out “ is the default behavior of all packet switching devices, Does a good job, if network is not saturated and is easy to be implemented.



## IP Quality of Service

- No QoS is necessary in case of over-provisioning
  - But can you economically justify it?
- Manages available bandwidth in case of congestion
  - But cannot create additional bandwidth on the fly
- Ensures certain upper limits for transmission parameters
  - Bounded maximum delay, jitter and loss
  - Assured minimum throughput
- Needs more performance at the network components
  - Hardware (ASIC), CPU, memory at Ethernet switches, IP routers, firewalls, etc.
- Needs monitoring
  - To understand what is going on in your network
  - To recognize trends for deploying additional bandwidth in time

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 71

Synchronous networks (also called "isochronous" networks) have no problem with quality of service: Each piece of data is transmitted in equal times and there are no "**bursts**" of data that would lead to sudden congestion. Real synchronous clocking of all network components means no need of any buffers.

The Internet on the other hand is an extremely inhomogeneous and asynchronous kind of network. It is a network of networks and the prevalent type of traffic is data traffic. Data traffic is bursty by nature.

What is meant by "Quality of Service" (QoS) actually? The total network bandwidth is shared by many users and applications. We demand for QoS in that some specific type of traffic should be handled "nicer" than others. When buffers in routers and switches are congested then delays and packet-drops occur. One solution is to over-provision the network, but this is expensive and the behavior is still unpredictable.

Thus QoS mechanisms are thought to **manage** the network resources in terms of bandwidth and queuing behavior, so delays can be limited.

IP QoS mechanisms ensure that the most important IP datagrams will get their sufficient traffic characteristic behavior (e.g. bounded maximum delay but still variable, bounded maximum jitter, zero loss or loss below certain percentage, assured minimum throughput) during times of congestions. Of course quality of service is determined by the weakest link between each two points of communications.

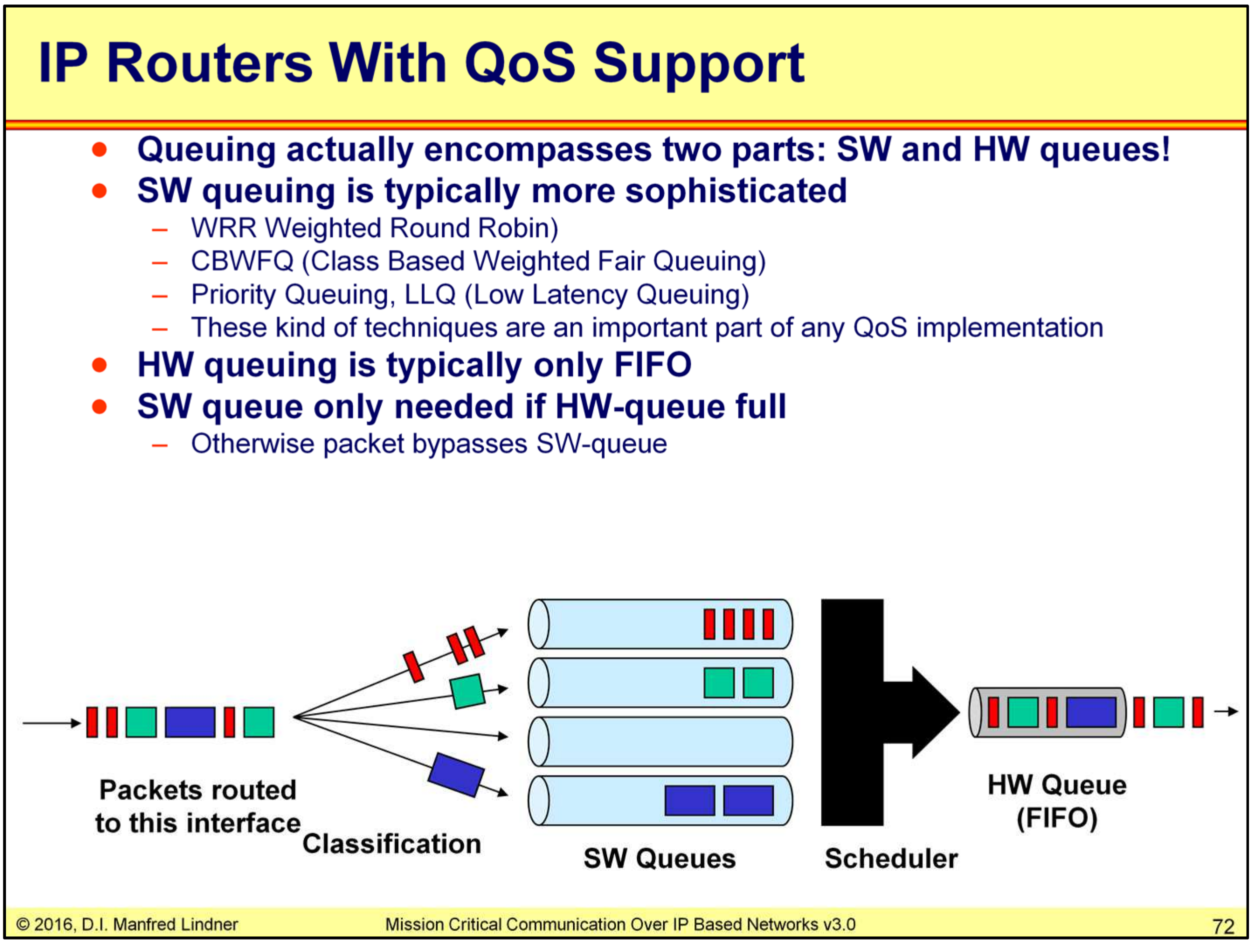
Important to know is that QoS cannot create additional bandwidth. You are still limited to the given physical bit-rate / bandwidth on the weakest network link between source and destination of traffic. QoS mechanisms can only manage the available bandwidth according to a QoS policy during time of congestion.

The negative side of these mechanisms is that less important traffic will suffer (= higher delay) or even will be killed (= loss because of no buffer resources are available). Hence it can be conducted that QoS can overcome only short-time temporary congestion but no long-time congestion. During long-time-congestion less important traffic will get into starvation so either this traffic has no real importance to the overall system or the bandwidth has to be increased anyway to keep such traffic flowing again.

QoS mechanisms need to be supported by the network components, hence such QoS enabled components need more performance (CPU) and in case of higher speed links hardware support (ASICS) especially for queuing. Software queuing may be sufficient up to 2-10 Mbps speed. Hardware queuing is a must for Fast Ethernet (100 Mbps) and higher.

QoS monitoring helps verifying the communication matrix and real behavior of distributed communication. It also allows base-lining and forecasts. Hence network performance can be increased by adding more bandwidth to the corresponding links in time in a proactive way.

Golden rule is to have more bandwidth as needed in the network (over-provisioning) and usage of the QoS system just to overcome temporary traffic bursts.



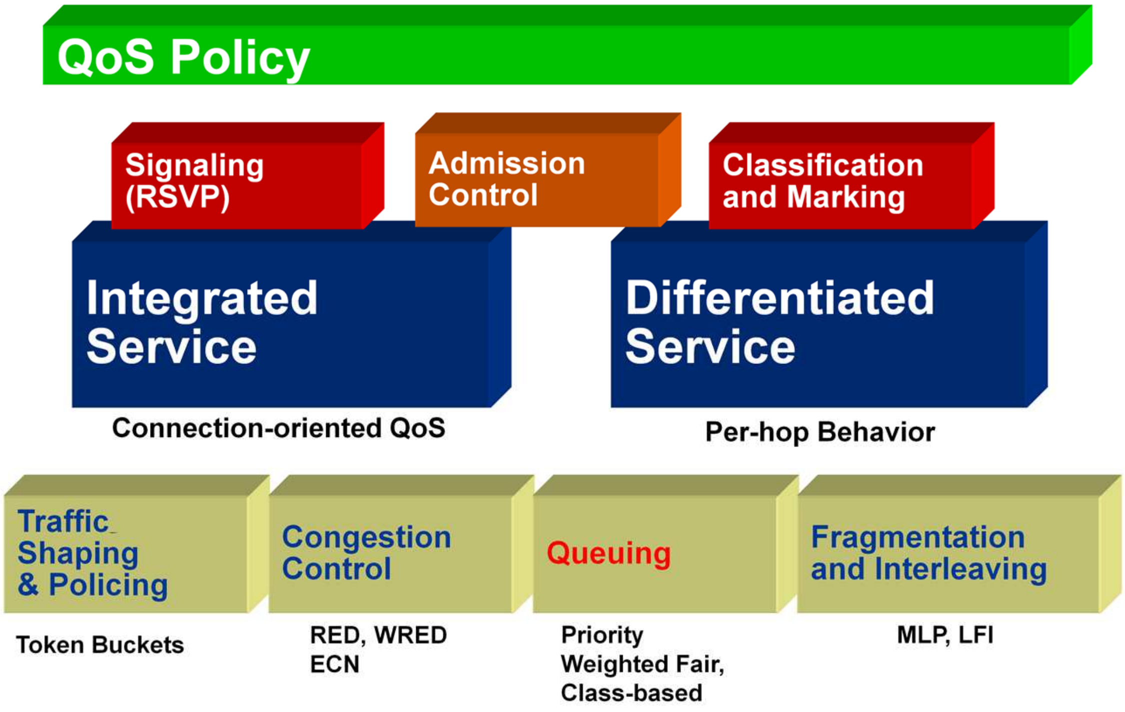
# IP Routers With QoS Support

- **Queuing actually encompasses two parts: SW and HW queues!**
- **SW queuing is typically more sophisticated**
  - WRR Weighted Round Robin)
  - CBWFQ (Class Based Weighted Fair Queuing)
  - Priority Queuing, LLQ (Low Latency Queuing)
  - These kind of techniques are an important part of any QoS implementation
- **HW queuing is typically only FIFO**
- **SW queue only needed if HW-queue full**
  - Otherwise packet bypasses SW-queue

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 72

Ethernet switches with QoS support are a little bit different. They often support advanced queuing techniques in hardware.

## Building Blocks for QoS



Mission Critical Communication Over IP Based Networks v3.0 73

The diagram above shows the main building blocks of a QoS design. Let's discuss it from the bottom to the top:

The first layer represents important tools which are almost always present in a QoS architecture:

**Traffic Shaping** (smooth bursts), **Congestion Control** (avoid bottlenecks and denial of service), **Queuing** are the most fundamental parts of QoS. **RED** (Random Early Discard) randomly drops segments before queue is full by utilizing TCP specific behavior of dynamically adjusting traffic throughput by reducing TCP window size. **WRED** (Weighted Random Early Discard) drops less important packets more aggressively than more important packets based on ToS/DSCP value. **ECN** **(**Early Congestion Notification) allows ECN capable routers to inform TCP sessions to reduce TCP window size before a packet drop occurs. **MLP** (Multilink PPP) and **LFI** (Link Fragment Interleaving) are techniques to provides load balancing functionality over multiple WAN links, while providing packet fragmentation and proper sequencing. MLP allows packets to be fragmented and the fragments to be sent at the same time over multiple point-to-point links to the same remote address. MLP reduces transmission latency across WAN links. LFI Interleaving on MLP allows large packets to be multilink encapsulated and fragmented into a small enough size to satisfy the delay requirements of real-time traffic. Small real-time packets are not multilink encapsulated and are sent between fragments of the large packets. The interleaving feature also provides a special transmit queue for the smaller, delay-sensitive packets, enabling them to be sent earlier than other flows.

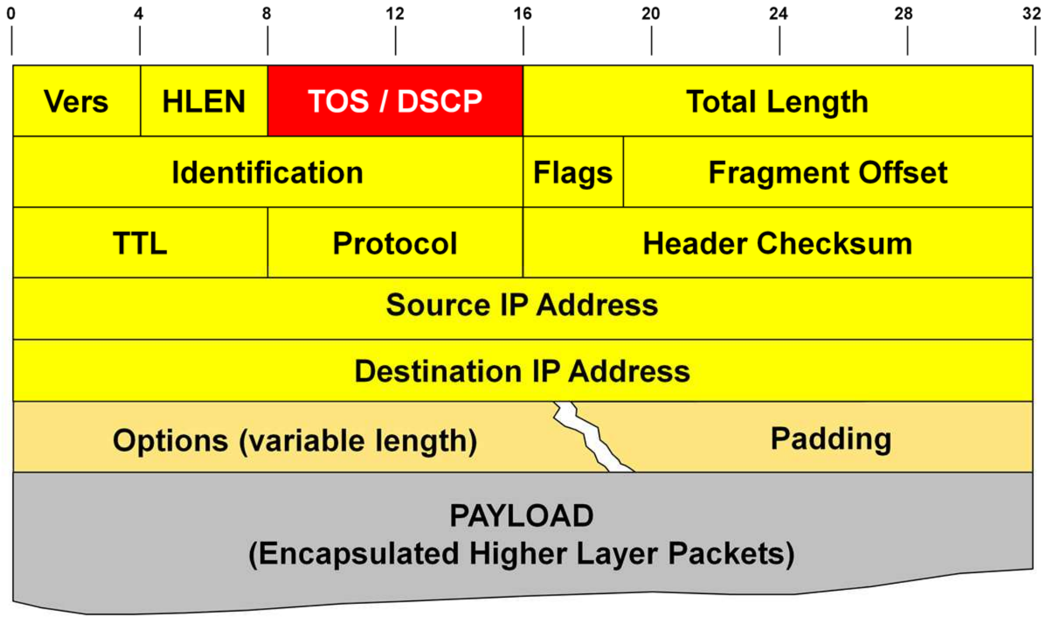
The next layer represents two totally different QoS architecture principles: **Integrated Service** ("**IntServ**") and **Differentiated Services** ("**DiffServ**"). IntServ (RFC 1633) is a connection-oriented approach but seems not to be scalable for large networks. DiffServ (RFC 2475) is connectionless and more scalable but QoS is difficult to control. Both technologies adjust the queuing parameters.

We will concentrate on DiffServ principle for our discussion of IP QoS.

On the next layer, the diagram presents the instances that are used to implement QoS: IntServ uses a signaling protocol called **RSVP** **(**Resource ReSerVation Protocol, RFC 2205) and DiffServ uses a **Classification** and **Marking** engine in the packet switches (routers). **Admission Control** assign QoS to dedicated users (dynamic on demand by usage of signaling for IntServ and static in a provisioning phase for DiffServ).

Finally on the top layer the **QoS Policy** constitutes a fundament for the whole QoS concept. Here all desires and rules are specified.

## IP Header Field TOS / DSCP
### (Used as Indication of QoS Service Class)



Vers | HLEN | TOS / DSCP | Total Length

Identification | Flags | Fragment Offset

TTL | Protocol | Header Checksum

Source IP Address

Destination IP Address

Options (variable length) | Padding

PAYLOAD
(Encapsulated Higher Layer Packets)

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 74

The IP header supports classification of service classes in the TOS (Type of Service, 8 bits.) or DSCP (Differentiated Services Code Point, 6 bits) field. Usage of original ToS (RFC 1430, 1349) was seldom implemented in the applications and only partly supported by routers. DSCP (RFC 2474) is the current way to implement DiffServ in the IP world by classifying IP datagrams to belong to a certain QoS class, which is handled by the routers according to the QoS policy.

ToS (old meaning):

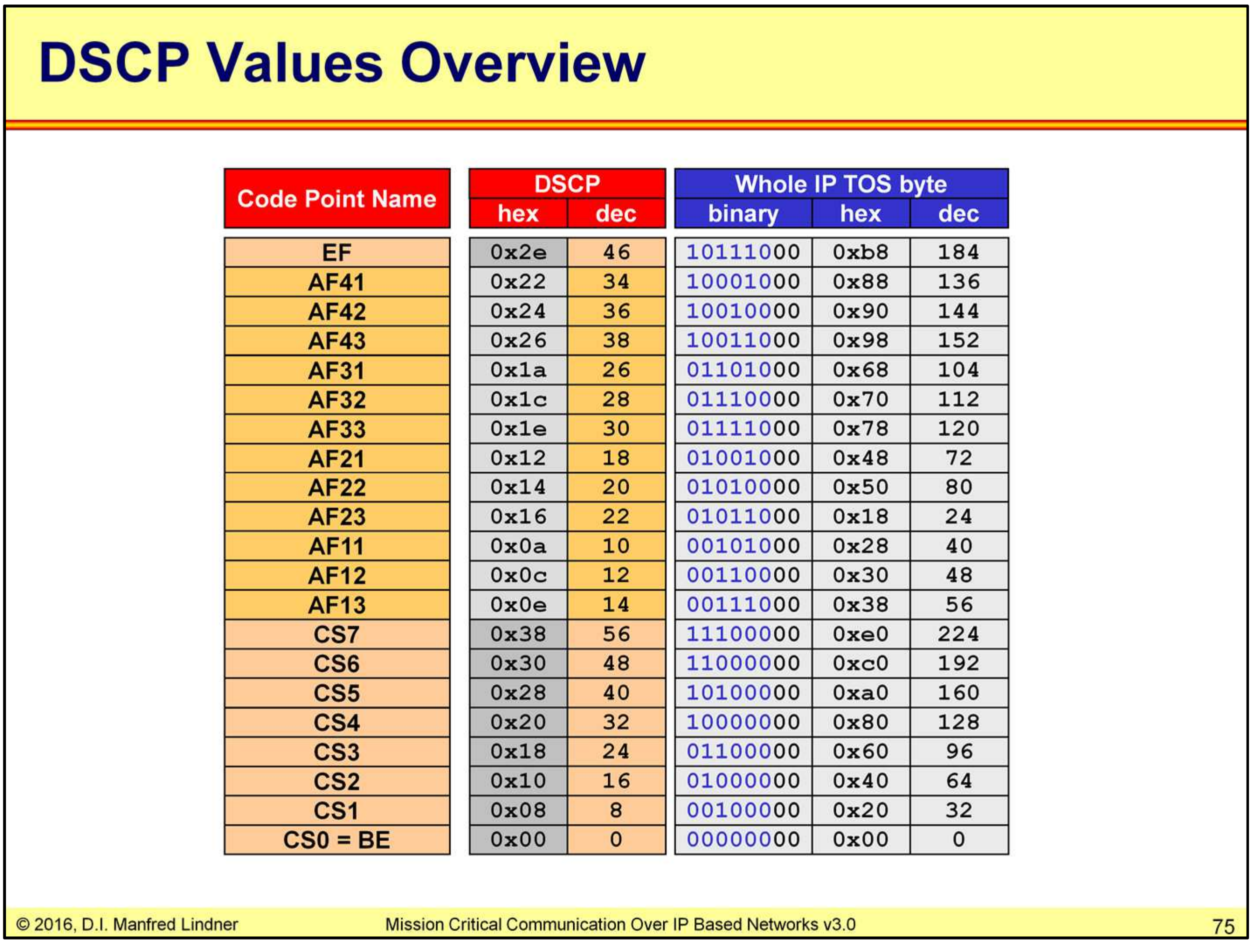
Tells the priority of a datagram (3 precedence bits) and the preferred network characteristics (4 bits, low delay D, high throughput T, high reliability R, low monetary cost B). Precedence bits define the handling of a datagram within the router e.g. priority within the input / output queues. D, T, R and C bits can be used to take a path decision for routing if multiple paths with different characteristics exist to the destination. That needs one routing table per characteristic which is not supported by current routing protocols anymore. TOS bits may be ignored by routers but may never lead to discarding a packet if the preferred service cannot be provided.

DSCP (current meaning for DiffServ):

Bits 0,1,2 – Class Selector: Again carries the 0-7 IP ToS precedence value. However there is a new definition of these values. See next slide.

Bits 3,4 – Drop Precedence: Carry the drop precedence which allows a further differentiation of traffic within the same class level. Note that this requires bit 5 to be zero.

Bits 6,7 – Currently Unused (or ECN…)



## DSCP Values Overview

| Code Point Name | DSCP | | Whole IP TOS byte | | |
|---|---|---|---|---|---|
| | hex | dec | binary | hex | dec |
| EF | 0x2e | 46 | 10111000 | 0xb8 | 184 |
| AF41 | 0x22 | 34 | 10001000 | 0x88 | 136 |
| AF42 | 0x24 | 36 | 10010000 | 0x90 | 144 |
| AF43 | 0x26 | 38 | 10011000 | 0x98 | 152 |
| AF31 | 0x1a | 26 | 01101000 | 0x68 | 104 |
| AF32 | 0x1c | 28 | 01110000 | 0x70 | 112 |
| AF33 | 0x1e | 30 | 01111000 | 0x78 | 120 |
| AF21 | 0x12 | 18 | 01001000 | 0x48 | 72 |
| AF22 | 0x14 | 20 | 01010000 | 0x50 | 80 |
| AF23 | 0x16 | 22 | 01011000 | 0x18 | 24 |
| AF11 | 0x0a | 10 | 00101000 | 0x28 | 40 |
| AF12 | 0x0c | 12 | 00110000 | 0x30 | 48 |
| AF13 | 0x0e | 14 | 00111000 | 0x38 | 56 |
| CS7 | 0x38 | 56 | 11100000 | 0xe0 | 224 |
| CS6 | 0x30 | 48 | 11000000 | 0xc0 | 192 |
| CS5 | 0x28 | 40 | 10100000 | 0xa0 | 160 |
| CS4 | 0x20 | 32 | 10000000 | 0x80 | 128 |
| CS3 | 0x18 | 24 | 01100000 | 0x60 | 96 |
| CS2 | 0x10 | 16 | 01000000 | 0x40 | 64 |
| CS1 | 0x08 | 8 | 00100000 | 0x20 | 32 |
| CS0 = BE | 0x00 | 0 | 00000000 | 0x00 | 0 |

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 75

Expedited Forwarding (EF):

DSCP 46 = 101 110 binary

For low delay, low loss, and low jitter

Defined in RFC 3246

Assured Forwarding (AF):

12 codepoints: 4 classes and 3 drop precedence each

Defined in RFC 2597

Guarantees a certain bandwidth to a traffic class.

If the traffic exceeds the committed bandwidth the drop probability is raised according to the specified drop precedence.

There are 12 different AF behavior code points consisting of 4 classes (AF1y to AF4y) and 3 drop probabilities (AFx1 to AFx3) for each class (low/med/hi)

Best Effort (BE):

000000 binary

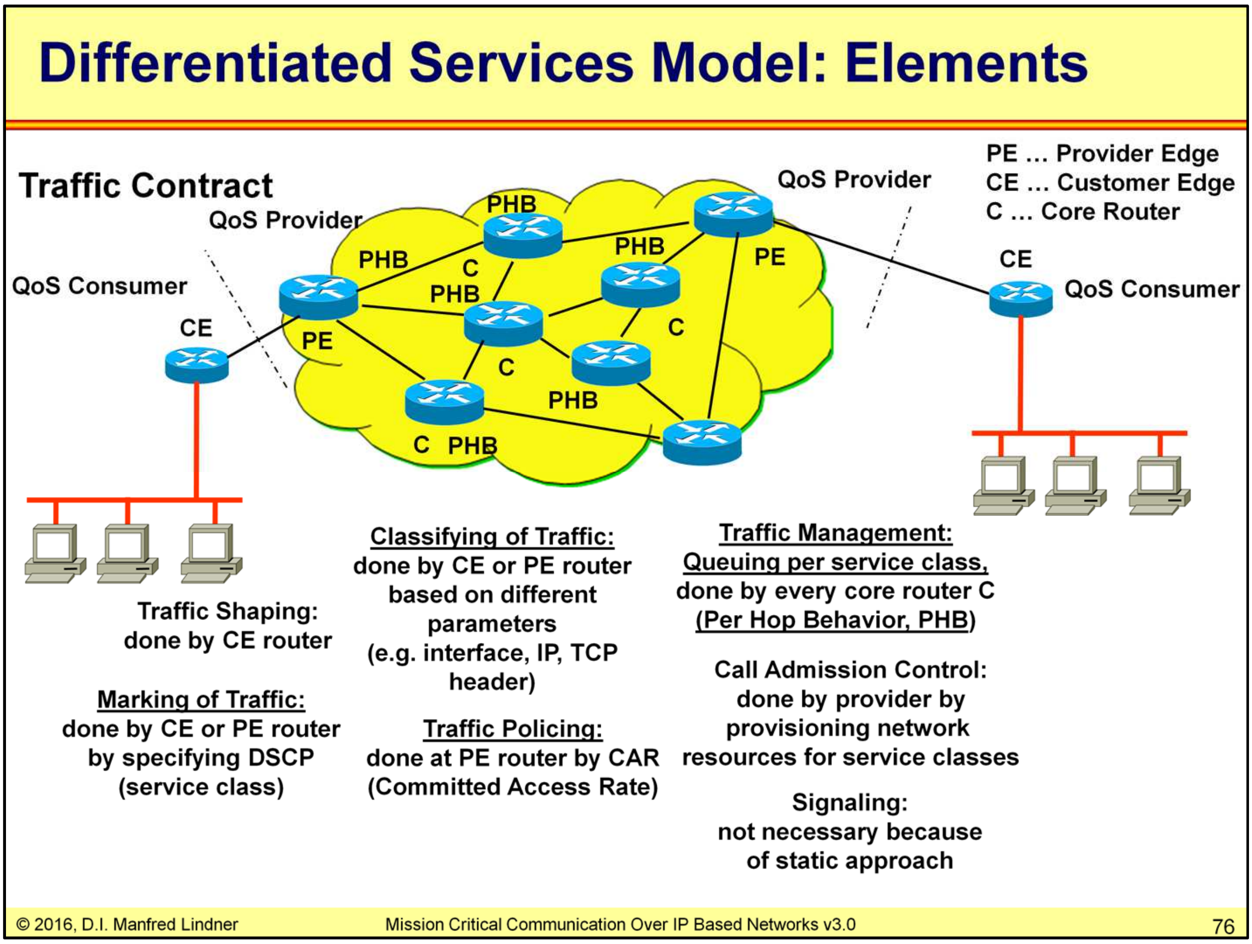
The legacy ToS IP Precedence values (0-3) can be directly mapped into the three Class Selector bits (0,1,2) with the three LSBs (3,4,5) set to zero

This results in the seven CSx values

CS0 = DSCP 00 = 000000

...

CS7 = DSCP 56 = 111000



## Differentiated Services Model: Elements

PE ... Provider Edge
CE ... Customer Edge
C ... Core Router

**Traffic Shaping:** done by CE router

**Marking of Traffic:** done by CE or PE router by specifying DSCP (service class)

**Classifying of Traffic:** done by CE or PE router based on different parameters (e.g. interface, IP, TCP header)

**Traffic Policing:** done at PE router by CAR (Committed Access Rate)

**Traffic Management:** Queuing per service class, done by every core router C (Per Hop Behavior, PHB)

**Call Admission Control:** done by provider by provisioning network resources for service classes

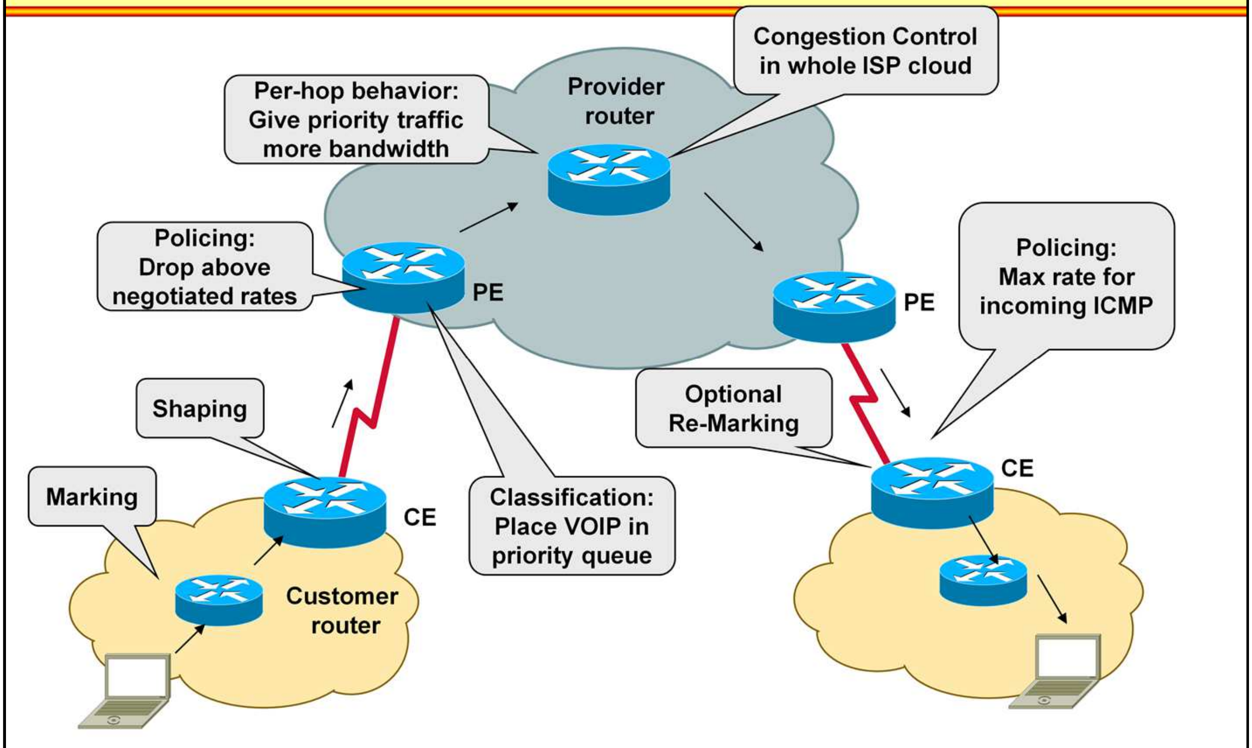
**Signaling:** not necessary because of static approach

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 76

QoS mechanism are traffic marking (TM, assigning an IP packet to a QoS service class), traffic classification (TC, assigning traffic into different queues of a router/switch), traffic policing (TP, controlling amount of traffic of a given service class and acting on violation of traffic contract for this service class), traffic shaping (TS, reforming traffic characteristics regarding burstiness and average throughput to fulfill a given traffic contract) and finally queuing strategy applied to queues maintained by a router (PHP, Per-Hop-Behavior). Typical queuing strategies are WFQ (Weighted Fair Queuing), PQ (Priority Queuing), FIFO (FirstInFirstOut Queuing), WRR (Weighted Round Robin), CBWFQ (Class Based WFQ) and LLQ (Low Latency Queuing - combination of PQ and CBWFQ).

Recommendation: Analyze the tradeoff between QoS management compared to over-provisioning in your solution. Think about if it is really worth to invest in technology and operational people instead of taking the more expensive wire. Especially testing is a challenge, because QoS comes only into action if there is congestion in the network. Producing of huge traffic may be a problem in a simple test-lab for verification of overall functionality and performance. Fixing it in the operational network might be not possible. A sound network monitoring is necessary in all cases.

# Differentiated Services Model: In Action



Per-hop behavior: Give priority traffic more bandwidth

Congestion Control in whole ISP cloud

Policing: Drop above negotiated rates

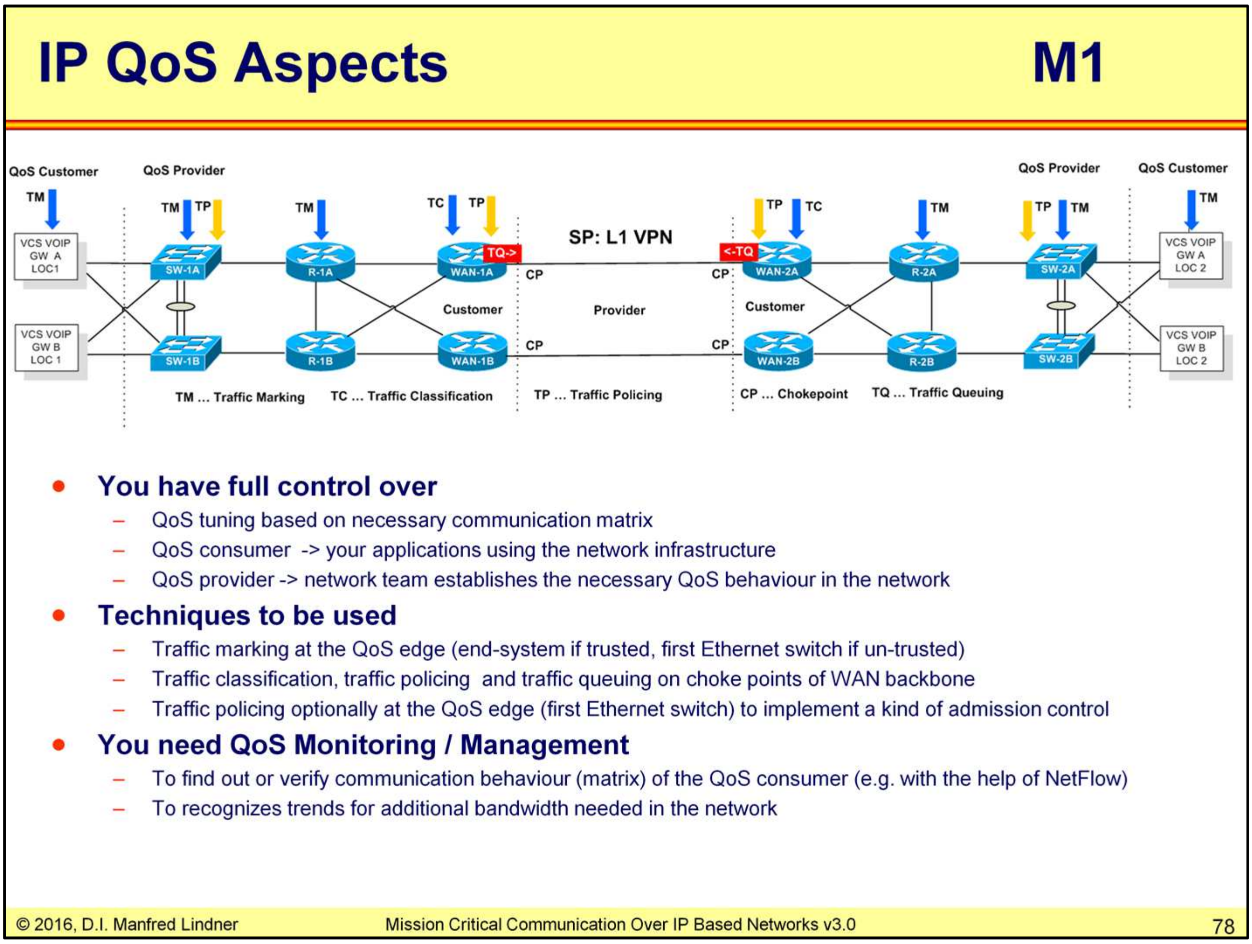
Policing: Max rate for incoming ICMP

Shaping

Optional Re-Marking

Marking

Classification: Place VOIP in priority queue



## IP QoS Aspects M1

- **You have full control over**
  - QoS tuning based on necessary communication matrix
  - QoS consumer -> your applications using the network infrastructure
  - QoS provider -> network team establishes the necessary QoS behaviour in the network
- **Techniques to be used**
  - Traffic marking at the QoS edge (end-system if trusted, first Ethernet switch if un-trusted)
  - Traffic classification, traffic policing and traffic queuing on choke points of WAN backbone
  - Traffic policing optionally at the QoS edge (first Ethernet switch) to implement a kind of admission control
- **You need QoS Monitoring / Management**
  - To find out or verify communication behaviour (matrix) of the QoS consumer (e.g. with the help of NetFlow)
  - To recognizes trends for additional bandwidth needed in the network

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 78

Depending on the operational model QoS mechanism can implemented with reasonable effort in case you have full control over the network (e.g. only L1-VPNs are provided by external partners). In case part of the network is operated by a service provider (L2-VPN or L3-VPN) QoS management and supervision becomes much more complicated.

IP QoS aspects for operational model M1:

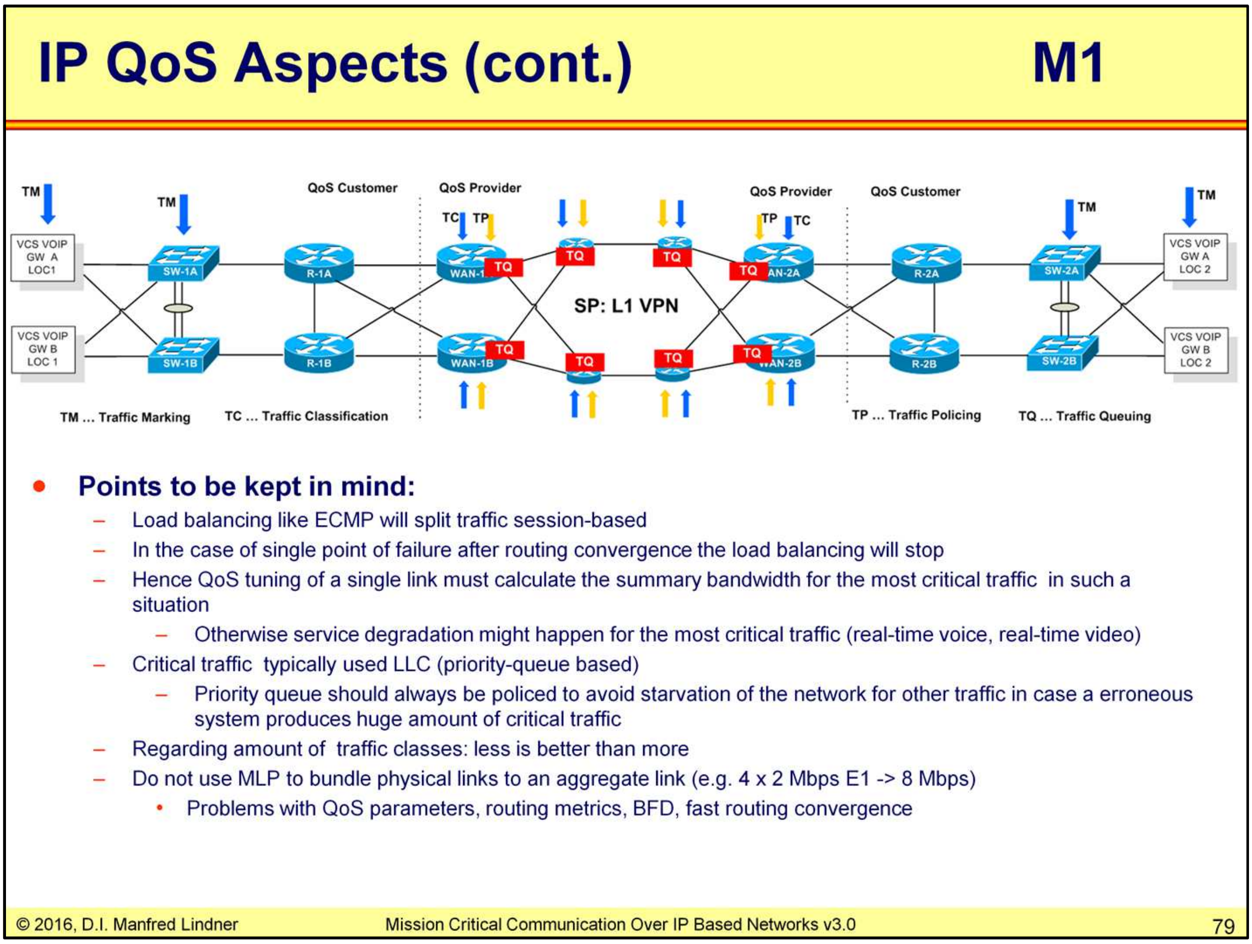
Model M1 is based on L1-VPN links with (dedicated) constant bandwidth and constant delay provided by an external service provider. Usually the bandwidth on the WAN links will be much slower (e.g. 1.5 to 8 Mbps) than the access speed at the location (Fast Ethernet, Gigabit Ethernet). So probability is very high that the WAN link becomes a real chokepoint.

The customer has full control over the network by managing all active L2 (Ethernet switches) and L3 components (IP routers). QoS mechanism can be implemented and tuned with reasonable effort based on the communication matrix of the distributed systems.

The communication matrix describes who is talking to whom (source and destination IP addresses, IP protocol type, TCP/UDP port numbers, and communication style such as unicast, multicast or broadcast) and the traffic statistics of the communication sessions (average bandwidth, burstiness). There is a (hidden) traffic contract between QoS consumers (the applications) and QoS provider (the overall network operated by the customer).

Traffic marking (TM) can be done either by the end-systems themselves or by the first Ethernet switch (preferred -> TM usually done in hardware) or IP router (second choice -> TM usually done in software) where end-systems are connected to. The decision where to do TM depends if end-system can be trusted or not. If they are not trusted the network should apply the service class in the IP DSCP field of bypassing IP packets. Optionally the first network component can apply also traffic policing (TP) to reduce amount of traffic of a given end-system to that what is agreed. Traffic classification (TC) and traffic queuing (TQ) is a must at the chokepoint (outgoing interface towards L1-VPN link at WAN-1x, WAN-2x routers).

Optionally traffic policing (TP) can be performed at the choke-point to eliminate any unwanted traffic during times of congestion without filling the queues of a service class. QoS monitoring (statistics about queue depths, dropped packets, average usage per QoS service class) can give you a feeling, what is really going on in your network either verifying the communication matrix or adjusting the knowledge about the communication behavior of your applications.



## IP QoS Aspects (cont.)

M1

- **Points to be kept in mind:**
  - Load balancing like ECMP will split traffic session-based
  - In the case of single point of failure after routing convergence the load balancing will stop
  - Hence QoS tuning of a single link must calculate the summary bandwidth for the most critical traffic in such a situation
    - Otherwise service degradation might happen for the most critical traffic (real-time voice, real-time video)
  - Critical traffic typically used LLC (priority-queue based)
    - Priority queue should always be policed to avoid starvation of the network for other traffic in case a erroneous system produces huge amount of critical traffic
  - Regarding amount of traffic classes: less is better than more
  - Do not use MLP to bundle physical links to an aggregate link (e.g. 4 x 2 Mbps E1 -> 8 Mbps)
    - Problems with QoS parameters, routing metrics, BFD, fast routing convergence

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 79

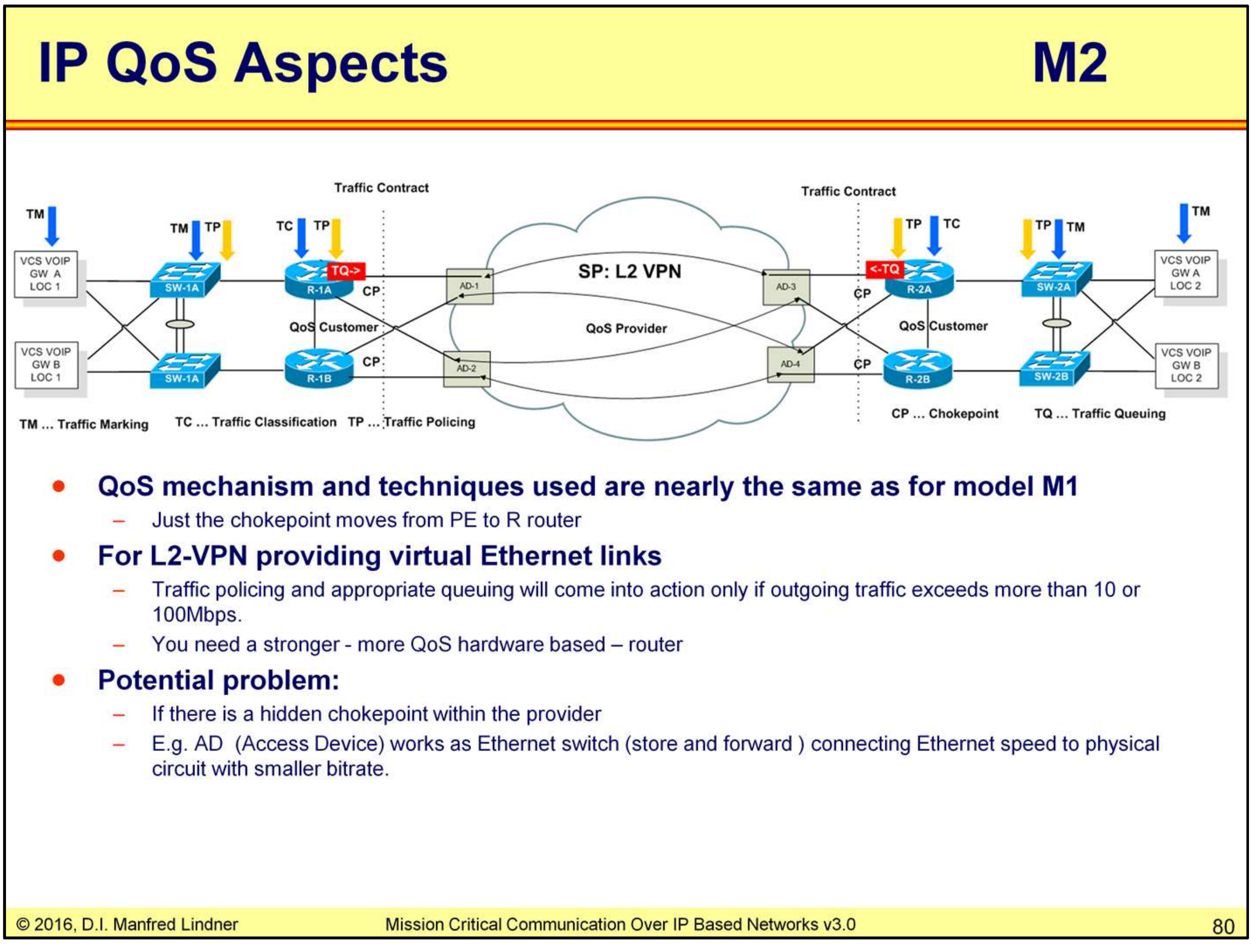
Attention: QoS monitoring gives you information about a service class. If more than one communication session is using a service class - that is the common approach - then you cannot see the details of a particular communication session. In such a case NetFlow monitoring performed on IP routers should additionally be used to see more details, which has to be correlated with QoS monitoring statistics. Depending on the quality of your network management system, that can be achieved with different effort. NetFlow allows you to see how many packets per communication session are passing across an IP router (routers R-1x or R-2x are candidates for NetFlow).

The pictures shows a typical distribution of QoS mechanism for supporting X locations and placing the border between QoS consumer and QoS provider on the border between LAN and IP WAN.

QoS tuning can become quite tricky in such a scenario. Remember ECMP where load balancing is performed on a session base. If one path fails the other will take over the traffic either immediately or after routing convergence succeeds. QoS tuning of a single link has to calculate such events. Therefore the summary bit rate at least for the most critical traffic (e.g. real-time voice) flowing between two locations has to be calculated and provisioned on such a link. Otherwise service degradation will happen in case of a failure. So load balancing sounds good for high availability but with from QoS point of view it is more complicated. Assure that your network has reasonable bandwidth in all possible scenarios of single point-of-failures independent from the physical network topology. That means that you have to support the sum of all critical traffic over all core physical links in the worst case.

Further critical traffic is typically priority queued to achieve low latency. If a device creates a huge amount of such traffic (either by a failure or part of a security attack) all other service classes will starved. Therefore priority queues and/or critical traffic should always be policed in order to avoid starvation of other traffic, which may be even not so critical but still important. In most cases you have only one priority queue available at IP routers. So splitting critical traffic (which needs low latency) into many service classes will not help.

Golden rule: Keep the amount of traffic classes low.



## IP QoS Aspects M2

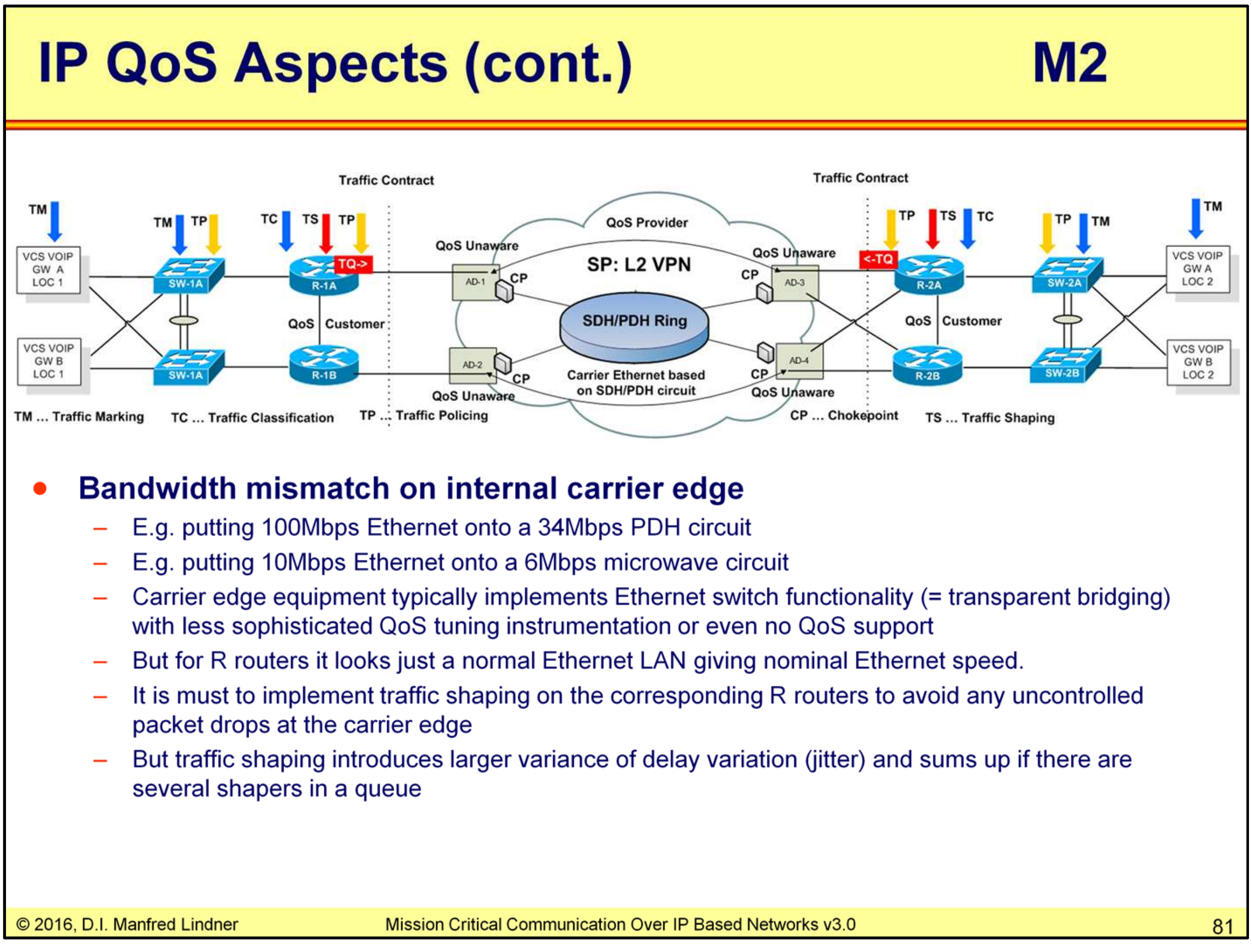
- **QoS mechanism and techniques used are nearly the same as for model M1**
  - Just the chokepoint moves from PE to R router
- **For L2-VPN providing virtual Ethernet links**
  - Traffic policing and appropriate queuing will come into action only if outgoing traffic exceeds more than 10 or 100Mbps.
  - You need a stronger - more QoS hardware based – router
- **Potential problem:**
  - If there is a hidden chokepoint within the provider
  - E.g. AD (Access Device) works as Ethernet switch (store and forward ) connecting Ethernet speed to physical circuit with smaller bitrate.

Mission Critical Communication Over IP Based Networks v3.0 80

IP QoS aspects for operational model M2:

Model M2 is based on L2-VPN links with constant access bandwidth but variable delay provided by an external service provider. One example is a VPLS service based on an IP-MPLS infrastructure with e.g. 10 or 100Mbps Ethernet links towards the customer. If outgoing traffic of a location can reach more than 10 or 100Mbps then QoS mechanism can be applied in nearly the same way as for model M1. Only traffic classification (TC) and traffic queuing will move from the WAN router to the R-1x, R2-x routers. Keep in mind that QoS mechanism applied for link speed higher than 10Mbps needs a stronger router (hardware based QoS).

Be careful if your traffic passes regions where no QoS support can be implemented. You will turn back to best-effort system over such region which may not satisfy your QoS requirements in case of congestion. Typically example are gateways between Ethernet and microwave or SDH circuits which behaves as transparent bridge (= Ethernet switch) store and forward packet switching device without knowing about L2 or L3 QoS hence doing best-effort FIFO queuing only. Next slide explains such a problem.



## IP QoS Aspects (cont.) M2

- **Bandwidth mismatch on internal carrier edge**
  - E.g. putting 100Mbps Ethernet onto a 34Mbps PDH circuit
  - E.g. putting 10Mbps Ethernet onto a 6Mbps microwave circuit
  - Carrier edge equipment typically implements Ethernet switch functionality (= transparent bridging) with less sophisticated QoS tuning instrumentation or even no QoS support
  - But for R routers it looks just a normal Ethernet LAN giving nominal Ethernet speed.
  - It is must to implement traffic shaping on the corresponding R routers to avoid any uncontrolled packet drops at the carrier edge
  - But traffic shaping introduces larger variance of delay variation (jitter) and sums up if there are several shapers in a queue

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 81

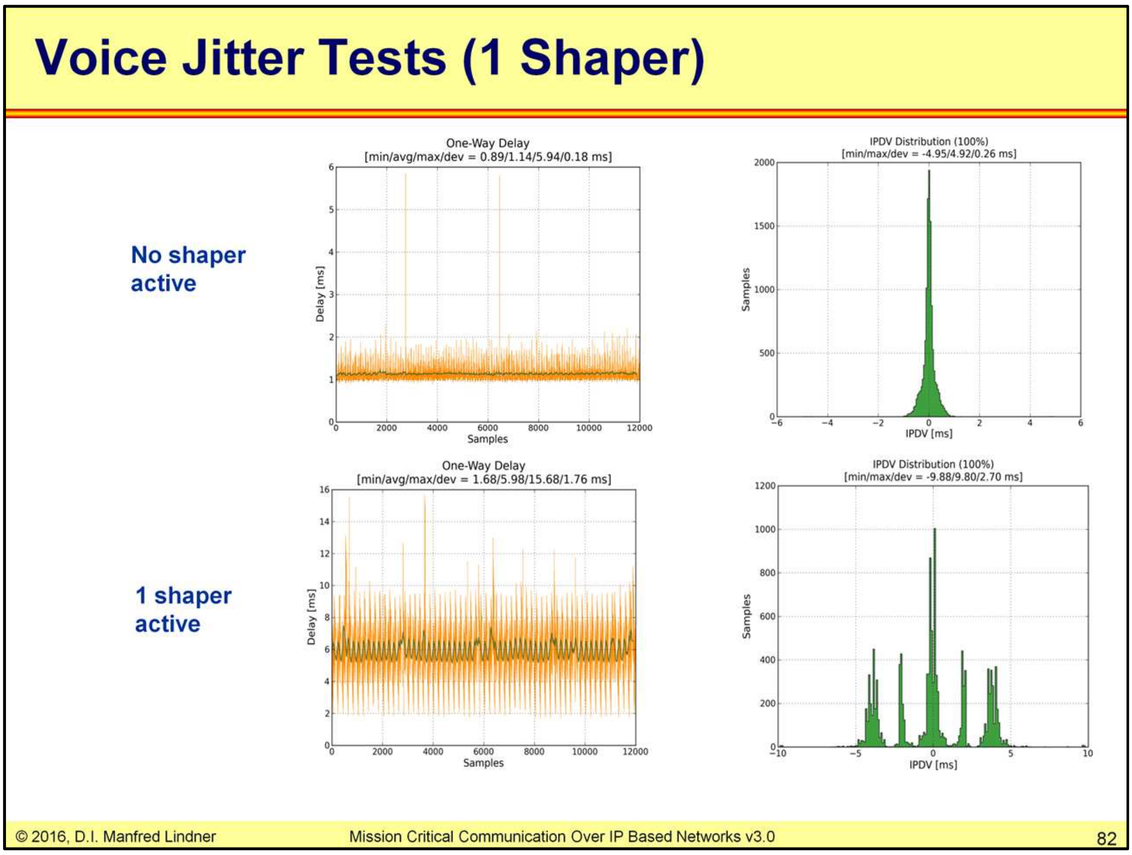
A L2-VPNs based on microwave/SAT modems or SDH-AD (Add/Drop) multiplexers offering their service on a corresponding Ethernet link to the customer routers can result in a speed mismatch between the Ethernet link and the microwave-link or SDH circuit (e.g. 100 Mbps Ethernet should use a 34 Mbps circuit).

In such a case the provider equipment acts as Ethernet switch performing packet switching between the different physical technologies. In case of packet bursts it can happen that buffering is not possible anymore and some packets get lost. Usually such provider equipment does not support L3 QoS mechanism. So if for example only 34Mbps are possible for a 100Mbps Ethernet link between R-1A and R-2A then any TQ mechanism will not work, because they will come into action if the 100Mbps are overloaded.

The only possible solution is usage of traffic shaping (TS) on R-1A and R-2A. Traffic shaping will configured in such a way that - whenever outgoing traffic exceeds 34 Mbps on the 100 Mbps Ethernet link - traffic shaping will backpressure this traffic allowing traffic queuing and traffic policing to act according the QoS policy. Of course traffic shaping - when active e.g. traffic above 34 Mbps - introduces additional delay and a more variable jitter. But it is the only possibility to deal with the QoS unaware provider equipment in between. The picture shows such a scenario.

Please recognize that the real chokepoint has moved to the provider device which is QoS unaware. At the customer routers TS simulates these chokepoints to get QoS mechanism into action one step before.

Attention: Traffic shapers produce a more statistical distributed delay variation (jitter). Therefore be careful not to add too many of such shapers in a sequence between your distributed systems relying on critical communication sessions.



# Voice Jitter Tests (1 Shaper)

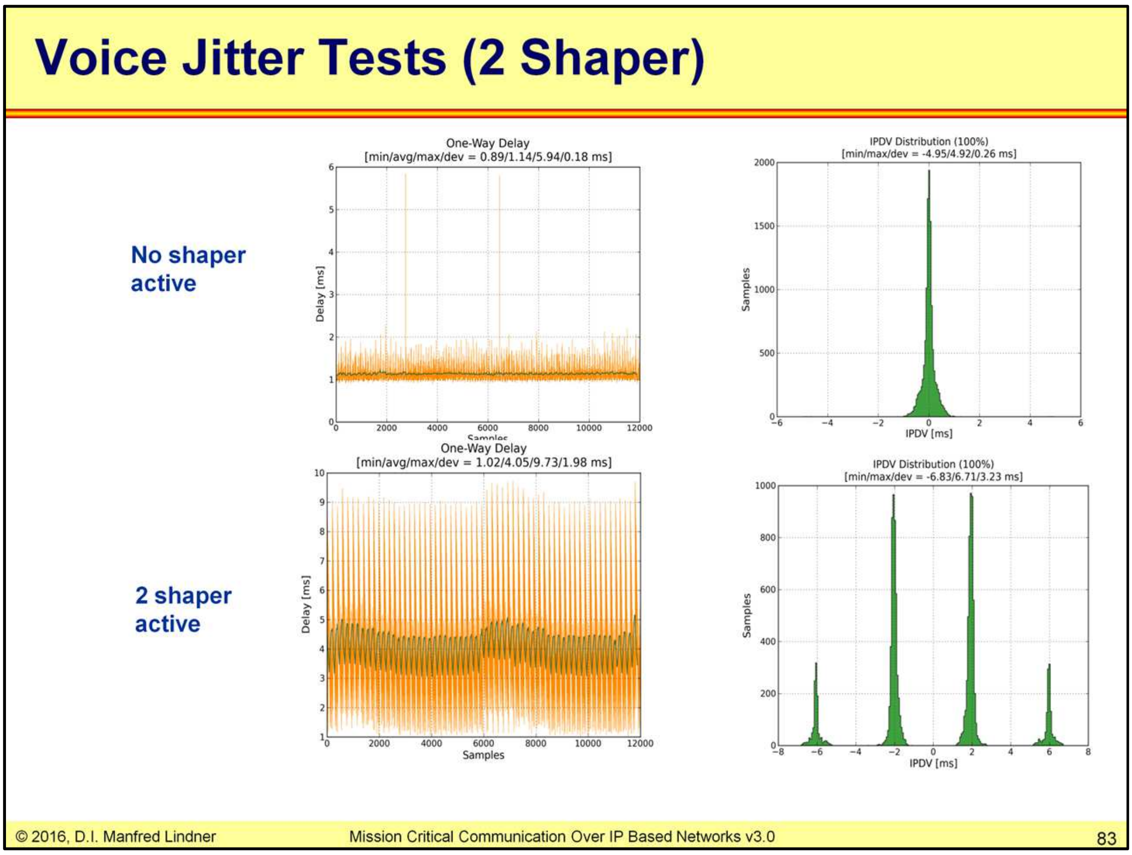
No shaper active

1 shaper active

© 2016, D.I. Manfred Lindner

Mission Critical Communication Over IP Based Networks v3.0

82



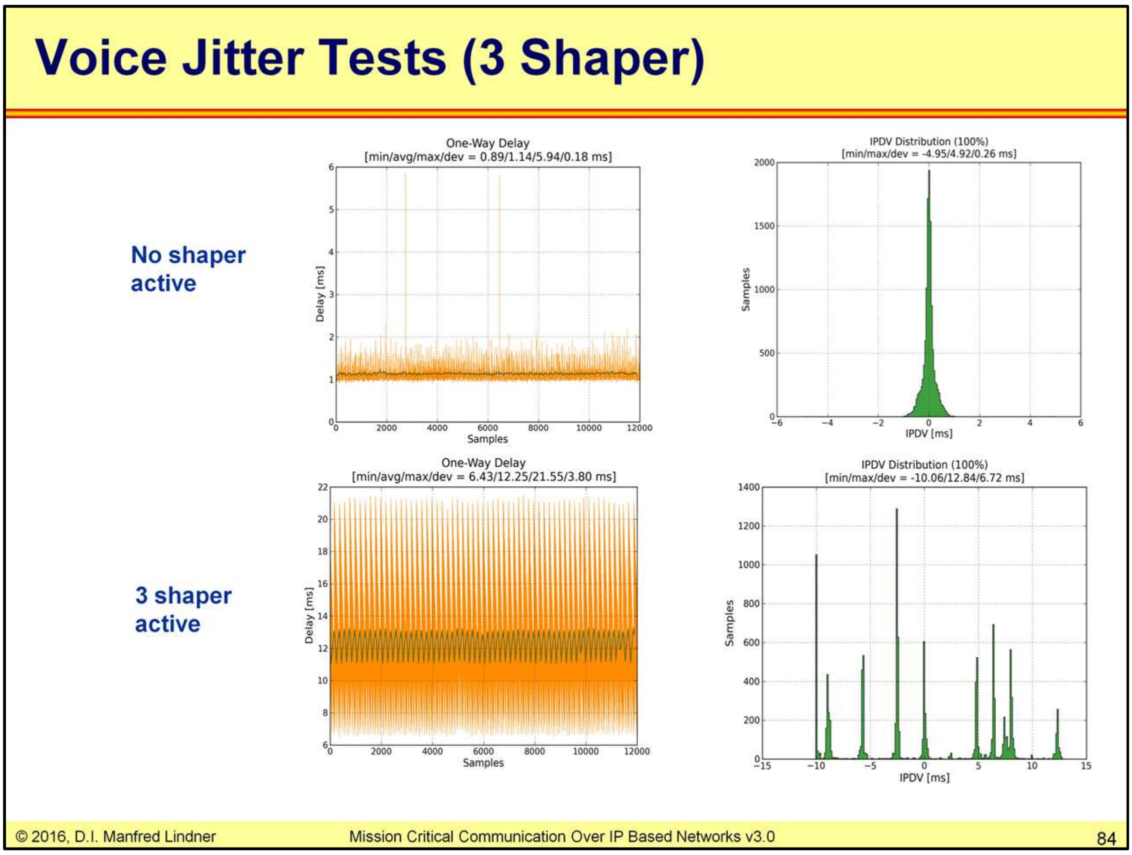
# Voice Jitter Tests (2 Shaper)

No shaper active

2 shaper active

Mission Critical Communication Over IP Based Networks v3.0

83

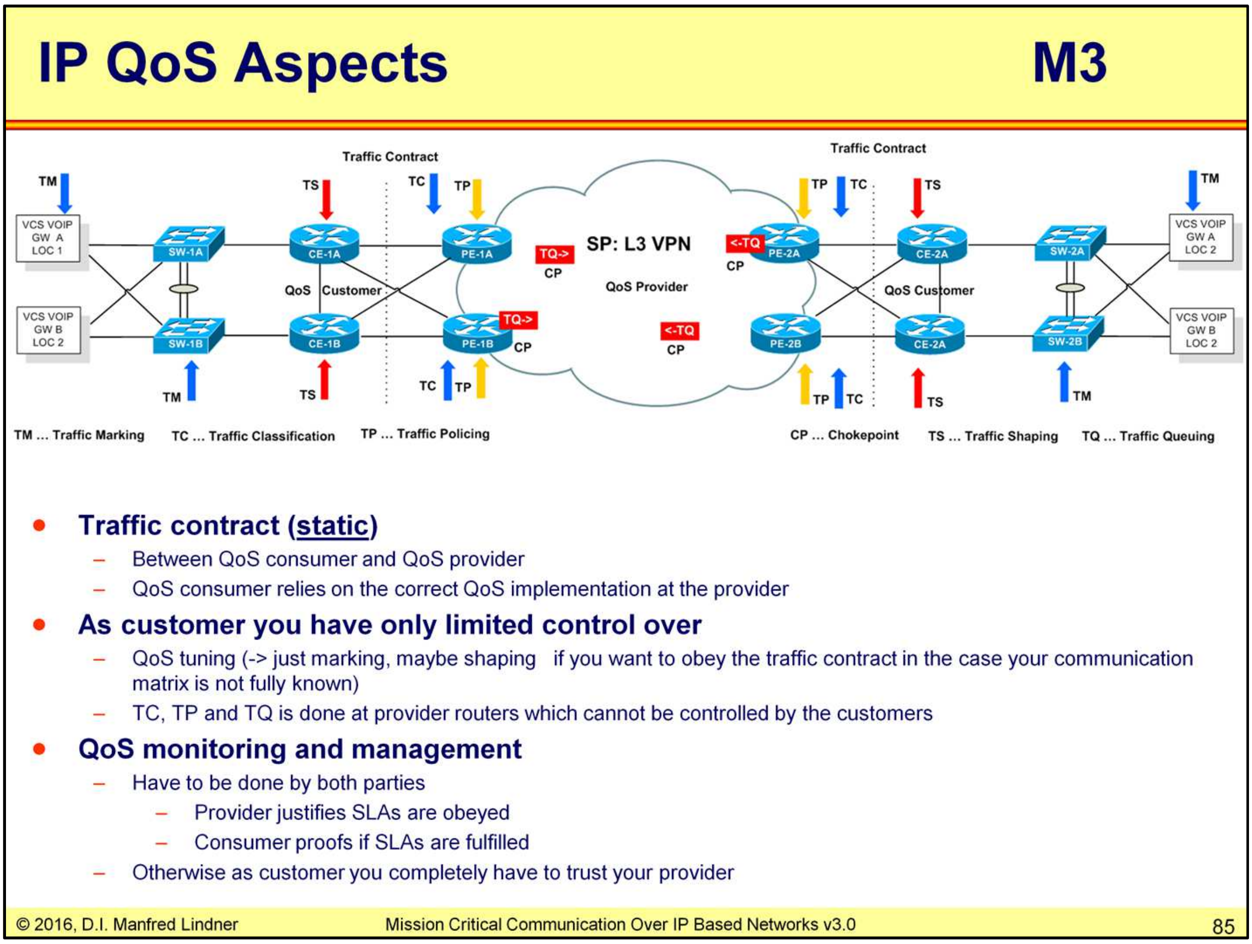


# Voice Jitter Tests (3 Shaper)

No shaper active

3 shaper active

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 84



## IP QoS Aspects M3

- **Traffic contract (static)**
  - Between QoS consumer and QoS provider
  - QoS consumer relies on the correct QoS implementation at the provider
- **As customer you have only limited control over**
  - QoS tuning (-> just marking, maybe shaping  if you want to obey the traffic contract in the case your communication matrix is not fully known)
  - TC, TP and TQ is done at provider routers which cannot be controlled by the customers
- **QoS monitoring and management**
  - Have to be done by both parties
    - Provider justifies SLAs are obeyed
    - Consumer proofs if SLAs are fulfilled
  - Otherwise as customer you completely have to trust your provider

© 2016, D.I. Manfred Lindner  Mission Critical Communication Over IP Based Networks v3.0  85

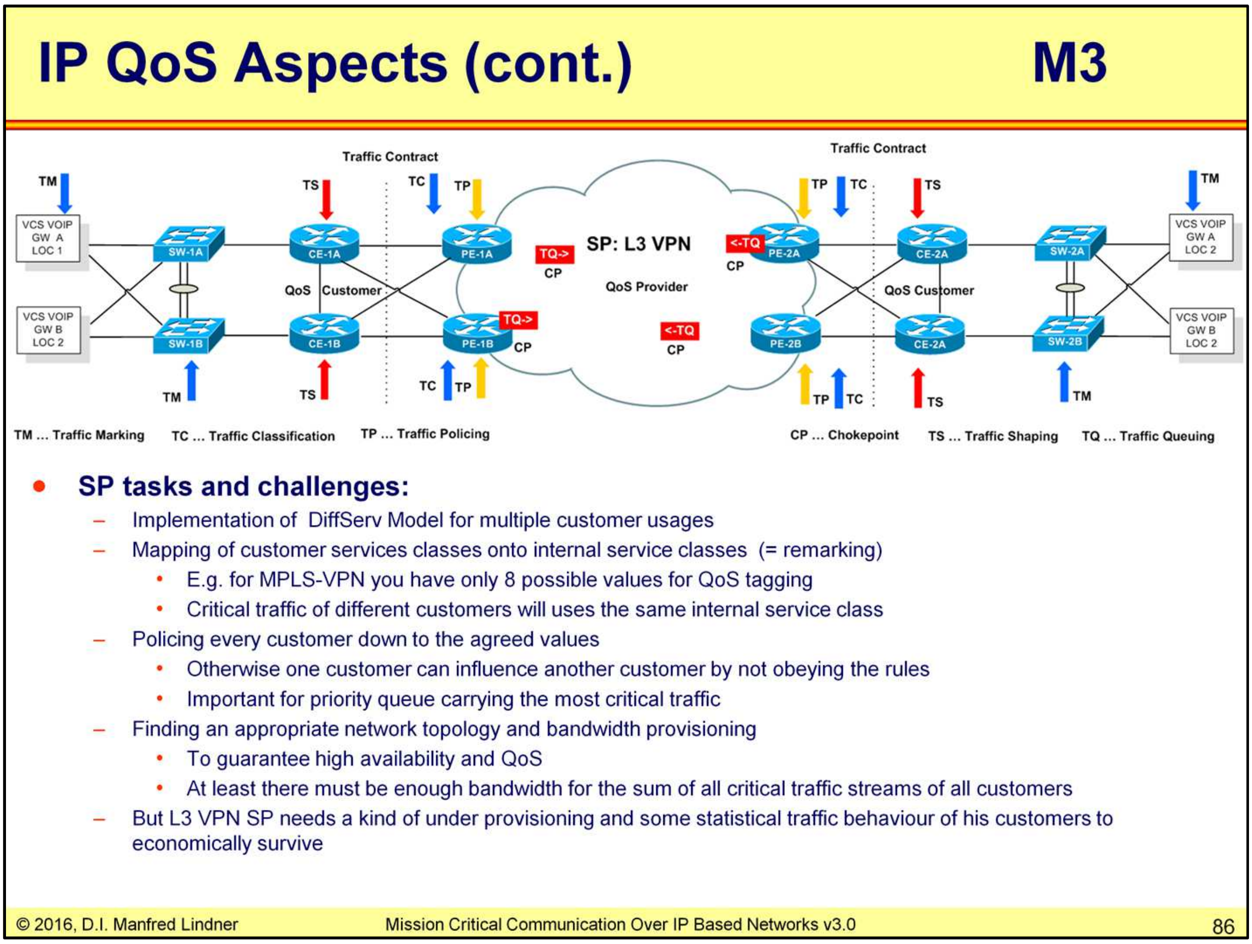
IP QoS aspects for operational model M3:

Model M3 is based on a L3-VPN provided by an external service provider. Nowadays such a L3-VPN is usually implemented by MPLS-VPN technology on top of an IP-MPLS packet switching network infrastructure. Without any IP QoS mechanism this is a best-effort network like the Internet (variable bandwidth or throughput and variable delay). Even if you are happy to find a provider offering IP QoS guarantees then this is still the most challenging approach.

