

CEITEC VUT využívá nový Multi-GPU systém pro pokročilé aplikace strojového učení

Superpočítače NVIDIA DGX, dodané společností M Computers, pomáhají vědcům z CEITEC VUT v Brně zpracovávat data z výrobních strojů a robotů v reálném čase. K dispozici budou i pro malé a střední podniky pro testování jejich AI aplikací.

Unikátní systémy pro vědce i průmyslové podniky

Vědecké pracoviště **CEITEC na Vysokém učení technickém v Brně (VUT)** představilo nově instalované výpočetní systémy **NVIDIA DGX A100** a **NVIDIA DGX H100**, které **dodala a nainstalovala společnost M Computers**. Tyto systémy umožní rozšíření možností ve výzkumu a aplikacích umělé inteligence. Spojují dvě generace technologie NVIDIA DGX, což zajišťuje výjimečnou výpočetní kapacitu a flexibilitu pro různé výzkumné a průmyslové aplikace.

Prof. Ing. Pavel Václavěk, Ph.D., vedoucí výzkumné skupiny Kybernetika a robotika a koordinátor výzkumného programu Průmyslová kybernetika, instrumentace a systémová integrace na CEITEC VUT, zdůrazňuje využití nových systémů v rámci projektů programu Digital Europe, jako jsou EDIH-DIGIMAT, zaměřený na digitalizaci a robotizaci výrobních společností, a AI TEF AI-MATTERS, síť testovacích prostředí pro ověřování AI v průmyslovém sektoru.

„V rámci našich služeb EDIH a TEF poskytujeme firmám možnost experimentovat s AI, vzdělávat se a testovat aplikace umělé

inteligence na špičkových systémech, které jsou součástí nově instalovaného superpočítače,“ vysvětluje prof. Václavěk. „Tak umožníme malým a středním podnikům do 499 zaměstnanců využívat pokročilé technologie za 100% dotovanou cenu. Naším cílem je i integrace systému DGX s dalšími technologiemi našeho RICAIP Testbedu Brno tak, abychom mohli zpracovávat data z výrobních strojů a robotů v reálném čase.“



Videoreportáž k dodanému řešení



Vysoký výkon i vyladěné prostředí

Nové systémy **NVIDIA DGX A100** a **NVIDIA DGX H100**, každý s osmi navzájem propojenými GPU akcelerátory a celkovou pamětí 640 GB, poskytují výkonné nástroje pro masivně paralelní výpočty, což je klíčové pro zpracování rozsáhlých datových setů získaných z dat s výrobních technologií.

Oba výpočetní uzly jsou propojeny sítí **InfiniBand** s přenosovou rychlostí až 200

Gb/s, což zaručuje mimořádně rychlou a efektivní komunikaci mezi systémy. Kromě vysokého výkonu poskytují tyto systémy také robustní softwarovou vrstvu, včetně předinstalovaného a vyladěného prostředí pro strojové učení, které umožňuje snadné a rychlé nasazení do provozu.

Další výhodou je přímé propojení s online databází nejpoužívanějších frameworků a knihoven pro AI, což umožňuje uživatelům snadno stahovat a používat různé softwarové nástroje ve formě kontejnerů, což urychluje vývoj a implementaci AI aplikací.

„Díky těmto systémům můžeme nabídnout firmám a našim vědcům přístup k nejmodernějším technologiím, což umožňuje rychlejší a efektivnější výzkum,“ dodává prof. Václavěk. Po instalaci kampusové 5G sítě je toto dalším doplněním infrastruktury RICAIP Testbedu Brno v této oblasti.

CEITEC VUT tak potvrzuje svou pozici předního vědeckého pracoviště v oblasti výzkumu a využití nejmodernějších technologií pro podporu vědy a průmyslu.

Technická specifikace

Systémy NVIDIA DGX nejsou jen špičkovým hardwarem, přicházejí i s inovativními vylepšeními pro snadnější správu infrastruktury a implementaci AI. Jsou vybaveny vyladěným prostředím Dockeru a operačním systémem DGX OS a kromě toho také nově nabízejí nástroj **NVIDIA Base Command**, který umožňuje efektivní správu celé infrastruktury. Tím se zjedno-

dušuje nasazování a implementace AI aplikací pro výzkumné a vývojové týmy.

