“Distinguished minimum latency traffic in a converged traffic environment” DMLT

Ludwig Winkel
Siemens AG

IEEE 802.3 Ethernet Working Group

IEEE 802.3 Tutorial, SG DMLT
July, 15, 2013
Geneva, CH

Agenda

- Welcome, Introduction (Ludwig)
- Market needs and Market potential (Dan)
- Architectural options (Norm)
- Technical and Economic feasibility (Pat)
- Conclusion (Ludwig)
- Discussion (Ludwig)

Reflector and Web

- CFI information posted at page <http://www.ieee802.org/3/cfi/request_1112_1.html> with a link to the presentation - the presentation itself can be found at the link <http://www.ieee802.org/3/cfi/1112_1/CFI_01_1112.pdf>.
- Study Group reflector

stds-802-3-DMLT@listserv.ieee.org

To subscribe to the DMLT-reflector, send an email to: ListServ@ieee.org with the following in the body of the message (do not include “<>”):
subscribe stds-802-3-DMLT <yourfirstname> <yourlastname>

- Study Group web page URL:

http://www.ieee802.org/3/DMLT/

Draft PAR (P802.3br) title & scope

- SG DMLT proposes a PAR title:

IEEE Standard for Ethernet

Amendment Specification and Management Parameters for

Interspersing Express Traffic.

- Scope:
  - The scope of this project is to specify additions to and appropriate modifications of IEEE Std 802.3 to add a support for interspersed express traffic.

ABSTRACT

There is a need for support of time sensitive traffic in a converged traffic environment in IEEE 802.3 networks that supports interspersed express traffic and the traditional normal traffic. This would help address the requirements in markets such as industrial and automotive control networking, where control data is time-sensitive and often requires minimum latency. This tutorial will examine the needs of time sensitive traffic in IEEE 802.3 networks, the support for interspersed express traffic and the ordinary traffic, and will provide background for the PAR proposed by the IEEE 802.3 Distinguished Minimum Latency Traffic (DMLT) Study Group.

Presenters/Authors

Name:	**Presentation:**	**Employer/Affiliation** (if different)**:**
Ludwig Winkel	Welcome, Introduction, Conclusions	Siemens
Dan Sexton	Market needs and Market potential	GE
Norm Finn	Architectural options	Cisco
Pat Thaler	Technical and Economic feasibility	Broadcom
Additional authors:		
Albert Tretter		Siemens
Christian Boiger		Self
Razvan Mihalache		Bosch

Market needs and market potential

Dan Sexton	GE

Potential Markets Served by DMLT

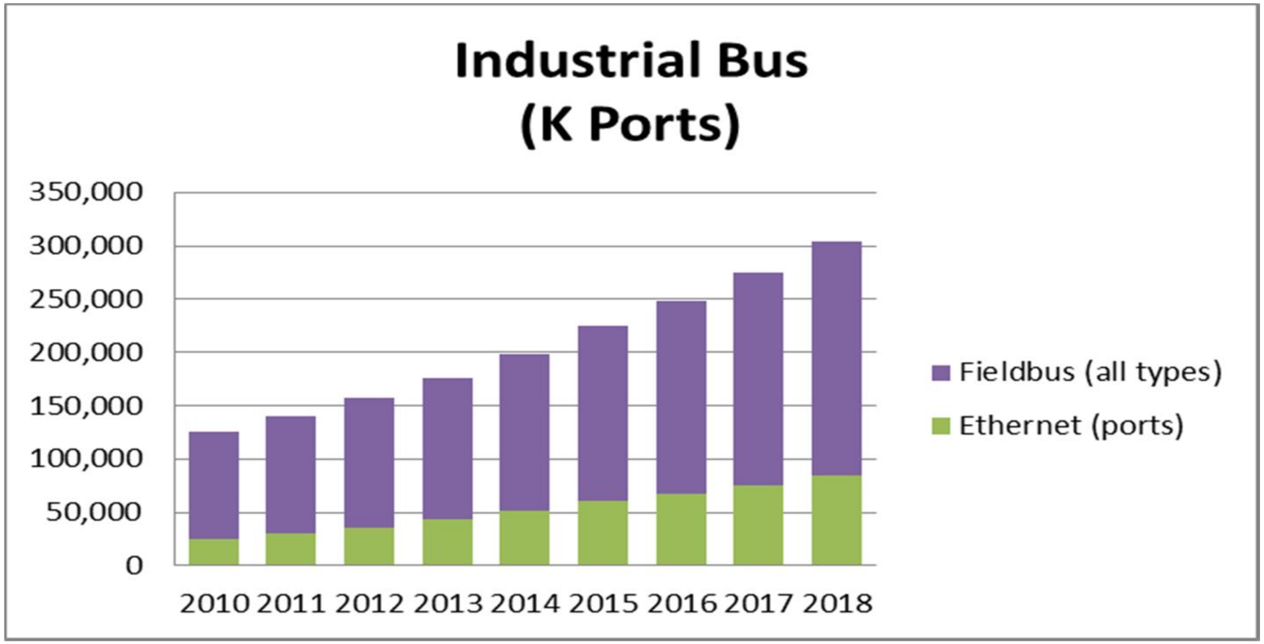
Industrial and Commercial Market

Introduction

- Ethernet use in industrial and commercial market is growing.
- About a dozen proprietary protocols currently serve the networking needs

Forecast

- Strong desire and need for converged traffic networking.
- Expect both conversion from field bus and growth of Ethernet to converge over time.



Source: Contributions from Hirschmann, Siemens and Broadcom

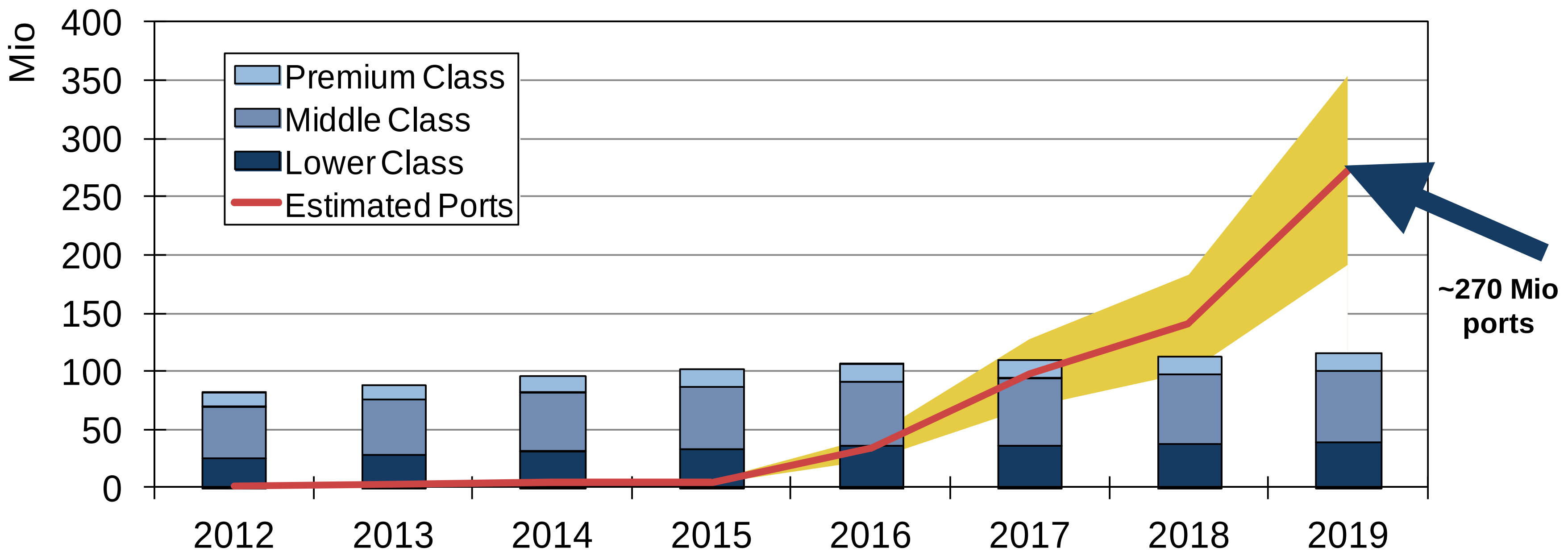
Automotive Ethernet Market

Introduction

- Ethernet use in automotive networks is now reality.
- Some mainstream in-car networks, e.g. CAN, Flexray, in use.

