Hewlett Packard Enterprise

HPE GreenLake for File (nejen) pro AI Factory

Ladislav Pecen

AIDAYS 2024
23.10.2024

HPE – historie, současnost i budoucnost protkaná systémy pro vývoj a provoz AI

- HPE je leader v HPC řešeních, masivně využívaných v AI
- HPE dlouhodobě roste v oblasti technologií pro AI – ucelená řešení i jednotlivé HW či SW prvky
- Strategické akvizice i vlastní vývoj v tomto směru
- Spolupráce s předními hráči v oblasti AI – např. NVIDIA
- Vlastní infrastrukturní řešení – compute i storage
- HPE GreenLake = „cloud-like zkušenost“

Determined AI

Pachyderm

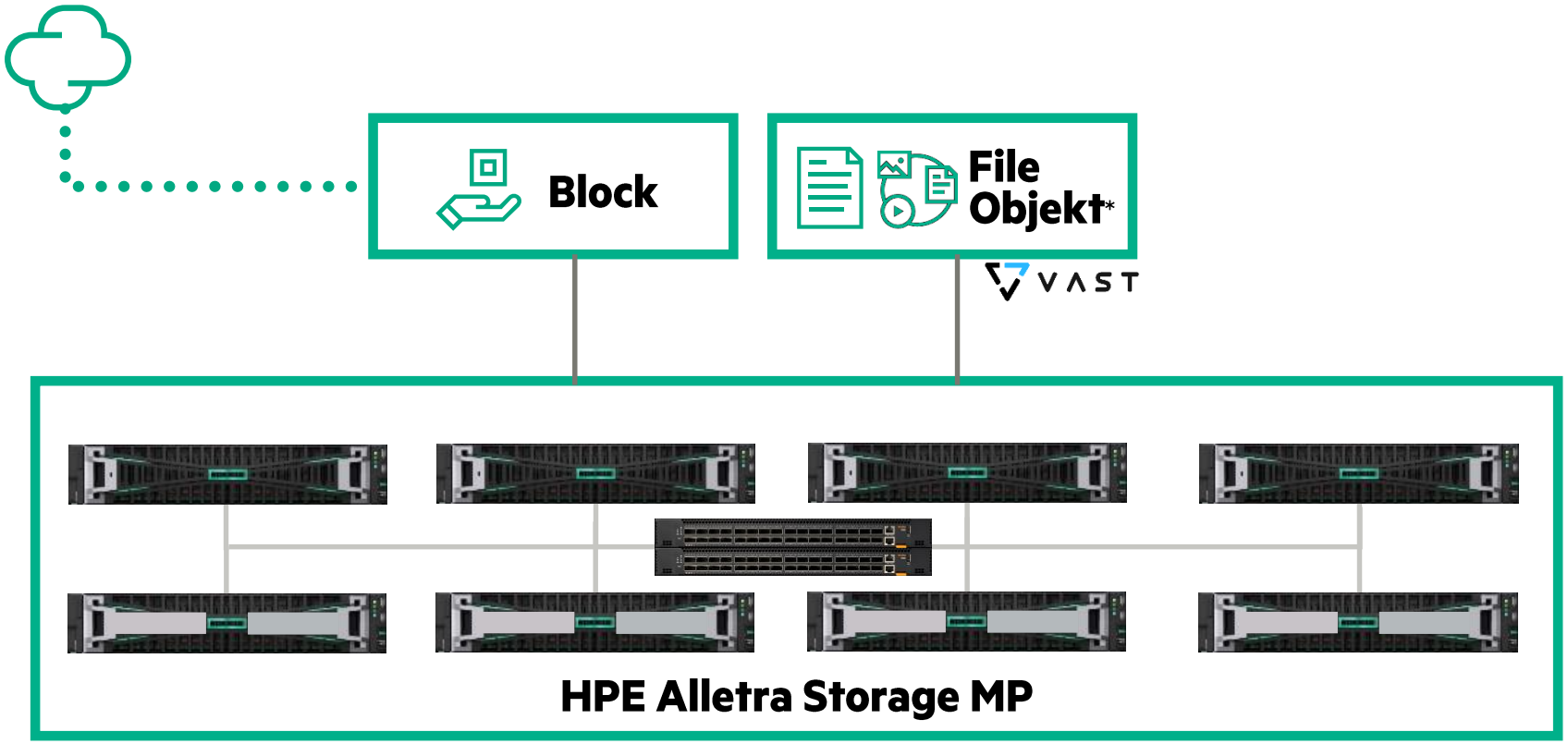
HPE GreenLake for File Storage jako univerzální úložiště nestrukturovaných dat

