Vigilent

Turning cooling into compute

How AI optimization of data hall cooling and the chiller plant frees power for compute in power-constrained data centers

~30%
cooling power savings from air and chiller plant optimization, often in the first week

94.6 → 99.8%
thermal SLA compliance at a global financial institution

~6 MW
of power expected to be freed for compute per 100 MW of IT[1]

Executive summary

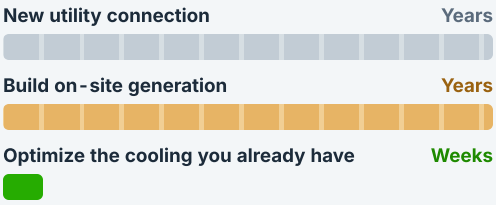
Power, not floor space, now sets the limit on data center growth. Operators have two familiar ways to get more: wait for a new utility connection, or build on-site generation behind the meter, such as fuel cells, gas engines or turbines. Both take years. There is a faster source: the power already flowing into the cooling plant.

Most of that power is spent by two systems: the air handlers that cool the data halls and the chiller plant that supplies them. Both are typically run for worst-case conditions. Fans run faster than needed, redundant units run all the time, setpoints are held cold to protect a few hot spots, and the chiller plant produces colder water than the halls actually require.

This paper shows a different approach: **closed-loop AI control that measures what every rack actually needs and runs the data hall and chiller plant as one system.** In live production, it has raised thermal compliance and cut cooling energy at the same time. The power it releases can be sold or used as IT capacity, which for a capacity-constrained operator is worth far more than the electricity saved.

> **Every megawatt the cooling plant consumes is a megawatt that can't run compute.**

THREE WAYS TO ADD CAPACITY



Software-led and non-disruptive. Works with existing BMS, air handlers and chillers from any vendor.

01 The power wall

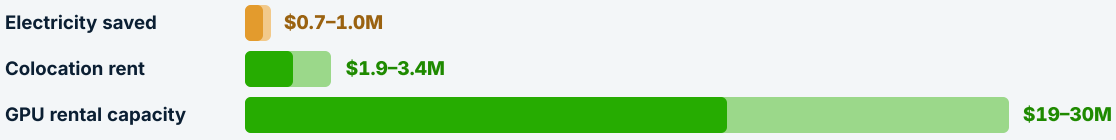
In major metros, grid congestion is limiting new supply, and new utility capacity can take years to arrive. Many operators are turning instead to on-site generation: fuel cells, gas engines or turbines installed behind the meter. That avoids depending on a new utility connection, but it requires permits, equipment, fuel infrastructure and electrical systems, and it also takes years.

For a capacity-constrained operator, the economics are stark. A freed megawatt is worth far more as sellable capacity than as an energy saving: roughly **$1.9–3.4M per year in colocation rent**, against $0.7–1.0M in electricity, and an order of magnitude more at GPU-rental rates.[2]

Cooling is the largest controllable load in the facility, and in most sites it is controlled the way it was commissioned: static setpoints, fixed fan speeds and rule-based chiller staging, with the data hall and the plant each managed by its own system.

The opportunity is to treat cooling as a **source of compute capacity** rather than a fixed overhead. It does not replace new supply, but every megawatt released from cooling is one that does not have to be bought from the utility or generated on site.

WHAT ONE FREED MEGAWATT IS WORTH PER YEAR[2]



02 Why cooling plants overcool

Overcooling is rarely a mistake. It is the rational response to limited information. Operators can't see which air handler is cooling which racks, so they run all of them. They can't predict where a hot spot will appear, so they hold supply air cold enough to protect the worst rack in the room. And the chiller plant, with no view into the data halls, produces water cold enough for the worst case.

The cost compounds. Fan power rises with roughly the cube of fan speed, so units running faster than needed waste far more energy than the extra airflow suggests. Redundant units running in parallel mix bypass air with return air, lowering return temperature and the efficiency of every coil. And every degree of unnecessary chilled-water cooling makes the chillers work harder.

COMMON SOURCES OF WASTED COOLING POWER

- **Too many units running, too much of the time,** including redundant units that add little cooling.
- **Fixed fan speeds** set for peak load, regardless of actual heat.
- **Cold setpoints** chosen to protect a handful of hot spots.
- **Bypass and recirculation** that lower return temperatures and coil efficiency.
- **Fixed chilled-water temperature** and rule-based chiller staging that ignore actual hall demand and weather.
- **Siloed controls:** the plant cannot safely raise water temperature without knowing how much reserve each hall has.

03 Optimizing the data hall

The first step is to replace assumptions with measurement. AI control of the data hall works in three steps:

1

Sense every rack

Rack-inlet temperature sensors show actual conditions across the whole room, not just at a few return-air points.

