

“High Scalability: Building bigger, faster,
more reliable websites*”

PLNOG, 15-16/01/2009
Warszawa, Marcin Mazurek

Allegro Group

* <http://highscalability.com/>



Agenda

Łańcuch pokarmowy.

QXL Poland Sp. z o.o.

Serwisy Aukcyjne



Serwisy Ogłoszeniowe



Płatności



Sklepy Internetowe



Struktura Działu Technicznego Grupy Allegro

- Zespół serwisów ogłoszeniowych

- OtoMoto
- OtoDom
- AlleWakacje

- Zespół systemów płatności

- Platnosci.pl
- Payu.pl
- Reks.pl
- Labfoto.pl

- Projekty

- PayBACK
- PayGSM
- iStore

- Dział Aplikacji Allegro

- Interfejs
- Międzynarodowy
- Projekty

- Dział Infrastruktury**

- HelpDesk**
- NOC**
- DBA**
- Sieć**
- System**

•Serwis
• wydajność

•Management
•Warehouse



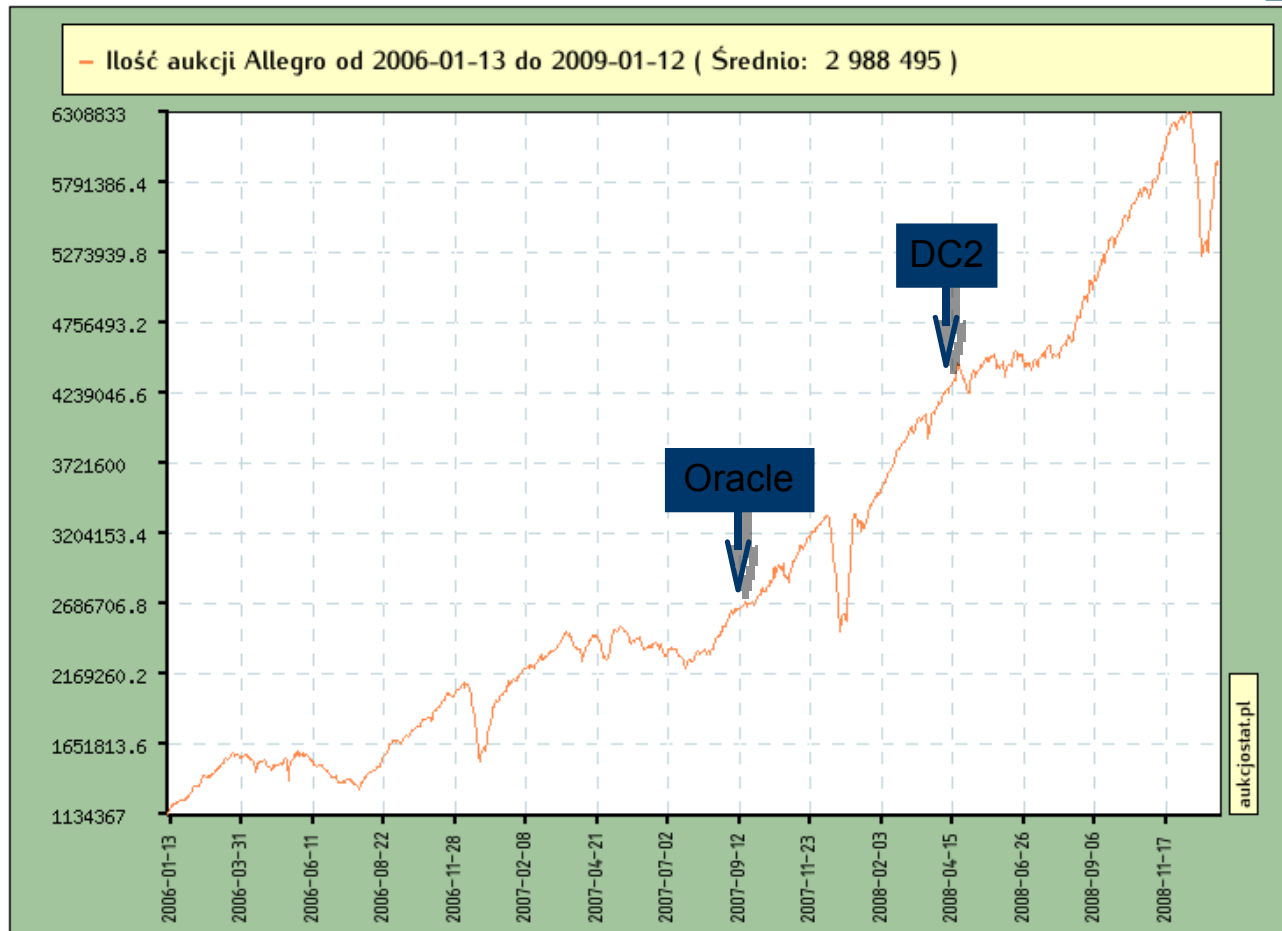
O Grupie Allegro od strony technicznej...

- Ok. 500 serwerów, w większości w technologii blade.
 - W 99% customizowany CentOS.
 - apache, lighttpd, squid, varnish
 - PHP, Ansi C, C++, Java
 - Oracle, MySQL, PostgreSQL
 - Cisco, IronPort, F5, Juniper
 - HP, IBM, SUN
 - IBM, 3PAR, OnStore
-
- Wysyłamy ok. 4 mln maili dziennie z powiadomieniami.
 - Ponad 200 mln obrazków (z dwóch miesięcy).
 - show_item – 30% - 120 tys/min.
 - Ruch HTTP
 - http requests: ok. 100 tys/sek
 - nowe http requests: ok. 20 000/sek
 - serwer http requests: ok. 2000/sek



Ilość wystawionych przedmiotów, ostatnie 2 lata.

