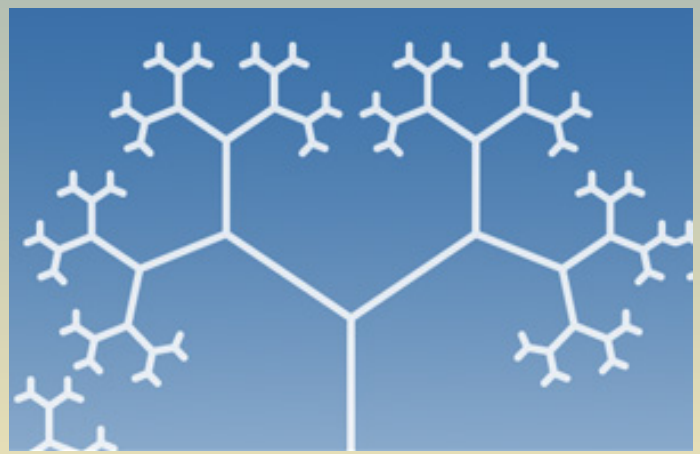
Tokutek®

Fractal Tree® Technology Overview

The Art of Indexing

Martín Farach-Colton
Co-founder & Chief Technology Officer



Not all indexing is the same

B-tree is the basis for almost all DB systems

- Data structure invented in 1972
- Has not kept up with hardware trends
  - Works poorly on modern rotational disks
  - Works poorly on SSD

Fractal Tree Indexes is the basis of TokuDB

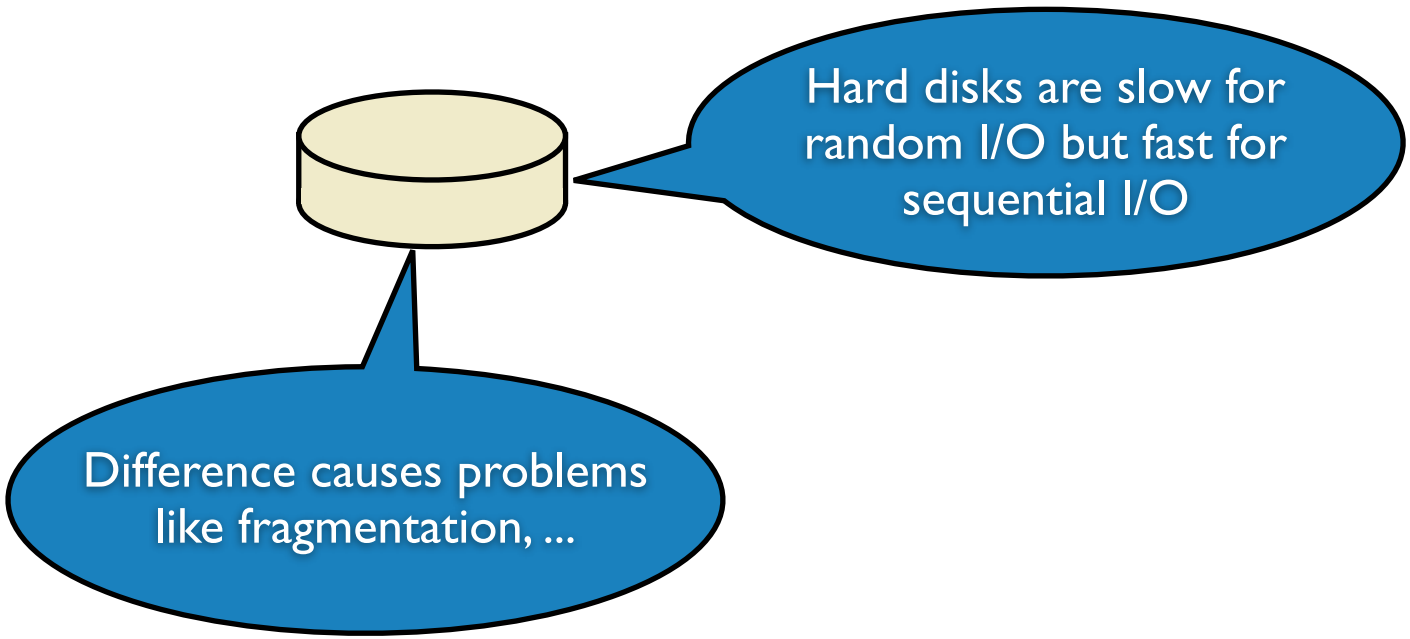
- Scales with hardware
- Fast Indexing ➔ More Indexing ➔ Faster Queries
- Great Compression
- No Fragmentation
- Reduced wear on SSDs

How do Fractal Tree Indexes outperform B-trees?

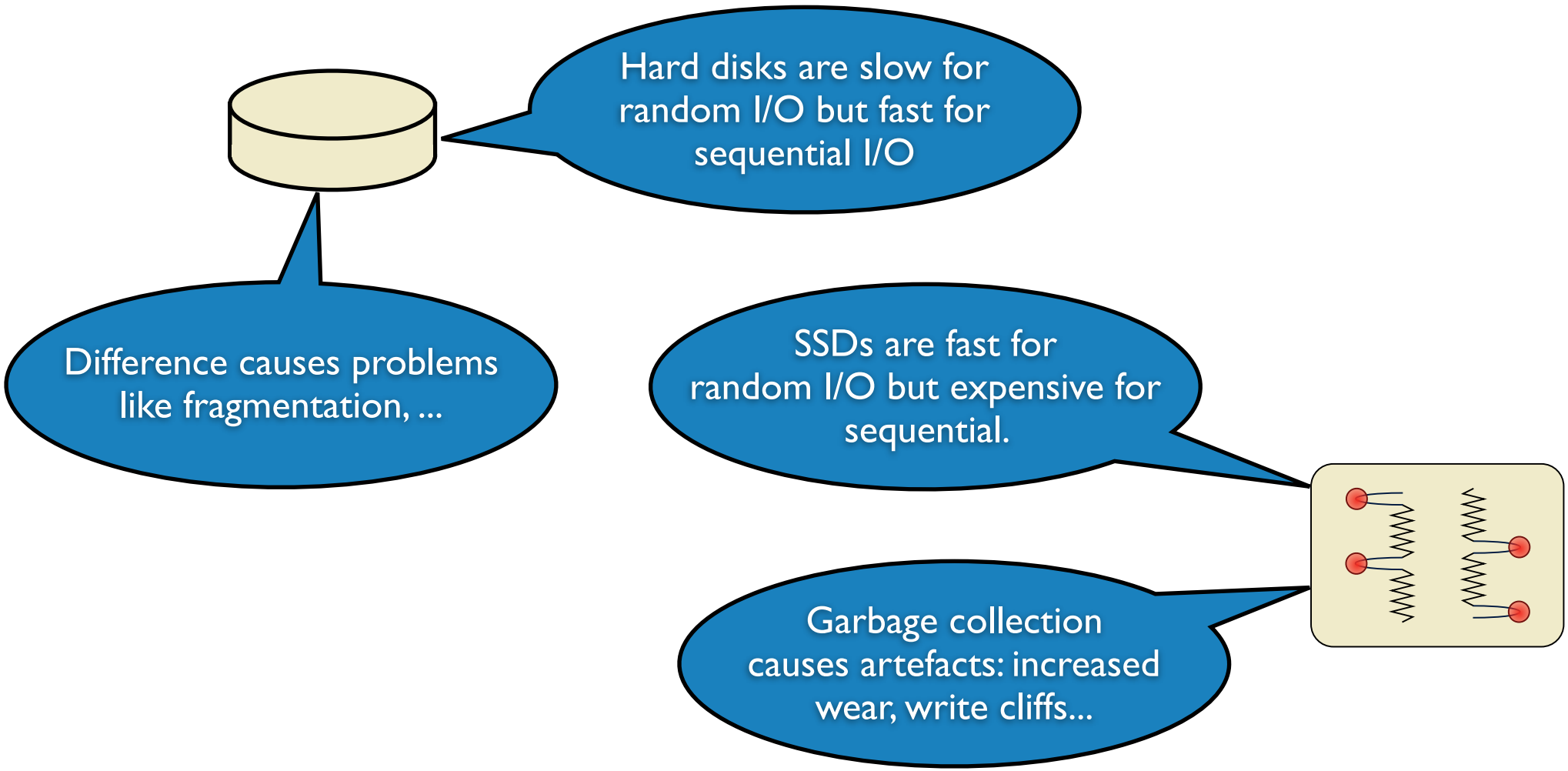
How do Fractal Tree Indexes outperform B-trees?

