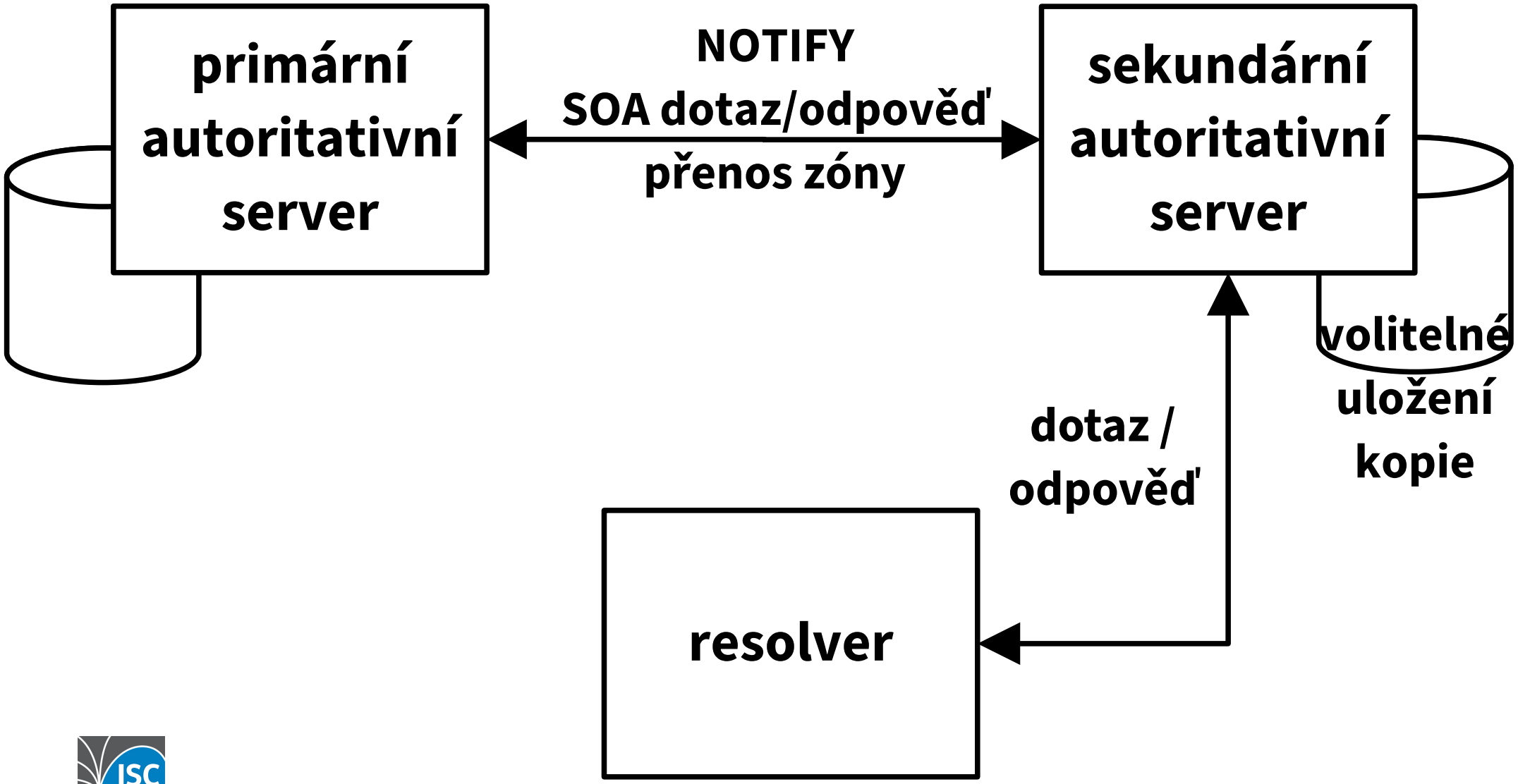
Měření výkonu přenosu DNS zón

Petr Špaček

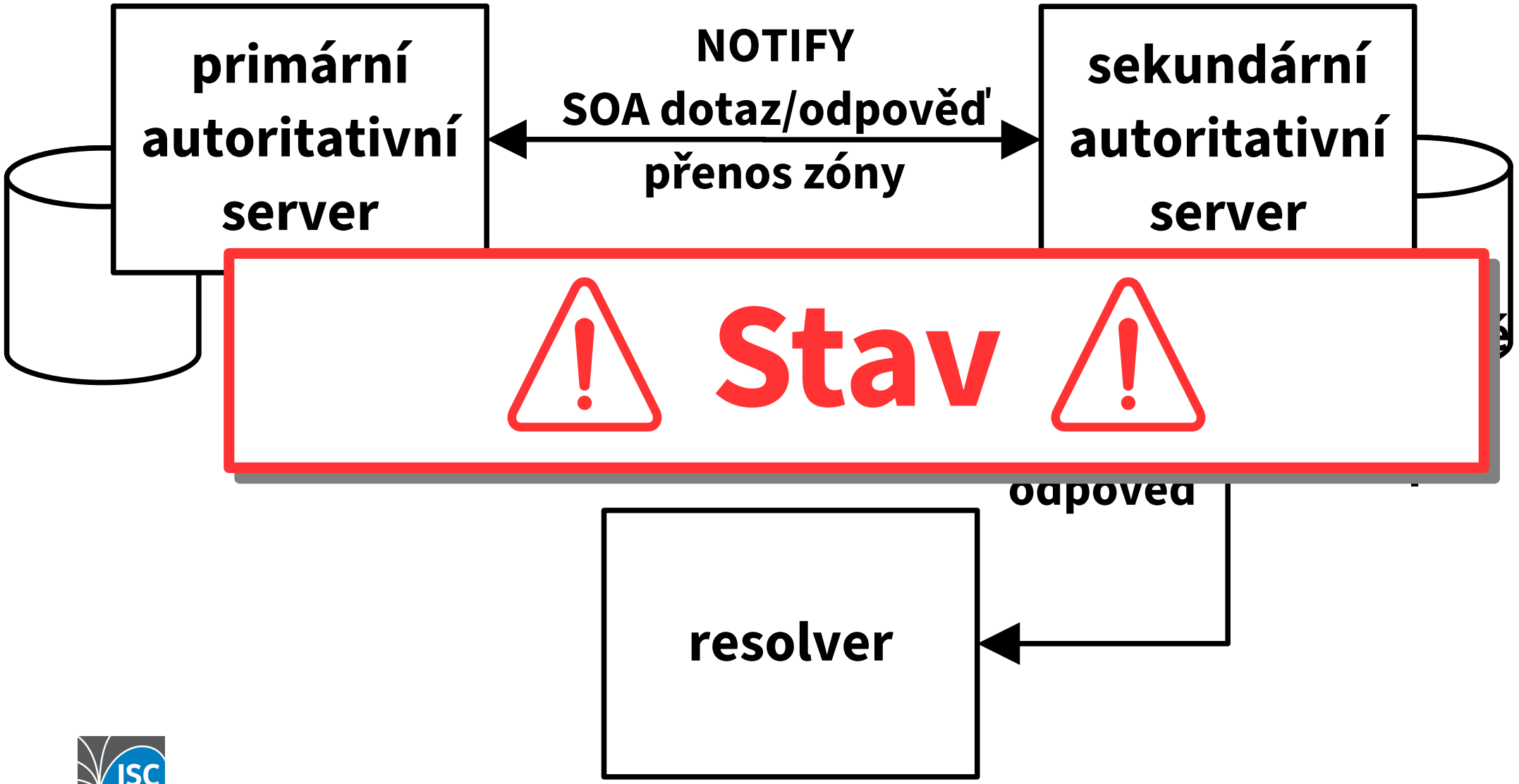
2025-01-21

pspacek@isc.org

XFR vs. DNS dotaz



XFR vs. DNS dotaz



Měříme …

- Max QPS / spotřebu paměti
  - v klidu / při změnách
- poměr primárních : sekunárních serverů
- latenci odpovědí
- čas propagace změn (SLA?)
- čas na start serveru
- …

