Hewlett Packard Enterprise

Brochure

HPE CRAY EX LIQUID-COOLED CABINET FOR LARGE-SCALE SYSTEMS

For the last 20+ years, most supercomputing solutions have shared a similar design—a large scale-out cluster of servers based on a leading processor vendor and connected by an industry-standard interconnect. This approach has worked well, providing the highest performance possible for challenges in scientific and engineering fields such as weather and climate, computational fluid dynamics, mechanical design, security, defense applications, and more.

But the next decade is bringing a new set of workloads and a dizzying array of potential processing solutions that challenge the high-performance computing (HPC) status quo. Relentless data growth coupled with a global business imperative for digital transformation is driving more and bigger workloads. Traditional modeling and simulation workloads are fusing with AI, analytics, and the Internet of Things (IoT) to create massive business-critical workflows.

Simultaneously, multiple x86, Arm®, GPU, and field programmable gate arrays (FPGA) vendors are expected to deliver compelling offerings. As a result, it's becoming increasingly difficult to predict, which future core processing architectures will offer the best value. It means HPC users will need systems that can:

1. Support a variety of compute platforms
2. Upgrade to new architectures as they become available

This confluence of factors has forced a complete rethinking of supercomputing architecture. It also signals a major technology inflection point and a new era for HPC. More than a simple speed milestone, the exascale era is a new set of capabilities for a new set of workloads that is transforming every field of inquiry.

The HPE Cray supercomputing architecture is the HPE answer to these new demands. It is an entirely new design. Compute, interconnect, software, and storage have been rethought and rearchitected to meet current and future system requirements across HPC, artificial intelligence (AI), and converged workloads. Built to be data centric, it runs fast, diverse workloads all at the same time. Hardware and software innovations limit system bottlenecks in processing, data movement, and I/O. It helps eliminates the distinction between clusters and supercomputers and provides a rich software and system interconnect in different form factors. It allows for multiple processor and accelerator architectures and a choice of system interconnect technologies, including our new HPE Slingshot interconnect.

HPE Cray EX Liquid-Cooled Cabinet

For customers requiring the greatest performance, density, and efficiency for large-scale systems, the HPE Cray supercomputer is available in liquid-cooled cabinetry, which supports direct liquid cooling of all components in a compact bladed configuration.

This HPE Cray EX cabinet architecture contains many innovative features that support the high wattage CPUs and GPUs (in excess of 500W), dramatically reducing interconnect cabling requirements and reducing operational expense significantly. The liquid-cooled infrastructure also results in a much more compact system architecture and helps minimize the use of more expensive optical interconnect cables over less expensive electrical ones.

Additionally, the HPE Cray EX infrastructure has been carefully designed to support multiple processor architectures and accelerator options while remaining forward compatible with next-generation CPU, GPU, and interconnect technologies for at least the next decade.

Key to the HPE Cray EX cabinet's flexibility is the bladed architecture for both compute and networking. It enables mixing and matching of various CPU and GPU technologies as well as providing a straightforward upgrade path to next-generation processors and interconnect capabilities. Critical for many users, the physical, software, and network compatibility of the various CPU and GPU blades allows late-binding decisions on compute platform choice as all are designed to plug into the same infrastructure.

