Presentation to USPTO

- The following slides were presented at a “Technology Fair” at the United States Patent and Trademark Office in Alexandria, Virginia, on June 11, 2013, by Norman Finn of Cisco Systems, to an audience largely composed of patent examiners.
- It was also emailed to a subset of IEEE 802.1 members.
- Michael Johas Teener, chair of the IEEE 802.1 TSN TG, suggested I post it to the 802.1 public website.

Deterministic Ethernet

Norman Finn
Cisco Fellow
Cisco Systems, Inc.
June 10, 2013

Abstract

Standards organizations, vendors, and users are very active in developing solutions to the problem of real-time control networks using Ethernet technology, both wired and wireless. This requires Ethernet to offer reliability and delivery time guarantees far beyond anything it has previously attempted.

Author

Norman Finn.

B.S. Caltech

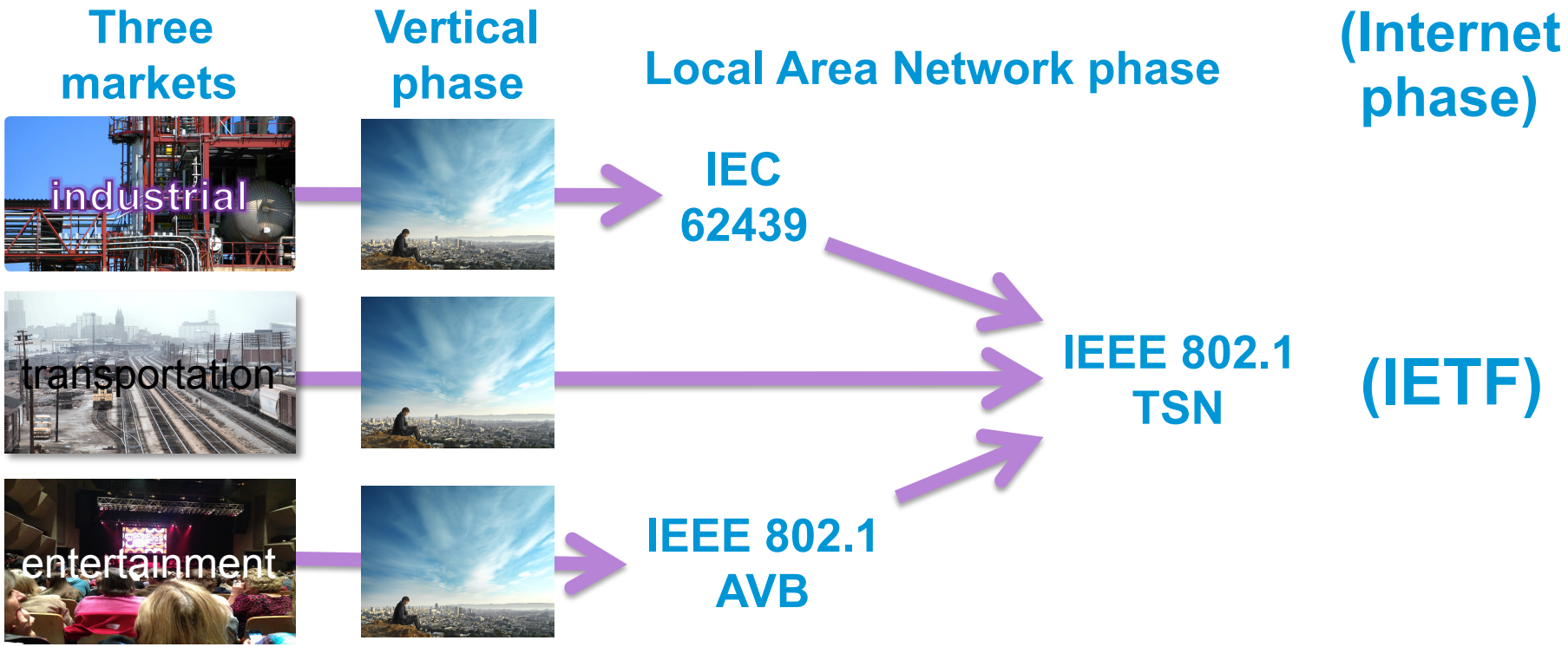
Cisco Fellow.

~60 patents awarded, ~25 in progress, mostly for Cisco Systems.

Chairs one of Cisco's 15+ patent idea review committees.

Active participant in standards activities, including IEEE 802.1.