CHARAKTER DAT	Úložiště pro nestrukturovaná data				
WORKLOAD	**Obecné použití**	**High Performance Datová analýza**	**Modelování a simulace**		**Umělá inteligence**
PŘÍKLADY POUŽITÍ	**Vysoce dostupné souborové úložiště** • Home/Project Dir. • Zdrojové kody, knihovny • Backup/Recovery/DR	**Mission Critical workloady** • Databázová data • In-memory analytické workloady s velkou sdílenou pamětí v rámci jednoho systému	**Model-driven** • Computer-aided engineering (CAE) in Aerospace, Automotive and Manufacturing • Algorithmic Trading in FSI	**Sensor-driven** • Seismic processing in Oil & Gas • Particle accelerator in High Energy Physics • Electron Microscopy in Life Science	**Trénování a provozování modelů AI** • Machine Learning • Deep Learning • Natural Language Processing (NLP) • Computer Vision • Speech Recognition • Provoz – inference nad hotovým AI modelem
DATOVÉ VZORY	• Vysoká dostupnost • Různé velikosti souborů • Datové služby (replikace, redukce atd) • Malé i velké datasety	• Mixed, málo predikovatelné zátěžové vzory • Různé velikosti souborů • Malé i velké datasety	• Random reads with Large Output • Varying size • Starts small and grows	• Large Input • Big Files • Enormous volumes of data	• Rozsáhlá vstupní data • Náhodný přístup (hlavně čtení) • Rychlý nárůst objemu dat