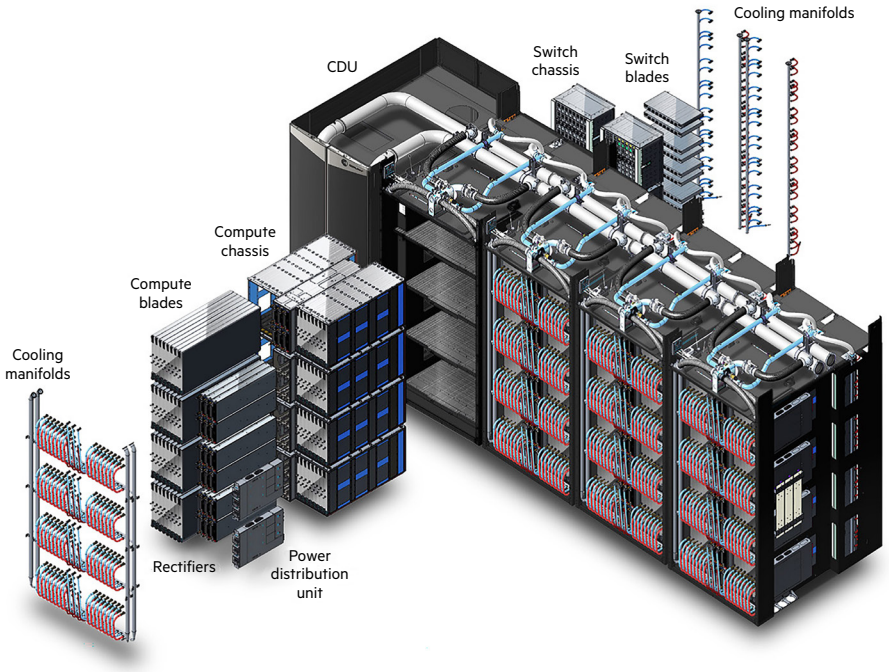

FIGURE 1. HPE Cray EX cabinet exploded view

HPE Cray EX cabinet architecture

The basic building blocks for the compute and switch architecture of the liquid-cooled cabinet are:

- Compute chassis: The compute chassis is a mechanical assembly housing up to eight compute blades. Each HPE Cray EX cabinet holds eight compute chassis, enabling up to 64 compute blades and up to 512 processors per cabinet. The compute blades are aligned vertically and inserted into the compute chassis from the front of the cabinet.
- Compute blade: The compute blade is the module in the compute chassis comprising the HPE Cray EX computing elements—CPUs, fabric connections, printed circuit boards, and cooling and power components. The initial compute blade contains four dual-socket AMD EPYC 7002 nodes.
- Switch chassis: The switch chassis is a mechanical assembly, which houses up to eight HPE Slingshot interconnect switch blades. Each HPE Cray EX cabinet holds eight HPE Slingshot switch chassis, enabling up to 64 switches per cabinet. Switch blades are aligned horizontally and are inserted into the switch chassis from the rear of the cabinet.
- Switch blade: The switch blade contains the HPE Slingshot fabric switch silicon, printed circuit board with connections for compute blades, and all components required for cooling and power. HPE Cray EX supports up to eight switch blades per switch chassis enabling up to 16 fabric connections per compute blade (two fabric connections per physical connector).

Each cabinet contains eight compute chassis and eight switch chassis designed for direct fabric connections from the switch blades to the compute blades without cabling or a midplane. The switch blade is direct liquid cooled as are the actual switch ports, which dissipate significant heat when active optical cables are used. The compute chassis holds up to eight compute blades and the switch chassis holds up to eight switch blades.

The compute chassis and the switch chassis are bolted together as shown in Figure 2. Direct connections between the switch blades and the compute blades are through connectors on each blade, which pass through the open space of a frame between the compute chassis and switch chassis. As the compute blades align vertically and the switch blades align horizontally, each switch blade can directly connect to each compute blade.

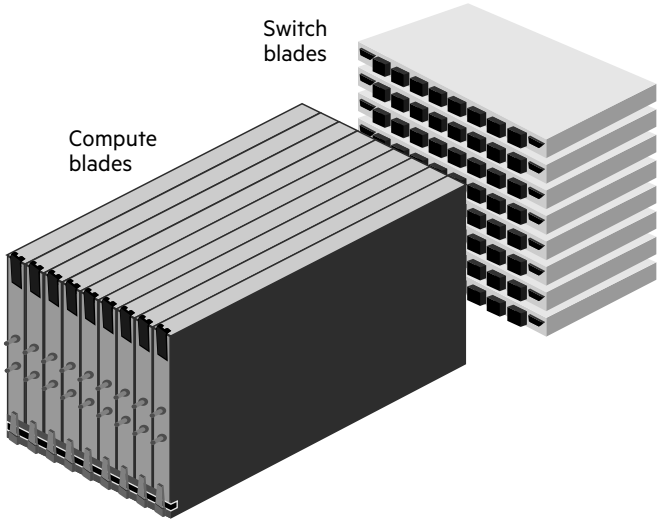

FIGURE 2. Compute blade and switch blade interface

HPE Cray EX425 compute blade detail

The initial liquid-cooled compute blade (HPE Cray EX425) will feature four dual-processor AMD EPYC 7002 servers. Announced future product plans include compute blades based on other CPU architectures and blades with GPUs. These future blades will be form factor compatible and will have similar interconnect capabilities.

