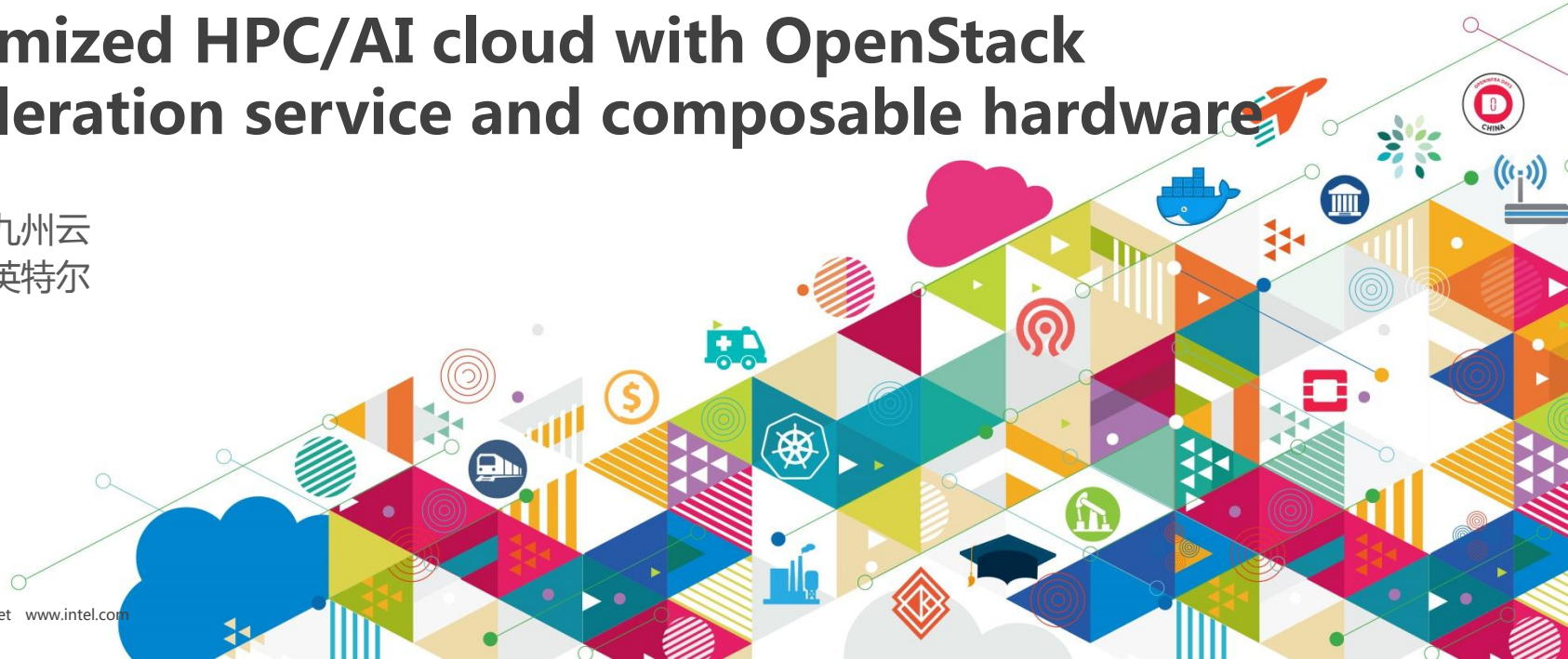
CHINA OpenInfra Days

CHINA OpenInfra Days Cloud | IT大咖说 知识共享平台

# Optimized HPC/AI cloud with OpenStack acceleration service and composable hardware

黄舒泉, 九州云
丁建峰, 英特尔



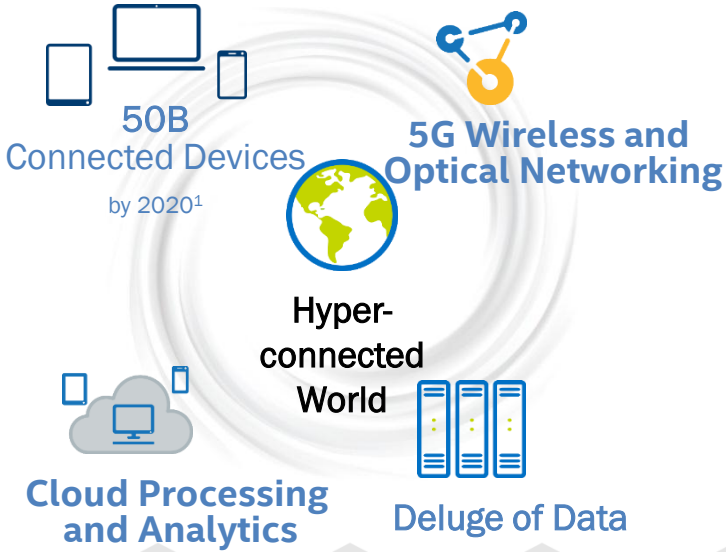
# Agenda

- AI/HPC Workload Challenges
- Dynamically aggregate hardware & accelerator – Intel® RSD
- OpenStack Acceleration Service – Cyborg
- Future Plan

# OpenStack for Scientific Research



- HPC/AI workloads make cloud acceleration a requirement rather than an interesting option.
- Hardware acceleration isn't new. GPU, ASIC, NVMe, FPGA, etc.
- Current issues to use accelerators in OpenStack



# Intel® RSD Vision



**Resource Pooling**
- NVMe over PCIe
- NVMe over Fabrics
- FPGA Accelerators

**Enable Solutions**
- Partner with OEMs & ISVs
- Interface with Orchestrators
- openstack / kubernetes
- Build Validated Solution Stacks

**Implement Standards**
- Compute and Rack API
- DMTF Redfish
- Storage Management API
- SNIA Swordfish
- Network Device Management API

Better Utilization
Greater Flexibility

\+

More Vendor choice
Lower cost of ownership

Benefits increase over time

# Intel® RSD Roadmap



Any forward looking information provided here is subject to change without notice