First because a QoS contract has to be defined and implemented between QoS customer (network where the distributed processes reside) and QoS provider (network part which is controlled by a service provider and hence QoS statistics of provider part is not accessible for the customer or may not be trusted by the customer).

Second because either the customer has to fully trust the provider or has to perform QoS monitoring on its own to have a proof about the SLA to be guaranteed by the service provider. Obey that QoS traffic contract has static nature. So changes of the contract cannot be done in seconds. In needs some preliminary lead time especially to upgrade bandwidth in the network.

Third customer has only limited control over QoS tuning. Only traffic marking (TM) and optionally traffic shaping (TS) can be controlled by the customer in its domain. The chokepoints are somewhere within the provider network but technically supervised by traffic classification (TC) and traffic policing (TP) at the PE routers. The picture shows a typical scenario.



## IP QoS Aspects (cont.) M3

TM ... Traffic Marking TC ... Traffic Classification TP ... Traffic Policing CP ... Chokepoint TS ... Traffic Shaping TQ ... Traffic Queuing

- **SP tasks and challenges:**
  - Implementation of DiffServ Model for multiple customer usages
  - Mapping of customer services classes onto internal service classes (= remarking)
    - E.g. for MPLS-VPN you have only 8 possible values for QoS tagging
    - Critical traffic of different customers will uses the same internal service class
  - Policing every customer down to the agreed values
    - Otherwise one customer can influence another customer by not obeying the rules
    - Important for priority queue carrying the most critical traffic
  - Finding an appropriate network topology and bandwidth provisioning
    - To guarantee high availability and QoS
    - At least there must be enough bandwidth for the sum of all critical traffic streams of all customers
  - But L3 VPN SP needs a kind of under provisioning and some statistical traffic behaviour of his customers to economically survive

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 86

Also the service provider has a quite a challenge and a lot of tasks when offering IP QoS to customers. Why?

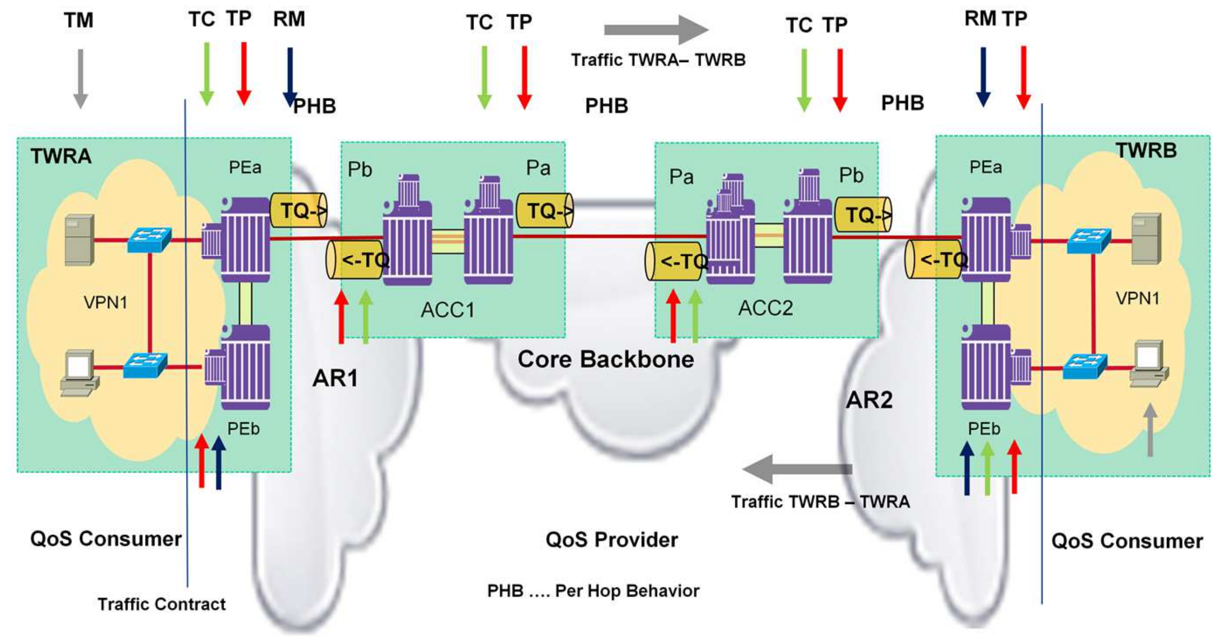
Implementation of payable IP QoS service based on Differentiated Services model for multiple customers is not so straight forward than implementing that model in a single domain such as our M1 network operational model. For example mapping of customer service classes agreed by the particular traffic contracts to only 8 QoS tags usable for MPLS (so called experimental Bits in the MPLS header) will lead to aggregation of a lot of critical customer traffic in just one service class used within the provider infrastructure. Remark: A service provider offering IP QoS based on LISP network technology would have 64 DSCP values instead of 8.

Policing every customer down to the agreed values is a must otherwise one misbehaving customer will influence other customers because of service class aggregation used within the provider. That is especially important for the priority queues carrying the most critical traffic.

Last but not least finding an appropriate network topology and bandwidth provisioning in order to guarantee high availability by usage of redundancy and automatic switchover together with QoS provisioning is not quite easy. Remember the discussion about ECMP and QoS in case of a failure. In the worst case every link in the core has to provide the sum of all bandwidth guarantees for all customers. But on the other side a service provider needs a kind of under-provisioning and some statistical traffic behavior of its customer (that means not all customers sending their guaranteed traffic all the time) to economically survive.

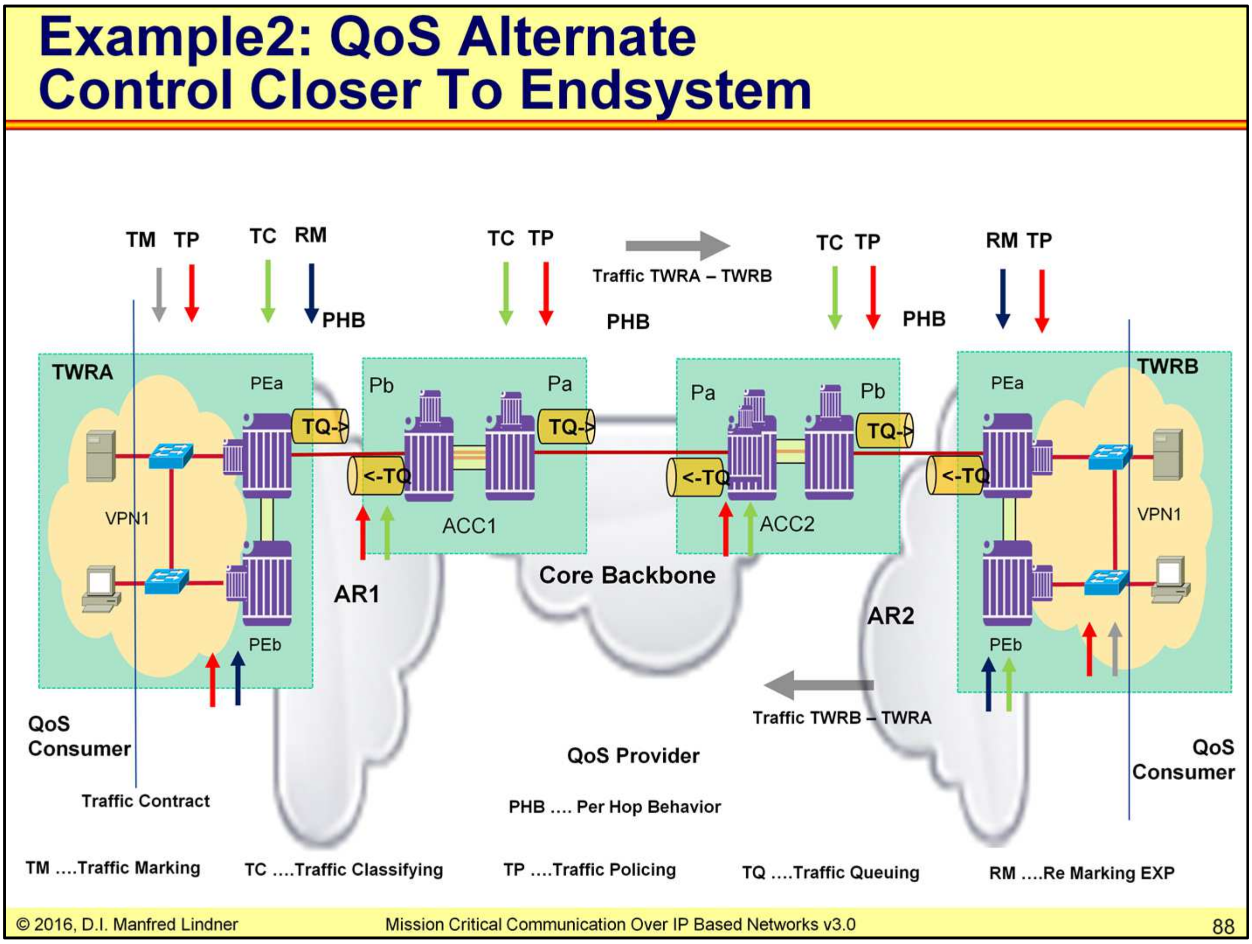
QoS monitoring is a must in order to proactively react by increasing bandwidth on the network links before customers complain about performance. QoS reports given to the customer for proofing SLAs need a certain kind of trust relationship between the customer and the provider. Trust is based on professional working and a real service oriented attitude. Both have its costs.

# Example1: QoS Functions Overview MPLS-VPN Based



TM ….Traffic Marking TC ….Traffic Classifying TP ….Traffic Policing TQ ….Traffic Queuing RM ….Re Marking EXP

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 87



# Example2: QoS Alternate Control Closer To Endsystem

TM ….Traffic Marking

TC ….Traffic Classifying

TP ….Traffic Policing

TQ ….Traffic Queuing

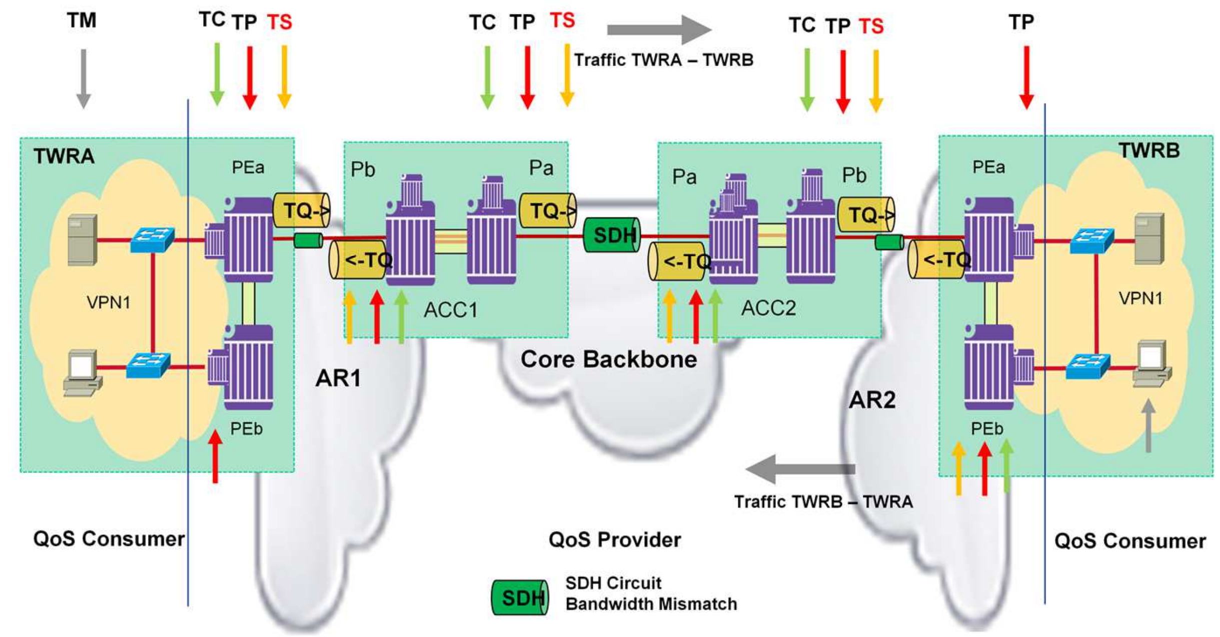
RM ….Re Marking EXP

© 2016, D.I. Manfred Lindner

Mission Critical Communication Over IP Based Networks v3.0

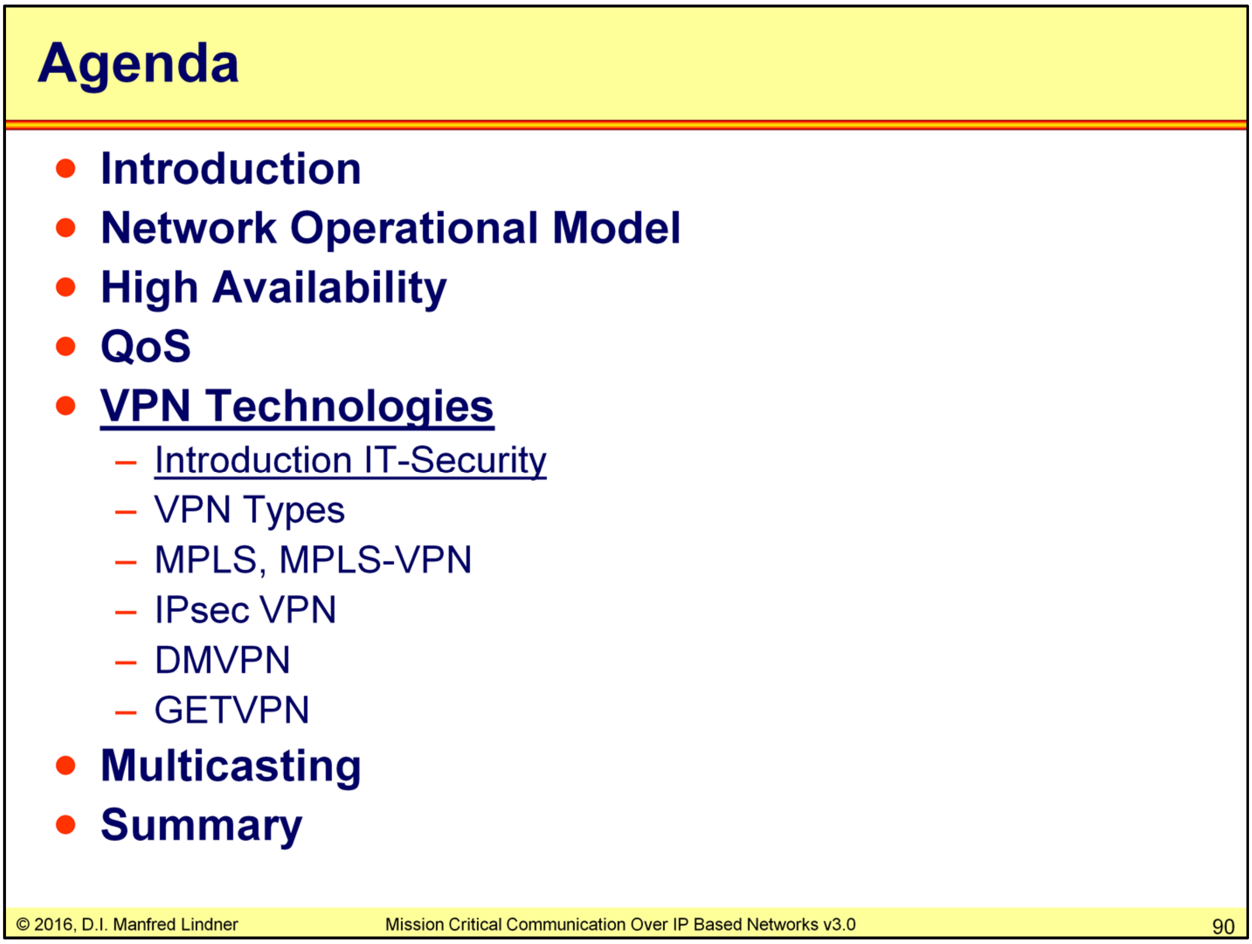
88

# Example 3: QoS Functions With Bandwidth Mismatch



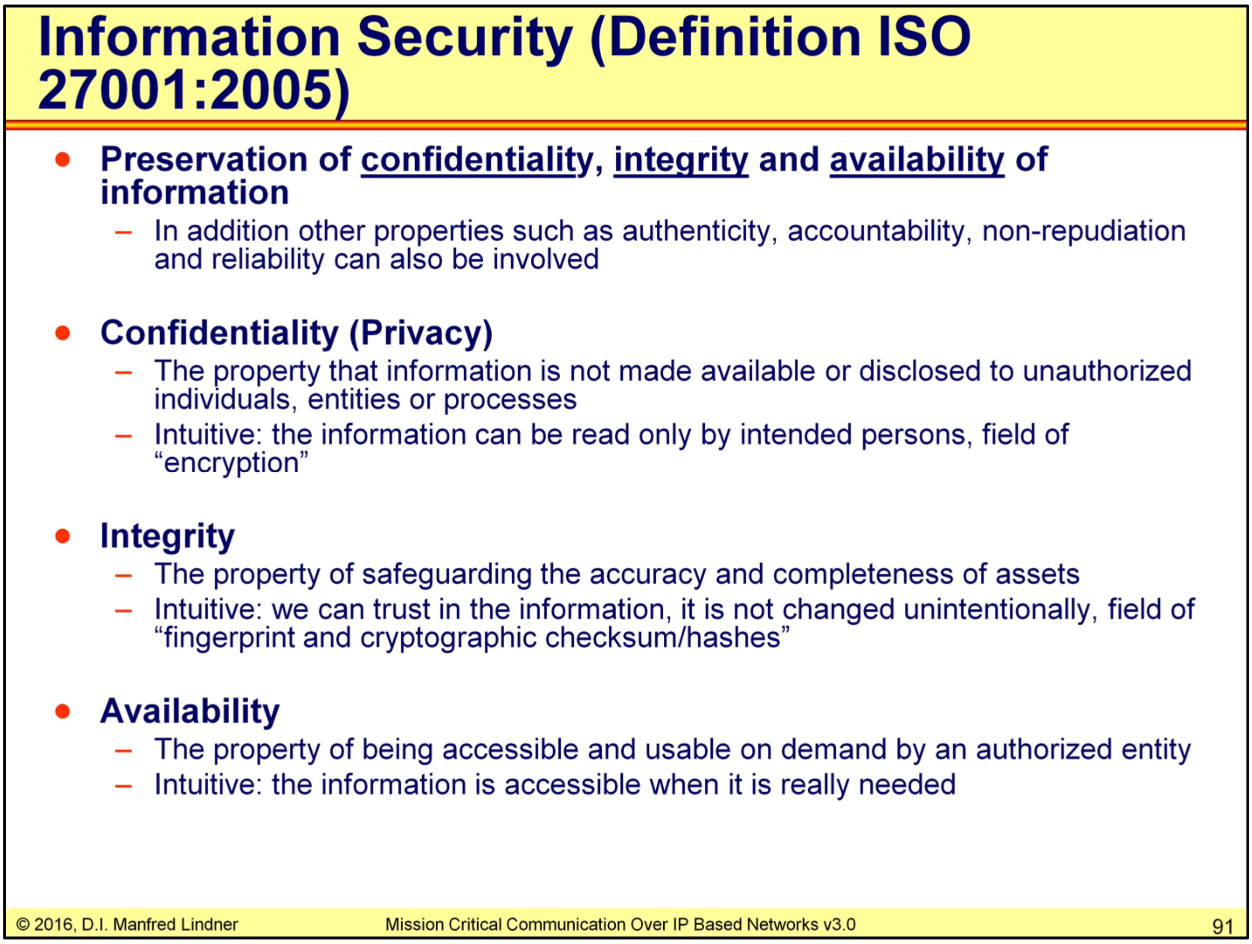
TM ....Traffic Marking TC ....Traffic Classifying TP ....Traffic Policing TQ ....Traffic Queuing TS ....Traffic Shaping

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 89



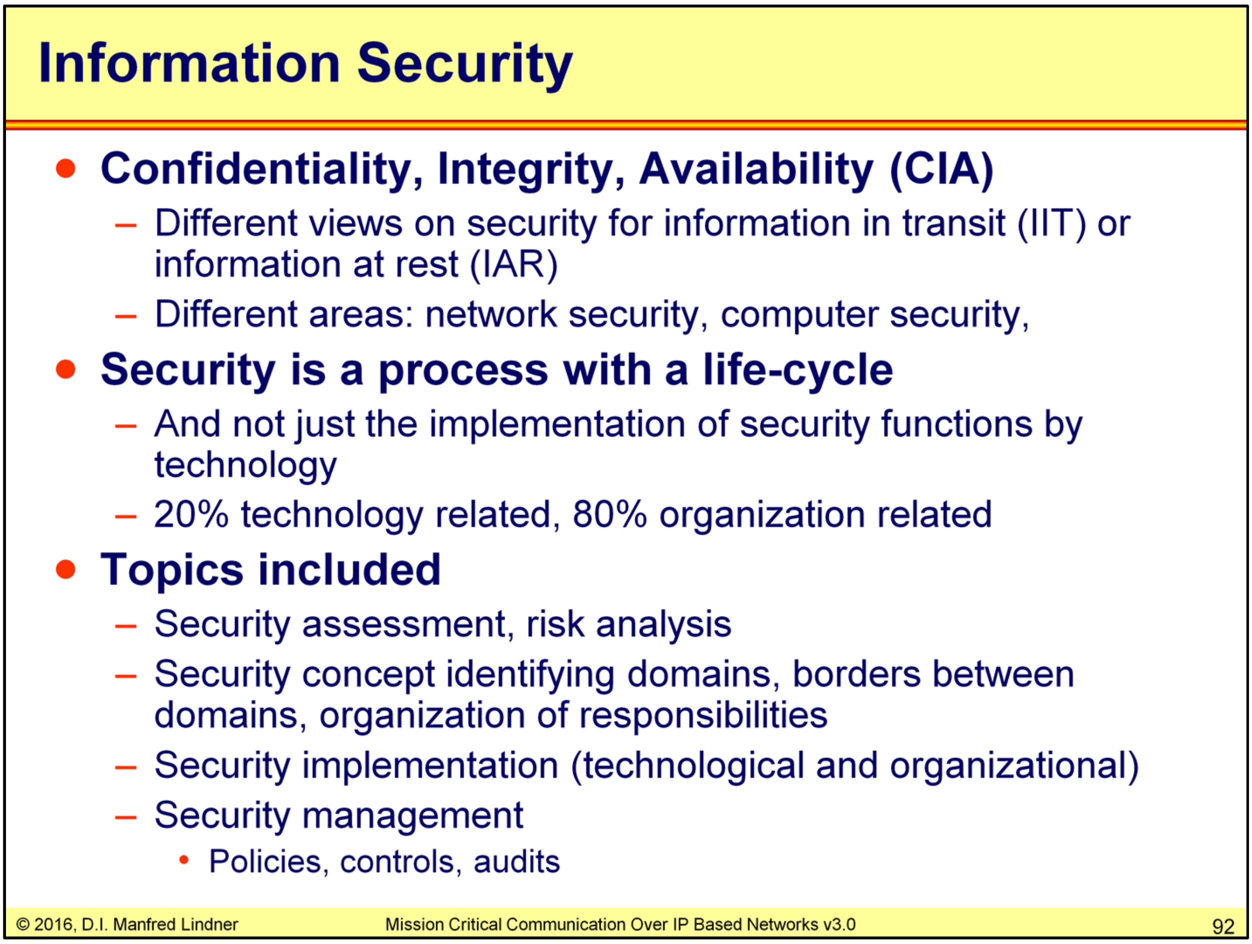
# Agenda

- Introduction
- Network Operational Model
- High Availability
- QoS
- VPN Technologies
  - Introduction IT-Security
  - VPN Types
  - MPLS, MPLS-VPN
  - IPsec VPN
  - DMVPN
  - GETVPN
- Multicasting
- Summary



# Information Security (Definition ISO 27001:2005)

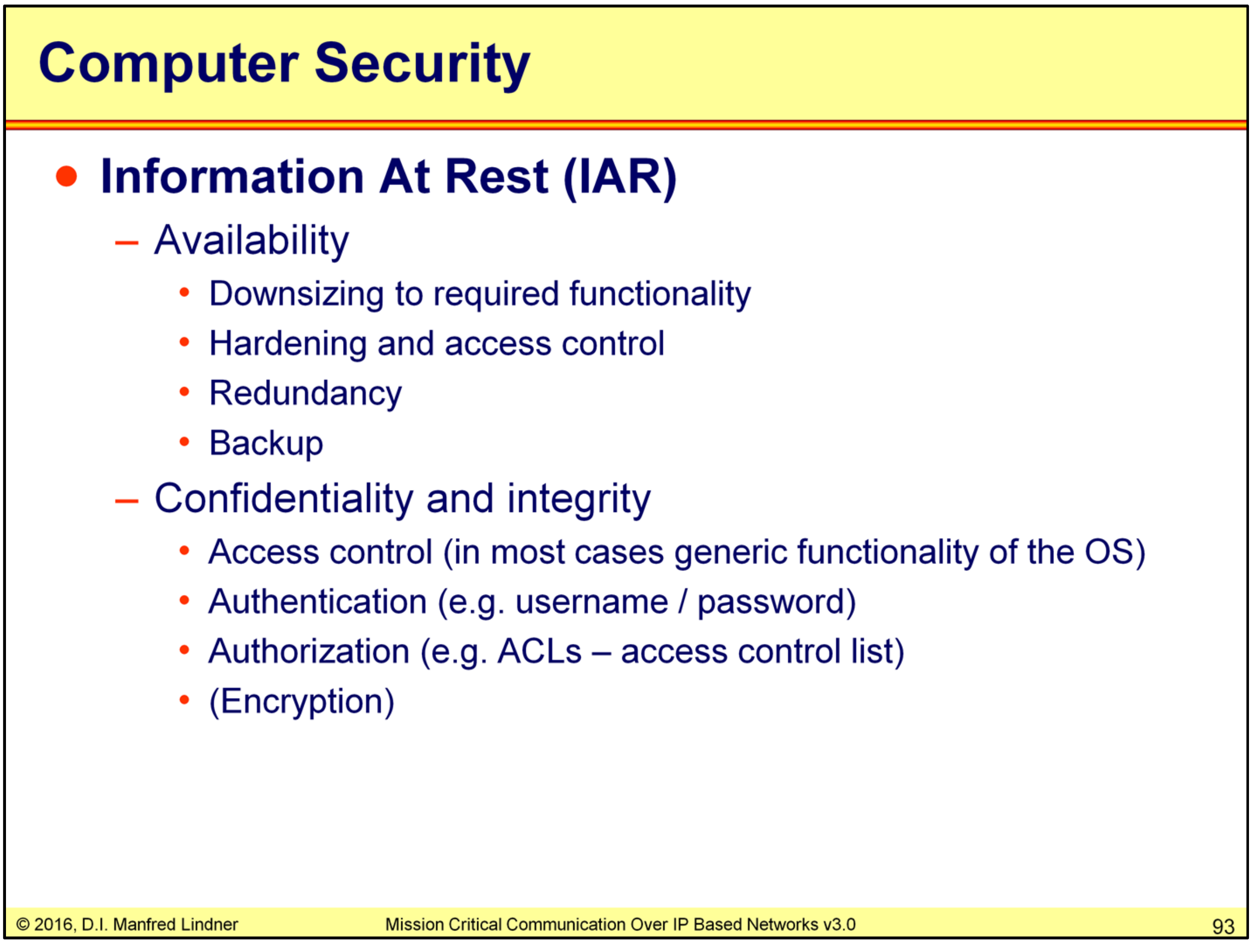
- **Preservation of confidentiality, integrity and availability of information**
  - In addition other properties such as authenticity, accountability, non-repudiation and reliability can also be involved
- **Confidentiality (Privacy)**
  - The property that information is not made available or disclosed to unauthorized individuals, entities or processes
  - Intuitive: the information can be read only by intended persons, field of “encryption”
- **Integrity**
  - The property of safeguarding the accuracy and completeness of assets
  - Intuitive: we can trust in the information, it is not changed unintentionally, field of “fingerprint and cryptographic checksum/hashes”
- **Availability**
  - The property of being accessible and usable on demand by an authorized entity
  - Intuitive: the information is accessible when it is really needed



## Information Security

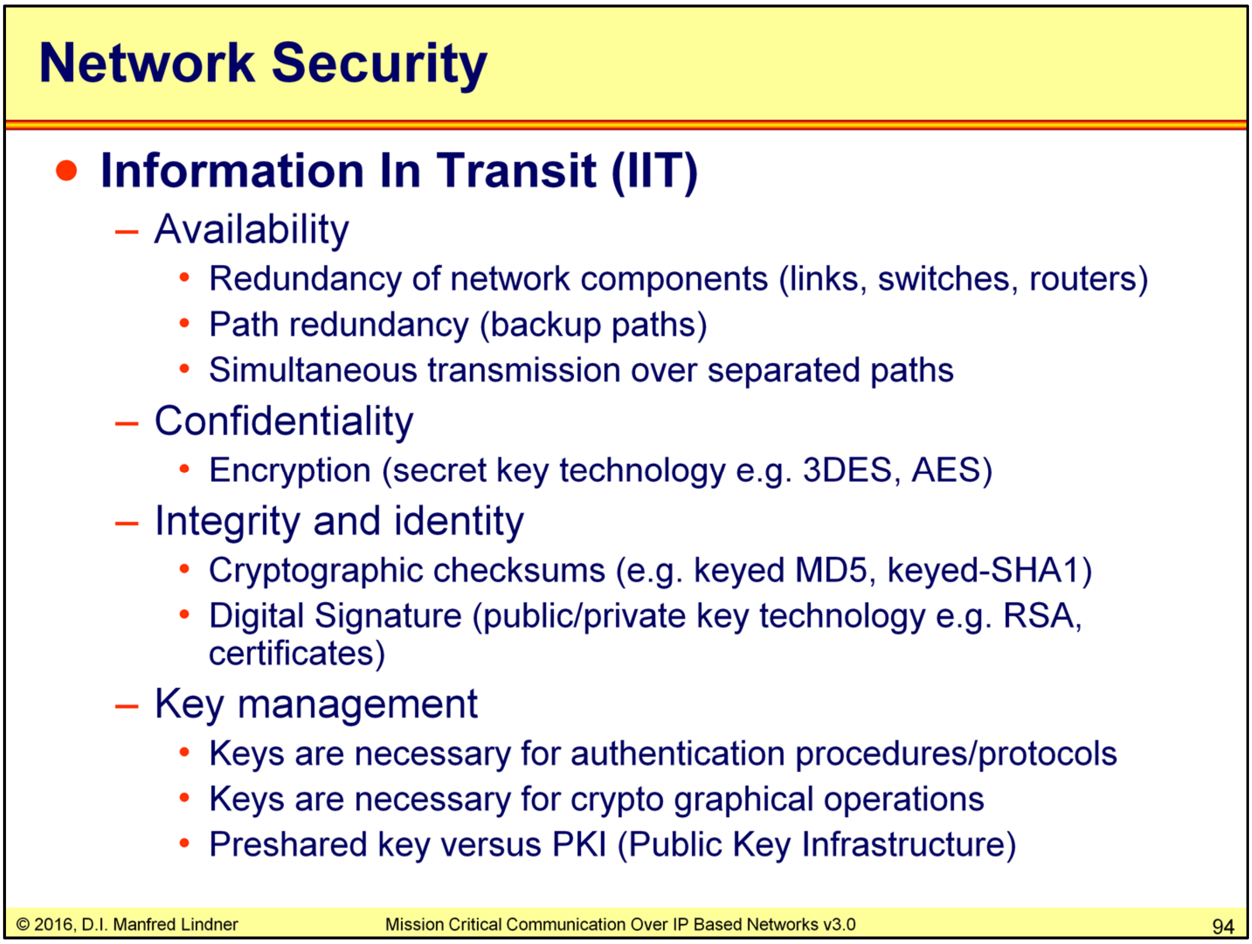
- Confidentiality, Integrity, Availability (CIA)
  - Different views on security for information in transit (IIT) or information at rest (IAR)
  - Different areas: network security, computer security,
- Security is a process with a life-cycle
  - And not just the implementation of security functions by technology
  - 20% technology related, 80% organization related
- Topics included
  - Security assessment, risk analysis
  - Security concept identifying domains, borders between domains, organization of responsibilities
  - Security implementation (technological and organizational)
  - Security management
    - Policies, controls, audits

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 92



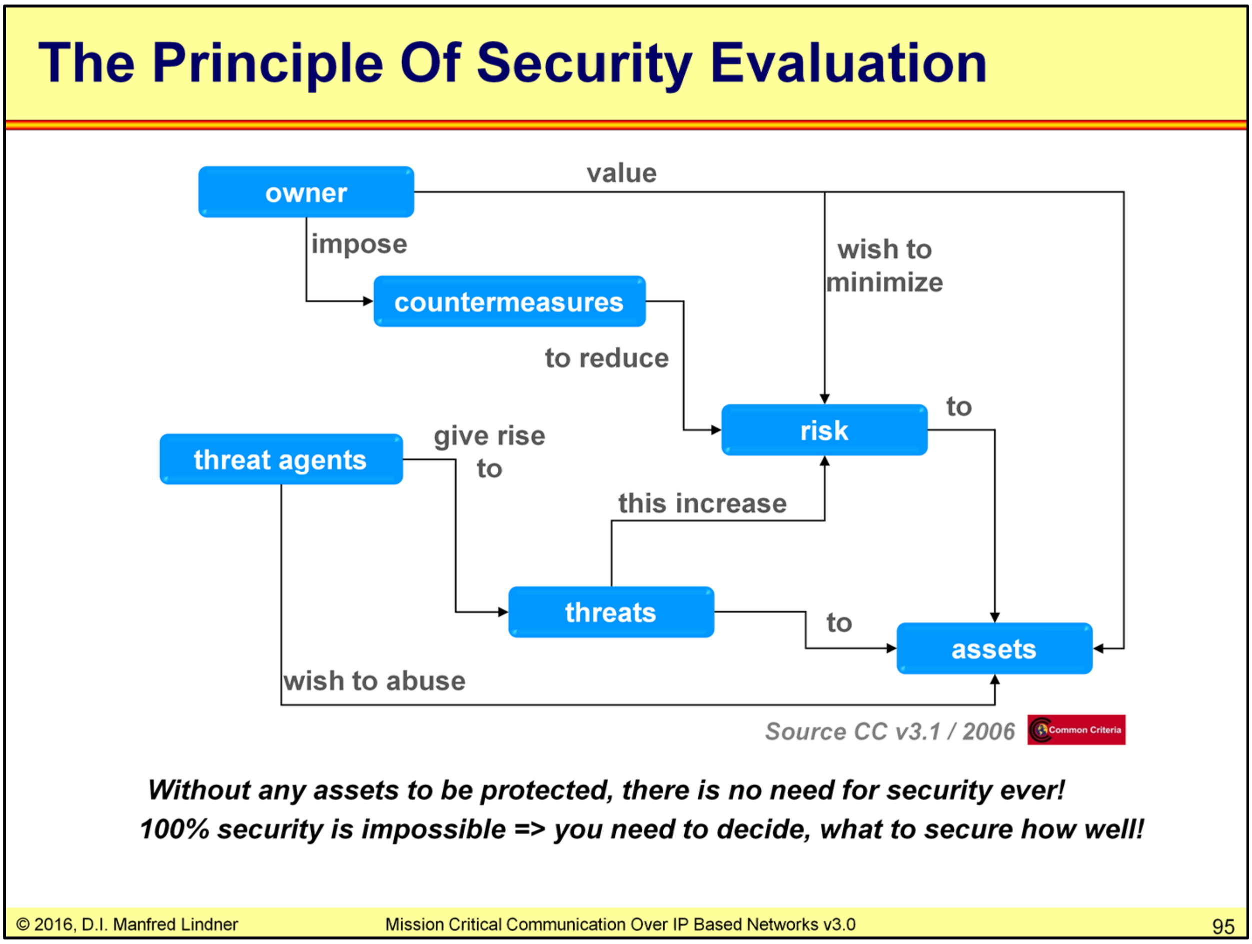
## Computer Security

- **Information At Rest (IAR)**
  - Availability
    - Downsizing to required functionality
    - Hardening and access control
    - Redundancy
    - Backup
  - Confidentiality and integrity
    - Access control (in most cases generic functionality of the OS)
    - Authentication (e.g. username / password)
    - Authorization (e.g. ACLs – access control list)
    - (Encryption)



# Network Security

- **Information In Transit (IIT)**
  - Availability
    - Redundancy of network components (links, switches, routers)
    - Path redundancy (backup paths)
    - Simultaneous transmission over separated paths
  - Confidentiality
    - Encryption (secret key technology e.g. 3DES, AES)
  - Integrity and identity
    - Cryptographic checksums (e.g. keyed MD5, keyed-SHA1)
    - Digital Signature (public/private key technology e.g. RSA, certificates)
  - Key management
    - Keys are necessary for authentication procedures/protocols
    - Keys are necessary for crypto graphical operations
    - Preshared key versus PKI (Public Key Infrastructure)



# The Principle Of Security Evaluation

owner

value

impose

countermeasures

wish to minimize

to reduce

risk

to

threat agents

give rise to

this increase

threats

to

assets

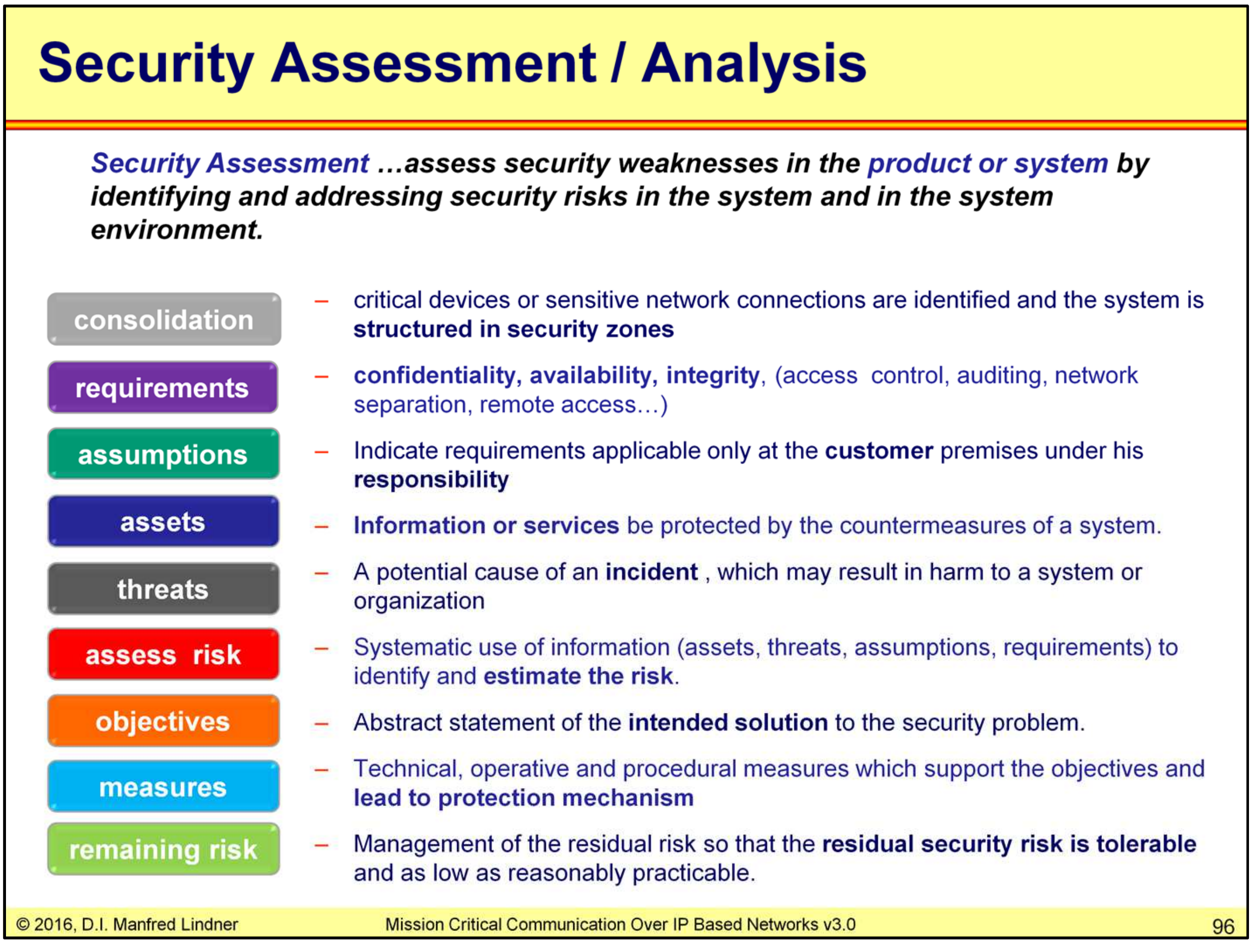
wish to abuse

*Source CC v3.1 / 2006*

Common Criteria

***Without any assets to be protected, there is no need for security ever!***

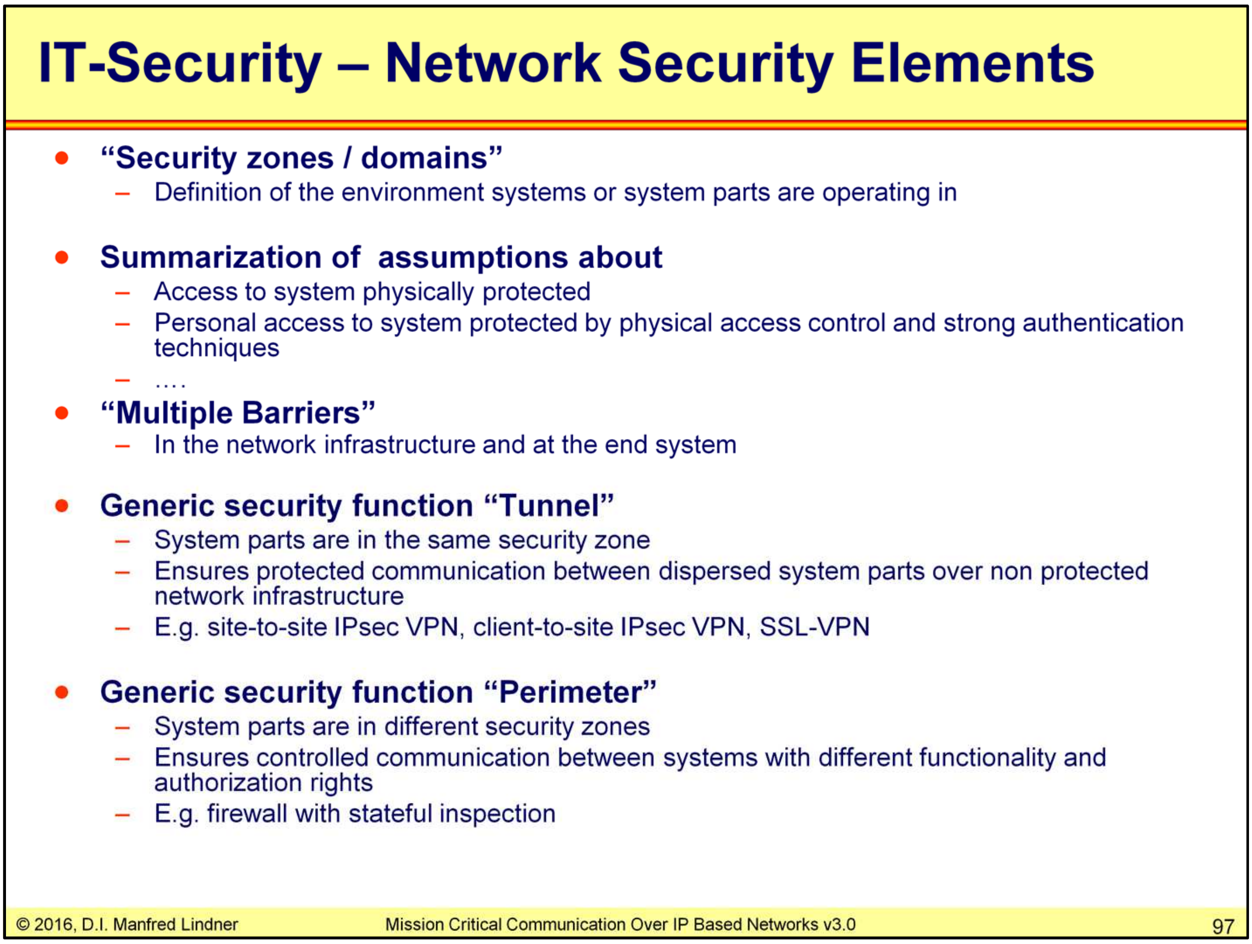
***100% security is impossible => you need to decide, what to secure how well!***



# Security Assessment / Analysis

***Security Assessment** …assess security weaknesses in the **product or system** by identifying and addressing security risks in the system and in the system environment.*

- consolidation – critical devices or sensitive network connections are identified and the system is **structured in security zones**
- requirements – **confidentiality, availability, integrity**, (access control, auditing, network separation, remote access…)
- assumptions – Indicate requirements applicable only at the **customer** premises under his **responsibility**
- assets – **Information or services** be protected by the countermeasures of a system.
- threats – A potential cause of an **incident** , which may result in harm to a system or organization
- assess risk – Systematic use of information (assets, threats, assumptions, requirements) to identify and **estimate the risk**.
- objectives – Abstract statement of the **intended solution** to the security problem.
- measures – Technical, operative and procedural measures which support the objectives and **lead to protection mechanism**
- remaining risk – Management of the residual risk so that the **residual security risk is tolerable** and as low as reasonably practicable.



## IT-Security – Network Security Elements

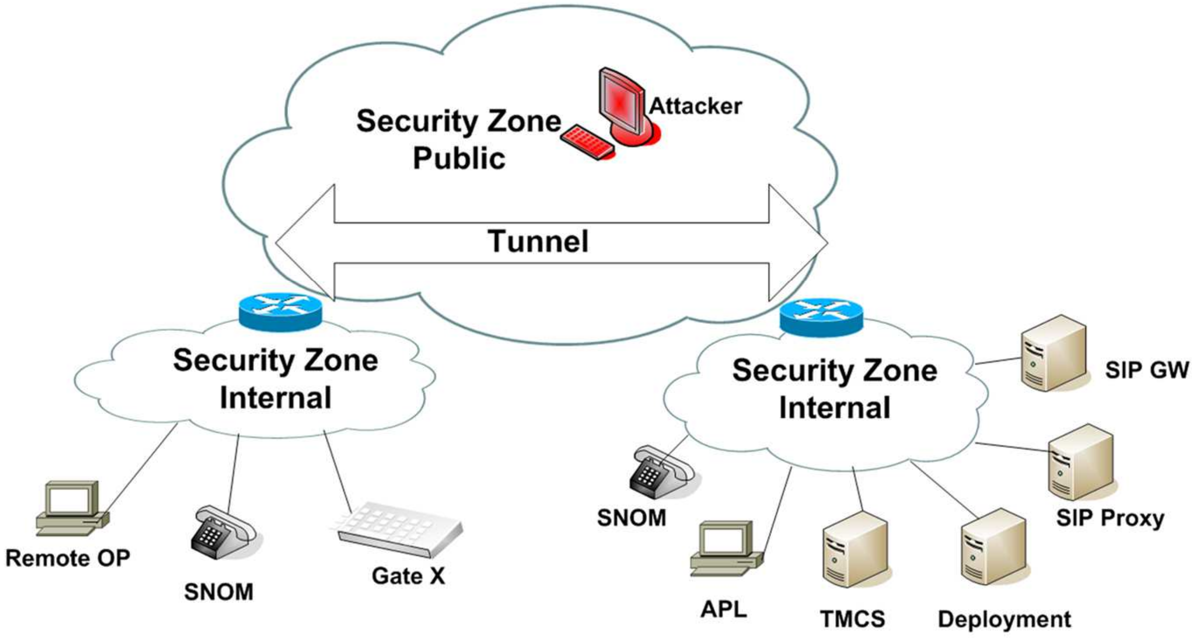
- “Security zones / domains”
  - Definition of the environment systems or system parts are operating in
- Summarization of  assumptions about
  - Access to system physically protected
  - Personal access to system protected by physical access control and strong authentication techniques
  - ….
- “Multiple Barriers”
  - In the network infrastructure and at the end system
- Generic security function “Tunnel”
  - System parts are in the same security zone
  - Ensures protected communication between dispersed system parts over non protected network infrastructure
  - E.g. site-to-site IPsec VPN, client-to-site IPsec VPN, SSL-VPN
- Generic security function “Perimeter”
  - System parts are in different security zones
  - Ensures controlled communication between systems with different functionality and authorization rights
  - E.g. firewall with stateful inspection

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 97

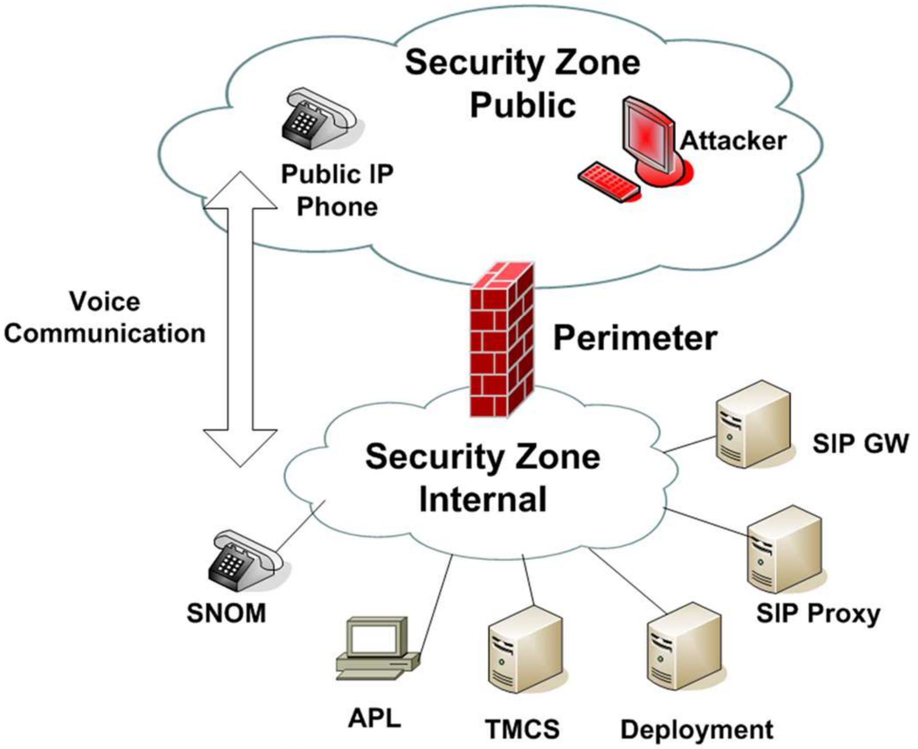
Security mechanisms can protect communication between systems (information in transit, IIT) and control access to systems hence protect information at rest (IAR). The former is performed by “tunneling” techniques which use crypto-graphical methods to ensure integrity and privacy of IP datagrams on transit. Examples for that are IPsec-VPN, SSL-VPN, and HTTPS. The later (IAR) is performed by “perimeter” mechanism which use policies based on filters and authentication mechanism to decide if traffic is allowed to pass the perimeter. Examples for that are stateful firewalls, white-lists, and access-lists typically in cooperation with security servers like radius in case of larger deployments.

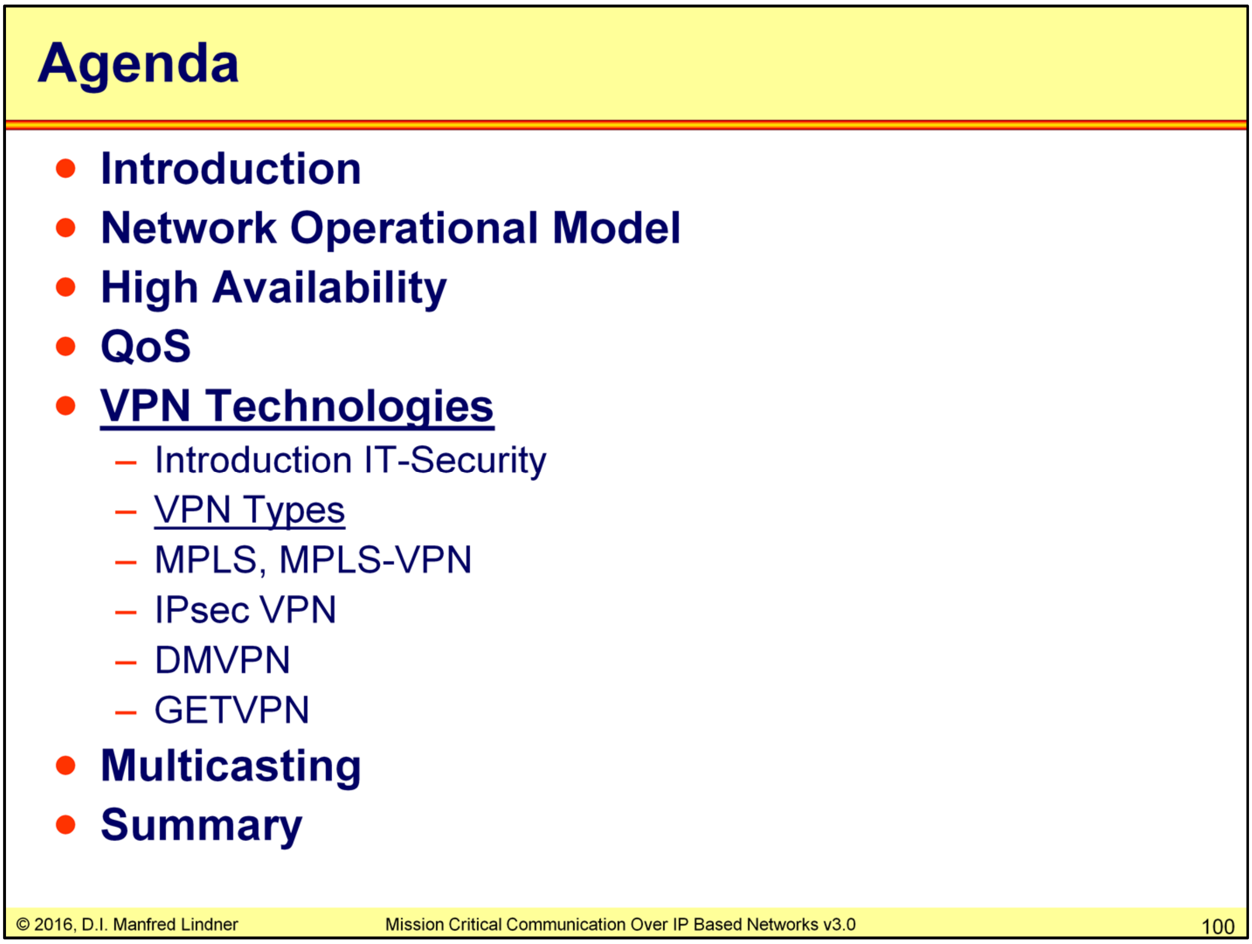
Based on security assessment and security policy typical topics to be covered by a security concept are clarification of security domains and the borders in between, clarification about organization of responsibilities, and finally agreement how security management should be established. Security implementation is always a combination of technological and procedural components. Only if all the initial work is done, finally the location of perimeter and tunnel mechanism can be concluded.

## Security Function Tunnel



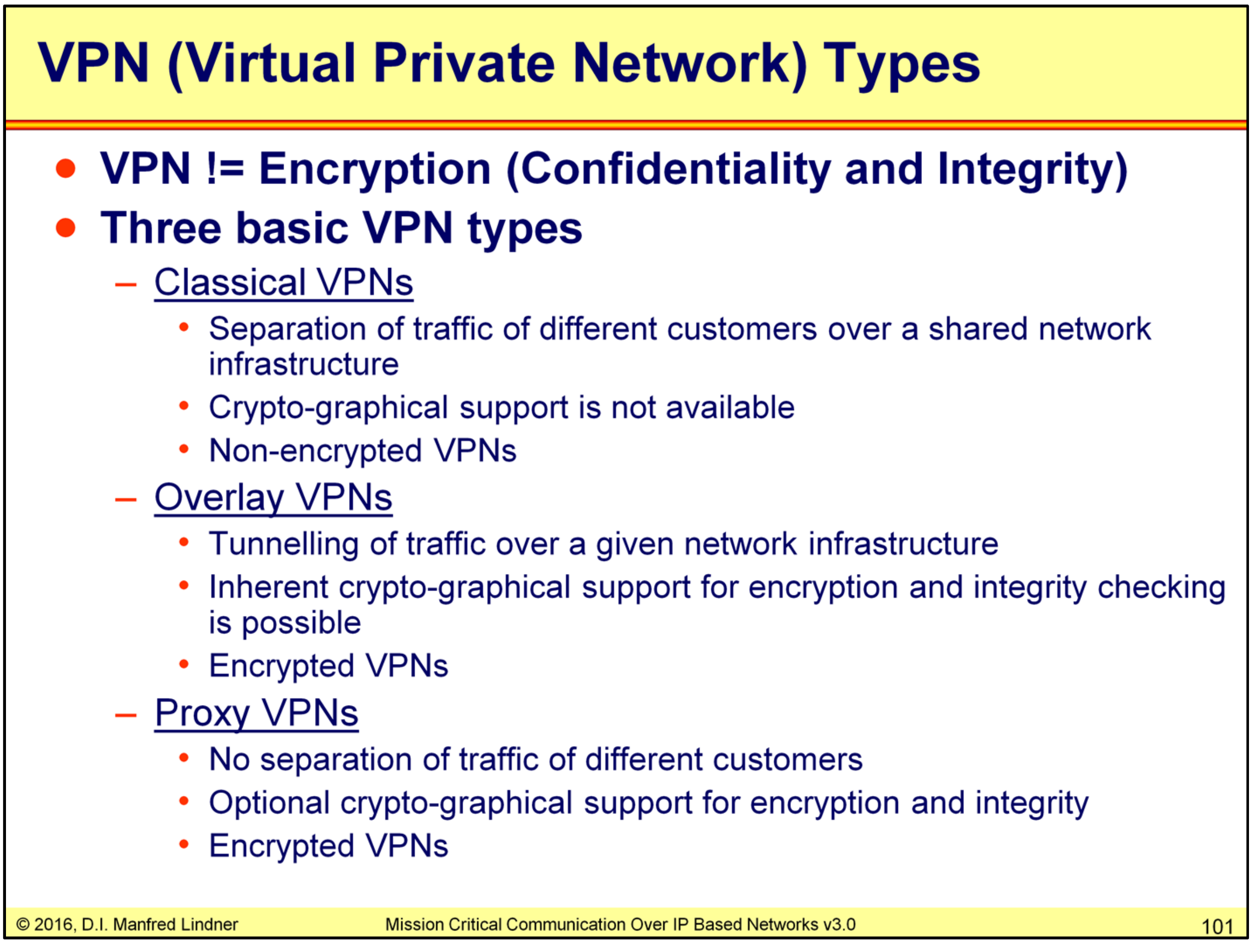
## Security Function Perimeter





## Agenda

- Introduction
- Network Operational Model
- High Availability
- QoS
- VPN Technologies
  - Introduction IT-Security
  - VPN Types
  - MPLS, MPLS-VPN
  - IPsec VPN
  - DMVPN
  - GETVPN
- Multicasting
- Summary



## VPN (Virtual Private Network) Types

- **VPN != Encryption (Confidentiality and Integrity)**
- **Three basic VPN types**
  - Classical VPNs
    - Separation of traffic of different customers over a shared network infrastructure
    - Crypto-graphical support is not available
    - Non-encrypted VPNs
  - Overlay VPNs
    - Tunnelling of traffic over a given network infrastructure
    - Inherent crypto-graphical support for encryption and integrity checking is possible
    - Encrypted VPNs
  - Proxy VPNs
    - No separation of traffic of different customers
    - Optional crypto-graphical support for encryption and integrity
    - Encrypted VPNs

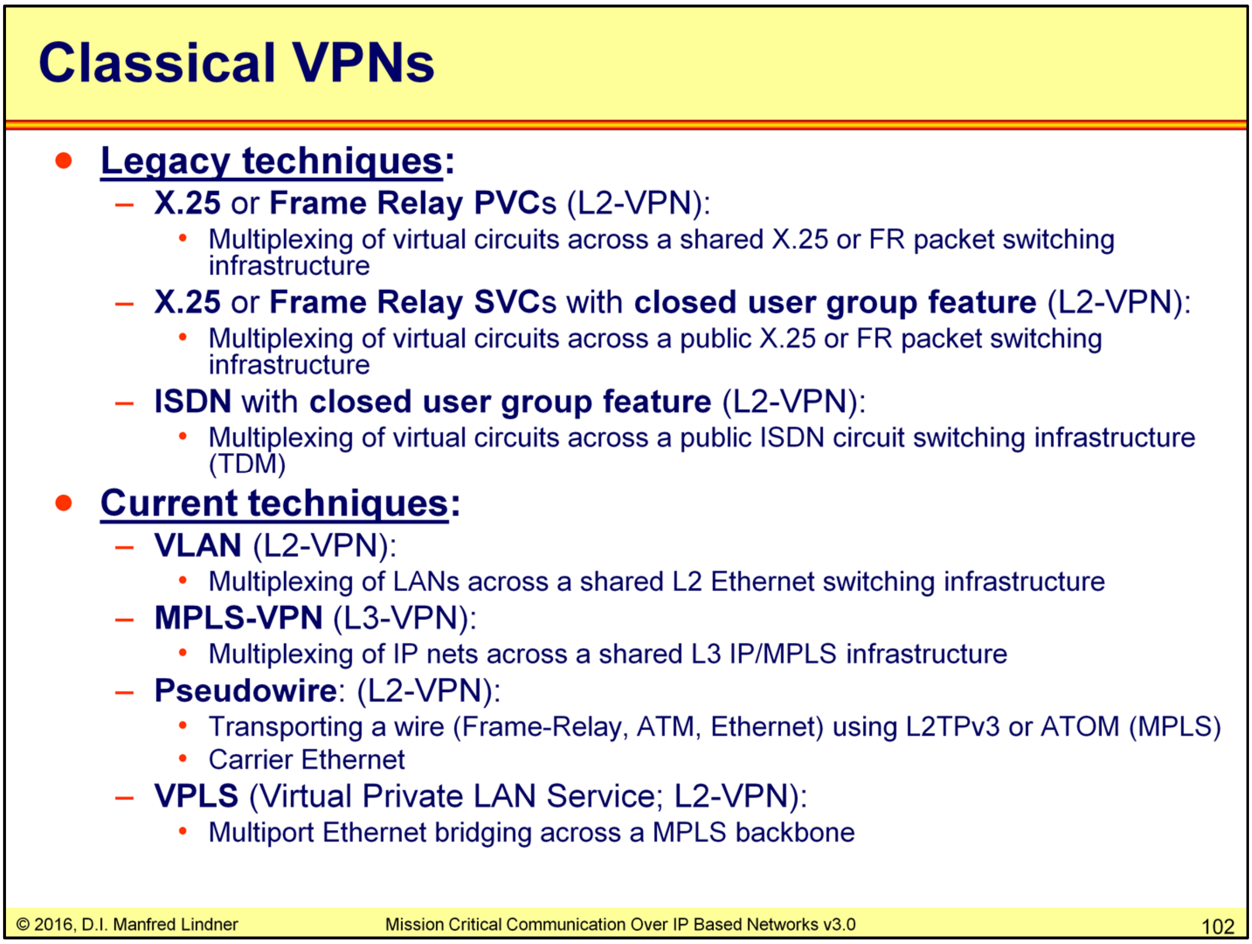
© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 101

It is important to know that VPN does not automatically include support for encryption and integrity.

