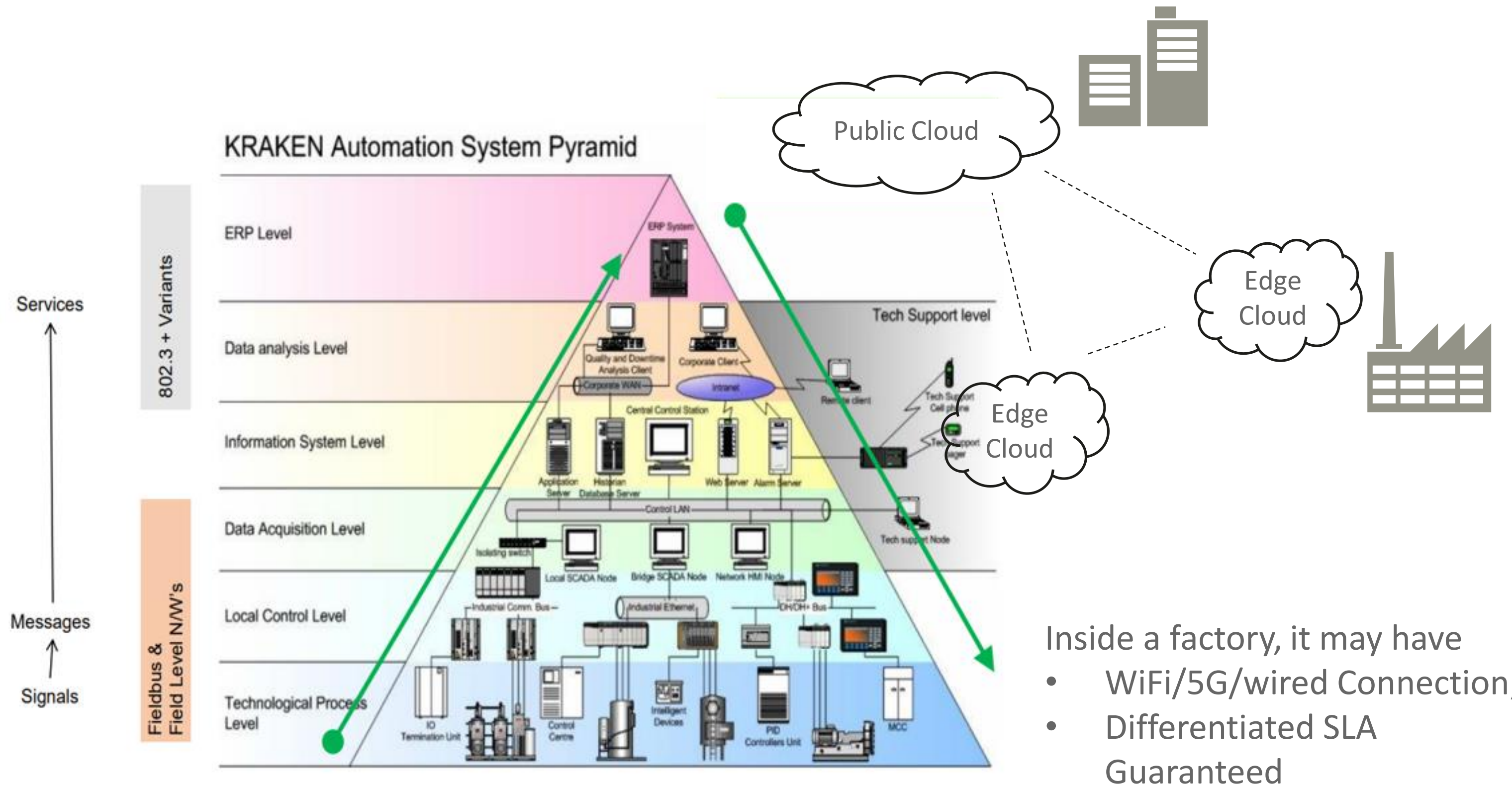
Industrial internet over service provider networks

