Wysoka dostępność w systemie Linux

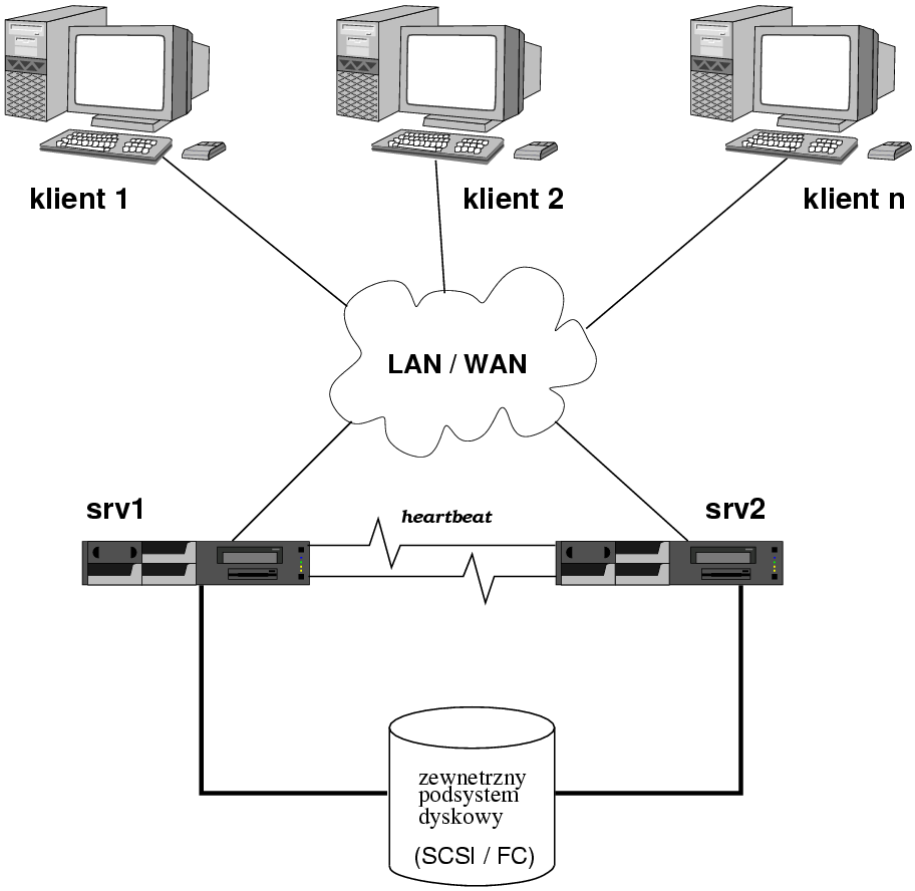
Marcin Owsiany

Zakres referatu

- HA pod Linuksem — klastry HA
- dostępne oprogramowanie usług uczestnictwa
- ,,HA solutions''
- współdzielenie danych

Klastry wysokiej dostępności

- Prawie wszystkie system HA na Linuksie to klastry wysokiej dostępności (koszty)



- Podsystemy klastra
  - Usługi uczestnictwa (ang. *membership*)
  - Usługi komunikacji
  - Zarządzanie klastrem (ang. *cluster management*)
  - Odgradzanie od zasobów (ang. *resource fencing*)
  - Monitorowanie zasobów
  - Współdzielenie/replikacja pamięci masowych

Failover — zagrożenia

- Mechanizm failover
  - wykrycie ,,odchodzącego'' węzła
  - przejęcie adresu IP
  - uzyskanie dostępu do współdzielonych danych (fsck ?)
  - uruchomienie usług
- Split-brain
  1. aktywny jest węzeł srv1
  2. srv1 ulega awarii (przestaje emitować puls i świadczyć usługi), ale nie umiera. System plików jest w nieokreślonym stanie (brudne bufory na srv1). Podobnie: przy utracie łączności.
  3. srv2 traci puls srv1, montuje dysk read/write, przejmuje funkcje węzła aktywnego
  4. srv1 nagle czuje się lepiej i zapisuje bufory na dysk
  5. Ooops!

