Publicado en Ciudad de México, a 21 de diciembre de 2020. el 22/12/2020

# [NVIDIA A100 se lanza en AWS, esto marca el comienzo de la próxima década de la Computación Cloud Acelerada](http://www.notasdeprensa.es)

## La nueva instancia Amazon EC2 P4d con A100 ya está disponible, las GPUs de NVIDIA cumplen una década al servicio de AWS

 Por IAN BUCK. https://blogs.nvidia.com/blog/2020/11/02/nvidia-a100-launches-on-aws/ La primera instancia de GPU de Amazon Web Services debutó hace 10 años, con la NVIDIA M2050. En ese momento, las aplicaciones basadas en CUDA se centraban principalmente en acelerar las simulaciones científicas. Todavía faltaba mucho para la aparición de la IA y el deep learning. Desde entonces, AWS agregó a su conjunto de instancias de GPU en el cloud la K80 (p2), K520 (g3), M60 (g4), V100 (p3/p3dn) y T4 (g4). Con la nueva instancia P4d disponible a partir de hoy, AWS prepara el camino para otra increíble década de computación acelerada con la GPU NVIDIA A100 Tensor Core más reciente. La instancia P4d ofrece el rendimiento más alto de AWS, la plataforma basada en GPU más rentable para las aplicaciones de computación de alto rendimiento y entrenamiento de machine learning. La instancia reduce hasta 3 veces el tiempo para entrenar modelos de machine learning con FP16 y hasta 6 veces con TF32, en comparación con la precisión FP32 determinada. También proporcionan un rendimiento de inferencia excepcional. El mes pasado, las GPUs NVIDIA A100 arrasaron en las evaluaciones de inferencia de MLPerf, ya que lograron un rendimiento 237 veces más rápido que las CPUs. Cada instancia P4d cuenta con ocho NVIDIA A100 GPUs y, con AWS UltraClusters, los clientes obtener acceso escalable y según demanda a más de 4,000 GPUs a la vez usando Elastic Fabric Adaptor (EFA) de AWS y al almacenamiento escalable y de alto rendimiento con Amazon FSx. P4d ofrece una red de 400 Gbps y usa tecnologías de NVIDIA, como: NVLink, NVSwitch, NCCL y GPUDirect RDMA, para acelerar aún más las cargas de trabajo de entrenamiento de deep learning. NVIDIA GPUDirect RDMA en EFA garantiza conexiones de red de baja latencia, ya que pasa los datos de una GPU a otra GPU entre servidores sin pasar por la CPU ni la memoria del sistema. Además, la instancia P4d es compatible con muchos servicios de AWS, incluidos Amazon Elastic Container Services, Amazon Elastic Kubernetes Service, AWS ParallelCluster y Amazon SageMaker. P4d también puede aprovechar todo el software contenerizado y optimizado, disponible en NGC, lo que incluye aplicaciones de HPC, frameworks de IA, modelos previamente entrenados, plantillas de Helm y software de inferencia, como: TensorRT y Triton Inference Server. Las instancias P4d también están disponibles en las regiones Este y Oeste de EE.UU. Pronto, también estarán disponibles en más regiones. Las instancias se pueden comprar On-Demand, con Savings Plans, con Reserved Instances o como Spot Instances. La primera década de la computación de cloud con GPU proporcionó más de 100 exaflops de computación de IA al mercado. Con el lanzamiento de la instancia Amazon EC2 P4d impulsada por las GPUs NVIDIA A100, la próxima década de la computación de cloud con GPU comenzará con el pie derecho. NVIDIA y AWS permiten que las aplicaciones continúen rompiendo los límites de la IA en una amplia variedad de especialidades. "Estamos ansiosos por ver cómo las aprovecharán los clientes". Visitar AWS y comenzar a usar las instancias P4d hoy mismo.

**Datos de contacto:**

Carlos Valencia

MKQ PR Agency

55 39 64 96 00

Nota de prensa publicada en: [https://www.mexicopress.com.mx/nvidia-a100-se-lanza-en-aws-esto-marca-el](http://www.notasdeprensa.es/educalivecom-disfruta-de-cursos-presenciales-y-clases-particulares-a-traves-de-internet)

Categorías: Nacional Inteligencia Artificial y Robótica Programación Hardware E-Commerce Software



[**http://www.mexicopress.com.mx**](http://www.notasdeprensa.es)