Distribuovaná NVME storage v hyperkonvergované DC infrastruktuře bez dopadu na běžný provoz

Pavel Mráček, Tomáš Procházka

SEZNAM.CZ

Prostředí a HW

Openstack

- Samostatné clustery - 3 DC => 12k serverů

  KO: 5c / 2175w, **OA:** 3c / 1436w, **NG:** 3c / 409w

  - Open Compute Project (OCS) blade:

  CPU:

  - 2x socket Intel (Xeon Silver 4114) => 20 cores [@40]
  - 1x AMD EPYC (7402p / 7443p) => 24 cores [@48]

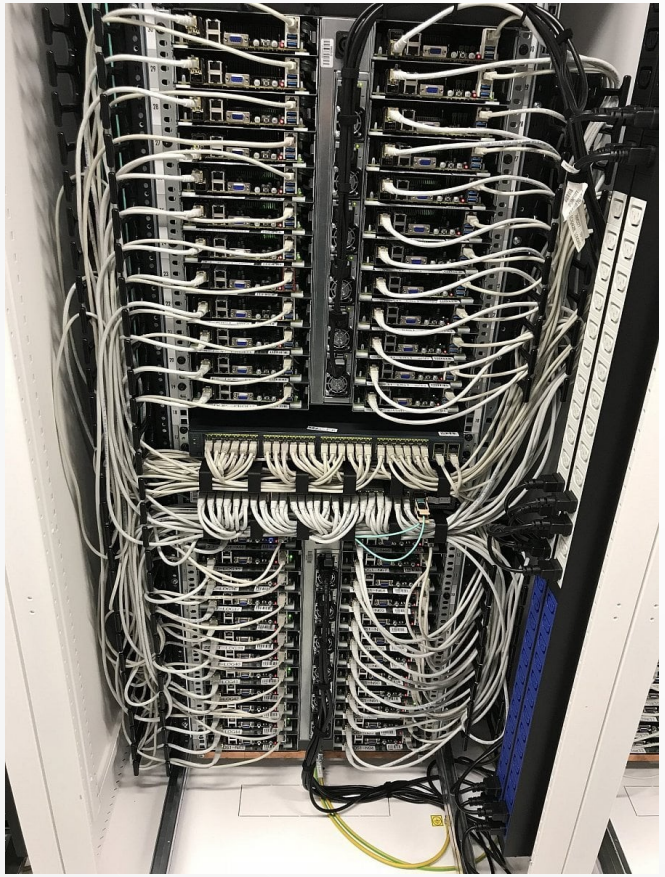
  RAM: 192GB / 256GB

  NIC: 10Gb / 25Gb

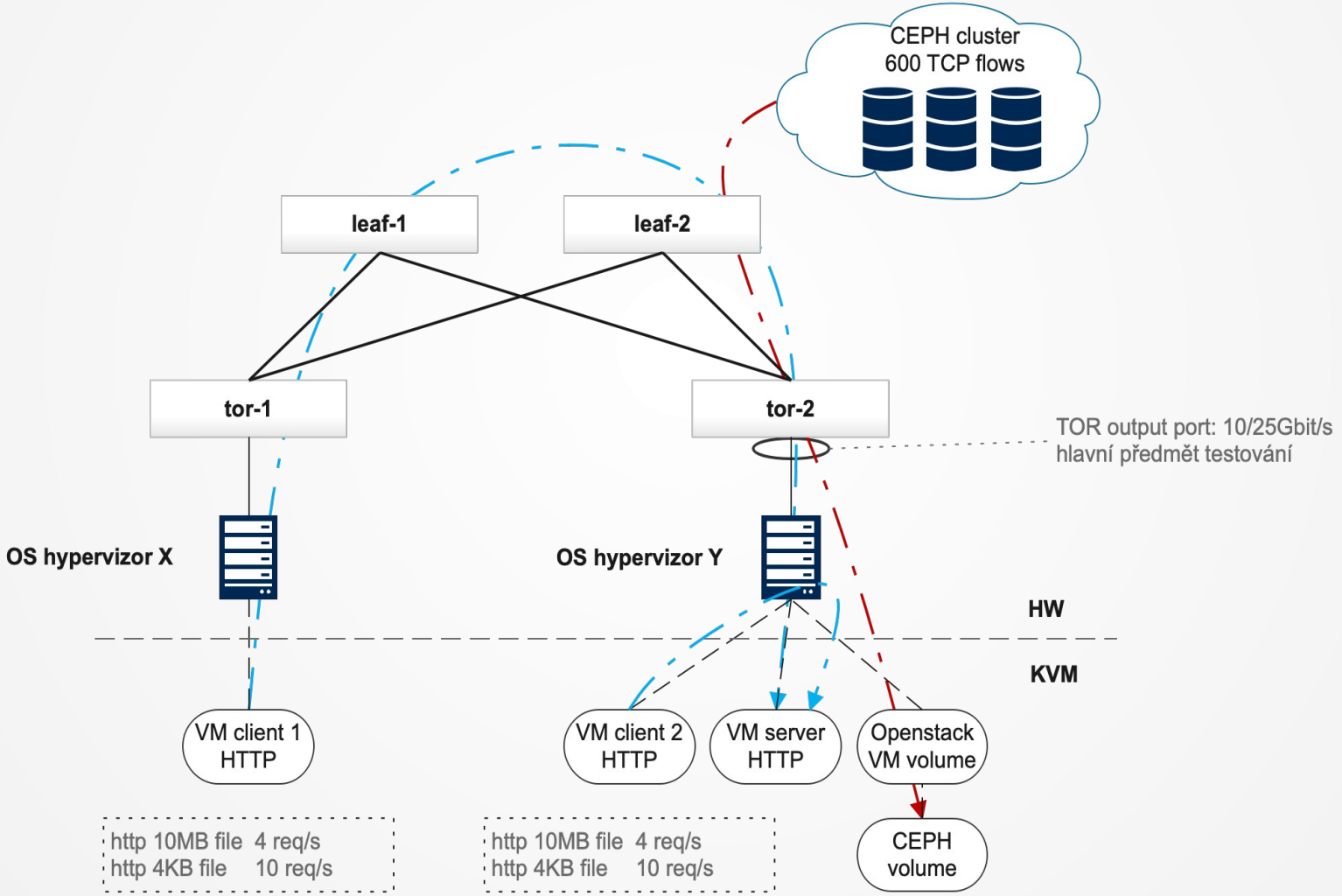
  Storage:

  - Cca ½ workeru osazena 6x (1|2|4)TB NVME

Openstack - HW

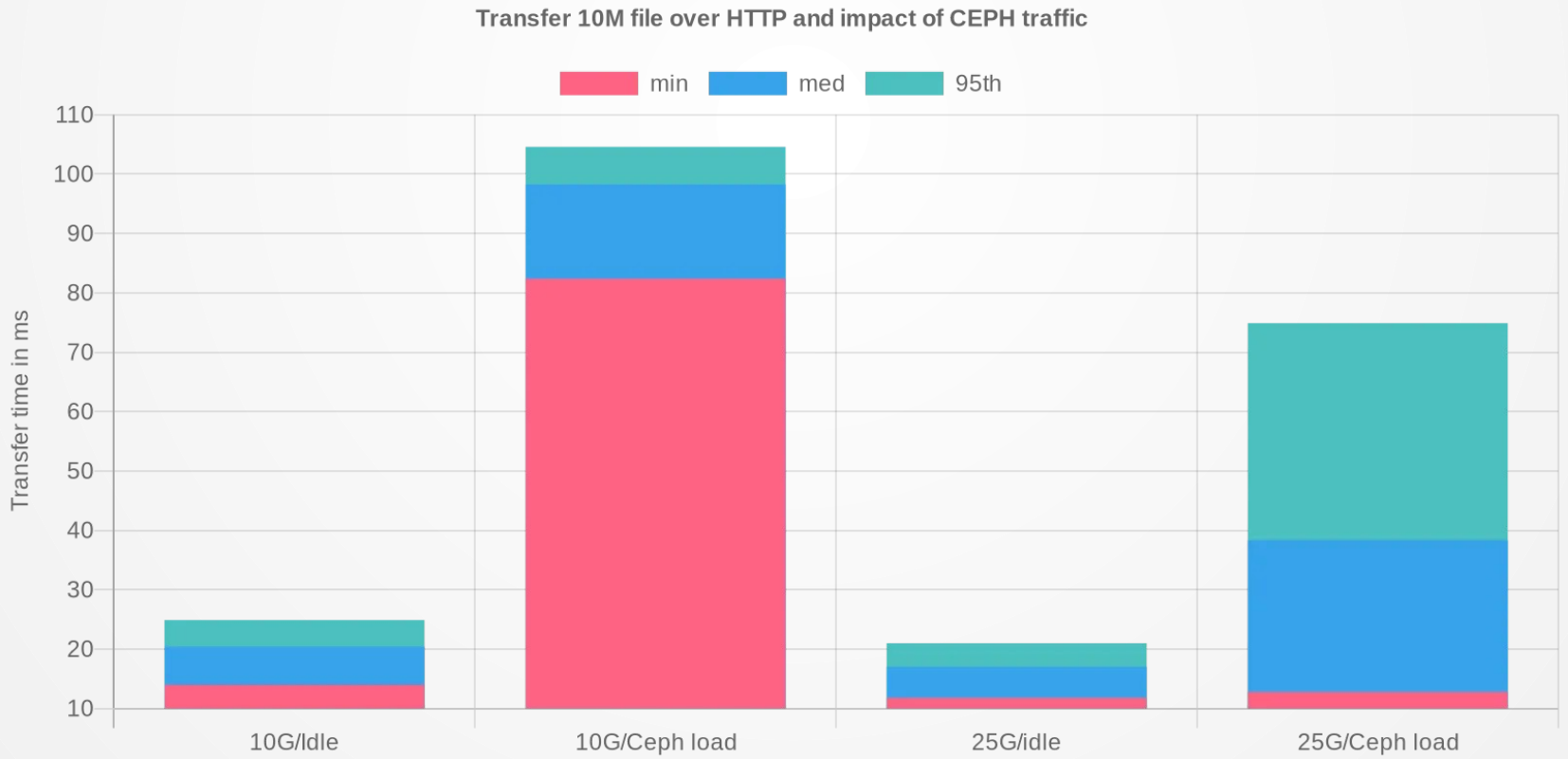


Jak velký problém může neřízený provoz být?

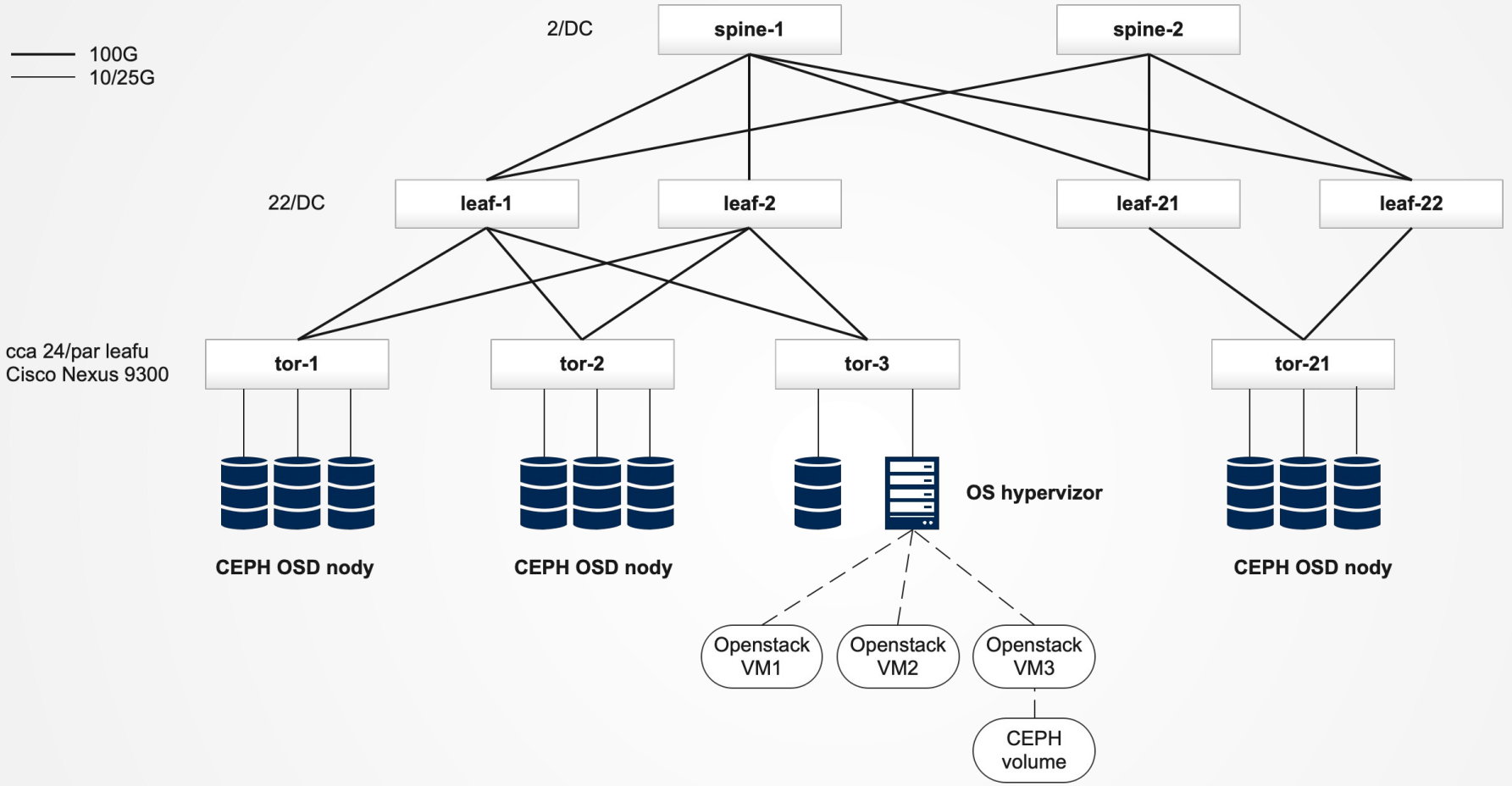


Dopad souběhu neřízeného provozu

- Provoz celého DC se sbíhá na portu k workeru.
- Odchozí port je tedy to úzké místo kde je potřeba provoz řídit.
- Storage bez limitace umí přicpat i 25Gb/s port a na 10Gbit/s je to ještě horší.



