Deadline based Forwarding

draft-peng-detnet-deadline-based-forwarding-07

Shaofu Peng	ZTE
Zongpeng Du	China Mobile
Kashinath Basu	Oxford Brookes University
Zuopin Cheng	New H3C
Dong Yang	Beijing Jiaotong University
Chang Liu	China Unicom

IETF-118 Nov 2023

Updates

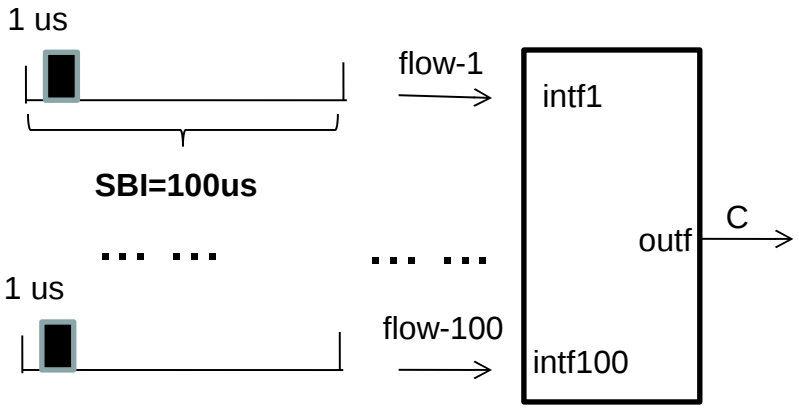
- Add more co-authors who are interested in this document.
- Recommend options (3 or 4) that are more suitable for large-scaling requirements.
- Give the relationship between D (planned residence time) and d (delay level), as well as SBI (service burst interval) and d.
- Supplement on-time scheduling mode, and the corresponding schedulability considerations.
- Give alternet QAR (queue allocation rule) for RPQ.
- Supplement the enqueue rule and dequeue rule for each option.
- Update the evaluation table.
- Supplement security considerations.

Motivations

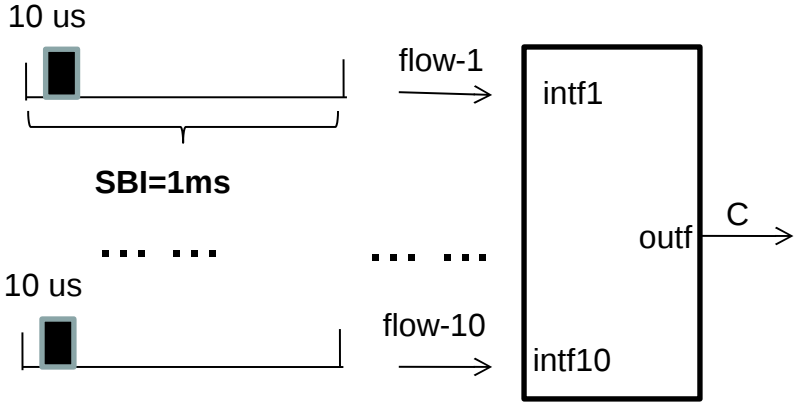
- Challenges of the existing queueing mechanisms:
  - TSN ATS/CBS come with a high latency variance, as the minimum latency is not affected by them. The worst-case latency is **overestimated**, basically inversely proportional to the service rate.
    - CBS can not even work independently, and should combine with re-shaping function (such as ATS) to avoid burstiness cascade.
  - TSN TAS requires time synchronization and has **scalability issues** on GCL calculation, update and installation.
  - TSN CQF requires time synchronization and relies on very small link delay. Although ECQF only requires frequence synchronization, but with **overprovision issues**.
  - The widely used priority based queuing scheme in IP/MPLS diff-serv network, may give better average latency, but with **worst case latency**.

- To meet the large scaling requirements, this document **introduce EDF (Earliest Deadline Forwarding) scheduling to DetNet Data Plane**, to uniformly provide bounded delay/jitter by in-time/on-time mode.
  - It belongs to TSN mechanism **category "Node-local SHAPING"**. More specifically, it is delay based scheduling described by [netcalc] book.