Resource fencing

- Metoda quorum — nie zawsze możliwa i nie zawsze skuteczna
- Resource fencing == ,,Odgradzanie od zasobów'' — przed przejęciem roli aktywnego węzła uniemożliwiamy poprzedniemu korzystanie z zasobów
- Dwa rodzaje:
  - hard fencing (in.: resource based) — zatrzymanie I/O na żądanie. Wymaga wsparcia ze strony sprzętu: np. SCSI reserve/release
  - STONITH (Shoot The Other Node In The Head) — twardy reset. Proste, tanie, uniwersalne.

Współdzielenie danych

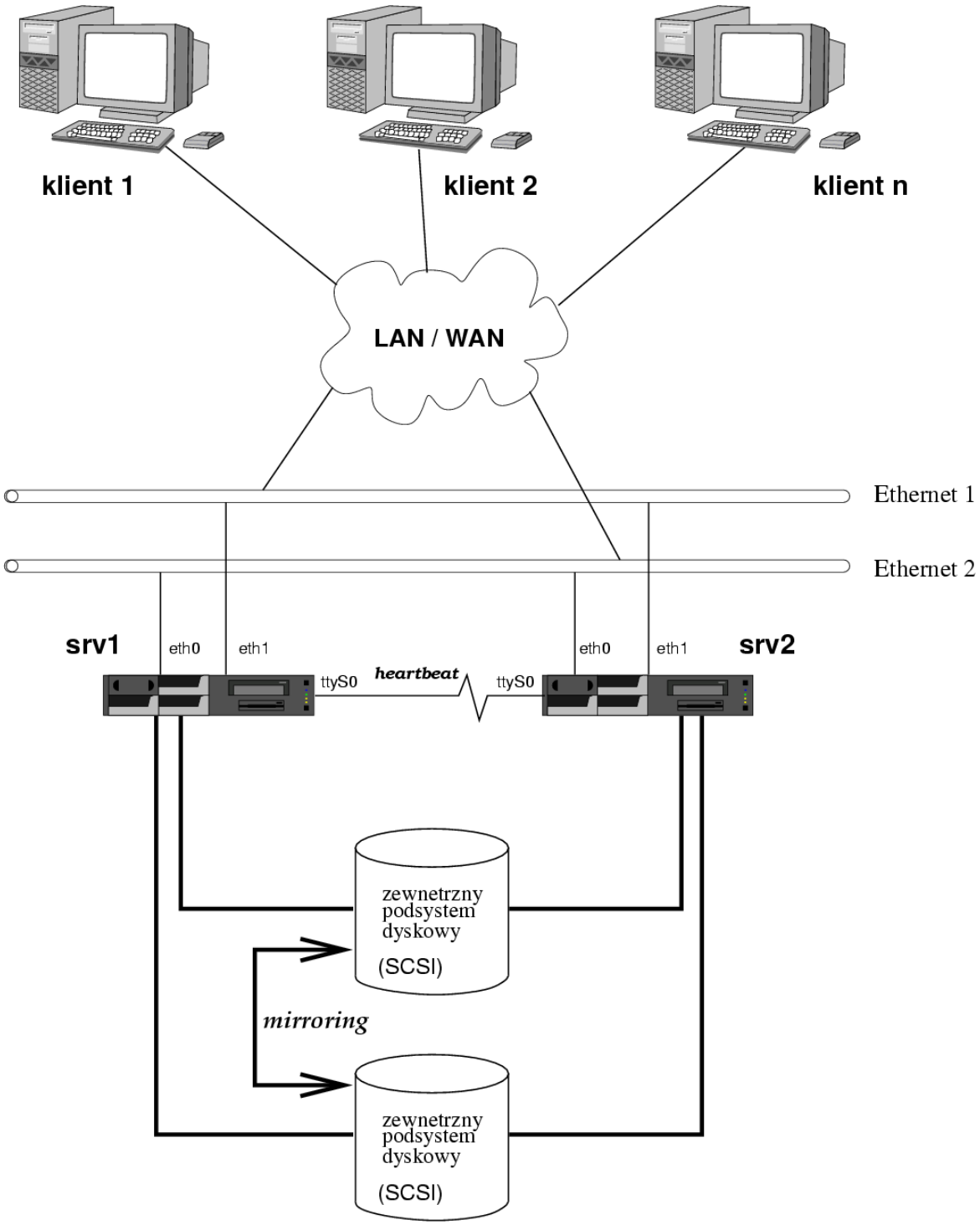
Kilka sposobów (w zależności od charakteru usług i sposobu modyfikacji):

