Loadbalancing a cachování v linuxu

Petr Medonos, Lukáš Heřbolt

O nás

Petr Medonos

- Bc. ININ VŠCHT
- 5 let v ETN
- RHCE
- databáze, performance, bezpečnost, návrh OA

- Datart, TO2, Fast, Allianz, Eagri, ...

O nás

Lukáš Heřbolt

- Bc. FEL ČVUT
- 1 rok v ETN
- 2 roky na KPGI FEL CVUT

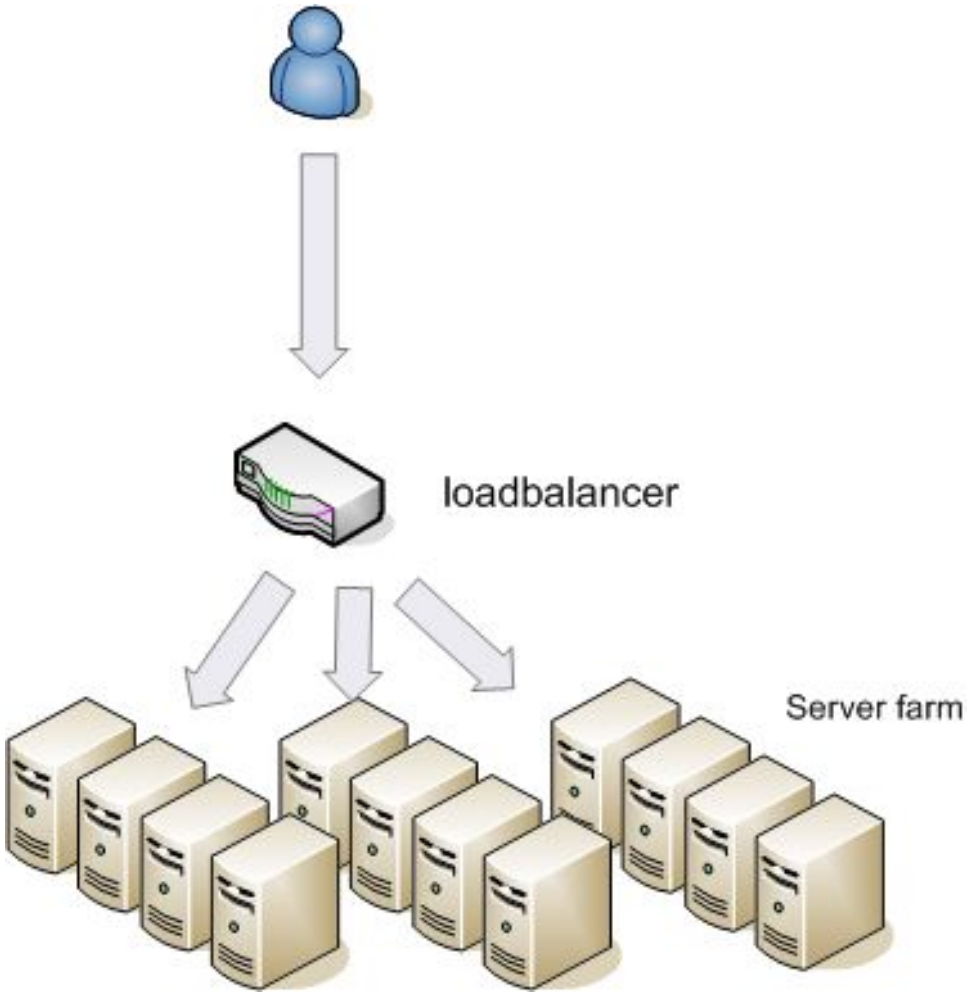
- TO2 - (extravyhody.cz, firemnitelefony.cz), Fast, mojeallianz.cz, moje.partners.cz,

O čem budem povídat?