Tongtong Wang	Huawei Technologies
Xueqin Jia	China Unicom
Huanan Chen	China Telecom
Ruibo Han	China Mobile

Agenda

- Industrial internet cases over carrier networks
  - Manufactory
  - Smart grid
  - Smart port
  - Network slicing
- Multiple approaches to provide guaranteed SLA (service level agreement) in carrier networks
  - Concept and comparison
- Suggestions in TSN for Service Provider Networks discussion
  - How different TSN Ethernet techniques fulfill industrial internet requirements
  - Recommendations for bounded latency/jitter/reliability
- Some initial research and recommendations

Enterprise-wide network over carrier networks



Traditional hierarchy for industrial network

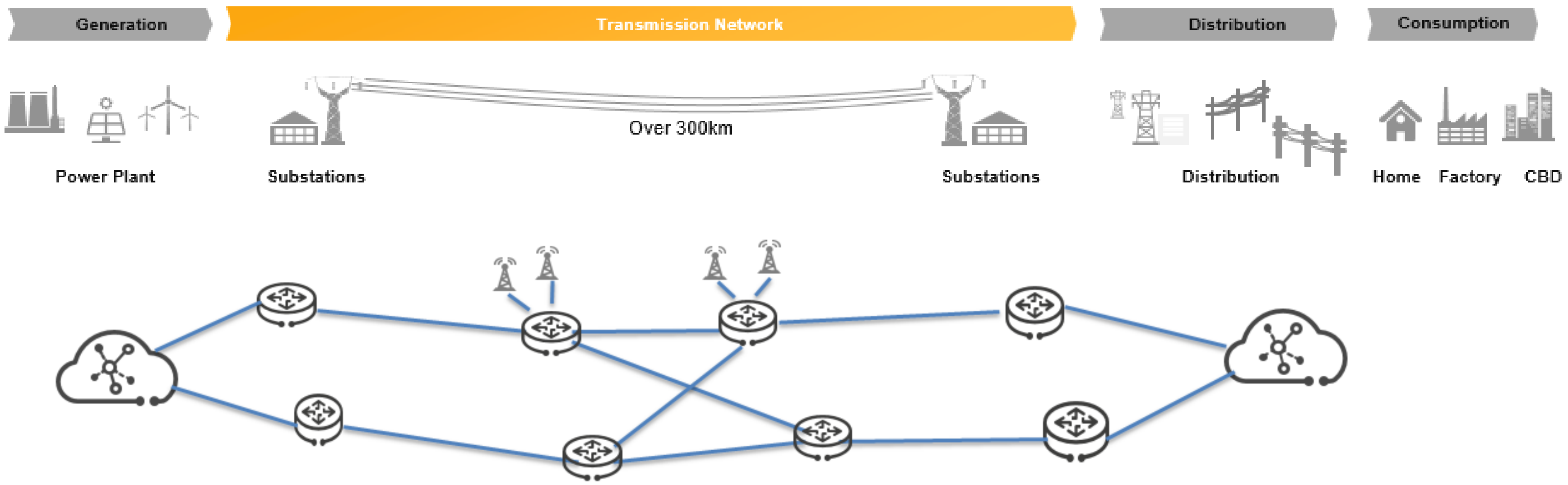
Inside a factory, it may have
- WiFi/5G/wired Connection;
- Differentiated SLA Guaranteed
- Network Slicing

(converged networking)

- Different network requirements on traditional industrial hierarchy.
- Service provider networks connect multiple remote factories/buildings.
  - Coordinated computation in public cloud
  - Remote monitoring
- SLA guaranteed service provider networks enable diverse vertical applications, e.g. to coordinate remote operations/manufactories.
- **Requirements:**
  - High reliability

http://www.ieee802.org/3/ad_hoc/ngrates/public/18_03/woods_nea_01_0318.pdf

Smart grid requirement over carrier networks



Strict requirements:

- ultra high reliability
- bounded low latency

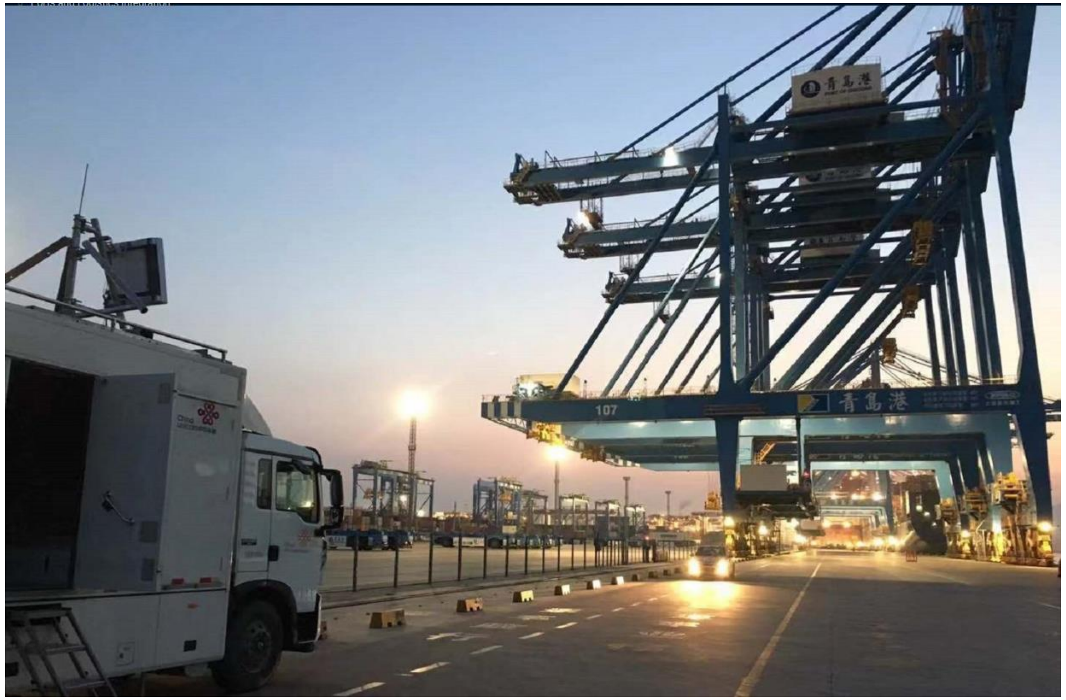
Traditionally smart grid monitoring and controlling applications are connected by separate networks.

5G URLLC networking enables Smart Grid tele-protection over carrier network, which requires strict latency bound (5ms) according to 3GPP TS 23.501.

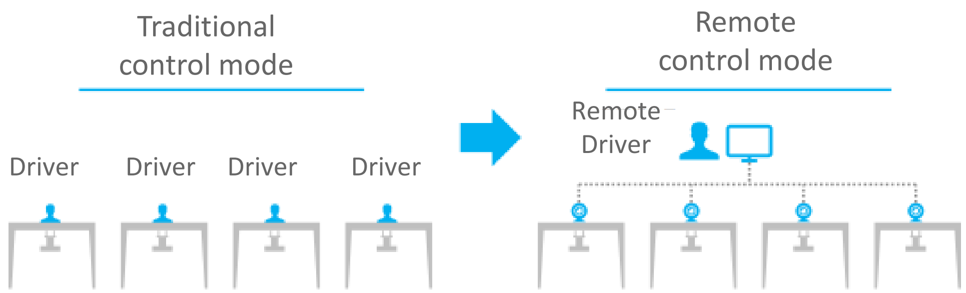
5QI Value	Resource Type	Default Priority Level	Packet Delay Budget (NOTE 3)	Packet Error Rate	Default Maximum Data Burst Volume (NOTE 2)	Default Averaging Window	Example Services
85		21	5 ms (NOTE 5)	10^{-5}	255 bytes	2000 ms	Electricity Distribution- high voltage (see TS 22.261 [2]). V2X messages (Remote Driving. See TS 22.186 [111], NOTE 16)

https://www.gsma.com/futurenetworks/wp-content/uploads/2020/02/5_Smart-Grid-Powered-by-5G-SA-based-Network-Slicing_GSMA.pdf

Smart port over small regional network



“An automated ship-to-shore (STS) crane that was operated via a 5G link to the control center and used to lift containers.”



