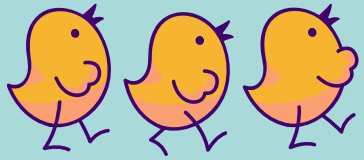



# 더 빠른 대규모 벡터 연산을 위해 FPGA를 이용해 가속을 해볼까?

김주현(메티스엑스) + 박진형(사이오닉AI)

# FPGA? 벡터 검색 가속?
# (CXL은 또 무엇?)

LangCon 2024

# Quick Intro

Language Conference

## Who?

MetisX

**CXL Computational Memory**를 개발하는 Fabless start-up

### 김주현

Co-founder, Chief Product Officer
프로세서, 시스템SW
하드웨어+소프트웨어 통합 설계
SoC, SSD, Memory, Database, AI, …

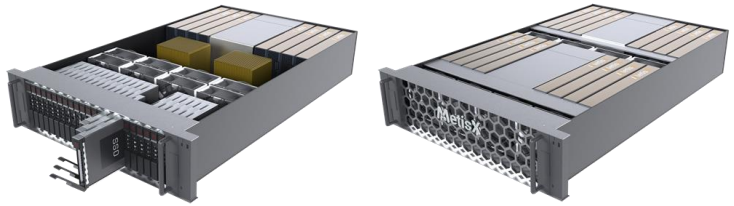
## CXL

CXL Compute Express Link

GPU/SSD를 꽂던 PCIe 슬롯에 메모리를!
**캐시 일관성 지원**
Storage처럼 Memory도 Disaggregation

### Computational Memory

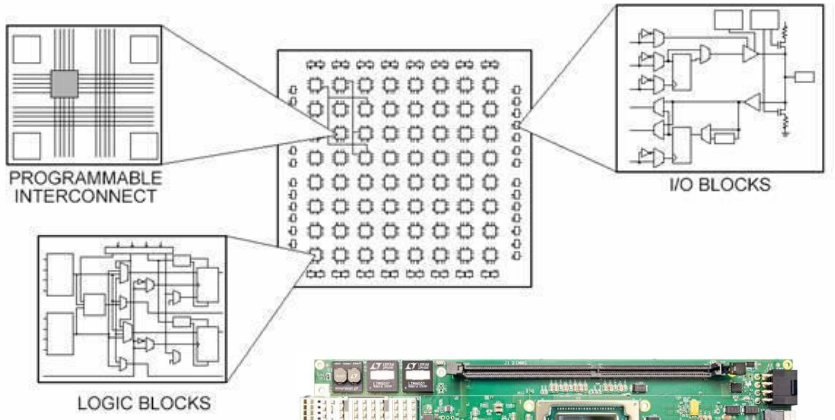
데이터 연산을 해주는 똑똑한 메모리
**Near Data Processing**
→ **벡터**/DB쿼리/ETL/…



## FPGA

### Field Programmable Gate Array

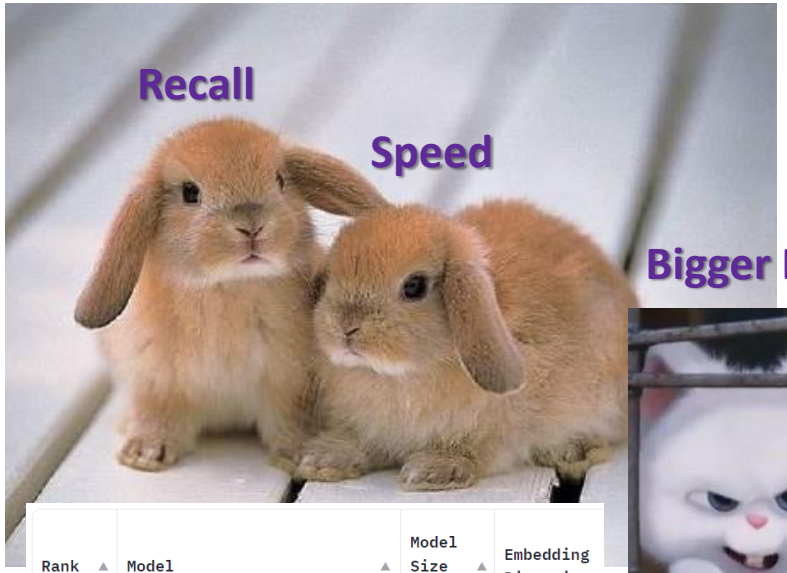
가속 하드웨어를 Programmable하게 만들어 볼 수 있는 칩




# 왜 벡터 HW 가속이 필요할까?

Language Conference

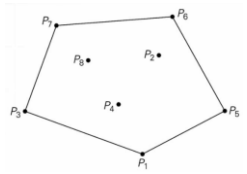
- CPU 보다 빠르고 정확하게, GPU 보다 크고 싸게!

