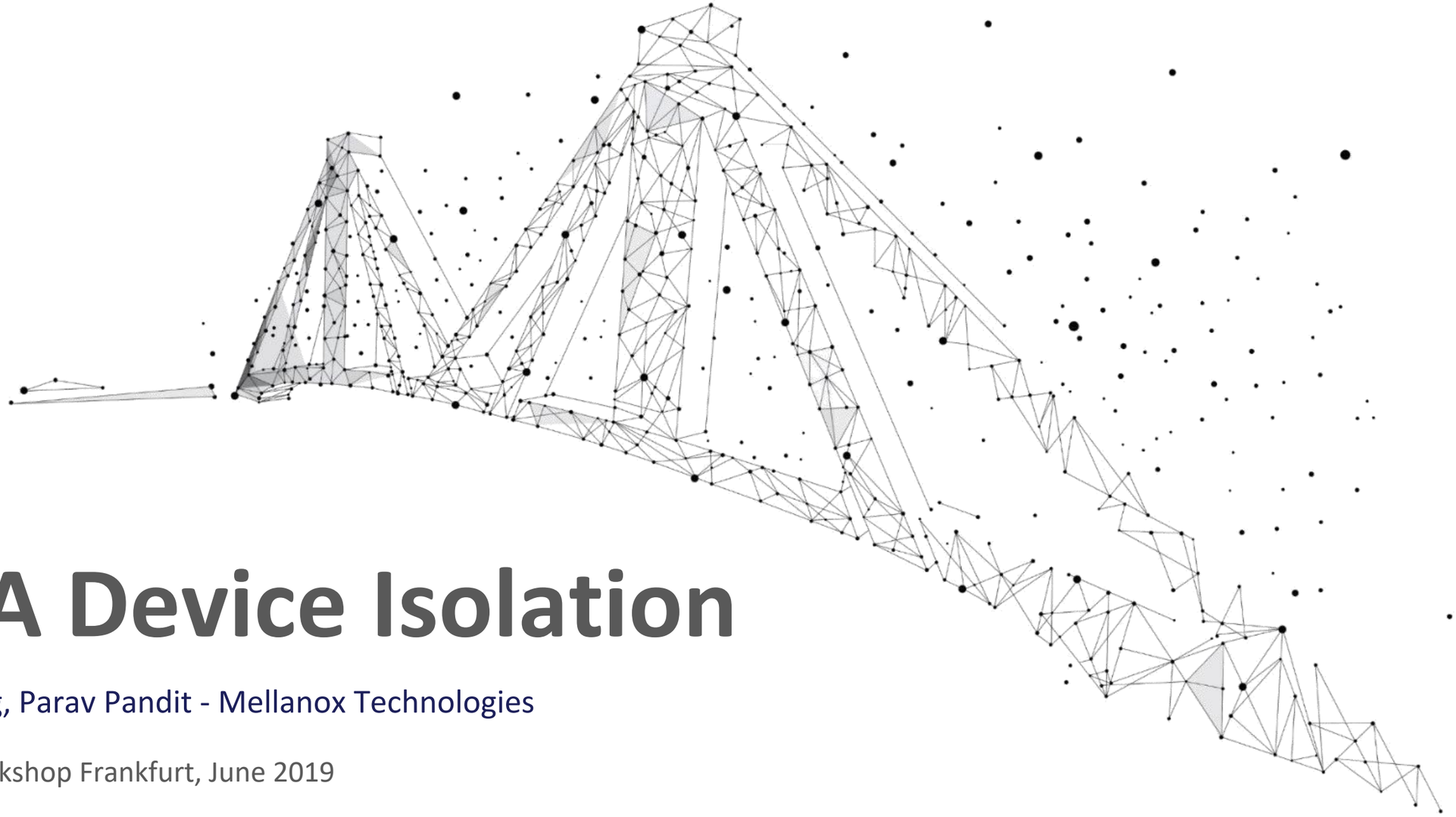
