Polish Network Operators' Group

IP Anycast

Ochrona i skalowanie usług sieciowych

Łukasz Bromirski
lbromirski@cisco.com

Agenda

- IP Anycast – co to jest?
- Jak wykorzystać IP Anycast?
- Rozważania projektowe dla DNS
- Rozważania projektowe dla HTTP
- Q&A

IP Anycast

Trochę o historii…

To nic nowego, ale nadal działa dobrze ☺

- Wykrywanie nowych usług – SNTPv4 (RFC 2030) oraz NTP "manycast" (RFC 4330)
- Pierwszy mechanizm przechodzenia na IPv6 (RFC 2893) a potem 6to4 (RFC 3068)
- Użycie anycastów dla RP – MSDP i PIM (RFC 4610)
- W IPv6 pierwsze RFC (RFC 1884, 2373 i 3513) zabraniały używania adresu anycastowego jako adresu źródłowego; zniesiono to ograniczenie w RFC 4291
- Zastosowanie anycastów w systemie DNS (RFC 3258) wspomina wykorzystanie "shared unicast"

IP Anycast

- Prosty, efektywny, szeroko używany ☺
- Niezależny od warstwy trzeciej

  zarówno IPv4 jak i IPv6 działają – po prostu

  RFC mówi o specjalnych adresach anycast – ale większość technik skalowania, w tym te najpowszechniejsze – wykorzystują podejście 'shared unicast'

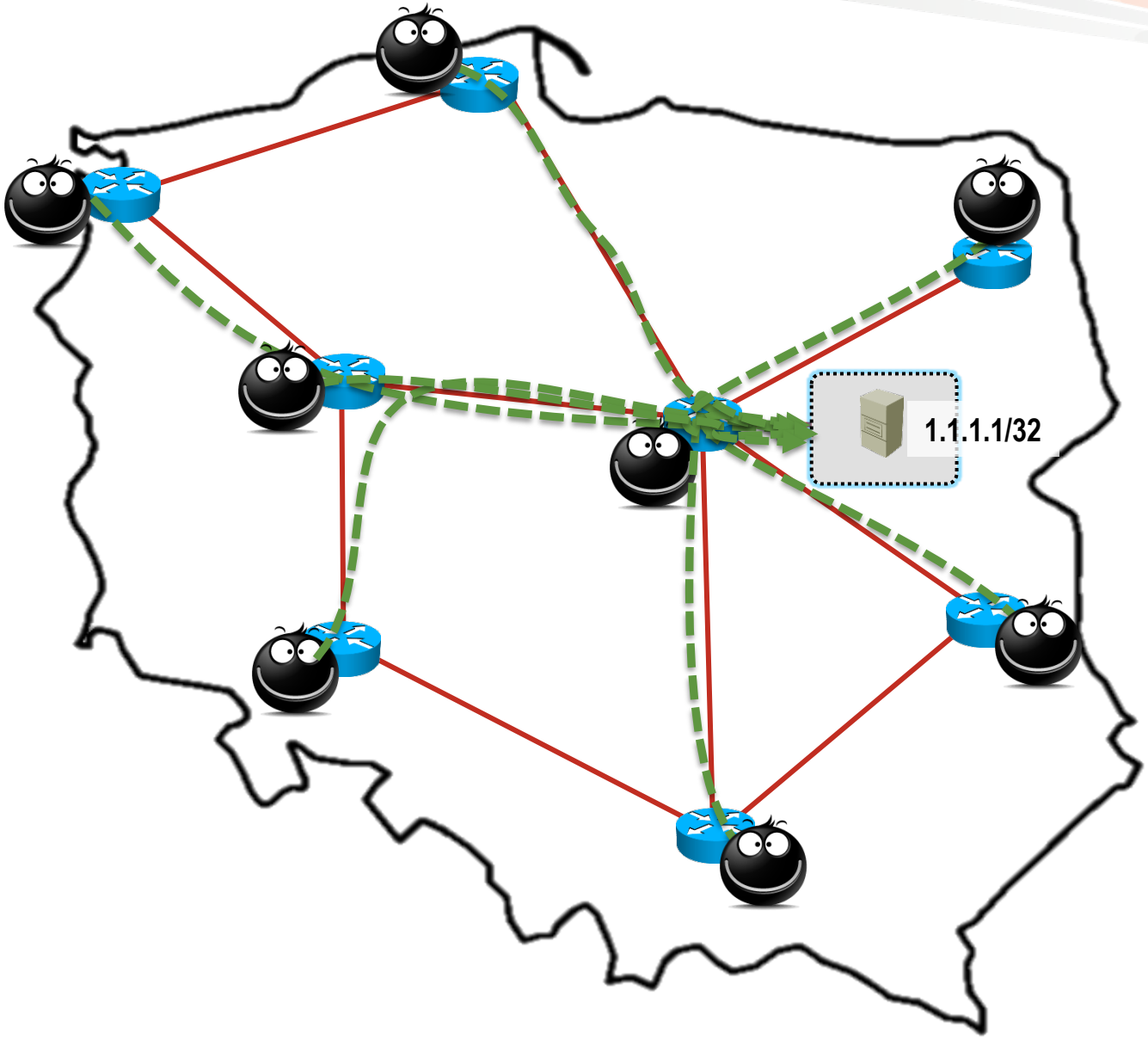
IP Anycast

- Jest bardzo uniwersalny
- Może być wykorzystywany zamiast lub równocześnie z:
  - równoważeniem obciążenia w L3/L4
  - dodawaniem usług IP w sposób tradycyjny
- W przeciwieństwie do powszechnego przekonania, IP Anycast może być również używany do skalowania usług TCP*

