

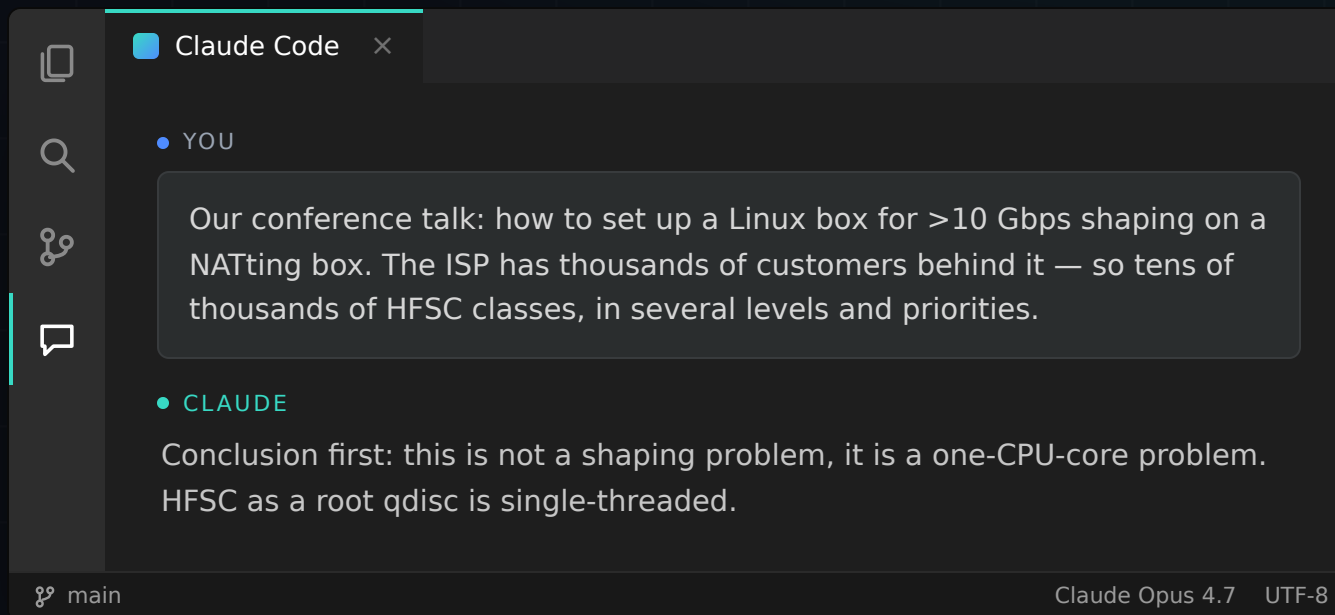
# Shaping ... ?

---

autor: Karel Řeřicha

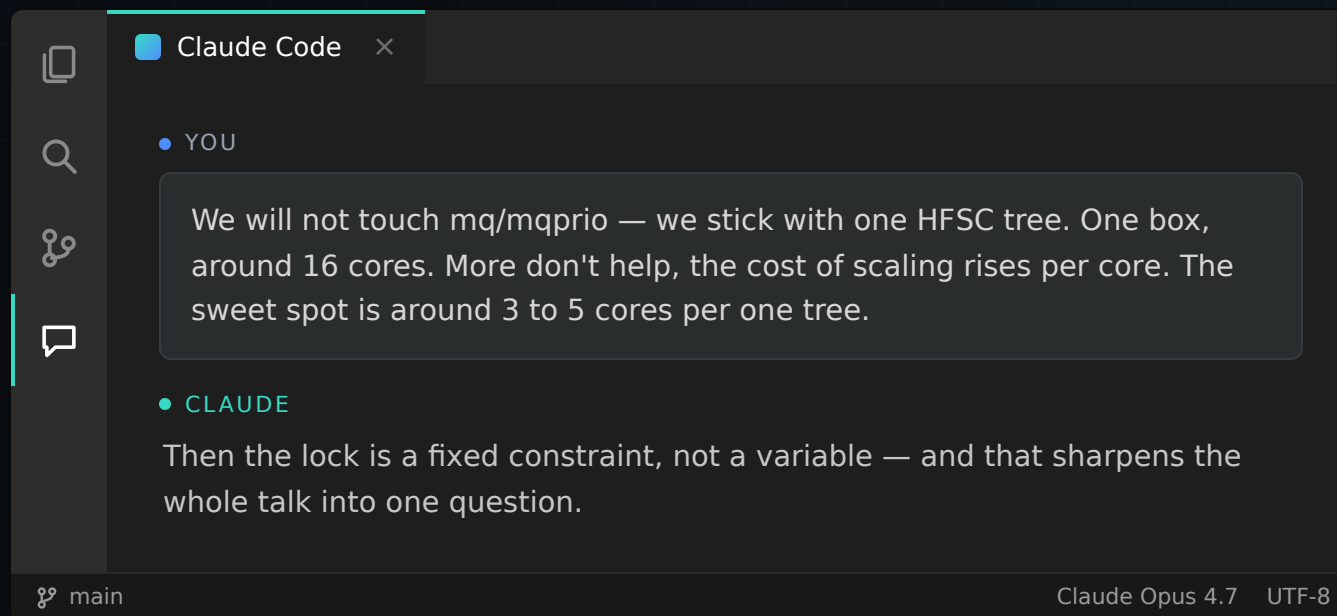
## Není to problém shapingu - je to problém jednoho jádra

