SSD vs. HDD / WAL, indexy a fsync

Prague PostgreSQL Developers Day 2012

Tomáš Vondra (tv@fuzzy.cz)

What a great day for science!

Otázky

DB = data + indexy + transakční log (WAL)

- Co umístit na samostatný disk?
- Co patří na HDD a co na SSD?
- Jakou fsync metodu zvolit?
- Jak to závisí na typu zátěže?

Charakteristiky disků

7.2k SATA spinners

- ~ 120 IOPS
- sekvenčně: ~ 100 MB/s
- náhodně: ~ 0.5 MB/s
- 512B / 4kB sektory, běžně kapacita několik TB

SSDs (Intel 320 120GB)

- (deseti)tisíce IOPS
- čtení: ~ 250 MB/s
- zápis: ~ 120 MB/s
- 512kB bloky, kapacita desítky / stovky GB

Consumer-level ...

- jen „obyčejné“ SATA-II disky
- demonstrace obecných I/O charakteristik
- aplikovatelné na srovnatelný hw
  - 15k SAS disky, podobné SATA SSD disky, …

- obtížnější aplikace na
  - specializovaná SSD (fusion-io, VeloDrive/Z-Drive, ...)
  - RAID pole s dobrým řadičem (write cache)
  - výrazně odlišný workload

Benchmarky

OLTP (pgbench)

- hodně „malých“ transakcí (typicky přes primární klíče)
- počet transakcí za vteřinu (tps)
- read-only nebo read-write (50:50, netypický poměr)

TPC-H

- velké dotazy nad velkým objemem dat
- počet vteřin (trvání loadu, vyhodnocení dotazu)
- GROUP BY, velké joiny, … (i hodně náhodného I/O)

data >> RAM

Kdybych měl tolik peněz

jako Bruce Wayne ...

Dosahovat výkon za hromady

peněz není velké umění.

výkon / cena

výkon / (cena / GB)

$ / GB

HDD (Maxtor DiamondMax 21)

- 500 GB SATA II 7.2k ~ 2000 Kč => **4 Kč / GB**

SSD (Intel 320)

- 120 GB SATA II ~ 4200 Kč => **40 Kč / GB**

kombinace (naivní model)

- vážený součet podle objemu (indexy, WAL, data, ...)
- 20% HDD + 80% SSD => **33 Kč / GB**
- 80% HDD + 20% SSD => **11 Kč / GB**

výkon / (cena / GB)

pgbench na HDD

- výkon 120 tps
- cena 4 Kč / GB
- metrika = 120 / 4 = 30

pgbench na SSD

- výkon 4000 tps
- cena 40 Kč / GB
- metrika = 4000 / 40 = 100

100 > 30

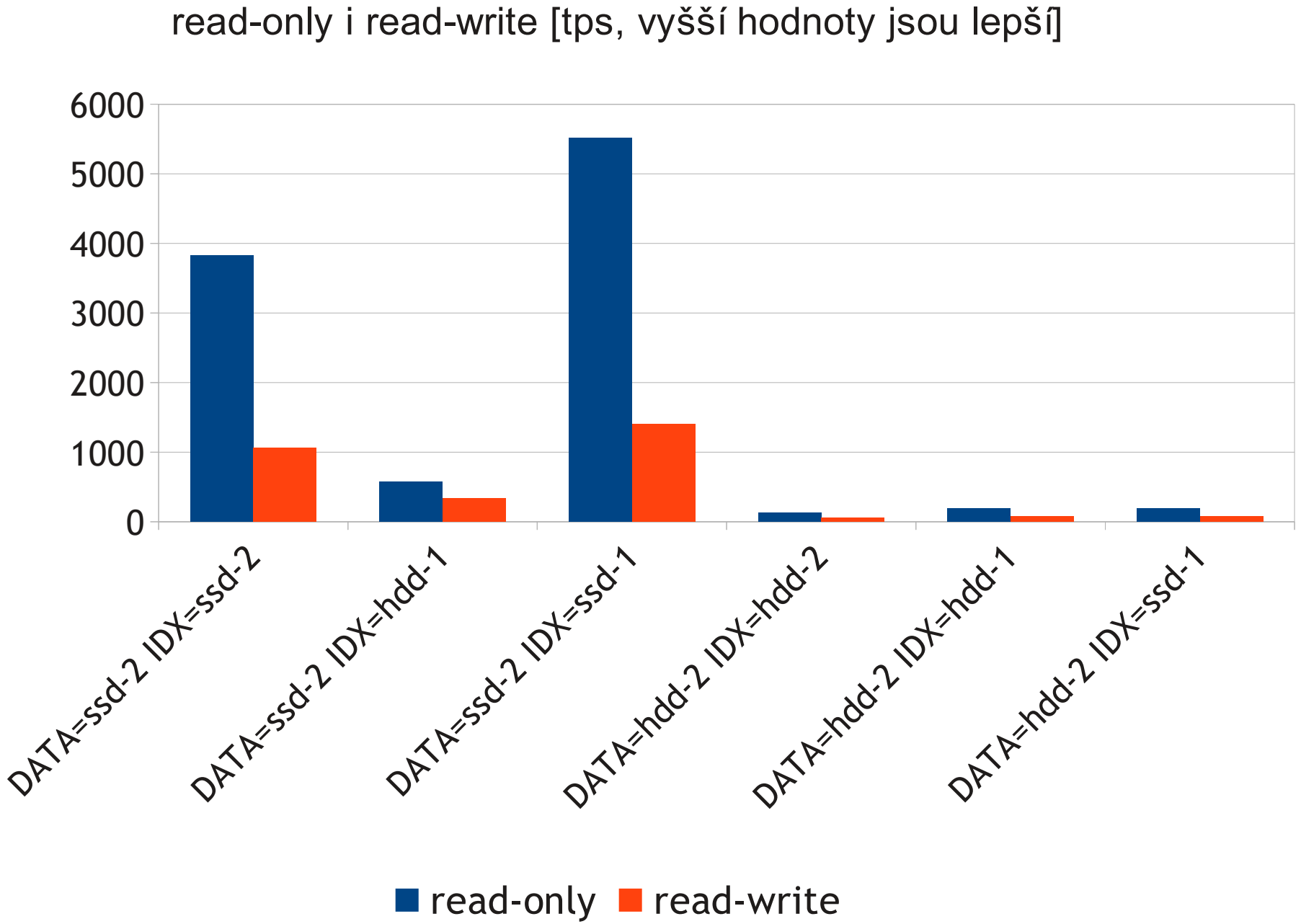
78,31% všech statistik je vymyšlených

Důležitá jsou vaše čísla!

data vs. indexy

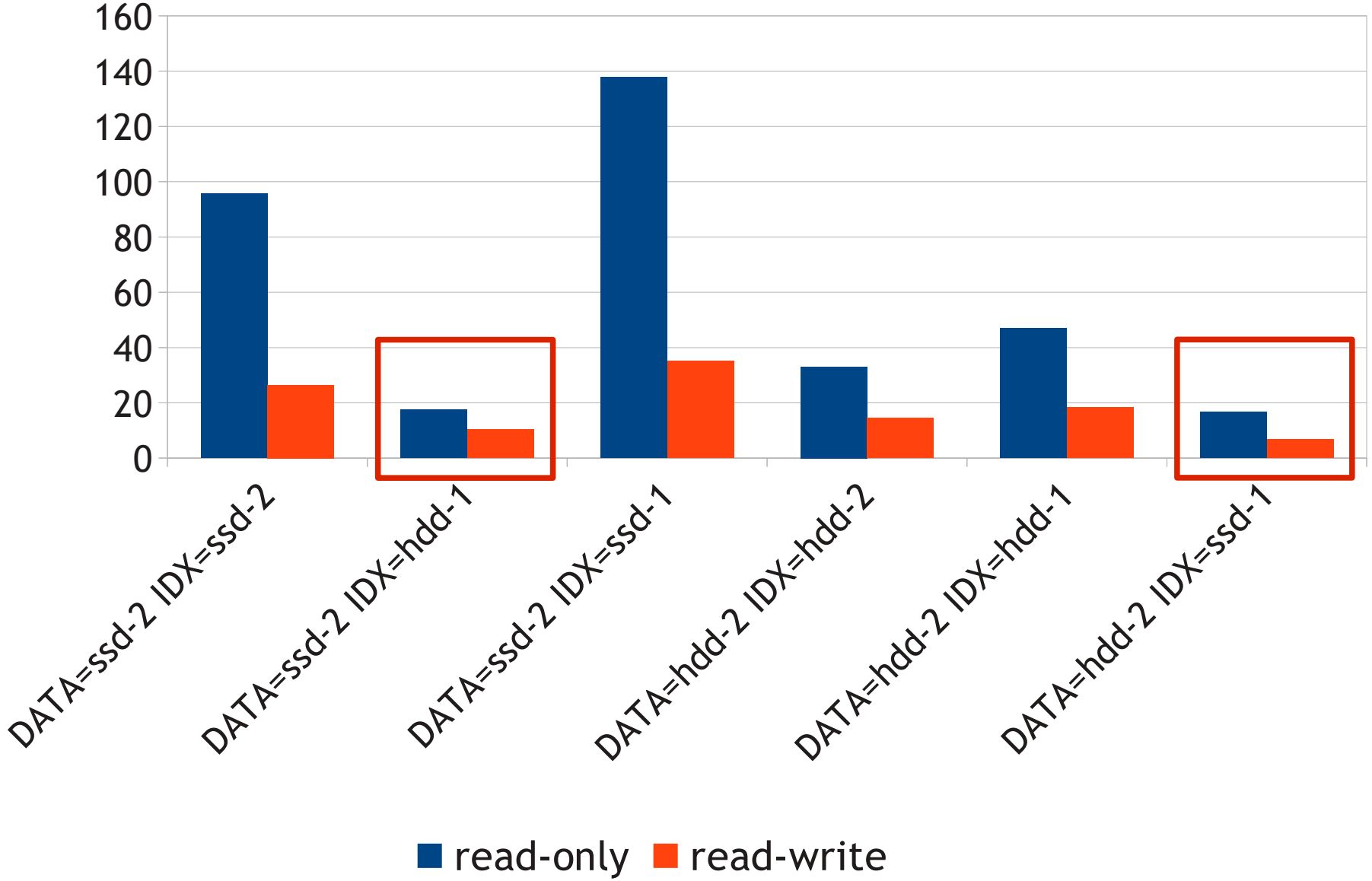
OLTP (pgbench)

pgbench s oddělením dat a indexů



pgbench s oddělením dat a indexů (80:20)

