Evolution of Rack Scale Architecture
Storage

Murugasamy (Sammy) Nachimuthu, Principal Engineer
Mohan J Kumar, Fellow
Intel Corporation
• Introduction to Intel® Rack Scale Design
• Storage in Intel® Rack Scale Design (Today and Future)
• Intel® Rack Scale Design Storage Orchestration
• Summary
Agenda

• Introduction to Intel® Rack Scale Design
• Storage in Intel® Rack Scale Design (Today and Future)
• Intel® Rack Scale Design Storage Orchestration
• Summary
Data Center Challenges
Infrastructure has not kept up with increasing business demands

Business Needs
- **Reduce** operational and capital expenses.
- **Deliver** new services in minutes, not months.
- **Optimize** data center based on real-time analytics.
- **Address** application workload needs with agility.
- **Scale** capacity without interruption.

Inefficiency
Less than 50% server utilization

Growth
Data growth doubles every 18 months

Agility
New services can take a week or more to provision

---

1. Worldwide and Regional Public IT Cloud Services 2013–2017 Forecast. IDC (August 2013) [idc.com/getdoc.jsp?containerId=242464](idc.com/getdoc.jsp?containerId=242464)
Intel® Rack Scale Design

Logical architecture for efficiently building and managing cloud infrastructure—and providing the simplest path to a software defined data center.

Benefits

- >25% decrease in capital costs
- Increase capacity/IT $
- Reduce time to cloud deployment

Increase performance per TCO$ & accelerate cloud adoption
1. Pooled systems
2. Pod management
3. Network fabric
4. Pod-wide Storage

Modular scalable management architecture
• Asset & location discovery
• Disaggregated resource management
• Composable system support
• Support compute, network, and storage,

Comprehensive management architecture
RSD Management - Example

- Location aware compute, network, storage
- Pooled System Management Engine per Rack Scale Drawer (Multi-node aggregation)
- Environmental Management using Rack Management Module

POD Manager

Rack Scale Rack

PSME – Pooled System Management Engine
RMM – Rack Management Module

NIC attached to external network
NIC attached to private management network
Switch
Services outside the Rack Scale Rack

Rack Private Management Network

Redundancy allowed, Not Shown
Agenda

• Introduction to Intel® Rack Scale Design
• Storage in Intel® Rack Scale Design (Today and Future)
• Intel® Rack Scale Design Storage Orchestration
• Summary
Storage Trends

Processor Node

Memory Tier

Local Tier

Hot Tier

Warm Tier

Cold Tier

COST-OPTIMIZED DATA ARCHIVE

SHARED CAPACITY STORE

SHARED WORKING SET

CACHED RKING SET

COLD TIER

WARM TIER

SHARED CAPACITY STORE

COST-OPTIMIZED DATA ARCHIVE

Processor

Node

APPS

DRAM

NVDIMM

NV Storage
(3D XPoint™, NAND, …)

HDD

Capacity

Performance
• Storage node is a collection of storage drives and storage controller
• Storage nodes are identified and exposed to Orchestration layer
• Comprehends storage pooling
• Today SAS/SATA drives are widely used
• Storage node based on Ethernet fabric
Local vs. Pooled Storage

Pooling Value

Consolidation of Storage

Consolidation of compute

IOPS/SVR > Local Capacity

IOPS/SVR = Local Capacity

Avg IOPS/SVR = Local Capacity

Avg IOPS/SVR < Local Capacity

Diverse Workload deployment

Local Deployment model?

Pooled Infrastructure cost

Op Ex Value of pooling

WL/SVR

Less than capacity deployed

More than capacity deployed

Local Capacity

Consolidation

Avg
IOPS/SVR

<

Operation Value of pooling

More than capacity deployed

Local Capacity

Pooled Infrastructure cost

Consolidation

Avg
IOPS/SVR

Less than capacity deployed

More than capacity deployed

Local Capacity

Consolidation

Avg
IOPS/SVR

<

Operation Value of pooling

More than capacity deployed

Local Capacity

Pooled Infrastructure cost

Consolidation

Avg
IOPS/SVR

Diverse Workload deployment
RSD Pooled NVMe Controller (PNC)

- Enable pooling of NVMe devices
- Assign high performance storage to nodes based on workload demand
- Prevent SPOF through host failover
- Enables ease of workload migration in hyperscale cloud environment
- Enables better utilization of DC resources by allowing composable high performance IO capacity
RSD Pooled NVMe Controller (PNC) (cont...)