Examples to Understand Per-hop Latency



- Admit 100 flows, each with packet size C*1us per service burst interval 100us.
- Service rate C has been consumed by all flows.
- The worst-case per-hop latency is 100us in the case of eligible arrivals.



- Admit 10 flows, each with packet size C*10us per service burst interval 1ms.
- Service rate C has **NOT** been consumed by all flows, in order to still get the worst-case per-hop latency 100us.
- Of course, more flows can be admitted, but with larger per-hop latency (e.g, 1ms).

We learned:

- Per-hop latency is generally determined by the admitted eligible burst aggregation and the service rate C (i.e., **two resource types: burst, bandwidth**).
- An important aspect of queueing mechanism is to ensure (or pick out) eligible arrivals.

Per-hop Latency Provided by EDF

- EDF can preset multiple per-hop latencies (i.e., delay levels), then limit the arrival constraint functions of all delay levels to meet the schedulability condition.
  - The arrival constraint function can be generally represented by leaky bucket: $A_i(t) = b_i + r_i*t$, for delay level d_i.
  - (b_i, r_i) is the (burst, bandwidth) resource pool for delay level d_i and can be allocated by flows that belongs to this delay level.
- Schedulability condition:
  - For sorted queue, the traffic constraint funciton is:

    $\sum_{i\geq 1} A_i(t-d_i) <= C*t$ (equatation-1)
  - For rotation priority queue, the traffic constraint function is:

    $A_1(t-d_1) + \sum_{i\geq 2} A_i(t+AT-d_i) <= C*t$, if d_i contains single D. (equatation-2)

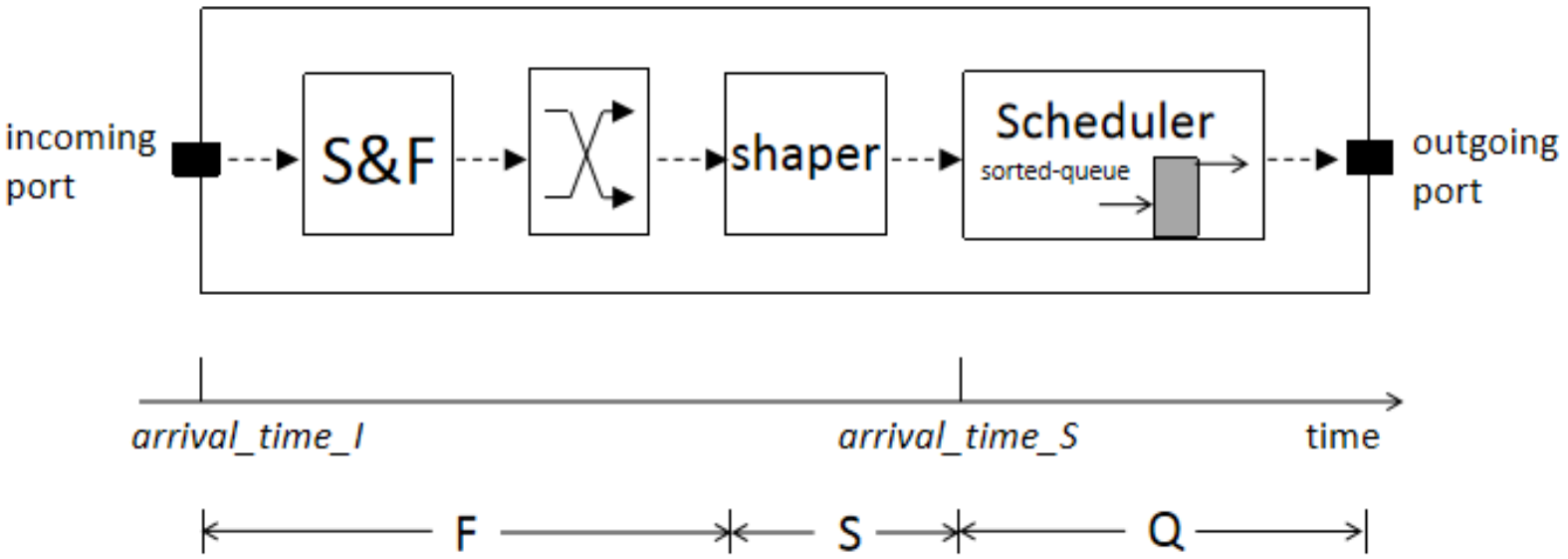
    $\sum_{i\geq 1} A_i(t+AT-d_i) <= C*t$, if d_i contains multiple D. (equatation-3)

    where A_i is the constraint function of delay level d_i, AT is the interval between adjacent delay levels, C is the service rate of the EDF scheduler, D is the planned residence time.

