CHAPTER 4

Data Center Basics: Building, Power, and Cooling

Internet and cloud services run on a planet-scale computer with workloads distributed across multiple data center buildings around the world. These data centers are designed to house computing, storage, and networking infrastructure. The main function of the buildings is to deliver the utilities needed by equipment and personnel there: power, cooling, shelter, and security. By classic definitions, there is little work produced at the data center. Other than some departing photons, all of the energy consumed is converted into heat. The delivery of input energy and subsequent removal of waste heat are at the heart of the data center's design and drive the vast majority of non-computing costs. These costs are proportional to the amount of power delivered and typically run in range of $10–20 per watt (see Chapter 6), but can vary considerably depending on size, location, and design.

4.1 DATA CENTER OVERVIEW

4.1.1 TIER CLASSIFICATIONS AND SPECIFICATIONS

The design of a data center is often classified using a system of four tiers [TSB]. The Uptime Institute, a professional services organization specializing in data centers, and the Telecommunications Industry Association (TIA), an industry group accredited by ANSI and comprised of approximately 400 member companies, both advocate a 4-tier classification loosely based on the power distribution, uninterruptible power supply (UPS), cooling delivery, and redundancy of the data center [UpIOS, TIA].

- *Tier I* data centers have a single path for power distribution, UPS, and cooling distribution, without redundant components.
- *Tier II* adds redundant components to this design (N + 1), improving availability.
- *Tier III* data centers have one active and one alternate distribution path for utilities. Each path has redundant components and is concurrently maintainable. Together they provide redundancy that allows planned maintenance without downtime.

L. A. Barroso et al., *The Datacenter as a Computer*

- *Tier IV* data centers have two simultaneously active power and cooling distribution paths, redundant components in each path, and are supposed to tolerate any single equipment failure without impacting the load.

The Uptime Institute's specification focuses on data center performance at a high level. The specification implies topology rather than prescribing a specific list of components to meet the requirements (notable exceptions are the amount of backup diesel fuel and water storage, and ASHRAE temperature design points [UpIT]). With the Uptime standards, there are many architectures that can achieve a given tier classification. In contrast, the TIA-942 standard is more prescriptive and specifies a variety of implementation details, such as building construction, ceiling height, voltage levels, types of racks, and patch cord labeling.

Formally achieving tier classification is difficult and requires a full review from one of the certifying bodies. For this reason most data centers are not formally rated. Most commercial data centers fall somewhere between tiers III and IV, choosing a balance between construction cost and reliability. Generally, the lowest individual subsystem rating (cooling, power, and so on) determines the overall tier classification of the data center.

Real-world data center reliability is strongly influenced by the quality of the organization running the data center, not just the design. Theoretical availability estimates used in the industry range from 99.7% for tier II data centers to 99.98% and 99.995% for tiers III and IV, respectively [TIA]. However, real-world reliability often is dominated by factors not included in these calculations; for example, the Uptime Institute reports that over 70% of data center outages are the result of human error, including management decisions on staffing, maintenance, and training [UpIOS]. Furthermore, in an environment using continuous integration and delivery of software, software-induced outages dominate building outages.

4.1.2 BUILDING BASICS

Data center sizes vary widely and are commonly described in terms of either the floor area for IT equipment or *critical power*, the total power that can be continuously supplied to IT equipment; Two thirds of U.S. servers were recently housed in data centers smaller than 5,000 ft^2 (450 square meters) and with less than 1 MW of critical power [EPA07, Koo11]. Large commercial data centers are built to host servers from multiple companies (often called co-location data centers, or "colos") and can support a critical load of tens of megawatts; the data centers of large cloud providers are similar, although often larger. Many data centers are single story, while some are multi-story (Figure 4.1); the critical power of some data center buildings can exceed 100 MW today.

Figure 4.1: Google's four-story cloud data center in Mayes County, Oklahoma.

At a high level, a data center building has multiple components. There is a *mechanical yard* (or a central utility building) that hosts all the cooling systems, such as cooling towers and chillers. There is an *electrical yard* that hosts all the electrical equipment, such as generators and power distribution centers. Within the data center, the main *server hall* hosts the compute, storage, and networking equipment organized into hot aisles and cold aisles. The server floor can also host *repair areas* for operations engineers. Most data centers also have separate areas designated for *networking*, including inter-cluster, campus-level, facility management, and long-haul connectivity. Given the criticality of networking for data center availability, the networking areas typically have additional physical security and high-availability features to ensure increased reliability. The *data center building* construction follows established codes around fire-resistive and non-combustible construction, safety, and so on [IBC15], and the design also incorporates elaborate *security* for access, including circle locks, metal detectors, guard personnel, and an extensive network of cameras.

Figure 4.2 shows an aerial view of a Google data center campus in Council Bluffs, Iowa. Figure 4.3 zooms in on one building to highlight some of the typical components in greater detail.

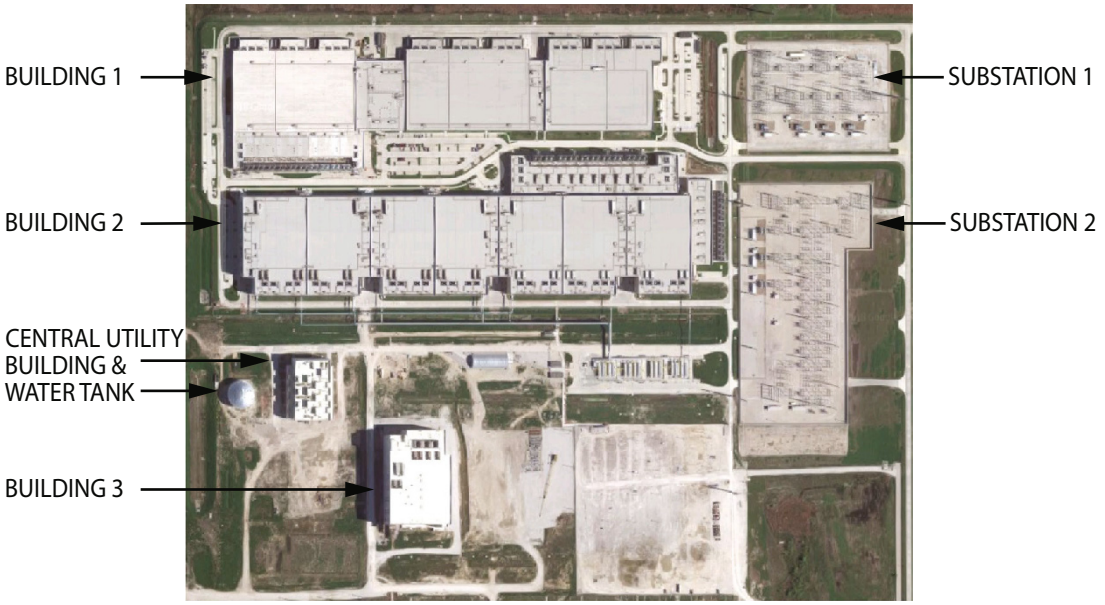

Figure 4.2: Aerial view of a Google data center campus in Council Bluffs, Iowa.

Figure 4.3: A Google data center building in Council Bluffs, Iowa, showing the mechanical yard, electrical yard, and server hall.

Figure 4.4 shows the components of a typical data center architecture. Beyond the IT equipment (discussed in Chapter 3), the two major systems in the data center provide power delivery (shown in red, indicated by numbers) and cooling (shown in green, indicated by letters). We discuss these in detail next.

