

ویژگی های کلیدی

- فرم فکتور: Rack 2U
- پردازنده / شتاب دهنده: NVIDIA GH200 Grace Hopper Superchip
- تعداد Superchip: یک یا دو GH200
- حداکثر هسته CPU: تا 144 هسته Arm Neoverse V2 در پیکربندی دوگانه
- حافظه CPU: تا 960GB LPDDR5X

دریافت مشاوره خرید ۰۲۱-۹۱۰۰۹۰۷۳

توضیحات محصول

سرور HPE ProLiant Compute DL384 Gen12 یک سرور سازمانی 2U است که به طور ویژه برای نسل جدید بارهای کاری هوش مصنوعی، مدل های زبانی بزرگ، استنتاج AI و محاسبات HPC طراحی شده است. برخلاف سرورهای سنتی که CPU و GPU بخش های نسبتاً مستقلی هستند، این مدل از معماری NVIDIA Grace Hopper استفاده می کند که CPU و GPU را با پهنای باند بسیار بالا به یکدیگر متصل می کند.

پشتیبانی از حداکثر دو GH200 Superchip، حافظه HBM3e فوق سریع و ذخیره سازی NVMe Gen5 باعث شده DL384 Gen12 برای اجرای مدل های بزرگ، Fine-tuning، AI Inference، Retrieval-Augmented Generation و پردازش های علمی حافظه محور گزینه ای تخصصی باشد. HPE این سرور را همچنین برای شبیه سازی های بزرگ، EDA و پیش بینی آب و هوا مناسب معرفی می کند.



ویژگی‌های مهم محصول:

1. معماری NVIDIA Grace Hopper

DL384 Gen12 از GH200 Grace Hopper Superchip استفاده می‌کند که CPU مبتنی بر Arm و GPU معماری Hopper را در یک پلتفرم منسجم ترکیب می‌کند و گلوگاه انتقال داده بین CPU و GPU را کاهش می‌دهد.

2. پشتیبانی از دو GH200

در بالاترین پیکربندی، دو Superchip به صورت NVIDIA GH200 NVL2 با یکدیگر کار می‌کنند و در مجموع 144 هسته Arm Neoverse V2 در اختیار سیستم قرار می‌دهند.

3. حافظه یکپارچه با ظرفیت بالا

پیکربندی دوگانه دارای 960GB حافظه LPDDR5X در بخش CPU و 288GB HBM3e در بخش GPU است؛ ترکیبی که حدود 1.2TB حافظه بسیار سریع برای مدل‌ها و دیتاست‌های بزرگ فراهم می‌کند.

4. پهنای باند بسیار بالای حافظه

حافظه HBM3e در پیکربندی GH200 NVL2 تا 9.8TB/s پهنای باند ارائه می‌دهد و ارتباط CPU-to-GPU از طریق NVLink-C2C در پیکربندی دوگانه به 1800GB/s می‌رسد.

توضیحات تکمیلی:

DL384 Gen12 بیش از آنکه یک سرور عمومی برای فایل‌سرور یا مجازی‌سازی متداول باشد، یک پلتفرم تخصصی برای Accelerated Computing است. HPE این مدل را برای AI Inference، Fine-tuning، مدل‌های بزرگ، RAG و Workloadهای حافظه‌محور معرفی می‌کند. یکی از نقاط تمایز این مدل، ارتباط مستقیم و پرسرعت CPU Grace با GPU Hopper است. این معماری نیاز به جابه‌جایی مداوم داده از مسیرهای سنتی CPU/GPU را کاهش داده و برای پردازش‌هایی که به ظرفیت و پهنای باند بالای حافظه نیاز دارند اهمیت زیادی دارد.



جدول مشخصات محصول

مقدار	مشخصه
HPE ProLiant Compute DL384 Gen12	مدل
Rack 2U	فرم فکتور
NVIDIA GH200 Grace Hopper	Superchip
1 یا 2 عدد	تعداد Superchip
Arm Neoverse V2 x144	حداکثر هسته CPU
تا 960GB LPDDR5X	حافظه CPU
تا 288GB HBM3e	حافظه GPU
تا 9.8TB/s	پهنای باند HBM
EDSFF NVMe Gen5	ذخیره سازی
تا 8 درایو، بسته به پیکربندی	حداکثر EDSFF
PCIe Gen5 / OCP 3.0	رابط توسعه
HPE iLO	مدیریت



مزایای محصول

1. مناسب مدل‌های بسیار بزرگ AI

ظرفیت بالای حافظه یکپارچه اجازه می‌دهد مدل‌هایی که روی سرورهای GPU معمولی با محدودیت حافظه مواجه می‌شوند، با انعطاف بیشتری اجرا شوند.

2. کاهش گلوگاه CPU و GPU

NVLink-C2C ارتباط مستقیم و پرسرعت بین Grace CPU و Hopper GPU ایجاد می‌کند و نیاز به انتقال سنتی داده بین حافظه CPU و GPU را کاهش می‌دهد.

3. عملکرد بالا در AI Inference

HPE این سرور را برای استنتاج LLM و بارهای Generative AI بهینه کرده و برای GH200 NVL2 حدود 8 پتافلاپس توان AI اعلام کرده است.

4. مناسب پردازش‌های حافظه‌محور

ترکیب LPDDR5X و HBM3e در کنار پهنای باند بسیار بالا، DL384 را برای Workloadهایی که سرعت دسترسی به داده در آن‌ها حیاتی است مناسب می‌کند.

5. توسعه شبکه پرسرعت

پشتیبانی از Ethernet، InfiniBand و BlueField امکان اتصال سرورها در Clusterهای AI با پهنای باند بالا را فراهم می‌کند.

6. تراکم پردازشی بالا در 2U

تمام این توان پردازشی در یک شاسی استاندارد 2U Rack و با سیستم خنک‌کنندگی هوایی ارائه می‌شود و برای دیتاسنترهایی که تراکم رک اهمیت دارد مناسب است.



سوالات متداول (FAQ)

1. سرور DL384 Gen12 برای چه کاربردهایی مناسب است؟

این مدل بیشتر برای HPC، RAG، Fine-tuning، LLM Inference، Generative AI، شبیه سازی علمی و پردازش های حافظه محور طراحی شده است.

2. DL384 Gen12 از چه پردازنده های استفاده می کند؟

این سرور به جای پردازنده های متداول Intel یا AMD، از NVIDIA GH200 Grace Hopper Superchip شامل Grace CPU و Hopper GPU استفاده می کند.

3. امکان نصب چند GH200 وجود دارد؟

سرور در پیکربندی Single GH200 یا Dual GH200 NVL2 ارائه می شود و در نسخه دوگانه دو Superchip از طریق NVLink به یکدیگر متصل هستند.

4. حداکثر حافظه DL384 Gen12 چقدر است؟

نسخه Dual GH200 دارای حداکثر 960GB LPDDR5X و 288GB HBM3e است؛ در مجموع حدود 1.2TB حافظه سریع در اختیار پلتفرم قرار می گیرد.

5. چه نوع Storage روی این سرور قابل استفاده است؟

DL384 Gen12 از EDSFF NVMe Gen5 پشتیبانی می کند و بسته به کانفیگ می تواند تا 8 درایو EDSFF داشته باشد. همچنین تا دو M.2 برای کاربردهای بوت قابل پشتیبانی است.