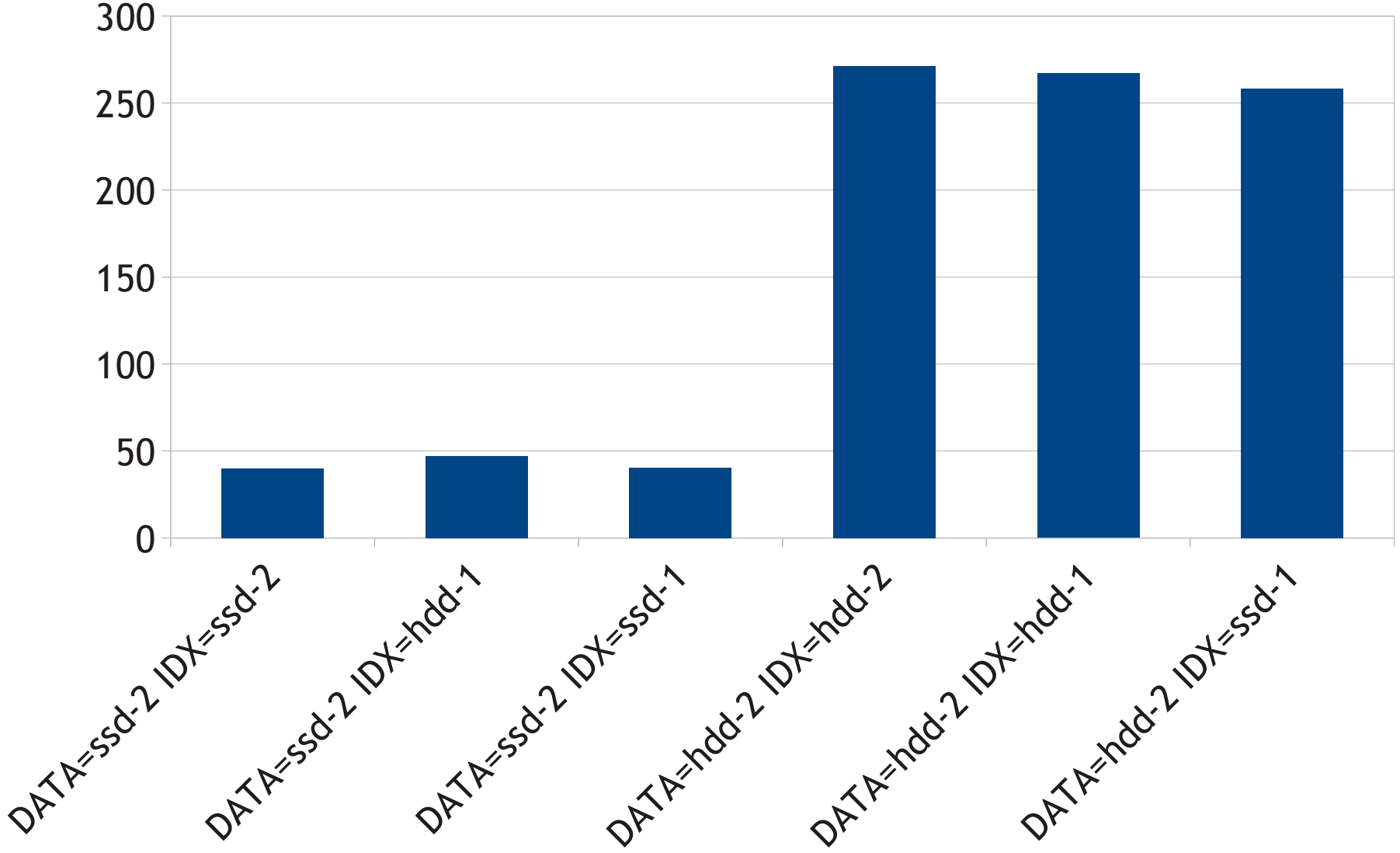
read-only i read-write [výkon/$, vyšší hodnoty jsou lepší]



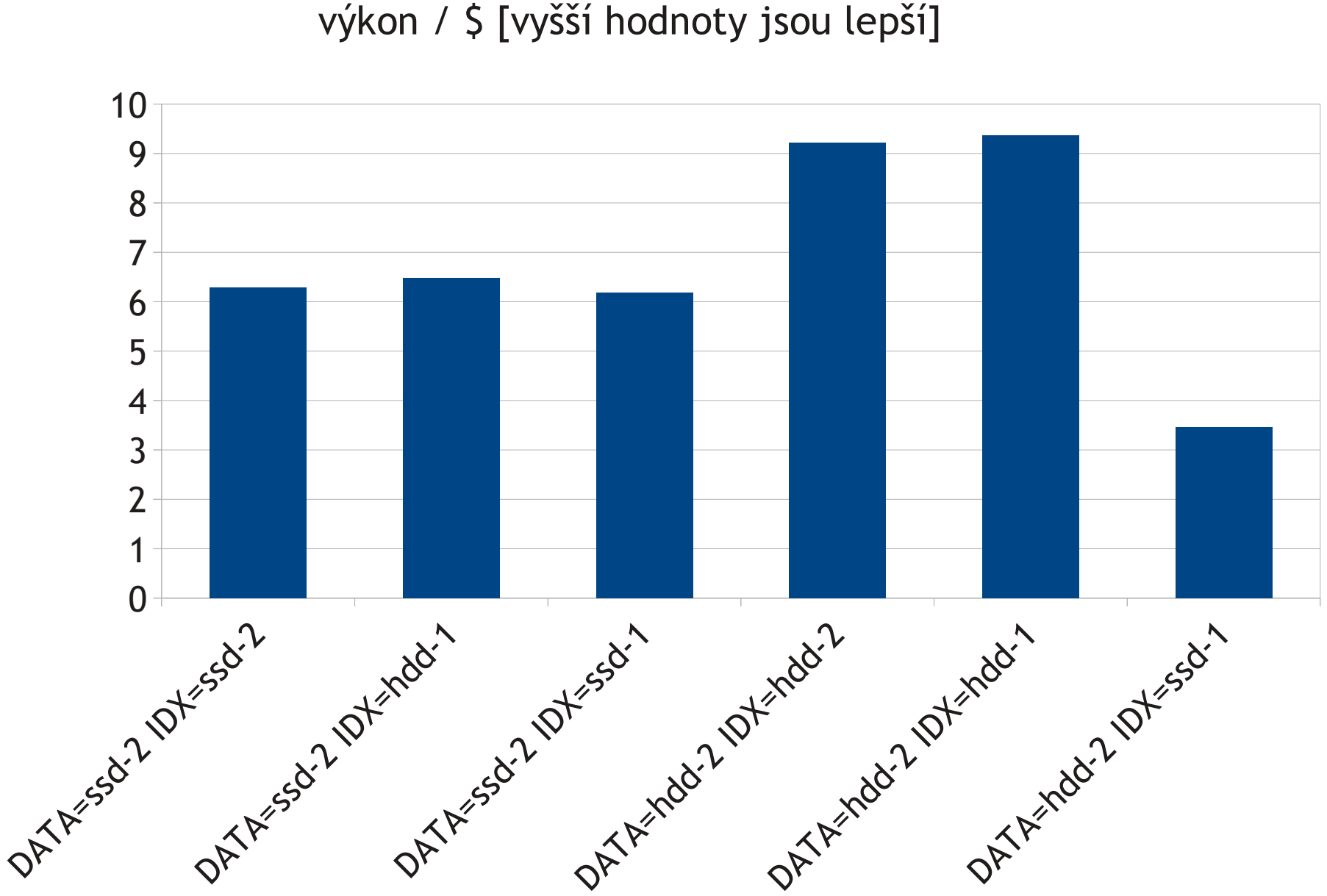
DSS/DWH (TPC-H)

TPC-H výkon / dotaz č. 17

počet vteřin na vyhodnocení [nižší hodnoty jsou lepší]



TPC-H výkon / dotaz č. 17 (80:20)



data vs. indexy

SSD

- výborná investice pro OLTP zátěž (pgbench)
- ne tak dobré pro DWS/DSS-like workload

HDD

- OLTP výkon nemůže konkurovat SSD (ani výkon/cena)
- v DWH často zdatně konkuruje SSD

kombinace SSD a HDD

- nic moc (když musíte, tak na SSD dejte data)