Figure 3 shows the HPE Cray EX425 compute blade. All components are direct liquid cooled.

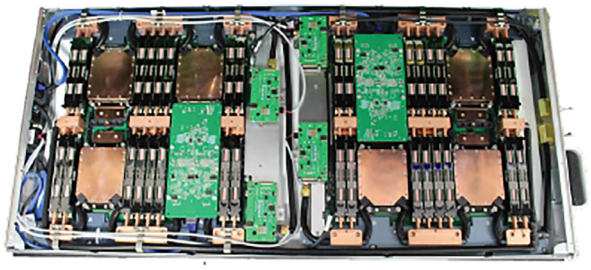

FIGURE 3. HPE Cray EX425 compute blade

The HPE Cray EX cabinet has no fans. (See Power and Cooling section for more details.)

Switch blade detail

The switch blade contains the HPE Slingshot fabric switch silicon, printed circuit board with connections for compute blades, and all components required for cooling and power. HPE Cray EX supports up to eight switch blades per switch chassis enabling up to eight fabric connections per compute blade.

HPE SLINGSHOT IN HPE CRAY EX CABINET

The HPE Slingshot interconnect is a new high-speed, purpose-built supercomputing network with a custom-designed 64-port switch that provides 12.8 Tb/s of bandwidth.

Compute blade network connections

The compute nodes in the HPE Cray EX compute blades interface to the network fabric via a mezzanine card with a PCIe interface to the CPUs. The mezzanine card for AMD EPYC 7002 four-node compute blade contains the HPE Slingshot connections for two nodes. As the compute blades support both single injection (one network connection per node) and dual injection (two network connections per node), the mezzanine cards are, therefore, deployed in groups of two or four per compute blade for single and double injection support. Future compute blades may support more injection ports.

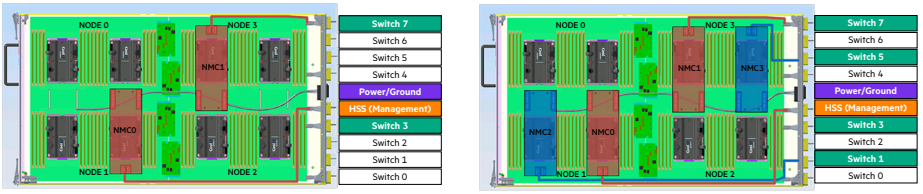

FIGURE 4. Compute node to switch blade connections

HPE Slingshot switch blade network connections

The liquid-cooled HPE Slingshot switch has 16 internal connections, which interface the compute blades (two per compute blade) and 48 external connections for switch-to-switch connections. In the case of the AMD EPYC 7002 four-node compute blade, two switch blades are required for a HPE Cray EX425 single injection port to each node. (Four connections to four nodes per compute blade; 32 connections to 32 nodes per chassis). Double injection requires four switch blades.

FIGURE 5. HPE Cray EX425 compute blade

FIGURE 6. HPE Slingshot switch blade

Building out the network

The liquid-cooled network is typically built out around switch groups with 16 switches with an all-to-all connection inside the switch group, although other configurations are supported. These groups are then linked in a Dragonfly topology that can scale to hundreds of cabinets and hundreds of thousands of nodes and enables endpoint-to-endpoint communication in three hops for all endpoints. Because of the system density in the liquid-cooled infrastructure, electrical cables can be used for all intra-switch group communication with optical connections only required for group-to-group. It reduces the cost of the overall network solution and improves reliability.