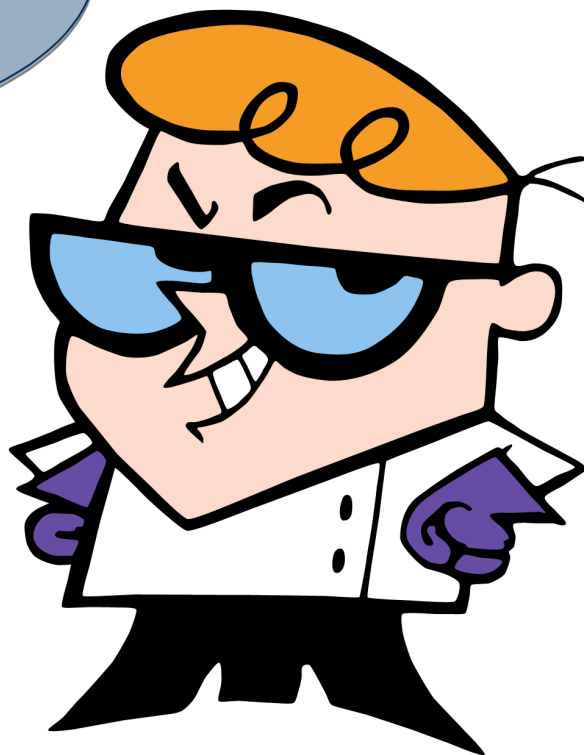
DC3



Źródło: <http://aukcjostat.pl/>

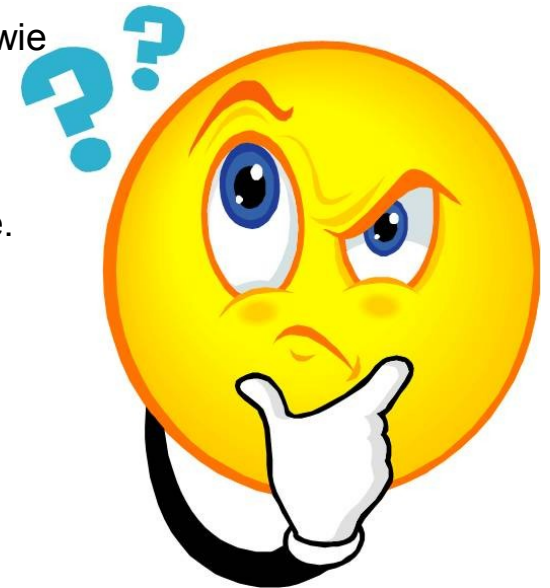
BĘDĘ WIELKI !!!

"All your base are belong to us !!!"

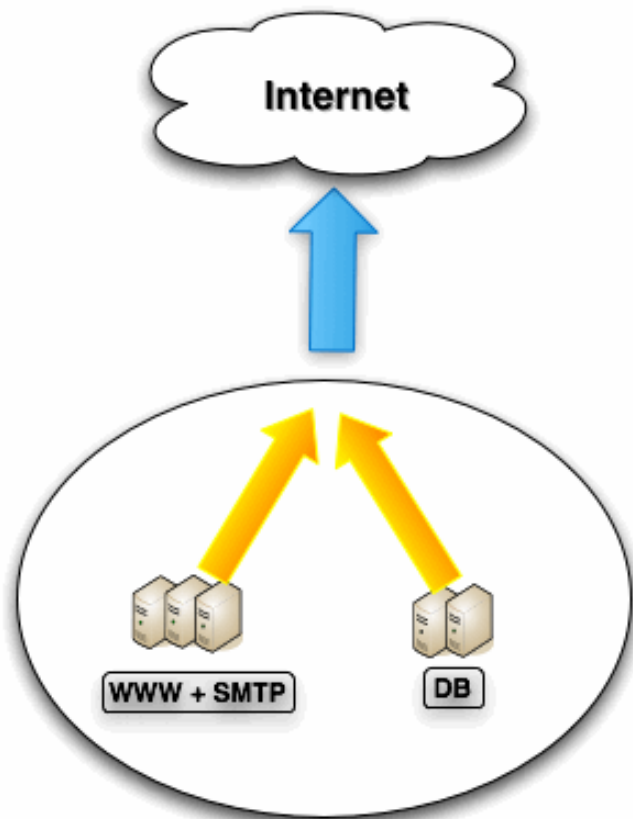


Problemy, z którymi musimy się zmierzyć.

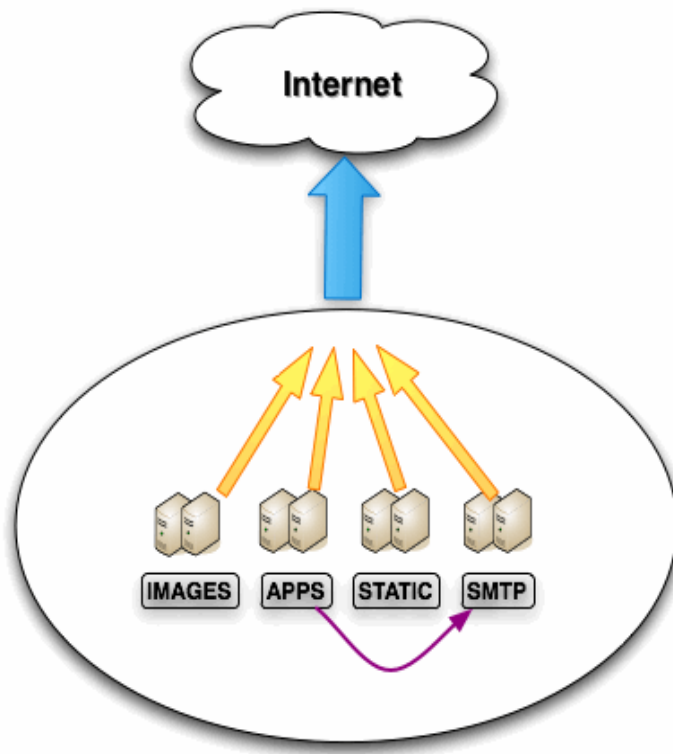
2. Skalowalność aplikacji (najlepiej liniowa).
3. Możliwość wyodrębnienia logicznych/funkcjonalnych części aplikacji (SOA).
4. Duży wolumen ruchu HTTP, który trzeba odpowiednio ukierunkować.
5. Dostarczenie jak najlepszej usługi przy użyciu możliwie małej ilości zasobów i kosztów.
6. Zarządzanie infrastrukturą.
7. Szybkie reagowanie na potrzeby i zmiany w serwisie.
8. Eliminowanie pojedynczych punktów awarii
9. Efektywny i użyteczny monitoring.
10. Dokumentacja i procedury.