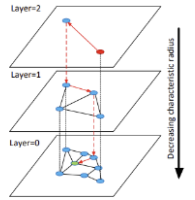



| Rank | Model | Model Size (GB) | Embedding Dimensions |
|---|---|---|---|
| 1 | SFR-Embedding-Mistral | 14.22 | 4096 |
| 2 | voyage-lite-02-instruct | | 1024 |
| 3 | GritLM-7B | 14.48 | 4096 |
| 4 | e5-mistral-7b-instruct | 14.22 | 4096 |
| 5 | GritLM-8x7B | 93.41 | 4096 |
| 6 | UAE-Large-V1 | 1.34 | 1024 |
| 7 | text-embedding-3-large | | 3072 |
| 8 | voyage-lite-01-instruct | | 1024 |
| 9 | Cohere-embed-english-v3.0 | | 1024 |
| 10 | multilingual-e5-large-instruc | 1.12 | 1024 |

Higher dimension
Larger dataset

https://huggingface.co/spaces/mteb/leaderboard







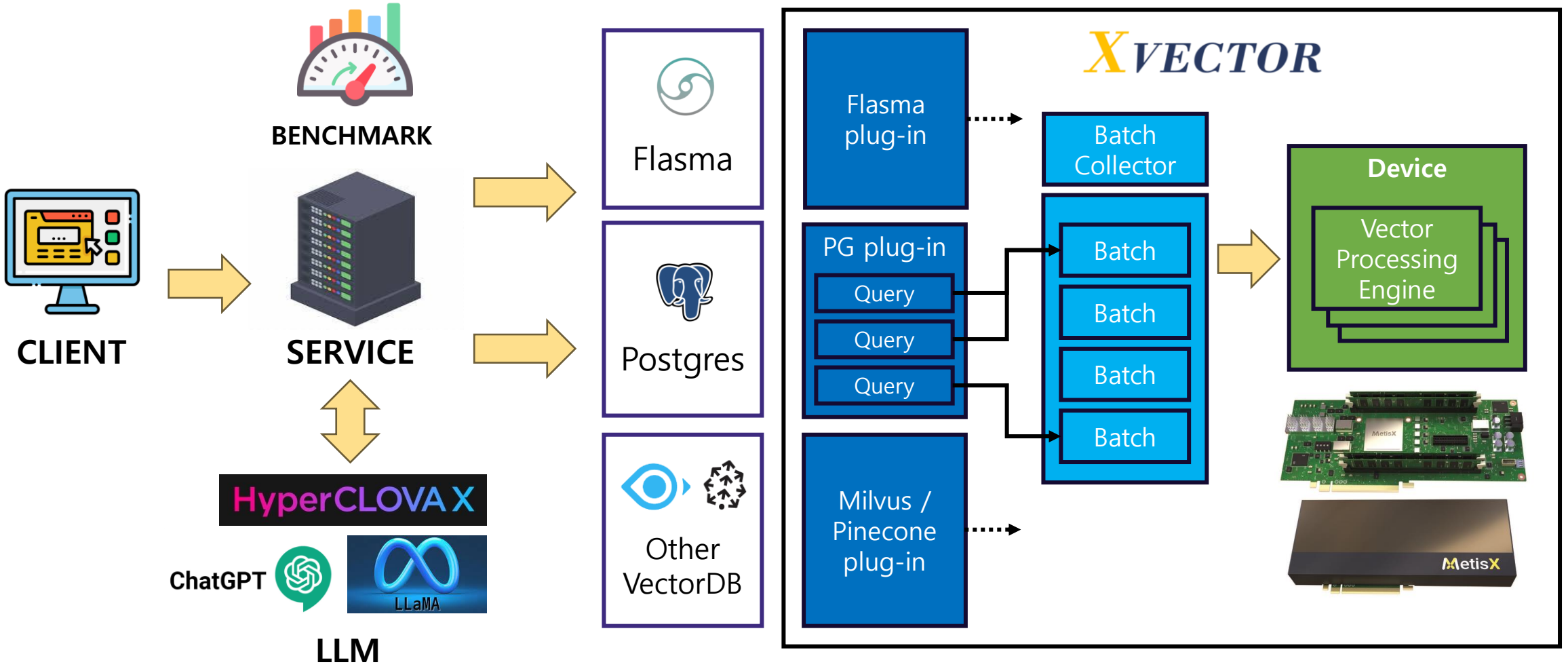
| | Recall | Speed | Size |
|---|---|---|---|
| Flat / Brute-force | 👍 | 👎 | 👎 |
| ANN (HNSW, …) | 👎 | 👍 | 👎 |
| GPU Accelerated Expensive! | 👍 | 👍 | 👎 |
| **CXL Near Memory Vector Engine** | 👍 | 👍 | 👍 |