Classical VPN techniques like Frame Relay, VLAN, MPLS-VPN, VPLS, and LISP-VPN concentrates just on separation of traffic of different customers over a shared infrastructure.

Overlay VPNs like IPsec, SSL-VPN, and DMVPN have inherent crypto-graphical support for encryption and integrity checking. They can be seen as security tunnels between security gateways or between end systems and may operate either over a shared environment or on top of a classical VPN. Typically based on a point-to-point security relationship these techniques have scalability and performance issues (number of relationships, keys, states, key management).

With GETVPN based on IPsec with group encryption (point-to-multipoint security relationship) these scalabilities problems can be solved. Therefore GETVPN as tunnel-less technology is also called Proxy VPN doing neither traffic separation nor using a tunnel but still achieve privacy/integrity for traffic in transit.



## Classical VPNs

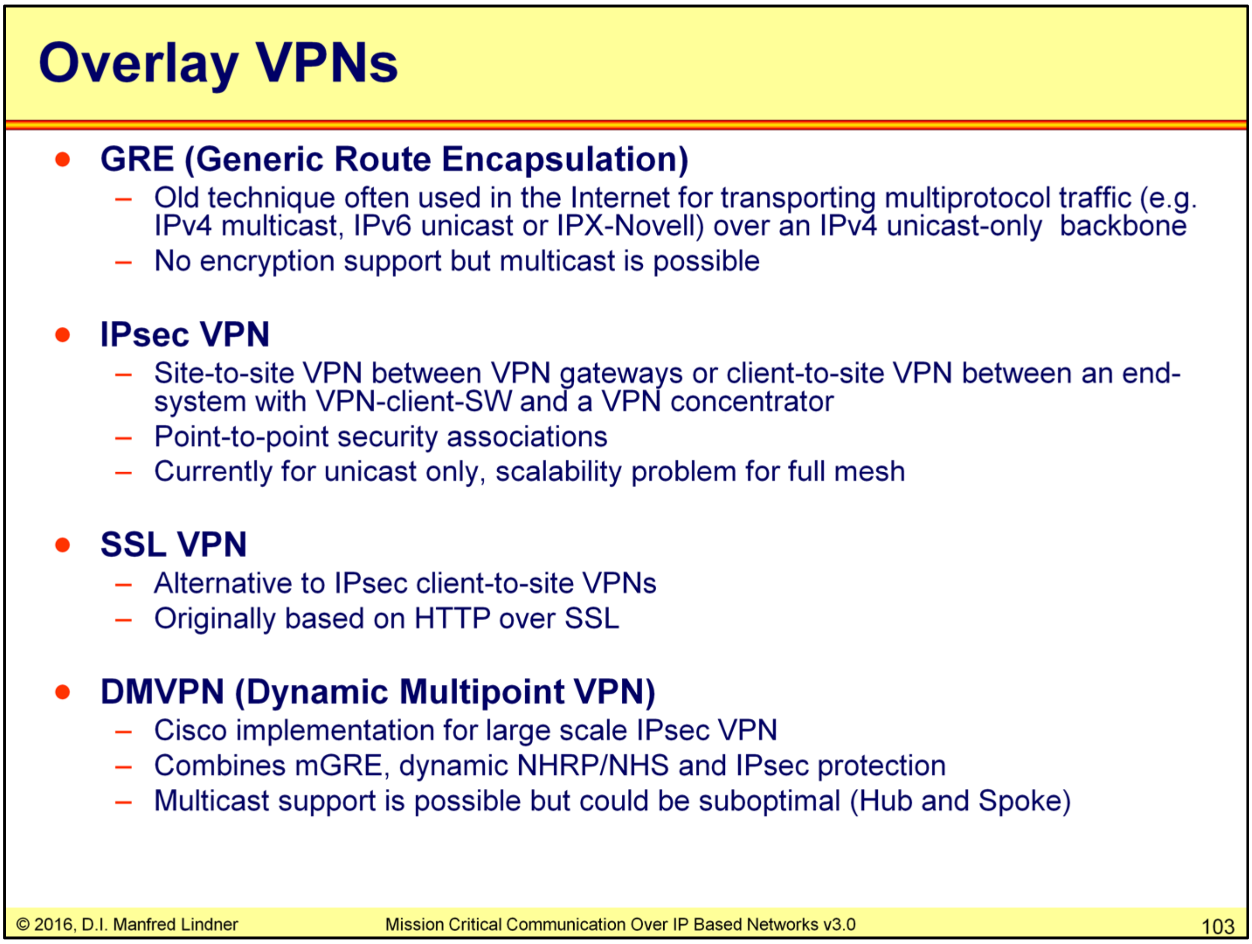
- **Legacy techniques:**
  - **X.25** or **Frame Relay PVCs** (L2-VPN):
    - Multiplexing of virtual circuits across a shared X.25 or FR packet switching infrastructure
  - **X.25** or **Frame Relay SVCs** with **closed user group feature** (L2-VPN):
    - Multiplexing of virtual circuits across a public X.25 or FR packet switching infrastructure
  - **ISDN** with **closed user group feature** (L2-VPN):
    - Multiplexing of virtual circuits across a public ISDN circuit switching infrastructure (TDM)
- **Current techniques:**
  - **VLAN** (L2-VPN):
    - Multiplexing of LANs across a shared L2 Ethernet switching infrastructure
  - **MPLS-VPN** (L3-VPN):
    - Multiplexing of IP nets across a shared L3 IP/MPLS infrastructure
  - **Pseudowire**: (L2-VPN):
    - Transporting a wire (Frame-Relay, ATM, Ethernet) using L2TPv3 or ATOM (MPLS)
    - Carrier Ethernet
  - **VPLS** (Virtual Private LAN Service; L2-VPN):
    - Multiport Ethernet bridging across a MPLS backbone

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 102

SVC … Switched Virtual Circuit

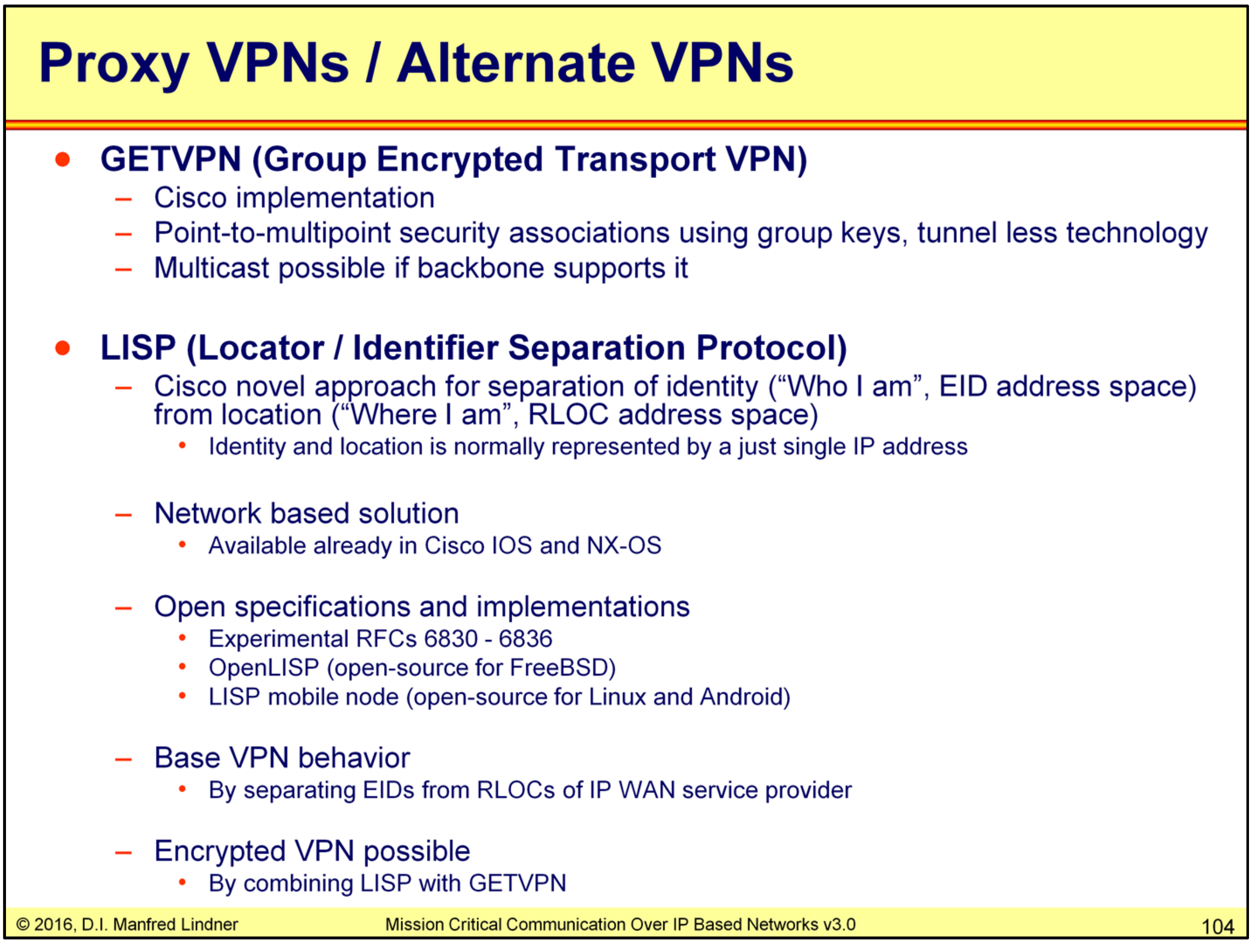
PVC … Permanent Virtual Circuit

ATOM … Any Transport Over MPLS



# Overlay VPNs

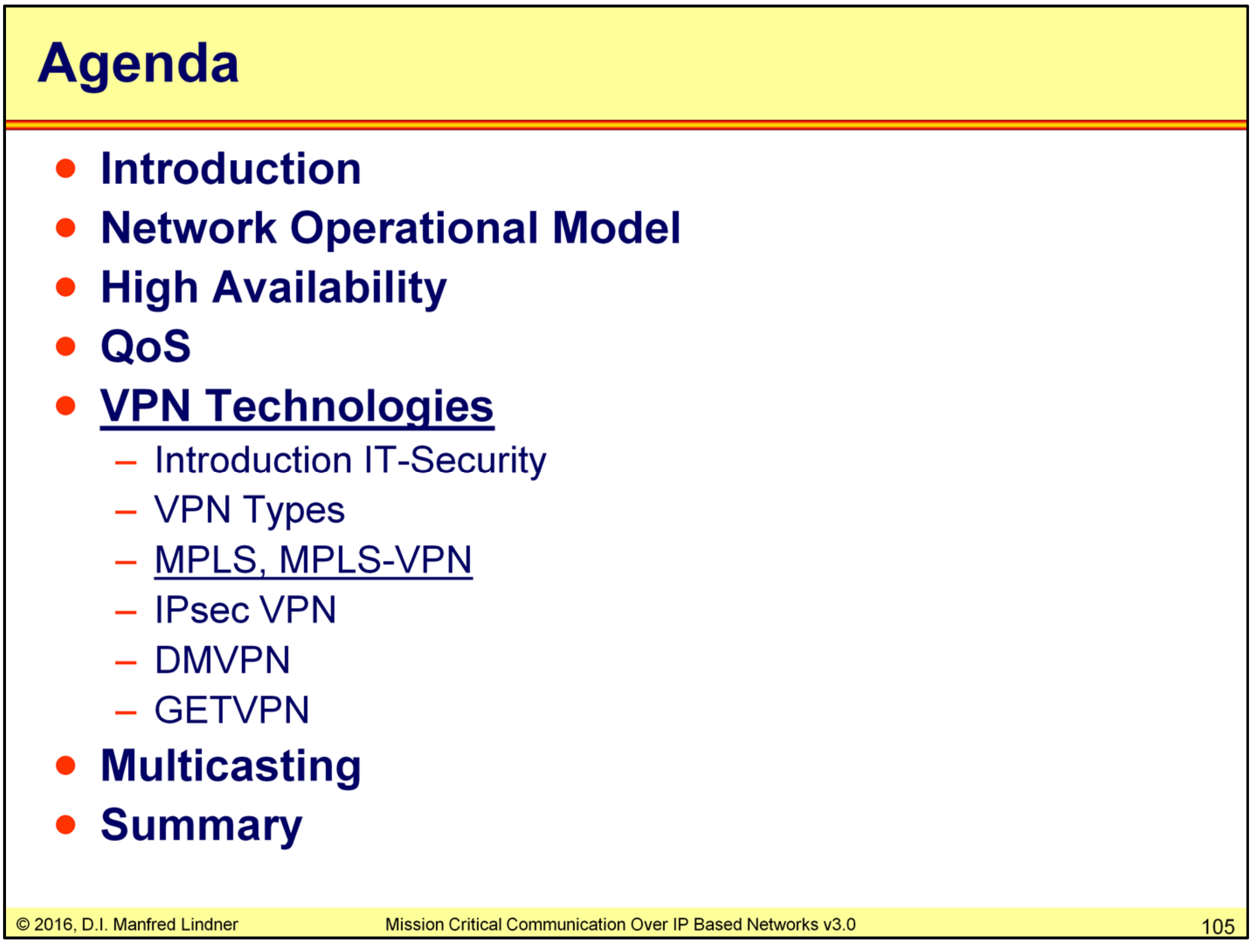
- **GRE (Generic Route Encapsulation)**
  - Old technique often used in the Internet for transporting multiprotocol traffic (e.g. IPv4 multicast, IPv6 unicast or IPX-Novell) over an IPv4 unicast-only backbone
  - No encryption support but multicast is possible
- **IPsec VPN**
  - Site-to-site VPN between VPN gateways or client-to-site VPN between an end-system with VPN-client-SW and a VPN concentrator
  - Point-to-point security associations
  - Currently for unicast only, scalability problem for full mesh
- **SSL VPN**
  - Alternative to IPsec client-to-site VPNs
  - Originally based on HTTP over SSL
- **DMVPN (Dynamic Multipoint VPN)**
  - Cisco implementation for large scale IPsec VPN
  - Combines mGRE, dynamic NHRP/NHS and IPsec protection
  - Multicast support is possible but could be suboptimal (Hub and Spoke)



## Proxy VPNs / Alternate VPNs

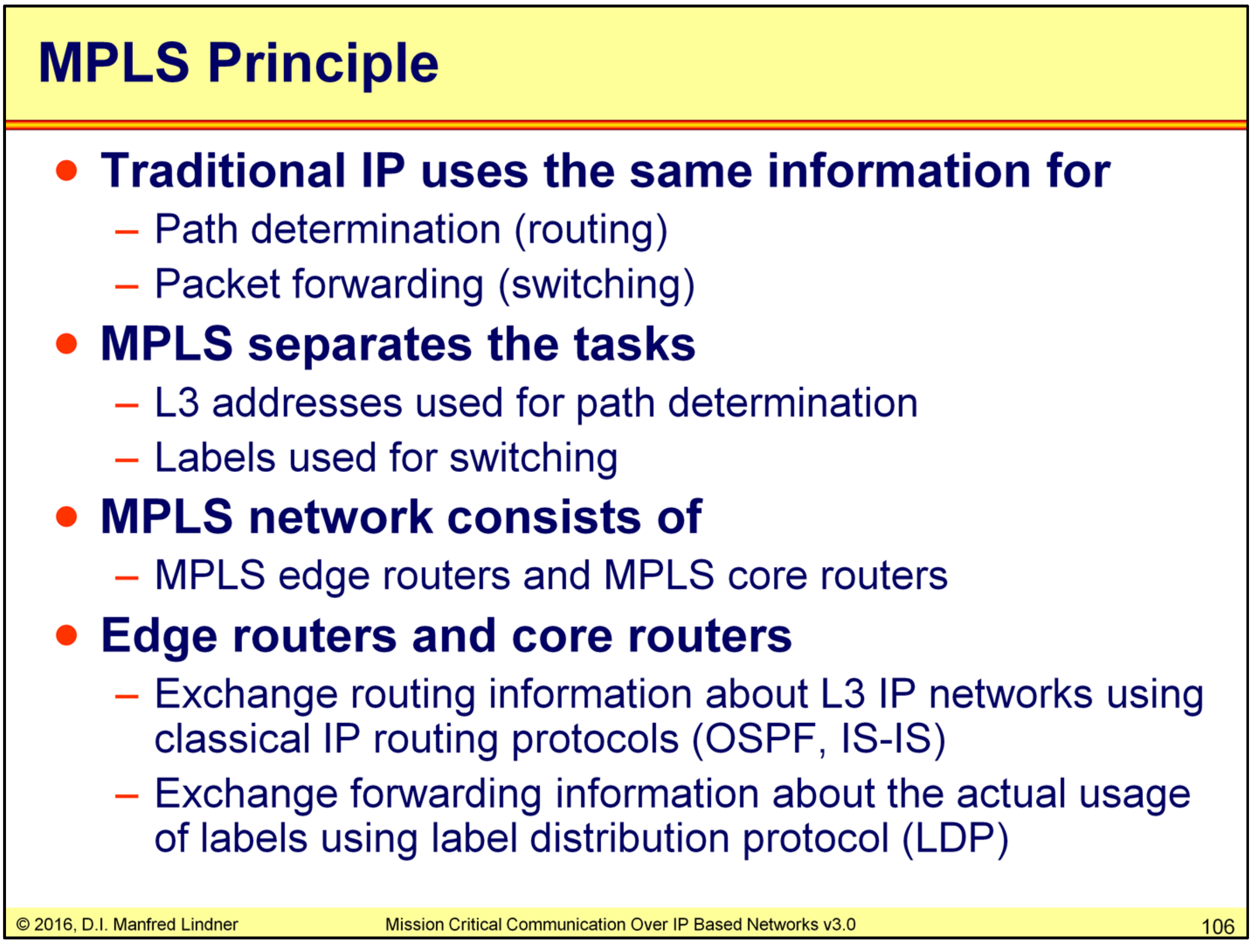
- **GETVPN (Group Encrypted Transport VPN)**
  - Cisco implementation
  - Point-to-multipoint security associations using group keys, tunnel less technology
  - Multicast possible if backbone supports it

- **LISP (Locator / Identifier Separation Protocol)**
  - Cisco novel approach for separation of identity (“Who I am”, EID address space) from location (“Where I am”, RLOC address space)
    - Identity and location is normally represented by a just single IP address
  - Network based solution
    - Available already in Cisco IOS and NX-OS
  - Open specifications and implementations
    - Experimental RFCs 6830 - 6836
    - OpenLISP (open-source for FreeBSD)
    - LISP mobile node (open-source for Linux and Android)
  - Base VPN behavior
    - By separating EIDs from RLOCs of IP WAN service provider
  - Encrypted VPN possible
    - By combining LISP with GETVPN



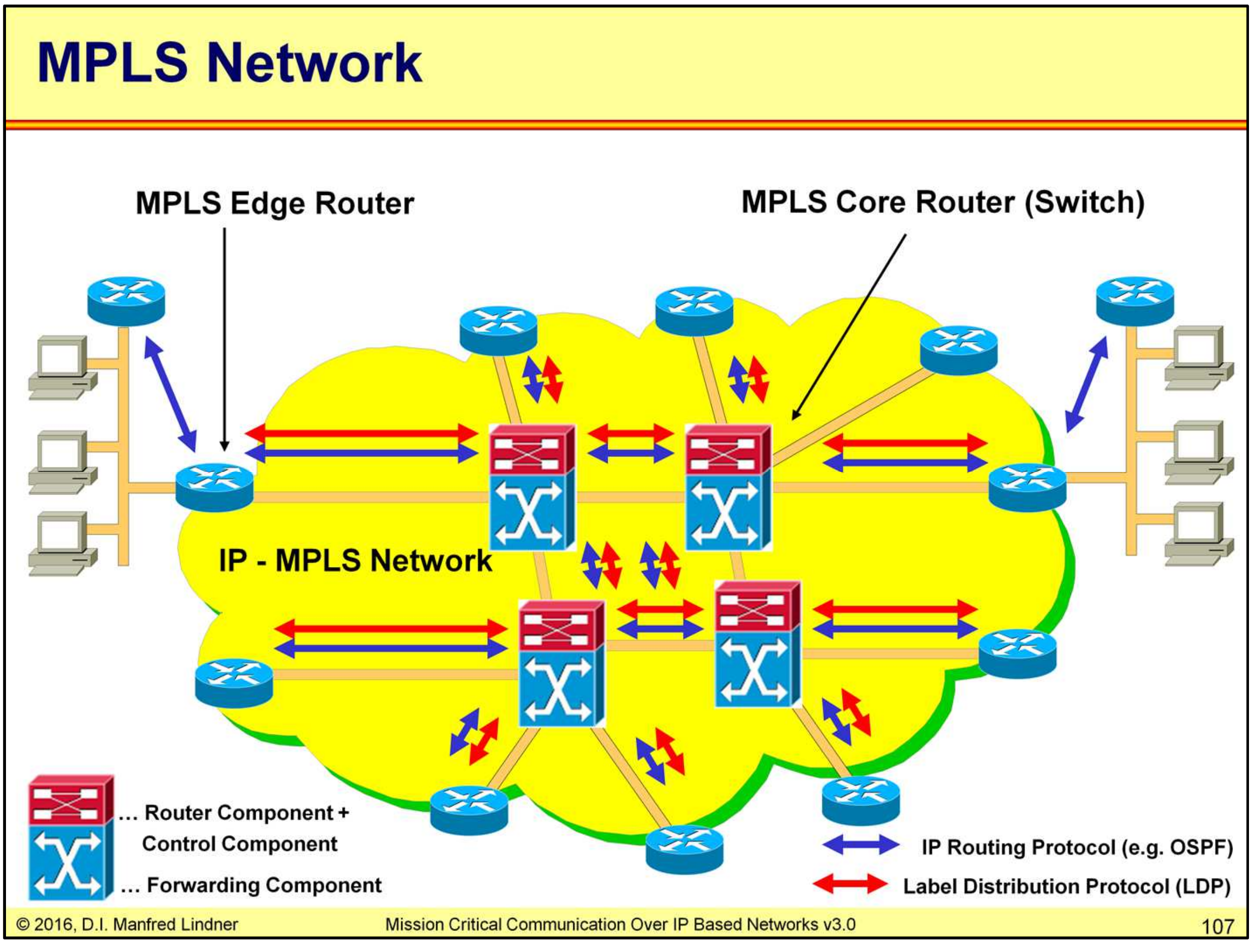
## Agenda

- Introduction
- Network Operational Model
- High Availability
- QoS
- VPN Technologies
  - Introduction IT-Security
  - VPN Types
  - MPLS, MPLS-VPN
  - IPsec VPN
  - DMVPN
  - GETVPN
- Multicasting
- Summary



## MPLS Principle

- Traditional IP uses the same information for
  - Path determination (routing)
  - Packet forwarding (switching)
- MPLS separates the tasks
  - L3 addresses used for path determination
  - Labels used for switching
- MPLS network consists of
  - MPLS edge routers and MPLS core routers
- Edge routers and core routers
  - Exchange routing information about L3 IP networks using classical IP routing protocols (OSPF, IS-IS)
  - Exchange forwarding information about the actual usage of labels using label distribution protocol (LDP)



## MPLS Network

© 2016, D.I. Manfred Lindner    Mission Critical Communication Over IP Based Networks v3.0    107

Routing component:

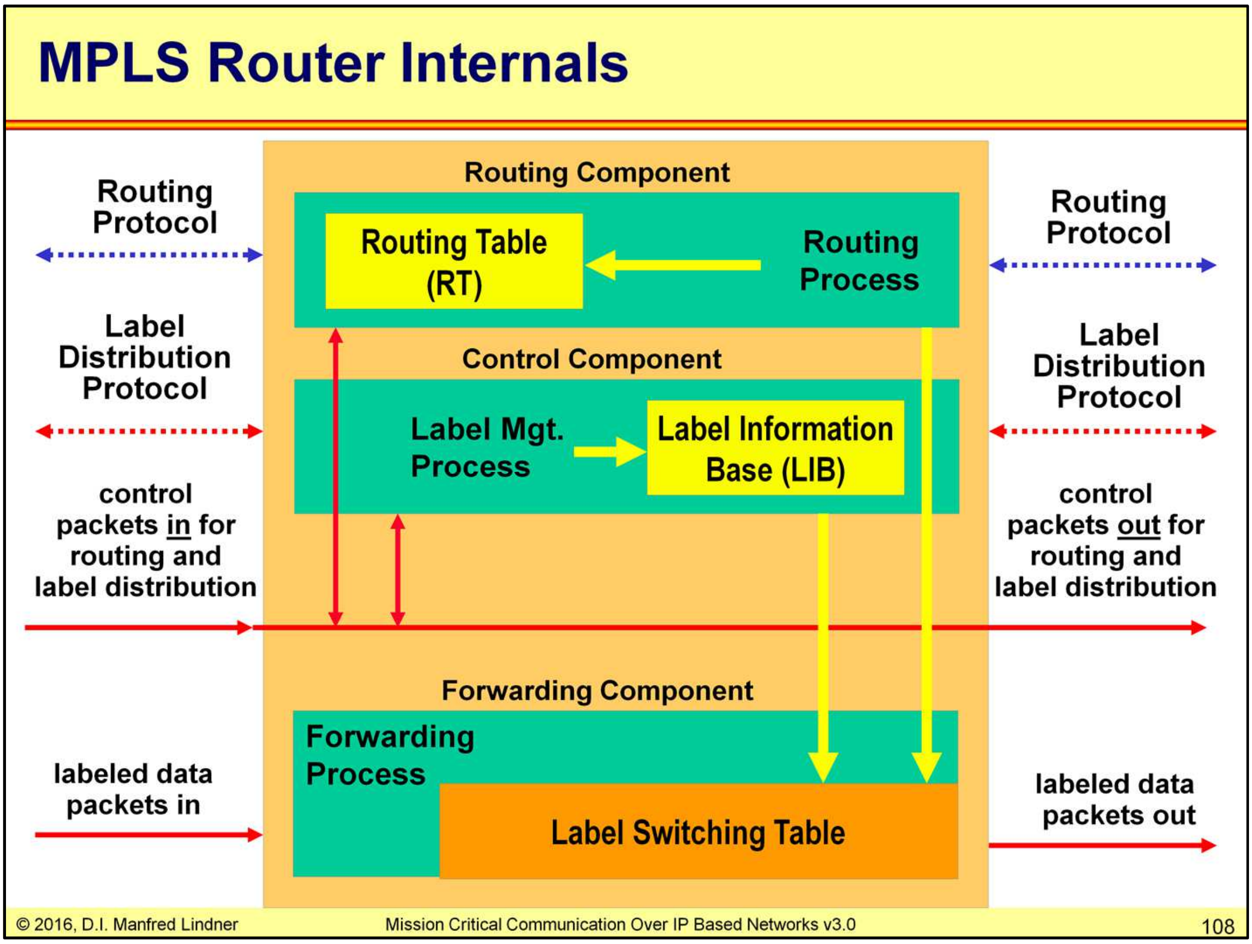
Still accomplishes traditional IP routing creating a classical IP routing table.

Control component:

Maintains correct label distribution among a group of label switched routers using LDP (Label Distribution Protocol) for communication among MPLS core routers and between MPLS core routers and MPLS edge routers. Information about labels is stored in the label information base (LIB)

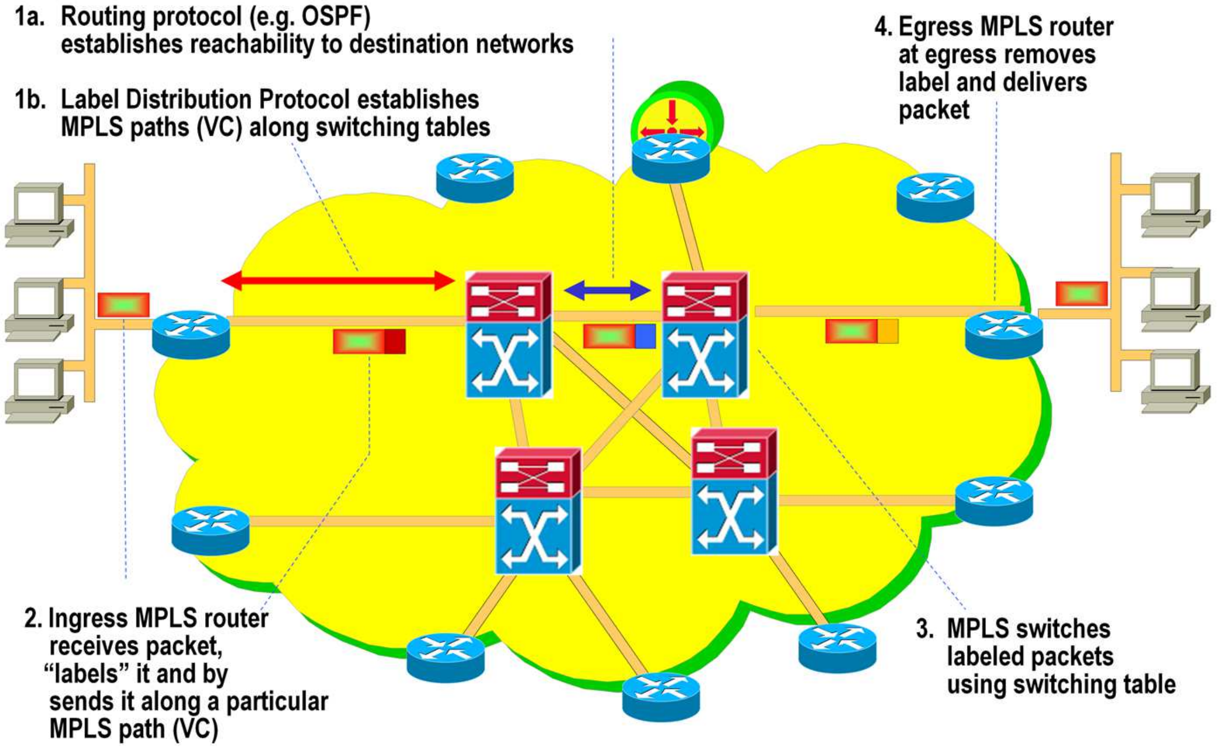
Forwarding component:

Uses label carried in the packet header together with a  label switching table maintained by MPLS Router to perform packet switching/forwarding in a classical VC (virtual circuit) switching style similar to ATM (Asynchronous Transfer Mode). Label swapping is done when a MPLS packet is forwarded by a MPLS router. The difference to ATM is that label switching table is based on the state of the IP routing table whereas an ATM switching table is maintained by explicit connection setup done by a signaling protocol.



## MPLS Router Internals

## MPLS Label Swapping



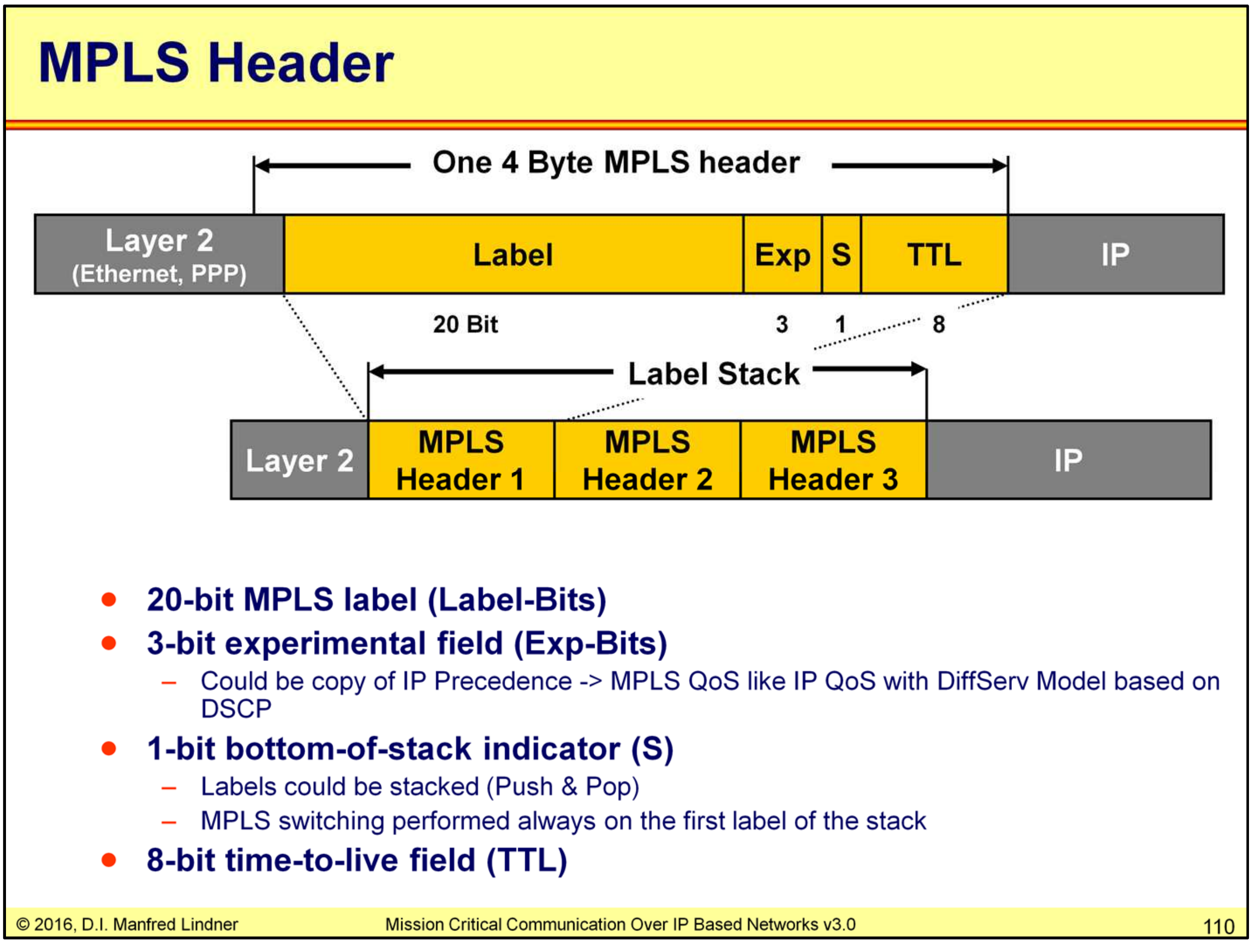
1a. Routing protocol (e.g. OSPF) establishes reachability to destination networks

1b. Label Distribution Protocol establishes MPLS paths (VC) along switching tables

2. Ingress MPLS router receives packet, “labels” it and by sends it along a particular MPLS path (VC)

3. MPLS switches labeled packets using switching table

4. Egress MPLS router at egress removes label and delivers packet



## MPLS Header

- **20-bit MPLS label (Label-Bits)**
- **3-bit experimental field (Exp-Bits)**
  - Could be copy of IP Precedence -> MPLS QoS like IP QoS with DiffServ Model based on DSCP
- **1-bit bottom-of-stack indicator (S)**
  - Labels could be stacked (Push & Pop)
  - MPLS switching performed always on the first label of the stack
- **8-bit time-to-live field (TTL)**

The MPLS Header is made up of four bytes and is located between the L2 header and the L3 header. The existence of an MPLS header is indicated by the layer two type field entry 0x8848.

The MPLS header is made up of a:

20 bit label field used for forwarding,

3 Experimental bits typically used to carry IP Precedence (IP TOS) settings,

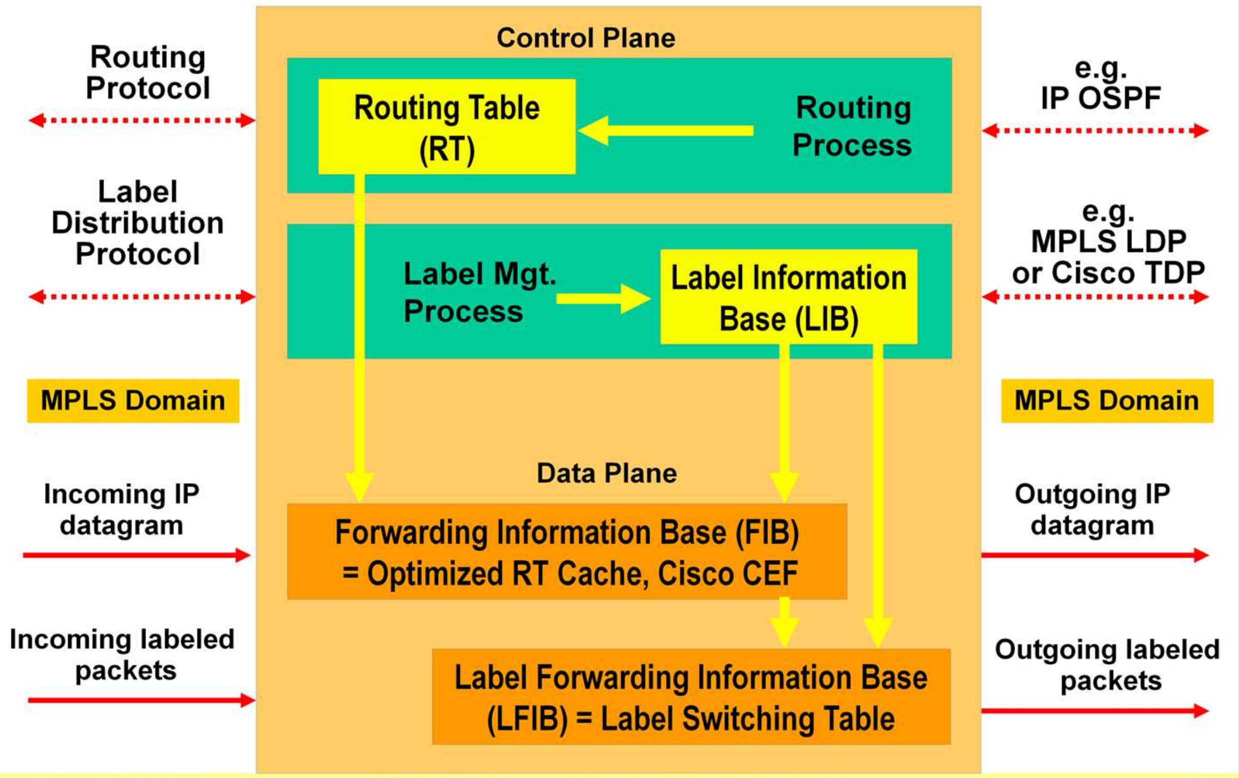
1 bit bottom of stack (0 indicates last label in the stack, 1 indicates there are some more labels on top of the bottom label)

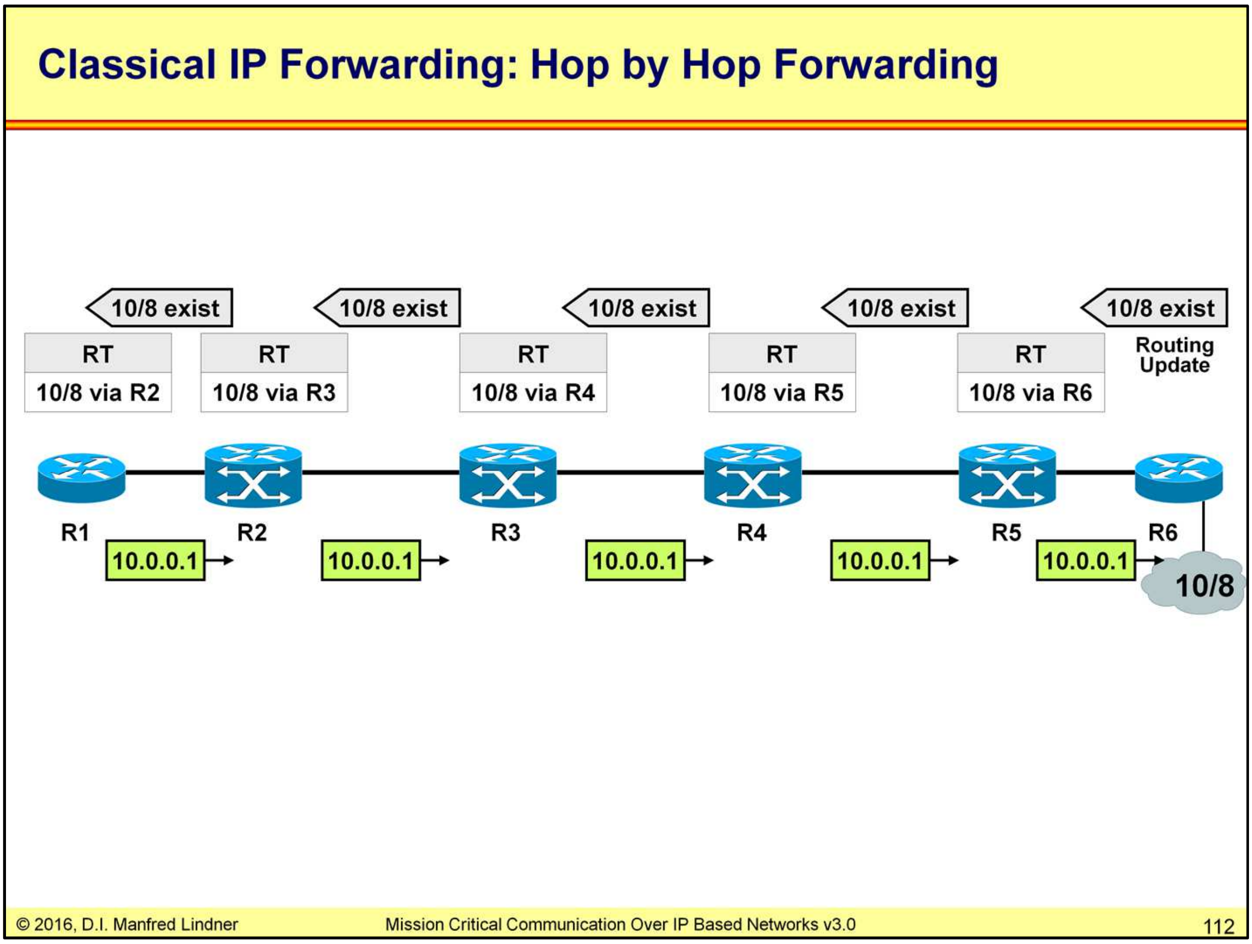
TTL field in which by default the IP TTL value is copied to when a label is inserted.

Note: The labels 0 to 15 are reserved. Therefore the lowest usable label number is 16 and the highest possible label is 1,048,575 (which is actually 2^20-1). Only four out of the 16 reserved labels had been defined (RFC 3032), which are: 0 "IPv4 Explicit Null Label", 1 "Router Alert Label", 2 "IPv6 Explicit Null Label", 3 "Implicit Null Label".

Several reasons lead to a label stack. For example, with MPLS VPNs, the top label identifies the egress router while a second label identifies the VPN itself. Thus the egress router can (as soon as the packet arrived) pop the outermost label and forward the packet to the right interface according to the inner label. Another example is MPLS Traffic Engineering (TE), where the outer label points to the TE tunnel endpoint and the inner label to the final destination itself.

## MPLS Router Internals (Cisco)

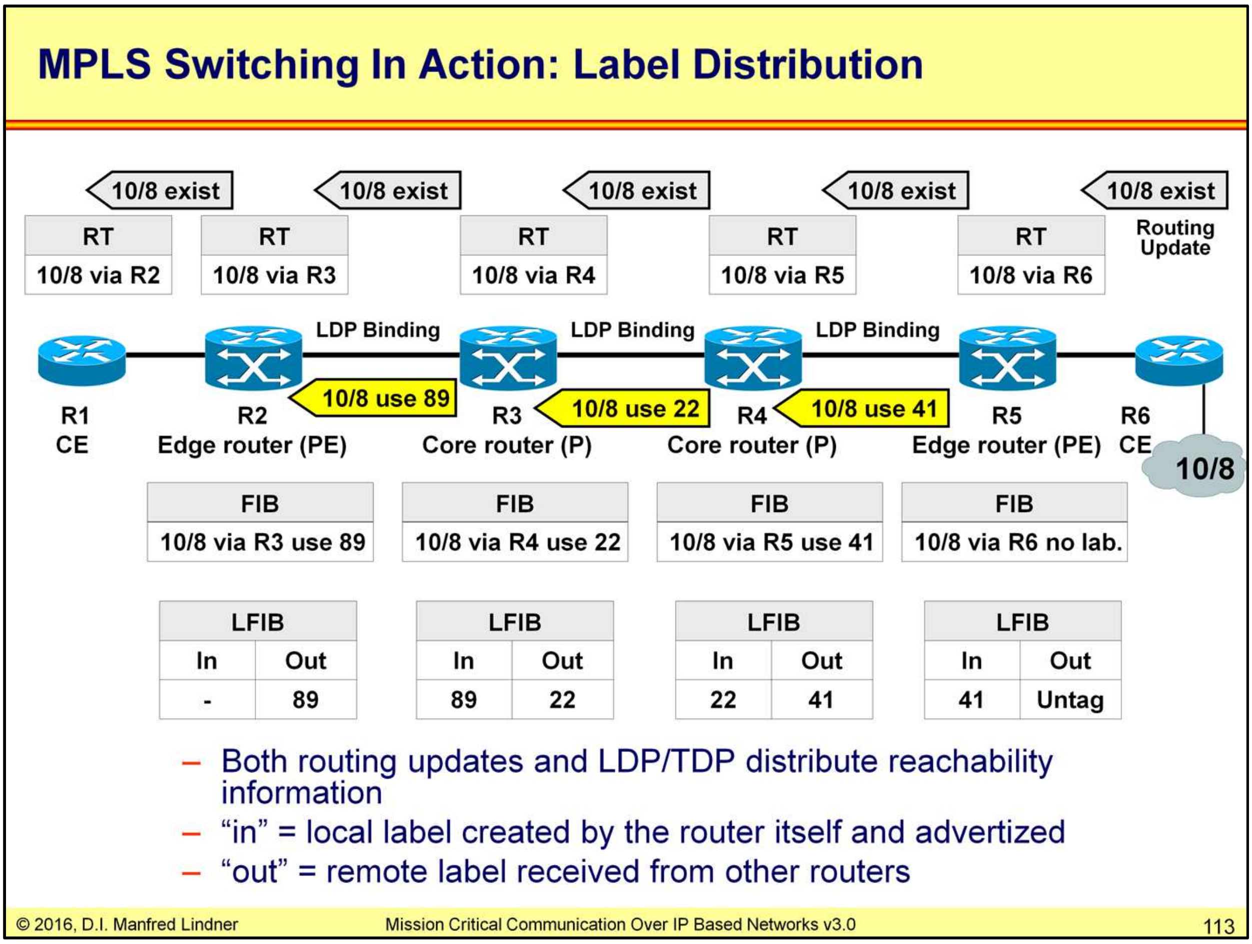




## Classical IP Forwarding: Hop by Hop Forwarding

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 112

The picture above shows classical IP hop-by-hop routing using signposts established by routing protocols and stored in the corresponding routing table.



## MPLS Switching In Action: Label Distribution

| RT |
|---|
| 10/8 via R2 |

| RT |
|---|
| 10/8 via R3 |

| RT |
|---|
| 10/8 via R4 |

| RT |
|---|
| 10/8 via R5 |

| RT |
|---|
| 10/8 via R6 |

| FIB |
|---|
| 10/8 via R3 use 89 |

| FIB |
|---|
| 10/8 via R4 use 22 |

| FIB |
|---|
| 10/8 via R5 use 41 |

| FIB |
|---|
| 10/8 via R6 no lab. |

| LFIB | |
|---|---|
| In | Out |
| - | 89 |

| LFIB | |
|---|---|
| In | Out |
| 89 | 22 |

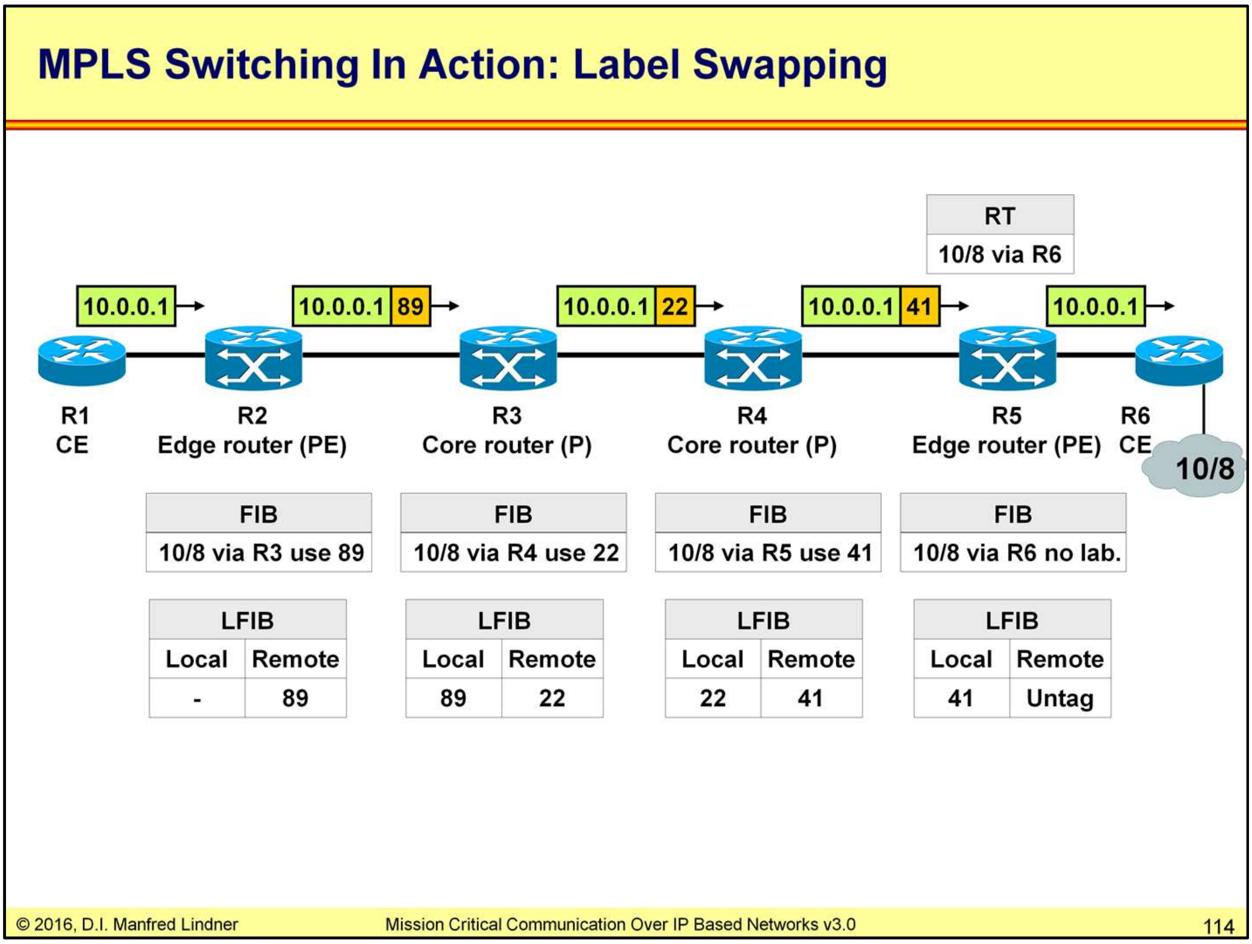
| LFIB | |
|---|---|
| In | Out |
| 22 | 41 |

| LFIB | |
|---|---|
| In | Out |
| 41 | Untag |

– Both routing updates and LDP/TDP distribute reachability information
– “in” = local label created by the router itself and advertized
– “out” = remote label received from other routers

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 113

The picture above shows how a label-switched path is established from left to the right. Both routing updates as well as a label distribution protocol (LDP or TDP) distribute reachability information for this destination network.



## MPLS Switching In Action: Label Swapping

| RT |
|---|
| 10/8 via R6 |

| FIB |
|---|
| 10/8 via R3 use 89 |

| FIB |
|---|
| 10/8 via R4 use 22 |

| FIB |
|---|
| 10/8 via R5 use 41 |

| FIB |
|---|
| 10/8 via R6 no lab. |

| LFIB | |
|---|---|
| Local | Remote |
| - | 89 |

| LFIB | |
|---|---|
| Local | Remote |
| 89 | 22 |

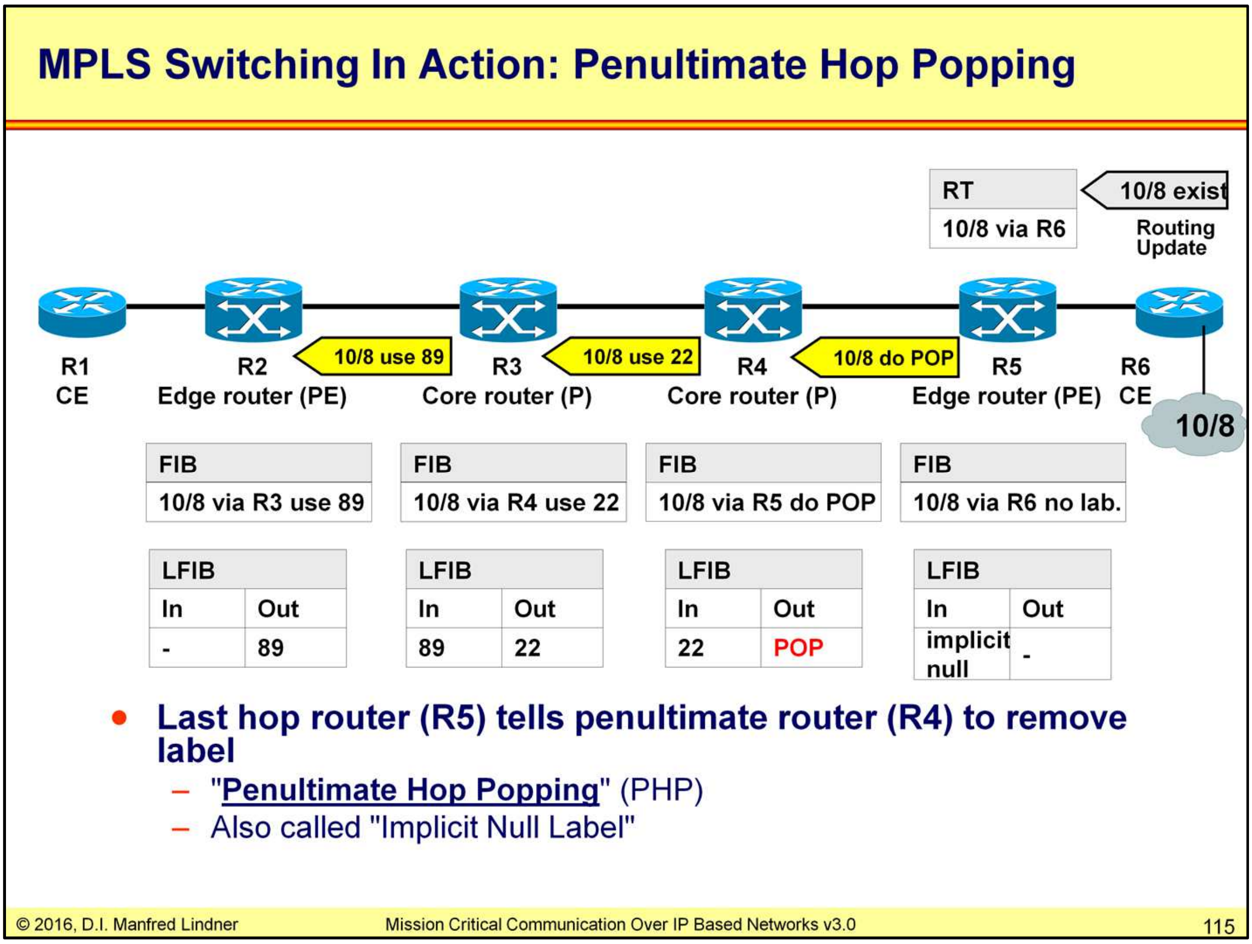
| LFIB | |
|---|---|
| Local | Remote |
| 22 | 41 |

| LFIB | |
|---|---|
| Local | Remote |
| 41 | Untag |

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 114

The picture above shows how packets can now be sent using a MPLS header. Label switching is performed on each hop (LSR) inside the provider domain (R2, R3, R4, R5). The LFIB tables are used to perform a fast lookup.

But R5 cannot find any outgoing label in its LFIB. After this unsuccessful lookup, R5 looks into the FIB and determines the next hop. Note that this double lookup would be done for every packet! Therefore it would be reasonable to remove the label even one hop earlier (the penultimate hop, R4) in order to leave R5's LFIB empty.



## MPLS Switching In Action: Penultimate Hop Popping

| FIB |
|---|
| 10/8 via R3 use 89 |

| FIB |
|---|
| 10/8 via R4 use 22 |

| FIB |
|---|
| 10/8 via R5 do POP |

| FIB |
|---|
| 10/8 via R6 no lab. |

| LFIB | |
|---|---|
| In | Out |
| - | 89 |

| LFIB | |
|---|---|
| In | Out |
| 89 | 22 |

| LFIB | |
|---|---|
| In | Out |
| 22 | POP |

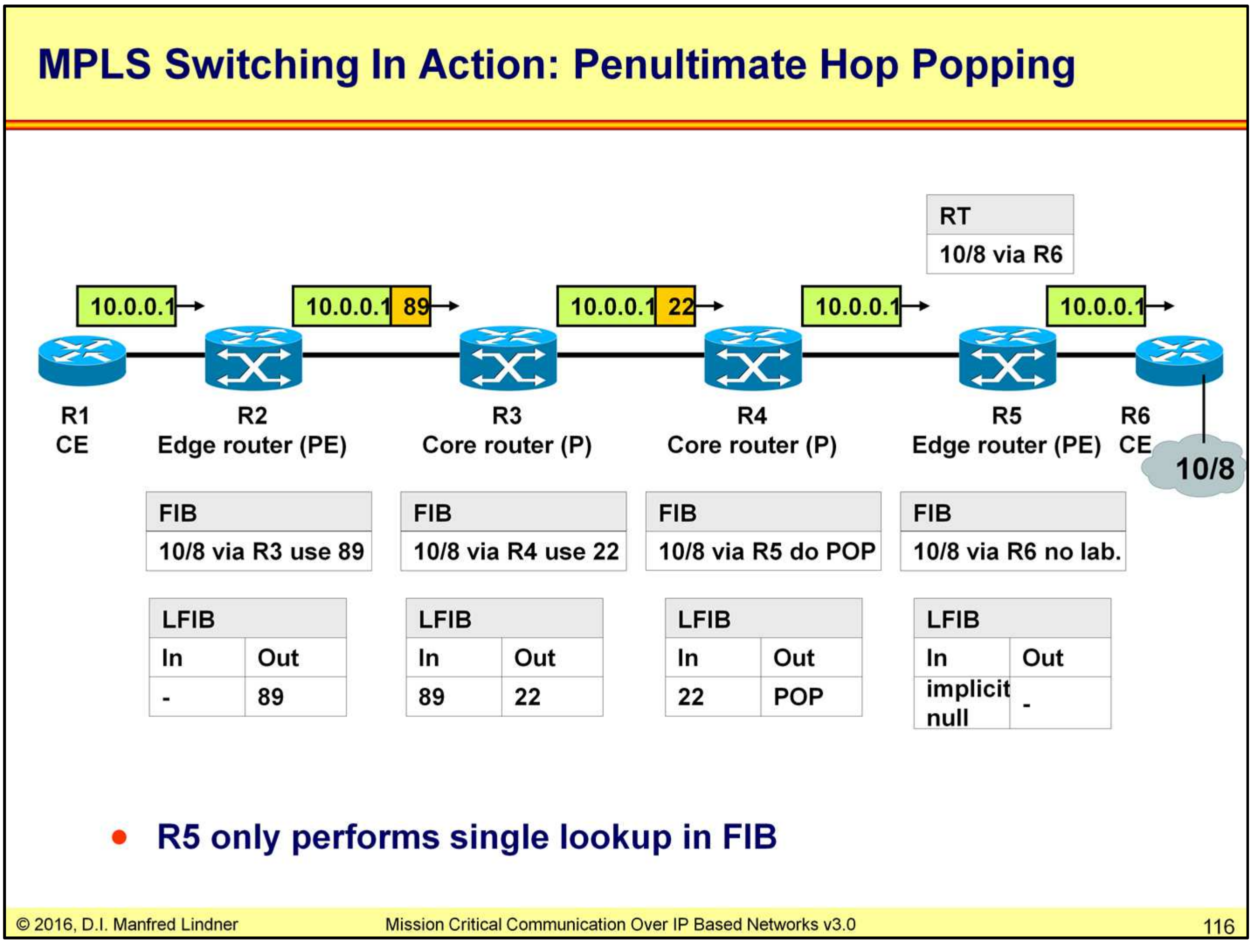
| LFIB | |
|---|---|
| In | Out |
| implicit null | - |

- **Last hop router (R5) tells penultimate router (R4) to remove label**
  - "Penultimate Hop Popping" (PHP)
  - Also called "Implicit Null Label"

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 115

In this scenario "Penultimate Hop Popping" (PHP) is illustrated. Now R5 does not allocate an incoming label for this destination but rather announces to R4 to use an "implicit null" label. It is also said, that R4 should perform the "POP" operation. The label number "3" had been reserved to represent the "do POP" command.

