TATA CONSULTANCY SERVICES

Building on belief

AI Workload Migration to Cloud: Server or Serverless?

Abstract

The future of the digital enterprise will be driven by artificial intelligence (AI). Global spending on AI is estimated to double, growing from $50.1 billion in 2020 to more than $110 billion[1] in 2024, according to International Data Corporation (IDC). As companies embrace AI, they are also investing more in cloud services and software as a service which make it easier to manage machine learning (ML) and deep learning (DL) workloads. Through 2023, AI will be one of the top workloads that drive IT infrastructure decisions[2], according to a Gartner prediction. However, deploying ML and DL workloads on the cloud incurs a recurring cost. In addition, cloud deployment comes with service level agreements (SLAs) to deliver low latency and high throughput, especially while inferencing the models for making predictions.

To ensure optimal workload deployment, organizations must choose cloud services that ensure high performance and are cost-effective. One approach is to use containers with virtual machines (VMs) or optimally configured instances in ML platforms. Serverless-architecture-based deployment is also a popular solution for cloud migration due to its high scalability and pay-per-use cost model. This white paper highlights the cost and performance tradeoffs between an ML platform and serverless deployment. This knowledge, will be highly beneficial in deploying applications such as recommender systems.

Workload Deployment on Cloud: Balancing Cost and Performance

On-premise deployment of applications requires an upfront investment to build compute capacity. In certain instances, due to peak workload, purchasing, and deploying hardware resources in excess to avoid SLA violations, results in higher maintenance costs. This is more likely to happen when resources are idle due to low load during non-peak time.

To avoid this, enterprises are migrating their ML and DL workloads to the cloud since it offers a highly flexible infrastructure, low operational management, and better performance. Cloud vendors also provide many services, platforms, and hardware instances for deploying these workloads. For enterprises, the choice should be based on application characteristics and design, performance requirements, and the available budget. This will ensure the right balance between cost and performance.

---

[1] *IDC, Worldwide Spending on Artificial Intelligence Is Expected to Double in Four Years, Reaching $110 Billion in 2024, According to New IDC Spending Guide; Aug 25 2020; Worldwide Spending on Artificial Intelligence Is Expected to Double in Four Years, Reaching $110 Billion in 2024, According to New IDC Spending Guide*

[2] *Gartner; Our Top Data and Analytics Predicts for 2021; Jan 12 2021; Gartner Blog Network*

Popular cloud vendors provide ML platforms for deploying AI workloads, such as SageMaker by Amazon Web Services (AWS), Microsoft Azure open-source MLFlow, etc., as shown in Figure 1. These platforms ease the process by providing automatic training, deployment, and inference of the models[3].