- wsparcie sprzętowe — zewnętrzne macierze SCSI/Fibre Channel
- automatyczna replikacja danych:
  - ciągła, jaką oferuje na przykład DRBD, albo NBD w połączeniu z jednym z systemów: RAID (moduł md), LVM lub EVMS
  - okresowa — wykonywana co pewien czas
  - ,,na żądanie'' — na przykład po aktualizacji stron WWW są one rozsyłane automatycznie na innych węzłach. (rsync?)
  - wykorzystująca własne mechanizmy replikacji danej aplikacji czy usługi — replikacja MySQL, OpenLDAP, DNS, NIS itp. . .
- sieciowe systemy plików — NFS, codafs czy InterMezzo

Kwestia systemu plików — najlepiej klastrowy, ewentualnie przynajmniej z journalem

Linux-HA http://linux-ha.org/: heartbeat

Implementacja usług uczestnictwa i komunikacji

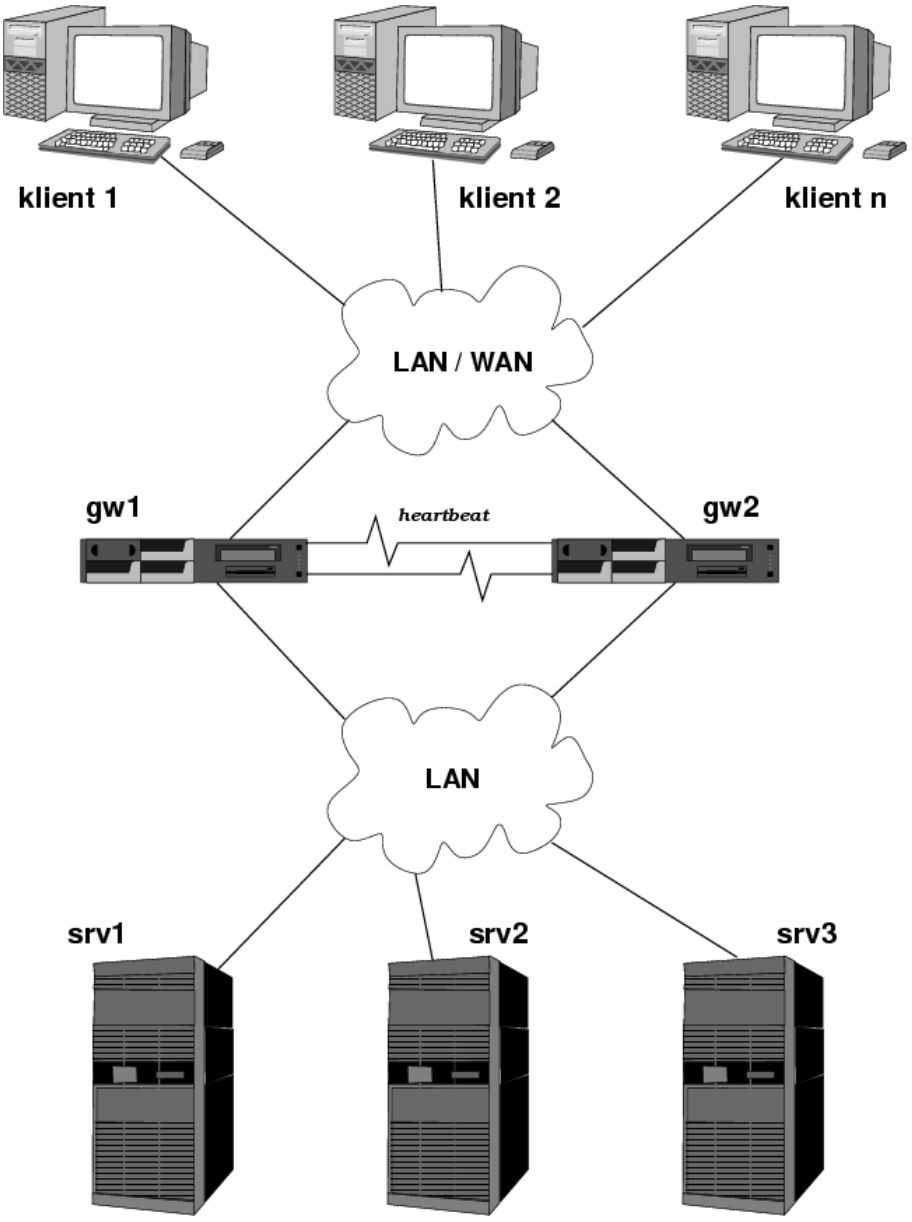



```
srv1 1.2.3.4 Filesystem::/dev/sda1::/data::ext2 \
apache samba
```