First, some facts about storage systems

Storage is quirky



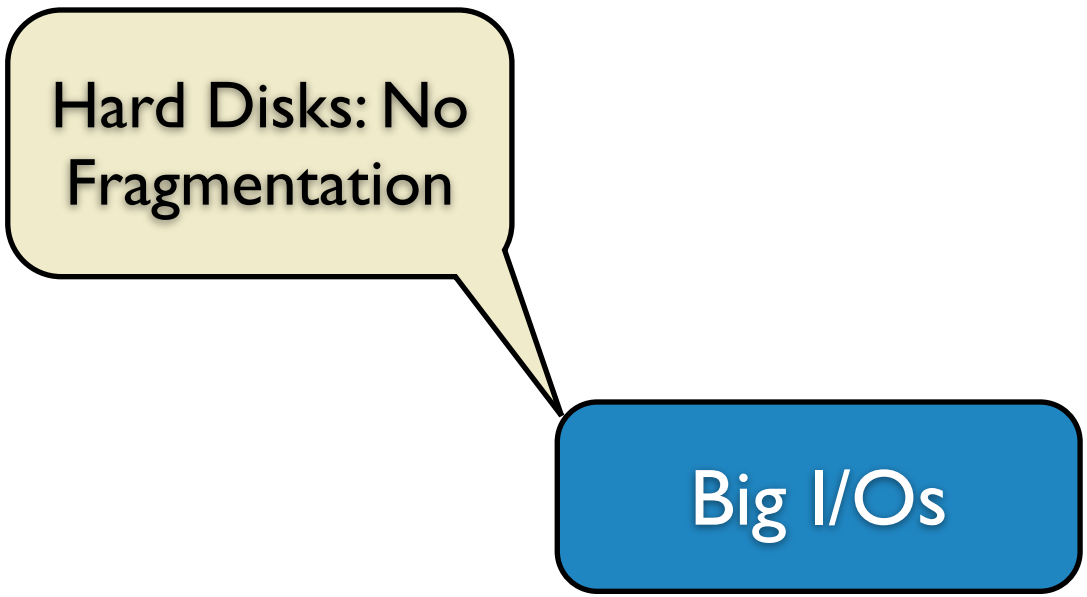
Storage is quirky



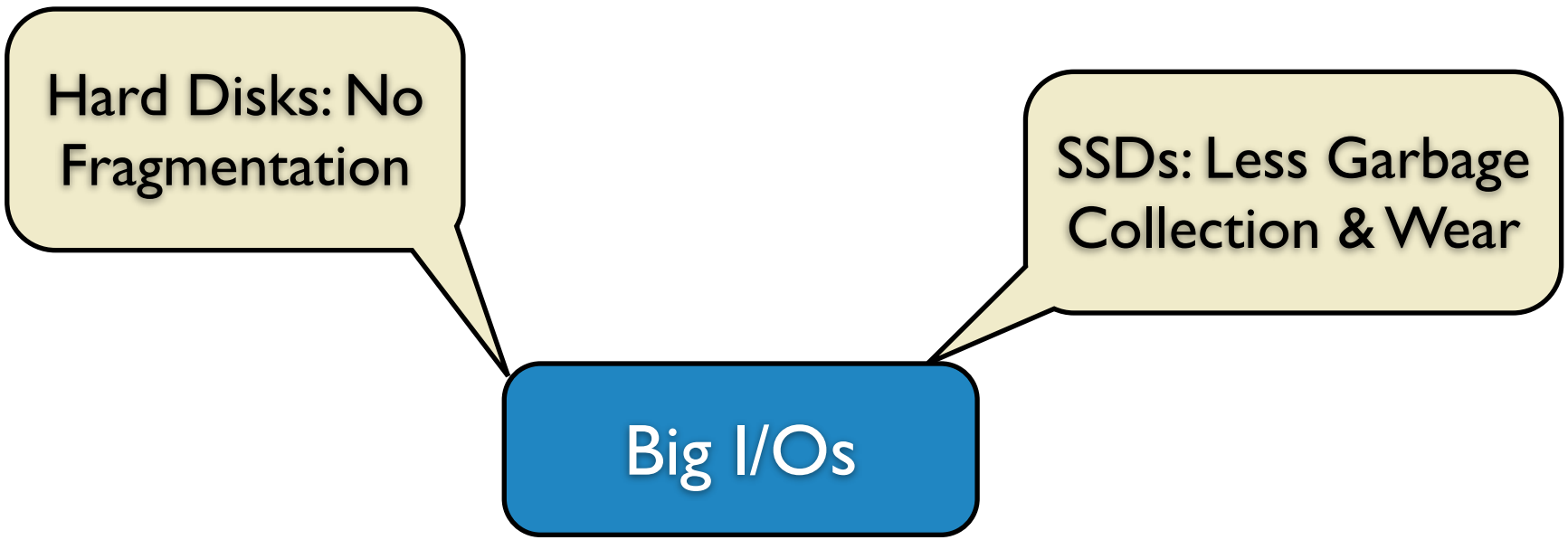
Storage ♡ Big Reads and Writes

Big I/Os

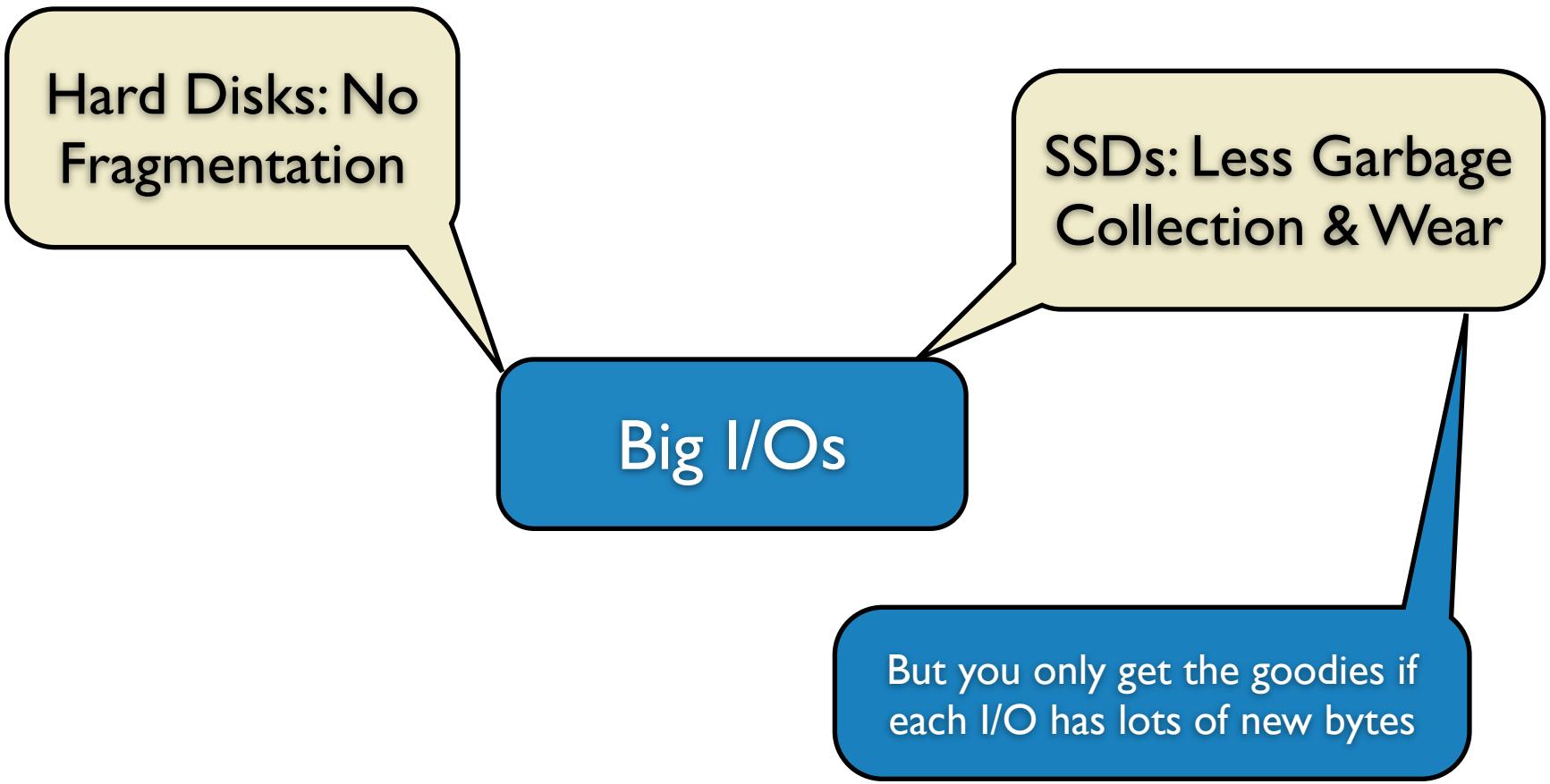
Storage ♡ Big Reads and Writes



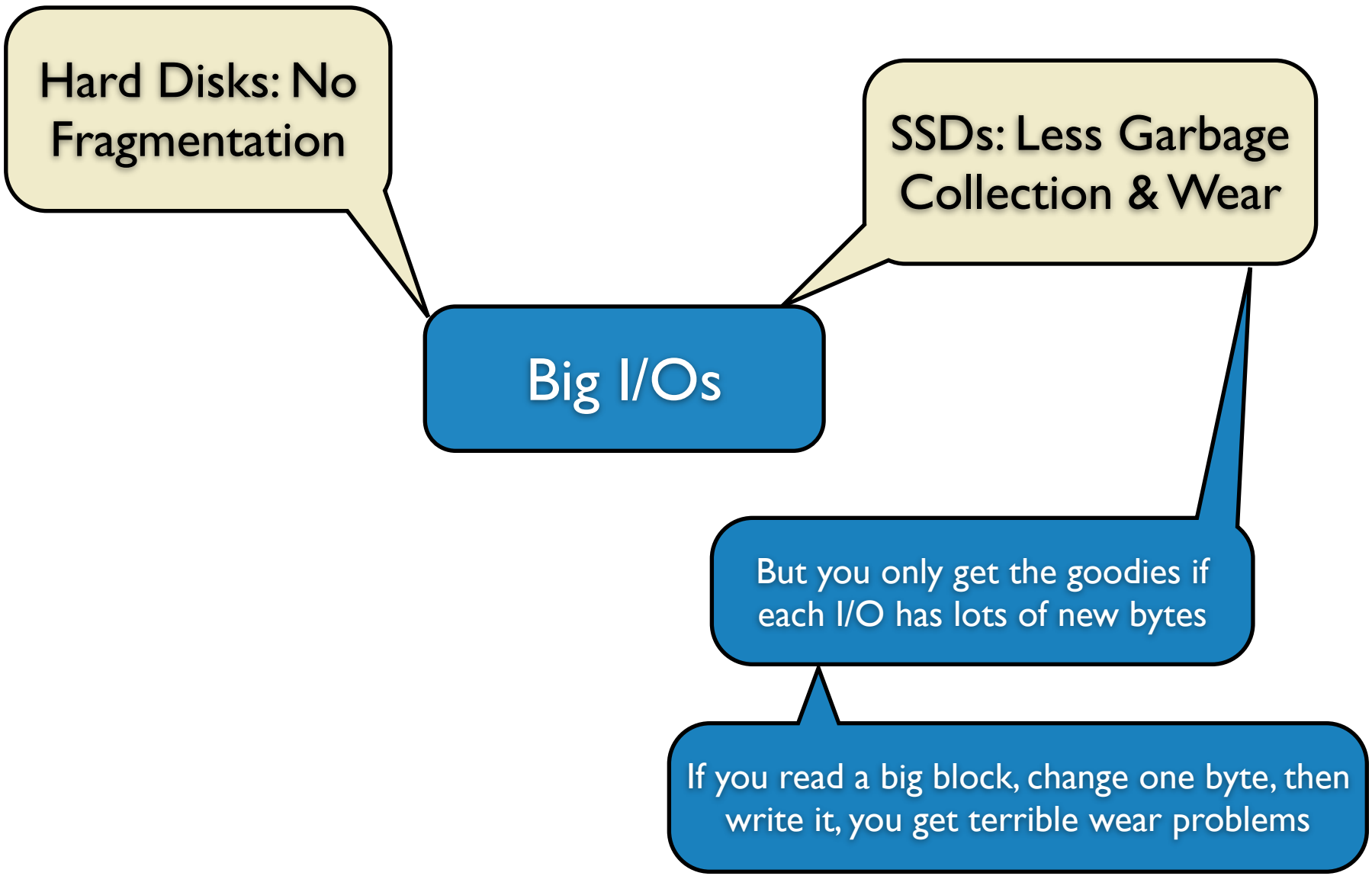
Storage ♡ Big Reads and Writes



Storage ♡ Big Reads and Writes

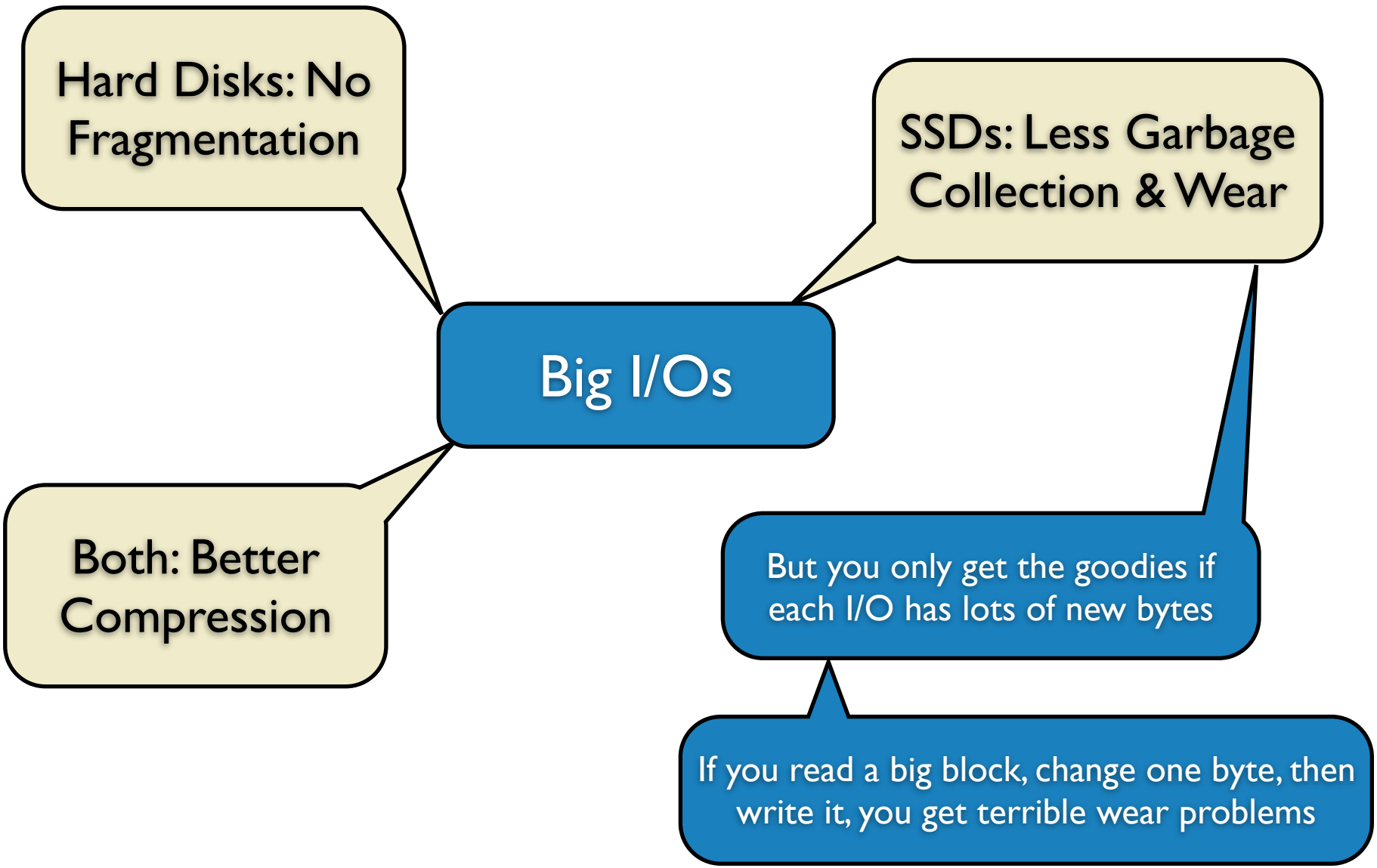


Storage ♡ Big Reads and Writes



Tokutek®

Storage ♡ Big Reads and Writes



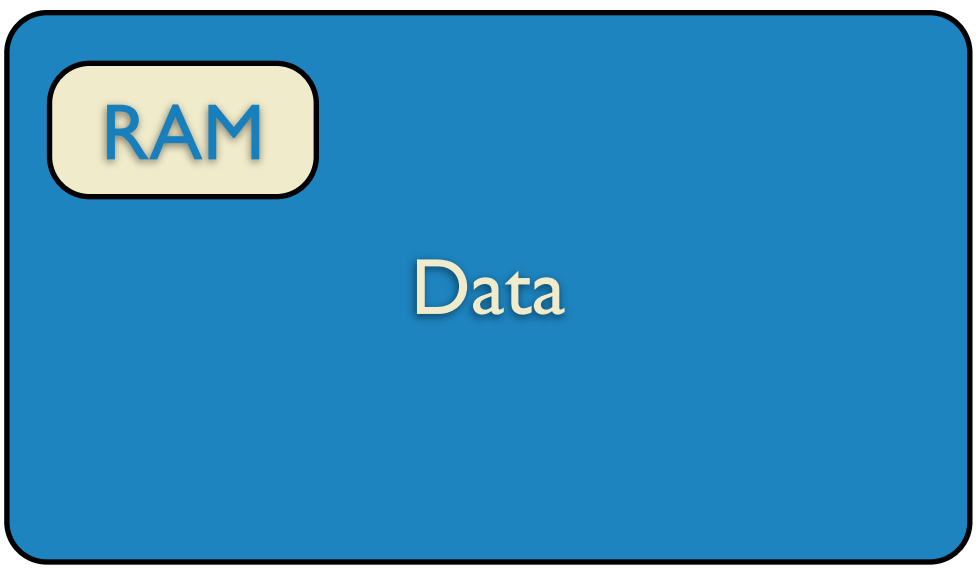
Tokutek®

Data is big, RAM is Small

Caching is great

- But you can't cache all your data
- For stuff not in memory, you have to go to disk

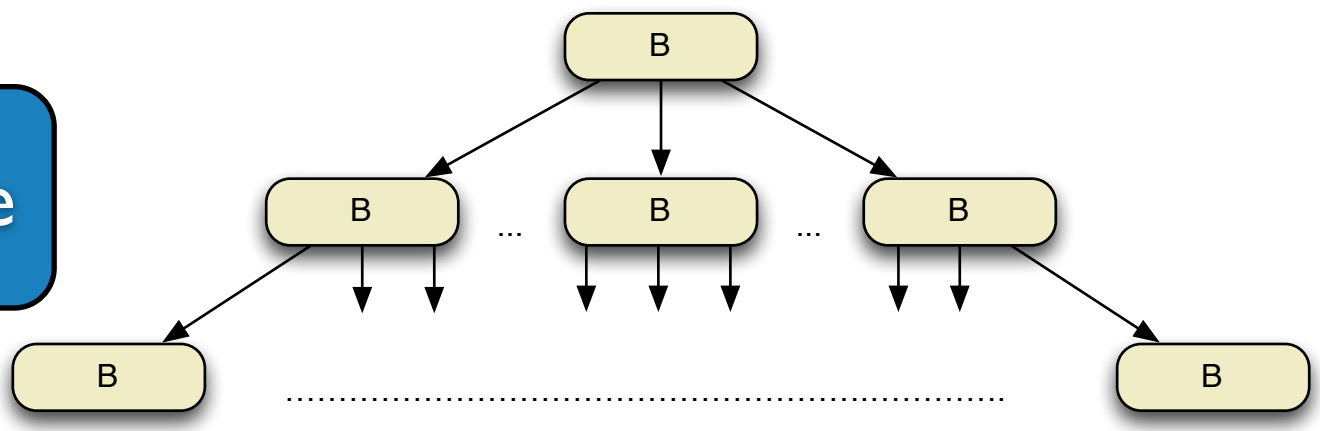
Goal: Do the best we can for the stuff on disk



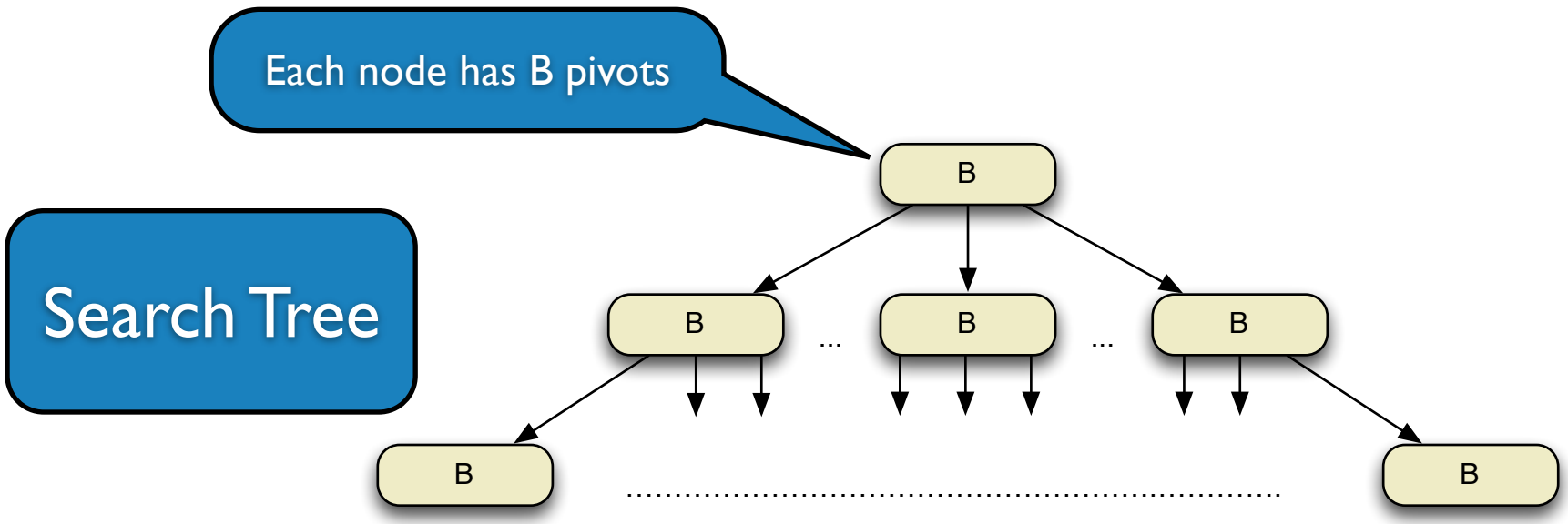
Now, What's a B-tree? & a Fractal Tree Index

What's a B-tree?

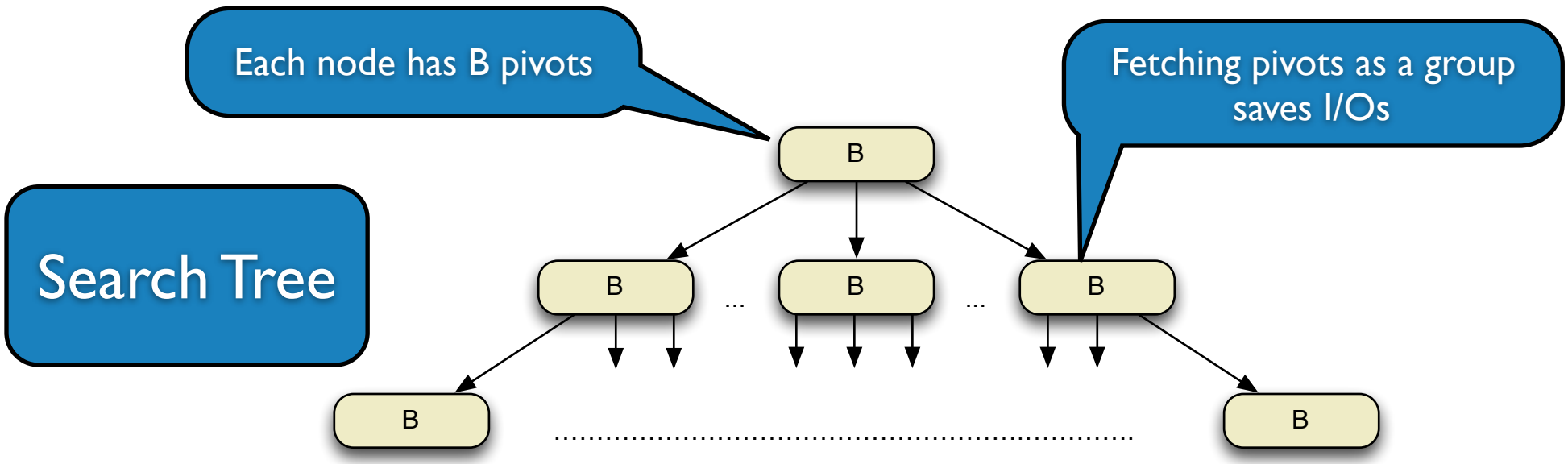
Search Tree



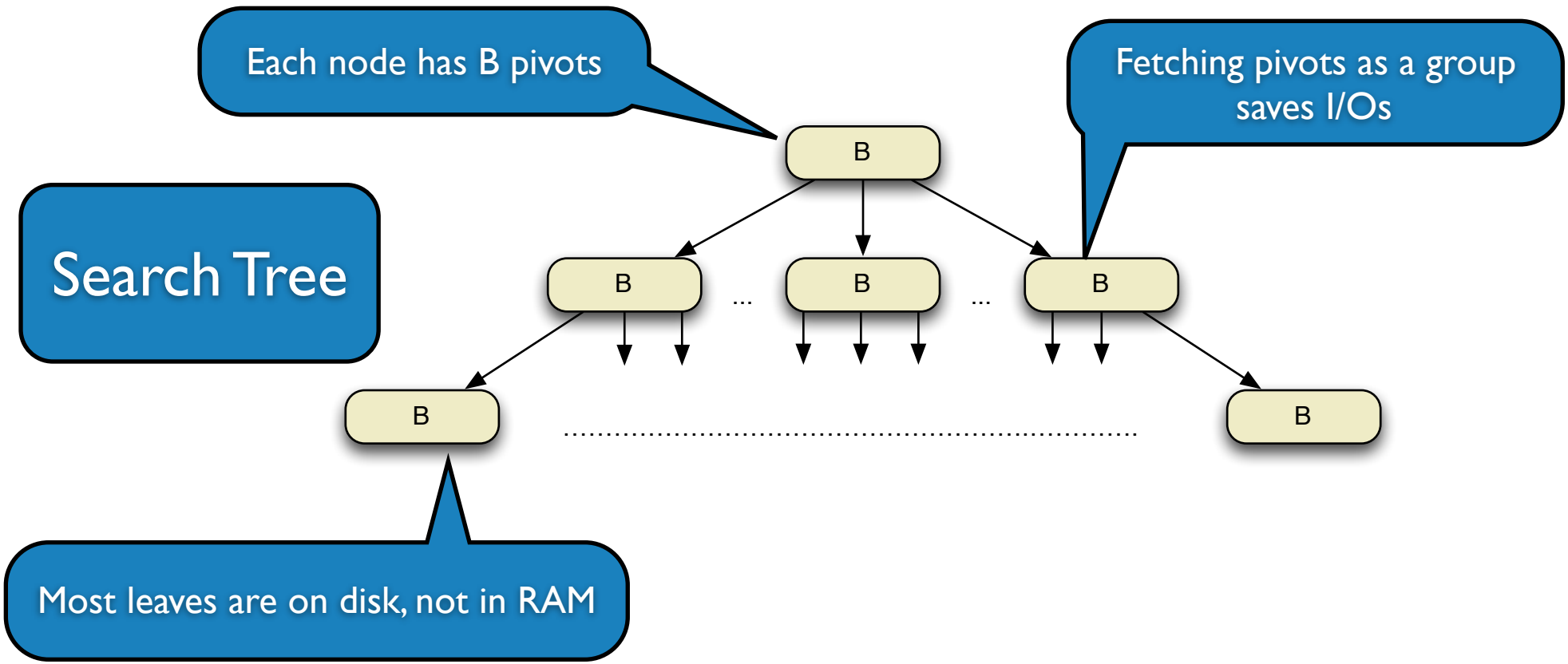
What's a B-tree?



What's a B-tree?

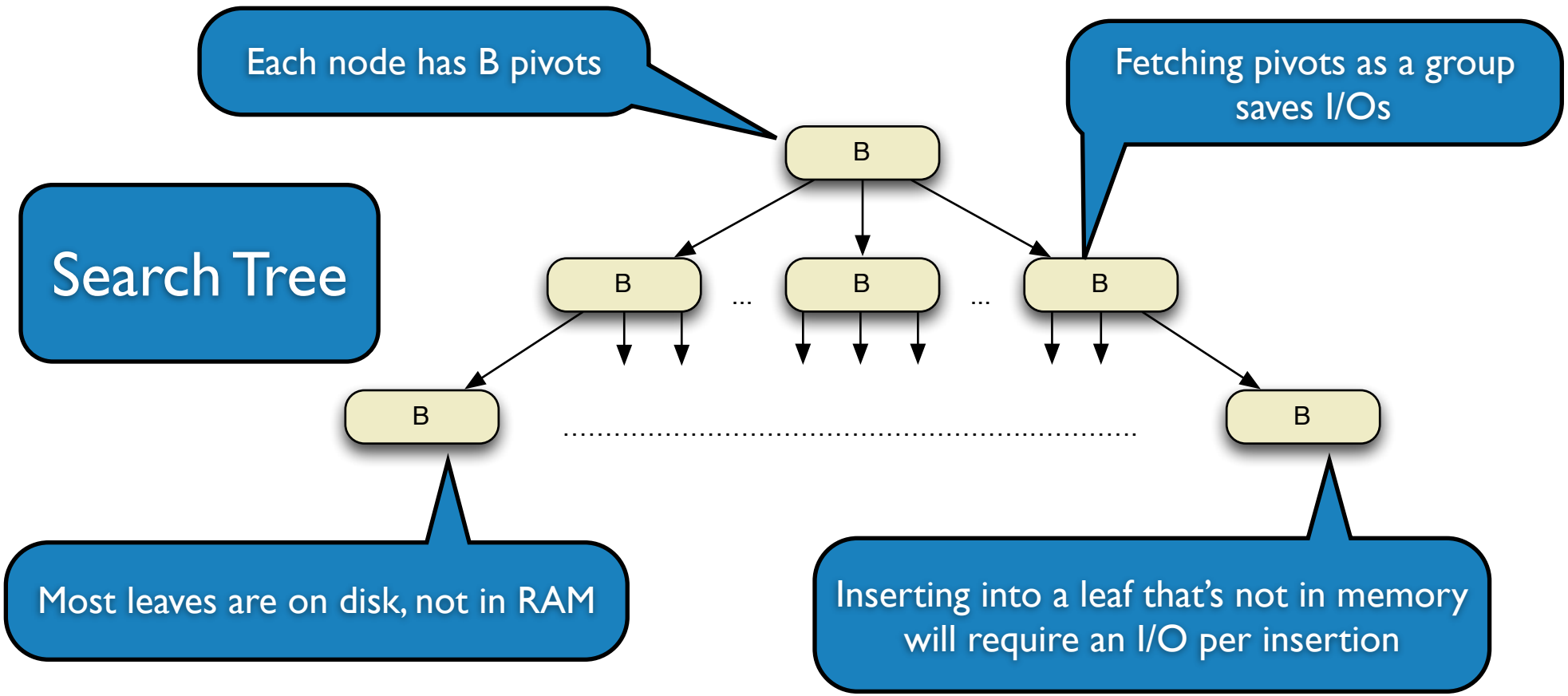


What's a B-tree?



Tokutek®

What's a B-tree?



