Database Systems on Modern CPU Architectures

Philipp Fent, André Kohn
Database Systems on Modern CPU Architectures
April 23, 2019
External Sort

Problem:

- Sorting an arbitrary amount of data, stored on disk
External Sort

Problem:

• Sorting an arbitrary amount of data, stored on disk
• Accessing disk is slow – so we do not want to access each value individually

Solution:

• Load pieces (called “runs”) of the data into main memory
• and sort them

With $m$ values fitting into main memory and $d$ values that should be sorted:

$$k = \lceil \frac{d}{m} \rceil$$

sorted runs
External Sort

Problem:
- Sorting an arbitrary amount of data, stored on disk
- Accessing disk is slow – so we do not want to access each value individually
- Sorting in main memory is fast

Solution:
- Load pieces (called “runs”) of the data into main memory and sort them
- With \( m \) values fitting into main memory and \( d \) values that should be sorted:
  \[ k = \lceil \frac{d}{m} \rceil \text{ sorted runs} \]
External Sort

Problem:

- Sorting an arbitrary amount of data, stored on disk
- Accessing disk is slow – so we do not want to access each value individually
- Sorting in main memory is fast – but main memory size is limited
External Sort

Problem:
- Sorting an arbitrary amount of data, stored on disk
- Accessing disk is slow – so we do not want to access each value individually
- Sorting in main memory is fast – but main memory size is limited

Solution:
- Load pieces (called “runs”) of the data into main memory
External Sort

Problem:
- Sorting an arbitrary amount of data, stored on disk
- Accessing disk is slow – so we do not want to access each value individually
- Sorting in main memory is fast – but main memory size is limited

Solution:
- Load pieces (called “runs”) of the data into main memory
- and sort them
External Sort

Problem:
- Sorting an arbitrary amount of data, stored on disk
- Accessing disk is slow – so we do not want to access each value individually
- Sorting in main memory is fast – but main memory size is limited

Solution:
- Load pieces (called “runs”) of the data into main memory
- and sort them
- With $m$ values fitting into main memory and $d$ values that should be sorted:
  $$k = \lceil \frac{d}{m} \rceil$$ sorted runs
Sort $k$ runs

Memory

Disk

\[=\text{unsorted}\]
Sort $k$ runs

Memory

Disk

= unsorted
Sort $k$ runs

Memory

Disk

\text{std::sort()}

= unsorted
Sort $k$ runs

Memory

Disk

= unsorted
Sort $k$ runs

Memory

Disk

= unsorted
Sort $k$ runs

Memory

\texttt{std::sort()}

Disk

\[
\begin{array}{c}
\text{\color{lightgray} = unsorted} \\
\end{array}
\]
Sort $k$ runs

Memory

Disk

\[\text{unsorted}\]
Sort $k$ runs

Memory

Disk

= unsorted
Sort $k$ runs

Memory

std::sort()

Disk

= unsorted
Sort $k$ runs

Memory

Disk
Sort $k$ runs
Merge

Memory

Disk
Merge

Memory

Disk
Merge
Merge

Memory

Disk
Merge

Memory

Disk
Merge
Merge

Memory

Disk
Merge

Memory

Disk
Merge

- Produces sorted result after $\log k$ phases
- Each merge phase reads the whole input data from disk
- Still expensive
$k$-way merge

Memory

Disk
$k$-way merge
$k$-way merge

Memory

Disk
$k$-way merge
$k$-way merge

Memory

Disk
$k$-way merge

Memory

Disk
Longer runs

- Minimal disk reads
- $k$-way merge might not fit memory
- Fall back to regular merge
- Produce longer runs with replacement selection
Replacement selection

