Politechnika Wrocławska

Wydział Elektroniki

Kurs: Inżynieria Ruchu 2

Temat projektu:

Porównanie scenariuszy No_Load_Balancing, Random, Server_Load w środowisku Riverbed Modeler

Autorzy projektu:

1. Dominik Modrzejewski
2. Paweł Ptaszyński

Wrocław, 2020

1 Load balancing – definicja

Load balancing jest techniką równoważenia obciążenia pomiędzy wieloma komputerami, dyskami, klastrami, połączeniami, procesorami i innymi zasobami sieciowymi. Tą technikę możemy wykorzystać do optymalizacji wykorzystania zasobów, obniżenia czasu odpowiedzi i uniknięcia przeładowania danego zasobu sieciowego.

1.1 W jakim celu istotna jest analiza sieciowa z wykorzystaniem techniki load balancing'u?

Analiza sieciowa z wykorzystaniem techniki load balancing'u pozwala na określenie, o ile zostaną zredukowane obciążenia na poszczególnych serwerach za pomocą techniki load balancing'u. Analiza ta również może zbadać jak zachowają się serwery, jeśli jeden lub kilka z nich ulegnie awarii. Całość tej analizy sieciowej powinna doprowadzić do ustalania konkretnych zasad równoważenia obciążania dla analizowanej sieci.

1.2 W jaki sposób można zmieniać konfigurację load balancer'a w tym projekcie?

Konfiguracja Load Balancer'a jest wykonywana na samym węźle modułu Load Balancer'a, który zawiera dwa atrybuty Load Balancer Address i Load Balancer Configuration.
W atrybucie Load Balancer Address można ustawić opcje Application: Destination Preference.
Pole Load Balancer Configuration służy do ustalenia zasad konfiguracji Load Balancingu'u. Należy wypełnić trzy kolumny Policy, Application i Candidate Server List.

W projekcie wykorzystaliśmy następujące dostępne scenariusze:

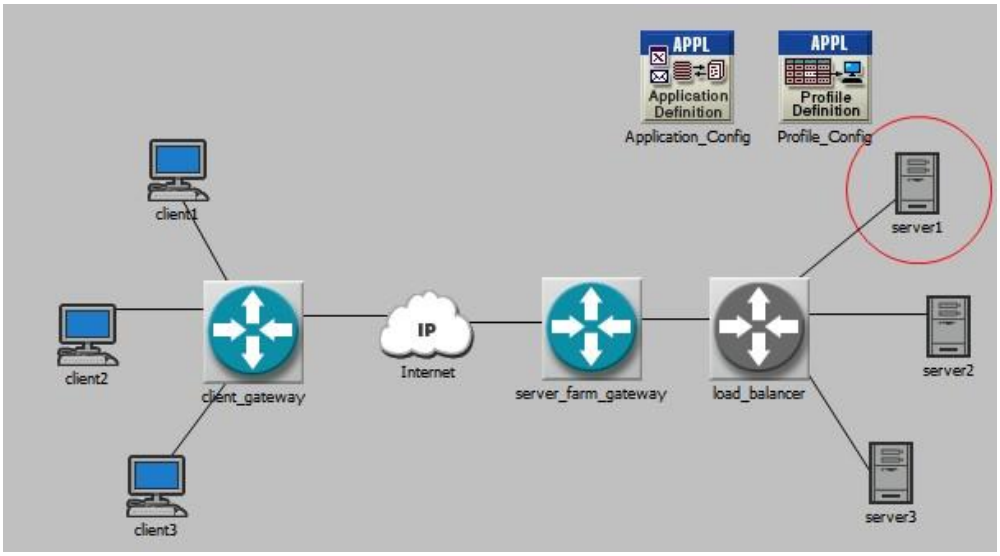
- No Load Balancing
- Random
- Server Load

1.3 Konfiguracja profilu użytkownika dla symulowanej sieci

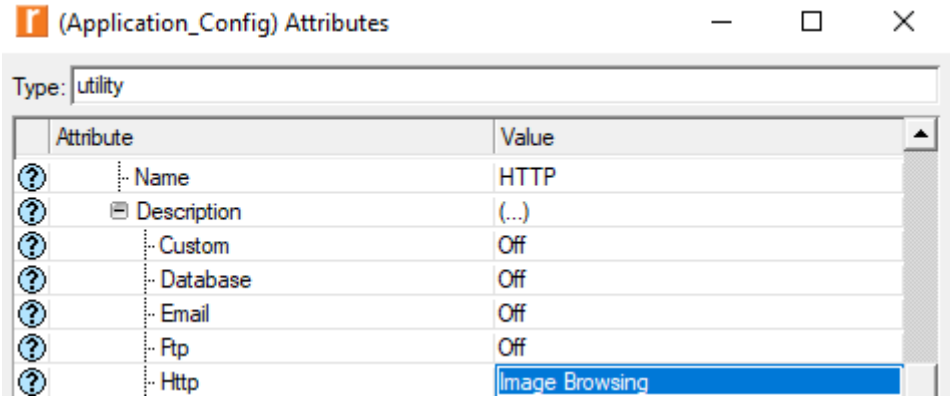
a) No Load Balancing

W scenariuszu No Load Balancing nie są wykorzystywane techniki load balancing'u. Wszystkie żądania HTTP są wysyłane do Server1. Server1 obsłuży cały obciążenie, podczas gdy inne serwery nie będą używane. Jeśli ten serwer ulegnie awarii, wszystkie żądania będą odrzucane do momentu jego naprawy. Analizowana sieć składa się z 3 klientów, którzy podłączeni są do bramy domyślnej oraz do Internetu.
Mamy tutaj również 3 serwery, które są podłączone do Load Balancera i do bramy domyślnej serwerów. Ta sieć również zawiera konfiguracje aplikacji i profilu.
Aplikacja w scenariuszu No Load Balancing jest skonfigurowana na Http Image Browsing. Profil użytkownika dla symulowanej sieci jest skonfigurowany na WebUser.

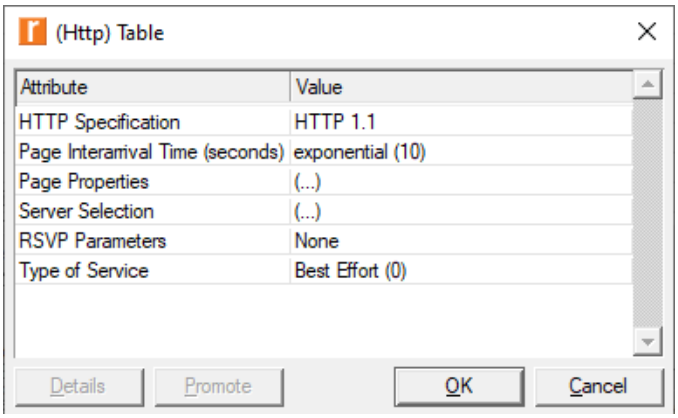
Na poniższych rysunkach została przedstawiona topologia scenariusza No Load Balancing oraz konfiguracja aplikacji oraz usług.



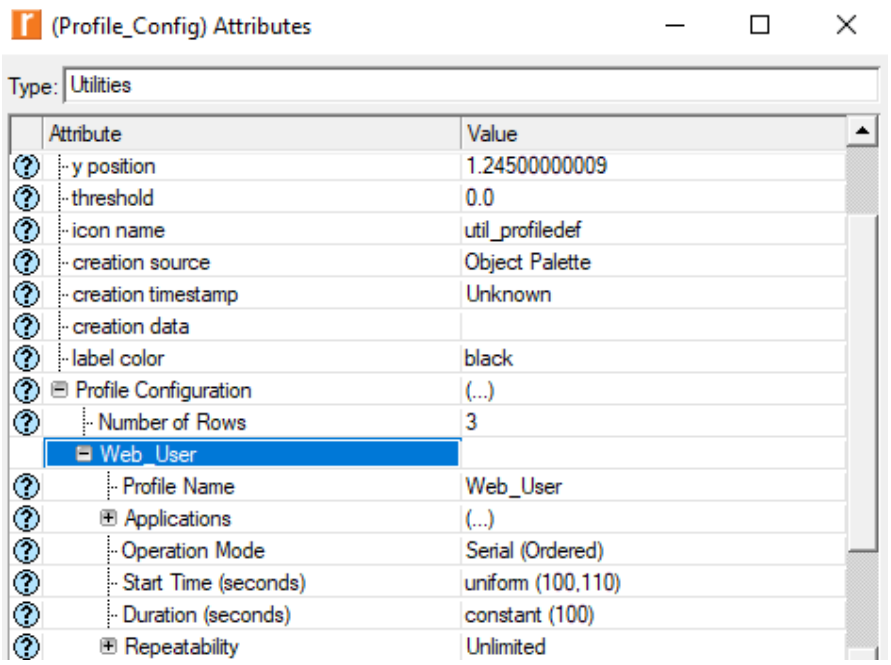
Rysunek 1 Scenariusz No Load Balancing – topologia sieciowa



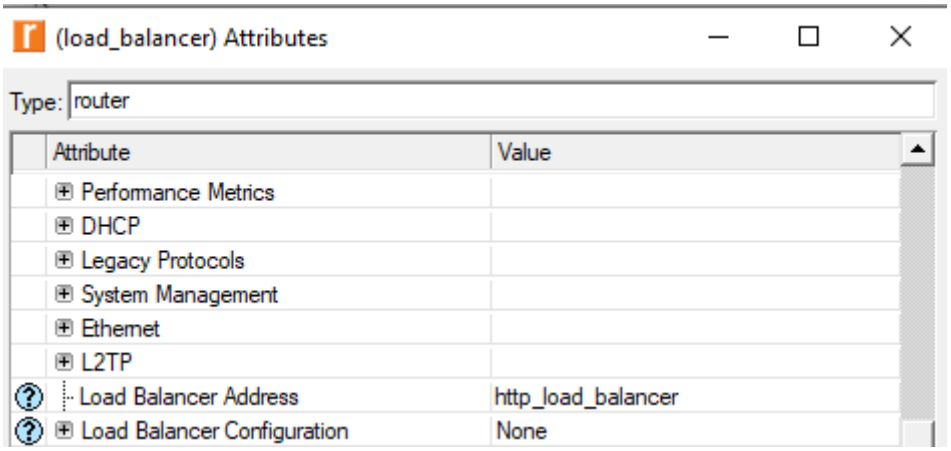
Rysunek 2 Konfiguracja aplikacji ustawiona na Image Browsing



