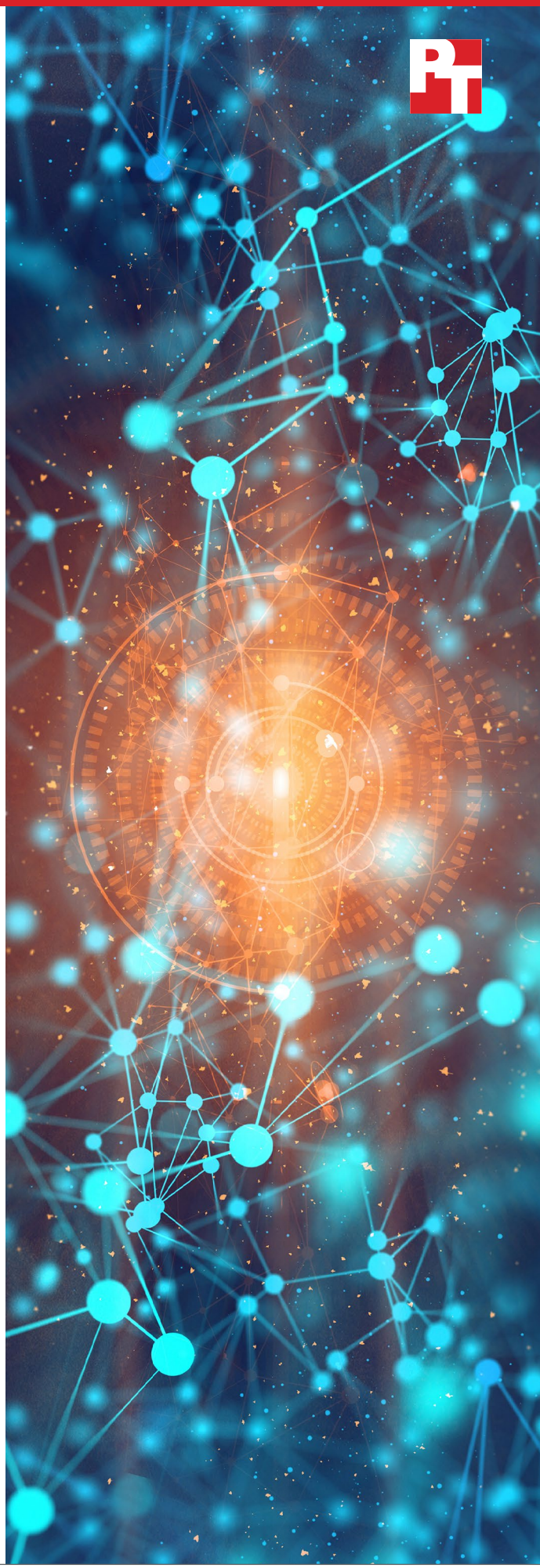
Droga do sukcesu w zakresie sztucznej inteligencji dzięki ofercie rozwiązań AI firmy Dell

Porównanie oferty rozwiązań Dell w zakresie sztucznej inteligencji z podobnymi ofertami Supermicro

Sztuczna inteligencja (AI) to nowa granica, która ma zmienić działalność biznesową w różnych branżach. Chociaż organizacje z różnych sektorów dociekają, w jaki sposób mogą wykorzystać sztuczną inteligencję (AI) do poprawy swoich operacji biznesowych, ważne jest, aby pamiętać, że wdrażanie AI i czerpanie korzyści z jej zastosowania nie dzieje się z dnia na dzień. Każda firma jest wyjątkowa, dlatego każda musi ocenić swoje dane i cele biznesowe, aby zobaczyć, w jaki sposób wykorzystanie sztucznej inteligencji może przynieść określone pożądane rezultaty. Współpraca z firmą taką jak Dell, która oferuje kompleksowe portfolio sztucznej inteligencji, w tym planowanie, przygotowywanie danych, odpowiedni wybór sprzętu, projektowanie modeli sztucznej inteligencji, testowanie koncepcji, architektury referencyjne i kompleksowa pomoc techniczna, mogą że przełożyć się na udane projekty związane ze sztuczną inteligencją.

Ze względu na wiele opcji dostępnych na rynku znalezienie partnera, który może pomóc we wszystkich tych decyzjach, może oznaczać różnicę między udanym wdrożeniem sztucznej inteligencji a kosztownym błędem. W tym artykule przyjrzymy się bliżej portfolio rozwiązań Dell i Supermicro w domenie sztucznej inteligencji w celu wyjaśnienia czytelnikom korzyści, jakie firma Dell może zapewnić w trakcie wdrażania sztucznej inteligencji przez klienta. Najpierw skupimy się na opcjach serwerów i maszyn obliczeniowych, gdzie każda z firm ma szeroką i zróżnicowaną ofertę dla klientów. Następnie dowiemy się, w jaki sposób firma Dell wykracza poza kwestie sprzętowe w przypadku firm poszukujących rozwiązań dla sektora edukacji, usług planowania, ekosystemów partnerskich i nie tylko.



Serwery i wyniki dotyczące wydajności obciążeń roboczych związanych ze sztuczną inteligencją

Serwery, czyli podstawowe elementy infrastruktury obliczeniowej, która obsługuje obciążenia robocze związane ze sztuczną inteligencją, mogą wykorzystywać procesory CPU, GPU lub jedne i drugie jako zasoby obliczeniowe w zależności od rozmiaru lub typu obciążenia roboczego. W przypadku większych lub bardziej wymagających obciążeń roboczych, takich jak HPC lub sztuczna inteligencja, procesory graficzne zapewniają najwyższą wydajność. Układy graficzne są dostępne w różnych formatach, w tym uniwersalnym PCIe, Open Compute Project Accelerator Module (OAM) oraz w autorskiej architekturze NVIDIA SXM, która obecnie zapewnia najwyższą wydajność.[1] Duże pojemności pamięci i cechy konstrukcyjne serwera, takie jak architektura chłodzenia i wydajność energetyczna, również wpływają na wydajność. Większość centrów danych nadal korzysta z chłodzenia powietrzem, co oznacza, że obciążenia robocze związane ze sztuczną inteligencją wymagają serwerów zbudowanych z myślą o jak najbardziej skutecznym chłodzeniu powietrzem. Poniżej przedstawiamy ofertę serwerów Dell PowerEdge z podziałem na podzespoły, opcje chłodzenia i inne elementy, a także opublikowane wyniki MLCommons® MLPerf®.

Wyniki testu

MLPerf® to pakiet testów porównawczych, który testuje wydajność sztucznej inteligencji zarówno na potrzeby trenowania, jak i wnioskowania. Aby organizacja mogła opublikować oficjalne wyniki MLPerf®, muszą być one zgodne z warunkami określonymi przez firmę odpowiedzialną za programowanie testu porównawczego — MLCommons®.[2] Te wytyczne dotyczące zgodności zawierają standardy, które ułatwiają porównywanie wyników. Do testowania wnioskowania MLPerf® używa zestawów danych Datacenter, Edge, Mobile i Tiny oraz raportuje wyniki sztucznej inteligencji i waty energii zużytej podczas testowania. Pakiet testów porównawczych wnioskowania obejmuje testowanie wielu popularnych modeli sztucznej inteligencji, uczenia maszynowego i uczenia głębokiego (patrz tabela 1).

Tabela 1. Modele sztucznej inteligencji, uczenia maszynowego i uczenia głębokiego zawarte w testach MLPerf® i typowe przypadki użycia dla każdego z nich. Źródło: Principled Technologies.

Typowe modele sztucznej inteligencji	Typowe zastosowania
ResNet	Model klasyfikacji obrazów, który pomaga komputerom się uczyć, zapamiętywać i identyfikować różne obrazy na potrzeby przypadków użycia, takich jak obrazowanie medyczne, moderowanie treści w mediach społecznościowych i rozpoznawanie twarzy.
RetinaNet	Typ wykrywania obiektów, który może obsłużyć większą złożoność w porównaniu z ResNet. Pomaga komputerom identyfikować i lokalizować obiekty na obrazach lub w klatkach wideo i klasyfikować je według ważności. Używany w takich dziedzinach jak autonomiczna jazda, technologia automatycznego wspomagania pojazdów, monitoring, rozpoznawanie twarzy.
3D-UNet	Właściwe dla segmentacji obrazów medycznych
RNN-T	Rozpoznawanie mowy w przypadkach użycia takich jak tłumaczenie automatyczne
BERT	Przetwarzanie języka naturalnego w przypadkach użycia takich jak streszczanie tekstu, tłumaczenie i automatyczne wykonywanie zadań
DLRM-v2-99.9	Model rekomendacji dla przypadków użycia takich jak ukierunkowane reklamy i spersonalizowane rekomendacje produktów
GPTJ-99 oraz 99.9	Duży model językowy (LLM) do przetwarzania języka naturalnego, który doskonale sprawdza się w generowaniu tekstu w przypadkach użycia takich jak chatboty i narzędzia SI oparte na czacie

Informacje o MLPerf

Wyniki MLPerf® obejmują kilka parametrów oprócz samych modeli sztucznej inteligencji, co może sprawić, że wiele danych będzie trzeba przeanalizować na jednym wykresie lub w jednej tabeli. Oto krótki opis tych parametrów:

- 99.0 i 99.9: te liczby odnoszą się do dokładności, do której model został wytrenowany. Im dokładniejsze mają być dane wyjściowe, tym bardziej złożony jest model i tym dłużej może trwać przetwarzanie danych.
- Próbki offline/s: tryb, w którym test porównawczy wysyła wszystkie zapytania na początku testu, symulując dane już obecne w systemie.
- Zapytania serwera/s: tryb, w którym test porównawczy wysyła zapytania przez cały czas trwania testu, symulując analizowanie strumienia danych na żywo.