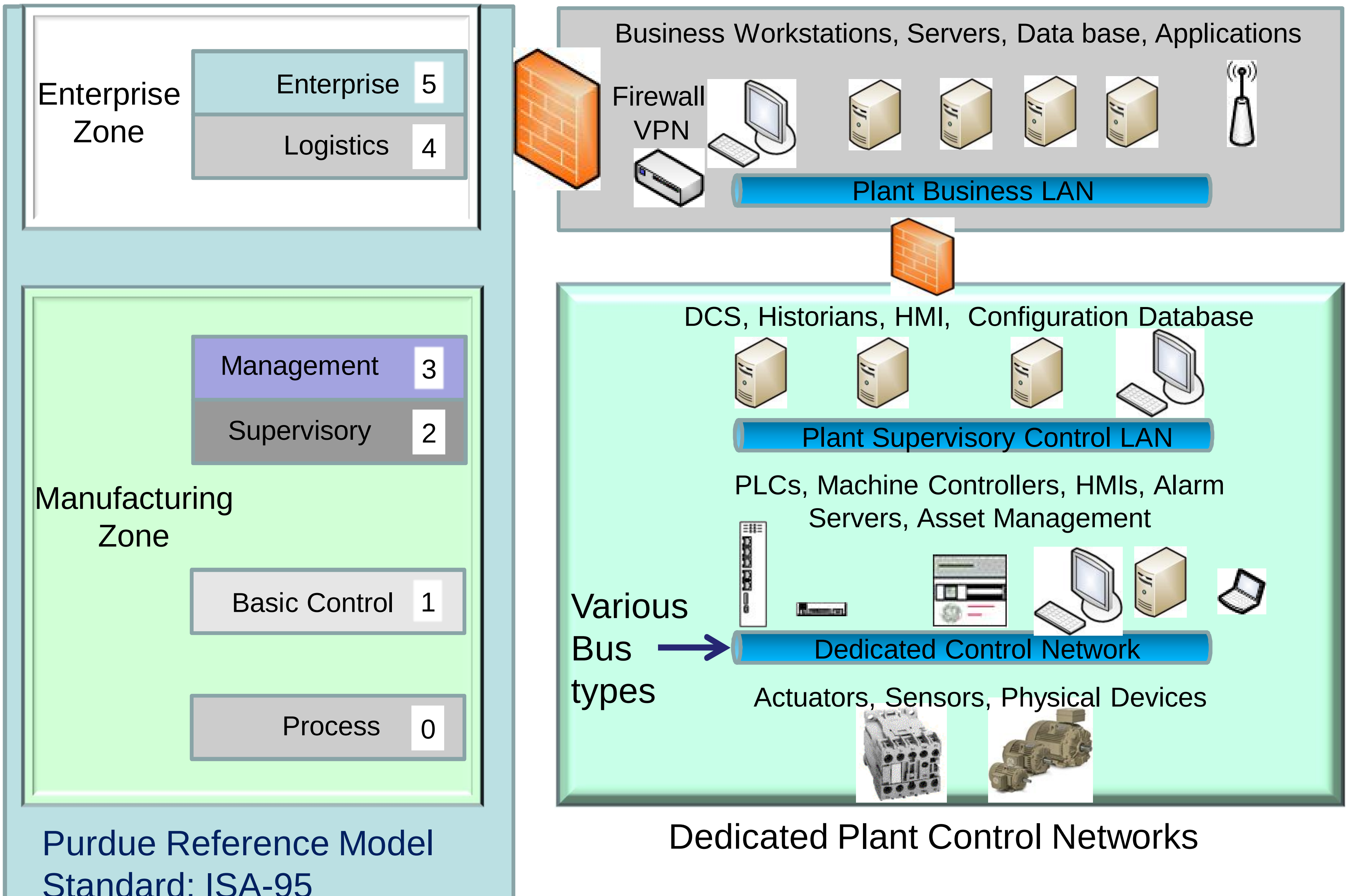
Forecast

- Strong desire and need for converged networking.
- Strong desire to interconnect mainstream in- car networks and emerging Ethernet networks.

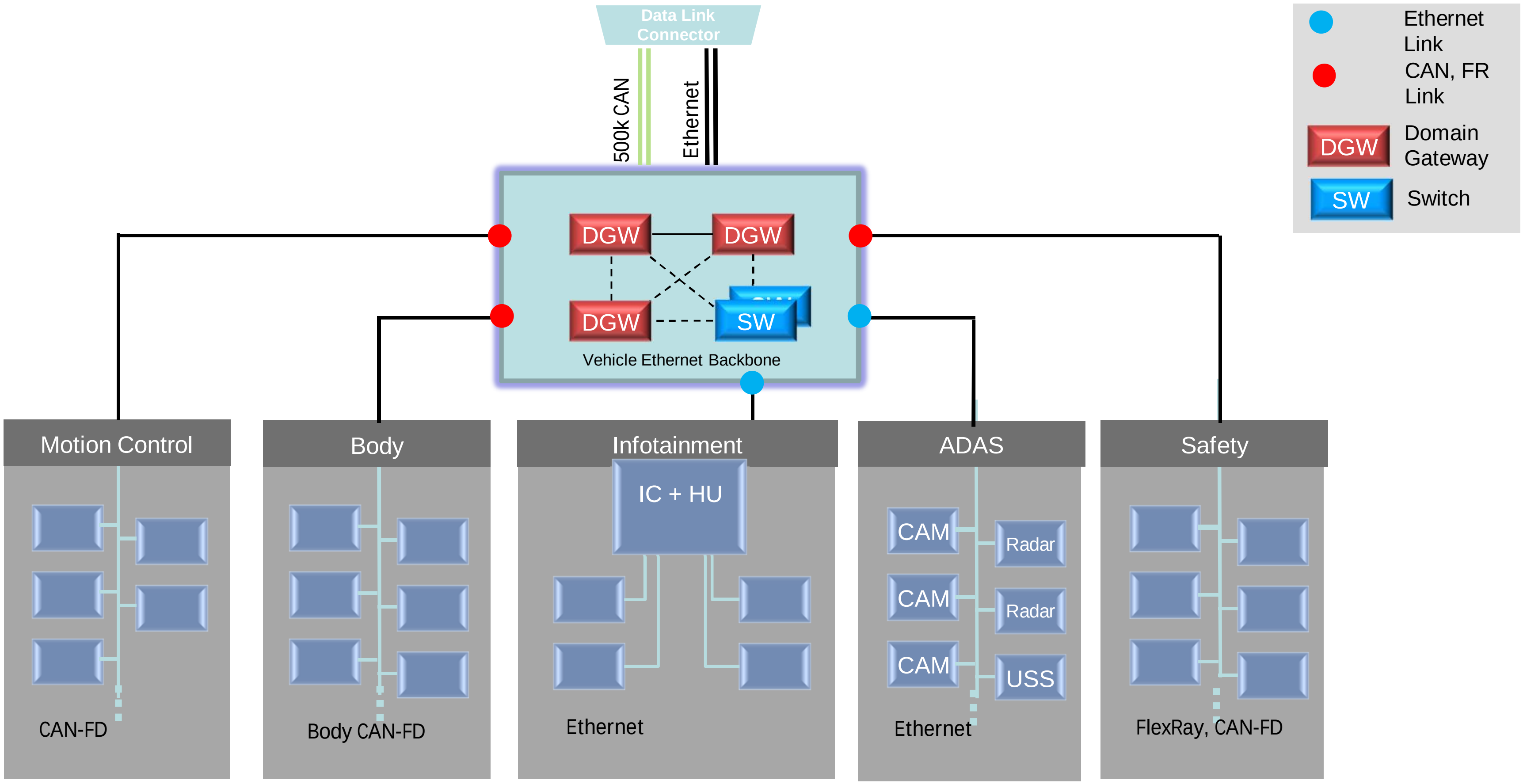


Networked equipment in a power plant

Typical Industrial Network Configurations

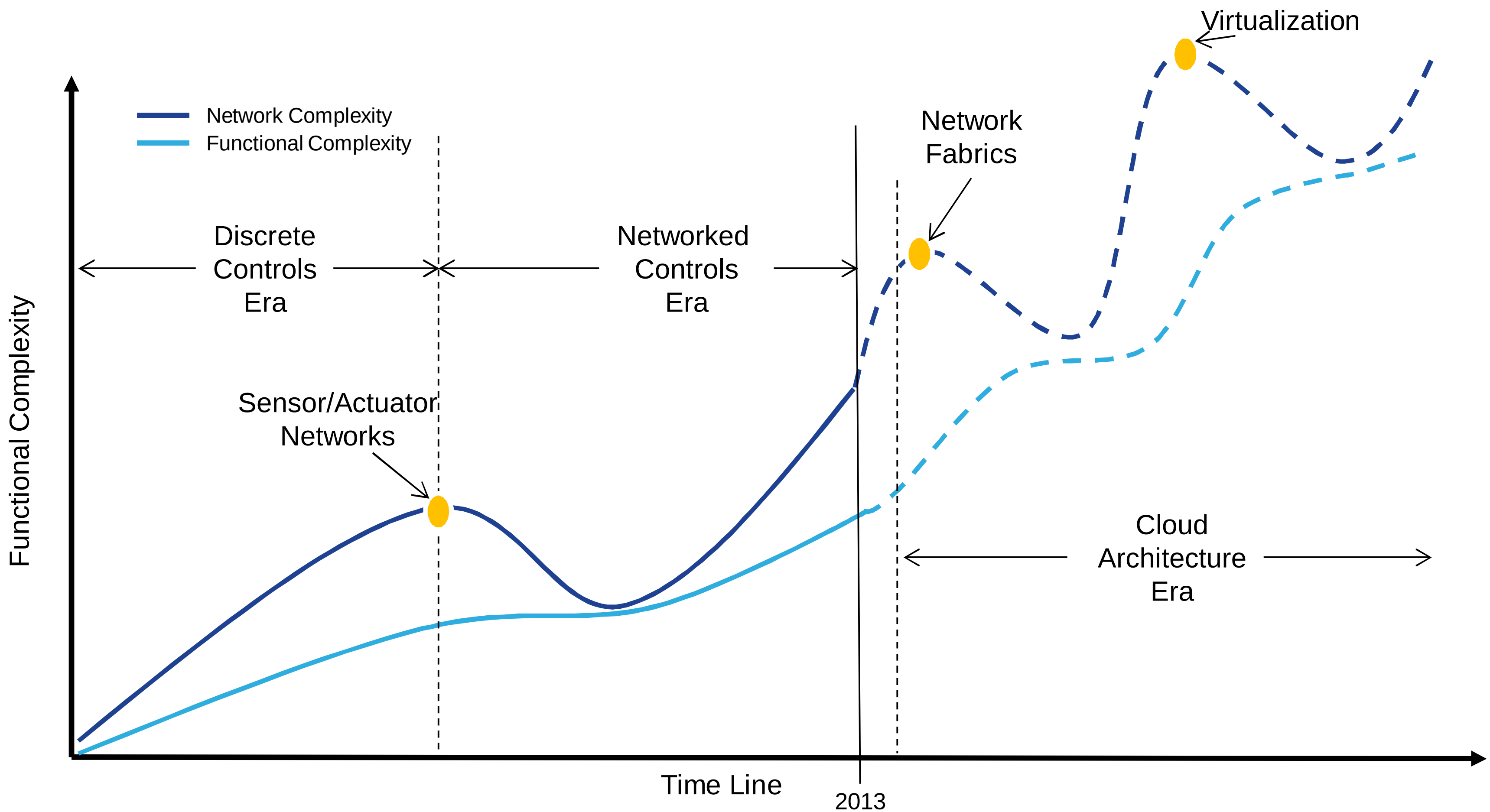


Automotive Network > 2018



Ethernet Backbone with Mixture of Control and Multi-media Traffic

Control Function and Network Complexity Progression



Control Systems in all market sectors perpetually increase in functional complexity.
Communications complexity limits functional capability.
Advanced communications architectures enable advances in controls.