Dále je součástí systému i softwarový stack **NVIDIA AI Enterprise** (NVAIE), který poskytuje kompletní sadu nástrojů pro vývoj a optimalizaci AI aplikací. Tato kombinace technologií usnadňuje a urychluje proces vývoje a nasazení AI řešení v rámci celé infrastruktury.



PARAMETR

NVIDIA DGX H100 640 GB

NVIDIA DGX A100 640 GB

GPUs	8 × NVIDIA H100 SXM5 80 GB	8 × NVIDIA A100 SXM4 80 GB
GPU memory	640 GB total	640 GB total
CPU	2 × Intel Xeon Platinum 8480C CPU, (112 jader) 2.00 GHz	2 × AMD Epyc 7742 (128 jader, 2.25GHz)
Výkon (tensor operace)	32 PetaFLOPS (FP8)	5 PetaFLOPS (FP16)
# CUDA jader	135 168	55 296
# Tensor jader	4 224	3 456
Multi-instančce GPU	56 instancí	56 instancí
RAM	2 TB	2 TB
HDD	OS: 2 × 1.92 TB NVMe data: 30 TB (8 × 3.84 TB) NVMe	OS: 2 × 1.92 TB NVMe data: 30 TB (8 × 3.84 TB) NVMe
Network	8 × ConnectX-7 400Gb/s InfiniBand 4 × ConnectX-7 200Gb/s Ethernet	8 × ConnectX-7 200Gb/s InfiniBand 4 × ConnectX-7 200Gb/s Ethernet
Max. spotřeba	10,2 kW	6,5 kW
Provedení	rack, 8U	rack, 6U

Softwarová výbava

Čím se nejvíce odlišují DGX systémy od bare-metal řešení je softwarová výbava. Všechny shodně nabízejí již předinstalovanou, a především výkonově vyladěnou prostředí pro strojové učení (např. Caffé, resp. Caffé 2, Theano, TensorFlow, Torch nebo MXNet) nebo intuitivní prostředí pro datové analytiky (NVIDIA Digits). To vše elegantně zabalené v Docker kontejnerech. Tyto neustále aktualizované kontejnery lze stáhnout na stránkách NVIDIA GPU Cloud (NGC).

NVIDIA GPU Cloud (NGC)

NVIDIA GPU Cloud (NGC) představuje katalog Docker obrazů nejpoužívanějších prostředí pro vývoj machine learning a deep learning aplikací, HPC aplikací nebo vizualizaci akcelerovanou NVIDIA GPU kartami. Nasazení těchto aplikací je pak otázkou zkopírování odkazu na příslušný

Docker obraz, jeho spuštění na DGX systému a stažení a spuštění v Docker kontejneru.

Jednotlivá vývojová prostředí — verze všech obsažených knihoven a frameworků, nastavení parametrů prostředí — jsou aktualizovaná a optimalizovaná NVIDIA pro nasazení na DGX systémech.

Takto vyladěné prostředí poskytuje dle NVIDIA o 30 % vyšší výkon pro aplikace v oblasti učení strojů proti aplikacím nasazeným jenom čistě na NVIDIA hardware. Hlavní výhodou předinstalovaného prostředí je ale rychlost nasazení, které se pohybuje v jednotkách hodin.

Podpora hardwaru i softwaru

Silnou stránkou nabízeného řešení NVIDIA je také podpora celého systému. Rychlá hardwarová podpora (v případě selhání některé z komponent) je samozřejmostí.

Zásadní je softwarová podpora pro celé

prostředí pro případ, že něco nefunguje tak, jak má. Zákazník má k dispozici stovky vývojářů připravených pomoci. Podpora je součástí nákupu všech systémů NVIDIA DGX. Je k dispozici na 3–5 let a po skončení této doby je možné ji dále prodloužit.



Společnost M Computers zastupuje společnost NVIDIA na českém trhu v Enterprise oblasti výpočetních akceleratorů a AI systémů. Jako první společnost ve střední a východní Evropě získala nejvyšší status ELITE PARTNER a zároveň i dvě ocenění NVIDIA AI Innovator a NVIDIA AI Champion. Při vyhlášení NVIDIA Partner Awards EMEA 2024 pak získala prestižní ocenění v kategorii Go to Market Excellence.