Veleúspěšný koncept Alletra storage MP, to je základ HPE GreenLake for File

HPE GreenLake platform - control plane, ovládání

Jedna HW platforma, různé typy úložišť – blokové nebo souborové nebo objektové



* Funkcionalita typu objekt umožněna pouze pro vybrané workloady

HPE GreenLake for File Storage škáluje výkon a kapacitu nezávisle

K tomu super jednoduché ovládání pomocí HPE GreenLake Platform

HPE GreenLake for File Storage OS

HPE Alletra MP architektura

HPE GreenLake for File Storage

Disaggregated shared everything

Na začátku malé prostředí, pak nezávislé škálování kapacity a výkonu

Efektivní redukce objemu dat

Lokální komprese, globální deduplikace, rovněž podobnostní redukce

Vysoká dostupnost

Odolnost proti výpadkům na všech vrstvách

Architektura Scale-out

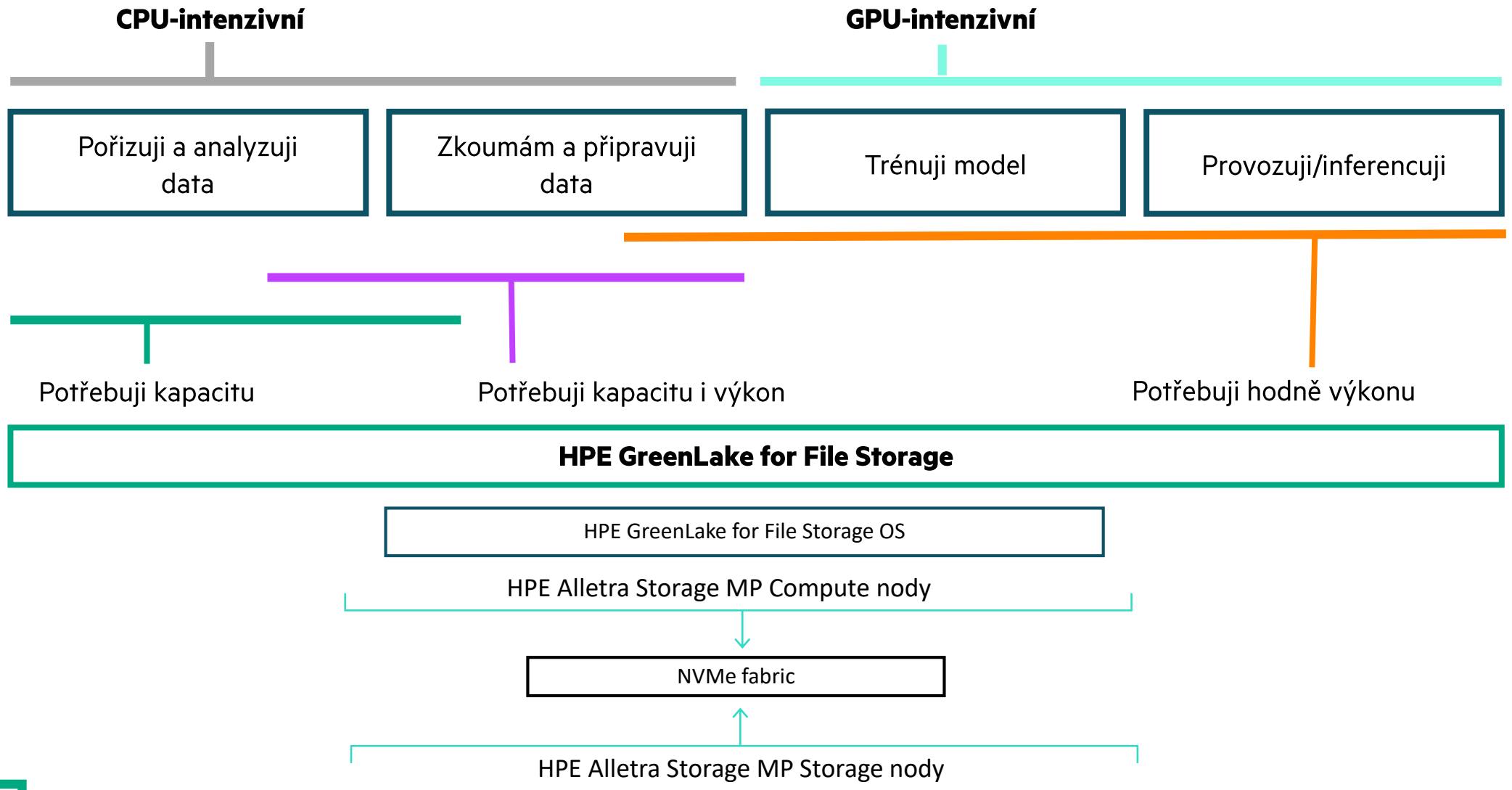
Přidám řadiče, dostanu adekvátně vyšší výkon

