Flash Memory Summit

Bring Intelligence to Your Database Storage

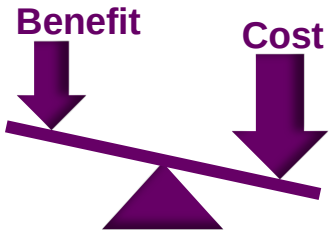
Tong Zhang

Chief Scientist, ScaleFlux Inc.

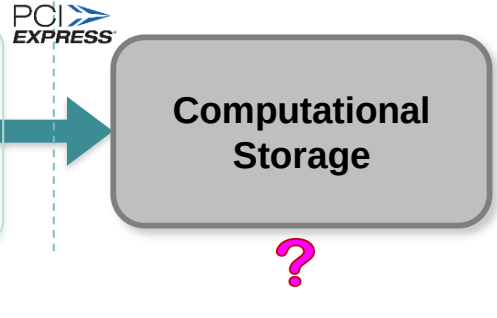
Professor, Rensselaer Polytechnic Institute (RPI)

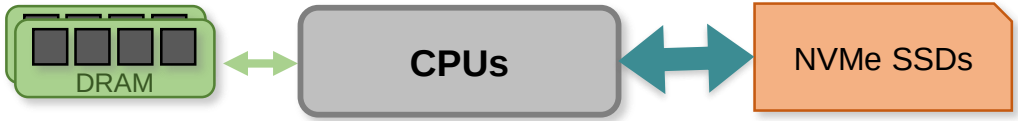
Computational Storage

- A very **simple** and **intuitive** idea
  - One option to architect a **heterogeneous computing** fabric
  - Entertained by the academia over the past 20 years: “Intelligent RAM”, “Active Disk”, …



Healthy CMOS scaling in the Good Old Days

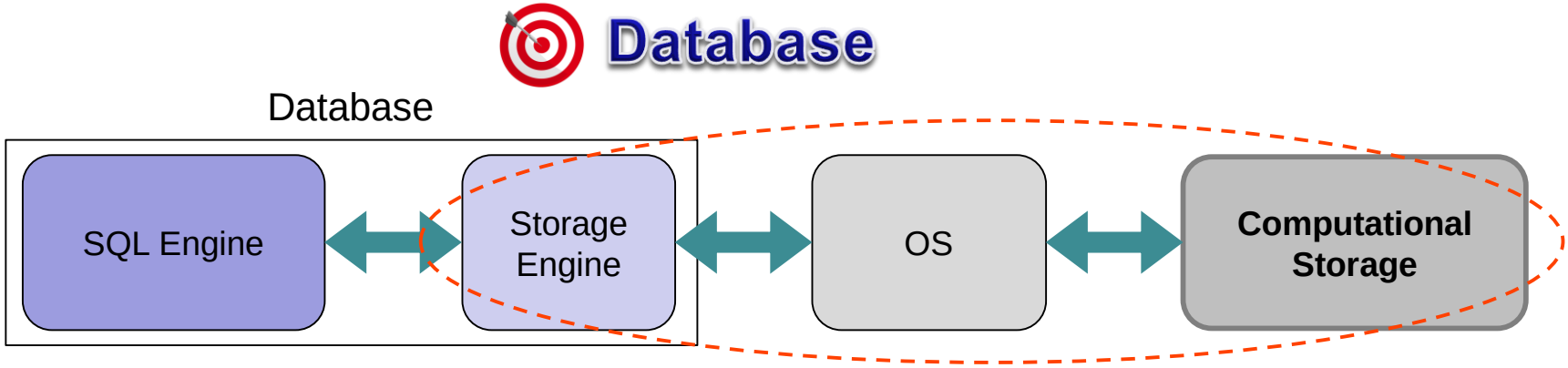




(Multi-threading, SIMD, …)

