The Cray® XC™ Supercomputer Series: Energy-Efficient Computing

Greg Pautsch, Duncan Roweth, and Scott Schroeder

Cray Inc.
Table of Contents

Introduction ............................................................................................................................................................................ 3
Energy Efficiency of a Supercomputer .................................................................................................................................... 4
Cray XC30 System Packaging: Balancing Density and Energy Use .................................................................................. 7
Cray XC30 Cooling: Innovation Driving Energy Efficiency .................................................................................................. 9
  Cooling of network, storage, and ancillary equipment ................................................................................................................. 12
  Room-neutral operation .......................................................................................................................................................... 12
Air-cooled Cray XC30 Systems ........................................................................................................................................... 14
Comparison of Transverse Cooling and Direct Liquid Cooling ............................................................................................... 15
Cray XC30 Power Train: Delivering Power to the Processors ................................................................................................. 15
Power Monitoring and Control: Optimize and Account for Energy Use ................................................................................... 17
Determining Cost of Operation ................................................................................................................................................ 19
Energy Efficiency Initiatives at Cray ....................................................................................................................................... 20
Conclusion ............................................................................................................................................................................. 21
References .............................................................................................................................................................................. 22
Acknowledgements ............................................................................................................................................................... 23
Introduction

As global energy costs climb, Cray has taken its long-standing expertise in optimizing power and cooling and focused it on developing overall system energy efficiency. The resulting Cray XC supercomputer series integrates into modern datacenters and achieves high levels of efficiency while minimizing system and infrastructure costs.

Effective use of energy is fundamental to the design of the processing nodes, the power train, and the cooling system. Fine-grain power consumption monitoring provides system operators and funding agencies with detailed information on energy usage. In this white paper we will demonstrate the energy-efficiency advancements of the Cray XC supercomputer. Specifically, we will discuss:

- Integration into state-of-the-art datacenters
- Transverse cooling system innovations
- Efficient power distribution
- Control and monitoring of power consumption
- Cray Inc. energy-efficiency initiatives

The Cray XC30 system is a distributed memory supercomputer developed as part of Cray’s participation in the Defense Advanced Research Project Agency’s (DARPA) High Productivity Computing System (HPCS) program. Capable of sustained multi-petaflops performance, the XC30 system’s hybrid architecture combines multiple processor technologies, a high performance network, distributed operating system, and a productive programming environment. Cray XC30 systems are designed to maximize energy efficiency while keeping system and infrastructure costs to a minimum.
Energy Efficiency of a Supercomputer

The energy efficiency of a supercomputer system can be characterized by three metrics: time-to-solution on a given workload; power consumed executing the workload, and the power usage effectiveness (PUE) of the datacenter housing the system. These metrics combine to form the energy to solution:

$$\text{Energy to solution} = \text{time to solution} \times \text{power consumption} \times \text{PUE}$$

Time-to-solution and power consumption are closely coupled. For example, the choice of processor can reduce time-to-solution at the expense of power consumption. Power consumption can also vary within individual jobs and over time as the set of jobs that make up a given workload start and stop.\(^1\)

Key elements of the Cray XC30 are designed to minimize time-to-solution for large problems — the choice of processors, the performance and scalability of the network, and the programming environment. These elements enhance energy efficiency by reducing the time taken to execute a given workload.

Other elements of the design enhance system utilization (e.g., boosting availability or reducing interference between jobs) thereby reducing the time-to-solution and hence the energy required to execute the workload as a whole. In addition, the physical implementation of the Cray XC30 system — packaging, power delivery, cooling system — is designed to provide power and extract the resulting heat efficiently. Overall, the Cray XC30 design minimizes total cost of ownership (TCO) and strikes a balance between initial capital and facilities costs and ongoing electricity costs.

The PUE metric characterizes the operating efficiency of the datacenter housing a Cray system and its peripherals. PUE is defined as the ratio of the total datacenter facility power to the power of the IT equipment on the floor. This ratio depends on a variety of factors including:

- Power delivery and cooling requirements of the system itself
- Quality of the datacenter cooling system
- External ambient conditions and their seasonal variations
- Variation in the application workload
- Size and power density of the datacenter

Datacenter infrastructure and supercomputer systems have developed largely independent of one another, but rising electricity costs have strengthened the case for closer integration. Today, high performance computing (HPC) systems are being designed to reduce operating costs and datacenters are being designed to provide power and cooling more efficiently.

The Uptime Institute recently reported an average datacenter PUE of between 1.8 and 1.89 and an increasing awareness in the industry of the issues related to its accurate measurement \([1]\). In a conventional datacenter with a PUE in this range, 50 percent of the power typically goes to the IT systems, 15 percent is lost in power conversion and distribution, and 35 percent is used for cooling. In contrast, the state-of-the-art for modern datacenters is a PUE of 1.1. For example, Google reports a global average PUE of 1.13 \([2]\), a figure that includes a number of older facilities. Cray XC30 systems can achieve these high levels of efficiency without requiring substantial investment in infrastructure.

