
Równoważenie proxy według obciążenia

Pula wirtualna wysyła maszyny z GPU
tam, gdzie potrzebują ich pracownicy

Dokument wyjaśniający dla działu IT

Cellule-PRO, wersja 0.2.158

Do zapamiętania

Wasze serwery Cellule-PRO (*pule*) to zwykłe serwery. Moc obliczeniową zapewniają maszyny z kartą graficzną (*proxy*). Zamiast na zawsze przypisywać każdą maszynę do jednej lokalizacji, możecie pozwolić jej zgłaszać się do *puli wirtualnej* klastra, która wysyła ją tam, gdzie czeka się na odpowiedzi: najpierw do puli, która nie ma już żadnego silnika, potem do puli najbardziej obciążonej. Gdy ruch się uspokoi, maszyna wraca. **Nic się nie przesuwa bez Waszej zgody:** domyślnie wszystko jest wyłączone.

Spis treści

1	W dwie minuty	2
2	Jak to działa	2
2.1	Trzy zasady puli wirtualnej	2
2.2	Droga proxy	3
2.3	A jeśli pula padnie?	3
3	Ustawienia krok po kroku	3
3.1	Krok 1 — wybierzcie proxy, które mogą się przesuwać	4
3.2	Krok 2 — włączcie przenoszenie według obciążenia	4
3.3	Krok 3 — opóźnienie przełączenia w razie awarii	4
3.4	Krok 4 — sprawdźcie w Nadzorze	5
4	Zalety i wady	5
4.1	Zalety	5
4.2	Wady i ograniczenia	5
5	Czy włączyć to u Was?	6
6	Najczęstsze pytania	6

1 W dwie minuty

Wyobraźcie sobie firmę taksówkową z trzema agencjami. Każda agencja ma swoje okienko (*pule*), ale samochody (*proxy*, Wasze maszyny z kartą graficzną) nie muszą przez cały dzień stać zaparkowane przed tym samym okienkiem. Rano jedna agencja jest zawałona pracą, a inna czeka na klientów; centrala radiowa (*pula wirtualna*) wysyła wtedy wolny samochód tam, gdzie klienci stoją w kolejce. Wieczorem samochód wraca do centrali.

Cellule-PRO robi dokładnie to samo z mocą obliczeniową:

- **Pule** przyjmują pytania pracowników swojej lokalizacji. To zwykle serwery Linux, karta graficzna nie jest w nich wymagana.
- **Proxy** to mocne maszyny, które tworzą odpowiedzi (jedna lub kilka kart graficznych, jeden lub kilka modeli).
- **Pula wirtualna** to jeden wspólny adres klastra (wirtualny adres IP, czyli VIP, przejmowany po kolei przez jedną z pul). To pod tym adresem zgłaszają się proxy.

Do zapamiętania

- **Przesuwają się tylko proxy.** Nigdy komputery pracowników, nigdy pule, nigdy dane.
- **Żadna pula nie zostaje pominięta.** Pula, która nie ma już żadnego silnika wnioskowania, jest obsługiwana jako pierwsza.
- **Każda pula zachowuje co najmniej jedno proxy na model:** nie rozbiera się jednej lokalizacji, żeby ubrać inną.
- **Dwa przełączniki, domyślnie wyłączone:** jeden dla klastra, jeden dla każdego proxy.

