

NeoTec Use Case Discussion

Nabeel Cocker
Red Hat

Luay Jalil
Verizon

March 19th, 2025

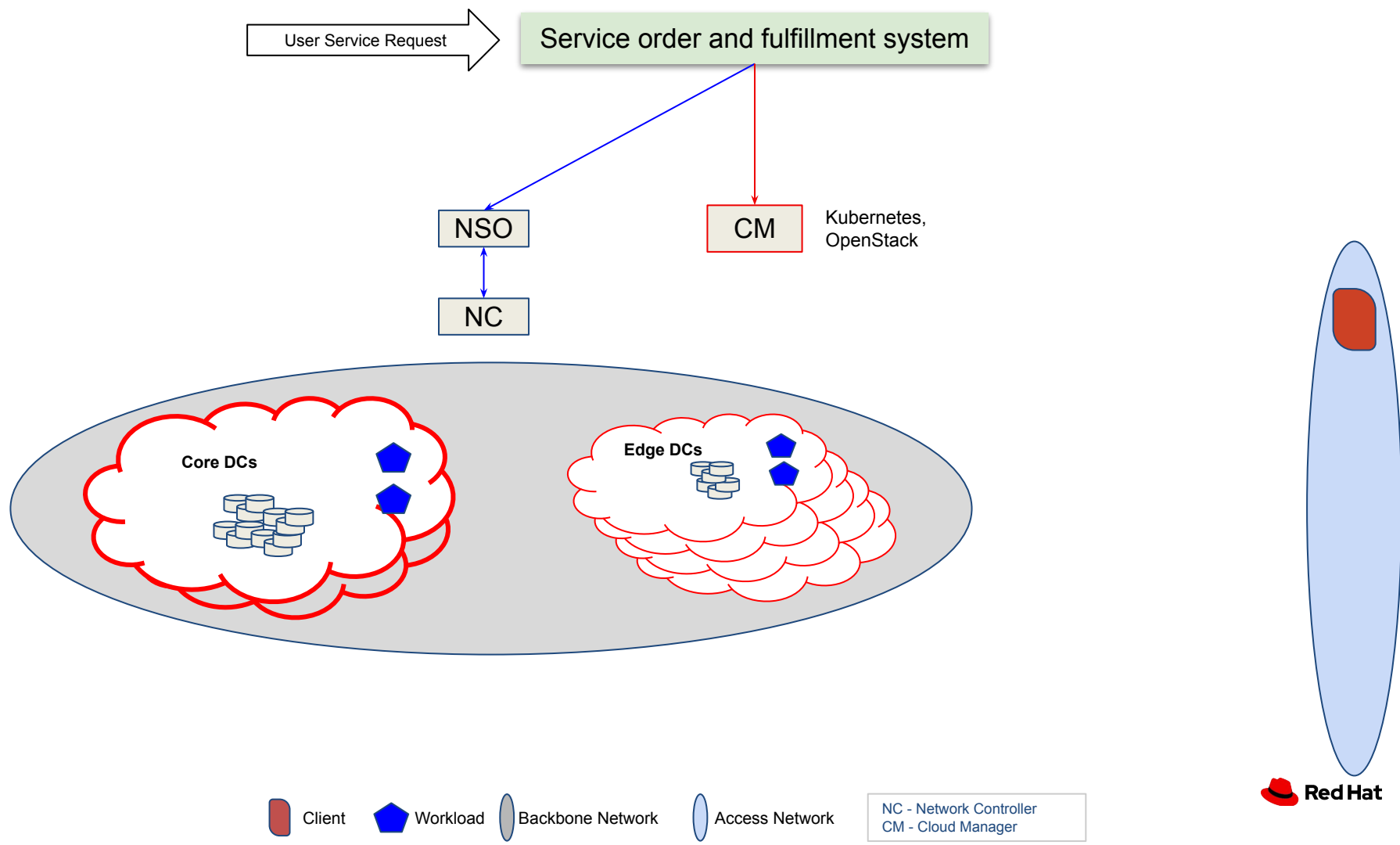
Edge computing

- Architecture that provides cloud computing capabilities at the edge of the network
- Placement of small footprint of compute resources closer to the end users or sources of data
- Main use case is reduced latency for delay sensitive services