- základní pojmy
- možnosti v linuxu
- EWA
- výkon

Základní pojmy

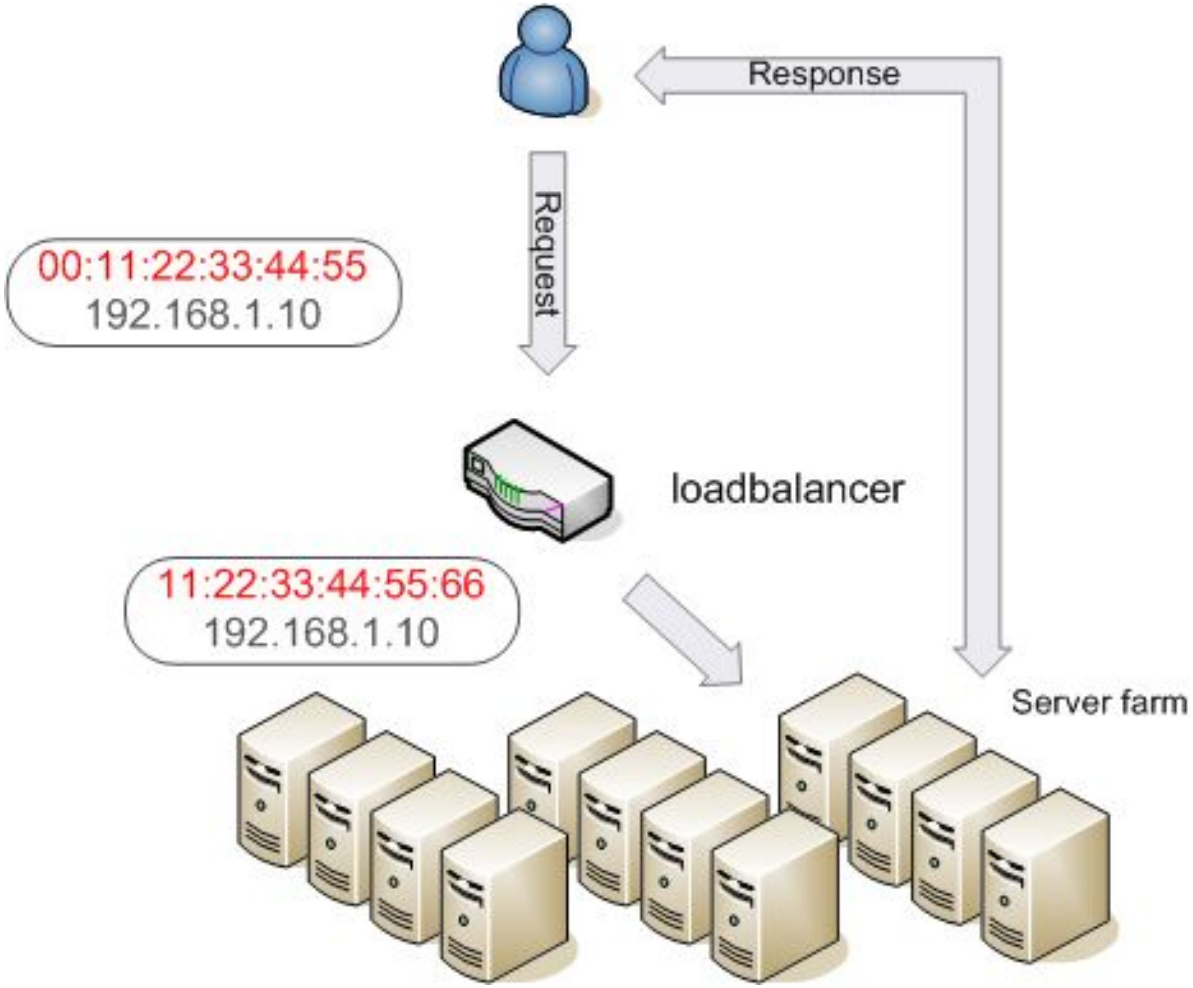
Co to je loadbalancer?



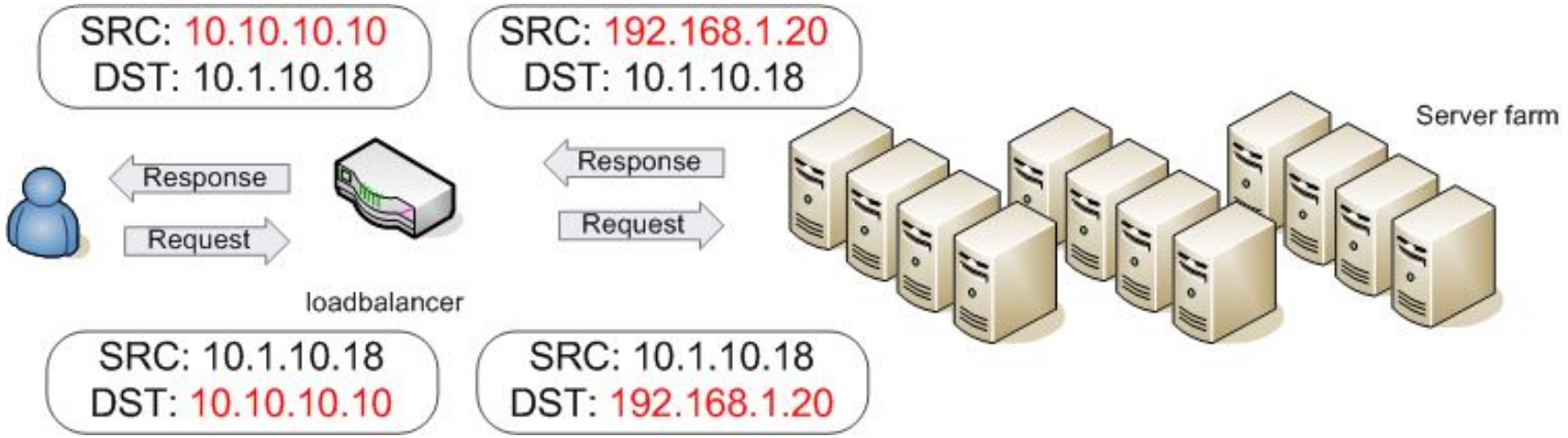
Módy zapojení

- Direct routing mód (L4)
- Routed mód (L4)
- Proxy (L7)

