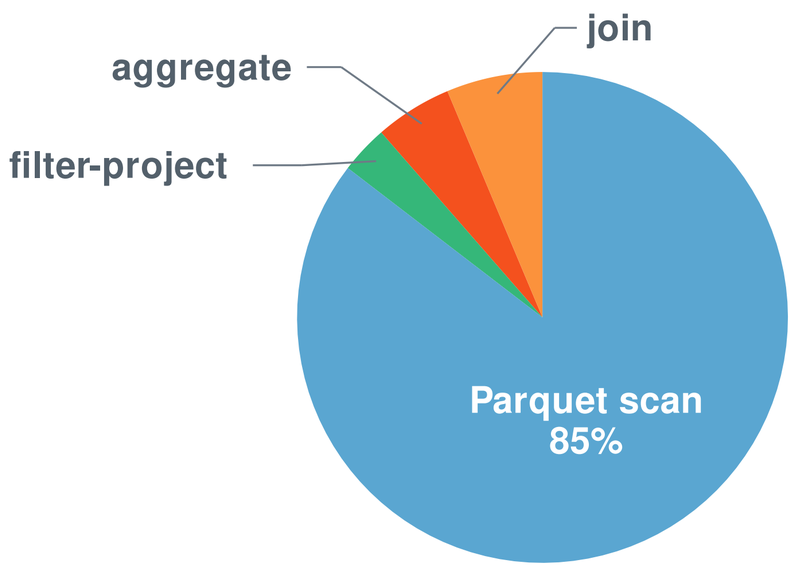
Do GPUs Really Need New Tabular File Formats?

Jigao Luo, Qi Chen, Carsten Binnig
Technische Universität Darmstadt

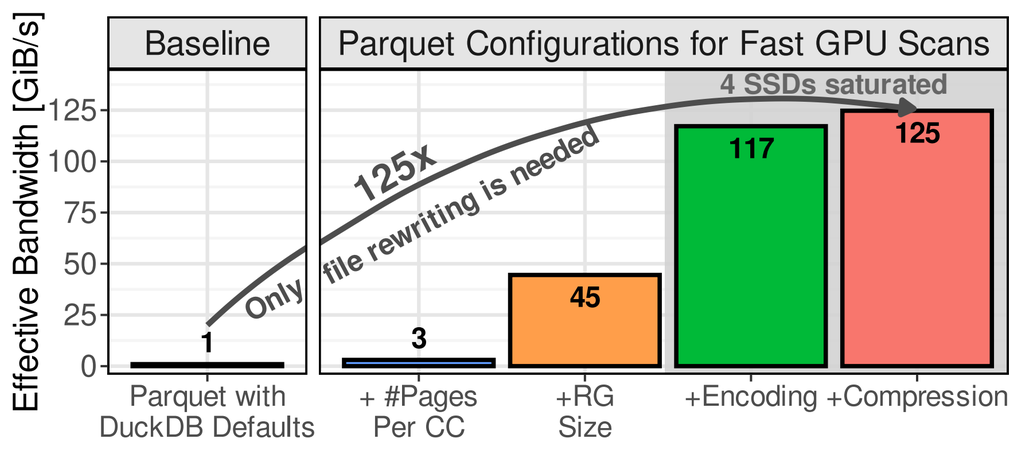
SYSTEMS

Motivation & Research Question



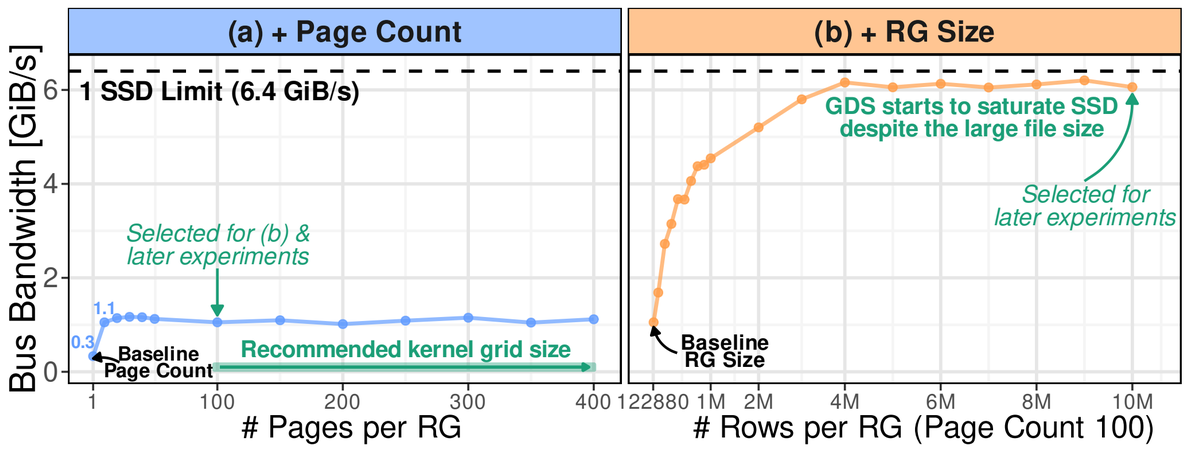
- Parquet: RowGroup→ColumnChunks→Pages (encoded&compressed)
- NVIDIA RAPIDS: **85%(!)** of TPC-H runtime in Parquet scan
- **Question: Do GPUs need new file formats?**

