

Neviditelní návštěvníci

problémy způsobené agresivními boty na webu

Petr Krčmář



28. března 2026



Uvedené dílo (s výjimkou obrázků) podléhá licenci Creative Commons Uveďte autora 3.0 Česko.

- linuxák od roku 1998
- správce serverů
- lektor a konzultant
- šéfredaktor [Root.cz](#)
- člen [vpsFree.cz](#)
- organizátor [LinuxDays](#)
- můj web je [petrkrcmar.cz](#)



Prezentace už teď na webu

www.petrkrcmar.cz

Začneme odjinud

- provozovat **poštovní server** je čím dál složitější
- přibývají různé vrstvy ochrany
 - SPF, DKIM, DMARC, TLS, RBL, greylisting, rate limiting...
- integrace externích služeb
 - blacklisty, reputační databáze, různá API...
- hrozby jsou dynamické
 - podoba spamu se rychle mění
 - boti se přizpůsobují
- monitoring, sledování, ladění, přizpůsobení parametrů
- výsledek = je to náročné, složité a nikdo to nechce dělat
- bohužel se nám totéž **děje i na webu**

Když zlobí boti

Idylická doba

- spustit vlastní web bylo dříve velmi snadné
- stačilo na jakémkoliv počítači nastartovat server
 - stačilo Raspberry Pi u vás v komoře
- provozovat třeba malý web pro pár nadšenců nebyl problém
- občas přišel bot od vyhledávače, ale to nebyl problém
 - takových botů bylo pár a chovaly se slušně
- později přišel komentářový spam, ale ten se dal řešit
 - spousta webů komentáře vůbec nemá
 - řešení podobné jako u e-mailu
- pak ovšem přišli **boti pro AI** a ti změnili všechno

Násilí a rabování

- 1 rozsah
 - svého bota si dnes provozuje kdekdo
 - nízká bariéra nasazení - klidně stovky tisíc instancí
- 2 agresivita
 - nehledí na škody a jedou hlava nehlava
 - chybějící limity - masivní bombardování cíle
- 3 špatný návrh
 - roboti jsou špatně napsaní a dělají nepořádek
 - chyby v logice - nekontrolované chování
- 4 distribuce
 - botnety jsou rozprostřeny přes mnoho poskytovatelů
 - anonymizace, orchestrace - obtížně se blokují
- 5 zneužití
 - boti často zneužívají legitimní zdroje
 - cloud, VPN, úniky klíčů do API, proxy služby, VPN

Co boti chtějí?

- nájezd má jediný cíl – vytěžit všechna data
 - snaha nakrmit nenasytnou chuť po trénovacích datech pro LLM
- velké i malé firmy chtějí všechno stáhnout
 - „máme právo na veřejná data a můžeme si je vzít“
- některé firmy to dělají viditelně (OpenAI, Meta, Google...)
 - existují jich ale tisíce – každý chce svůj model
- **dynamický web** umí vygenerovat stovky tisíc stránek
 - WordPress je v tom přeborník – archivy, kategorie, tagy...
 - mailing listy, gitovské repozitáře, databáze...

Vedlejší škody

Co se děje?

- dramaticky zvýšené zatížení serveru
 - CPU, paměť, I/O a síťové toky prudce rostou
- vyčerpání zdrojů
 - krátkodobé nebo trvalé vyčerpání – selhání nových požadavků
- zvýšení latence
 - uživatelé sledují zpomalení webu
- vyčerpání úložiště
 - logy, keš nebo uploady zaplní diskový prostor
- falešné alarmy
 - zaplavení IDS/IPS a SIEM signálů způsobí únavu z alertů

Konkrétní pozorování

- server s 2000 požadavky za sekundu
 - pokročilou filtrací pokles na 50 za sekundu
- lokální web s tisíci návštěvami denně
 - najednou stovky tisíc požadavků z celého světa
- roboti nereagují na HTTP kódy a kešovací hlavičky
 - sto tisíc požadavků na stránku vracející 404
- charakteristikou to připomíná DDoS
 - Datový Drtič online Služeb
- **více než 90 % provozu na webu je dnes šum!**

Tak dělejme jednodušší weby!

