Deadline Queueing Mechanism

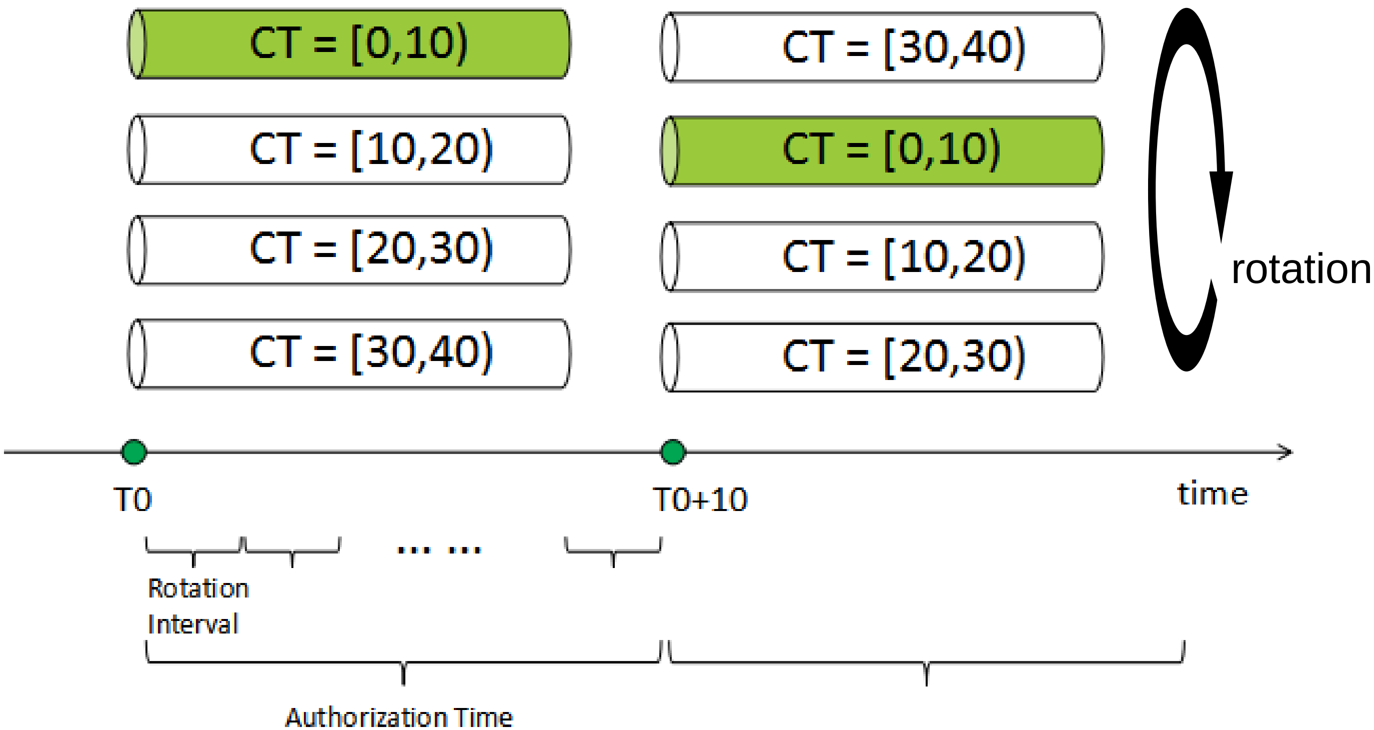
Shaofu Peng ZTE
Peng Liu China Mobile
Dong Yang BJTU

What is Deadline Queueing Mechanism ?

- Deadline Queueing Mechanism is a variants of EDF (Earliest Deadline Forwarding) scheduling.
    - An EDF scheduler, which always selects the packet with the shortest deadline for transmission, is an **optimal scheduler** for a bounded delay service in the sense that it can support the delay bounds for any set of connections that can be supported by some other scheduling method [1].
    - However, a **disadvantage** of EDF scheduler is that queued packets must be sorted according to their deadlines, requiring a search operation whenever a new packet arrvies at the scheduler.
- Here, the meaning of “variants” is:
    - using FIFO queue to replace LIST for ease of hardware implementation [2].
    - without requiring queued packets to be sorted strictly.
        - instead, multiple queues are sorted by priority and dynamically rotate the priority of each queue.
    - introducing non-work-conserving behavior, for jitter-sensitive service, besides the existing work-conserving behavior of EDF.
    - using latency compensation to replace re-shaper inside the network.

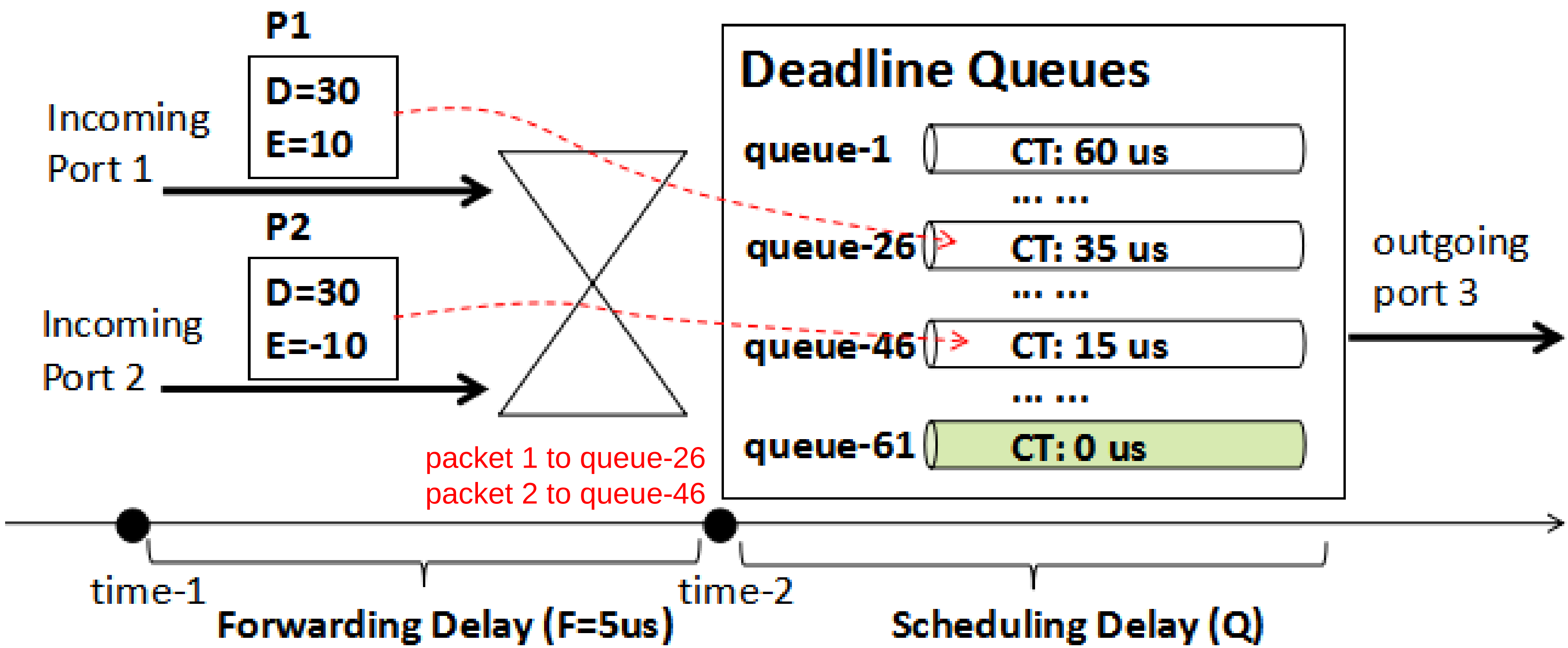
[1] C. L. Liu and J. W. Layland, “Scheduling algorithms for multiprogramming in a hard real time environment,” J. ACM, vol. 20, no. 1, pp.46-61, Jan. 1973.
[2] J. Liebeherr, D. E. Wrege and D. Ferrari, "Exact admission control for networks with a bounded delay service", IEEE/ACM Trans. Netw., vol. 4, no. 6, pp. 885-901, Dec. 1996.