Życie serwisu w kilku krokach cz. 1



Życie serwisu w kilku krokach cz. 2



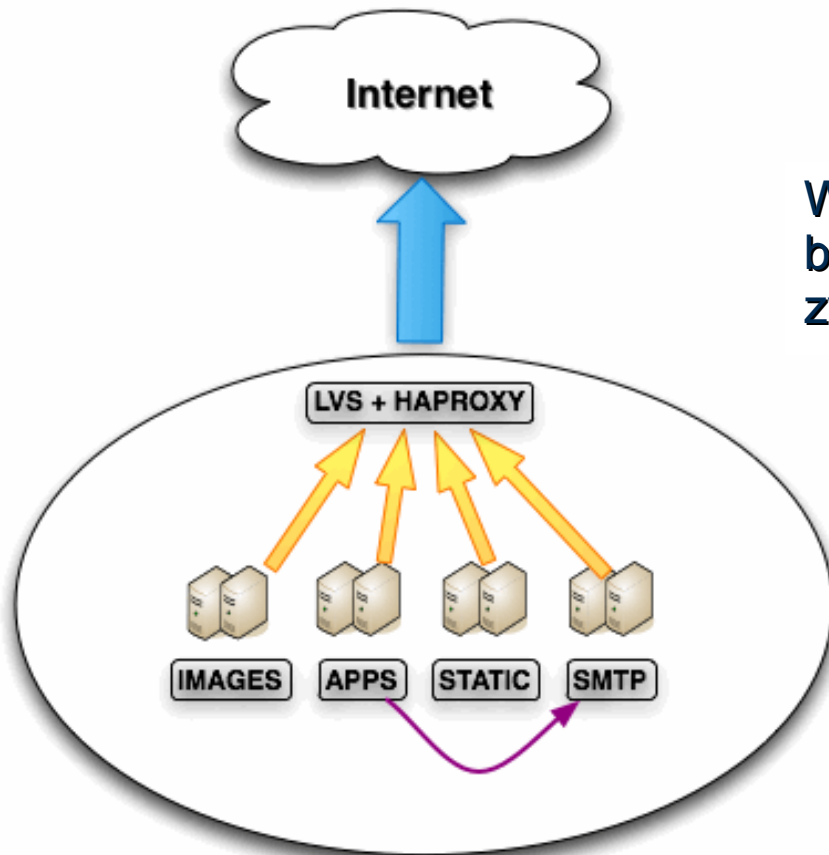
<http://www.bede-wielki.pl>

<http://static.bede-wielki.pl>

<http://images.bede-wielki.pl>

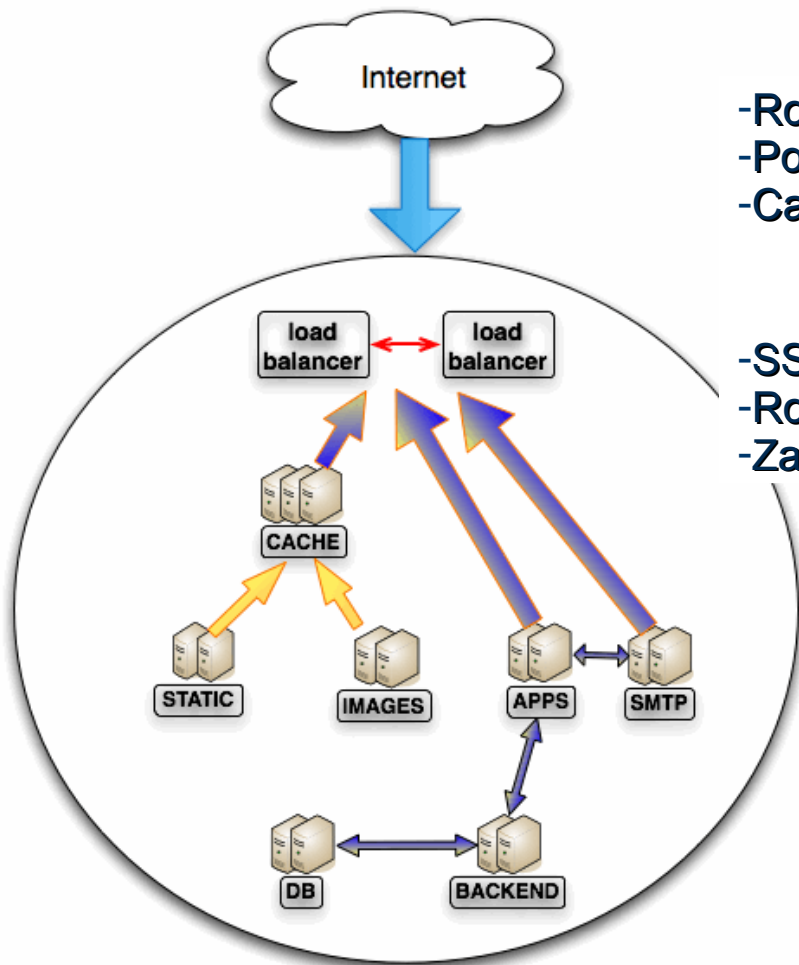
smtp.bede-wielki.pl

Życie serwisu w kilku krokach cz. 3



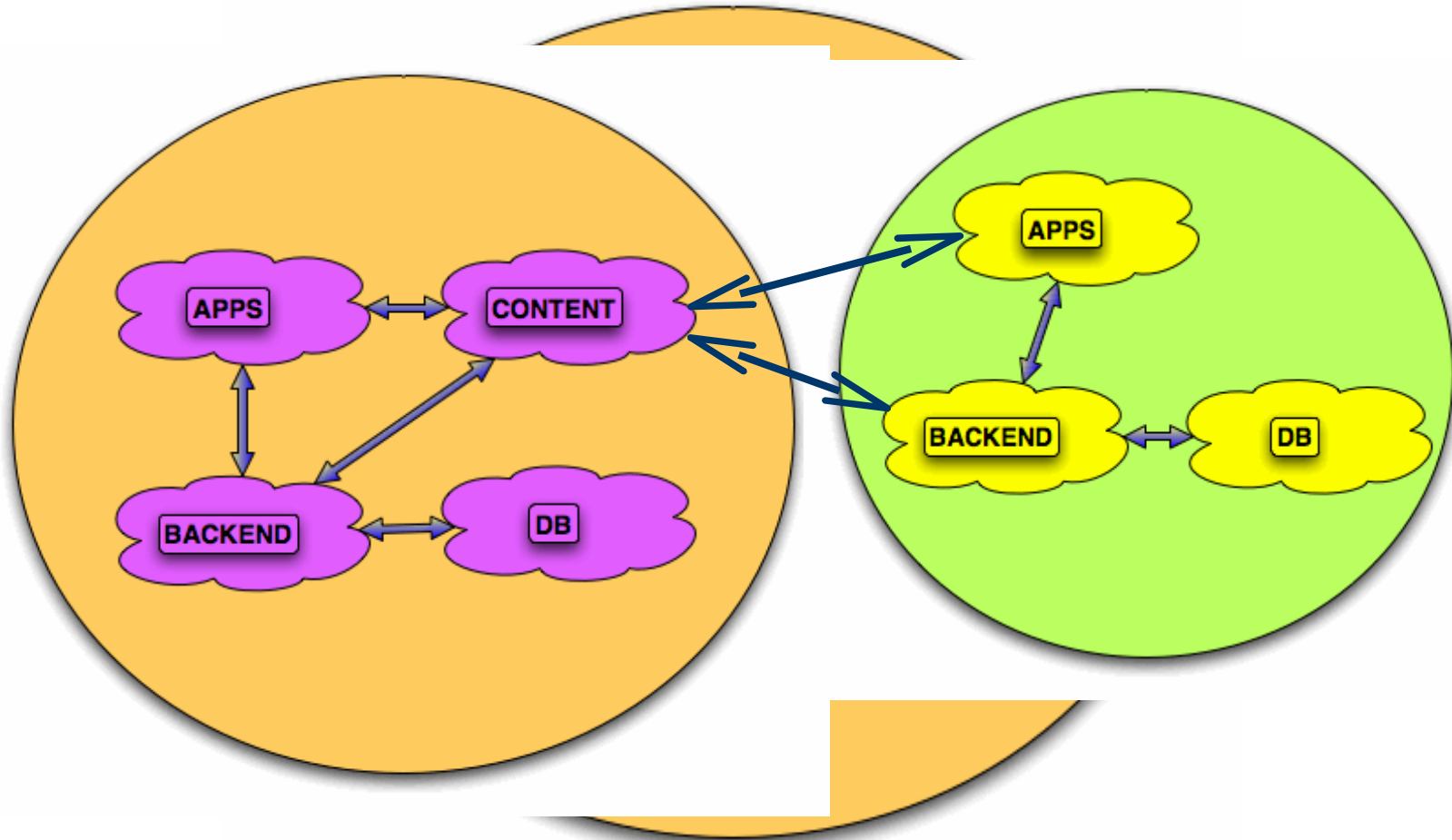
Wykorzystanie software load balancerów pozwalających na zwiększenie ich redundancji.