data vs. WAL

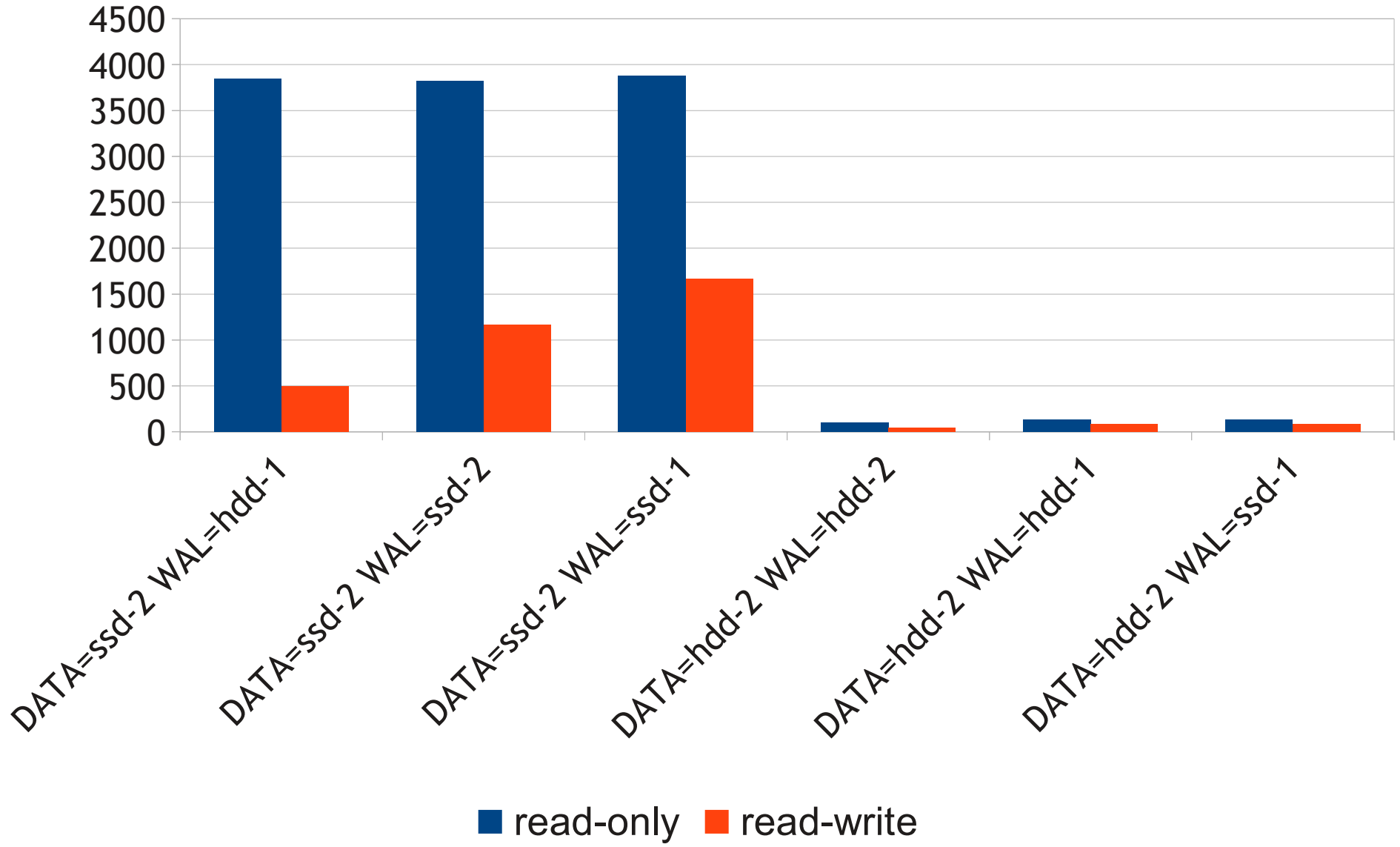
Write Ahead Log

- aka transakční log
  - aka XLog
  - aka REDO
- informace o modifikacích dat
- životně důležité pro recovery
- při dotazování se nepoužívá

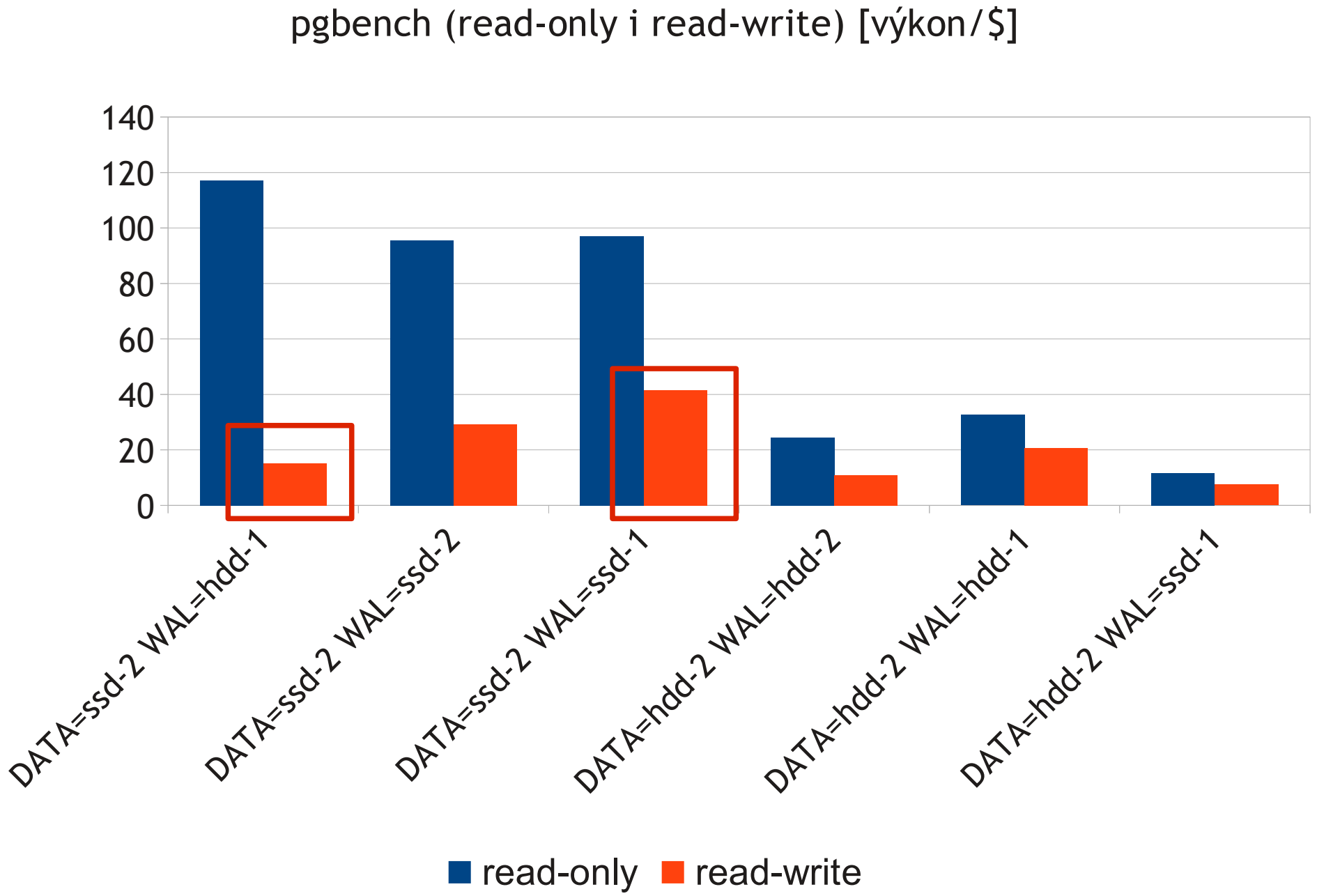
OLTP (pgbench)

oddělení dat a WAL

pgbench (read-only i read-write) [počet transakcí za vteřinu]



oddělení dat a WAL



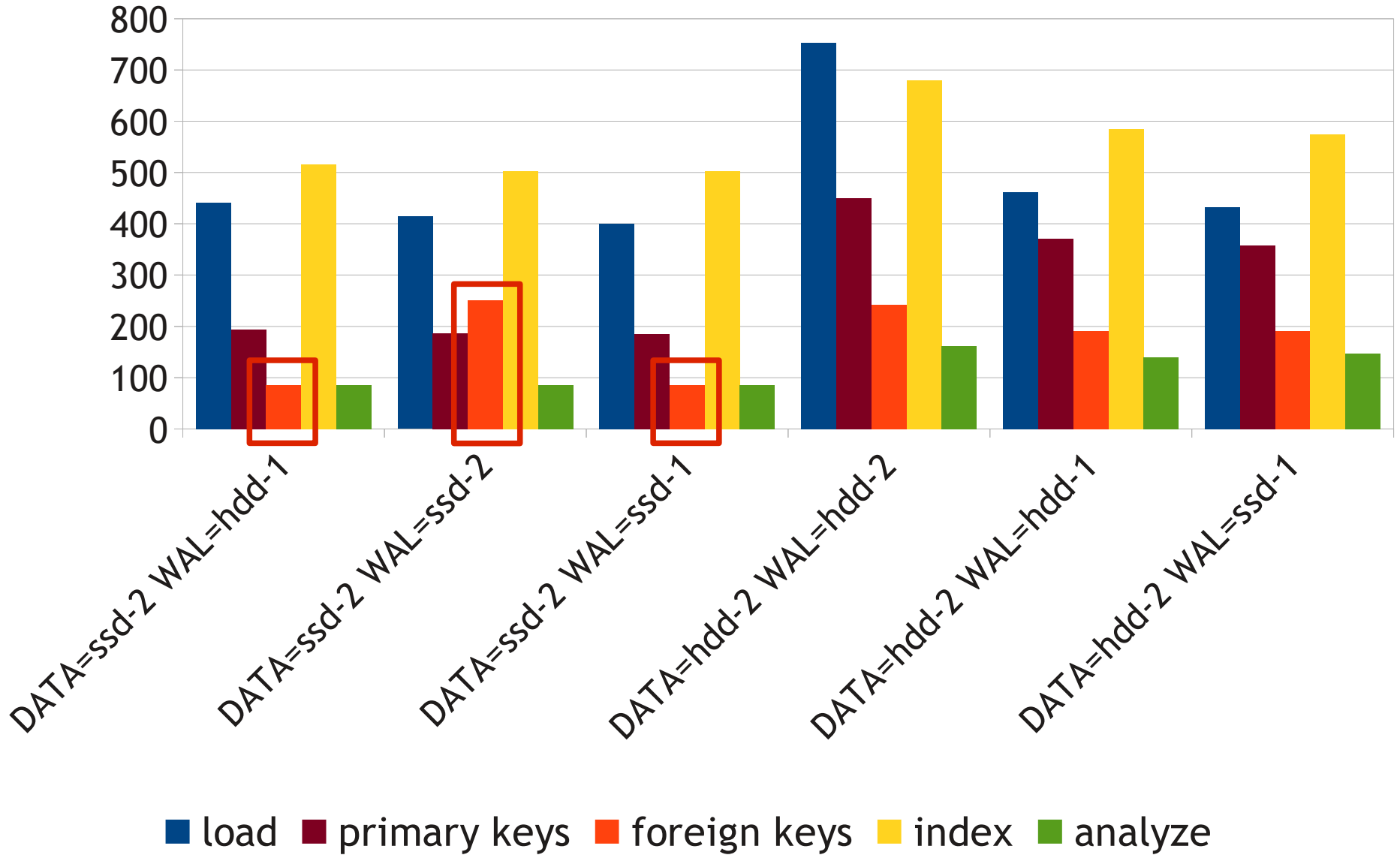
data vs. WAL / OLTP

- data – spousta náhodných přístupů => SSD
- WAL – sekvenční zápisy => HDD
- WAL i data na stejném disku => terror :-(
  - databáze musí seekovat (přístup k datům)
  - **důsledek**: zápisy do WAL pak nejsou sekvenční
- skvělá je kombinace data na SSD a WAL na HDD
  - rozdíl vs. SSD+SSD ~ rychlost sekvenčního zápisu (HDD má jen 70 MB/s, SSD má 130 MB/s)
  - s rychlejším HDD diskem (RAID) => asi ideál :-)

