WHITE PAPER

Data Center
Rack Scale Design

Intel® Rack Scale Design Architecture

Executive Overview

IT managers, cloud services providers and communications companies are all facing unprecedented challenges from new digital business models, rapid data center growth, and technology changes. The traditional data center built with servers configured for specific applications can't keep pace with the demands of the Digital Economy. This paper describes the Intel® Rack Scale Design (Intel® RSD) hyperscale reference architecture, an open, interoperable approach to *composable disaggregated infrastructure (CDI),* the technology on which next generation hyperscale data centers will be built.

Composable Disaggregated Infrastructure

Intel RSD is an open industry blueprint for the transition to a software-defined, composable, disaggregated infrastructure (CDI). The goal of Intel RSD is to enable more flexibility, manageability, economy and openness for telecommunications, cloud and enterprise service providers.

The key concept behind CDI is to break down today's servers with their fixed ratios of compute, storage and networking resources into separate resource pools that can be interconnected, or "composed," on demand into logical systems or "nodes" optimized for specific applications. When a composed node is no longer needed, its resources can be released back to the resource pools for use by another workload. This enables more efficient use of hardware resources and the ability to quickly react to changes in workload requirements. The advancing speed of fabrics like 100G Ethernet, Intel® Omni-Path and optical interconnects make disaggregation and composition possible without sacrificing performance. With comprehensive, open management APIs, composability

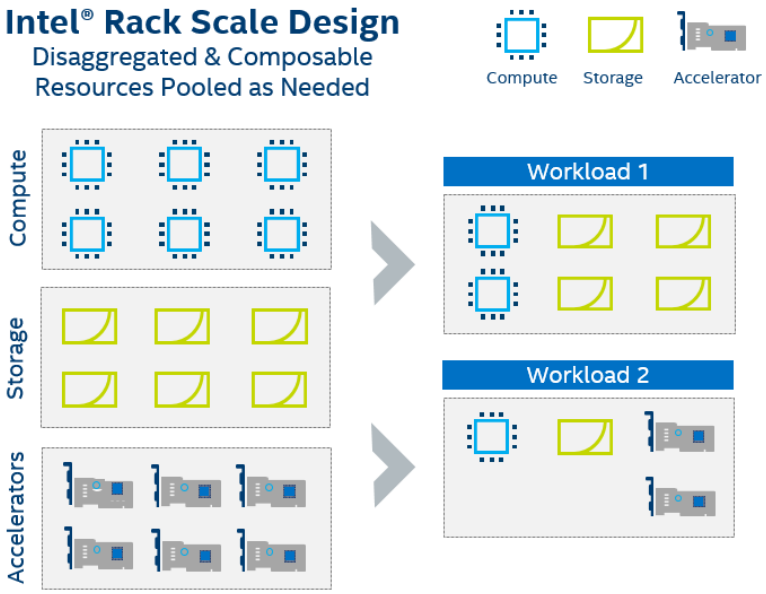


Figure 1. Intel® RSD is a disaggregated architecture that enables user specified systems to be composed on the fly.

enables intelligent software to dynamically optimize data center hardware for best performance and utilization in real time. To summarize, the benefits of CDI include:

- Faster, easier scale out and lower refresh costs with savings in both capex and opex
- Greater agility in application development, provisioning, and life cycle management
- Higher efficiency due to better resource utilization, reduced overprovisioning and dynamic workload tuning, all leading to lower TCO
- Lower capex resulting from independent upgrade cycles (i.e., only the targeted resource needs to be replaced, not a whole server)—you get more capacity per dollar spent
- Optimized performance via custom configurations including fast non-volatile memory and accelerators
- More automated infrastructure management, and more efficient use of staff

Intel® Rack Scale Design Architecture Overview

As noted above, Intel RSD is a disaggregated architecture, meaning that various data center hardware resources, such as compute modules, non-volatile memory modules, hard disk (HDD) storage modules, FPGA modules, and networking modules, can be installed individually within a rack. These can be packaged as blades, sleds, chassis, drawers or larger physical configurations—it's up to the solution provider. Resources being managed will have at least two network connections: a high-bandwidth data connection (Ethernet or other fabric), and a separate out-of-band Ethernet link to a dedicated management network. Typically (but not necessarily), these two networks connect to top-of-rack switches. Through these switches, a rack can be connected to other racks making up a management domain called a "pod." Pods are linked together using whatever network topology is determined by the data center administrators.