Aby uzyskać więcej informacji na temat wyników MLCommons® i MLPerf®, zobacz https://mlcommons.org/benchmarks/inference-datacenter/.

Wyniki zawarte w tym raporcie pochodzą z wyników MLPerf® v3.1 Inference Datacenter opublikowanych na stronie MLCommons® w listopadzie 2023 r.[3] Wyniki te obejmują zgłoszenia od producentów technologii i dostawców usług w chmurze i obejmują szereg konfiguracji. W porównaniu z publicznie dostępnymi informacjami od Supermicro serwery Dell PowerEdge uzyskały porównywalne wyniki. Tabela 2 zawiera szczegółowe informacje o serwerze.

Tabela 2. Serwery Dell i Supermicro zawarte w wynikach MLCommons® MLPerf® 3.1 results published as of November 2023. MLPerf 3.1 opublikowanych w listopadzie 2023 r. Źródło: Principled Technologies.

Zgłaszający	Model serwera	Liczba i model procesorów graficznych	Opis
Dell[4]	PowerEdge XE9680	8 × NVIDIA H100 SXM	Do trenowania sztucznej inteligencji i wnioskowania przy dużych obciążeniach, takich jak duże modele językowe
	PowerEdge XE9640	4 × NVIDIA H100 SXM	Do trenowania dużych modeli sztucznej inteligencji w centrach danych o dużej gęstości chłodzonych cieczą
	PowerEdge XE8640	4 × NVIDIA H100 SXM	Do obsługi tradycyjnych szkoleń AI, HPC i aplikacji do analizy danych w obudowie 4U dla centrów danych chłodzonych powietrzem
Supermicro[5]	AS-8125GS-TNHR	8 × NVIDIA H100 SXM	Do trenowania sztucznej inteligencji na dużą skalę i obciążeń roboczych HPC z procesorami AMD
	SYS-821GE-TNHR	8 × NVIDIA H100 SXM	Do trenowania sztucznej inteligencji na dużą skalę i obciążeń roboczych HPC z procesorami Intel
	SYS-421GU-TNXR	4 × NVIDIA H100 SXM	Modułowa konstrukcja zapewniająca elastyczność obsługi obciążeń roboczych HPC i sztucznej inteligencji

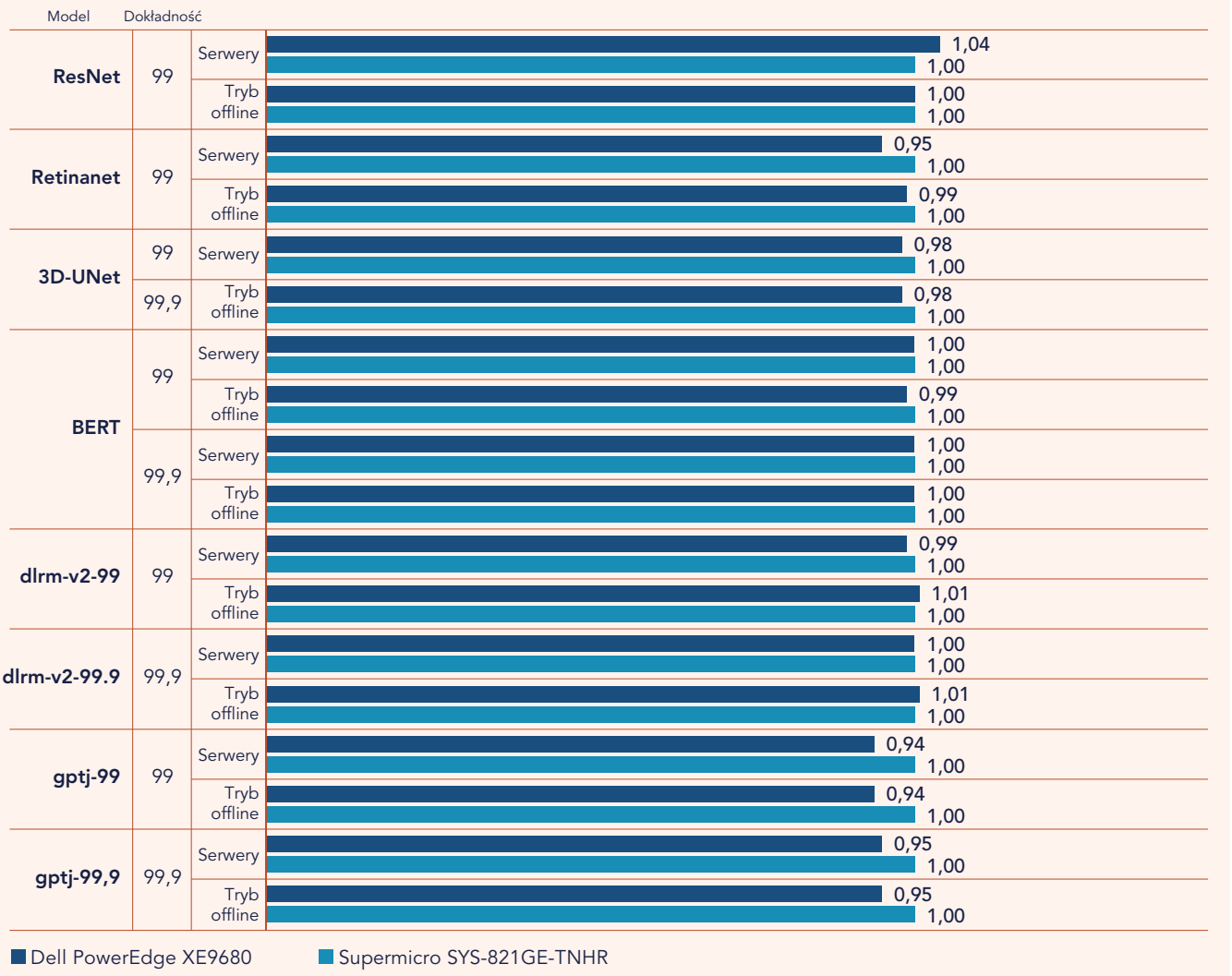
Ponieważ zarówno firma Dell, jak i Supermicro przedstawiły wyniki z identycznymi konfiguracjami układów graficznych, porównanie wydajności jest proste. Jak pokazują rysunki 1 i 2, wspólne konfiguracje obu dostawców prowadzą do wyników, które w dużej sobie dorównują. W przypadku innych konfiguracji, takich jak te pokazane na ilustracjach 3 i 4, firma Dell przewyższyła Supermicro w modelu gptj-99.9 w testach 4-GPU. Należy pamiętać, że o ile firma Dell przesłała wyniki dla wszystkich dostępnych modeli ze wszystkimi trzema serwerami, firma Supermicro tego nie zrobiła. Porównamy tylko modele, dla których oba serwery mają wyniki. Aby zobaczyć pełny zestaw wyników firmy Dell, odwiedź stronę wyników MLCommons® MLPerf®.

Wyniki z ośmioma procesorami graficznymi

Serwer Dell PowerEdge XE9680 obsługuje do ośmiu procesorów graficznych NVIDIA H100 SXM5 w celu przyspieszenia SI oraz maksymalnie dwa skalowalne procesory Intel® Xeon® czwartej generacji. Rodzina produktów PowerEdge XE ma modułową architekturę obsługującą procesory graficzne NVIDIA SXM4 lub SXM5 bądź zestawy procesorów graficznych Open Compute Project Accelerator Module (OAM), co może zwiększyć wydajność w porównaniu ze standardowymi procesorami graficznymi PCIe. Serwer Dell PowerEdge XE9680 jest również wyposażony w akcelerator AMD Instinct™ MI300X.[6] Model PowerEdge XE9680 to kompaktowy serwer z ośmioma procesorami NVIDIA H100 SXM5 zajmujący zaledwie 6U miejsca w szafie serwerowej.

