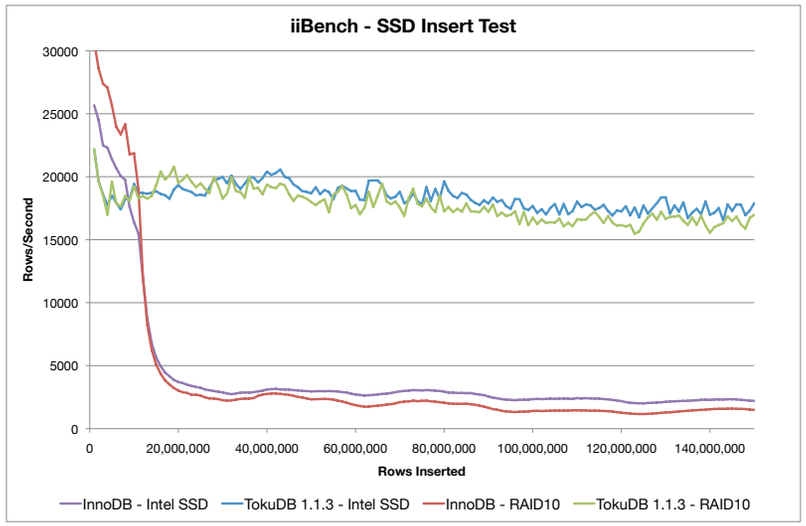
A Performance Puzzle: B-Tree Insertions are Slow on SSDs

or

What Is a Performance Model for SSDs?

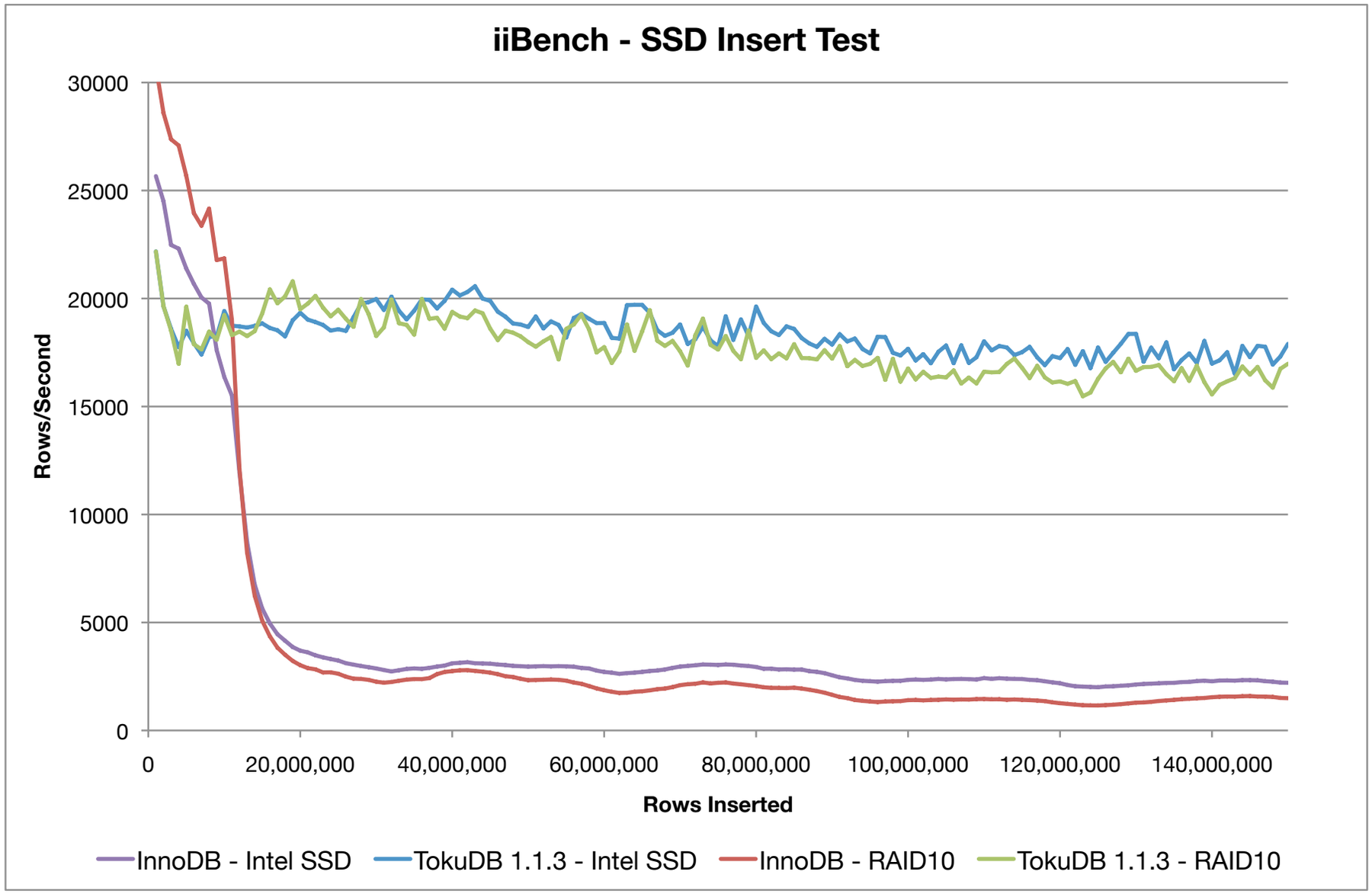
Bradley C. Kuszmaul
MIT CSAIL, & Tokutek



Motivation: I want to understand SSD performance so I can design fast data structures.