Tokutek®

B-tree Delivery Service

If fast memory is like walking across a room

- Each update in a B-tree is a walking trip from
  - New York

B-tree Delivery Service

If fast memory is like walking across a room

- Each update in a B-tree is a walking trip from
  - New York to St Louis

B-tree Delivery Service

If fast memory is like walking across a room

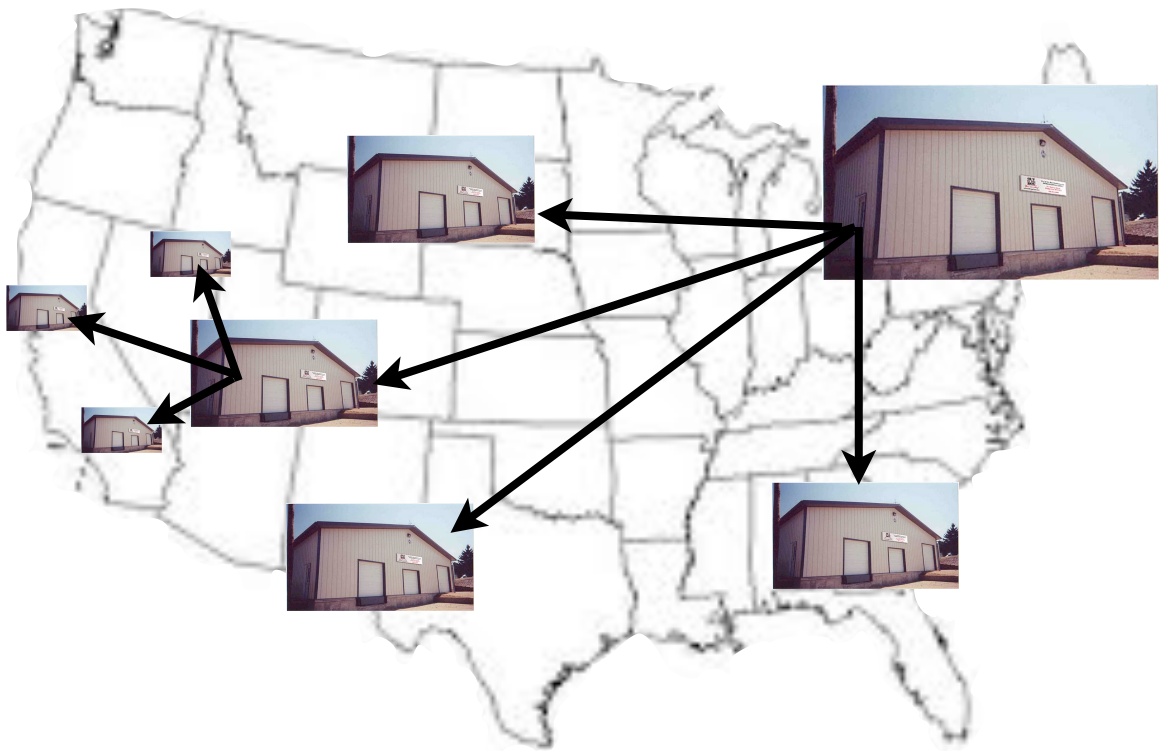
- Each update in a B-tree is a walking trip from
  - New York to St Louis

Tokutek®

Real-world delivery

Keep regional warehouses

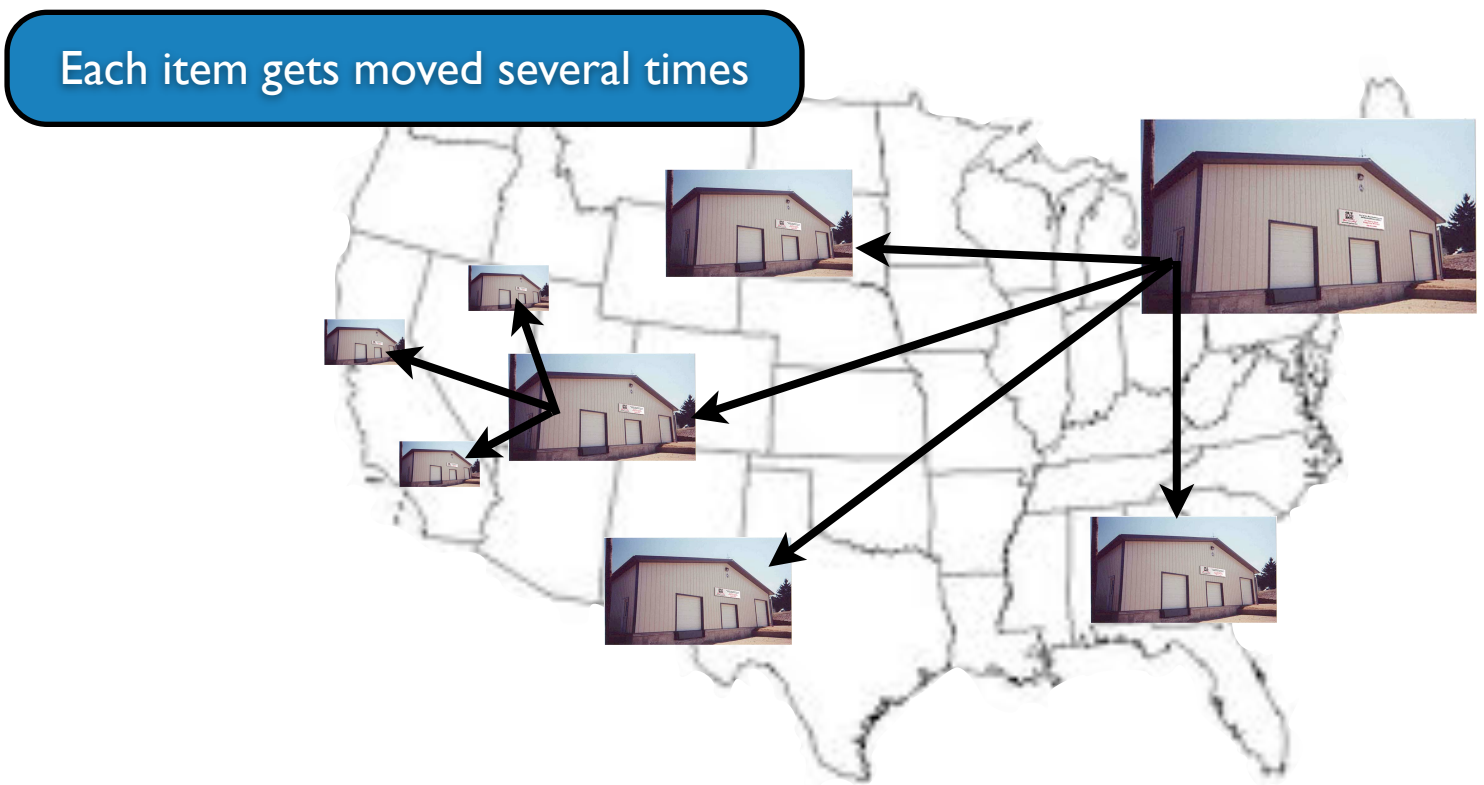
- Only move stuff when you can move a lot



Real-world delivery

Keep regional warehouses

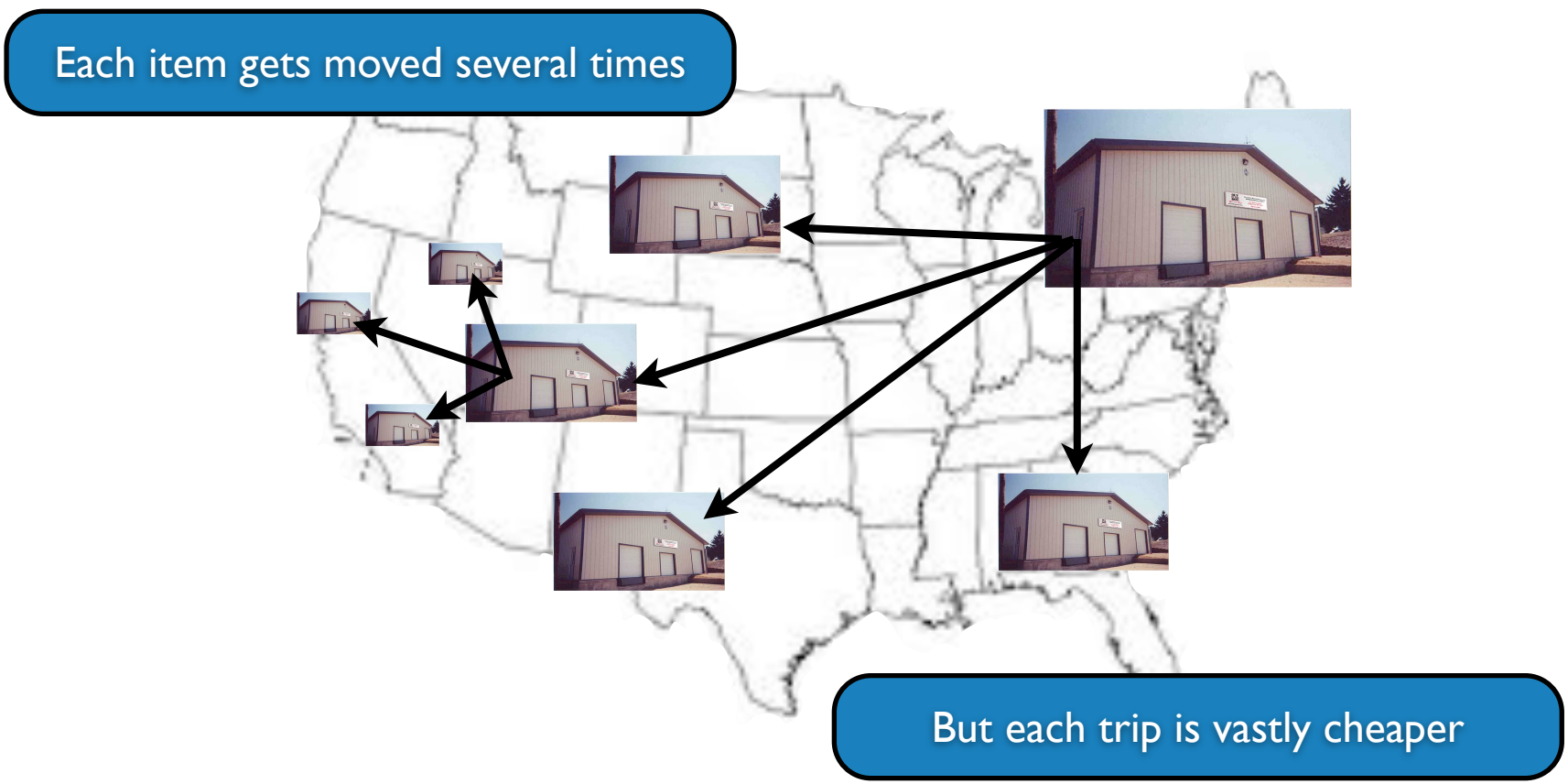
- Only move stuff when you can move a lot