- nápad: když budeme mít jednoduché weby, nevadí nám to
 - předpoklad: statické stránky podáváme rychle, nezatěžuje nás to
- myšlenka dobrá, ale stále je tu problém s **přenosem dat**
- konkrétní [příběh firmy Read the Docs](#)
- provozují servery s dokumentací k různým projektům
- jeden jediný bot jim stáhl **73 TB dat** za měsíc
 - trvalý tok 218 Mbit/s celý měsíc
- bot se dostal na URL mimo kešující CDN (chyba)
- stahoval opakovaně tisíckrát každý velký soubor
- stálo je to 5000 dolarů na nákladech za přenosy
 - firma naštěstí reagovala, chybu opravila a škodu nahradila
- později někdo zneužil Facebook content downloader
 - stáhl dalších 10 TB dat, Facebook je nekontaktovatelný

Strčme to za CDN!

- existují veřejné služby, které nabízejí filtraci
 - Cloudflare, Fastly, Bunny, GCore...
- je to levné, jednoduché a rychlé řešení
- ale přicházíme tím o **malý internet**
 - stáváme se závislími na externích službách
 - podřizujeme se rozhodování velkých subjektů
- tohle se stalo u e-mailu a vážně to hrozí webu

Protiopatření

Slušné požádání

- s boty komunikujeme pomocí souboru robots.txt
- můžeme je slušně požádat, aby odešli
- tohle funguje opět jen na ty slušné
 - s těmi ale vlastně nemáme problém
- problémoví boti obsah souboru ignorují
 - jejich autorům je váš názor ukradený
- pomůže to jen omezeně a spíše vůbec ne

robots.txt

```
User-agent: GPTBot  
Disallow: /
```

```
User-agent: bingbot  
Allow: /  
Crawl-delay: 10
```

Limit na úrovni IP

- webové servery dovolují limitovat počty požadavků pomocí IP
 - Nginx má modul `ngx_http_limit_req_module`
 - metoda děravého kyblíku, po překročení počítá a pak zastaví
- funguje proti jednoduchým stahovačům z jednoho bodu
 - nebo třeba proti pokusům o injeckáže a podobně
- AI boti ale využívají obrovské množství rozsahů z celého světa
 - často koordinují postup z různých sítí či cloudů
 - běžně přichází každý dotaz z **jiné adresy** - nepřekročí limit
- tohle vlastně taky nefunguje, mají moc adres a jsou distribuovaní

Omezení kadence

```
limit_req_zone $binary_remote_addr zone=one:10m rate=1r/s;
```

Blokace podle států

- webové servery mají podporu pro GeoIP
 - balíčky `libnginx-mod-http-geoip` a `geoip-database`
- můžeme vpustit jen uživatele využívající rozsahy vybraného státu
 - nebo naopak blokovat jen konkrétní země
- nehodí se pro weby s mezinárodním publikem
- velké riziko **falešné positivity**
 - nerozpoznáme bota od skutečného uživatele
 - databáze nejsou vždy stoprocentně přesné
 - uživatelé mohou cestovat a nechceme je vyhodit
- při rozumném využití může pomoci, dopad je omezený

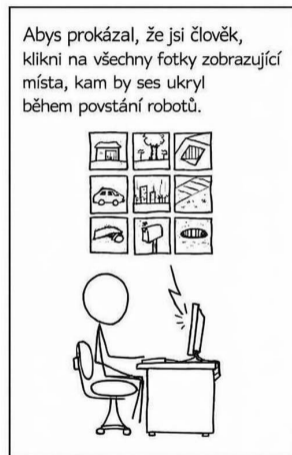
Blokace provozu z datacenter

- můžeme využít různých sbírek IP adres z datacenter
 - například iplists.firehol.org
- podobně existují sbírky notoricky známých sítí s boty
 - například blocklist.de
- zablokujeme tím tisíce rozsahů (stovky milionů IP)
- seznamy je třeba pravidelně aktualizovat, situace se mění
- pozor na přátelské aktivity (vyhledávače), ty chceme
- podle mých zkušeností účinná strategie
 - stále ovšem nezachytí vše, hrubé síto