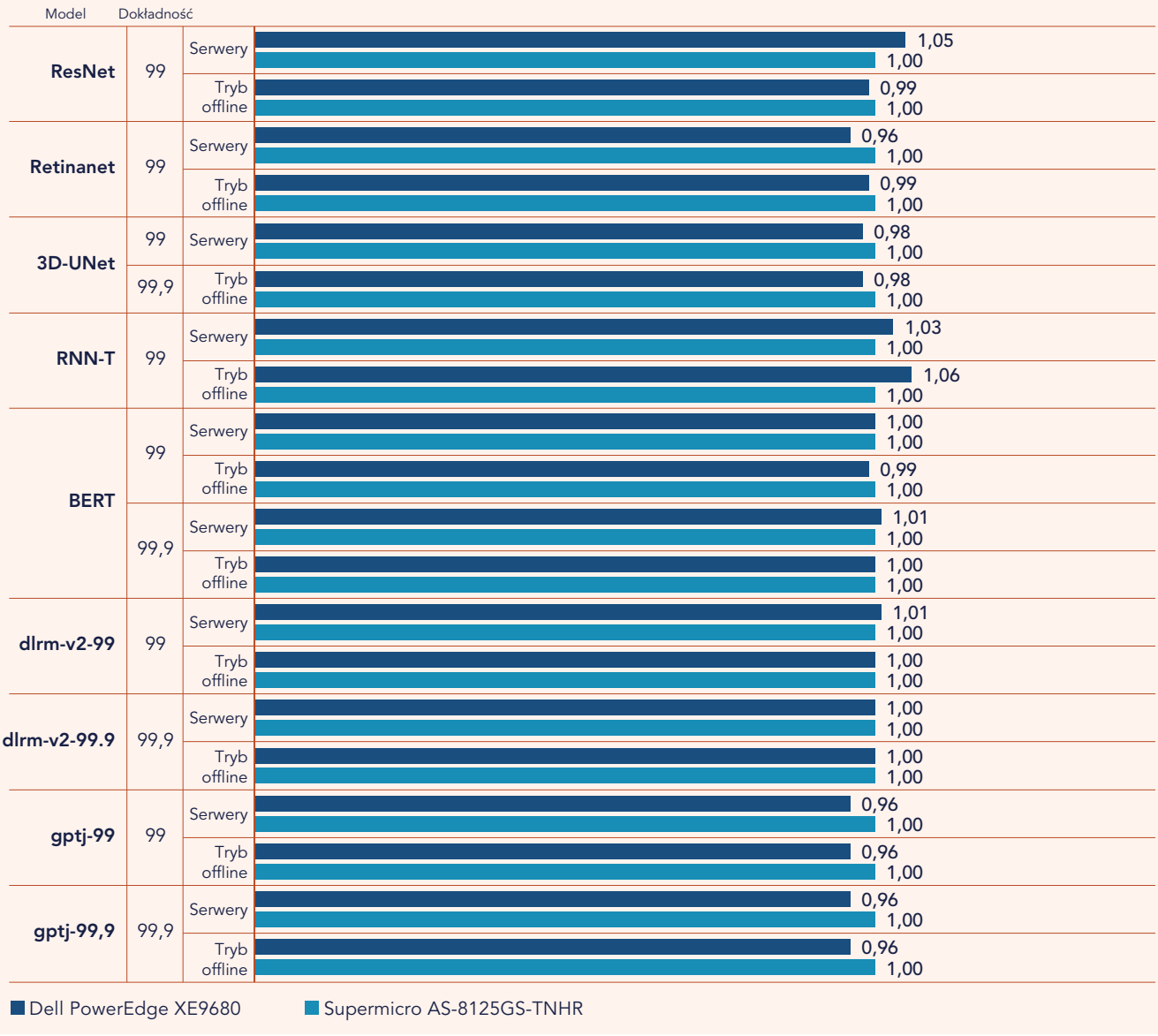
Natomiast serwer Supermicro z 8 GPU wymaga o 33% więcej miejsca w szafie serwerowej przy 8U, a jednocześnie oferuje format SXM dla procesorów graficznych NVIDIA. Różnica w rozmiarze oznacza, że można zmieścić siedem serwerów Dell PowerEdge XE9680 w szafie serwerowej w porównaniu z tylko pięcioma serwerami Supermicro. Na rysunkach 1 i 2 porównaliśmy wyniki serwera Dell PowerEdge XE9680 z dwiema konfiguracjami serwera Supermicro 8-GPU: SYS-821GE-TNHR z procesorami Intel i AS-8125GS-TNHR z procesorami AMD. Zwracamy uwagę, że firma Supermicro nie przesłała wyników dla RNN-T na SYS-821GE-TNHR, dlatego wykluczamy ten model z wykresu na rysunku 1.

Znormalizowane wyniki MLPerf®: Dell PowerEdge XE9680 z procesorem H100 SXM5 w porównaniu z Supermicro SYS-821GE-TNHR z procesorem H100 *(im większa wartość, tym lepiej)*



Rysunek 1. Opublikowano wyniki MLPerf® dla serwerów Dell PowerEdge XE9680 i Supermicro SYS-821GE-TNHR według stanu na dzień 29/11/2023. Oba układy wykorzystują format SXM procesorów graficznych NVIDIA H100. Źródło: Principled Technologies na podstawie danych z MLCommons®.[7,8]

Znormalizowane wyniki MLPerf®: Dell PowerEdge XE9680 z procesorem H100 SXM5 w porównaniu z Supermicro AS-8125GS-TNHR z procesorem H100 SXM5 *(im większa wartość, tym lepiej)*



Rysunek 2. Opublikowano wyniki MLPerf® dla serwerów Dell PowerEdge XE9680 i Supermicro AS-8125GS-TNHR według stanu na dzień 29/11/2023. Oba układy wykorzystują format SXM procesorów graficznych NVIDIA H100. Źródło: Principled Technologies na podstawie danych z MLCommons®.[9,10]

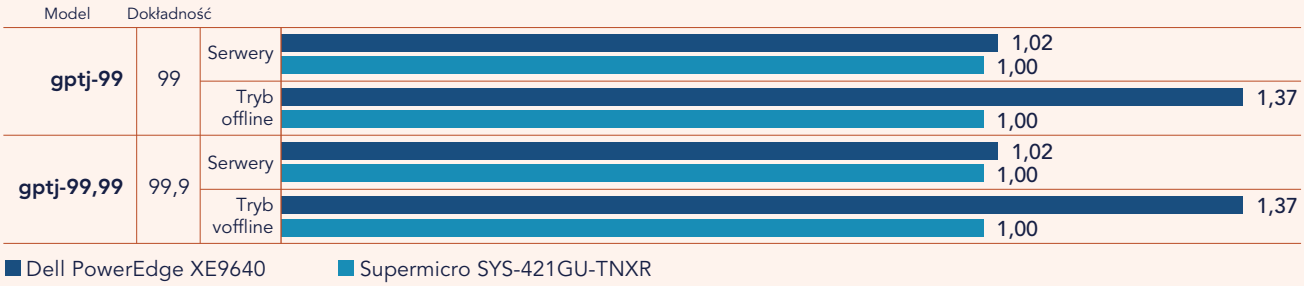
Jak pokazują te wyniki, wybór serwera Dell PowerEdge XE9680 zapewni podobną wydajność w przypadku wielu obciążeń roboczych związanych z wnioskowaniem AI przy jednoczesnym zajmowaniu mniejszej ilości miejsca w centrum przetwarzania danych.

Wyniki z czterema procesorami graficznymi

Model Dell PowerEdge XE9640 2U może być właściwym wyborem, gdy kluczowe znaczenie ma oszczędność energii lub miejsca w centrum danych. Dzięki maksymalnie czterem procesorom graficznym NVIDIA H100 SXM serwer XE9640 zapewnia o połowę mniejszą moc obliczeniową niż model XE9680 na przestrzeni o dwie trzecie mniejszej.[11] Gęsto upakowany serwer Dell PowerEdge XE9640 jest wyposażony w technologię Dell Smart Cooling, która zawiera szereg rozwiązań termicznych, w tym bezpośrednie chłodzenie cieczą dla procesorów i kart graficznych.[12] Obudowa 2U serwera PowerEdge XE9640 jest wyposażona w ulepszone mechanizmy przepływu powietrza, w tym większe wentylatory i radiatory, które pomagają chłodzić inne ważne podzespoły, takie jak karty PCIe i pamięć.[13]

Supermicro oferuje starszy serwer SYS-220GQ-TNAR+, który zwiera cztery procesory graficzne NVIDIA A100 HGX w obudowie 2U, ale nie udało nam się znaleźć żadnych serwerów Supermicro 2U z czterema najnowszymi procesorami GPU H100 HGX, które pasowałyby do PowerEdge XE9640.[14] Serwer 4-GPU z procesorami GPU HGX H100 NVIDIA Supermicro w zgłoszeniach MLPerf® to SYS-421GU-TNXR, który jest serwerem 4U. Jak wspomniano wcześniej, firma Supermicro przesłała wyniki MLPerf® 3.1 tylko dla SYS-421GU-TNXR w modelu sztucznej inteligencji gptj-99.9, więc nie możemy porównać ich z PowerEdge XE9640 w innych modelach. Jednak w opublikowanych wynikach serwer PowerEdge XE9640 pokonał serwer Supermicro w testach offline, osiągając nawet 1,37-krotnie lepszy wynik (patrz rysunek 3).

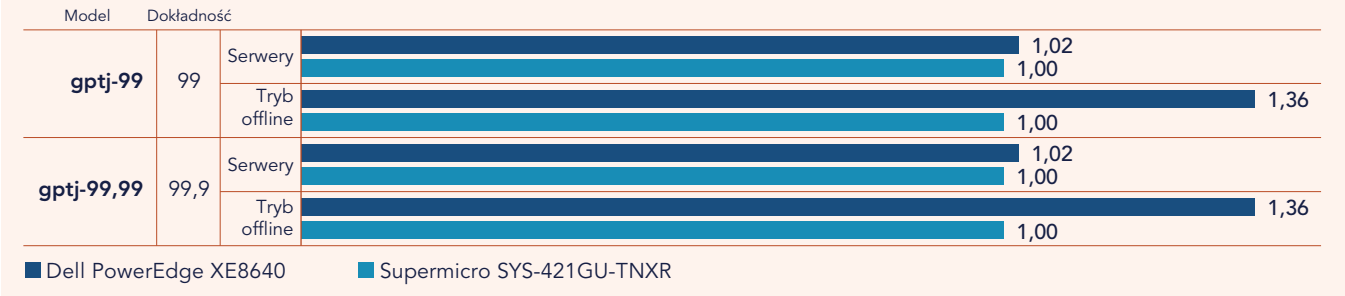
Znormalizowane wyniki MLPerf®: Dell PowerEdge XE9640 z procesorem H100 SXM5 w porównaniu z Supermicro SYS-421GU-TNXR z procesorem H100 SXM5 *(im większa wartość, tym lepiej)*



Rysunek 3. Opublikowano wyniki MLPerf® dla serwerów Dell PowerEdge XE9640 i Supermicro SYS-421GU-TNXR według stanu na dzień 29/11/2023. Oba układy wykorzystują format SXM procesorów graficznych NVIDIA H100. Źródło: Principled Technologies na podstawie danych z MLCommons®.[15,16]

Możemy również porównać wyniki serwera Supermicro SYS-421GU-TNXR z serwerem Dell PowerEdge XE8640, 4-GPU 4U, który także obsługuje procesory graficzne NVIDIA H100 HGX. Mimo że serwer PowerEdge jest większy niż PowerEdge XE9640, XE8640 nie wymaga bezpośredniego chłodzenia cieczą, co stanowi kompromis między technologią gęstości a chłodzeniem w centrach przetwarzania danych, które nie mają dostępu do chłodzenia wodą. Serwer PowerEdge XE8640 jest dostępny z układem chłodzenia procesorów powietrzem i układem chłodzenia procesorów graficznych powietrzem wspomaganym cieczą, który nie wymaga dopływu wody do szafy serwerowej.[17] Serwer Dell PowerEdge XE8640 jest wyposażony w najnowsze skalowalne procesory Intel Xeon 4. generacji i do 4 TB pamięci do obsługi dużych zbiorów danych oraz złożonych obliczeń typowych dla sztucznej inteligencji i analizy danych.[18] Dzięki obudowie 4U serwer PowerEdge XE8640 jest podobny do Supermicro SYS-421GU-TNXR zarówno pod względem gęstości, jak i możliwości GPU. Jednak podobnie jak w przypadku serwerów PowerEdge XE9640, serwer Dell PowerEdge XE8640 uzyskuje lepsze wyniki gptj-99 w testach offline niż serwer Supermicro (rysunek 4).