- ▶ Začal jsem popisem problému: >10 Gbps na NAT boxu.
- ▶ Desítky tisíc HFSC tříd, tisíce zákazníků.
- ▶ Čekal jsem výčet optimalizací.
- ▶ Místo toho přišlo přerámování celého zadání.



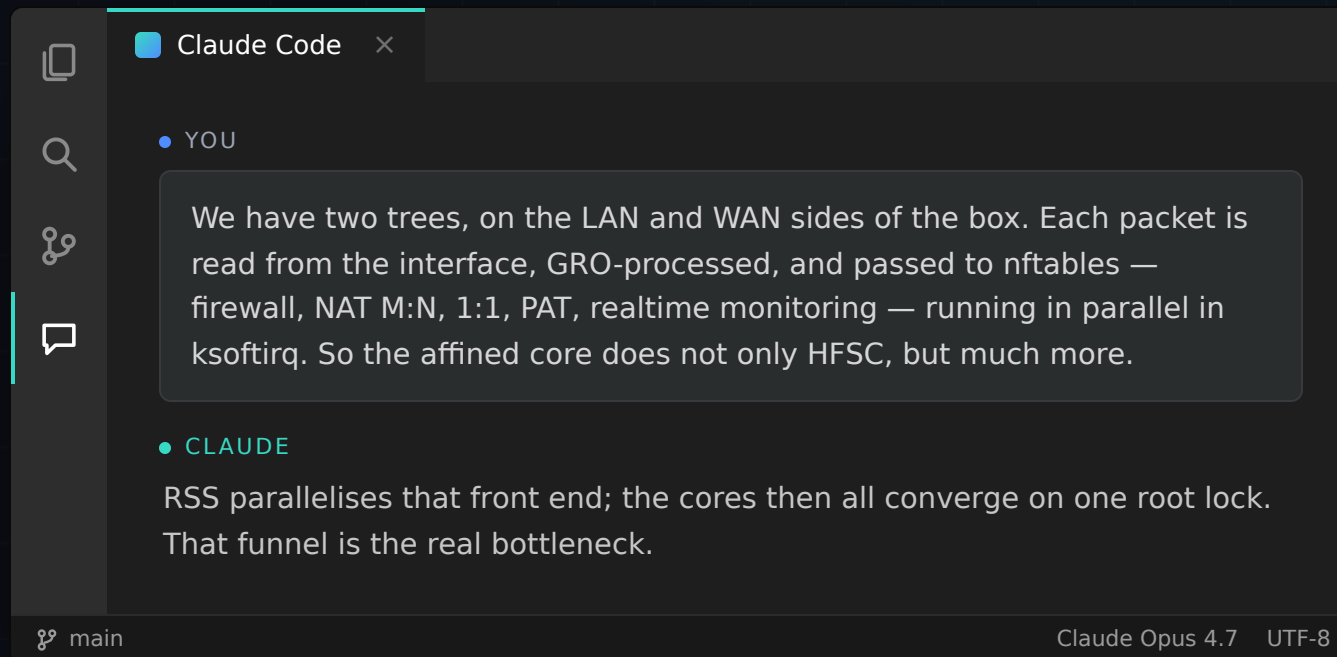
## Jeden strom, ~16 jader - a zámek jako konstanta

- ▶ Upřesnil jsem mantinely: jeden strom, ~16 jader.
- ▶ Zmínil jsem náš sweet spot 3-5 jader na strom.
- ▶ Reakce: pak je spinlock konstanta, ne proměnná.
- ▶ To zaostřilo celou další debatu.



