November 2022

CONVERGED RACK ARCHITECTURE (CRA) FOR DATA CENTER

Anant Thakar

Rakesh Bhatia

CONVERGED RACK ARCHITECTURE (CRA) FOR DATA CENTER

AUTHORS:
Anant Thakar
Rakesh Bhatia

ABSTRACT

The data center hardware (HW) infrastructure in public and private clouds is going through a paradigm shift due to the demands of a heterogenous computing model. The lower-cost economics of a private cloud need to be augmented with the operational simplicity of a public cloud. The Open Compute Project (OCP) started with the goal of simplifying the compute aspect and they did succeed to some extent, but enterprise adoption has been slow due to the HW complexity of various elements and version compatibility. Since the evolution of OCP, a number of technology changes have occurred, and a new infrastructure standard is necessary to tackle the heterogenous computing model along with a thermal and high-speed input/output (I/O) interconnection nexus. Techniques are presented herein that support a converged rack architecture (CRA). Such an architecture encompasses a universal rack I/O paradigm comprising embedded opto-electrical data I/O posts, centralized power and cooling, and a common rack management unit (RMU) which allows a rack itself to be thought of as a compute unit where different compute, network, and storage components may be composed in a virtual domain on the fly and then decomposed when they are not required. In essence, a CRA manages and treats a rack frame as a complete converged system of compute, network, and storage elements and not as disparate entities.

DETAILED DESCRIPTION

As an initial matter, it will be helpful to confirm an element of nomenclature. In a (e.g., data center) equipment rack, the measure "U" stands for a ‘unit’ of vertical rack space (or equivalently, "rack unit" (RU)) that is approximately 1.75 inches high. A typical rack frame is 75 to 80 inches tall (depending upon the manufacturer) which means that it can hold 42 1U-sized devices, or 21 2U-sized devices, or etc. In other words, a rack frame may be defined as a 42U mechanical enclosure that provides critical base services (such as

power, management, etc.) and into which different components (such as routers, switches, and servers) may be installed

A data center rack, as described above, comprises compute servers, storage facilities, and networking boxes. The electromechanical structure that is employed in designing such units is usually in 1U, 2RU, and xRU terms, with each unit enclosed in its own metal envelope. For example, in a compute server the choice is typically 1U or 2RU. Alternatively, a multi-U blade chassis may comprise 7RU. An information technology (IT) buyer must determine, upfront, whether to buy a rack server or a blade system for provisioning. Once a decision is taken, it is expensive to go back and forth even if the workload demand changes. Besides compute servers, a rack consists of storage, network switches, routers, and specialized network services such as a firewall and server load balancing (SLB) appliances. All of these elements are interconnected using expensive input/output (I/O) cabling at the back side of a rack frame. These divergent yet connected systems have separate control and management domains. On the hardware (HW) side, common to these boxes are data I/O connectivity cables, power supplies and associated cables, and high-speed fans for thermal management. Within such an arrangement there is a significant replication of HW resources which adds to the carbon footprint of a data center.

All of the above-described attributes present a number of challenges.

A first challenge concerns reducing the number of replicated HW elements. Examples include intra-rack fame cabling for the electrical I/O interconnection for both data and management, the multitude of power supplies and associated cabling, fan modules, logic replication, etc.

A second challenge concerns offering operational simplicity in managing a rack frame as a converged system and not as just a collection of elements. For example, of interest is how one may dynamically define and configure a virtual blade chassis and/or a virtual rack along with all of the services (such as networking, SLB, routing, etc.) without form factor limitations from common HW elements.

A third challenge concerns minimizing the replication of control and management plane HW and software (SW) while the majority of functions (such as fans, power management, board resets, firmware updates, security compliance, application programming interfaces (APIs), logs, and Wake-on-LAN (WoL) capabilities) remain the

same. The implementation of SW and HW varies and brings with it its own challenge in terms of software updates, defects, and security compliance.

The HW infrastructure that is driving cloud computing and enterprise computing is primarily based on an Open Compute Project (OCP) for cloud vendors and proprietarily products such as blade and multi-modular niche appliance servers for enterprise vendors. Enterprises and the new edge computing market wish to lower the cost economics of cloud-based computing while, at the same time, obtaining the simplicity of on-premise or edge deployments all while embracing the open-sourced HW infrastructure. The OCP started with the goal of simplifying the compute aspect and they did succeed to some extent, but enterprise adoption has been slow due to the HW complexity of various elements and version compatibility. Since the evolution of OCP, a number of technology changes have taken place and a new infrastructure standard is necessary to tackle the heterogenous computing model along with the thermal and high-speed I/O interconnection nexus. Broadly speaking, it is necessary to think beyond a physical blade or a 1U/2U server modular chassis concept and consider a rack itself as the compute unit where the compute, network, and storage elements may be composed in a virtual domain, on the fly, and then decomposed when not required.