Energy proportionality — the principle that energy consumption should increase in proportion to system load — is widely adopted in industry-standard servers where utilization levels are often low. Cray

\(^1\) Power consumption generally varies over the course of a job and between jobs. In this paper we assume the average power consumption over the runtime of the job rather than integrating over time. Power efficiency mechanisms serve to reduce this average figure.
systems typically have 90-plus percent utilization, but significant savings can still be made through energy proportionality. The Intel® Xeon® processors used in XC30 systems show a high degree of proportionality as load increases from idle to maximum — a 70 percent dynamic range in the power consumed is typical. In addition, some memory or network-bound applications can run efficiently with the processors in reduced power states. Where resources are unused (e.g., a set of nodes idle between interactive jobs), an XC30 system can automatically select a low-power state. Today’s memory, networks, and storage systems tend to have a lower dynamic power range; savings of up to 25 percent are typical as load drops to idle. The power efficiency of these components is an active area of research for future systems.

Modern processors require substantial amounts of power; 350 watts per node is typical for a pair of high performance processors, their memory, and support chips. A high density system requires a power train that can deliver 75 to 100 kilowatts per cabinet. The power train must also convert down from a datacenter’s high voltage AC supply to the low DC voltages used by the processor. Power conversion losses grow with number of levels of conversion and the efficiency of each level. Transmission losses grow with distance and the square of the current. Cray XC30 systems minimize these losses, distributing power at high voltage and converting to processor core voltages as close to the processors as possible. Under normal load, the power train achieves 80 percent or higher efficiency end to end.

Fan and pump motors, computer room air conditioning (CRAC) units, and chillers are the main consumers of power within a datacenter cooling system. CRAC units cool air using cold water and are typically only 60-80 percent efficient. Chillers provide cold water to the CRAC units or directly to the IT systems. Since chillers can contribute as much as 0.4 to the PUE, their elimination is often the single biggest step toward a more efficient total system. An efficient liquid-cooled supercomputer that minimizes the need for CRAC units can reduce the PUE to 1.5 or 1.6. Where that system can run with unchilled water the PUE can be as low as 1.1.

Figure 1 illustrates the components of a conventional cooling loop — cooling tower, chillers, CRAC units, coolant distribution units (CDU), and HPC system. Figure 2 illustrates an equivalent cooling loop that operates without chillers.

---

2 The Cray XC30 is the first model in the XC series. Each node uses a pair of Intel Xeon processors.
Figure 2: Datacenter design using free cooling with warm water.

A key element of this approach is known as “free cooling.” The computer system is cooled with warm water which is itself cooled using low temperature external air. Electricity is consumed by pumps and fans so while the term “free” cooling is in common use; a more accurate description of the technique is “cooling without mechanical refrigeration.”

This approach works well in cold climates; new datacenters may be constructed without chillers, saving substantially on facilities and running costs. Older facilities can install a separate free cooling loop with a relatively quick payback, especially when designed to take advantage of incentives from power utilities and local governments. In warmer climates chillers will be required, but mechanical cooling can be partially or completely disabled easily and dynamically when air temperatures are low, saving on running costs.

The Edinburgh University Advanced Computing Facility reported free cooling of the HECToR Cray XE6 system active 76 percent of the year (see [21]), with some level of free cooling available on all 52 weeks of the year (data for 2011). Edinburgh is at latitude 56 degrees north with average daytime air temperature varying between 3°C and 15°C. Savings made from free cooling of the HECToR system are estimated to be approximately $208,000 (£130,000) per year.

Cray XC30 systems use higher inlet water temperatures than the previous generation Cray XE™ and Cray XK™ supercomputers and generate a higher ∆T (the difference between input and output water temperatures) enabling free cooling systems to operate at higher efficiency, for more of the year, and in warmer climates.

The National Energy Research Supercomputer Center (NERSC) located in Berkeley, California reported that Bay Area conditions are favorable for cooling without mechanical refrigeration all year round (see [15]). In worst case conditions, which occur for just a few hours per year, air can be provided at 74°F and water at 75°F. The new datacenter for the Computational Research and Theory Facility (CRT) is being constructed without chillers. The Cray XC30 system is designed to operate within this environment, providing a state-of-the-art supercomputer system to NERSC users at the highest degree of energy efficiency. NERSC’s existing computer center was upgraded with a separate warm water cooling system with no mechanical cooling to cool the XC30 system while the new center is being constructed.

In June 2010, the Commonwealth Scientific and Industrial Research Organisation (CSIRO) received funding from the Australian Government’s Education Investment Fund to develop a direct-heat geothermal demonstration site. The Pawsey geothermal supercomputer cooling project will use
geothermal energy from a hot sedimentary aquifer to provide cooling to the Pawsey High-Performance Computing Centre, and cooling and heating to the co-located CSIRO facility. CSIRO estimates that the aquifer's water temperature is sufficient for its planned purposes at a target depth of 3 kilometers [see [16]]. Cray XC30 supercomputers are being installed at the Pawsey site for use by the Australian research community and as part of the Australian Square Kilometre Array (SKA) pathfinder project [16].

Energy efficiency can be improved further. PUE values below 1.0 are possible if the heat generated by the computer system can be recovered. The Royal Institute of Technology (KTH) in Sweden reports that as part of the PDC Center for High-Performance Computing’s ongoing work to reduce its environmental footprint they are engaging in a heat re-use project. Excess heat from their Cray XE6 supercomputer is being used to heat a nearby building on the KTH campus. [3] Calculations show that between 60-70 percent of the energy from the XE6 system can be re-used during the cool seasons saving on heating and cooling costs. Altogether PDC sends around 1,300 megawatt hours of energy per year to the KTH building. This provides cooling cost savings of approximately €80,000 per year. The building was already equipped with heat re-use from recycled air. The heat re-use mechanisms have an efficiency of around 50 percent; KTH makes savings of €40,000 per year on heating costs as a result. Total savings per year are around €120,000 for a normal year.