Overview of Deadline Queueing Mechanism



- Each deadline queue has **CT** (Count-down Time) that is decreased by **TI** (rotation interval), and **AT** (Authorization Time) that is for sending duration.
  - AT = N * TI, where N >= 1.
  - TI can equal to AT, but it is usually a smaller value to sensitively perceive service differences.
  - Based on the maximum deadline that a node can provide, the larger the AT value, the fewer queues are required.

- A packet with Allowable Queueing Delay **(Q),** computed by Planned Residence Time **(D)** and Accumulated Residence Deviation **(E)**, will put to a deadline queue, meeting CT ≤ Q < CT+AT.
  - **Q = D + E - F**
    - where, D = (E2E_delay - total_link_delay) / hops
    - D also termed as delay level.
  - Note that if AT is close to 0, then it is exactly the traditional EDF.

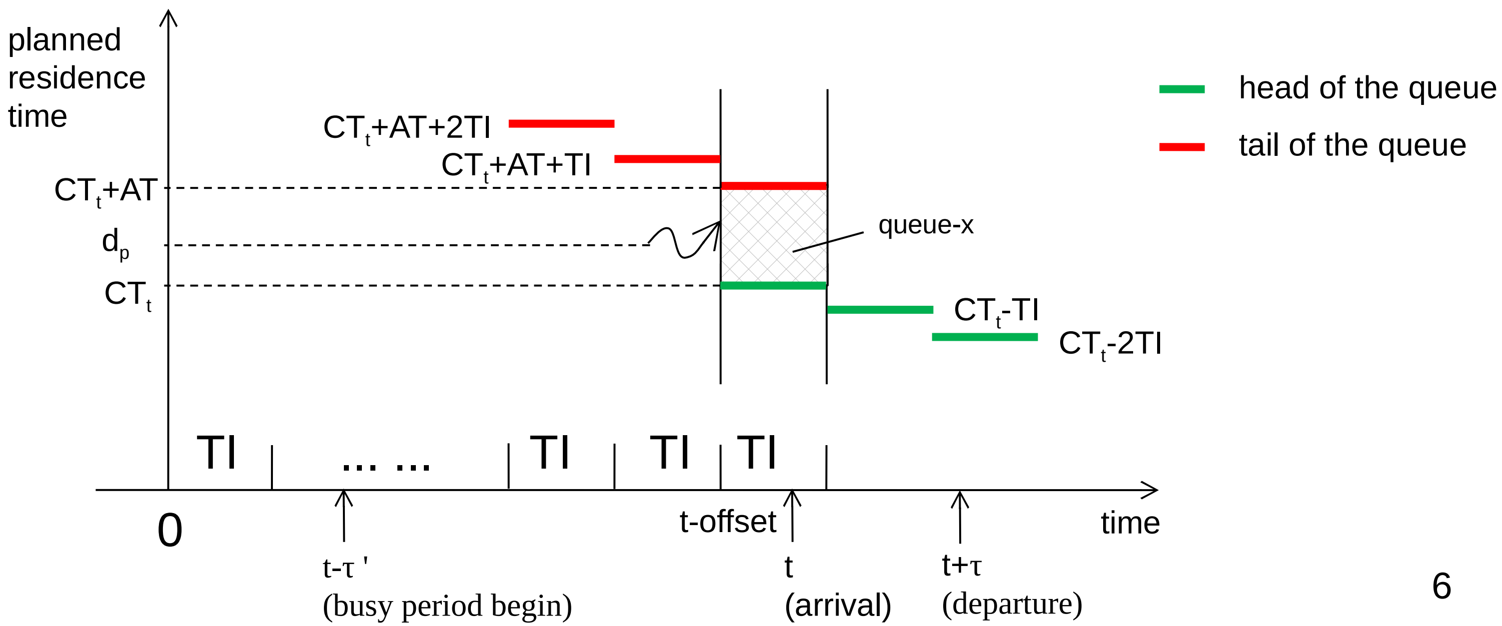


- There are two types of deadline schedulers, each with isolated queues, isolated service rate.
  - In-time scheduler
    - Work-conserving behavior, packets is sent ASAP before their deadline.
    - If the CT value of the queue becomes 0, it has highest priority and imediately get sending authorization.
    - If the CT value of the queue is not 0, it can also get the opportunity to send packets as long as queues with higher priority are empty.
  - On-time scheduler
    - Non-work-conserving behavior, packets is held until their deadline.
      - Due to the time required for packets sending, it **should pay attention to the extent of exceeding the deadline**.
      - Thus strict on-time & loose on-time are further introduced.
    - If the CT value of the queue becomes 0, it has highest priority and imediately get sending authorization.
    - If the CT value of the queue is not 0, it can NOT request sending authorization.

How to Guarantee Delay ?