Zrychlení
výkon, který roste dle potřeby

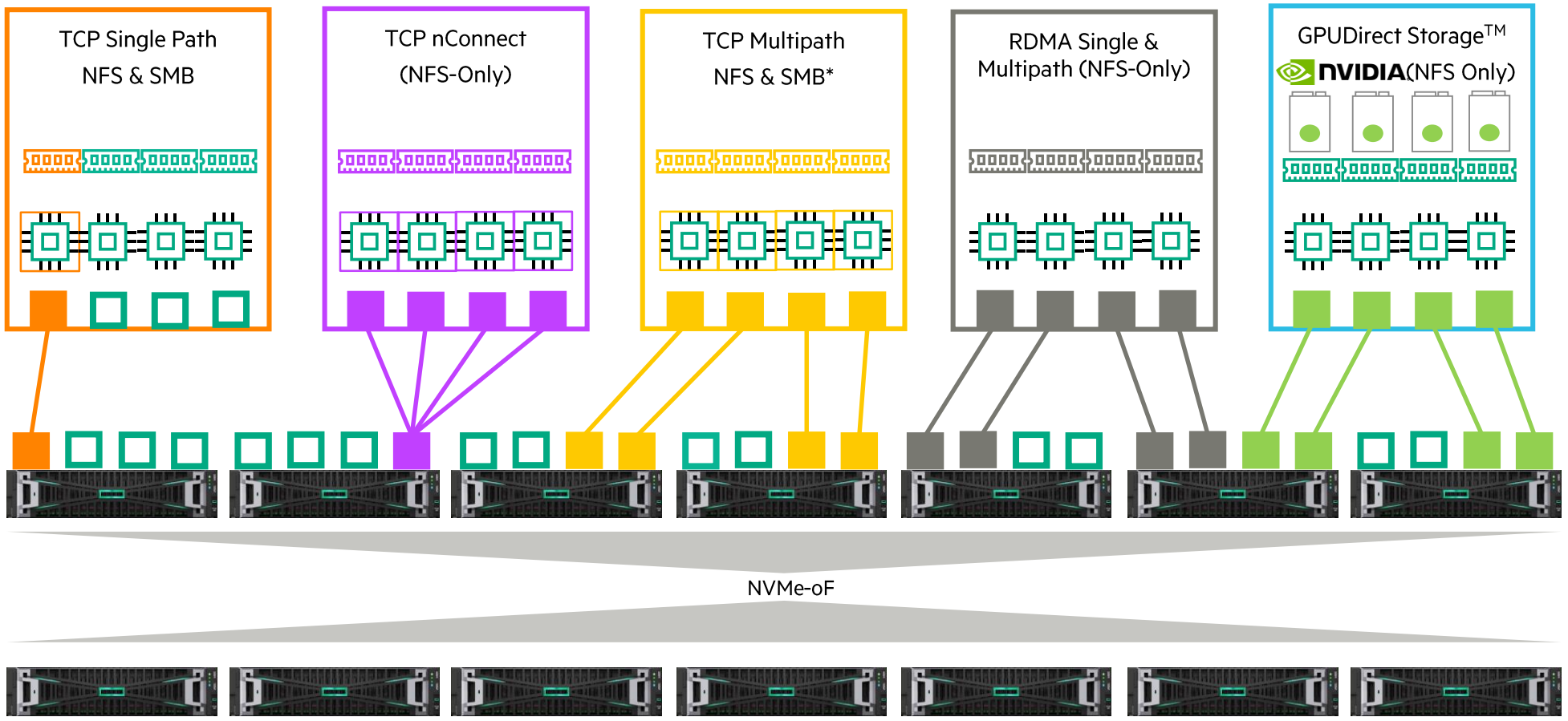
Zjednodušení
intuitivní ovládání, jako v cloudu

Rozšíření
použití pro většinu scenářů s nestrukturovanými daty

Tato architektura se náramně hodí pro všechny fáze AI projektu



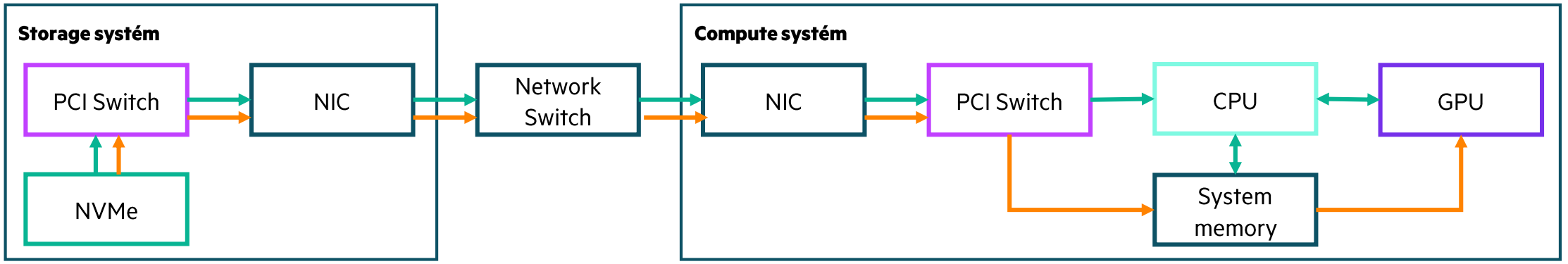
K datům se dá jednoduše dostat různými NAS scenáři současně



