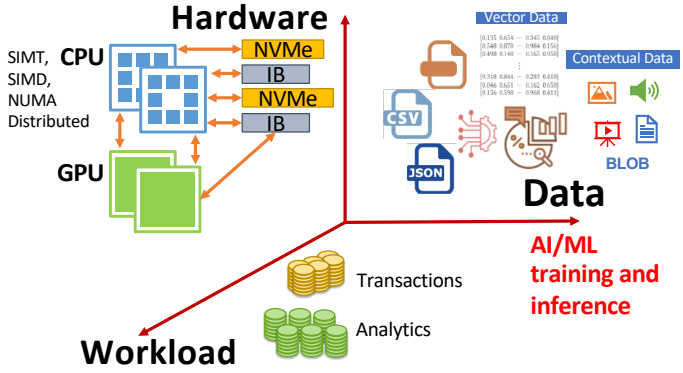
Taming Heterogeneity in the AI-Driven Data Landscape

Data-Intensive Applications and Systems Laboratory

Catching up with an Evolving Landscape

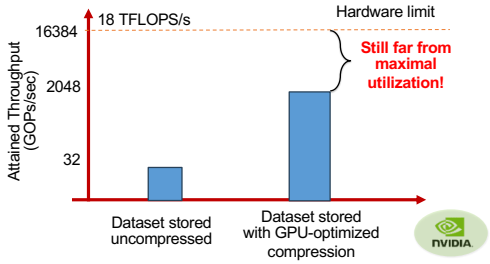


Next generation systems must adapt!

Classical Analytical Workloads and GPUs

```
SELECT sum(l_extendedprice * (1 - l_discount)) AS revenue
FROM customer JOIN orders JOIN lineitem
WHERE o_orderdate >= '1993-01-01'
  AND o_orderdate < '1994-01-01';
```

NVIDIA A40
4 SSD drives
Proteus engine
SSB Q1.1
SF 1000

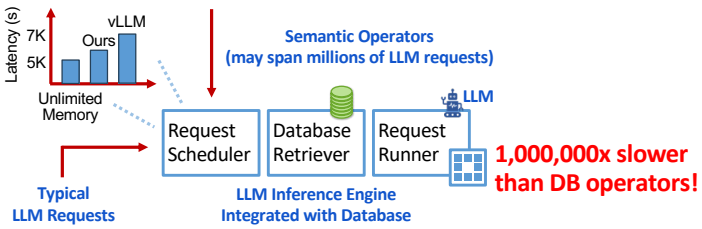


Classical workloads are interconnect bound, and struggle even with GPU optimized compression

AI-Augmented Workloads with LLMs

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

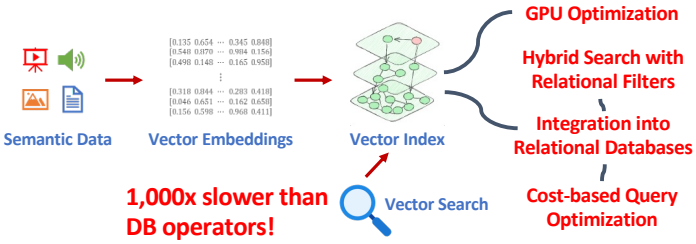
```
SELECT * FROM my_course_history AS m, this_semester_course AS c WHERE
AI_FILTER('{c.course} extends the knowledge I learned from {m.course}');
```
Semantic Join



Semantically rich data processing should be efficient

Vector Search & Semantic Similarity

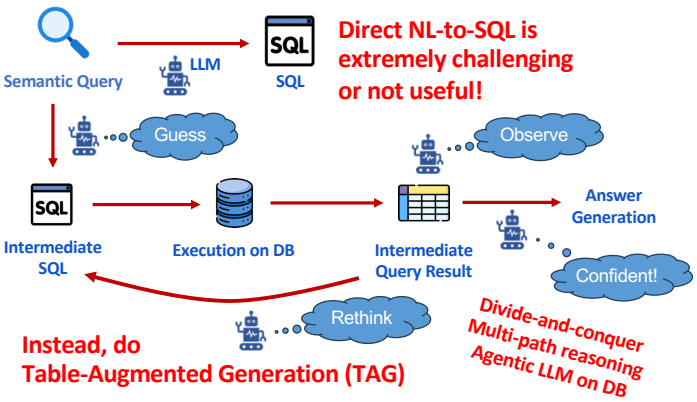
```
SELECT * FROM my_course_history AS m, this_semester_course AS c WHERE
AI_FILTER('{c.course} is similar to {m.course}');
```
Semantic Similarity Join



High-dimensional vector search: a new bottleneck

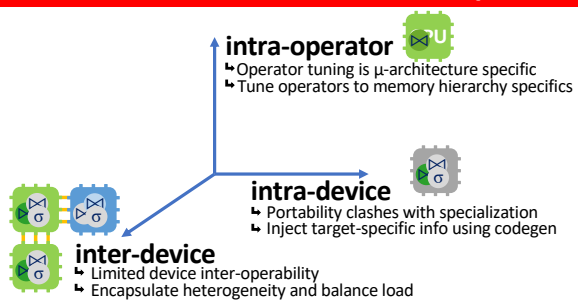
New AI-Driven Workloads to Leverage DB?

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



Find good reasoning paths leveraging SQLs

JIT as a Solution to Handle New Operators?



Traits in Heterogeneous Servers

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

Encapsulate transitions in operators

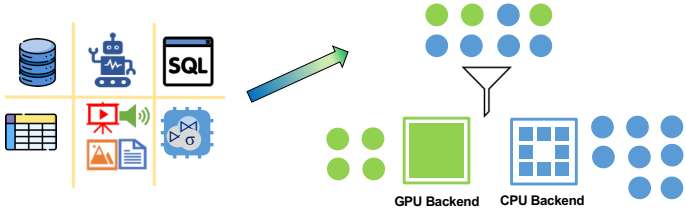
Efficient execution via accelerator-level parallelism

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

Complex Data, and AI Workloads

Code Generation

Runtime & Execution on GPU-accelerated servers



proteusdb.com