Past vedle pasti

- **Kontrola testovacího prostředí!**
- Viz CSNOG 2024
  - Smysluplné měření kapacity DNS serverů
- "Echo server" na přenos zóny nestačí
- Konfigurace, logy, formáty – chybí standardy
- Stav …
- TSIG, TCP/TLS/…

Nejmenší zóna

Čas se limitně se blíží k nule … doufám

Nejmenší zóna

- SOA RR + 1 x NS RR
- prodleva přenosu zóny =
  (čas načtení zóny) – (čas startu serveru)

Nejmenší zóna

- SOA RR + 1 x NS RR
- prodleva přenosu zóny =
  (čas načtení zóny) – (čas startu serveru)
- prodleva = **-0,354 sekundy**

Nejmenší zóna

- SOA RR + 1 x NS RR
- prodleva přenosu zóny =

⚠ Přesnost údajů ⚠

- načtení 2024-10-14T12:40:32Z
- start 2024-10-14T12:40:32.646Z

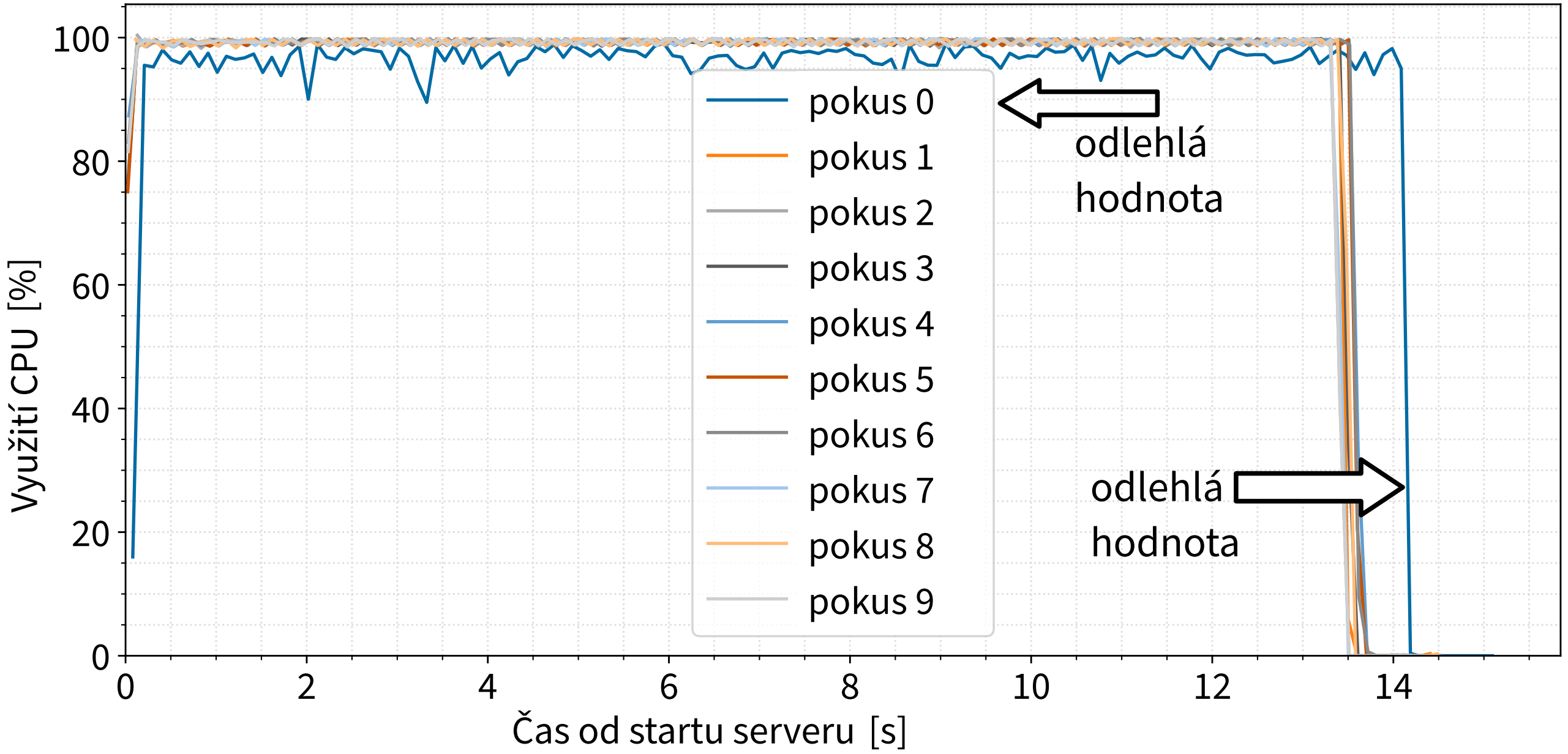
TLD

Pár vteřin? Deset? Dvacet??