Życie serwisu w kilku krokach cz. 4

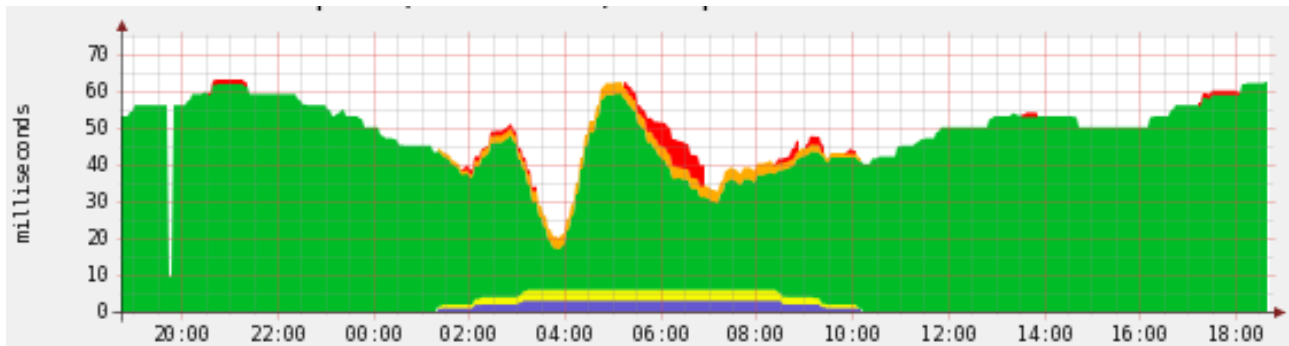


- Rozbicie kontentu na niezależne farmy
- Podział funkcjonalny aplikacji
- Cachowanie
 - show_item 31% [średnio pewnie 2 zdj na stronie]
 - showcat + search 23% [po 50 zdjęć na stronie]
- SSL offload
- Rozdzielenie backend od aplikacji
- Zaawansowany load balancing

Service Oriented Architecture



Cache everything !!!

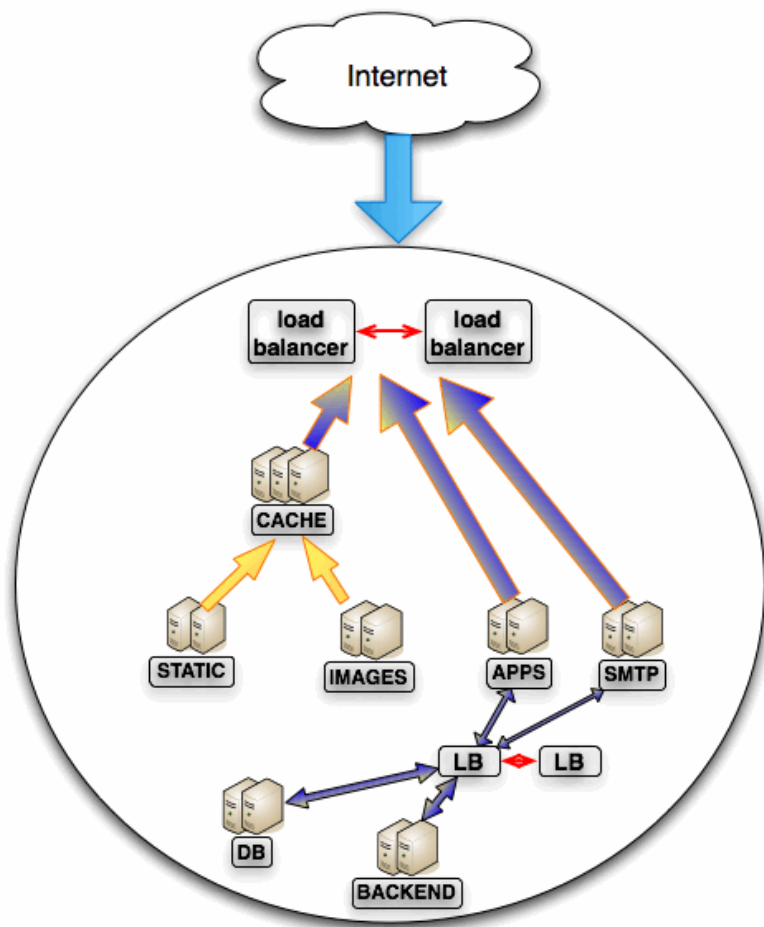


| | | | | | |
|-------------------------|-------|----|------|----|------|
| Overall Service Time: | Curr: | 0 | Avg: | 1 | Max: |
| Hit Service Time: | Curr: | 0 | Avg: | 1 | Max: |
| Miss Service Time: | Curr: | 62 | Avg: | 46 | Max: |
| Near Miss Service Time: | Curr: | 0 | Avg: | 1 | Max: |
| ICP Query Service Time: | Curr: | 0 | Avg: | 0 | Max: |
| DNS Service Time: | Curr: | 0 | Avg: | 1 | Max: |

Varnishtop:

| | |
|------------|--------------------------------|
| 9.59 TxURL | /public/magazin_cn/export.html |
| 6.79 TxURL | /public/click/export.html |
| 1.59 TxURL | /public/hry/export1.html |
| 1.37 TxURL | /public/t-mobile/export3.html |
| 0.75 TxURL | /public/online/a2_240.html |
| 0.69 TxURL | /public/mailru/export_b.html |
| 0.64 TxURL | /public/mailru/export_m.html |
| 0.57 TxURL | /public/search/b.html |
| 0.57 TxURL | /public/mailru/export_h.html |

Życie serwisu w kilku krokach cz. 5



Zastosowanie load balancerow
(także w backend).

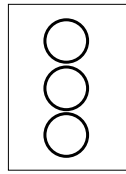
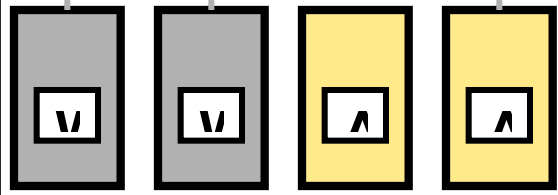


"High Scalability Building bigger, faster, more reliable websites".

