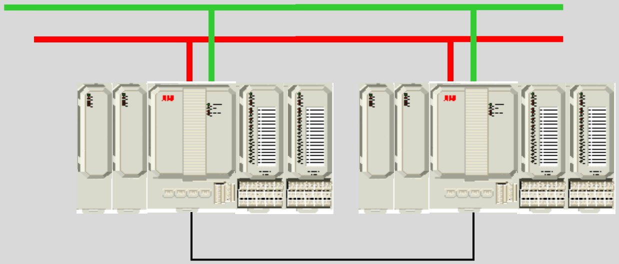
INTERNATIONAL
ELECTROTECHNICAL
COMMISSION

picture from shawnhallwachs' photostream

IEC 62439-3 §4

PRP - Parallel Redundancy Protocol

An IEC standard for a seamless redundancy method using parallel networks, applicable to hard-real time Industrial Ethernet



Prof. Dr. Hubert Kirrmann, Solutil, Switzerland

2019 February 17

Highly available automation networks

Automation networks require a high availability to ensure continuous plant operation.

Beyond high quality elements and good maintenance, availability is increased by *redundant elements* (links, bridges, interfaces or devices).

Redundancy activation after a failure by traditional protocols such as RSTP cost *recovery delays*, during which the network is unavailable. The plant must be able to tolerate such interruption of service.
This applies not only to fault situations, but also to removal and insertion of redundancy.

IEC SC65C WG15 standardized in the IEC 62439 Suite several methods to implement high availability networks, divided into two main categories:

- "redundancy in the network", using singly attached devices, i.e. devices attached to a single bridge only, while the bridges implement redundancy (e.g. using RSTP or a Ring Protocol such as MRP)

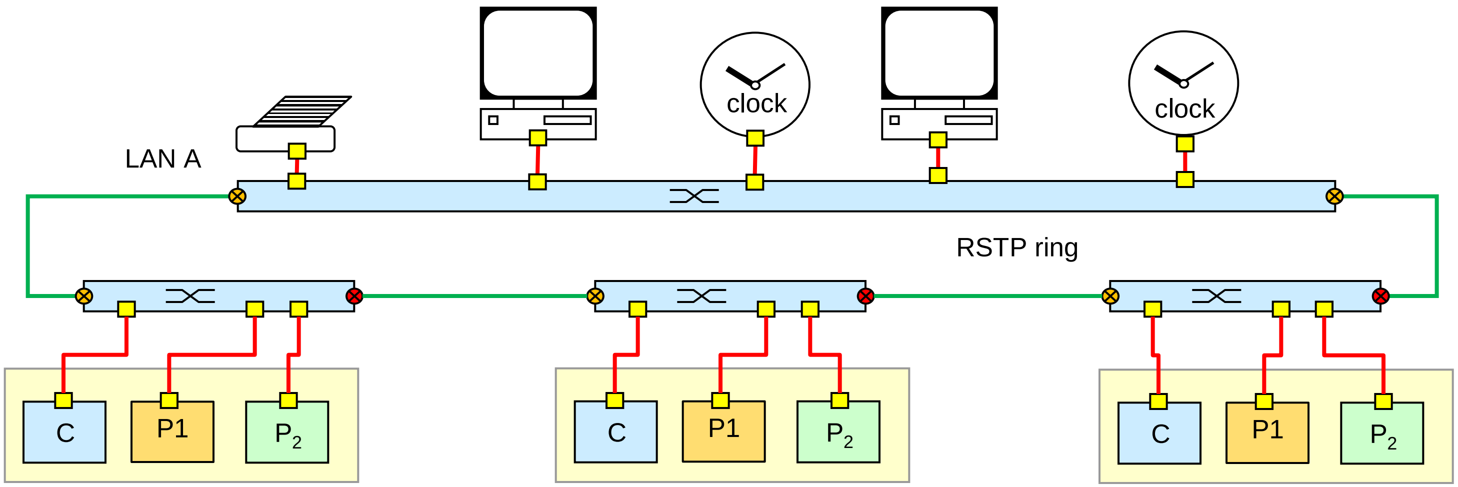
- "redundancy in the devices", using doubly attached devices, i.e. devices with two network interfaces attached to redundant networks elements

This presentation describes the **Parallel Redundancy Protocol** (PRP), a "redundancy in the devices" method, that provides seamless operation in case of failure or reintegration.

Redundancy in the network and redundancy in the nodes

Redundancy in the network

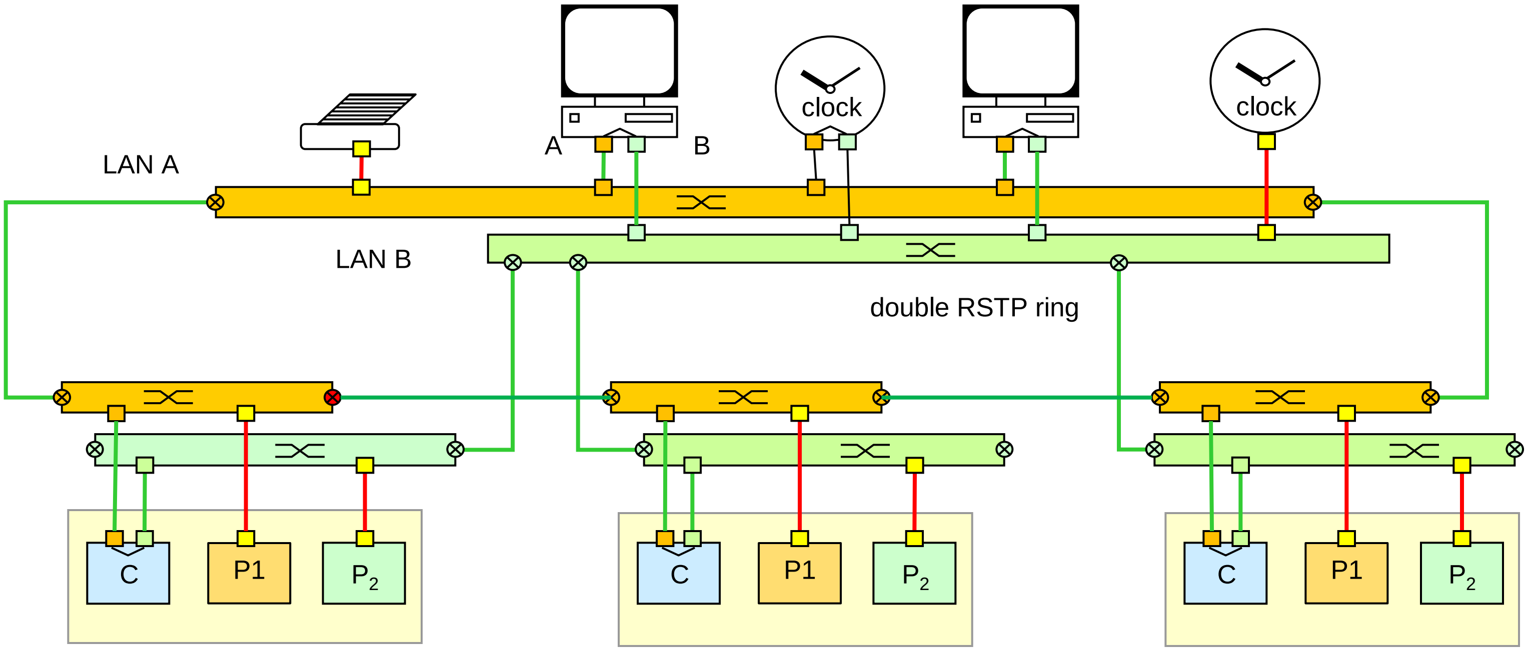
devices attached to the RSTP ring can sustain the failure of any of the four ring links (green), but there are 14 unsecured links (red). Availability is not improvement significantly



Redundancy in the nodes

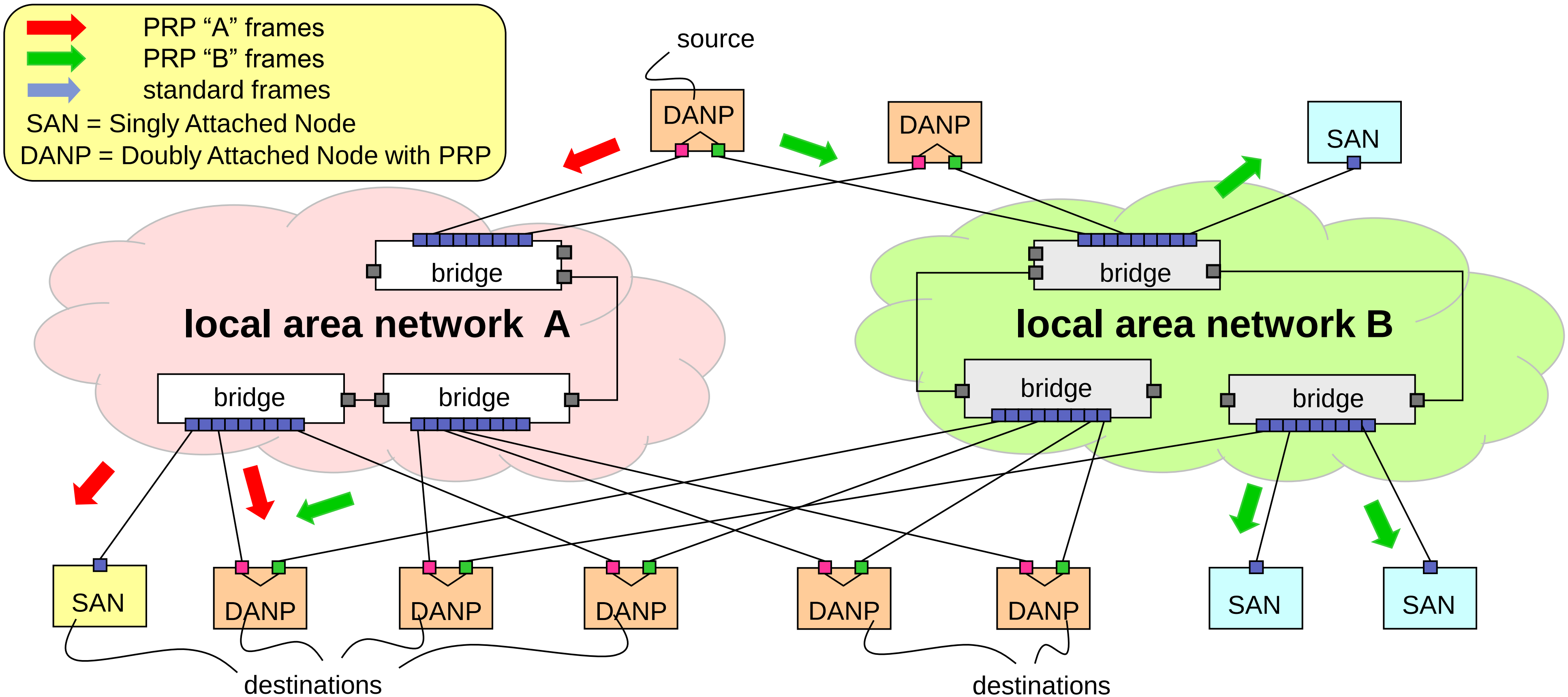
all doubly attached nodes can overcome any failure of a link (green) or bridge.

