WHITE PAPER

cadence

Maximize Your ROI by Reclaiming Data Center Lost Capacity

Effectiveness of long-term data center management solutions, maximizing efficiency via capacity recovery

By Dave King, Cadence

As the importance of data center performance, efficiency, and sustainability grows, owners and operators must move beyond “quick fixes” when managing their data center and IT deployments. These temporary solutions are neither effective nor sustainable. Long-term solutions and effective management over time are required to achieve meaningful efficiency gains. Capacity planning is one such discipline that employs predictive tools to recover lost capacity and improve data center efficiency. This whitepaper delves into the problem of lost capacity and its financial implications while offering a solution through effective data center management and predictive analysis.

Contents

Introduction

Although long-term planning is crucial, the data center industry often operates under tight deadlines and frequently prioritizes short-term fixes to meet the growing demand for new IT deployment. Short-term solutions can result in underutilization or ineffective utilization of capacity, reducing financial returns and negatively impacting the environmental sustainability of the data center. According to Gartner, existing data centers typically only utilize 50-60% of rack capacity[1], meaning significant capacity is left unutilized. Considering the huge investment in building and operating a data center coupled with the challenges of building new facilities, this value is incredibly low, wasting a significant amount of investment. Furthermore, sources have forecasted a shortage of resources to build new data centers, which are predicted to fall short of future demand[2],meaning current data centers could be forced into maximizing the utilization of pre-existing space before building new ones.

However, maintaining data center resilience and considering complex factors like airflow can make it risky to implement long-term solutions without a data center management tool. To benefit from the improved efficiency and sustainability solutions on offer, owner-operators need to carefully and scientifically assess their impact before hard-wiring them into the physical data center.

Multiple metrics, such as available space, power, and network ports, are considered when assessing data center capacity; however, data center cooling is the hardest metric to quantify and predict accurately. IT equipment cannot be deployed should any of these metrics reach their limit. Inefficient deployment of IT equipment can decrease the available capacity for future deployments and lead to capacity fragmentation over time.

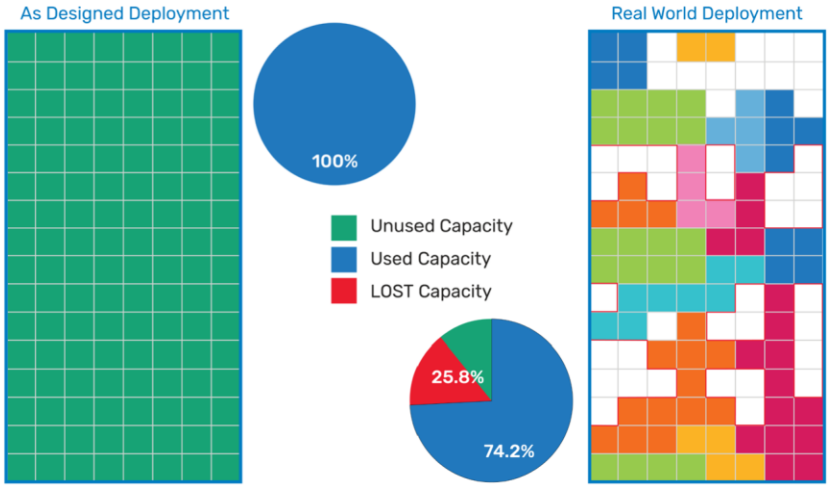

Figure 1: Figurative example of data center fragmentation using blocks

However, long-term planning coupled with effective organization and analysis of deployment from the beginning of an operational data center life cycle can prevent unanticipated restrictions in future deployment and, therefore, ensure capacity can be utilized more efficiently. A new wave of operational data centers has embraced a forward-thinking approach by implementing data center management tools that utilize computational fluid dynamics (CFD). These help translate the data centers' complex systems into a comprehensible format. Any proposed future changes can be analyzed, including cost-benefit analysis, risk assessment, and environmental performance evaluation.

This paper proposes a framework for identifying capacity loss in data centers, as well as outlining the negative financial and environmental impact of data center capacity loss. In turn, this paper introduces the implementation of a long-term capacity planning strategy that employs Cadence Reality DC, a CFD-powered solution for data center management using digital twin technology. A case study will be conducted on a real-world project to determine the potential return on investment (ROI) achieved through effective capacity utilization made possible by Cadence Reality DC.

