NGINX-Plus on Intel IPU managed at scale by Red Hat

Security Appliance for Enterprise Edge AI models

Intel:
Swati Mittal – Solutions Architect
Arun Paneri – SW Product Manager

Red Hat:
Balazs Nemeth, Phd – Senior Principal Software Engineer

F5:
Paul Pindell – Principal Architect,Technology Alliances

Solution Features

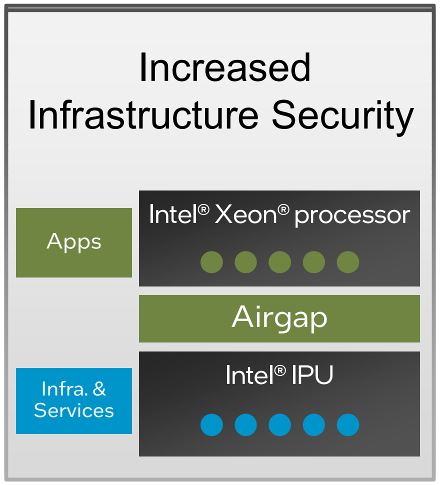
- ✓ Industry’s **first** OpenShift dpu-operator solution
- ✓ **IPU solutions ecosystem:** Dell, RedHat and F5
- ✓ **IPU capabilities** : IPU Service Function Chaining, IPU integration with MicroShift
- ✓ **Deploy ability**: Hands free deployment using RedHat dpu-operator
- ✓ **Monitoring:** A single pane of glass to monitor the workloads.
- ✓ **Availability:** Solution will be available as Tech Preview in OCP 4.19 – June 20

IPU Highlights

- Intel IPU E2100 features and capabilities
- Architecture with IPU on Red Hat OpenShift
- AI Inference Deployment with NGINX-plus offloaded to IPU
- Service Function Chaining on IPU with Red Hat OpenShift
- Demo of the solution

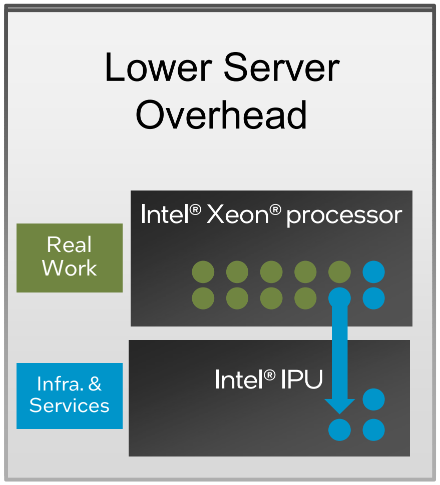
Intel® IPU Value Proposition

Security


Increased Infrastructure Security

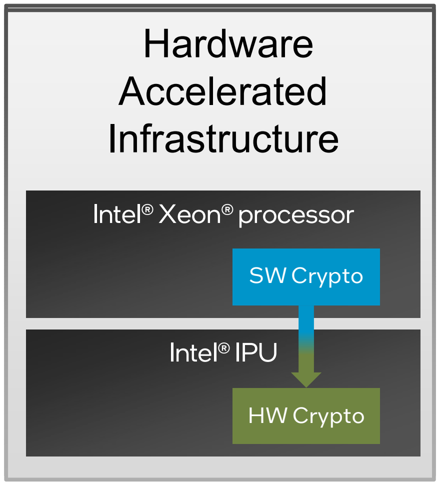
Application & Tenant Isolation from Infrastructure

Infrastructure Offload


Lower Server Overhead

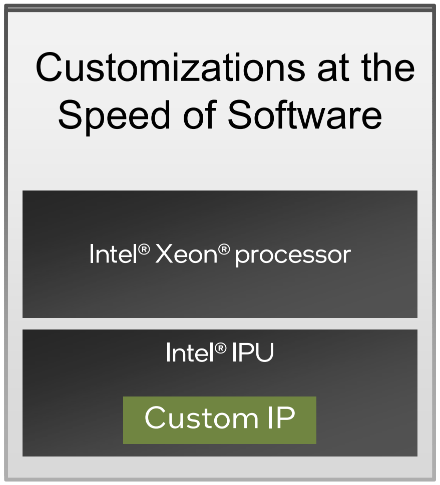
IPUs Reduce Host Compute Cycles Doing Infrastructure Work