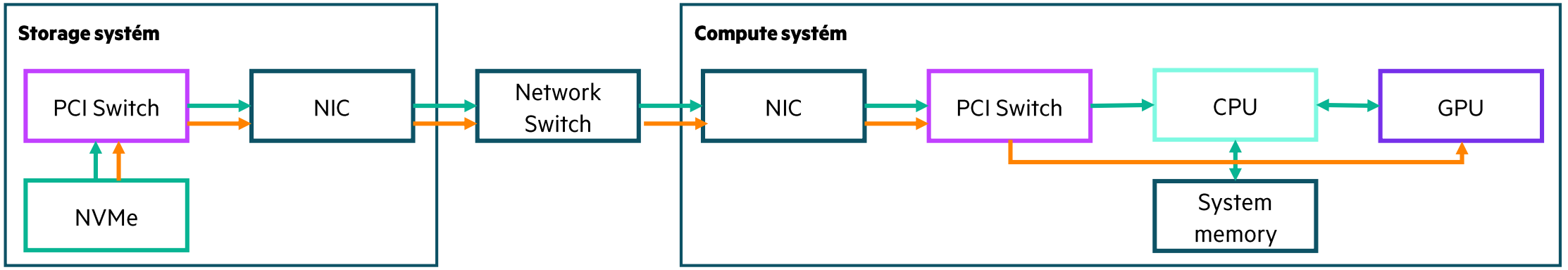
* SMB multichannel

Data jedou z pole rovnou do GPU paměti – to je Nvidia GPUDirect®

Tradiční RDMA datový tok



Nvidia GPUDirect datový tok



Signalizace
Datový tok

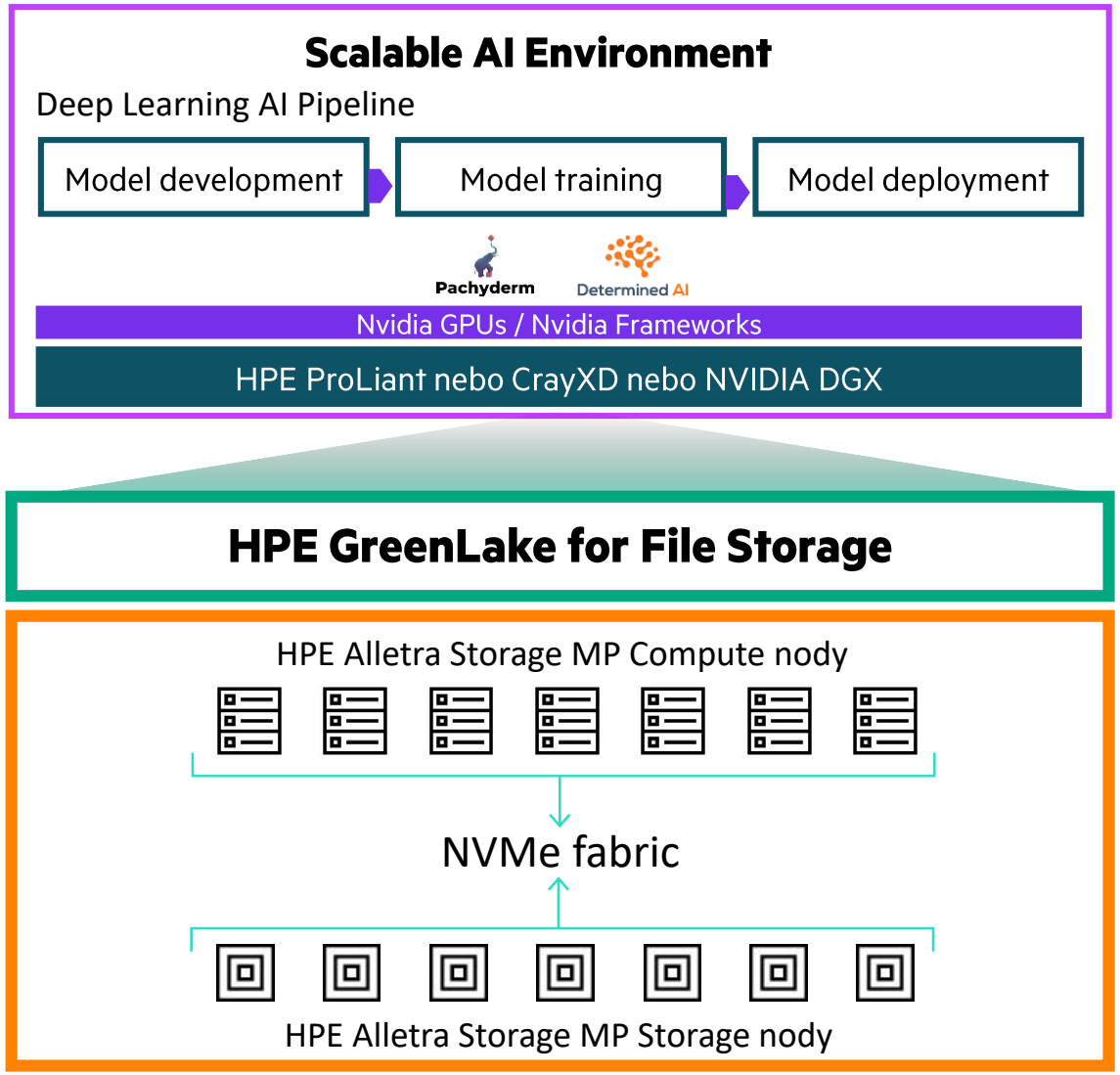
GPU Direct Storage násobně snižuje end-to-end latenci