To address the challenge that was described above, techniques are presented herein that support a new standardized converged rack architecture (CRA) that spans a common electrical I/O interconnection nexus among various components such as power, cooling, and electrical I/O.

The main features of a CRA, which will be described and illustrated in the below narrative, include prewired(less) I/O hub posts carrying copper or silicon photonics; the four rack posts embedding the active logic plus power; compute, network, and storage HW cell units; smart centralized cooling; no hot/cold isles; centralized power (e.g., 48 volt (V) or 54V) distribution across a rack frame; an ability to create virtual blade servers or U servers from common HW cell units; and a rack management unit (RMU).

According to the techniques presented herein, a CRA comprises a series of major blocks as depicted in Figure 1, below.

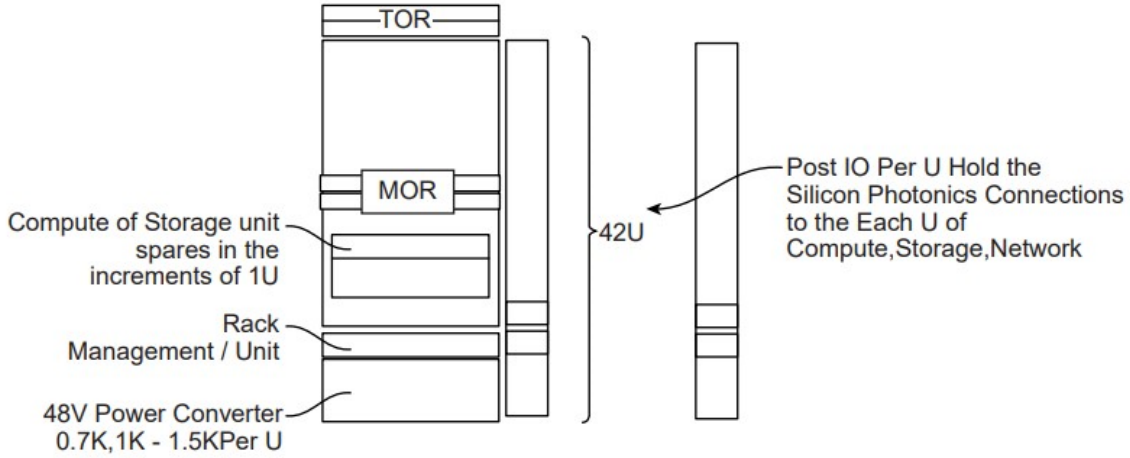

Figure 1: Exemplary CRA Organization

As illustrated in Figure 1, above, a CRA may encompass a 42RU chassis comprising 36U (2 x 18U) slots for compute, storage, and router components; 3RU slots for a power converter (e.g., in the range of 20K to 30K); a 1U slot for an RMU (supporting boot installation, provisioning, etc.); and 2U (2 x 1U) slots for a middle of the rack (MoR) section. Alternatively, such a chassis may comprise 36U (2 x 18 x 1U) slots for compute components; 2U (2 x 1U) slots for a MoR switch; a 1U (2 x ½ width) slot for an RMU; and 3U slots for a power converter. Additionally, the posts of a CRA chassis may be integrated with photonic or copper technology I/O. Further, a CRA may support switched general I/O (such as Video Graphics Array (VGA) ports, serial ports, Universal Serial Bus (USB) ports, etc.). A CRA may comprise a direct current (DC)-ready 48V power distribution block for each RU. A CRA may also contain centralized cooling (comprising rear rack door fans and a heat exchanger thus eliminating the need for individual fan units thereby improving energy efficiency, yielding a simpler unit design, providing higher reliability, and offering operational simplicity).

According to aspects of the techniques presented herein, and as described and illustrated above, a CRA frame is a 42U mechanical structure which can accommodate U-level HW cell units. The posts of a rack may be mounted with embedded I/O hubs which are prewired(less) for all of the connectivity and power. Such an approach eliminates the need for intra-frame cables for connecting heterogenous elements within a rack, as are

often used in some solutions that can be cable intensive while also potentially lacking available bandwidth.