DSS/DWH (TPC-H)

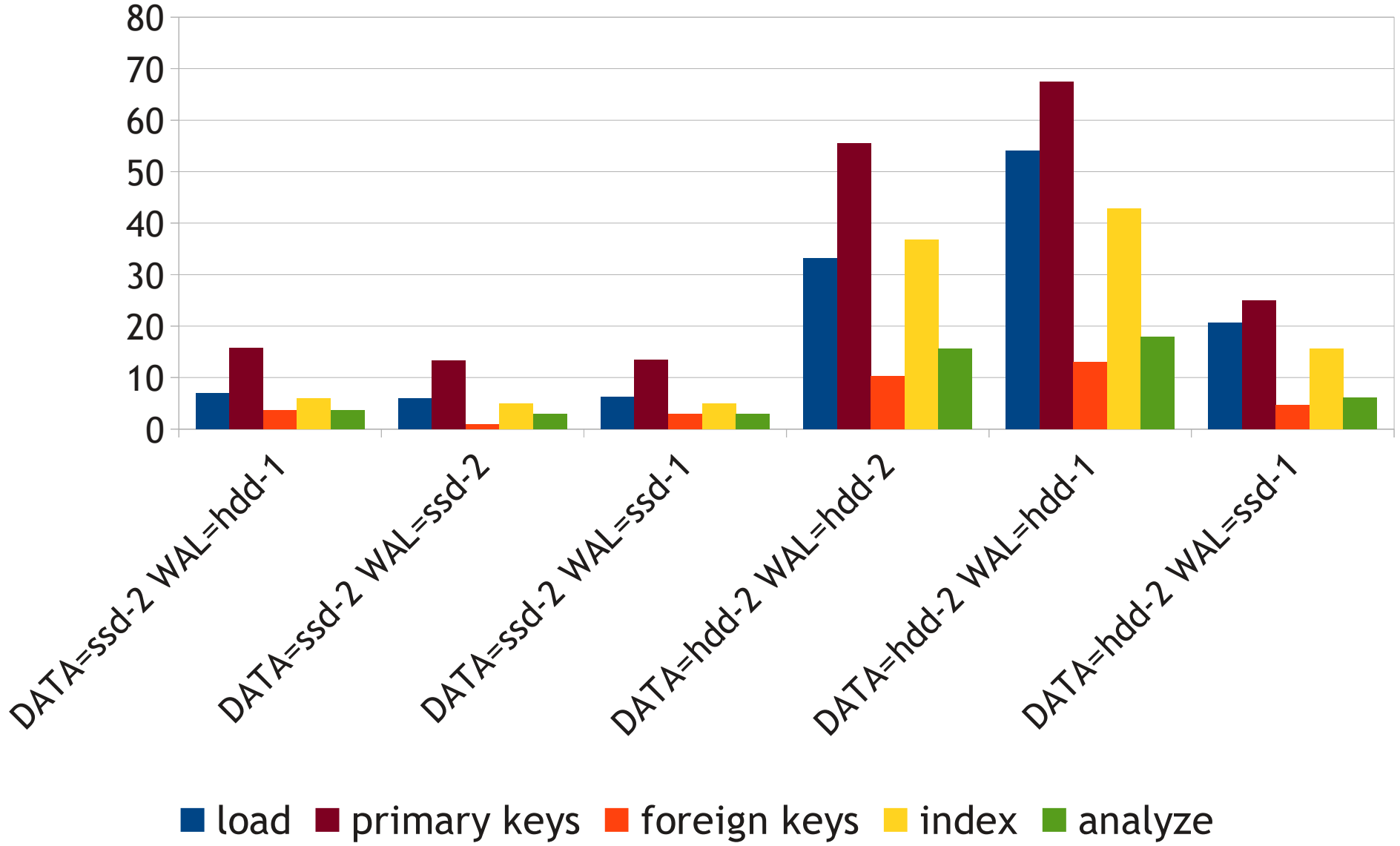
TPC-H benchmark / load

příprava dat (load, apod.) [počet vteřin, nižší hodnoty jsou lepší]



TPC-H benchmark / load (80:20)

příprava dat (load, apod.) [výkon/$, vyšší hodnoty jsou lepší]



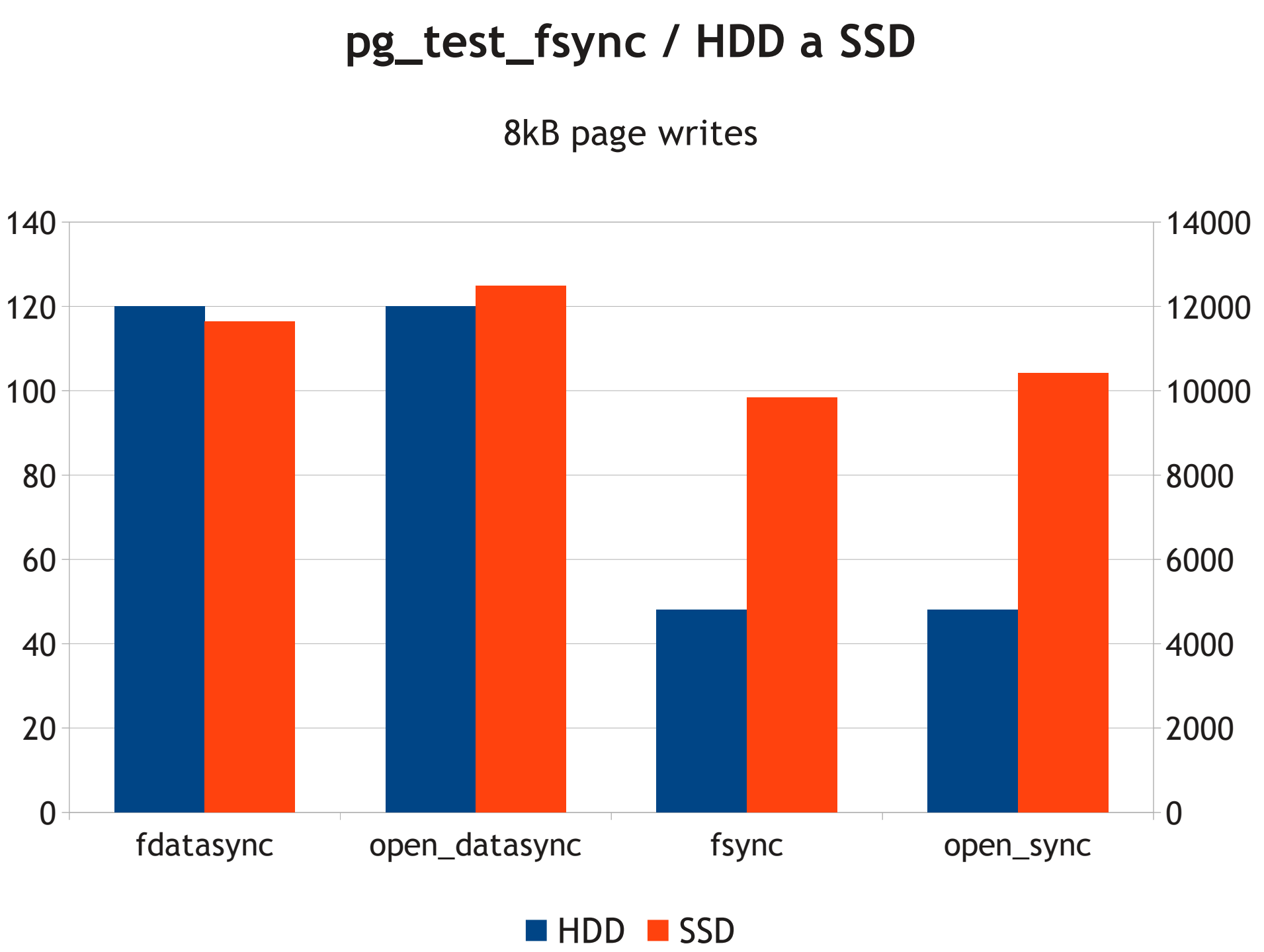
data vs. WAL / TPC-H setup

- WAL nijak neovlivňuje dotazy
- výkon/$ - pevné disky bezpečně vedou
- nižší výkon <=> nižší sekvenční rychlost
- rychlejší HDD => výkon srovnatelný se SSD