Only doubly attached nodes improve availability significantly.
Singly attached nodes (red links) are not protected.



PRP principles

PRP Principle Network Architecture



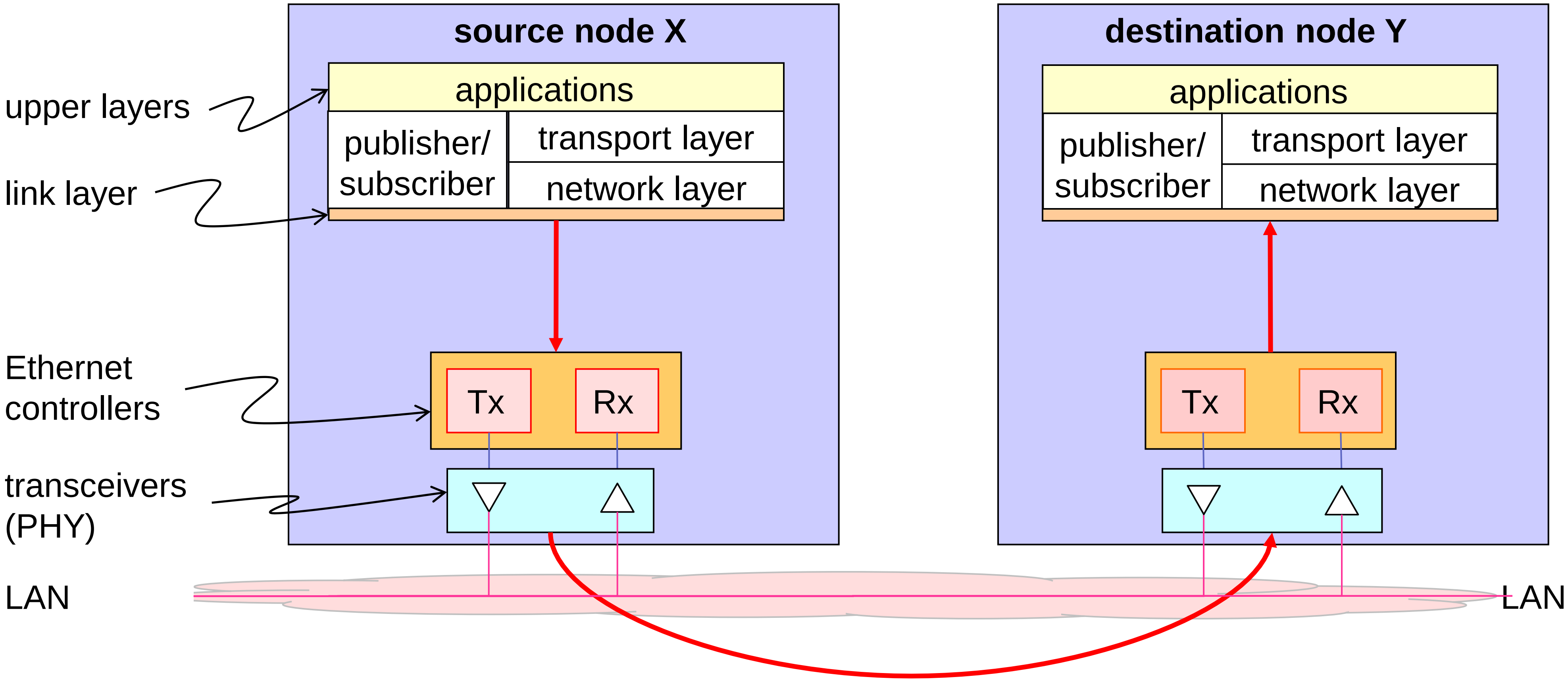
Two Ethernet networks (LANs) , completely separated, of similar topology operate in parallel.
Each doubly attached node with PRP (=DANP) has an interface to each LAN.
A DANP source sends a frame simultaneously on both LANs.
A DANP destination receives in normal operation both frames and discards the duplicate.
A singly attached destination only receives one.
If a LAN fails, a DANP destination continues to operate with the frames from the other LAN.

PRP characteristics

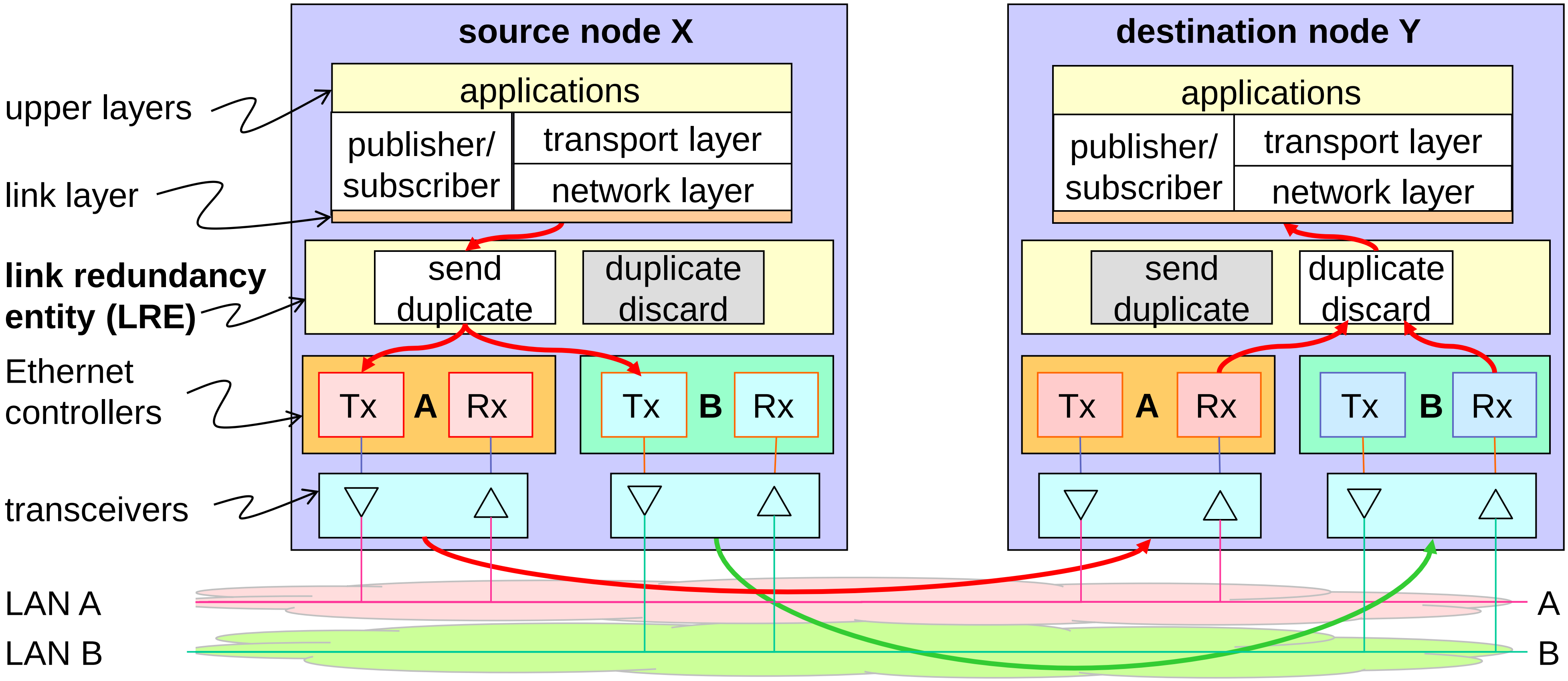
PRP allows to attach nodes to a network by redundant network interfaces and links:

- Tolerates any single network component (link, bridge) failure
- Provides seamless operation (zero nanoseconds switchover) in case of network failure.
- Supports time-critical communication
- Increases probability of successful delivery
- Is Invisible to the application
- Is applicable to any Industrial Ethernet on layer 2 or above
- Can be used with any topology (tree, ring,…)
- Does not rely on higher layer protocols to operate or be configured
- Allows mixing devices with single and double network attachment on the same LAN,
- Allows to connect laptops and workstations to the network with standard Ethernet adapters-
- Uses off-the shelf components (network interfaces, controllers, bridges and links)
- Supervises redundancy continuously for device management
- Is suitable for hot swap - 24h/365d operation in substations
- Designed particularly for substation automation, high-speed drives and transportation.

