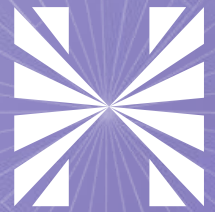
HAYSTACK



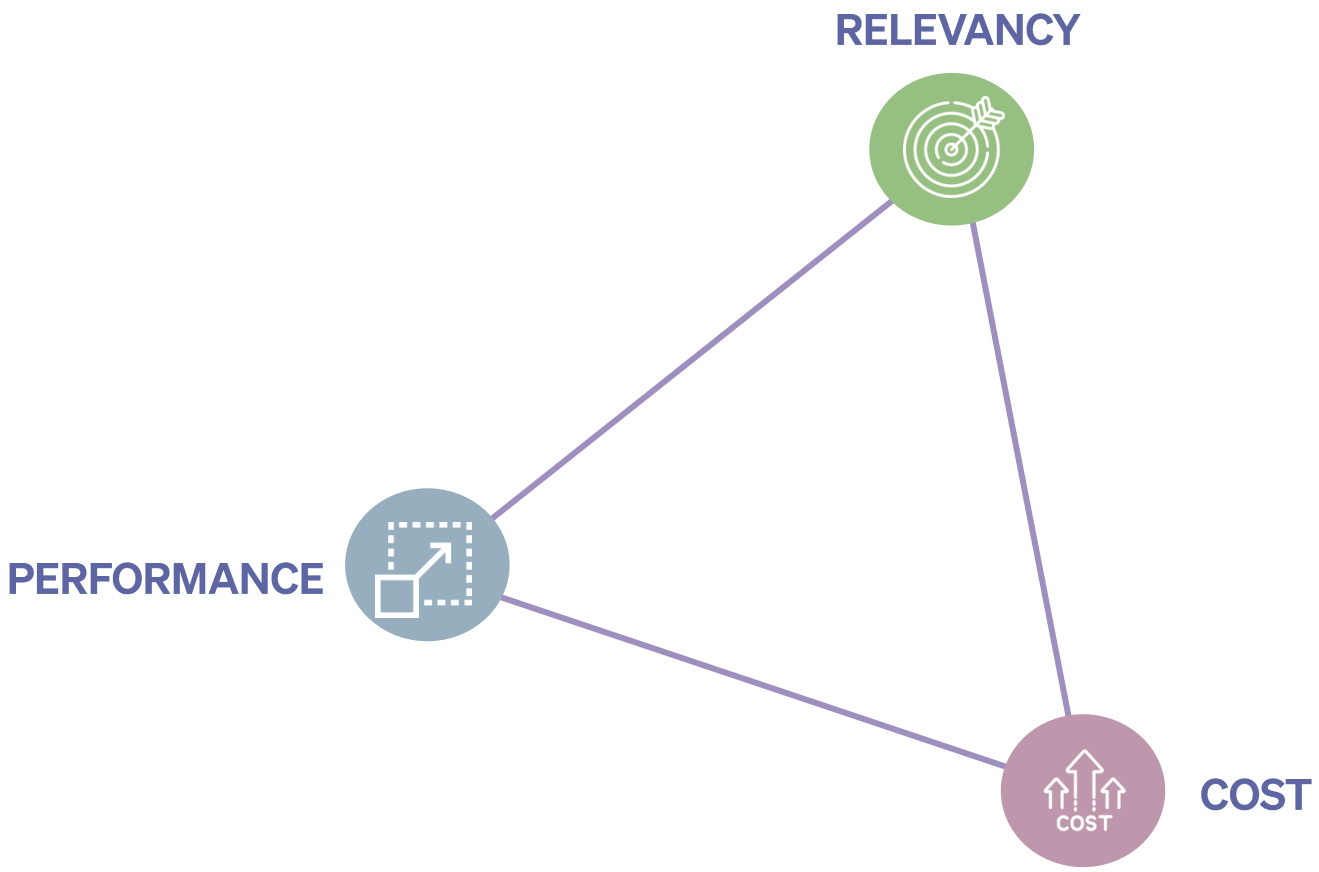
HYPERSPACE

Breaking Search Performance Limits with Domain-Specific Computing

linkedin.com/company/hyperspace-db/

ohadl@hyper-space.io

Tradeoff



About Me

- Technology builder
- Obsessed with Data, AI and Performance
- Built Data-Intensive applications from zero to 1 million users

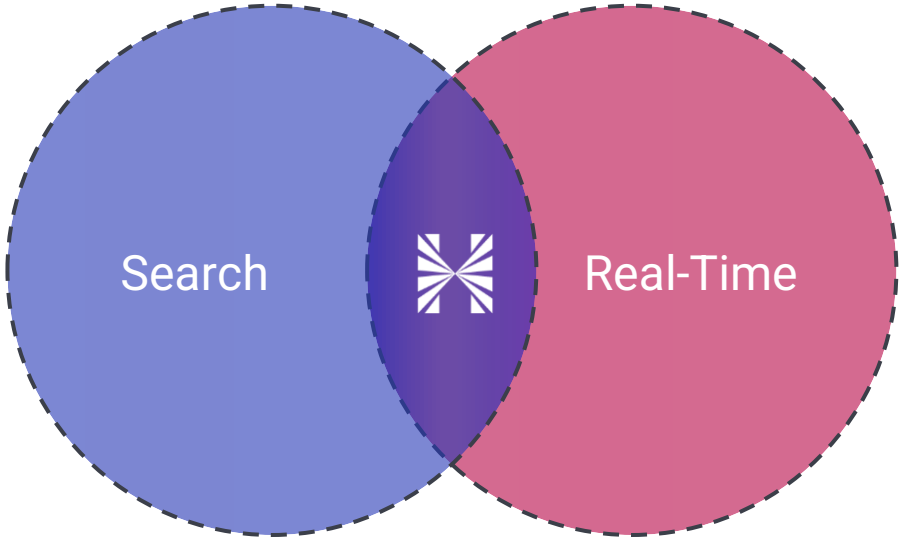
- Based in Tel Aviv
- Father for 3 amazing daughters
- Love Scuba diving, Hiking and recently Skiing

- **Founded Hyperspace to build the world's fastest search database**

Ohad Levi
Co-Founder and CEO

Today's talk

- Real-Time Search:
  - Use Cases
  - Market Trends
- The Challenges with Achieving Real-Time Search at Scale
- Domain specific computing
- Programming a Dedicated Chip for Search
- Results & Benchmarks



Why is Real-Time Search Important?

Amazon

Each 100ms Page Load = 1% in Revenue

Amazon study: Every 100ms in Added Page Load Time Cost 1% in Revenue

Last updated: AUGUST 10, 2021

Back in 2006, Amazon found that every 100ms in added page load time cost them 1% in sales. This has now become one of the most referenced data points surrounding page speed and web performance — standing the test of time as a clear example for why having a fast site is important.

For context, a 1% loss of annual revenue for Amazon in 2006 would have been around $107 million. Today, this would be about $3.8 billion!

Google

Web search latency directly impact user engagement

Speed Matters for Google Web Search

Jake Brutlag
Google, Inc.
June 22, 2009

Abstract – **Experiments demonstrate that increasing web search latency 100 to 400 ms reduces the daily number of searches per user by 0.2% to 0.6%. Furthermore, users do fewer searches the longer they are exposed. For longer delays, the loss of searches persists for a time even after latency returns to previous levels.**

Google runs experiments on search traffic to understand and improve the search experience. A series of such experiments injected different types of server-side delay into the search results page load in order to understand the impact of latency on user behavior. In a given experiment, one group of users experienced the delay, while a second group served as the control. Across the experiments, the type of delay, the magnitude of the delay, and experiment perceived differently by users due to the degree of partial rendering on the page.

All other things being equal, more usage, as measured by number of searches, reflects more satisfied users. Table 1 gives the average daily searches per user over the experiment duration for the experiment group relative to the control group.

