Efficient Disk-to-Disk Sorting: 
A Case Study in 
the Decoupled Execution Paradigm

Hassan Eslami, Anthony Kougkas, Maria Kotsifakou, Theodoros Kasampalis 
Kun Feng, Yin Lu, William Gropp, Xian-He Sun, Yong Chen, Rajeev Thakur

University of Illinois at Urbana-Champaign
Illinois Institute of Technology
Texas Tech University
Argonne National Laboratory

eslami2@illinois.edu
Challenges in data-intensive computing

• Excessive **data movement** from storage layers to compute nodes

• Proposal #1: using faster storage devices (SSDs, PCMs, etc.)
  
  • Still does not solve the IO-wall problem

• Proposal #2: moving some computational units to data
  
  • Often **limited capability** of computational units, **reduced flexibility** in programming
A promising solution: the Decoupled Execution Paradigm (DEP)
Contributions

• Investigate the benefits of the DEP architecture
• Propose an efficient algorithm for disk-to-disk sorting in DEP
• Analytically model the proposed sorting algorithm
Disk-to-disk sorting
Disk-to-disk sorting
Disk-to-disk sorting
Disk-to-disk sorting
Disk-to-disk sorting
Disk-to-disk sorting
Disk-to-disk sorting
Disk-to-disk sorting
Disk-to-disk sorting
Disk-to-disk sorting
Disk-to-disk sorting
Disk-to-disk sorting
Disk-to-disk sorting in the DEP architecture
Disk-to-disk sorting in the DEP architecture

Compute Nodes (CN)

Compute-side Data Nodes (CDN)

Storage-side Data Nodes (SDN)

Interconnect

Read
Disk-to-disk sorting in the DEP architecture

Interconnect

Storage-side Data Nodes (SDN)

Compute-side Data Nodes (CDN)

Compute Nodes (CN)

Read

Sort
Disk-to-disk sorting in the DEP architecture

Compute Nodes (CN)

Compute-side Data Nodes (CDN)

Interconnect

Storage-side Data Nodes (SDN)

Read
Sort
Split
Disk-to-disk sorting in the DEP architecture
Disk-to-disk sorting in the DEP architecture

Receive

Read
Sort
Split
Compress
Send

Compute Nodes (CN)
Compute-side Data Nodes (CDN)
Storage-side Data Nodes (SDN)

Interconnect
Disk-to-disk sorting in the DEP architecture

Receive
Decompress

Read
Sort
Split
Compress
Send

Compute Nodes (CN)

Compute-side Data Nodes (CDN)

Interconnect

Storage-side Data Nodes (SDN)
Disk-to-disk sorting in the DEP architecture

Receive
Decompress

Read
Sort
Split
Compress
Send

Compute Nodes (CN)

Compute-side Data Nodes (CDN)

Interconnect

Storage-side Data Nodes (SDN)
Disk-to-disk sorting in the DEP architecture

Receive
Decompress

Read
Sort
Split
Compress
Send

Interconnect

Compute Nodes (CN)

Compute-side Data Nodes (CDN)

Storage-side Data Nodes (SDN)
Disk-to-disk sorting in the DEP architecture

Receive
Decompress
Merge

Read
Sort
Split
Compress
Send

Compute Nodes
(CN)

Compute-side
Data Nodes
(CDN)

Storage-side
Data Nodes
(SDN)

Interconnect
Disk-to-disk sorting in the DEP architecture

Receive
Decompress
Merge

Read
Sort
Split
Compress
Send

Compute Nodes (CN)

Interconnect

Compute-side Data Nodes (CDN)

Storage-side Data Nodes (SDN)
Disk-to-disk sorting in the DEP architecture

Receive
Decompress
Merge

Compress

Read
Sort
Split
Compress
Send

Compute Nodes (CN)

Compute-side Data Nodes (CDN)

Interconnect

Storage-side Data Nodes (SDN)
Disk-to-disk sorting in the DEP architecture

- Compute Nodes (CN)
- Compute-side Data Nodes (CDN)
- Storage-side Data Nodes (SDN)