## 2016

### Intel RSD v1.2

- Open and Modern HW Management APIs (Redfish*)
- Pod-Level Architecture APIs
- Networked-Storage Services APIs

**MODERN MANAGABILITY**

Available Now

## 2017

### Intel RSD v2.1

- Physical Storage Disaggregation and Composability APIs
- Storage Pooling over PCIe (Direct-Attach)
- Pooled Node Controller Discovery

**NVMe POOLING OVER PCIe**

Available Now

### Intel RSD v2.2

- Intel Xeon Scalable Processor Support
- Out-of-Band Telemetry APIs
- TPM Support
- FPGA Discovery over PCIe

**ADVANCED MANAGABILITY**

Available Now

## 2018

### Intel RSD v2.3

- NVMe over Fabrics* (Ethernet, RDMA)
- Standards-Based Storage Mgmt. (SNIA Swordfish*)
- Telemetry for NVMe over Fabrics*

**NVMe POOLING OVER ETHERNET**

In Development
Target: Q2 2018

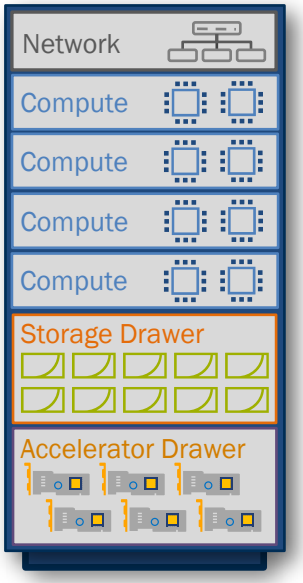
### *Features under investigation:*

- Intel persistent memory support
- FPGA Accelerator Pooling over PCIe
- Network card pooling
- Standards-Based Network Mgmt. *(Yang-to-Redfish*)*

** Other names and brands may be claimed as property of others.*
*** Dates above refer to delivery of Intel RSD reference code. OEM delivery of solutions are typically launched 1-2 quarters later*

