Wysokowydajny system składowania plików graficznych ze zdjęciami użytkowników portalu

Tomasz Paszkowski

Agenda

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Informacje o spółce Nasza Klasa Sp. z o.o.
- Pierwsze wersje systemu zdjęć N-K
- Opis ogólny systemu zdjęć
- Charakterystyka techniczna:
  - Load balancing
  - Frontend
  - Backend
- Następne kroki
- Inne rozwiązania

Agenda

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

Informacje o spółce Nasza Klasa Sp. z o.o.

Nasza Klasa Sp. z o.o. Spółka

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Część grupy FORTICOM posiadającej serwisy społecznościowe min. w Rosji: odnoklassniki.ru, Litwie: one.lt,videogaga.lt, Łotwie: one.lv,videogaga.lv
- Bogate doświadczenie oraz know-how wniesione przez Forticom pozwoliło nam przejść bardzo płynie od drobnej firmy do dużej spółki internetowej
- Od początku działalności spółka intensywnie rozwija się w sposób organiczny
- Operujemy z trzech biur: Wrocław,Kraków,Warszawa
- Zatrudniamy ponad 120 osób w takich działach jak
  - IT
  - Finanse,Księgowość,Administracja (w tym HR)
  - Marketing oraz PR
  - Sprzedaży

- Rozwijamy i utrzymujemy najpopularniejszy serwis społecznościowy w polskim internecie

Nasza Klasa Sp. z o.o.
No. 1 w Polskim internecie

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

Źródło: Megapanel PBI/Gemius, czerwiec 2009, grupa: internauci w wieku 7+

	Użytkownicy	Odsłony	Czas na użytkownika
nasza-klasa.pl	***11 640 468***	***10 440 971 023***	***9 h 17 min***
Grupa Google	15 474 727	4 120 521 558	7 h 50 min
Grupa Allegro.pl	10 784 557	3 894 936 928	3 h 57 min
Grupa Onet.pl	12 380 590	3 776 197 970	5 h 11 min
Grupa Wirtualna Polska	10 803 086	2 494 419 422	4 h 11 min
Grupa INTERIA.PL	10 268 501	1 273 816 957	2 h 56 min
YouTube.com	9 217 498	1 226 900 717	3 h 48 min
Grupa o2.pl	9 612 326	1 224 590 247	3 h 10 min
Grupa Gazeta.pl	9 910 182	851 517 174	1 h 35 min
wikipedia.org	7 728 637	214 229 742	39 min 9 s

Nasza Klasa Sp. z o.o. Infrastuktura

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

Dwie serwerownie: Poznań, Warszawa

- ~ 1000 serwerów
- ~ 100 macierzy dyskowych fiber channel (SAS,SATA,FC)
- ~ 220TB dostępnej przestzenii dyskowej w RAID10
- ~ 8Gb/s ruchu do sieci Internet
- ~ 200 000 zapytań na sekunde kierowanych do warstwy frontendowej w szczycie
- Większość naszego system zbudowana na
- oprogramowaniu opensource
  - Nginx,Apache,LightHttpd
  - MySQL,Memcache
  - Linux Debian
  - LVS,Haproxy,Squid
- Sporo oprogramowania napisaliśmy samodzielnie
  - NKDB,NkCache

Agenda

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

Pierwsze wersje systemu zdjęć N-K

Początki

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Jak każdym szanujący się startup zaczynaliśmy od serwera wirtualnego, gdzie cała aplikacja znajdowała się w jednym drzewie
- W Pierwszym tygodniu wyczerpany cały roczny limit transferu
- W następne kilka dni wyczerpany transfer na kolejny rok
- Przenosiny na dedykowany serwer w niemieckiej serwerowni
- Problemy wydajnościowe i decyzja o wynajęciu dedykowanego serwera pod zdjęcia, upload z aplikacji PHP po NFS-ie
- Dołożenie kolejnych serwerów pod zdjęcia nastąpiło bardzo szybko, upload z aplikacji PHP nadal po NFS-ie (każdy serwer PHP ma podmontowany każdy serwer ze zdjęciami po NFS-ie)
- Po osiągnięciu poziomu kilku serwerów PHP i kilku serwerów pod zdjęcia zapanowanie nad odpowiednim eksportem NFS staje się bardzo trudne ;)
- Przejście na system MogileFS (po pozytywnych doświadczeniach z memcache od tego samego zespołu developerów)
- Po roku mamy 160 serwerów obsługujących zdjęcia za bardzo nie równym obciążeniem (sieć,pojemność,wydajność)