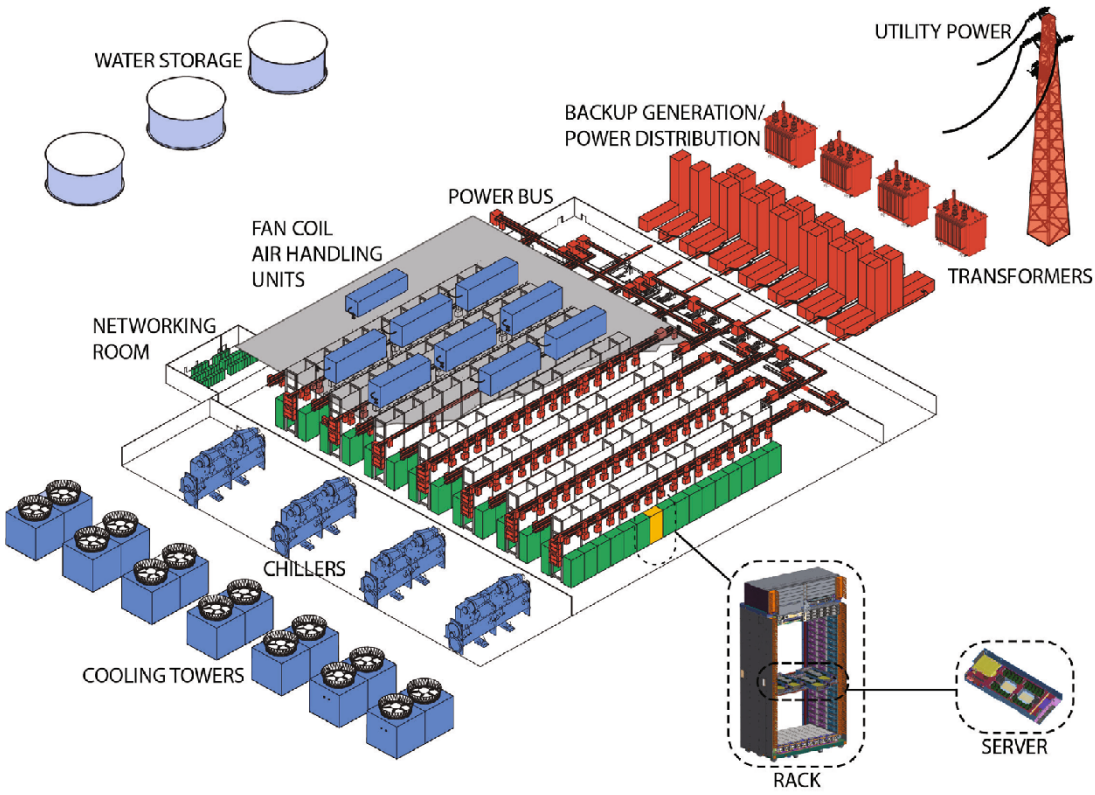

Figure 4.4: The main components of a typical data center.

4.2 DATA CENTER POWER SYSTEMS

Power enters first at a utility substation (not shown) which transforms high voltage (typically 110 kV and above) to medium voltage (typically less than 50 kV). Medium voltage is used for site-level distribution to the primary distribution centers (also known as unit substations), which include the primary switchgear and medium-to-low voltage transformers (typically below 1,000 V). From here, the power enters the building with the low-voltage lines going to the uninterruptible power supply (UPS) systems. The UPS switchgear also takes a second feed at the same voltage from a set of diesel generators that cut in when utility power fails. An alternative is to use a flywheel or alternator assembly, which is turned by an electric motor during normal operation, and couples to a diesel motor via a clutch during utility outages. In any case, the outputs of the UPS system are routed to the data center floor where they are connected to Power Distribution Units (PDUs). PDUs are the last layer in the transformation and distribution architecture and route individual circuits to the computer cabinets.

4.2.1 UNINTERRUPTIBLE POWER SYSTEMS (UPS)

The UPS typically combines three functions in one system.

- First, it contains a transfer switch that chooses the active power input (either utility power or generator power). After a power failure, the transfer switch senses when the generator has started and is ready to provide power; typically, a generator takes 10–15 s to start and assume the full rated load.
- Second, it contains some form of energy storage (electrical, chemical, or mechanical) to bridge the time between the utility failure and the availability of generator power.
- Third, it conditions the incoming power feed, removing voltage spikes or sags, or harmonic distortions in the AC feed. This conditioning can be accomplished via "double conversion."

A traditional UPS employs AC–DC–AC double conversion. Input AC is rectified to DC, which feeds a UPS-internal bus connected to strings of batteries. The output of the DC bus is then inverted back to AC to feed the data center PDUs. When utility power fails, input AC is lost but internal DC remains (from the batteries) so that AC output to the data center continues uninterrupted. Eventually, the generator starts and resupplies input AC power.

Traditional double-conversion architectures are robust but inefficient, wasting as much as 15% of the power flowing through them as heat. Newer designs such as line-interactive, delta-conversion, multi-mode, or flywheel systems operate at efficiencies in the range of 96–98% over a wide range of load cases. Additionally, "floating" battery architectures such as Google's on-board UPS [Whi+] place a battery on the output side of the server's AC/DC power supply, thus requiring only a small trickle of charge and a simple switching circuit. These systems have demonstrated efficiencies exceeding 99%. A similar strategy was later adopted by the OpenCompute UPS [OCP11], which distributes a rack of batteries for every four server racks, and by Google's high-availability rack systems, which contain servers powered from a rack-level DC bus fed from either modular, redundant rectifiers or modular, redundant battery trays.

Because UPS systems take up a sizeable amount of space, they are usually housed in a room separate from the data center floor. Typical UPS capacities range from hundreds of kilowatts up to two megawatts or more, depending on the power needs of the equipment. Larger capacities are achieved by combining several smaller units.

It's possible to use UPS systems not only in utility outages but also as supplementary energy buffers for power and energy management. We discuss these proposals further in the next chapter.

4.2.2 POWER DISTRIBUTION UNITS (PDUS)

In our example data center, the UPS output is routed to PDUs on the data center floor. PDUs resemble breaker panels in residential houses but can also incorporate transformers for final voltage adjustments. They take a larger input feed and break it into many smaller circuits that distribute power to the actual servers on the floor. Each circuit is protected by its own breaker, so a short in a server or power supply will trip only the breaker for that circuit, not the entire PDU or even the UPS. A traditional PDU handles 75–225 kW of load, whereas a traditional circuit handles a maximum of approximately 6 kW (20 or 30 A at 110–230 V). The size of PDUs found in large-scale data centers is much higher, however, corresponding to the size of the largest commodity backup generators (in megawatts), with circuits sometimes corresponding to high-power racks ranging in the tens of kW capacity. PDUs often provide additional redundancy by accepting two independent ("A" and "B") power sources and are able to switch between them with a small delay. The loss of one source does not interrupt power to the servers. In this scenario, the data center's UPS units are usually duplicated on A and B sides, so that even a UPS failure will not interrupt server power.

In North America, the input to the PDU is commonly 480 V 3-phase power. This requires the PDU to perform a final transformation step to deliver the desired 110 V output for the servers, thus introducing another source of inefficiency. In the EU, input to the PDU is typically 400 V 3-phase power. By taking power from any single phase to neutral combination, it is possible to deliver a desirable 230 V without an extra transformer step. Using the same trick in North America requires computer equipment to accept 277 V (as derived from the 480 V input to the PDU), which unfortunately exceeds the upper range of standard power supplies.