A Technology Trajectory



- From individual market needs – **in the past**
- To vertical standards – **now in use**
- To Local Area Network standards – **now being developed**
- To Internet standards – **before too long**

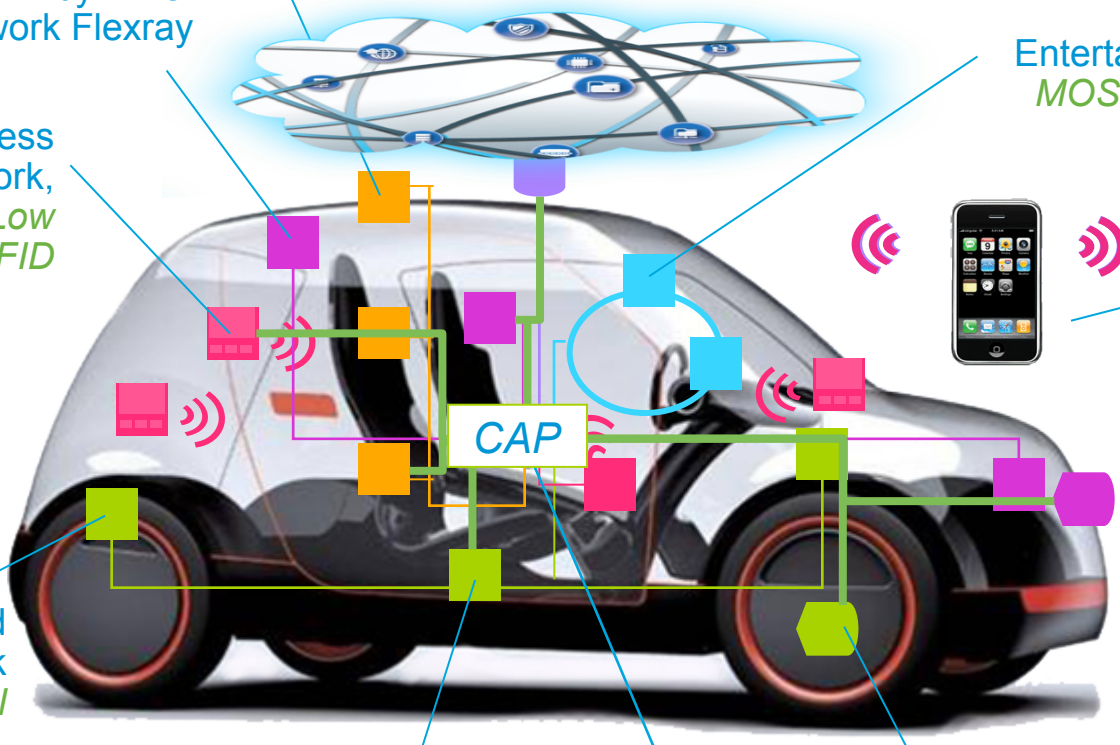
Customers have similar needs

- Heavy audio/video users: Television studios, movie studios, recording studios, sports event producers, stadiums, theme parks.
  - Live, real-time data is the end product.
- Industrial process and machinery control users: Assembly lines, chemical processing, paper mills, food processing.
  - Complex processes with feedback loops with sensors and actuators.
  - Wind turbine synchronization, printing press paper feeding, copier machine.
- Vehicles: Automobiles, trains, and planes.
  - Fuel/air ratios, anti-skid braking, collision avoidance, landing gear retraction.
- **Time synchronization**
- **Guaranteed delivery of a data packet within a guaranteed time window.**

Standards Development Organizations (SDOs)

- SDOs and the Standards Process
  - Computer communication requires either standards or a monopoly.
  - Vending companies, consuming companies, universities, and government agencies meet to create computer communications standards.
  - **A rising tide lifts all boats.** Standards expand the market, benefiting (almost) all participants.
- Verticals: Paid company memberships (BBF, MEF)
  - Single industry, only paid members can use the IPR.
- RAND: Reasonable And Non-Discriminatory (IEEE, IETF)
  - Participants in the standards promise to license the use of their IPR to those implementing the standard on a reasonable and non-discriminatory basis.
- Treaty and national organizations: Mandated by law (ITU, DIN)
  - It's against the law in some countries to violate some of these standards.

Automobiles are computer networks!



Automobiles are computer networks!