[1] "Your Data Center is Old. Now What? - Gartner." 03 May. 2021, https://www.gartner.com/smarterwithgartner/your-data-center-is-old-now-what

[2] "Europe may struggle to build enough new datacenters – report." 25 Jul. 2023, https://www.theregister.com/2023/07/25/aggreko_datacenter_demand_europe/

Explaining the Gap in Data Center Capacity

During the data center design phase, engineers will specify the design capacity based on the commissioned facility power. This power is derived from the needs of the colocation provider, hyperscale, or enterprise customer. The design capacity is typically a single kilowatt value, which is based on the system's required power and the power to efficiently cool the data center, with an additional agreed redundancy. Another variable, "available capacity," depends on four resource components: space, power, network, and cooling.

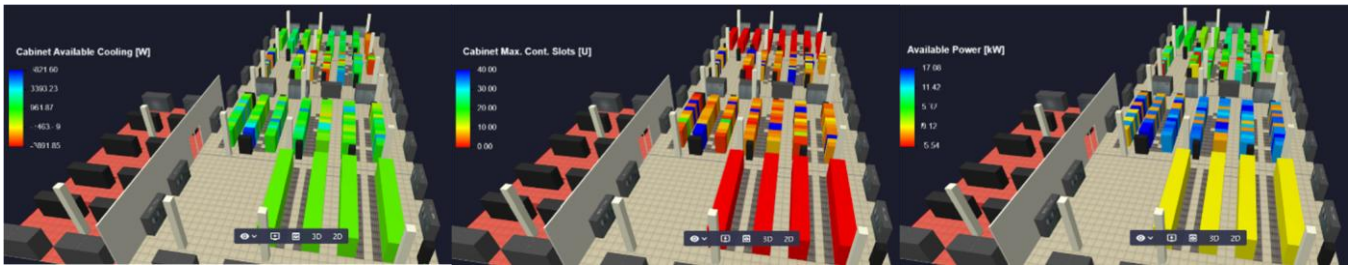

Figure 2: Image and relative results of capacity availability for a data center in operations, considering cooling, space and power, generated using Cadence Reality DC Digital Twin

The data center designer will then distribute the power to a number of cabinets with a predetermined layout while ensuring individual cabinets are at acceptable power densities. The cooling system will then be sized to remove the expected heat, keeping to the desired resilience. Finally, the power system and network will be connected to ensure distributed power with effective failure strategies are in place. Typically, this will assume a notional power per rack with limited variation except perhaps for expected low-density and high-density zones or variations due to areas being of different types, such as servers, storage, and network.

However, although this conceptual design is required to test and validate the concept, the one thing we know about the design is that, for most data centers, the conceptual design will never become reality. This is because the actual IT to be deployed is normally unknown at the design stage, and while the data center design is fixed, the operating data center undergoes many small and large changes throughout its lifecycle, notably continually changing IT deployments and cabinet layout shifts. On average, over the 15 to 20-year lifecycle of a data center, IT will be refreshed every three to four years.[3] As a result, the original design assumptions will be gradually broken, which, over time, will lead to significant differences in the distribution of power, network availability, cooling, and rack U-slot space. The Uptime Institute provides examples of data center refreshes and the effects on data center power and cooling. They also identify the trends in increasing power consumption in newer servers and how this may further increase negative effects on capacity for an operational data center.[4]

In an operational data center, if any one of the four key components is fully consumed, deploying new IT equipment becomes impossible. This may even occur before any of the resources are fully consumed because the resources do not coincide for any given location. Therefore, the available capacity in each rack becomes practically zero. For all the resource components, deployment availability can be inspected visually, except for cooling, which is invisible. Some operators intentionally or unintentionally ignore cooling availability, but this practice will likely compromise resilience and increase the risk of IT equipment overheating over time. In fact, not checking that all the resources available for your chosen location are resiliently available may result in an unseen accident waiting to happen in the case of a facility infrastructure failure.

This capacity challenge produces a cascading effect. Each deployment, if done inefficiently, will cause fragmentation and thus restrict availability for the next deployment. The individual types of resource fragmentation in space, power, cooling, and network will be discussed later in the paper.

The resulting expanding gap between designed available capacity and operational capacity is quantified as "lost" capacity. Lost capacity affects data center businesses both financially and environmentally. While almost all data centers will fail to reach their maximum intended capacity, the extent of such failure is significant.