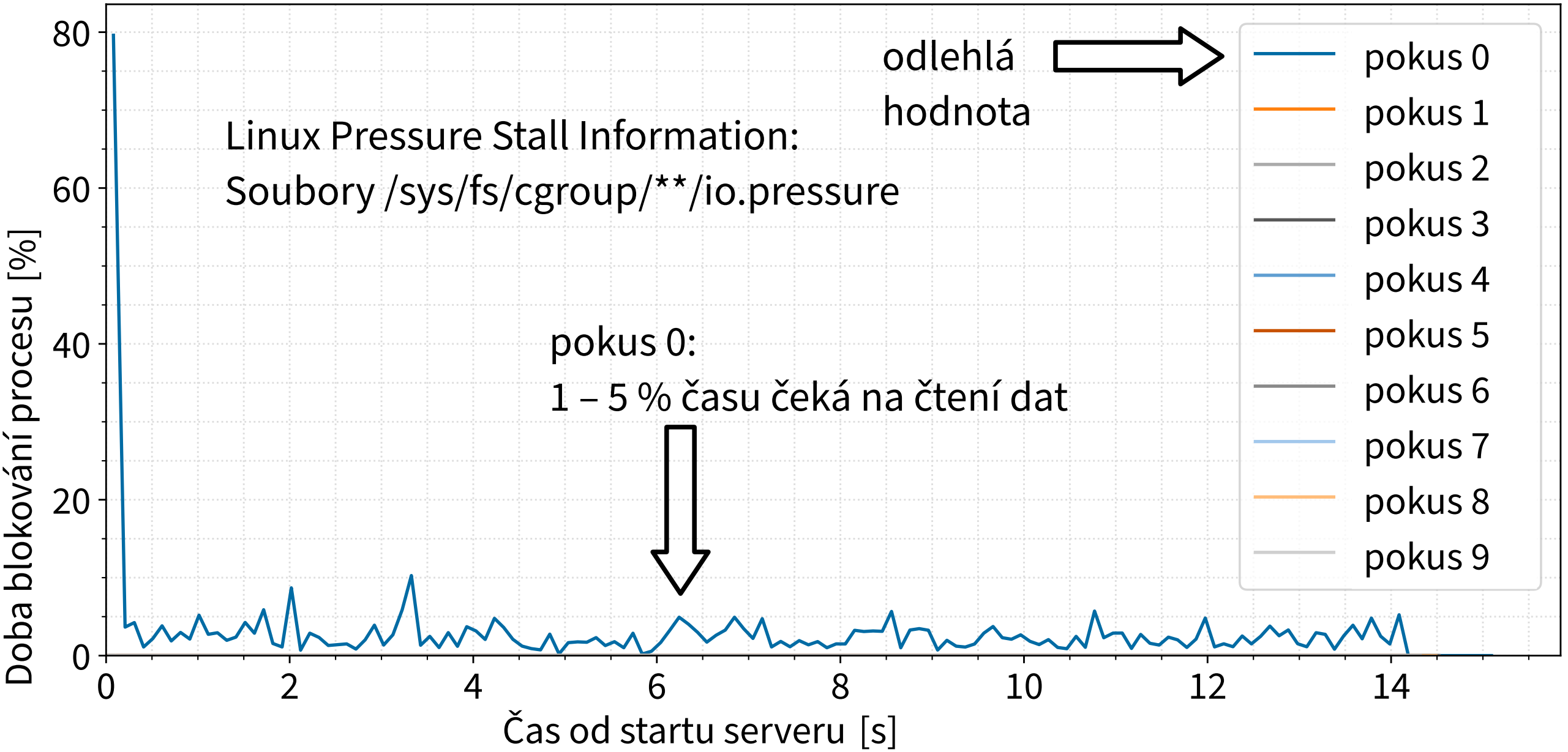
TLD AXFR

- ch. TLD
- 14 mil. záznamů
- 691 MB
- 47 700 DNS zpráv
- ~ 26,5 sekund přes UDP+TCP
  - ± 0,6 %, 10 opakování

