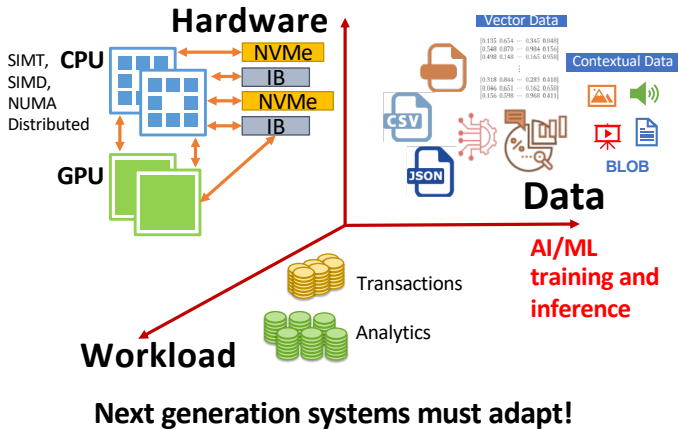


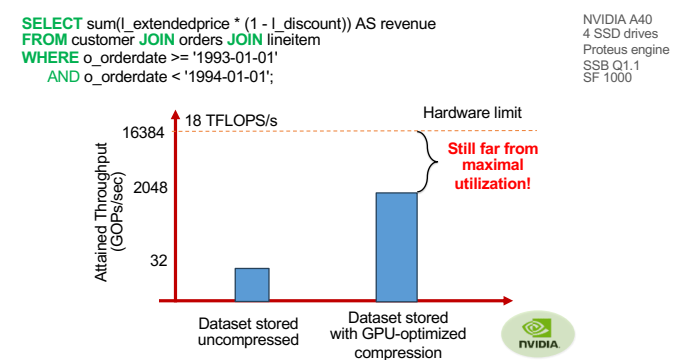
## Taming Heterogeneity in the AI-Driven Data Landscape

Data-Intensive Applications and Systems Laboratory

### Catching up with an Evolving Landscape



### Classical Analytical Workloads and GPUs



Classical workloads are interconnect bound, and struggle even with GPU optimized compression

### AI-Augmented Workloads with LLMs

```
SELECT * FROM my_photo_library WHERE AI_FILTER('I am smiling with my cat');
```

Semantic Filter

```
SELECT * FROM my_course_history AS m, this_semester_course AS c WHERE AI_FILTER('{c.course} extends the knowledge I learned from {m.course}');
```

Semantic Join

Latency (s)

7K, 5K, vLLM, Ours

Unlimited Memory

Typical LLM Requests

Semantic Operators (may span millions of LLM requests)

Request Scheduler, Database Retriever, Request Runner

LLM Inference Engine Integrated with Database

1,000,000x slower than DB operators!

### Vector Search & Semantic Similarity

```
SELECT * FROM my_course_history AS m, this_semester_course AS c WHERE AI_FILTER('{c.course} is similar to {m.course}');
```

Semantic Similarity Join

Semantic Data

Vector Embeddings

Vector Index

1,000x slower than DB operators!

GPU Optimization

Hybrid Search with Relational Filters

Integration into Relational Databases

Cost-based Query Optimization

Vector Search

### Semantically rich data processing should be efficient

### New AI-Driven Workloads to Leverage DB?

Q: Summarize the reviews of all CS courses in the last year and suggest courses with good reviews

Semantic Query → LLM → SQL → SQL → Execution on DB → Intermediate Query Result → Answer Generation

Direct NL-to-SQL is extremely challenging or not useful!

Guess, Observe, Rethink

Confident!

Instead, do Table-Augmented Generation (TAG)

Divide-and-conquer Multi-path reasoning Agentic LLM on DB

Find good reasoning paths leveraging SQLs

### High-dimensional vector search: a new bottleneck

### JIT as a Solution to Handle New Operators?

intra-operator

- Operator tuning is  $\mu$ -architecture specific
- Tune operators to memory hierarchy specifics

intra-device

- Portability clashes with specialization
- Inject target-specific info using codegen

inter-device

- Limited device inter-operability
- Encapsulate heterogeneity and balance load

Traits in Heterogeneous Servers

	control	data	
gpu $\leftrightarrow$ cpu	heterogeneity	granularity	(un)pack
router	parallelism	locality	mem-move

Encapsulate transitions in operators

Efficient execution via accelerator-level parallelism

### Fast AI Enhanced Analytics through JIT Code Generation & GPU-Acceleration

