**Specyfikacja Warunków Zamówienia**

**CEL**

Celem niniejszego zamówienia jest zapewnienie Zamawiającemu dostępu do elastycznej i skalowalnej usługi wynajmu infrastruktury obliczeniowej wyposażonej w wysokowydajne układy GPU, przeznaczonej do zastosowań wymagających dużej mocy obliczeniowej. Usługa ma umożliwić prowadzenie zaawansowanych obliczeń, takich jak trening modeli sztucznej inteligencji i uczenia maszynowego (AI/ML) oraz symulacje w dogodnym modelu rozliczeniowym. Dzięki temu Zamawiający będzie mógł dynamicznie skalować wykorzystane zasoby obliczeniowe w zależności od bieżących potrzeb, płacąc proporcjonalnie do rezerwacji zasobów.

**PRZEDMIOT ZAMÓWIENIA**

Przedmiotem zamówienia jest **usługa wynajmu i udostępnienia elastycznej, skalowalnej infrastruktury obliczeniowej z akceleratorami GPU** do celów związanych z wykorzystaniem modeli sztucznej inteligencji. W ramach usługi Wykonawca zapewni Zamawiającemu zdalny dostęp do zasobów obliczeniowych (maszyn wirtualnych lub serwerów fizycznych) wyposażonych w wyspecyfikowane poniżej układy GPU, wraz z niezbędną infrastrukturą towarzyszącą.

Usługa obejmuje dostarczenie odpowiedniej mocy obliczeniowej na żądanie, zgodnie z potrzebami Zamawiającego, oraz utrzymanie środowiska informatycznego umożliwiającego pełne wykorzystanie tych zasobów. W szczególności Wykonawca zobowiązuje się zapewnić dostęp do instancji obliczeniowych o wymaganych parametrach sprzętowych, z pełnym dostępem administracyjnym dla Zamawiającego, a także dostęp do wymaganej przestrzeni dyskowej i preinstalowanych narzędzi programistycznych zgodnie z dalszymi wymaganiami technicznymi.

**WYMAGANIA TECHNICZNE**

Oferowana usługa musi spełniać następujące minimalne wymagania techniczne:

1. **Sprzęt GPU:** Zastosowane procesory graficzne (GPU) muszą charakteryzować się wydajnością klasy co najmniej **NVIDIA H100** (wydajność: 980 TFLOPS – FP32, 1975 TFLOPS - FP16/BF16) lub równoważnej. Każdy GPU powinien dysponować **minimum 80 GB** własnej pamięci VRAM o bardzo wysokiej przepustowości (co najmniej **3 TB/s**). Wymagane jest, aby układy GPU posiadały specjalizowane jednostki do przyspieszania obliczeń macierzowych i AI (np. **Tensor Cores** lub równoważne technologie w przypadku innego producenta). Dopuszcza się układy nowszej generacji o równoważnych lub wyższych parametrach (jeśli dostępne na rynku w okresie realizacji zamówienia).
2. **Konfiguracja i skalowalność:** Usługa musi umożliwiać elastyczne konfigurowanie zasobów GPU. W szczególności wymaga się możliwości przydzielenia od **1 do 8 GPU** w ramach pojedynczej instancji (maszyny wirtualnej lub serwera fizycznego) w zależności od potrzeb konkretnego zadania. Wykonawca powinien zapewnić opcję jednoczesnego uruchomienia wielu instancji/maszyn, co pozwoli Zamawiającemu korzystać z większej liczby GPU równolegle, jeżeli zajdzie taka potrzeba. Skalowanie zasobów (zarówno zwiększanie, jak i zmniejszanie liczby aktywnych instancji lub przypisanych GPU) powinno być możliwe dynamicznie, na żądanie Zamawiającego, z minimalnym czasem oczekiwania.
3. **System operacyjny i dostęp:** Udostępnione instancje obliczeniowe muszą obsługiwać systemy operacyjne z **jądrem Linux**. W szczególności wymaga się pełnej kompatybilności co najmniej z następującymi dystrybucjami: **Ubuntu**, **Arch Linux** oraz **Fedora** (dopuszczalne są również inne dystrybucje Linuksa po potwierdzeniu kompatybilność z oprogramowaniem realizującym obliczenia). Zamawiający musi posiadać **pełne uprawnienia administratora (root)** w systemie operacyjnym każdej instancji, co umożliwi samodzielną instalację oprogramowania i konfigurację środowiska. Dostęp do instancji musi być zapewniony zdalnie poprzez **protokół SSH**, umożliwiający zdalną pracę na wynajętych maszynach.
4. **Środowisko AI/ML:** Wykonawca powinien zapewnić wsparcie dla zastosowań sztucznej inteligencji i uczenia maszynowego. Preferowane jest udostępnienie **gotowych obrazów maszyn lub kontenerów** z preinstalowanym oprogramowaniem niezbędnym do pracy z AI/ML. W szczególności w środowisku powinny być dostępne aktualne wersje popularnych frameworków, takich jak **PyTorch** i **TensorFlow**, wraz z odpowiednimi bibliotekami (np. **CUDA**, **cuDNN** dla układów NVIDIA) oraz sterownikami GPU. Jeżeli gotowe obrazy nie są dostarczane, Wykonawca musi zapewnić, że instalacja wymaganego oprogramowania na instancjach (przez Zamawiającego) będzie możliwa i nieograniczona, a wszelkie niezbędne sterowniki do obsługi GPU są dostępne i kompatybilne z wybranym systemem operacyjnym i wyżej wymienionym oprogramowaniem.
5. **Pamięć masowa:** Usługa musi zapewniać **trwałą przestrzeń dyskową** o pojemności co najmniej **1 TB**, dostępną dla instancji obliczeniowych Zamawiającego. Przestrzeń ta powinna umożliwiać **trwałe przechowywanie danych** – tzn. dane zachowują integralność i dostępność niezależnie od ewentualnego wyłączenia lub restartu instancji obliczeniowych. Preferowane jest, aby przestrzeń dyskowa była zrealizowana w oparciu o szybkie dyski (np. technologia SSD/NVMe) zapewniające odpowiednią wydajność operacji we/wy. Wykonawca powinien gwarantować możliwość rozszerzenia przestrzeni dyskowej w razie wzrostu zapotrzebowania oraz zapewnić mechanizmy zabezpieczające dane.
6. **Model rozliczeń (pay-per-use):** Oferowana usługa musi być świadczona w modelu rozliczeniowym opartym o rzeczywiste wykorzystanie zasobów. **Rozliczanie za rzeczywiste zużycie** oznacza, że Zamawiający ponosi koszty proporcjonalne do użycia infrastruktury – np. czasu działania instancji z GPU, liczby wykorzystanych GPU-godzin/dni – zamiast stałej opłaty niezależnej od wykorzystania. Wszelkie opłaty stałe (abonamentowe) po stronie Zamawiającego powinny być ograniczone do niezbędnego minimum albo nie występować wcale. Wymagane jest, by Wykonawca dostarczał czytelne i szczegółowe raporty zużycia zasobów, które będą podstawą do naliczania opłat (zgodnie z zasadami opisanymi w sekcji **WARUNKI DODATKOWE** dotyczącej monitoringu i rozliczalności).

**WARUNKI DODATKOWE**

Ponadto, oferowana usługa musi spełniać następujące dodatkowe warunki i zapewnienia:

1. **Gwarancja dostępności (SLA):** Wykonawca powinien zapewnić wysoki poziom dostępności usługi, potwierdzony zapisami w **Umowie o Poziomie Usługi (SLA)**. Wymagany minimalny poziom dostępności to **99,5%** w skali miesiąca (lub wyższy), z uwzględnieniem ewentualnych zaplanowanych okien konserwacyjnych. SLA powinno precyzować zasady postępowania w przypadku awarii lub przerw w świadczeniu usługi, w tym ewentualne mechanizmy kompensacyjne lub kary umowne dla Wykonawcy za niedotrzymanie gwarantowanego poziomu dostępności.
2. **Lokalizacja danych:** Wszystkie dane przetwarzane i przechowywane w ramach usługi muszą **znajdować się na terenie Unii Europejskiej**. Wykonawca zobowiązany jest wskazać lokalizację (kraje i centra danych), w których będą przechowywane dane Zamawiającego oraz odbywać się będą obliczenia. Preferowane jest, aby infrastruktura była zlokalizowana na terenie **Polski lub innego kraju UE** zapewniającego pełną zgodność z regulacjami dotyczącymi ochrony danych (w szczególności **RODO**). Dane Zamawiającego nie mogą być przekazywane ani przechowywane poza obszarem UE bez uprzedniej pisemnej zgody Zamawiającego.
3. **Certyfikaty bezpieczeństwa:** Wykonawca musi posiadać aktualne certyfikaty potwierdzające wysoki standard bezpieczeństwa informacji i zarządzania usługą. W szczególności wymagany jest certyfikat **ISO/IEC 27001** (System Zarządzania Bezpieczeństwem Informacji) lub równoważny certyfikat w zakresie bezpieczeństwa informacji. Dodatkowo, mile widziane jest posiadanie certyfikatów **ISO/IEC 27017** (bezpieczeństwo usług cloud) oraz **ISO/IEC 27018** (ochrona danych osobowych w chmurze) lub innych równoważnych świadectw potwierdzających bezpieczeństwo środowiska chmurowego dostawcy. Wykonawca powinien przestrzegać wszystkich obowiązujących przepisów prawa w zakresie ochrony danych i bezpieczeństwa oraz stosować najlepsze praktyki branżowe (np. szyfrowanie danych w spoczynku i w transmisji, właściwe mechanizmy autoryzacji dostępu do zasobów itp.).
4. **Monitoring i rozliczalność:** Usługa musi zapewniać mechanizmy monitorowania wykorzystania zasobów oraz transparentnego rozliczania. Wykonawca udostępni Zamawiającemu narzędzia (np. panel administracyjny, API lub raporty cykliczne) pozwalające na bieżące śledzenie parametrów pracy instancji i zużycia zasobów, takich jak obciążenie GPU/CPU, ilość wykorzystanej pamięci RAM, zajętość przestrzeni dyskowej, transfer danych itp. Dane z monitoringu powinny stanowić podstawę do naliczania opłat i być dostępne dla Zamawiającego w celu weryfikacji prawidłowości rozliczeń. Rozliczenia przedstawiane Zamawiającemu muszą być **jasne, weryfikowalne i oparte o zmierzone zużycie**. Wykonawca powinien zapewnić również mechanizmy alarmowania lub powiadomień (np. gdy zużycie zbliża się do określonych progów) w celu ułatwienia Zamawiającemu kontrolowania kosztów i efektywnego zarządzania wykorzystaniem usługi.

Dokument ten określa minimalne wymagania oraz oczekiwania Zamawiającego wobec usługi. Wykonawcy składający ofertę powinni upewnić się, że wszystkie powyższe warunki zostaną spełnione. Jakiekolwiek odstępstwa od wymagań muszą zostać wyraźnie zaznaczone w ofercie i mogą skutkować jej odrzuceniem. Zamawiający zastrzega sobie prawo do weryfikacji oferowanych parametrów oraz do odrzucenia oferty niespełniającej wymagań określonych w niniejszej specyfikacji.

**Współczynnik wzrostu wydajności**

|  |  |
| --- | --- |
| **Model karty** | **Współczynnik wzrostu wydajności** |
| **A100** | **1** |
| **H100** | **5.03** |
| **H200** | **5.83** |
| **B100** | **12.48** |
| **B200** | **16.07** |
| **B300** | **23.31** |

Jak czytać tą tabelkę: w ocenie przyjmiemy, że np. Karta B300 jest ponad 18 razy wydajniejsza od karty A100 więc 1 godzina pracy karty B300 wykona tyle samo obliczeń co 18 godzin pracy A100.

Dla modeli GPU nieuwzględnionych w opublikowanym zestawieniu, Zamawiający zastrzega sobie prawo do samodzielnego określenia odpowiedniego współczynnika wydajności na etapie oceny ofert, na podstawie danych technicznych i dostępnych benchmarków. Tak ustalone wartości będą miały zastosowanie wyłącznie w bieżącej procedurze i służą wyłącznie do celów porównawczych i umożliwienie rzetelnego porównania każdej z otrzymanych ofert.

**KRYTERIA OCENY**

Oferty będą oceniane na podstawie poniższych **Ceny**  (stawka za godzinę pracy jednej jednostki GPU o wymaganych parametrach podzielona przez współczynnik wzrostu wydajności).

**Formularz ofertowy**

|  |  |  |  |
| --- | --- | --- | --- |
| **Typ Karty** | **CENA**  **Cena brutto za godzinę (pln) przy zakupie do 50 godzin miesięcznie** | **CENA**  **Cena brutto za godzinę (pln) przy zakupie do 100 godzin miesięcznie** | **CENA**  **Cena brutto za godzinę (pln) przy zakupie do 200 godzin miesięcznie** |
|  |  |  |  |