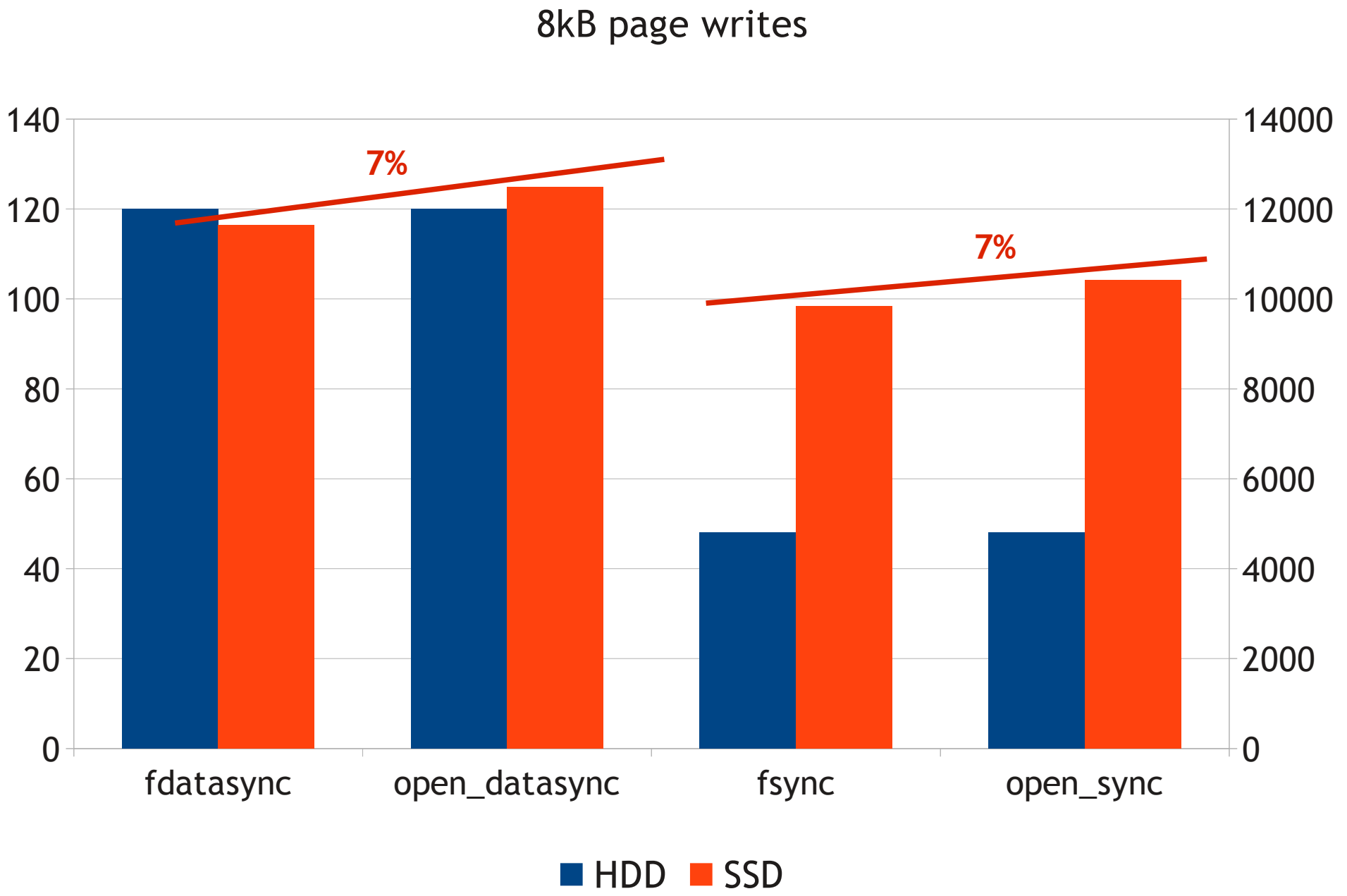
fsync metody

pg_test_fsync

- srovnání chování různých metod fsyncu
  - fsync / open_sync
  - fdatasync / open_datasync
- podpora závisí na
  - souborovém systému, verzi kernelu, …
- vztah na „direct I/O“
  - standardně „write + write + … + fsync“ (fs cache)
  - obcházení page cache (ne vždy užitečná - řadiče)


pg_test_fsync / HDD a SSD
8kB page writes
140
120
100
80
60
40
20
0
14000
12000
10000
8000
6000
4000
2000
0
fdatasync
open_datasync
fsync
open_sync
HDD
SSD

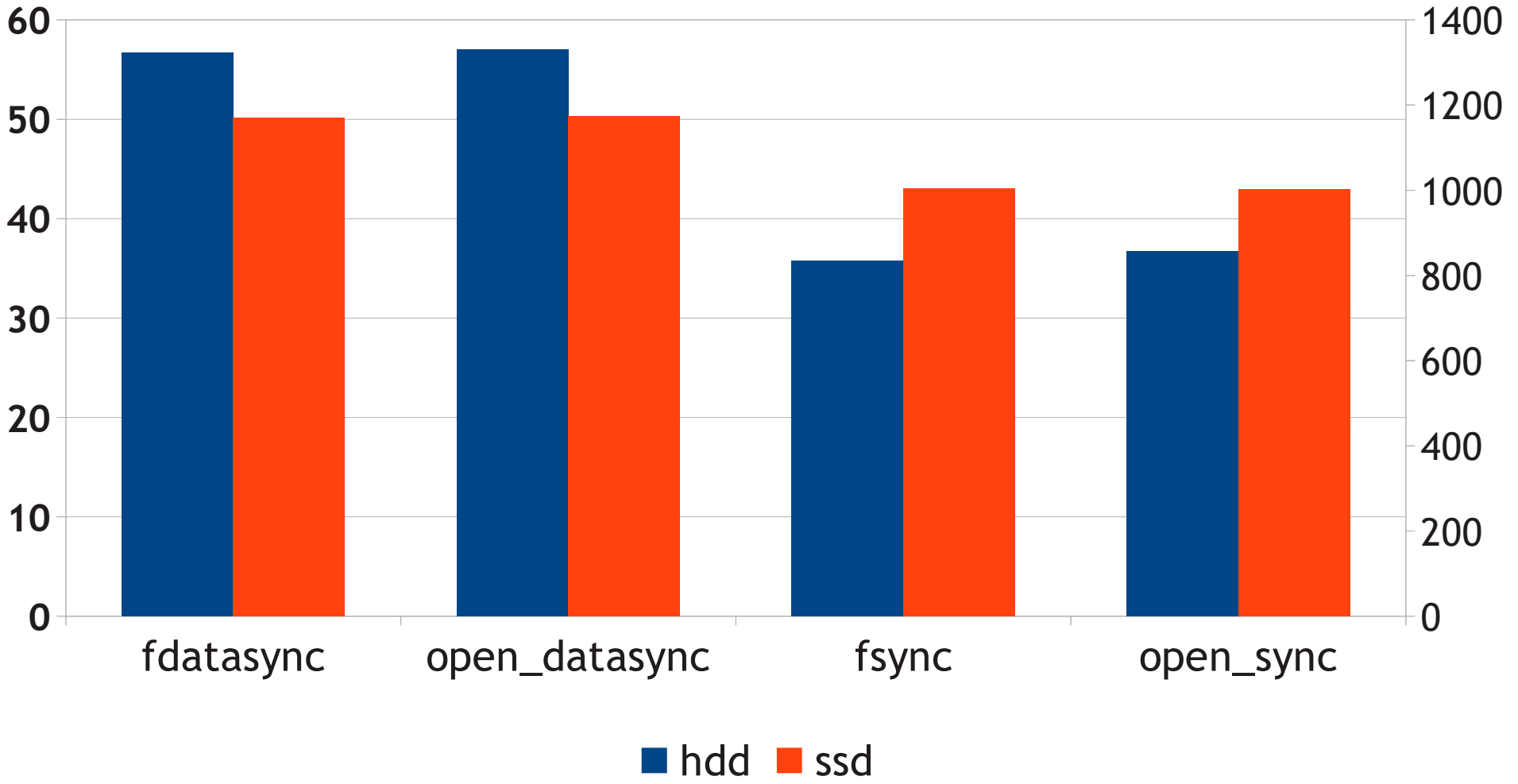
pg_test_fsync / HDD a SSD



OLTP (pgbench)

SSD vs. HDD / read-write pgbench

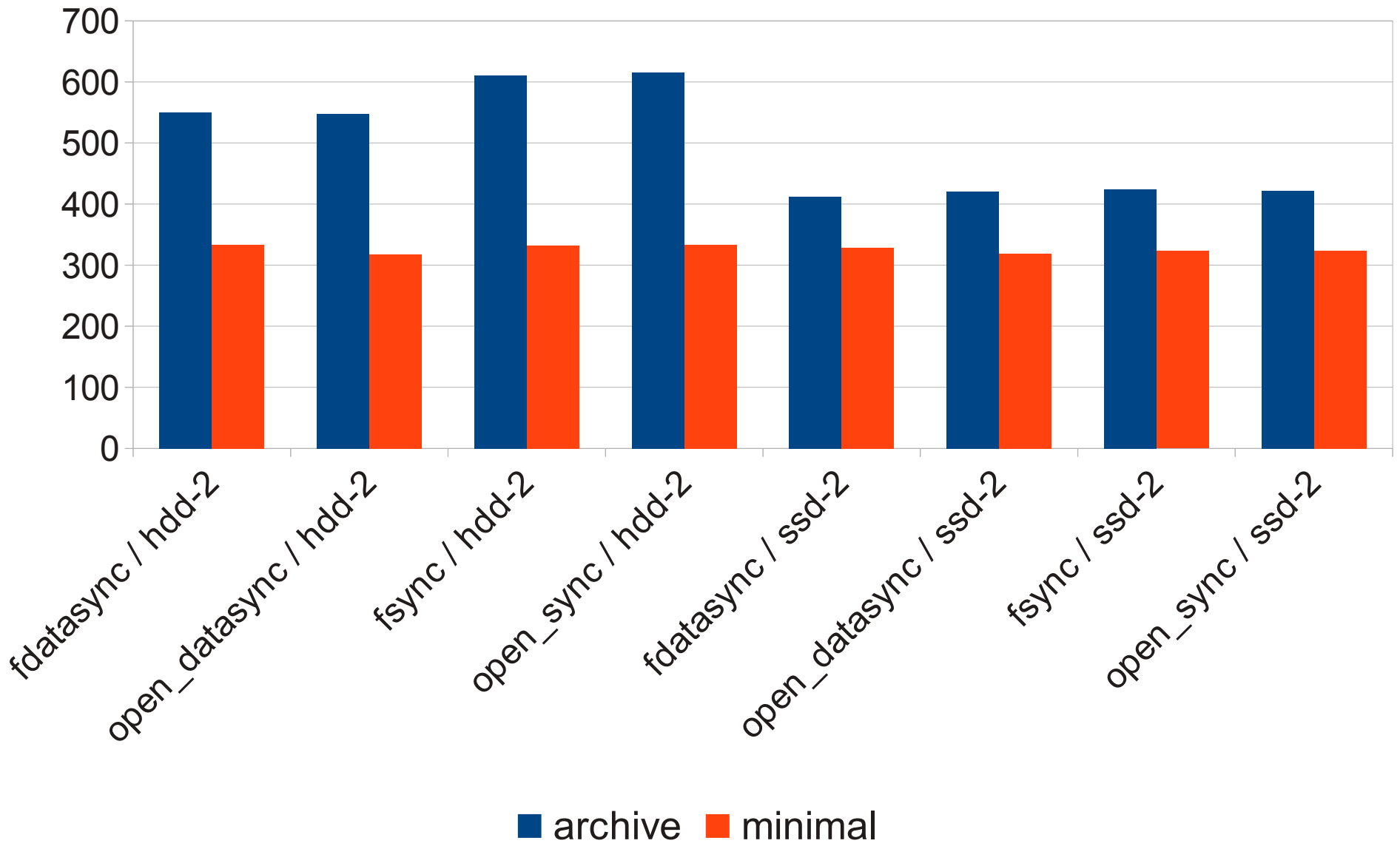
počet transakcí za vteřinu [vyšší hodnoty jsou lepší]



DSS/DWH (TPC-H)