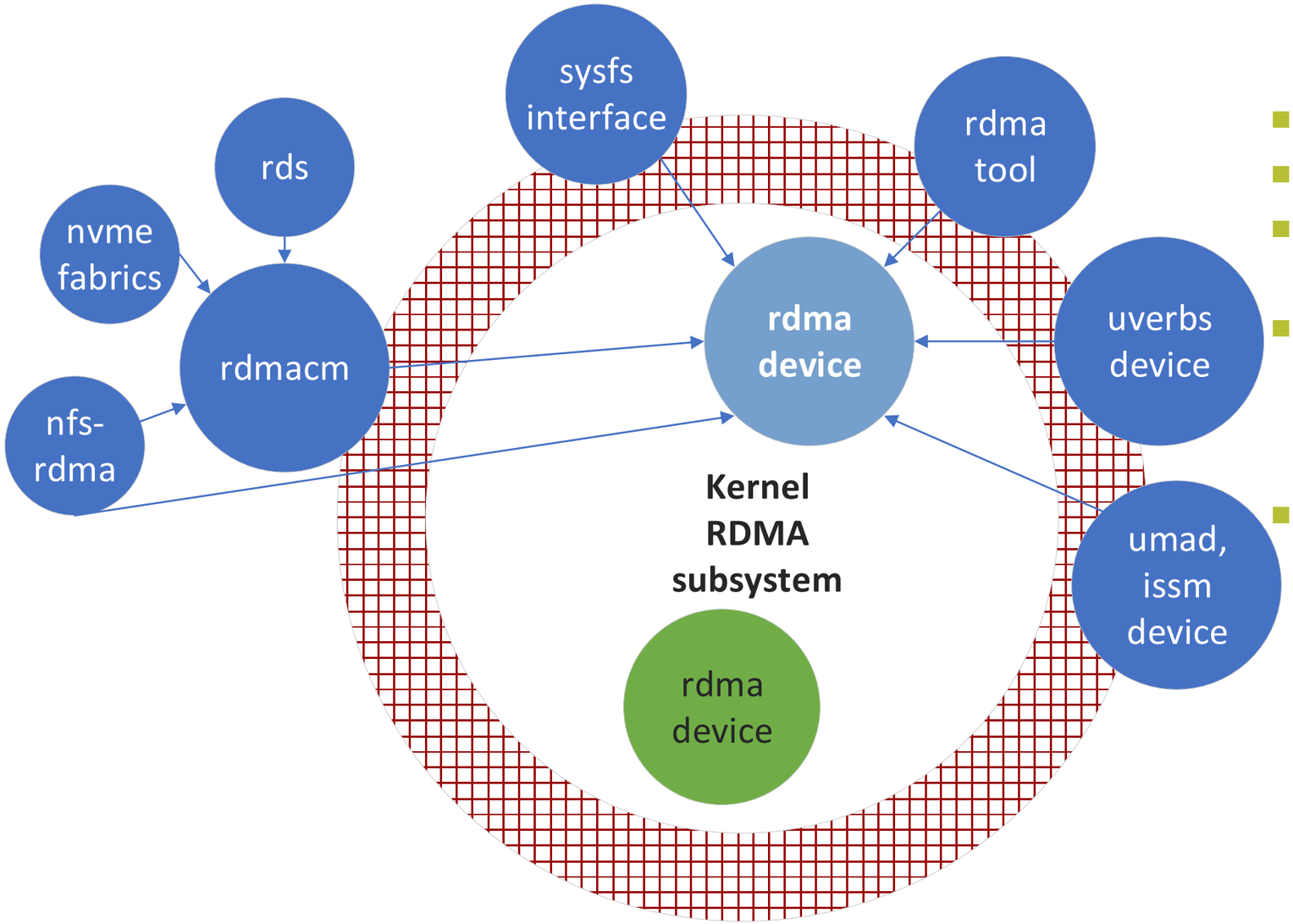