Option-1: Rate-controlled + Sorted-queue

- Rate-controlled: re-shaping (RFC2212) or Interleaved shaping (ATS).
- Sorted-queue: e.g, PIFO

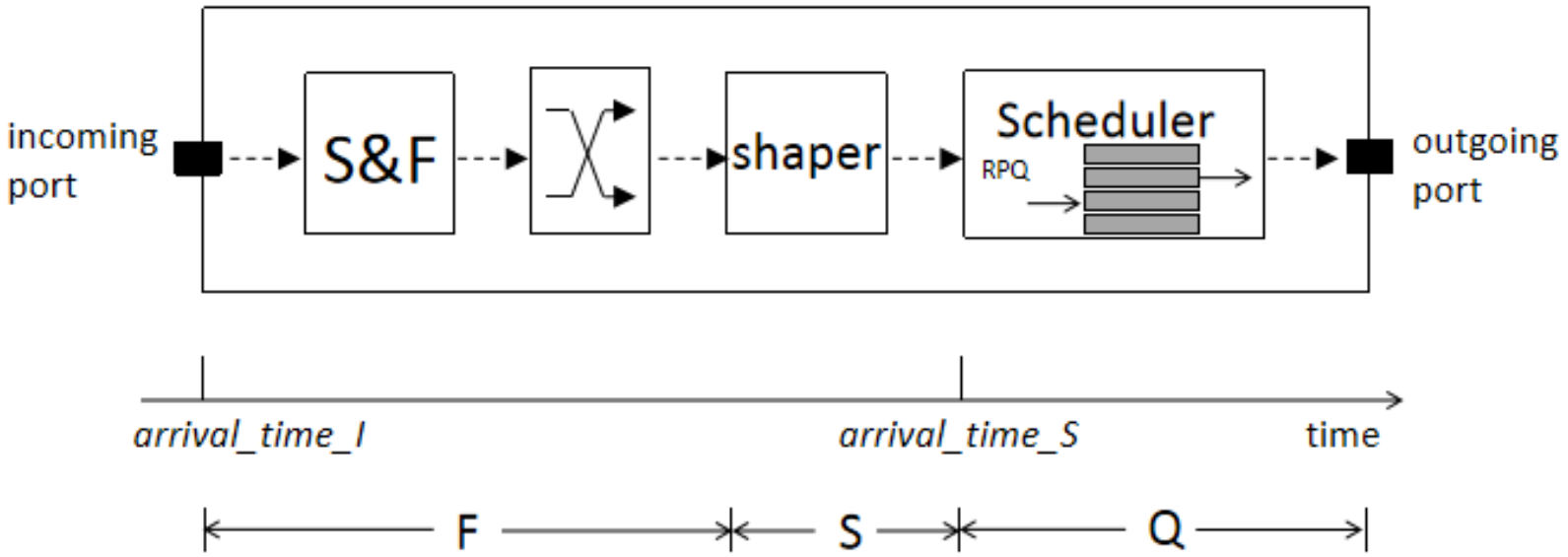
- Due to rate-controlled, all flows arrived at EDF scheduler are **eligibility**, the schedulability condition met.
  - Packet is put to the PIFO according to **arrival_time_I + D**, where arrival_time_I is the time arrived at the incoming port, D is the planned residence time.



Option-2: Rate-controlled + RPQ

- Rate-controlled: re-shaping (RFC2212) or Interleaved shaping (ATS).
- RPQ: rotation priority queues, each with count-down time (CT). The smaller the CT, the higher the priority.

