Seeking PostgreSQL (2013)

Mechanical Drive Physics

- Head seeking time
  - 3 to 12 ms
  - Small SAS drives faster than Big SATA drives
- Rotation
  - 15K to 5400 RPM
  - 250 to 90 Rotations/Second
  - 4 to 11 ms
- I/O operations per second (IOPS)
  - Average head seek plus ½ rotation

Throughput

- 10 ms per seek is 100 seeks/second
  - AKA 100 IOPS
- PostgreSQL pages are 8192 bytes each
- 100 / sec * 8192 = 0.8 MB/s

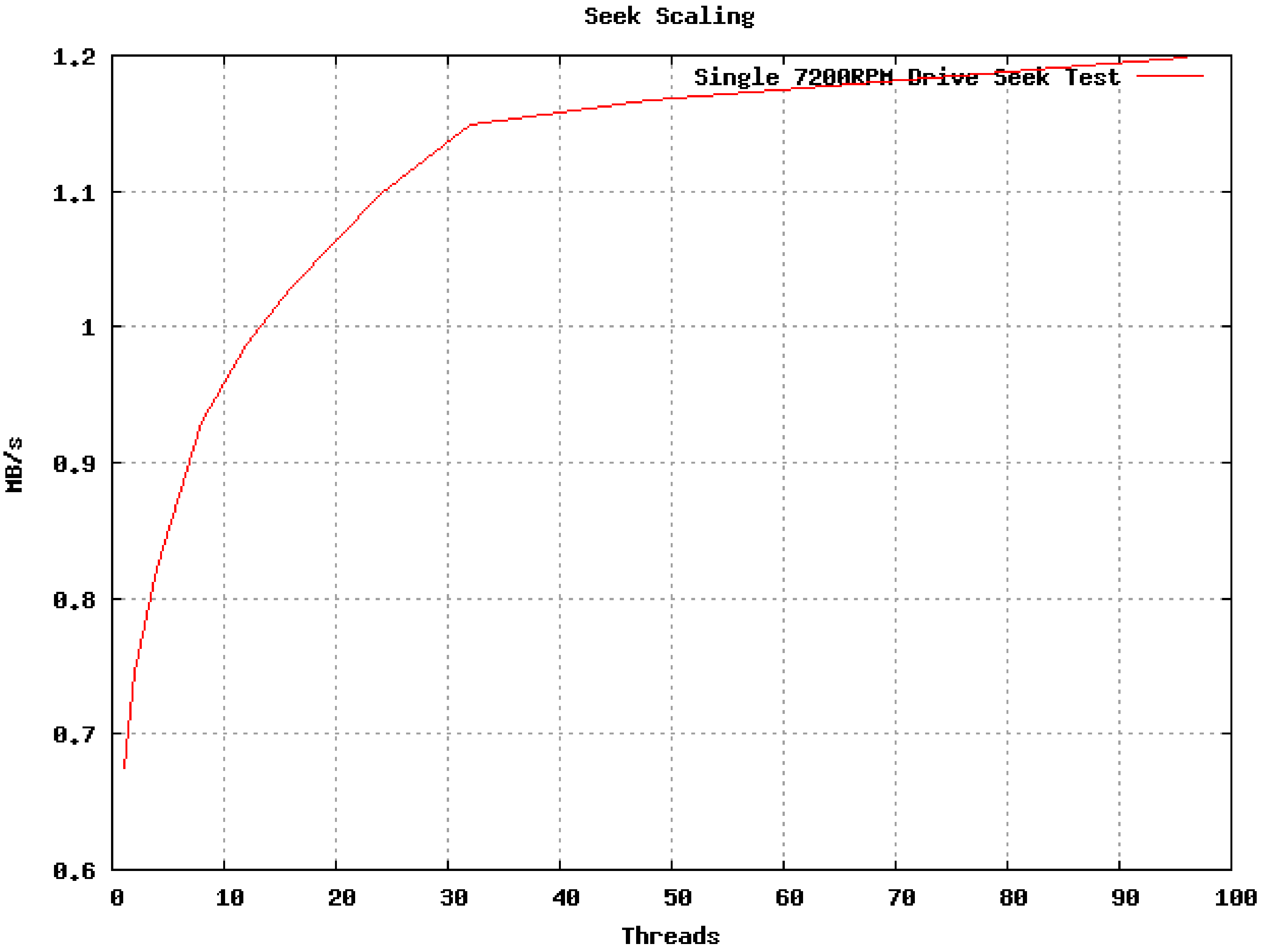
Optimizations

- Elevator sorting
  - Native Command Queueing
  - Typically 32 request queue
- Read/write combining
- Read-ahead
- Non-volatile write caches
  - http://wiki.postgresql.org/wiki/Reliable_Writes
  - Look for the battery

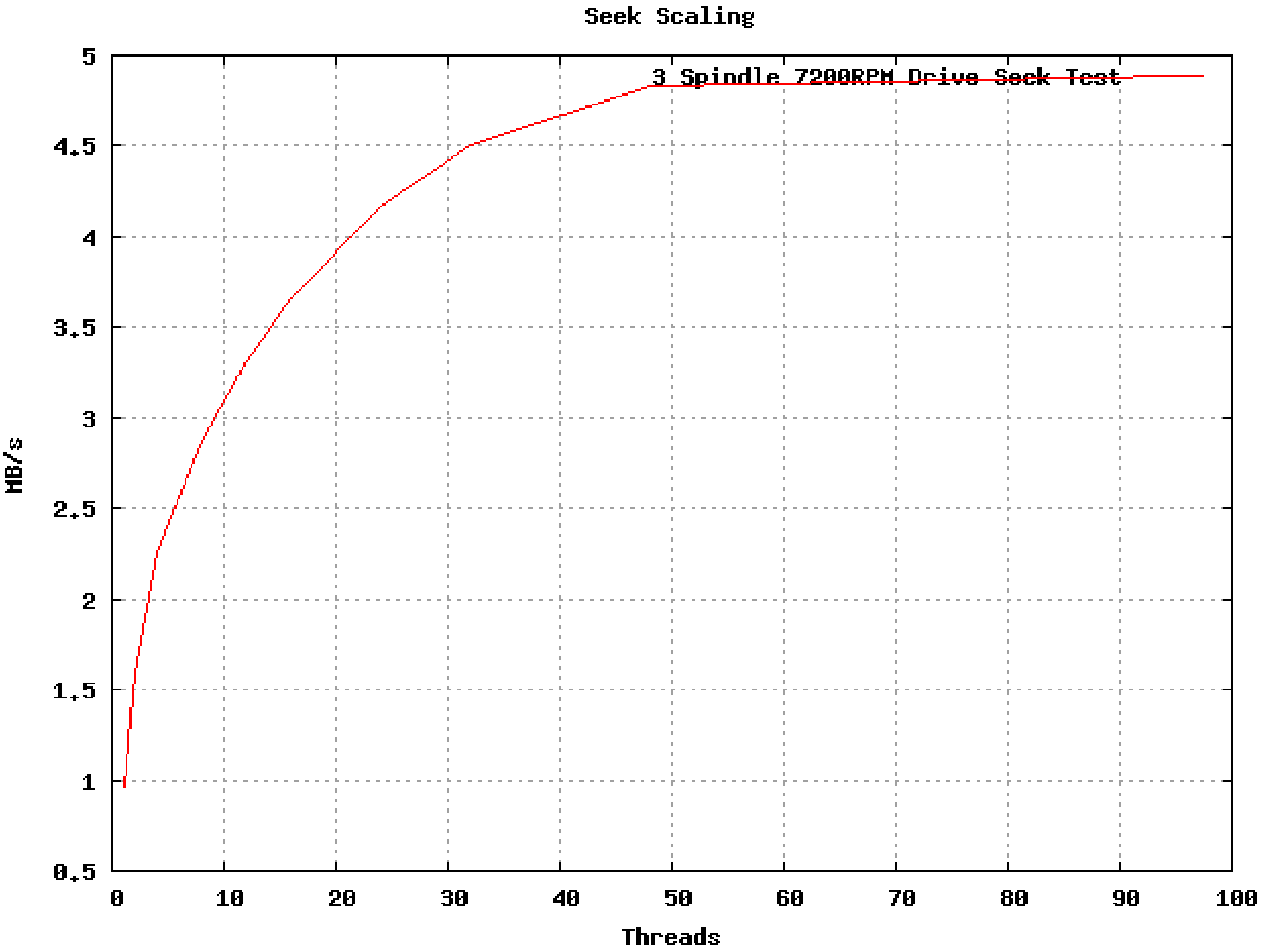
Seek scaling

- https://github.com/gregs1104/seek-scaling/
- Executes using sysbench
- Cache clearing code is Linux only
- Simple disk seeks
- Fixed size
    - Test sizes need to match
- Variable number of clients

