shaping tomorrow with you

Unlock Infrastructure Capabilities with Intel ® Rack Scale Architecture and OpenStack

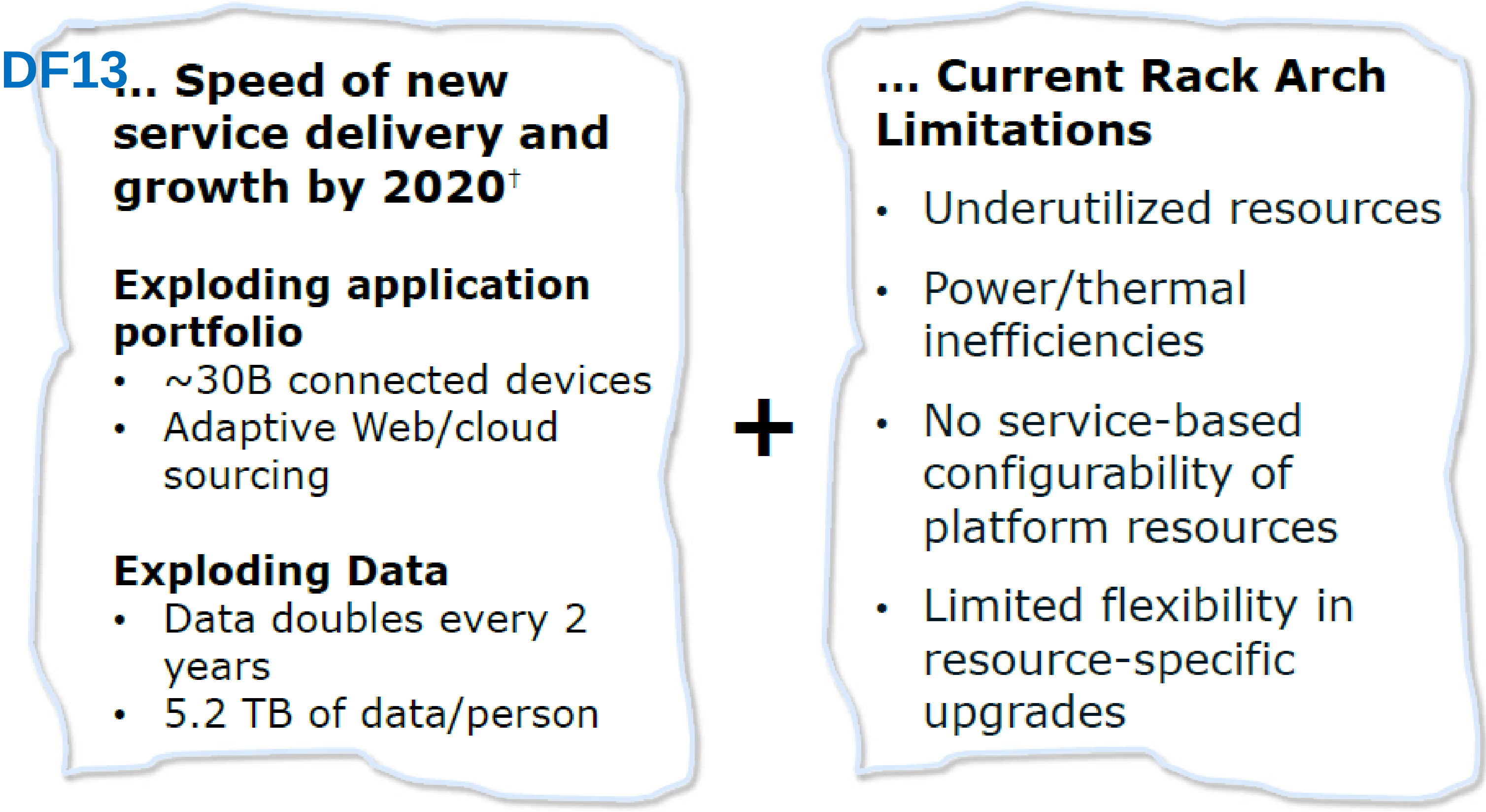
Dieter Kasper
Fujitsu Distinguished Engineer
CTO Enterprise Platform Services

2016-10-06 v2

Past

RSA/Rack Scale Architecture Motivation

IDF13



… Speed of new service delivery and growth by 2020[†]

Exploding application portfolio

- ~30B connected devices
- Adaptive Web/cloud sourcing

Exploding Data

- Data doubles every 2 years
- 5.2 TB of data/person

+

… Current Rack Arch Limitations

- Underutilized resources
- Power/thermal inefficiencies
- No service-based configurability of platform resources
- Limited flexibility in resource-specific upgrades

=

Need flexible rack scale architecture to dynamically match server to service

1st step to a Modular Architecture: Blade Server

FUJITSU

First commercial blade server shipped in 2001 by RLX-Technologies

- Values
  - Share key infrastructure components: Chassis, Power, Cooling, Fabric
  - Simplify cable management
  - Simplify manageability, serviceability
  - Compute density
  - Improved TCO

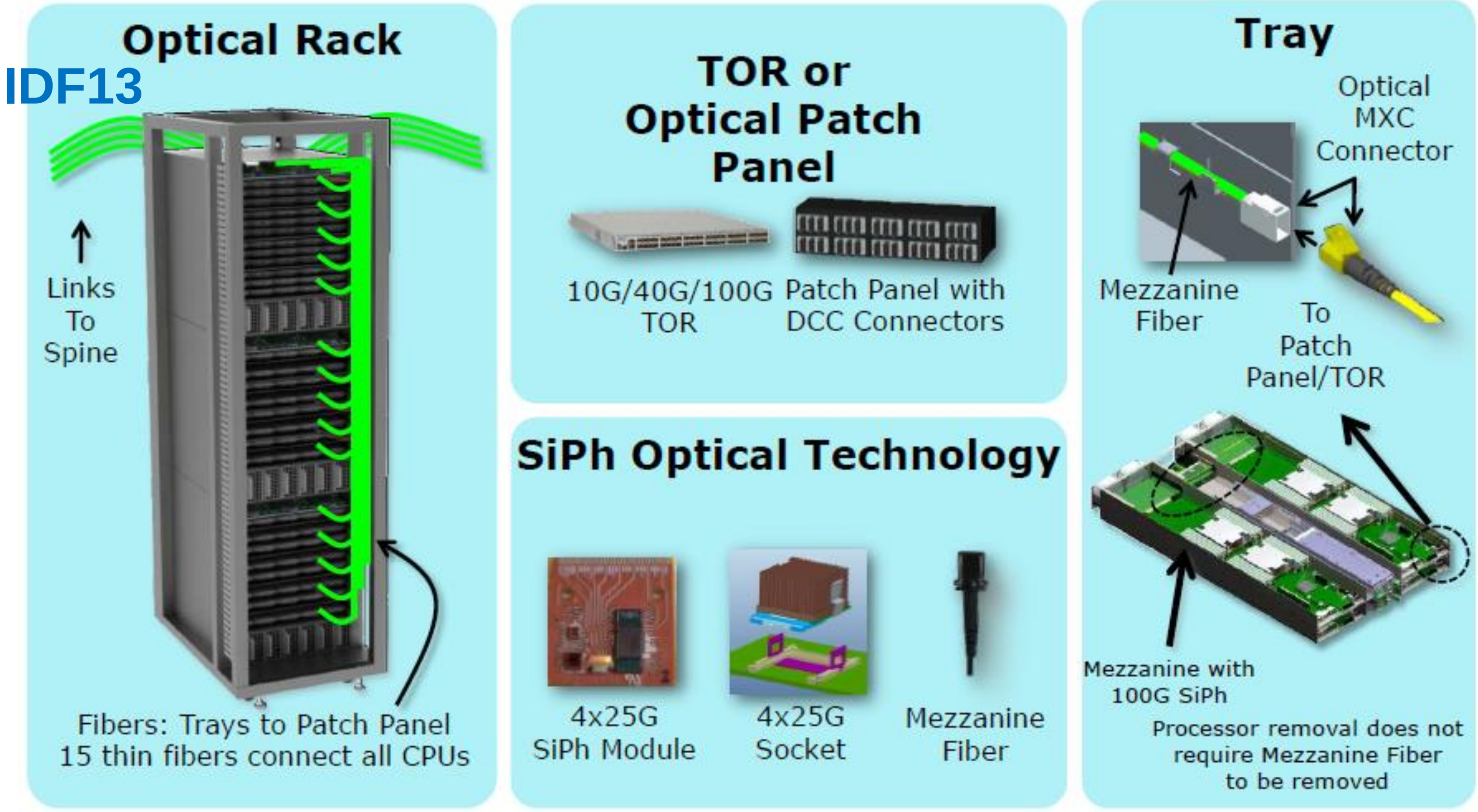
- Limitations
  - Limited core processors
  - Limited I/O-Slots and bandwidth
  - Limited internal Storage … SAN infrastructure required
  - No / limited support of FibreChannel, Infiniband
  - Special form factor I/O-Switches are released later

Rack Scale Architecture Vision

2013

- The RSA idea is more than a “Blade across a Rack”
- Disaggregate the basic building blocks of a server:
  compute, memory, storage, I/O (FC, Eth, IB, …)
- Pool and Compose new servers from these building blocks in a dynamic, flexible and agile manner

RSA - Silicon Photonics for Disaggregation



IDF13

Optical Rack

Links To Spine

Fibers: Trays to Patch Panel
15 thin fibers connect all CPUs

TOR or Optical Patch Panel

10G/40G/100G TOR

Patch Panel with DCC Connectors

SiPh Optical Technology

4x25G SiPh Module

4x25G Socket

Mezzanine Fiber

Tray

Optical MXC Connector

Mezzanine Fiber

To Patch Panel/TOR

Mezzanine with 100G SiPh

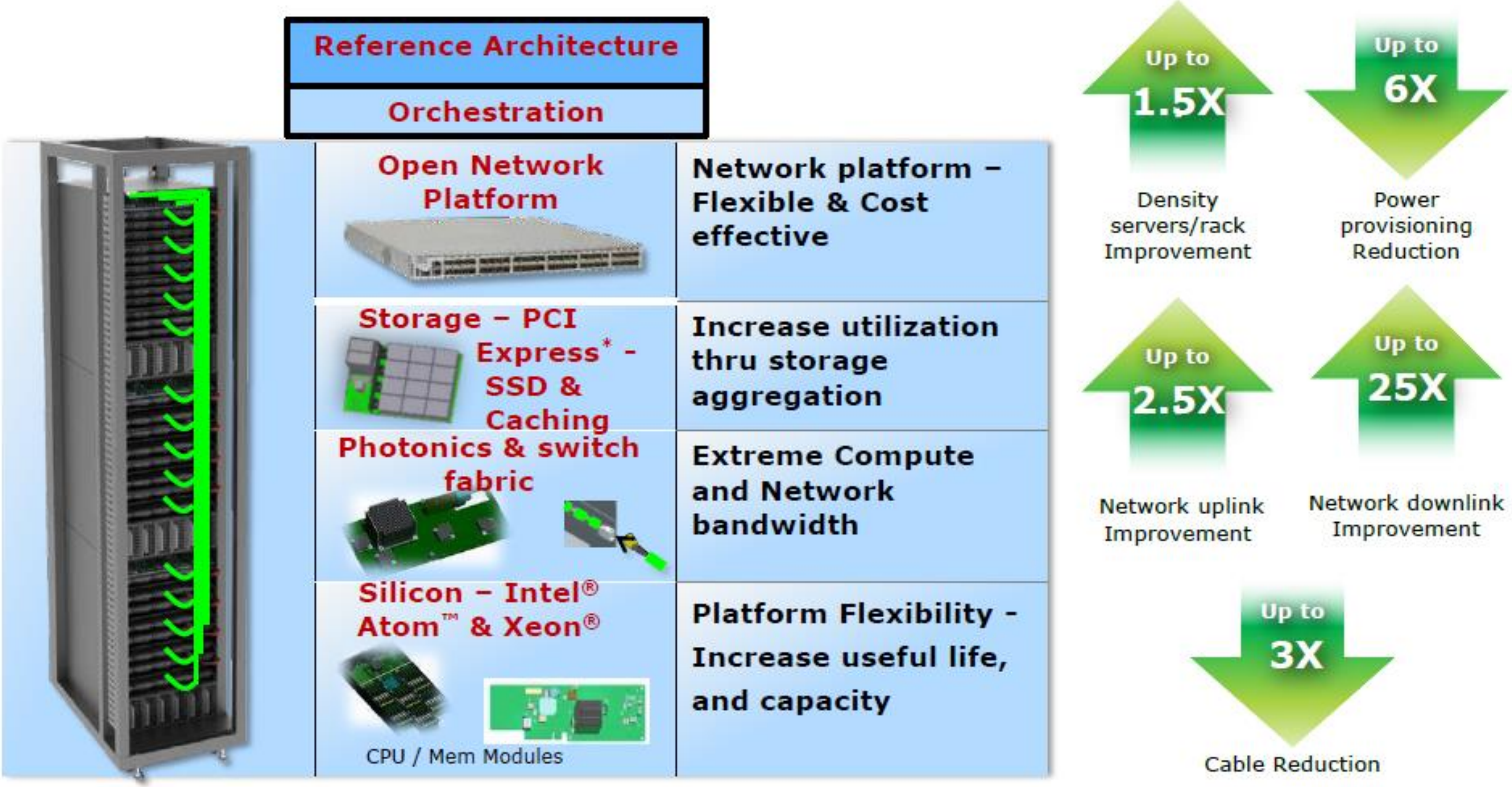
Processor removal does not require Mezzanine Fiber to be removed

=

Rack is configured once,
ready for multi-generational use

Rack Scale Architecture Value

IDF13



Reference Architecture

Orchestration

Open Network Platform	Network platform – Flexible & Cost effective
Storage – PCI Express* - SSD & Caching	Increase utilization thru storage aggregation
Photonics & switch fabric	Extreme Compute and Network bandwidth
Silicon – Intel® Atom™ & Xeon® CPU / Mem Modules	Platform Flexibility - Increase useful life, and capacity