Singly attached nodes



PRP Node Operation



send on both LANs: the LRE sends each frame simultaneously on LAN A and LAN B.

transfer: Frames over LAN A and B transit with different delays (or could not arrive at all)

receive from both LANs: the node receives both frames, the LRE between link layer and Ethernet controllers handles the frames and can filter duplicates. Both lines are treated equal.

Why a layer 2 redundancy ?

Most industrial Ethernet networks (e.g. Profinet) operate with a layer 2 (Link Layer) protocol.

One motivation for this is the use of the publisher-subscriber method, that relies on broadcast of source-addressed data within the MAC broadcast domain.

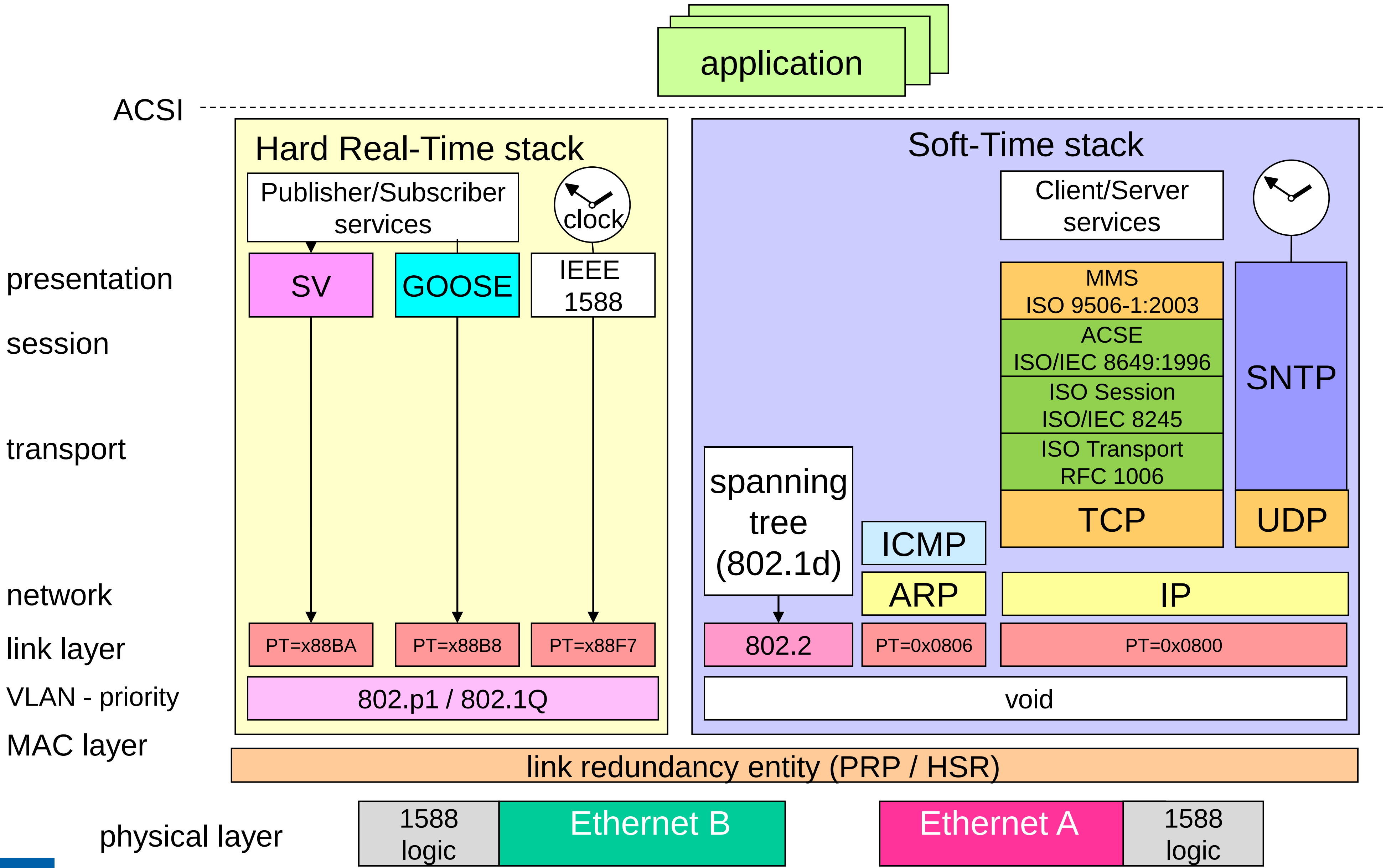
This excludes a redundancy scheme based on routers (Layer 3).

Each node in PRP has the same MAC address on both network interfaces.

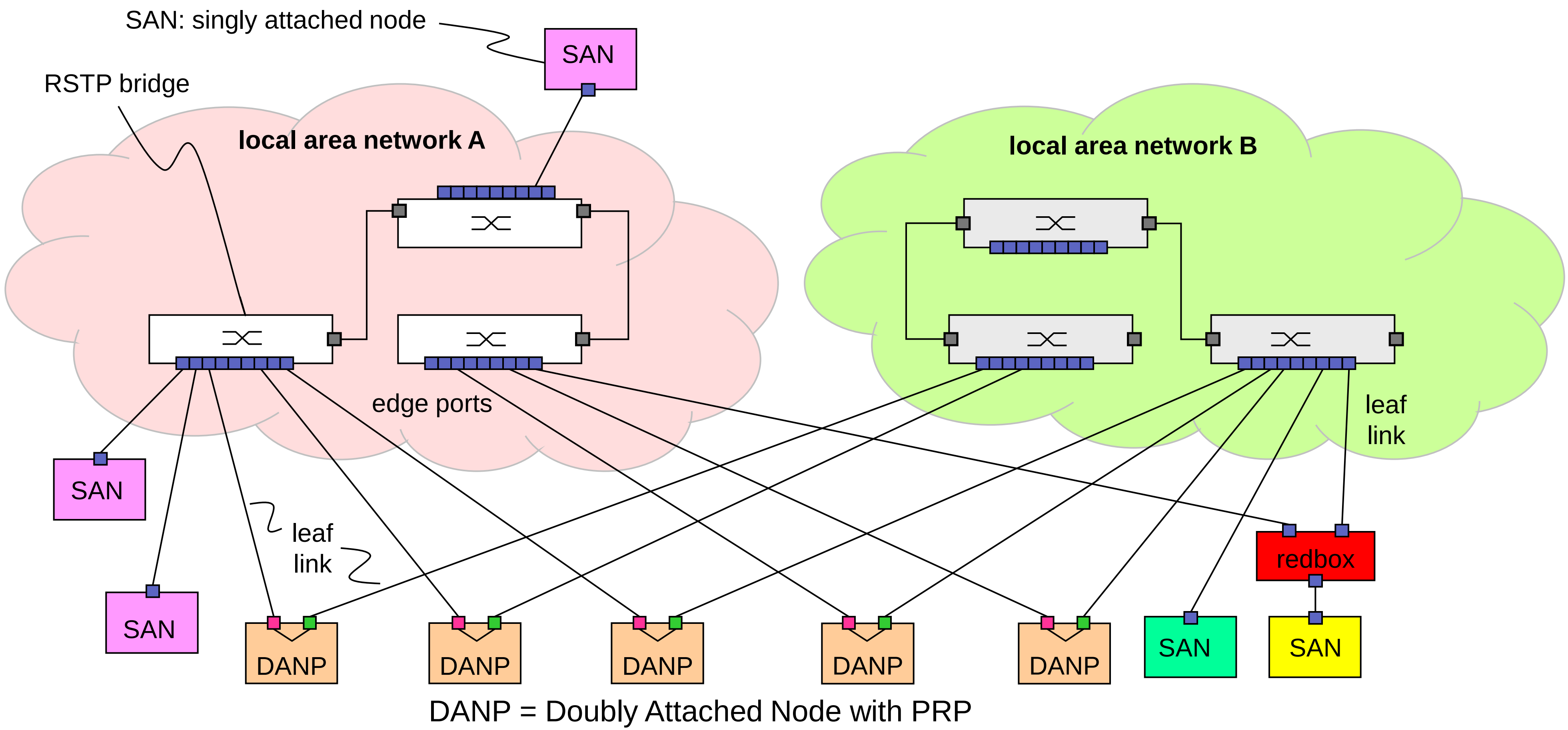
Therefore, management protocols such as ARP operate as usual and assign that MAC address to the IP address(es) of that node. Tools based on SNMP also operate as usual.

Redundancy methods must operate on Layer 2 (Logical Link) to address Industrial Ethernets with time-critical stacks built on Layer 2, such as IEC61850.