Short-stroked 7200RPM Disk



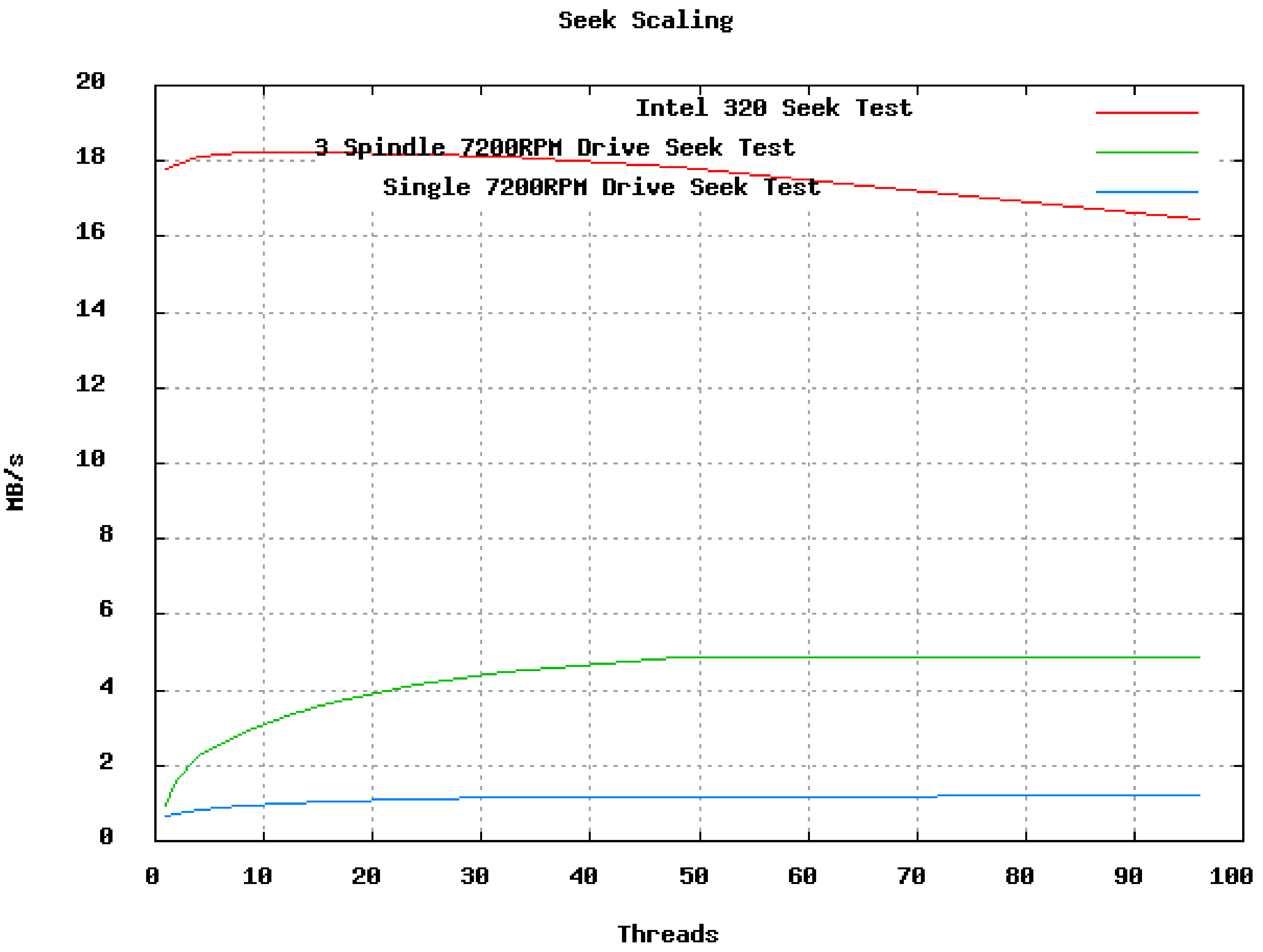
3 Disk RAID-0


Seek Scaling
3 Spindle 7200RPM Drive Seek Test
MB/s
Threads
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
10
20
30
40
50
60
70
80
90
100

Silicon State Devices (SSD)

- AKA Flash RAM drives
- Intel 320 Series SSD
  - Enterprise 710 series mainly longer lifetime
  - Up to 270MB/s reads!
  - Up to 47K Read IOPS!

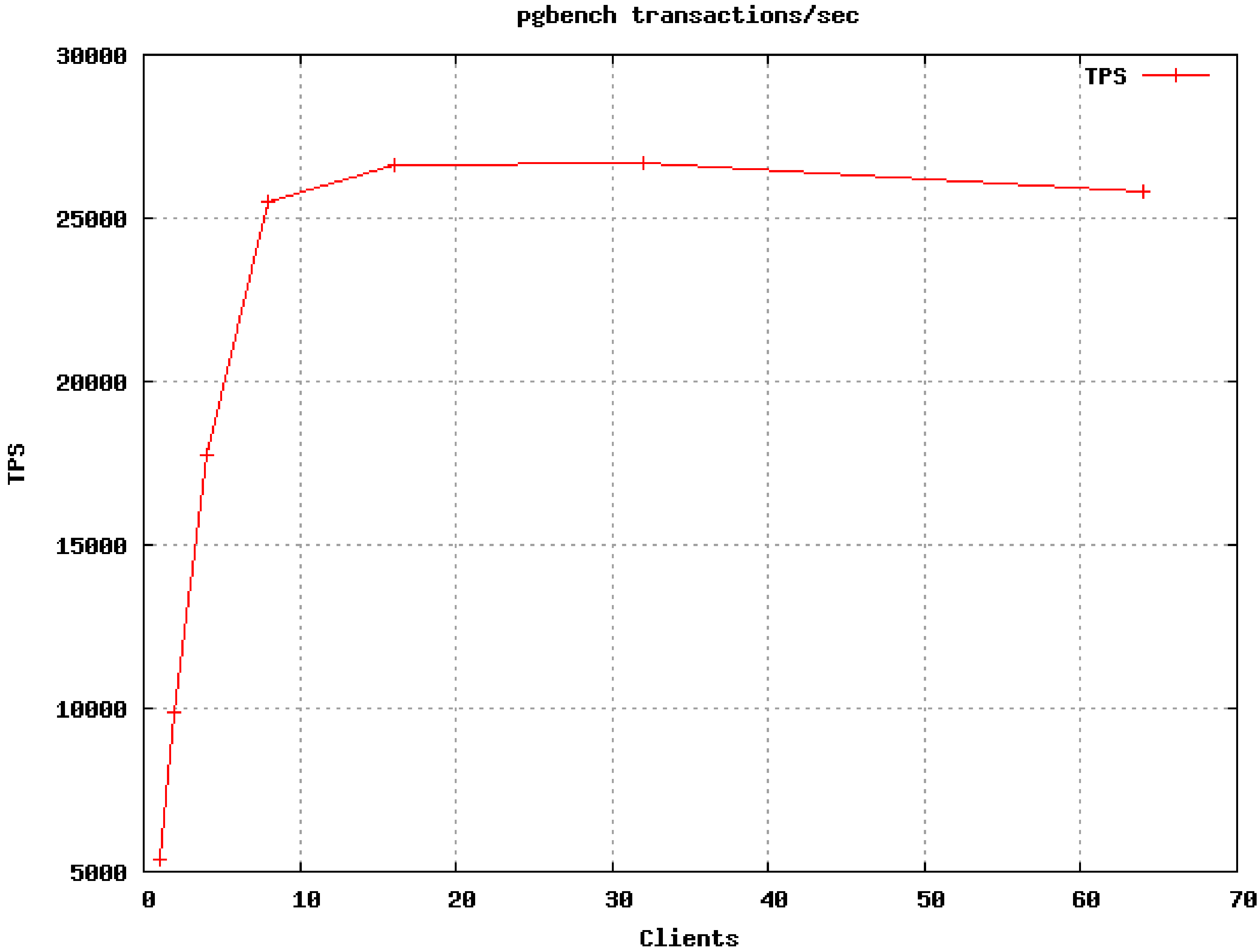
Up to no good


Seek Scaling
Intel 320 Seek Test
3 Spindle 7200RPM Drive Seek Test
Single 7200RPM Drive Seek Test
MB/s
0
2
4
6
8
10
12
14
16
18
20
Threads
0
10
20
30
40
50
60
70
80
90
100