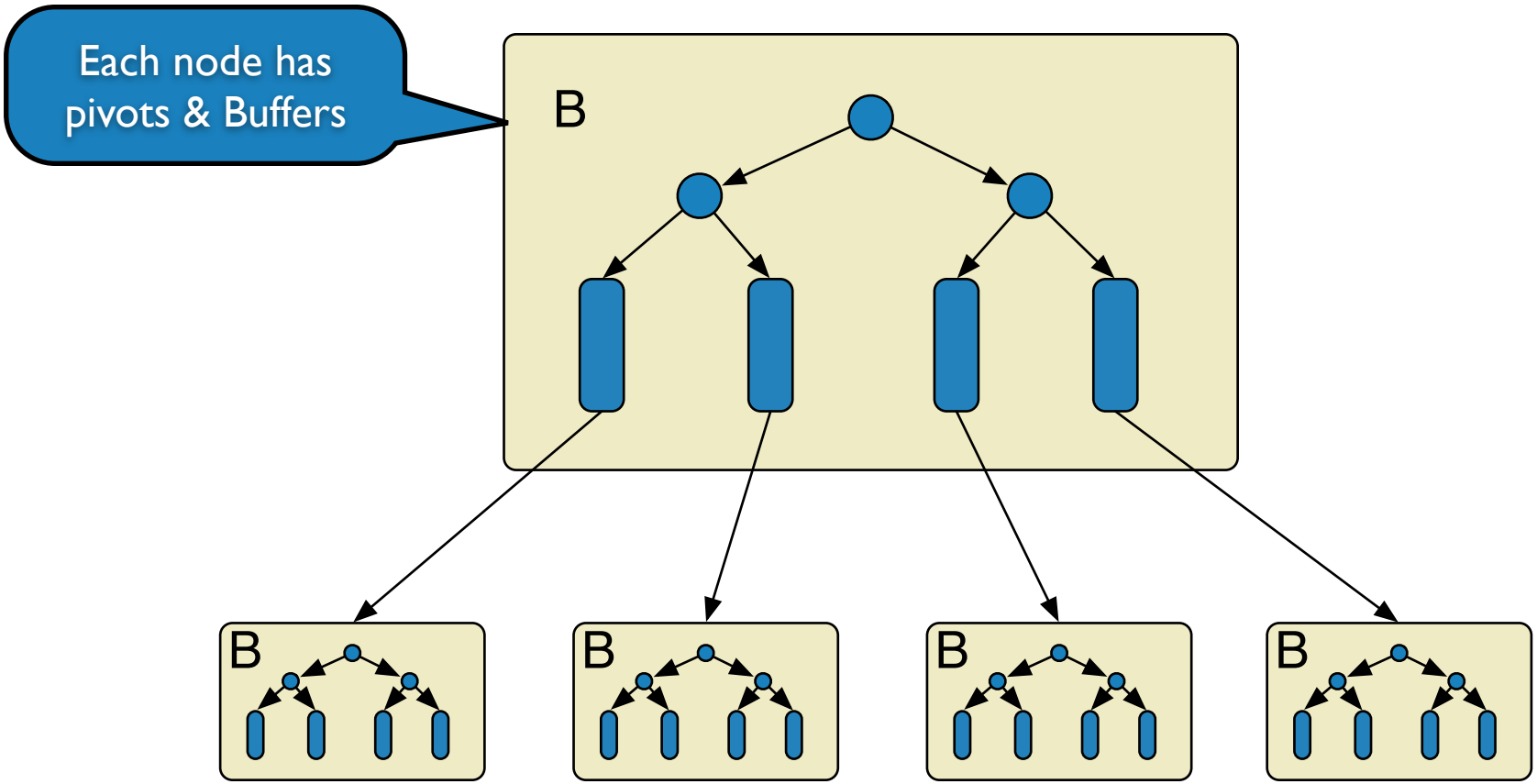
Real-world delivery

Keep regional warehouses

- Only move stuff when you can move a lot

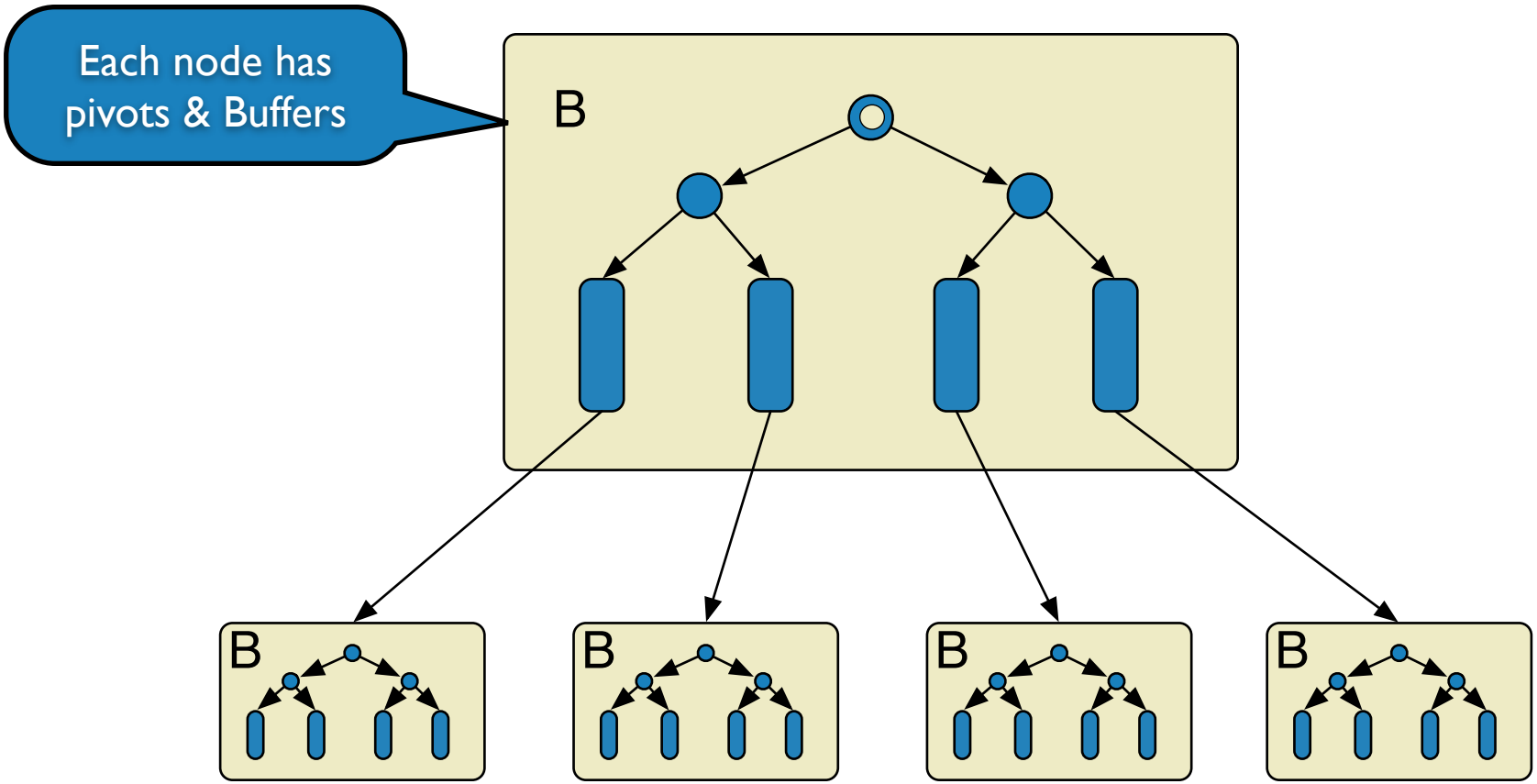


Fractal Tree Indexes



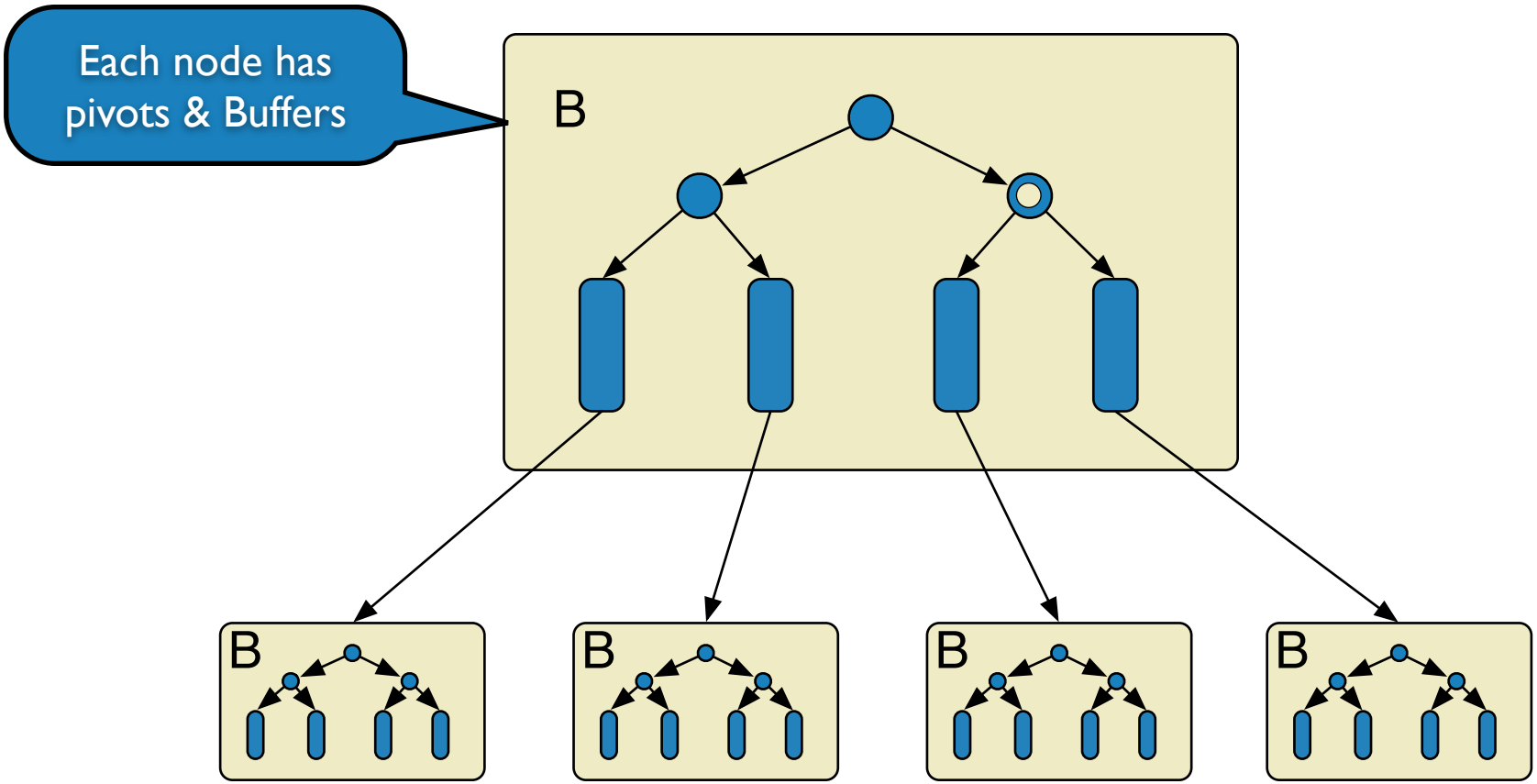
Tokutek®

Fractal Tree Indexes



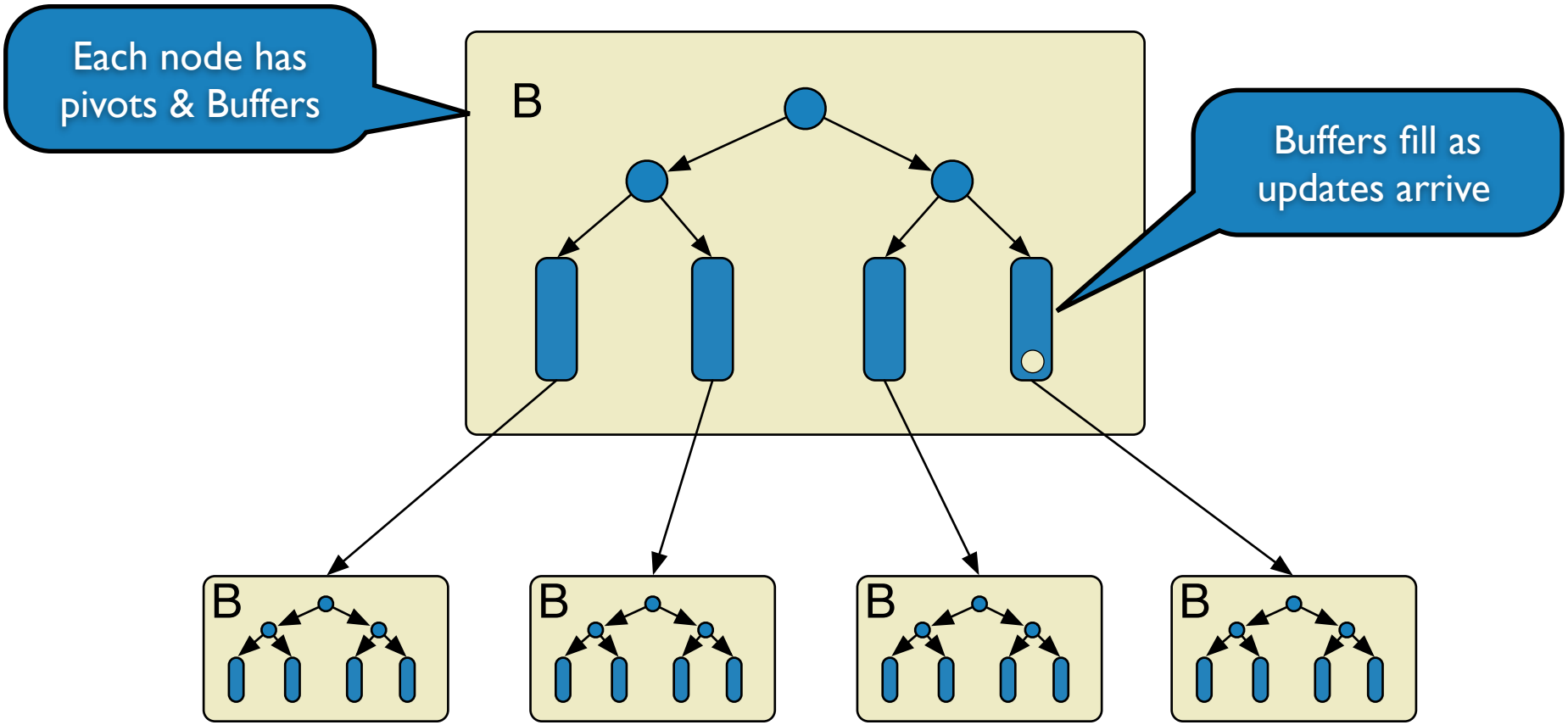
Tokutek®

Fractal Tree Indexes



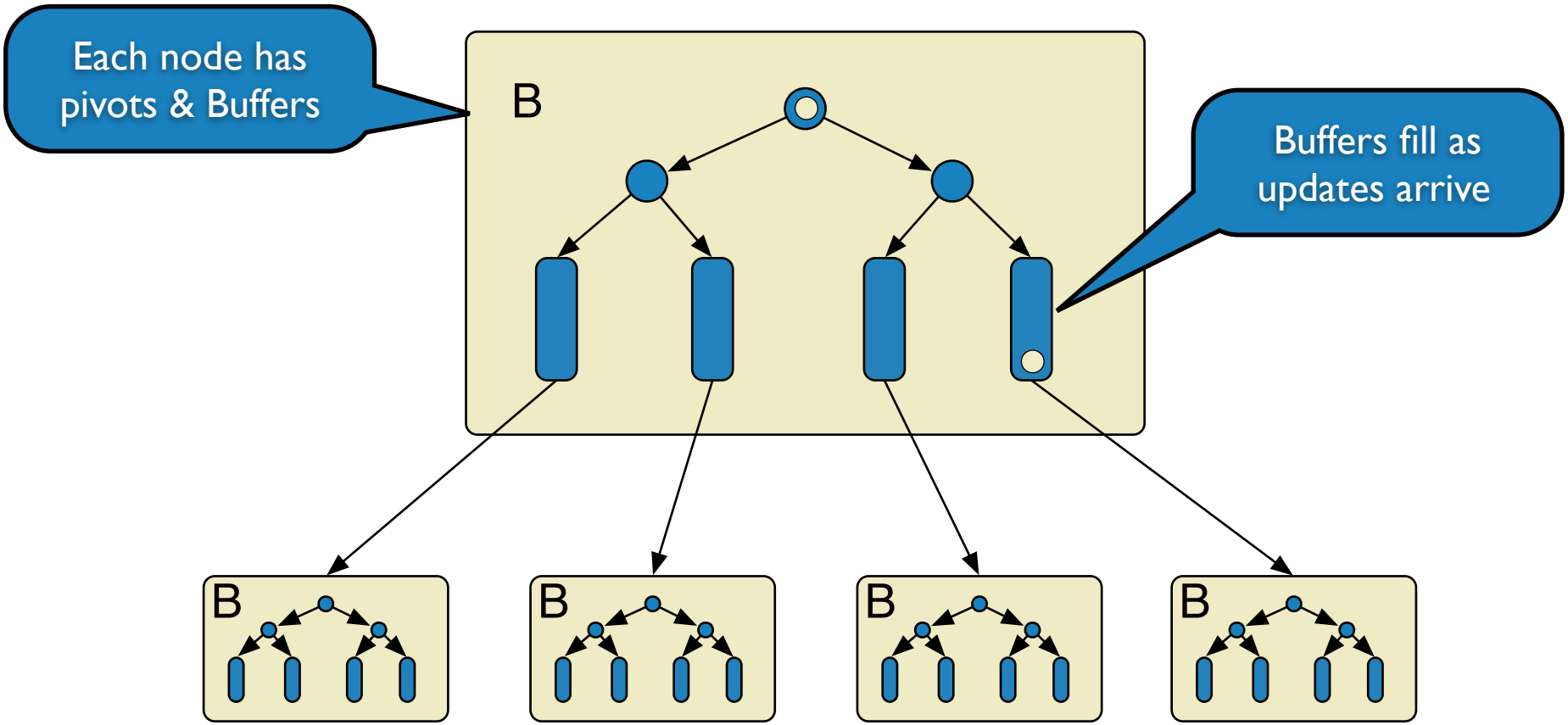
Tokutek®

Fractal Tree Indexes



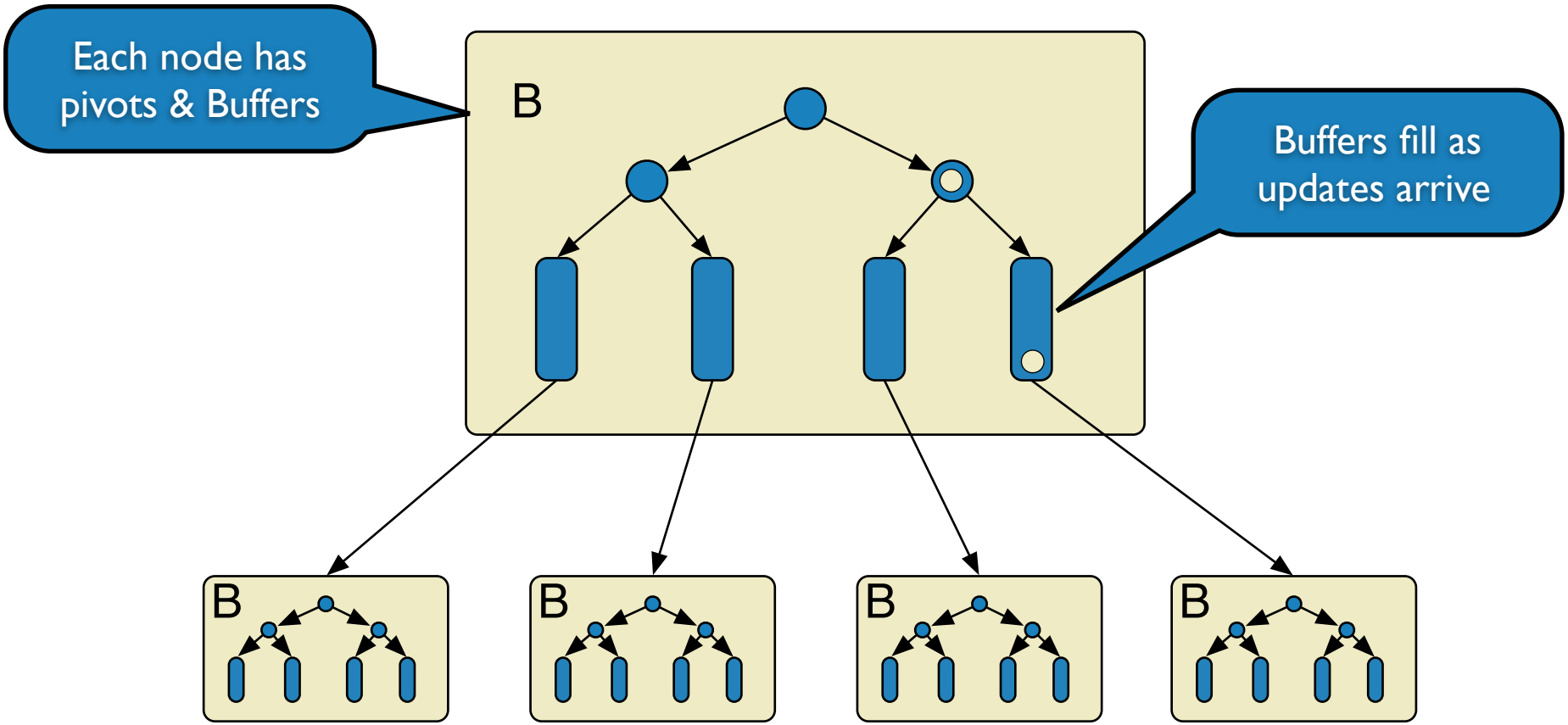
Tokutek®

Fractal Tree Indexes



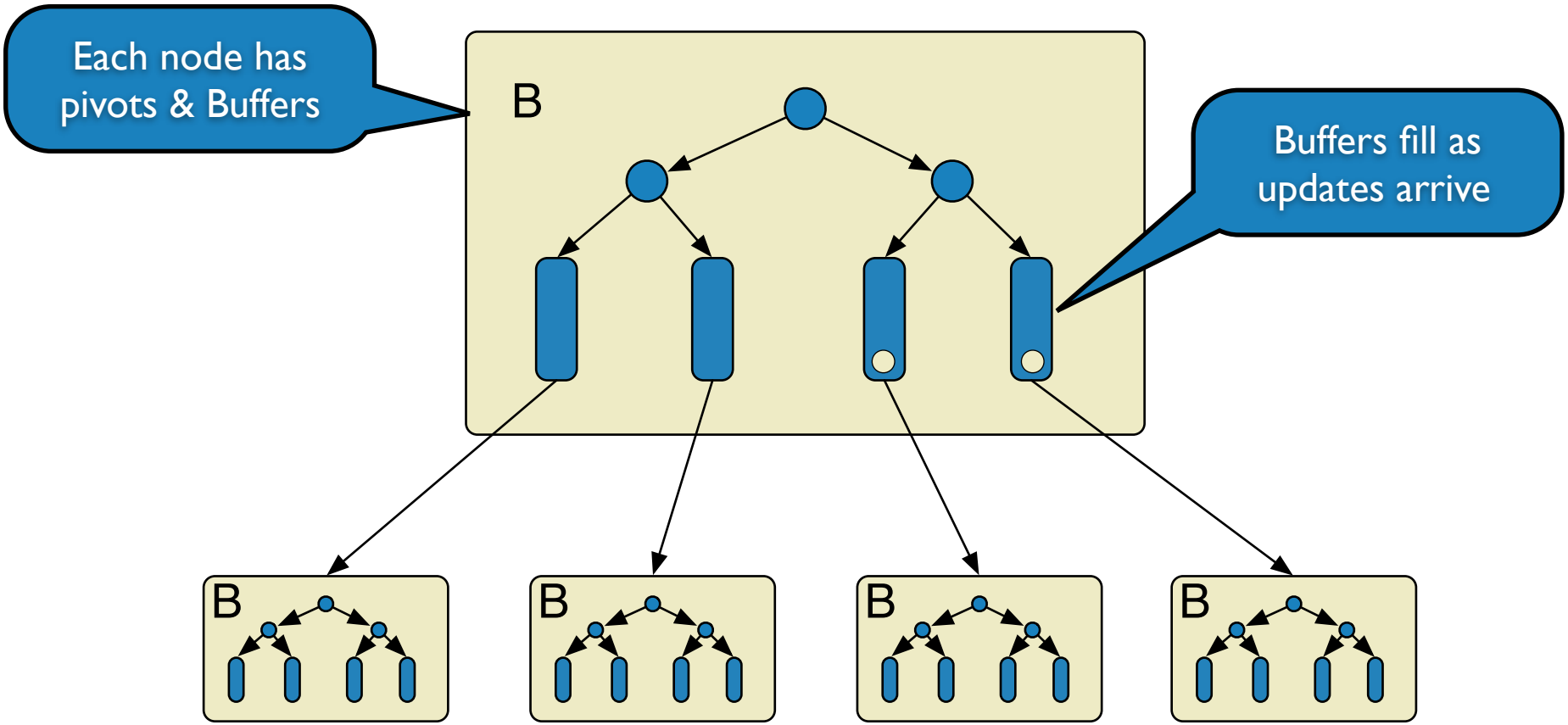
Tokutek®

Fractal Tree Indexes



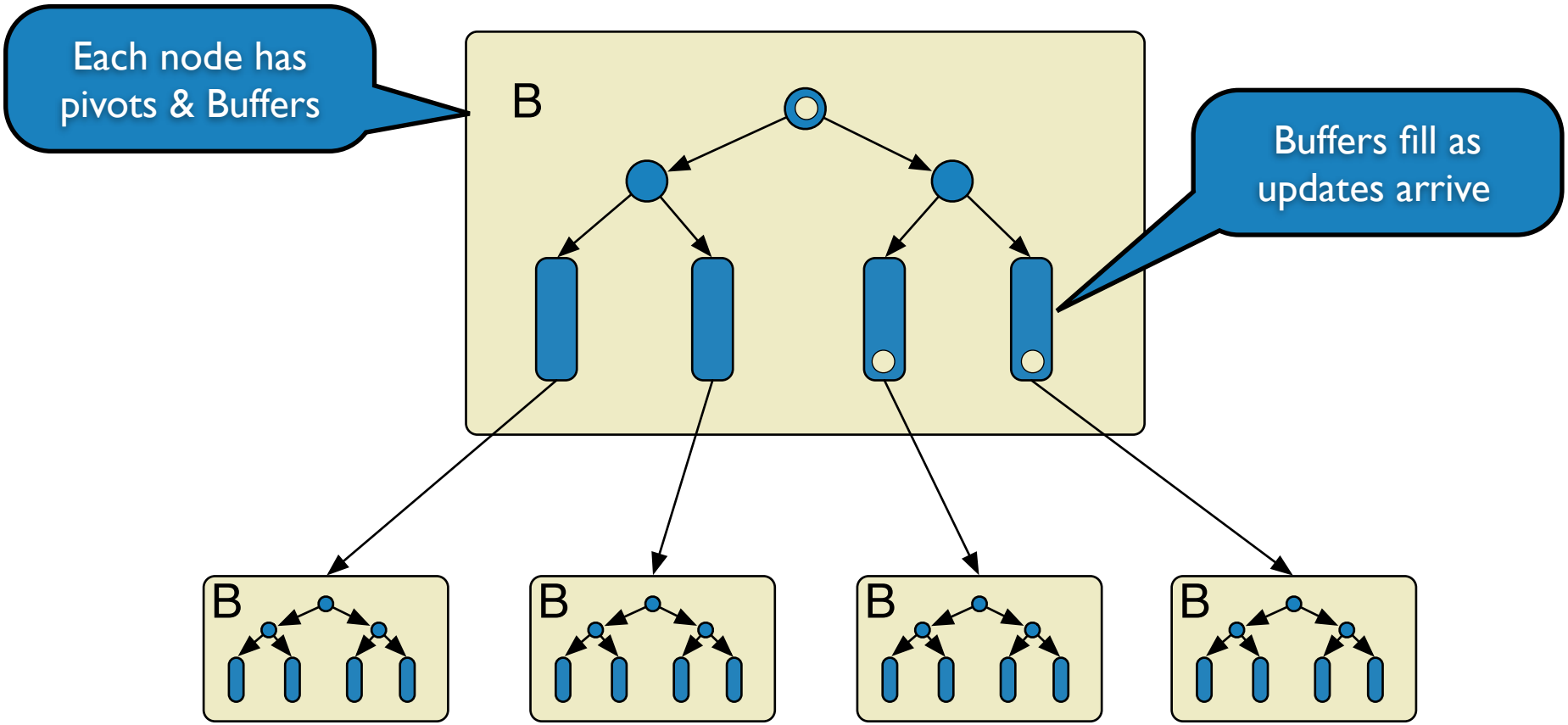