Example: Consider a data center designed to accommodate 1 MW of equipment for 15 years. However, after just 6 years, the data center has 50% stranded capacity and cannot safely house more equipment. This means that the business can only achieve

[3] "The data center life story - DCD - DatacenterDynamics." 21 Jul. 2017, https://www.datacenterdynamics.com/en/analysis/the-data-center-life-story/

[4] "Uptime Institute Data Center Capacity Trends Survey 2022
https://datacenter.uptimeinstitute.com/rs/711-RIA-145/images/CapacityTrendsSurvey.Report.02032023.pdf

500 kW despite paying for 1,000 kW. There is an additional demand of 500 kW that cannot be met. As a result, the business will have to invest in building a new data center to meet this demand, resulting in significant capital expenditure. Essentially, the business will be paying twice for the same demand. This situation highlights the true cost of capacity loss in a data center.

Capacity planning and forecasting tools work on the assumption that all the unused capacity is available; however, this is never the case. Let's look at how Space, Power, Cooling, and Network can be fragmented and hence lost for future deployments

Space

When assessing whether equipment can be deployed in a facility based on available space, relying solely on the total U-slot space is insufficient; the amount of contiguous U-space is also a crucial factor to consider. For instance, if only a few 1U-sized slots are available in the entire data center and the required deployment for IT is 2U, deployment cannot be carried out. This extreme example highlights the importance of prioritizing contiguous space over total space.

Typically, the amount of contiguous U-space decreases over the operational lifecycle due to the "fragmentation" of U-slot space caused by the deployment and decommissioning of IT equipment.

Example: A new 4U server needs to be installed next week and the current state of the cabinet suggests that slot 28 is the first available U-slot. However, one of the 4U servers at the bottom of the cabinet is planned to be decommissioned tomorrow. By the installation date, the actual first U-slot available will be slot 10. The installer does not know this, and the plan is finalized to install the IT in slot 28, which results in the fragmentation of the remaining available U-slots, which are now divided into two sections rather than one contiguous block (see Figure 3).

Figure 3: Example of fragmentation in continuous cabinet U-slot space

Deployment plans are generally made using short-term thinking, only considering the current state of each rack without considering how it will impact available space for future deployments. We often see cabinets filled in what appears to be a haphazard manner, but this is just the result of many years of deployment/decommission cycles.

We see the same issue on the data center floor, especially in colocation data centers. When colocation providers divide and assign data center floor space for customers, capacity is easily lost. As customers leave and new customers enter, the space is further and further divided (usually into cages) until less useful space is left for each consequent deployment. Colocation providers are then tasked with moving existing customers to reclaim lost capacity.

Power

When designing a data center, the total power is divided in various ways, including via connected remote power panels that distribute power to the racks or bus bars with tap-offs. Some designs may call for higher power availability in certain areas, such as high-density or network zones. Power will be distributed according to specific areas of the data center and types of stationed equipment, such as networking, storage, and servers. Whilst the power distribution in the room will not be evenly distributed throughout the room design, it still assumes a certain uniformity at the cabinet level.

However, for operational data centers, the newly installed IT equipment will have varying power densities and sizes, and the relationship between the amount of space required and the power draw differs for various technologies. For example, rack-mounted storage systems use large numbers of low-powered disk shelves and can fill an entire cabinet without reaching the design power. In contrast, servers and blades can use up an entire cabinet power in less than half the available U-space.

This mismatch between different IT equipment leads to several issues. Cabinets containing lots of low-power equipment will run out of space before they run out of power, leaving the remaining power unavailable. Conversely, cabinets with denser processing equipment will run out of power before the space in the cabinet is used up.