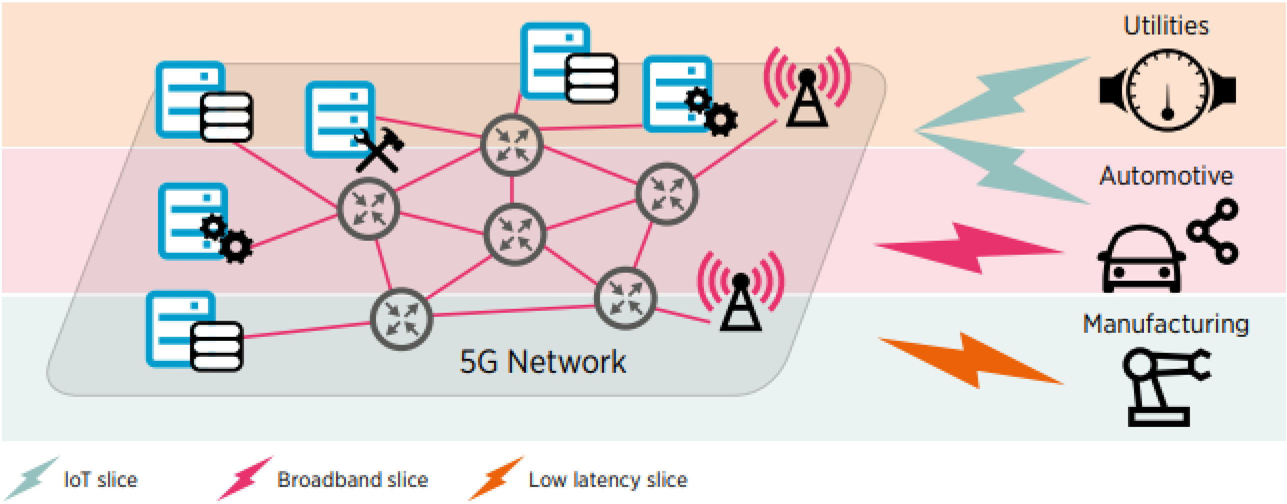
- To improve shipping efficiency and safety, video surveillance and AI detection are used to help controllers.
- Onsite remote controlling requires strict latency bound (30ms)
- Requirements
  - ➢ High Reliability
  - ➢ Bounded latency

https://www.maritime-executive.com/article/5g-smart-port-system-trialed-at-qingdao

Network slicing over carrier networks

- 5G-ACIA documents describe network slicing in factories, in order to support multiple applications in converged network.
- IETF discussion about slicing architecture over all possible transport networks. https://tools.ietf.org/html/draft-ietf-teas-enhanced-vpn-05
- Network slicing relates to differentiate SLA guarantee in a converged network.

5G networks subdivided into virtual networks each optimised for one business case



Service provider network characteristics:

1. Security
2. Scalability
3. Maintainability
4. Reliability
5. SLA guarantee capability
6. ..

https://www.gsma.com/futurenetworks/wp-content/uploads/2017/11/GSMA-An-Introduction-to-Network-Slicing.pdf

Multiple approaches to provide SLA guarantee

- Multiple possible approaches on different network layers

OSI Layer	Data unit	SLA guarantee Methods
Network	Packets	VPN/ DetNet over TSN_sublayer/Others
Data Link	Frames	TSN / Qos
Physical	Bits	FlexE/ Optical Transport Network
Medium	Wavelength/links	Dedicate links/ Wavelength

To achieve

Network KPI:

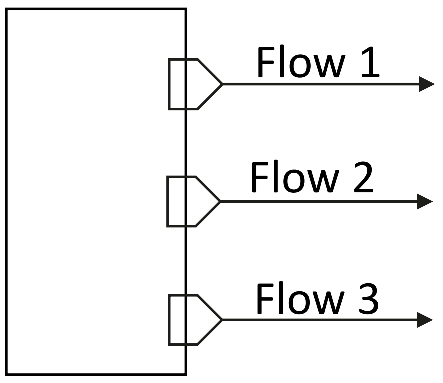
- Bounded latency;
- Bounded jitter;
- Reliability;
- Packet loss ratio;

- Lower layer methods provide more strict resource separation, less chance of interference from other users;
  - Advantages： Security / easier for OAM / ..

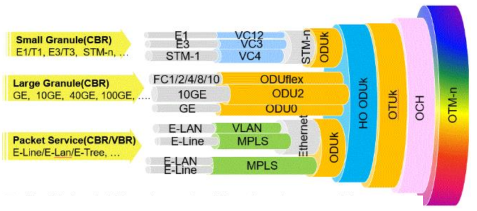
- Higher layer methods provide more statistical multiplexing capability, more cost efficiency for best effort traffics;
  - Advantages：Cost efficiency / flexibility / scalability / ..

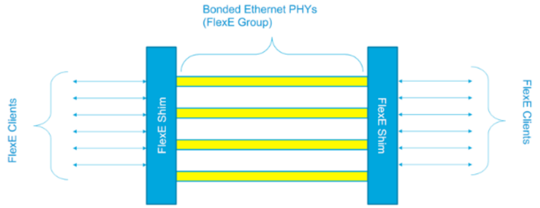
Queueing and forwarding plane for bounded latency

Examples on latency guarantee approaches



Separate physical links / ports





Separate frame encapsulation/ like in OTN , FlexE

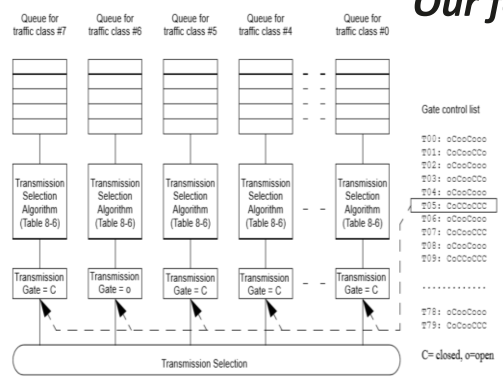
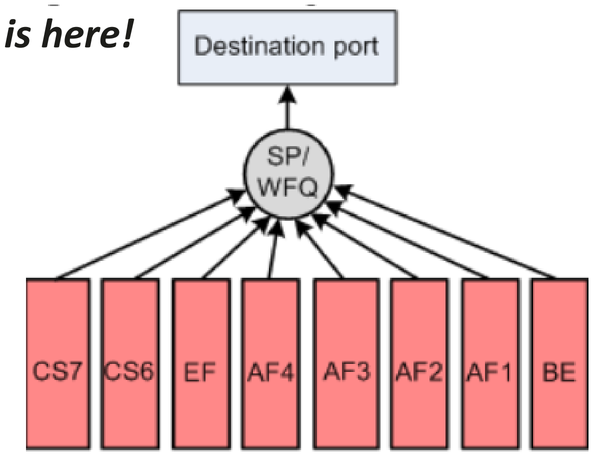


Figure 8-14—Transmission selection with gates

802.1Qbv Scheduled traffic

Our focus is here!