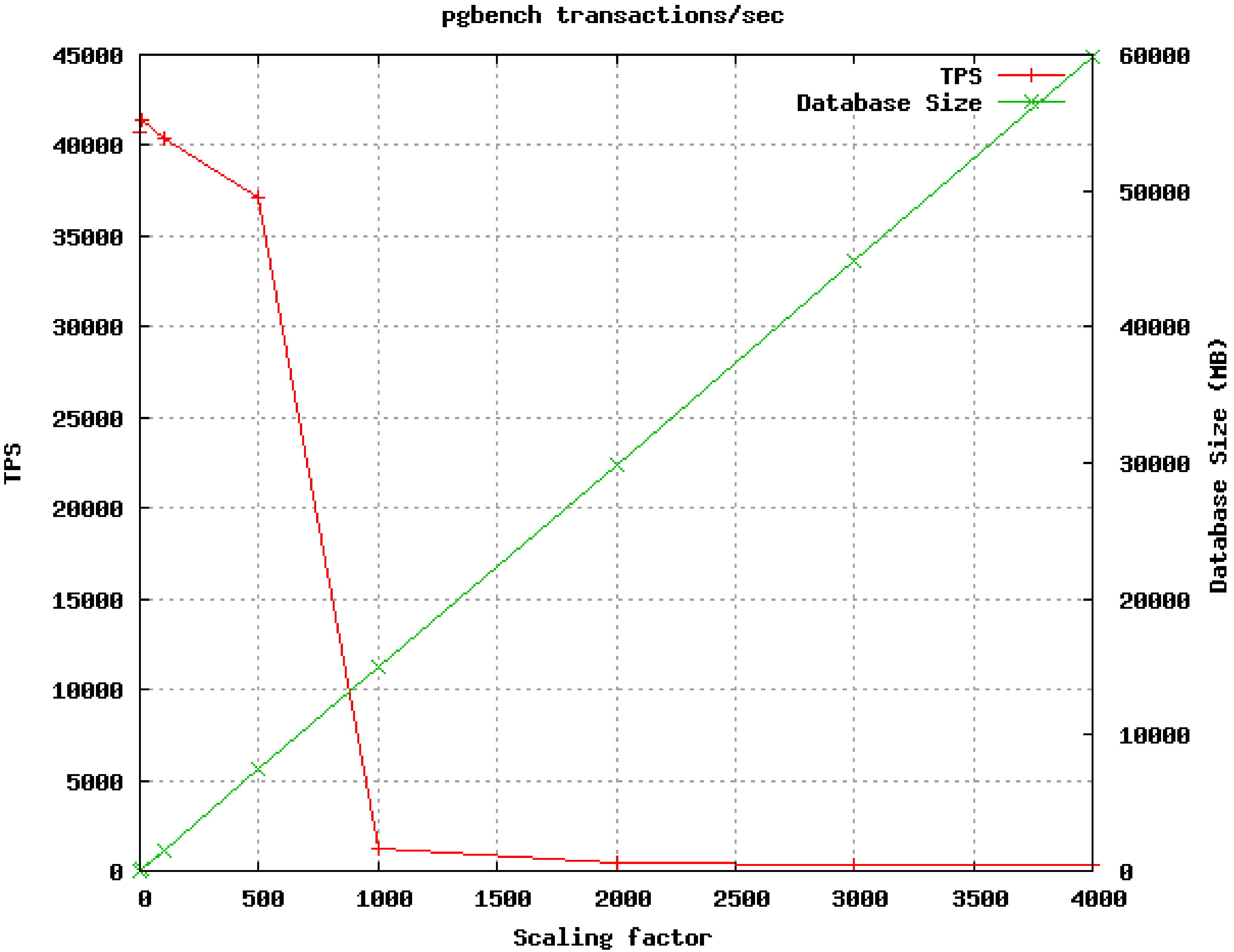
Database tests

- pgbench
- PostgreSQL 9.0
  - 9.1 mostly the same
  - 9.2 very different on larger servers
- 4 Hyperthreaded cores = 8 threads
- Server with 16GB of RAM
- 2 PCI-E slots with storage controllers
- 7 drive bays
- Scientific Linux 6.0, XFS filesystems