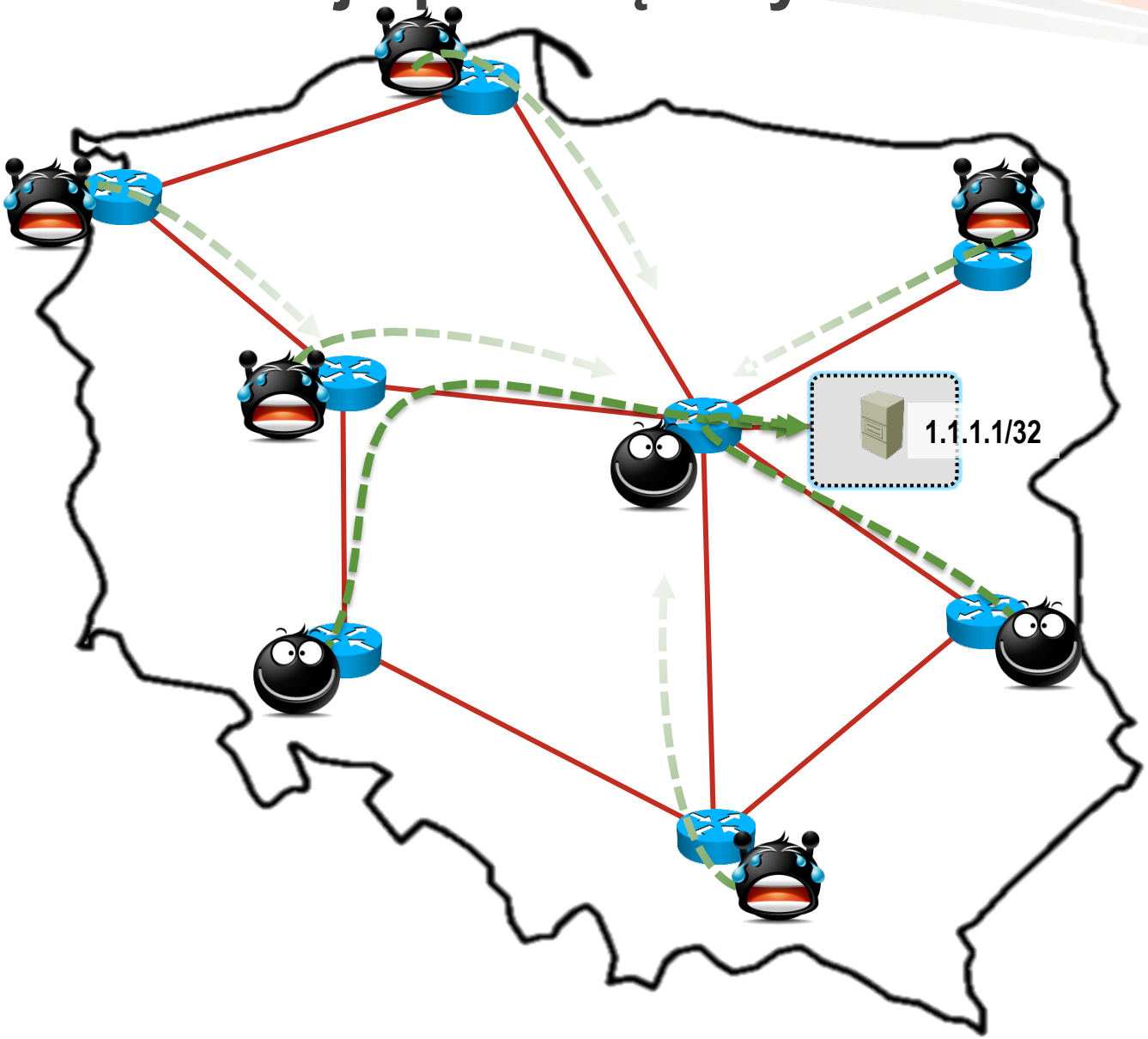
* więcej o tym później, pamiętajcie jednak że Wasze rezultaty mogą się różnić

