Outline

Who am I [Github]

Past Projects & Experiences

- VBASE (OSDI’23)
- Misc stuffs

Standing on Past Projects

My Interests in Research

Who am I

Zhizhen (Cathy) Cai

What am I

- USTC Graduate
- Former MSR system research group intern
  - Co-author of OSDI'23 paper *VBASE*
- USTC Linux User Group vice-president and Linux fan
- TA of Operating Systems and ...

I am a system researcher love "**hacking**" What is hacking?

I like systems for AI not AI for systems (but they are important, huh?)

Next: Past experience

My Past Things

So, why am I applying for PhD in systems?

- My first research internship in Big Data Analysis Lab (BDAA in USTC)
  - However, I soon found me not so into NLP/ML and quitted

- Then, TAships and Linux User Group
  - OS(Honored) and Web Info TA
  - We are all interested in Systems!

- MSR Asia System Research Group

VBASE

- VBASE is a database with efficient vector query support
- We discovered the relaxed monotonicity shared by main vector indices, and leverage this to embed vector index into traditional DB (PostgreSQL)

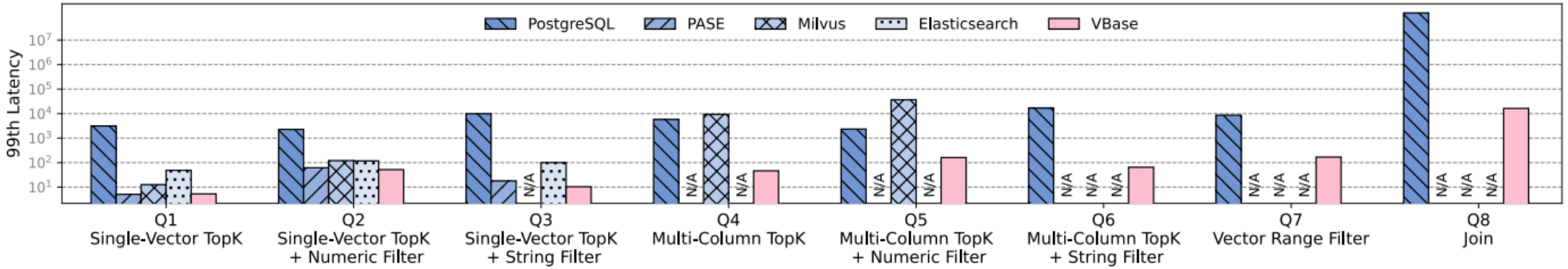


Figure 6: 99th Percentile Query Latency (ms)

Vector Search Background

- kNN: find k nearest points to query point

```
select * from images order by distance(q_embedding, embeddings)
ASC limit K;
```

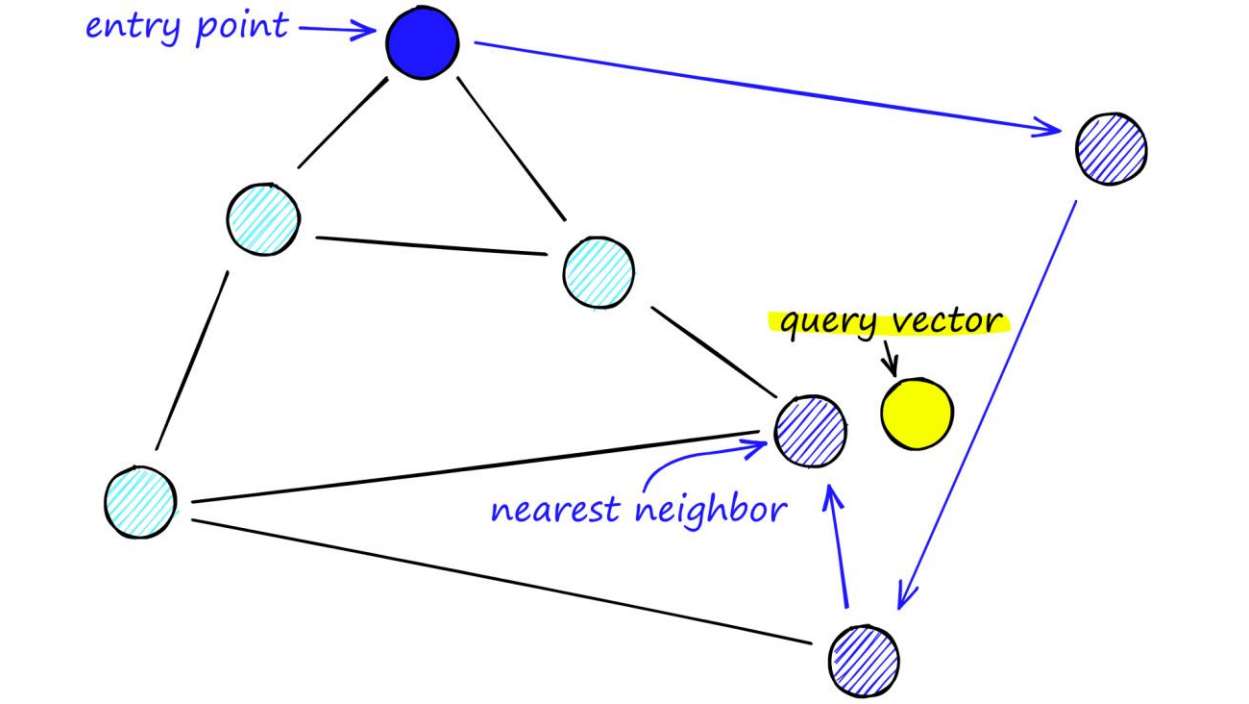
- Range Search: find all points with distance < r to query point

```
select * from images where distance(q_embedding, embeddings) <
R;
```