Rysunek 3 Ustawienie wyświetlania stron co 10 sekund

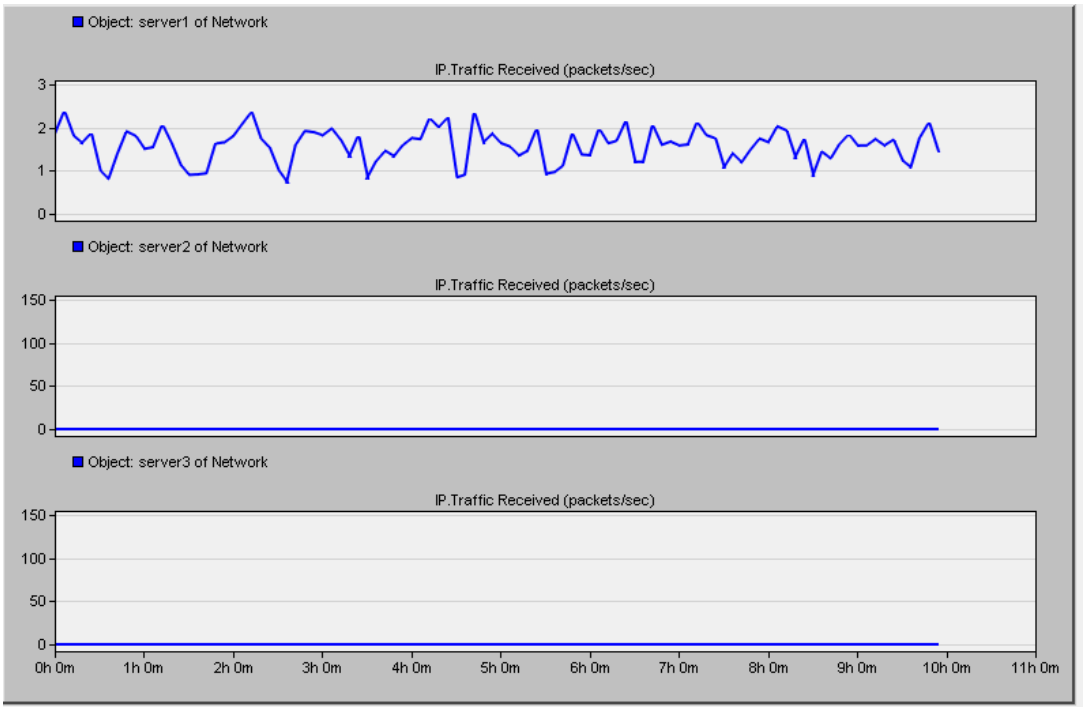


Rysunek 4 Ustawienie profilu użytkownika



Rysunek 5 Scenariusz No Load Balancing - konfigruacja Load Balancera

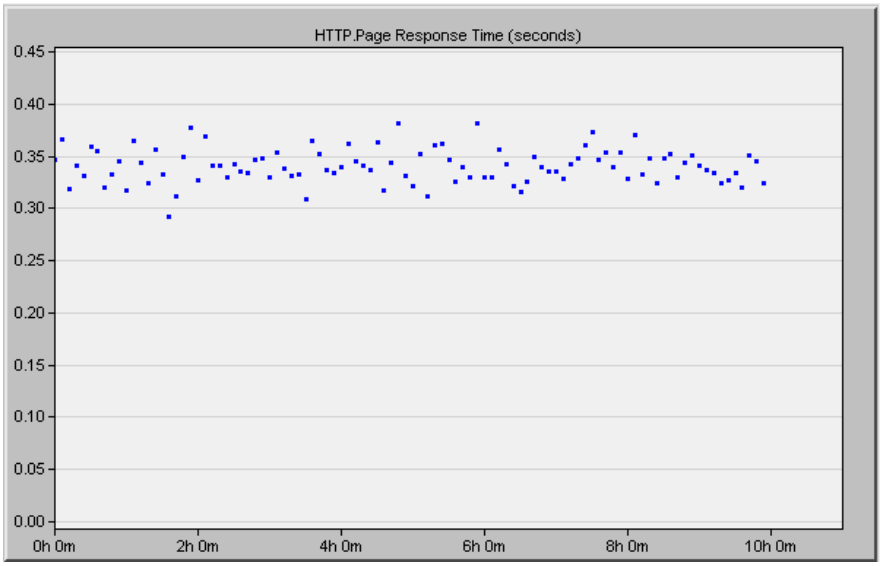
Po konfiguracji została przeprowadzona symulacja, która dotyczyła ruchu otrzymanego. Poniższy rysunek przestawia symulacje ruchu otrzymanego z podziałem na serwery.



Rysunek 6 **Scenariusz No Load Balancing – symulacja ruchu otrzymanego**

Symulacja na rysunku nr 6 pokazuje, że cały ruch sieciowy jest obsługiwany tylko przez Server1, gdyż w tym scenariuszu nie została uwzględniona żadna technika równoważenia obciążenia.

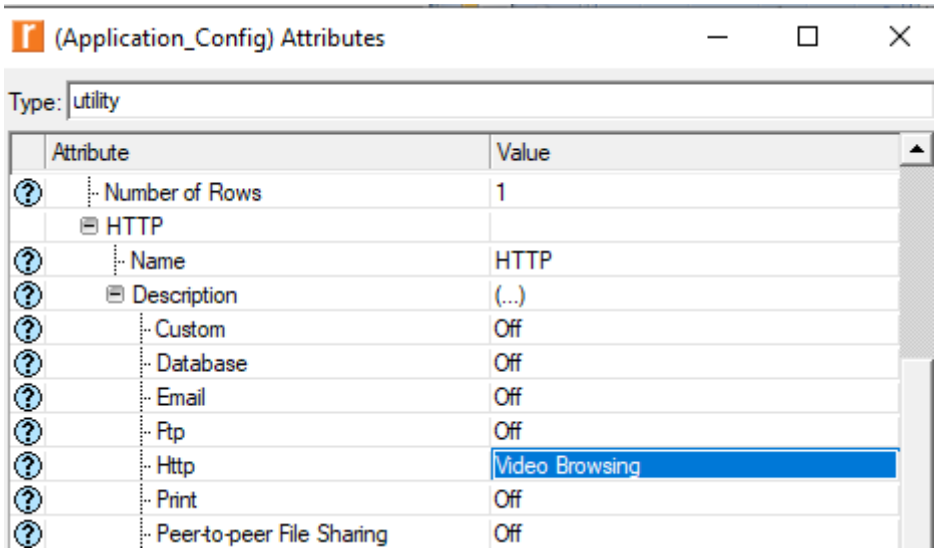
Kolejna symulacja dotyczy czasu odpowiedzi strony, gdy użytkownicy otwierają strony internetowe co 10 sekund.



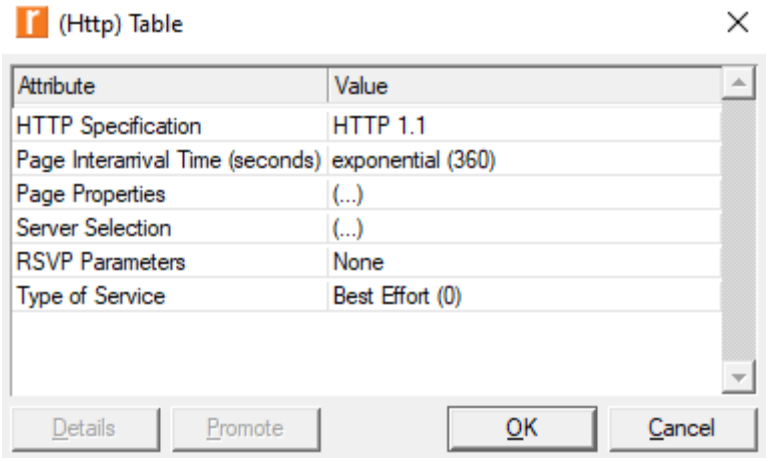
Rysunek 7 Scenariusz No Load Balancing – symulacja czasu wczytywania się stron internetowych

Gdy w ustawieniach każdy klient otwiera stronę z obrazami co 10 sekund, to czas oczekiwania na otwarcie strony wynosi maksymalnie około 0.38 sekundy, co jest akceptowalne przez zwykłego użytkownika.

Aby zwiększyć obciążenie naszej sieci oraz serwerów zmieniamy ustawienia profilu aplikacji z Image Browsing na Video Browsing, wszelkie dokonane zmiany zostały pokazane poniżej.

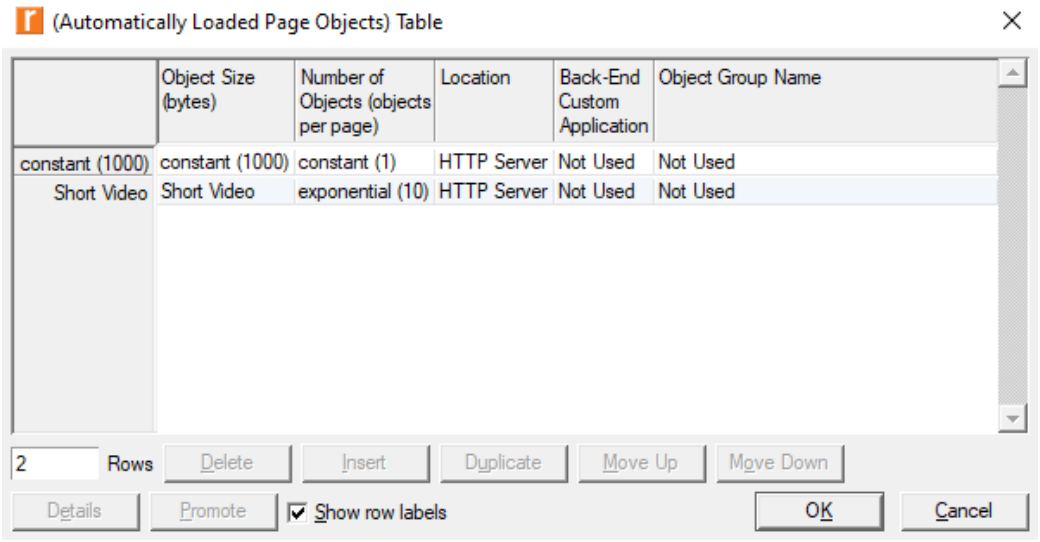


Rysunek 8 Zmiana ustawienia konfiguracji aplikacji



Rysunek 9 Ustawienia aplikacji Video Browsing

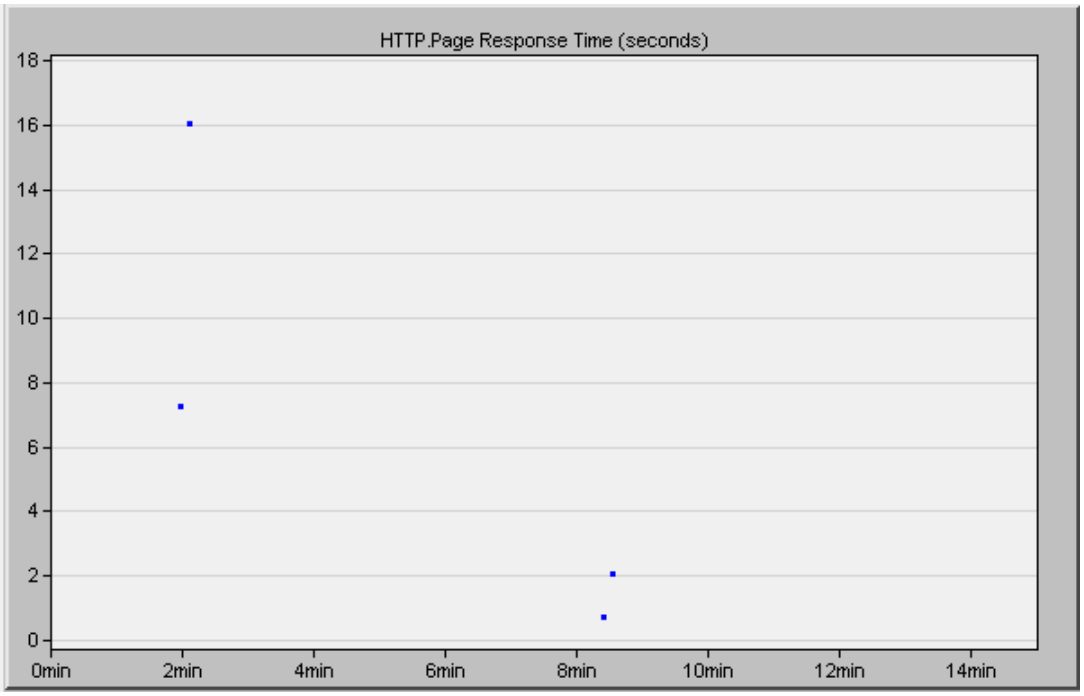
Każdy użytkownicy będą włączali filmy co 360 sekund.