Wykorzystanie anycastu IP

Typowe zastosowanie

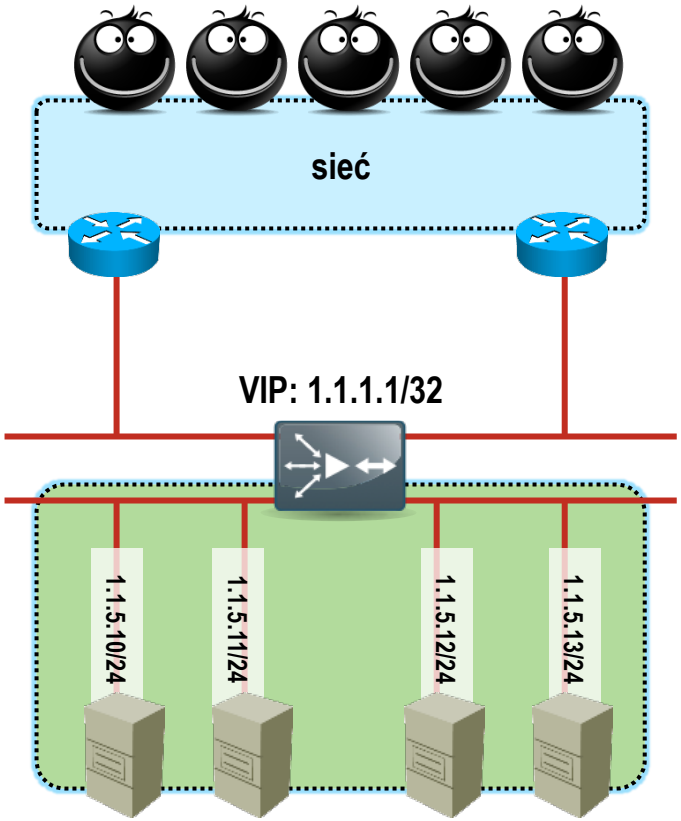


Serwer zostaje przeciążony

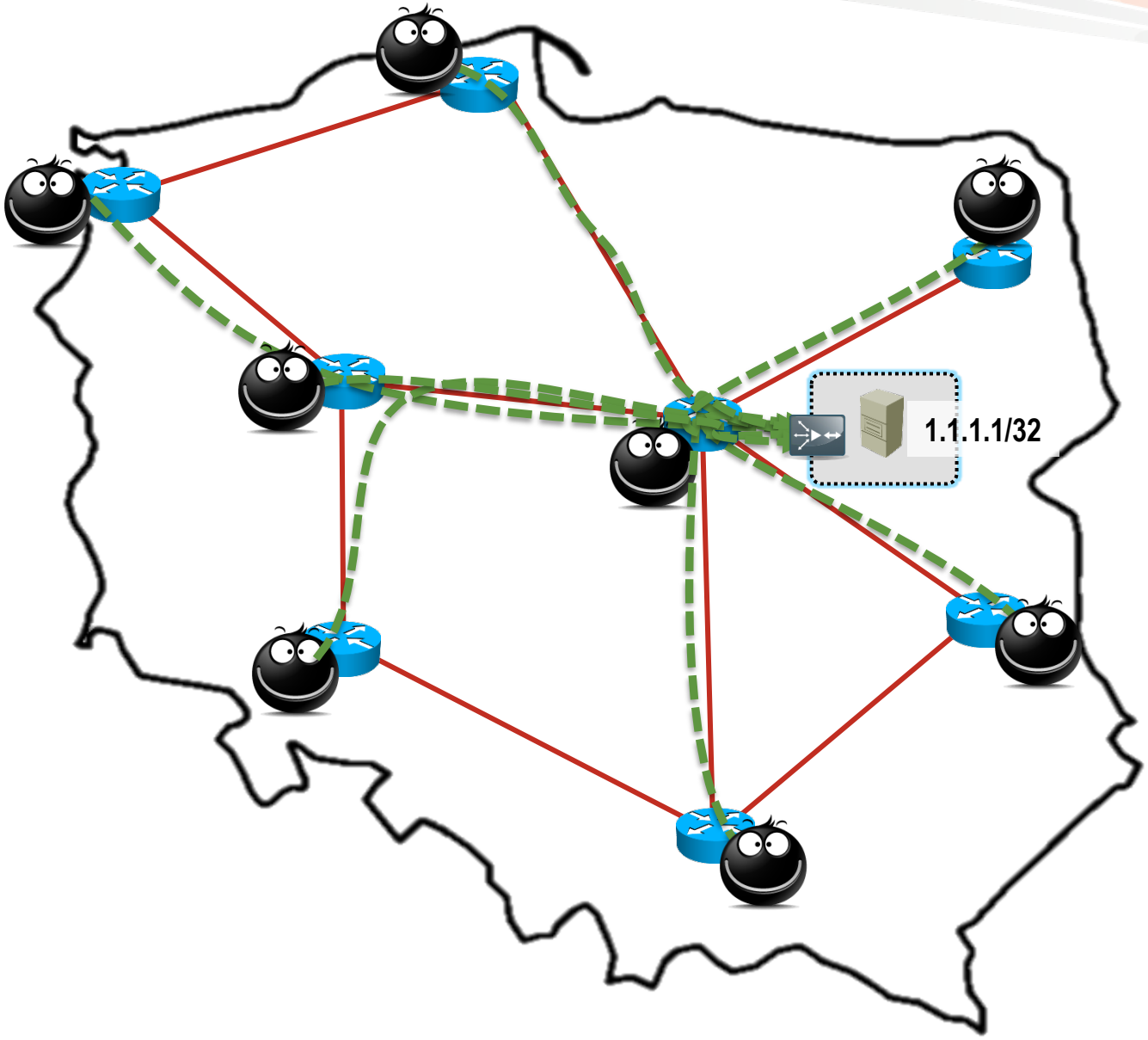


Rozwiązanie pierwsze: rośniemy wszerz

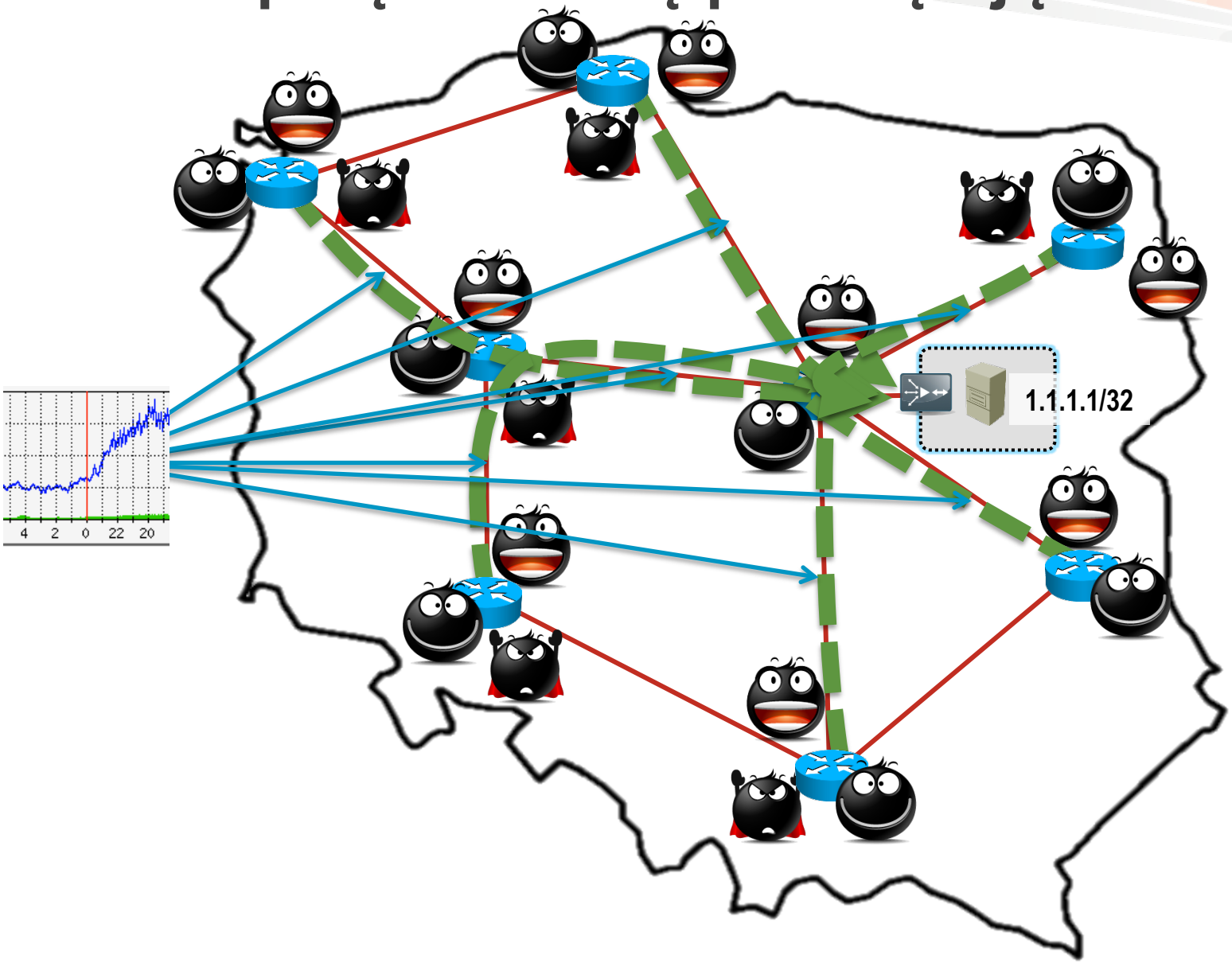
"Zainstalujmy loadbalancer i dodajmy serwerów"