Początki

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Zainwestowaliśmy bardzo dużo czasu w rozpracowanie MogileFS, min. przepisujemy trackera w C aby rozwiązać problem wydajnościowy oraz nie udolne komunikowanie się z bazą
- Z pierwotnej funkcjonalności tracker finalnie używalismy tylko 20-30% jego funkcjonalności pozostałe elementy zaimplementowaliśmy sami (min. keszowanie wyników zapytań w memcache)
- Centralna baza danych z lokalizacjami zdjęć się nie sprawdza, przez moment pomagamy sobie replikacją (kilka serwerów slave)
- W serwerach slave mamy problem z replication lag, postanawiamy wszystkie nowo dodane zdjęcia dodawać od razu do memcache
- Na bazie doświadczeń budujemy idea naszego systemu zdjęć i przystępujemy do jego implementacji
- Przenosiny do nowego systemu połączone z przeniesieniem systemu zdjęć do nowej serwerowni w Polsce

Agenda

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

Opis ogólny systemu zdjęć

System zdjęć
Usługa

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Podstawowym elementem naszego serwisu jest usługa hostingu zdjęć
- Odnawialny m-c limit na transfer zdjęć (nie ma potrzebny kasowania starych zdjęć)
- Możliwość dodawania wielu zdjęć naraz
- Możliwość komentowania zdjęć
- Możliwość oceniania zdjęć
- Możliwość polecania zdjęć
- Możliwość grupowania zdjęć w albumy
- Możliwość pinezkowania zdjęć
- Powiadomienie znajomych o nowo dodanym zdjęciu, komentarzu do zdjęcia itp. itd
- Zdjęcia prezentowane użytkownikom w trzech wersjach
  - Miniaturka
  - Duża wersja zdjęcia dopasowana tak aby mieściła się na naszej stronie
  - Oryginalna wersja

System zdjęć
Założenia projektowe

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Zagwarantowane równomierne obciążenie każdego z elementów systemu
- Mocno rozbudowane mechanizmy keszujące gwarantujące najwyższą wydajność systemu
- Minimalizacja operacji odczytu/zapisu do dysku jako operacji najbardziej czasochłonnej
- System prosty w obsłudze rozbudowie i migracji (w tym geograficznej z jednej serwerowni do drugiej)
- Możliwość lokalizacji poszczególnych elementów systemu (backend,frontend) w geograficznie oddalonych od siebie lokalizacjach z łącznością tylko przez sieć Internet
- Odporność na awarie pojedynczych elementów systemu
- Możliwość odtworzenia całego systemu w skończonym czasie po klęskach żywiołowych
- Bezpieczeństwo danych na pierwszym miejscu

System zdjęć
Charakterystyka ruchu

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Największą popularność mają zdjęcia dodane najwcześniej
- W większości przypadków prezentujemy tylko miniaturki zdjęć (powiadomienia,lista znajomych,boksy z ostatnio dodanymi zdjęciami)
- Hit rate dla cache miniaturek zdjęć powyżej 98% !
- Aby to osiągnąć konieczne jest przemyślane zaprojektowanie pamięci cache.
- Ilość zapytań o zdjęcia utrzymuje się na stały poziomie, nie ma potrzeby skalowania wydajnościowej jedynie pojemnościowej (stare zdjęcia przestają być oglądane ale cały czas pozostają dostępne). Dzięki temu mamy bardzo duży zapas wydajnościowy na backendzie (o wydajności decyduje ilość dostępnych dysków).
- Bardzo ważne jest równomierne rozproszenie nowo dodawanych zdjęć po równo na każdy z węzłów
- Każde zdjęcie naszego użytkownika traktujemy jak dane (BLOB), które nie mogą absolutnie ulec zniszczeniu czy utracie
- Wszystkie zdjęcia dostępne pod adresem: photos.nasza-klasa.pl
- 3,5Gb/s ruchu do sieci Internet w szczycie
- 100 000 zapytań na sekundę w szczycie

Agenda

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

Charakterystyka techniczna:
Load balancing

Load balancing
L3, LVS + ECMP

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- photos.nasza-klasa.pl dla świata to jest pojedynczy adres IP: 91.206.173.228
- Load balancing rozpoczyna się już na routerach brzegowych, poprzez rozproszenie zapytań na klaster serwerów LVS za pomocą tradycyjnego ECMP (L3 load balancing)
- Routery brzegowe kierują ruch na virtualne adresy VRRP (keepaliaved) serwerów LVS
- Keepalived skonfigurowany w trybie backup-backup, aby uniknąć flapowania serwisu w przypadku problemu z jedną z maszyn, na każde trzy maszyny LVS stawiamy jedną nadmiarową
- Każdy serwer LVS składa się z czterech kart sieciowych, ruch przychodzący kierowany jest do każdego z serwerów LVS po trzech kartach sieciowych (optymalizacja wykorzystania przerwań na serwerach z procesorami wielordzeniowymi)
- Serwery LVS wyposażone w karty sieciowe z MSI-X (cztery IRQ dla jednej karty sieciowej)
- Serwery LVS pracują w trybie direct routing
- LVS rozpraszają ruch na klaster serwerów HAPROXY
- Detekcją martwych serwerów HAPROXY zajmuje się ldirector