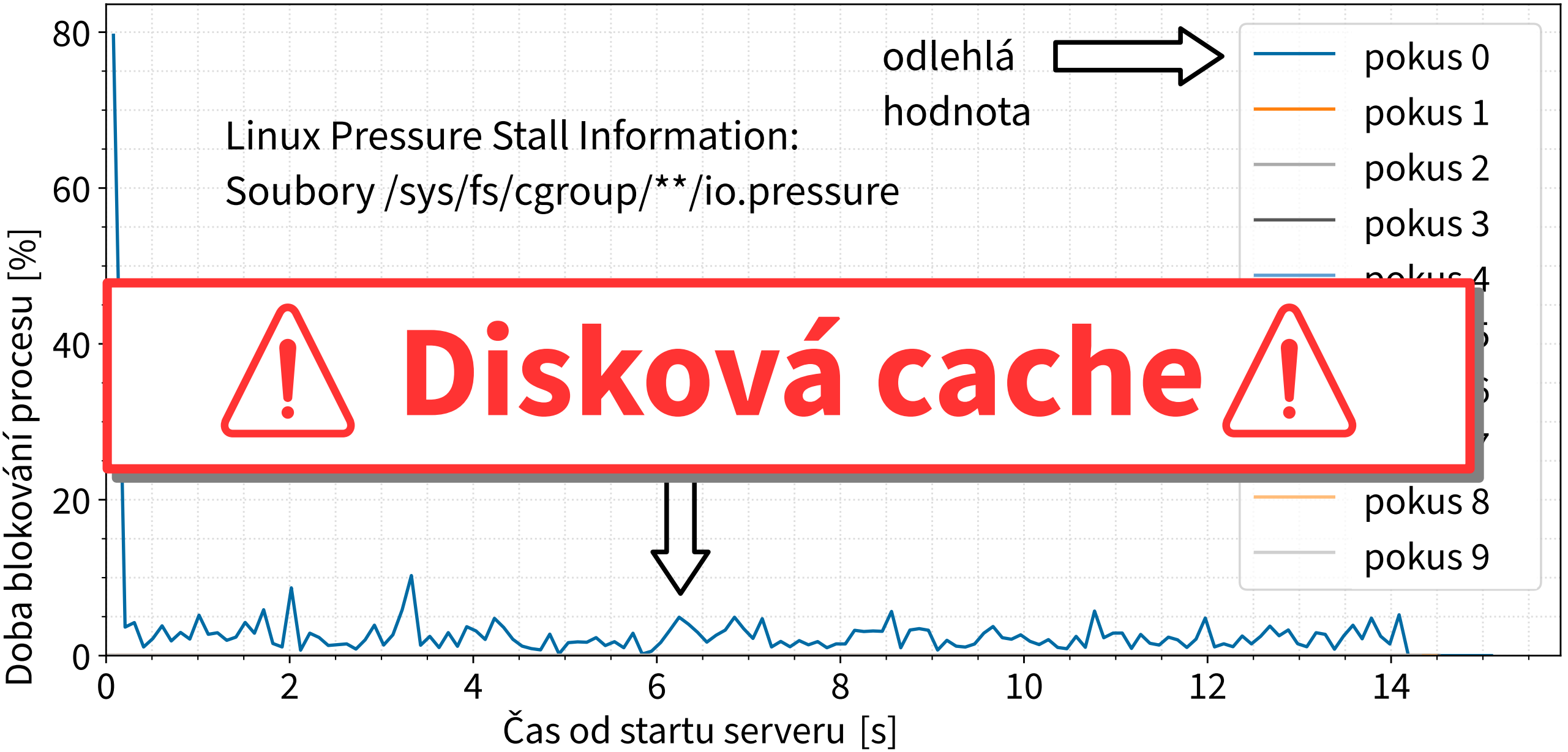
TLD, primární server, načítání, CPU



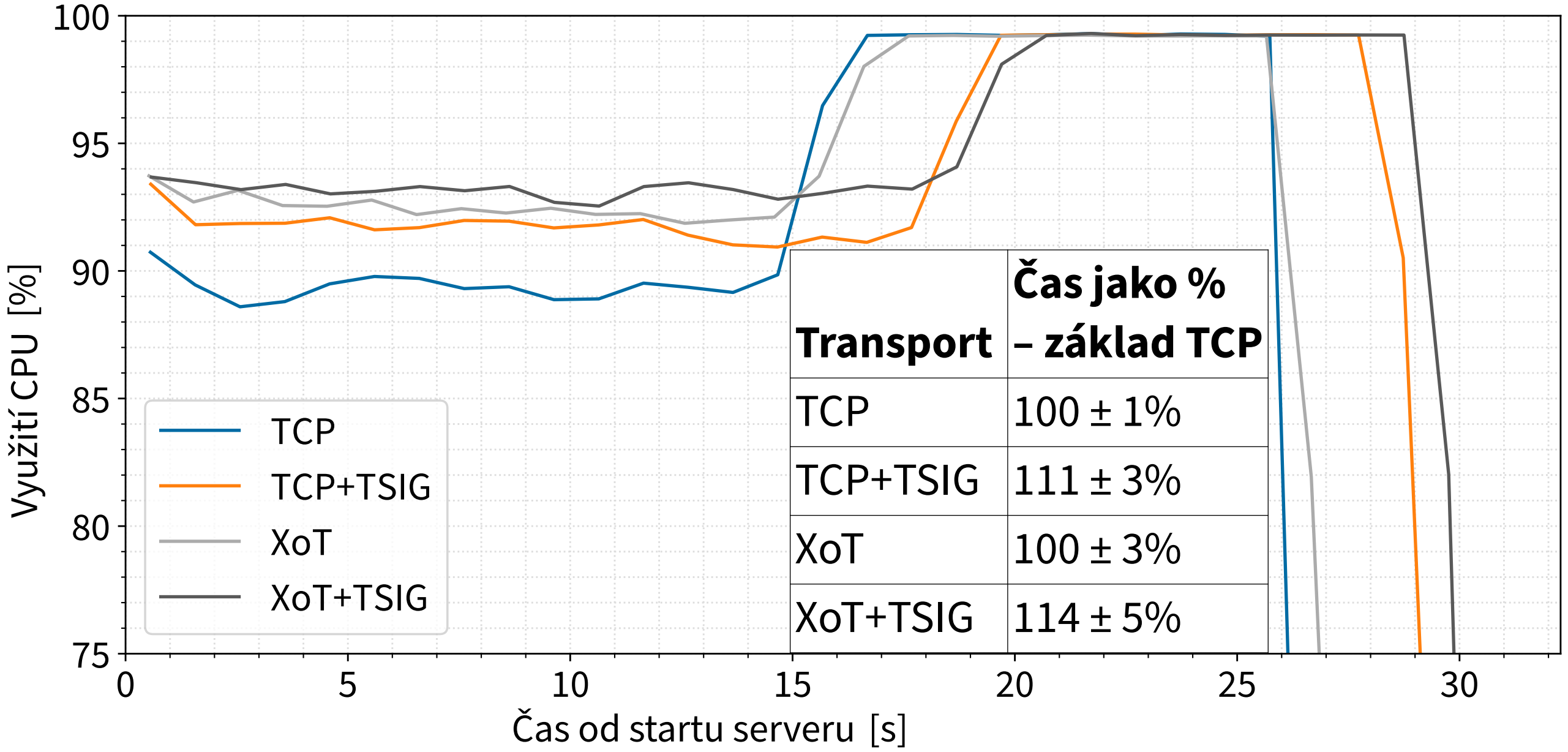
Načítání zóny, čekání na vstup



Načítání zóny, čekání na vstup



TLD, sekundární, využití CPU a čas



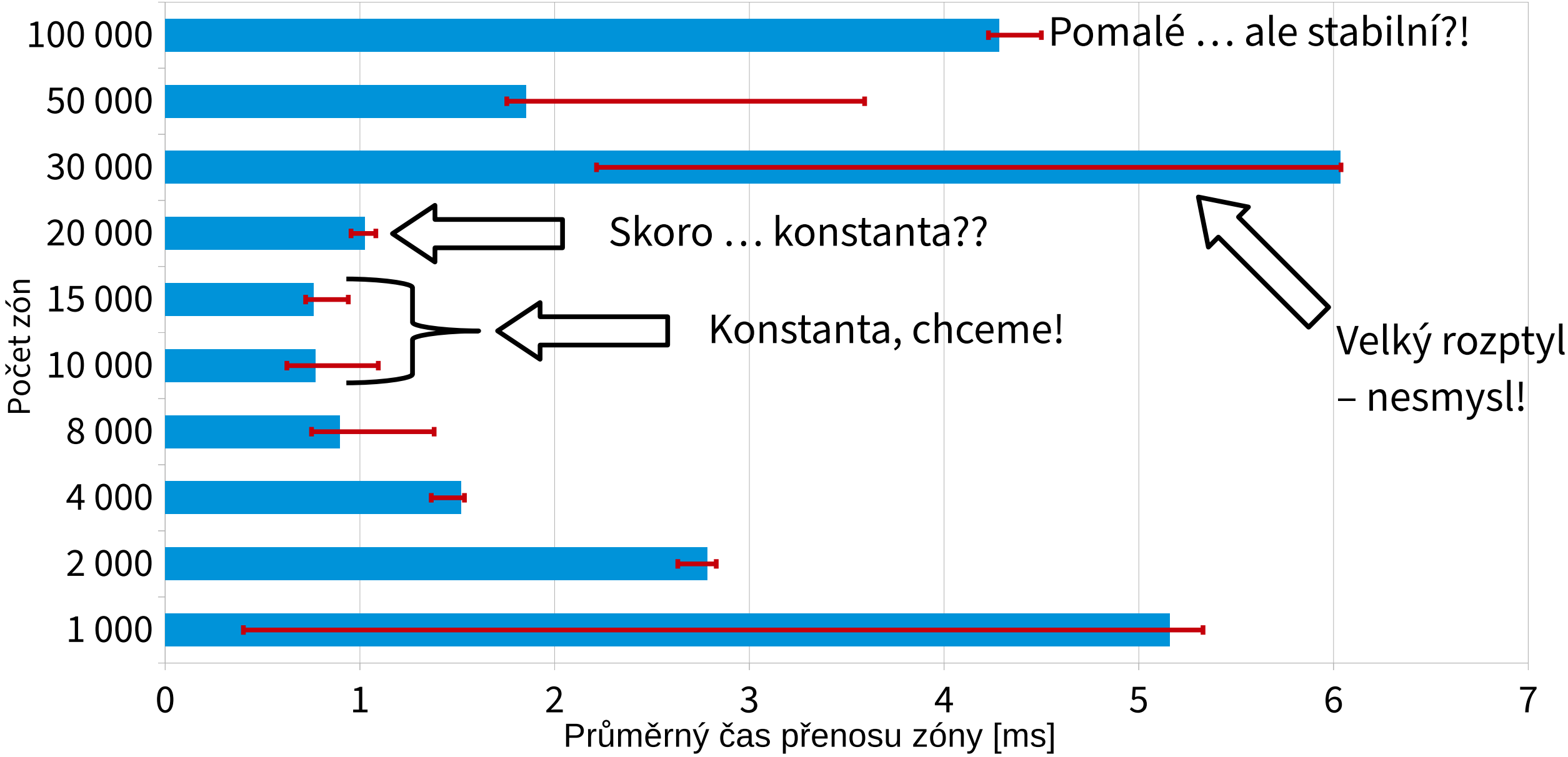
Transport	Čas jako % – základ TCP
TCP	100 ± 1%
TCP+TSIG	111 ± 3%
XoT	100 ± 3%
XoT+TSIG	114 ± 5%