# Intel® RSD - Disaggregation

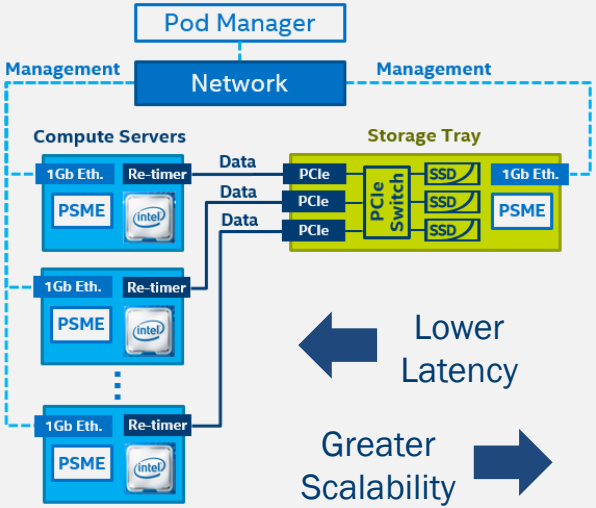
## Disaggregation



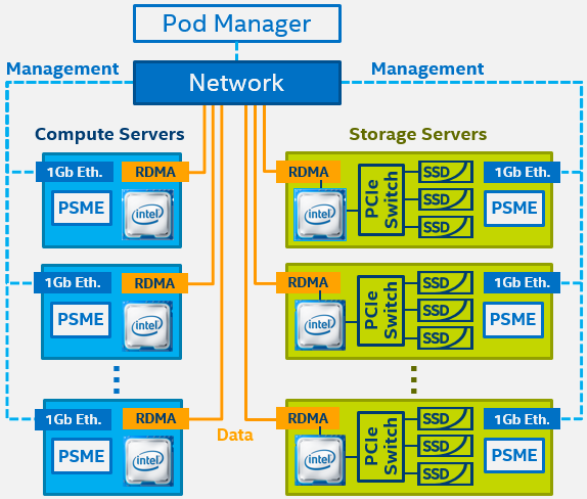

Spend less up front and save $$ over time

## Two implementation options for hardware disaggregation

### Storage Pooling over PCIe




### NVMe over Fabrics (Ethernet)



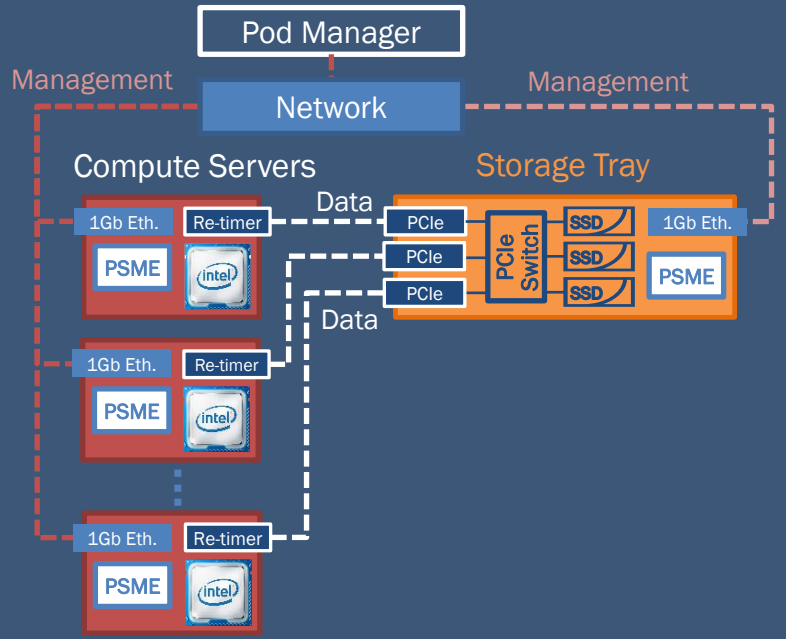
# Storage Pooling – PCIe Direct Attach



**Direct Attach PCIe Storage Pooling**
requires PCIe Re-timer cards in Compute Servers

| | |
|---|---|
| Storage Pool | 16 drives typically |
| Compute Radix | Generally 8-16 x4 |
| Latency | Add ~1 µSec per I/O |
| Bandwidth | <50% oversubscribed |
| Use Case | Hottest Tier Storage |
| Cost | Low |