Znormalizowane wyniki MLPerf®: Dell PowerEdge XE8640 z procesorem H100 SXM5 w porównaniu z Supermicro SYS-421GU-TNXR z procesorem H100 SXM5 *(im większa wartość, tym lepiej)*



Rysunek 4. Opublikowano wyniki MLPerf® dla serwerów Dell PowerEdge XE8640 i Supermicro SYS-421GU-TNXR według stanu na dzień 29/11/2023. Oba układy wykorzystują format SXM procesorów graficznych NVIDIA H100. Źródło: Principled Technologies na podstawie danych z MLCommons®.[19,20]

Jak zaobserwowaliśmy, wyniki MLPerf® dotyczące wydajności serwerów Supermicro i Dell opartych na procesorach graficznych są ogólnie podobne, z zauważalną przewagą serwerów Dell PowerEdge XE8640 i PowerEdge XE9640 w jednym modelu AI.

Ponieważ wydajność to tylko jeden z aspektów wdrożenia sztucznej inteligencji, przyjrzeliśmy się również pozostałej ofercie sztucznej inteligencji firm Dell i Supermicro, od stacji roboczych, pamięci masowej i sieci po usługi, wsparcie, edukację i inne. Okazało się, że portfolio Dell AI jest szersze niż oferta Supermicro i oferuje rozwiązania odpowiadające na wiele wyzwań, z którymi firmy borykają się poza wydajnością obliczeniową.

Klienckie stacje robocze i pamięć masowa

Dodatkowe opcje obliczeniowe ze stacjami roboczymi

Niektóre przypadki użycia sztucznej inteligencji wymagają innego podejścia obliczeniowego, a nie każdy, kto potrzebuje dostępu do sprzętu zdolnego do obsługi sztucznej inteligencji, może pozostać związany z centrum danych. Naukowcy pracujący w laboratoriach mogą nie mieć miejsca na szafę serwerową, a dla osób pracujących na brzegu sieci noszenie dużego, ciężkiego systemu stacjonarnego jest niepraktyczne. W tym miejscu do gry wchodzą stacje robocze. Oferta rozwiązań Dell AI obejmuje kilka stacji roboczych Precision z obsługą sztucznej inteligencji, w tym konfiguracje w obudowach typu tower, mobilnych i montowanych w szafie serwerowej, które spełniają różne potrzeby.[21] Natomiast Supermicro nie oferuje mobilnych stacji roboczych, które zapewniają łatwy dostęp w ruchu. Oferta stacji roboczych z procesorami graficznymi składa się z kilku różnych konfiguracji typu tower, a jak twierdzi firma Supermicro, niektóre z nich użytkownicy mogą montować w szafach serwerowych.[22] Nasze badania wykazały, że istnieją dwie stacje robocze do montażu w szafie serwerowej, ale wydają się one być starsze, dostępne tylko z procesorami graficznymi NVIDIA A100 i prawdopodobnie nie są już sprzedawane.[23] Jeśli Twoja firma potrzebuje elastyczności w zakresie rodzaju i mobilności wdrażanych przez nią stacji roboczych GPU, oferta rozwiązań Dell AI jest lepiej dopasowana do tych potrzeb.

Kwestie dotyczące magazynu (pamięci masowej)

Pamięć masowa może być równie ważna jak moc obliczeniowa w kontekście obsługi obciążeń związanych ze sztuczną inteligencją. Więcej danych zwiększa dokładność modelu sztucznej inteligencji, ale przechowywanie ogromnych zbiorów danych i zarządzanie nimi może stanowić wyzwanie dla wielu centrów danych. Ponadto systemy pamięci masowej gotowe do pracy ze sztuczną inteligencją muszą z łatwością obsługiwać wiele różnych typów danych, ponieważ modele są zwykle trenowane przy użyciu danych nieustrukturyzowanych.[24] Aby zapewnić pojemność i skalowalność zestawów danych sztucznej inteligencji, uczenia maszynowego i uczenia głębokiego, firma Dell oferuje serię PowerScale™ do przechowywania plików oraz rozwiązanie Elastic Cloud Storage (ECS) lub definiowane programowo narzędzie ObjectScale do obiektowej pamięci masowej.

Oferta serwerów NAS PowerScale typu all-flash firmy Dell obejmuje opcje pojemności od 3,84 TB do 720 TB pojemności fizycznej na węzeł, przy czym klastrowane pojemności all-flash osiągają 186 PB pojemności fizycznej. Dzięki elastyczności i skalowalności PowerScale może obsługiwać szeroką gamę klientów i przypadków użycia sztucznej inteligencji.[25] Wszystkie trzy modele PowerScale typu all-flash — F200, F600 i F900 — obsługują wbudowaną kompresję oraz deduplikację danych w celu zwiększenia wydajności pamięci masowej.[26] Każdy model pamięci masowej PowerScale korzysta z systemu plików Dell OneFS™, który wykorzystuje zaawansowane zasady tworzenia warstw dające pewność, że najczęściej wykorzystywane dane znajdują się na najbardziej wydajnych warstwach pamięci masowej.[27] Firma Dell oferuje również oprogramowanie OneFS w AWS Marketplace z pamięcią masową APEX File Storage dla platformy AWS. Klienci mogą korzystać z OneFS z instancjami obliczeniowymi AWS, aby zapewnić spójne środowisko użytkownika z tymi samymi funkcjami dostępnymi w lokalnych macierzach OneFS.[28]

Oferta pamięci masowej Supermicro obejmuje serwery pamięci masowej — montowane w szafie serwerowej, gęste serwery pamięci masowej o różnych rozmiarach i gęstości.[29] Aby uzyskać pamięć masową do przechowywania plików od firmy Supermicro, klienci muszą wybierać spośród różnych ofert pamięci masowej zdefiniowanej programowo innych firm, takich jak WekaIO, Scality RING lub OSNEXUS.[30] Podczas gdy Scality RING i OSNEXUS zawierają opcje przechowywania plików w opisach platform, WekaIO wydaje się być podstawową opcją dla klientów szukających podstawowej pamięci masowej plików. Supermicro oferuje kilka architektur referencyjnych obejmujących szeroką gamę zastosowań, ale klienci muszą mieć subskrypcję lub licencję na oprogramowanie WekaIO, co może zwiększyć ogólny koszt rozwiązania.[31]

Opcje obiektowej pamięci masowej firmy Dell obejmują rozwiązanie Dell ECS Enterprise Object Storage, które „stworzono specjalnie do przechowywania nieustrukturyzowanych danych w skali chmury publicznej”.[32] Wraz z wbudowaną kompatybilnością z obiektową pamięcią masową Amazon S3 umożliwiającą korzystanie z funkcjonalności chmury hybrydowej węzły pamięci masowej ECS zapewniają pojemność do 14 PB na szafę.[33] Podobnie jak w przypadku przechowywania plików oferty przechowywania obiektów Supermicro wymagają konfiguracji innej firmy. Platforma OSNEXUS to połączona platforma pamięci masowej do przechowywania plików, bloków i obiektów, podczas gdy rozwiązanie Scality RING łączy pamięć masową plików i obiektów; jedno i drugie wymaga licencji od dostawców zewnętrznych.[34][35] Do samego przechowywania obiektów klienci mogą kupić rozwiązanie Supermicro dla Quantum ActiveScale do przechowywania obiektów w chmurze prywatnej z subskrypcją oprogramowania Quantum.[36] W chwili pisania tego tekstu nie udało nam się znaleźć żadnych elastycznych opcji zużycia/wydatków operacyjnych oferowanych przez Supermicro.

Ponieważ oferta pamięci masowej Supermicro wymaga od klientów współpracy z zewnętrznymi dostawcami oprogramowania, mogą oni ponieść dodatkowe koszty licencji lub subskrypcji i mogą napotkać trudności w zakresie pomocy technicznej, rozwiązywania problemów i nie tylko. Oferta rozwiązań AI firmy Dell zapewnia klientom pamięci masowych pojedyncze, niezawodne rozwiązanie w zakresie usług i wsparcia w każdym aspekcie ich rozwiązania pamięci masowej.

