SPECYFIKACJA INFRASTRUKTURY ZAMAWIAJĄCEGO W ZAKRESIE WSPARCIA SI/ML/HPC

Lp.	Obszar	Wymaganie
1.	System	1.1. Oferowany System musi znajdować się na liście niezależnie zweryfikowanych Systemów przez MLCommons. 1.1.1. Oferowana konfiguracja Systemu musi składać się z co najmniej jednej zweryfikowanej przez MLCommons konfiguracji Systemu, przy czym dopuszczalne jest zwielokrotnienie tej konfiguracji poprzez dostarczenie większej liczby węzłów. Zamawiający oczekuje, że system zostanie dostarczony w konfiguracji nie więcej niż 31 węzłów compute, zawierających każdy nie więcej niż 8 akceleratorów GPU wspierających obliczenia, dostarczonych w szafach rack wraz z niezbędnymi komponentami zarządzania systemem, dystrybucji zasilania, sieci LAN oraz sieci storage. 1.1.1.1. Oferowany System musi być wyposażony w akceleratory SI o potwierdzonym w benchmarku MLCommons wyniku dla modelu llama3.1-405b powyżej 100 Queries/s dla pojedynczego akceleratora (uzyskany wynik Systemu z kolumny llama3.1-405b, Server, Queries/s podzielony przez wartość z kolumny #of Accelerators i następnie podzielony przez wartość z kolumny #of Nodes) 1.1.1.2. Oferowany System musi być wyposażony w akceleratory SI o potwierdzonym w benchmarku MLCommons wyniku dla modelu mixtral-8x7b powyżej 15000 Tokens/s dla pojedynczego akceleratora (uzyskany wynik Systemu z kolumny mixtral-8x7b, Server, Tokens/s podzielony przez wartość z kolumny #of Accelerators i następnie podzielony przez wartość z kolumny #of Nodes)
2.	Wydajność	• Łączna wydajność oferowanego Systemu musi gwarantować co najmniej 2 EFLOPS FP8 (2000 petaFLOPS FP8) trenowania modeli SI oraz co najmniej 4 EFLOPS FP8 (4000 petaFlops FP8) inferencji modeli SI. 2.1. Deklarowana wydajność oferowanego Systemu musi być potwierdzona przez producenta.
3.	Procesory	3.1. Oferowany System musi posiadać procesory w standardzie x86 - 64 bity do uruchamiania systemu operacyjnego.

Centrum e-Zdrowia
ul. Stanisława Dubois 5A
00-184 Warszawa

tel.: +48 22 597-09-27
fax: +48 22 597-09-37
biuro@cez.gov.pl | www.cez.gov.pl

NIP: 5251575309
REGON: 001377706

Zadanie „Platforma Usług Inteligentnych” jest realizowane w ramach projektu “e-Zdrowie KPO”, współfinansowanego ze środków Unii Europejskiej w ramach instrumentu na rzecz Krajowego Planu Odbudowy i Zwiększania Odporności dla przedsięwzięć realizowanych w ramach inwestycji D1.1.2 „Przyspieszenie procesów transformacji cyfrowej ochrony zdrowia poprzez dalszy rozwój usług cyfrowych w ochronie zdrowia”, będącej elementem komponentu D „Efektywność, dostępność i jakość systemu ochrony zdrowia”

