Flash Memory Summit

Hybrid Transactional/Analytical Processing over Computational Storage Drives

Kecheng Huang*, Zhaoyan Shen**, Zili Shao*, Tong Zhang†, Feng Chen‡

* The Chinese University of Hong Kong ** Shandong University

† ScaleFlux Inc. ‡ Louisiana State University

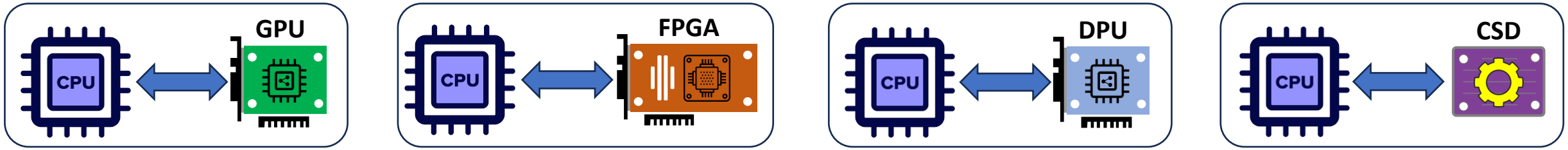
Presenter: Tong Zhang, ScaleFlux Inc.

Innovate Database in Heterogeneous Computing Era

Flash Memory Summit

. . . ORACLE MySQL PostgreSQL MariaDB snowflake . . .

influxdb mongoDB Cassandra Microsoft SQL Server AEROSPIKE



Off-load certain database processing tasks from CPU

❌ Significant changes to databases

❌ Questionable ROI

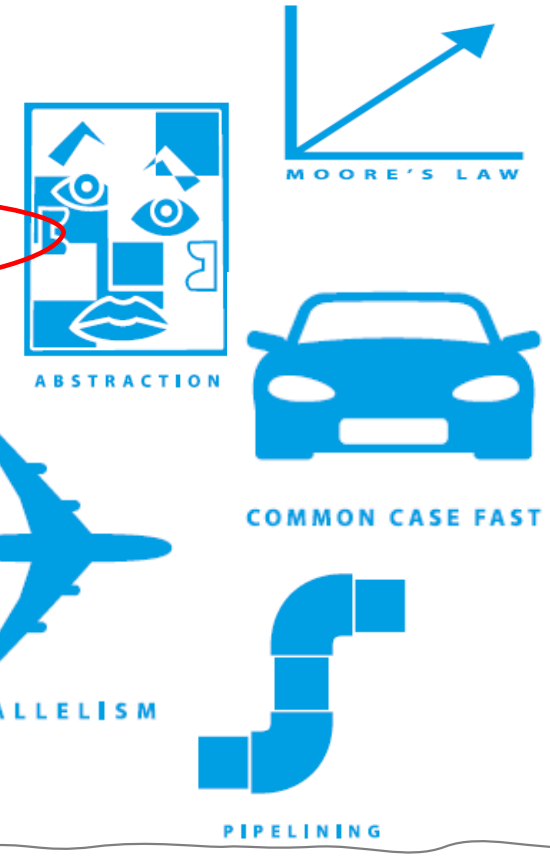
❌ Hardware vendor lock-in

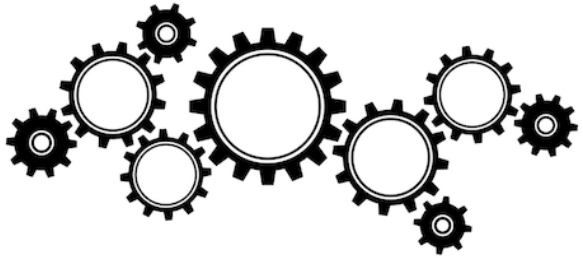
Innovate Database in Heterogeneous Computing Era

Eight Great Ideas in Computer Architecture (Patterson & Hennessy)

1. Design for ***Moore's Law***
2. Use ***abstraction*** to simplify design
3. Make the ***common case fast***
4. Performance *via* ***parallelism***
5. Performance *via* ***pipelining***

...