Content switching



BIG-IP



Database System

SPEED

- SSL Acceleration
- Quality of Service
- Connection Pooling
- Intelligent Compression
- L7 Rate Shaping
- Content Spooling/Buffering
- TCP Optimization
- Content Transformation

SECURITY

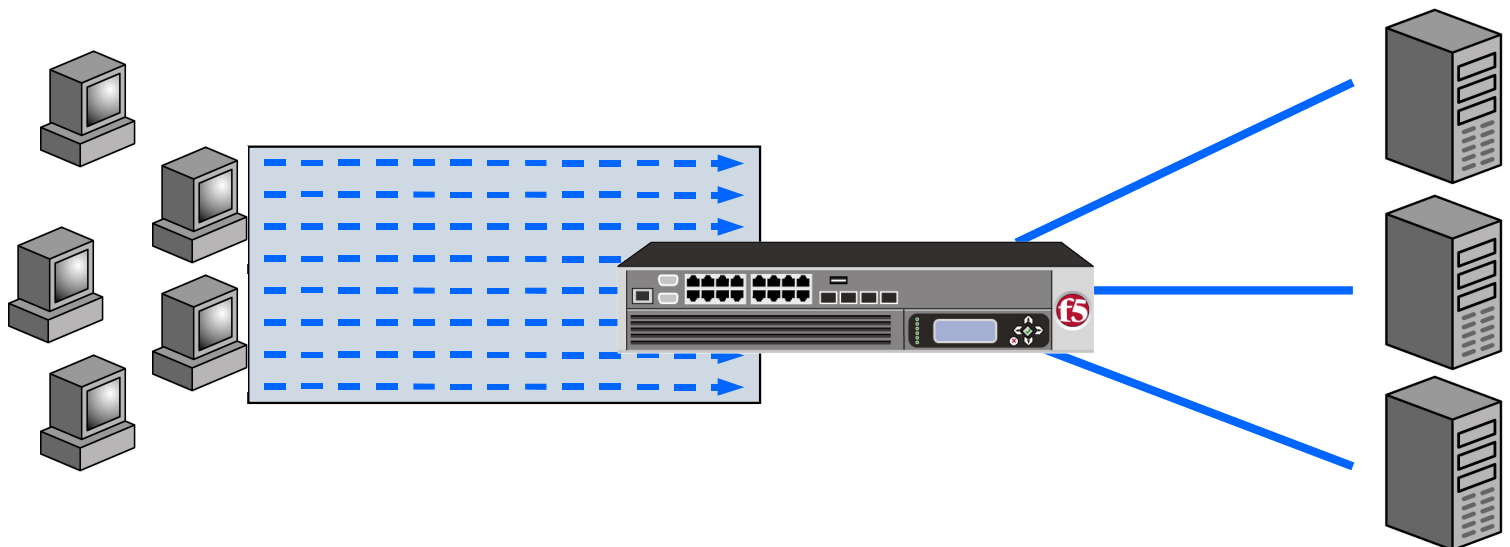
- DoS and SYN Flood Protection
- Network Address/Port Translation
- Application Attack Filtering
- Certificate Management
- Resource Cloaking
- Advanced Client Authentication
- Firewall - Packet Filtering
- Selective Content Encryption
- Cookie Encryption

AVAILABILITY

- Comprehensive Load Balancing
- Advanced Application Switching
- Customized Health Monitoring
- Intelligent Network Address Translation
- Intelligent Port Mirroring
- Universal Persistence

OneConnect™ – Connection Pooling

- ❖ Increase server capacity by 30%
 - Aggregates massive number of client requests into fewer server side connections
- ❖ Transformations form HTTP 1.0 to 1.1 for Server Connection Consolidation
- ❖ Maintains Intelligent load balancing to dedicated content servers



F5 - iRule

```
rule redirect_error_code {
  when HTTP_REQUEST {
    set my_uri [HTTP::uri]
  }
  when HTTP_RESPONSE {
    if {[HTTP::status] == 500} {
      HTTP::redirect http://192.168.33.131$my_uri
    }
  }
}
```

```
rule protect_content {
  when HTTP_RESPONSE_DATA {
    set payload [HTTP::payload [HTTP::payload length]]

    # Find and replace SSN numbers.
    reesub -all {\d{3}-\d{2}-\d{4}} $payload "xxx-xx-xxxx" new_response

    # Replace only if necessary.
    if {$new_response != 0} {
      HTTP::payload replace 0 [HTTP::payload length]
      $new_response
    }
  }
}
```

```
when HTTP_REQUEST {
  log local0. "VSERVER=[IP::local_addr]
  IP=[IP::client_addr] HOST=[HTTP::host]
  URI=[HTTP::uri]"

  HTTP::header insert "RealIP" [IP::client_addr]

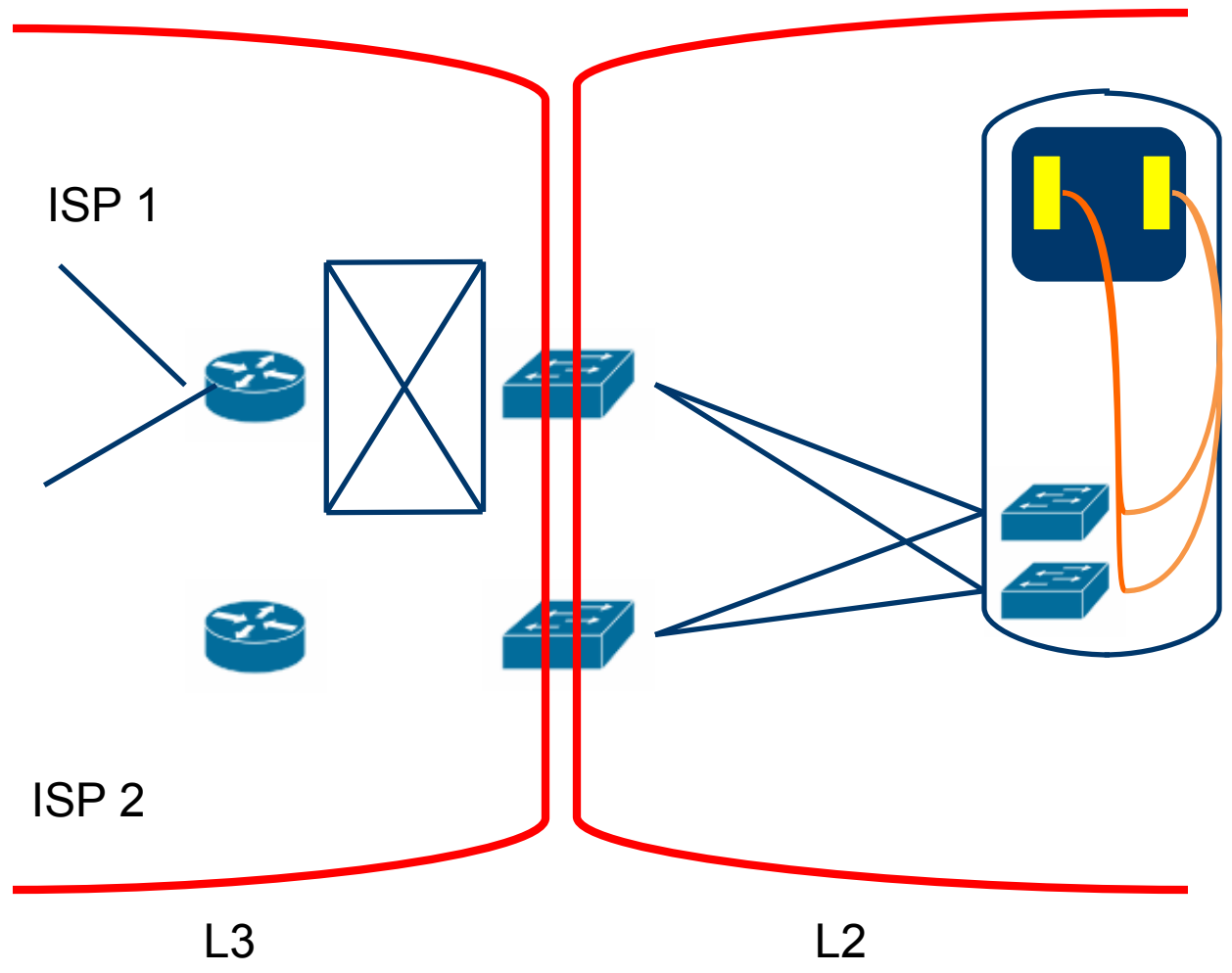
  if {[HTTP::uri] contains "/original/"} {
    pool original
  }
}
```

```
when CLIENT_ACCEPTED {
  TCP::collect
}
when CLIENT_DATA {
  #
  # Do a regex search and replace of binary TCP data
  #
  if {[regexp -indices "\x61\x62\x63\x64\x65\x66" [TCP::payload] firstmatch]} {
    set matchlen [expr [lindex $firstmatch 1] - [lindex $firstmatch 0] + 1]
    set replacement [binary format c* {97 98 99 0 100 101 102}]
    TCP::payload replace [lindex $firstmatch 0] $matchlen $replacement
    TCP::release
  }
}
```