Infrastructure Acceleration


Hardware Accelerated Infrastructure

IPUs Can Accelerate Some Applications

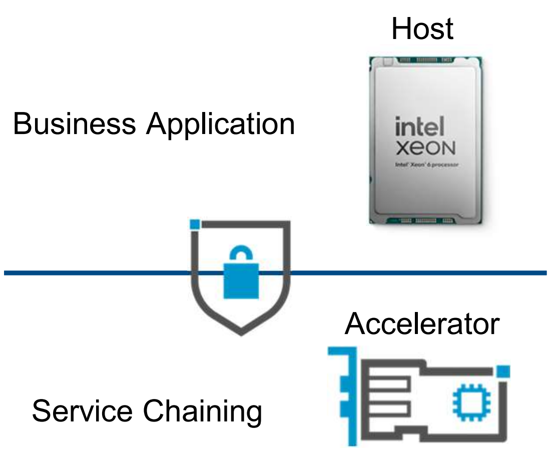
Feature Velocity


Customizations at the Speed of Software

IPUs Provide Reconfigurability and Programmability

Factors affecting Total Cost of Ownership (TCO)

Customer vs. Infrastructure apps



Host

Business Application

Service Chaining

Accelerator

Focus on what drives revenue

Enterprise Class Management

DMTF Redfish

Single Pane of Glass

OPEN PROGRAMMABLE INFRASTRUCTURE PROJECT

Red Hat Enterprise Linux

Everything at Scale

No vendor lock-in

Open and Heterogeneous

swati.mittal@intel.com

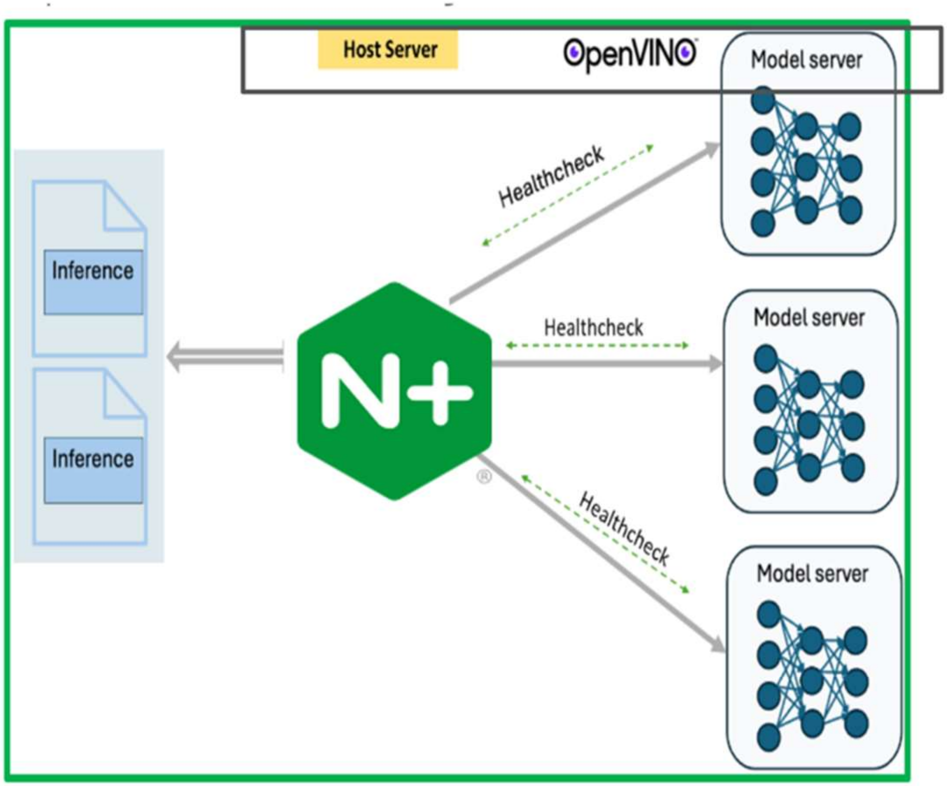
Leading Scalability and Efficiency

HOSTED	ACCELERATED	MANAGED	DELIVERED
DELL Technologies	intel®	Red Hat	f5®
Server with iDRAC monitoring	IPU E2100 MKP adapter OpenVINOModel servers	RHEL OS OpenShift Cluster and Microshift Custer	NGNIX-Plus