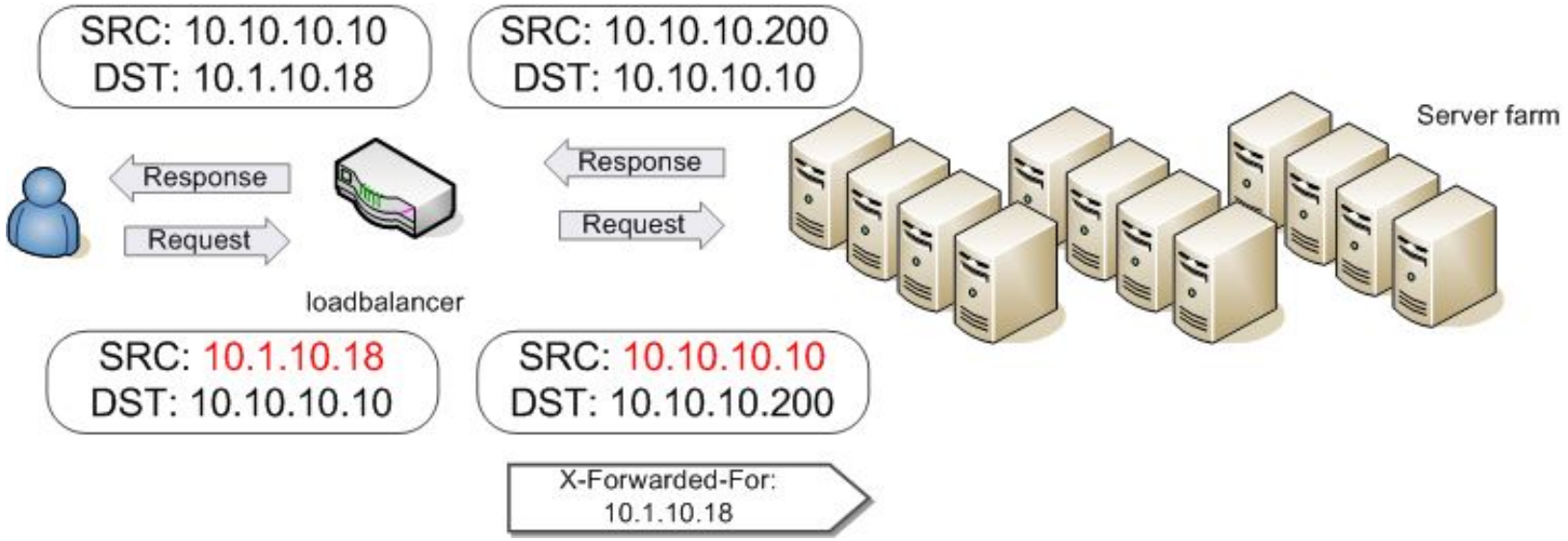
Direct routing



Routed mód

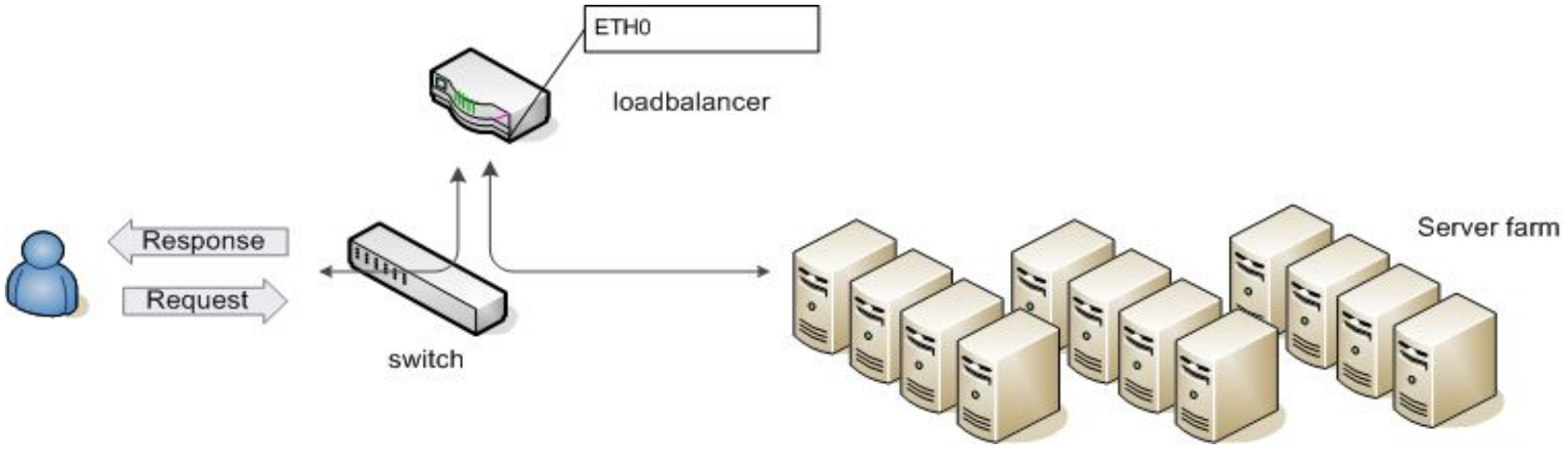

SRC: 10.10.10.10
DST: 10.1.10.18
SRC: 192.168.1.20
DST: 10.1.10.18
Server farm
Response
Request
Response
Request
loadbalancer
SRC: 10.1.10.18
DST: 10.10.10.10
SRC: 10.1.10.18
DST: 192.168.1.20