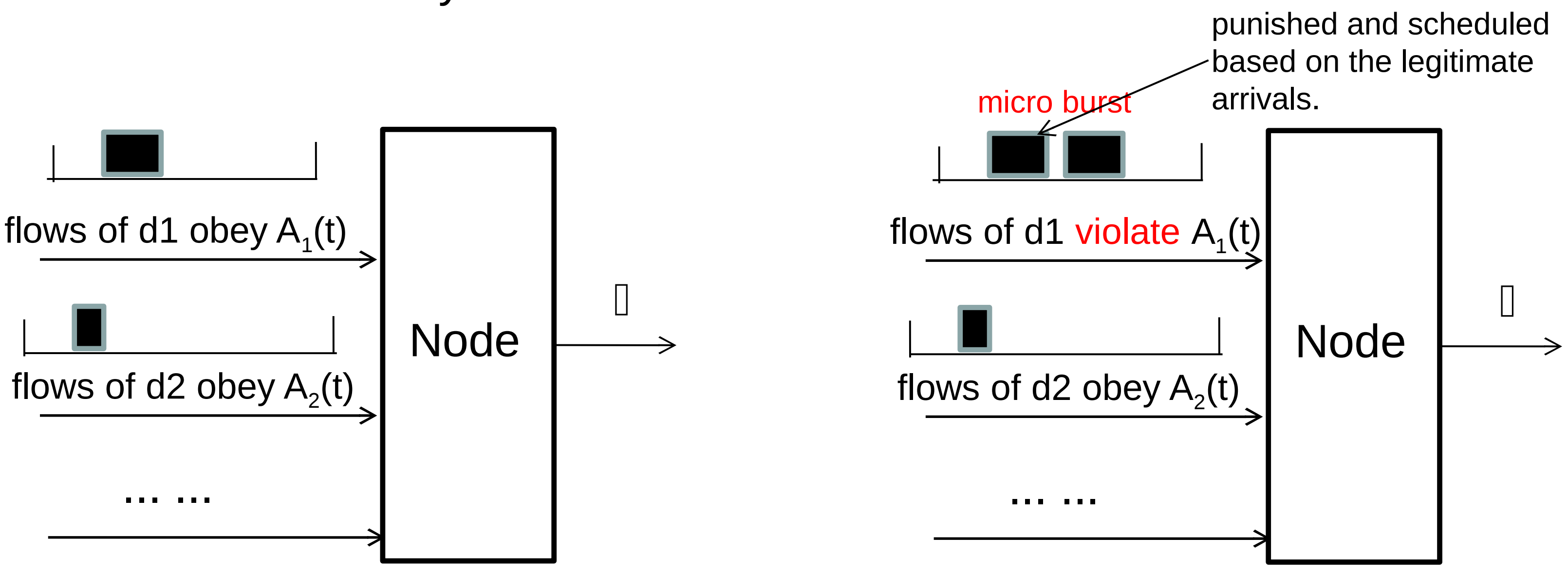
- Schedulability Condition
  - Observe a specific node, suppose that the deadline scheduler of the outgoing port has a service rate C, which can be less than or equal to the port bandwidth. For all concurrent arrvied flows, with arrival constraint function $A_i(t)$ and planned residence time d_i for flow i, as long as the following condition is meet, the scheduler can successfully send all packets before their deadline:

$$C * t \geq A_1(t - d_1) + \sum_{i=2}^{N} A_i(t + AT - d_i) + M$$



- proof (with work-conserving behavior)
  - Suppose that the observed packet, with planned residence time $\mathbf{d_p}$, arrives at the scheduler at time $\mathbf{t}$ and leaves the scheduler at time $\mathbf{t+\tau}$. It will be inserted to the physical queue-x with CT_t, meeting CT_t <= $\mathbf{d_p}$ < CT_t+AT, at the current rotation interval TI with starting time **t-ofst** and end time **t-ofst+TI**.
  - Suppose that $\mathbf{t-\tau'}$ is the beginning of the busy period closest to $\mathbf{t}$. Then, the amount of packets within interval [$\mathbf{t-\tau'}$,$\mathbf{t+\tau}$] that must be scheduled before the observed packet include the following four components:
    - □ $\sum A_i[\mathbf{t-\tau'}, t]$, for all flow i with CT_t ≤ $\mathbf{d_i}$ < CT_t+AT.
    - □ $\sum A_i[\mathbf{t-\tau'}, t\text{-}ofst+TI-(d_i-CT\_t-AT+TI)]$, for all flow i with $\mathbf{d_i}$ ≥ CT_t+AT.
    - □ $\sum A_i[\mathbf{t-\tau'}, t+CT\_t-d_i]$, for all flow i with $\mathbf{d_i}$ < CT_t
    - □ deduct the traffic that has been sent during the busy period. i.e., C*($\mathbf{\tau+\tau'}$)
  - The workload is □+□+□-□, which is less than $A_1(\mathbf{\tau'+d_p}-d_1) + \sum_{i\geq 2} A_i(\mathbf{\tau'+d_p}+AT-d_i) - C^*(\mathbf{\tau'+d_p})$.
    - Let it further less than zero, and denote $\mathbf{\tau'+d_p}$ as t, then get the condition.

- Therefore, it is important that the actual arrival traffic of each delay level (d_i) on the node must not exceed its constraint function, otherwise, it should:
  - re-shape traffic to obey the constraint function.
  - or, reward or punish the priority of the traffic by latency compensation, to identify the legitimate emergency traffic (**suggested**).
- The value of the parameters of the constraint function of each delay level construct the delay resource.



(a) legitimate arrival

(b) ilegitimate arrival

Conditions for Leaky Bucket Constraint Function

- For the case that n types of planned residence delay levels (d_1, d_2,..., d_n) is supported, and each level d_i has the leaky bucket arrival curve $A_i(t) = b_i + r_i * t$, we have the following conditions:

$$b_1 \le C*d_1 - M$$

$$b_1+b_2 + r_1*(d_2-d_1) + r_2*AT \le C*d_2 - M$$

$$b_1+b_2+b_3 + r_1*(d_3-d_1)+r_2*(d_3-d_2) + (r_2+r_3)*AT \le C*d_3 - M$$

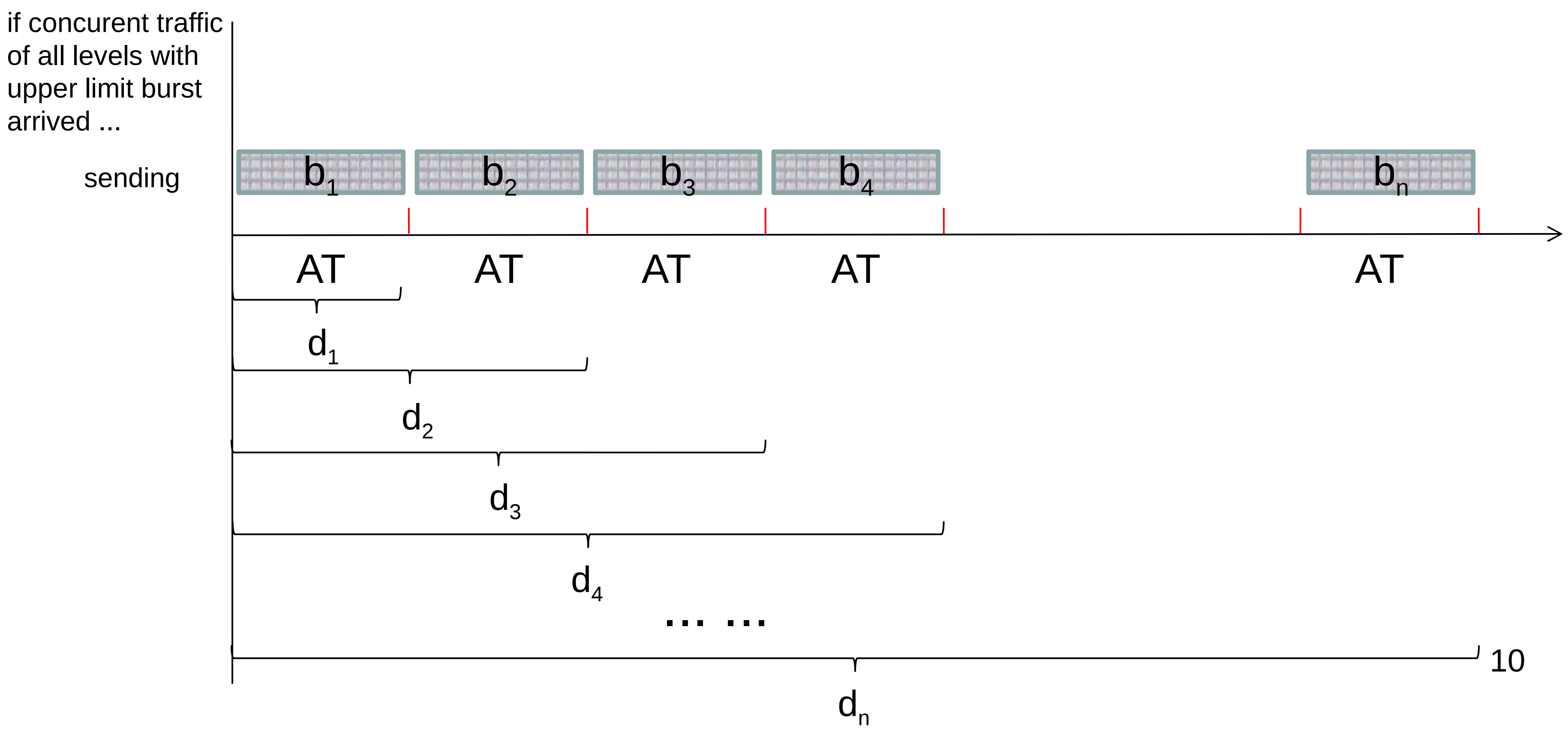
... ...

$$\sum b_i + r_1*(d_n-d_1)+r_2*(d_n-d_2)+...+ r_{n-1}*(d_n-d_{n-1}) + (r_2+...+r_n)*AT \le C*d_n - M$$

- For each level d_i of the link, the parameters (b_i, r_i) is called its delay resource pool that can be reserved by the service.
  - the delay resource is the enhancement of the traditional bandwidth resource, with additional burst resources.