Shipping NOW with v2.1



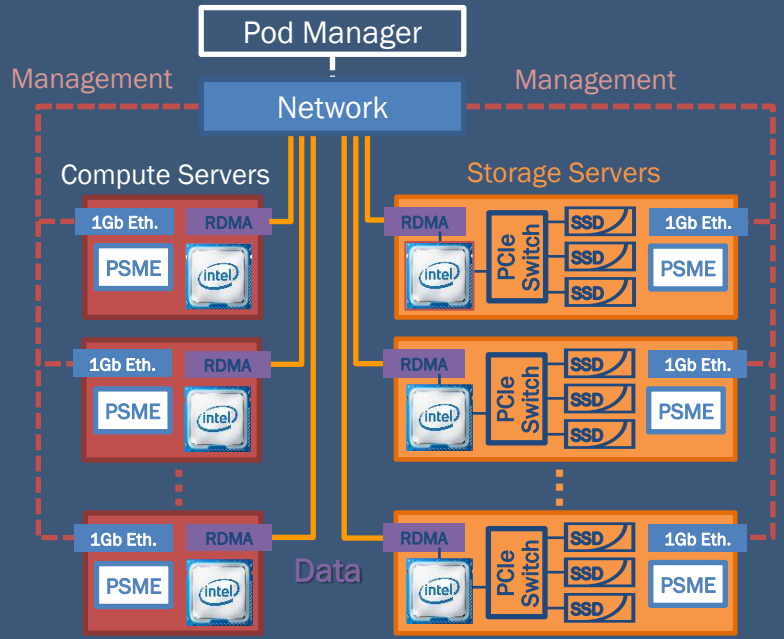
# Storage Pooling – NVMe over Fabric



**Software-based** Storage Pooling over Ethernet using "NVMe over Fabrics" protocol

| | |
|---|---|
| Storage Pool | Unlimited |
| Compute Radix | Unlimited |
| Latency | Add ~20 µSec per I/O |
| Bandwidth | May be oversubscribed |
| Use Case | Warm Tier Storage |
| Cost | Medium |