Reuse of the hot water generated by computer systems can be an attractive option in cases where a new datacenter and offices are co-located. For instance, the new facility at NERSC will use waste heat from computer systems in the basement to heat the office building above.

Retrofitting heat reuse to existing buildings is generally expensive and may not be cost effective. The Cray XC30 system is designed to reduce the cost of operation for sites where full or partial free cooling is possible, but heat reuse mechanisms are not yet in place. Cray can undertake a detailed study of the costs of cooling a Cray system at your site. Please contact your Cray representative for details.

The Cray XC30 system has been developed in anticipation of an environment in which running costs as well as processor time are fully accounted for — we expect users to be given an energy bill at the end of every job. In addition to designing energy-efficient systems, Cray is developing tools to monitor, report, and optimize power use, enabling sites to operate their systems so as to minimize energy consumption.

**Cray XC30 System Packaging: Balancing Density and Energy Use**

The performance of Cray systems has increased substantially over recent years [12]. A significant element of this growth in performance has come through use of increasing numbers of nodes. We expect this trend to continue with increased emphasis on sustained performance [13]. Meeting these demands in a cost- and energy-efficient fashion requires dense packaging of nodes. The Cray XC30 system is designed to support up to 92,544 individual compute nodes — three times the size of the National Center for Supercomputing Applications’ Blue Waters system, the largest Cray system shipped to date [13].

Cray XC30 systems are constructed from a mix of compute and I/O blades. Four dual-socket nodes are packaged on each compute blade along with the Aries network router that connects the blades [10]. Sixteen blades are packaged in a chassis, (see Figure 3) with the chassis backplane providing connectivity.

The chassis are connected by electrical cables within cabinets and groups. The maximum extent of electrical connectivity is limited by the signal speed (14 gigabits per second). The high density packaging of XC30 systems maximizes the number of nodes that can be connected via these electrical links. In the liquid-cooled system, groups of six chassis are housed in a pair of 36 inch (900 millimeter) cabinets.

---

3 At the time of writing the saving of €120,000 is equivalent to $160,000 per annum.
cabinets with each two-cabinet group providing up to 384 nodes (see Figure 3). These two-cabinet groups are the building blocks for larger systems. They are connected using high bandwidth active optical cables (see Figure 4). With 192 dual socket nodes per cabinet, the high density design maximizes use of the low cost, low power electrical links.

Figure 3: Cray XC30 chassis. Each chassis houses 64 dual-socket nodes and the first stage of network. Each cabinet contains three chassis, together with the power distribution, water cooling, and control systems. Cabinet pairs are connected by electrical cables to form a 384-node group. Optical links connect these groups.

Figure 4: Cray XC30 system two cabinet group. Local electrical cables (striped) and global optical cables (green).
Cray XC30 blades support a pair of processor daughter cards (PDC). Each such card provides two dual-socket Intel Xeon processor nodes. Other variants of the PDC support current and future processors, coprocessors, and accelerators. The modular structure and the adaptive nature of the XC30 system defray infrastructure costs and associated facilities costs over multiple generations of processors.

An alternate blade design comprising a pair of single-socket nodes each with two PCI-Express Gen3 interfaces provides system services and I/O. Up to 50 percent of the blades in a system can be I/O blades — although in most systems the I/O blades make up between five and 10 percent of the total.

In addition to the high-density compute cabinets, XC30 installations typically include Cray storage [20] and a number of industry-standard servers housed in external 19 inch (600 millimeter) racks.

**Cray XC30 Cooling: Innovation Driving Energy Efficiency**

Cray XC30 systems are cooled using horizontal side-to-side “transverse” airflow with air moving in series through all cabinets in a row. Water coils between each cabinet remove the heat, cooling the air before it enters the next cabinet. Airflow is maintained by horizontally mounted fans in separate blower cabinets.

Figure 5 illustrates the transverse cooling system. Each cabinet houses 48 horizontally mounted blades in three chassis, a power distribution unit (PDU), a cooling water coil, and the hardware supervisory system (HSS) control processor. Blower cabinets are positioned between each pair of cabinets.

**Figure 5: Transverse cooling.** Air blows horizontally along a row of cabinets cooling the blades. Water coils in each compute cabinet extract heat. Inlet and outlet air temperatures can be balanced to make the system room neutral.

The XC30 cooling system is extremely scalable. Each row of cabinets forms a self-contained unit. The cooling system is room neutral\(^4\), eliminating CRAC units other than those required for support of other

\(^4\) Small amounts of heat may be rejected to air if the ambient air temperature is high relative to the input water temperature.
equipment and dehumidification\(^5\). Additionally, transverse cooling eliminates the need for hot and cold aisles and the need to supply air to each cabinet. The result is a drastic reduction in overall air movement, reducing the requirement for air handlers, reducing noise levels, and increasing datacenter efficiency.

The large surface area available between racks allows the use of large water coils, increasing cooling capacity; more heat can be extracted per unit volume of water. This feature leads to less difference between the water and air temperatures (a “closer approach”) and a greater change in water temperature, a higher \(\Delta T\).