HPTS 2009

Poor MYSQL B-Tree SSD Performance?



Surprisingly, disk almost as good as SSD. InnoDB's insertion buffer helps. Not CPU bound.

Intel X25E Specifications

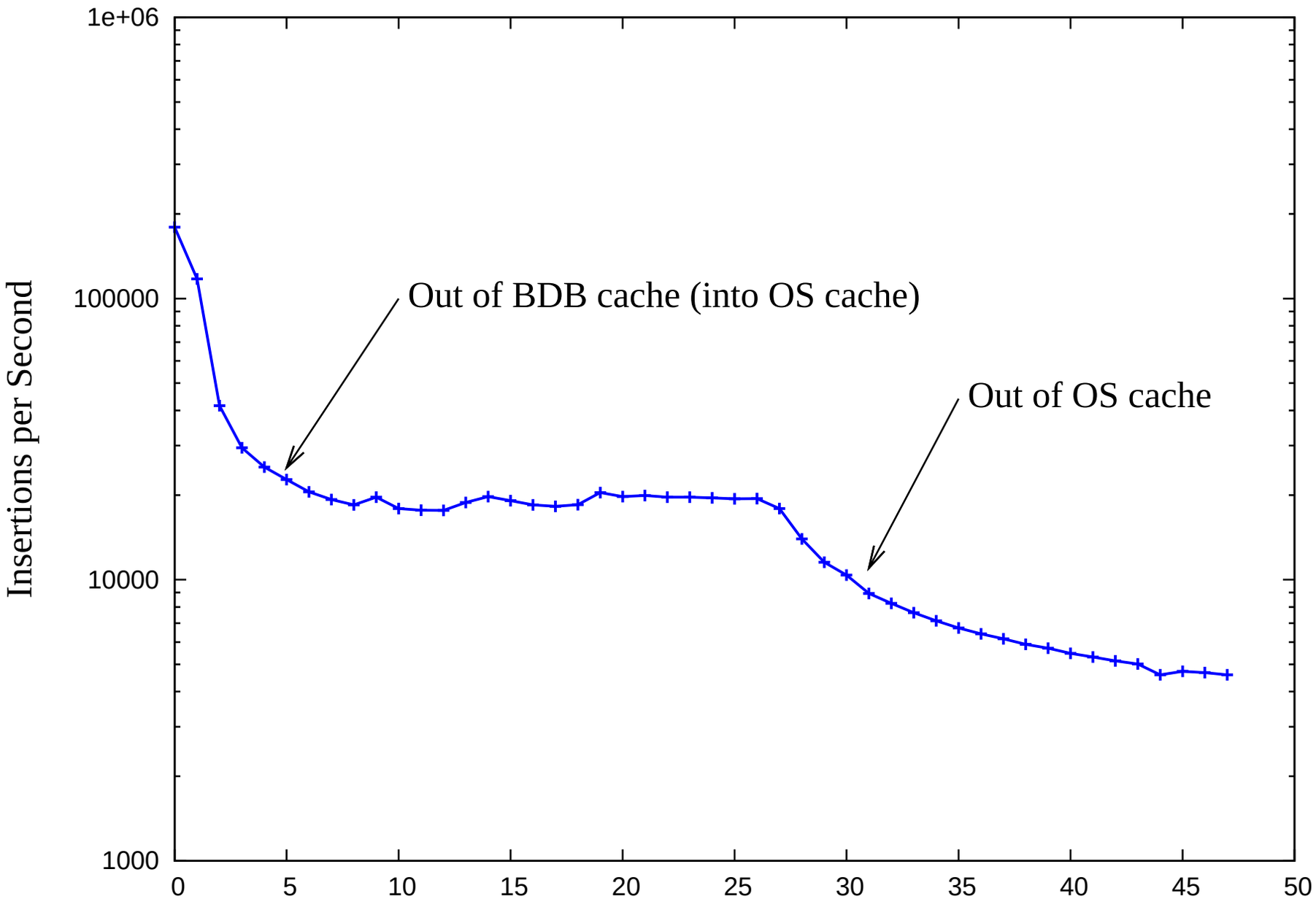
- Read bandwidth up to 250 MB/s.
- Write bandwidth up to 170 MB/s.
- Random 4KB read rate: 35 KIO/s.
- Random 4KB write reate: 3.5 KIO/s.