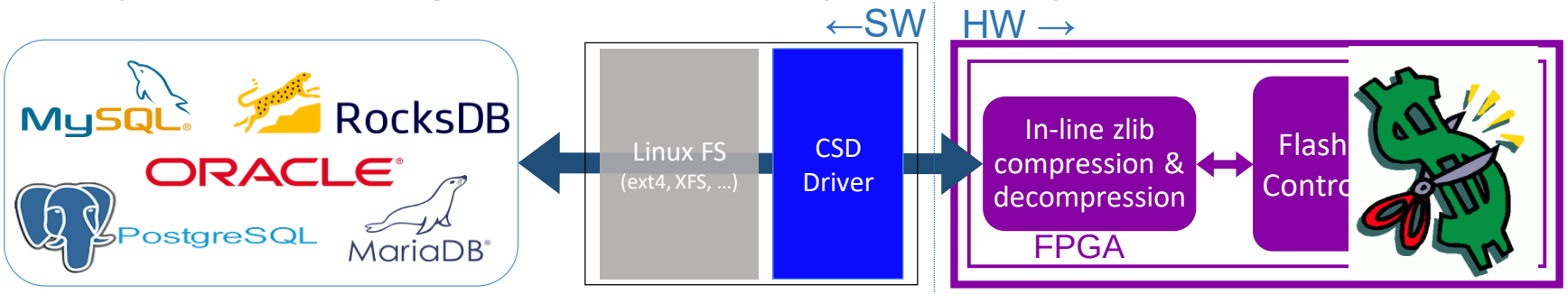
Computational Storage

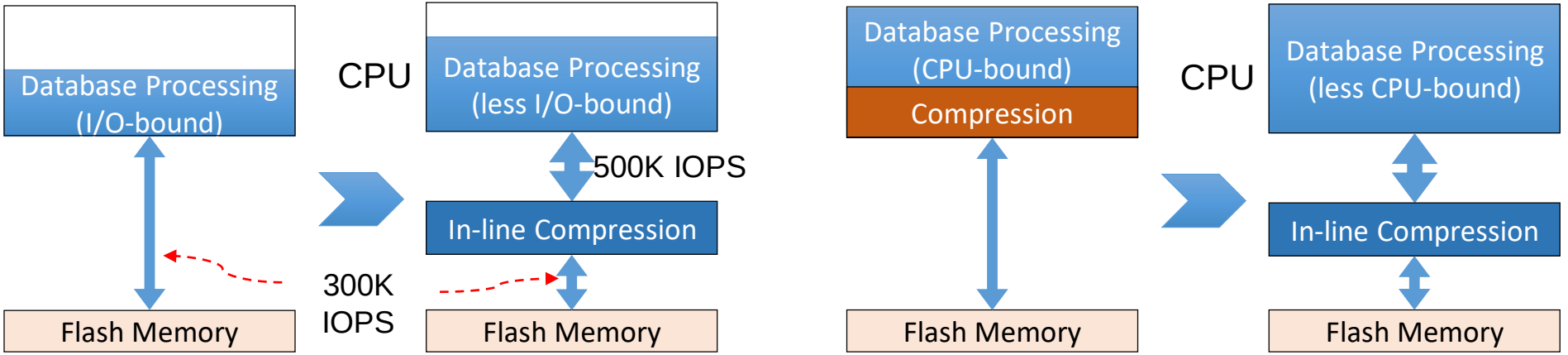
- Reduce cost:
  - Make in-storage computation as ***transparent*** as possible
  - Avoid any changes to the **core structure/algorithm** of applications
- Improve benefit:
  - Focus on mainstream, compute/IO-intensive applications



Computational Storage for Database

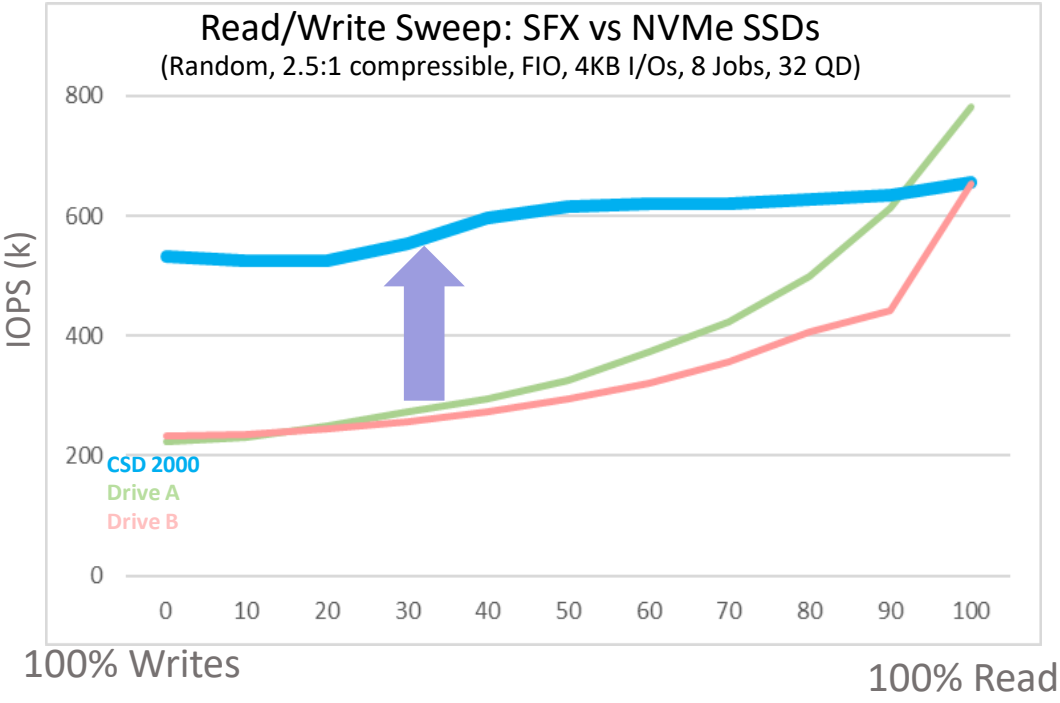
1. Computational storage with **in-line transparent compression**





