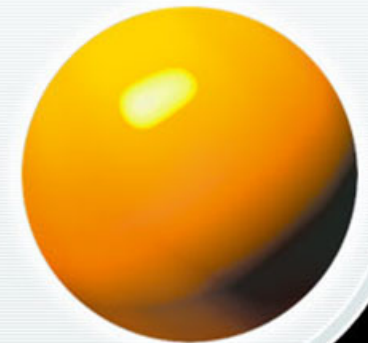




# Onet Moduły

Pawel.andrejas@portal.onet.pl



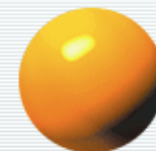
# Agenda:

- I. Rys historyczny
- II. Problemy skali
- III. Onet Moduły
- IV. Co dalej



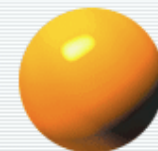
# Start

- 1996 – nowa marka na rynku  
Internet głównie na uczelniach, w firmach i domach – modemy
- jedna lokalizacja, jedno połączenie z Internetem



## Rozwój 1996 - 1999

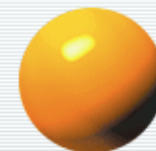
- kolejne zmiany lokalizacji, podyktowane wzrostem zapotrzebowania na miejsce dla ludzi i serwerów
- zasilanie - 12 kVA
- wzrost ruchu, kolejne łącza operatorskie



## Dalszy rozwój...

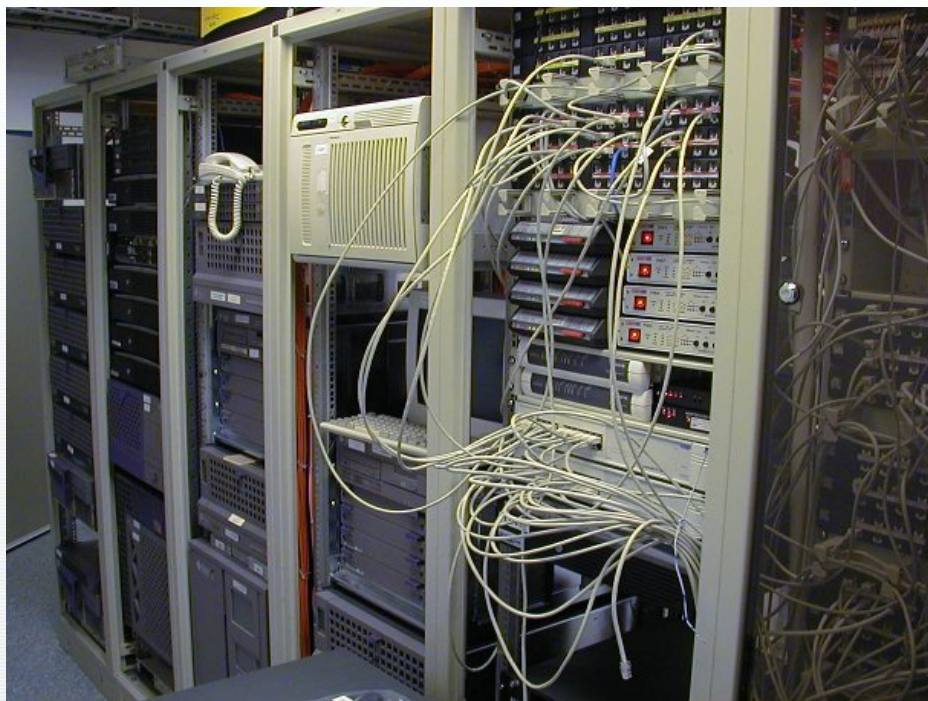


- 2002: wchodzimy do nowo wybudowanej serwerowni
- większe bezpieczeństwo
- więcej miejsca (dla ponad tysiąca serwerów)
- 320 kVA mocy
  
- rozwój usług, coraz większe przepustowości, usługi video
- więcej sprzętu, coraz większe zapełnienie serwerowni
- 2006 – pierwsze kolokacje

