# Introduction

This document aims to explore the Total Cost of Ownership (TCO) savings achievable using NeuroBlade's SQL Processing Unit (SPU) to accelerate data analytics workloads. By offloading critical processing tasks to specialized hardware, the SPU achieves remarkable acceleration, particularly for complex and resource-intensive queries, leading to significant cost savings in the data center. The paper provides a high-level overview of potential TCO savings, emphasizing that NeuroBlade can offer a more detailed analysis tailored to specific customer environments. This detailed analysis can include server configuration, test setup, cluster size, and other relevant costs such as power, cooling, space, and maintenance.

## Background

Data centers are integral to modern computing infrastructure, but they come with significant costs, including server equipment, power, cooling, and space. With the growth of AI/ML workloads and the exponential growth of data, there has been a corresponding increase in power consumption for data analytics workloads. This surge drives the urgent need to reduce costs while improving computing power, highlighting the importance of efficient data center operations.

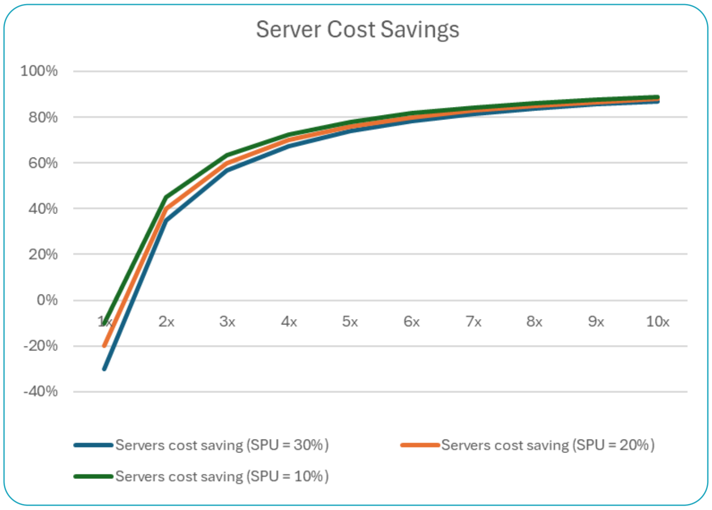
To address this, the industry is moving towards solutions that offer higher performance per watt. Technologies like the NeuroBlade SPU, which can accelerate workloads while consuming less power, are becoming increasingly valuable. Reducing operational costs while enhancing computing capabilities is a primary concern for data center operators. This need drives the adoption of innovative solutions that offer significant TCO savings. As described in the following sections, integrating NeuroBlade's SPU can lead to higher efficiency and substantial cost reductions in several key areas.

## Server Equipment Cost Reduction

When considering TCO savings, one of the primary factors is the reduction in the number of servers required to perform the same workload. The NeuroBlade SPU can accelerate data analytics workloads to such an extent that multiple unaccelerated servers can be replaced by a single accelerated server.

The acceleration ratio (N:1) denotes that N unaccelerated servers can be replaced by one accelerated server. For instance, an acceleration ratio of 3:1 means three unaccelerated servers can be replaced by one accelerated server. However, the SPU introduces an additional cost of approximately 20%, thus the effective ratio becomes N:1.2, meaning that N servers can be replaced by 1.2 servers.

Using N=3 as an example, the effective reduction in server equipment is represented by the ratio 3:1.2. This implies a server equipment savings of (3-1.2)/3 = 1.8/3 = 60%. The figure below shows the server equipment cost savings as a function of the acceleration speed (N=1,2, …10) and includes three different graphs showing the impact of the SPU overhead on the total server cost. As shown, an acceleration of 2x to 5x achieves the best impact value in server cost efficiency. Additionally, at an acceleration speedup of less than 1.3x, the efficiency improvement is minimal.

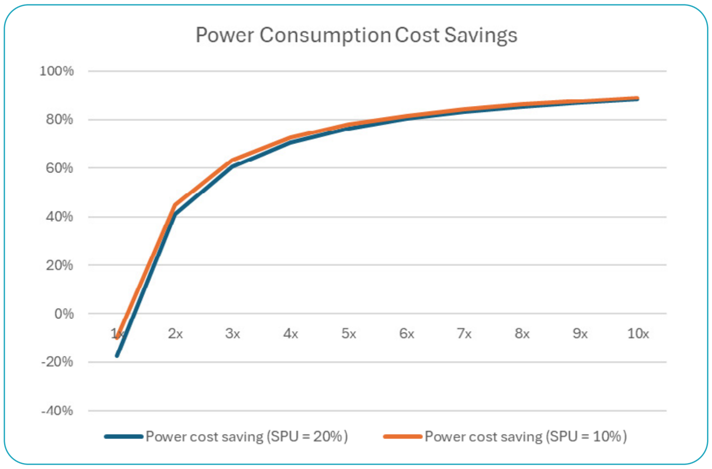


## Power Consumption Cost Savings

Power consumption costs represent a significant part of the TCO in data centers. This is a limiting factor for the growth in compute power since data centers are challenged to grow the power supply to meet the compute demand. Therefore, performing the same task with fewer compute resources not only saves equipment costs but also unlocks compute capacity growth in data centers.