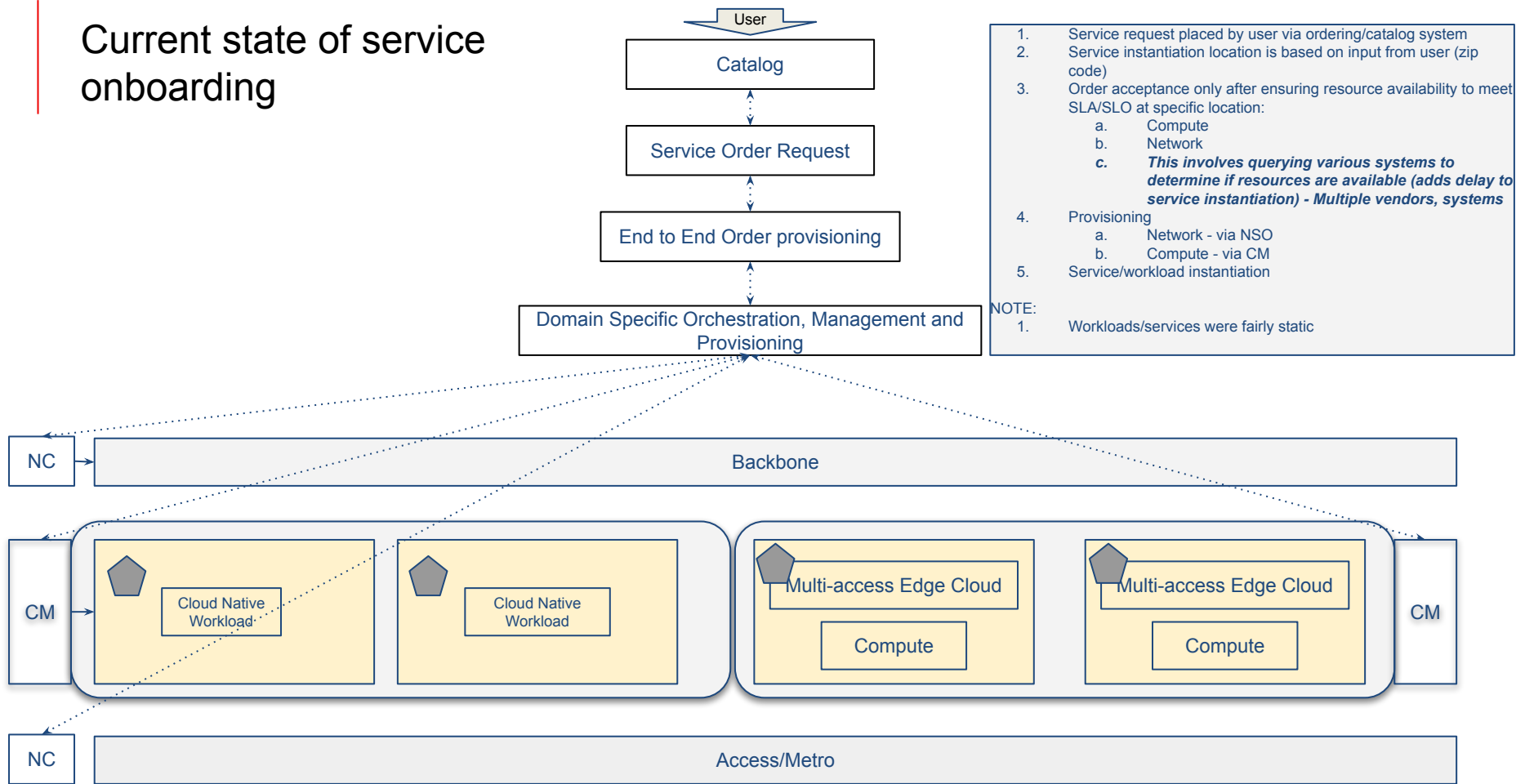
Example services

- AI
 - Online inference
 - Model instantiated close to the data
 - Facial/voice recognition at security checkpoints or building access (offices, malls etc)
 - Traffic patterns vary during the day and week
 - Office hours, weekends
 - Latency sensitive, interactive
 - Dynamic scaling based on traffic volume
- UPF
 - Scheduled and unscheduled traffic increase
 - Dynamic scaling at ports, docks, event venues
- Caching on-demand (content, scheduled/event based)