Memory

Disk
Replacement selection

insert 98

Memory

98

Disk
Replacement selection

insert 74

Memory

74 98

Disk
Replacement selection

insert 59

Memory

59 74 98

Disk
Replacement selection

insert 76

Memory

59 74 76 98

Disk
Replacement selection

insert 38

Memory

| 38 | 59 | 74 | 76 | 98 |

Disk
Replacement selection

insert 66

| Memory | 38 59 66 74 76 98 |

| Disk |
Replacement selection

insert 10

<table>
<thead>
<tr>
<th>Memory</th>
<th>10 38 59 66 74 76 98</th>
</tr>
</thead>
</table>

Disk
Replacement selection

insert 34

| Memory | 10 | 38 | 59 | 66 | 74 | 76 | 98 |

Disk
Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>34 38 59 66 74 76 98</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk</td>
<td>10</td>
</tr>
</tbody>
</table>
Replacement selection

insert 47

<table>
<thead>
<tr>
<th>Memory</th>
<th>34 38 59 66 74 76 98</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Disk</th>
<th>10</th>
</tr>
</thead>
</table>
Replacement selection

| Memory | 38 47 59 66 74 76 98 |
| Disk   | 10 34 |
Replacement selection

insert 53

<table>
<thead>
<tr>
<th>Memory</th>
<th>38</th>
<th>47</th>
<th>59</th>
<th>66</th>
<th>74</th>
<th>76</th>
<th>98</th>
</tr>
</thead>
</table>

| Disk   | 10 | 34 |
Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>47</th>
<th>53</th>
<th>59</th>
<th>66</th>
<th>74</th>
<th>76</th>
<th>98</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk</td>
<td>10</td>
<td>34</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Replacement selection

**Memory**

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>47</td>
<td>53</td>
<td>59</td>
<td>66</td>
<td>74</td>
<td>76</td>
</tr>
</tbody>
</table>

**Disk**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>34</td>
<td>38</td>
</tr>
</tbody>
</table>

insert 86
Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>53</th>
<th>59</th>
<th>66</th>
<th>74</th>
<th>76</th>
<th>86</th>
<th>98</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk</td>
<td>10</td>
<td>34</td>
<td>38</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Replacement selection

insert 84

<table>
<thead>
<tr>
<th>Memory</th>
<th>53</th>
<th>59</th>
<th>66</th>
<th>74</th>
<th>76</th>
<th>86</th>
<th>98</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk</td>
<td>10</td>
<td>34</td>
<td>38</td>
<td>47</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>59</th>
<th>66</th>
<th>74</th>
<th>76</th>
<th>84</th>
<th>86</th>
<th>98</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk</td>
<td>10</td>
<td>34</td>
<td>38</td>
<td>47</td>
<td>53</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>59</th>
<th>66</th>
<th>74</th>
<th>76</th>
<th>84</th>
<th>86</th>
<th>98</th>
</tr>
</thead>
</table>

| Disk    | 10  | 34  | 38  | 47  | 53  |

insert 19
## Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>Disk</th>
</tr>
</thead>
<tbody>
<tr>
<td>66 74 76 84 86 98 19</td>
<td>10 34 38 47 53 59</td>
</tr>
</tbody>
</table>
Replacement selection

insert 30

<table>
<thead>
<tr>
<th>Memory</th>
<th>66</th>
<th>74</th>
<th>76</th>
<th>84</th>
<th>86</th>
<th>98</th>
<th>19</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Disk</th>
<th>10</th>
<th>34</th>
<th>38</th>
<th>47</th>
<th>53</th>
<th>59</th>
</tr>
</thead>
</table>
Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>74</th>
<th>76</th>
<th>84</th>
<th>86</th>
<th>98</th>
<th>19</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk</td>
<td>10</td>
<td>34</td>
<td>38</td>
<td>47</td>
<td>53</td>
<td>59</td>
<td>66</td>
</tr>
</tbody>
</table>
Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>74 76 84 86 98 19 30</td>
</tr>
</tbody>
</table>

insert 63

<table>
<thead>
<tr>
<th>Disk</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10 34 38 47 53 59 66</td>
</tr>
</tbody>
</table>
Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>76</th>
<th>84</th>
<th>86</th>
<th>98</th>
<th>19</th>
<th>30</th>
<th>63</th>
</tr>
</thead>
</table>