Hands—free Infrastructure Offload that seamlessly integrates into your Datacenter

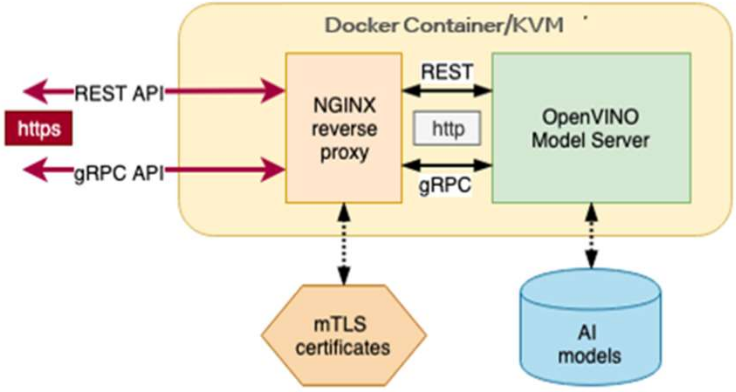
swati.mittal@intel.com

Securing AI models on the host (without IPU)



- Incoming client requests are reverse proxied by NGINX to different Open Vino Model servers on the host
- All NGINX related crypto ops are done on the host

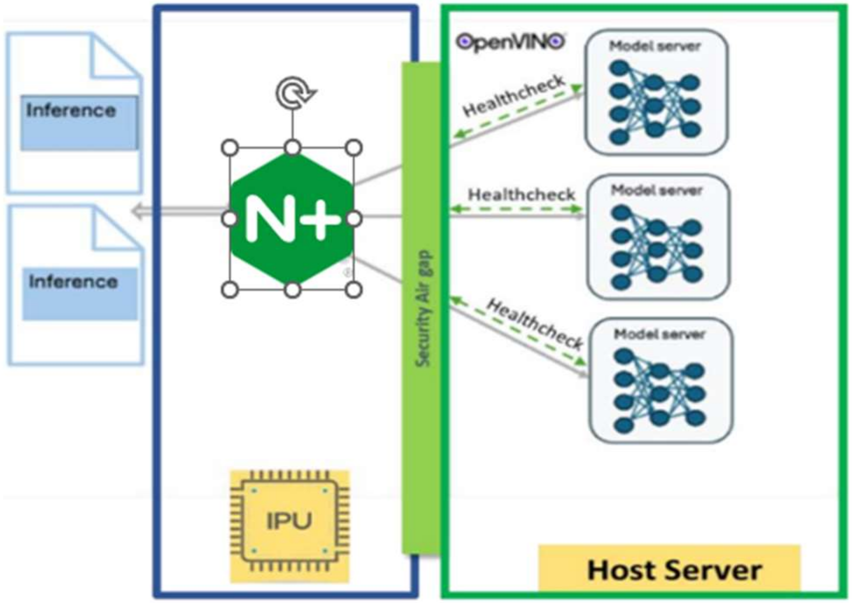
NGINX and OpenVino Model running in the same container