As described and illustrated above, a CRA supports a range of common HW cell units. Such units may include compute, storage, and memory elements; a pool of graphics processing units (GPUs); an SLB facility; a firewall; and specialized network services. These HW cell units may be inserted at the U level into a rack frame and may plug into the prewired I/O posts. Additionally, common HW cell units may be defined in terms of half- or full-size for compute and storage elements. Common HW cell units weigh less and are more efficient, as the rest of the alternating current (AC) power and cooling is centralized, thus greatly simplifying the design.

As indicated in Figure 1, above, a network switch may be implemented in the MoR section instead of at the usual top of the rack (ToR) location. Such an approach makes equidistant the signal integrity (SI) load in the case of electrical I/O connections. All of the interconnected traffic between different HW cell units may flow through such a switch, which may support low latency and a converged fabric type.

A CRA may encompass a common power unit with an embedded backup. An entire rack frame may be distributed with 48V DC power, thus eliminating the need for individual power supply units (PSUs). Such an approach incurs less power loss when compared to an AC/DC 12V distribution. Alternatively, an entire rack frame may be distributed with 48V or54V DC from a ToR-based centralized AC-to-DC converter that is fed from a grid PSU, thus eliminating the need for discrete AC-to-DC PSU units and simplifying the associated cabling.

A CRA may also include an RMU to support asset management, discovery, boot image storage and management, an integrated terminal server, a switched keyboard, video, mouse (KVM) facility, configuration management, etc. An RMU module unifies an asset discovery process, management, and telemetry. It may also support operating system (OS) boot provisioning, a configuration manager, and an overall gatekeeper. Further, it may also serve as a debugging tool by integrating a terminal server, KVM, etc. into the entre rack frame, thus providing one unified view of a rack frame rather than having a separate management console for each HW cell unit.

On the cooling front, the rear door of a CRA may integrate fan modules to move the air from the front to the back. Such a centralized approach reduces the power consumption that is required for cooling and offers a more efficient way to cool an entire rack. Importantly, there are no fans on the HW cell units. A CRA may encompass a hybrid cooling design (with the ability to work for both air-cooled and liquid-cooled IT equipment), a hybrid heat exchanger design (with the ability to work with a rack-contained liquid distribution unit (LDU) or with datacenter facilities having external chilled water where available), and a self-contained liquid- or air-cooled industry standard rack for usage in datacenters. It is estimated that a 15% power savings may be realized with a liquid-to-air rack-level closed loop system.

Figures 2a through 2d, below, illustrate different elements of an exemplary CRA arrangement according to the techniques presented herein.

