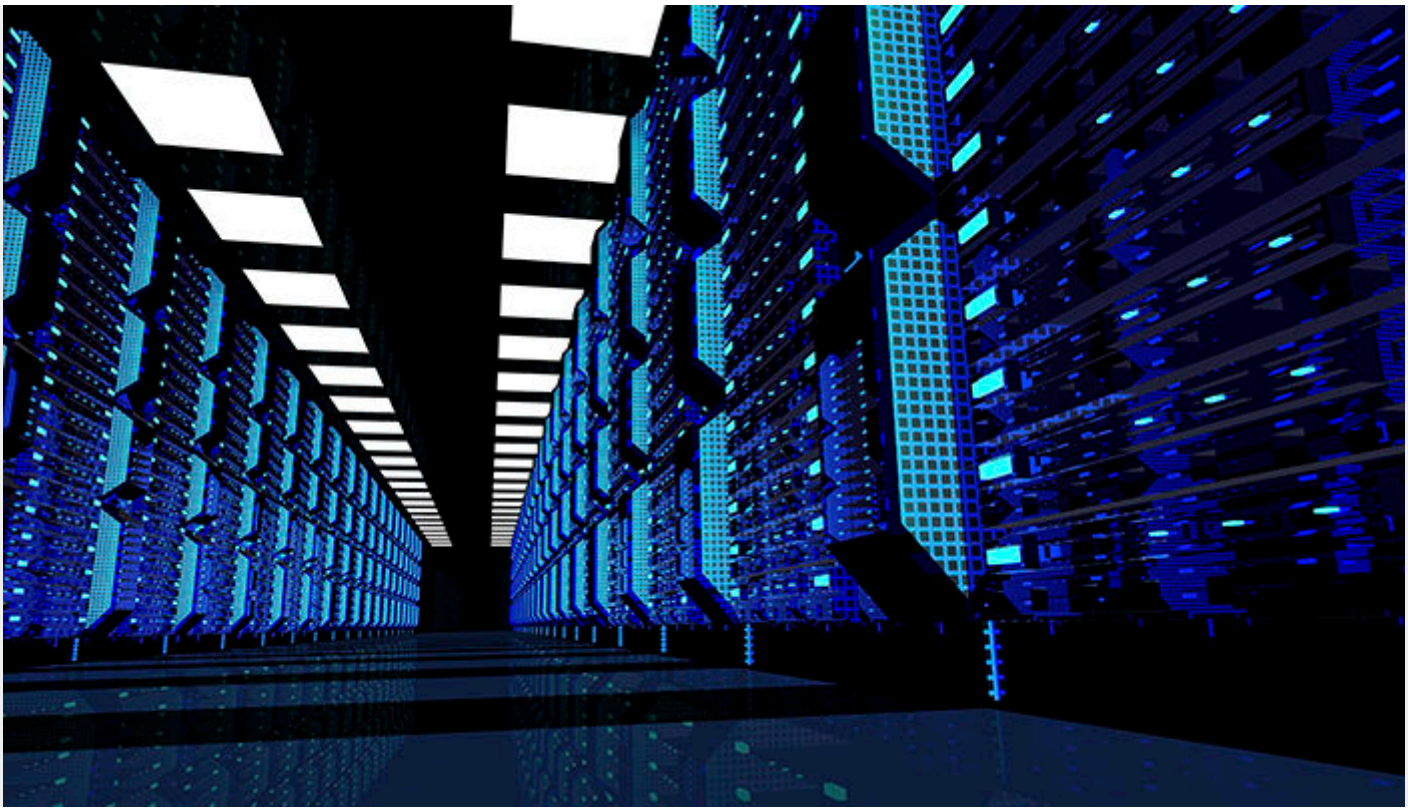
HOW GOOGLE IS USING AI FOR DATA CENTER COOLING



At the heart of all Google services, is a vast network of servers and computing resources that promise continually improving end-user experience, performance and availability. These resources are responsible for performing search queries, transferring data and delivering computing services for millions of users at any given moment, around the world. In order to maintain optimal performance of the computing infrastructure, the data center must maintain an optimal room and server hardware temperature.

