K8S CNI FOR XPU OFFLOADING

OPI Community Discussion

Toshi Kani, HPE

Jan 2023

DESIGN & ASSUMPTIONS

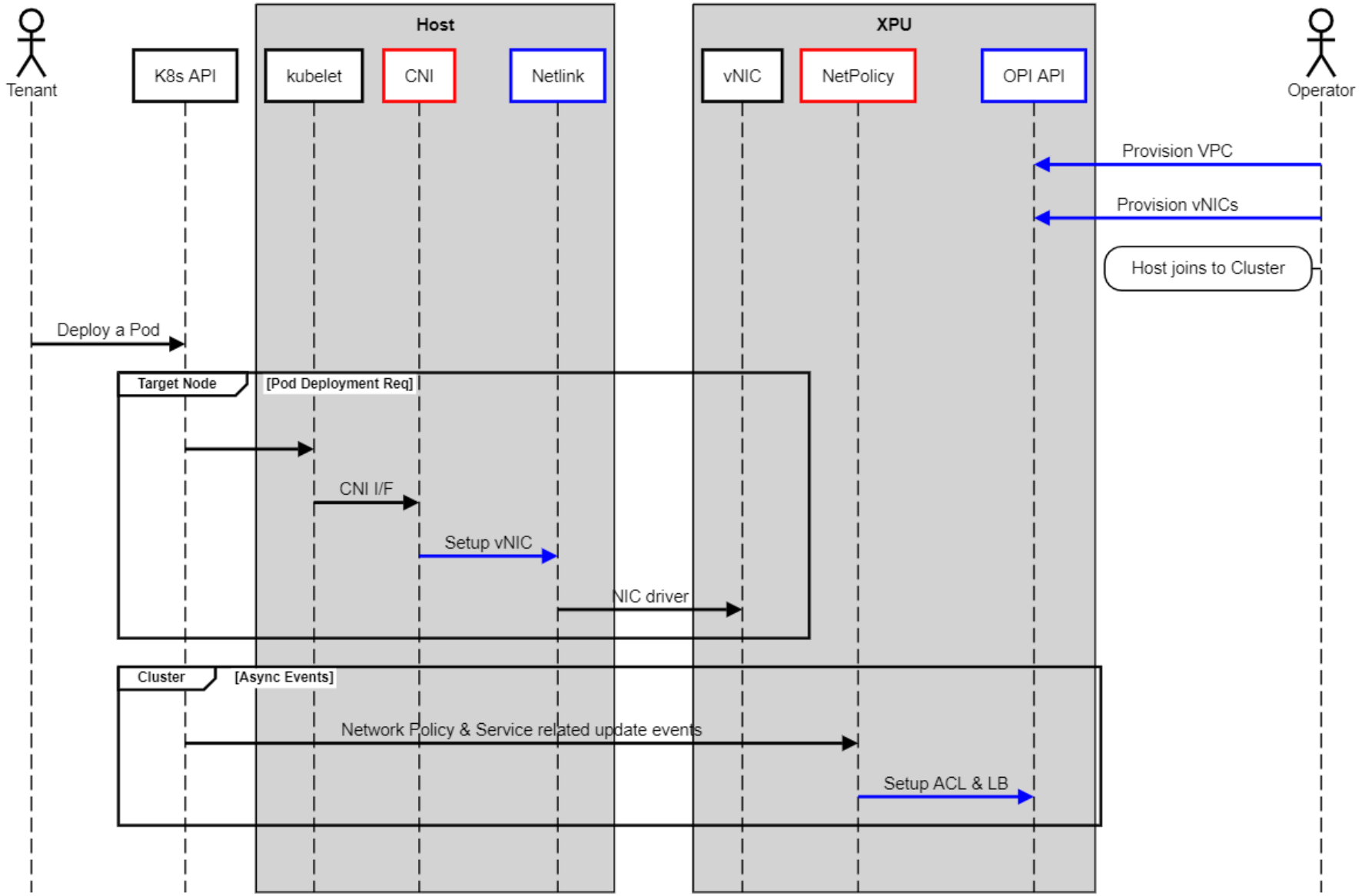
- Cloud use-cases need tenant/operator role separation
  - VPC (as tenant overlay network) isolates vNIC interfaces from the operator underlay network
- K8s CNI package calls OPI API and netlink API
  - OPI API – XPU API that sets up XPU for network offloading (See the next page)
  - Netlink API – Linux API that sets up host NIC interfaces, e.g., link state, IP, MTU, routing. No API change
- OPI API abstracts vendor-specific XPU API
  - Vendor-specific API includes XPU Network OS & SDK APIs unless they are adopted as OPI API
- K8s CNI package has the following main modules for XPU offloading
  - K8s CNI – CNI plugin, which is called from kubelet and sets up vNIC as Pod network interface
  - Network Policy – Controller that monitors K8s API and sets up XPU for K8s Network Policy offloading
  - Kube-proxy – Controller that monitors K8s API and sets up XPU for K8s Service (as load-balancing) offloading
- The diagrams cover static & dynamic vNIC provisioning flows as example setups with the same APIs
  - Static – Assignable vNICs are pre-provisioned to VPC. The host then joins to K8s cluster as a worker node
  - Dynamic – vNIC is provisioned to VPC on demand for increased flexibility
- The diagrams may apply to multiple CNI types. Necessary code changes vary in CNI types
  - E.g., Calico Felix (as Network Policy module) allows plugin data-plane drivers, such as OPI driver