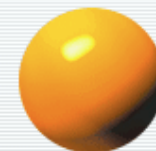


# Pierwsza dedykowana serwerownia – AD 1999

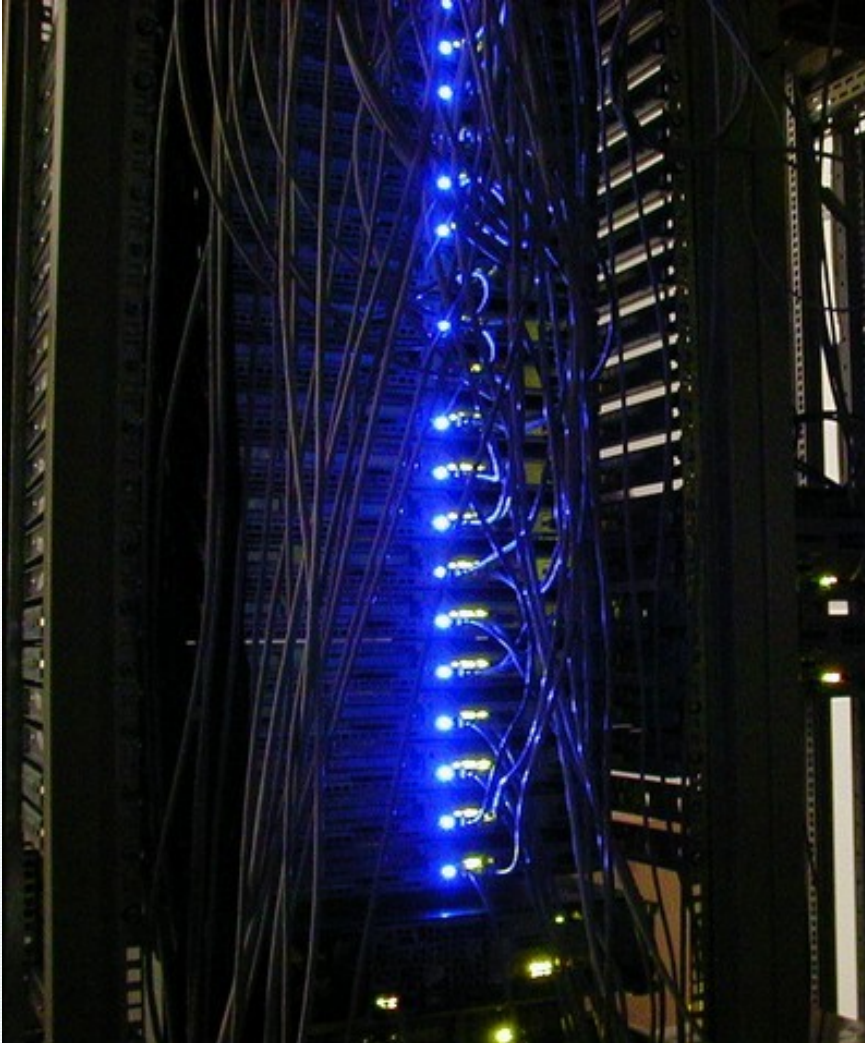
- 10 szaf serwerowych
- 32 kVA mocy
- ponad 100 serwerów



- kolejne setki Mbps
- znów mało miejsca
- SPOF?



# Onet.pl do AD 2008:



- dwie własne serwerownie
- więcej urządzeń w kolokacji niż „u siebie”
- budynek biurowy z serwerownią backoffice
- dodatkowe biura w Krakowie i Warszawie
- 1500 serwerów
- 8 dużych routerów, kilkadziesiąt switchy rackowych i drugie tyle w paczkach blade



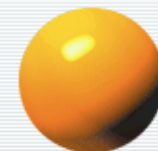
# Agenda:

- I. Rys historyczny
- II. **Problemy ze skalą**
- III. Onet Moduły
- IV. Co dalej



# Rok 2006

- 2 duże serwerownie w tym jedna kolokowana, w każdej z nich jest po kilkaset serwerów.
- Mamy coraz większe farmy serwerów http, reverse proxy, aplikacyjnych, klastrów bazodanowych.
- Stopień skomplikowania zależności między aplikacjami wzrasta.
- Dziesiątki Vlanów.
- Jeden duży nie routowalny Vlan „NAS” do którego wpięty jest każdy serwer
- Aplikacje głównie komunikują się po „NASie”
- Request flow wielokrotnie krąży między DC



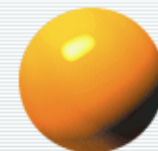
## L2 czy L3?