Rysunek 10 Wybór typu treści wideo

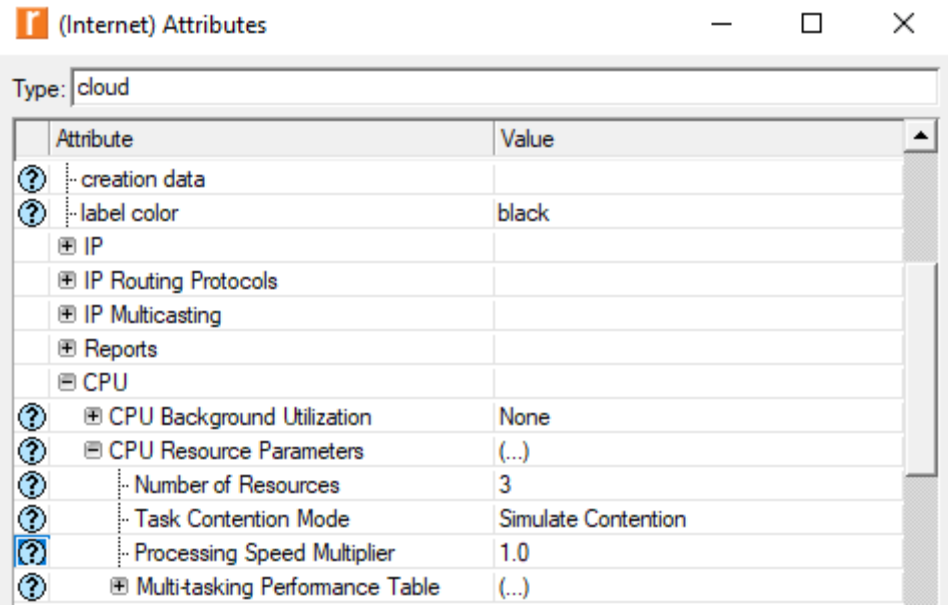
Zdecydowaliśmy, że nasi użytkownicy będą oglądać krótkie filmy.

Następnie przeprowadziliśmy symulacje czasu ładowania się stron, gdy użytkownicy oglądali krótkie filmy wideo. Wyniki tej symulacji znajdują się na rysunku 11.



Rysunek 11 Scenariusz No Load Balancing – symulacja wczytywania się stron wideo

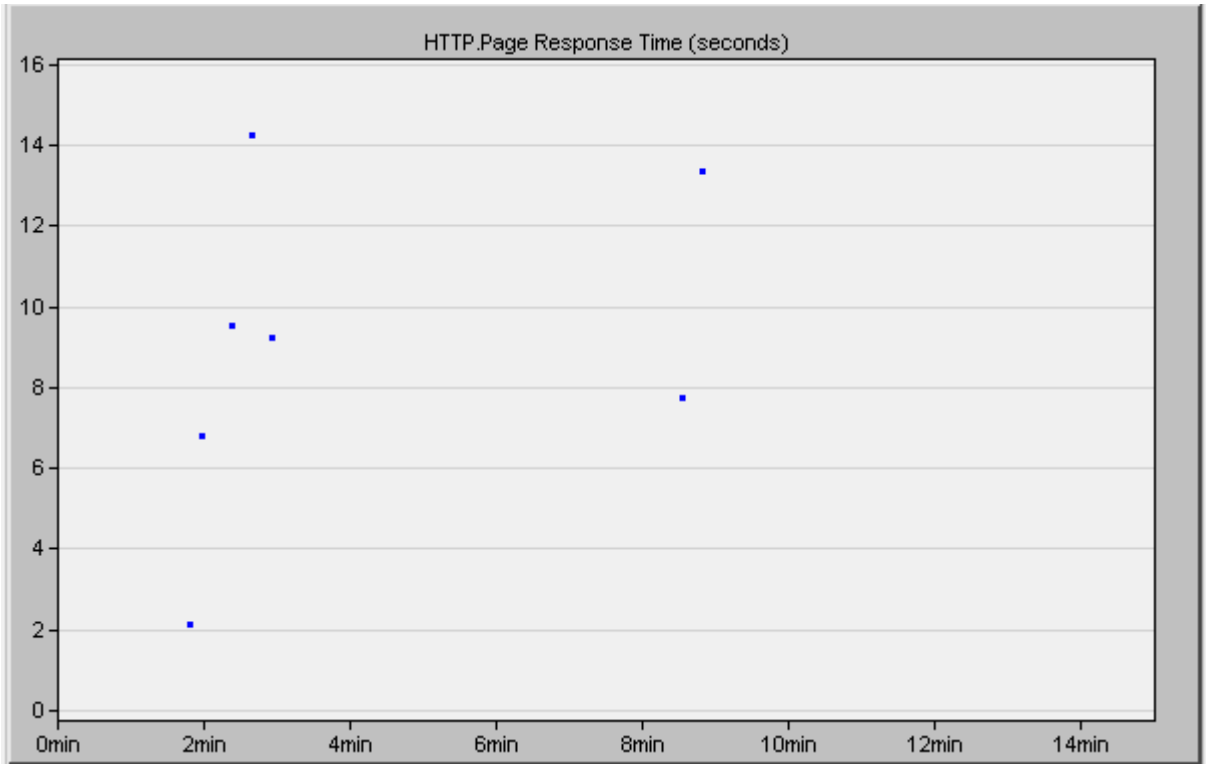
Gdy zmienimy definicję naszej aplikacji na przeglądanie wideo co 360 sekund, to widzimy, że czas oczekiwania na otwarcie strony z filmem wideo może wynosić ponad 16 sekund. Jest to długi czas oczekiwania i przeciętnego użytkownika zirytuje tak długie oczekiwanie, dlatego powinniśmy spróbować zwiększyć wydajność urządzeń sieciowych.
Modyfikacją, jaką dokonaliśmy jest zwiększenie wydajności urządzeń sieciowych.



Rysunek 12 Scenariusz No Load Balancing - zwiększenie wydajności urządzeń sieciowych

Aby zobaczyć, czy polepszyło to czasy oczekiwania na załadowania się strony, zrobiliśmy taką symulację ponownie.

Zwiększyliśmy wydajność urządzeń sieciowych, co pozwoliło niewielkie przyspieszenie działania stron z treścią wideo, co obrazuje rysunek nr 13.



Rysunek 13 Scenariusz No Load Balancing - zwiększenie wydajności urządzeń sieciowych

Widzimy, że czasy oczekiwania na odpowiedź strony zmalały do wartości około 14 sekund, co jest wartością jeszcze zbyt dużą dla użytkownika.

Brak modułu równoważenia obciążania pomiędzy serwerami i kierowanie całego obciążenia sieci na server 1 jest przyczyną zjawiska zbyt długiego wczytywania się stron z treścią wideo.

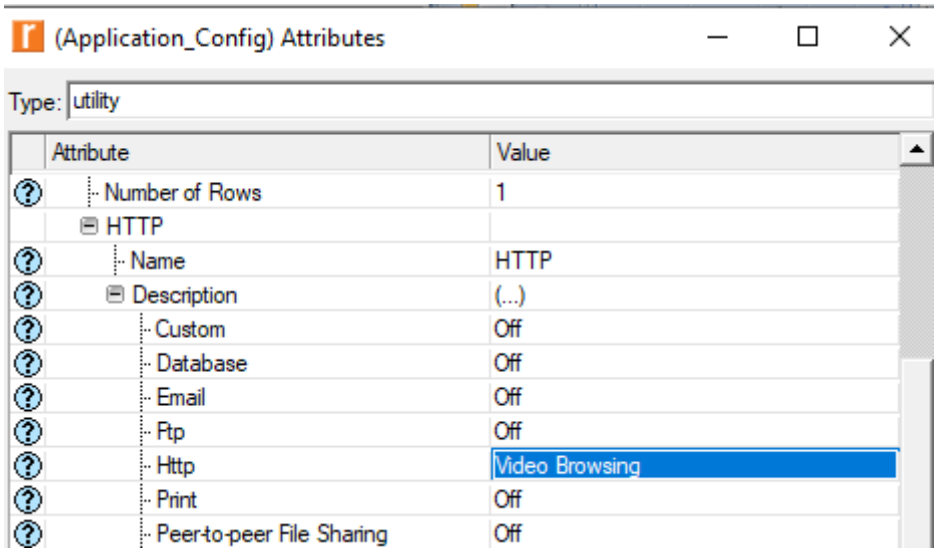
b) Random

Scenariuszu Random moduł Load Balancing'u został tak skonfigurowany, że wybiera losowo między dostępnymi serwerami aplikacji. Został skonfigurowany tak, że korzysta z server3 dwa razy częściej niż z dwóch pozostałych serwerów. Server3 będzie w mniejszym stopniu wykorzystywany, niż jakby miał obsługiwać całe obciążenie. Analizowana sieć składa się z 3 klientów, którzy podłączeni są do bramy domyślnej oraz do Internetu. Mamy tutaj również 3 serwery, które są podłączone do Load Balancera i do bramy domyślnej serwerów. Ta sieć również zawiera konfiguracje aplikacji i profilu.

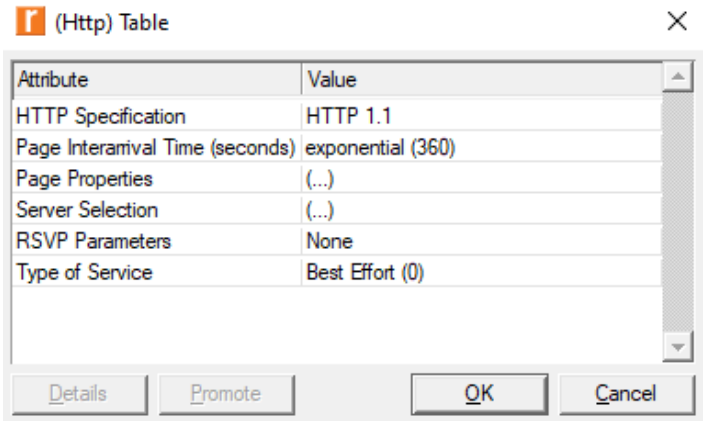
Aplikacja w scenariuszu Random jest skonfigurowana na Video Browsing .

Profil użytkownika dla symulowanej sieci jest skonfigurowany na WebUser.