Assign Drive 2 to Node 1

Assign Drive 1, Drive 3 to Node 2

Logical effect of the assignment
• Configures the NVMe Over Ethernet Targets and binds the host and targets
  • Network access path established
  • Storage NameSpace assigned
  • Security policies established
• Assign QOS to the NVMe over Ethernet traffic
• Monitors drive health
### NVDIMM Types

<table>
<thead>
<tr>
<th>NVDIMM-N</th>
<th>NVDIMM-F</th>
<th>NVDIMM-P</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Diagram" /></td>
<td><img src="image2.png" alt="Diagram" /></td>
<td><img src="image3.png" alt="Diagram" /></td>
</tr>
</tbody>
</table>

- **NVDIMM-N**
  - Only DRAM is addressable by SW
  - NV Media acts as backup for DRAM
  - NV Media not addressable
  - At least 1:1 Capacity Ratio between DRAM & NV Media
  - Tracks DRAM latency & memory channel BW for Read and Write

- **NVDIMM-F**
  - No DRAM
  - NV Media is directly addressable via Window mechanism
  - Tracks NV Media latency
  - Benefits from memory channel bandwidth

- **NVDIMM-P**
  - Combination of NVDIMM-N and NVDIMM-F
  - Flash memory beyond that needed for persistence is accessible as block
### RSD Comprehends Memory Including NVDIMMs

- **DIMM Type and Sizes**
  - **Two NVDIMM-N 16GB**
    - DIMM1
    - DIMM2
  - **Three DDR4 DIMMs 8GB**
    - DIMM3
    - DIMM4
    - DIMM5
  - **One NVDIMM-F 8GB**
    - DIMM6

- **Memory Chunks**
  - **#1** – 8GB Volatile
  - **#2** – Not present
  - **#3** – 8GB PMEM
  - **#4** – 2GB Mirrored Volatile
  - **#5** – 8GB Volatile
  - **#6** – 8GB Volatile
  - **#7** – 8GB Block

---

**Flash Memory Summit**

**RSD Memory - Example**

- **CPU1**
  - DRAM DIMM
  - NVDIMM_F
  - NVDIMM_N

- **CPU2**
  - DRAM DIMM
  - NVDIMM_N

- **NVDIMM-N**
  - Region1
    - V (4GB)
  - Region2
    - V (4GB)
  - Region3
    - PM (4GB)
  - Region4
    - PM (4GB)

- **NVDIMM-N**
  - Region1
    - V (4GB)
  - Region2
    - V (4GB)
  - Region3
    - PM (4GB)
  - Region4
    - PM (4GB)

- **DIMM1**
  - Region1
    - V (4GB)
  - Region2
    - V (4GB)
  - Region3
    - PM (4GB)

- **DIMM2**
  - Region1
    - V (4GB)
  - Region2
    - V (4GB)
  - Region3
    - V (4GB)

- **DIMM3**
  - Region1
    - V (2GB)
  - Region2
    - V (2GB)
  - Region3
    - V (4GB)

- **DIMM4**
  - Region1
    - V (2GB)
  - Region2
    - V (2GB)
  - Region3
    - V (4GB)

- **DIMM5**
  - Region1
    - V (8GB)
  - Region2
    - V (8GB)

- **DIMM6**
  - Region1
    - V (8GB)
Agenda

• Introduction to Intel® Rack Scale Design
• Storage in Intel® Rack Scale Design (Today and Future)
• Intel® Rack Scale Design Storage Orchestration
• Summary
Intel® Rack Scale Design Orchestration

Orchestration Layer

- Compute Controller
- SDN Controller
- SDS Controller

Intel® Rack Scale Design Resources (PODM)

Service/Tenant Portal

- Users
- Compute
- Switch
- Storage
• Rack Scale Storage related elements are exposed to the Pod Manager through Intel® Rack Scale Design API
  • RSD API exposes storage availability zones to Orchestration layer
• Orchestration participates in configuring and managing the pooled storage configuration
• Introduction to Intel® Rack Scale Design
• Storage in Intel® Rack Scale Design (Today and Future)
• Intel® Rack Scale Design Storage Orchestration
• Summary
Evaluation of Rack Scale Design

New Memory/Storage Hierarchies, new Technologies

Resource Pooling

Simplified Platform Management

Efficiency, Agility, Optimized

Reduce TCO
Maximize resource utilization
Solve the Storage Bottleneck
Heterogeneous Composition

Today
Tomorrow
Future

Santa Clara, CA
August 2016
Summary

• Intel® Rack Scale Design delivers the Next Generation Data Center Architecture and comprehends advancements in storage

• Intel® Rack Scale Design references available online –

• Intel® Rack Scale Design overview, whitepaper and specifications