Example liquid-cooled network

- 16-switch group
- 2 switches per chassis for single injection to 32 compute nodes (8 compute blades)
- 16 switches per cabinet for single injection to 256 compute nodes (64 compute blades)
- Scales to hundreds of cabinets with maximum 3-hop connection for any endpoint to any endpoint

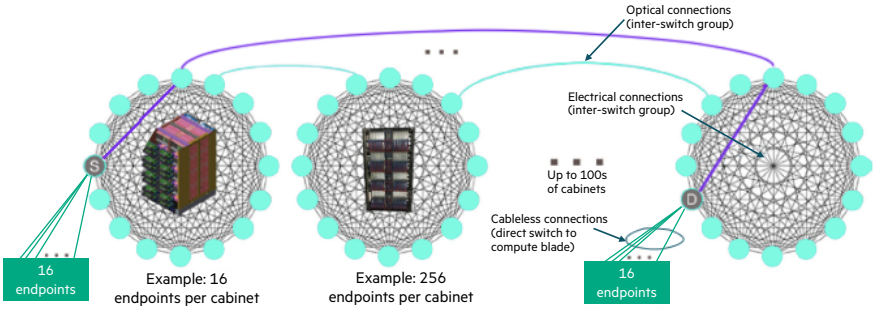

FIGURE 7. Example of Dragonfly topology in HPE Slingshot switches

POWER AND COOLING

As the liquid-cooled cabinet can support up to 300 KW of power, careful attention has been given to HPE Cray EX power and cooling.

Power: Each cabinet has a series of PDUs and rectifiers, which convert incoming 480V or 400V 3-phase AC power into 380V DC power for distribution to the individual compute and switch blades. A series of DC-to-DC converters on the compute and switch blades convert the incoming 380V DC power first into 48V DC and then subsequently into the appropriate DC voltages for the various components. The HPE Cray EX cabinet supports either top or bottom feed for incoming power.

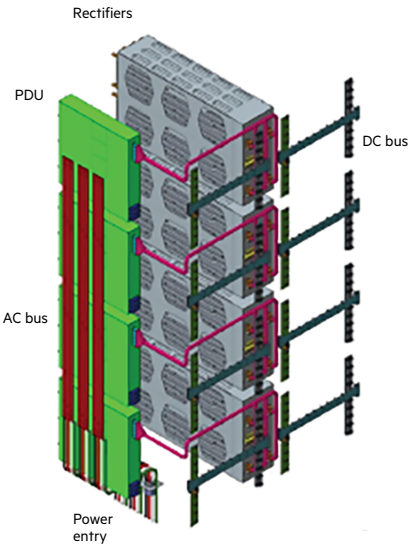

FIGURE 8. Liquid-cooled PDU

Cooling: The HPE Cray EX cabinet and all components are completely cooled by the liquid cooling loops that run through the compute infrastructure. The cooling distribution unit (CDU) cools the liquid and removes the heat from the system via a heat exchanger with data center water.

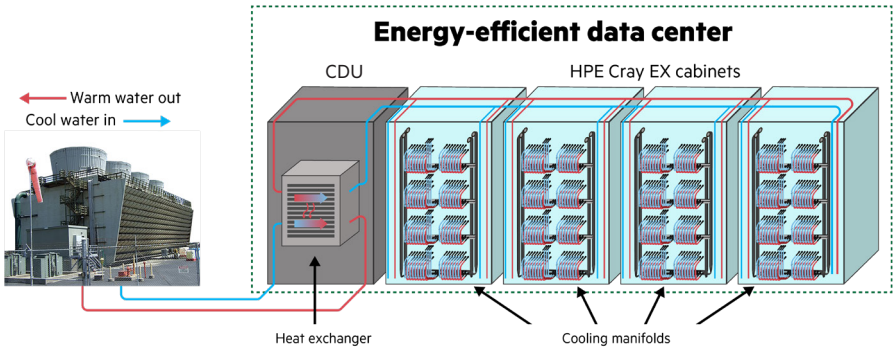

FIGURE 9. HPE Cray EX liquid cooling flow

The overall cooling loop is a closed loop that originates in the CDU. One CDU can support up to four liquid-cooled cabinets. The CDU maintains the cooling liquid at the specified temperature and removes the heat via a heat transfer mechanism to data center water. The CDU requires an inlet water temperature of up to 32°C, which helps eliminate the need for chillers in many environments and further lowers energy usage. The exact data center warm water cooling technology that's been deployed is environmentally dependent and will vary according to climate.