- další seznamy IP adres vykazujících podezřelé chování
- například český projekt [NERD](#) z dílny sdružení Cesnet
 - reputační data poskytuje i [Turris](#)
- lze filtrovat podle různých kategorií, událostí, skóre a podobně
- získáme další zdroj potenciálně problémových IP adres
- vysoké riziko **falešné positivity** a omezení uživatelů
 - je nebezpečné prostě je bez dalšího blokovat
- seznamy jsou dynamické a adresa se na ně může dostat různě
- mnoho uživatelů je dnes za NAT, lze je snadno odříznout

CAPTCHA

- dokaž, že jsi člověk...
 - ... a ještě k tomu jasnovidec
- velmi **otravná** záležitost
- vysoká míra falešné positivity
- tohle nikdo nechce podstupovat
- navíc opakovaně na každém webu!
- vyhodíte spoustu lidí



Jde to i jinak...

XKCD 2228

ROOT.CZ vpsFree.cz

Důkaz prací

Anubis

- egyptský bůh pohřebišť
- soudce a ochránce mrtvých
- zobrazovaný jako muž s hlavou šakala
- soudí zemřelé – váží jejich srdce
 - hledá svědomí, pravdu a spravedlnost
- pokud je srdce těžší, hříšník je zatracen
 - lehčí srdce může do věčného života
- postavíme soudce ke svým serverům



Jeff Dahl, CC by-sa

ROOT.CZ vpsFree.cz

Ověřujeme, že nejste robot

Making sure you're not a bot!



Calculating...
Difficulty: 4, Speed: 0kH/s



► Details

Protected by [Anubis](#) From [Techaro](#). Made with ❤️ in .

Mascot design by [CELPHASE](#).

This website is running Anubis version v1.25.0.

Proof of work (PoW)

- aktuální trend pro rozdělování botů a lidí
- na rozdíl od CAPTCHA neobtěžuje uživatele
- prohlížeč musí vyřešit **výpočetní operaci**
- uživatel zaznamená jen krátkou prodlevu
 - ukáže se mu informační obrazovka
 - poté se automaticky zobrazí web
 - po uživateli se nechce žádná akce
- botům by skenování výrazně zvýšilo zátěž
 - boti nespouštějí JavaScript, jen stahují
 - nevyřeší hádanku → nedostanou se k obsahu

- vznikl v lednu 2025 (mladý projekt)
- funguje jako reverzní proxy
 - TLS terminace → Anubis → web
- novému návštěvníku se ukáže Anubis
- dostane výzvu a musí zjistit správnou odpověď
- pokud odpoví, dostane propustku (JSON web token)
 - ta se uloží do cookie a pustí ho dál
 - je platná omezenou dobu
 - digitálně podepsaná dočasnou ed25519
- klient s platnou propustkou prochází bez přerušení

Co se počítá?

- úloha je velmi podobná algoritmu Bitcoinu
- zadavatel vytvoří náhodnou výzvu (challenge)
- klient k ní má za úkol přidat vlastní řetězec (nonce)
 - výsledek poté zahašuje pomocí SHA-256
- výsledkem musí být řetězec začínající **daným počtem nul**
 - výchozí hodnota je 5, určuje obtížnost úkolu
- řešitelné jen hrubou silou - musíte se trefit
 - server pak může odpověď snadno ověřit jedním hašem

Funkce

```
const hash = await sha256(`${challenge}${nonce}`);
```

Co když nemám JavaScript?