Więcej informacji o rozwiązaniach Dell APEX

Dla klientów, którzy chcą korzystać z pamięci masowej plików jako usługi, firma Dell oferuje usługi APEX Data Storage Services, które obejmują pamięć masową plików, bloków i kopii zapasowych. Korzystając z konsoli Dell APEX Console, klienci mogą zamawiać nowe subskrypcje, dostosowywać i monitorować pojemność pamięci masowej i nie tylko. Według firmy Dell rozwiązanie to umożliwia „uzyskanie łatwości i zwinności obsługi chmury dzięki większej kontroli nad aplikacjami i danymi”.[37]

Aby dowiedzieć się więcej na temat rozwiązań Dell APEX, odwiedź stronę https://www.dell.com/en-us/dt/apex/storage/data-storage-services/index.htm

Opcje sieci

Sieć jest kolejnym ważnym elementem infrastruktury sztucznej inteligencji. Ponieważ wiele obciążeń roboczych związanych ze sztuczną inteligencją działa na dużych klastrach serwerów, które wymagają stałej komunikacji między sobą i pamięcią masową, potrzebują one solidnej sieci, aby uniknąć wąskich gardeł. Jeśli wydajność sieci jest niewystarczająca dla obciążeń związanych ze sztuczną inteligencją, czas trenowania i wnioskowania wydłuży się, spowalniając przetwarzanie danych i uzyskiwanie szczegółowych informacji. Firma Dell oferuje przełączniki PowerSwitch Data Center montowane na górze szafy serwerowej (ToR) oraz moduły we/wy PowerEdge MX do sieci Ethernet i sieci szkieletowych.[38] Oferta PowerSwitch obejmuje zakres od 1 GbE do 400 GbE, aby sprostać różnym potrzebom. Ponadto przełączniki Dell PowerSwitch z serii Z oferują połączenia 100 GbE i 400 GbE zoptymalizowane pod kątem struktur typu leaf/spine.[39]

Firma Supermicro oferuje również przełączniki z portami Ethernet do 400 GbE dla przełączników ToR i innych zastosowań, takich jak centrum danych w strukturze spine i leaf.[40] Ponadto usługi sieciowe Della zapewniają kilka korzyści związanych z łatwością użytkowania i elastycznością, których Supermicro nie oferuje. Usługi takie jak Dell Fabric Design Center mogą pomóc uniknąć niedopasowania sieci, luk lub nieefektywności, pomagając klientom w planowaniu i wdrażaniu sieci z automatyzacją.[41] W przypadku określonych środowisk, takich jak konfiguracje VMware VxRail, VMware ESXi i Dell PowerStore, firma Dell oferuje usługi SmartFabric, które umożliwiają wdrażanie infrastruktury zdefiniowanej programowo i zarządzanie cyklem życia. Dzięki PowerStore usługi SmartFabric Services mogą zautomatyzować „nawet 99% zadań związanych z łącznością LAN poprzez zastosowanie sieci szkieletowej typu plug-and-play".[42] Usługi takie jak te zapewniające automatyzację, wskazówki i nie tylko w zakresie projektowania i wdrażania sieci, wspierają klientów podczas procesu adopcji sztucznej inteligencji.

Usługi, szkolenia i nie tylko

Głównym wyzwaniem niezwiązanym ze sprzętem we wdrażaniu sztucznej inteligencji jest potrzeba wewnętrznej wiedzy eksperckiej w zakresie strategii, planowania, przygotowywania danych i zarządzania nimi. Zarządzanie i utrzymywanie obciążeń roboczych związanych ze sztuczną inteligencją wymaga unikatowych zasobów wiedzy, w tym zarówno bardziej tradycyjnej wiedzy eksperckiej w zakresie sprzętu, jak i operacji uczenia maszynowego i analizy danych. Osoby projektujące i wdrażające strategię sztucznej inteligencji potrzebują również głębokiego zrozumienia unikalnych celów operacyjnych firmy, aby nowe obciążenia robocze związane ze sztuczną inteligencją spełniały te cele.[43]

Kolejną istotną przeszkodą może być bezproblemowa integracja sztucznej inteligencji z istniejącymi systemami operacyjnymi. Ta integracja wymaga strategicznego dopasowania nowych technologii sztucznej inteligencji do obecnych procesów biznesowych, zapewniającego, że wprowadzenie sztucznej inteligencji nie zakłóca ustalonych przepływów pracy. Współpraca z firmą taką jak Dell, która zapewnia kilka zoptymalizowanych, zweryfikowanych architektur referencyjnych rozwiązań, kursów, opcji zarządzania i dużego ekosystemu partnerskiego, może ułatwić wdrożenie sztucznej inteligencji.

Usługi profesjonalne w zakresie sztucznej inteligencji

W zakresie edukacji i planowania firma Dell oferuje szereg usług przeznaczonych specjalnie dla sztucznej inteligencji.[44] Usługi firmy Dell wspierające wdrażanie sztucznej inteligencji obejmują doradztwo, przygotowywanie danych, wdrażanie, wsparcie i edukację, a każda z nich jest ukierunkowana na konkretne aspekty wdrażania sztucznej inteligencji. Usługi doradcze firmy Dell w zakresie generatywnej sztucznej inteligencji pomagają klientom tworzyć mapę drogową identyfikującą przypadki użycia i pomagającą usprawnić procesy.[45] Podobnie usługi wdrożeniowe w zakresie generatywnej sztucznej inteligencji obejmują warsztaty z profesjonalistami firmy Dell, które umożliwiają przegląd potrzeb i unikalnych wyzwań w celu określenia wstępnie wytrenowanego modelu dla danej firmy i przeprowadzenia sesji transferu wiedzy w celu szkolenia personelu IT.[46] Firma Dell oferuje również usługi wdrożeń, skalowania i zarządzane w zakresie generatywnej sztucznej inteligencji, które zapewniają różne poziomy wsparcia i szkoleń, aż do w pełni zarządzanej infrastruktury sztucznej inteligencji. Dzięki temu pracownicy działów IT mogą skupić się na modelach i danych, podczas gdy firma Dell zarządza sprzętem.[47] Usługi ProSupport zapewniają optymalną wydajność systemu i niezbędną pomoc w zakresie sprzętu oraz oprogramowania w bieżących operacjach sztucznej inteligencji, rozwiązując problemy techniczne.[48]

Usługi edukacyjne są integralną częścią wspierania umiejętności i wiedzy niezbędnych do wykorzystania sztucznej inteligencji. Oferta szkoleń firmy Dell obejmuje kompleksowe programy szkoleniowe w zakresie analizy danych, zaawansowane certyfikaty analityczne i warsztaty dotyczące określonych technologii sztucznej inteligencji, takich jak uczenie maszynowe.[49]

Usługi Supermicro są natomiast w większości ograniczone do rozwiązywania problemów, instrukcji obsługi, autoryzacji zwrotu towaru (RMA) i gwarancji.[50] W ofercie Supermicro AI nie udało nam się znaleźć żadnych usług w zakresie projektowania, wdrażania, zarządzania ani edukacji. W przypadku firm, które poszukują partnera szkoleniowego w zakresie wdrażania sztucznej inteligencji, firma Dell jest oczywistym wyborem.

Współpraca partnerska z innymi firmami w zakresie obciążeń roboczych związanych ze sztuczną inteligencją

Firmy Dell Technologies i NVIDIA współpracują, aby oferować projekty Dell Validated Design, które mają na celu zapewnienie kompleksowego rozwiązania w zakresie generatywnej sztucznej inteligencji w środowiskach biznesowych. Ten projekt tworzy skalowalną, wysokowydajną infrastrukturę opartą na technologiach i oprogramowaniu Dell i NVIDIA oraz strukturę modeli sztucznej inteligencji, która umożliwia firmom tworzenie i uruchamianie niestandardowych modeli sztucznej inteligencji. Rozwiązanie umożliwia klientom szybkie uruchamianie obciążeń roboczych związanych z generatywną sztuczną inteligencją.[51] Aby dowiedzieć się więcej, zapoznaj się z sekcją Dell Validated Design poniżej.

Firma Dell nawiązała współpracę z kilkoma firmami w celu ulepszenia aplikacji technologicznych opartych na sztucznej inteligencji. Dzięki Hugging Face firma Dell ułatwia konfigurowanie dużych modeli językowych (LLM) na miejscu. Partnerstwo to łączy wiedzę ekspercką w zakresie sztucznej inteligencji Hugging Face z serwerami i systemami pamięci masowej firmy Dell. Portal Hugging Face firmy Dell udostępnia narzędzia do prostego, bezpiecznego wdrażania modeli sztucznej inteligencji Hugging Face typu open source. Celem jest ciągłe doskonalenie tych modeli systemów firmy Dell, zwiększanie wydajności i obsługa nowych aplikacji sztucznej inteligencji.[52]

Firmy Dell i Starburst pracują nad wydajnym, skalowalnym jeziorem danych, które integruje analizę Starburst z technologią w domenach obliczeń i pamięci masowej firmy Dell, oferując pojedynczy punkt dostępu do wszystkich źródeł danych dla narzędzi sztucznej inteligencji i uczenia maszynowego. Klienci będą mogli skorzystać z tego partnerstwa, aby eliminować silosy danych.[53]

Według naszych badań Supermicro ma znacznie bardziej ograniczone partnerstwa w zakresie sztucznej inteligencji. SIMA.ai i Supermicro nawiązały współpracę w celu opracowania Supermicro SYS-E300-13AD, kompaktowego serwera brzegowego ML zaprojektowanego z myślą o przetwarzaniu wielostrumieniowej analizy wideo. Ten serwer, wyposażony w potok SIMA.ai ML na chipie, skutecznie obsługuje wiele kanałów wideo, zmniejsza całkowity koszt użytkowania oraz poprawia niezawodność i bezpieczeństwo. Serwer oferuje ustawienie obliczeniowe przeznaczone do przetwarzania i analizy wielu strumieni wideo, zapewniając inteligencję brzegową odpowiednią do różnych zastosowań korporacyjnych.[54]