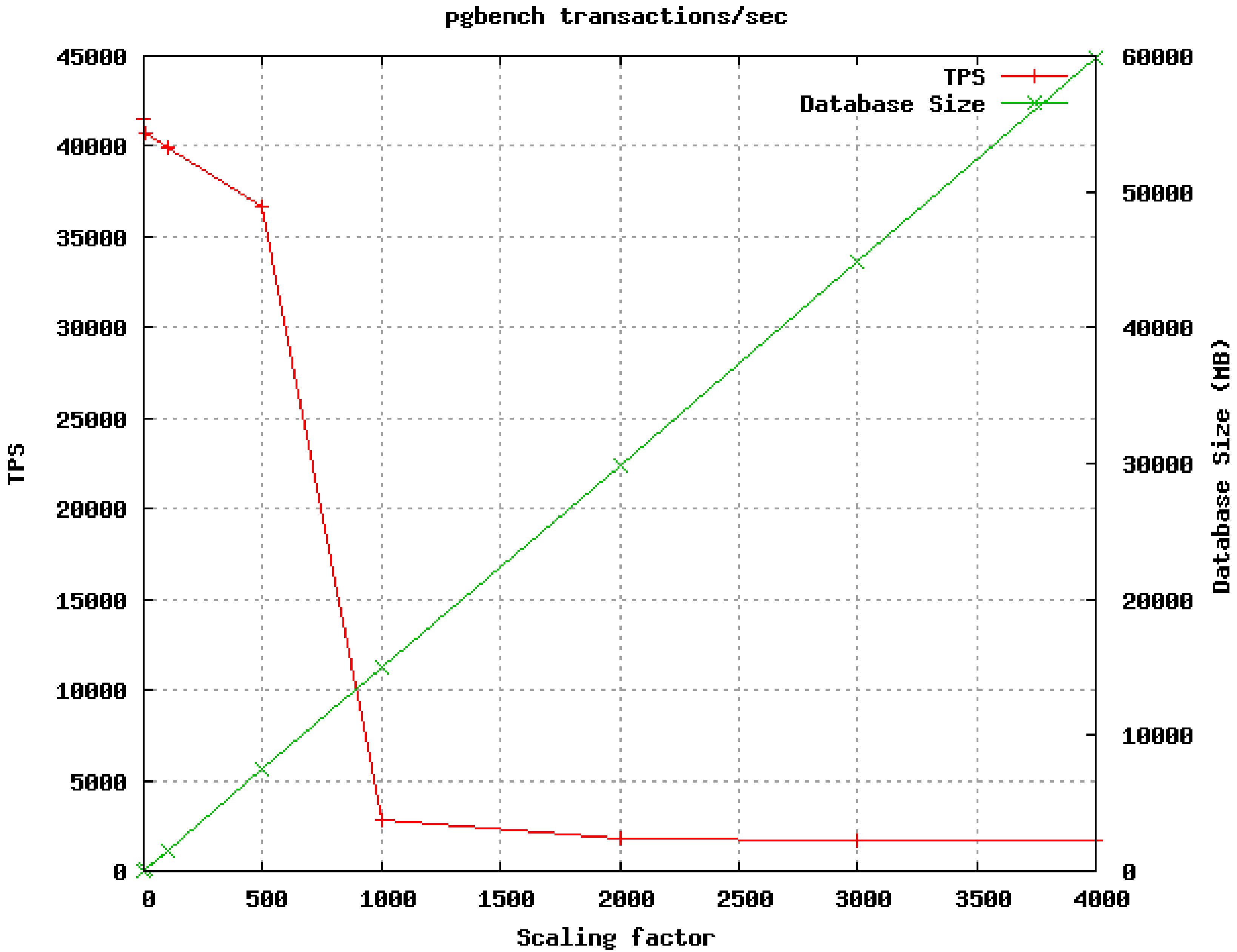
SELECT-only Client scaling



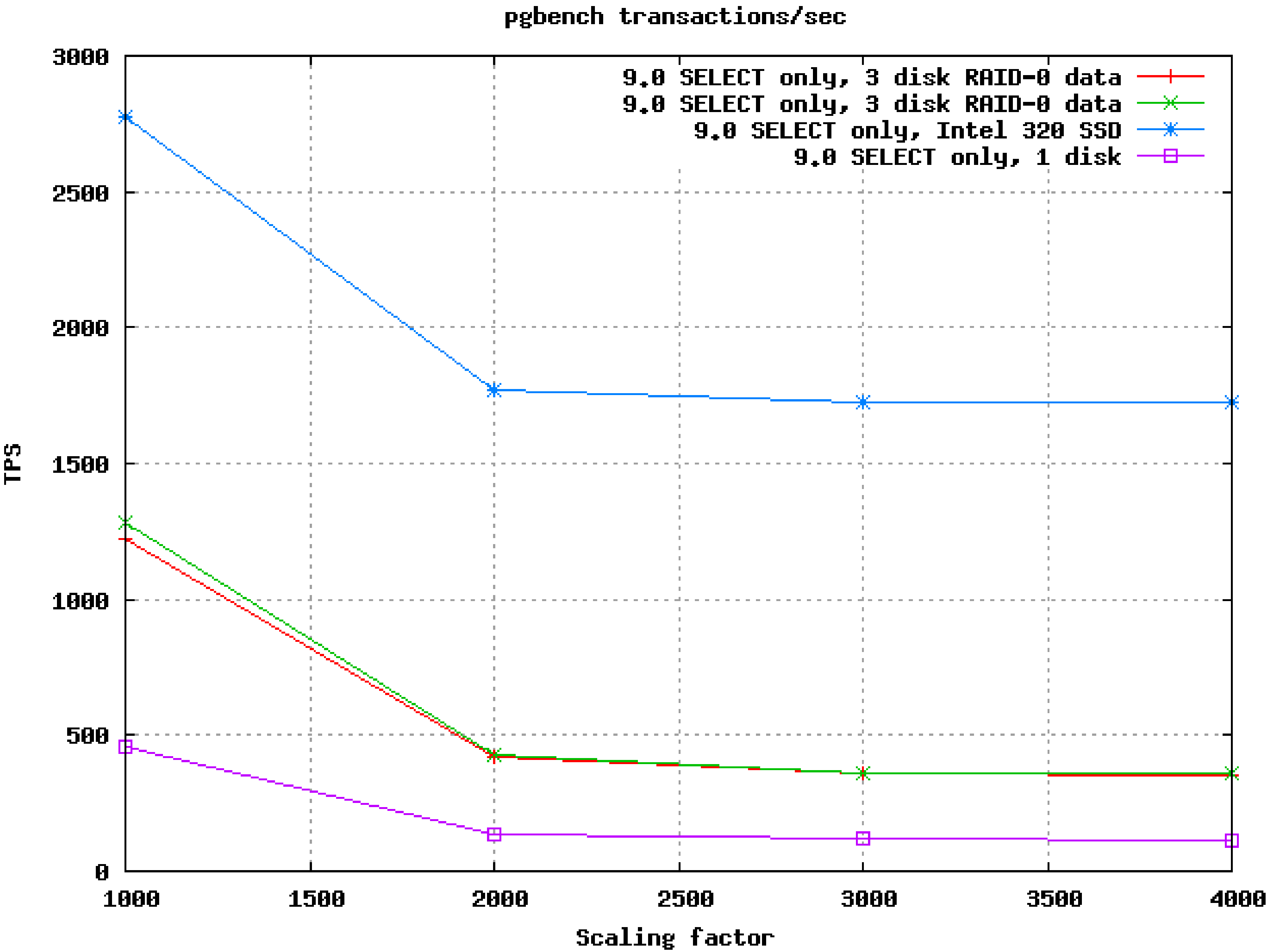
SELECT-only, 3 disk RAID-0



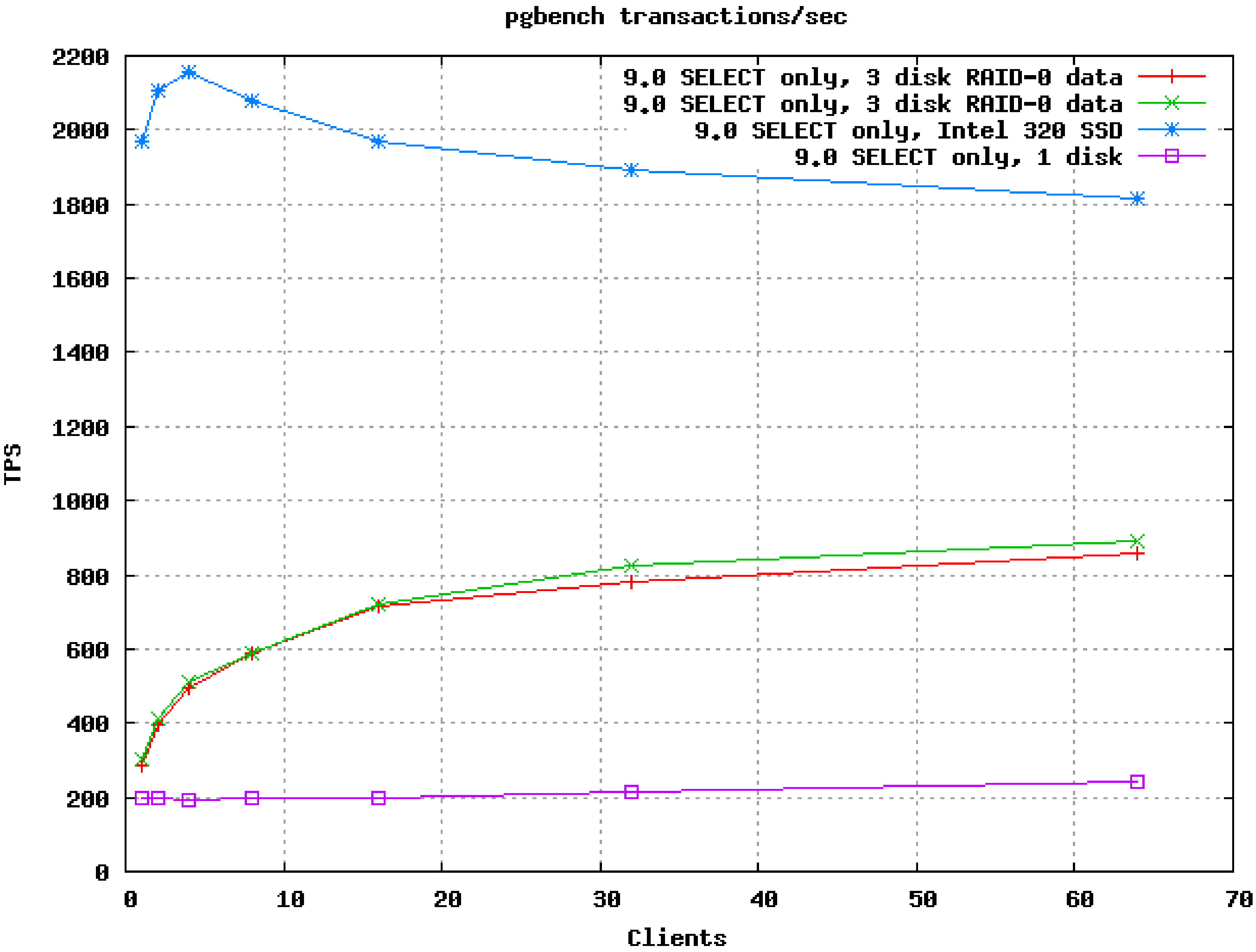
Intel 320 SSD



Big data!


pgbench transactions/sec
TPS
Scaling factor
3000
2500
2000
1500
1000
500
0
1000
1500
2000
2500
3000
3500
4000
9.0 SELECT only, 3 disk RAID-0 data
9.0 SELECT only, 3 disk RAID-0 data
9.0 SELECT only, Intel 320 SSD
9.0 SELECT only, 1 disk