- Computers had to go into automobiles.
  - Emission standards
  - Satellite radios
  - Automatic Braking
- Computers need to talk to sensors/actuators and to each other
  - Air temperature, pressure, humidity, engine temperature
  - Fuel injectors, brake cylinders
  - Engine control, chassis control, infotainment
- **Verticals** developed different digital communications media:
  - FlexRay
  - CAN Controller Area Network
  - MOST Media Oriented Systems Transport
  - USB Universal Serial Bus
  - Wi-Fi IEEE 802.11 wireless
- **All of these are now in every car built. That's too much!**

Entertainment

- Same story for audio/video digital distribution
- Originally analog
  - NTSC, PAL, coaxial cable
- Then, digital
  - HD-TV, DVI, HDMI
- But there is a big demand in studios, stadiums, and theme parks for a **video infrastructure** that is **converged with the data infrastructure**.

Industrial process and machine control

- First, proprietary digital links (some with vertical standards)
  - Modbus
  - Profibus
  - DF-1
- Then proprietary variants of Ethernet and vertical standards
  - TTTech
  - Profinet
- Then protocols that use “dumb” bridges and “smart” stations
  - IEC 62439 HSR: High-availability Seamless Redundancy
  - IEC 62439 PRP: Parallel Redundancy Protocol

Standards: IEEE Bridging versus IETF Routing

- Ethernet bridges defined in IEEE 802.1 working group
  - Bridges are an Ethernet (and Wi-Fi) technology
  - Addresses are 48-bit MAC addresses assigned by manufacturer.
  - MAC addresses are “who,” not “where”.
  - Default is to flood everywhere, hoping to find station.
  - **Local in scope.**
- Routers defined in various IETF (Internet Engineering Task Force) working groups.
  - Routers use any and all physical media.
  - Addresses are 32-bit or 128-bit IP addresses assigned by network admin.
  - IP addresses are “where.” IPv6 128-bit addresses can also be “who.”
  - Default is to drop a packet if the router doesn’t know where to send it.
  - **Global in scope.**
- IP packets often ride over (inside) Ethernet frames.

Common problem statement

- **Ethernet** – LAN is right size, ubiquitous in enterprise, inexpensive, nearly good enough without change.
- **Clock synchronization** – Every station must agree on what time it is to within 10 ns – 1 µs.
- **High delivery probability** – From 10^{-5} packet loss rate to 10^{-12}
- **Latency** – Worst-cast delivery time across network must be guaranteed, 100 µs – 50 ms.
- **Compatibility with Wi-Fi** – Though not at 10^{-12} loss rate!
- **Convergence** – Critical and best-effort traffic share same network.
- **Flexibility** – Must be possible to alter the configuration without laboratory work.

First round of results – 2002 to 2010

- IEEE 1588 Precision Clock Synchronization
- IEEE 802.1BA AVB profile
- IEEE 802.1AS Plug-and-play profile of 1588
- IEEE 802.1Qat Stream Reservation
- IEEE 802.1Qav Traffic Shaper

IEEE 1588

- Before 1588, the usual way to synchronize clocks was to insert a timestamp into a packet as it is transmitted, then insert another as the packet is relayed and/or finally received. (E.g., ITU-T Y.1731)
  - Requires re-spin of every port ASIC every time any standard changes.
  - Level at which data can be inserted is far from the actual PHY (physical) layer, and especially, above MAC-layer encryption, leading to inaccuracy.
- 1588 uses separate “Sync” and “Followup” packets.
  - Sync is not changed as it moves. Hardware takes a locally-meaningful nS timestamp when the packet leaves or arrives on the wire (or aether).
  - Software (or firmware) reads timestamp, converts to universal format, transmits Followup very soon after Sync.
- Now, hardware only has to be told “Remember when you sent this packet,” and hardware only has to say, “This is when I received this packet.” Only software worries about packet formats.