Up to 1.5X

Density servers/rack Improvement

Up to 6X

Power provisioning Reduction

Up to 2.5X

Network uplink Improvement

Up to 25X

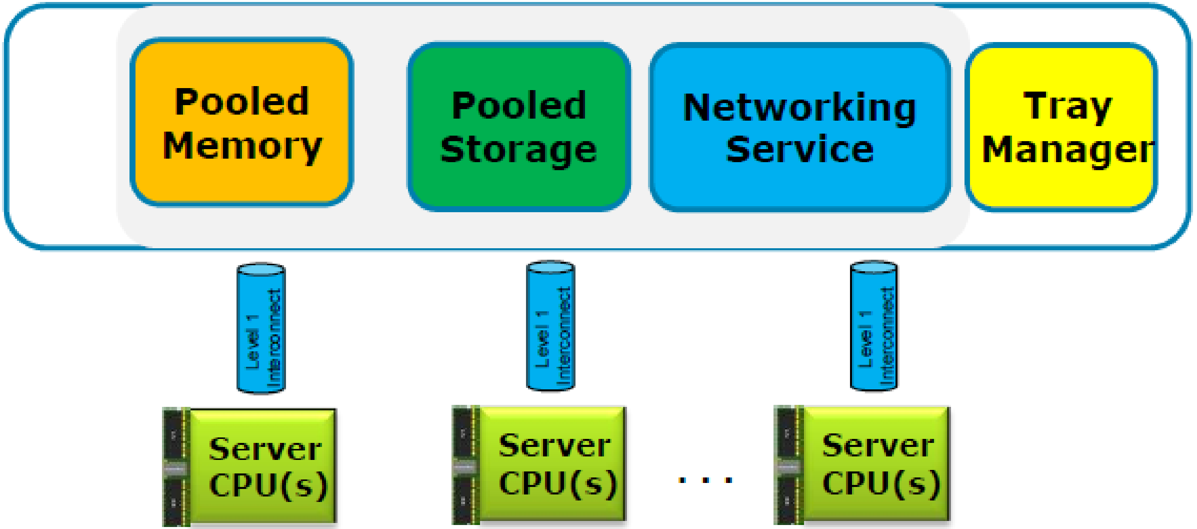
Network downlink Improvement

Up to 3X

Cable Reduction

RSA – Value of Pooled Resources

IDF13

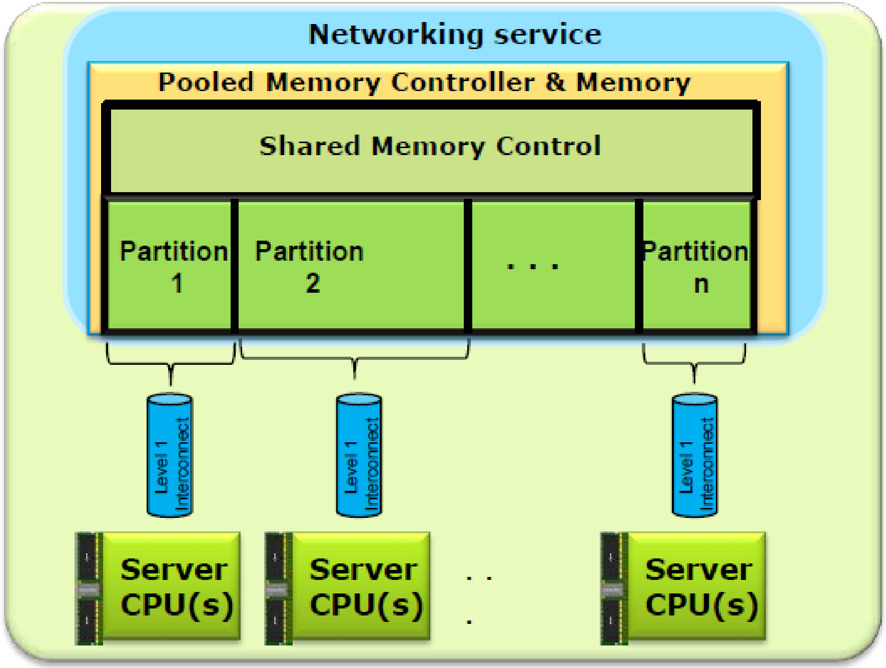


- Configurable, density optimized cloud solution
- Right sizes server resources to service workload dynamically
- Resource Pooling enables a flexible Cloud Architecture
- Enables software innovation through features such as memory sharing

Shared Memory Pool Solutions

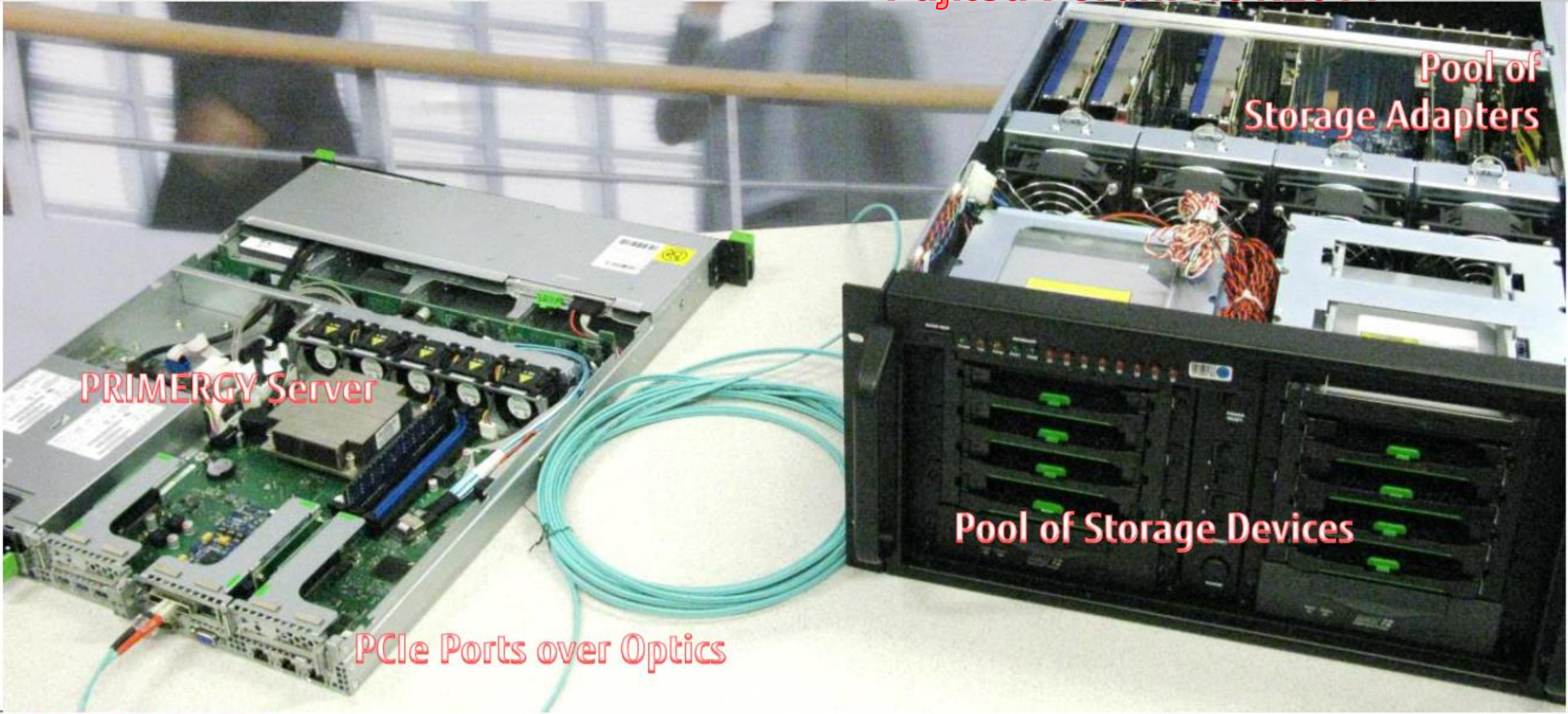
IDF13

- Large disaggregated memory pool using standard DIMMs / NV-DIMMs
- Apportionable memory to nodes based on workload demand
- Support per node partition and shareable partition(s)
- Sharable partitions can be used for VM Migration and other advanced functions



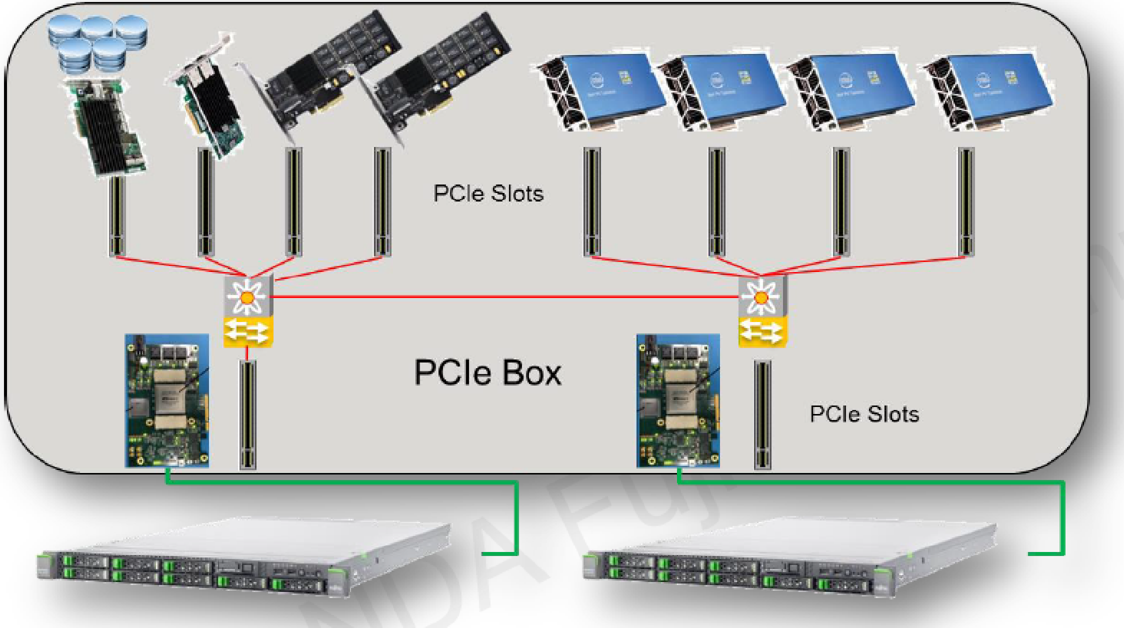
Remote Pool of Storage

Fujitsu Forum Nov.2014



Proof of Concept Configuration Fujitsu Forum

Fujitsu Forum Nov.2014



- 2 x RX200 S8 as Compute Nodes
- 2 x MXC Connector and ClearCurve® Fibre
- 1 x PCIe I/O Box
- 4 x Optical Engine and FPGA PCIe gasket on PCIe card
  - Placed inside the server and the PCIe I/O Box
- Two PLX PCIe Switches inside the PCIe I/O Box
- 1 x10GBase T adapter from Intel
- 1 x Cougar LSI RAID Controller connected to 8 HDD'S inside the PCIe I/O Box.
- Two FusionI/O PCIe SSD's with 1,5 and 1 Terabyte connected to the RX200 S8
- 4 x Intel XEON Phi™ connected to the RX200 S8
- Microsoft Windows 2012 R2 for the video workload
- Linux RedHat 6.4 for the GPGPU workload

Past	2013
Present	**2016**

RSA became RSD = Rack Scale Design

- “RSA” is owned by DELL/EMC
  - The RSA encryption algorithm was developed in 1977 by Ron Rivest, Adi Shamir and Leonard Adleman
  - RSA Security was acquired by EMC Corporation in 2006
- “Intel – RSA” was mainly based on Hardware technologies
- “Intel – RSD” is more Software configuration minded

Speculation:

- Technology issues ?
- Business / Cost issues ?
  - Fast-IT = Cloud-DCs are growing faster than Enterprise-DCs
  - Cloud-DCs are Ethernet based: 10/25/50/100 GbE

Transition from RSA to RSD

2013 - RSA	2016 - RSD	
CPU pool	-	
MEM pool	-	
Optical PCIe / Switch	-	Cu based
IO consolidation	partly	
PCIe SSD pool	yes	
Storage pool	yes	
(FPGA pool)	yes	
Pod-/Rack-/Drawer-Mgmt	yes	

RSD Status

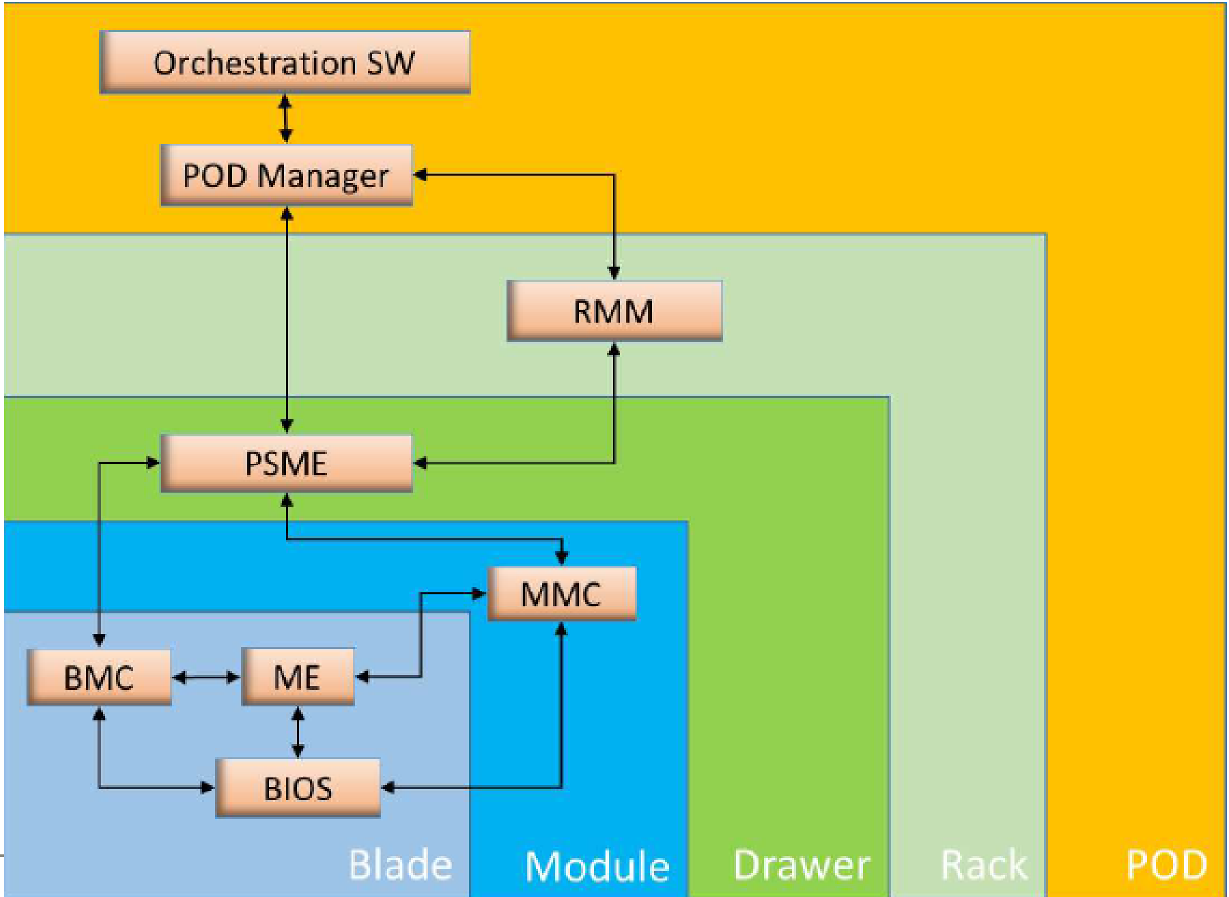
- Reference Code and specs are complete, code released for open source development http://itpeernetwork.intel.com/intel-rack-scale-design-now-ready-open-source-development/
- Worked with industry partners to extend DMTF† Redfish™ to support management of Memory, CPU, PCIe, Local Storage and Network
- Working with SNIA™ to extend Redfish to comprehend managing data storage and storage services
- Pod Manager implementation is complete and productized by partners
- Intel® Rack Scale Design Aligned Ecosystem with Multiple RSD Vendor Solutions