Cabinet	1	2	3	4	5
Total Available U-slot Space	17U	19U	30U	21U	21U
The largest block of contiguous U-slots	8U	8U	26U	8U	10U
Available power	-1.2 kW	1.7 kW	0 kW	2kW	1.7kW
Available for the deployment of 11 U 1.4kW	Not available	Not available	Not available	Not available	Not available

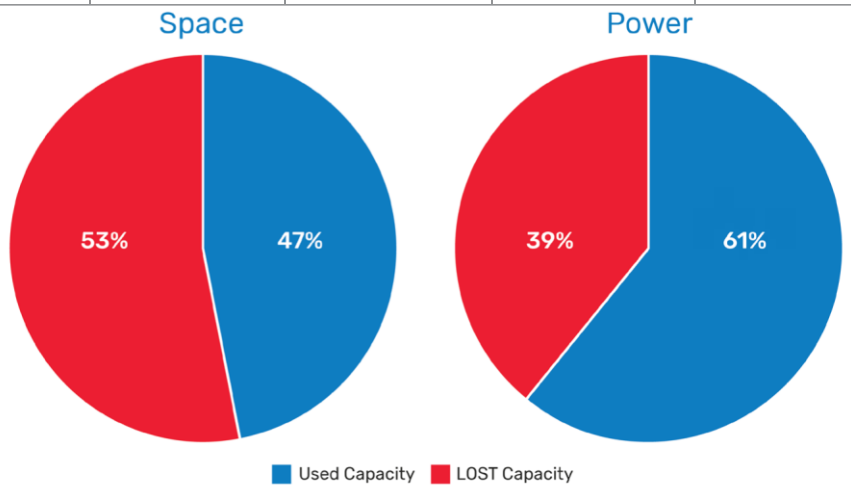

Figure 4: Example of space and power fragmentation

Example: Figure 4 shows five cabinets with differing loads.

There is a demand to deploy a new HPC cluster into these cabinets, requiring 11U of rack space and 1.4kW of power. Because of how the cluster components communicate, they must be placed in a contiguous block space.

These constraints mean that none of the cabinets can accommodate the new equipment, both due to space and power; therefore, the remaining capacity is unavailable and lost in this scenario.

Additionally, phase load balancing can cause fragmentation in available capacity. One of the facility manager's main goals is to ensure data center resilience and prioritize a balanced phase load over capacity.

Example: A 90kW PDU will have 30kW available on each of the three phases, but if the cabinets on phase one are filled only with 10kW of power, then to keep the design capacity, the 20kW remaining from phase one would have to be migrated to other cabinets or the 20kW of capacity will be lost. The facility manager will try to avoid this imbalance as much as possible, often at the expense of usable capacity.

Network

The data network design will consider many aspects, including functional areas like the Main or Horizontal Distribution Area (MDA or HDA), network topology, structured or unstructured cabling, cable types, and patching, to name a few. This design will eventually be translated to network equipment and patching, as well as allocating the ports available to certain networks or future expansions. The hardware installed will directly impact how many ports are consumed in the operational phase.

Example: Rackmount server equipment can use as many as ten copper ports for around 500W of processing power, whereas a blade chassis can provide up to 5kW of processing power with the same number of ports. If a data center is provisioned with a port density of 48 network ports and 5kW of power per cabinet, then deploying a blade chassis would utilize all the power but leave 80% of the ports unused, and the rack mount servers could use all the ports, but only 50% of the power.

Cooling

As previously discussed, cooling is the data center's most challenging and complex capacity metric. First, airflow is invisible, which makes it difficult to manage. Second, understanding where the cooling is distributed, why hot spots occur, and how to affect improvements to the air delivery is almost impossible without specific physics-based data center simulation software. Even for a "stationary" data center design, calculating cooling is difficult without the help of specially designed tools.

Due to the uniformity in space and power in the data center design, the cooling design aims to have a similar degree of uniformity, evenly distributing airflow to each piece of IT to match the distribution of the other resources. Thermal management strategies are designed to match cooling supply with IT cooling demand, ensuring all IT equipment has relatively equal temperatures (often measured as inlet temperature) necessary to operate and effectively remove exhaust heat. Essentially, the thermal management system is designed to cool IT equipment to the temperatures necessary to operate. This all assumes a notional power dissipation per rack and, importantly, notional airflow per kW.