IEEE 802.1 AVB → TSN

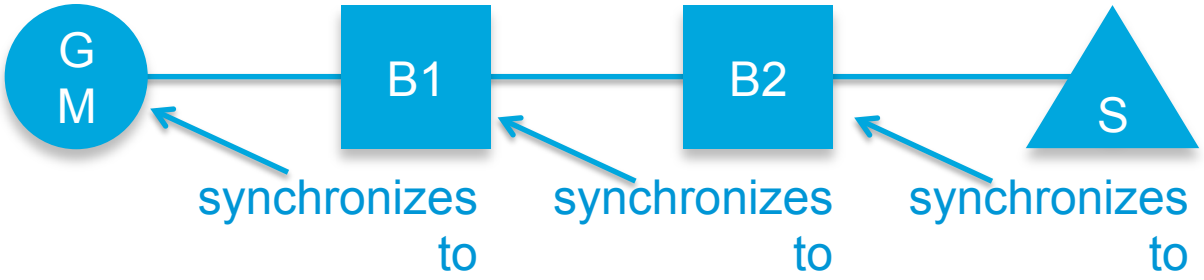
- IEEE Audio Video Bridging (AVB) Task Group formed in 2006 to work on converting the home to Ethernet and Wi-Fi, instead of speaker wire, coax, and wired and wireless HDMI.
- (Didn't happen.)
- But "professional" AV – TV studios, recording studios, theme parks, stadiums, on-site events, and theaters – did pick up on it.
- In 2012, name changed to Time-Sensitive Networking, as industrial and transportation markets joined the existing infotainment markets in looking for standard Ethernet solutions.

IEEE 802.1 AVB: 802.1AB

- Describes requirements for bridges and end stations to be compliant with AVB requirements.
- Tells which options in 802.1Q are needed and not needed.
- Tells what protocols and options an end station needs to implement.
- Specifies default configuration parameters to be set at the time a bridge or end station is manufactured.
- Specifies which of the 8 bridge priority levels are for AVB traffic.
- If followed, guarantees a plug-and-play network of modest size will deliver packets, with a reasonable probability, within 2 ms or 50 ms, as selected by the end stations' applications.

IEEE 802.1 AVB: 802.1AS

- Plug-and-play “profile” of IEEE 1588.
- Automatically selects a “Grand Master” clock – the one that claims the most accuracy (e.g., atomic clock, GPS sync, WWV sync, crystal oscillator).
- Automatically constructs a distribution tree from the Grand Master to all clock users.
- Every bridge (or router) along the path synchronizes with the Grand Master.



- **Accuracy good to better than 1 µS.**

IEEE 802.1 AVB: 802.1Qat

- Uses Link Layer Discovery Protocol (802.1AB [not BA]) to determine which end stations and bridges are (not) 802.1BA compliant, and defends network capabilities against them.
  - All traffic from non-compliant stations or bridges at an AVB priority level are remapped to another, best-effort, priority level, to protect the AVB queues.
- Runs Stream Reservation Protocol to reserve bandwidth for AVB streams.
  - End stations register as “Talkers” and/or “Listeners”.
  - Bridges notify end stations of the AVB traffic classes’ abilities.
  - A Talker requests permission from the nearest bridge for permission to transmit an AVB stream with a certain destination MAC address and a certain bandwidth, on a specific AVB traffic class.
  - The bridges propagate Talker and Listener registrations, allocate resources, and grant or refuse Talkers permission to transmit.

IEEE 802.1 AVB: 802.1Qat

- Why make reservation?
  - Assumption is that applications (e.g. a video recorder or camera) know what streams are needed.
  - This makes network plug-and-play.
- What if the network says, “No!”
  - The network cannot supply bandwidth that does not exist.
  - Talker either does not transmit stream, or transmits the stream on a best-effort traffic class, instead of an AVB class.
- Special features:
  - High-ranking reservations (e.g., 911 calls) can bump other reservations.
  - If network topology changes (new link, failed bridge, etc.), reservations are remade automatically. There may be a brief disruption of guarantees, and reservations can be granted or denied as a result of the network acquiring more, or losing some, resources.