Common theme is dynamic service instantiation and life cycle management

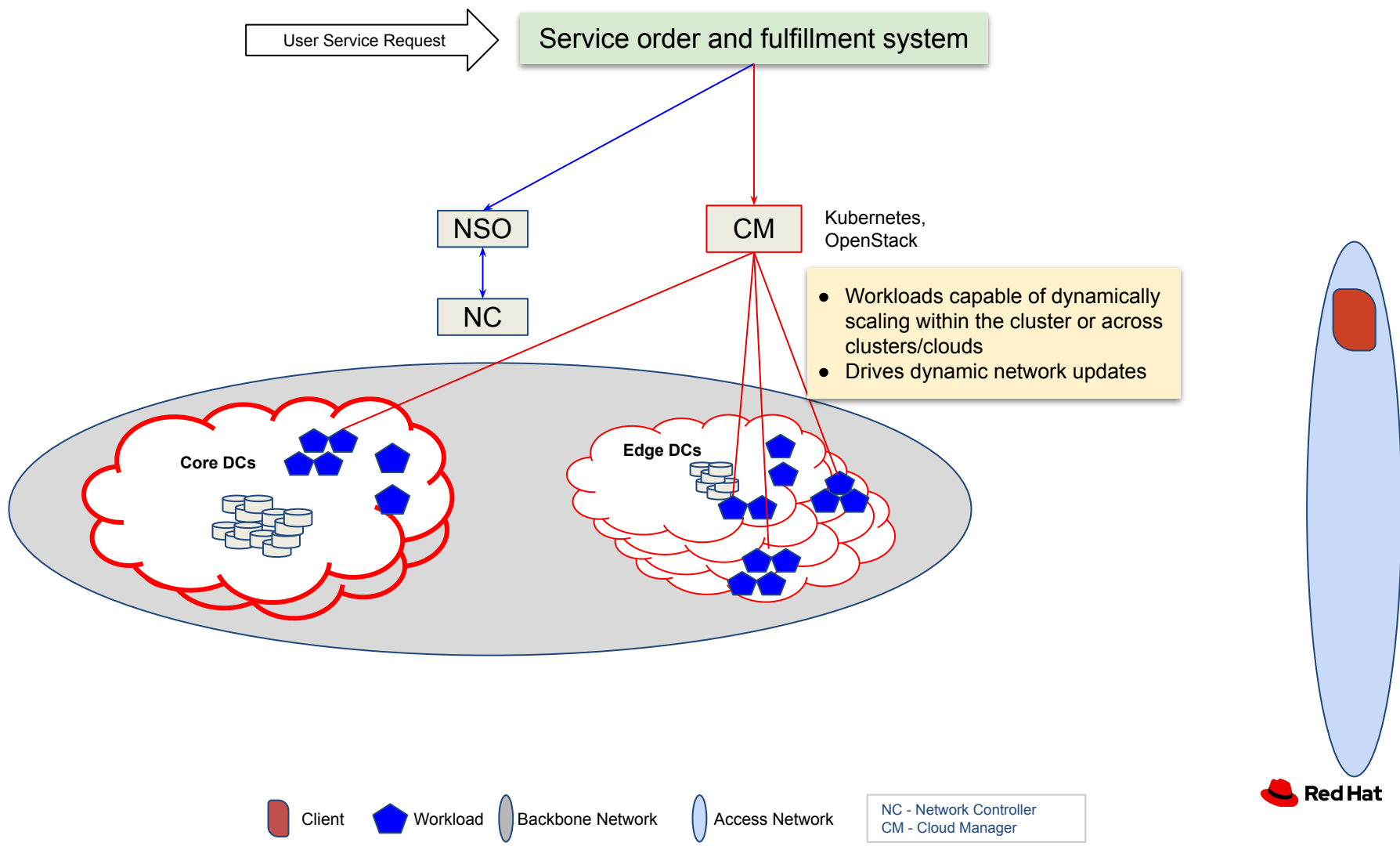