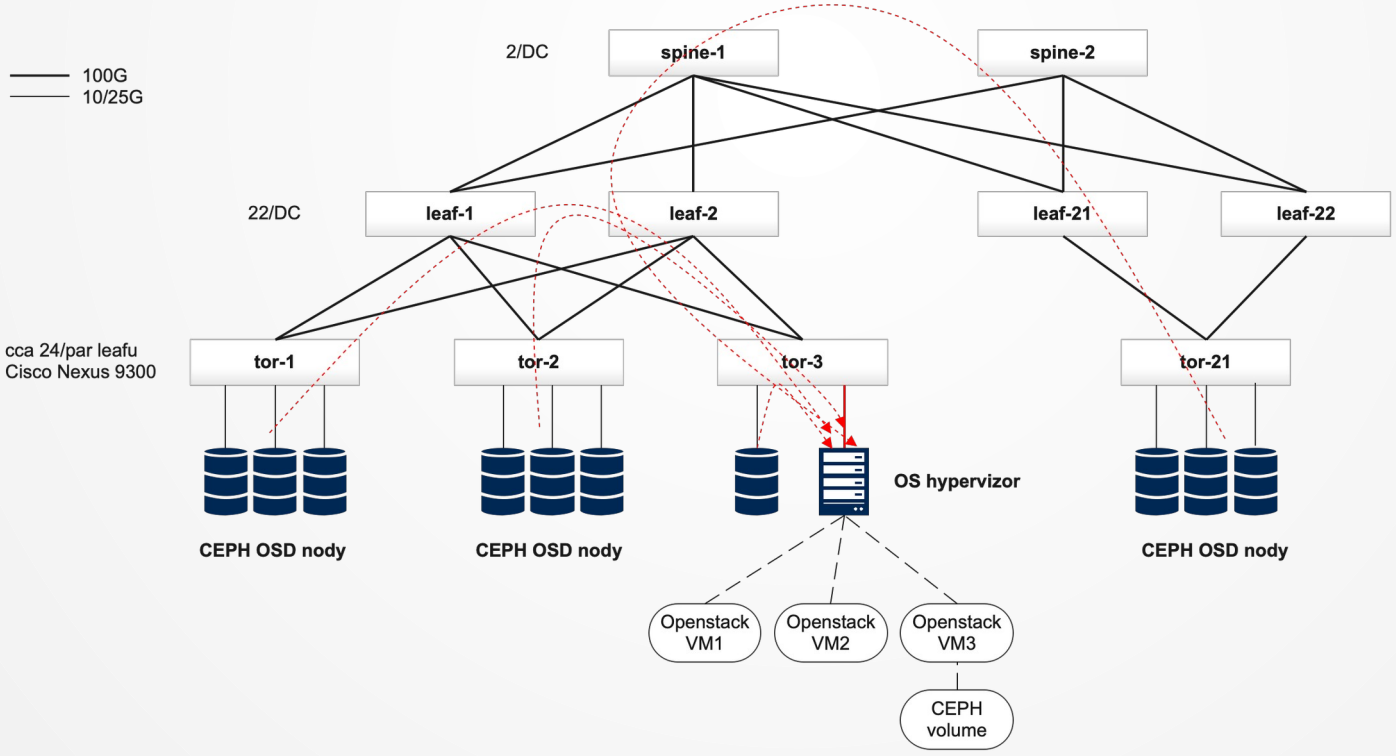
Výchozí stav DC infrastruktury



- Infrastruktura, jeji vytížení a redundance, QoS

CEPH storage výzva

- Komunikace N:1, ucpaný port hypervizoru – vysoké latence
- Datově velké toky (elephants) sdílí výstupní queue na portu s malými, krátkými toky (mice) – drop obou typu provozů stejnou mírou
- Řešení? QoS, kapacita, storage network?



Jak z toho ven?