Coming in **v2.3**

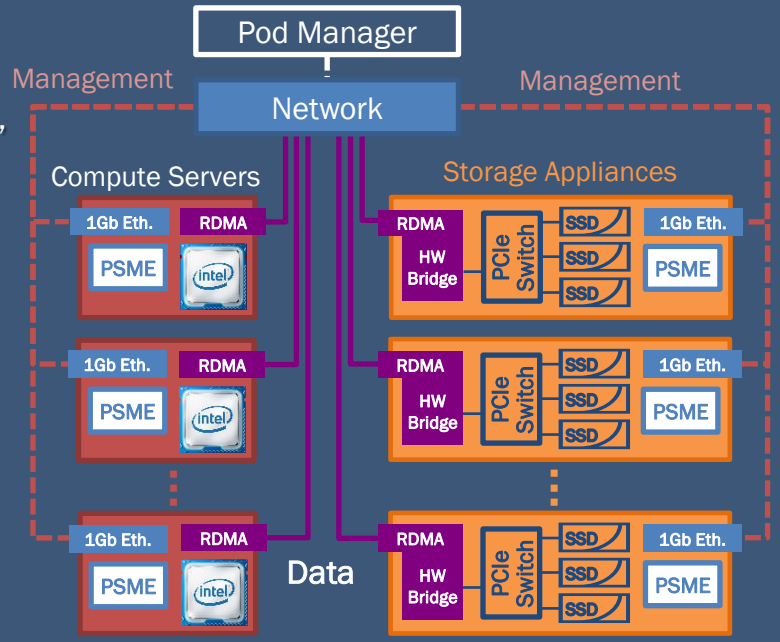


# Storage Pooling – NVMe over Fabric



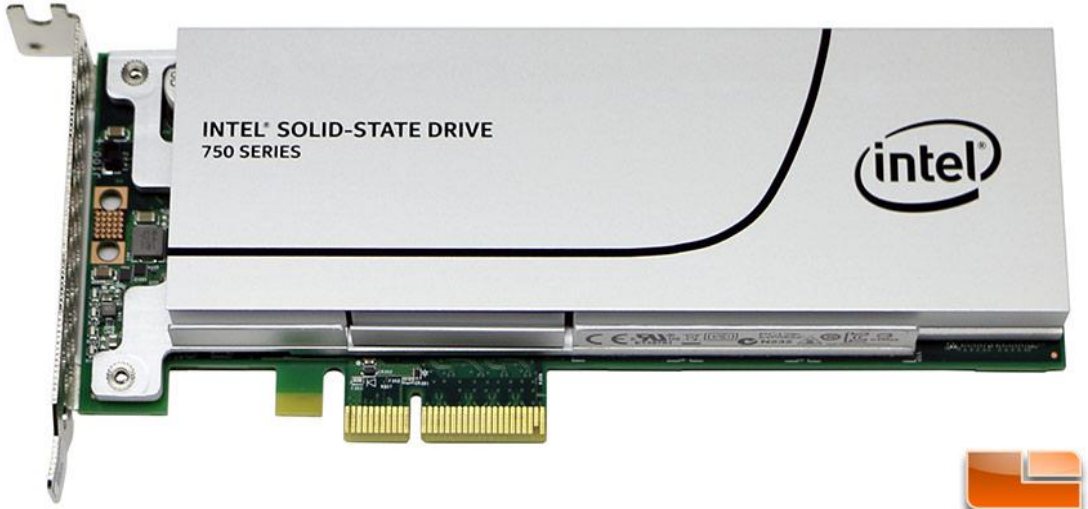
**Hardware-based** Storage Pooling over Ethernet using "NVMe over Fabrics" protocol

| | |
|---|---|
| Storage Pool | Unlimited |
| Compute Radix | Unlimited |
| Latency | Add ~2 µSec per I/O |
| Bandwidth | May be oversubscribed |
| Use Case | Warm Tier Storage |
| Cost | Low |