		• Każdy węzeł przetwarzający zadania SI w ramach oferowanego Systemu musi być wyposażony w co najmniej 100 rdzeni procesorów x86 – 64 bity.
4.	Pamięć operacyjna	4.1. Oferowany System musi łącznie posiadać co najmniej 60 TB pamięci operacyjnej. • Każdy węzeł przetwarzający zadania SI w ramach oferowanego Systemu musi posiadać co najmniej 1 TB pamięci operacyjnej.
5.	Pamięć masowa	• Oferowany System musi zawierać skalowalną pamięć masową dedykowaną do rozwiązań SI, certyfikowaną przez producenta GPU. 5.1. Pojemność oferowanego systemu pamięci masowej musi obejmować co najmniej 1 PB dostępnej surowej przestrzeni na dane. 5.2. Deklarowana wydajność systemu pamięci masowej musi zapewniać co najmniej 180 GB/S dla sekwencyjnego odczytu oraz 120 GB/s sekwencyjnego zapisu. 5.3. Oferowany system pamięci masowej musi umożliwiać instalację w standardowej szafie sprzętowej o rozmiarze do 4 RU. 5.4. Oferowany system pamięci masowej musi charakteryzować się nie większym zużyciem prądu niż 5 KW. 5.5. Oferowany system pamięci masowej musi być wyposażony w co najmniej 16 interfejsów sieciowych o przepływności co najmniej 200 Gbit/s. 5.6. Oferowany system pamięci masowej musi wspierać standard RDMA oraz połączenia dedykowane dla GPU Systemu. 5.7. Oferowany system pamięci masowej musi umożliwiać konfigurację globalnych dysków Hot Spare. 5.8. Oferowany system pamięci masowej musi zawierać awaryjne zasilanie z akumulatora lustrzanej pamięci podręcznej o parametrach pojemnościowych i wydajnościowych zapewniających możliwość zapisu danych z pamięci podręcznej na przestrzeń dyskową w razie utraty zasilania i wymuszenia wyłączenia całego systemu. 5.9. Oferowany system pamięci masowej musi zapewniać redundantne, aktywne kontrolery. 5.10. Oferowany system pamięci masowej musi być wysoce dostępny i reduntanty. W szczególności kontrolery, moduły/porty IO, kable, zasilacze, wentylatory itp.

Centrum e-Zdrowia
ul. Stanisława Dubois 5A
00-184 Warszawa

tel.: +48 22 597-09-27
fax: +48 22 597-09-37
biuro@cez.gov.pl | www.cez.gov.pl

NIP: 5251575309
REGON: 001377706

KRAJOWY PLAN ODBUDOWY
Rzeczpospolita Polska
Sfinansowane przez Unię Europejską NextGenerationEU
Ministerstwo Zdrowia

Zadanie „Platforma Usług Inteligentnych" jest realizowane w ramach projektu "e-Zdrowie KPO", współfinansowanego ze środków Unii Europejskiej w ramach instrumentu na rzecz Krajowego Planu Odbudowy i Zwiększania Odporności dla przedsięwzięć realizowanych w ramach inwestycji D1.1.2 „Przyspieszenie procesów transformacji cyfrowej ochrony zdrowia poprzez dalszy rozwój usług cyfrowych w ochronie zdrowia", będącej elementem komponentu D „Efektywność, dostępność i jakość systemu ochrony zdrowia"