QoS: Strict Priority/Weight Fair Queueing

Layer 0 resource isolation	***Layer 1 resource isolation***	***Layer 2+ resource isolation on shared networks***

- Different Qos/TSN schedulers and shapers, either synchronized or unsynchronized , provide differentiated service levels (SLA) on shared network resources;
- This analysis leads to network slicing concepts, to divide network and share among users/applications;
- Multiple queueing and forwarding techniques are capable to support network slicing with multiple levels of service guarantee;
- These approaches can be combinational solution in service provider networks to satisfy specific requirements and constraints;

TSN for Service Provider Networks discussions

Most interested in,

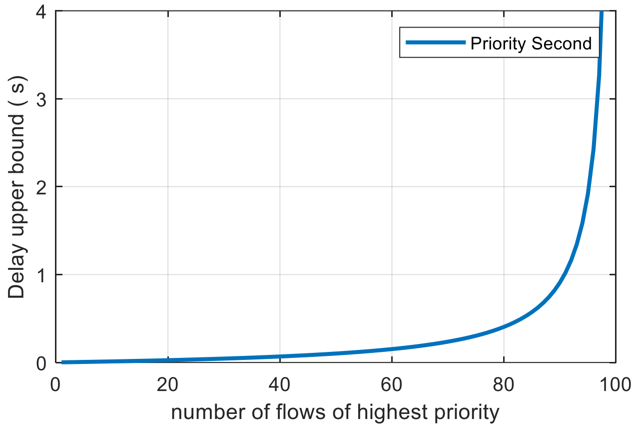
- How TSN techniques in carrier networks can help fulfill industrial internet requirements
- Suggestions and Comparisons of TSN techniques on
  - latency/jitter
  - reliability
  - Scalability
  - Others
- Recommendations for
  - bounded latency
  - Jitter
  - High Reliability
  - ..

Thank you.

Latency analysis – strict priority

Based on Network calculus methodology,
Setup arrival model as: α (t) = burst_size +rate * t;

With Strict priority scheduler, theoretically high priority traffic suffer no interference from low priority traffics. Consider only same class traffic multiplexing as blind multiplexing. Service curve can be modeled as:
$\beta(t) = (C - \Sigma FlowRate_{other})^* \; (t - (\Sigma b + L_0)/(C - \Sigma FlowRate_{other}))$;

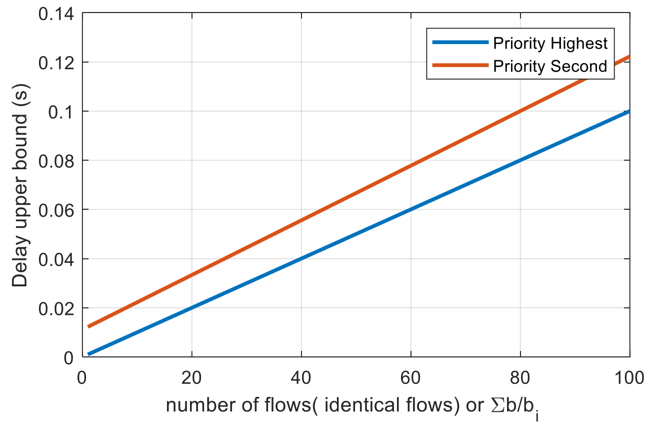


Observations on bounded latency provided by strict priority:

- Low bounded latency is achievable when high priority traffic is constrained in low utilization .
- With increasing utilization of high priority traffics, latency bound deteriorates quickly.
- Recommendation: Strict priority scheduling fits well for bounded low latency applications with low utilization. e.g. smart grid tele-protection application

If high priority traffic is constrained, say 10%, second priority class get bounded latency:

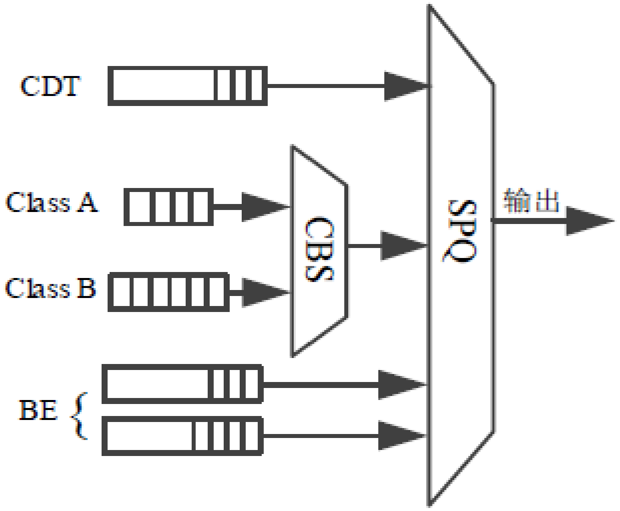
$$\beta_k^{SP}(t) = \left(C - \sum_{j=k+1}^{N} \sum_i \rho_{i,j}\right) * \left[t - \frac{\sum_{j=k+1}^{N} \sum_i \sigma_{i,j} + \max(L_{max,i,j})}{C - \sum_{j=k+1}^{N} \sum_i \rho_{i,j}}\right]^+$$



When bandwidth for a queue is guaranteed, its queueing latency increasing linearly with aggregated burst size.
Recommendation: constraints on aggregation scale and burst size;

http://www.ieee802.org/1/files/public/docs2020/dd-grigorjew-strict-priority-latency-0320-v02.pdf,
Discussion on achieve bounded latency with edge shaping and simple priority scheduling

Latency analysis – Credit based shaper



Credit based shaper algorithm defined in Std 802.1 Qav, combines strict priority scheduling and shapers on SR class A/B;

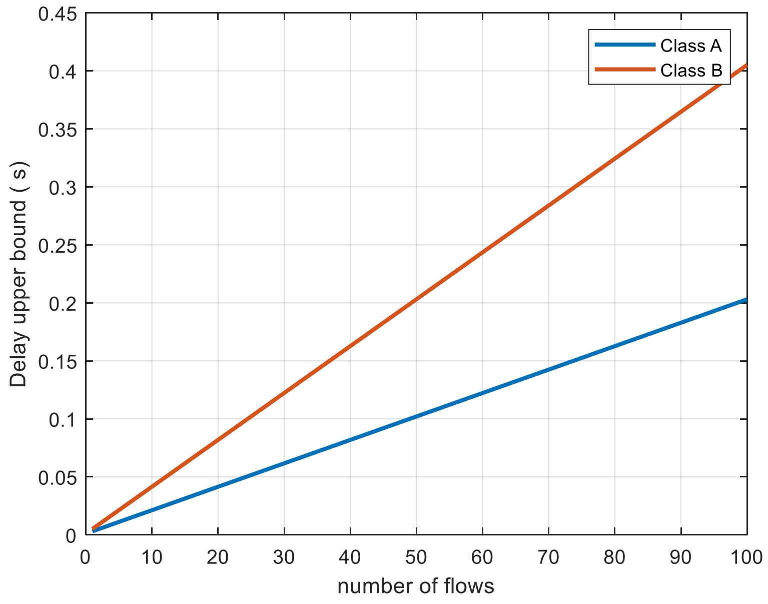
- Priority: Class CDT > SR Class A > SR Class B ;
- Parameters $idleslope_{a/b}$, $sendslope_{a/b}$, while (idleslope – sendslope) = C;
- $I_a + I_b < 75\% * C$;

SR Class A $$\beta_A(t) = R_A \times \left(t - \frac{L_0 + \sum b_{CDT} + \sum \rho_{CDT} \frac{\max(L_A, L_B, L_0)}{C}}{C}\right)$$

$$R_A = I_A \frac{C - \sum \rho_{CDT}}{C}。$$

SR Class B $$\beta_B(t) = R_B \times \left(t - \frac{L_0 + L_A + \sum b_{CDT} - L_0 \frac{I_A}{I_A - C} + \sum \rho_{CDT} \frac{\max(L_A, L_B, L_0)}{C}}{C}\right)$$

$$R_B = I_B \frac{C - \sum \rho_{CDT}}{C}。$$



If burst size and data rate is constrained ,
latency on SR Class A/B increase with total burst size;