Implicit Null Label and hence POP upstream sent out only for directly connected networks or aggregates of advertising router



## MPLS Switching In Action: Penultimate Hop Popping

| RT |
|---|
| 10/8 via R6 |

| FIB |
|---|
| 10/8 via R3 use 89 |

| LFIB | |
|---|---|
| In | Out |
| - | 89 |

| FIB |
|---|
| 10/8 via R4 use 22 |

| LFIB | |
|---|---|
| In | Out |
| 89 | 22 |

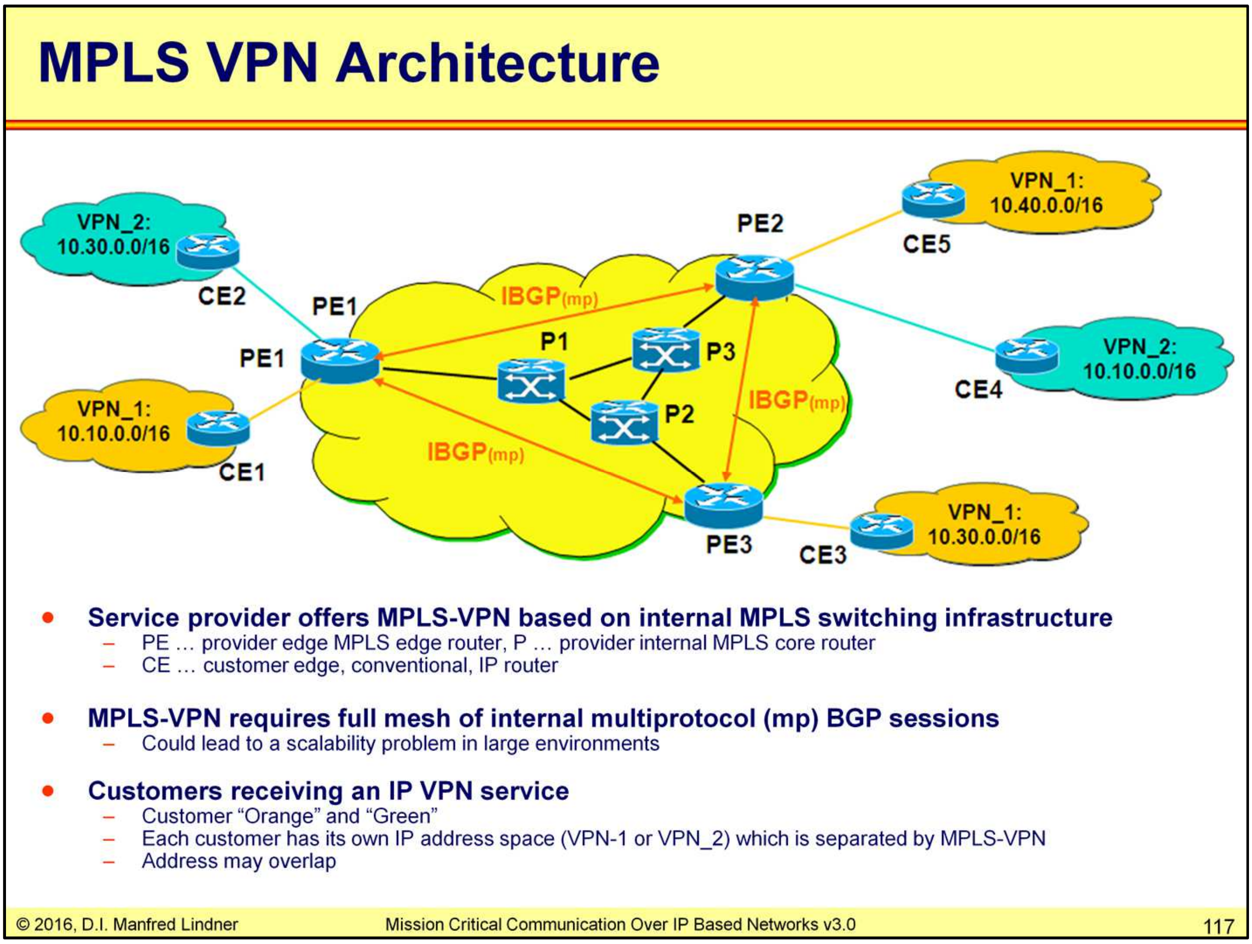
| FIB |
|---|
| 10/8 via R5 do POP |

| LFIB | |
|---|---|
| In | Out |
| 22 | POP |

| FIB |
|---|
| 10/8 via R6 no lab. |

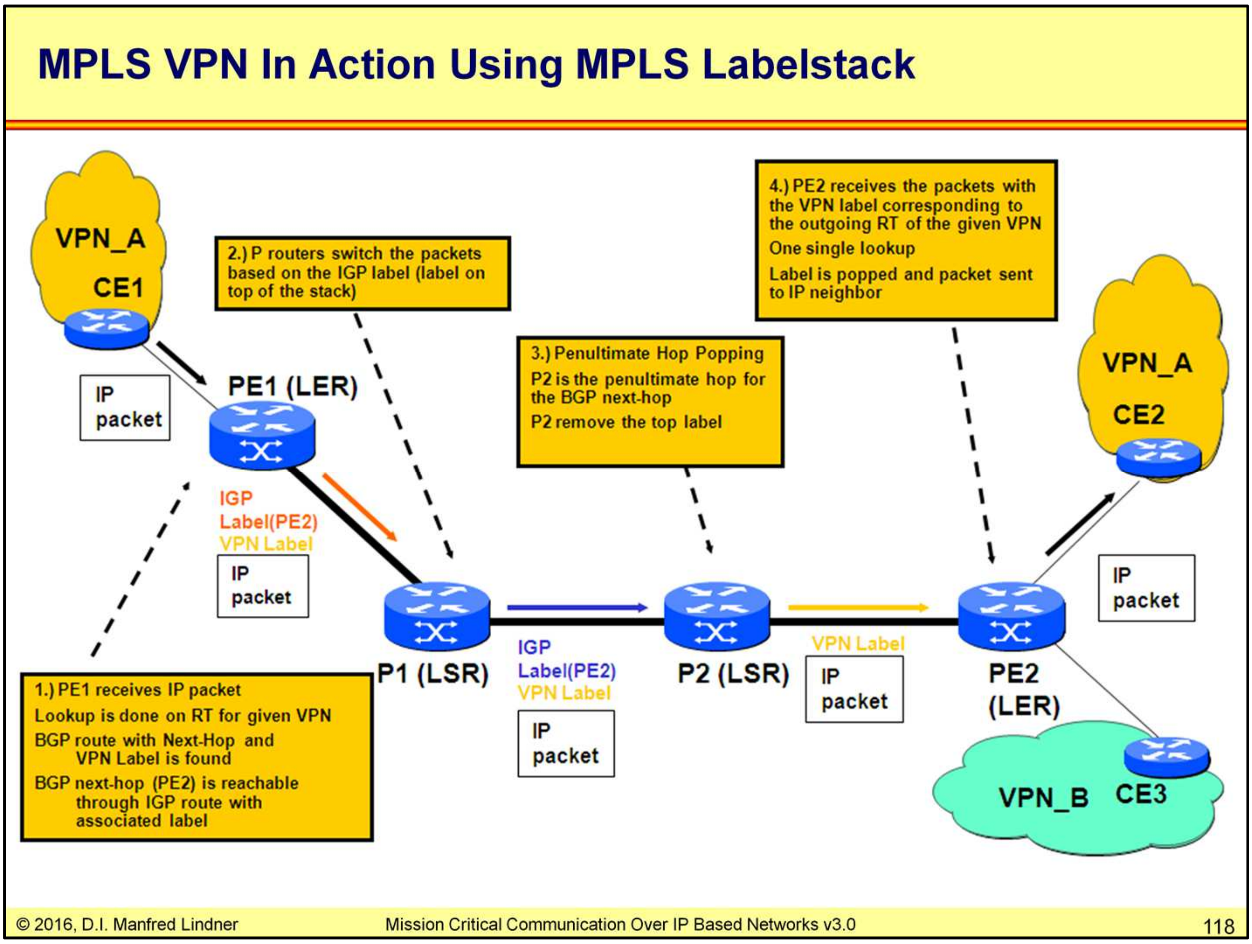
| LFIB | |
|---|---|
| In | Out |
| implicit null | - |

- R5 only performs single lookup in FIB



## MPLS VPN Architecture

- **Service provider offers MPLS-VPN based on internal MPLS switching infrastructure**
  - PE … provider edge MPLS edge router, P … provider internal MPLS core router
  - CE … customer edge, conventional, IP router
- **MPLS-VPN requires full mesh of internal multiprotocol (mp) BGP sessions**
  - Could lead to a scalability problem in large environments
- **Customers receiving an IP VPN service**
  - Customer “Orange” and “Green”
  - Each customer has its own IP address space (VPN-1 or VPN_2) which is separated by MPLS-VPN
  - Address may overlap



## MPLS VPN In Action Using MPLS Labelstack

1.) PE1 receives IP packet

Lookup is done on RT for given VPN

BGP route with Next-Hop and VPN Label is found

BGP next-hop (PE2) is reachable through IGP route with associated label

2.) P routers switch the packets based on the IGP label (label on top of the stack)

3.) Penultimate Hop Popping

P2 is the penultimate hop for the BGP next-hop

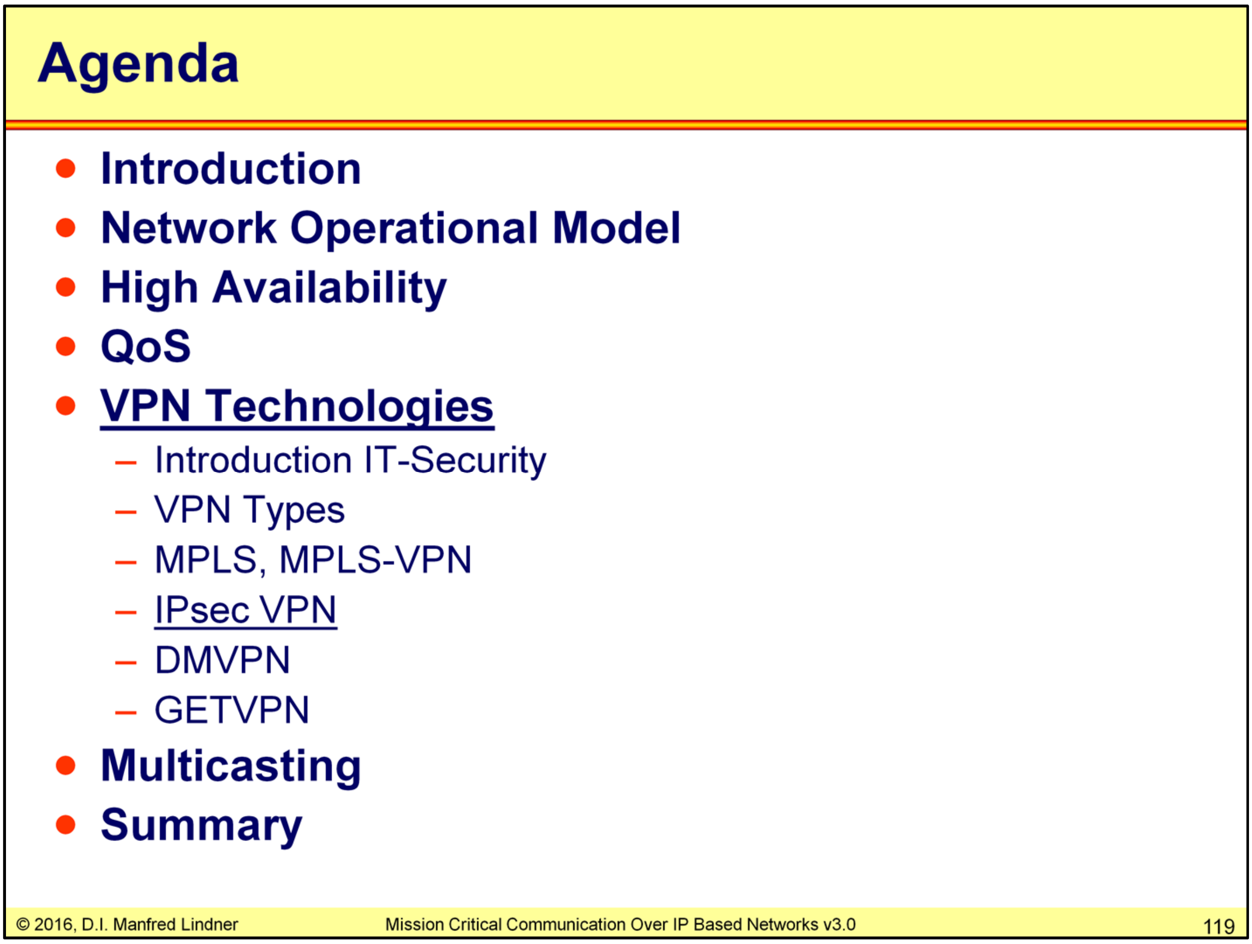
P2 remove the top label

4.) PE2 receives the packets with the VPN label corresponding to the outgoing RT of the given VPN

One single lookup

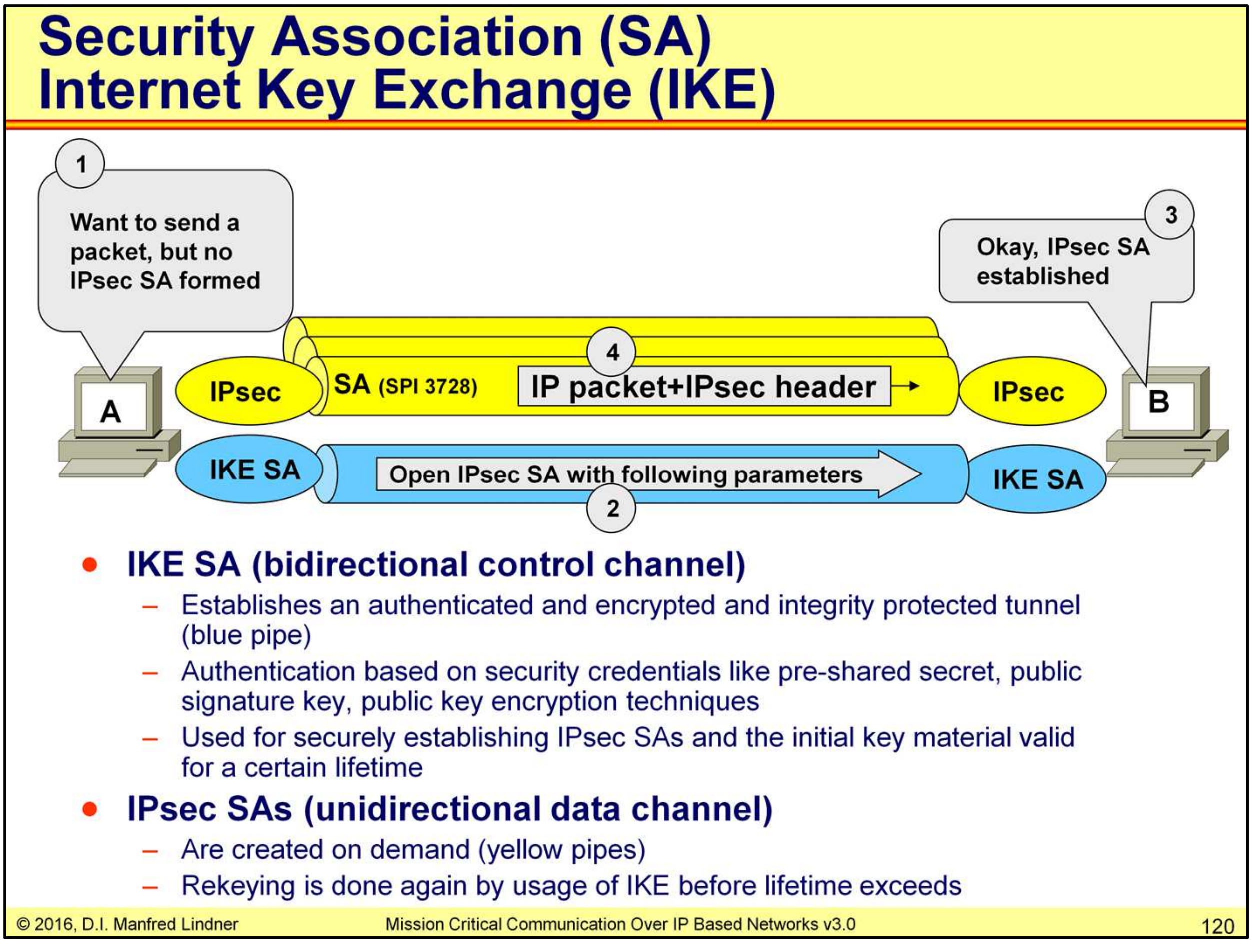
Label is popped and packet sent to IP neighbor

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 118



# Agenda

- Introduction
- Network Operational Model
- High Availability
- QoS
- VPN Technologies
  - Introduction IT-Security
  - VPN Types
  - MPLS, MPLS-VPN
  - IPsec VPN
  - DMVPN
  - GETVPN
- Multicasting
- Summary



## Security Association (SA)
## Internet Key Exchange (IKE)

1
Want to send a packet, but no IPsec SA formed
3
Okay, IPsec SA established
A
B
IPsec
SA (SPI 3728)
4
IP packet+IPsec header
IPsec
IKE SA
Open IPsec SA with following parameters
2
IKE SA

- **IKE SA (bidirectional control channel)**
  - Establishes an authenticated and encrypted and integrity protected tunnel (blue pipe)
  - Authentication based on security credentials like pre-shared secret, public signature key, public key encryption techniques
  - Used for securely establishing IPsec SAs and the initial key material valid for a certain lifetime
- **IPsec SAs (unidirectional data channel)**
  - Are created on demand (yellow pipes)
  - Rekeying is done again by usage of IKE before lifetime exceeds

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 120

What is IPsec?

IPsec enables a system to select required security protocols, determine the algorithms to use for the services, and put in place any cryptographic keys required. IPsec protects one path between a pair of hosts, between a pair of security gateways, or between a security gateway and a host.

Elements of IPsec:

1) Security protocols (for traffic security):

Authentication header (AH) defined in RFC 2402 obsoleted by RFC 4302, 4305 since Dec 2005.

Encapsulating security payload (ESP) defined in RFC 2406 obsoleted by RFC 4303, 4305 since Dec 2005.

2) Cryptographic Algorithms:

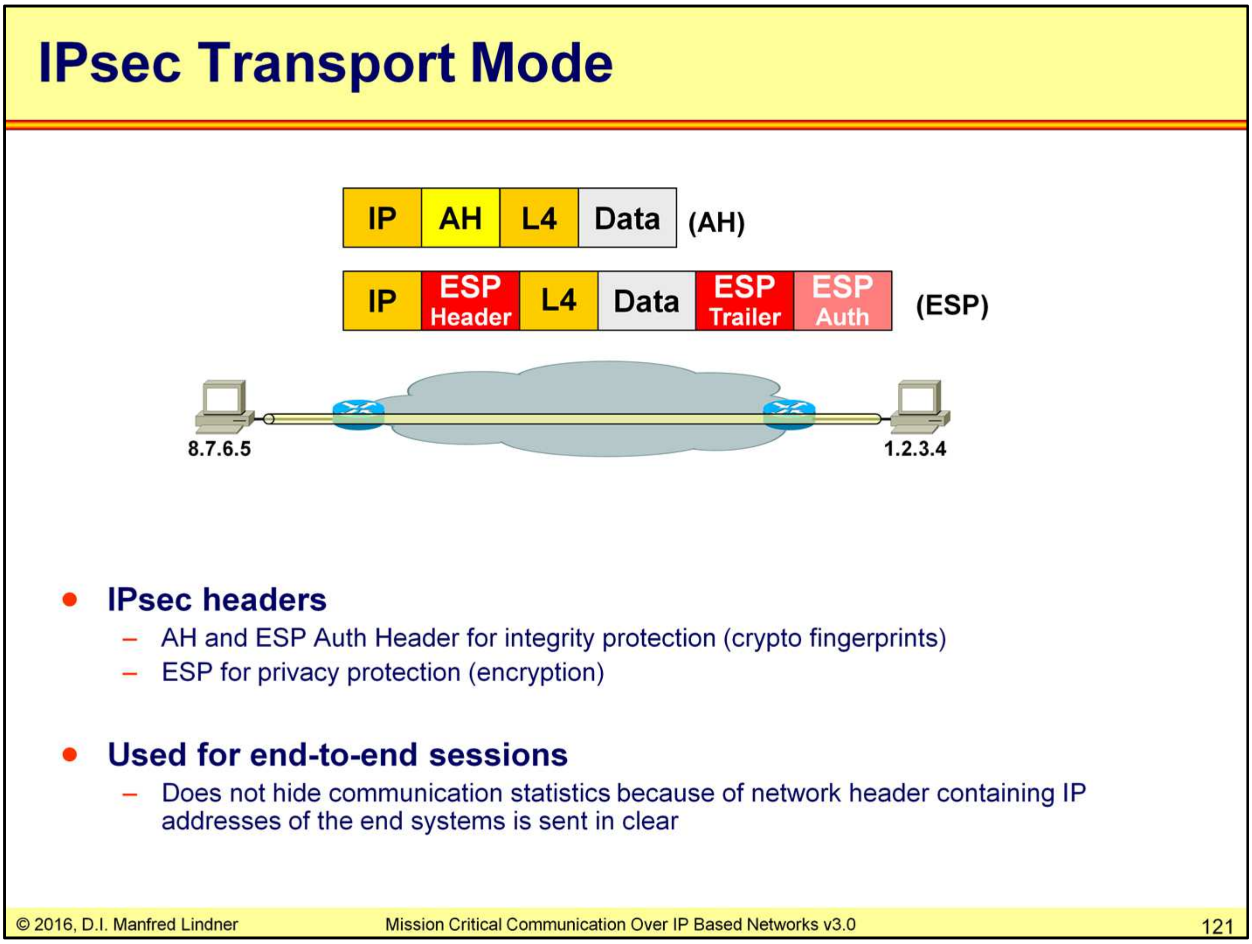
Mandatory implementation requirements for ESP and AH are defined in RFC 4305.

Secret-key algorithms are used so far because of performance reason.

HMAC-SHA1, HMAC-MD5, AES-XCBC-MAC, DES-CBC, 3DES-CBC, AES-CBC, AES-CTR are defined in separate RFC' (see  4305, 2403, 2404, 2405, 2451, 3566, 3602, 3686)

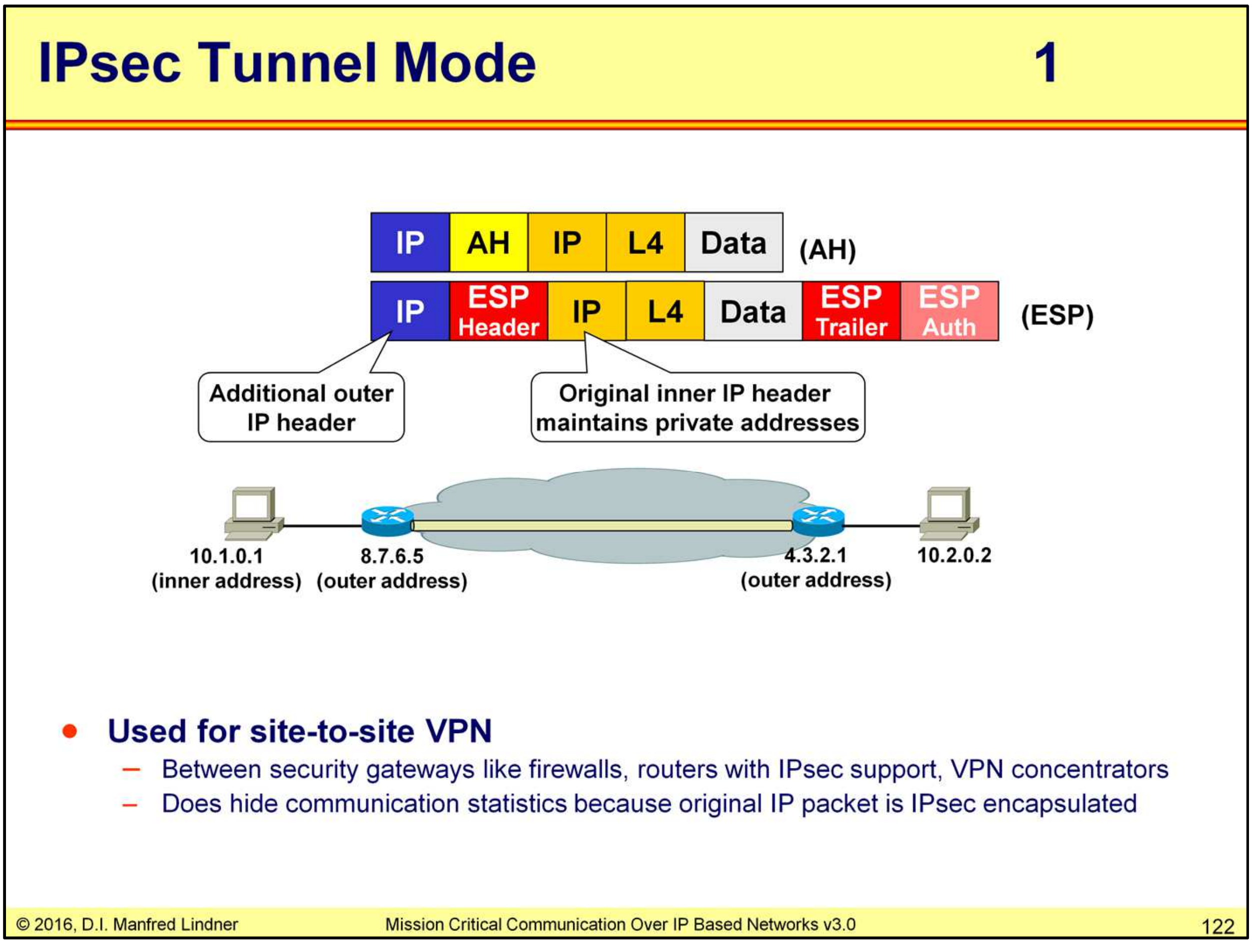
3) Management of security associations (SA) and keys:

Manual management for static and small environments. Automatic management for scalable environments by ISAKMP (Internet Security Association and Key Management Protocol defined in RFC 2408 obsoleted by RFC 4306 since Dec 2005) and Internet Key Exchange (IKE, defined in RFC 2409 aka IKEv1, obsoleted by RFC 4306 since Dec 2005 -> IKEv2).



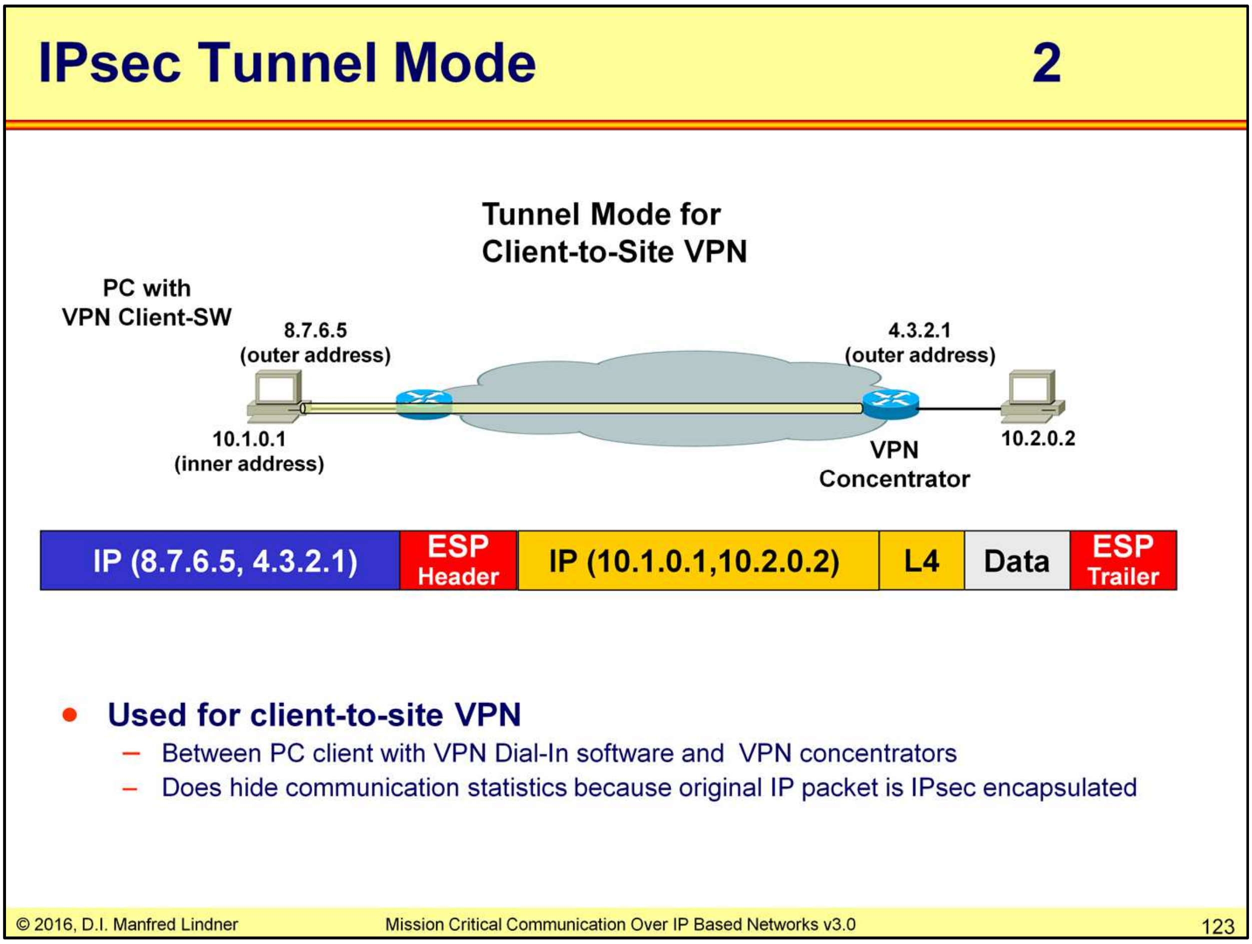
# IPsec Transport Mode

- **IPsec headers**
  - AH and ESP Auth Header for integrity protection (crypto fingerprints)
  - ESP for privacy protection (encryption)
- **Used for end-to-end sessions**
  - Does not hide communication statistics because of network header containing IP addresses of the end systems is sent in clear



# IPsec Tunnel Mode 1

- **Used for site-to-site VPN**
  - Between security gateways like firewalls, routers with IPsec support, VPN concentrators
  - Does hide communication statistics because original IP packet is IPsec encapsulated



## IPsec Tunnel Mode 2

| | | | | | |
|---|---|---|---|---|---|
| IP (8.7.6.5, 4.3.2.1) | ESP Header | IP (10.1.0.1,10.2.0.2) | L4 | Data | ESP Trailer |

- Used for client-to-site VPN
  - Between PC client with VPN Dial-In software and VPN concentrators
  - Does hide communication statistics because original IP packet is IPsec encapsulated

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 123

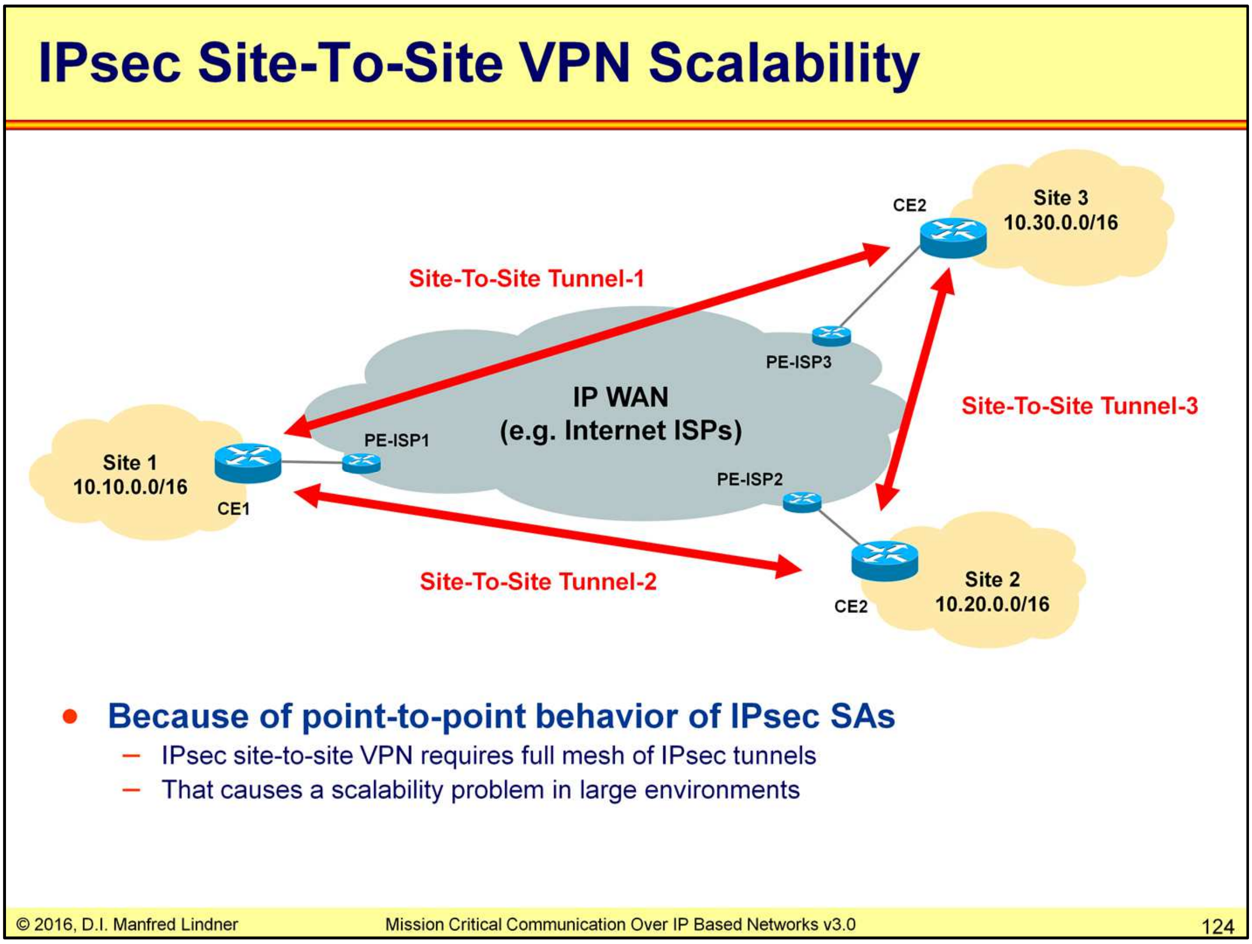
IPsec for site-to-site VPN often uses pre-shared secrets for authentication of IKE peers. Why? Usage of certificates means maintaining a kind of PKI (Public Key Infrastructure) and at least a private CA (Certification Authority) server is needed. On the other side VPN router/concentrator can physically protected in the data center

With mobile PC with IPsec for client-to-cite VPN there is a different situation. Mobile PCs calling from insecure places. Preshared secret may be compromised hence configuration and maintenance overhead if number of clients is high. Therefore a combination of IPsec, well-known RAS authentication techniques (PPP with EAP, RFC 3748) and X-AUTH is used.

Client dials-in, authenticates itself at a authentication server (VPN concentrator) using a group-key (could be a preshared secret) and a first secure tunnel is established. Now the VPN concentrator using X-AUTH asks the user of the PC for the credentials (e.g. username/password). If ok, then PC is allowed to come in. The necessary IPsec configuration is pushed from the VPN concentrator to the client to establish the final secure tunnel. Also the internal IP address of the mobile-PCs is assigned by the VPN concentrator and maybe other security settings like host-based firewall setting, OS healthy checks or disabling splitted tunneling are forced by the VPN concentrator.

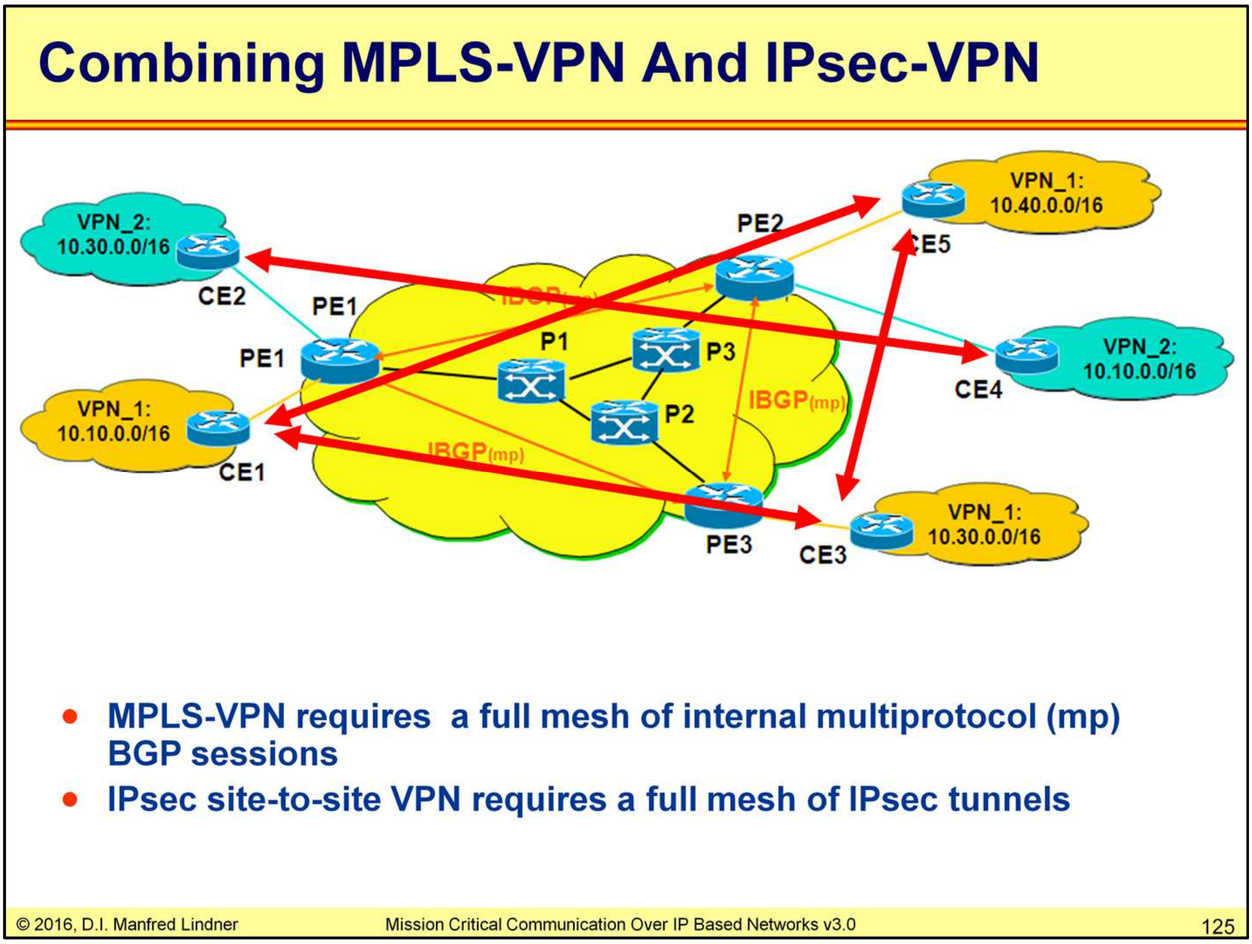
X-AUTH exchange as an add-on to IKEv1. X-AUTH exchange is an inherent optional part of IKEv2.

NAT (Network Address Translation) or N(P)AT (NAT with port address translation) causes problems in case of IPsec AH or ESP. You need a NAT traversal technique in such a case. NAT-T encapsulates IPsec stuff into a UDP or TCP packet.



# IPsec Site-To-Site VPN Scalability

- **Because of point-to-point behavior of IPsec SAs**
  - IPsec site-to-site VPN requires full mesh of IPsec tunnels
  - That causes a scalability problem in large environments



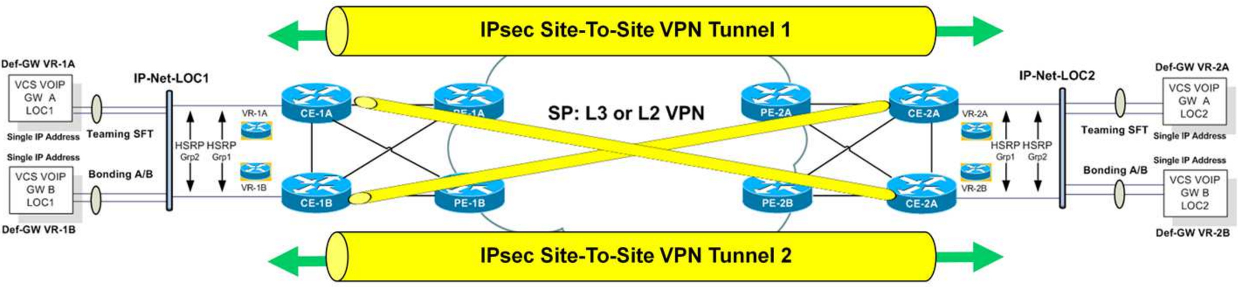
# Combining MPLS-VPN And IPsec-VPN

- MPLS-VPN requires a full mesh of internal multiprotocol (mp) BGP sessions
- IPsec site-to-site VPN requires a full mesh of IPsec tunnels

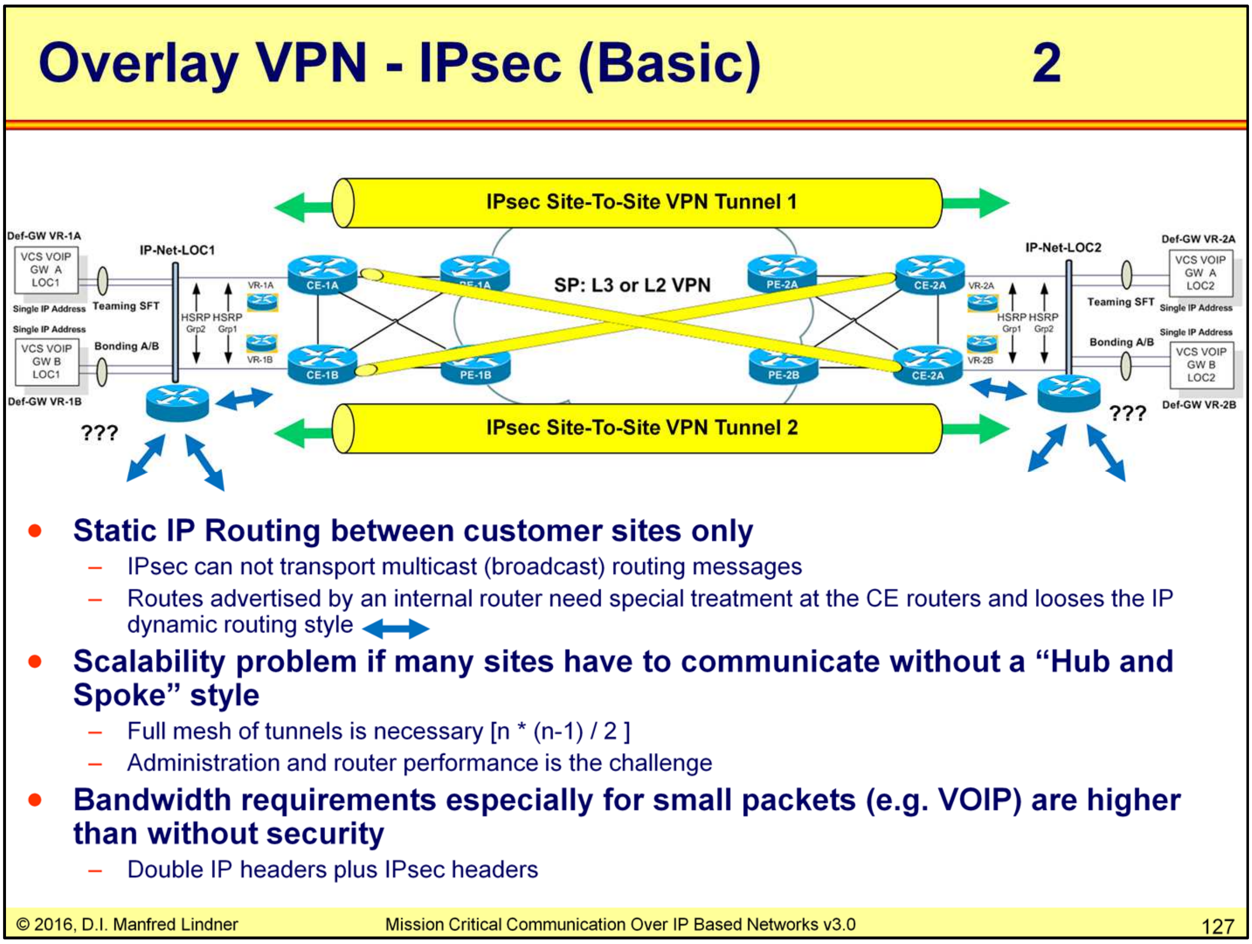
© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 125

Combining MPLS VPN (Separation) and IPsec VPN (Encryption/Integrity)

## Overlay VPN - IPsec (Basic) 1

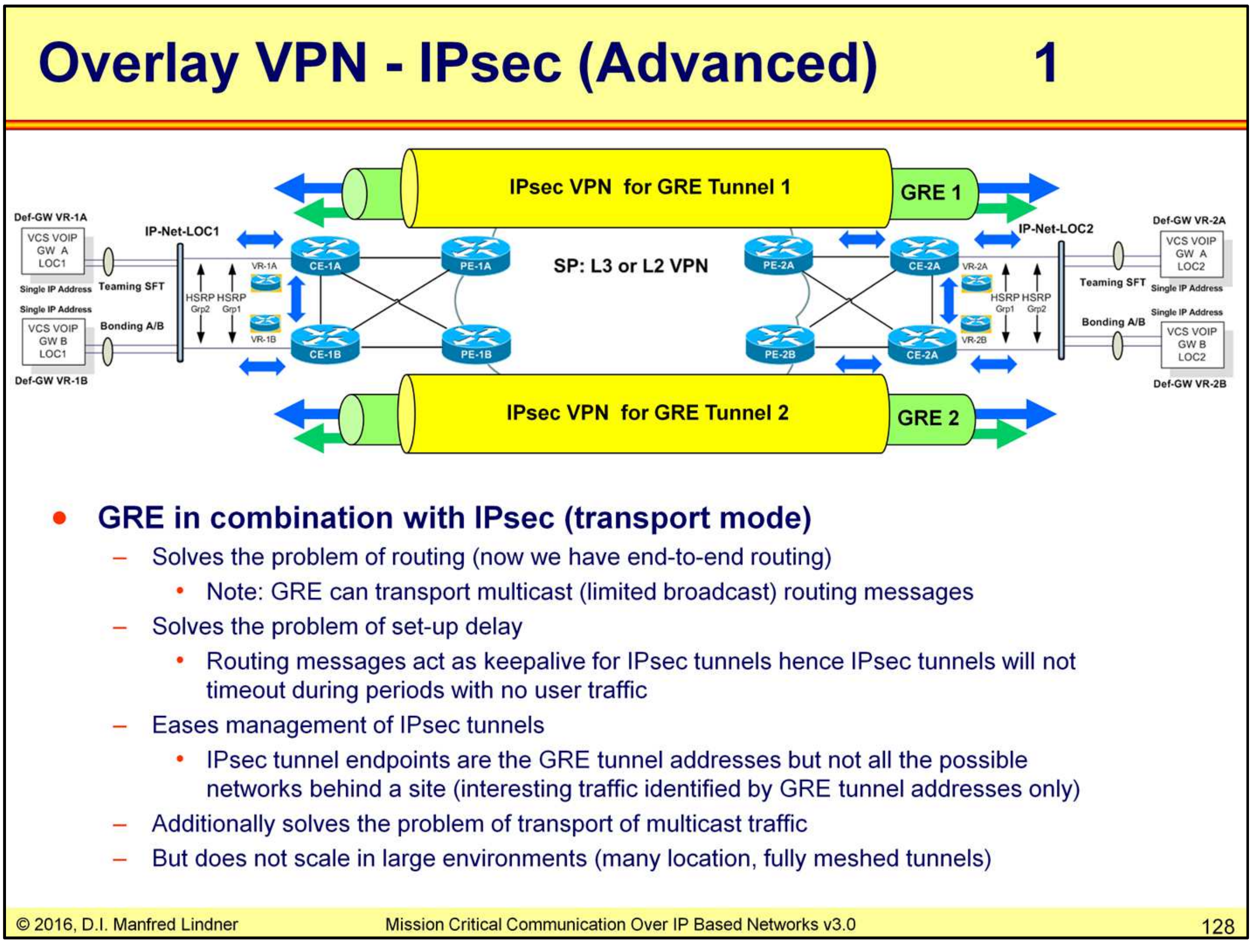


- **IPsec Management**
  - Task of the customer on the CE routers
- **Traffic between IP-Net-LOC1 and IP-Net-LOC2** ↔
  - Will be protected by IPsec site-site VPN (tunnel-mode) from CE-1A to CE-2A ⬭
  - Other tunnels on picture above for redundancy
  - Interesting traffic (= to be encrypted traffic) has to be specified (worst case: every net-id combination)
- **Attention: IPsec is a kind of "Dial-Up" technique**
  - Set-up of an IPsec tunnel is triggered by "interesting traffic"
  - Hence the problem of set-up delay of so far not used tunnels arises in case of a failure



## Overlay VPN - IPsec (Basic) 2

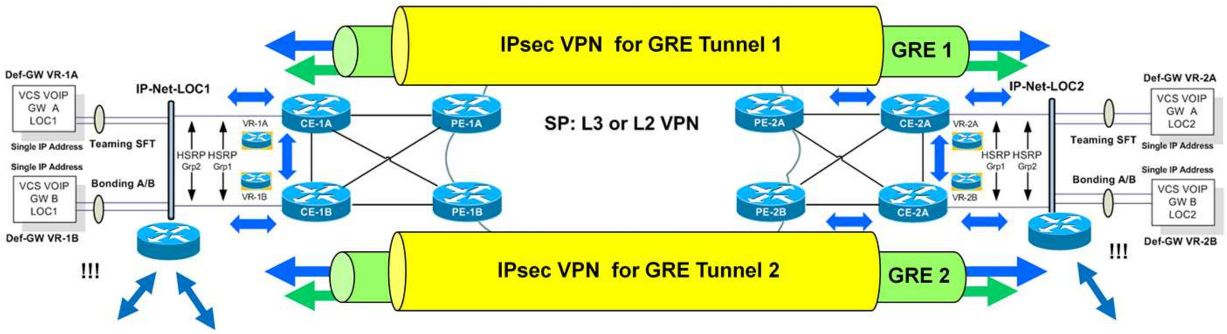
- **Static IP Routing between customer sites only**
  - IPsec can not transport multicast (broadcast) routing messages
  - Routes advertised by an internal router need special treatment at the CE routers and looses the IP dynamic routing style ⟷
- **Scalability problem if many sites have to communicate without a “Hub and Spoke” style**
  - Full mesh of tunnels is necessary [n * (n-1) / 2 ]
  - Administration and router performance is the challenge
- **Bandwidth requirements especially for small packets (e.g. VOIP) are higher than without security**
  - Double IP headers plus IPsec headers



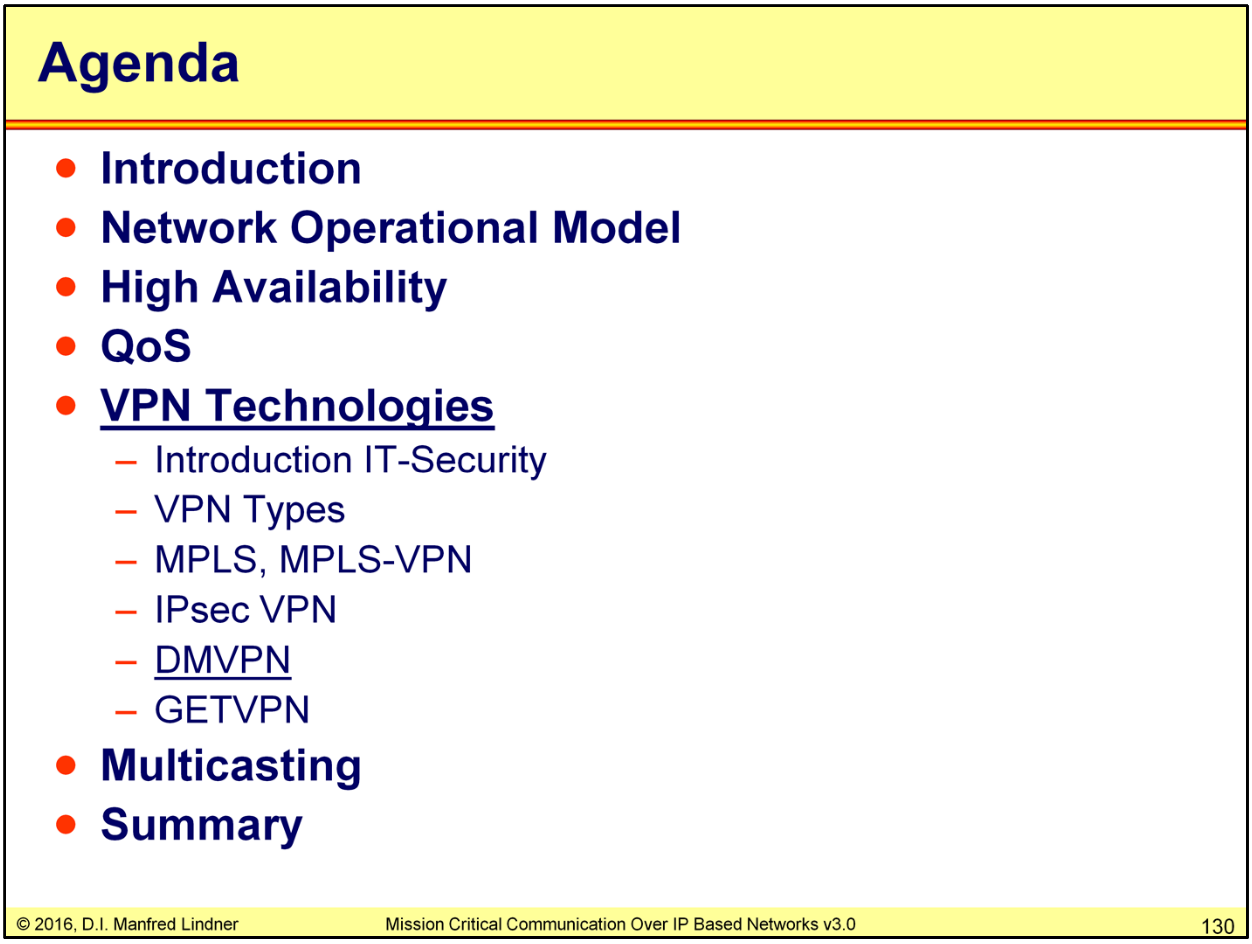
## Overlay VPN - IPsec (Advanced) 1

- **GRE in combination with IPsec (transport mode)**
  - Solves the problem of routing (now we have end-to-end routing)
    - Note: GRE can transport multicast (limited broadcast) routing messages
  - Solves the problem of set-up delay
    - Routing messages act as keepalive for IPsec tunnels hence IPsec tunnels will not timeout during periods with no user traffic
  - Eases management of IPsec tunnels
    - IPsec tunnel endpoints are the GRE tunnel addresses but not all the possible networks behind a site (interesting traffic identified by GRE tunnel addresses only)
  - Additionally solves the problem of transport of multicast traffic
  - But does not scale in large environments (many location, fully meshed tunnels)

# Overlay VPN - IPsec (Advanced) 2

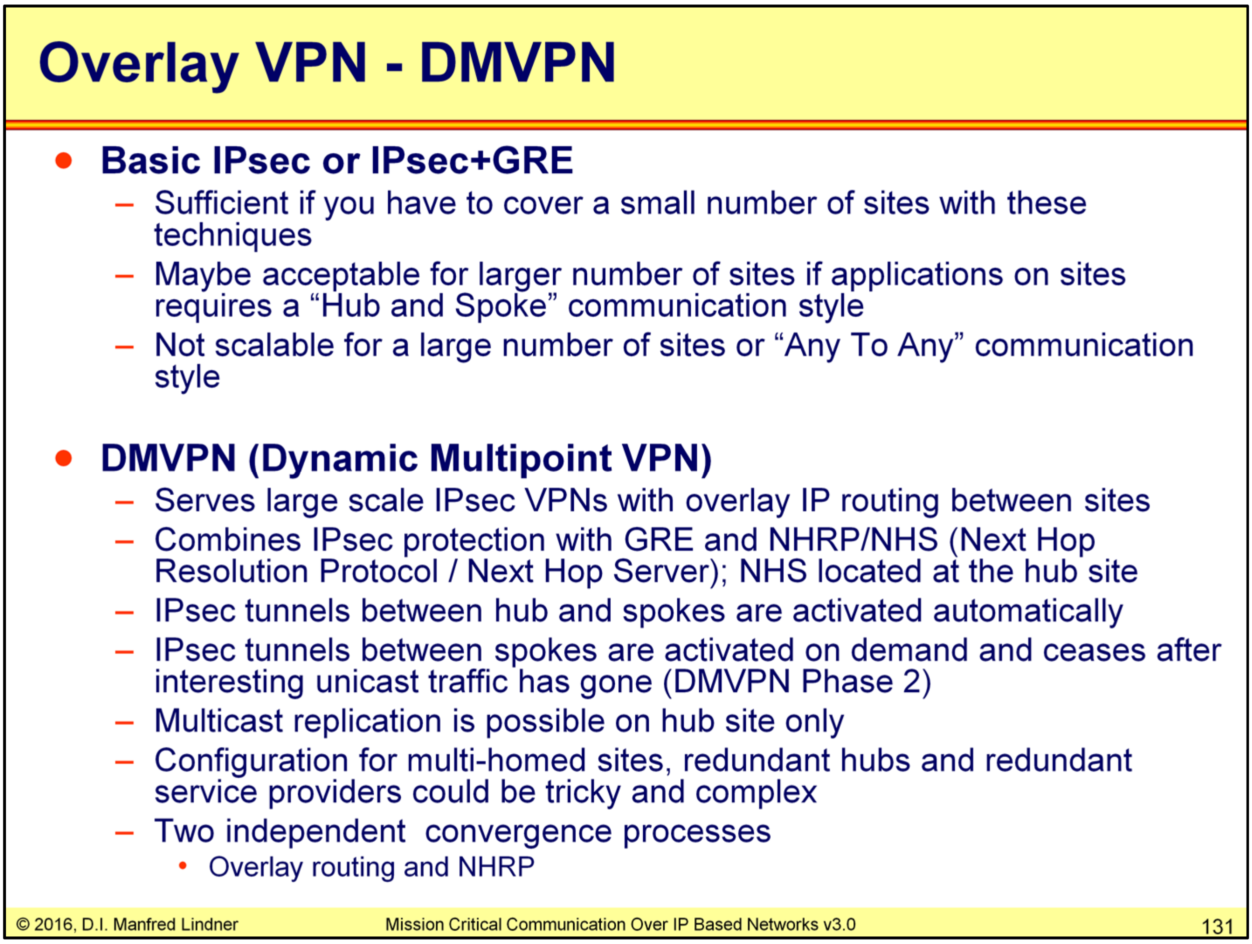


- **Dynamic Routing in the Overlay Network**
  - Eases management of routing
    - No static routes necessary
    - Full view of all sites and their networks in the overlay network
    - Service provider independent routing
  - Can improve routing convergence
    - Even if the routing convergence of the SP provider is too slow for the applications of the customer
    - By tuning routing parameters of the end-to.-end overlay routing protocol



# Agenda

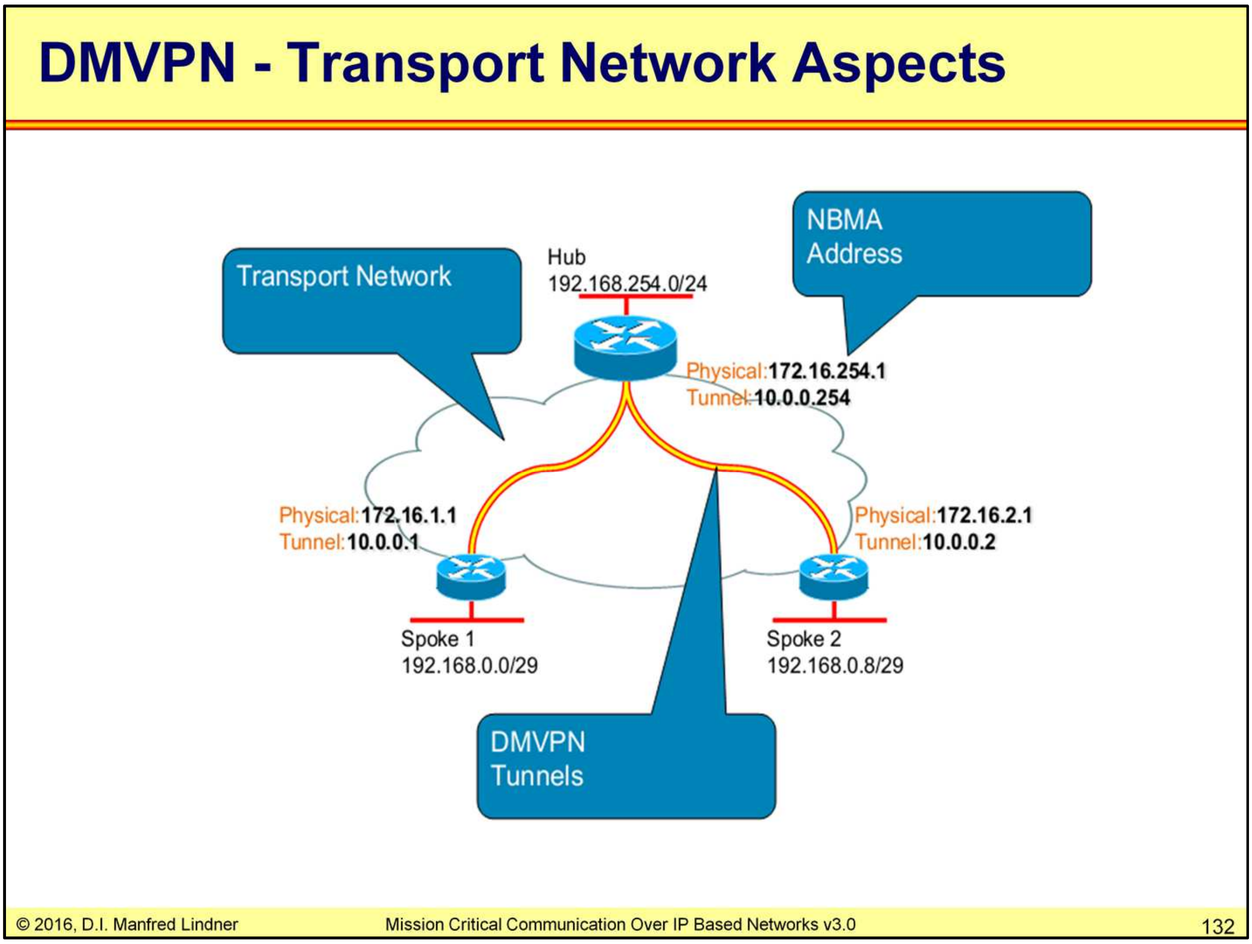
- Introduction
- Network Operational Model
- High Availability
- QoS
- VPN Technologies
  - Introduction IT-Security
  - VPN Types
  - MPLS, MPLS-VPN
  - IPsec VPN
  - DMVPN
  - GETVPN
- Multicasting
- Summary



# Overlay VPN - DMVPN

- **Basic IPsec or IPsec+GRE**
  - Sufficient if you have to cover a small number of sites with these techniques
  - Maybe acceptable for larger number of sites if applications on sites requires a “Hub and Spoke” communication style
  - Not scalable for a large number of sites or “Any To Any” communication style

- **DMVPN (Dynamic Multipoint VPN)**
  - Serves large scale IPsec VPNs with overlay IP routing between sites
  - Combines IPsec protection with GRE and NHRP/NHS (Next Hop Resolution Protocol / Next Hop Server); NHS located at the hub site
  - IPsec tunnels between hub and spokes are activated automatically
  - IPsec tunnels between spokes are activated on demand and ceases after interesting unicast traffic has gone (DMVPN Phase 2)
  - Multicast replication is possible on hub site only
  - Configuration for multi-homed sites, redundant hubs and redundant service providers could be tricky and complex
  - Two independent convergence processes
    - Overlay routing and NHRP



# DMVPN - Transport Network Aspects

Transport Network

Hub
192.168.254.0/24

NBMA
Address

Physical:172.16.254.1
Tunnel:10.0.0.254

Physical:172.16.1.1
Tunnel:10.0.0.1

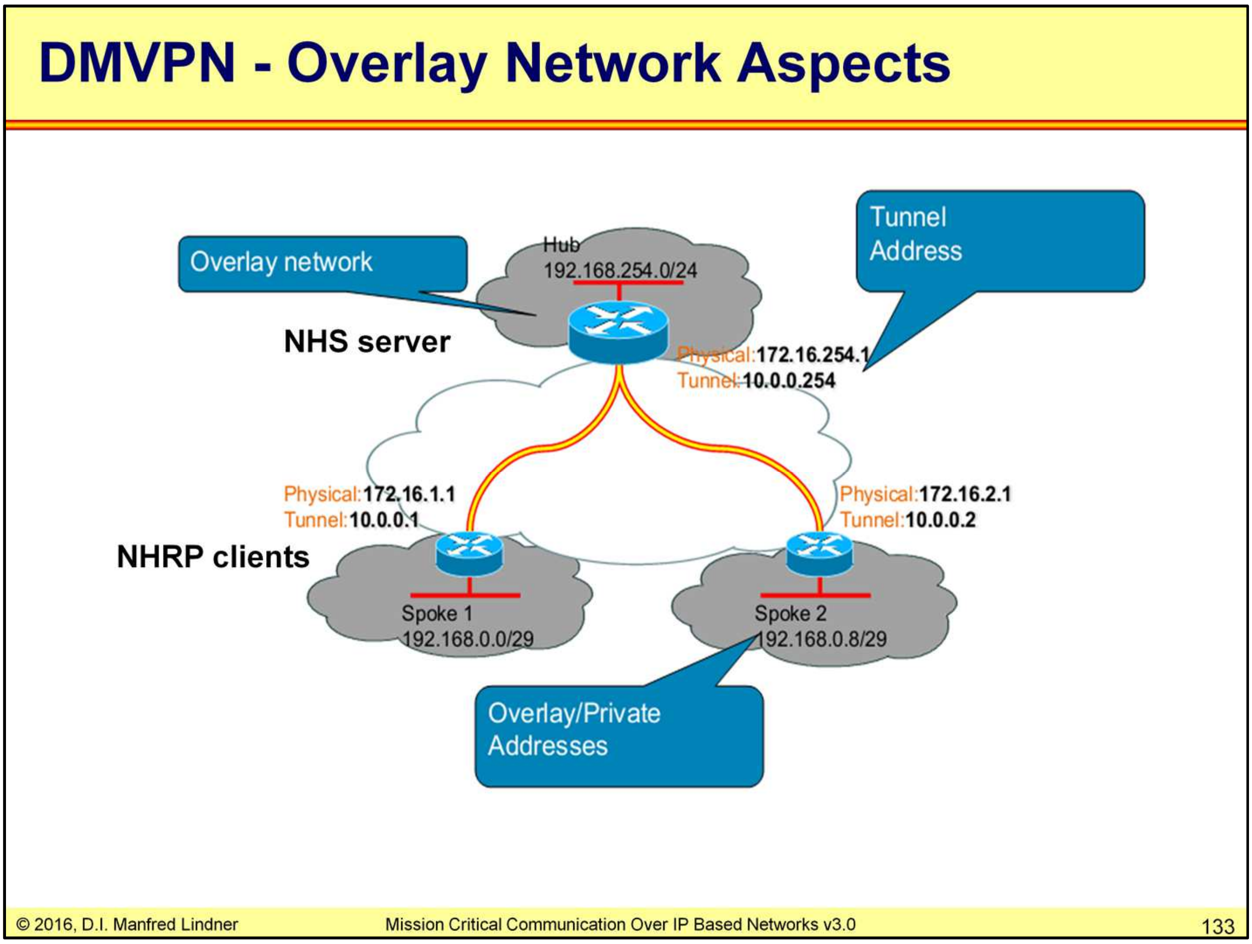
Physical:172.16.2.1
Tunnel:10.0.0.2

Spoke 1
192.168.0.0/29

Spoke 2
192.168.0.8/29

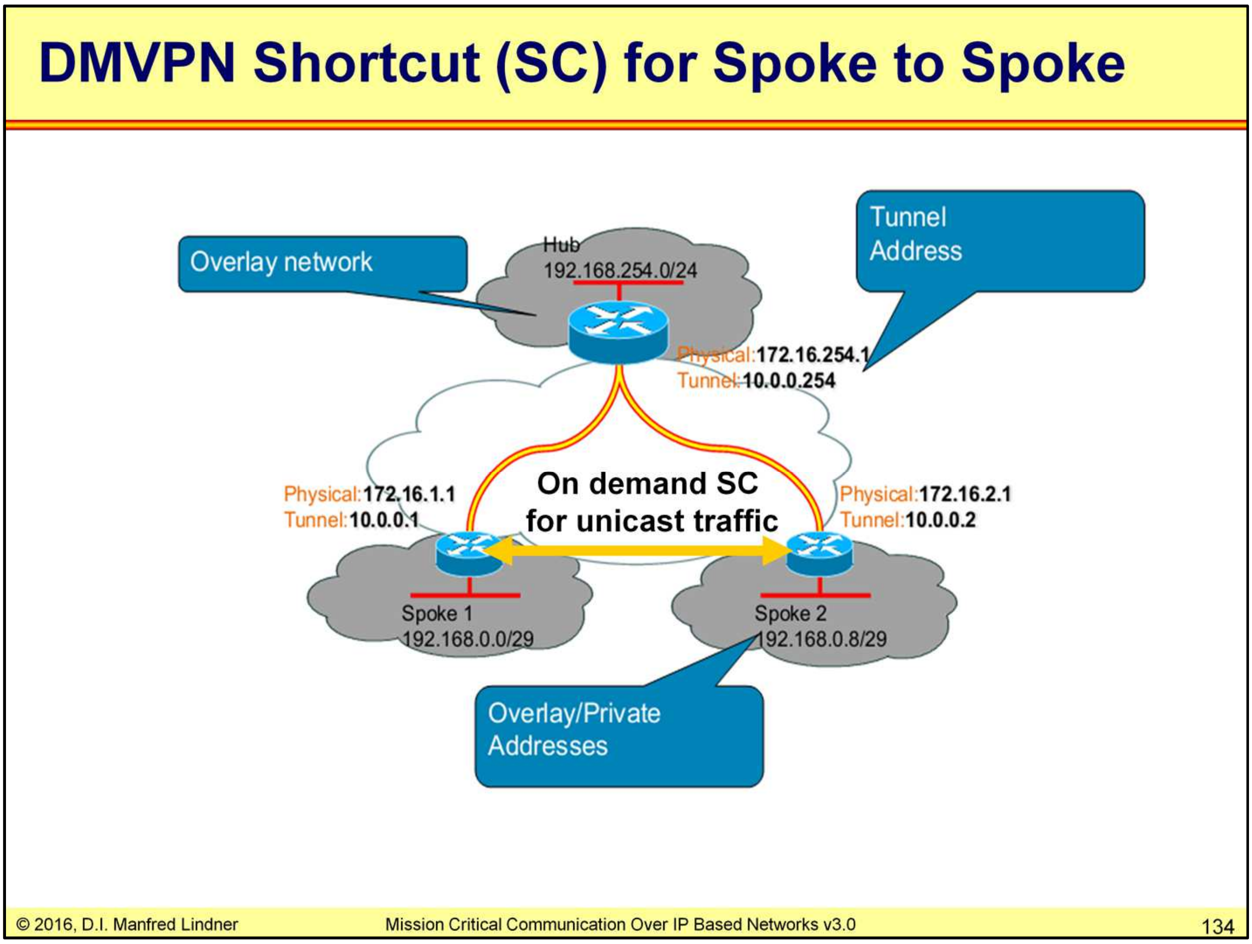
DMVPN
Tunnels

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 132



# DMVPN - Overlay Network Aspects

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 133



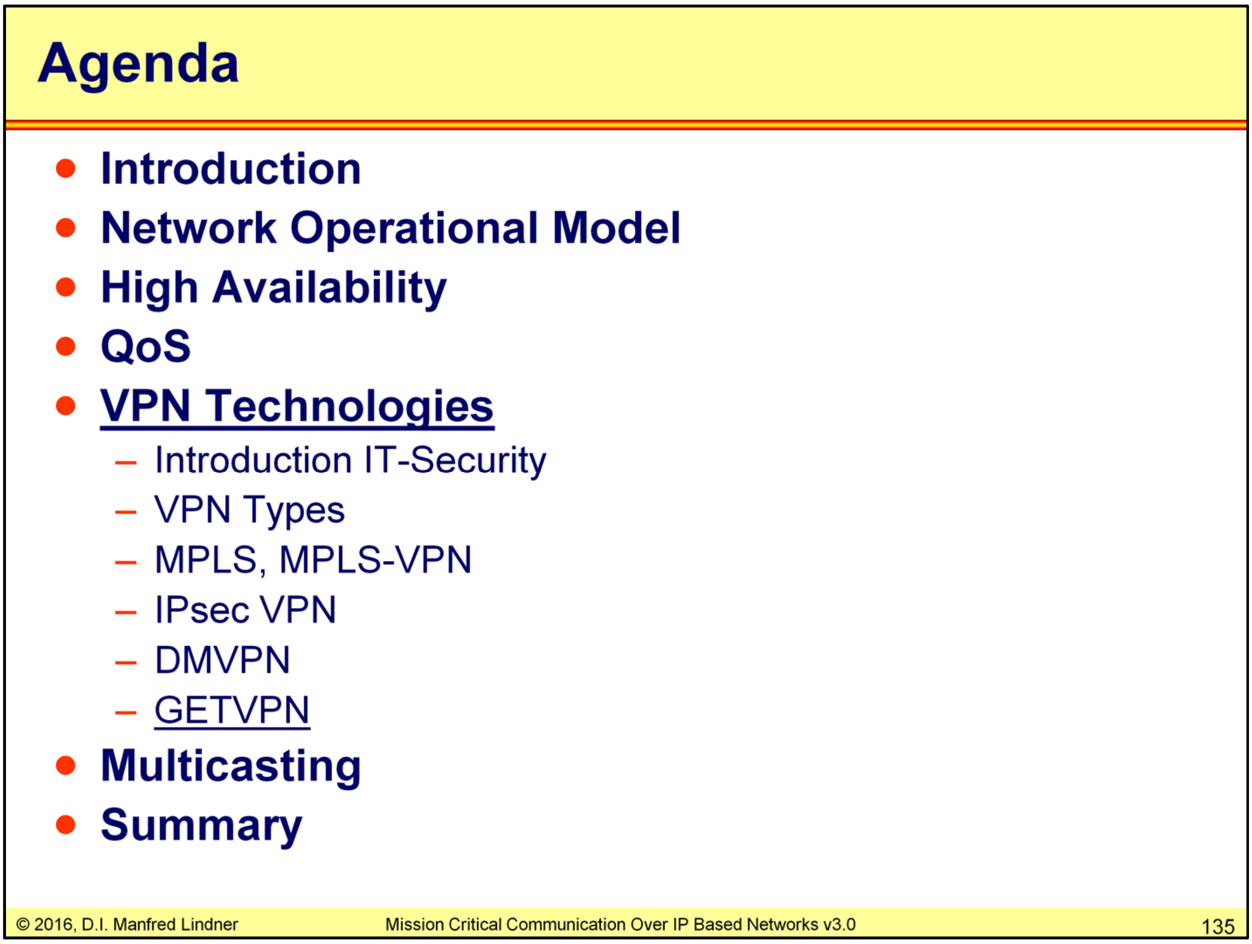
## DMVPN Shortcut (SC) for Spoke to Spoke

