Persistent Memory in Kubernetes*

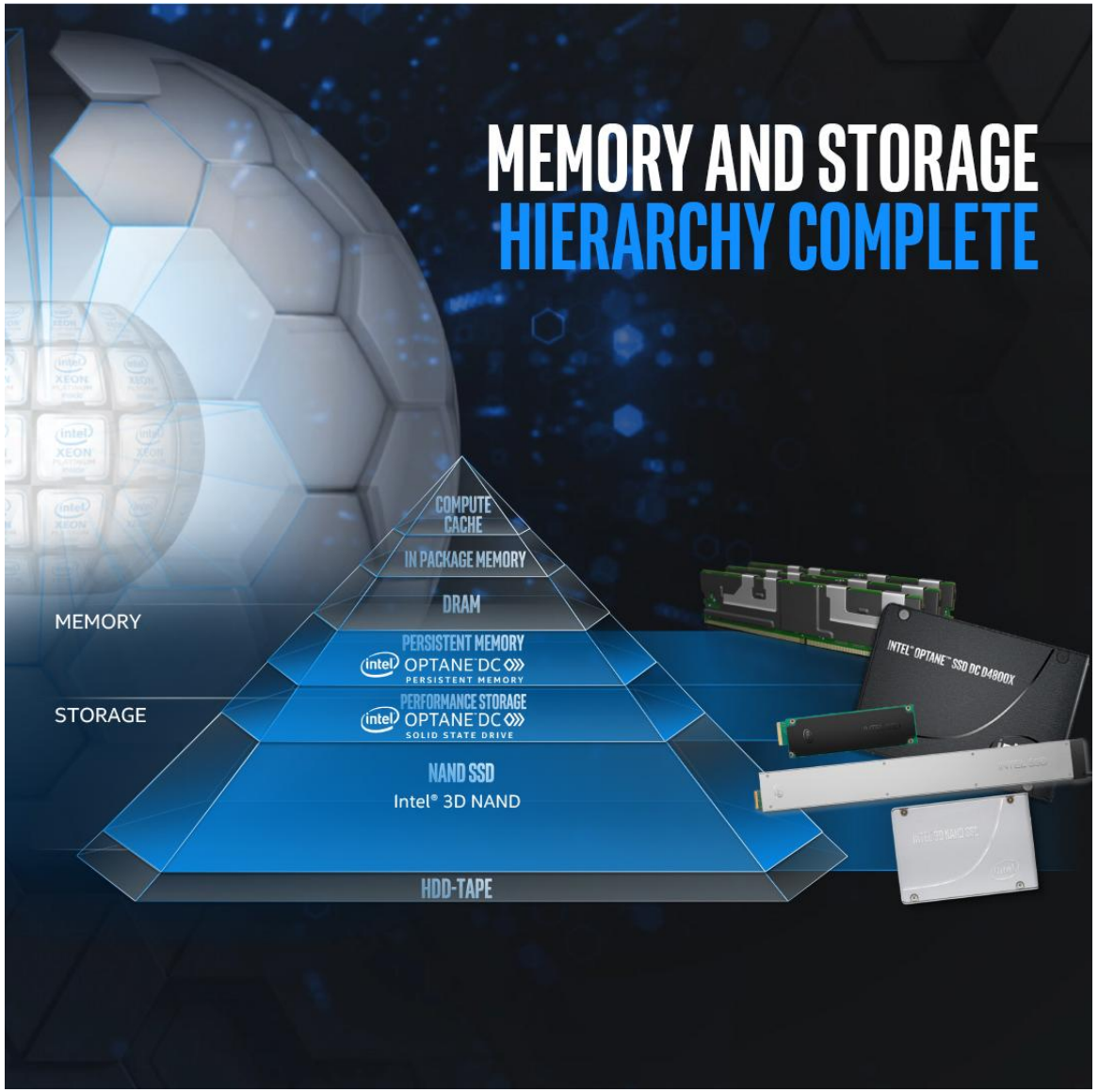
Patrick Ohly, Intel

*Other names and brands may be claimed as the property of others.