Tokutek®

Fractal Tree Indexes



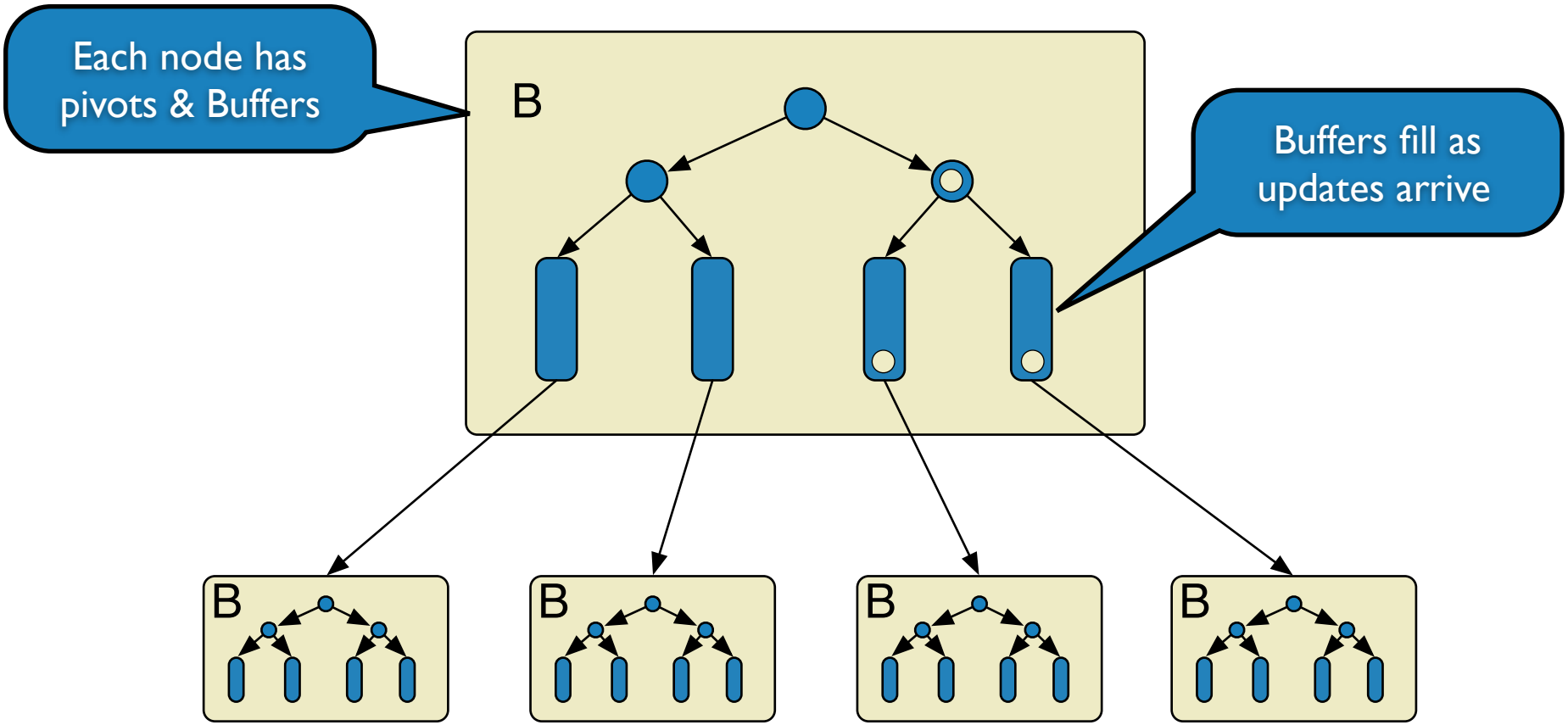
Tokutek®

Fractal Tree Indexes



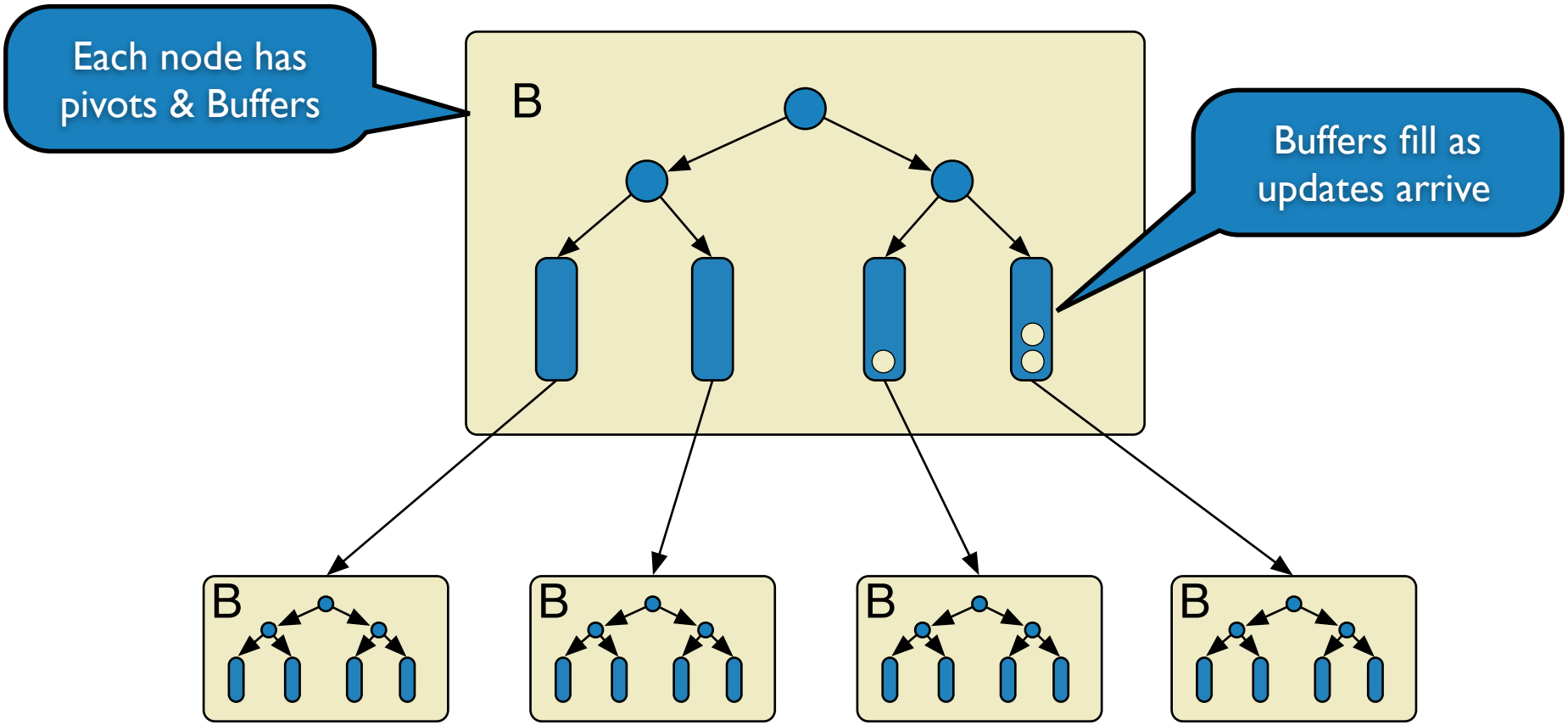
Tokutek®

Fractal Tree Indexes



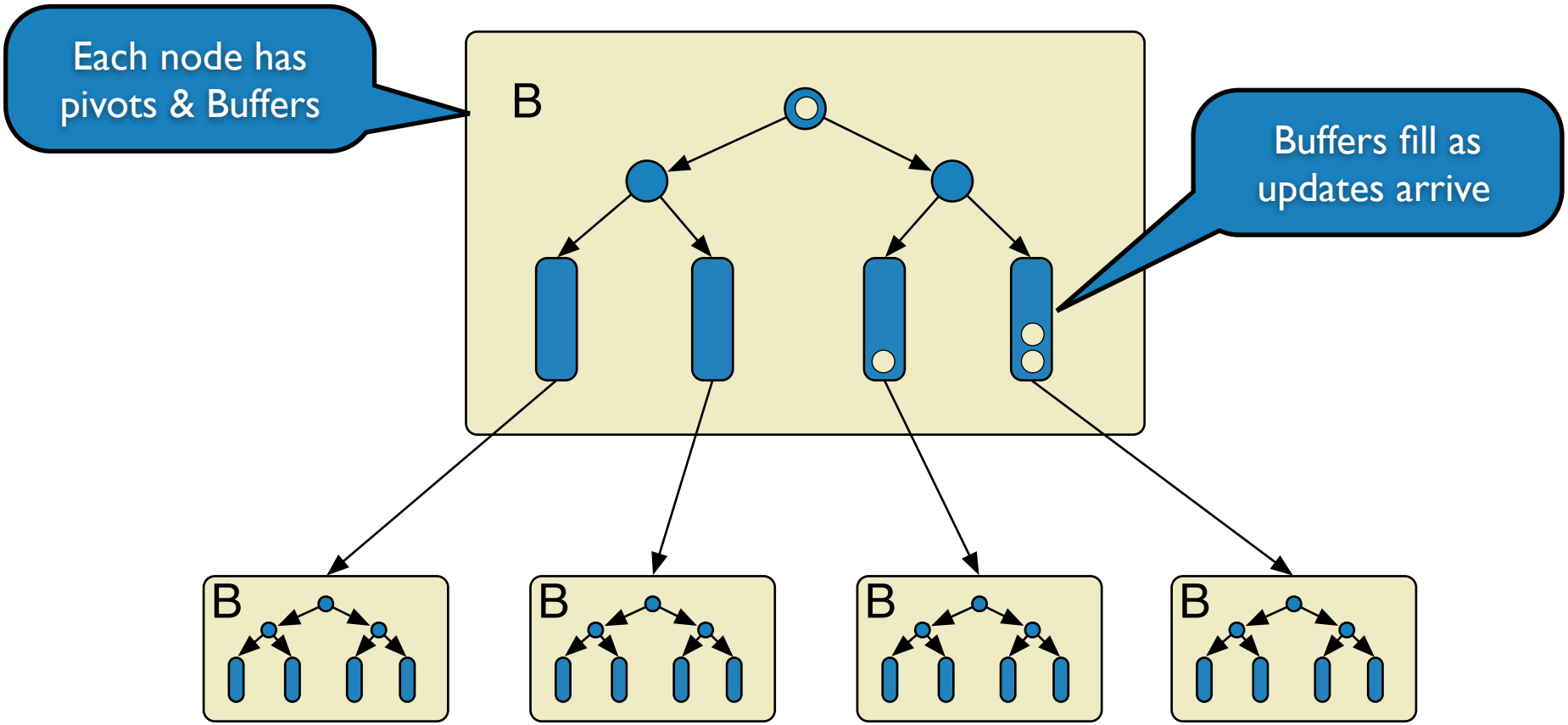
Tokutek®

Fractal Tree Indexes



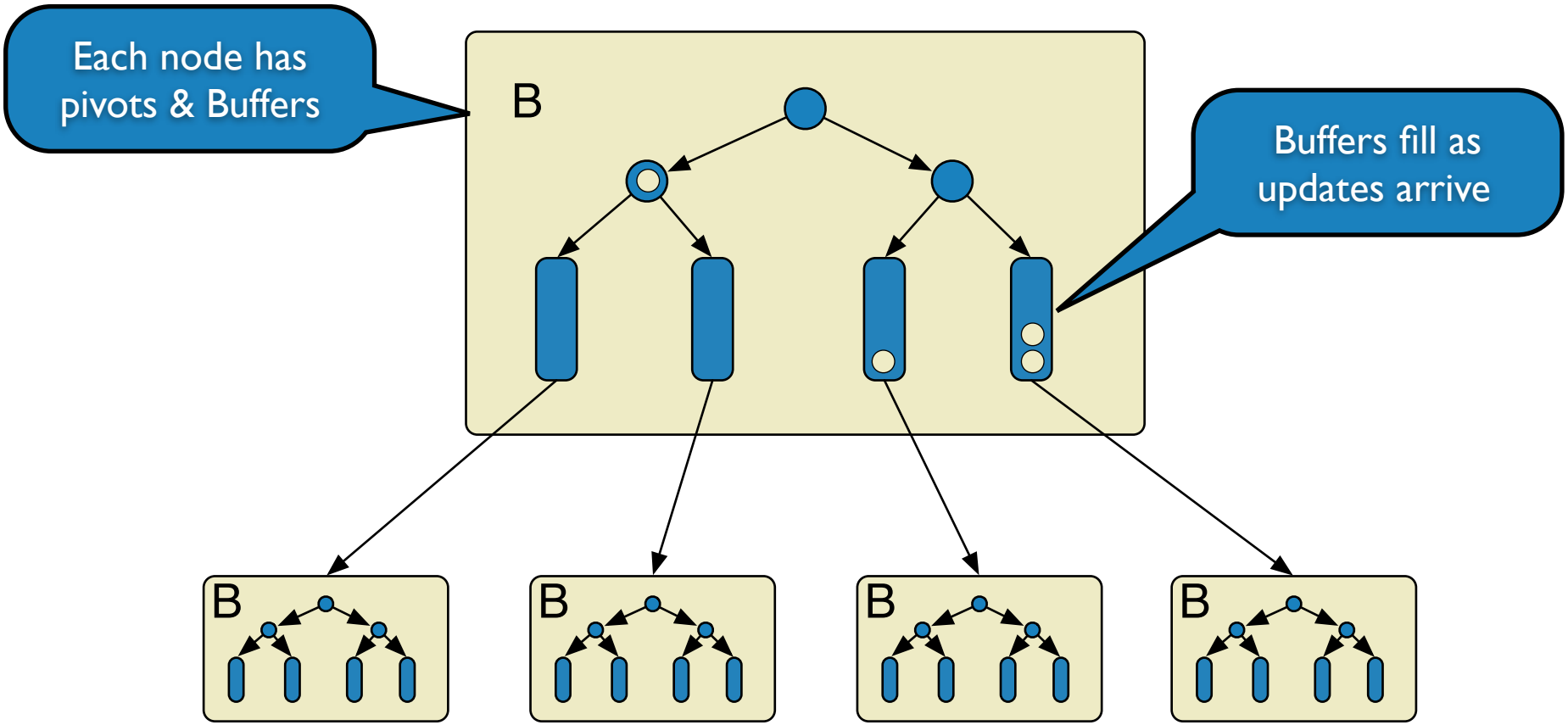
Tokutek®

Fractal Tree Indexes



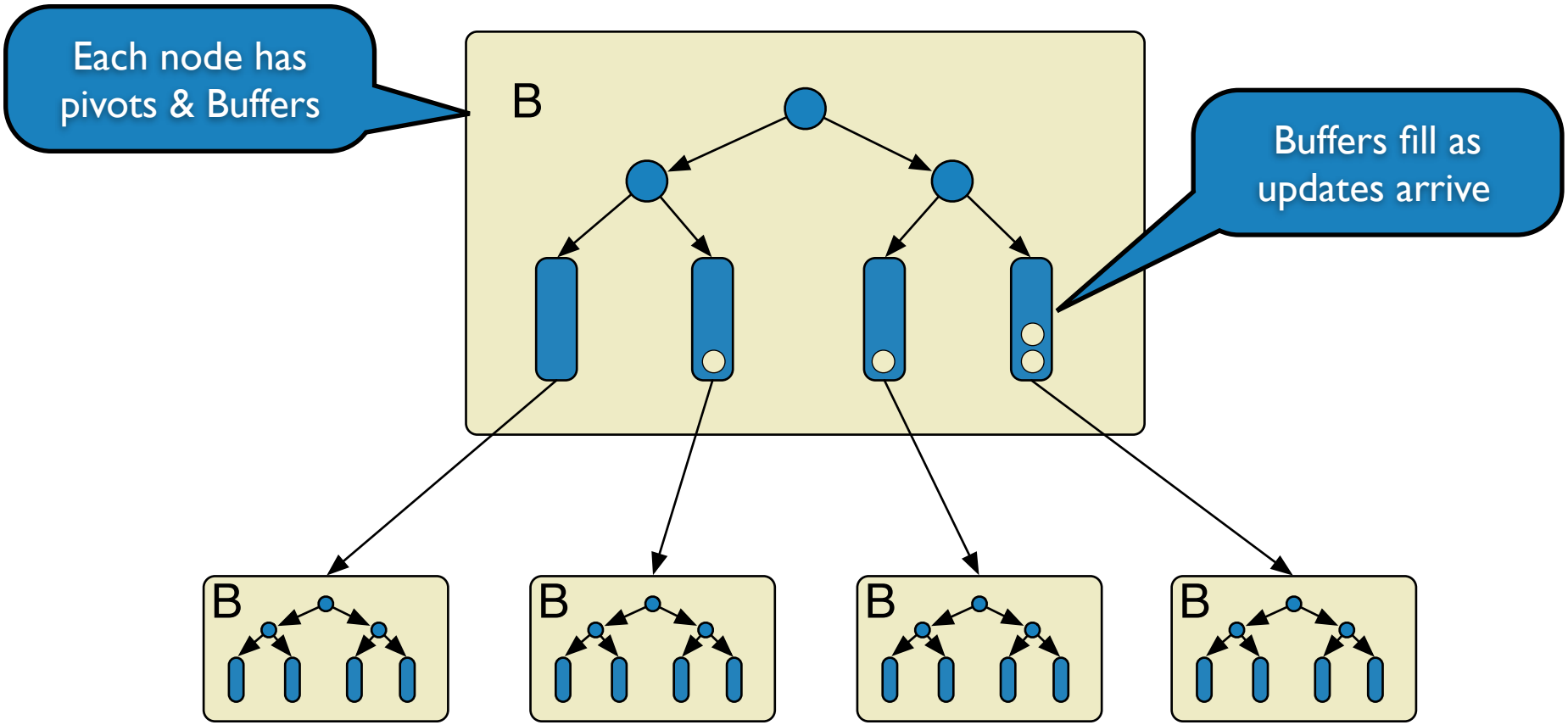
Tokutek®

Fractal Tree Indexes



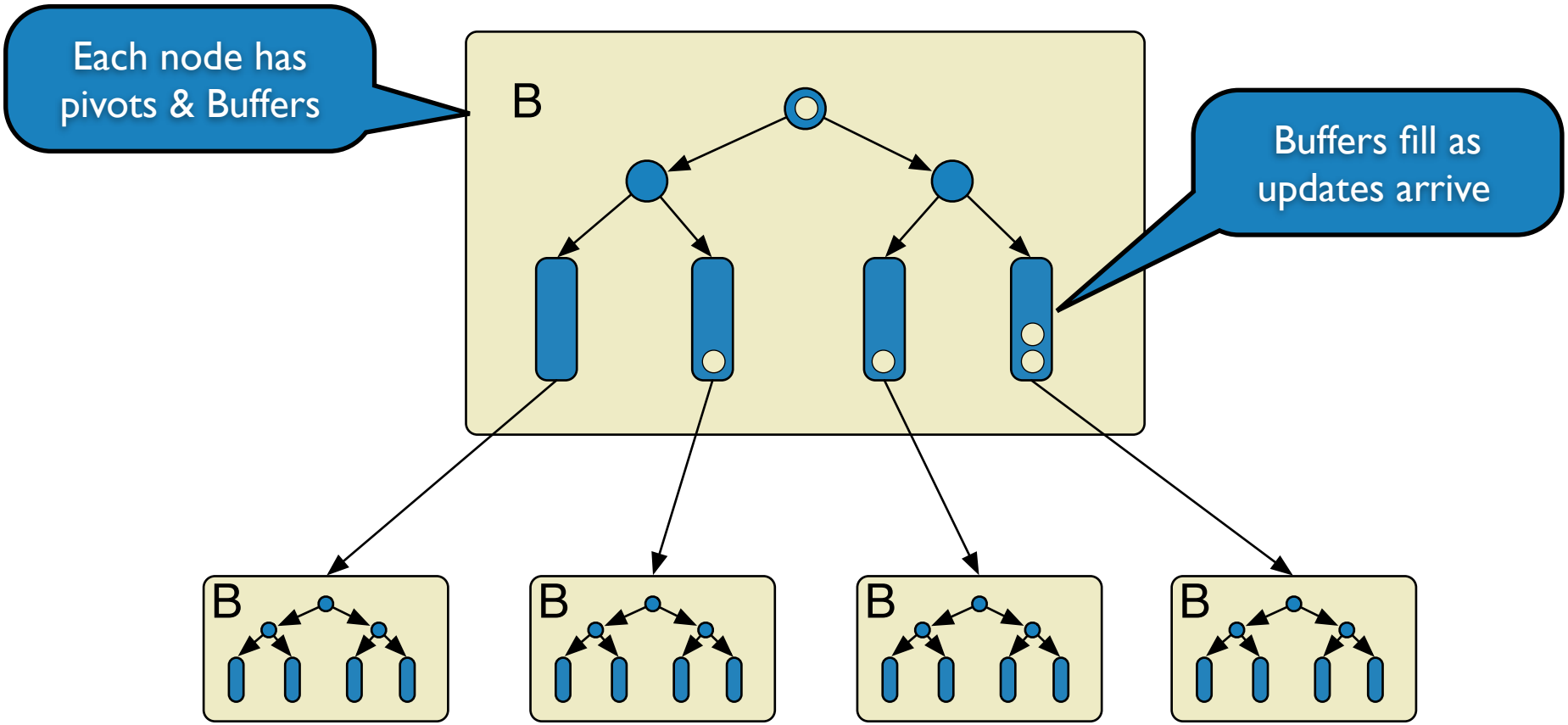
Tokutek®

Fractal Tree Indexes



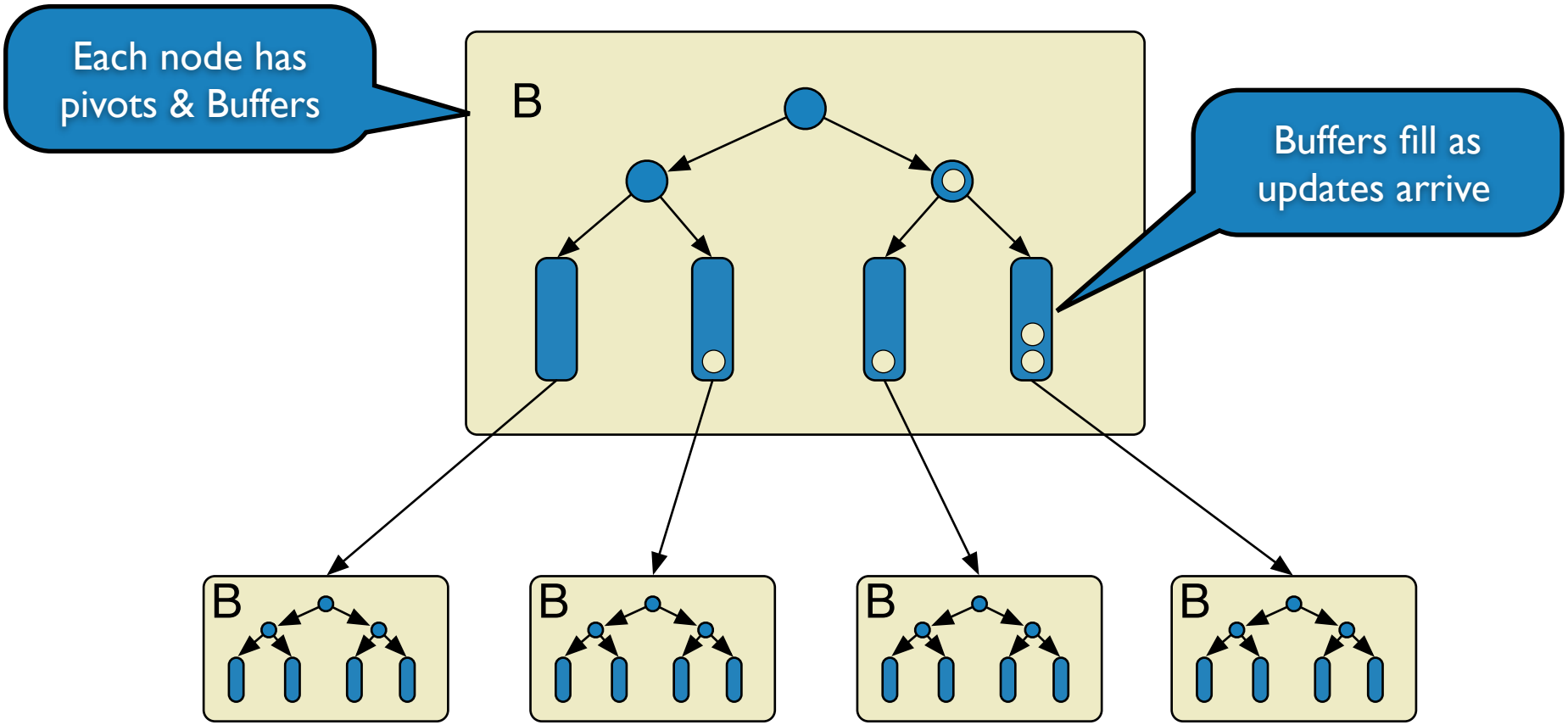
Tokutek®

Fractal Tree Indexes



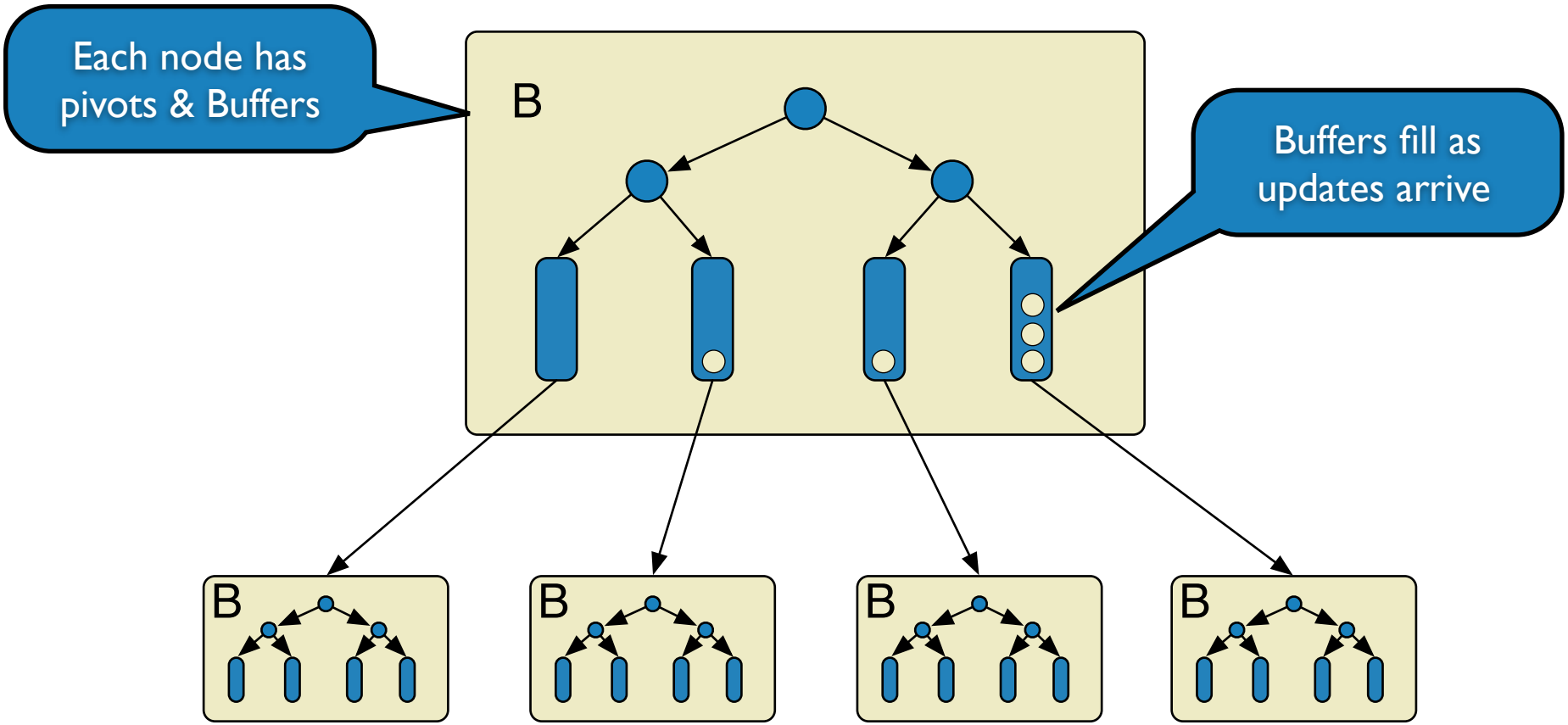
Tokutek®