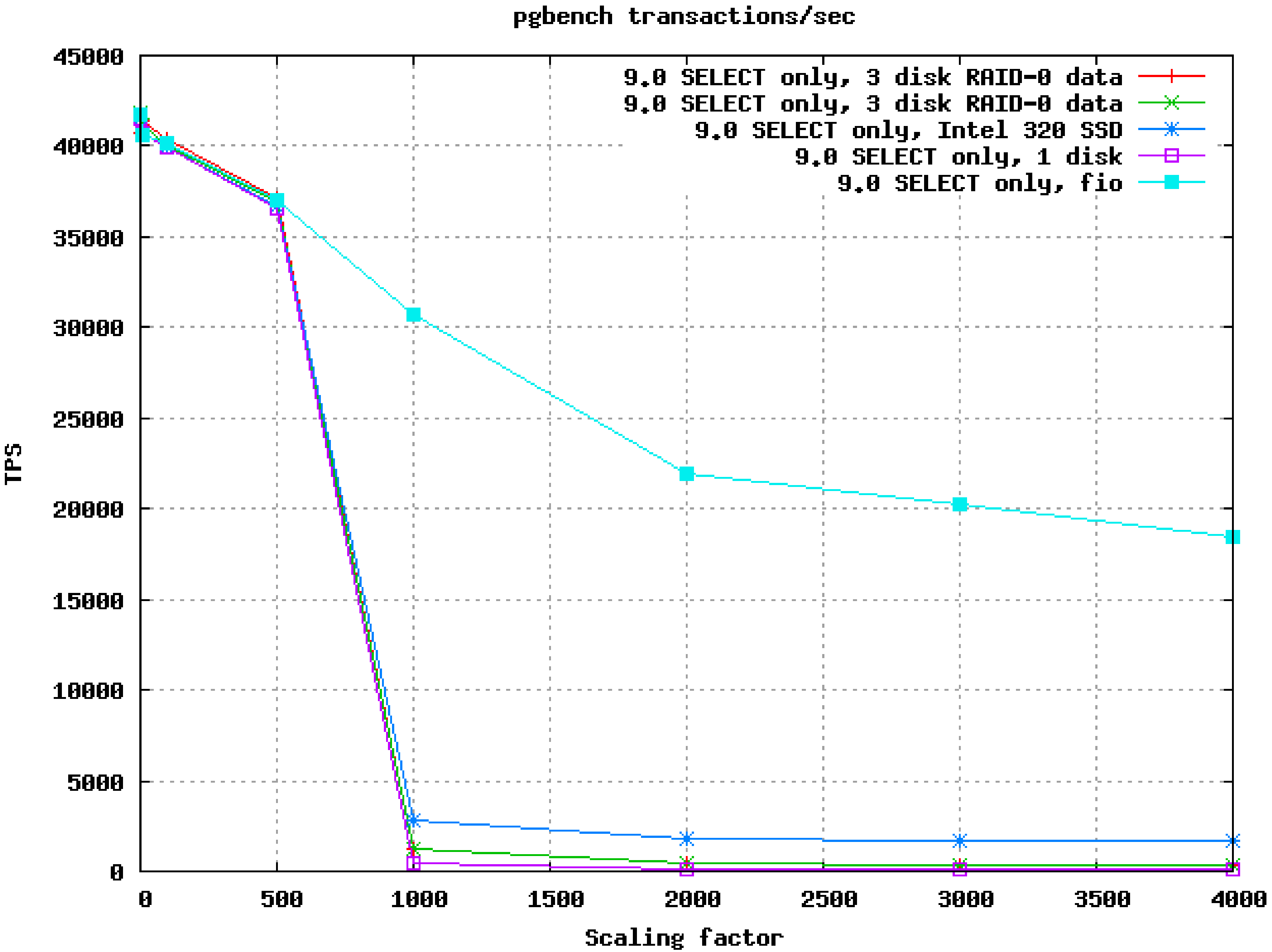
Concurrency



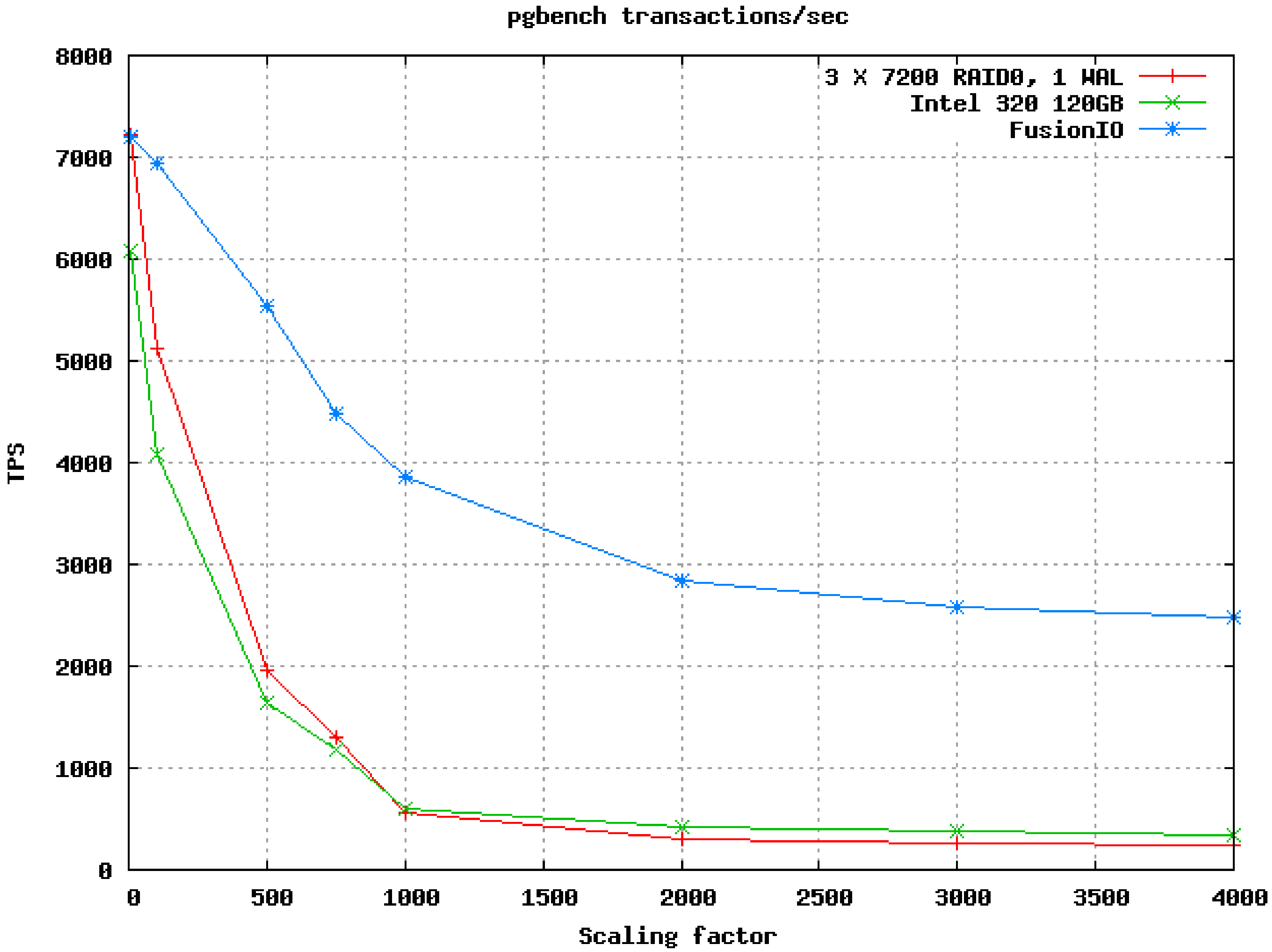
Money *can* buy you scaling

- PCI-E flash cards
- Fusion-io, TMS RAMSAN, Virident
- Many channels of flash
- Many dollars of cash
  - Typically >$10K each for small capacities