RDMA Device Isolation

Dror Goldenberg, Parav Pandit - Mellanox Technologies

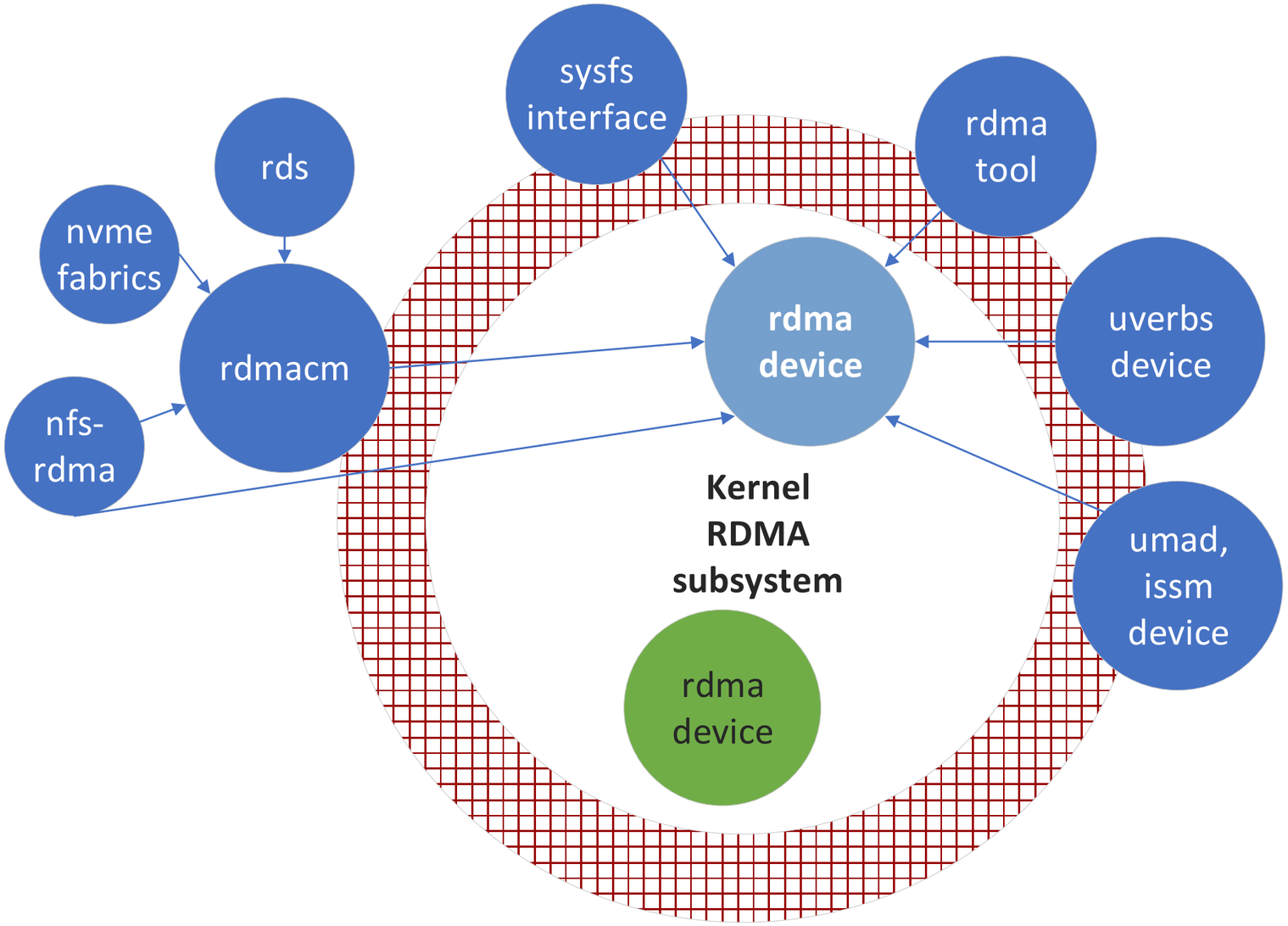
ISC Container Workshop Frankfurt, June 2019

RDMA Device Access Paths



- RDMA Connection Manager (CM)
- Verbs (via uverbs char device)
- Rdmatool (via netlink sockets)
  - Resource, connection information
- Sysfs file interface
  - Counters
  - Network addresses (IP/GID/LID)
  - More..
- UMAD char device
  - MAD packets
  - Device, address information