Computational Storage for Database

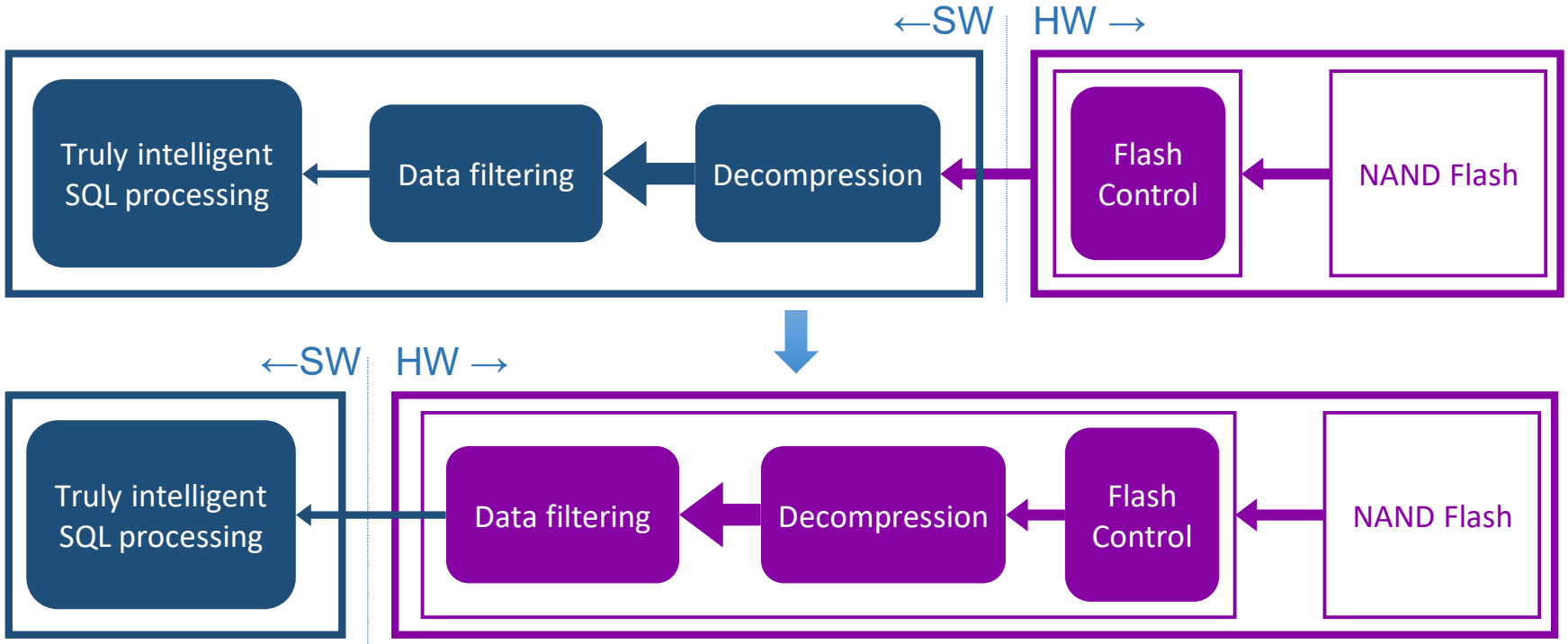
1. Computational storage with **in-line transparent compression**



- ✓ Consistent throughput across R/W mix
- ✓ No burden on CPU for running compression
- ✓ Reduced Write Amplification
  - ➢ Better endurance
- ✓ Performance leadership for key applications
  - ➢ OLTP (65-95% Reads)
  - ➢ Mixed Read/Write (50-90% Reads)
  - ➢ Write-Intensive (50%+ Writes)

Computational Storage for Database

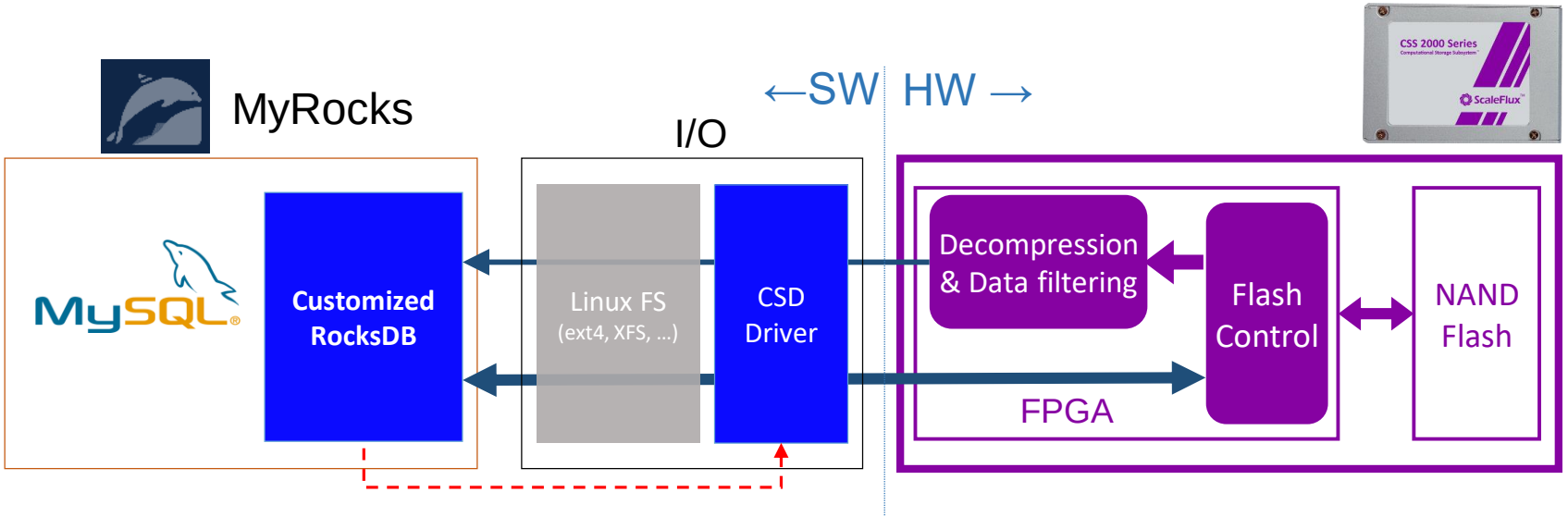
2. Computational storage with **in-line data filtering**