Real-world data centers contain many variants of the simplified design described here. These include the "paralleling" of generators or UPS units, an arrangement where multiple devices feed a shared bus so the load of a failed device can be picked up by other devices, similar to handling disk failures in a RAID system. Common paralleling configurations include N + 1 (allowing one failure or maintenance operation at a time), N + 2 (allowing one failure even when one unit is offline for maintenance), and 2N (fully redundant pairs).

4.2.3 COMPARISON OF AC AND DC DISTRIBUTION ARCHITECTURES

The use of high-voltage DC (HVDC) on the utility grid presents advantages for connecting incompatible power grids, providing resistance to cascading failures, and long-distance transmission efficiency. In data centers, the case for DC distribution is centered around efficiency improvements, increased reliability from reduced component counts, and easier integration of distributed generators with native DC outputs. In comparison with the double-conversion UPS mentioned above, DC systems eliminate the final inversion step of the UPS. If the voltage is selected to match the DC primary stage of the server power supply unit (PSU), three additional steps are eliminated: PDU transformation, PSU rectification, and PSU power factor correction.

Figure 4.5 compares AC and DC distribution architectures commonly used in the data center industry. State-of-the-art, commercially available efficiencies (based on [GF]) are shown for each stage of the "power train." The overall power train efficiency using state-of-the-art components remains a few percent higher for DC distribution as compared to AC distribution; this difference was more pronounced in data centers with older components [Pra+06]. Note that the AC architecture shown corresponds to the voltage scheme commonly found in North America; in most other parts of the world the additional voltage transformation in the AC PDU can be avoided, leading to slightly higher PDU efficiency.

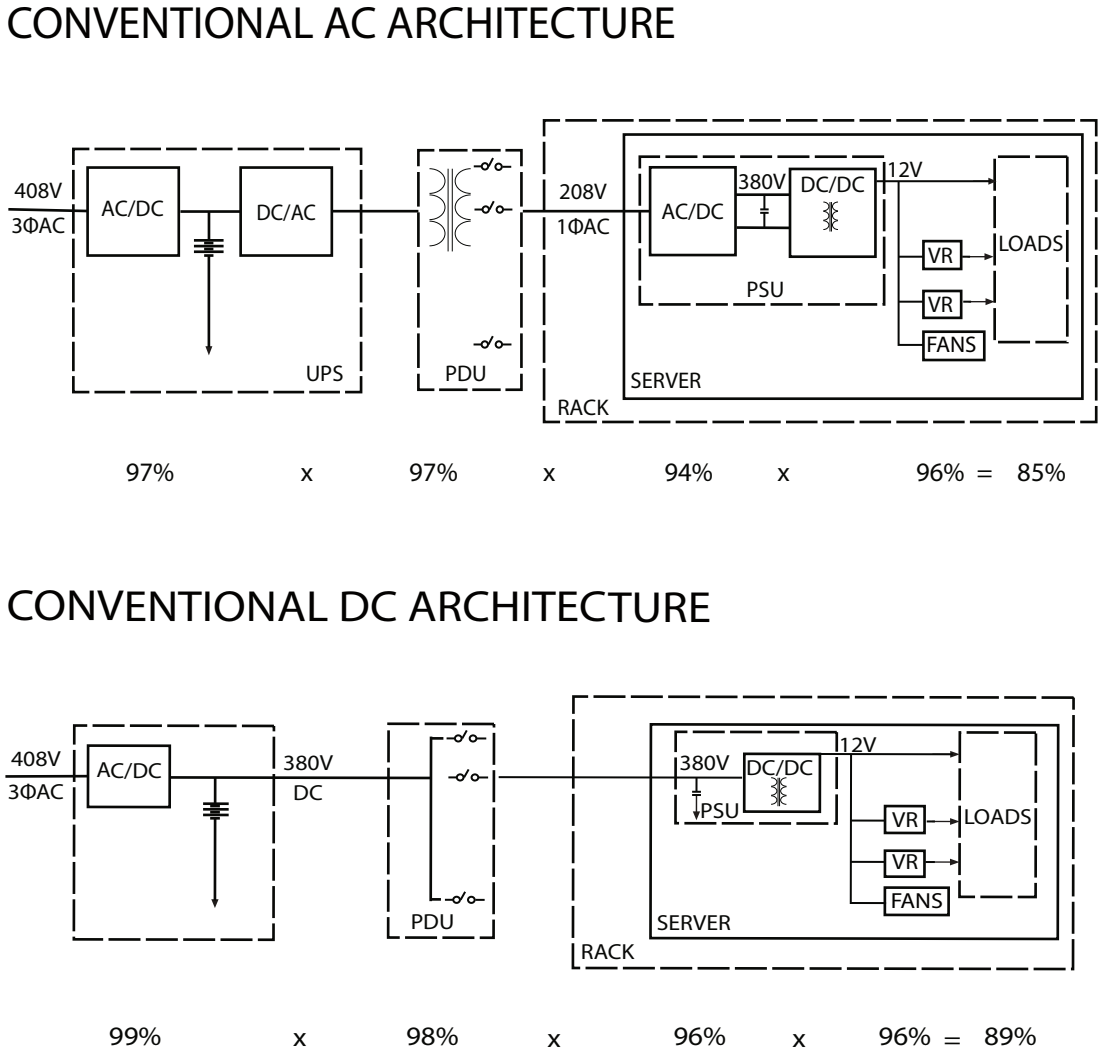

Figure 4.5: Comparison of AC and DC distribution architectures commonly employed in the data center industry.

The placement of batteries as parallel sources near the load can virtually eliminate UPS losses: ~0.1% compared to ~1%–3% for an in-line UPS with 480 V AC input. In typical Google designs, batteries are included either on server trays or as modules of a rack in parallel with a DC bus, and this has allowed elimination of the upstream UPS, further increasing the power train efficiency.

Commercial DC equipment for data centers is available, but costs remain higher than for comparable AC equipment. Similarly, the construction of large data centers involves hundreds and sometimes thousands of skilled workers. While only a subset of these will be electricians, the limited availability of DC technicians may lead to increased construction, service, and operational costs. However, DC power distribution is more attractive when integrating distributed power generators such as solar photovoltaic, fuel cells, and wind turbines. These power sources typically produce native DC and are easily integrated into a DC power distribution architecture.

4.3 EXAMPLE: RADIAL POWER DISTRIBUTION WITH REDUNDANCY

A conventional AC power distribution scheme for a large data center is shown in Figure 4.6. This topology is known as "radial" because power fans out to the entire data center floor from a pair of medium voltage buses that provide redundancy in case of loss of a utility feed. Low voltage (400–480 V) AC power is supplied to the data center floor by many PDUs, each fed by either a step-down transformer for utility power or a backup generator. In addition, power availability is greatly enhanced by an isolated redundant PDU. This module is identical to the others, but needs to carry load only when other low voltage equipment fails or needs to be taken temporarily out of service.

4.4 EXAMPLE: MEDIUM VOLTAGE POWER PLANE

An interesting modern architecture for data center power distribution is Google's medium voltage power plane (Figure 4.7), which allows for sharing of power across the data center. High availability at the building level is provided by redundant utility AC inputs. Building-level transformers step the voltage down to a medium voltage of 11–15 kV for further distribution through the building's electrical rooms. For backup power, a "farm" of many medium voltage generators are paralleled to a bus, and automated systems consisting of breakers and switches select between the utility and generator sources. Redundant paths exist from both utility and generator sources to many unit substations. Each unit substation steps down the voltage to approximately 400 V AC for distribution to a row of racks on the data center floor.