Table 1: Experiment Impact on Daily Searches Per User

Type of Delay	Magnitude	Duration	Impact
Pre-header	50 ms	4 weeks	—
Pre-header	100 ms	4 weeks	−0.20%
Post-header	200 ms	6 weeks	−0.29%
Post-header	400 ms	6 weeks	−0.59%
Post-ads	200 ms	4 weeks	−0.30%

HYPERSPACE

Booking

An increase of 30% in latency costs about 0.5% in conversion rates

Lesson 4: prediction serving latency matters

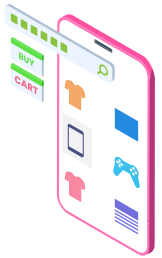
Here we have yet another data point on the impact of performance on business metrics. In a experiment introducing synthetic latency, Booking.com found that an increase of about 30% in latency cost about 0.5% in conversion rates "*a relevant cost for our business*".

> This is particularly relevant for machine learned models since they require significant computational resources when making predictions. Even mathematically simple models have the potential of introducing relevant latency.

Booking.com go to some lengths to minimise the latency introduced by models, including horizontally scaled distributed copies of models, a in-house developed custom linear prediction engine, favouring models with fewer parameters, batching requests, and pre-computation and/or caching.

Real-Time Search Applications

E-commerce



Recommendation engines

Fraud Prevention

Marketplaces

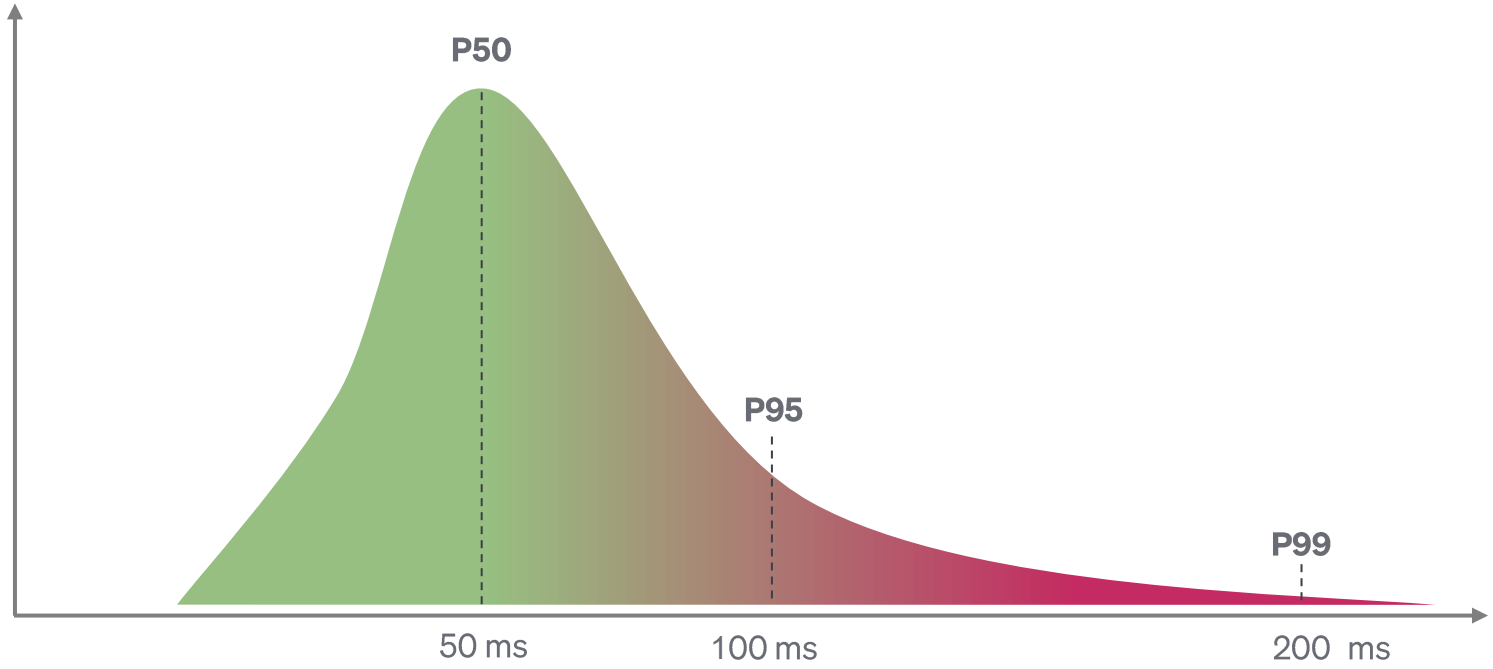
Cyber / Bot detection

Challenges

Big Data — Compute-Intensive — Real-Time

What is Real-Time Search?

- Removing network and pre/post processing tasks, search query typically has 100ms to run
- Latency limits are being dictated by user experience & engagement
- Maintaining low latency below 100ms at scale optimization and compromises