2

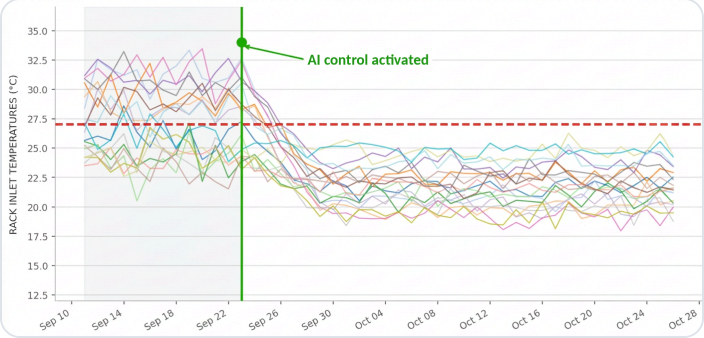
Learn the influences

Machine learning maps how each air handler affects each rack, and keeps the model current as IT load and layout change.

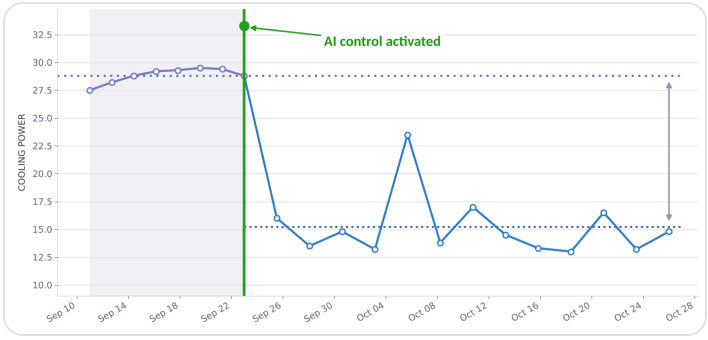
3

Act continuously

Slow fans, adjust setpoints and turn off units that aren't needed, while keeping every rack inside its thermal limit.



Rack inlet temperatures, before and after. Before activation (shaded), many racks exceed the 27°C limit (red). After activation (green line), all racks stay below it.



Cooling power, before and after. Measured cooling power drops by about half immediately after activation and stays down.

The two results come from the same control actions. Because the AI knows which racks each unit affects, it can remove cooling where it isn't needed and add it precisely where it is. **Compliance and efficiency improve together**, and every kilowatt the fans no longer draw is available to IT.

04 Proven where uptime matters most

The approach is proven at scale across more than 1,500 facilities, from global banks running trading-floor operations to telecom operators and colocation providers. Results are measured in live production and simulations:

99.8%
thermal SLA compliance, up from 94.6%, at a global financial institution with 15,000+ racks

43.6%
cooling energy and carbon savings across 500+ cooling units from multiple vendors

2.2 MW
of power expected to be freed for compute by data hall air optimization

3.9 MW
expected to be freed for compute by chiller plant optimization

The same bank cut the number of cooling units in constant operation by 60% and widened cooling ΔT from 10.2°C to 14.6°C, roughly 40% more heat removed per unit of airflow from equipment already installed.

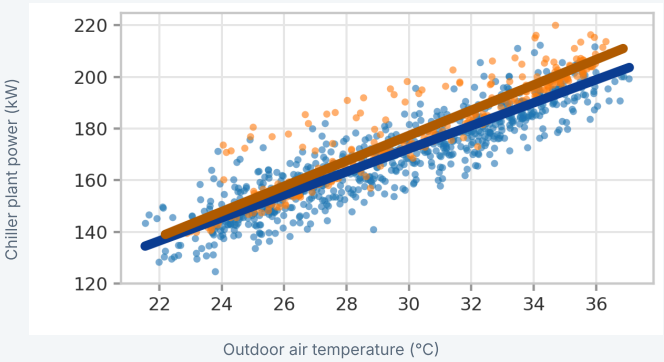
05 Optimizing the chiller plant

The chiller plant, with its chillers, pumps and cooling towers, is usually the largest single cooling load. It is also where static control leaves the most on the table, because the best operating point shifts constantly with IT load, outdoor temperature and humidity.

AI control of the plant monitors power, cooling load and weather, and continuously adjusts chilled-water temperature, chiller staging, pumps and towers to find the most efficient combination that still meets the data halls' needs. Raising chilled-water supply temperature by 5°C or more can improve chiller efficiency (COP) by over 20%.[3]

The key enabler is information from the data halls. The plant can only raise water temperature safely if it knows how much cooling reserve each hall has, and in particular the hall with the least. When the air and plant controls share that measurement, the plant can run as warm as the halls allow, and no warmer.

CHILLER PLANT POWER VS. OUTDOOR TEMPERATURE



Measured chiller plant data from one site. With AI control, plant power is lower at every outdoor temperature.

