Maximizing price for performance on the Cloud:

How to get the best value from accelerator backed instances on the Cloud

Ashok Ramachandran
Brian Ray
Brandon Wong

Office of the CTO- Eviden Cloud

Contents

Introduction and executive summary

The article examines the rising costs and challenges of using accelerator-backed instances like GPUs and ASICs in cloud computing for HPC and AI applications. It highlights the importance for IT decision-makers in balancing cost and performance in cloud platforms such as AWS, GCP, and Azure. The focus is on achieving optimal price-to-performance ratios without compromising quality and reliability. It suggests utilizing AI, FinOps, and expert services to optimize cloud costs and performance effectively.

Cloud computing has enabled businesses to innovate and scale faster than ever. However, as the demand for high-performance computing (HPC) and artificial intelligence (AI) applications grows, so does the cost of cloud resources. Accelerator backed instances, such as GPUs and ASICs, offer significant speed and efficiency advantages over traditional CPUs, but they also come with a hefty price tag.

As organizations increasingly migrate to the cloud, the quest for optimal price-to-performance ratios has become a critical consideration. Here, we discuss the escalating costs associated with accelerator-backed instances for GPUs and ASICs, delving into the challenges across major cloud providers like AWS, GCP, and Azure while focusing on the pivotal role played by IT Decision-Makers such as CTOs, CIOs, and IT managers responsible for making decisions about cloud services, infrastructure, and expenditures.

So, how can you find the best price for performance on the cloud, without compromising on quality, performance, and reliability? In this article, we will explore the challenges and opportunities of using accelerator backed instances on the cloud, and how IT Decision-Makers within the customer's organization can use AI, FinOps, and Eviden's expertise to optimize your cloud cost and performance.

VMs - CPUs	Accelerator Backed Instances - GPUs
Central processing unit	Graphic processing unit
4 to 8 cores	100s / 100s of cores
Low Latency	High Throughput
Best for Serial Processing	Best for Parallel Processing
Interactive Process Task Executing	Distributed Process Parallel Task Execution
CU Sequential Programming Execution	SW based CPU to GPU Functional Transformation

Table 1: CPU Vs. GPU Features Comparison

With the combined leverage of AI and FinOps, IT Decision-Makers could save up to 40% on their cloud spending, while maintaining or improving the performance and quality of their AI models and services.

The top 5 challenges faced by IT decision-makers

As an IT decision-maker, you are responsible for making decisions about cloud services, infrastructure, and expenditures, that can have a significant impact on your organization's performance, competitiveness, and innovation. However, when it comes to accelerator backed instances, you may face some of the following challenges:

- **Visibility and transparency** into the cloud costs and performance of accelerator backed instances, across different cloud providers, regions, and availability zones.
- **Control and flexibility** over the cloud resources and configurations, that can affect the performance and the cost of accelerator backed instances, such as the type, number, and generation of the devices, the operating system, the network bandwidth, and the storage capacity.
- **Alignment and collaboration** between the financial and the operational teams, that can lead to inefficiencies, waste, and overspending on the cloud resources and services.
- **Expertise and guidance** on how to optimize the cloud spending and performance for accelerator backed instances, and how to leverage the best optimization techniques and tools, such as resizing, scaling, scheduling, spot instances, reserved instances, and savings plans.
- **Awareness and understanding** of the latest trends and developments in the cloud computing landscape, and how to take advantage of the new and more powerful devices, such as the NVIDIA A100 Tensor Core GPU, the AWS Inferentia ASIC, and the Google Cloud TPU.

Why are costs increasing for accelerator-backed instances for GPUs and ASICs when compared with new generation virtual machines, across AWS, GCP, and Azure clouds?

Accelerator backed instances, powered by **GPUs** and **ASICs**, are pivotal for high-performance computing tasks. However, the costs associated with these instances have witnessed a surge when compared to new-generation virtual machines. There are several underlying factors contributing to this trend across major cloud platforms, and into the evolving cloud pricing dynamics.

Though GPU computing was once primarily associated with gaming and graphical rendering, it has grown into the main driving force of high-performance computing in many different scientific and engineering fields. Accelerator backed instances are specialized cloud resources that are designed to handle compute-intensive tasks, such as **machine learning, deep learning, computer vision and natural language processing**. They use dedicated hardware devices, such as GPUs and ASICs, to accelerate the processing of data and algorithms, resulting in faster and more accurate outcomes. GPUs and ASICs are also more energy-efficient than CPUs, which means they can reduce the **carbon footprint** of cloud computing.