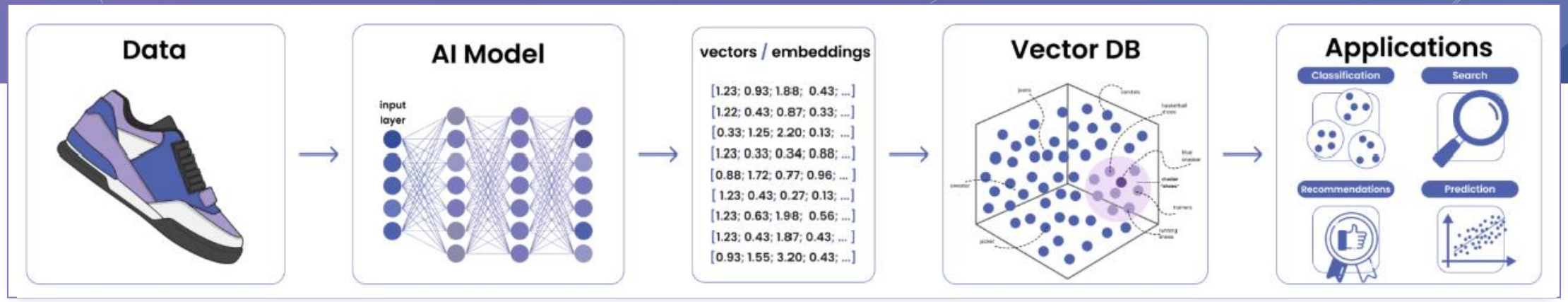
Real-Time Search < 100ms

HYPERSPACE

Trends in Search Applications

Vector Search

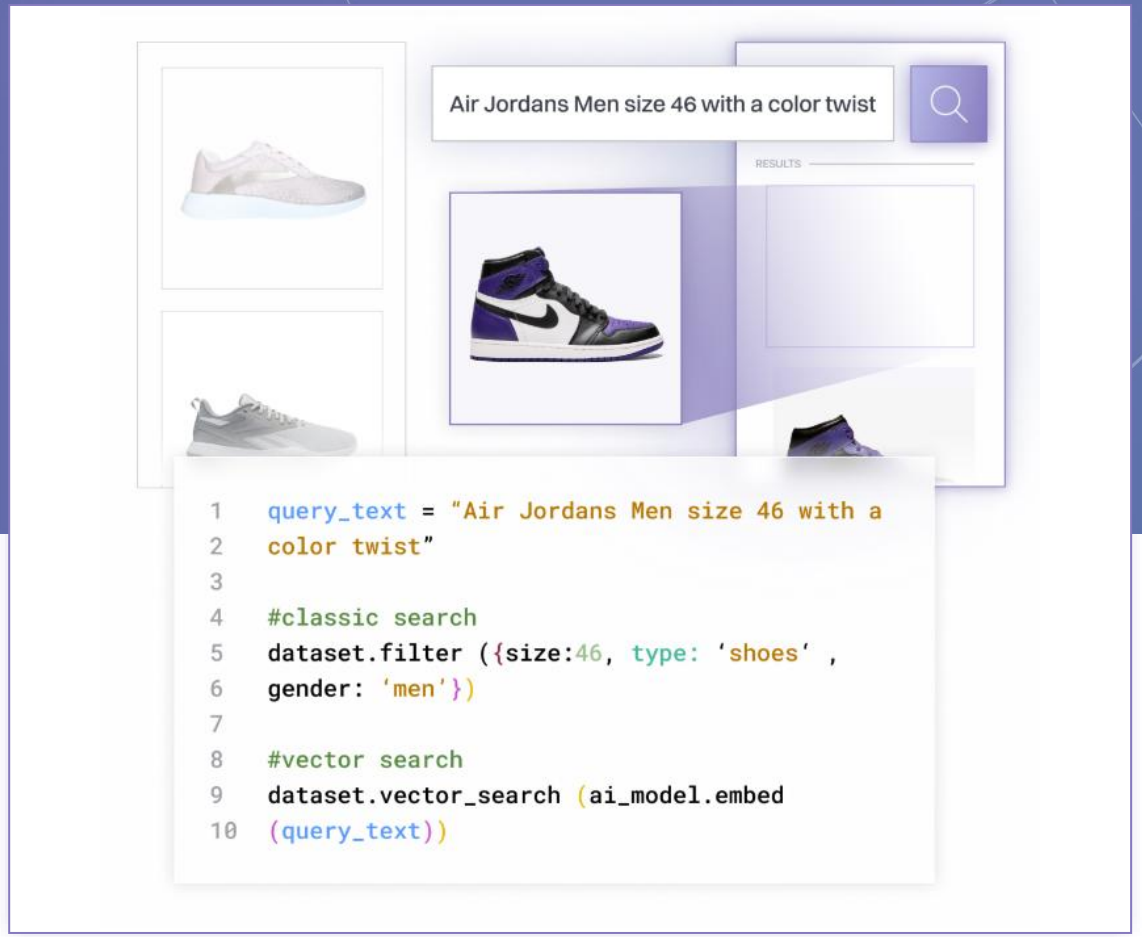
- Everything2vec trend - by representing almost every object as a vector, it then becomes available for similarity search
- Search Algorithms: k-Nearest Neighbor (k-NN), Approximate Nearest Neighbor (ANN)
- Neural search / Semantic search



Hybrid Search

Running state-of-the-art search queries, combining vector search with the classic search functions:

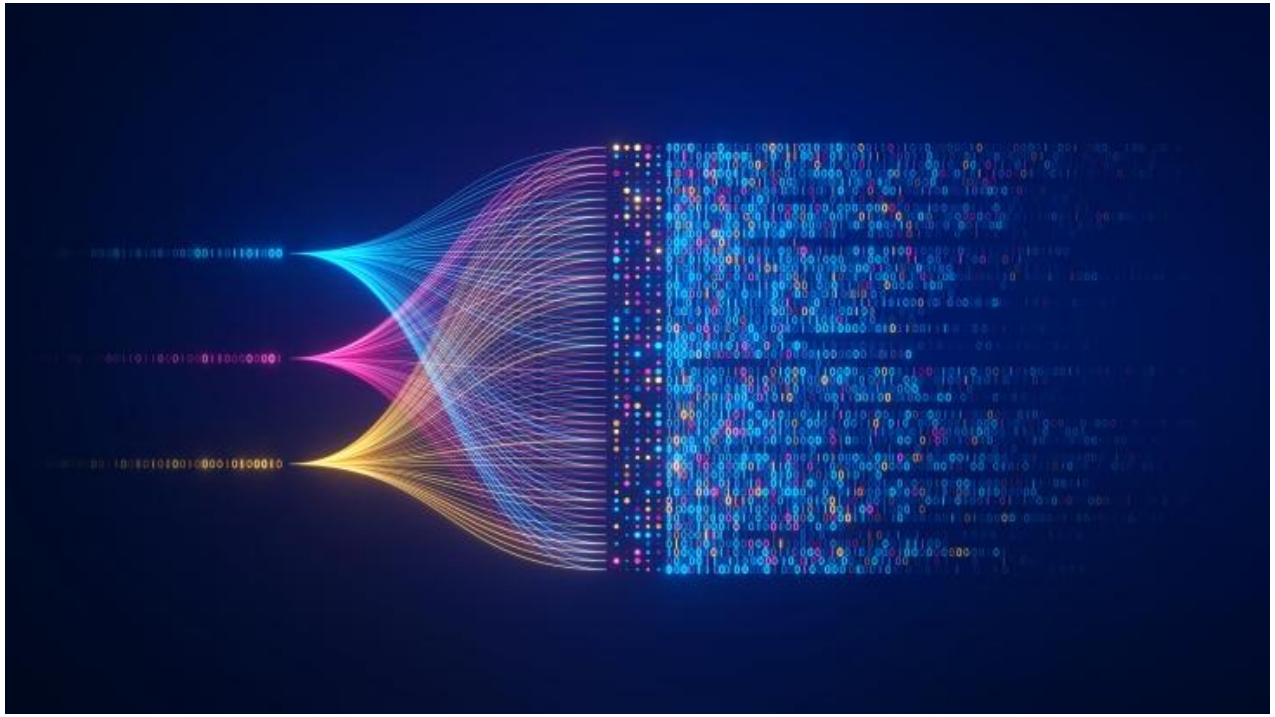
- Metadata Filters
- TF-IDF/BM25
- Aggregations



HYPERSPACE

Generative Search

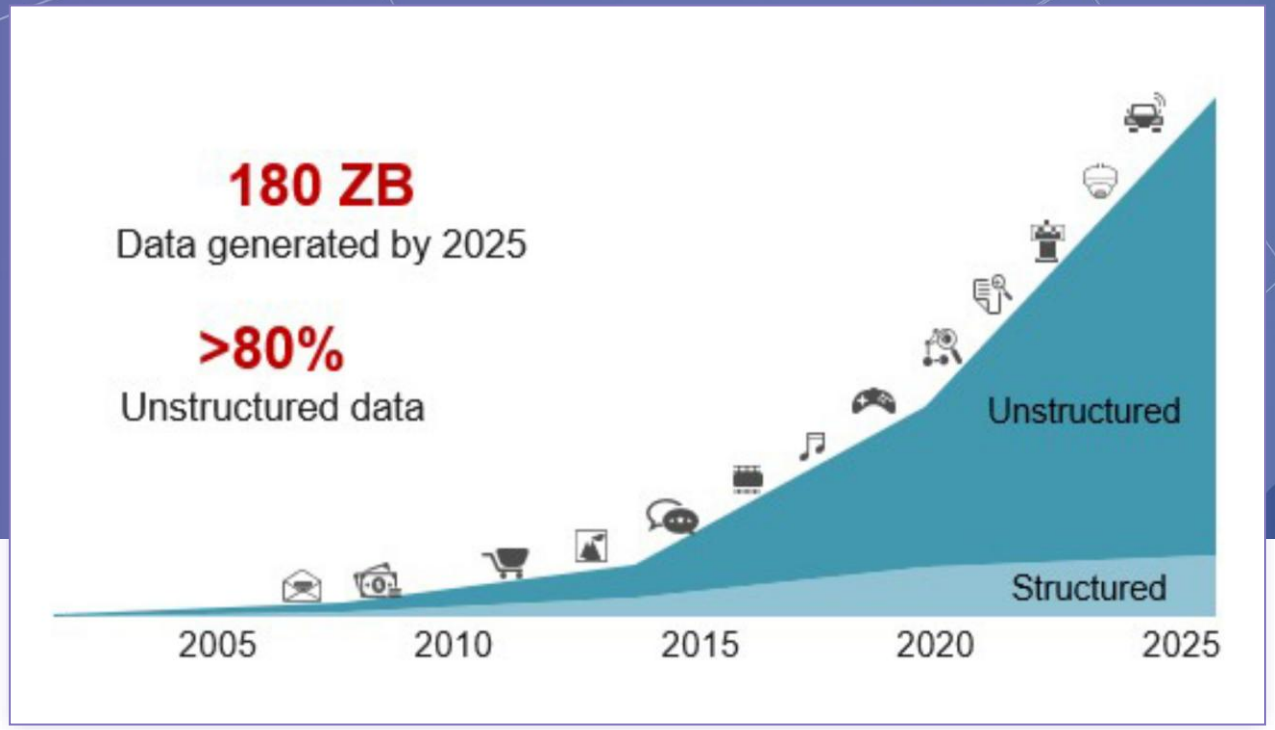
- Enrich search relevancy with large language models (LLMs) to create next-gen personalized search experiences
- Enable AI-powered search models that closely emulate the complex processes of human cognition