		5.11. Oferowany system pamięci masowej musi obsługiwać dyski samoszyfrujące (Self Encrypting Device) bez wpływu na wydajność. 5.12. Oprogramowanie systemu pamięci masowej musi zapewniać wysokowydajny, równoległy system plików wraz z kontrolą integralności danych.
6.	Dyski twarde	• Każdy węzeł przetwarzający zadania SI oferowanego Systemu musi posiadać zainstalowane co najmniej 2 dyski min. 1TB SSD służące do uruchamiania systemu oraz co najmniej 8 dysków SSD NVMe o pojemności min. 3 TB.
7.	Wsparcie dla systemów operacyjnych	7.1. Oferowany System musi zostać dostarczony wraz systemem operacyjnym Linux, którego zaoferowana dystrybucja jest wspierana i certyfikowana przez producenta Systemu. • System operacyjny musi zawierać optymalizacje jądra systemu, w tym wsparcie dla bezpośredniego dostępu akceleratorów SI (tj. GPU) do pamięci masowych 7.2. Oferowany System musi umożliwić wdrożenie w pełni odizolowane od internetu (ang. Air-Gapped) w celu umożliwienia bezpiecznej aktualizacji sytemu operacyjnego wraz z oprogramowaniem bazowym. Wymagane jest zapewnienie mechanizmów umożliwiających instalację i aktualizację oprogramowania bez konieczność zapewnienia dostępu do internetu czy usług w chmurze producenta Systemu.
8	Interfejsy sieciowe	• Każdy węzeł przetwarzający zadania SI oferowanego Systemu musi posiadać co najmniej 4 interfejsy sieciowe Infiniband NDR o przepływności co najmniej 400 Gbit/s służące do komunikacji pomiędzy węzłami przetwarzającymi zadania SI. • Każdy węzeł przetwarzający zadania SI musi posiadać 2 dedykowane interfejsy Infiniband NDR o przepływności co najmniej 200 Gbit/s do komunikacji z zewnętrzną pamięcią masową. 8.1. Każdy węzeł przetwarzający zadania SI musi posiadać min. 2 dedykowane interfejsy In-Band Management oraz jeden dedykowany interfejs Out-Band Management o przepływności co najmniej 200 Gbit/s do zarządzania. 8.2. Wykonawca dostarczy System z następującym ukompletowaniem urządzeń sieciowych wyprodukowanych przez producenta akceleratorów GPU: 8.2.1. Przełącznik InfiniBand, 64 porty NDR 400 Gb/s, 32 porty OSFP 400 Gb/s, łączna przepustowość min. 51 Tb/s, redundancja zasilania, zajętość szafy rack max 1U – 18 szt.

Centrum e-Zdrowia
ul. Stanisława Dubois 5A
00-184 Warszawa

tel.: +48 22 597-09-27
fax: +48 22 597-09-37
biuro@cez.gov.pl | www.cez.gov.pl

NIP: 5251575309
REGON: 001377706

KRAJOWY PLAN ODBUDOWY
Rzeczpospolita Polska
Sfinansowane przez Unię Europejską NextGenerationEU
Ministerstwo Zdrowia

Zadanie „Platforma Usług Inteligentnych” jest realizowane w ramach projektu “e-Zdrowie KPO”, współfinansowanego ze środków Unii Europejskiej w ramach instrumentu na rzecz Krajowego Planu Odbudowy i Zwiększania Odporności dla przedsięwzięć realizowanych w ramach inwestycji D1.1.2 „Przyspieszenie procesów transformacji cyfrowej ochrony zdrowia poprzez dalszy rozwój usług cyfrowych w ochronie zdrowia”, będącej elementem komponentu D „Efektywność, dostępność i jakość systemu ochrony zdrowia”

		8.2.2. Unified Fabric Manager, zarządzanie siecią, 2 porty NDR 400 Gb/s, redundancja zasilania, zajętość szafy rack 1U – 4 szt. 8.2.3. Przełącznik sieci Ethernet 800 Gbe, 64 porty OSFP, 1 port SFP28, łączna przepustowość min. 51 Tb/s, redundancja zasilania, zajętość szafy rack max 2U – 4 szt. 8.2.4. Przełącznik sieci Ethernet 1GBase-T/100GbE, 48 portów RJ45 1Gb/s, 2 porty QSFP28 100Gb/s, przepustowość min. 440 Gb/s, redundancja zasilania, zajętość szafy rack max 1U – 7 szt. 8.2.5. Wszystkie porty w urządzeniach wymienionych powyżej muszą być aktywne bez konieczności zakupu dodatkowych licencji. 8.2.6. Komplet oryginalnych wkładek oraz interfejsów sieciowych zapewniających maksymalną przepustowość dla wszystkich urządzeń sieciowych. 8.2.7. Komplet kabli zapewniających realizację połączeń wewnętrznych pomiędzy węzłami Systemu.
9.	Akceleratory GPU	9.1. Oferowany System powinien umożliwiać wykorzystanie akceleratorów GPU najnowszej (na dzień składania ofert) generacji oferowanej przez producenta GPU będących częścią składową węzłów przetwarzania oferowanego Systemu. Oznacza to, że nawet jeżeli obecne wymagania wydajnościowe pozwolą na zastosowanie akceleratorów starszej generacji niż dostępne w dniu składania ofert, oferowany System musi umożliwić obsługę najnowszej generację akceleratorów, która będzie dostępna w dniu składania ofert. Każdy węzeł Systemu musi zapewnić min. 1,4 TB pamięci GPU typu HBM3 łącznie.