However, they are also more expensive than regular VMs, which use CPUs as their primary processing units. The cost of accelerator backed instances depends on several factors, such as the type, number, and generation of the devices, the cloud provider, the region, the availability zone, the operating system, the network bandwidth, the storage capacity, and the duration of the usage.

For example, according to the latest pricing information from AWS, GCP and Azure, a single GPU instance can cost anywhere from $0.7 to $24.48 per hour, depending on the configuration and the provider. In contrast, a single CPU instance can cost as low as $0.01 per hour, depending on the configuration and the provider.

Moreover, the cost of accelerator backed instances is increasing over time, as the cloud providers introduce new and more powerful devices, such as the **NVIDIA A100 Tensor Core GPU, the AWS Inferentia ASIC, and the Google Cloud TPU.** In addition, the surge in popularity of Gen AI based services have also led to the increase in need of computing resources and related costs. These devices offer superior performance and features, such as higher memory bandwidth, larger memory capacity, faster interconnects, and support for mixed-precision arithmetic. However, they also come with a higher price tag, which can make them inaccessible or unaffordable for some business cases.

Therefore, utilizing a repeatable frame can help us find the best price for performance on the cloud for accelerator backed instances is a challenging and complex task, that requires careful analysis and comparison of the available options, as well as the understanding of the specific requirements and objectives of the customer's business.

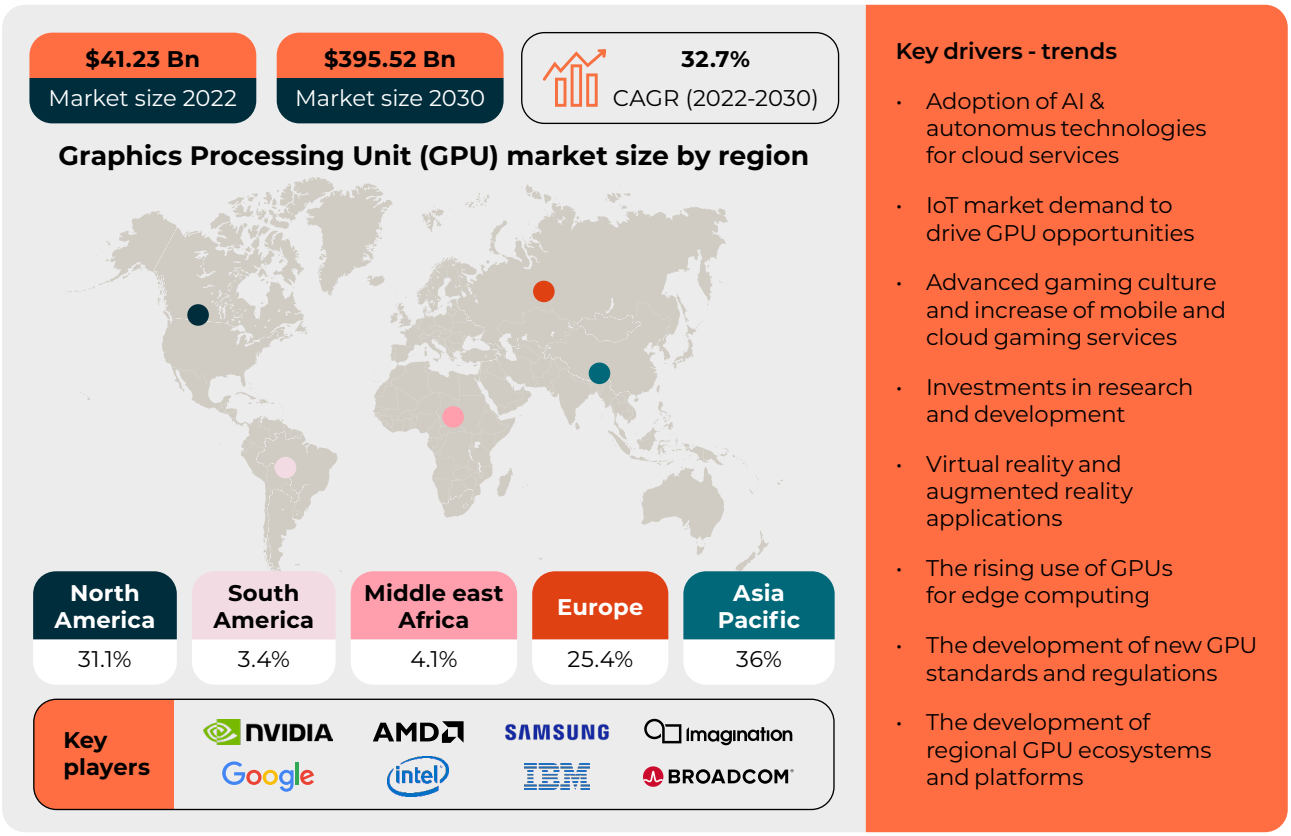


Figure 1: GPU Market Size, Players, Growth and Key Drivers and Trends
(Sources: Insight Partners, Gartner, Mordor Intelligence, GVR)