The liquid cooling is routed to individual blades and components in the liquid-cooled cabinets through a series of manifolds that distribute the cooling liquid from the primary piping that runs from the CDU into the individual blades and switches and then returns the heated liquid back to the return piping to the CDU to be re-cooled. The cabinet has cooling manifolds in the front for the compute blades and cooling manifolds at the back for the switch blades. Additionally, liquid-cooling structures cool the rectifiers that also connect to the primary piping.

Connections to and from the compute and switch blades are quick connect and dripless and allow a blade to be removed for servicing without shutting down the entire system.

Cold plates remove heat from the CPUs directly. The NIC mezzanine cards, when present above the CPUs, are also cooled by the same CPU cold plates.

Splitter blocks take fluid from the cooling loop and provide some flow to capillary tubes, which direct liquid through the DIMM field. Figure 10 shows the compute blade with DIMM cold plate structures along with the splitter blocks. Switch blade cooling is similar.

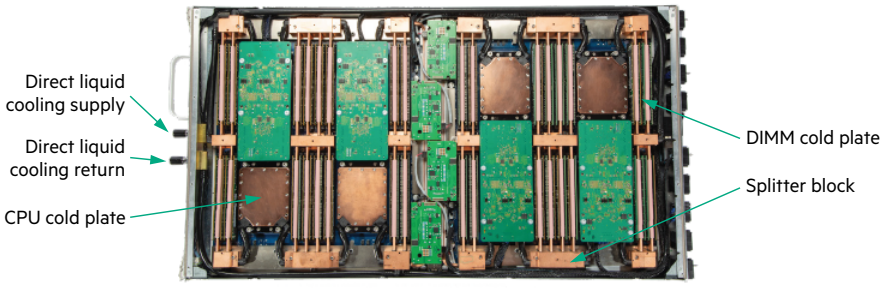

FIGURE 10. HPE Cray EX liquid cooling flow

Given the efficiencies of liquid cooling as compared to air, the budget for power and cooling for a liquid-cooled supercomputer can be significantly less than a similar sized air-cooled installation. Additionally, as the liquid-cooled cabinet can accept data center water up to 32˚C, it can offer more flexibility in data center water cooling technologies (for example, removing of chiller).

CONCLUSION

The HPE Cray EX liquid-cooled architecture delivers the HPE Cray supercomputing experience in a highly integrated and flexible form factor with high performance, scale, efficiency, and value possible with today and tomorrow's processor and GPU technologies.

- **Performance:** Direct liquid cooling supports the high-powered processing elements connected with HPE Slingshot high-speed network for HPC and AI workloads. Coupled with the HPE Cray programming and runtime environment, the liquid-cooled cabinet delivers the high performance at both the node and system level.
- **Scale:** Scale to hundreds of cabinets and hundreds of thousands of nodes. Cableless networking inside the chassis leads to minimal external cables and optical cables compared to other manufacturers.
- **TCO:** Save operating expenses related to electricity and water usage over the lifetime of the product.
- **Flexibility:** Flexible, yet highly integrated liquid-cooled infrastructure allows for a wide choice of compute platforms, upgradable networking solutions, future compatibility, and the ability to make late-binding processor decisions.
- **Reliability:** Minimal use of cables, no mechanical moving parts (fans), and highly reliable power supplies and cooling solutions, which prevent the possibility of overheating, contribute to the overall reliability of the platform. Combined with HPE Cray System Management, the HPE Cray EX cabinet delivers improved system uptime for large-scale systems as compared to similarly sized air-cooled solutions.

LEARN MORE AT

hpe.com/us/en/compute/hpc/supercomputing/cray-exascale-supercomputer.html

a50002389ENW, August 2020