Cray XC30 processor modules use a graduated heat sink pitch (see Figure 6); a technique developed on Cray XE systems to ensure that all processors operate within the same thermal envelope\(^6\). On processors closest to the cooling coils, the heat sinks require only a small number of widely spaced fins. The number of heat sink fins increases on processors farther away from the cooling coils where air temperatures are higher, efficiently and effectively normalizing processor socket temperatures, regardless of their location in the system cabinet.

Figure 6: Cray XC30 processor module. Each processor module provides a pair of dual-socket nodes. A graduated heat sink pitch ensures that all processors operate within the same thermal envelope.

Each cabinet has an independent cooling coil connected directly to the facilities water supply. The HSS includes a microcontroller in each cabinet that provides both environmental monitoring\(^7\) and control. The controller adjusts water flow rates by opening and closing a valve so as to maintain a constant outlet air temperature. Environmental data is sent to the system management workstation for logging and subsequent analysis.

\(^5\) An optional preconditioning cabinet is added on the input to each row where dehumidification is required.

\(^6\) In a tightly coupled parallel application processors must operate at the same speed. This requirement can be relaxed when the application uses an asynchronous programming model.

\(^7\) The HSS logs data on statistics including ambient air temperature, relative humidity, dew point, blower status, air velocity, air and water inlet and outlet temperatures, water pressure differential across the cooling coils, and power consumption. Data is collected on the system management workstation.
Water cooling system parameters depend on the processors and memory used, the input water temperature, the datacenter air temperature, and the altitude. For example, a system with 130-watt processors, situated at sea level and with input water at 68°F (20°C) requires a flow rate of 38 gallons per minute (2.4 liters per second) to maintain an output water temperature of 83°F (28.5°C). Liquid-cooled XC30 systems support input water supply temperatures in the range 41°F (5°C) to 75°F (25°C). The cooling system can tolerate changes in inlet water temperature that occur with free cooling as outside air temperatures change [4].

**Figure 7: Cray XC30 water cooling system.** Each cabinet has an independent water cooling system fed by facilities water and controlled by the HSS.

Blower cabinets are located between each pair of compute cabinets to maintain air flow. Each blower cabinet contains six independent and easily accessible fan modules.

**Figure 8: Cray XC30 blower cabinet with optional preconditioning coil.** Each blower cabinet contains six independent fan modules. Fans can be hot swapped while the system is operating.
The Cray XC30 system uses highly reliable blowers, but with hundreds of them in a large system failures will occur. It provides comprehensive and inexpensive protection against fan failure, increasing the speed of the other fans to maintain air velocity along the row of cabinets. The system can operate in this fashion until replacement components are available. Systems with five or more blower cabinets (10 compute cabinets) per row are naturally N+1 redundant. The blower cabinet can be withdrawn and the fan replaced while the system is in operation (see Figure 8). For smaller systems, an additional blower cabinet on the exhaust end of the row makes the blower cabinets N+1 redundant. Each blower cabinet provides airflow for compute nodes drawing 150–180 kilowatts. The power required to do this is between 4.1 and 5.2 kilowatts, or 14 watts per node — a contribution of just 0.03 to the PUE\(^8\). Power required for cooling is approximately 40 percent lower than the industry norm.

The Cray XC30 cooling system does not require datacenter air to be kept cold. In line with current design principles [2], the XC30 operates at warm ambient air temperatures. Where the air temperature exceeds 77°F (25°C), a preconditioner is inserted in the first blower cabinet in each row, cooling the air as it enters the system (see Figure 8). A preconditioner is also used in conditions of high relative humidity. The Cray XC30 system requires that the dew point\(^9\) be at least 4°F (2°C) below the inlet water temperature. A preconditioner is required when this condition cannot be met.

XC30 systems operate on standard facilities water\(^10\) without the need for glycol\(^11\), other additives, or refrigerants. As such, the system conforms to all relevant environmental standards. By eliminating or reducing the need for chillers and greatly reducing the need for CRAC units, power consumption is reduced, the range of acceptable water inlet temperatures is increased, and the ΔT is increased.

**Cooling of network, storage, and ancillary equipment**

The Cray XC30 network is comprised of a single Aries ASIC per blade together with printed circuit board and backplane links, electrical cables, and active optical cables (AOC). The network ASICs are cooled by the transverse airflow. Each Cray XC30 cabinet contains 120 AOCs; heat output from each of these devices is approximately 3 watts. A stable operating temperature reduces error rates in the electro-optical convertors and extends their lifetime. The XC30 supercomputer uses a system of heat pipes and direct liquid cooling in the back of each cabinet to cool the optics.

In addition to the high density compute cabinets, XC30 systems typically include a number of standard 19 inch (600 millimeter) cabinets containing Cray storage systems, industry-standard servers, the boot RAID, and other peripherals. The heat output of each of these cabinets ranges from 5 to 17 kilowatts depending on their configuration. The number of such cabinets and hence their total heat output is generally small in comparison with that of the compute cabinets. Storage and server cabinets can be air cooled or fitted with water-cooled doors [7].

**Room-neutral operation**

The Cray XC30 cooling system can be adjusted to be room neutral, or in some circumstances, room negative, allowing the system to cool small amounts of peripheral equipment. However, the cooling system is not designed to dehumidify datacenter air. This task should be handled by a computer room air handler.