Spousta malých zón

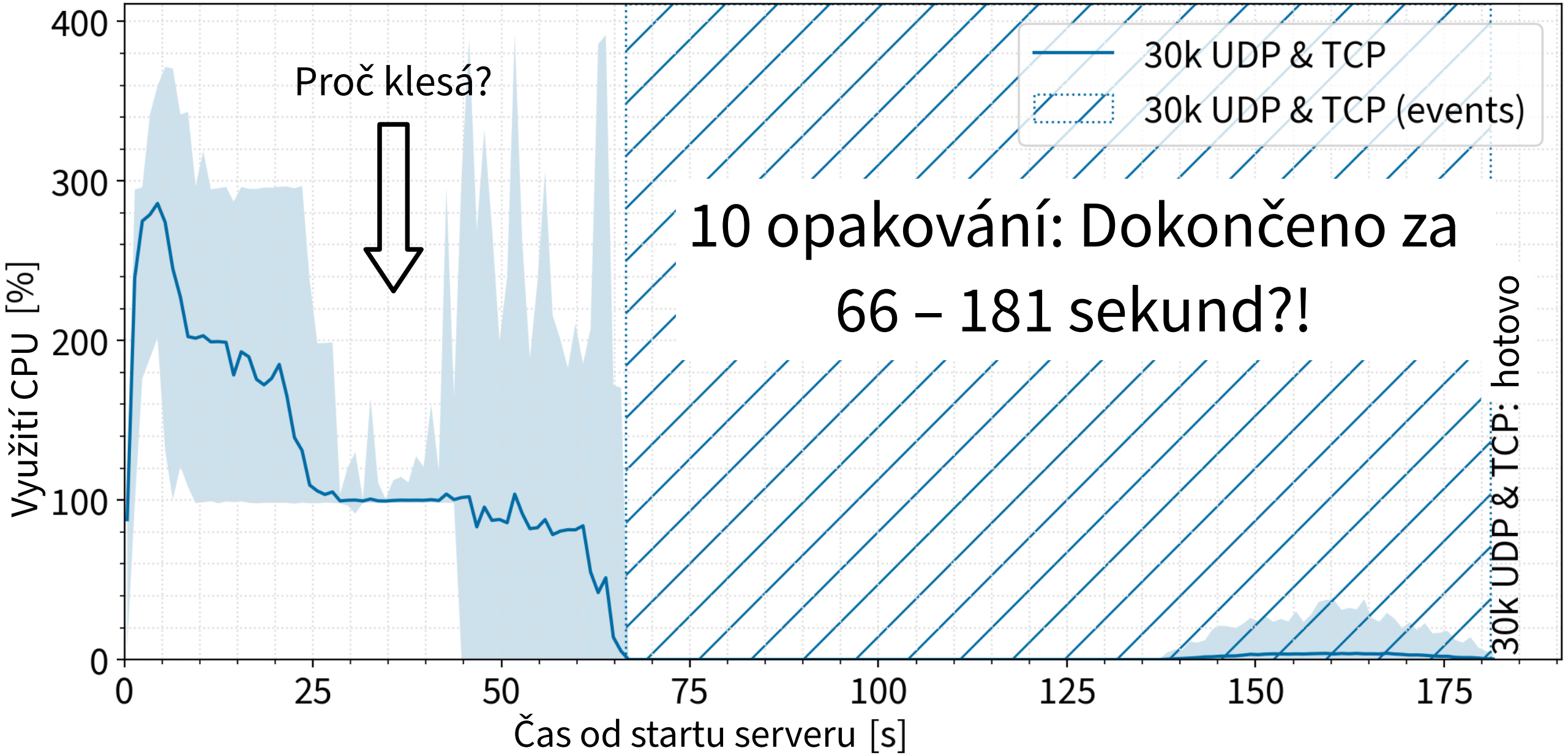
Spousta malých zón

- 1, 2, 4, 8, 10, 15, 20, 30, 50, 100 tisíc zón
- Typický DNS hosting
- Studený start sekundárního serveru
  - Nemá zóny na disku
- **Celkový čas úměrný počtu zón?**
- **Vliv transportu?**

UDP & TCP, čas na 1 zónu



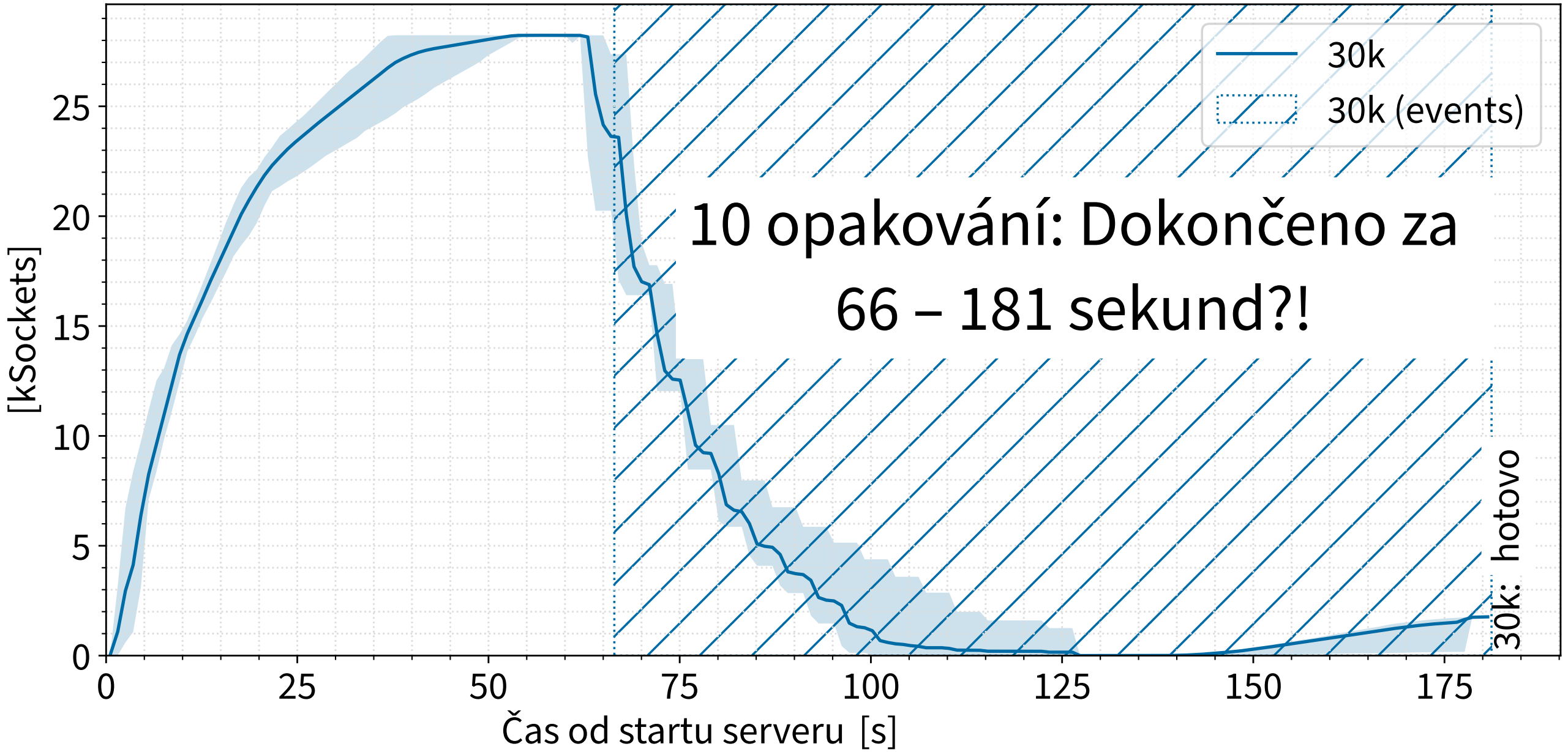
30 k zón: Využití CPU



30 k zón: sekundání server – log

- ~ 38 sekund po startu
- **`failed to connect: address not available`**
  - … poprvé