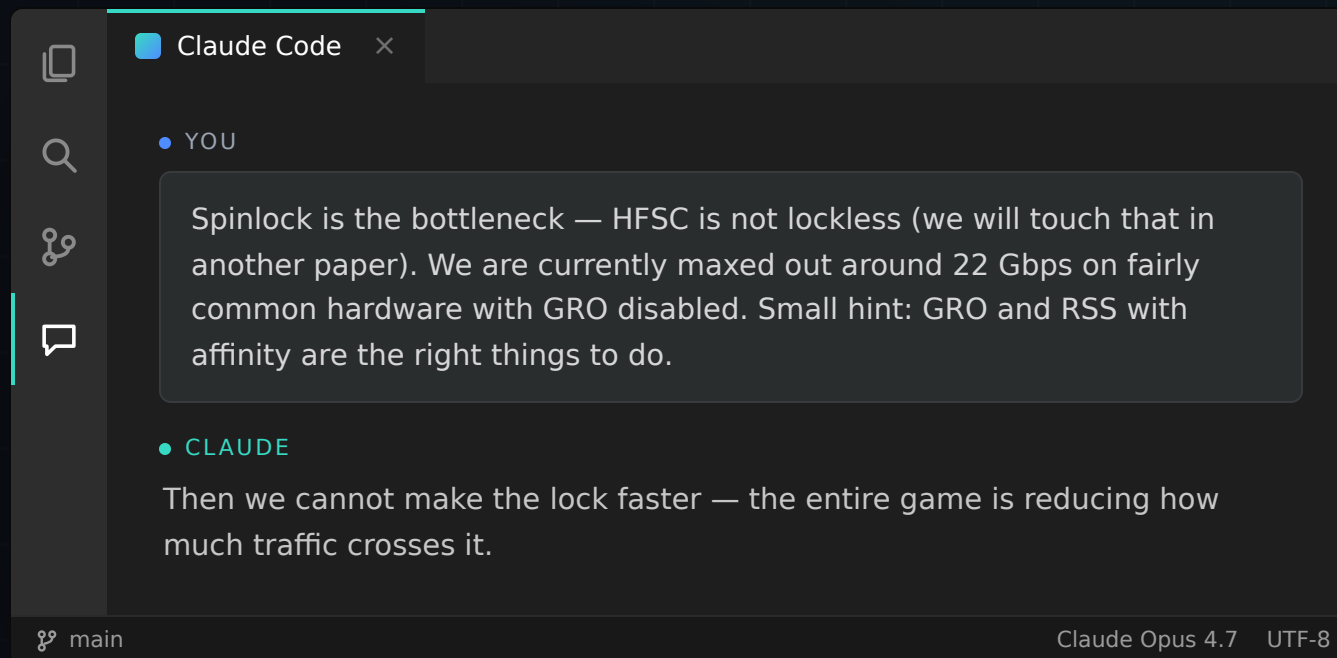
## Dva stromy a celá cesta paketu

- ▶ Popsal jsem reálnou pipeline paketu na boxu.
- ▶ Dva stromy, GRO → nftables → ksoftirq.
- ▶ Affinované jádro nedělá zdaleka jen HFSC.
- ▶ Potvrzeno: RSS paralelizuje, zámek je trychtýř.