IEEE 802.1 AVB: 802.1Qav

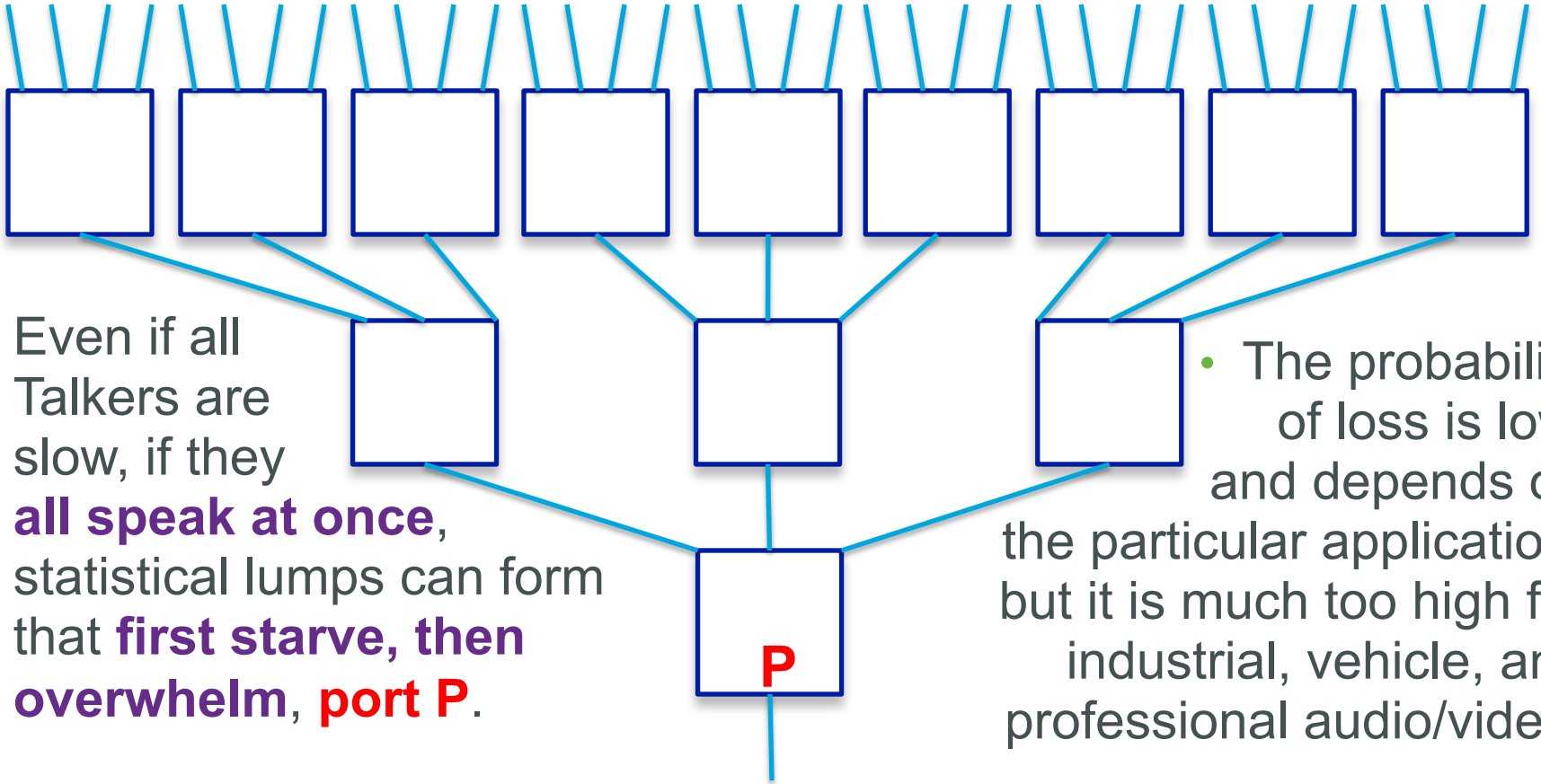
- Defines a variation of “classic” traffic shapers defined by IETF and MEF.
  - One FIFO queue per port per AVB traffic class (typically 2). **Not** a queue per AVB stream.
  - Queue shaper accumulate credit at Talker-registered stream rate while packets are waiting in an AVB queue, and expends credit at (line rate)—(stream rate) while transmitting.
  - Whenever port is idle, highest-priority AVB queue with a packet and credit ≥ 0 transmits.
  - When queue is empty, credit ← 0.
- As reservations are made via SRP, each queue is configured for the total bandwidth of all streams passing though that port in that queue’s traffic class.

They've gone about as fer as they kin go

- AVB is plenty good enough for the home, but TSN (industrial, transportation, and professional audio/video) networks need:
  - **More accurate time synchronization.**
  - **Better protection against network and end station failures.**
  - **Much better guarantees against late delivery and packet loss.**
- What causes late delivery and packet loss?
  - Bridge (router) or link **failures** lose or corrupt packets.
  - **Cosmic rays** or other radiation zapping memory cells lose or corrupt packets.
  - **Congestion**, triggered by random synchrony of transmission events on different links, causes excessive buffer delay or packet loss.

Congestion loss

- **Traffic shapers** do not **guarantee** latency or 0 congestion loss because of the “fan in problem.”



- Even if all Talkers are slow, if they **all speak at once**, statistical lumps can form that **first starve, then overwhelm**, **port P**.

- The probability of loss is low, and depends on the particular application, but it is much too high for industrial, vehicle, and professional audio/video.