Proxy

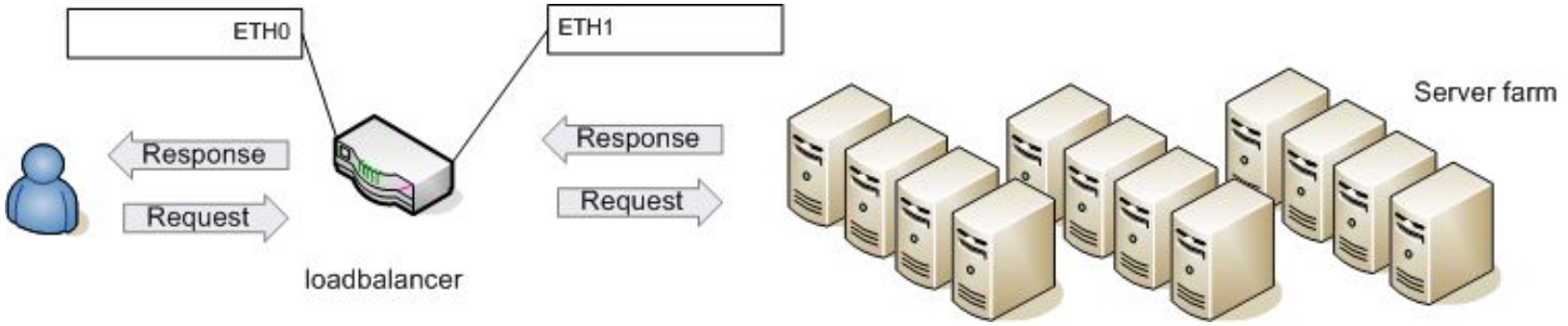

SRC: 10.10.10.10
DST: 10.1.10.18
SRC: 10.10.10.200
DST: 10.10.10.10
Server farm
Response
Request
Response
Request
loadbalancer
SRC: 10.1.10.18
DST: 10.10.10.10
SRC: 10.10.10.10
DST: 10.10.10.200
X-Forwarded-For:
10.1.10.18

Zapojení v síti

- one arm



- two arm



Co dnes musí umět loadbalancer?

- balancovat
- SSL terminace
- cachování a komprese obsahu
- snížení počtu spojení
- pokrývání výpadků
- L7 QoS a řazení do fronty
- HA
- jiné protokoly než http
- IPv6
- bezpečnost
- ...

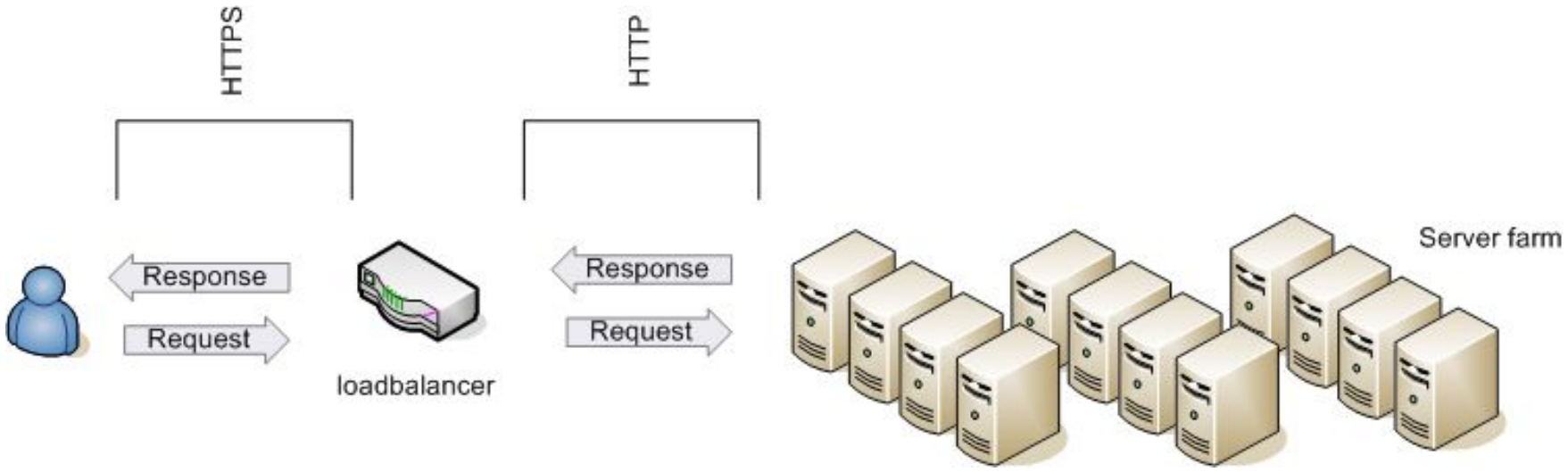
Loadbalancing - algoritmy

- round robin
- source IP
- least connection
- weighted [round robin|least connection|source IP]
- first
- random
- ...