The power plane architecture offers several advantages with respect to traditional radial architectures. First, a large pool of diverse workloads can increase the opportunity for power oversubscription [Ran+06], discussed in the following chapter. Roughly speaking, Google's power plane architecture doubles the quantity of IT equipment that can be deployed above and beyond a data

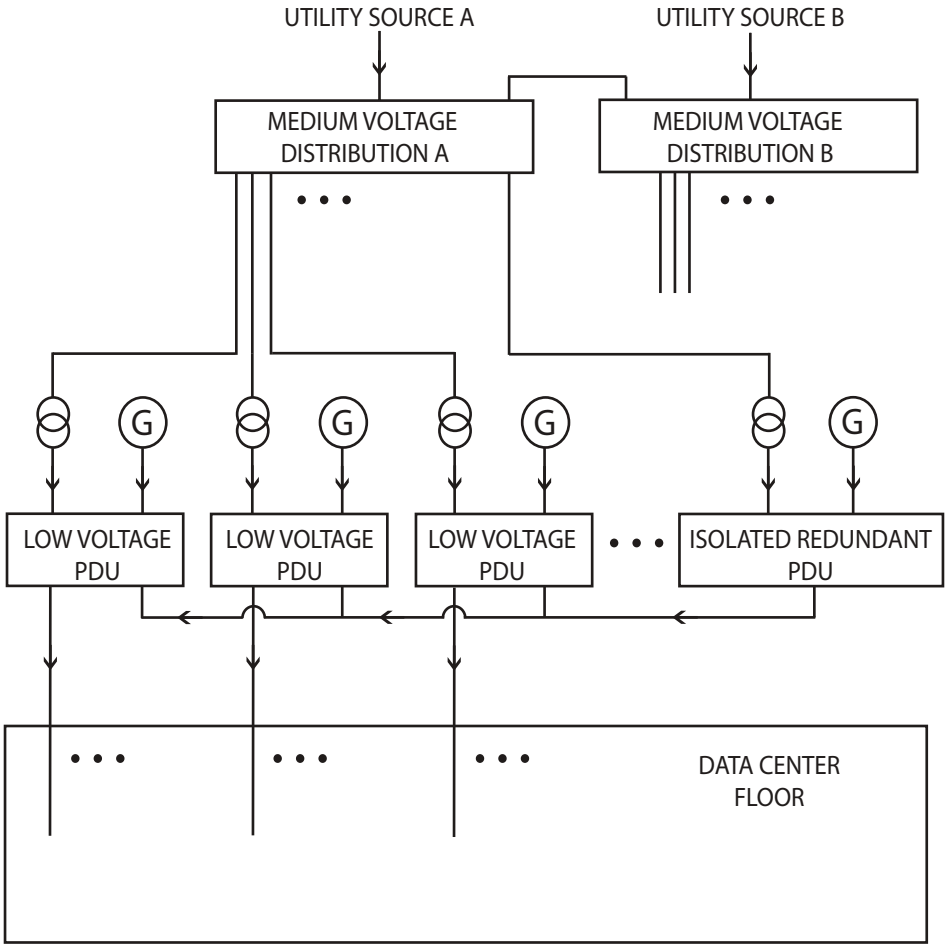

Figure 4.6: A radial power architecture with generators distributed among many low voltage PDUs. Each PDU has its own backup generator indicated by a "G." Power loss due to failures of low voltage equipment is greatly mitigated by the presence of an isolated redundant PDU, which can take the place of any other PDU as a low voltage source.

center's critical power capacity. Second, the generator farm offers resilience against generator failures with a minimum of redundant equipment. Finally, power is more fungible across the entire data center floor: with appropriate sizing of both medium- and low-voltage distribution components, a high dynamic range of deployment power density can be supported without stranding power. This is an important benefit given that rack power varies substantially depending on the type of IT equipment within the rack. For example, storage-intensive racks consume much less power than compute-intensive racks. With traditional radial power architectures, a low power density in one region of the data center floor can result in permanently underused infrastructure. The medium-voltage power plane enables power sharing across the floor: high power racks in one region can compensate for low power racks in another region, ensuring full utilization of the building's power capacity.

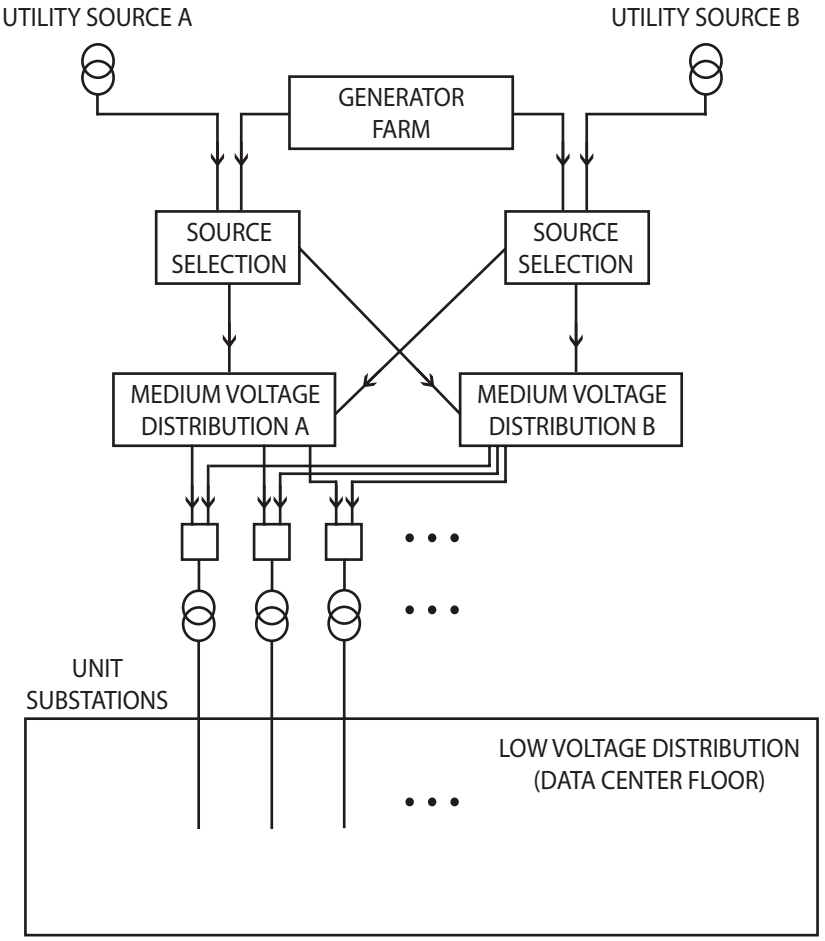

Figure 4.7: Concept for medium-voltage power plane architecture.

4.5 DATA CENTER COOLING SYSTEMS

Data center cooling systems remove the heat generated by the equipment. To remove heat, a cooling system must employ some hierarchy of loops, each circulating a cold medium that warms up via some form of heat exchange and is somehow cooled again. An open loop replaces the outgoing warm medium with a cool supply from the outside, so that each cycle through the loop uses new material. A closed loop recirculates a separate medium, continuously transferring heat to either another loop via a heat exchanger or to the environment; all systems of loops must eventually transfer heat to the outside environment.

The simplest topology is fresh air cooling (or air economization)—essentially, opening the windows. Such a system is shown in Figure 4.8. This is a single, open-loop system that we discuss in more detail in the section on free cooling.

