Mellanox Container Journey

Dror Goldenberg, Parav Pandit

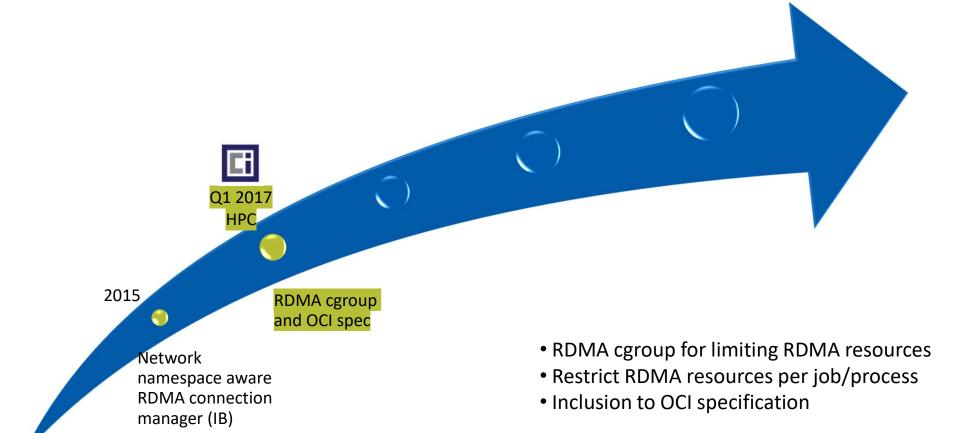Mellanox Technologies

June 2019

Start of RDMA container support

2015

Network namespace aware RDMA connection manager (IB)

- RDMA Connection manager supporting network namespaces
- InfiniBand and IPoIB netdevices

Extending with RDMA cgroup

Q1 2017
HPC

2015

RDMA cgroup
and OCI spec

Network
namespace aware
RDMA connection
manager (IB)

- RDMA cgroup for limiting RDMA resources
- Restrict RDMA resources per job/process
- Inclusion to OCI specification
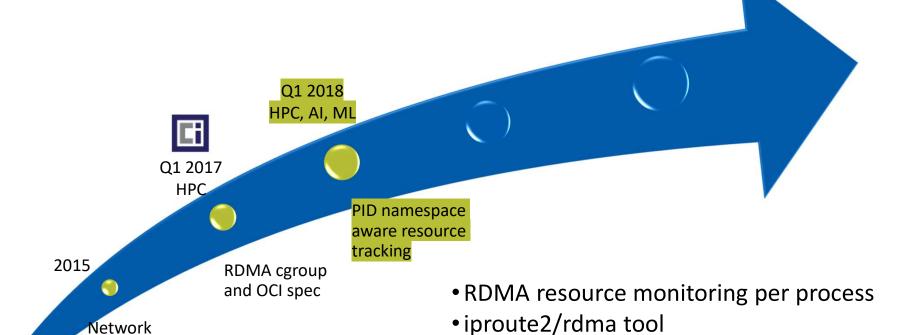
PID namespace isolation

Q1 2018
HPC, AI, ML

Q1 2017
HPC

PID namespace
aware resource
tracking

2015

RDMA cgroup
and OCI spec

Network
namespace aware
RDMA connection
manager (IB)

- RDMA resource monitoring per process
- iproute2/rdma tool
- RDMA resources isolated per processes of a PID namespace

RoCE network namespace support

Q2 2018
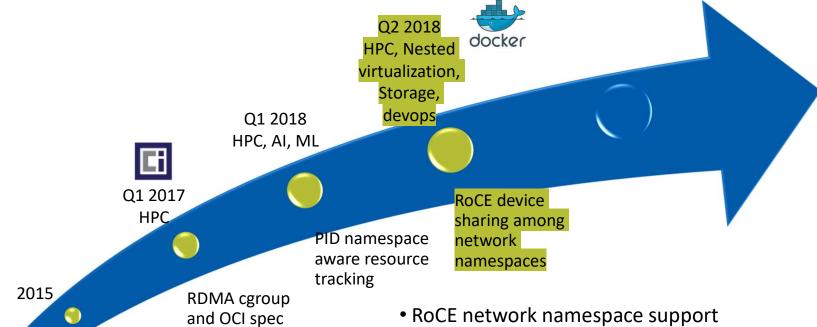HPC, Nested virtualization, Storage, devops