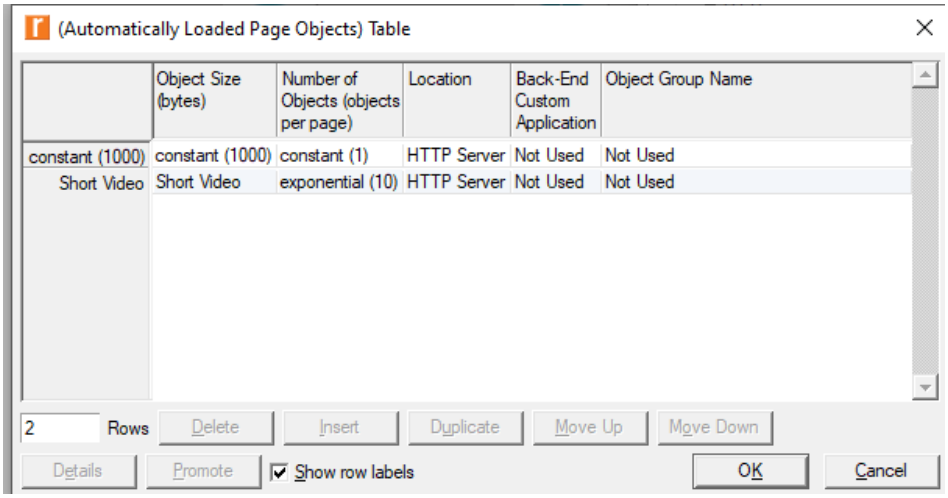
Na rysunku nr 14 oraz rysunku nr 15 przedstawiamy konfiguracje scenariusza Random wraz z jego topologią.



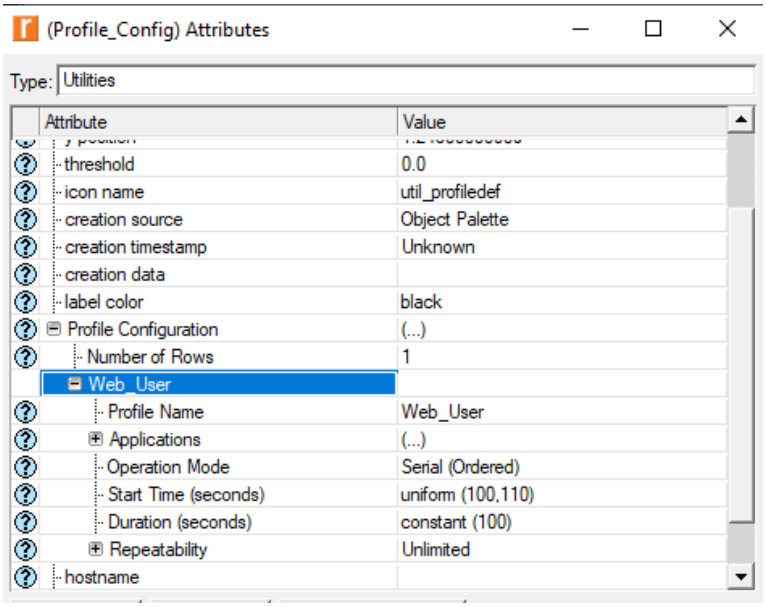
Rysunek 14 Ustawienie profilu aplikacji



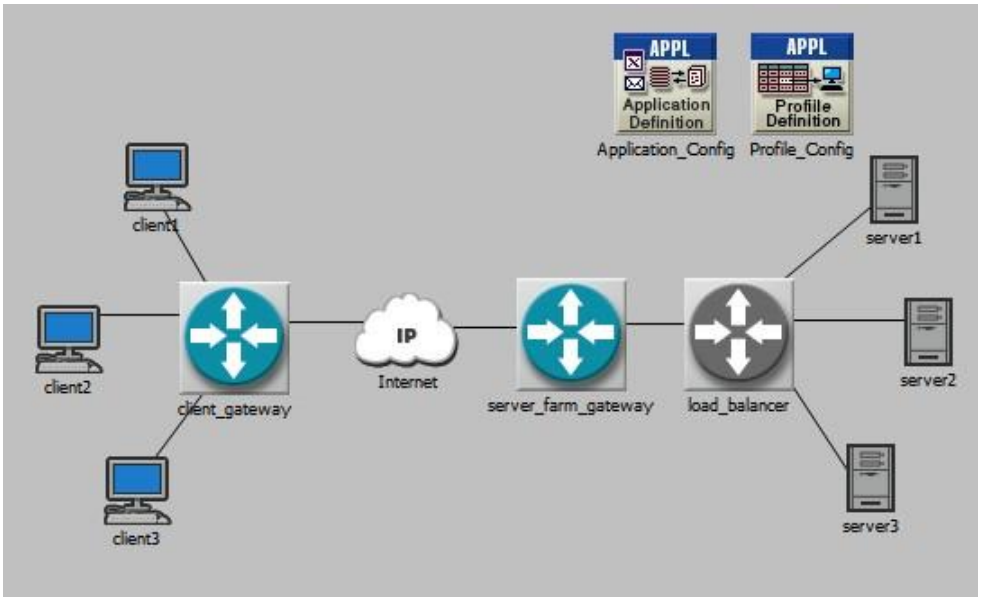
Rysunek 15 Ustawienia aplikacji Video Browsing



Rysunek 16 Wybór typu treści wideo

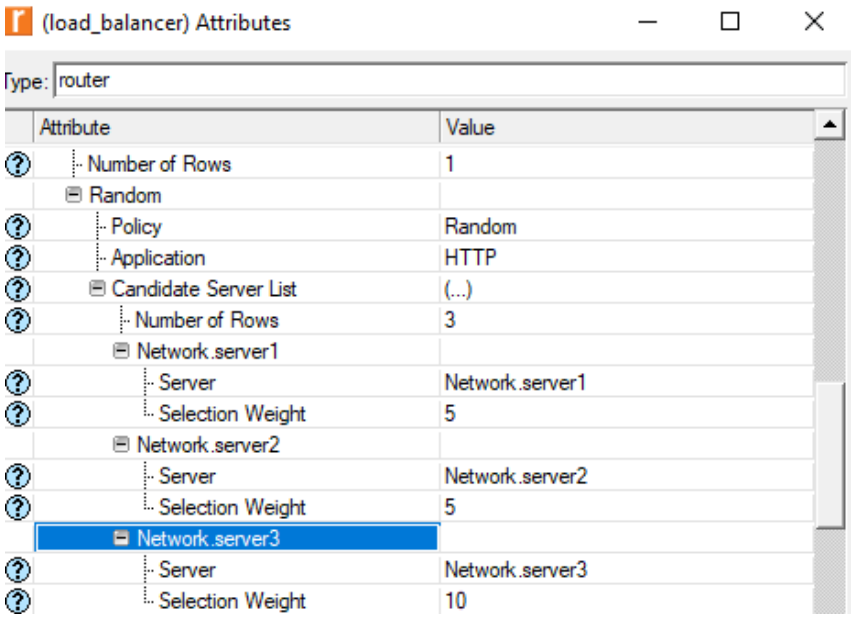


Rysunek 17 Konfiguracja profilu użytkownika



Rysunek 18 Scenariusz Random - Topologia sieciowa

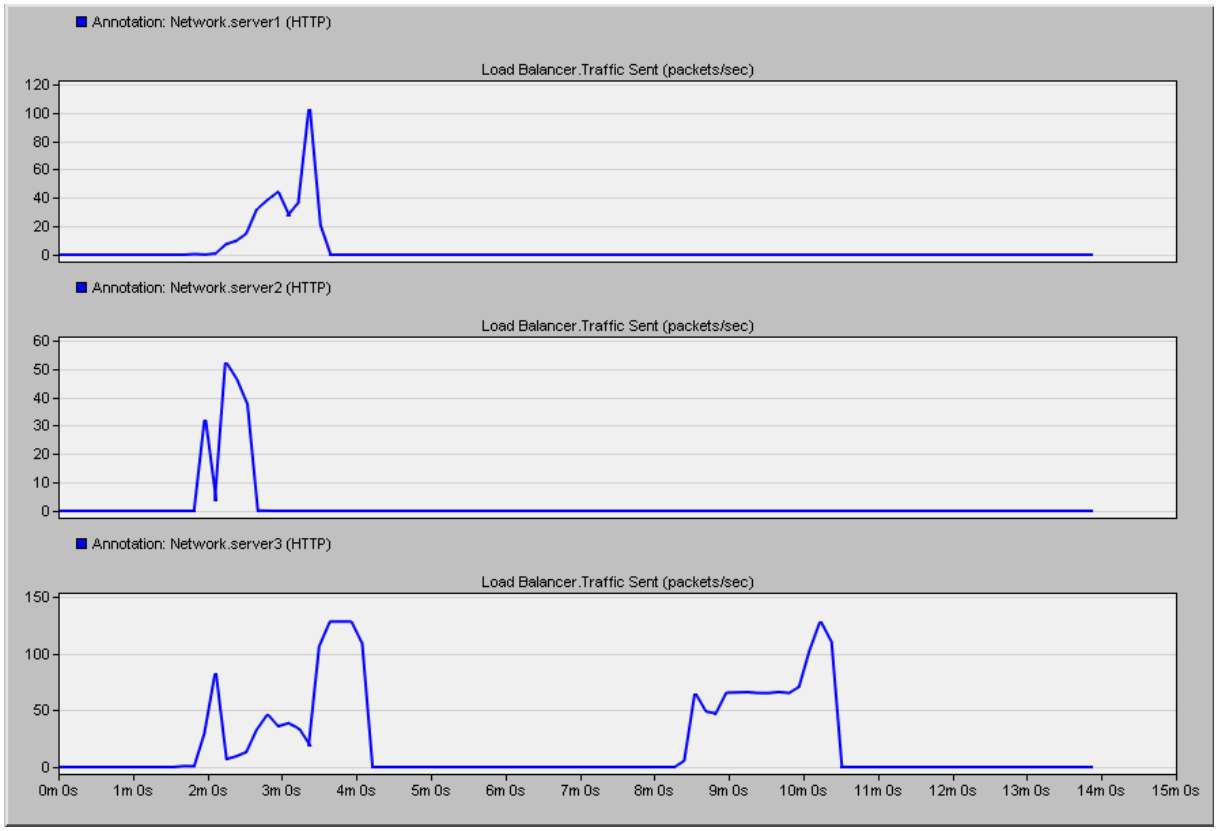
Poniżej przedstawiamy konfiguracje Load Balancera w scenariszu Random, został on tak skonfigurowany, użytkownicy będą korzystać z server3 dwa razy częściej niż z dwóch pozostałych serwerów.



Rysunek 19 Ustawienia Load Balancera

Warto zwrócić uwagę, że atrybut Selection Weight ma zastosowanie tylko wtedy, gdy technika równoważenia obciążenia jest ustawiona na „round-robin” lub „random”.
Służy on do określania serwera do wyboru.
Prawdopodobieństwo wyboru danego serwera to: (waga wyboru) / (waga całkowita).