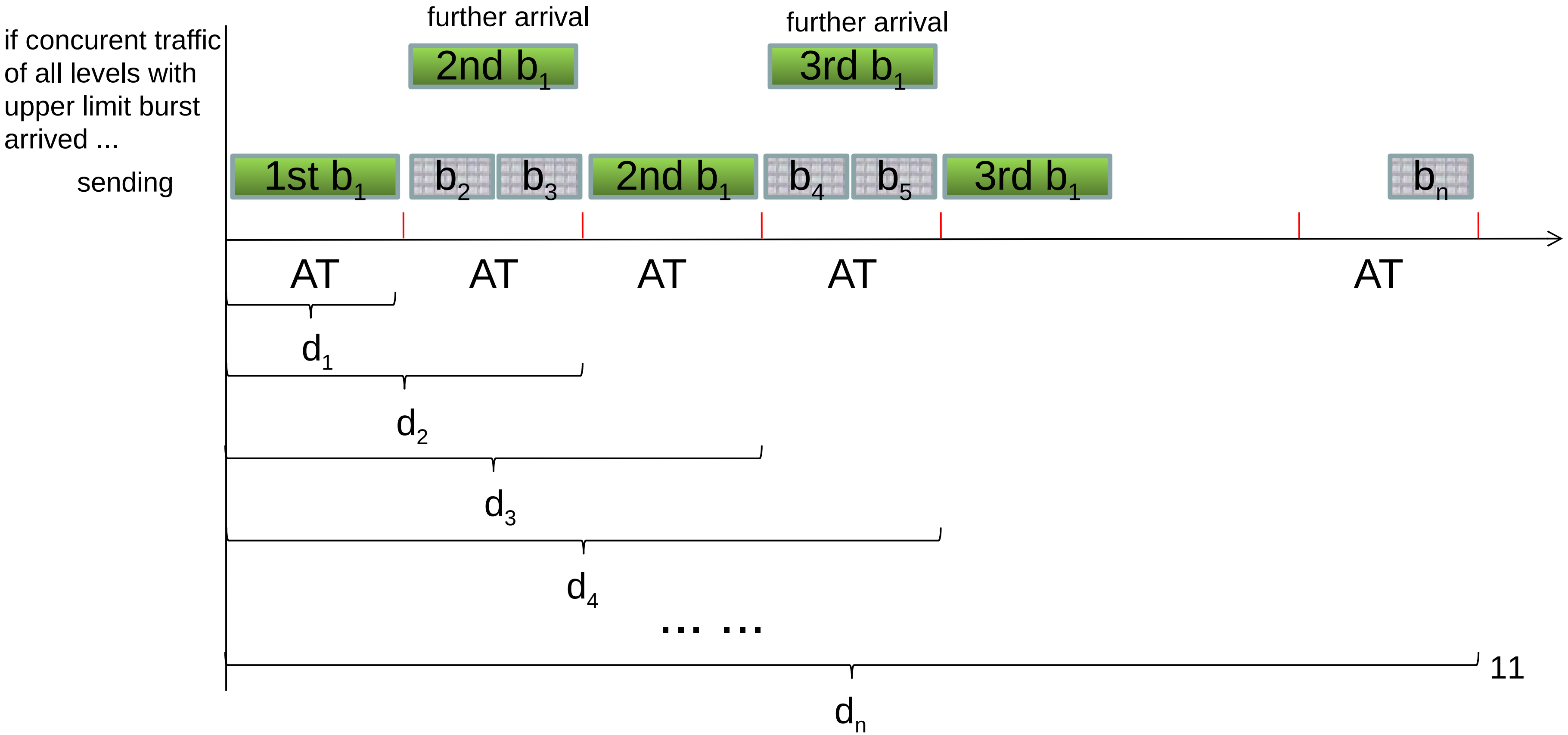
- Example 1

  If the bandwidth of all levels is approximately 0 (due to the large burst interval of the services they serve), each delay level can provide burst resource with full value, i.e., C*AT - M.

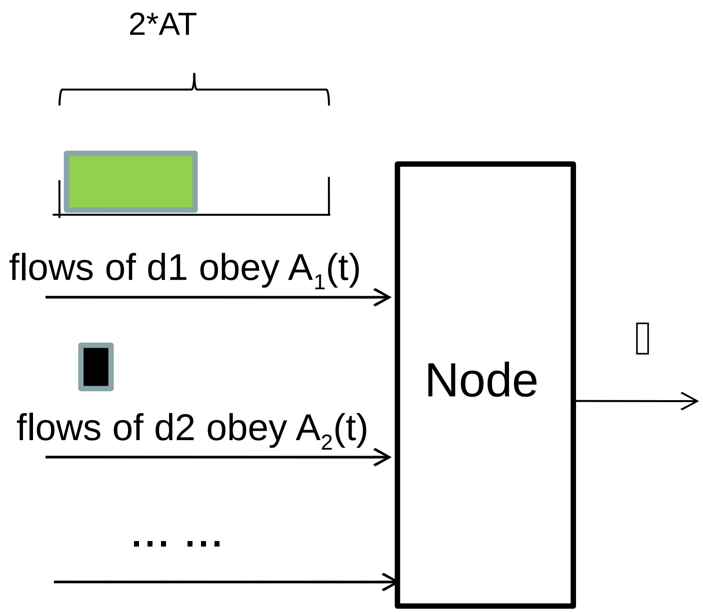


- Example 2

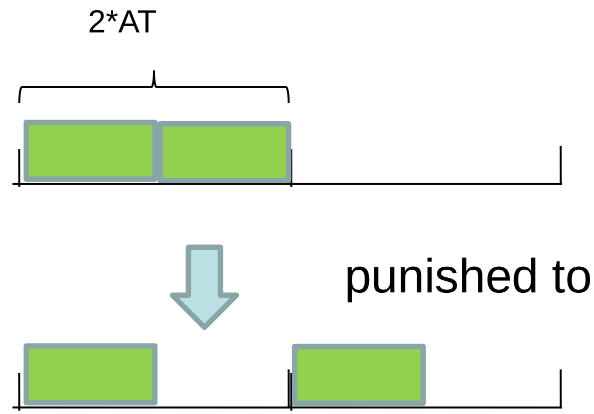
  If the delay level d1 has burst resource with full value, and bandwidth resource C/2 (i.e., burst with full value per 2*AT), all other delay levels have the approximate bandwidth 0, then all other levels may have burst resource with half of the full value:



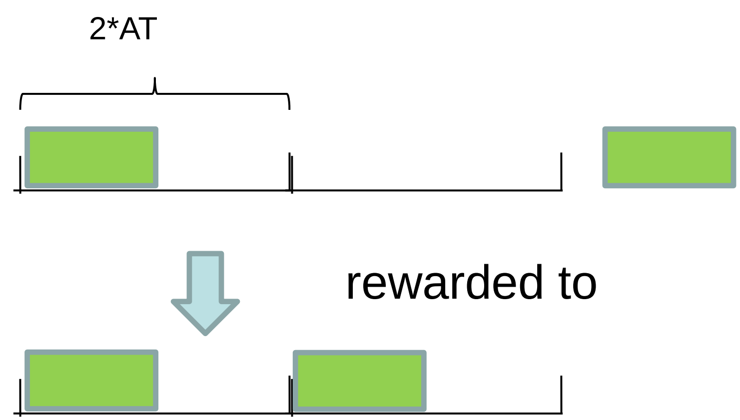
- arrival traffic of Example 2



(a) legitimate arrival



(b) ilegitimate arrival



(c) late arrival

Note:

- The legitimate arrival for each level contains the concurrent aggregate traffic for that level.
- For in-time case, (c) does not occur in reality, only exists in theory.

Conditions for On-time Mode

- The on-time mode does not cause the arrival curve to exceed the expected traffic constraint function, however, it is non-work-conserving and waste the opportunity to send packets, cause that some packets may exceed their deadline in the extreme case, e.g, each concurrent b_i is full value (i.e., C*AT - M) .
  - suggest in-time mode for low delay service.
  - suggest in-time or loose on-time mode applied on the transit nodes, and strict on-time on the egress, for low delay jitter service.
    - queueing rule of loose on-time: CT ≤ Q-k*AT < CT+AT, k ≥ 1

	in-time	**loose on-time**	**strict on-time**
when sum(all b_i) equals to C*AT - M (traffic well-distributed)	before deadline	before deadline	near deadline, ±AT
	before deadline	before deadline	partially exceed deadline
when each b_i equals to C*AT - M	before deadline	partially exceed deadline	exceed deadline, AT~d_n

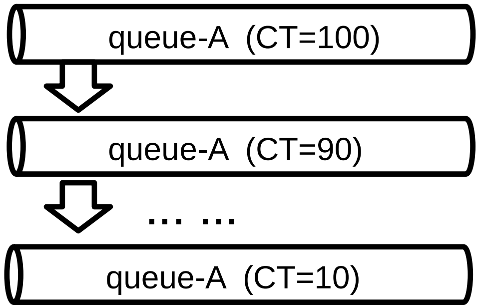
Buffer size design

- Each deadline queue is not bound to a fixed delay level (d_i), and it will actually store all levels of traffic during its CT decrement process.

  e.g:
  - At T0, d_{100} traffic arrived and inserted to queue-A with CT = 100
  - At T0+10us, d_{90} traffic arrived and inserted to the same queue-A with CT = 90

    ... ...
  - At T0+90us, d_{10} traffic arrived and inserted to the same queue-A with CT = 10

  That is, each level (d_i) of traffic arrived is r_i * AT, where r_i is the averaged bandwidth of level d_i, AT is 10 us.