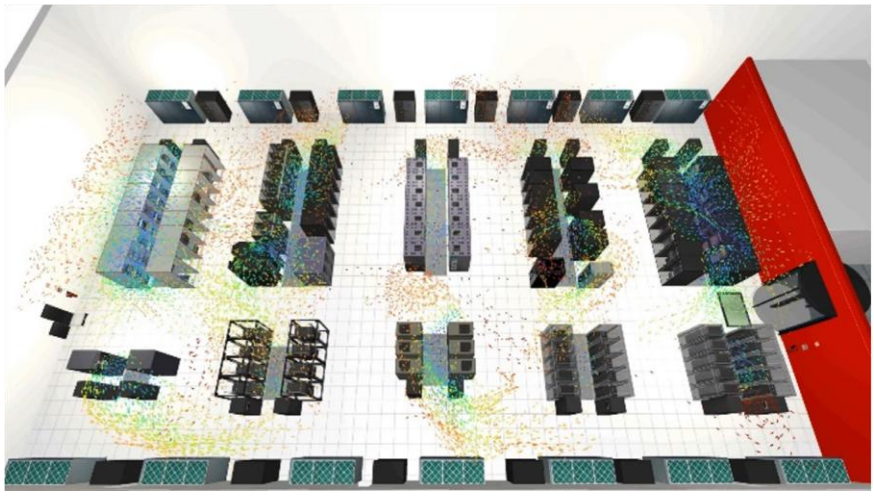

Figure 5: Cadence Reality DC Digital Twin with uneven airflow distribution shown with streamline plots of the airflow.

Assessing the cooling effectiveness of an operational data center is considerably more difficult due to its dynamic nature. Uniformity in cooling is drastically affected by minimal changes in the operational phase, down to rack level, leading to changes in the entire data center's airflow and cooling. There are many factors that cause fragmentation in the cooling uniformity throughout data center operations, including: ← The deployment of different pieces of IT equipment: Each piece of IT requires varying amounts of air to flow in different directions to cool the chips inside. This variation happens not just between different types of IT equipment but between manufacturers and even between different generations of the same model. ←

Equipment location: The ability of the cooling system to deliver the correct amount of air to the inlets of all the IT devices in the facility depends on where it is located. ← Airflow distribution to certain areas or contained rows: This can cause a mismatch between cooling required and cooling delivered. In this scenario, IT in areas with less air than required will likely overheat, while areas with excess cooling will be running inefficiently. ← Power usage: Power utilization varies throughout the day, based on user application usage, which can increase and decrease the demand for cooling. ← Poor leakage management: Cable penetrations, missing blanking plates, poor containment construction, or poor cabinet choices allow cooling to bypass IT and return to the cooling system. ← Cooling controls: Whether it is cooling unit control schemes negating the effectiveness of other cooling units, poor sensor locations, or ineffective control logic, poor control strategies can compromise available cooling.

Figure 6: Cabinet-level exhaust airflow pattern for a design cabinet (on the right) containing the same equipment with an evenly distributed load and an operational cabinet (on the left) with differing IT and power densities

Another consideration for data center design and operational capacity that is the top priority for the facility manager is resilience, or ensuring the data center can safely withstand a failure scenario. Power and data networks are quite straightforward in designing and operating at a level of redundancy, with many utilizing a standard 2N architecture for failure. However, understanding how a facility will respond to a cooling failure or maintenance schedule is considerably more difficult if one does not employ simulation to predict how the facility will respond.

The issue is compounded by the lack of accountability from both facility managers and IT departments for regulating airflow and cooling in the data center. In many organizations, the facility manager will ensure that the cooling system is adequate, but they also have no control over IT deployment. Equally, the IT teams will have control over the IT deployment, but the cooling system will be outside of their remit. Due to this contradiction, airflow management can inadvertently be neglected, with no one taking responsibility. The ultimate consequence is the inadequate planning of deployments, which causes hotspots and a loss of capacity.

Combined Capacity in Network, Power, Space, and Cooling

It's insufficient to consider the four key metrics of the data center individually as they all impact each other, making the problem far more complex. For instance, network ports or power availability can force deployment to an area that may negatively impact the cooling system's performance. Therefore, it's crucial to carefully evaluate and balance the different parts of the data center ecosystem to optimize and decipher a long-term cost-benefit analysis on each deployment.

