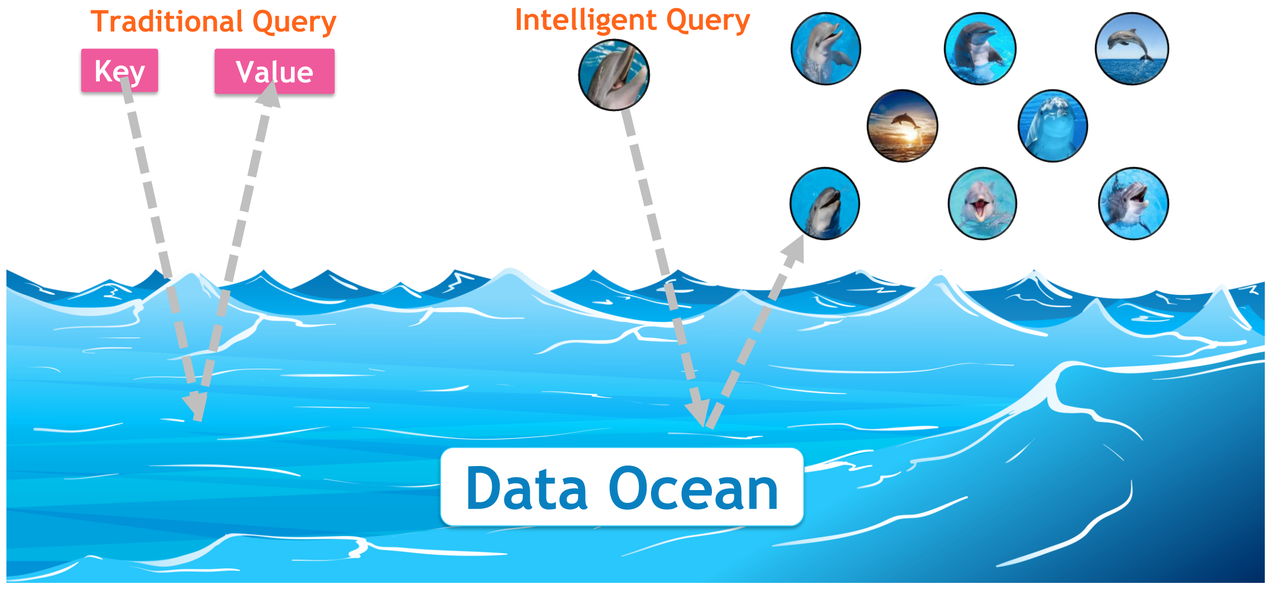
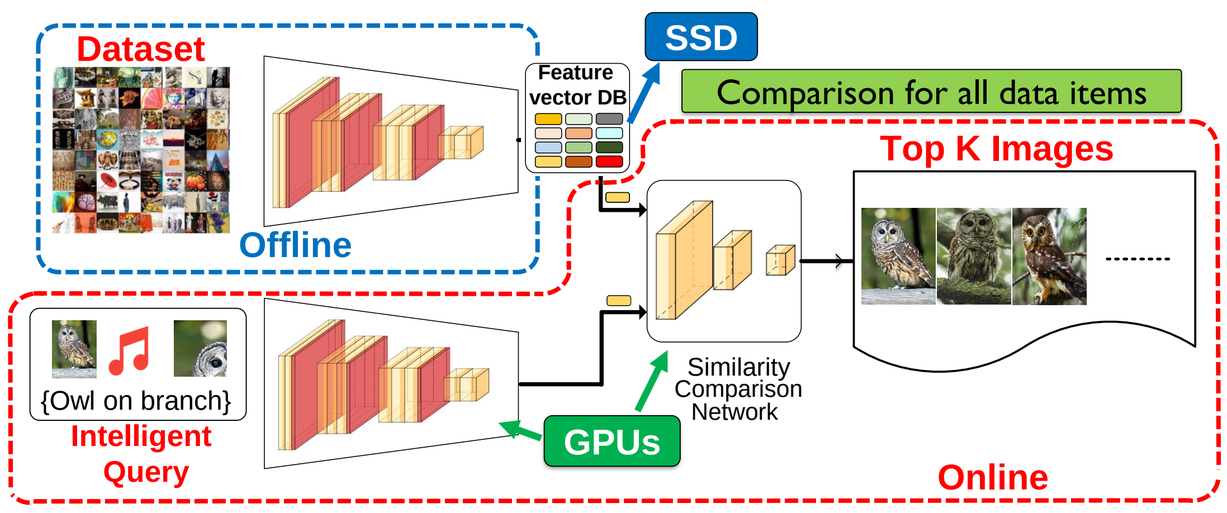
# DeepStore: In-Storage Acceleration for Intelligent Queries

