A Compute Resources Oriented Scheduling Mechanism based on Dataplane Programmability

draft-li-coinrg-compute-resource-scheduling-00

https://datatracker.ietf.org/doc/draft-li-coinrg-compute-resource-scheduling/

Presenter: Kehan Yao

Backgrounds

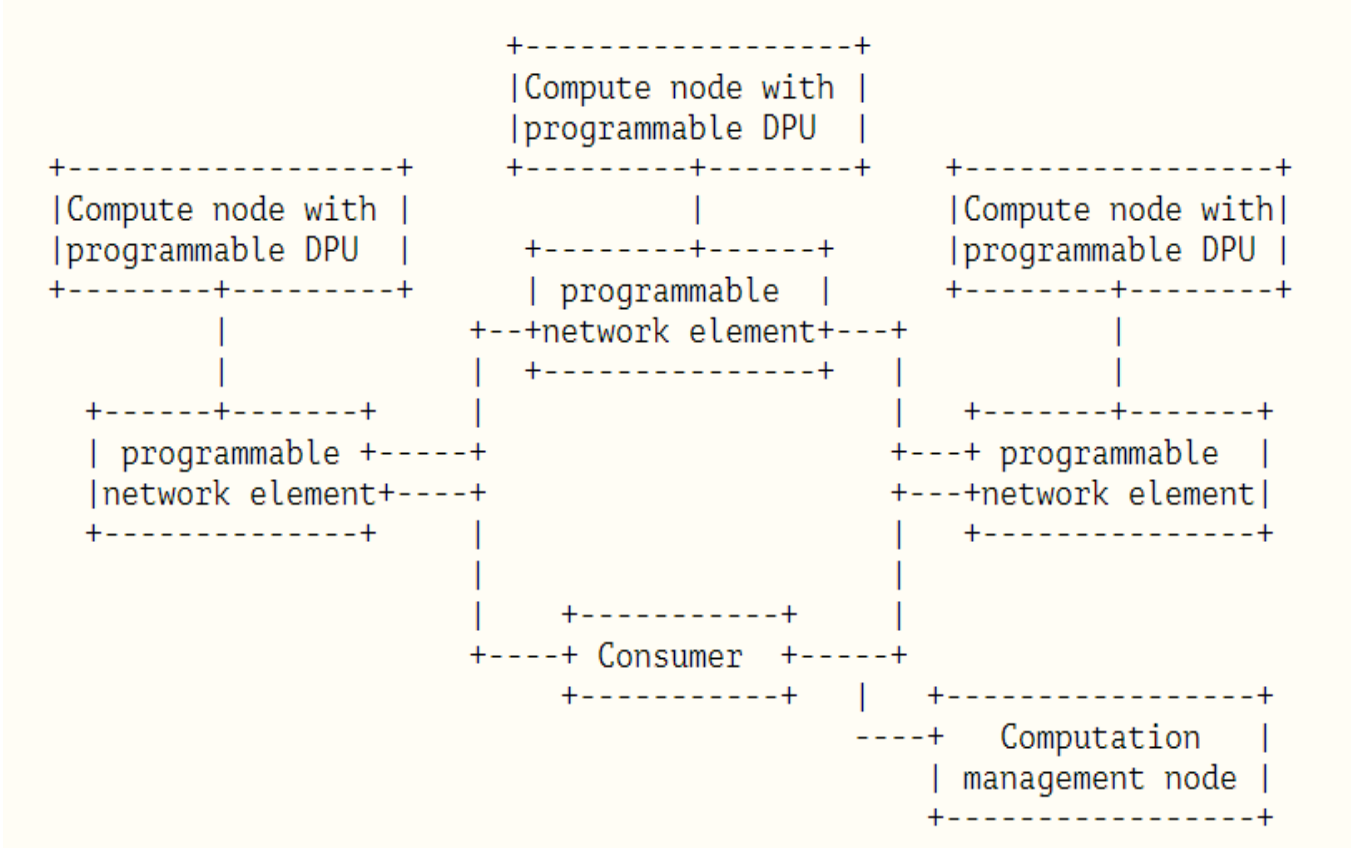
- Massive data computation moves from cloud to edge
- Moore’s Law is gradually reaching its limitation
- Domain specific chip architecture is raising up, like GPU, DPU and programmable Switch ASIC

Questions

- How to realize Cloud-Edge-Terminal coordination?
- How to effectively use different compute resources?