1. w jednej serwerowni – nie ma problemu
  
1. w dwóch serwerowniach – rozciągnięcie L2 jest najbezpieczniejszym sposobem na wejście do drugiej serwerowni.
  - miejscami pojawia się L3
  - globalna zmiana na L3 jest trudna, ponieważ wymaga dużych zmian po stronie aplikacyjnej – łatwiej dokładać, niż modernizować
  
1. kolokacje i kolejne DC – tu zaczynają się schody



# Pojawiają się problemy

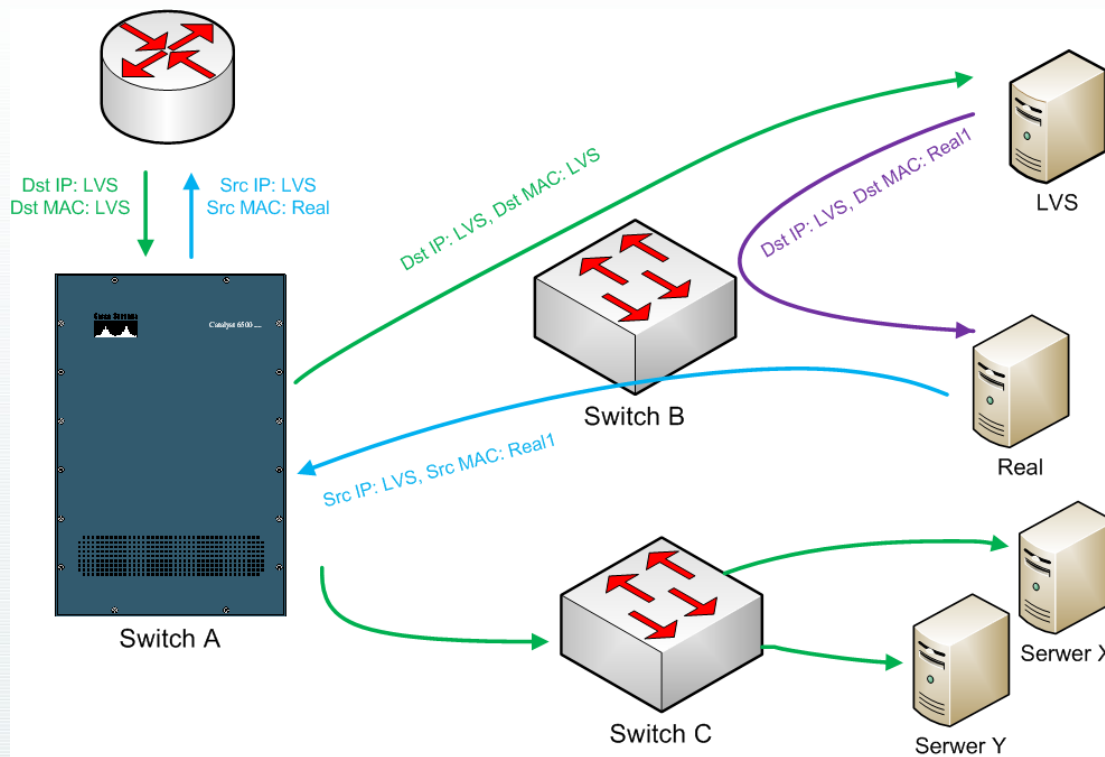
- Trudno izolować awarie
  - Problemy z dużym serwisem, odbijają się na dostępności reszty. (duża farma serwerów)
  - Awaria aplikacji używana w każdym serwisie ale nie istotna dla podawania serwisu (sonda, ivona) powoduje czkawkę na wszystkich serwisach.
  - LVS pracują w DR, wystawienie adresu na serwerze bez odpowiednich wpisów powoduje awarie.
  - Dużo serwerów zadaniowych (SPOFów)



# Dalsze Problemy...

Rozrastanie się L2 i problemy skali:

- sztormy broadcastowe
- MAC aging...



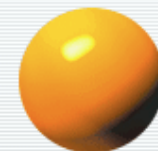
# Awaria jednego DC?

- Nierówno rozłożone farmy serwerów między DC
- Poszczególne klastry LVSów w jednej lokalizacji
- L2 rozciągnięte między DC („NAS”)

Gdy tracimy połączenie między DC lub mamy awarię 1 DC, aby przywrócić do działania top10 potrzeba dużego wysiłku administratorów i programistów.

- Postawienie brakujących klastrów lvs
- Zrównoważenie klastrów aplikacyjnych
- Odcięcie niepotrzebnych usług „nas” nie mamy firewall, wstawiamy adres który resetuje połączenia do danej usługi

Testy NRD (Niezawodność i redundancja ) są niemożliwe.

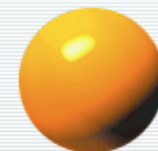


## Budowa nowego DC

W perspektywie mamy wybudować nowe DC i zmigrować się do niego.