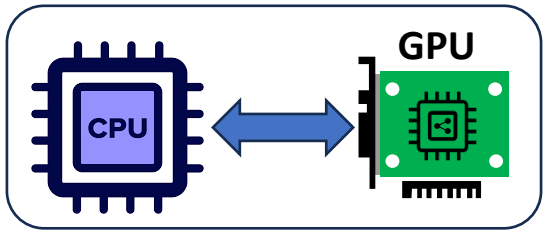


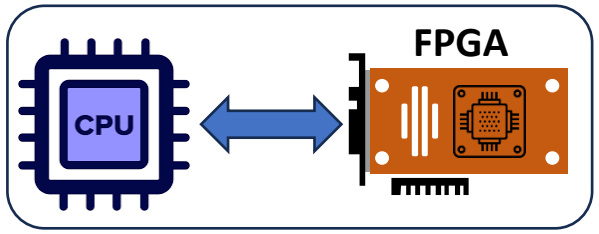


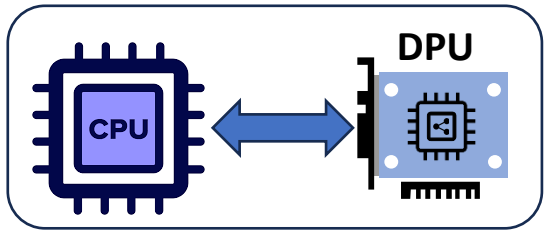
Computation offloading in heterogenous computing

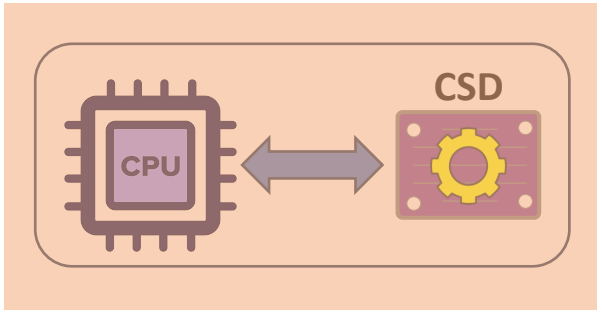
Breaks the principle of abstraction

Innovate Database in Heterogeneous Computing Era





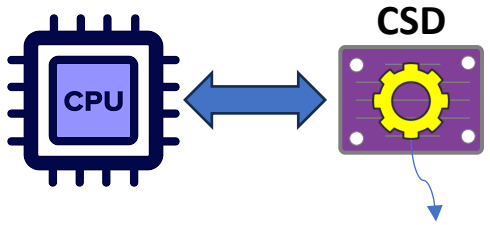




Off-load certain database processing tasks from CPU

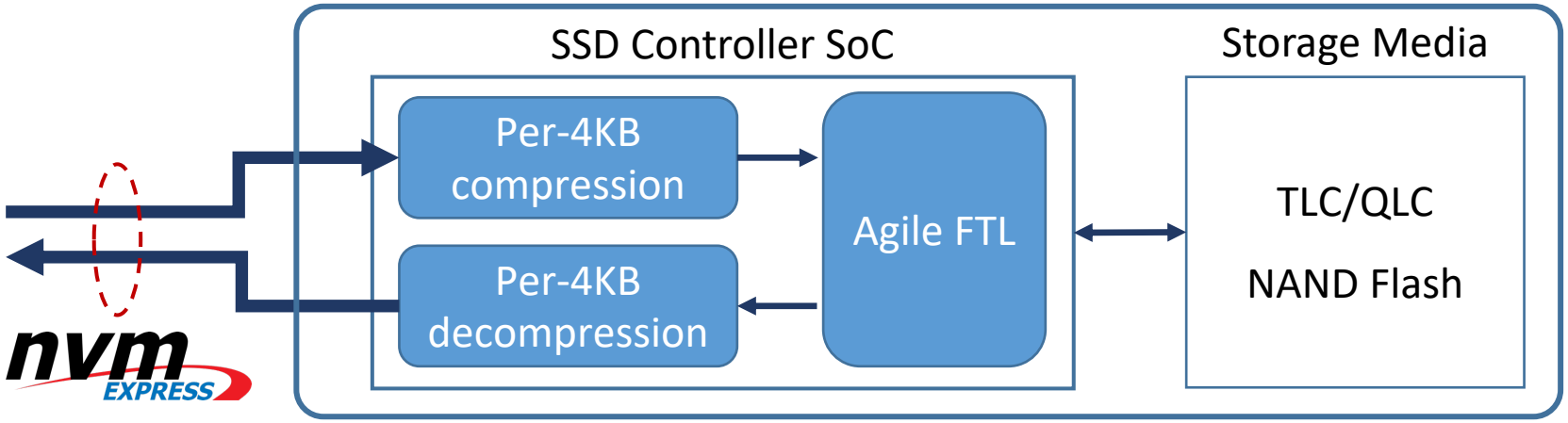
Innovate database in heterogeneous computing era **without** computation off-loading!

Drop the Computation Off-loading Mindset!