Q1 2018
HPC, AI, ML

Q1 2017
HPC

RoCE device sharing among network namespaces

PID namespace aware resource tracking

2015

RDMA cgroup and OCI spec

Network namespace aware RDMA connection manager (IB)
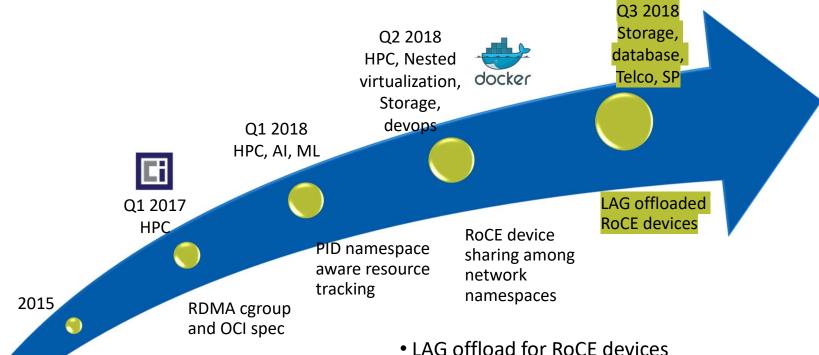
- RoCE network namespace support
- Single RDMA device shared among multiple net namespaces
- Docker default macvlan driver, ease of deploy
- Devops, development, nested virtualization with VM, Containers use case
- Docker SR-IOV network plugin

Link/Network Resilient RoCE LAG devices

Q3 2018
Storage,
database,
Telco, SP

Q2 2018
HPC, Nested
virtualization,
Storage,
devops

Q1 2018
HPC, AI, ML

Q1 2017
HPC

LAG offloaded
RoCE devices

2015

RoCE device
sharing among
network
namespaces

PID namespace
aware resource
tracking

RDMA cgroup
and OCI spec

Network
namespace aware
RDMA connection
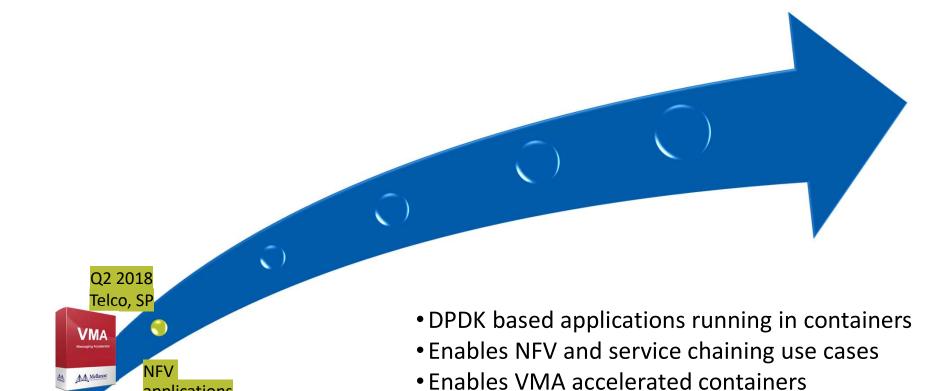manager (IB)

- LAG offload for RoCE devices
- Bandwidth scaling with single RDMA and netdevice
- Transparent to container applications
- Leverages existing SR-IOV infrastructure

Enable DPDK based applications in container

Q2 2018
Telco, SP

NFV
applications
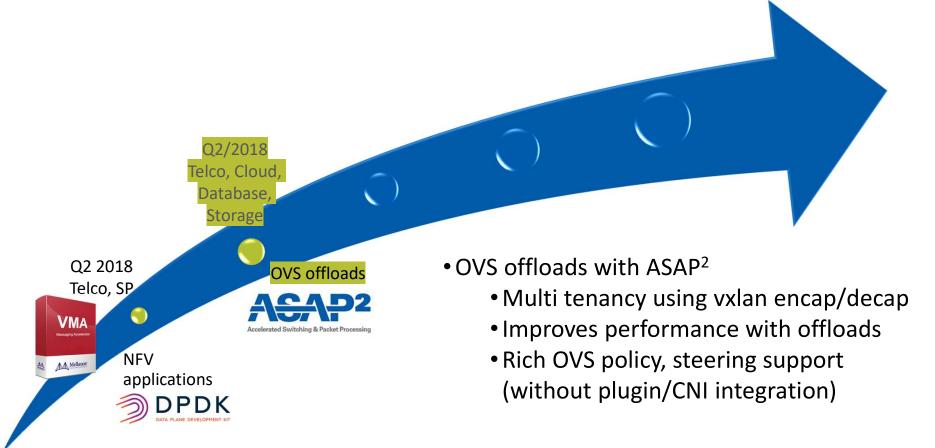
- DPDK based applications running in containers
- Enables NFV and service chaining use cases
- Enables VMA accelerated containers