- But we cannot scan the whole data table for 100% recall
- Existing approximate vector index solutions:
  - HNSW, IVFPQ, LSH, …
  - Graph Based, Tree Based and Hash Based

Vector Search Background

- Example of graph based index: Navigable Small-World Graph



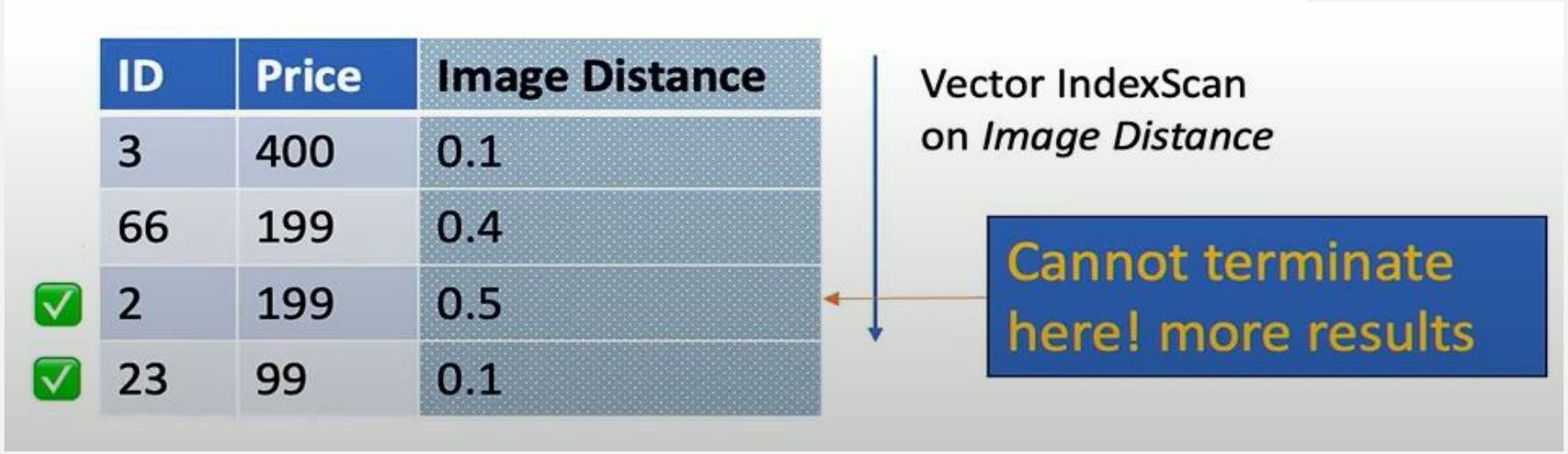
Pinecone: Hierarchical Navigable Small Worlds (HNSW)

- After enough points are visited, NSW aggregate the results

Why Vector DBs didn't Perform Well

- **Iterative** or **ordered**, can't we really have both?
- The dress example:

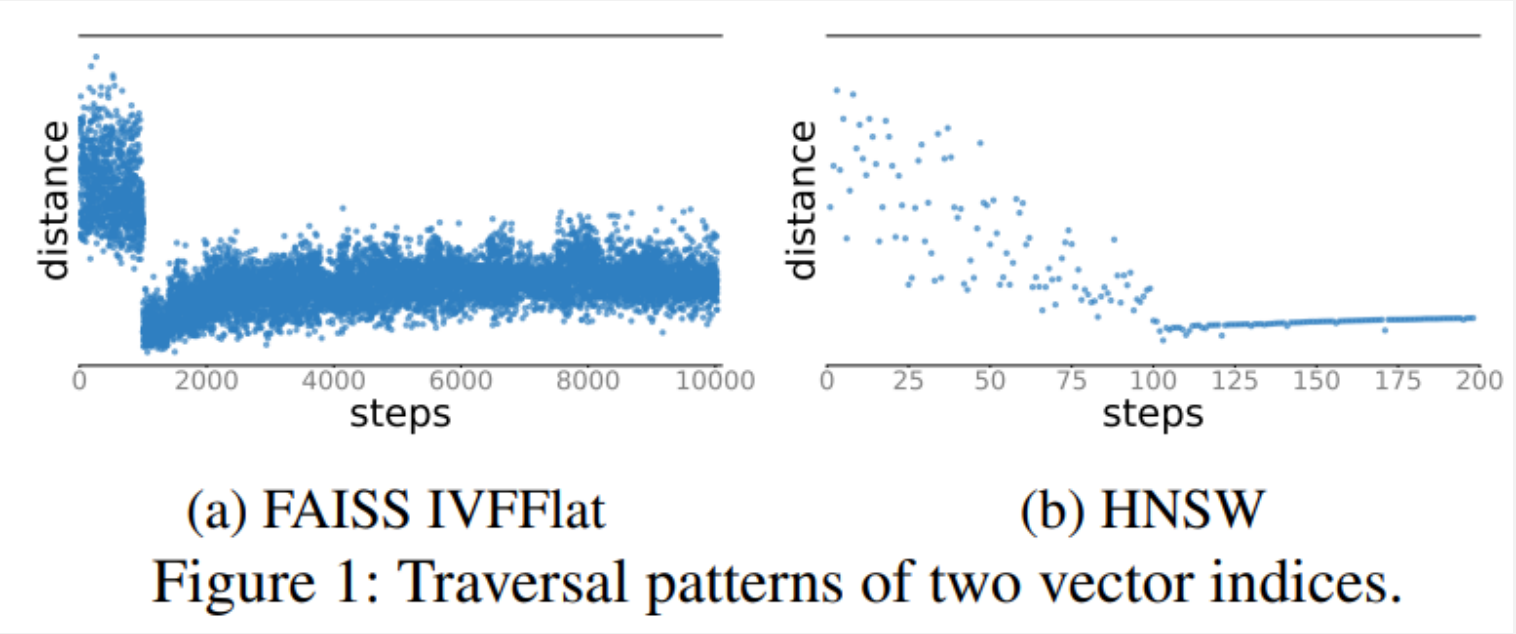
```
Select ID from dress where price < 200 order by distance(q_image, images) limit 2;
```



- DB engine leverage iterative model to efficiently apply multiple filters, but vector indices don't

Why Vector DBs Didn't Perform Well

- Previous Work (PASE): ordered but not iterative
  - 👎 Set an internal K', through trial and error: k = 2 → 4 → … till we have enough results
- Can we eat the cake and have it too?
  - VBASE: Yes, by exploiting the relaxed order shared in main vector indices