Idea & Contributions



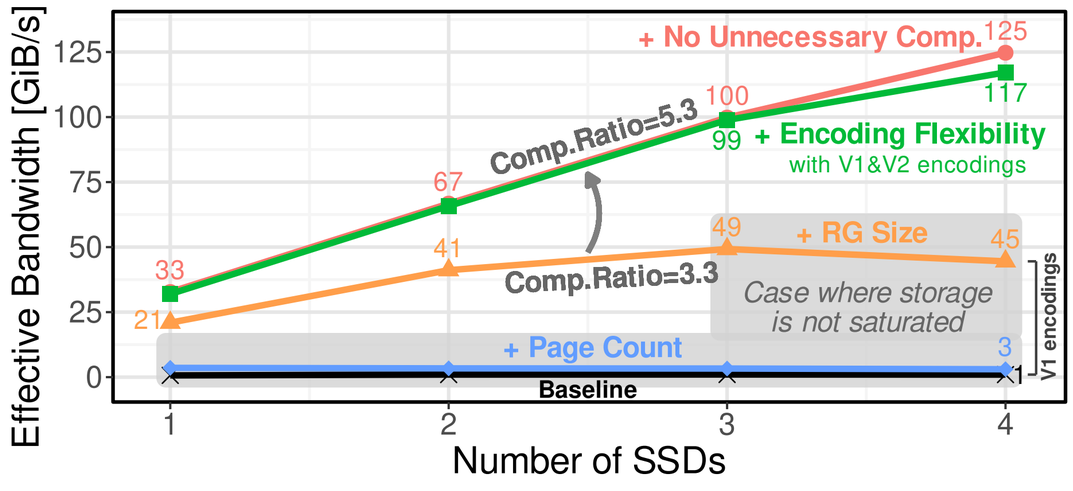
- Configuration Defaults: inherited from CPU best practices
- Issue: CPU defaults **!=** GPU-optimal
- Misconception: Parquet is often wrongly criticized
- Up to **125x** speedup after GPU-aware rewriting → 4 Insights

Insight 1 & 2: # Pages in CC & RG Size



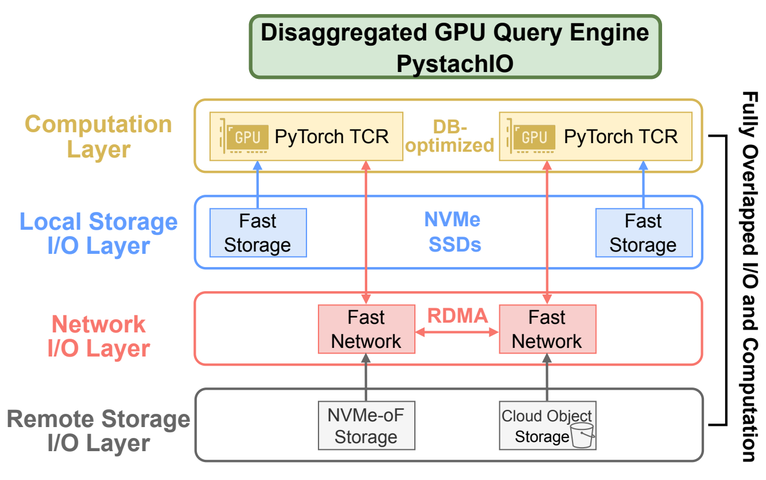
- # Pages in ColumnChunk → grid size of GPU decoding kernel
  - **Insight 1**: Avoid too few pages to prevent low GPU util.
- GPU I/O stack: MB-level I/O sizes saturate SSDs
  - **Insight 2**: Use million-row RG size for storage saturation