Flow:
- Receive
- Decompress
- Merge
- Compress
- Send
- Read
- Sort
- Split
- Compress
- Send

Interconnect
Disk-to-disk sorting in the DEP architecture

- Compute Nodes (CN)
- Compute-side Data Nodes (CDN)
- Storage-side Data Nodes (SDN)
- Interconnect

Steps:
- Receive
- Decompress
- Merge
- Compress
- Send
- Receive
- Store
- Read
- Sort
- Split
- Compress
- Send
Disk-to-disk sorting in the DEP architecture

Barrier

Compute Nodes (CN)

Compute-side Data Nodes (CDN)

Interconnect

Storage-side Data Nodes (SDN)
Disk-to-disk sorting in the DEP architecture

Read

Serve CN requests

Compute Nodes (CN)

Compute-side Data Nodes (CDN)

Interconnect

Storage-side Data Nodes (SDN)
Disk-to-disk sorting in the DEP architecture

Read
Decompress

Serve CN requests

Compute Nodes (CN)

Compute-side Data Nodes (CDN)

Storage-side Data Nodes (SDN)
Disk-to-disk sorting in the DEP architecture

Read
Decompress
Merge

Serve CN requests

Compute Nodes (CN)
Compute-side Data Nodes (CDN)
Storage-side Data Nodes (SDN)
Disk-to-disk sorting in the DEP architecture

Read
Decompress
Merge

Compress

Serve CN requests

Compute Nodes (CN)

Compute-side Data Nodes (CDN)

Interconnect

Storage-side Data Nodes (SDN)
Disk-to-disk sorting in the DEP architecture
Disk-to-disk sorting in the DEP architecture

- Compute Nodes (CN)
- Compute-side Data Nodes (CDN)
- Storage-side Data Nodes (SDN)
- Interconnect

Read
Decompress
Merge

Compress
Send

Serve CN requests

Receive
Decompress
Disk-to-disk sorting in the DEP architecture
<table>
<thead>
<tr>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read Phase</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Barrier</strong></td>
</tr>
<tr>
<td><strong>Write Phase</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Compute Nodes (CN)</strong></td>
</tr>
<tr>
<td><strong>Compute-side Data Nodes (CDN)</strong></td>
</tr>
<tr>
<td><strong>Storage-side Data Nodes (SDN)</strong></td>
</tr>
</tbody>
</table>
Evaluation

Switch 0

Switch 1

Switch 2

22 Nodes

21 Nodes

21 Nodes
Evaluation
## Micro-benchmarks

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>PFS Latency</td>
<td>2.8 ms</td>
</tr>
<tr>
<td>PFS BW</td>
<td>111/162 MiB/s</td>
</tr>
<tr>
<td>SSD Latency</td>
<td>5 us</td>
</tr>
<tr>
<td>SSD BW</td>
<td>368 MiB/s</td>
</tr>
<tr>
<td>Memory BW</td>
<td>2.8 GiB/s</td>
</tr>
<tr>
<td>Network BW</td>
<td>10.9 MiB/s</td>
</tr>
<tr>
<td>Compression BW</td>
<td>520 MiB/s</td>
</tr>
<tr>
<td>Decompression BW</td>
<td>604 MiB/s</td>
</tr>
</tbody>
</table>
Evaluation of the model

Effect of Number of CDNs and SDNs (with 4CNs)
Weak scalability

![Graph showing total execution time vs. number of nodes (CNs+CDNs+SDNs)]

- **Actual**
- **Model**

- Total Execution Time (s)
- Number of Nodes (CNs+CDNs+SDNs)
Throughput

Number of Nodes (CNs+CDNs+SDNs) vs Sort Rate (MiB/s)

- Actual
- Model
- DEP Optimal
- Non-DEP Optimal
Conclusion

• The Decoupled Execution Paradigm helps in reducing the data movement, hence increasing the performance of disk-to-disk sorting

• Our disk-to-disk algorithm…
  • … exploits programmable units closer to file system
  • … compresses data before sending it over network or writing it to storage
  • … overlaps most of computation and communication with IO
Thank You!

eslami2@illinois.edu
github/heslami/dep-sort