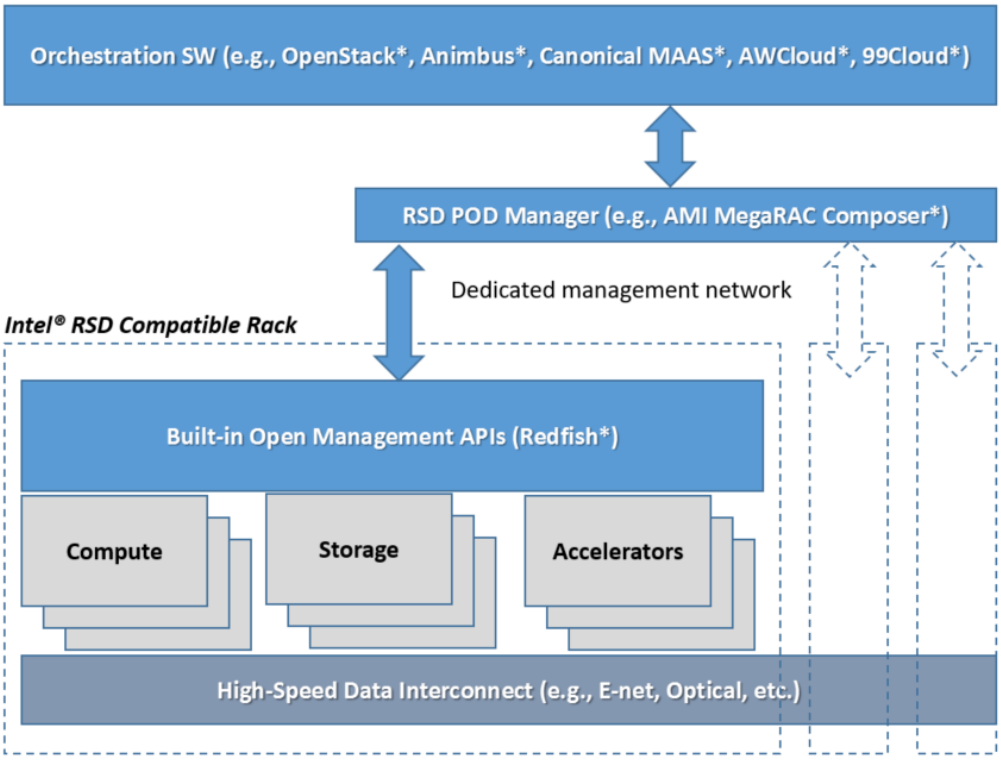


Figure 2. Intel RSD compatible hardware components communicate with RSD Pod Manager (PODM), which carries out requests for higher level orchestration software via a common set of open APIs based on the DMTF Redfish framework.

The rack data fabric (e.g., 100 Gb Ethernet, Infiniband or optical interconnect) provides data links between resource modules, so a system or "node" can be composed from available resources in the rack. For example, a user may want to compose a node with a 2-socket Intel® Xeon compute module, a 500 Gb Intel® Optane™ module, an Intel® FPGA module and a 4 TB HDD storage module to meet the requirements for a specific application. Once composed, the node can be provisioned with a bare metal stack (operating system and application code), a virtual machine environment such as KVM or VMware*, or a container environment like Docker*. An orchestration software service running somewhere in the data center can access the racks within a pod via common APIs to deploy and manage applications. But how does all this get coordinated in the data center environment?