DMTF = Distributed Management Task Force

RSD Terminology

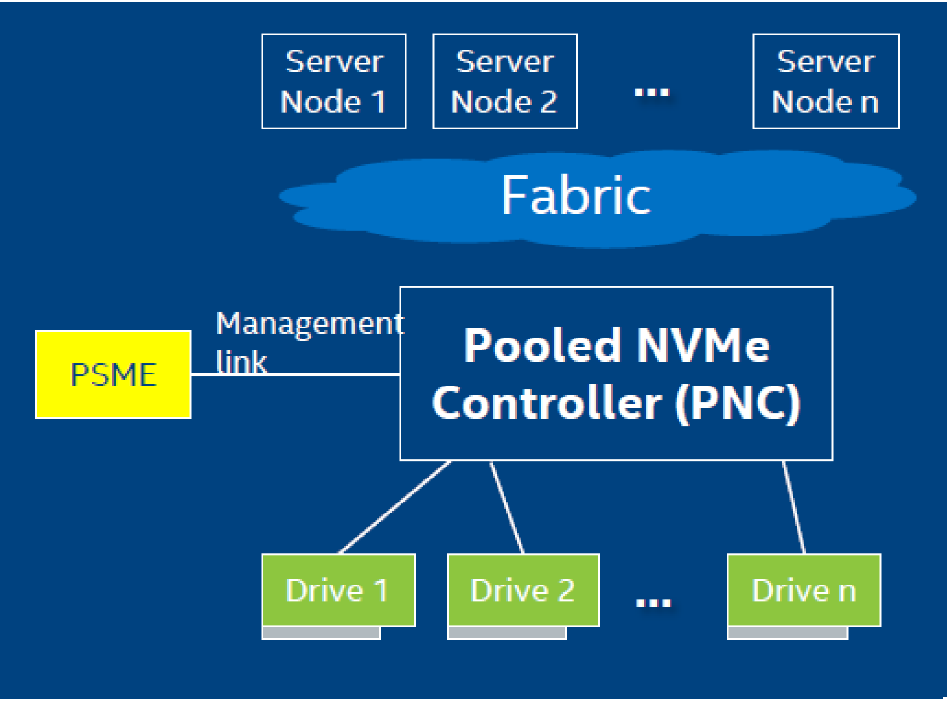
BMC	Baseboard management controller
CPP	Control Plane Processor = host to run the PSME on ref platform
DMC	Drawer management controller, where PSME functionality is implemented
EORS	End-of-Rack-Switch
MMC	Module management controller, manage the blades in the module
Node	any compute node, such as Xeon or Atom processor
PNC	Pooled NVMe Controller
POD	Collection of Racks
PODM	POD manager
PSME	Pooled System Management Engine = Micro controller responsible for configuration of shared and pooled Recourses (MEM, Storage, Nodes, SDN)
RMM	Rack management module
TORS	Top-of-Rack-Switch