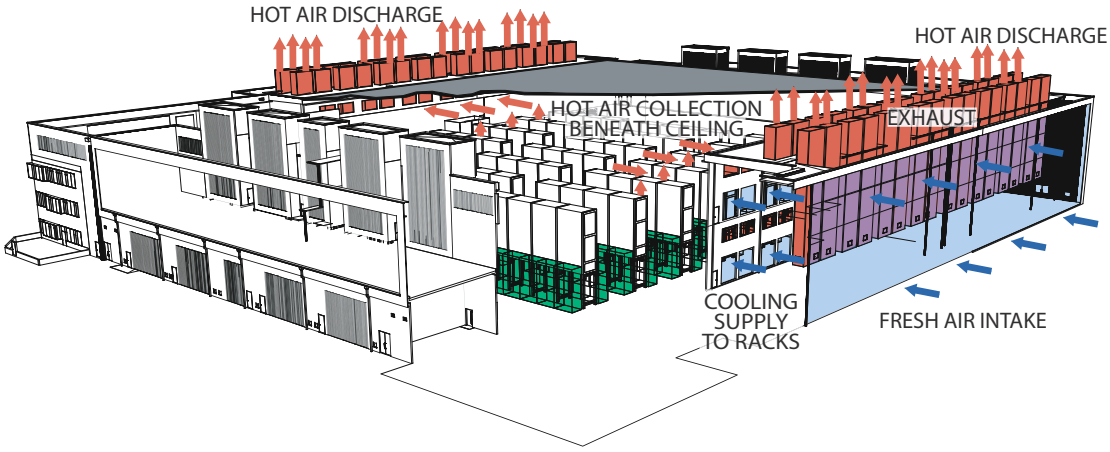

Figure 4.8: Airflow schematic of an air-economized data center.

Closed-loop systems come in many forms, the most common being the air circuit on the data center floor. Its function is to isolate and remove heat from the servers and transport it to a heat exchanger. As shown in Figure 4.9, cold air flows to the servers, heats up, and eventually reaches a heat exchanger to cool it down again for the next cycle through the servers.

Typically, data centers employ raised floors, concrete tiles installed onto a steel grid resting on stanchions two to four feet above the slab floor. The underfloor area often contains power cables to racks, but its primary purpose is to distribute cool air to the server racks. The airflow through the underfloor plenum, the racks, and back to the CRAC (a 1960s term for *computer room air conditioning*) defines the primary air circuit.

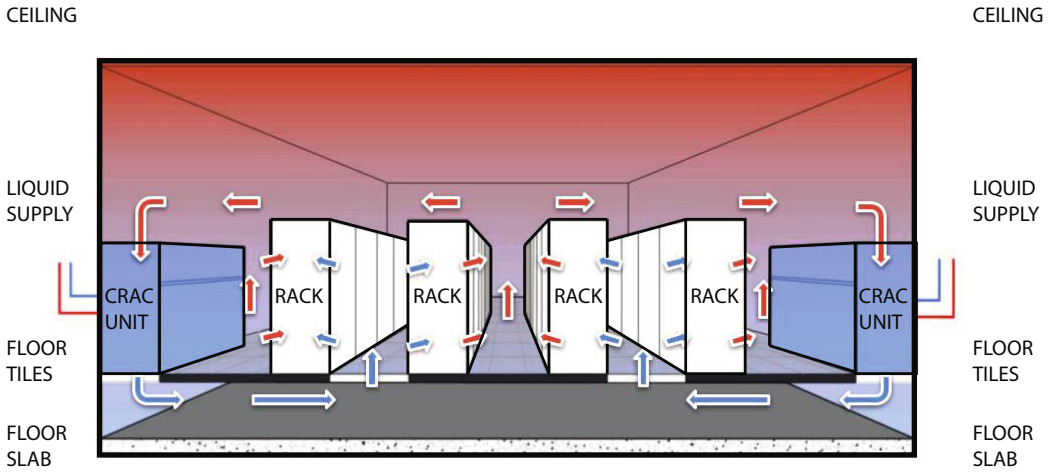

Figure 4.9: Raised floor data center with hot-cold aisle setup (image courtesy of DLB Associates [Dye06]).

The simplest closed-loop systems contain two loops. The first loop is the air circuit shown in Figure 4.9, and the second loop (the liquid supply inside the CRACs) leads directly from the CRAC to external heat exchangers (typically placed on the building roof) that discharge the heat to the environment.

A three-loop system commonly used in large-scale data centers is shown in Figure 4.10. The first *datacenter floor loop* involves circulating air that is alternately cooled by fan coils and heated by IT equipment on the data center floor. In the *process loop*, warm water from the fan coils returns to the cooling plant to be chilled and pumped back to the fan coils. Finally, the *condenser water loop* removes heat received from the process water through a combination of mechanical refrigeration by chiller units and evaporation in cooling towers; the condenser loop is so named because it removes heat from the condenser side of the chiller. Heat exchangers perform much of the heat transfer between the loops, while preventing process water from mixing with condenser water.

Each topology presents tradeoffs in complexity, efficiency, and cost. For example, fresh air cooling can be very efficient but does not work in all climates, requires filtering of airborne particulates, and can introduce complex control problems. Two-loop systems are easy to implement, relatively inexpensive to construct, and offer isolation from external contamination, but typically have lower operational efficiency. A three-loop system is the most expensive to construct and has moderately complex controls, but offers contaminant protection and good efficiency when employing economizers.

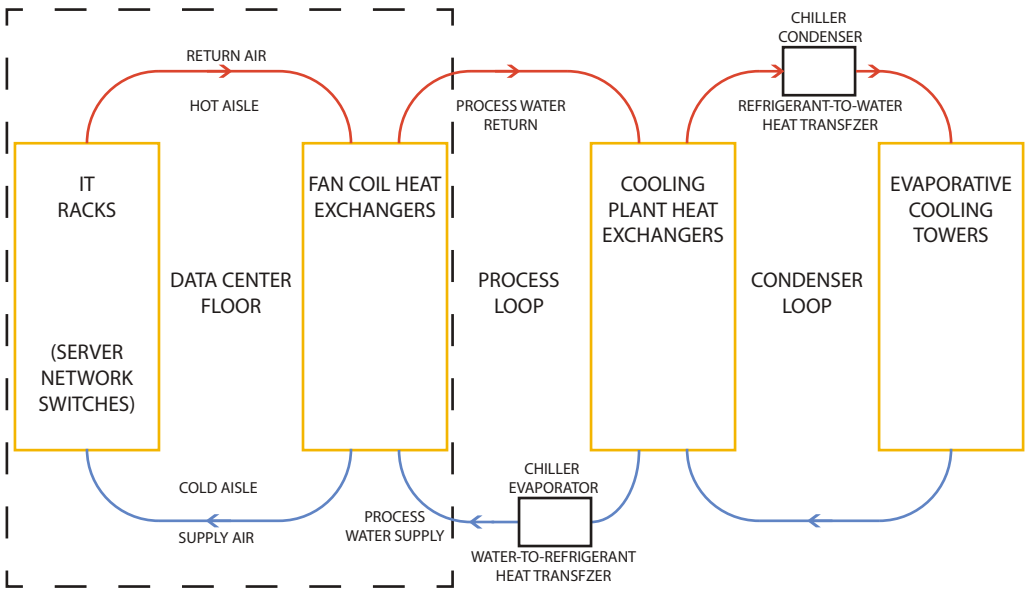

Figure 4.10: Three-loop data center cooling system. (Note that in favorable weather conditions, the entire data center heat load can be removed by evaporative cooling of the condenser water; the chiller evaporator and chiller condenser heat transfer steps then become unnecessary.)