Typical Power Utility (IEC 61850) stack



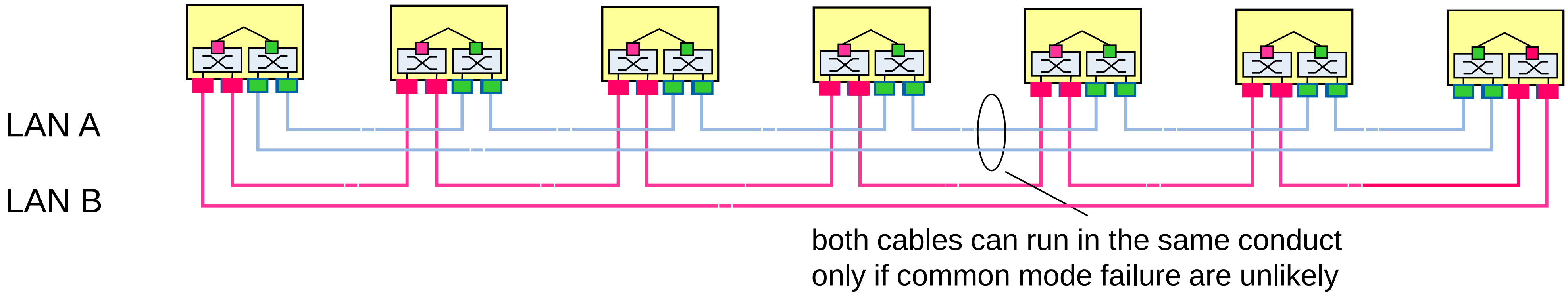
Attachment of non-redundant nodes (SAN)



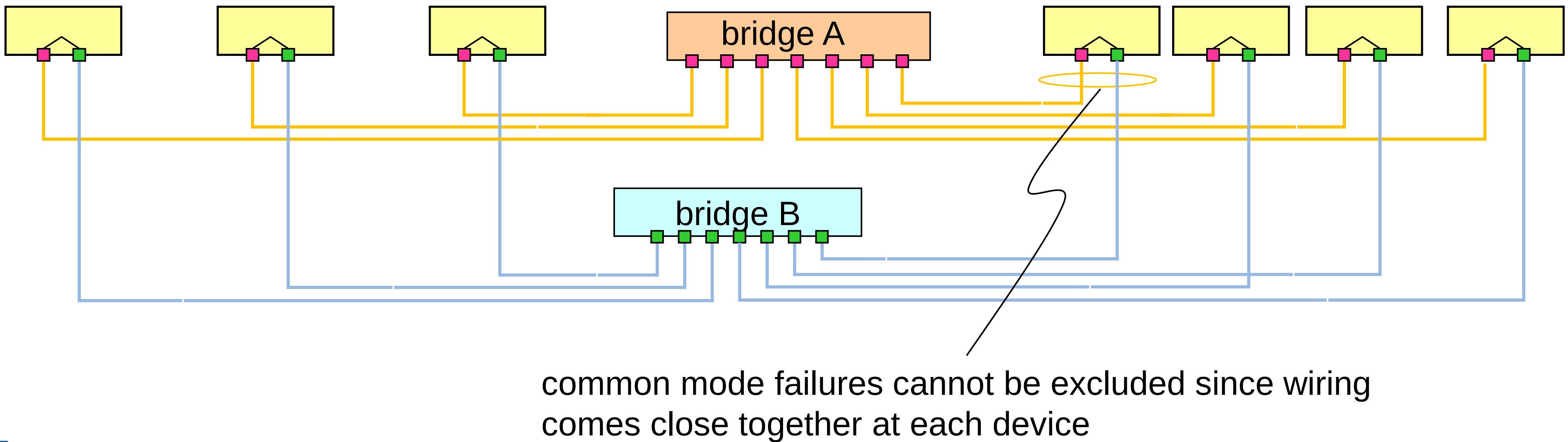
Singly attached nodes (SAN) are preferably all attached to the same network.
Network traffic is asymmetrical, but this does not affect redundant operation.
Singly attached nodes can also be attached through a “redundancy box” (**red box**)

PRP cabling rules: keep the LANs separate

Party-Line topology (bridge element in nodes)



Star topology (one bridge per LAN)



How PRP discards duplicates

Filtering: duplicate discard and duplicate accept

A node receives the same frame twice if both redundant LANs are fault-free.

There is in theory no need to discard duplicates at the link layer.

Any communication or application software must be able to deal with duplicates, since bridging networks (e.g. 802.1D RSTP) could generate duplicates:

- Most applications work on top of TCP, which was designed to discard duplicates
- Applications on top of UDP or Layer 2 protocols (publisher/subscriber) must be able to ignore duplicates because they rely on a connectionless communication.

PRP can operate without duplicate filtering (“Duplicate Accept” mode, used for testing).

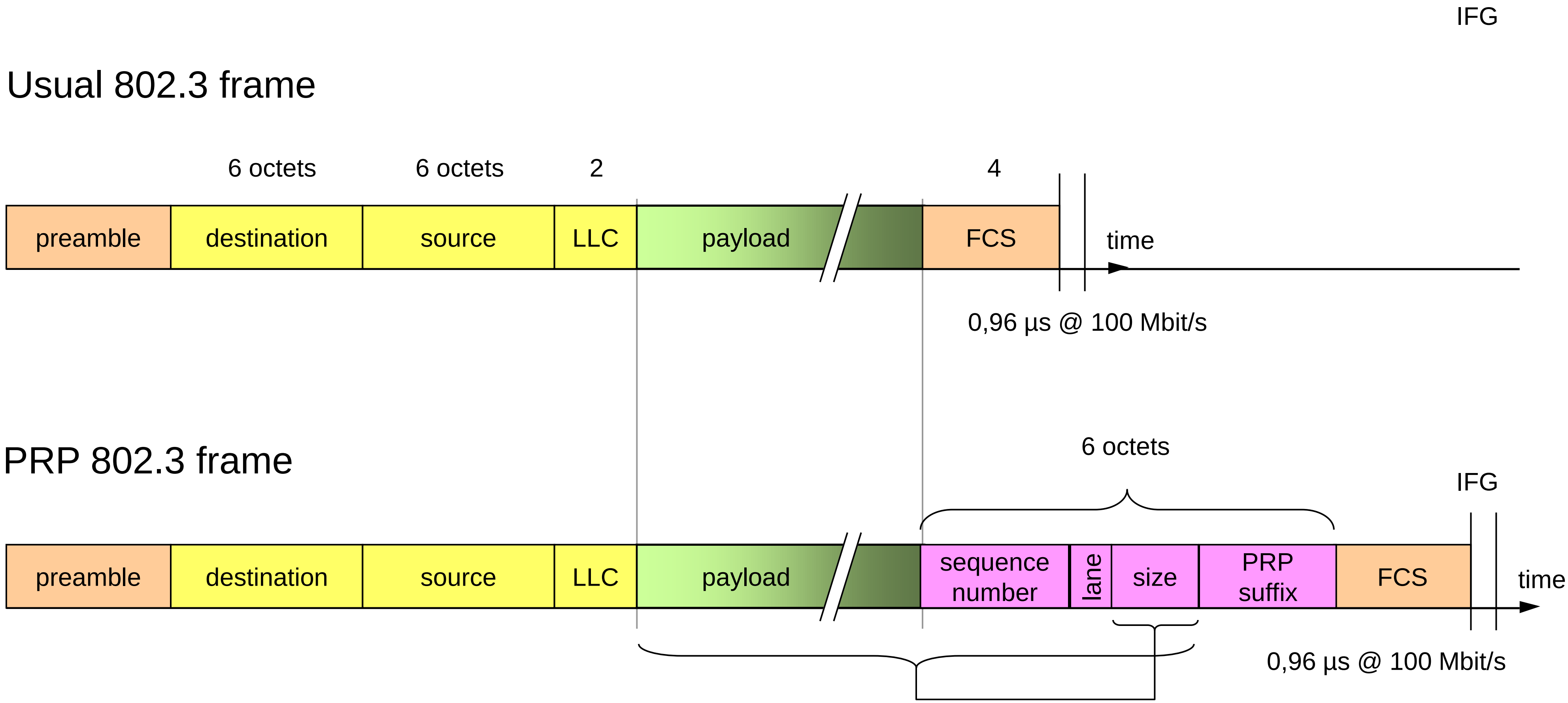
PRP uses a “Duplicate Discard” mode, helpful:

1) to offload the processors, especially when communication controllers are used.

2) to supervises redundancy, bus errors, partner nodes and topology. This turns out to be the major motivation.

For this, a unique identifier for the frames is needed.

Introducing transparently the PRP trailer



The extension by 6 octets is invisible to normal applications

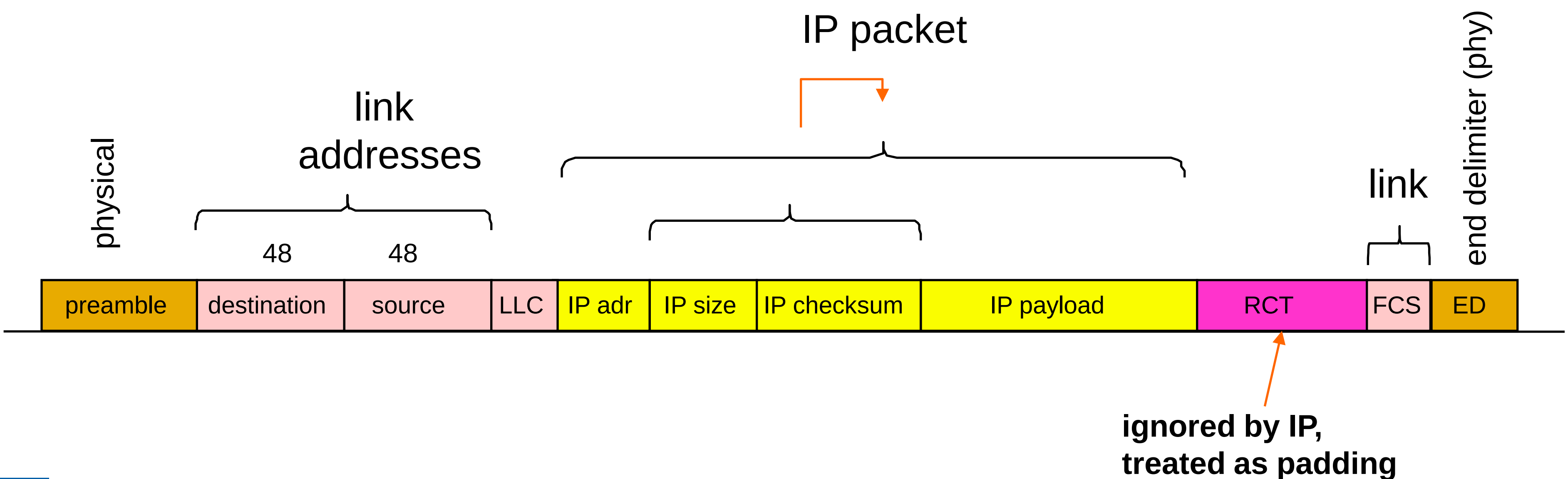
Why the PRP redundancy trailer works with SANs

Putting the redundancy control trailer after the payload allows SANs, i.e. nodes not aware of PRP (singly attached nodes such as laptops) to understand PRP frames.