Computational Storage for Database

2. Computational storage with **in-line data filtering**

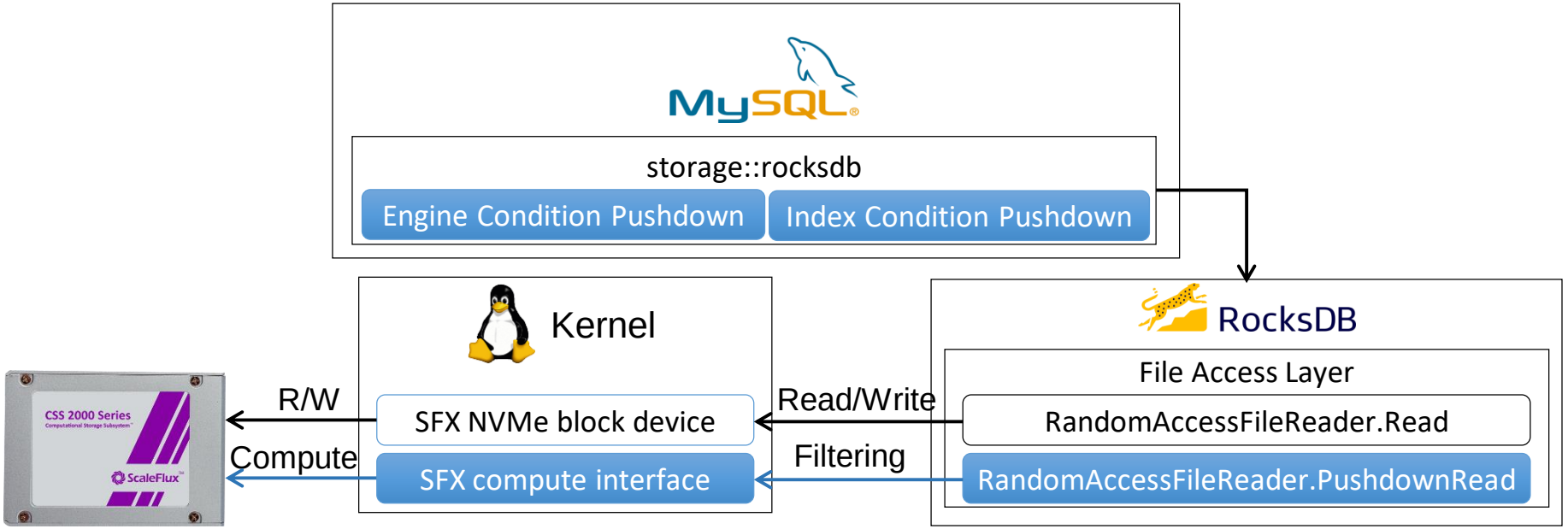
➡ Empower OLTP-oriented databases with more efficient analytics support



