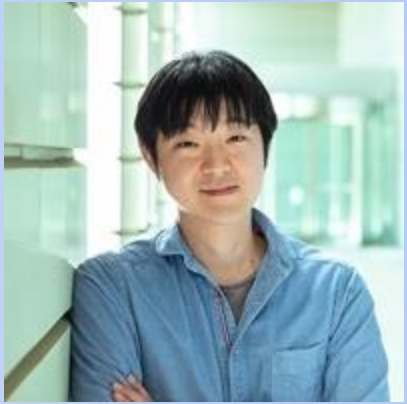
CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA

Dressing-up your cluster for AI in minutes with a portable network CR

CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA

November 12, 2024
Salt Lake City



Tatsuhiro Chiba
Senior Technical Staff Member
IBM Research

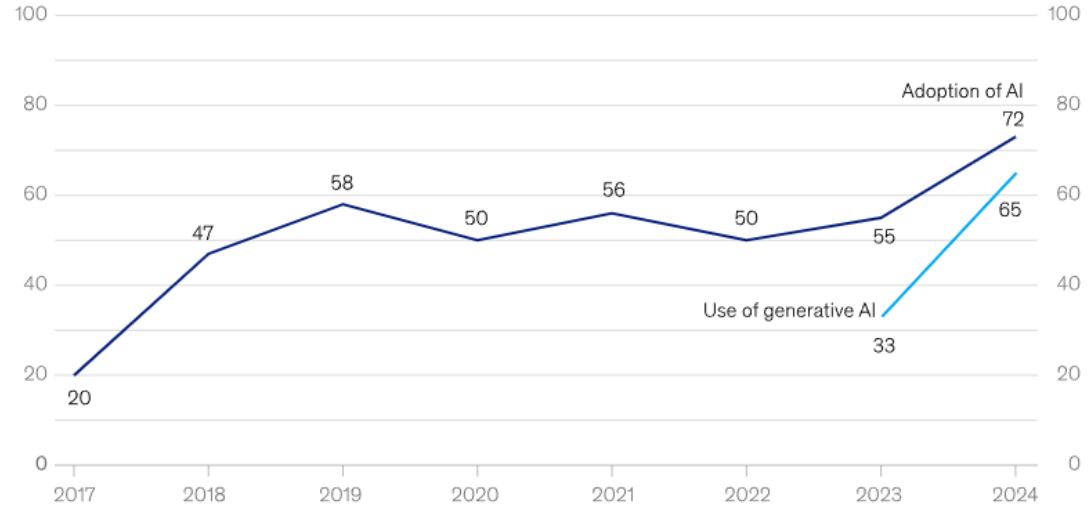
Sunyanan Choochotkaew
Staff Research Scientist
IBM Research

AI adoption has increased dramatically, especially after the rise of generative AI.

2x Use of GenAI in 2024

AI adoption worldwide has increased dramatically in the past year, after years of little meaningful change.

Organizations that have adopted AI in at least 1 business function,[1] % of respondents



[1]In 2017, the definition for AI adoption was using AI in a core part of the organization's business or at scale. In 2018 and 2019, the definition was embedding at least 1 AI capability in business processes or products. Since 2020, the definition has been that the organization has adopted AI in at least 1 function.
Source: McKinsey Global Survey on AI, 1,363 participants at all levels of the organization, Feb 22–Mar 5, 2024

McKinsey & Company

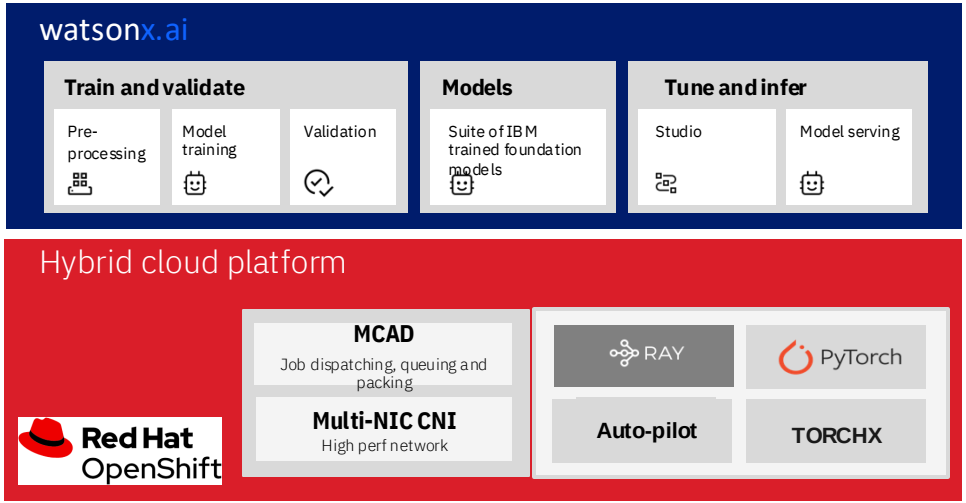
Trending: AI workloads in Containers

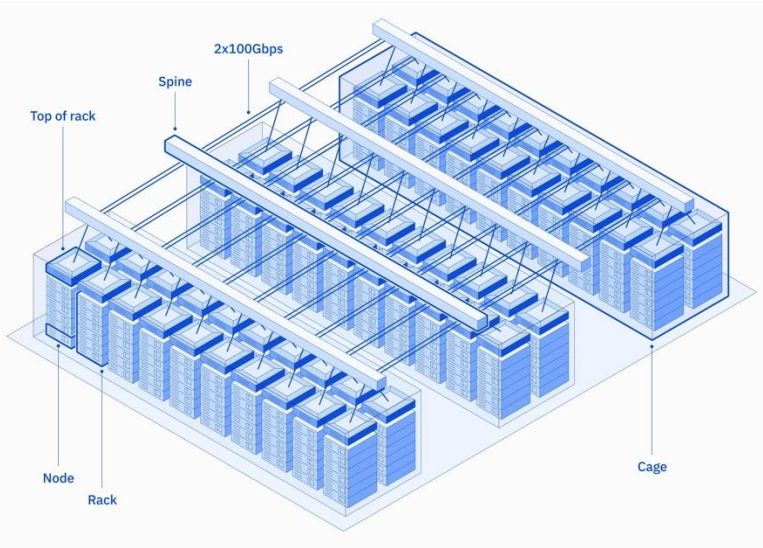
- Consistent Environments
- Rapid Experimentation
- Simplified Dependency Management