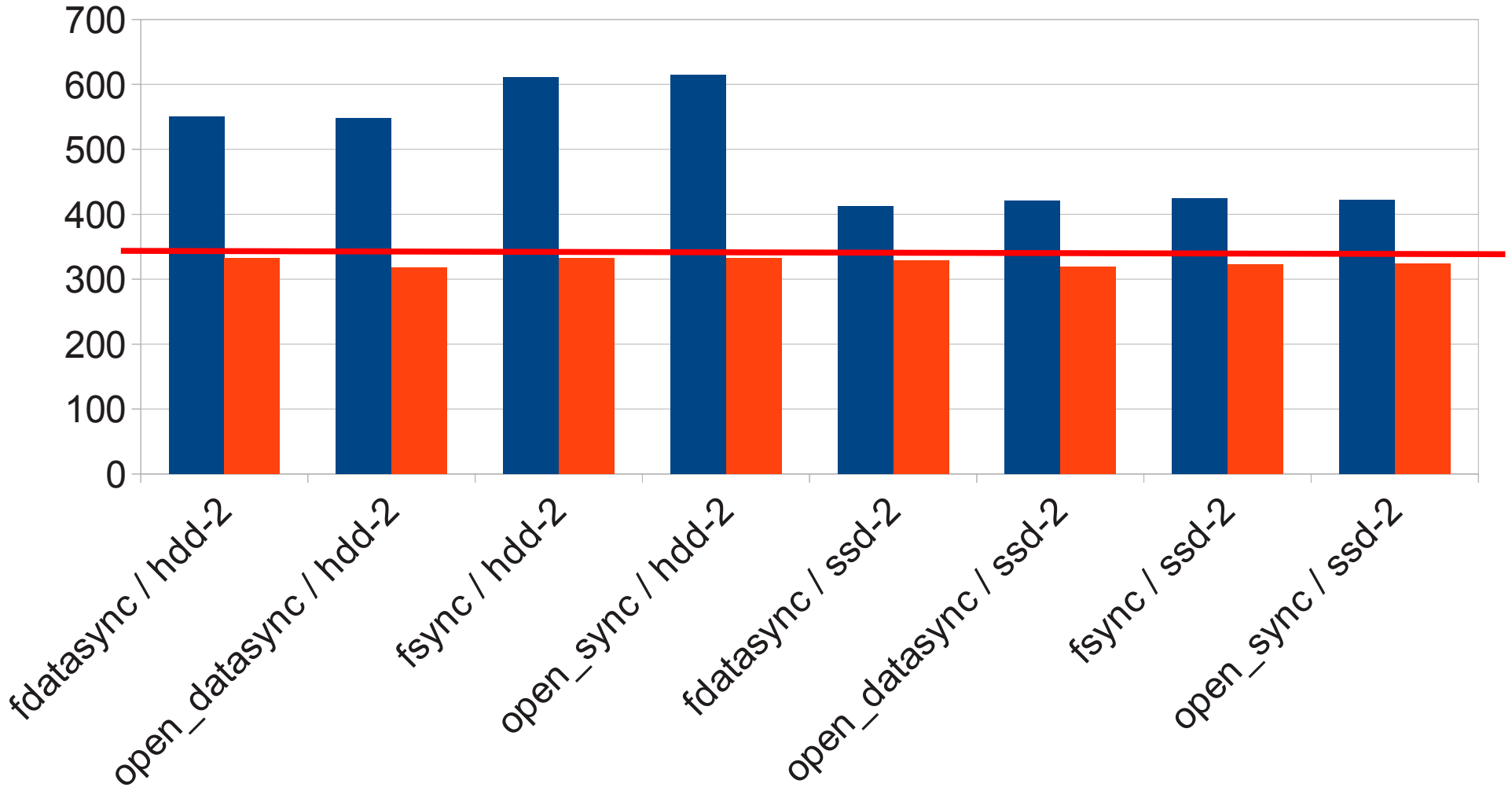
TPC-H set-up / load dat

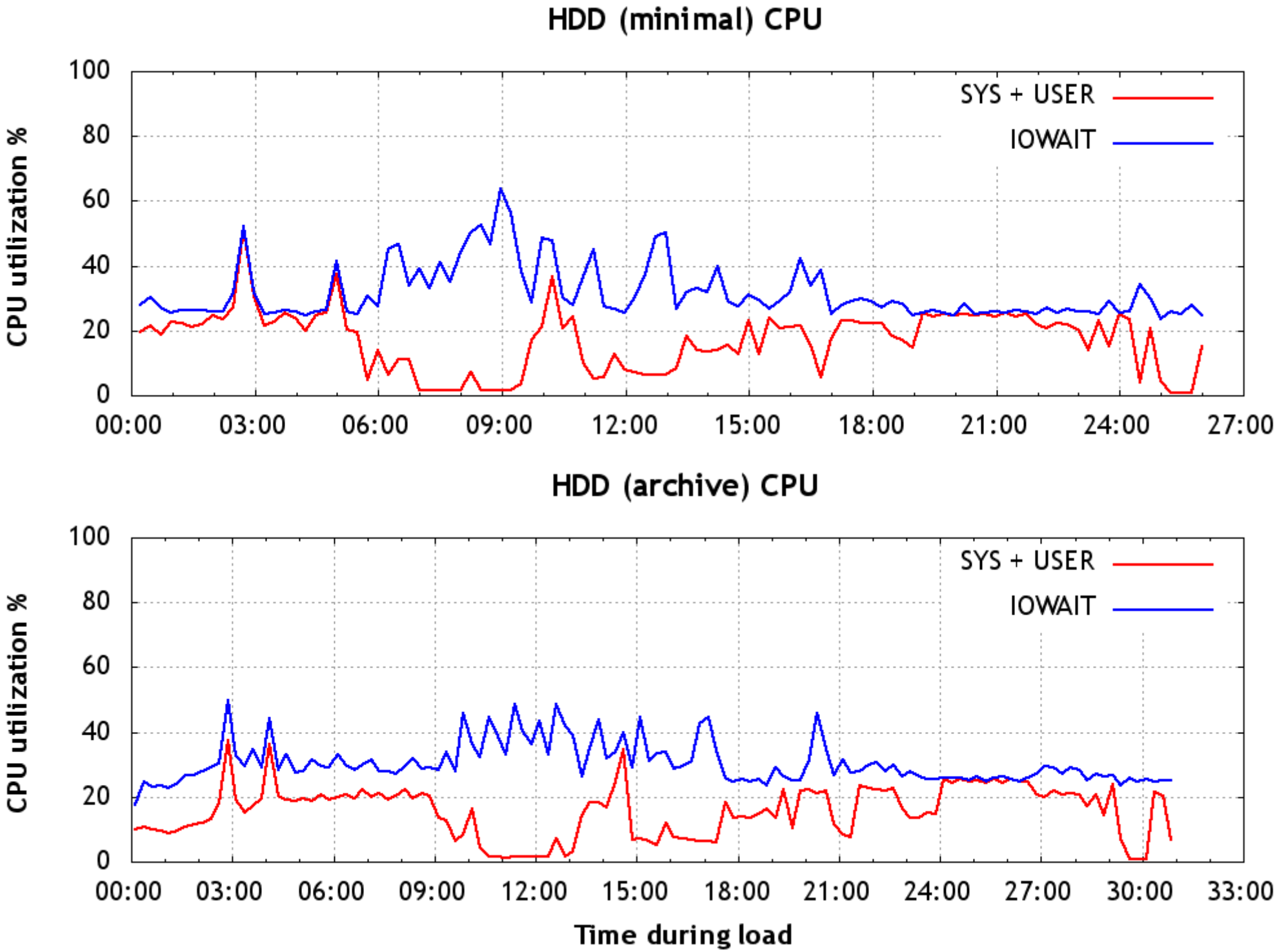
různé fsync metody / wal_level [počet vteřin, nižší hodnoty jsou lepší]



TPC-H set-up / load dat

různé fsync metody / wal_level [počet vteřin, nižší hodnoty jsou lepší]




HDD (minimal) CPU
CPU utilization %
100
80
60
40
20
0
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
24:00
27:00
SYS + USER
IOWAIT
HDD (archive) CPU
CPU utilization %
100
80
60
40
20
0
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
24:00
27:00
30:00
33:00
SYS + USER
IOWAIT
Time during load

fsync metody

- značný rozdíl mezi sync a datasync ~ 20%
- open_* - žádný měřitelný efekt
- wal_level = minimal
  - cca 30% zrychlení při loadu dat
  - pokud nepotřebujete „archive“

Cloud?

There’s a rumor that Stewart Smith and I might do a Q&A about databases in the cloud. If it happens, it will involve lots of pessimism and swearing.