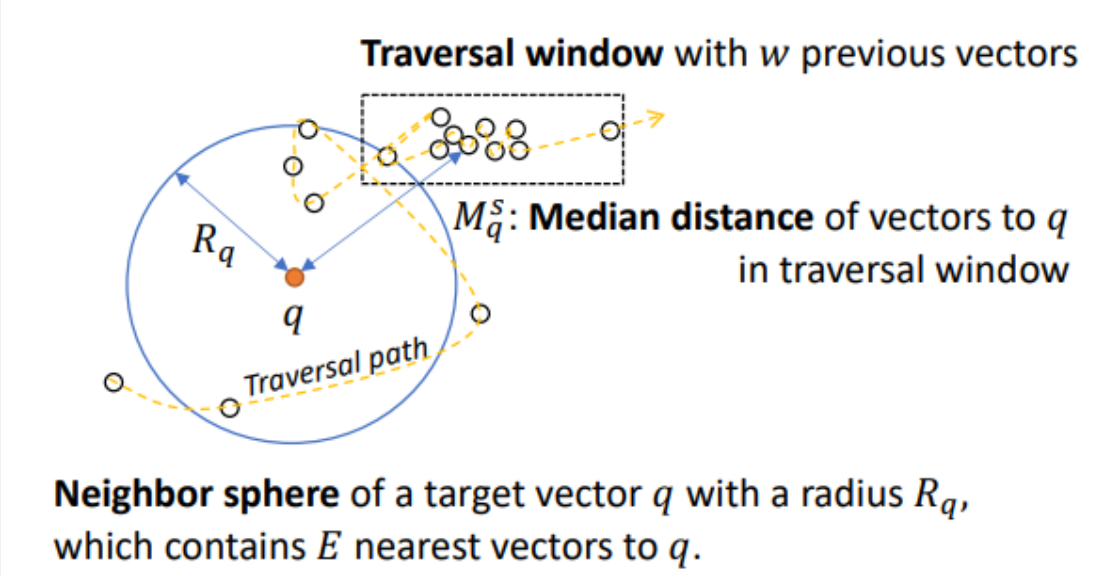
(a) FAISS IVFFlat (b) HNSW

Figure 1: Traversal patterns of two vector indices.

Next: vector index intro

VBASE's solution

- Use traversal window to identify the pivot in 2-phase scanning
  - Once enough out-of-range points visited, stop



- Now we can adapt DB's volcano model, return results iteratively and efficiently 👍

What's next

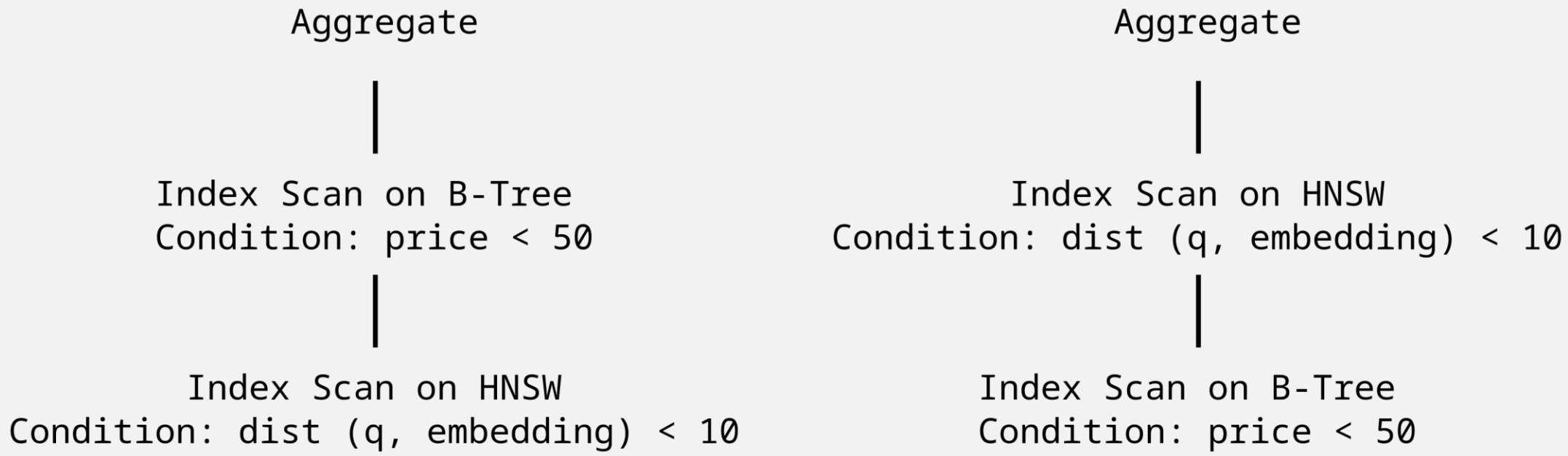
- There are a lot of possibility if we support iterative model:

- Better support to existing query:
  - Hybrid query (the dress example)
- New types of query:
  - Vector join (widely used in auto-tagging)

- Further optimization:
  - My work here!

Query Planning Intro

- Alike traditional queries, vector queries also have multiple plans.