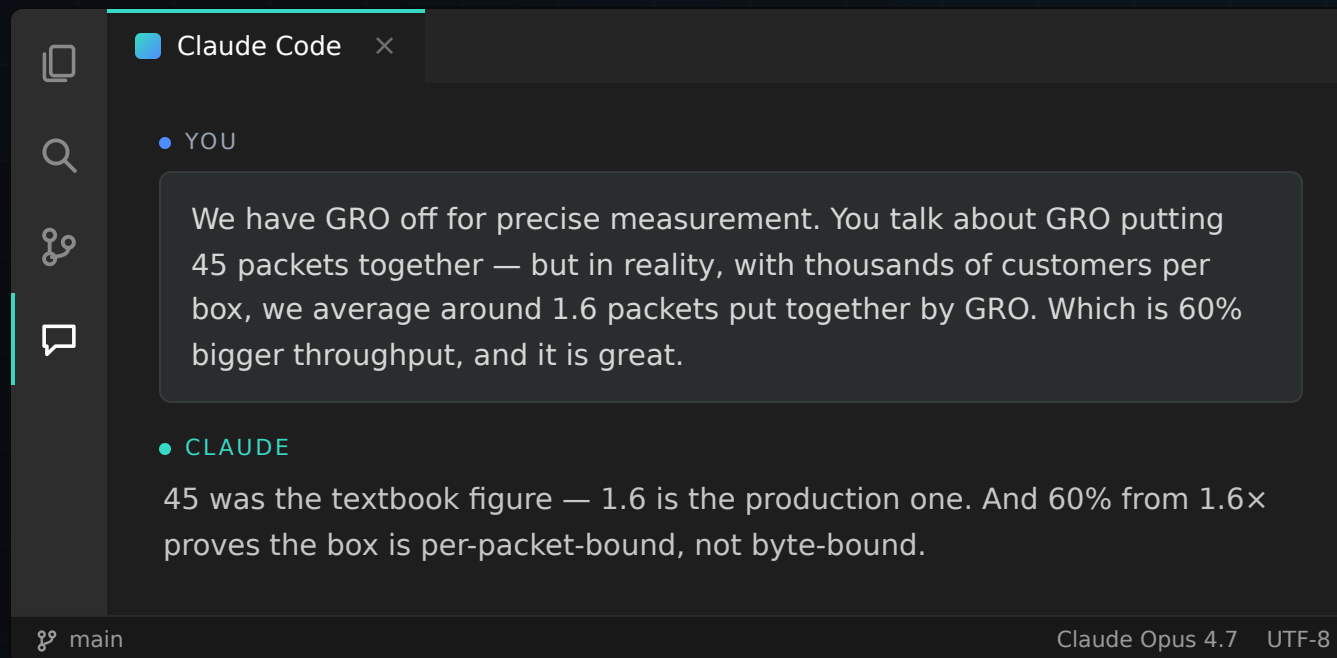
## Spinlock je bottleneck - a je daný

- ▶ Řekl jsem natvrdo: spinlock je bottleneck.
- ▶ A že teď jedeme ~22 Gbps s vypnutým GRO.
- ▶ Dal jsem hint: GRO a RSS s afinitou.
- ▶ Shoda: zámek nezrychlíme, jen ho míň zatížíme.



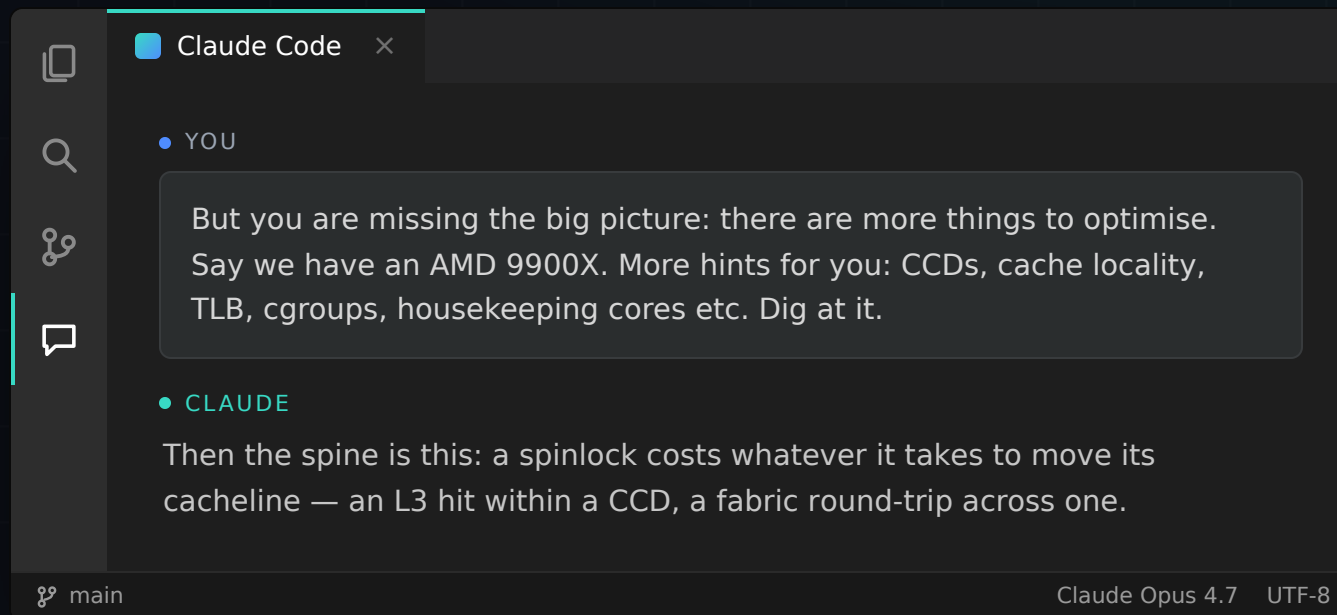
## GRO: učebnice slibuje 45, realita dává 1.6

- ▶ Opravil jsem učebnicové číslo 45.
- ▶ V realitě ISP máme průměr ~1.6 paketu.
- ▶ I to dává +60 % propustnosti.
- ▶ Z toho odvodil: box je vázaný na pakety.