Overlay network

Tunnel Address

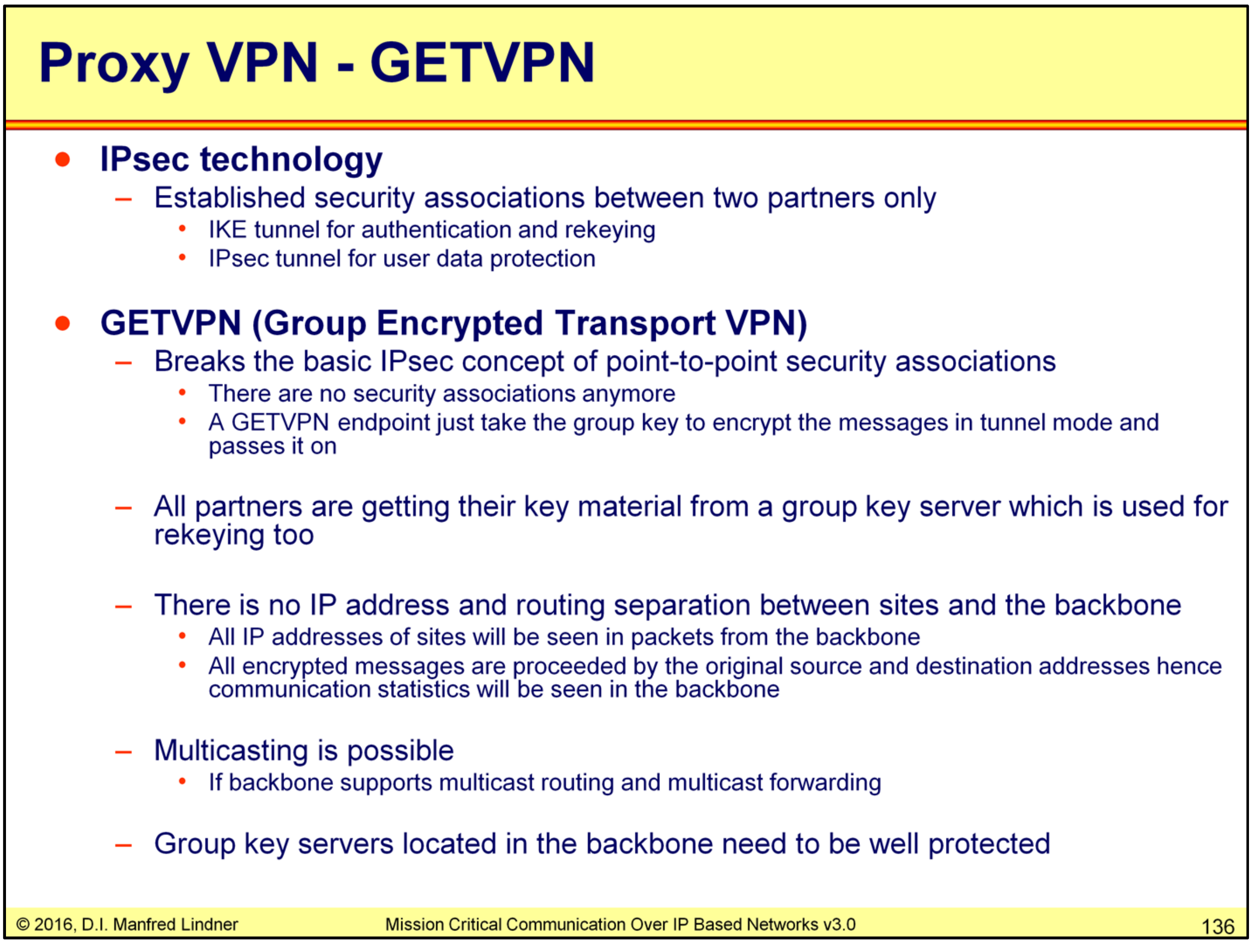
On demand SC for unicast traffic

Overlay/Private Addresses



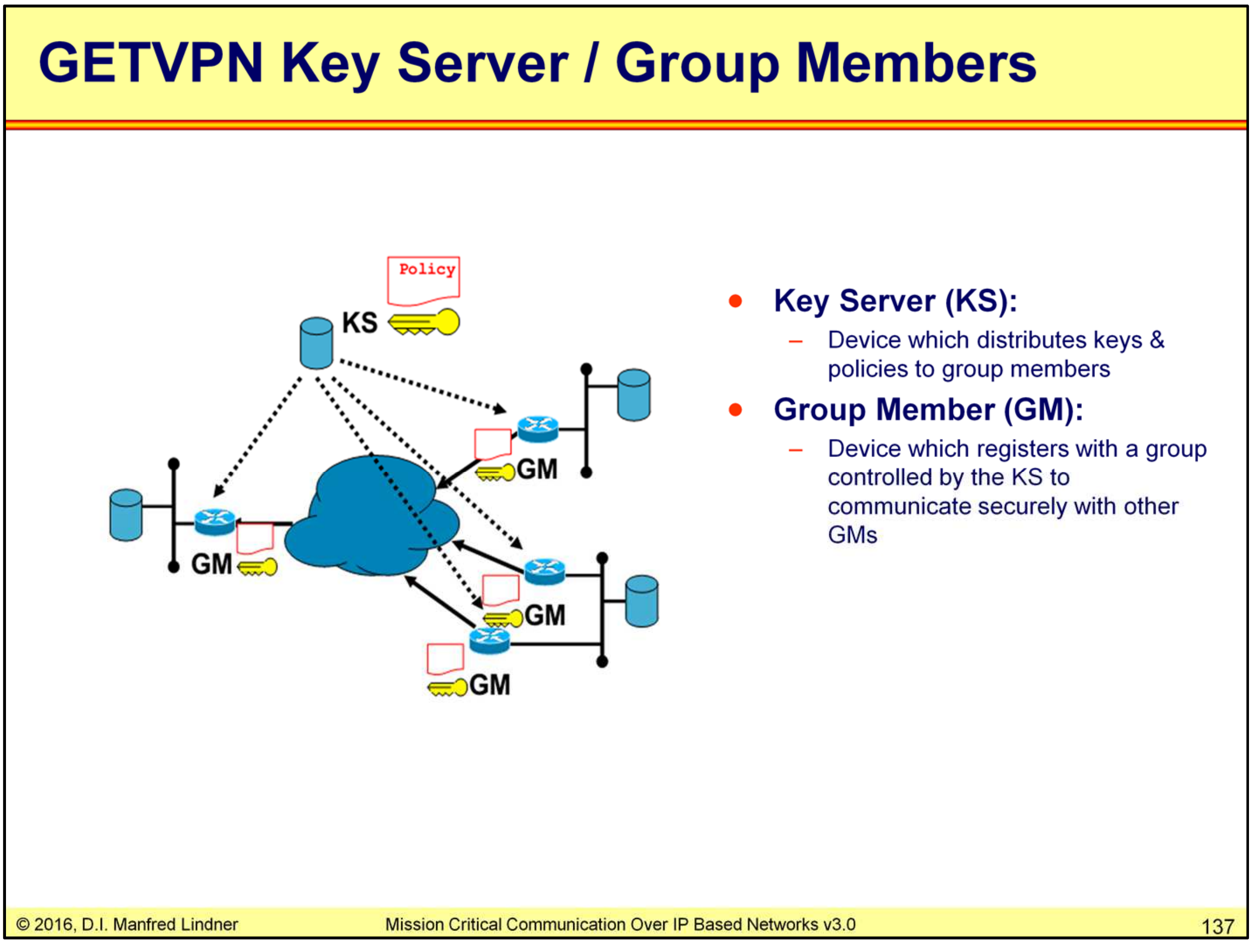
# Agenda

- **Introduction**
- **Network Operational Model**
- **High Availability**
- **QoS**
- **VPN Technologies**
  - Introduction IT-Security
  - VPN Types
  - MPLS, MPLS-VPN
  - IPsec VPN
  - DMVPN
  - GETVPN
- **Multicasting**
- **Summary**



# Proxy VPN - GETVPN

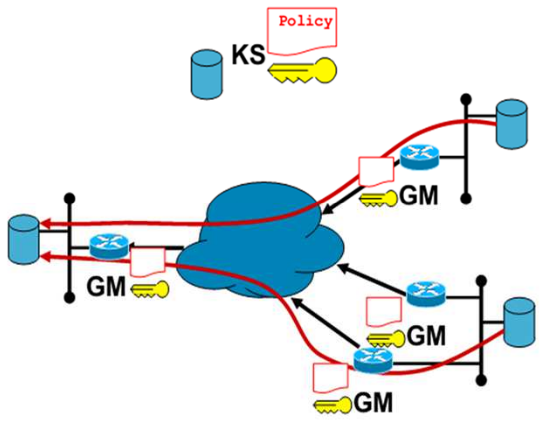
- **IPsec technology**
  - Established security associations between two partners only
    - IKE tunnel for authentication and rekeying
    - IPsec tunnel for user data protection
- **GETVPN (Group Encrypted Transport VPN)**
  - Breaks the basic IPsec concept of point-to-point security associations
    - There are no security associations anymore
    - A GETVPN endpoint just take the group key to encrypt the messages in tunnel mode and passes it on
  - All partners are getting their key material from a group key server which is used for rekeying too
  - There is no IP address and routing separation between sites and the backbone
    - All IP addresses of sites will be seen in packets from the backbone
    - All encrypted messages are proceeded by the original source and destination addresses hence communication statistics will be seen in the backbone
  - Multicasting is possible
    - If backbone supports multicast routing and multicast forwarding
  - Group key servers located in the backbone need to be well protected



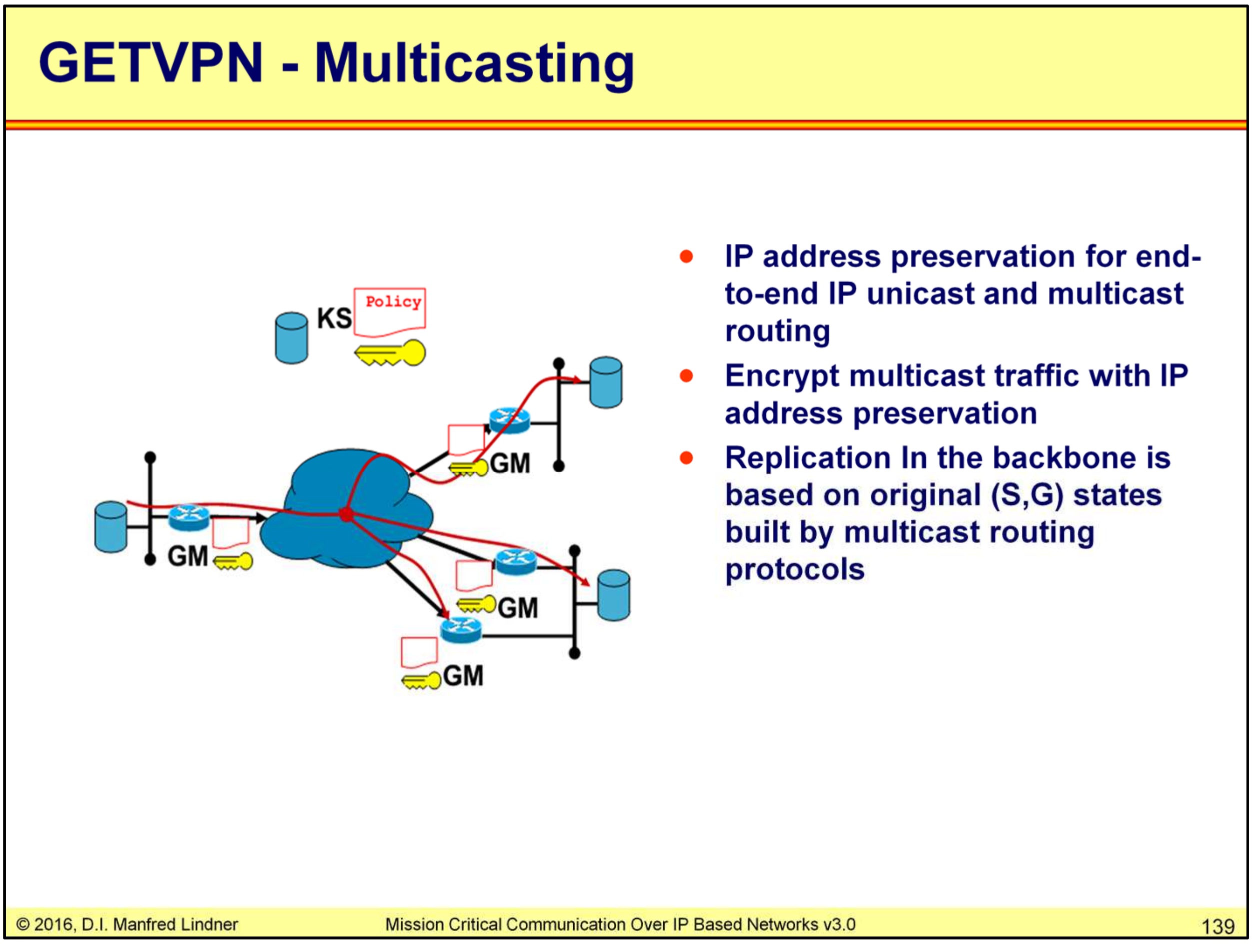
# GETVPN Key Server / Group Members

- **Key Server (KS):**
  - Device which distributes keys & policies to group members
- **Group Member (GM):**
  - Device which registers with a group controlled by the KS to communicate securely with other GMs

# GETVPN - Security Protection



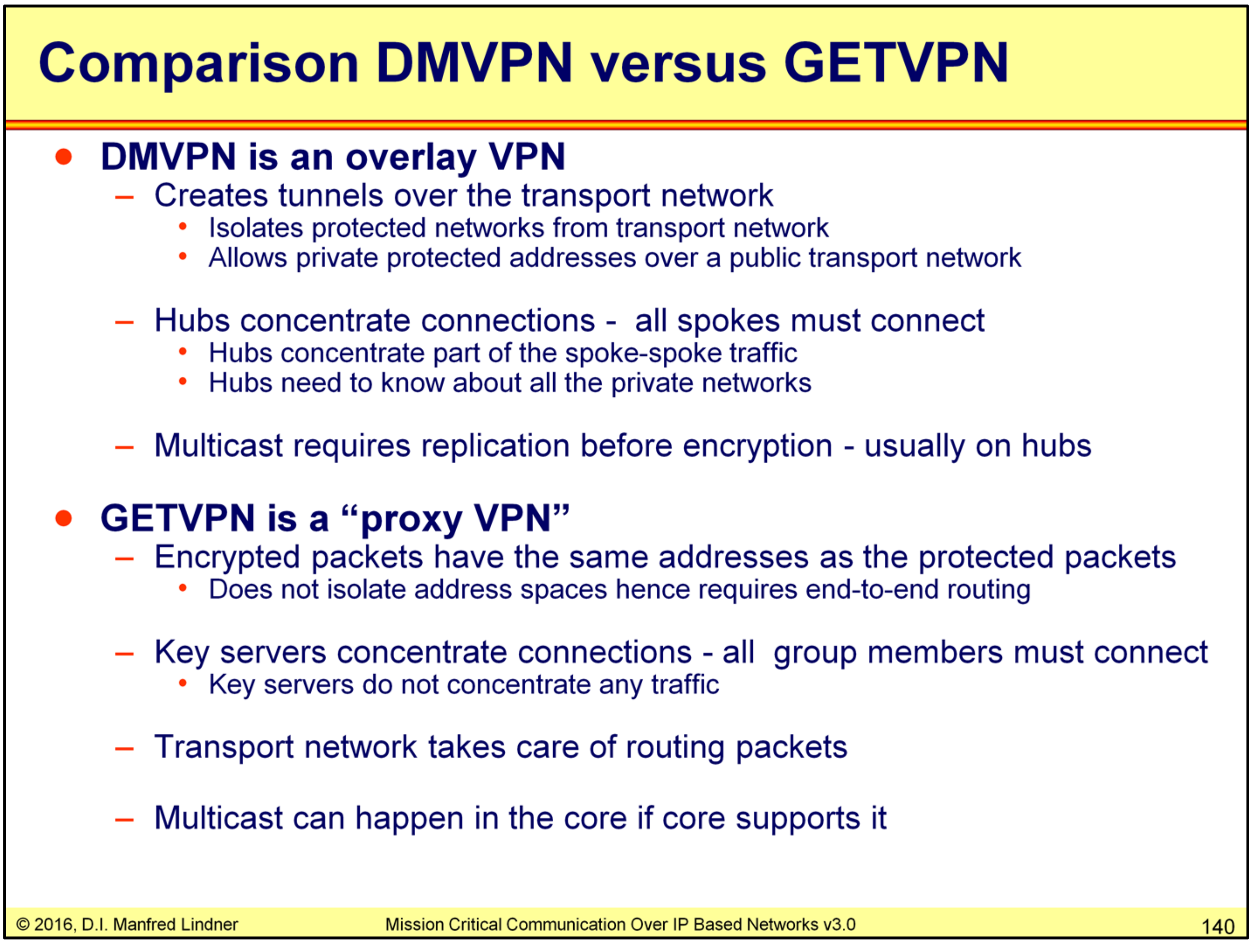
- Receiver does not know the potential encryption sources
- Receiver assumes that legitimate group members obtain Traffic Encryption Key from key server for the group
- Receiver can authenticate the group membership



# GETVPN - Multicasting

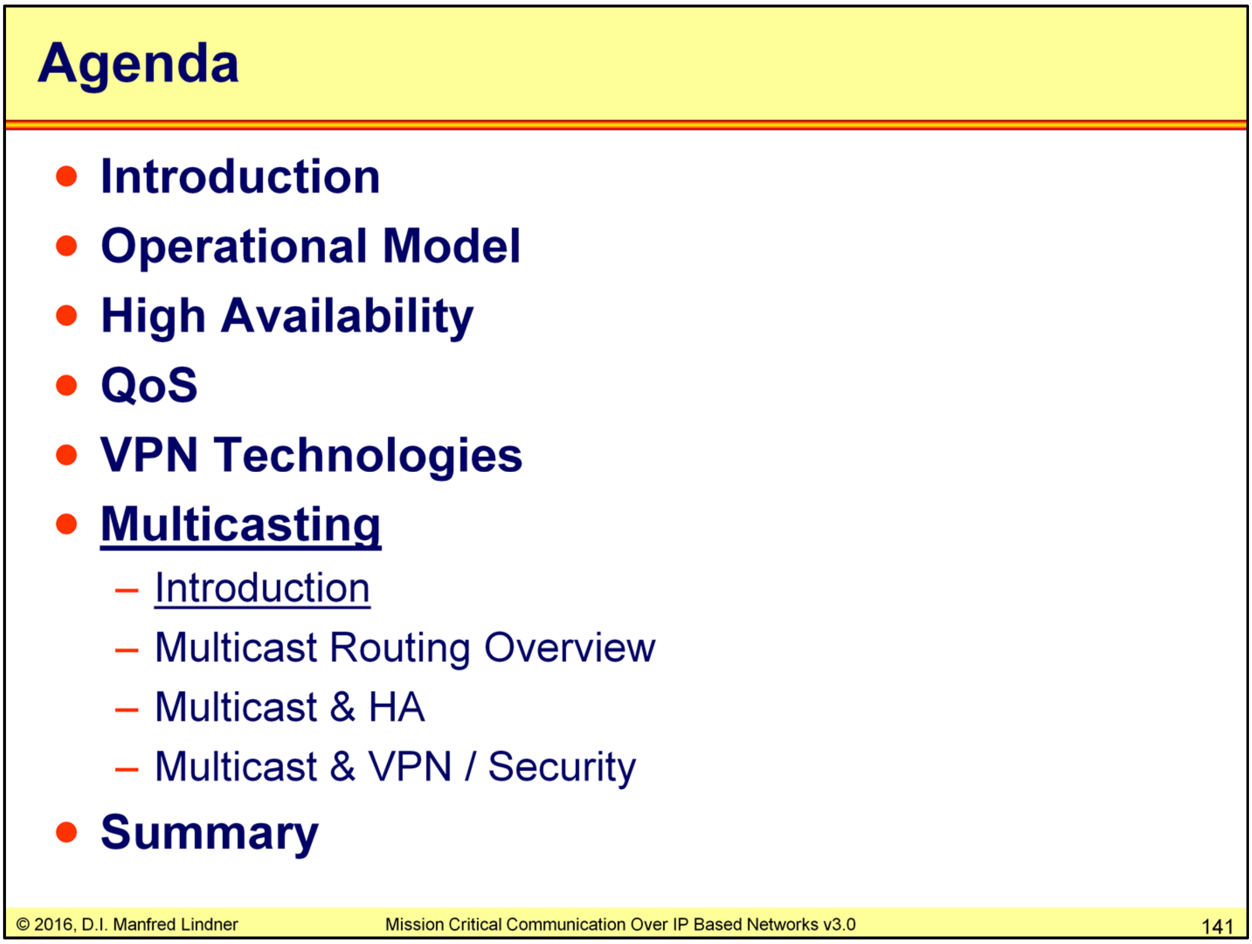
- IP address preservation for end-to-end IP unicast and multicast routing
- Encrypt multicast traffic with IP address preservation
- Replication In the backbone is based on original (S,G) states built by multicast routing protocols

The slide shows how a single multicast message is passed on along the multicast distribution tree in the backbone network. Of course in such a case an end-to-end multicast routing environments is necessary (meaning IGMP multicast group membership protocol and PIM-SM multicast routing protocol).



## Comparison DMVPN versus GETVPN

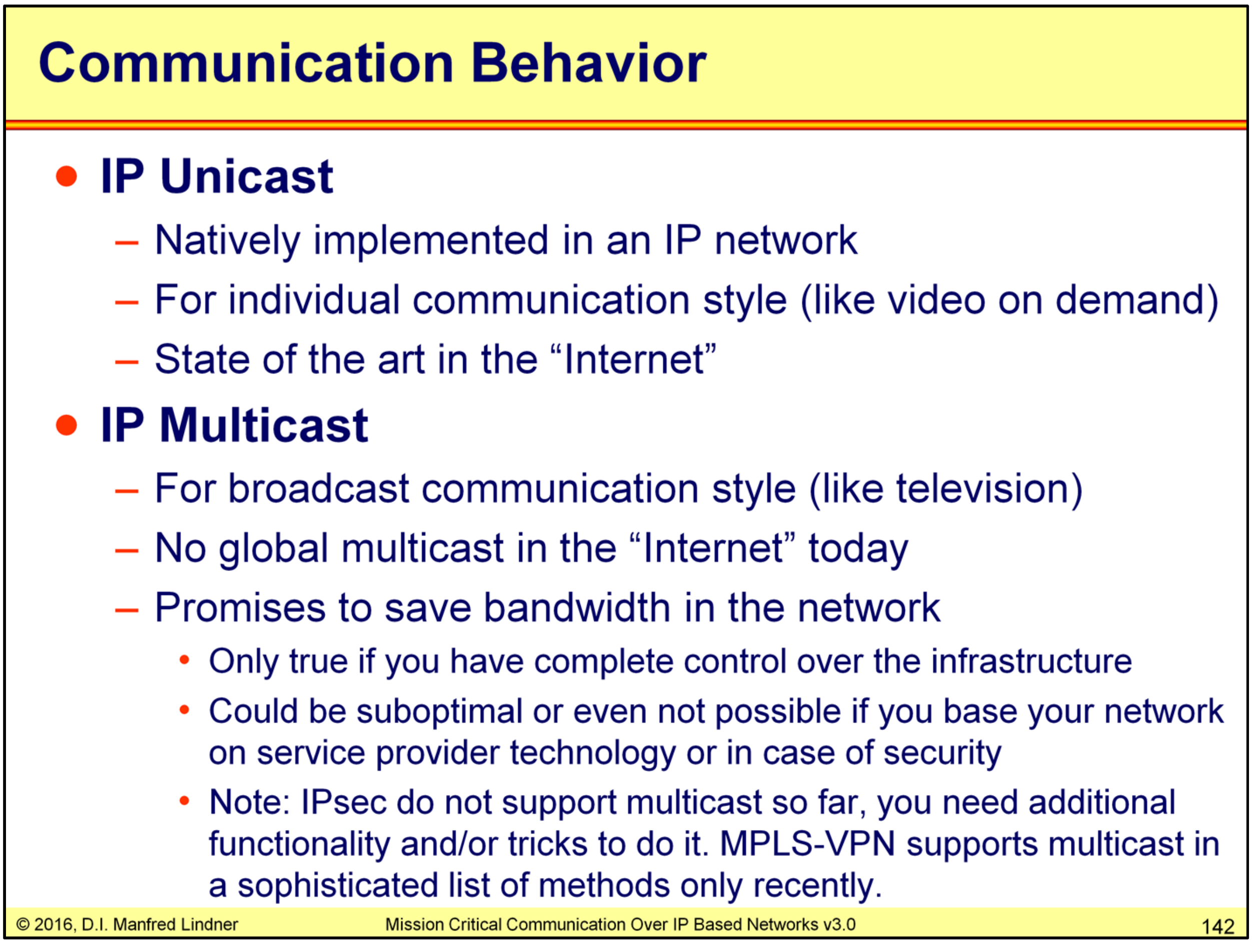
- **DMVPN is an overlay VPN**
  - Creates tunnels over the transport network
    - Isolates protected networks from transport network
    - Allows private protected addresses over a public transport network
  - Hubs concentrate connections - all spokes must connect
    - Hubs concentrate part of the spoke-spoke traffic
    - Hubs need to know about all the private networks
  - Multicast requires replication before encryption - usually on hubs
- **GETVPN is a “proxy VPN”**
  - Encrypted packets have the same addresses as the protected packets
    - Does not isolate address spaces hence requires end-to-end routing
  - Key servers concentrate connections - all group members must connect
    - Key servers do not concentrate any traffic
  - Transport network takes care of routing packets
  - Multicast can happen in the core if core supports it



## Agenda

- Introduction
- Operational Model
- High Availability
- QoS
- VPN Technologies
- Multicasting
  - Introduction
  - Multicast Routing Overview
  - Multicast & HA
  - Multicast & VPN / Security
- Summary

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 141



## Communication Behavior

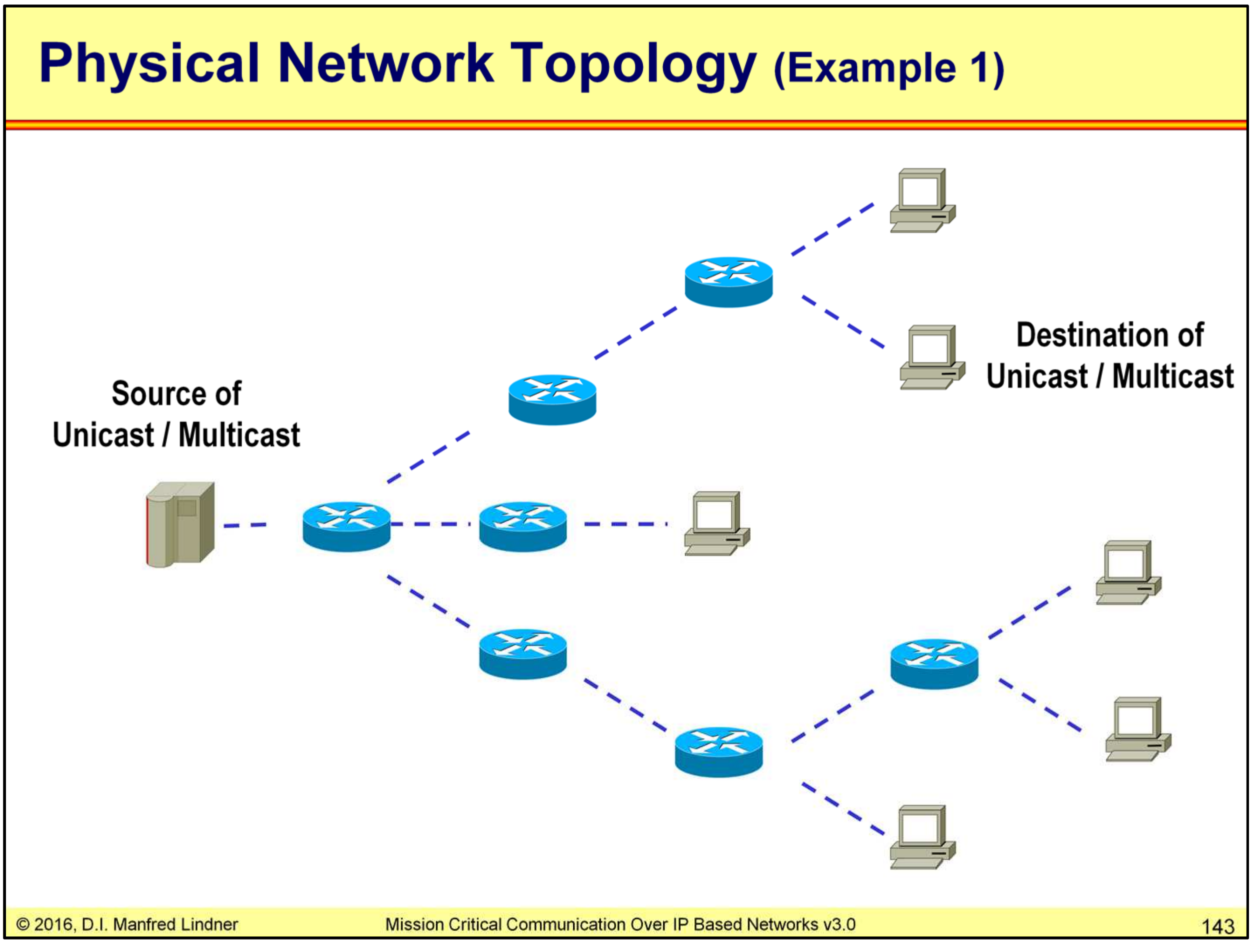
- **IP Unicast**
  - Natively implemented in an IP network
  - For individual communication style (like video on demand)
  - State of the art in the “Internet”
- **IP Multicast**
  - For broadcast communication style (like television)
  - No global multicast in the “Internet” today
  - Promises to save bandwidth in the network
    - Only true if you have complete control over the infrastructure
    - Could be suboptimal or even not possible if you base your network on service provider technology or in case of security
    - Note: IPsec do not support multicast so far, you need additional functionality and/or tricks to do it. MPLS-VPN supports multicast in a sophisticated list of methods only recently.

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 142

Unicast is the usual style for individual communication between two processes from one source to one destination. It does not matter if communication is unidirectional or bidirectional. It is natively implemented in any IP network by unicast destination-based IP routing. Because unicast is per se a point-to-point relationship any protocol providing reliability by retransmissions of lost messages (like TCP) and any security protocol acting on point-to-point security relationship (like IPsec) providing privacy and integrity can be applied on top of it.

Multicast is an optional style for unidirectional communication from one sender process to a group of receiver processes. The goal is to save communication bandwidth in the network by sending out a message for the group just once and the network will duplicate the message at appropriate points in the network in order that every multicast group member gets exactly one copy of the original message. The decision for forwarding a multicast message by an IP multicast enabled router is based on the source address (from where did the message come) and the knowledge about where group members addressed by the multicast group address are located in the network. Compared to IP unicast routing it is “routing on the head”. Multicast routing tries to find the best interface to the source! On the other hand, traditional unicast routing wants to determine the best interface to the destination.

Because of the multicast style putting a reliable protocol like TCP on top of it is not possible. So IP multicast can provide a best-effort service only. Adding privacy and integrity by a security protocol is only possible if the security protocol uses group encryption keys (like in GETVPN).



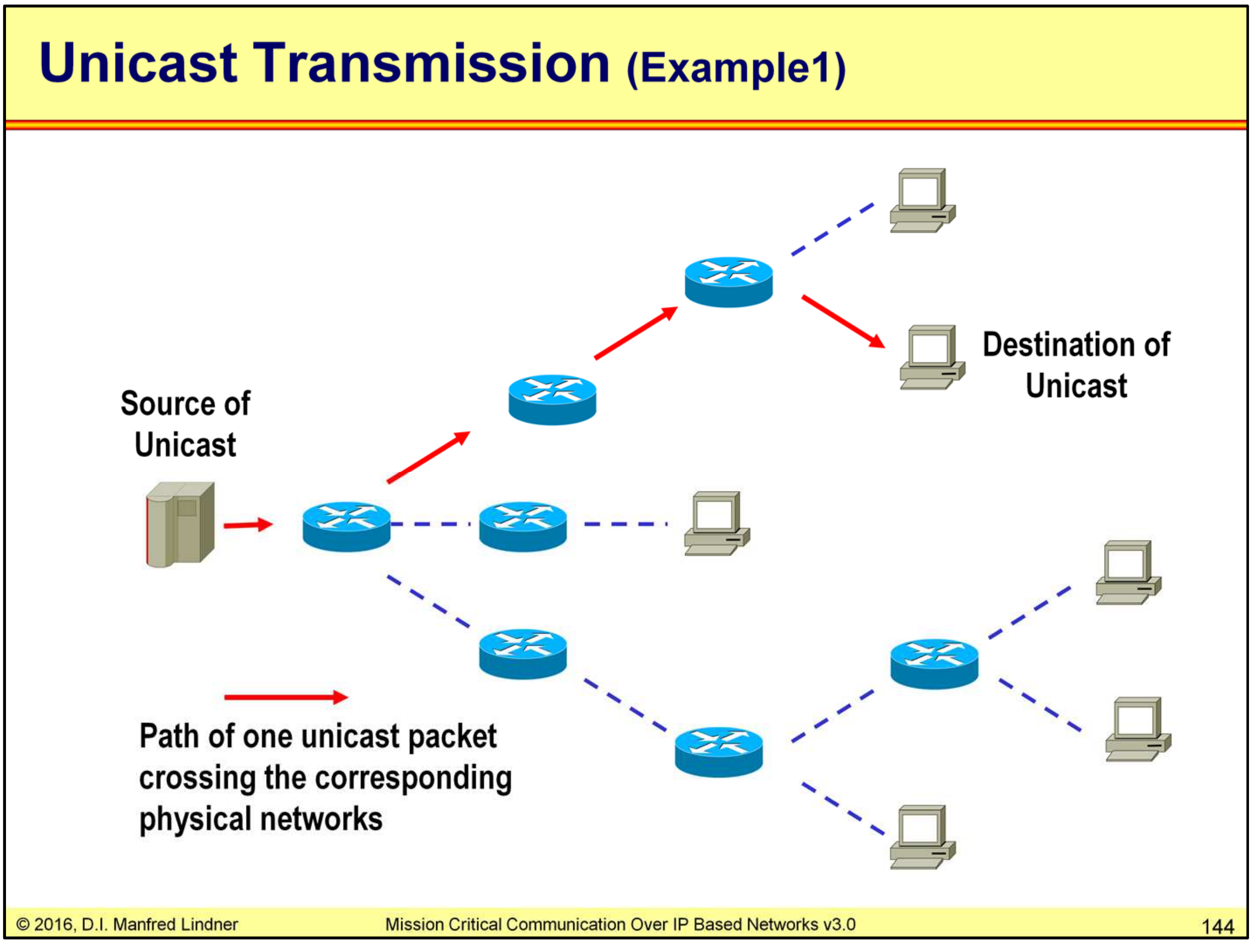
## Physical Network Topology (Example 1)

Source of Unicast / Multicast

Destination of Unicast / Multicast

Mission Critical Communication Over IP Based Networks v3.0 143

We want to discuss the base principles of unicast and multicast using the above physical network topology. The dashed blue lines are the physical network links (e.g. Ethernet or serial).



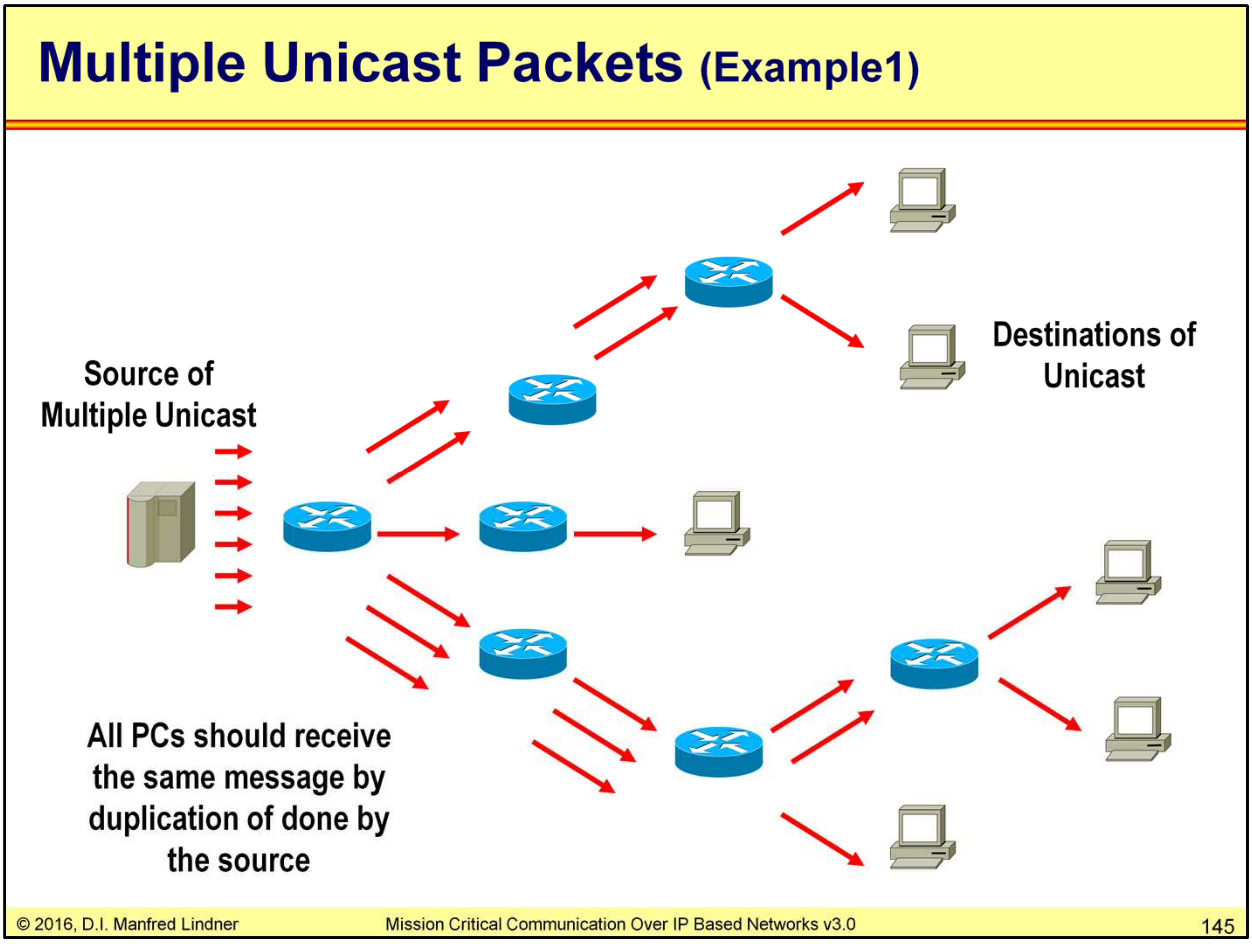
## Unicast Transmission (Example1)

Source of Unicast

Destination of Unicast

Path of one unicast packet crossing the corresponding physical networks

The picture shows how an IP datagram address to a unicast destination address is forwarded hop-by-hop based on the unicast IP routing table.



## Multiple Unicast Packets (Example1)

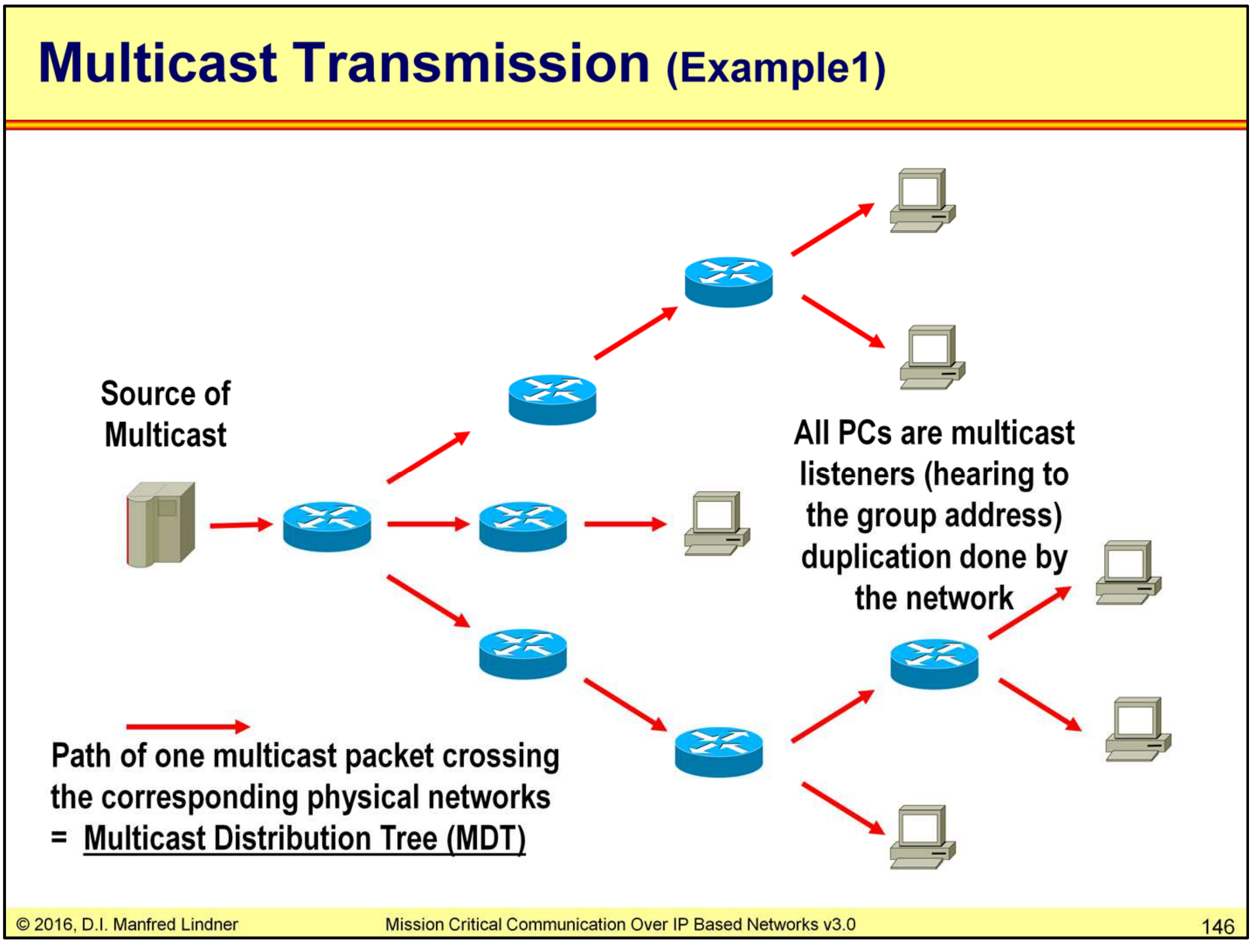
Source of Multiple Unicast

Destinations of Unicast

All PCs should receive the same message by duplication of done by the source

The picture shows how we can emulate a kind of group addressing by using multiple unicast IP datagram generated by the source station and addressed to multiple unicast destination address. In our example all PCs on the right should receive the same information. As you can see parts of the physical network will get more traffic because of this duplication done by the source.

The closest interface to the source is necessary in order to check whether a multicast packet arrived indeed on the upstream interface - an interface which belongs to the MDT. This check is called "Reverse Path Forwarding" (RPF). Using RPF, each router that receives a multicast packet checks (using its routing table) whether this receiving interface is actually the closest to the source. If this is the case then this interface is an upstream interface and the packet can be forwarded (flooded) through all other interfaces. Cisco routers perform a RPF check every 5 seconds by default.



## Multicast Transmission (Example1)

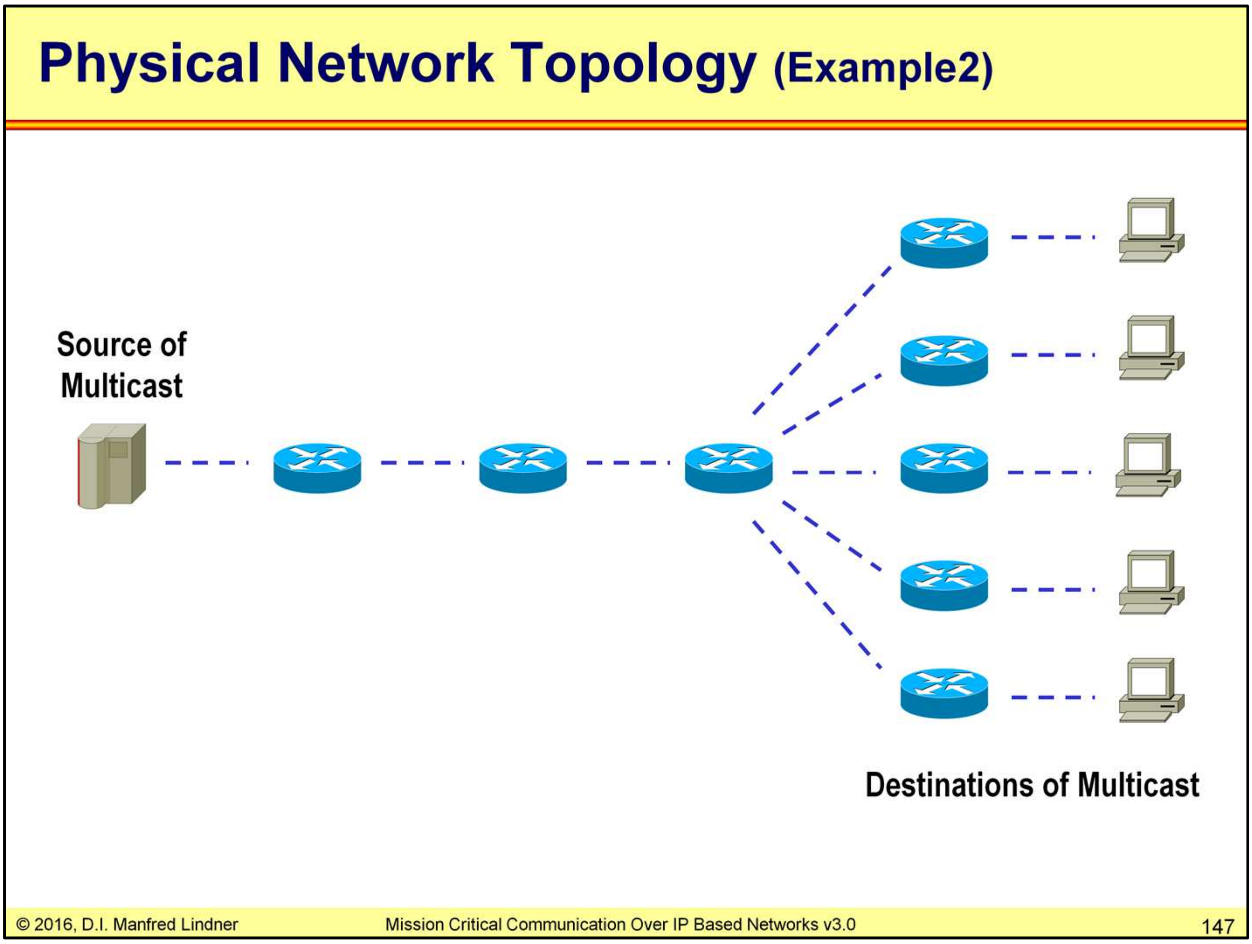
Source of Multicast

All PCs are multicast listeners (hearing to the group address) duplication done by the network

Path of one multicast packet crossing the corresponding physical networks = Multicast Distribution Tree (MDT)

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 146

If multicast transport is enabled in the network a single IP datagram generated by the source station and addressed to the multicast group address will just follow the multicast distribution tree (MDT) from the source to all multicast listeners. The duplication is done by the multicast routers whenever a fork in the occurs at that tree. The red arrows in the picture represent the multicast distribution tree. Multicast routing techniques will establish a MDT based on the knowledge first where receivers for a given multicast group address are located (the domain of local IGMP – Internet Group Management Protocol) and second if the multicast IP datagram arrives on an interface which is on the shortest path from the source to the corresponding router (domain of PIM – Protocol Independent Multicast routing protocols).



## Physical Network Topology (Example2)

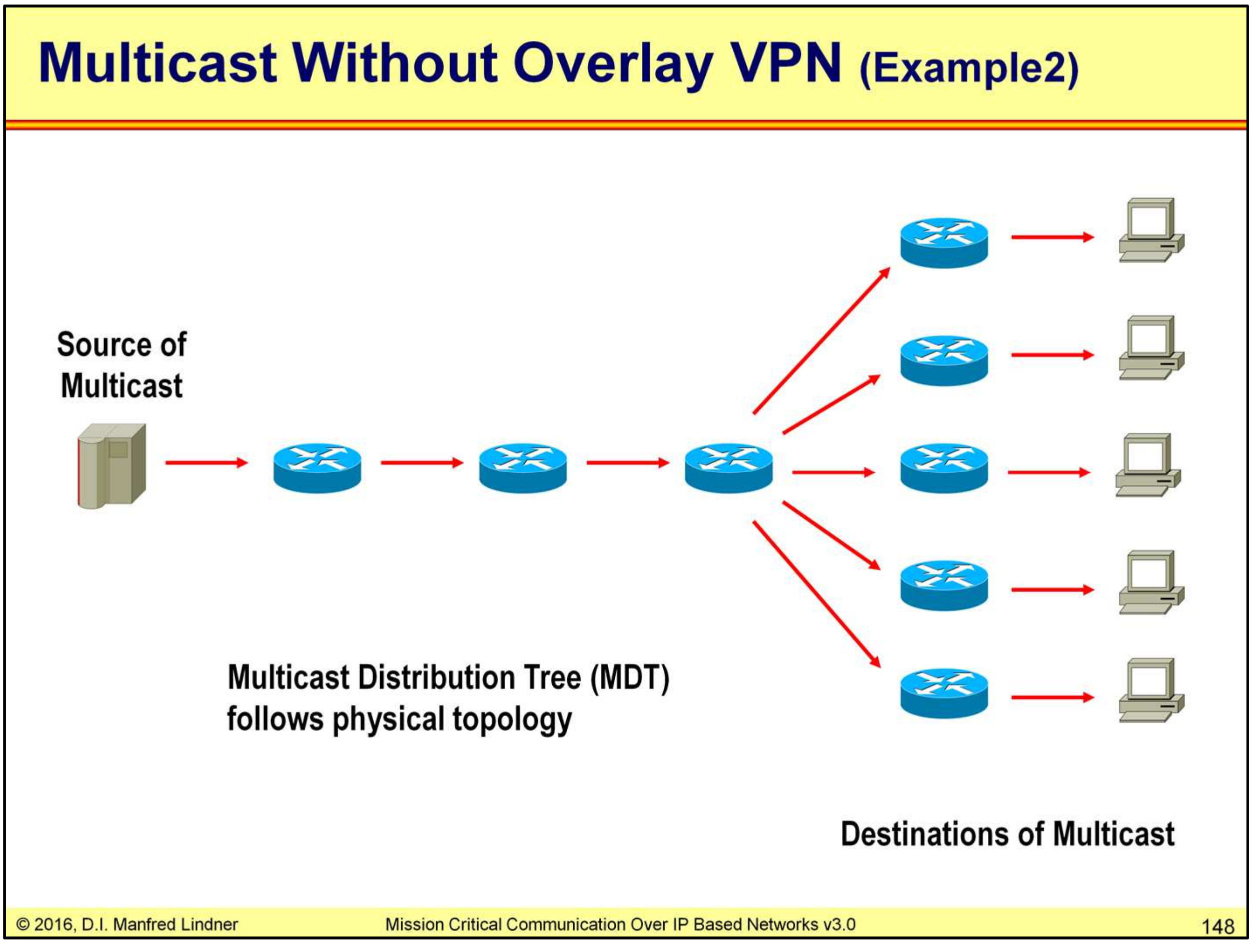
© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 147

For the following discussions we want to use another physical network topology together with an overlay network. Reason for that overlay maybe a network unable to transport multicast or security.

Adding privacy and integrity for multicast by means of a security protocol is only possible if the security protocol uses either group encryption keys (like in GETVPN) or multicast traffic is forced into GRE-IPsec tunnel as described already in the routing sections of IPsec VPN technologies

On the other side, if you create - caused by such requirements - an overlay VPN on your own like usage of DMVPN or GRE-IPsec, IP multicast may be supported. But even if it is supported in most cases it will be suboptimal regarding saving of bandwidth, because an overlay has no knowledge of the underlying network topology. For example imagine a meshed network offering potential for bandwidth saving by multicast technology. If you but a star-shaped overlay topology on top of it (that is what DMVPN is doing) it is clear that has limited benefit. It will save maybe some bandwidth in the overlay tunnels but that does not mean that is an advantage in the underlay network. Remember that tunnels leaving from a single end-point (e.g. router as GRE-IPsec VPN overlay entry point) will cause duplication of multicast on every tunnel even if this router is connected just with one physical link to the IP WAN.

The following slides show first the difference between MDT (Multicast Distribution Tree) following the physical topology and MDT following the logical topology caused by overlay and second problems involved (duplication on just a single physical link) when applying multicast on top of an overlay VPN.



## Multicast Without Overlay VPN (Example2)

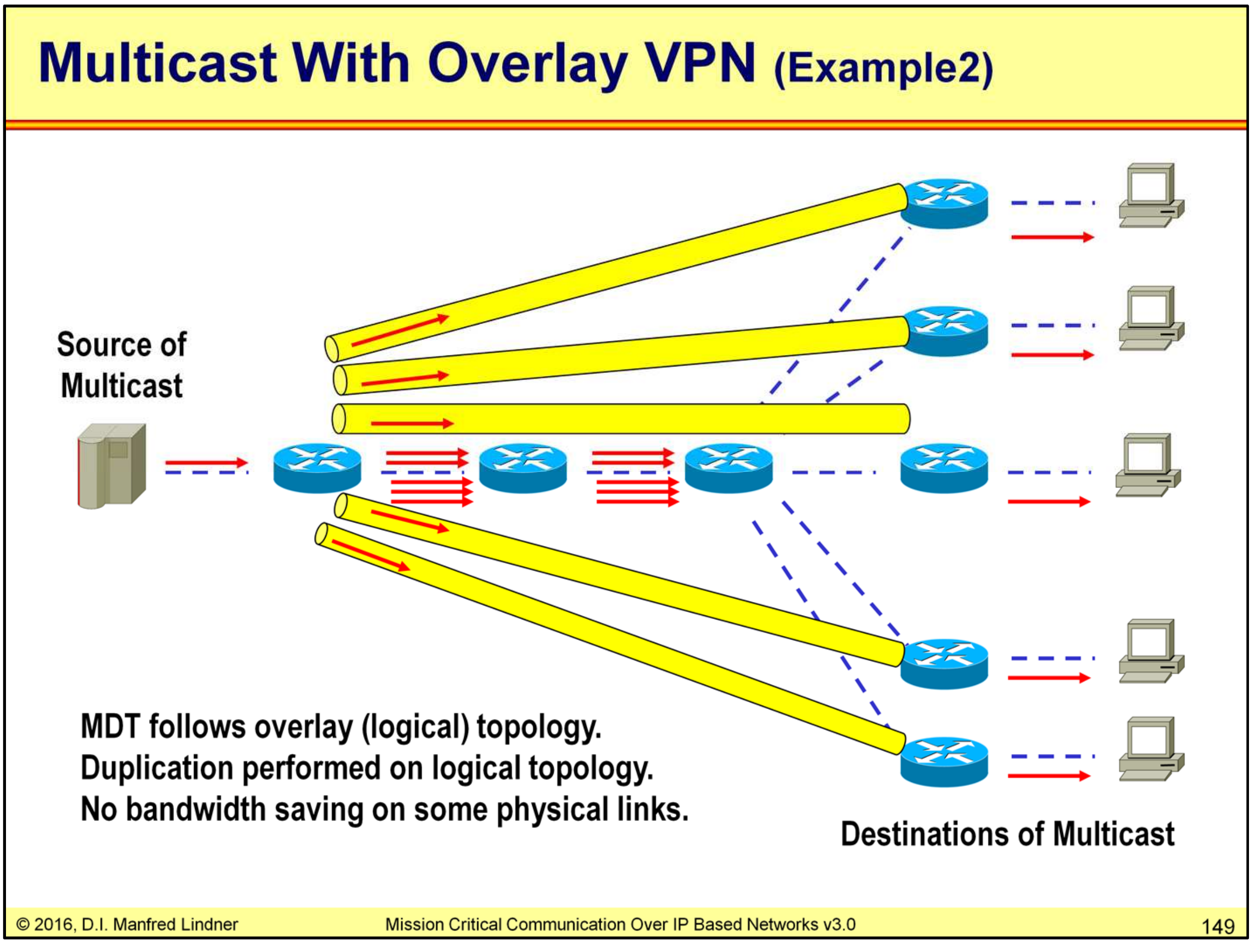
Source of Multicast

Multicast Distribution Tree (MDT) follows physical topology

Destinations of Multicast

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 148

This pictures shows a MDT (Multicast Distribution Tree) following the physical topology and appropriate duplication at the fork router two hops apart from the multicast listeners.

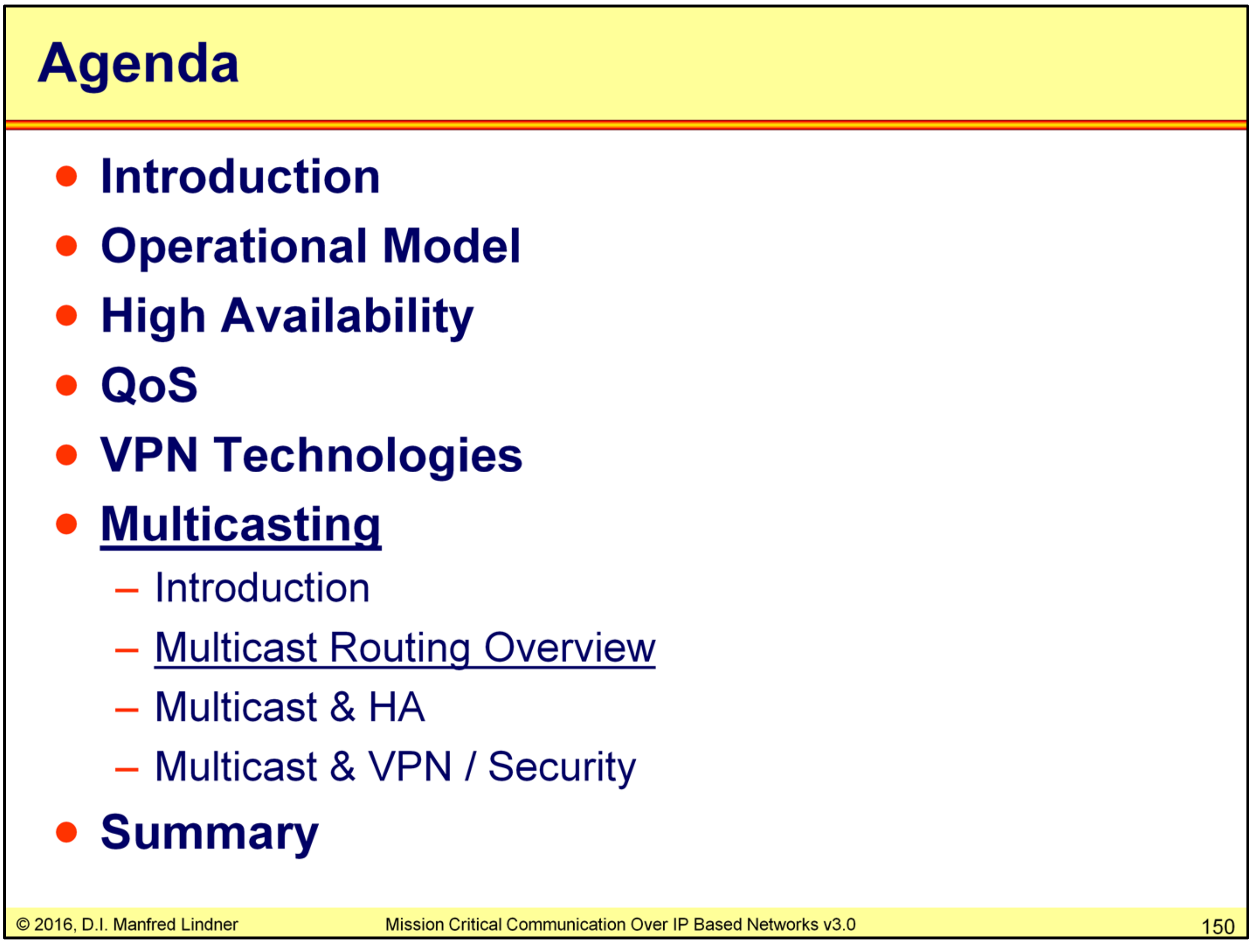


## Multicast With Overlay VPN (Example2)

MDT follows overlay (logical) topology.
Duplication performed on logical topology.
No bandwidth saving on some physical links.

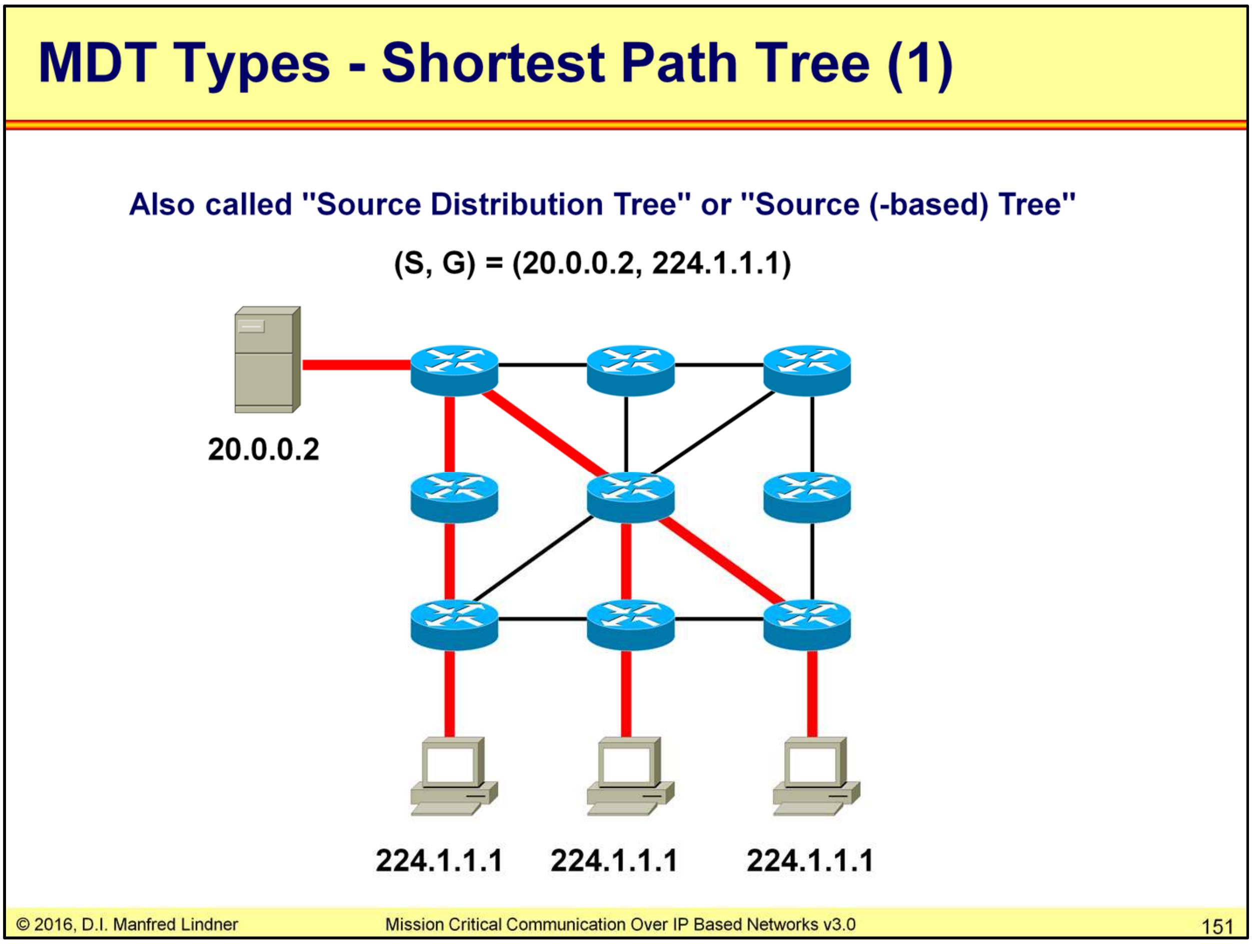
© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 149

This pictures shows a MDT (Multicast Distribution Tree) following the logical (overlay) topology. Duplication is done at the logical fork router (entry point into overlay tunnels). This causes in the physical network a duplication of a single multicast datagram and hence wastes bandwidth on the links between first and third router (seen from the source).