Additionally, generators (and sometimes UPS units) provide backup power for most mechanical cooling equipment because the data center may overheat in a matter of minutes without cooling. In a typical data center, chillers and pumps can add 40% or more to the critical load supported by generators, significantly adding to the overall construction cost.

CRACs, chillers, and cooling towers are among the most important building blocks in data center cooling systems, and we take a slightly closer look at each below.

4.5.1 COMPUTER ROOM AIR CONDITIONERS (CRACS)

All CRACs contain a heat exchanger, air mover, and controls. They mostly differ by the type of cooling they employ:

- direct expansion (DX);
- fluid solution; and
- water.

A DX unit is a split air conditioner with cooling (evaporator) coils inside the CRAC, and heat-rejecting (condenser) coils outside the data center. The fluid solution CRAC shares this basic

architecture but circulates a mixture of water and glycol through its coils rather than a phase-change refrigerant. Finally, a water-cooled CRAC connects to a chilled water loop.

CRAC units pressurize the raised floor plenum by blowing cold air into the underfloor space, which then escapes through perforated tiles in front of the server racks. The air flows through the servers and is expelled into a "hot aisle." Racks are typically arranged in long rows that alternate between cold and hot aisles to reduce inefficiencies caused by mixing hot and cold air. In fact, many newer data centers physically isolate the cold or hot aisles with walls [PF]. As shown in Figure 4.9, the hot air produced by the servers recirculates back to the intakes of the CRACs, where it is cooled and exhausted into the raised floor plenum again.

4.5.2 CHILLERS

A water-cooled chiller as shown in Figure 4.11 can be thought of as a water-cooled air conditioner.

Figure 4.11: Water-cooled centrifugal chiller.

Chillers submerge the evaporator and condenser coils in water in two large, separate compartments joined via a top-mounted refrigeration system consisting of a compressor, expansion valve, and piping. In the cold compartment, warm water from the data center is cooled by the evaporator coil prior to returning to the process chilled water supply (PCWS) loop. In the hot compartment, cool water from the condenser water loop is warmed by the condenser coil and carries the heat away

to the cooling towers where it is rejected to the environment by evaporative cooling. Because the chiller uses a compressor, a significant amount of energy is consumed to perform its work.

4.5.3 COOLING TOWERS

Cooling towers (Figure 4.12) cool a water stream by evaporating a portion of it into the atmosphere. The energy required to change the liquid into a gas is known as the latent heat of vaporization, and the temperature of the water can be dropped significantly given favorable dry conditions. The water flowing through the tower comes directly from the chillers or from another heat exchanger connected to the PCWS loop. Figure 4.13 illustrates how it works.

Figure 4.12: Data center cooling towers.

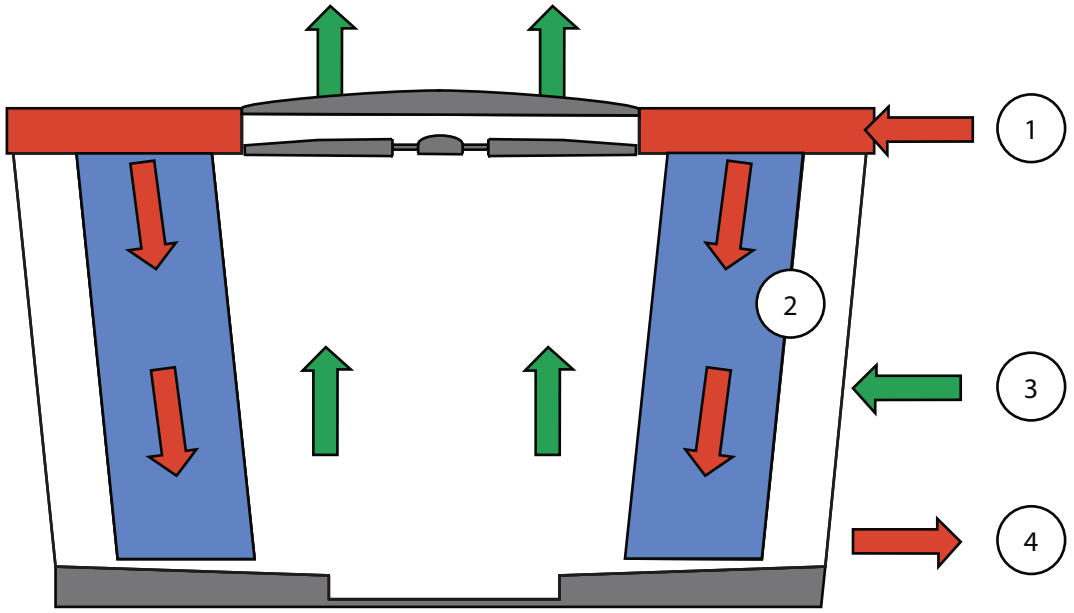

Figure 4.13: How a cooling tower works. The numbers correspond to the associated discussions in the text.

1. Hot water from the data center flows from the top of the cooling tower onto "fill" material inside the tower. The fill creates additional surface area to improve evaporation performance.

2. As the water flows down the tower, some of it evaporates, drawing energy out of the remaining water and reducing its temperature.

3. A fan on top draws air through the tower to aid evaporation. Dry air enters the sides and humid air exits the top.

4. The cool water is collected at the base of the tower and returned to the data center.

Cooling towers work best in temperate climates with low humidity; ironically, they do not work as well in very cold climates because they need additional mechanisms to prevent ice formation on the towers and in the pipes.

4.5.4 FREE COOLING

Free cooling refers to the use of cold outside air to either help produce chilled water or directly cool servers. It is not completely free in the sense of zero cost, but it involves very low-energy costs compared to chillers.

As mentioned above, air-economized data centers are open to the external environment and use low dry bulb temperatures for cooling. (The dry bulb temperature is the air temperature measured by a conventional thermometer). Large fans push outside air directly into the room or the raised floor plenum when outside temperatures are within limits (for an extreme experiment in this area, see [AM08]). Once the air flows through the servers, it is expelled outside the building. An air-economized system can be very efficient but requires effective filtering to control contamination, may require auxiliary cooling (when external conditions are not favorable), and may be difficult to control. Specifically, if there is a malfunction, temperatures will rise very quickly since air can store relatively little heat. By contrast, a water-based system can use a water storage tank to provide a significant thermal buffer.

Water-economized data centers take advantage of the wet bulb temperature [Wbt]. The wet bulb temperature is the lowest water temperature that can be reached by evaporation. The dryer the air, the bigger the difference between dry bulb and wet bulb temperatures; the difference can exceed 10°C, and thus a water-economized data center can run without chillers for many more hours per year. For this reason, some air-economized data centers employ a hybrid system where water is misted into the airstream (prior to entering the data center) in order to take advantage of evaporation cooling.

Typical water-economized data centers employ a parallel heat exchanger so that the chiller can be turned off when the wet bulb temperature is favorable. Depending on the capacity of the cooling tower (which increases as the wet bulb temperature decreases), a control system balances water flow between the chiller and the cooling tower.

Yet another approach uses a radiator instead of a cooling tower, pumping the condenser fluid or process water through a fan-cooled radiator. Similar to the glycol/water-based CRAC, such systems use a glycol-based loop to avoid freezing. Radiators work well in cold climates (say, a winter in Chicago) but less well at moderate or warm temperatures because the achievable cold temperature is limited by the external dry bulb temperature, and because convection is less efficient than evaporation.