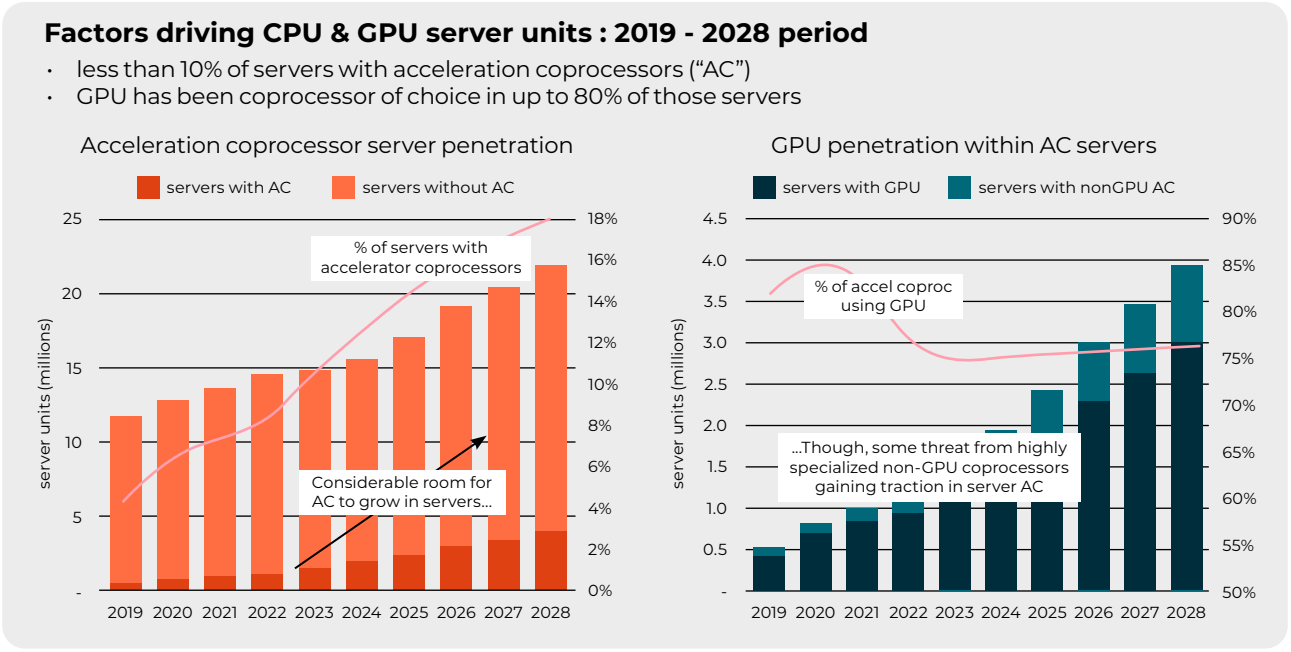


Figure 2: Factors driving CPU and GPU Server Units: 2019 – 2028 (Source: Yole Intelligence)

The solution

Selecting the right cloud platform, with the appropriate resource combinations to achieve the right cost-to-performance balance

The solution to finding the best price for performance on the cloud for accelerator backed instances is to use a combination of AI, FinOps, and Eviden's expertise. Below are the key elements of the solution:

AI: AI emerges as a powerful ally in the pursuit of optimal cloud resource utilization. By leveraging AI algorithms, IT Decision-Makers can analyze vast datasets to identify the ideal cloud platform and resource combinations that strike the right balance between cost and performance. The application of AI in decision-making processes and its role in optimizing cloud infrastructure is key to achieving the right balance.

AI can help choose the right cloud platform, with the appropriate resource combinations, by using data-driven and automated methods, such as **benchmarking, modeling, optimization, and recommendation**. AI can help evaluate, predict, simulate, and select the optimal or near-optimal cloud resources, that satisfy the performance, the cost objectives, and constraints.