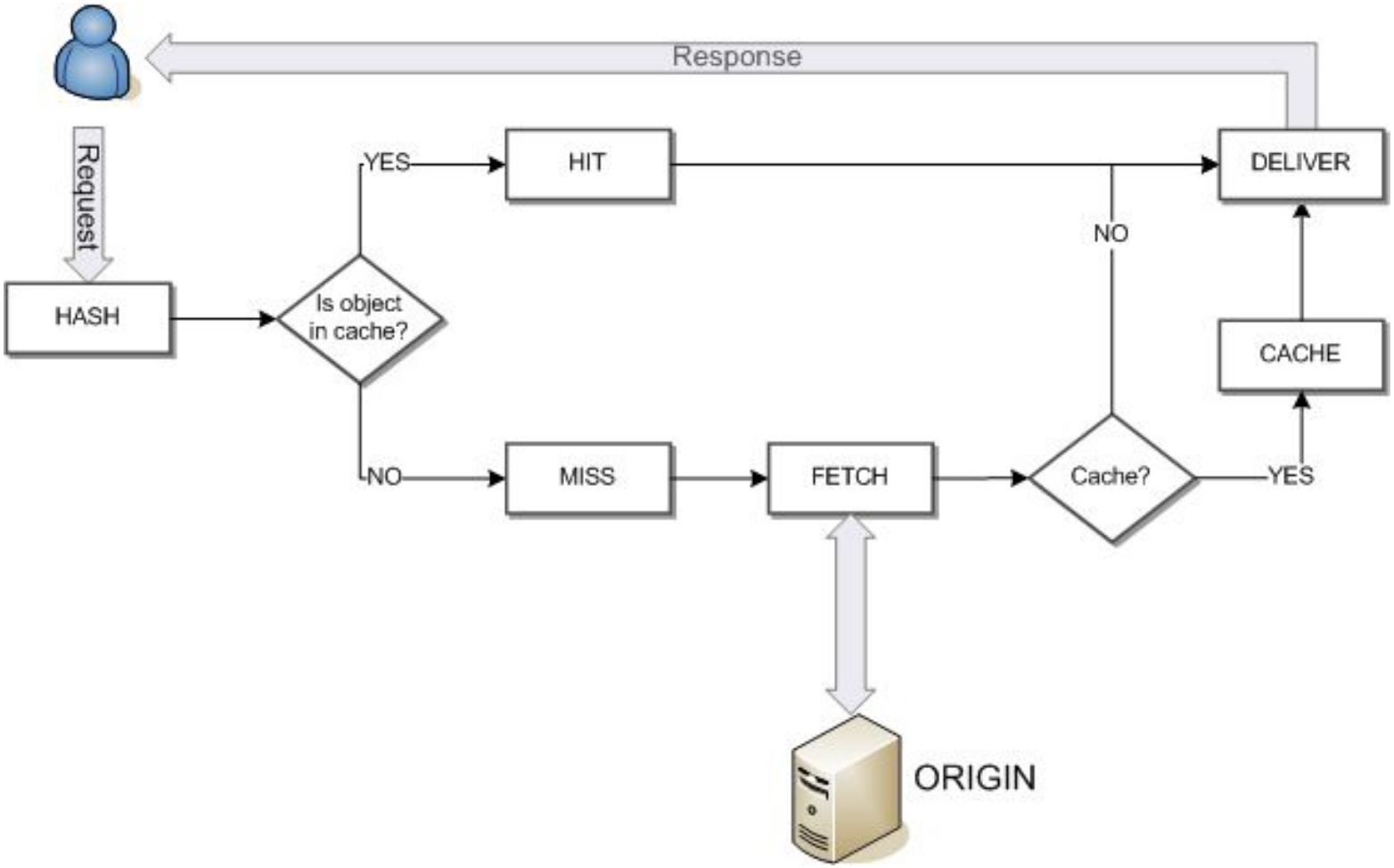
Perzistence

- je možné poslat uživatele na jakýkoli server?
- source IP
- sticky session

SSL terminace



Cachování


Response
Request
HASH
Is object in cache?
YES
HIT
DELIVER
NO
CACHE
NO
MISS
FETCH
Cache?
YES
ORIGIN

Cachování

- cachovatelný obsah
- HIT rate
- platnost objektu (TTL)
  - hlavičky `expires` nebo `Cache-Control: public, max-age=600`
  - nastavení vlastního TTL
- snížení počtu spojení

Pokrývání výpadků

- monitoring serverů
- reakce na výpadek
- pokrytí z cache

Možnosti v linuxu

Software pro balancer v linuxu

- **Balancing**
  - HAProxy
  - Pound
  - IPVS
  - Nginx
- **Cache**
  - Varnish
  - Apache Traffic Server (ATS)
  - Nginx
- **HA**
  - Keepalived
  - Pacemaker
- **SSL terminace**
  - Nginx
  - Stunnel
  - Apache