Fusion-io ioDrive 80GB



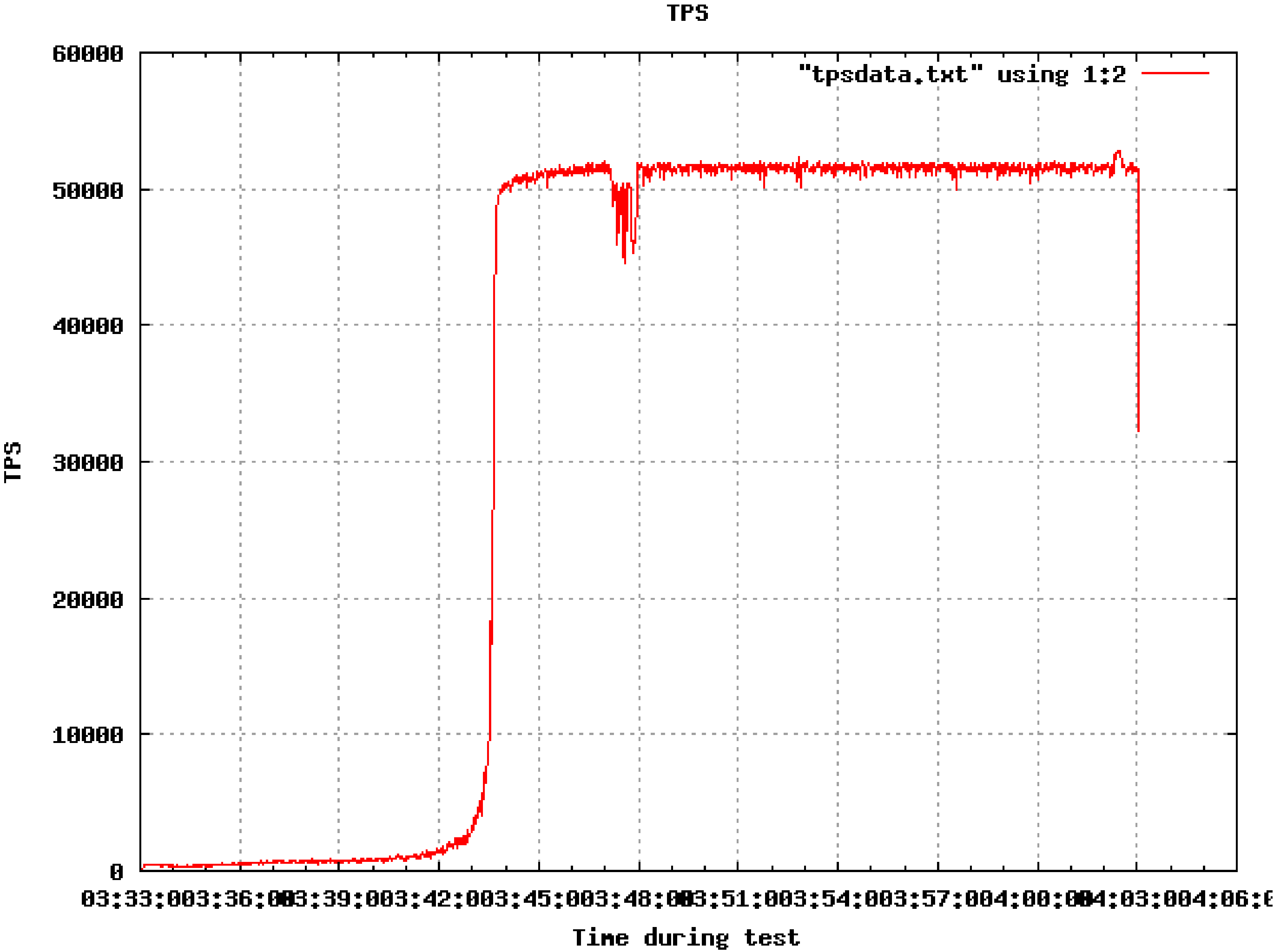
pgbench TPC-B writes



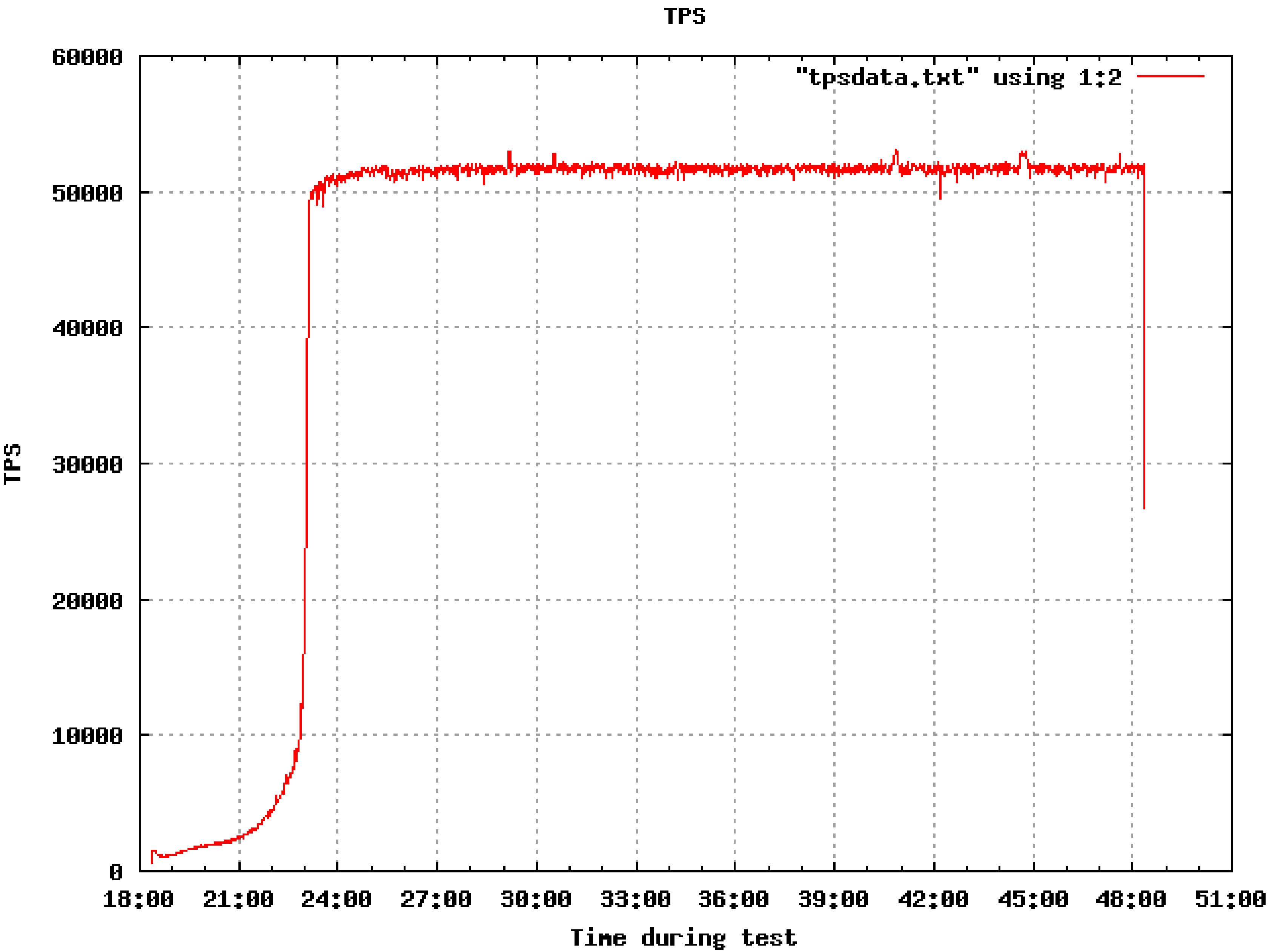
Cache refill

- Server has been restarted
- No cached information
- 7.5GB database, 32 clients
- Possible to do 50K TPS when in memory
- How long until original performance?

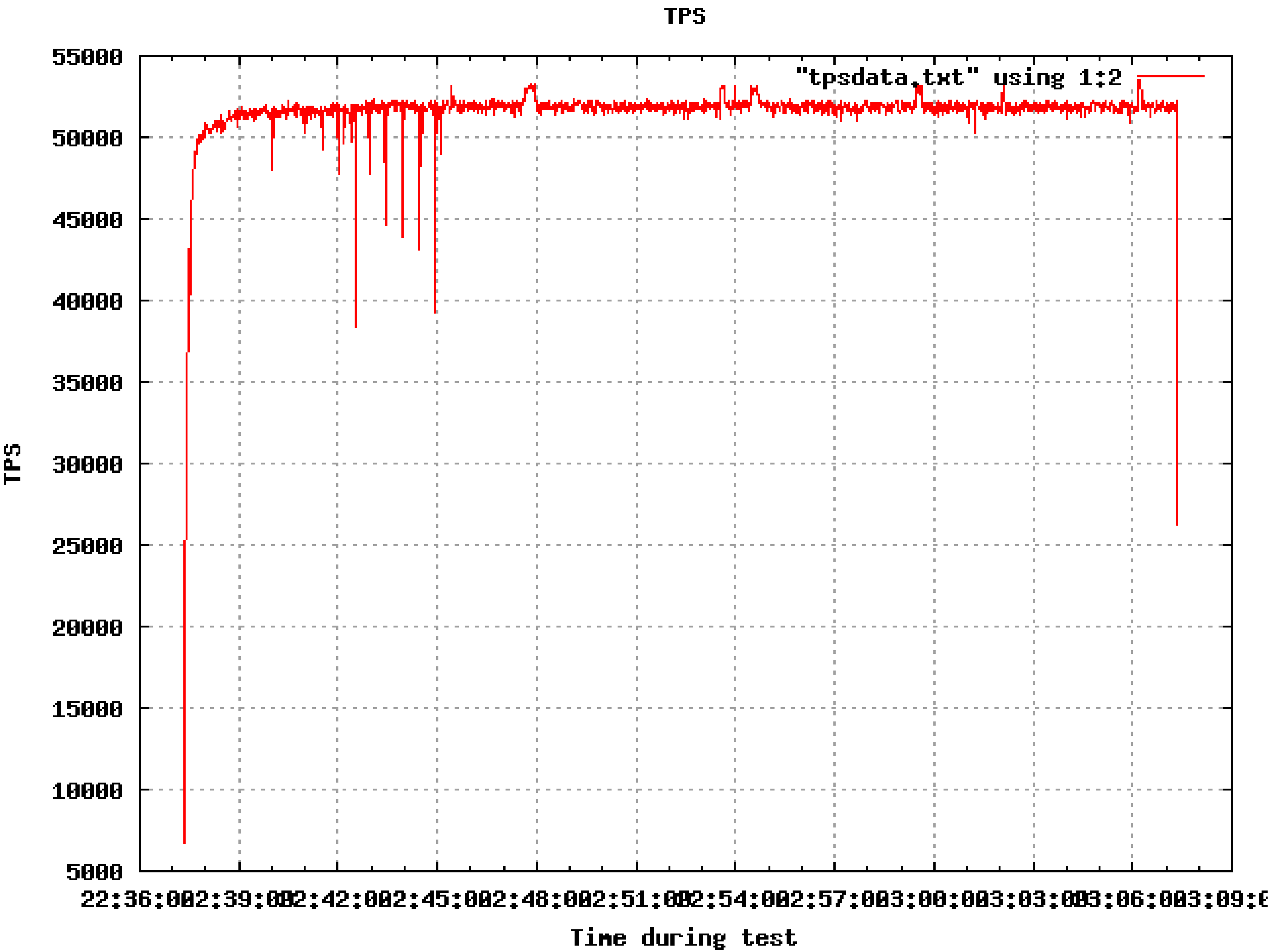
3-disk RAID0: 11 minutes



Intel 320: 5 minutes



Fusion-io: 20 seconds



Measured refill rates

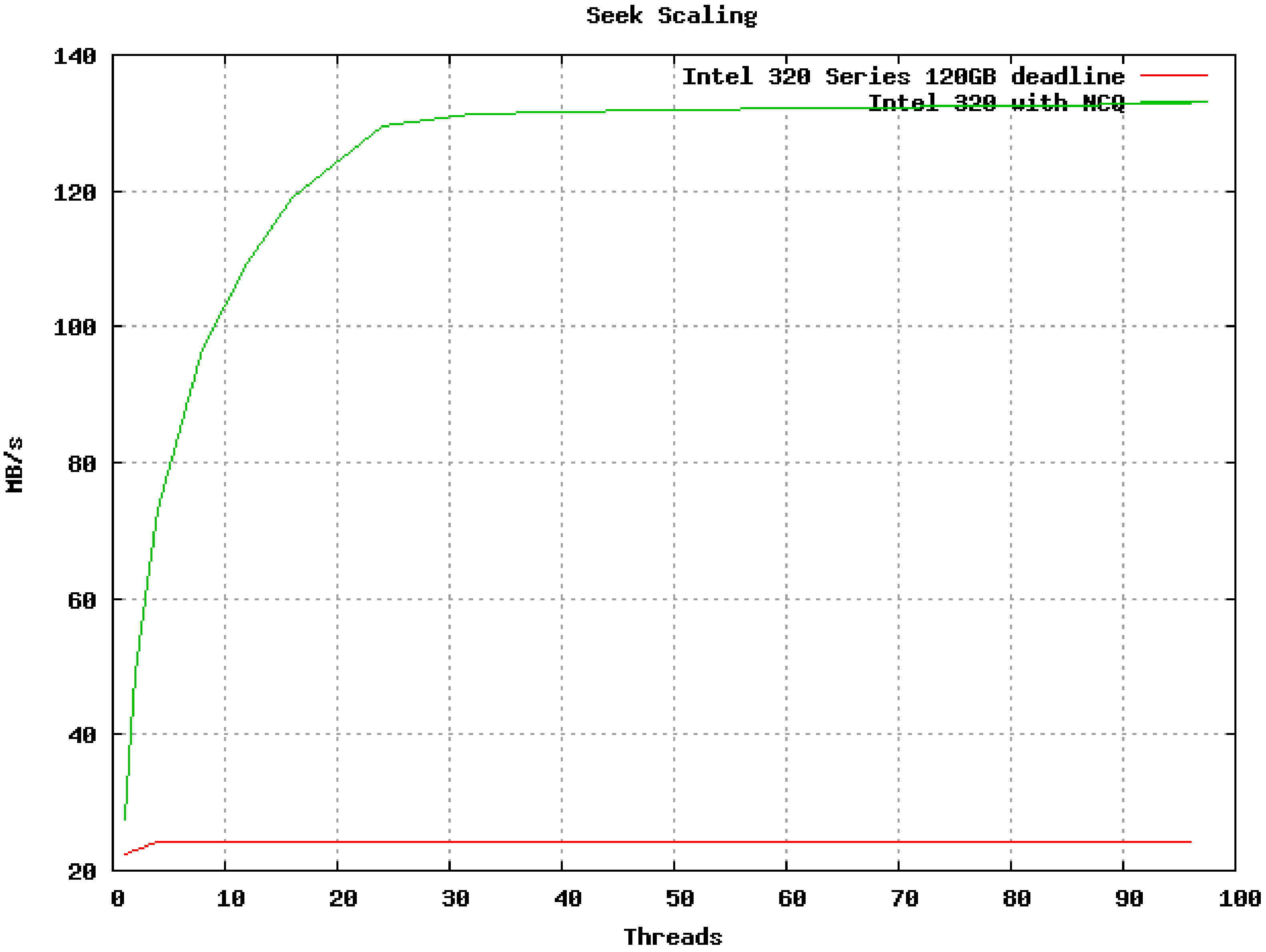
- 3 disk RAID-0: 7 to 15MB/s
- Intel SSD: 29 to 32MB/s
- Fusion-io ioDrive: 583 to 621MB/s

Oops!