IBM Cloud Vela:
IBM's first AI-optimized, cloud-native supercomputer

- Vela is an A100 supercomputer built on top of IBM Cloud VPC (operational by May 2022)
- 8 x A100 GPUs with 80G of memory
- Virtualization with baremetal performance (<~5%)
- 400G Ethernet network (RoCE / GPU Direct RDMA)
- Supports training models in Billions of parameters that process Trillions of tokens

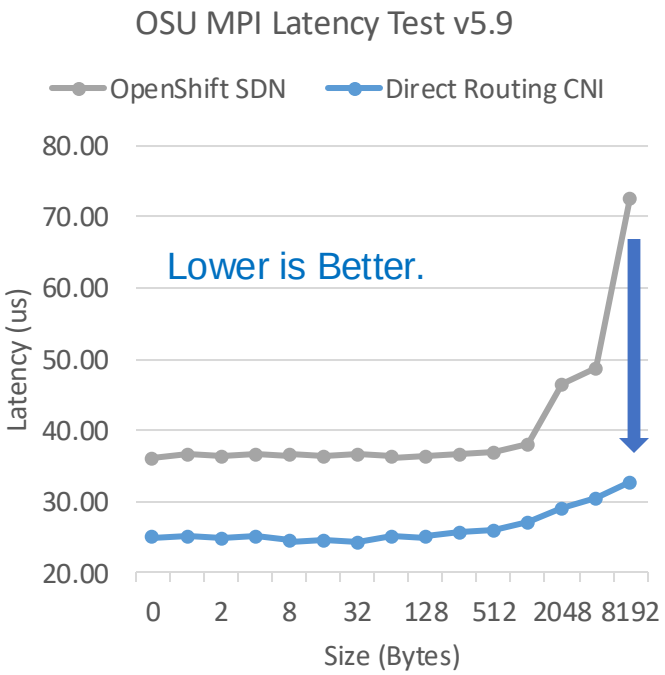
- Multi-NIC CNI dynamically provides **a pod direct communication** with a secondary network