- Queueing a scheduling na portech k serveru + congestion avoidance mechanismy na N9K s NXOS
- ECN (Explicit Congestion Notification)
  - Notifikace při zahlcení portu v cestě datového toku, neprovede se drop
  - Výhoda pro mice toky (control messages, query, responses)
  - ECN na N9K/NXOS – použij WRED nebo AFD (Cisco proprietary)

- WRED a ECN
  - Po překročení min thresholdu dojde k označení
  - O zahlcení portu ví switch – provede notifikaci
  - Používají se dva bity v ToS poli v IP hlavičce packetu
  - Musí podporovat koncové systémy (klient, server) a switch

ECT Bit	CE Bit	význam
0	0	neumí ECN
0	1	ECN podporováno
1	0	ECN podporováno
1	1	detekováno zahlcení

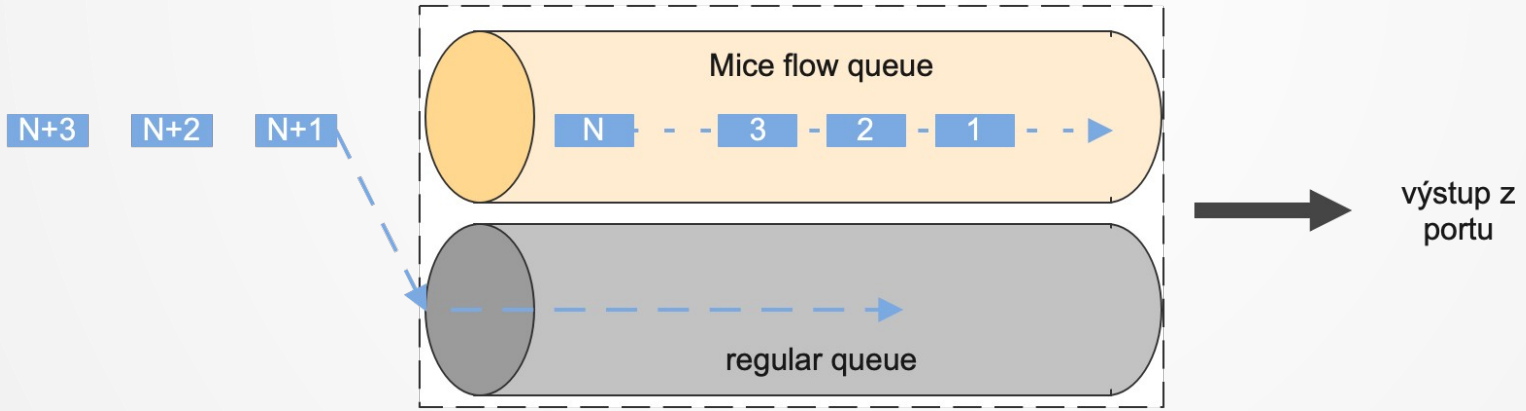
- AFD (Approximate Fair Drop)

  - AQM, udržuje část bufferu volného pro mice toky
  - rozlišuje (na základě tzv ETRAP) velké datové toky (elephant) a malé (mice)
  - Mice toky neřeší, Elephant zahazuje – dle četnosti/frekvence packetu na vstupu do fronty
  - ETRAP definuje Elephant toky (podle počtu byte, délky toku a objemu dat) a registruje je do tzv ETRAP flow tabulky
  - také lze využít ECN
  - definice ETRAP