Why all this effort?
Why can't you just … ?

- Overprovision a faster link (this is done, but …)
  - This works much better in isolated networks, than in converged networks.
  - Vehicles are limited to 100Mb/s – 1Gb/s (future) by noise immunity.
  - Uncompressed audio/video streams are too fast to overprovision.
    - (Audio/video is mission critical??? It is to a TV studio!)
- Use duplicate networks (this is done, but …)
  - Both networks can have the same congestion characteristics.
  - Separate critical and ordinary networks → 3 or 4 networks for critical redundancy → demand for load sharing → duplicate converged networks.
  - Vehicles cannot afford the weight/cost penalty of multiple networks.
  - Separate critical/ordinary networks are a training/maintenance/operations/ logistics pain.

Back to the future

- Telephone technology has always offered Constant Bit Rate service using Time Division Multiplexing (TDM) techniques. Asynchronous Transfer Mode (ATM) offered CBR service.
- What does TDM mean to the customer?
  - Customer purchases a service with *X* Mb/s.
  - Customer supplies data at a rate ≤ *X* Mb/s.
  - Data gets through with a 10^{-8}–10^{-10} packet loss rate due to cosmic rays.
  - No data is lost due to congestion, because bandwidth is allocated using TDM.
- What do packet networks offer today?
  - Traffic shaping achieves 10^{-3}–10^{-5} packet loss rate due to congestion in a busy network. (Maybe 10^{-6} loss rate with gross overprovisioning.)

"Mission-critical" == packet loss rate in the 10^{-9}–10^{-14} range (or better).

How does IEEE 802.1 TSN meet these needs, while supporting convergence?

- Some needs and solutions are discussed, but not standardized, in IEEE 802.1 meetings.
- IEEE Std 802.1aq – ISIS protocol for bridging
- Project P802.1ASbt – Improvements to 802.1AS time sync.
- Project P802.1Qbu – Preemptive transmission.
- Project P802.1Qbv – Time scheduled transmissions.
- Project P802.1Qca – Multiple path creation..
- Project P802.1CB – Duplicated and re-collapsed data streams.
- Project P802.1Qcc – Revised Stream Reservation Protocol.

Paranoia

- Vendors must be very careful about bridge design to ensure that memory failures cannot result in a corrupted packet with a valid checksum. (*caveat emptor*)
- Protection against misbehaving nodes, stations, and links requires that the end stations and network nodes all be paranoid:
  - Timed input gates and filters for critical traffic in order to guard against failures that cause nodes to become chatty, and to prevent pollution of critical resources by poseur non-critical traffic.
  - Application data checking and modeling (beyond the scope of the network).
  - Security, often involving cryptographic techniques, against various threats (a deep subject, not further addressed).
  - Individual packets failing tests are discarded. Ports attached to stations or bridges that repeatedly transmit bad data are blocked from receiving. That is, any misbehavior becomes one or more link failures.

IEEE Std 802.1aq Shortest Path Bridging

- Spanning tree algorithms have defined bridges until 2012.
- Now, bridging can be controlled using the ISIS (Intermediate System to Intermediate System) protocol defined by ISO/IEC.
- This is a link state protocol, not a distance vector protocol.
  - Distance vector (spanning tree) is global information distributed to one's local neighbors.
  - Link state (ISIS) is information about one's local neighbors distributed globally.
- In the worst case, ISIS converges faster than spanning tree.
- ISIS allows all paths in the network to be used, and allows multicasts and broadcasts to take the best path from the source to all listeners.
- All new work in 802.1 TSN is based on ISIS.

P802.1ASbt Improvements to time sync

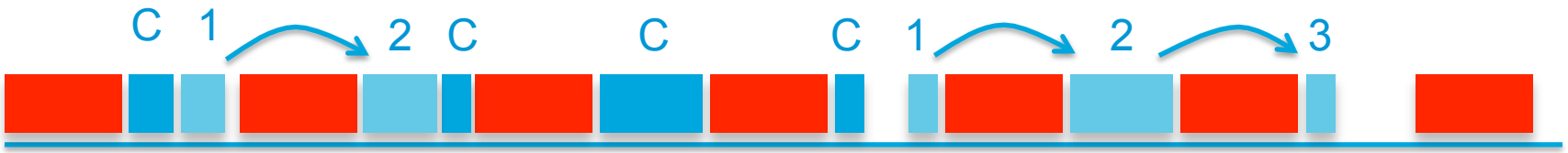
- (This is the author's opinion on a work very much in progress.)
- The TSN Task Group has plans for supporting:
  - Multiple Grand Master clocks so that stations can switch to a backup faster.
  - Separate paths for distributing the Grand Masters' time, to minimize the chance that a single link failure will disconnect a station from the master.
  - Improvements in accuracy.
  - Various minor technical improvements.
  - Improved distribution of standardization between the IEEE 1588 "root" document and the various profiles being generated by IEEE 802.1, ITU-T, IETF, and other organizations.