LB x CACHE x SSL TERMINATION

	ATS	HAProxy	NGINX	Squid	Varnish	Apache
Worker threads	Y	N	Y	N	Y	Y
Multi-process	N	Y	Y	Y	N	Y
Plugin APIs	Y	N	Y	part	Y	Y
Forward Proxy	Y	N	N	Y	N	Y
Reverse Proxy	Y	Y	Y	Y	Y	Y
Load Balancer	weak	Y	weak	weak	weak	weak
Persistent Cache	Y	N	Y	Y	Y*	Y
ESI	Y	N	N	Y	Y	N
ICP	Y	N	N	Y	N***	N
Keep-Alive	Y	Y	Y	Y	Y	Y
SSL	Y	Y**	Y	Y	N	Y

* experimental
** development 1.5dev17
*** lze nastavit přes VCL

LoadBalancing

	HAProxy	IPVS(LVS)	NGINX	Pound	KTCPVS	Pen
L4	Y	Y	N	N	N	N
L7	Y	N	Y	Y	Y	Y
SPLICE	>=1.4	?	N	N	?	N
SSL	1.5	N	Y	Y	N	N

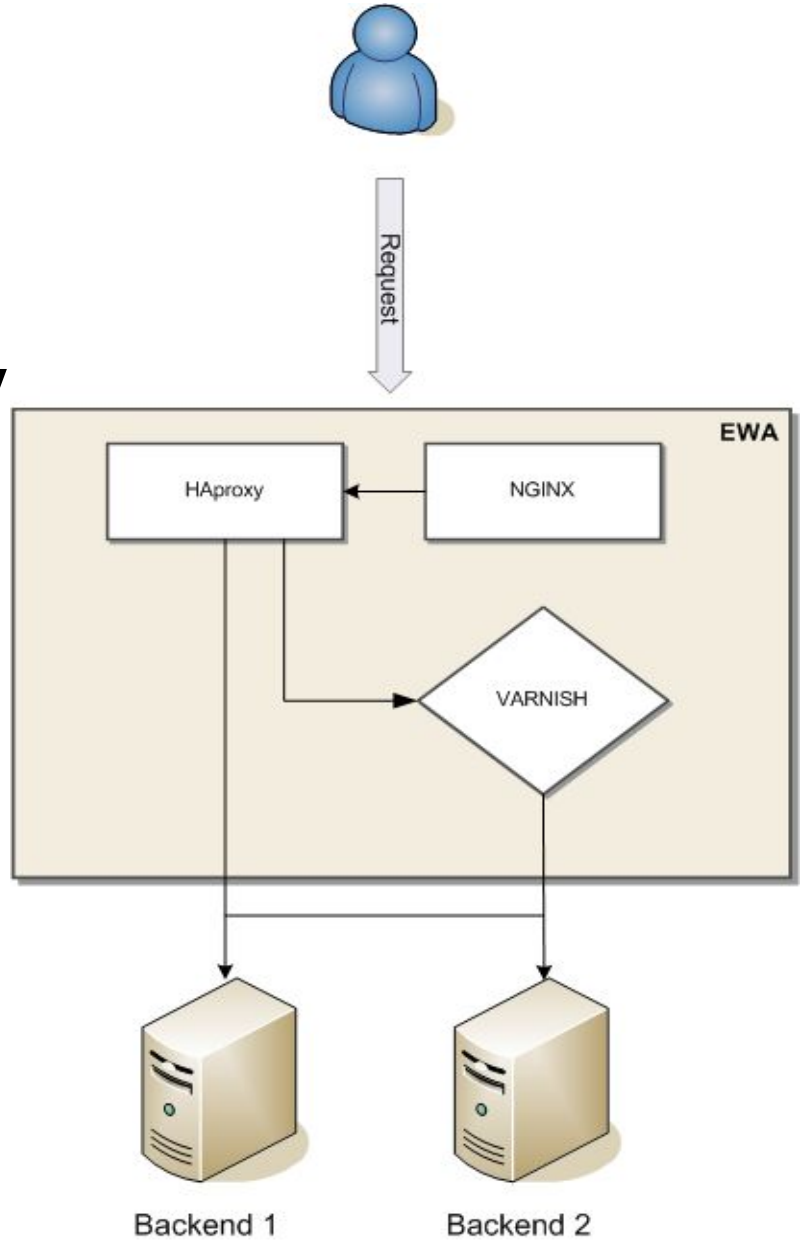
EWA (Etnetera Web Akcelerátor)

Hardware

- Fujitsu PRIMERGY CX400, CX270
- Multi-Node server
  - 2 (až 4) servery v jednom 2U chassis
  - snazší instalace

Software

- loadbalancing: **haproxy**
- ssl terminace: **nginx**
- cachování: **varnish**
- HA: **keepalived**