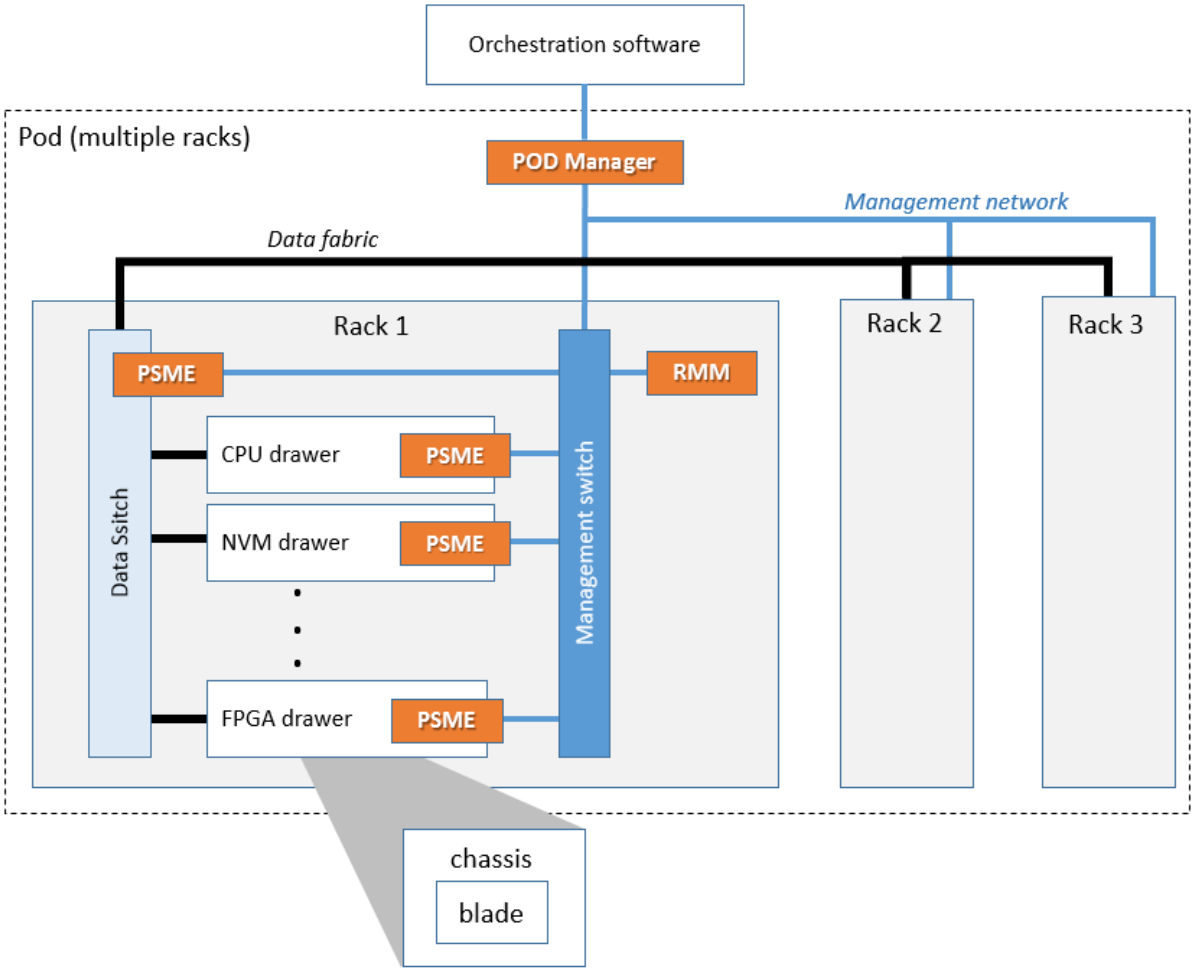


Figure 3. Intel RSD employs software components at the module (PSME), rack (RMM) and pod, i.e., multiple rack, levels (PODM) communicating with each other and higher level orchestration software via a dedicated Ethernet management network.

Composition and other management actions are enabled by software with open APIs compatible with DMTF's Redfish RESTful framework. At the lowest level, each resource module in a rack will have a software component called the Pooled System Resource Manager (PSME). This can be executed on baseboard management controllers (BMCs) built into the resource module, or on other processors in the rack. There may be one or multiple PSMEs in a rack, and the exact location of PSMEs is up to the manufacturer—PSMEs could be in each blade, in each chassis, in each drawer, or there could be one PSME for the entire rack—as long as the implementation provides PSME functionality on behalf of each contained resource and exposes the specified APIs to higher levels via the Ethernet management network. The integration of the PSME components to the physical hardware is implementation

dependent, e.g., it could be implemented in a firmware layer in each blade, or in a multi-blade chassis controller, for example.

The next layer in the architecture is the Rack Manager Module (RMM). RMM communicates with the PSMEs and takes responsibility for a variety of rack-oriented functions including distributing security certificates to PSMEs, managing rack power and cooling, reporting environmental and health status, and designating and reporting the physical location of each asset within the rack.

At the next level, Pod Manager (PODM) software manages functions within and across racks for a single pod. The southbound PODM APIs interact with the PSMEs and the RMMs. The northbound PODM APIs are exposed to the user's choice of orchestration software, such as OpenStack*, VMware or other commercial, open source or custom DCIM software.

Interacting with the PSMEs, PODM discovers all the resources within each rack and records asset information in its database. This information defines resource pools that may be used to compose nodes. PODM exposes asset information to orchestration software via its own northbound APIs. Orchestration software can then request PODM to compose nodes on its behalf based on a set of application requirements, such as processor type, number of sockets, minimum DRAM, required NVM and HDD storage capacity, etc. In response, PODM identifies unused pooled resources that meet the minimum requirements and makes requests to the PSMEs to perform operations required to compose the node (e.g., setting switch parameters, mapping IP addresses, creating name spaces and storage volumes, etc.). It then returns information to the orchestration layer to enable provisioning and workload distribution on the composed node.