Challenges of Achieving Real-Time Search at Scale

Data is growing exponentially

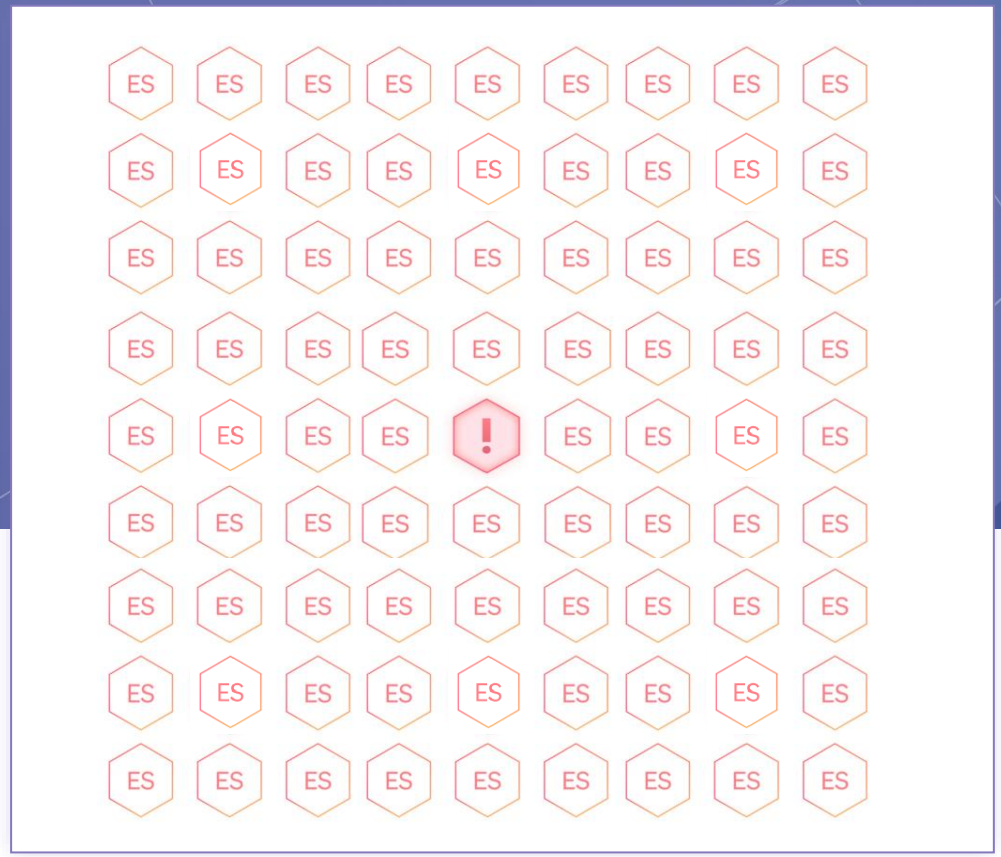
- Searching through billions of items is becoming a standard
- 80% of the world's data is becoming unstructured



Source: IDC report - 80% of the world's data is unstructured

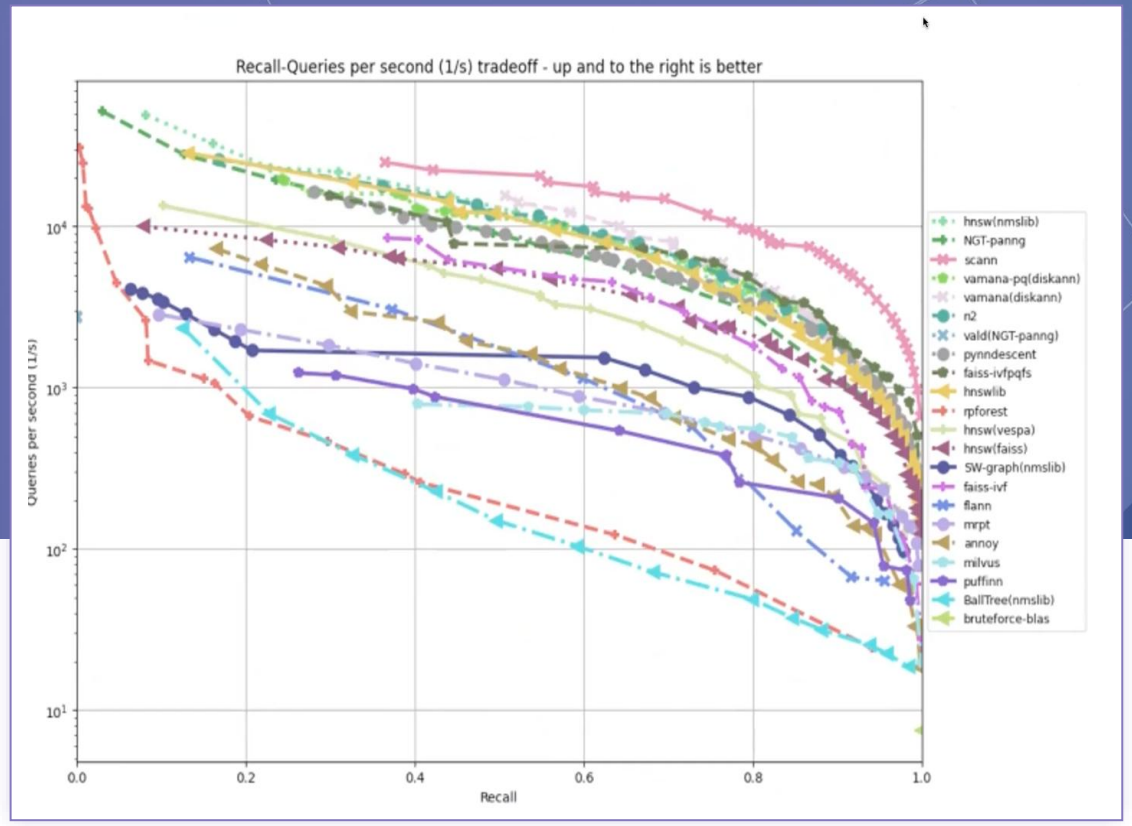
Infrastructure costs are soaring

- Large indexes kept in-memory for high-performance
- High query throughput during surges (e.g. black Friday)
- Results in extremely large clusters of powerful servers



Performance vs. Accuracy Tradeoff

- Performance limitations forcing sub-optimal methods such as Approximate Nearest Neighbor (ANN) instead of k-nearest neighbors (k-NN)
- ANN algorithms have to balance between performance (queries per second) and accuracy (recall)