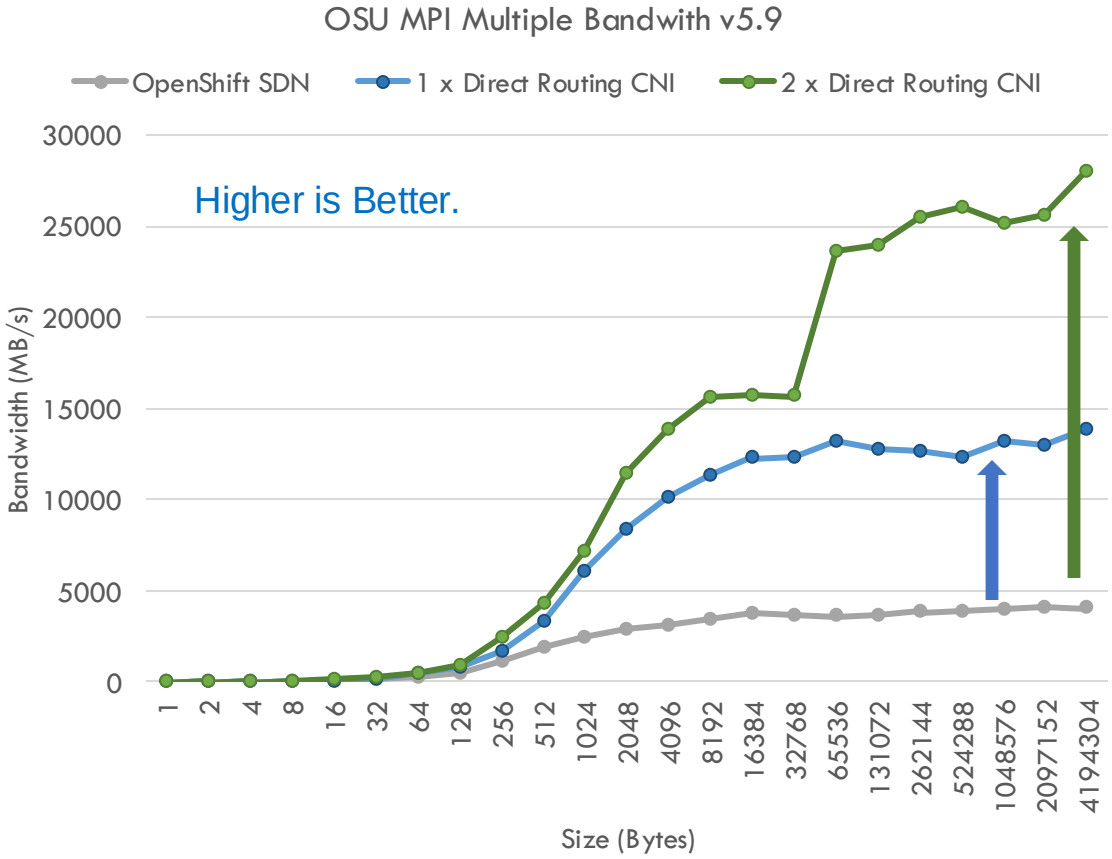




A Common Goal of AI Networking Solution in Containers

Native Performance at Scale

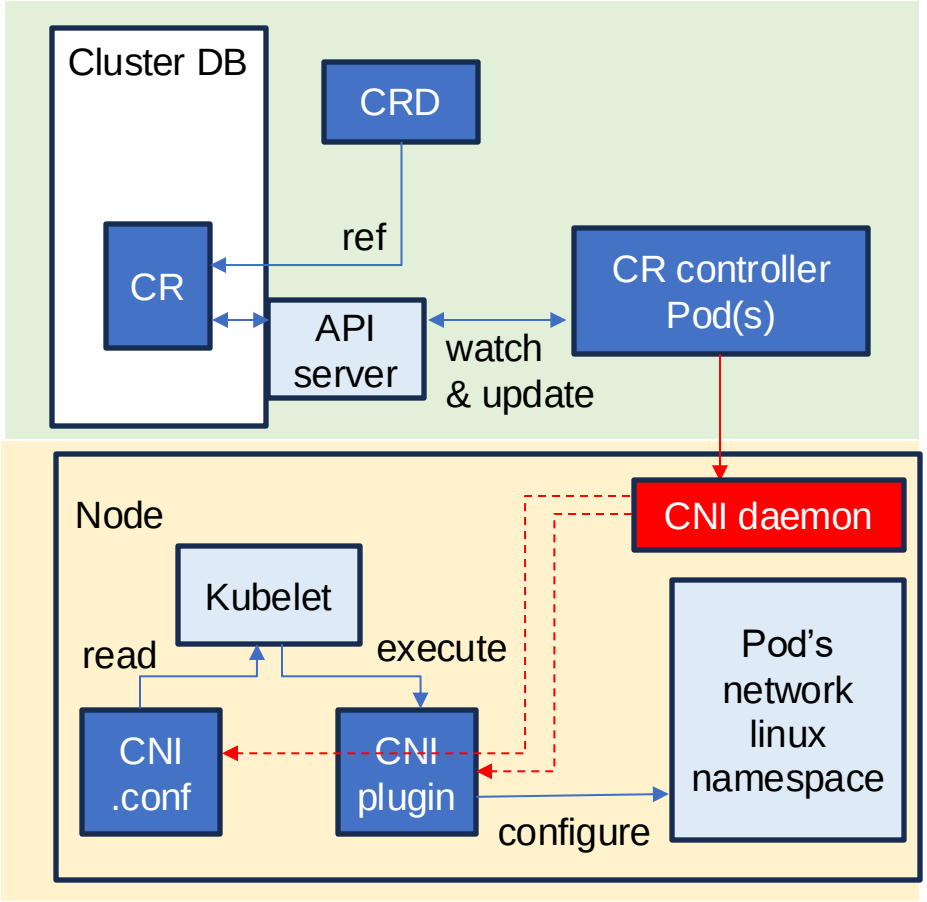




Content

- Background – *CNI, CR*
- Reasons behind “A Portable” Network CR
- Idea – *in a very simplified version*
- Our implementation – *Multi-NIC CNI v1*
- What’s Next?

Network CR



CR (custom resource) is commonly used to serve inputs to CNI for more flexibiltiy in network configuration.

Multus has **NetworkAttachmentDefinition CR** to delegate execution for secondary networks

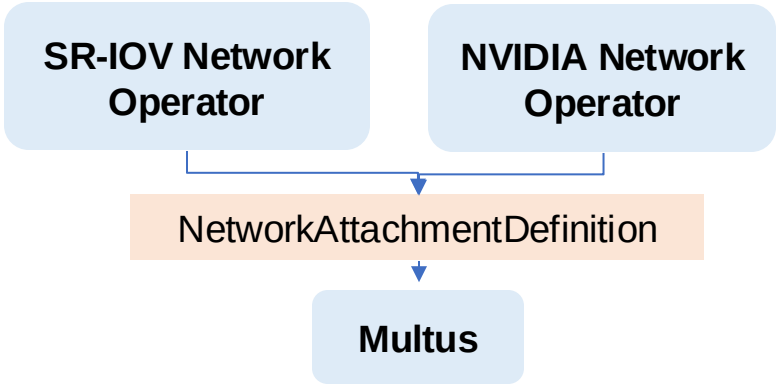
SR-IOV Network Operator has SriovNetwork CR

NVIDIA Network Operator has MacvlanNetwork, HostDeviceNetwork, IPoIBNetwork

Our CNI names Multi-NIC CNI, our CR would name ???

Basic of CNIs for High Performance Network Device

- Provide a pod direct communication of high performance network device, separating from the primary network, with a secondary network attachment by Multus.
- **Multus** attaches secondary networks by network plugins,
  - defined in NetworkAttachmentDefinition CR,
  - to be annotated in Pod Annotation.



1. Define NetworkAttachmentDefinition CR

```
apiVersion: "k8s.cni.cncf.io/v1"
kind: NetworkAttachmentDefinition
metadata:
  name: macvlan-conf
spec:
  config: '{
      "cniVersion": "0.3.0",
      "type": "macvlan",
      "master": "eth0",
      "mode": "bridge",
      "ipam": {
        "type": "host-local",
        "subnet": "192.168.1.0/24",
        "rangeStart": "192.168.1.200",
        "rangeEnd": "192.168.1.216",
        "routes": [
          { "dst": "0.0.0.0/0" }
        ],
        "gateway": "192.168.1.1"
      }
    }'
```

2. Pod Annotation

```
apiVersion: v1
kind: Pod
metadata:
  name: samplepod
  annotations:
    k8s.v1.cni.cncf.io/networks: macvlan-conf
```

As-is solution

Pod direct communication is a key of Pod's networking performance.
However, direct routing is not always available as-is, *especially on virtual cloud*.

	TCP/IP with L3	ENI on AWS	RoCE v2
CNI & Routing solution	Host Routing + IPVLAN	VPC API for IP Registration	Host Device + Host IP
⇩ Complexity ⇩	Options: 1. A central pool with per-pod IP routing configuration → *not scale* 2. **Per-host per-interface pool** with static configuration	IP in **valid range of scheduled host**	Static IP **for each host and each device**
#CR for N host with M interfaces	**N**xM	**N**	**N**xM

Limitation 1 of as-is solution

Host-dependent IP requirement → Host must be known at Pod annotating point.

1. Define NetworkAttachmentDefinition CR (s)

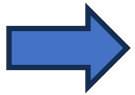
2. Pod Annotation

Multiple NetworkAttachmentDefinitions for direct routing (as-is)

NetAttachDef for Host A

NetAttachDef for Host B

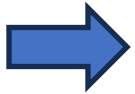
NetAttachDef for Host C



Pod
annotations:
k8s.v1.cni.cncf.io/networks: *either A or B or C*

A portable network CR (to-be)