GPUDirect Storage významně navyšuje datovou propustnost

HPE GL4F představuje skvělý datový podvozek. S certifikací DGX BasePOD

Bezpečné úložiště (nejen) pro NVIDIA BasePOD

- Fialový obdélník může být řada věcí, např.:
  - HPE Machine Learning Platform
    - MLDM, MLDE, MLIS
    - Obecná HW infrastruktura, HPE Proliant, CrayXD
  - Referenční architektura NVIDIA DGX BasePOD
  - **HPE Private Cloud AI**



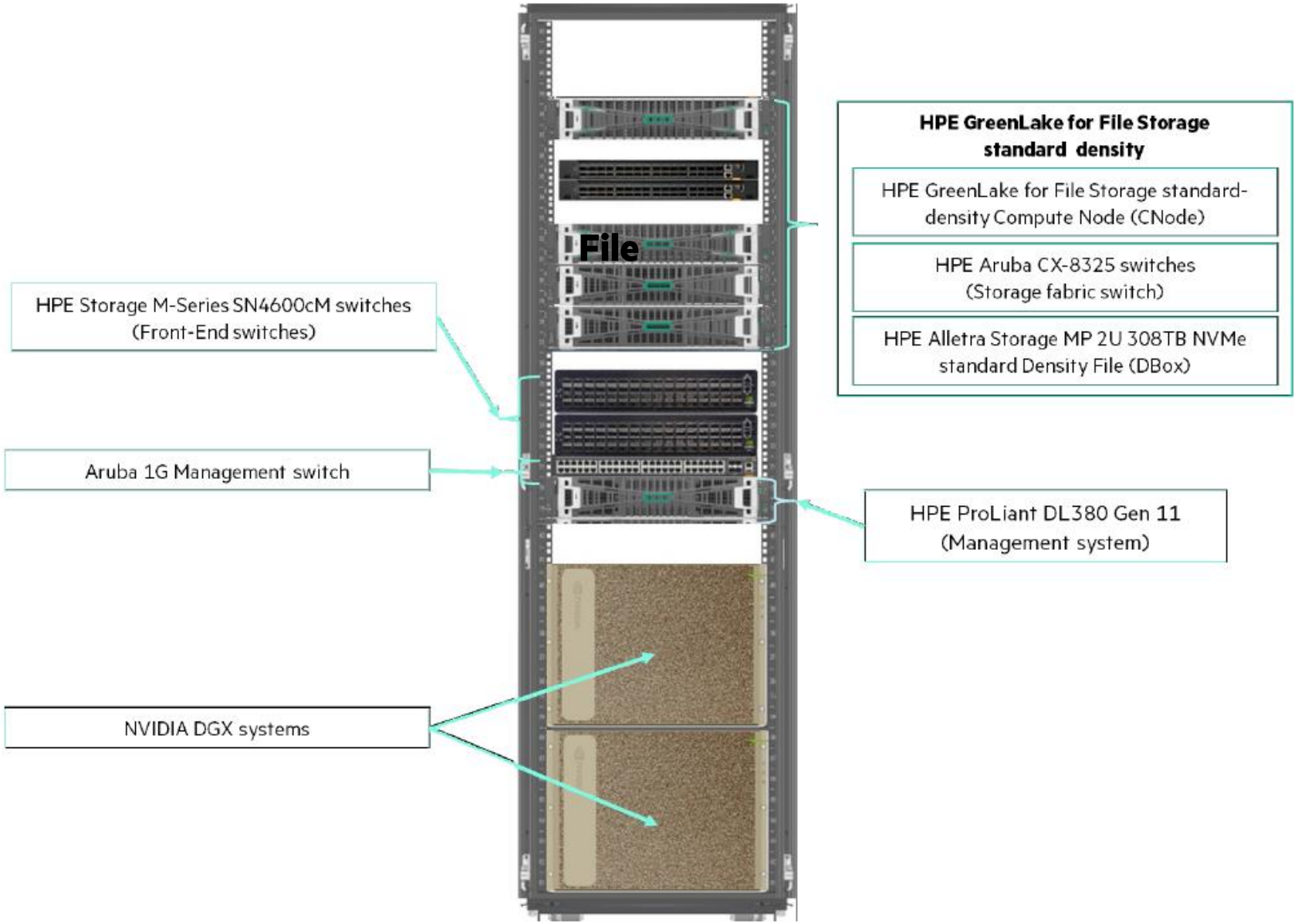
NVIDIA DGX BasePOD s HPE GreenLake for File Storage