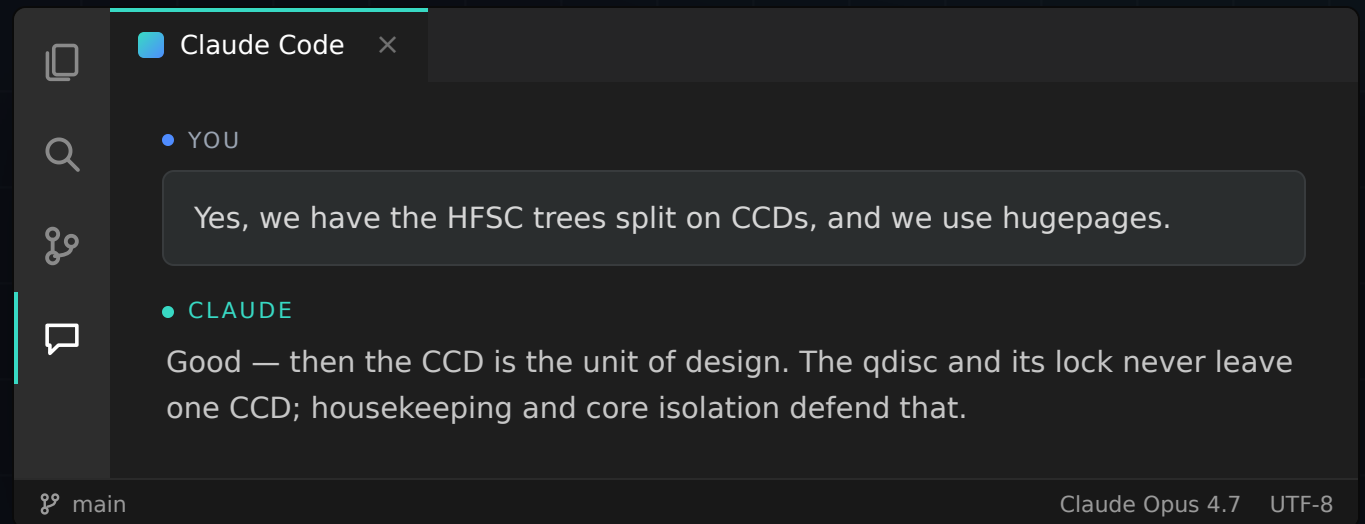
## „Chybí ti velký obraz“ - chiplet, CCD, cache

- ▶ Řekl jsem: díváš se moc úzce.
- ▶ Dal jsem hardware (AMD 9900X) a témata.
- ▶ CCD, cache locality, TLB, cgroups, housekeeping.
- ▶ „Dig at it“ - a konverzace se zlomila.