To calculate the power cost savings, we can use a similar approach as we used for the server cost based on the acceleration value. Assuming the SPU adds a power overhead to a server with a baseline power consumption, this translates to a small overhead but on a smaller cluster size (a smaller number of accelerated servers can achieve the same task as a larger number of unaccelerated servers).

For example, with an unaccelerated server's power consumption of 750W and an SPU power consumption of up to 150W, the overhead on the accelerated server is 20% of the power usage. As shown in the figure below, despite the additional power consumption of the SPU, the reduced number of servers needed to complete the same task results in a significant reduction in overall power consumption for the server cluster.



The calculation follows a similar approach as the previous section. For an acceleration ratio of N:1 (N unaccelerated servers can be replaced by one accelerated server), the effective power consumption acceleration for a server with a 20% SPU power overhead is N:1.2. The figure below shows the power cost saving as a function of the acceleration speed (N=1, 2, …10) for SPU power consumption of 10% and 20%.

## Data Center Cooling Costs

Cooling costs are another significant component of the TCO in data centers. Effective cooling is essential to maintain optimal operating temperatures for servers and other equipment, preventing overheating and ensuring reliable performance. As data centers scale the size and capacity, the cooling requirements and associated costs also increase.

By integrating NeuroBlade's SPU, data centers can achieve substantial cooling cost savings. The primary reason for this is the reduction in the number of servers needed to perform the same workload. Fewer servers generate less heat, which in turn reduces the burden on the cooling systems.

To calculate cooling cost savings, consider the power consumption reduction achieved by using accelerated servers. The amount of heat generated by the servers is directly proportional to their power consumption. Therefore, reducing power consumption also reduces the heat output, which translates to lower cooling requirements.

For instance, if an unaccelerated server consumes 750W and an accelerated server with an SPU consumes 900W (750W + 150W overhead), but you need fewer servers due to the acceleration, the overall heat generated is less. Suppose you replace three unaccelerated servers with one accelerated server; the power consumption changes from 2250W (3 x 750W) to 900W, resulting in a significant reduction in heat output.

Cooling systems typically consume additional power to remove the generated heat. If the cooling system requires 0.2 of power to dissipated server power, the cooling power needed for unaccelerated servers would be 450W (2250W * 0.2). For the accelerated server, the cooling power required would be 180W (900W * 0.2). This results in a cooling power saving of 270W (450W - 180W) or 60% (270/450).

## Other Cost Savings

In addition to savings on server equipment, power consumption, and cooling, integrating NeuroBlade's SPU into data centers can lead to significant cost reductions in other areas. These include space, maintenance, and management costs, all of which contribute to the overall TCO.

### Other Equipment Savings

Reducing the number of servers also means fewer associated equipment needs, further contributing to cost savings. These include costs for Top-of-Rack switches, cabling, racks, and rack mounts, significantly reducing the TCO of the data center. The savings highly depend on the cluster size and the number of servers per rack. For example, assuming approximately 20 servers per rack and an equipment cost of $5,000 per rack, with 30,000 servers, the expected savings are $7.5 million ($5,000 * 30,000 / 20) over the cluster's lifetime, which translates to approximately $1.8 million per year.

### Space Savings

Reducing the number of servers required to perform the same workload directly impacts the physical space needed in a data center. Fewer servers mean less rack space, which reduces the data center's physical footprint, lowering real estate costs and enabling quicker data center expansion. With each rack taking about 2.5 sq ft and each sq ft costing about $100-500, saving N:1.2 rack space leads to significant TCO savings, especially in clusters with several thousand nodes.

### Improved Airflow

With fewer servers generating heat, there is better airflow, which can improve cooling efficiency and further reduce cooling costs. This parameter is not included in our model since it highly depends on the specific data center characteristics and is thus hard to quantify for a generic case.

### Maintenance and Management Savings

With fewer servers to manage, the costs associated with hardware maintenance, including replacement parts and labor, are reduced. Additionally, improved reliability and performance of the accelerated servers can lead to less downtime, which is critical for maintaining business continuity. As a result, IT staff can manage the infrastructure more efficiently, allowing them to focus on strategic initiatives rather than routine maintenance.