Native in-storage functions

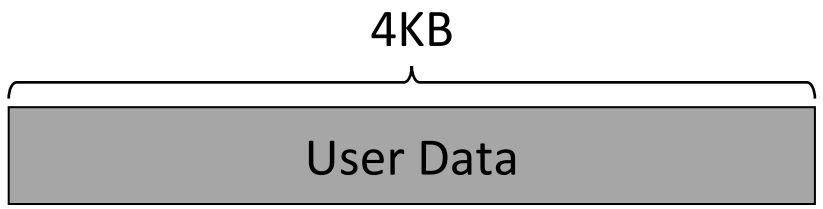
- Transactional memory/storage
- Performance profiling and counting
- Malware detection
- Snapshot
- Security
- Continuous data protection
- Compression
- Data deduplication
- Data transcoding & transformation
- . . .



- ✓ 100% compliant with NVMe
- ✓ Zero changes to I/O stack
- ✓ Transparently reduce cost & improve IOPS

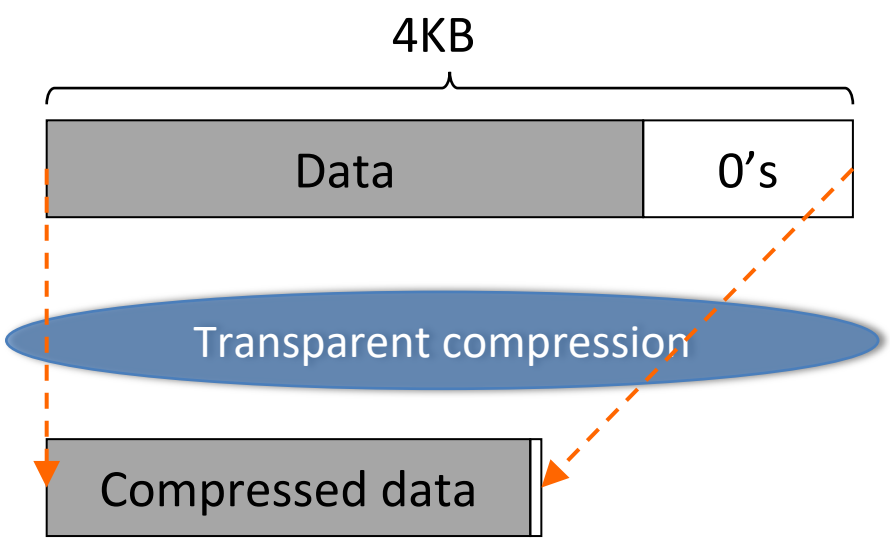
CSD with In-line Transparent Compression

Where the Innovation Potential Come From?

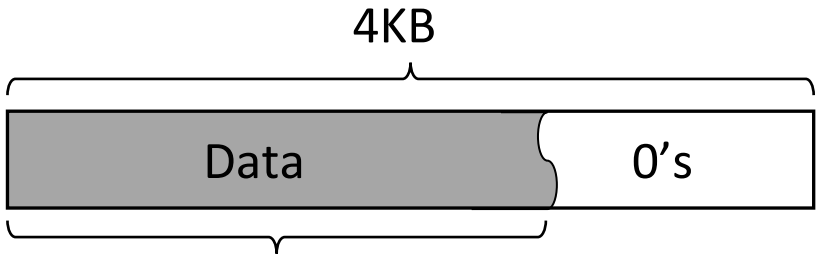


Fixed-size block I/O

Data Management SW



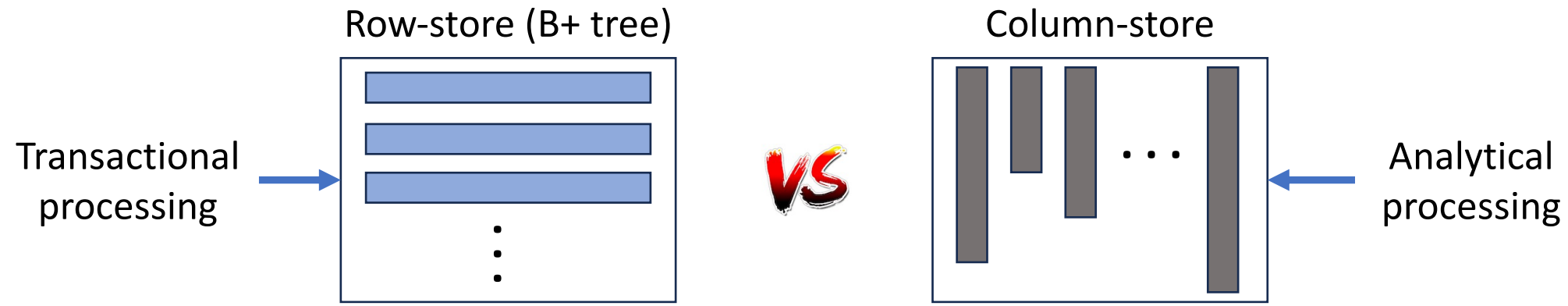
Unnecessary to completely fill each 4KB sector with user data