RDMA Device - The Need for Isolation

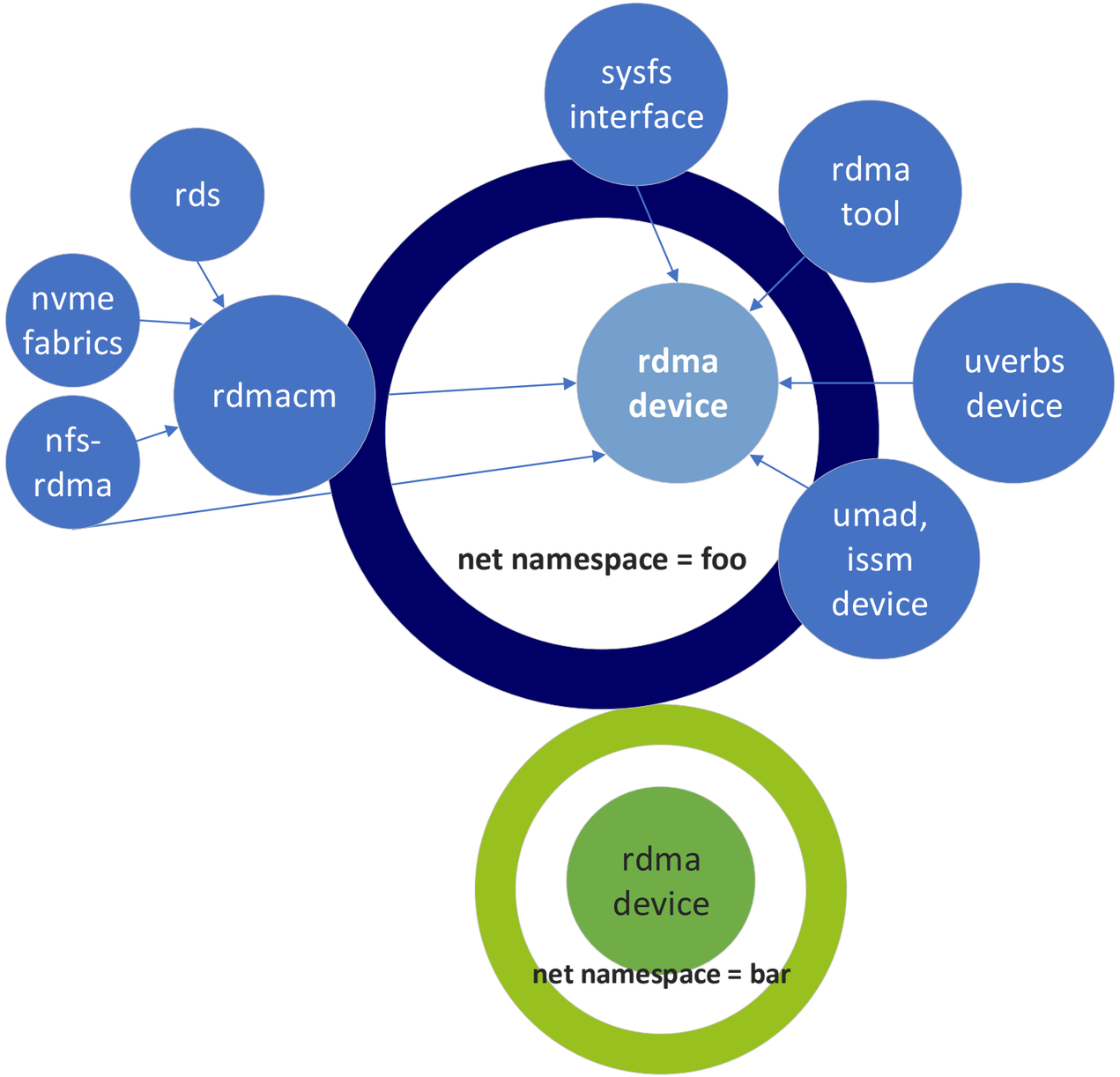


- Device cgroup - char devices ACL
  - Too coarse for network level
- RDMA cgroup - # of resources
  - Does not control the network access
- RDMA is yet another network device