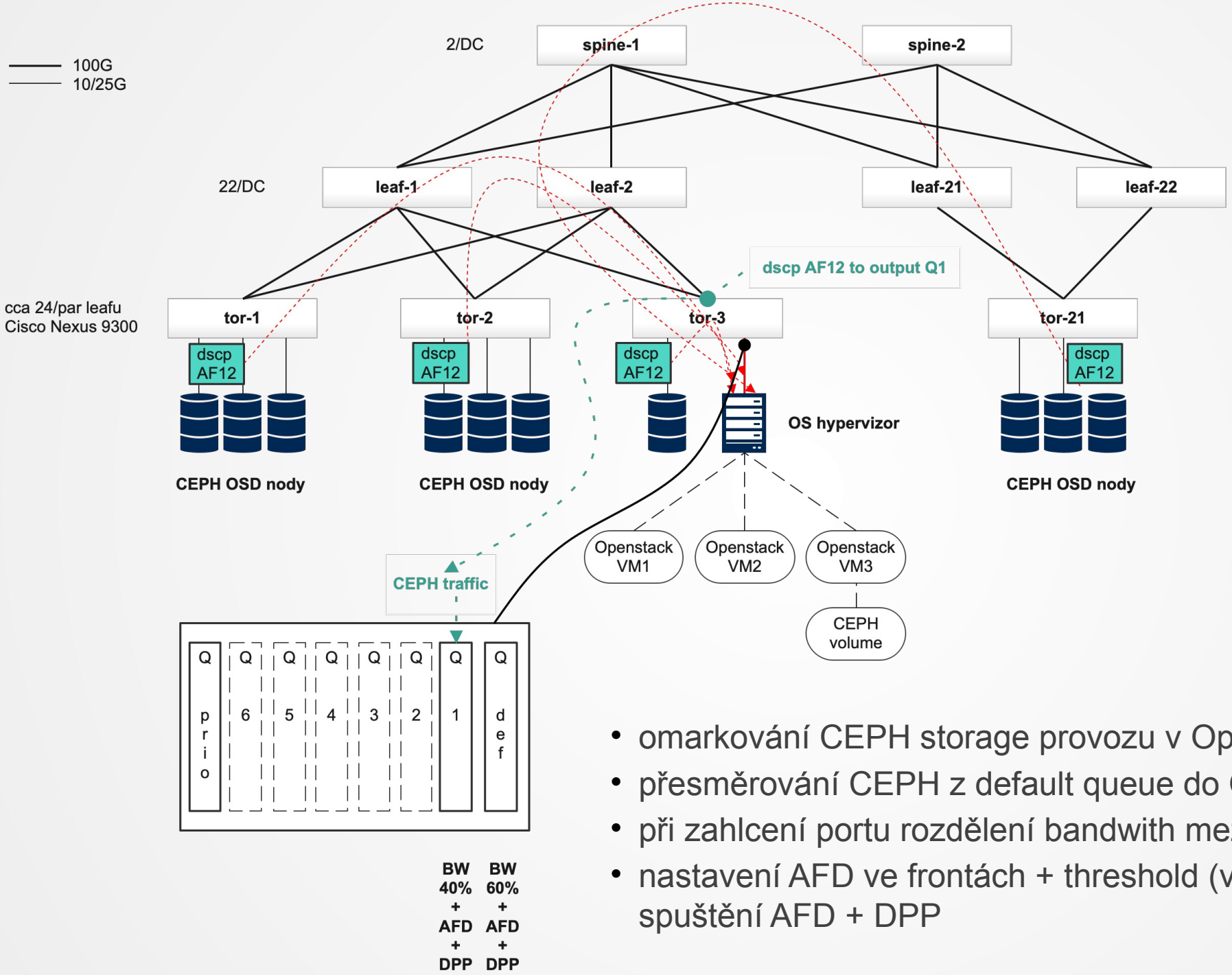
```
hardware qos etrap age-period 20 usec
hardware qos etrap byte-count 350000
hardware qos etrap bandwidth-threshold 170 bytes
```

- # DPP (Dynamic Packet Prioritization)

  - Maximální “akcelerace” malých, krátkých toků například přes prioritní frontu
  - Umožňuje poslat prvních N packetů z flow do jiné výstupní fronty než-li danému flow náleží
  - Po dosažení limitu N jdou další packety flow zařazeny do fronty, ve které jsou běžně odbavovány



