

Huawei Cloud na veletrhu KubeCon EU 2024: příchod inteligentní éry díky nepřetržitým inovacím s otevřeným zdrojovým kódem

26.3.2024 - Lavanda Wang | PROTEXT

Umělá inteligence představuje klíčovou výzvu pro paradigma nativního cloudu.

Cloudové technologie v posledních letech způsobily revoluci v tradičních IT systémech a urychlily digitální pokrok v oblastech, jako jsou internet a vládní služby. Nativní cloud díky správě mikroslužeb přinesl nové možnosti, jako jsou bleskový prodej a agilní provoz, například DevOps. Tyto změny mají zásadní dopad na životy lidí i rychlý růst a široké přijetí umělé inteligence, včetně rozsáhlých modelů, se staly základem průmyslové inteligence.

Podle průzkumu společnosti Epoch z roku 2023 se výpočetní výkon potřebný pro základní modely zvyšuje každých 18 měsíců desetinásobně, což je pětikrát rychleji než tempo růstu předpovězené Moorovým zákonem pro obecné výpočty. Vznik tohoto "nového Moorova zákona" v důsledku umělé inteligence a převaha rozsáhlých modelů umělé inteligence představuje výzvu pro technologie nativního cloudu. Dennis Gu ve svém projevu nastínil následující klíčové body:

- Nízké průměrné využití GPU/NPU zvyšuje náklady na trénování a odvozování umělé inteligence.
- Častá selhání velkých tréninkových shluků modelu snižují efektivitu tréninku.
- Složitá konfigurace rozsáhlých modelů má za následek vysoké nároky na vývoj umělé inteligence.
- Ruku v ruce s nasazením rozsáhlého odvozování umělé inteligence přichází riziko nepředvídatelného zpoždění přístupu koncových uživatelů a objevují se potenciální problémy s ochranou osobních údajů.

Inovace v oblasti umělé inteligence ze strany společnosti Huawei Cloud nabízí vývojářům nápady, jak řešit různé výzvy.

Rostoucí velikost modelů umělé inteligence vyžaduje více výpočetní kapacity, což představuje výzvu pro technologie nativního cloudu, ale také nabízí příležitosti pro inovace v tomto odvětví. Dennis Gu se podělil o příběhy týkající se inovací v oblasti umělé inteligence společnosti Huawei Cloud a nabídl vývojářům referenční bod pro řešení výzev.

Společnost Huawei Cloud pomocí KubeEdge – cloudové nativní platformy pro edge computing – vytvořila platformu pro plánování a správu více robotů. Pomocí této platformy mohou uživatelé s využitím příkazů v přirozeném jazyce říkat platformě, co má dělat, a systém bude koordinovat několik robotů na okraji ve snaze splnit složité úkoly. Systém je navržen s třídílnou architekturou (cloud, okrajový uzel a robot), která řeší problémy, jako je porozumění přirozenému jazyku, efektivní plánování a správa více robotů či správa přístupu k robotům různých typů. Prostřednictvím rozsáhlých modelů provádí příkazy v přirozeném jazyce a předvídá dopravu, přiděluje úkoly a plánuje trasy. Třídílná architektura výrazným způsobem zvyšuje flexibilitu robotické platformy, zlepšuje efektivitu řízení o 25 %, zkracuje dobu potřebnou pro nasazení systému o 30 % a zkracuje dobu potřebnou pro nasazení nových robotů z měsíců na pouhé dny.

Hlavní službou jedné z předních platforem pro sdílení obsahu v Číně, kterou využívá více než 100

milionů aktivních uživatelů měsíčně, jsou doporučeni na domovské stránce. Tato funkce vychází z modelu s téměř 100 miliardami parametrů. K trénování tohoto modelu platforma využívá tréninkový clutser s tisíci výpočetními uzly, včetně stovek ps a pracovníků na jednu tréninkovou úlohu. Existuje proto velká poptávka po lepším plánování topologie, vysokém výkonu a velké propustnosti. Volcano, projekt s otevřeným zdrojovým kódem, rozšiřuje podporu pro úlohy umělé inteligence nebo strojového učení v systému Kubernetes a nabízí řadu zásad pro správu úloh a pokročilé plánování. Volcano zahrnuje algoritmy, jako je plánování s ohledem na topologii, plnění přihrádek a plánování s ohledem na dohodu o úrovni služeb (SLA), což vede k 20% zlepšení celkového výkonu tréninku a výraznému snížení složitosti provozu a údržby platformy.

Bezserverová umělá inteligence stojí v čele vývoje nativního cloudu.

Řada podniků a vývojářů čelí výzvě, jak efektivně a spolehlivě provozovat aplikace umělé inteligence a zároveň minimalizovat provozní náklady. Společnost Huawei Cloud přišla s řešením tohoto problému, když identifikovala klíčové požadavky na cloudové nativní platformy umělé inteligence a zavedla nový koncept nazvaný Serverless AI.