Fractal Tree Indexes



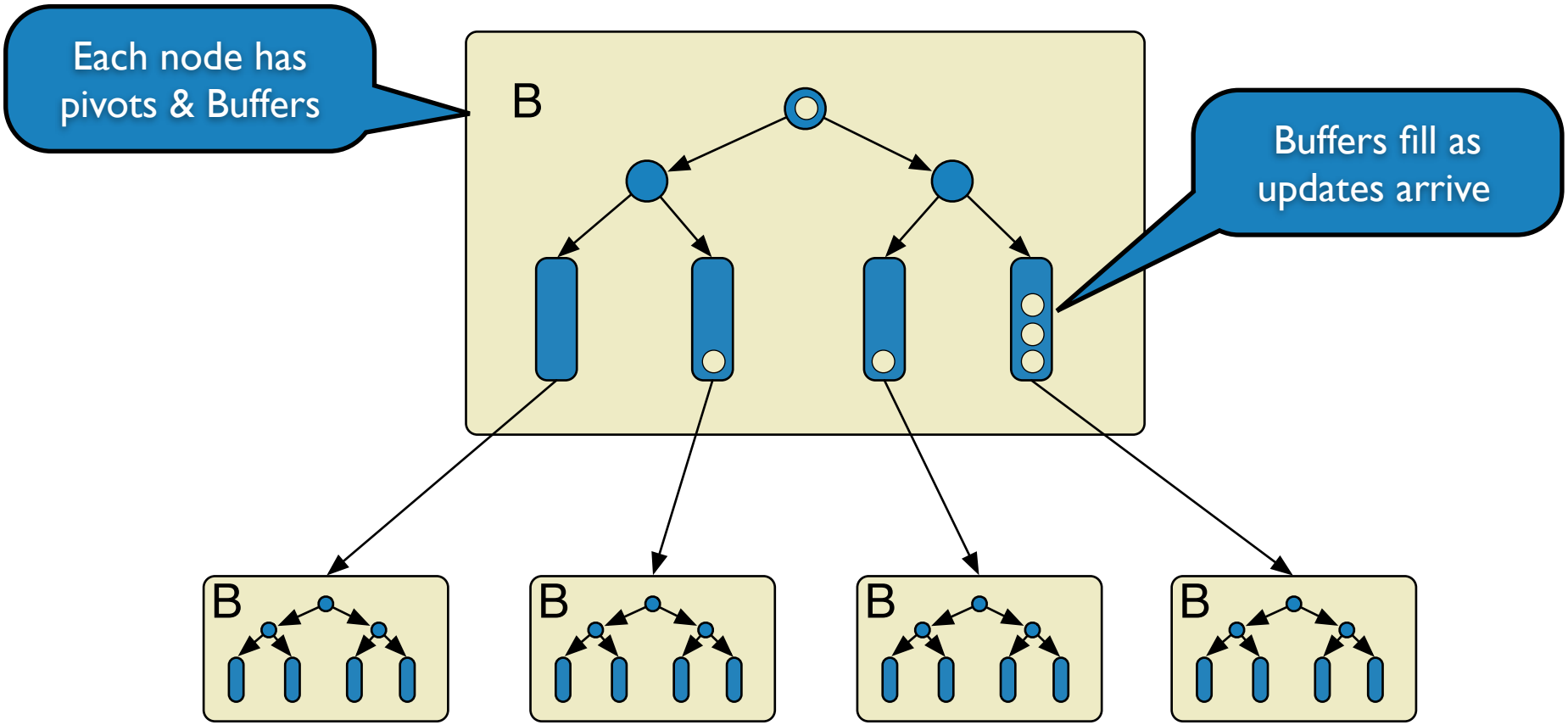
Tokutek®

Fractal Tree Indexes



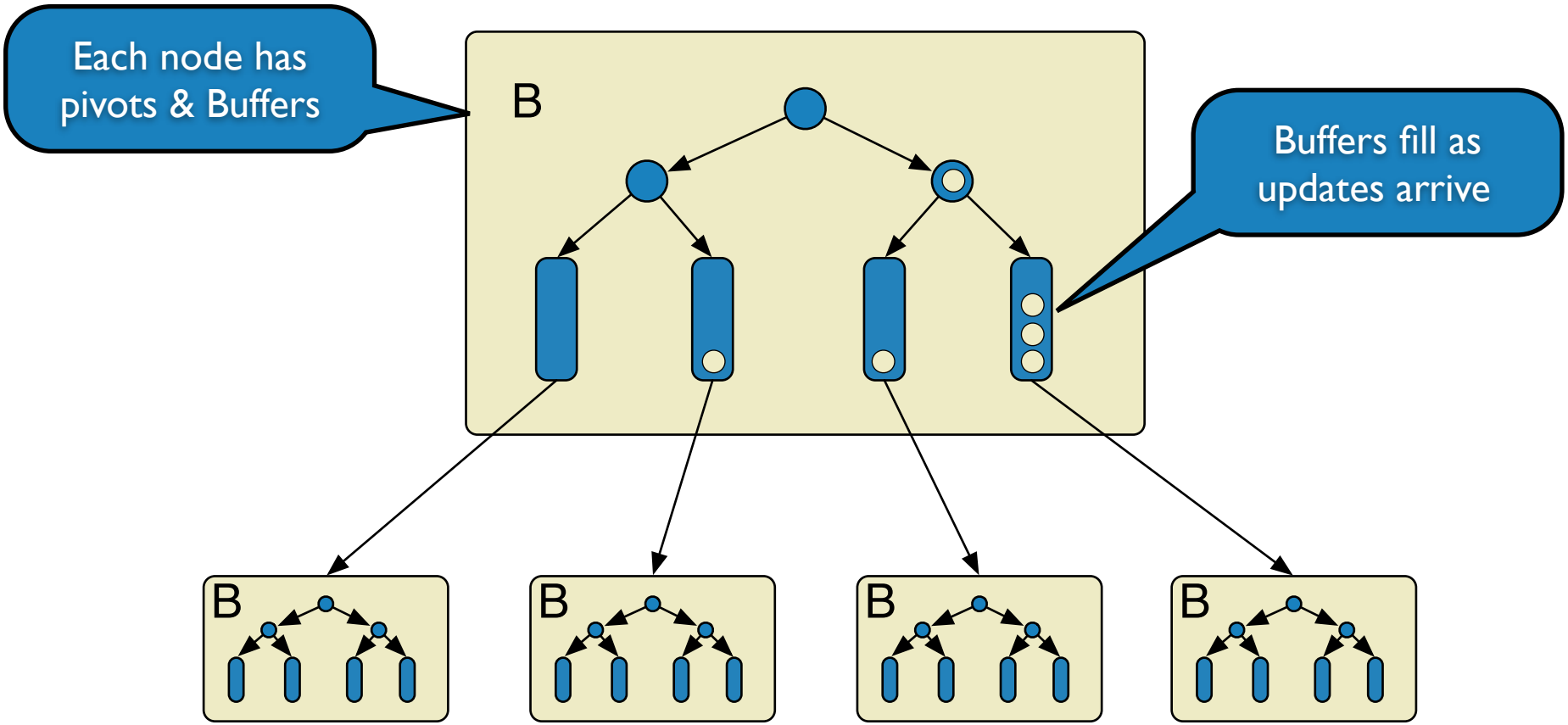
Tokutek®

Fractal Tree Indexes



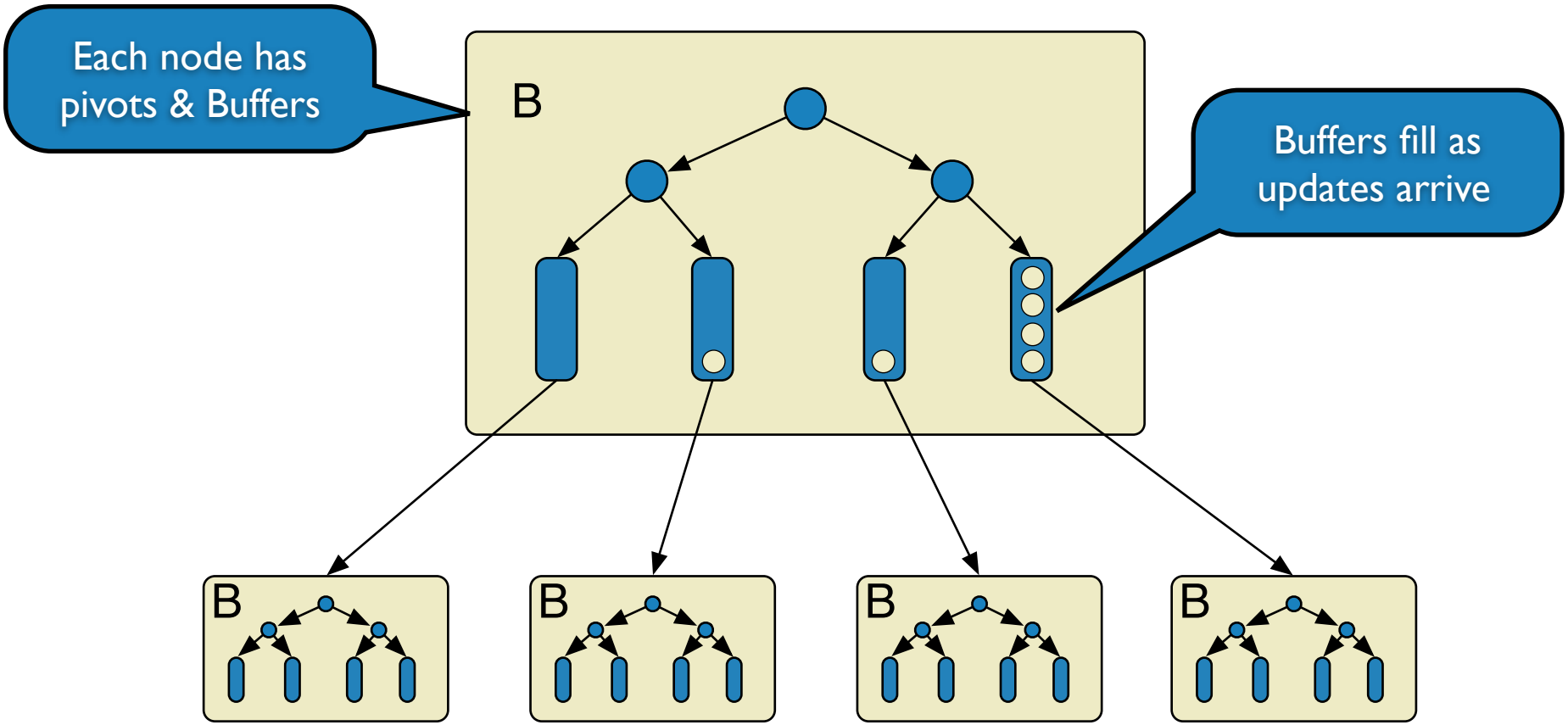
Tokutek®

Fractal Tree Indexes



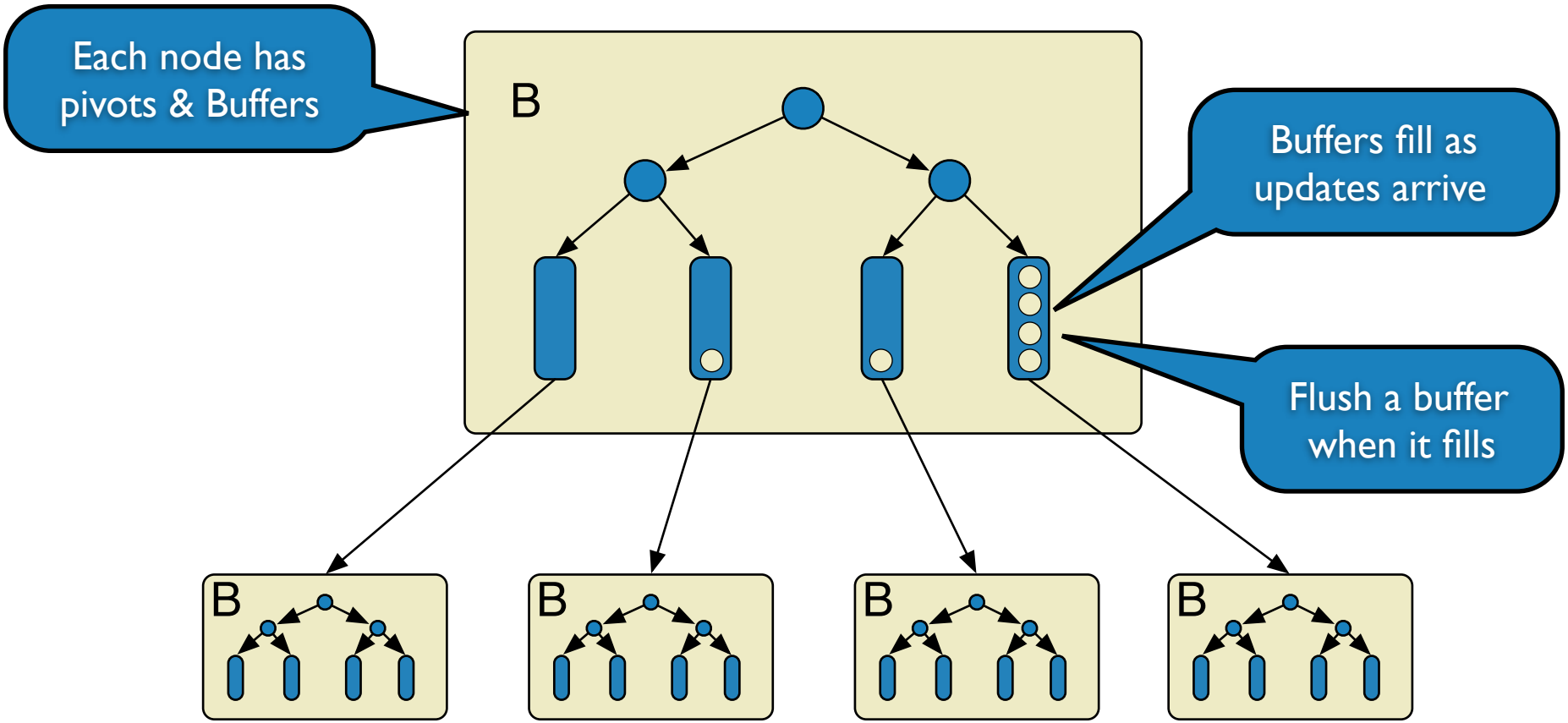
Tokutek®

Fractal Tree Indexes



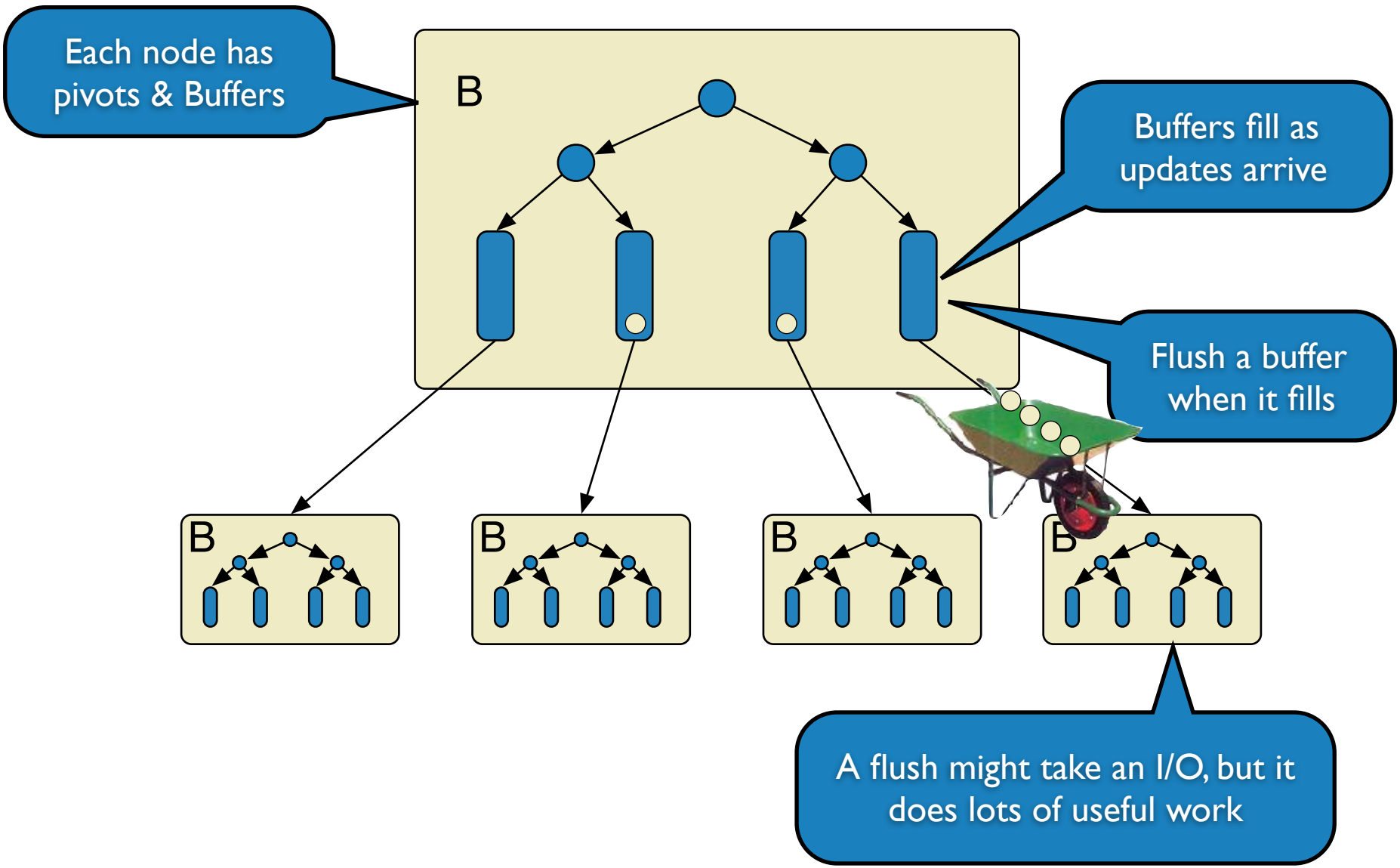
Tokutek®

Fractal Tree Indexes



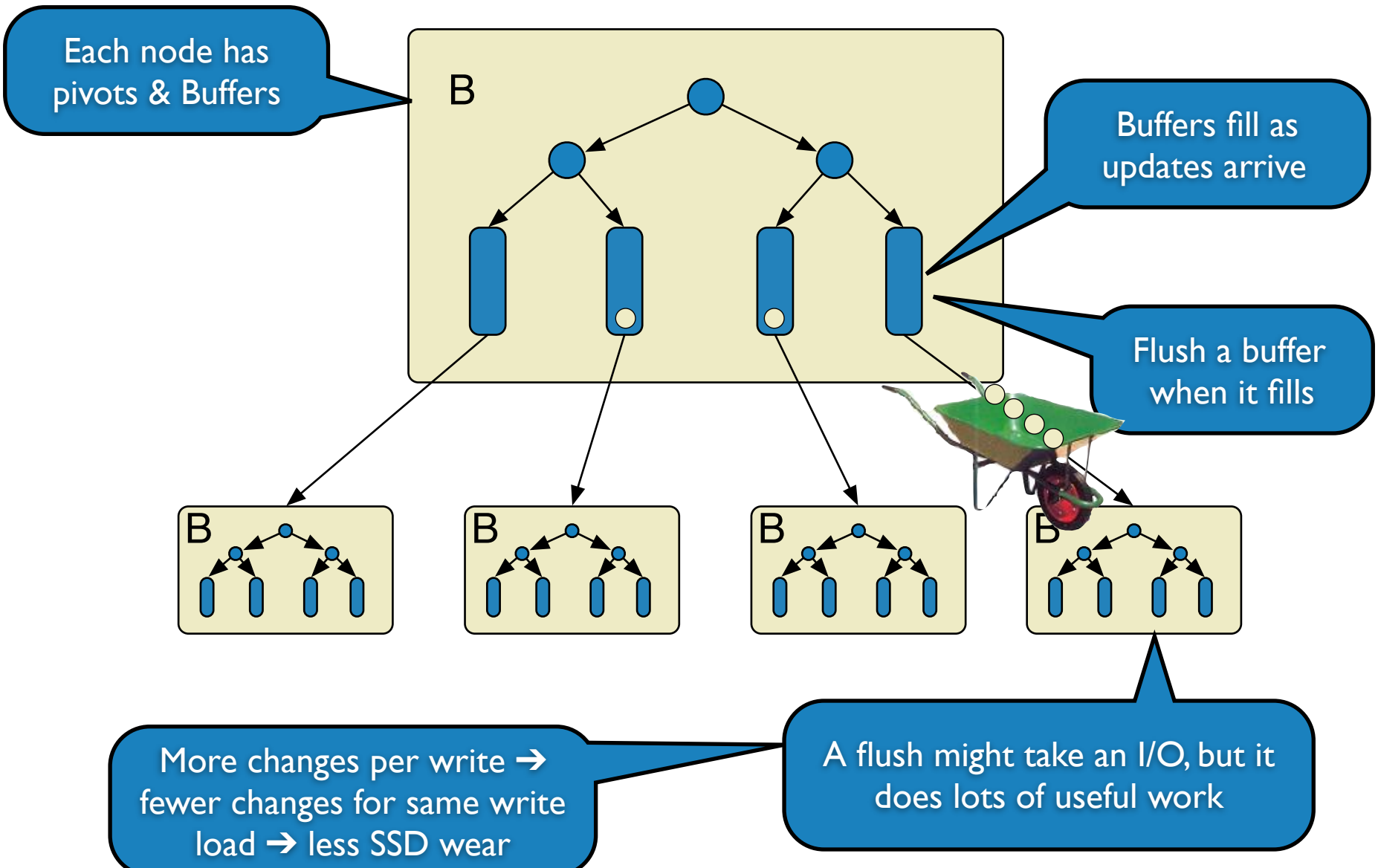
Tokutek®

Fractal Tree Indexes



Tokutek®

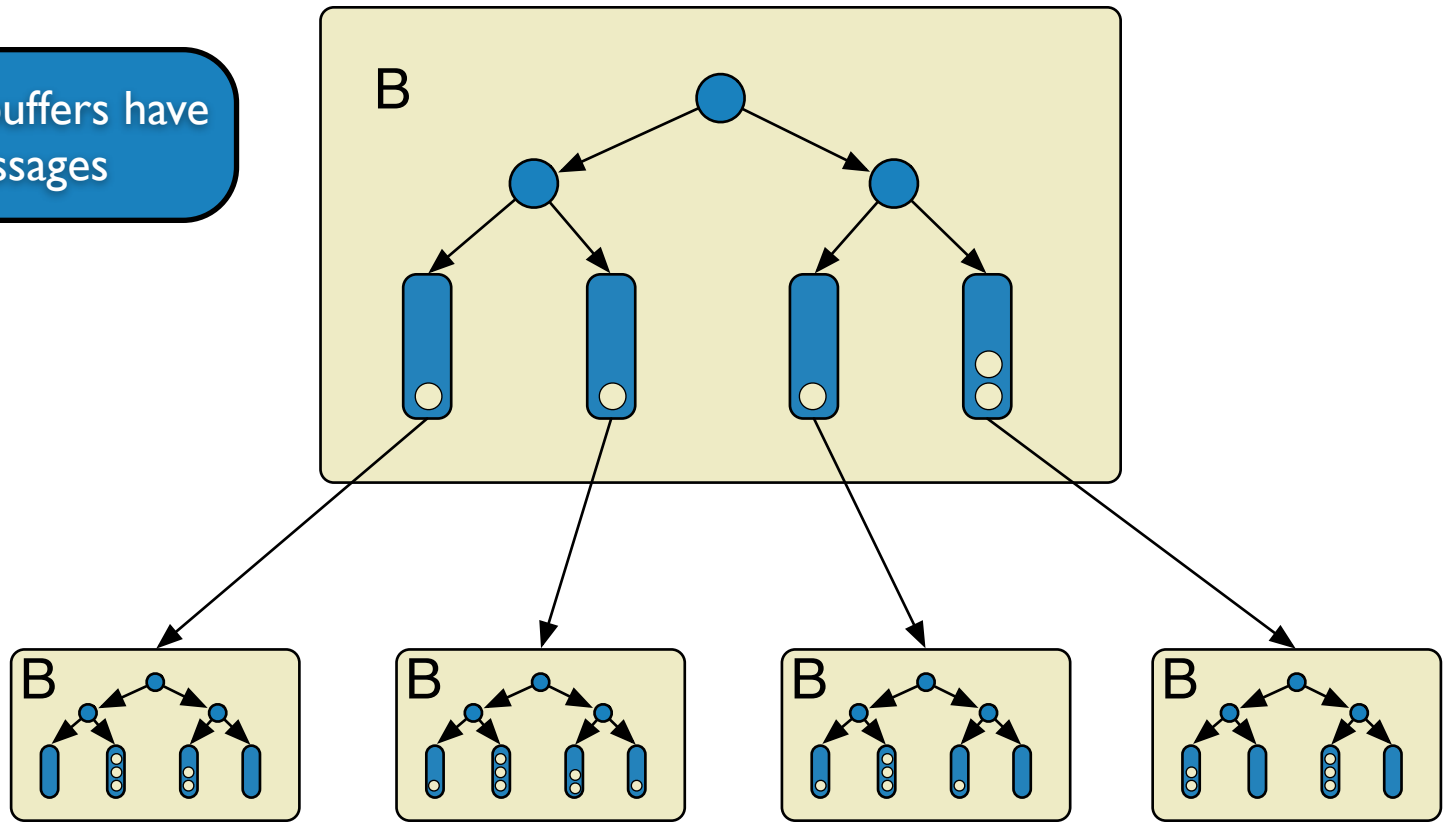
Fractal Tree Indexes



Tokutek®

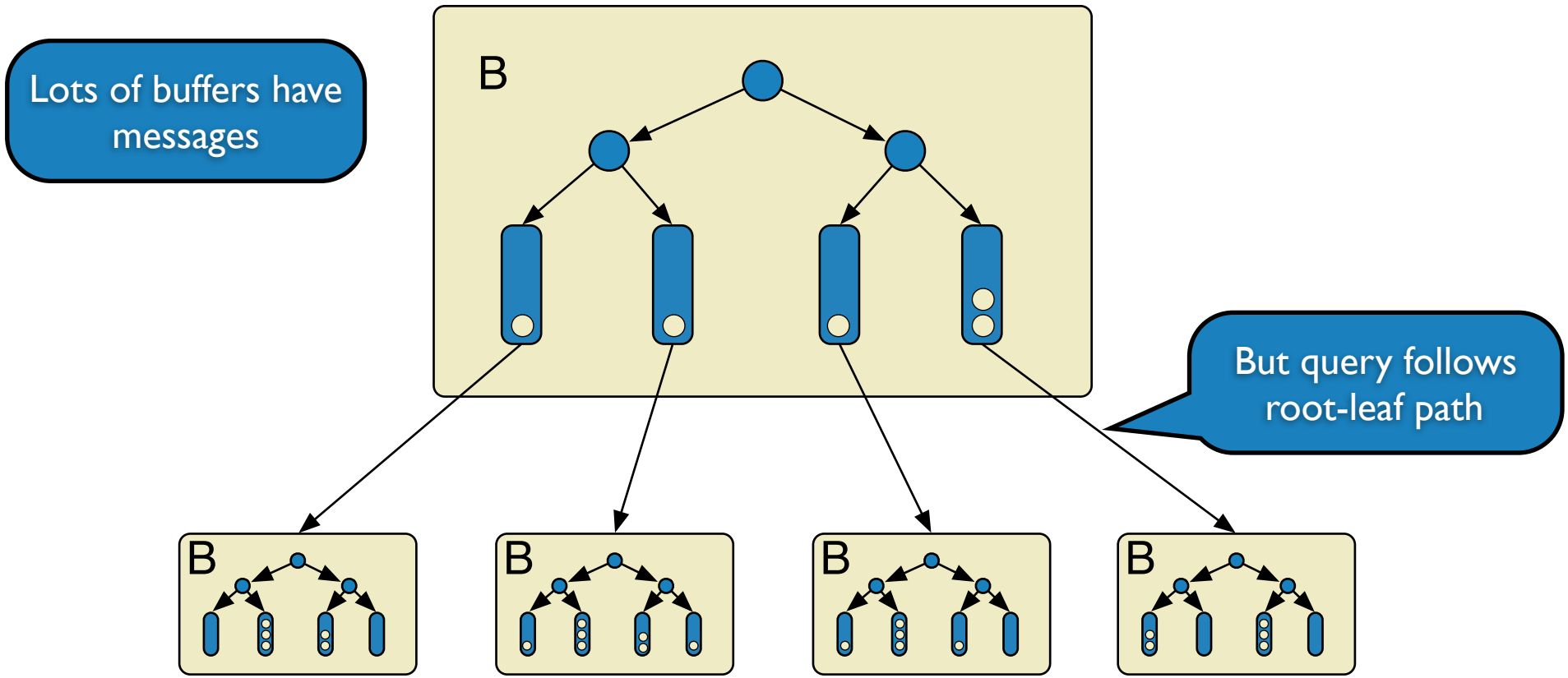
Fractal Tree Indexes: Queries