Query execution trees, my Bachelor thesis

Next: visualization

My Work in VBASE

- As multiple plans are feasible, we can choose which to execute
  - How? Estimate the execution cost in advance

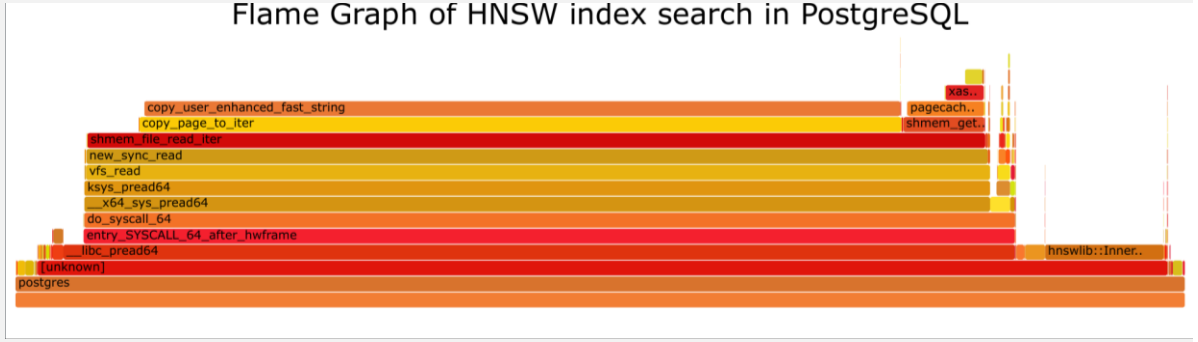
- I designed and implement a 2-layer cost model for vector index search
  - 2-layer here correspond to 2-step behavior in vector search (approach then apart)

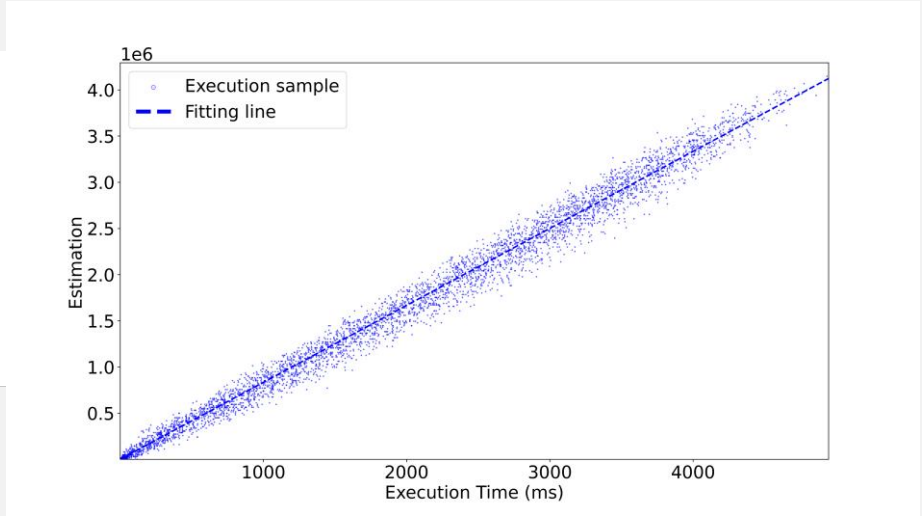
$C = C_{start} + C_{iter}$; C_{start} and C_{iter} are index-dependent

- (A lot of formula and evaluation following, I designed the cost model for each index, profiled, corrected and verified them)

My Work in VBASE

- Trivial things?
  - They are invisible in paper but took me 90% time 🤣
  - Browsing papers, communications, meetings and code demo
  - Fixing server bombed ~~by me~~ and then by another intern
  - Coding, Searching for documentation, **not documented?**
  - Profiled my advisor's code and sped them up by 100%