2 Jak to działa

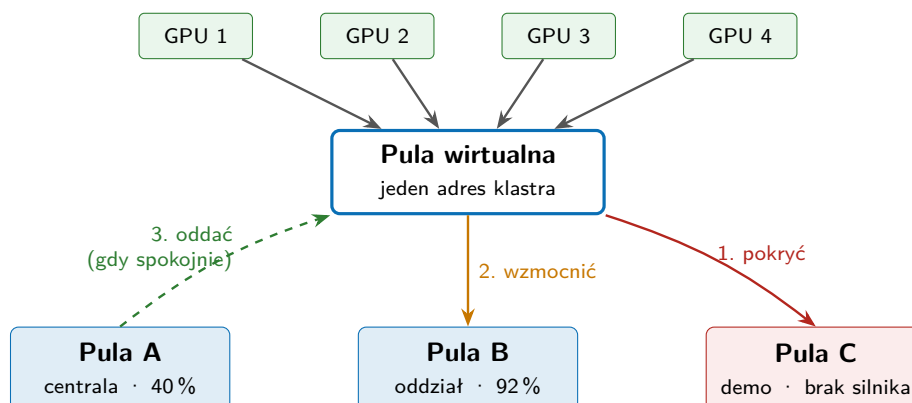
2.1 Trzy zasady puli wirtualnej

Co minutę pula, która ma adres wirtualny, sprawdza proxy *przybyłe przez ten adres* oraz obciążenie każdej puli klastra. Stosuje trzy zasady, w tej kolejności:

Zasada	Co się dzieje
1. Pokryć	Pula klastra nie ma żadnego silnika wnioskowania (pula demonstracyjna bez karty graficznej, świeżo zainstalowana pula, pula, z której odeszły maszyny): dostaje mobilne proxy w pierwszej kolejności , nawet jeśli sama pula wirtualna jest obciążona.
2. Wzmocnić	Pula przekracza próg górny (domyślnie 80 %): dostaje mobilne proxy, pod warunkiem że pula wirtualna bez tego proxy sama pozostanie poniżej progu górnego.
3. Oddać	Pula pozostaje poniżej progu dolnego (domyślnie 60 %) przez minimalny czas (5 min): oddaje puli wirtualnej proxy, które jej powierzono. Nigdy ostatniego proxy danego modelu, nigdy jedynego proxy puli.

Obciążenie puli to odsetek jej silników wnioskowania zajętych odpowiadaniem: 3 zajęte silniki z 4 = 75 %. Pula, która nie ma żadnego silnika, liczy się jako 100 %: nie może nikomu odpowiedzieć.

2.2 Droga proxy



Pracownicy każdej lokalizacji nadal piszą do swojej puli; miejsce zmieniają tylko maszyny z GPU.

Pula zajmie się tym sama

Proxy, które dostaje polecenie przeniesienia, **najpierw kończy trwające odpowiedzi** (do jednej minuty), a potem w ciągu kilku sekund łączy się z pulą docelową przez szyfrowane połączenie. Pracownicy nie widzą niczego poza odpowiedzią, która przychodzi szybciej.

2.3 A jeśli pula padnie?

- **Pada pula, która ma adres wirtualny:** w ciągu kilku sekund adres przechodzi na inną pulę (keepalived). Proxy, które nadal działają, łączą się ponownie pod ten sam adres; nowa pula wirtualna rozmieszcza je na nowo. Żadna centralna baza nie ginie: takiej bazy nie ma.
- **Pada inna pula:** proxy, które w niej pracowały, przełączają się na inną pulę klastra po upływie opóźnienia przełączenia (domyślnie 5 min, do ustawienia); dalej więc odpowiadają, a pula wirtualna rozmieszcza te, które wracają przez jej adres.
- **Pula wraca bez żadnego silnika:** zasada „pokryć” wysyła jej proxy już w następnej minucie.

3 Ustawienia krok po kroku

Cztery ustawienia, w tej kolejności. Pierwsze trzy wykonuje się **tylko raz, z dowolnej puli**: są one kopiowane na wszystkie pule klastra.

Najpierw: instalujcie proxy na adresie puli wirtualnej

Przy instalacji proxy jako adres puli podajcie **wirtualny adres klastra** (lub jego nazwę domenową), a nie adres konkretnej puli. Proxy zainstalowane na adresie konkretnej puli pozostaje do niej przypisane: działa, ale pula wirtualna nie będzie go rozmieszczać.

3.1 Krok 1 — wybierzcie proxy, które mogą się przesuwać

Zakładka „Połączone workery”, karta „Proxy przenoszone według obciążenia”. Zaznaczcie kolumnę „Mobilny” dla każdej maszyny, którą zgadzacie się przenosić.