Disk bandwidth (5 disk RAID):

- Read/Write bandwidth about 400 MB/s.
- Random Read/Write rate: 600/s (with the wind at your back.)

So why isn't the SSD giving InnoDB a 6x performance boost?

MySQL Complex ⇒ Measure Berkeley DB



Trending to 4500 writes per second (still dropping...)

Berkeley DB too complex ⇒ Try File I/O

Method:

- Build a 12GB file on a machine with 3GB RAM.
- Perform random reads and writes of various sizes.
- Build a performance model.

Still strange and unpredictable.

A Performance Model

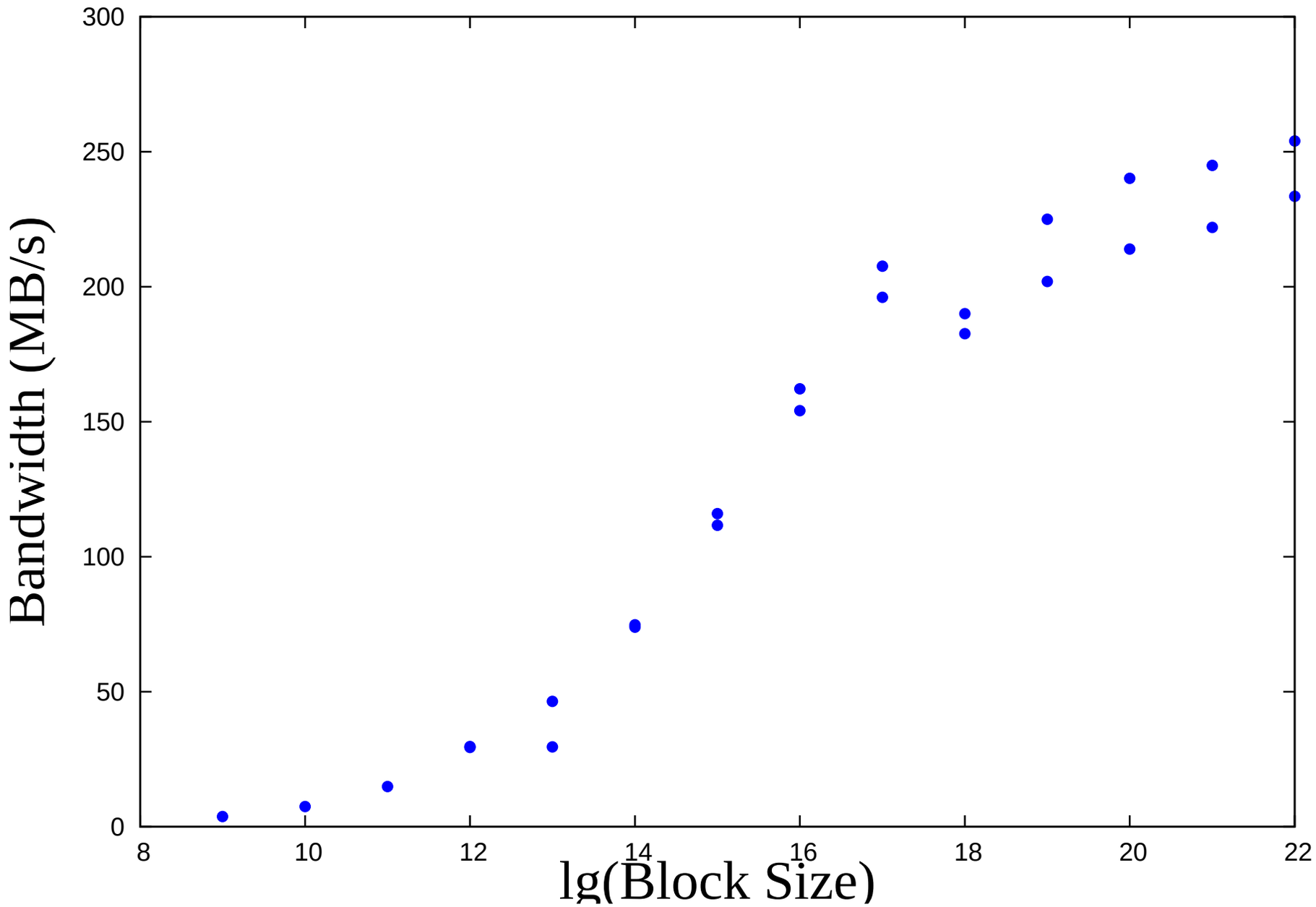
Can I make the following performance model work?
When reading a block of size B,

- There is a startup cost, S, ("seek time")
- There is bandwidth, W, ("transfer rate").