FinOps: FinOps is a set of best practices and principles, that aim to align the financial and the operational aspects of cloud computing and enable IT Decision-Makers to manage and optimize their cloud spending and performance, in a collaborative and agile way. FinOps can help achieve **visibility, accountability, and optimization** for the cloud resources, by collecting and aggregating data, allocating and distributing costs, applying optimization techniques and tools, and establishing roles and responsibilities.

Eviden's expertise: This is the knowledge and experience of expert teams who specialize in cloud cost optimization and performance management and help IT Decision-Makers navigate the complex landscape of cloud computing. Eviden's expertise can help leverage the best technologies and methodologies, and provide cutting-edge and state-of-the-art solutions and services, that deliver value and satisfaction.

Eviden can partner with IT Decision-Makers in different ways, which aims to help them achieve their cloud goals, but differ in the level of involvement and responsibility of the GSI and the customer:

- In a hybrid model, where Eviden and the customer share the responsibility of managing the cloud infrastructure. Here, Eviden performs the benchmarking, consulting, and training services, as well as the ongoing monitoring, maintenance, and optimization of the cloud infrastructure. The customer provides the requirements, feedback, and approval, as well as the access, resources, and collaboration. This is suitable for customers who want to leverage the expertise and experience of Eviden, while still maintaining some control and visibility over their cloud infrastructure.

- In an Empowerment model, where Eviden takes the initial role and the customer taking the final role. Eviden performs the benchmarking, consulting, and training services, as well as the initial setup and configuration of the cloud infrastructure. Through a knowledge sharing process, customer takes over the ongoing monitoring, maintenance, and optimization of the cloud infrastructure. This is suitable for customers who want to develop their own capabilities and expertise in cloud computing, while still benefiting from the guidance and support of Eviden.

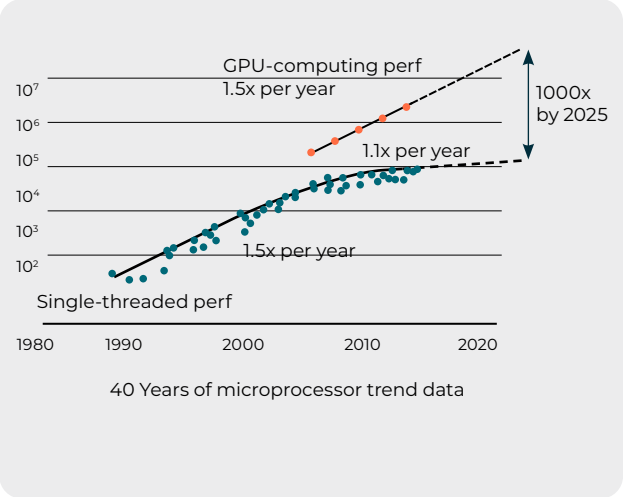


Figure 3 (b): GPU vs. CPU Speed Comparison when handling these large computations (Source: OrboGraph)

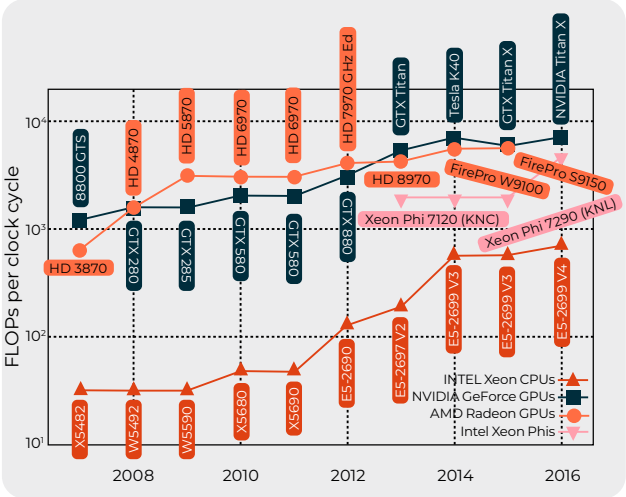


Figure 3 (b): GPU vs. CPU Speed Comparison when handling these large computations (Source: OrboGraph)

Eviden's perspective

Eviden, as a leading GSI, provides its unique insights into navigating the intricate landscape of price-to-performance optimization on the cloud. This section outlines Eviden's approach, methodologies, and tools that empower organizations to make informed decisions, ensuring maximum value from their cloud investments.

Eviden's vision is to become the leading and **trusted** partner for customers and IT Decision-Makers, who want to leverage accelerator backed instances for their HPC and AI applications, on the cloud.

Eviden's mission is to help customers and IT Decision-Makers find the best price for performance on the cloud, without compromising on **quality and reliability**.