The XC30 system has two modes of operation:

---

\(^8\) The power required to move air for cooling forms part of the mechanical load in a true measure of PUE. It is often considered to be part of the IT load, lowering the PUE at the expense of the IT load.

\(^9\) Dew point is the temperature at which air becomes saturated. The closer the air temperature is to the dew point, the higher the relative humidity.

\(^10\) Datacenter cooling water systems require corrosion inhibitors and other treatments that prevent growth of bacteria. The Cray XC30 does not require more.

\(^11\) Please contact Cray for information on water flow rates where glycol is used.
• **Automatic room neutral.** For most installations the system monitors the inlet air temperature on each row and controls the water flow rates to match that outlet temperature at the row exit.

• **Manual control.** This mode allows users to set the exhaust air temperature to a value in the 65-90°F (18-32°C) range. Manual control allows the XC30 system to provide cooling for small amounts of equipment elsewhere in the room. In cases where a system is sited in a humid area with no air-cooled equipment in the room, the exit air temperature may be set so as to reject just enough heat to the computer room to allow an air handler to provide dehumidification.

In order to maintain room-neutral operation, the inlet water temperature must be at least 7°F (4°C) below the ambient air temperature. Where the inlet water temperature is not sufficiently below the ambient air temperature, the temperature of the exhaust air from the last cabinet in each row will exceed that of the ambient air. For example, a system with eight cabinets per row, input water at 66°F (19°C), and an ambient air temperature of 71°F (21°C), the output air temperature will be 73°F (23°C) and approximately 8 kilowatts is output to air per row of cabinets. This figure falls as the ambient air temperature is increased to 73°F. The heat output of a system will also vary as the workload changes. Thermal inertia in the water cooling system provides stable control over time, but results in fluctuations in output air temperature as the work load changes.

![Figure 9: The Cray XC30 “Piz Daint” system at the Swiss National Supercomputing Centre.](image)

The Swiss National Supercomputing Centre (CSCS) took delivery of one of the first large Cray XC30 systems in December 2012. A major upgrade of the system took place in October 2013 to create Europe’s most powerful supercomputer and the first supercomputer with sustained petaflop performance in Switzerland. Dubbed “Piz Daint,” the system is a hybrid design combining the scalar performance of Intel Xeon E5 processors with the floating point performance and energy efficiency of NVIDIA® Tesla® K20X GPU accelerators. Early results show sustained performance of 4.2 petaflops on DCA++, a quantum Monte Carlo code used to simulate high-temperature superconductors. Regional climate simulations using a new implementation of the COSMO model [22] show a seven-fold reduction in energy to solution. Operating power for Piz Daint ranges from 1200 kilowatts for an ensemble of COSMO jobs to 1700 kilowatts for a single large DCA++ job. Cold water (43°F, 6°C) extracted from Lake Lugano provides cooling for the datacenter [17].
Transverse cooling alleviates problems arising from cables obstructing air flow. It also frees up space at the rear of the cabinets, simplifying cable routing. Transverse cooling improves reliability, availability, and serviceability (RAS) of XC30 systems. Power distribution and cooling systems continue to operate in the presence of failure. Blades, processor modules, voltage regulators, and fan modules can all be swapped while the system is operating and without risk to the cooling system. Mean time to repair is reduced as the blades are directly accessible from the front of the system. Reliability is improved by reducing the number of components required to cool the system.

Air-cooled Cray XC30 Systems

For datacenters that do not support liquid cooling, the Cray XC30 system is available in an air-cooled variant — the Cray XC30-AC system. It can also be used for small test and development systems where the heat output is low.

The XC30-AC system supports up to 512 nodes in configurations of one to eight cabinets. A single XC30 chassis housing up to 16 compute blades is mounted vertically in each cabinet. An axial turbo blower provides bottom-to-top air flow cooling the blades. These single chassis computing cabinets can be effectively cooled with air only, removing requirements for liquid infrastructure, plumbing, and supplies.

The streamlined XC30-AC cabinets have a smaller physical footprint and require less power (up to 37 kilowatts per cabinet). There are also options for 208V operation as well as the 480V level common to the liquid-cooled XC30 systems. Cray XC30-AC systems are designed for use with computer room ambient air, under floor air, or a combination of the two. The air-cooled XC30 systems leverage the same HPC-optimized OS, software tools, interconnect, compute and I/O blades as the liquid-cooled configurations. Applications with demanding performance requirements and/or high-density supercomputing will likely target the XC30 liquid-cooled configurations, while smaller, more economical technical enterprise applications may be serviced with air cooled systems.
Comparison of Transverse Cooling and Direct Liquid Cooling

Liquid cooling is necessary for high power devices that cannot be air cooled\textsuperscript{12}. Direct liquid cooling may also assist in increasing density, and hence the number of nodes that can be connected using low cost electrical links. In such systems, coolant is taken directly to the device. Increasing the input water temperature enables a system to operate without chillers under a wider range of circumstances and can also make heat recovery simpler\textsuperscript{13}.