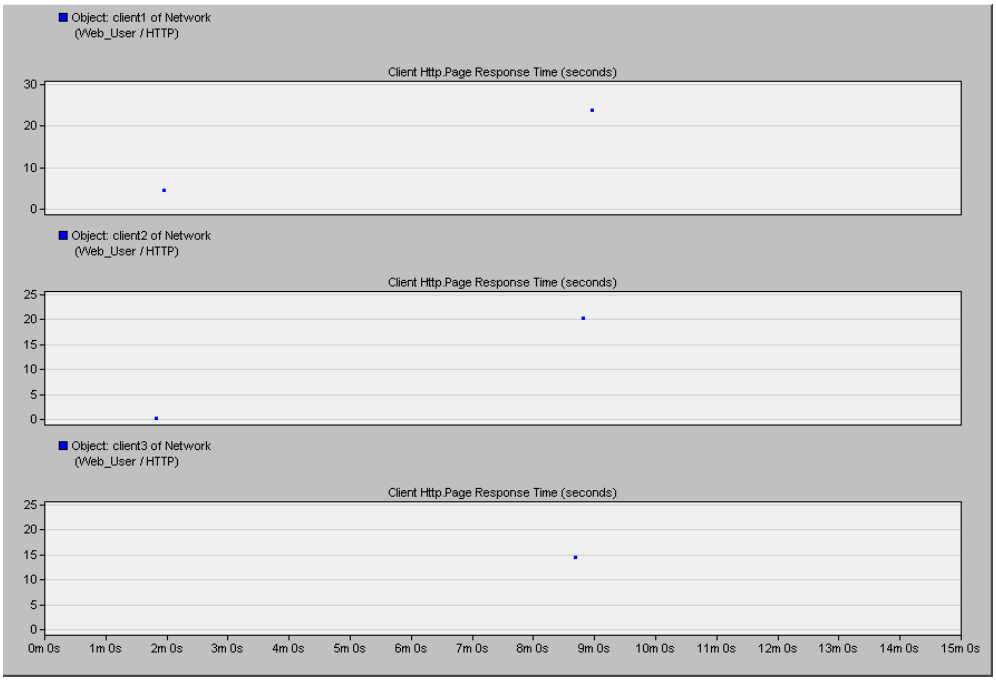
Aby udowodnić kierowanie 2 razy większego ruchu na server 3, niż na server 1 i server 2 została przeprowadzona symulacja ruchu wysłanego przez Load Balancer.



Rysunek 20 Symulacja ruchu wysłanego przez Load Balancer

Scenariusz ten zakłada, że moduł obciążenia sieci ma rozkładać losowo obciążenie, lecz w tym przypadku moduł został tak skonfigurowany, że 2 razy więcej obciążenia jest obsługiwane przez Server3. Co potwierdza rysunek nr 20, gdzie widzimy, że Server3 częściej odbiera pakiety od pozostałych serwerów.

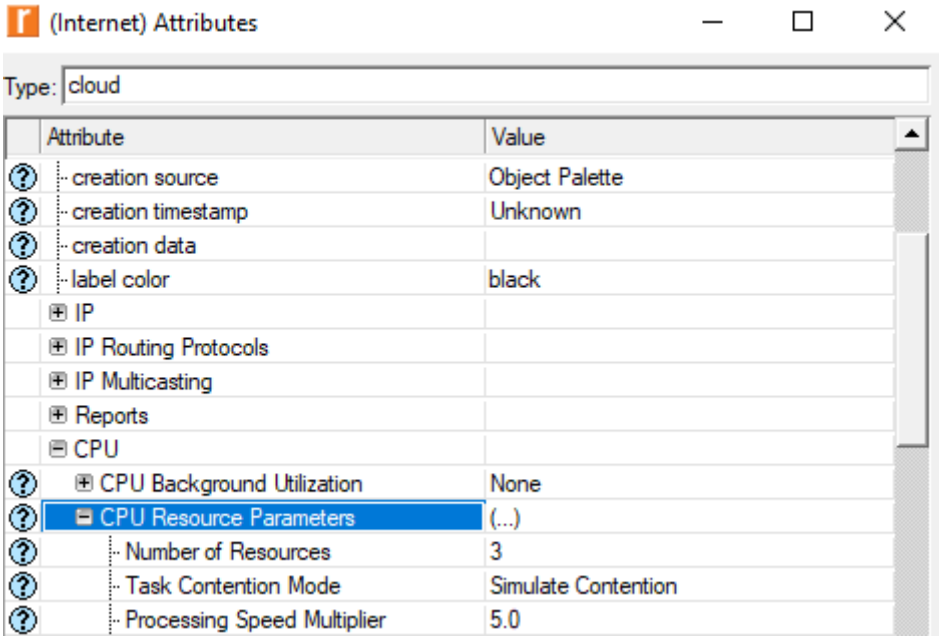
W celu porównanie z scenariuszem No Load Balancing przeprowadzona została symulacja czasu oczekiwania na załadowanie się strony z treścią wideo.



Rysunek 21 Scenariusz Random – symulacja wczytywania się stron wideo

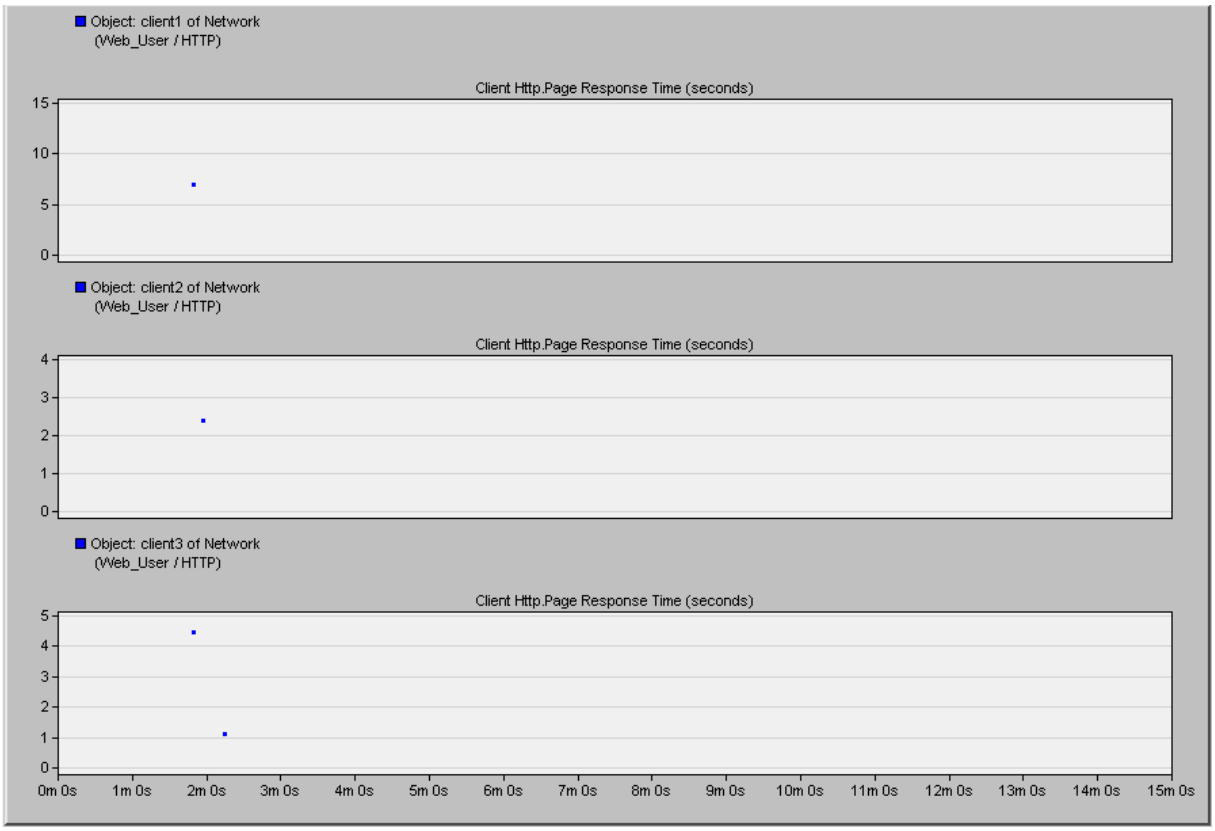
W scenariuszu Random, gdy obciążenie jest rozdzielane, maksymalny czas wczytywania się strony wynosi nawet 20 sekund, jest to gorszy wynik w porównaniu do scenariusza No Load Balancing.

W celu poprawy uzyskiwanych czasów możemy podobnie jak postąpiliśmy w scenariuszu No Load Balancing zwiększyć wydajność urządzeń sieciowych.



Rysunek 22 Zwiększenie parametrów urządzeń sieciowych

Po modyfikacji urządzeń sieciowych została przeprowadzona ponowna symulacja czasu oczekiwania na załadowanie się strony.



Rysunek 23 Czas oczekiwania na załadowanie się strony po ulepszeniu łącza

Poprzez zwiększenie wydajności urządzeń sieciowych udało nam się znacząco zmniejszyć czas oczekiwania na załadowania się strony, która na widocznym powyżej wykresie osiąga wartość około 5 sekund. Nie jest to czas, który zaspokoi każdego użytkownika, ale widzimy, że przyczyną takiej sytuacji jest losowe rozłożenie obciążenia pomiędzy serwerami, gdyż na server3 jest kierowane dwa razy większe obciążenie niż na server1 i server2

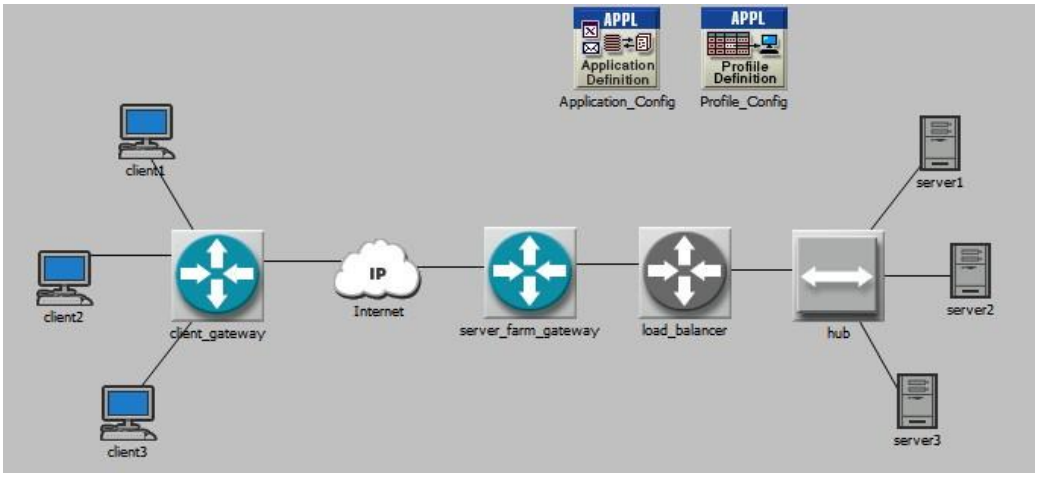
c) Server Load

W scenariuszu Server Load moduł Load Balancing'u wybierze serwer o najmniejszym obciążeniu. Ten scenariusz ilustruje także inną konfigurację, w której moduł Load Balancing'u może być podłączony do serwerów.