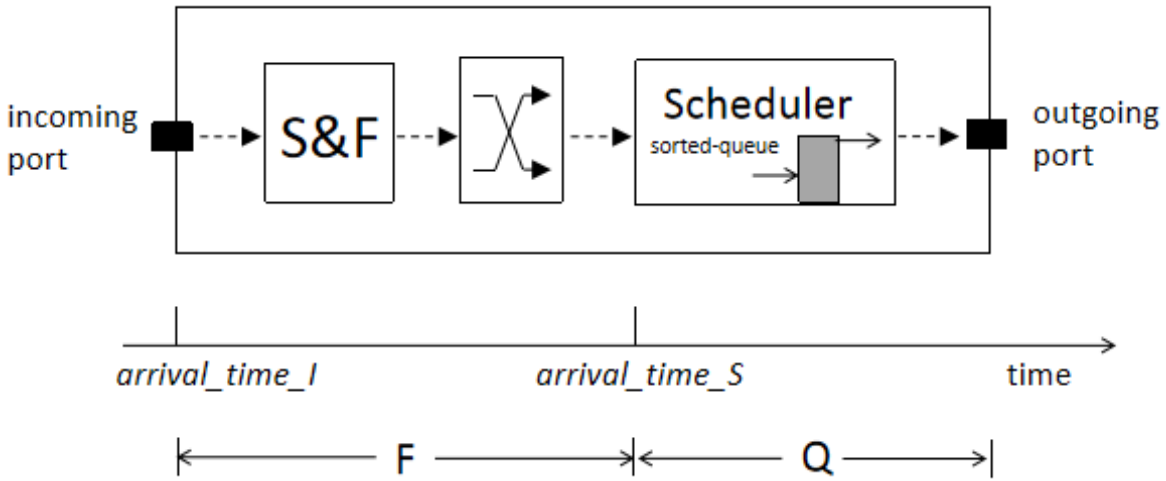
- Due to rate-controlled, all flows arrived at EDF scheduler are **eligibility**, the schedulability condition met.
  - Packet is put to the RPQ according to **CT ≤ Q < CT+AT**, where AT is the delay level interval, Q is the allowable queueing delay (Q = D-F).



Option-3: Latency Compensation + Sorted-queue (RECOMMONDED for core-stateless)

- Latency compensation is based on latency deviatoin (**E**).
- Sorted-queue: e.g, PIFO

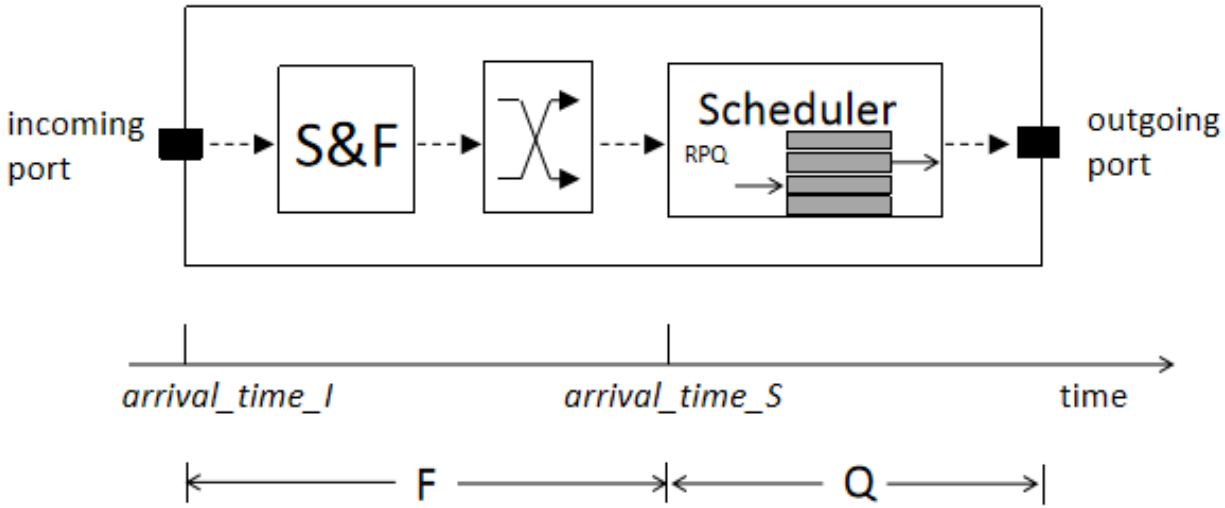
- Due to latency compensation, all flows arrived at EDF scheduler are distinguished as **eligibility** or ineligibility, eligibility arrivals always scheduled first.
  - Packet is put to the PIFO according to **arrival_time_I + D + E**, where arrival_time_I is the time arrived at the incoming port, D is the planned residence time, E is the latency deviation.