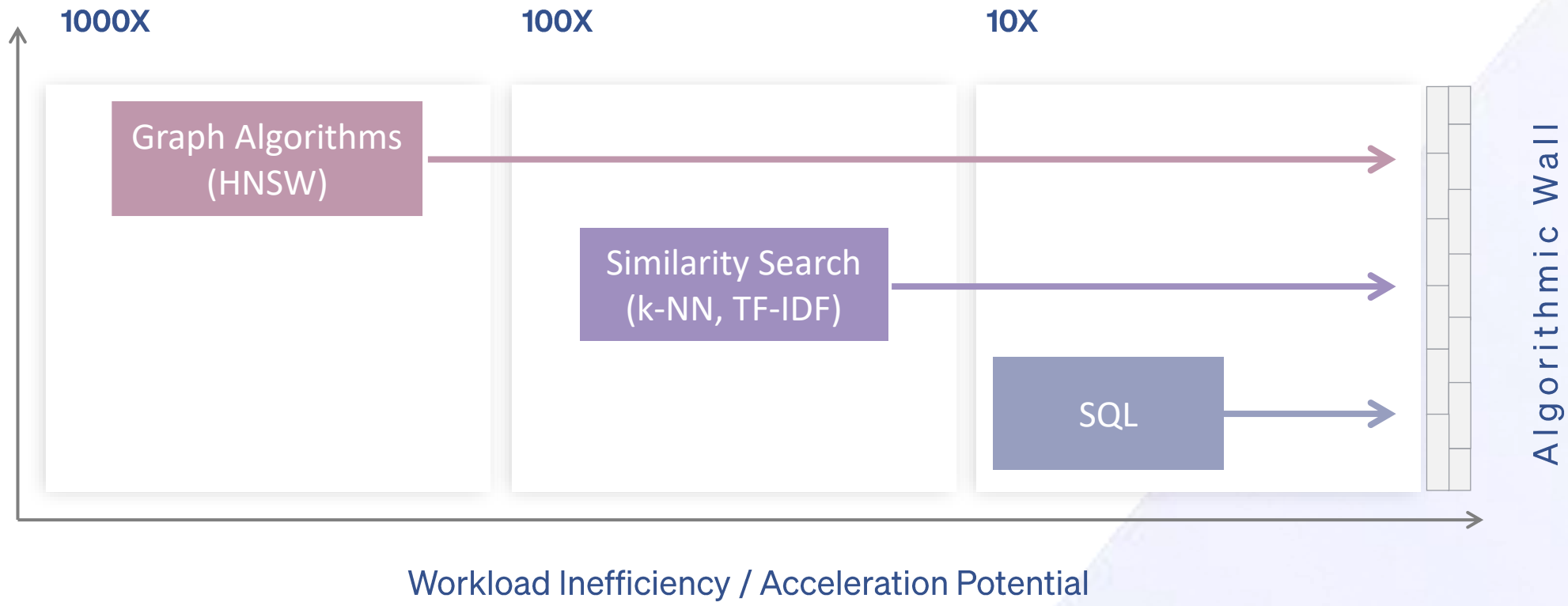


Source: Ann-benchmarks.com

Domain-Specific Computing

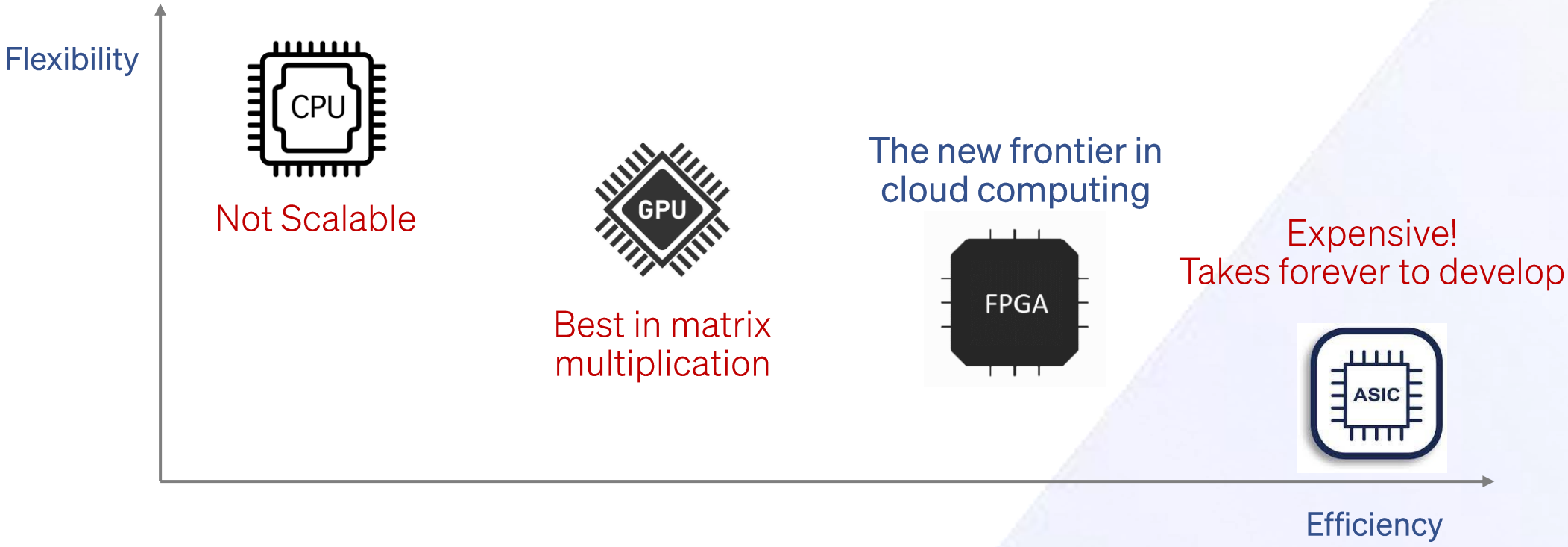
Software Inefficiency

- Software-only solutions don't scale, especially when dealing with complex algorithms, large datasets, or where the software is not well parallelized.
- CPUs are designed to handle a wide range of tasks, but they are not optimized for any specific workloads



Workload-Specific Acceleration

- AI is approaching computational limits, forcing companies to pay millions of dollars for powerful servers
- FPGA: optimal solution for domain specific computing



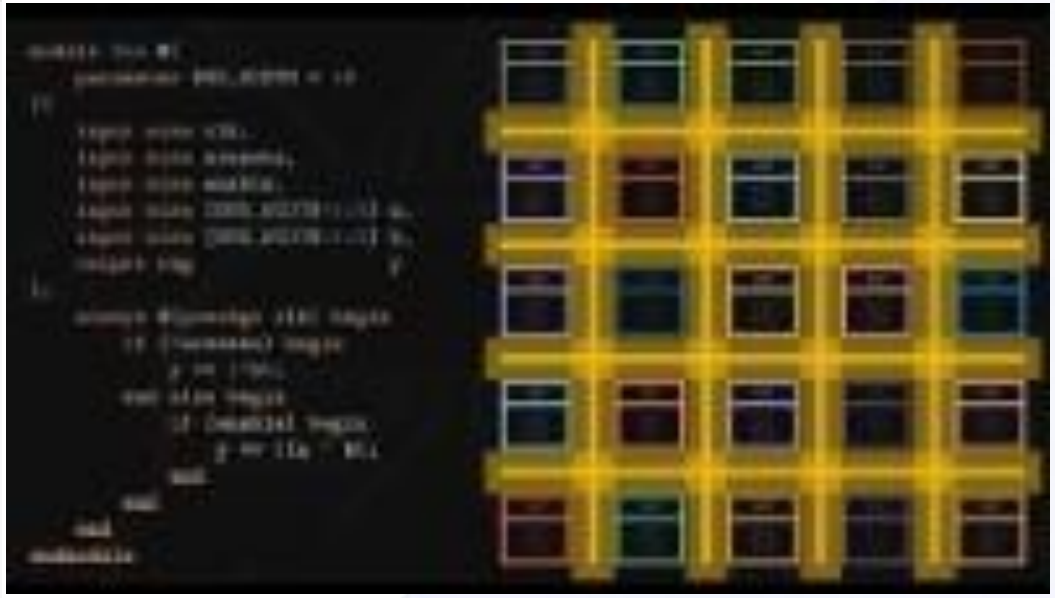
What is FPGA?

Explain in brief, what is FPGA and what is it good for?

FPGA stands for Field-Programmable Gate Array. It is a type of integrated circuit that can be programmed to perform a specific set of functions or tasks. Unlike other types of integrated circuits, which are designed for specific functions, an FPGA can be reprogrammed multiple times to perform different functions, making it a versatile and flexible solution for many different applications.