P802.1Qbv Time-scheduled transmission

- (This is the author's opinion on a work very much in progress.)
- This is the "Time Domain Multiplexing" part.
- Allows queues to be turned on or off on a detailed schedule that repeats over and over.
- A packet cannot start transmission unless it can finish (or be preempted) by the end of the transmission window.

P802.1Qbu Preemptive transmission

- (This is the author's opinion on a work very much in progress.)
- Convergence does **not** mean, "Non-critical traffic is unimportant." We must offer useful levels for Quality of Service to non-critical traffic.
- Time-scheduled transmission of mission-critical data using TDM techniques also means that the mission-critical transmissions can be made very regular, which in turn, means that best-effort transmissions are very regular.
- If we then preempt best-effort traffic, then the existing 802.1Qbv shaped traffic can still make its probabilistic guarantees, which are satisfactory for all non-professional uses, based on lowered available bandwidth.



P802.1Qbz/.11ak Wired/wireless bridging

- (This is the author's opinion on a work very much in progress.)
- This project will support an integrated bridge / Wireless Access Point, or a bridge with one (or more) wireless stations as ports.
- Instantly, all of the features of 802.1 AVB and TSN networks become available to Wi-Fe, including time-scheduled transmissions, multiple paths, packet replication and elimination, etc. (Though probably not preemption.)
- Of course, the reliability of Wi-Fi links is, by nature, far lower than that of wired links.
- Nevertheless, the techniques discussed, here, can improve Wi-Fi reliability to the point that it will be useful for some real-time applications, and better suited to many enterprise etwork apps.

P802.1Qca – Multiple path creation

An aside:

- There are many existing ways to ensure a reliable network in the face of failures of nodes or links in the network:
  - Spanning trees or ISIS or other routing protocols that work over any topology, prevent endless forwarding of packets in a loop, and ensure the detection of and recovery from outright failures of links, bridges, or routers.
  - "Protection switching" methods that work over any topology, but use administratively-established primary and alternate paths for every conversation. Frequent transmission of "OAM" packets ensures that all paths are continually monitored, and data traffic is switched from one to another after the detection of a failure.
- A third way is supported, today, using IEC 62439, with limitations:
  - Send identical data along two different paths. Let the recipient throw away the excess. No fault detection time, no reaction time. Lack of reception along one path triggers repairs in human time.

P802.1Qca – Multiple path creation

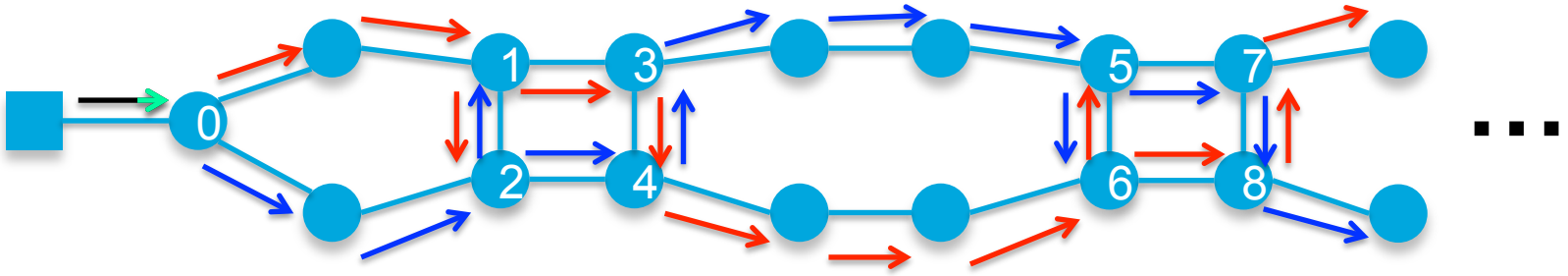
- (This is the author's opinion on a work very much in progress.)
- Clean failures and checksum errors can be handled by using redundant data paths **not** repaired by a routing/bridging protocol.
  - Sending data redundantly along two or more paths eliminates any requirement to detect and react to a failure, and hence eliminates the time required to detect and react. End points simply don't care.
- Protection against misbehaving nodes, stations, and links is assisted by redundancy:
  - Misbehavior, when detected, can be mitigated by isolating the misbehavior, and letting redundancy preserve the network operation.
- Redundant paths, typically more than two, and typically involving path separation by (at least) time, space, and frequency, can bring the reliability of wireless links up to meet the needs of at least some critical applications.