30 k zón: TCP TIME_WAIT sockety



TCP: krok 1

- spojení: (zdrojová IP, **zdrojový port**, cílová IP, cílový port)

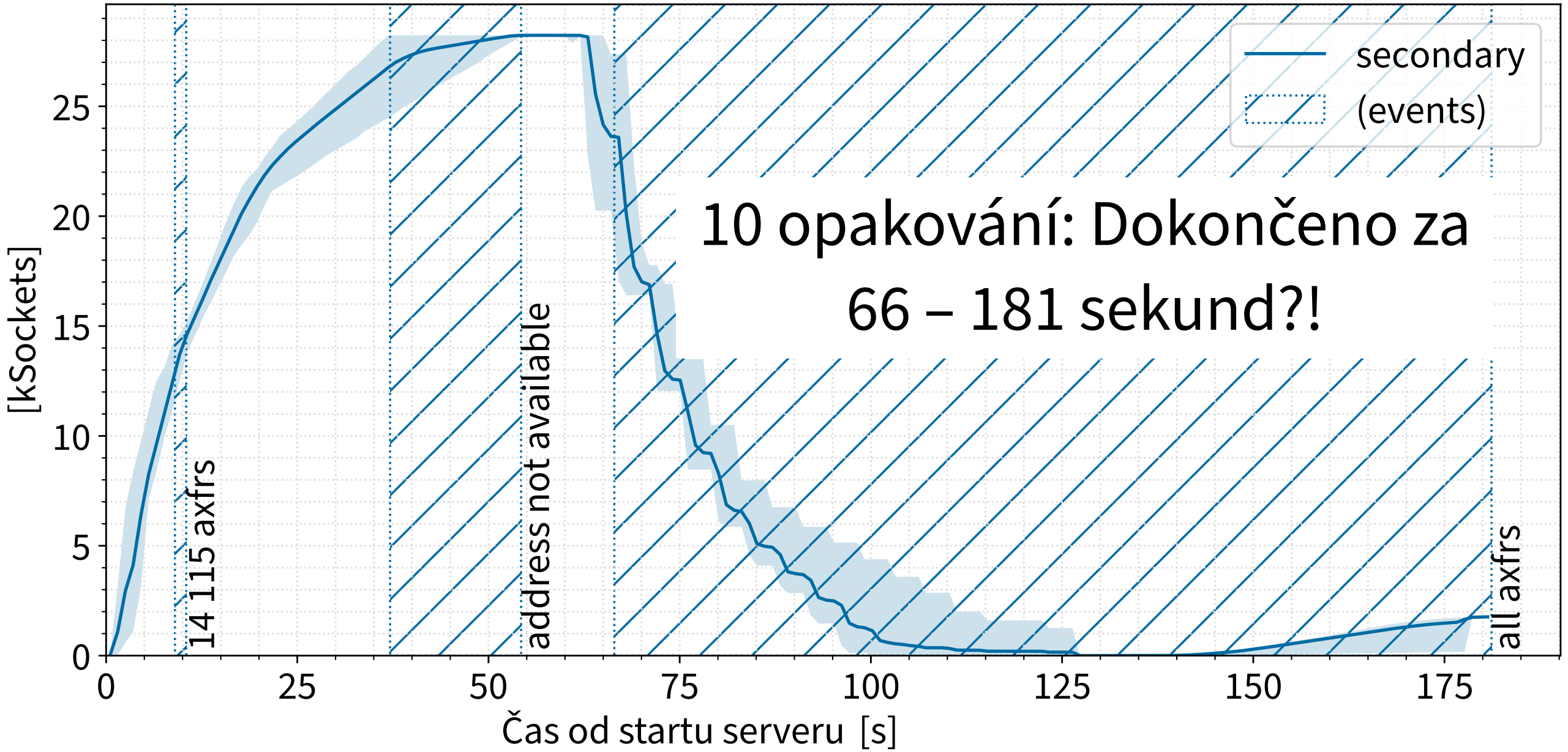
TCP: krok 2

- spojení: (zdrojová IP, **zdrojový port**, cílová IP, cílový port)
- Rozsah portů pro odchozí spojení ("ephemeral")
  - `$ sysctl net.ipv4.ip_local_port_range`
  - `net.ipv4.ip_local_port_range = 32768 60999`
    - **28 231 portů – *výchozí hodnota***