- Nie możemy rozciągnąć L2 do nowego DC
- Zakup dużej ilości serwerów jako setupu migracyjnego – nieekonomiczne.
- Długi proces migracyjny, również nieekonomiczny -> płacimy za kolokacje

Pojawia się pomysł budowy Modułów.....



# Co dalej?

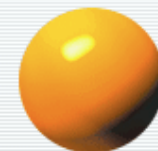


Zmiany...



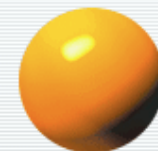
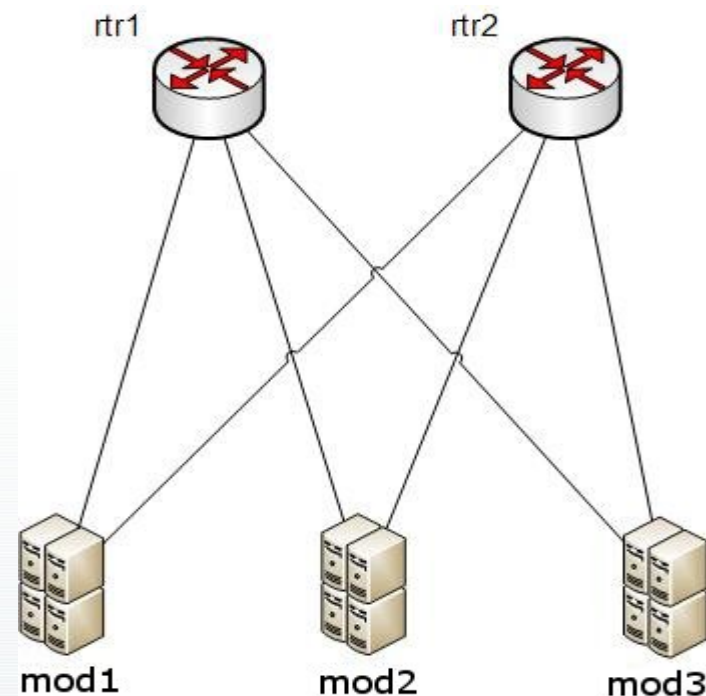
# Agenda:

- I. Rys historyczny
- II. Problemy skali
- III. **Onet Moduł**
- IV. Co dalej



# Onet Moduł

- Wydzielony fragment infrastruktury technicznej Onetu.
- Może funkcjonować niezależnie od reszty infrastruktury.
- Komunikacja po L3, z wykorzystaniem routingu dynamicznego
- możliwość przejęcia usług innego modułu
- redundantne połączenia do warstwy agregacji