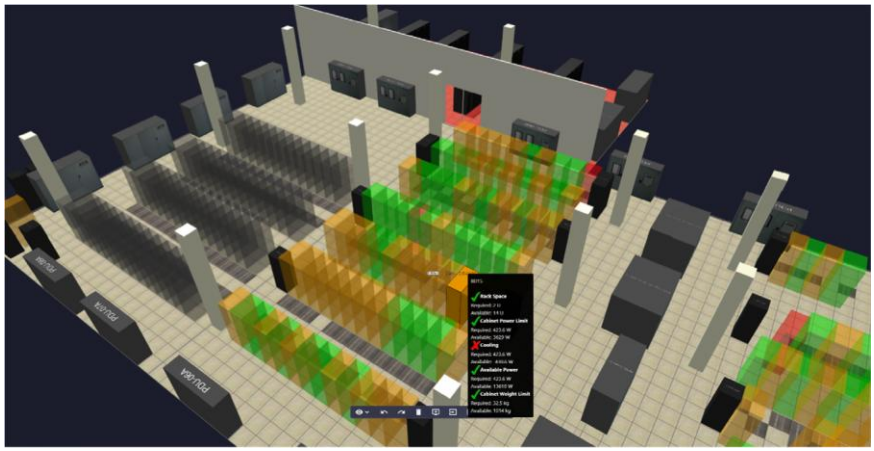

Figure 7: Cadence Reality DC Digital Twin presenting available rack space and power availability for IT deployment but insufficient cooling.

The Operational Cost to Underutilized Capacity

Businesses commission new data centers due to an increased demand for computing power. The IT facility notifies the business of the required computing power needed to meet demands. Then, a data center will be commissioned with design parameters and maximum capacity, which fulfill this required power, as discussed in more detail earlier in the paper. There are many costs to a data center, but these can be categorized into two groups: capital expenditure (CAPex) and operational expenditure (OPex).

CAPex refers to the cost of fixed assets that are essential and permanent. OPex refers to the estimated costs incurred during the operational lifecycle of the data center and is not fixed. It can also be subdivided into other groups, including:

- Fixed operational costs – maintenance costs, staffing costs, and taxes
- The load-dependent operational expenditure – utility bills required to power and cool the installed load
- Cost of turning inventory, such as IT equipment

At the commissioning stage, the OPex is often estimated to be lower than the actual OPex costs that the data center will sustain throughout the entirety of the data center's lifecycle.[5] This increase in predicted OPex can be caused by unforeseen increases in application and power requirements, which prematurely depletes capacity. Additionally, inefficient data center use will cause capacity to be used up faster than expected, which forces businesses to buy more capacity through new data centers or the cloud. This is explored in more detail in the following section with examples of how underutilized capacity can lead to wasted CAPex and increases in OPex, using a sanitized real-world case study.

Capacity Utilization Case Study

The following case study investigates a hypothetical data center commissioned to have 1.3MW and an expected lifespan of 15 years. The CAPex of this case is approximated to be $50M, calculated using the total cost of ownership (TCO) calculator offered by the Uptime Institute.[6] This calculator uses the static annuity method. Parameters such as the design capacity (1.3MW), cooling system, UPS architecture, and redundancy level are input to receive a cost summary.

The IT equipment cost can be ignored from the calculations as the processing requirements of the business will not change if the data center does not reach capacity. This case study assumes that the remaining IT equipment that cannot be deployed because of unavailable capacity will be deployed in another facility.

The expenditure will be recorded annually for the ROI calculations. As the CAPex is a fixed cost, it can be equally divided into the expected 15 years of the data center's lifespan, which equates to an annual rate of $6.3M.

The fixed operational costs, unlike CAPex, may vary over time, although they are generally out of the control of the data center owner/operator. In this case, the fixed operating costs are calculated to be $3.1M annually.

[5] "Using a Total Cost of Ownership (TCO) Model for Your Data Center." 01 Oct. 2013, https://www.datacenterknowledge.com/archives/2013/10/01/using-a-total-cost-of-ownership-tco-model-for-your-data-center/

[6] "A Simple Model for Determining True Total Cost of Ownership for Data …." https://datacenters.lbl.gov/sites/default/files/%28TUI3011B%29SimpleModelDeterminingTrueTCO.pdf

The load-dependent operational expenditure is the only real variable, and it depends on the amount of load installed in the facility. This variable is also somewhat under the control of the data center owner/operator. Inefficient use of the data center's capacity can increase the load-dependent operational expenditure significantly. More efficient use of the data center and better-utilized capacity can reduce this predicted cost. As previously discussed in the paper, the typical capacity utilization (characterized as utilized maximum load) of a data center was rounded to 50% for this example. Applying this, a 50% load will cost $0.9M per year. By contrast, theoretically, utilizing a 100% load will cost $1.4M per year, which is slightly lower than 50% because the calculation also includes the inefficiency of the plant.