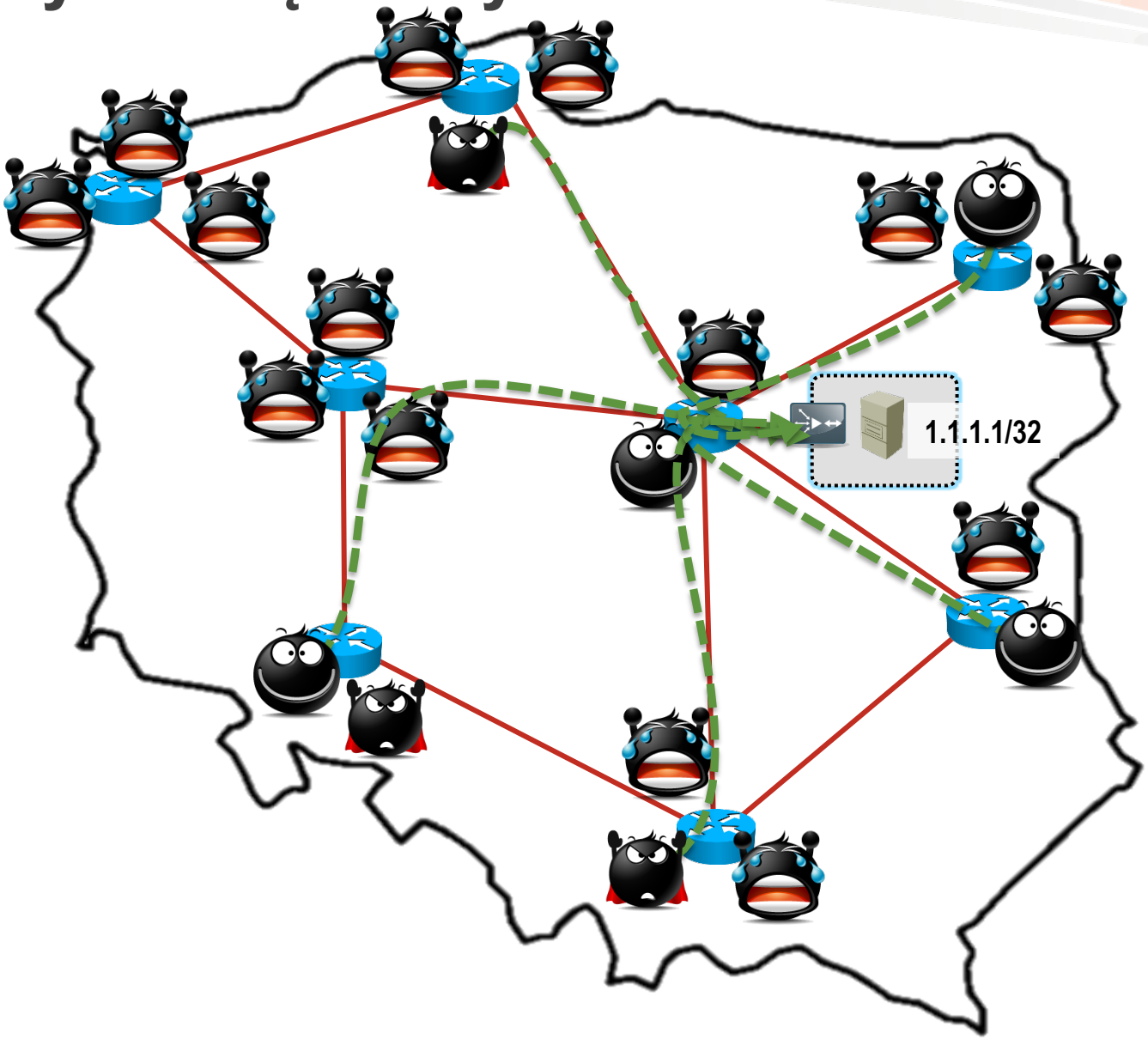
Po zastosowaniu VIPa i serwerów



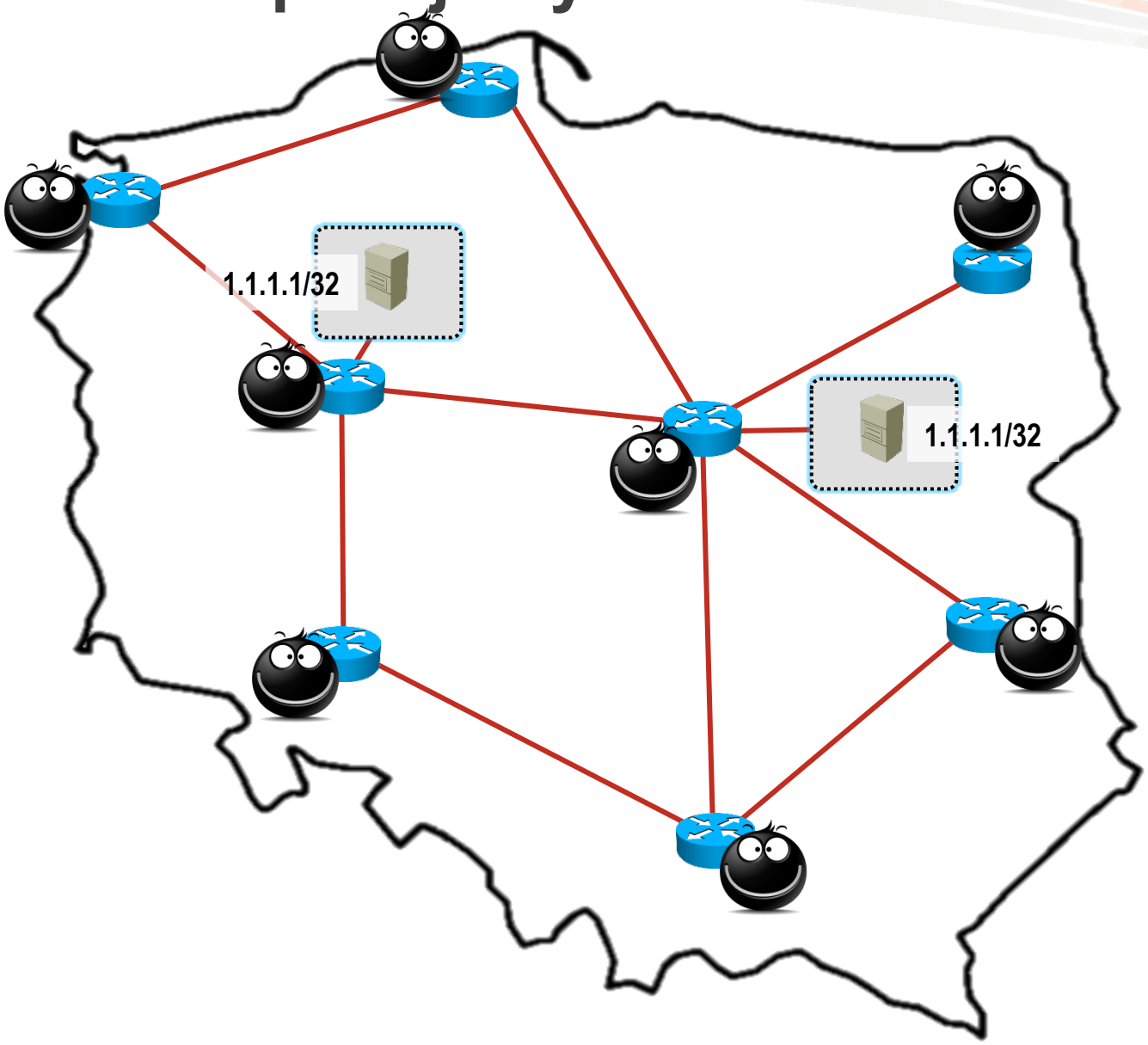
…ale połączenia się przeciążają…



…i tylko część użytkowników ma dostęp



A może zreplikujemy to samo IP?