NGINX and Model Server on the same VM

Client authentications for incoming server inference requests done on the Host

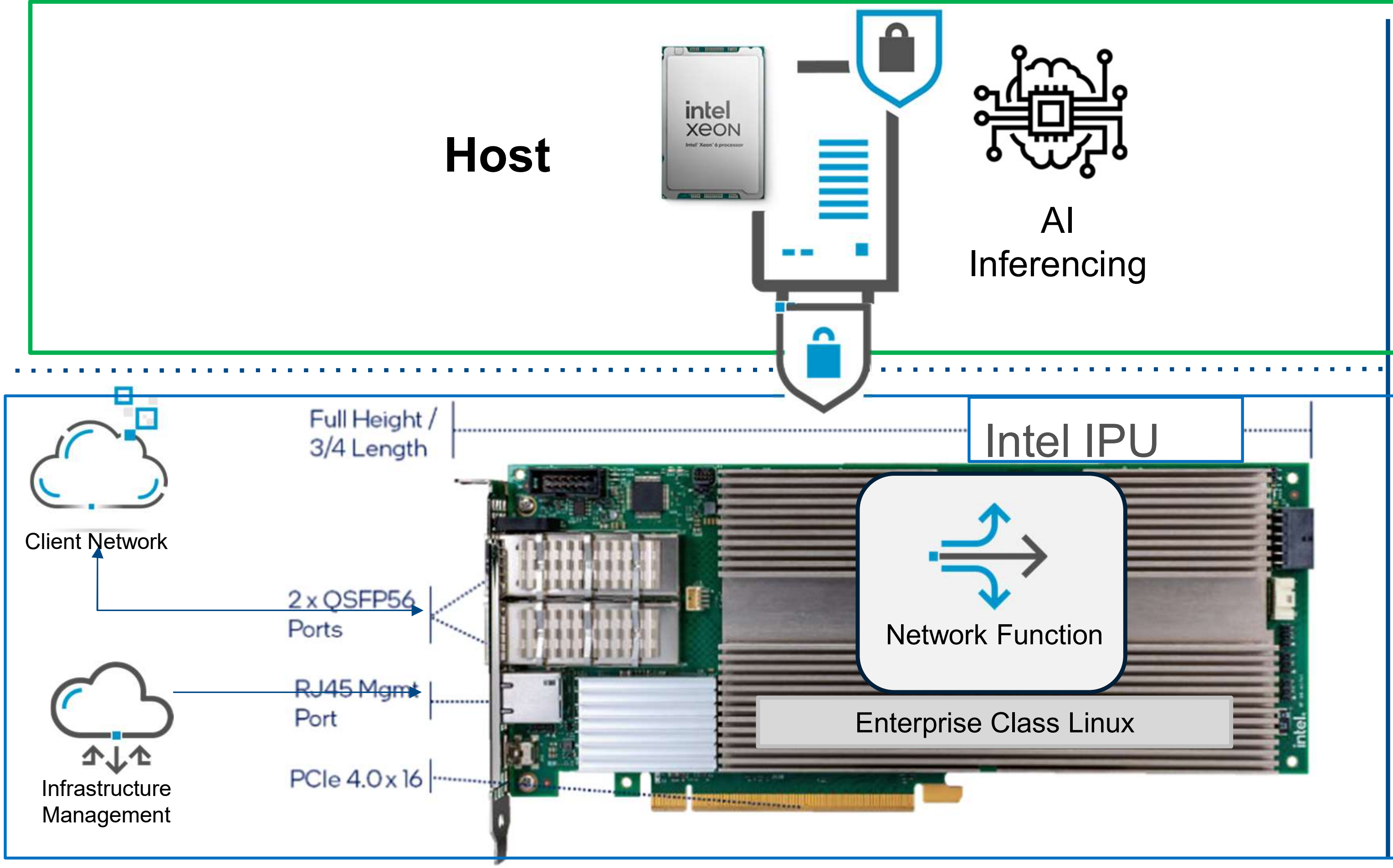
Securing AI models on the host with IPU



Security Air Gap between Model Server and NGINX

- NGINX user authentication related CPU intensive crypto ops run on IPU
- IPUs provide an isolated execution environment separate from the host's CPU
- Freed host CPU cycles available for AI/application workloads.