Persistent Memory (PMEM)

- Persistent
- Byte-addressable
- Read performance close to DRAM
- Higher capacity than DRAM

- Available in a DIMM form-factor
  - Intel® Optane™ persistent memory since 2019
  - Enhanced variants with higher bandwidth in 2020
  - Up to **6TB** in two-socket systems



Using PMEM

- Memory Mode:
  - Set up in BIOS,
    transparent to operating system and applications
  - DRAM used as cache for PMEM,
    not addressable separately

- App Direct Mode:
  - PMEM and DRAM both usable
  - Applications can:
    - choose where to store data
    - use persistency to speed up restarts
  - Example: memcached with restartable cache

- Storage over App Direct:
  - XFS and ext4 provide enhanced file IO

Resources

- `ipmctl`:
  - manage Intel® Optane™ persistent memory
- `ndctl`:
  - vendor-independent
  - manage **regions** and **namespaces**
- pmem.io:
  - Developer site
  - Persistent Memory Development Kit (PMDK)
  - libmemkind

Integration into Kubernetes

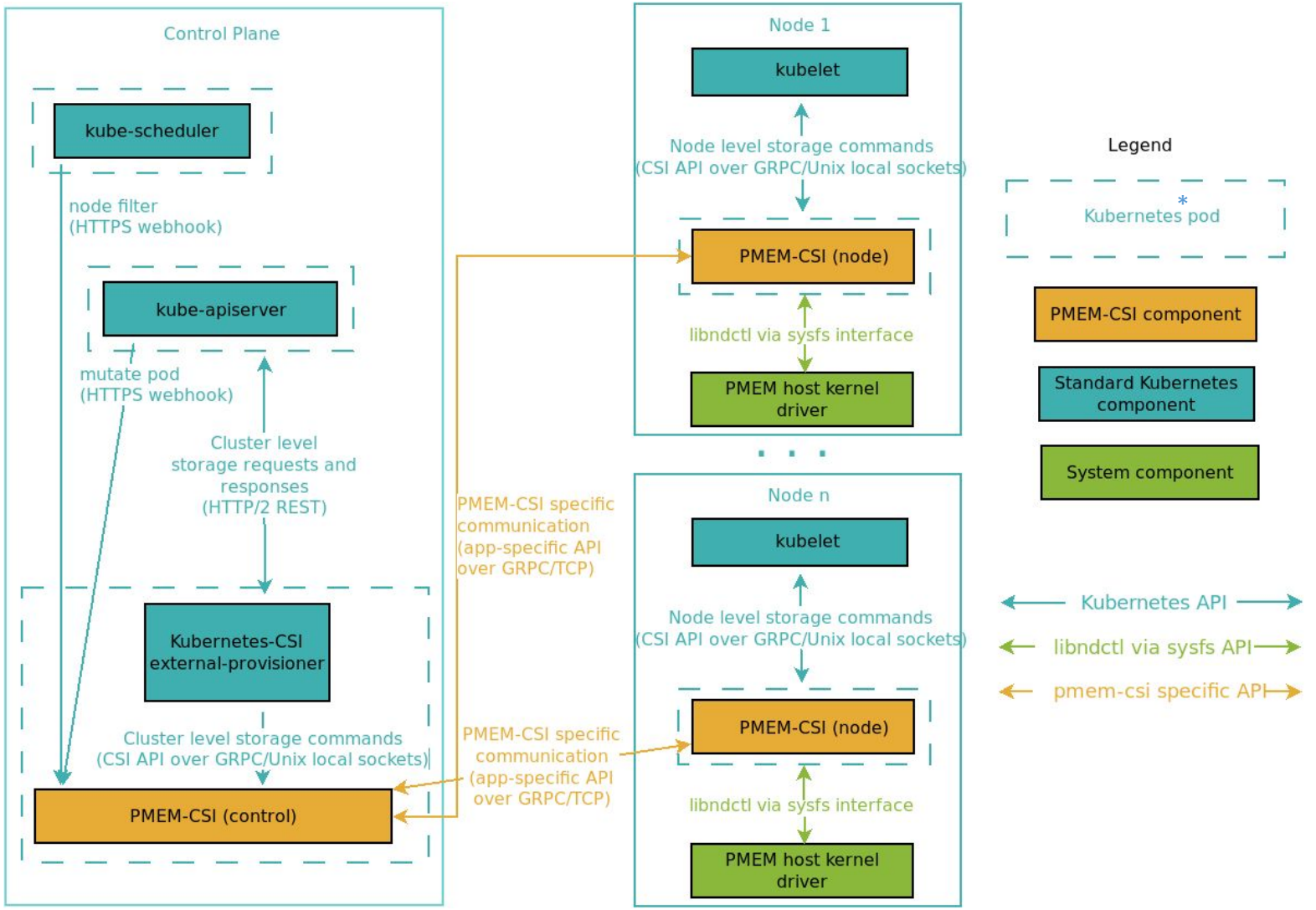
PMEM-CSI for AppDirect Mode
https://github.com/intel/pmem-csi

- Container Storage Interface (CSI) driver
- Dynamically provisions volumes with
  - libndctl:
    - as namespaces
    - may suffer from fragmentation
  - LVM:
    - as logical volumes
    - must allocate PMEM in advance
- Creates and mounts ext4 or XFS or provides raw block device