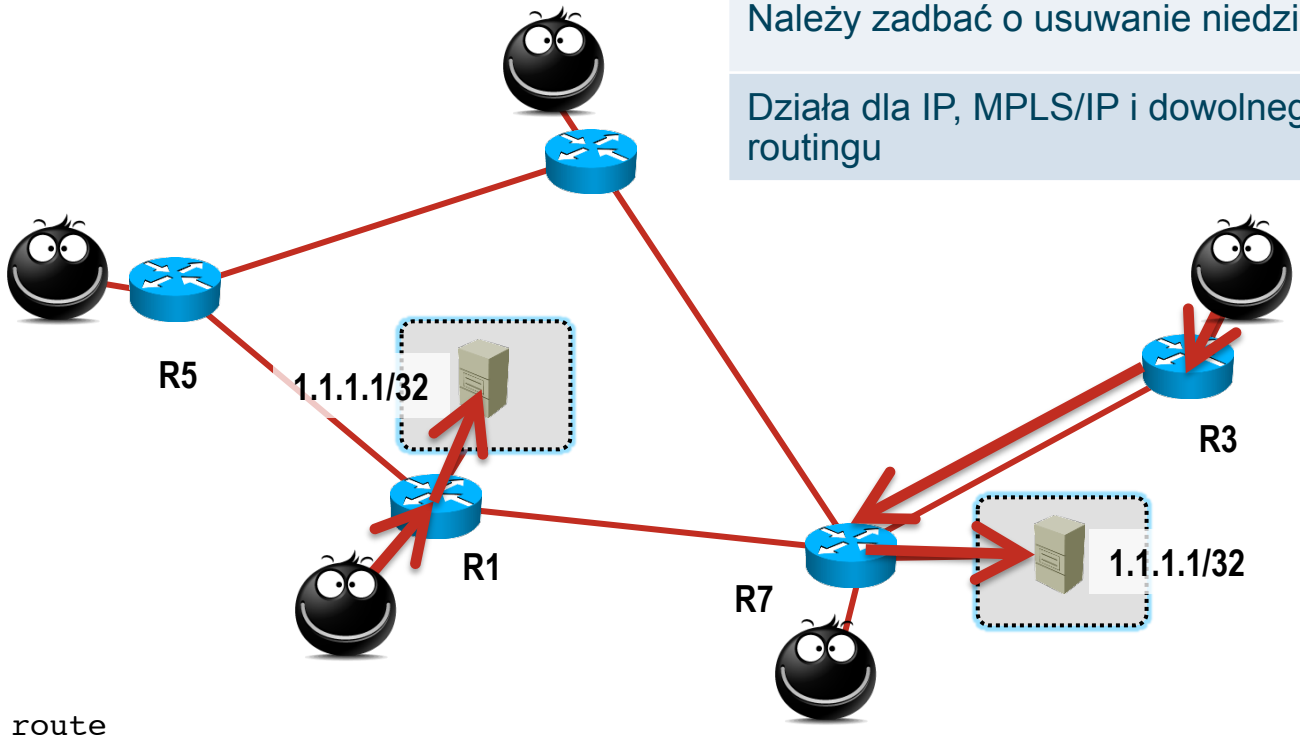
Zreplikowana usługa

Mechanizm routingu wybierze drogę o najmniejszym koszcie pomiędzy klientem a serwerem

Każdy z hostów zostanie skierowany do "najbliższej" instancji usługi

Należy zadbać o usuwanie niedziałających instancji

Działa dla IP, MPLS/IP i dowolnego protokołu routingu




```
R1> show ip route
I    1.1.1.1/32 [115/10] via 192.168.11.6
```

```
R3> show ip route
I    1.1.1.1/32 [115/20] via 192.168.73.7
```

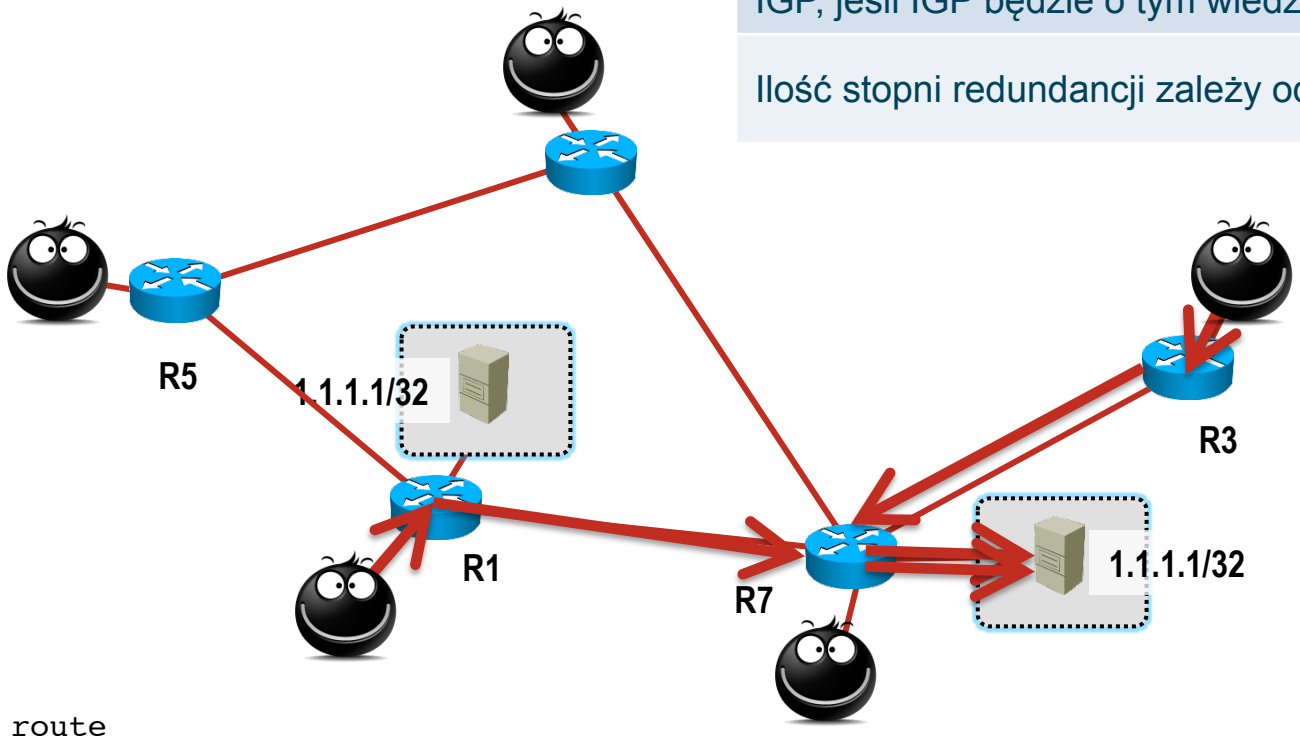
```
R7> show ip route
I    1.1.1.1/32 [115/10] via 192.168.77.6
```

Zreplikowana usługa

Replikacja zapewni wysoką dostępność

"Śmierć" jednej instancji spowoduje zniknięcie jej z IGP, jeśli IGP będzie o tym wiedział

Ilość stopni redundancji zależy od ilości instancji




```
R1> show ip route
I    1.1.1.1/32 [115/20] via 192.168.17.7

R7> show ip route
I    1.1.1.1/32 [115/10] via 192.168.77.6
```