Solution for Edge AI inferencing

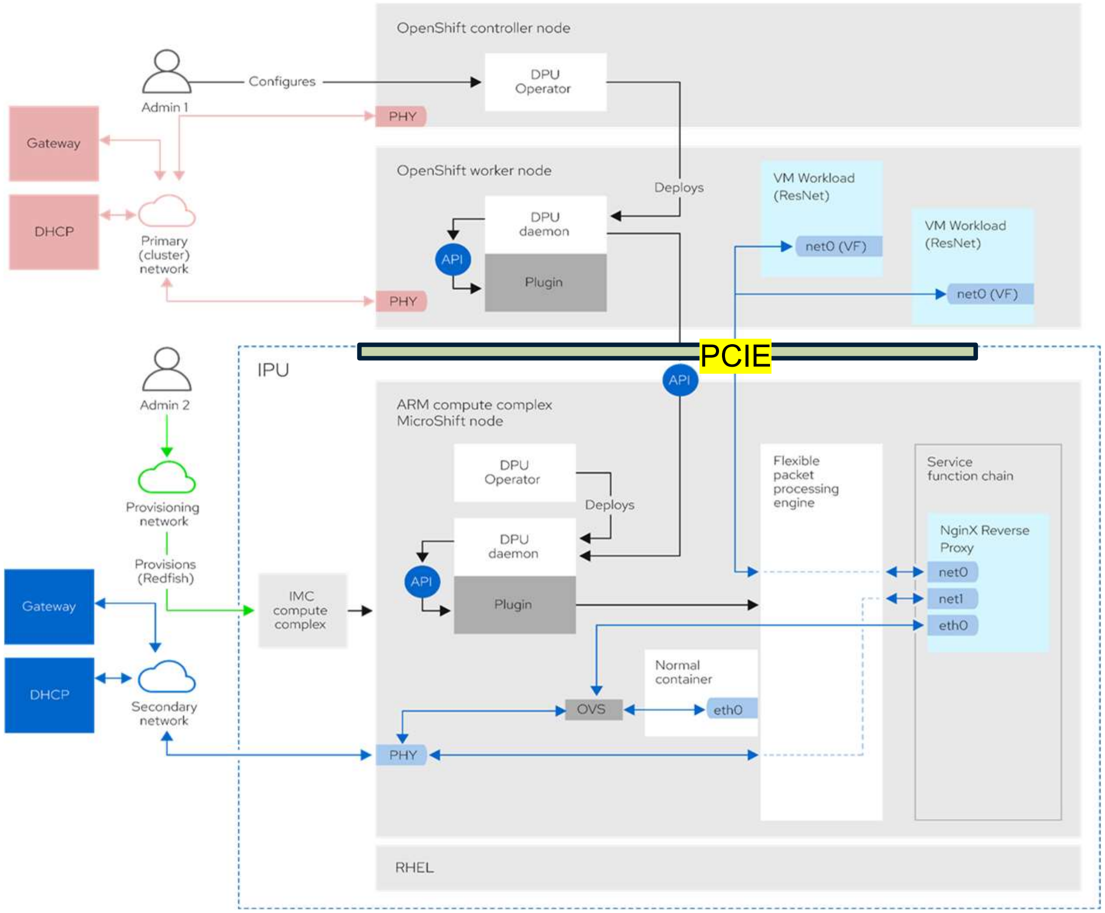


- Scale Out AI
- No changes to application
- Cluster Management

- Plug and Play
- Air gapped Infrastructure
- Saves CPU Cores
- Cluster management of IPU’s
- Enterprise ready OS

AI Inferencing On Scale with Offloaded NGINX Plus on IPU with OpenShift Cluster

- OpenShift uses the **dpu-operator** to run Infrastructure workloads directly on DPU's as **Containerized Network Functions** (CNFs).
- The dpu-operator programs the IPU's dedicated P4 Packet Processing Engine (via a **vendor plugin**) for accelerated packet processing at scale.



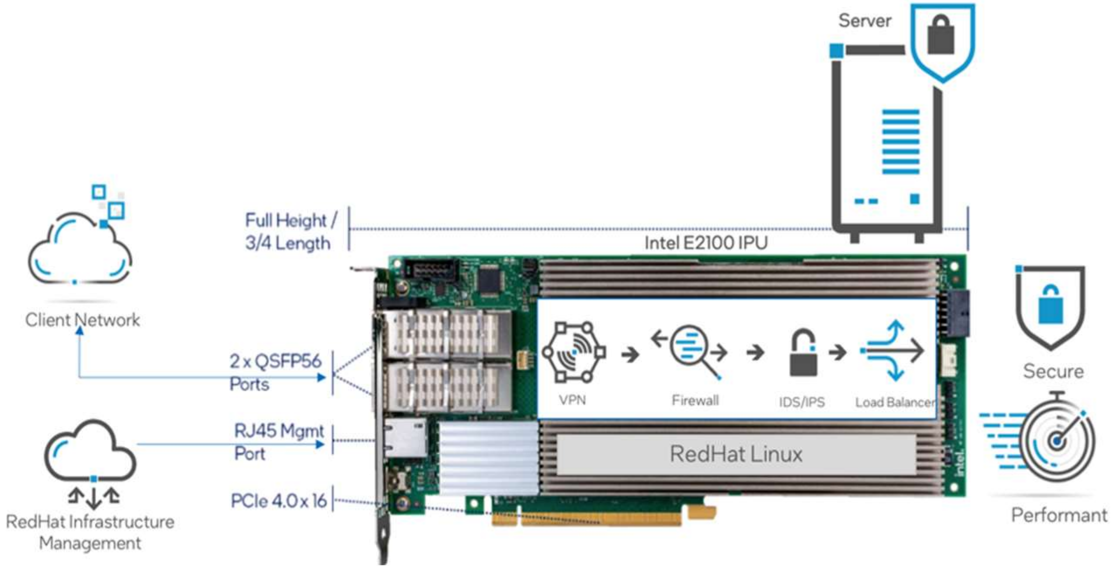
IPU Service Function Chaining (SFC)