Połączone workery › Proxy przenoszone według obciążenia			
Proxy	Model	Połączony	Mobilny
gpu-salle-serveur-1	35B	z tą pulą	☒ przenośny
gpu-salle-serveur-2	35B	z tą pulą	☒ przenośny
gpu-vision	vision	z tą pulą	☐ zostaje w swojej puli

Rada: zostawcie *w swojej puli* maszynę, którą lokalizacja musi mieć bez względu na wszystko (na przykład jedyną maszynę danej lokalizacji do odczytu dokumentów), a mobilnymi uczynicie maszyny „wsparcia”.

3.2 Krok 2 — włączcie przenoszenie według obciążenia

Zakładka „Równoważenie workerów”, karta „Obciążenie klastra i rozmieszczanie proxy”. Zaznaczcie „Przeń proxy według obciążenia (klastry)”, sprawdźcie progi, a potem zapiszcie.

Równoważenie workerów › Obciążenie klastra i rozmieszczanie proxy

☒ **Przeń proxy według obciążenia (klastry)** włączone (dotyczy tylko proxy „mobilnych”)

Próg górny % Próg dolny % Minimalny czas w puli s

Ustawienie	Domyślnie	Skutek
Próg górny	80 %	Powyżej niego pula dostaje wsparcie. Niżej = wsparcie wcześniej, więcej przeniesień.
Próg dolny	60 %	Poniżej niego pula oddaje wsparcie. Zachowajcie co najmniej 15–20 punktów różnicy względem progu górnego, aby uniknąć przenoszenia tam i z powrotem.
Minimalny czas w puli	300 s	Czas, przez który proxy zostaje tam, gdzie je wysłano, zanim będzie mogło znów się przenieść. Dłużej = stabilniej.

Tabela nad tymi ustawieniami pokazuje dla każdej puli jej obciążenie, oczekujące pytania i silniki (zajęte / wszystkie): obserwujcie ją przez kilka dni przed włączeniem, aby dobrać progi do Waszej działalności.

3.3 Krok 3 — opóźnienie przełączenia w razie awarii

Zakładka „Ustawienia zaawansowane”, kategoria „Asystent działu IT (alerty)”, ustawienie „Opóźnienie przed przełączeniem proxy na inną pulę (sekundy)”.

Ustawienia zaawansowane › Asystent działu IT (alerty)

Opóźnienie przed przełączeniem proxy na inną pulę (sekundy) od 30 do 3600

Krótkie (30–60 s): szybkie przełączenie przy prawdziwej awarii, ale kilkusekundowa przerwa w sieci może przenieść proxy. Długie (5 min i więcej): większa stabilność. To opóźnienie nie dotyczy proxy zainstalowanych na adresie puli wirtualnej, które podążają za adresem w ciągu kilku sekund.

3.4 Krok 4 — sprawdźcie w Nadzorze

Otwórzcie „Nadzór”: przeniesiona maszyna pojawia się pod nową pulą przy następnym odświeżeniu. Historię poleceń (proxy, skąd, dokąd, stan) można przeglądać w zakładce „Równoważenie workerów”, karta „Historia migracji”, a każda decyzja puli wirtualnej jest zapisywana w jej dzienniku wraz z powodem (*pokryć*, *wzmocnić* lub *oddać*).

Uwaga

Profil instalacji obowiązuje także przy rozmieszczaniu: proxy nigdy nie jest wysyłane do puli o innym profilu (dokument „Profile instalacji”). Klaster w pełnym air-gapie nigdy więc nie używa maszyny puli otwartej na świat zewnętrzny.

4 Zalety i wady

4.1 Zalety

Mniej kart graficznych do kupienia	Maszyna obsługuje lokalizację, która właśnie jej potrzebuje, zamiast stać bezczynnie w spokojnej lokalizacji. Wymiarujecie park maszyn na szczyt całego klastra, a nie na sumę szczytów każdej lokalizacji.
Żadna lokalizacja bez odpowiedzi	Pula, która traci swoje maszyny, albo nowa lokalizacja bez karty graficznej jest obsługiwana w pierwszej kolejności.
Odporność na awarie	Pula pada: jej maszyny dalej pracują i zostają rozdzielone na nowo. Brak serwera centralnego, brak jednej wspólnej bazy.
Tylko jeden adres do podania	Każde proxy instaluje się z adresem klastra; dodanie lokalizacji nie wymaga ponownej instalacji żadnej maszyny.
Kontrola zostaje u Was	Dwa przełączniki domyślnie wyłączone, wybór maszyna po maszynie, regulowane progi, najwyżej 6 przeniesień na godzinę, historia każdego polecenia.
Nic innego się nie przesuw	Ani komputery, ani dane, ani konta: pracownicy zawsze piszą do swojej puli.