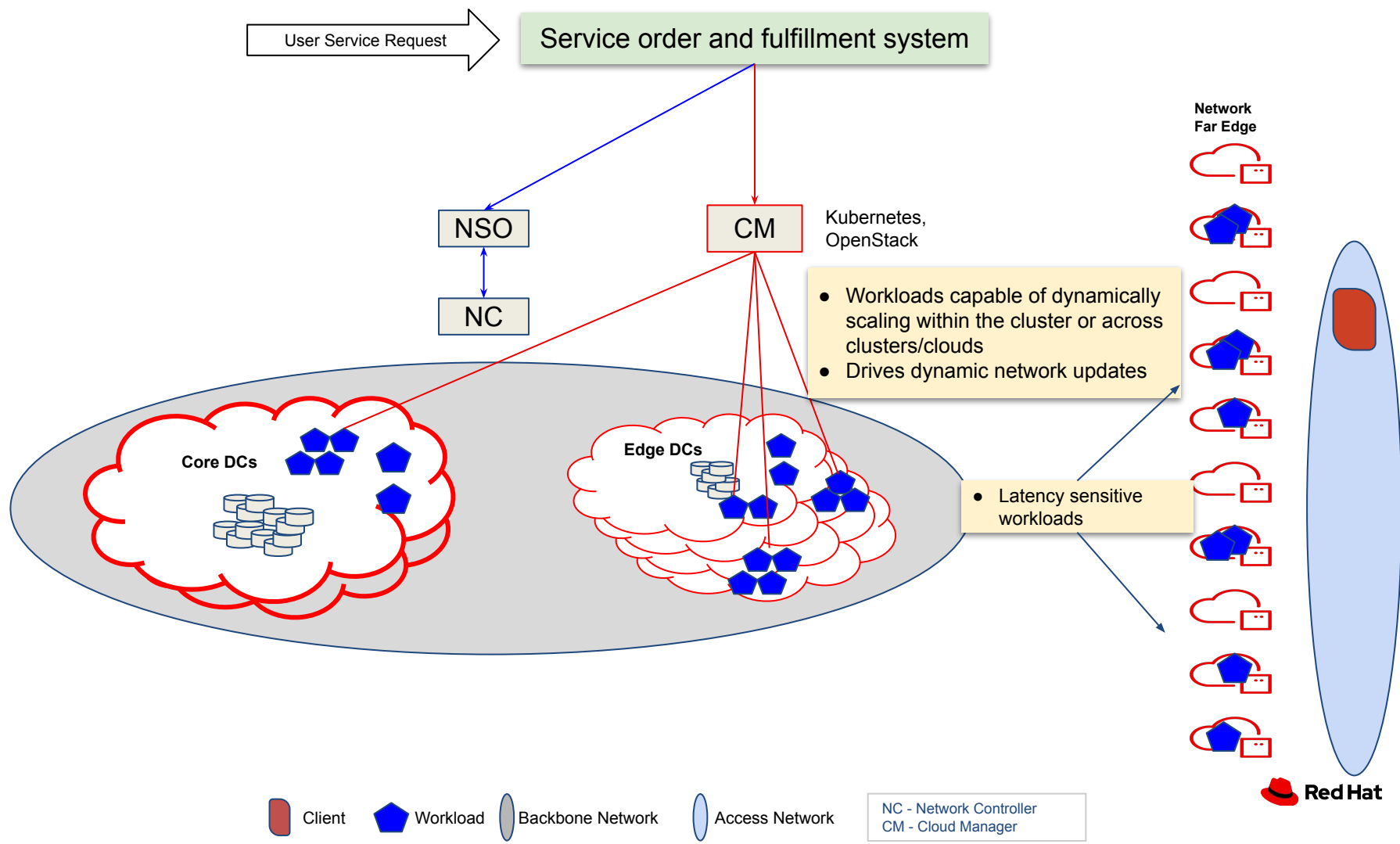
Current state of service onboarding



Cluster API (e.g, k8s API)