At Eviden, we can help IT Decision-Makers choose the right cloud platform, with the appropriate resource combinations, by using data-driven and automated methods, such as **benchmarking, modeling, optimization, and recommendation**.

At Eviden, we leverage our:

- **AI-powered benchmarking platform** to collect and analyze data from various cloud resources and generate comprehensive and customized reports and dashboards for our customers.
- **AI-powered modeling platform** to build and validate accurate and robust models of different cloud resources and provide our customers with reliable and actionable insights and forecasts.
- **AI-powered optimization platform** to solve complex and multi-objective optimization problems and provide our customers with optimal or near-optimal cloud resource solutions.
- **AI-powered recommendation platform** to learn and understand the preferences and the requirements of our customers and provide them with personalized and tailored cloud resource recommendations.

Backed by these capabilities, Eviden provides various FinOps services to ensure organizations can progress and achieve desired business outcomes. From an advisory and consulting perspective, Eviden's FinOps Consulting Packages consist of Cost Optimizer and Transformation Jumpstart. Cost Optimizer is a service that specifically homes in on key workloads with the goal of ensuring they have been designed and run with optimal price-to-performance in mind. This can be explored at a further at the next tier, Transformation Jumpstart – where FinOps culture and maturity within the organization are both assessed in depth, and the relationship between cloud performance and business objectives are realized. Additionally, Eviden provides full FinOps managed services that are focused on managing key elements of cost optimization. This assists customers in maintaining good cost management hygiene throughout the lifetime of these applications and initiatives, while continually monitoring the cloud technology landscape to capitalize on the best price-to-performance scenarios.

Industry use case for a hypothetical cloud customer

Examining a hypothetical scenario, this section presents a real-world industry use case. It illustrates how IT Decision-Makers within the organization of a cloud customer can benefit from implementing the strategies discussed earlier, showcasing tangible results in terms of cost savings and improved performance.

To illustrate how AI and FinOps can help find the best price for performance on the cloud for accelerator backed instances, let us consider a **hypothetical cloud customer**, who is a **large e-commerce company**. Now the customer's CTO wants to use accelerator backed instances for its ML and deep learning applications, such as **product recommendation, image recognition, and sentiment analysis**.

Following are the CTO's objectives and constraints:

- Achieve **high performance** and **accuracy** for its ML and deep learning applications and meet the expectations and the satisfaction of their end customers.
- **Minimize the cost** of its accelerator backed instances and stay within its cloud budget and forecast.
- Leverage the **best cloud platform** and the **best resource combination**, that suits their ML and deep learning applications, and their business needs and goals.
- Have **flexibility** and **scalability** for their accelerator backed instances and be able to adjust them according to the changing workload and demand.

In comes Eviden, a GSI that specializes in cloud cost optimization and performance management, helps the customer's CTO achieve their objectives and constraints, using its AI and FinOps solutions, through the following structured approach and steps:

Performance and the cost of different accelerator backed instances are measured and compared, across different cloud providers, regions, availability zones, and resource configurations, using standardized tests and metrics, such as the MLPerf benchmark. Subsequently, comprehensive, and customized reports and dashboards are generated that show the performance and the cost results and rankings of different accelerator backed instances and highlight the best options for the CTO's specific use cases and workloads.

Mathematical and statistical models are created and tested of the behavior and the characteristics of different accelerator backed instances, using historical and real-time data. This is used to predict and simulate the performance and the cost of different accelerator backed instances, under different scenarios and conditions, and estimate the impact of various factors, such as the workload type, size, and duration, the resource type, number, and generation, the cloud provider, region, and availability zone, and the operating system, network, and storage.

The result is reliable and actionable insights and forecasts that can help plan and prepare for the cloud needs and demands. Finding the optimal or near-optimal accelerator backed instances that satisfy the performance, and the cost objectives and constraints of the CTO is key, using mathematical and heuristic algorithms. This can help minimize the cost and maximize the performance of the accelerator backed instances, by finding the best trade-offs and compromises among the available options and adjusting the resource allocation and utilization according to the changing needs and demands.

Further, most suitable, and relevant accelerator backed instances are ranked that match the preferences and the requirements of the CTO, using machine learning and data mining techniques. Personalized and tailored cloud resource recommendations are provided that can help discover and explore new and better accelerator backed instances and make informed and confident decisions.

In addition, Eviden uses its FinOps solutions to help the customer gain visibility, assign accountability, and optimize the cloud spending and performance for its accelerator backed instances.

Another real-world customer example

To illustrate how AI, FinOps, and Eviden's expertise can help you get the best value from accelerator backed instances on the cloud, let us consider an example of a cloud customer, who is facing some of the challenges mentioned above.