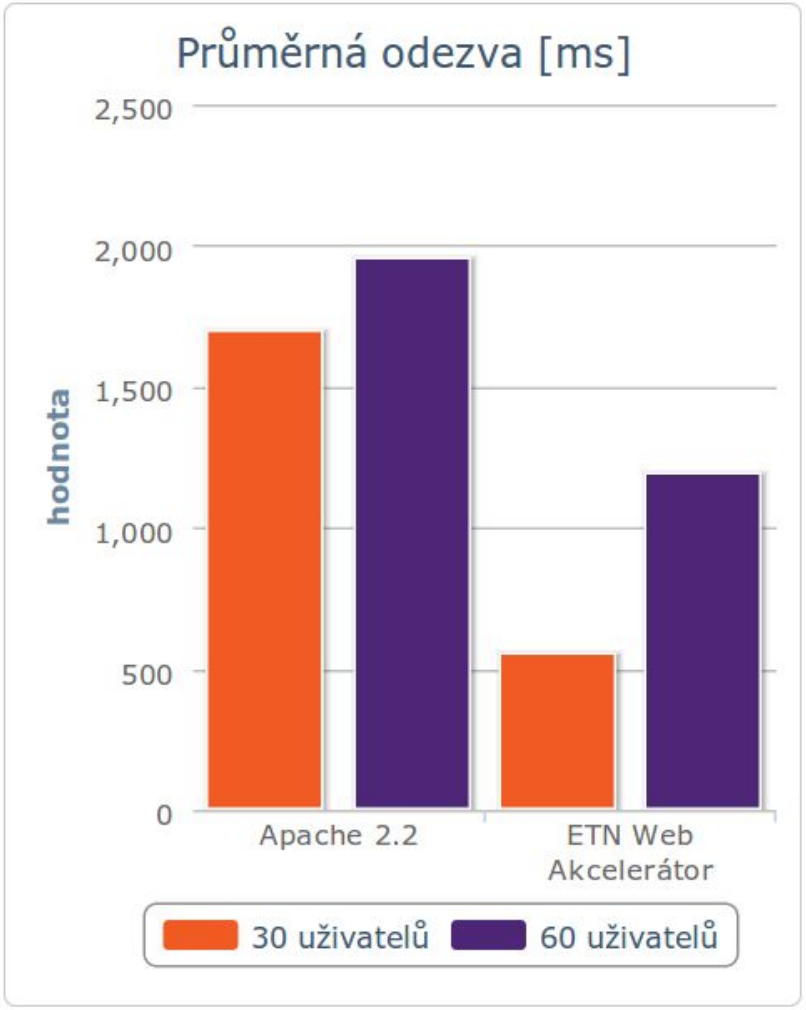
Konfigurační podvozek

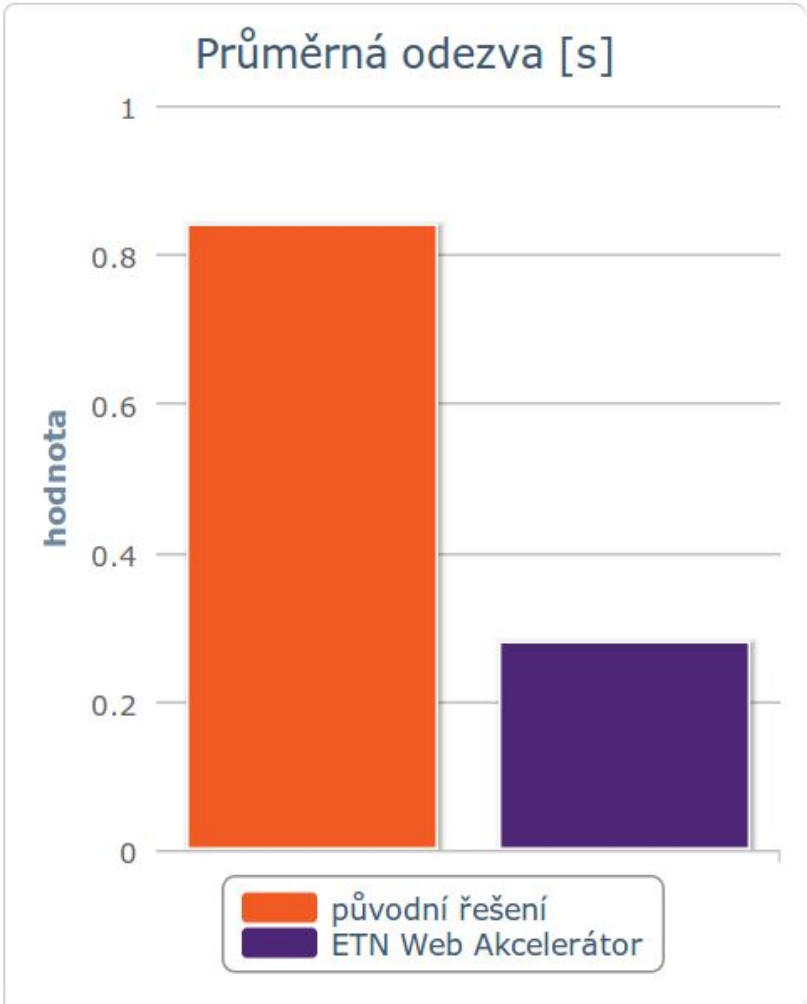
- **PUPPET!**
- dva "přepínatelné" módy:
  - standalone - vlastní lokální puppetmaster
  - zapojené do centrálního puppetmasteru

- **CLI rozhraní -** interaguje s puppetem
- **GUI rozhraní**
  - používá CLI, integruje statistiky
  - wizardy pro nastavení běžných věcí