Figure 8 compares the total estimated yearly cost for the example data center at both 50% and 100% utilized load. The variable cost is the load-dependent operational expenditure, and it is also the smallest cost. Therefore, this small increase in cost (approximately 5%) allows the owner/operator to retrieve 50% of prepaid capacity.

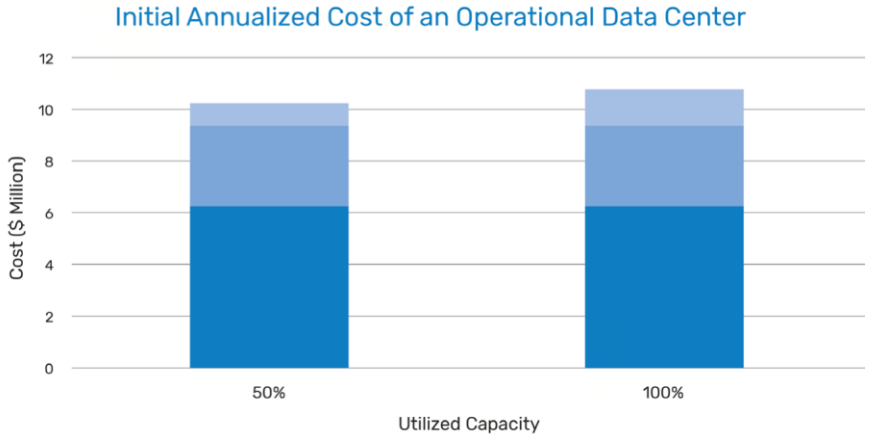

Figure 8: Annualized cost of an operational data center running at 50% and 100% capacity utilization

As mentioned earlier, the IT demand from the business remains the same. For example, with only 50% utilized capacity, there is still a requirement for another 0.65MW of processing power to equate to the full 1.3MW of the design, which means another identical data center will be built to satisfy the business need. The same problems will arise, and the second facility will also be considered full at 50%. Therefore, the total cost of 1.3MW of processing power is double the original estimate.

Figure 9 shows the total annual cost of a new data center that was built to match the load.

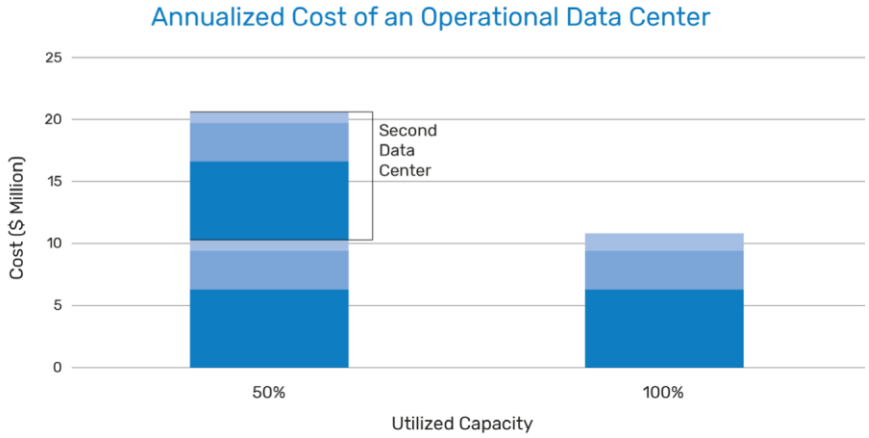

Figure 9: A: Annualized cost of an operational data center running at 50% with two data centers and a singular data center running at 100% capacity utilization

This increases the cost per kilowatt processing power ($/kW). For a typical 50% utilized data center, the cost is almost twice that of a fully utilized single data center in this example (as seen in Figure 10).

Effectively, as each data center is not fully utilized, an owner/operator pays the extra cost for the underutilization twice, which is significantly higher than the cost of a single, fully utilized facility.

