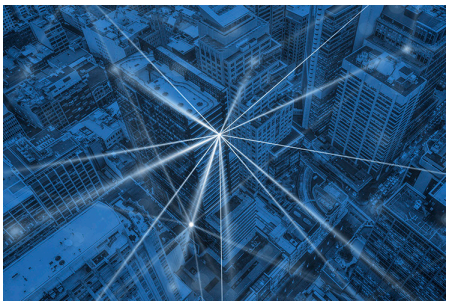
FEATURE BRIEF

Data Center Solutions
Kubernetes with Container Bare Metal

Enhanced Platform Awareness in Kubernetes

Enhanced Platform Awareness (EPA) is available in Kubernetes environments to improve performance and determinism for network functions virtualization (NFV) use cases



Enhanced Platform Awareness (EPA) represents a methodology targeting intelligent platform capability, configuration and capacity consumption. EPA delivers improved and deterministic application performance, and input/output throughput.

Key EPA capabilities are now available in Kubernetes to bring these benefits to container-deployed services. Some of these EPA features include Single Root I/O Virtualization (SR-IOV), CPU pinning (via CPU Manager for Kubernetes) and huge pages. Node Feature Discovery functionality has been added to Kubernetes to enable EPA features.

Node Feature Discovery identifies hardware features and advertises them to Kubernetes resource schedulers

Node Feature Discovery (NFD) works with the Kubernetes resource scheduler to identify server platform features and enable advanced Kubernetes scheduling. Important features such as SR-IOV, Intel® Advanced Vector Extensions 512 or Intel Turbo Boost may be available on certain nodes but will be not visible to the resource scheduler when it places applications.

An instance of NFD runs on each Kubernetes node and detects hardware features at the node level. Once it discovers the hardware features, the application assigns a label for each capability identified on the node and the availability of that capability. The labels are saved on the master and users of the system can request these labels via the Kubernetes Pod Specification so that an application lands on a node that has the required features.

SR-IOV provides dedicated VNF access to networking

One of the key capabilities detected by NFD is single root I/O virtualization (SR-IOV), a feature not available natively in Kubernetes. SR-IOV provides I/O virtualization that makes a single PCIe device (typically a network interface card (NIC)) appear as many network devices in the Linux kernel. In Kubernetes, this capability results in network connections that can be separately managed and assigned to different pods.

SR-IOV introduces the concept of virtual functions (VFs) that represent a regular PCIe physical function (PF) to a VNF. An example is a Kubernetes pod with multiple containers that are assigned access to their own Ethernet VF, which are all mapped to a specific physical Ethernet port. Performance is improved as packets move directly between the NIC and the pod.

In Kubernetes, SR-IOV is implemented by utilizing an SR-IOV Container Network Interface (CNI) plugin that lets the pod attach directly to the VF. For higher network performance, the DPDK mode in the SR-IOV CNI plugin allows the container to bind the VF to a Data Plane Development Kit (DPDK) driver. DPDK is an open source set of software drivers and libraries that accelerates the virtual network packet fast path.

CPU Manager for Kubernetes (CMK) delivers predictable network performance and workload prioritization

CPU pinning enables benefits such as maximizing cache utilization, eliminating operating system thread scheduling overhead as well as coordinating network I/O by guaranteeing resources. This delivers predictable performance, which is a key challenge of virtualization. Performance is impacted by an application being placed on different cores by the Linux scheduler in order to accommodate other applications running on the system.

CPU pinning creates an affinity between an application and a designated CPU core. A related feature is CPU isolation, which blocks other applications from running on that core. This is effective in isolating VNFs from "noisy neighbors," which are other applications that run on the same core and consume significant CPU cycles of the priority workload impacting performance.

In Kubernetes, CPU Manager for Kubernetes (CMK) manages these features. CMK is available as open source software on the Intel github. It supports a set of CPU management features including CPU pinning and isolation. CMK offers CPU pinning at the pod level and additionally offers thread-level process CPU affinity. There is an ongoing upstream development effort to add these features natively into Kubernetes. The result of phase 1 of this effort is native pod-level CPU pinning.

In a performance tests run using Intel® Xeon® Scalable processors[1], CPU pinning successfully improved packet throughput across a wide range of packet sizes (see Figure 1)

Native huge page management in Kubernetes enables the discovery, scheduling and allocation of huge pages as a resource

Intel Architecture processors leverage a page-based mechanism for converting virtual addresses to physical addresses. Huge page management is an EPA feature that is native to Kubernetes and increases supported page size from 4KB to up to 1GB.

The benefits of huge pages include reduced translation lookaside buffer (TLB) table size and improved lookup times and a reduction in TLB misses. These features provide a very positive impact on applications that require low latency access to memory, such as VNFs that leverage DPDK for high-throughput network performance. Huge page management can also benefit database applications because they require low-latency access to large amounts of memory.

Conclusion

EPA features including NFD, CMK, huge page support and SR/IOV deliver the high-performance packet processing that is critical to NFV workloads. As seen in Table 1, when these EPA features are applied, there is significant performance improvement across a wide range of packet sizes.

For more information on what Intel is doing with containers, go to https://networkbuilders.intel.com/network-technologies/intel-container-experience-kits.

Packet size	Throughput without CPU pinning and CPU isolation (average, Gbps)	Packet rate without CPU pinning and CPU isolation (average, Mpps)	Throughput with CPU pinning and CPU isolation (average, Gbps)	Packet rate with CPU pinning and CPU isolation (average, Mpps)	Performance gain
64B	6.59	9.85	12.20	18.30	185.27%
128B	11.00	9.31	20.36	17.22	185.03%
256B	10.19	4.61	18.86	8.56	185.12%
512B	13.48	3.18	22.79	5.35	169.15%
1024B	13.51	1.60	25.07	3.00	185.55%
1518B	13.51	1.10	25.10	2.00	185.77%

Figure 1: Benchmark tests showing performance improvement from CPU pinning and CPU isolation. See footnote for benchmark test system description.

[1] Performance testing conducted by Intel. Server included two Intel® Xeon® Gold Processor 6138T 2.00 GHz processors with 192GB (12 x 16 GB Micron) of RAM and two 25GbE ports on the Intel® Ethernet Network Adapter XXV710-DA2. Intel DC P3700 SSD storage was also part of the server.

SKU 336564-002 Enhanced Platform Awareness in Kubernetes - Feature Brief