4.5.5 AIR FLOW CONSIDERATIONS

Most data centers use the raised floor setup discussed above. To change the amount of cooling delivered to a particular rack or row, we exchange perforated tiles with solid tiles or vice versa. For cooling to work well, the cold airflow coming through the tiles should match the horizontal airflow through the servers in the rack. For example, if a rack has 10 servers with an airflow of 100 cubic feet per minute (CFM) each, then the net flow out of the perforated tile should be 1,000 CFM (or higher if the air path to the servers is not tightly controlled). If it is lower, some of the servers will receive cold air while others will ingest recirculated warm air from above the rack or other leakage paths.

Figure 4.14 shows the results of a Computational Fluid Dynamics (CFD) analysis for a rack that is oversubscribing the data center's airflow.

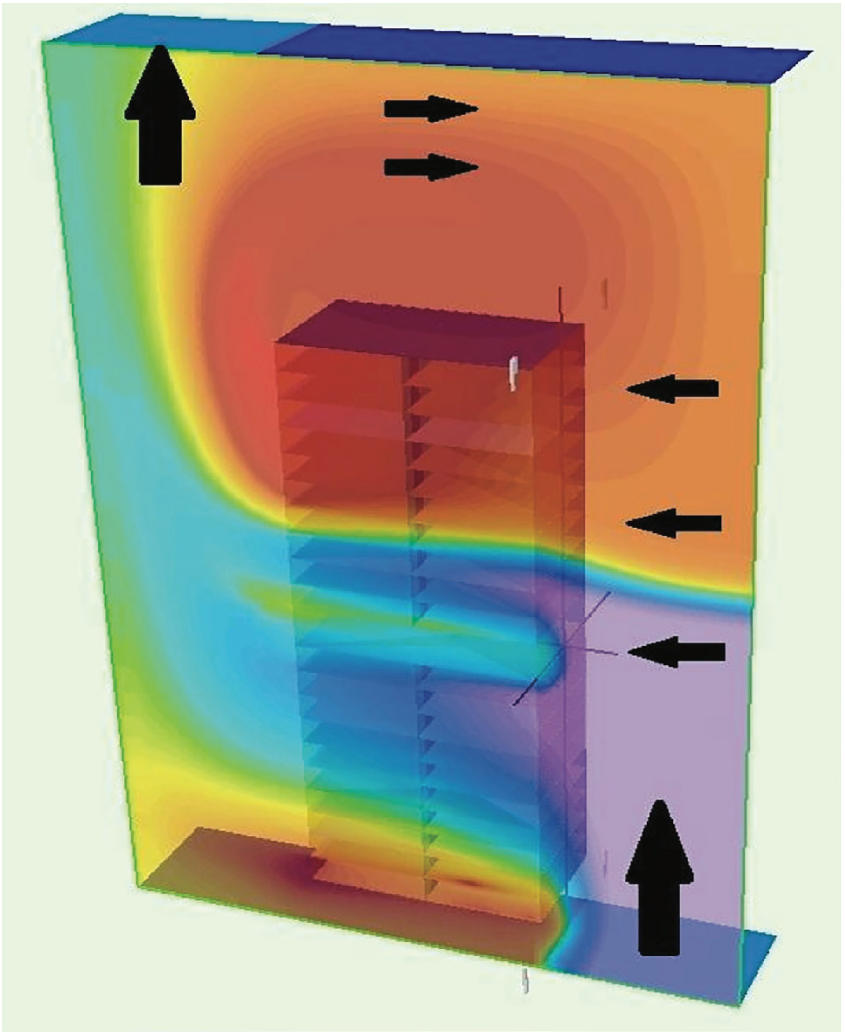

Figure 4.14: CFD model showing recirculation paths and temperature stratification for a rack with under-provisioned airflow.

In this example, recirculation across the top of the rack causes the upper servers to ingest warm air. The servers on the bottom are also affected by a recirculation path under the rack. Blockages from cable management hardware cause a moderate warm zone about halfway up the rack.

The facility manager's typical response to such a situation is to lower the temperature of the CRAC output. That works, but increases energy costs significantly, so it's better to fix the underly-

ing problem instead and physically separate cold and warm air as much as possible, while optimizing the path back to the CRACs. In this setup the entire room is filled with cool air (because the warm exhaust is kept inside a separate plenum or duct system) and, thus, all servers in a rack will ingest air at the same temperature [PF].

Air flow limits the power density of data centers. For a fixed temperature differential across a server, a rack's airflow requirement increases with power consumption, and the airflow supplied via the raised floor tiles must increase linearly with power. That in turn increases the amount of static pressure needed in the underfloor plenum. At low power densities, this is easy to accomplish, but at some point the laws of physics start to make it economically impractical to further increase pressure and airflow. Typically, these limitations make it hard to exceed power densities of more than 150–200 W/sq ft without substantially increased cost.

4.5.6 IN-RACK, IN-ROW, AND LIQUID COOLING

In-rack cooling can increase power density and cooling efficiency beyond the conventional raised-floor limit. Typically, an in-rack cooler adds an air-to-water heat exchanger at the back of a rack so the hot air exiting the servers immediately flows over coils cooled by water, essentially short-circuiting the path between server exhaust and CRAC input. In-rack cooling might remove part or all of the heat, effectively replacing the CRACs. Obviously, chilled water needs to be brought to each rack, greatly increasing the cost of plumbing. Some operators may also worry about having water on the data center floor, since leaky coils or accidents might cause water to spill on the equipment.

In-row cooling works like in-rack cooling except the cooling coils aren't in the rack, but adjacent to the rack. A capture plenum directs the hot air to the coils and prevents leakage into the cold aisle. Figure 4.15 shows an in-row cooling product and how it is placed between racks.

Finally, we can directly cool server components using cold plates, that is, local, liquid-cooled heat sinks. It is usually impractical to cool all compute components with cold plates. Instead, components with the highest power dissipation (such as processor chips) are targeted for liquid cooling while other components are air-cooled. The liquid circulating through the heat sinks transports the heat to a liquid-to-air or liquid-to-liquid heat exchanger that can be placed close to the tray or rack, or be part of the data center building (such as a cooling tower).

Figure 4.15: In-row air conditioner.

In spite of the higher cost and mechanical design complexity, cold plates are becoming essential for cooling very high-density workloads whose TDP per chip exceeds what is practical to cool with regular heatsinks (typically, 200–250 W per chip). A recent example is Google's third-generation tensor processing unit (TPU): as shown in Figure 4.16, four TPUs on the same motherboard are cooled in series on a single water loop.

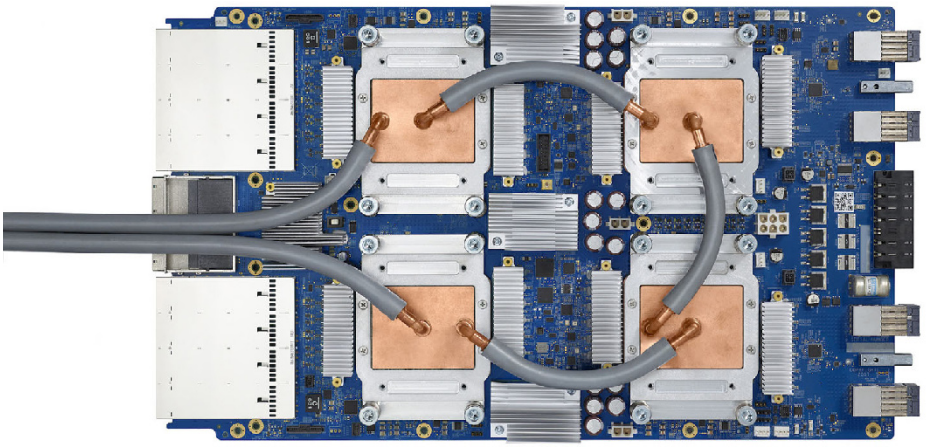

Figure 4.16: Copper cold plates and hose connections provide liquid cooling for Google's third-generation TPU.

