

ارائه برنامه‌های کاربردی هوش مصنوعی در لبه شبکه از طریق تجمیع بارهای کاری هوش مصنوعی و خدمات شبکه 5G



مقدمه

کسب‌وکارها امروزه با چالش‌های قابل‌توجهی در بهبود تجربه مشتریان مواجه هستند، در حالی که باید هزینه کل مالکیت (TCO) خود را نیز کاهش دهند. این چالش‌ها نیازمند راه‌حل‌های نوآورانه‌ای هستند که نه تنها امنیت را تضمین کنند، بلکه کارایی خدمات را نیز افزایش دهند، چرا که زیرساخت‌های سنتی فناوری اطلاعات غالباً برای پاسخگویی به نیازهای برنامه‌های کاربردی مدرن و مشتری‌محور کافی نیستند. این نیازها منجر به گرایشی به سمت هوش مصنوعی لبه‌ای شده است؛ جایی که پردازش داده‌ها نزدیک‌تر به منبع داده، به صورت لحظه‌ای، با تأخیر کمتر و حفظ حریم خصوصی بیشتر انجام می‌شود.

چالش‌ها

محدودیت‌های زیرساختی:

کسب‌وکارها هنگام تلاش برای پیاده‌سازی برنامه‌های هوش مصنوعی لبه‌ای در کنار زیرساخت‌ها و فرآیندهای موجود خود، با مشکلات فنی مواجه می‌شوند. تنظیمات فعلی فناوری اطلاعات اغلب به دلیل سخت‌افزارهای قدیمی، توانایی پردازش کافی برای مدیریت کارآمد برنامه‌های پیشرفته هوش مصنوعی را ندارند، که این امر منجر به افزایش تأخیر و کاهش کارایی فناوری‌های مشتری‌محور می‌شود.

هزینه کلی عملیات:

جداسازی زیرساخت‌های مخابرات 5G از سیستم‌های محاسباتی هوش مصنوعی ناکارآمد است. یکپارچه‌سازی بارهای کاری هوش مصنوعی با شبکه‌های 5G از طریق یک پلتفرم محاسباتی یکپارچه و با عملکرد بالا، توانایی‌های فنی را بهبود می‌بخشد و هزینه‌های مربوط به تجهیزات، مصرف انرژی و فضای فیزیکی را کاهش می‌دهد.

تجربه مشتری:

ارائه‌دهندگان خدمات 5G باید خدمات شخصی‌سازی شده و حساس به زمان ارائه دهند که انتظارات مشتریان را برآورده کند. بنابراین، همان‌طور که پیش‌تر اشاره شد، گرایشی به سمت هوش مصنوعی لبه‌ای شکل گرفته است، جایی که پردازش داده‌ها به صورت لحظه‌ای، نزدیک‌تر به منبع داده، با تأخیر کمتر و حفظ حریم خصوصی بیشتر انجام می‌شود.

راه حل

تجمع هوش مصنوعی در لبه شبکه با استفاده از پلتفرم هوش مصنوعی لبه‌ای و شبکه 5G، تحول دیجیتال را تسریع کرده و ارزش آفرینی را بهبود می‌بخشد. با تخصص پیشرو در ارائه هوش مصنوعی سازمانی بر روی پلتفرم‌های شتاب‌یافته توسط GPU، شرکت لنر به‌خوبی برای این پارادایم جدید محاسباتی آماده است. این پلتفرم‌ها به سازمان‌ها، اپراتورهای مخابراتی و ارائه‌دهندگان خدمات ابری (CSP) ها این امکان را می‌دهند که به‌صورت یکپارچه 5G را با اکوسیستم هوش مصنوعی لبه‌ای ترکیب کرده یا gNB شبکه 5G را به یک مرکز داده لبه‌ای برای بارهای کاری هوش مصنوعی تبدیل کنند.

ECA-6040

سرور قدرتمند هوش مصنوعی لبه‌ای **ECA-6040** یک سرور 2U با شاسی کوتاه است که برای استنتاج و آموزش لحظه‌ای در لبه شبکه 5G طراحی شده است. این سرور با معماری محاسباتی شتاب‌یافته توسط GPU و قابلیت vRAN 5G به‌عنوان یک پلتفرم نرم‌افزاری تعریف‌شده ارائه می‌شود که با سایر بارهای کاری هوش مصنوعی یکپارچه شده است.