ABWB Inc. (customer name masked) is a leading company in the field of computer vision, that develops and deploys AI models for various applications, such as face recognition, object detection, and scene segmentation. ABWB Inc. uses accelerator backed instances on the cloud, to train and run their AI models, and to provide their services to their clients. However, ABWB Inc. is struggling with the following issues:

- ABWB Inc. is spending a lot of money on the cloud resources, but they are not sure if they are getting the best performance and quality for their AI models.
- ABWB Inc. is using different cloud providers, regions, and availability zones, for their accelerator backed instances, but they are not sure if they are choosing the optimal combinations, that can meet their performance and cost requirements.
- ABWB Inc. is not able to monitor and manage their cloud spending and performance, in a granular and consistent way, and they are not able to allocate and optimize their cloud resources, according to their business needs and goals.

ABWB Inc. decides to engage Eviden, a leading GSI that specializes in cloud cost optimization and performance management and get a free consultation and a customized solution for their organization.

Eviden uses AI and FinOps to help ABWB Inc. achieve the following outcomes:

Eviden uses AI to analyze the historical and current data of ABWB Inc.'s cloud usage and performance, and to identify the best cloud platform and resource combinations, that can deliver the optimal performance and quality for their AI models, at the lowest possible cost. Eviden uses FinOps to provide ABWB Inc. with a dashboard and a report, that show them the detailed breakdown of their cloud spending and performance, across different cloud providers, regions, and availability zones, and that highlight the areas of improvement and optimization.

Eviden uses FinOps to help ABWB Inc. to implement the best optimization techniques and tools, such as resizing, scaling, scheduling, spot instances, reserved instances, and savings plans, that can help them reduce their cloud spending and improve their cloud performance, in a flexible and agile way. Eviden helps ABWB Inc. to stay updated and informed about the latest trends and developments in cloud computing, and to help them migrate and leverage the new and more powerful computing devices, that can enhance their AI models and services.

As a result of using Eviden's solution, ABWB Inc. was able to achieve the following benefits:

- ABWB Inc. was able to save up to **40%** on their cloud spending, while maintaining or improving the performance and quality of their AI models and services.
- ABWB Inc. was able to gain visibility and transparency into their cloud costs and performance, and to align and optimize their cloud resources, according to their business needs and goals.
- ABWB Inc. was able to take advantage of the latest and most advanced devices, that can boost their AI capabilities and competitiveness.

This is just one example of how AI and FinOps can help you get the best value from accelerator backed instances on the cloud. If you want to learn more about how Eviden can help you achieve similar or better results, you can contact them and get a free consultation and a customized solution for your organization.

Business value and conclusion

The main business value and benefits of using AI and FinOps for finding the best price for performance on the cloud, for accelerator backed instances, are:

- **Cost savings:** AI and FinOps can help IT Decision-Makers reduce their cloud spending and optimize their cloud budget, by finding the most cost-effective and efficient cloud resources, and applying the best optimization techniques and tools.
- **Performance improvement:** AI and FinOps can help IT Decision-Makers improve their cloud performance and accuracy, by finding the most suitable and relevant cloud resources, and adjusting them according to the workload and demand.
- **Flexibility and scalability:** AI and FinOps can help IT Decision-Makers achieve flexibility and scalability for their cloud resources, by finding the best cloud platform and the best resource combination and enabling them to adapt to the changing needs and demands.
- **Decision support:** AI and FinOps can help IT Decision-Makers make informed and confident decisions, by providing them with data-driven and automated methods, such as benchmarking, modeling, optimization, and recommendation, and giving them reliable and actionable insights and forecasts.
- **Collaboration and alignment:** AI and FinOps can help IT Decision-Makers collaborate and align their financial and operational aspects of cloud computing, by providing them with visibility, accountability, and optimization, and establishing the roles, responsibilities, and ownership, of the cloud users and the cloud managers.

Our conclusion emphasizes the value of combining AI, FinOps, and Eviden's expertise to maximize cloud computing efficiency. It offers actionable insights for improving cloud cost management and achieving the best price-to-performance ratios. The discussion includes a case study of a large e-commerce company utilizing accelerator-backed instances like GPUs and ASICs for ML and deep learning. It highlights how Eviden's solutions can assist IT decision-makers like CTOs in meeting their goals within budget constraints and shares Eviden's perspective on cloud pricing, values, and principles.

