Copyright ©1999 Timothy Howard Merrett
Permission to make digital or hard copies of part or all of
this work for personal or classroom use is granted without fee
provided that copies are not made or distributed for profit or
commercial advantage and that copies bear this notice and
full citation in a prominent place. Copyright for components
of this work owned by others than T. H. Merrett must be
honoured. Abstracting with credit is permitted. To copy
otherwise, to republish, to post on servers, or to redistribute
to lists, requires prior specific permission and/or fee. Re-
quest permission to republish from: T. H. Merrett, School
of Computer Science, McGill University, fax 514 398 3883.

The author gratefully acknowledges support from the tax-
payers of Québec and of Canada who have paid his salary
and research grants while this work was developed at McGill
University, and from his students (who built the implement-
ations and investigated the data structures and algorithms)
and their funding agencies.

T. H. Merrett

©99/11
The “Curse of Dimensionality”

(OLAP, Feature Vectors, ..)

What happens to small activities in many dimensions?

Say $a = 0.0001 = \frac{1}{10,000}$
Say $f = 2$ for each dimension.

In 1-D effective activity is 0.5:

$$\frac{1}{10,000}$$

T. H. Merrett
In 2-D effective activity is 0.25: \( \frac{1}{100} \times \frac{1}{100} \)

In 4-D effective activity is 0.0625: \( \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \times \frac{1}{10} \)

In 16-D effective activity is 1!: 
\[
0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56 \\
0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56 \times 0.56
\]

Note that \( a = 0.0001 \) is a breakeven activity, e.g., for \( R = 100, \rho = 1,000,000 \). Any \( a_{eff} \) over this means use sequential!

T. H. Merrett
Above assumes

1. The range query has same *shape* as the data space.

2. \( f_i = f \) and space is hypercube of side 1.

3. The data distribution is the product of the axial distributions.

Can be calculated generally using “fractional ceiling”,

\[
\text{ceil}(f, x) = g/f, \text{ where } 0 \leq (g - 1)/f < x \leq g/f \leq 1:
\]

\[
a_{\text{eff}} = (\text{ceil}(f, a^{1/d}))^d
\]

T. H. Merrett
Activity blowup:
Applies to any $d$-dim. paging that partitions the axes. Assumes (1) data distribution is Cartesian product, (2) range query, space are hypercubes.

<table>
<thead>
<tr>
<th>d</th>
<th>$a^{1/d}$</th>
<th>n</th>
<th>$a_{\text{eff}}$</th>
<th>n</th>
<th>$a_{\text{eff}}$</th>
<th>n</th>
<th>$a_{\text{eff}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.0001</td>
<td>2</td>
<td>.5</td>
<td>5</td>
<td>.2</td>
<td>10</td>
<td>.1</td>
</tr>
<tr>
<td>2</td>
<td>.01</td>
<td>4</td>
<td>.25</td>
<td>25</td>
<td>.04</td>
<td>100</td>
<td>.01</td>
</tr>
<tr>
<td>4</td>
<td>.1</td>
<td>16</td>
<td>.06</td>
<td>625</td>
<td>.002</td>
<td>1004</td>
<td>.0001</td>
</tr>
<tr>
<td>8</td>
<td>.31</td>
<td>256</td>
<td>.004</td>
<td>3.9105</td>
<td>.0007</td>
<td>1008</td>
<td>.0007</td>
</tr>
<tr>
<td>16</td>
<td>.56</td>
<td>64K</td>
<td>1</td>
<td>1.51011</td>
<td>.0003</td>
<td>1008</td>
<td>.0003</td>
</tr>
<tr>
<td>32</td>
<td>.75</td>
<td>4.3109</td>
<td>1</td>
<td>.0008</td>
<td></td>
<td>.0008</td>
<td></td>
</tr>
<tr>
<td>64</td>
<td>.87</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>.001</td>
<td></td>
</tr>
<tr>
<td>128</td>
<td>.93</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>.96</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>512</td>
<td>.98</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>1024</td>
<td>.99</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

N.B. $f = \infty$ (or every field is key): $a_{\text{eff}} = a$
$a$: activity; $a_{\text{eff}}$: effective activity due to paging;
$f$: number of page partitions per axis;
n: number of pages.

T. H. Merrett ©99/2
This is a danger (given the three assumptions) for any method involving multidimensional grids.

But not for trees. E.g., kd-tries are tries are one-dimensional.
Experimental Results on Data Dimensions

![Graph showing average #pages accessed vs. dimension for R*-trees, X-trees, and zoom tries. The X-axis represents the dimension, and the Y-axis represents the average #pages accessed. The graph includes a legend indicating the different tree types.]

Exact match: file size=6.4M bytes, BLK_SIZE=4096 bytes

X. Y. Zhao
Range query: file size=6.4MB, BLK_SIZE=4096B, activity=0.2, uniform data

- R*–trees
- X–trees
- zoom tries