HPE GreenLake for File Storage
Certifikovaná a bezpečná úložná komponenta

Akcelerovaná compute vrstva – NVIDIA DGX H200
NVIDIA Base Command
NVIDIA AI Enteprise

Data Network Fabric – HPE M-series SN4600 switche

Backend Fabric – HPE Aruba CX8325 switche



HPE Private Cloud AI

Vlajková loď HPE pro provoz/inferenci genAI scenářů

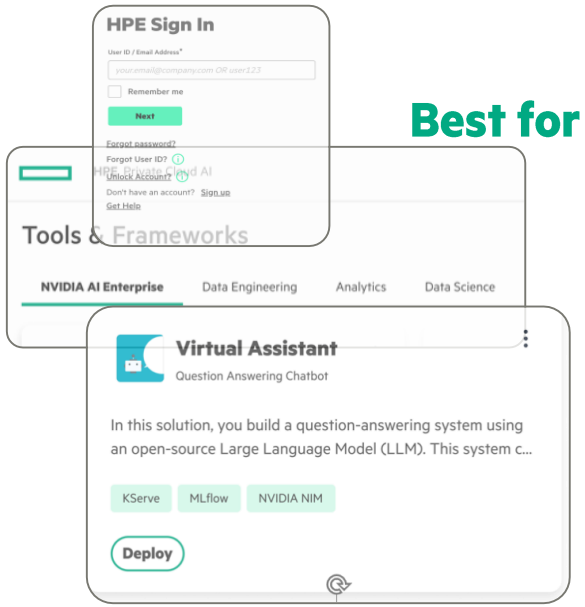
NVIDIA AI Computing by HPE

HPE Private Cloud AI

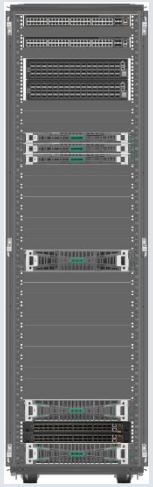
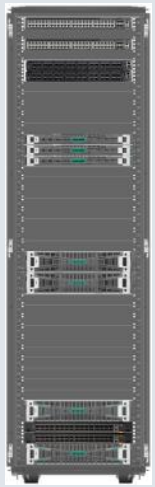
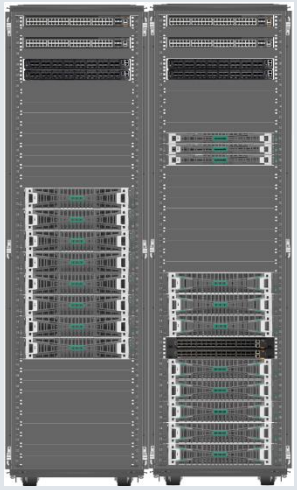
Hewlett Packard Enterprise | NVIDIA

HPE Private Cloud AI – end-to-end stack optimalizovaný pro AI.