Figure 2a: Standard 1400 mm Rack

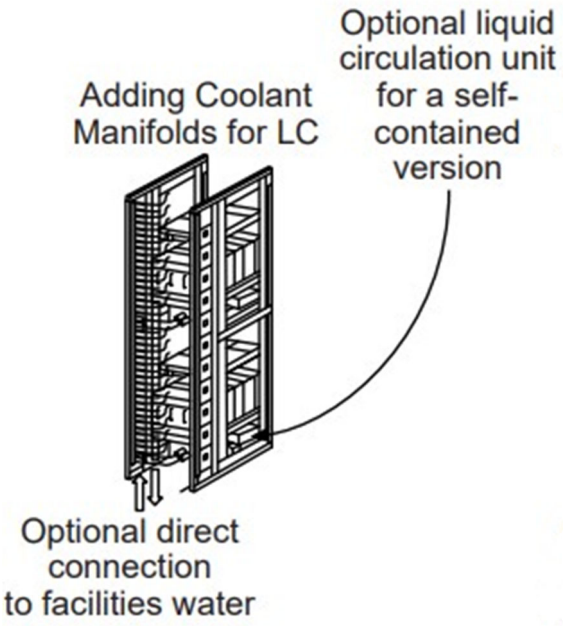

Figure 2b: Coolant Manifolds for LC

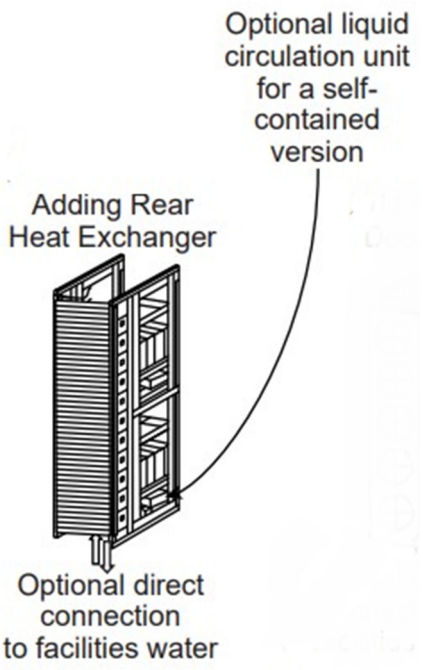

Figure 2c: Rear Heat Exchanger

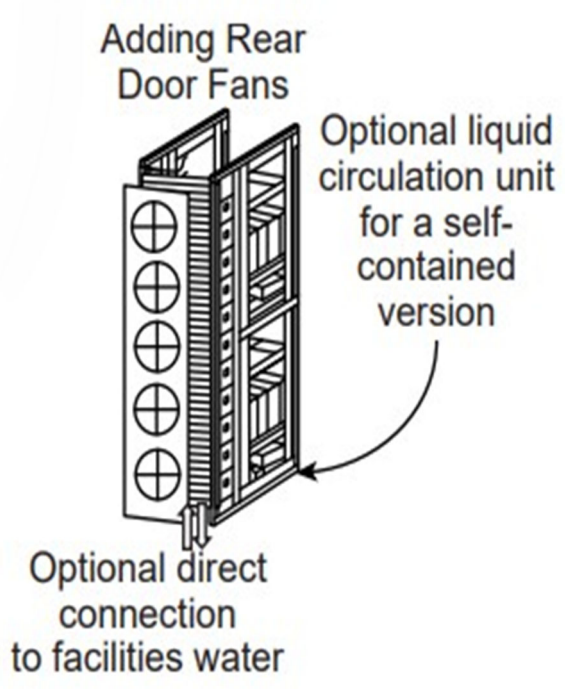

Figure 2d: Rear Door Fans

Each of the Figures 2a through 2d, above, depict a different element of an exemplary CRA arrangement for an 80+ kilowatt (kW) air- or liquid-cooled universal rack. Figure 2a illustrates an industry standard 1400-millimeter (mm) rack that may, according to aspects of the techniques presented herein, be modified to add a number of options. Figure 2b depicts the addition of optional coolant inlet and outlet manifolds every 1U which may support a quick-connect capability for IT equipment (but which will also work for IT equipment that is air cooled and which does not require liquid cooling). Figure 2c illustrates the addition of an optional rear door heat exchanger (which increases exchanger surface area) which may be cooled either by facilities water or by a rack self-contained coolant loop (e.g., a water or refrigerant-based heat exchanger), which can allow fully loaded (e.g., > 40 kW) racks to be deployed in datacenters with a Power Usage Effectiveness (PUE) close to 1.0. Figure 2d depicts the addition of optional rear door mounted fans which, in combination with liquid cooling, may eliminate the need for internal fans in IT equipment.

Central to the techniques presented herein is the design of intelligent I/O hubs that may be embedded inside of a rack's mechanical posts. Figure 3, below, illustrates elements of such an approach according to the aspects of the presented techniques.