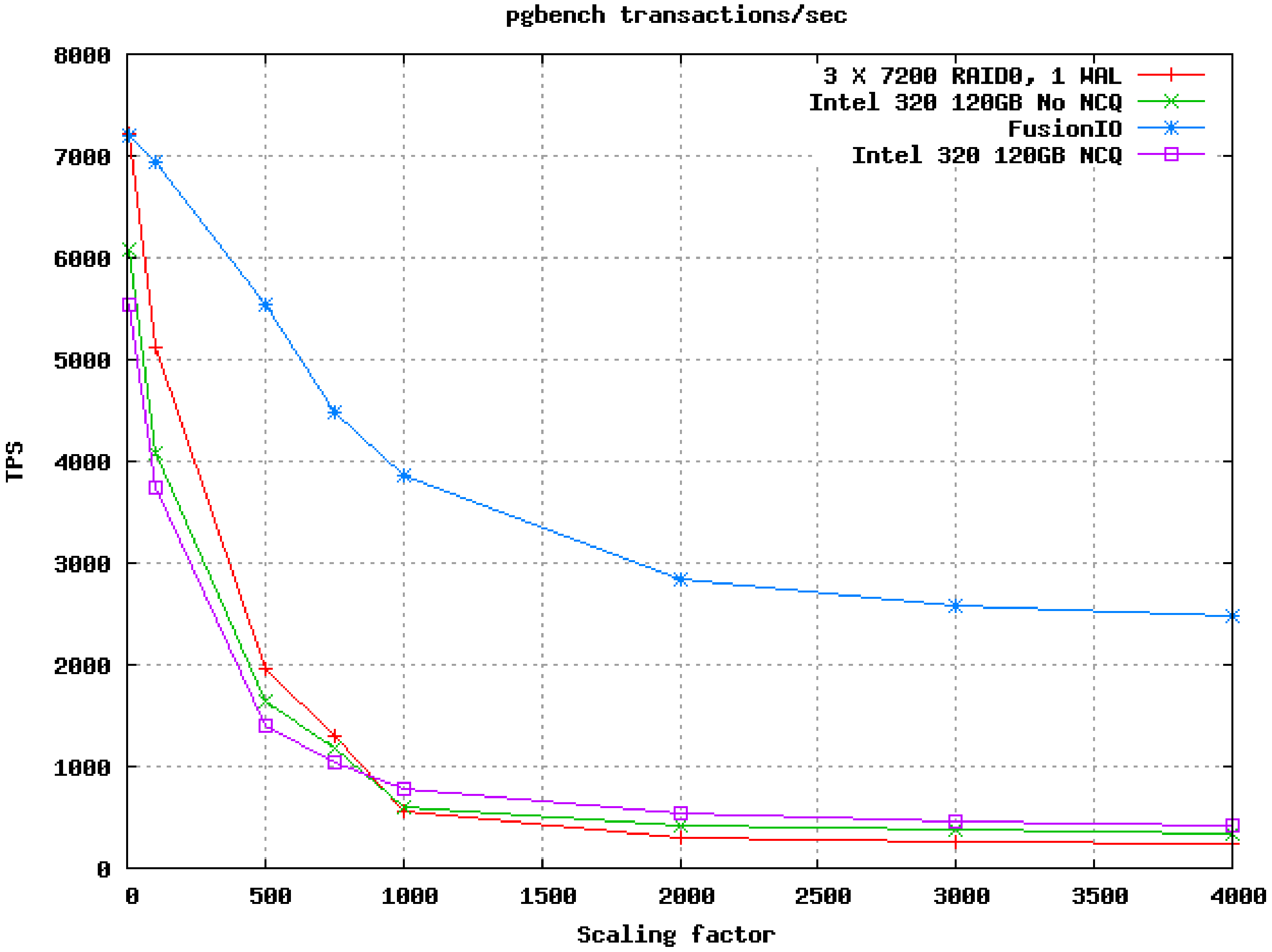
- Intel 320 Series drive didn't enable NCQ
- Should have scaled smoothly to handle 32 concurrent readers
- Instead rate was flat, showing no queue
- Motherboard BIOS fix enabled NCQ
- Check Linux with:

```
cat /sys/block/sdb/device/queue_depth
```