The simple model is thus

$$T_R = S_R + B/W_R$$
$$T_W = S_W + B/W_W$$

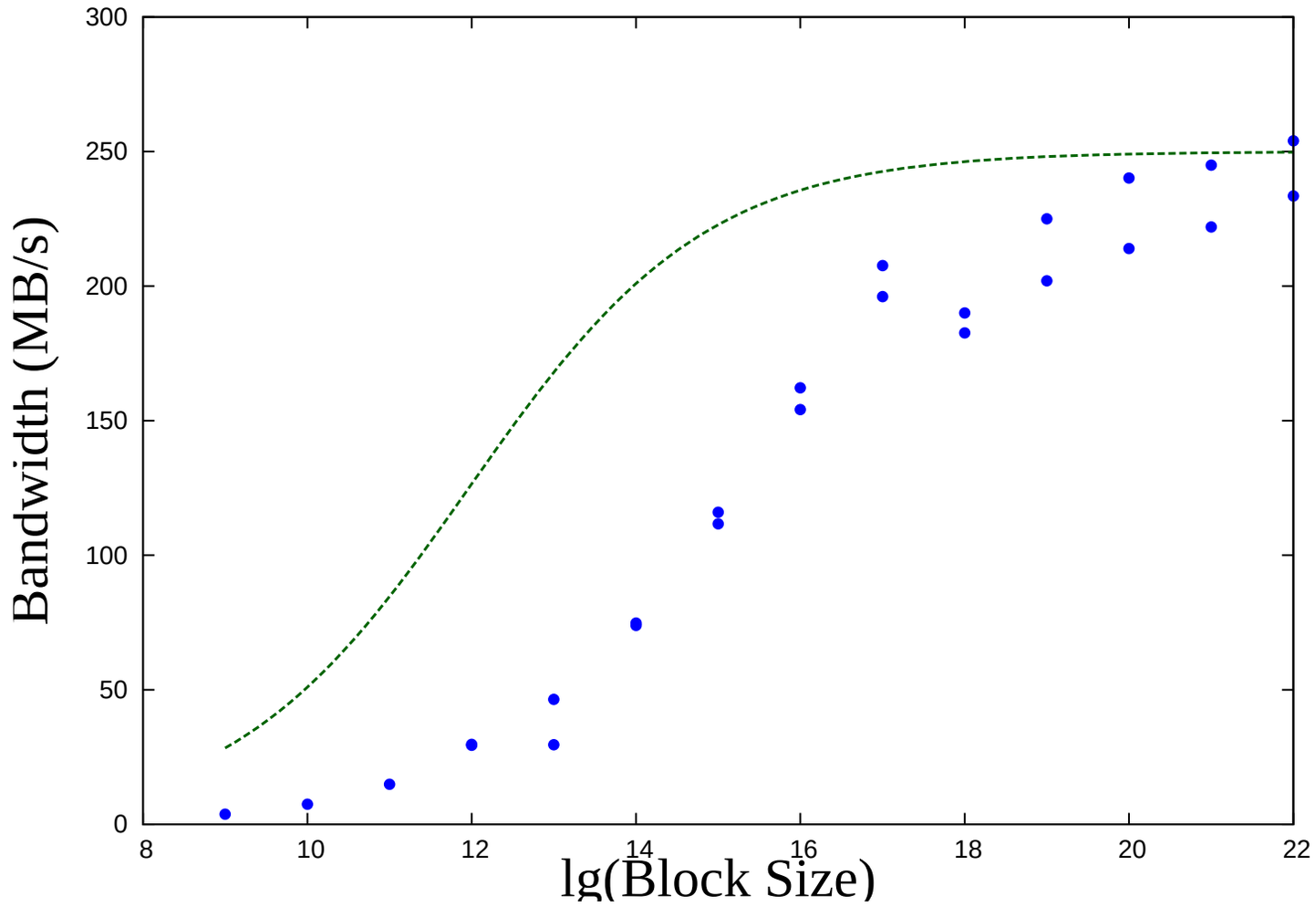
Read Performance as a Function of Block Size



A Model from the Data Sheet?

The Manufacturer's 35 KIO/s and 250 MB/s suggests this (poor) model:

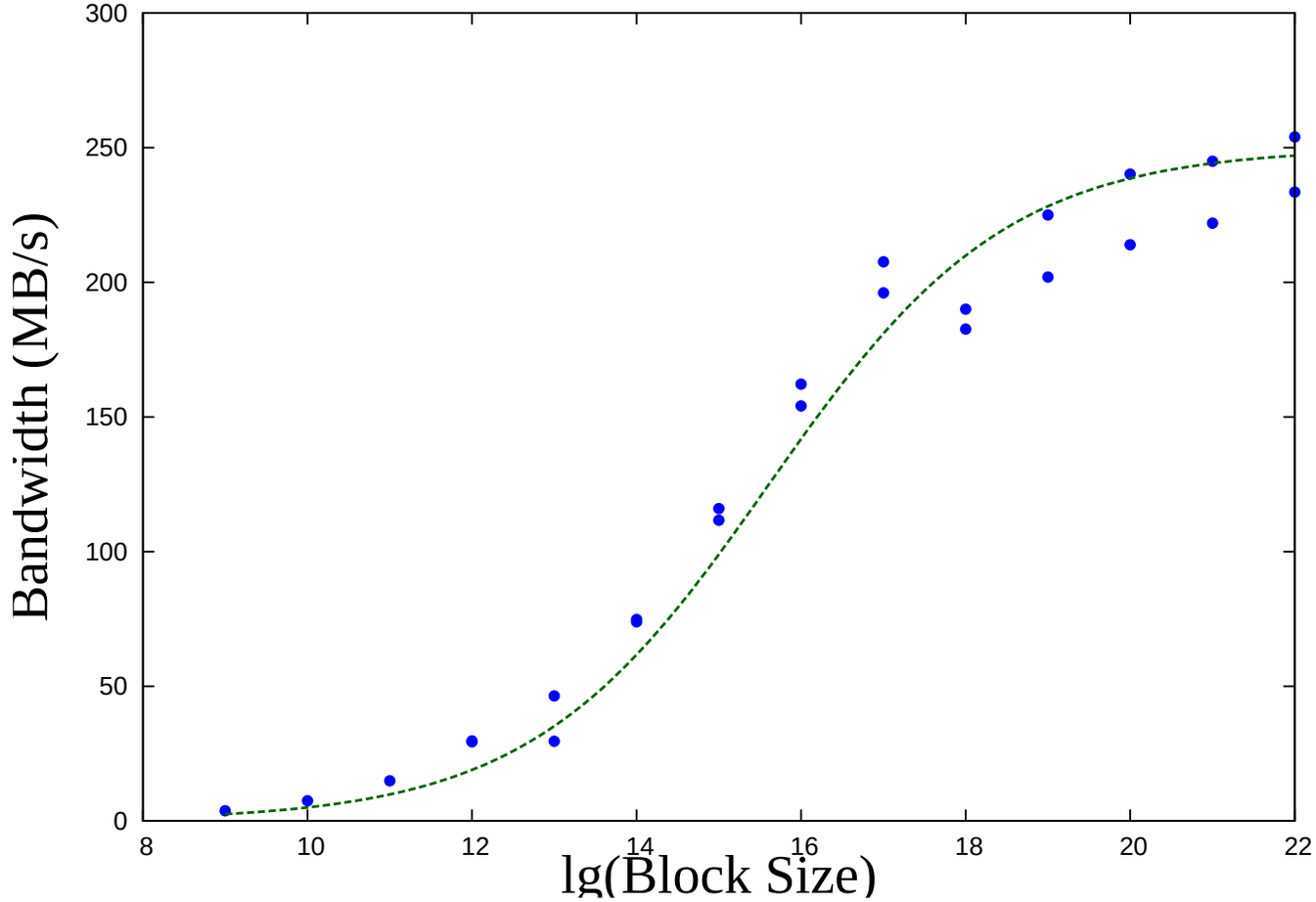
$$T_B = 16\mu\text{s} + B/(250\text{MB/s}).$$



A Model from the Data Sheet?

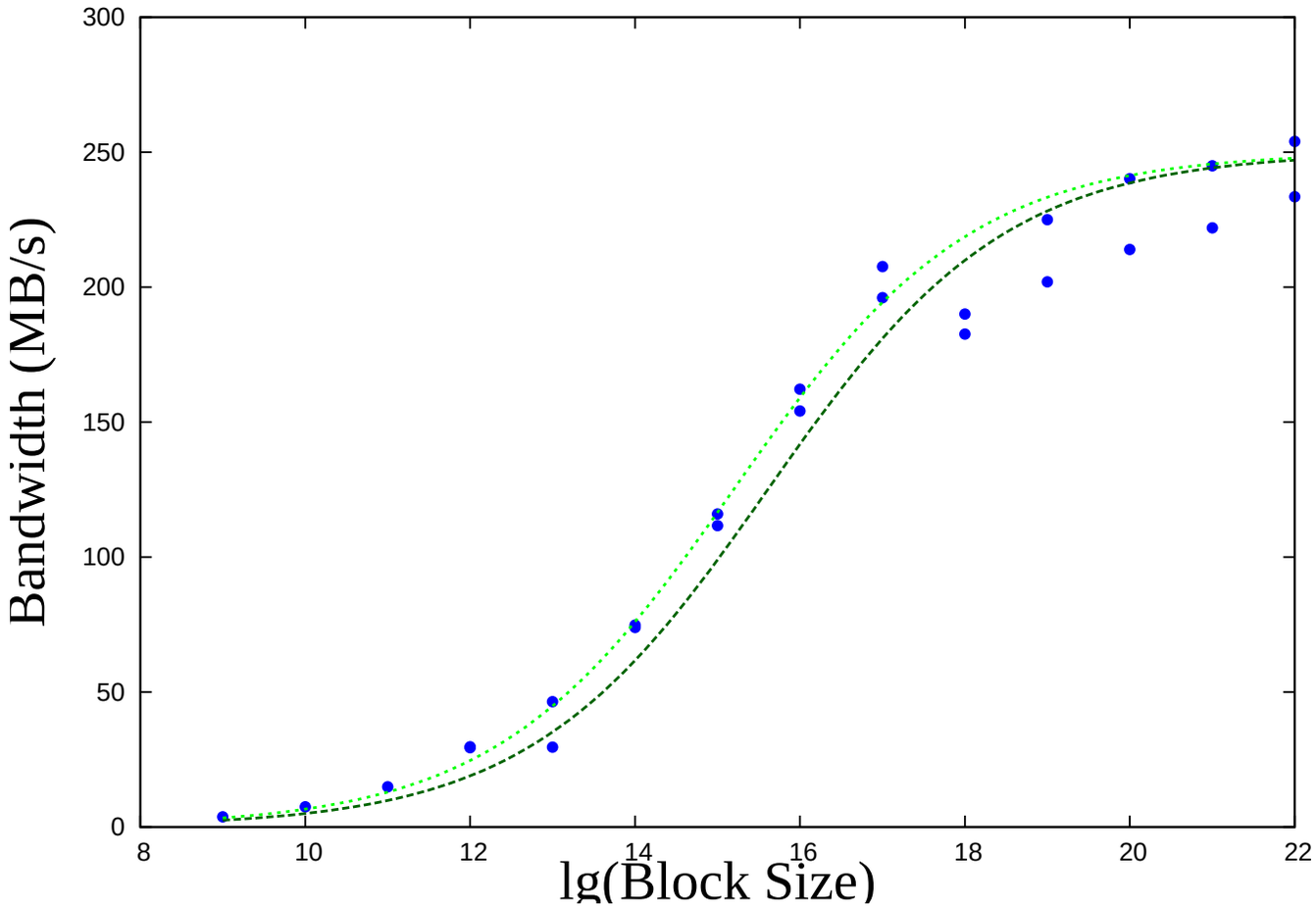
The bandwidth looks good, but I never saw anything like 35,000 IO/s on any workload. Actual read performance is about 10,000 IO/s:

$$T_B = 200\mu\text{s} + B/(250\text{MB/s}).$$



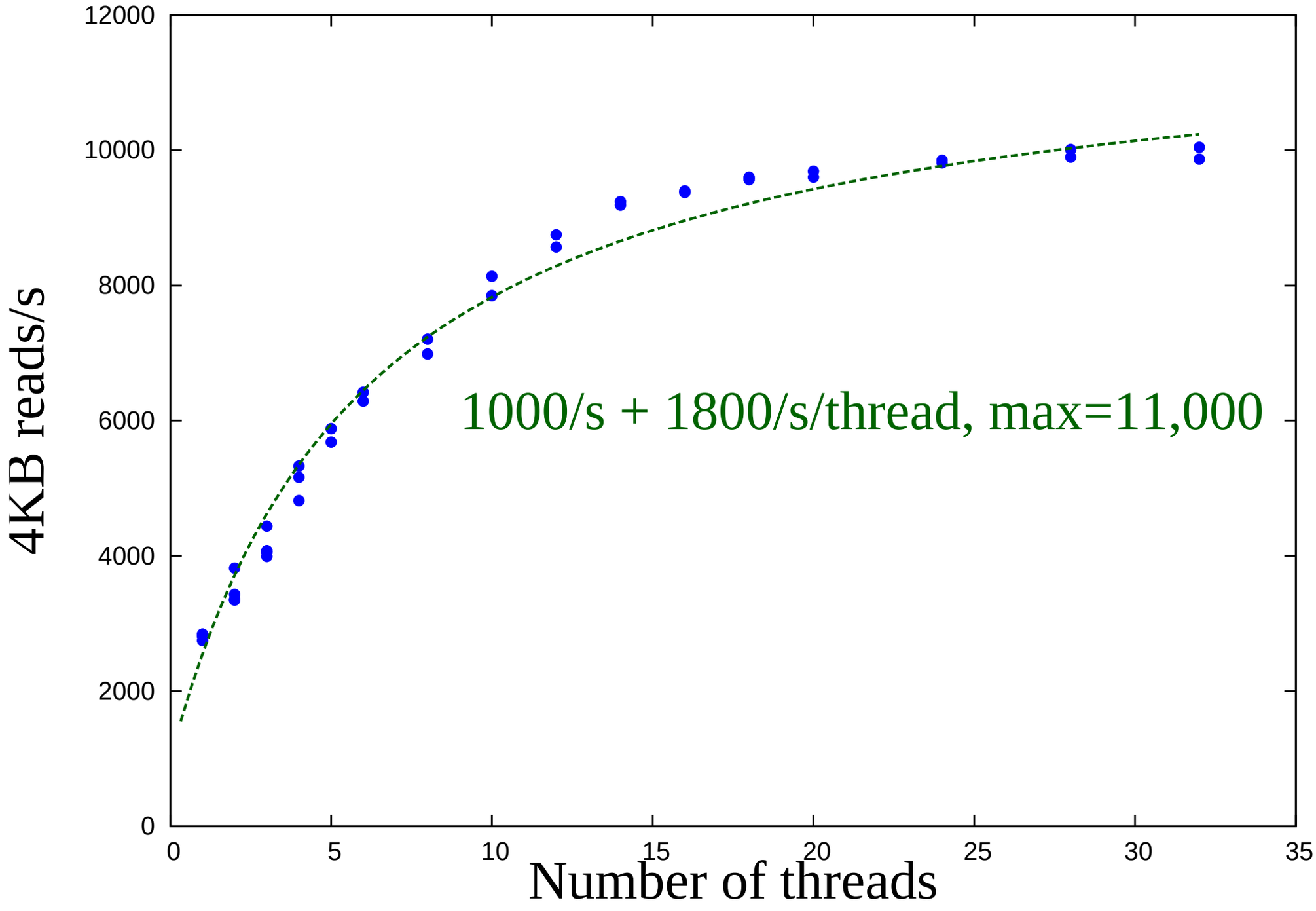
Small Blocks A Little Misleading

For block sizes of less than 