4.5.7 CONTAINER-BASED DATA CENTERS

Container-based data centers go one step beyond in-row cooling by placing the server racks inside a container (typically 20 or 40 ft long) and integrating heat exchange and power distribution into the container as well. Similar to in-row cooling, the container needs a supply of chilled water and uses coils to remove all heat. Close-coupled air handling typically allows higher power densities than regular raised-floor data centers. Thus, container-based data centers provide all the functions of a typical data center room (racks, CRACs, PDU, cabling, lighting) in a small package. Figure 4.17 shows an isometric cutaway of Google's container design.

Like a regular data center room, containers must be accompanied by outside infrastructure such as chillers, generators, and UPS units to be fully functional.

To our knowledge, the first container-based data center was built by Google in 2005 [GInc09], and the idea dates back to a Google patent application in 2003. However, subsequent generations of Google data centers have moved away from containers and instead incorporate the same principles at a broader warehouse level. Some other large-scale operators, including Microsoft [Micro] and eBay [eBay12], have also reported using containers in their facilities, but today they are uncommon.

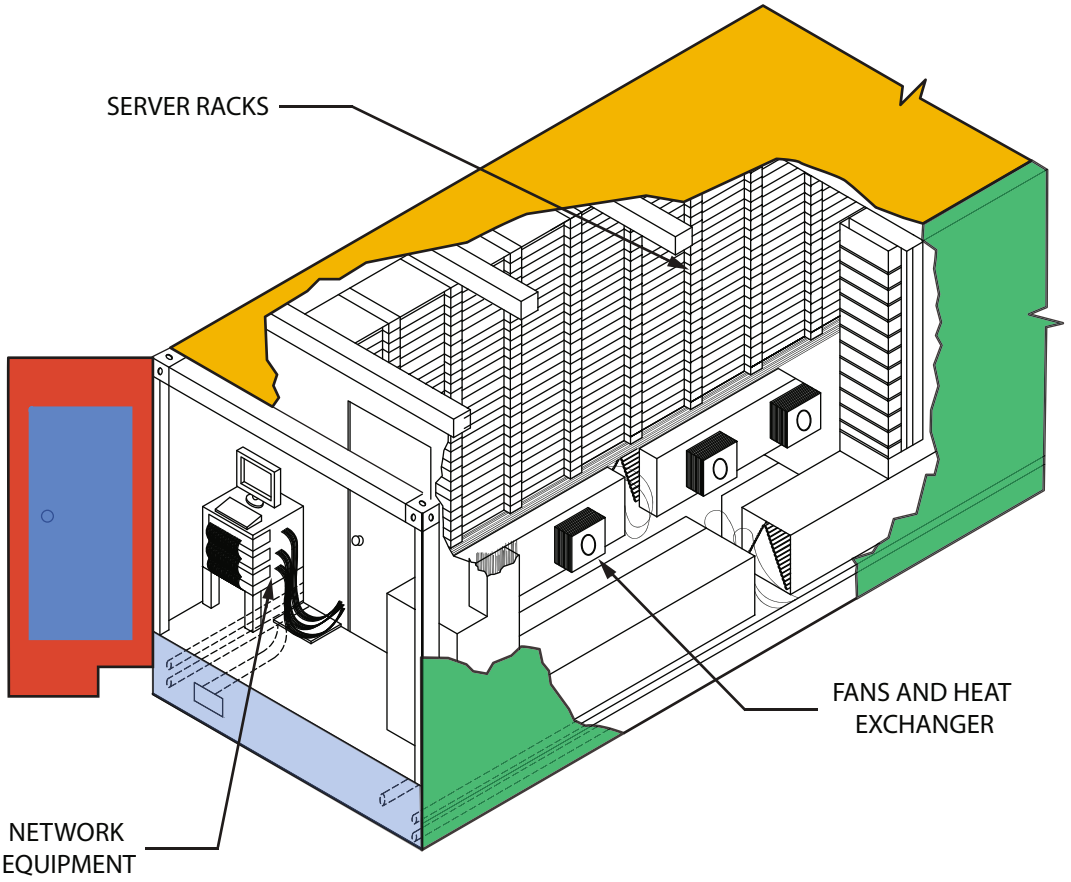

Figure 4.17: Google's container design includes all the infrastructure of the data center floor.

4.6 EXAMPLE: GOOGLE'S CEILING-MOUNTED COOLING FOR THE DATA CENTER

Figure 4.18 illustrates the main features and air flow of Google's overhead cooling system. This represents one variation on the efficient hot aisle containment that has become prevalent in the data center industry. Tall, vertical hot aisle plenums duct the exhaust air from the rear of the IT racks to overhead fan coils. The fan coils receive chilled process water from an external cooling plant; this water flows through multiple tube passages attached to fins, absorbing heat from the incoming hot air. Blowers in the fan coil units force the cooled air downward into the cold aisle, where it enters the intakes of servers and networking equipment. Together with the cooling plant and process

water loops, this air loop comprises a highly-efficient, end-to-end cooling system that consumes energy amounting to <10% of the energy consumed by the IT equipment.

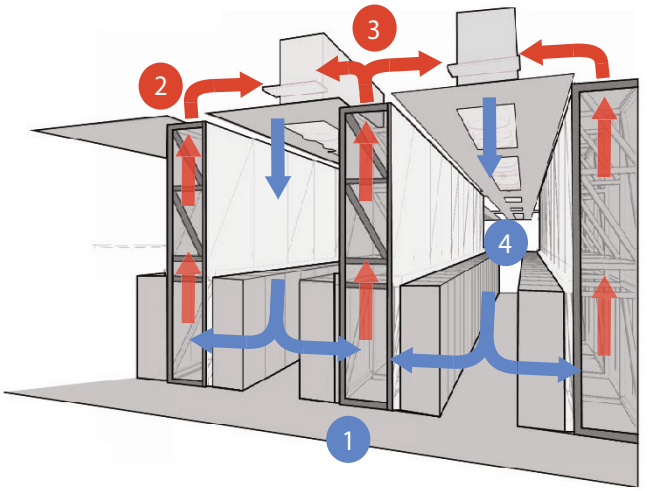

Figure 4.18: Cross-sectional view of a cold aisle and associated hot air plenums in a Google data center. (1) Hot exhaust from IT equipment rises in a vertical plenum space. (2) Hot air enters a large plenum space above the drop ceiling. (3) Heat is exchanged with process water in a fan coil unit, which also (4) blows the cold air down toward the intake of the IT equipment.

4.7 SUMMARY

Data centers power the servers they contain and remove the heat generated. Historically, data centers have consumed twice as much energy as needed to power the servers, but when best practices are employed this overhead shrinks to 10–20%. Key energy saving techniques include free-cooling (further boosted by raising the target inlet temperature of servers), well-managed air flow, and high-efficiency power distribution and UPS components.