Industrial and Automotive Needs

Performance "Guarantees"

There are no guarantees - the more "9s" the better
Low Latency – on time delivery, small mean
Low Jitter (Low Latency Variation, small σ)

Reliability/Availability:

There are no guarantees - the more "9s" the better
Redundancy/Availability, low MTTR (mean time to repair)
Accuracy
Security

A variety of Network Topologies

Star, Tree, Daisy Chain, Ring, Mesh.
Multiple hops – deep networks > 7 hops

Harsh Environment:

Operating temperature -40C to 105C no fans.
High Vibration & Shock
Noise
Water/Salt/Dust/Dirt/Snow/Ice – etc.
Stringent EMC requirements (Near HV switching)

Low life cycle cost/Low unit costs.

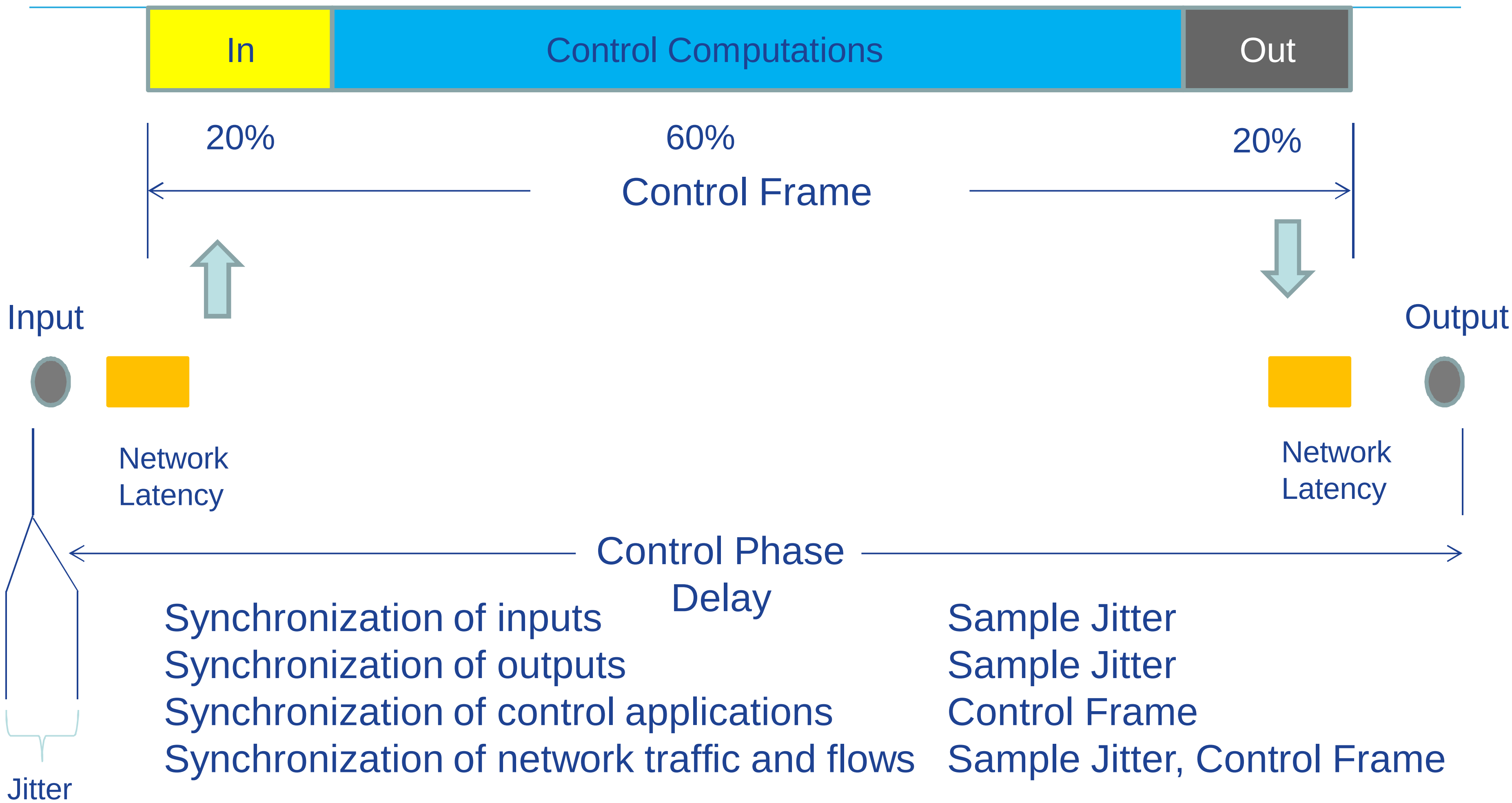
Cost pressures in this market can exceed commercial markets
UTP cabling

Application Protocols for Control

Motion	Industrial	Transportation	Distribution
Ethercat	Modbus TCP	ARINC 664	IEC 61850
SERCOS III	OPC	Flexray	IEC 60870
Powerlink	Profinet	AS6802 (TTE)	DNP 3.0
Profinet IRT	Foundation Fieldbus HSE	CAN	
SynqNet	Ethernet/IP	MOST	
IEEE 1588	NTP		

Note: There are many other proprietary protocols not on this list

The Basic Control Cycle



Overall Phase Delay Drives System Controllability/Stability

Why DMLT?

What it is:

A Technique to allow various types of application traffic to coexist.

What it isn't:

A new physical layer.
A replacement for existing Ethernet.

Simplifies the convergence of various applications into a single platform by guaranteeing each gets the quality of service needed

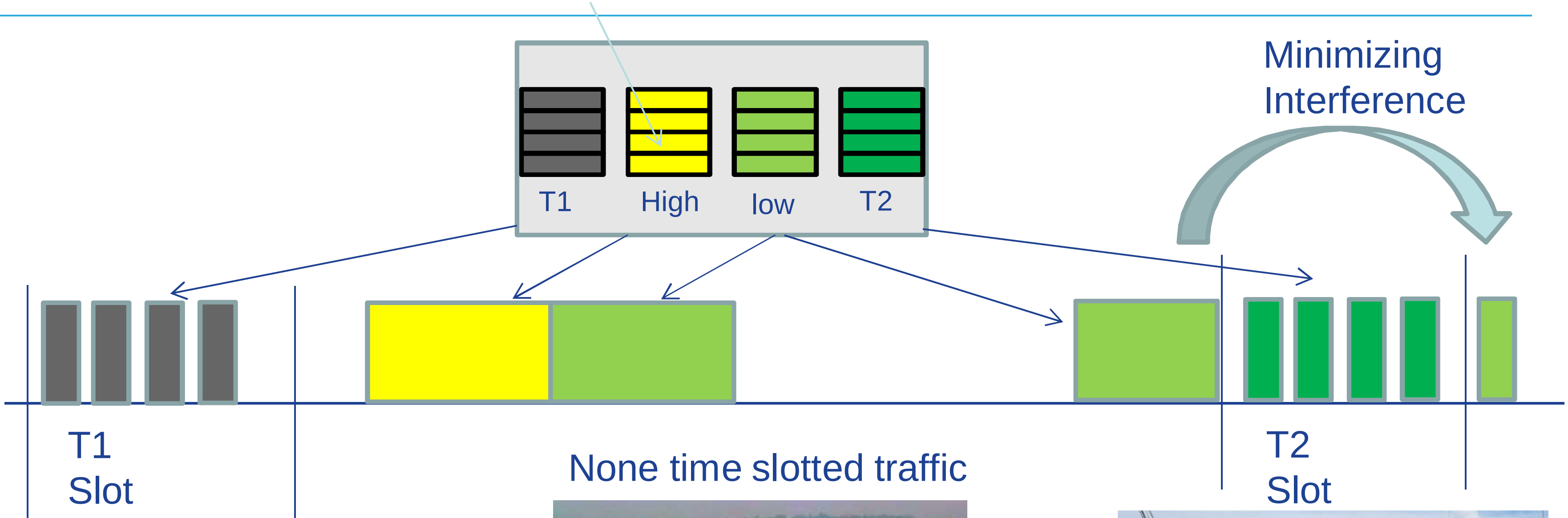
Why is it important to us?

1. We can support various application protocols over a single network (Ethernet IP, Profinet, OPC, IECC61850, FF-HSE + Audio/Video).
2. We can engineer networks to meet our control needs while also supporting non-control/deterministic applications.
3. Better clock synchronization.
4. Low development effort.
5. Better real time behavior for non-scheduled critical traffic.
6. Keeps us on the path with the technology providers – non-traditional Ethernet applications can participate in the future of Ethernet.

The argument for converged networks

- Security concerns – we can no longer rely on the perimeter/firewalls for security. Some fieldbuses have inadequate security. Multiple access databases are unmanageable.
- Simplicity – Maintenance, management, diagnostics & tools.
- Connectivity everywhere.
- Maintenance & Design Personnel training, simpler network design & effort.
- Flexibility, easier upgrades & enhancements
- Future proofing (faster upgrades and retrofits)
- Standards convergence, Open Systems

Why Converged Traffic Networks



Time Cyclic
Control Traffic

None Real time Traffic

- Logging
- Alerting

Real time – non cyclic Traffic

- Critical Alarms
- Discrete/event control

Traffic types in converged networks

- Business transactions/Internet Access
- Physical Security, Plant Maintenance
  - Video, Voice
- Business Communications
  - VOIP
- Asset management/Inventory tracking
  - Blob transfers, database access
- Control
  - Supervisory, Discrete, Process, Coordinated, Safety Critical
  - High speed, periodic & aperiodic, alarms and alerts
  - High reliability, availability

Why one single Network for all Communication Services

Only one network means:

- Reduced possibility of network failures
  - wire breaks, reduced confusion in case of maintenance
- Reduced installation costs
  - fewer cables and connectors, lower installed costs and faster startups
- Enables smaller devices
  - reduced space for connectors, lower power consumption (only half the number of PHYs needed)
- Reduced maintenance costs
  - easier to understand and to maintain, less personnel training
- Only one interface in the devices
  - only one MAC address, only one IP address, easier to understand and to maintain, easier coordination of the communication relations in the stack and application layer in the devices, more direct access to data.