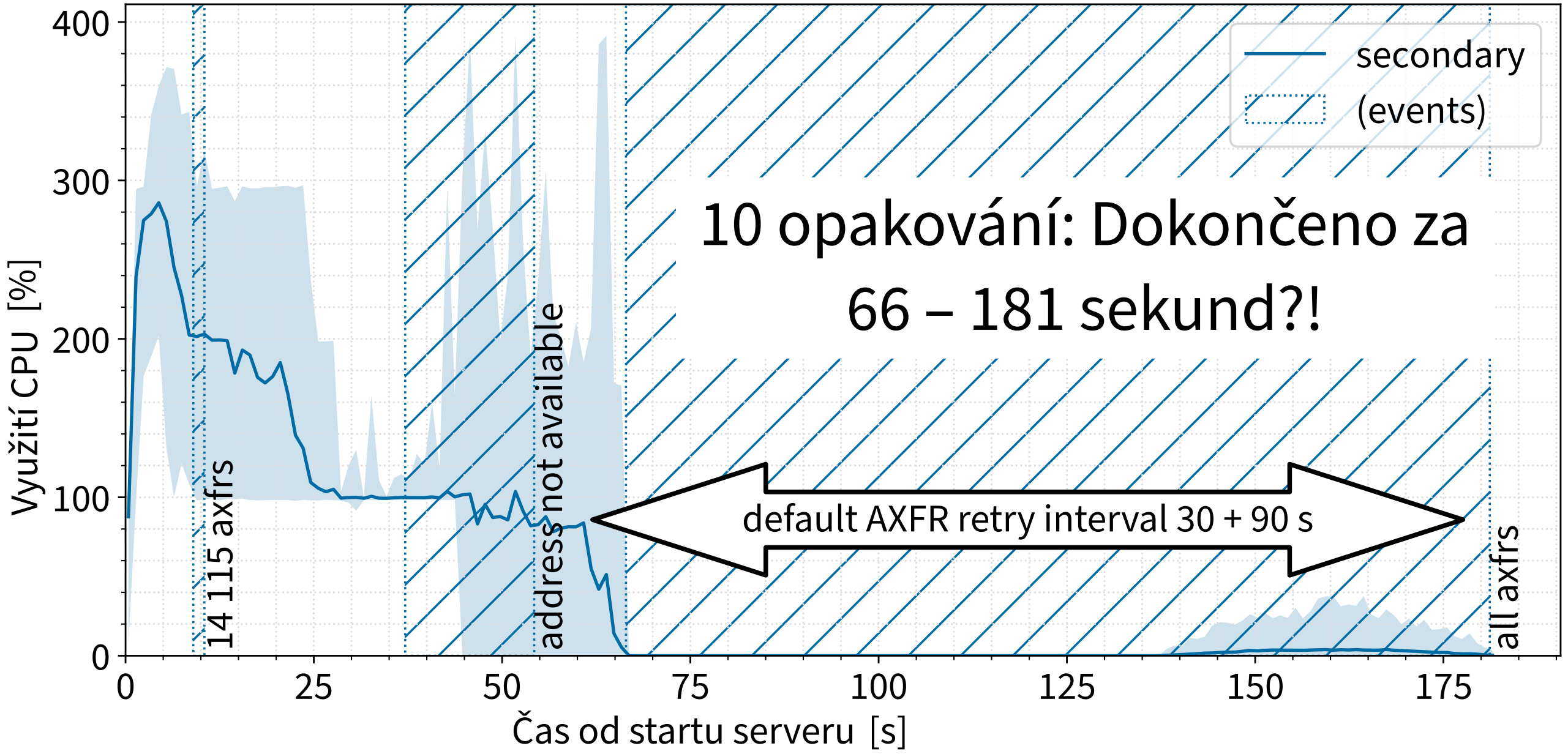
TCP: krok 3

- spojení: (zdrojová IP, **zdrojový port**, cílová IP, cílový port)
- Rozsah portů pro odchozí spojení ("ephemeral")
  - `$ sysctl net.ipv4.ip_local_port_range`
  - `net.ipv4.ip_local_port_range = 32768 60999`
    - **28 231 portů – *výchozí hodnota***
- `$ ss -t -o state time-wait`
  - `… timer:(timewait,60sec,0)`
  - výchozí hodnota: 60 s

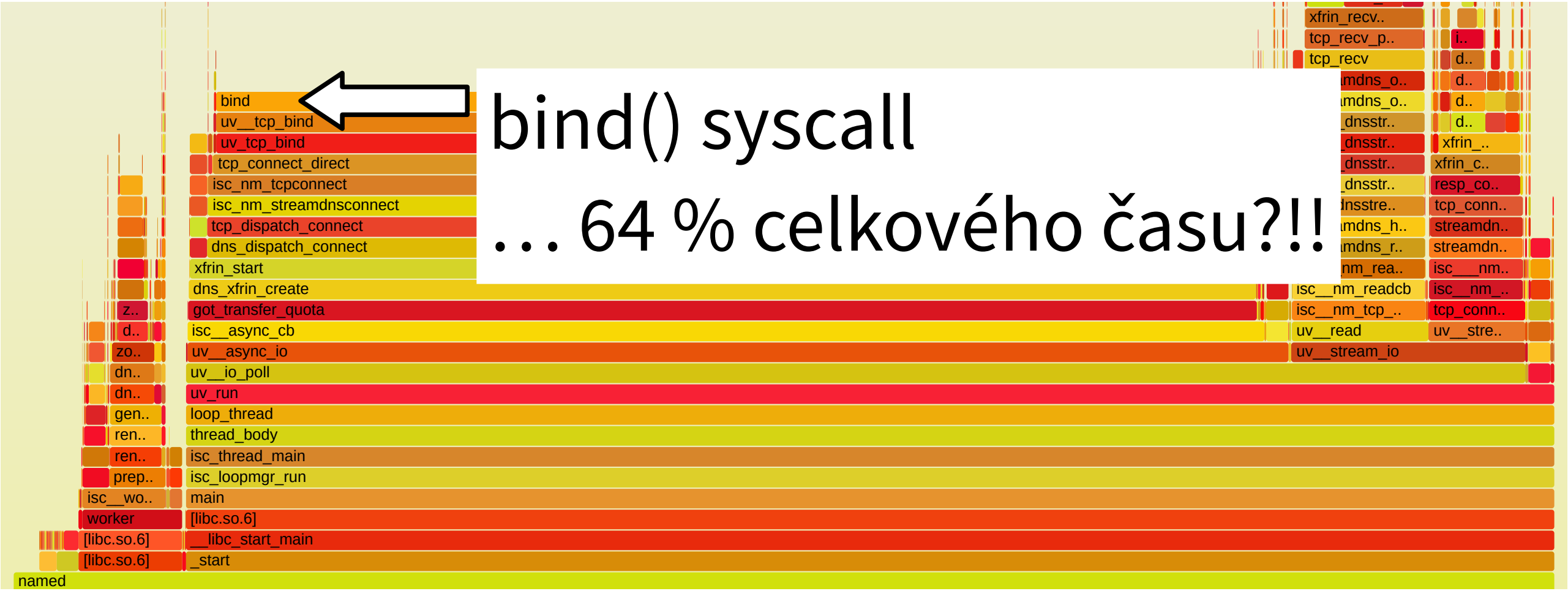
30 k zón: TCP TIME_WAIT spojení



30 k zón: Využití CPU



30 k zón: CPU profil



TCP vrací úder

- Linux Plumbers Conference 2023: **connect() – why you so slow?!**
  - https://lpc.events/event/17/contributions/1593/
  - https://youtu.be/J5Hm6PrJWI4?t=19000