OPI API FUNCTIONS FOR K8S CNI XPU OFFLOADING

- Provision VPC
  - Setup overlay network (e.g., VXLAN), tunnel endpoint, underlay routing, etc.
  - Setup IPSec for underlay network (e.g., VXLAN over IPSec) – Optional feature
- Provision vNIC
  - Allocate vNIC as an SR-IOV (or SIOV) NIC VF, which is an L2 NIC interface in the overlay network
    - Static provisioning may require some offloading setup, e.g., traffic shaping, at Pod deployment as a separate step
- Setup Stateful Firewall (ACL)
  - Setup ingress & egress stateful L3/4 firewall, similar functionality as AWS Security Groups (SG)
    - Policy rule with a group target, e.g., iptables IPSet and AWS SG, is highly desirable
- Setup Stateful Load-Balancing (LB)
  - Setup stateful L3/4 load-balancing with DNAT & SNAT

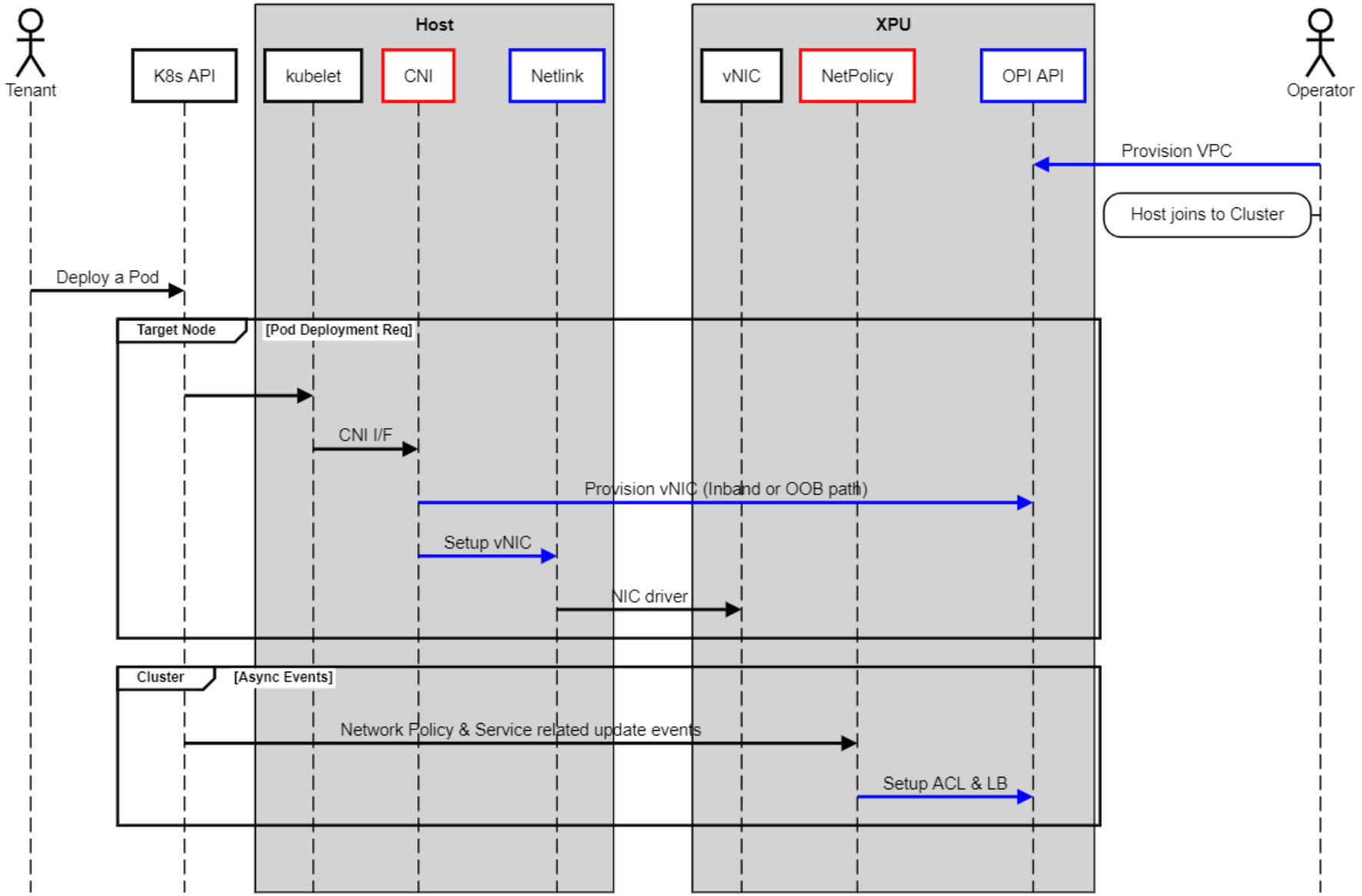
Note: OPI API interface definition is intentionally left for separate discussions

STATIC VNIC PROVISIONING FLOW



Note: Kube-proxy is omitted for brevity.
The sequence is shared with NetPolicy

DYNAMIC VNIC PROVISIONING FLOW



Note: Impl may vary in detail., e.g., alt. provision vNIC in the background to replenish minimal assignable vNICs