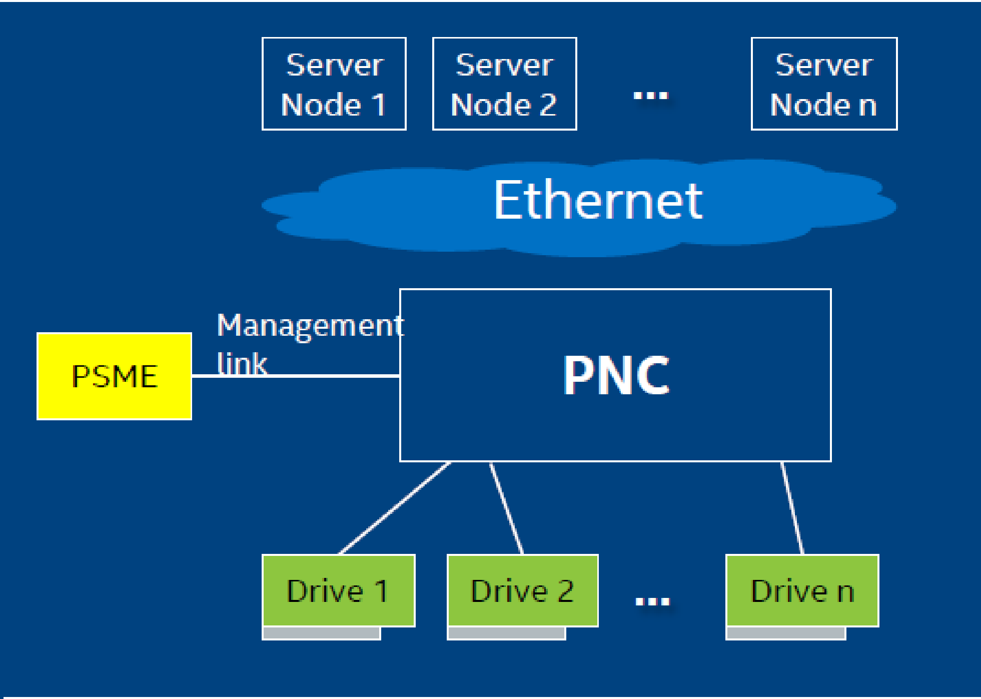
POD logical hierarchy



RSD Pooled Storage

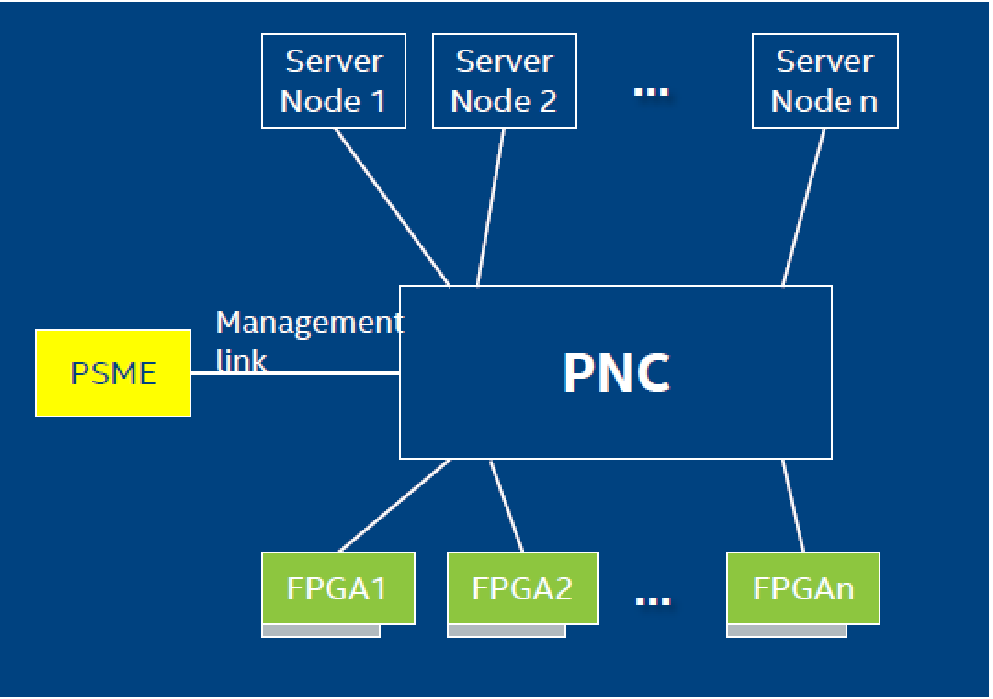
IDF16



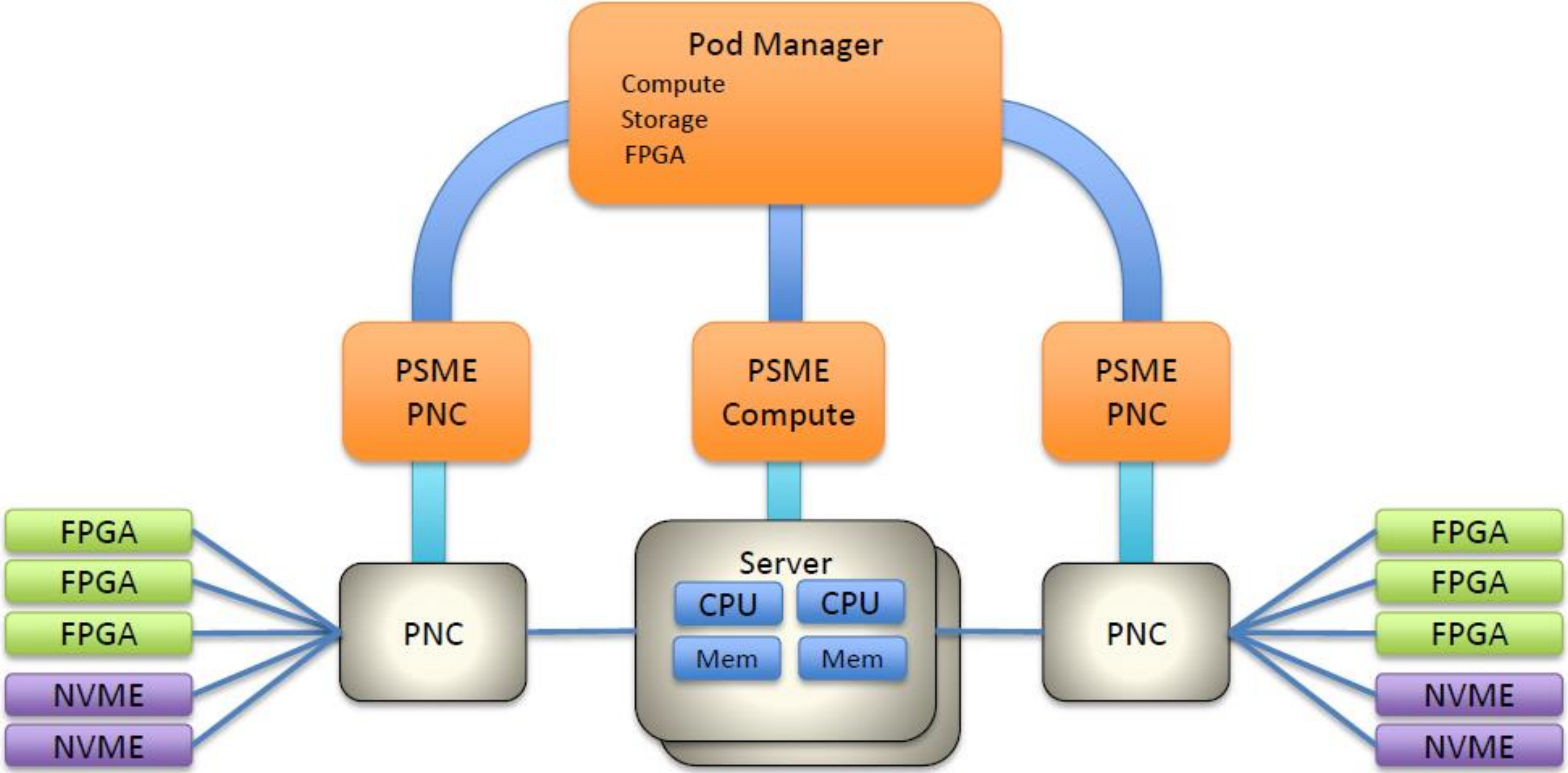


RSD Pooled FPGA

IDF16

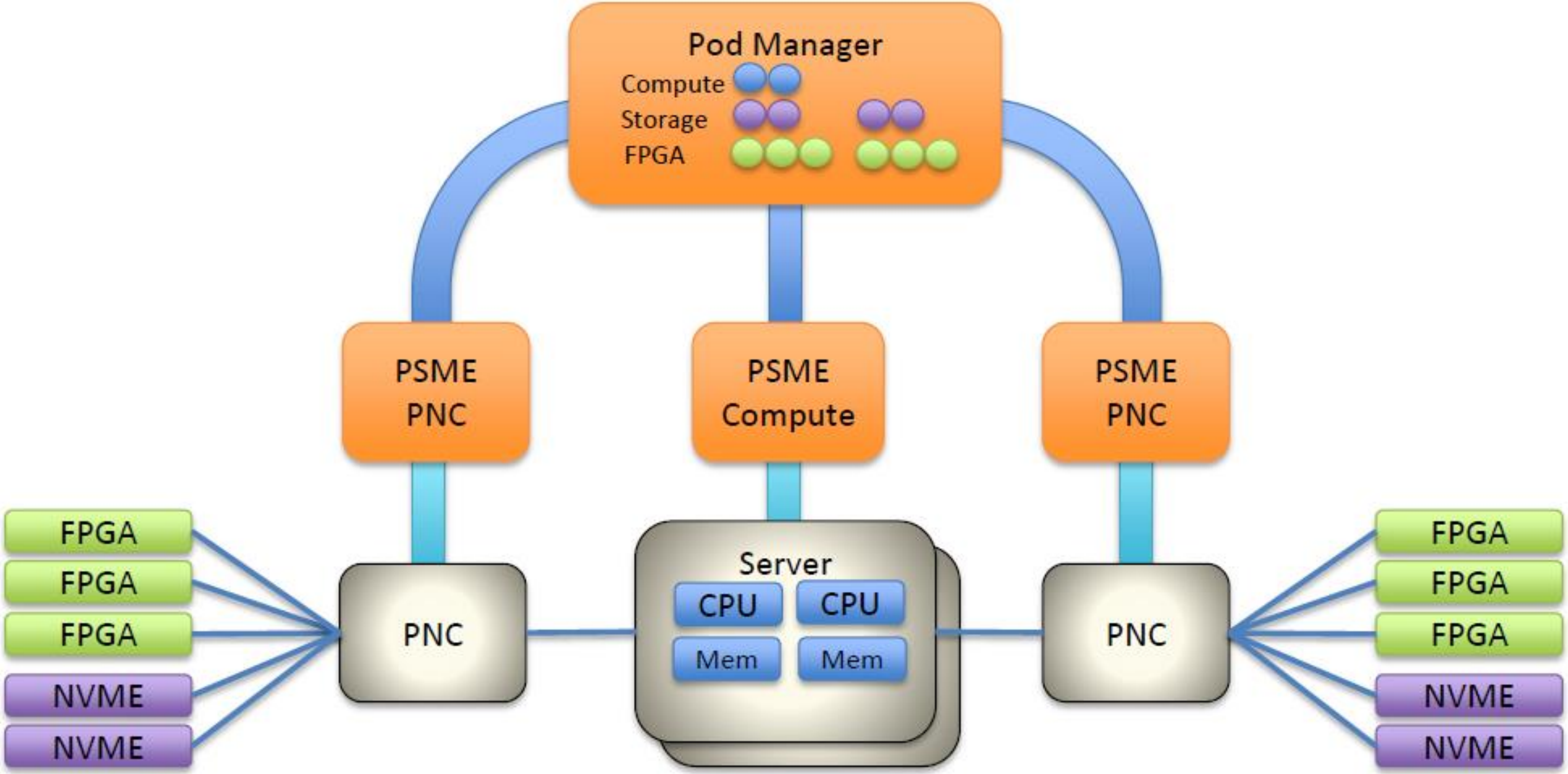


RSD Composition in Action (contd.)



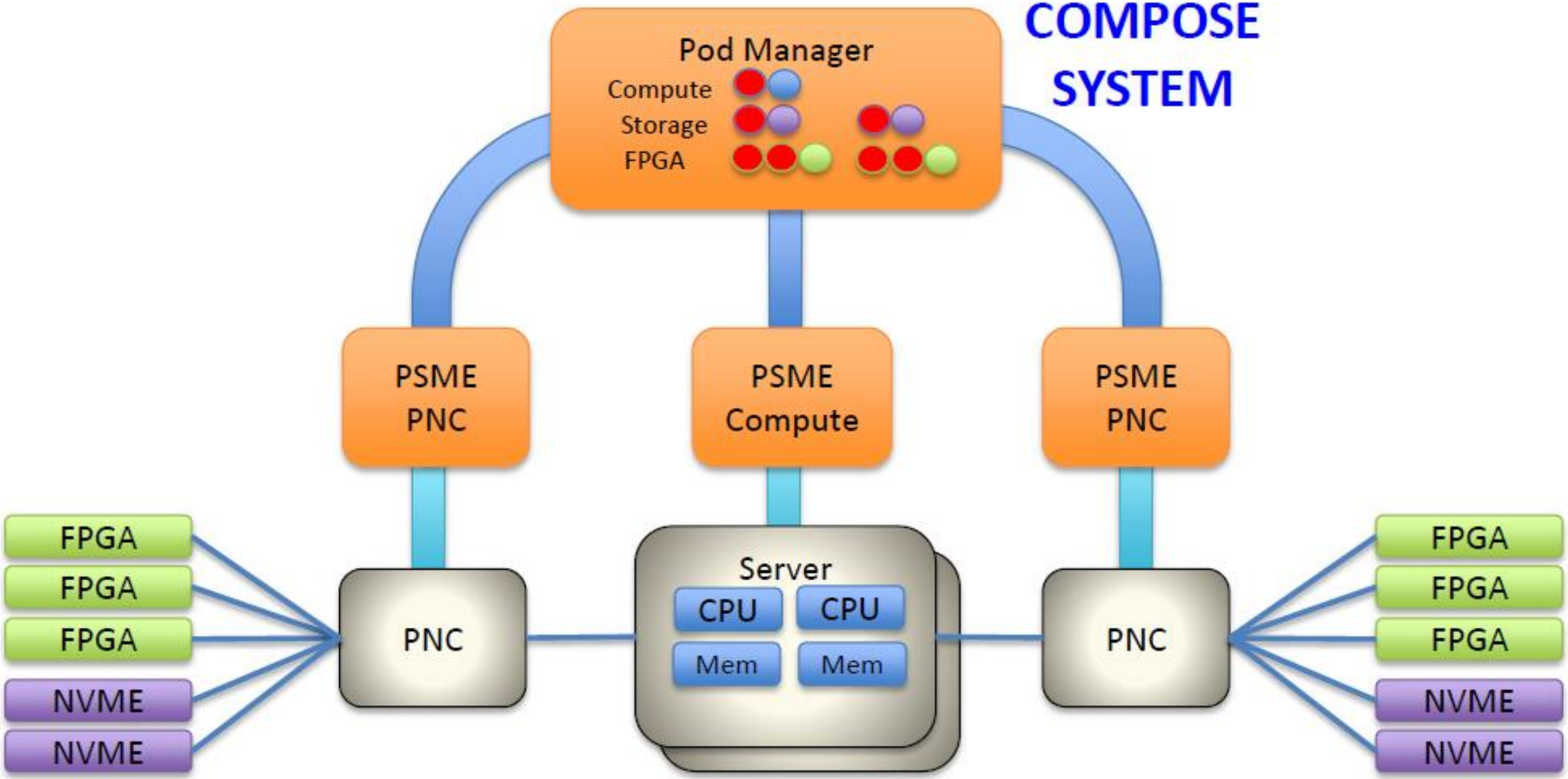
IDF16
INTEL DEVELOPER FORUM

RSD Composition in Action (contd.)



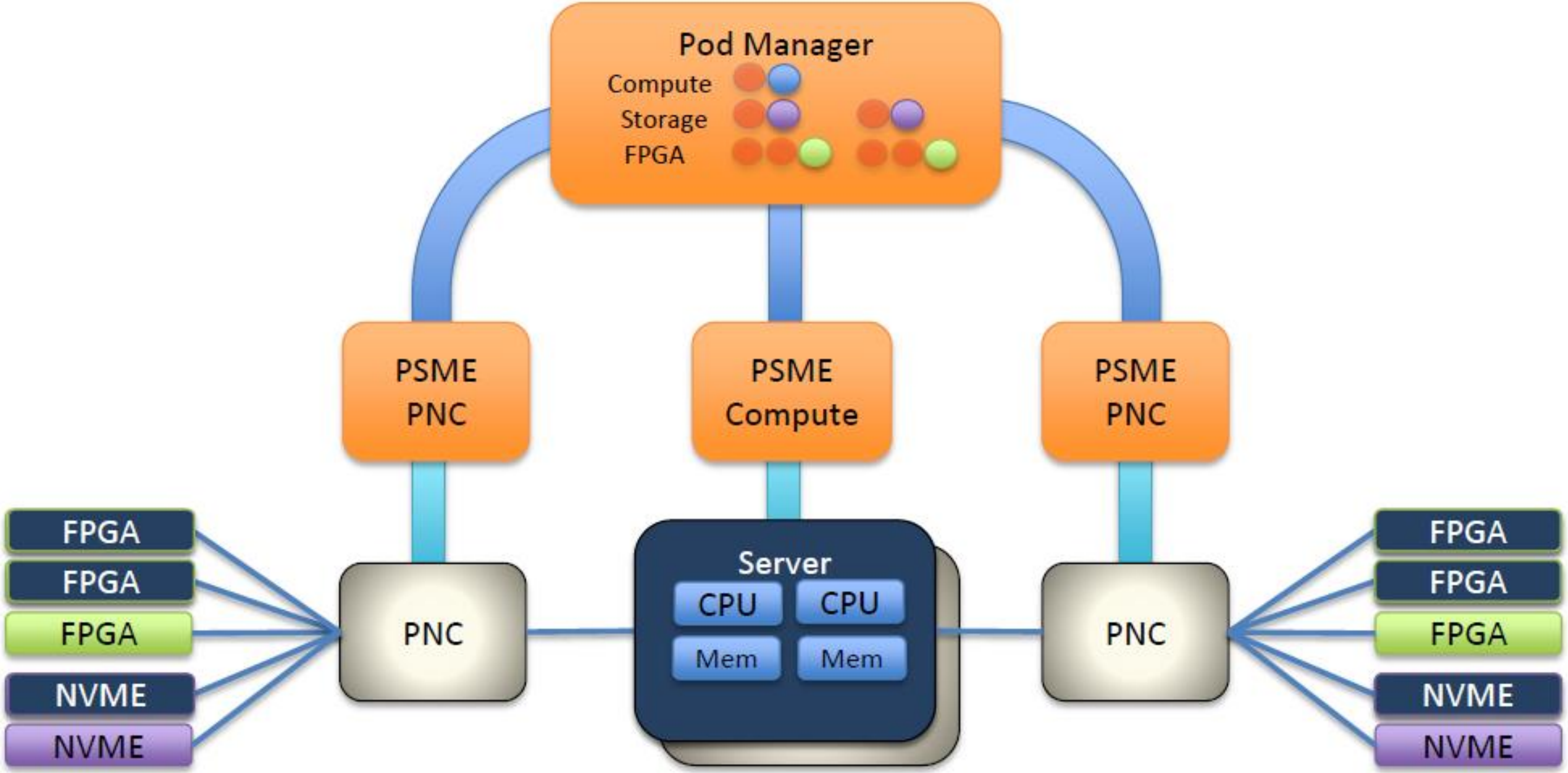
IDF16
INTEL DEVELOPER FORUM

RSD Composition in Action (contd.)

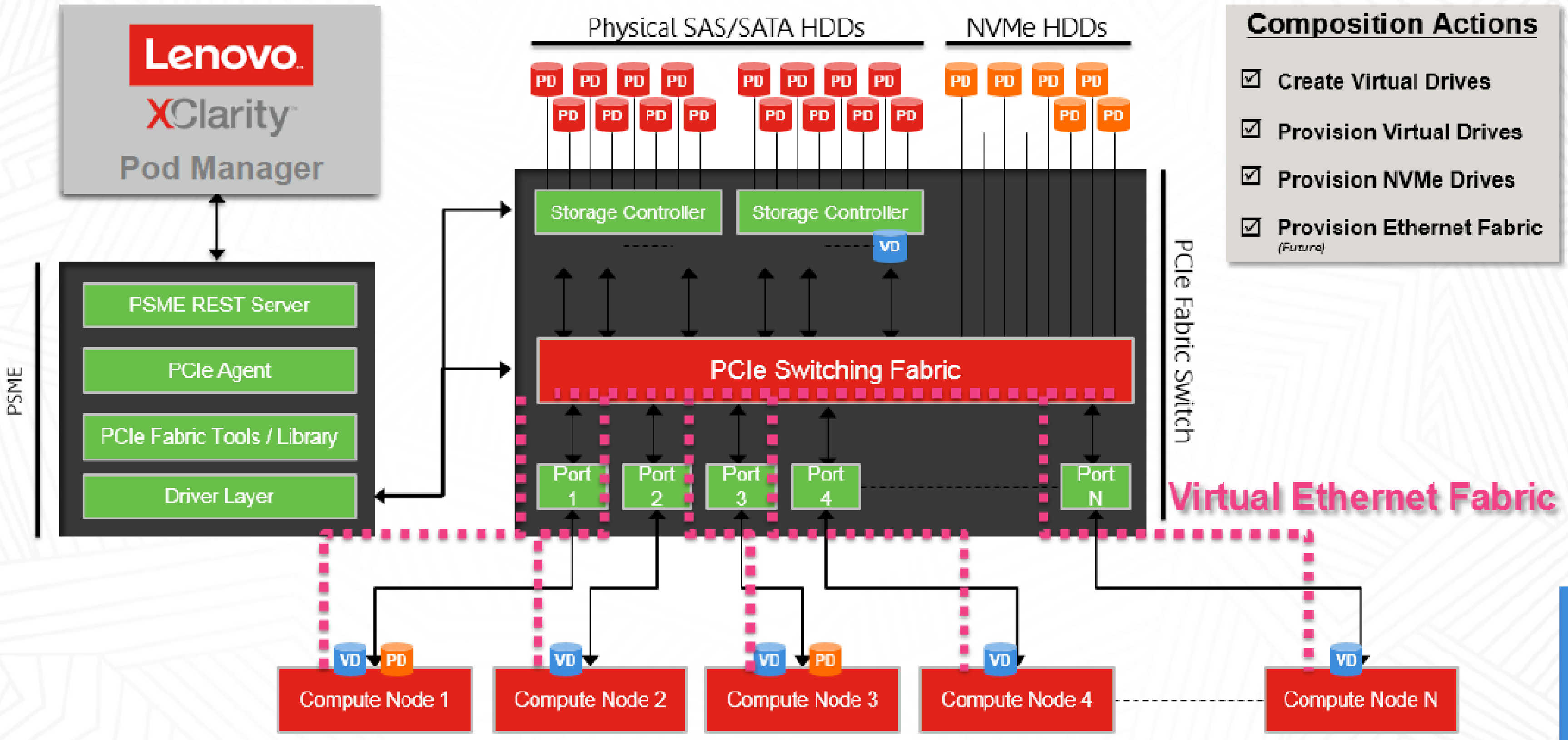


IDF16
INTEL DEVELOPER FORUM

RSD Composition in Action (contd.)