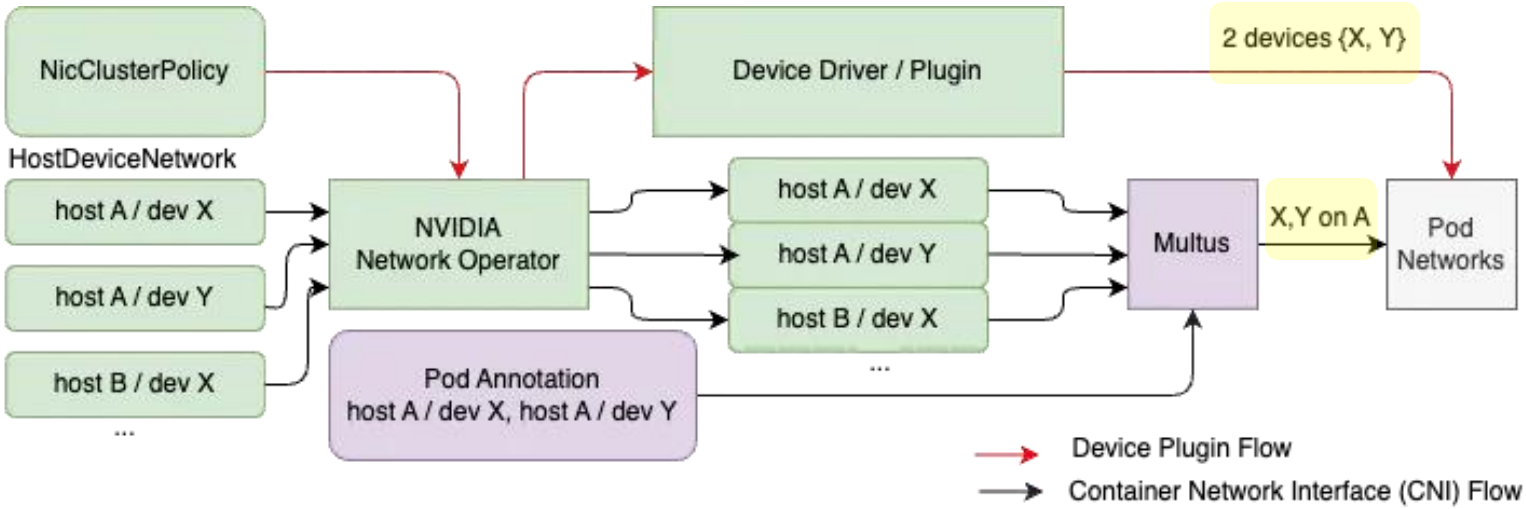
NetAttachDef X



Pod
annotations:
k8s.v1.cni.cncf.io/networks: X

Limitation 2 of as-is solution (RoCE v2)

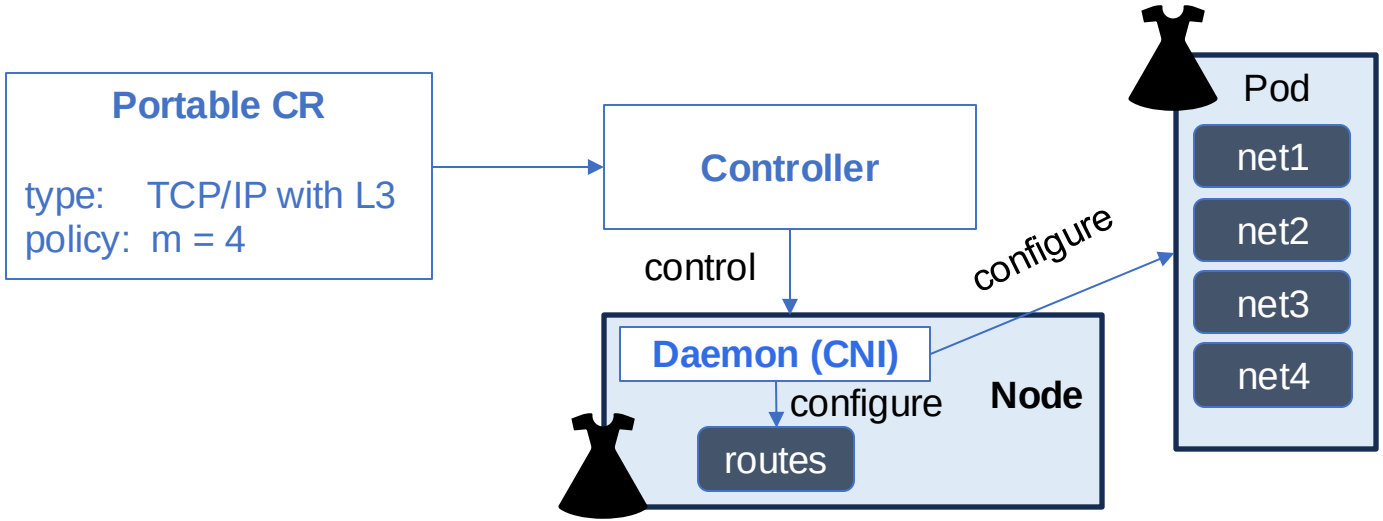
RoCE device plugin does not aware of Pod's annotation
→ Misorder of IP assignment for interface-dependent valid range



Idea of a portable network CR to dress up your cluster

1. User annotates **a single network for *m* secondary interfaces**.

2. The rest tasks (interface discovery, routing, IP assignment) are **automated**.

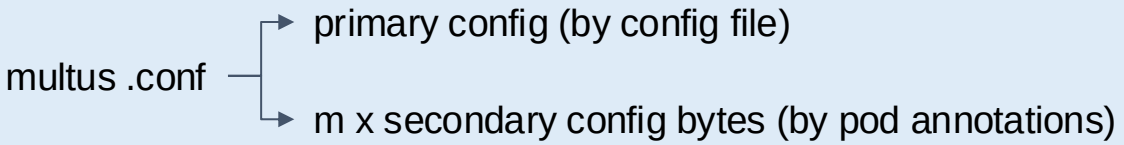


Key Design

An implementation to unpack single configuration to multiple configurations <u>after executions of scheduler and device manager</u>