این پلتفرم از حداکثر **1,024GB حافظه DDR5** پشتیبانی کرده و از پردازنده‌های **Intel Xeon Scalable نسل چهارم** بهره می‌برد که فناوری **Intel vRAN Boost** و تا **64 هسته پردازشی** قدرتمند را ارائه می‌دهد. این سرور مجهز به یک اسلات **OCP 3.0** و چهار اسلات توسعه **PCIe** است که قابلیت پشتیبانی از کارت‌های **NVIDIA GPU** و **Smart NIC** متعددی از جمله **NVIDIA L40S**، **NVIDIA L4 GPU** و کارت‌های **Bluefield 3/ConnectX-7 NIC Smart NIC** را دارد.

ECA-6040 برای کاربردهایی مانند شهرهای هوشمند، حمل‌ونقل هوشمند، سلامت هوشمند، هوش تجاری در خرده‌فروشی، تبدیل ویدئو (Video Transcoding) و اتوماسیون صنعتی ایده‌آل است.

ویژگی‌های ECA-6040

سرور **ECA-6040** با برخورداری از پورت‌های I/O در دسترس از جلو، طراحی بدون رنگ، امنیت سخت‌افزاری **TPM 2.0** و مدیریت از راه دور **iPMI**، یک پلتفرم با عملکرد بالا برای ساخت و استقرار خدمات هوش مصنوعی لبه‌ای باز، کارآمد و امن بر روی شبکه‌های خصوصی 5G ارائه می‌دهد.

سایر ویژگی‌های برجسته این سرور شامل **4 سینی HDD/SSD با سایز 2.5 اینچ**، **دو اسلات داخلی M.2**، منبع تغذیه اضافی **1,600W AC** و **6 فن هوشمند قابل تعویض** برای کنترل حرارتی بهینه است.

نتایج

با پشتیبانی از چندین GPU و کارت‌های Smart NIC، سرور **ECA-6040** قادر است به‌طور همزمان بارهای کاری پیچیده هوش مصنوعی را مدیریت کند. این امر باعث می‌شود که تمام برنامه‌های هوش مصنوعی پردازش داده‌های خود را به‌صورت محلی انجام دهند، که نتیجه آن زمان پاسخ‌دهی سریع‌تر و کاهش وابستگی به خدمات ابری است.

با امکان استفاده از یک پلتفرم سخت‌افزاری یکپارچه مانند **ECA-6040** شرکت لنر، دستگاه‌های لبه‌ای چندمنظوره می‌توانند بارهای کاری متعدد هوش مصنوعی را به‌صورت همزمان و با منابع محدود اجرا کنند. این امر از طریق تقسیم‌بندی منابع، ایزوله‌سازی و مدیریت از راه دور امکان‌پذیر می‌شود.

مجهز به فناوری **Intel vRAN Boost** که در پردازنده‌های **Intel Xeon Scalable نسل چهارم** جدیدترین نسل این شرکت گنجانده شده است، سرور **ECA-6040** عملکرد بی‌نظیر و کارایی انرژی بالایی را برای مجازی‌سازی RAN ارائه می‌دهد، که تضمین‌کننده استفاده بهینه از منابع در استقرارهای **5G O-RAN** است.

نتیجه گیری

با استقرار یک پلتفرم ایده آل برای تجمیع بارهای کاری مانند **ECA-6040**، کسب و کارها به طور چشمگیری زمان توسعه و هزینه های کل عملیات را کاهش می دهند. مشخصات و پیکربندی های **ECA-6040** تضمین کننده یکپارچگی بدون مشکل نرم افزاری و گسترش مقیاس پذیر سخت افزاری است که تطابق کامل با زیرساخت های موجود فناوری اطلاعات را فراهم می کند. علاوه بر این، با تجمیع هوش مصنوعی و 5G از طریق **ECA-6040** لnr، فضای سخت افزاری و هزینه های مرتبط نیز به طور قابل توجهی کاهش می یابد.

ECA-6040

2U 19" Appliance With 5th Gen Intel® Xeon® Scalable Processors

CPU	Intel® Xeon® Processor Scalable Family (Codename Sapphire Rapids-SP/EMR-SP/Rapids-EE)
Chipset	Intel® C741