However, direct liquid cooling also has disadvantages. It increases the capital cost of the system and the facility. Direct water cooling also increases the complexity of the cooling system, increasing the risk of failure and making maintenance more difficult. Furthermore, direct liquid cooling is only partially effective; the processors typically generate 60-70 percent of the heat output with the memory, network devices, voltage regulators, and power supplies generating the rest. Direct liquid cooling can be extended to cover some of these components at additional expense and with diminishing return, but the heat rejected to air remains significant. Relying on conventional air cooling for 30-40 percent of the heat load — several megawatts in a large system — is not an energy-efficient option\textsuperscript{14}. A large system will require two cooling loops, one using hot water (30-40°C) for the processors and, optionally, the memory; and a second loop using warm water (18-25°C) to cool the other components.

In the majority of cases, increased costs for the system and the facility offset much of the benefit of direct liquid cooling. The bulk of the savings are made by operating without chillers or by reducing the number of hours per year requiring chillers. The XC30 system is designed to meet these objectives.

Cray XC30 Power Train: Delivering Power to the Processors

HPC systems draw power provided by the facility (typically high-voltage AC) and need to deliver it to the processors and other devices at low voltage (core voltages of 1-volt DC are typical) in an efficient manner. Power conversion losses increase with the number of conversion levels and the efficiency of each level. Transmission losses grow with the distance and the square of the current, making it important to convert to core voltages as close to the processor as possible. The Cray XC30 power system addresses these issues efficiently (over 80 percent of the wall socket power is delivered to the devices) while maximizing system availability and minimizing time to repair.

Cray XC30 systems use 480/400-volt AC — the standard power distribution system for large datacenters\textsuperscript{15,16}. The power distribution unit (PDU) in the base of each cabinet converts voltage from the supply to 52-volt DC (see Figure 11). From here power is fed to the three chassis hosts and the 48 blades in each cabinet. In addition, there is a 5-volt DC always on supply to the cabinet controller and associated HSS components. Voltage regulator modules (VRM) on the PDCs and blades convert from 52-volt DC to 12-volt DC for supply to the processors and the Aries network devices. A final stage of VRMs on the processor modules converts from 12-volt DC to the core voltages.

\textsuperscript{12} The 1200 watt multi-chip modules used in Cray X1 systems were an extreme example, using direct phase-change cooling.
\textsuperscript{13} Increased water temperatures make heat recovery simpler, but output temperatures are too low for efficient generation of electricity.
\textsuperscript{14} If chillers and CRAC units are required for 30 percent of the heat load the PUE will increase by around 0.2.
\textsuperscript{15} Step-up transformers are required in Japan where 200-volt circuits are the norm.
\textsuperscript{16} See the Cray XC site planning guides\textsuperscript{[14]} for details of the circuits required.
The Cray XC30 system PDUs use up to 36, 3-kilowatt 80 PLUS Platinum-certified rectifiers per cabinet, converting from 480/400-volt AC to 52-volt DC. They achieve an efficiency of 95 percent or better when running between 30 and 90 percent of full load.

In an XC30 system with dual-socket Intel Xeon nodes, the peak power load per cabinet varies from 75 to 90 kilowatts depending on the processor selected and the memory configuration. Future system designs may support processors with higher power requirements, e.g. accelerators/coprocessors with a thermal design point of 200-250 watts. The number of rectifiers used in each PDU is configured to meet the power load of the cabinet and maintain a high level of efficiency. Table 1 summarizes power conversion efficiencies.

<table>
<thead>
<tr>
<th>Conversion</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rectifiers convert from 480/400V AC to 52V DC</td>
<td>95%</td>
</tr>
<tr>
<td>Voltage regulators convert from 52V DC to 12V DC</td>
<td>95%</td>
</tr>
<tr>
<td>Voltage regulators convert from 12V DC to logic levels</td>
<td>89%</td>
</tr>
<tr>
<td>Overall efficiency</td>
<td>81%</td>
</tr>
</tbody>
</table>

**Table 1: Cray XC30 power conversion efficiencies**

Overall efficiency in power delivery and conversion exceeds 80 percent across a wide range of operating loads (see Figure 12). Improvements to the efficiency of the power distribution and conversion system result in a 4 to 5 percent savings over the Cray XE, the previous generation of Cray product. The power delivery efficiency is significantly higher than that of industry-standard rackmount servers. As a result, energy costs are lower than those of other systems using the same processors.
The XC30 power train provides N+1 redundancy in the power distribution components. Rectifiers can be hot swapped while the system is operating normally. N+1 redundancy in the blade VRMs protects against failures that would otherwise impact all four nodes on the blade.

Compute and I/O blades can be warm swapped. Standard operating procedure is to drain the workload from the affected nodes. Once the jobs have finished the blade is swapped out, the repair is completed, and the blade is returned to the system. The nodes are then rebooted, tested, and returned to service. VRMs on the blades and processor modules are socketed to allow easy replacement in the field. Impact on the system is minimal; it operates with a reduced number of nodes for a limited period of time.

Power Monitoring and Control: Optimize and Account for Energy Use

The Cray XC30 system provides comprehensive monitoring — logging where power is consumed and by which users. Power consumption is sampled periodically and reported via the HSS. As each job completes, the monitoring subsystem logs both the energy consumed by the job and the CPU time\(^ {17} \). This infrastructure enables Cray XC30 sites to account the total cost of each job. Energy efficiency measures such as the Green500 report performance on the Linpack benchmark in megaflops per watt [18], an interesting measure, but not one that necessarily relates to sustained performance. The Cray XC30 allows users, administrators, and funding agencies to measure and account for the energy efficiency of their systems on their production workload.