Výsledky

http://ewa.etnetera.cz





Výkon

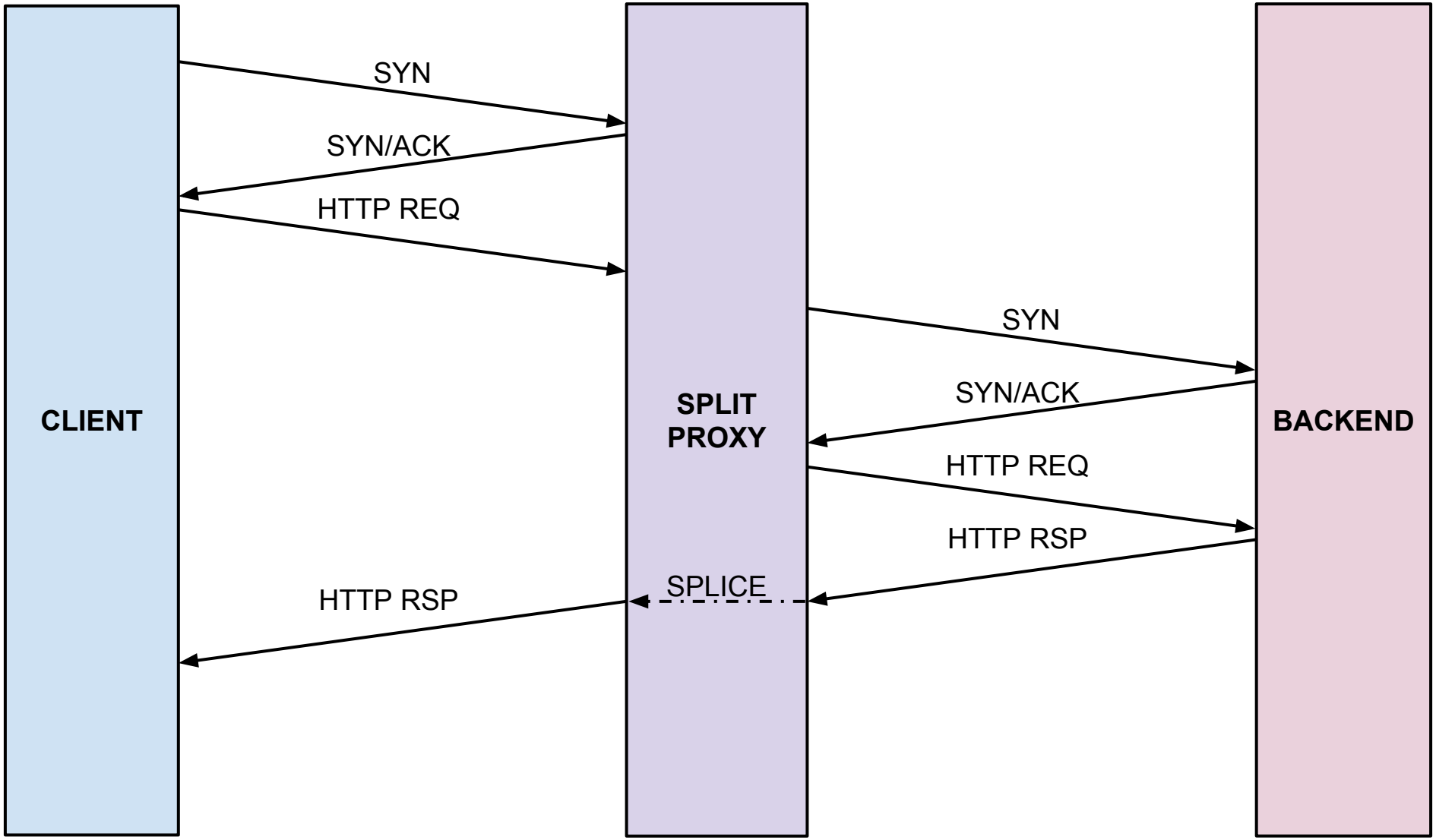
Výkon

- syn-cookie
  - `net.ipv4.tcp_syncookies = 1`
- local port range
  - normálně: 32768 - 61000
  - `net.ipv4.ip_local_port_range = 1025 65535`
- zvětšení conntrack tabulky
  - `net.netfilter.nf_conntrack_max = 2000000`
  - `net.nf_conntrack_max = 2000000`
  - velikost conntrack bucketu
    - `echo 500000 > /sys/module/nf_conntrack/parameters/hashsize`

Výkon

- TCP congestion control
  - `net.ipv4.tcp_congestion_control`
  - `net.ipv4.tcp_available_congestion_control`
    - vegas
    - bic
    - cubic
    - reno
    - westwood
- splice()
  - originally: 2.6.17 (buggy)
  - reimplemented: 3.5

Výkon - splice()



Výkon - budoucnost

- TCP Fast Open (TFO)
  - client-side 3.6
  - server-side 3.7
- TCP friends (patch from google to 3.6.1)
- TCP Cookie Transactions
  - syn-flood protection
  - rfc 6013 (experimental)

Q&A?

petr.medonos@etnetera.cz

lukas.herbolt@etnetera.cz