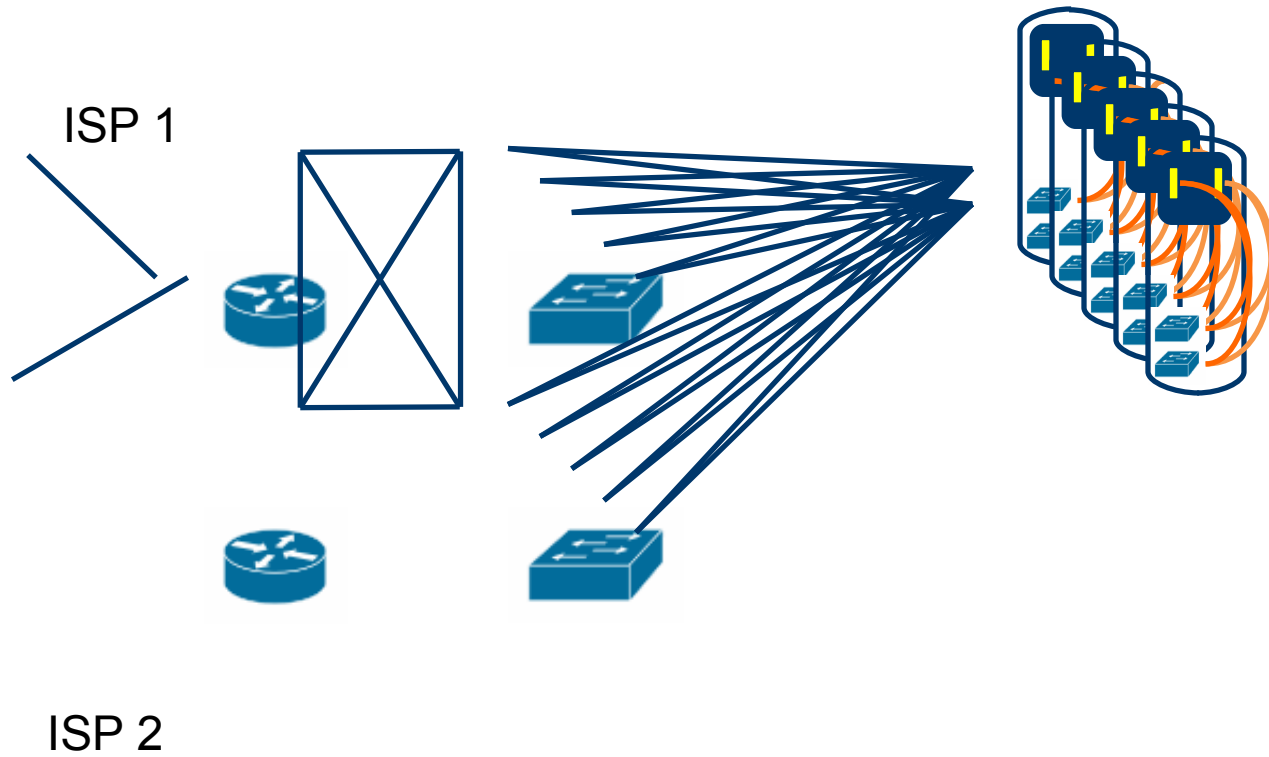
Kiedy brakuje nam czasu/wiedzy.



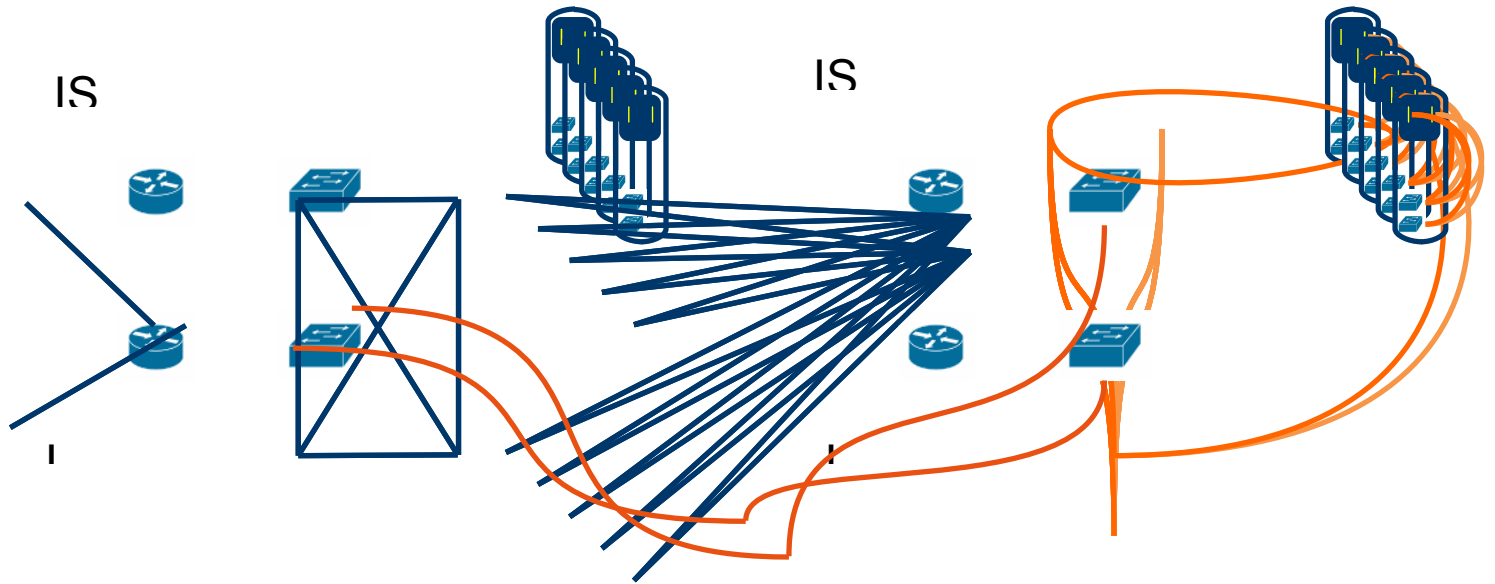
Sieć



Sieć



Sieć



Ostrożnie z ogn^AH^AH... STP, VTP, itp.