Power consumption is monitored at the blade level with each blade reporting usage by the nodes and the network. Data is aggregated by the cabinet controllers and reported out-of-band to the system management workstation where it is logged for subsequent processing (see Figure 13). Information on the allocation of nodes to jobs is logged at the same time, enabling energy consumption to be added to the data collected for each job.

---

\(^{17}\) System management workstation logs data. How this data is reported varies between work load managers.
Figure 13: Power usage data logged by a four-cabinet Cray XC30 system. System collects data for each node and job. Idle load is approximately 22 kilowatts per cabinet. Load rises to 60-70 kilowatts per cabinet as jobs execute.

XC30 systems provide the ability to monitor and control the power. Power controls are available for both the system and the user. With electricity pricing varying widely between countries and suppliers, Cray designs anticipate operating patterns in which XC30 sites optimize their use of power to match their electricity contracts. Third-party workload management products can build on this infrastructure to schedule power hungry jobs when electricity is plentiful or available at reduced cost.

Performance state (P-state) settings are used to control dynamic voltage frequency scaling (DVFS) in modern processors [for details see description of advanced configuration and power interface in reference [6]]. Users can control the P-state at application launch, locking in to the highest power state (P1) or selecting a lower power state in which each processor alters its power consumption dynamically. For highly processor-bound workloads the best option is generally to use P1, minimizing time to solution. However, for memory or network bandwidth-limited codes, locking the cores into P1 may end up with similar time-to-solution, but at a higher energy cost. For example, researchers from Sandia National Laboratory report performance of their adaptive multi-grid (AMG) solver:

"AMG demonstration on 6,144 nodes of ORNL’s Jaguar shows that managing P-states allows for a 32% decrease in energy used while only increasing time to solution by 7.5%." [11]

The Cray XC30 system allows users to make power consumption measurements themselves, optimizing for both performance and power consumption.

Optimizing major codes for both energy efficiency and performance is a difficult but increasingly important process — especially for applications being designed for the next generation of systems. The Intel Xeon processors used in XC30 systems support the running average power limit (RAPL) model for high frequency in-band power monitoring and power capping [8]. The Cray system provides statistics
on total energy consumption and instantaneous power consumption in-band. This data can be accessed directly via the kernel sysfs interface or through a performance counter library interface.

Support for P-state control is integrated into the Cray Linux Environment (CLE). As jobs complete, the nodes enter the IDLE_UP state and a higher P-state is applied, reducing power consumption until the next job is started. Support for tickless idle (a Linux kernel feature that eliminates periodic timer ticks when processors are idle) reduces power consumption while such nodes are not in use. Nodes that remain idle can transition to the IDLE_DOWN state, in which they are no longer operational and power consumption is minimized. CLE provides application interfaces that control use of these idle states, enabling an external workload manager to tune the number of nodes available to match the operating conditions.

Cray XC30 supports system power capping. Total power consumption of the system can be limited to comply with utility contract maximum figures or recurrent budgets. The power capping mechanism can be used to limit risk of budget overrun in the event of unforeseen increases in electricity costs. Power profiles are provided for each operating period describing how P-states will be limited for each type of node so as to reduce power consumption in the event of reaching power thresholds.

**Determining Cost of Operation**

The contribution of power and cooling to the total cost of ownership depends on the power consumption of the system, its utilization, the efficiency of the cooling system, and the details of the electricity contract. These factors are complex and vary between sites. A number of simplifying assumptions are necessary in order to make an initial estimate:

\[
\text{Annual Energy Use Cost} = (8760 \text{ hours/year}) \times (\text{Average Utility Rate in } \$/\text{kilowatt hour}) \times (\text{Average Equipment Power in kilowatts}) \times \text{PUE}
\]

- Peak power consumed by the XC30 compute cabinets ranges from 75-90 kilowatts depending on their configuration. We assume a figure of 85 kilowatts for this calculation. Cray representatives can provide this information for a specific system.

- Actual power consumption will vary according to the workload. This figure can be measured while a system is in operation or as part of the benchmarking process. We assume a figure of 75 kilowatts for this calculation.

- Actual power consumption will vary according to the utilization. This number is generally assumed to be high, 90-plus percent, with a 70 percent reduction in power for nodes in the IDLE_UP state and a 90 percent reduction for the IDLE_DOWN state. We make a conservative assumption of 95 percent power utilization for this calculation.

- The cost of cooling varies widely with some sites achieving a high proportion of free cooling and others constrained to operate in older, less efficient facilities. The PUE may range from 1.1 and 1.4. The XC30 system is designed to enable operation at the low end of this range. We vary this factor to illustrate its effect.

- Utility rates and contracts vary widely. We assume energy costs of $0.11/kilowatt hour in the US, €0.12/kilowatt hour in Europe (equivalent to $0.16/kilowatt hour) and ¥14/kilowatt hour in Japan (equivalent to $0.14/kilowatt hour).

---

18 Hybrid systems using Intel® Xeon® Phi™ coprocessors or NVIDIA Tesla GPU accelerators may consume more power per node.  
19 U.S. HPC community uses a guide figure for annual electricity costs of $1 million per megawatt ($0.11/kWh).  
20 Electricity costs vary across Europe with some HPC sites in Germany paying €0.15/kWh.
With this set of assumptions the annual cost of operating an XC30 compute cabinet containing 384 Intel Xeon E5 processors and 24 terabytes of memory is shown in Figure 14 below.