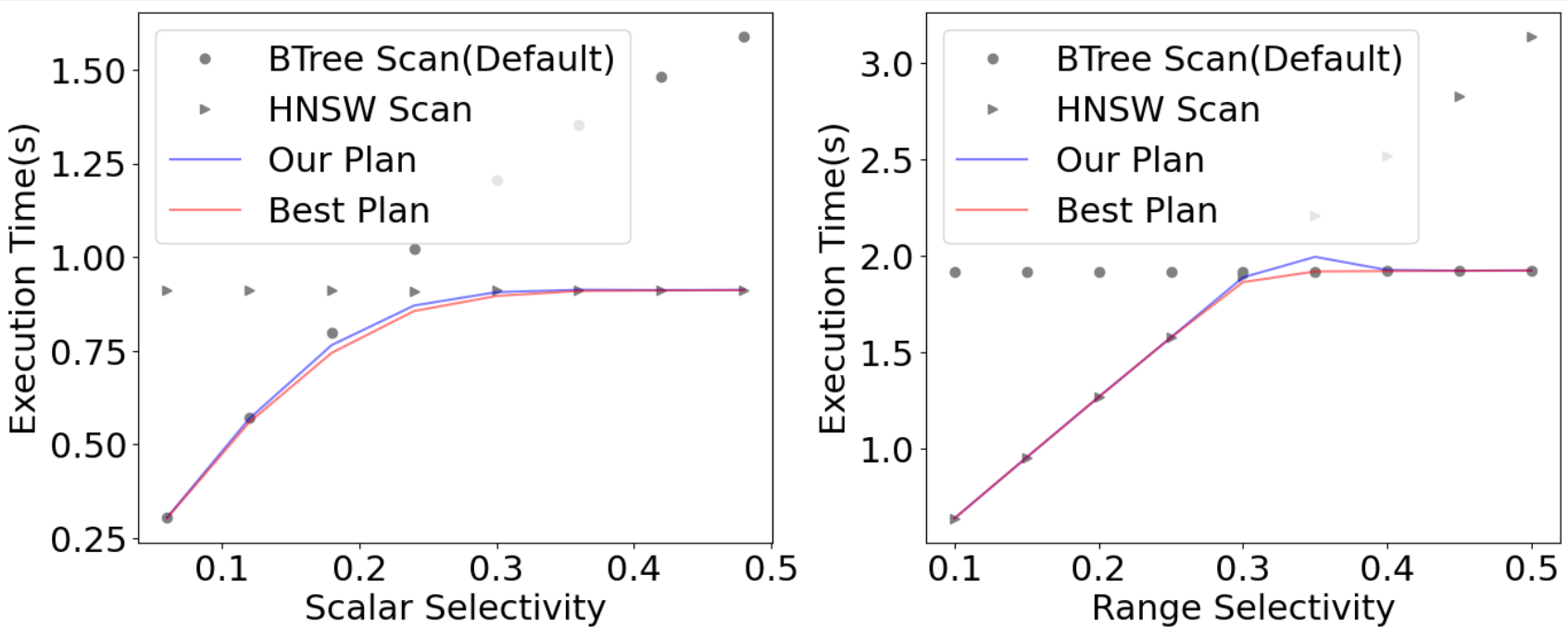




Next: Inspiration

Trivia?

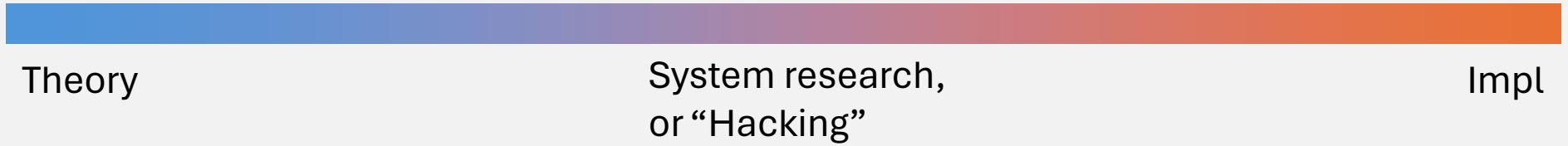
- That's what research (and esp. system research) like in most of the time
- And they formed a **solid basis** for my work, and VBASE's performance



VBASE guided by **my** planning

Why I'm Inspired

- What is system research to me



  - Hacking: The **art** of bridging ideas and code
- What am I?
  - Explorer
  - System "hacker": Combine different ideas and cook interesting systems
    - Pushing existing solutions to extreme: performance, reliability, …
  - Any examples?
    - my class proj: ROBDD [Github]
    - Papers/Projects?

Next: My non-academic expectation

What I’m Expecting (non-academic ones)

- WLB?
  - I’d like regular & balanced schedule (and a little bit flexibility is appreciated)
  - WFH is ok

- Communications?
  - The more the better, but high entropy is required

Q & A

Thank You!