Vikram Sharma Mailthody, Zaid Qureshi, Weixin Liang, Ziyan Feng, Simon Garcia de Gonzalo, Youjie Li, Hubertus Franke*, Jinjun Xiong*, Jian Huang, Wen-mei Hwu
University of Illinois at Urbana-Champaign, Urbana, IL 61801
* IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598

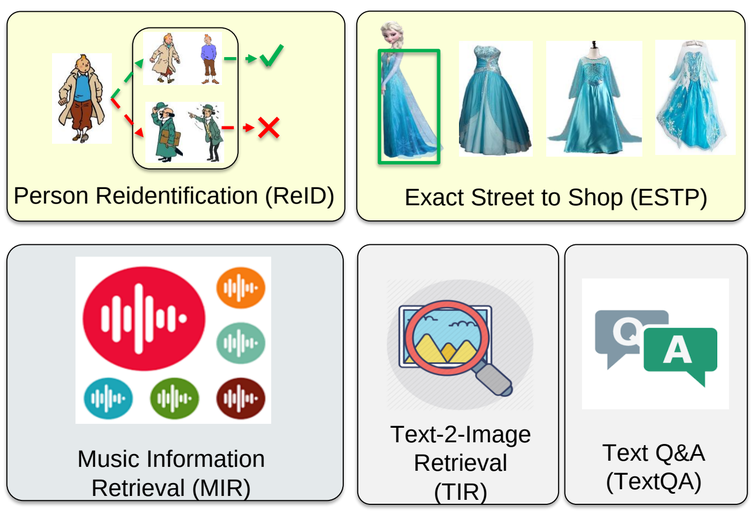
## Intelligent Queries Next Generation Workload