Summary: Industrial Requirements for Interspersed Traffic

<u>Performance requirements for Interspersed Traffic:</u>

– **Minimum latency: < 3µsec max per hop accumulated latency (GE – min frame)**

– Guaranteed latency, low jitter

– Topology independent

– Typical data size (payload size): 40 - 300 bytes

– Range of transmission period: 31.25µs – 100ms and aperiodic

– Scheduled Traffic & Alarm has higher priority than Reserved Traffic and Best Effort Traffic

– Low cost, Low power, Low complexity

* These are our best estimates derived from multiple use cases of the current and future industrial applications.

Summary: Industrial Requirements for Interspersed Traffic

Preconditions for performance requirements

– Network topology: Star, Tree, Daisy Chain, Ring, Mesh

– Network attributes

- Maximum 64 hops +
- Maximum number of nodes (bridged end stations & end stations): up to 2000
- Yields as many as 64 hops and 2000 devices, perhaps more.
- Maximum cable length: Standard length for Tx and Fx are required
- High available network - seamless redundancy for critical Traffic

– Payload size of Reserved Traffic (e .g diagnostic data): ~400 bytes

– Payload size of Best Effort Traffic: 1500 bytes

* These are our best estimates derived from multiple use cases of the current and future industrial applications.

Main Benefits of DMLT

- Better network utilization for scheduled traffic (More capacity).
- Lower latency for High Priority, critical asynchronous (non-scheduled) traffic.
- Lower cost and power consumption (for equivalent performance).
- Better environmental characteristics.

Architectural options

Norm Finn	Cisco

Summary of this segment

- Current and forthcoming 802.1 Classes of Service offered by bridges, and why they are needed for Deterministic Networking.
- We need express / interspersed transmission.
  - Primary need is to support convergence of scheduled, rate limited, and best-effort traffic on the same network.
- Interspersion must be defined near the MAC layer by 802.3, and may be defined above the MAC layer by 802.1, as well.

802.1 Classes of Service: **today**

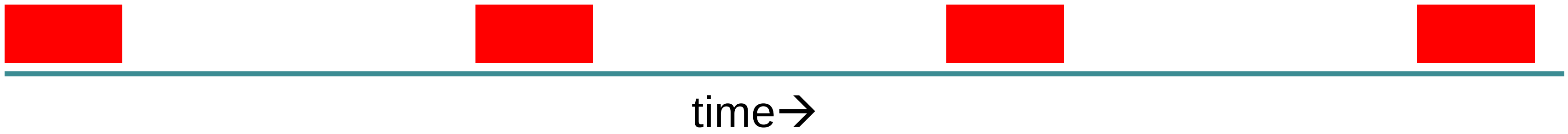
- **Traffic shaped** (AVB: Audio Video Bridging)
- **Weighted best-effort** (all ordinary traffic)
- **Traffic shaped queues are ahead of all other queues in terms of absolute priority**, because:
  1. The shaper guarantees the lower-priority queues a predictable latency.
  2. Highest priority gives shaped traffic predictable latency. Mostly.

802.1 Classes of Service: **new**

- **Scheduled mission-critical**
  - Every queue has a gate, admitting it to the transmission selector, that is controlled by a repeating schedule that can be synchronized with other bridges' schedules.
  - Presumably, those gates are used to:
    - Provide times during which only selected mission-critical queues are able to transmit.
    - Provide "guard bands" – times before scheduled mission-critical transmissions when no frame can be transmitted from any queue.
  - This supports data streams with jitter < 100 bit times.

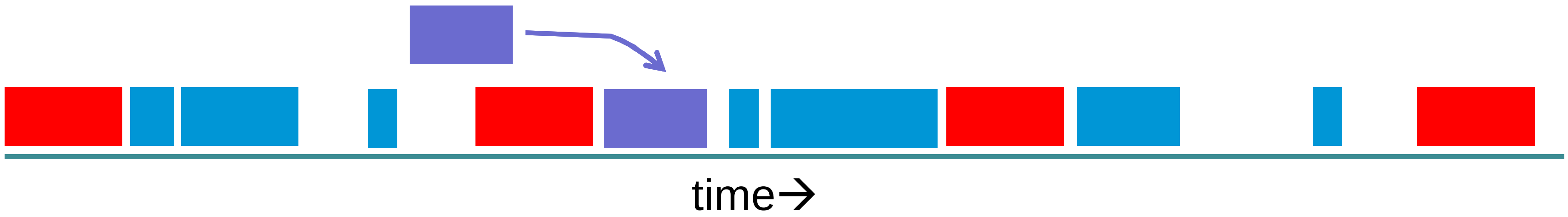
802.1 Classes of Service: **new**

- **Scheduled mission-critical**
- **Why?** Because schedule transmissions are the only way known to achieve 0 congestion loss in the presence of potentially interfering traffic.



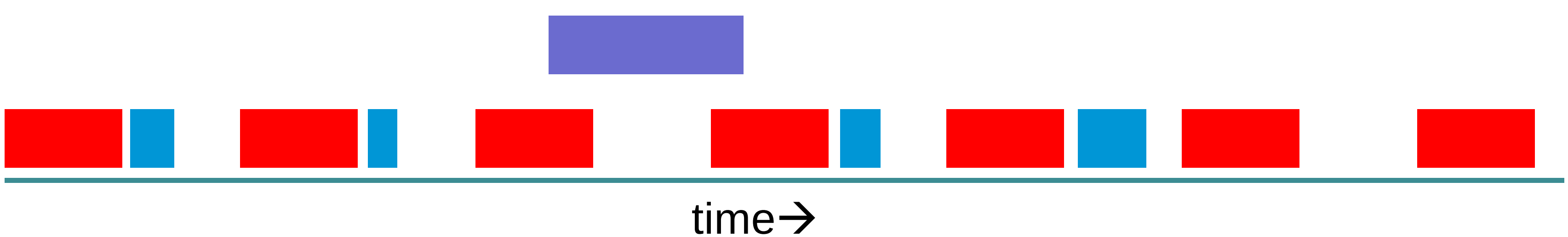
802.1 Classes of Service

- **Scheduled mission-critical**
- **All other traffic**
  - **Must sometimes wait**, in order to not delay a scheduled transmission



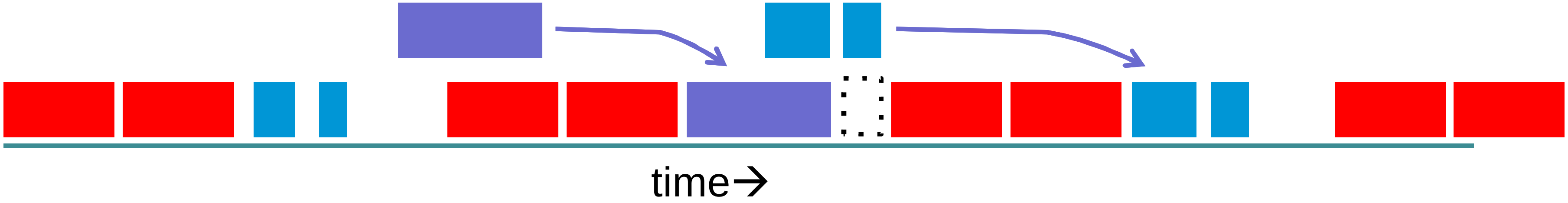
802.1 Classes of Service

- **Lots of** scheduled mission-critical
- **Some other traffic doesn't fit!**



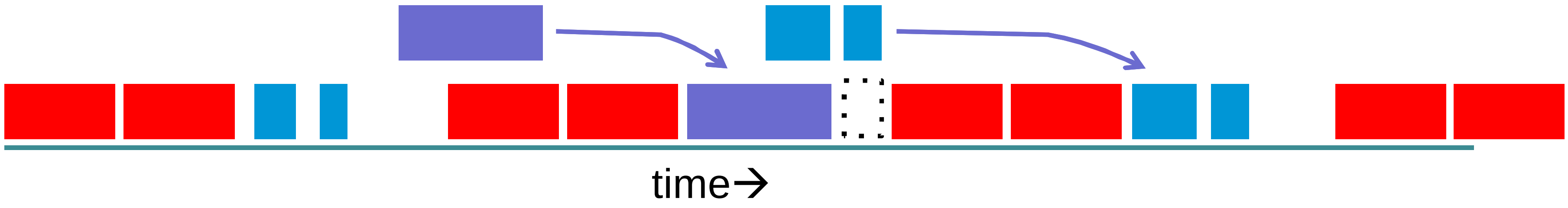
802.1 Classes of Service

- **Lots of** scheduled mission-critical
- **Some other traffic doesn’t fit!**
- Perhaps you rearrange the schedule, but now fitting is difficult – it’s a **packing problem**.