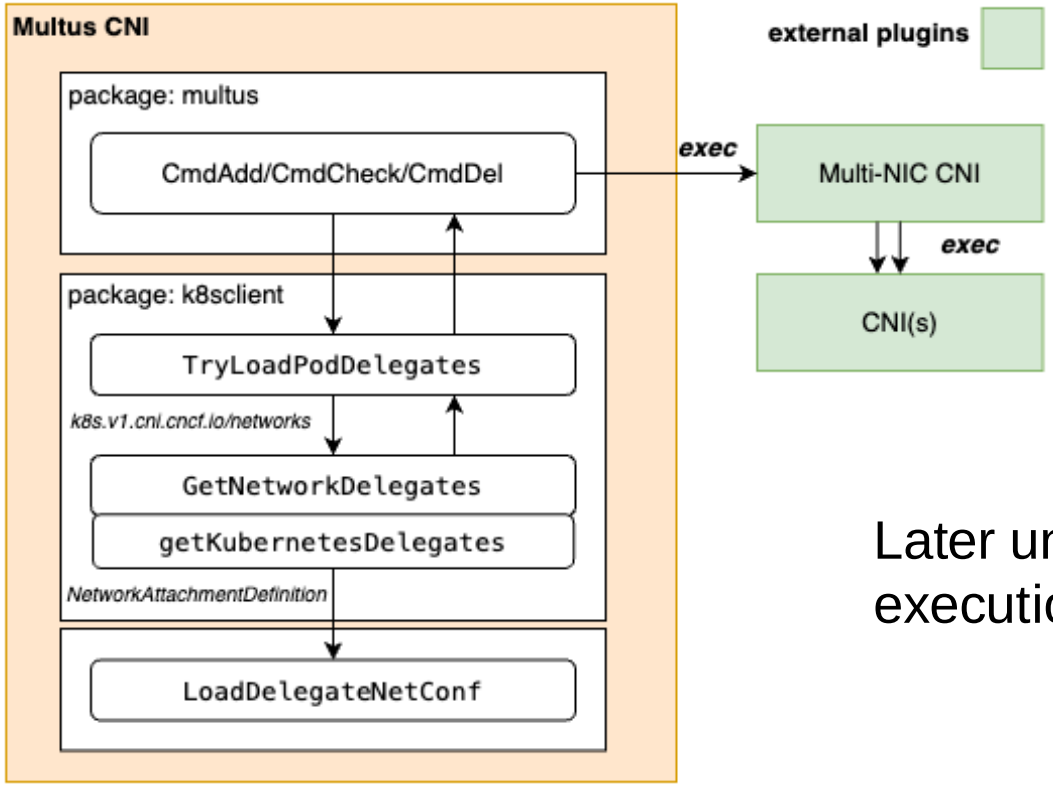
Remark:

Multus does 1-to-m configurations.



Even so, it unpacks the configuration *<u>before host is scheduled.</u>*

Multi-NIC CNI – Mechanism Behind

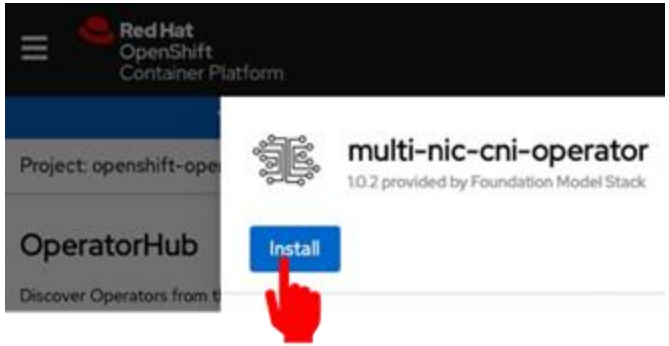


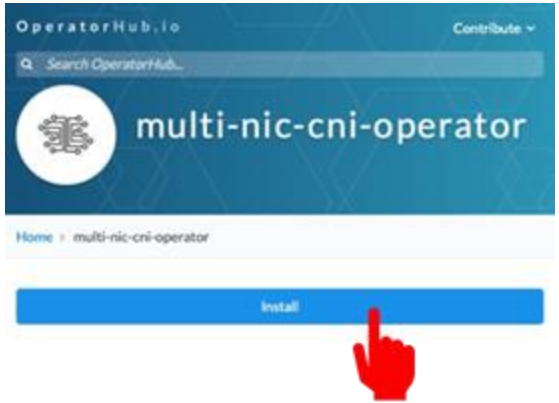
Later unpack and delegate CNI executions after Multus delegation

Multi-NIC CNI – 3 magic steps

CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA

1. Install Operator





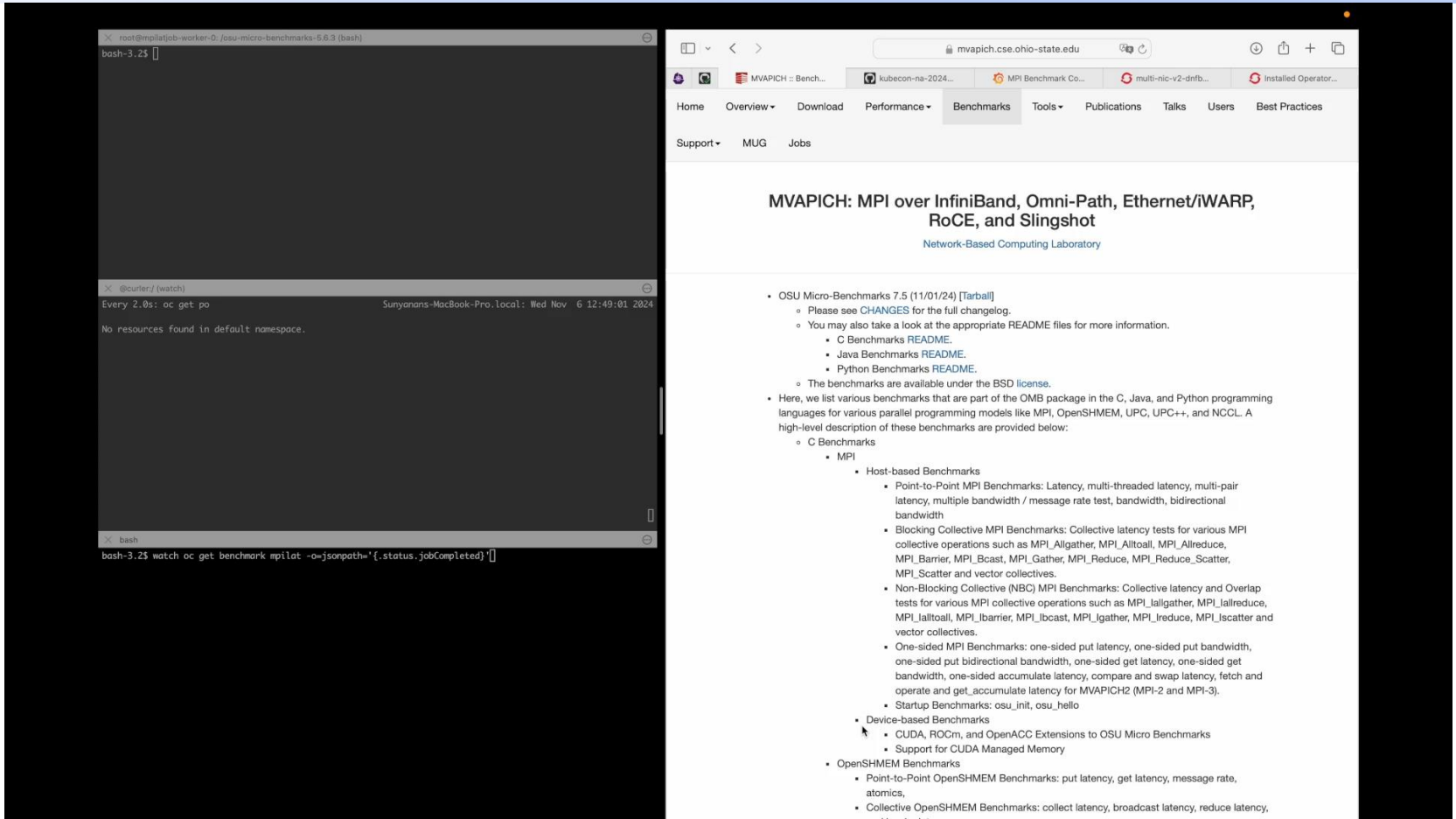
2. Deploy MultiNicNetwork (for TCP/IP L3 cluster)

```yaml
apiVersion: multinic.fms.io/v1
kind: MultiNicNetwork
metadata:
  name: multi-nic-sample
spec:
  subnet: "192.168.0.0/16"
  ipam: |
    {
      "type": "multi-nic-ipam",
      "hostBlock": 6,
      "interfaceBlock": 2,
      "vlanMode": "l3"
    }
  multiNICIPAM: true
  plugin:
    cniVersion: "0.3.0"
    type: ipvlan
    args:
      mode: l3
```