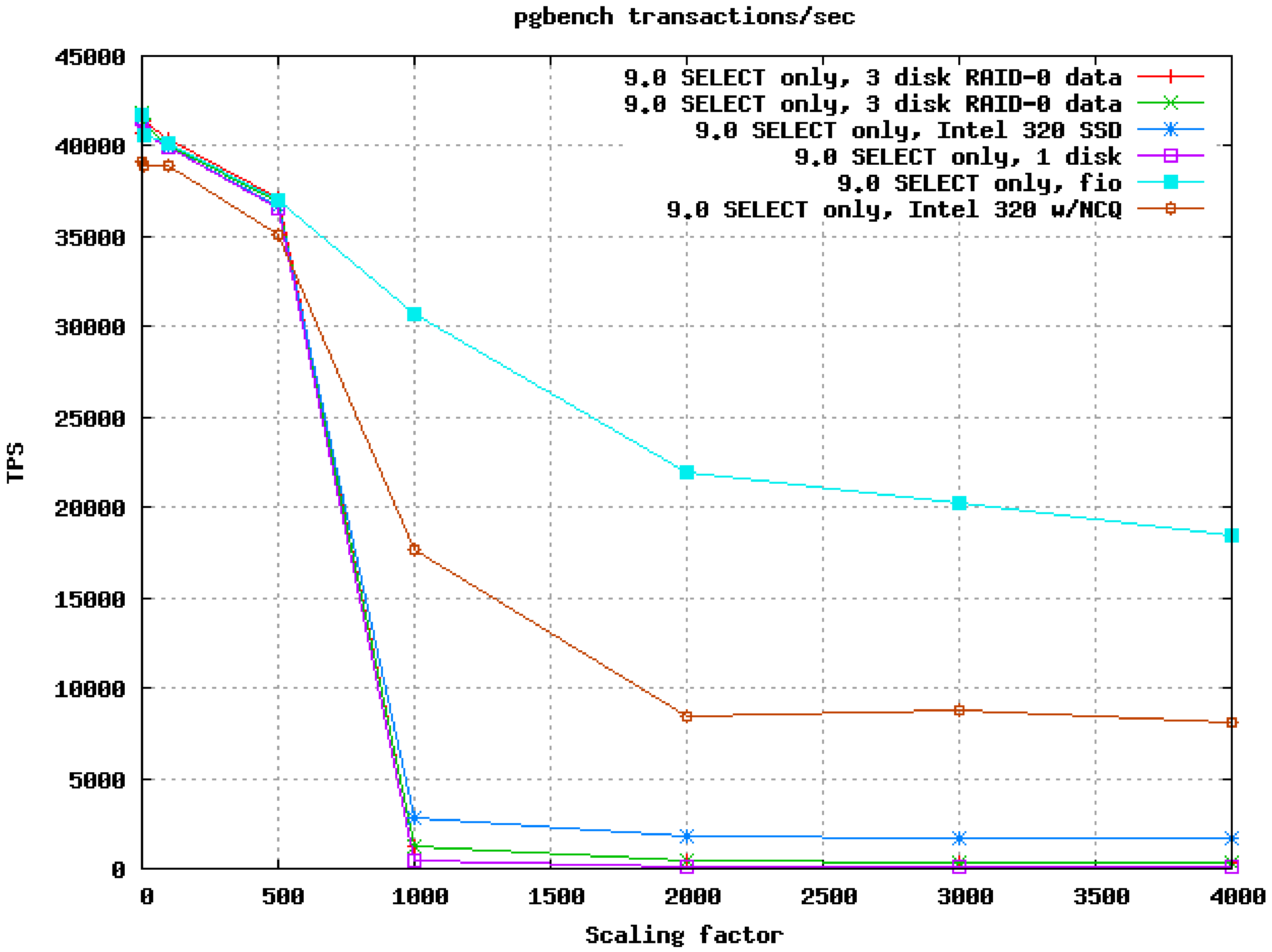
Intel 320 NCQ Speedup



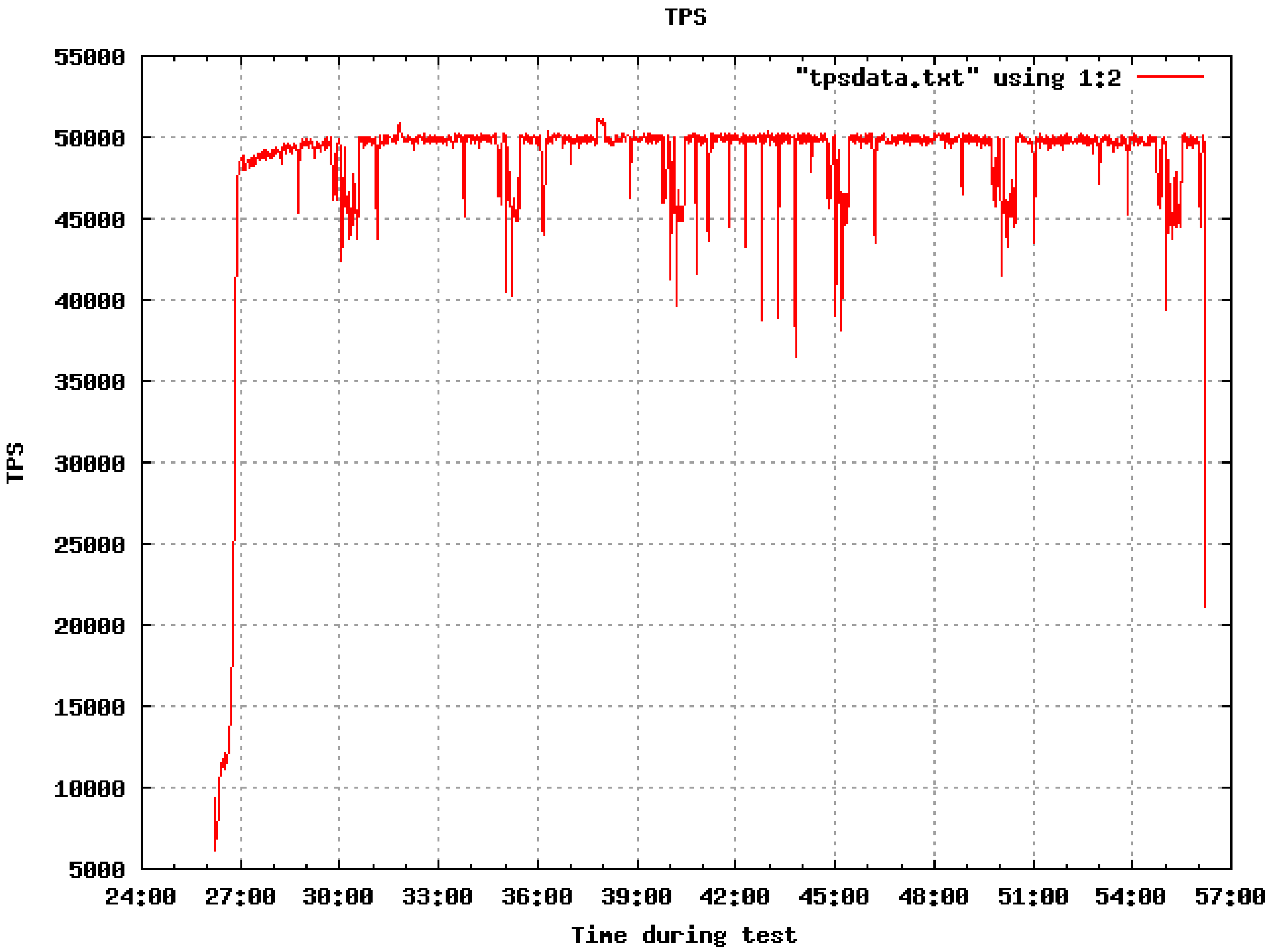
pgbench TPC-B writes



Random reads



Intel 320 w/NCQ: 1 minute refill



Measured refill rates

- 3 disk RAID-0: 7 to 15MB/s
- Intel SSD without NCQ: 29 to 32MB/s
- Intel SSD with NCQ: 160 to 192MB/s
- Fusion-io ioDrive: 583 to 621MB/s

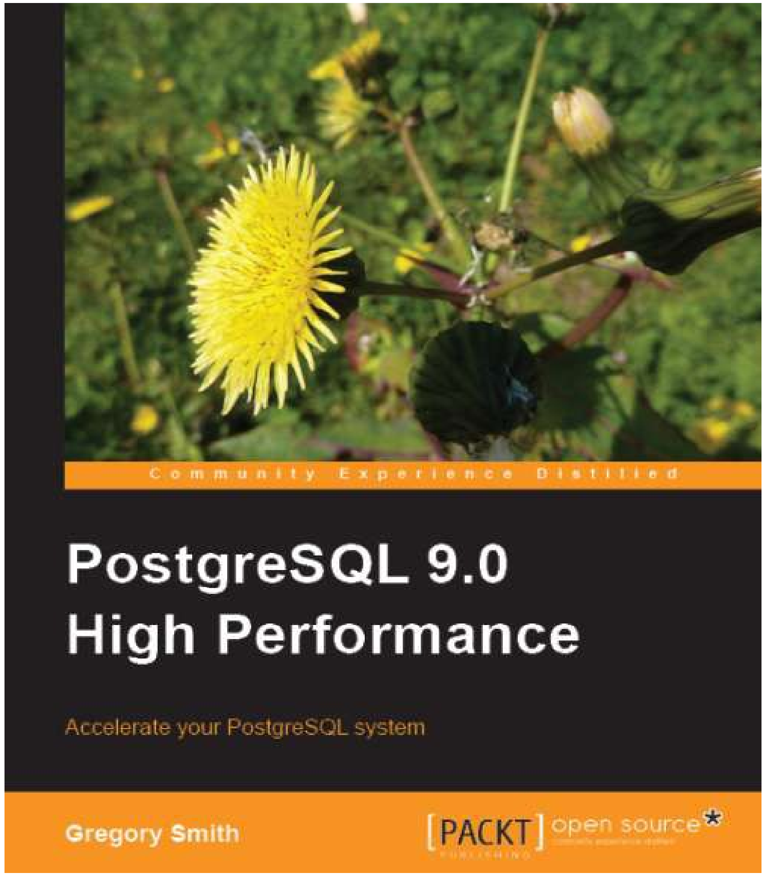
PostgreSQL Papers

- Greg Smith greg@2ndQuadrant.com
- Talks: http://www.2ndquadrant.com/en/talks/
- Blog: http://blog.2ndquadrant.com/
- Twitter: @postgresperf

-
  -

PostgreSQL Books & Talks

http://www.2ndquadrant.com/books



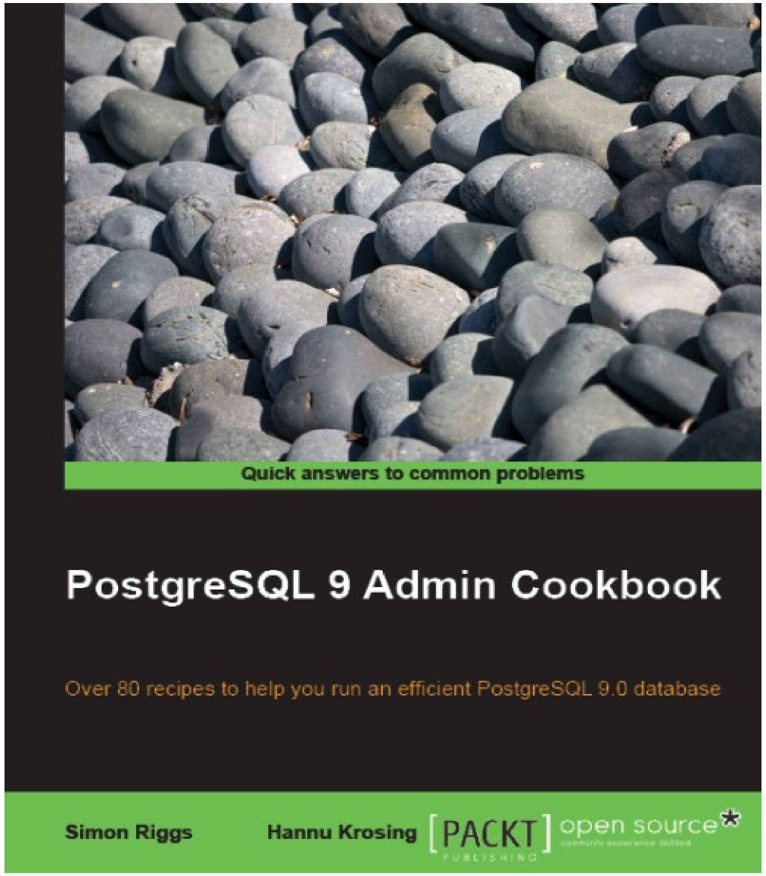
Community Experience Distilled

PostgreSQL 9.0 High Performance

Accelerate your PostgreSQL system

Gregory Smith

[PACKT] open source*



Quick answers to common problems

PostgreSQL 9 Admin Cookbook

Over 80 recipes to help you run an efficient PostgreSQL 9.0 database

Simon Riggs

Hannu Krosing

[PACKT] open source*