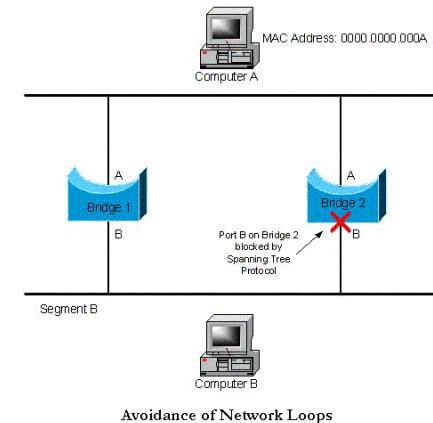
Spanning Tree:

- korzystanie z RSTP / MST.
- zweryfikowanie miejsc istotnych dla STP; root, designated ports.
- zmniejszanie wpływu działania STP na sieć, rozbijanie sieci na mniejsze domeny, rozdzielone L3.

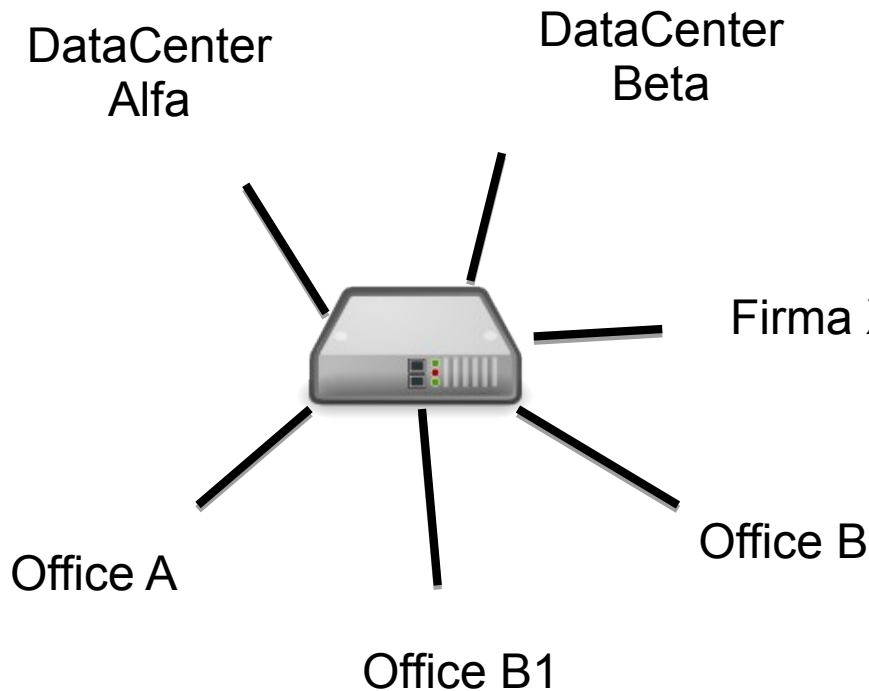
Ostrożnie z VTP.

Przygotowanie do 10G (FCoE).

Rozbijanie ruchu na wyizolowane vlany.



Planuj wykorzystanie adresów.



DC A: 10.1.0.0/16

DC B: 10.2.2.0/16

DC *: 10.0.0.0/14

VPN DC A: 91.21.4X./27

VPN DC B: 91.21.5X./27

Biura:

10.4.0.0/16 – Office A

10.5.1.0/24

10.5.2.1/24

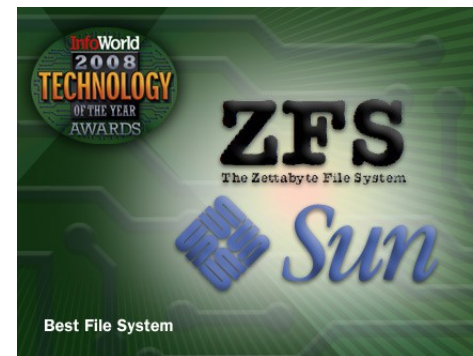
10.5.0.0/16 – Office BX

10.4.0.0/14 - Offices

Przechowywanie contentu.

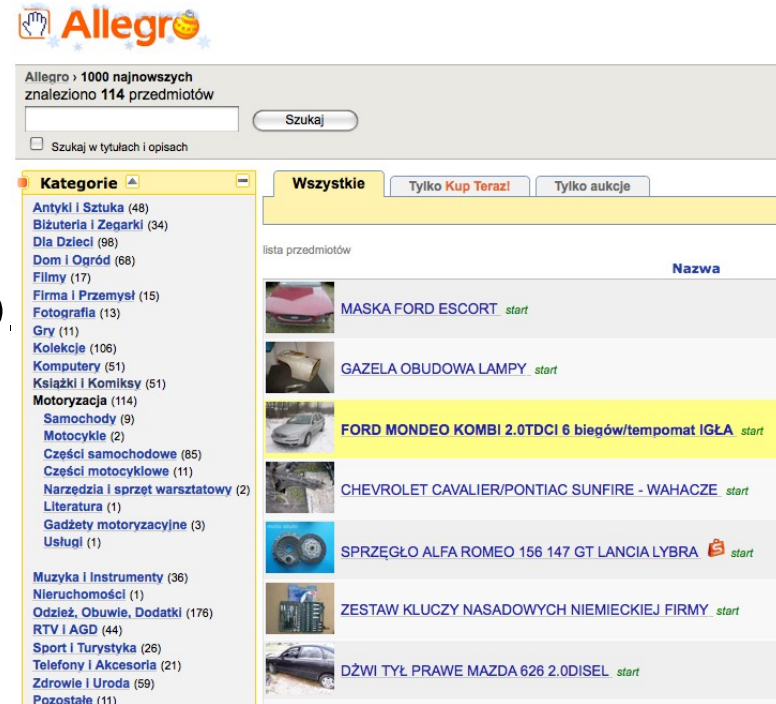
Rzeczy do rozważenia:

- Wielkość filesystemu
- Zapewnienie kopii danych
- Ilość IO/sek
- Struktura danych na FS
- Bezpieczeństwo FS
- Warstwa cachująca
- Wykorzystanie CDN
- Block size
- inodes