FPGAs are used in a wide range of applications, including digital signal processing, image processing, and telecommunications, among others. They are particularly useful in applications that require high-speed processing, low latency, and low power consumption. FPGAs can also be used to create custom logic circuits, which can be optimized for specific tasks or applications, providing better performance and energy efficiency than off-the-shelf solutions.

FPGAs are also popular in the field of hardware acceleration, where they are used to accelerate tasks that would otherwise be performed by software running on a general-purpose processor. By offloading these tasks to an FPGA, the overall system can achieve better performance, lower latency, and lower power consumption.



How can FPGA help improve ChatGPT?

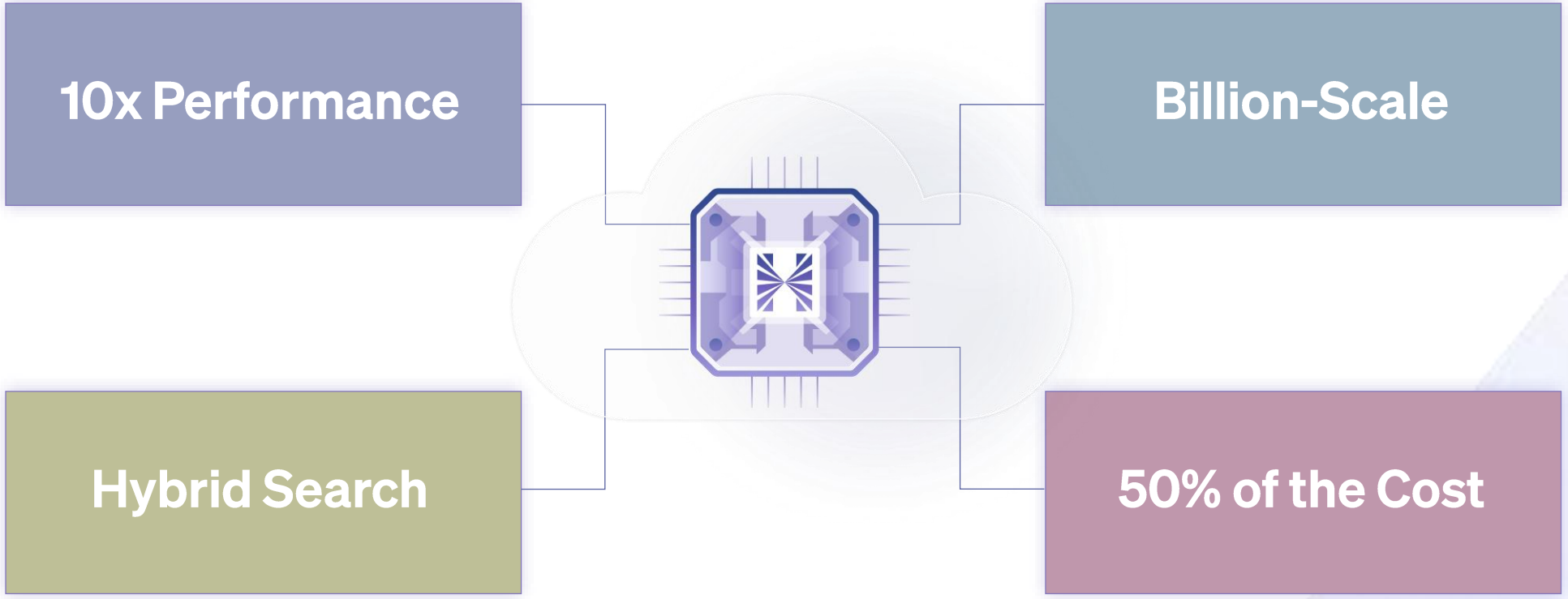
FPGAs could potentially improve ChatGPT in several ways:

1. Acceleration: FPGAs can be used to accelerate the inference of deep learning models like ChatGPT. By offloading the compute-intensive operations to the FPGA, ChatGPT can potentially achieve faster response times and handle more concurrent user requests.
2. Customization: FPGAs can be programmed to implement custom logic, allowing ChatGPT to be optimized for specific tasks or applications. For example, if ChatGPT is being used to generate responses in a particular domain, the FPGA could be programmed with specialized hardware that is designed to process the relevant data more efficiently.
3. Lower power consumption: FPGAs can be more power-efficient than general-purpose CPUs or GPUs. By using FPGAs to accelerate ChatGPT, it may be possible to reduce the overall power consumption of the system, which could be beneficial in applications where power consumption is a critical factor.
4. Improved scalability: FPGAs can be used to implement scalable architectures that can handle increasing workloads. This could be particularly useful for ChatGPT, which is designed to handle large volumes of user requests. By using FPGAs, ChatGPT could potentially scale up to handle more requests without incurring a significant increase in latency.

In summary, FPGAs have the potential to accelerate ChatGPT, customize it for specific tasks, reduce power consumption, and improve scalability, all of which could result in a better user experience.

Designing a Search Processing Unit (SPU®)

Leverage domain specific computing to deliver unmatched search performance in data-intensive applications, surpassing the limitations of traditional software-based solutions.

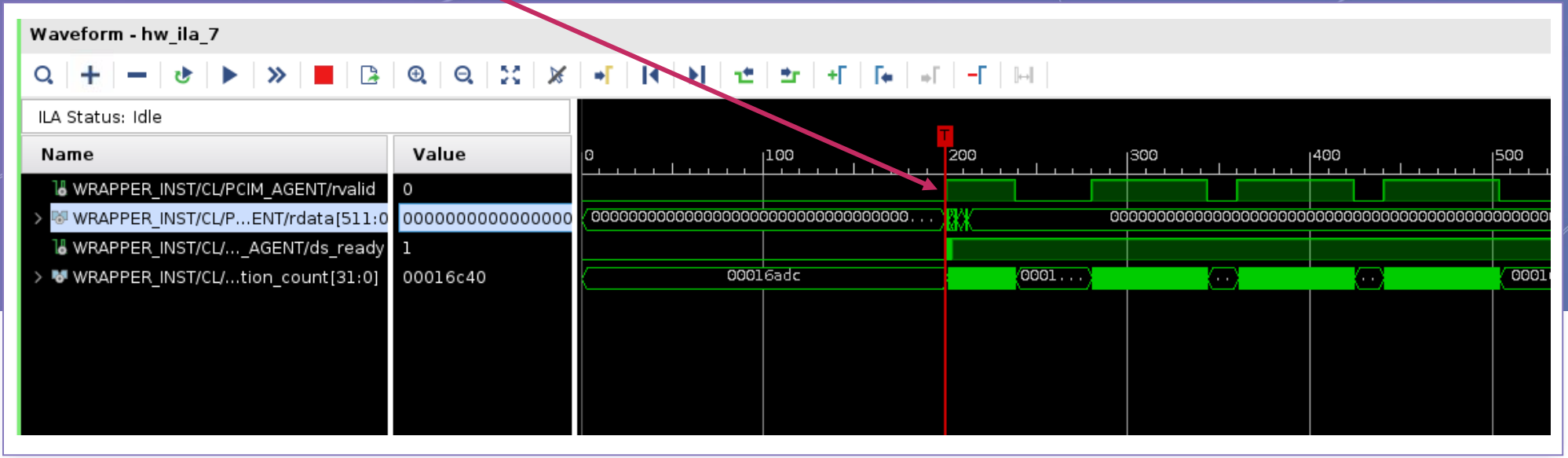


Cloud Opportunity

- Hyperspace SPU® is a cloud-native managed database powered by AWS F1 instances (cloud-based FPGAs)
- Available via Elasticsearch-compatible API, delivering unmatched search performance in data-intensive applications