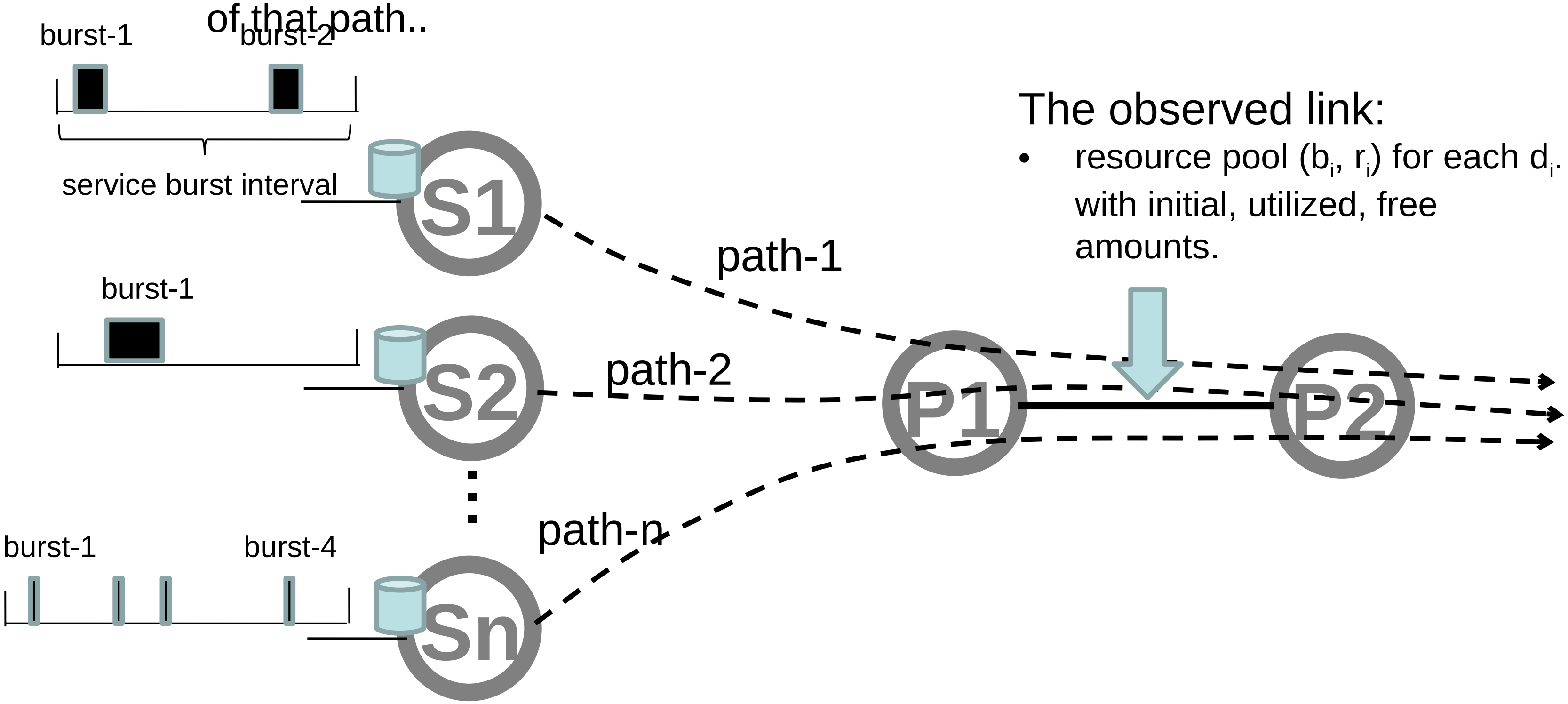
- Considering the stability condition $\sum r_i \leq C$, then the buffer size is designed to C*AT -M, where r_i is the bandwidth resource of level d_i, C is the service rate of the deadline scheduler, M is the maximum interference packet size.

  The burst resource of any level must be less than the full-value, i.e., C*AT - M.

  The burst resource of each level will be more small if the bandwidth of other higher priority level can not be negligible.