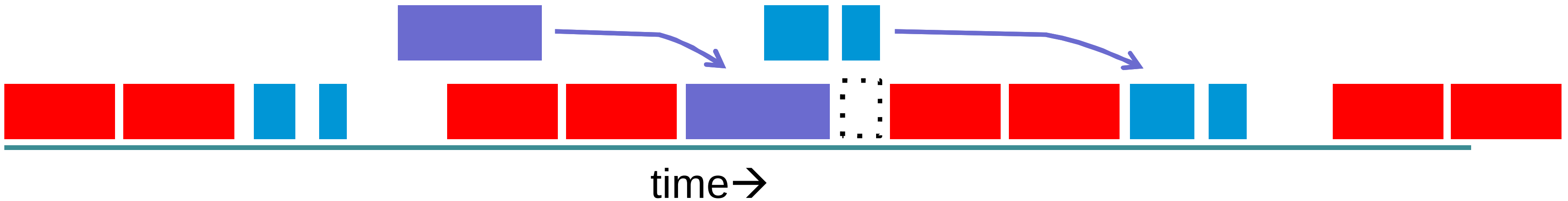
802.1 Classes of Service: **new**

- The critical question for converged traffic becomes:
- **Given a schedule, what is the bandwidth available best-effort and to the shaped (AVB) queues?**



802.1 Classes of Service: **new**

- The critical question for converged traffic becomes:
- Given a schedule, what is the bandwidth available to the best-effort and to the shaped (AVB) queues?
- **Without express/interspersed traffic, the answer is, “I don’t know”.**

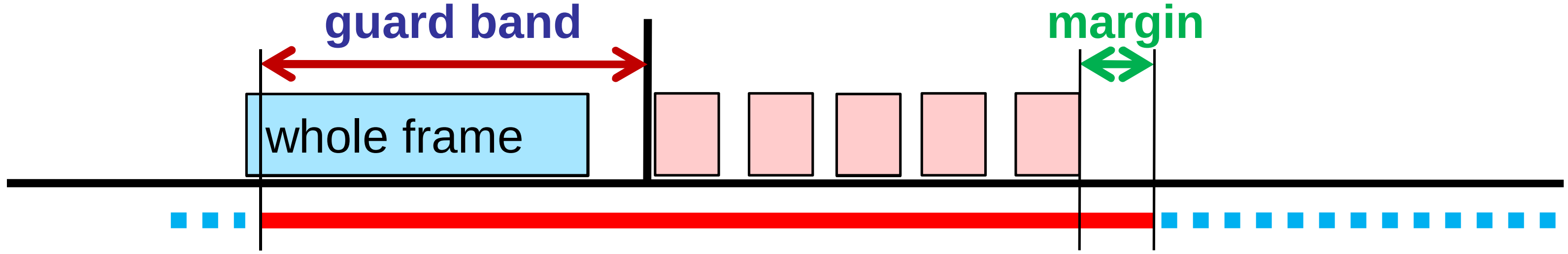


802.1 Classes of Service: **new**

- The critical question for converged traffic becomes:
- Given a schedule, what is the bandwidth available to best-effort / shaped queues?
- **With express/interspersed traffic, it is, "The bandwidth remaining after subtracting the schedule, minus a bit for worst-case interspersion overhead."**

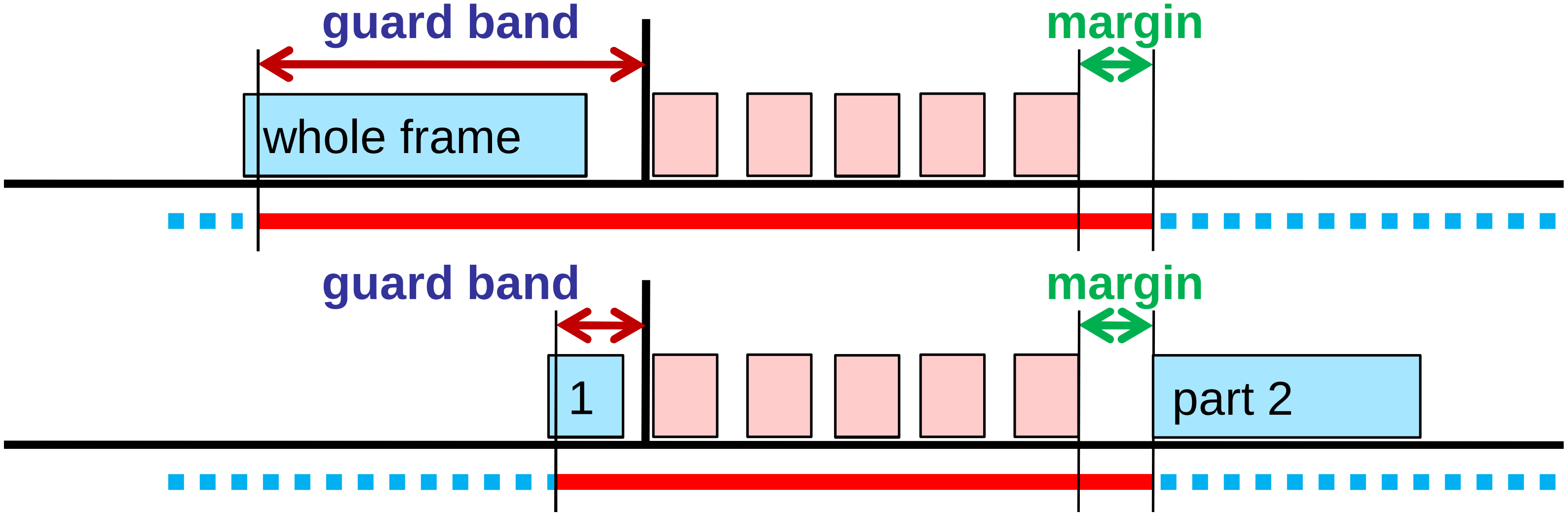
The 24 rule

- In order to transmit a mission-critical frame at a specific time, I must apply a “guard band” on the port, and not transmit any frame for a maximum frame size time before the scheduled transmission.



The 24 rule

- **If express/interspersed traffic is used, the guard band gets much smaller.**



The 24 rule

- The guard band also gets smaller (in time) if I go to the **next-higher speed** link. In that case, I get a **10x** smaller guard band.
- Assuming a 127-byte worst case for an non-interspersable frame, if I have **interspersion**, I get a (2000 + 20) / (127 + 20) = **13.7x** smaller guard band.
- Assuming a 64-byte worst case, I get a (2000+20) / (64 + 20) = **24x** smaller guard band.

The 24 rule

- The guard band also gets smaller (in time)

good if I go to the **next-higher speed** link. In that case, I get a **10x** smaller guard band.

- Assuming a 127-byte worst case for an non-interspersable frame, if I have **interspersion**, I get a (2000 + 20) / (127 + 20) = **13.7x** smaller guard band.

better

- Assuming a 64-byte worst case, I get a

best (2000+20) / (64 + 20) = **24x** smaller guard band.

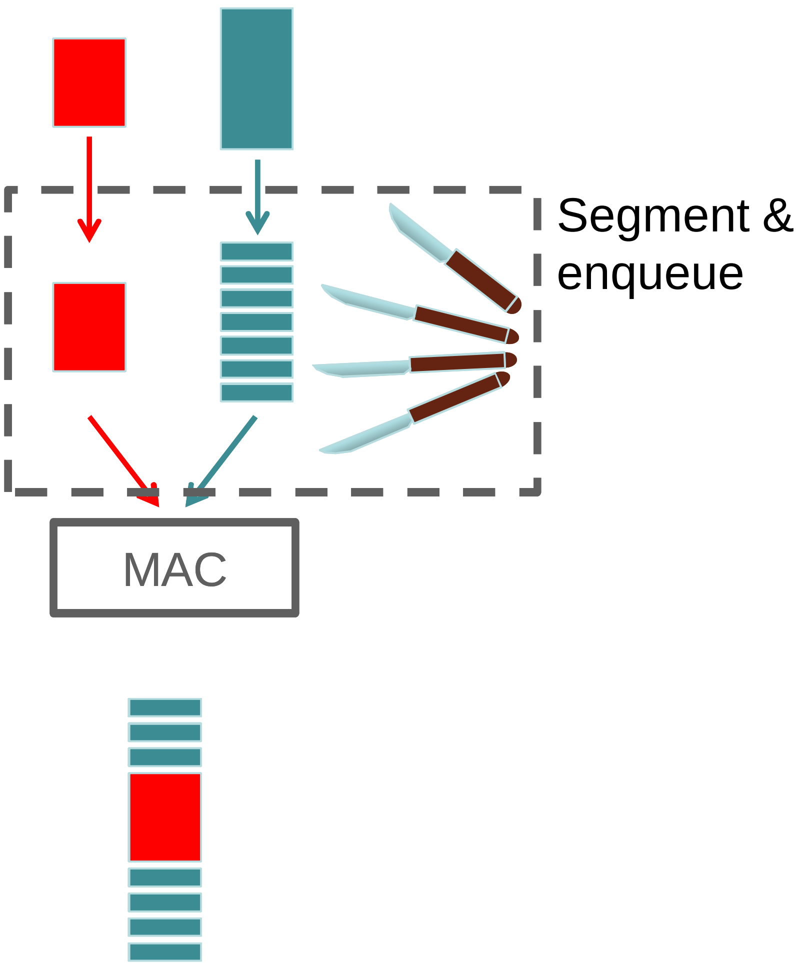
Above or near the MAC?

We will look at two interspersion alternatives that we believe are broadly representative of all schemes so far proposed:

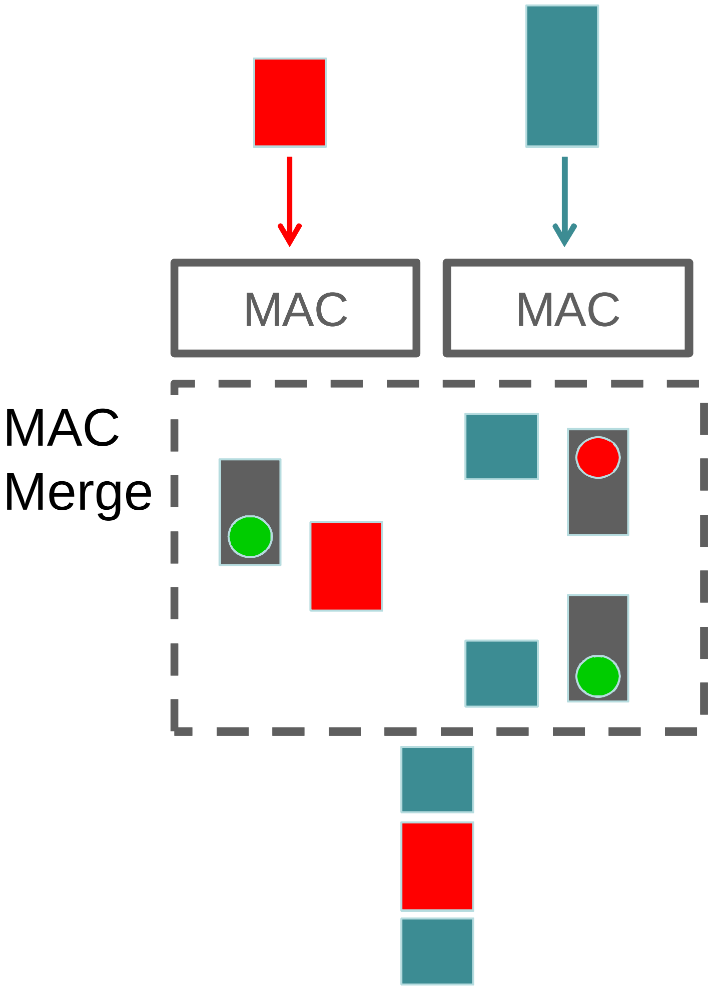
- **Preemption:** Performed **near** the MAC layer by an 802.3 function. Fragments of preempted frames are **not** valid frames under the current 802.3.
- **Segmentation:** Performed **above** the MAC layer by an 802.1 function. Fragments of segmented frames **are** valid 802.3 frames.