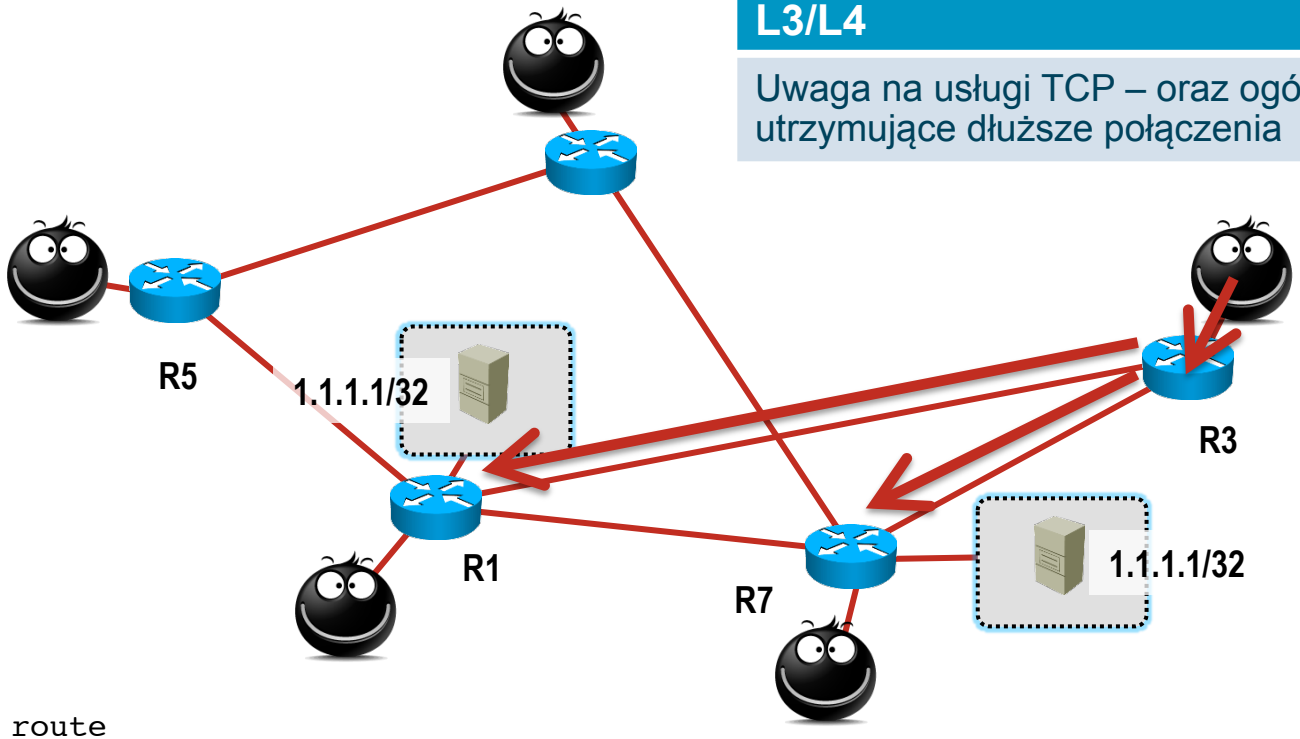
```
R3> show ip route
I    1.1.1.1/32 [115/20] via 192.168.73.7

R7> show ip route
I    1.1.1.1/32 [115/10] via 192.168.77.6
```

Zreplikowana usługa

Automatyczne rozkładanie obciążenia w L3/L4

Uwaga na usługi TCP – oraz ogólnie, na usługi utrzymujące dłuższe połączenia




```
R3> show ip route
I    1.1.1.1/32 [115/20] via 192.168.13.1
                         via 192.168.37.7
```

IP Anycast w służbie DNS

Problemy tradycyjne

- Brak stosowania "best practices"
  - Serwery DNS usytuowane są zwykle w jednym miejscu (z punktu widzenia topologii sieci) co oznacza, że stają się łatwym celem dla ataków i zwiększa ryzyko ich zniknięcia w przypadku awarii
    - wyeliminowanie serwerów i łącz do nich nigdy nie było prostsze
  - Nawet jeśli serwery DNS znajdują się w różnych miejscach sieci, przyciągają różny ruch – do różnych IP
    - atak na 2-3 adresy lub podsieci jest nadal bardzo prosty do przeprowadzenia
    - przełączenie się z pierwszego adresu na drugi przy rozwiązywaniu nazw nadal nie wychodzi dobrze większości popularnych systemów operacyjnych

Problemy tradycyjne

- Skalowalność:
  - większość klientów DNS (resolverów) w stosowanych powszechnie systemach operacyjnych nie radzi sobie dobrze z więcej niż 1-2 serwerami DNS
    - czasami można dodać więcej, ale rzadko robi to realną różnicę*
    - możesz serwować różne adresy DNS w różnych częściach swojej sieci, ale jest to trudne do zarządzania i utrzymania
  - można też dodać lokalny load-balancing na wiele różnych adresów IP – sprzętowo lub programowo
    - to jednak też dodaje złożoności do całości rozwiązania

Kto używa IP anycastu dla DNS?

- "Prawie" wszyscy! ☺

OpenDNS®

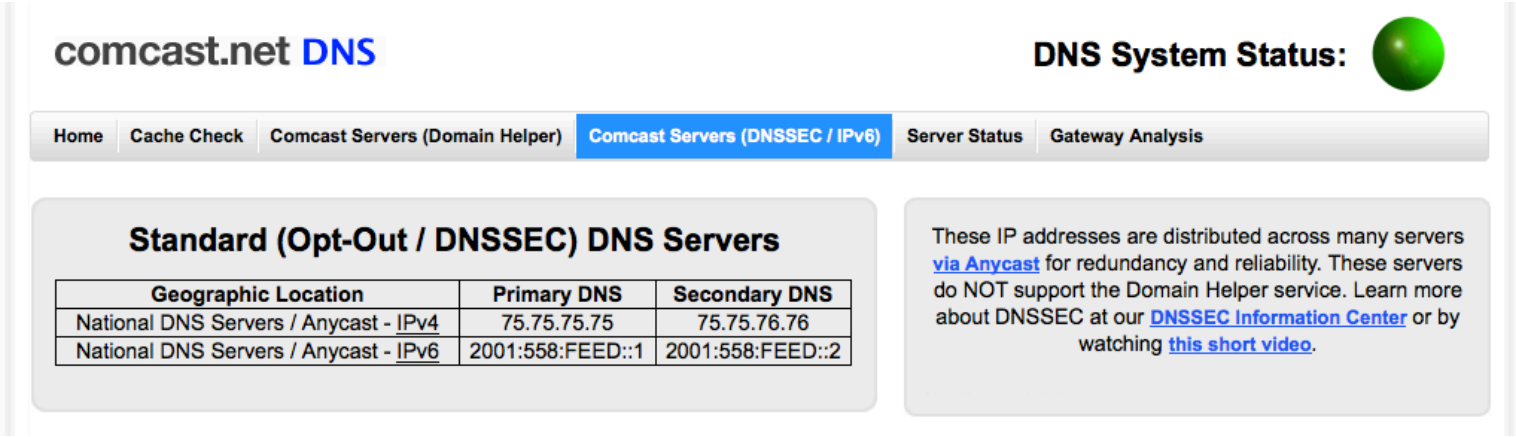
Google Public DNS IP addresses

The Google Public DNS IP addresses (IPv4) are as follows:

- 8.8.8.8
- 8.8.4.4

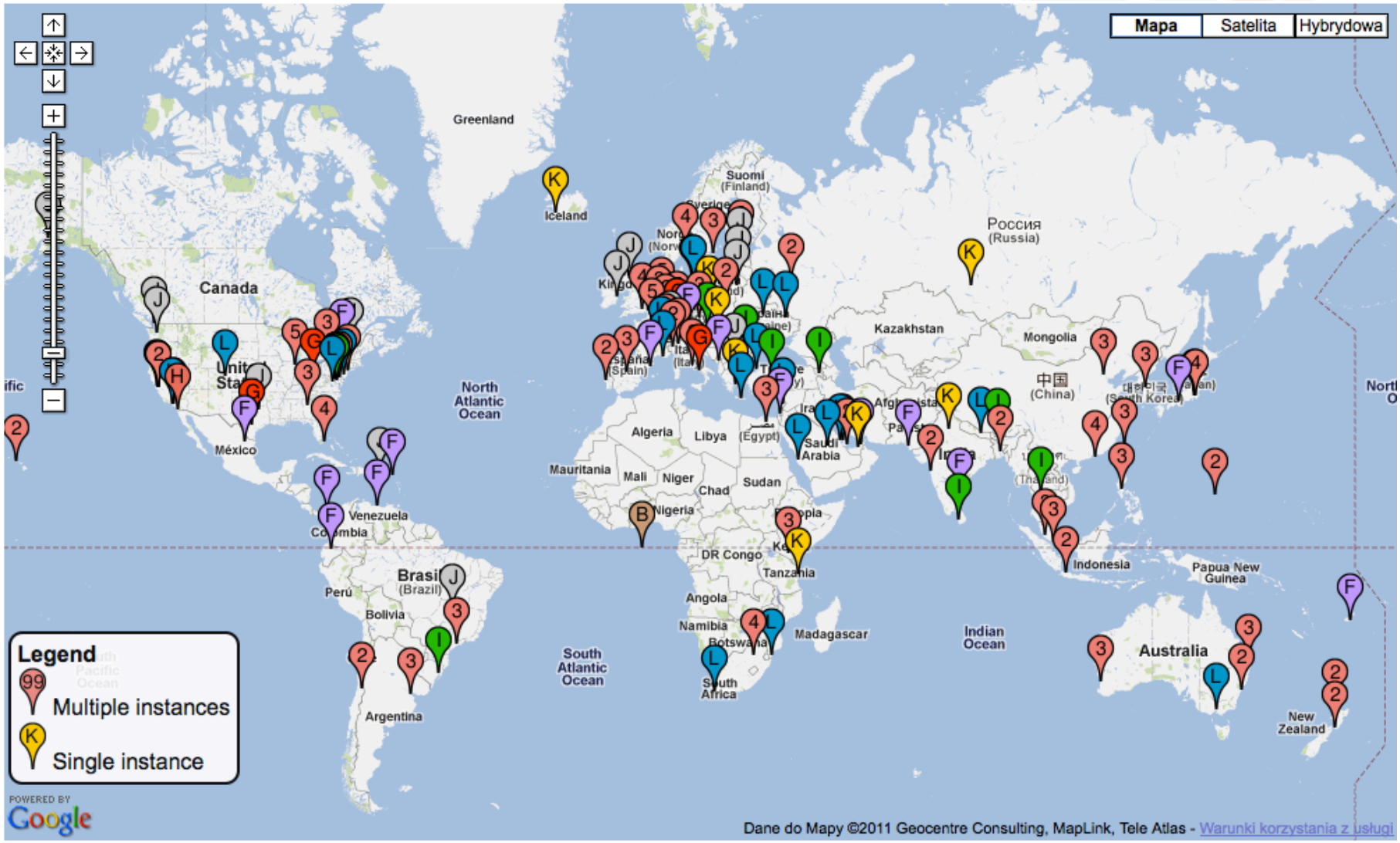
The Google Public DNS IPv6 addresses are as follows:

- 2001:4860:4860::8888
- 2001:4860:4860::8844



* Mapa CAIDA dla DNS: http://www.caida.org/research/dns/influence-map/

Mapa root serwerów DNS



http://www.root-servers.org/

Ochrona serwerów root DNS anycastem

ICANN

Factsheet

Root server attack on 6 February 2007

On 6 February 2007, starting at 12:00 PM UTC (4:00 AM PST), for approximately two-and-a-half hours, the system that underpins the Internet came under attack. Three-and-a-half hours after the attack stopped, a second attack, this time lasting five hours, began.

At least six root servers were attacked but only two of them were noticeably affected: the "g-root", which is run by the U.S. Department of Defense and is physically based in Ohio, and the "l-root" run the Internet Corporation for Assigned Names and Numbers (ICANN), which is physically based in California.

The reason why these two were particularly badly affected was because they are the only root servers attacked that have yet to install Anycast (a further three root servers without Anycast were not attacked this time).

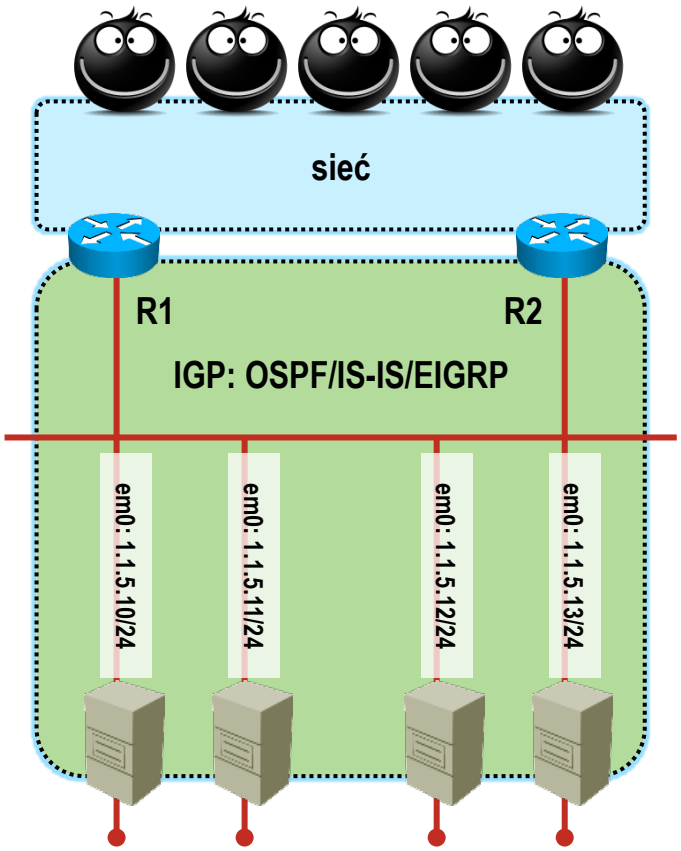
Rozwiązanie IP anycastu dla DNS

W detalach

Tablica routingu R1

```
R1# sh ip route

*> 1.1.1.1/32
     via 1.1.5.10
     via 1.1.5.11
     via 1.1.5.12
     via 1.1.5.13
```