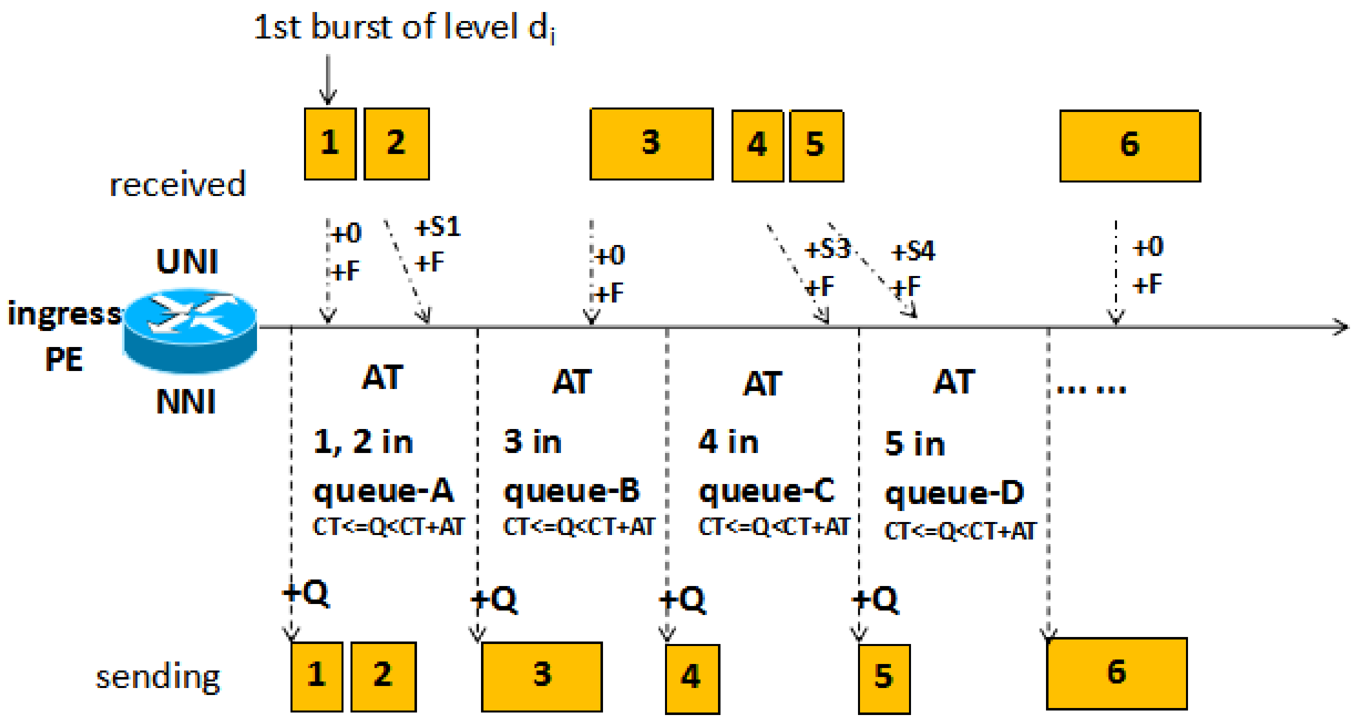
Admission Control on the Ingress

- Traffic regulation on the imcoming port of the ingress node.
  - Leaky bucket depth(CBS) is set to cover the reserved burst resource.
  - Leaky bucket rate(CIR) is set to cover the reserved bandwidth resource.
- All path j through the link allocate delay level d_i resources, $\sum(b_{ij}, r_{ij})$, is less than the delay resource pool (b_i, r_i) of that link's delay level d_i.
  - all service k of delay level d_i over path j contribute the reserved resources (b_{ij}, r_{ij}) of that path..



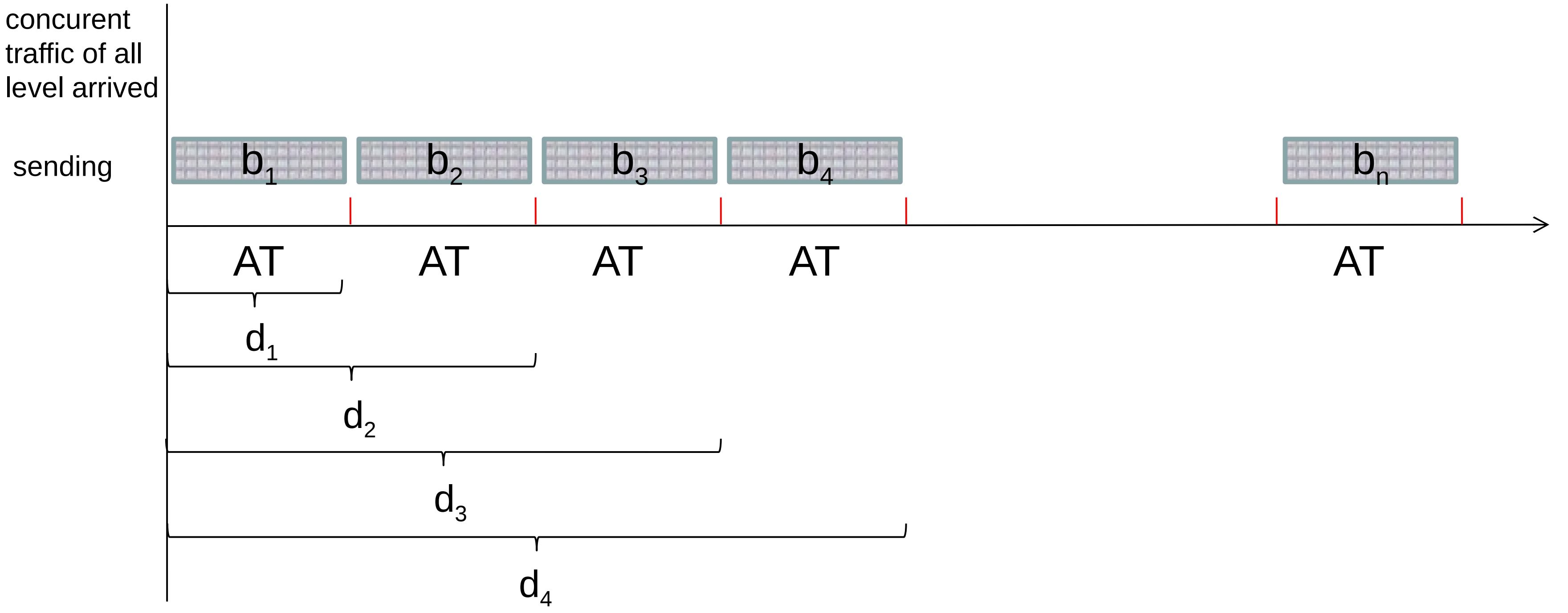
- The leaky bucket shaping essentially make all bursts within the service burst interval evenly distributed within that interval, which may be inconsistent with the original arrival curve of the service flow.
  - Some packets may face more shaping delay than others, which cannot be included in the latency compensation formula and need to be addressed.

admision control example:



Overprovision Analysis

- According to the schedulability condition, each delay level d_i has its own resources pool (b_i, r_i).
  - In the extreme case, each b_i can be the full value, i.e., C*AT - M, the scheduling procedure maybe like:



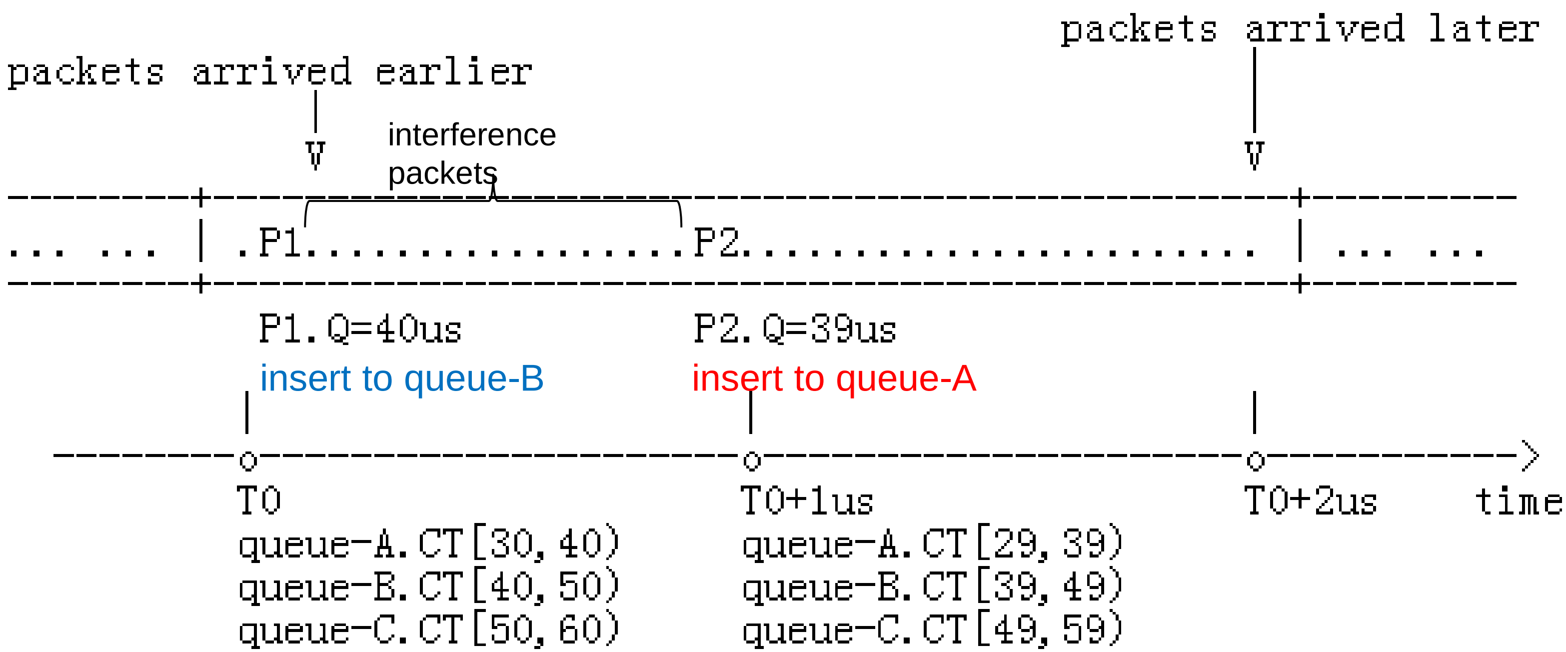
  - However, each level does not require the overprovision bandwidth b_i / AT. The bandwidth resource of each level is separate from the burst resource.
    - The requirece bandwidth of service is still according to the burst size per burst interval.

Summary of How to Guarantee Delay

- Partition delay resource for each delay level on the outgoing port according to the schedulability condition.
  - i.e., preset parameters of the arrival constraint function.
- Execute delay resource reservation on each port of each path on the control plane.
- Execute admission control on the network entry, to let the admited traffic obey the constraint function.
- Execute latency compensation in the network, to identify the legitimate arrived traffic that still obey the constraint function.

Disorder by latency compensation

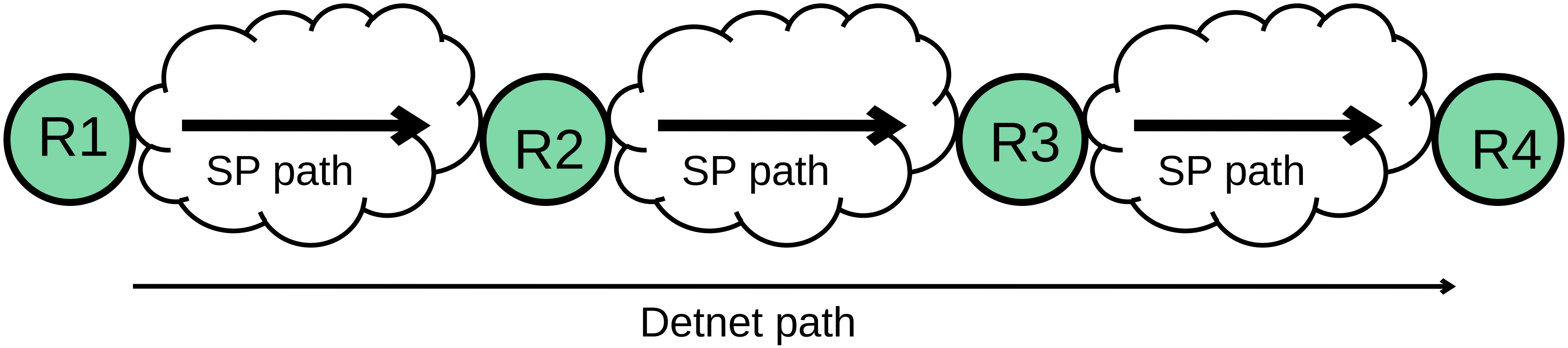
- The CT of the queue is not changed in real-time, but gradually with the decreasing step TI, it may cause two packets belonging to the same burst to be out of sequence.



- Addressing:
  - by POF(Packet Ordering Function)
  - by OGO(order guarantee object)

Partially Upgraded Scenarios

- A DetNet path consists of several upgraded border nodes (R1, R2, R3, R4) and traditional SP domain.
  - SP path should contain fewer hops, with the bounded latency per hop [1][2].
  - Unreasonable setting of DS field may get a large latency.
  - The legacy device in the SP domain may or not measure and insert the deadline information into packets.
- The border nodes decide whether to speed up or hold according to the residence deviation information of the packet.
  - If the traffic experiences less latency within the SP domain, on-time mode will work to achieve the end-to-end jitter target.



[1] A. Grigorjew et al., "Bounded Latency with Bridge-Local Stream Reservation and Strict Priority Queuing," 2020 11th International Conference on Network of the Future (NoF), Bordeaux, France, 2020, pp. 55-63
[2] Zhi-Li Zhang, Zhenhai Duan and Y. T. Hou, "Fundamental trade-offs in aggregate packet scheduling," Proceedings Ninth International Conference on Network Protocols. ICNP 2001, Riverside, CA, USA, 2001, pp. 129-137

Matching Evaluation of Requirements

- Checklist

Requirement items	Evaluation
3.1. Tolerate Time Asynchrony	No time synchronization.
3.2. Support Large Single-hop Propagation Latency	Not affected by link propagation dely.
3.3. Accommodate the Higher Link Speed	Each link sets AT independently according to its speed.
3.4. Be Scalable to The Large Number of Flows	No states per flow on the transit nodes. No overprovision issues.
3.5. Prevent Flow Fluctuation from Disrupting Service	Distinguish fluctuation flow by latency compensation.
3.6. Tolerate Failures of Links or Nodes and Topology Changes	No relationship with queueing mechanism...
3.7. Support Enhancement of Queuing Mechanisms	In-time mode for low latency, on-time mode for low jitter.
4.1. Support Aggregated Flow Identification	Defined delay level (d_1, d_2, ..., d_n)
4.2. Support Information used by Functions ensuring Deterministic Latency	Defined delay resource for each level, protocl extensions to advertise and reserve resource (TBD).
4.3. Support Redundancy Related Fields	No relationship with queueing mechanism...
4.4. Support Explicit Path Selection	No relationship with queueing mechanism...

Questions & Comments ?