## Agenda

- Introduction
- Operational Model
- High Availability
- QoS
- VPN Technologies
- Multicasting
  - Introduction
  - Multicast Routing Overview
  - Multicast & HA
  - Multicast & VPN / Security
- Summary

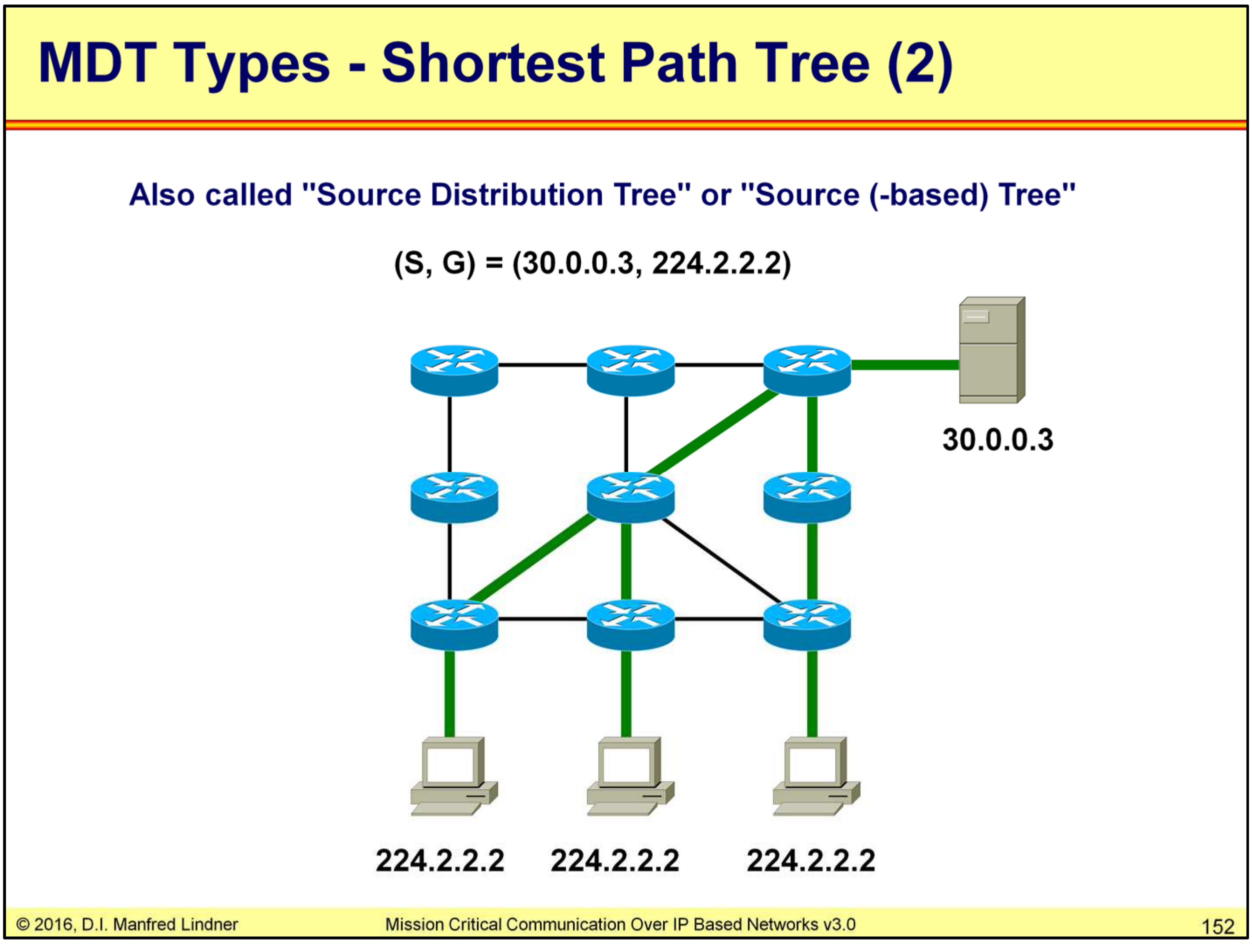


# MDT Types - Shortest Path Tree (1)

Also called "Source Distribution Tree" or "Source (-based) Tree"

(S, G) = (20.0.0.2, 224.1.1.1)

"Shortest Path Trees" (SPT) are also called "Source Distribution Trees" or "Source Trees". The basic idea is that a separate tree is created for each single source. The picture above shows only one source based tree. This distribution method consumes much memory in the involved routers—of order O(S · G)—but it provides optimal paths from source to all receivers and minimizes delay.



## MDT Types - Shortest Path Tree (2)

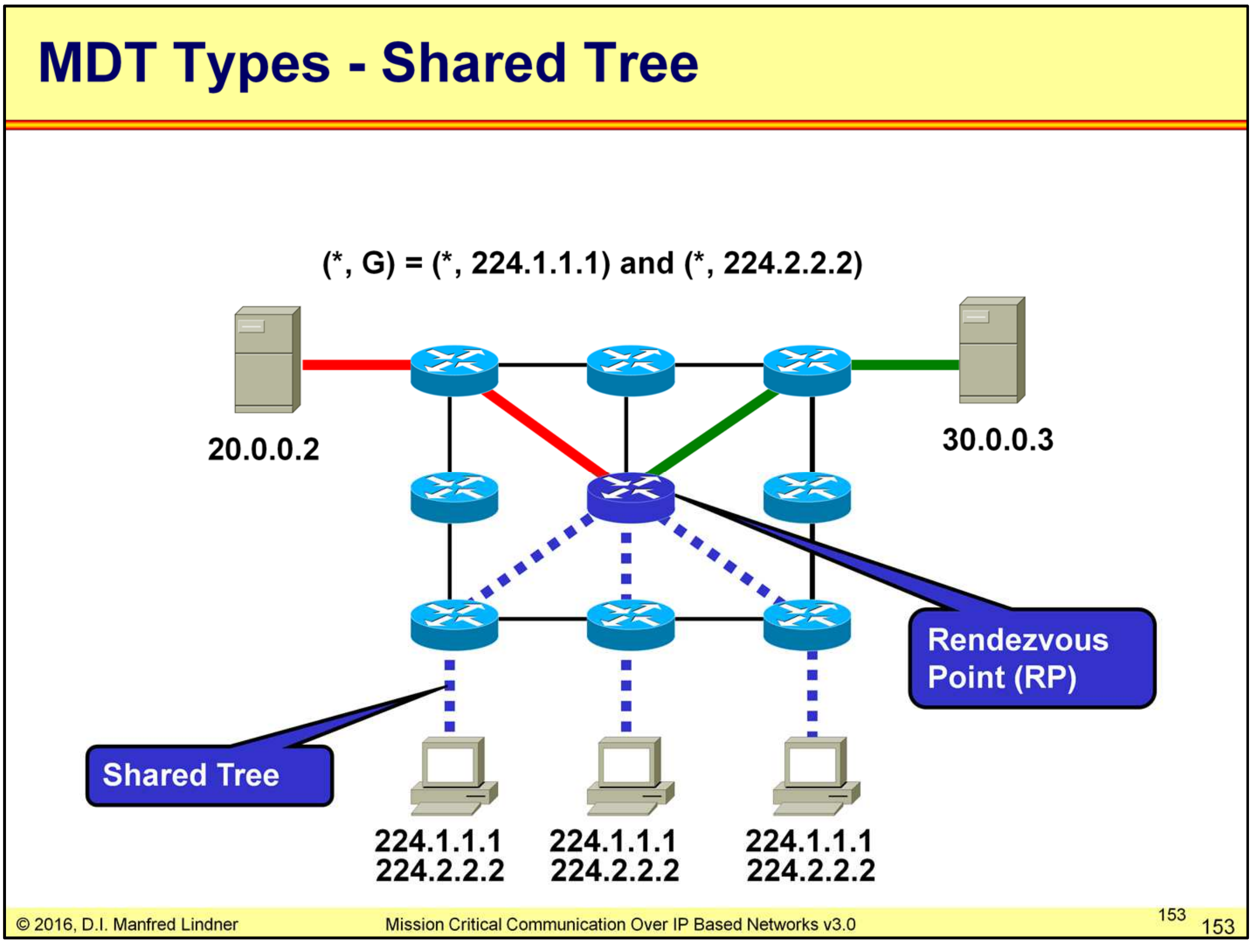
Also called "Source Distribution Tree" or "Source (-based) Tree"

(S, G) = (30.0.0.3, 224.2.2.2)

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 152

This picture shows another SPT. Each SPT is identified by a pair of addresses, that is, the unicast source address S and the multicast group address G, thus (S, G). The SPT principle is the most implemented method, found in DVMRP, PIM-DM, and other protocols. Note that each router must maintain state information for each (S, G) combination, including timers, interface lists, etc (suppose there are hundreds of sources and hundreds of groups…)

The main point here is that a separate SPT is built for every source "S" sending to group "G". Traffic is forwarded via the shortest path from the Source!

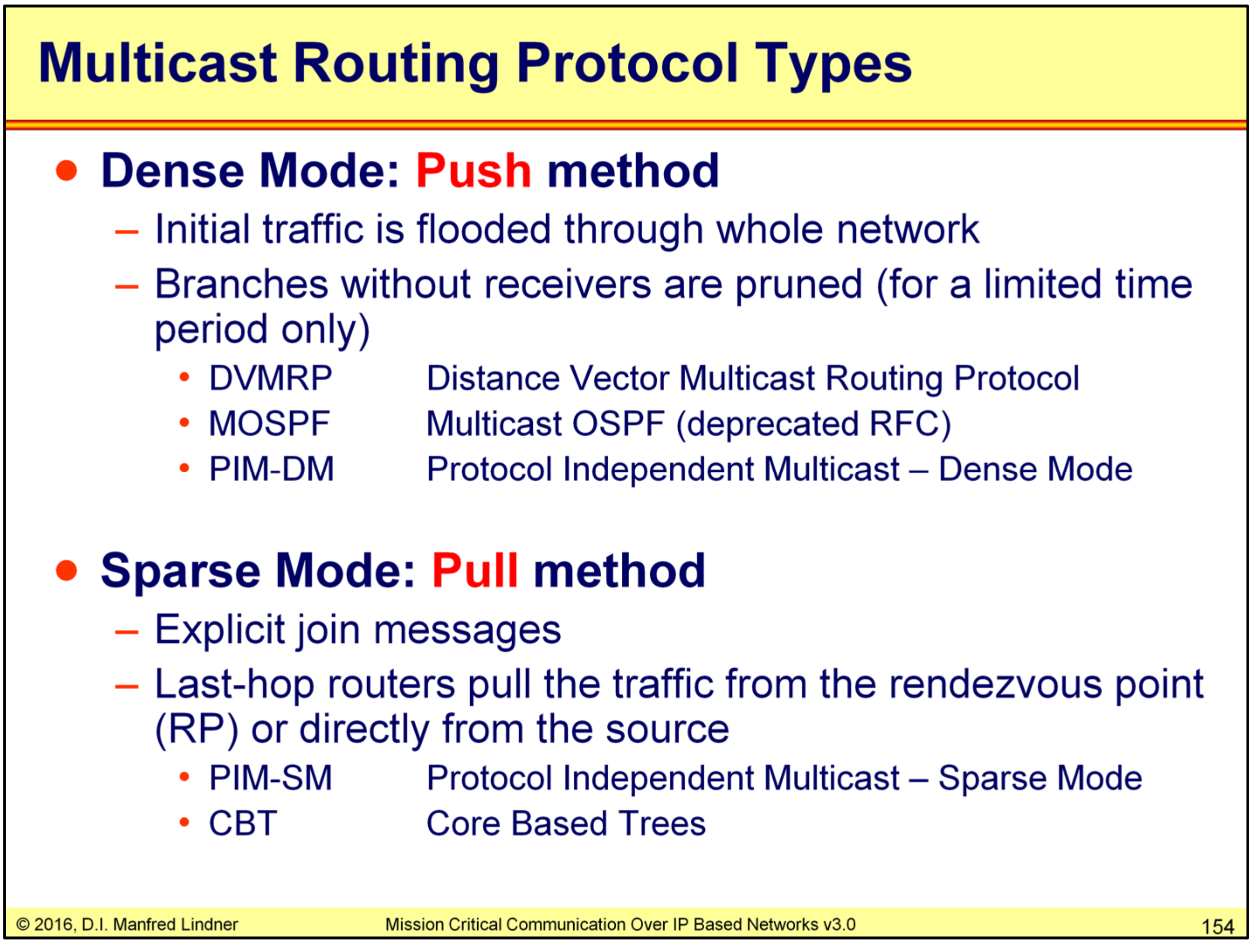


## MDT Types - Shared Tree

(*, G) = (*, 224.1.1.1) and (*, 224.2.2.2)

Mission Critical Communication Over IP Based Networks v3.0 153 153

Shared trees utilize a so-called "Rendezvous Point" (RP), which distributes multicast traffic to its attached receivers. The idea is similar as the supermarket principle: "Customers should not have to visit every manufacturer but rather buy everything at the shop around the corner." In this sense, the RP acts as supermarket and offers multicast traffic from several sources. Typically, each RP is a leaf of a SPT, which is rooted at a source. That is, the shared tree principle is mostly used in combination with a SPT. Shared trees consume memory of order O(G) but might result in sub-optimal paths from the source to all receivers. Furthermore they may introduce extra delay. Thus, only a clever combination of both SPT and shared trees might be most efficient.



## Multicast Routing Protocol Types

- **Dense Mode: Push method**
  - Initial traffic is flooded through whole network
  - Branches without receivers are pruned (for a limited time period only)

| | |
|---|---|
| DVMRP | Distance Vector Multicast Routing Protocol |
| MOSPF | Multicast OSPF (deprecated RFC) |
| PIM-DM | Protocol Independent Multicast – Dense Mode |

- **Sparse Mode: Pull method**
  - Explicit join messages
  - Last-hop routers pull the traffic from the rendezvous point (RP) or directly from the source

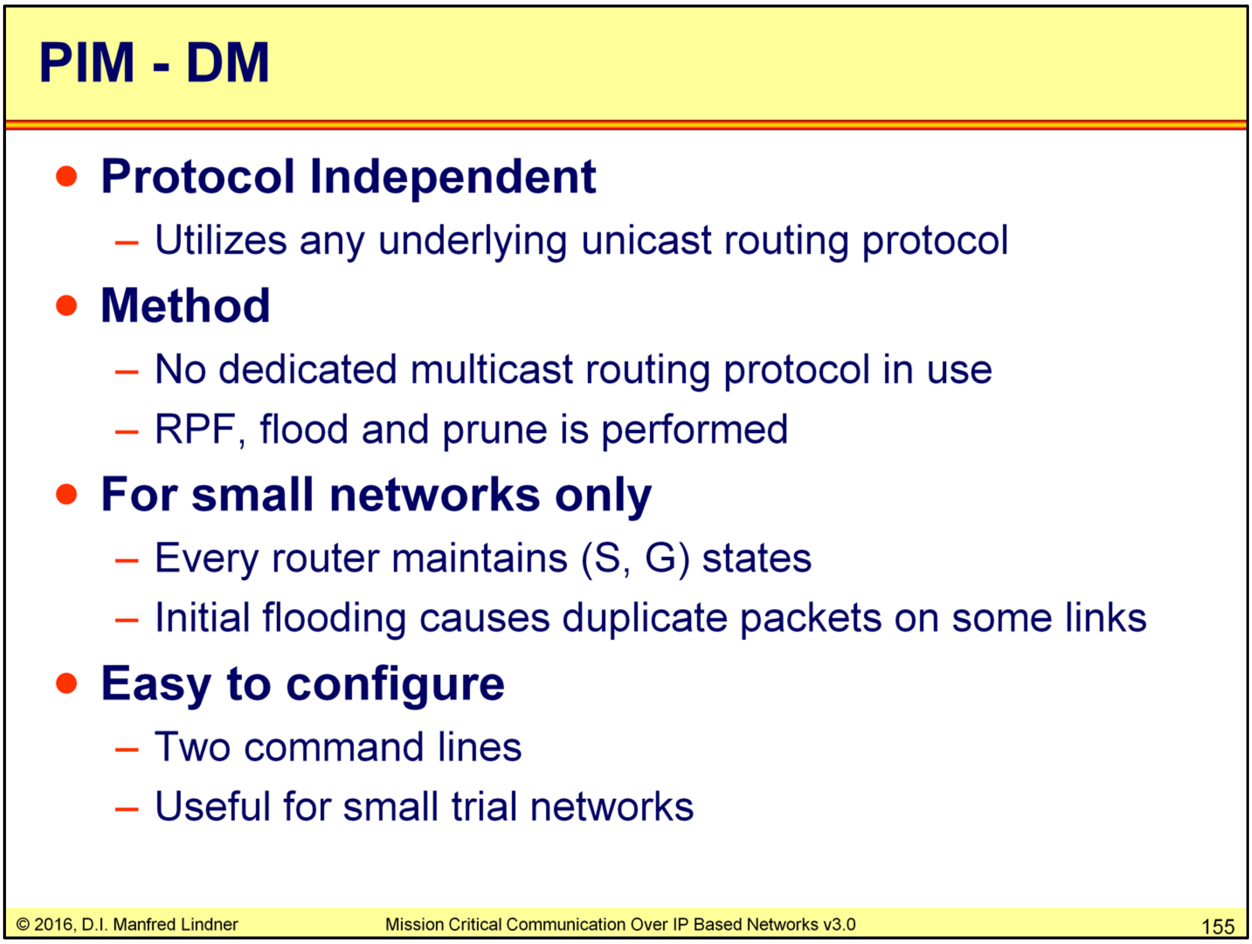
| | |
|---|---|
| PIM-SM | Protocol Independent Multicast – Sparse Mode |
| CBT | Core Based Trees |

© 2016, D.I. Manfred Lindner   Mission Critical Communication Over IP Based Networks v3.0   154

Multicast routing protocols are either dense mode or sparse mode.

The dense mode principle uses a "push" method to create the distribution tree. Multicast packets are flooded throughout the network and each router creates its outgoing interface list (OIL) using the RPF check and "prune" messages to cut off unnecessary branches of the tree. That is, after the initial flood, branches without receivers are pruned. But after a timeout, traffic is flooded throughout the network again. Typically every 3 minutes a flood and prune occurs.

The sparse mode principle uses the opposite method in that routers which want to be part of the tree must send explicit "join" messages. Thus, the sparse mode supports a "pull" method for tree establishment. Note: Branches without receivers never get any multicast traffic!



## PIM - DM

- **Protocol Independent**
  - Utilizes any underlying unicast routing protocol
- **Method**
  - No dedicated multicast routing protocol in use
  - RPF, flood and prune is performed
- **For small networks only**
  - Every router maintains (S, G) states
  - Initial flooding causes duplicate packets on some links
- **Easy to configure**
  - Two command lines
  - Useful for small trial networks

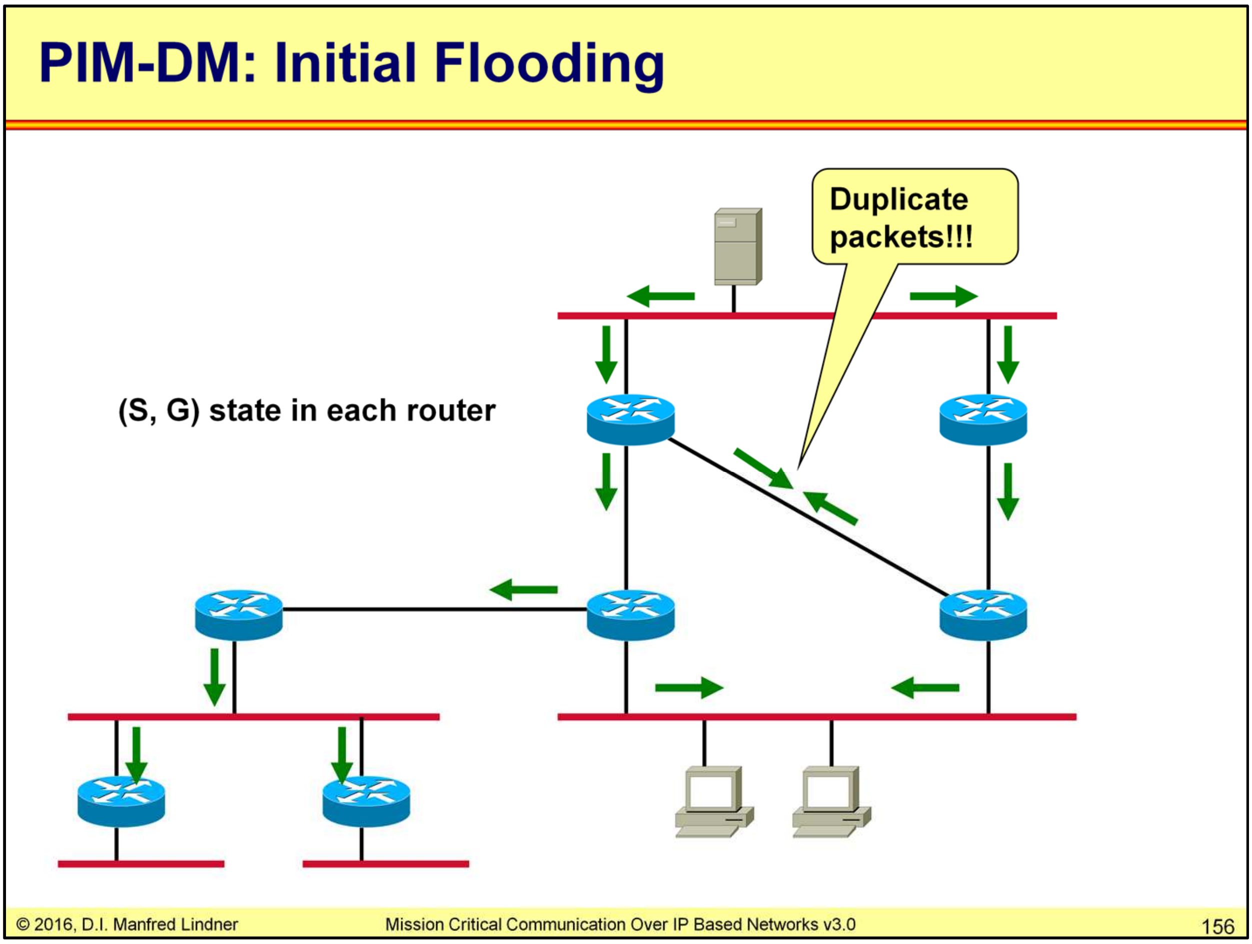
© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 155

The Protocol Independent Multicast - Dense Mode (PIM-DM) supports any underlying unicast routing protocols, including static, RIP, IGRP, EIGRP, IS-IS, BGP, and OSPF.

When a PIM-DM router receives multicast traffic via its (upstream) RPF interface it forwards the multicast traffic to all of its PIM-DM neighbors.

But then, the next-hop routers might receive packets also on non-RPF interfaces! Clearly this silly method results in duplicate packets on some links. These non-RPF flows are normal for the initial flooding of data and will be corrected by a PIM DM pruning mechanism.

Special "assert" messages are used to prune another routers interface. Flood and prune is performed every 3 minutes. If the metric is equal, then the highest IP address on an interface wins.



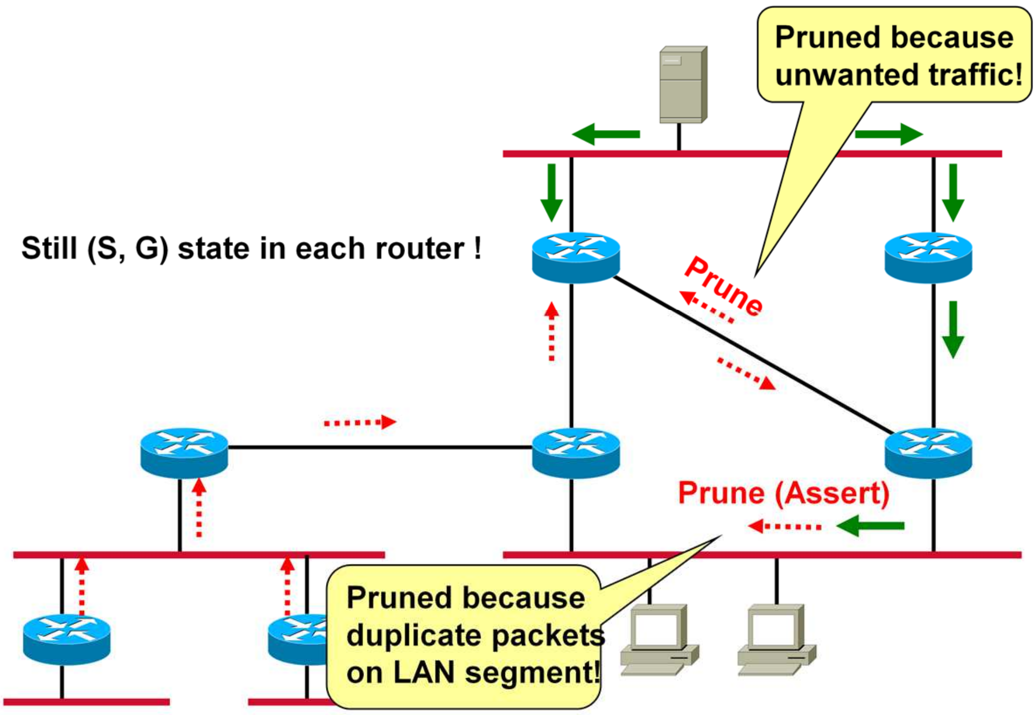
## PIM-DM: Initial Flooding

Duplicate packets!!!

(S, G) state in each router

The example above shows some routers which receive packets on non-RPF interfaces. The routers will discard these packets because only packets received through the upstream interface are considered as good packets. Duplicate packets can occur on some links during the initial flooding of data and will be removed by a PIM DM pruning mechanism, following in the next step. Also note, that each router must maintain a (S, G) state.

# PIM-DM: Pruning

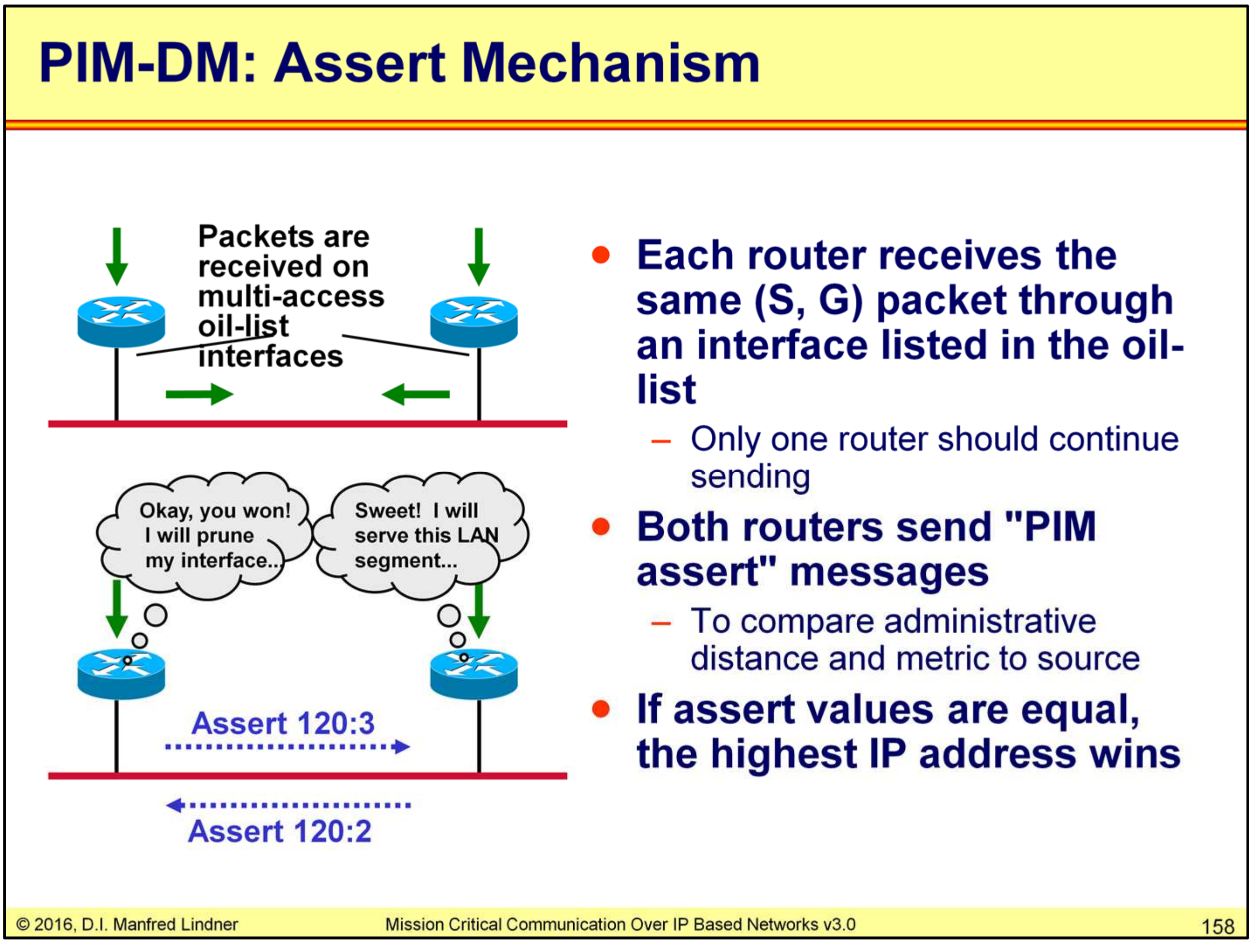


Pruned because unwanted traffic!

Still (S, G) state in each router !

Pruned because duplicate packets on LAN segment!

Pruning occurs after the initial flooding (which is done every 3 minutes by default) and serves for two purposes: First, branches can be cut off when there are no further receivers downstream; secondly, a router can use a so-called assert message to stop another router from sending packets to its non-upstream (i. e. non-RPF-) interface. The latter method forces the other router to prune its own interface. Again: The prune state lasts three minutes by default. Then a new flooding occurs over all links! Note, that each router must still maintain the (S, G) state.



## PIM-DM: Assert Mechanism

Packets are received on multi-access oil-list interfaces
Okay, you won! I will prune my interface...
Sweet! I will serve this LAN segment...
Assert 120:3
Assert 120:2

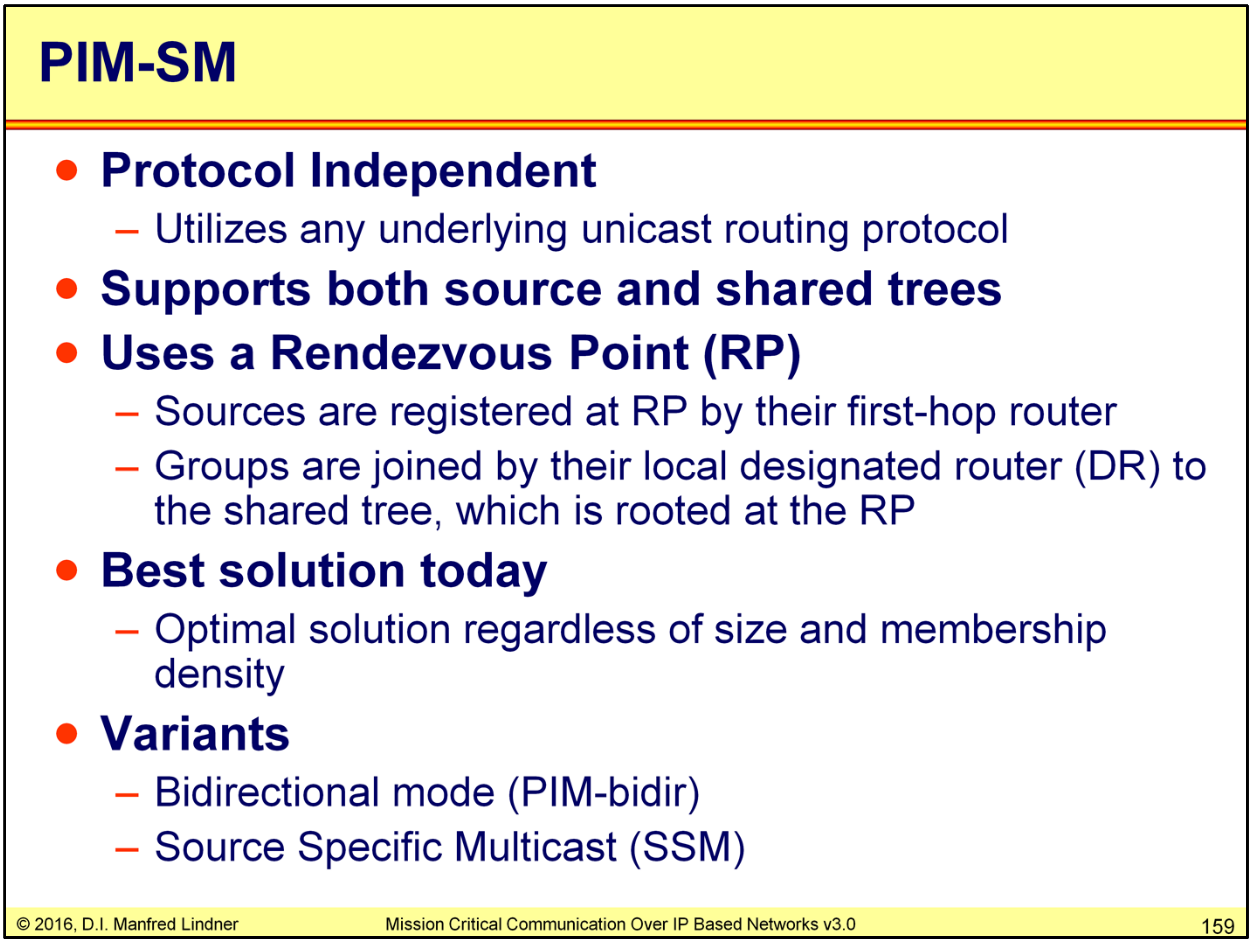
- Each router receives the same (S, G) packet through an interface listed in the oil-list
  - Only one router should continue sending
- Both routers send "PIM assert" messages
  - To compare administrative distance and metric to source
- If assert values are equal, the highest IP address wins

Mission Critical Communication Over IP Based Networks v3.0 158

The PIM assert mechanism is used to eliminate duplicate flows on the same multi-access segment. The assert mechanism is only performed when duplicate packets appear on this link.

When a router receives a (S, G) packet via a multi-access interface which is listed in the (S, G) oil-list, then it will send an assert message, telling the other router a so-called assert value.

The assert value contains both the administrative distance of this router and the metric toward the source. The administrative distance is evidentially the high-order part of this assert value. Obviously the other router sends also an assert message. Now both routers compare these values to determine who has the best path (i. e. lowest value) to the source. If both values are the same, the highest IP address is used as tiebreaker. Losing routers prune their interface, whereas the winning router continues to forward multicast traffic onto the LAN segment.



## PIM-SM

- **Protocol Independent**
  - Utilizes any underlying unicast routing protocol
- **Supports both source and shared trees**
- **Uses a Rendezvous Point (RP)**
  - Sources are registered at RP by their first-hop router
  - Groups are joined by their local designated router (DR) to the shared tree, which is rooted at the RP
- **Best solution today**
  - Optimal solution regardless of size and membership density
- **Variants**
  - Bidirectional mode (PIM-bidir)
  - Source Specific Multicast (SSM)

The Protocol Independent Multicast – Sparse Mode (PIM-SM) has been defined in RFC 2362 and is the most useful multicast protocol today. PIM-SM relies on a explicit pull concept. Traffic is only forwarded to receivers that ask for it (i. e. send a join message).

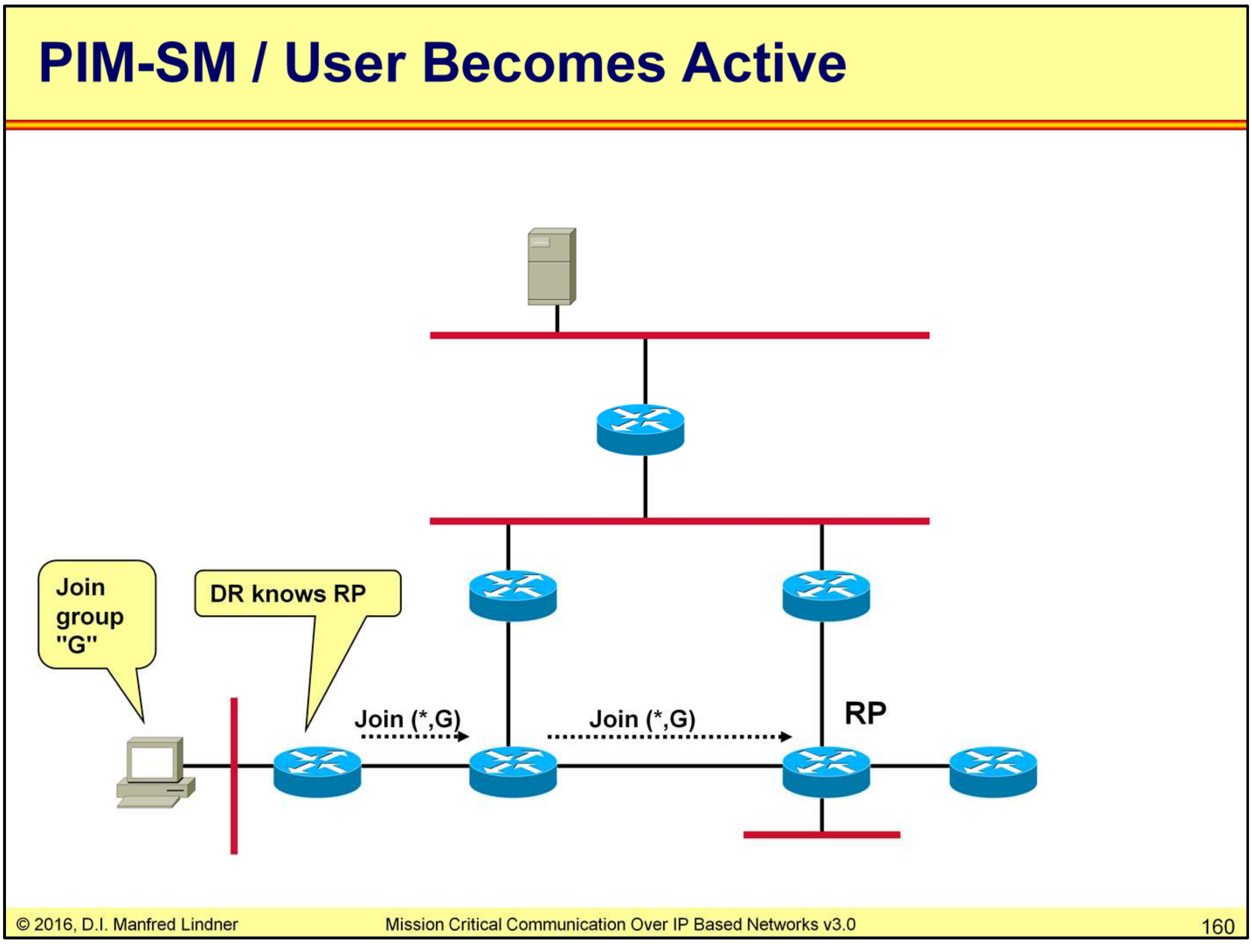
PIM-SM utilizes a Rendezvous Point (RP) which roots a shared tree to the groups. The groups are joined by their local designated router (DR) to this shared tree. Basically, PIM SM uses shared distribution trees, but it may also switch to the source rooted distribution tree.

Sources are registered to RPs by so called register packets, which are created by the first hop routers, closest to the source. A single copy of the multicast packet is sent through the RP to the registered receivers. Group members are joined to the shared tree by their local designated router. A shared tree that is built this way is always rooted at the RP.

By the way: PIM-SM is the only solution recommended by Cisco.

The bidirectional PIM mode (PIM-bidir) had been designed for many-to-many applications such as needed for conferencing and whiteboarding purposes.

The Source Specific Multicast (SSM) is a variant of PIM-SM that only builds source specific shortest path trees. This solution does not need an active RP and uses the source-specific group address range 232/8.



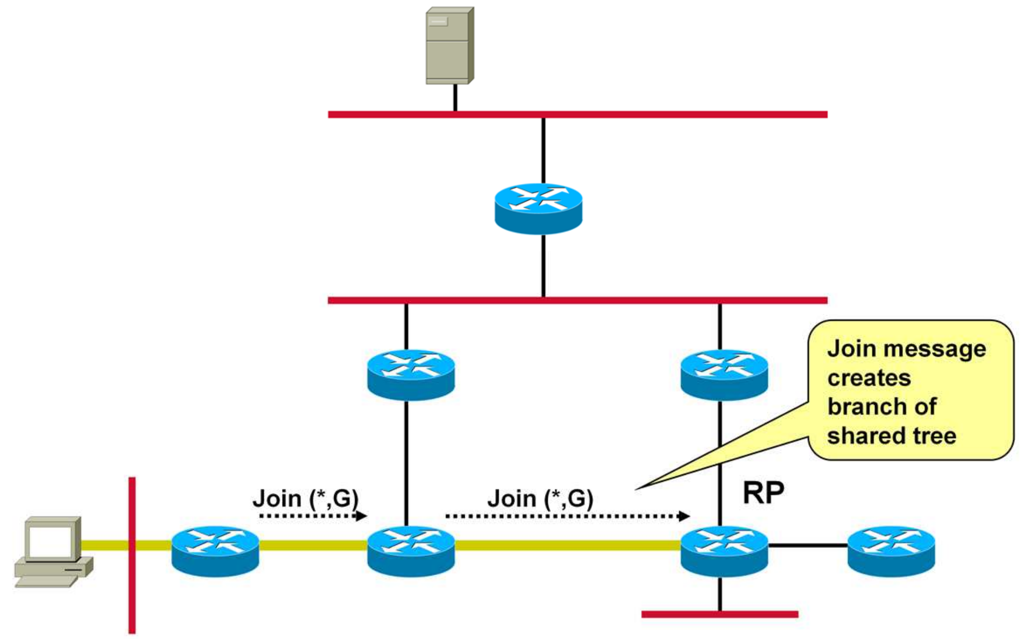
## PIM-SM / User Becomes Active

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 160

User joins group: The picture above shows how a receiver tells its designated router (DR) that he becomes active and wants to listen to group G. This is done using IGMP on the local LAN segment.

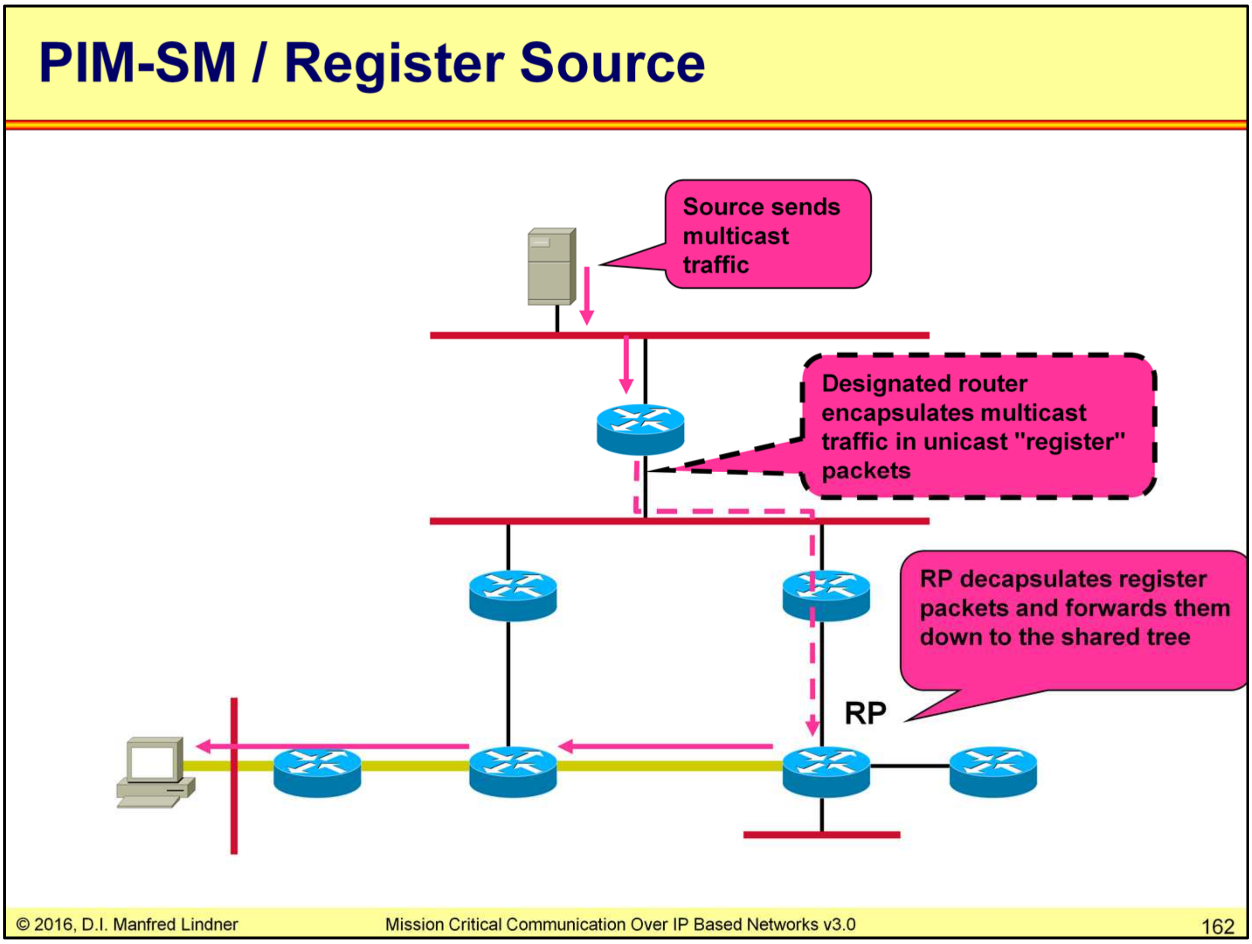
The DR sends a Join (*, G) to the RP. Obviously the DR must know the IP address of the RP. Obviously the DR does not need to know the IP address of the source. Obviously the receiver should at least know what he wants to listen to.

## PIM-SM / Create Shared Tree



© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 161

Shared tree to RP: This (*, G) join message is forwarded hop-by-hop toward the RP and hereby a branch of the shared tree is established. Now multicast traffic for group G may flow down the shared tree to the receiver.



## PIM-SM / Register Source

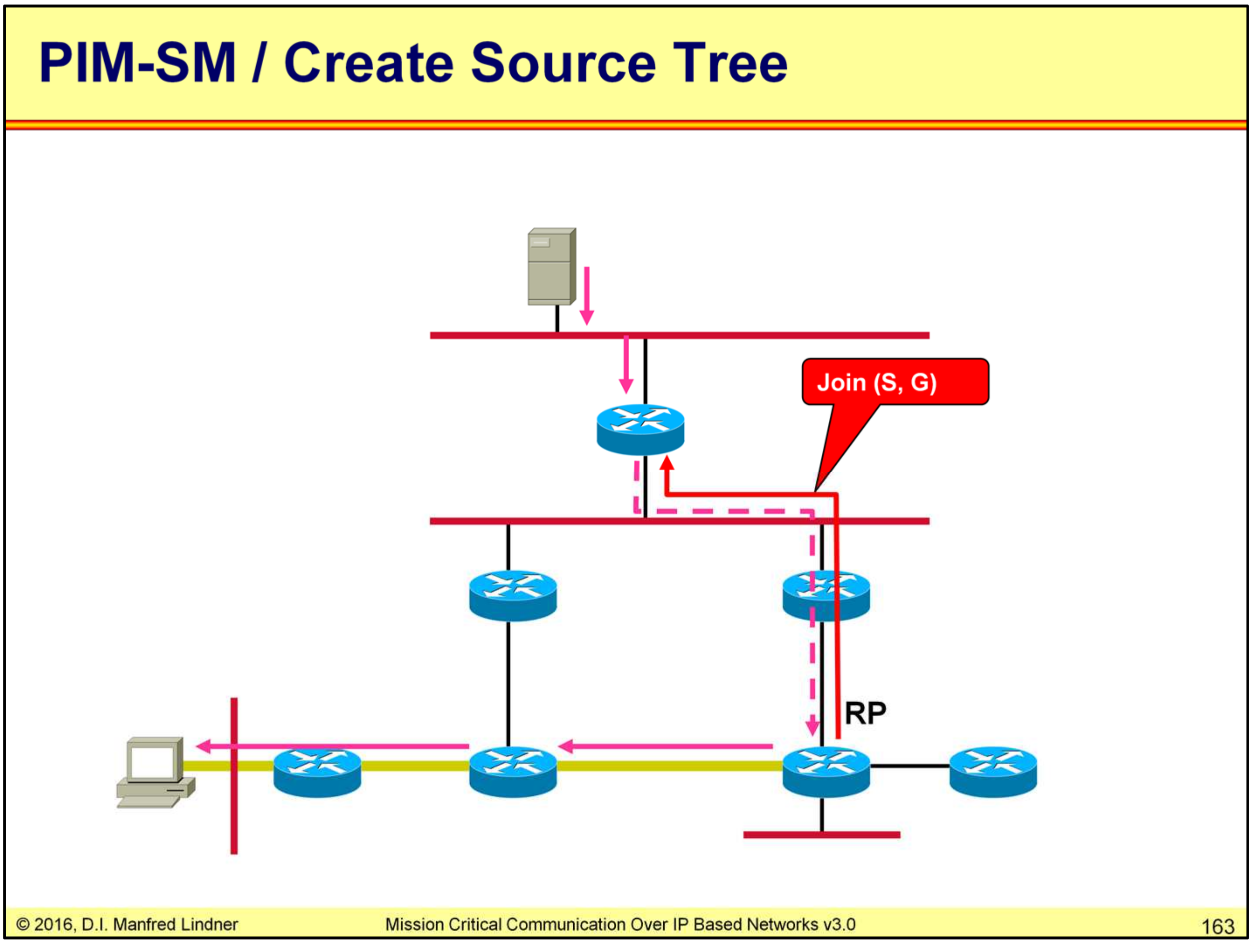
Source sends multicast traffic

Designated router encapsulates multicast traffic in unicast "register" packets

RP decapsulates register packets and forwards them down to the shared tree

DR registers at RP: The source (for G) becomes active and sends multicast packets, which are encapsulated by the first router (DR) into unicast packets. These "register" packets are sent to the RP. Obviously this DR must also know the IP address of the RP.

The RP decapsulates this packets and forwards the multicast packets (which had been carried inside the register packets) downstream to the group G.

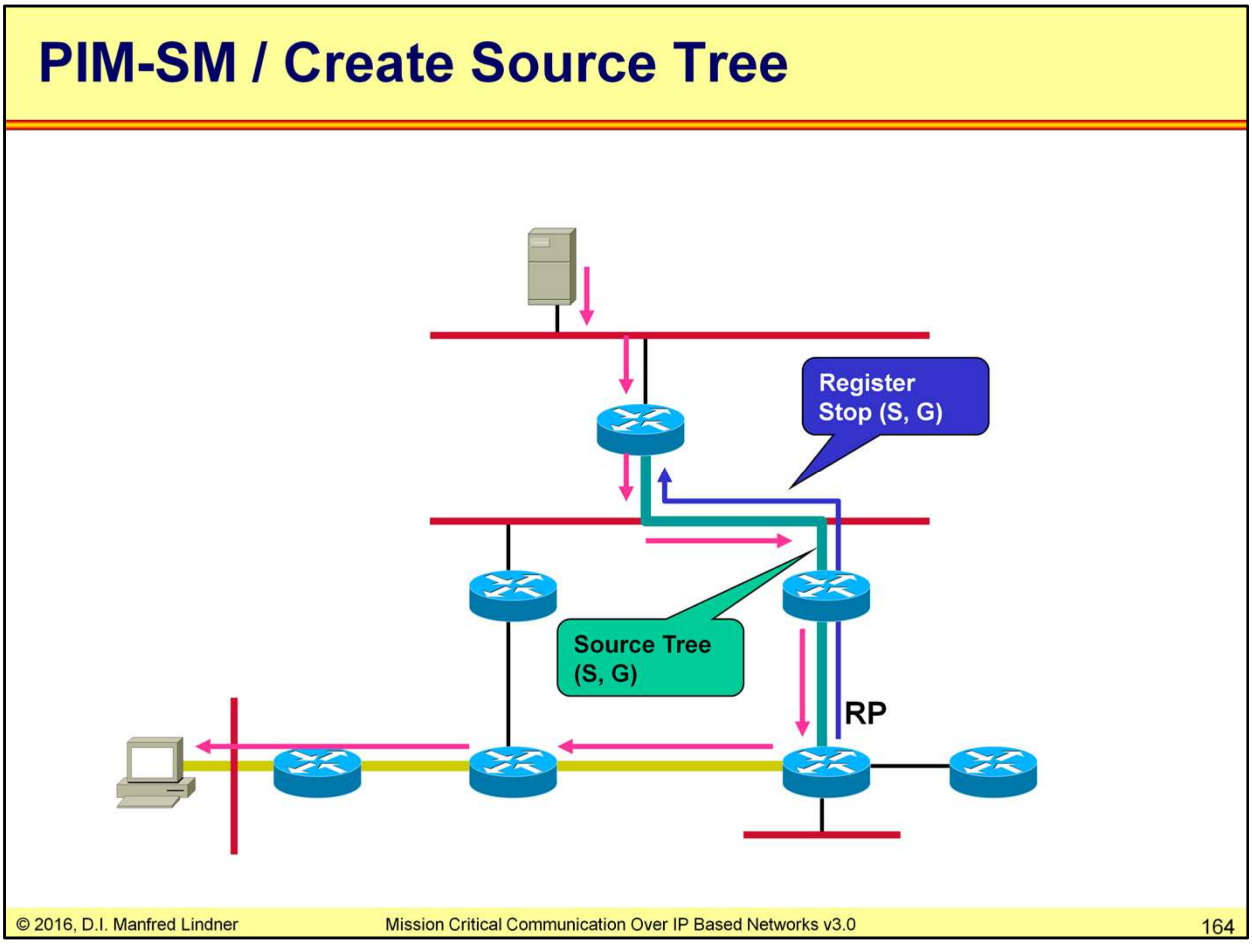


# PIM-SM / Create Source Tree

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 163

RP joins SPT: Now the RP creates a shortest-path tree (SPT) by sending an (S, G) join toward the source. Now (S, G) states are created in all routers along this new SPT path.

Note: Also the RP must maintain a (S, G) state.



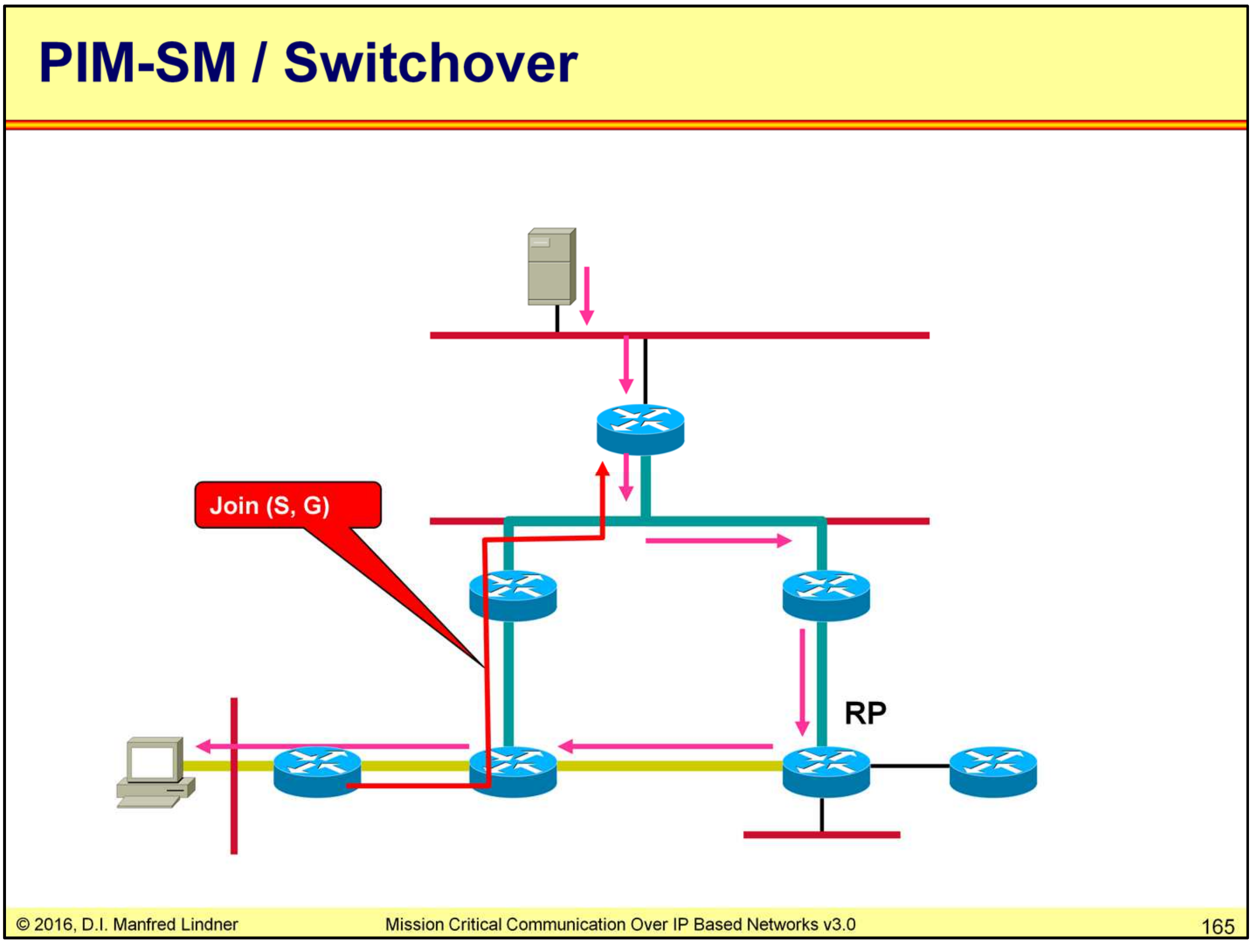
## PIM-SM / Create Source Tree

Register Stop (S, G)

Source Tree (S, G)

RP

RP stops registering: As soon as native multicast packets arrive at the RP (over the newly established SPT) the RP sends a "Register Stop (S, G)" message to the first-hop router, in order to stop the sending of unnecessary register packets.



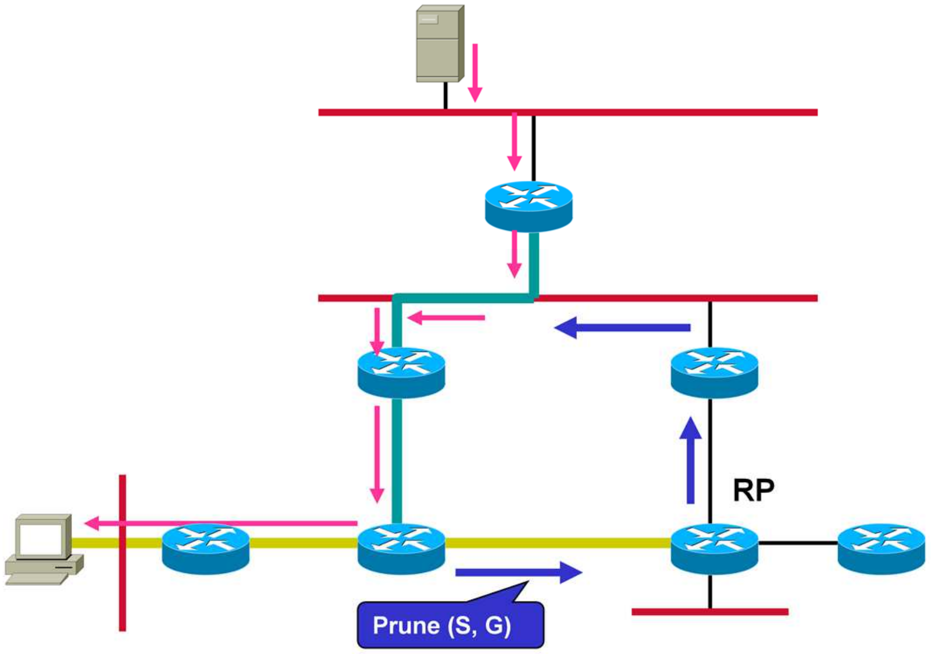
## PIM-SM / Switchover

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 165

Shortcut: PIM-SM is able to switchover to the shortest connection to the source. That is, last-hop routers (i.e. routers with directly connected members) can switch to the Shortest-Path Tree and bypass the RP if the traffic rate is above a configured threshold called the “SPT-Threshold”.

Note: The default value of the SPT-Threshold in Cisco routers is zero. Therefore the default behavior for Cisco PIM-SM leaf routers is to immediately join the SPT to the source as soon as the first packet arrives via the (*,G) shared tree.

## PIM-SM / Pruning

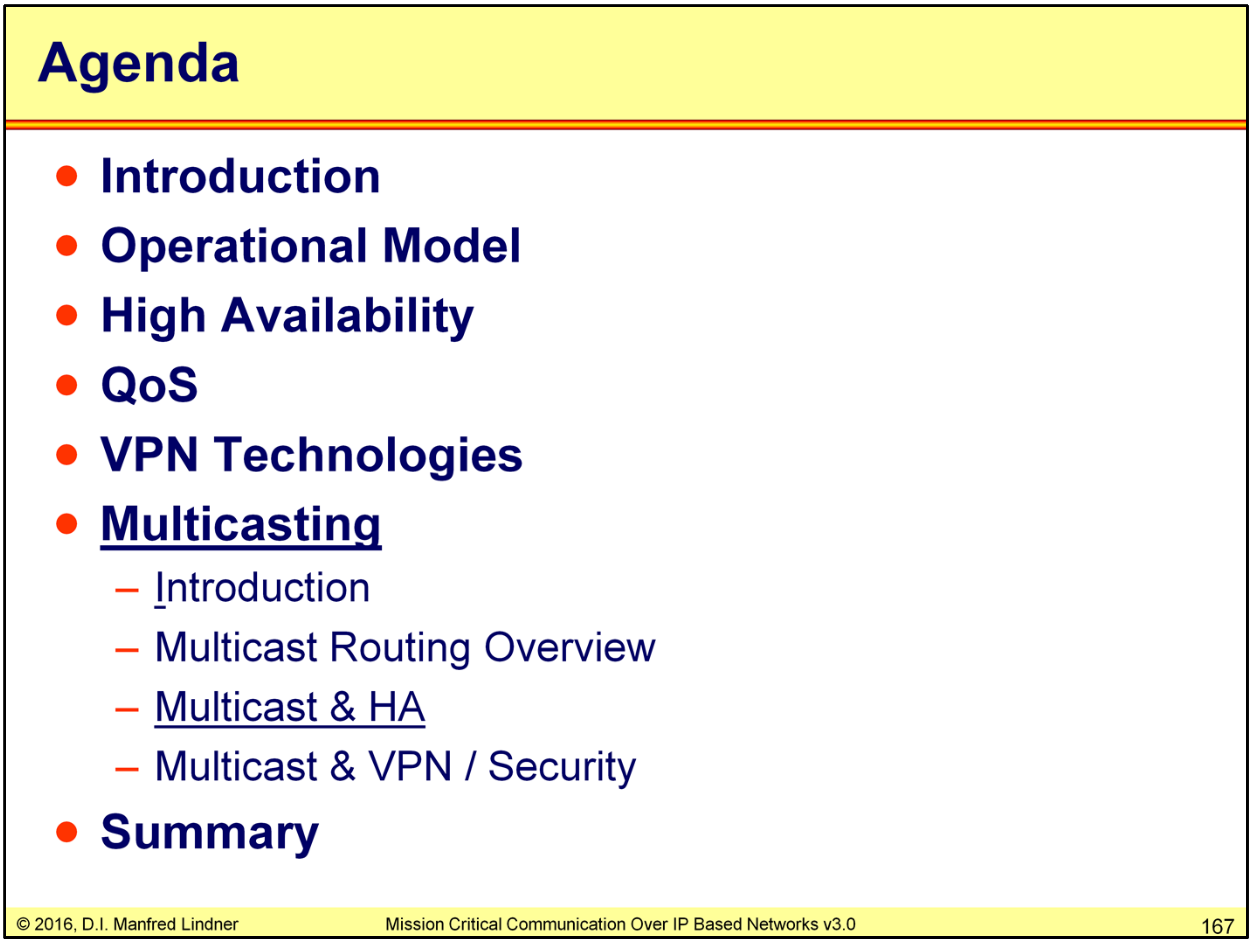


Disconnect from RP: Now, special (S, G) RP-bit prune messages are sent up the shared tree to prune only the (S, G) traffic from this shared tree. This prune is important to avoid duplicate packets.

RP may disconnect from source DR: When the (S, G) prune (with RP-bit set) arrives at the RP the RP sends (S, G) prune messages back toward the source to stop the unnecessary (S, G) traffic. Note: Of course the RP may only do this if the RP has received an (S, G) RP-bit prune via all branches, i. e. no receiver on the shared tree wants to receive the (S, G) traffic from the RP anymore.

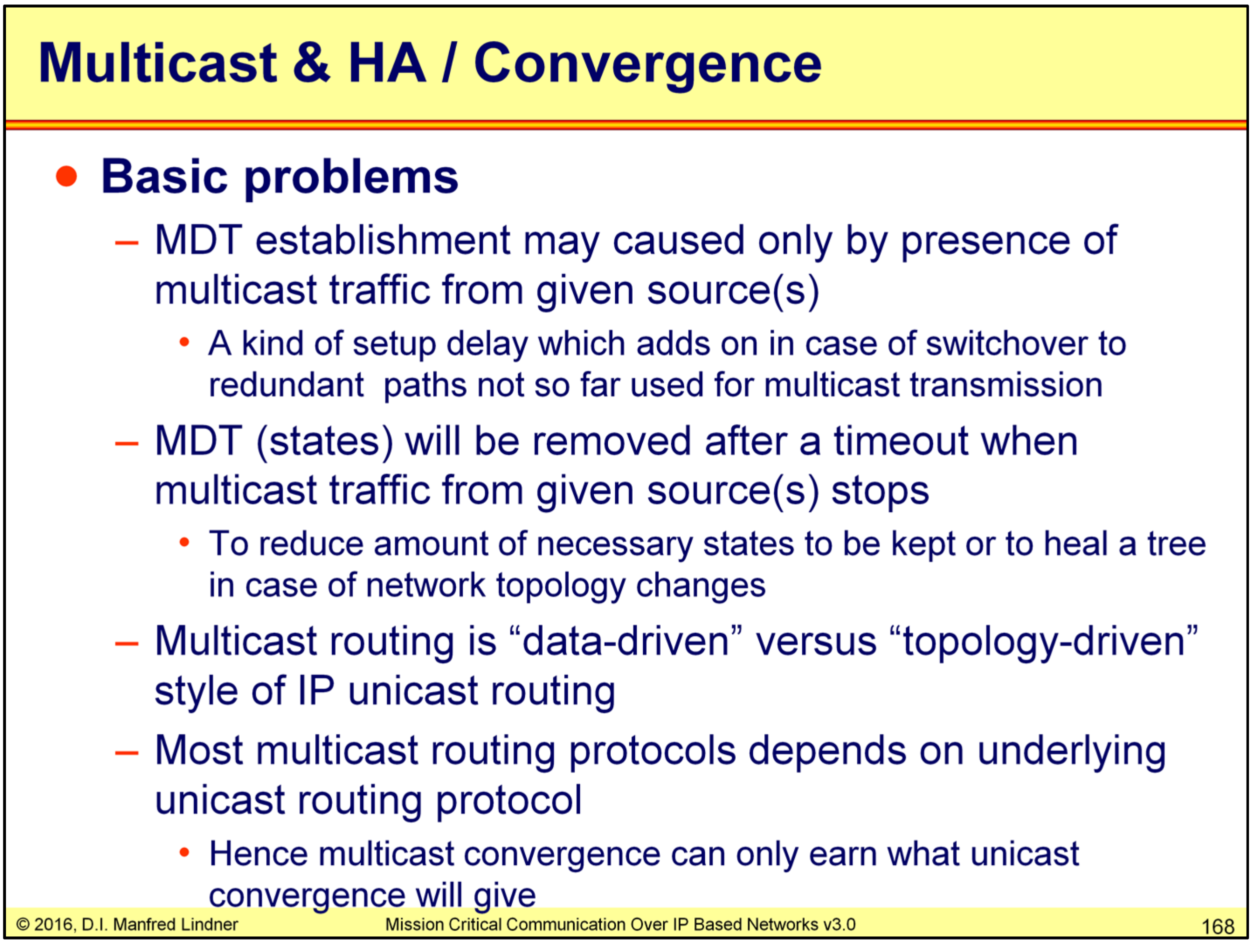
Now we learned:

PIM-SM can also create SPT (S, G) trees but in a much more economical way than PIM-DM (fewer forwarding states). PIM-SM is efficient, even for large scale multicast domains. PIM-SM can be efficiently used for both sparse and dense distribution of multicast receivers. There is no need to flood multicast traffic at any time. On the other hand a RP is needed at least for the initial setup of a MDT.