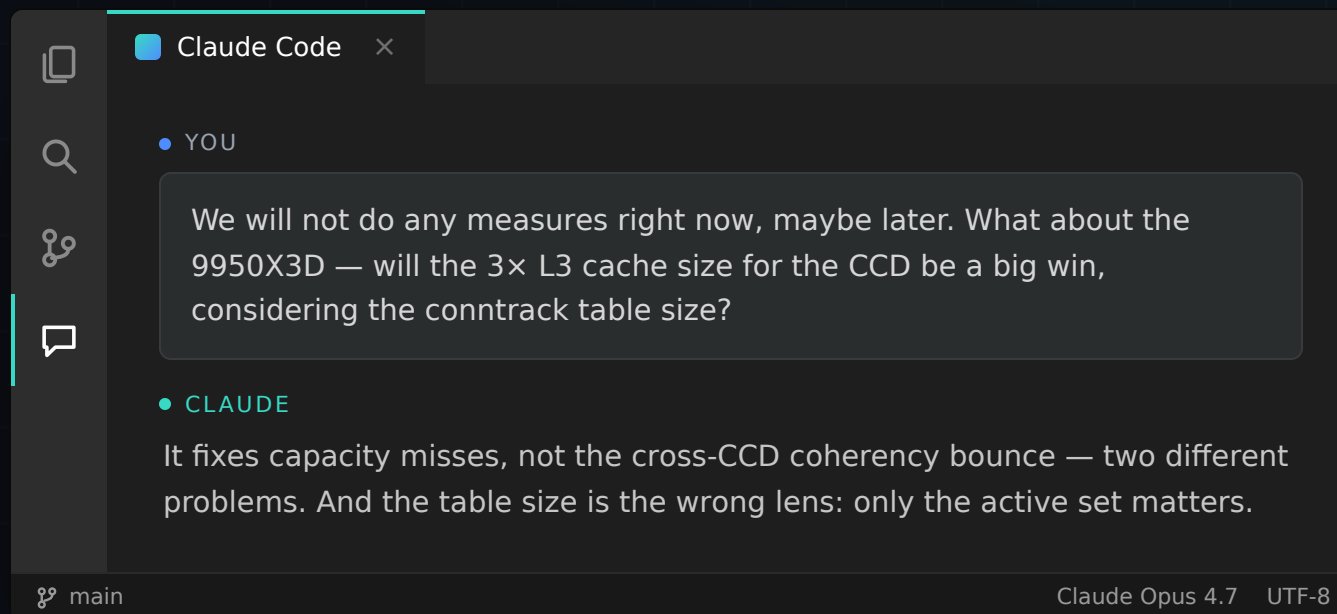
## Stromy připnuté po CCD, hugepages

- ▶ Potvrdil jsem: stromy máme připnuté po CCD.
- ▶ A že používáme hugepages.
- ▶ Tím se potvrdil celý model návrhu.
- ▶ CCD jako jednotka – zámeček svůj CCD neopustí.



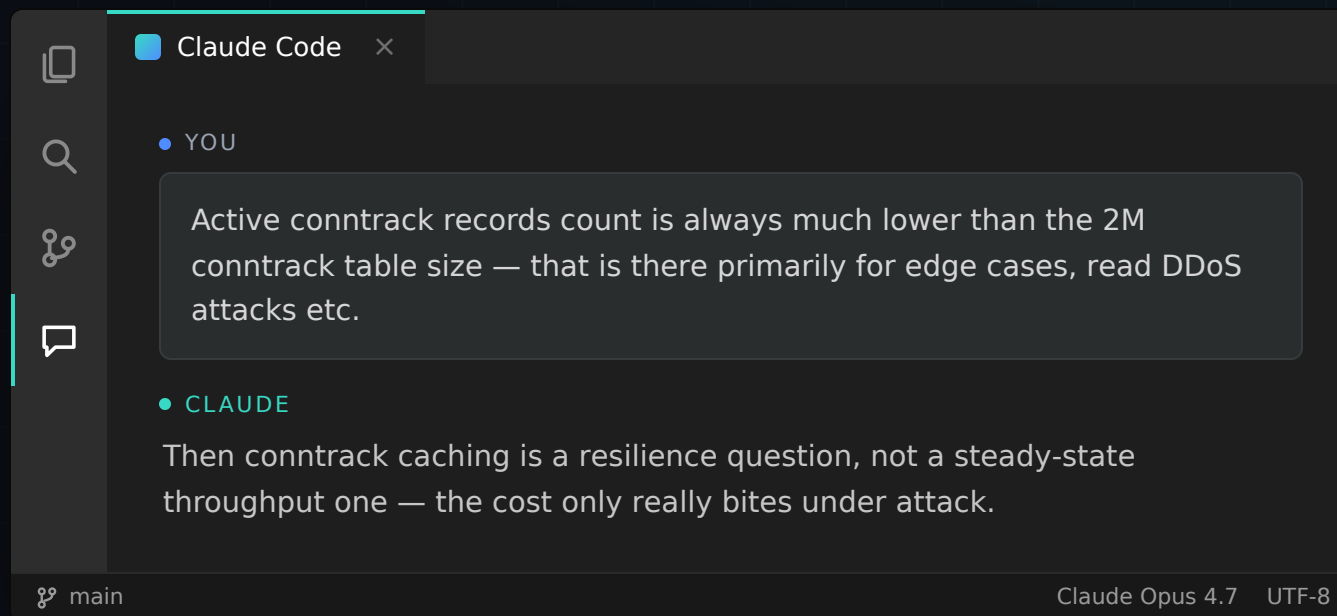
## 9950X3D: pomůže 3x větší L3?

- ▶ Zeptal jsem se na 9950X3D a 3x větší L3.
- ▶ Jestli to pomůže vzhledem k velikosti contracku.
- ▶ Odpověď rozlišila dva různé problémy:
- ▶ kapacita (V-cache pomůže) vs. koherence (ne).