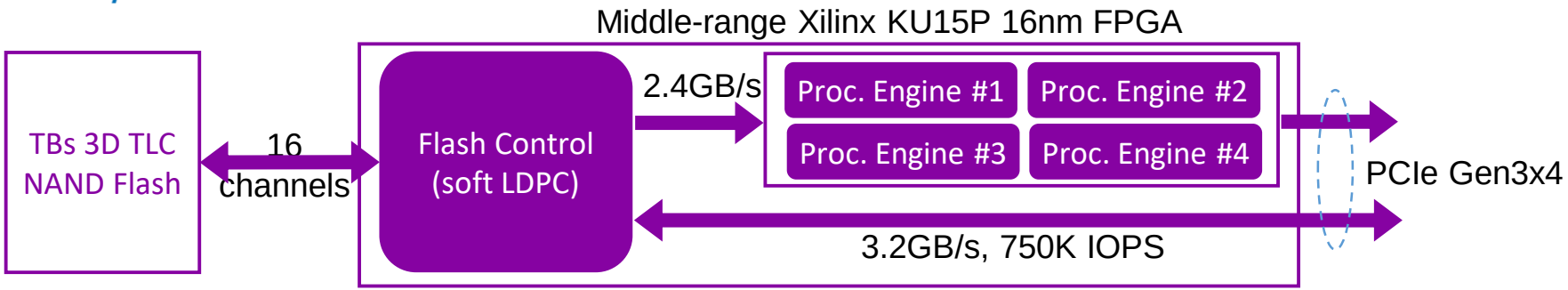
Computational Storage for Database

MyRocks appears to be an ideal first target

- ✓ MySQL built-in support of **Engine Condition Pushdown** & **Index Condition Pushdown**
- ✓ Elegant RocksDB data structure/format ➔ Simplify hardware implementation

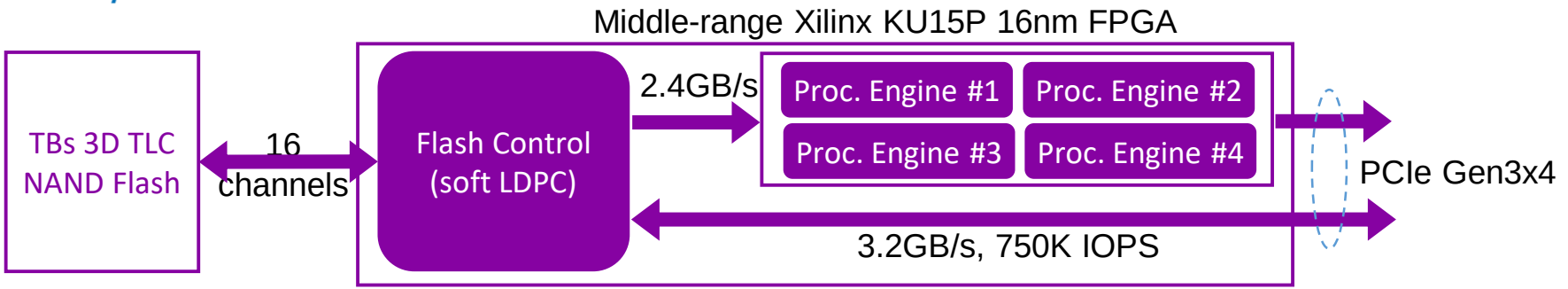


Computational Storage for Database



- Implementation: MySQL 5.6 & RocksDB 5.18
- Support in-line Snappy decompression, filter conditions: =, !=, >, <, >=, <=, !Null, Null
- Engine condition pushdown (ECP): Direct comparison between non-indexed columns and constants inside computational storage devices
- Index condition pushdown (ICP): Condition evaluation on the index inside computational storage devices