Lots of buffers have messages



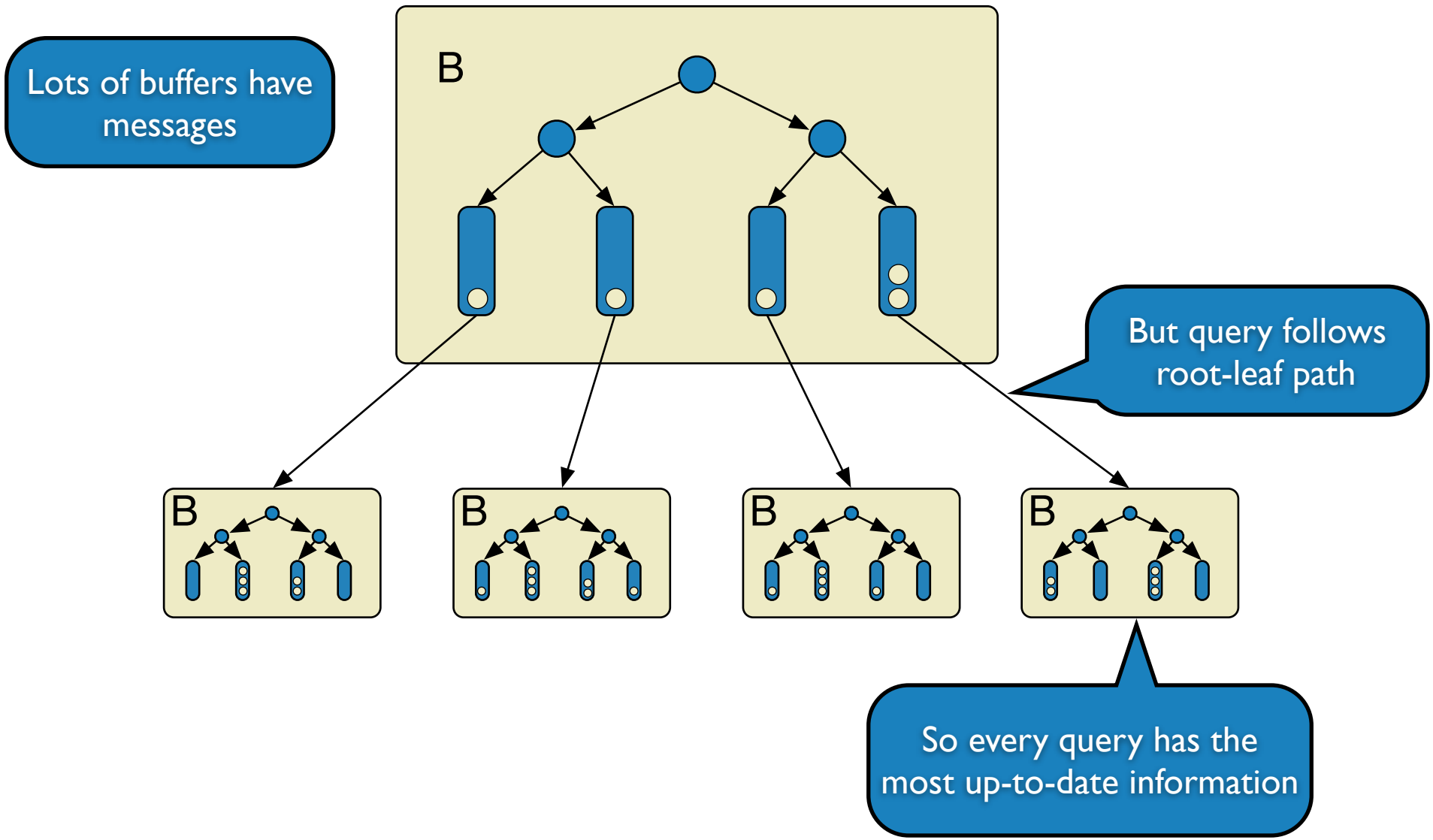
Tokutek®

Fractal Tree Indexes: Queries



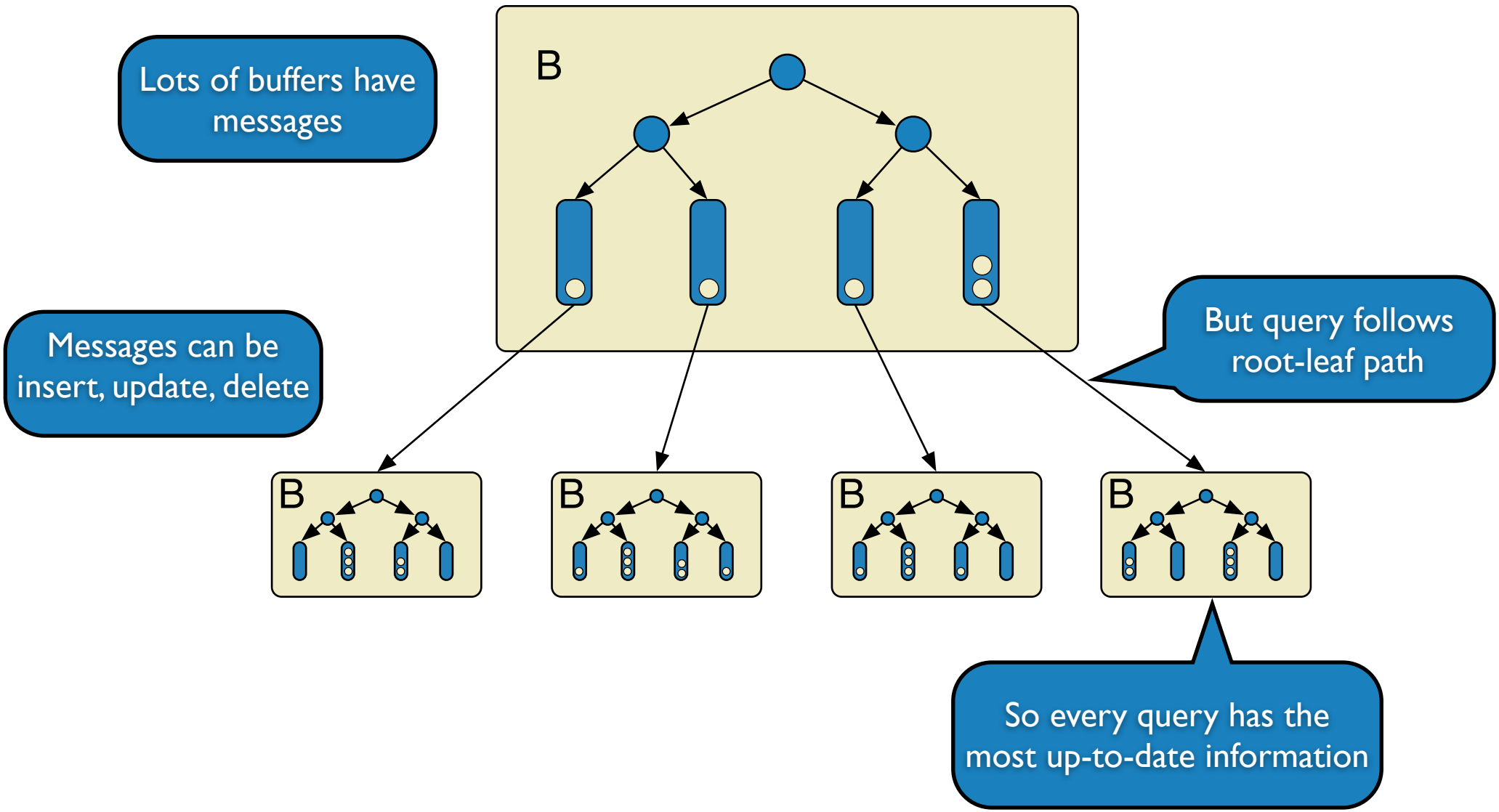
Tokutek®

Fractal Tree Indexes: Queries



Tokutek®

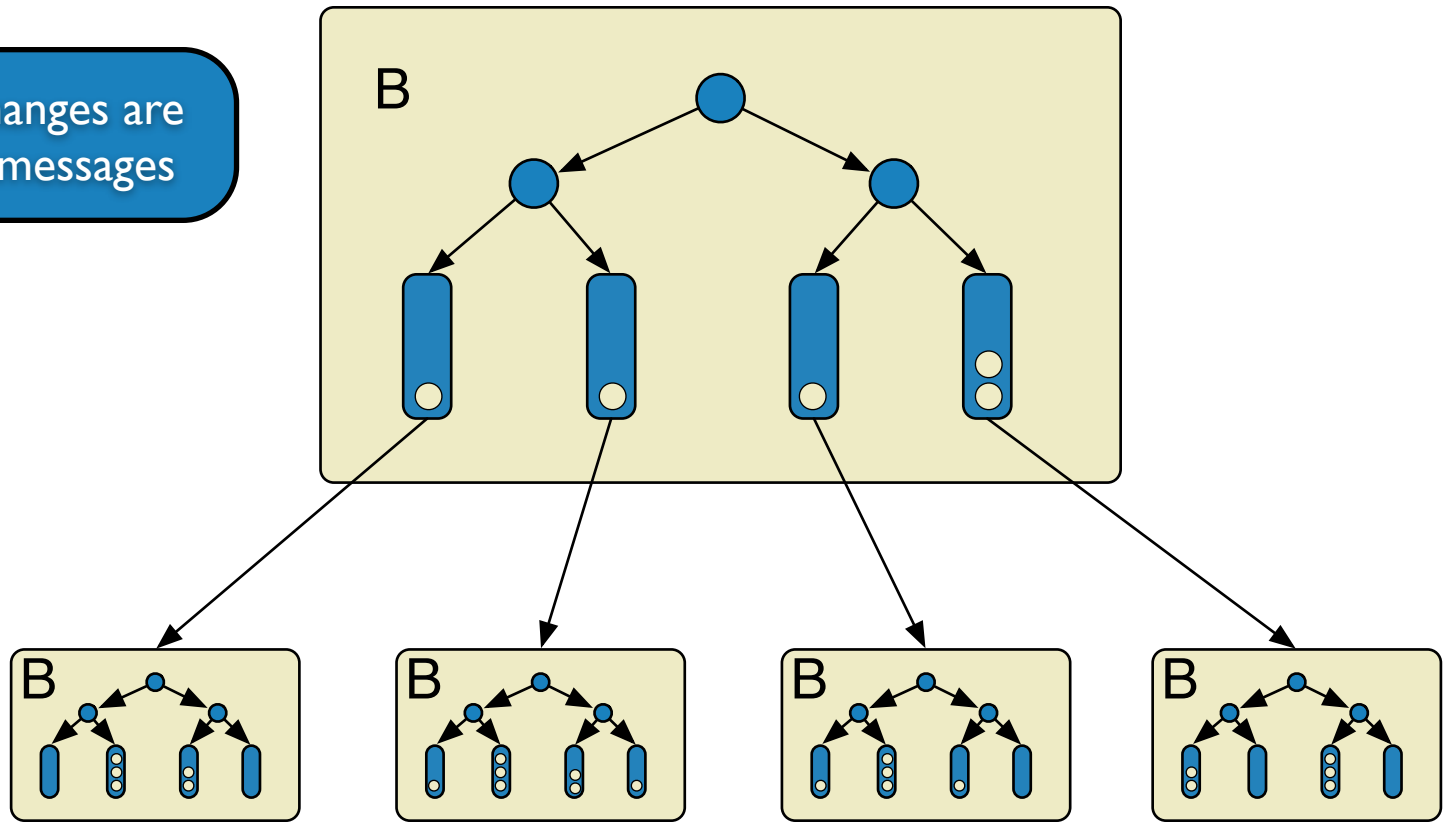
Fractal Tree Indexes: Queries



Tokutek®

Fractal Tree Indexes: Schema Changes

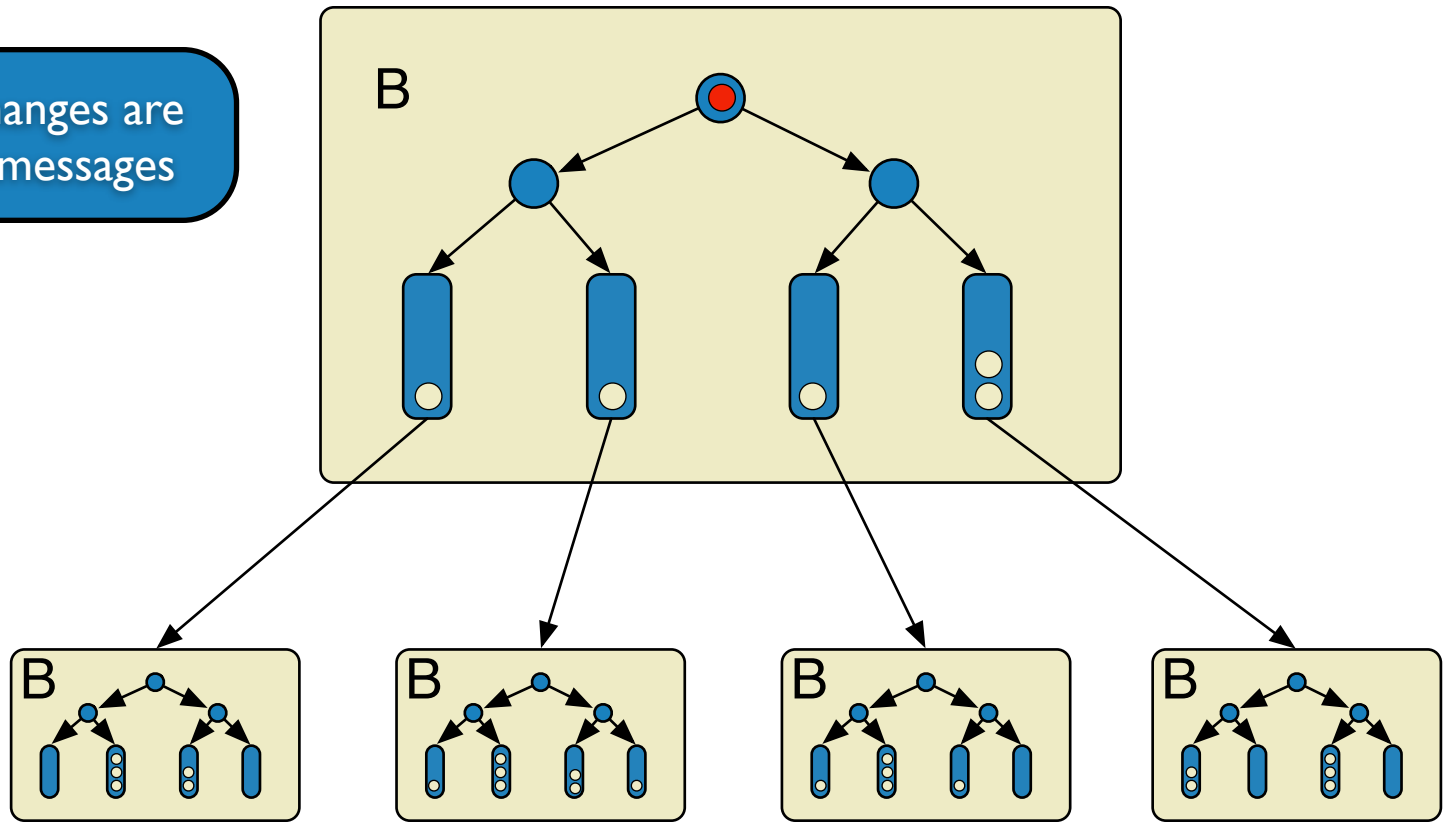
Schema Changes are broadcast messages



Tokutek®

Fractal Tree Indexes: Schema Changes

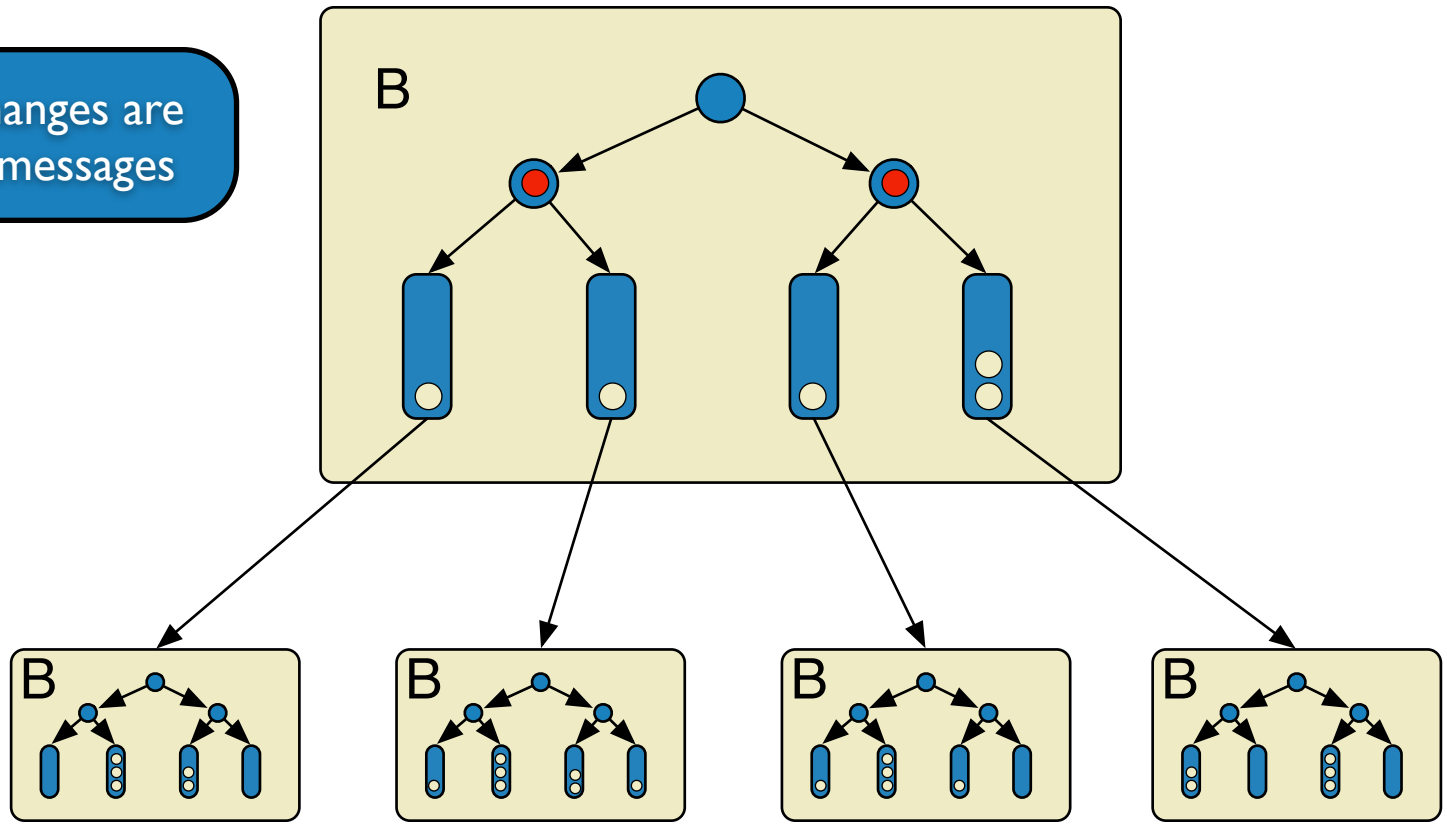
Schema Changes are broadcast messages



Tokutek®

Fractal Tree Indexes: Schema Changes