3. Attach Network to Pod

```yaml
metadata:
  annotations:
    k8s.v1.cni.cncf.io/networks: multi-nic-sample
```

Multi-NIC CNI - DEMO



Multi-NIC CNI – Currently Supported Solutions

TCP/IP with L3

```
apiVersion: multinic.fms.io/v1
kind: MultiNicNetwork
metadata:
  name: multinic-ipvlanl3
spec:
  subnet: "192.168.0.0/16"
  ipam: |
    {
      "type": "multi-nic-ipam",
      "hostBlock": 8,
      "interfaceBlock": 2,
      "vlanMode": "l3"
    }
  multiNICIPAM: true
  plugin:
    cniVersion: "0.3.0"
    type: ipvlan
    args:
      mode: l3
```

ENI on AWS

```
apiVersion: multinic.fms.io/v1
kind: MultiNicNetwork
metadata:
  name: multinic-aws-ipvlan
spec:
  ipam: |
    {
      "type": "multi-nic-ipam",
      "hostBlock": 8,
      "interfaceBlock": 2,
      "vlanMode": "l2"
    }
  multiNICIPAM: true
  plugin:
    cniVersion: "0.3.0"
    type: aws-ipvlan
    args:
      mode: l2
```

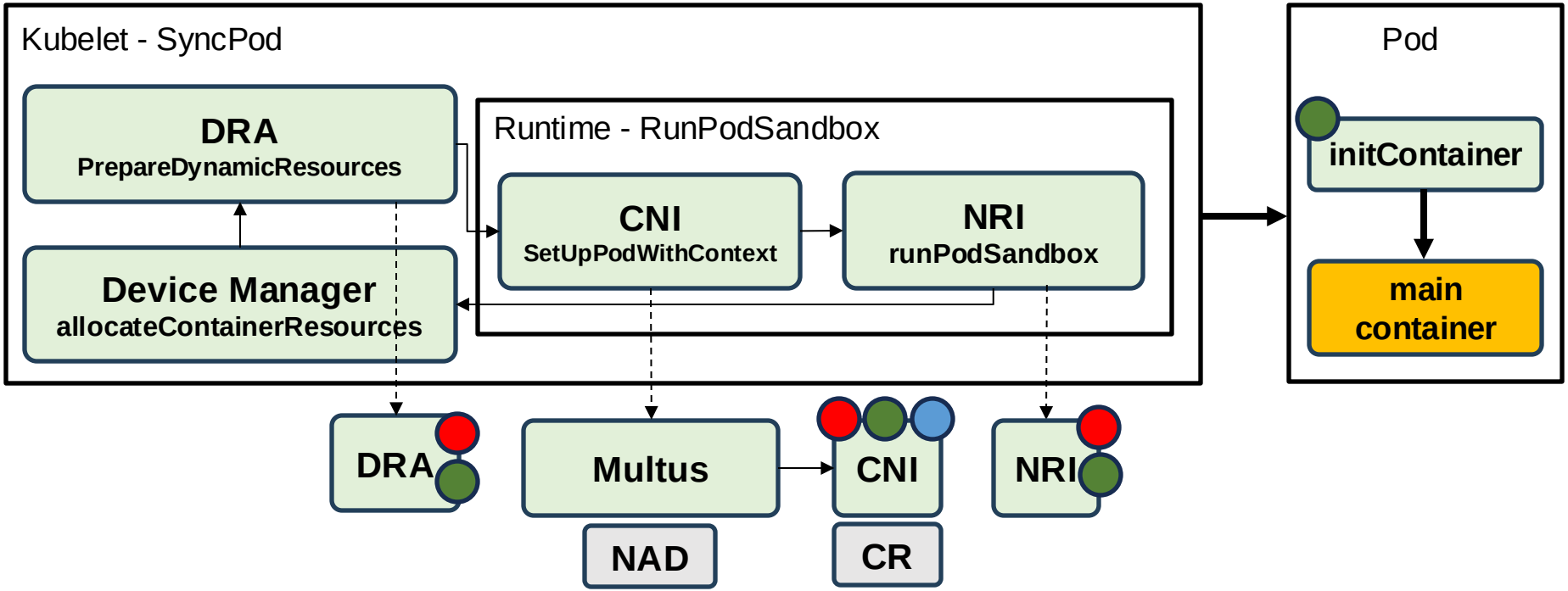
RoCE v2

```
apiVersion: multinic.fms.io/v1
kind: MultiNicNetwork
metadata:
  name: multinic-mellanox-hostdevice
spec:
  subnet: ""
  ipam: |
      {
        "type": "host-device-ipam"
      }
  multiNICIPAM: false
  plugin:
    cniVersion: "0.3.1"
    type: mellanox
```