Node Composition Workflow -- Provision Ethernet Fabric



Composition Actions

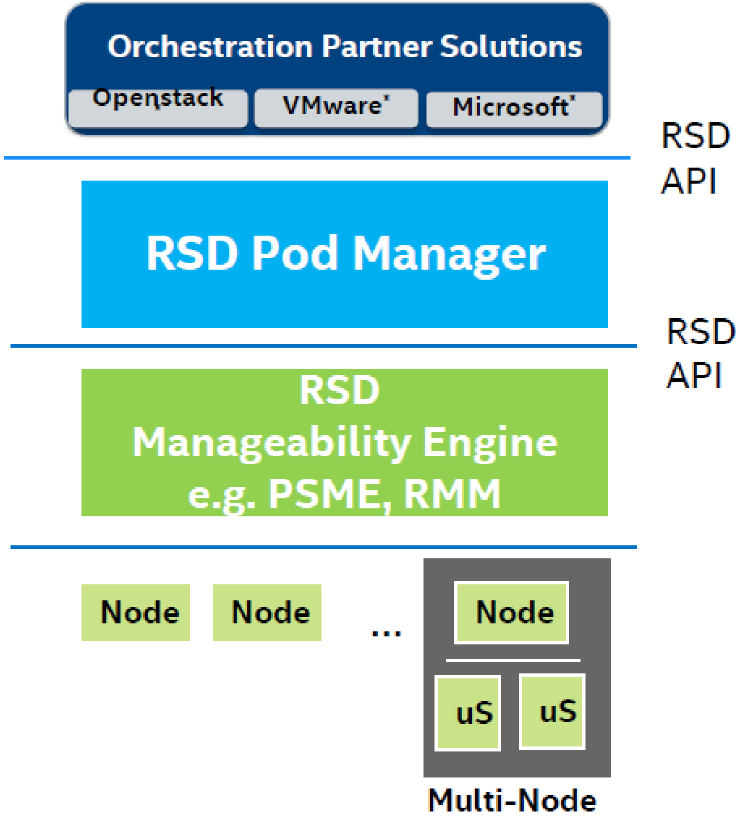
☑ Create Virtual Drives

☑ Provision Virtual Drives

☑ Provision NVMe Drives

☑ Provision Ethernet Fabric *(Future)*

RSD and Orchestration

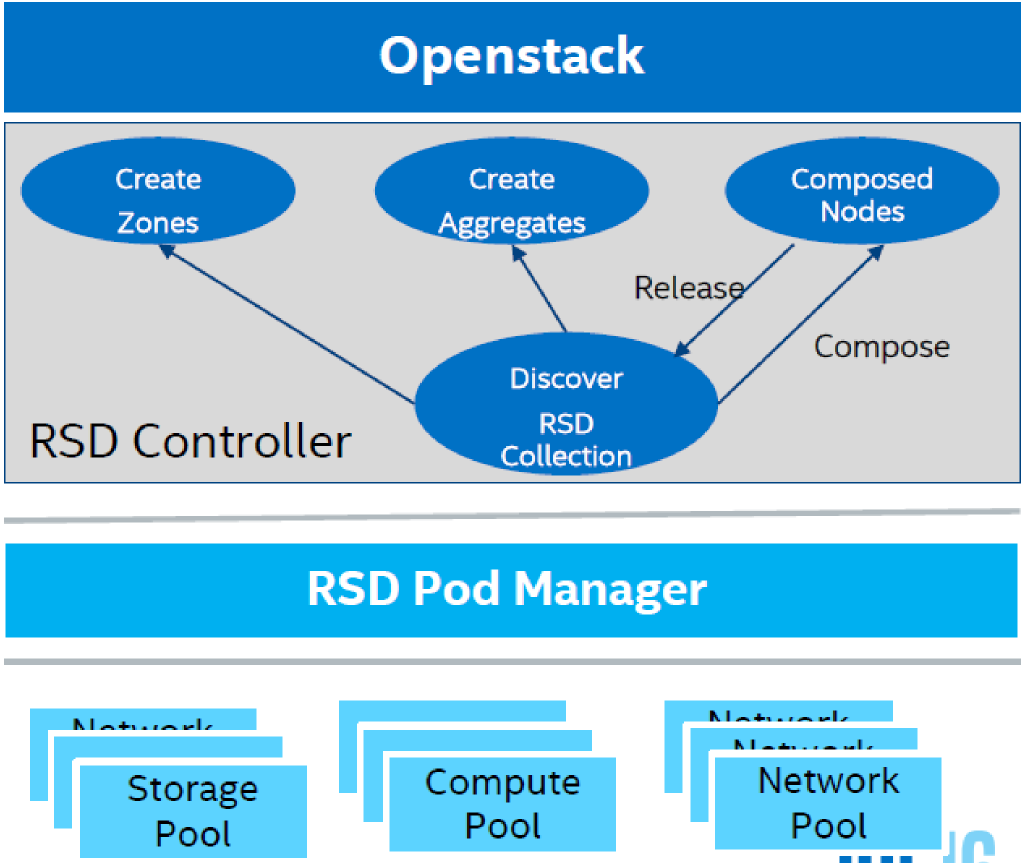


- RSD Pod manager designed to interface with multiple orchestration stacks
- RSD provides physical resource and capabilities discovery across vendor implementations
- Location aware placement
- Enables composition of pooled systems for agile orchestration
- Supports stateful lifecycle management

IDF16 INTEL DEVELOPER FORUM

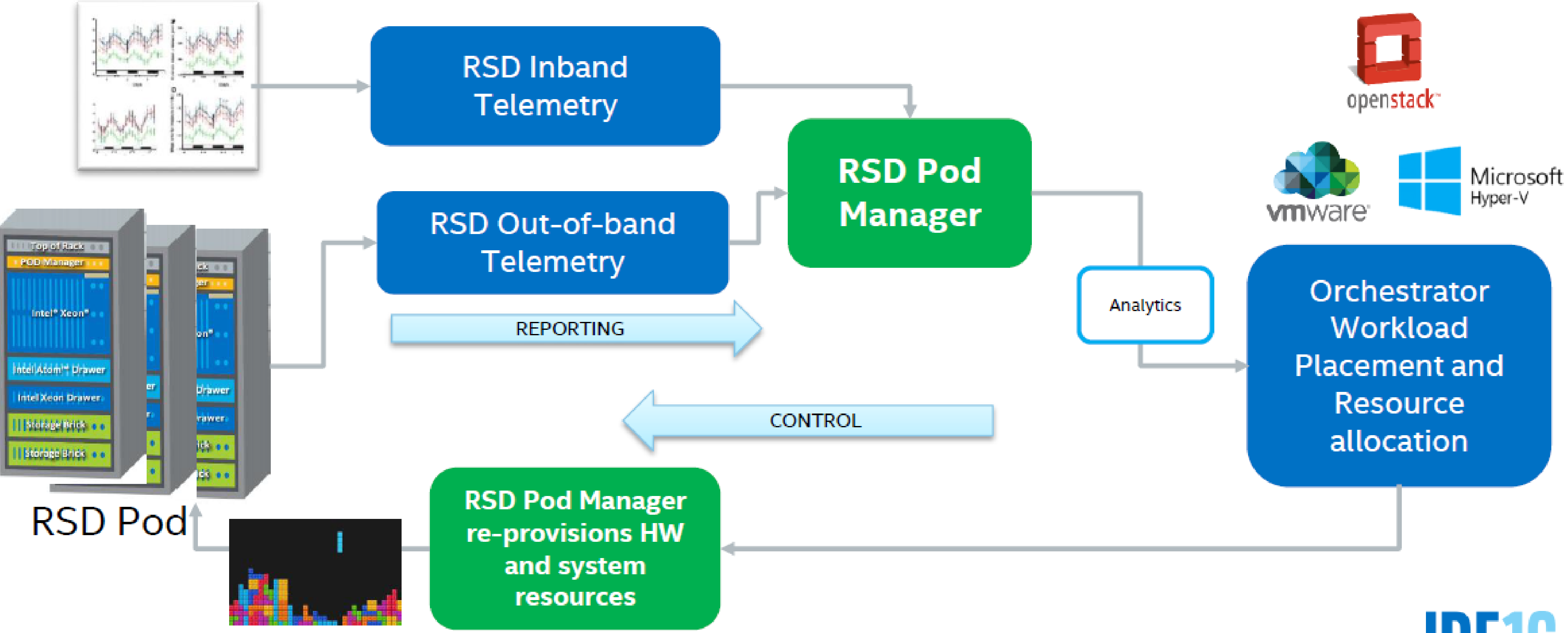
RSD Controller

- RSD controller solution which uses RSD APIs to interface RSD Pod Manager to Openstack
- Elastic HW lifecycle management for Pooled infrastructure – Compose and Release
- Automatic deployment of bare metal systems using "discovery" and "compose" APIs



RSD Controller = https://wiki.openstack.org/wiki/Valence

RSD Telemetry Flow



RSD Inband Telemetry

RSD Out-of-band Telemetry

REPORTING

RSD Pod Manager

Analytics

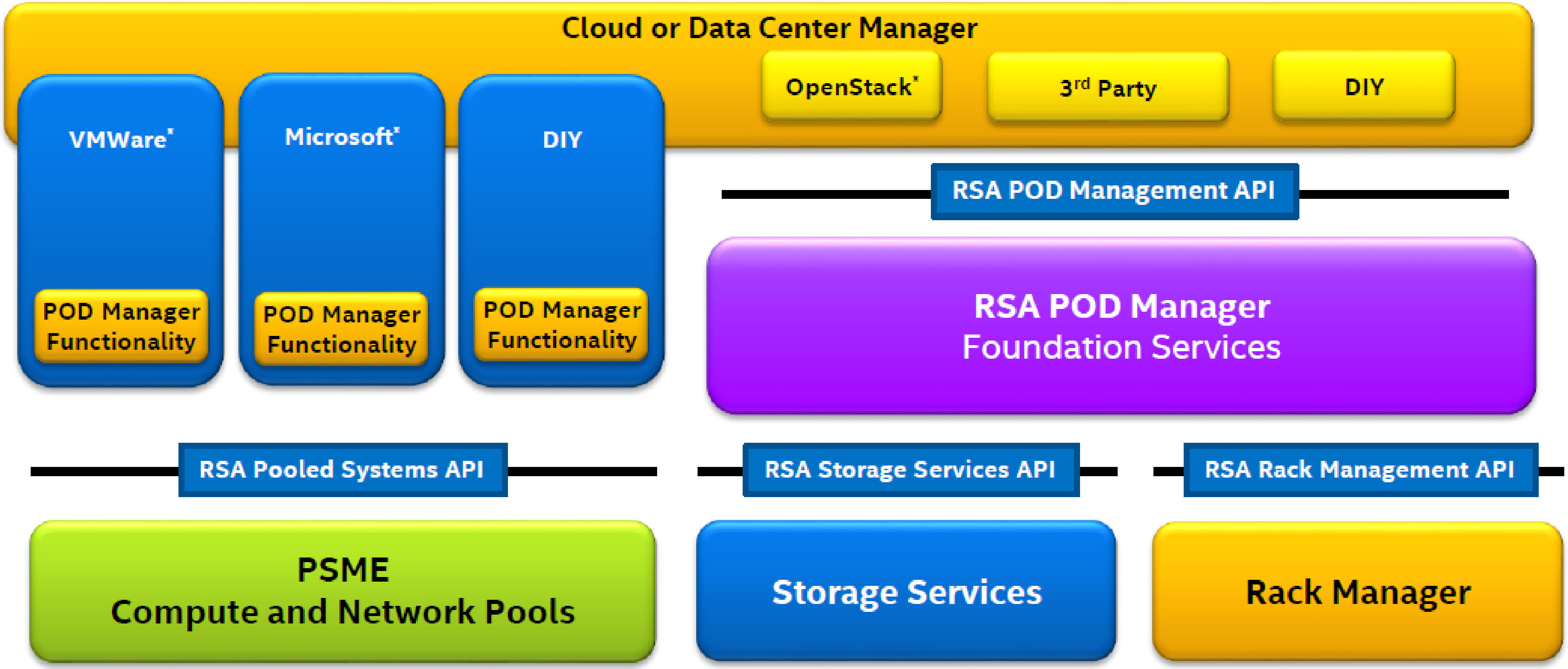
Orchestrator Workload Placement and Resource allocation