4096, the OS first does a read (of 4KB) and then a write.

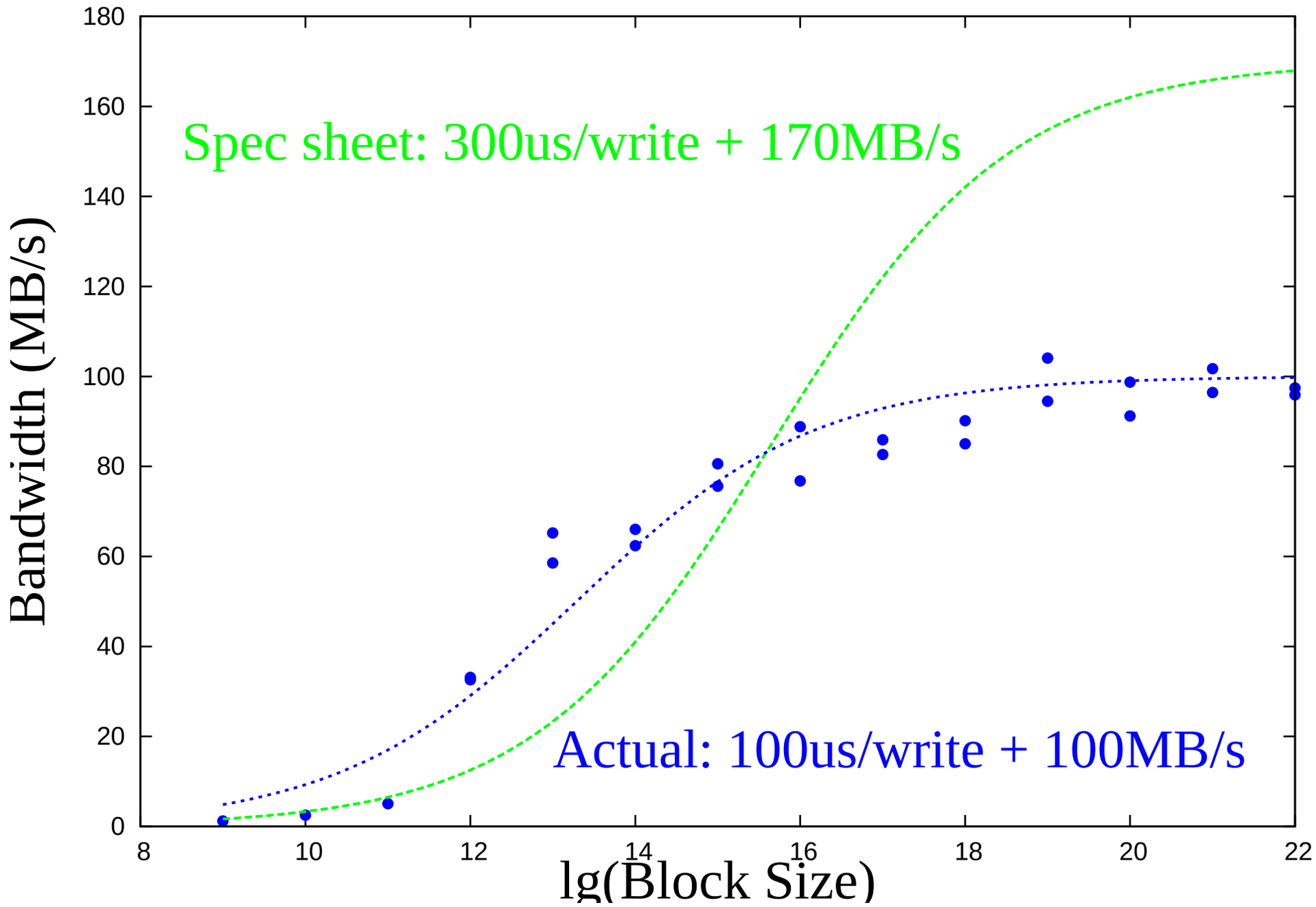


Surprisingly, this doesn't affect the curve much.