Dell Validated Designs

Aby wyeliminować niepewność w przypadku rozwiązań sprzętowych z zakresu sztucznej inteligencji, firma Dell oferuje architektury referencyjne zatwierdzone w laboratoriach, zoptymalizowane pod kątem różnych obciążeń roboczych związanych ze sztuczną inteligencją i innych. Te projekty Validated Design obejmują koncepcje architektoniczne, pełne przeglądy rozwiązań, weryfikacje wydajności i inne weryfikacje laboratoryjne potwierdzające możliwości rozwiązania w obciążeniu roboczym, dla którego zostało ono zaprojektowane. Te obciążenia obejmują środowiska wirtualizowane, MLOps, uczenie maszynowe, sztuczną inteligencję konwersacyjną, wnioskowanie generatywnej sztucznej inteligencji, dostrajanie modelu generatywnej sztucznej inteligencji, NVIDIA Fleet Command i OpenShift AI.[55]

Jednym z przykładów jest rozwiązanie AI for Virtualized Environments Validated Design, które łączy sztuczną inteligencję z obsługą VMware ze środowiskiem NVIDIA AI Enterprise w infrastrukturze Dell, optymalizując działanie sztucznej inteligencji w środowiskach wirtualnych.[56] Validated Design Guide zawiera wyniki wydajności pokazujące trening modelu ResNet, który stanowi potwierdzenie, że projekt działa, i pokazuje, jakiego rodzaju wydajności klienci mogą się spodziewać.[57] Weryfikacje te zapewniają klientom wartość wykraczającą poza zwykłą listę składników sprzętu, które współpracują ze sobą — wyjaśniają modele koncepcyjne, zalecają konfiguracje i przeprowadzają klientów przez rozważania i oczekiwania dotyczące wydajności.[58]

Firma Supermicro oferuje rozwiązania na podstawie konkretnych zastosowań, ale nie jest w stanie osiągnąć poziomu architektury referencyjnej, jaki zaobserwowaliśmy w przypadku projektów Dell Validated Design. Zamiast polecać specjalnie zaprojektowane rozwiązanie do konkretnych obciążeń roboczych firma Supermicro dzieli swoje serwery i procesory GPU na kategorie, takie jak wnioskowanie i szkolenia sztucznej inteligencji, wizualizacja i projektowanie HPC/AI itd.[59] W broszurach i arkuszach danych kategorie te składają się z kilku serwerów i procesorów graficznych, które rekomendują jako najlepiej nadające się do danego zadania, kilku przypadków użycia oraz list kluczowych technologii i sugestii dotyczących oprogramowania.[60] W przeciwieństwie do projektów Dell Validated Design wydaje się, że nie obejmują one architektur sieciowych ani danych dotyczących wydajności czy weryfikacji. Firma Supermicro oferuje również kilka projektów referencyjnych, które zapewniają bardziej szczegółową architekturę referencyjną dla niektórych rozwiązań związanych ze sztuczną inteligencją, takich jak wielkoskalowe szkolenie AI z rozwiązaniem chłodzenia cieczą, które firma Supermicro opublikowała we współpracy z NVIDIA.[61] Klienci mogą być w stanie znaleźć bardziej dogłębną architekturę referencyjną dla konkretnego scenariusza, ale w czasie naszych poszukiwań znaleźliśmy tylko trzy: architekturę chłodzenia cieczą, o której wcześniej wspomnieliśmy, architekturę stacji roboczych ze sztuczną inteligencją i architekturę RedHat OpenShift.[62]

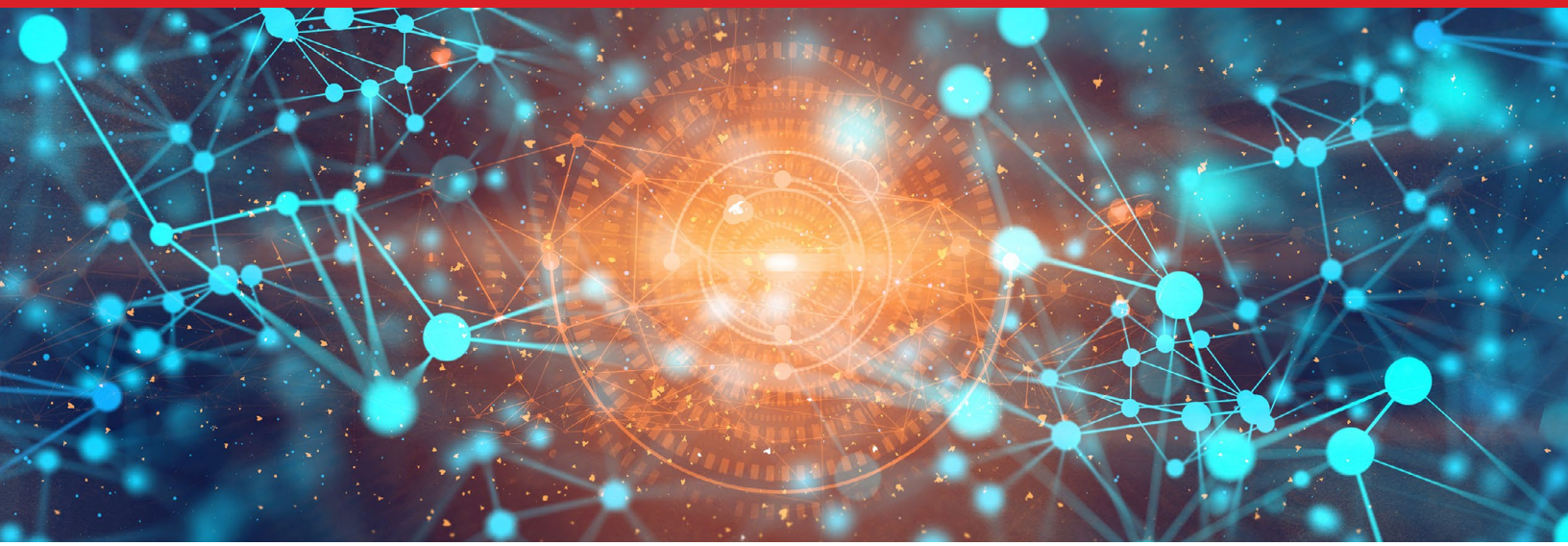
Ogólnie stwierdziliśmy, że projekty Dell Validated Design obejmowały więcej obciążeń roboczych związanych ze sztuczną inteligencją i oferowały bardziej szczegółowe wskazówki niż oferty Supermicro.

Usługi zarządzania i iDRAC

Według raportu Principled Technologies z kwietnia 2023 r. kontroler iDRAC (Integrated Dell Remote Access Controller) oferuje kilka zaawansowanych funkcji za pośrednictwem interfejsu IPMI (Supermicro Intelligent Platform Management Interface), zwłaszcza w zakresie automatyzacji, zabezpieczeń i konfiguracji.[63] W tabeli 3 przedstawiono porównanie funkcji zarządzania Dell i Supermicro z tego raportu, które pokazuje, w jaki sposób kontroler iDRAC może zapewnić łatwiejsze wdrażanie, łatwiejsze aktualizacje oprogramowania wewnętrznego i więcej funkcji zabezpieczeń niż Supermicro IPMI. Należy pamiętać, że niektóre ustalenia mogły ulec zmianie od czasu pierwotnej publikacji.

Tabela 3. Podsumowanie porównania Principled Technologies narzędzi do zarządzania firm Dell i Supermicro z kwietnia 2023 r. Niektóre ustalenia mogły ulec zmianie od czasu publikacji. Źródło: Principled Technologies https://facts.pt/V5fDf06.

	Różnice w dostępie do narzędzi do zarządzania firmy Dell	Porównanie
Łatwiejsze aktualizacje oprogramowania sprzętowego *IDRAC9 a Supermicro IPMI* *OME a Supermicro SSM*	• Automatyczne aktualizacje online za pomocą kontrolera iDRAC9 z opcjami planowania • OME umożliwia tworzenie niestandardowych repozytoriów oprogramowania wewnętrznego i może aktualizować oprogramowanie wewnętrzne systemu BIOS, BMC i innych komponentów serwera bez dodatkowych narzędzi lub agentów	• Automatyczne aktualizacje w kontrolerze iDRAC są konfigurowane w **zaledwie 74 sekundy** • Supermicro IPMI **nie ma** dostępnej funkcji automatycznych aktualizacji, więc administratorzy muszą je przeprowadzać ręcznie • SSM obsługuje tylko aktualizacje oprogramowania wewnętrznego systemu BIOS i BMC i wymaga SUM do aktualizacji innych komponentów.
Więcej funkcji bezpieczeństwa *IDRAC9 a Supermicro IPMI* *OME a Supermicro SSM*	• Kontroler iDRAC9 oferuje funkcję MFA i dynamiczne wyłączanie portów USB bez przestojów systemu • OME oferuje zarówno kontrolę dostępu opartą na rolach (RBAC), jak i kontrolę dostępu opartą na zakresie (SBAC), aby ograniczyć zarządzanie urządzeniami do podzbioru grup urządzeń.	• Supermicro IPMI nie ma **funkcji MFA** • Supermicro IPMI wymaga ponownego uruchomienia systemu i przejścia do konfiguracji systemu BIOS w celu wyłączenia portów USB • Supermicro SSM oferuje RBAC, ale nie bardziej restrykcyjne SBAC
Łatwiejsze zarządzanie cyklem życia *OME a Supermicro SSM*	• Pełne zarządzanie cyklem życia bez agentów za pośrednictwem OME w celu ułatwienia zarządzania i monitorowania	• SSM wymaga agenta SuperDoctor5 w celu uzyskania szczegółowych lokalnych wskaźników kondycji systemu i programu Supermicro Update Manager (SUM) w celu aktualizacji dodatkowych komponentów
Łatwiejsze wdrażanie serwerów *IDRAC9 a Supermicro IPMI*	• Importowanie pełnego profilu serwera Dell w zaledwie 12 krokach za pomocą kontrolera iDRAC9 • Solidne opcje konfiguracji systemu BIOS z wykorzystaniem kontrolera iDRAC9 z 52 funkcjami systemu BIOS i obsługą konfiguracji komponentów takich jak karta sieciowa RAID i kontroler iDRAC.	• Firma Supermicro IPMI pozwoliła nam zapisać i przywrócić tylko konfigurację IPMI, a nie cały profil serwera • Kontroler iDRAC9 ma 52 funkcje systemu BIOS, a interfejs IPMI **nie oferuje opcji konfiguracji systemu BIOS**
Więcej opcji raportowania i analizy *IDRAC9 a Supermicro IPMI* *OME a Supermicro SSM*	• Kontroler iDRAC9 oferuje streaming telemetrii, który umożliwia użytkownikom łatwe wysyłanie danych serwera do narzędzi analitycznych, takich jak Splunk • OME wysyła dane telemetryczne bezpośrednio do CloudIQ w celu łatwiejszego monitorowania	• IPMI oferuje tylko funkcje SYSLOG, których administratorzy mogą używać do wysyłania wiadomości w celu agregacji i analizy • SSM nie ma rozwiązania do zarządzania opartego na chmurze równoważnego Dell CloudIQ
Więcej funkcji zrównoważonego rozwoju *OME a Supermicro SSM*	• Więcej wskaźników monitorowania w programie OME Power Manager, w tym danych dotyczących śladu węglowego	• SSM ma mniej solidne wskaźniki wykorzystania i **nie ma sposobu na śledzenie śladu węglowego**
Więcej sposobów na monitorowanie *OME a Supermicro SSM*	• Zarządzanie serwerami Dell z dowolnego miejsca za pomocą aplikacji mobilnej OpenManage • Monitorowanie urządzeń innych firm za pomocą OME przy użyciu adresów IP serwera i poświadczeń z obsługą importowania BAZ MIB SNMP innych firm	• SSM **nie ma aplikacji mobilnej** • SSM **nie zezwala na** monitorowanie urządzeń innych firm za pomocą adresów IP serwera