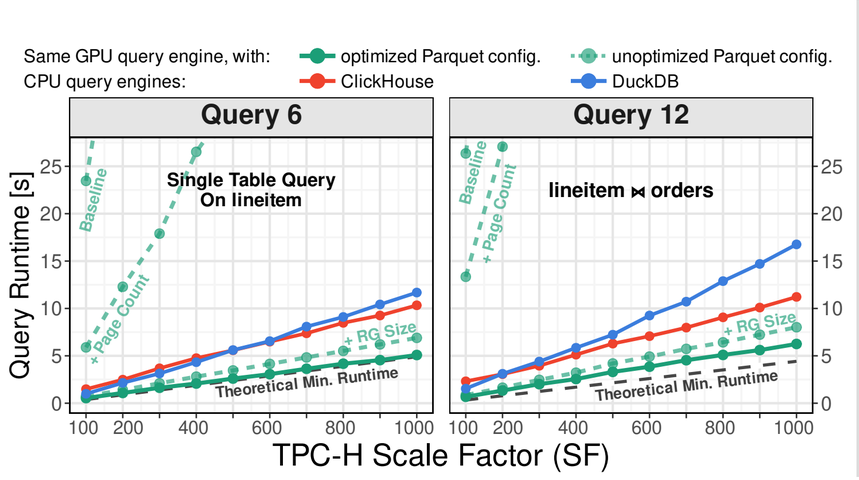
Insight 3 & 4: Encoding & Compression

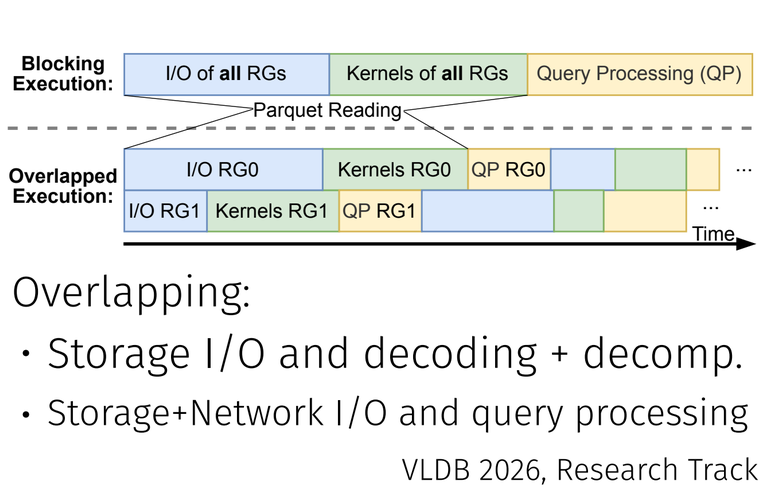


- Search for the min-size encoding per ColumnChunk
  - **Insight 3**: Flexible encoding improves compression ratio
- Some compression gives zero size reduction
  - **Insight 4**: Avoid these compression cases

More: PystachIO, Our Distributed Disaggregated GPU Query Engine





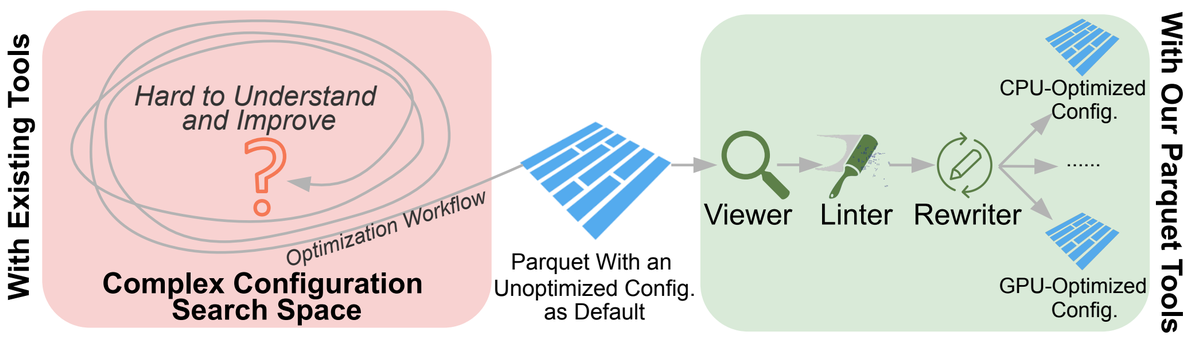


Overlapping:

- Storage I/O and decoding + decomp.
- Storage+Network I/O and query processing

VLDB 2026, Research Track

More: Automating Parquet Optimization Workflow



- Parquet tools:
  - Viewer: understanding configurations
  - Linter: gathering suggestions
  - Rewriter: applying suggested configurations
- To **new file formats**: Are there any tools available?

Conclusion & Takeaway

- GPU-aware Parquet configuration matters: 4 Insights
- Parquet performs well once optimized
- GPU-optimized config. **!=** CPU-optimized config.
- Understand & optimize Parquet **before** replacing it!

Paper

Rewriter Code

Personal Website