Specyfikacja oprogramowania Zamawiającego.

Centrum e-Zdrowia
ul. Stanisława Dubois 5A
00-184 Warszawa

tel.: +48 22 597-09-27
fax: +48 22 597-09-37
biuro@cez.gov.pl | www.cez.gov.pl

NIP: 5251575309
REGON: 001377706

KRAJOWY PLAN ODBUDOWY
Rzeczpospolita Polska
Sfinansowane przez Unię Europejską NextGenerationEU
Ministerstwo Zdrowia

Zadanie „Platforma Usług Inteligentnych” jest realizowane w ramach projektu “e-Zdrowie KPO”, współfinansowanego ze środków Unii Europejskiej w ramach instrumentu na rzecz Krajowego Planu Odbudowy i Zwiększania Odporności dla przedsięwzięć realizowanych w ramach inwestycji D1.1.2 „Przyspieszenie procesów transformacji cyfrowej ochrony zdrowia poprzez dalszy rozwój usług cyfrowych w ochronie zdrowia”, będącej elementem komponentu D „Efektywność, dostępność i jakość systemu ochrony zdrowia”

Lp.	Obszar	Wymaganie
1.	Bazowe oprogramowanie wsparcia SI	• Bazowe oprogramowanie wsparcia SI musi posiadać wystawiony przez producenta tego oprogramowania lub organizację posiadającą stosowne prawa autorskie, certyfikat lub inny równoważny dokument poświadczający możliwość jego bezproblemowego i bezkonfliktowego wdrożenia i eksploatacji na infrastrukturze obliczeniowej oferowanego Systemu. W szczególności oferowane oprogramowanie musi być certyfikowane dla: 1.1.1. Architektury i generacji oferowanego Systemu. 1.1.2. Rodzaju i wersji oferowanego systemu operacyjnego. 1.1.3. Platformy orkiestracji kontenerów Kubernetes oraz kontenerów w technologii Docker lub OCI wdrażanych bezpośrednio na systemie operacyjnym, poza środowiskiem orkiestracji Kubernetes. • Licencja na oferowane oprogramowanie wsparcia SI musi zapewnić dostęp do kodówczy to to samo co wyżej? źródłowych oraz dokumentacji umożliwiającej budowanie aplikacji i ich rozwój z wykorzystaniem dostarczonych narzędzi i bibliotek. • Oferowane oprogramowanie SI musi posiadać funkcjonalności zarządzania infrastrukturą systemową, oprogramowaniem orkiestracyjnym, aby umożliwić skalowanie przetwarzania SI. W szczególności oprogramowanie musi posiadać: 1.1.4. Funkcjonalność zapewniającą automatyzację zarządzania sterownikami akceleratorów GPU udostępnioną jako Operator dla platformy orkiestracji kontenerów Kubernetes. 1.1.5. Katalog Kontenerów typu Docker wraz ze wsparciem automatyzacji wdrożenia w standardzie Helm Chart. 1.1.6. Funkcjonalność zapewniającą w klastrze Kubernetes dynamiczne przydzielanie akceleratorów GPU. Musi umożliwiać wielu użytkownikom i zadaniom współdzielenie pojedynczego GPU, alokując im ułamki mocy obliczeniowej w zależności od potrzeb. • Funkcjonalność zapewniająca w klastrze Kubernetes agregację akceleratorów GPU pozwalającą na łączenie mocy wielu GPU dla pojedynczego, intensywnego zadania (np. trenowanie modeli SI). 1.1.7. Funkcjonalność zapewniającą w klastrze Kubernetes automatyczne skalowanie poprzez dynamiczne dostosowanie alokacji zasobów GPU, CPU, pamięci w zależności od potrzeb uruchomionych zadań.