Need to protect RDMA devices

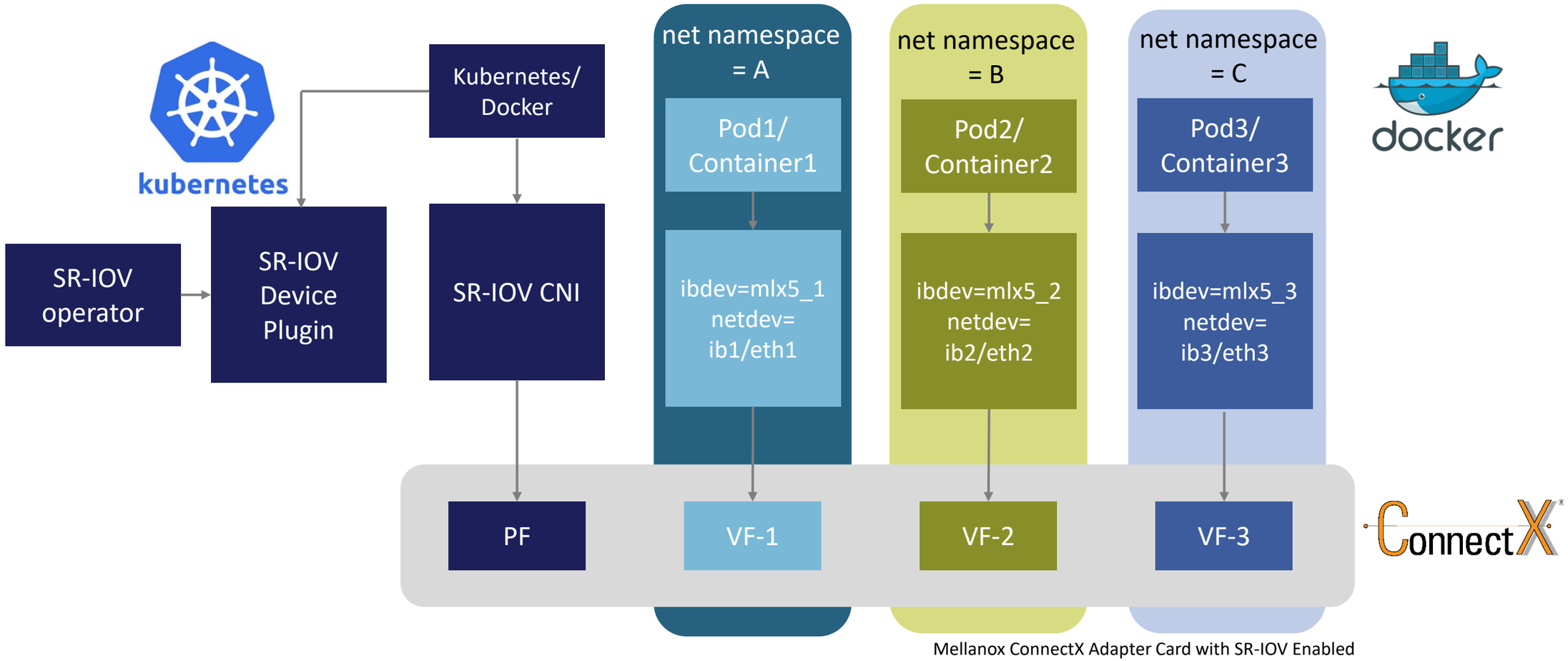
- At the network level
- In a reliable, unified, deterministic way
- Fit in existing orchestration frameworks (CNI, device plugin…)
- Future proof for new apps, interfaces, APIs
- Backward compatible

RDMA Devices in Net Namespace



- Isolation ring to access the RDMA device
  - Use existing net namespace of Linux kernel
- RDMA network namespace modes
  - Exclusive or shared
  - Via netlink
- Default as shared mode (backward compatible)
- RDMA device associated with net namespace
  - New netlink command
- Integrates with CNI and device plugin of K8s, Docker network plugin extension