Dennis Gu během svého vystoupení vysvětlil, že bezserverová umělá inteligence je navržena tak, aby zjednodušila složité úlohy trénování a odvozování pomocí inteligentního doporučování paralelních zásad, což vývojářům usnadňuje její používání. Obsahuje také funkci automatického rozšíření adaptivního GPU/NPU dynamicky upravující přidělování prostředků na základě změn pracovní zátěže v reálném čase a zajišťující efektivní provádění úloh. Kromě toho je v bezserverové umělé inteligenci k dispozici bezporuchový clutser GPU/NPU, takže se již vývojáři nemusí obávat, že by hardwarové poruchy mohly přerušit služby. A co je nejdůležitější, bezserverová umělá inteligence je kompatibilní s běžnými architekturami umělé inteligence, takže vývojáři mohou snadno integrovat své stávající nástroje a modely umělé inteligence.

Bezserverová umělá inteligence je také velmi významným pokrokem pro poskytovatele cloudových služeb. Bezserverová umělá inteligence přináší řadu výhod, jako jsou lepší využití GPU/NPU, efektivnější hybridní pracovní zátěž pro trénování, odvozování a vývoj i ekologicky šetrné výpočty díky lepší energetické účinnosti, takže můžete ušetřit peníze za elektřinu. Bezserverová umělá inteligence navíc umožňuje sdílení GPU/NPU mezi více nájemci v různých prostorech nebo v různém čase, což zvyšuje míru opětovného využití prostředků. Nejdůležitějším aspektem bezserverové umělé inteligence je její schopnost zajišťovat garantovanou kvalitu služeb (QoS) a SLA pro úlohy školení i odvozování, což zajišťuje stabilní a velmi kvalitní služby.

Bezserverová umělá inteligence využívá flexibilní vrstvu plánování prostředků, která je postavena na virtualizovaném operačním systému. Tato vrstva zapouzdřuje základní funkce aplikačních rámců do vrstvy zprostředkování aplikačních prostředků. Dennis Gu představil referenční architekturu pro bezserverovou umělou inteligenci. Domnívá se, že tento návrh architektury umožňuje bezserverové umělé inteligenci automatické řízení rozsáhlých zdrojů. To zahrnuje přesnou analýzu vzorců využití prostředků, sdílení prostředků z heterogenních hardwarových fondů a zajištění odolnosti proti chybám během úloh tréninku umělé inteligence skrze virtualizaci GPU/NPU a živou migraci zátěže. Kromě toho vícerozměrné plánování a adaptivní pružné škálování zlepšují využití prostředků.

Techničtí odborníci ze společnosti Huawei Cloud na dílčím fóru poznamenali, že počet úloh spojených s umělou inteligencí nebo strojovým učením, které běží na platformě Kubernetes, neustále roste. Výsledkem je, že řada společností buduje platformy umělé inteligence nativního cloudu na několika clutsech Kubernetes, které jsou rozmístěny v datových centrech a na různých typech GPU. Aplikace Karmada a Volcano zvládají inteligentně plánovat pracovní zátěž GPU ve více clusterech, podporují přenos chyb a zajišťují konzistenci a efektivitu v rámci clusterů i mezi nimi. Kromě toho mohou vyvažovat využití prostředků v celém systému a kvalitu služeb pracovních úloh s

různými prioritami, aby se vypořádaly s výzvami správy rozsáhlých a heterogenních prostředí GPU.

Karmada nabízí okamžitou a spolehlivou automatickou správu aplikací ve scénářích s více cloudy a hybridními cloudy. Stále více uživatelů pomocí aplikace Karmada vytváří přizpůsobitelná a efektivní řešení v produkčních prostředích. Aplikace Karmada byla v roce 2023 oficiálně povýšena na inkubační projekt CNCF a komunita se těší, že se k ní připojí další partneři a vývojáři.

Volcano Gang Scheduling je řešení pro scénáře distribuovaného tréninku umělé inteligence a velkých dat, přičemž řeší problémy nekonečného čekání a slepých uliček v distribuovaných tréninkových úlohách. Zásadou topologie úloh a plánování s ohledem na vstupy a výstupy se minimalizuje zpoždění přenosu distribuovaných tréninkových úloh, čímž se tréninkový výkon zlepší o 31 %. Kromě toho minResources řeší spor o zdroje mezi ovladačem Spark a spouštěčem ve scénářích s vysokou souběžností, optimalizuje stupeň paralelismu a zvyšuje výkon o 39,9 %.

Dennis Gu je přesvědčen, že klíčem ke zvýšení produktivity umělé inteligence je agilita cloudových technologií a inovace heterogenních výpočetních platforem umělé inteligence. Společnost Huawei Cloud se věnuje inovacím s otevřeným zdrojovým kódem a prostřednictvím spolupráce s kolegy z oboru se snaží nastolit novou éru inteligence.

Foto – https://mma.prnewswire.com/media/2370741/Dennis_Gu_Chief_Architect_Huawei_Cloud.jpg

KONTAKT: Lavanda Wang, lavanda.wang@huawei.com

<https://www.ceskenoviny.cz/tiskove/zpravy/huawei-cloud-na-veletrhu-kubecon-eu-2024-prichod-inteligentni-ery-diky-nepretrzitym-inovacim-s-otevrenym-zdrojovym-kodem/2498017>