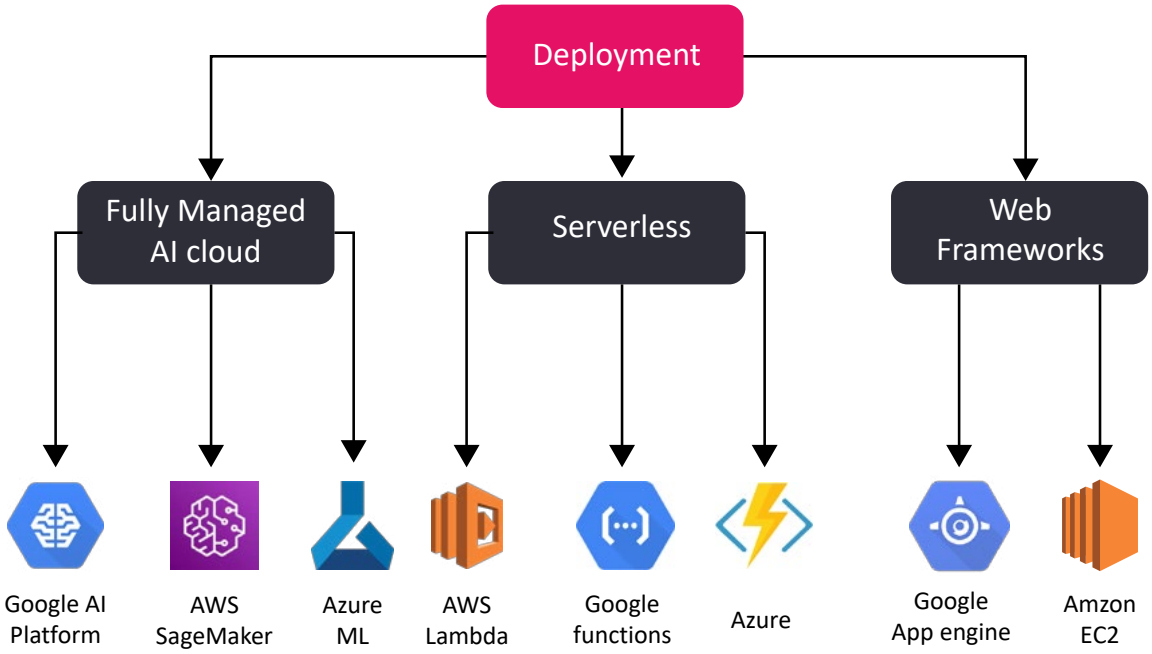

Figure 1: Platforms for deploying AI workloads

These platforms from cloud vendors also provide various computing, memory-intensive CPU (central processing unit), and GPU (graphics processing unit) instances for deploying ML and DL models. Defining appropriate scaling policies helps address SLA violations due to dynamically varying workloads. However, frequent scaling up and down the VM instances due to 'bursty' or variable workload could increase the cost, and degrade application performance.

Serverless platforms are emerging as an alternate option for applications built using models based on ML and DL for inference[4]. The autoscaling feature in the serverless platform allows up-scaling and down-scaling instances as the workload increases or decreases. Moreover, billing is based on the actual usage, unlike VM billing instances, where the cost depends on the lease duration, irrespective of the actual usage. However, there are certain inherent constraints in serverless platforms, such as statelessness and limited compute power. In addition, serverless instances have a 15-minute lifetime and suffer from a cold start problem. This results in high latency for the first inference request served by the model. Popular cloud vendors such as AWS Lambda, Azure Functions, Google function, and the open-source platform OpenWhisk offer serverless platform services.

---

[3] *Dheeraj Chahal, Ravi Ojha, Sharod Roy Choudhury, and Manoj Nambiar; ACM Digital Library; Migrating a Recommendation System to Cloud Using ML Workflow; April 2020; https://doi.org/10.1145/3375555.3384423*

[4] *D. Chahal, M. Ramesh, R. Ojha and R. Singhal; Institute of Electrical and Electronics Engineers; High Performance Serverless Architecture for Deep Learning Workflows, Aug 2021; High Performance Serverless Architecture for Deep Learning Workflows | IEEE Conference Publication | IEEE Xplore*

How to Choose the Right Cloud Deployment Platform for AI Applications: A Use Case

The following use case of a recommender system will help understand the tradeoffs between cost and performance for an on-premise, ML platform, and serverless architecture-based deployment of an AI application. A recommender system based on a graph neural network uses products chosen in the past to recommend the next product. The three scenarios for analyzing deployment of the recommender system are as follows:

- On-premise on Intel servers containing four and eight physical cores with 256 GB memory.
- ML platform (such as AWS SageMaker) with ml.c5.xlarge (four cores) and ml.c5.2xlarge (eight cores) compute family CPU instances and gdn.ml.xlarge and gdn.ml.2xlarge family GPU instances.
- Serverless instances (AWS Lambda) with 3GB memory resulting in two cores on each instance.

Figure 2 shows the response time comparison between the on-premise, ML platform (CPU and GPU), and serverless platform. At a lower workload, the on-premise deployment results in the least response time compared to the ML and serverless platform. However, as the workload increases, the response time increases for on-premise and ML platforms. This is primarily due to the fixed number of resources and compute capacity (number of cores) available for on-premise servers and slow scalability of the ML platform.