PODM has the ability to power on and boot modules, and initiate bare-metal provisioning on behalf of the orchestration layer. Alternately, a software environment like OpenStack or VMware may have its own bare metal provisioning service that can work with POD Manger. Once basic provisioning is completed, the orchestration layer can install a virtual environment on the composed node, or just run a bare metal application stack. For special workloads, PODM can compose nodes that include a variety of accelerators, such as Intel FPGAs. It can also assemble resources such as Intel® Optane™ storage-class memory, Intel® 3D NAND SSDs and HDDs to provide the basis for a high-performance hot-warm-cold storage hierarchy.

The Intel RSD Roadmap

Moving the industry to adopt a new data center architecture and a comprehensive set of common management APIs is not a single event—it's a journey. It takes time to develop, validate and approve standards. Suppliers need to integrate the new blueprint into their product lines, and customers adopt the new infrastructure over time as part of their planning roadmap.

Intel, along with Broadcom*, Dell*, Ericsson*, Hewlett-Packard Enterprise*, Lenovo*, Supermicro* and VMWare*, was one of the original founders and promoters of the DTMF's* Scalable Platforms Management Forum. This group, launched in September 2014, is responsible for the Redfish* standard, and associated extensions such as Swordfish* (based on SNIA* standards) for storage management, and the Yang*-to-Redfish effort for network management. Intel RSD is based on and extends the Redfish standard. Intel extensions are submitted to the Redfish committee for inclusion in upcoming versions of the standard.** Over time there have been several releases on the Intel RSD roadmap and each release has specific value for end users. Here is a summary of the releases to date and a preview of upcoming releases on the Intel RSD roadmap (subject to change):

Intel RSD 1.2—replaces the outdated IPMI management standard with a modern, scalable, secure management framework based on Redfish internet-friendly RESTful APIs and JSON style payloads. Redfish simplifies the management of resources in a hyperscale environment, fixes security vulnerabilities inherent in IPMI, adds capabilities needed to manage new technologies, and provides an open interface for interoperability across data center hardware and management software.

Table 1. There have been several releases on the Intel® RSD roadmap, and each release has specific value for end users. Future releases will continue to build on the existing baseline and provide additional capabilities.

Version 1.2	Version 2.1	Version 2.2
Key Benefits		
Modern Manageability	Storage Pooling over PCIe	Advanced Management Features (Telemetry and TPM)

• Modern and Open Hardware Management APIs (*Redfish**)
• Rack-level Power and Cooling Management
• Pod-Level Architecture and Orchestration Capability
	• Physical Storage Disaggregation and Composability
	• Storage Pooling over PCIe (Direct-Attach)
	• Automated System Conformance Test Suite
		• Intel® Xeon® Processor Scalable Family support:
		  • Out-of-Band Telemetry
		  • Trusted Platform Module (TPM) Support

Intel RSD 2.1—enables disaggregation of CPU/DRAM and NVMe storage resources, enabling NVMe* pooling over high-speed PCIe interconnects. This provides greater flexibility and improved utilization of NVMe drives for applications without compromising bandwidth and latency. It also allows users to refresh CPU/DRAM or storage (HDD or NVM) independently to improve capacity and performance without replacing and re-provisioning entire servers. Studies have shown that this can reduce refresh cost by as much as 50%[1].

In version 2.1, Intel also provides a Conformance Test Suite (CTS), which can be used by OEMs and other suppliers to validate that Intel RSD-based systems are fully compliant with the specifications. Compliance testing provides customers with greater assurance that their systems harness the full benefits of Intel RSD architecture, including interoperability for those considering a mix of Intel RSD compliant racks procured from different vendors.

Intel RSD 2.2—supports the Intel® Xeon® Scalable Family platform and out of band (OOB) discovery of features such as Trusted Platform Module (TPM), telemetry and PCIe FPGA.

OOB telemetry provides detailed information about the health and utilization of infrastructure components to aid in capacity, performance and power/thermal management. The agentless and OOB implementation to gather telemetry allows very minimal performance impact on running applications. This version also discovers and provisions servers with add-in-card PCIe FPGA so that orchestration software can leverage Intel RSD APIs to assign FPGA resources for specific application workloads.