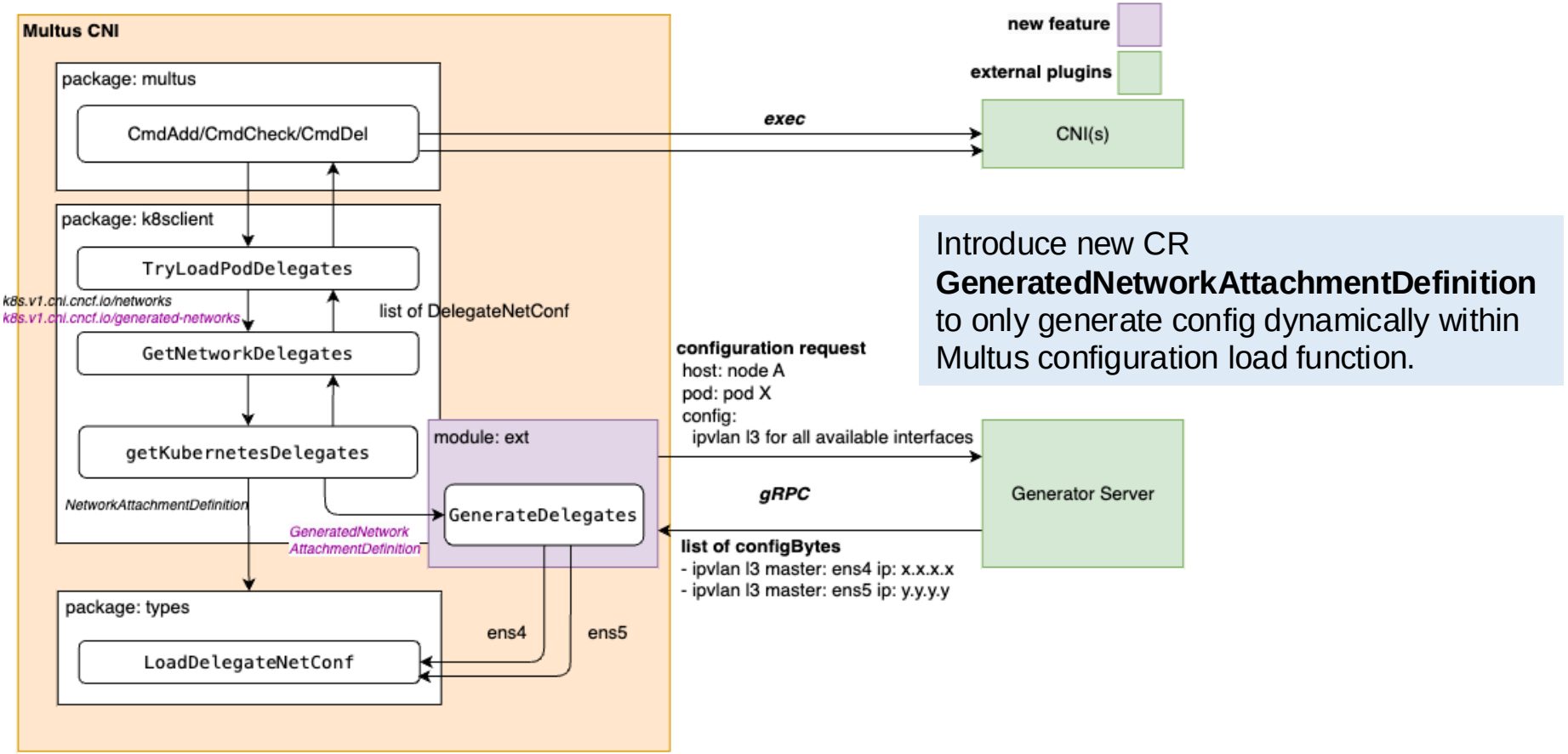
Alternative implementation – with current k8s

Key Functions

- select interfaces and assign dynamic host-dependent IP
- configure Pod's network namespace
- register the IP/configure route for IP