Řešení

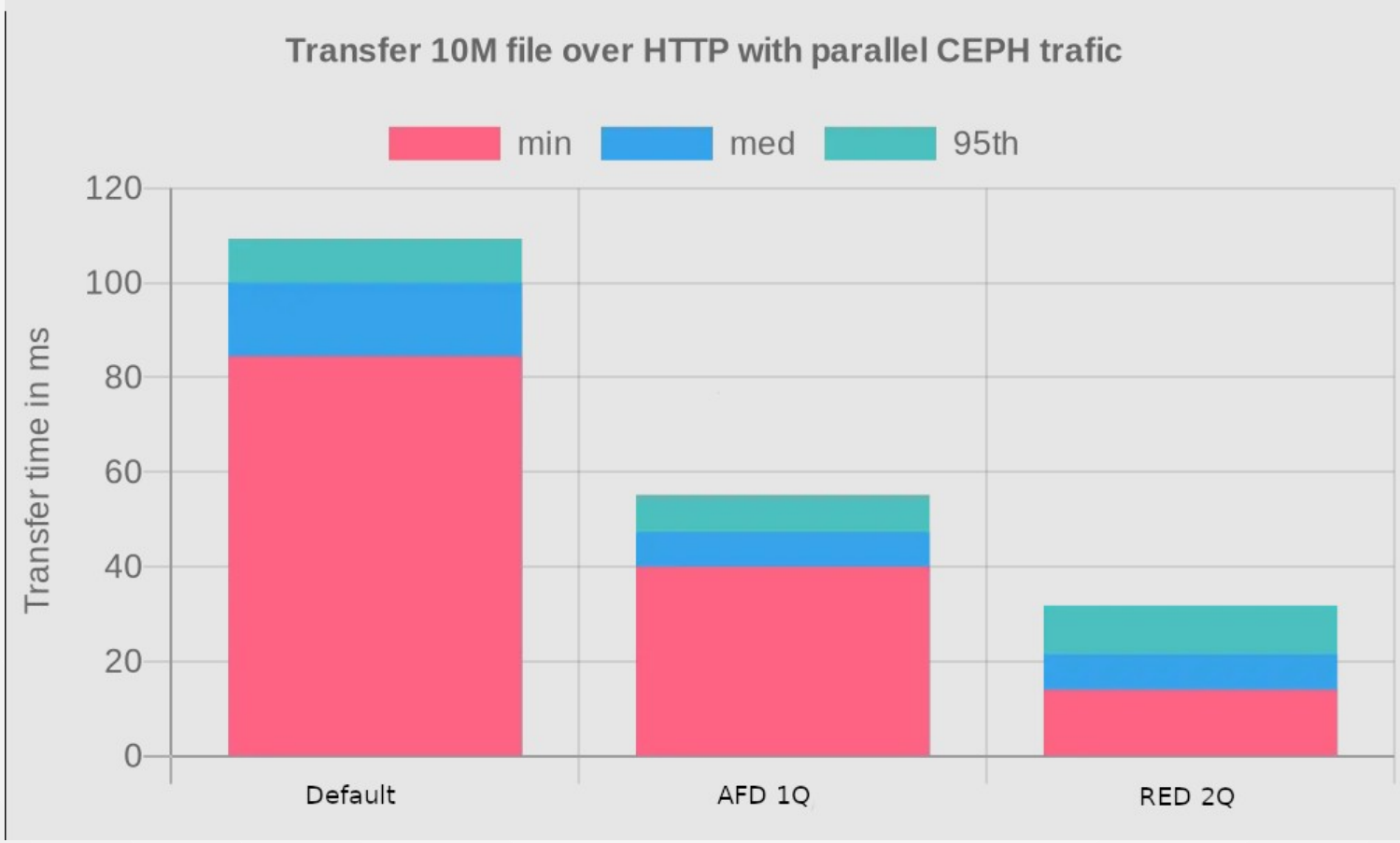


- omarkování CEPH storage provozu v Openstacku
- přesměrování CEPH z default queue do Q1 – na základě DSCP
- při zahlcení portu rozdělení bandwith mezi Q1 a Q def
- nastavení AFD ve frontách + threshold (velikost fronty) pro spuštění AFD + DPP