# SionicAI + MetisX 콜라보

Language Conference

Sionic AI

MetisX

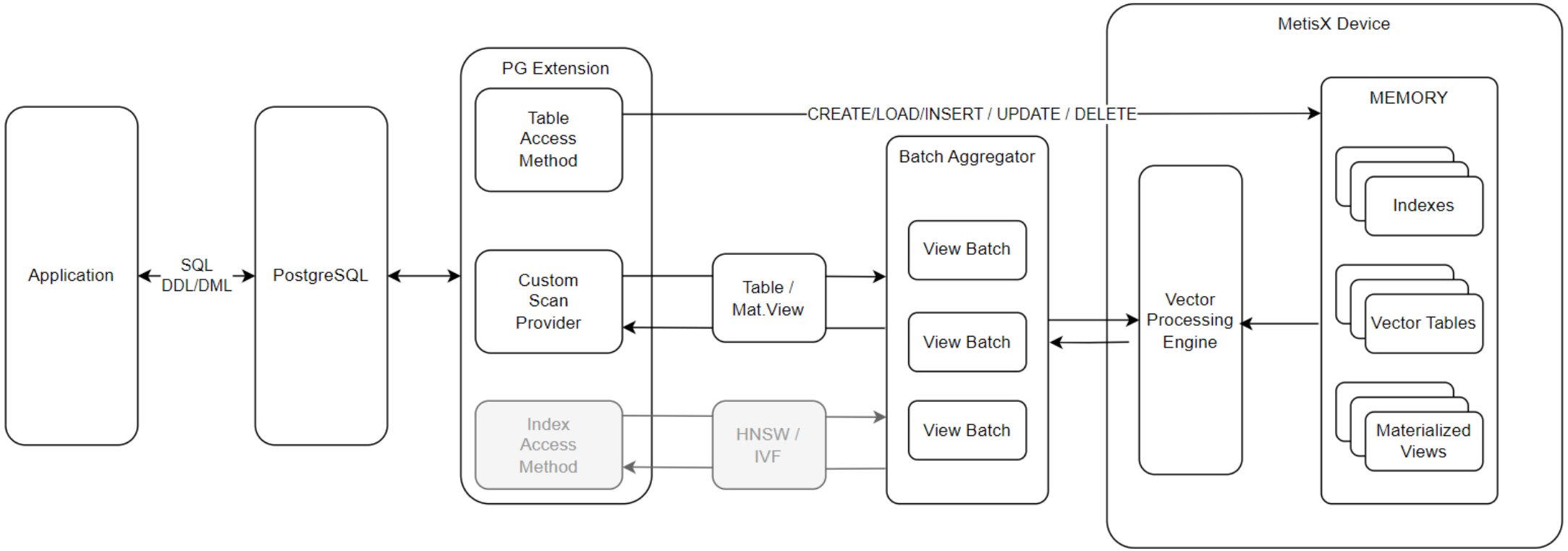


# pg-Xvector 구조



- PG Extension을 이용하여 Vector Table / Index Scan 가속
- Table,View 기반 Full Search / IVF Index 지원

기존 pg_vector 에 **pg_Xvector** plug-in 만 설치하면 적용 끝! (no schema/sql change!!)




https://www.postgresql.org/docs/current/custom-scan.html
https://www.postgresql.org/docs/current/tableam.html
https://www.postgresql.org/docs/current/indexam.html

# 벡터 쿼리 성능

Language Conference

- 400 QPS(FPGA) >> 260 QPS(3090GPU) (쿼리 당 벡터 100,000개 비교해서 Top-K를 뽑았을 때)
- 카드당 벡터 데이터베이스 크기 ~512GB x 슬롯 추가 확장
- ASIC(2025년 출시 예정)에서는 성능 10배, 용량 4배 이상 - Higher Recall & Speed, Bigger Data

**Xilinx FPGA**
**512GB**

Vector search 성능 : 40M Vector/s @ 2K dim.(fp32)

Target vector 100,000개 환산 400 Query/s (12 batch)