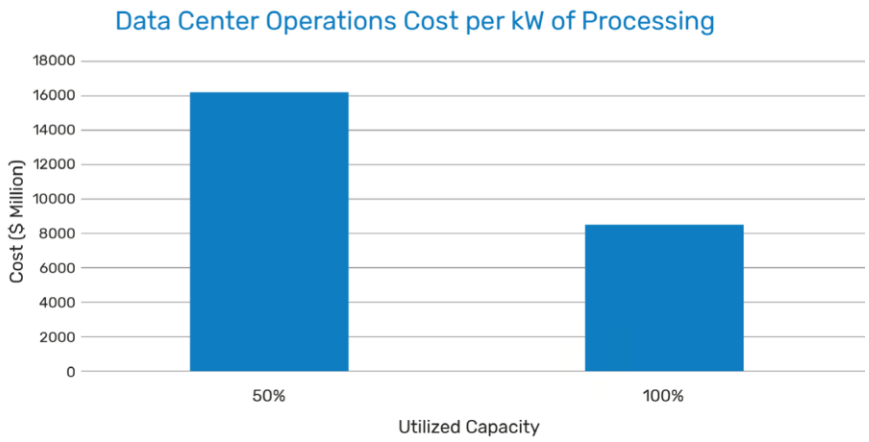

Figure 10: Data center operations cost per kW of processing for a data center operating at 50% and 100% utilized capacity

It is important to understand that no data center can reach 100% capacity utilization, but this example illustrates how much underutilizing data center capacity can really cost. It also demonstrates the ROI a data center owner/operator can achieve by better utilizing data center capacity.

If a data center recovers stranded capacity, the costs associated with utility, etc., are negated, as it is already utilizing existing cooling. The only cost to incur comes from the changes made to recover the stranded capacity, such as the cost of installing new equipment, any increase in power only for cooling infrastructure, and manual labor.

How to Avoid Lost Capacity

Completely avoiding lost capacity is impossible for any data center, but it can be significantly minimized. To avoid lost capacity, every change made to the data center and deployment implemented needs to be effectively evaluated with the long-term consequences to capacity taken into consideration. This is especially true when optimizing deployment and analyzing the fragmentation of space, power, network, and cooling.

There is a solution to the capacity problem. Cadence Reality DC Digital Twin combines modern data center management systems and simulates complex operational factors, such as cooling and airflow, using computational fluid dynamics (CFD). This tool evaluates current data center capacity for power, space, network, and cooling while showing insights into future deployment. Users can then choose more effective deployments for capacity utilization while ensuring data center resilience.

Losing some capacity cannot be avoided, but using a tool such as Cadence Reality DC Digital Twin can effectively understand each deployment's consequences. Capacity can be maximized, which will lead to both financial savings for the business and more efficient data centers for the environment. Finally, a proper cost-benefit analysis can be conducted, which helps forecast the costs for the next data center build.

Conclusion

In an ideal world, the data center's four components of capacity would reach 100% utilization simultaneously. The way to achieve this is to design the facility with a specific plan and barely deviate. However, in today's mission-critical facility, this is rarely an option. Effectively keeping pace with technological improvements means that business plans will continually evolve. To ensure that this evolution does not sacrifice capacity and greatly reduce the return on the investment of the facility, each deployment needs to be assessed before implementation. Physics-based simulation techniques present a scientific way for data center operators to accurately gain this foresight into the current and future state of their facility.

It is important to note that this whitepaper focuses on capacity utilization in the enterprise sector; however, these methods can also be used similarly for other data center owners/operators, like colocation providers.

cadence®

Cadence is a pivotal leader in electronic systems design and computational expertise, using its Intelligent System Design strategy to turn design concepts into reality. Cadence customers are the world's most creative and innovative companies, delivering extraordinary electronic products from chips to boards to complete systems for the most dynamic applications. **www.cadence.com**

For more Information, please contact

ALPHA-NUMERICS
cadence® CHANNEL PARTNER

ALPHA-Numerics GmbH
Römerstraße 32
56355 Nastätten
Germany

info@alpha-numerics.de
+49 6772 969 3430
www.alpha-numerics.gmbh

As a pioneer in the introduction of digital twins of electronic devices in data centres, ALPHA-Numerics offers powerful software solutions for designers and operators. This software, developed by the American company Cadence Design Systems, is based on virtual 3D models that calculate the prediction of cooling air paths and thermal conditions (CFD technology) and can display them in a 3D model.

Digital twins of data centres open up new perspectives that go beyond the classic functionalities of DCIM software for the daily management of computer or telecommunications rooms. They are valuable allies in reducing their ecological footprint, a major challenge in the face of climate change. They enable the identification of sustainable solutions for the use of AI (artificial intelligence) or edge computing. At the electronic equipment level, our solutions enable designers to design and validate innovative ways of cooling ever more compact and highly heat-emitting systems.