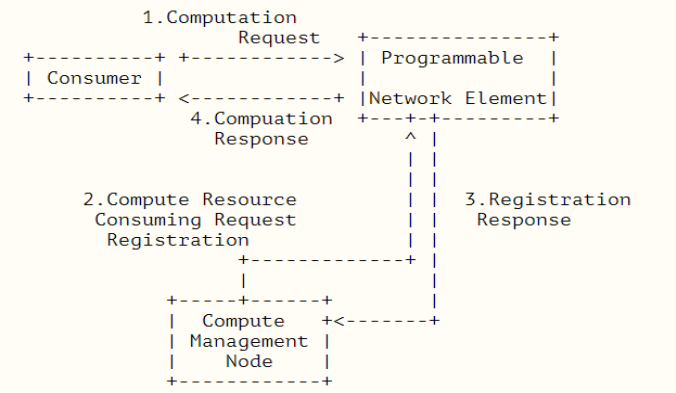
Motivation

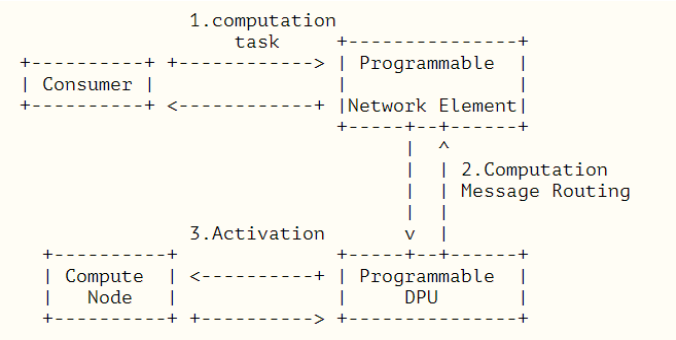
- Compute resources scheduling towards better C-E-T coordination.
- A typical topology describing the process in which a consumer (i.e terminal) wants to request for some compute resources.
- Programmable devices can be leveraged in finishing the task in network dataplane.

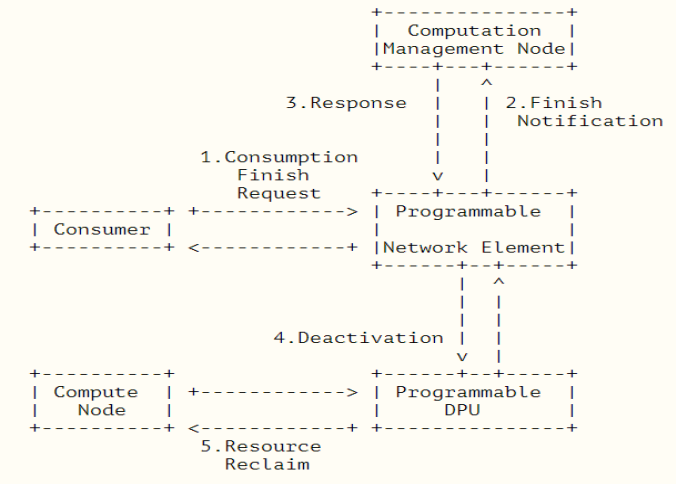


Design

- Step 1：
  - Computation Request Procedure
  - Specific compute node ID and compute resources would be allocated for the consumer.
- Step 2:
  - Computation Activation procedure
  - Programmable device will help manage the lifetime of the compute node, operational states including creating, running, pausing…
- Step 3:
  - Finish Consumption of compute resources
  - Close the lifetime management
  - Tell the management node to collect the resources.







Relative research

- We have a program together with Tsinghua working on In network computing
- In-network computing on security as an example
  - In-Network Scanner with Programmable Switches (paper accepted by Hotnets 2021)

    - state-of-the-art network scanners are implemented in commodity servers;
    - they are usually located at network edges;
    - Not fast, bandwidth waste and higher possibility in dropping packets.

    - in-network scanners are implemented in programmable switches;
    - located in network;
    - Fast, scalable, and bandwidth saving.

Relative research

- Other ongoing research
  - In-network acceleration of Machine learning systems
  - Fine-grained in-network measurement
    - Network telemetry in data plane driven by Programmable Switches

Thanks!