Segmentation vs. Preemption

Segmentation **above** MAC



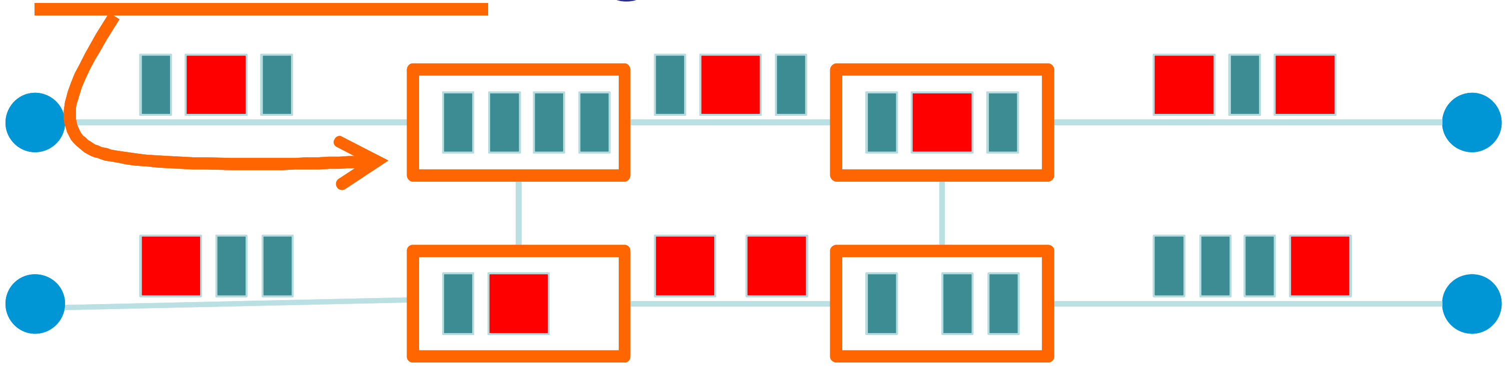
Preemption **near** MAC



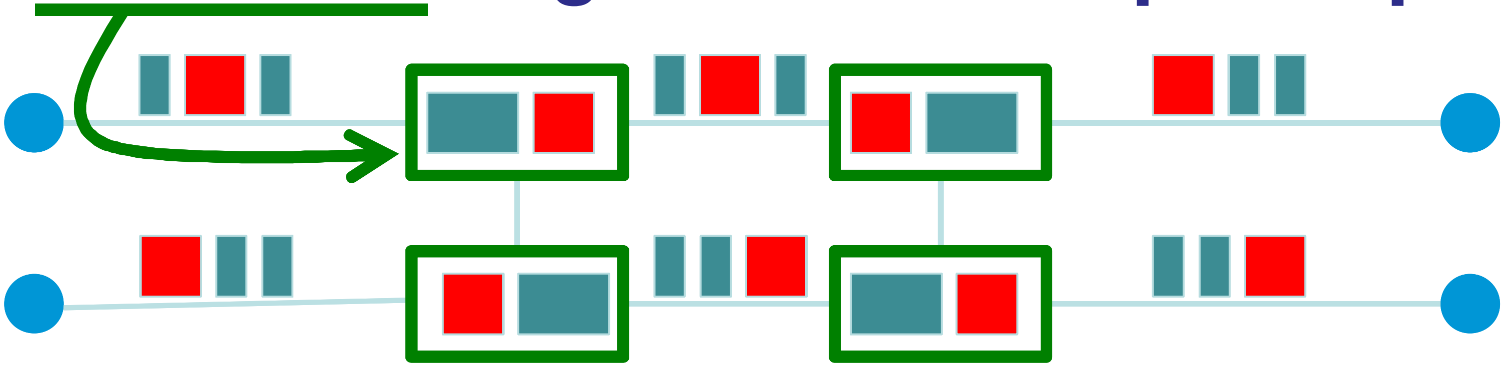
End-to-end or link-local?

- **End-to-end segmentation**

Everything gets jumbled



- **Link-local segmentation** or **preemption**



Frames reconstituted each hop

Reassembly process

- Segmentation/preemption at one point implies reassembly at some other point. The last possible point is the receiving end station.
- Segments must be buffered at the reassembly point until all the segments for a given frame have been received, or until the reassembly function gives up on the frame, and discards the incomplete set of segments.

Reassembly resources

What resources are required for reassembly by either method?

- Buffer space for frames to be reassembled.
- Means to detect frames that cannot be reassembled, and to recover the buffer.
- Means to recover the “original” CRC.

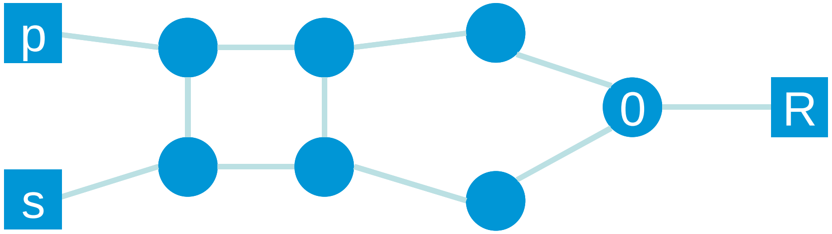
Resources needed for preemption only?

- State for interrupted MACsec or CRC functions.

Segmentation of critical frames

- We can assume that every mission-critical data stream is known to every bridge, either by configuration or by run-time protocols. This includes the source(s), destination(s), bandwidth, priority, etc.
- Therefore, the resources required to reassemble critical data streams can be known, and either the resources allocated or transmission permission denied, before they are used.

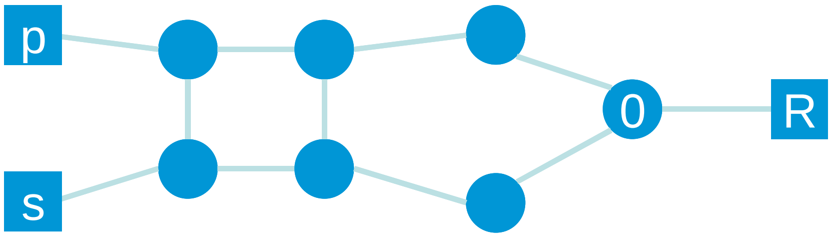
End-to-end **critical** reassembly



Easy to do!

- Sources p and s can be segmenting frames for either Bridge 0 or receiver R to reassemble. Segments for different streams can arrive in any order.
- But, since segments for a given stream are in order, then reassembly requires no timers; if frame *n* is missing fragments, the arrival of frame *n*+1 indicates that *n* can be discarded.
- No need for this has been shown to 802.1.

End-to-end **ordinary** reassembly



Hard to do!

- Streams of **best-effort** frames are **not** signaled, and **not** configured, beforehand.
- Whether two frames belong to the same stream is often unknown, and can be unknowable, to a bridge.
- Therefore, every segmented best-effort frame is an individual item.

End-to-end **ordinary** reassembly

Given the nature of best-effort traffic:

- The number of reassembly buffers is unbounded.
- Every reassembly buffer requires a timer, to recycle the buffer if a fragment is lost.
- This is not supportable.
- **Ordinary best-effort frames cannot feasibly be reassembled after the segments have been mingled in a network.**

Link-local reassembly

- **Whether doing Preemption or Segmentation, link-local reassembly is easier than for the end-to-end case:**
- There is one source and one destination.
- The buffer resources are limited to the number of levels of preemption / segmentation allowed. No valid use case for more than 2 buffers has been presented.
- No timers are required, as the arrival of a subsequent segment is sufficient to release the resources.

Near-the-MAC Link-local (Preemption)

- ☹ The MACsec facility may have to pause mid-fragment, remember its state, and resume after the fragment resumes.
  - – We mention this for the sake of transparency; it may or may not be part of any 802 standard.
- ☺ There is no chance that fragments can be interspersed, as for the end-to-end case.

Near-the-MAC Link-local (Preemption)

- ☺ A preemptable frame can start transmission, and the point at which it is interrupted, if any, can be determined at a later point in time.
  - **This minimizes the latency of the express frame.**
  - We will assume a penalty of 28 bytes per interruption (see Thaler’s slides), 20 bytes preamble and gap, 4 bytes CRC, 4 bytes other.

Above-the-MAC Link-local (Segmentation)

- ☺ MACsec can, presumably, be independent of the segmentation process.
- ☹ There is **a significant chance that fragments can be interspersed**, as for the end-to-end case.
  - – All it takes is a buffered repeater == a bridge that runs no protocols == $10 at the local electronics store.
  - – While not encouraged by 802.1, nor supported by the standards, they are a fact of life in industrial applications and vendors must deal with them.

Above-the-MAC Link-local (Segmentation)

- ☹ **Every interspersable frame must be segmented before transmission**, else there is no benefit to the procedure.
  - Assuming that segmentation requires a 4-byte tag per segment, a 2000-byte maximum frame requiring 2020 byte times to transmit, when divided into 64-byte segments, requires 3864 byte times to transmit, == 91% penalty = 52% line rate.
  - 127-byte segments = 19 pieces = 2793 byte times to transmit, 38% penalty, 72% line rate.