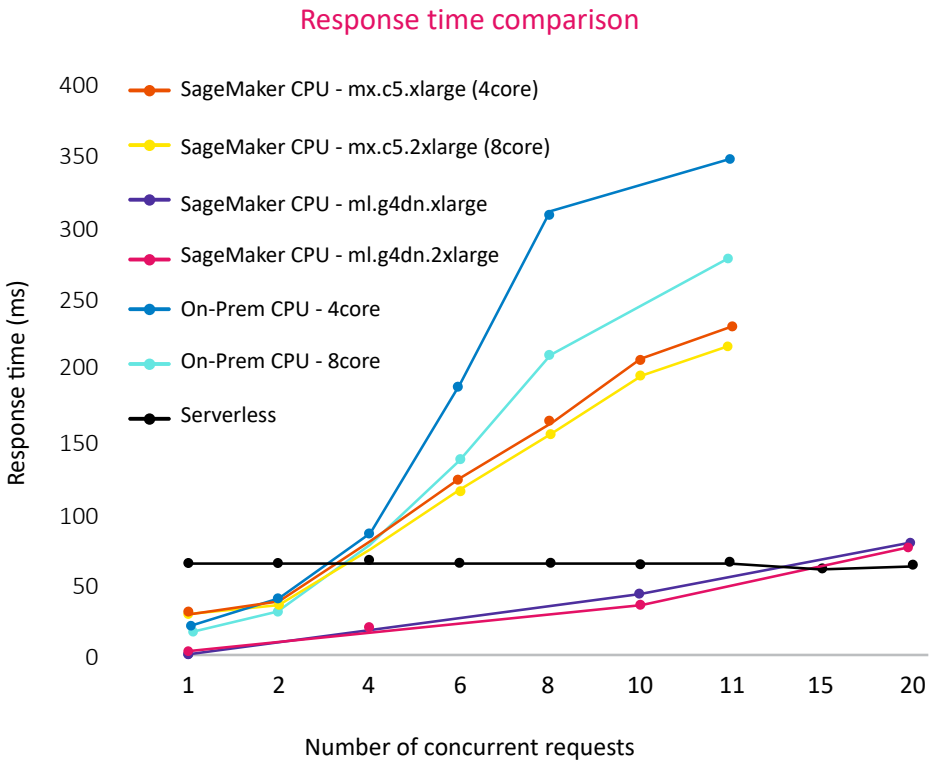

Figure 2: Performance comparison between the on-premise, ML (CPU and GPU instances) and serverless platforms

The response time for serverless architecture remains the same with an increase in workload.

On the other hand, response time for the serverless platform remains steady even with an increase in workload. This is because, with an increase in inference requests, the number of instances serving these requests increases proportionally. Moreover, new serverless instances are spawned faster compared to those on the ML platform.

One of the challenges of cloud migration is comparing the cost efficiency of different services across varying cost models. The cost of ML platform instances is fixed and independent of the load served. In the serverless option, costs change with the workload and new instances. Figure 3 shows the cost comparison of instances used for inference in the ML and serverless platforms. The ML platform is cost-effective when the request rate served by instances is high. However, a serverless platform is a better choice if the resources reserved for inferencing are underutilized. The intersection of the serverless platform with the instances in Figure 3, represents the breakeven point. ML platform instances can serve workloads with request rates above this breakeven point with low latency and cost. Conversely, workloads with a request rate below the intersection point means that the ML platform instances are underutilized, resulting in higher costs. In such a scenario, the serverless platform is a cost-effective solution due to its pay-per-use cost model.