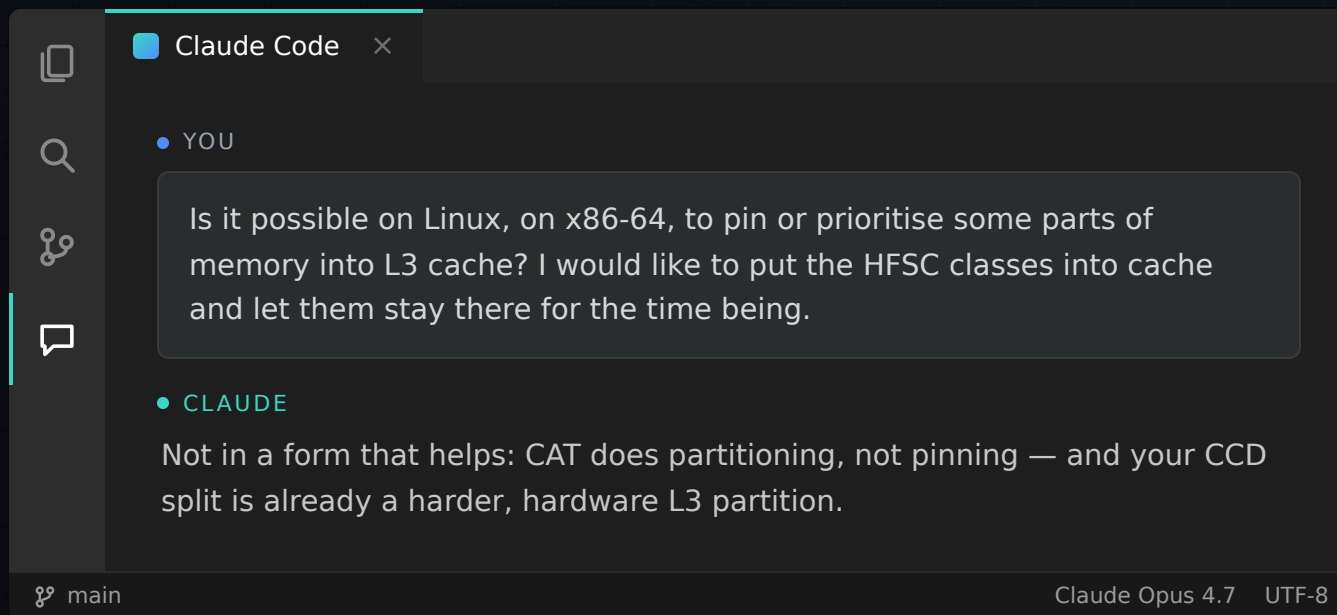
## Conntrack je otázka odolnosti, ne propustnosti

- ▶ Doplnil jsem: aktivních conntrack záznamů je málo.
- ▶ 2M tabulka je rezerva pro DDoS a edge cases.
- ▶ Tím se conntrack stal otázkou odolnosti –
- ▶ ne otázkou propustnosti v ustáleném stavu.



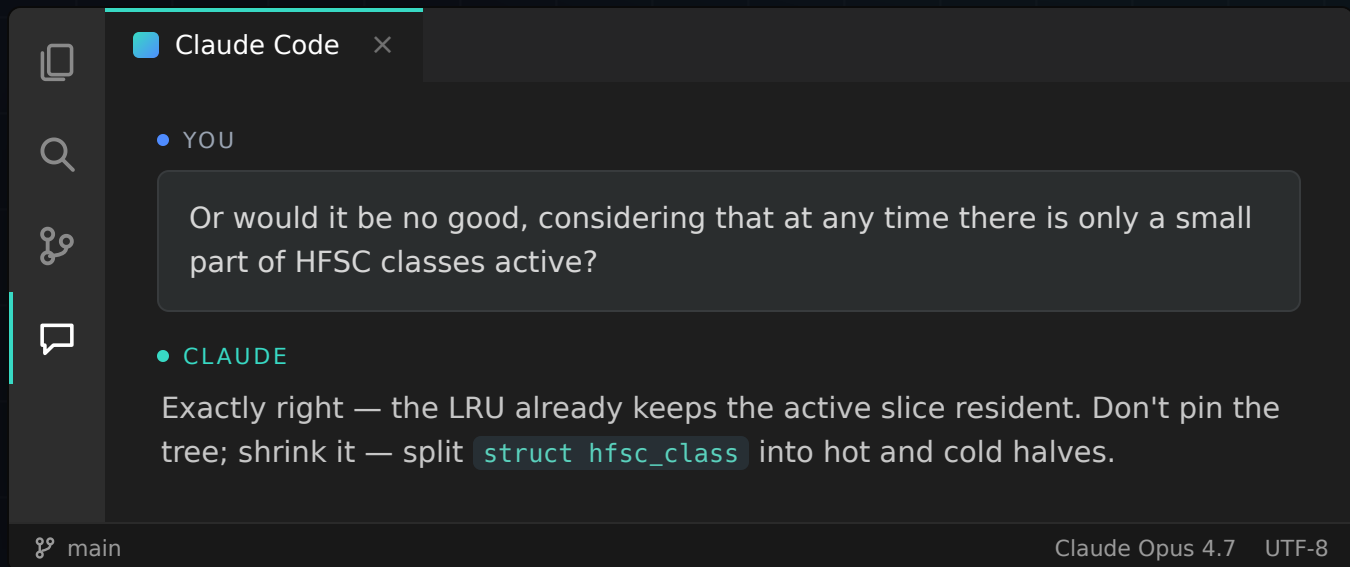
## Lze připnout paměť do L3 cache?