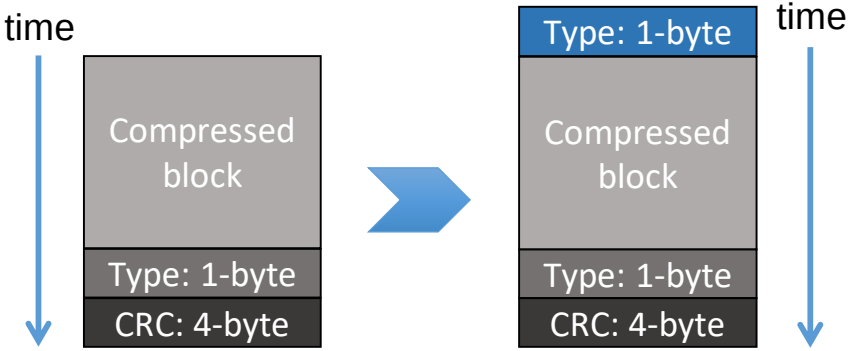
Computational Storage for Database



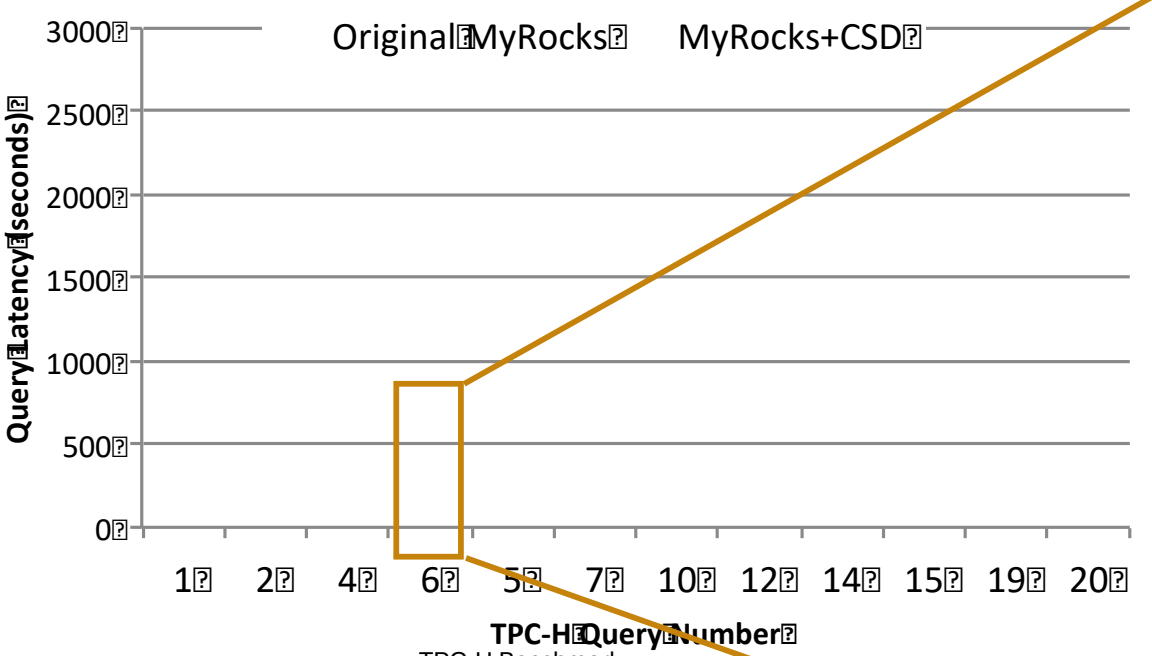
Performance benefit → Hardware utilization → Parallel & batch processing