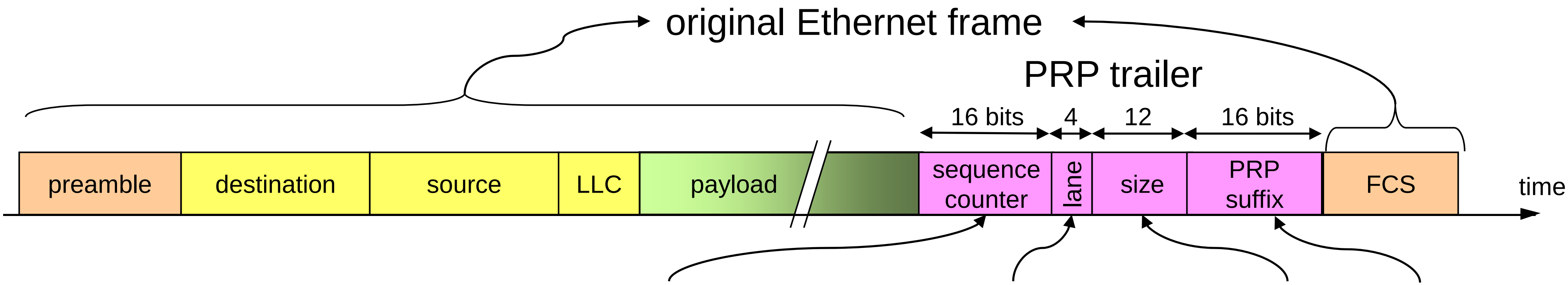
All singly attached nodes ignore octets between the payload and the frame check sequence since they consider it as padding.

To this effect, all (well-designed) protocols built on layer 2 have a size control field and a separate checksum.

Example IP packet:



Discarding duplicate frames with a sequence counter



- each frame is extended by a sequence counter, a lane indicator, a size field and a suffix * inserted after the payload to remain invisible to normal traffic.

- the sender inserts the same sequence counter into both frames of a pair, and increments it by one for each frame sent under the PRP protocol.

- the receiver keeps track of the sequence counter for each for each source MAC address it receives frames from. Frames with the same source and counter value coming from different lanes are ignored.

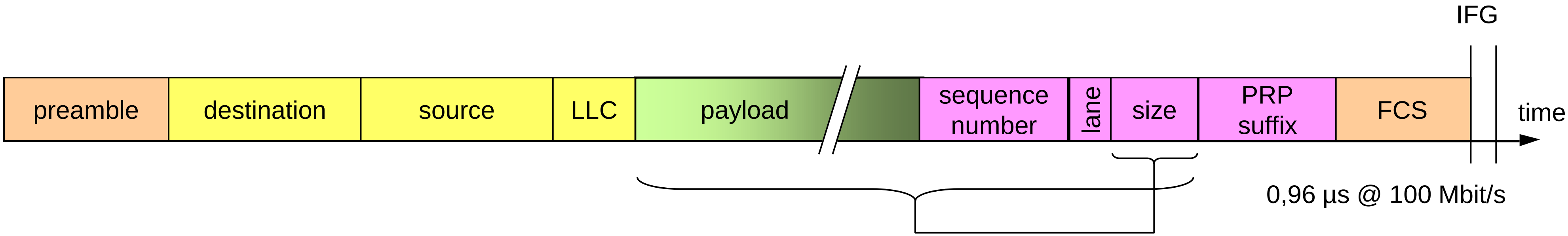
- for supervision, each node keeps a table of all other nodes in the network, based on observation of the network. This allows to detect nodes absence and bus errors at the same time.

* not present in PRP-0

Distinguishing redundant from non-redundant frames

The receiver detects if a frame is sent by a doubly attached node running PRP or is sent by a singly attached node with no redundancy trailer.

To this purpose, a receiver identifies the PRP frames using suffix and the size field. If the suffix is present and the last 12 bits of the frame match the size of the payload, the receiver assumes that this frame is a PRP frame and a candidate for discarding.



In a frame sent by a singly attached node, assuming random data in the payload, there is a probability of 2^{-28} that the size field matches the payload.
However, a frame will not be discarded unless a second frame arrives from the other lane from that same source and sequence number and with the correct line identifier, which is not possible since a singly attached node is attached to one lane only (half network).

NOTE: why not the size of the whole frame? because the size varies when VLAN tags are attached.

Sequence number build-up

A PRP node increments the sequence number for each frame (pair) sent.

A PRP receiver uniquely identifies a frame by {MACaddress, SequenceNumber}

Due to bridge filtering, a receiver only receives part of the traffic (especially if one part of the traffic is multicast) and could see gaps in the received sequence numbers from the same source.

It is also possible that the same frame comes more than twice in case of reconfiguration of RSTP.

To this effect, a receiver keeps a list of received frame identifiers and sequence numbers (as a table, hash table, or numbering scheme).

NOTE: the former version PRP0 had a sequence number per {sourceMAC, destinationMAC} tuple.

Principle of duplicate detection

IEC 62439-3 does not specify the method used to discard duplicates since it depends on the implementation (e.g. FPGA or software), but it specifies rules:

Any scheme that detects duplicates must fulfill the requirements:

1) never discard a legitimate frame as a duplicate
2) discard nearly all duplicates (some drop-out are unavoidable)

An entry in the list must be purged before a different frame with the same MAC and SequenceNumber can come again.

An entry in the list must be purged before a node that reboots can start sending frames with any sequence number. To this purpose, a node shall not reboot faster than the time it takes to flush all entries from the table of all its destinations.

An entry in the sequence number list resides at least $T_{resideMin}$ and at most $T_{resideMax}$ (values are implementation-dependent)

One cannot rely on the second occurrence of a pair {identifier, sequence} to retire an entry from the list since the second occurrence may never come.

Duplicate detection: wrap-around of the sequence number

The duplicate discard bases on sequence numbers to uniquely identify a frame.

The sequence number has a fixed size (i. e. 16 bits) so it wraps around after a number of frames (65535).

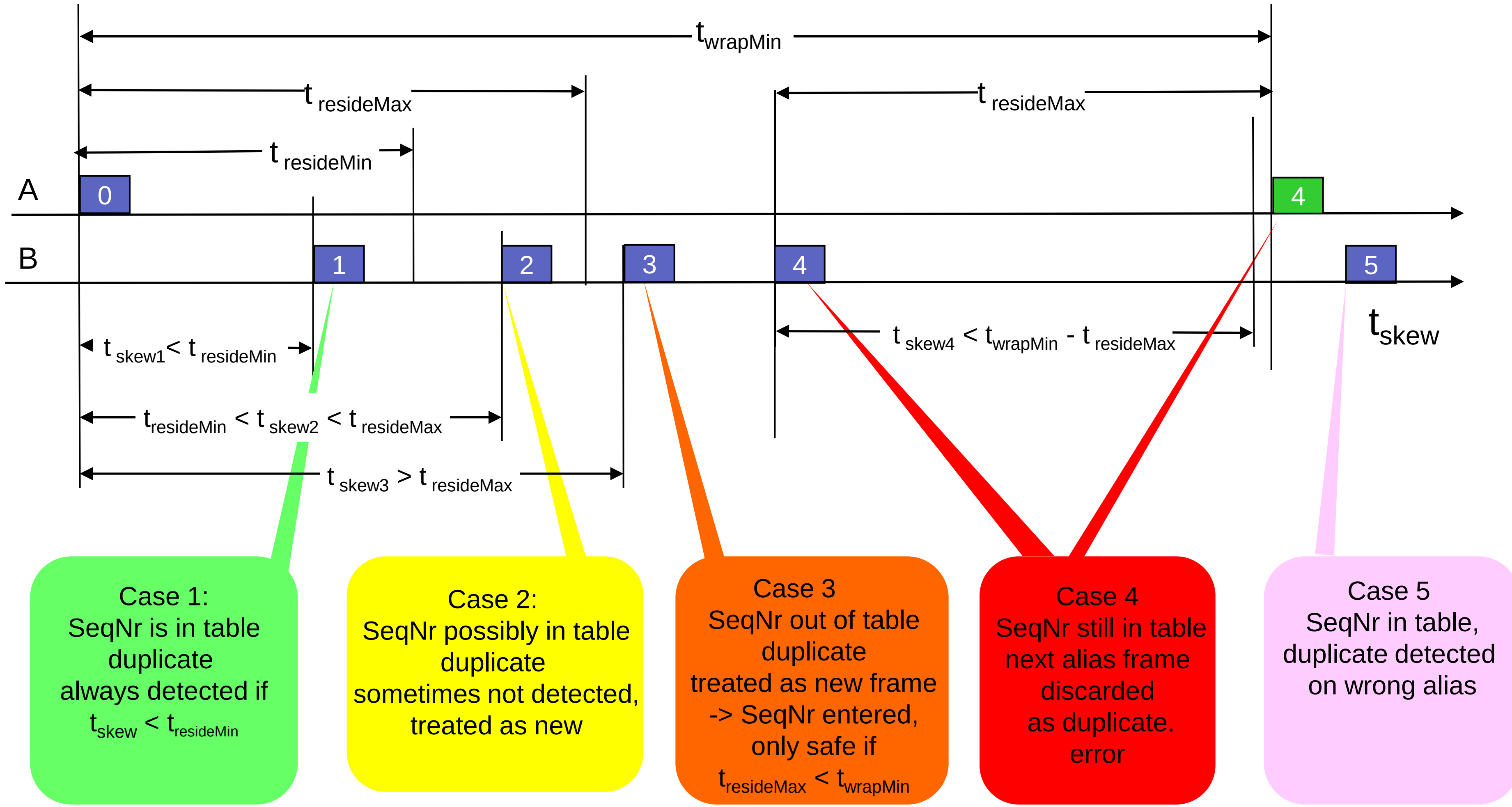
The minimum wrap-around time $t_{wrapMin}$ happens when a node sends a series of consecutive minimal length frames (an unrealistic, worst-case)