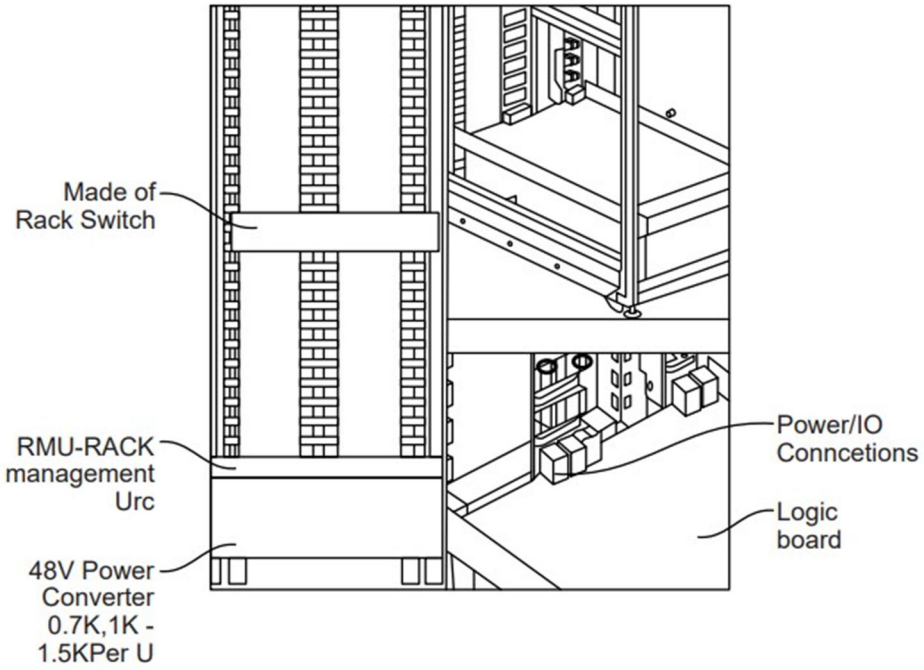

Figure 3: Rack Frame as Converged Chassis

As depicted in Figure 3, above, a frame's posts may serve as intelligent I/O hubs into which compute, network, and storage components may plug. Such posts may contain embedded copper or photonic technology (thus eliminating the need for intra-rack cables), may contain a MoR switch, may contain centralized cooling (comprising rear rack door fans and a heat exchanger thus eliminating the need for individual fan units thereby improving energy efficiency, yielding a simpler unit design, providing higher reliability, and offering operational simplicity) employing a thermally neutral design, may provide 48V distribution in the rack for each slot, and may comprise an RMU.

As described and illustrated above, a CRA's posts (according to aspects of the techniques presented herein) may carry data connectivity to a fabric in a MoR switch, may support miscellaneous I/O connectivity to an RMU, and may provide 48V (or, alternatively, 54V) power distribution.

Such an arrangement eliminates the need for all intra-frame cabling, individual (e.g., AC-to-DC) PSUs, and per-slot cooling. A CRA rack frame may come prewired(less) for installation, so that no new wiring needs to be completed (in contrast to the current practice) consequently saving a significant amount of time and reducing installation costs.

As described previously, common HW cell units may be defined in terms of quarter, half, or full size for compute and storage elements. Common HW cell units weigh less and are more efficient as they integrate only the compute or storage logic elements, with the rest of the AC power and cooling being centralized, thus greatly simplifying the design. An agile or composable infrastructure may be constructed on the fly and then decomposed for different workloads based on the HW cell units.

Figure 4, below, depicts additional elements of an exemplary CRA according to the techniques presented herein and reflective of the above discussion.

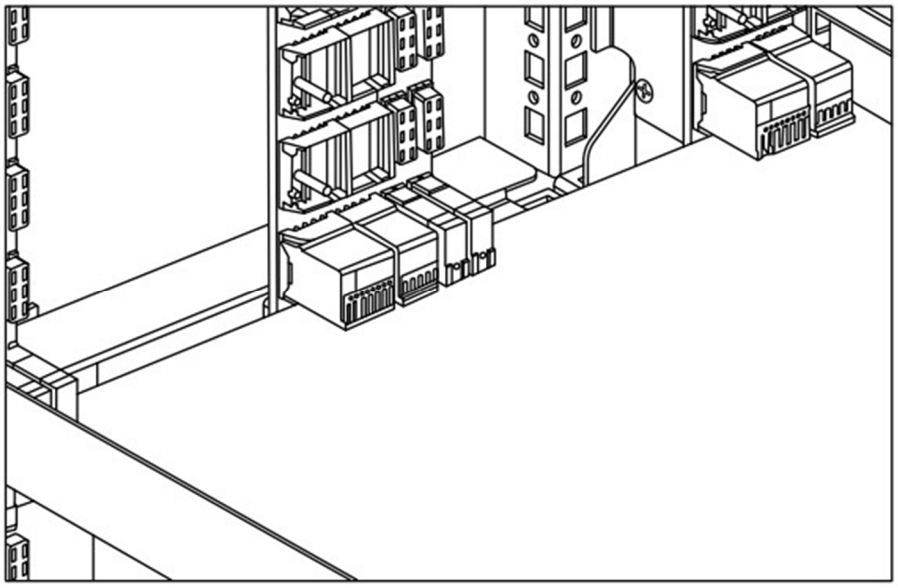

Figure 4: Further Aspects of an Exemplary CRA

A frame RMU may manage discovery, power, configurations, and the ability to create virtual blades and/or individual blades from the same HW through just a few mouse clicks. An operational mode may be seamlessly configured, by policy or by users, through just the click of a button in a management user interface (UI).

A power unit that is located within a CRA may convert grid power to (e.g., 48V) DC power and distribute the same across the frame for all HW cell units. Such an approach eliminates the need for discrete AC-to-DC PSU units and all of the associated cabling. Additionally, the system may be backed up at a common location for an entire rack frame.