CONTROL

RSD Pod Manager re-provisions HW and system resources

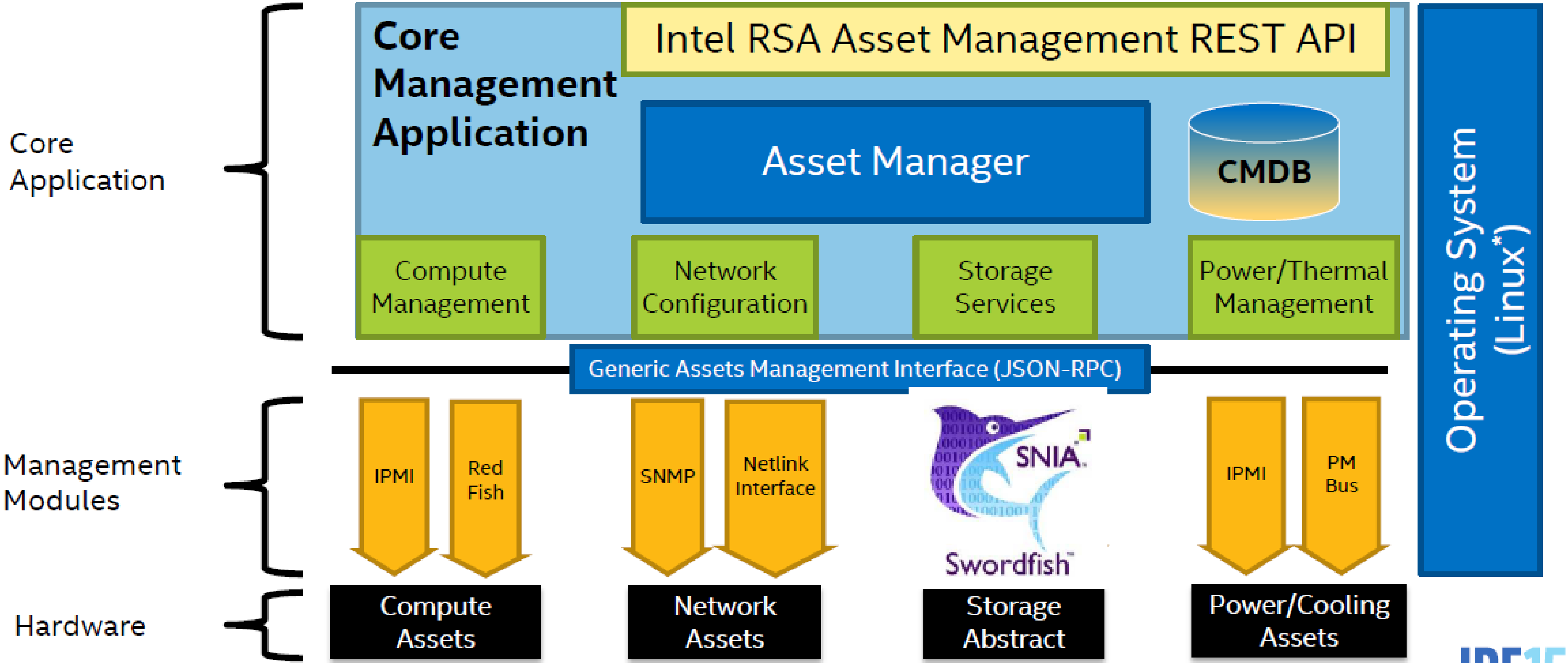
RSD Pod

Intel® Rack Scale Architecture Software Stack



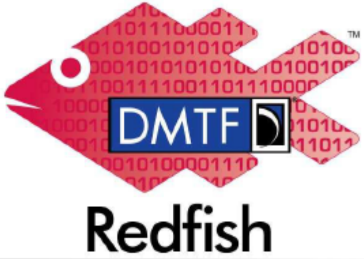
RSA = Intel® Rack Scale Architecture

Intel® Rack Scale Architecture Software/Firmware Code Flexible Architecture



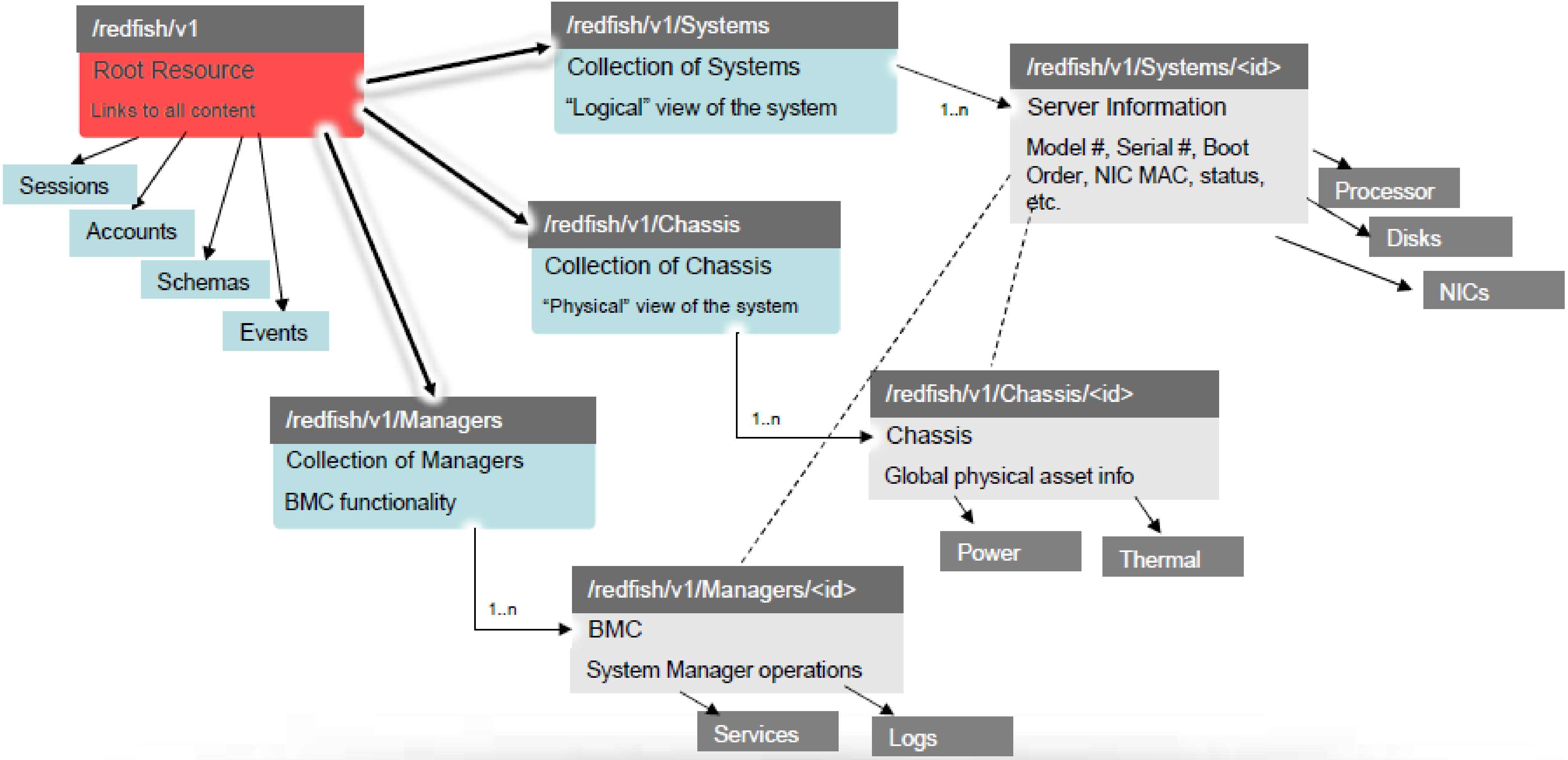
What is RedFish ?

- Industry Standard RESTful API for IT Infrastructure
  - HTTPS in JSON format based on Odata v4
  - Equally useable by Apps, GUIs, and Scripts
  - Schema-backed but human-readable



- First release focused on Servers
  - A secure, multi-node capable replacement for IPMI-over-LAN
  - Add devices over time to cover customer use cases & technology
    - Direct attach storage, PCIe and SAS switching, NVDIMMs, Multifuction Adapters, Composability
  - Intended to meet OCP Remote Machine Management Requirements

RedFish Resource Map



/redfish/v1

Root Resource

Links to all content

Sessions

Accounts

Schemas

Events

/redfish/v1/Systems

Collection of Systems

“Logical” view of the system

/redfish/v1/Systems/<id>

Server Information

Model #, Serial #, Boot Order, NIC MAC, status, etc.

Processor

Disks

NICs

/redfish/v1/Chassis

Collection of Chassis

“Physical” view of the system

/redfish/v1/Chassis/<id>

Chassis

Global physical asset info

Power

Thermal

/redfish/v1/Managers

Collection of Managers

BMC functionality

/redfish/v1/Managers/<id>

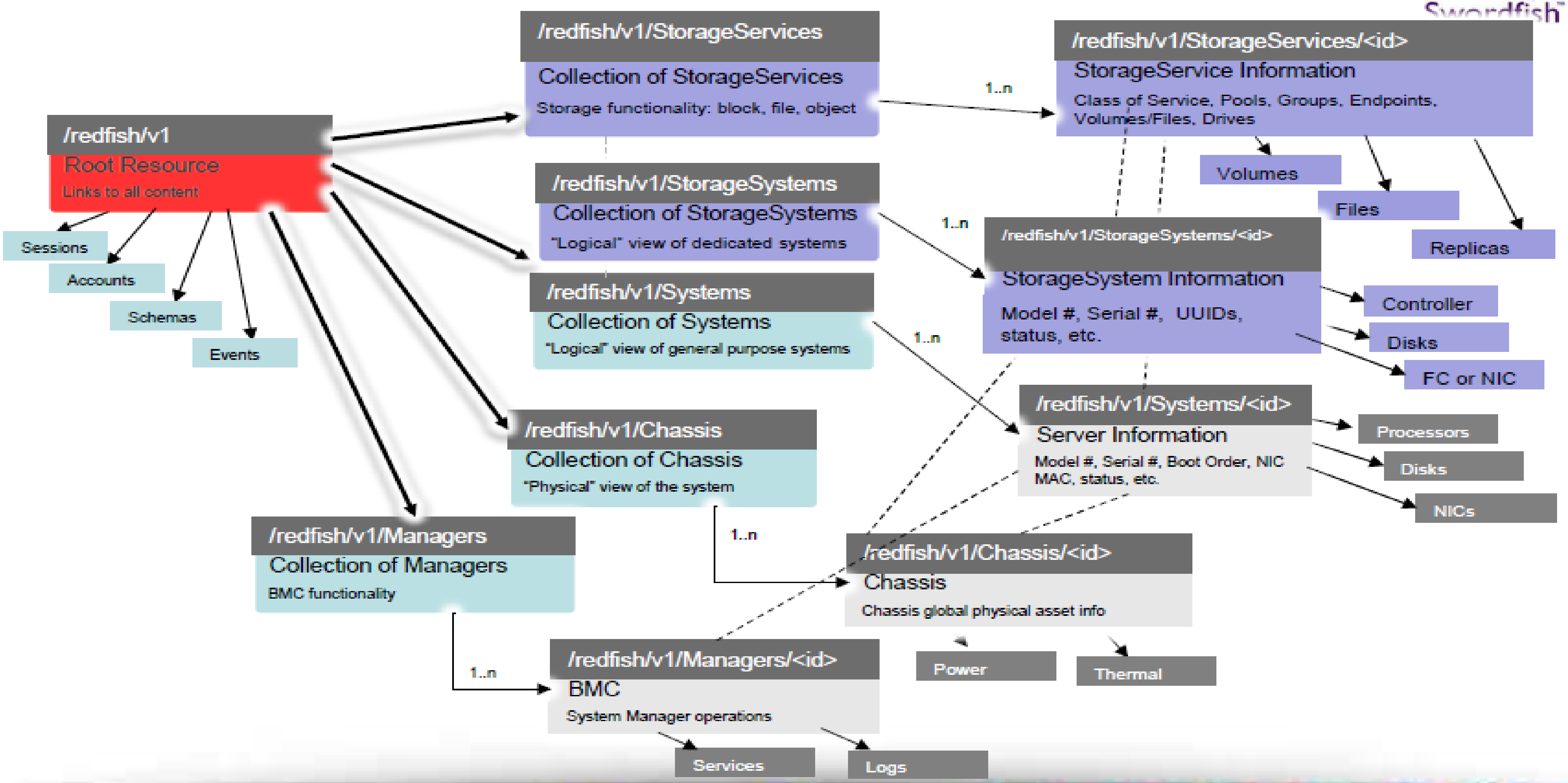
BMC

System Manager operations

Services

Logs

Adding Storage to RedFish



/redfish/v1

Root Resource

Links to all content

Sessions

Accounts

Schemas

Events

/redfish/v1/StorageServices

Collection of StorageServices

Storage functionality: block, file, object

/redfish/v1/StorageSystems

Collection of StorageSystems

"Logical" view of dedicated systems

/redfish/v1/Systems

Collection of Systems

"Logical" view of general purpose systems

/redfish/v1/Chassis

Collection of Chassis

"Physical" view of the system

/redfish/v1/Managers

Collection of Managers

BMC functionality

1..n

/redfish/v1/StorageServices/<id>

StorageService Information

Class of Service, Pools, Groups, Endpoints, Volumes/Files, Drives

Volumes

Files

Replicas

1..n

/redfish/v1/StorageSystems/<id>

StorageSystem Information

Model #, Serial #, UUIDs, status, etc.

Controller

Disks

FC or NIC

1..n

/redfish/v1/Systems/<id>

Server Information

Model #, Serial #, Boot Order, NIC MAC, status, etc.