The minimum wrap around time can calculated:
@ 100 Mbit/s : ((8 + 64 + 6 + 12) * 8) × 65536 = 7.2 µs × 65536 = 472 ms
@ 1 Gbit/s: ((8 + 64 + 6 +12) * 8) × 65536 = 0,72 µs × 65536 = 47 ms.

This means that a legitimate frame is detected as duplicate when the second frame of a pair is delayed so that it arrives within that time of the legitimate frame (see next slide).

However, since a practical implementation is limited by the memory size, especially in FPGAs, it is likely that the memory will roll-over before the maximum wrap around time is reached.

Duplicate Detection in function of skew



Rule for reliable discard: $t_{skew} < t_{resideMin}$

Rule for probable discard: $t_{skew} < t_{resideMax}$

Rule for safe accept: $t_{residenceMax} < t_{aliasRepMin}$ ideally: $t_{residenceMin} = t_{residenceMax} = t_{wrapMin} /2$

Proposal: $t_{residenceMax}$ = 500 ms

Discussion

The time $t_{wrapMin}$ is the shortest time it takes for the same sequence number to appear again, so a legitimate frame could be rejected after this time because of an alias with a former frame.

Since the shortest frame in 100 Mbit/s lasts 6,72 µs and the sequence number wraps at 65’536, the value of $t_{wrapMin}$ is 440 ms (At 1 Gbit/s, it is 44 ms). This case is extreme, since no device should send a stream of contiguous short frames.

This means that a new frame could be rejected as “duplicate” after this time.

Entries must be aged out below this time, i.e. an entry should not remain longer than 440ms at 100 Mbit/s (resp. 44 ms in 1 Gbit/s) *for the same MAC source address*
Therefore, a frame entry shall not reside longer than this time in the duplicate table,
The condition is $T_{resideMax} < {}_{TwrapMin}$.

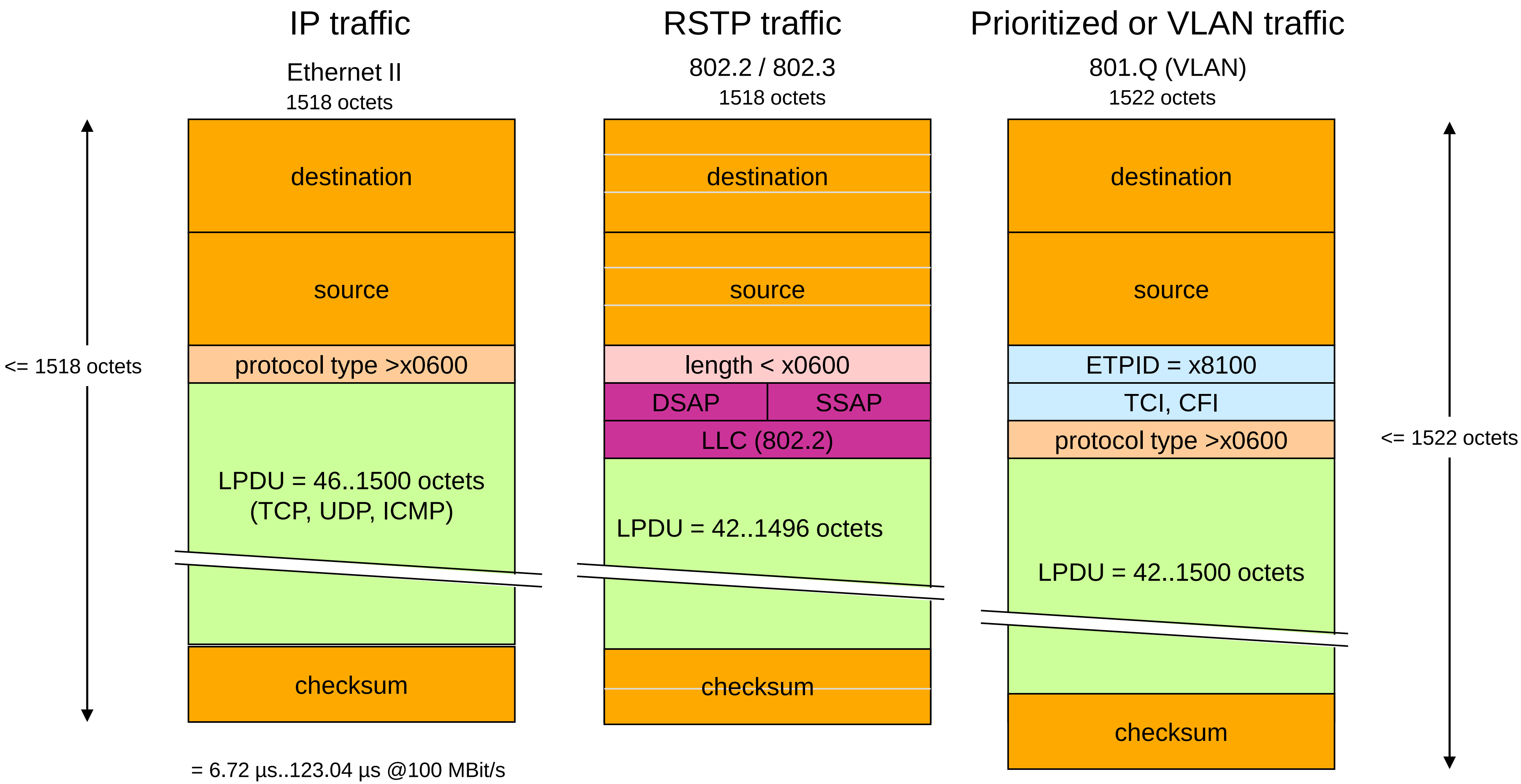
This value (EntryForgetTime) is independent from the network speed and set to 440 ms.

However, if a node reboots, it start sending frames with a random sequence number.
For that reason, a node that reboots is constrained not to reboot before EntryForgetTime.

The quiet time (NodeRebootInterval) was set to 500 ms for all speeds

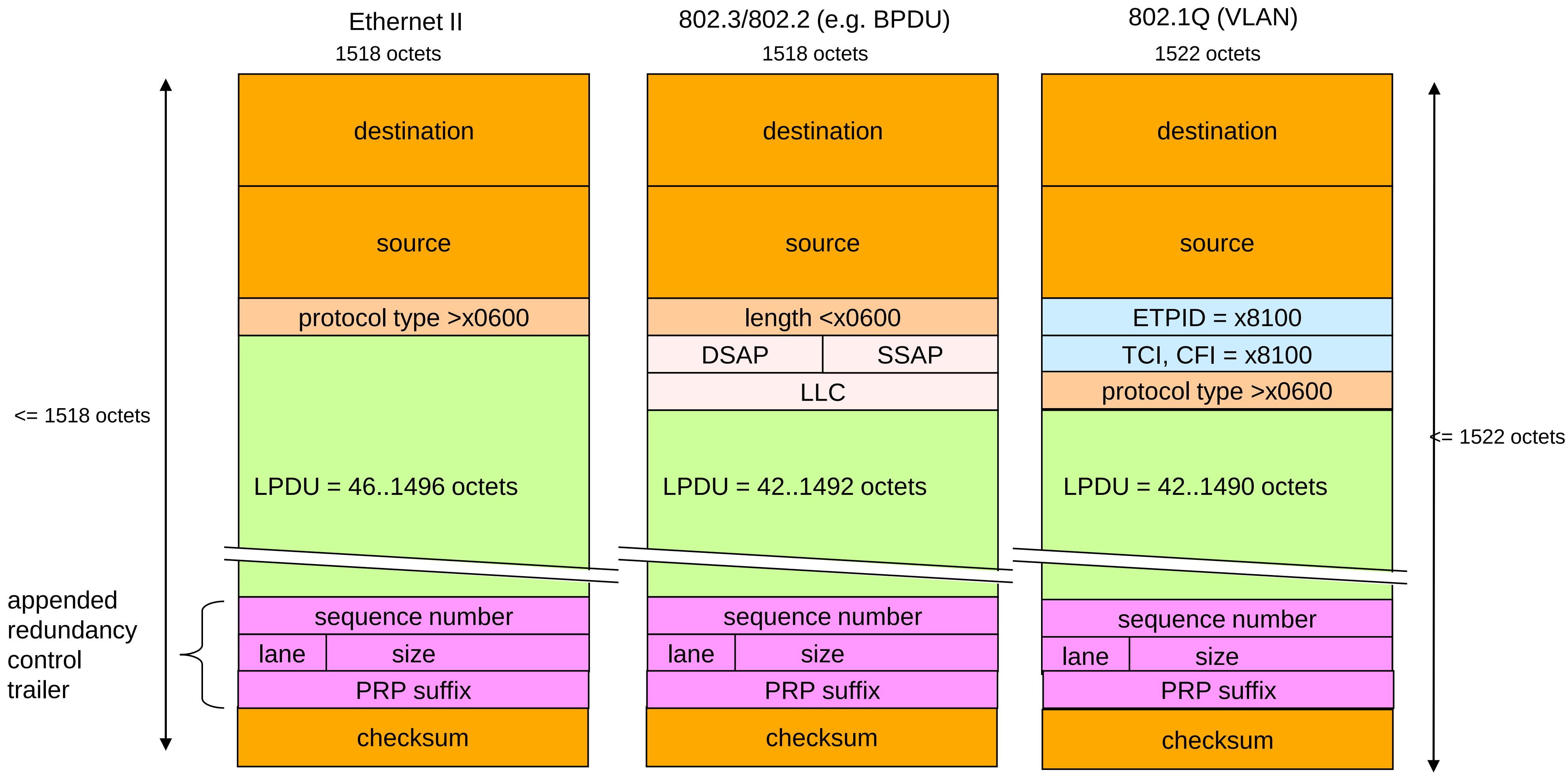
PRP Frame formats

Frame formats used in (all) Industrial Ethernets (SAN)