Call to action

If you are an IT Decision-Maker, who wants to use accelerator backed instances for your HPC and AI applications, on the cloud, you need to consider the following call to actions:

- Contact Eviden, a leading GSI, that can help you find the best price for performance on the cloud, for accelerator backed instances, using its AI and FinOps solutions.
- Request a free consultation and a demo, to see how Eviden can help you measure and compare, predict, and simulate, optimize, and select, and recommend and rank the best cloud resources, for your specific use cases and workloads.
- Start your cloud optimization journey, with Eviden, and enjoy the benefits of cost savings, performance improvement, flexibility and scalability, decision support, and collaboration and alignment.

As an IT Decision-Maker, here are some key take-aways for you to remember:

- Accelerator backed instances, such as GPUs and ASICs, are essential for HPC and AI applications, but they are also costly and complex to manage.
- The cost of accelerator backed instances is increasing over time, as the cloud providers introduce new and more powerful devices, such as the NVIDIA A100 Tensor Core GPU, the AWS Inferentia ASIC, and the Google Cloud TPU.
- AI can help you find the best cloud platform and resource combinations, that suit your HPC and AI applications, and your business needs and goals.
- FinOps can help you manage and optimize your cloud spending and performance, by providing you with visibility, accountability, and optimization.
- Eviden can help you navigate the complex and dynamic landscape of cloud computing and provide you with the best practices and principles, that can help you achieve your cloud goals and objectives.

If you want to learn more about how to get the best value from accelerator backed instances on the cloud, you can contact Eviden, a leading GSI that specializes in cloud cost optimization and performance management and get a free consultation and a customized solution for your organization.

Appendix A: Additional success stories – cost optimizer

A media organization from the UK, QWERTY Inc. (customer name masked), had various FinOps challenges associated with strong growth and a general lack of understanding or control around cloud consumption. Due to many acquisitions, there were a wide range of ways of working leading to sprawl and little to no unified policies or governance with cloud. As they continued to grow, QWERTY were also focused on new technologies necessitated by new revenue streams – all the while, app teams had little-to-no reassurance that they were planning or executing efficiently. This, combined with a lack of control, unified process, policies, and governance, resulted in climbing spend with no priority placed on reigning it in.

QWERTY Inc. then decided to work with Eviden for a Cost Optimizer engagement. This short but effective engagement allowed QWERTY Inc. to get significant insights on key areas of focus for cost optimization and acted as a starting point for process improvement and unification, along with driving accountability between teams.

QWERTY Inc. and Eviden selected 3 of the most impactful workloads that encompassed **70%** of QWERTY Inc.'s spend and conducted individual deep-dive workshops and cost and usage reviews with the application teams. Detailed reports were shared with each of the involved stakeholders and high-level process and structure reviews were also conducted and shared. Immediately, this provided the application teams with direct visibility to begin optimizing, enabled by Eviden's tooling trial. Focal points from a cost optimization perspective for these workloads comprised of storage lifecycle management, version control and a guidance needed for starting a dedicated FinOps team - though QWERTY Inc. is likely to engage Eviden for further FinOps managed services in this regard.

The key outcomes for QWERTY Inc. following the Cost Optimizer engagement were:

- Application teams beginning the process of cost optimization and reductions with newfound visibility and understanding of their cloud consumption.
- Cost optimization recommendations for the 3 workloads totalling up to approximately **$265,000**, of which QWERTY Inc. immediately began implementation on AWS ElastiCache rightsizing, and RDS reservations
- Storage lifecycle management policy implementations that drove approximately **$36,000** in additional savings

Appendix B: Additional success stories – transformation jumpstart

An insurance organization from the UK, XYZ Co. (customer name masked), had significant difficulties in developing an effective FinOps practice in the wake of massively increasing cloud costs. They had an existing FinOps team which were struggling to make an impact and were ultimately receiving very little sponsorship within the organization. This was proving to be very problematic as cloud cost increases were climbing, despite intending to be a small proportion of the overall budget. Even though there was a major focus on cost-cutting organization-wide, it was not being realized in cloud the same way as it was elsewhere in the organization.