OVS offloads using ASAP² technology

Q2/2018
Telco, Cloud,
Database,
Storage

Q2 2018
Telco, SP

VMA
Messaging Accelerator

NFV
applications

DPDK
DATA PLANE DEVELOPMENT KIT

OVS offloads

ASAP2
Accelerated Switching & Packet Processing

- OVS offloads with ASAP²
  - Multi tenancy using vxlan encap/decap
  - Improves performance with offloads
  - Rich OVS policy, steering support
    (without plugin/CNI integration)

Kubernetes support for RDMA

Q3/2018
HPC, AI, ML,
Database,
Cloud

Q2/2018
Telco, Cloud,
Database,
Storage

Q2 2018
Telco, SP

OVS offloads

K8s support

VMA
Messaging Accelerator

NFV
applications

DPDK
DATA PLANE DEVELOPMENT KIT

ASAP2
Accelerated Switching & Packet Processing

docker

CNI

- SR-IOV CNI plugin
- SR-IOV device plugin
- Shared mode HCA device plugin

RDMA device isolation in network namespace

May 2019
HPC, AI, ML, Database, Cloud

Q3/2018
HPC, AI, ML, Database, Cloud

Q2/2018
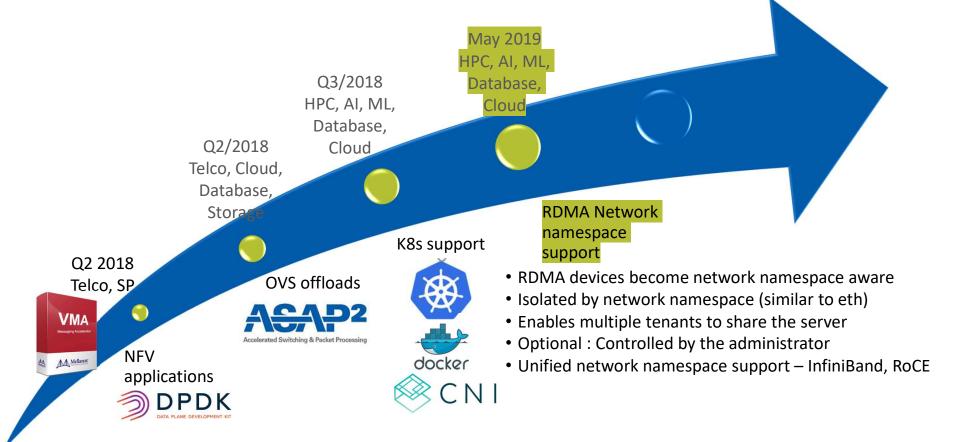Telco, Cloud, Database, Storage

RDMA Network namespace support

K8s support

Q2 2018
Telco, SP

VMA
Messaging Accelerator

OVS offloads

ASAP²
Accelerated Switching & Packet Processing

- RDMA devices become network namespace aware
- Isolated by network namespace (similar to eth)
- Enables multiple tenants to share the server
- Optional : Controlled by the administrator
- Unified network namespace support – InfiniBand, RoCE

NFV applications

DPDK
DATA PLANE DEVELOPMENT KIT

docker

CNI

Openshift device plugin and CNI integration

May 2019
HPC, AI, ML,
Database,
Cloud

Q3/2018
HPC, AI, ML,
Database,
Cloud

Q2/2018
Telco, Cloud,
Database,
Storage

RDMA Network
namespace
support

K8s support

Q2 2018
Telco, SP

OVS offloads

VMA
Messaging Accelerator

NFV
applications

DPDK
DATA PLANE DEVELOPMENT KIT

ASAP²
Accelerated Switching & Packet Processing

docker

CNI

- Redhat Openshift SR-IOV device plugin support for RDMA merged
- Enables Openshift Telco users to use DPDK and RDMA

Thank You