SPU® - under the hood

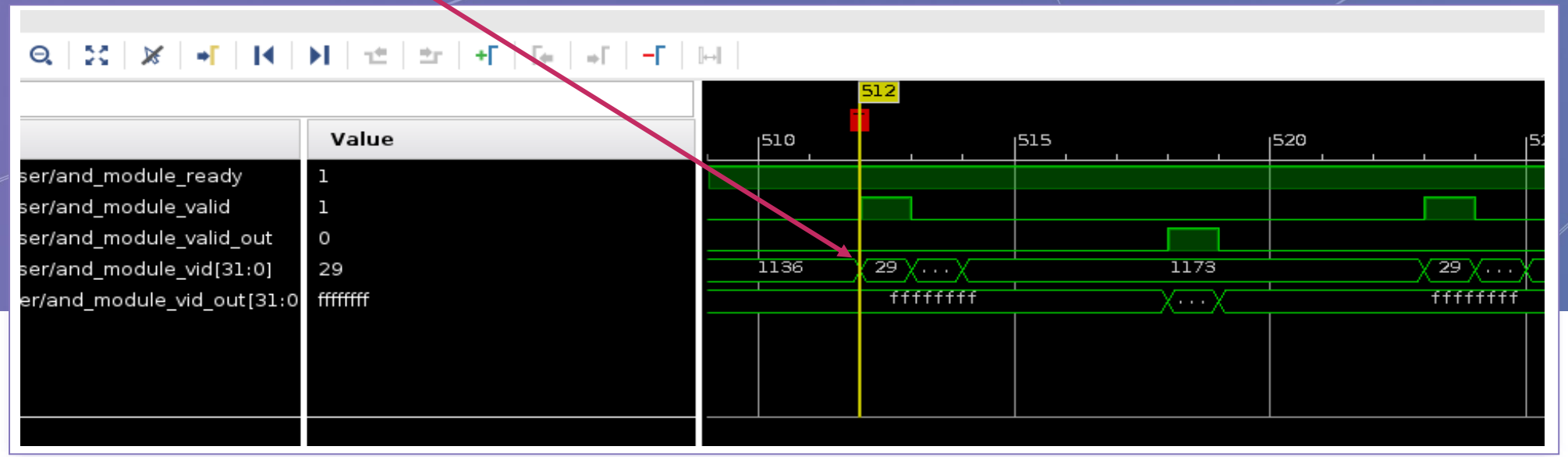
Incoming search received in the SPU



* Cycle by cycle view of the data processed in the SPU in nanosecond resolution.

SPU® - under the hood

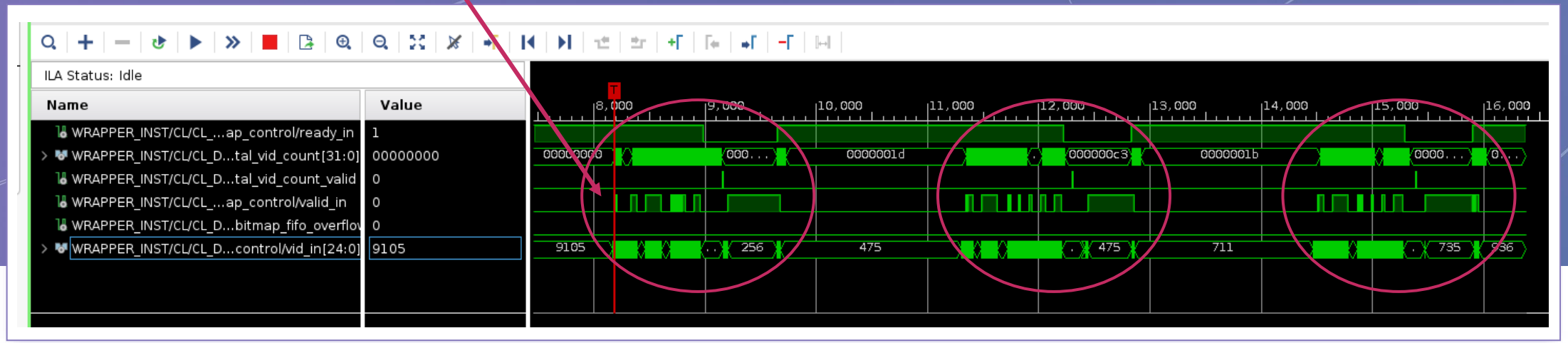
Document fetched from the index



* Cycle by cycle view of the data processed in the SPU in nanosecond resolution.

SPU® - under the hood

Documents handled in the candidate generation process

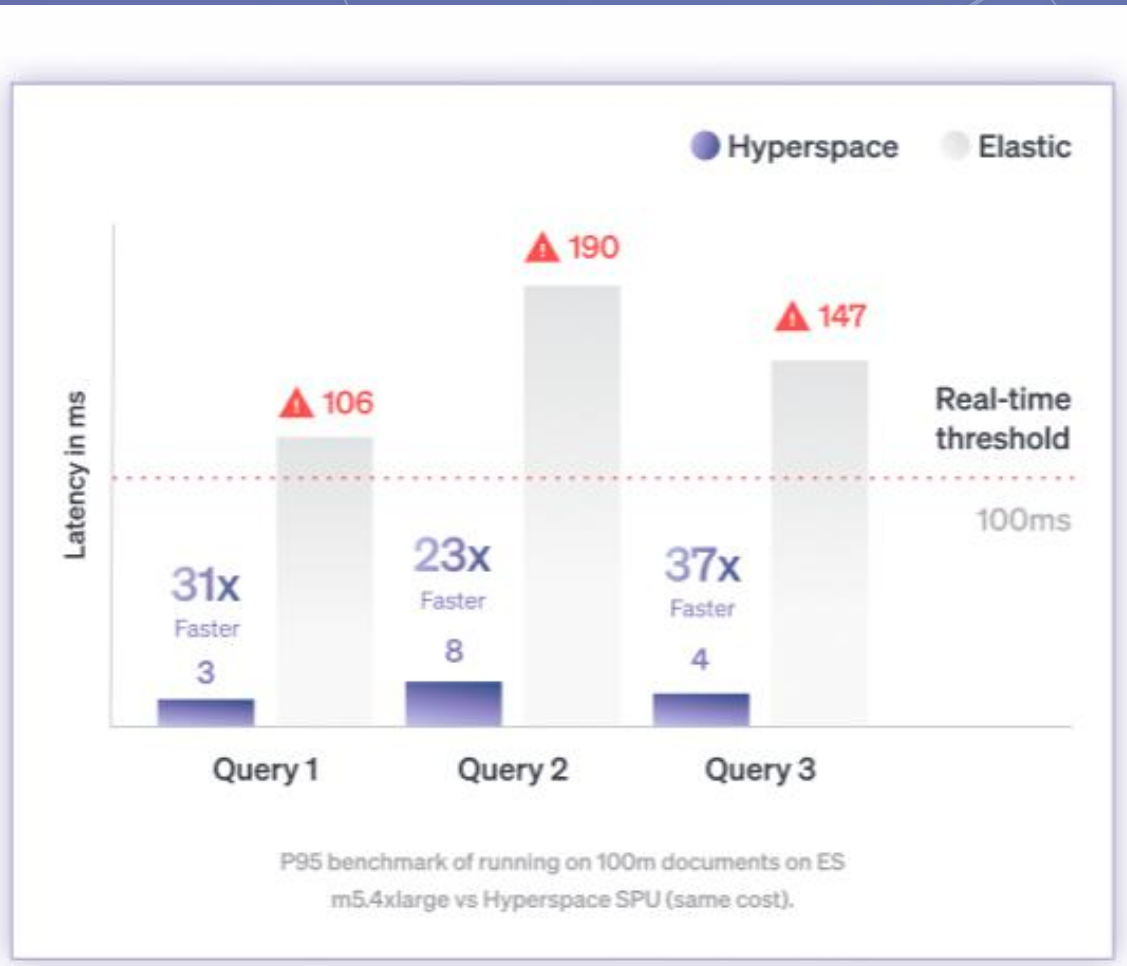


* Cycle by cycle view of the data processed in the SPU in nanosecond resolution.

