VIJETNAMSKI LINUX CLUSTER

Dinko Korunić (InfoMAR)

Što se babi snilo...

- naručitelj – Vijetnamska vlada & USAID
- svrha – dokumentiranje i javna dostupnost državnih reformi
- tipični zahtjevi.. idealni za Linux platformu ☺
  - visoka redundancija (hardverska i aplikativna)
  - centralizirani spremnik sadržaja
  - otpornost na DoS napade, ispade poslužitelja, itd.
  - nadogradnja/servisiranje poslužitelja bez utjecaja na ostatak sustava (frontend, backend, itd.)
  - konkurentna finalna cijena (hardver, postavljanje, održavanje, edukacija)
  - lako horizontalno skaliranje i proširenje kapaciteta
  - otvorenost i lako mijenjanje komponenti
- realizirana većina zahtjeva, projekt građen tijekom 3 mjeseca te approx utrošeno 1000 inženjer-sati

Čemu sve to?

- sadržaj:
  - stotine novinara unosilo dokumente svakodnevno
  - 110 GB statičkog sadržaja (dominantan!):
    - dokumenti (dosjei)
    - višerazinski međuspremnici samog CMS-a
  - 3+ milijuna datoteka, 1+ milijun direktorija
  - 300 tisuća objekata u samom CMS-u
  - dinamički sadržaj.. desetine milijuna redova u tablicama sa metadata opisima, ~30GB InnoDB
- CMS – eZ Publish, implementirao Netgen
- basics 101: broj datoteka/direktorija utiče drastično na performanse ovisno o datotečnom sustavu i raspoređenosti datoteka/direktorija...

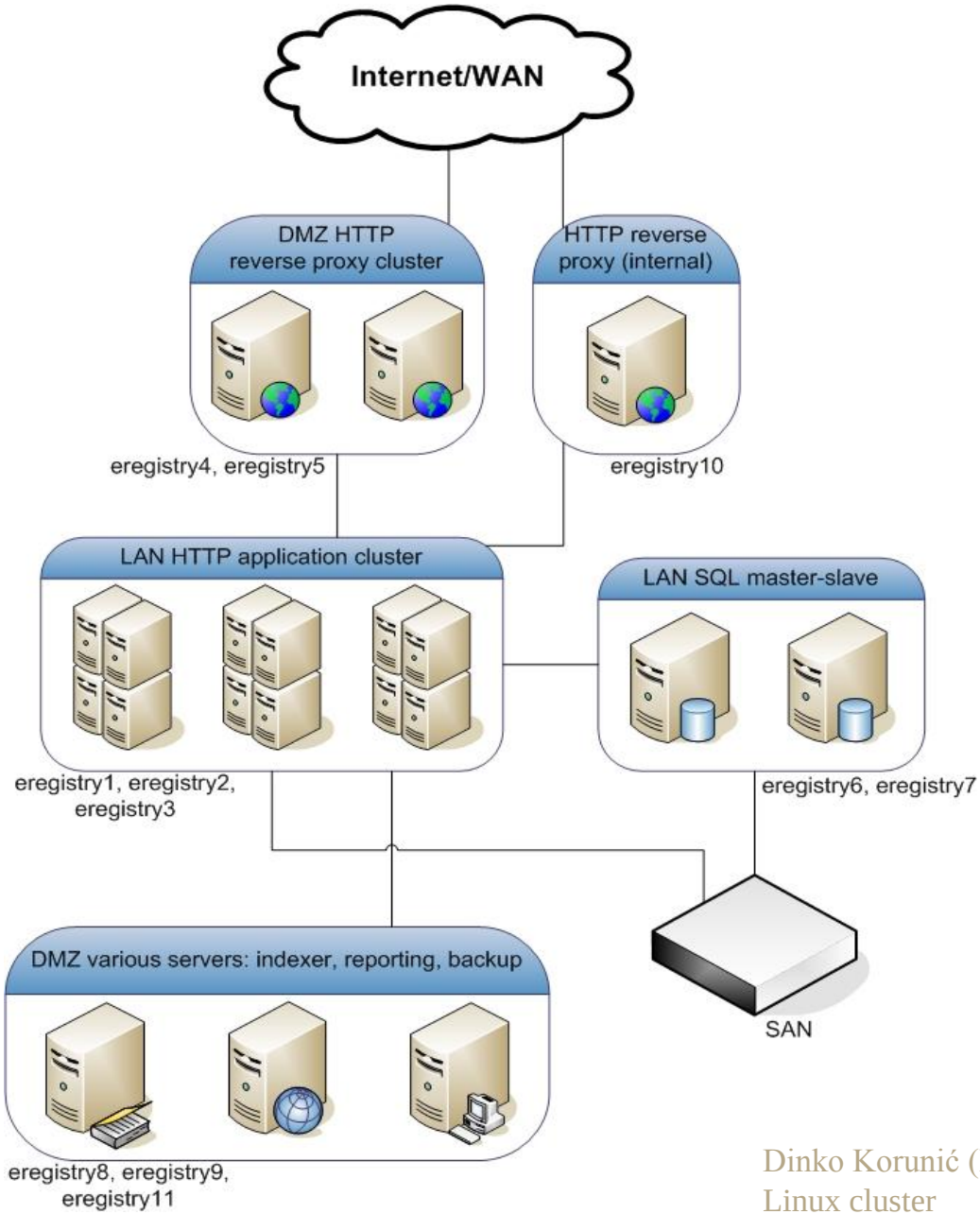
Mi bi brzo, jeftino i dobro...