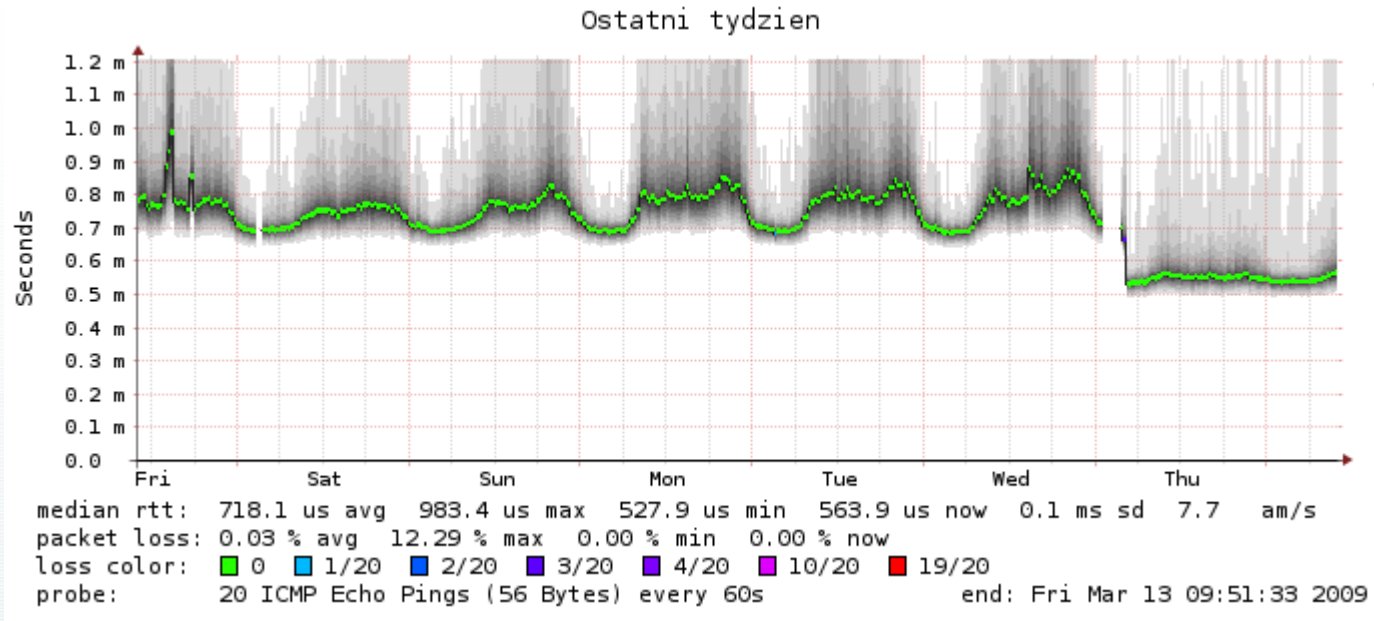


# Zalety Modułów

1. Zwiększenie niezależności usług/serwisów
2. Zwiększenie stabilności
3. Możliwości wywiezienia do innej serwerowni
4. Możliwość przełączania usług między modułami
5. Ułatwienie migracji
6. Izolacja awarii
7. Brak SPOFów

