Toward Disaggregated Data Systems in the AI Hardware Era

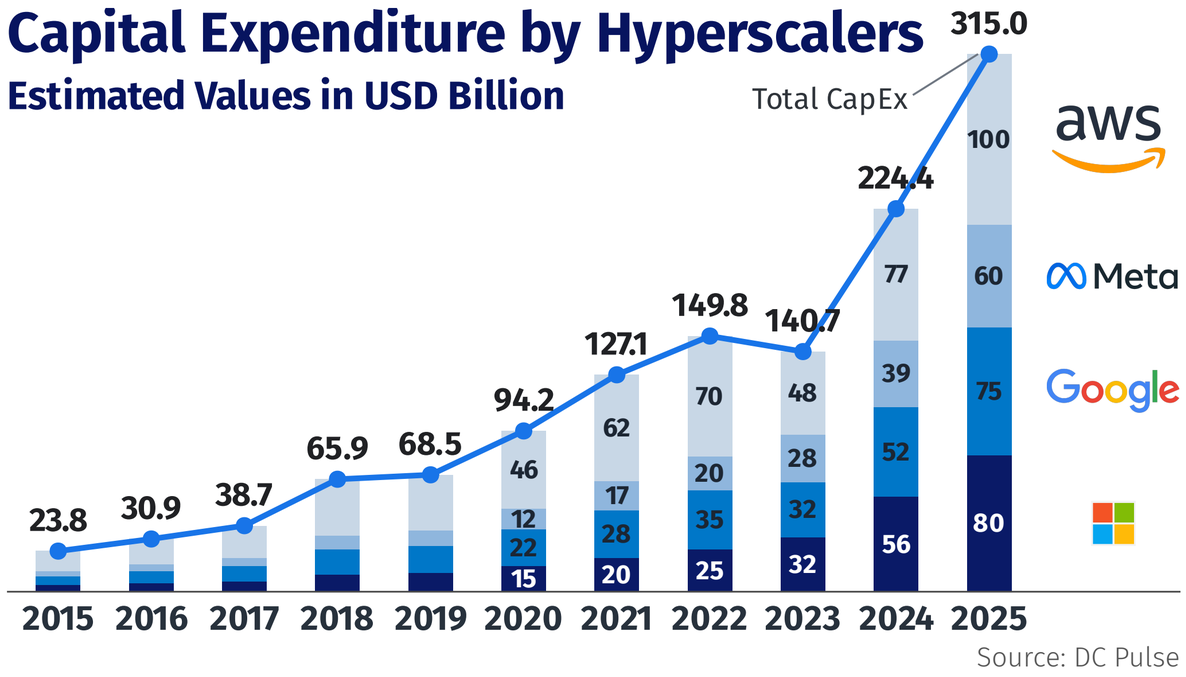
Jigao Luo

Technische Universität Darmstadt

AI Investment Shapes Modern & Future HW

Capital Expenditure by Hyperscalers

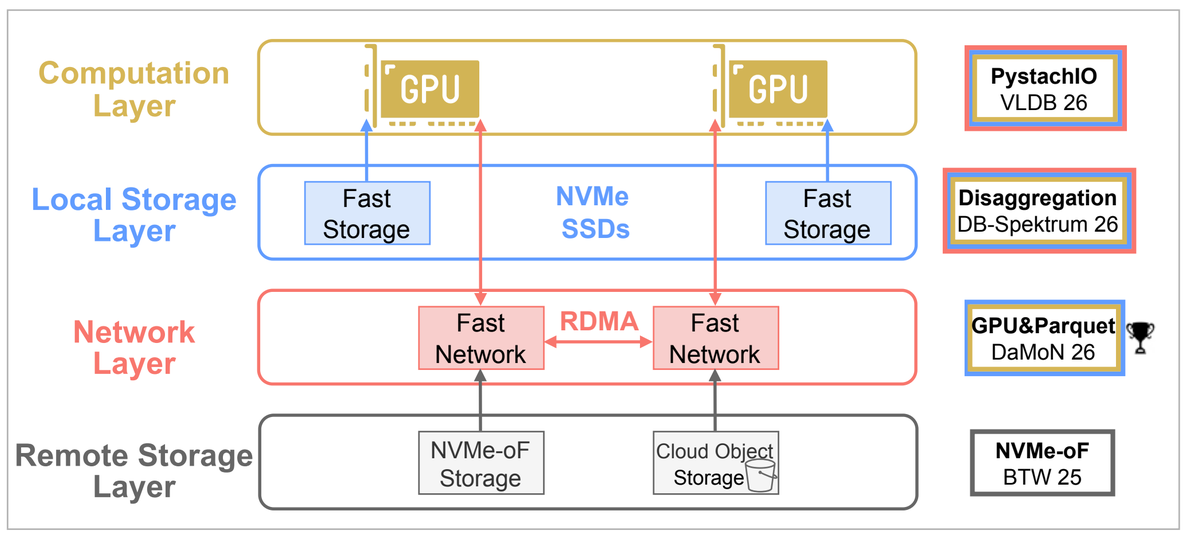
Estimated Values in USD Billion



Source: DC Pulse

- CPU performance plateau → outside the AI spotlight
- Data centers shifting toward accelerators, e.g., GPUs
- **How do we ride the AI investment wave for DB?**