Wnioski

Jeśli chodzi o projektowanie, wdrażanie, zarządzanie i utrzymywanie rozwiązań SI w Twojej firmie, należy wziąć pod uwagę wiele czynników. Aby mądrze zainwestować i w pełni wykorzystać możliwości rozwiązania sztucznej inteligencji, możesz poszukać dostawcy, który może być nie tylko dostawcą sprzętu. Z naszych badań wynika, że firma Dell oferuje usługi, które mogą pomóc każdemu partnerowi w całym procesie. Dlatego warto rozważyć inwestowanie w rozwiązania w zakresie sztucznej inteligencji firmy Dell.

---

1. Vipera, „NVIDIA's H100 and A100 GPU Cards: Exploring the Intricacies of SXM and PCI-E Connections", dostęp 5 stycznia 2024 r., https://www.viperatech.com/unraveling-the-mysteries-sxm-vs-pci-e-connections-in-nvidias-high-end-h100-and-a100-gpus/.
2. MLCommons, „MLPerf Inference: Datacenter Benchmark Suite Results", dostęp 5 stycznia 2024 r., https://mlcommons.org/en/inference-datacenter-31/.
3. MLCommons, „MLPerf Inference: Zestaw wyników benchmarków dla centrów danych."
4. Dell, „PowerEdge XE Servers", dostęp 5 stycznia 2024 r., https://www.dell.com/en-us/dt/servers/specialty-servers/poweredge-xe-servers.htm.
5. Supermicro, „Next Leap of AI Infrastructure is Here", dostęp 5 stycznia 2024 r., https://www.supermicro.com/en/accelerators/nvidia.
6. Dell, „PowerEdge XE9680", dostęp 5 stycznia 2024 r., https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe9680-spec-sheet.pdf.
7. Zweryfikowany wynik MLPerf® w wersji 3.1 — wnioskowanie zamknięte. Pobrane z https://mlcommons.org/benchmarks/inference-datacenter/ 5 grudnia 2023 r., wpis 3.1-0069. Nazwa i logo MLPerf są zarejestrowanymi i niezarejestrowanymi znakami towarowymi MLCommons Association w Stanach Zjednoczonych i innych krajach. Wszelkie prawa zastrzeżone. Użycie bez upoważnienia jest surowo wzbronione. Więcej informacji można znaleźć na stronie www.mlcommons.org.
8. Zweryfikowany wynik MLPerf® w wersji 3.1 — wnioskowanie zamknięte. Pobrane z https://mlcommons.org/benchmarks/inference-datacenter/ 5 grudnia 2023 r., wpis 3.1-0135. Nazwa i logo MLPerf są zarejestrowanymi i niezarejestrowanymi znakami towarowymi MLCommons Association w Stanach Zjednoczonych i innych krajach. Wszelkie prawa zastrzeżone. Użycie bez upoważnienia jest surowo wzbronione. Więcej informacji można znaleźć na stronie www.mlcommons.org.
9. Zweryfikowany wynik MLPerf® w wersji 3.1 — wnioskowanie zamknięte. Pobrane z https://mlcommons.org/benchmarks/inference-datacenter/ 5 grudnia 2023 r., wpis 3.1-0069. Nazwa i logo MLPerf są zarejestrowanymi i niezarejestrowanymi znakami towarowymi MLCommons Association w Stanach Zjednoczonych i innych krajach. Wszelkie prawa zastrzeżone. Użycie bez upoważnienia jest surowo wzbronione. Więcej informacji można znaleźć na stronie www.mlcommons.org.
10. Zweryfikowany wynik MLPerf® w wersji 3.1 — wnioskowanie zamknięte. Pobrane z https://mlcommons.org/benchmarks/inference-datacenter/ 5 grudnia 2023 r., wpis 3.1-0132. Nazwa i logo MLPerf są zarejestrowanymi i niezarejestrowanymi znakami towarowymi MLCommons Association w Stanach Zjednoczonych i innych krajach. Wszelkie prawa zastrzeżone. Użycie bez upoważnienia jest surowo wzbronione. Więcej informacji można znaleźć na stronie www.mlcommons.org.

11. Dell, „PowerEdge XE9640 Rack Server", dostęp 5 stycznia 2024 r., godz.ttps://www.dell.com/en-us/shop/ipovw/poweredge-xe9640.
12. Accelsius, „Enabling the AI Revolution with Liquid Cooling", dostęp 5 stycznia 2024 r., https://www.accelsius.com/blog/enabling-the-ai-revolution-with-liquid-cooling.
13. Dell, „Dell PowerEdge XE9640 Technical Guide", dostęp 5 stycznia 2024 r., https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe9640-technical-guide.pdf.
14. Supermicro, „GPU Server Systems", odwiedzony 5 stycznia 2024 r. , https://www.supermicro.com/en/products/gpu?pro=pl_grp_type%3D1.
15. Zweryfikowany wynik MLPerf® w wersji 3.1 — wnioskowanie zamknięte. Pobrane z https://mlcommons.org/benchmarks/inference-datacenter/ 5 grudnia 2023 r., wpis 3.1-0067. Nazwa i logo MLPerf są zarejestrowanymi i niezarejestrowanymi znakami towarowymi MLCommons Association w Stanach Zjednoczonych i innych krajach. Wszelkie prawa zastrzeżone. Użycie bez upoważnienia jest surowo wzbronione. Więcej informacji można znaleźć na stronie www.mlcommons.org.
16. Zweryfikowany wynik MLPerf® w wersji 3.1 — wnioskowanie zamknięte. Pobrane z https://mlcommons.org/benchmarks/inference-datacenter/ 5 grudnia 2023 r., wpis 3.1-0133. Nazwa i logo MLPerf są zarejestrowanymi i niezarejestrowanymi znakami towarowymi MLCommons Association w Stanach Zjednoczonych i innych krajach. Wszelkie prawa zastrzeżone. Użycie bez upoważnienia jest surowo wzbronione. Więcej informacji można znaleźć na stronie www.mlcommons.org.
17. Dell, „PowerEdge XE8640", dostęp 5 stycznia 2024 r., https://www.delltechnologies.com/asset/en-us/products/servers/technical-support/poweredge-xe8640-spec-sheet.pdf.
18. Dell, „PowerEdge XE8640 Rack Server", dostęp 5 stycznia 2024 r., https://www.dell.com/en-us/shop/ipovw/poweredge-xe8640.
19. Zweryfikowany wynik MLPerf® w wersji 3.1 — wnioskowanie zamknięte. Pobrane z https://mlcommons.org/benchmarks/inference-datacenter/ 5 grudnia 2023 r., wpis 3.1-0066. Nazwa i logo MLPerf są zarejestrowanymi i niezarejestrowanymi znakami towarowymi MLCommons Association w Stanach Zjednoczonych i innych krajach. Wszelkie prawa zastrzeżone. Użycie bez upoważnienia jest surowo wzbronione. Więcej informacji można znaleźć na stronie www.mlcommons.org.
20. Zweryfikowany wynik MLPerf® w wersji 3.1 — wnioskowanie zamknięte. Pobrane z https://mlcommons.org/benchmarks/inference-datacenter/ 5 grudnia 2023 r., wpis 3.1-0133. Nazwa i logo MLPerf są zarejestrowanymi i niezarejestrowanymi znakami towarowymi MLCommons Association w Stanach Zjednoczonych i innych krajach. Wszelkie prawa zastrzeżone. Użycie bez upoważnienia jest surowo wzbronione. Więcej informacji można znaleźć na stronie www.mlcommons.org.
21. Dell, „Artificial Intelligence (AI) technologies powered by Dell Precision workstations", dostęp 5 stycznia 2024 r. , https://www.dell.com/en-us/dt/ai-technologies/index.htm?hve=explore+dell+precision+for+ai#tab0=0.
22. Supermicro, „Super Workstations", dostęp 5 stycznia 2024 r., https://www.supermicro.com/en/products/superworkstation .
23. Supermicro, „Rackmount Workstations", dostęp 5 stycznia 2024 r., https://www.supermicro.com/en/products/rackmount-workstations.
24. Stephen Pritchard, „Storage requirements for AI, ML and analytics in 2022", dostęp 05 stycznia 2024 r., https://www.computerweekly.com/feature/Storage-requirements-for-AI-ML-and-analytics-in-2022.
25. Dell, „PowerScale AI-Ready Data Platform", dostęp 5 stycznia 2024 r., https://www.dell.com/en-us/shop/powerscale-family/sf/powerscale.
26. Dell, „Dell PowerScale All-Flash", dostęp 05 stycznia 2024 r., https://www.delltechnologies.com/asset/en-us/products/storage/technical-support/h15963-ss-powerscale-all-flash-nodes.pdf.
27. Dell, „Dell PowerScale OneFS Software Features", dostęp 05 stycznia 2024 r., https://www.delltechnologies.com/asset/en-us/products/storage/technical-support/h18275-onefs-software-features-data-sheet.pdf.
28. Dell, „Dell ECS Enterprise Object Storage", dostęp 5 stycznia 2024 r., https://www.dell.com/en-us/dt/storage/ecs/index.htm#tab0=0&tab1=0&accordion0.
29. Supermicro, „Accelerating AI Data Pipelines", dostęp 5 stycznia 2024 r., https://www.supermicro.com/en/products/storage.
30. Supermicro, „Supermicro Software-Defined Storage and Memory Solutions", dostęp 5 stycznia 2024 r., https://www.supermicro.com/en/solutions/software-defined-storage.
31. Supermicro, „Supermicro WEKA Distributed Storage Solution", odwiedzony 5 stycznia 2024 r ., https://www.supermicro.com/en/solutions/wekaio .
32. Dell, „Dell ECS Enterprise Object Storage", dostęp 5 stycznia 2024 r., https://www.dell.com/en-us/dt/storage/ecs/index.htm#tab0=0&tab1=0&accordion0.
33. Dell, „Dell ECS Enterprise Object Storage"
34. Supermicro, „Supermicro OSNEXUS Software-Defined Storage Solution", dostęp 5 stycznia 2024 r., https://www.supermicro.com/en/solutions/osnexus.