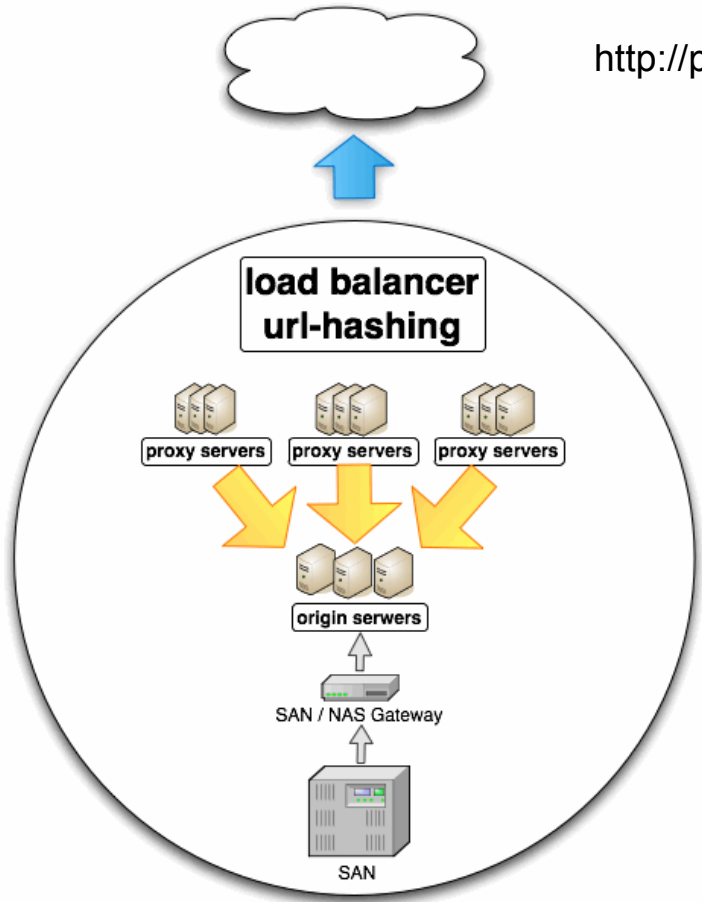
Przechowywanie contentu w QXL.

1. Bardzo duża ilość plików, ponad 200 mln (z dwóch miesięcy).
2. Szybkie zmiany -> zmiana obrazka na aukcji -> zamiana w proxy serwerach.
3. Pliki w różnych rozmiarach.
4. Problem z cachowaniem.
5. Problem z systemami plików.
6. Duża liczba requestów http -> IO.
7. Content nie może zniknąć.



The screenshot shows the Allegro website interface. At the top, the Allegro logo is visible. Below it, a search bar contains the text "Allegro > 1000 najnowszych znaleziono 114 przedmiotów". A search button labeled "Szukaj" is to the right. Below the search bar, there are filters for "Kategorie" and "Wszystkie" (with sub-filters "Tylko Kup Teraz!" and "Tylko aukcje"). The "Kategorie" list includes: Antyki i Sztuka (48), Biżuteria i Zegarki (34), Dla Dzieci (98), Dom i Ogród (68), Filmy (17), Firma i Przemysł (15), Fotografia (13), Gry (11), Kolekcje (106), Komputery (51), Książki i Komiksy (51), Motoryzacja (114), Samochody (9), Motocykle (2), Części samochodowe (85), Części motocyklowe (11), Narzędzia i sprzęt warsztatowy (2), Literatura (1), Gadżety motoryzacyjne (3), Usługi (1), Muzyka i Instrumenty (36), Nieruchomości (1), Odzież, Obuwie, Dodatki (176), RTV i AGD (44), Sport i Turystyka (26), Telefony i Akcesoria (21), Zdrowie i Uroda (59), Pozostałe (11). The main content area shows a list of items with a "Nazwa" header. The first item is "MASKA FORD ESCORT" with a "start" button. Other items include "GAZELA OBUDOWA LAMPY", "FORD MONDEO KOMBI 2.0TDCI 6 biegów/tempomat IGŁA", "CHEVROLET CAVALIER/PONTIAC SUNFIRE - WAHACZE", "SPRZĘGŁO ALFA ROMEO 156 147 GT LANCIA LYBRA", "ZESTAW KLUCZY NASADOWYCH NIEMIECKIEJ FIRMY", and "DŹWI TYŁ. PRAWA MAZDA 626 2.0DISEL".

Obrazki

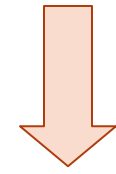


<http://photos03.allegro.pl/photos/128x96/424/82/80/424828054>



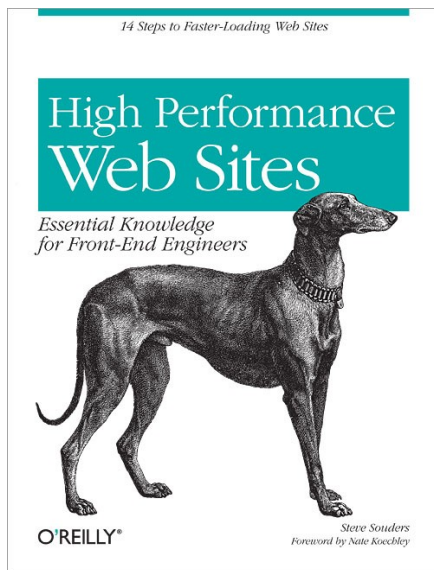
CRC32(URI)

$X \% 3$



Określone (zawsze te same) serwery cache.

Tunowanie I utrzymanie.



1. Make Fewer HTTP Requests
2. Use a Content Delivery Network
3. Add an Expires Header
4. Gzip Components
5. Put CSS at the top
6. Move Scripts to the Bottom
7. Avoid CSS Expressions
8. Make JavaScript and CSS External
9. Reduce DNS Lookups
10. Minify Javascript
11. Avoid Redirects
12. Remove Duplicate Scripts
13. Configure ETags

Narzędzia.

Monitoring:

- cacti
- nagios
- collectd
- cflowd
- SolarWinds
- Gomez

Zarządzanie:

- Sauron <http://sauron.jyu.fi/>
- Altiris
- rancid

Tikety

- Request Tracker
- OTRS

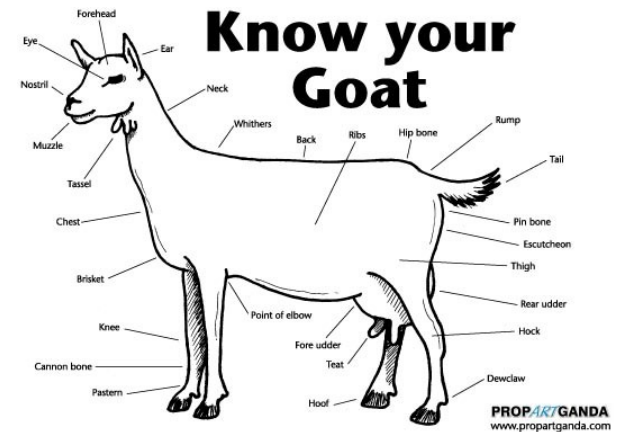


Szybyfowe prace...

```
while (true)
{
    identify_and_fix_bottlenecks();
    drink();
    sleep();
    notice_new_bottleneck();
}
```

* <http://highscalability.com/youtube-architecture>

- Keep It Simple Stupid!
- Szukaj nowych rozwiązań.
- Myśl długoterminowo (zapewnij skalowalność).
- Przygotuj się na wzrost.
- Skup się na rzeczach ważnych.
- Nie niedoceniaj narzędzi monitorujących.
- Korzystaj z wiedzy innych.
- ...



Koniec...

PYTANIA?

- <http://highscalability.com/youtube-architecture>
- <http://www.infoq.com/presentations/Second-Life-Ian-Wilkes>
- http://www.facebook.com/note.php?note_id=39391378919&

- <http://www.mysqlperformanceblog.com/>

- <http://www.datacenterknowledge.com/>
- <http://highscalability.com/>
- <http://www.ajaxperformance.com/>

- <http://developer.yahoo.com/yslow/>
- <http://getfirebug.com/>
- <http://www.gomez.com/>