Výsledky

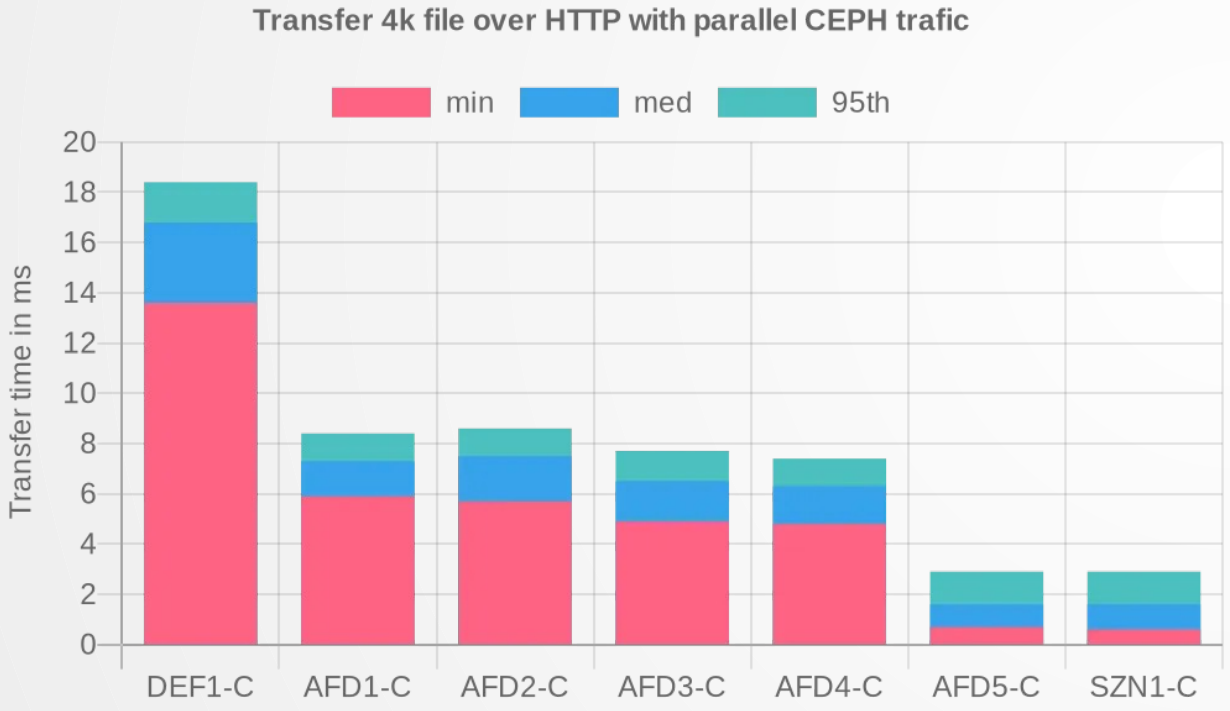
AFD 1Q vs RED 2Q

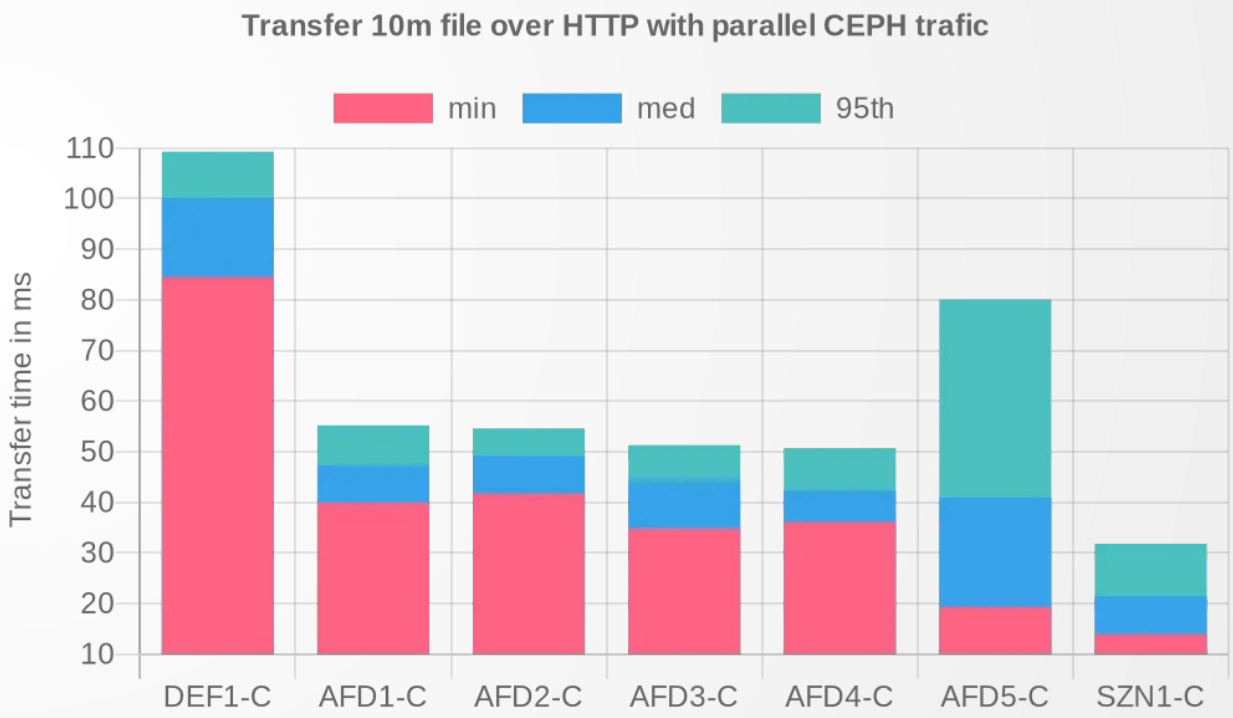
- Použití jen AFD by bylo implementačně snadné a univerzalní
- Ale rozdělení do 2 queues (RED) dávalo výsledky podobné nezatíženému portu!



AFD 1Q + tunning vs AFD 2Q

- Chování AFD jde ladit
- Nejlepší výsledky ale pořád dává rozdělení do více queues





AFD1-4 - různé parametry AFD 1Q
AFD5 - AFD + DDP
SZN1 - AFD 2Q

Jak rychlé to může být?

DISK random read	1M Q64	1M Q1	4k Q64	4k Q1
ssd mikron R1	1070MB/s (1k)	381MB/s (0.36k)	374MB/s (91k)	19MB/s (4.6k)
ssd intel R1	1108MB/s (1k)	291MB/s (0.27k)	503MB/s (123k)	35MB/s (8.5k)
nvme micron R1	**3750MB/s** (3.5k)	900MB/s (0.85k)	326*-650**MB/s (80-158k)	17-32MB/s (4-7.6k)
RBD replicated	**2540MB/s** (2.5k)	226MB/s (0.2k)	212MB/s (52k)	4MB/s (1k)

DISK random write	1M Q64	1M Q1	4k Q64	4k Q1
ssd mikron R1	523MB/s (0.5k)	470MB/s (0.45k)	245MB/s (60k)	34MB/s (8k)
ssd intel R1	497MB/s (0.47k)	413MB/s (0.4k)	238MB/s (58k)	44MB/s (11k)
nvme micron R1	1050MB/s (1k)	550MB/s (0.5k)	80*-500MB/s** (20-130k)	65MB/s (15k)
RBD replicated	**2900MB/s** (2.7k)	117MB/s (0.1k)	102MB/s (25k)	1.7MB/s (0.4k)

Hodnoty v zavorce jsou IO/s
*) Xeon(R) Silver 4210R
**) AMD EPYC 7443P

Díky za pozornost