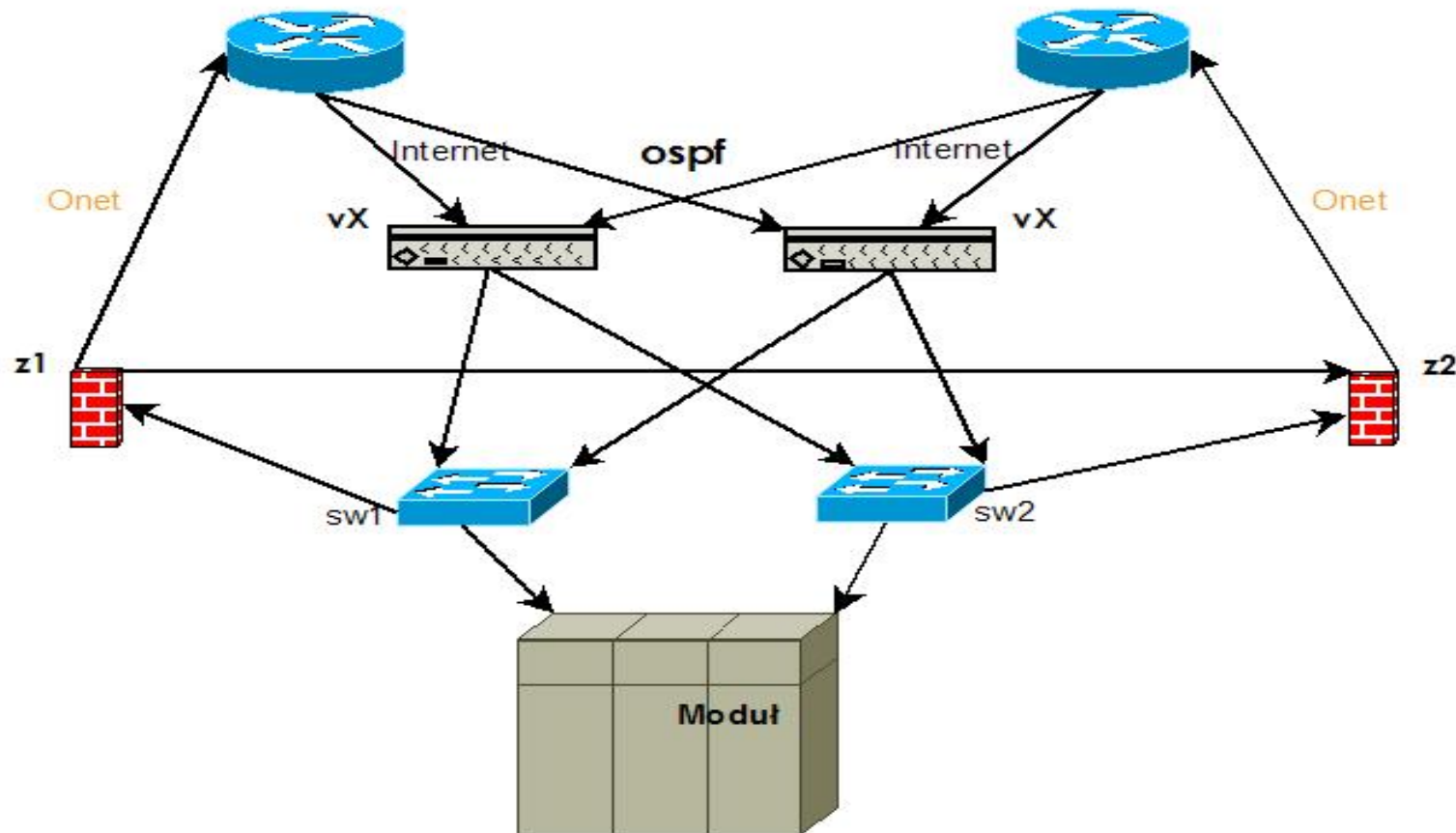


# L2 v L3 między DC



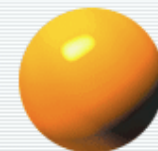
# Architektura Modułu

Onet Moduł - Hardware



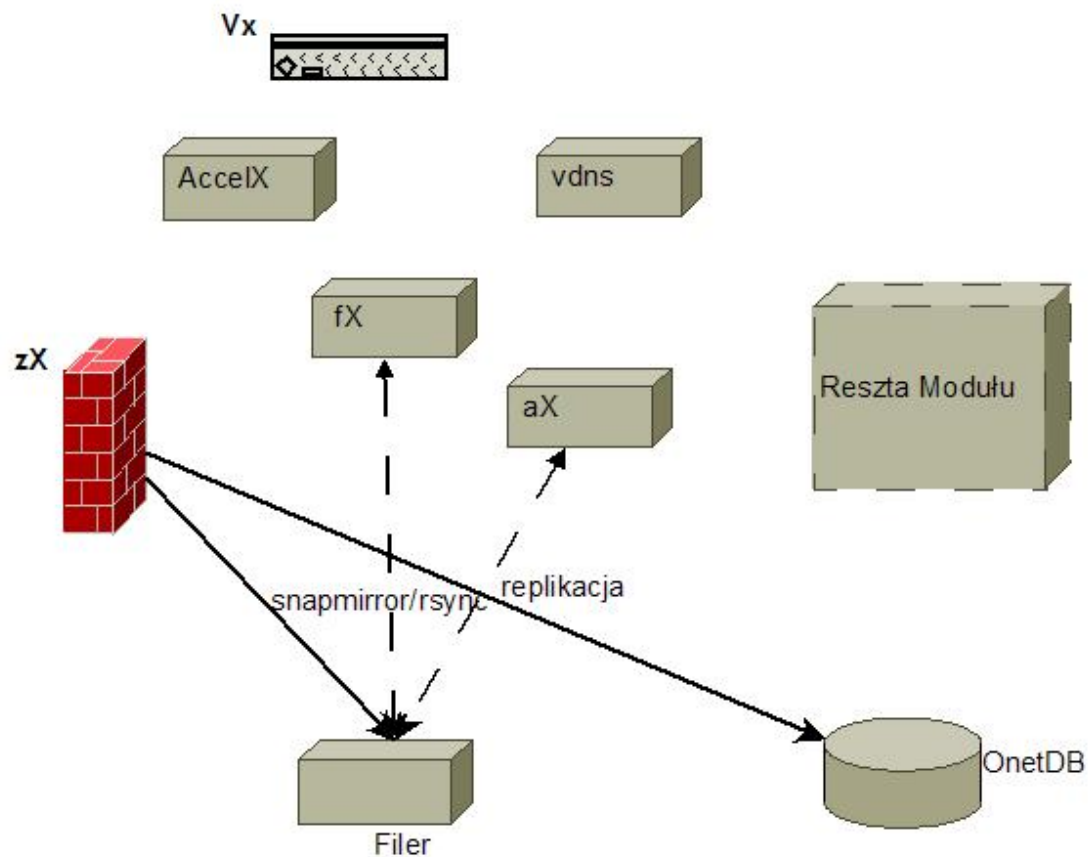
# Architektura (Hardware)

1. Pełna redundancja sprzętowa.
2. N+1 switch (LAN modułu)
  - na serwerach skonfigurowany bonding.
  - dynamiczny routing wewnątrz modułu. (rip)
3. N+1 LVS
4. 2 zX –(pływający gateway)
  - główna funkcje router i firewall
5. 2 routery – połączone LVSy, zX.
  - ospf między nimi (metryki)



# Architektura (Software)

Onet Moduł - Software



# Architektura (Software)

1. Synchronizacja kodu snapmirror/rsync.
2. Replikacja natywna baz.
3. Komunikacja z modułami przez zX.
4. Firewall na zX.
  - kontrola i odcinanie usług.
5. Brak dostępu do „NAS”.
6. System monitoringu na każdym module raportujący do centrali
7. Błędy php w monitorze.