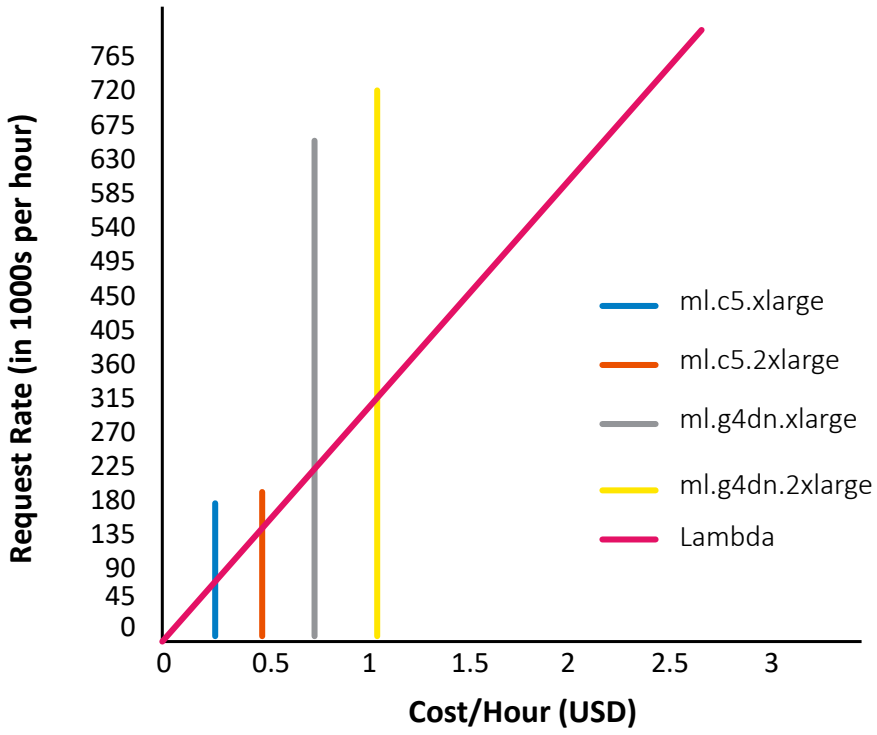

Figure 3: Cost comparison between the ML and serverless platform with increasing workload (request rate)

> From a cost perspective, a serverless platform is a better choice if the resources reserved for inferencing are underutilized.

Assessing the Performance-Cost Ratio Across Multiple Dimensions

There are certain scenarios where an on-premise and ML-platform-based deployment delivers better performance and cost ratio. For example, when a workload consists of many requests and resource utilization remains high for a long duration. Serverless platforms work better for short-duration workloads with low resource utilization.

The on-premise and ML-platform-based deployment on cloud platforms performs better than serverless platforms for smaller workloads. As the workload increases, the application performance drops due to limited scalability and resource constraints. Such workloads can be migrated to the cloud to leverage the extensive and flexible infrastructure. ML platforms on the cloud are easy to deploy but are slow when it comes to scaling operations. Serverless platforms provide high scalability and result in cost-effective deployment, particularly for 'bursty' or variable workloads.

The use of serverless platforms in conjunction with ML platforms might be advantageous for dynamically varying workloads. In this approach, the workload can be balanced between ML platform and serverless instances based on the instantaneous utilization of the resources, as depicted in Figure 4.

> The choice of the ideal cloud deployment platform should be based on application characteristics, which extract high performance at optimal costs. Compute requirement, memory usage and expected workload are some the parameters for comparison.