35. ASBIS, „Supermicro Solution for Scality RING", dostęp 05 stycznia 2024 r., https://news.asbis.com/news/suppliers/supermicro-renewed-the-line-of-scality-ring-solution.

36. Supermicro, „Rozwiązanie Supermicro dla Quantum ActiveScale", dostęp 5 stycznia 2024 r. https://www.supermicro.com/en/solutions/activescale.

37. Dell, „Scalable and Elastic Storage as-a-Service", dostęp 5 stycznia 2024 r. https://www.dell.com/en-us/dt/apex/storage/data-storage-services/.

38. Dell, „Flip the Switch to Open Networking with PowerSwitch", dostęp 5 stycznia 2024 r. , https://www.dell.com/en-us/dt/networking/.

39. Dell, „Dell PowerSwitch Data Center Switches", dostęp 5 stycznia 2024 r. , https://www.dell.com/en-us/dt/networking/data-center-switches/.

40. Supermicro, „SSE-T7132S - 400Gb Ethernet Switch", dostęp 5 stycznia 2024 r., https://www.supermicro.com/en/products/accessories/Networking/SSE-T7132SR.php.

41. Dell, „Dell EMC Networking SmartFabric Services Deployment with VxRail 4.7 — Fabric Design Center", dostęp 5 stycznia 2024 r., https://infohub.delltechnologies.com/l/dell-emc-networking-smartfabric-services-deployment-with-vxrail-4-7-1/fabric-design-center-26/.

42. Dell, „Dell SmartFabric Services", dostęp 5 stycznia 2024 r., https://www.dell.com/en-us/dt/networking/smartfabric/.

43. Penny Madsen, „Scaling Skills for AI: Lessons from Early Adopters", dostęp 5 stycznia 2024 r., https://www.delltechnologies.com/asset/en-us/products/servers/industry-market/idc-brief-importance-of-skills-for-ai-dell.pdf.

44. Dell, „Design Guide — Generative AI in the Enterprise — Model Customization — Overview", dostęp 5 stycznia 2024 r. https://infohub.delltechnologies.com/l/design-guide-generative-ai-in-the-enterprise-model-customization/overview-5381/.

45. Dell, „Design Guide—Generative AI in the Enterprise – Model Customization—Advisory Services for Generative AI", dostęp 5 stycznia 2024 r., https://infohub.delltechnologies.com/l/design-guide-generative-ai-in-the-enterprise-model-customization/advisory-services-for-generative-ai-1/.

46. Dell, „Design Guide—Generative AI in the Enterprise – Model Customization—Adoption Services for Generative AI", dostęp 5 stycznia 2024 r., https://infohub.delltechnologies.com/l/design-guide-generative-ai-in-the-enterprise-model-customization/adoption-services-for-generative-ai-1/.

47. Dell, „Poradnik projektowania — Generatywna sztuczna inteligencja w przedsiębiorstwie — Personalizacja modeli — Usługi wdrożeniowe dla generatywnej sztucznej inteligencji".

48. Dell, „Artificial Intelligence (AI) Ready Solution Services", dostęp 5 stycznia 2024 r., https://www.dell.com/en-us/dt/services/solutions/artificial-intelligence-services.htm.

49. Dell, „Comprehensive AI Training Modules Tailored for You", dostęp 5 stycznia 2024 r., https://education.dell.com/content/emc/en-us/home/training/aiml.html.

50. Supermicro, „Usługi i wsparcie", dostęp 5 stycznia 2024 r. , https://www.supermicro.com/en/support.

51. Travis Vigil, „Dell and NVIDIA: Bring Generative AI to the Enterprise", dostęp 5 stycznia 2024 r., https://www.dell.com/en-us/blog/dell-and-nvidia-bringing-generative-ai-to-the-enterprise/.

52. Dell, „Dell Technologies and Hugging Face to Simplify Generative AI with On-Premises IT", dostęp 05 stycznia 2024 r., https://www.dell.com/en-us/dt/corporate/newsroom/announcements/detailpage.press-releases~usa~2023~11~20231114-dell-technologies-and-hugging-face-to-simplify-generative-ai-with-on-premises-it.htm.

53. Richard DeMare, „Starburst and Dell expand partnership to accelerate AI efforts with more intelligent data collection", dostęp 5 stycznia 2024 r., https://www.starburst.io/blog/starburst-and-dell-expand-partnership-to-accelerate-ai-efforts-with-more-intelligent-data-collection/.

54. Business Wire, „SIMA.ai and Supermicro announce Partnership to Accelerate Power Efficient ML at the Edge", dostęp 5 stycznia 2024 r., https://www.businesswire.com/news/home/20231129609794/en/SiMa.ai-and-Supermicro-Announce-Partnership-to-Accelerate-Power-Efficient-ML-at-the-Edge.

55. Dell, „Dell AI solutions", dostęp 5 stycznia 2024 r., https://www.dell.com/en-us/dt/solutions/artificial-intelligence/index.htm#accordion0&tab0=0.

56. Dell, „Unlock the power of AI in virtualized environments", dostęp 5 stycznia 2024 r., https://www.delltechnologies.com/asset/en-us/products/ready-solutions/briefs-summaries/ai-vxrail-powerscale-brief.pdf.

57. Dell, „Design Guide — Virtualization GPU for AI with VMware and NVIDIA Based on Dell Infrastructure — Performance Results", dostęp 5 stycznia 2024 r. https://infohub.delltechnologies.com/l/design-guide-virtualizing-gpus-for-ai-with-vmware-and-nvidia-based-on-dell-infrastructure-1/performance-results-15/.

58. Dell, „Design Guide — Virtualizing GPU for AI with VMware and NVIDIA Based on Dell Infrastructure — Design Considentions", dostęp 5 stycznia 2024 r. https://infohub.delltechnologies.com/l/design-guide-virtualizing-gpus-for-ai-with-vmware-and-nvidia-based-on-dell-infrastructure-1/design-considerations-105/.

59. Supermicro, „Accelerate Every Workload”, dostęp 5 stycznia 2024 r., https://www.supermicro.com/en/solutions/ai-deep-learning.

60. Supermicro, „Supermicro Enterprise AI Inference & Training”, dostęp 5 stycznia 2024 r., https://www.supermicro.com/datasheet/Datasheet_AI-Workloads_Enterprise_AI_Inferencing_and_Training.pdf.

61. Supermicro, „SUPERMICRO RACK SCALE SOLUTIONS: LARGE SCALE AI TRAINING WITH LIQUID COOLING”, dostęp 5 stycznia 2024 r., https://www.supermicro.com/solutions/Solution-Brief_Rack_Scale_AI.pdf.

62. Supermicro, „Accelerate Every Workload”, dostęp 5 stycznia 2024 r., https://www.supermicro.com/en/solutions/ai-deep-learning.

63. Principled Technologies, „Dell management tools made server deployment and updates easier, offered more comprehensive security, and provided more robust infrastructure analytics”, dostęp 5 stycznia 2024 r., https://www.principledtechnologies.com/Dell/Management-tools-vs-Supermicro-0423.pdf.

Zobacz oryginalną, angielską wersję tego raportu na stronie https://facts.pt/q9p46K9

Projekt ten został zlecony przez Dell Technologies.

Principled Technologies®

Facts matter.®