Option-4: Latency Compensation + RPQ
(RECOMMONDED for core-stateless)

- Latency compensation is based on latency deviatoin (**E**).
- RPQ: rotation priority queues, each with count-down time (CT). The smaller the CT, the higher the priority.

- Due to latency compensation, all flows arrived at EDF scheduler are distinguished as **eligibility** or ineligibility, eligibility arrivals always scheduled first.
  - Packet is put to the RPQ according to **CT ≤ Q < CT+AT**, where AT is the delay level interval, Q is the allowable queueing delay (Q = D+E-F).



On-time Scheduling Mode for Option 3/4

- The above option 3/4 described is work-conserving (in-time mode), may get E2E latency less than D*hops, but may have large jitter.
- On-time mode is based on planned residence time (**D**) and latency deviatoin (**E**), to get low jitter.
  - On-time mode may eliminate burst accumulation, making buffer design more simple.
  - On-time mode will not break the constraint function of flow.
- On-time mode can be further implemented in two methods:
  - 1) The packet is scheduled by D+E with on-time mode. In this case, the E2E latency may be in the range [D*hops, D*hops+d_i], that is, it may exceed packet's deadline but the exceeding valus is less than delay level value.
  - 2) The packet is scheduled by E with on-time mode firstly (scheduler-I), then scheduled by D with in-time mode (scheduler-II). In this case, the E2E latency may be in the range [D*hops-d_i, D*hops].
  - Both methods provide jitter of delay level value (d_i).

Alternet Queue Allocation Rule for RPQ

- It may further let a RPQ queue (act as the virutal parent queue) contain multiple sub-queues, each for a delay level.
  - The physical sub-queue with small delay level (e.g, 10us) is ranked before the physical sub-queue with large delay level (e.g, 20us).
  - So that for two packets with the same Q but different D, we can decide to firstly schedule the packet with the smallest D.
  - This is similar to PIFO using planned residence time (D) as tiebreaker when two packets have the same rank value.

- Why?
  - The reason is based on analysis of extreme scenarios that multiple delay levels of bursts arrive sequentially, with lower priority burst arriving first and higher priority burst arriving later, and then simultaneously releasing flood. In this case, it is necessary to ensure that the higher priority burst is sent first to meet its deadline.

Evaluation

Requirement items	Evaluation	Notes
3.1. Tolerate Time Asynchrony	Yes	No full time synchronization needed, only need frequency sync(3.1.3).
3.2. Support Large Single-hop Propagation Latency	Yes	The eligibility arrival of flows is independent with the link propagation delay.
3.3. Accommodate the Higher Link Speed	Partial	The higher service rate, the more burst resource may provided by each delay level, and more buffer space is needed. And, extra instructions to calculate E ....
3.4. Be Scalable to The Large Number of Flows and Tolerate High Utilization	Yes	Multiple delay levels, each with limited delay resources, can support lots of flows, without overprovision. Utilization may reach 100% link bandwidth. The unused bandwidth of the high delay level can be used by the low levels or best-effort flows.
3.5. Tolerate Failures of Links or Nodes and Topology Changes	N/A	No relationship with queueing mechanism...
3.6. Prevent Flow Fluctuation	Yes	Flows are permitted based on the resources reservation of delay levels, and isolated from each other.
3.7. Be scalable to a Large Number of Hops with Complex Topology	Yes	E2E latency is liner with hops , from ultra-low to low latency by multiple delay levels. E2E jitter is low by on-time mode.
3.8 Support Multi-Mechanisms in Single Domain and Multi-Domains	N/A	No relationship with queueing mechanism...

Next step

- The content is basically mature and detailed for implementation, and we would like to request WG adoption.
- Any questions/comments ?

Thank you!