Block-by-block iterative processing → Look-ahead batch processing

Sync call → Mixed sync/async call



Computational Storage for Database



TPC-H Benchmark
CentOS v. 7.2.1511
100GB Database

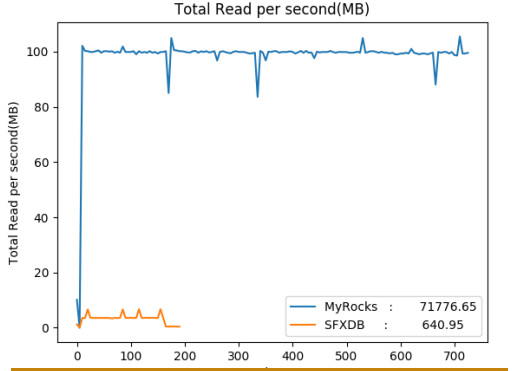
Dual
E5 2620 V3
@ 2.2GHZ

128GB
DRAM

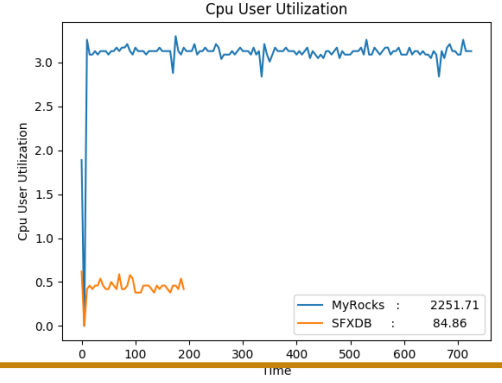
MySQL Server

MySQL v5.6

Massive Data Movement Savings

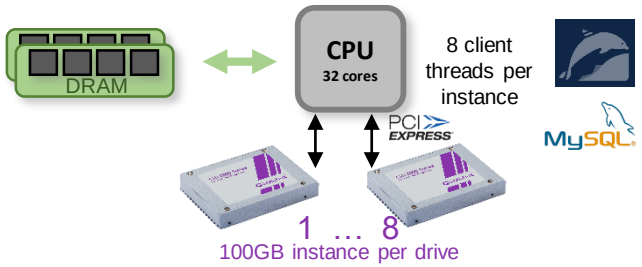


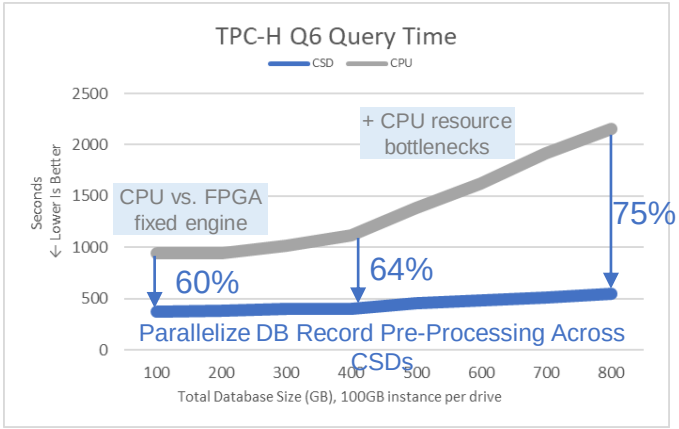
Massive CPU Processing Savings

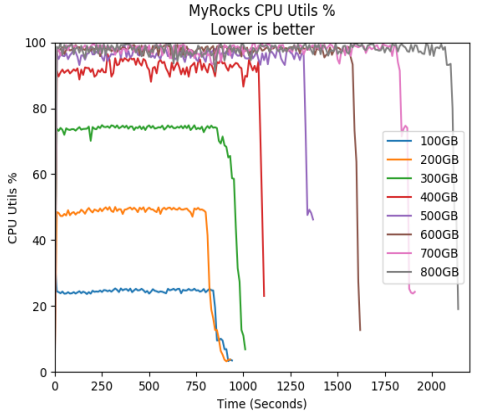


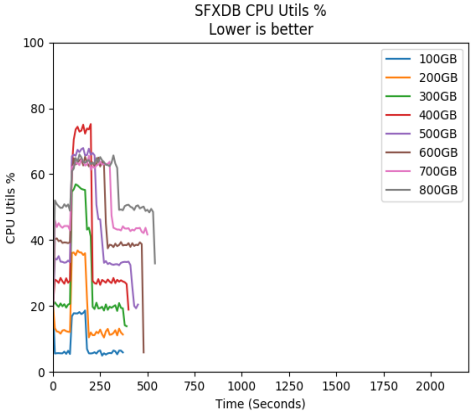
Flash Memory Summit

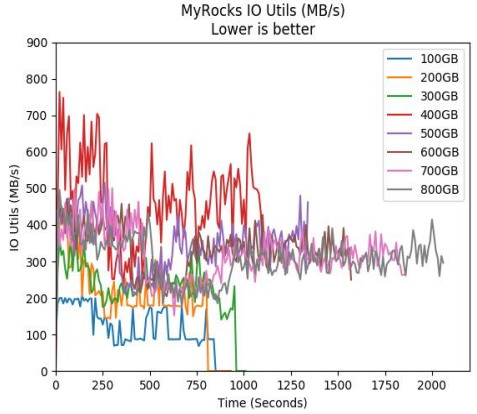
Computational Storage for Database

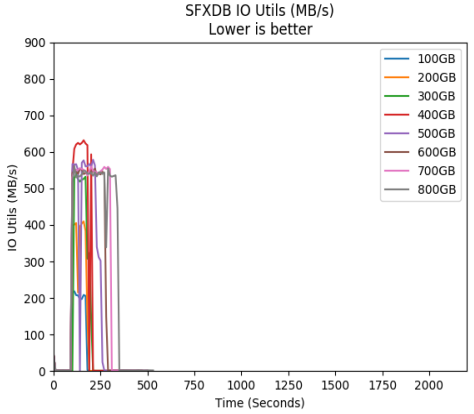






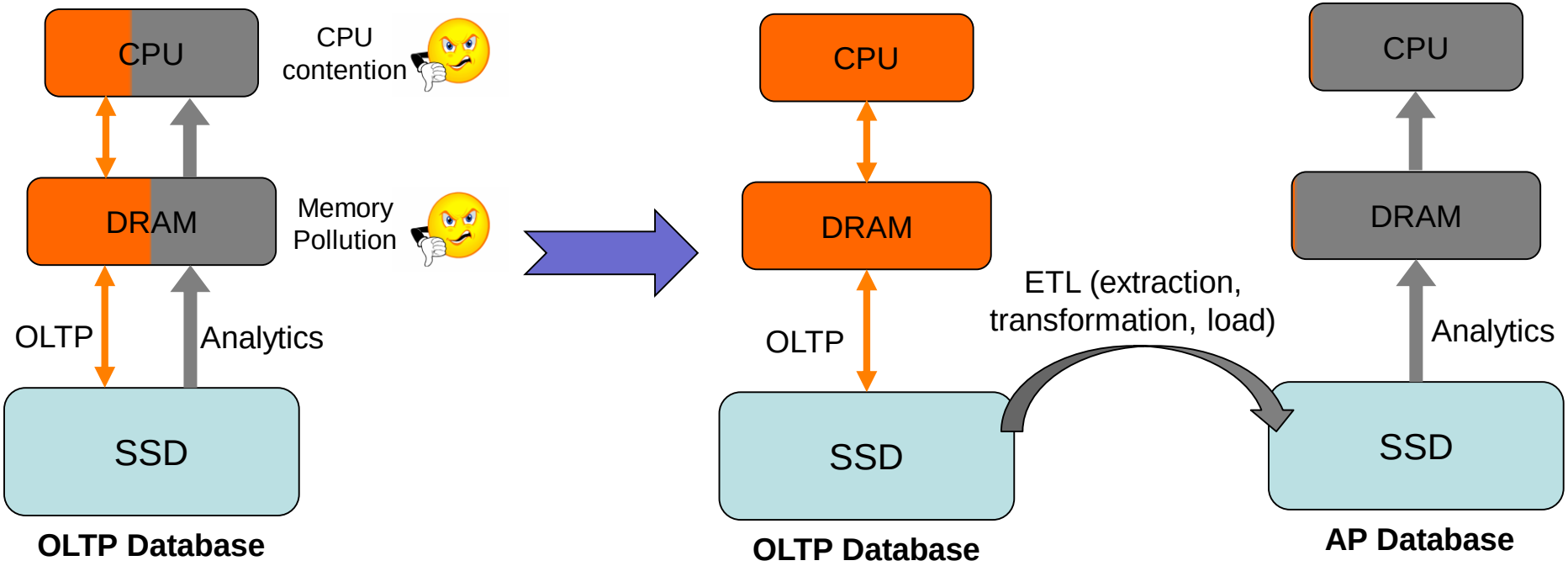






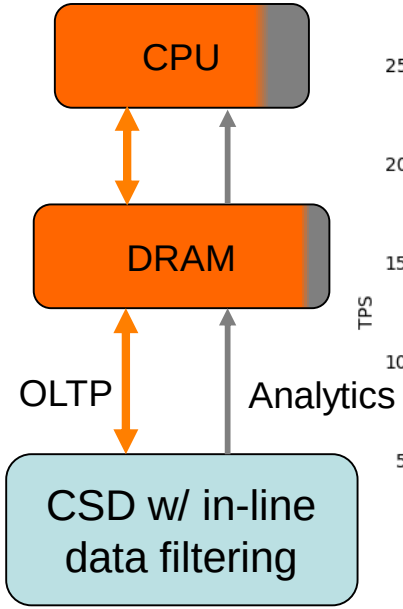
Computational Storage for Database