More power for IT
Every kilowatt the plant no longer draws is available to compute.

Less water
Lower cooling load means less cooling-tower evaporation.

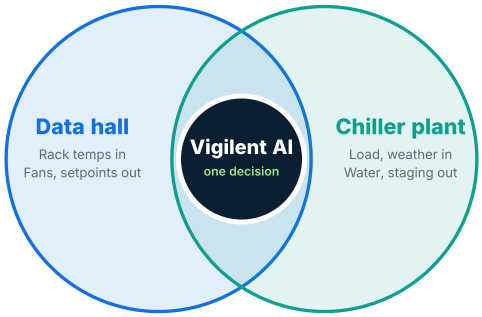
Less wear
Fewer compressor and fan cycles extend equipment life.

Deferred capex
More capacity from equipment already installed.

06 One AI for the whole cooling plant

Optimizing the data hall and the chiller plant separately tends to push problems from one to the other. Warmer chilled water saves plant energy but can erode the halls' reserve; aggressive fan savings can force the plant to work harder.

Vigilent runs both on one AI. When it changes chilled-water temperature, it knows the effect on the reserve cooling capacity of every hall, and accounts for it in the same decision. The result is efficiency, reliability and capacity improving together. As halls add liquid-cooled racks, the same approach extends to the CDUs they rely on.



The data hall and chiller plant share one AI, so each change in one is planned against its effect on the other.

07 What the freed power is worth

For a modeled 100 MW IT facility, data hall air and chiller plant optimization together are expected to free about 6 MW for compute.[1] Valued by how the operator uses it:

$4–6M a year
as electricity saved, at $0.7–1.0M per MW

$12–21M a year
as colocation capacity, at $1.9–3.4M per MW

$116–183M a year
as GPU-rental capacity, at $19–30M per MW

The data hall measurements also show where reserve cooling capacity exists, so new IT load can be placed with confidence rather than held back by conservative margins.

08 Safe by design: industrial AI for critical facilities

Handing cooling decisions to software is a reasonable concern for any operator. Vigilent's control is built on proven industrial techniques (supervised learning, model predictive control and extremum seeking), not generative AI. It optimizes within explicit limits and cannot improvise.

1. **Bounded by your rules.** Model predictive control enforces operator-defined constraints on every decision.
2. **Fail-safe and guard mode.** Unexpected conditions trigger a fall-back to operator-defined safe defaults.
3. **Premises-based.** Control and data can stay on-site, with no public-cloud dependency.
4. **Fully observable.** Operators set objectives and overrides, and can see why each decision was made.
5. **Open and vendor-neutral.** Integrates with any BMS and cooling equipment over standard protocols.
6. **Human experts included.** Regular reviews with Vigilent engineers keep results improving over time.

09 Getting started

1. **Assess**
   Vigilent engineers quantify available capacity, redundancy, overcooling and the power your plant could release, site by site.
2. **Deploy**
   Software-led and non-disruptive, using existing controls and sensors where possible. Deployment takes weeks, with value from day one.
3. **Sustain and extend**
   Regular expert reviews and health report cards keep gains compounding across the data halls and the chiller plant.

Five questions for your next capacity review

1. **How many air handlers run all the time,** and how many are actually needed to hold every rack within limits?
2. **How cold is our supply air and chilled water** compared with what IT equipment requires?
3. **Does our chiller plant know how much reserve each data hall has,** or does it run for the worst case?
4. **How many megawatts** do our fans and chiller plant draw at peak?
5. **What is a freed megawatt worth to us** as sellable capacity or GPU revenue?

About Vigilent

Vigilent invented autonomous AI control for data center cooling and has more than 20 years of AI-driven innovation and 52 patents in control and optimization. Its software runs in more than 1,500 facilities in 37 countries, serving colocation providers, financial institutions, hyperscalers and telecom operators.

Find the megawatts in your cooling plant

Talk to a Vigilent engineer.

vigilent.com

sales@vigilent.com · +1 510 524 8480 · www.vigilent.com

Notes and sources. 1. Vigilent modeling for a 100 MW IT facility: about 2.2 MW from data hall air optimization and 3.9 MW from chiller plant optimization; site results vary with climate, design, redundancy and baseline operation. 2. Value per MW per year: colocation rent $1.9–3.4M; electricity $0.7–1.0M; GPU-rental capacity $19–30M. Sources: Digital Realty 2Q26 supplemental; CBRE H1 2026 (Northern Virginia asking rates); EIA 2026 industrial rates; Vigilent estimates. 3. Vigilent analysis of chiller performance at higher chilled-water temperatures. Customer results for the financial institution and the charts on pages 2 and 3 are measured in live production; individual results vary. © 2026 Vigilent Corporation. All rights reserved.