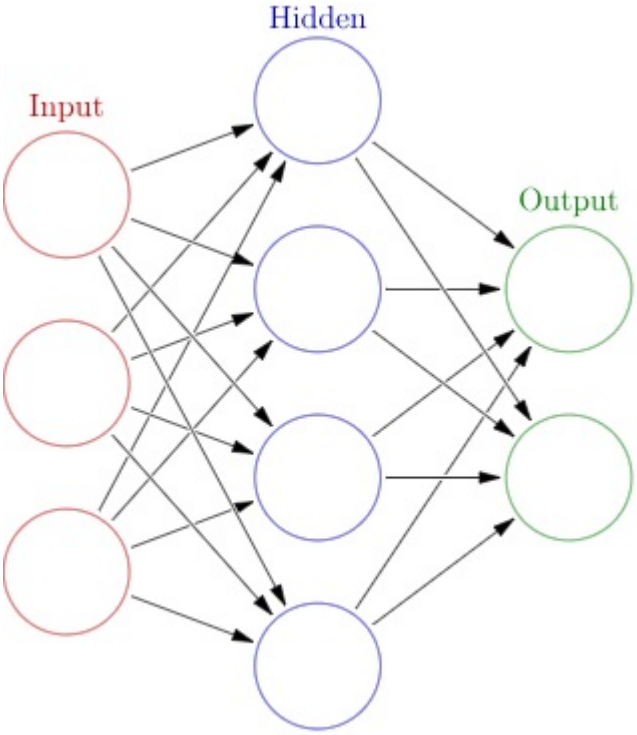
The cooling system essentially draws heat from data center equipment and its surrounding environment, and replaces it with cool air or fluids to reduce the temperature of the hardware. While improvements in cooling system technologies have allowed the Internet company to improve power efficiency, the pace of improvements has reduced in recent years. Instead of regularly reinvesting in new cooling technologies to pursue diminishing returns, the Internet giant has resorted to Artificial Intelligence to efficiently manage the cooling operations of its data center infrastructure.

Why Neural Networks?

The cooling system in a modern large-scale data center regulates several parameters in guiding the flow of heat and cooling to achieve maximum efficiency. These parameters including temperatures, cooling performance, energy consumption and cooling fluid flow characteristics, among others, are interconnected and impact the overall efficiency of the cooling system. Traditional engineering methodologies may not suffice to accurately model these complex interdependencies.

Google's implementation of AI to address this challenge involves the use of Neural Networks, a methodology that exhibits cognitive behavior to identify patterns between complex input and output parameters. For instance, a small change in ambient air temperature may require significant variations of cool air flow between server aisles, but the process may not satisfy safety and efficiency constraints among certain components. This relationship may be largely unknown, unpredictable and behave nonlinearly for any manual or human-supervised control system to identify and counteract effectively.

A Brief Primer on Google's Neural Network for Data Center Cooling:



Source: Wikipedia

Consider the input variables of the algorithm as neurons. Multiple neurons are grouped into a single layer. Multiple layers exist and neurons between successive layers are connected through a weighted computation process. Simply put, the information from a neuron in one layer is multiplied by an appropriate weight, or a mathematical calculation, before supplying it as an input to a neuron in the next layer. This interaction between multiple neurons through successive layers is performed as a final converged solution is reached at the output. The error in this solution is identified, the neuron connection weights are adjusted accordingly, and the process is repeated until an acceptable solution is reached. Through this process, the neural network learns how the system should behave in response to nonlinear changes to each input parameter.

The cooling system of Google's data center is equipped several sensors providing real-time information on various components, server workloads, power consumption and the ambient conditions. The neural network takes the instantaneous, average, total or meta-variable values from these sensors to process with the neural network. For the initial training of the neural network, Google uses arbitrary values for the input variables such that an impact of every parameter is seen in subsequent neural network algorithm.

Through each iteration of the neural network process, the cost function, or error in the neuron value is identified by comparing the actual output with the predicted values. A certain amount of tradeoff

value is also considered, to allow for a gradual convergence of the solution through successive iterations. Without this tradeoff, the neural network may fail to predict future values reliably.

The error is propagated back to the appropriate neurons of previous layers, adjusting the weight of each neuron connection between the layers to yield accurate output in future iterations. It may take thousands of iterations before these weighted connections are adjusted to yield accurate output based on true nonlinear characteristics of every neuron at the input parameter layer.

According to Google's white paper documenting the methodology, the implementation computes 2 years' worth of data to fully train the neural network. The data sets are randomized to reduce potential biasing of the neural network. The algorithm is thoroughly tested and validated against the behavior of known data sets, input parameters and output yield.

The algorithm is also used to simulate data center environments and impact of configuration changes to the physical environment. The algorithm allows data center operators to perform a sensitivity analysis on individual parameters, identifying and verifying the behavior of every component within the data center cooling system.