Link-local summary

issue	Above the MAC segmentation	Near the MAC preemption
% lost BW interrupting 2000-byte frames, 64-byte min. segment	Constant 48% of best-effort traffic	No interruptions 0% Max interruptions 33% One interruption 1%
% lost BW interrupting 2000-byte frames, 127-byte min. segment	Constant 32% of best-effort traffic	No interruptions 0% Max interruptions 19% One interruption 1%
Gain in latency for 64-byte min. segment	2020/84 = 24x	2020/84 = 24x
Gain in latency for 127-byte min. segment	2020/147 = 13.7x	2020/147 = 13.7x
Gain in latency for 10x link speed increase	10x	10x
Buffered repeater makes segmentation / preemption not work	Presence of repeater may not be detectable before application starts.	Repeater can be detected before application starts.

Summary of this segment

- Current and forthcoming 802.1 Classes of Service offered by bridges, and why they are needed for Deterministic Networking.
- We need express / interspersed transmission.
  - Primary need is to support convergence of scheduled, rate limited, and best-effort traffic on the same network.
- Interspersion must be defined near the MAC layer by 802.3, and may be defined above the MAC layer by 802.1, as well.

Technical and economic feasibility

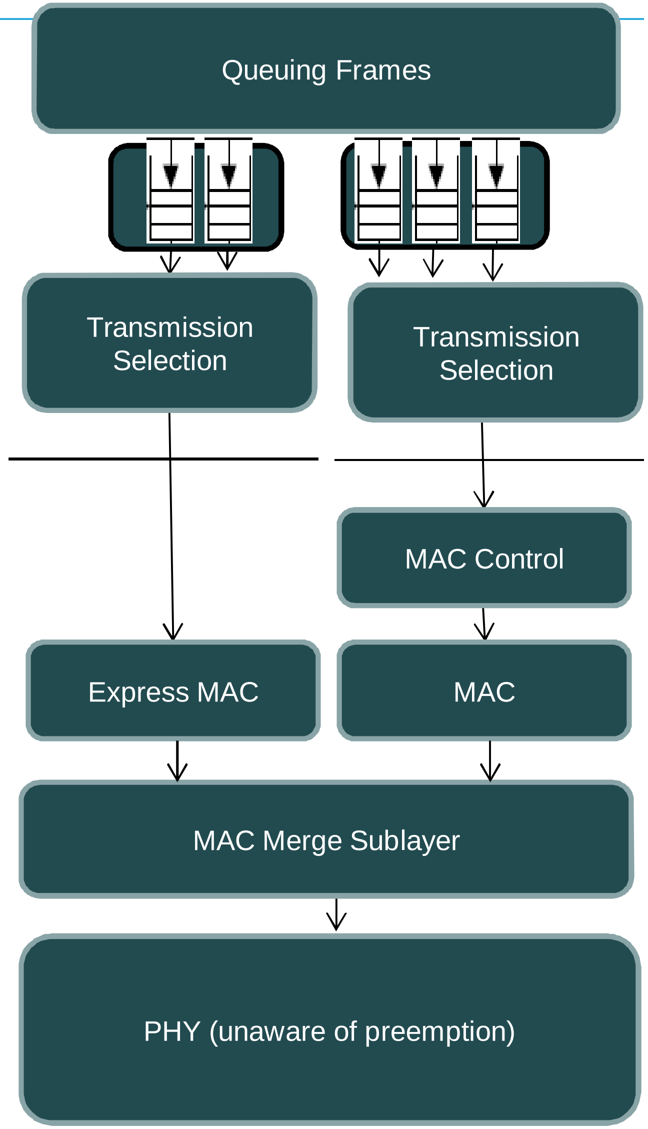
Pat Thaler	Broadcom

Methodology

- Feasibility will be demonstrated by showing a feasible solution that meets all the objectives
- This isn't meant to indicate that this is the only solution nor does it assume that the task group will adopt this solution

A Potential Architecture

- MAC Merge sublayer
  - Provides lower latency for express traffic
  - Preserves frame integrity
  - Minimizes impact on throughput
  - Is transparent to existing non-deprecated PHYs above 10 Mb/s
  - Doesn't change MAC operation



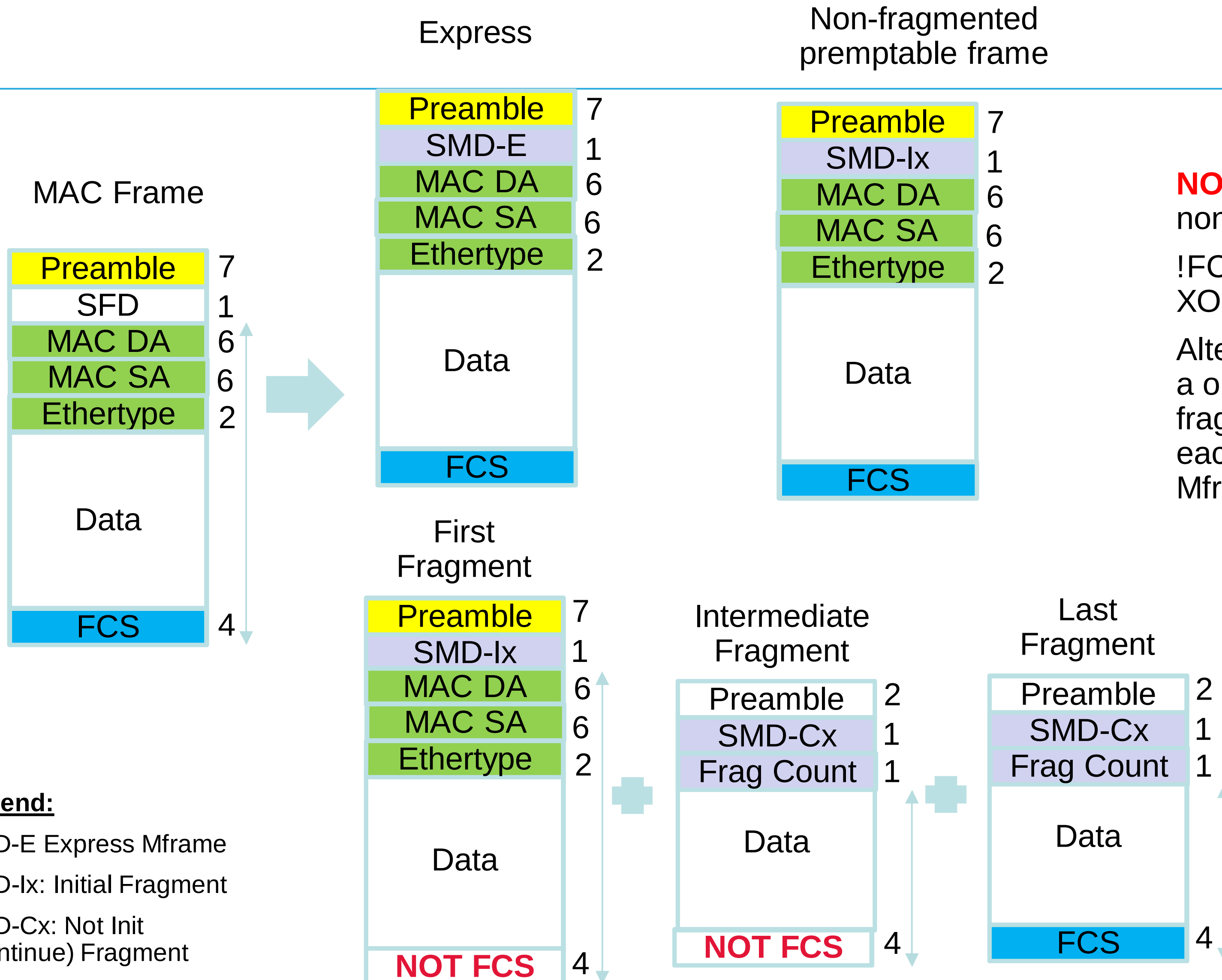
Terminology

- Express – a non-premptable frame
- Premptable – a premptable frame
- Mframe – the transmission unit from the MAC Merge sublayer
  - Looks like a frame to the PHY layer
  - May contain a whole Express frame, a whole premptable frame or a fragment

Mframe encapsulation goals

- Minimize fragmentation overhead
- Distinguish between Express frames and premptable frames
- To protect against misassembly
  - Carry frame number for start of a premptable frame
  - Carry frame number and fragment number for any subsequent fragments
  - Both are circular count of 4
- Frame CRC is generated by MAC and not altered by MAC Merge

Mframe encapsulations



PHY compatibility lower bound
Frag HDR + error detect

NOT FCS indicates non-final fragment.

!FCS in use. Use FCS XOR FF00

Alternatively, could use a one-byte fragmentation trailer on each non Express Mframe

Legend:

SMD-E Express Mframe

SMD-Ix: Initial Fragment

SMD-Cx: Not Init (Continue) Fragment

SMD and Frag Count byte encodings

Mframe type	Frame #	SMD
SMD-E	NA	0x33
SMD-Ix Premptable frame start	0	0x66
	1	0xCC
	2	0xFF
	3	0xAA
SMD-Cx Non-initial fragment	0	0xE1
	1	0xD2
	2	0x1E
	3	0x2D