Selena Deckelmann

http://chesnok.com

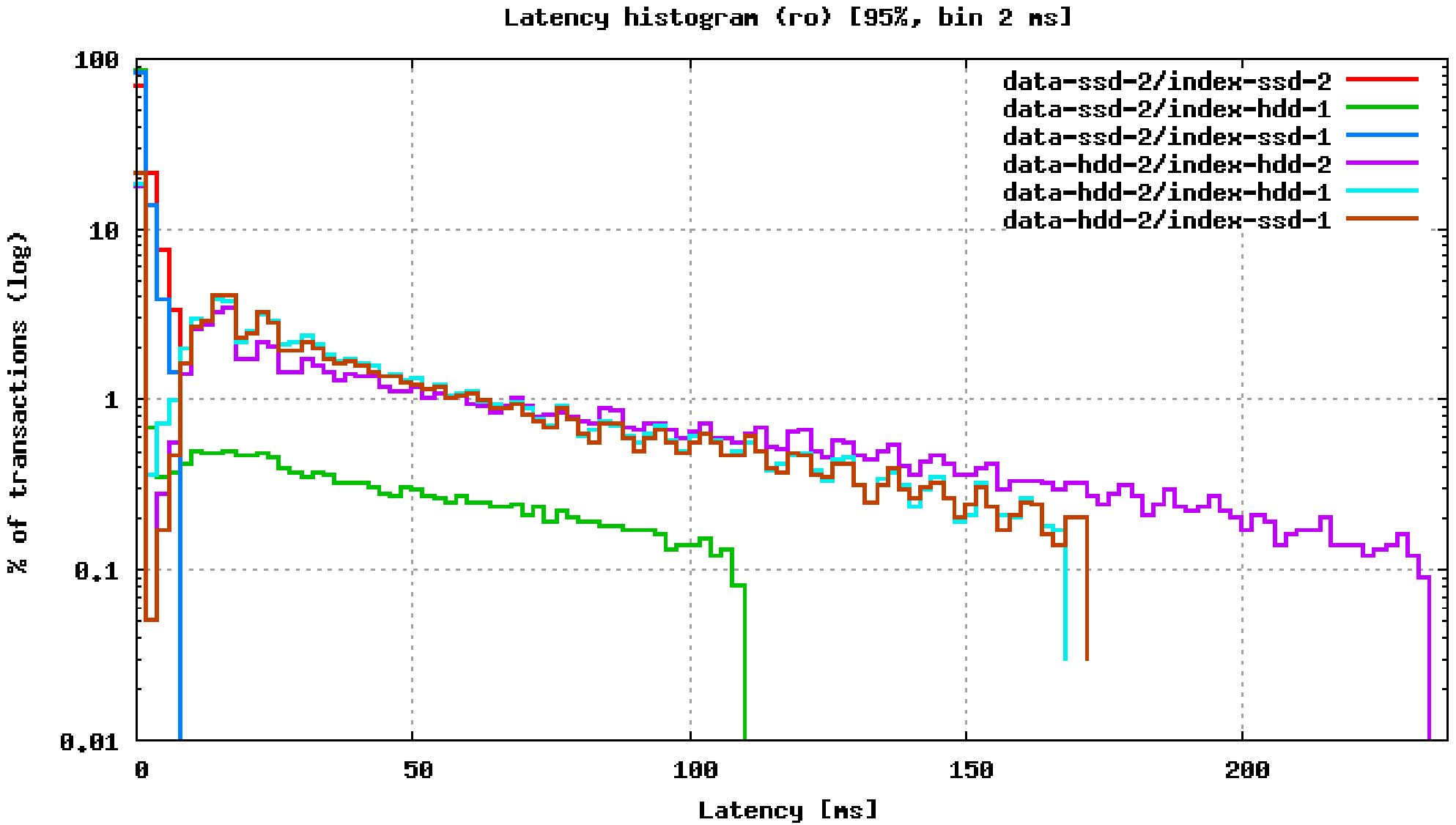
Závěrem …

Závěrem ...

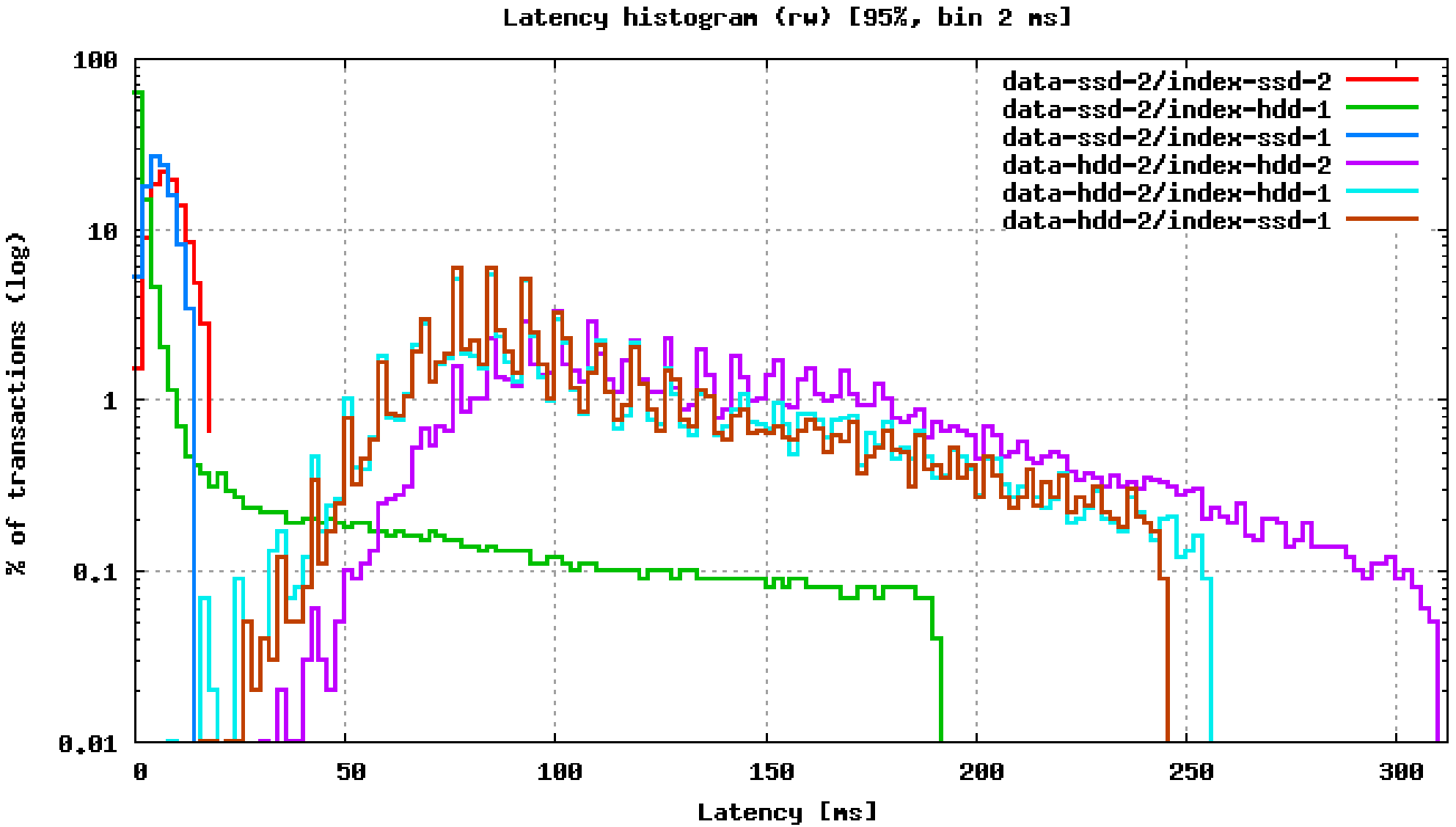
- přínos SSD závisí na workloadu (OLTP)
  - pgbench je extrém (zápis:čtení ~ 1:1)
- SSD výrazně více zatěžují CPU než HDD
- spolehlivost SSD je otázka
- HDD i SSD se skládají z „bloků“
  - HDD mají plotny, SSD mají moduly (např. 40 GB)
  - více „bloků“ => vyšší sekvenční výkon
- rychlejší disky (15k SAS) – horší cena/výkon

Každý systém má bottleneck ...

