- There is only one VNF type

- There is an infinite queue to keep the requests before processing by the VNF

- There are a set of different service types

- Each service type is described by

* is the arrival rate of the demands, which is changing over time
* is the service rate
* is the maximum delay that this service can tolerate, where
* is the penalty function if the QoS requirement is violated (it can be a constant number, a linear function and exponential or …)

- The request from type

* arrives at time (which is specified by )
* goes into the queue
* when arrives to the front of the queue, it is picked up and processed by an instance (the processing time is determined by )
* leaves the system at time

- The SLA is violated if , in this case, the cost is

- Each instance specified by three functions , , and (…) (the inputs can be whatever)

* is the instantiation cost (e.g., license cost, …)
* is the cost of using the instance (e.g., power consumption cost, …) that depends on time
* (…) is the instantiation time (can be a constant number or depends on different parameters)

- There is an scaler. It monitors the queue, … and at time triggers instantiation, creating a new instance takes (…) and incurs cost (…), at time (…), the new instance is ready. If the scaler decides to shutdown the instance at time , the total cost will be

- The objective is to find the best number of instances to minimize

- The first scenario to simulate is the “no scaling”, where the scaler at time creates a number of instances and does not change them later in site of changing overtime