Most Industrial Ethernet allow a mix of 802.2, 802.3, etc…resulting in 3 different frame types...

PRP frames: redundancy control after the payload (DANP)



The additional six octets could generate oversize frames of more than 1522 octets.
Although these frames are accepted by all certified bridges (up to 1535 octets), some older bus controllers do not accept oversize frames, and therefore the sender should reduce the payload by playing on the LSDU size in the network layer (if the frame is going to be tagged).

Is appending a redundancy trailer in line with IEEE 802.3 ?

PRP’s Duplicate Discard mode requires that frames be stretched by 6 octets.

All protocols must be able to deal with data in the frames after the payload, since padding is part of the Ethernet protocol.

E.g. IP has its own size control and checksum – data after the IP payload are ignored.

The maximum size of the payload must be reduced by six octets to match the size of 1500 octets foreseen by IEEE 802.3-2002 § 4.4.2.1

This would not be necessary for the bridges, since all commercial bridges permit longer frames to support double VLAN tagging (Q-in-Q).

However, some older Network Interfaces for PCs do not accept longer frames.

Network analyzers can deal with PRP, the corresponding PRP extension already has been appended to EtherReal (http://www.ethereal.com) / Wireshark

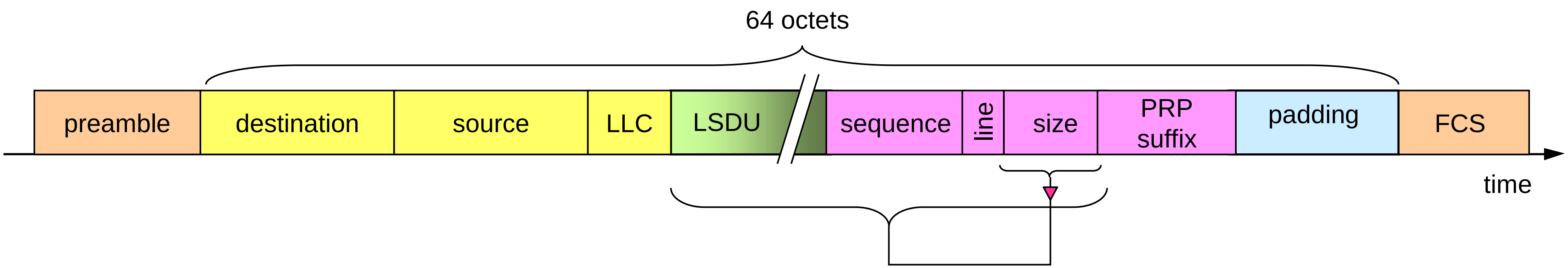
In networks where this could become a problem, the source should reduce of the LSDU size, e.g. at the IP level.

Redundancy Control Trailer (RCT) and padding

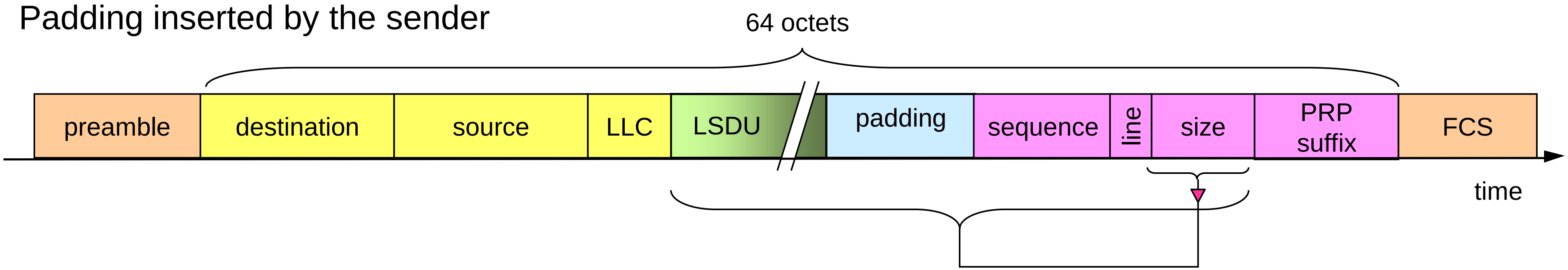
Small frames use padding to meet the minimum frame size of 64 octets.

Since padding can be introduced by bridges, the sender should always insert the padding itself to reduce the decoding burden (otherwise, the receiver has to search backwards for a matching field)

Padding introduced automatically: do not use

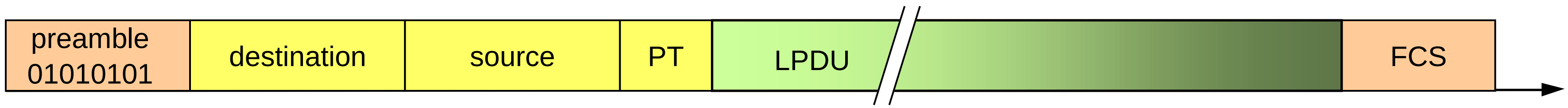


Padding inserted by the sender

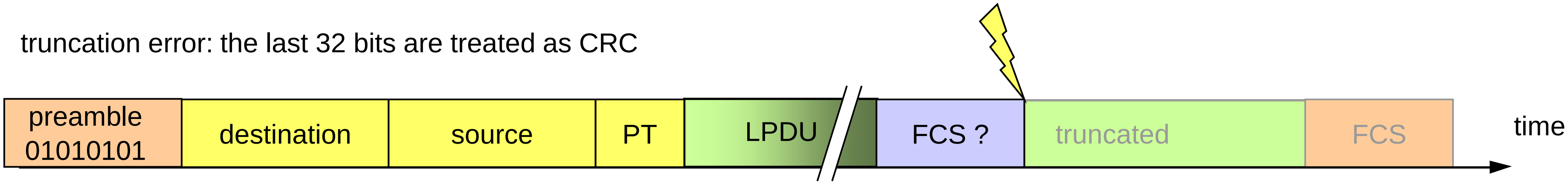


Why a size field is also useful

Ethernet frames have, contrary to IEC 8802.3 frames, only a Hamming Distance of 1 against synchronization errors (frames truncation), since they do not contain a length field.



truncation error: the last 32 bits are treated as CRC

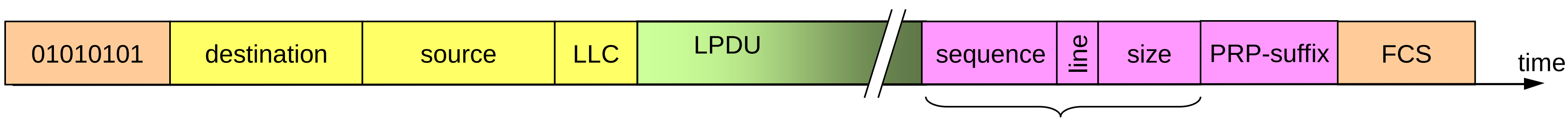


only the application (if it is aware of it) can detect the problem....

"We detected a mismatch between TCP and Ethernet checksum in up to one frame in 400"
(Siggcom 2000, J. Stone & C. Patridge @standford, bbn)

To improve protection of the frames, a size field is appended to the frames.

A node that receives a PRP frame from a known partner with incorrect size can flag an error.



PRP network management

Network supervision and detection of lurking faults

PRP checks continuously all paths.

In order not to rely on application cyclic data for this, each DANP sends periodically a supervision frame that indicates its state.