# Agenda

- Introduction
- Operational Model
- High Availability
- QoS
- VPN Technologies
- Multicasting
  - Introduction
  - Multicast Routing Overview
  - Multicast & HA
  - Multicast & VPN / Security
- Summary



## Multicast & HA / Convergence

- **Basic problems**
  - MDT establishment may caused only by presence of multicast traffic from given source(s)
    - A kind of setup delay which adds on in case of switchover to redundant  paths not so far used for multicast transmission
  - MDT (states) will be removed after a timeout when multicast traffic from given source(s) stops
    - To reduce amount of necessary states to be kept or to heal a tree in case of network topology changes
  - Multicast routing is “data-driven” versus “topology-driven” style of IP unicast routing
  - Most multicast routing protocols depends on underlying unicast routing protocol
    - Hence multicast convergence can only earn what unicast convergence will give

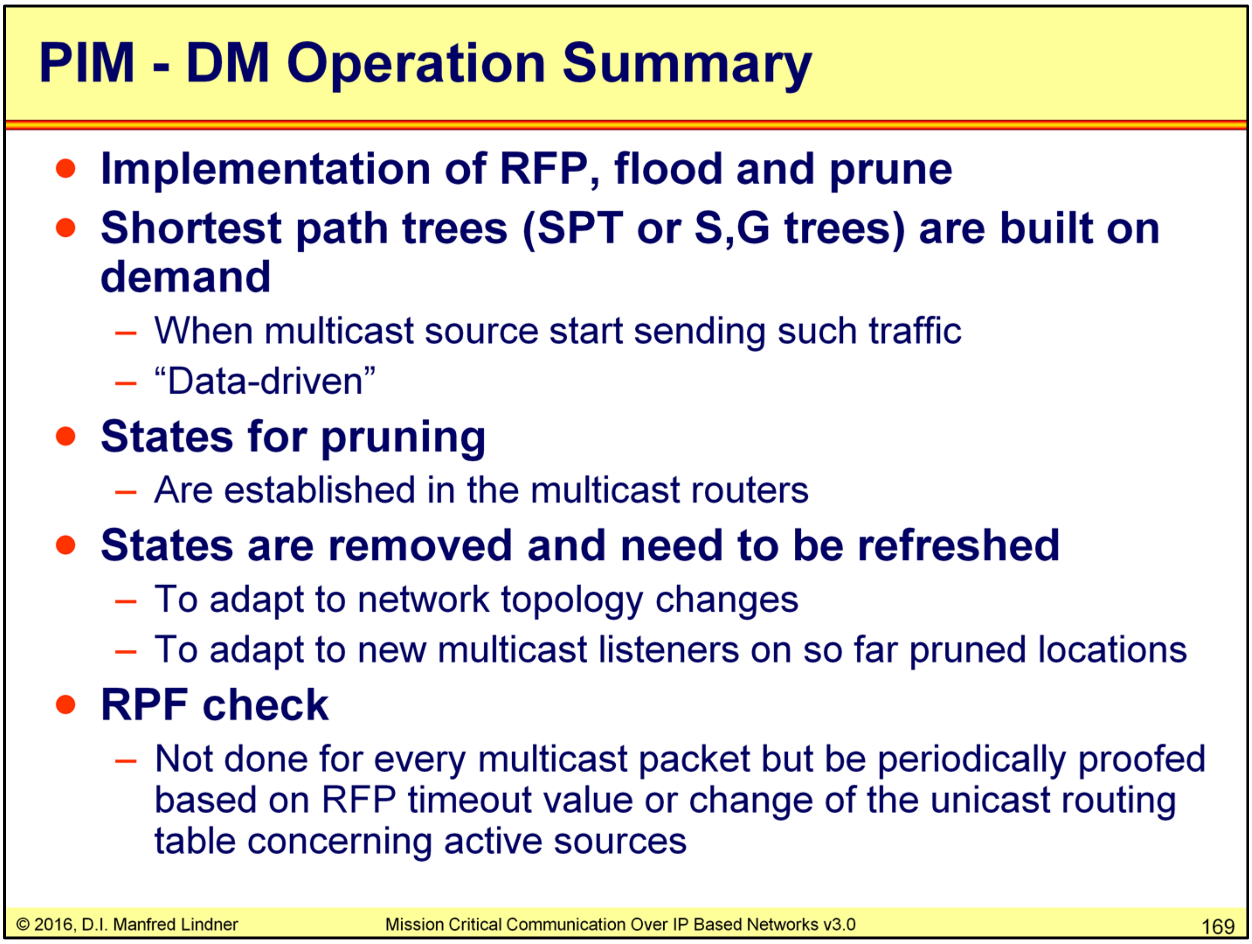
© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 168

High availability based on redundancy of components / physical paths together with automatic switchover (rerouting techniques) needs more careful investigations compared to HA for unicast communication. Depending on the particular multicast routing technique used there are three basic problems.

First problem: Multicast distribution trees (MDT) are established when multicast traffic from a given source appears in the network for the first time. So we have a kind of setup delay which may not be tolerable for HA.

Second problem: If a multicast traffic stops then after a timeout the states representing the multicast distribution tree in the network are removed. So verifying that a multicast distribution tree is available is not just as easy as with unicast IP routing. Unicast IP routing tables are established concerning reachability in the network (“topology-driven”) but need no unicast traffic in order to be erected. Multicast distribution trees are typically “data-driven” to express that multicast data traffic is necessary for establishing of “multicast IP routing table”. The reason for that: Source IP address of multicast traffic may change over time (e.g. if several sources at different places contributes multicast data in a conference style for the same group address). In principle a multicast distribution tree is a shortest path tree from the source to all known multicast receivers. If you have many sources then there are many such shortest path trees spanned over the network. In a huge network to calculate that for all possible source locations does not scale. Hence it is data-driven and not topology-driven.

Third problem: Most multicast routing depends on the availability of traditional unicast IP routing in order to verify the shortest path towards the source of multicast traffic or towards the rendezvous point. So in case of a network failure only after convergence of the unicast routing the additional tasks for erecting a new multicast distribution tree can be done. Hence multicast routing inherits time behavior of the unicast routing in a similar way as a customer inherits routing convergence behavior of the service provider network in case of a L3-VPN model.

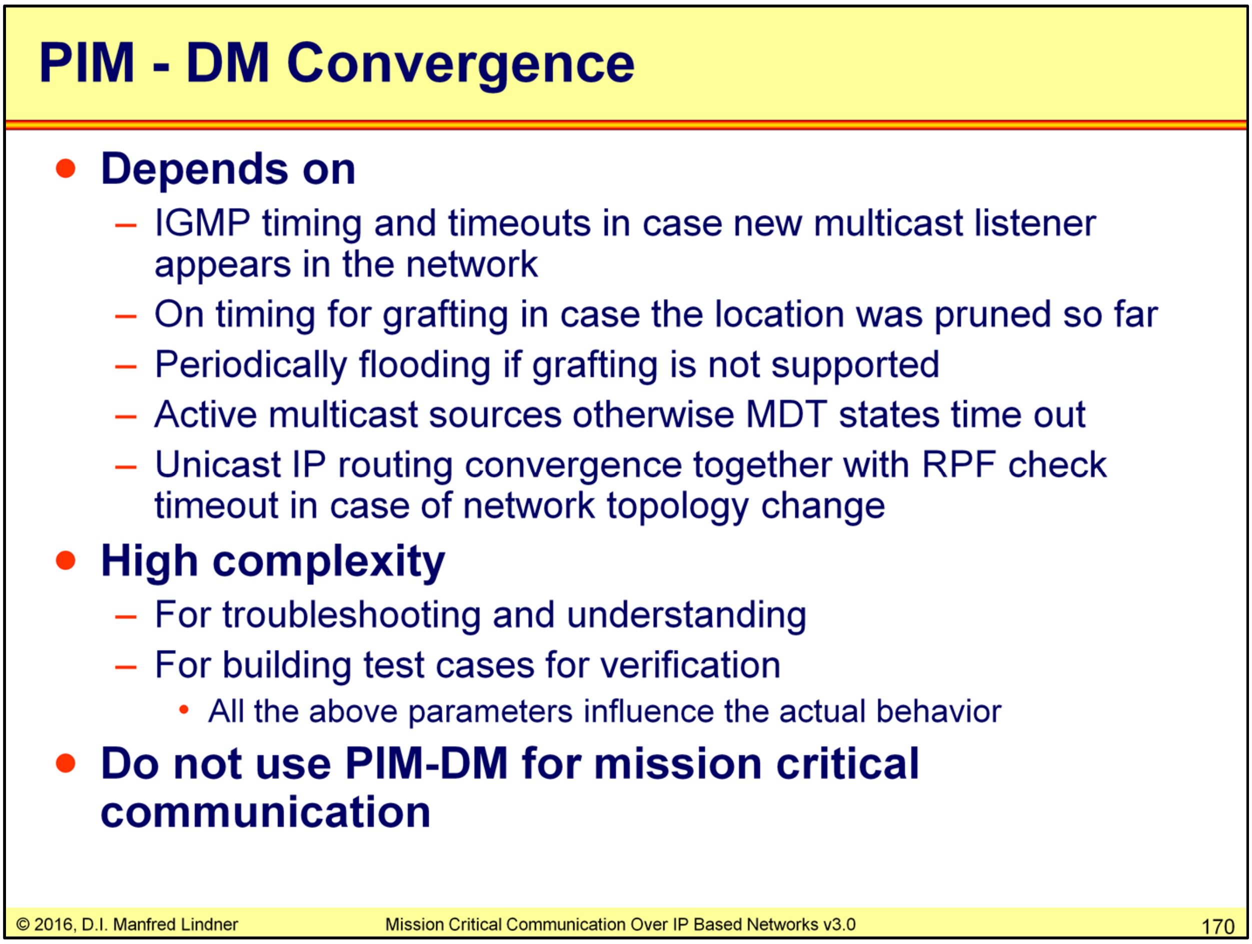


## PIM - DM Operation Summary

- **Implementation of RFP, flood and prune**
- **Shortest path trees (SPT or S,G trees) are built on demand**
  - When multicast source start sending such traffic
  - "Data-driven"
- **States for pruning**
  - Are established in the multicast routers
- **States are removed and need to be refreshed**
  - To adapt to network topology changes
  - To adapt to new multicast listeners on so far pruned locations
- **RPF check**
  - Not done for every multicast packet but be periodically proofed based on RFP timeout value or change of the unicast routing table concerning active sources

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 169

PIM DM (Protocol Independent Multicast routing Dense Mode): Basically it is an implementation of RPF (reverse path forwarding) check which looks if a multicast packet arrives on the interface which is on the shortest path to the multicast source. If so, the multicast packet will be duplicated to all other interfaces. If not - meaning the multicast packets arrives on an interface which is not on the on the shortest path to the multicast source - the multicast packet will be discarded. Using this technique a shortest path from the source to all network parts can be established. Now - after multicast traffic actually arrives from the source - a downstream router can decide based on IGMP, if it needs that multicast traffic or not. If not it will send to the upstream router a "prune" routing message to stop this unwanted traffic and the upstream router will install a state to stop duplication traffic on the interface towards downstream router. Downstream and upstream are terms used according to the shortest path tree to the multicast source also called SPT (Source-Path-Tree) or S,G tree.



## PIM - DM Convergence

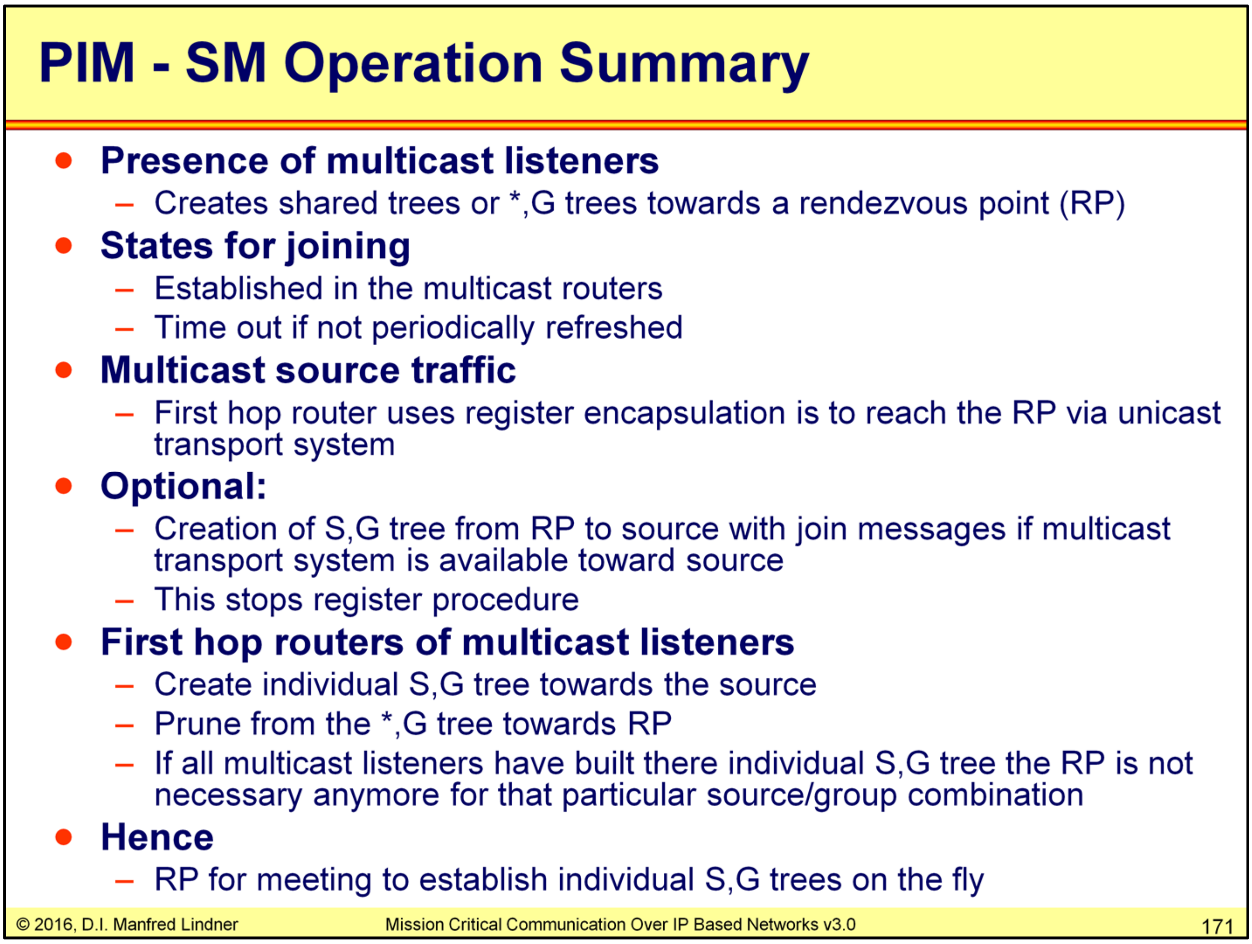
- **Depends on**
  - IGMP timing and timeouts in case new multicast listener appears in the network
  - On timing for grafting in case the location was pruned so far
  - Periodically flooding if grafting is not supported
  - Active multicast sources otherwise MDT states time out
  - Unicast IP routing convergence together with RPF check timeout in case of network topology change
- **High complexity**
  - For troubleshooting and understanding
  - For building test cases for verification
    - All the above parameters influence the actual behavior
- **Do not use PIM-DM for mission critical communication**

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 170

Now let us analyze the PIM DM in more detail concerning routing convergence. It need IGMP (Internet Group Membership Protocol) to get aware which multicast groups are necessary at a given location. IGMP allows a multicast listener to tell the local multicast router to which multicast addresses he/she is interested in. Based on IGMP a multicast router can make its decision if a given multicast address needs support (forwarding of multicast traffic) or not.

What happens if group membership changes on a pruned part of the SPT? States are removed after a certain timeout (3 minutes per default) and multicast traffic will flood the network again using RPF. So downstream routers have a new chance to decide if to prune or not. If 3 minutes are too long there is an additional method of “grafting” to speed up. Whenever a downstream router has pruned for a multicast group address but then recognizes via IGMP that a local multicast listener appears, it can send a graft message to the upstream router to immediately disable the state hence get the corresponding multicast traffic without waiting for 3 minutes refresh interval.

So far we have considered only setup and refresh of a MDT for a given multicast group with given source(s). What happens if there is failure in the network and a change of topology? When multicast traffic is still produced by the source, RPF will establish a new SPT after IP unicast routing has converged. RPF checks depend on shortest path calculation done by the unicast routing protocol to the multicast source hence the new SPT will reach a consistency after IP unicast has reached consistency and the downstream routers have pruned the trees accordingly. During time of change in the network we will experience flooding of multicast packets which may introduce additional delay for convergence.



# PIM - SM Operation Summary

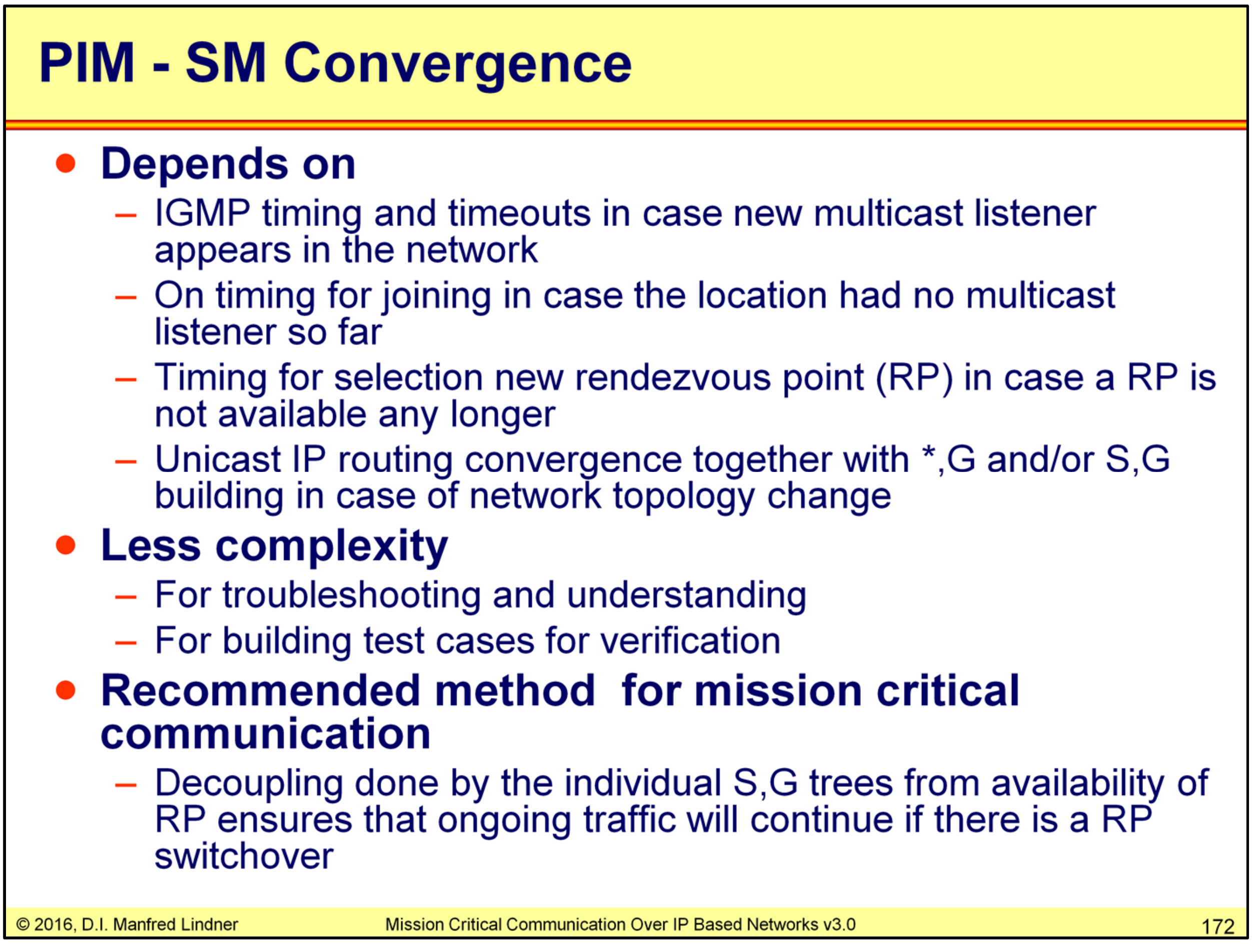
- **Presence of multicast listeners**
  - Creates shared trees or *,G trees towards a rendezvous point (RP)
- **States for joining**
  - Established in the multicast routers
  - Time out if not periodically refreshed
- **Multicast source traffic**
  - First hop router uses register encapsulation is to reach the RP via unicast transport system
- **Optional:**
  - Creation of S,G tree from RP to source with join messages if multicast transport system is available toward source
  - This stops register procedure
- **First hop routers of multicast listeners**
  - Create individual S,G tree towards the source
  - Prune from the *,G tree towards RP
  - If all multicast listeners have built there individual S,G tree the RP is not necessary anymore for that particular source/group combination
- **Hence**
  - RP for meeting to establish individual S,G trees on the fly

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 171

PIM SM (Protocol Independent Multicast routing Spares Mode): This kind of multicast routing was developed to overcome the deficits of PIM DM namely first periodical flooding the whole network with multicast traffic in order to refresh or rebuilt the multicast distribution trees and second the necessary amount of state information in the network.

Multicast routers know by configuration or auto-discovery about IP address of a rendezvous point acting as core. Multicast listeners are supported starting from the rendezvous point with a “shared” multicast distribution tree (so called *,G tree in comparison to S,G tree). Every router - after recognizing by IGMP that certain group address is required by its local listeners - will sent a multicast “Join” message upstream to the rendezvous point. For determination of upstream unicast routing is used to find the shortest path to the rendezvous point. The next upstream router - depending on if the *,G tree was already erected towards it or not – will either move on the “Join” message towards the rendezvous point or will add the interface where the “Join” was received to the so far established *,G tree. Establishment of the *,G tree does not depend on presence of multicast traffic for a certain group. It is “topology-driven” based on the knowledge where the rendezvous point and the potential multicast listeners are located. The *,G tree will be refreshed by periodical “Join” messages from the downstream routers as long as there are some listeners present. If no listeners are known anymore the *,G tree will disappear. If now a multicast source starts sending multicast traffic the first router - knowing about the rendezvous point - will encapsulate that multicast message in a unicast message (“Register” message) destined for the rendezvous point. The rendezvous point will decapsulate the message in will forward it in multicast style along the *,G tree to all listeners registered so far. If another source from a different location starts sending of multicast traffic for the same group, the traffic will arrive encapsulated at the rendezvous point and will take the same *,G tree to reach all listeners of that group. Therefore the name “shared” tree where all source of multicast traffic take the same multicast distribution tree. In PIM DM there would be two or more different S,G trees if two or more sources are present at the same time. PIM SM scales much better in the case that most multicast listeners are in a local distance to the rendezvous point (sparse mode). Hence multicast sources far away from the rendezvous point need no multicast transport system in the network because of unicast style of “Register” messages. On the other side if a full multicast network is available then optionally the rendezvous point can establish a S,G tree using “Join” messages towards the upstream router. For determination of upstream again unicast routing is used to find the shortest path to the source. S;G tree expresses that this multicast distribution tree is used only for multicast traffic from that particular source S for a given multicast group G.

There is another optimization for a multicast router which has joined the *,G tree. On presence of incoming multicast traffic arriving at the *,G tree a multicast router of such a location can decide to built a (private) S,G tree towards the source. After establishment of this source-based tree the router can leave the *,G tree. Now multicast traffic towards this router uses a separate S,G tree from the source location to the corresponding location on its own. Cisco PIM SM multicast routers make this decision based on the amount of multicast traffic arriving on the *.G tree. The default value is 0% meaning a private S,G tree is immediately established. Now you can fully understand the name rendezvous point. It used just to meet and direct multicast traffic to your location. Later multicast traffic will use the private S,G tree and the rendezvous point is not necessary anymore.



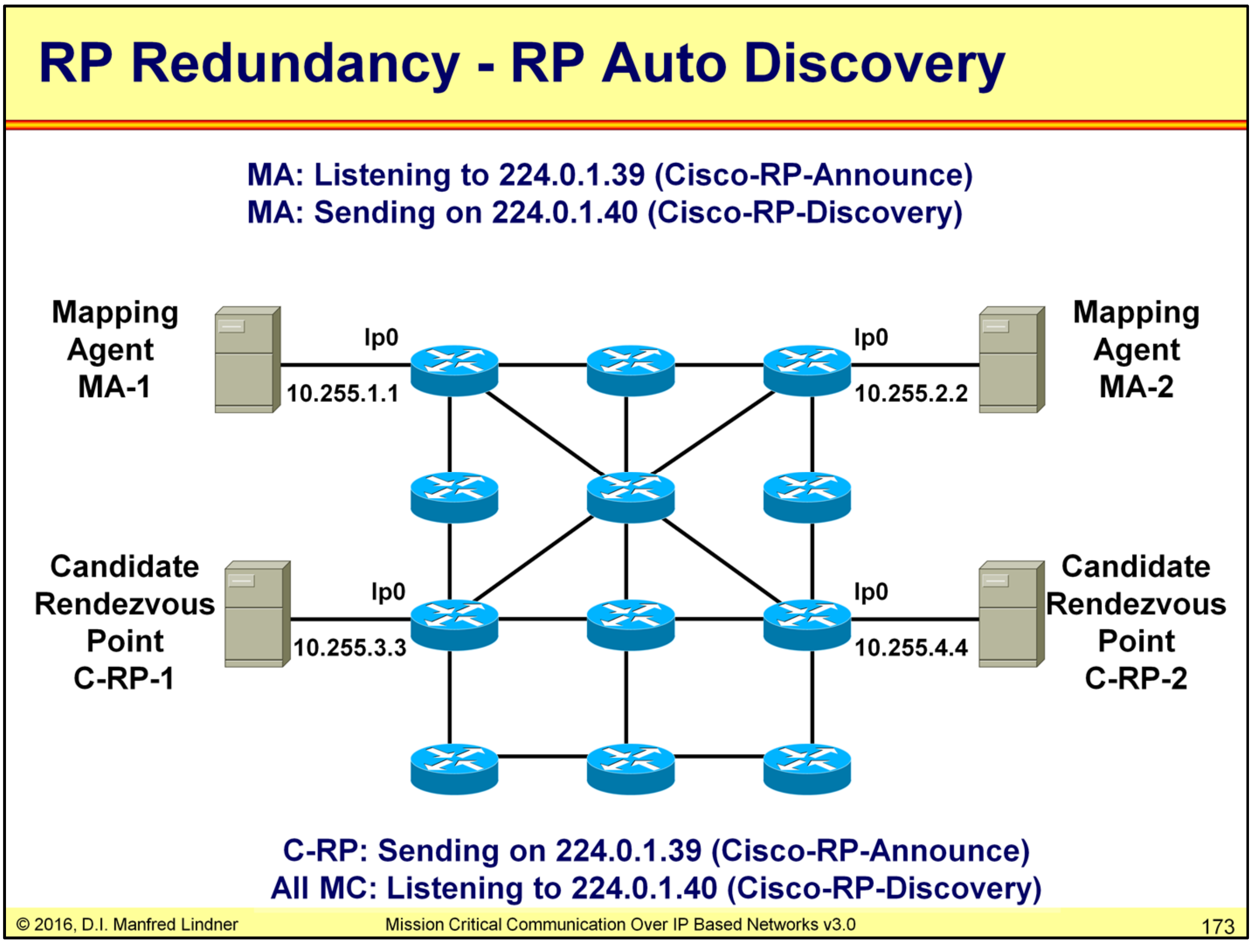
## PIM - SM Convergence

- **Depends on**
  - IGMP timing and timeouts in case new multicast listener appears in the network
  - On timing for joining in case the location had no multicast listener so far
  - Timing for selection new rendezvous point (RP) in case a RP is not available any longer
  - Unicast IP routing convergence together with *,G and/or S,G building in case of network topology change
- **Less complexity**
  - For troubleshooting and understanding
  - For building test cases for verification
- **Recommended method for mission critical communication**
  - Decoupling done by the individual S,G trees from availability of RP ensures that ongoing traffic will continue if there is a RP switchover

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 172

What happens in case of a network failure influencing the already installed trees? The multicast router will recognize that the upstream interface is not anymore on the shortest path towards the rendezvous point or source, will remove the state and will initiate a new "Join" after the IP unicast routing has converged. Actual timing when *,G and S,G trees are available again depends on if a failure can be directly detected by the router (e.g. interface down) or if the router can recognize a change by looking to the unicast routing table, where changes depend on unicast routing protocol or BFD timeouts. But again multicast routing convergence depends on underlying unicast routing convergence plus necessary time to build the new distribution trees.

Another critical point is the presence of rendezvous point. If you use the technique of private S,G trees then it has no influence on ongoing multicast streams. Only new multicast streams which need to meet at rendezvous point first cannot be transported if the rendezvous point is absent. If you do not use the technique of private S,G trees then of course the outage of a single rendezvous point (RP) is a disaster.



## RP Redundancy - RP Auto Discovery

MA: Listening to 224.0.1.39 (Cisco-RP-Announce)
MA: Sending on 224.0.1.40 (Cisco-RP-Discovery)

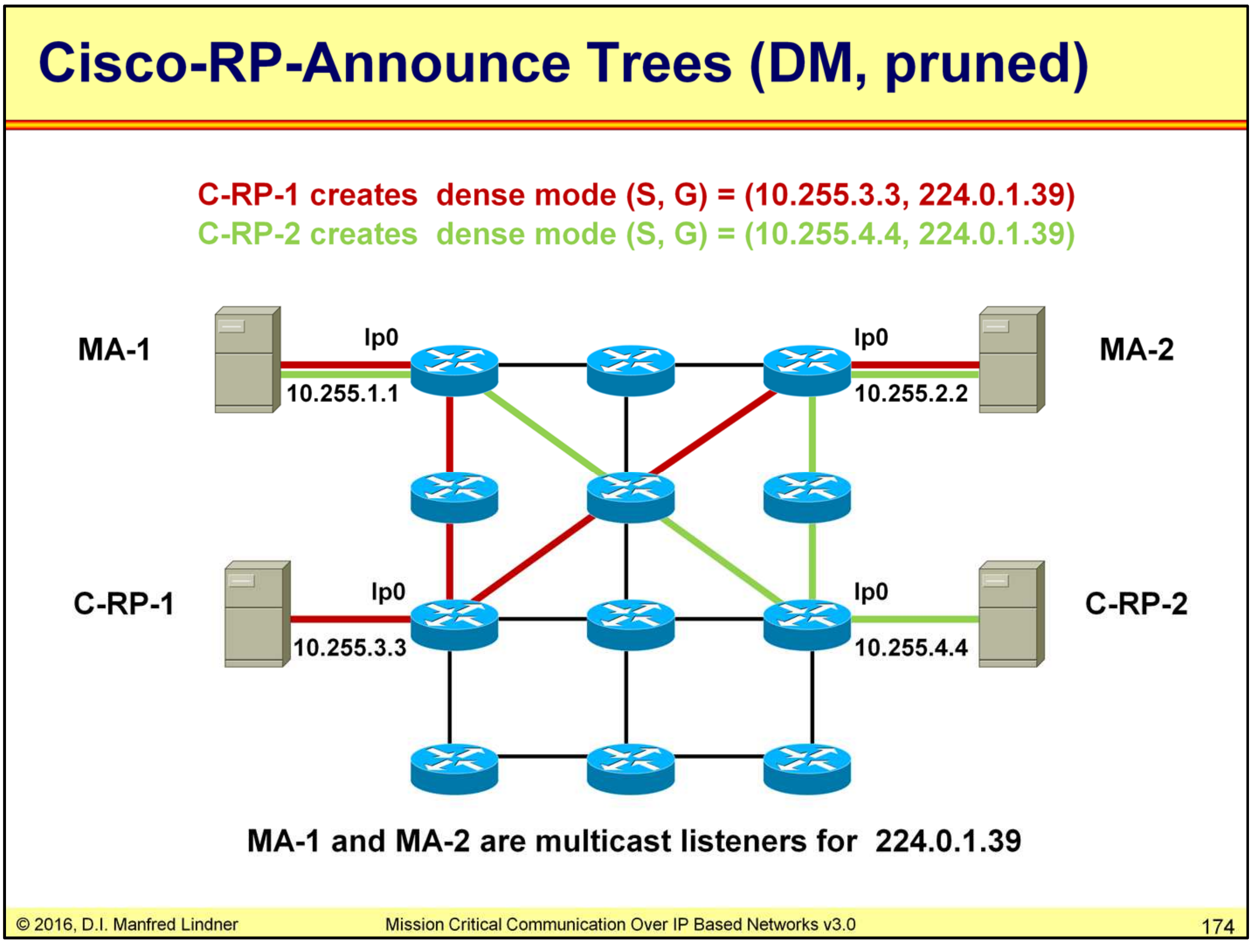
C-RP: Sending on 224.0.1.39 (Cisco-RP-Announce)
All MC: Listening to 224.0.1.40 (Cisco-RP-Discovery)

© 2016, D.I. Manfred Lindner    Mission Critical Communication Over IP Based Networks v3.0    173

Cisco solves the problem of RP being a of single point of failure with a “RP Auto-Discovery” procedure. The timing of the procedure has to be tuned to reach seconds convergence time.

RP-Auto-Discovery works as follow. Some routers are configured as mapping agents (MAs), some routers are configured as candidate rendezvous points (C-RPs). MAs use multicast dense mode techniques to discover potential C-RPs. One of the C-RPs is declared as the actual RP by the MAs, the others are standby RPs. Multicast routers are informed by the MAs about the actual RP. If there is change of presence of a RP or MA protocol timeouts will detect that change. Of course that will influence the overall availability of the multicast routing system.

Technique can also be used to provide a kind of load balancing concerning RP. So RP-1 can cover a certain range of IP multicast groups and RP-2 can cover another range of IP multicast groups. Both can act as backup for the other.



# Cisco-RP-Announce Trees (DM, pruned)

C-RP-1 creates dense mode (S, G) = (10.255.3.3, 224.0.1.39)
C-RP-2 creates dense mode (S, G) = (10.255.4.4, 224.0.1.39)

MA-1 and MA-2 are multicast listeners for 224.0.1.39

C-RP is a candidate for a multicast rendezvous point

MA is mapping agent allowing C-RPs to register

Auto-RP Procedure 1:

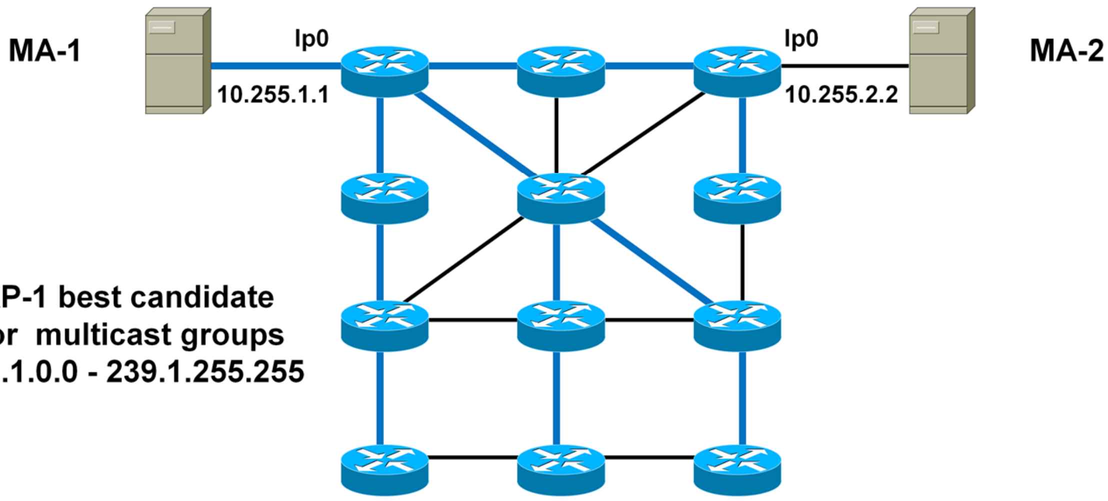
C-RPs advertize themselves and their corresponding MC groups for which they are responsible to the Cisco-RP-Announce 224.0.1.39 multicast group.

MAs listens to the Cisco-RP-Announce 224.0.1.39 multicast group, select the best candidate for RP (highest IP address).

In order to allow the Auto-RP procedure to work (to allow MA routers to join the 224.0.1.39 group and all routers to join the 224.0.1.40 group in all circumstances) a new mode was created by Cisco: the so called “ip pim spares-dense mode”. As long as no RP is known every router is using dense-mode. If RP is known routers switchover to spares-mode operation.

# Cisco-RP-Discovery Tree 1 (DM)

MA-1 creates dense mode (S, G) = (10.255.1.1, 224.0.1.40)



RP-1 best candidate for multicast groups 239.1.0.0 - 239.1.255.255

All multicast routers are multicast listeners for 224.0.1.40

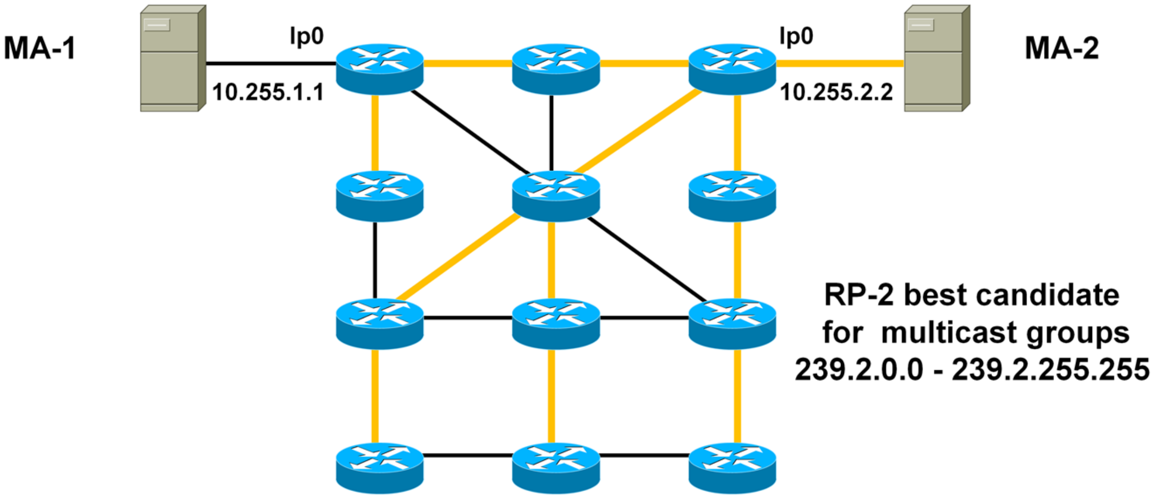
Auto-RP Procedure 2:

MAs select the best candidate for RP (highest IP address as tie-breaker) for a given multicast group range and announce this using Cisco-RP-Discovery 224.0.1.40 multicast group.

Any router configured for PIM-SM listens to 224.0.1.40 and learns the correct RP for each known group.

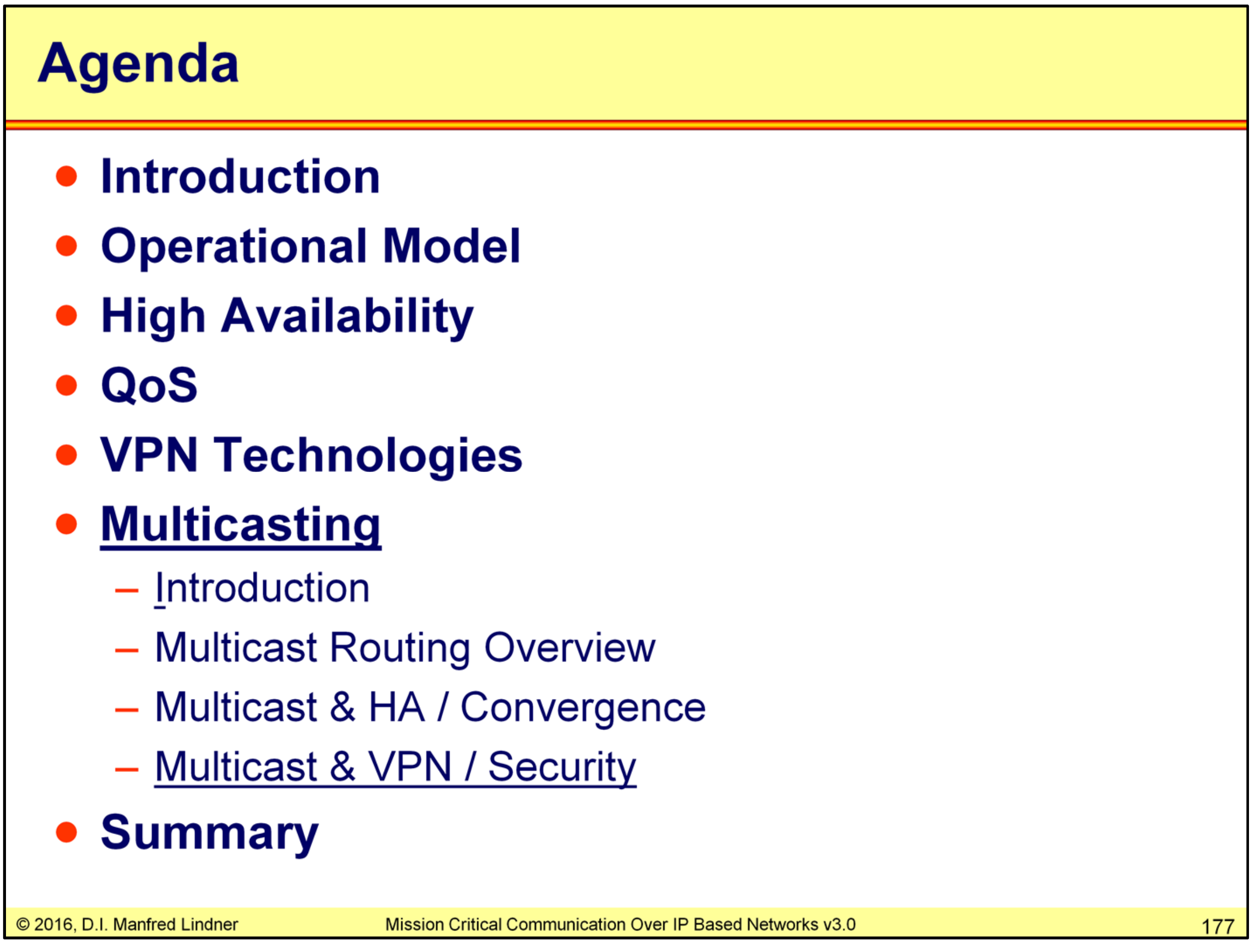
# Cisco-RP-Discovery Tree 2 (DM)

MA-2 creates  dense mode (S, G) = (10.255.2.2, 224.0.1.40)



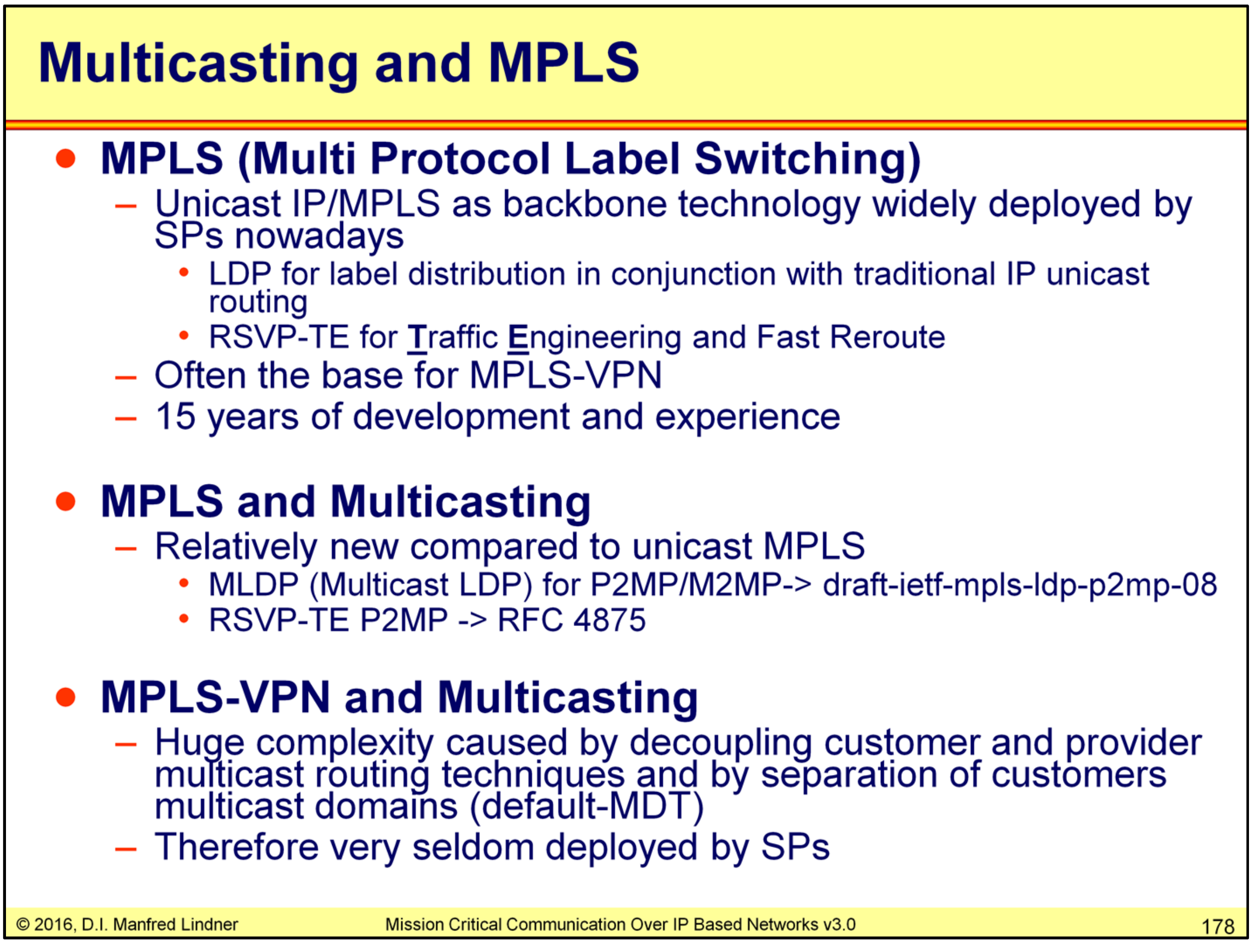
RP-2 best candidate for  multicast groups 239.2.0.0 - 239.2.255.255

All multicast routers are multicast listeners for  224.0.1.40



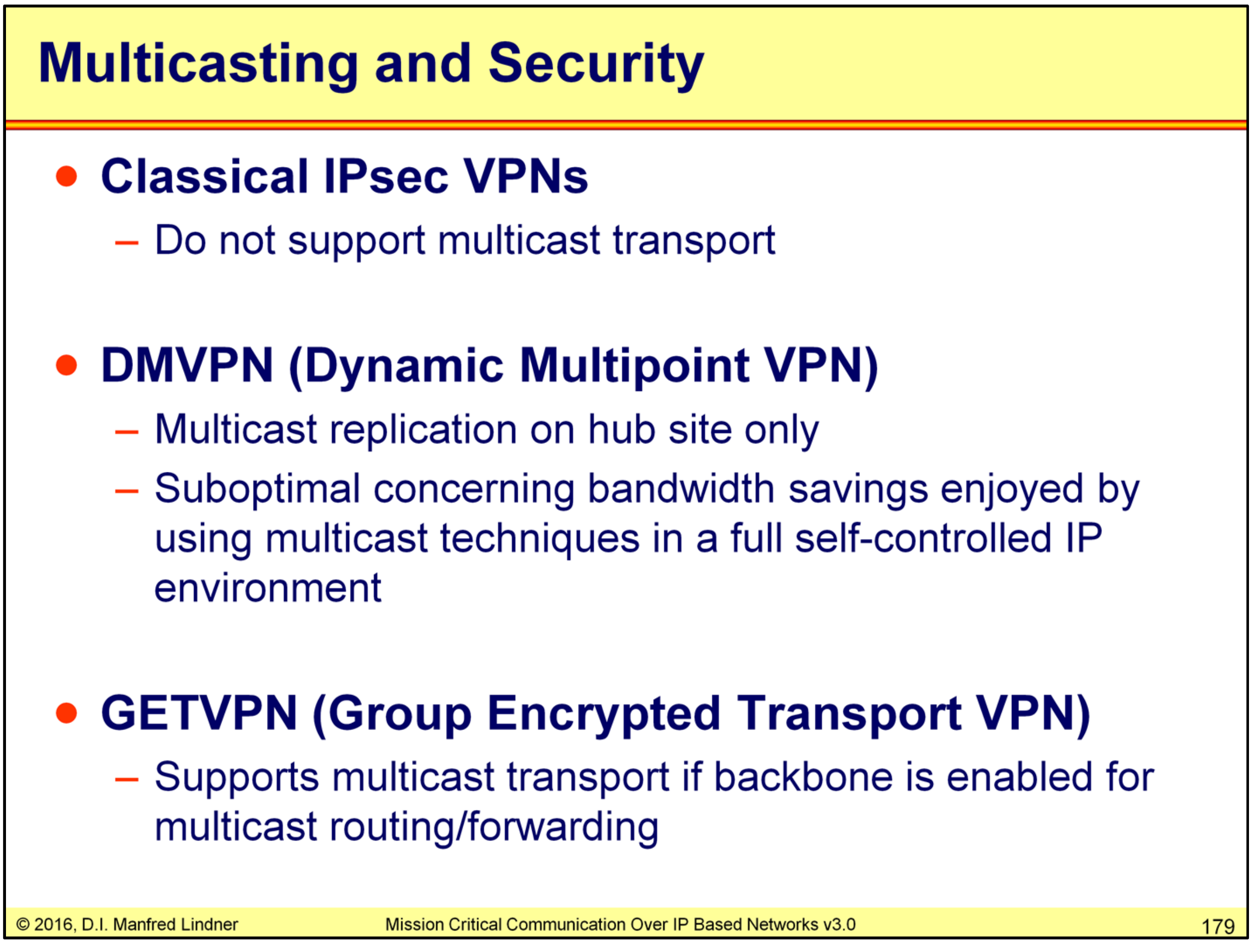
# Agenda

- Introduction
- Operational Model
- High Availability
- QoS
- VPN Technologies
- Multicasting
  - Introduction
  - Multicast Routing Overview
  - Multicast & HA / Convergence
  - Multicast & VPN / Security
- Summary



## Multicasting and MPLS

- **MPLS (Multi Protocol Label Switching)**
  - Unicast IP/MPLS as backbone technology widely deployed by SPs nowadays
    - LDP for label distribution in conjunction with traditional IP unicast routing
    - RSVP-TE for Traffic Engineering and Fast Reroute
  - Often the base for MPLS-VPN
  - 15 years of development and experience
- **MPLS and Multicasting**
  - Relatively new compared to unicast MPLS
    - MLDP (Multicast LDP) for P2MP/M2MP-> draft-ietf-mpls-ldp-p2mp-08
    - RSVP-TE P2MP -> RFC 4875
- **MPLS-VPN and Multicasting**
  - Huge complexity caused by decoupling customer and provider multicast routing techniques and by separation of customers multicast domains (default-MDT)
  - Therefore very seldom deployed by SPs

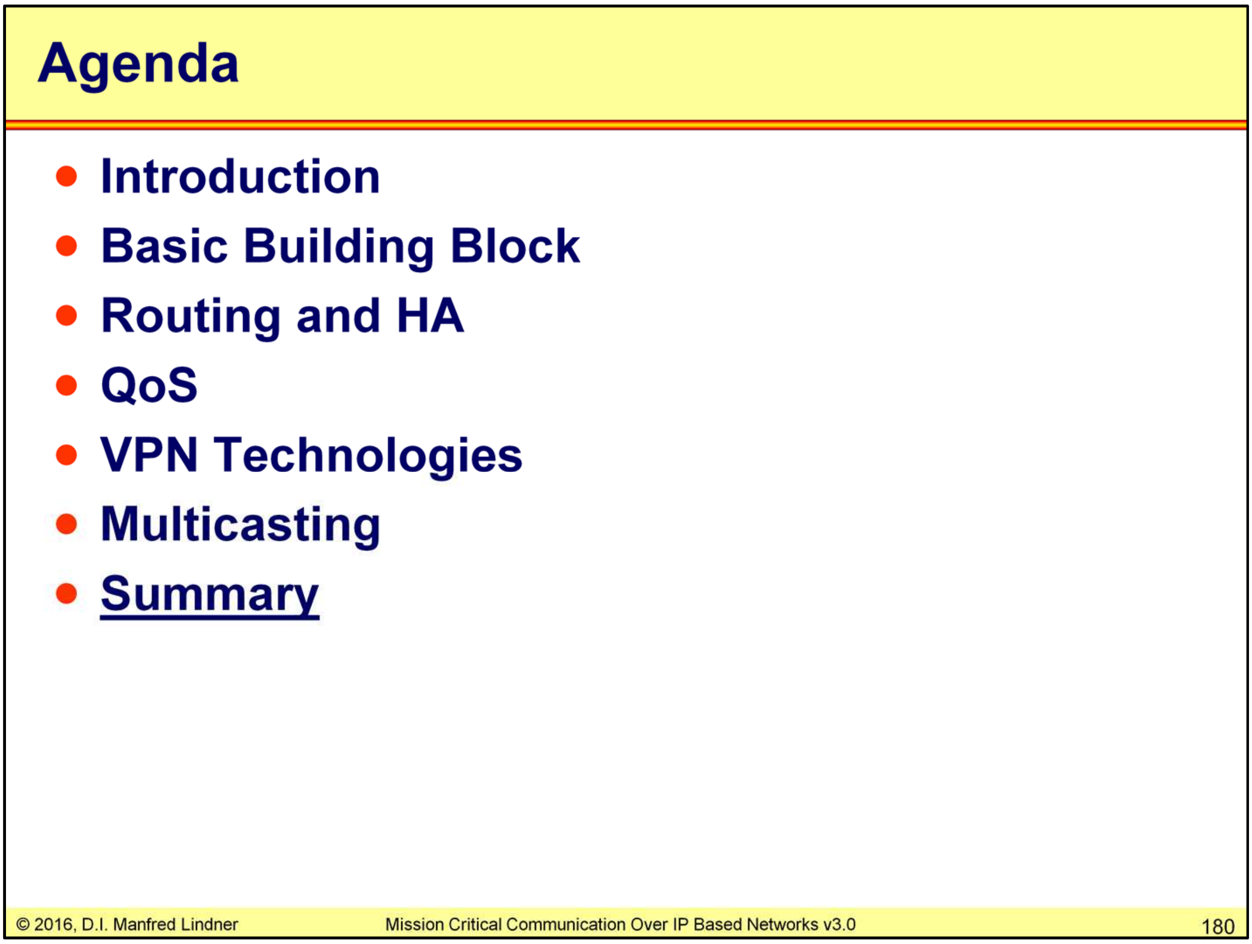


## Multicasting and Security

- **Classical IPsec VPNs**
  - Do not support multicast transport
- **DMVPN (Dynamic Multipoint VPN)**
  - Multicast replication on hub site only
  - Suboptimal concerning bandwidth savings enjoyed by using multicast techniques in a full self-controlled IP environment
- **GETVPN (Group Encrypted Transport VPN)**
  - Supports multicast transport if backbone is enabled for multicast routing/forwarding

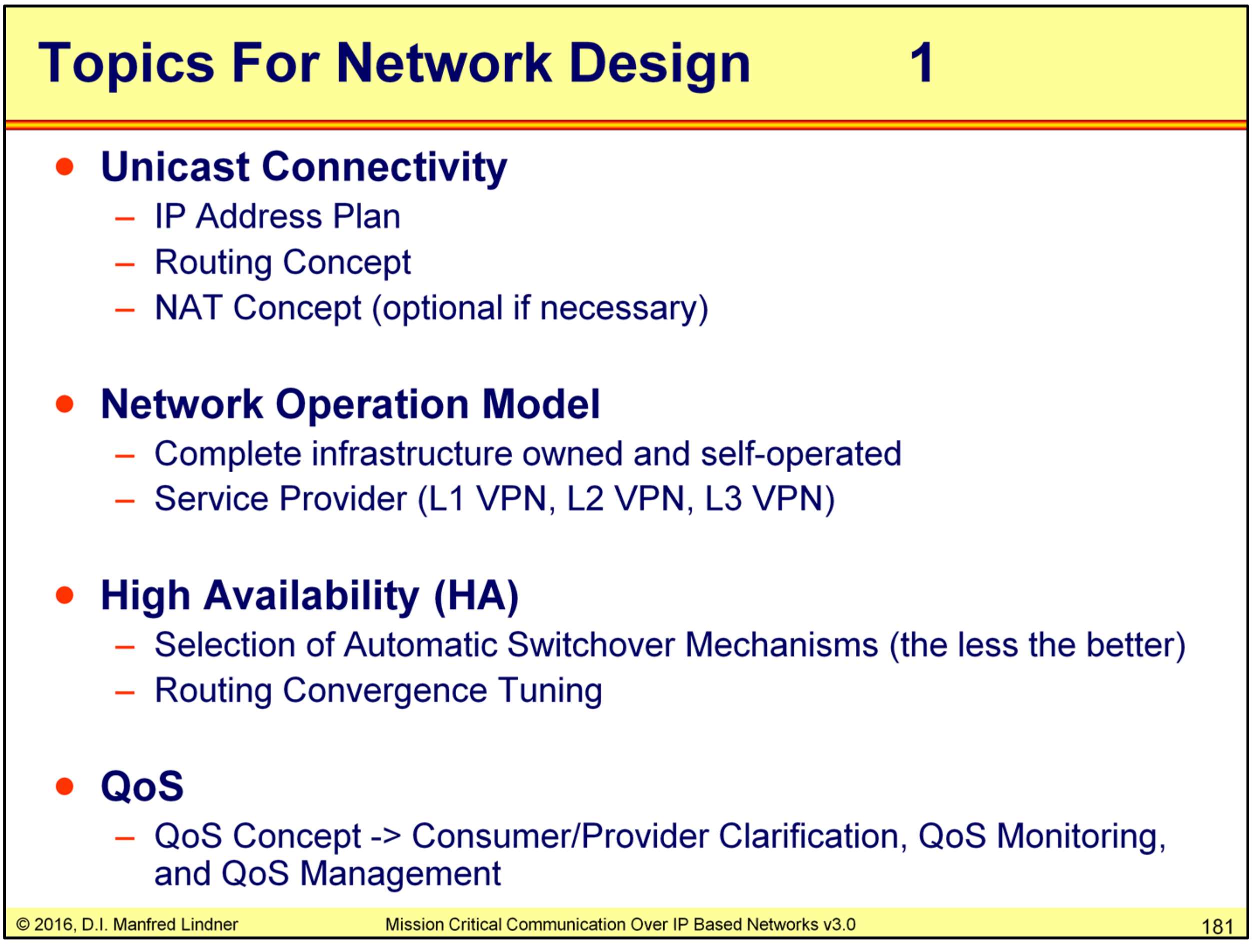
© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 179

Unfortunately overlay VPNs have either no support for multicast (IPsec) or only suboptimal support for multicast (DMVPN). Even most classical VPNs like MPLS-VPN do not offer support for multicast. So be careful when enforcing a multicast communication style in your solution. It depends on the VPN technique involved, if it is possible or not. GETVPN - because of group encryption - can be combined with IP multicast without any problem.