As noted previously, on the cooling front a rear CRA door may integrate fan modules to move the air from the front to the back. Such a centralized approach reduces the power consumption as well as provides a more efficient way to cool the entire rack. A CRA may be liquid cooling ready, with quick connections, and may employ an intra-rack LDU or a data center's facility. Additionally, modular cooling elements may be added based on a data center's needs for deployment by specific customers. Thus, the cooling solution may adapt to a particular customer's needs while preserving the key element of centralized and efficient cooling.

As described and illustrated in the above narrative, a CRA comprises a number of important aspects, as discussed herein. To begin, and in contrast to existing solutions, a CRA's frame posts may be prewired(less) with copper or silicon photonics to support an I/O nexus thus eliminating the need for replicated resources (such as intra-frame cables for data or for power) and the need for a backplane across an entire 42RU rack. As a result, all of the different (router, network, compute, etc.) elements have well-defined entry and exit points for opto-electrical I/O for I/O interconnection.

Further, the use of centralized functions (such as power, cooling, etc.) allows for a cost amortization to be realized over an entire rack. Additionally, the elimination of AC PSUs (and the validation effort that accompanies the same) yields a simpler compliance model. It also becomes easier to complete upgrade operations as new technology integration becomes available. For example, it is far easier to change a rack's post compared to changing a chassis backplane.

Moreover, a CRA offers an architecture that provides simplified and dense logic (i.e., more sockets for each U) HW cell units. No longer must a distinction be made between a blade versus a 1U/2RU rack. Additionally, it is possible to dynamically virtualize compute elements to behave like a blade server or individual rack units. Further, a CRA offers a thermally neutral design with future proofing through a support for liquid cooling.

A CRA also offers an easier deployment with no ‘heavy lifting’ of a chassis or cable reconfiguration. Such an approach results in structurally well-balanced weight distribution. An RMU can also support asset management, discovery, boot image storage and management, an integrated terminal server, and KVM switching; provides for a unified management pane (in contrast to the existing separate management platforms managing common tasks such as assets, boot image and storage, etc.); allows for sandboxing control for each IT DevOps that is responsible for its domain; and supports rack-level telemetry and control.

The total cost of ownership (TCO) can also be improved since only the few common HW cell units need to be qualified from an OS and application certification perspective. Such a benefit also results in a carbon footprint reduction. Further, the cost of the centralized functions (such as power, cooling, etc.) may be amortized over an entire rack and not just at a 1U or multiple chassis level. For example, consider the impact that not requiring all of the AC cables and AC PSU development at each product level would have on qualification and compliance savings. Further, any potential safety and/or electromagnetic compatibility (EMC) concerns can likely be addressed through workforce training, further development/research, or the like.

Finally, a CRA takes the different unified computing models to the next level of innovation by tackling all of the white-label original design manufacturer (ODM) vendors and preserving margins in the cutthroat rack server market.

As described and illustrated in the above narrative, a CRA provides a first mover's advantage in this field. While the OCP, even today, still thinks in terms of U (and hence the elements in an OCP approach are still a chassis or 1U or 2U enclosures), the techniques presented herein support a converged rack with built-in ‘smarts.’

One of the most significant innovations within the presented techniques, with respect to an OCP approach, concerns the embedding of (opto-electrical) intelligent data I/O in a rack frame thus eliminating the backplane itself from the design. All of the I/O connectivity happens between the elements through such embedded I/O hubs. It is the standardization of I/O (which usually comprises Ethernet (such as 10-100 Gigabit KR or multi-terabit optical technology optical technology) as the backbone of the converged frame architecture.

Another significant innovation within the presented techniques concerns an RMU. Such a facility helps to manage and control all of the different entities (such as compute, network, and storage elements; services; etc.) in a unified manner and at a rack level. Such an approach breaks from the traditional OCP focus of thinking about just a compute perspective.

In summary, techniques are presented herein that support a CRA. Such an architecture encompasses a universal rack I/O paradigm comprising embedded opto-electrical data I/O posts, centralized power and cooling, and a common RMU which allows a rack itself to be thought of as a compute unit where different compute, network, and storage components may be composed in a virtual domain on the fly and then decomposed when they are not required. In essence, a CRA manages and treats a rack frame as a complete converged system of compute, network, and storage elements and not as disparate entities. Moreover, the various cooling options described herein are compatible with American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) guidelines.