Need for Re-Design & Contributions

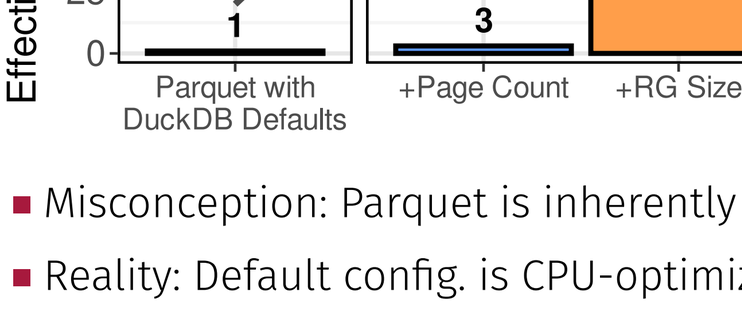


- Issue: Existing GPU DBs remain single-node only
- **Research Question: How to build a GPU DB for disaggregation?**
- Challenges from fast networks and fast storage
- Focus shifting from computation to I/O

GPU Query Engine: *PystachIO* Over Fast Networks and Fast Storage

Query Processing:

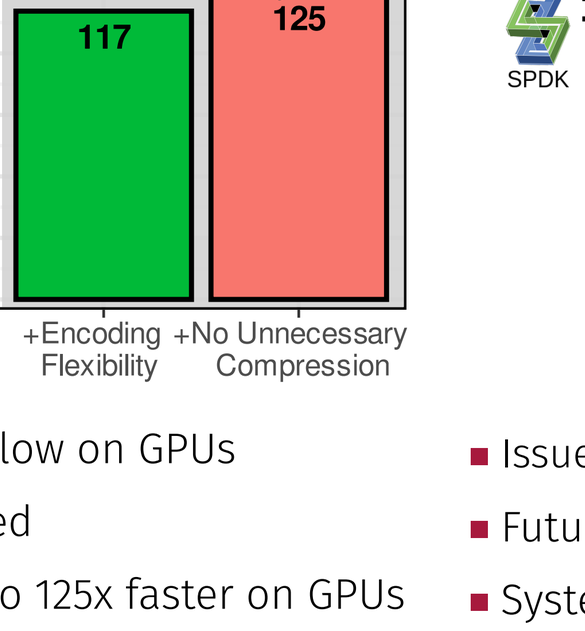
- GPUs idle during slow I/O
- Overlap compute and I/O for GPU util.



- Storage I/O ↔ decoding + decomp.
- Network I/O ↔ partition + hashjoin
- Storage + Network I/O ↔ QP

(↔ := Overlapping)

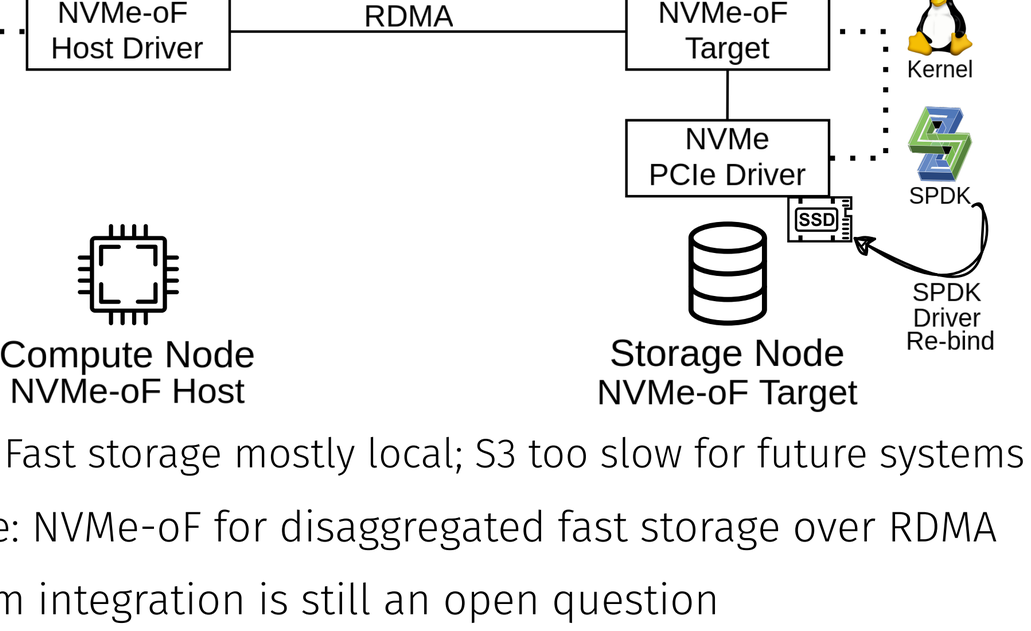
Storage:



- Faster and more scalable
- But also fewer OOMs

Research Impact: Merged in RAPIDS

Storage & Network with I/O Queue Coordination:



- Storage: base tables
- HBM: I/O Queues & operators
- Storage → Queues → Network → QP: Full overlap

(→ := Data Flow Path)

VLDB 26 & Datenbank-Spektrum 26

Storage File Format: Reading Parquet on GPUs

- Misconception: Parquet is inherently slow on GPUs
- Reality: Default config. is CPU-optimized
- Fix: GPU-aware reconfiguration → up to 125x faster on GPUs

Research Impact: Remove Parquet Bottleneck **DaMoN 26, Best Short Paper**

Remote Storage Protocol: NVMe Over Fabrics (NVMe-oF)

- Issue: Fast storage mostly local; S3 too slow for future systems
- Future: NVMe-oF for disaggregated fast storage over RDMA
- System integration is still an open question

BTW 25

Conclusion

- GPUs and disaggregation reshape the data systems stack
- Focus shifting from computation to I/O: storage & network
- Compute-I/O overlap is key to disaggregated system design

Future Work

- File Formats: Plenty of room for GPU-aware optimization
- NVMe-oF: Integration into our system *PystachIO*
- Workloads: Disaggregated GPU systems beyond OLAP or DB