P802.1CB Stream replication/elimination

- (This is the author's opinion on a work very much in progress.)
- Uses either the multiple paths of 802.1Qca or protocol-controlled paths provided by ISIS or any other protocol.
- Either sending end station (if station has multiple ports) or first bridge (if station has only one port) replicates every packet, marking each copy with the same serial number.
- Either receiving end station(s) or last bridge eliminates all but the first-received copy of each packet.
- This can happen multiple times within a network.

P802.1CB Stream replication/elimination

- (This is the author's opinion on a work very much in progress.)
- Multiple replication/elimination for added resiliency:



- Bridge 0 generates two copies of special packets from the source.
- Bridges 1, 4, 6, and 7 pass on red copies, and generate one red copy from whichever (blue or red) is received first.
- Bridges 2, 3, 5, and 8 pass on blue copies, and generate one blue copy from whichever (blue or red) is received first.
- **Multiple failures are required to prevent successful delivery.**

P802.1Qcc Revised Stream Reservation Protocol

- (This is the author's opinion on a work very much in progress.)
- Different Layer 2 protocols (e.g., ITU-T rings, ISO rings, IEEE spanning tree or SPB) can run on different VLANs, creating multiple virtual networks on one physical infrastructure.
- Layer 2 bridging and Layer 3 routing can both be desirable in a relatively small network.
- So, many independent methods can operate simultaneously to decide to which output port a received packet is to be forwarded.
- But, when it comes to Quality of Service (QoS, meaning priorities, shapers, weighted queues, etc.) and especially Time Division Multiplexing (TDM), Layer 2 and Layer 3 and protocols disappear **-- there are only boxes and links. There can be only one choice for what packet is transmitted next.**

P802.1Qcc Revised Stream Reservation Protocol

- (This is the author's opinion on a work very much in progress.)
- Furthermore, in many applications, we do not want to have software in every device specially tailored to the specific use case (a particular machine in a particular factory). The end station may not even know whether it is connected to a bridge or a router.
- We need, therefore, a means to express a station's need for:
  - Network services: Multi-pathing, packet replication, maximum latency, allowed loss ratio, etc.
  - Stream specifics: Maximum bandwidth, largest/smallest packets, minimum interval, L2 and L3 destination addresses, upper layer protocols, etc.
- In a manner that is **independent of whether the network device(s) adjacent to the station are bridges or routers**.
- This is what P802.1Qcc will provide.

P802.1Qcc Revised Stream Reservation Protocol

- (This is the author's opinion on a work very much in progress.)
- Finally, there are multiple ways to do the complex computations required to pull together all of the stations' requests, and determine whether all can be fulfilled, and how to fulfill them:
  - Peer-to-peer protocols run among the bridges, routers, and end-stations.
  - An all-seeing all-knowing server (i.e. the IETF "Path Computation Element").
  - Off-line calculations for a particular application, downloaded to all the devices in the network while not operational.
- At this stage, the final version of 802.1Qcc may incorporate any or all of these methods.

The NP-complete problem

- The complexity of the calculations required to produce a detailed schedule of transmissions in a network in order to assign each its time slot, so that congestion losses are impossible, is an NP-complete problem, which means that the amount of work required grows exponentially (or faster!) as the number of stations, network nodes, and links grows. This makes the ideal solution impossible to achieve.
- But, if one relaxes the problem constraints, e.g. offering worse fixed latency guarantees, or less fine-grained bandwidth measurements, then the problem can be simplified. However, it becomes much more likely that the simplified method cannot find a solution, when a solution actually does exist.

The bottom line

- Were there a perfect answer, then one patent would cover it.
- There is not. There are lots of ways to slice the problem up to offer various trade-offs between complexity and completeness of the solution.
- That means there are still lots of patent applications to come in this area.
- That means plenty of work for the USPTO, and plenty of patent application bonuses for inventors. **We all win!**

Questions?

Thank you.