

Redundancy

Like hospitals, data centers require a high level of redundancy to ensure they never go down, and the generators are part of this preparedness. Redundancy is often also created through multiple power feeds. The servers have backup capacity, too. For instance, your inbox is stored on several servers so that if one fails, your emails remain accessible.

Mechanical yards

The mechanical yards operate similarly but typically focus on cooling rather than on power redundancy. On the ground, you'll find chillers that cool the data center. At some data centers, they are on the roof.

Heat is a byproduct of computer processing. The servers within data centers must remain cool to function correctly. There are multiple ways of to achieve this cooling, but for the purposes of this paper, we focus on three main categories: air-cooled chillers, water-cooled chillers, and evaporative cooling.

A highly efficient **air-cooled chiller system** is sealed in a closed loop and uses little to no water—less than the amount used in three single-family homes.

Both water-cooled and evaporative systems use more water and consume less electricity than air-cooled chiller systems. In an open-loop **water-cooled chiller system**, cooling towers evaporate water for cooling purposes. **Evaporative cooling** is primarily achieved by large air-handling units that move air

across a wet medium to evaporate water, which removes heat and cools the air in the data hall.

Data center structure

Let's now make our way through the security checkpoint and into the data center building itself. Inside, we find the **data hall**, where the servers sit in rows of tall racks along a series of aisles fed by **electrical distribution equipment**. Data halls are the *pièce de résistance* of any data center. Such a structure is typically designed from the data hall outward, with all supporting infrastructure responding directly to its needs.

If you examine the racks from the floor upward, you see bundles of colorful fiber optic cables (inside of which are thin strands of glass) connecting servers to switches and routers. These cables enable high-speed data transfer. Without them, we would be unable to access the web or many of the apps we use daily. The cables are organized in lanes that converge at massive **switches**.

Inside the data center building, you also find electrical rooms that house other critical power supply infrastructure. Battery-powered **uninterruptible power supply (UPS) systems** sit in large cabinets against the wall. In the event of a utility outage, these systems feed power to **power distribution units (PDUs)** inside the data hall. UPS systems are the first-used backup power source and often prevent use of the generators outside.

The importance of clustering

The internet is a network of interconnected networks, and so are data centers. **Clustering** refers to the practice of linking the servers at multiple data centers with high-speed, low-latency connections so they work together as a unified system. This interconnected group is called a cluster, and each server in the cluster is called a **node**.

Importantly, clustering requires physical proximity: locating multiple data centers close to one another geographically so that optical signals in fiber cables don't lose strength. That's why a huge portion of global internet traffic passes through major hubs such as northern Virginia's "Data Center Alley."

Clustering also requires interconnection—linking these data centers through high-speed, low-latency connections.

Reduced latency

One primary reason for data center clustering is to reduce **latency**, the time it takes for information to travel from its source to its destination. When data centers are geographically distant from each other, latency increases, which leads to slower application performance for end users, including the dreaded "spinning wheel of death." When data centers are

located near each other, data can travel shorter distances, and this information exchange runs faster. Due to quicker response times, the user experience improves.

Improved reliability

Clustering enhances redundancy. If one data center server goes offline, others in the cluster can take over, ensuring continuous service availability. Without clusters, one server failure could cause an outage with international implications. In the appendix, we explore this concept further through a case study.

Load balancing

Clustered data centers allow for more efficient **load balancing**, the process of distributing network traffic across multiple servers at interconnected data centers. This practice prevents any single server or center from becoming overwhelmed.

Shared infrastructure

Data centers require conduits for power and, in some cases, water. When data centers cluster, they benefit from shared power and cooling infrastructure while also reducing the need for long-distance fiber optic connections.