Tablica routingu R2

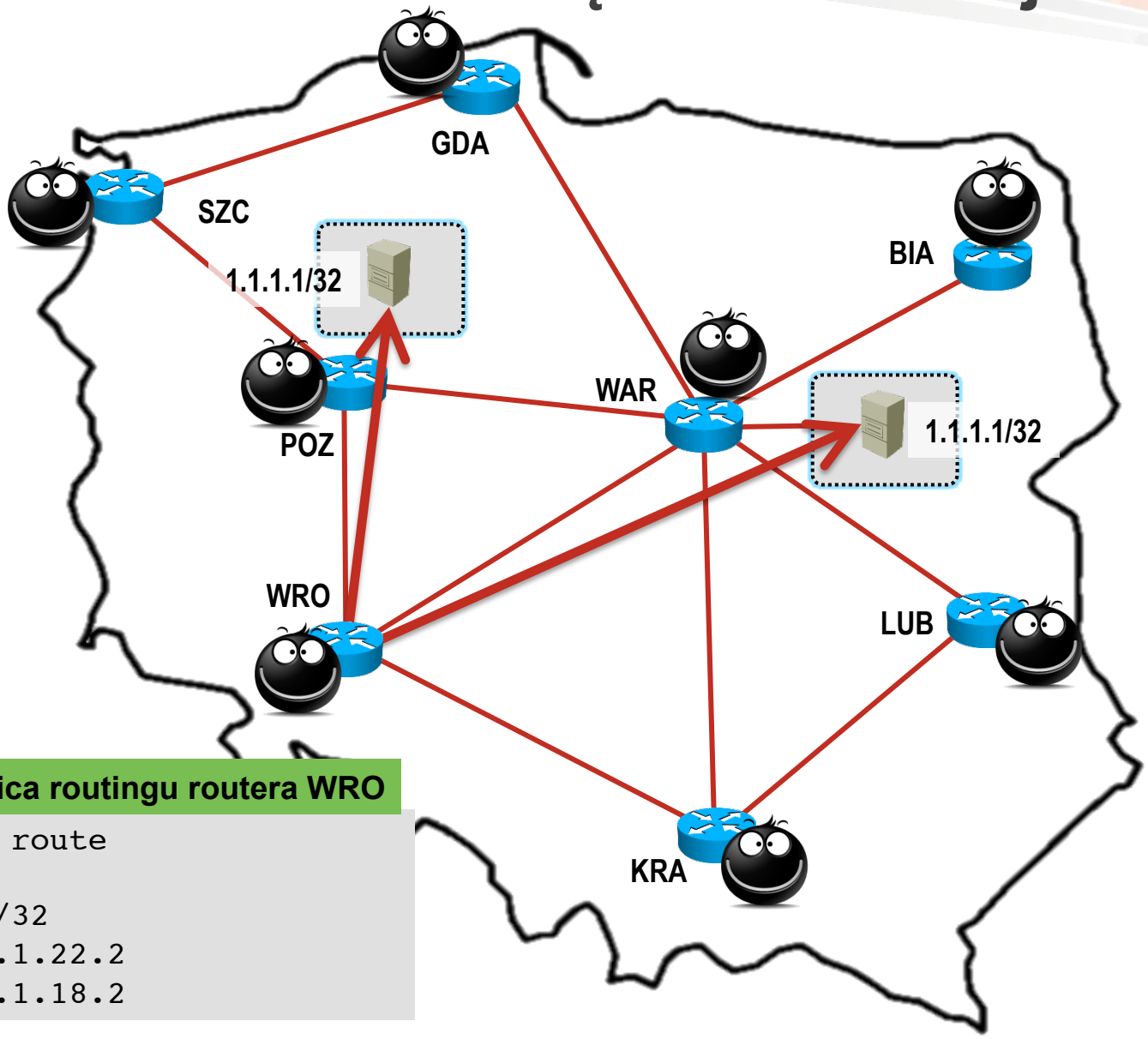
```
R2# sh ip route

*> 1.1.1.1/32
     via 1.1.5.10
     via 1.1.5.11
     via 1.1.5.12
     via 1.1.5.13
```

zapytanie o adres serwera DNS

```
;; ANSWER SECTION:
ns1.anycasted.com.      86400   IN A    1.1.1.1
```

Równoważenie obciążenia w dużej sieci



Tablica routingu routera WRO

```
WRO# sh ip route

*> 1.1.1.1/32
     via 1.1.22.2
     via 1.1.18.2
```

Rozwiązanie IP anycastu dla DNS

Konfiguracja systemu DNS - named

- Wymagana jest uważna konfiguracja

  odpowiadamy na pytania z adresu anycast, przypisanego do interfejsu loopback0 (1.1.1.1/32 & 2001:db8::100:100)

  sami pytamy, przeprowadzamy transfery stref/etc z adresu fizycznego interfejsu lub innego interfejsu loopback

NS1-1 named.conf

```
options {
  listen-on { 1.1.1.1; };
  listen-on-v6 { 2001:db8::100:100; };
  query-source address 1.1.5.10;
  query-source-v6 2001:db8::34:254;
  transfer-source 1.1.5.10;
  transfer-source-v6 2001:db8::34:254;
};
```

NS1-2 named.conf

```
options {
  listen-on { 1.1.1.1; };
  listen-on-v6 { 2001:db8::100:100; };
  query-source address 1.1.5.11;
  query-source-v6 2001:db8::34:254;
  transfer-source 1.1.5.11;
  transfer-source-v6 2001:db8::34:254;
};
```

Rozwiązanie IP anycastu dla DNS

Konfiguracja – quagga/OpenOSPFd

- Pomiędzy serwerem DNS a routerami należy uruchomić protokół IGP – pakiet routingu do wyboru

quagga.conf

```
interface em0
  ip address 1.1.5.11/24
!
interface lo0
  ip address 1.1.1.1/32
```

ospfd.conf

```
router ospf
  ospf router-id 1.1.5.11
  network 1.1.5.0/24 area 10
  redistribute connected route-map PF-DNS
!
route-map PF-DNS permit 10
  match ip address prefix-list dns-ospf
!
ip prefix-list dns-ospf permit 1.1.1.1/32
```

IP Anycast w służbie HTTP

Skalowanie usług TCP w ogólności

- Ilość zmian topologii wpływających na "bliskość" instancji powinna być ograniczona
  - zmiana topologii oznacza w tej sytuacji zerwanie sesji TCP
  - obejście problemu jest związane zwykle z
    - synchronizacją stanów pomiędzy serwerami (bolesne, "ciężkie" - złożone)
    - nauczenie aplikacji klienta lub urządzenia po drodze (CPE) by dawało sobie z tym radę (działa w rzeczywistych zastosowaniach*)
- Jeśli pomiędzy klientem a instancjami mamy wiele równoległych ścieżek, mogą pojawić się problemy
  - można w prosty** sposób zmienić koszty połączeń tak aby uniknąć problemu jeśli występuje

* Studium przypadku dla streamingu TCP dla usług IP anycast, NANOG #37

Usługa HTTP z IP anycast

- Dla długo trwających sesji HTTP, w szczególności z wykorzystaniem różnego rodzaju mechanizmów śledzących, jak np. cookies – przy zmianie topologii może pojawić się problem
  - Technologie typu AJAX są lepsze do tego typu zastosowań
- Jeśli topologia zmieni się pomiędzy kolejnymi zapytaniami klienta, mogą pojawić się pewne problemy
  - np. niezsynchronizowane/uaktualnione cookies, powodujące wyrzucenie/złe przekierowanie klienta
  - można to oczywiście obejść wykonując synchronizację z niezależnych adresów

Gdzie znaleźć więcej informacji?

- Wykorzystanie anycastu z TCP

  Anycast-aware transport for content delivery networks

  http://dl.acm.org/citation.cfm?id=1526750

- Wdrażanie anycastów z TCP

  NANOG #37

  http://www.nanog.org/meetings/nanog37/abstracts.php

- BCP 126 / RFC 4786

  Operation of Anycast Services

  http://tools.ietf.org/html/rfc4786

- Rozważania architekturalne dla IP Anycast

  http://tools.ietf.org/html/draft-mcpherson-anycast-arch-implications-00

Q&A

Łukasz Bromirski
lbromirski@cisco.com