4.2 Wady i ograniczenia

Sieć między lokalizacjami	Maszyna wysłana do odległej lokalizacji odpowiada przez Wasze połączenie między lokalizacjami (szyfrowane). Przy wolnym łączu każda odpowiedź przychodzi trochę później niż lokalnie; treść pytań i odpowiedzi przechodzi przez to łącze.
----------------------------------	---

Kilka sekund przejścia	Podczas przenoszenia maszyna kończy swoje odpowiedzi, a potem łączy się ponownie: przez tę krótką chwilę nie przyjmuje nowych pytań.
Pomiar na razie ogólny	Obciążenie mierzy się dla całej puli, a nie model po modelu: pula przeciążona odczytem dokumentów, ale spokojna przy tekście, może nie dostać właściwego wsparcia. Pomiar dla każdego modelu jest planowany w jednej z kolejnych wersji.
Potrzebny jest adres wirtualny	Bez adresu wirtualnego (jedna pula albo brak keepalived) nie ma puli wirtualnej, a więc nie ma rozmieszczania.
Bez włączenia nic się nie dzieje	Przełącznik wyłączony: pula bez silnika zostaje bez silnika. W Nadzorze planowany jest alert „ta pula nie ma żadnego silnika wnioskowania”.
Maszyny zmieniające lokalizację	Przy audycie maszyna, która odpowiedziała na pytanie, nie musi już być maszyną danej lokalizacji. Historia poleceń pozwala to ustalić.

5 Czy włączyć to u Was?

Wasza sytuacja	Rada
Jedna pula	Niepotrzebne: nie ma dokąd wysłać maszyny.
Kilka pul w tej samej lokalizacji (ten sam budynek, szybka sieć)	Tak. Zysk jest największy, a koszt sieciowy zerowy.
Kilka lokalizacji połączonych dobrym łączem	Tak , zostawiając w każdej lokalizacji jedną maszynę „w swojej puli” i czyniąc mobilnymi maszyny wsparcia.
Lokalizacje połączone wolnym łączem	Ostrożnie: podnieście próg górny (85–90 %) i wydłużcie minimalny czas, aby przenosić tylko przy prawdziwym przeciążeniu albo gdy pula nie ma silnika.
Pula demonstracyjna lub testowa	Tak: często nie ma karty graficznej, a zasada „pokryć” używa jej maszyny.

6 Najczęstsze pytania

Czy maszyna może odejść, gdy jej użytkownicy są w trakcie rozmowy?

Przed odejściem kończy trwające odpowiedzi (do jednej minuty). Sama rozmowa jest przechowywana przez pulę pracownika, która się nie przesuwa.

Czy można zabronić przenoszenia konkretnej maszyny?

Tak: zostawcie jej pole „Mobilny” niezaznaczone. To stan domyślny każdej nowej maszyny.

Czy pula wirtualna jest „nadrzędna”?

Nie. To tymczasowa rola związana z adresem wirtualnym: jeśli pula, która go ma, się zatrzyma, inna go przejmuje i działa dalej, nic nie tracąc, bo nie ma żadnego stanu centralnego.

Co się dzieje, gdy dwie pule są jednocześnie mocno obciążone?

Pula wirtualna obsługuje najpierw pulę bez żadnego silnika, potem najbardziej obciążoną, jedno polecenie naraz, nigdy nie przekraczając sama progu górnego i nigdy nie zabierając ostatniego proxy danego modelu.

Czy moje stanowiska pracy (CPU pracowników) są przenoszone?

Nigdy. Przenosić można tylko proxy, czyli dedykowane maszyny z kartą graficzną.