PMEM-CSI Status

- v0.5.0, August 2019:
  - First public release
  - Container images at https://hub.docker.com/u/intel
- v0.6.0, December 2019:
  - Raw block volumes
  - CSI ephemeral inline volumes:
    created and destroyed together with the pod,
    ideal for **local, non-persistent scratch space**
- v0.7.0, June 2020:
  - Installation via operator: simpler configuration and updates
  - Scheduler extensions: avoid nodes with insufficient storage
  - Kata Containers* are supported
- v0.8.0, October 2020:
  - The core features of PMEM-CSI are now production-ready:
    - tested on Kubernetes* 1.17, 1.18, 1.19
    - up- and downgrade testing, version skew testing
  - Metrics support for Prometheus*

Local Storage: PMEM-CSI



*Other names and brands may be claimed as the property of others.

Local Storage: Kubernetes*

Alpha features in 1.19:

- Storage Capacity Tracking:
  - Publish capacity information via apiserver
  - Use that information for pod scheduling

- Generic ephemeral volumes:
  - Works with unmodified CSI drivers
  - Supports storage capacity tracking, snapshot restore, cloning, …

Under investigation:

- external-provisioner on each node

Call to Action

- Watch some PMEM-CSI demos
  https://01.org/kubernetes/demos

- Try out PMEM-CSI
  - on a virtual QEMU cluster
  - with the memcached example

- Try the new Kubernetes* 1.19 alpha features

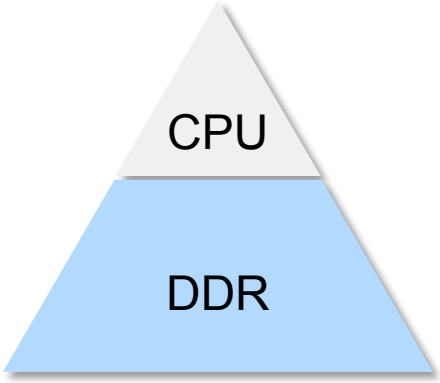
- Provide feedback:
  - patrick.ohly@intel.com
  - Kubernetes Slack, #csi and #sig-storage

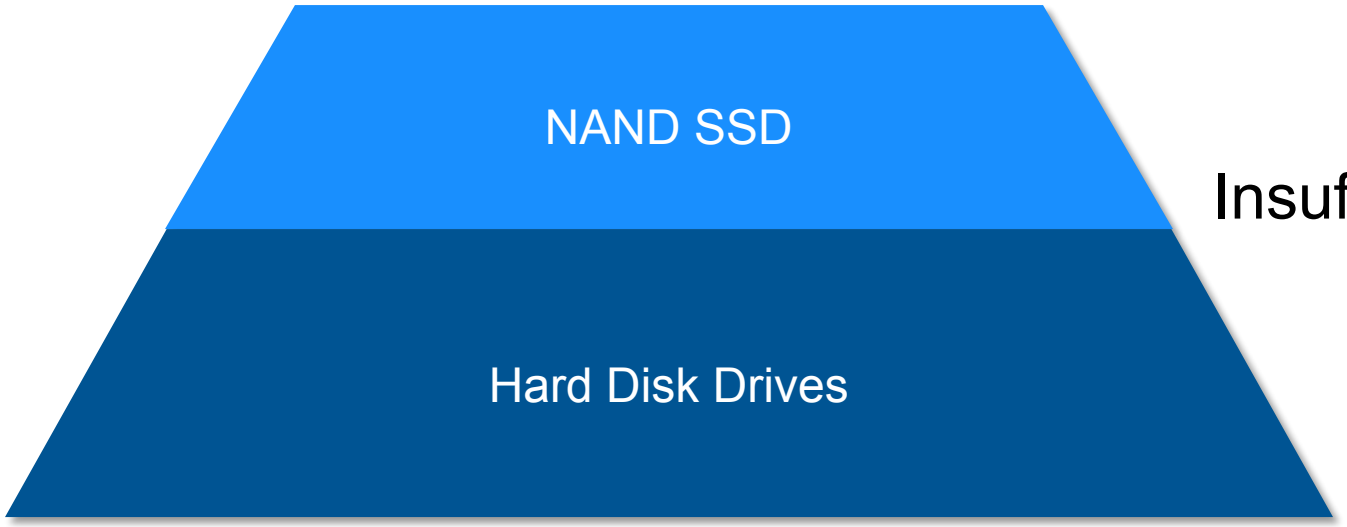
Notices and Disclaimers


KEEP CLOUD NATIVE
EVERYWHERE
KubeCon
CloudNativeCon
North America 2020
Virtual

The Memory/Storage Gap

Expensive
Capacity limited
No persistency





Latency
Insufficient performance