Up to 10,000 reads/s with Multithreading



Write Performance



Read-Write Performance

Mixing reads and writes gives the worst of both.

	startup (S)	bandwidth (W)
read	$200\mu s$	$250 MB/s$
write	$100\mu s$	$100 MB/s$
mixed	$200\mu s$	$100 MB/s$

What Block Size To Use?

For point-queries, B-trees are insensitive to block size. As soon as you have any reasonable fanout you do well.

For range queries, the block size is important.

Tension:

- Large block sizes make range queries faster.
- Large block sizes make point queries slower.

Half-power point

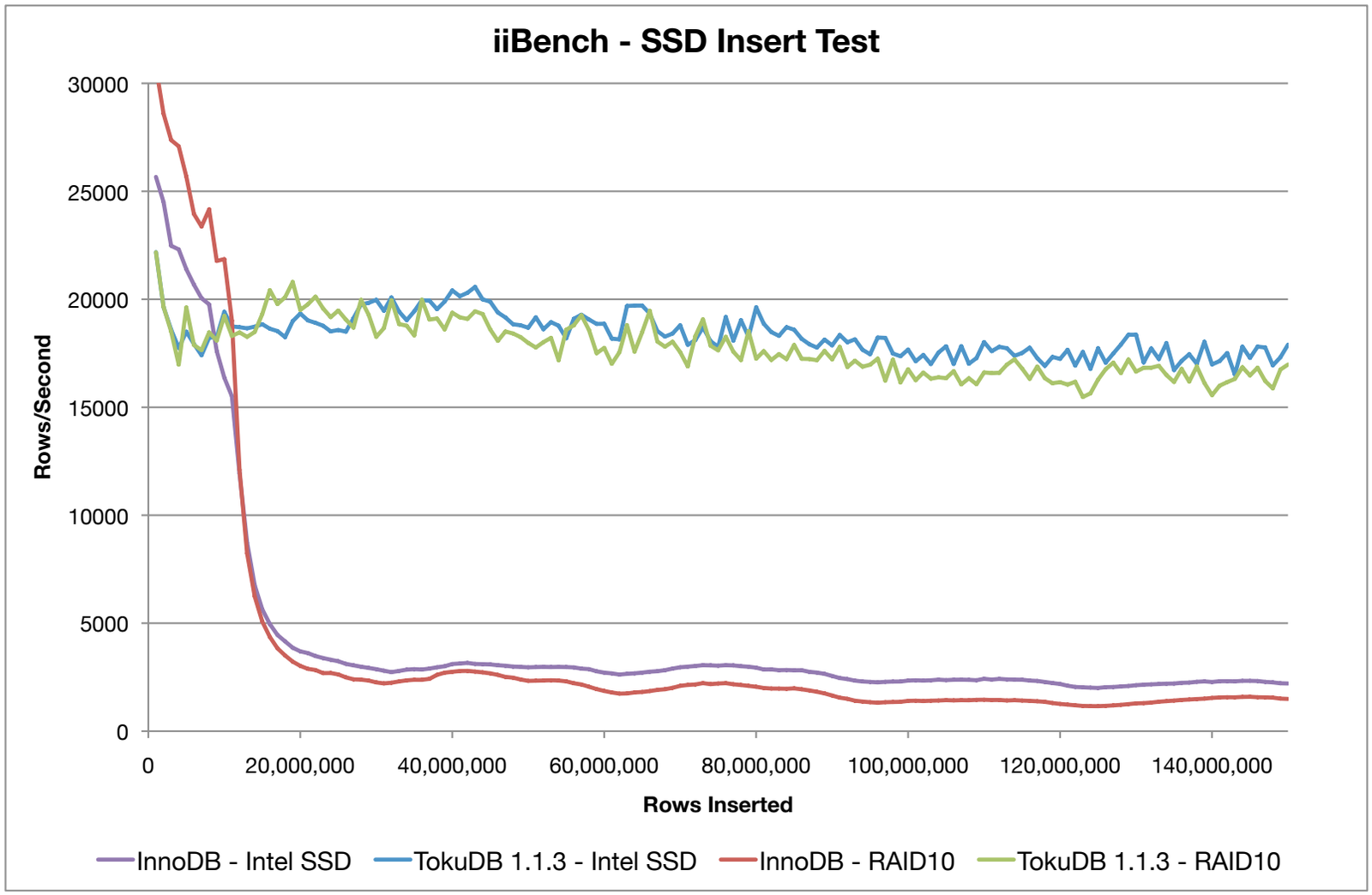
Idea: Set block size so that half the time is accounted for the "seek time", and half the time by "bandwidth".

"Half Power Point"

	SSD	Rotating Disk
read	50KB	0.5MB–1MB
write	10KB	0.5MB–1MB
read/write mix	21KB	0.5MB–1MB

Cache-Oblivious Approach

- Use data structures that are fast for any block size.
- Can also speed insertions without slowing searches.



Tokutek's MySQL storage engine uses these ideas.