- previše sadržaja za shared-nothing strojeve (i problem sinkronizacije) – rješenje FC ili iSCSI SAN + clusterfs
- iSCSI SAN:
  - niska inicijalna ulaganja (GE switchevi),
  - osrednje performanse: ~1000 IOPS max, realno 500 IOPS (RAID5 10krpm SAS)
  - RAID1 – MySQL baza
  - RAID5 – pretežno serviranje statičkog sadržaja (85%) i nešto povremenog pisanja; potrebno minimizirati (max IOPS)
  - proširenja/ubrzanja: više šasija (DS3300), flashcopies, kopiranje među šasijama, više kontrolera, više diskova pa transformacija u RAID10, itd.

Što sa sporim storageom...

- obzirom na osrednje SAN performanse, potrebno je...
- minimizirati pisanje:
  - ezmutex prosječno 40ak readova/lockova/pisanja po HTTP upitu: flock leak vodi do 10.000x usporenja!
  - umjesto cluster-wise flock() za sinkronizaciju koristiti Memcached semafore
- minimizirati broj HTTP upita koji pogađa backendove:
  - Varnish HTTP akcelerator/proxy
  - višerazinski međuspremnici samog CMS-a...
- problem: autenticirani korisnici i sessioni
  - Varnish za stripanje kolačića i keširanje svih statičkih te dijela dinamičkih (PDF) sadržaja
  - u prosjeku 80% hit ratio

Sistemska implementacija



Sistemska implementacija

- Linux za "krojeno" rješenje
- HTTP reverse proxy-cachevi (Varnish):
  - uobičajeni DNS RR + HA IP klaster: automatska migracija IP adresa, međusobni nadzor, nadzor nadležnog usmjernika
  - load-balanceri: slučajna raspodjela prema backendovima + kontinuirana provjera stanja backend poslužitelja
  - integrirani lokalni međuspremnici serviranog sadržaja
  - anti-DoS mogućnosti
- iSCSI (Open-iSCSI) prema centralnom SAN-u:
  - IBM DS3300 kao entry-level SAN iSCSI model
  - iSCSI + multipathing + OCFS2 (Web) / Ext3 (SQL)
  - očekivano 500-600 IOPS

Sistemska implementacija (2)

- Web aplikacijski poslužitelji:
  - cluster-aware datotečni sustav kroz postojeću infrastrukturu (iSCSI vs. FC)
  - OCFS2 za zajednički Web pool – samostalno rješenje bez dodatnih klaster-servisa
  - rezultat – međusobno nezavisni! (servisiranje, nadogradnje)
  - lako proširenje s novim poslužiteljima (online)
- SQL poslužitelji:
  - Ext3 zadovoljava performanse
  - ne pretjerano jednostavno proširenje, ali nije usko grlo
  - budućnost: multimaster, SQL proxy, IPVS, itd.
  - plan: ABA ciklička replikacija – propao zbog problema sa replikacijom

Sigurnost, nadzor, redundancija

- nadzor i alerting:
  - sigurnosni pregled (zaštitne sume, logovi): OSSEC HIDS
  - kontinuitet rada servisa (testiranje): Monit
  - centralno sistemsko i aplikativno nadziranje, reporting, praćenje performansi: ZenOSS
  - praćenje autentikacije: Fail2Ban
- redundancija:
  - 2x iSCSI portovi na poslužiteljima (multipath)
  - 2x LAN portovi na poslužiteljima (bonding)
  - 2x mrežni preklopnici u klaster načinu
  - 4x iSCSI portovi na SAN (multipath, 2x kontroler)
  - 2x mgmt portovi na SAN
  - te uobičajeno (2x PSU, LightPath dijagnostika, itd...)

Loša iskustva

- MySQL 5.0:
  - redovni SBR replication #FAIL
  - transakcija od mastera kolidira sa stanjem tablica na slaveu?! problem sa rollbackom na slaveu?
- MySQL 5.1:
  - OOPS na master serveru (i reboot): slave server cannot continue replication from impossible position...
  - konverzija sa 5.0 na 5.1 zahtijeva reimport…
  - nužno redovno pratiti replikaciju kroz ZenOSS
- Debian Lenny kerneli:
  - sporadični OOPS
  - nestabilni kernel moduli: GFS, GFS2
  - OCFS2 je OK, ali ne backportaju se patchevi...

Još loših iskustva...

- Debian... je bio loš izbor za enterprise cluster distribuciju

- Debian RHCS:
  - staro/neodržavano/nepodržano razvojno stablo - verzija 2
  - nema službenog aktivnog Debian maintainera
  - won't fix bugovi: redovni segmentation faultovi, ispadanje nodeova, freejanje memorijskih polja pa kasnije korištenje u ispisu(!), …
  - pad RHCS uzrokuje razne popratne pojave: nemogućnost umounta GFS, čistog reboota stroja...
  - tijekom preprodukcijskog testiranja prosječno 4 potpuna ispada u 7 dana

Još... radi li išta dobro?!

- GFS i GFS2:
  - spori, nestabilni, vrlo loše skaliranje s brojem datoteka i file-lockova
  - GFS2: OOPS na svim čvorovima odmah u prvih 5 sati preprodukcijskog testiranja
  - GFS: drastično se usporava se s vremenom, brojem file-lockova, brojem datoteka, uber-loše performanse u produkciji, Apache procesi redovno u D stanju
- OCFS2:
  - povremeni file-deletion / file-lock bugovi + flock usporenja
  - povremeni ali rijetki OOPSevi (1.5.0 verzija jedino)
  - povremeni problem sa ne-otpuštanjem heartbeata pri umountu (samo najnoviji kerneli...)
  - nema directory hashinga

EOF

- pitanja, komentari, sugestije, diskusija, flamewar ☺
- kontakt:
  - OS/security: Dinko Korunić dinko.korunic@infomar.hr
  - CMS: Ivo Lukač ivo@netgen.hr