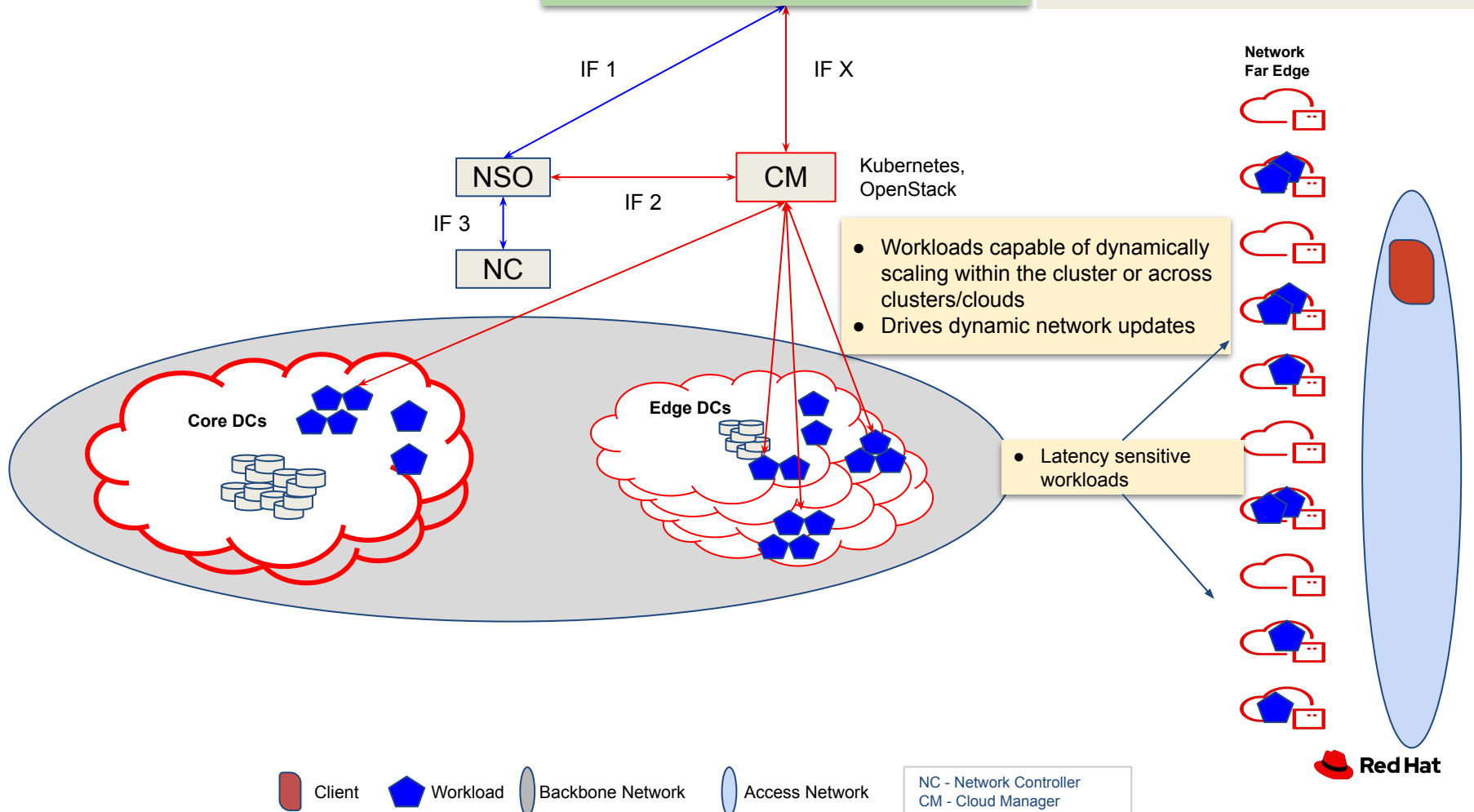
NC - Network Controller
CM - Cloud Manager