RDMA Devices in Network Namespaces

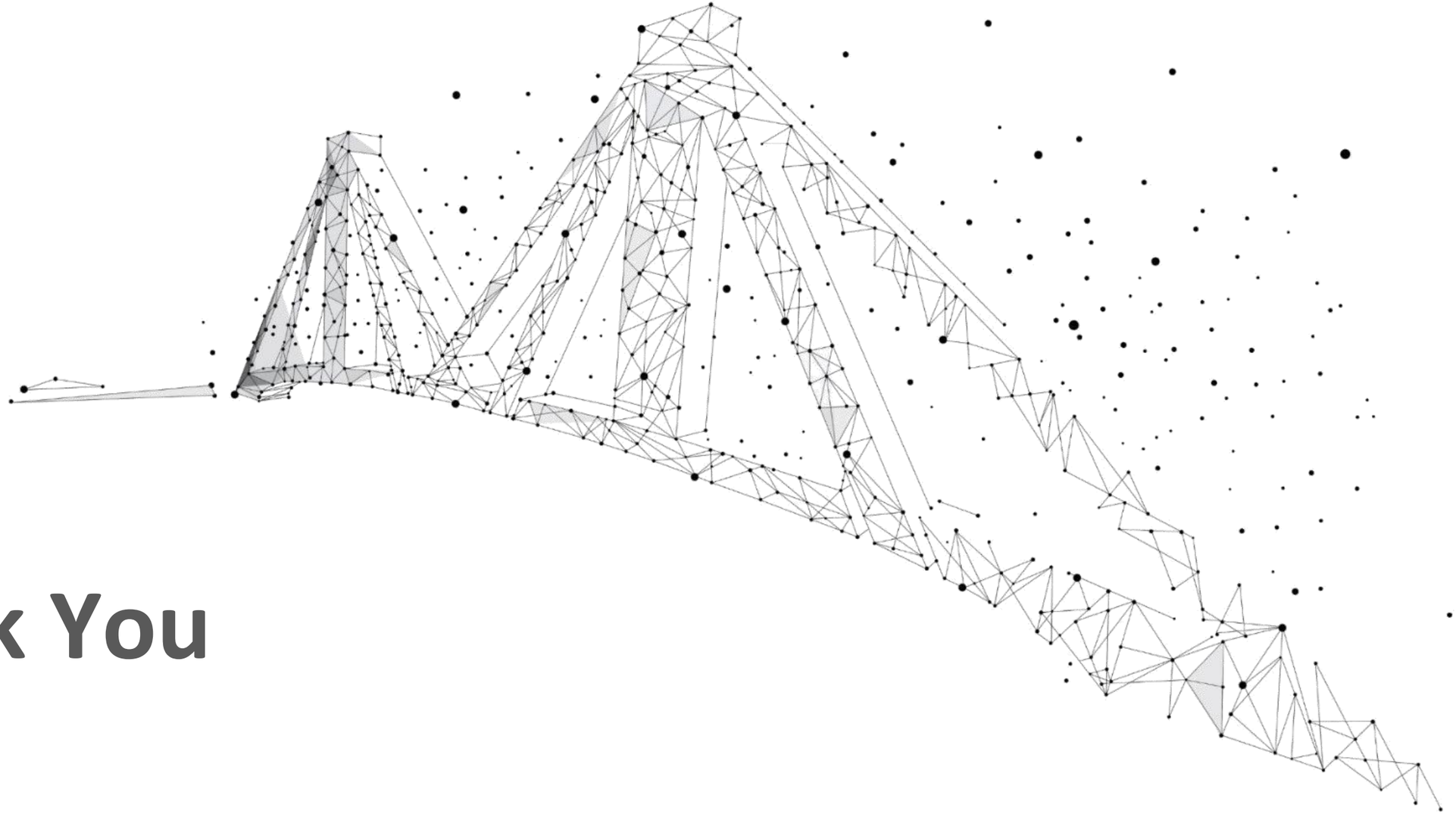


Mellanox ConnectX Adapter Card with SR-IOV Enabled

- Every container/POD has an IB device (mlx5_1,2,3) and netdevice
- Isolation is done on the net namespace level

Additional Information…

- Examples
  - Query, Change RDMA subsystem mode
    - $ rdma system show
    - $ rdma system set netns exclusive
  - Move RDMA device to new network namespace
    - $ ip netns add foo
    - $ rdma dev set mlx5_1 netns foo
- Current status (6/15/2019)
  - Merged to upcoming Linux kernel 5.2 and iproute2/rdma tool
  - Merged to netlink golang library
- Ahead of us
  - Integrate to docker sr-iov plugin
  - Integrate to SR-IOV operator and CNI plugin



Thank You