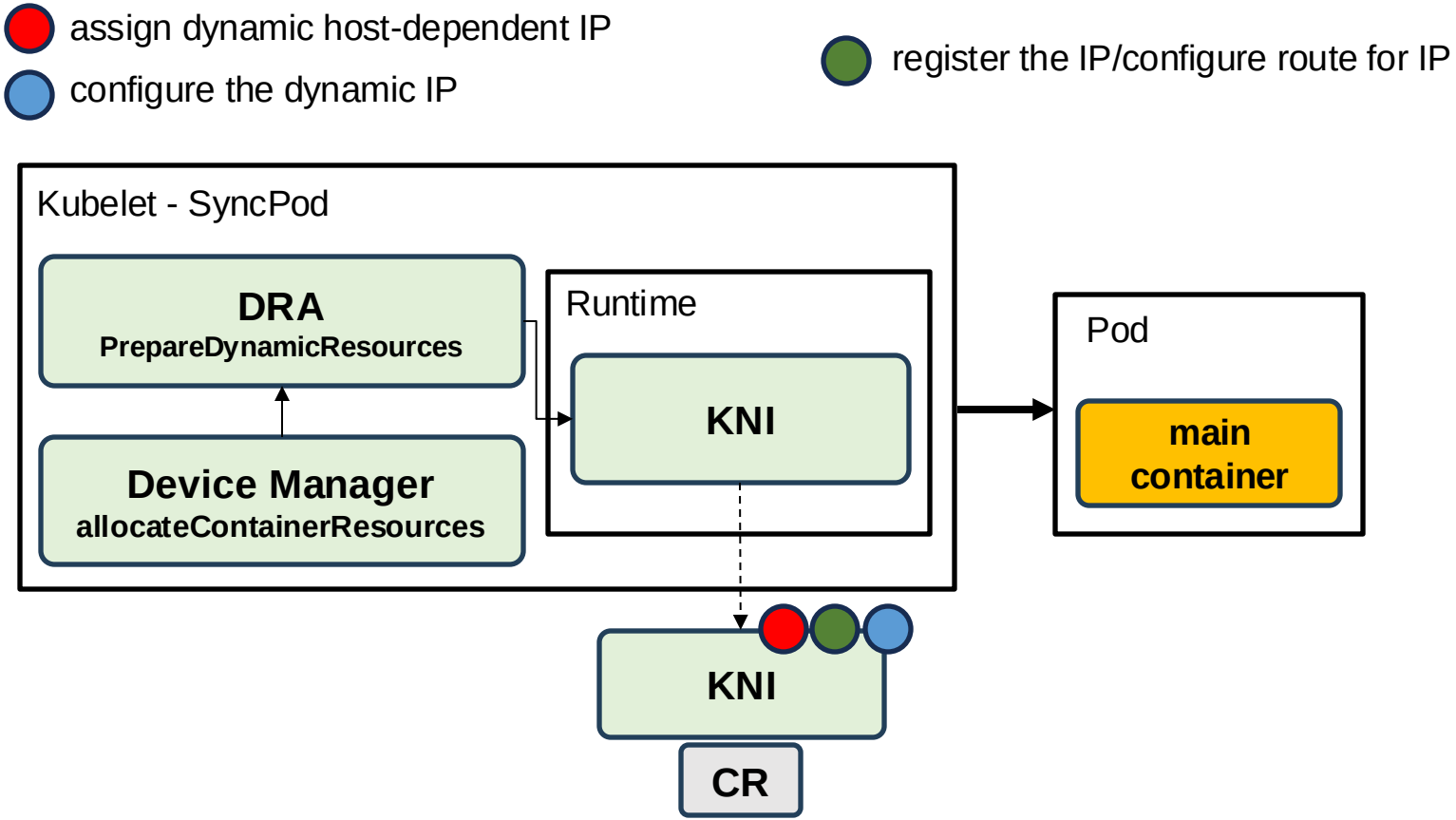


Alternative implementation – with Multus extension



Introduce new CR **GeneratedNetworkAttachmentDefinition** to only generate config dynamically within Multus configuration load function.

Alternative implementation – with future k8s

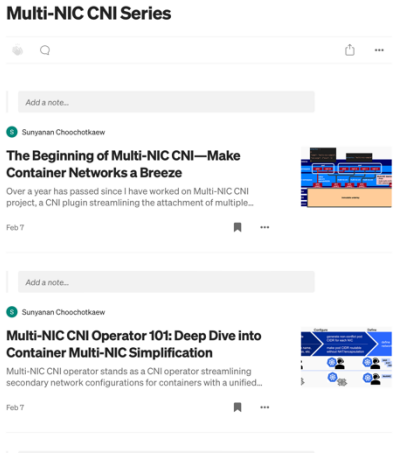


Resources & Team

Open Source Project

https://github.com/
foundation-model-stack/
multi-nic-cni

Medium Blogs



The infrastructure powering IBM's Gen AI model development

Talia Gershon* Seetharami Seelam* Brian Belgodere* Milton Bonilla* Lan Hoang Danny Barnett I-Hsin Chung Apoorve Mohan Ming-Hung Chen Lixiang Luo Robert Walkup Constantinos Evangelinos Shweta Salaria Marc Dombrowa Yoonho Park Apo Kayi Liran Schour Alim Alim Ali Sydney Pavlos Maniotis Laurent Schares Bernard Metzler Bengi Karacali-Akyamac Sophia Wen Tatsuhiro Chiba Sunyanan Choochotkaew Takeshi Yoshimura Claudia Misale Tonia Elengikal Kevin O'Connor Zhuoran Liu Richard Molina Lars Schneidenbach James Caden Christopher Laibinis Carlos Fonseca Vasily Tarasov Swaminathan Sundararaman Frank Schmuck Scott Guthridge Jeremy Cohn Marc Eshel Paul Muench Runyu Liu William Pointer Drew Wyskida Bob Krull Ray Rose Brent Wolfe William Cornejo John Walter Colm Malone Clifford Perucci Frank Franco Nigel Hinds Bob Calio Pavel Druyan Robert Kilduff John Kienle Connor McStay Andrew Figueroa Matthew Connolly Edie Fost Gina Roma Jake Fonseca Ido Levy Michele Payne Ryan Schenkel Amir Malki Lion Schneider Aniruddha Narkhede Shekeba Moshref Alexandra Kisin Olga Dodin Bill Rippon Henry Wrieth John Ganci Johnny Colino Donna Habeger-Rose Rakesh Pandey Aditya Gidh Aditya Gaur Dennis Patterson Samsuddin Salmani Rambilas Varma Rumana Rumana Shubham Sharma Aditya Gaur Mayank Mishra Rameswar Panda Aditya Prasad Matt Stallone Gaoyuan Zhang Yikang Shen David Cox Ruchir Puri Dakshi Agrawal
IBM Research

Drew Thorstensen Joel Belog Brent Tang Saurabh Kumar Gupta Amitabha Biswas Anup Maheshwari Eran Gampel Jason Van Patten Matthew Runion Sai Kaki Yigal Bogin Brian Reitz Steve Pritko Shahan Najam Surya Nambala Radhika Chirra Rick Welp Frank DiMitri Felipe Telles Amilcar Arvelo King Chu Ed Seminaro Andrew Schram Felix Eickhoff William Hanson Eric Mckeever Dinakaran Joseph Piyush Chaudhary Piyush Shivam Puneet Chaudhary Wesley Jones Robert Guthrie Chris Bostic Rezaul Islam Steve Duersch Wayne Sawdon John Lewars Matthew Klos Michael Spriggs Bill McMillan George Gao
IBM Infrastructure

Ashish Kamra Gaurav Singh Marc Curry Tushar Katarki Joe Talerico Zenghui Shi Sai Sindhur Malleni Erwan Gallen
Red Hat

*Corresponding Authors:
tsgersho@us.ibm.com, sseelam@us.ibm.com, bmbelgod@us.ibm.com, bonillam@us.ibm.com

Thank you !

Sched.com - Rate & Feedback

Thank you to our Diamond Sponsors!

CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA

IBM

Red Hat

Thank you to our Platinum Sponsor!

Thank you to our Gold Sponsor!

CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA

Thank you to our Recording Sponsor!

CLOUD NATIVE & KUBERNETES AI DAY NORTH AMERICA