GIRONA NOTICIES.COM
+ Periodisme

Clockwork.io presenta una nueva clase de tolerancia a fallos para acabar con el desperdicio de GPU

Comunicados | 11/03/2026 | 00:00

La nueva solución TorchPass aborda un desafío multimillonario en la infraestructura de IA; utiliza migración en vivo de GPU para mantener el entrenamiento de IA a gran escala en funcionamiento ante fallos de hardware en lugar de obligar a costosos reinicios

La nueva solución TorchPass aborda un desafío multimillonario en la infraestructura de IA; utiliza migración en vivo de GPU para mantener el entrenamiento de IA a gran escala en funcionamiento ante fallos de hardware en lugar de obligar a costosos reinicios

Clockwork.io, líder en Software-Driven AI Fabrics™, una capa de software programable y neutral respecto a proveedores que optimiza clústeres de GPU a gran escala para observabilidad en tiempo real, tolerancia a fallos y rendimiento determinista, anunció hoy la disponibilidad general de TorchPass Workload Fault Tolerance. Esta nueva clase de tolerancia a fallos impulsada por software elimina uno de los modos de fallo más costosos en el entrenamiento de IA a gran escala: los reinicios catastróficos de trabajos provocados por fallos de infraestructura.

Ofrecido como una capacidad central de la plataforma FleetIQ de Clockwork.io, TorchPass aplica los principios de Software-Driven AI Fabrics al entrenamiento distribuido, utilizando migración en vivo de GPU para permitir que las cargas de trabajo continúen ejecutándose a través de fallos de GPU, interrupciones de red, errores de controladores e incluso caídas completas de nodos, sin reinicios desde checkpoints ni pérdida de progreso.

"Las empresas están invirtiendo miles de millones en chips de nueva generación, y sin embargo los costes de ejecutar trabajos de IA distribuidos siguen estando enormemente inflados porque el ecosistema ha aceptado el fallo como una constante", afirmó Suresh Vasudevan, CEO de Clockwork.io. "Construimos TorchPass para rechazar fundamentalmente esa premisa. En lugar de tratar el fallo como inevitable y reiniciar después, TorchPass hace que los fallos de infraestructura sean invisibles para la carga de trabajo: el entrenamiento continúa a través de los fallos de forma transparente, en software. Para un despliegue típico de 2.048 GPU, eso se traduce en más de 6 millones de dólares al año en capacidad de cómputo recuperada. Esto es exactamente lo que nuestro enfoque de Software-Driven AI Fabric fue diseñado para ofrecer: infraestructura de IA tolerante a fallos".

Dylan Patel, fundador y CEO de SemiAnalysis, coincidió en que los trabajos de entrenamiento a gran escala están limitados por interrupciones. "A medida que se despliegan clústeres Blackwell

con un dominio NVL72, y mirando al futuro con el dominio NVL576 de Rubin Ultra, la idea de que un solo error de GPU o una fluctuación en un enlace de red pueda detener toda una ejecución es totalmente inaceptable", afirmó Patel. "TorchPass resuelve un enorme desafío de fiabilidad de clústeres: proporciona conmutación por error transparente y migración en vivo de cargas de trabajo que mantiene alto el MFU, lo que a su vez impulsa una mejor economía de GPU".

Por qué el entrenamiento de IA falla a escala

El entrenamiento distribuido de IA sigue siendo una de las cargas de trabajo más propensas a fallos en la infraestructura moderna. A medida que crece el tamaño de los clústeres, la fragilidad aumenta de forma pronunciada. Investigaciones de Meta FAIR muestran que el tiempo medio hasta el fallo desciende a 7,9 horas en un clúster de 1.024 GPU y a solo 1,8 horas en uno de 16.384 GPU. Esto significa que, para la mayoría de las grandes empresas centradas en IA o nubes de IA, los reinicios provocados por fallos son completamente inevitables, lo que convierte este problema en una gran barrera para escalar el impacto de la IA.

Cada fallo obliga a los trabajos de entrenamiento a retroceder hasta el checkpoint más reciente, descartando minutos u horas de trabajo completado y perdiendo tiempo adicional en intervención manual, reprovisión de recursos y reinicio del entrenamiento. Estos reinicios limitan silenciosamente la utilización de GPU, lo que convierte la fiabilidad en uno de los mayores costes ocultos de la infraestructura de IA.

TorchPass aborda este problema al gestionar de forma proactiva los fallos costosos de cargas de trabajo de IA, resolviéndolos antes de que el trabajo se detenga o necesite reiniciarse. Fundamental para empresas que ejecutan grandes cargas de trabajo de IA y para nubes de IA por igual, TorchPass mejora drásticamente la fiabilidad de las cargas de trabajo y la utilización de los clústeres. Para las nubes de IA, que ahora pueden solucionar problemas en GPU afectadas mientras mantienen la ejecución del entrenamiento según lo previsto, esto se traduce en mejores SLA para los clientes y una economía global de nube de IA más favorable, mejorando su capacidad para proteger márgenes y ofrecer nuevos modelos antes.

"Gestionar la producción de cómputo en clústeres de GPU a gran escala es vital para garantizar que estamos ofreciendo capacidad fiable a nuestros clientes. Al usar TorchPass contamos con el respaldo de una empresa que se centra en la resiliencia como si fuera una función central del negocio: sustituye cualquier GPU específica que falle y mantiene el resto del trabajo en marcha, en lugar de permitir que un pequeño problema afecte a nuestras operaciones a gran escala", afirmó David Power, CTO de Nscale.

Habilitando la próxima generación de infraestructura de IA

Al convertir la fiabilidad en una capacidad definida por software en lugar de una limitación de hardware, TorchPass proporciona la confianza operativa necesaria para desplegar sistemas de nueva generación altamente acoplados, como los NVIDIA GB200 y NVIDIA GB300 NVL72, así como futuros sistemas a escala de rack, donde las arquitecturas densas amplifican el coste incluso de pequeños fallos.

TorchPass se basa en el lanzamiento previo de Network Fault Tolerance de Clockwork.io, que aplica los mismos principios de Software-Driven AI Fabric a la resiliencia de red mediante el redireccionamiento transparente del tráfico alrededor de fallos de enlace.

Para obtener más información sobre el lanzamiento de TorchPass, visitar al equipo de Clockwork.io en persona en NVIDIA GTC 2026 del 16 al 19 de marzo, en el stand n.º 205, o visitar https://clockwork.io.

Acerca de Clockwork.io

Clockwork.io es pionera en Software-Driven AI Fabrics™, ofreciendo una capa de software programable que hace que los clústeres de IA a gran escala sean observables, deterministas y resilientes por diseño para impulsar el progreso continuo de las cargas de trabajo y la máxima utilización del clúster. Su plataforma FleetIQ permite a las empresas entrenar, desplegar y servir las cargas de trabajo de IA más exigentes del mundo de forma más rápida, fiable y a menor coste. Empresas como Uber, Wells Fargo, Nebius, Nscale y White Fiber confían en Clockwork.io para impulsar su infraestructura de IA. Más información en www.clockwork.io.

Autor: Redacció

https://gironanoticies.com/comunicado/290850-clockworkio-presenta-una-nueva-clase-de-tolerancia-a-fallos-para-acabar-con-el-desperdicio-de-gpu.htm