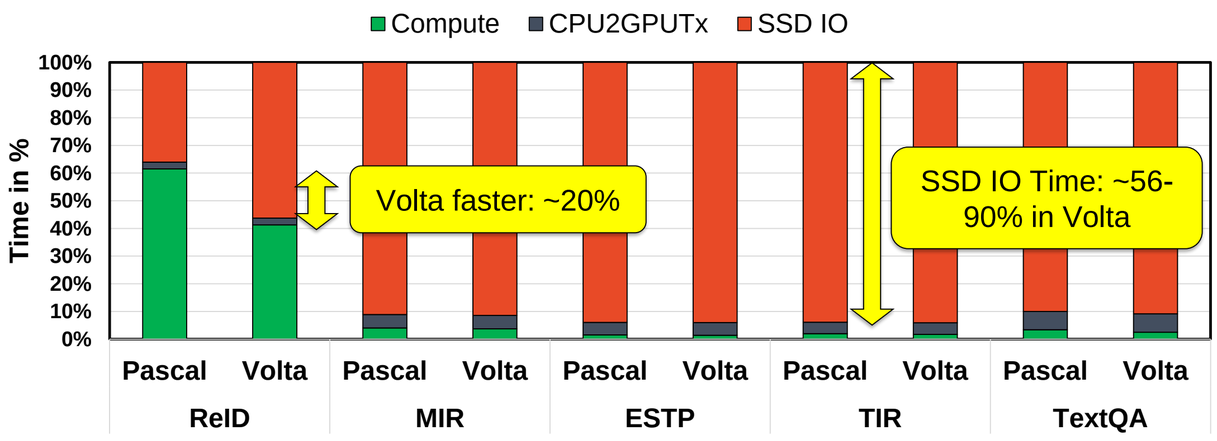

## State-of-art Intelligent Query System




## Intelligent Query Applications Studied




## Intelligent Queries System Requirement



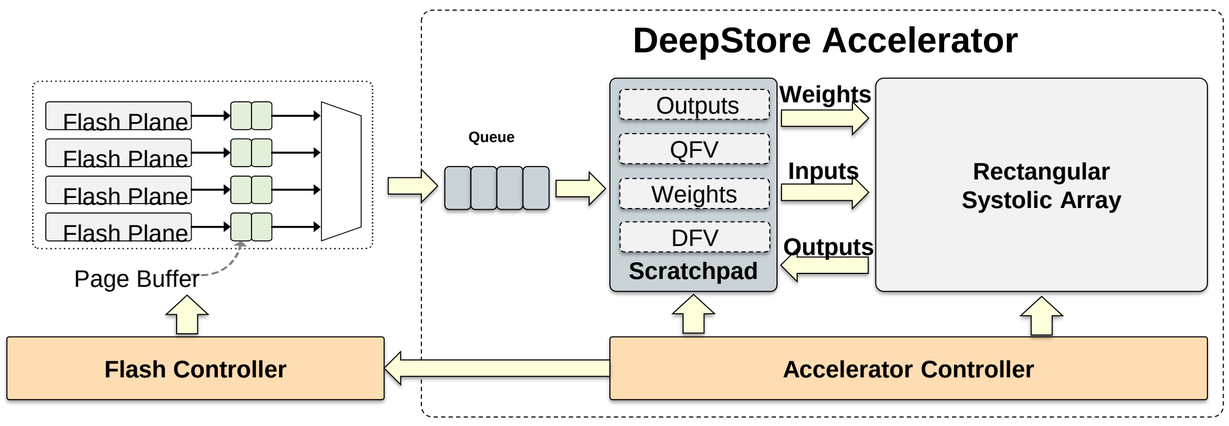

**Removal of SSD I/O Bandwidth**

**Compute in Storage**

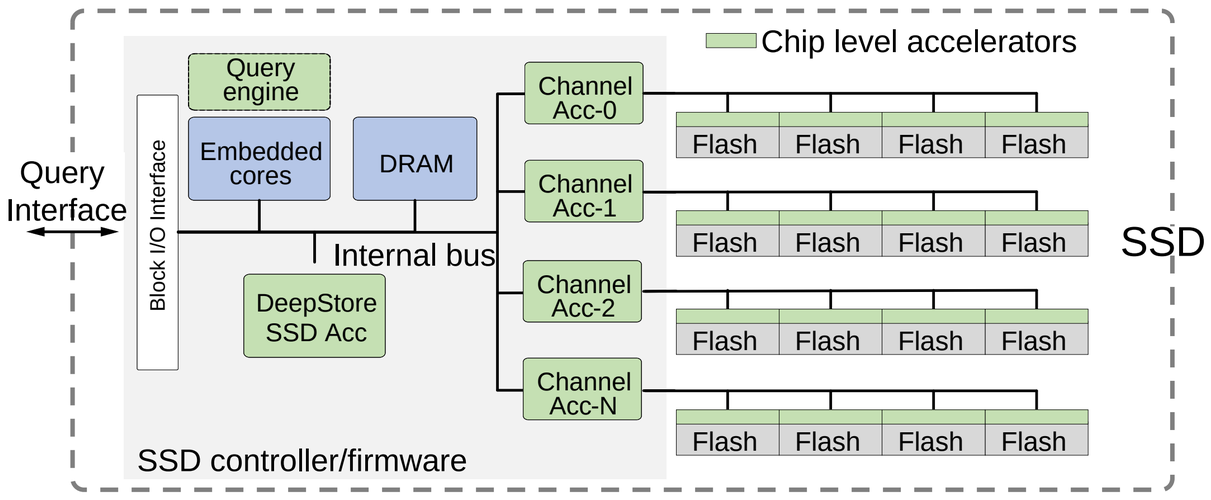
**High throughput Compute**

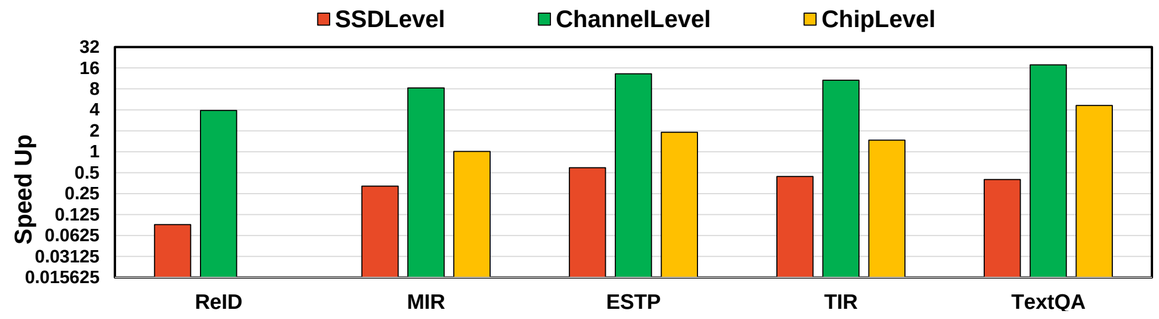
**Add DL accelerators to SSD**

## DeepStore Accelerator Design




## DeepStore Accelerator Placement



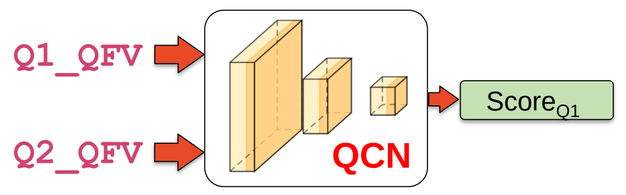

**DeepStore's channel level design performs best due to parallelism exploited and high reuse of weights**

## Query Cache

**Optimization: exploit temporal locality and semantic similarity of queries**

```
Q1. A brown dog is running in the sand
Q2. A brown dog plays at the beach
```