## Agenda

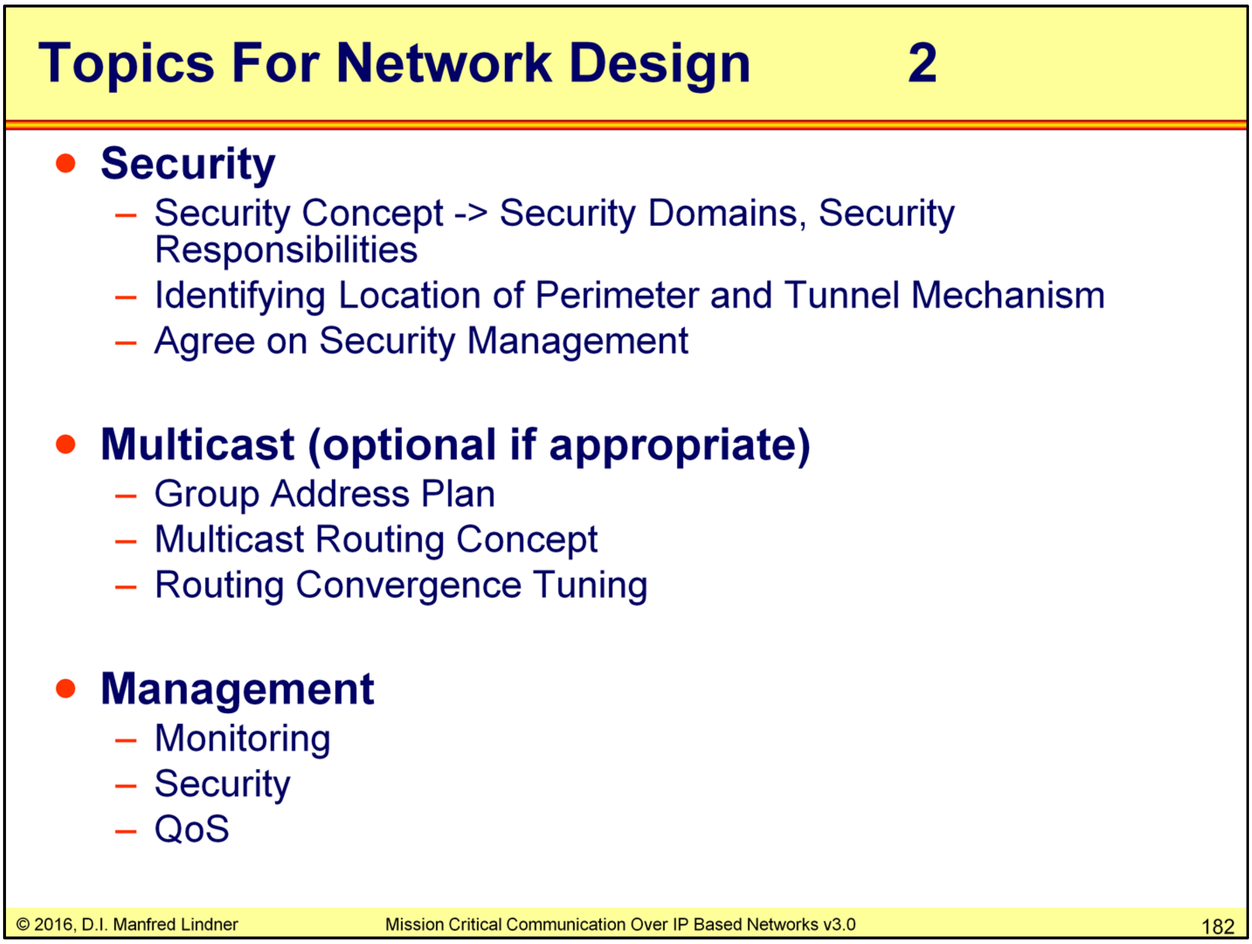
- Introduction
- Basic Building Block
- Routing and HA
- QoS
- VPN Technologies
- Multicasting
- Summary



## Topics For Network Design 1

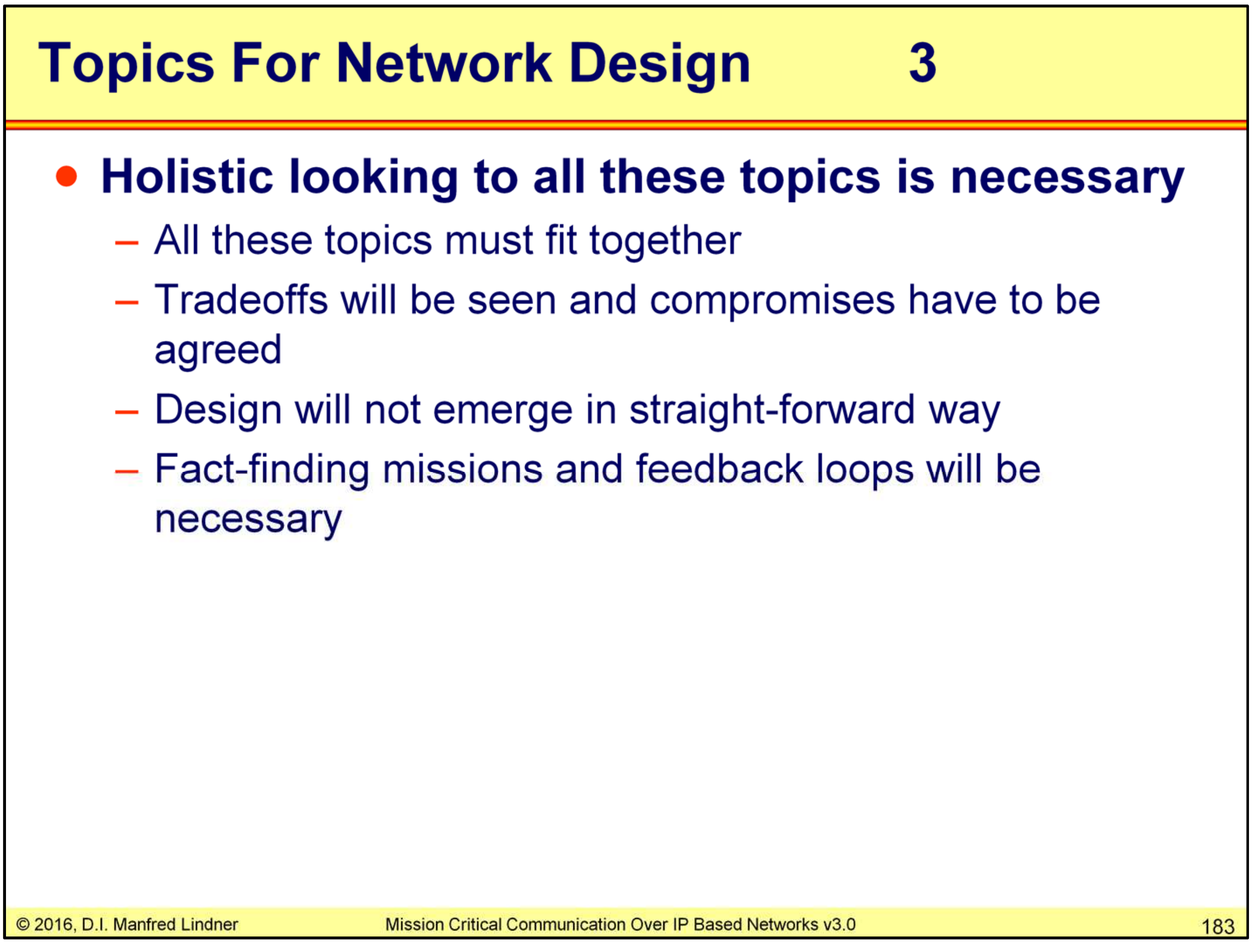
- **Unicast Connectivity**
  - IP Address Plan
  - Routing Concept
  - NAT Concept (optional if necessary)
- **Network Operation Model**
  - Complete infrastructure owned and self-operated
  - Service Provider (L1 VPN, L2 VPN, L3 VPN)
- **High Availability (HA)**
  - Selection of Automatic Switchover Mechanisms (the less the better)
  - Routing Convergence Tuning
- **QoS**
  - QoS Concept -> Consumer/Provider Clarification, QoS Monitoring, and QoS Management

© 2016, D.I. Manfred Lindner Mission Critical Communication Over IP Based Networks v3.0 181



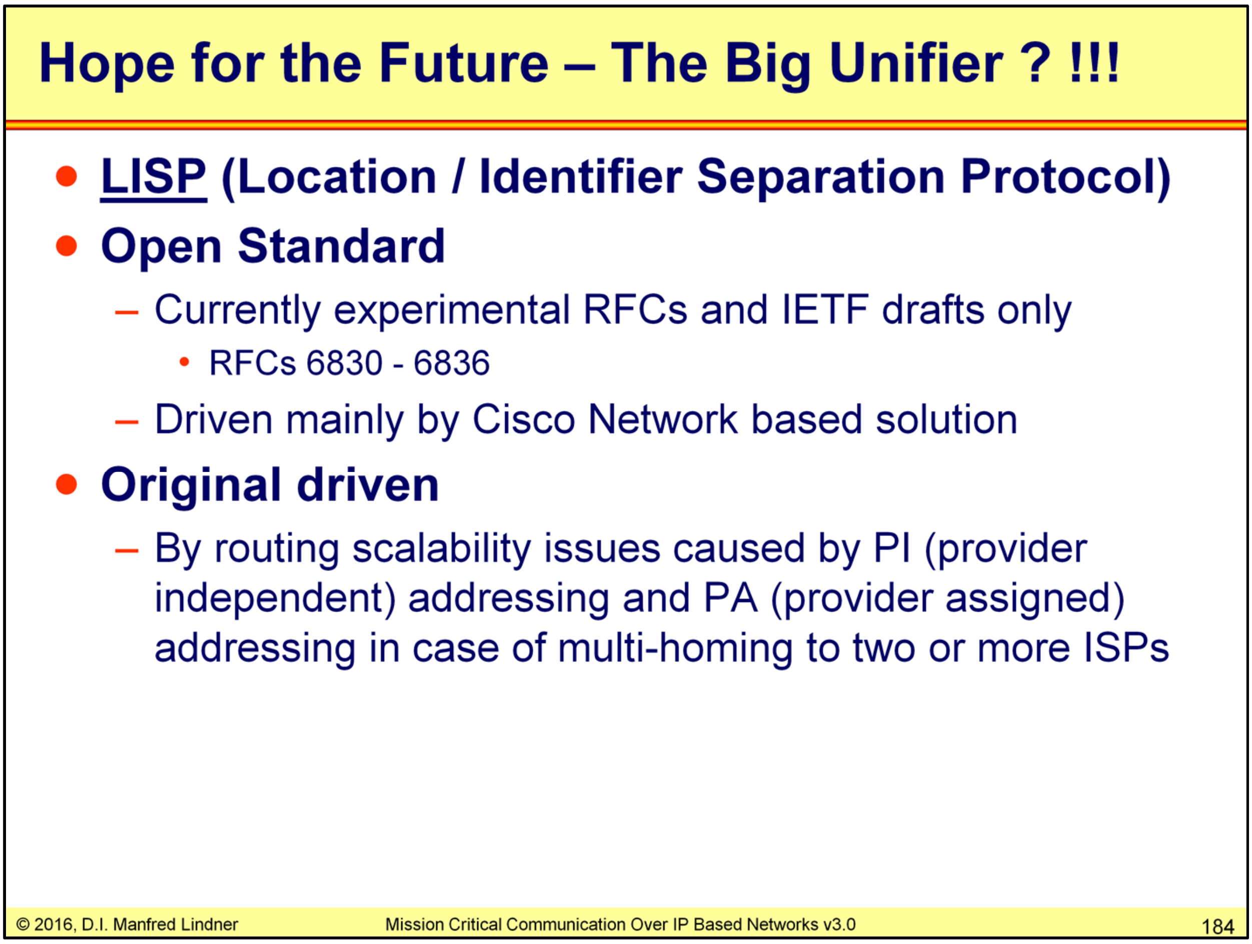
# Topics For Network Design 2

- **Security**
  - Security Concept -> Security Domains, Security Responsibilities
  - Identifying Location of Perimeter and Tunnel Mechanism
  - Agree on Security Management
- **Multicast (optional if appropriate)**
  - Group Address Plan
  - Multicast Routing Concept
  - Routing Convergence Tuning
- **Management**
  - Monitoring
  - Security
  - QoS



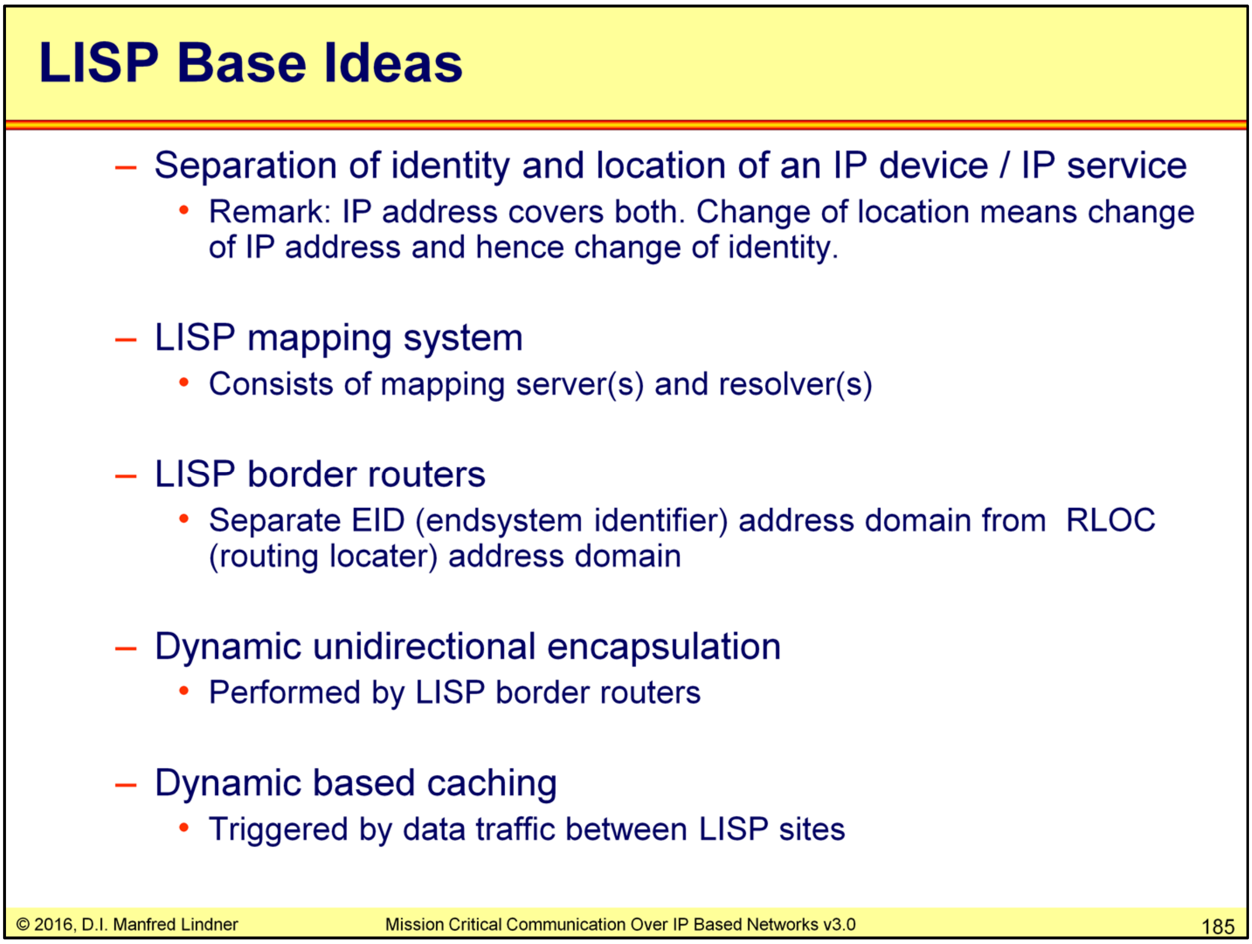
## Topics For Network Design 3

- **Holistic looking to all these topics is necessary**
  - All these topics must fit together
  - Tradeoffs will be seen and compromises have to be agreed
  - Design will not emerge in straight-forward way
  - Fact-finding missions and feedback loops will be necessary



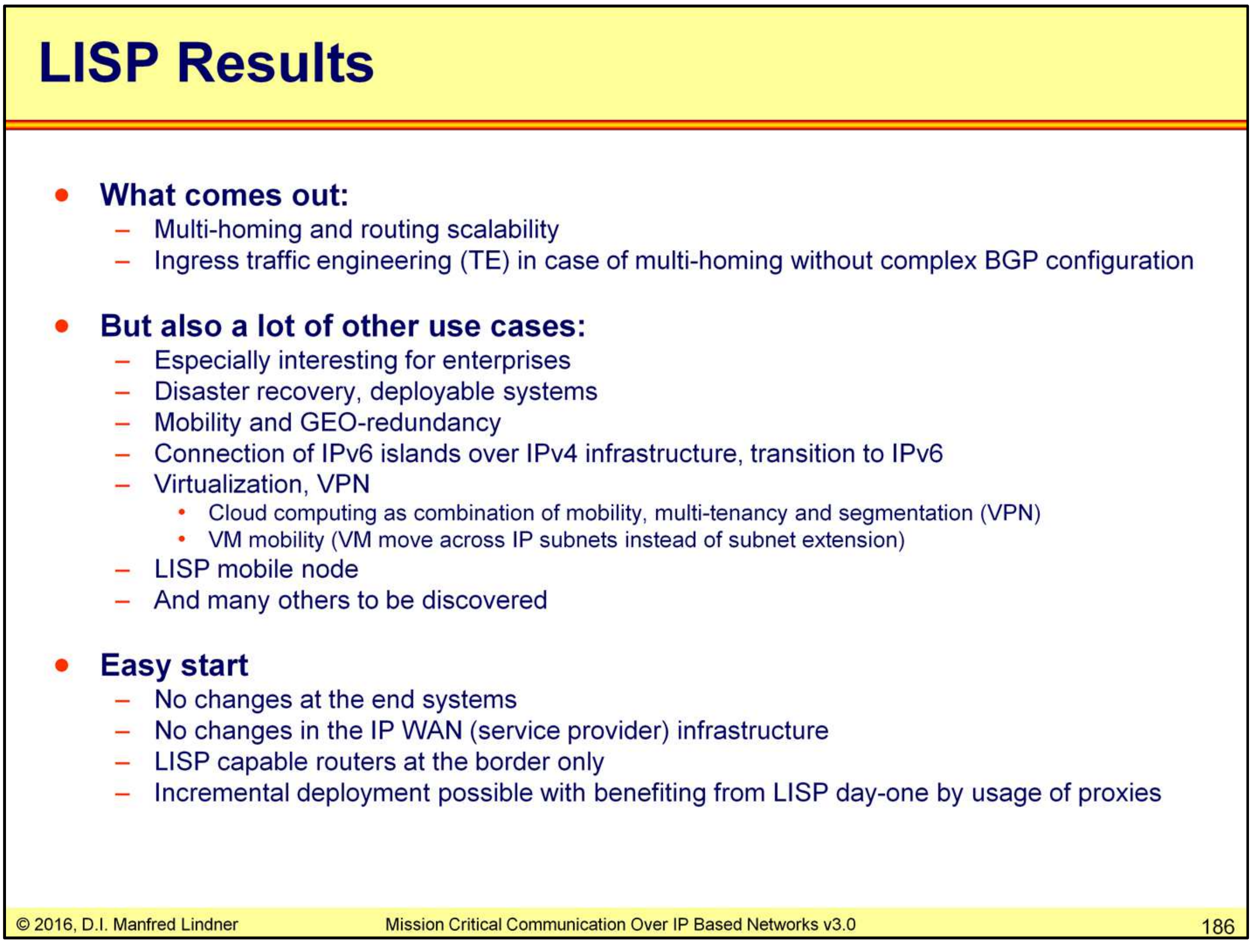
# Hope for the Future – The Big Unifier ? !!!

- LISP (Location / Identifier Separation Protocol)
- Open Standard
  - Currently experimental RFCs and IETF drafts only
    - RFCs 6830 - 6836
  - Driven mainly by Cisco Network based solution
- Original driven
  - By routing scalability issues caused by PI (provider independent) addressing and PA (provider assigned) addressing in case of multi-homing to two or more ISPs



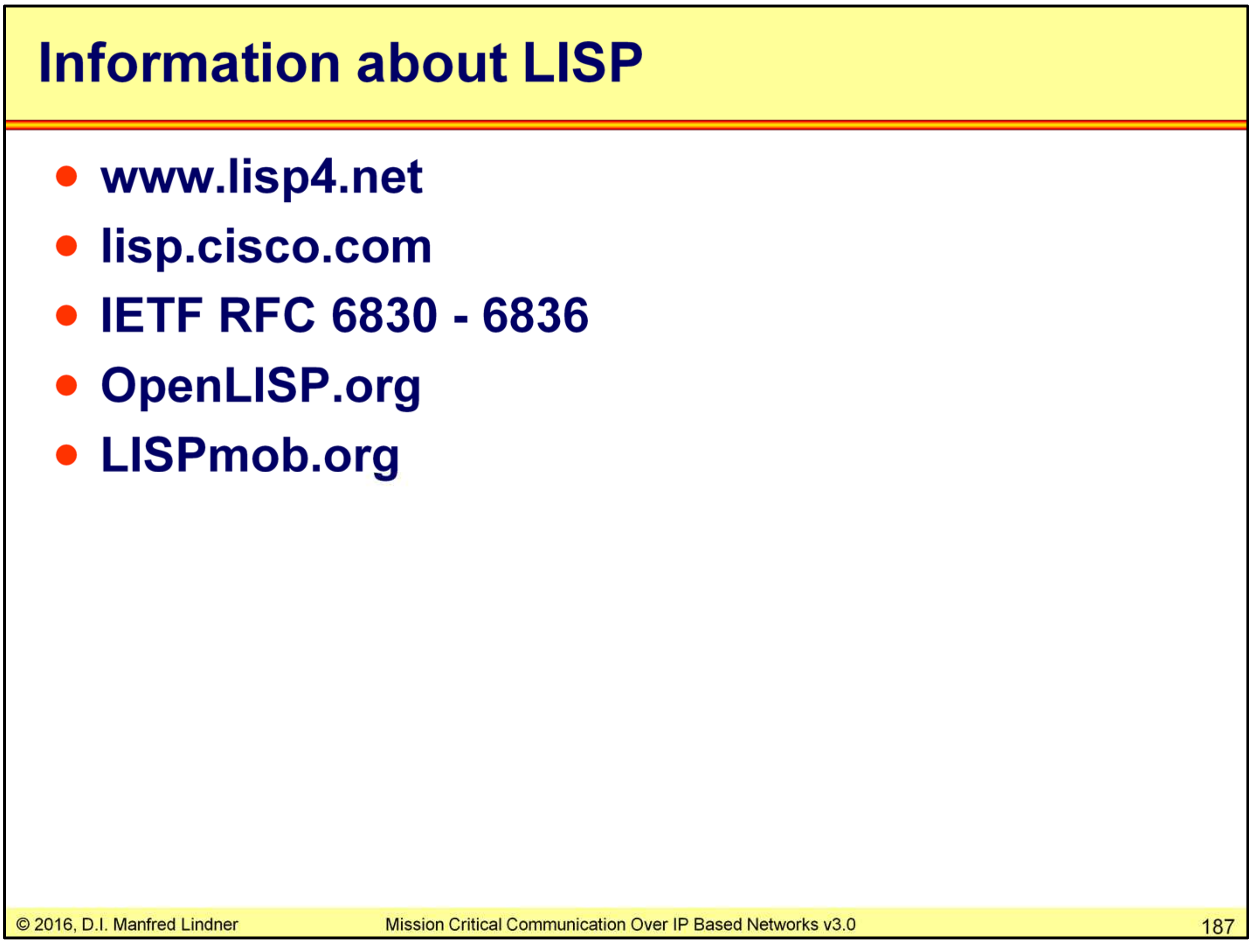
## LISP Base Ideas

- Separation of identity and location of an IP device / IP service
  - Remark: IP address covers both. Change of location means change of IP address and hence change of identity.
- LISP mapping system
  - Consists of mapping server(s) and resolver(s)
- LISP border routers
  - Separate EID (endsystem identifier) address domain from RLOC (routing locater) address domain
- Dynamic unidirectional encapsulation
  - Performed by LISP border routers
- Dynamic based caching
  - Triggered by data traffic between LISP sites



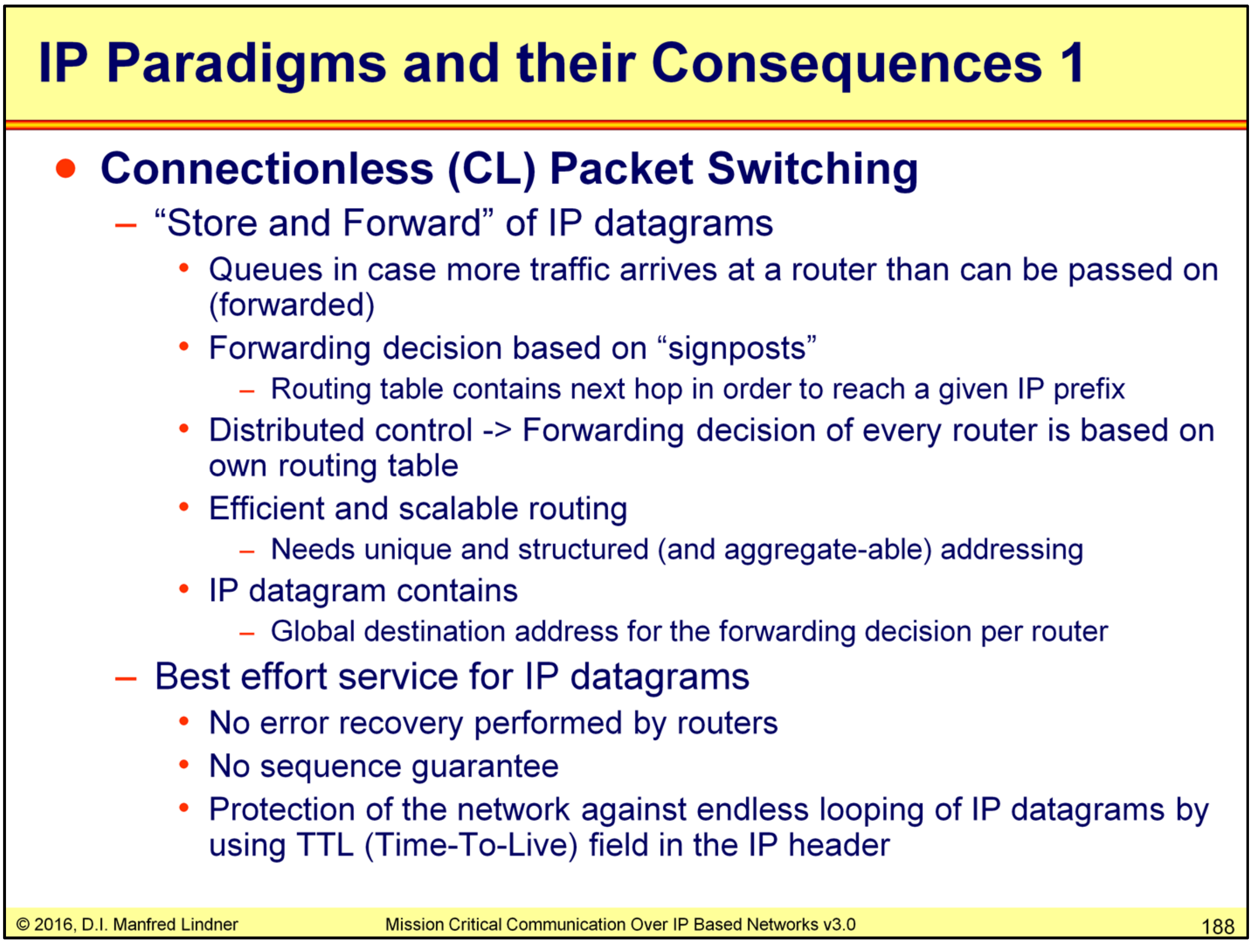
# LISP Results

- **What comes out:**
  - Multi-homing and routing scalability
  - Ingress traffic engineering (TE) in case of multi-homing without complex BGP configuration
- **But also a lot of other use cases:**
  - Especially interesting for enterprises
  - Disaster recovery, deployable systems
  - Mobility and GEO-redundancy
  - Connection of IPv6 islands over IPv4 infrastructure, transition to IPv6
  - Virtualization, VPN
    - Cloud computing as combination of mobility, multi-tenancy and segmentation (VPN)
    - VM mobility (VM move across IP subnets instead of subnet extension)
  - LISP mobile node
  - And many others to be discovered
- **Easy start**
  - No changes at the end systems
  - No changes in the IP WAN (service provider) infrastructure
  - LISP capable routers at the border only
  - Incremental deployment possible with benefiting from LISP day-one by usage of proxies



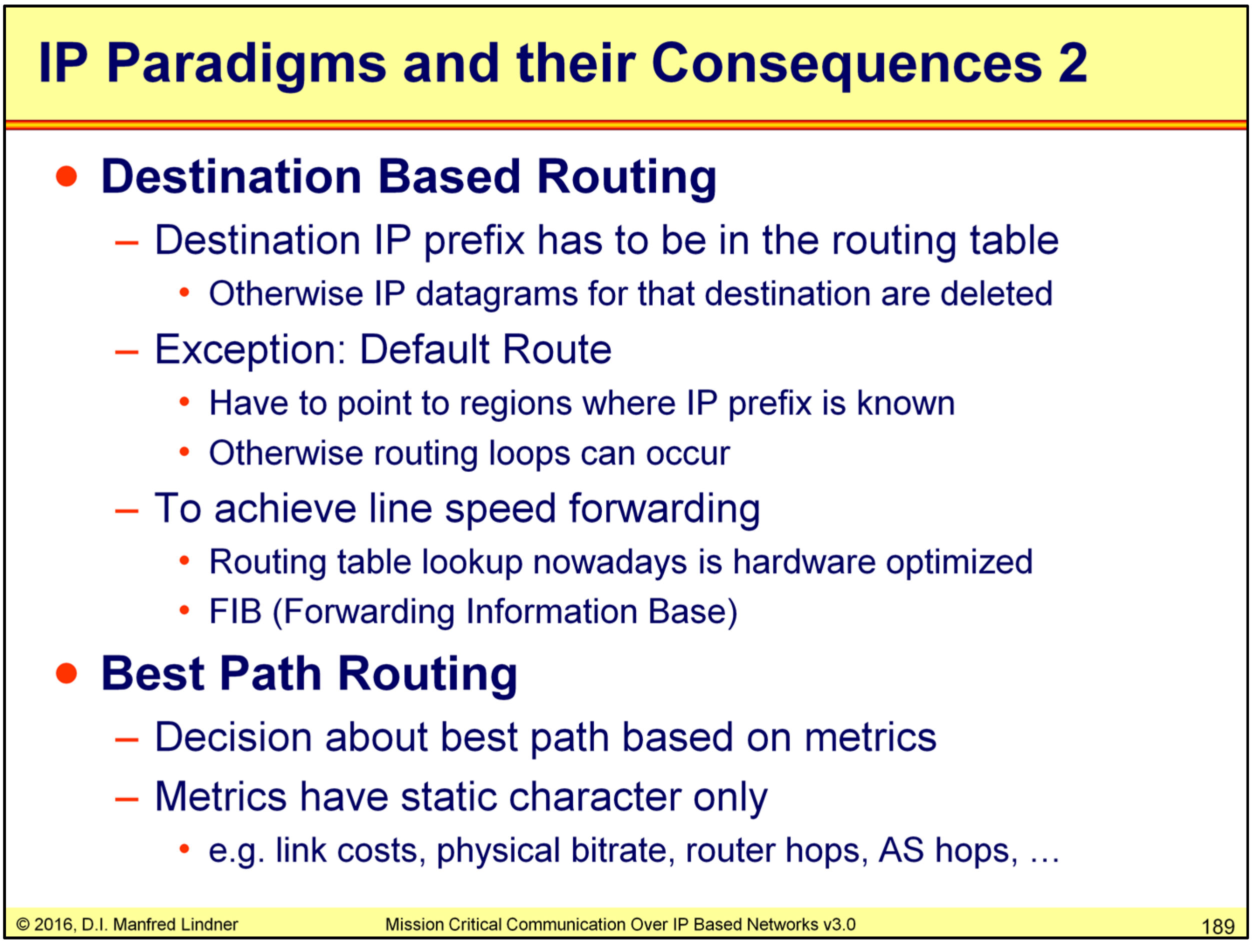
## Information about LISP

- www.lisp4.net
- lisp.cisco.com
- IETF RFC 6830 - 6836
- OpenLISP.org
- LISPmob.org



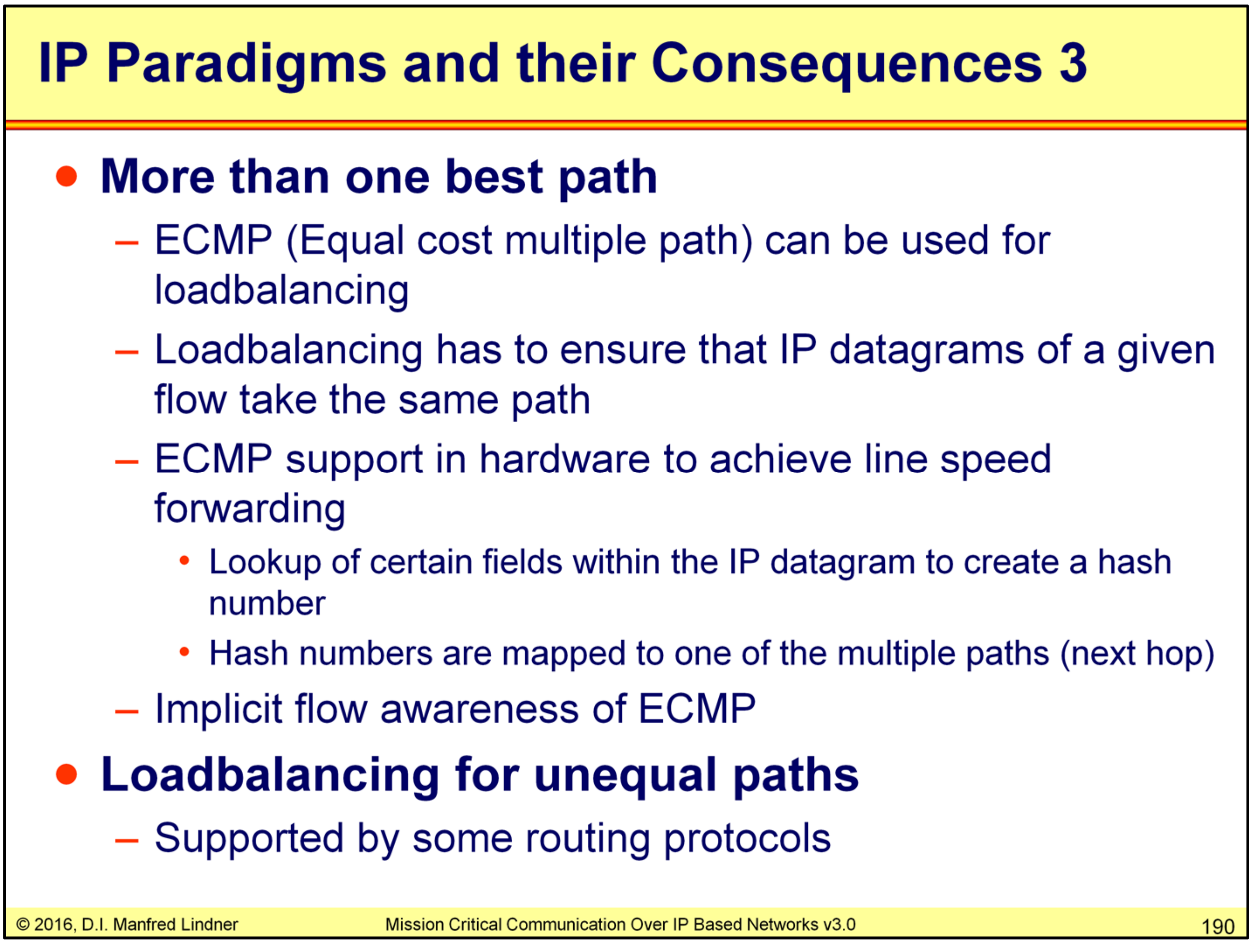
# IP Paradigms and their Consequences 1

- **Connectionless (CL) Packet Switching**
  - “Store and Forward” of IP datagrams
    - Queues in case more traffic arrives at a router than can be passed on (forwarded)
    - Forwarding decision based on “signposts”
      - Routing table contains next hop in order to reach a given IP prefix
    - Distributed control -> Forwarding decision of every router is based on own routing table
    - Efficient and scalable routing
      - Needs unique and structured (and aggregate-able) addressing
    - IP datagram contains
      - Global destination address for the forwarding decision per router
  - Best effort service for IP datagrams
    - No error recovery performed by routers
    - No sequence guarantee
    - Protection of the network against endless looping of IP datagrams by using TTL (Time-To-Live) field in the IP header



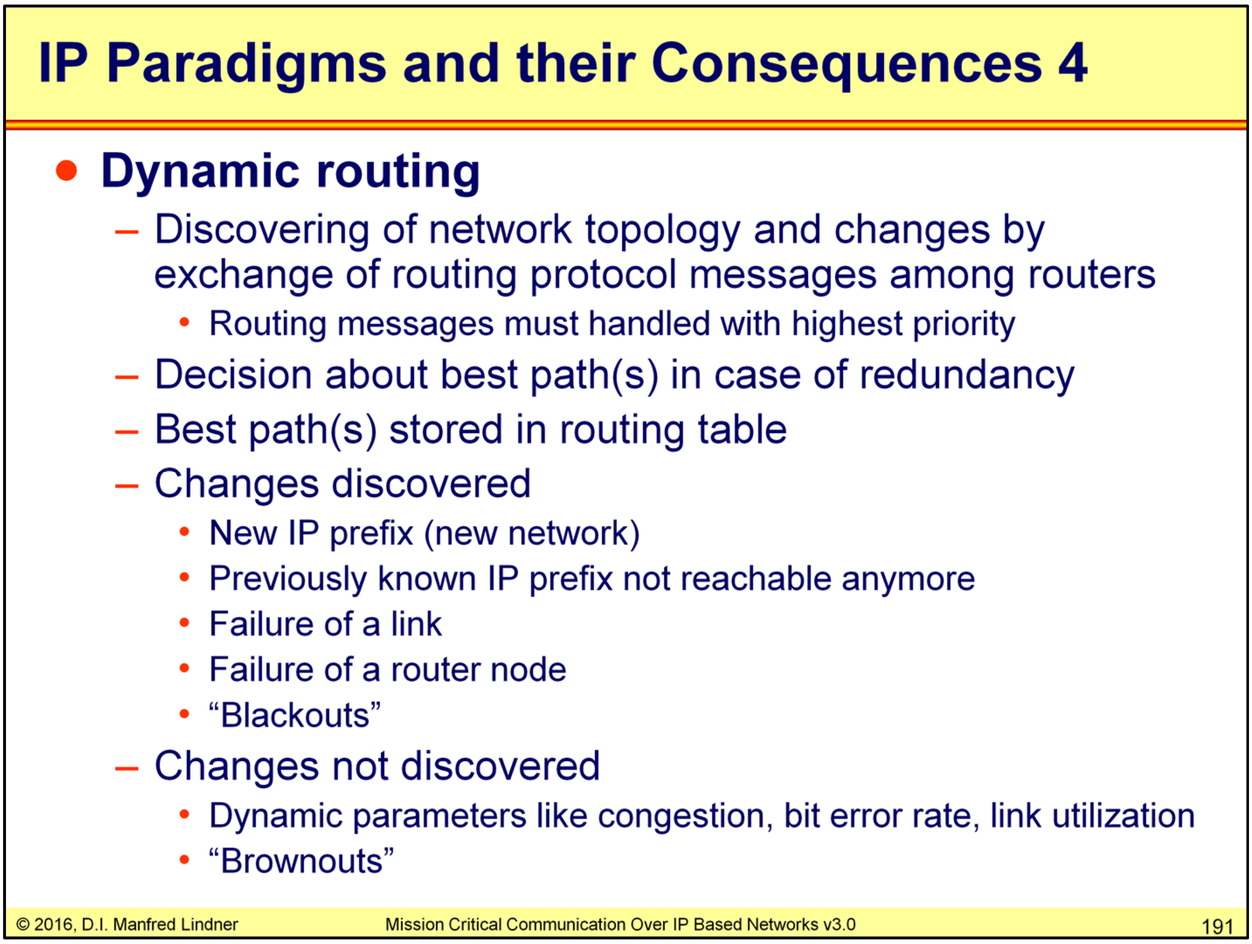
# IP Paradigms and their Consequences 2

- **Destination Based Routing**
  - Destination IP prefix has to be in the routing table
    - Otherwise IP datagrams for that destination are deleted
  - Exception: Default Route
    - Have to point to regions where IP prefix is known
    - Otherwise routing loops can occur
  - To achieve line speed forwarding
    - Routing table lookup nowadays is hardware optimized
    - FIB (Forwarding Information Base)
- **Best Path Routing**
  - Decision about best path based on metrics
  - Metrics have static character only
    - e.g. link costs, physical bitrate, router hops, AS hops, …



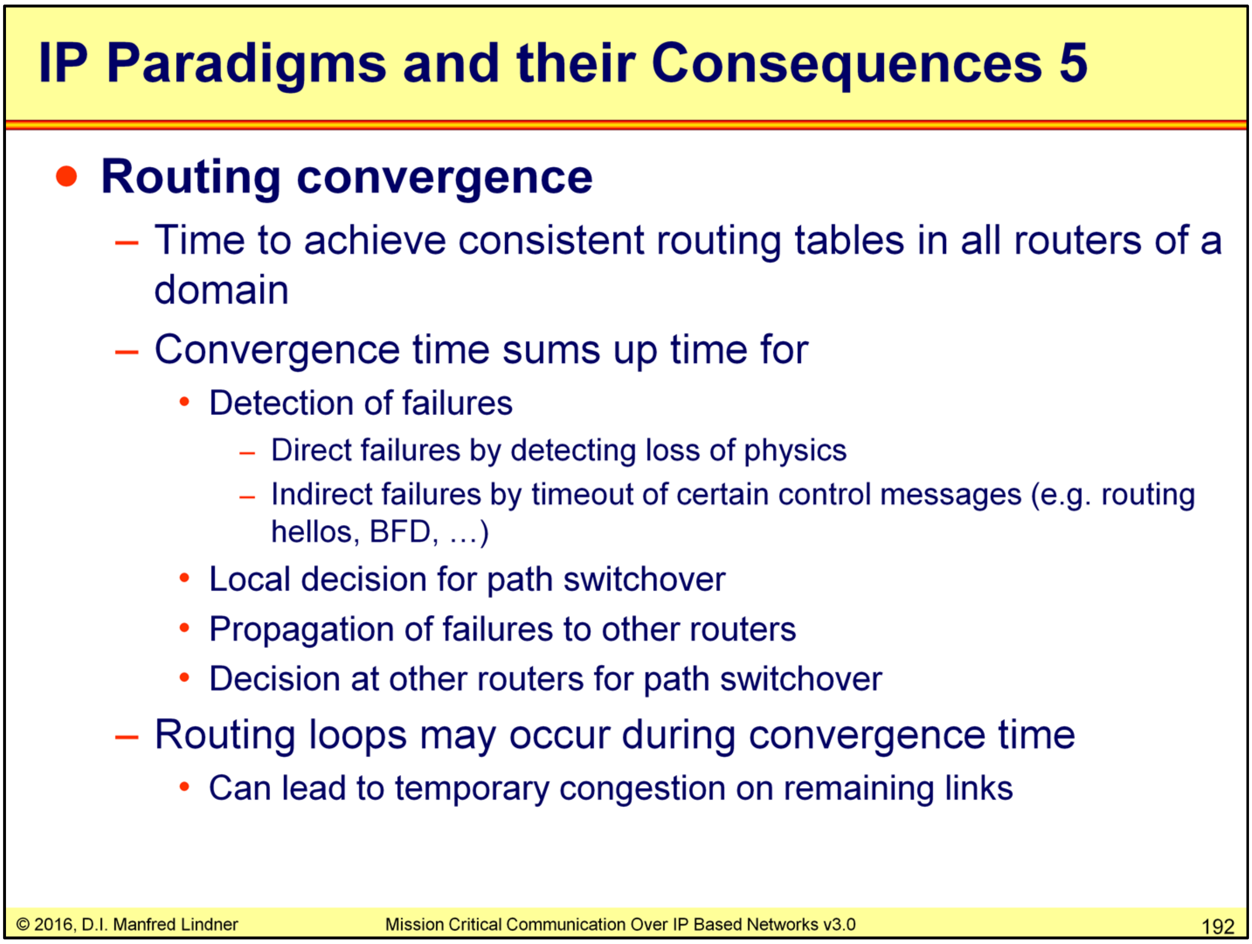
## IP Paradigms and their Consequences 3

- **More than one best path**
  - ECMP (Equal cost multiple path) can be used for loadbalancing
  - Loadbalancing has to ensure that IP datagrams of a given flow take the same path
  - ECMP support in hardware to achieve line speed forwarding
    - Lookup of certain fields within the IP datagram to create a hash number
    - Hash numbers are mapped to one of the multiple paths (next hop)
  - Implicit flow awareness of ECMP
- **Loadbalancing for unequal paths**
  - Supported by some routing protocols



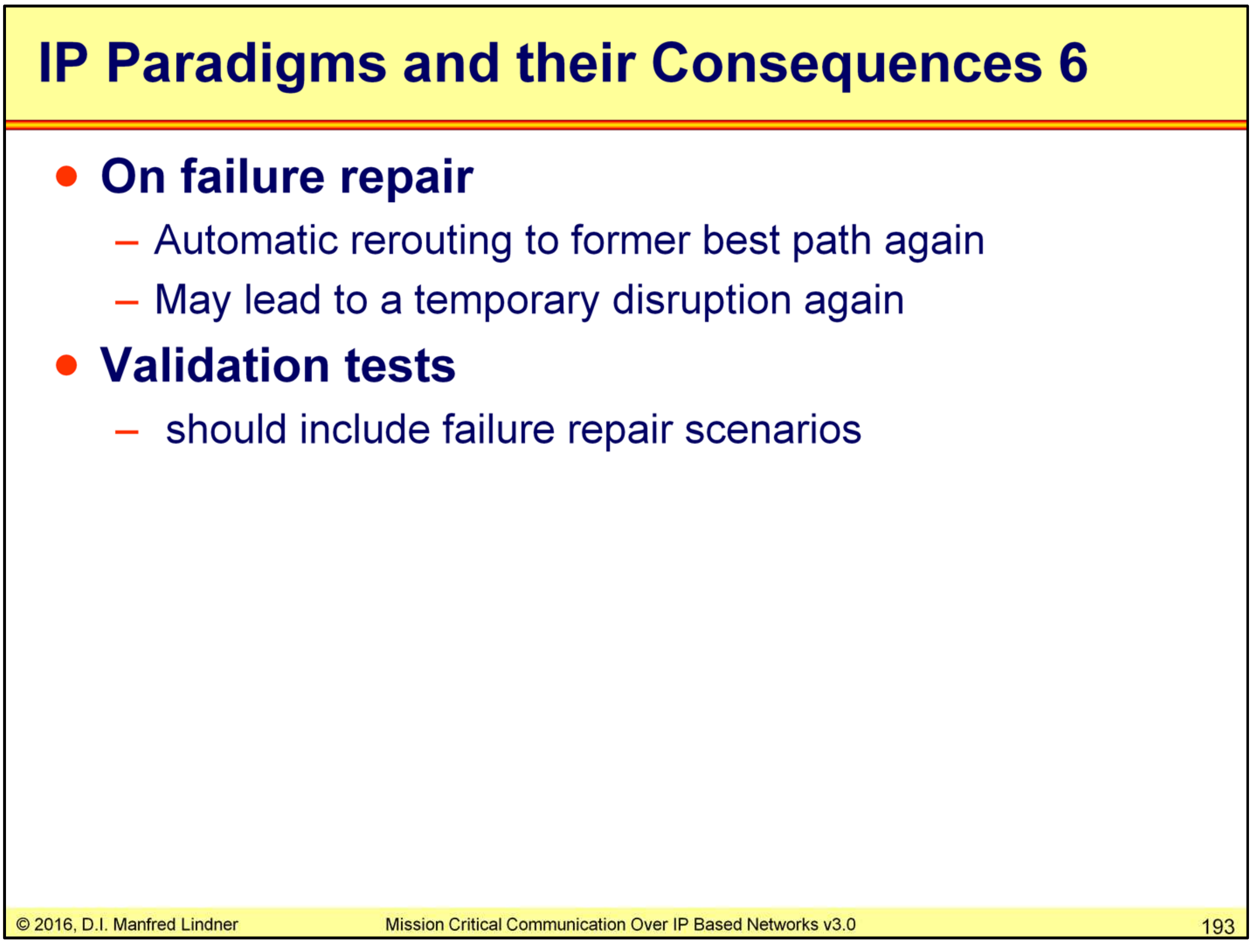
# IP Paradigms and their Consequences 4

- **Dynamic routing**
  - Discovering of network topology and changes by exchange of routing protocol messages among routers
    - Routing messages must handled with highest priority
  - Decision about best path(s) in case of redundancy
  - Best path(s) stored in routing table
  - Changes discovered
    - New IP prefix (new network)
    - Previously known IP prefix not reachable anymore
    - Failure of a link
    - Failure of a router node
    - “Blackouts”
  - Changes not discovered
    - Dynamic parameters like congestion, bit error rate, link utilization
    - “Brownouts”



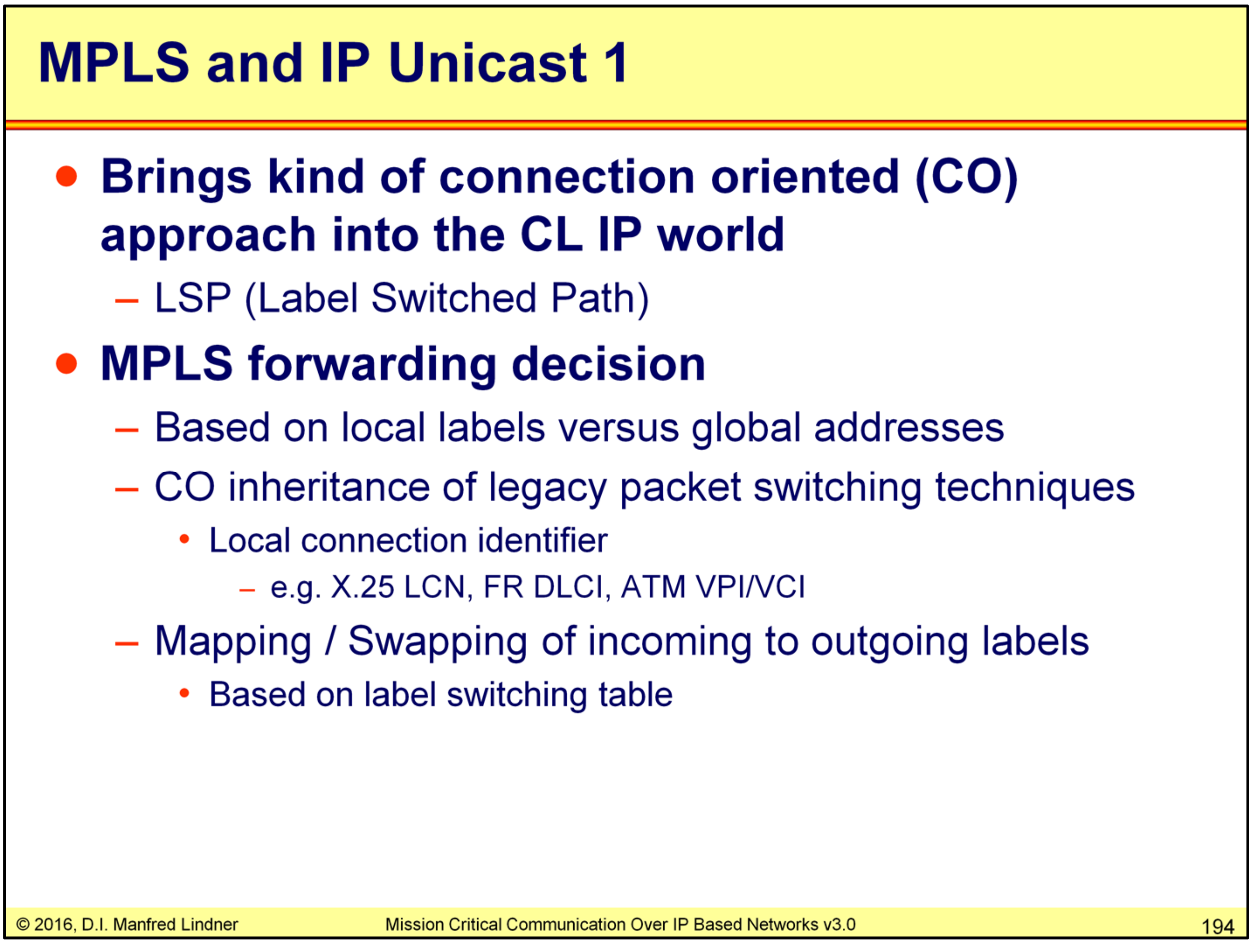
# IP Paradigms and their Consequences 5

- **Routing convergence**
  - Time to achieve consistent routing tables in all routers of a domain
  - Convergence time sums up time for
    - Detection of failures
      - Direct failures by detecting loss of physics
      - Indirect failures by timeout of certain control messages (e.g. routing hellos, BFD, …)
    - Local decision for path switchover
    - Propagation of failures to other routers
    - Decision at other routers for path switchover
  - Routing loops may occur during convergence time
    - Can lead to temporary congestion on remaining links



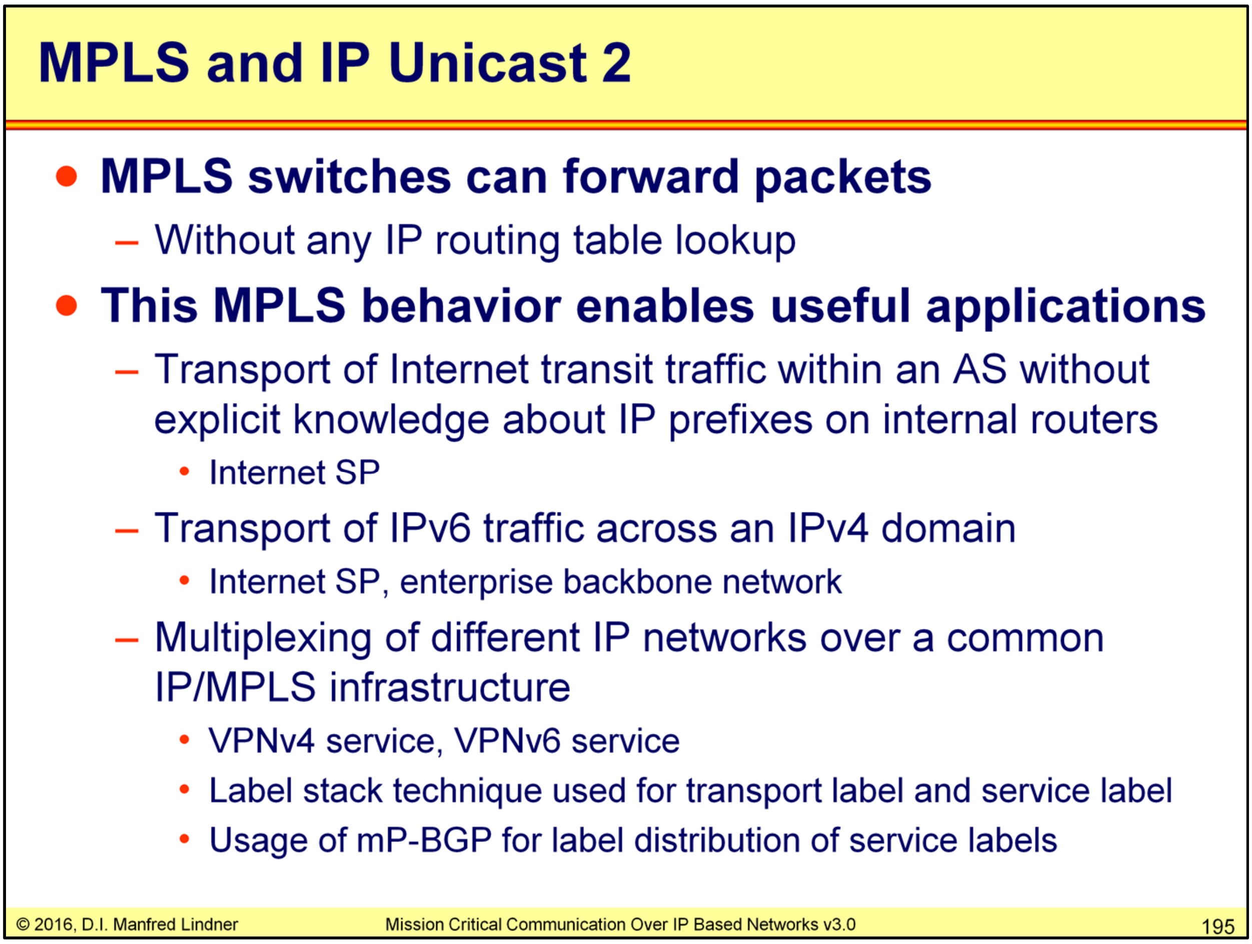
# IP Paradigms and their Consequences 6

- **On failure repair**
  - Automatic rerouting to former best path again
  - May lead to a temporary disruption again
- **Validation tests**
  - should include failure repair scenarios



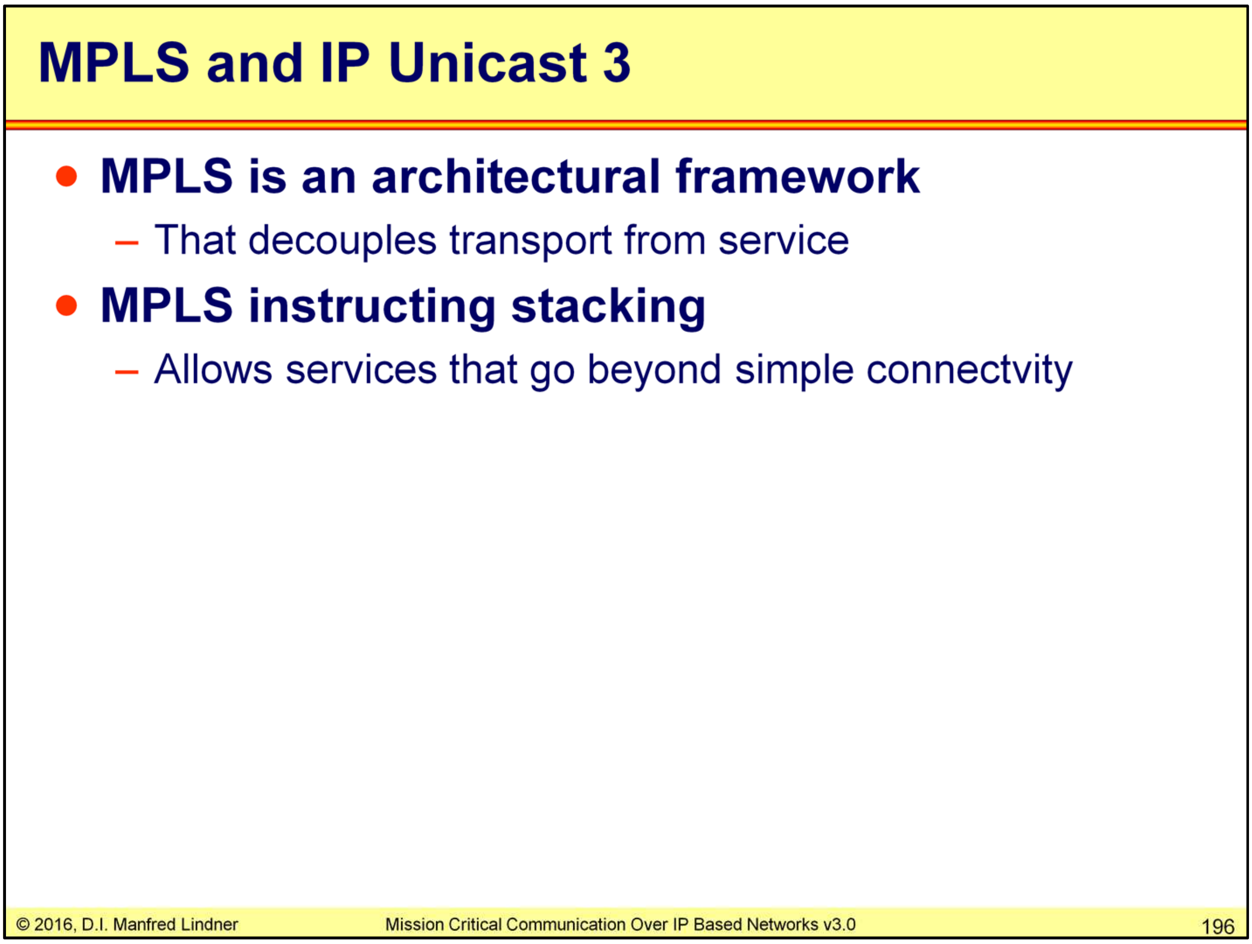
## MPLS and IP Unicast 1

- **Brings kind of connection oriented (CO) approach into the CL IP world**
  - LSP (Label Switched Path)
- **MPLS forwarding decision**
  - Based on local labels versus global addresses
  - CO inheritance of legacy packet switching techniques
    - Local connection identifier
      - e.g. X.25 LCN, FR DLCI, ATM VPI/VCI
  - Mapping / Swapping of incoming to outgoing labels
    - Based on label switching table



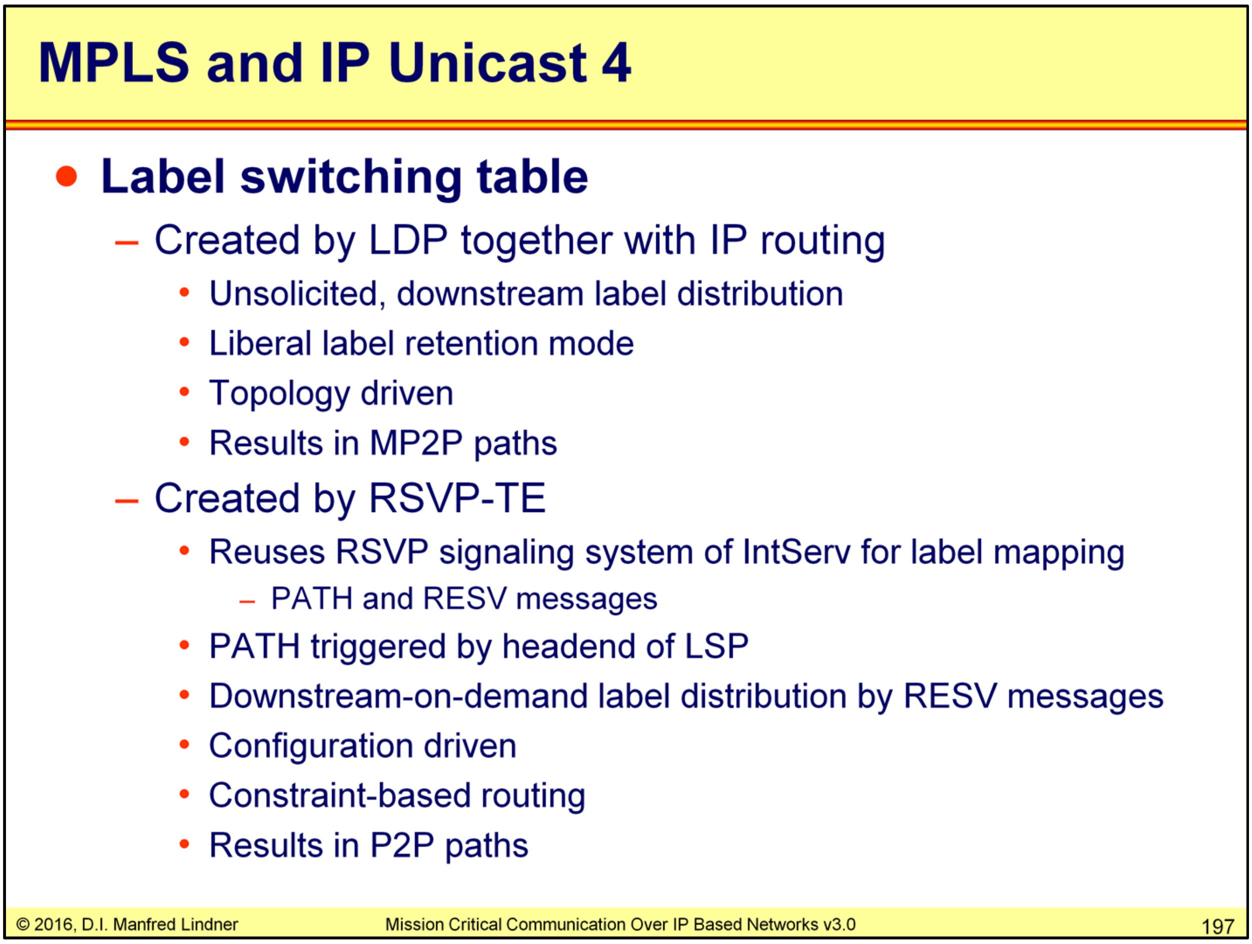
## MPLS and IP Unicast 2

- **MPLS switches can forward packets**
  - Without any IP routing table lookup
- **This MPLS behavior enables useful applications**
  - Transport of Internet transit traffic within an AS without explicit knowledge about IP prefixes on internal routers
    - Internet SP
  - Transport of IPv6 traffic across an IPv4 domain
    - Internet SP, enterprise backbone network
  - Multiplexing of different IP networks over a common IP/MPLS infrastructure
    - VPNv4 service, VPNv6 service
    - Label stack technique used for transport label and service label
    - Usage of mP-BGP for label distribution of service labels



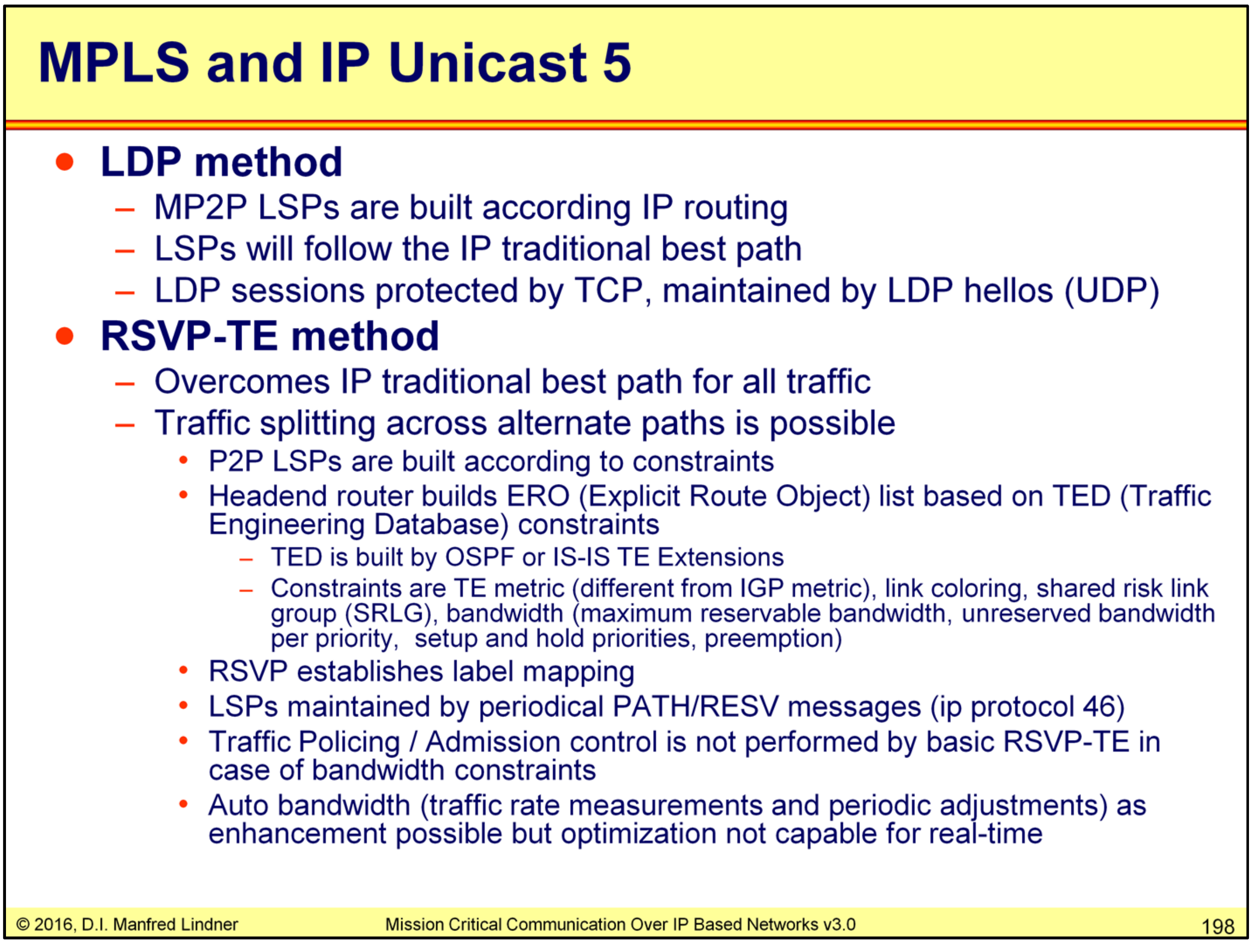
## MPLS and IP Unicast 3

- **MPLS is an architectural framework**
  - That decouples transport from service
- **MPLS instructing stacking**
  - Allows services that go beyond simple connectvity



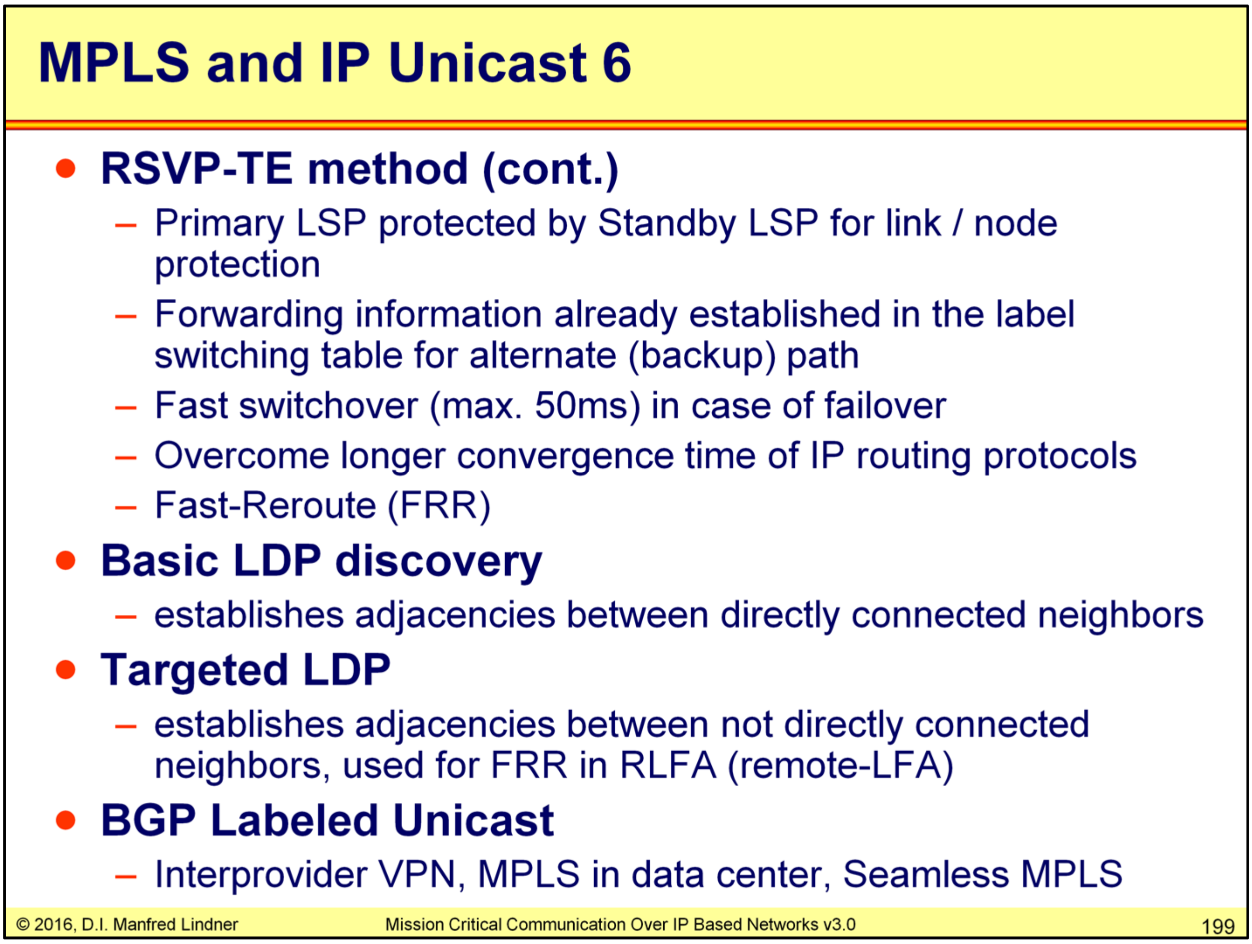
# MPLS and IP Unicast 4

- **Label switching table**
  - Created by LDP together with IP routing
    - Unsolicited, downstream label distribution
    - Liberal label retention mode
    - Topology driven
    - Results in MP2P paths
  - Created by RSVP-TE
    - Reuses RSVP signaling system of IntServ for label mapping
      - PATH and RESV messages
    - PATH triggered by headend of LSP
    - Downstream-on-demand label distribution by RESV messages
    - Configuration driven
    - Constraint-based routing
    - Results in P2P paths



## MPLS and IP Unicast 5

- **LDP method**
  - MP2P LSPs are built according IP routing
  - LSPs will follow the IP traditional best path
  - LDP sessions protected by TCP, maintained by LDP hellos (UDP)
- **RSVP-TE method**
  - Overcomes IP traditional best path for all traffic
  - Traffic splitting across alternate paths is possible
    - P2P LSPs are built according to constraints
    - Headend router builds ERO (Explicit Route Object) list based on TED (Traffic Engineering Database) constraints
      - TED is built by OSPF or IS-IS TE Extensions
      - Constraints are TE metric (different from IGP metric), link coloring, shared risk link group (SRLG), bandwidth (maximum reservable bandwidth, unreserved bandwidth per priority, setup and hold priorities, preemption)
    - RSVP establishes label mapping
    - LSPs maintained by periodical PATH/RESV messages (ip protocol 46)
    - Traffic Policing / Admission control is not performed by basic RSVP-TE in case of bandwidth constraints
    - Auto bandwidth (traffic rate measurements and periodic adjustments) as enhancement possible but optimization not capable for real-time



# MPLS and IP Unicast 6

- **RSVP-TE method (cont.)**
  - Primary LSP protected by Standby LSP for link / node protection
  - Forwarding information already established in the label switching table for alternate (backup) path
  - Fast switchover (max. 50ms) in case of failover
  - Overcome longer convergence time of IP routing protocols
  - Fast-Reroute (FRR)
- **Basic LDP discovery**
  - establishes adjacencies between directly connected neighbors
- **Targeted LDP**
  - establishes adjacencies between not directly connected neighbors, used for FRR in RLFA (remote-LFA)
- **BGP Labeled Unicast**
  - Interprovider VPN, MPLS in data center, Seamless MPLS