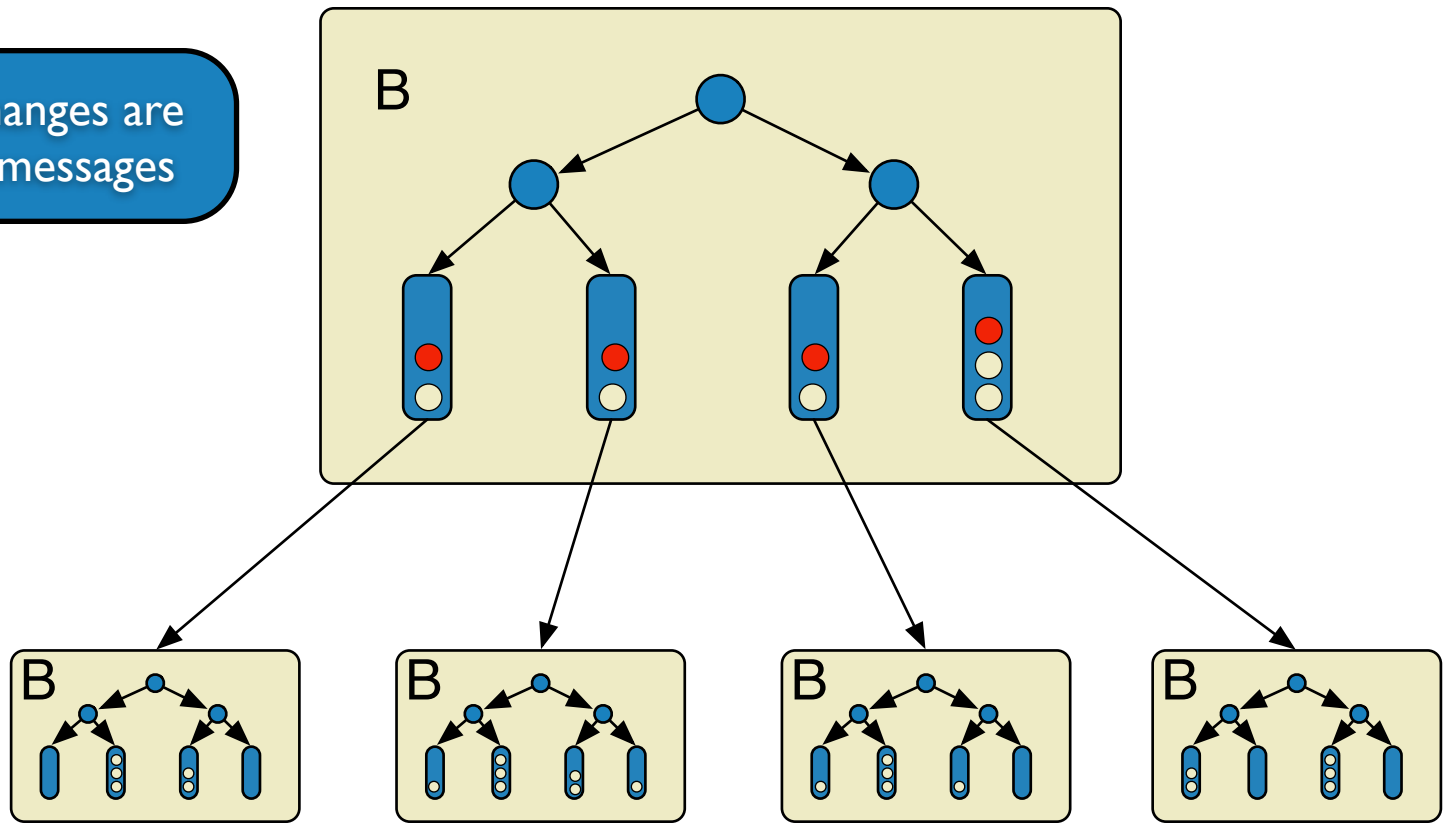
Schema Changes are broadcast messages



Tokutek®

Fractal Tree Indexes: Schema Changes

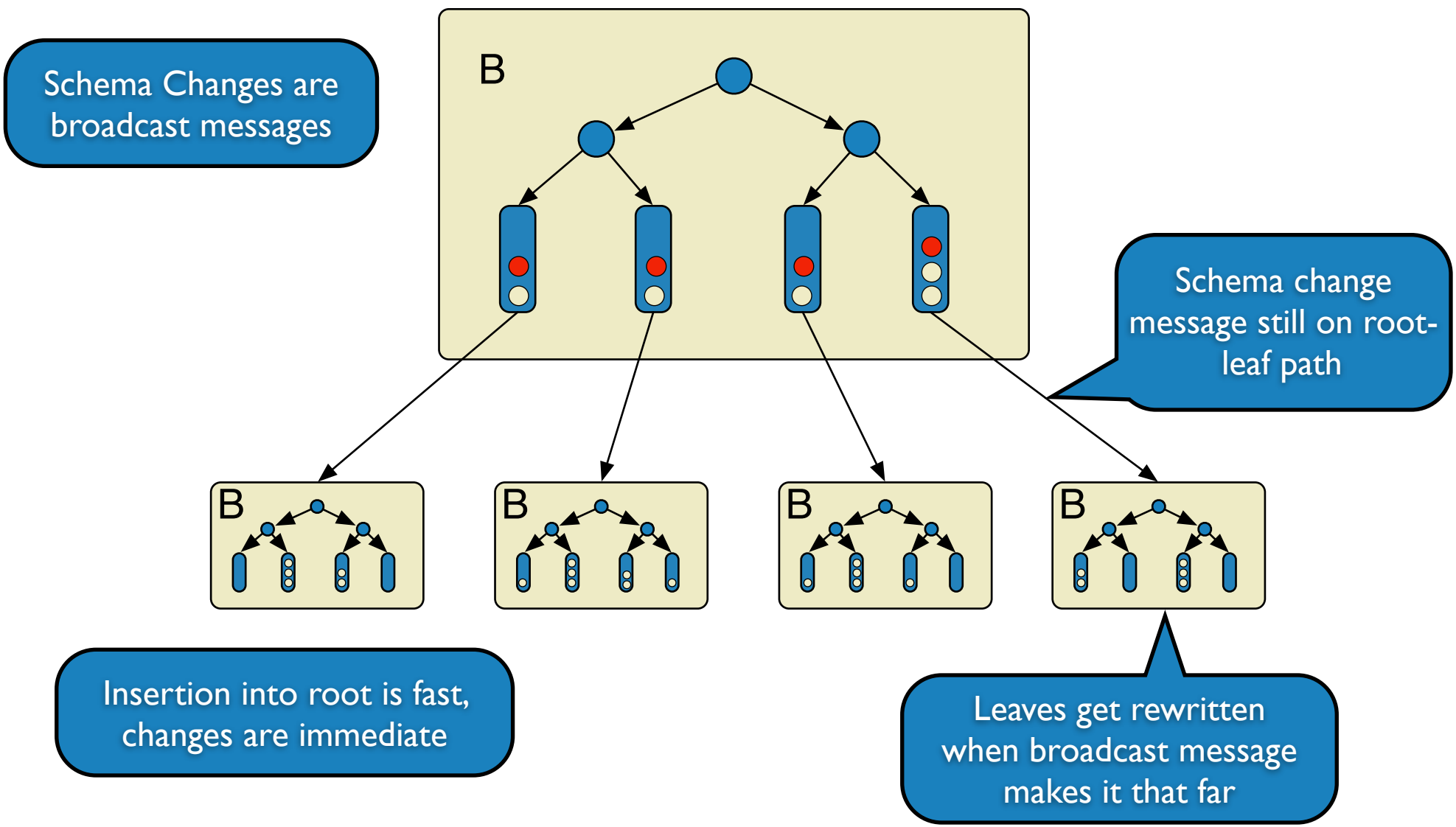
Schema Changes are broadcast messages



Insertion into root is fast, changes are immediate

Tokutek®

Fractal Tree Indexes: Schema Changes



Tokutek®

Analysis

Delivery system gives goodies:

- Messages get moved, but each I/O pays for a lot of movement
- You get very fast inserts
- You get Hot Schema Changes

Each flush carries lots of useful information

- So it's worth it to make nodes big
- No fragmentation, Better Compression
- Much better wear on SSDs