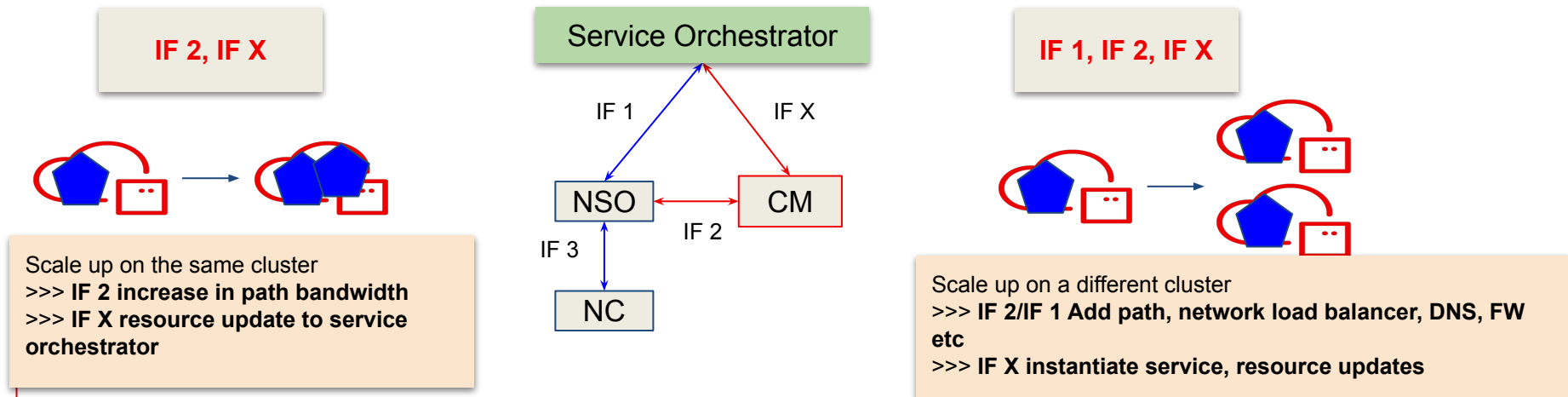
Service Orchestrator

Admission, Placement, Lifecycle management



Telco cloud manager and network service orchestrator

- Cloud manager
 - Life cycle management of workload/service
 - Scheduling and placement of service/workload on specific cloud/cluster and/or node
 - Pool of resources that are allocatable for workloads
 - Placement and scheduling based on workload resource requirements
- Workload can scale upto limit/quota (example trigger: CPU, queue occupancy)
 - Cloud:
 - Service scaled up e.g., replicas/deployment/stateful set scaled up OR service instantiated on a different cluster/cloud
 - NOTE: currently the placement of the workload does not take into account network path metrics but rather compute metrics only
 - Network:
 - Scale up bandwidth, instantiate additional path with same network requirements, updates to DNS, network load balancer, FW etc



Service Orchestrator to CM and NSO

- Orchestrator determines placement based on:
 - Service requirements, available cloud capabilities, etc
 - Can the workload/service be instantiated at the required SLA/SLO without impacting existing services?
 - IF 1**
 - ***Requires current network resource availability information (based on metrics/telemetry or API)***
 - *Cloud vs node*
 - IF X**
 - ***Requires current compute resource availability information (based on metrics/telemetry or API)***
 - *Cloud vs node especially when it comes to NUMA awareness*
 - ***Note: this is location dependant information***
 - Business logic:
 - Pricing
 - priority vs other workload/customer workloads for preemption
 - If there is a mix of GPUs, prioritize using GPUs with a higher performance delivered per unit of power
 - Green energy
 - If admissible
 - Trigger workflow:
 - NSO to provision path/network service
 - CM to reserve compute resources, instantiate service, LCM of application
 - Orchestrator enters the lifecycle management state for the service

What is the ask?

- Definition of network resources that can be exposed to the CM to facilitate more optimized workload/service placement/scheduling
 - How are these exposed/consumed (API?)
- Definition of cloud resources to be exposed to the network orchestrator
- The above requires working through specific use cases to get a better understanding on the specifics
 - Complete life cycle of a service

- Support the initiative and will actively participate