Distributed Text Index Construction

Patrick Dinklage, Johannes Fischer, Florian Kurpicz

Department of Computer Science
Chair 11: Algorithm Engineering
Algorithmic Foundations and Education in Computer Science
Distributed Text Index Construction

Patrick Dinklage, Johannes Fischer, Florian Kurpicz

Department of Computer Science
Chair 11: Algorithm Engineering
Algorithmic Foundations and Education in Computer Science
Distributed Text Index Construction

Patrick Dinklage, Johannes Fischer, Florian Kurpicz

Department of Computer Science
Chair 11: Algorithm Engineering
Algorithmic Foundations and Education in Computer Science
Distributed Text Index Construction

Patrick Dinklage, Johannes Fischer, Florian Kurpicz

Department of Computer Science
Chair 11: Algorithm Engineering
Algorithmic Foundations and Education in Computer Science
DISTRIBUTED (COMPUTING)

In Practice

- Synchronization is expensive
- Load-balancing is important
- Communication & memory efficiency

(by courtesy of lido.tu-dortmund.de)
but what if?
but what if?
but what if?
but what if?

<table>
<thead>
<tr>
<th>Words</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1, #3, #7, ?...</td>
<td>#1, #3, #7, ?...</td>
</tr>
<tr>
<td>#2, #3, ?...</td>
<td>#2, #3, ?...</td>
</tr>
<tr>
<td>#2, #4, #5, ?...</td>
<td>#2, #4, #5, ?...</td>
</tr>
<tr>
<td>#1, #240, ?...</td>
<td>#1, #240, ?...</td>
</tr>
<tr>
<td>#3, #7, #9, ?...</td>
<td>#3, #7, #9, ?...</td>
</tr>
<tr>
<td>#4, #5, #6, ?...</td>
<td>#4, #5, #6, ?...</td>
</tr>
</tbody>
</table>
... but what if?
**SUFFIX ARRAYS & WAVELET TREES**

- Sorting suffixes?
- Lots of data dependencies?
- /kurpicz/dssss

- Working on bits?
- Sorting is too much?
- /pdinklag/distwt
Evaluated suffix array construction on the LiDO3 cluster.

Weak scaling experiments (only DNA) with 90 MB per PE.

![Graph showing throughput and memory usage for different number of PEs.]

- Throughput in MB/s
- Memory usage in GB/20p
- PEs range from 1 to 64
- Lines represent different algorithms:
  - dDivSufSort
  - dPD
  - PSAC
  - DC3/7/13
Evaluated wavelet tree construction on the LiDO3 cluster.
Weak scaling experiments (only DNA) with 53 MB per PE.
Evaluated wavelet tree construction on the LiDO3 cluster.

Weak scaling experiments (only DNA) with 53 MB per PE.

Throughput [GiB/s] vs. PEs [20 \cdot p]

Memory [GiB/2op] vs. PEs [20 \cdot p]

- bsort
- dd
- dsplit
- dynbsort

Thank You