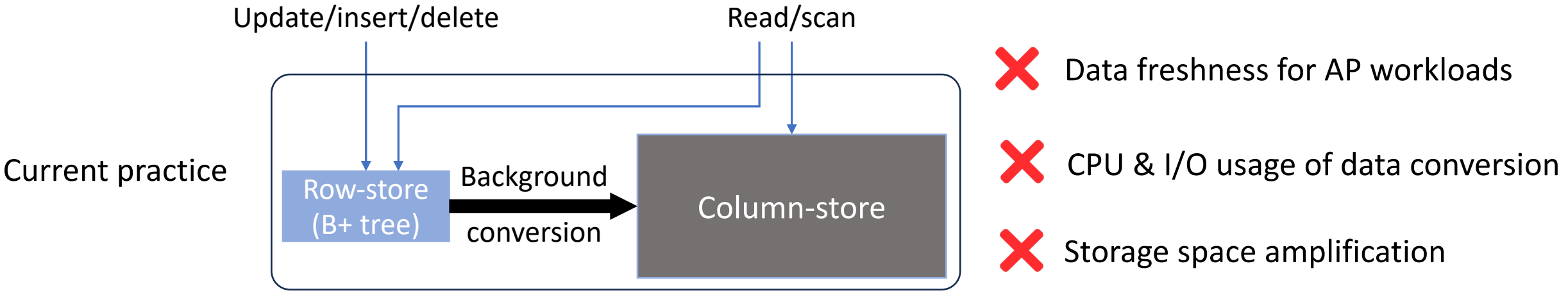


Arbitrary data size

Virtually variable-size block I/O

Innovate Database over SSDs w/ Transparent Compression



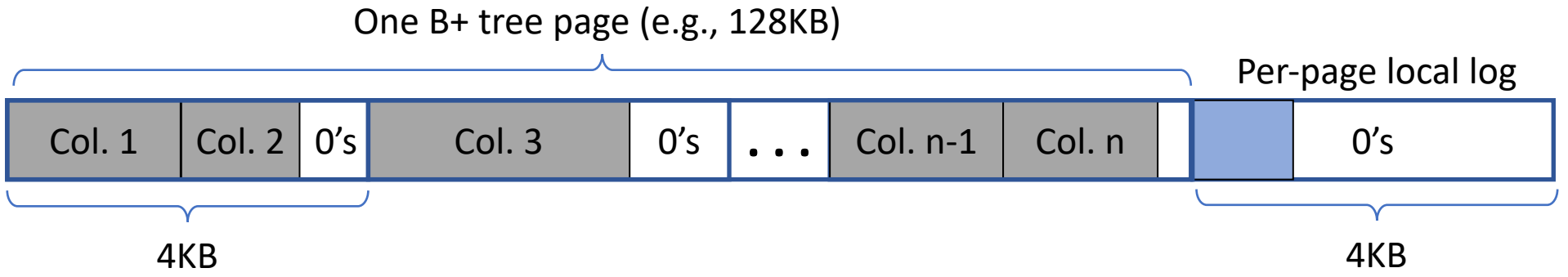


HTAP (Hybrid Transactional/Analytical Processing)

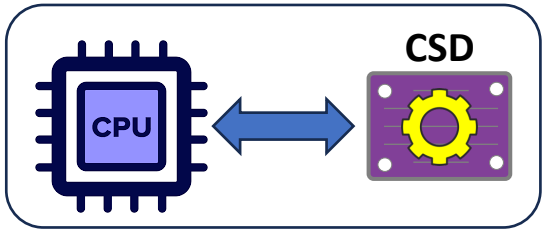
A **single B+ tree** for both TP and AP

Large B+ tree page size + Intra-page 4KB-aligned column-store

- High-performance AP over B+ tree
- Worse B+ tree write amplification



Conclusion



Explicit computation off-loading → Innovate database **without** computation off-loading

- ❑ Innovate database over 100% NVMe-compliant CSD with built-in transparent compression
- ❑ A single B+ tree to serve HTAP (hybrid transactional/analytical processing)
  - ✓ Most effective **single data structure** for HTAP enabled by CSD with built-in transparent compression