Results & Benchmarks

Breaking the Limits of Search

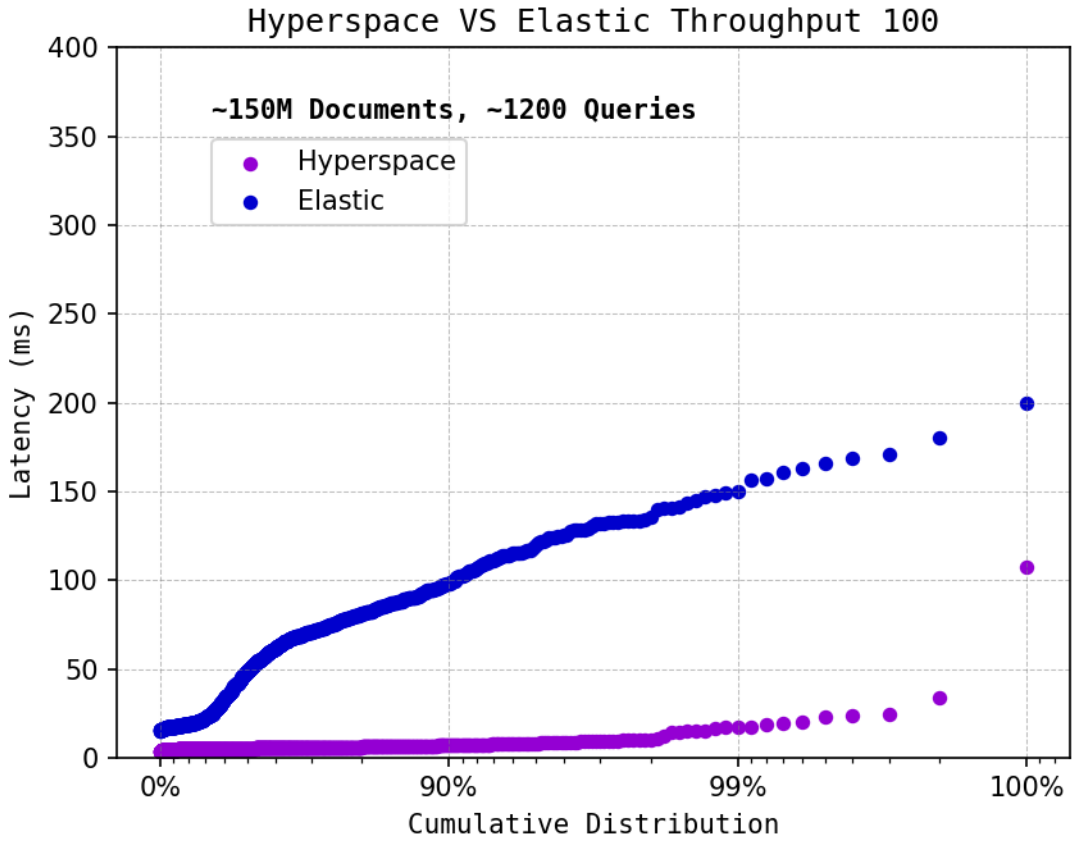
- Running complex search queries in milliseconds!
- Perform deeper search at a fraction of the current compute costs



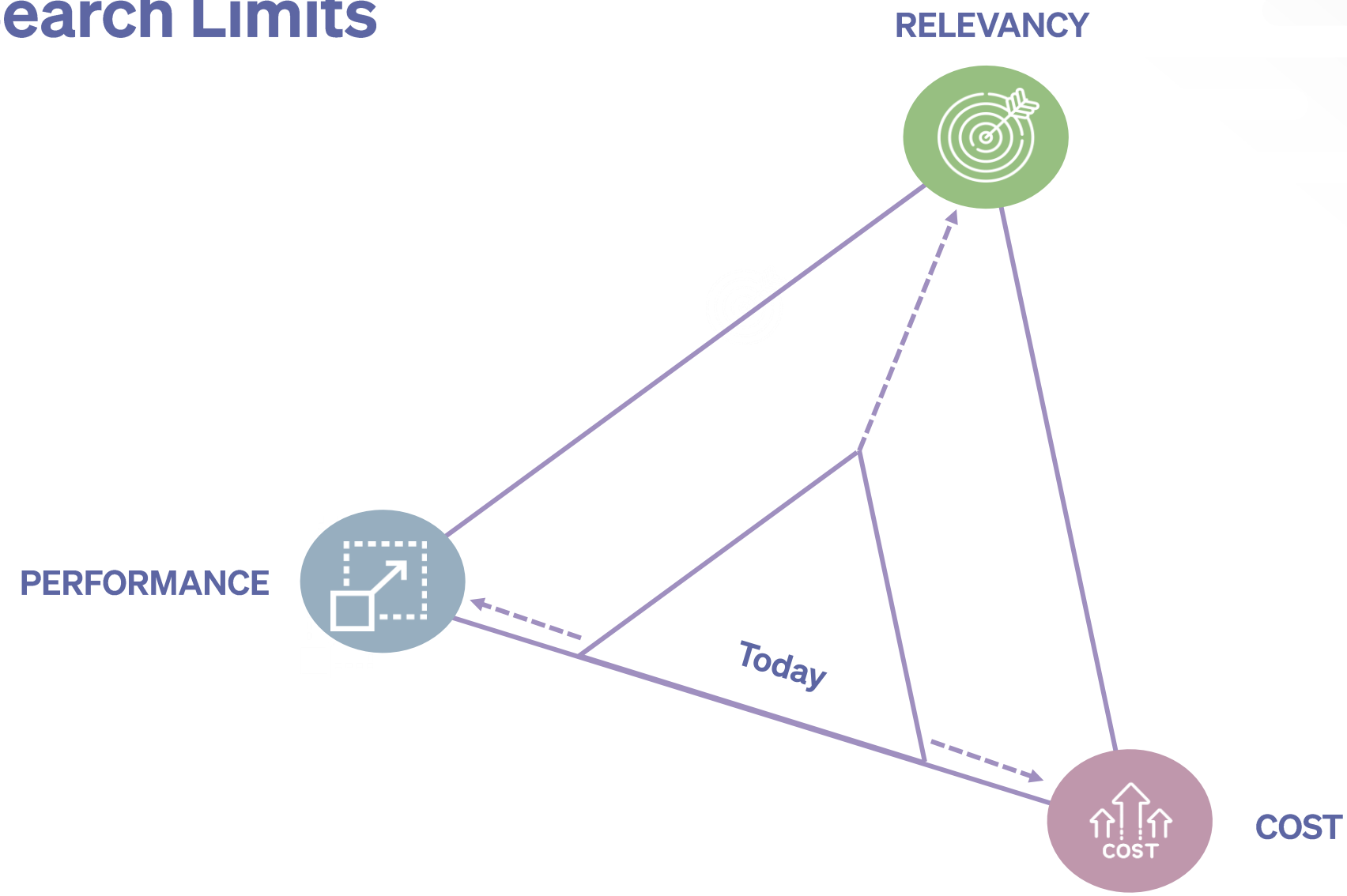
HYPERSPACE

Hyperspace SPU® vs. Elastic running on CPU

- 150m documents
- Query includes filters, TF/IDF and aggregations
- Running 1,000 queries
- 100 QPS



Breaking Search Limits



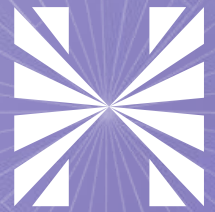
Hyperspace

- Building the world’s fastest search database, breaking today’s search limits
- Based in Tel-Aviv
- 20 employees
- Beta ready
- Emerging out of stealth
- Running high-scale pilots at domain market leaders

How can you help?

- Follow us on LinkedIn
- Give us feedback
- Join our beta program

HAYSTACK



HYPERSPACE

Let’s Connect

https://www.linkedin.com/in/ohad-levi/

ohadl@hyper-space.io