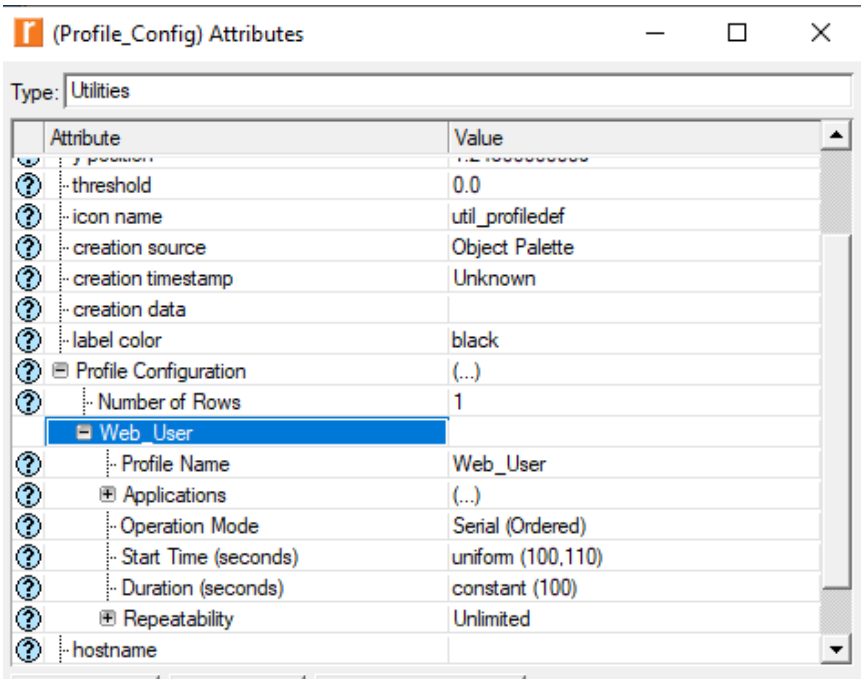
Analizowana sieć składa się z 3 klientów, którzy podłączeni są do bramy domyślnej oraz do Internetu. Mamy tutaj również 3 serwery, które są podłączone koncentratora oraz do Load Balancera i do bramy domyślnej serwerów. Ta sieć również zawiera konfiguracje aplikacji i profilu.

Aplikacja w scenariuszu Random jest skonfigurowana na Video Browsing . Profil użytkownika dla symulowanej sieci jest skonfigurowany na WebUser.

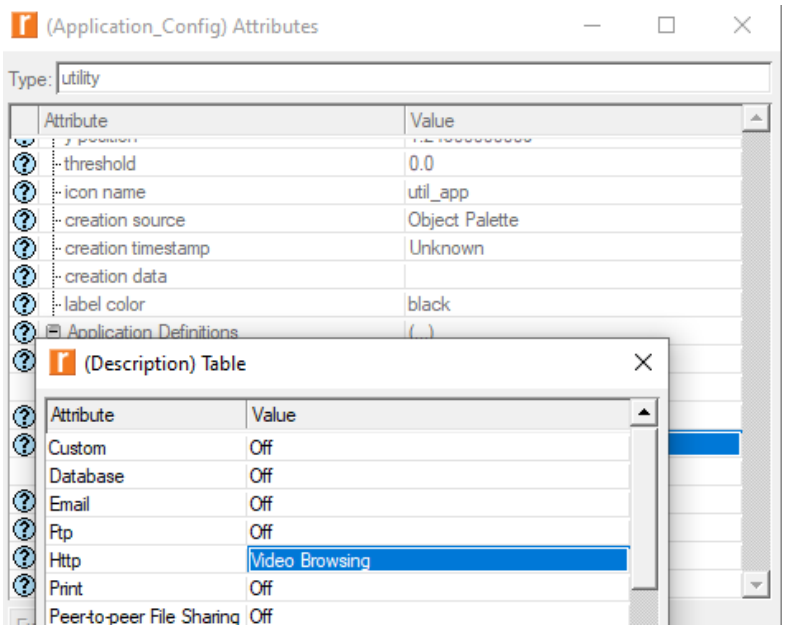
Poniżej przedstawiona zostanie topologia oraz konfiguracja scenariusza Server Load.



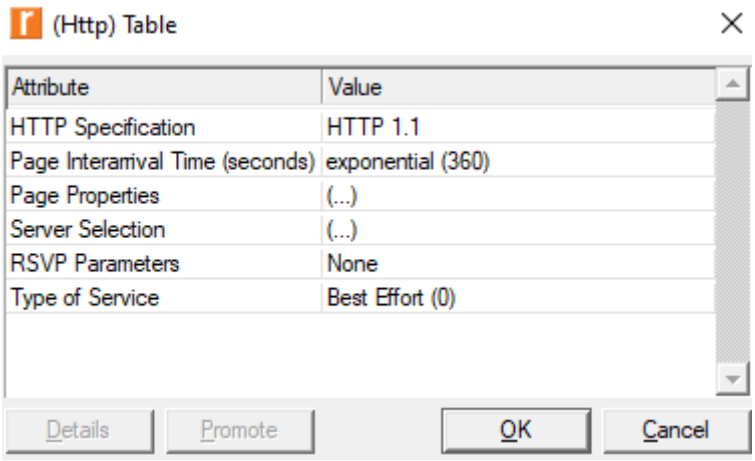
Rysunek 24 Scenariusz Server Load – topologia sieciowa



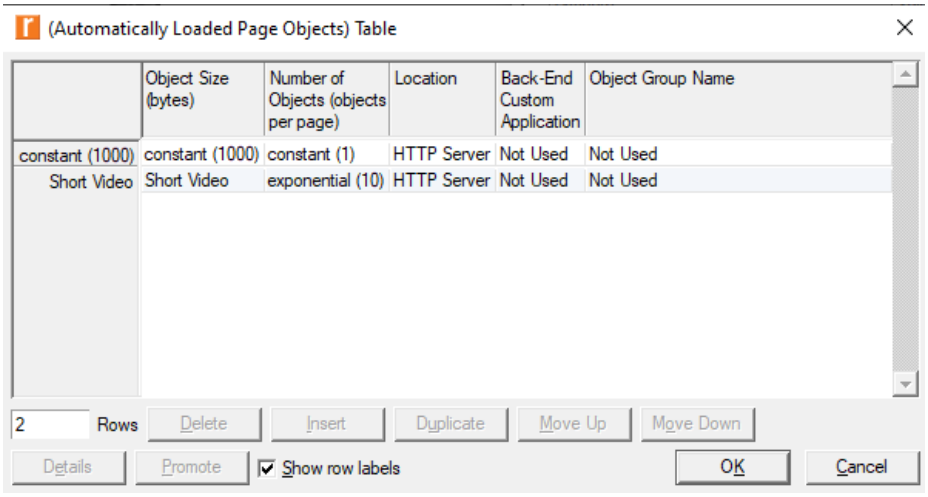
Rysunek 25 Ustawienie konfiguracji profilu



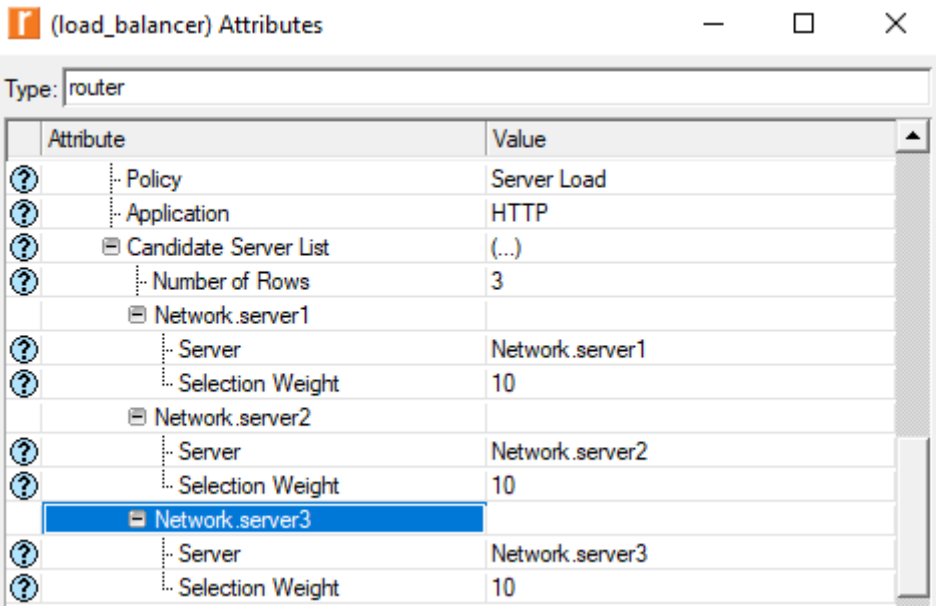
Rysunek 26 Konfiguracja aplikacji, wybór Video Browsing