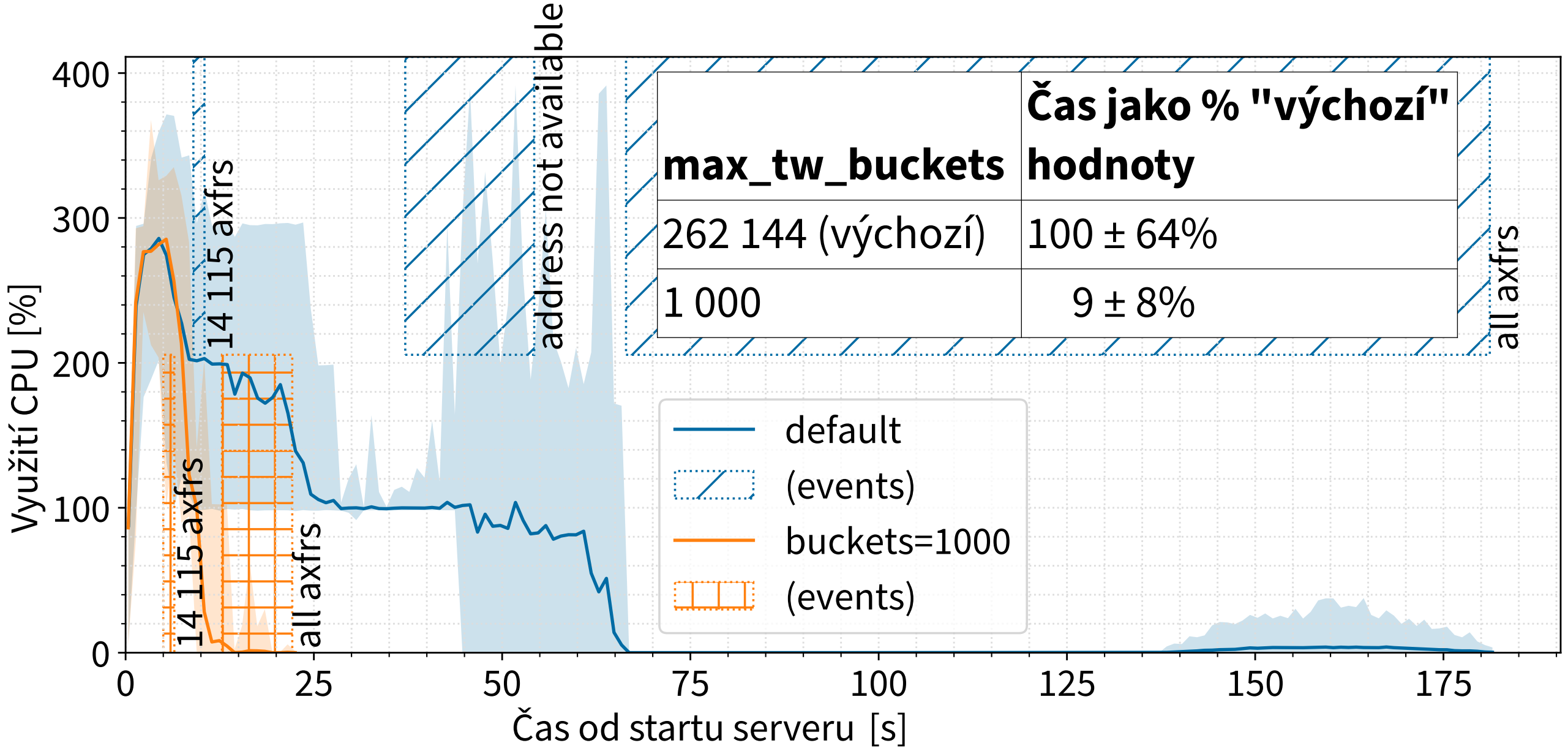
Linux sysctl tcp_max_tw_buckets

- tcp_max_tw_buckets - INTEGER
- Maximal number of timewait sockets held by system simultaneously.
- **If this number is exceeded time-wait socket is immediately destroyed** and warning is printed. This limit exists only to prevent simple DoS attacks,
- **you _must_ not lower the limit artificially**,
- but rather increase it (probably, after increasing installed memory), if network conditions require more than default value.
- ☠ `$ sysctl -w net.ipv4.tcp_max_tw_buckets=1000` ☠

Linux sysctl tcp_max_tw_buckets

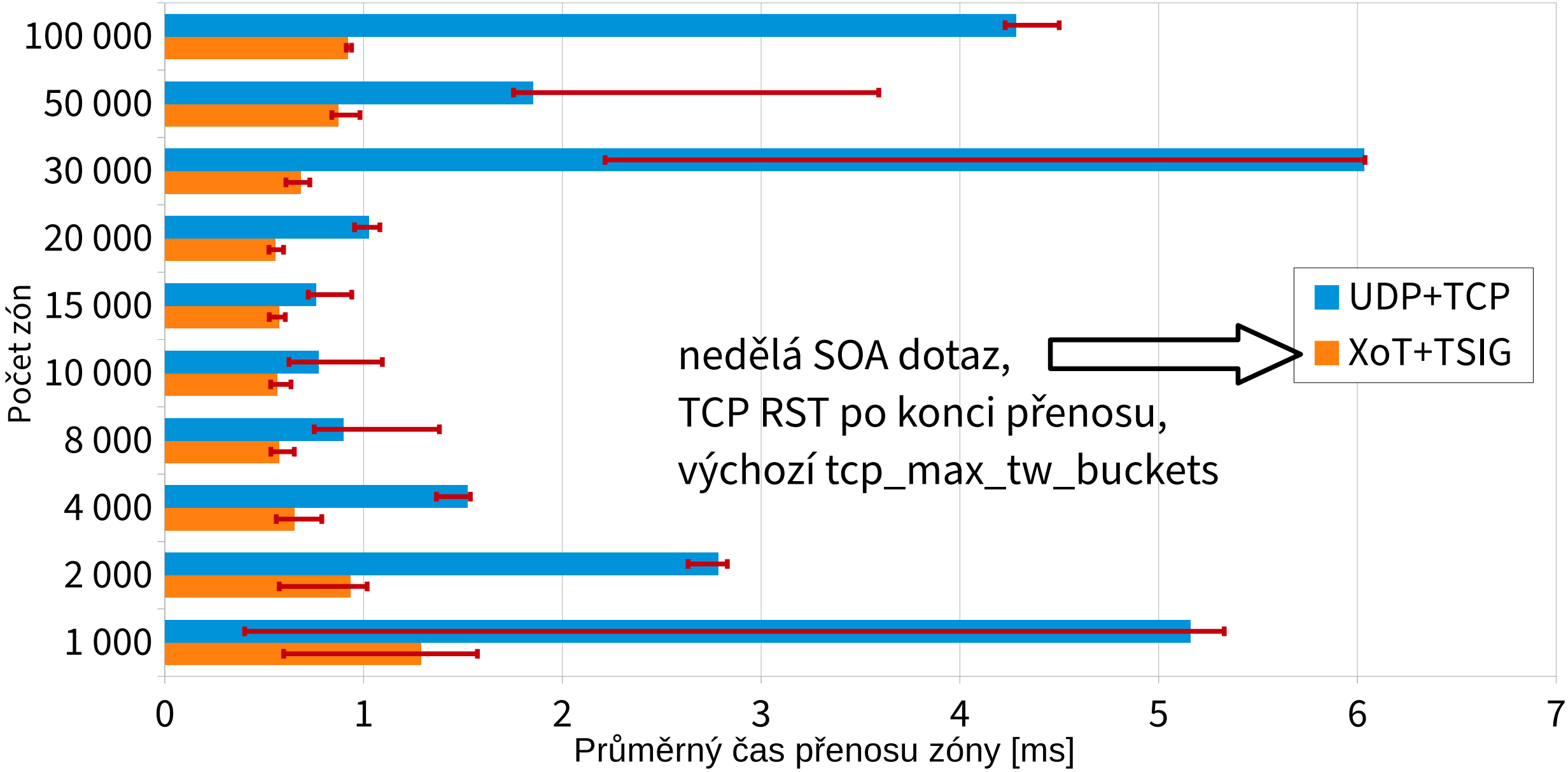
- tcp_max_tw_buckets - INTEGER
- Maximal number of timewait sockets held by system simultaneously.
- **If this number is exceeded time-wait socket is immediately destroyed** and warning is printed. This limit exists only to prevent simple DoS attacks,
- **you _must_ not lower the limit artificially**,
- f
- **Tak určitě …** ☠

CPU load, 30 k zones, tcp_max_tw_buckets

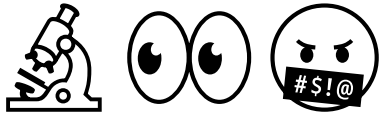


max_tw_buckets	Čas jako % "výchozí" hodnoty
262 144 (výchozí)	100 ± 64%
1 000	9 ± 8%

Průměrný čas přenosu: TCP vs. XoT



Závěrem

- Kontrola testovacího prostředí 
- **Zapomeňte na lineární chování** 
- Hodně zón => hodně TCP => hodně ladění 
  - Vylepšení protokolu? Znovupoužití spojení?
- XFR přes TLS může být rychlé
  - … náhodou