Intel RSD 2.3 (expected release in first half of 2018)—enables composition of CPU/DRAM and storage resources using NVMe over Fabrics*with RDMA on Ethernet. This provides a high-performance, long-distance interconnect that extends scalability across racks, enabling NVMe pooling for practically unlimited provisioning of "hot and warm tier storage" capacity for an application. This version also implements SNIA Swordfish management APIs to support storage management functions for NVMe over Fabrics targets.

Future releases—will support disaggregation and pooling of FPGA accelerators over PCIe interconnects. The next generation PCIe FPGA card from Intel (expected in the first quarter of 2018) will provide a scalable, OOB management interface compatible with Intel RSD. It will support discovery, composition and decomposition of FPGAs, and provision greenbit/bluebit streams via Intel RSD APIs.

Upcoming releases will also support discovery, provisioning and configuration of Intel's next generation persistent memory on the Intel® Xeon® Scalable Family refresh platform, and will consolidate software defined network functionality by recasting Yang interfaces into the Redfish RESTful API framework. Over time Intel RSD will continue

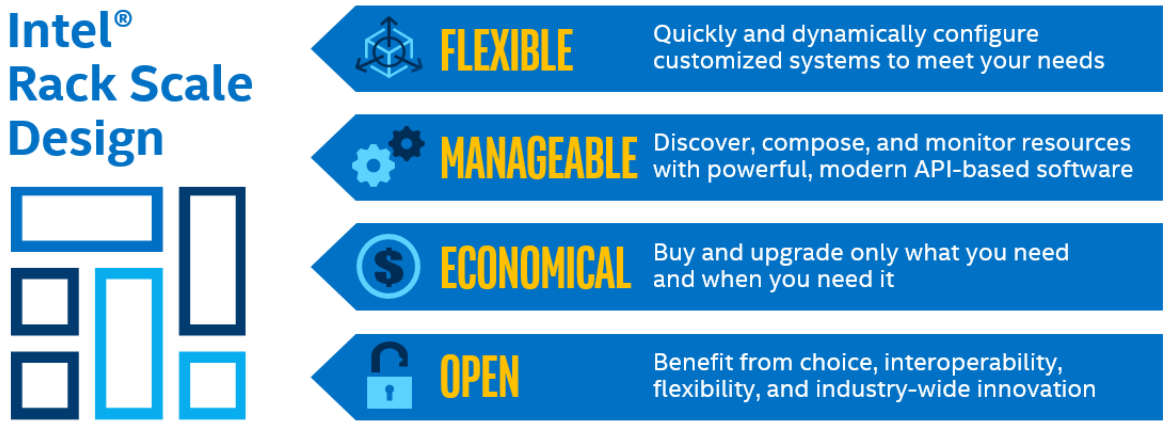
to extend disaggregation and composability to additional resources such as Intel® Nervana™ family processors, and will adapt as needed to enable new segments such as high performance computing (HPC).

Product Availability

Intel RSD based products are available today from many suppliers including Dell EMC*, Ericsson*, HPE*, Huawei*, Inspur*, Quanta Cloud Technology*, Radisys*, Supermicro*, Wiwynn* and others. For more information, please go to intel.com/intelrsd.

Conclusion

Intel RSD is an open, industry-wide hyperscale reference architecture for the transition to composable, disaggregated infrastructure (CDI) with full software control. In an environment of explosive data center growth, Intel RSD offers greater agility and interoperability, while reducing TCO. Learn more about how Intel RSD can accelerate your transition to CDI at: intel.com/intelrsd. Or, contact your local Intel representative to discuss how Intel can help you to meet the challenge of the Digital Transformation. For more information, please go to intel.com/intelrsd.



---

[1]Disaggregated Server Architecture, Shesha Krishnapura, Intel Fellow and Intel IT CTO, 2017.

****A Note on Redfish Compatibility**

It is important to note that Intel RSD is aligned with the Redfish industry standards effort. Intel RSD employs Redfish and extends it to support composable disaggregated infrastructure. Intel RSD extensions are regularly submitted to the Redfish Scalable Platforms Management Forum as proposals for inclusion in the official Redfish standard. This process results in some delay between the latest Intel RSD APIs and the official Redfish APIs, but it is designed to ensure convergence between Intel RSD and Redfish over time. The Forum is also working on future Redfish extensions such as Swordfish (based on SNIA standards) for storage management, and Yang-to-Redfish for network management, based on the IETF Yang network management standard.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps. No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. Copyright © 2018 Intel Corporation. All rights reserved. Intel, the Intel logo, Xeon, Xeon Phi, Nervana and Optane are trademarks of Intel Corporation in the U.S. and/or other countries.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request. No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document. Intel technologies' features and benefits depend on system configuration and may require enabled hardware or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer.

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