- SFC chaining in HW and SW
  - Low latency
  - Optimized Traffic Flow
  - Flexibility
  - Improved efficiency
  - Automation and Orchestration with OpenShift
  - Support multi vendor applications
  - Reduces need for multiple appliances

Examples

- OVS offload (HW)-->Reverse proxy (SW)
- Firewall (HW) —> IDS (SW)→ Load balancers
- Load Balancers (SW) → Telemetry(SW)

OpenShift DPU operator enables SFC in HW and SW transparently to the user.

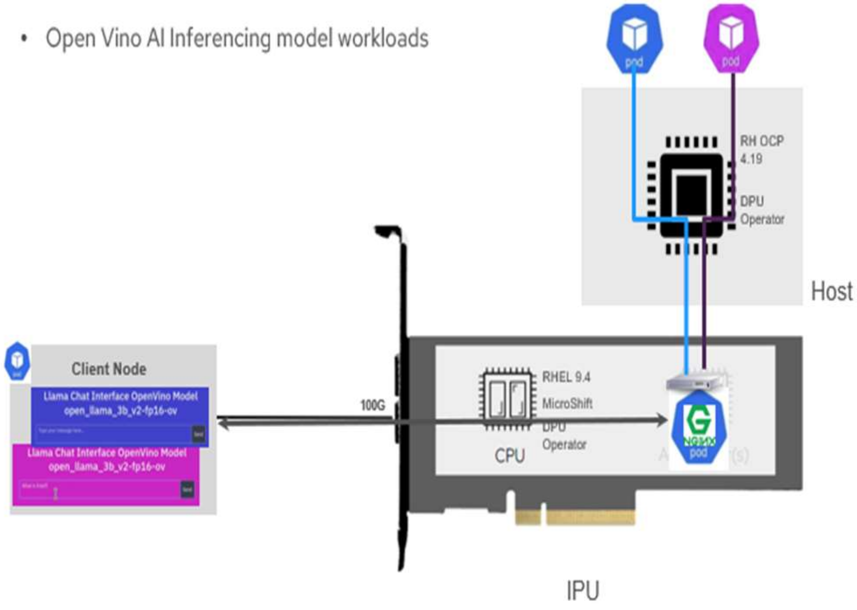


Intel® Infrastructure Processing Unit Adapter E2100-CCQDA2HL

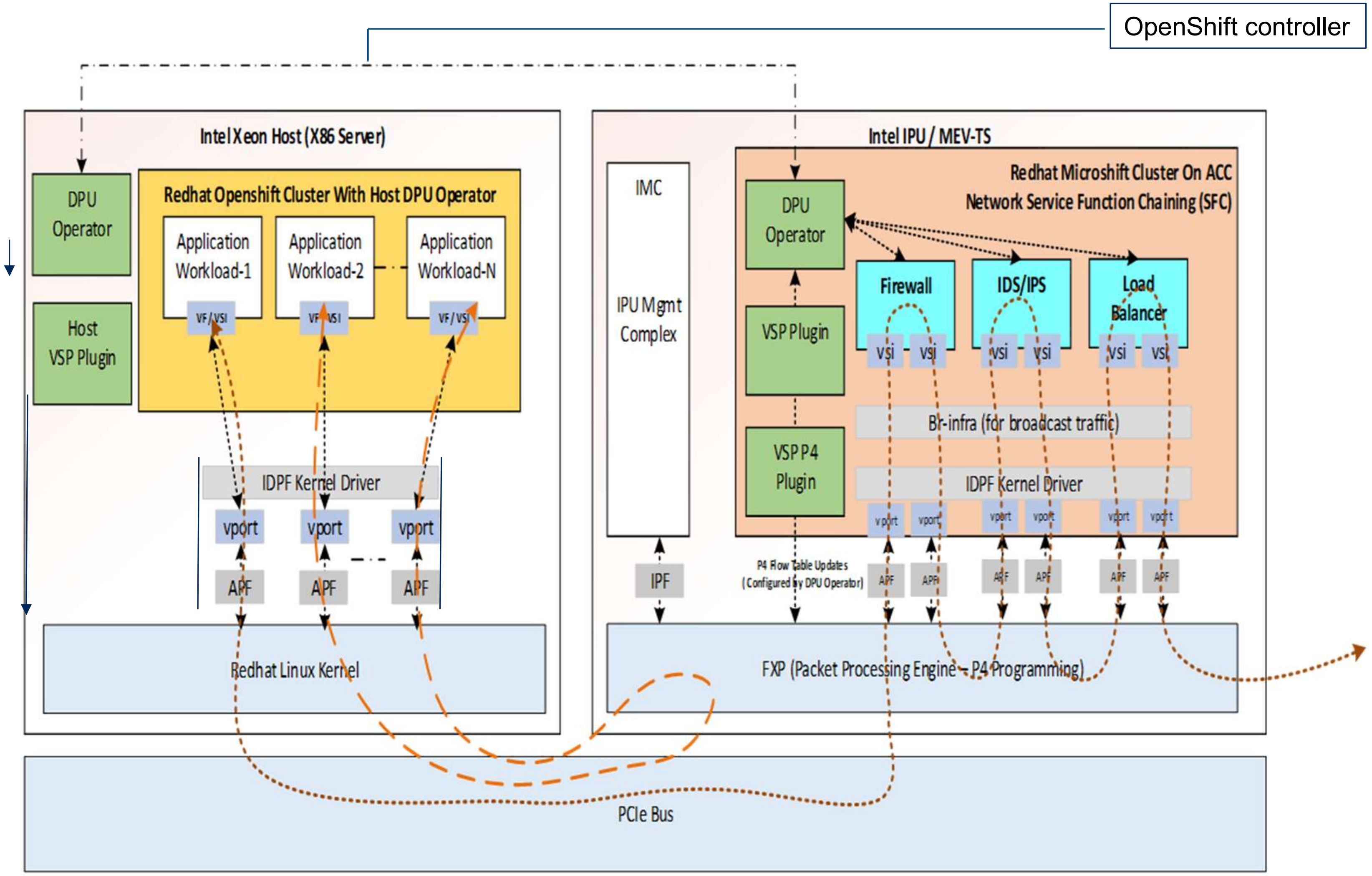
AI Inferencing at scale - Application Pod on host, Infrastructure pod on IPU

Deployment Steps

- Add the host node to the OpenShift cluster as a worker node.
- Add IPU to MicroShift cluster.
- Use OpenShift to manage IPU resources
- Deploy the NGINX pod onto the IPU.
- NGINX will reverse-proxy and load-balance remote client traffic to host pods.
- Run AI workloads on the host.
- Monitor both workloads via the OpenShift GUI.


• Open Vino AI Inferencing model workloads
pod
pod
RH OCP 4.19
DPU Operator
Host
Client Node
Llama Chat Interface OpenVino Model open_llama_3b_v2-fp16-ov
Send
Llama Chat Interface OpenVino Model open_llama_3b_v2-fp16-ov
Send
100G
RHEL 9.4
MicroShift
DPU Operator
CPU
NGINX
pod
IPU

Offload Network Functions on IPU: Deploy and Chain NFs with Red Hat OpenShift Operator



Thanks to the team

Team at Intel :

Swati Mittal, Naren Mididaddi, Arun Kumar V, Bandyopadhyay Sayan, Arun Paneri, Nishant Lodha, Scott Taylor

Team at F5 :

Paul Pindell, Sanjay Shitole

Team at RedHat:

Balazs Nemeth, Korry Nguyen

Deploy It Yourself: Intel IPU (Tech Preview) on OpenShift 4.19