# Architektura (software)

1. Lvs pracujący w trybie
  - acceleratory – Masq
  - reszta – DR
  - SureAlievD – do testowania reali
2. Nat na Lvs'ach dla wybranych usług
3. Zmiany w aplikacjach , są świadome w którym module się wykonują.
4. Usługi poza modułowe, w wypadku utraty połączenia są zaślepiane (serwisy podają się dalej)

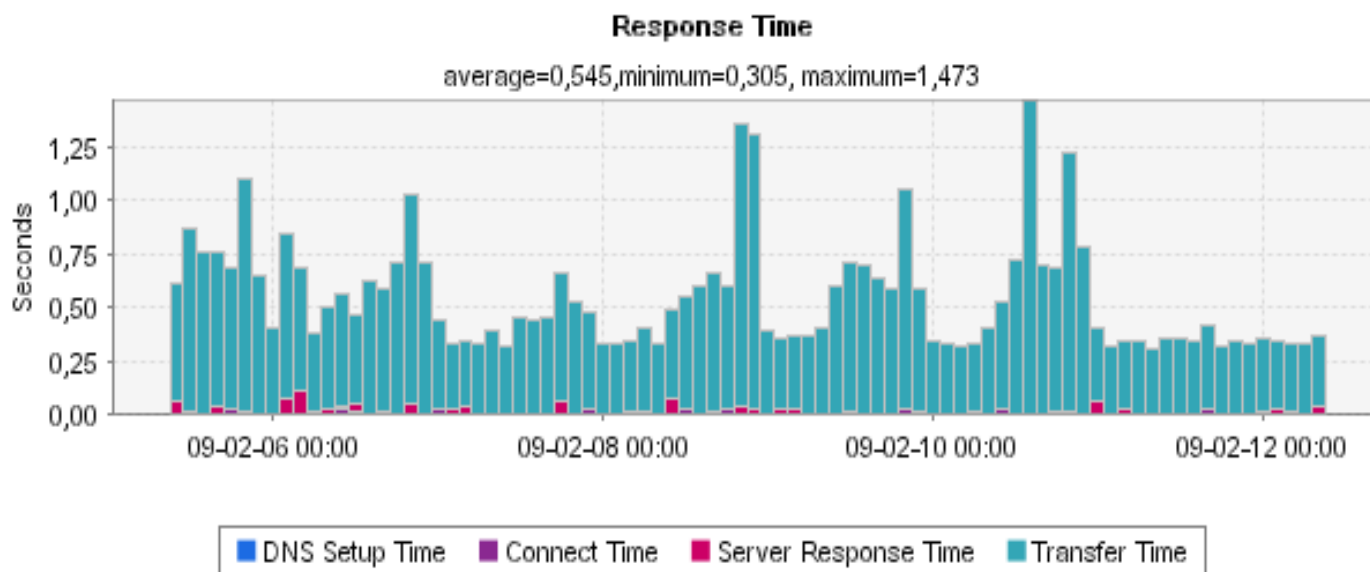


# Początki...

1. Pierwszy moduł powstaje w 2006 roku
  - Odbiega od założeń, ma możliwość komunikacji z „Nasem”
  - brak ospfa (infrastruktura sieciowa jeszcze nie gotowa)
  - Są jeszcze SPOFy, brak pełnej redundancji sprzętowej
1. Po tych doświadczeniach w 2008 powstaje w pełni redundantny moduł
2. Rozpoczynamy testy „NRD” 2x do roku



# Migracja serwisu na moduł

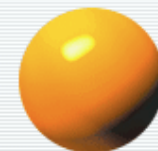


## 2009 migracja do nowego DC

Umiemy budować moduły ale potrzebujemy narzędzia który przyspieszy i zautomatyzuje jego budowę.