Processors

Disks

NICs

1..n

/redfish/v1/Chassis/<id>

Chassis

Chassis global physical asset info

Power

Thermal

1..n

/redfish/v1/Managers/<id>

BMC

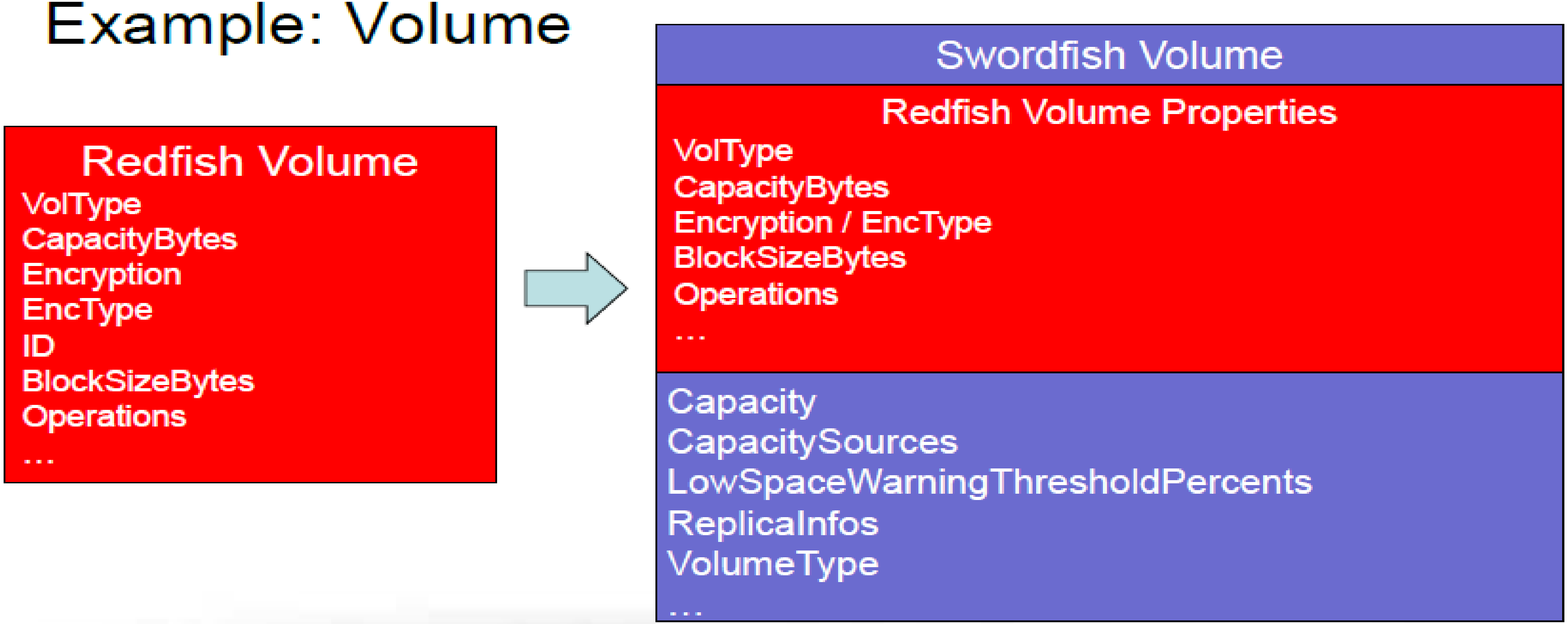
System Manager operations

Services

Logs

Seamless Extension of Redfish to Swordfish

- Make Swordfish a seamless extension of Redfish local storage schema
- Example: Volume



Redfish Volume

VolType
CapacityBytes
Encryption
EncType
ID
BlockSizeBytes
Operations
…

Swordfish Volume

Redfish Volume Properties

VolType
CapacityBytes
Encryption / EncType
BlockSizeBytes
Operations
…

Capacity
CapacitySources
LowSpaceWarningThresholdPercents
ReplicaInfos
VolumeType
…

Useful Links

https://www.dmtf.org/standards/redfish

http://redfish.dmtf.org/redfish/v1/mockup/

http://www.snia.org/forums/smi/swordfish

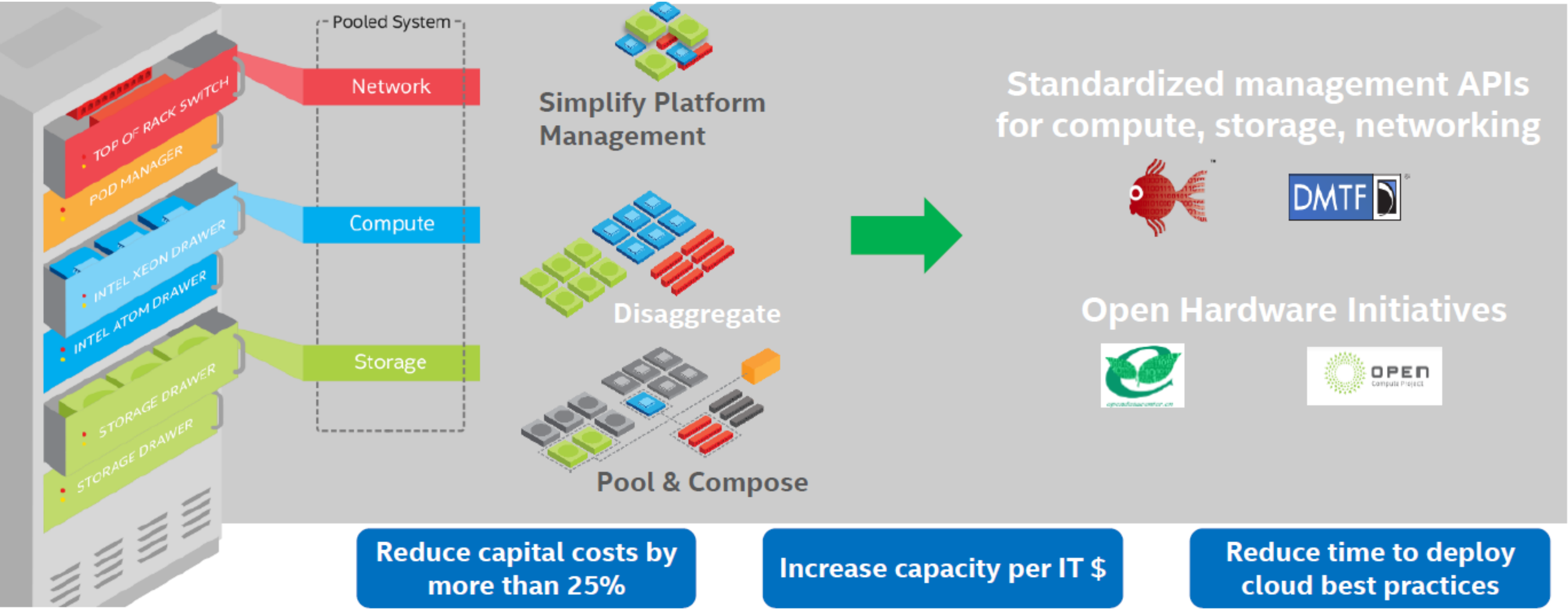
http://www.nvmexpress.org/specifications/

https://wiki.openstack.org/wiki/Valence

■ Past	2013	
■ Present	2016	
■ **Future**	**?**	**… and Summary**

RSD Summary I

IDF16 http://www.intel.com/content/www/us/en/architecture-and-technology/rack-scale-design-overview.html



Simplify Platform Management

Disaggregate

Pool & Compose

Standardized management APIs for compute, storage, networking

Open Hardware Initiatives

Reduce capital costs by more than 25%

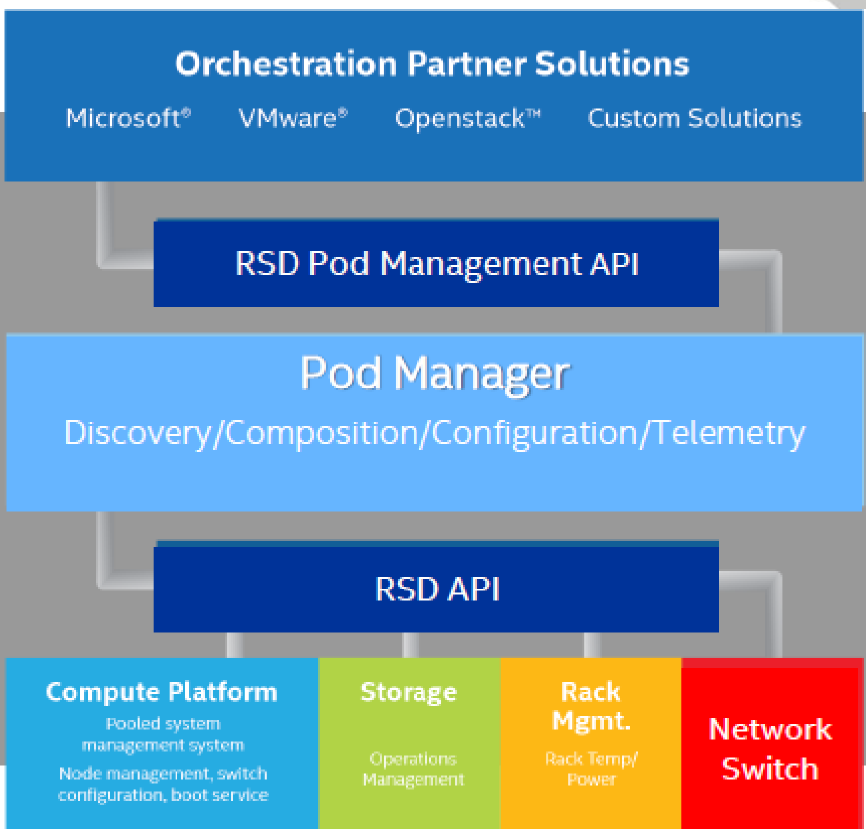
Increase capacity per IT $

Reduce time to deploy cloud best practices

RSD Summary II

Management Software Framework

- Asset & location discovery
- Disaggregated resource management
- Composable system support
- Support compute, network, and storage
- Built using DMTF† Redfish™



Orchestration Partner Solutions

Microsoft® VMware® Openstack™ Custom Solutions

RSD Pod Management API

Pod Manager

Discovery/Composition/Configuration/Telemetry

RSD API

Compute Platform

Pooled system management system

Node management, switch configuration, boot service

Storage

Operations Management

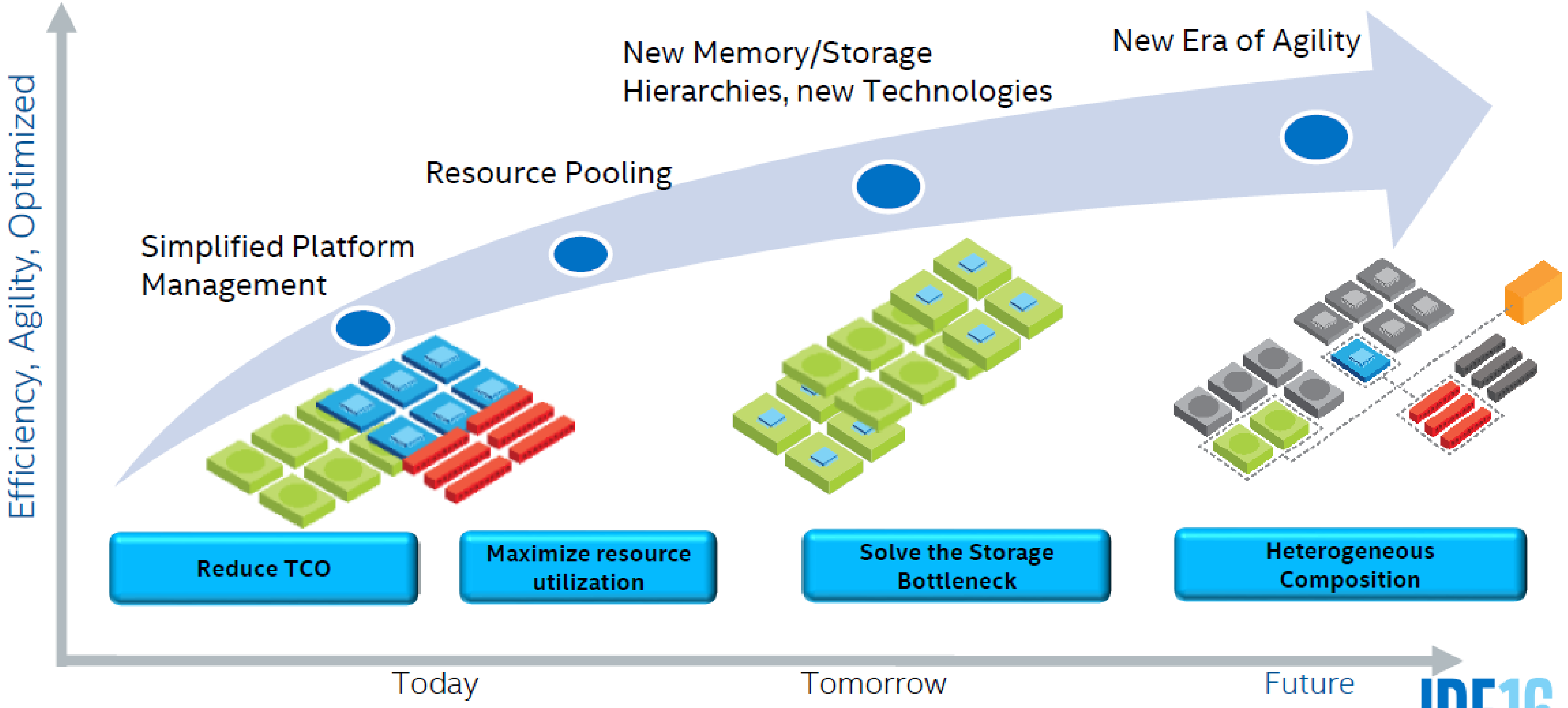
Rack Mgmt.

Rack Temp/ Power

Network Switch

Comprehensive management architecture

Evolution of Rack Scale Design



Efficiency, Agility, Optimized
Simplified Platform Management
Resource Pooling
New Memory/Storage Hierarchies, new Technologies
New Era of Agility
Reduce TCO
Maximize resource utilization
Solve the Storage Bottleneck
Heterogeneous Composition
Today
Tomorrow
Future

IDF16 INTEL DEVELOPER FORUM

My view of RSD future

- RSD will be based on commodity interconnects
- PCIe fabric based on Cu will stay in a niche
- Ethernet (25/50/100/200 GbE) with optical lines will dominate the Rack interconnect
- NVMf (NVMe over Fabric) will enable a flexible storage pool assignment

- RedFish and SwordFish will become the standard interfaces for IT infrastructure
- The integration of RDS with OpenStack will further evolve