Linux Virtual Server

LVS (`http://www.LinuxVirtualServer.org/`) kładzie nacisk przede wszystkim na HPC, ale daje też HA

- redundantny director — eliminacja SPOF
- `mon` lub `ldirectord` lub `keepalived` — monitorowanie usług i modyfikacja tablic routingu



,,Gotowe" rozwiązania

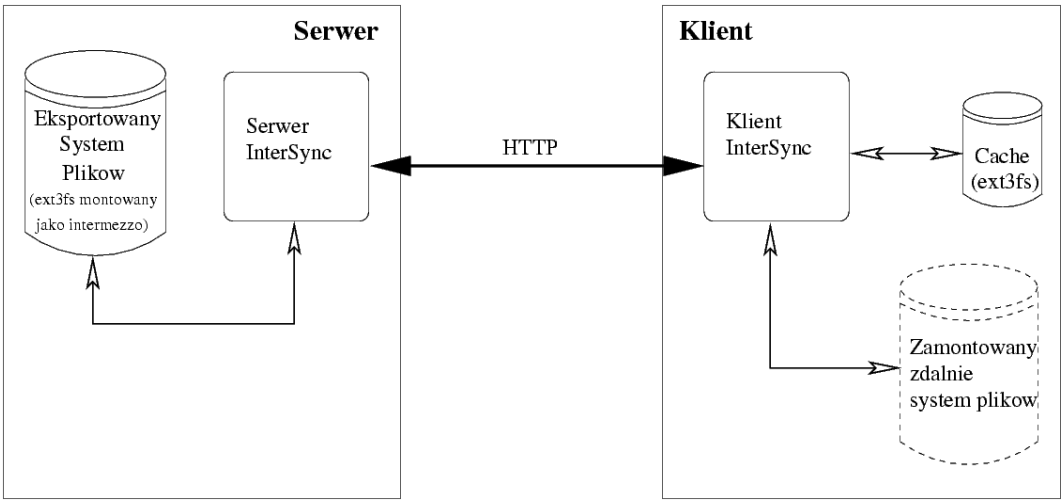
- Red Hat, Inc. — Piranha (LVS + nanny + pulse + GUI)
- TurboLinux — TurboCluster
- UltraMonkey — LVS + heartbeat + ldirectord
- Mission critical Linux — Convolo cluster — NetGuard edition, Kimberlite
- Motorola, Inc — Advanced HA: próba uzyskania 6NINES (30 sek downtime/rok)
  - platforma MXP: magistrala PICMG 2.16 + serwery ,,blade"
  - ,,Managed object model" + ,,Distributed event management"
  - Distributed Inter-System Communication System (DISCS)
  - Checkpoint services
- Hewlett-Packard — HP Multi-Computer/Serviceguard