Empower OLTP-oriented databases with more efficient analytical processing support

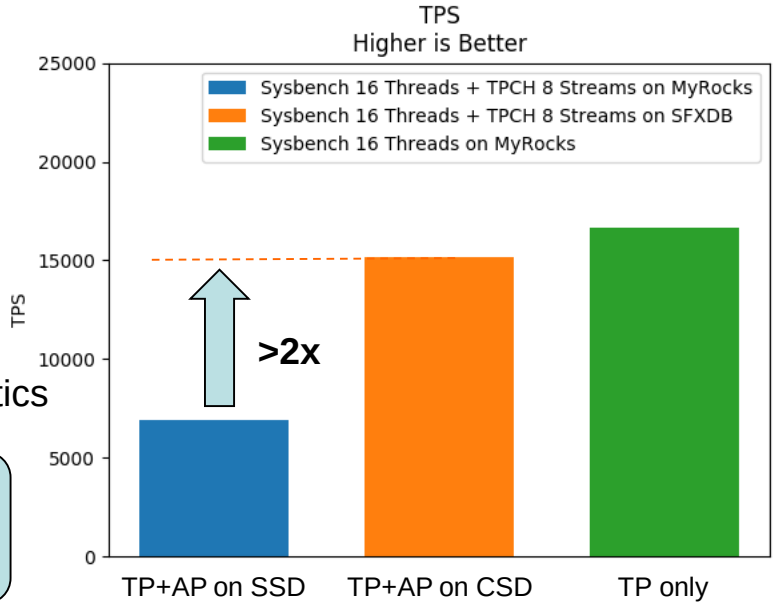


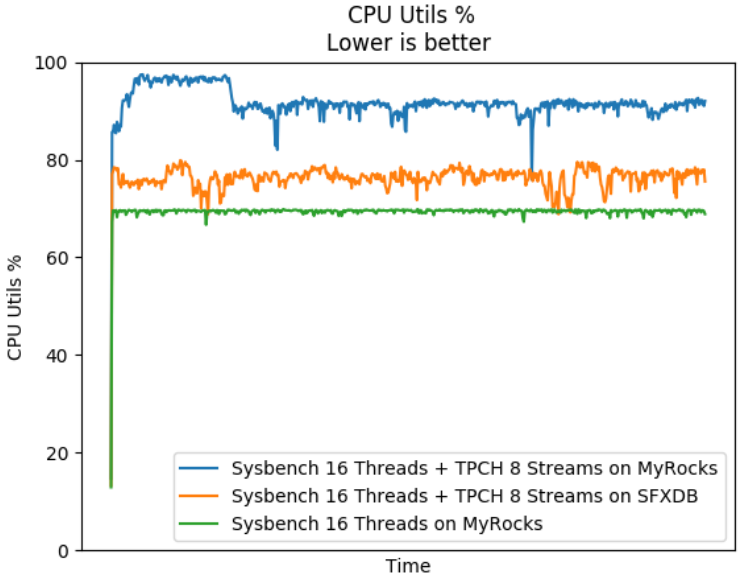
Computational Storage for Database

Much less CPU contention & memory pollution ➔ much less impact on OLTP performance



OLTP Database





Sysbench OLTP read-only + TPC-H Q6

Flash Memory Summit

Summary & Call to Action

- **Computational Storage is Coming finally!**
- **Database: One ideal target of computational storage**
  - General-purpose: **In-line transparent compression**
  - Database-specific: **In-line predicate pushdown** (successful demonstration with MyRocks on computational storage)
- **Very large design space to be explored**
  - Cross-layer innovation across software and hardware
  - Close collaboration across industry sectors

ScaleFlux™

Computational Storage for Data-Driven Applications