Mohammadpour E, Stai E, Mohiuddin M, et al. Latency and backlog bounds in time-sensitive networking with credit based shapers and asynchronous traffic shaping //2018 30th International Teletraffic Congress (ITC 30)

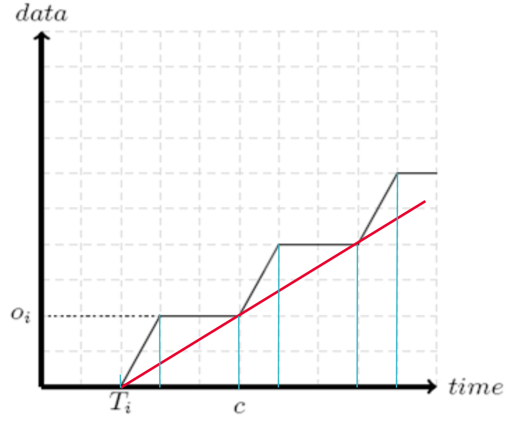
Latency analysis – Time Aware Shaper/TDM similar

$\beta(t) = \frac{C}{n}[t - (n-1)(T_c + Gb)]^+$

- Assume time windows of same width Tc;
- Guardband $Gb = \frac{L}{C}$;
- n time windows open in rotation;

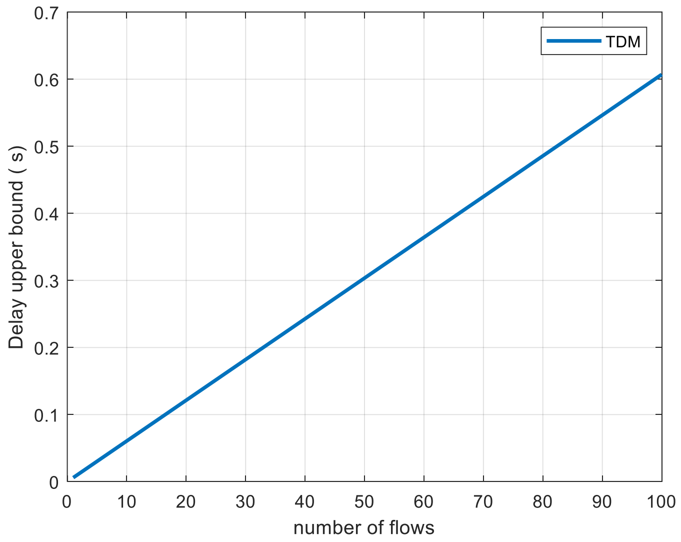
Worst Case Delay = (n-1)(T_C + Gb) + b*n/C;

- Worst case delay increases with number of time windows;
- TAS/TDM methods have larger lower bound, since it can not share idle bandwidth among competing traffics.
- TAS/TDM provide smaller jitter, given reasonable configurations.



(c) TDMA service curve ($\upsilon_{c,o_i,T_i} \otimes \beta(t)$)

Figure 3.1: Illustration for TDMA



Summary on latency analysis

No universal method fits all delay critical scenarios.

For traffic type of token bucket，$\alpha(t) = b + rt$; generally，

- Strict Priority fits for low bounded latency and low bandwidth applications -- 5G smart grid tele-protection is good example;
- Weighted Round Robin fits for bounded latency and bandwidth guaranteed applications.
- TAS/TDM fits for bandwidth guaranteed and low bounded jitter applications; //considering CQF variants.

For traffic type of periodic traffics，$\alpha(t)$ = TSPEC;

- With global time sync, TAS/TDM fits well for low bounded latency and low jitter；roughly, Worst Case Delay = $T_C + Gb$;
- Coordination of Network cycle and Application cycle in Industrial automation environment probably will be a good example, at cost of time synchronization.

Further analysis

- From single hop delay to path delay.
- Consider examples with combinational approaches.
    - E.g. Weighted Round Robin + Strict Priority
- Compare from multiple performance perspectives, not only on latency analysis. Consider more on jitter, packet loss, reliability;
- Discussion on example use cases and requirement.
- Other suggestions?