Systemy plików

- Global FileSystem firmy Sistina — klastrowy system plików (dostęp R/W z wielu węzłów)
  - bezpośrednie I/O — nadaje się dla zaawansowanych RDBMS
  - dobra skalowalność: 4x lepsza (?)
  - architektura OmiLock — wydajne, skalowalne blokowanie, wybór mechanizmu blokowania
  - quoty — miękkie i twarde
  - journaling
  - rozproszone metadane
  - zgodność z POSIX
- Coda — NFS on steroids
  - praca bez połączenia z siecią
  - wolnodostępny — liberalna licencja
  - cache po stronie klienta
  - replikacja po stronie serwera
  - model bezpieczeństwa umożliwiający autoryzację, szyfrowanie i kontrolę dostępu
  - umożliwia dalszą pracę nawet w przypadku uszkodzenia części sieci serwerów
  - adaptacja do przepustowości sieci
  - dobra skalowalność
  - dobrze zdefiniowana semantyka, nawet w przypadku uszkodzenia sieci

InterMezzo

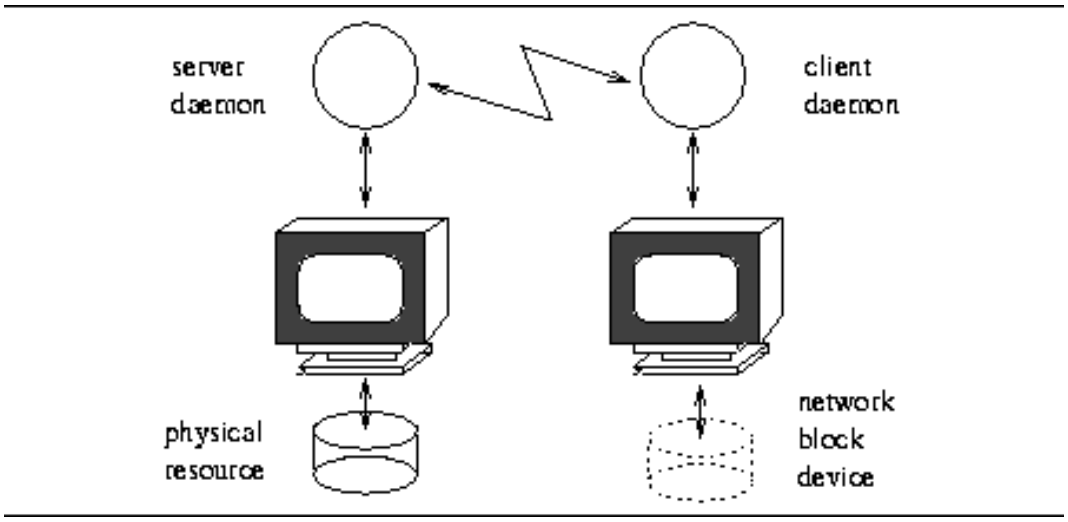
- zainspirowany przez codafs, ale zupełnie przeprojektowany i zaimplementowany od zera
- przeznaczony specjalnie dla HA, nadaje się do replikacji serwerów, komputerów przenośnych, zarządzania oprogramowaniem systemowym na dużych klastrach itp. . .
- implementowany jako ,,nakładka'' na ,,prawdziwy'' system plików



- użycie protokołu HTTP ma na celu uzyskiwanie rozszerzeń ,,za darmo'' (np. SSL, autoryzacja)

Network Block Device

- NBD (Pavel Machek) i ENBD (Peter Breuer)
- moduł jądra + serwer umożliwające wykorzystanie połączenia TCP jako urządzenia blokowego (`/dev/nd0`). Urządzenie takie można traktować jako zwykłą partycję, (RAID, gdy po drugiej stronie jest to partycja)



- niskopoziomowy — na partycji zamontowanej przez NBD można umieścić dowolny system plików
- z drugiej strony, w danej chwili tylko jeden klient może bezpiecznie montować system plików NBD do zapisu i odczytu.
- DRBD — podobna implementacja, ale służy do replikacji