(1) Shared Power/Cooling, HDD-Pooling integration

(2) PCIe NVMe pool, pooled FPGA,
SDI Orchestration layer integration

(3) NVMe over Ethernet pool

(4) GPGPU pools

(5) Memory pool ??
3D-Xpoint / DRAM

FUJITSU

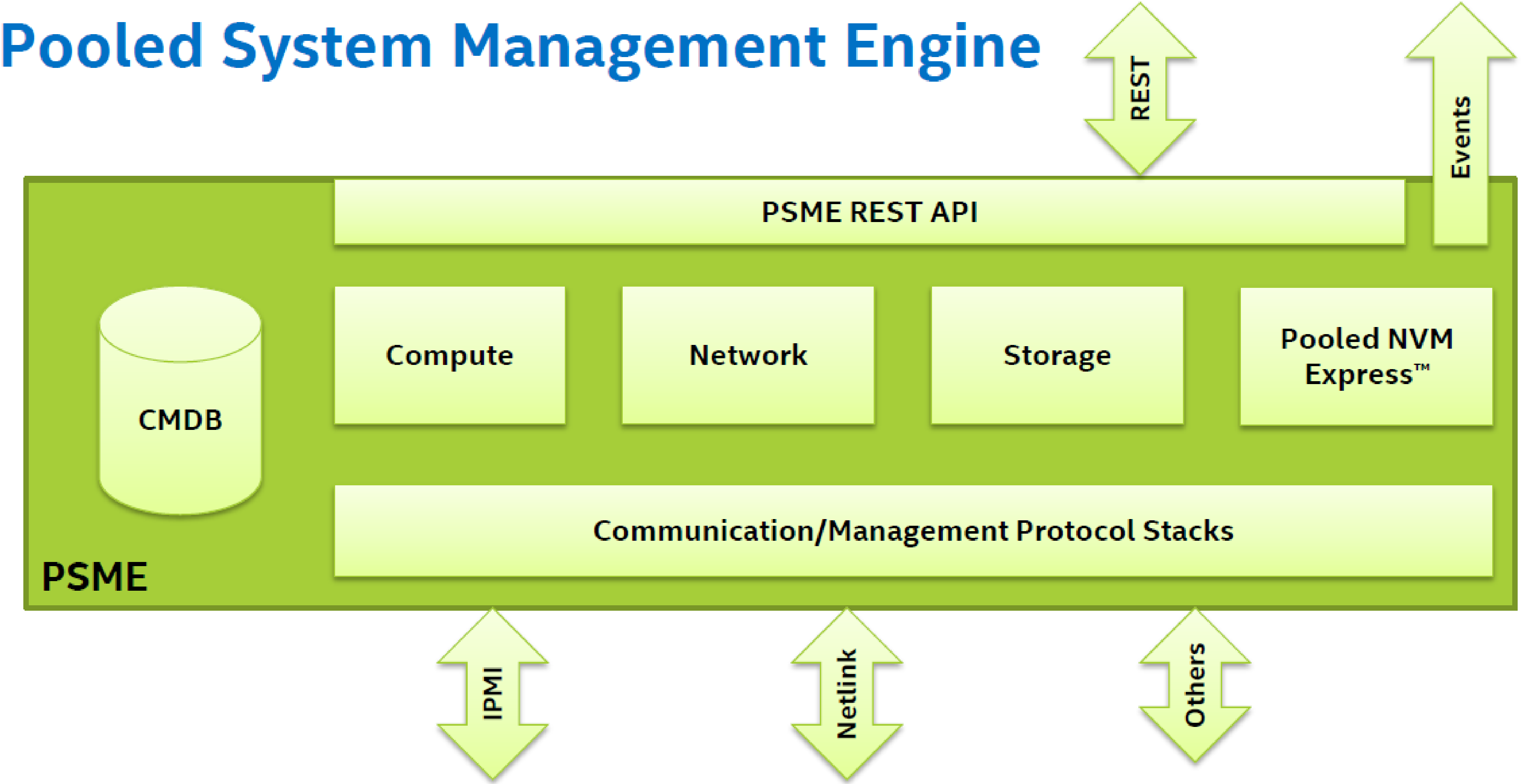
shaping tomorrow with you

Fujitsu Technology Solutions

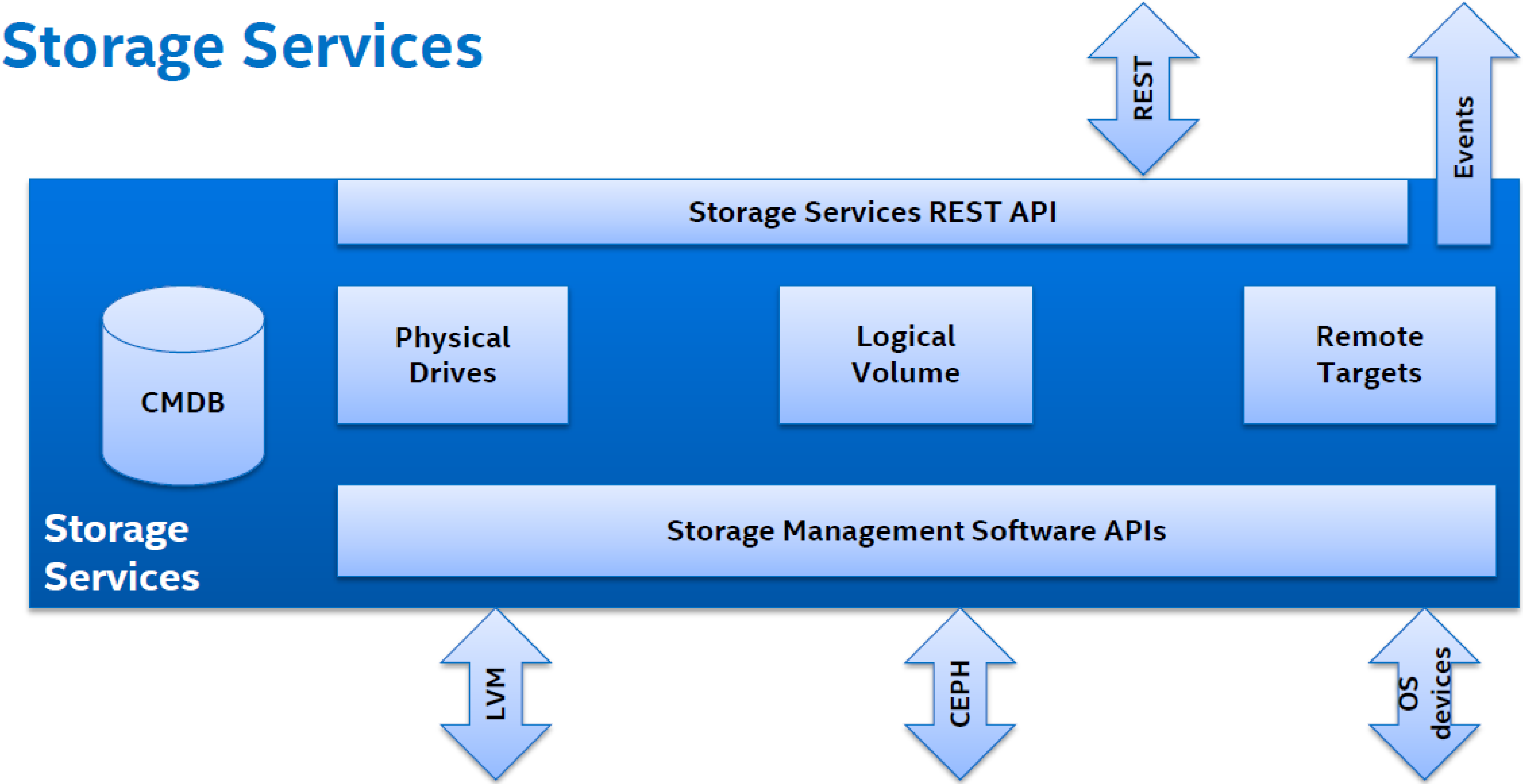
Dieter.Kasper@ts.fujitsu.com

v8

Pooled System Management Engine

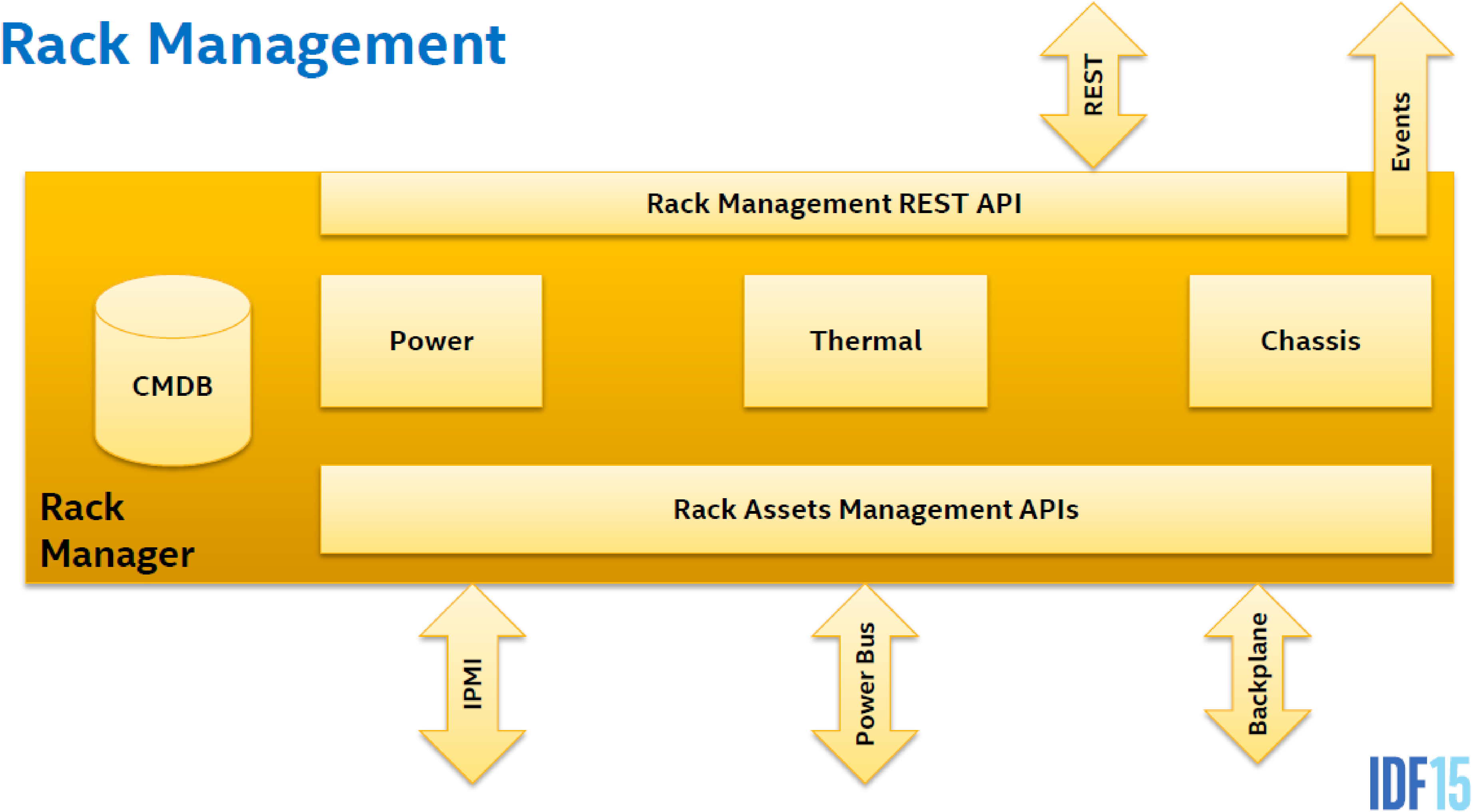


Storage Services

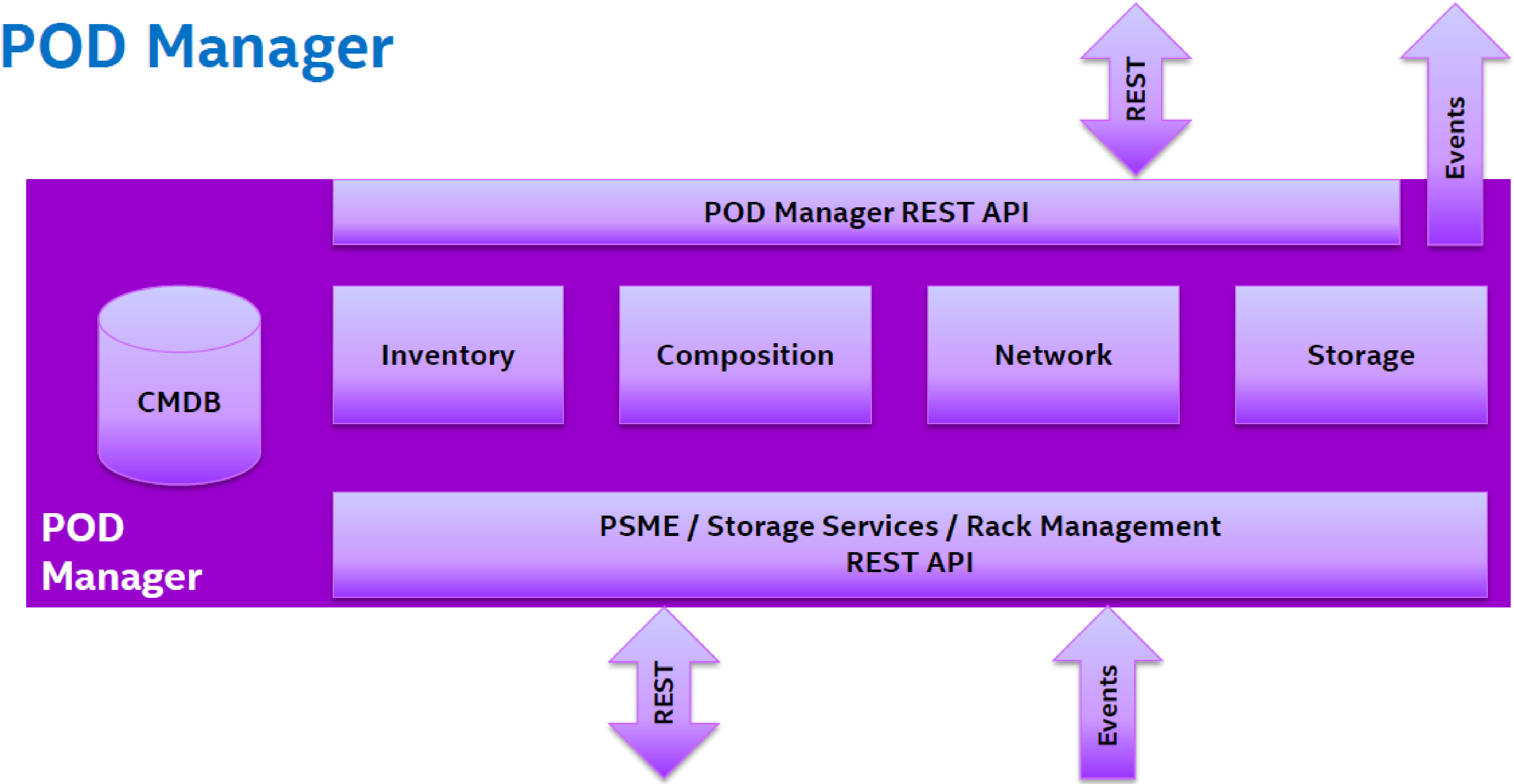


IDF15 INTEL DEVELOPER FORUM

Rack Management



POD Manager


REST
Events
POD Manager REST API
CMDB
Inventory
Composition
Network
Storage
POD Manager
PSME / Storage Services / Rack Management REST API
REST
Events