# Accelerators in Cloud Infrastructure



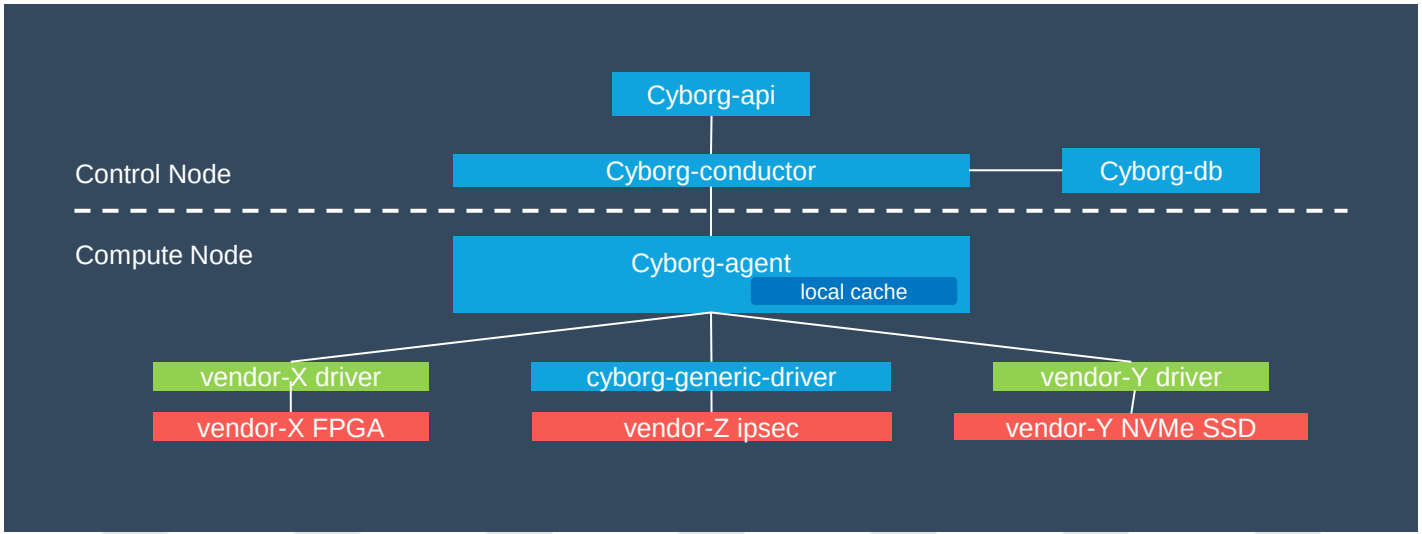



GPU

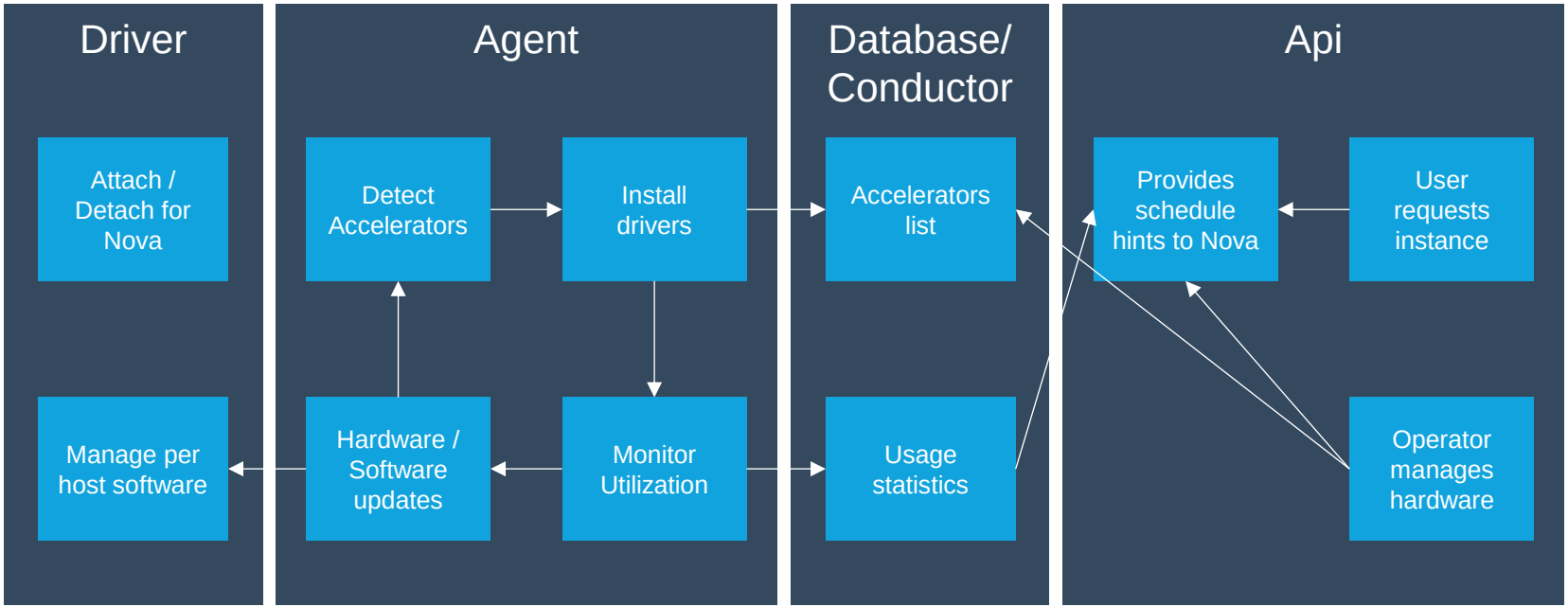


# OpenStack Acceleration Service – Cyborg

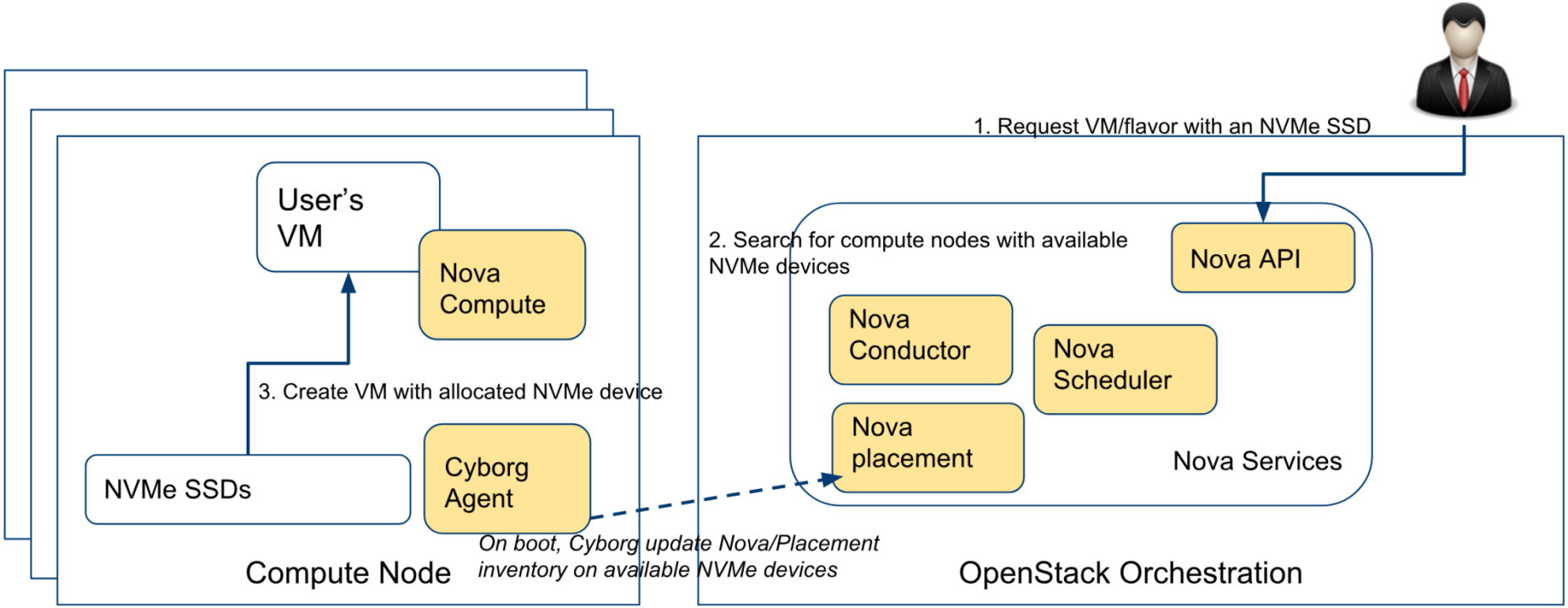
Cyborg is an OpenStack project that aims to provide a general purpose management framework for acceleration resources (i.e. various types of accelerators such as Crypto cards, GPU, FPGA, **NVMe/NOF SSDs**, ODP, DPDK/SPDK and so on). So Cyborg will be a good choice to manage NVMe high-speed storage devices in Intel RSD rack, by considering it as one kind of accelerator.



# The Workflow of Cyborg Services



# How Cyborg supports NVMe Devices



CHINA OpenInfra Days

CHINA OpenInfra Days Cloud | IT大咖说 知识共享平台

# Thank You

www.99cloud.net www.intel.com

# Demo

- Dynamically compose hardware and accelerator to meet AI/HPC requirement.
- Network Topology auto discovery by leveraging an enhanced telemetry solution
- VM to be deployed into the host sever with the big available network bandwidth per default/flavor policy
- Update flavor policy with specified network bandwidth setting and verify how target VM to be deployed