- ▶ Zeptal jsem se: lze připnout paměť do L3?
- ▶ Chtěl jsem držet HFSC třídy v cache natrvalo.
- ▶ Odpověď: CAT je partitioning, ne pinning.
- ▶ A CCD split je už tak tvrdší HW partition.



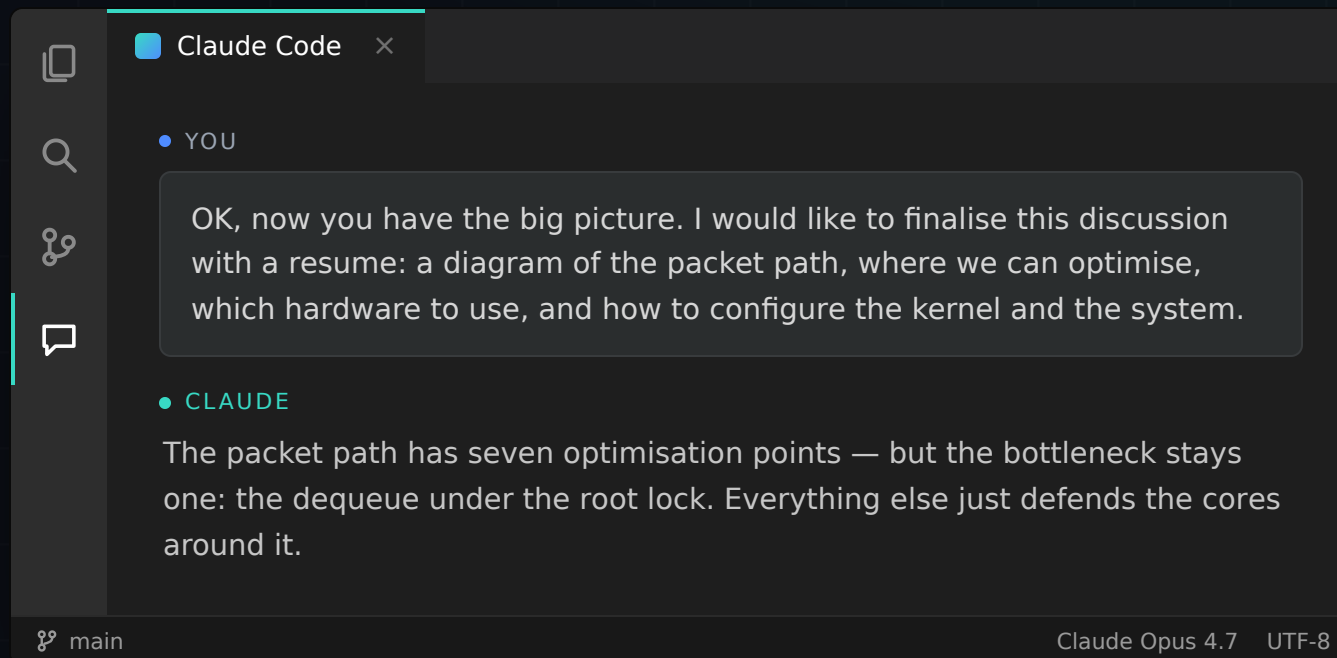
## Aktivní je vždy jen malá část tříd

- ▶ Sám jsem nadhodil: aktivní je jen část tříd.
- ▶ A to byla správná intuice.
- ▶ L3 jako hardwarová LRU drží horké třídy sama.
- ▶ Řešení není připínat – ale zmenšit working set.



## Závěr: cesta paketu a kde optimalizovat

- ▶ Na závěr jsem požádal o shrnutí konverzace.
- ▶ Diagram cesty paketu a optimalizační body.
- ▶ Hardware, konfigurace kernelu i systému.
- ▶ Bottleneck ale zůstává: dequeue pod zámkem.



**Děkuji za pozornost,  
Happy shaping!**