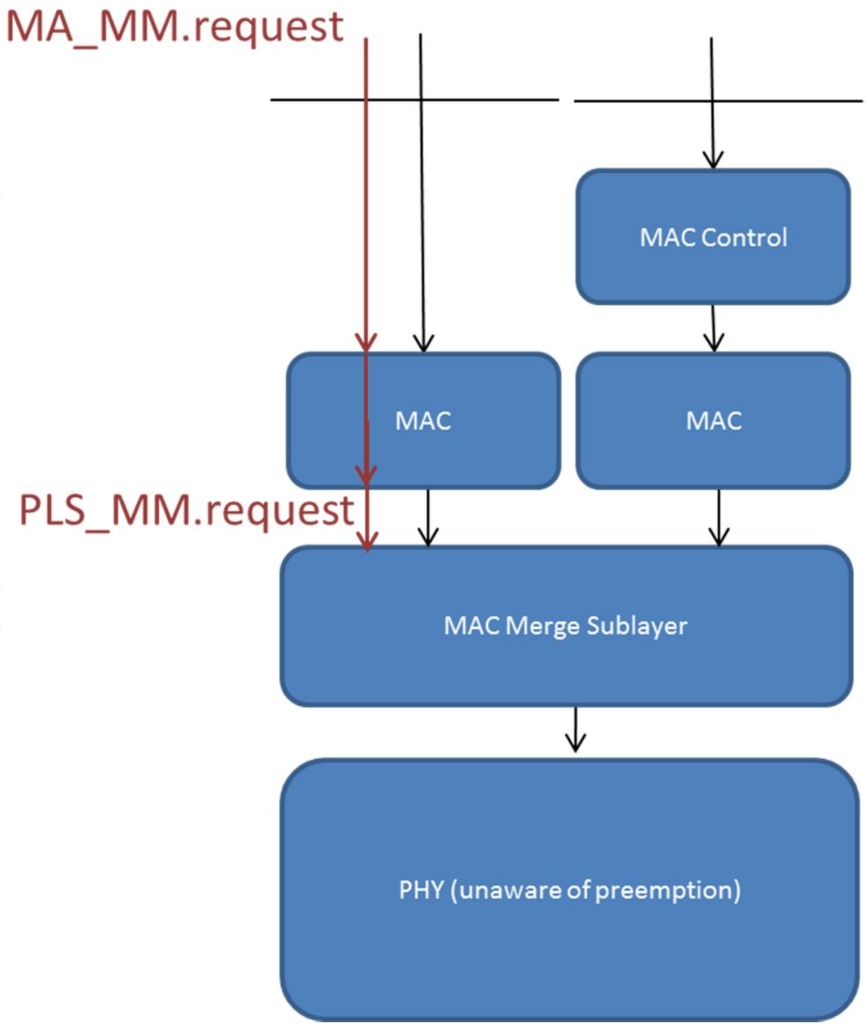
Frag Count	Frag
0	0x66
1	0xCC
2	0xFF
3	0xAA

Providing support for guardband

- The MAC Client is responsible for the schedule
  - knows when scheduled traffic should arrive
- Guard band is provided by preempting traffic before the scheduled traffic arrives
- When scheduled frame arrives, it can be transmitted immediately
- Guardband is signaled to MAC Merge with a primitive that tunnels through the MAC

MA_MM.request and PLS_MM.request

- MA_MM.request: Additional primitive on Express MAC client service interface
- PLS_MM.request: Additional primitive on interface between Express MAC and MAC Merge sublayer



MA_MM.request(hold_req) and PLS_MM.request(hold_req)

- Express MAC operation is not affected by this primitive except to send the primitive on the lower layer interface
- hold_req parameter takes one of two values:
  - hold – asserts hold variable in MAC Merge sublayer
  - release – clears hold value in MAC Merge sublayer
- MAC Merge preempts whenever hold = TRUE or Express MAC PLS_DATA.request has a bit to transmit.

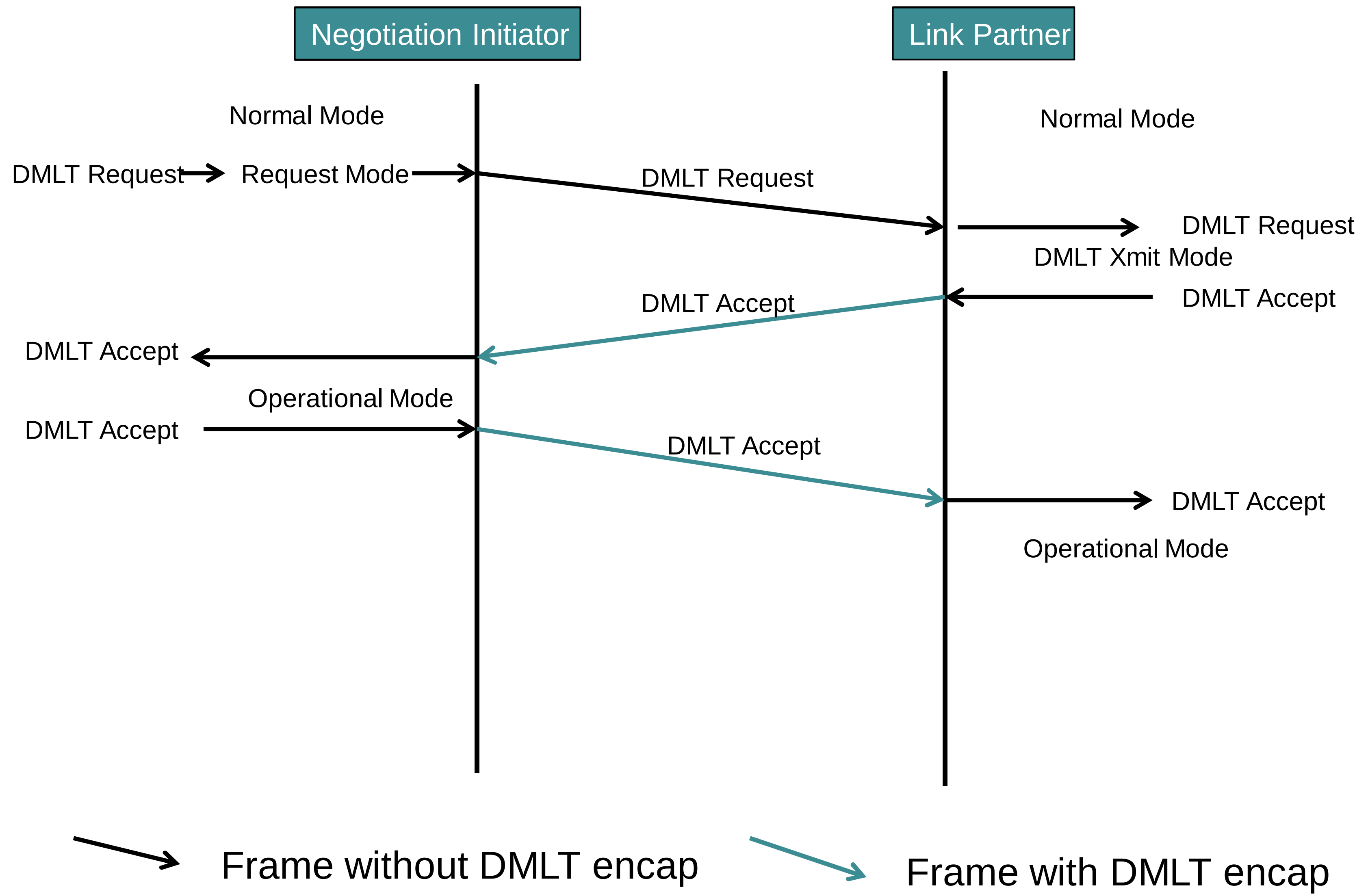
Negotiation Goals

- Should be PHY independent – only some IEEE 802.3 PHYs support auto-negotiation
- Should ensure that Mframes are only sent on links that can understand them
  - Prevent enabling on a link with a buffered repeater that doesn't block nearest neighbor address
- Can use MAC Control frames to accomplish this.

Operating modes

- Normal mode – Express MAC and MAC Merge sublayer are disabled.
- Request mode –
  - Express frame xmit is blocked
  - MAC Merge sublayer is in detect mode
    - Any frames received with normal SFD are sent to MAC
    - Once MAC Merge encapsulation is detected, receive side of MAC Merge sublayer is enabled
- Transmit mode
  - Transmission of data frames is blocked
  - MAC Merge transmit side is enabled
  - MAC Merge receive behaves same as Request mode
- Operational mode – Express MAC and MAC Merge are operational

DMLT negotiation



Recovery if MAC Merge enable fails

- If DMLT Detect Timeout expires before DMLT operation is established,
  - Transmit MAC Control DMLT fail frame
  - Start DMLT recovery timer
  - Disable DMLT MAC, DMLT MAC Control and MAC Merge sublayer
  - When DMLT recovery time expires, enable transmission of data frames

Conclusion

These presentations reflect the draft PAR, 5C, and objectives created by the SG DLMT.

The market needs from industrial automation, automotive and others should be recognized by IEEE 802.3.

The architecture options and the Technical and Economic feasibility were well presented and hopefully convinced the audience that the Study Group DMLT finalized his job so that a TF IEEE 802.3br can be approved at IEEE 802.3 closing plenary.

Before providing you a chance for questions, please let me show you the objectives for this project.

Objectives (1) – Approved in SG

1. Preserve the IEEE 802.3 Ethernet frame format at the MAC client service interface.
2. Preserve minimum and maximum frame size of the current IEEE 802.3 standard.
3. Use the Clause 4/4a MAC without alteration.
4. Support full duplex point-to-point operation only.
5. Support a speed of 100 Mb/s and above at the MAC/PLS service interface.
6. Preserve relevant MAC/PLS service interface.
7. Does not degrade (increase) undetected bit error ratio (BER) at the MAC/PLS service interface.

Objectives (2) – Approved in SG

8. Provide affirmative assurance that both end of the link have this capability before operating in this mode.
9. Provide a mechanism for reduced access latency where the reduced access latency is significantly less than one maximum packet transmit time.
10. Maximum latency for DMLT frame transmission (ahead of the non-DMLT frame) will be as close to the minimum packet size + IPG (1st and last) as practically possible.
11. Quantify the maximum access latency of the DMLT transmit path.
12. Provide two MAC service interfaces at each end of the DMLT link, as the means to distinguish between the DMLT and the ordinary traffic.
    – Optional MAC Control sub-layer shall be confined to the ordinary MAC Service Interface.

Objectives (3) – Approved in SG

13. Address the impact between Energy-Efficient Ethernet and DMLT operation.
14. This project will be media independent.
15. Require no changes to existing Point-To-Point full-duplex PHYs.
16. Consider providing, at the MAC Client Service interface, a primitive that holds the transmit path in the express position.
17. “M-Frame in the wild” should be constructed such that it will not be forwarded by non-DMLT-capable devices.
    - Buffer repeater e.g. legacy TPMR would be “ in the wild“.

Objectives – OPEN & Consideration

1. Preserve [Clause 4/4a] frame format on the respective physical medium.
2. Support of the Point to Multipoint (P2MP) is not a goal.
   – Downstream Support presents fewer challenges.

THANK YOU

for your attention