Results

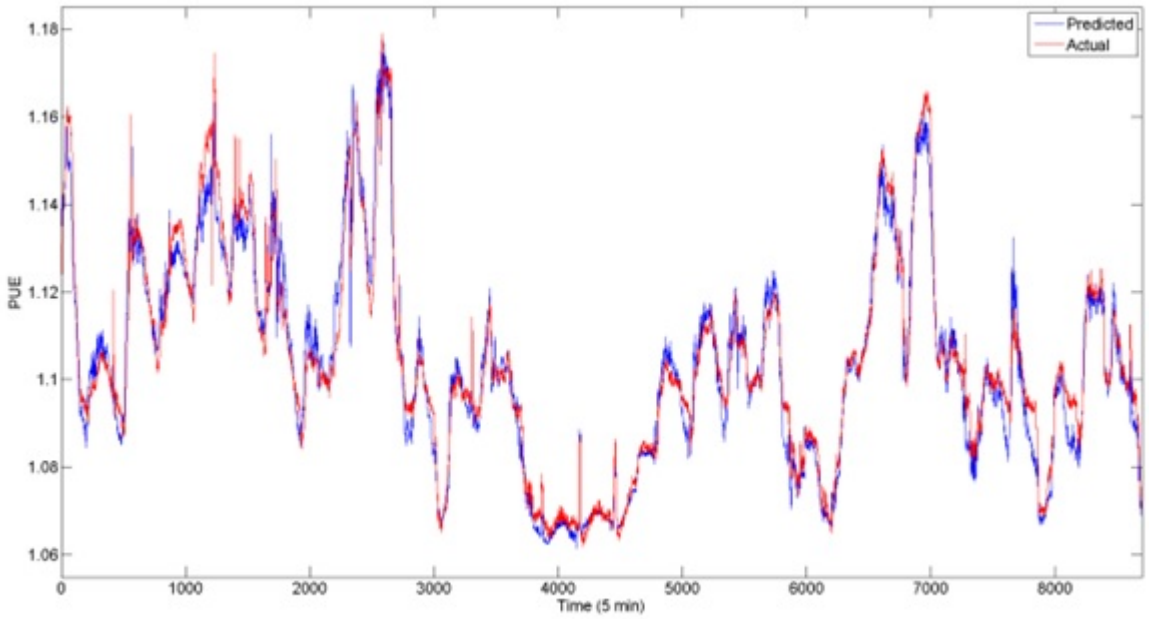


Fig. 3 Predicted vs actual PUE values at a major DC.

Source: Google

Google's implementation of the neural network reduced the error to 0.004 Power Utilization Effectiveness (PUE) or 0.34-0.37 percent of the PUE value. The error percentage is expected to further reduce as the neural network processes new data sets and validates the results against the actual system behavior. These numbers translate into a 40 percent energy savings for the data center cooling system.

The graph below demonstrates how the neural network implementation delivered PUE improvements over the years. Since these results are aggregated from multiple data centers operating under different environmental and technical constraints, the optimal implementation of the machine learning algorithm promises even better improvements in comparison with traditional control system implementations. The neural network algorithm is one of many methodologies that Internet companies including Google may have implemented for data center cooling applications.

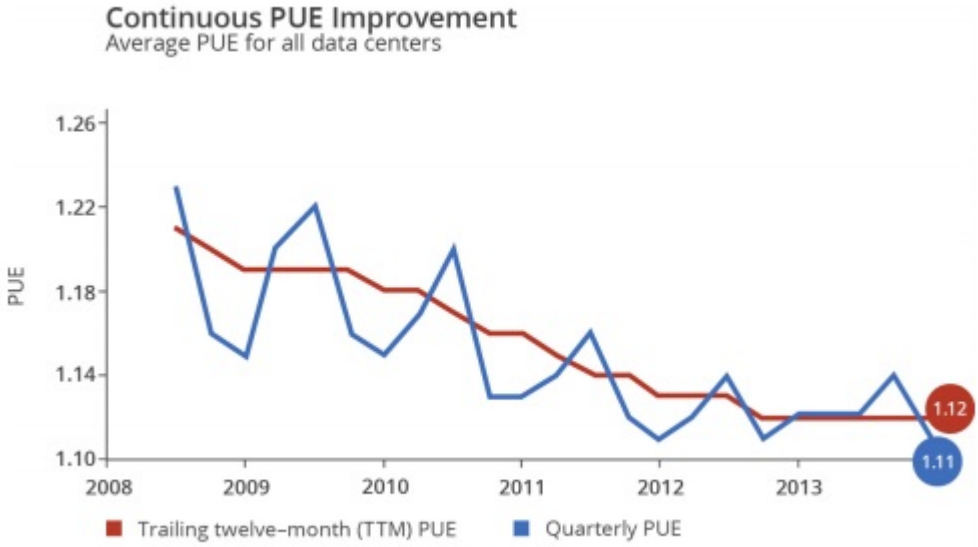


Fig 1. Historical PUE values at Google.

Source: Google

As the company edges closer to powering its entire infrastructure by 100 percent renewable power sources, artificial intelligence technologies promise unprecedented improvements in energy utilization. The energy consumption and cooling demands of Google's growing infrastructure will continue to increase as servers perform faster computing operations, store more data and handle exploding volumes of data traffic from around the world.

And while Google's research publication analyzes a performance between the period 2008-2014, the company is expected to have continued the improvement trajectory in its data center infrastructure management capabilities. This is just an example of how AI complements humans in addressing the key challenges associated with operating the most innovative and advanced data center technologies in the world. A similar behavior is well demonstrated through Google's AI implementation for diverse use cases such as autonomous vehicle projects, search, advertisement and more.