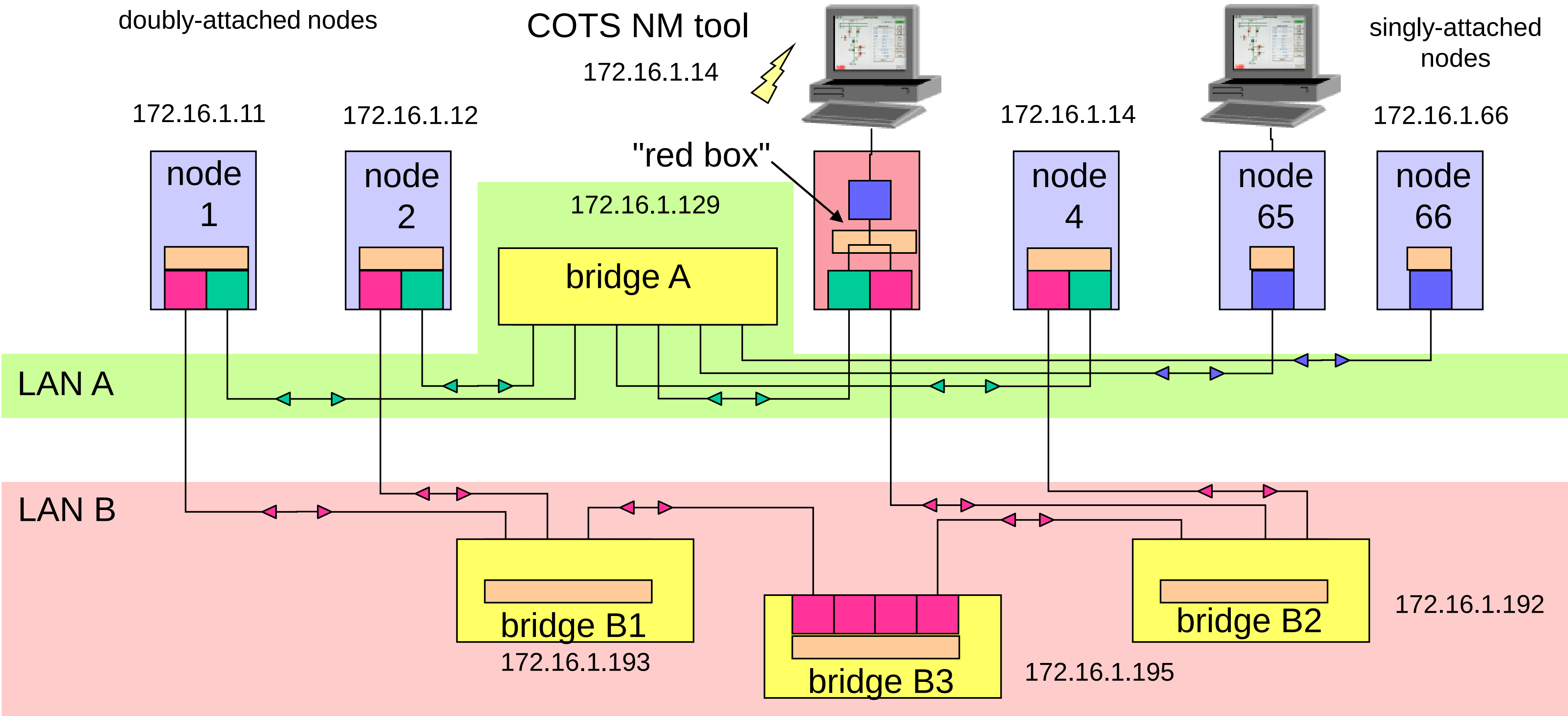
The period is relatively long (some seconds) since the supervision frame is not needed for switchover, but only to check dormant redundancy.

All nodes should keep a node table of all detected partners and registers the last time a node was seen as well as the number of received frames from each other node over both interfaces. Since this may increase the complexity, simple devices are exempted from keeping a node table.

The duplicate discard mode allows to keep track of all nodes in the network.

Changes to the topology are communicated over SNMP or to the Link Management Entity.

Network management



bridges are **single-attached devices** and have different IP addresses on each LAN, although their function can be the same as that of a corresponding bridge on the other network.
Each PRP node has an SNMP agent to keep track of redundancy.

PRP technology

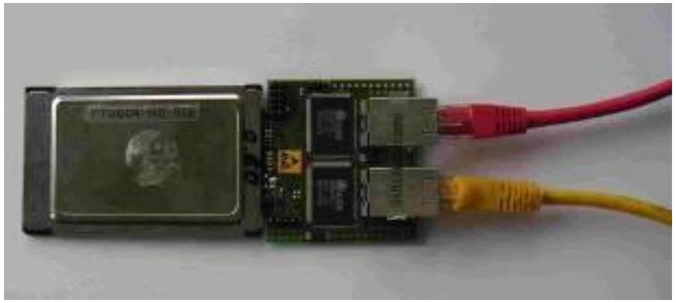
Maturity of technology

PRP is supported since 2007 by ABB and the Zurich University of Applied Sciences at Winterthur (ZHAW).

ZHAW built an independent Linux DANP using only the IEC document as specification.

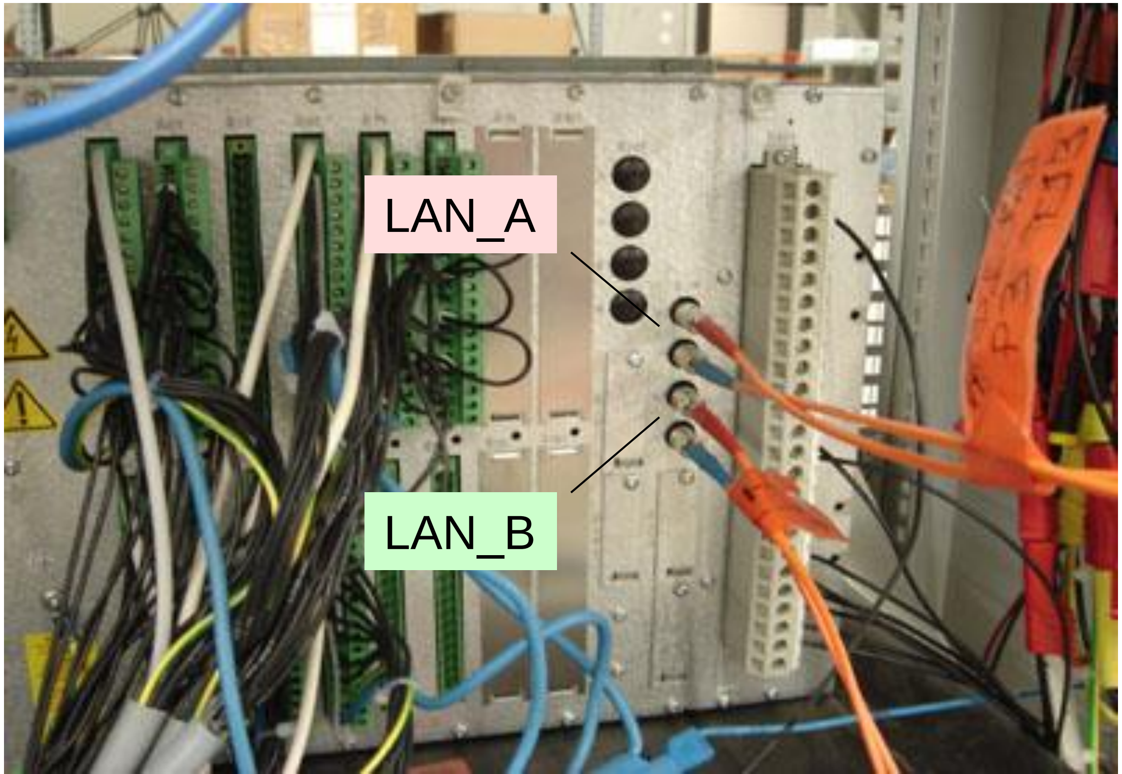
Interoperability between VxWorks, Linux and Windows implementations was tested in 2008.

PRP has been adopted by IEC 61850 as the redundancy principle for the Station Bus.



1st prototype (2005)
PRP is here handled by the main processor.
Co-processor or FPGA can execute PRP as well.

PRP in Factory Acceptance Test - picture courtesy ABB



PRP: Pros & Cons

+ seamless operation (zero switchover time)

+ application-independent, suits any Industrial Ethernet network.

+ uses standard bridges and protocols unmodified (ARP, DHCP, TCP/IP…)

+ allows attachment of nodes with any single port node to the network (with no redundancy)

+ does not breach the fail-independence of the redundant networks

+ supervises constantly the redundancy in “duplicate discard” mode (both LANs are active)

+ monitors actual topography (over network management / SNMP)

+ compatible with IEEE 1588 – a redundant clock profile is defined

- requires complete doubling of the network, two Ethernet controllers and a special driver per node (about twice the costs – but this is the price to pay for any full redundancy)

- restricted to a layer 2 broadcast domain (not a limitation in Industrial Ethernet)

- requests that singly attached nodes that need to communicate with each other are connected all to the same LAN (or through a “red box”)

- cost six bytes overhead in a frame – but oversize frames should become IEEE standard.

PRP extensions

The principle of PRP has been applied to IP networks.

In this case, the sequence number already exists since IPv4 supports segmentation (identification field).

In IPv6, a sequence number is also available in the flow label.

However, there are no applications in utility automation that make use of this capabilites.

One reason is that the expected availability increase has to be matched against the costs of defining disjoint paths in a network.

PRP vs. other standards

State of standardization and other solutions in IEC 62439

The “Highly Available Automation Networks” IEC SC65C WG15 selected PRP and HSR as one of its redundancy methods, along with:

MRP (Siemens-Hirschmann) implements “redundancy in the network” with singly attached devices attached to a ring, with moderate increase in availability and disruption delay of 200 ms-500 ms. It is interesting if the bridges are integrated in the devices, but this limits topology to a simple ring of up to 50 bridges.

CRP (Honeywell/Fieldbus Foundation) implements – like PRP – “redundancy in the devices”, offers the same availability as PRP, but has disruption times of 200 ms – 2s. It allows to connect singly attached devices to both network halves, but costs aggregated links in the (mandatory) root bridges.

CRP (SupCon, China) is a ring redundancy protocol which competes with MRP and uses a tight clock synchronization to support time-slotted real-time traffic.

Only PRP and HSR provide zero recovery time.

INTERNATIONAL
ELECTROTECHNICAL
COMMISSION