Maintenance and management savings also depend on the specific data center characteristics, making them difficult to quantify for a generic case. Consequently, these savings are not included in our generic model.

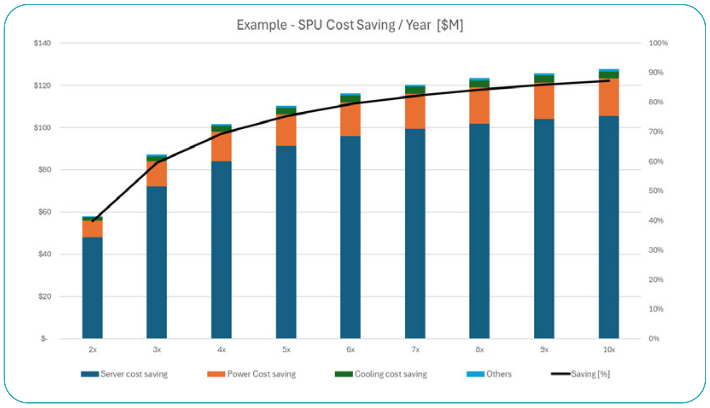
## Quantitative Example for SPU Cost Saving

To illustrate the cost-saving potential of NeuroBlade's SPU, let's consider a hypothetical data center scenario with the following expenditures:

- **Total Yearly Server Spend:** $120,000,000 (for ~6000 new nodes)
- **Total Yearly Power Spend:** $20,000,000 (for a cluster size of 25,000-30,000 nodes)
- **SPU Cost Overhead on Server Cost:** 20%
- **Server Power:** 750W
- **SPU Power:** 150W

The figure below shows the estimated per-year cost-saving breakdown for different SPU acceleration speed-up values (1x - 10x). As shown in this example, the yearly cost saving can range from $50 million per year to over $120 million per year for a reasonably sized cluster deployment.

- **Server Cost Saving:** As the SPU acceleration speed increases, the number of required servers decreases proportionally, leading to significant savings in server expenditures. For instance, at a 3x acceleration, the server cost saving is approximately $72 million.
- **Power Cost Saving:** Similarly, fewer servers result in lower power consumption, leading to substantial power cost savings. For example, at a 3x acceleration, the power cost saving is around $15 million.
- **Total Cost Saving:** Combining the server and power cost savings provides the total yearly cost-saving potential. At a 3x acceleration, the total cost saving is approximately $87 million per year.



This quantitative example demonstrates the substantial cost-saving potential of NeuroBlade's SPU, highlighting how data centers can achieve significant TCO reductions while enhancing performance and efficiency. The exact savings will vary depending on the specific characteristics and scale of the data center deployment.

## Summary

This white paper has explored the substantial Total Cost of Ownership (TCO) savings achievable through the integration of NeuroBlade's SQL Processing Unit (SPU) in data centers. By offloading critical processing tasks to specialized hardware, the SPU provides remarkable acceleration for data analytics workloads, particularly for complex and resource-intensive queries.

### Key areas of cost savings include:

- **Server Equipment Costs:** The acceleration capabilities of the SPU reduce the number of servers needed to perform the same workload, leading to significant savings in server expenditures.
- **Power Consumption Costs:** Fewer servers result in lower power consumption, which translates to substantial savings in power costs.
- **Cooling Costs:** The reduction in the number of servers and the associated decrease in heat generation improve cooling efficiency, further reducing cooling costs.
- **Other Costs:** Fewer servers mean reduced physical space requirements, lower maintenance needs, and more efficient management, allowing IT staff to focus on strategic initiatives.

Overall, the integration of NeuroBlade's SPU offers comprehensive TCO reductions, enhancing the performance and efficiency of data center operations. By adopting this advanced technology, data centers can meet the growing demands of data analytics workloads while achieving significant cost savings. This white paper provides a high-level overview, but more detailed analysis tailored to specific customer environments can yield even more accurate and substantial savings, considering server configurations, test setups, cluster sizes, and other relevant factors.performance and efficiency. The exact savings will vary depending on the specific characteristics and scale of the data center deployment.

# About NeuroBlade

NeuroBlade is unlocking data analytics by introducing its SQL Processing Unit (SPU) accelerator. This innovative technology significantly enhances query processing speed and scalability, offering up to a 100-fold improvement in performance-per-cost for Data Analytics workloads. Emphasizing a Compute Made For Analytics approach, NeuroBlade's advanced processor design optimizes throughput for petabyte-scale data, enabling queries to run exponentially faster.

Founded in 2018 by veterans of the systems, storage, and data analytics industries, NeuroBlade is headquartered in Tel Aviv, Israel, and has an office in Palo Alto, California, with operations in Taipei, Taiwan. **For more information, visit www.neuroblade.com.**