Rysunek 27 Wybór odtwarzania filmów co 360 sekund

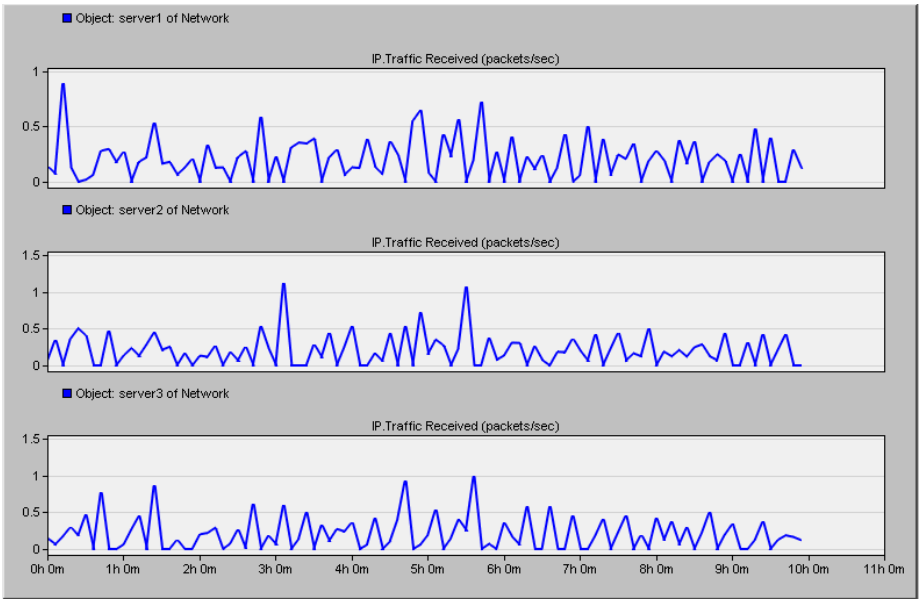


Rysunek 28 Wybór typu treści wideo



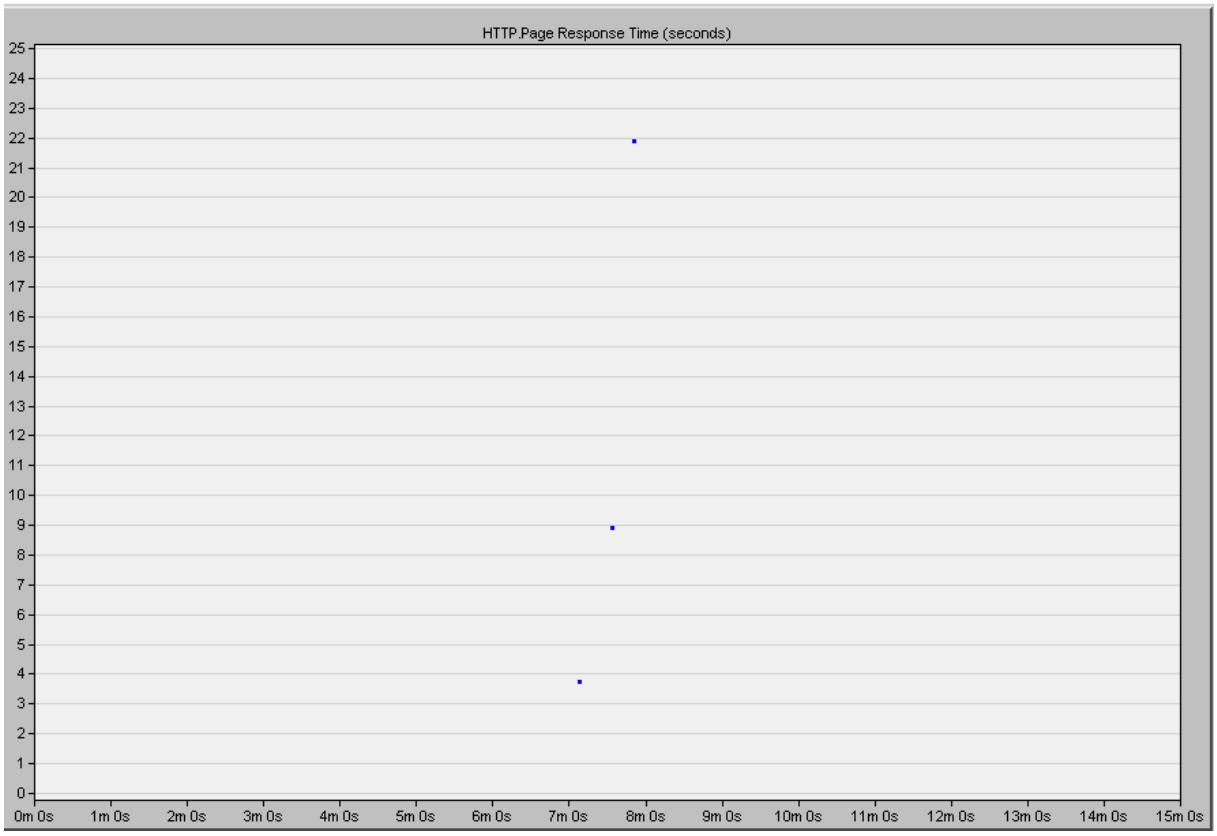
Rysunek 29 Ustawienia Load Balancera

Widzimy, że wagi poszczególnych serwerów są takie same, także na każdy serwer powinno być kierowane takie samo obciążenie. Potwierdza to poniższa symulacja ruchu otrzymanego.



Rysunek 30 Scenariusz Server Load – symulacja otrzymanych pakietów

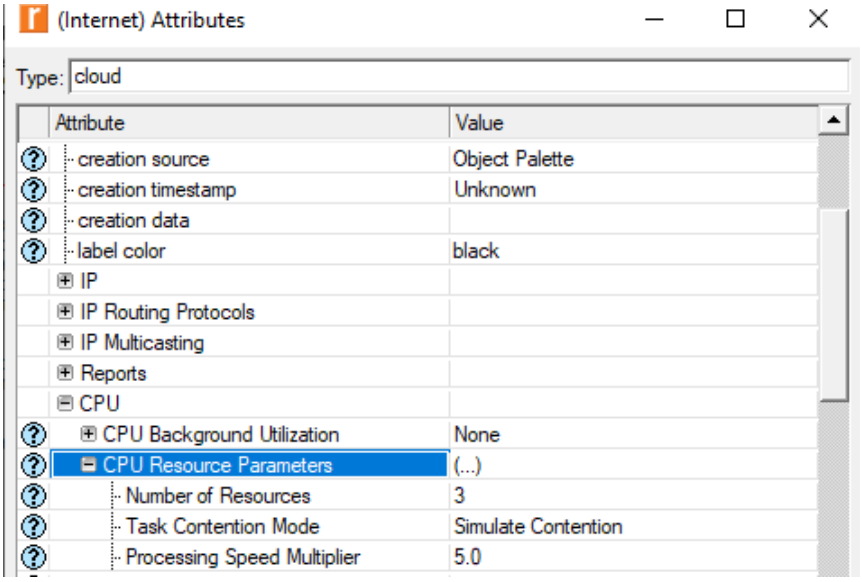
Bez żadnych modyfikacji scenariusza Server Load przeprowadzamy symulacje czasu odpowiedzi na załadowanie się stron z treścią wideo.



Rysunek 31 Scenariusz Server Load - Czas odpowiedzi na załadowanie się strony internetowej z treścią wideo

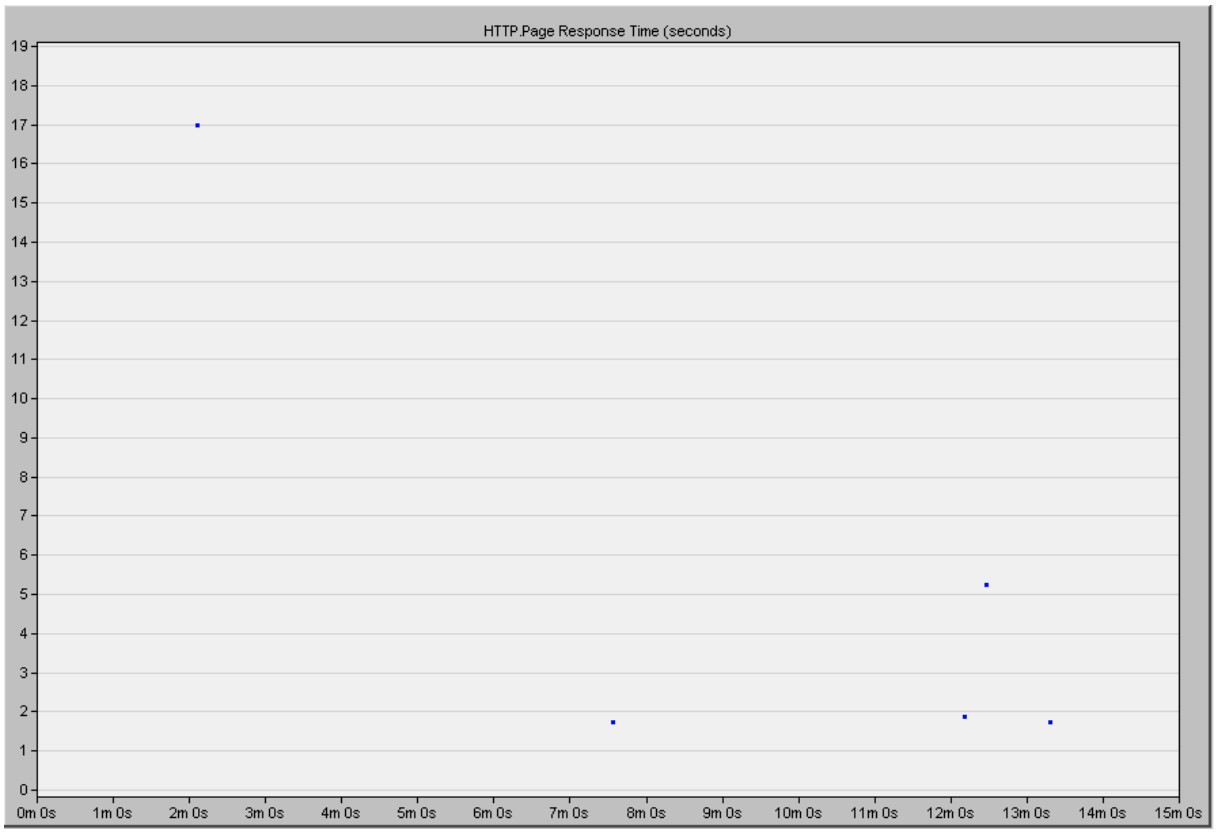
Bez zmieniania domyślnej konfiguracji sprzętowej i sieciowej omawianego scenariusza otrzymujemy maksymalnie czas odpowiedzi na załadowanie się strony wynoszący mniej niż 23 sekund, co jest podobnym czasem do scenariusza Random.

W celu redukcji tego czasu możemy zwiększyć wydajność urządzeń sieciowych.



Rysunek 32 Zwiększenie wydajności urządzeń sieciowych

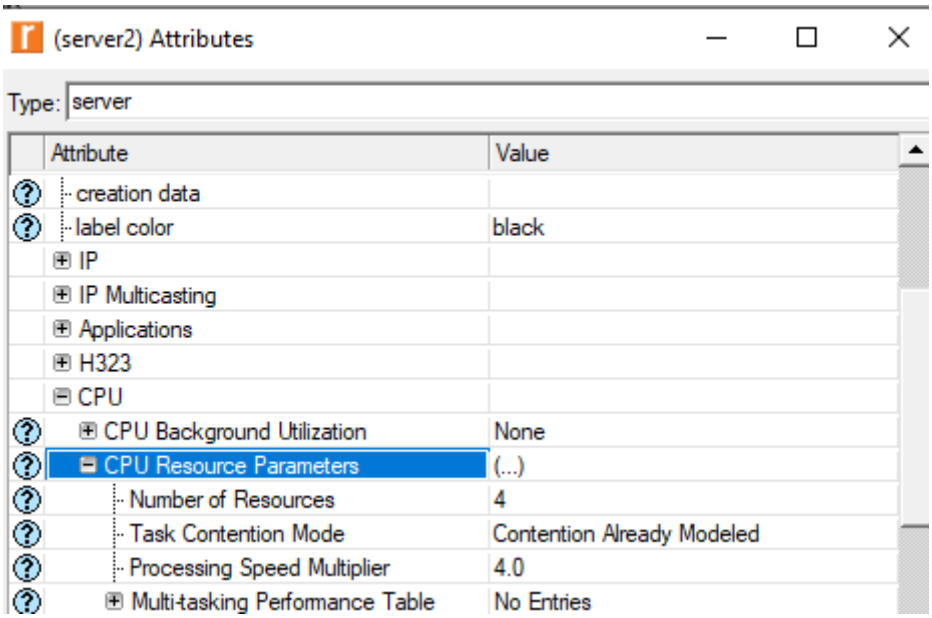
Po zwiększeniu wydajności urządzeń sieciowych przeprowadzamy ponowną symulację czasu oczekiwania na załadowanie się strony.