**Semantically similar query**

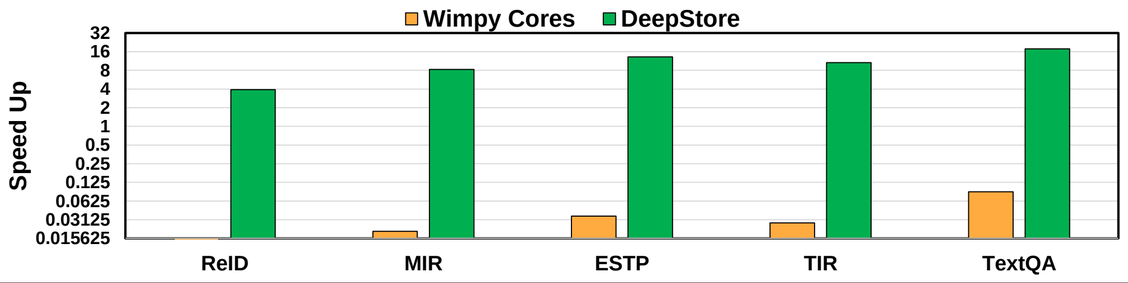



Query Cache

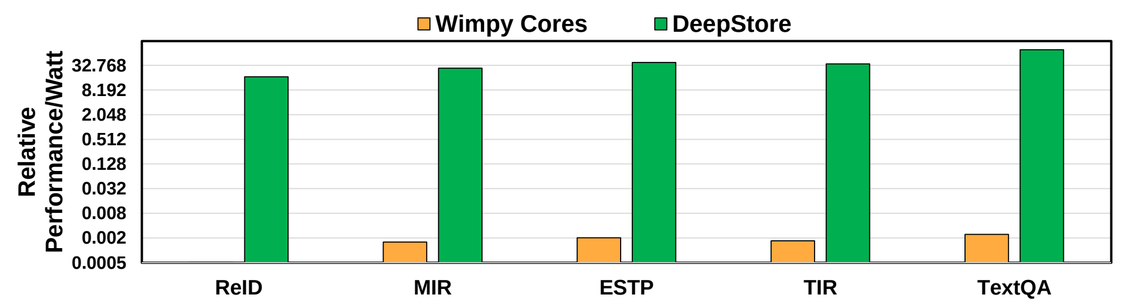
| Valid | QFV(Tag) | TopKFV |
|---|---|---|
| 1 | Q1_QFV | Top1, Top2… |
| 1 | Q8_QFV | Top1, Top2… |
| … | … | … |
| … | … | … |

**Hit: Take from query cache. Miss: Execute SCN over database**

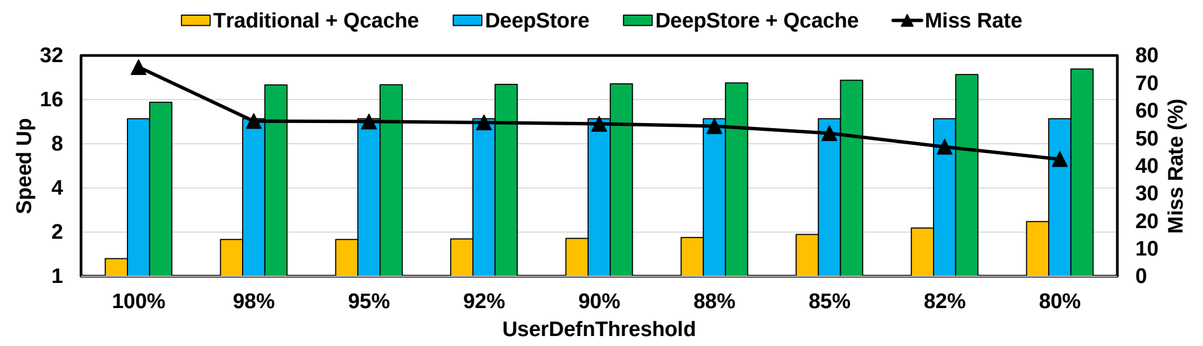
## Results




**DeepStore performs up to 17x faster due to the removal of I/O bottleneck, parallelism exploited, and high reuse of weights**




**DeepStore is up to 78x more efficient than GPU for intelligent queries**




**Query Cache in DeepStore can significantly improve performance due to DeepStore's lower mis penalty**