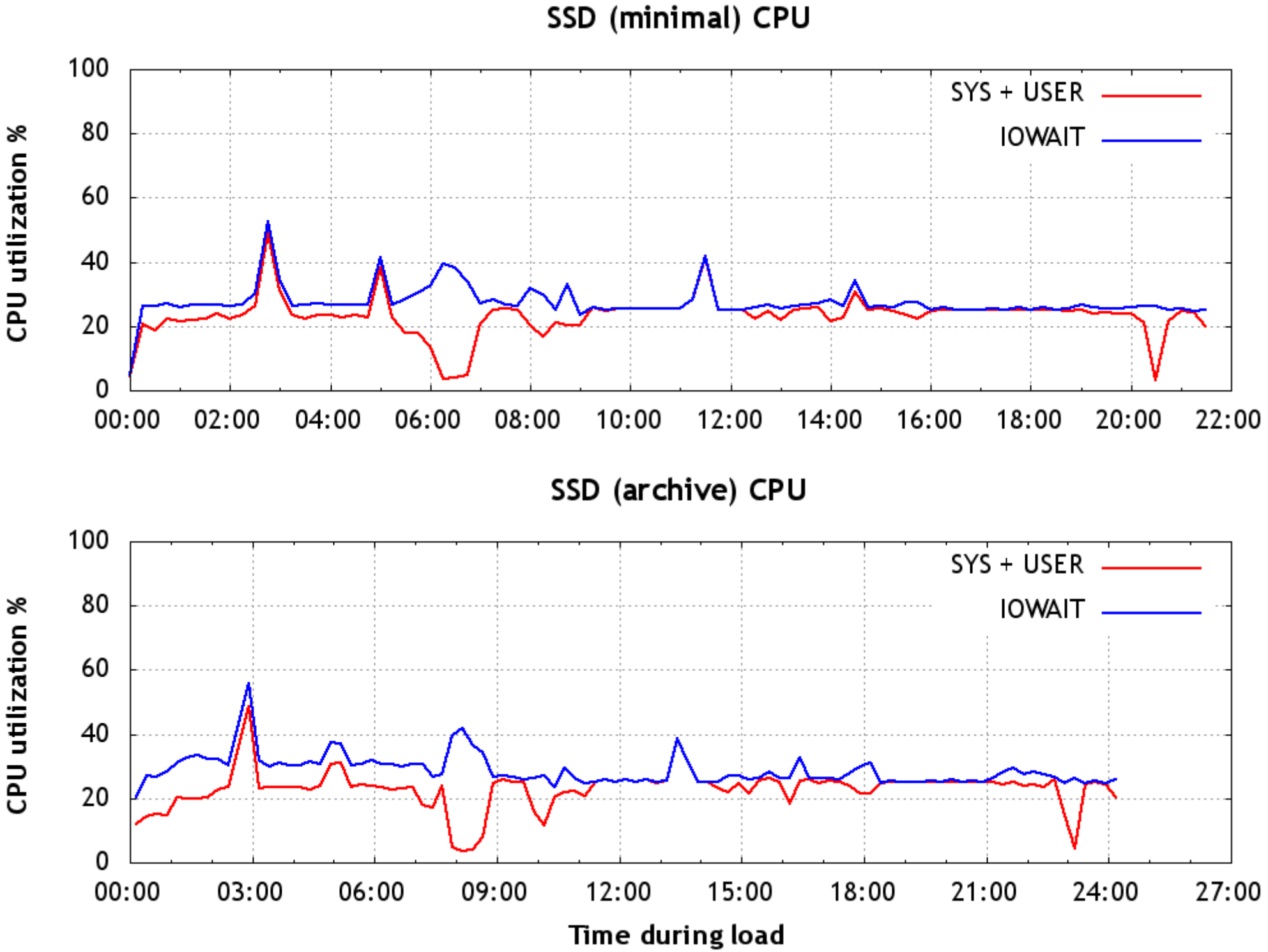
Read-only latence



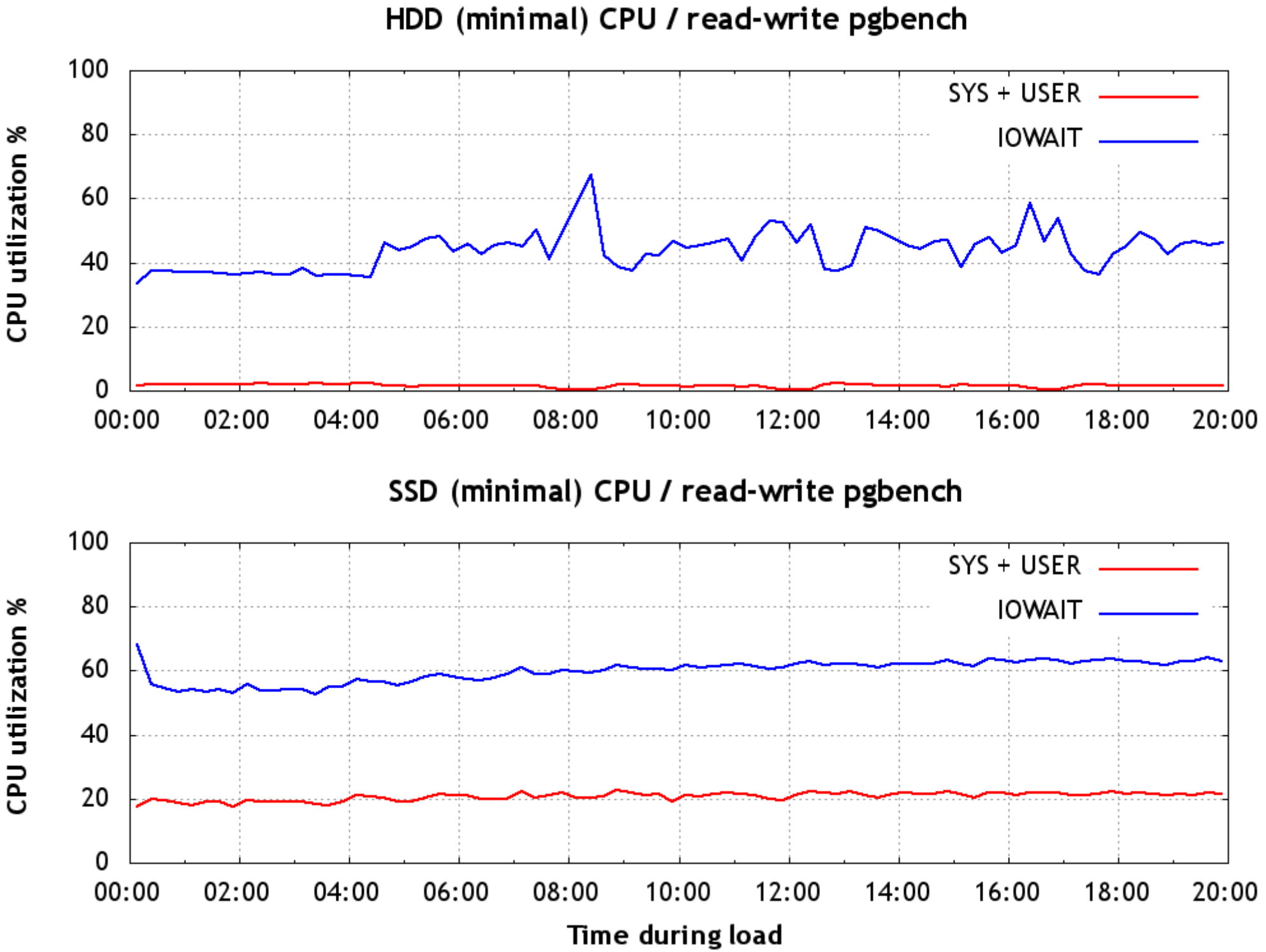
Read-write latence

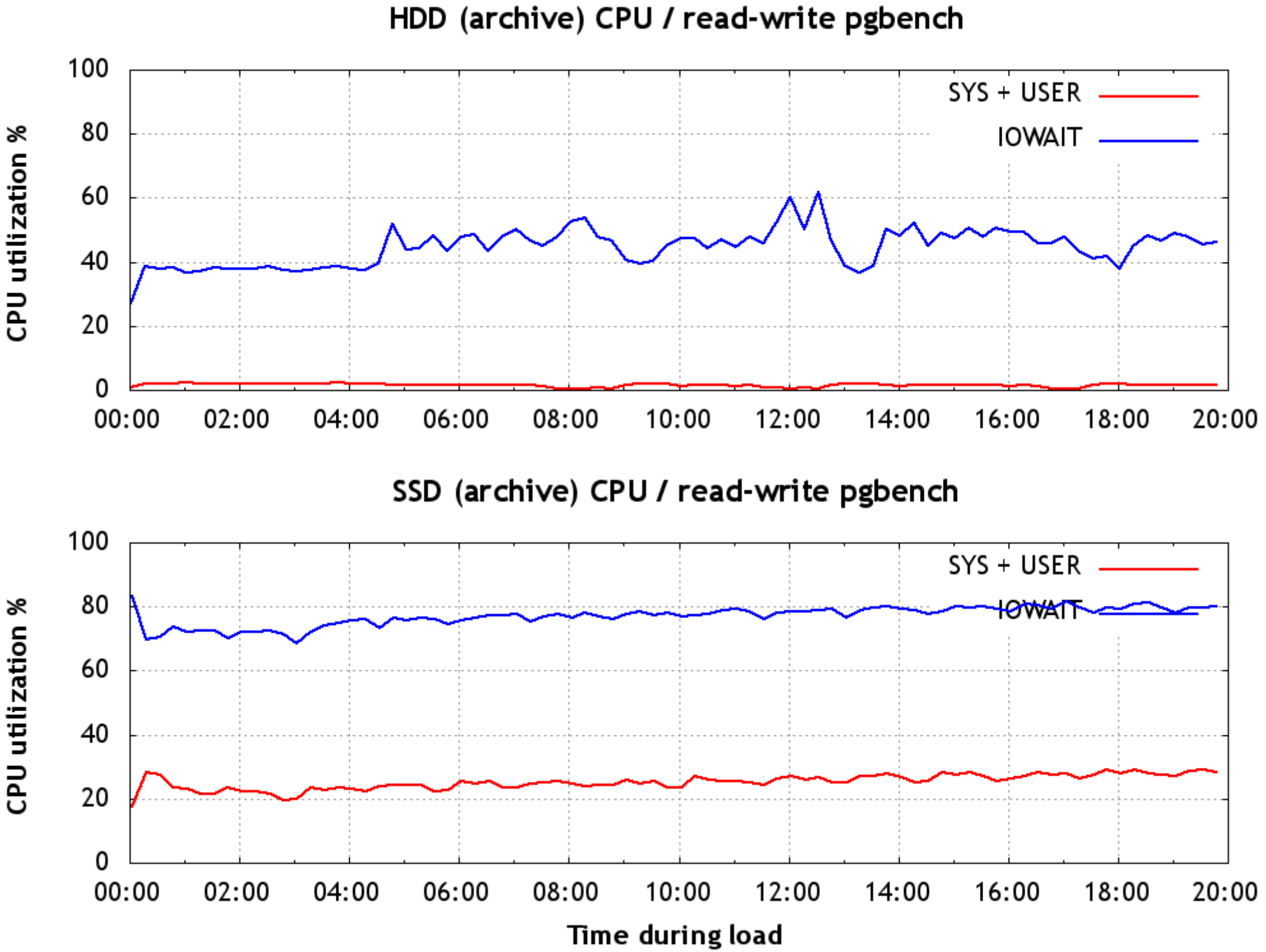


load na SSD


SSD (minimal) CPU
CPU utilization %
100
80
60
40
20
0
SYS + USER
IOWAIT
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
SSD (archive) CPU
CPU utilization %
100
80
60
40
20
0
SYS + USER
IOWAIT
00:00
03:00
06:00
09:00
12:00
15:00
18:00
21:00
24:00
27:00
Time during load

r/w pgbench


HDD (minimal) CPU / read-write pgbench
CPU utilization %
100
80
60
40
20
0
SYS + USER
IOWAIT
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
SSD (minimal) CPU / read-write pgbench
CPU utilization %
100
80
60
40
20
0
SYS + USER
IOWAIT
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
Time during load


HDD (archive) CPU / read-write pgbench
CPU utilization %
100
80
60
40
20
0
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
SYS + USER
IOWAIT
SSD (archive) CPU / read-write pgbench
CPU utilization %
100
80
60
40
20
0
00:00
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
SYS + USER
IOWAIT
Time during load

We care more about your data
than our benchmarks …

… and you should too!

Ne tak zřejmé ...

- SSD vám nezaručí vyšší výkon
  - závisí na workloadu (OLTP/DWH, read/write)
  - závisí na části databáze na SSD (indexy, WAL)
  - mohou být levnější varianty (hot standby, ...)
  - závisí na mount options (barriers, …)

- SSD výrazně více zatěžují CPU
  - menší iowait => větší sys/user time
  - každý systém má bottleneck

Ne tak zřejmé ...

spolehlivost SSD je otázka

- málo dat, každé nové SSD je víceméně „jiné“
- S.M.A.R.T. víceméně jen pro wearout :-(
- failuje hlavně firmware (řadič) – nepředvídatelné
- příklad: Intel 8MB bug

RAID a SSD

- ne každý RAID umí spolupracovat (TRIM)
- stejně jako u HDD stavět RAID z více šarží
- firmware je většinou stejný

SSD a HDD

HDD

- skládají se z ploten
- počet ploten, vyšší hustota => vyšší rychlost
- náhodné operace dány rychlostí otáčení
- rychlost se mění i podle pozice na disku

SSD

- skládají se z modulů (např. Intel á 40GB)
- více modulů => vyšší sekvenční rychlost
- náhodné operace dány řadičem

Rychlejší disky (15k)?

cca 2x vyšší výkon

- 7.2k SATA - 120 IOPS
- 15k SAS - 250 IOPS (až 400)

4x vyšší cena

- 300 GB 7.2k SATA - 9600 Kč
- 300 GB 15k SAS - 2100 Kč

celkově ~ 2x horší metrika (tps / $ / GB)