![Figure 14: Model cost of operation. Annual cost of operation for an XC30 compute cabinet with 384 Intel Xeon E5 processors and 24 terabytes of memory. Assumes utility rates of $0.11 in the U.S., $0.14(¥14) in Japan, and $0.16(€0.12) in Europe.](image)

A saving of 0.1 in the PUE results in savings of approximately $6,800 per cabinet per year or $100,000 per thousand nodes over three years. The XC30 design allows a wide range of Cray sites to achieve such savings without burdening the capital cost of the system or the facility cooling infrastructure. Please note that this is only a model calculation. Cray can undertake a detailed cost of ownership calculation for a specific configuration as part of a site review.

**Energy Efficiency Initiatives at Cray**

We prioritize energy efficiency in powering and cooling our systems. As a company we also promote energy efficiency in the manufacture, operation, and eventual decommissioning of these systems. Examples of energy-efficient facilities include:

- Our Chippewa Falls, Wisconsin facility is heated primarily by the heat generated in the assembly, bring up, and testing of customer production systems. Heat pumps used to move warm air

---

21 Japanese electricity prices were similar to those in the U.S. but have risen 20 percent since the March 2011 Tohoku earthquake.

22 Equivalent savings are approximately €8,000 per cabinet per year in Europe and ¥90,000 in Japan.
through the facility heating infrastructure are connected to the cooling water loops. During winter months the heat passes through the pumps prior to running through the cooling towers.

- We completed a project recently to replace the high pressure sodium lighting in a 50,000 square foot warehouse with 80 high-efficiency, fluorescent fixtures with motion sensors. Since the change, energy usage has dropped 75 percent. Not only does the new lighting system reduce energy consumption, the air conditioning costs have been reduced as well.

- We received a $250,000 grant recently from Focus on Energy, a program funded by Wisconsin utilities to help eligible companies and residents tackle cost-effective energy efficiency and renewable energy projects. While we have received lighting upgrade rebates previously through the program, this project was more significant. The power and cooling demand of XC30 systems made it necessary to upgrade the mechanical and electrical infrastructure for two of the production checkout bays in Chippewa Falls. By purchasing 1,460 tons of high efficiency chillers and 1,200-ton heat exchangers for “free air” cooling in the colder months, we will save significant money on electricity in addition to receiving the largest grant allowed by the Focus on Energy program.

- We seek to reduce our energy footprint by sourcing as many components locally as possible. Devices such as processors and memory are manufactured worldwide, but we source many of the larger components from suppliers close to our Chippewa Falls manufacturing facility.

- We work closely with suppliers to refine and implement environmentally sound packaging materials that maximize reuse and recycling and minimize packaging material volume, handling, waste, and related costs.

As an organization Cray seeks to reduce our environmental footprint through energy efficiency in our products and operations, reuse of parts and packaging, recycling, and minimizing use of exotic materials.

Conclusion

The Cray XC30 system provides an energy efficient platform for state-of-the-art computation that maximizes sustained performance per watt. Energy efficiency is optimized through use of processors and networking that minimize time to solution, a power train that delivers power efficiently, and a cooling system that makes direct use of facilities water. Overall power usage effectiveness can be as low as 1.1 in a modern datacenter. Peak performance of a Cray XC30 system ranges from 99 to 283 teraflops per cabinet depending on the choice of node. Performance per watt ranges from 1.3 to 3.2 gigaflops per watt. The modular structure of the Cray XC30 system allows infrastructure costs to be defrayed over multiple generations of processors. Cray Linux Environment accounts energy use of individual jobs and provides users with the information required to optimize for energy to solution.

---

23 Cray customers installing XC30 systems have also received substantial grants from energy efficiency rebate programs run by their local power utilities [15].

24 Performance figures at the time of writing, November 2013.
References


[2] For information on Google datacenters, see www.google.com/about/datacenters


[4] The Cray XC30 site planning guides provide detailed system data on water-cooled and air-cooled XC30 systems. See http://docs.cray.com for up-to-date copies of these documents.


[7] For more information on Cray Sonexion storage see www.sonexion.com


[12] For information on the Top500 supercomputer sites see www.top500.org

[13] For information on sustained petaflop performance see www.ncsa.illinois.edu/BlueWaters

[14] For information on the history of Cray supercomputers and their cooling systems see www.computerhistory.org


[16] For information on the Pawsey Centre see www.ivec.org

[17] For information on the Swiss National Supercomputer Centre see www.cscs.ch

[18] For information on the Green500 list see www.green500.org

[19] For information on Cray Cluster solutions see www.cray.com/cluster

[20] For information on Cray storage systems see www.cray.com/storage


Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency under its Agreement No. HR0011-07-9-0001. Only a small number of the many Cray employees that developed the Cray XC30 system were involved in the preparation of this paper. The authors would like to acknowledge the enormous effort made by the entire team.

© 2013 Cray Inc. All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior permission of the copyright owners.

Cray is a registered trademark, and the Cray logo, Cray XC, Cray XE6, and Cray Sonexion are trademarks of Cray Inc. Other product and service names mentioned herein are the trademarks of their respective owners.