Centrum e-Zdrowia
ul. Stanisława Dubois 5A
00-184 Warszawa

tel.: +48 22 597-09-27
fax: +48 22 597-09-37
biuro@cez.gov.pl | www.cez.gov.pl

NIP: 5251575309
REGON: 001377706

KRAJOWY PLAN ODBUDOWY
Rzeczpospolita Polska
Sfinansowane przez Unię Europejską NextGenerationEU
Ministerstwo Zdrowia

Zadanie „Platforma Usług Inteligentnych" jest realizowane w ramach projektu "e-Zdrowie KPO", współfinansowanego ze środków Unii Europejskiej w ramach instrumentu na rzecz Krajowego Planu Odbudowy i Zwiększania Odporności dla przedsięwzięć realizowanych w ramach inwestycji D1.1.2 „Przyspieszenie procesów transformacji cyfrowej ochrony zdrowia poprzez dalszy rozwój usług cyfrowych w ochronie zdrowia", będącej elementem komponentu D „Efektywność, dostępność i jakość systemu ochrony zdrowia"

2.	Katalog Kontenerów	2. W Katalogu Kontenerów konieczne jest zapewnienie: 2.1. Serwera inferencji SI wspierającego MLfow, TensorFlow, PyTorch. 2.2. Narzędzia do gromadzenia i wizualizacji logów w ramach platformy. 2.3. Środowiska rozwoju modeli generatywnej sztucznej inteligencji. Środowisko musi wspierać rozwój multi modalnych modeli włączając Llama 2, Falcon, CLIP, LLAVA, GPT, T5, BERT, MoE, RETRO. Środowisko musi zapewniać pretrenowane, bazowe modele w obszarze automatycznego rozpoznawania mowy, przetwarzania języka naturalnego (NLP), zamiany tekstu na mowę oraz analizy obrazów (typu Computer Vision). Środowisko musi dostarczać funkcjonalności wspierające strojenie modeli typu Supervised Fine-Tuning (SFT) oraz Reinforcement Learning from Human Feedback (RLHF). 2.4. Środowisko budowy potoków analitycznych w oparciu o akceleratory GPU umożliwiające wysokopoziomowe programowanie w oparciu o język Python. Środowisko musi zawierać akcelerowane na GPU biblioteki cuDF, CuML, cuGraph, cuVS, RMM, RAFT lub równoważne.

Centrum e-Zdrowia
ul. Stanisława Dubois 5A
00-184 Warszawa

tel.: +48 22 597-09-27
fax: +48 22 597-09-37
biuro@cez.gov.pl | www.cez.gov.pl

NIP: 5251575309
REGON: 001377706

KRAJOWY PLAN ODBUDOWY
Rzeczpospolita Polska
Sfinansowane przez Unię Europejską NextGenerationEU
Ministerstwo Zdrowia

Zadanie „Platforma Usług Inteligentnych” jest realizowane w ramach projektu “e-Zdrowie KPO”, współfinansowanego ze środków Unii Europejskiej w ramach instrumentu na rzecz Krajowego Planu Odbudowy i Zwiększania Odporności dla przedsięwzięć realizowanych w ramach inwestycji D1.1.2 „Przyspieszenie procesów transformacji cyfrowej ochrony zdrowia poprzez dalszy rozwój usług cyfrowych w ochronie zdrowia”, będącej elementem komponentu D „Efektywność, dostępność i jakość systemu ochrony zdrowia”