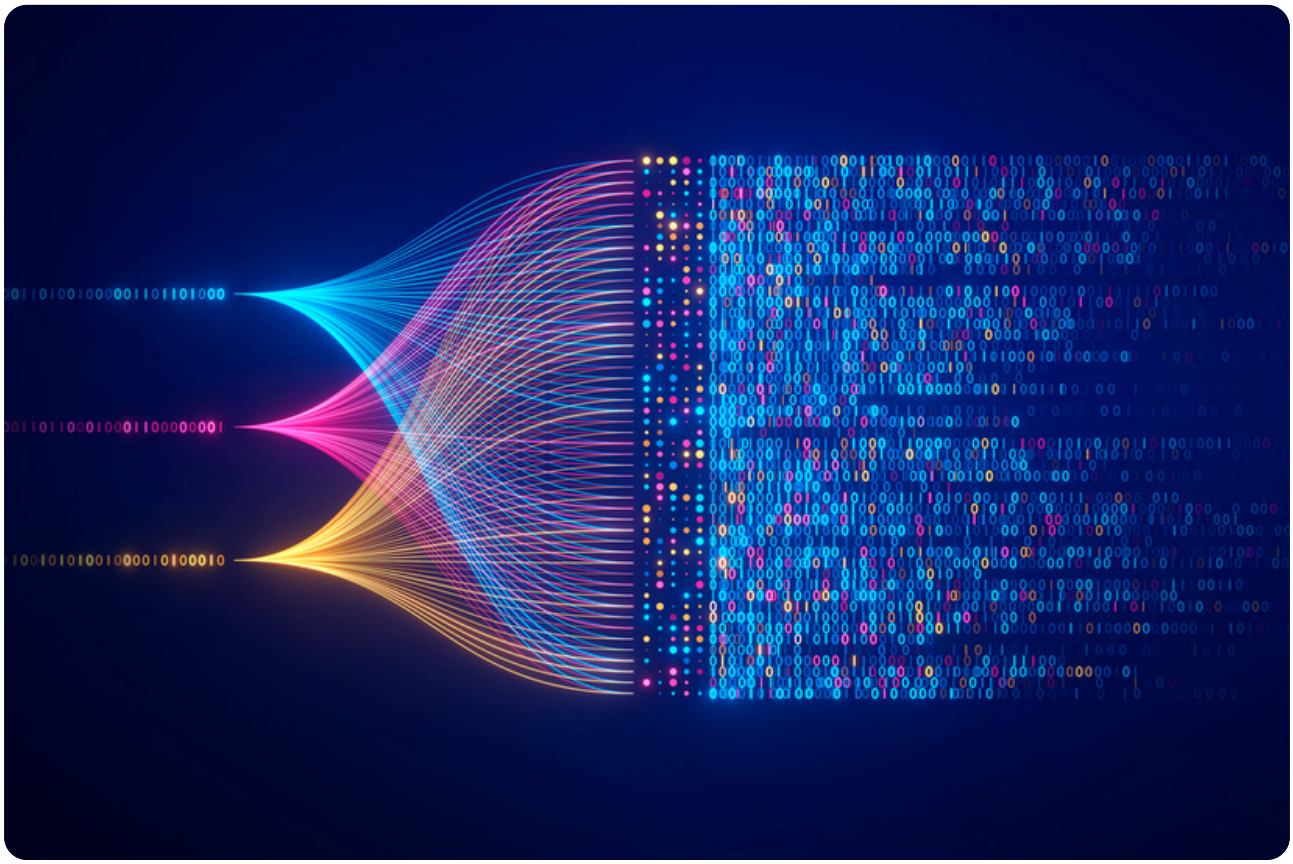
XYZ Co. worked with Eviden on a Transformation Jumpstart engagement to begin the process of strengthening its FinOps practice and developing its maturity. This 2-month engagement allowed XYZ Co. to uncover where their areas for improvement were and enabled the FinOps team to get accountability and buy-in from cloud engineering stakeholders.

Eviden worked with the customer to conduct in-depth analysis of their cloud estate and overall FinOps processes. Visibility was the first step, and this engagement allowed Eviden to improve actual cost and usage visibility through recommended tooling and implementing processes for alerting. This enabled the engineering teams to view the results of their day-to-day activities and impacts of cloud decision-making, driving awareness and emphasizing their effect on the budget. Furthermore, complete cost analysis and cost optimization reports were shared with key stakeholders, along with detailed architecture reviews of their largest applications.

Budget and forecasting modifications were suggested to improve on a cumbersome existing budget and forecasting process, and sustainability concepts were also introduced to drive added accountability.

Overall, the impact of this engagement resulted in the following outcomes:

- Total annual rate optimization recommendations were approximately **$1.2 million**
- Database optimizations were suggested, largely focusing on RDS instances that had been ignored, totaling approximately **$756,000**
- Savings recommendations on the two largest workloads alone totaled approximately **$632,000** ($252,000 in usage optimization and $380,000 in rate optimization)



Contact the authors:
ashok.ramachandran@eviden.com
brian.ray@eviden.com
brandon.wong@eviden.com

Connect with us

eviden.com