Vector database size : max 512GB

**10x**

4000 QPS, max 2TB

**MetisX CXL**
**Computational**
**Memory**

**NVIDIA RTX 3090**
**24GB**


number of target vectors = 100,000 https://github.com/sionic-ai/flasma
table value = RPS(request per second)
table value 0 -> Cuda Out Of Memory

| batch/dim | 100 | 300 | 512 | 768 | 1024 | 2048 | 30522 |
|---|---|---|---|---|---|---|---|
| 1 | 393 | 389 | 383 | 365 | 355 | 192 | 31 |
| 2 | 822 | 705 | 555 | 432 | 394 | 241 | 42 |
| 4 | 1420 | 968 | 684 | 505 | 459 | 260 | 0 |
| 8 | 2187 | 1242 | 836 | 605 | 506 | 0 | 0 |
| 16 | 3006 | 1442 | 936 | 655 | 0 | 0 | 0 |
| 32 | 3750 | 1579 | 1001 | 0 | 0 | 0 | 0 |
| 64 | 4184 | 1645 | 0 | 0 | 0 | 0 | 0 |
| 128 | 4424 | 0 | 0 | 0 | 0 | 0 | 0 |

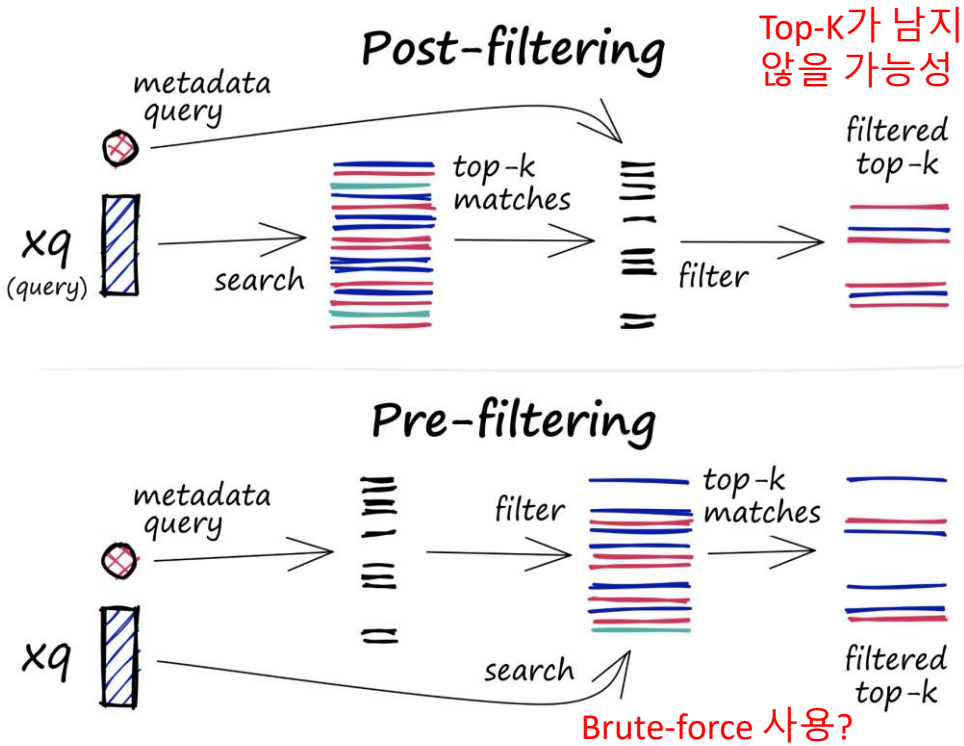
**"메모리 용량 문제"**

# RAG를 더 똑똑하게

Language Conference

그러나, 벡터 거리만으로 찾기에는 현실 세계가 너무 복잡/방대합니다.

- WHERE clause in vector search,
- `top-k where {“volume” > 14 and “department” == “finance”}`




Top-K가 남지 않을 가능성

Brute-force 사용?

Filter <-> ANN Index 상충

**“In-Query Filtering w/ metadata”**

```
index.query(queries=[xq.tolist()], top_k=3,
            filter={'lang': {'$eq': 'en'}})
```

```
{
  'context': 'Il Collegio di Ingegneria è stato istituito nel 1920,
  'metadata': {'lang': 'it', 'title': ''}
}
{
  'context': 'La KU School of Engineering è una scuola di ingegneri
  'metadata': {'lang': 'it', 'title': ''}
}
{
  'context': 'The KU School of Engineering is an ABET accredited, p
  'metadata': {'lang': 'en', 'title': 'University_of_Kansas'}
}
```

https://www.pinecone.io/learn/vector-search-filtering/