Rysunek 33 Czas odpowiedzi na załadowanie się strony

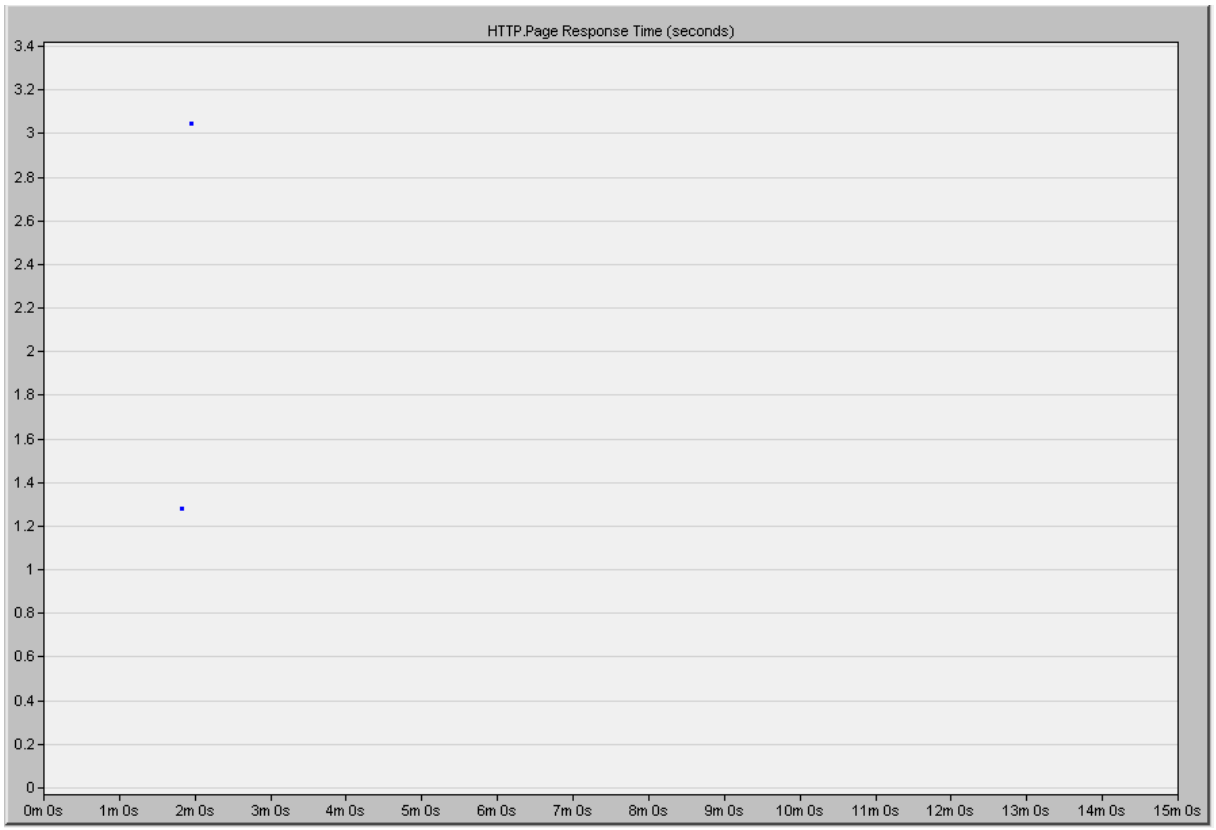
Po zwiększeniu wydajności urządzeń sieciowych uzyskujemy czas odpowiedzi, który maksymalnie sięga 17 sekund.

Aby dodatkowo zredukować czas wczytywania się stron możemy ulepszyć procesory serwerów, by mógłby realizować więcej operacji w tym samym czasie



Rysunek 34 Zwiększenie wydajności procesorów serwerów

Po ulepszeniu procesorów serwerów oraz zwiększeniu wydajności łącza ponawiamy symulacje.



Rysunek 35 Czas oczekiwania - po ulepszeniu łącza i procesorów

W scenariuszu Server Load po zwiększeniu wydajności procesorów naszych serwerów udało się uzyskać wystarczające wyniki oczekiwania na załadowanie się strony z filmem wideo. W

najgorszym przypadku jest to około 3 sekundy. Jednak konfiguracja aplikacji jest ustawiona na krótkie filmy wideo, w przypadku większego obciążenia potrzebowalibyśmy większej liczby serwerów. Widać również, że scenariusz Server Load jest optymalnym rozwiązaniem obciążania naszej sieci.

2 Ocena szybkości ładowania się stron – współczynnik Largest Contentful Paint

Aby ocenić poziom szybkości ładowania się stron możemy posłużyć się trzema progami oceny. Przy przeprowadzanych testach uwzględnia się szybkość Internetu i kraj pochodzenia użytkownika odwiedzającego badaną stronę:
Progi oceny są następujące:

Dobrze: 0-2,5 sekundy
Wymaga poprawy: 2,5 sekundy-4 sekundy
Słabo: 4 sekundy i więcej

Rysunek 36 Metryka Largest Contentful Paint

Metryka podaje czas renderowania największego obrazu, bloku tekstu, treści wideo widocznej na stronie.
Te dane pochodzą z raportu o korzystaniu z przeglądarki Google Chrome (CrUX). Raport CrUX jest aktualizowany co miesiąc i jest publicznie dostępny. Ten raport publiczny stanowi źródło aktualnych i historycznych danych o szybkości ładowania się stron w Internecie.

Ważne jest aby strony internetowe mieściły się w zakresie 0-2.5 sekundy. Należy testy przeprowadzić zarówno dla urządzeń desktopowych jak i mobilnych.

2.1 Metryka Apdex

Apdex (Application Performance Index) jest otwartym standardem pomiaru wydajności aplikacji programowych w obliczeniach. Jego celem jest przekształcenie pomiarów w spostrzeżenia na temat satysfakcji użytkowników, poprzez określenie jednolitego sposobu analizy i raportowania stopnia, w jakim mierzona wydajność spełnia oczekiwania użytkowników. Został on opracowany przez sojusz firm.

Metoda Apdex przekształca wiele pomiarów w jedną liczbę w jednolitej skali od 0 do 1 (0 = brak zadowolonych użytkowników, 1 = wszyscy użytkownicy zadowoleni). Uzyskany wynik Apdexu jest liczbowym miernikiem zadowolenia użytkowników z wydajności aplikacji korporacyjnych. Metryka ta może być wykorzystywana do raportowania dowolnych źródeł pomiarów wydajności użytkownika końcowego, dla których zdefiniowano cel wydajności.

Wzór Apdexu to liczba zadowolonych próbek plus połowa próbek tolerowanych plus żadna z próbek sfrustrowanych, podzielona przez wszystkie próbki:

$$\mathrm{Apdex}_t = \frac{\mathrm{SatisfiedCount} + (\mathrm{ToleratingCount} * 0.5) + (\mathrm{FrustratedCount} * 0)}{\mathrm{TotalSamples}}$$

.

gdzie podskrypt t jest czasem docelowym, a czas tolerowany przyjmuje się jako czterokrotność czasu docelowego. Łatwo jest więc zauważyć, że stosunek ten jest zawsze bezpośrednio związany z postrzeganiem przez użytkowników zadowalającej reakcji aplikacji.

Przykład: zakładając cel wydajnościowy 3 sekundy lub lepszy i tolerowany standard 12 sekund lub lepszy, biorąc pod uwagę zbiór danych ze 100 próbkami, gdzie 60 jest poniżej 3 sekund, 30 jest pomiędzy 3 a 12 sekundami, a pozostałe 10 jest powyżej 12 sekund, wynik Apdex jest:

$$\mathrm{Apdex}_3 = \frac{60 + (30 * 0.5) + (10 * 0)}{100} = 0.75$$

.

Wzór Apdex jest równoważny średniej ważonej, gdzie zadowolony użytkownik otrzymuje wynik 1, użytkownik tolerowany otrzymuje wynik 0,5, a użytkownik sfrustrowany otrzymuje wynik 0.

Dodatkowo poniżej została obliczona metryka Apdex dla danych uzyskanych w analizowanych scenariuszach.

Tabela 1 Metryka Apdex dla analizowanych scenariuszy

Scenariusz	Aptex
No Load Balancing	0
Random	0.5
Server Load	1

3 Wnioski i podsumowanie

W celu łatwiejszej wizualizacji uzyskanych wyników została sporządzona tabela, zawierająca czasy ładowania się stron dla poszczególnych scenariuszy.

Tabela 2 Uzyskane czasy ładowania się stron dla analizowanych scenariuszy

Scenariusz	Średni czas ładowania się strony z treścią wideo [s]
No Load Balancing	14
Random	5
Server Load	3

Analizując czasy uzyskane w tabeli 2 widzimy, że czas ładowania najkrótszy jest dla scenariusza Server Load – wynika to głownie z równomiernego rozłożenia obciążenia wskutek czego serwery nie są przeciążone i mogą pracować z optymalną prędkością. Scenariusz Random pozwolił nam na uzyskanie średniego czasu 5 sekund na załadowanie się strony, dla bardziej cierpliwych użytkowników i mniej wymagający również jest to akceptowalny czas. Scenariusz No Load Balancing pozwolił na uzyskanie czasu ładowania się stron około 14 sekund. Jest to za długi czas i dlatego warto używać technik równoważenia obciążenia, by uzyskiwać krótsze czasy oczekiwania na załadowania się stron wideo.

Odnosząc się do współczynika LCP - oceny szybkości ładowania się stron wideo widzimy, że w scenariusz No Load Balancing udało nam się uzyskać najmniejszy czas 14 sekund co nie jest akceptowane w metryce LCP i zostałby przydzielony próg strony słabej. Jeżeli chodzi o scenariusz Random tam czas ładowania się stron z treścią wynosił około 5 sekund – to również jest słaby wynik, lecz zbliżony do progu wymagającego poprawy działania strony.

Scenariusz Server Load pozwolił na osiągniecie czasu ładowania się strony wynoszącej około 3 sekundy co pozwalało by na wpisanie takiej strony do progu wymagającego poprawy.
Analizując wszystkie te scenariusze widać, że żadnemu nie udało się uzyskać średniego czasu, poniżej 2,5 sekundy co pozwoliło by znaleźć się najlepszym progu.
Jednak zależy zauważyć, ze badania są przeprowadzone dla wszystkich stron, a strony z treści wideo zawierają zdecydowanie więcej zasobów niż zwykłe treści z teksem.
Gdy w początkowej symulacji analizowaliśmy stronę z obrazkami w scenariuszu No Load Balancing udało nam się uzyskać czasy około 0,40 sekundy, co spełniało by próg strony dobrej.

Kolejna metryka Apdex pozwoliła nam określić poziom zadowolenia użytkowników, gdy mieli by oni korzystać z analizowanych scenariuszy. Odnosząc się do tabeli 1 widzimy, że gdy użytkownicy by korzystali z usług zdefiniowanych to byli by oni nie zadowoleni. W scenariuszu Random połowa użytkowników była by usatysfakcjonowana z poziomu świadczenia usług, a w scenariuszu Server Load, każdy użytkownik byłby zadowolony.

Należy zwrócić uwagę, że każda metryka zwraca uwagę na inne aspekty oraz należy podkreślić fakt, że odczucia każdej osoby są subiektywne oraz należy odróżnić przeciętnego użytkownika, od osoby, która potrzebuje niezawodności i szybkości działania usług, np. w celu pracy.

4 Bibliografia

1) https://web.dev/lcp/#what-is-lcp
2) https://www.thinkwithgoogle.com/intl/en-cee/feature/testmysite/faq/?section=2#section-2
3) https://en.wikipedia.org/wiki/Apdex