| Disk    | 10 | 34 | 38 | 47 | 53 | 59 | 66 | 74 |
Replacement selection

insert 83

<table>
<thead>
<tr>
<th>Memory</th>
<th>76</th>
<th>84</th>
<th>86</th>
<th>98</th>
<th>19</th>
<th>30</th>
<th>63</th>
</tr>
</thead>
</table>

| Disk   | 10  | 34  | 38  | 47  | 53  | 59  | 66  | 74  |
Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>83</th>
<th>84</th>
<th>86</th>
<th>98</th>
<th>19</th>
<th>30</th>
<th>63</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk</td>
<td>10</td>
<td>34</td>
<td>38</td>
<td>47</td>
<td>53</td>
<td>59</td>
<td>66</td>
</tr>
</tbody>
</table>
Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>83</th>
<th>84</th>
<th>86</th>
<th>98</th>
<th>19</th>
<th>30</th>
<th>63</th>
</tr>
</thead>
</table>

| Disk   | 10  | 34  | 38  | 47  | 53  | 59  | 66  | 74  | 76  |

insert 37
Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>84</th>
<th>86</th>
<th>98</th>
<th>19</th>
<th>30</th>
<th>37</th>
<th>63</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk</td>
<td>10</td>
<td>34</td>
<td>38</td>
<td>47</td>
<td>53</td>
<td>59</td>
<td>66</td>
</tr>
</tbody>
</table>
Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>84</th>
<th>86</th>
<th>98</th>
<th>19</th>
<th>30</th>
<th>37</th>
<th>63</th>
</tr>
</thead>
</table>

| Disk   | 10  | 34  | 38  | 47  | 53  | 59  | 66  | 74  | 76  | 83  |

insert 65
Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>86 98 19 30 37 63 65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk</td>
<td>10 34 38 47 53 59 66 74 76 83 84</td>
</tr>
</tbody>
</table>
Replacement selection

Insert 29

<table>
<thead>
<tr>
<th>Memory</th>
<th>86</th>
<th>98</th>
<th>19</th>
<th>30</th>
<th>37</th>
<th>63</th>
<th>65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk</td>
<td>10</td>
<td>34</td>
<td>38</td>
<td>47</td>
<td>53</td>
<td>59</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>74</td>
<td>76</td>
<td>83</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>98</th>
<th>19</th>
<th>29</th>
<th>30</th>
<th>37</th>
<th>63</th>
<th>65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk</td>
<td>10</td>
<td>34</td>
<td>38</td>
<td>47</td>
<td>53</td>
<td>59</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>74</td>
<td>76</td>
<td>83</td>
<td>84</td>
<td>86</td>
<td>95</td>
<td></td>
</tr>
</tbody>
</table>
Replacement selection

insert 45

<table>
<thead>
<tr>
<th>Memory</th>
<th>98</th>
<th>19</th>
<th>29</th>
<th>30</th>
<th>37</th>
<th>63</th>
<th>65</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Disk</th>
<th>10</th>
<th>34</th>
<th>38</th>
<th>47</th>
<th>53</th>
<th>59</th>
<th>66</th>
<th>74</th>
<th>76</th>
<th>83</th>
<th>84</th>
<th>86</th>
<th>95</th>
</tr>
</thead>
</table>
Replacement selection

<table>
<thead>
<tr>
<th>Memory</th>
<th>19</th>
<th>29</th>
<th>30</th>
<th>37</th>
<th>45</th>
<th>63</th>
<th>65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disk</td>
<td>10</td>
<td>34</td>
<td>38</td>
<td>47</td>
<td>53</td>
<td>59</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>74</td>
<td>76</td>
<td>83</td>
<td>84</td>
<td>86</td>
<td>95</td>
<td>98</td>
</tr>
</tbody>
</table>