Jednotná zkušenost prostřednictnvím HPE GreenLake platform

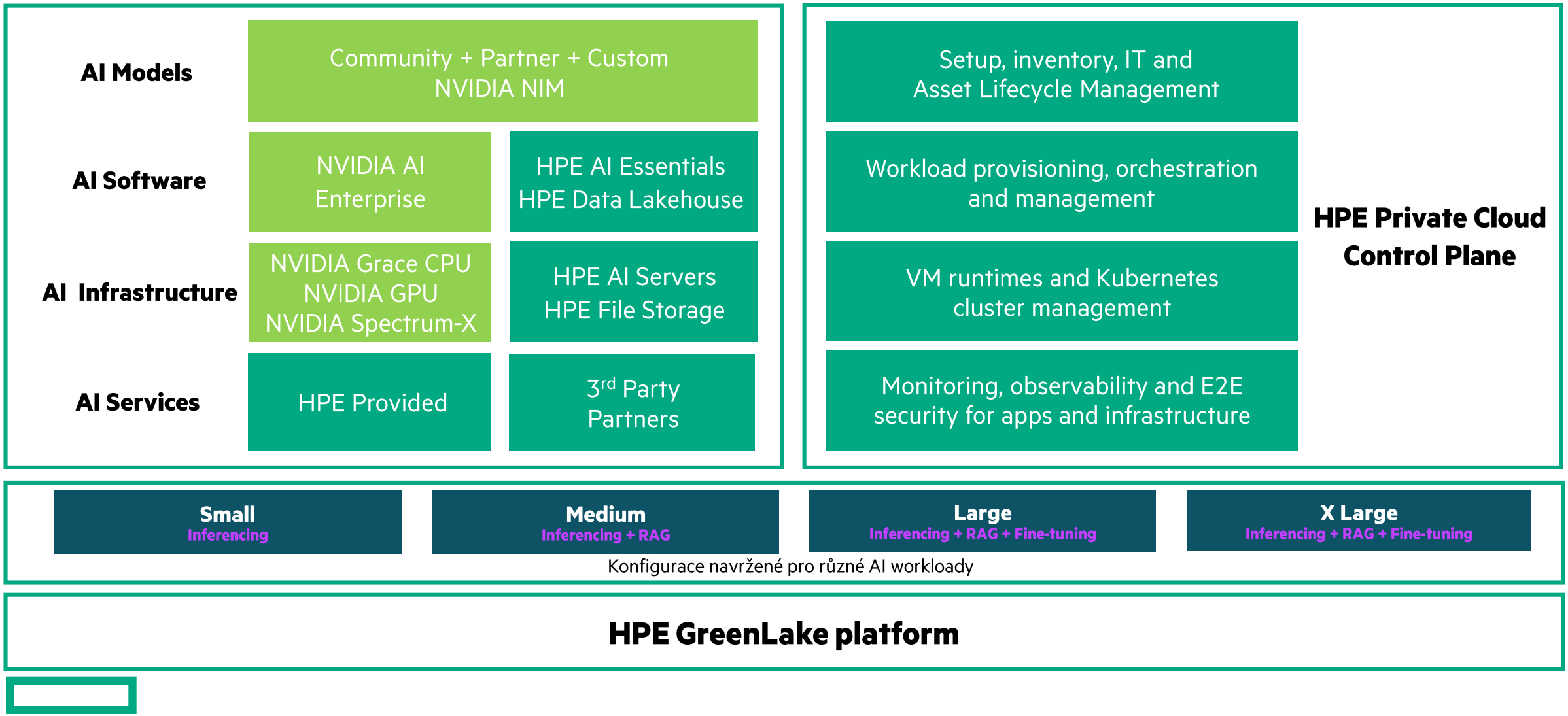


Best for

	Inferencing	Inferencing + RAG	Inferencing + RAG + Fine-tuning
			
	Small	**Medium**	**Large**
Compute	4 or 8 NVIDIA L40S GPUs	8 or 16 NVIDIA L40S GPUs	16 or 32 NVIDIA H100 NVL GPUs
Storage	109TB File/Object	217TB File/Object	670TB File/Object
Networking	100GbE NVIDIA Networking	200GbE NVIDIA Networking	400GbE NVIDIA Networking
Power	up to 8 kW rack	up to 17.7 kW	up to 25 kW x 2

NVIDIA AI Computing by HPE

HPE Private Cloud AI



HPE a business AI jde dobře dohromady.

Technologie i ucelené systémy (nejen) pro AI factory

- Akcelerované compute technologie jako standardní základ
- Vychytaná řešení ukládání dat pro všechny fáze AI projektu – např. **HPE GreenLake for File**
- K tomu specifická softwarová řešení z vlastní dílny – **HPE Machine Learning Platform**
- A samozřejmě předchystaná řešení – jak referenční architektury např. pro Nvidia BasePOD, tak hotové appliance pro provoz (inferenci) AI scenářů
  – **HPE Private Cloud AI**

Děkuji za pozornost!

ladislav.pecen@hpe.com