Loadbalancing
L3-L4, HAPROXY

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- HAPROXY terminuje połączenie od użytkownika i zestawia nowe żądanie do warstwy frontendowej
- HAPROXY jest bardzo wydajnym i stabilnym rozwiązaniem do rozpraszania ruch z zaawansowanymi funkcjami kontroli i kierowania ruchem
- W celu wykorzystania wielordzeniowych procesorów HAPROXY również wyposażone jest w cztery karty sieciowej i otrzymuje pakiety przychodzące po trzech z nich
- W celu optymalizacji pamięci cache w frontendzie na URL-u zapytania o zdjęcie liczony jest hash i dzięki temu żadania o wydzielone grupy zdjęć trafiają zawsze na ten sam węzeł frontendu (każde zdjęcie keszowane jest tylko w jednym z węzłów)
- Url hashing gwarantuje nam również równomierne rozłożenie obciążenia pomiędzy każdy z węzłów (ruch sieciowych + ilość zapytań)
- HAPROXY odpowiada u nas defacto za normalizacje zapytań oraz odpowiedzi (wycięcie zbędnych bądź niebezpiecznych informacji)
- HAPROXY aktywnie sprawdza dostępność serwerów frontendowych (HTTP check) i w przypadku wykrycia awarii jednego z nich przestaje tam kierować ruch

Agenda

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

Charakterystyka techniczna: Frontend

Frontend Download

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- photos.nasza-klasa.pl/numer_galerii/numer_zdjecia/typ_zdjecia/suma_kontrolna.jpeg
- W URL-u zakodowane mamy
  - Numer galerii
  - Numer zdjęcia
  - Typ zdjęcia
  - Suma kontrola (wyliczana również z md5 liczonego na zawartości zdjęcia)
- Zadaniem frontendu jest
  - Weryfikacja poprawności URL-a poprzez przeliczenie sumy kontrolnej
  - Jeżeli nie ma zdjęcie w pamięci kesz to pobranie go z backendu
  - Blokada nie pożądanych zdjęć
  - Wykrywanie martwych węzłów backendu i przełączanie się na zapasowy w przypadku awarii
- W chwili obecnej frontend dysponuje ~ 2TB pamięci kesz służącej wyłącznie do przechowywania zdjęć użytkowników, nie mieszczące się wpisy usuwane LRU
- Oprogramowaniem napędzającą tą warstwę jest Squid + nasza własna aplikacji podpięta do niego przez url redirector (stdin/stdout)

Frontend Upload

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Aplikacja w PHP przystosowana do przyjmowania uploadu wielu zdjęć naraz, defacto zintegrowana z główną aplikacja portalową
- Zadaniem aplikacji jest
  - Pobranie zdjęcia/zdjęć od użytkownika
  - Transraiting (imagemagick)
  - Usunięcie niebezpiecznych elementów plików graficznych (EXIF)
  - Wygenerowania trzech form pliku ze zdjęciem (miniaturka,duży,oryginalny format)
  - Zapisanie zdjęcia w backendzie

Agenda

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

Charakterystyka techniczna: Backend

Backend

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Zadaniem backendu jest przechowywanie plików ze zdjęciami, oraz ich serwowanie do warstwy frontendowej
- Ilość zapytań dochodzących do tej warstwy jest bardzo mała z racji bardzo dużej pamięci kesz warstwy frontendowej
- Każdy z serwerów backend ma dodatkowo dedykowane dla siebe do 16GB pamięci cache.
- Backend zbudowany jest w sposób nadmiarowy, każdy element węzla jest zdublowany z racji dużej ilości danych i co jest związane z tym długiego czasu odtworzenia z backupu
- Wszystkie dane z każdego z węzłów backendu backupowane w oddalonej geograficznie lokalizacji z pełnym zestawem danych, służącym do przywrócenia pełnej funkcjonalności usługi w krótkim czasie po totalnym zniszczeniu DC (pożar,trzęsienie ziemi,powódź ...)
- Warstwa frontendowa odpowiedzialna jest za upload plików ze zdjęciami do każdego z elementów