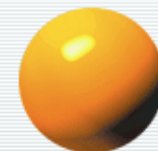
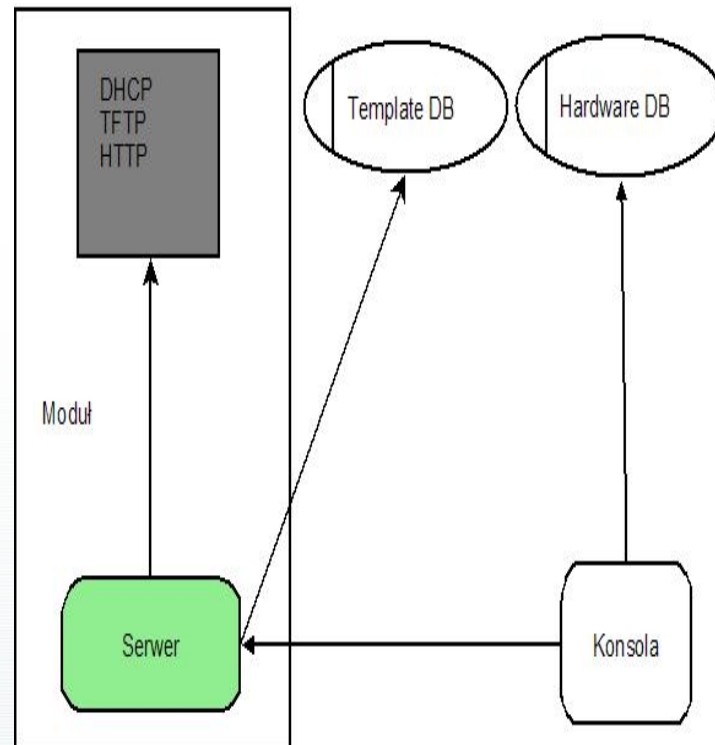
- Powstaje PRN (Portal Root NFS)
  - Centralne miejsce do automatycznej instalacji serwerów.
  - Baza hardware.
  - Dhcp, tftp, http
  - Template (serwisy, pliki (tagi), aliasy, dziedziczenie, podział partycji, raid)
  - Serwery (ip i mac)
  - Obrazy poszczególnych klastrów
  - Zdalny update plików na serwerach (historia plików)

Moduł możemy uruchomić w ciągu kilku godzin



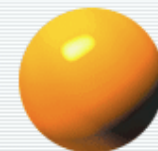
# Jak działa PRN..

- PRN znajduje serwer w bazie sprzętu.
- PRN przestawia bootowanie serwera na siec i restartuje serwer.
- Serwer ściąga instalator po tftp
- Instalator ściąga dane z DB
- Instalator konfiguruje, raid, dyski, filesystem
- Instalator ściąga obraz po http, rozpakowuje , instaluje bootloader zmienia ustawienia bootowania i restart



# Podsumowanie.

- Ospf i wystawianie do 16 ścieżek danego adresu IP (ECMP)
- Migracja usług między modułami
- Serwis na n+1 modułów w stanie active/active
- 80% serwisów podajemy z modułów
- W nowym DC tylko moduły, obecnie jest ich 10
- Czas odtworzenie modułu jest pomijalny w stosunku do czasu wstawienia hardware do szaf i skonfigurowania sieci



# Agenda:

- I. Rys historyczny
- II. Rozwój i problemy sieci
- III. Nowa struktura
- IV. **Co dalej**

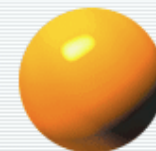


# Kolejny krok

## Problemy...

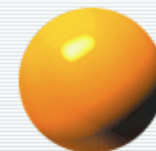
- Skalujemy się na eventy, większe koszty utrzymania modułu
- Moduły nie mogą „pożyczać” sprzętu
- Klastry serwerów nieoptymalnie wykorzystują hardware

Potrzebujemy rozwinąć idee modułów..



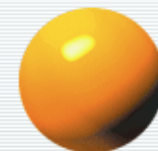
# Wirtualizacja

1. Optymalne wykorzystanie sprzętu
2. Rozwój PRNa
  - automatyczna instalacja virtualnych maszyn
  - Konsola zarządzająca virtualkami
  - Live migration



# Cloud

- Możliwość „pływania” zasobów między modułami
  - Jednolity sprzęt BladeCenter może posiadać 4 switchy, rezygnujemy z redundancji, podpinamy do 4 modułów
  - PRN zarządza zasobami
  - Zmiany na serwerach tylko za pomocą PRNa
  - Konsola do administracji loadbalancerami
  - Automatyczne przechodzenie serwerów w standby/active w zależności od obciążenia



# Pytania?



Dziękuję za uwagę.  
Pawel.Andrejas@portal.onet.pl