- Anubis je připraven na nasazení různých výzev
- implementována je výzva **Meta Refresh** nevyžadující JavaScript
 - mnohem jednodušší, jen automatické načtení stránky
- využívá volbu `http-equiv="refresh"` vynucující znovunačtení
- funguje ve všech prohlížečích, ale dá se obejít

Pravidlo pro Meta refresh

```
- name: generic-browser
  user_agent_regex: >-
    Mozilla|Opera
  action: CHALLENGE
  challenge:
    difficulty: 1          # počet sekund pro načtení
    algorithm: metarefresh # úloha nevyžadující JavaScript
```

Pravidla

- Anubis se řídí pravidly napsanými v YAML
- možné posuzovat: URL, user-agent a HTTP hlavičky
- lze filtrovat i podle IP adresy (rozsahů)
- tři možné verdikty: allow, challenge nebo deny
- lze měnit obtížnost výzvy a algoritmus (rychlý vs. náročný)

Pravidla

```
- name: generic-bot-catchall
  user_agent_regex: (?i:bot|crawler)
  action: CHALLENGE
  challenge:
    difficulty: 16
    algorithm: slow
```

- Anubis přichází s vlastní knihovnou pravidel
 - využívá možnosti vložení (include) mezi soubory
- obsahuje užitečná nastavení pro začátek
- odlišuje některé hodné boty podle IP adres
 - například zná slušné vyhledávače
 - naopak cíleně blokuje notoricky známé boty
- lze ji použít jako základ a poté rozšířit vlastními pravidly
- nezapomeňte na externí služby, které používáte
 - monitoring, optimalizace obrázků, localhost...

Propojení s GeoIP

- u jazykově specifických webů lze využít
- Anubis přímo neumí, ale web server může signalizovat
 - pošleme mu hlavičku s kódem země
- například české uživatele můžeme pustit bez ověření

Hlavička Nginx

```
proxy_set_header X-GeoIP-Country-Code "$geoip_country_code";
```

Pravidlo

```
- name: allow-cz-users
  headers_regex:
    X-GeoIP-Country-Code: CZ
  action: ALLOW
```

Propojme to všechno

Kombinace různých metod

- zablokovat známé boty a datacentra
- limitovat velký počet dotazů dle IP
- pustit uživatele z vybraných států
 - volitelně omezit jiné státy
- ostatním nasadit Anubis
 - nezablokujeme žádného legitimního uživatele
 - někteří si jen občas chvíli počkají

Konfigurovatelnost

- výhoda: je to silně konfigurovatelné
 - možno přizpůsobit konkrétním potřebám
 - lze zahrnout GeoIP, user-agenty, rozsahy IP...
 - můžeme více chránit některé části webu
 - naopak můžeme odkrýt ty necitlivé
- nevýhoda: je to silně konfigurovatelné
 - musíme se tím zabývat
 - ladění zabere nějakou dobu
 - boti se budou pravděpodobně přizpůsobovat
 - budeme muset časem nějak dál upravovat

Shrnutí

- blokáce známých datacenter zastaví zhruba polovinu dotazů
- Anubis vytřídí dalších 90 % zbývajících provozu
 - při otevření přístupu uživatelům místního státu
- úplně zmizely obrovské špičky
- zátěž je vyrovnaná, server se nepřetěžuje
- výsledek: funguje to podle očekávání

Můžou se přizpůsobit?

- ano a pravděpodobně to udělají
- budeme se přizpůsobovat taky
 - vyměníme úlohu, přidáme další testy
- možná budeme muset dělat fingerprinting
 - včetně toho, jak prohlížeč vykresluje písmo
- nebo budeme muset limitovat vypočítanou cookie
 - jedna cookie stáhne jen několik stránek
- bohužel to asi bude nekonečný boj

Boti tu jsou a budou

- snaží se být nenápadní
- vydávají se za běžné prohlížeče
- využívají mnoho obrovských rozsahů IP adres
- je jich hodně a jsou agresivní
- zřejmě bude ještě hůř
- nespíme a máme metody, jak se jim bránit

- Domovská stránka projektu Anubis
- Dokumentace k Anubisu
- Zbavte se na webu otravných robotů, vyžeňte je nástrojem go-away
- Fighting the AI scraperbot scourge
- Anubis sends AI scraperbots to a well-deserved fate
- AI crawlers need to be more respectful

Otázky?



Petr Krčmář
petr.krcmar@iinfo.cz