Backend

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Sprzętem tworzącym element węzła jest serwer wyposażony w 16GB pamięci RAM, karte HBA oraz przyłączonymi do niego macierzami fiberchannel
- Węzły podzielone są dwie części, jedna służy do obsługi miniaturek druga do obsługi pozostałych formatów zdjęć
- Miniaturki przechowujemy na macierzach z dyskami SAS
- Pozostałe formaty przechowujemy na macierzach z dyskami SATA bądź dyskami SATA z interfejsem SAS (nowość, szybsze od swoich odpowiedników SATA), wszystkie LUN-y w RAID10
- System skalujemy na pojemność ponieważ użytkownicy regularnie dodają nowe zdjęcia nie kasując starych
- Mamy ogromny zapas mocy (I/O) w tej części systemu z racji ciągłej konieczności dokładania nowych macierzy z dyskami twardymi
- Średnio m-c wymieniamy około 20 dysków twardych SATA, które ulegają awarii

Backend

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Backend podzielony na tzw. strefy (zony). W tej chwili mamy 1024 strefy dla dużych zdjęć oraz 256 stref dla miniaturek
- Przynależność zdjęcia do strefy wyliczona na podstawie funkcji hashującej przyjmującej jako parametr:
  - Numer galerii
  - Numer zdjęcia w galerii
  - Typ zdjęcia
  - Md5 z zawartości pliku ze zdjęciem
- Dzięki strefom uzyskujemy równomierne obciążenie każdego z węzłów pod względem:
  - Pojemnościowym
  - Obciążeniowym
  - Poziomu ruchu sieciowego

Backend NKFS

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Pliki ze zdjęciami trzymamy w napisanym przez nas rozwiązaniu określanym mianem NKFS
- Każda ze stref reprezentowana jest przez pojedynczy plik na systemie plików XFS, w którym zawarte są wszystkie jej zdjęcia
- Struktura pliku jest bardzo prosta i zawiera kolejne zdjęcia
- Na każdym z serwerów backendowych uruchomiony jest serwery MySQL, który zawiera dane o ofssetach poszczególnych zdjęć w plikach stref.
- Nowe zdjęcie dopisywane jest na końcu pliku
- Zdjęcia do frontendu serwujemy za pomocą serwera LightHttpd i modułu do obsługi NKFS napisanego w C
- Upload zdjęć od frontendu przyjmowany jest również LightHttpd (WEBDAV) + moduł do obsługi NKFS w C
- Nasze testy wykazały około 30% mniejsze zapotrzebowanie na I/O rozwiązania NKFS w stosunku do tradycyjnego podejścia trzymania każdego zdjęcia w osobnym pliku w systemie XFS.
- W przypadku potrzeby przeniesienie strefy na inny węzeł przegrywamy po prostu jeden plik na inny serwer :-)

Agenda

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

Następne kroki

Następne kroki

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Squid nie skaluje się w systemach wielordzeniowych, grupa FORTICOM stworzyła własne oprogramowanie frontend w JAVA i jesteśmy już po etapie dostosowania tego oprogramowania do naszych wymagań i uruchamiania pierwszych testów produkcyjnych
- Komunikacja HTTP pomiędzy frontendem a backendem nie jest najbardziej optymalną, dlatego jako serwer backendowy użyjemy również oprogramowania FORTICOM napisanego w JAVA wykorzystującego własny autorski system komunikacji z kilkukrotnie mniejszym narzutem od HTTP
- Rozwiązaniem z MySQL do trzymania indeksu nie jest idealne dlatego zamiast obecnego rozwiązania (MySQL + plik z danymi) będziemy wykorzystywać BerkleyDB (indeks + dane), dodatkowo uzyskamy za darmo możliwość replikacji i to pozwoli zdjąć obowiązek z frontendu polegający na dostarczeniu zdjęcia w trzy miejsca węzła (dwie kopie produkcyjne + backup)
- JAVA doskonale skaluje się w środowiskach wielordzeniowych
- Intensywnie testujemy dyski SSD (na razie w macierzach fiberchannel) pod kątem zastosowania ich jako storage dla miniaturek zdjęć dzięki temu powinna pojawić się możliwość eliminacji warstwy keszującej. W przyszłości powinna się możliwość eliminacji macierzy dyskowych i zastąpienia ich dwoma dyskami SSD w serwerze (1TB SSD już jest !!!!!)

Następne kroki

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Nasze testy SSD w macierzach FC pokazują rezultaty na poziomie 35 000 IOPS na 12 dyskach !!!!!!!!!!!!
- Dyski SSD pobierają <1W natomiast dyski SAS/SATA >10W ! Ogromna ilość oszczędzonej energii pozwoli znacznie zmniejszyć TCO oraz gęstość upakowania w szafie (24 dyski w 3U instalowane od frontu)

Inne rozwiązania

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

- Haystack, Facebook http://www.facebook.com/note.php?note_id=76191543919
- MogileFS, Danga http://www.danga.com/mogilefs/

nasza-klasa.pl
PORTAL DLA LUDZI Z KLASĄ

Pytania ?