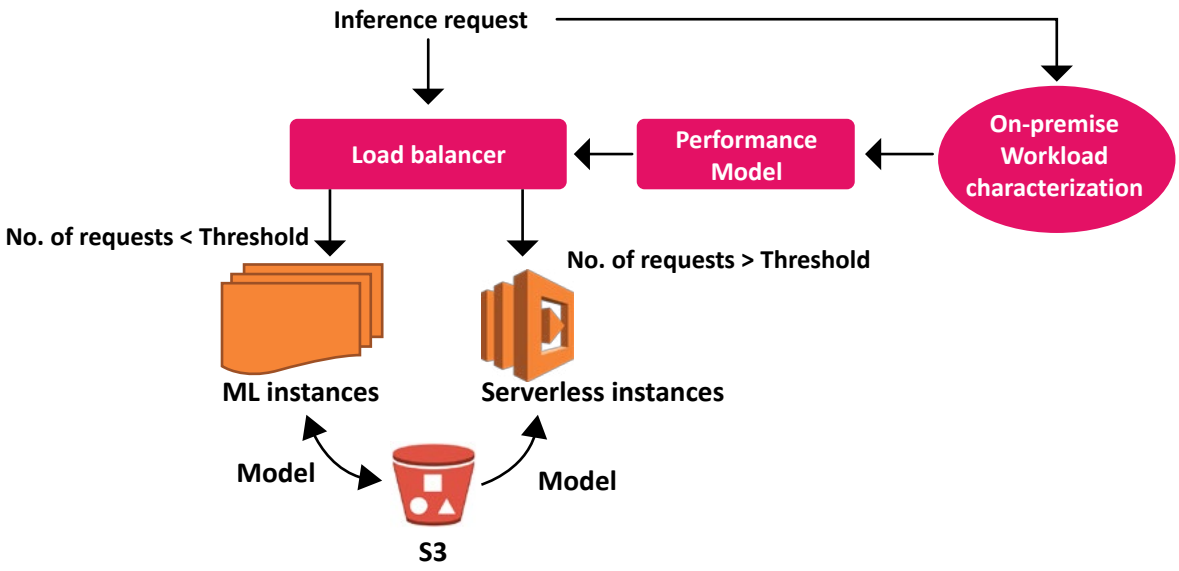

Figure 4: Load balancing across cloud services

The Future of AI Lies in Serverless Computing

AI-driven platforms are the future, enabling businesses to make quick and innovative decisions, and transforming the customer experience. When deployed correctly, serverless computing is ideal for handling ML and DL workloads, offering dramatic cost savings due to execution-based pricing and better resource utilization. It enables rapid prototyping of new and complex AI services by eliminating the time spent on managing the server's infrastructure. In the future, cloud vendors may provide GPUs within serverless platforms which would increase the computation power significantly and deliver state-of-the-art AI capabilities. Furthermore, it will support compute intensive jobs such as training large deep learning models. Going forward, a serverless architecture may be the driving force behind AI innovation.

About the authors

Dheeraj Chahal is Senior Scientist at TCS Research. He received a PhD from Clemson University, South Carolina, USA. Prior to joining TCS, he was Staff Software Engineer at IBM's Systems and Technology group. His research interests include high performance computing, cloud computing, and performance modeling. He has played a key role in the development of tools such as PerfExt++, Decide, and MAPLE at TCS. Chahal has published more than 20 research papers in reputed international conferences and has filed several patents.

Mayank Mishra is a Scientist at TCS Research, where he creates frameworks for accelerating ML/DL training and inference pipelines. His research interests include cloud computing, computer networks, and high performance computing. Mayank received a PhD and M.Tech. from IIT Bombay. Prior to joining TCS Research, he was a post-doctoral research fellow at the Department of Electrical and Computer Engineering at Iowa State University. He has several years of industry experience at Cisco Systems and Webaroo (a startup focused on offline search).

TATA CONSULTANCY SERVICES

Awards and accolades

#1 IN CUSTOMER SATISFACTION 2020/21 EUROPE Whitelane Research

top EMPLOYER NORTH AMERICA 2021 CERTIFIED EXCELLENCE IN EMPLOYEE CONDITIONS

top EMPLOYER GLOBAL 2020 CERTIFIED EXCELLENCE IN EMPLOYEE CONDITIONS

ASIA'S OUTSTANDING COMPANIES POLL 2020 ASIAMONEY

STEVIE AWARDS 2020

CONGRATULATION CIO100 WINNERS 2020

FinanceAsia ASIA'S BEST COMPANIES 2020

DIGITAL IMPACT AWARDS WINNER

CANADIAN HR AWARDS 2020

THE CIVIC 50 POINTS OF LIGHT

FTSE4Good

CDP DISCLOSURE INSIGHT ACTION

2020 Brand Finance® Awards
TOP 3 IT SERVICES BRAND

2020 Brand Finance® Awards
FASTEST GROWING IT SERVICES BRAND FOR THE DECADE 2010 - 2020

AWARDED Superbrands BRITAIN'S CHOICE 2020

AWARDED Superbrands AMERICA'S CHOICE 2020

Contact

For more information on Research and Innovation visit https://www.tcs.com

Email: innovation.info@tcs.com

About Tata Consultancy Services Ltd (TCS)

Tata Consultancy Services is a purpose-led transformation partner to many of the world's largest businesses. For more than 50 years, it has been collaborating with clients and communities to build a greater future through innovation and collective knowledge. TCS offers an integrated portfolio of cognitive powered business, technology, and engineering services and solutions. The company's 500,000 consultants in 46 countries help empower individuals, enterprises, and societies to build on belief.

Visit **www.tcs.com** and follow TCS news **@TCS**.

Corporate Marketing | Design Services | M | 02 | 22